Exploratory Data Analysis - Final Project

Team 10
Prashanth Thirukkurungudi Sekar, Pravin Sundar,
Barathwaaj Parthasarathy, Bhavna Sinha

April 23, 2019

1 Introduction

Occurrence of expensive health conditions that account towards high medical costs are seldom and hard to predict. Premiums set by insurance companies has a direct dependence on a person's age, fitness, habits, history of illness etc. Through this project we will try to explore and identify key factors that affect a person's medical charges.

Research Question : What is the relationship between BMI and Medical charges for people of different age groups and smoking habits?
BMI is a person's weight in kilograms (kg) divided by his or her height in meters squared.[1]

2 Data Description

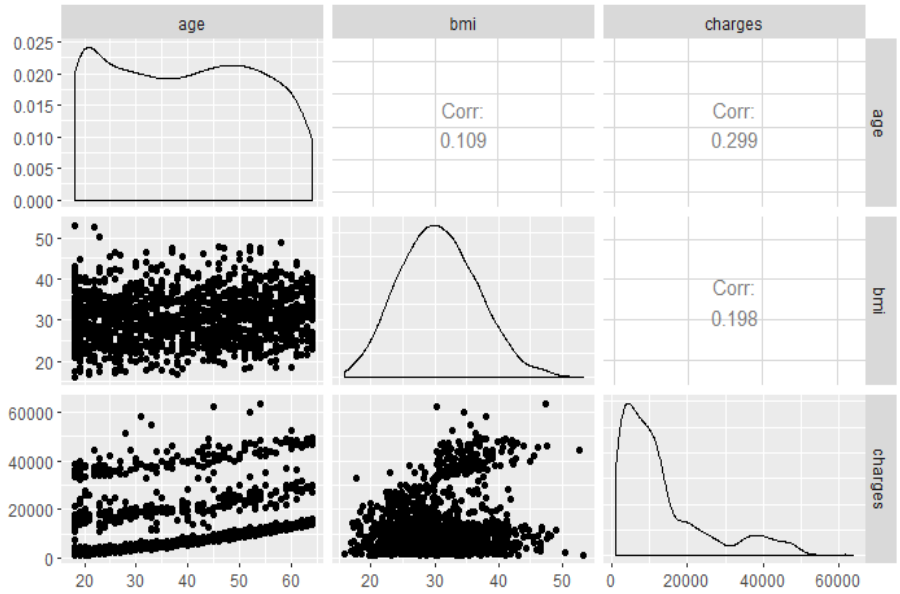
This dataset, simulated using U.S. Census Bureau, approximates to real-world conditions thus helping us understand the research question. The file contains 1,338 examples of beneficiaries' medical charges and their details for the calendar year [2] [3].

The features are:

1. **Age** - The person's Age
2. **Sex** - Gender Male or Female
3. **BMI** - The person's Body Mass Index
4. **No of children or dependents** - The number of children or dependents covered by the insurance plan
5. **Smoker** - Weather the person smokes or not
6. **Region** - Where does the person live (south East, South West, North East, North West)
7. **Health Charges** - The money spent in Dollars for health purposes over a calender year

3 Descriptive Analysis

Looking at the distribution of each variable of interest:



It can be observed that Age and BMI are quite well behaved but Medical charges seems to be right skewed. Going forward in the project, a log transformation of this variable will be used. Looking at the variation of age with charges, an increasing linear trend can be observed. In the variation of BMI with charges, two distinct clusters can be observed after a BMI value of 30.

