Insurance Charges Prediction

Title: CUNY SPS MDS DATA621_Final Project”

Author: Charles Ugiagbe

Date: 5/24/2023

Part 1 - Introduction

Medical expenses are any costs incurred in the prevention or treatment of injury or disease. To realize their profit, insurance companies must charge a higher premium than the amount paid to the insured. For this reason, insurance companies invest a lot of time, effort, and money in creating models that are able to accurately predict health care costs/charges. In order to fulfill this mission, we will first analyze the factors that influence medical loads and secondly try to build an adequate model and optimize its performance. For this study, our objective are:

- Determine if the mean insurance charges of Smokers in the dataset is different from the mean charges of Non - smokers
- Formulate a multiple Regression model or predicting the insurance charges of individuals

Part 2 - Data

Data Source:

Data is from kaggle public datasets and can be found online here: https://www.kaggle.com/mirichoi0218/insurance (https://www.kaggle.com/mirichoi0218/insurance)

Type of study:

This is an observational study as there is no control group.

Cases:

There are 7 variables and 1338 observations in the dataset. six(6) of the Variable in the dataset are potential predictor of the of the 7th variables (Insurance charges). There are no missing value in any of the observation. Each observation represents the likely variable that play vital roles in determining the insurance charge. The variables are explained below.

- Age: the age of the insured (recipients).
- Sex: sex of insured persons; “male” or “female”.
- bmi: body mass index, providing an understanding of the body, relatively high or low weights relative to height, objective body weight index (kg / m ^ 2) using the height / weight ratio.
- children: number of children covered by health insurance / number of dependents.
- smoker: does the insured smoke or not.
- region: the recipient’s residential area in the United States; northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance.

Response Variable (Dependent Variable)

The Dependent variable is the Insurance Charges and its numerical

Predictor Variables (Independent Variables):

There are six(6) independent used. They independent variables are: Age(numeric), sex(categorical), BMI(numeric), Children(numeric), Smoker(categorical), Region(categorical)

Part 3 - Exploratory data analysis

Data Preparation

```
# load the required libraries
library(tidyverse)
library(magrittr)
library(Amelia)
library(corrplot)
library(cowplot)
library(gridExtra)
```

load the Data and view the head

```
url <- "https://raw.githubusercontent.com/omocharly/DATA606_PROJECT/main/insurance.csv"
insurance <- read_csv(url)
```

```
## Rows: 1338 Columns: 7
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr (3): sex, smoker, region
## dbl (4): age, bmi, children, charges
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(insurance)
```

```
## # A tibble: 6 × 7
##     age sex      bmi children smoker region    charges
##   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
## 1    19 female  27.9        0 yes    southwest  16885.
## 2    18 male    33.8        1 no     southeast   1726.
## 3    28 male    33          3 no     southeast   4449.
## 4    33 male    22.7        0 no     northwest  21984.
## 5    32 male    28.9        0 no     northwest   3867.
## 6    31 female  25.7        0 no     southeast   3757.
```

Take a glimpse look at the data structure

```
glimpse(insurance)
```

```
## Rows: 1,338
## Columns: 7
## $ age      <dbl> 19, 18, 28, 33, 32, 31, 46, 37, 37, 60, 25, 62, 23, 56, 27, 1…
## $ sex      <chr> "female", "male", "male", "male", "male", "female", "female",…
## $ bmi      <dbl> 27.900, 33.770, 33.000, 22.705, 28.880, 25.740, 33.440, 27.74…
## $ children <dbl> 0, 1, 3, 0, 0, 0, 1, 3, 2, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0…
## $ smoker   <chr> "yes", "no", "no", "no", "no", "no", "no", "no", "no", "no", …
## $ region   <chr> "southwest", "southeast", "southeast", "northwest", "northwes…
## $ charges  <dbl> 16884.924, 1725.552, 4449.462, 21984.471, 3866.855, 3756.622,…
```

The Data has 7 variable and 1,338 Observation.

We will convert the variables sex, children, region, smoker to the type factor which corresponds to the categorical variables for easy analysis:

```
insurance$sex %<>% as.factor()
insurance$children %<>% as.factor()
insurance$region %<>% as.factor()
insurance$smoker %<>% as.factor()
```

Take the summary of the data

```
summary(insurance)
```

```
##       age            sex           bmi        children smoker
##  Min.   :18.00   female:662   Min.   :15.96   0:574    no :1064
##  1st Qu.:27.00   male  :676   1st Qu.:26.30   1:324    yes: 274
##  Median :39.00                Median :30.40   2:240
##  Mean   :39.21                Mean   :30.66   3:157
##  3rd Qu.:51.00                3rd Qu.:34.69   4: 25
##  Max.   :64.00                Max.   :53.13   5: 18
##        region       charges
##  northeast:324   Min.   : 1122
##  northwest:325   1st Qu.: 4740
##  southeast:364   Median : 9382
##  southwest:325   Mean   :13270
##                  3rd Qu.:16640
##                  Max.   :63770
```

put BMI into categories

```
insurance %<>% mutate(bmi_cat = cut(bmi,
  breaks = c(0, 18.5, 25, 30, 60),
  labels = c("Under Weight", "Normal Weight", "Overweight", "Obese")
))
```

