Project

2022-04-04

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

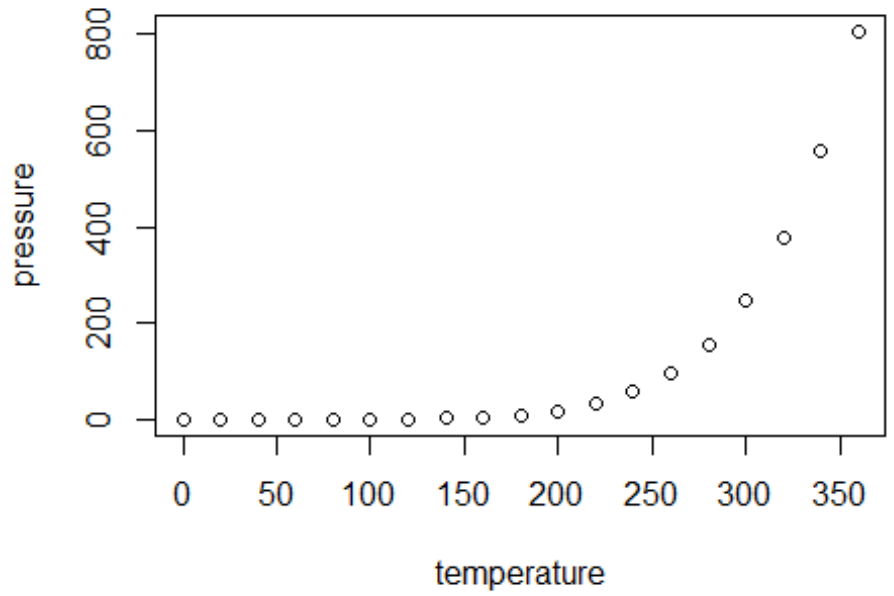
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
setwd("~/CIND820")
library(readr)
Dataset <- read_csv("C:/Users/annac/OneDrive/Desktop/Data Analytics/CIND820/D
ataset.csv")

## Rows: 46464 Columns: 20
## -- Column specification --------------------------------------------------
------
## Delimiter: ","
## chr (13): GEO, DGUID, Sex, Age at admission, Immigrant admission category,
Y...
## dbl  (5): REF_DATE, UOM_ID, SCALAR_ID, VALUE, DECIMALS
## lgl  (2): SYMBOL, TERMINATED
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this m
essage.

str(Dataset)

## spec_tbl_df [46,464 x 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ REF_DATE                    : num [1:46464] 2006 2007 2008 2009 2010 ..
.
##  $ GEO                         : chr [1:46464] "Canada" "Canada" "Canada"
```

```
"Canada" ...
##  $ DGUID                       : chr [1:46464] "2016A000011124" "2016A0000
11124" "2016A000011124" "2016A000011124" ...
##  $ Sex                         : chr [1:46464] "Both sexes" "Both sexes" "
Both sexes" "Both sexes" ...
##  $ Age at admission            : chr [1:46464] "Total, Age at admission" "
Total, Age at admission" "Total, Age at admission" "Total, Age at admission"
...
##  $ Immigrant admission category: chr [1:46464] "Total, immigrant admission
category" "Total, immigrant admission category" "Total, immigrant admission c
ategory" "Total, immigrant admission category" ...
##  $ Years since admission       : chr [1:46464] "0 years since admission" "
0 years since admission" "0 years since admission" "0 years since admission"
...
##  $ Income type                 : chr [1:46464] "Wages, salaries and commis
sions" "Wages, salaries and commissions" "Wages, salaries and commissions" "W
ages, salaries and commissions" ...
##  $ Statistics                  : chr [1:46464] "Total count" "Total count"
"Total count" "Total count" ...
##  $ UOM                         : chr [1:46464] "Persons" "Persons" "Person
s" "Persons" ...
##  $ UOM_ID                      : num [1:46464] 249 249 249 249 249 249 249
249 249 249 ...
##  $ SCALAR_FACTOR               : chr [1:46464] "units" "units" "units" "un
its" ...
##  $ SCALAR_ID                   : num [1:46464] 0 0 0 0 0 0 0 0 0 0 ...
##  $ VECTOR                      : chr [1:46464] "v1028809067" "v1028809067"
"v1028809067" "v1028809067" ...
##  $ COORDINATE                  : chr [1:46464] "1.1.1.1.1.1.1" "1.1.1.1.1.
1.1" "1.1.1.1.1.1.1" "1.1.1.1.1.1.1" ...
##  $ VALUE                       : num [1:46464] 154640 145895 151290 155340
169745 ...
##  $ STATUS                      : chr [1:46464] NA NA NA NA ...
##  $ SYMBOL                      : logi [1:46464] NA NA NA NA NA NA ...
##  $ TERMINATED                  : logi [1:46464] NA NA NA NA NA NA ...
##  $ DECIMALS                    : num [1:46464] 0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   REF_DATE = col_double(),
##   ..   GEO = col_character(),
##   ..   DGUID = col_character(),
##   ..   Sex = col_character(),
##   ..   `Age at admission` = col_character(),
##   ..   `Immigrant admission category` = col_character(),
##   ..   `Years since admission` = col_character(),
##   ..   `Income type` = col_character(),
##   ..   Statistics = col_character(),
##   ..   UOM = col_character(),
##   ..   UOM_ID = col_double(),
##   ..   SCALAR_FACTOR = col_character(),
```

```
##   ..   SCALAR_ID = col_double(),
##   ..   VECTOR = col_character(),
##   ..   COORDINATE = col_character(),
##   ..   VALUE = col_double(),
##   ..   STATUS = col_character(),
##   ..   SYMBOL = col_logical(),
##   ..   TERMINATED = col_logical(),
##   ..   DECIMALS = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
head(Dataset)
```

```
## # A tibble: 6 x 20
##   REF_DATE GEO    DGUID Sex   `Age at admiss~` `Immigrant adm~` `Years sin
ce a~`
##      <dbl> <chr>  <chr> <chr> <chr>            <chr>            <chr>
## 1     2006 Canada 2016~ Both~ Total, Age at a~ Total, immigran~ 0 years si
nce a~
## 2     2007 Canada 2016~ Both~ Total, Age at a~ Total, immigran~ 0 years si
nce a~
## 3     2008 Canada 2016~ Both~ Total, Age at a~ Total, immigran~ 0 years si
nce a~
## 4     2009 Canada 2016~ Both~ Total, Age at a~ Total, immigran~ 0 years si
nce a~
## 5     2010 Canada 2016~ Both~ Total, Age at a~ Total, immigran~ 0 years si
nce a~
## 6     2011 Canada 2016~ Both~ Total, Age at a~ Total, immigran~ 0 years si
nce a~
## # ... with 13 more variables: `Income type` <chr>, Statistics <chr>, UOM <
chr>,
## #   UOM_ID <dbl>, SCALAR_FACTOR <chr>, SCALAR_ID <dbl>, VECTOR <chr>,
## #   COORDINATE <chr>, VALUE <dbl>, STATUS <chr>, SYMBOL <lgl>,
## #   TERMINATED <lgl>, DECIMALS <dbl>
```

```
summary(Dataset)
```

```
##     REF_DATE        GEO               DGUID               Sex
##  Min.   :2006   Length:46464       Length:46464       Length:46464
##  1st Qu.:2008   Class :character   Class :character   Class :character
##  Median :2011   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2011
##  3rd Qu.:2014
##  Max.   :2016
##
##  Age at admission   Immigrant admission category Years since admission
##  Length:46464       Length:46464                 Length:46464
##  Class :character   Class :character             Class :character
##  Mode  :character   Mode  :character             Mode  :character
##
##
```

```
##
##
##  Income type         Statistics            UOM                UOM_ID
##  Length:46464       Length:46464       Length:46464       Min.   : 81
##  Class :character   Class :character   Class :character   1st Qu.: 81
##  Mode  :character   Mode  :character   Mode  :character   Median :165
##                                                           Mean   :165
##                                                           3rd Qu.:249
##                                                           Max.   :249
##
##  SCALAR_FACTOR        SCALAR_ID    VECTOR           COORDINATE
##  Length:46464       Min.   :0   Length:46464       Length:46464
##  Class :character   1st Qu.:0   Class :character   Class :character
##  Mode  :character   Median :0   Mode  :character   Mode  :character
##                     Mean   :0
##                     3rd Qu.:0
##                     Max.   :0
##
##      VALUE           STATUS           SYMBOL        TERMINATED        DECI
MALS
##  Min.   :     0   Length:46464       Mode:logical   Mode:logical   Min.
:0
##  1st Qu.:  4379   Class :character   NA's:46464     NA's:46464     1st Qu.
:0
##  Median :  9000   Mode  :character                                 Median
:0
##  Mean   : 19474                                                    Mean
:0
##  3rd Qu.: 20565                                                    3rd Qu.
:0
##  Max.   :198810                                                    Max.
:0
##  NA's   :27072
```

```
Dataset$DGUID <- NULL
Dataset$DECIMALS <- NULL
Dataset$TERMINATED <- NULL
Dataset$SYMBOL <- NULL
Dataset$STATUS <- NULL
Dataset$UOM_ID <- NULL
Dataset$SCALAR_ID <- NULL
Dataset$SCALAR_FACTOR <- NULL
Dataset$UOM_ID <- NULL

head(Dataset)
```

```
## # A tibble: 6 x 12
##   REF_DATE GEO    Sex        `Age at admissi~` `Immigrant adm~` `Years sin
ce a~`
##      <dbl> <chr>  <chr>      <chr>             <chr>            <chr>
```

```
## 1     2006 Canada Both sexes Total, Age at ad~ Total, immigran~ 0 years si
nce a~
## 2     2007 Canada Both sexes Total, Age at ad~ Total, immigran~ 0 years si
nce a~
## 3     2008 Canada Both sexes Total, Age at ad~ Total, immigran~ 0 years si
nce a~
## 4     2009 Canada Both sexes Total, Age at ad~ Total, immigran~ 0 years si
nce a~
## 5     2010 Canada Both sexes Total, Age at ad~ Total, immigran~ 0 years si
nce a~
## 6     2011 Canada Both sexes Total, Age at ad~ Total, immigran~ 0 years si
nce a~
## # ... with 6 more variables: `Income type` <chr>, Statistics <chr>, UOM <c
hr>,
## #   VECTOR <chr>, COORDINATE <chr>, VALUE <dbl>
```

```
Dataset$VECTOR <- NULL
Dataset$COORDINATE <- NULL
```

```
head(Dataset)
```

```
## # A tibble: 6 x 10
##   REF_DATE GEO    Sex        `Age at admissi~` `Immigrant adm~` `Years sin
ce a~`
##      <dbl> <chr>  <chr>      <chr>             <chr>            <chr>
## 1     2006 Canada Both sexes Total, Age at ad~ Total, immigran~ 0 years si
nce a~
## 2     2007 Canada Both sexes Total, Age at ad~ Total, immigran~ 0 years si
nce a~
## 3     2008 Canada Both sexes Total, Age at ad~ Total, immigran~ 0 years si
nce a~
## 4     2009 Canada Both sexes Total, Age at ad~ Total, immigran~ 0 years si
nce a~
## 5     2010 Canada Both sexes Total, Age at ad~ Total, immigran~ 0 years si
nce a~
## 6     2011 Canada Both sexes Total, Age at ad~ Total, immigran~ 0 years si
nce a~
## # ... with 4 more variables: `Income type` <chr>, Statistics <chr>, UOM <c
hr>,
## #   VALUE <dbl>
```

```
Dataset$Sex <- as.factor(Dataset$Sex)
str(Dataset$Sex)
```

```
##  Factor w/ 4 levels "Both sexes","Females",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
as.numeric(Dataset$Sex)
```

```
Dataset$`Age at admission` <- as.factor(Dataset$`Age at admission`)
Dataset$`Years since admission` <- as.factor(Dataset$`Years since admission`)
Dataset$`Income type` <- as.factor(Dataset$`Income type`)
```

```r
Dataset$Statistics <- as.factor(Dataset$Statistics)
Dataset$UOM <- as.factor(Dataset$UOM)

str(Dataset$`Age at admission`)

