

ANALYSIS ON NO-SHOW MEDICAL APPOINTMENT ANALYSIS FOR BRAZILIAN GOVERNMENT DATA

Logistic Regression to analyze factors influencing show/drop rates of medical patients at appointments

TABLE OF CONTENTS

1 Introduction.....	2
2 Dataset Description	
2.1 About Data.....	2
2.2 Data Processing.....	3
3 Exploratory Data Analysis	
3.1 EDA on non-factor variables	4
3.2 EDA on factor variables	6
4 Logistic Regression Analysis.....	7
5 Conclusion and discussion.....	8
6 References.....	9
7 Appendix A-R Code	
7.1 Data Familiarity	
7.1.1 Data Visualize.....	10
7.1.2 Data Clean.....	11
7.2 Exploratory Analysis	
7.2.1-Age vs no show status variable.....	14
7.2.2-Gender vs No show status Variable.....	15
7.2.3-Handicap vs no show status variable.....	17
7.2.4- EDA on other factor variables.....	18
7.2.5- EDA on appointment day vs No Show.....	21
7.2.6- EDA on Neighborhood vs No Show.....	22
7.3- Logistic Regression.....	24
7.4- Logistic Regression with training and test data set.....	27
7.5- Model Evaluation.....	29

1- INTRODUCTION

A person makes doctor appointment, receives the instruction but doesn't show up at the scheduled time. There are many factors influencing this decision and we can understand why this happens in this study. The Brazilian public health system, known as SUS for Unified Health System is acronym in Portuguese, is one of the largest health systems in the world [1]. It represents government investment of more than 9% of GDP. However, its operation is not homogeneous, and quality of life has different perceptions from different regions of country concerning medical health care.

This analysis will use the no-show appointments dataset which collects information from medical appointments in Brazil and is focused on the question of whether patients show up for their appointment or not, including a set of characteristics about the patient. This analysis aims to outline some possible reasons for patient no-showing at the scheduled appointments, as well as get insights about the Brazilian public health system. Since the project aims to predict information about binary variable, such analysis would follow Logistic regression modeling technique. The dataset is extracted from Kaggle [2]. To accomplish the goal, first step is exploring the data followed by exploratory data analysis on variables in dataset, in an iterative process, in order to draw some conclusions about the subject. Some of the hypothesis to be explored are:

- What are the neighborhoods with the highest no-showing rate?
- Is the data balanced in relation to the interest class (showing/no-showing to appointments)?
- How each factor influences the no-showing variable and what is expected difference in their proportions?

2- DATASET DETAILS

2.1 About Data

The dataset on medical appointments have 110527 samples and 14 associated variables. The target variable for the study is show/no-show variable which is analyzed against other variables as listed below:

- PatientID(num)- Identification of the patient
- AppointmentID(num)- Identification of each appointment
- Gender(character)- Gender of the patient
- Scheduled Day/DataMarcacaoConsulta(datetime)- The day of the actual appointment, when they must visit the doctor.
- Appointment Day/DataAgendamento(datetime)-The day someone called or registered the appointment; this is before appointment.
- Age(num)- age of the patient
- Neighborhood(character)- The place where appointment takes place
- Scholarship(num)- Brazilian government gives this scholarship as a subsidy for very low-income families if they have, they children going to school and other more specific requirements. Further reference can be found at the link [3]

- Hypertension(num)- Whether the patient suffers from hypertension or not
- Diabetic(num)- Diabetic patient or not
- Alcoholism(num)- Alcoholic patient or not
- Handicap(num)- Handicapped levels 0-4 where 0 signifies minimum number of handicaps.
- SMS_received(num)- 1 or more text messages sent to client or not
- No-show(character)- 0(means patient showed up) or 1(means patient dint show up)

Another important information about the source of data that is relevant to this study is if this data came from a government agency or from a compiled data of each health care unit. According to data description [2], the data is about public health unit, it is not mentioned from which government area they belong to (municipal, state or federal). The absence of such information can result in a biased analysis, since the complex and decentralized Brazilian health system brings inevitable conflicts and contradictions between its associated entities, as identified in [4].

2.2 Data Processing

Head of the dataset is found and can be seen in section 7.1.1 of Appendix A. The dataset variables are renamed to more meaningful names and it is possible to notice the dataset has no missing values, since all the attributes have 110527 instances. Level 'No' of no-show variable is renamed to Showed Up and 'Yes' to not showed up as this is important to the study. A new column called day is created which would help analyze how is the show/no-show variable dependent on weekday and weekend appointment. Certain character variable such as gender, neighborhood, scholarship, hypertension, diabetes, alcoholism, handicap, SMS_received is converted into factor type for easier interpretation. It is not possible to have -1 age or more than 100 years. This is analyzed appropriately at the exploratory data analysis section, but ages under zero are a mistake hence removed. The corresponding code is given in appendix section 7.1.2 of Appendix A. Also, a final snapshot of data can be seen in Figure 1 below.

```

Min. 1st Qu. Median Mean 3rd Qu. Max.
0.00 18.00 37.00 37.09 55.00 115.00
Classes 'tbl_df', 'tbl' and 'data.frame': 110526 obs. of 17 variables:
$ patient_id : num 2.99e+13 5.59e+14 4.26e+12 8.68e+11 8.84e+12 ...
$ appointment_id : num 5642903 5642503 5642549 5642828 5642494 ...
$ gender : Factor w/ 2 levels "Female", "Male": 1 2 1 1 1 1 1 1 ...
$ schedule_day : POSIXct, format: "2016-04-29 18:38:08" "2016-04-29 16:08:27" "2016-04-29 16:19:04" "2016-04-29 17:29:31" ...
$ appointment_day: POSIXct, format: "2016-04-29" "2016-04-29" "2016-04-29" "2016-04-29" ...
$ age : num 62 56 62 8 56 76 23 39 21 19 ...
$ neighborhood : Factor w/ 81 levels "AEROPORTO", "ANDORINHAS", ...: 40 40 47 55 40 59 26 26 2 13 ...
$ scholarship : Factor w/ 2 levels "0", "1": 1 1 1 1 1 1 1 1 1 ...
$ hypertension : Factor w/ 2 levels "0", "1": 2 1 1 1 2 2 1 1 1 1 ...
$ diabetes : Factor w/ 2 levels "0", "1": 1 1 1 1 2 1 1 1 1 1 ...
$ alcoholism : Factor w/ 2 levels "0", "1": 1 1 1 1 1 1 1 1 1 1 ...
$ handicap : Factor w/ 5 levels "0", "1", "2", "3", ...: 1 1 1 1 1 1 1 1 1 1 ...
$ sms_received : Factor w/ 2 levels "0", "1": 1 1 1 1 1 1 1 1 1 1 ...
$ no_show_status: Factor w/ 2 levels "not showed up", ...: 2 2 2 2 2 2 1 1 2 2 ...
$ day : chr "Friday" "Friday" "Friday" ...
$ date : Date, format: "2016-04-29" "2016-04-29" "2016-04-29" "2016-04-29" ...
$ days : Ord.factor w/ 7 levels "Sun" < "Mon" < "Tue" < ...: 6 6 6 6 6 6 6 ...

```

Figure 1: Final Structure of Data before starting EDA

3- EXPLORATORY DATA ANALYSIS (EDA)

3.1 EDA on non-factor variables

The study started with analyzing data about age factor and it was interesting to note data has all samples from all ages but primarily patients range between 5-60 years. Initially, proportion of patients showing and not showing is plotted along with boxplot to determine if there is any difference in mean age of patients from both categories. The plot as shown in Figure 2 clearly depicts this difference and it was found statistically mean age of patients who showed up for appointments is 34 and that of not showing up is 37. This suggested some significance for target variable hence T-test was conducted to verify the results and similar readings were obtained as show in section 7.2.1 of Appendix A. Histogram for proportion of showed up patients highlighted another point that number of patients showing up for medical appointments are higher in the age 60 to 80 as compared to patient age under 40. Section 7.2.1 of Appendix A covers all these details

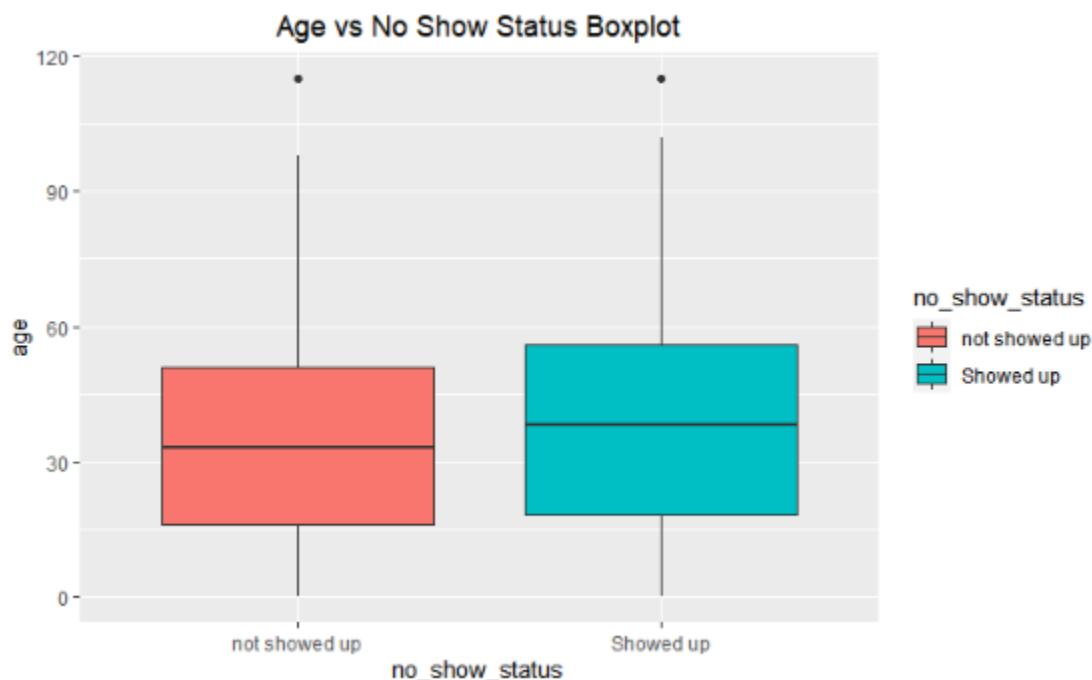


Figure 2: Age distribution for patients showing difference in mean age between show and no-show appointments

Next factor analyzed was gender in reference to section 7.2.2. It is evident from same section that nearly 25% of patients don't show up for their appointments. There are more samples from female patients as compared to male patients hence their corresponding proportion was calculated to determine no-show rate. This also suggested females tend to have significantly more medical issues in Brazil. The percentage of females showing up for their appointment is 79% which is almost same as percentage of male patients showing up. For better visualization of data, proportion of no-show rate was plotted which gave similar results as expected and can be referred in section 7.2.2. T-test suggested similar results and hence the variable was found not significant for target variable. Handicap levels is another interesting variable which as per Brazilian government has 0-4 levels where 0 signifies least number of handicaps. Generally one

can expect a patient having higher number of handicaps would show up for appointments regularly but on analyzing the data as in section 7.2.3 of Appendix A there was not much difference in proportion of no-show rates for handicapped category and majority of patients were found to have almost no handicap. Pearson Chi squared test was conducted to verify significance of this variable and on contrary to expectation it was not influencing the target variable and proportion distribution for handicap variable was found to be even as given in Figure 3 below.