- Under Weight: bmi<18.5
- Normal Weight: 18.5<bmi<25
- Overweight : 25=<bmi<30
- Obese: bmi>=30

Missing values

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

	count missing
age	0
sex	0
bmi	0
children	0
smoker	0
region	0
charges	0
bmi_cat	0

The dataset contains no missing values

Description of categorical variables

```
categ_cols <- insurance %>% select_if(~ class(.) == "factor")
for (col in names(categ_cols)) {
  t <- insurance %>%
    group_by_(col) %>%
    summarise(count = n()) %>%
    mutate(frequency = paste0(round(100 * count / sum(count), 0), "%")) %>%
    knitr::kable("html", align = "lcc") %>%
    kableExtra::kable_styling(full_width = F, position = "left") %>%
    print()
}
```

```
## Warning: `group_by_()` was deprecated in dplyr 0.7.0.
## ℹ Please use `group_by()` instead.
## ℹ See vignette('programming') for more help
```

sex	count	frequency
female	662	49%
male	676	51%

```
## Warning: `group_by_()` was deprecated in dplyr 0.7.0.
## ℹ Please use `group_by()` instead.
## ℹ See vignette('programming') for more help
```

children	count	frequency
0	574	43%
1	324	24%
2	240	18%
3	157	12%
4	25	2%
5	18	1%

```
## Warning: `group_by_()` was deprecated in dplyr 0.7.0.
## ℹ Please use `group_by()` instead.
## ℹ See vignette('programming') for more help
```

smoker	count	frequency
no	1064	80%
yes	274	20%

```
## Warning: `group_by_()` was deprecated in dplyr 0.7.0.
## ℹ Please use `group_by()` instead.
## ℹ See vignette('programming') for more help
```

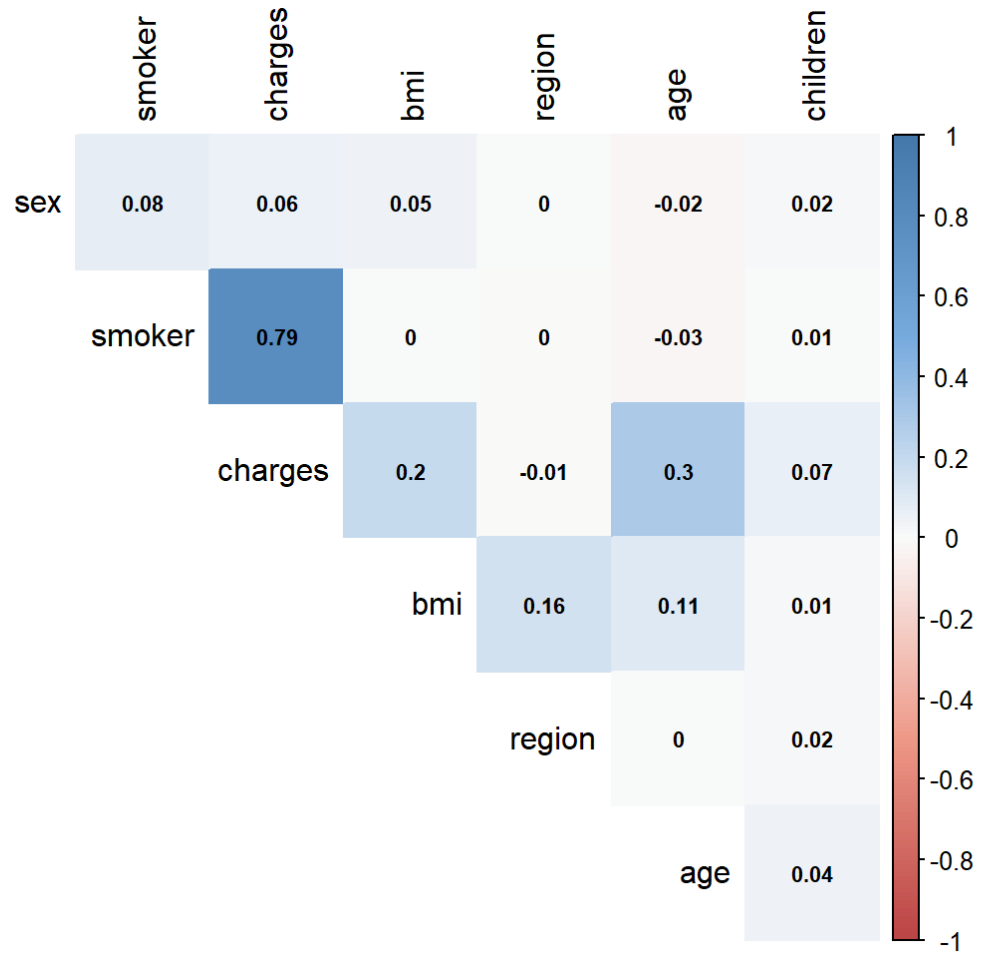
region	count	frequency
northeast	324	24%
northwest	325	24%
southeast	364	27%
southwest	325	24%

```
## Warning: `group_by_()` was deprecated in dplyr 0.7.0.
## ℹ Please use `group_by()` instead.
## ℹ See vignette('programming') for more help
```

bmi_cat	count	frequency
Under Weight	21	2%
Normal Weight	226	17%
Overweight	386	29%
Obese	705	53%