##  Factor w/ 4 levels "20 to 24 years",..: 4 4 4 4 4 4 4 4 4 4 ...

str(Dataset$`Years since admission`)

##  Factor w/ 11 levels "0 years since admission",..: 1 1 1 1 1 1 1 1 1 1 ...

str(Dataset$`Income type`)

##  Factor w/ 6 levels "All income","Employment insurance benefits",..: 6 6 6
6 6 6 6 6 6 6 ...

str(Dataset$Statistics)

##  Factor w/ 4 levels "Mean with income",..: 3 3 3 3 3 3 3 3 3 3 ...

str(Dataset$UOM)

##  Factor w/ 2 levels "Dollars","Persons": 2 2 2 2 2 2 2 2 2 2 ...

as.numeric(Dataset$`Age at admission`)

as.numeric(Dataset$`Years since admission`)

as.numeric(Dataset$`Income type`)

as.numeric(Dataset$Statistics)

str(Dataset)

## spec_tbl_df [46,464 x 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ REF_DATE                    : num [1:46464] 2006 2007 2008 2009 2010 ..
.
##  $ GEO                         : chr [1:46464] "Canada" "Canada" "Canada"
"Canada" ...
##  $ Sex                         : Factor w/ 4 levels "Both sexes","Females"
,..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Age at admission            : Factor w/ 4 levels "20 to 24 years",..: 4
4 4 4 4 4 4 4 4 4 ...
##  $ Immigrant admission category: chr [1:46464] "Total, immigrant admission
category" "Total, immigrant admission category" "Total, immigrant admission c
ategory" "Total, immigrant admission category" ...
##  $ Years since admission       : Factor w/ 11 levels "0 years since admiss
ion",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Income type                 : Factor w/ 6 levels "All income","Employme
nt insurance benefits",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ Statistics                  : Factor w/ 4 levels "Mean with income",..:
3 3 3 3 3 3 3 3 3 3 ...
##  $ UOM                         : Factor w/ 2 levels "Dollars","Persons": 2
```

```
2 2 2 2 2 2 2 2 2 ...
##  $ VALUE                       : num [1:46464] 154640 145895 151290 155340
169745 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   REF_DATE = col_double(),
##   ..   GEO = col_character(),
##   ..   DGUID = col_character(),
##   ..   Sex = col_character(),
##   ..   `Age at admission` = col_character(),
##   ..   `Immigrant admission category` = col_character(),
##   ..   `Years since admission` = col_character(),
##   ..   `Income type` = col_character(),
##   ..   Statistics = col_character(),
##   ..   UOM = col_character(),
##   ..   UOM_ID = col_double(),
##   ..   SCALAR_FACTOR = col_character(),
##   ..   SCALAR_ID = col_double(),
##   ..   VECTOR = col_character(),
##   ..   COORDINATE = col_character(),
##   ..   VALUE = col_double(),
##   ..   STATUS = col_character(),
##   ..   SYMBOL = col_logical(),
##   ..   TERMINATED = col_logical(),
##   ..   DECIMALS = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>

library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

rename(Dataset, c("Year" = "REF_DATE"))

## # A tibble: 46,464 x 10
##     Year GEO    Sex        `Age at admission`  `Immigrant adm~` `Years sin
ce a~`
##    <dbl> <chr>  <fct>      <fct>               <chr>            <fct>
##  1  2006 Canada Both sexes Total, Age at admi~ Total, immigran~ 0 years si
nce a~
##  2  2007 Canada Both sexes Total, Age at admi~ Total, immigran~ 0 years si
nce a~
```

```
##  3  2008 Canada Both sexes Total, Age at admi~ Total, immigran~ 0 years si
nce a~
##  4  2009 Canada Both sexes Total, Age at admi~ Total, immigran~ 0 years si
nce a~
##  5  2010 Canada Both sexes Total, Age at admi~ Total, immigran~ 0 years si
nce a~
##  6  2011 Canada Both sexes Total, Age at admi~ Total, immigran~ 0 years si
nce a~
##  7  2012 Canada Both sexes Total, Age at admi~ Total, immigran~ 0 years si
nce a~
##  8  2013 Canada Both sexes Total, Age at admi~ Total, immigran~ 0 years si
nce a~
##  9  2014 Canada Both sexes Total, Age at admi~ Total, immigran~ 0 years si
nce a~
## 10  2015 Canada Both sexes Total, Age at admi~ Total, immigran~ 0 years si
nce a~
## # ... with 46,454 more rows, and 4 more variables: `Income type` <fct>,
## #   Statistics <fct>, UOM <fct>, VALUE <dbl>

rename(Dataset, c("Age at Admission" = "Age at admission"))

## # A tibble: 46,464 x 10
##    REF_DATE GEO    Sex        `Age at Admiss~` `Immigrant adm~` `Years sin
ce a~`
##       <dbl> <chr>  <fct>      <fct>            <chr>            <fct>
##  1     2006 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  2     2007 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  3     2008 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  4     2009 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  5     2010 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  6     2011 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  7     2012 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  8     2013 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  9     2014 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
## 10     2015 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
## # ... with 46,454 more rows, and 4 more variables: `Income type` <fct>,
## #   Statistics <fct>, UOM <fct>, VALUE <dbl>

rename(Dataset, c("Years Since Admission" = "Years since admission"))
```

```
## # A tibble: 46,464 x 10
##    REF_DATE GEO    Sex        `Age at admiss~` `Immigrant adm~` `Years Sin
ce A~`
##       <dbl> <chr>  <fct>      <fct>            <chr>            <fct>
##  1     2006 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  2     2007 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  3     2008 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  4     2009 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  5     2010 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  6     2011 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  7     2012 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  8     2013 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  9     2014 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
## 10     2015 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
## # ... with 46,454 more rows, and 4 more variables: `Income type` <fct>,
## #   Statistics <fct>, UOM <fct>, VALUE <dbl>

rename(Dataset, c("Income Type" = "Income type"))

## # A tibble: 46,464 x 10
##    REF_DATE GEO    Sex        `Age at admiss~` `Immigrant adm~` `Years sin
ce a~`
##       <dbl> <chr>  <fct>      <fct>            <chr>            <fct>
##  1     2006 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  2     2007 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  3     2008 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  4     2009 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  5     2010 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  6     2011 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  7     2012 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  8     2013 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  9     2014 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
```

```
nce a~
## 10     2015 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
## # ... with 46,454 more rows, and 4 more variables: `Income Type` <fct>,
## #   Statistics <fct>, UOM <fct>, VALUE <dbl>

rename(Dataset, c("Unit of Analysis" = "UOM"))

## # A tibble: 46,464 x 10
##    REF_DATE GEO    Sex        `Age at admiss~` `Immigrant adm~` `Years sin
ce a~`
##       <dbl> <chr>  <fct>      <fct>            <chr>            <fct>
##  1     2006 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  2     2007 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  3     2008 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  4     2009 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  5     2010 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  6     2011 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  7     2012 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  8     2013 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  9     2014 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
## 10     2015 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
## # ... with 46,454 more rows, and 4 more variables: `Income type` <fct>,
## #   Statistics <fct>, `Unit of Analysis` <fct>, VALUE <dbl>

rename(Dataset, c("Value" = "VALUE"))

## # A tibble: 46,464 x 10
##    REF_DATE GEO    Sex        `Age at admiss~` `Immigrant adm~` `Years sin
ce a~`
##       <dbl> <chr>  <fct>      <fct>            <chr>            <fct>
##  1     2006 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  2     2007 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  3     2008 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  4     2009 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  5     2010 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
```

```
##  6     2011 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  7     2012 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  8     2013 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
##  9     2014 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
## 10     2015 Canada Both sexes Total, Age at a~ Total, immigran~ 0 years si
nce a~
## # ... with 46,454 more rows, and 4 more variables: `Income type` <fct>,
## #   Statistics <fct>, UOM <fct>, Value <dbl>
```

```
Dataset$`Immigrant admission category` <- NULL
Dataset$GEO <- NULL
```

```
library(dplyr)
rename(Dataset, c( "Year" = "REF_DATE"))
```

```
## # A tibble: 46,464 x 8
##     Year Sex    `Age at admiss~` `Years since a~` `Income type` Statistics
UOM
##    <dbl> <fct>  <fct>            <fct>            <fct>         <fct>
<fct>
##  1  2006 Both ~ Total, Age at a~ 0 years since a~ Wages, salar~ Total cou~
Pers~
##  2  2007 Both ~ Total, Age at a~ 0 years since a~ Wages, salar~ Total cou~
Pers~
##  3  2008 Both ~ Total, Age at a~ 0 years since a~ Wages, salar~ Total cou~
Pers~
##  4  2009 Both ~ Total, Age at a~ 0 years since a~ Wages, salar~ Total cou~
Pers~
##  5  2010 Both ~ Total, Age at a~ 0 years since a~ Wages, salar~ Total cou~
Pers~
##  6  2011 Both ~ Total, Age at a~ 0 years since a~ Wages, salar~ Total cou~
Pers~
##  7  2012 Both ~ Total, Age at a~ 0 years since a~ Wages, salar~ Total cou~
Pers~
##  8  2013 Both ~ Total, Age at a~ 0 years since a~ Wages, salar~ Total cou~
Pers~
##  9  2014 Both ~ Total, Age at a~ 0 years since a~ Wages, salar~ Total cou~
Pers~
## 10  2015 Both ~ Total, Age at a~ 0 years since a~ Wages, salar~ Total cou~
Pers~
## # ... with 46,454 more rows, and 1 more variable: VALUE <dbl>
```

```
rename(Dataset, c("Age at Admission" = "Age at admission"))
```

```
## # A tibble: 46,464 x 8
##    REF_DATE Sex       `Age at Admiss~` `Years since a~` `Income type` Stat
istics
```

```
##       <dbl> <fct>     <fct>            <fct>            <fct>         <fct
>
##  1     2006 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  2     2007 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  3     2008 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  4     2009 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  5     2010 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  6     2011 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  7     2012 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  8     2013 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  9     2014 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## 10     2015 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## # ... with 46,454 more rows, and 2 more variables: UOM <fct>, VALUE <dbl>

rename(Dataset, c("Years Since Admission" = "Years since admission"))

## # A tibble: 46,464 x 8
##    REF_DATE Sex       `Age at admiss~` `Years Since A~` `Income type` Stat
istics
##       <dbl> <fct>     <fct>            <fct>            <fct>         <fct
>
##  1     2006 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  2     2007 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  3     2008 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  4     2009 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  5     2010 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  6     2011 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  7     2012 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  8     2013 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  9     2014 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## 10     2015 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
```

```
l cou~
## # ... with 46,454 more rows, and 2 more variables: UOM <fct>, VALUE <dbl>

rename(Dataset, c("Income Type" = "Income type"))

## # A tibble: 46,464 x 8
##    REF_DATE Sex       `Age at admiss~` `Years since a~` `Income Type` Stat
istics
##       <dbl> <fct>     <fct>            <fct>            <fct>         <fct
>
##  1     2006 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  2     2007 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  3     2008 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  4     2009 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  5     2010 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  6     2011 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  7     2012 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  8     2013 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  9     2014 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## 10     2015 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## # ... with 46,454 more rows, and 2 more variables: UOM <fct>, VALUE <dbl>

rename(Dataset, c("Unit of Analysis" = "UOM"))

## # A tibble: 46,464 x 8
##    REF_DATE Sex       `Age at admiss~` `Years since a~` `Income type` Stat
istics
##       <dbl> <fct>     <fct>            <fct>            <fct>         <fct
>
##  1     2006 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  2     2007 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  3     2008 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  4     2009 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  5     2010 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  6     2011 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
```

```
##  7     2012 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  8     2013 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  9     2014 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## 10     2015 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## # ... with 46,454 more rows, and 2 more variables: `Unit of Analysis` <fct
>,
## #   VALUE <dbl>
```

