Final Project

Shaya Engelman

2024-04-02

```r
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.3.2
```

```r
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2

## Warning: package 'stringr' was built under R version 4.3.2

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(summarytools)
```

```
## Warning: package 'summarytools' was built under R version 4.3.3

##
## Attaching package: 'summarytools'
##
## The following object is masked from 'package:tibble':
##
##     view
```

```r
library(ggcorrplot)
```

Research Question

Can we predict which patients will show up for their appointments versus those who will either cancel last minute or not show up at all?

Justification

No-shows are a common problem in healthcare. They lead to wasted resources, lost revenue, and can have a negative impact on patient outcomes. By identifying patients who are at risk of not showing up for their appointments, healthcare providers can take steps to reduce the number of no-shows and improve patient outcomes. Additionally, in cases where multiple no-shows seem likely, providers can plan for how best utilize their expected time for other uses.

My wife is a Registered Dietitian and her most common complaint is no-shows. She has a limited number of appointments available each day and when a patient doesn't show up, it's a lost opportunity to help someone else.

Data

This dataset was found on Kaggle here https://www.kaggle.com/datasets/joniarroba/noshowappointments/data. It contains information on over 100,000 medical appointments in Brazil. The dataset includes information on patient demographics, medical history, and whether or not the patient showed up for their appointment. I am currently in contact with my wife's employer to see if I can get access to their data to see if I can use it for this project. I haven't received a final answer yet so I do not know for sure if I will be able to use this data. If I do not get it, I will use the data from Kaggle and possibly explore other avenues of finding more data.

```r
data <- read.csv("C:\\Users\\shaya\\OneDrive\\Documents\\Final Project\\Data\\KaggleV2-May-2016.csv")
kable(head(data))
```

PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No.show
2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	0	0	No
5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	0	No
4.262962e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	0	0	No
8.679512e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	0	0	No

PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No.show
8.841186e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0	0	0	No
9.598513e+13	5626772	F	2016-04-27T08:36:51Z	2016-04-29T00:00:00Z	76	REPÚBLICA	0	1	0	0	0	0	No

We can see the dataset contains the following variables: It has the patient ID, Appointment ID, their gender, the day the appointment was scheduled, appointment day, age, neighborhood, scholarship status, hypertension, diabetes, alcoholism, handicap status, SMS received as a reminder, and whether or not the patient showed up for their appointment. This dataset is collected from a hospital in Brazil and scholarship status refers to whether or not the patient is enrolled in the Bolsa Familia program which is a social welfare program in Brazil.

```
data[!complete.cases(data),] #check for incomplete rows
```

```
##  [1] PatientId      AppointmentID  Gender         ScheduledDay   AppointmentDay
##  [6] Age            Neighbourhood  Scholarship    Hipertension   Diabetes
## [11] Alcoholism     Handcap        SMS_received   No.show
## <0 rows> (or 0-length row.names)
```

```
str(data) #check for data types
```

```
## 'data.frame':    110527 obs. of  14 variables:
##  $ PatientId     : num  2.99e+13 5.59e+14 4.26e+12 8.68e+11 8.84e+12 ...
##  $ AppointmentID : int  5642903 5642503 5642549 5642828 5642494 5626772 5630279 5630575 5638447 5629
##  $ Gender        : chr  "F" "M" "F" "F" ...
##  $ ScheduledDay  : chr  "2016-04-29T18:38:08Z" "2016-04-29T16:08:27Z" "2016-04-29T16:19:04Z" "2016-04
##  $ AppointmentDay: chr  "2016-04-29T00:00:00Z" "2016-04-29T00:00:00Z" "2016-04-29T00:00:00Z" "2016-04
##  $ Age           : int  62 56 62 8 56 76 23 39 21 19 ...
##  $ Neighbourhood : chr  "JARDIM DA PENHA" "JARDIM DA PENHA" "MATA DA PRAIA" "PONTAL DE CAMBURI" ...
##  $ Scholarship   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Hipertension  : int  1 0 0 0 1 1 0 0 0 0 ...
##  $ Diabetes      : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ Alcoholism    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Handcap       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ SMS_received  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ No.show       : chr  "No" "No" "No" "No" ...
```

We can see from the above that our data is very clean and doesn't require much work. We have no missing values and all of our data types are correct. The target variable is the No-show column, which is a character type column with strings "Yes" and "No". We will need to convert this to a binary variable for our model. Sex is also a character column with "M" and "F" values. We will need to convert this to a binary variable as well. The day the appointment was scheduled and the appointment day are both character columns and we will need to convert these to date columns.

