UNIVERSIDADE FEEVALE

WILLIAM JACKSON DA COSTA

INCADATABR: UMA BIBLIOTECA EM R PARA MANIPULAÇÃO DE DATASETS DO INCA

Novo Hamburgo

2019

WILLIAM JACKSON DA COSTA

INCADATABR: UMA BIBLIOTECA EM R PARA MANIPULAÇÃO DE DATASETS DO INCA

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do grau de Bacharel em Ciência da Computação pela Universidade Feevale

Orientador: Dr. Juliano Varella de Carvalho

Novo Hamburgo

2019

AGRADECIMENTOS

Gostaria de agradecer a todos os que, de alguma maneira, contribuíram para a realização desse trabalho de conclusão, em especial:

Ao meu noivo que conviveu comigo diariamente, minha gratidão pelo apoio emocional e paciência nos períodos mais difíceis do trabalho.

Aos meus pais, por sempre me incentivarem e não me deixarem desistir desta longa jornada e que não mediram esforços para que eu chegasse até esta etapa da minha vida.

Ao meu orientador, por toda a paciência e dedicação em suas correções e por todo o incentivo no decorrer de todo o trabalho.

RESUMO

De acordo com os dados divulgados em 2017 pela Organização Mundial da Saúde (OMS), a cada ano 8,8 milhões de pessoas morrem em decorrência do câncer. Devido a isso, diversas entidades ao redor do mundo realizam a coleta, processamento e distribuição dos dados relacionados aos casos de câncer com o objetivo de promover estudos e pesquisas sobre a doença. No Brasil, embora os dados do Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA) sejam de domínio público, os pesquisadores encontram algumas dificuldades na análise e processamento destes. Entre estas, destacam-se a dificuldade em realizar o download dos arquivos e a necessidade de usar um aplicativo exclusivo para *Windows* desenvolvido pelo Departamento de Informática do Sistema Único de Saúde do Brasil (DATASUS). Devido a estas dificuldades, a biblioteca desenvolvida neste trabalho tem como objetivo facilitar a interação de profissionais da área da saúde e Tecnologia da Informação (TI), que possuam conhecimento básico de programação, com os Registros Hospitalares de Câncer (RHC), um dos *datasets* disponibilizados pelo INCA. Este *dataset* é composto por 40 variáveis e dispõe de registros de casos de câncer entre os anos de 1985 e 2017. A biblioteca facilita o processo de importação, exportação e interação dos dados fornecidos pelo INCA, com o intuito de auxiliar no processo de análise dos dados. Além disso, a biblioteca permite a geração de gráficos de forma dinâmica.

Palavras-chave: Registro de casos de câncer. Mineração de dados. Base de dados INCA. Biblioteca na linguagem R. Registros Hospitalares de Câncer.

ABSTRACT

According to data released in 2017 by the World Health Organization (WHO), each year 8.8 million people die from cancer. For this reason, several entities in the world a process conduct collection, analyzing and controlling cancer cases with the goal of promoting studies and research on a disease. In Brazil, although *Instituto Nacional de Câncer José Alencar Gomes da Silva* (INCA) data are in the public domain, researchers find some difficulties in analyzing and processing them. Among these difficulties are the download of files and the need to use a exclusive Windows application developed by *Departamento de Informática do Sistema Único de Saúde do Brasil* (DATASUS). Due to these difficulties, the library developed in this work aims to facilitate the interaction of professionals in the area of health and information technology, who have basic programming knowledge, with the Hospital Registry of Cancer (RHC), one of the datasets made available by INCA. This dataset is composed of 40 variables and has records of cancer cases between the years of 1985 and 2017. The library facilitates the process of importing, exporting and interacting the data provided by INCA, with the aim of assist in the process of data analysis. In addition, the library allows the generation of graphics in a dynamic way.

Keywords: Registry of cancer cases. Data Mining. INCA database. Library in the R language. Hospital Based Cancer Registries

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

AIDS	*Acquired Immunodeficiency Syndrome*
CACON	Centros de Assistência de Alta Complexidade em Oncologia
CBO	Classificação Brasileira de Ocupações
CETED	Centro de Tecnologias Digitais
CID-O	Classificação Internacional de Doenças para Oncologia
CONPREV	Coordenação de Prevenção e Vigilância
CRAN	*Comprehensive R Archive Network*
DATASUS	Departamento de Informática do Sistema Único de Saúde do Brasil
dbf	*dBase database file*
HIV	*Human Immunodeficiency Virus*
IARC	*International Agency for Research on Cancer*
IBGE	Instituto Brasileiro de Geografia e Estatística
IDE	*Integrated Development Environment*
INCA	Instituto Nacional de câncer José Alencar Gomes da Silva
LSI	Laboratório de Sistemas Integráveis
OMS	Organização Mundial da Saúde
ONU	Organização das Nações Unidas
RCBP	Registros de câncer de Base Populacional
RHC	Registros Hospitalares de câncer
SQL	*Structured Query Language*
SUS	Sistema Único de Saúde
TI	Tecnologia da Informação
TNM	*TNM Classification of Malignant Tumours*
UBS	Unidades Básicas de Saúde
UF	Unidade Federativa
UICC	União Internacional Contra o câncer
UNACON	Unidades de Assistência de Alta Complexidade em

Oncologia

WHO *World Health Organization*

SUMÁRIO

1 INTRODUÇÃO

Atualmente as aplicações computacionais tornaram-se parte integrante de nosso cotidiano. Alinhado a isso, o armazenamento de dados aumentou de forma considerável. Camilo e Silva (2009) explicam que a redução no custo de armazenamento evidenciou esta tendência presente desde o surgimento de sistemas computacionais: o armazenamento de dados como um dos principais objetivos das instituições que fazem uso de softwares de gestão financeira, gestão de conhecimento, entre outros.

Este avanço, aliado a utilização de ferramentas que automatizam a coleta de dados, permitiu que as instituições passassem a acumular grandes volumes de dados de diferentes tipos: arquivos de texto, planilhas, bancos de dados, *data warehouse*, entre outros. Da Costa Côrtes, Porcaro e Lifschitz (2002) em seu estudo sobre mineração de dados, descrevem as instituições como extremamente eficientes em capturar, organizar e armazenar grandes quantidades de dados, em suas operações diárias ou pesquisas científicas.

Estas instituições perceberam que a velocidade de coleta é superior a velocidade de análise e processamento dos dados coletados. Cardoso e Machado (2008) apontam que este atraso na análise gera um problema e uma contradição, pois as organizações por possuírem uma grande quantidade de dados, têm uma falsa sensação de que estão bem informadas. Em contrapartida, o atraso na análise dos dados faz com que a informação e o conhecimento extraídos destes já não se adéquem à realidade da instituição. Em um cenário ainda pior, não se extrai nenhuma informação útil dos dados coletados. Larose e Larose (2014) afirmam que o problema hoje não está na falta de informações suficientes, nem na ineficiência da transmissão destas.

> *Estamos, de fato, inundados com dados na maioria dos campos. Em vez disso, o problema é que não há suficientes analistas humanos treinados disponíveis que são habilidosos em traduzir todos esses dados para o conhecimento. (LAROSE, 2005, v. 2, p. 14, tradução nossa).*

Neste contexto é importante saber diferenciar dados, informação e conhecimento. De acordo com da Silva, Peres e Boscarioli (2016), considera-se como um dado, todo valor documentado ou obtido a partir de alguma medição; a informação

é gerada quando um sentido semântico ou significado é atribuído aos dados; e o conhecimento, por sua vez, é atingido quando for possível tomar decisões baseadas nas informações adquiridas.

Na área da saúde esse cenário não é diferente. Com a crescente adoção de sistemas informatizados por Unidades Básicas de Saúde (UBS) e hospitais, o volume de dados cresceu consideravelmente. De acordo com o DATASUS (2017), até o final de 2018, 100% das UBS irão aderir ao prontuário eletrônico. Com isso, uma grande quantidade de dados passará a ser registrada diariamente.

Todos esses dados são agrupados em bancos, de acordo com sua relevância a uma determinada área e servem como base para análises de qualidade dos serviços de saúde e auxiliam na tomada de decisão.

> *[...] os sistemas de informação de serviços de saúde são aqueles cujo propósito é selecionar os dados pertinentes a esses serviços e transformá-los na informação necessária para o processo de decisões, próprio das organizações e indivíduos que planejam, financiam, administram, proveem, medem e avaliam os serviços de saúde. (MORAES, 1994, p. 26 apud THAINES et al, 2009, p. 2).*

Dentre estas diversas bases de dados, o Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA), através do Programa de Epidemiologia e Vigilância do câncer e seus Fatores de Risco, disponibiliza informações relacionadas aos casos de câncer registrados no Brasil, que são agrupados com diferentes propósitos e particularidades em *datasets*. Essas informações permitem análises sobre a incidência e mortalidade por câncer no Brasil, e garantem o conhecimento detalhado sobre a situação da doença no País.

> *Nesse contexto, a vigilância de câncer torna-se um componente estratégico para o planejamento eficiente e efetivo dos programas de controle de câncer. A vigilância de câncer no Brasil é fundamentalmente baseada a partir de informações dos Registros de Câncer de Base Populacional (RCBP), os quais fornecem informações sobre o perfil da incidência de câncer em residentes de determinada área geográfica coberta pelo RCBP e dos Registros Hospitalares de câncer (RHC). (PINTO et al., 2012, p. 114).*

São disponibilizados pelo INCA os Registros Hospitalares de Câncer (RHC), que são de domínio público e podem ser baixados diretamente no *site* da instituição, sem a necessidade de um cadastro. O Registro de Câncer de Base Populacional

(RCBP) é disponibilizado para consulta no site da instituição, mas para realizar o *download* é necessário enviar uma solicitação formal ao instituto.

Embora os *datasets* do RHC, disponibilizados pelo INCA, sejam de domínio público, e estejam disponíveis no *site* da instituição, estes são de difícil acesso. Para realizar o *download* dos dados é necessário um longo processo, com uma grande quantidade de cliques. Além disso, a instituição organiza os dados em arquivos com casos de câncer categorizados por ano, fazendo com que o *download* seja ainda mais trabalhoso.

Outra dificuldade existente está no acesso e análise dos dados, o qual é realizado por meio de uma ferramenta disponibilizada pelo DATASUS para a tabulação deles: *Tabwin*. Como sugerido pelo próprio nome, este é um aplicativo exclusivo para *Windows*. Além disso, a ferramenta não possui código aberto. Desta forma, os pesquisadores tornam-se dependentes de um pequeno grupo de desenvolvedores do governo para obtenção de melhorias no tabulador.

Devido às limitações no processo de análise de dados, diversos pesquisadores têm buscado por ferramentas que os permitam realizar estas análises e manipulações destes dados de forma simplificada e flexível. Entre as ferramentas que têm ganho maior espaço nos últimos anos está a linguagem de programação R. Esta é uma linguagem livre e multiplataforma e geralmente é executada de maneira performática até mesmo em computadores com configuração de hardware modestas. Estas características devem-se principalmente ao processo de instalação base da *Integrated Development Environment* (IDE) normalmente utilizada, o *R Studio*. Nesta IDE são carregados somente métodos essenciais para o uso da ferramenta e da linguagem (LANDEIRO, 2011).

Para realização de tarefas mais complexas, normalmente é necessário realizar a instalação de pacotes adicionais (LANDEIRO, 2011). Pacotes são bibliotecas de funções encapsuladas que podem ser importadas na linguagem R para a análise e manipulação de dados. Desta forma, essa linguagem pode ser facilmente estendida através destes componentes (BEASLEY, 2004), garantindo novos métodos e funcionalidades, sem a necessidade de o usuário reescrever o código.

Este trabalho construiu um pacote desenvolvido na linguagem R para manipulação dos dados fornecidos pelo INCA sobre os casos de câncer registrados no Brasil, baseando-se nos *datasets* do RHC. Como motivação para o desenvolvimento deste pacote foram consideradas as dificuldades de acesso aos

dados e trabalhosos ajustes destes, antes de realizar o processo de análise; e ainda, a inexistência de ferramentas que permitam realizar a geração de gráficos de forma personalizada com estes dados.

Devido à natureza da formação de pesquisadores da área da saúde, a construção desta biblioteca observará diversos fatores, entre eles podemos destacar: (a) a limitação técnica da maior parte do público da biblioteca; (b) os dados presentes nos *datasets* do INCA, considerados relevantes por pesquisadores; (c) a possibilidade de gerar visualizações de forma simplificada, entre outros.

Esse trabalho foi dividido em 5 capítulos, primeiro é esta introdução. O segundo apresenta parte da história do câncer, percorrendo registros ao longo de toda a história até os dias atuais, apresentando alguns dados e estimativas sobre a doença no Brasil e no mundo, expondo os benefícios e a complexidade de pesquisas baseadas em dados de uma maneira geral, apontando as principais dificuldades encontradas neste processo. O terceiro capítulo descreve o *dataset* que será utilizado para a construção da biblioteca e detalha as variáveis disponíveis para tabulação. O quarto capítulo descreve a estrutura do pacote construído, apresentando as funcionalidades disponíveis. O quinto capítulo apresenta um breve manual de uso do pacote INCADATABR, destinado a usuários da área da saúde, onde explica-se o processo de instalação do compilador da linguagem e da *IDE* RStudio, instalação do pacote através de arquivo e o uso das funções presentes neste.

2 REGISTRO DE CASOS DE CÂNCER

Câncer é o nome dado a um conjunto de mais de 100 doenças, cuja principal característica é o crescimento anormal e desordenado de células e, por consequência, o surgimento de tumores malignos com o potencial de invadir ou se espalhar para outras partes do corpo, algo conhecido como metástase (INCA, 2011).

O câncer é uma doença extremamente agressiva que causa uma grande quantidade de mortes em todo o mundo. A doença é conhecida pela humanidade desde longa data, há relatos feitos por egípcios, persas e indianos no século XXX a.C. Contudo, por volta do século IV a.C, estudos da escola hipocrática grega definiram a doença tal como conhecemos hoje. Caracterizando-a como um “tumor duro que, muitas vezes, reaparece depois de extirpado, ou que se alastrava para diversas partes do corpo levando à morte”, sendo associada, neste período, a um desequilíbrio dos fluidos que compunham o organismo (TEIXEIRA, 2007).

De acordo com Oliveira *et al.* (2017), o primeiro registro moderno de câncer teve início em Hamburgo, Alemanha, em 1926. Em 1937, a vigilância do câncer chegou ao Brasil com a criação de um centro especializado para o tratamento da doença no Rio de Janeiro. Em 1983, com o objetivo de difundir o conhecimento sobre o câncer e traçar um perfil da doença que auxiliasse na detecção precoce desta, o Instituto Nacional de Câncer (atual Instituto Nacional de Câncer José Alencar Gomes da Silva - INCA) deu início ao funcionamento do primeiro Registro Hospitalar de Câncer (RHC) do Brasil (MINISTÉRIO DA SAÚDE (BR); INCA, 2012), motivado pela necessidade de mapear os casos de câncer no país e a relevância do estudo de epidemiologia da doença.

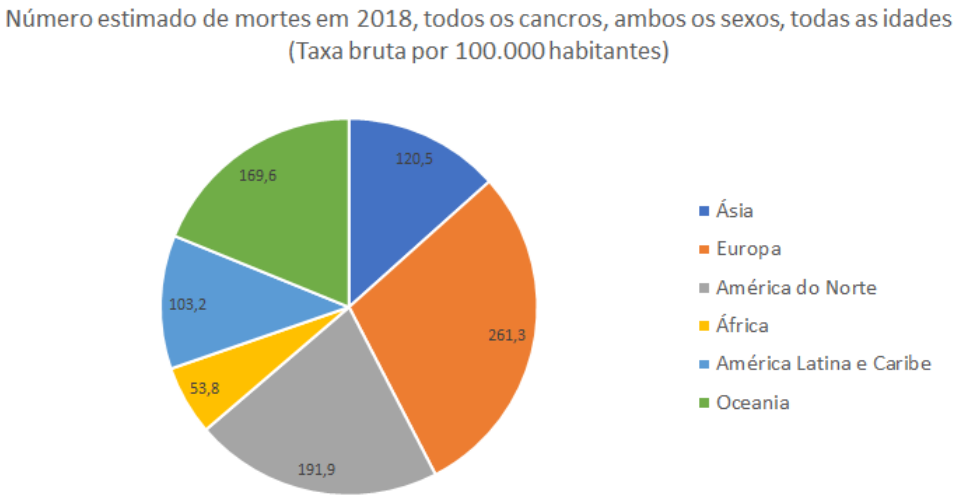
O principal objetivo da instituição era construir um sistema mais estruturado que coletasse os dados de todos os pacientes atendidos em hospitais e gerasse informações que possibilitassem avaliar a eficácia do diagnóstico e tratamento dessas pessoas (OLIVEIRA *et al.*, 2017). De acordo com o (MINISTÉRIO DA SAÚDE (BR); INCA, 2012) “todo esse empreendimento era justificado pela reconhecida importância dos registros, já evidenciada por organizações internacionais, apontados como instrumentos da luta contra o câncer”.

Embora grandes avanços tenham ocorrido no tratamento do câncer, o número de mortes causadas pela doença aumentou. Dados da OMS divulgados em 2017, indicam que entre os anos 2000 e 2015 houve um aumento de 31% no número de

mortes por conta do câncer (INCA, 2017b). De acordo com a OMS (2017a), o câncer é responsável por uma em cada seis mortes no mundo. Este número é 2,5 vezes maior do que as mortes em decorrência de complicações relacionadas a HIV/AIDS, tuberculose e malária combinadas (ONU, 2017).

De acordo com a OMS, em 2018 o câncer foi responsável por 9,6 milhões de morte, sendo a segunda principal causa de morte no mundo. A Figura 1 apresenta o número estimado de mortes por 100.000 habitantes, agrupado por região. De acordo com a ONU (2017) mais de 14 milhões de pessoas desenvolvem câncer todos os anos. Este número deve subir para 21 milhões de pessoas até 2030.

Figura 1 – Número estimado de mortes por 100 mil habitantes em 2018 por região



Fonte: WHO (2018)

No Brasil, este cenário não é diferente, o INCA (2018a) estima cerca de 1,2 milhão de novos casos de câncer no Brasil para o biênio 2018-2019, ou seja, cerca de 575,55 novos casos a cada 100.000 habitantes. Devido ao crescimento do número de casos de câncer ao redor do mundo, diversas instituições têm investido em projetos que auxiliem no desenvolvimento de pesquisas sobre o câncer, entre essas instituições podemos citar *Cancer Research UK*, *Terry Fox Research Institute*, *American Society of Clinical Oncology* e *International Agency for Research on Cancer* e Instituto Nacional de Câncer José Alencar Gomes da Silva.

O INCA age como um órgão auxiliar do Ministério da Saúde no desenvolvimento de programas de combate e controle do câncer (INCA, 2007). O instituto coordena ações integradas para a prevenção e o controle da doença no

Brasil, entre estas ações estão: assistência médico-hospitalar e a atuação em áreas estratégicas, como prevenção e detecção precoce, formação de profissionais especializados, desenvolvimento da pesquisa e geração de informação epidemiológica (INCA, 2018).

Através do Programa de Epidemiologia e Vigilância do câncer e seus Fatores de Risco, o INCA disponibiliza informações agrupadas com diferentes propósitos e particularidades. Essas informações permitem análises sobre a incidência e mortalidade do câncer no Brasil, e garantem o conhecimento detalhado sobre a situação da doença no País e seus fatores de risco.

> *Após tantos anos de trabalho, os Registros de câncer assumiram destaque e responsabilidade crescente como instrumento de apoio à formulação da política nacional de prevenção e controle do câncer; ao planejamento da assistência oncológica, em âmbito nacional e regional; ao processo administrativo hospitalar; e à elaboração de trabalhos científicos. (KLIGERMAN, 2001, p. 1).*

Os dados coletados por hospitais de todo o país são unificados em uma única base de dados, o RHC. Estes dados então são disponibilizados no *site* do INCA, sendo agrupados por ano. Estes *datasets* incluem informações referentes aos pacientes e sobre o estágio da doença. Entre estes dados podemos citar: idade, sexo, raça/cor, local de nascimento, escolaridade, estado conjugal atual, histórico de consumo de bebida alcoólica, histórico de consumo de tabaco, estadiamento clínico do tumor, de acordo com um padrão globalmente reconhecido para classificar a extensão da disseminação do câncer – a *TNM Classification of Malignant Tumours* (TNM), entre outras. São disponibilizados pela instituição *datasets* com registros entre os anos de 1985 e 2016 (com exceção de 1987).

O RHC tem como objetivo reunir os dados dos pacientes com diagnóstico de câncer nos hospitais em que foram diagnosticados e/ou tratados. Com base nestes dados é possível conhecer o perfil dos pacientes que chegam à unidade e como foram diagnosticados e tratados. Além disso, o RHC serve como ferramenta para avaliação da eficácia dos serviços oferecidos pelas clínicas que realizaram os atendimentos, possibilitando o planejamento e a melhoria de sua qualidade. O RHC também é útil como fonte de dados para pesquisas e vigilância epidemiológica. (OLIVEIRA *et al*, 2017).