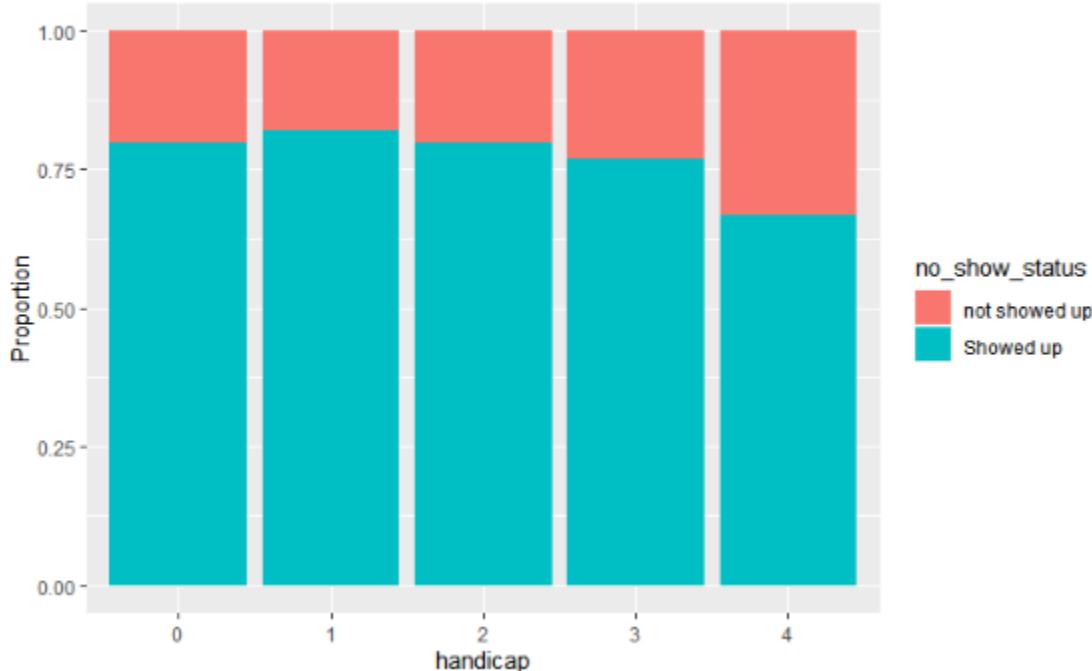


Figure 3: Proportion of no-show rate for handicap level 0-4 across Brazilian patients in public health care

Section 7.2.5 in Appendix A of study reveals number of appointments varies across weekdays and weekends. As mentioned earlier in report a new factor called 'day' was formed from scheduled date factor which revealed significantly lesser medical appointments on weekends. Maximum number of appointments are fixed on Tuesday and Wednesday while not much appointments are made on Saturday and around 20% of appointments are not successful for each day on average. Ordered factor days is also created to visualize no-show rates for each day of week and their relative proportion show that day is significant in influencing the decision of not showing up for interview. Statistical test confirmed this hypothesis as well. Neighborhood could be considered another factor based on which mostly patients tend to miss their meeting. To determine this hypothesis a table depicting no-show rates for each one of them was formed and clearly large difference was visible for different region. Based on this table as in section 7.2.6 of Appendix A, a proportion graph was plotted to figure which areas had maximum no-show count. As noted in same section Ilhas Oceanicas De Trindade region has maximum no-show counts of patients which as per data is 99% which tells none of the patients in this area show up on their schedules interview. Following the similar case is Santos Dumont area having 75% rate but interestingly areas such as Maria Ortiz and Jardim Camburi are observed to have more than

1000 failed medical appointments. Most of the neighborhoods have no show rates ranging between 20-25% but there was no single area that has 100% successful appointments.

3.2 EDA on factor variables

Usually the most convenient form of information about patients which depict his health state form most significant basis in determining the seriousness towards medical issues. Factor like diabetes, hypertension and alcoholism are indicators of this hypothesis. Other factors like having scholarship for medical appointment and text message received would also help determine respective rates. This is in reference to section 7.2.4 in Appendix A of the study. The patients showing up in public health care in Brazil show only 10-12% have scholarships, 5% are alcoholic and less than 25% have diabetes and hypertension indicating good health conditions of country. The proportion of patients showing and not showing was plotted for each category as shown in Figure 4 below.

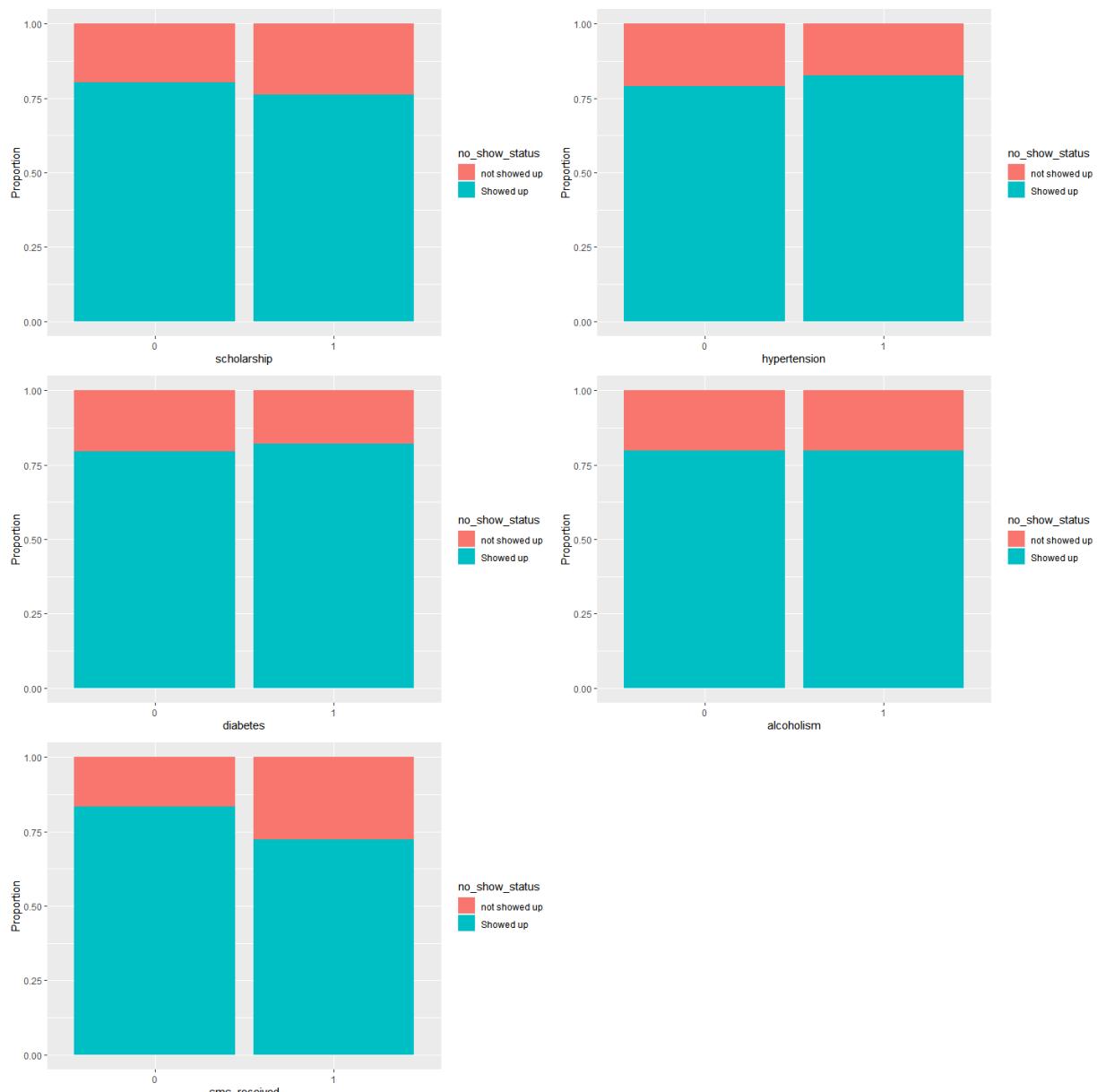


Figure 4: Proportion of no-show rate for factors such as diabetes, alcoholism, hypertension, SMS and scholarship.

Contrary to general hypothesis it was evident from the individual plots that only alcoholism, diabetes and hypertension affect very less in determining the cause of the study. For verification Pearson Chi Squared test were performed as shown in section 7.2.4 in Appendix A. Significantly lower p-values were found for all these factors except alcoholism which is clearly showing no dependence with no-show status of patient. These forms important piece of information in determining modeling factors and we should expect similar results from model selection approaches as well. There was interesting point to note that in cases where patients received SMS alert more than 25% failed to show up still. Though on average for irrespective of the medical factors and scholarship available around 20-25% patients still don't come for their appointment.