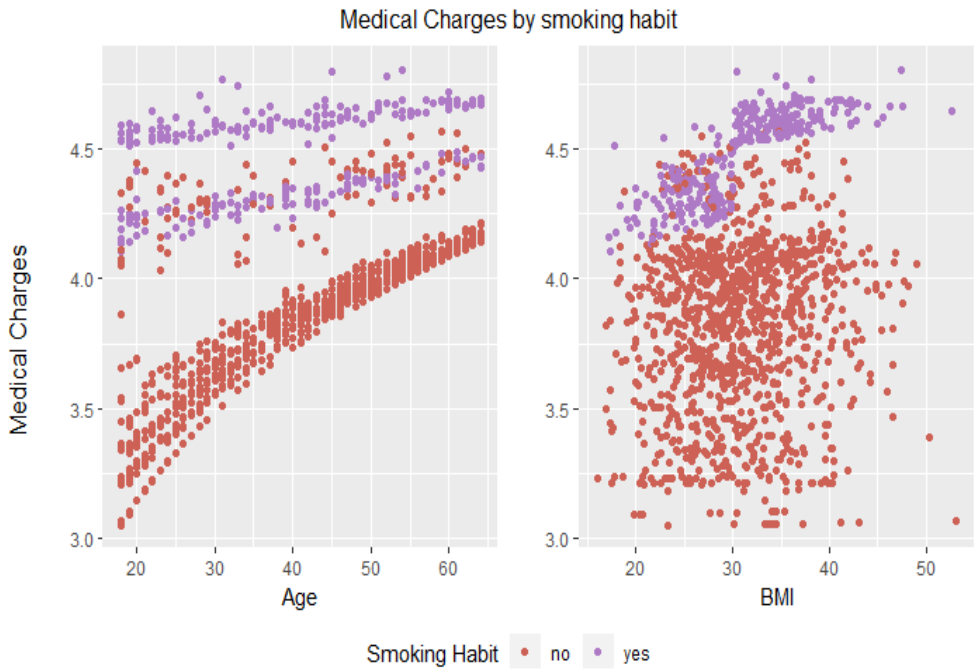
Before proceeding further with the descriptive analysis we will categorize "age" and "BMI" to understand the trend across various age and BMI groups respectively.
The following are the Age-Categories created:

- **Young Adult** - Ages between 18 and 35
- **Senior Adult** - Ages between 36 and 55
- **Elder** - Older than 56

The following are the BMI-Categories created [4]:

- **Under Weight** - BMI less than 18.5
- **Normal Weight** - BMI between 18.5 to 24.9
- **Over Weight** - BMI between 25 to 29.9
- **Obese** - BMI greater than 30

3.1 Charges vs Age and BMI



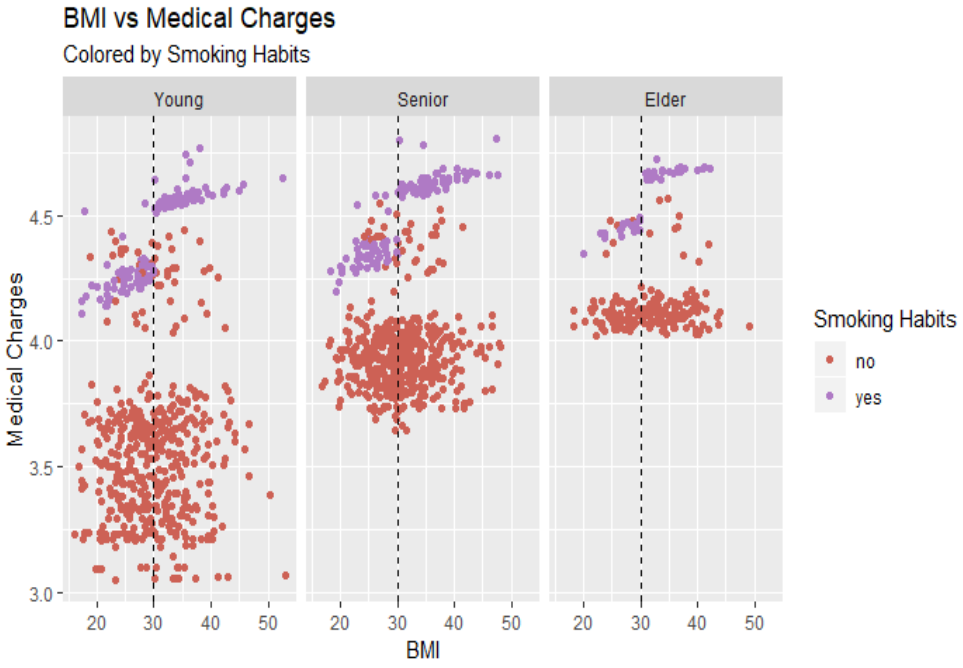
The medical charges for smokers are much higher than that of the non-smokers irrespective of their age (or) BMI.