Os arquivos são disponibilizados no formato *dbf* e exigem, para a visualização e manipulação dos dados, a utilização da aplicação *TabWin*. Essa ferramenta foi desenvolvida pelo DATASUS e serve para a importação de tabulações geradas pelo aplicativo TABNET (também desenvolvido pelo DATASUS).

O *TabWin* teve sua primeira versão beta (0.07) lançada em 15/07/1997 e sua primeira versão estável (1.0) em 08/03/1998. Desde então, o DATASUS realiza diversas implementações na ferramenta, que atualmente encontra-se na versão 3.6b lançada em 05/07/2010. Por não ser de código aberto, a ferramenta acaba trazendo uma série de limitações aos usuários que a utilizam, conforme descrito por Petruzalek (2016).

> *As ferramentas disponibilizadas no sítio do DATASUS são ferramentas livres, porém de código fechado e disponíveis exclusivamente para a plataforma Windows, o que impõe severas limitações à implementação de melhorias e customizações no software de tabulação, ao desenvolvimento de processos automatizados (batches) e na escalabilidade das plataformas de software, além de vincular o desenvolvimento de qualquer produto derivado a um único sistema operacional. (PETRUZALEK, 2016, p. 2).*

Um problema comum identificado ao utilizar as bases de dados abertas é a falta de padronização nos diversos *datasets* disponibilizados. De acordo com Pinto *et al* (2012), somente a partir do ano 2000 um sistema padronizado para a entrada e armazenamento dos dados do RHC foi adotado. Esta falta de padronização nos registros anteriores ao ano 2000 acaba dificultando a utilização destes registros disponibilizados pelo INCA.

Outra dificuldade na utilização de dados disponibilizados por instituições em seus portais é desconhecer a qualidade e confiabilidade dos dados. Conforme descrito por Pinto *et al* (2012), garantir a qualidade e a confiabilidade dos sistemas de informação é fundamental na vigilância em saúde. Segundo os autores, na literatura, não há um consenso associado à qualidade de dados.

Os dados do RHC, disponibilizados pelo INCA como *datasets,* já foram avaliados em diversos momentos. Pinto *et al* (2012) descreveram os dados presentes nestes registros como com boa completude e consistência das informações, porém, indicam que algumas variáveis apresentaram completude ruim. Wollmann (2018) indicou que a base do INCA atingiu o nível de 3 estrelas (nível médio) no modelo utilizado em seu trabalho para avaliação, sendo indicada como de grande valia para

os pesquisadores na área da saúde, uma vez que, disponibiliza informações detalhadas sobre os casos de câncer nas últimas três décadas. Além disso, Wollmann (2018) indica que o INCA analisa os dados coletados e realiza a remoção de registros em duplicidade, garantindo assim, uma base íntegra e de boa qualidade.

Este capítulo apresentou um panorama sobre a situação do câncer no Brasil e no Mundo e tem como objetivo contextualizar a situação da doença e a forma como é realizada a coleta de dados sobre a doença. Além disso, este capítulo apresentou o INCA e a maneira como os dados coletados pela instituição podem ser utilizados como ferramenta de apoio a decisão para o desenvolvimento de programas de controle e prevenção da doença. Além disso, foram descritas algumas das dificuldades enfrentadas por pesquisadores para utilizar dados em suas pesquisas.

3 CARACTERÍSTICAS DO *DATASET* DO RHC DISPONIBILIZADO PELO INCA

Embora exista grande abundância de dados, pesquisadores encontram algumas dificuldades em desenvolver projetos cujo objetivo é realizar a análise de dados. Entre esses problemas podemos citar o complexo processo de obtenção dos dados, muitas vezes associado a burocracia, má estruturação, inexistência de uma estrutura para armazenamento dos dados coletados ou a omissão destes dados por parte do governo e instituições.

Grandes avanços ocorreram nos últimos anos, motivados principalmente pelo movimento de dados abertos e no Brasil à sanção da Lei 12.527/2011, também conhecida como Lei de Acesso à Informação. Esta lei motivou a criação do Portal Brasileiro de Dados Abertos, uma plataforma que permite que todos possam encontrar e utilizar os dados e as informações públicas. Assim, diversas instituições começaram a disponibilizar bases de dados que podem ser utilizadas em pesquisas e como ferramenta de apoio a decisões.

> *Em 18 de novembro de 2011 foi sancionada a Lei de Acesso à Informação Pública (Lei 12.527/2011) que regula o acesso a dados e informações detidas pelo governo. Essa lei constitui um marco para a democratização da informação pública, e preconiza, dentre outros requisitos técnicos, que a informação solicitada pelo cidadão deve seguir critérios tecnológicos alinhados com as "3 leis de dados abertos". (PORTAL BRASILEIRO DE DADOS ABERTOS, [201-?], p. 1).*

O uso de dados em pesquisa também é dificultado pelo número limitado de ferramentas, as quais ainda exigem que o pesquisador tenha conhecimentos intermediários e avançados em informática. Estes conhecimentos acabam sendo requeridos, em virtude da necessidade de realizar ajustes sobre os dados que serão utilizados na pesquisa, seja devido a existência de dados incorretos, inválidos ou duplicados. Estes ajustes são necessários, pois a existência de dados inconsistentes pode impedir a utilização de diversos registros, inviabilizar o uso de um determinado *dataset*, ou ainda contribuir para análises errôneas.

> *Os dias atuais caracterizam-se por profundas e constantes mudanças, onde é crescente e cada vez mais acelerada a inovação tecnológica, colocando à*

> *disposição dos profissionais e usuários, os mais diversos tipos de tecnologia, tais como: tecnologias educacionais, tecnologias gerenciais e tecnologias assistenciais. (BARRA et al, 2006, p. 2).*

Atualmente, os dados representam muito mais do que apenas informações do mundo real, há muito conhecimento disponível nestes dados. De acordo com Camilo e Silva (2009), embora sejam realizados grandes investimentos por diversas organizações e instituições no processo de coleta de dados, pouca informação útil é identificada nestes dados. Amaral (2016) indica que, devido a abundância de dados, o mundo jamais será o mesmo, uma vez que, se há alguns anos, para produzir, armazenar e analisar dados eram necessários equipamentos gigantescos, hoje a produção e armazenamento de dados é algo comum que faz parte de nosso cotidiano.

As organizações e instituições reconhecem que os dados devem ser usados como ferramenta de apoio para a tomada de decisão, porém, em diversos casos, estes são coletados e jamais utilizados. Este cenário é descrito por Han, Pei e Kamber (2011) como uma situação rica em informações, mas pobre em conhecimento, uma vez que, estes conjuntos de dados superaram a capacidade humana de compreensão, impossibilitando a identificação de padrões e relações existentes entre eles. Com isso, os grandes depósitos de dados tornam-se "túmulos de dados" que raramente serão acessados.

No domínio científico da computação, a inferência de algumas informações valiosas nos dados observados é apontada por Pujari (2001) como um problema, ressaltando a importância do desenvolvimento de ferramentas e mecanismos que auxiliem neste processo. Na área da saúde, diversas iniciativas para a modernização foram realizadas nos últimos anos. No final de 2016, o Ministério da Saúde e Ministério da Ciência, Tecnologia, Inovações e Comunicações assinaram um acordo de cooperação para estimular o desenvolvimento de estudos e novas soluções tecnológicas na área da saúde no país (BRASIL, 2016).

Além disso, de acordo com o DATASUS (2017), o Ministério da Saúde estabeleceu a meta de, até o fim de 2018, 100% das Unidades Básicas de Saúde (UBS) do Brasil aderirem ao prontuário eletrônico. Estas iniciativas demonstram a relevância e importância do uso de tecnologia para a coleta de informações, que poderão ser utilizadas como ferramenta para elevar a eficiência e a qualidade dos atendimentos realizados pelo Sistema Único de Saúde (SUS).

No controle e prevenção do câncer no Brasil, o INCA age como um órgão auxiliar do Ministério da Saúde. A instituição tem investido constantemente no aprimoramento do processo e metodologia de coleta e análise dos dados, cujo objetivo é garantir a qualidade e a confiabilidade dos dados disponibilizados pela instituição, algo classificado como fundamental na vigilância em saúde.

> *A disponibilidade de informações com base em dados confiáveis é condição essencial para a análise de situação e subsídio para tomada de decisões. Para análise do desempenho do sistema de saúde são consideradas e aplicadas medidas-sínteses (indicadores) que vistos em conjunto, servem para a vigilância da situação de atenção ao paciente, em todos os níveis. (REBELO et al, 2008).*

O instituto também é responsável por produzir as estimativas de câncer no país, utilizando como base para estas estimativas dados oriundos de diversas fontes oficiais. Estas estimativas são utilizadas para o planejamento de ações de prevenção e controle do câncer no Brasil.

> *[...] o Brasil produz as estimativas para a incidência de câncer desde 1995, com aprimoramento metodológico constante para o seu cálculo, a partir da melhoria da quantidade, qualidade e da atualidade das informações dos RCBP, dos RHC e do Sistema de Informações sobre mortalidade (SIM) (INCA, 2017a, p. 8).*

De acordo com o instituto, embora existam limitações, acredita-se que as estimativas divulgadas sejam capazes de dimensionar o impacto do câncer no país, garantindo o planejamento de ações de prevenção e controle da doença. De acordo com Stewart e Wild apud INCA (2018, p. 25) “Informações sobre a ocorrência de câncer e seu desfecho são requisitos essenciais para programas nacionais e regionais para o controle do câncer, além de pautar a agenda de pesquisa sobre câncer”.

Além disso, o INCA é responsável pelo Programa de Epidemiologia e Vigilância do câncer e seus Fatores de Risco, sendo este programa encarregado pela construção da base nacional consolidada dos Registros de Câncer de Base Populacional (RCBP) e Registros Hospitalares de Câncer (RHC). O principal objetivo destes dois registros é fornecer um quadro detalhado sobre a situação do câncer no Brasil. Tais informações podem ser utilizadas como parâmetro para definições de programas de combate e prevenção à doença.

> *As ações nacionais de Vigilância do câncer têm como objetivo conhecer com detalhes o atual quadro do câncer no Brasil. A vigilância do câncer é realizada por meio da implantação, acompanhamento e aprimoramento dos Registros de câncer de Base Populacional e dos Registros Hospitalares de câncer (centros de coleta, processamento, análise e divulgação de informações sobre a doença, de forma padronizada, sistemática e contínua). Os registros possibilitam conhecer os novos casos e realizar estimativas de incidência do câncer, subsídios fundamentais para o planejamento das ações locais de prevenção e controle da doença de acordo com cada região. (INCA, 2018a, p. 1).*

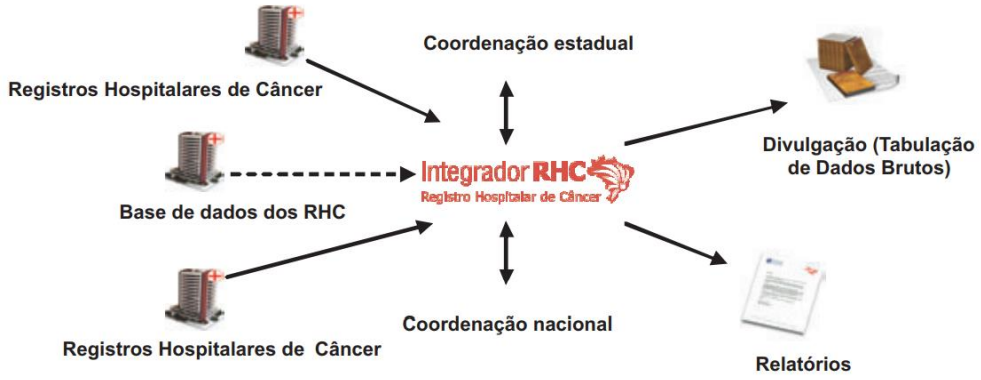
O INCA centraliza em seu *site*[1] registros coletados de 260 hospitais brasileiros (CASATI, 2012). Estes dados são unificados utilizando uma ferramenta desenvolvida pela própria instituição, por meio da Divisão de Informação da Coordenação de Prevenção e Vigilância (CONPREV) e da Divisão de Tecnologia da Informação da Coordenação de Ações Estratégicas, em parceria com o Laboratório de Sistemas Integráveis (LSI) da Escola Politécnica da Universidade de São Paulo (BRASIL, 2007). Conhecido como *IntegradorRHC* esta é uma ferramenta Web, multiplataforma, que em conjunto com a ferramenta de coleta dos dados, *SisRHC*, é denominado como RHC Brasil.

> O *IntegradorRHC* se apresenta como uma ferramenta pioneira, permitindo a consolidação de bases de dados hospitalares sobre câncer, agilizando o acesso à informação. Dentro do processo de democratização da informação, a qualidade e a oportunidade formam a base fundamental para conhecimento da realidade desta doença no País, contribuindo para definição de políticas públicas. (BRASIL, 2007, p. 868).

De acordo com Brasil (2007), o *IntegradorRHC* abrange desde a captação de dados dos RHC (utilizando o *SisRHC* para garantir a consistência dos dados presentes nos registros), até a consolidação das informações. Por fim, esses dados são disponibilizados para análise, pesquisa, tabulações e exportações. Este fluxo é apresentado na Figura 2. De acordo com Thuler, Bergmann e Casado (2012), o objetivo do *IntegradorRHC* é consolidar as informações hospitalares dos diversos RHC distribuídos em todo o país e divulgá-los à comunidade científica de forma sistemática e contínua.

[1] INCA ([20–]a)

Figura 2 – Fluxo de informações no *IntegradorRHC*



Fonte: INCA (2011)

Os registros de câncer foram consolidados nas últimas décadas como pilares fundamentais para a vigilância epidemiológica da incidência da doença. Além disso, os registros de câncer são fontes imprescindíveis para o desenvolvimento de pesquisas epidemiológicas e clínicas, e planejamento e avaliação das ações de controle da doença (BRASIL, 2007).

A biblioteca desenvolvida neste trabalho foi pensada com base no problema descrito por Pujari (2001), considerando a importância do desenvolvimento de ferramentas e mecanismos que auxiliem no processo de aquisição de conhecimento sobre os dados fornecidos pelo INCA. A biblioteca se baseia nos dados presentes no RHC devido a sua relevância, uma vez que análises realizadas sobre o RHC podem contribuir para o planejamento de ações nas áreas de educação e detecção precoce da doença.

O RHC é disponibilizado pelo INCA ao público geral, sem a necessidade de registro. Este *dataset* possui 40 variáveis e registros entre os anos de 1985 e 2017 (com exceção 1987 que não possui registros para *download* no site da instituição).

No portal do *IntegradorRHC* é disponibilizado pelo INCA um documento denominado notas técnicas. Neste documento são apresentadas informações técnicas sobre o RHC, como: a origem dos dados existentes, os sistemas de classificação de doença adotado, o fluxo e atualização dos dados e a descrição das variáveis disponíveis para tabulação no RHC.

O INCA (2018a) descreve o RHC como um desafio para o Brasil, devido às suas dimensões continentais. Para coletar estes dados, o instituto aposta em centros de informação.

> *O registro nacional de câncer é um desafio para países em desenvolvimento, especialmente para o Brasil com suas dimensões continentais. A estratégia tem sido manter e fortalecer centros de informação (Registros de Câncer de Base Populacional e Hospitalares - RCBP/RHC) que permitam monitorar a situação do câncer como parâmetro para todo o país; e ainda, dentro dessa lógica, por meio das estimativas de câncer, seja possível obter informações atualizadas e aplicáveis às necessidades estratégicas do país (INCA, 2017a, p. 8).*

De acordo com o INCA (2011), são cadastrados no registro nacional de câncer apenas casos que se adequem a Classificação Internacional de Doenças para Oncologia (CID-O). Os registros anteriores a 2005 utilizavam a CID-O/2 e posteriores a este ano adotam o CID-O/3.

> *São consideradas elegíveis para cadastro as doenças com comportamento classificado, na Classificação Internacional de Doenças para Oncologia (CID-O), como malignas, neoplasias in situ e alguns tumores de comportamento benigno, incerto ou desconhecido que tenham sido considerados como de interesse científico para o registro. (INCA, 2011, p. 2)*

Com relação ao fluxo e atualização dos dados do *IntegradorRHC*, este é regido pelo artigo 5º da Portaria GM/MS nº 741 de dezembro de 2005. Neste artigo, é estabelecido que Centros de Assistência de Alta Complexidade em Oncologia (CACON), Centros de Referência de Alta Complexidade em Oncologia e Unidades de Assistência de Alta Complexidade em Oncologia (UNACON) devem enviar os dados coletados anualmente, seguindo o calendário estabelecido pela portaria. Para realizar o envio é utilizado o *IntegradorRHC*. Após isso, uma base de casos analíticos é criada com todos os dados recebidos. Esta base então é submetida a análises para remoção de casos com mais de um registro. Este processo é realizado em dois níveis: no nível estadual, este processo é realizado pela Coordenação Estadual da Vigilância do câncer; e no nível nacional, este processo é realizado pelo INCA.

As variáveis disponíveis para tabulação no RHC são separadas entre variáveis de preenchimento obrigatório (de 1 até 35) e variáveis de preenchimento opcional (de 36 até 40). Abaixo são listadas estas variáveis, juntamente com uma breve descrição sobre cada uma delas, assim como os valores possíveis para cada uma no RHC.

Estas informações foram obtidas no documento de notas técnicas[2] e no dicionário de dados[3], ambos disponibilizados pelo INCA.

1. **Tipo do caso:** Esta variável indica se o caso do registro foi classificado como analítico ou não analítico. Neste campo os valores possíveis são 1 para analítico e 2 para não analítico.
2. **Ano da primeira consulta:** Esta variável se refere ao ano em que a primeira consulta relacionada ao tumor foi realizada. O valor é armazenado no formato dd/mm/aaaa.
3. **Unidade Hospitalar:** Esta variável permite identificar a Unidade Hospitalar que prestou assistência oncológica ao paciente e cadastrou o caso. O valor gravado nesta variável é o código da unidade hospitalar, de acordo com a Tabela de Clínicas do *SisRHC*.
4. **Município da Unidade Hospitalar:** Esta variável identifica o município em que está instalada a Unidade Hospitalar que prestou assistência oncológica ao paciente e cadastrou o caso. O valor gravado nesta variável é o código do município, de acordo com a tabela do Instituto Brasileiro de Geografia e Estatística (IBGE).
5. **UF da Unidade Hospitalar:** Esta variável identifica a Unidade Federativa em que está localizada a Unidade Hospitalar que prestou assistência oncológica ao paciente e cadastrou o caso. O valor gravado nesta variável é a sigla correspondente a Unidade Federativa, composta por dois caracteres.
6. **Sexo:** Esta variável identifica o sexo do paciente. Os valores admitidos são 1 para masculino e 2 para feminino.
7. **Faixa Etária**: Esta variável permite identificar a faixa etária do paciente. As faixas etárias (em anos) estão agrupadas nas seguintes categorias: De 00 a 04 anos; De 05 a 09 anos; De 10 a 14 anos; De 15 a 19 anos; De 20 a 24 anos; De 25 a 34 anos; De 35 a 44 anos; De 45 a 54 anos; De 55 a 64 anos; De 65 a 74 anos; De 75 a 84 anos; 85 ou mais; sem informação.
8. **Faixa etária infantil detalhada:** Esta variável se refere a faixa etária infantil (em anos) e abrange as faixas de zero aos dezoito anos. Ela possibilita que

[2] INCA ([20–?]b)
[3] INCA ([20–?]c)

o usuário tabule os dados por cada idade ou agrupe as faixas etárias da forma que considerar mais adequada.

9. **Faixa etária infantil:** Esta variável se refere a agrupamentos de idade entre crianças cuja idade vai de zero aos dezoito anos. As faixas de idade estão agrupadas nas seguintes categorias: De 00 a 02 anos; de 03 a 07 anos; de 08 a 14 anos; 15 anos; 16 anos; 17 anos e 18 anos.
10. **Local de nascimento:** Esta variável identifica a Unidade Federativa de nascimento do paciente. O valor gravado nesta variável é a sigla correspondente a Unidade Federativa, composto por dois caracteres, sendo que, para estrangeiros é utilizada a sigla EX.
11. **Raça/Cor:** Esta variável permite identificar a raça/cor do paciente. O valor gravado nesta variável são os atributos adotados pelo IBGE que classificam raça/cor nas seguintes categorias: branca, preta, amarela, parda e indígena e sem informação.
12. **Grau de Instrução:** Esta variável permite identificar a escolaridade do paciente. As categorias disponíveis são: analfabeto; 1º grau incompleto; 1º grau; 2º grau, superior e sem informação.
13. **Clínica de 1º Atendimento:** Esta variável permite identificar o serviço médico especializado responsável pela matrícula e atendimento inicial ao paciente no hospital. O valor gravado nesta variável é o código da unidade hospitalar, de acordo com a Tabela de Clínicas do *SisRHC*[4].
14. **Clínica de tratamento:** Esta variável permite identificar a clínica onde o tratamento do paciente foi efetivamente iniciado, sendo que, caso o tratamento foi realizado por mais de uma clínica, o RHC considera qual a clínica assumiu o papel primordial no tratamento. O valor gravado nesta variável é o código da unidade hospitalar, de acordo com a Tabela de Clínicas do *SisRHC*.
15. **UF de procedência:** Esta variável permite identificar o estado em que o paciente reside. O valor gravado nesta variável é a sigla correspondente a Unidade Federativa, composto por dois caracteres, sendo que, para estrangeiros o valor usado é EX.