```
rename(Dataset, c("Value" = "VALUE"))
```

```
## # A tibble: 46,464 x 8
##    REF_DATE Sex       `Age at admiss~` `Years since a~` `Income type` Stat
istics
##       <dbl> <fct>     <fct>            <fct>            <fct>         <fct
>
##  1     2006 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  2     2007 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  3     2008 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  4     2009 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  5     2010 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  6     2011 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  7     2012 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  8     2013 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  9     2014 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## 10     2015 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## # ... with 46,454 more rows, and 2 more variables: UOM <fct>, Value <dbl>
```

```
head(Dataset)
```

```
## # A tibble: 6 x 8
##   REF_DATE Sex        `Age at admiss~` `Years since a~` `Income type` Stat
istics
##      <dbl> <fct>      <fct>            <fct>            <fct>         <fct
>
## 1     2006 Both sexes Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## 2     2007 Both sexes Total, Age at a~ 0 years since a~ Wages, salar~ Tota
```

```
l cou~
## 3     2008 Both sexes Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## 4     2009 Both sexes Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## 5     2010 Both sexes Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## 6     2011 Both sexes Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## # ... with 2 more variables: UOM <fct>, VALUE <dbl>
```

```
Mean <- mean(Dataset$VALUE, na.rm = TRUE)
Dataset$VALUE[is.na(Dataset$VALUE)]= Mean
summary(Dataset$VALUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   11400   19474   19474   19474  198810
```

```
summary(Dataset)
```

```
##     REF_DATE                Sex                        Age at admission
##  Min.   :2006   Both sexes    :11616   20 to 24 years         :11616
##  1st Qu.:2008   Females       :11616   35 to 44 years         :11616
##  Median :2011   Males         :11616   55 to 64 years         :11616
##  Mean   :2011   Sex not stated:11616   Total, Age at admission:11616
##  3rd Qu.:2014
##  Max.   :2016
##
##               Years since admission                          Income type
##  0 years since admission : 4224     All income                     :7744
##  1 years since admission : 4224     Employment insurance benefits  :7744
##  10 years since admission: 4224     Investment income              :7744
##  2 years since admission : 4224     Self-employment income         :7744
##  3 years since admission : 4224     Social welfare benefits        :7744
##  4 years since admission : 4224     Wages, salaries and commissions:7744
##  (Other)                 :21120
##               Statistics         UOM            VALUE
##  Mean with income  :11616   Dollars:23232   Min.   :     0
##  Median with income:11616   Persons:23232   1st Qu.: 11400
##  Total count       :11616                   Median : 19474
##  Total with income :11616                   Mean   : 19474
##                                             3rd Qu.: 19474
##                                             Max.   :198810
##
```

```
as.numeric(Dataset$`Age at admission`)
```

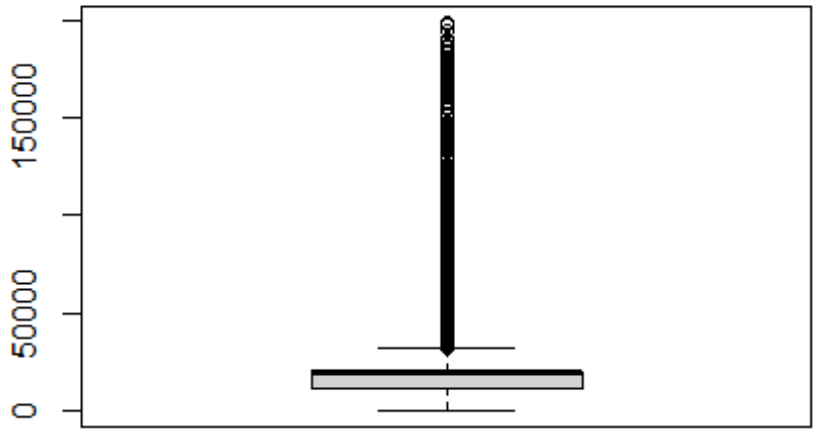
```
as.numeric(Dataset$`Years since admission`)
```

```
as.numeric(Dataset$`Income type`)
```

```
as.numeric(Dataset$Statistics)
```

```
as.numeric(Dataset$UOM)

as.numeric(Dataset$REF_DATE)

Dataset

## # A tibble: 46,464 x 8
##    REF_DATE Sex       `Age at admiss~` `Years since a~` `Income type` Stat
istics
##       <dbl> <fct>     <fct>            <fct>            <fct>         <fct
>
##  1     2006 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  2     2007 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  3     2008 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  4     2009 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  5     2010 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  6     2011 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  7     2012 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  8     2013 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
##  9     2014 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## 10     2015 Both sex~ Total, Age at a~ 0 years since a~ Wages, salar~ Tota
l cou~
## # ... with 46,454 more rows, and 2 more variables: UOM <fct>, VALUE <dbl>

str(Dataset)

## spec_tbl_df [46,464 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ REF_DATE            : num [1:46464] 2006 2007 2008 2009 2010 ...
##  $ Sex                 : Factor w/ 4 levels "Both sexes","Females",..: 1
1 1 1 1 1 1 1 1 ...
##  $ Age at admission    : Factor w/ 4 levels "20 to 24 years",..: 4 4 4 4
4 4 4 4 4 4 ...
##  $ Years since admission: Factor w/ 11 levels "0 years since admission",..
: 1 1 1 1 1 1 1 1 1 ...
##  $ Income type         : Factor w/ 6 levels "All income","Employment insu
rance benefits",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ Statistics          : Factor w/ 4 levels "Mean with income",..: 3 3 3
3 3 3 3 3 3 3 ...
##  $ UOM                 : Factor w/ 2 levels "Dollars","Persons": 2 2 2 2
2 2 2 2 2 2 ...
##  $ VALUE               : num [1:46464] 154640 145895 151290 155340 169745
...
```

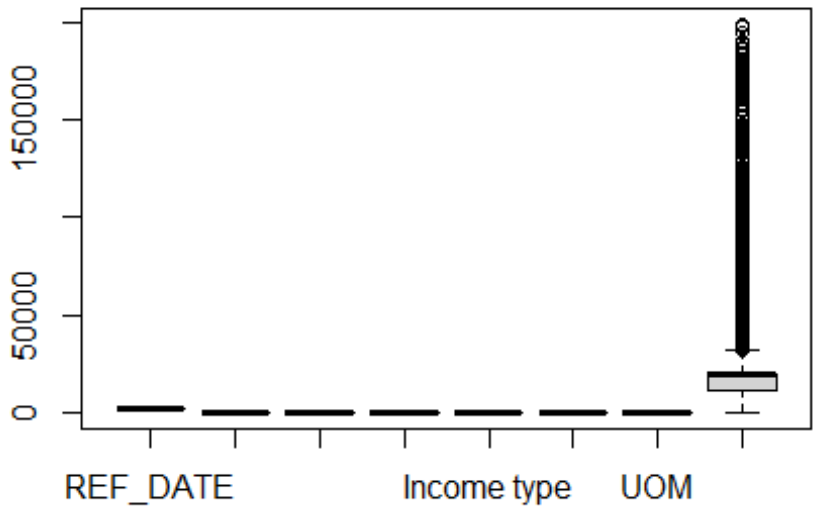
```
##  - attr(*, "spec")=
##   .. cols(
##   ..   REF_DATE = col_double(),
##   ..   GEO = col_character(),
##   ..   DGUID = col_character(),
##   ..   Sex = col_character(),
##   ..   `Age at admission` = col_character(),
##   ..   `Immigrant admission category` = col_character(),
##   ..   `Years since admission` = col_character(),
##   ..   `Income type` = col_character(),
##   ..   Statistics = col_character(),
##   ..   UOM = col_character(),
##   ..   UOM_ID = col_double(),
##   ..   SCALAR_FACTOR = col_character(),
##   ..   SCALAR_ID = col_double(),
##   ..   VECTOR = col_character(),
##   ..   COORDINATE = col_character(),
##   ..   VALUE = col_double(),
##   ..   STATUS = col_character(),
##   ..   SYMBOL = col_logical(),
##   ..   TERMINATED = col_logical(),
##   ..   DECIMALS = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>

Dataset$`Age at admission` <- as.numeric(Dataset$`Age at admission`)
Dataset$`Years since admission` <- as.numeric(Dataset$`Years since admission`
)
Dataset$`Income type` <- as.numeric(Dataset$`Income type`)
Dataset$Statistics <- as.numeric(Dataset$Statistics)
Dataset$UOM <- as.numeric(Dataset$UOM)
Dataset$REF_DATE <- as.numeric(Dataset$REF_DATE)

boxplot(Dataset$VALUE)
```



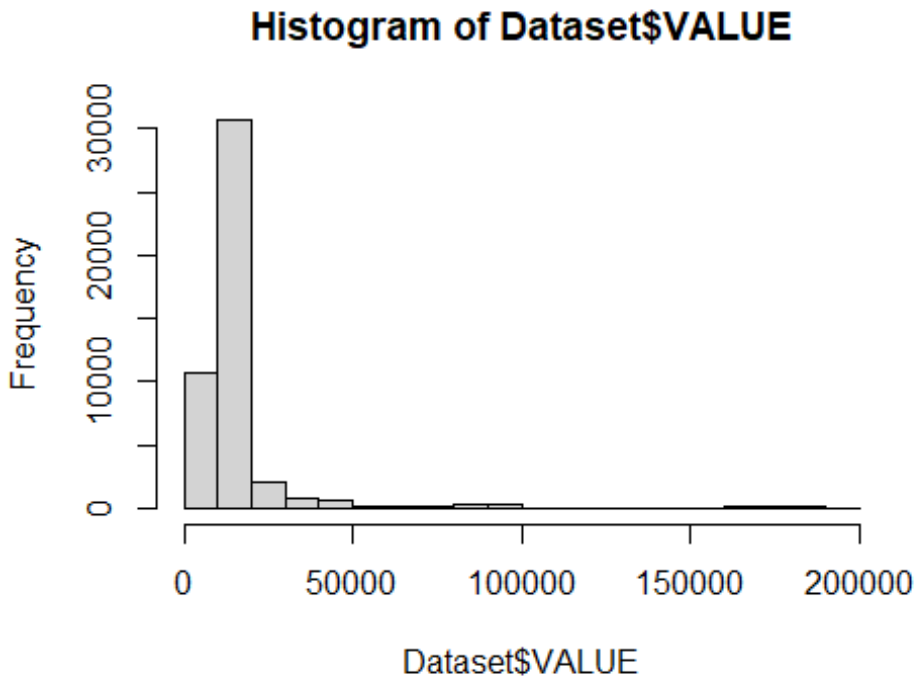

```
str(Dataset)

## spec_tbl_df [46,464 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ REF_DATE             : num [1:46464] 2006 2007 2008 2009 2010 ...
##  $ Sex                  : Factor w/ 4 levels "Both sexes","Females",..: 1
1 1 1 1 1 1 1 1 1 ...
##  $ Age at admission     : num [1:46464] 4 4 4 4 4 4 4 4 4 4 ...
##  $ Years since admission: num [1:46464] 1 1 1 1 1 1 1 1 1 1 ...
##  $ Income type          : num [1:46464] 6 6 6 6 6 6 6 6 6 6 ...
##  $ Statistics           : num [1:46464] 3 3 3 3 3 3 3 3 3 3 ...
##  $ UOM                  : num [1:46464] 2 2 2 2 2 2 2 2 2 2 ...
##  $ VALUE                : num [1:46464] 154640 145895 151290 155340 169745
...
##  - attr(*, "spec")=
##   .. cols(
##   ..   REF_DATE = col_double(),
##   ..   GEO = col_character(),
##   ..   DGUID = col_character(),
##   ..   Sex = col_character(),
##   ..   `Age at admission` = col_character(),
##   ..   `Immigrant admission category` = col_character(),
##   ..   `Years since admission` = col_character(),
##   ..   `Income type` = col_character(),
##   ..   Statistics = col_character(),
##   ..   UOM = col_character(),
##   ..   UOM_ID = col_double(),
##   ..   SCALAR_FACTOR = col_character(),
```

```
##   ..   SCALAR_ID = col_double(),
##   ..   VECTOR = col_character(),
##   ..   COORDINATE = col_character(),
##   ..   VALUE = col_double(),
##   ..   STATUS = col_character(),
##   ..   SYMBOL = col_logical(),
##   ..   TERMINATED = col_logical(),
##   ..   DECIMALS = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>

