

BIG DATA NO MERCADO DE SEGUROS DE VEÍCULOS: ESTUDO DE CASO PARA A REGIÃO METROPOLITANA DO RIO DE JANEIRO

Orlando Fonseca Guilarte¹

Rodrigo dos Santos Targino²

Pedro Guilherme Costa Ferreira³

Resumo

Este trabalho tem o intuito de realizar uma análise estatística exploratória para o mercado de seguro de veículos do Rio de Janeiro dos anos de 2012 a 2016 utilizando as ferramentas que a linguagem R oferece para trabalhar com *Big Data*. Serão implementados algoritmos para trabalhar com o grande fluxo de informação proveniente das seguradoras. Foi verificado que esta linguagem possui pacotes muito úteis para o trabalho com Big Data que podem ser aplicados especificamente ao mercado de seguro.

Palavras-chave: mercado de seguro brasileiro, big data, análise exploratório, sparklyr

Abstract

This work aims to make an initial statistical exploratory analysis for the car insurance market of Rio de Janeiro from year 2012 to the year 2016 and to use the tools that R offers to work with Big Data. Algorithms will be implemented to work with a large flow of information coming from insurance companies. It has been found that R has very useful tools for working with Big Data which can be applied specifically to the insurance market.

Keywords: brazilian insurance market, big data, exploratory analysis, sparklyr

Introdução

As tecnologias de informação estão em constante evolução e a quantidade de dados gerados diariamente é cada vez maior. Processar e armazenar esse grande volume de informações é uma tarefa complexa que implica a aplicação de procedimentos de Big Data.

¹ Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). Fundação Getúlio Vargas (FGV), orlando.guilarte@fgv.br

² Fundação Getúlio Vargas (FGV), rodrigo.targino@fgv.br

³ Fundação Getúlio Vargas (FGV), pedro.guilherme@fgv.br

Big Data refere-se ao volume, à velocidade e à granularidade dos conjuntos de dados. São dados que trazem desafios em volume (tamanho), velocidade, variedade (formatos), veracidade (precisão), entre outras.

Por exemplo, no mercado de seguros de automóveis, se tem um grande volume de contratos por ano, para garantir à parte segurada cobertura contra danos causados por acidentes com veículos. No ato de contratação ou renovação do contrato junto à seguradora, cada segurado deve preencher um questionário, revelando sua idade, sexo, CEP da residência, etc. Com base nestas e outras informações, a seguradora calcula o risco associado ao contratante e, conseqüentemente, o valor do prêmio que lhe será cobrado (Richter,2016). Para analisar todos os contratos por anos das distintas seguradoras é preciso utilizar ferramentas elaboradas especialmente para o trabalho com grandes conjuntos de dados. Nesse sentido, a tecnologia desenvolvida na área da Big Data tornou-se o grande facilitador da mudança do setor de seguros (Boobier,2016).

Conforme estudo da consultoria PwC (PwC,2012), as seguradoras que explorarem técnicas de Big Data terão uma vantagem competitiva significativa em relação aos seus pares. Estas irão superar as barreiras do limite de custo e armazenamento, conseguindo uma interação muito mais dinâmica e acessível com os grandes dados, de acordo com as necessidades do negócio (PowerData,s.d.). Técnicas relacionadas à Big Data fornecem muitas possibilidades para os atuários, pois a ciência dos dados e a ciência atuarial se reforçam mutuamente. Mais dados permitem uma base mais rica para análise matemática atuarial e grandes conjuntos de dados levam a uma abordagem dinâmica de gerenciamento de riscos (Institute of Actuaries in Belgium,2015)

A linguagem R, software de excelência para estatísticos, tem entre seus pacotes o *sparklyr* (Luraschi,2018), especialmente desenvolvido como entorno para Big Data, e o pacote *dplyr* (Wickham,2017), muito útil para manipular os dados conforme em (Grolemund,2017). Estas serão as ferramentas para trabalhar com todos os dados disponibilizados pelas seguradoras.

