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 tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

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

What percentage of patients miss their medical appoitment 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

```r
str(med_data)
```

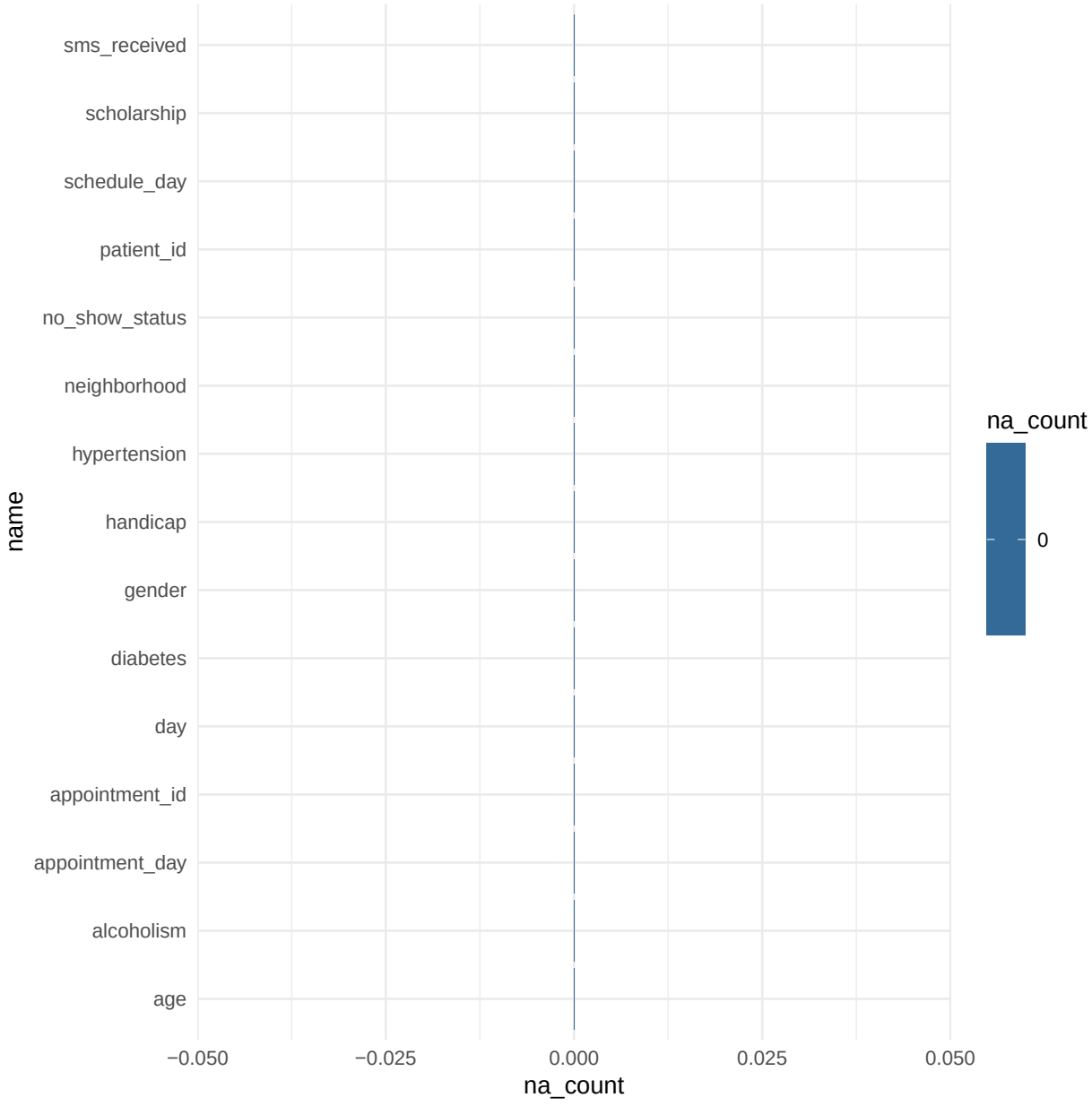
```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 110527 obs. of  14 variables:
##  $ PatientId     : num  2.99e+13 5.59e+14 4.26e+12 8.68e+11 8.84e+12 ...
##  $ AppointmentID : num  5642903 5642503 5642549 5642828 5642494 ...
##  $ Gender        : chr  "F" "M" "F" "F" ...
##  $ ScheduledDay  : 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

```r
# 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"
```

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

```r
# 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

```
# coverting 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
##
##
```

```
##
##
```

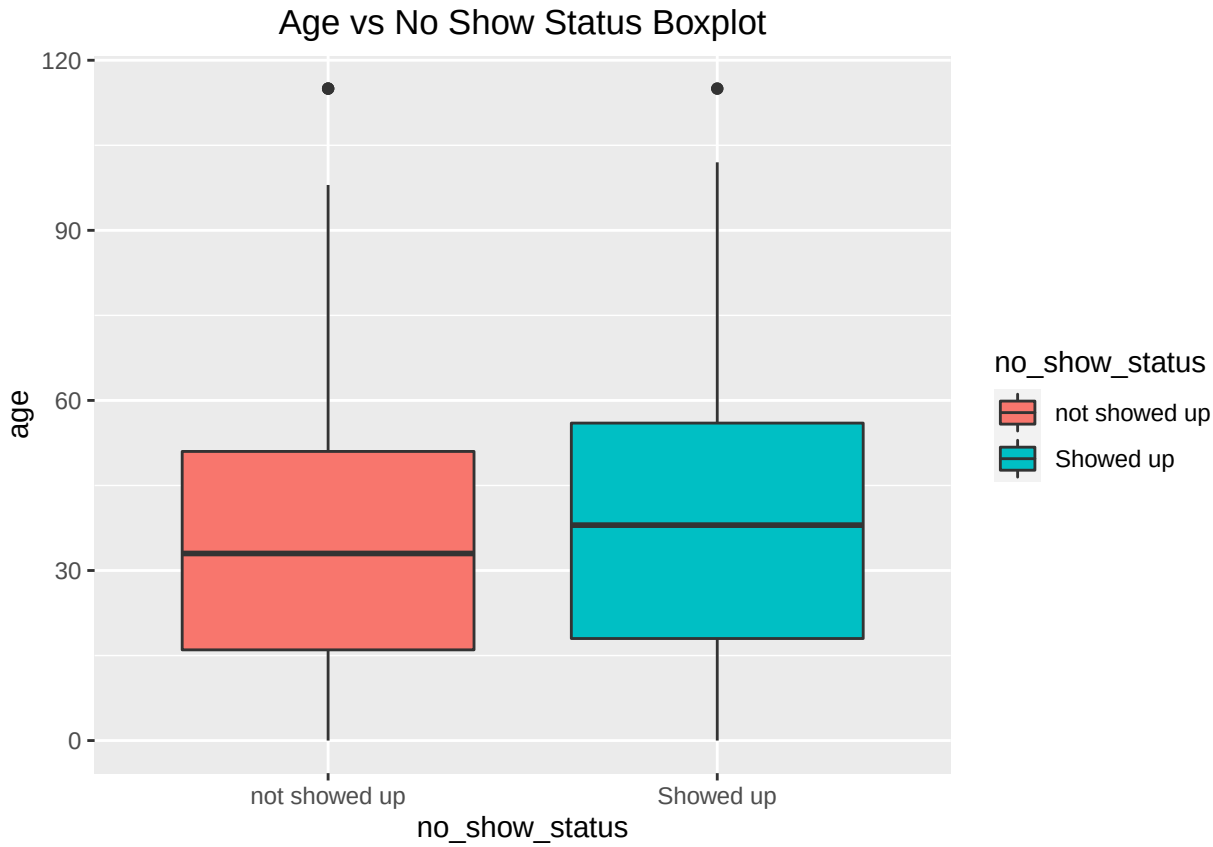
```
# removing age = -1
med_data<-med_data[!(med_data$age<0),]
summary(med_data$age)
```

```
##    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 appoitnments
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.

