Lawerence Addae

15/03/2024

FACTORS THAT AFFECT MEDICAL INSURANCE COSTS IN THE US

Abstract

This report explores the relationship between personal attributes, geographic factors, and medical insurance charges among 1338 US citizens using the Health Insurance Dataset. The study investigates variables such as age, gender, BMI, family size, smoking habits, and region, aiming to understand their impact on healthcare expenses and develop predictive models for estimating insurance costs. The analysis comprises five main tasks: visualization of variable distributions and summary statistics, testing the independence of predictor variables, linear regression modeling to describe influential variables, assessing differences in central tendencies based on insurance charges, and evaluating disparities in interval predictor variables across geographical regions. Findings from these analyses provide insights into the factors driving medical insurance charges.

Table of Contents

Introduction

Unlike many European countries, US citizens do not have access to free health care, albeit this has its own set of issues. As a result, many Americans rely on private firms, 66%, or public companies, 36% (Katherine Keisler-Starkey, 2022), for medical or health insurance.

“Health insurance is a contract between a company and a consumer. The company agrees to pay all or some of the insured person's healthcare costs in return for payment of a monthly premium.” (Kagan, 2023). However, some health insurance plans include annual coverage limits, meaning they will only pay up to a particular amount for medical bills during a given year. If these limits are exceeded, the patient may be liable for any additional expenses. However, many plans have abolished or increased these limits in response to Affordable Care Act (ACA) rules. This company concept is quite lucrative. It takes advantage of the lack of free health care and the worry of incurring an enormous American medical bill in the event of an incident. Out of 300.9 million persons with health insurance in 2021 (census), 21%, or 62 million, were injured. As a result, these corporations grabbed the remaining funds.

Although the likelihood of suffering an injury is quite modest, 18% of the 336,997,624 population in 2021 was injured (team, 2021). However, as the popular phrase goes, "Better safe than sorry".

However, in this research, the main focus will not be the worthiness of health insurance or the challenges associated with a non-free health care system, but on the factors that determine the amount charged for said insurance.

Description of dataset

While medical insurance is typically costly, certain factors can cause the monthly or annual premium to increase or decrease. This information can be extremely beneficial for individuals who require coverage or are considering purchasing coverage. For this investigation, three categorical factors were selected: gender (male or female), smoking status (yes or no), and area (northwest, northeast, southwest, and southeast); a continuous variable: BMI; and two discrete variables: age and number of children. 1338 people between the ages of 18 and 64 were chosen to perform this study, which used statistical analysis to assess the presence of any association between these variables and the amount charged.

Nevertheless, it is possible that the utilised data does not entirely capture the precise impact of every variable on the billed amount. This is because the number of individuals chosen is infinitesimally small when compared to the number of people with health insurance. As a result, the findings obtained may be erroneous and biased, implying that the outcome of the correlation between a variable and the amount charged cannot be predicted with high accuracy.

Descriptive analysis of the dataset

Table 1 displays the summary statistics for the numerical variables.

Table 1 - Summary statistics of the numerical variables

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max	SD
Age	18	27	39	39.21	51	64	14.05
BMI	15.96	26.30	30.40	30.66	34.69	53.13	6.10
Children	0	0	1	1.095	2	5	1.21

Charges	1122	4740	9382	13270	16640	63770	12110

Table 2 displays the frequency of categorical variables.

Table 2 - Categorical variables frequency

Variable	Male	Female	Smoker	Non-Smoker	Northeast	Northwest	Southeast	Southwest
Quantity	676	662	274	1064	324	325	364	325

Appendix I contains the distribution of every independent variable. Even though, according to the summary statistics in Table 1, the minimum age sampled is 18 and the maximum age is 64, which gives a wide spread of the age variable whose mean value is 39, it is evident from the bar chart in Figure() that there are more people between the ages of 18 and 23. This impacts the data and can lead to bad interference and biassed estimates.

Additionally, the dataset contains several outliers, as indicated by the substantial standard deviations (refer to Table 1). Outliers, or observations that differ dramatically from the rest of a dataset, can distort relationships between variables, resulting in inaccurate interpretations in regression or correlation studies. They can also lower the effectiveness of statistical tests and models by introducing noise and decreasing the power to identify actual effects. To deal with outliers once they are identified, they can be deleted or winsorized which is “replacing the smallest and largest values with the observations closest to them” (Hargrave, 2023).