Objetivo

Neste contexto, este artigo irá apresentar as ferramentas que o R oferece para trabalhar com Big Data e fazer uma análise estatística exploratória para o mercado de seguro de veículos do Rio de Janeiro dos anos de 2012 a 2016.

Material e Método

A Superintendência Brasileira de Seguros Privados (SUSEP), sendo o órgão responsável pelo controle e fiscalização dos mercados de seguro, coleta semestralmente dados estatísticos referentes a carteira de automóveis das companhias seguradoras que operam no Brasil. As seguradoras são obrigadas a enviar essas informações dentro dos prazos estipulados pela SUSEP (Richter,2016), as quais incluem dados referentes a apólices vigentes (bases R_AUTO) e sinistros ocorridos (bases S_AUTO), ou seja, são informações derivadas dos microdados, descritos no nível da apólice. Todos os dados se encontram em formato .csv. Para o presente trabalho o período de análise será de 2012 até 2016. Somando as bases tem-se um tamanho de 64.7 GB.

O processamento de Big Data precisa de ferramentas e técnicas próprias para tratar grandes volumes de dados. Considerando-se distribuir os dados e os cálculos, o Hadoop-Spark, como visto em (The Apache Software Foundation,2018), é uma das melhores alternativas, pois permite um processamento distribuído de grande volume de dados, como por exemplo, a série temporal das tabelas R_AUTO, que contém em média cerca de 6 GB de dados para cada semestre.

Para a implementação em linguagem R, o pacote *dplyr* mostra muitas vantagens para manipular os dados, sendo mais simples e flexível para essa tarefa. O *dplyr* pode ser utilizado para filtrar as linhas das tabelas, reordená-las, adicionar novas colunas para formação de categorias e reduzir cada grupo em uma única linha sob alguma medida estatística de interesse.

Para a interface entre R e Spark pode ser utilizado o pacote *sparklyr*, o que permite escrever código R com *dplyr* para ser executado num cluster. Desta forma, pode ser estabelecida uma conexão local para testar o funcionamento ou remota se já se tem clusters instalados e configurados. O RStudio oferece suporte integrado para trabalhar com *sparklyr* e com Spark DataFrame, que é o dado em Spark equivalente ao dataframe em base R.

Os passos do algoritmo criado são os seguintes⁴:

- 1) Fazer a conexão com Spark usando *sparklyr*;
- 2) Para cada tabela R_AUTO dos distintos anos:
 - 2.1- Copiar os dados para Spark;

⁴ O código detalhado dos algoritmos para a análise estatística em linguagem R encontram-se em anexo.

- 2.2- Fazer as operações requeridas com *dplyr* (select, filter, mutate, arrange, group by + summarise);
- 2.3- Coleta de dados do Spark para R e armazenar os resultados;
- 3) Visualização dos resultados;
- 4) Fechar a conexão com Spark;

Resultados e Discussões

Através do software R 3.4.1 (The R Foundation, 2017) e RStudio 1.0.143 (RStudio, 2018), foram feitos os distintos testes utilizando os pacotes *sparklyr* na versão 0.7.0 e *dplyr* na versão 0.7.4, num computador com microprocessador i5-3330 e 8GB de RAM. Todos os testes apresentados foram realizados via conexão local, o seja, com conexão ao cluster local. Mesmo sem acesso à um cluster remoto foi possível observar ganhos expressivos na velocidade das funções aplicadas em *data frames* muito pesados, em comparação à análise dos dados na memória RAM. Note que para a maioria dos anos foi impossível carregar o arquivo R_AUTO, dadas as limitantes de capacidade da RAM.