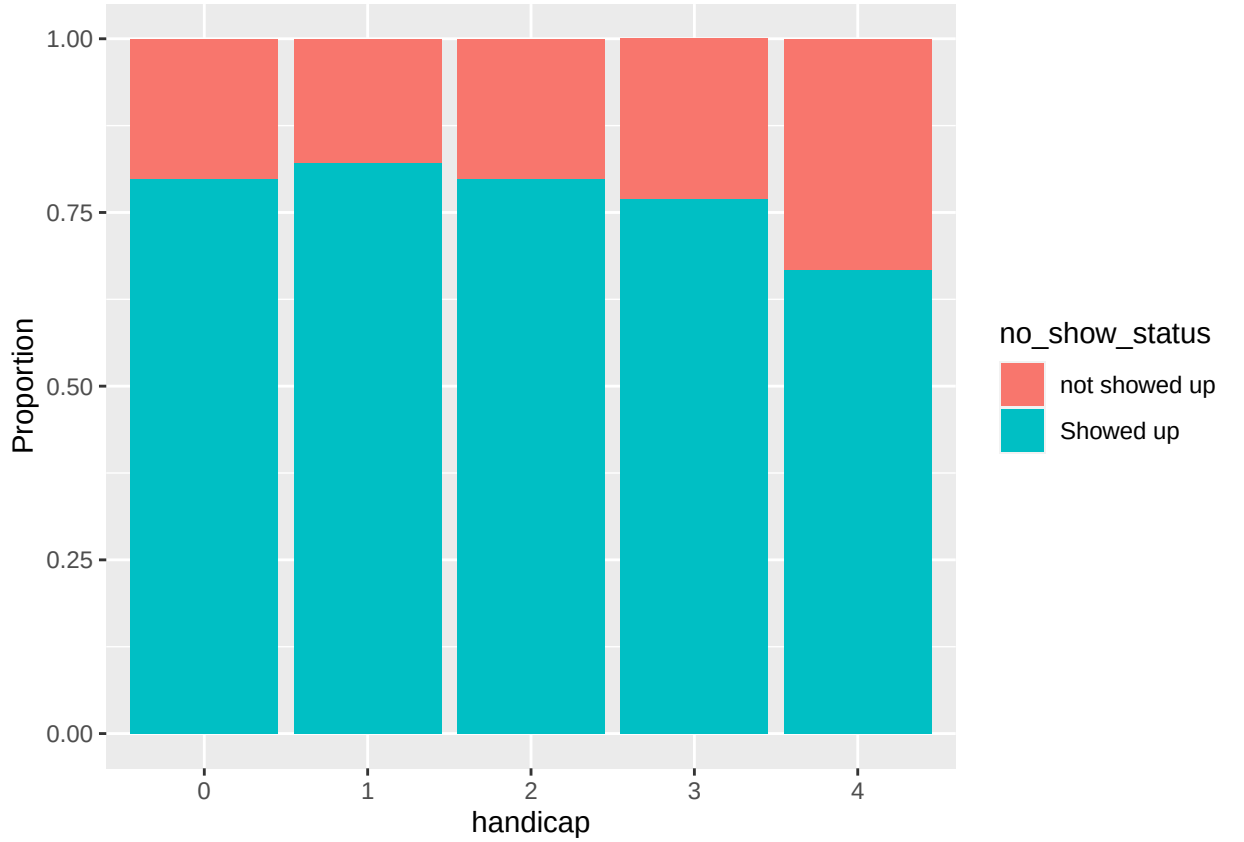
```
# Statistical test for significance of gender
# H0: gender factor is not significant in predicting show/no show factor
chisq.test(table(med_data$gender,med_data$no_show_status), correct = FALSE)# without continuity correct
```

```
##
##  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

```
# Visulizing 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 variabl esuggests show/no-show factor is not really dependent on it . Furthur we can verify it using chi squared test