Visualisation

Correlation matrix



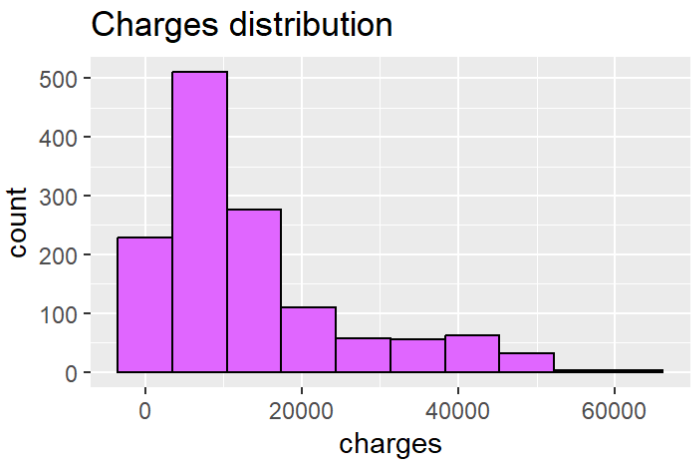
The variables most correlated with the charges are “smoker”, “age” and “bmi”.

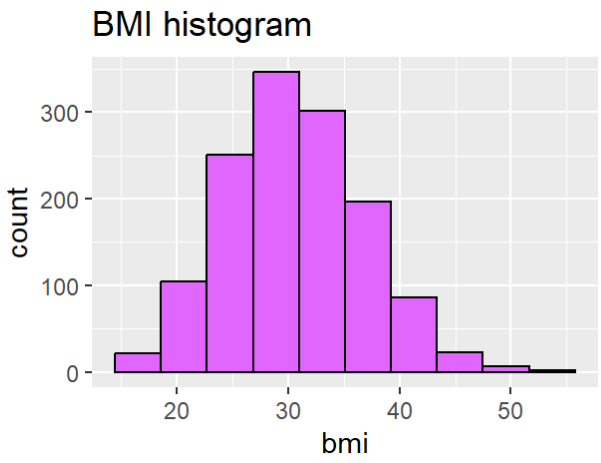
Visualization of Charges, Age, Smoker and BMI distribution

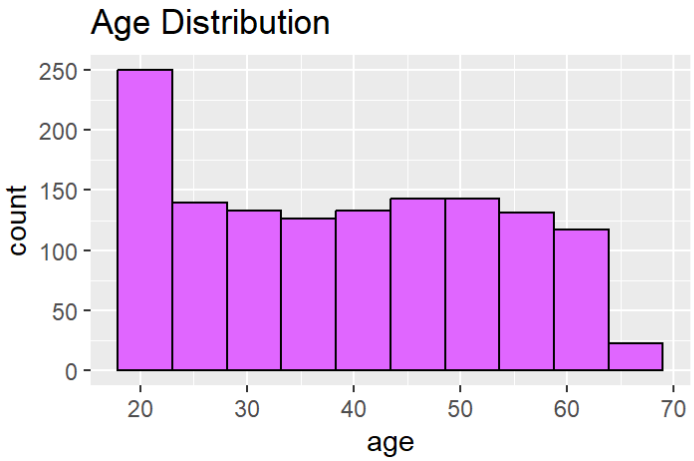
```
d1<-ggplot(data = insurance,aes(x=charges)) + geom_histogram(color="black", fill="mediumor
chid1", bins=10)+
labs(title="Charges distribution")
d2<-ggplot(data = insurance,aes(x=bmi)) + geom_histogram(color="black", fill="mediumorchid
1", bins=10)+
labs(title="BMI histogram")
d3<-ggplot(data = insurance,aes(x=age)) + geom_histogram(color="black", fill="mediumorchid
1", bins=10)+
labs(title="Age Distribution")
d4<-ggplot(data = insurance,aes(x=smoker)) + geom_bar(color="black", fill="mediumorchid1",
bins=10)+
labs(title="Smoker Distribution")
```

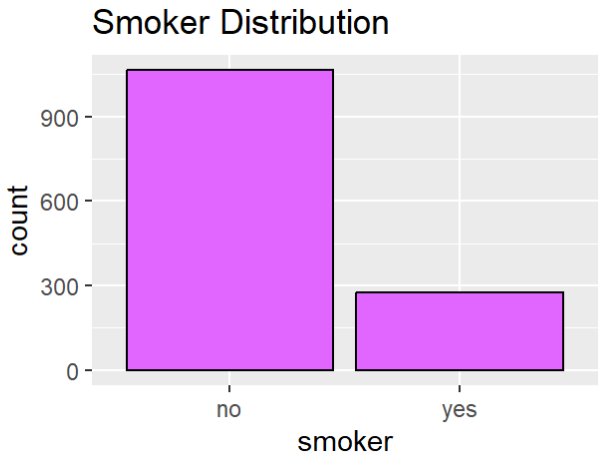
```
## Warning in geom_bar(color = "black", fill = "mediumorchid1", bins = 10):
## Ignoring unknown parameters: `bins`
```

```
plot_grid(d1, d2, d3, d4, rel_widths = c(1.15, 1),ncol = 2,
  align = "hv")
```