Para fazer a conexão com Spark (no passo 1 do algoritmo antes citado) se utilizam as seguintes linhas de código R⁵:

```
config <- spark_config()  
sc <- spark_connect(master = "local", config = config)
```

Para copiar os dados para Spark (passo 2.1) é utilizada a função *spark_read_csv*, pelo fato dos dados estarem já nesse formato:

```
nome_R <- spark_read_csv(sc, nome_R, path = paste0(pathR, ".csv"), delimiter =  
";", memory = FALSE)
```

Se considerou somente os dados das tabelas referentes a Região Metropolitana de Rio de Janeiro, caracterizados pelo valor "18" no campo REGIAO, as apólices com cobertura compreensiva, caracterizados pelo valor "1" no campo COBERTURA e as apólices criadas no período, o seja, sem contar suas modificações ou atualizações, caracterizadas pelo valor ""0000000000" no campo ENDOSSO. Esses filtros iniciais correspondem ao passo 2.2 do algoritmo, e serão atribuídos para a variável *data_filter* como apresentado código R dos Anexos.

⁵ Se preciso, uma configuração mais detalhada pode ser especificada

Para fazer a coleta de dados e salvar o resultado numa variável de nome *DadoCole* (passo 2.3) é chamada a função *collect()*:

```
DadoCole<- data_filter%>% collect()
```

A visualização dos dados pode ser feita utilizando *barplot* ou qualquer outra função para esse fim, e finalmente a conexão com Spark tem que ser finalizada com a função *spark_disconnect(sc)*, no passo 3 e 4 respectivamente.

Com relação à visualização dos resultados, os períodos de tempo de estudo seguem o formato ano(1) para o primeiro semestre do ano e ano(2) para o segundo. Por exemplo, o período 2015(2) corresponde ao segundo semestre do citado ano.

Para uma primeira análise exploratória dos dados pretende-se determinar no tempo se os homens fazem mais contratos com as seguradoras que as mulheres. A barra dos valores relativa aos homens (M) terá a cor azul e a relativa as mulheres (F) a cor roxa. Como mostrado na Figura 1, foi justificado para todo o período de estudo, chegando a ter uma diferença de apólices de 277.042 mil como aconteceu no 2015(1). É preciso salientar que o algoritmo demorou somente 31 minutos para executar o passo 2, que é o passo que mais tempo consome.