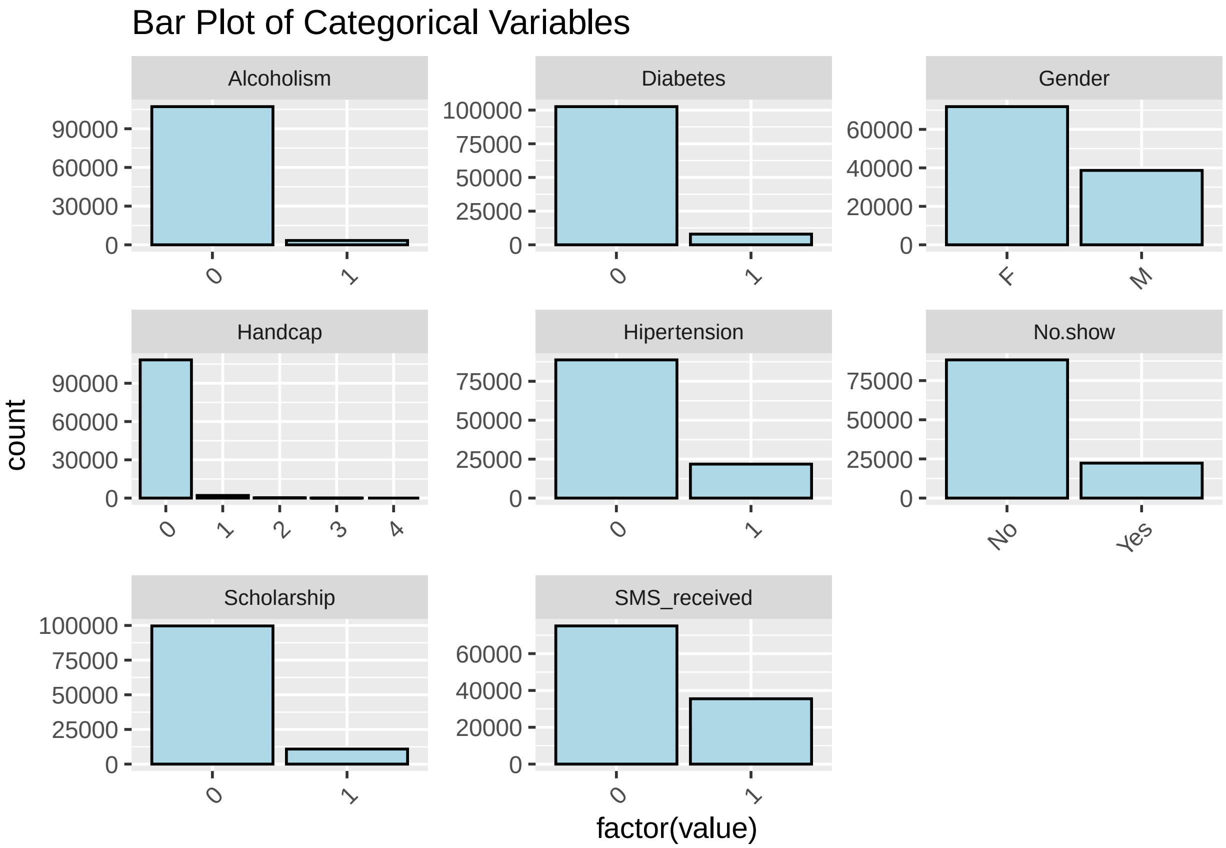
```
descr(data)
```

```
## Non-numerical variable(s) ignored: Gender, ScheduledDay, AppointmentDay, Neighbourhood, No.show
## Descriptive Statistics
## data
## N: 110527
##
##                          Age   Alcoholism   AppointmentID    Diabetes     Handcap   Hipertension
## ----------------- ----------- ------------ --------------- ----------- ----------- --------------
##              Mean       37.09         0.03      5675305.12        0.07        0.02           0.20
##           Std.Dev       23.11         0.17        71295.75        0.26        0.16           0.40
##               Min       -1.00         0.00      5030230.00        0.00        0.00           0.00
##                Q1       18.00         0.00      5640285.00        0.00        0.00           0.00
##            Median       37.00         0.00      5680573.00        0.00        0.00           0.00
##                Q3       55.00         0.00      5725524.00        0.00        0.00           0.00
##               Max      115.00         1.00      5790484.00        1.00        4.00           1.00
##               MAD       28.17         0.00        62490.11        0.00        0.00           0.00
##               IQR       37.00         0.00        85238.00        0.00        0.00           0.00
##                CV        0.62         5.65            0.01        3.59        7.26           2.02
##          Skewness        0.12         5.47           -1.24        3.32        8.27           1.52
##       SE.Skewness        0.01         0.01            0.01        0.01        0.01           0.01
##          Kurtosis       -0.95        27.93            3.74        8.99       82.55           0.32
##           N.Valid   110527.00    110527.00       110527.00   110527.00   110527.00      110527.00
##         Pct.Valid      100.00       100.00          100.00      100.00      100.00         100.00
##
## Table: Table continues below
##
##
##
##                             PatientId   Scholarship   SMS_received
## ----------------- ------------------- ------------- --------------
##              Mean  147496265710394.09          0.10           0.32
##           Std.Dev  256094920291739.06          0.30           0.47
##               Min            39217.84          0.00           0.00
##                Q1    4172457111246.00          0.00           0.00
##            Median   31731838713978.00          0.00           0.00
##                Q3   94393811856983.00          0.00           1.00
##               Max  999981631772427.00          1.00           1.00
##               MAD   45979143102074.08          0.00           0.00
##               IQR   90219106453983.00          0.00           1.00
##                CV                1.74          3.03           1.45
##          Skewness                1.97          2.70           0.77
##       SE.Skewness                0.01          0.01           0.01
##          Kurtosis                2.58          5.29          -1.41
##           N.Valid           110527.00     110527.00      110527.00
##         Pct.Valid              100.00        100.00         100.00
```