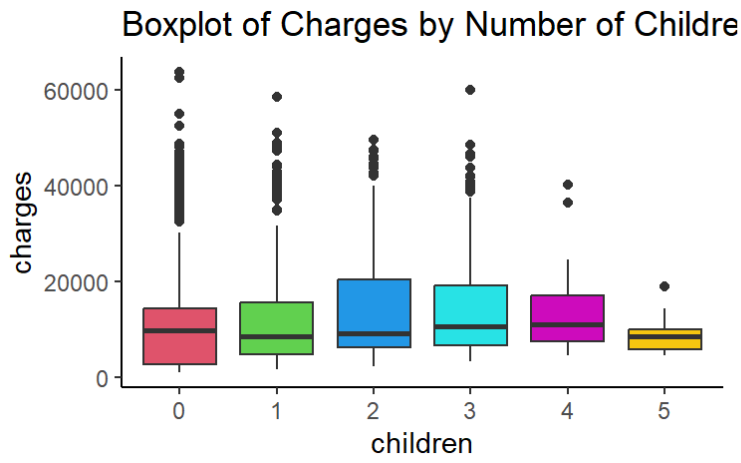


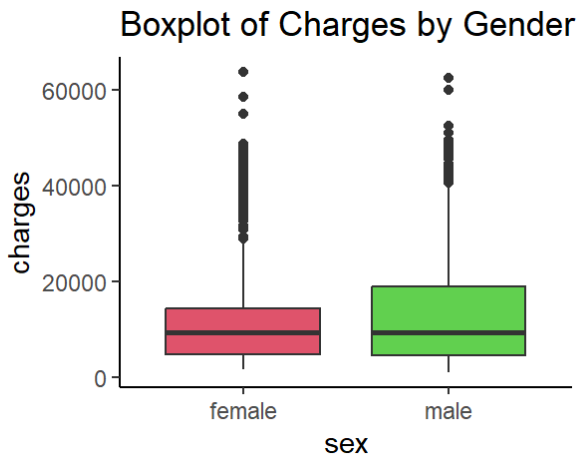


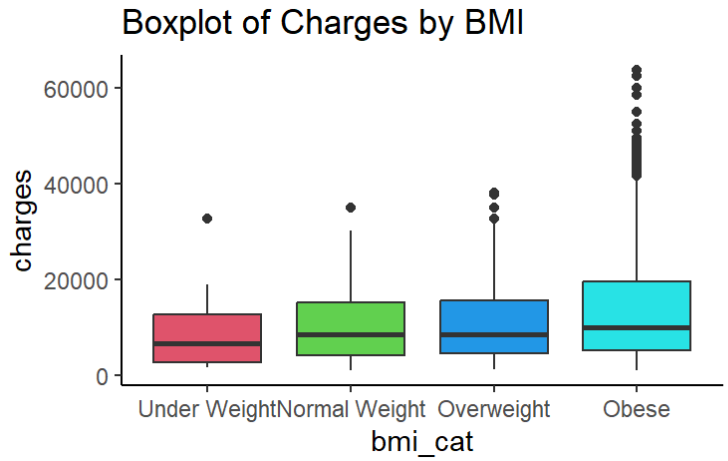
Observations :

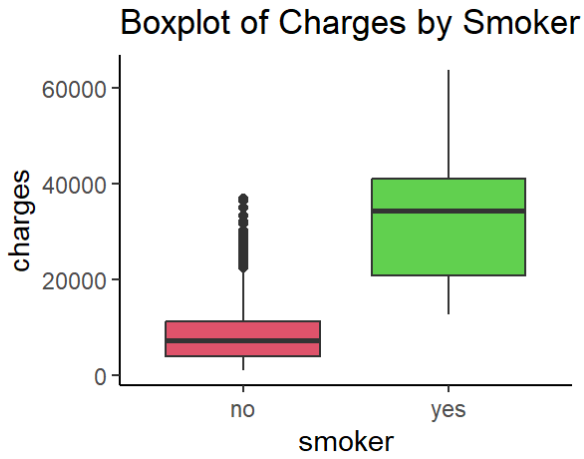
- The distributions of variables :
  - The age distribution of individuals is relatively the same, except for the 18 and 19 year olds who have a higher proportion.
  - The distribution of bmi is apparently normal centered around 30.
  - The distribution of charges is negatively asymmetric.
- We notice an effect of these variables on the charges, which we will explore more in depth later.
- No significant dependency between: age & bmi, smoker & bmi, age & smoker.

```
p1<-ggplot(data = insurance,aes(as.factor(children),charges)) + geom_boxplot(fill = c(2:
7)) +
  theme_classic() +  xlab("children") +
  ggtitle("Boxplot of Charges by Number of Children")
p2<-ggplot(data = insurance,aes(sex,charges)) + geom_boxplot(fill = c(2:3)) +
  theme_classic() + ggtitle("Boxplot of Charges by Gender")
p3<-ggplot(data = insurance,aes(bmi_cat,charges)) + geom_boxplot(fill = c(2:5)) +
  theme_classic() + ggtitle("Boxplot of Charges by BMI")
p4<-ggplot(data = insurance,aes(smoker,charges)) + geom_boxplot(fill = c(2:3)) +
  theme_classic() + ggtitle("Boxplot of Charges by Smoker")
plot_grid(p1, p2, p3, p4, rel_widths = c(1.25, 1),ncol = 2,
  align = "hv")
```