[4] INCA ([20–?]d)

16. **Procedência:** Esta variável permite identificar o município em que o paciente reside. O valor gravado nesta variável é o código do município, de acordo com a tabela do IBGE.
17. **Ocupação:** Esta variável permite identificação da ocupação do paciente. O RHC descreve em seu manual que, esta variável retrata a ocupação preponderante do paciente e não apenas a que ele exerce no momento da matrícula. Nesta variável é gravado o código da Classificação Brasileira de Ocupações (CBO) do Ministério do Trabalho.
18. **Ano do diagnóstico:** Esta variável permite identificar o ano em que o diagnóstico do câncer foi realizado. O valor gravado nesta variável segue o formado “aaaa”. Esta data pode ser anterior ou posterior a primeira consulta.
19. **Origem do encaminhamento**: Esta variável permite identificar à origem do encaminhamento do paciente à Unidade Hospitalar. Os valores possíveis para esta variável são: SUS; não SUS; veio por conta própria e sem informação.
20. **Diagnóstico e Tratamento Anterior:** Esta variável permite identificar se houve diagnóstico e tratamento do tumor em período anterior a entrada do paciente no hospital. As categorias disponíveis são: sem diagnóstico/sem tratamento; com diagnóstico/sem tratamento; com diagnóstico/com tratamento; outros e sem informação.
21. **Base do diagnóstico:** Esta variável permite identificar através de qual método foi realizado o diagnóstico da doença. As categorias disponíveis são: exame clínico e patologia clínica; exame por imagem; endoscopia; cirurgia exploradora/necropsia; citologia ou hematologia; histologia da metástase; histologia do tumor primário; sem informação.
22. **Exames para diagnóstico:** Esta variável indica os exames relevantes para o diagnóstico e planejamento da terapêutica do tumor. As categorias disponíveis são: exame clínico e patologia clínica; exame por imagem; endoscopia e cirurgia exploradora; anatomia patológica; sem informação.
23. **Localização primária:** Esta variável indica a localização primária do tumor. Para este campo são utilizados os códigos da Classificação Internacional de Doenças para Oncologia (CID-O/2 para os casos com data da primeira consulta anterior a 01/01/2005 e CID-O/ 3 para os casos com data da

primeira consulta a partir de 01/01/2005). O código utilizado é composto por 3 caracteres (de C00 a C80).

24. **Localização primária detalhada:** Esta variável indica a subcategoria de localização primária do tumor. Assim como na localização primária, são utilizados os códigos do CID-O, porém, os códigos possuem 4 caracteres (de C00.0 a C80.9).
25. **Localização primária grupo:** Esta variável permite identificar o agrupamento de categorias da localização primária do tumor a qual o tumor se adequa. A codificação é feita utilizando a CID-O.
26. **Localização primária provável:** Esta variável permite que a localização primária seja indicada quando não se tem certeza da localização primária do tumor. Também utiliza a CID-O, com 3 caracteres (de C00 a C80).
27. **Tipo Histológico:** Esta variável armazena valores referentes a caracterização da estrutura celular do tumor (morfologia do tumor) através de exame microscópico. A codificação do tipo histológico é feita utilizando a CID-O e este código é composto por 5 caracteres, sendo que, os 4 primeiros caracteres (que variam de 8000 a 9989) indica o tipo celular e o último caractere está relacionado ao comportamento biológico do tumor.
28. **Lateralidade:** Esta variável é preenchida somente quando o tumor atinge órgãos que possuem par (como mama, pulmão, rim, entre outros). O objetivo de estudar a frequência de tumores em órgãos múltiplos. As categorias disponíveis são: direita; esquerda; bilateral; não se aplica e sem informação.
29. **Tumor Primário Múltiplo:** Esta variável indica se ocorreram múltiplos tumores primários (em um determinado órgão ou em órgãos diferentes). As categorias disponíveis são: sim; não e duvidoso.
30. **Estadiamento (TNM):** Esta variável se refere à avaliação da extensão da neoplasia maligna antes do tratamento. Para o estadiamento dos tumores é utilizada a *TNM Classification of Malignant Tumours* (TNM) da União Internacional Contra o câncer - UICC. Para os casos anteriores a 2005 é utilizada a 5ª Edição do TNM e para os casos a partir do ano 2005 é utilizada a 6ª edição do TNM.

31. **Estadiamento grupo**: Esta variável é semelhante a variável estadiamento, os estágios são agrupados nas seguintes categorias: 0, I, II, III, IV, A, B, C e D. Assim como no Estadiamento, é utilizada a TNM da UICC.
32. **Ano do primeiro tratamento:** Esta variável indica o ano em que foi realizado o primeiro tratamento recebido na unidade hospitalar. O valor é salvo no formato aaaa.
33. **Primeiro tratamento recebido no hospital:** Esta variável indica o primeiro tratamento realizado para o tumor. As categorias disponíveis são: nenhum; cirurgia; quimioterapia (QT); radioterapia (RXT); hormonioterapia (HT); transplante de medula óssea (TMO); categorias que se referem à combinação dessas modalidades de tratamentos; outros procedimentos terapêuticos; sem informação.
34. **Razão para não tratar:** Esta variável indica o motivo pelo qual não foi realizado tratamento. As categorias disponíveis são: recusa do tratamento; doenças avançadas; falta de condições clínicas; outras doenças associadas; abandono do tratamento; complicações do tratamento; óbito; outras; não se aplica e sem informação.
35. **Estado da doença ao final do primeiro tratamento:** Esta variável indica a situação da doença após a realização do primeiro tratamento. As categorias disponíveis para essa variável são: sem evidência da doença (remissão completa); remissão parcial; doença estável; doença em progressão; fora de possibilidade terapêutica; óbito; não se aplica e sem informação.
36. **Estado conjugal:** Esta variável indica o estado conjugal atual do paciente. Os valores disponíveis são: casado; solteiro; desquitado/separado/divorciado; viúvo e sem informação. Esta variável não se refere a situação legal do casal.
37. **Ano da triagem:** Esta variável indica o ano em que o paciente teve seu primeiro contato com a unidade hospitalar relacionado ao tumor. O valor é armazenado no formato aaaa.
38. **História Familiar de câncer:** Esta variável armazena informação sobre a existência de casos de câncer na família. As categorias disponíveis são: sim, não e sem informação.

39. **Alcoolismo:** Esta variável armazena o histórico do paciente com relação ao consumo de bebidas alcoólicas. As categorias disponíveis são: Sim; não; não se aplica e sem informação.
40. **Tabagismo:** Esta variável armazena o histórico do paciente com relação ao consumo de tabaco. As categorias disponíveis são: sim; não; não se aplica e sem informação.

Ao longo deste capítulo, foram apresentadas as responsabilidades do INCA e os *datasets* disponibilizados pela instituição. Como o objeto de estudo deste trabalho, o RHC foi especificado, isto é, o fluxo dos dados e as variáveis disponíveis neste *dataset* foram detalhados.

4 O PACOTE INCADATABR

O pacote (biblioteca) que foi construído tem o objetivo de auxiliar pesquisadores na área da saúde ou outros pesquisadores que necessitem interagir e realizar análises dos dados disponibilizados pelo INCA. O objetivo do pacote é permitir que pesquisadores possam realizar análises utilizando a linguagem R, evitando diversos pré-processamentos necessários, caso outras ferramentas fossem utilizadas.

O pacote permite que usuários com conhecimentos básicos na linguagem R possam utilizá-lo, tornando a análise mais flexível. Desta forma, o *INCADATABR* torna-se uma outra alternativa às análises realizadas com a ferramenta *TabNet*.

A linguagem R foi escolhida devido a sua relevância na área de análise de dados e principalmente devido ao crescente número de usuários. De acordo com uma pesquisa realizada pelo site (KDNUGGETS*,* 2012*)*, 30,7% dos usuários mencionaram ter utilizado a linguagem R em processos de análise e mineração de dados. Este número é 7,4% maior do que o registrado na pesquisa realizada no ano anterior. Além disso, em uma pesquisa realizada pela (*REXER ANALYTICS*, 2015) a linguagem R foi a mais utilizada entre os 1220 cientistas avaliados, sendo que, 76% dos entrevistados relatam o uso de R e 36% identificam a linguagem R como sua principal ferramenta.

Ao longo deste capítulo serão apresentados alguns trabalhos relacionados ao tema, a modelagem do pacote e as funcionalidades do mesmo, bem como um tutorial de uso das funções para profissionais da área da saúde e aspectos importantes sobre a biblioteca.

4.1 TRABALHOS RELACIONADOS

Com o objetivo de entender e levantar os diversos requisitos para a construção de um pacote para a análise de dados relacionados ao Câncer, pesquisou-se sobre iniciativas voltadas a esta área. Embora o número iniciativas encontradas que se relacionassem com a proposta deste trabalho seja pequeno, identificou-se duas iniciativas que direcionaram a construção deste pacote. Este levantamento foi realizado através de uma busca por pacotes disponibilizados na *Comprehensive R Archive Network* (CRAN), repositório oficial de pacotes da linguagem R e em buscas no acervo de trabalhos desenvolvidos na Universidade Feevale.

O trabalho proposto por Medinger (2017), um egresso do curso de Sistemas de Informação da Universidade Feevale, apresentado como requisito à obtenção do grau de Bacharel em Sistemas de Informação. Esse trabalho apresenta o DataVis INCA, um *dashboard* para análise dos *datasets* de RHC, o qual foi utilizado como alicerce do pacote *INCADATABR*. O outro trabalho encontrado foi proposto por Bulow (2017) e serviu como parâmetro para a escolha do modelo de arquitetura e modularização do pacote *INCADATABR*. As duas propostas serão detalhadas nas seções seguintes com o objetivo de contextualizar e dar um panorama atual de ferramentas já existentes cujo objetivo é similar ao da ferramenta proposta neste trabalho.

4.1.1 DATAVIS INCA

Como mencionado anteriormente, para o desenvolvimento do pacote *INCADATABR* utilizou-se como base a ferramenta proposta por Medinger (2017). O autor, em seu trabalho, construiu uma ferramenta que permite gerar visualizações dinâmicas sobre os dados disponibilizados pelo INCA.

A ferramenta é conhecida como *DataVis INCA* e tem como objetivo permitir a geração de gráficos baseados nos dados fornecidos pelo INCA diretamente no navegador, sem que o usuário necessite escrever nenhuma linha de código, isto é, a geração é realizada de forma automática de acordo com os parâmetros selecionados pelo usuário na aplicação. As Figuras 3, 4 e 5 apresentam algumas interfaces da aplicação e exemplos de gráficos gerados pela mesma.

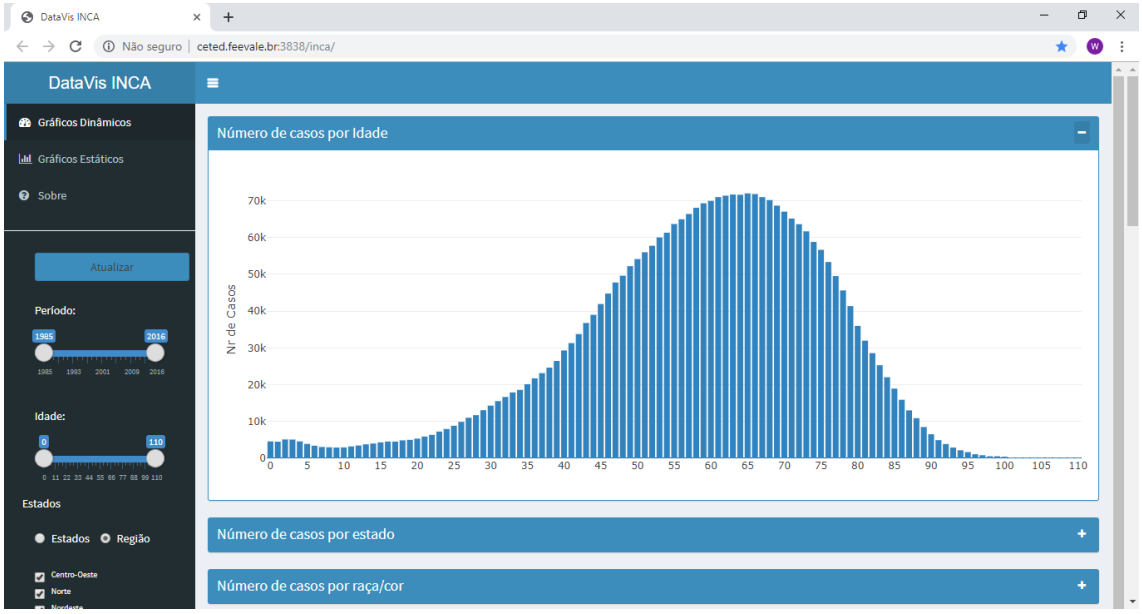
Em seu trabalho, Medinger (2017) descreve que a linguagem R foi escolhida para o desenvolvimento do *DataVis INCA* devido a sua capacidade em gerar visualizações em diversos formatos. Além disso, a existência de diversas bibliotecas já disponíveis para este tipo de atividade foi levada em consideração para a escolha desta linguagem.

Durante as fases iniciais de desenvolvimento as bibliotecas *shiny* e *ggplot2* foram utilizadas por Medinger (2017) na construção do *front-end* e plotagem dos gráficos. No entanto, para melhorar a aparência da ferramenta e dos gráficos gerados, a biblioteca *Plotly* foi adotada.

A adoção da biblioteca *Plotly* garantiu à ferramenta alguns *add-ons* como: a possibilidade de filtrar as variáveis que serão exibidas no gráfico após sua plotagem, a exibição de informações adicionais ao posicionar o ponteiro do *mouse*, a

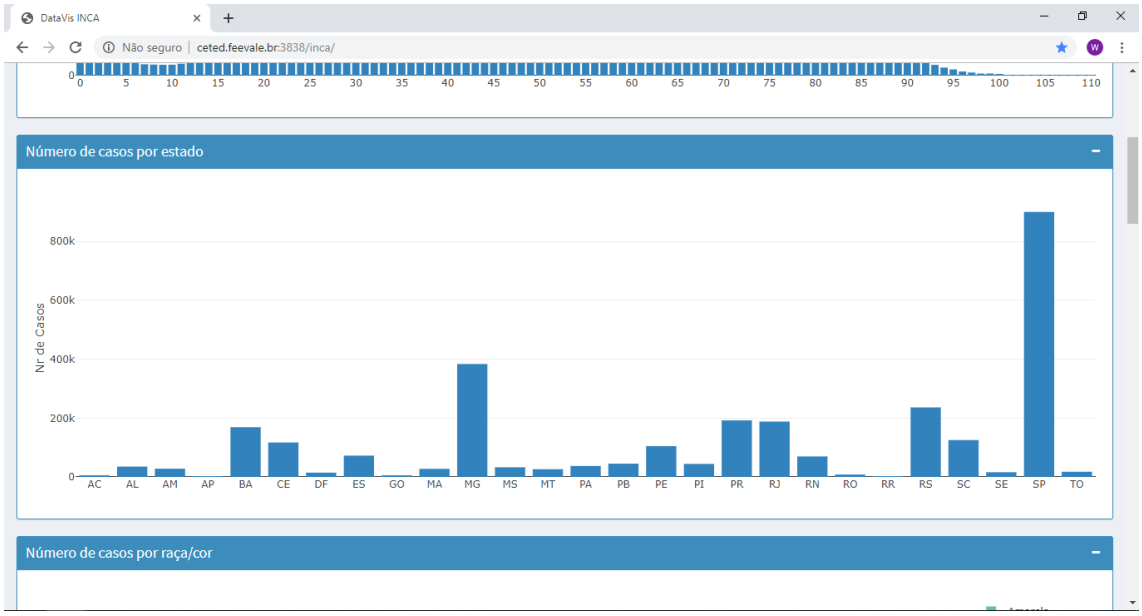
possibilidade de dar *zoom* sobre parte do gráfico e a exportação do gráfico gerado como imagem no formato *png*.

Figura 3 – Tela inicial da ferramenta *DataVis INCA* proposta por Medinger (2017).



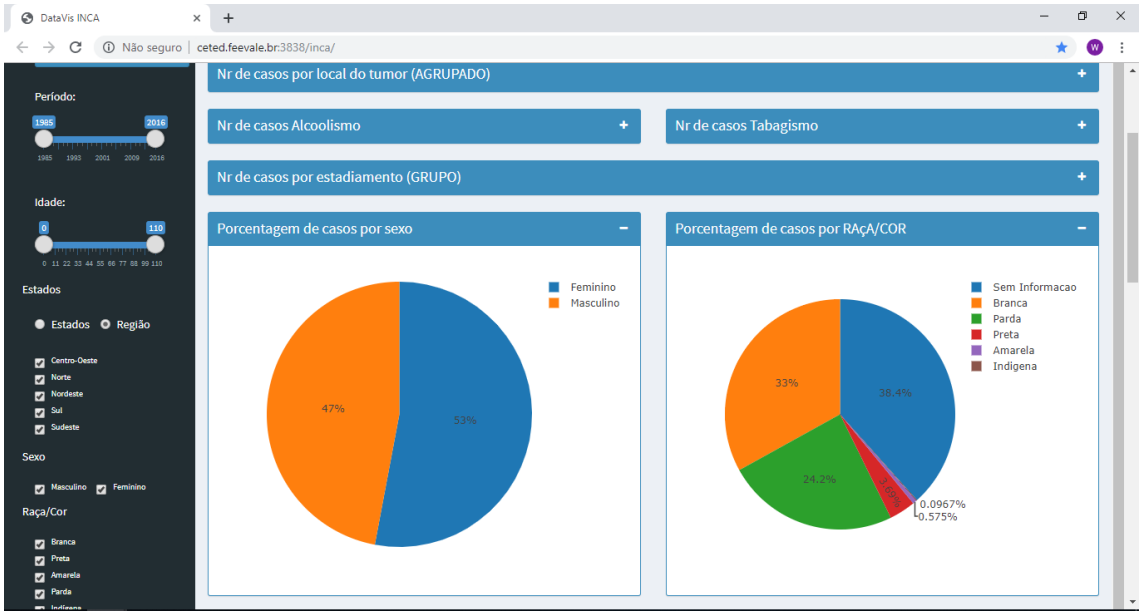
Fonte: Adaptado pelo Autor

Figura 4 – *Dashboard DataVis INCA* com menu “encolhido”



Fonte: Adaptado pelo Autor

Figura 5 – Exemplo de gráfico pizza do *DataVis INCA*.



Fonte: Adaptado pelo Autor

Os dados utilizados pela ferramenta proposta Medinger estão armazenados em um servidor do Centro de Tecnologias Digitais (CETED) da Universidade Feevale, sendo este um centro da instituição cujo objetivo é integrar os projetos de ensino, pesquisa e extensão ligados à área de TI. Através deste são coordenadas ações que aproxima a universidade da comunidade.

4.1.2 INCADATA

Durante o desenvolvimento do pacote, identificou-se no *Comprehensive R Archive Network* (CRAN) um projeto semelhante ao *INCADATABR*. Desenvolvido por Bulow (2017), o INCADATA é composto por um conjunto de funções desenvolvidas em R cujo objetivo é prover funcionalidades básicas para manipular dados do INCA e dos centros regionais de câncer da Suécia. Entre as funcionalidades disponíveis na versão 0.8.1 do pacote destacam-se:

***as.incadata*:** esta função tem como objetivo realizar a conversão de dados oriundos de fontes diversas de forma forçada e automática para o tipo apropriado. Este processo é realizado através da identificação das colunas ou através do conteúdo presente na variável. Além disso, a função faz a conversão dos nomes de variáveis para minúsculas.

***as.Dates*:** Esta função reconhece os formatos de data utilizados em *datasets* fornecidos por instituições suecas. Na documentação desta função cita-se INCA e Rockan.

***best_match*:** Esta função tem como objetivo identificar automaticamente possíveis erros no input de dados de um formulário e realizar a correção de forma automática. De acordo com a documentação, este processo é realizado através da utilização *fuzzy string matching*, uma técnica de processamento de linguagem natural onde o objetivo é identificar *strings* que correspondem a um padrão aproximado. Esta técnica é utilizada por corretores ortográficos, como o Microsoft Word.

O pacote desenvolvido por Bulow é *open source* e está disponível em um repositório do *bitbucket* e se distancia do *INCADATABR* devido ao fato de não gerar nenhum tipo de visualização dos dados.