Dataset$Sex <- as.numeric(Dataset$Sex)
str(Dataset)

## spec_tbl_df [46,464 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ REF_DATE             : num [1:46464] 2006 2007 2008 2009 2010 ...
##  $ Sex                  : num [1:46464] 1 1 1 1 1 1 1 1 1 ...
##  $ Age at admission     : num [1:46464] 4 4 4 4 4 4 4 4 4 ...
##  $ Years since admission: num [1:46464] 1 1 1 1 1 1 1 1 1 ...
##  $ Income type          : num [1:46464] 6 6 6 6 6 6 6 6 6 ...
##  $ Statistics           : num [1:46464] 3 3 3 3 3 3 3 3 3 ...
##  $ UOM                  : num [1:46464] 2 2 2 2 2 2 2 2 2 ...
##  $ VALUE                : num [1:46464] 154640 145895 151290 155340 169745
...
##  - attr(*, "spec")=
##   .. cols(
##   ..   REF_DATE = col_double(),
##   ..   GEO = col_character(),
##   ..   DGUID = col_character(),
##   ..   Sex = col_character(),
##   ..   `Age at admission` = col_character(),
##   ..   `Immigrant admission category` = col_character(),
##   ..   `Years since admission` = col_character(),
##   ..   `Income type` = col_character(),
##   ..   Statistics = col_character(),
##   ..   UOM = col_character(),
##   ..   UOM_ID = col_double(),
##   ..   SCALAR_FACTOR = col_character(),
##   ..   SCALAR_ID = col_double(),
##   ..   VECTOR = col_character(),
##   ..   COORDINATE = col_character(),
##   ..   VALUE = col_double(),
##   ..   STATUS = col_character(),
##   ..   SYMBOL = col_logical(),
##   ..   TERMINATED = col_logical(),
##   ..   DECIMALS = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>

cor(Dataset)
```

```
##                            REF_DATE         Sex Age at admission
## REF_DATE                1.000000000  0.00000000        0.0000000
## Sex                     0.000000000  1.00000000        0.0000000
## Age at admission        0.000000000  0.00000000        1.0000000
## Years since admission   0.000000000  0.00000000        0.0000000
## Income type             0.000000000  0.00000000        0.0000000
## Statistics              0.000000000  0.00000000        0.0000000
## UOM                     0.000000000  0.00000000        0.0000000
## VALUE                  -0.009873198 -0.07097935        0.2154485
##                        Years since admission Income type Statistics       U
OM
## REF_DATE                           0.0000000 0.000000000  0.0000000 0.00000
00
## Sex                                0.0000000 0.000000000  0.0000000 0.00000
00
## Age at admission                   0.0000000 0.000000000  0.0000000 0.00000
00
## Years since admission              1.0000000 0.000000000  0.0000000 0.00000
00
## Income type                        0.0000000 1.000000000  0.0000000 0.00000
00
## Statistics                         0.0000000 0.000000000  1.0000000 0.89442
72
## UOM                                0.0000000 0.000000000  0.8944272 1.00000
00
## VALUE                              0.0314035 0.004756614  0.0480968 0.13811
35
##                              VALUE
## REF_DATE                -0.009873198
## Sex                     -0.070979352
## Age at admission         0.215448501
## Years since admission    0.031403504
## Income type              0.004756614
## Statistics               0.048096797
## UOM                      0.138113459
## VALUE                    1.000000000

boxplot(Dataset)
```

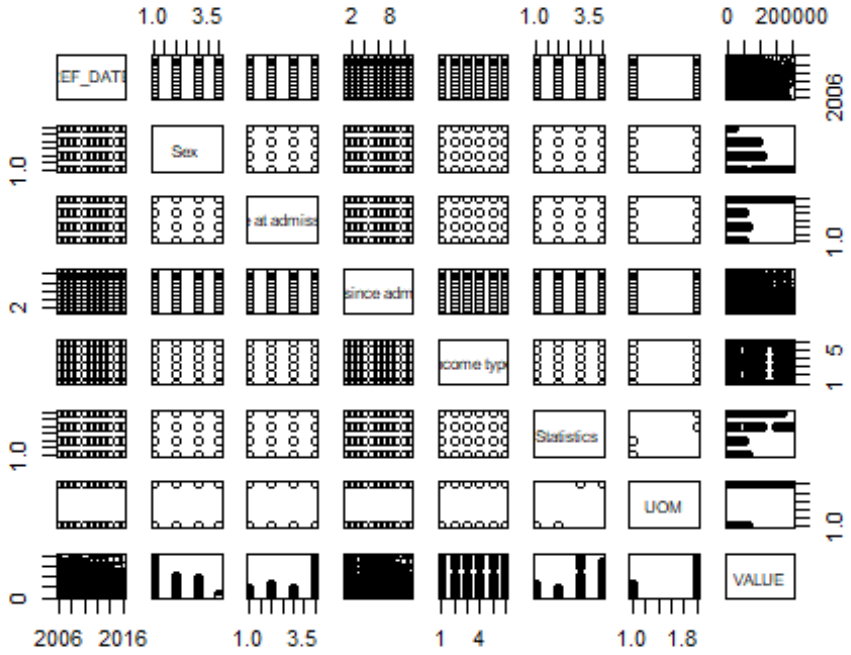



```
hist(Dataset$VALUE)
```




```
princomp(Dataset, cor=TRUE, score=TRUE)
```

```
## Call:
## princomp(x = Dataset, cor = TRUE, scores = TRUE)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    C
omp.8
## 1.3837285 1.1038677 1.0000000 1.0000000 1.0000000 1.0000000 0.8752560 0.31
73302
##
##  8  variables and  46464 observations.

summary(Dataset)

##     REF_DATE         Sex       Age at admission Years since admission
##  Min.   :2006   Min.   :1.00   Min.   :1.00     Min.   : 1
##  1st Qu.:2008   1st Qu.:1.75   1st Qu.:1.75     1st Qu.: 3
##  Median :2011   Median :2.50   Median :2.50     Median : 6
##  Mean   :2011   Mean   :2.50   Mean   :2.50     Mean   : 6
##  3rd Qu.:2014   3rd Qu.:3.25   3rd Qu.:3.25     3rd Qu.: 9
##  Max.   :2016   Max.   :4.00   Max.   :4.00     Max.   :11
##   Income type    Statistics        UOM          VALUE
##  Min.   :1.0   Min.   :1.00   Min.   :1.0   Min.   :     0
##  1st Qu.:2.0   1st Qu.:1.75   1st Qu.:1.0   1st Qu.: 11400
##  Median :3.5   Median :2.50   Median :1.5   Median : 19474
##  Mean   :3.5   Mean   :2.50   Mean   :1.5   Mean   : 19474
##  3rd Qu.:5.0   3rd Qu.:3.25   3rd Qu.:2.0   3rd Qu.: 19474
##  Max.   :6.0   Max.   :4.00   Max.   :2.0   Max.   :198810

plot(Dataset)
```




```
max(Dataset$VALUE)

## [1] 198810

min(Dataset$VALUE)

## [1] 0

cor(Dataset, method= "pearson")

##                            REF_DATE         Sex Age at admission
## REF_DATE                1.000000000  0.00000000        0.0000000
## Sex                     0.000000000  1.00000000        0.0000000
## Age at admission        0.000000000  0.00000000        1.0000000
## Years since admission   0.000000000  0.00000000        0.0000000
## Income type             0.000000000  0.00000000        0.0000000
## Statistics              0.000000000  0.00000000        0.0000000
## UOM                     0.000000000  0.00000000        0.0000000
## VALUE                  -0.009873198 -0.07097935        0.2154485
##                        Years since admission Income type Statistics       U
OM
## REF_DATE                           0.0000000 0.000000000  0.0000000 0.00000
00
## Sex                                0.0000000 0.000000000  0.0000000 0.00000
00
## Age at admission                   0.0000000 0.000000000  0.0000000 0.00000
00
## Years since admission              1.0000000 0.000000000  0.0000000 0.00000
```

```
00
## Income type                        0.0000000 1.000000000  0.0000000 0.00000
00
## Statistics                         0.0000000 0.000000000  1.0000000 0.89442
72
## UOM                                0.0000000 0.000000000  0.8944272 1.00000
00
## VALUE                              0.0314035 0.004756614  0.0480968 0.13811
35
##                                        VALUE
## REF_DATE                        -0.009873198
## Sex                             -0.070979352
## Age at admission                 0.215448501
## Years since admission            0.031403504
## Income type                      0.004756614
## Statistics                       0.048096797
## UOM                              0.138113459
## VALUE                            1.000000000
```

```
cor(Dataset, method = "spearman")
```

```
##                          REF_DATE       Sex Age at admission
## REF_DATE                1.0000000 0.0000000        0.0000000
## Sex                     0.0000000 1.0000000        0.0000000
## Age at admission        0.0000000 0.0000000        1.0000000
## Years since admission 0.0000000 0.0000000        0.0000000
## Income type             0.0000000 0.0000000        0.0000000
## Statistics              0.0000000 0.0000000        0.0000000
## UOM                     0.0000000 0.0000000        0.0000000
## VALUE                   0.1673294 0.1106724        0.1077653
##                         Years since admission Income type  Statistics
UOM
## REF_DATE                           0.0000000  0.00000000  0.00000000 0.0000
000
## Sex                                0.0000000  0.00000000  0.00000000 0.0000
000
## Age at admission                   0.0000000  0.00000000  0.00000000 0.0000
000
## Years since admission              1.0000000  0.00000000  0.00000000 0.0000
000
## Income type                        0.0000000  1.00000000  0.00000000 0.0000
000
## Statistics                         0.0000000  0.00000000  1.00000000 0.8944
272
## UOM                                0.0000000  0.00000000  0.89442719 1.0000
000
## VALUE                              0.1821219  0.01397637 -0.04405075 0.0388
791
##                              VALUE
## REF_DATE                  0.16732941
```

```
## Sex                    0.11067242
## Age at admission       0.10776530
## Years since admission  0.18212189
## Income type            0.01397637
## Statistics            -0.04405075
## UOM                    0.03887910
## VALUE                  1.00000000
```

```
summary (lm(formula= Dataset$VALUE ~ Dataset$REF_DATE + Dataset$Sex + Dataset
$`Age at admission` + Dataset$`Years since admission` + Dataset$`Income type`
+ Dataset$Statistics + Dataset$UOM))
```

```
##
## Call:
## lm(formula = Dataset$VALUE ~ Dataset$REF_DATE + Dataset$Sex +
##     Dataset$`Age at admission` + Dataset$`Years since admission` +
##     Dataset$`Income type` + Dataset$Statistics + Dataset$UOM)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30892  -8869  -2085   4265 165243
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   125174.04   55786.97   2.244   0.0249 *
## Dataset$REF_DATE                 -62.23      27.74  -2.243   0.0249 *
## Dataset$Sex                    -1265.33      78.46 -16.127  < 2e-16 ***
## Dataset$`Age at admission`      3840.74      78.46  48.951  < 2e-16 ***
## Dataset$`Years since admission`  197.93      27.74   7.135 9.81e-13 ***
## Dataset$`Income type`             55.51      51.36   1.081   0.2798
## Dataset$Statistics             -6723.85     175.44 -38.325  < 2e-16 ***
## Dataset$UOM                    18953.16     392.30  48.313  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18910 on 46456 degrees of freedom
## Multiple R-squared:  0.1001, Adjusted R-squared:  0.09995
## F-statistic: 738.1 on 7 and 46456 DF,  p-value: < 2.2e-16
```

```
library(RCurl)
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.1.3
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.1.3
```

```r
cv_train <- sample(nrow(Dataset), floor(nrow(Dataset)*0.7))
Train <- Dataset[cv_train,]
Test <- Dataset[-cv_train,]