From the above output, several important observations can be made about the dataset. Firstly, the average age of patients attending appointments is approximately 37 years, with a considerable spread indicated by a standard deviation of 23 years. There appears to be an outlier in age with a minimum value of -1, which requires further investigation. The dataset includes patients with various medical conditions, such as hypertension (mean prevalence of 20%) and diabetes (7% prevalence), while alcoholism is relatively uncommon (3% prevalence). The majority of patients did not receive SMS reminders (mean proportion of 32%). The statistics also reveal extreme values in the Handcap variable, with a maximum value of 4, suggesting potential data integrity issues or varying definitions of disability levels. Further exploration and cleaning of the dataset are recommended to ensure the accuracy and reliability of subsequent analyses.

Exploratory Data Analysis

```
selected_columns <- c("Gender", "Scholarship", "Hipertension", "Diabetes", "Alcoholism", "Handcap", "SMS
data |>
  select(all_of(selected_columns)) |>
  gather(key = "variable", value = "value") |>
  ggplot(aes(x = factor(value))) +
  geom_bar(fill = 'lightblue', color = 'black') +
  facet_wrap(~ variable, scales = 'free') +
  theme(strip.text = element_text(size = 8),
       axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Bar Plot of Categorical Variables")
```

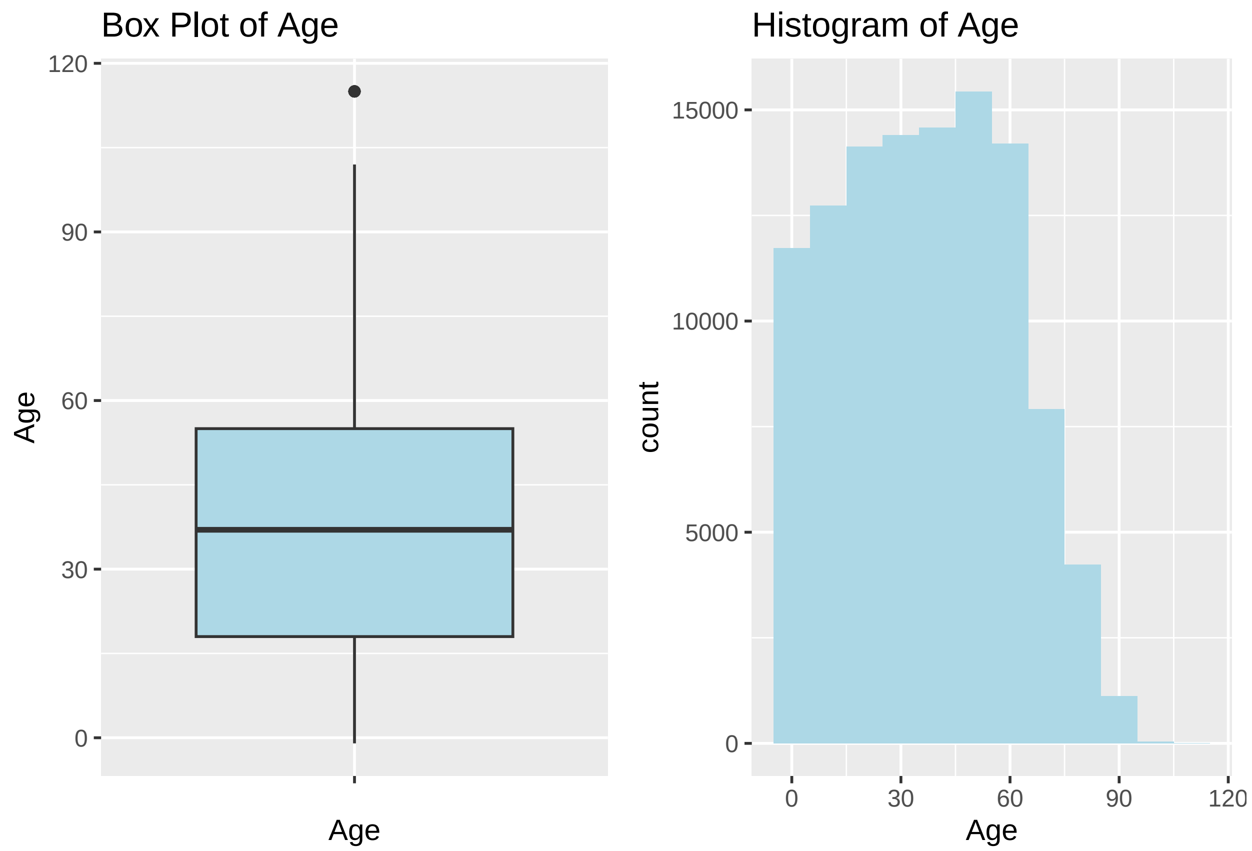


The bar plots above show the distribution of the categorical variables in the dataset. The variabls all seem to have heavy class imbalances. This will havee to be considered when building our model. It also again shows us something wrong with the Handcap variable as it has a value of 4 which should not be possible. We will need to investigate this further.