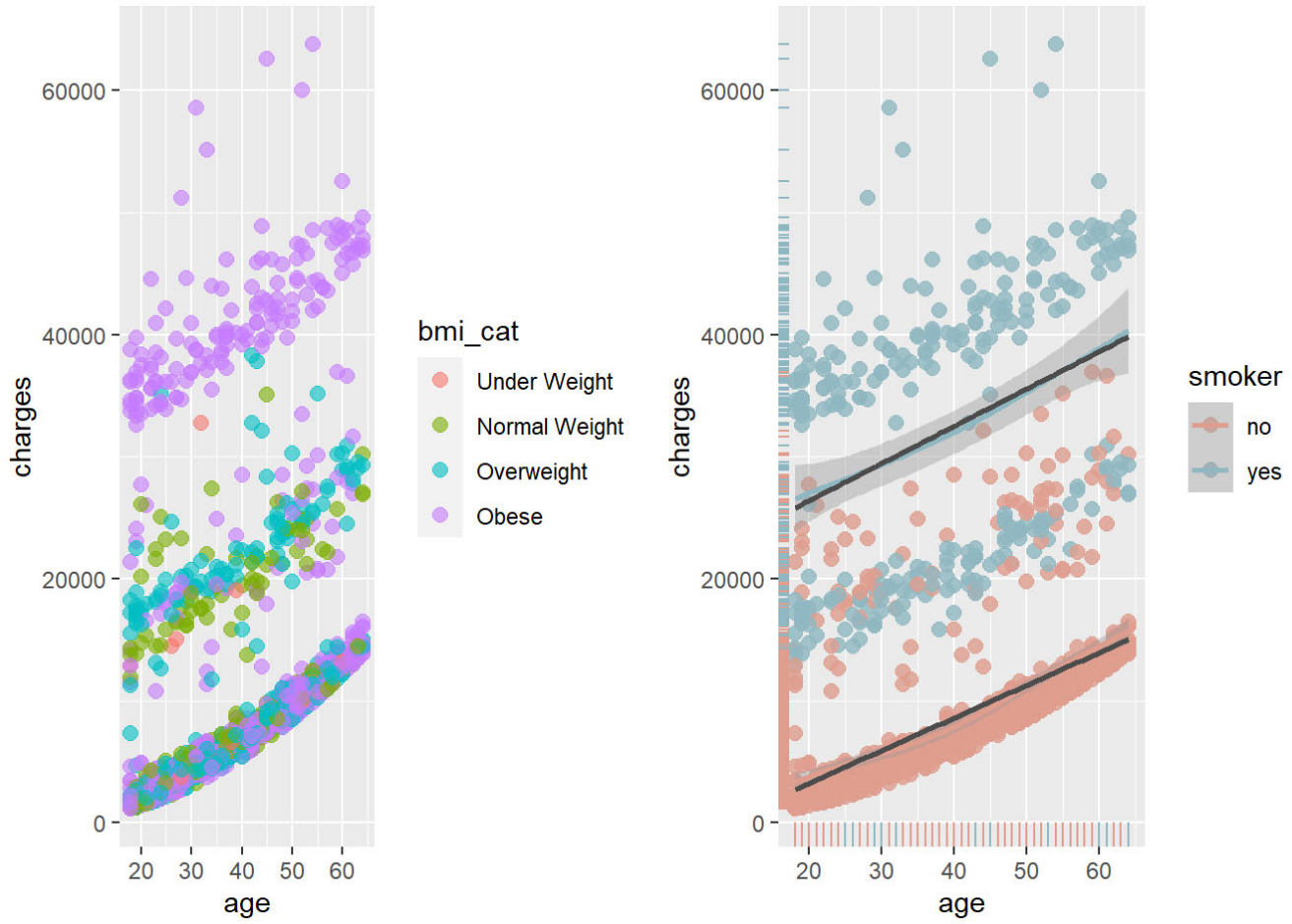




Interactions between age, bmi and smoking and their impact on medical charges

```
g1 <- insurance %>%
  ggplot(aes(x = age, y = charges, col = bmi_cat)) +
  geom_point(alpha = 0.6, size = 2.5)
g2 <- insurance %>%
  ggplot(aes(x = age, y = charges, col = smoker)) +
  geom_point(alpha = 0.8,size = 2.5) +
  scale_color_manual(values = c("#e09e8f", "#90b8c2")) +
  geom_rug() +
  geom_smooth() +
  geom_smooth(
    data = filter(insurance, smoker == "yes"),
    col = "grey30",
    method = lm,
    se = FALSE
  ) +
  geom_smooth(
    data = filter(insurance, smoker == "no"),
    col = "grey30",
    method = lm,
    se = FALSE
  )
grid.arrange(g1, g2, nrow = 1)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



The charges are linked to age by an almost linear relationship at three levels:

a first group which is characterized by the highest charges, it is completely made up of obese smoker individuals. a second group which is characterized by the lowest charges, it consists entirely of non-smoking individuals and a normal bmi distribution. and a third non-homogeneous group which requires more exploration.

We can also see that - for the three levels - the older the customers, the higher their charges.

Part 4 - Inference

Test the mean insurance charges of smoker and non-smoker

Hypothesis Question

H_0: *There is no different between the mean insurance charges between smoker and Non-smoker.*

H_A: *There is different between the mean insurance charges between smoker and Non-smoker.*