```
# Pearson Chi squared test: HO- 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.

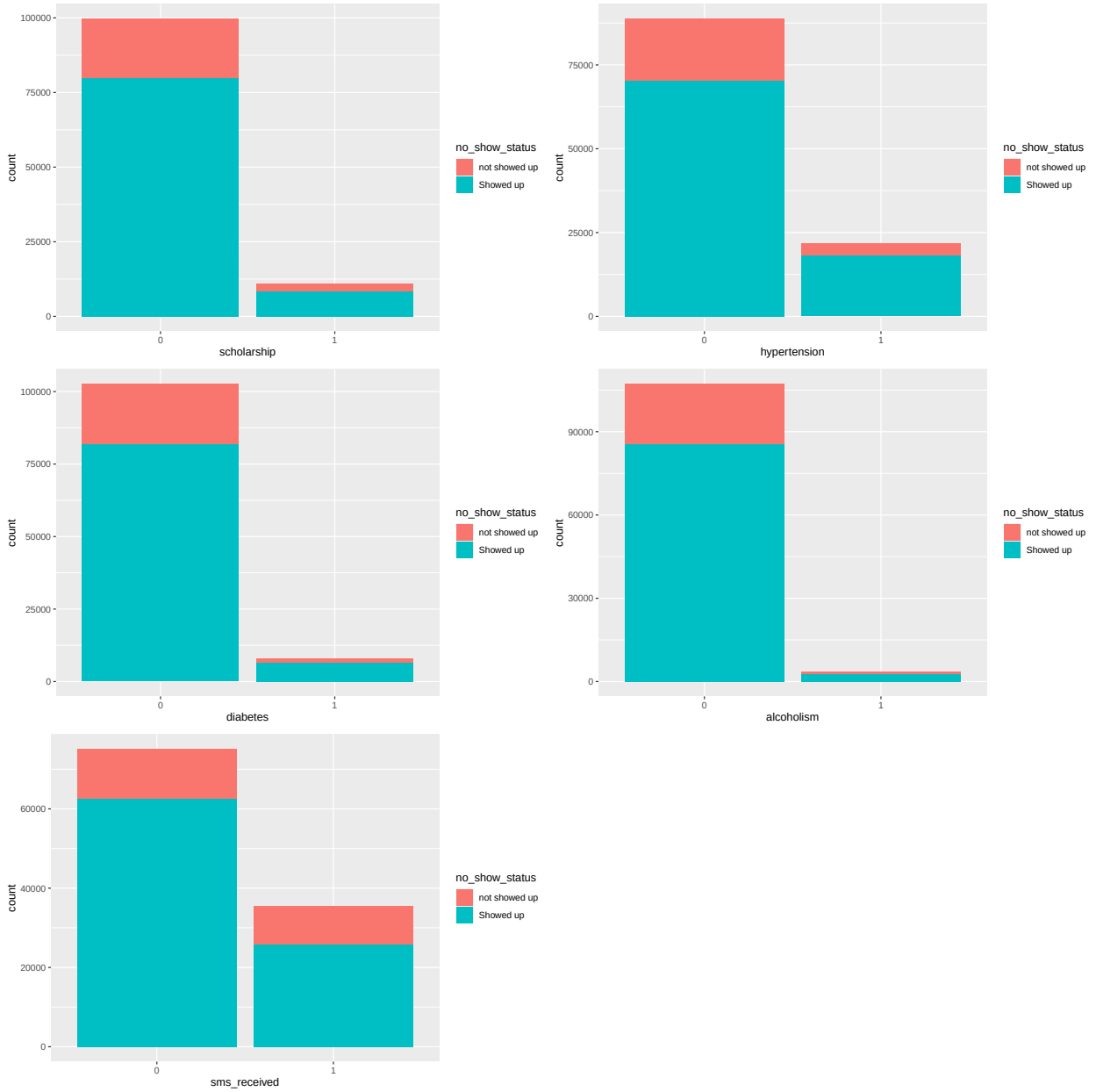
7.2.4- EDA on other factor variables

```
library(gridExtra)
```

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

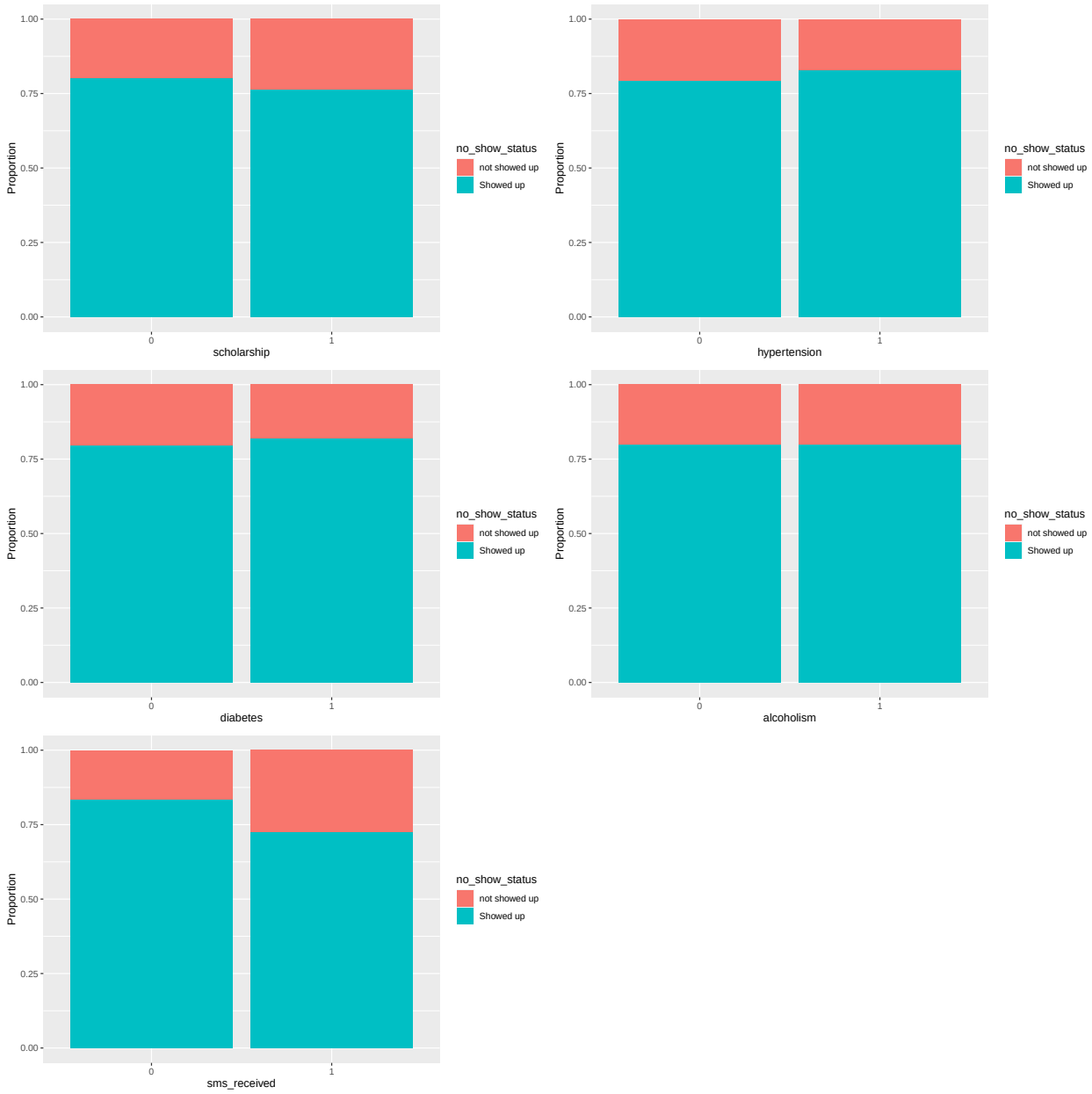
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
# 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
```

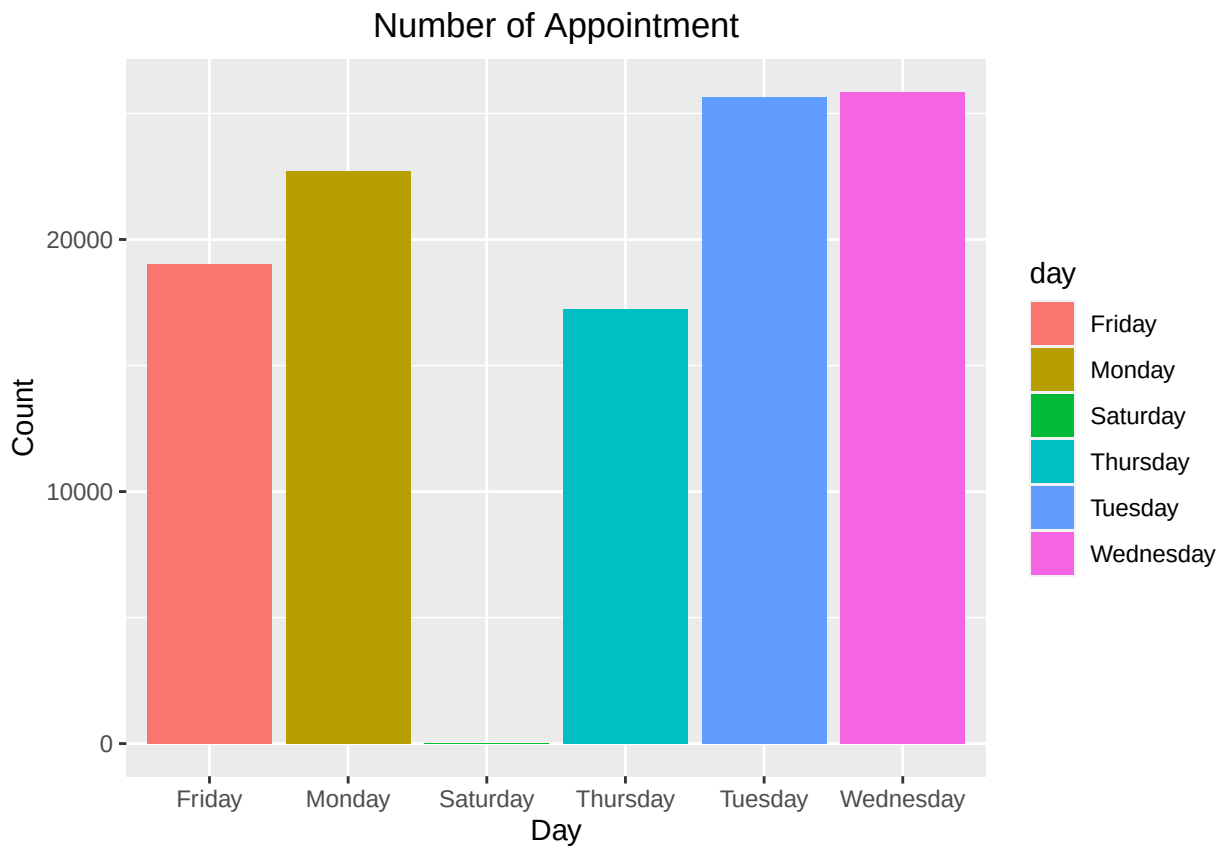