4.2 MODELAGEM DO PACOTE INCADATABR

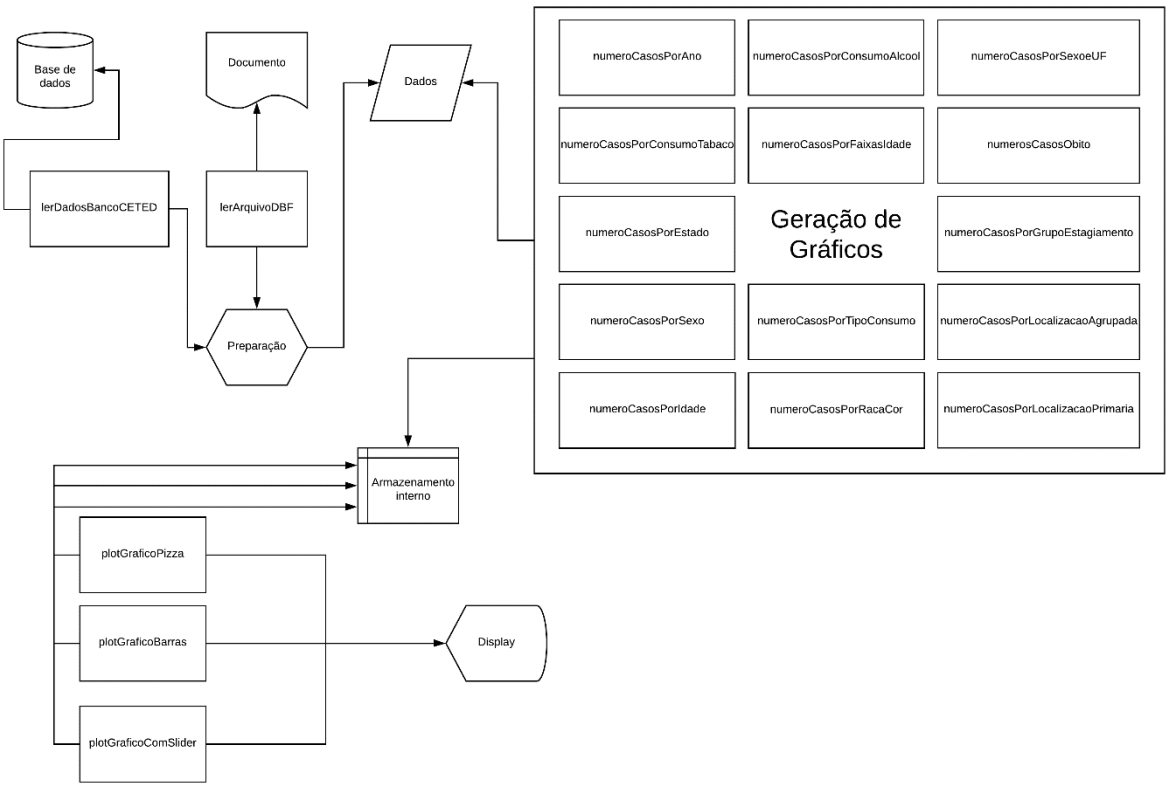
O pacote foi construído baseado no modelo de arquitetura monolítica, um modelo descrito em Engenharia de Software como uma arquitetura em que a aplicação é projetada sem modularidade externa, ou seja, sem a preocupação de construir uma aplicação que possa vir a ser um módulo para uma outra aplicação (LUCIO, 2017).

Devido a isto, o pacote foi segmentado em 5 grupos, sendo eles: exportação de dados, importação de dados, manutenção de dados, geração de gráficos e funções úteis. O diagrama apresentado na Figura 6 apresenta a arquitetura do pacote de forma simplificada. Ressalta-se que, o modo de uso de todas as funções existentes no pacote será detalhado nas subseções 4.2.1, 4.2.2, 4.2.3, 4.2.4, 4.2.5 e no tutorial disponível na seção 5 deste trabalho.

O código do pacote foi disponibilizado em um repositório público no Github[5].

[5] https://github.com/costawilliam/incadatabr

Figura 6 – Arquitetura do pacote (simplificada).



Fonte: Adaptado pelo Autor

4.2.1 GRUPO DE EXPORTAÇÃO DE DADOS

Este grupo tem como objetivo centralizar a responsabilidade pela exportação de dados utilizada pelas funções. Inicialmente, o grupo compunha-se por 4 funções, isto é, criou-se uma função para a exportação de cada formato de arquivo. Ao longo do desenvolvimento do pacote, optou-se por unificá-las em uma única função e permitir que o usuário indique o formato em que deseja exportar os dados através da passagem de parâmetro.

4.2.2 GRUPO DE IMPORTAÇÃO DE DADOS

Este grupo tem como objetivo centralizar a responsabilidade pela importação de dados para a linguagem R, isto é, através das funções presentes neste grupo é possível instanciar *datasets* com dados oriundos de arquivos fornecidos pelo INCA ou oriundos diretamente do banco de dados hospedado no servidor do CETED (Centro

de Tecnologias Digitais da Universidade Feevale). O grupo é composto por duas funções, sendo elas:

lerArquivoDBF: Esta função tem como objetivo a criação de um *dataframe* através de um arquivo com a extensão "*.dbf*" fornecido pelo INCA. A função simplifica a importação deste tipo de arquivo para a linguagem R, visto que, exige apenas que o usuário passe o caminho deste arquivo como parâmetro.

***lerDadosBancoCETED*:** Esta função destaca-se devido a flexibilidade que garante ao pacote, uma vez que, com o uso desta o pacote permite que o usuário possa realizar a geração de gráficos de forma simplificada, não sendo necessário que o *download* dos *datasets* do INCA seja realizado. A função se responsabiliza pela obtenção dos dados e instanciação dos *dataframes* que serão utilizados para a geração de gráficos.

4.2.3 GRUPO DE MANUTENÇÃO DE DADOS

Este grupo tem como objetivo centralizar a responsabilidade de manutenção do banco de dados do CETED e é composto de uma função única nomeada como *inserirRegistrosCeted*. Esta função será utilizada pelo responsável pelo banco de dados hospedado no servidor do CETED, com o objetivo para manter o banco de dados sempre atualizado. Essa função permitirá que a importação de registros existentes em novos *datasets* (posteriores a 2016) seja realizada de forma simplificada. Entre as responsabilidades desta função estão: Validação da existência dos registros para o ano que está sendo importado e ajustes em algumas variáveis, como a remoção de registros que possuem valores ausentes para variáveis importantes para o pacote e a conversão de dados para o formato adequado.

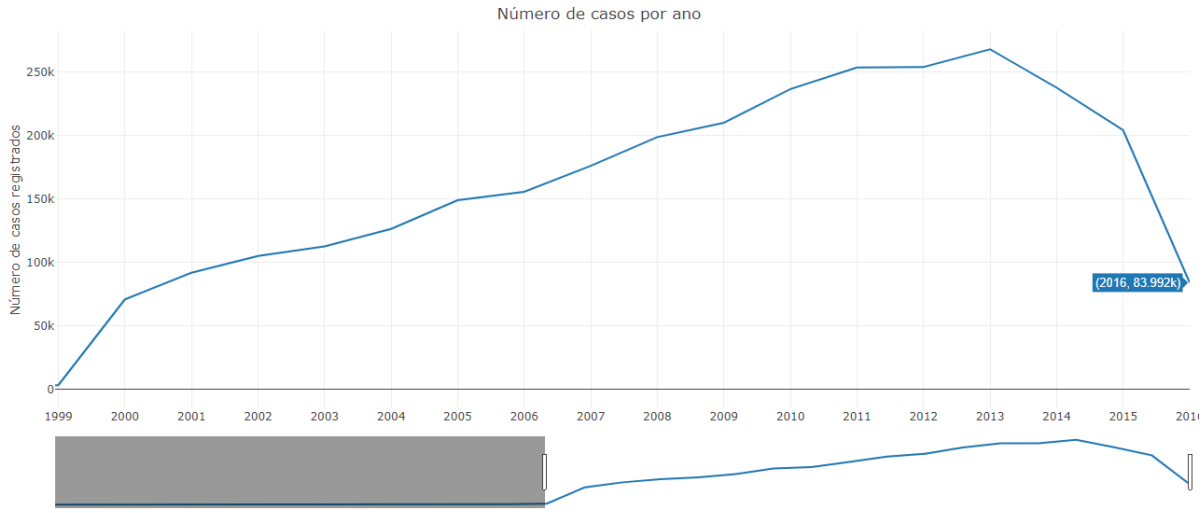
4.2.4 GRUPO DE GERAÇÃO DE GRÁFICOS

Este grupo tem como objetivo centralizar a responsabilidade pela geração dos gráficos e pode ser classificado como o principal grupo do pacote, visto que, será com as funções deste grupo que o usuário interagirá. O grupo é composto por 14 funções sendo estas:

***numeroCasosPorAno*:** Esta função tem como objetivo permitir ao usuário quantificar o número total de casos de câncer registrados em um determinado ano. O

gráfico gerado permite ao usuário ter um panorama geral sobre a situação da doença no país. A Figura 7 apresenta um dos gráficos gerados por esta função.

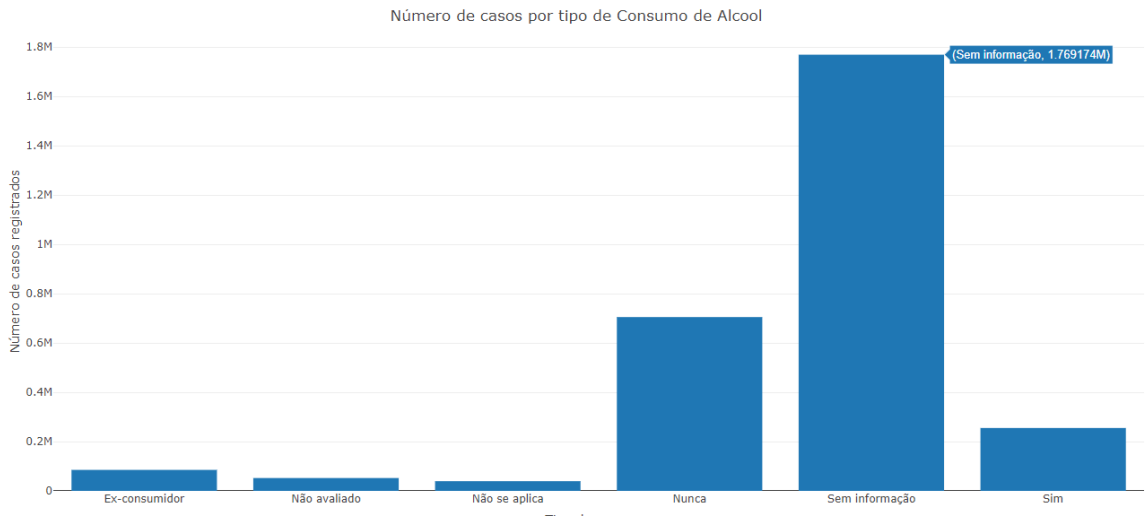
Figura 7 – Gráfico com *slider* da função *numeroCasosPorAno.*



Fonte: Adaptado pelo Autor

numeroCasosPorConsumoAlcool: Esta função tem como objetivo permitir ao usuário quantificar o número total de casos de câncer registrados baseado no consumo de álcool. Esta variável foi elencada como candidata para uma função visto que, de acordo com o INCA (2019), o consumo de bebidas alcoólicas favorece o desenvolvimento de diversos tipos de câncer. A Figura 8 apresenta o gráfico de barras gerado pela função, através deste gráfico é possível observar que, apesar da importância desta variável, uma grande parte dos registros presentes no *dataset* do INCA não possui esta informação.

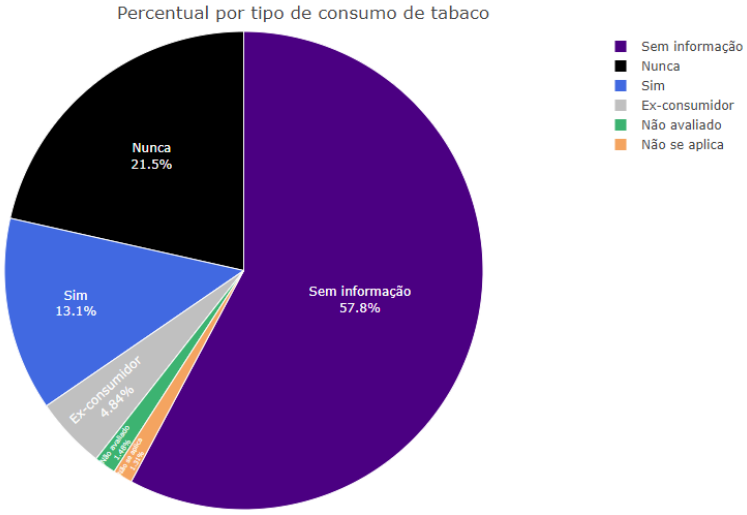
Figura 8 – Gráfico com barras da função *numeroCasosPorConsumoAlcool.*



Fonte: Adaptado pelo Autor

numeroCasosPorConsumoTabaco: Esta função tem como objetivo permitir ao usuário quantificar o número total de casos de câncer registrados baseado no consumo de tabaco. Assim como a variável consumo de álcool, o consumo de tabaco foi definido como variável para uma função visto que, de acordo com a INCA (2019), o tabaco fumado é responsável por 90% dos cânceres de pulmão. Além disso, em uma análise realizada pelo *Centers for Disease Control and Prevention* (2018) indicou que o tabagismo pode ser responsável pela ocorrência câncer em diversas partes do corpo, sendo citadas em sua análise: Bexiga, sangue (leucemia mieloide aguda), colo do útero, cólon e reto (colorretal), esôfago, rim, uretra, laringe, fígado, orofaringe (inclui partes da garganta, língua, palato mole e amígdalas), pâncreas, estômago, traqueia, brônquios e pulmão. A Figura 9 apresenta o gráfico de pizza gerado pela função e permite que o usuário perceba que, a variável que indica o consumo de tabaco não foi informada em uma grande quantidade dos registros do *dataset* do INCA.

Figura 9 – Gráfico com pizza da função *numeroCasosPorConsumoTabaco.*



Fonte: Adaptado pelo Autor

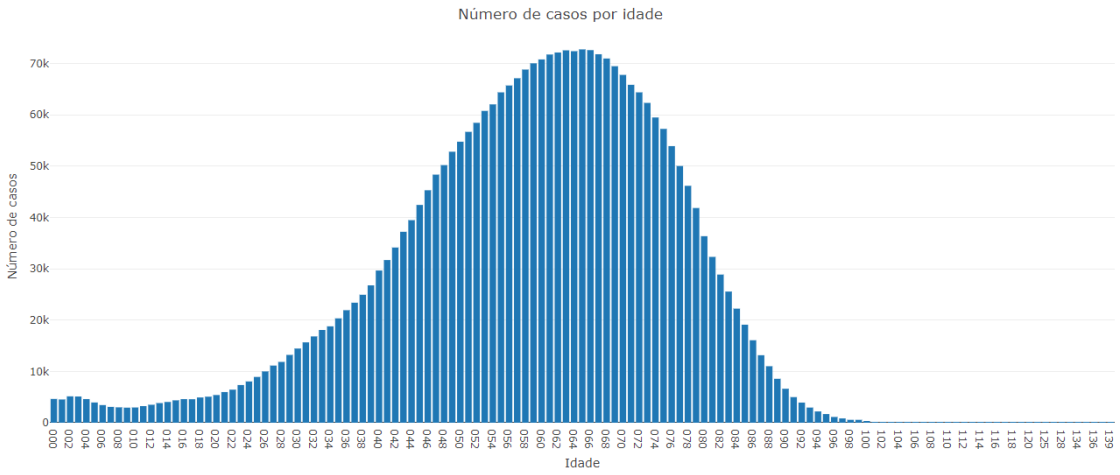
numeroCasosPorEstado*:* Esta função tem como objetivo permitir ao usuário quantificar o número total de casos de câncer por Unidade Federativa e assim identificar os estados com maior ocorrência de câncer no país. A Figura 10 apresenta o gráfico gerado por esta função

Figura 10 – Gráfico com barras da função *numeroCasosPorEstado*.

Fonte: Adaptado pelo Autor

***numeroCasosPorIdade*:** Esta função tem como objetivo permitir ao usuário quantificar o número total de casos por idade. O gráfico de barras gerado por esta função pode ser visualizado na Figura 11.

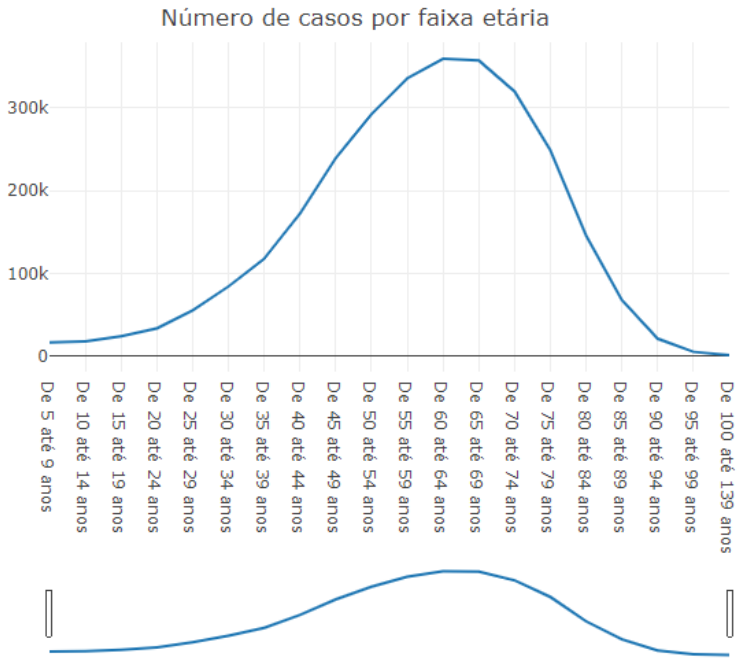
Figura 11 – Gráfico de barras da função *numeroCasosPorIdade*.



Fonte: Adaptado pelo Autor

***numeroCasosPorFaixasIdade*:** Esta função tem como objetivo permitir ao usuário quantificar o número total de casos por faixa etária. A função permite que o próprio usuário indique as faixas que serão utilizadas através da passagem de parâmetro. Devido a isso, esta função pode ser utilizada como um complemento para a análise do número de casos por idade. A Figura 12 apresenta o gráfico de linha com *slider* gerado pela função.

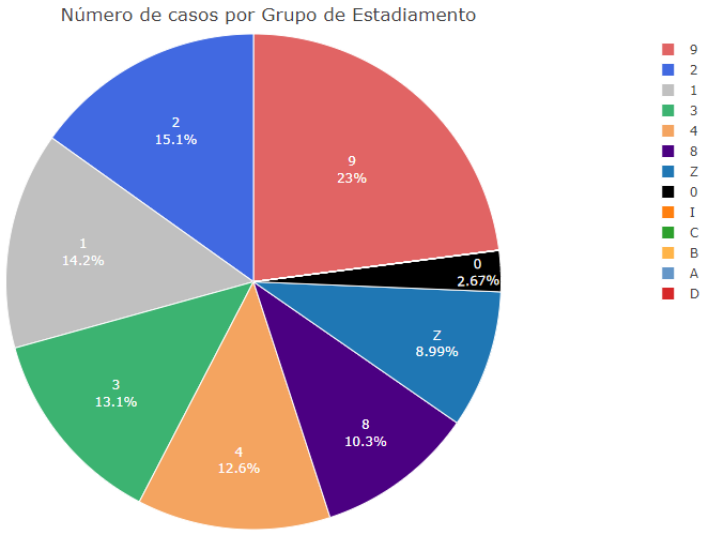
Figura 12 – Gráfico de linha com slider da função numeroCasosPorFaixasIdade.



Fonte: Adaptado pelo Autor

***numeroCasosPorGrupoEstadiamento*:** Como detalhado na seção 3 deste trabalho, a variável estadiamento é utilizada para a avaliação da extensão da neoplasia maligna antes do tratamento. Este gráfico permite ao usuário quantificar o número de casos por grupo de estadiamento. A Figura 13 apresenta o gráfico de pizza gerado por esta função.