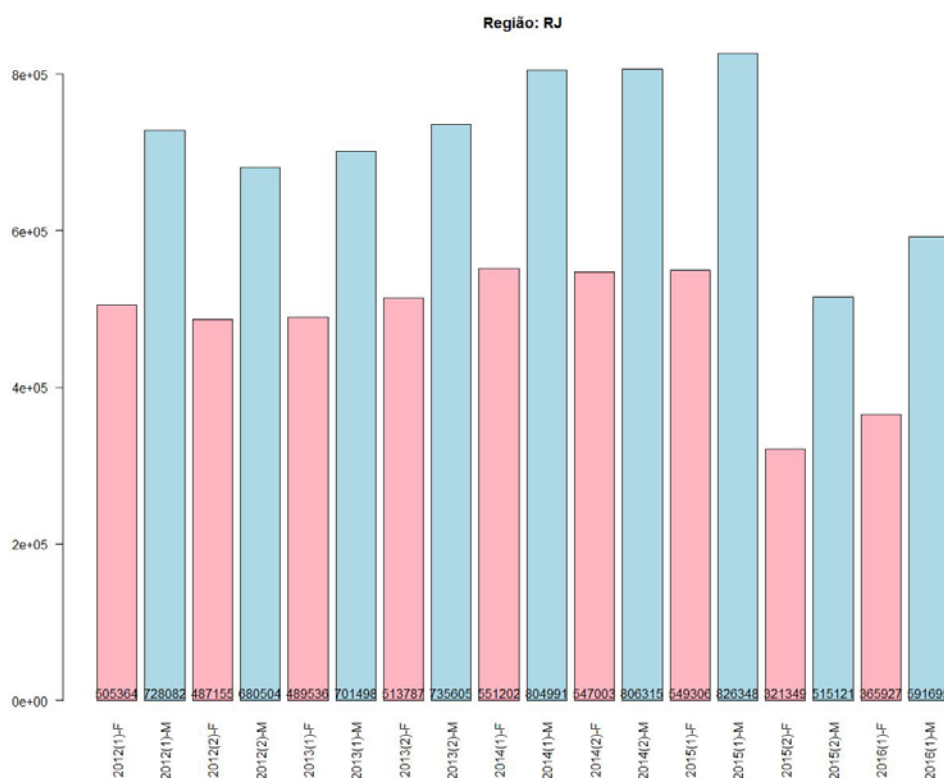


Figura 1 – Quantidade de apólices por sexo

Há evidência de que homens custam, em média, bem caro para as seguradoras. Para este trabalho, foi analisado somente indivíduos do sexo masculino, e pretende-se determinar quanto custam em média para as seguradoras arcar em caso de acidentes de colisão com indenização integral aos segurados, que pertencem ao grupo etário dos 30 a 40 anos de idade. A Figura 2 a seguir mostra, primeiramente, a quantidade das apólices do sexo masculino nessa faixa etária, onde é possível visualizar um comportamento crescente em geral para todos os anos, só tendo uma queda representativa do 2015(1) para 2015(2). A mesma foi elaborada após de aplicar os filtros. Por exemplo, pode-se observar que as quantidades de observações para o primeiro semestre do ano 2012 correspondia a 33879915 linhas, porém depois de aplicar os filtros foi reduzida a 140354 linhas.

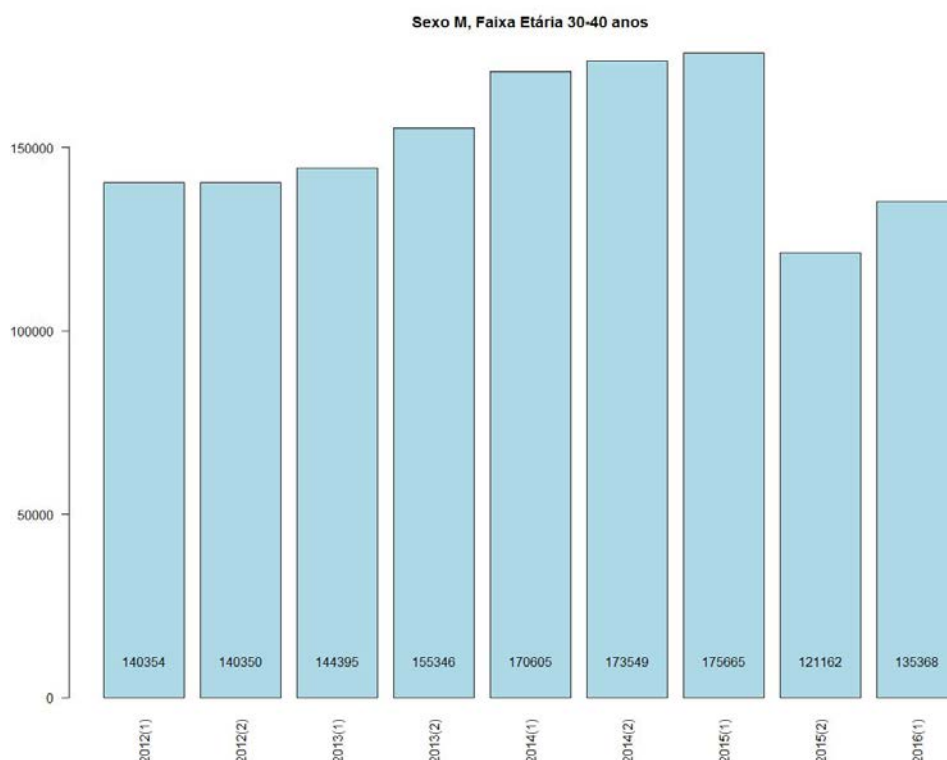


Figura 2 – Quantidade de apólices do sexo masculino e faixa etária 30-40 anos

Como observado na Figura 3, o valor médio das indenizações totais para homens entre 30 e 40 anos está situado a partir de R\$5000, com períodos em que os valores ultrapassem R\$6000.