```
t.test(insurance[which(insurance$smoker=="yes"), "charges"],
       insurance[which(insurance$smoker=="no"), "charges"], alternative="two.sided", var.e
qual= TRUE)
```

```
##
##  Two Sample t-test
##
## data:  insurance[which(insurance$smoker == "yes"), "charges"] and insurance[which(insur
ance$smoker == "no"), "charges"]
## t = 46.665, df = 1336, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  22623.17 24608.75
## sample estimates:
## mean of x mean of y
## 32050.232  8434.268
```

With the p-value less than 0.05 we can reject the null hypothesis of equal mean charges between smoker and non-smoker and accept the alternative hypothesis.

Model for insurance charges using age, bmi and smoker

```
model <- lm(charges ~ age + bmi + smoker, data = insurance)
summary(model)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + smoker, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12415.4  -2970.9   -980.5   1480.0  28971.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11676.83     937.57  -12.45   <2e-16 ***
## age            259.55      11.93   21.75   <2e-16 ***
## bmi            322.62      27.49   11.74   <2e-16 ***
## smokeryes    23823.68     412.87   57.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6092 on 1334 degrees of freedom
## Multiple R-squared:  0.7475, Adjusted R-squared:  0.7469
## F-statistic:  1316 on 3 and 1334 DF,  p-value: < 2.2e-16
```

The linear model for predicting the score based on age, bmi and smoking status is given by:

charges = -11676.83 + 259.55(age) + 322.62(bmi) + 23823.68(smokeryes)

The multiple r-squared is 74.75%. We add another variable children to see if it will give a better proportion of variance for charges. so we add children.

Model for Charges using age, bmi, smoker and children.