Figura 13 – Gráfico de pizza da função numeroCasosPorGrupoEstadiamento

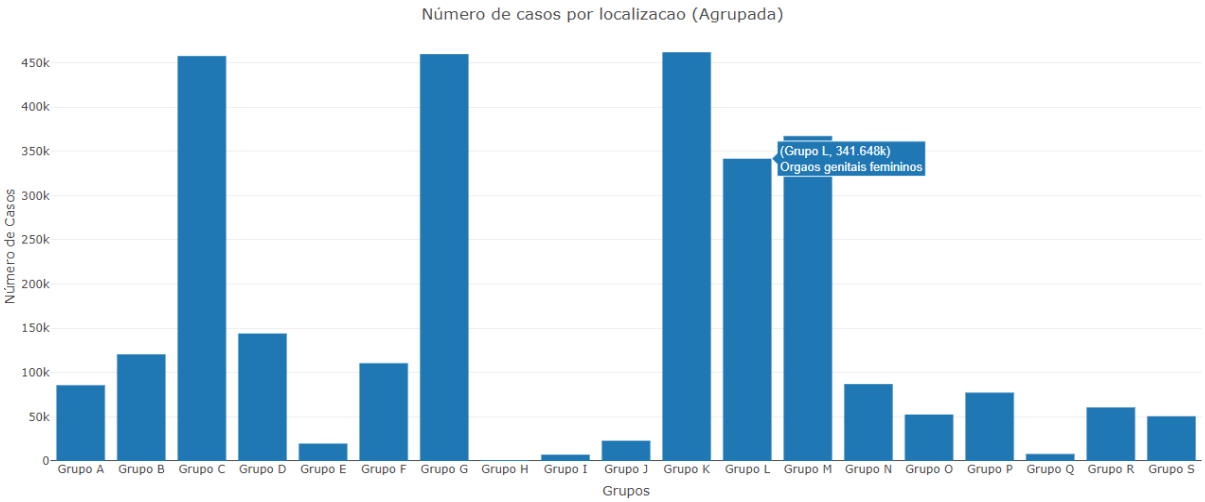


Fonte: Adaptado pelo Autor

***numeroCasosPorLocalizacaoAgrupada*:** Esta função tem como objetivo permitir ao usuário quantificar o número total de casos de acordo com a grupo de localização primária do tumor e, assim, realizar uma análise de quais os agrupamentos tem o maior número de casos.

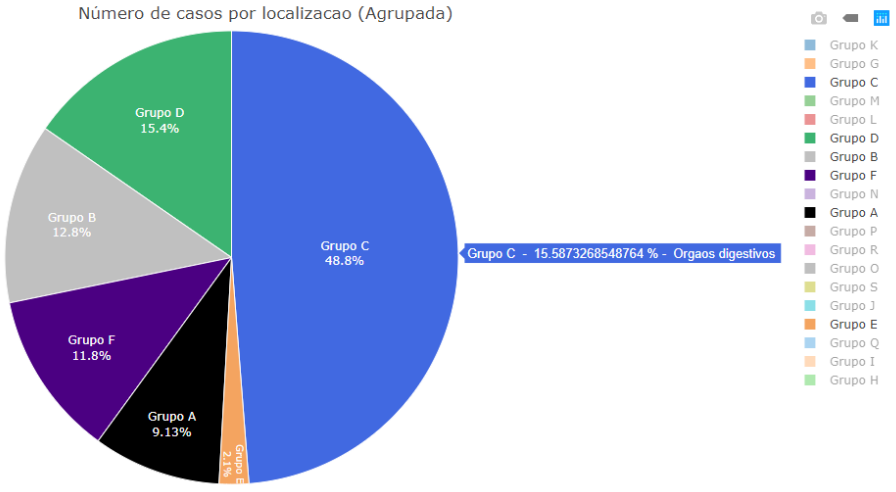
Com o objetivo de simplificar a compreensão do gráfico, ao posicionar o mouse sobre um dos grupos, será exibida a descrição deste grupo, por exemplo, ao posicionar o cursor do mouse sobre o grupo L, será exibido ao usuário a informação de que este grupo corresponde aos casos de câncer em órgãos genitais femininos, já ao posicionar o cursor do mouse sobre o grupo C será exibido ao usuário que este grupo corresponde ao número de casos de câncer em órgãos do sistema digestivo. O gráfico de barras gerado por esta função pode ser visualizado na Figura 14 e na Figura 15 pode-se visualizar o gráfico de pizza desta função com apenas alguns grupos selecionados. Nestas Figuras, pode-se visualizar o balão exibido ao posicionar o mouse sobre um dos grupos com a descrição deste grupo.

Figura 14 – Gráfico de barras da função numeroCasosPorLocalizacaoAgrupada.



Fonte: Adaptado pelo Autor

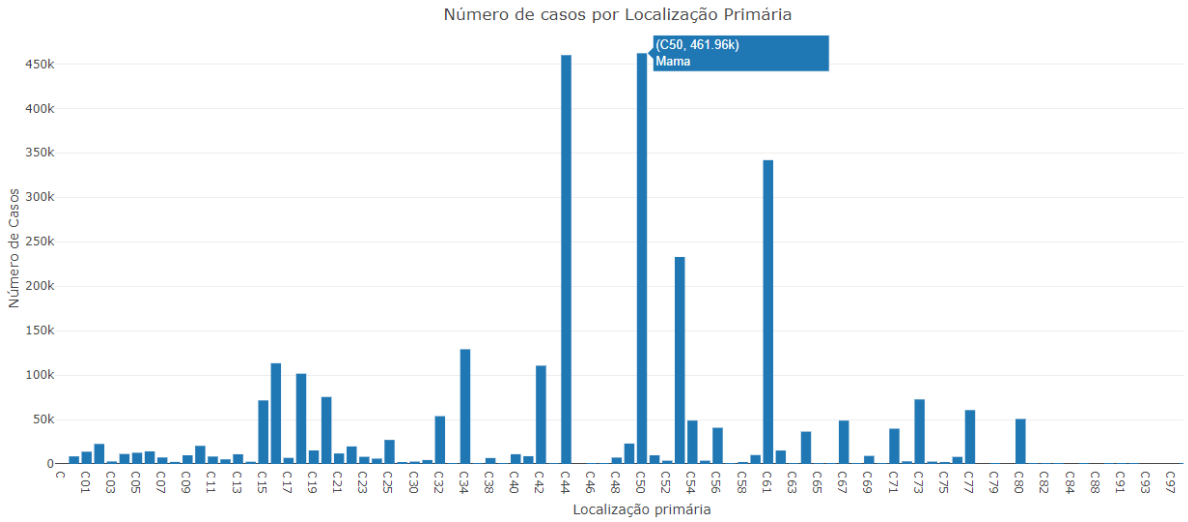
Figura 15 – Gráfico de pizza da função *numeroCasosPorLocalizacaoAgrupada*



Fonte: Adaptado pelo Autor

***numeroCasosPorLocalizacaoPrimaria*:** Esta função tem como objetivo permitir ao usuário quantificar o número total de casos de acordo com a localização primária do tumor, permitindo assim, que o usuário tenha acesso ao número de casos de cada localização primária e possa identificar a localização primária mais comum. A Figura 16 apresenta o gráfico de barras gerado por esta função.

Figura 16 – Gráfico de barras da função *numeroCasosPorLocalizacaoPrimaria.*



Fonte: Adaptado pelo Autor

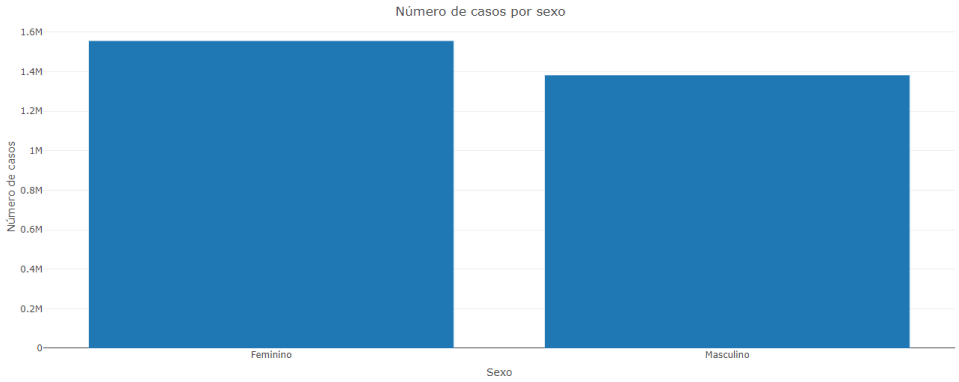
numeroCasosPorRacaCor: Esta função tem como objetivo permitir ao usuário quantificar o número de casos de acordo com a raça/cor do paciente. Assim como em outras variáveis, através do gráfico percebe-se que, em diversos registros presentes no *dataset* do INCA, a variável está definida como “sem informação”. A Figura 17 exibe o gráfico gerado pela função, diretamente no R Studio, com a passagem de alguns parâmetros.

Figura 17 – Utilização da função *numeroCasosPorRacaCor* dentro da *IDE* R Studio.

Fonte: Adaptado pelo Autor

***numeroCasosPorSexo*:** Esta função tem como objetivo permitir ao usuário quantificar o número de casos de câncer por sexo, com o intuito de identificar o sexo mais vulnerável à ocorrência de câncer. Embora esta função desconsidere o tipo de tumor avaliado, permite ao usuário identificar o sexo mais afetado pelo câncer. A Figura 18 apresenta o gráfico gerado por esta função.

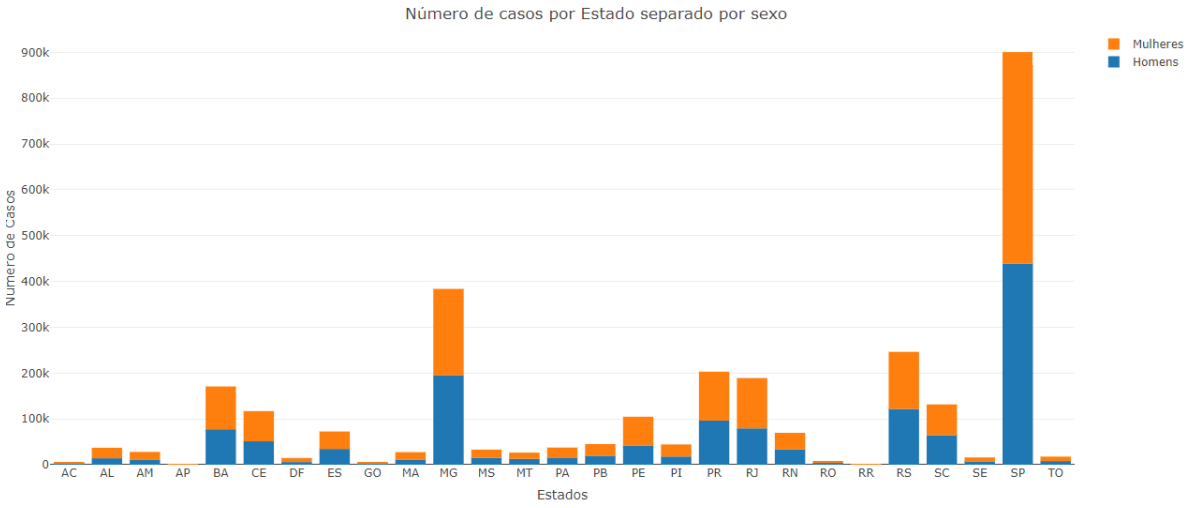
Figura 18 – Gráfico de barras da função *numeroCasosPorSexo*.



Fonte: Adaptado pelo Autor

***numeroCasosPorSexoeUF*:** Esta função acaba sendo um complemento para as funções *numeroCasosPorEstado* e *numeroCasosPorSexo* visto que, consiste em uma combinação das variáveis *UFUH* e *SEXO*. A Figura 19 apresenta o gráfico gerado por esta função. Através desta é possível realizar uma comparação entre os casos de câncer entre homens e mulheres e os estados da federação.

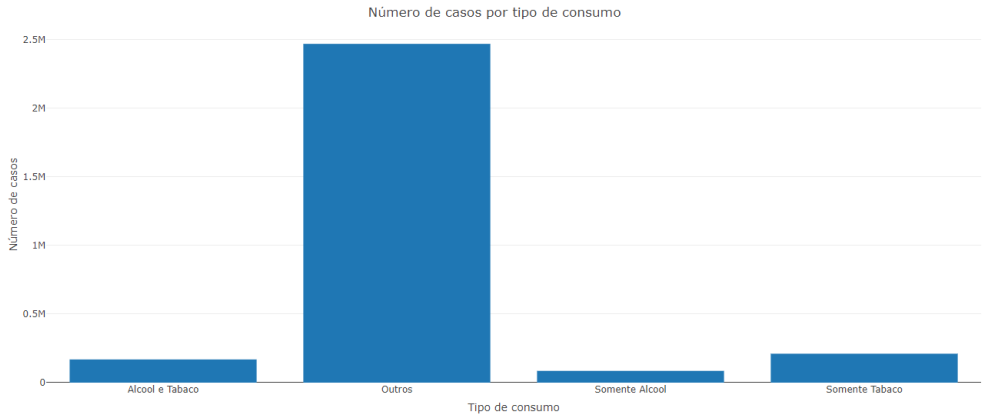
Figura 19 – Utilização da função *numeroCasosPorRacaCor*.



Fonte: Adaptado pelo Autor

***numeroCasosPorTipoConsumo*:** Esta função foi desenvolvida com o objetivo de permitir identificar a relação entre o número de casos de câncer e o tipo de consumo de substâncias diretamente ligadas ao risco de câncer. Para a geração deste gráfico foi necessário realizar a análise das combinações possíveis das variáveis alcoolismo e tabagismo presentes no *dataset* do INCA. A Figura 20 apresenta o gráfico de barras gerado por esta função.

Figura 20 – Gráfico de barras da função *numeroCasosPorTipoConsumo*



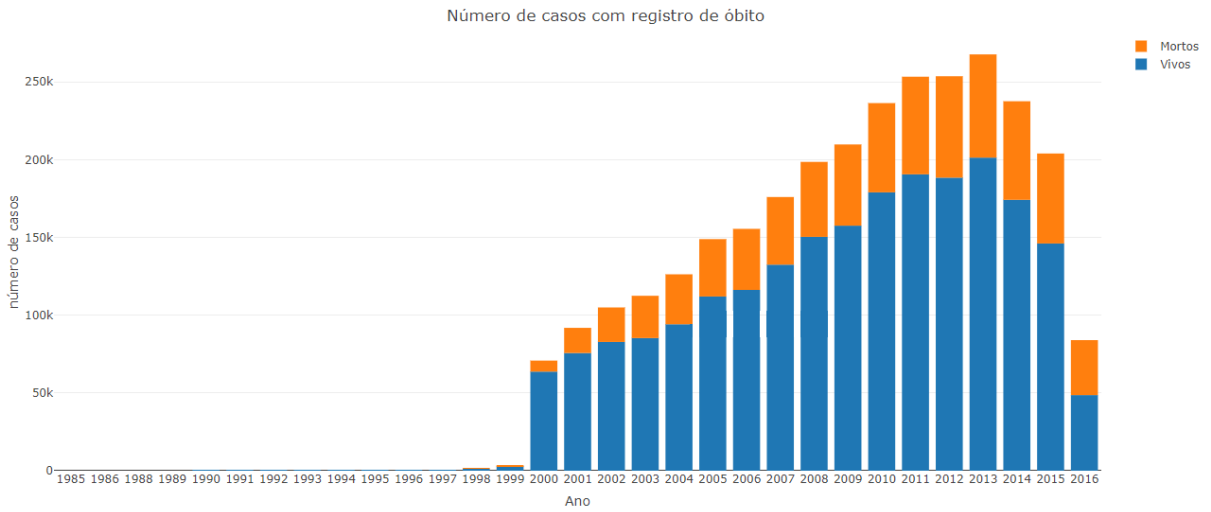
Fonte: Adaptado pelo Autor

***numerosCasosObito*:** Esta função foi desenvolvida devido aos *feedbacks* dos entrevistados por Medinger (2017) em seu trabalho. Com esta função permite-se que o usuário tenha acesso ao número de casos em que houve registro de óbito.

> *Utilizar mais entradas da base de dados do INCA deixando o trabalho mais completo. Seria interessante a inclusão de atributos que constam na base do INCA e que possibilitariam a realização de mais análises. Informações importantes como as listadas abaixo: DATAOBITO – data do óbito, dado para a determinação da taxa de mortalidade e taxa de cura (percentual de pessoas que morrem por conta da doença e que sobrevivem sem novo diagnóstico após cinco anos do término do tratamento). (MEDINGER, 2017, P. 94)*

Para obter esta informação a função baseia-se na variável “data óbito”, isto é, trata-se a data, preenchida no *dataset* no formado dd/mm/aaaa para a extração do ano do óbito e após isso quantifica-se a quantidade de casos com óbito por ano. A Figura 21 apresenta o gráfico gerado por esta função.

Figura 21 – Gráfico de barras da função *numerosCasosObito*



Fonte: Adaptado pelo Autor

4.2.5 GRUPO DE FUNÇÕES GENÉRICAS

Este grupo tem como objetivo centralizar funções classificadas como genéricas, isto é, funções que possam ser chamadas em diversos pontos do pacote. O principal objetivo deste grupo é reaproveitamento de código, visto que, a utilização destas funções evita retrabalho e simplifica a realização de manutenções, sendo este um conceito amplamente difundido no mercado e na academia.

> *[...] é vislumbrada a necessidade de orientar o foco do desenvolvimento para a produção de componentes reutilizáveis de software, que possibilitariam a construção de novas aplicações a partir da integração de componentes previamente implementados e devidamente testados (GUIZZARDI, 2000, p.17 apud GALL et al., 1995)*

Outro benefício ao adotar este conceito está na redução da complexidade das funções, uma vez que, cada função terá uma única responsabilidade. Dentro do pacote *INCADATABR*, isto é aplicado nas funções de geração de gráfico, visto que, cada função será responsável por preparar os dados necessários para geração de um gráfico, mas o processo de plotagem será delegado a uma das funções que possui esta responsabilidade presente neste grupo, isto é, bastará realizar chamada de uma função com os parâmetros requeridos para que o gráfico seja plotado.

Esse grupo é composto por 10 funções, sendo elas:

calcularPercentual: Esta função tem como objetivo abstrair a complexidade do cálculo de percentual. Esta função foi definida como uma função genérica, visto que, este processo é realizado por todas as funções que realizam a geração de gráficos de pizza.

converterFatorParaCaracter**:** Esta função tem como objetivo converter colunas do tipo *factor* (um tipo de dado do R utilizado para lidar com variáveis categorizadas) para *caractere* (formato de dado equivalente a *String* na linguagem Java) e é utilizada devido a uma característica da função *read.dbf* presente no pacote *foreign*, visto que, esta função foi utilizada para conversão de arquivos ".dbf" em *datasets* no R.

converterFatorParaInteiro**:** Esta função é semelhante a função *converterFatorParaCaracter*. Seu objetivo é converter variáveis do f*actor para o tipo numeric* (que pode ser tratado como *integer* no R).

obterDados**:** Esta função responsabiliza-se por criar um *driver* de conexão com o banco de dados PostgreSQL, instanciar a conexão com o banco de dados do CETED e executar uma *query* recebida como parâmetro. O *resultset* desta query será retornado como *dataframe* e este será utilizado para geração dos gráficos.

plotGraficoBarras**:** Esta função responsabiliza-se pela montagem e plotagem dos gráficos de barras. Monta-se o gráfico baseado nos dados recebidos como parâmetro pela função. Para este processo, optou-se pela utilização do pacote *Plotly*, visto que, este traz diversos recursos interativos interessantes para os gráficos gerados, como por exemplo a possibilidade de realizar o *download* do gráfico gerado como imagem, a possibilidade de dar *zoom* sobre uma determinada área do gráfico, entre outros.

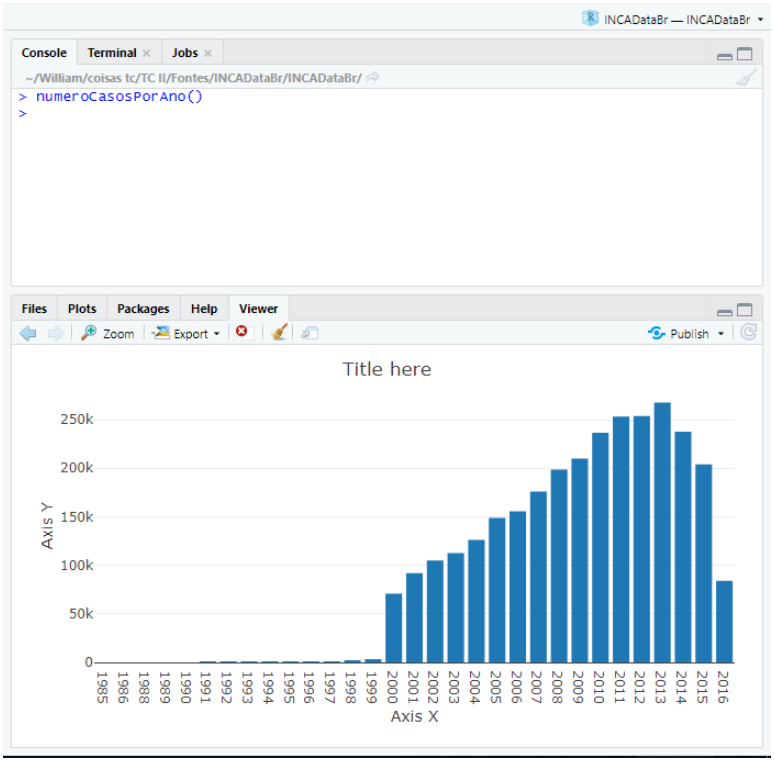
plotGraficoComSlider: Esta função responsabiliza-se pela montagem e plotagem dos gráficos de linhas, comumente utilizados para demostrar a evolução e/ou decréscimo de uma sequência numérica. Assim como na função *plotGraficoBarras*, optou-se em utilizar o *pacote Plotly*.

plotGraficoPizza: Esta função responsabiliza-se pela montagem e plotagem dos gráficos de pizza. Assim como na função plotGraficoBarras, optou-se em utilizar o *pacote Plotly*.

tratarParametros: Esta função responsabiliza-se em iniciar os parâmetros utilizados pelas funções existentes no grupo de geração de gráficos. Optou-se por criar esta função, para permitir que o usuário possa gerar gráficos sem ter que passar

uma grande quantidade de parâmetros (os parâmetros que não foram indicados pelo usuário na chamada de qualquer função de geração de gráfico, serão definidos com valores pré-estabelecidos). Com isso, o usuário poderá gerar um gráfico de forma simplificada. A Figura 22 exemplifica esta situação, nesta pode ser visto que foi feita a chamada da função *numeroCasosPorAno*, sem a passagem de parâmetros e observa-se que o título do gráfico e rótulos dos eixos X e Y estão com valores genéricos.