Statement of problem

The purpose of this research is to examine the association between personal attributes, geographic factors, and medical insurance charges among 1338 US individuals. The study involves exploring changes in central tendencies across groups based on insurance charges, as well as examining variable dependencies, and comparing logistic regression to multiple linear regression for predicting healthcare expenses. Through standard statistical procedures, the study aims to draw valid conclusions from the findings.

Statistical analysis

Test for Independence

First, the premise of independent predictor variables is examined. Testing for independence of predictor variables, also known as testing for multicollinearity, is critical because when predictor variables are highly linked, it can be difficult to interpret the model's coefficients. High multicollinearity can result in unstable coefficient estimates, making it impossible to isolate the unique impacts of each predictor variable on the outcome. Different approaches were utilised to test this because not all factors are quantifiable. Pearson correlation was used to calculate the independence between numerical variables, which can be achieved using the following equation:

$$r = \frac{\sum(x_i - \bar{x})^2(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} \qquad \textit{Equation 1}$$

The p-value linked with the correlation coefficient (r) is determined through hypothesis testing. The null hypothesis (H_0) stipulates that the predictor variables are independent of the other variable (i.e., the correlation coefficient is zero), while the alternative hypothesis (H_a) states that the predictor variable is dependent on other variables (i.e., the correlation coefficient is not zero).

The test statistic for the correlation coefficient (r) follows a t-distribution under the null hypothesis, and the formula for the test statistic is:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \text{Equation 2}$$

The Pearson correlation was also used to determine the independence of categorical and numerical variables, although ordinal values were assigned to the categorical variables.
To evaluate the independence of two categorical variables, the chi-squared test was used, which is represented by equation 3.

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad \textit{Equation 3}$$

Table 3 - test for independence between predictor variables

Variable	Groups	P-values	Method
Age	Sex BMI Children Smoker Region	0.44591 6.19E-05 0.12049 0.36048529 0.938034	Pearson correlation Pearson correlation Pearson correlation Pearson correlation Pearson correlation
Sex	BMI Children Smoker Region	0.08998 0.53049 0.006548 0.9329	Pearson correlation Pearson correlation Chi-squared test Chi-squared test
BMI	Children Smoker Region	0.64101 0.89098503 6.86E-09	Pearson correlation Pearson correlation Pearson correlation
Children	Smoker Region	0.77915957 0.544805	Pearson correlation Pearson correlation
Smoker	Region	0.06172	Chi-squared test

The decision-making process assumes a minimum P-value (α) of 5%. Therefore, from Table 3 it can be determined that the null hypothesis was rejected by Age-BMI, Sex-Smoker, and BMI-Region. The p-values for each pair of variables are less than α, indicating a strong relationship between them. These connections support the hypothesis that younger people have a lower BMI because they engage in more physical activities, smoking is more common among men, and eating habits vary by location, all of which affect the BMI of the population. As a result, no variable will be left out when determining the dependency on health insurance charges.

Table 4 - independence between predictor variables and charges

Variable	Age	Sex	BMI	Children	Smoker	Region
P-value	4.89E-29	0.036133	2.46E-13	0.012852	8.27E-283	0.820518

In Table 4 the identical null hypothesis is evaluated, but this time the dependent variable was the amount charged, and the Pearson correlation was utilised for all predictor variables. In this scenario, the variables age, sex, BMI, children, and smoker all reject the null hypothesis. This means that the amount charged was determined by these variables. To determine which variables had a significant impact on insurance prices, the Pearson coefficient was used, which assesses the linear relationship between variables and outcomes. The matrix in Figure 1 demonstrates that the key effectors are the individual's age, BMI, and smoking status. This was further demonstrated by the scatter plots in Appendix II.