Train_model <- lm(VALUE~REF_DATE+Sex+`Age at admission`+`Years since admissio
n`+`Income type`+Statistics+UOM, data=Train)
Prediction_model <- predict(Train_model, interval="prediction", newdata=Test)
summary(Train_model)
```

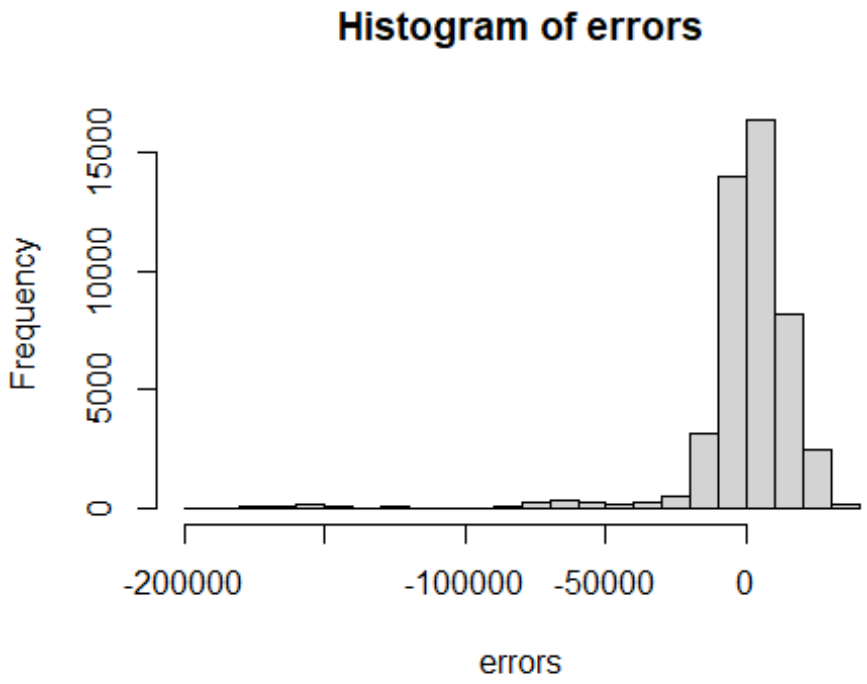
```
##
## Call:
## lm(formula = VALUE ~ REF_DATE + Sex + `Age at admission` + `Years since ad
mission` +
##     `Income type` + Statistics + UOM, data = Train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30971  -8858  -2111   4258 165304
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             98618.99   66868.94   1.475    0.140
## REF_DATE                  -49.14      33.25  -1.478    0.139
## Sex                     -1256.47      93.96 -13.373  < 2e-16 ***
## `Age at admission`       3834.08      93.96  40.807  < 2e-16 ***
## `Years since admission`   200.40      33.30   6.018 1.78e-09 ***
## `Income type`              87.33      61.59   1.418    0.156
## Statistics              -6731.58     210.06 -32.045  < 2e-16 ***
## UOM                     19041.05     470.56  40.465  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18940 on 32516 degrees of freedom
## Multiple R-squared:  0.09996,    Adjusted R-squared:  0.09976
## F-statistic: 515.9 on 7 and 32516 DF,  p-value: < 2.2e-16
```

```r
summary(Prediction_model)
```

```
##       fit             lwr              upr
##  Min.   : 4227   Min.   :-32906   Min.   :41360
##  1st Qu.:14898   1st Qu.:-22231   1st Qu.:52027
##  Median :19509   Median :-17621   Median :56638
##  Mean   :19478   Mean   :-17652   Mean   :56608
##  3rd Qu.:24012   3rd Qu.:-13118   3rd Qu.:61143
##  Max.   :34540   Max.   : -2593   Max.   :71672
```

```r
errors <- Prediction_model[,"fit"] - Dataset$VALUE
```

```
## Warning in Prediction_model[, "fit"] - Dataset$VALUE: longer object length
is
## not a multiple of shorter object length

hist(errors)
```




```
rmse <- sqrt(sum((Prediction_model[,"fit"]- Dataset$VALUE)^2)/nrow(Test))

## Warning in Prediction_model[, "fit"] - Dataset$VALUE: longer object length
is
## not a multiple of shorter object length

rmse

## [1] 38168.99

library(MASS)
library(leaps)

full <- lm(Dataset$VALUE ~Dataset$REF_DATE+Dataset$Sex+Dataset$`Age at admiss
ion`+Dataset$`Years since admission`+Dataset$`Income type`+Dataset$Statistics
+Dataset$UOM)
null <- lm(Dataset$VALUE~1, data=Dataset)
stepF <- stepAIC(null, scope=list(lower=null, upper=full), direction="forward
", trace=TRUE)

## Start:  AIC=919991.6
## Dataset$VALUE ~ 1
```

```
##
##                                   Df  Sum of Sq        RSS    AIC
## + Dataset$`Age at admission`       1 8.5676e+11 1.7601e+13 917785
## + Dataset$UOM                      1 3.5208e+11 1.8105e+13 919099
## + Dataset$Sex                      1 9.2990e+10 1.8364e+13 919759
## + Dataset$Statistics               1 4.2697e+10 1.8415e+13 919886
## + Dataset$`Years since admission`  1 1.8202e+10 1.8439e+13 919948
## + Dataset$REF_DATE                 1 1.7992e+09 1.8456e+13 919989
## <none>                                          1.8457e+13 919992
## + Dataset$`Income type`            1 4.1761e+08 1.8457e+13 919993
##
## Step:  AIC=917785.2
## Dataset$VALUE ~ Dataset$`Age at admission`
##
##                                   Df  Sum of Sq        RSS    AIC
## + Dataset$UOM                      1 3.5208e+11 1.7249e+13 916848
## + Dataset$Sex                      1 9.2990e+10 1.7508e+13 917541
## + Dataset$Statistics               1 4.2697e+10 1.7558e+13 917674
## + Dataset$`Years since admission`  1 1.8202e+10 1.7582e+13 917739
## + Dataset$REF_DATE                 1 1.7992e+09 1.7599e+13 917782
## <none>                                          1.7601e+13 917785
## + Dataset$`Income type`            1 4.1761e+08 1.7600e+13 917786
##
## Step:  AIC=916848.3
## Dataset$VALUE ~ Dataset$`Age at admission` + Dataset$UOM
##
##                                   Df  Sum of Sq        RSS    AIC
## + Dataset$Statistics               1 5.2516e+11 1.6723e+13 915414
## + Dataset$Sex                      1 9.2990e+10 1.7156e+13 916599
## + Dataset$`Years since admission`  1 1.8202e+10 1.7230e+13 916801
## + Dataset$REF_DATE                 1 1.7992e+09 1.7247e+13 916845
## <none>                                          1.7249e+13 916848
## + Dataset$`Income type`            1 4.1761e+08 1.7248e+13 916849
##
## Step:  AIC=915413.7
## Dataset$VALUE ~ Dataset$`Age at admission` + Dataset$UOM + Dataset$Statist
ics
##
##                                   Df  Sum of Sq        RSS    AIC
## + Dataset$Sex                      1 9.2990e+10 1.6630e+13 915157
## + Dataset$`Years since admission`  1 1.8202e+10 1.6705e+13 915365
## + Dataset$REF_DATE                 1 1.7992e+09 1.6722e+13 915411
## <none>                                          1.6723e+13 915414
## + Dataset$`Income type`            1 4.1761e+08 1.6723e+13 915414
##
## Step:  AIC=915156.6
## Dataset$VALUE ~ Dataset$`Age at admission` + Dataset$UOM + Dataset$Statist
ics +
##     Dataset$Sex
##
```

```
##                                     Df  Sum of Sq        RSS    AIC
## + Dataset$`Years since admission`  1 1.8202e+10 1.6612e+13 915108
## + Dataset$REF_DATE                 1 1.7992e+09 1.6629e+13 915154
## <none>                                          1.6630e+13 915157
## + Dataset$`Income type`            1 4.1761e+08 1.6630e+13 915157
##
## Step:  AIC=915107.7
## Dataset$VALUE ~ Dataset$`Age at admission` + Dataset$UOM + Dataset$Statist
ics +
##     Dataset$Sex + Dataset$`Years since admission`
##
##                          Df  Sum of Sq        RSS    AIC
## + Dataset$REF_DATE        1 1799225877 1.6610e+13 915105
## <none>                                 1.6612e+13 915108
## + Dataset$`Income type`   1  417605050 1.6612e+13 915109
##
## Step:  AIC=915104.7
## Dataset$VALUE ~ Dataset$`Age at admission` + Dataset$UOM + Dataset$Statist
ics +
##     Dataset$Sex + Dataset$`Years since admission` + Dataset$REF_DATE
##
##                          Df Sum of Sq       RSS    AIC
## <none>                                1.661e+13 915105
## + Dataset$`Income type`   1 417605050 1.661e+13 915105
```

```
summary(stepF)
```

```
##
## Call:
## lm(formula = Dataset$VALUE ~ Dataset$`Age at admission` + Dataset$UOM +
##     Dataset$Statistics + Dataset$Sex + Dataset$`Years since admission` +
##     Dataset$REF_DATE, data = Dataset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30753  -8875  -2087   4265 165104
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     125368.33   55786.78   2.247   0.0246 *
## Dataset$`Age at admission`        3840.74      78.46  48.951  < 2e-16 ***
## Dataset$UOM                      18953.16     392.30  48.313  < 2e-16 ***
## Dataset$Statistics               -6723.85     175.44 -38.325  < 2e-16 ***
## Dataset$Sex                      -1265.33      78.46 -16.127  < 2e-16 ***
## Dataset$`Years since admission`    197.93      27.74   7.135 9.81e-13 ***
## Dataset$REF_DATE                   -62.23      27.74  -2.243   0.0249 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18910 on 46457 degrees of freedom
```

```
## Multiple R-squared:  0.1001, Adjusted R-squared:  0.09995
## F-statistic:   861 on 6 and 46457 DF,  p-value: < 2.2e-16
```

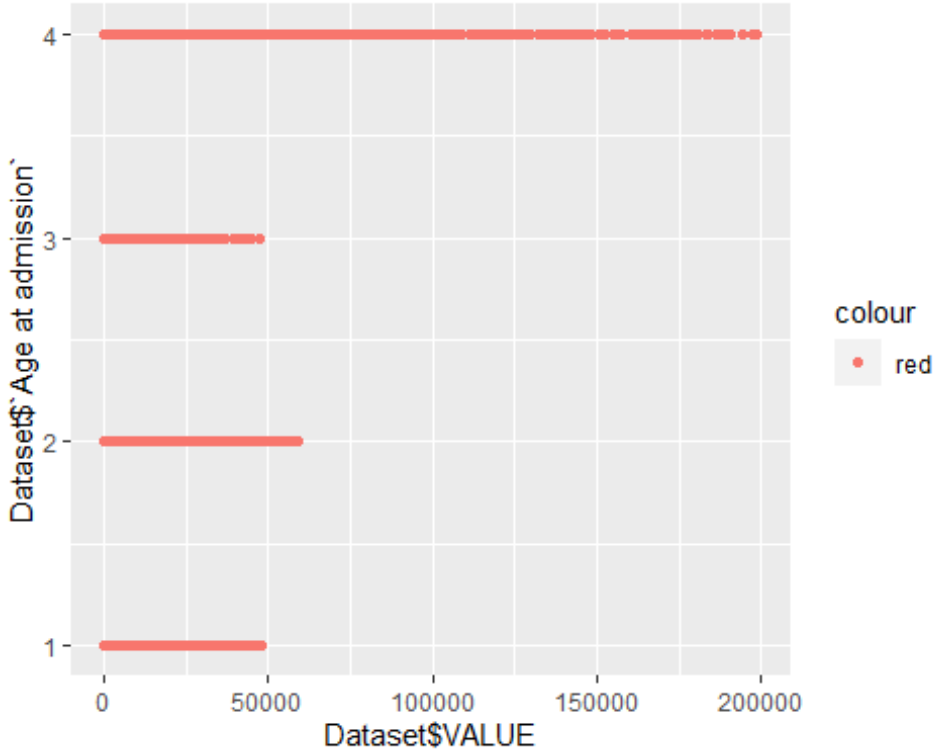
```
full <- lm(Dataset$VALUE~Dataset$REF_DATE+Dataset$Sex+Dataset$`Age at admissi
on`+Dataset$`Years since admission`+Dataset$`Income type`+Dataset$Statistics+
Dataset$UOM)
stepB <- stepAIC(full, direction="backward", trace=TRUE)
```