In the graph on the left, for non-smokers, medical charges increases in a non-linear manner as age increases. While for people with smoking habits, it can be observed that there are two distinct clusters with a marginal increase in medical charges as age increases. The two distinct clusters can be due to their BMI values and this will be explored in detail in the subsequent analysis.

In the graph on the right, for non-smokers there is not any clear trend (or) pattern between medical charges and BMI. However for smokers, it can be observed that there is an increasing trend for medical charges with BMI.

We will now look at these variations by the categories created for age and BMI to extract some more interesting insights.

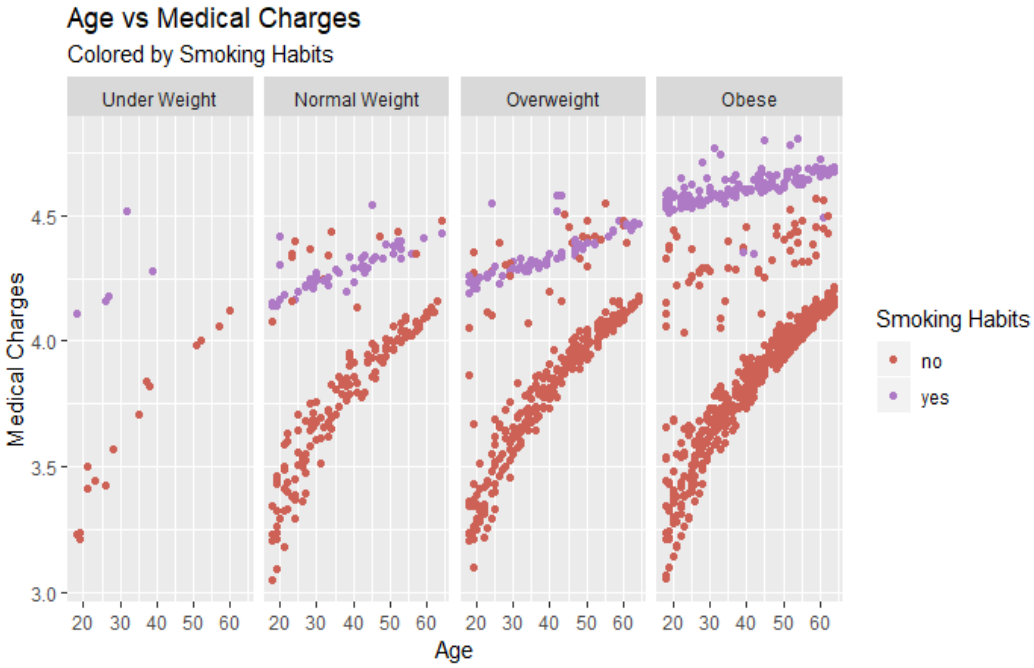
3.2 Charges v/s BMI by Age colored by Smoking Habits



In all the three age groups, it can be observed that smokers incur higher charges than non-smokers. Also, it can be observed that for non smokers there is an increase in medical charges with age categories. The young adults have the least charges while the elder people incur maximum charges.

From the scattered distribution with respect to BMI, it is apparent that charges do not increase with BMI for non-smokers belonging to an age category. However, for smokers, a spike in charges is seen in all three age groups when BMI is greater than 30, which marks the body mass index for obesity, as we can see two distinct clusters to the left and right side of the virtical line.