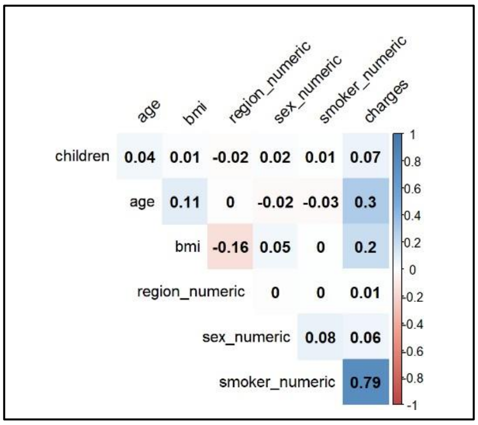


Figure 1 - Variable independence using Pearson correlation coefficient

Test for differences using t-test

The test for differences in central tendency, often known as the t-test, is used to determine whether two groups' means differ statistically significantly. In this scenario, the test is used to compare the mean of the variables in the 'High' and 'Low' charge groups, which are based on medical insurance charges.
The formula for the two-sample t test (also known as the Student's t-test) is provided below.

$$t = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{(s^2(\frac{1}{n_1}+\frac{1}{n_2}))}} \qquad \textit{Equation 4}$$

In this formula, t represents the t-value, x_1, and x_2 are the means of the two groups being compared, s_2 is the pooled standard error of the two groups, and n_1 and n_2 are the numbers of observations in each group. A higher t value indicates that the difference between group means is bigger than the pooled standard error, implying a more significant difference between the groups.
However, a few assumptions will be made to conduct this test:

- Observations within each charge group (high and low) should be independent of one another.
- Data variances within groups should be equal. If the variances are unequal, the test results may be unreliable, requiring corrections such as Welch's t-test.

The null hypothesis (H_0) is that the true difference in averages between groups High and Low is 0. The alternative hypothesis (H_a) instead is that the true difference differs from zero. For decision-making, α is considered to be 0.05.

Table 5 – Decisions on t-test findings

Variable	T-value	P-value	Decision
Age	21.822	< 2.2e-16	As the p-value is less than α, H_0 is rejected, indicating that the true difference in means between the groups is not equal to zero.
Sex	-0.10928	0.913	As the p-value is greater than α, H_0 is not rejected, indicating that the true difference in means between the groups is equal to zero.
BMI	3.2992	0.000995	As the p-value is less than α, H_0 is rejected, indicating that the true difference in means between the groups is not equal to zero.
Children	0.70289	0.4822	As the p-value is greater than α, H_0 is not rejected, indicating that the true difference in means between the groups is equal to zero.
Smoker	21.526	< 2.2e-16	As the p-value is less than α, H_0 is rejected, indicating that the true difference in means between the groups is not equal to zero.
Region	1.7333	0.08327	As the p-value is greater than α, H_0 is not rejected, indicating that the true difference in means between the groups is equal to zero.

The statistical analysis in Table 5 used a 95 percent confidence interval to estimate the difference in averages between the 'High' and 'Low' groups based on medical insurance charges. These findings further emphasise the importance of age, BMI, and smoking as determinants of medical insurance charges, with older people, those with higher BMIs, and smokers typically paying more than their peers.

Test for differences using ANOVA

ANOVA is another technique for detecting differences. The acronym ANOVA stands for Analysis of Variance. ANOVA is a statistical test created by Ronald Fisher in 1918. Simply expressed, it

determines whether there are statistical differences between the means of three or more independent groups. In this case, it is utilised to determine whether there are statistically significant differences in the central tendency of interval predictor variables, such as BMI and Age, in relation to the geography variable (region).
This is done by "assessing the relative size of variance among group means (between group variance) compared to the average variance within groups (within group variance)" (Kim, 2014). ANOVA calculates an F-statistic, which compares the variation in group means to the variation within groups. If the variation between group means is substantially greater than the variation within groups, the F-statistic will be high, suggesting that the means of at least one pair of groups differ significantly.

The following assumptions will be made:

- Data variances within each group should be equal. If the variances are unequal, the test findings may be unreliable, requiring corrections.
- Residuals (differences between observed and anticipated values) should be independent. To ensure the model captures all essential information, the residuals should not show any systematic trend.
- Observations within each group (specified by geography) should be independent of one another.

The null hypothesis (H_0) in this scenario is that the mean values of the interval predictor variables, age and BMI, are identical across all areas (northwest, northeast, southwest, and southeast). The alternative hypothesis (H_a) states that at least one pair of interval predictor variables' mean values differs considerably across regions. Again, α is assumed to be 0.05.