```
final_model <- lm(charges ~ age + bmi + smoker + children, data = insurance)
summary(final_model)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + smoker + children, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12097.1  -2922.6   -950.7   1551.0  29566.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12093.32     947.78 -12.760  < 2e-16 ***
## age            258.08      11.91  21.665  < 2e-16 ***
## bmi            319.80      27.38  11.682  < 2e-16 ***
## smokeryes    23796.71     412.05  57.752  < 2e-16 ***
## children1      368.77     421.57   0.875 0.381868
## children2     1626.51     466.56   3.486 0.000506 ***
## children3      996.95     547.80   1.820 0.068997 .
## children4     2984.36    1239.60   2.408 0.016197 *
## children5      899.13    1453.36   0.619 0.536250
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6064 on 1329 degrees of freedom
## Multiple R-squared:  0.7508, Adjusted R-squared:  0.7493
## F-statistic: 500.4 on 8 and 1329 DF,  p-value: < 2.2e-16
```

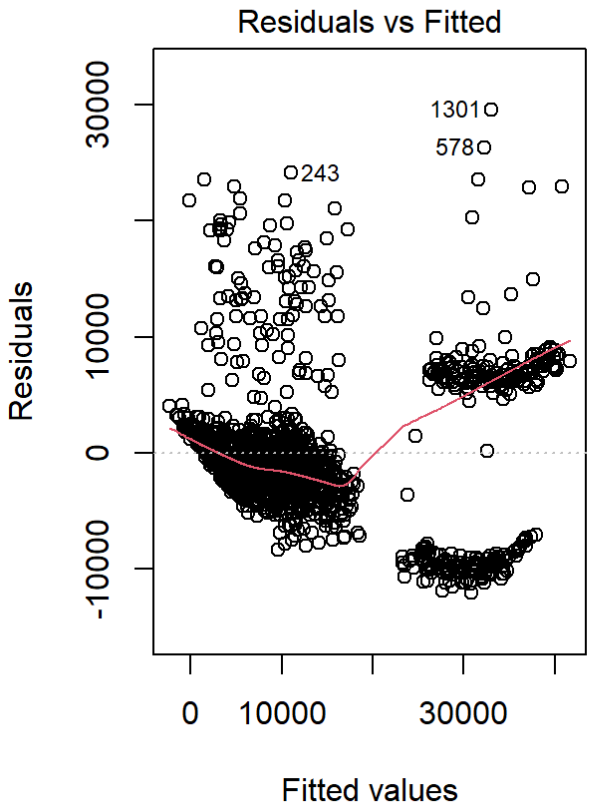
The linear model for predicting the score based on the final model that I settled on is given by:

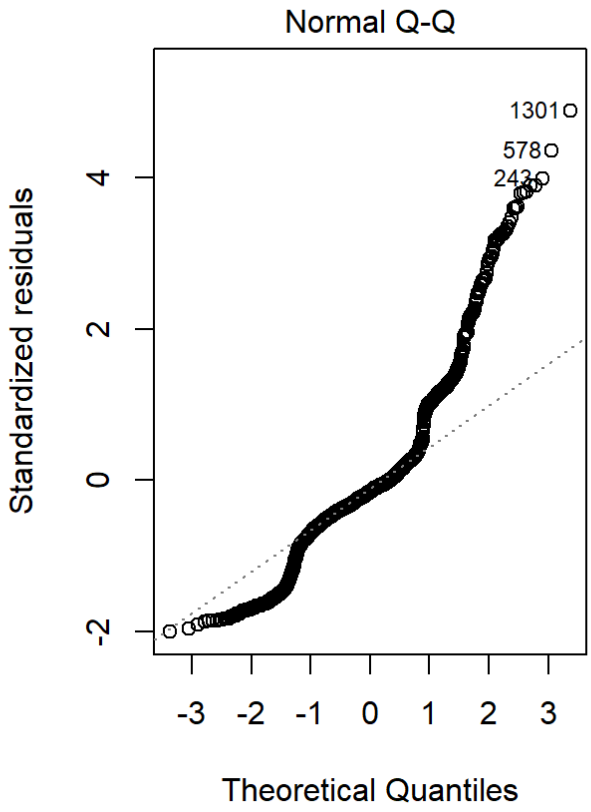
charges = -12093.32 + 258.08(age) + 319.80(bmi) + 23796.71(smokeryes) + 368.77(children1) + 1626.51(children2) + 996.95(children3) + 2984.36(children4) + 899.13(children5)

We get a Higher R-squared for Charges when we added Children to the variable. so, we choose the later as our final model

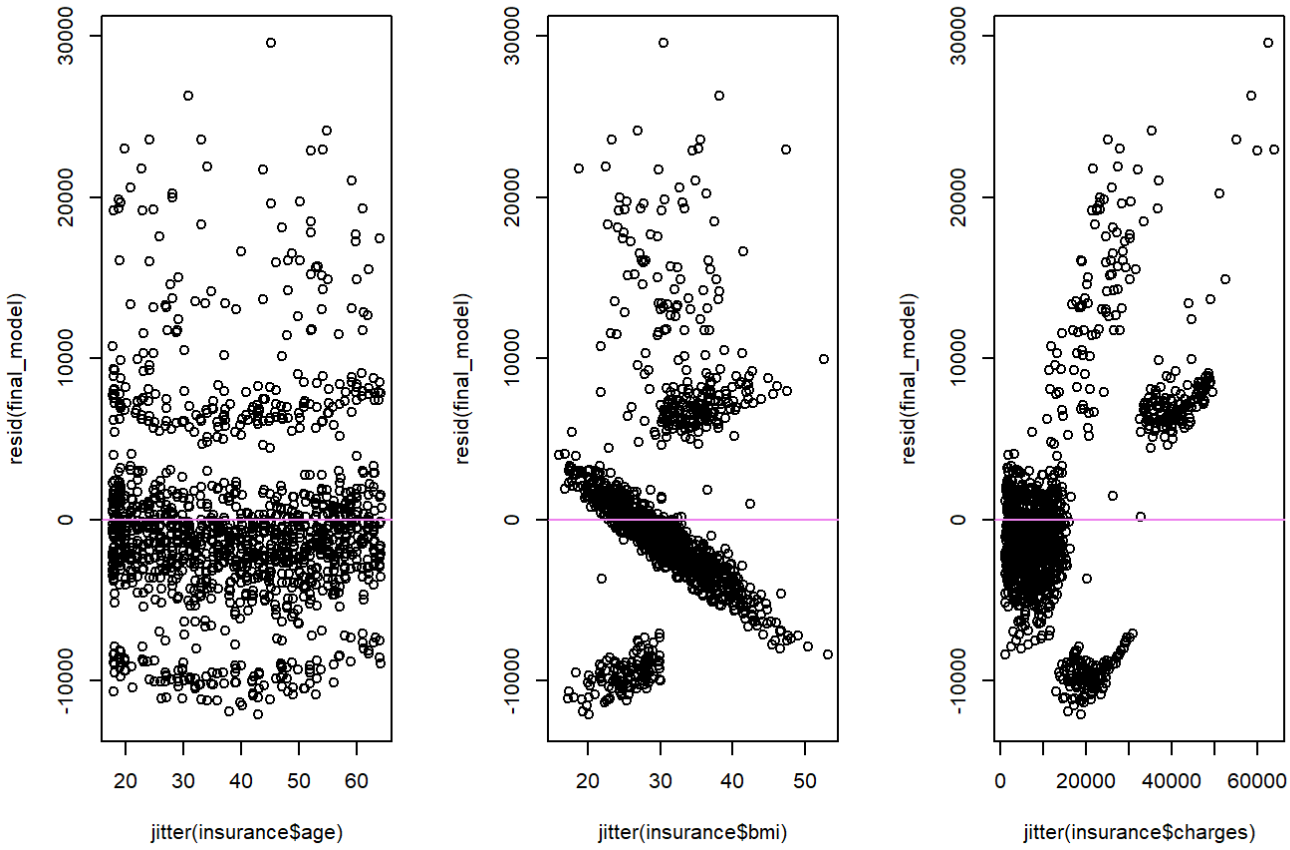
To test reliablility of linear model using diagnostic plots.