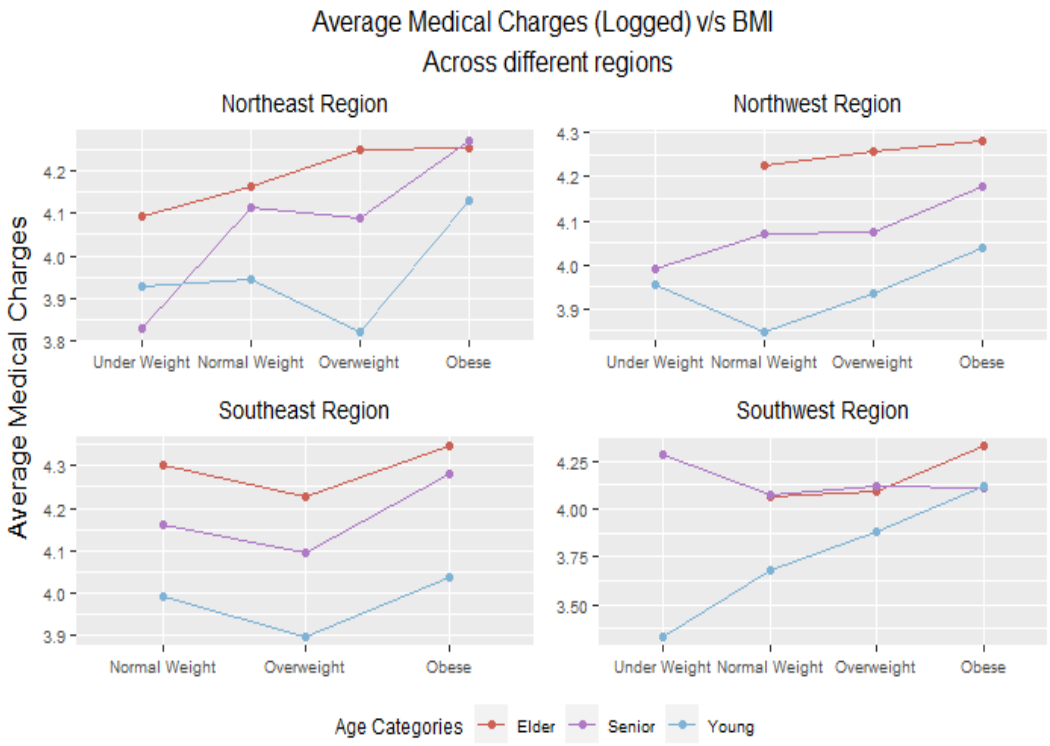
3.3 Charges v/s Age by BMI colored by Smoking Habits



We can see that there are very few data points in the under weight category and the number of data points increase as we go from one category to another.

From the plot, the increasing trend in the medical charges as the age increases for both smokers and non-smokers holds good in all the BMI categories.
It can be observed that among the four categories of BMI, smokers who are in the Obese category have higher medical charges as compared to the smokers from the other three BMI categories through all the ages.
This also explains two bands of smokers (data points) in the graph between age and medical charges that we observed in section 3.1.

3.4 Charges by Region



From the plots, it can be observed that there is no pattern in the plots other than the charges is maximum for Obese people in all the regions. Also we do not have any data for under weighted elders in Northwest and Southwest regions. It can also be seen that the region which incurred the least medical charges is Southwest. However, since the plots show no pattern or relationship of importance, it can be concluded that regions do not affect the medical charges.