Table 6 - ANOVA test results

Variable	F value	Pr(>F)	Decision
Age	0.08	0.971	As the p-value is above α, H_0 is rejected. Therefore, the mean values of the interval predictor variables are equal across all regions
BMI	39.49	<2e-16	As the p-value is below α, H_0 is rejected. Therefore, it can be concluded that there are significant differences in the mean values of the interval predictor variable across the regions.

In summary, Table 6 shows that there are no significant differences in age between areas, however, there are substantial disparities in BMI. This implies that regional factors influence BMI but not necessarily age in the dataset. Tukey's HSD test will be used to conduct more tests to support this claim.

Post-hoc test

Post-hoc tests are statistical procedures that are used after an initial analysis, such as ANOVA, to make particular comparisons among groups. When ANOVA returns a significant result, it indicates that there is a difference between the groups, though, it does not explain which groups are different. Post-hoc tests can help to uncover these unique discrepancies. For example, in this scenario, the ANOVA test for BMI indicated significant differences between regions. As a result, a post-hoc test can be performed to determine which pairs of locations have significantly different BMIs.
There are various types of post-hoc tests, the most common of which is Tukey's HSD (Honestly Significant Difference). Tukey's HSD test analyses all possible pairs of means and accounts for multiple comparisons.
This is accomplished by:

1. Calculating the difference between the means of all conceivable group pairs.
2. Determining the standard error of the differences.
3. Creating confidence intervals for each difference.
4. Comparing differences to the crucial value (based on the number of groups and desired confidence level).
5. Statistical significance is determined when the difference between two groups' means exceeds the critical value.

The main advantage of post-hoc testing is a more detailed understanding of group differences than the omnibus ANOVA test. However, it is critical to recognize the increased likelihood of Type I error due to multiple comparisons, so adjustments, such as the Bonferroni correction, can be used to control the familywise error rate.

The null and alternative hypotheses for this test are similar to those used in the ANOVA test. As a result, H_0 states that there is a substantial difference in the variables, age or BMI, between a pair of regions, whereas H_a states that there is no difference in the variables between a pair of regions. The P-value (α) threshold is 0.05.

Table 7 - Post-hoc test of age differences in regions

Regions	Diff	P Adj	Decision
Northwest-Northeast	-0.071595	0.9999022	Since the p-value is greater than 0.05, H_0 is rejected. It can be deduced that there is no significant difference in age between the regions.
Southeast-Northeast	-0.328958	0.9900359	Since the p-value is greater than 0.05, H_0 is rejected. It can be deduced that there is no significant difference in age between the regions.
Southwest-Northeast	0.186866	0.9982766	Since the p-value is greater than 0.05, H_0 is rejected. It can be deduced that there is no significant difference in age between the regions.
Southeast-Northwest	-0.257363	0.9951516	Since the p-value is greater than 0.05, H_0 is rejected. It can be deduced that there is no significant difference in age between the regions.
Southwest-Northwest	0.258462	0.9954741	Since the p-value is greater than 0.05, H_0 is rejected. It can be deduced that there is no significant difference in age between the regions.
Southwest-Southeast	0.515824	0.9633994	Since the p-value is greater than 0.05, H_0 is rejected. It can be deduced that there is no significant difference in age between the regions.

Table 8 - Post-hoc test of BMI differences in regions

Regions	Diff	P Adj	Decision
northwest-northeast	0.026282	0.9999328	Since the p-value is greater than 0.05, H_0 is rejected. It can be deduced that there is no significant difference in BMI between the regions.
southeast-northeast	4.182486	0.0000000	Since the p-value is lower than 0.05, H_0 is not rejected. It can be deduced that there is a significant difference in BMI between the regions.
southwest-northeast	1.423112	0.0106965	Since the p-value is lower than 0.05, H_0 is not rejected. It can be deduced that there is a significant difference in BMI between the regions.

southeast-northwest	4.156204	0.0000000	Since the p-value is lower than 0.05, *H*$_0$ is not rejected. It can be deduced that there is a significant difference in BMI between the regions.
southwest-northwest	1.396831	0.0127393	Since the p-value is lower than 0.05, *H*$_0$ is not rejected. It can be deduced that there is a significant difference in BMI between the regions.
Southwest-southeast	-2.759374	0.0000000	Since the p-value is lower than 0.05, *H*$_0$ is not rejected. It can be deduced that there is a significant difference in BMI between the regions.