```
## Start:  AIC=915105.5
## Dataset$VALUE ~ Dataset$REF_DATE + Dataset$Sex + Dataset$`Age at admission
` +
##     Dataset$`Years since admission` + Dataset$`Income type` +
##     Dataset$Statistics + Dataset$UOM
##
##                                    Df  Sum of Sq        RSS    AIC
## - Dataset$`Income type`             1 4.1761e+08 1.6610e+13 915105
## <none>                                           1.6610e+13 915105
## - Dataset$REF_DATE                  1 1.7992e+09 1.6612e+13 915109
## - Dataset$`Years since admission`   1 1.8202e+10 1.6628e+13 915154
## - Dataset$Sex                       1 9.2990e+10 1.6703e+13 915363
## - Dataset$Statistics                1 5.2516e+11 1.7135e+13 916550
## - Dataset$UOM                       1 8.3454e+11 1.7445e+13 917381
## - Dataset$`Age at admission`        1 8.5676e+11 1.7467e+13 917440
##
## Step:  AIC=915104.7
## Dataset$VALUE ~ Dataset$REF_DATE + Dataset$Sex + Dataset$`Age at admission
` +
##     Dataset$`Years since admission` + Dataset$Statistics + Dataset$UOM
##
##                                    Df  Sum of Sq        RSS    AIC
## <none>                                           1.6610e+13 915105
## - Dataset$REF_DATE                  1 1.7992e+09 1.6612e+13 915108
## - Dataset$`Years since admission`   1 1.8202e+10 1.6629e+13 915154
## - Dataset$Sex                       1 9.2990e+10 1.6703e+13 915362
## - Dataset$Statistics                1 5.2516e+11 1.7136e+13 916549
## - Dataset$UOM                       1 8.3454e+11 1.7445e+13 917380
## - Dataset$`Age at admission`        1 8.5676e+11 1.7467e+13 917439
```

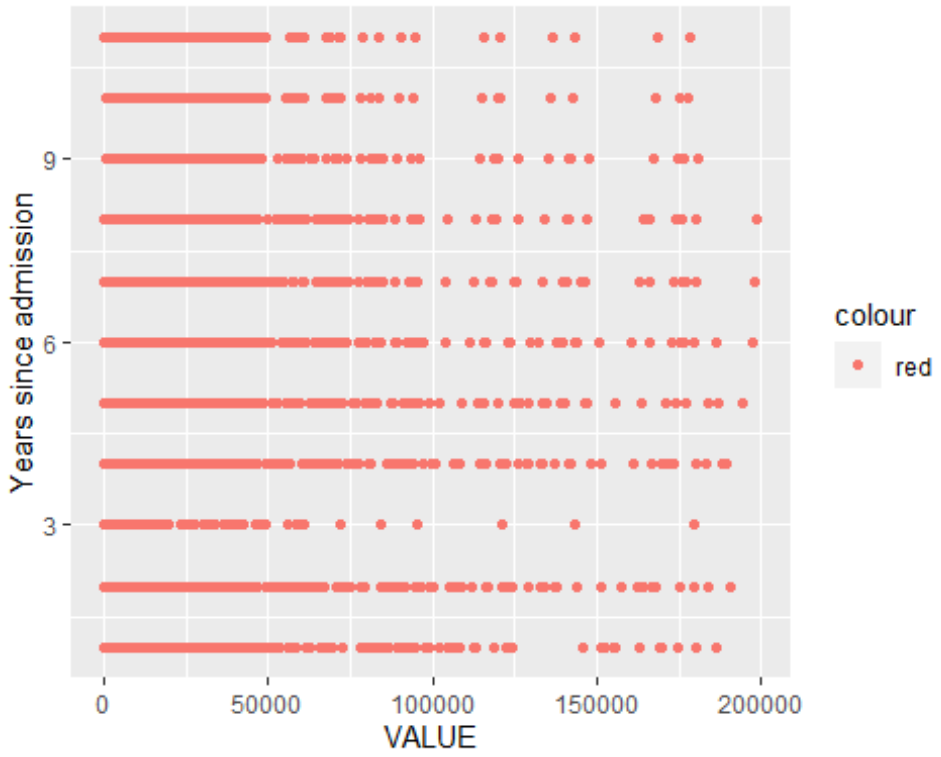
```
summary(stepB)
```

```
##
## Call:
## lm(formula = Dataset$VALUE ~ Dataset$REF_DATE + Dataset$Sex +
##     Dataset$`Age at admission` + Dataset$`Years since admission` +
##     Dataset$Statistics + Dataset$UOM)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30753  -8875  -2087   4265 165104
##
## Coefficients:
```

```
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    125368.33   55786.78   2.247   0.0246 *
## Dataset$REF_DATE                  -62.23      27.74  -2.243   0.0249 *
## Dataset$Sex                     -1265.33      78.46 -16.127  < 2e-16 ***
## Dataset$`Age at admission`       3840.74      78.46  48.951  < 2e-16 ***
## Dataset$`Years since admission`   197.93      27.74   7.135 9.81e-13 ***
## Dataset$Statistics              -6723.85     175.44 -38.325  < 2e-16 ***
## Dataset$UOM                     18953.16     392.30  48.313  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18910 on 46457 degrees of freedom
## Multiple R-squared:  0.1001, Adjusted R-squared:  0.09995
## F-statistic:   861 on 6 and 46457 DF,  p-value: < 2.2e-16

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.3

ggplot(Dataset, aes(Dataset$VALUE, Dataset$`Age at admission`,  col ="red"))+
geom_point()
```




```
ggplot(Dataset, aes(VALUE, `Years since admission`,  col ="red"))+geom_point(
)
```

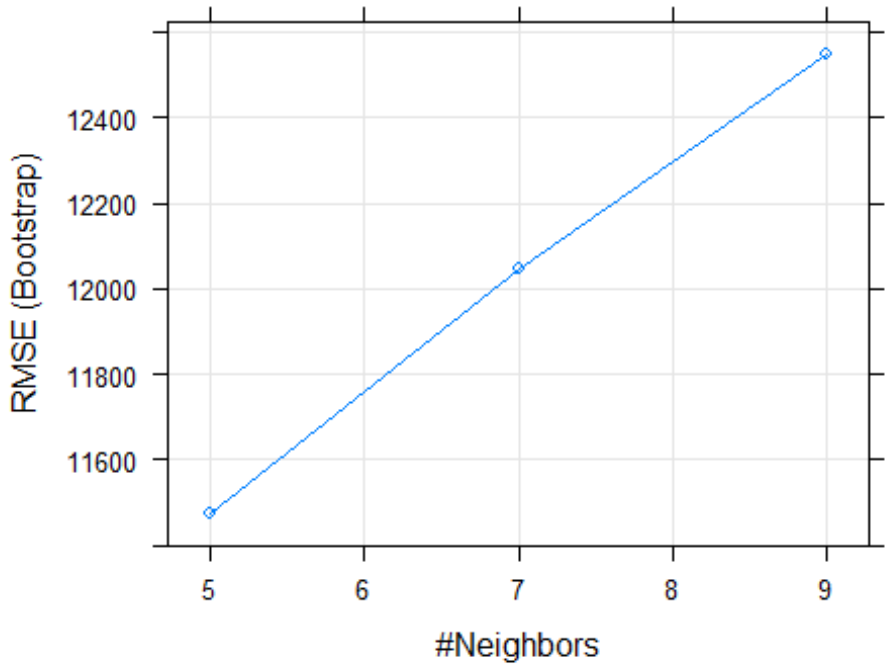



```
library(caret)

## Warning: package 'caret' was built under R version 4.1.3

## Loading required package: lattice

model_Reg <- knnreg(Dataset$VALUE~Dataset$REF_DATE+Dataset$Sex+Dataset$`Age a
t admission`+Dataset$`Years since admission`+Dataset$`Income type`+Dataset$St
atistics+Dataset$UOM, data= Dataset)
model_Reg

## 5-nearest neighbor regression model

set.seed(1)
Training <- createDataPartition(Dataset$VALUE, p= .70, list= FALSE)
training_ <- Dataset[Training,]
testing_ <- Dataset[-Training,]

library(Metrics)

## Warning: package 'Metrics' was built under R version 4.1.3

##
## Attaching package: 'Metrics'

## The following objects are masked from 'package:caret':
##
##     precision, recall
```

```
model_RegTr <- train(VALUE~., data=training_, method='knn')
model_RegTr
```

```
## k-Nearest Neighbors
##
## 32526 samples
##     7 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 32526, 32526, 32526, 32526, 32526, 32526, ...
## Resampling results across tuning parameters:
##
##   k  RMSE      Rsquared   MAE
##   5  11474.07  0.6814075  4447.038
##   7  12046.55  0.6625394  4872.494
##   9  12548.42  0.6447056  5230.346
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 5.
```

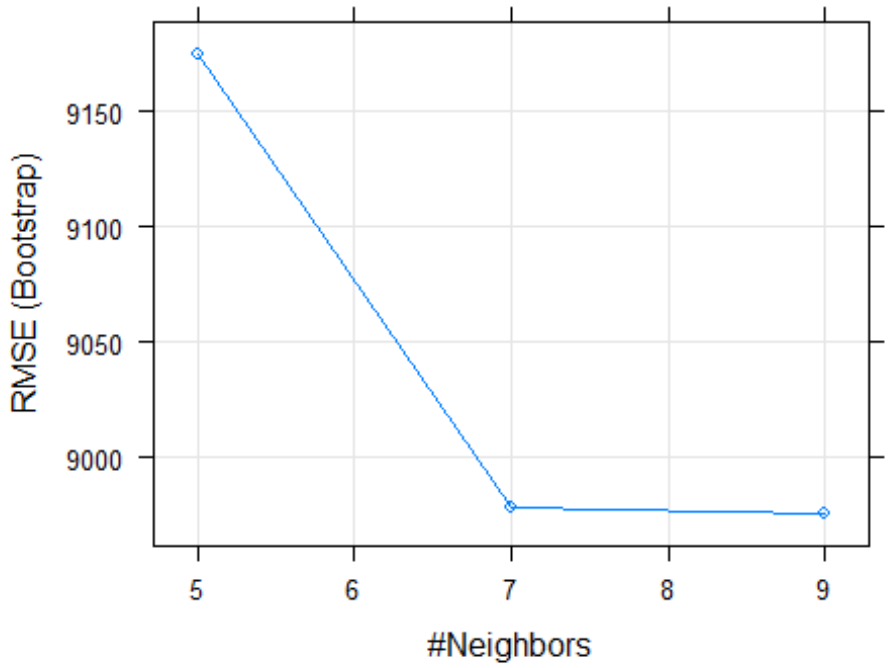
```
plot(model_RegTr)
```




```
model_RegTrp <- train(VALUE~., data=training_, method='knn', preProcess=c("ce
nter", "scale"))
model_RegTrp
```

```
## k-Nearest Neighbors
##
## 32526 samples
##     7 predictor
##
## Pre-processing: centered (7), scaled (7)
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 32526, 32526, 32526, 32526, 32526, 32526, ...
## Resampling results across tuning parameters:
##
##   k  RMSE      Rsquared   MAE
##   5  9174.648  0.7840895  3238.261
##   7  8978.190  0.7932470  3393.637
##   9  8975.777  0.7940605  3546.619
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 9.
```

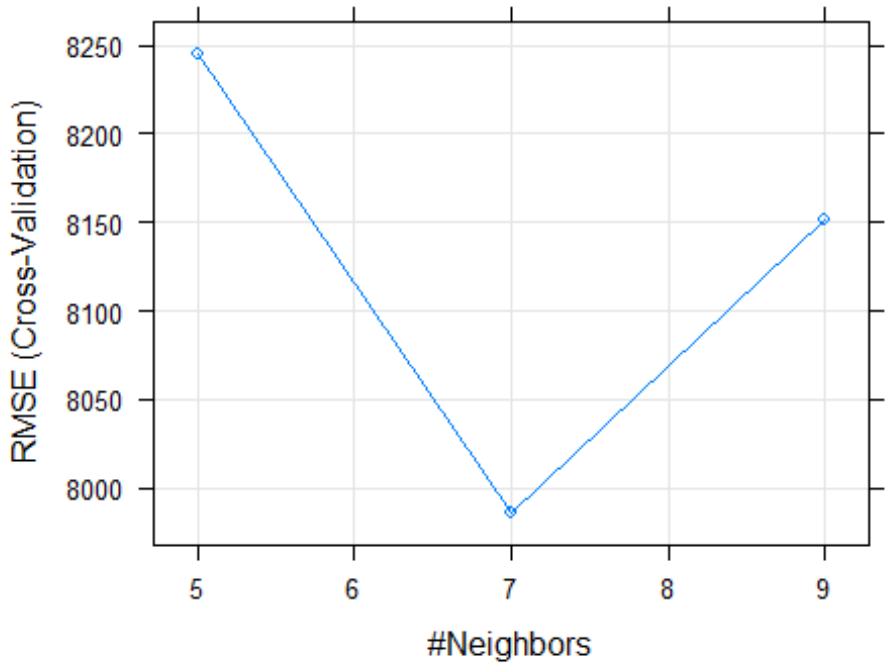
```
plot(model_RegTrp)
```