```
par(mfrow=c(1,2))
plot(final_model,c(1,2))
```

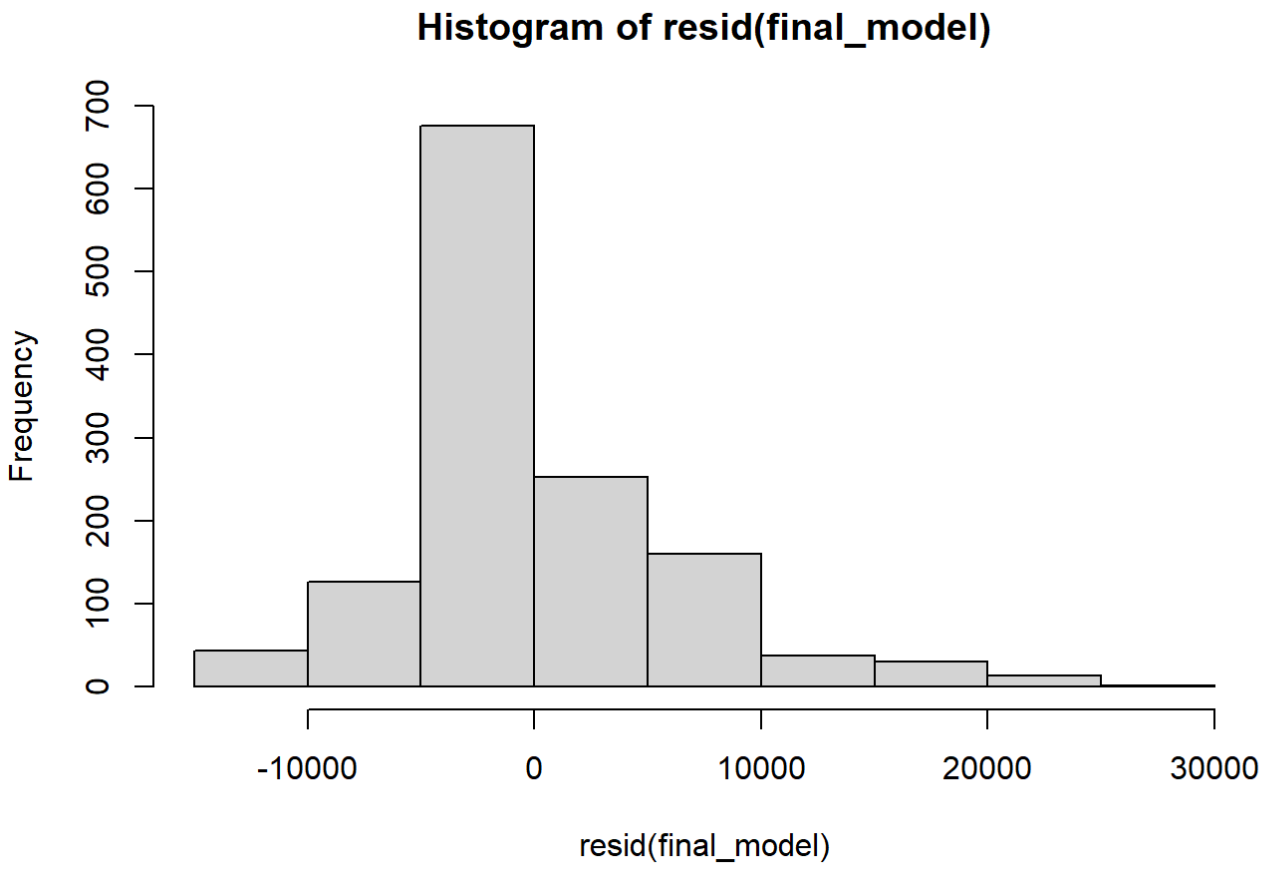





```
par(mfrow=c(1,3))
plot(jitter(insurance$age), resid(final_model))
abline(h=0, col="violet")
plot(jitter(insurance$bmi), resid(final_model))
abline(h=0, col="violet")
plot(jitter(insurance$charges), resid(final_model))
abline(h=0, col="violet")
```




```
hist(resid(final_model))
```



1. Linearty: For the quantitative variables `age` , `bmi` , `charges` : The residuals are most likely to be randomly dispersed, no obvious shapes or patterns are found.
2. Nearly normal residuals The histogram of the residuals shows a normal distribution. The qq plot shows the residuals are mostly line along on the normal line.The normal residual condiction is somewhat met.
3. Constant variability The majority of residuals are distributed between -1 and 1. The constant variability apprears to be met.

Based on the three observation above, the linear model is reliable.

Part 5 - Conclusion

From the Exploratory data analysis, it was discovered smoking status, bmi and age are the highest predictor of insurance charges. smokers, insurer with bmi>30 and high numbers of age are the greater contributor to insurance charges.Also, it was discover in the t-test to compare the mean charges of smoker and non-smoker that there is huge different between the mean insurance charges of smokers and non-smokers. We were able to see that the condition for the multiple regression plot is reasonable.