These post-hoc results, presented in Tables 7 and 8, give particular comparisons between areas for both Age and BMI variables. The findings confirm that there is no age difference among the participants in the dataset across all four locations. Furthermore, the disparities in BMIs between regions are also confirmed:

- The BMI in the Southeast region is significantly higher compared to the Northeast region.
- The BMI in the Southwest region is significantly higher compared to the Northeast region.
- The BMI in the Southeast region is significantly higher compared to the Northwest region.
- The BMI in the Southwest region is significantly higher compared to the Northwest region.
  The BMI in the Southwest region is significantly lower compared to the Southeast region.

As previously stated, this could be due to differences in eating habits, or it could be due to the natural topography, such as mountains, that prompts the inhabitants of certain places to be more physically active than people in other locations.

Regression analysis

Regression analysis is a fundamental statistical approach for identifying and quantifying correlations between variables in datasets. It lets researchers to understand how changes in independent variables affect changes in a dependent variable. Fitting regression models allows analysts to assess the influence of many factors on the outcome of interest and get useful insights into the underlying processes driving the observed data. Regression analysis has a wide range of applications, including understanding the causes of healthcare expenditures. It is an important tool for making data-driven decisions and testing hypotheses.

The formula for linear regression can be expressed as:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon \tag*{\textit{Equation 5}}$$

Where:

- y is the dependent variable.
- x_1, x_n are the independent variables.
  β_0, β_1 are the coefficients of the model, representing the effect of each independent variable on the dependent variable.
- ε is the error term, which is the difference between the observed values of the dependent variable and the values predicted by the model.

The purpose of linear regression is to determine the coefficient values that minimise the sum of squared errors between the observed and predicted values of the dependent variable. The least squares method is commonly used to do this.

As previously tested, the BMI and age variables are correlated, hence a PLS regression is applied. PLS regression is particularly beneficial when the predictors are multicollinear. The primary purpose of PLS regression is to maximise the covariance between the predictor and responder variables by dividing both matrices into a collection of latent variables or components. These components capture the common information between the predictor variables and the response variable, which is then used to construct a predictive model.

The Kernel Partial Least Squares (kernelPLS) regression is employed in this case. It improves the standard Partial Least Squares (PLS) regression method by operating in a higher-dimensional feature space using kernel functions. This transformation enables more flexible modelling of nonlinear interactions between predictors and response variables. Unlike traditional PLS regression, which assumes linear correlations, kernelPLS may detect complicated interactions and nonlinearities in data.

Results and findings

The linear regression model's null hypothesis states that a change in the variable results in a statistical difference in the cost of health insurance, whilst the alternative hypothesis is that a change in the variable does not results in a statistical difference in the amount charged.

Table 9 - Linear regression model

Variable	Estimate	Pr(>|t|)	Decision
Age	259.5475	5.24E-90	The p-value is lower than α, therefore the null hypothesis is not rejected. Meaning that the variable has a statistically significant effect on the cost.
BMI	322.6151	2.42E-30	The p-value is lower than α, therefore the null hypothesis is not rejected. Meaning that the variable has a statistically significant effect on the cost.
Smoker	23823.68	0	The p-value is lower than α, therefore the null hypothesis is not rejected. Meaning that the variable has a statistically significant effect on the cost.

Table 10 - PLS model results

Variables	Age	BMI	Smoker
X	83.68	99.92	100.00
Charges	10.16	12.43	74.75

Table 9 shows that all three variables (age, BMI, and smoking status) have a linear relationship with the amount charged. This means that, for example, if age increases and the other variables remain constant, the cost of insurance rises. Specifically, the coefficient for "age" is 259.5475, suggesting that the model forecasts an increase in charges of around $259.55.