```
predictions_ = predict(model_RegTrp, newdata=testing_, interval = "prediction
")
summary(predictions_)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     177   12798   19474   19329   19474  185712
```

```
rmse_ <- rmse(testing_$VALUE, predictions_)
rmse_

## [1] 8256.209

R2 <- cor(predictions_, testing_$VALUE)^2
R2

## [1] 0.8401004

set.seed(1)
cross_validation <- trainControl(method="CV", number=10)

set.seed(1)
Model_cv <- train(VALUE~., data=training_, method='knn', preProcess= c("cente
r", "scale"), trControl=cross_validation)
Model_cv

## k-Nearest Neighbors
##
## 32526 samples
##     7 predictor
##
## Pre-processing: centered (7), scaled (7)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 29274, 29273, 29273, 29273, 29274, 29274, ...
## Resampling results across tuning parameters:
##
##   k  RMSE      Rsquared   MAE
##   5  8244.859  0.8254507  3111.160
##   7  7986.169  0.8385622  3218.821
##   9  8151.592  0.8328720  3415.836
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 7.

plot(Model_cv)
```




```
predictions_a = predict(Model_cv, newdata=testing_, interval="prediction")
summary(predictions_a)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     172.9   12478.4   19474.4   19385.5   19474.4 187260.0

rmse_a <- rmse(testing_$VALUE, predictions_a)
rmse_a

## [1] 8159.73

R2a <- cor(predictions_a, testing_$VALUE)^2
R2a

## [1] 0.8420954

set.seed(1)
Model_cva <- train(VALUE~., data=training_, method='knn', trControl=cross_val
idation)
Model_cva

## k-Nearest Neighbors
##
## 32526 samples
##     7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
```

```
## Summary of sample sizes: 29274, 29273, 29273, 29273, 29274, 29274, ...
## Resampling results across tuning parameters:
##
##   k  RMSE      Rsquared   MAE
##   5  10115.21  0.7849304  4116.906
##   7  11781.23  0.7170213  4932.901
##   9  12521.12  0.6844175  5437.250
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 5.
```

```
predictions_a2 = predict(Model_cva, newdata=testing_)
summary(predictions_a2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   14683   19474   19319   19474  182215
```

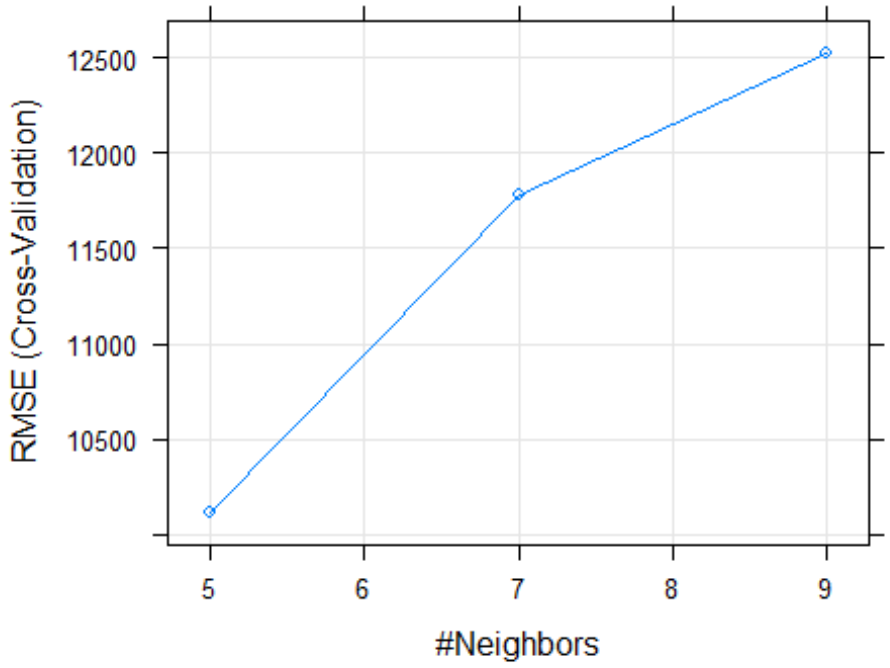
```
rmse_a2 <- rmse(testing_$VALUE, predictions_a2)
rmse_a2
```

```
## [1] 9594.846
```

```
R2a2 <- cor(predictions_a2, testing_$VALUE)^2
R2a2
```

```
## [1] 0.8262025
```

```
plot(Model_cva)
```



```
str(Dataset)
```

```
## spec_tbl_df [46,464 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ REF_DATE             : num [1:46464] 2006 2007 2008 2009 2010 ...
##  $ Sex                  : num [1:46464] 1 1 1 1 1 1 1 1 1 ...
##  $ Age at admission     : num [1:46464] 4 4 4 4 4 4 4 4 4 ...
##  $ Years since admission: num [1:46464] 1 1 1 1 1 1 1 1 1 ...
##  $ Income type          : num [1:46464] 6 6 6 6 6 6 6 6 6 ...
##  $ Statistics           : num [1:46464] 3 3 3 3 3 3 3 3 3 ...
##  $ UOM                  : num [1:46464] 2 2 2 2 2 2 2 2 2 ...
##  $ VALUE                : num [1:46464] 154640 145895 151290 155340 169745
...
##  - attr(*, "spec")=
##   .. cols(
##   ..   REF_DATE = col_double(),
##   ..   GEO = col_character(),
##   ..   DGUID = col_character(),
##   ..   Sex = col_character(),
##   ..   `Age at admission` = col_character(),
##   ..   `Immigrant admission category` = col_character(),
##   ..   `Years since admission` = col_character(),
##   ..   `Income type` = col_character(),
##   ..   Statistics = col_character(),
##   ..   UOM = col_character(),
##   ..   UOM_ID = col_double(),
##   ..   SCALAR_FACTOR = col_character(),
##   ..   SCALAR_ID = col_double(),
##   ..   VECTOR = col_character(),
##   ..   COORDINATE = col_character(),
##   ..   VALUE = col_double(),
##   ..   STATUS = col_character(),
##   ..   SYMBOL = col_logical(),
##   ..   TERMINATED = col_logical(),
##   ..   DECIMALS = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 4.1.3
```

```
model = rpart(VALUE~., data=Dataset)
model
```

```
## n= 46464
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##   1) root 46464 1.845738e+13  19474.39
##     2) Age at admission< 3.5 34848 2.304999e+12  16091.55 *
```

```
##     3) Age at admission>=3.5 11616 1.455723e+13  29622.91
##       6) UOM< 1.5 5808 3.767205e+11  16894.85 *
##       7) UOM>=1.5 5808 1.229868e+13  42350.97
##        14) Sex>=1.5 4356 3.517025e+12  32713.61
##          28) Statistics>=3.5 2178 5.751819e+11  21829.36 *
##          29) Statistics< 3.5 2178 2.425801e+12  43597.86
##            58) Sex>=3.5 726 1.893904e+10  18026.54 *
##            59) Sex< 3.5 1452 1.694773e+12  56383.52
##             118) REF_DATE>=2010.5 792 8.337574e+11  41489.28
##               236) Years since admission>=2.5 648 4.937165e+11  32586.17 *
##               237) Years since admission< 2.5 144 5.753902e+10  81553.28 *
##             119) REF_DATE< 2010.5 660 4.744838e+11  74256.62 *
##        15) Sex< 1.5 1452 7.163334e+12  71263.05
##          30) Statistics>=3.5 726 1.253292e+12  38610.22
##            60) Income type>=1.5 605 5.668959e+11  29918.38
##             120) Income type< 5.5 484 1.281742e+10  19434.83 *
##             121) Income type>=5.5 121 2.881085e+11  71852.61 *
##            61) Income type< 1.5 121 4.121569e+11  82069.39 *
##          31) Statistics< 3.5 726 4.361909e+12 103915.90
##            62) REF_DATE>=2010.5 396 2.151729e+12  69701.63
##             124) Years since admission>=2.5 324 1.270465e+12  49304.04
##               248) Years since admission>=6.5 180 1.439596e+11  24725.91 *
##               249) Years since admission< 6.5 144 8.818510e+11  80026.70
##                 498) REF_DATE>=2013.5 72 1.557880e+11  33499.44 *
##                 499) REF_DATE< 2013.5 72 4.143338e+11 126554.00
##                   998) Years since admission< 3.5 18 2.382280e-22  19474.3
9 *
##                   999) Years since admission>=3.5 54 1.391489e+11 162247.2
0 *
##             125) Years since admission< 2.5 72 1.398420e+11 161490.80 *
##            63) REF_DATE< 2010.5 330 1.190341e+12 144973.00
##             126) Years since admission>=9.5 60 3.558550e+11  96419.69
##               252) REF_DATE>=2008.5 24 1.270549e-21  19474.39 *
##               253) REF_DATE< 2008.5 36 1.190318e+11 147716.60 *
##             127) Years since admission< 9.5 270 6.616083e+11 155762.60
##               254) Years since admission< 3.5 90 3.712921e+11 124810.50
##                 508) Years since admission>=2.5 30 1.227965e+11  51463.51
*
##                 509) Years since admission< 2.5 60 6.405462e+09 161484.00
*
##               255) Years since admission>=3.5 180 1.609819e+11 171238.60 *
```

```
Dataset$`Age at admission` <- as.numeric(Dataset$`Age at admission`)
Dataset$`Years since admission` <- as.numeric(Dataset$`Years since admission`
)
Dataset$`Income type` <- as.numeric(Dataset$`Income type`)
Dataset$Statistics <- as.numeric(Dataset$Statistics)
Dataset$UOM <- as.numeric(Dataset$UOM)
Dataset$REF_DATE <- as.numeric(Dataset$REF_DATE)
Dataset$Sex <- as.numeric(Dataset$Sex)
```

```
library(janitor)
```

```
## Warning: package 'janitor' was built under R version 4.1.3
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
Dataset2 <- clean_names(Dataset)
names(Dataset2)
```

```
## [1] "ref_date"              "sex"                   "age_at_admission"
## [4] "years_since_admission" "income_type"           "statistics"
## [7] "uom"                   "value"
```