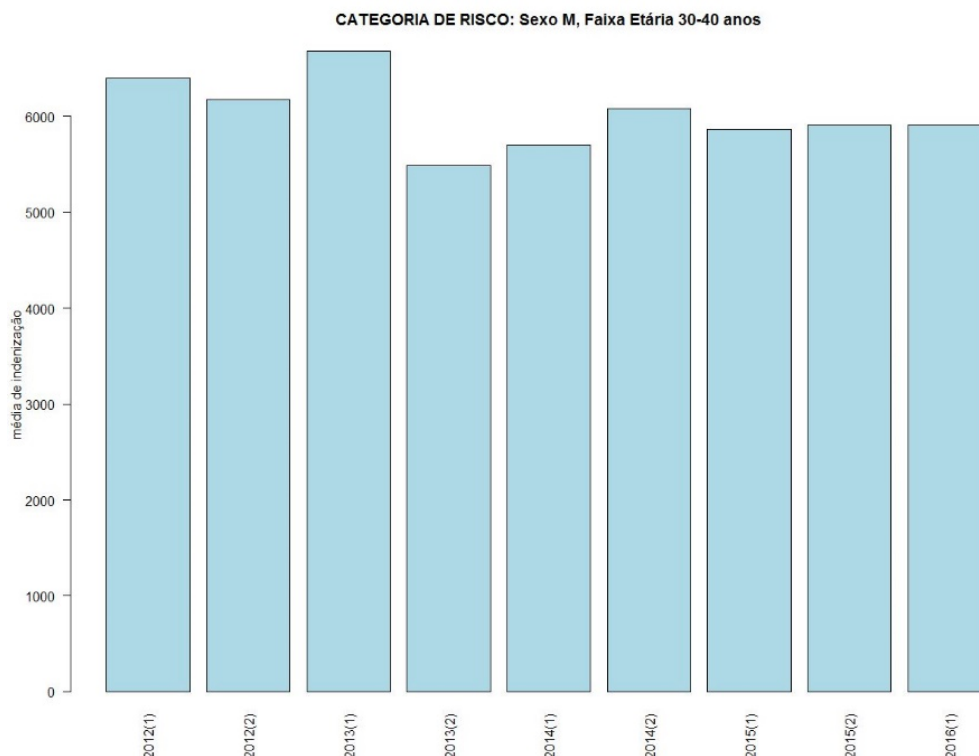


Figura 3 – Média das indenizações para apólices do sexo masculino e faixa etária 30-40 anos

Conclusão

Este trabalho teve como objetivo usar as ferramentas que o R oferece para trabalhar com Big Data, utilizando as bases R_AUTO e S_AUTO relativas ao período de 2012 a 2016. Os pacotes *dplyr* e *sparklyr* foram os que permitiram realizar a análise estatística exploratória para o mercado de seguro de veículos do Rio de Janeiro. Com o algoritmo proposto buscou-se identificar os principais passos para carregar grandes conjuntos de dados, transformá-los e visualizar os resultados. Como recomendação para trabalhos futuros podem-se utilizar outras variáveis como a importância segurada ou valor do prêmio para realizar outras análises, que podem incluir criação de cluster ou regressão, funções incluídas também no *sparklyr*.