The data shown in Table 10 indicate the variance. The variance of the predictor variables (X) reflects how effectively the model captures their variability. In contrast, for the response variable (charges), the variation explained by each component reveals how well the model predicts the variability in the response variable based on the predictors. Higher percentages suggest that the model performs better in predicting the response variable.

Variance is a crucial subject in statistics and data analysis because it provides vital information about data dispersion as well as the performance of statistical models in explaining and forecasting variability. In PLS regression, variance is used to assess the model's ability to capture variability in both predictors and response variables.

Conclusion

Based on the comprehensive analysis of the available datasets and the application of diverse statistical methodologies, this investigation highlights the pivotal factors influencing the escalation of health insurance charges. Notably, age, BMI, and, significantly, smoking habits emerge as the primary drivers. The statistical significance of these predictor variables, indicated by their small p-values falling below the designated threshold, underscores their substantial impact on insurance costs.

References

Hargrave, M., 2023. *Investopedia.* [Online]
Available at: https://www.investopedia.com/terms/w/winsorized_mean.asp
[Accessed 29 February 2024].

Kagan, J., 2023. *Investopedia.* [Online]
Available at: https://www.investopedia.com/terms/h/healthinsurance.asp
[Accessed 26 February 2024].

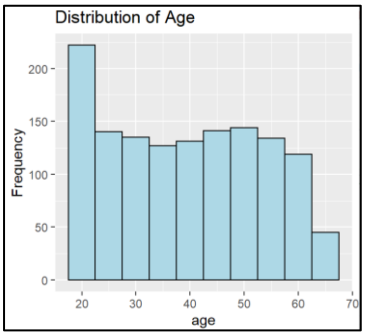
Katherine Keisler-Starkey, L. N. B. a. R. A. L., 2022. *Census.* [Online]
Available at: https://www.census.gov/library/publications/2023/demo/p60-281.html#:~:text=Highlights,91.7%20percent%20or%20300.9%20million
[Accessed 26 February 2024].

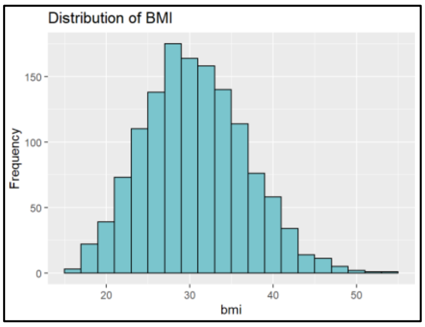
Kim, H.-Y., 2014. *National library of medicine.* [Online]
Available at:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3916511/#:~:text=The%20ANOVA%20method%20assesses%20the,groups%20(within%20group%20variance)
[Accessed 9 March 2024].

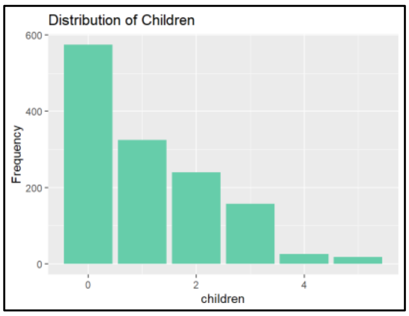
team, M., 2021. *Macrotrends.* [Online]
Available at: https://www.macrotrends.net/global-metrics/countries/USA/united-states/population
[Accessed 26 February 2024].

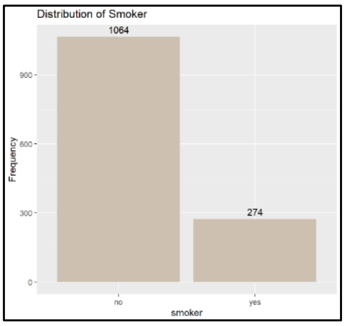
Appendices

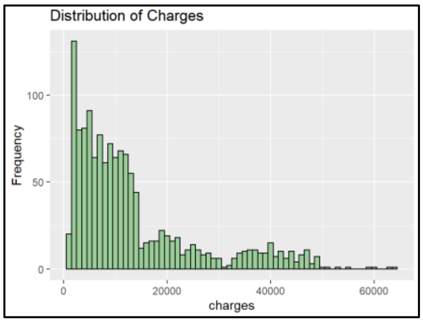
Appendix I – Distribution of the variables