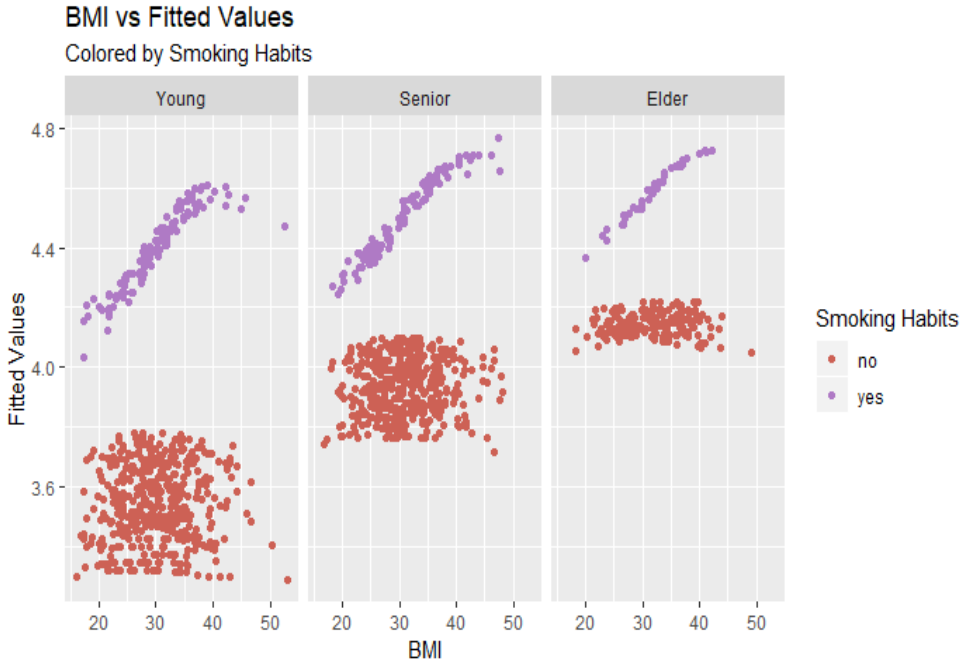
4 Predictive Analysis

We will explore multiple models to capture the maximum variance. Using the variables Age, BMI, Smoking habit and sex, we fit linear, GAM and loess models with medical charges (log) as the dependant variable. We included interactions between age, bmi, sex and smoking habits to capture the non linearity observed during the descriptive analysis.

The Linear and GAM Model captured about 80% of the variance in the data. On the other hand, the Loess model explained approximately 82% of the variance in the data.

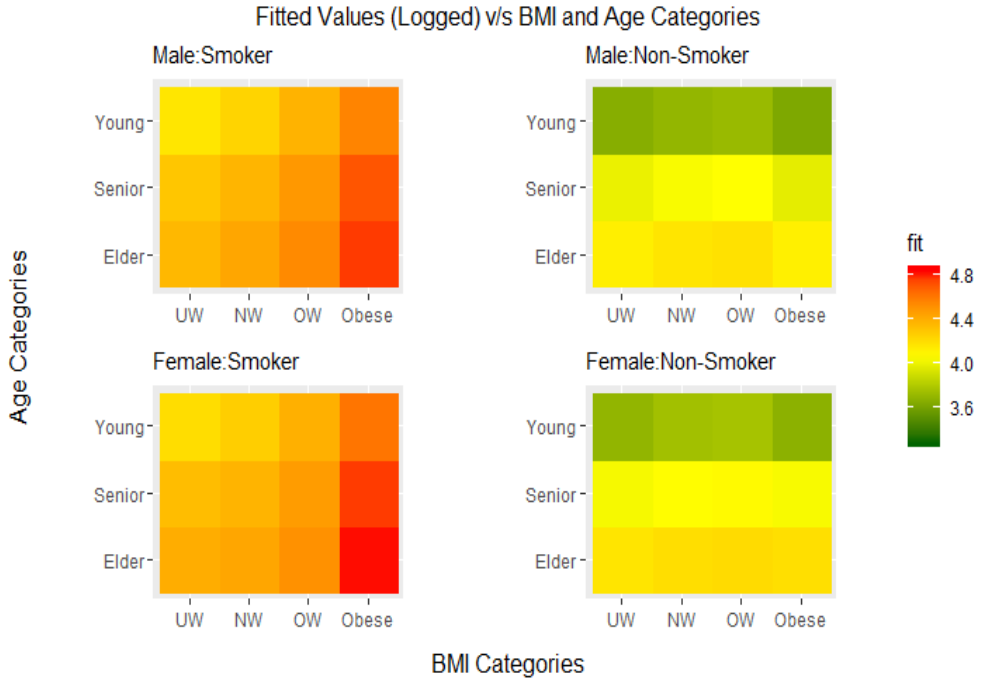
We will proceed with the loess model and explore the variation of its fitted values with the independent variables.