4- LOGISTIC REGRESSION ANALYSIS

The hypothesis we aimed to answer were covered in EDA section of study which also gave enough information required to start logistic modeling from data to determine significant factors in determining no-show rate of medical appointments for Brazilian government. Scheduled day and appointment day are initially not taken into consideration based on assumption these do not influence the target variable. Logistic model is fit, and summary was obtained. The sign of coefficients for each factor was clear indicator of how they vary w.r.t no_show_status variable. We can observe negative coefficients for scholarship true, diabetes, alcoholism, handicap, SMS_received, weekend day Saturday indicating number of no shows of patients increases if these are true for patient pointing to the fact that more is their value or if present in patient data would not tend to show up for appointment. This further confirms our EDA on these factors individually. While for male gender, weekdays, hypertension and age factors number if this were true or more generally indicate showing up more for same appointments. These results are in reference to section 7.3 from Appendix A. Coefficients from same summary indicate that the odds of showing up for appointment reduces by 13.4% on Saturday keeping all other factors constant like age, gender, scholarship yes or no, hypertension, diabetes and alcoholism. If the person has hypertension the odds of showing up for appointment increases by 6.9% if patient keeping all other factors fixed. This seems reasonable if patient already has some medical ailment , he tends to show up for appointment. Significance of each predictor was determined using Drop1 function corresponding to same section which gave significantly higher values for gender and handicap variable hence suggesting them to be dropped from further analysis. Another logistic model was fit after removing these variables and following the summary for p-values all were found to be significant to be included in model. The data was divided into training and testing set for better accuracy in ratio 7:3. Dataset obtained from after removing predictors not relevant to study was modeled using logistic regression shown in section 7.4 of Appendix A. AIC and VIF functions were used to determine appropriate model selection technique and if any factors were causing inflation. Tuesday, Wednesday and Hypertension had high VIF which was more than 10 indicating they should be removed from further analysis. Final model obtained after AIC was conducted only had age, scholarship, SMS-received and day factor. The sign of coefficients indicates only on Saturdays, patients having scholarship and those receiving SMS tend to show up less for scheduled appointment and patients having appointment on weekdays tend to show more. Target variable is not dependent on other factors hence cannot be judged while predicting the accuracy, sensitivity or AUC metric. The final model had class imbalance

between showed and not showed up group though section 7.5 of Appendix A reveals that 80% of patients in Brazil public health cares attend their medical appointment and rest fail to show up on decided date. Normal QQ plot is also plotted for data but no linear relationship is observed and there are huge clusters of corresponding to target value of Yes or No which can be clearly seen in same section. Final confusion matrix figure for the model is shown in Table 1 below.

	Actual No show	Actual Show
Predicted No show	6696	26462
Predicted Show	0	0

Table 1: Confusion Matrix for Logistic Regression for no show medical appointment data

5- CONCLUSION AND DISCUSSION

The study has answered the hypothesis which we aimed to solve, and we can arrive at conclusions based on analysis. This analysis sought to explore the medical appointment no-show's dataset, containing more than 100000 appointments and of which approximately 30% of the patients have not attended. There is more female data present which indicate greater medical issues concerning women of the related regions. There are considerably significant number of appointments for every age group and not specific to old people. There is very less data available for weekends which tend to indicate health cares are booked mostly on weekdays and one can assume weekends are mostly dedicated to emergency cases. There was another interesting data revelation about text messages sent to patients which were half of total number of cases. This might indicate preference to illness, prepaid appointments or availability of contact. In this way, the data exploration presented in this file had the purpose to show up some insights on the possible causes of this appointment missing. It is important to notice, though, that with the provided data and the applied methods it is not possible to affirm any statement with statistical validity nor causal relations. On the other hand, the insights obtained through this analysis can, instead, inform and direct in-depth research on the subject, in order to validate the causes of the appointment missing and find ways to eradicate them. Various improvements can be done in data collection like specific illness case, cost of medical treatment, private unit information, doctor qualification, address of patient to improve modeling on no-show rate for these appointments. Managing unbalanced data would require under and over sampling which is beyond the scope of this study but can be taken up as well.

6- REFERENCES

- [1] Wikipedia contributors. Brazil. Wikipedia, The Free Encyclopedia. February 9, 2018, 22:06 UTC. Available at: <https://en.wikipedia.org/w/index.php?title=Brazil&oldid=824851232>. Accessed in February 10, 2018.
- [2] Kaggle. *Medical Appointment No Shows*. Reviewed dataset. Available at: <https://www.kaggle.com/joniarroba/noshowappointments>. Accessed in February 10, 2018.
- [3] Wikipedia contributors. Brazil. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Bolsa_Fam%C3%ADlia
- [4] Victora, C. G., Barreto, M. L., Leal, M. D. C., Monteiro, C. A., Schmidt, M. I., Paim, J. S., ... & Reichenheim, M. E. (2011). *Condições de saúde e inovações nas políticas de saúde no Brasil: o caminho a percorrer* (in Portuguese). Available at: <https://repositorio.ufba.br/ri/handle/ri/5973>. Accessed in February 12, 2018.

7- APPENDIX A (R-CODE)

Statistics_Final

Jasleen Kaur

03/04/2020

```
options(warn=-1)
library(tidyverse)

## -- Attaching packages ----

## v ggplot2 3.3.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr    0.8.3
## v tidyverse 1.0.0    v stringr  1.4.0
## v readr   1.3.1     vforcats  0.4.0

## -- Conflicts -----
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()   masks stats::lag()
```

What percentage of patients miss their medical appointment and why?

7.1-Data Familiarity

```
# read the table
med_data <- read_csv("KaggleV2-May-2016.csv")

## Parsed with column specification:
## cols(
##   PatientId = col_double(),
##   AppointmentID = col_double(),
##   Gender = col_character(),
##   ScheduledDay = col_datetime(format = ""),
##   AppointmentDay = col_datetime(format = ""),
##   Age = col_double(),
##   Neighbourhood = col_character(),
##   Scholarship = col_double(),
##   Hipertension = col_double(),
##   Diabetes = col_double(),
##   Alcoholism = col_double(),
##   Handcap = col_double(),
##   SMS_received = col_double(),
##   `No-show` = col_character()
## )
```

7.1.1 Data Visualize

```

str(med_data)

## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and '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  : num  5642903 5642503 5642549 5642828 5642494 ...
## $ Gender         : chr  "F" "M" "F" "F" ...
## $ ScheduledDay   : POSIXct, format: "2016-04-29 18:38:08" "2016-04-29 16:08:27" ...
## $ AppointmentDay: POSIXct, format: "2016-04-29" "2016-04-29" ...
## $ Age            : num  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     : num  0 0 0 0 0 0 0 0 0 0 ...
## $ Hipertension    : num  1 0 0 0 1 1 0 0 0 0 ...
## $ Diabetes        : num  0 0 0 0 1 0 0 0 0 0 ...
## $ Alcoholism      : num  0 0 0 0 0 0 0 0 0 0 ...
## $ Handcap         : num  0 0 0 0 0 0 0 0 0 0 ...
## $ SMS_received    : num  0 0 0 0 0 0 0 0 0 0 ...
## $ No-show         : chr  "No" "No" "No" "No" ...
## - attr(*, "spec")=
##   .. cols(
##     .. PatientId = col_double(),
##     .. AppointmentID = col_double(),
##     .. Gender = col_character(),
##     .. ScheduledDay = col_datetime(format = ""),
##     .. AppointmentDay = col_datetime(format = ""),
##     .. Age = col_double(),
##     .. Neighbourhood = col_character(),
##     .. Scholarship = col_double(),
##     .. Hipertension = col_double(),
##     .. Diabetes = col_double(),
##     .. Alcoholism = col_double(),
##     .. Handcap = col_double(),
##     .. SMS_received = col_double(),
##     .. `No-show` = col_character()
##   .. )

```

7.1.2 Data Clean

```

# changing columns name
names(med_data)<- c('patient_id','appointment_id','gender','schedule_day',
                      'appointment_day','age',
                      'neighborhood','scholarship','hypertension','diabetes',
                      'alcoholism','handicap',
                      'sms_received','no_show_status')

# In data description no_show:'No' means patients showed up and 'yes'
#means patients dint show up . Hence giving clear names
med_data$no_show_status[med_data$no_show_status == 'No'] <- 'Showed up'
med_data$no_show_status[med_data$no_show_status == 'Yes'] <- 'not showed up'
med_data$no_show_status <- as.factor(med_data$no_show_status)

# replace names for gender levels
med_data$gender[med_data$gender == 'M'] <- "Male"
med_data$gender[med_data$gender == 'F'] <- "Female"

```

```

med_data$day <- weekdays(as.Date(med_data$appointment_day))

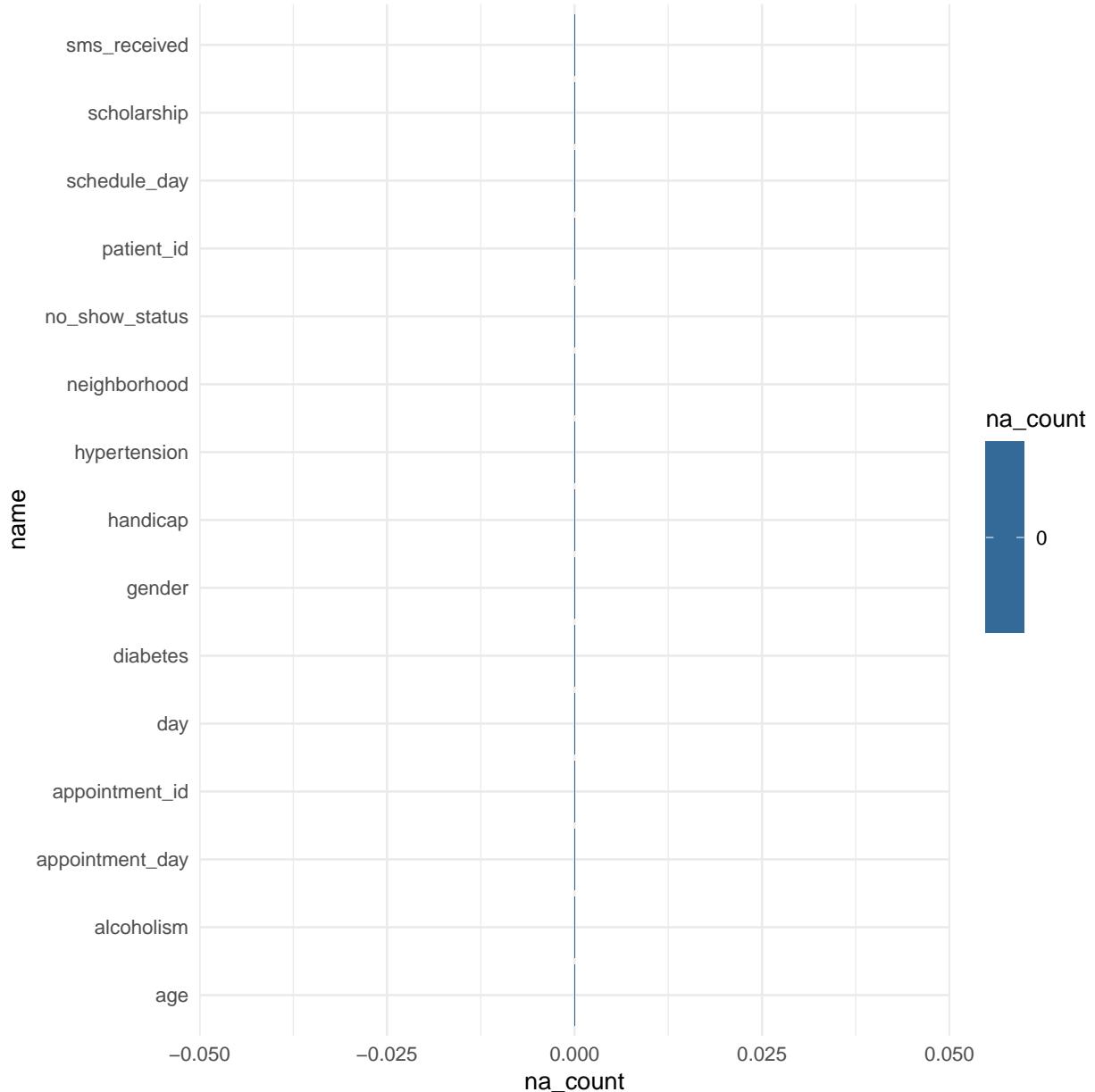
# Check missing values
na_count <- sapply(med_data, function(y) sum(length(which(is.na(y)))))

na_count <- data.frame(na_count)
na_count$name<-rownames(na_count)

# Visualize NA values in data frame
p<- ggplot(data=na_count, aes(x=name,y=na_count,fill=na_count))+  

  geom_bar(stat="identity")+
  theme_minimal()
p + coord_flip()