```
chisq.test(table(med_data$sms_received, med_data$no_show_status), correct = FALSE) # sms_received
```

```
##
##  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))
```

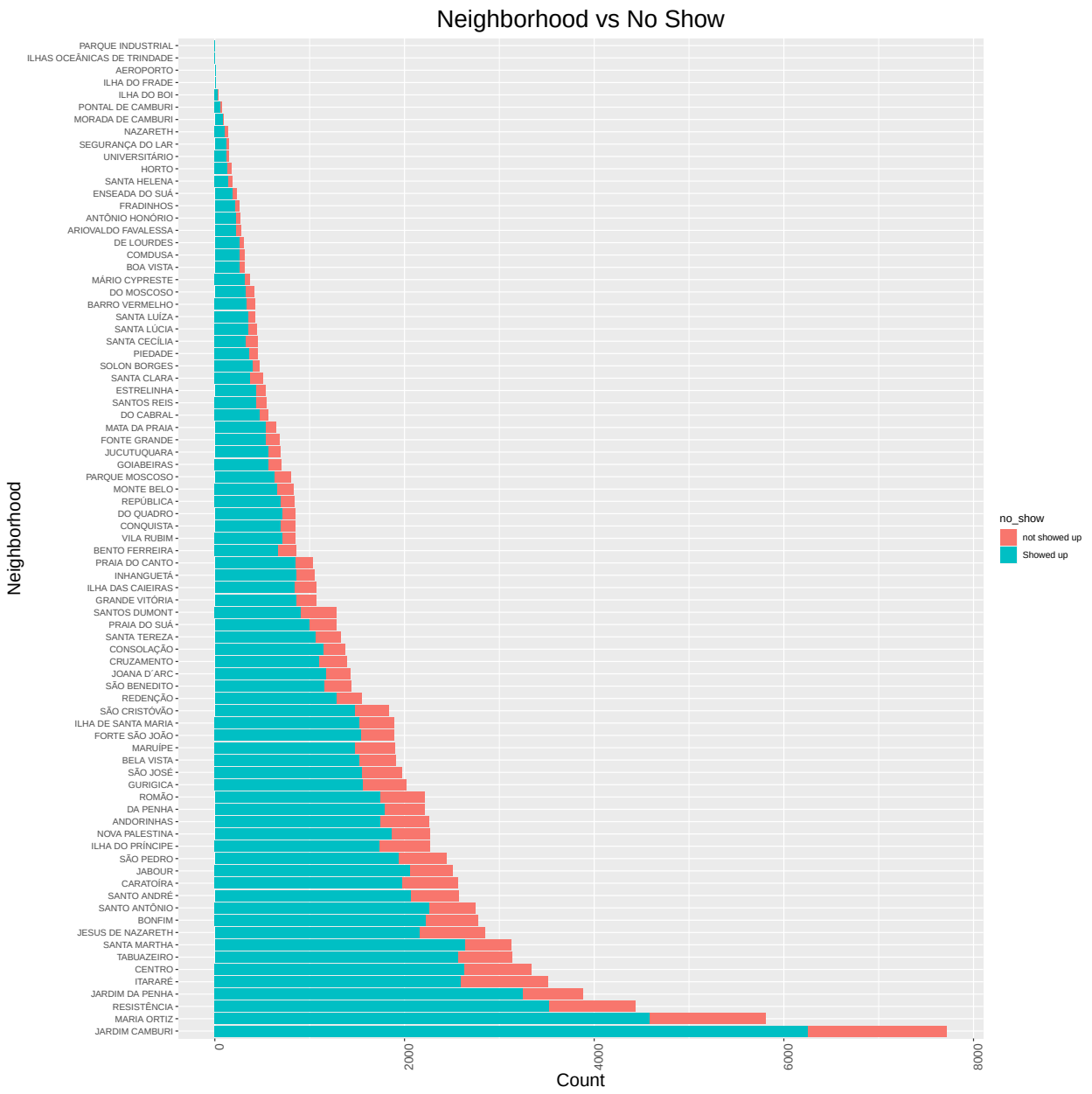
```
##
##  Pearson's Chi-squared test
##
## data:  table(med_data$no_show_status, med_data$day)
## X-squared = 27.48, df = 5, p-value = 4.599e-05
```

-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