Referências

- Boobier, T. **Analytics for Insurance: The Real Business of Big Data**. John Wiley & Sons, Ltd. 2016
- Grolemund, G.; Wickham, H. **R for Data Science: Import, Tidy, Transform, Visualize, and Model Data**. O'Reilly Media. 2017. Disponível em: <http://r4ds.had.co.nz/>
- Institute of Actuaries in Belgium. **BIG DATA: An actuarial perspective**. 2015. Disponível em: https://www.iabe.be/sites/default/files/bijlagen/big_data_paper_full_v009.pdf. Acesso em: 8 fev. 2018.
- Luraschi, J.; Kuo, K.; Ushey, K.; Allaire, JJ. **Package sparklyr**. R package version 0.7.0. 2018. Disponível em: <https://cran.r-project.org/web/packages/sparklyr/sparklyr.pdf>
- PowerData. **9 estratégias definitivas de Big Data para el sector Seguros**. Disponível em: <http://cdn2.hubspot.net/hub/239039/file-1958725915-pdf/docs>. Acesso em: 2 fev. 2018
- PwC. **Insurance 2020: Turning change into opportunity**. 2012. Disponível em: <https://www.pwc.com/gx/en/insurance/pdf/insurance-2020-turning-change-into-opportunity.pdf>. Acesso em: 2 fev. 2018
- Richter, G. **Uma análise da relação entre gênero, idade e sinistros no seguro de automóveis: o caso brasileiro**. FGV. 2016
- RStudio. **Take control of your R code**. 2018. Disponível em <https://www.rstudio.com/products/rstudio/>
- The Apache Software Foundation. **Apache Spark: Lightning-fast cluster computing**. 2018. Disponível em: <https://spark.apache.org/>
- The R Foundation. **The R Project for Statistical Computing**. 2017. Disponível em: <https://www.r-project.org/>
- Wickham, H.; Francois, R.; Henry, L; Müller, K. **dplyr: A Grammar of Data Manipulation**. R package version 0.7.4. 2017. Disponível em: <https://cran.r-project.org/web/packages/dplyr/index.html>

Anexo

```
###Código R para gerar os resultados da Figura 1:
library(sparklyr)
library(dplyr)
library(stringr)
library(lubridate)

#PASSO 1
spark_install(version = "2.2.0", hadoop_version = "2.7")
config <- spark_config()
sc <- spark_connect(master = "local", config = config)

PATH<- "BD/"
QuantSexoXAnos<- data.frame(Anos=character(0),SEXO=character(0), total=
numeric(0))
x <-
c("2012A","2012B","2013A","2013B","2014A","2014B","2015A","2015B","2016A")

#PASSO 2
for (i in 1:length(x))
{
  var <- x[i]
  P <- paste0(PATH,var)
  P <- paste0(P,"/")
  nome_R <- paste0("R_AUTO_", var)
  pathR<- paste0(P,nome_R)
```



```
#PASSO 2.1
nome_R<-spark_read_csv(sc,nome_R,path= paste0(pathR,".csv"),delimiter =
";",memory=FALSE)

#PASSO 2.2
data_filter<- nome_R %>% filter(REGIAO == "18",COBERTURA == "1",ENDOSSO ==
"0000000000", SEXO %in% c("F","M")) %>% select(cod_apo,SEXO) %>%
group_by(cod_apo, SEXO) %>% distinct(cod_apo, SEXO) %>% group_by(SEXO)%>%
summarise(total = n())

#PASSO 2.3
DadoCole<- data_filter%>% collect()
newvar<- str_replace(var,"A","(1)")
newvar<- str_replace(newvar,"B","(2)")
Dado <- DadoCole %>% mutate(Anos = paste(newvar,SEXO,sep="-"))%>%
arrange(SEXO)
QuantSexoXAnos<- rbind(QuantSexoXAnos, Dado)
nome_R <- paste0("R_AUTO_", var)
dplyr::db_drop_table(sc,nome_R)
}

#PASSO 3
cols <- c("lightblue", "lightpink")[(grepl("F", QuantSexoXAnos$Anos))+1]
bp<-barplot(QuantSexoXAnos$total,col=cols,names.arg =
QuantSexoXAnos$Anos,xlab="",ylab="",main="Região: RJ",las=2)
text(bp, 10000, labels =QuantSexoXAnos$total)

#PASSO 4
spark_disconnect(sc)
```