```



no NA values found in dataset

```
# converting some columns to factors
med_data <- mutate_at(med_data, vars('gender',
  'neighborhood',
  'scholarship',
  'hypertension',
  'diabetes',
  'alcoholism',
  'handicap',
  'sms_received'), as.factor)
```

```
# checking summary
summary(med_data)
```

```
##      patient_id      appointment_id      gender
##  Min.   :3.922e+04   Min.   :5030230   Female:71840
##  1st Qu.:4.173e+12   1st Qu.:5640286   Male   :38687
##  Median :3.173e+13   Median :5680573
##  Mean   :1.475e+14   Mean   :5675305
##  3rd Qu.:9.439e+13   3rd Qu.:5725524
##  Max.   :1.000e+15   Max.   :5790484
##
##      schedule_day      appointment_day
##  Min.   :2015-11-10 07:13:56   Min.   :2016-04-29 00:00:00
##  1st Qu.:2016-04-29 10:27:01   1st Qu.:2016-05-09 00:00:00
##  Median :2016-05-10 12:13:17   Median :2016-05-18 00:00:00
##  Mean   :2016-05-09 07:49:15   Mean   :2016-05-19 00:57:50
##  3rd Qu.:2016-05-20 11:18:37   3rd Qu.:2016-05-31 00:00:00
##  Max.   :2016-06-08 20:07:23   Max.   :2016-06-08 00:00:00
##
##      age          neighborhood      scholarship hypertension
##  Min.   :-1.00      JARDIM CAMBURI : 7717   0:99666   0:88726
##  1st Qu.: 18.00      MARIA ORTIZ     : 5805   1:10861   1:21801
##  Median : 37.00      RESISTÊNCIA   : 4431
##  Mean   : 37.09      JARDIM DA PENHA: 3877
##  3rd Qu.: 55.00      ITARARÉ       : 3514
##  Max.   :115.00      CENTRO        : 3334
##                  (Other)        :81849
##      diabetes      alcoholism handicap sms_received      no_show_status
##  0:102584   0:107167  0:108286  0:75045      not showed up:22319
##  1: 7943    1: 3360   1: 2042   1:35482      Showed up   :88208
##                  2:    183
##                  3:     13
##                  4:     3
##
##      day
##  Length:110527
##  Class :character
##  Mode  :character
##
##
```

```

##  

##  

# removing age = -1  

med_data<-med_data[!(med_data$age<0),]  

summary(med_data$age)

##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.  

##      0.00   18.00  37.00   37.09  55.00 115.00

```

7.2-Exploratory Analysis

7.2.1-Age vs no show status variable

```

# Visualize age vs no show status variable
library(ggplot2)

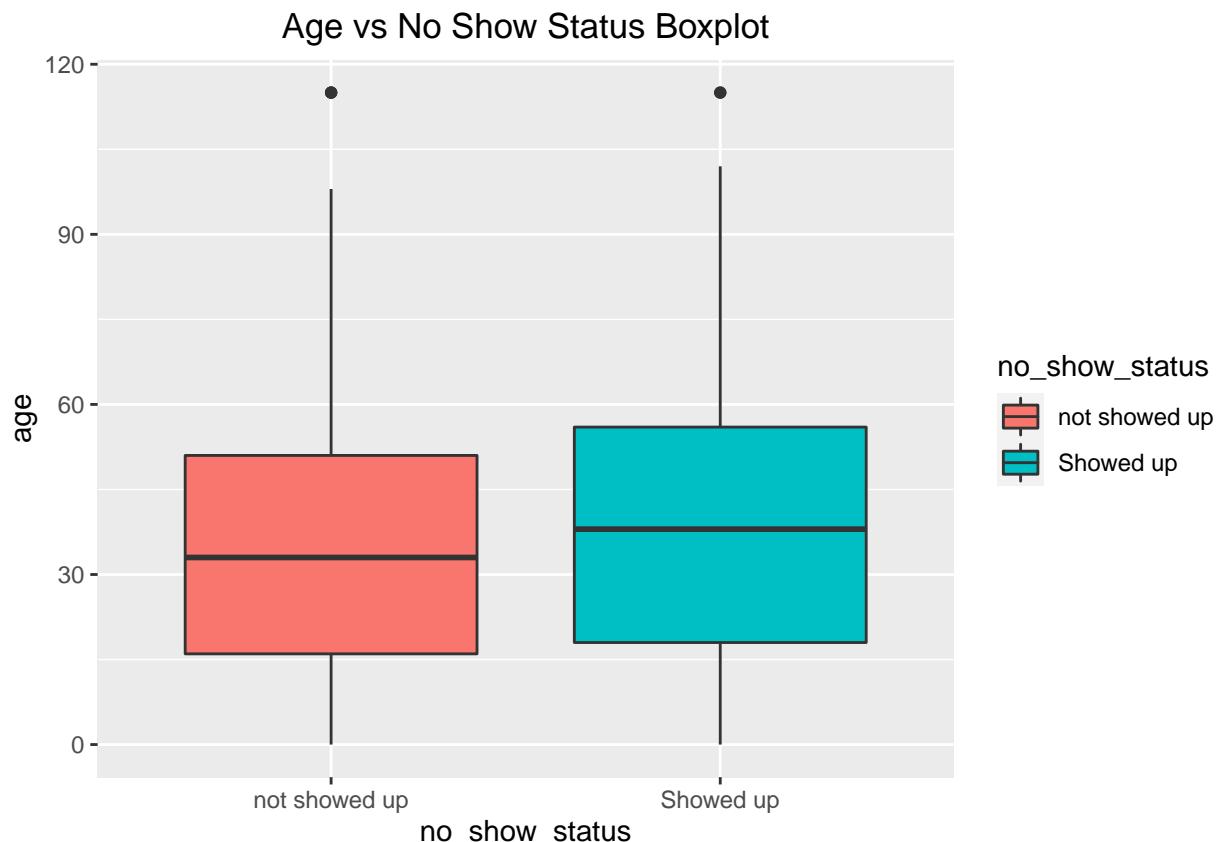
ggplot(med_data, aes(x = no_show_status, y = age, fill = no_show_status))+  

  geom_boxplot() +  

  ggtitle("Age vs No Show Status Boxplot") +  

  theme(plot.title = element_text(hjust = 0.5))

```



- Number of patients showing up is less for all age groups

- From age vs no show status histogram proportion of showed up patients are higher in the age range 60 to 80 than patient age under 40
- Mean age of patients not showing up for treatment is less than mean age of patients who showed up for treatment

```
# Exact mean value statistics of ages for show/no show
dplyr::select(med_data, age, no_show_status) %>%
  group_by(no_show_status) %>%
  summarise(age_mean = mean(age))
```

```
## # A tibble: 2 x 2
##   no_show_status age_mean
##   <fct>          <dbl>
## 1 not showed up     34.3
## 2 Showed up        37.8
```

- This signifies there is difference in age of patients showing up and not showing up

```
# Conducting test for significance of age factor; as age is numeric and
# no_show_status is a binary factor;
# independent 2 group t-test
# H0: True difference in means of show age and no show age = zero; age factor
# is not significant
# in determining show,no show factor in medical appointments
t.test(med_data$age ~ med_data$no_show_status)
```

```
##
##  Welch Two Sample t-test
##
## data: med_data$age by med_data$no_show_status
## t = -20.831, df = 36143, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -3.799602 -3.146073
## sample estimates:
## mean in group not showed up      mean in group Showed up
##                      34.31767           37.79050
```

- There is significant difference in mean age of showed and not showed up patients suggesting age is significant factor in determining medical appointment show status.