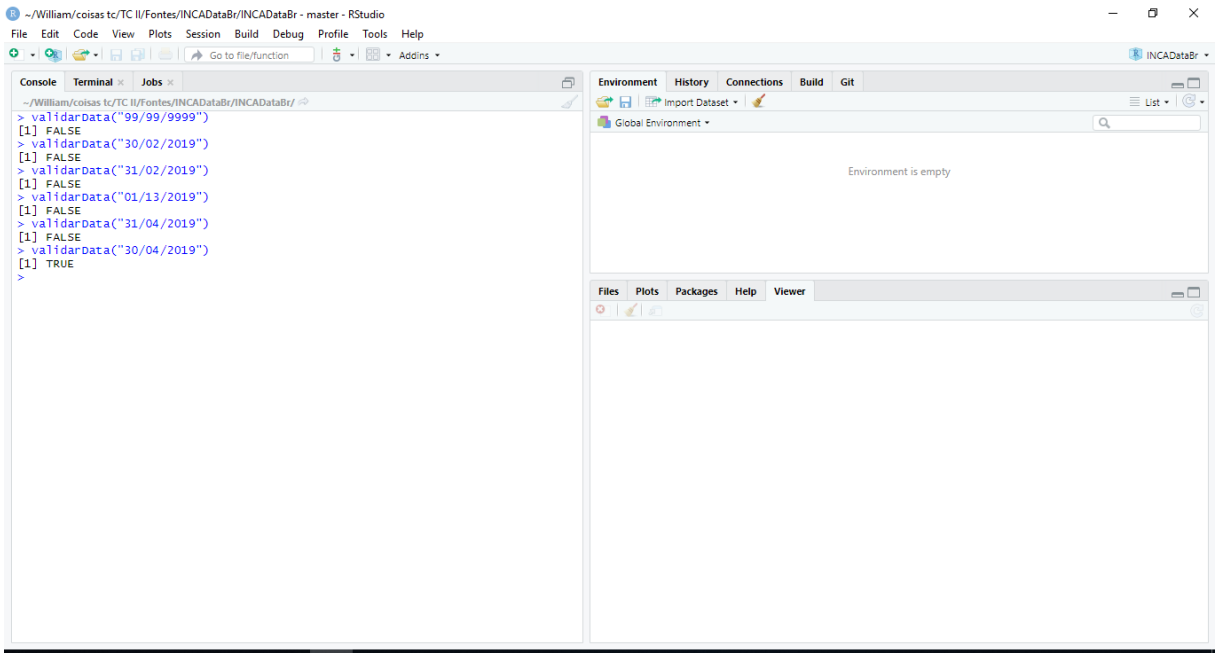
Figura 22 – Gráfico gerado no RStudio sem a passagem de parâmetros.



Fonte: Adaptado pelo Autor

validarData: Esta função tem como objetivo realizar a validação de uma data, sendo que, caso a data seja inválida retorna-se *false* e caso contrário, retorna-se *true*. A Figura 23 apresenta testes realizados com esta função através do console do RStudio.

Figura 23 – Teste da função validarData.



Fonte: Adaptado pelo Autor

verificarPacote: Esta função tem como objetivo verificar se um determinado pacote já se encontra instalado. Caso o pacote já esteja instalado, a função não fará nada, caso não esteja, o pacote será instalado. O objetivo é garantir que pacotes necessários à utilização do *INCADATABR* estejam instalados. Caso não estejam, o próprio pacote realizará a instalação, tornando essa etapa transparente ao usuário.

5 TUTORIAL DE USO DAS FUNÇÕES PARA PROFISSIONAIS DA ÁREA DA SAÚDE

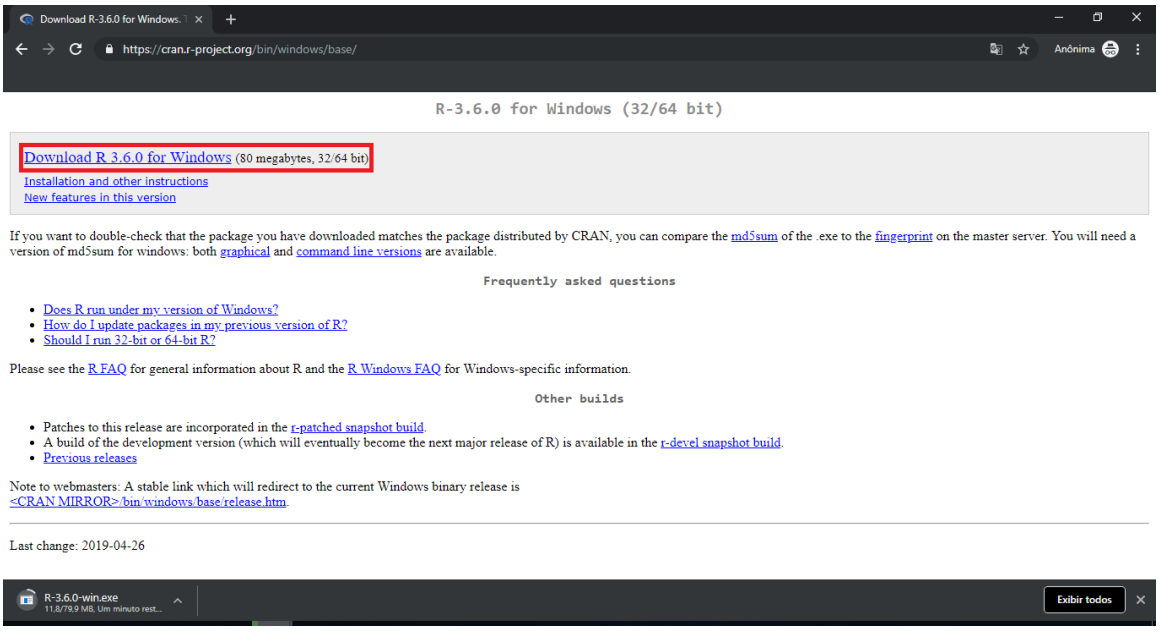
Este capítulo tem como objetivo esclarecer o método de funcionamento do pacote *INCADATABR*. Ao longo dele, é explicado o processo de instalação do pacote e o uso de suas funções. O pacote permite que o usuário realize a geração dos gráficos de duas formas, sendo elas: Utilizando registros do banco de dados do CETED e utilizando dados oriundos de um arquivo *dbf* fornecido pelo INCA. O uso destes dois métodos será explicado na subseção 5.1.4.

Este tutorial é destinado a usuários da plataforma Windows. Para realizar a instalação do pacote é necessário que o usuário já possua o R instalado e recomenda-se a utilização da *IDE* RStudio. Os procedimentos de instalação serão detalhados nas 3 seções seguintes.

5.1.1 INSTALAÇÃO DA LINGUAGEM R

O processo de instalação do R é bastante simples e será detalhado nesta subseção. O primeiro passo é realizar o download do instalador do R no site oficial (https://cran.r-project.org/bin/windows/base/). Ao acessar a página clique no link destacado na Figura 24.

Figura 24 – Site oficial da linguagem R.



Fonte: Adaptado pelo Autor

Após concluir o *download*, localize o diretório em que o arquivo foi salvo e dê um duplo clique sobre ele para que a instalação seja iniciada. Ao iniciar o instalador, será necessário selecionar o idioma desejado e clicar em OK. Feito isso, clique no botão “próximo” 6 vezes, até que o instalador inicie a descompactação dos arquivos no diretório indicado durante o processo de instalação.

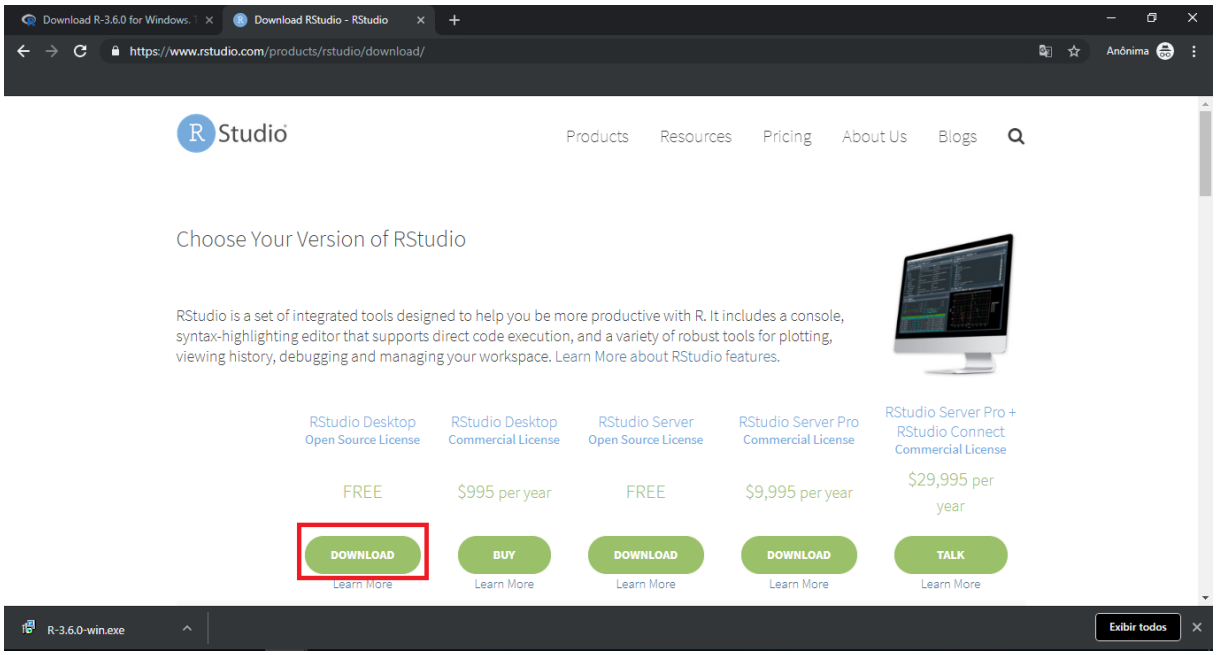
Aguarde que o processo de instalação seja finalizado e quando isto ocorrer clique no botão “Concluir”. Ressalta-se que o processo de instalação pode levar alguns minutos. Após concluir este processo, é possível realizar a instalação da *IDE* RStudio, descrita na subseção seguinte.

5.1.2 INSTALAÇÃO RSTUDIO

Após realizar a instalação do R, será necessário realizar a instalação da *IDE RStudio*, este processo será especificado nesta subseção.

Para realizar o *download* da *IDE*, acesse o site oficial desta[6] e localize o botão da versão desktop adequada (de acordo com o licenciamento desejado). Neste caso, será realizada a instalação da versão gratuita, conforme destacado na Figura 25.

Figura 25 – *Site* oficial da *IDE RStudio*



Fonte: Adaptado pelo Autor

[6] https://www.rstudio.com/products/rstudio/download/

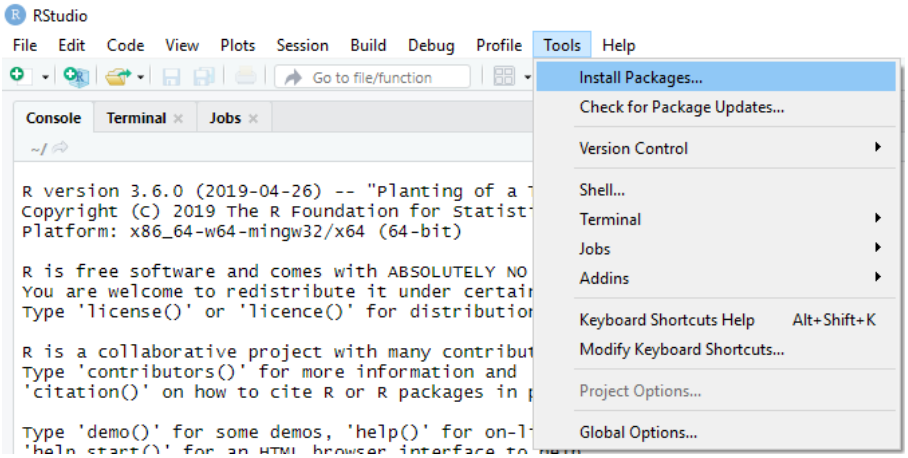
Ao clicar sobre o botão, será direcionado para a área do site em que as versões disponíveis são listadas. Selecione a versão para Windows. Ao fazer isto, o processo de download do instalador da *IDE* será iniciado.

Após o download concluído, localize o diretório em que este foi salvo e dê um duplo clique para que o processo de instalação seja iniciado. A *IDE* possui um processo de instalação simples, basta clicar em Próximo duas vezes e em seguida permitir que a instalação seja iniciada, clicando sobre o botão Instalar. Feito isso, aguarde até que os arquivos sejam descompactados. Este processo pode levar alguns minutos. Ao concluir este processo, é possível seguir para a instalação do pacote, especificada na subseção seguinte.

5.1.3 INSTALAÇÃO DO PACOTE VIA ARQUIVO

Após realizar a instalação do *RStudio*, localize-o em seu menu iniciar e inicie a aplicação. Feito isso, clique no menu "*Tools"* e em seguida na opção "*Install Packages"*, conforme exibido na Figura 26.

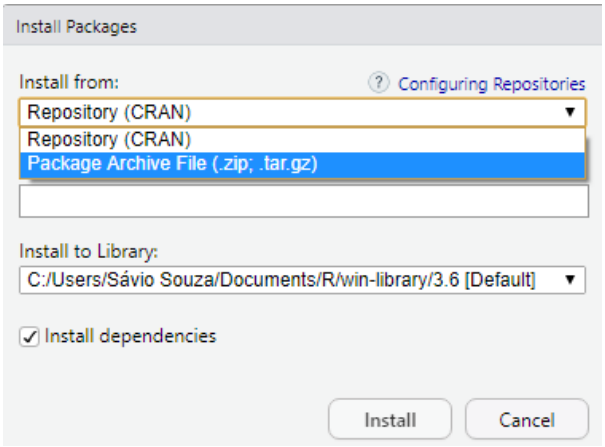
Figura 26 – Menu de Instalação de pacotes na *IDE RStudio*.



Fonte: Adaptado pelo Autor

Feito isso, será aberta uma tela, onde deverá ser indicada a origem do pacote que será instalado, neste caso, utilizaremos a opção "*Package Archive File* (.zip;.tar.gz)", como pode ser visto na Figura 27

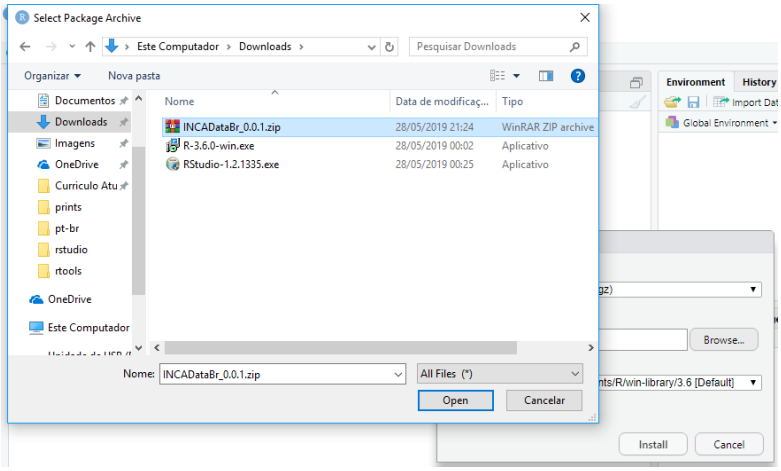
Figura 27 – Tela de instalação de pacotes na *IDE RStudio*



Fonte: Adaptado pelo Autor

Após isso, clique sobre o botão "*Browse*" e localize o arquivo *zip* do pacote, conforme exemplificado na Figura 28.

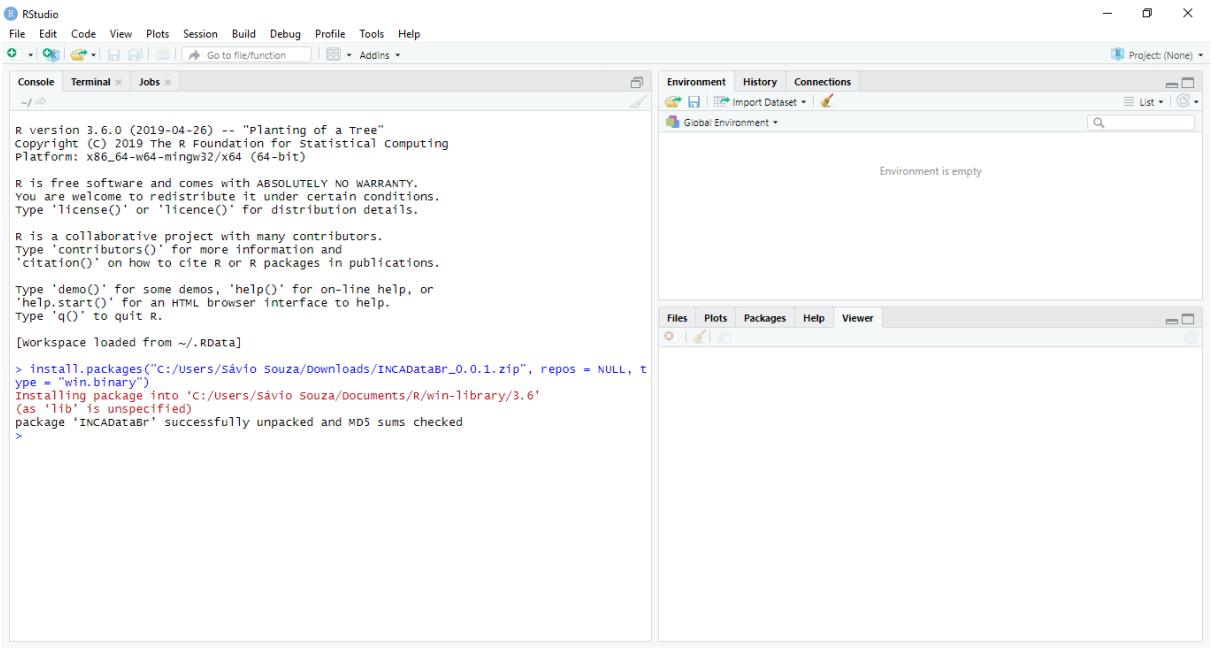
Figura 28 – Seleção do arquivo zip do pacote



Fonte: Adaptado pelo Autor

Em seguida, clique em "*Install*" e aguarde que o processo de instalação seja concluído. O processo será concluído quando o console da IDE voltar a exibir um ">", como pode ser visto na Figura 29.

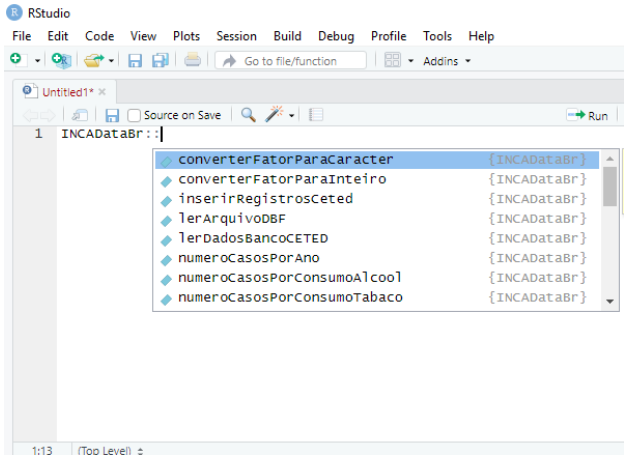
Figura 29 – Instalação do pacote concluída



Fonte: Adaptado pelo Autor

Após a instalação do pacote, deve ser aberto um novo arquivo de Script R através do menu "*File > New File > R Script"* e neste digitar o nome do pacote seguido de 2 dois-pontos e a IDE automaticamente listará as funções disponíveis no pacote INCADATABR, como pode ser visto na Figura 30.

Figura 30 – Auto completar da IDE *RStudio*.