```
library(caret)
library(rpart)
library(mlbench)
```

```
## Warning: package 'mlbench' was built under R version 4.1.3
```

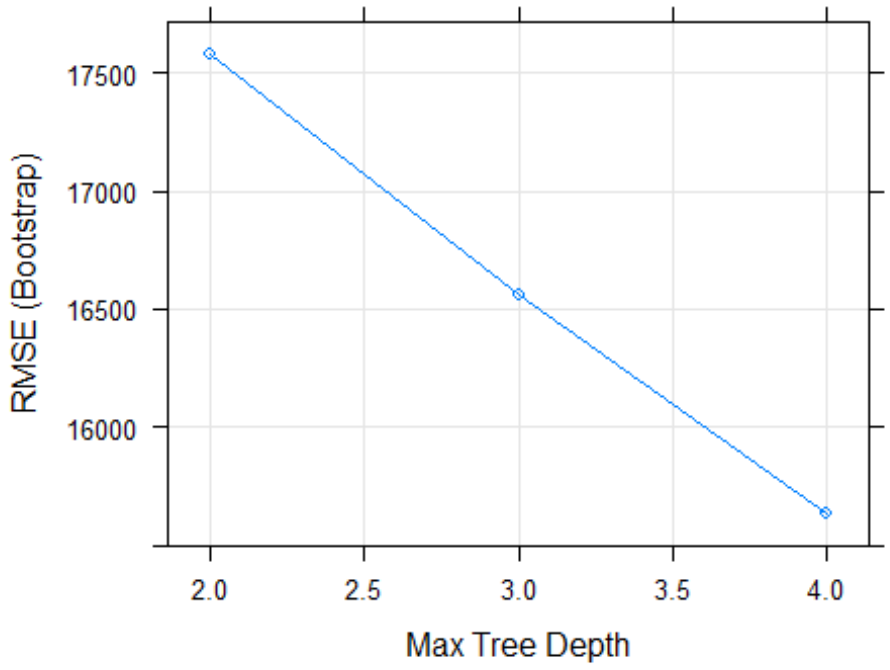
```
data(Dataset2)
```

```
## Warning in data(Dataset2): data set 'Dataset2' not found
```

```
set.seed(1)
model_T0 <- train(value~ref_date+sex+ age_at_admission+years_since_admission+
income_type+statistics+uom,
                  data = Dataset2, method = 'rpart2')
model_T0
```

```
## CART
##
## 46464 samples
##     7 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 46464, 46464, 46464, 46464, 46464, 46464, ...
## Resampling results across tuning parameters:
##
##   maxdepth  RMSE      Rsquared   MAE
##   2         17578.33  0.2283667  9780.447
##   3         16561.01  0.3158230  9518.228
##   4         15633.89  0.3909382  9083.262
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was maxdepth = 4.
```

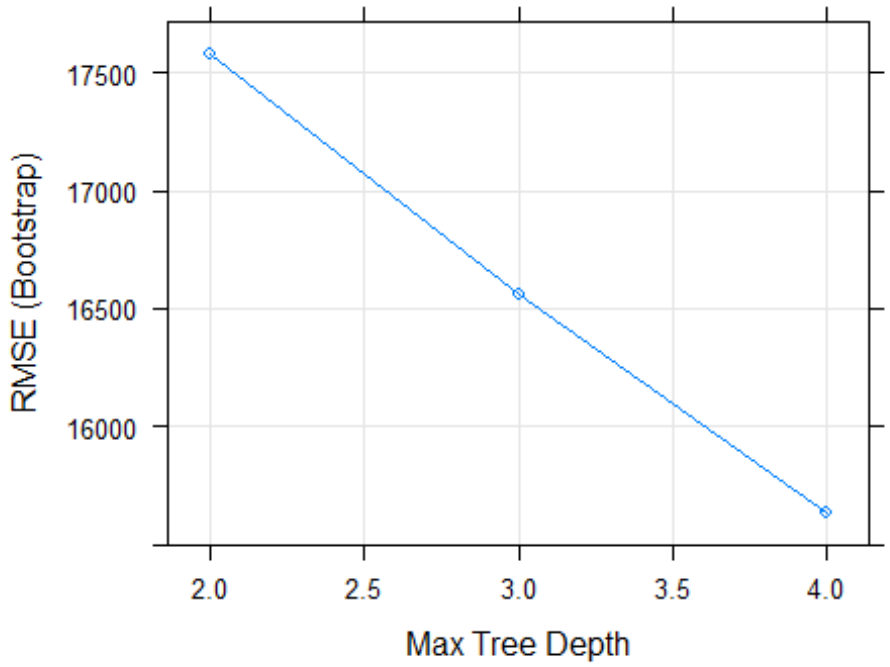
```
plot(model_T0)
```




```
set.seed(1)
model_T1 <- train(value~., data = Dataset2, method = "rpart2", preProcess = c
("center", "scale"))
model_T1
```

```
## CART
##
## 46464 samples
##     7 predictor
##
## Pre-processing: centered (7), scaled (7)
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 46464, 46464, 46464, 46464, 46464, 46464, ...
## Resampling results across tuning parameters:
##
##   maxdepth  RMSE      Rsquared   MAE
##   2         17578.33  0.2283667  9780.447
##   3         16561.01  0.3158230  9518.228
##   4         15633.89  0.3909382  9083.262
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was maxdepth = 4.
```

```
plot(model_T1)
```




```
set.seed(1)
intraining1 <- createDataPartition(Dataset2$value, p = 0.70, list = FALSE)
training1 <- Dataset2[intraining1,]
testing1 <- Dataset2[-intraining1, ]

set.seed(1)
model_T2 <- train(value~., data = training1, method = "rpart2", preProcess =
c("center", "scale"))
model_T2

## CART
##
## 32526 samples
##     7 predictor
##
## Pre-processing: centered (7), scaled (7)
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 32526, 32526, 32526, 32526, 32526, 32526, ...
## Resampling results across tuning parameters:
##
##   maxdepth  RMSE      Rsquared   MAE
##   2         17480.62  0.2181907  9778.142
##   3         16350.84  0.3155506  9447.933
##   4         15645.85  0.3742648  9091.924
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was maxdepth = 4.
```

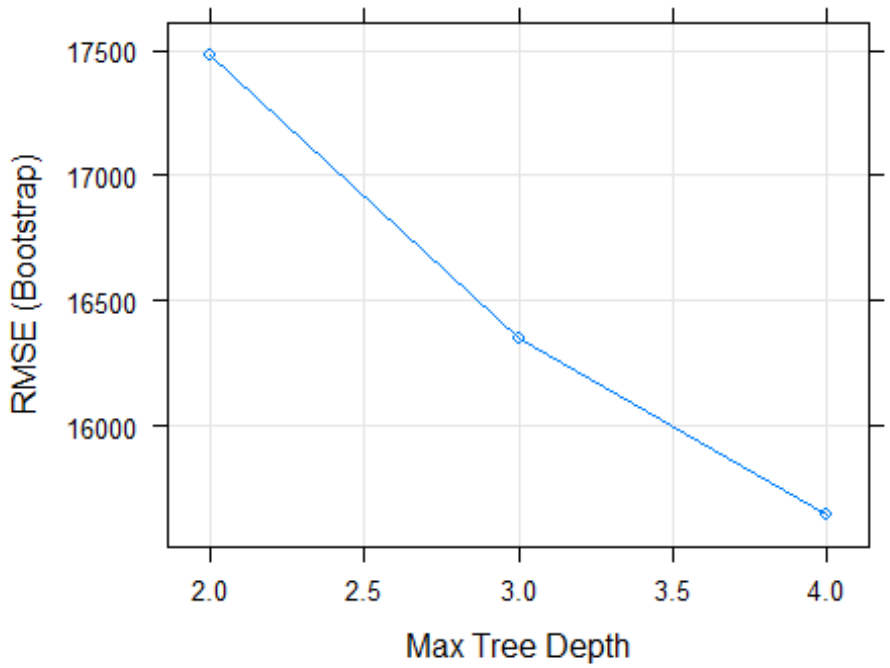
```
predictions_1 = predict(model_T2, newdata = testing1)
summary(predictions_1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16092   16092   16092   19323   16092  100771
```

```
rmse1 <- rmse(testing1$value, predictions_1)
rmse1
```

```
## [1] 15579.8
```

```
plot(model_T2)
```



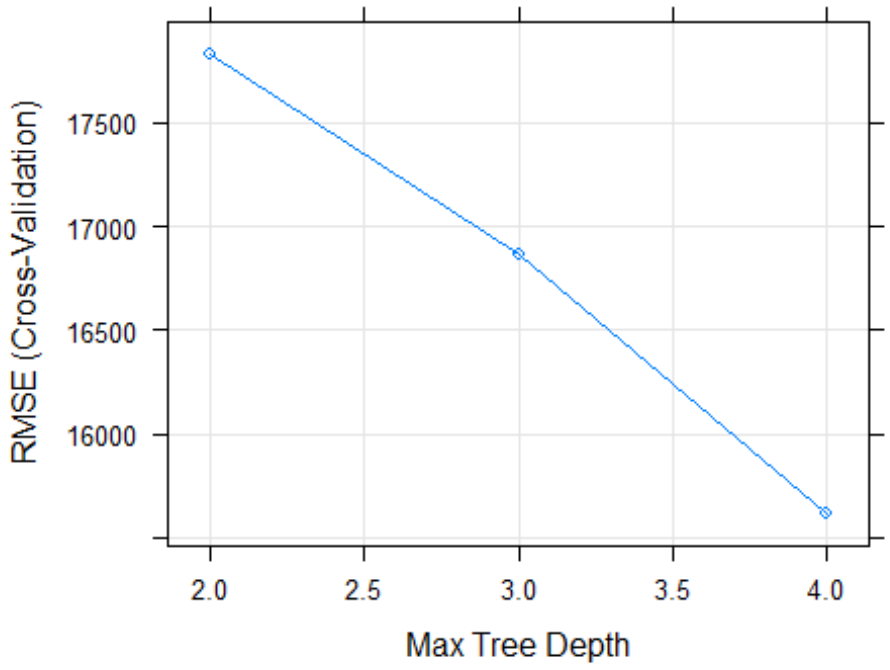

```
r2 <- cor(predictions_1, testing1$value)^2
r2
```

```
## [1] 0.415127
```

```
set.seed(1)
ctrl <- trainControl(method = "cv", number = 10)
model_T3 <- train(value~., data = training1, method = "rpart2", preProcess =
c("center", "scale"), trControl = ctrl)
model_T3
```

```
## CART
##
## 32526 samples
##     7 predictor
```

```
##
## Pre-processing: centered (7), scaled (7)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 29274, 29273, 29273, 29273, 29274, 29274, ...
## Resampling results across tuning parameters:
##
##   maxdepth  RMSE      Rsquared   MAE
##   2         17828.11  0.1858513  9877.689
##   3         16860.39  0.2727197  9667.532
##   4         15618.05  0.3760191  9106.360
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was maxdepth = 4.

plot(model_T3)
```




```
predictions_2 = predict(model_T3, newdata = testing1)
summary(predictions_2)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16092   16092   16092   19323   16092  100771

rmse2 <- rmse(testing1$value, predictions_2)
rmse2

## [1] 15579.8
```

```
r3 <- cor(predictions_2, testing1$value)^2
r3

## [1] 0.415127

library(earth)

## Warning: package 'earth' was built under R version 4.1.3

## Loading required package: Formula

## Loading required package: plotmo

## Warning: package 'plotmo' was built under R version 4.1.3

## Loading required package: plotrix

## Loading required package: TeachingDemos

## Warning: package 'TeachingDemos' was built under R version 4.1.3

library(Formula)
library(plotmo)
library(plotrix)
library(TeachingDemos)

set.seed(1)
tuneGrid <- expand.grid(degree = 1, nprune = c(2, 11, 10))
model_4 <- train(value~., data = training1, method = "earth", preProcess = c(
"center", "scale"), trControl = ctrl, tuneGrid = tuneGrid)
model_4

## Multivariate Adaptive Regression Spline
##
## 32526 samples
##     7 predictor
##
## Pre-processing: centered (7), scaled (7)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 29274, 29273, 29273, 29273, 29274, 29274, ...
## Resampling results across tuning parameters:
##
##   nprune  RMSE      Rsquared    MAE
##    2      18898.88  0.08499003  10069.11
##   10      17837.13  0.18486289  10255.50
##   11      17822.38  0.18618976  10239.27
##
## Tuning parameter 'degree' was held constant at a value of 1
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were nprune = 11 and degree = 1.

predictions_3 = predict(model_4, newdata = testing1)
summary(predictions_3)
```

```
##        y
##  Min.   : 4687
##  1st Qu.:12556
##  Median :18657
##  Mean   :19249
##  3rd Qu.:24819
##  Max.   :45956
```

```
rmse3 <- rmse(testing1$value, predictions_3)
rmse3
```

```
## [1] 18188.64
```

```
r4 <- cor(predictions_3, testing1$value)^2
r4
```

```
##        [,1]
## y 0.1971899
```

```
plot(model_4)
```