7.2.2-Gender vs No show status Variable

```
# total patient table
cat("PATIENT TABLE")

## PATIENT TABLE

table(med_data$no_show_status)
```

```

##  

## not showed up      Showed up  

##          22319       88207  

cat("\n\n\n","GENDER TABLE\n")  

##  

##  

##  

##  GENDER TABLE  

# gender table  

table(med_data$gender, med_data$no_show_status)  

##  

##          not showed up Showed up  

##  Female        14594     57245  

##  Male          7725      30962  

cat("\n\n\n","GENDER PROPORTION TABLE\n")  

##  

##  

##  

##  GENDER PROPORTION TABLE  

# proportion table for gender  

prop.table(table(med_data$gender, med_data$no_show_status), margin = 1)  

##  

##          not showed up Showed up  

##  Female      0.2031487 0.7968513  

##  Male        0.1996795 0.8003205

```

- Nearly 25% of patients dont show for scheduled medical appointment
- It seems no significant difference in gender showing or not showing for medical appointment.

```

# Statistical test for significance of gender  

# H0: gender factor is not significant in predicting show/no show factor  

chisq.test(table(med_data$gender,med_data$no_show_status), correct = FALSE)# without continuity correct

```

```

##  

##  Pearson's Chi-squared test  

##  

##  data:  table(med_data$gender, med_data$no_show_status)  

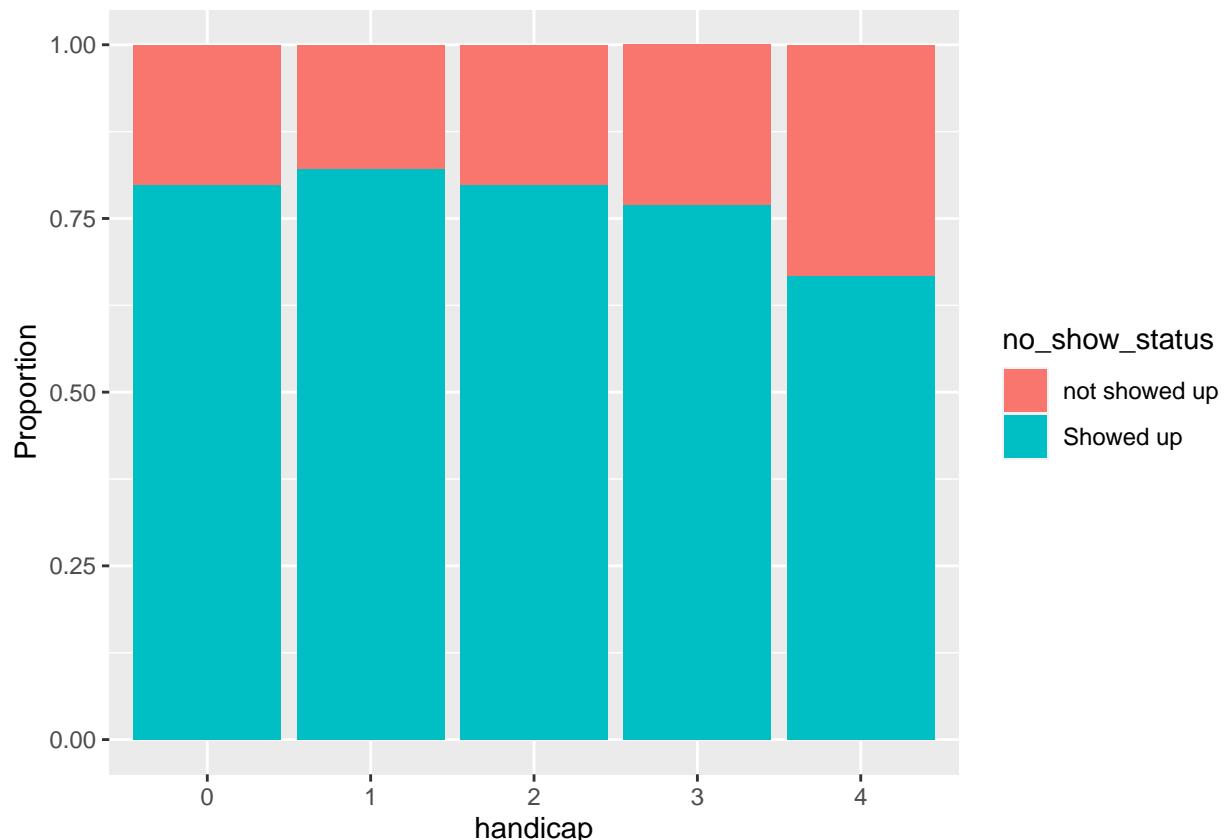
##  X-squared = 1.8779, df = 1, p-value = 0.1706

```

- From the Pearson's chi squared test p-value is higher than alpha value(~0.05) suggesting gender factor is not significant in predicting show no show of patients.

7.2.3-Handicap vs no show status variable

```
# Visualizing proportion of show/no-show for each level of handicap
ggplot(med_data)+geom_bar(aes( handicap, fill = no_show_status), position = position_fill())+
  ylab('Proportion')
```



- Proportion graph for handicap variable suggests show/no-show factor is not really dependent on it . Furthur we can verify it using chi squared test

```
# Pearson Chi squared test: H0- handicap factor is not significant in determining
# show/no-show status for medical appointment
chisq.test(table(med_data$handicap, med_data$no_show_status), correct = FALSE) #handicap
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(med_data$handicap, med_data$no_show_status)
## X-squared = 7.0356, df = 4, p-value = 0.134
```

-pvalue for handicap factor is more than alpha(~0.05) suggesting handicap is not significant in determining target factor.

7.2.4- EDA on other factor variables

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

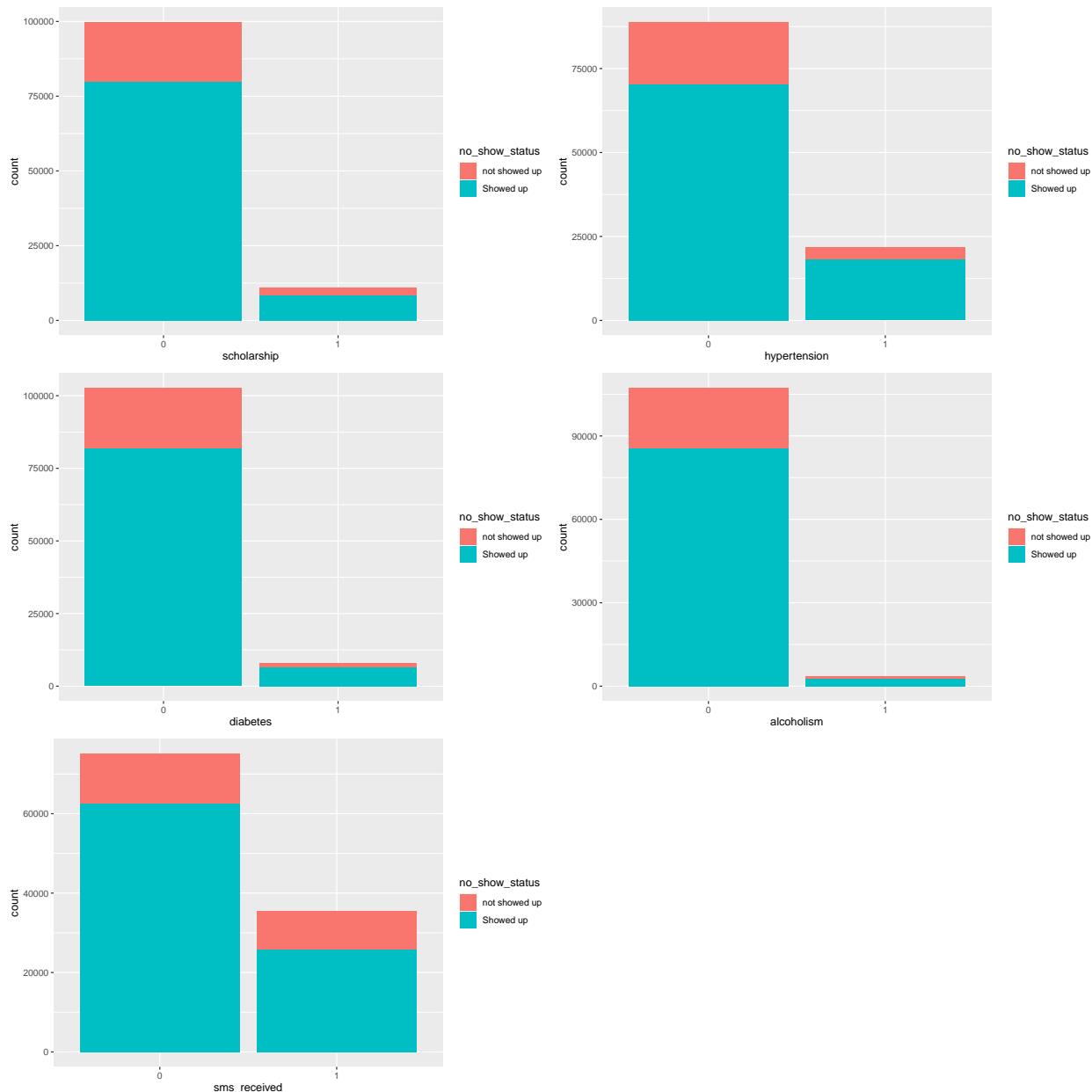
```

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

# Visualizing significance of other factors for show/no show target factor
g1 <- ggplot(med_data)+geom_bar(aes(scholarship, fill = no_show_status))
g2 <- ggplot(med_data)+geom_bar(aes(hypertension, fill =no_show_status))
g3 <- ggplot(med_data)+geom_bar(aes(diabetes, fill = no_show_status))
g4 <- ggplot(med_data)+geom_bar(aes (alcoholism, fill =  no_show_status))
g5 <- ggplot(med_data)+geom_bar(aes(sms_received, fill = no_show_status))

grid.arrange(g1,g2,g3,g4,g5, nrow = 3)