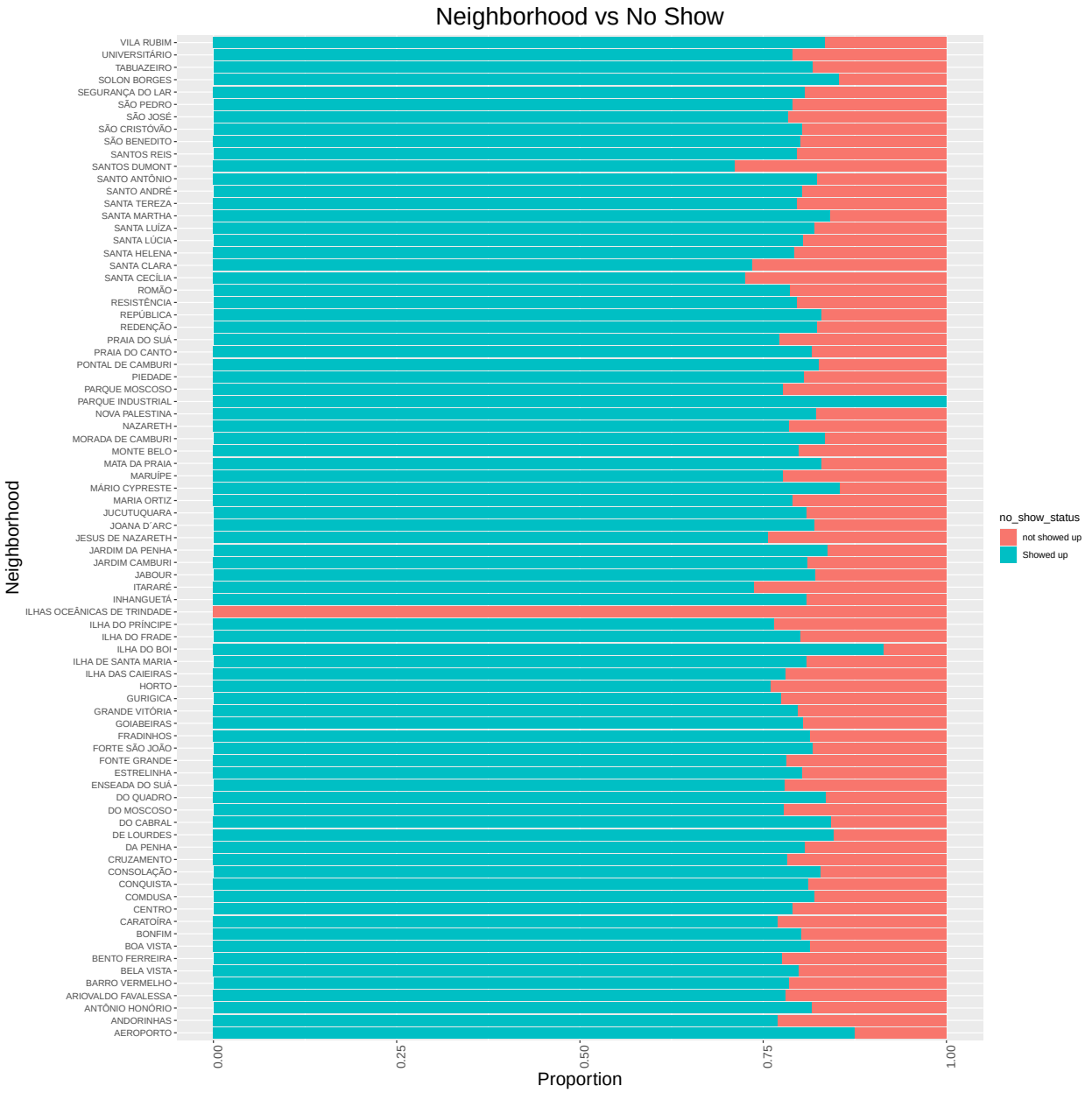
```
df_neighbor <- data.frame(table(med_data$neighborhood, med_data$no_show_status))
names(df_neighbor) <- c("neighborhood", "no_show", 'Count')
```

```
# 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()
```



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")+
    ylab('Proportion')+
    xlab('Neighborhood')+
    theme(plot.title = element_text(hjust = 0.5))+
    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_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 )
sumary(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 predcitor using drop1 fucntion
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 )
sumary(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
## scholarship1  -0.18828202  0.02429380  -7.7502 9.174e-15
## hypertension1  0.06691631  0.02457387   2.7231  0.006468
## diabetes1     -0.08356367  0.03412551  -2.4487  0.014337
## alcoholism1   -0.13303247  0.04448370  -2.9906  0.002784
## sms_received1 -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') )
sumary(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 alcholism and hypertension
logit_model_3 <- glm(no_show_status ~ age + scholarship +
    sms_received + day ,  data = train, family =binomial(link = 'logit') )
sumary(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))
```