We can see that, the model illustrates what we observed in the data with a higher degree of assertion towards smoking habits, from the large gap seen between the medical charges for the respective categories. The fitted values clearly indicate the increasing trend between age and medical charges, with a little bit of non linearity.
In each of the BMI categories, we can see that for non smokers, the medical charges are peaked at around 4.2.
This is the minimum charge incurred for Normal and Over weighted smokers. And in the case of obese smokers the medical charges starts at around 4.4.



Looking at the fitted values against BMI across different age groups above, for a non smoker belonging to an age group the medical charges does not have a clear trend with BMI. But, for a smoker a clear non linear increase in medical charges with increase in BMI can be observed.

To understand the variation of medical charges a little more, we will look into the last variable, sex. Graph 1, in appendix shows the variation of medical charges with gender and smoking habits in the original data.
It can be observed that there is no difference in the medical charges for smokers but for non smokers, males tend to have higher charges compared to females.
We will try to look at differences in medical charges between the different age groups and BMI groups for males and females separately depending on their smoking habits.



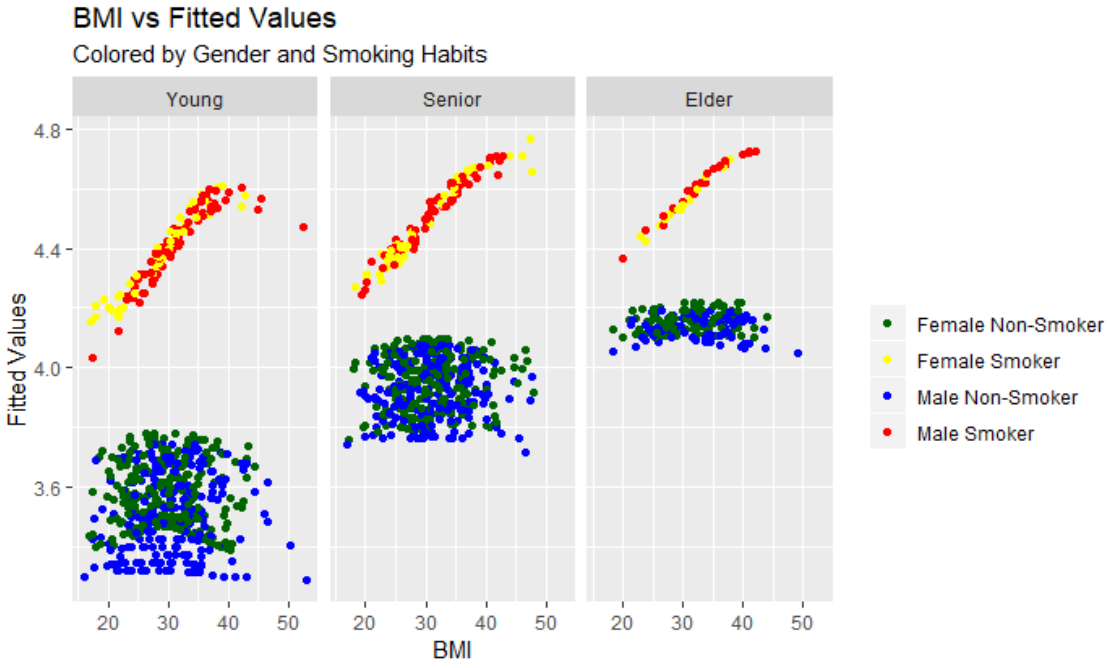
The above graph illustrates that the obese smokers irrespective of their gender incur the highest medical cost. We can see that there is not a lot of difference that can be observed between male and female within each category (smokers / non smokers). The right part of the graph shows that for non smokers, medical charges increases with increase in age and BMI.