```



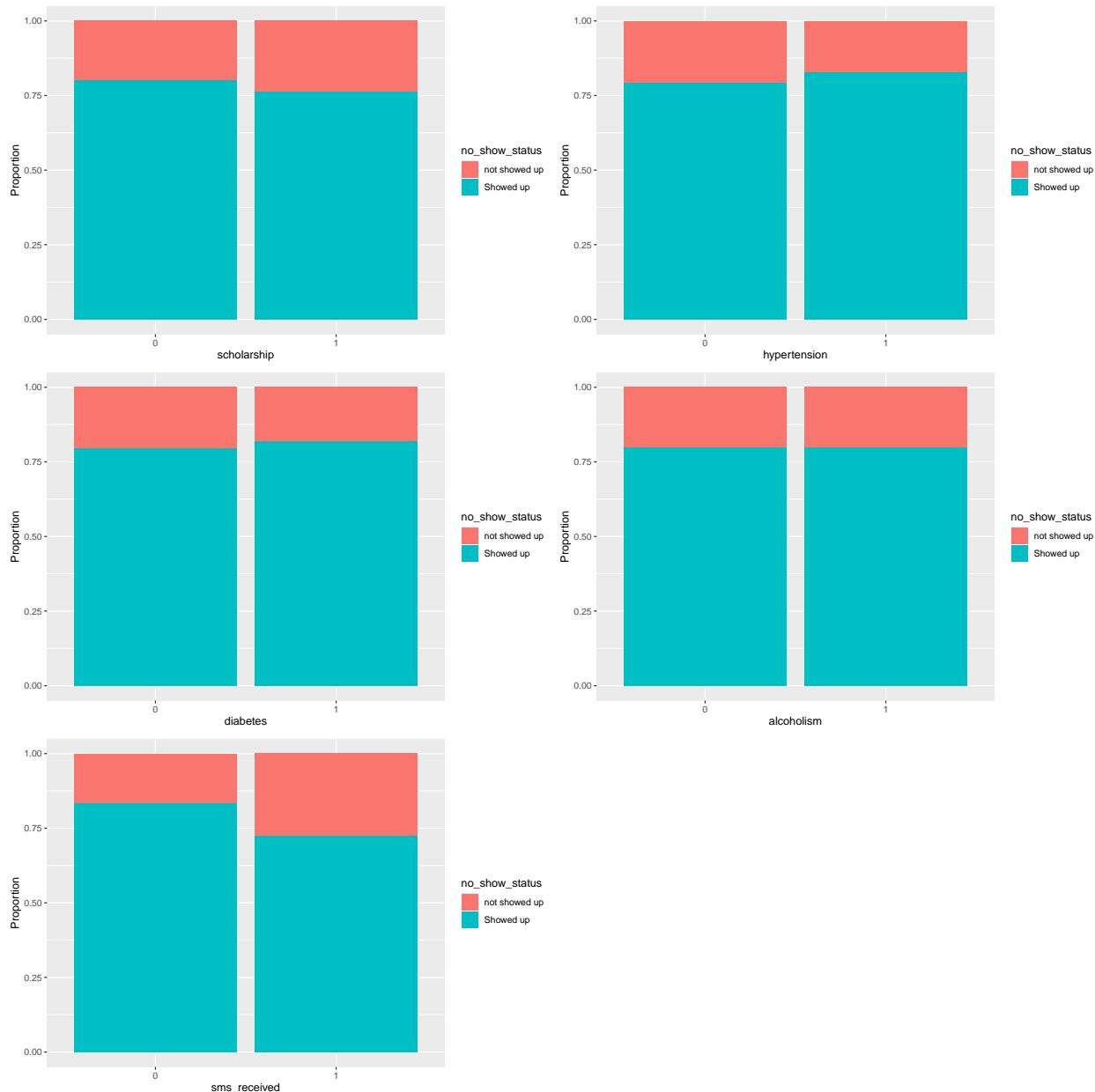
- The levels for each factor variables are not clearly depicting the proportion of show/no-show so finding their proportions

```

gg1 <- ggplot(med_data)+geom_bar(aes(scholarship, fill = no_show_status),
                                , position = position_fill())+ ylab('Proportion')
gg2 <- ggplot(med_data)+geom_bar(aes(hypertension, fill =no_show_status),
                                , position = position_fill())+ylab('Proportion')
gg3 <- ggplot(med_data)+geom_bar(aes(diabetes, fill = no_show_status),
                                , position = position_fill())+ ylab('Proportion')
gg4 <- ggplot(med_data)+geom_bar(aes (alcoholism, fill = no_show_status),
                                , position = position_fill())+ylab('Proportion')
gg5 <- ggplot(med_data)+geom_bar(aes(sms_received, fill = no_show_status),
                                , position = position_fill())+ ylab('Proportion')

grid.arrange(gg1,gg2,gg3,gg4,gg5, nrow = 3)

```



- From these plots except alcholic factor all others seems to have significant effect on show/no show status. We can determine significance using chi squared for each one of them

```
# H0: Factor is significant in determining show/no-show for patient
chisq.test(table(med_data$scholarship, med_data$no_show_status), correct = FALSE) # scholarship

## 
## Pearson's Chi-squared test
##
## data: table(med_data$scholarship, med_data$no_show_status)
## X-squared = 93.811, df = 1, p-value < 2.2e-16

chisq.test(table(med_data$hypertension, med_data$no_show_status), correct = FALSE) # hypertension

## 
## Pearson's Chi-squared test
##
## data: table(med_data$hypertension, med_data$no_show_status)
## X-squared = 140.89, df = 1, p-value < 2.2e-16

chisq.test(table(med_data$diabetes, med_data$no_show_status), correct = FALSE) # diabetes

## 
## Pearson's Chi-squared test
##
## data: table(med_data$diabetes, med_data$no_show_status)
## X-squared = 25.473, df = 1, p-value = 4.486e-07

chisq.test(table(med_data$alcoholism, med_data$no_show_status), correct = FALSE) #alcoholism

## 
## Pearson's Chi-squared test
##
## data: table(med_data$alcoholism, med_data$no_show_status)
## X-squared = 0.0042829, df = 1, p-value = 0.9478

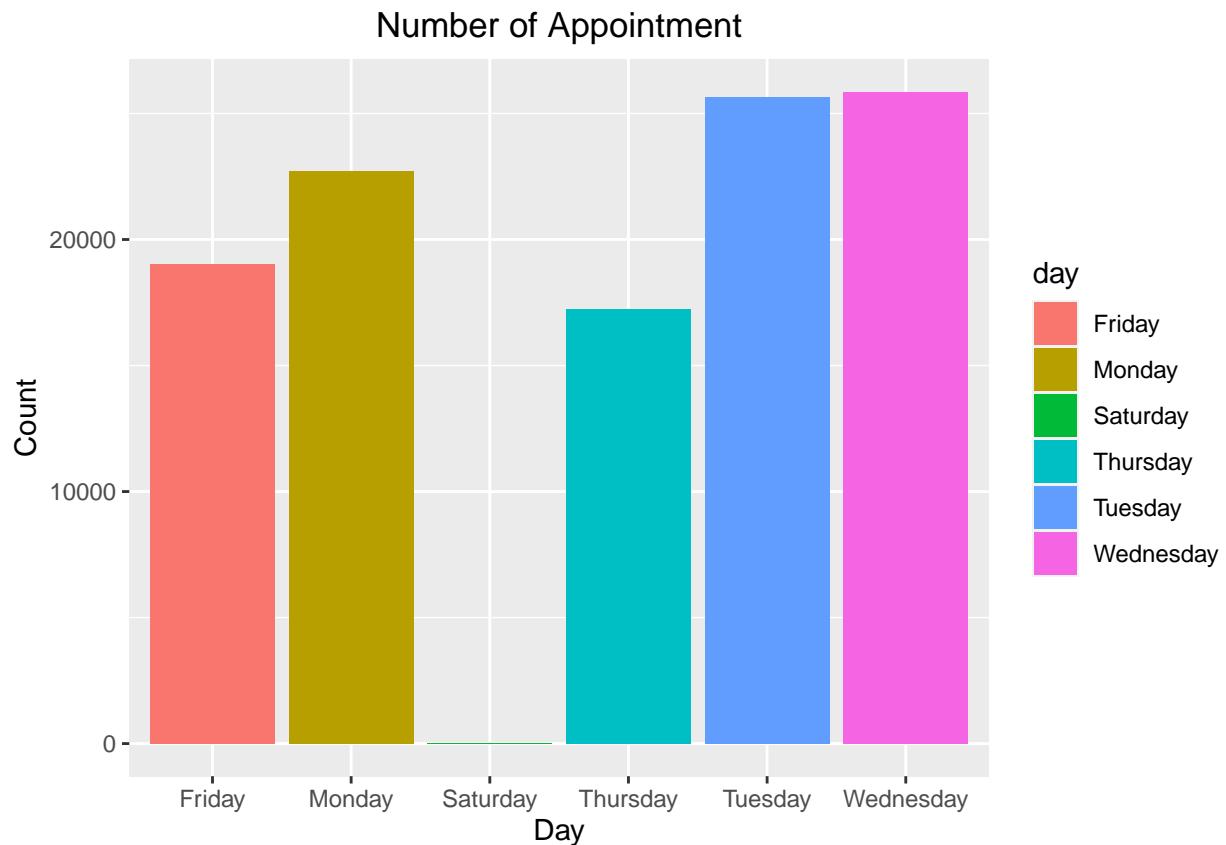
chisq.test(table(med_data$sms_received, med_data$no_show_status), correct = FALSE) # sms_received

## 
## Pearson's Chi-squared test
##
## data: table(med_data$sms_received, med_data$no_show_status)
## X-squared = 1766.7, df = 1, p-value < 2.2e-16
```

- p-values from chi squared test suggest that cholarship, hypertension, diabetes, sms_received group are significant in determining show/no-show for medical appointment and as proportion graphs suggested alcohol factor is not significant

7.2.5- EDA on appointment day vs No Show

```
ggplot(med_data)+geom_bar(aes(day, fill = day))+  
  ggtitle("Number of Appointment") +  
  ylab('Count') +  
  xlab('Day') +  
  theme(plot.title = element_text(hjust = 0.5))
```



-Number of appointment differ across week. Some day like Wednesday and Tuesday make more appointment than other. Statistics give exact information below.

```
library(lubridate)

##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##      date

# make days column, with label true
med_data$date <- as.Date(med_data$appointment_day)
med_data$days <- wday(med_data$date, label=TRUE)

# days column
table(med_data$days, med_data$no_show_status)
```

```

## not showed up Showed up
## Sun 0 0
## Mon 4690 18024
## Tue 5152 20488
## Wed 5093 20774
## Thu 3338 13909
## Fri 4037 14982
## Sat 9 30

```

- Weekends tend to have lesser number of appointments and weekdays have more . Tuesday and wednesday are busy days mostly. Day factor seem to be significant for determining show no show factor.
- Saturday has highest number of no shows. Seems like weekend appointment tend to cancel more

```
# H0: weekday is significant in determining target factor
chisq.test(table(med_data$no_show_status,med_data$day))
```

```

## 
## Pearson's Chi-squared test
## 
## data: table(med_data$no_show_status, med_data$day)
## X-squared = 27.48, df = 5, p-value = 4.599e-05

```

-Since p value is significantly low suggesting showing up in appointment day is dependent on which day the appointment is.