We can visualize the 'Age' variable using a box plot to identify any potential outliers or unusual patterns. We

```
age_box_plot <- data |>
  select("Age") |>
  ggplot(aes(x = "", y = Age)) +
  geom_boxplot(fill = 'lightblue') +
```

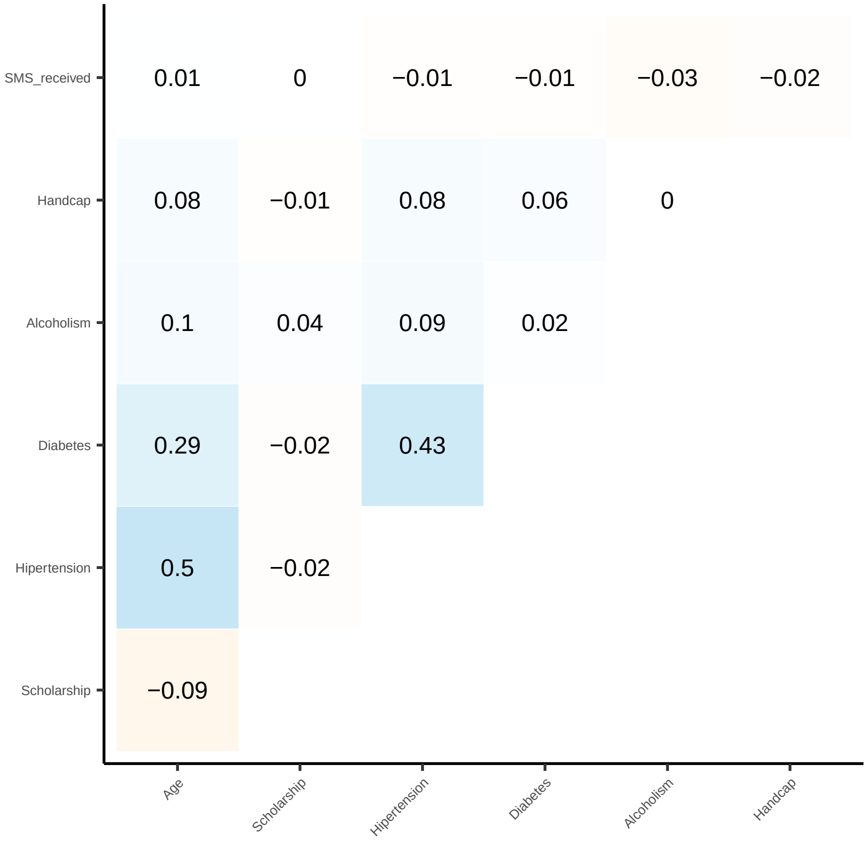
```r
  labs(title = "Box Plot of Age", x = "Age") +
  theme(strip.text = element_text(size = 8))

age_histogram <- data |>
  select("Age") |>
  ggplot(aes(x = Age)) +
  geom_histogram(binwidth = 10, fill = 'lightblue') +
  labs(title = "Histogram of Age", x = "Age") +
  theme(strip.text = element_text(size = 8))

grid.arrange(age_box_plot, age_histogram, ncol = 2)
```



The above plots show some alarming results. The box plot shows a minimum age of -1 which is not possible. The histogram shows a peak at 0 which is extremely unlikely as that would mean that a massively disproportionate number of patients are newborns. This amount of 0s is likely due to missing data and skews our summary of the mean age. We will need to investigate this further.

We will now check for correlation betweent the variables to see if these is any multicollinearity between them we need to be aware of.

```r
q <- cor(data[,c("Age", "Scholarship", "Hipertension", "Diabetes", "Alcoholism", "Handcap", "SMS_receive
ggcorrplot(q, type = "upper", outline.color = "white",
           ggtheme = theme_classic(),
           colors = c("orange", "white", "skyblue"),
           lab = TRUE, show.legend = F, tl.cex = 5, lab_size = 3)
```


SMS_received
0.01
0
−0.01
−0.01
−0.03
−0.02
Handcap
0.08
−0.01
0.08
0.06
0
Alcoholism
0.1
0.04
0.09
0.02
Diabetes
0.29
−0.02
0.43
Hipertension
0.5
−0.02
Scholarship
−0.09
Age
Scholarship
Hipertension
Diabetes
Alcoholism
Handcap