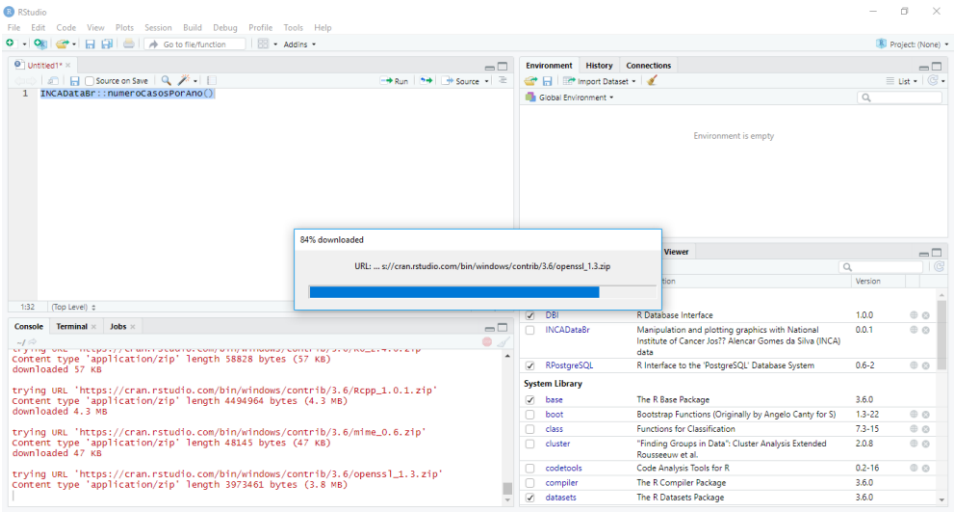


Fonte: Adaptado pelo Autor

Na primeira execução do *INCADATABR*, será realizado automaticamente o *download* e instalação de pacotes utilizados por ele. Este processo poderá levar alguns minutos e só será realizado na primeira execução, visto que, para posteriores

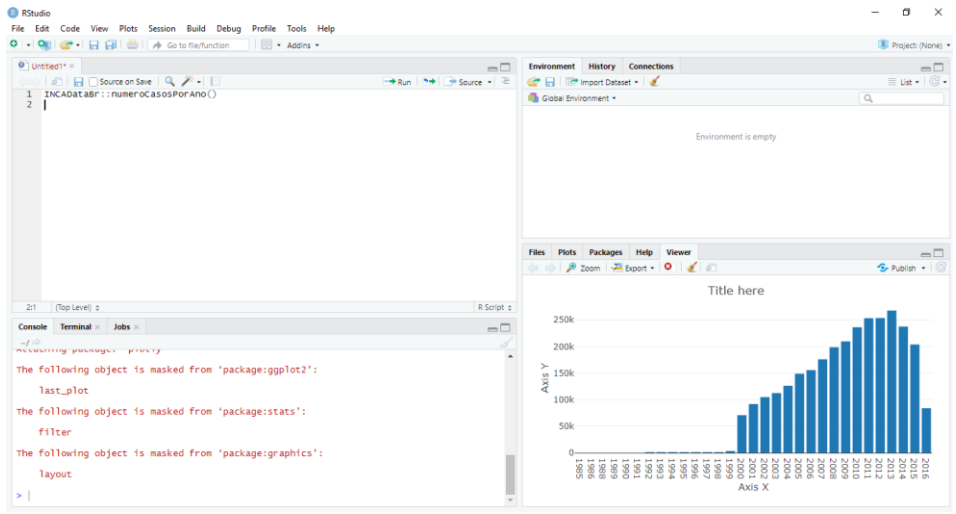
execuções os pacotes necessários já estarão instalados. As Figuras 31 e 32 demonstram este processo.

Figura 31 – *INCADATABR* realizado o *download* e instalação dos pacotes necessários para a execução da função.



Fonte: Adaptado pelo Autor

Figura 32 – Processo finalizado e gráfico gerado na *IDE.*



Fonte: Adaptado pelo Autor

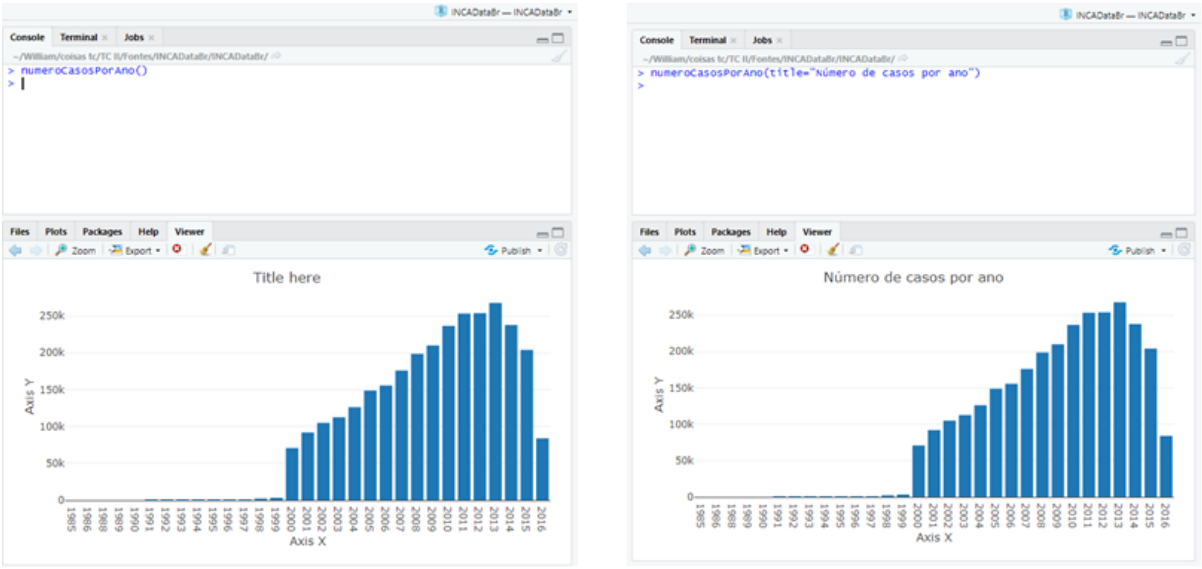
Após realizar a instalação do pacote, o usuário poderá utilizar todas as funções disponíveis neste.

5.1.4 USO DE FUNÇÕES DO PACOTE COM A PASSAGEM DE PARÂMETROS

Como mencionado neste trabalho, o pacote *INCADATABR* foi desenvolvido com o intuito de ser simples para o usuário, com isto, o usuário pode realizar a geração de gráficos sem a passagem de nenhum parâmetro. Porém, com o intuito de ser mais completo, o pacote permite que o usuário utilize uma série de parâmetros, abaixo estes serão apresentados. Salienta-se que estes parâmetros são de uso opcional.

Title: Este parâmetro permite que o usuário personalize o título do gráfico gerado. Caso este parâmetro seja omitido pelo usuário, o gráfico será gerado com o título padrão "*Title here*". A Figura 33 apresenta o gráfico quando o parâmetro é omitido (esquerda) e quando o parâmetro está presente na chamada da função (direita).

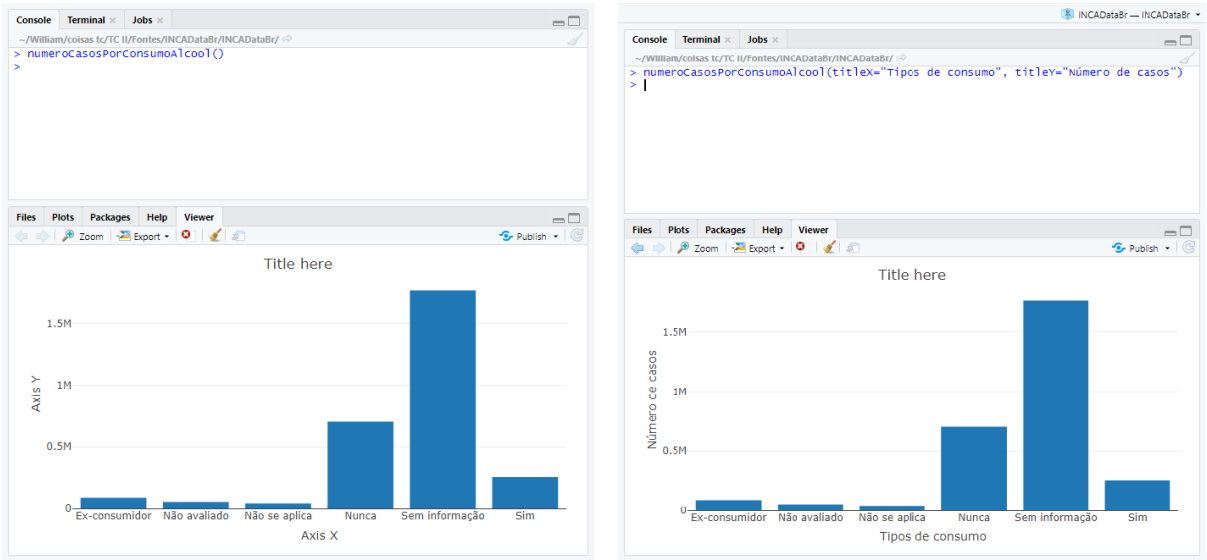
Figura 33 – Gráfico gerado com a omissão do parâmetro *Title* (esquerda) e com o parâmetro presente na chamada da função (direita)



Fonte: Adaptado pelo Autor

TitleX* e *TitleY: Estes parâmetros permitem ao usuário personalizar o texto presente nos eixos X e Y do gráfico. Caso estes parâmetros sejam omitidos, irão receber os valores padrão (*Axis X* e *Axis Y*). A Figura 34 apresenta o gráfico quando os parâmetros são omitidos (esquerda) e quando os parâmetros estão presentes na chamada da função (direita). Vale lembrar que, caso o usuário queria, este pode omitir apenas um dos parâmetros.

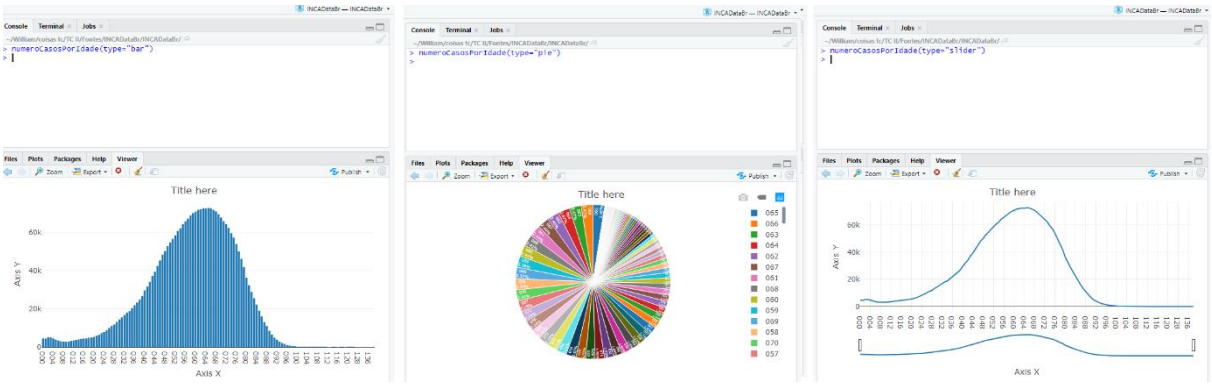
Figura 34 – Gráfico gerado com os parâmetros *TitleX e TitleY* sendo omitidos pelo usuário (esquerda) e com a passagem de ambos (direta)



Fonte: Adaptado pelo Autor

Type: O parâmetro *type* define o tipo de gráfico que será gerado (barras, pizza ou linha). Caso o parâmetro seja omitido, o parâmetro receberá o valor padrão "*bar*" que irá realizar a geração de um gráfico de barras. A Figura 35 apresenta os gráficos gerados com os 3 valores possíveis (*bar*, *pie* e *slider*).

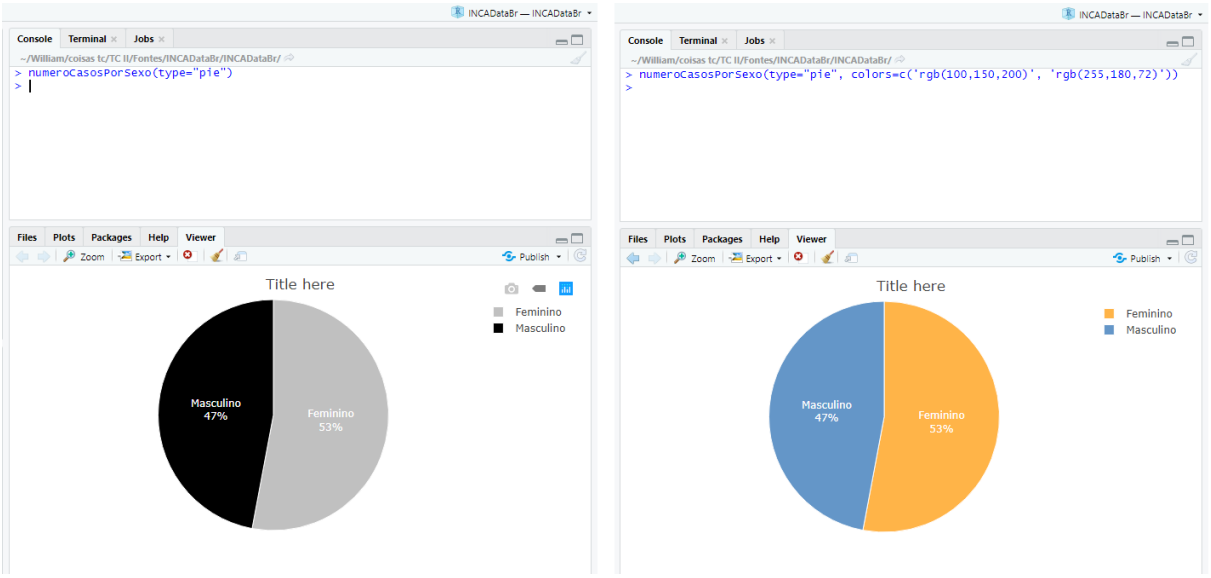
Figura 35 – Gráfico gerado com os valores possíveis para o parâmetro *Type* *bar* (direita), *pie* (centro) e *slider* (esquerda)



Fonte: Adaptado pelo Autor

***Colors*:** Permite que o usuário indique um vetor de cores (RGB) para a geração do gráfico. Este parâmetro é utilizado somente nos gráficos de pizza. A Figura 36 exemplifica o uso deste parâmetro.

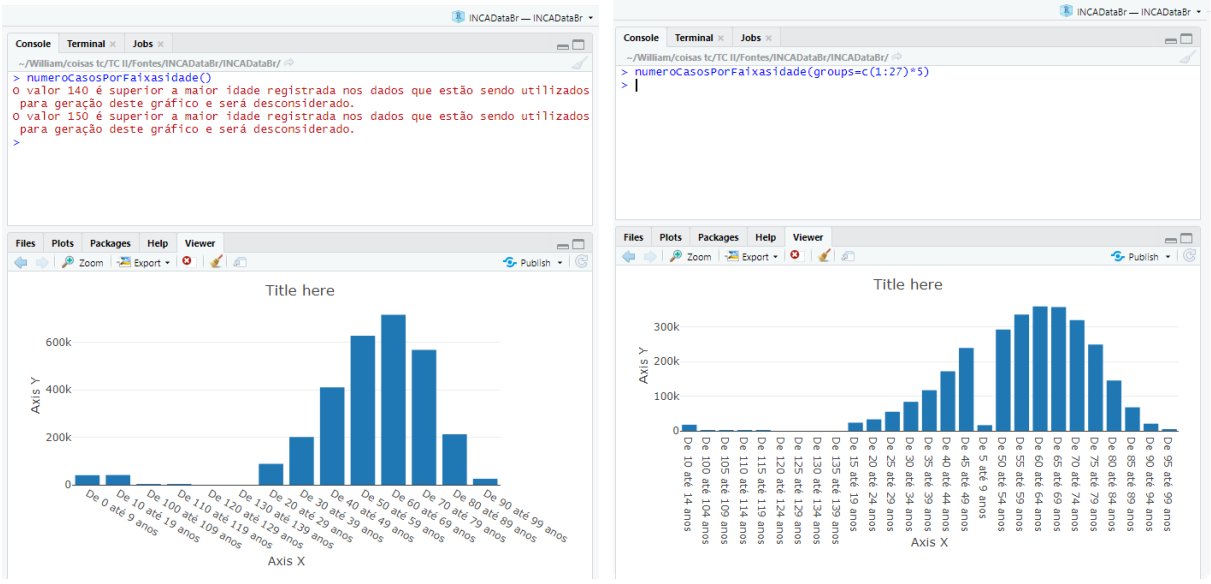
Figura 36 – Gráfico gerado com o parâmetro *colors* omitido e com a passagem deste.



Fonte: Adaptado pelo Autor

Groups: Este parâmetro é utilizado pela função *numeroCasosPorFaixasidade* e permite que o usuário crie quantas faixas desejar. Caso o parâmetro seja omitido, os grupos serão de 10 anos (de 0 até 9, de 10, até 19...). Por padrão, são criados 15 grupos de 10 anos. A função irá exibir um alerta caso a idade máxima registrada seja inferior aos valores indicados pelo usuário. A Figura 37 exibe o gráfico sendo gerado sem a passagem (esquerda) e com a passagem (direita) do parâmetro *groups.*

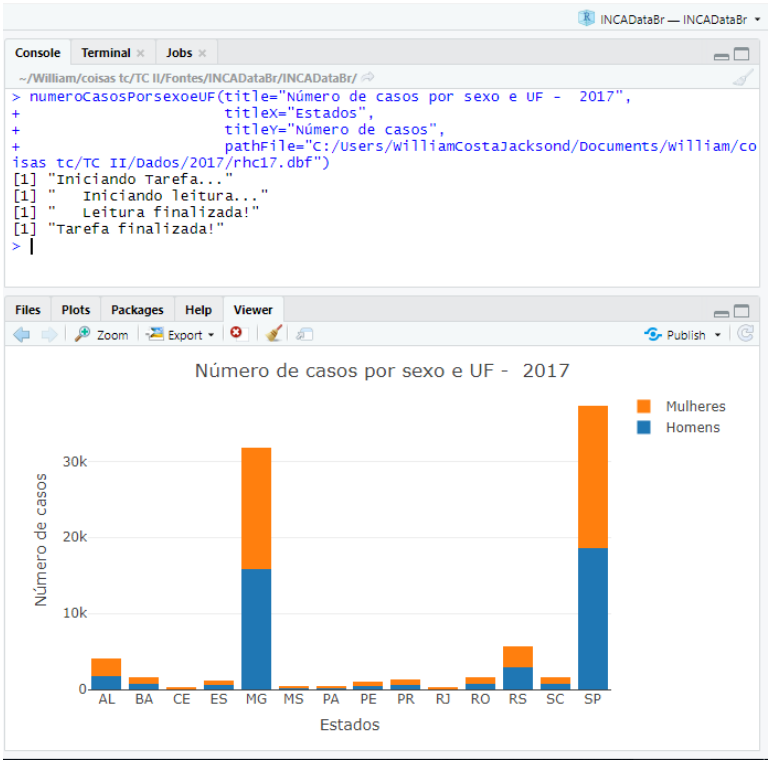
Figura 37 – Gráfico gerado com o parâmetro *groups* omitido e com a passagem deste



Fonte: Adaptado pelo Autor

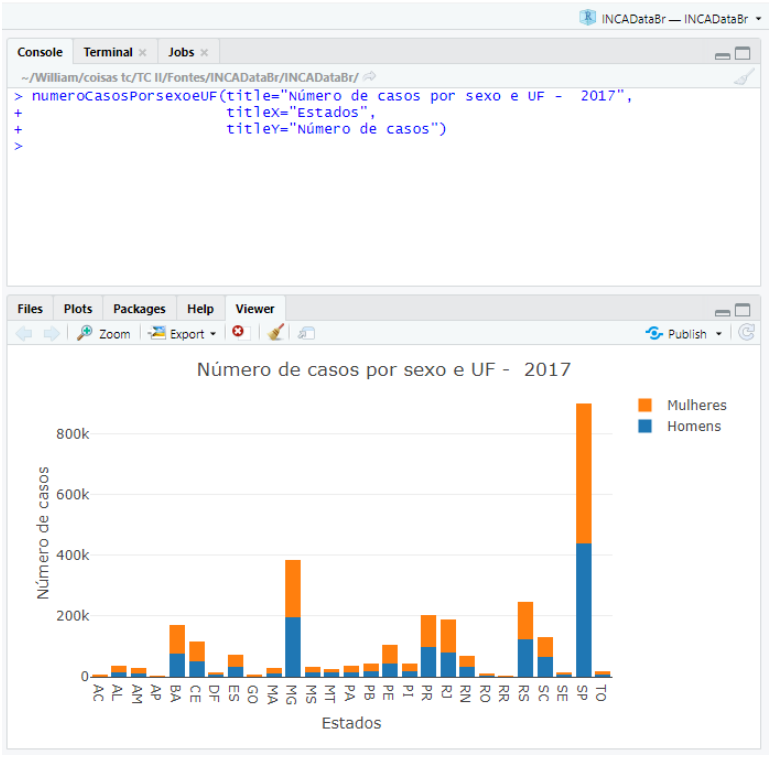
***PathFile*:** Como mencionado na subseção 5.1.1, o pacote *INCADATABR* permite que o usuário realize a geração de gráficos utilizando os registros que estão no banco de dados do CETED ou utilizando dados oriundos de arquivos *dbf* fornecidos pelo INCA. Esta escolha será realizada com base no parâmetro *PathFile*. Caso o usuário indique este parâmetro, o pacote tentará realizar a leitura dos dados do arquivo e então irá realizar o processamento sobre estes dados para realizar a geração do relatório. Caso o usuário omita este parâmetro, o pacote utilizará os registros disponíveis no banco de dados hospedado pelo CETED. A Figura 38 exemplifica a utilização de uma das funções com base em dados disponíveis em um arquivo *dbf* e a Figura 39 apresenta a geração do mesmo gráfico, baseado nos registros disponíveis no banco de dados hospedado no servidor do CETED.