Código R (PASSOS 2.2 e 2.3) para gerar os resultados da Figura 3:

```
data_LJ <- left_join(nome_R,nome_S,by=
c("cod_apo","ITEM","ENDOSSO","REGIAO"))
data_selec <- data_LJ %>%
select(REGIAO,SEXO=SEXO.x,DATA_NASC,CAUSA,INICIO_VIG,INDENIZ,COBERTURA=COBE
RTURA.x,D_AVI)

data_filter<- data_selec %>% filter(REGIAO == "18",
SEXO=="M",CAUSA=="4",COBERTURA == "1") %>%
select(DATA_NASC,INICIO_VIG,INDENIZ,D_AVI)

DadoCole<- data_filter%>% collect()
DadoCole$INICIO_VIG<- ymd(as.vector(DadoCole$INICIO_VIG))
DadoCole$DATA_NASC<- ymd(as.vector(DadoCole$DATA_NASC))
DadoCole$INDENIZ <- as.numeric(DadoCole$INDENIZ)
DadoCole$IDADE<- (DadoCole$INICIO_VIG-DadoCole$DATA_NASC)/365.25
DadoCole<- DadoCole%>%filter(IDADE >= 18)
DadoCole$FAIXAETARIA = rep(0,dim(DadoCole)[1])
DadoCole$FAIXAETARIA[DadoCole$IDADE>=25&DadoCole$IDADE<30]=1
DadoCole$FAIXAETARIA[DadoCole$IDADE>=30&DadoCole$IDADE<40]=2
DadoCole$FAIXAETARIA[DadoCole$IDADE>=40&DadoCole$IDADE<50]=3
DadoCole$FAIXAETARIA[DadoCole$IDADE>=50&DadoCole$IDADE<60]=4
DadoCole$FAIXAETARIA[DadoCole$IDADE>=60]=5
DadoCole$IDADE<-as.numeric(DadoCole$IDADE)

Dado<- DadoCole%>%filter(FAIXAETARIA == 2, INDENIZ >0)
```

```
Dado <- Dado %>% mutate(D_AVI = ymd(as.vector(D_AVI)) )
Dado2<- Dado %>% mutate(Mes = month(D_AVI),Ano =
paste0("y",year(D_AVI)),In =0 ) %>% select(INDENIZ,Mes,Ano,In)

for(j in 1:nrow(Dado2))
{
  Mes <- as.integer(Dado2[j,"Mes"])
  Ano <- as.character(Dado2[j,"Ano"])
  Y<- paste0("y",str_sub(var, 1, 4))
  if(is.na(Mes)) {
    Dado2[j,"In"] = (Dado2[j,"INDENIZ"]/IPCA[12,Y]) *IPCA[12,"y2016"]
  }
  else {
    Dado2[j,"In"] = (Dado2[j,"INDENIZ"]/IPCA[Mes,Ano])
*IPCA[12,"y2016"]
  }
}
MediaInden<- Dado2 %>% summarise(m<- mean(In))
MedInd[i]<- as.numeric(MediaInden)
```

Código R (PASSOS 2.2 e 2.3) para gerar os resultados da Figura 2:

```
data_filter<- nome_R %>% filter(REGIAO == "18",COBERTURA == "1",ENDOSSO ==
"0000000000", SEXO == "M") %>% select(cod_apo,DATA_NASC,INICIO_VIG)

DadoCole<- data_filter%>% collect()
DadoCole$INICIO_VIG<- ymd(as.vector(DadoCole$INICIO_VIG))
DadoCole$DATA_NASC<- ymd(as.vector(DadoCole$DATA_NASC))

Dado<- DadoCole %>% filter(FAIXAETARIA == 2) %>% group_by(FAIXAETARIA)
%>% summarise(total = n()) %>% select(total)
QuantSexoXAnos[i] <- Dado$total
```