To understand the effect of gender in the distribution of the fitted values, the smokers and non-smokers are categorized into male and female. The below two graphs shows the fitted values colored by the different combinations of smoking habits and gender.

It can be observed that the two genders are more or less spread evenly among all the categories with no unique trend separating the two.

It is worth noting that the fitted values of medical charges for male non smokers are lesser than that of female non smokers while the original data shows otherwise, which could be attributed towards overfitting.


AGE vs Fitted Values
Colored by Gender and Smoking Habits
Under Weight
Normal Weight
Overweight
Obese
Fitted Values
4.8
4.4
4.0
3.6
20 30 40 50 60
Age
Female Non-Smoker
Female Smoker
Male Non-Smoker
Male Smoker


BMI vs Fitted Values
Colored by Gender and Smoking Habits
Young
Senior
Elder
Fitted Values
4.8
4.4
4.0
3.6
20 30 40 50
BMI
Female Non-Smoker
Female Smoker
Male Non-Smoker
Male Smoker

5 Conclusion

A clear trend (or) relationship was discovered between BMI and Medical charges for people of different age groups and smoking habits. Some of the key inferences that were obtained through this analysis are :

- Medical charges increases as people's age and BMI value increases
- A person's smoking habit irrespective of the gender, results in increasing the medical charges when compared to a non-smoker

For future analysis, "number of dependents" (size of the family) and it's interactions with other variables can be considered as a factor to determine it's effect on medical charges. Also, history of illness plays a major role in impacting the medical charges of a person. An external append of this variable could help in determining more substantial factors that affect the medical charges that a person incurs.

References

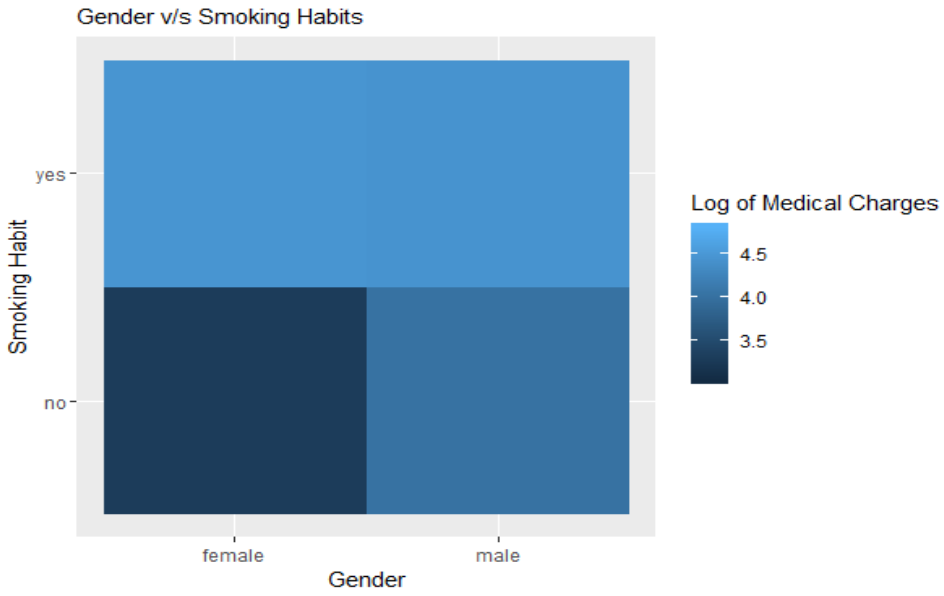
[1] MedicineNet - https://www.medicinenet.com/script/main/art.asp?articlekey=16125

[2] Machine Learning with R - https://edu.kpfu.ru/pluginfile.php/278552/mod_resource/content/1/MachineLearningR__Brett_Lantz.pdf

[3] Kaggle Data - https://www.kaggle.com/mirichoi0218/insurance

[4] American Cancer Society - https://www.cancer.org/cancer/cancer-causes/diet-physical-activity/body-weight-and-cancer-risk/adult-bmi.html

6 Appendix



Graph 1