7.2.6- EDA on Neighborhood vs No Show

```
df_neighbor <- data.frame(table(med_data$neighborhood, med_data$no_show_status))
names(df_neighbor) <- c("neighborhood", "no_show", 'Count')
```

```

# visualization on neighbourhood
gg_neigh<-ggplot(df_neighbor)+  

  geom_bar(aes(x = reorder(neighborhood, -Count), y = Count, fill = no_show),  

    stat = 'identity')+  

  theme(axis.text.x = element_text(size= 12, angle = 90, hjust = 1))+  

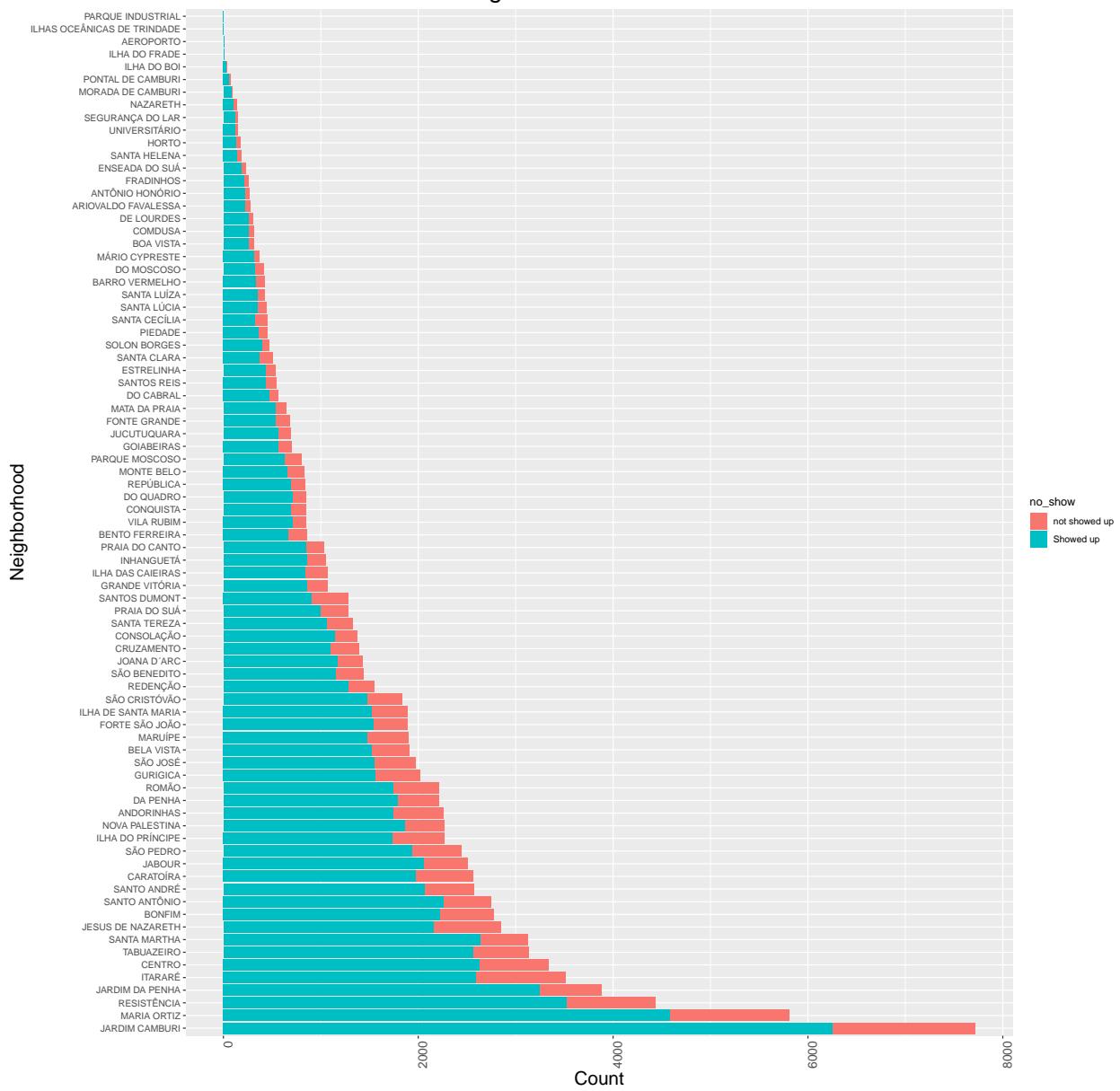
  ggtitle("Neighborhood vs No Show")+
  ylab('Count')+
  xlab('Neighborhood')+
  theme(plot.title = element_text(hjust = 0.5, size = 24))+  

  theme(axis.title.y = element_text(size =18))+  

  theme(axis.title.x = element_text(size =18))
gg_neigh+coord_flip()

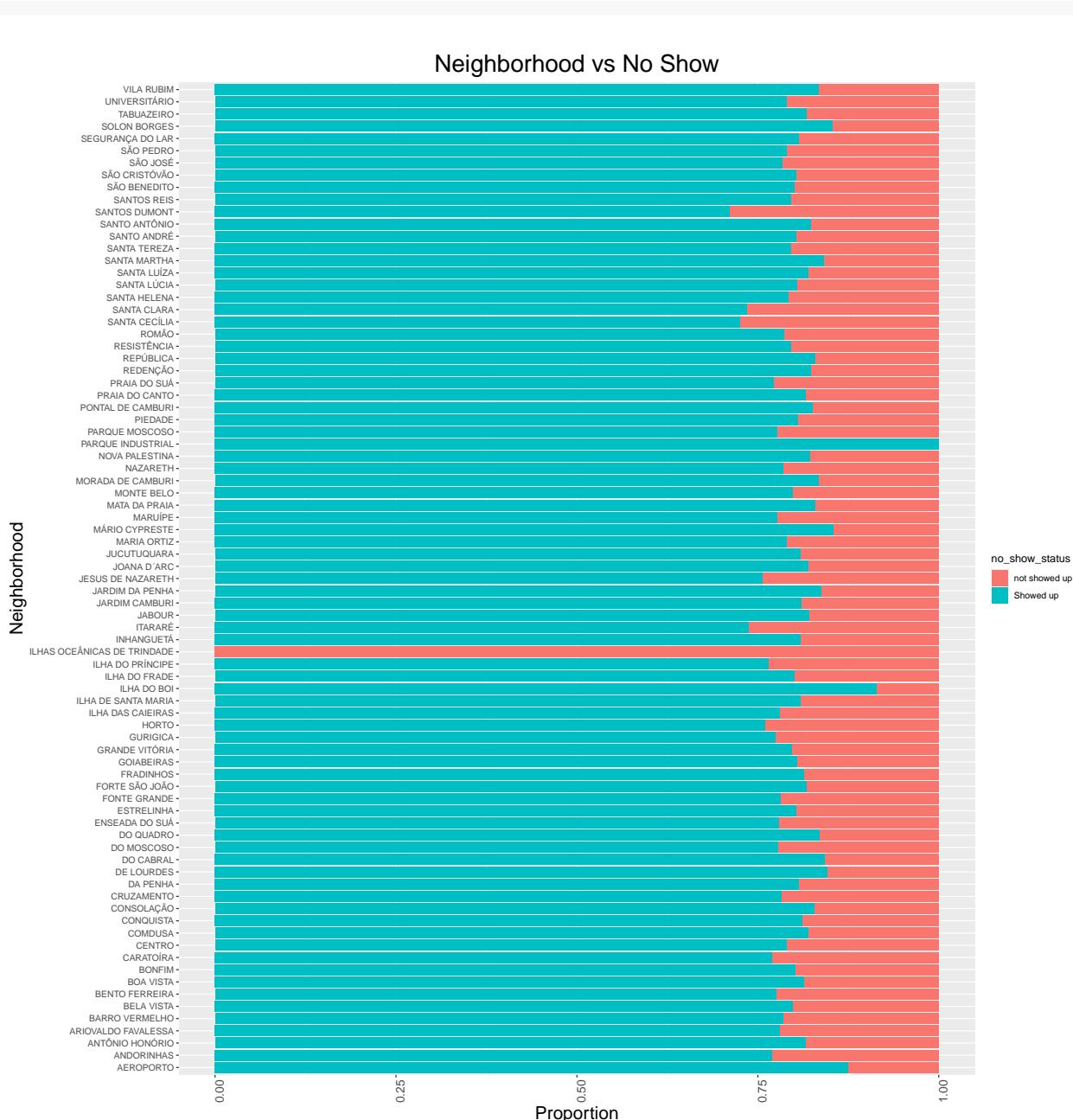
```

Neighborhood vs No Show



There is no clear analysis on how neighbourhood is affecting appointment no show status so plotting proportions

```
# proportion
gg_neighbour<-ggplot(med_data)+  
  geom_bar(aes(x = neighborhood, fill = no_show_status), position = position_fill())+  
  theme(axis.text.x = element_text(size= 12, angle = 90, hjust = 1))+  
  ggtitle("Neighborhood vs No Show")+\n  ylab('Proportion')+\n  xlab('Neighborhood')+\n  theme(plot.title = element_text(hjust = 0.5))+\n  theme(plot.title = element_text(hjust = 0.5, size = 24))+\n  theme(axis.title.y = element_text(size =18))+\n  theme(axis.title.x = element_text(size =18))  
gg_neighbour+coord_flip()
```



-ILHAS OCEANICAS DE TRINDADE neighbourhood has maximum no show counts of patients followed by SANTOS DUMONT area