Figura 38 – Geração de gráfico utilizando dados de um arquivo *dbf*



Fonte: Adaptado pelo Autor

Figura 39 – Geração de gráfico utilizando dados do banco de dados do CETED



Fonte: Adaptado pelo Autor

5.2 ASPECTOS IMPORTANTES

Nesta seção serão detalhados aspectos importantes sobre o pacote *INCADATABR*, como vantagens em utilizar o pacote, limitações existentes nele, estimativas de tempo de processamento das funções, problemas enfrentados durante o processo de desenvolvimento e, finalizado com propostas para trabalhos futuros.

5.2.1 BENEFÍCIOS DO USO DO PACOTE

O pacote *INCADATABR* permite a visualização de informações relativas aos casos de câncer registrados no Brasil, baseado nos dados fornecidos pelo INCA (RHC) de forma simples e rápida. Com o uso do pacote, um usuário com conhecimentos básicos em R poderá realizar a geração de gráficos relativamente complexos com o mínimo esforço. Outro benefício é o tempo de processamento e geração destes gráficos, visto que, em poucos segundos, eles já estarão disponíveis para o usuário.

5.2.2 LIMITAÇÕES

O pacote apresenta limitações nos filtros existentes para a geração de gráficos, isto é, atualmente os filtros não estão disponíveis para todas as funções presentes no pacote e o número de filtros ainda é pequeno.

Outra limitação que pode ser apontada está no número de variáveis utilizadas para a geração de gráficos, visto que, o *dataset* do INCA possui um grande número de variáveis e atualmente 14 são utilizadas para a geração de gráficos.

5.2.3 TEMPO DE PROCESSAMENTO E PROBLEMAS DURANTE O PROCESSO DE DESENVOLVIMENTO

Ao iniciar o desenvolvimento do pacote, concebeu-se a ideia de permitir que o usuário utilizasse arquivos *dbf* fornecidos pelo INCA para a geração dos gráficos e que, fosse possível utilizar os registros de *datasets* do INCA já importados em um banco de dados *PostgresSQL* hospedado no servidor do CETED.

Ao iniciar o trabalho de desenvolvimento do pacote para permitir que as funções presentes neste utilizassem as informações armazenadas no banco de dados do CETED deparou-se com um problema de performance devido ao grande volume de registros existente na base de dados (atualmente há mais de 2.000.000 de registros na base de dados mencionada). As limitações de velocidade de tráfego de dados existentes hoje gerariam lentidão na geração dos gráficos, visto que, para cada função executada seria necessário trafegar todos estes dados através da Internet. Isso deixaria o servidor ocupado por um grande período de tempo para cada função que fosse executada pelo usuário.

Devido a isso, buscou-se por soluções alternativas. A estratégia inicial fora que, em uma primeira execução, todos os mais de 2.000.000 de registros fossem trafegados via rede e armazenados em um arquivo local. Com essa estratégia, a lentidão na geração dos gráficos seria limitada a primeira execução. Iniciou-se o desenvolvimento deste modelo, porém, após alguns testes, observou-se que desta forma o usuário jamais teria seus dados atualizados, isto é, mesmo que uma nova carga de dados fosse realizada no banco de dados do CETED, o usuário estaria sempre consumindo os dados de sua primeira execução.

Após isso, analisou-se os dados utilizados pelas funções de geração de gráfico e observou-se que o conjunto completo jamais era utilizado, uma vez que, as funções

sempre buscavam quantificar o número de casos semelhantes, baseadas em uma ou mais variáveis, ou seja, o grande número de registros era trafegado via rede com o objetivo de quantificar o número de casos.

A partir disto, buscou-se alternativas que permitissem ter acesso ao número de casos semelhantes, baseado em uma ou mais variáveis e identificou-se que a execução de uma consulta no banco de dados, que retornasse a quantidade de registros agrupado por uma ou mais variáveis atenderia a necessidade de todas as funções que já haviam sido definidas. Por exemplo, ao invés de realizar o download de todos os registros para então encontrar o número de casos de câncer registrado por ano, utilizamos uma consulta utilizando a linguagem *Structured Query Language* (SQL) que realizasse a contagem de casos por ano e retornasse esta informação ao R, para que então, o processamento destas informações e geração do gráfico fossem realizados. A partir disto, realizou-se alguns testes para verificar a performance deste tipo de operação em um banco de dados remoto e com resultados positivos. Este modelo foi adotado para a obtenção dos dados utilizados nas funções do pacote.

5.2.4 TRABALHOS FUTUROS

Como mencionou-se na subseção 4.4.2, o pacote possui algumas limitações e estas podem ser exploradas em possíveis trabalhos futuros. Acredita-se que o maior ganho estaria na expansão das variáveis utilizadas para a geração de gráficos, visto que, com isto o usuário poderia ter novas visualizações e percepções sobre os dados fornecidos pelo INCA. Ainda, a adição de novos filtros para as funções já existentes permitiria ao usuário realizar uma análise mais detalhada de dados que ele classifica como importantes.

6 CONCLUSÃO

O estudo bibliográfico realizado nesta primeira etapa do trabalho permitiu identificar a relevância da biblioteca que foi desenvolvida para pesquisadores da área da saúde. Também foi possível compreender a situação da doença no país e no mundo e conhecer um pouco sobre iniciativas que buscam promover as pesquisas nesta área.

A estrutura dos *datasets* do RHC disponibilizado pelo INCA foi avaliada e nesta análise foi possível identificar que estes possuem informações de grande relevância para a realização de estudos sobre a doença, seu estado atual e acompanhamento do cenário epidemiológico do câncer no Brasil.

Ao longo do desenvolvimento da segunda etapa deste trabalho foram realizadas pesquisas para identificar a melhor forma de desenvolver o pacote e de como estruturá-lo. Após essa avaliação, a implementação da biblioteca foi realizada utilizando a linguagem R. Para isso foi necessário conhecimento técnico em computação e conhecimento na sobre o dataset do INCA.

A modelagem do pacote adotada permitiu ao usuário realizar a geração de uma série de gráficos de forma simplificada em sua *IDE,* com tempo de execução satisfatório, porém. Os filtros existentes nas funções ainda são limitados e a implementação de novos filtros trará melhorias ao pacote.

O pacote atingiu os objetivos definidos, visto que: (a) atende usuários com pouco conhecimento em computação, pois, para utilizar as funções de geração o usuário pode simplesmente realizar a chamada da função sem a passagem de nenhum parâmetro ou dado; (b) foram avaliados e selecionados atributos presentes nos *datasets* do INCA baseado em informações descritas por profissionais da saúde no trabalho de Medinger (2017) e documentos do próprio INCA; (c) foram realizadas visualizações para diversas variáveis existentes nos *datasets* disponibilizados pelo INCA sem a necessidade de ajustes por parte do usuário e este estará disponível para profissionais da saúde com um complemento as ferramentas já utilizadas, como, por exemplo, o TabWin.

Embora ainda existem diversos pontos em que o pacote possa ser melhorado, como a ampliação do número de variáveis utilizadas para geração de gráficos e adição de novos filtros, o pacote tem potencial de tornar-se uma ferramenta relevante para

profissionais da área da saúde devido a facilidade de uso e ao bom desempenho na geração dos gráficos.

REFERÊNCIAS BIBLIOGRÁFICAS

AMARAL, Fernando. **Introdução à Ciência de Dados: mineração de dados e big data**. Alta Books Editora, 2016.

BARRA, Daniela Couto Carvalho et al. **Evolução histórica e impacto da tecnologia na área da saúde e da enfermagem**. Revista Eletrônica de Enfermagem, v. 8, n. 03, p. 422-430, 2006. Disponível em: <http://ww.fen.ufg.br/revista/revista8_3/pdf/v8n3a13.pdf>. Acesso em: 01 novembro 2018.

BEASLEY, Colin Robert. **BIOESTATÍSTICA USANDO R: APOSTILA DE EXEMPLOS PARA O BIÓLOGO**. 2004. Disponível em: <https://cran.biodisk.org/doc/contrib/Beasley-BioestatisticaUsandoR.pdf>. Acesso em: 14 novembro 2018.

BRASIL. Esplanada dos Ministérios. **Integração de informações dos registros de câncer brasileiros**. Rev Saúde Pública, v. 41, n. 5, p. 865-68, 2007. Disponível em: <https://www.scielosp.org/pdf/rsp/2007.v41n5/865-868>. Acesso em 05 novembro 2018.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas**. Universidade Federal de Goiás (UFC), p. 1-29, 2009. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 22 agosto 2018.

CARDOSO, Olinda Nogueira Paes; MACHADO, Rosa Teresa Moreira. **Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras**. Revista de administração pública, v. 42, n. 3, p. 495-528, 2008. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-76122008000300004&lng=en&nrm=iso>. Acesso em: 21 agosto 2018.

CASATI, Murilo Furtado Mendonça et al. **Epidemiologia do câncer de cabeça e pescoço no Brasil: estudo transversal de base populacional**. Rev Bras Cir Cabeça Pescoço, v. 41, n. 4, p. 186-91, 2012. Disponível em: <http://www.sbccp.org.br/wp-content/uploads/2014/11/REVISTA-SBCCP-41-4-artigo-07.pdf>. Acesso em: 04 novembro 2018.

DA COSTA CÔRTES, Sérgio; PORCARO, Rosa Maria; LIFSCHITZ, Sérgio. **Mineração de dados-funcionalidades, técnicas e abordagens**. PUC, 2002. Disponível em <ftp://ftp.inf.puc-rio.br/pub/docs/techreports/02_10_cortes.pdf>. Acesso em: 22 agosto 2018.

DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017.

DATASUS. **Saúde lança edital para informatizar 100% das unidades básicas do SUS**. 2017 Disponível em: <http://datasus.saude.gov.br/noticias/atualizacoes/1112-saude-lanca-edital-para-informatizar-100-das-unidades-basicas-do-sus>. Acesso em: 19 agosto 2018.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining: concepts and techniques**. Elsevier, 2011.

INCA. **ABC do câncer: abordagens básicas para o controle do câncer**. 2011. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/abc_do_cancer.pdf>. Acesso em: 16 agosto 2018.

INCA. **CÓDIGO DE CLÍNICAS - SisRHC**. Rio de Janeiro: INCA, [20–?]d. Disponível em: <https://irhc.inca.gov.br/RHCNet/downloadTabWin!document.action?doc=tabela.clinicas>. Acesso em: 28 outubro 2018.

INCA. **Dia Mundial sem Tabaco - 2019: Tabaco e saúde pulmonar**. 2019. Disponível em: <https://www.inca.gov.br/campanhas/dia-mundial-sem-tabaco/2019/tabaco-e-saude-pulmonar>. Acesso em: 13 junho 2019.

INCA. **Dicionário das variáveis da base de dados do SisRHC disponível para download no IRHC**. Rio de Janeiro: INCA, [20–?]c. Disponível em: <https://irhc.inca.gov.br/RHCNet/downloadTabWin!document.action?doc=dicionario.dados>. Acesso em: 27 outubro 2018.

INCA. **Estimativa 2018: incidência de câncer no Brasil. Instituto Nacional de câncer José Alencar Gomes da Silva. Coordenação de Prevenção e Vigilância**. Rio de Janeiro: INCA, 2017a. Disponível em:

<http://www1.inca.gov.br/estimativa/2018/estimativa-2018.pdf>. Acesso em: 17 agosto 2018.

INCA. **IntegradorRHC**. Rio de Janeiro: INCA, [20–?]a. Disponível em: <https://irhc.inca.gov.br/RHCNet/>. Acesso em: 25 outubro 2018.

INCA. **IntegradorRHC: Ferramenta para a Vigilância Hospitalar de Câncer no Brasil**. Rio de Janeiro: INCA, 2011. Disponível em: <http://www1.inca.gov.br/inca/Arquivos/folder_integradorrhc.pdf>. Acesso em: 29 outubro 2018.

INCA. **Mortes por câncer aumentaram 31% no Brasil em 15 anos, diz OMS**. 2017b. Disponível em: <http://www.inca.gov.br/wps/wcm/connect/agencianoticias/site/home/noticias/2017/mortes_por_cancer_aumentaram_31_no_brasil_em_15_anos_diz_oms>. Acesso em: 16 novembro 2018.

INCA. **Notas técnicas**. Rio de Janeiro, 2 [20–?]b Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/sobreinca/site/oinstituto> Acesso em: 25 outubro 2018.

INCA. **Sobre o instituto**. Rio de Janeiro, 2007 Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/sobreinca/site/oinstituto> Acesso em: 25 outubro 2018.

INCA. **Vigilância do Câncer e seus Fatores de Risco**. Instituto Nacional de câncer José Alencar Gomes da Silva. Coordenação de Prevenção e Vigilância. 2018a. Disponível em: <https://www.inca.gov.br/acesso-a-informacao/acoes-e-programas/vigilancia-do-cancer-e-seus-fatores-de-risco>. Acesso em: 25 outubro 2018.

KDNUGGETS. **What Analytics, Data mining, Big Data software you used in the past 12 months for a real project?**. 2012. Disponível em <https://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html>. Acesso em: 02 junho 2019.

KLIGERMAN, Jacob. **Registro Hospitalar de câncer no Brasil - Hospital Based Cancer Registries in Brazil**. Revista Brasileira de Cancerologia, v. 47, n. 4, 2001. Disponível em:

<https://rbc.inca.gov.br/site/arquivos/n_47/v04/pdf/editorial.pdf>. Acesso em: 27 outubro 2018.

LANDEIRO, Victor Lemes. **Introdução ao uso do programa R**. 2011. Disponível em: <https://cran.r-project.org/doc/contrib/Landeiro-Introducao.pdf>. Acesso em: 14 novembro 2018.

LAROSE, Daniel T.; LAROSE, Chantal D. **Discovering knowledge in data: an introduction to data mining**. John Wiley & Sons, 2014. Disponível em: <https://pdfs.semanticscholar.org/f415/6a05a47fdeda30638e10954d3674cc056ab6.pdf>. Acesso em: 23 agosto 2018.

LAROSE; D. T.; **Discovering knowledge in data An Introduction to Data Mining**. 2. Ed. Hoboken: John Wiley & Sons Inc., 2005. Disponível em: <https://pdfs.semanticscholar.org/f415/6a05a47fdeda30638e10954d3674cc056ab6.pdf>. Acesso em 22 Agosto de 2018.

MEDINGER; Dieison. **VISUALIZAÇÃO DE DADOS DE PACIENTES COM CÂNCER A PARTIR DA BASE DE DADOS DO INCA**. 2017. Disponível em: < https://tconline.feevale.br/NOVO/tc/files/0002_4368.pdf >. Acesso em: 11 junho 2019.

MINISTÉRIO DA SAÚDE (BR); INCA. **Informação dos registros hospitalares de câncer como estratégia de transformação: perfil do Instituto Nacional de câncer José Alencar Gomes da Silva em 25 anos**. 2012. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/inca/Informacao_dos_registros_hospitalares.pdf>. Acesso em: 30 outubro 2018.

OLIVEIRA, Andrea Santos de et al. **Registros Hospitalares de câncer em Pernambuco: da Gestão ao Registro**. Revista Brasileira de Cancerologia, v. 63, n. 1, p. 21-28, 2017. Disponível em: <http://www.inca.gov.br/Rbc/n_63/v01/pdf/05-artigo-registros-hospitalares-de-cancer-em-pernambuco-da-gestao-ao-registro.pdf>. Acesso em: 27 outubro 2018.

OMS. **Estimated number of deaths in 2018, all cancers, both sexes, all ages**. 2018. Disponível em: <https://gco.iarc.fr/today/online-analysis-table?v=2018&mode=population&mode_population=continents&population=900&populations=900&key=asr&sex=0&cancer=39&type=1&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&nb_items=5&gro

up_cancer=1&include_nmsc=1&include_nmsc_other=1>. 2018. Acesso em 22 outubro 2018.

OMS. **What is cancer?**. 2017a. Disponível em: <http://www.who.int/cancer/en/>. Acesso em: 23 outubro 2018.

ONU. **OMS: câncer mata 8,8 milhões de pessoas anualmente no mundo. Organizações das Nações Unidas no Brasil**, 2017. Disponível em: <https://nacoesunidas.org/oms-cancer-mata-88-milhoes-de-pessoas-anualmente-no-mundo>. Acesso em: 26 agosto 2018.

PETRUZALEK, Daniela. **READ.DBC - UM PACOTE PARA IMPORTAÇÃO DE DADOS DO DATASUS NA LINGUAGEM R**. XV Congresso Brasileiro de Informática em Saúde, 2016 Disponível em: <http://docs.bvsalud.org/biblioref/2018/07/906543/anais_cbis_2016_artigos_completos-601-606.pdf>. Acesso em: 23 outubro 2018.

PINTO, I. V. et al. **Completude e consistência dos dados dos registros hospitalares de câncer no Brasil.** Cad Saúde Coletiva, v. 20, n. 1, p. 113-20, 2012. Disponível em: <http://www.cadernos.iesc.ufrj.br/cadernos/images/csc/2012_1/artigos/CSC_v20n1_113-120.pdf>. Acesso em 20 Agosto 2018.

PORTAL BRASILEIRO DE DADOS ABERTOS. **Sobre o dados.gov.br**. [201-?]. Disponível em: <http://dados.gov.br/pagina/sobre>. Acesso em: 14 outubro 2018.

PRODANOV, Cleber Cristiano; DE FREITAS, Ernani Cesar. Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico-2ª Edição. Editora Feevale, 2013.

PUJARI, Arun K. **Data mining techniques**. Universities press, 2001.

REBELO, Marise Souto et al. **Aplicação de Indicadores Selecionados aos Dados dos Registros Hospitalares de câncer para Planejamento e Gestão Hospitalar**. 2008. Disponível em: <http://www1.inca.gov.br/vigilancia/docs/Epi%202008/Aplica%C3%A7%C3%A3o%20de%20Indicadores%20Selecionados%20aos%20Dados%20dos%20Registros%20Hospitalares%20de%20C%C3%A2ncer%20para%20Planejamento%20e%20Gest%C3%A3o%20Hospitalar.pdf>. Acesso em: 03 novembro 2018.

REXER ANALYTICS. **Data Science Survey**, 2015. Disponível em: <http://www.rexeranalytics.com/data-science-survey.html>. Acesso em: 02 junho 2019.

TEIXEIRA, Luiz Antonio. **De doença desconhecida a problema de saúde pública: o INCA e o controle do câncer no Brasil**. Rio de Janeiro: Ministério da Saúde, 2007. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/doenca_desconhecida_saude_publica.pdf>. Acesso em: 22 outubro 2018.

THAINES, Geovana Hagata de Lima Souza et al. **Produção, fluxo e análise de dados do sistema de informação em saúde: um caso exemplar**. Texto and Contexto Enfermagem, v. 18, n. 3, p. 466, 2009. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-07072009000300009&lng=en&nrm=iso>. Acesso em 24 agosto 2018.

THULER, Luiz Claudio Santos; BERGMANN, Anke; CASADO, Letícia. Base Secundaria. **Perfil das pacientes com câncer do colo do útero no Brasil, 2000-2009: estudo de base secundária**. Revista brasileira de cancerologia, v. 58, n. 3, p. 351-357, 2012. Disponível em: <http://www.inca.gov.br/Rbc/n_58/v03/pdf/04_artigo_perfil_pacientes_cancer_colo_utero_brasil_2000_2009_estudo_base_secundaria.pdf>. Acesso em: 05 novembro 2018.

WHO. **Estimated number of deaths in 2018, all cancers, both sexes, all ages**. 2018. Disponível em: <http://gco.iarc.fr/today/online-analysis-table?v=2018&mode=population&mode_population=continents&population=900&populations=900&key=asr&sex=0&cancer=39&type=1&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&nb_items=5&group_> Acesso em: 05 dezembro 2018.

WOLLMANN, Cássio. **ANÁLISE DA QUALIDADE DE DUAS FERRAMENTAS QUE MANIPULAM DADOS SOBRE O CÂNCER E A BASE DE DADOS DO INCA**. Disponível em: <https://tconline.feevale.br/NOVO/tc/files/0002_4534.docx>. Acesso em: 27 outubro 2018.