7.3- Model Selection and Predictor Significance

```
# considering important factors
med_data_2 <- dplyr::select(med_data, age, gender, scholarship, hypertension, diabetes,
                             alcoholism, handicap,sms_received, day, no_show_status)
med_data_2 <- mutate_at(med_data_2, vars(day), as.factor)
str(med_data_2)
```

```

## Classes 'tbl_df', 'tbl' and 'data.frame': 110526 obs. of 10 variables:
## $ age : num 62 56 62 8 56 76 23 39 21 19 ...
## $ gender : Factor w/ 2 levels "Female","Male": 1 2 1 1 1 1 1 1 1 1 ...
## $ scholarship : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
## $ hypertension : Factor w/ 2 levels "0","1": 2 1 1 1 2 2 1 1 1 1 ...
## $ diabetes : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
## $ alcoholism : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
## $ handicap : Factor w/ 5 levels "0","1","2","3",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ sms_received : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
## $ day : Factor w/ 6 levels "Friday","Monday",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ no_show_status: Factor w/ 2 levels "not showed up",...: 2 2 2 2 2 2 1 1 2 2 ...

library(faraway)
log_model_1 <- glm(no_show_status ~ . ,family = binomial(link = 'logit'), data = med_data_2 )
summary(log_model_1)

##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.25631179  0.02355356 53.3385 < 2.2e-16
## age              0.00656250  0.00039285 16.7050 < 2.2e-16
## genderMale      0.01848034  0.01629266  1.1343  0.256680
## scholarship1   -0.18453145  0.02452045 -7.5256 5.247e-14
## hypertension1   0.06719226  0.02459479  2.7320  0.006296
## diabetes1       -0.08323481  0.03413837 -2.4382  0.014762
## alcoholism1    -0.13903678  0.04480999 -3.1028  0.001917
## handicap1        -0.01014531  0.05902362 -0.1719  0.863528
## handicap2        -0.13951084  0.18636164 -0.7486  0.454097
## handicap3        -0.30201039  0.66841267 -0.4518  0.651390
## handicap4        -0.54516959  1.23972347 -0.4398  0.660117
## sms_received1   -0.66754921  0.01564034 -42.6812 < 2.2e-16
## dayMonday        0.10276547  0.02446534  4.2005 2.664e-05
## daySaturday      -0.14466814  0.38587059 -0.3749  0.707725
## dayThursday      0.15981834  0.02649627  6.0317 1.622e-09
## dayTuesday       0.19918675  0.02408981  8.2685 < 2.2e-16
## dayWednesday     0.18329495  0.02400285  7.6364 2.234e-14
##
## n = 110526 p = 17
## Deviance = 108899.95262 Null Deviance = 111205.15739 (Difference = 2305.20477)

```

- We can observe negative coefficients for scholarship, true, diabetes, alcoholism, handicap, sms_received, weekend day saturday indicating number of no shows of patients increases if these are true for patient. While for male gender, weekdays, hypertension and age factors number of no shows are lesser.

```
exp(coef(log_model_1))
```

```

## (Intercept)          age   genderMale scholarship1 hypertension1
## 3.5124429      1.0065841    1.0186522    0.8314938    1.0695011
## diabetes1      alcoholism1    handicap1    handicap2    handicap3
## 0.9201351      0.8701960    0.9899060    0.8697836    0.7393304
## handicap4      sms_received1    dayMonday    daySaturday   dayThursday
## 0.5797435      0.5129642    1.1082315    0.8653094    1.1732977
## dayTuesday      dayWednesday
## 1.2204099      1.2011686

```

- We can interpret coefficients for:

- (a) saturday - 0.8653094 ODDS- the odds of showing up for appointment reduces by 13.4% on saturday keeping all other factors constant like age,gender, scholarship yes or no,hypertension,diabetes,alcoholism etc
 (b)hypertension-1.0695011 ODDS- the odds of showing up for appointment increases by 6.9% if patient has hypertension keeping all other factors fixed. Seems reasonable if patient already has some medical ailment , he tends to show up for appointment.

```
# we can even test significance of individual predictor using drop1 function
drop1(log_model_1,test = "Chi")
```

```
## Single term deletions
##
## Model:
## no_show_status ~ age + gender + scholarship + hypertension +
##   diabetes + alcoholism + handicap + sms_received + day
##             Df Deviance    AIC      LRT Pr(>Chi)
## <none>          108900 108934
## age            1  109183 109215  282.71 < 2.2e-16 ***
## gender         1  108901 108933   1.29  0.256444
## scholarship    1  108955 108987  55.32 1.022e-13 ***
## hypertension   1  108907 108939   7.50  0.006178 **
## diabetes       1  108906 108938   5.89  0.015209 *
## alcoholism     1  108909 108941   9.40  0.002167 **
## handicap        4  108901 108927   0.95  0.917735
## sms_received   1  110691 110723 1791.41 < 2.2e-16 ***
## day            5  108986 109010   86.15 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Gender and handicap p-values are significantly higher indicating we can drop them from our model as they are not significant

```
med_data_3 <- dplyr::select(med_data, age, scholarship, hypertension, diabetes,
                             alcoholism, sms_received, day, no_show_status)
log_model_2 <- glm(no_show_status ~ . ,family = binomial(link = 'logit'), data = med_data_3 )
summary(log_model_2)
```

```
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.26485218  0.02223331 56.8900 < 2.2e-16
## age          0.00650425  0.00038938 16.7042 < 2.2e-16
## scholarship -0.18828202  0.02429380 -7.7502 9.174e-15
## hypertension  0.06691631  0.02457387  2.7231  0.006468
## diabetes     -0.08356367  0.03412551 -2.4487  0.014337
## alcoholism    -0.13303247  0.04448370 -2.9906  0.002784
## sms_received -0.66814088  0.01562054 -42.7732 < 2.2e-16
## dayMonday     0.10293040  0.02446472  4.2073 2.584e-05
## daySaturday   -0.14562921  0.38581797 -0.3775  0.705835
## dayThursday   0.15992426  0.02649561  6.0359 1.581e-09
## dayTuesday    0.19910174  0.02408902  8.2652 < 2.2e-16
## dayWednesday  0.18329972  0.02400240  7.6367 2.228e-14
##
## n = 110526 p = 12
## Deviance = 108902.15180 Null Deviance = 111205.15739 (Difference = 2303.00560)
```

7.4- Logistic Regression with training and test data set

```
library(caTools)
set.seed(100)
indices = sample.split(med_data_3$no_show_status, SplitRatio = 0.7)
train = med_data_3[indices,]
test = med_data_3[!(indices),]

logit_model_1 <- glm(no_show_status ~ . , data = train, family =binomial(link = 'logit') )
# Stepwise selection
library("MASS")

## 
## Attaching package: 'MASS'

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

model_2<- stepAIC(logit_model_1, direction="both")

## Start: AIC=76144.38
## no_show_status ~ age + scholarship + hypertension + diabetes +
##      alcoholism + sms_received + day
##
##          Df Deviance AIC
## - diabetes     1    76122 76144
## <none>           76120 76144
## - alcoholism   1    76126 76148
## - hypertension  1    76127 76149
## - scholarship   1    76162 76184
## - day           5    76181 76195
## - age           1    76329 76351
## - sms_received  1    77458 77480
##
## Step: AIC=76144.35
## no_show_status ~ age + scholarship + hypertension + alcoholism +
##      sms_received + day
##
##          Df Deviance AIC
## <none>           76122 76144
## + diabetes       1    76120 76144
## - hypertension   1    76128 76148
## - alcoholism     1    76128 76148
## - scholarship    1    76164 76184
## - day            5    76183 76195
## - age            1    76329 76349
## - sms_received   1    77459 77479
```

```
vif(logit_model_1)
```

```
##          age scholarship1 hypertension1      diabetes1    alcoholism1
## 8.948484     5.783948     10.642574     8.778435     6.497357
## sms_received1 dayMonday   daySaturday   dayThursday  dayTuesday
## 5.883414     10.886425     6.189557    10.237957    11.495219
## dayWednesday
## 11.543298
```

- $vif > 10$ suggests inflation of factor in this case weekdays are getting inflated

```
# Model from AIC obtained
```

```
logit_model_2 <- glm(no_show_status ~ age + scholarship + hypertension + alcoholism +
  sms_received + day , data = train, family =binomial(link = 'logit') )
summary(logit_model_2)
```

```
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.2665112 0.0266719 47.4848 < 2.2e-16
## age         0.0066046 0.0004626 14.2771 < 2.2e-16
## scholarship1 -0.1877767 0.0288441 -6.5101 7.512e-11
## hypertension1  0.0628202 0.0275242  2.2824 0.0224681
## alcoholism1   -0.1304371 0.0536683 -2.4304 0.0150809
## sms_received1 -0.6875587 0.0186545 -36.8575 < 2.2e-16
## dayMonday     0.0990773 0.0292941  3.3822 0.0007192
## daySaturday    -0.1645879 0.5080276 -0.3240 0.7459574
## dayThursday    0.1503555 0.0317908  4.7295 2.250e-06
## dayTuesday     0.1914993 0.0288525  6.6372 3.197e-11
## dayWednesday   0.1951276 0.0288213  6.7702 1.286e-11
##
## n = 77368 p = 11
## Deviance = 76122.34789 Null Deviance = 77842.69539 (Difference = 1720.34751)
```

- hypertension and alcoholism have significantly higher p-values suggesting they are not significant in determining show/no-show status of appointment.

```
# Removing alcoholism and hypertension
```

```
logit_model_3 <- glm(no_show_status ~ age + scholarship +
  sms_received + day , data = train, family =binomial(link = 'logit') )
summary(logit_model_3)
```

```
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.25886723 0.02646750 47.5628 < 2.2e-16
## age         0.00702763 0.00039986 17.5751 < 2.2e-16
## scholarship1 -0.18889502 0.02880271 -6.5582 5.445e-11
## sms_received1 -0.68675995 0.01864181 -36.8398 < 2.2e-16
## dayMonday     0.09890574 0.02929276  3.3765 0.0007343
## daySaturday    -0.15221819 0.50773479 -0.2998 0.7643308
## dayThursday    0.15052495 0.03178871  4.7352 2.189e-06
## dayTuesday     0.19167885 0.02885029  6.6439 3.055e-11
## dayWednesday   0.19496395 0.02881951  6.7650 1.333e-11
##
## n = 77368 p = 9
## Deviance = 76132.84121 Null Deviance = 77842.69539 (Difference = 1709.85419)
```

- Most of the p-values are significant suggesting possible significance in model building

7.5- Model Evaluation

```
#predicted probabilities of appointment miss for test data
test_predicted = predict(logit_model_3, type = "response", newdata = test)
summary(test_predicted)
```

```
##      Min. 1st Qu. Median     Mean 3rd Qu.     Max.
## 0.5963  0.7550  0.8159  0.7985  0.8458  0.8940
```

```
pred_test <- ifelse(test_predicted>0.5,"No show","show")
tab <- table(predicted = pred_test, actual = test$no_show_status)
tab
```

```
##           actual
## predicted not showed up Showed up
##   No show          6696      26462
```

Model is not like as we expected because there is class imbalance in between showed up and not showed up group. There are 80 % patients who show up and 20 % those who don't show up for appointment

```
# observe residuals
qqnorm(residuals(logit_model_3))
```