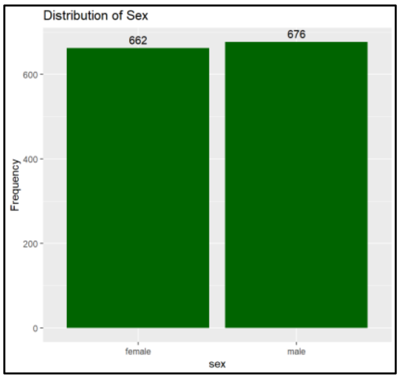


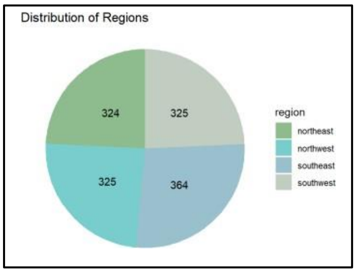




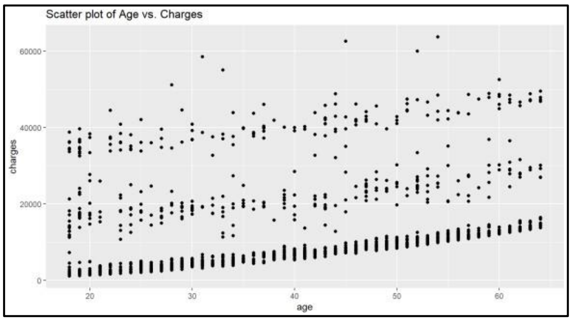


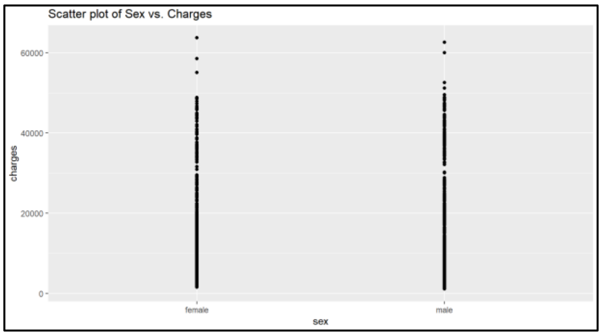


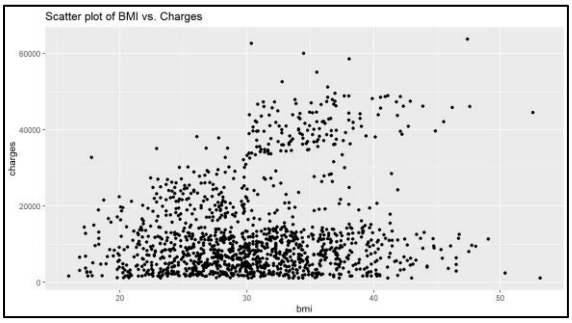


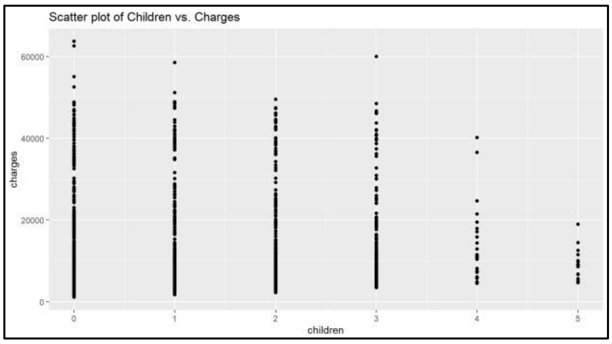


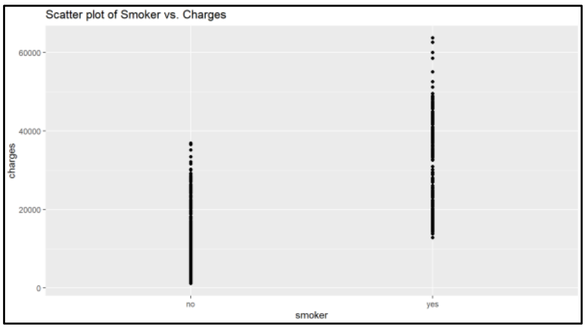
Appendix II – Variables independence

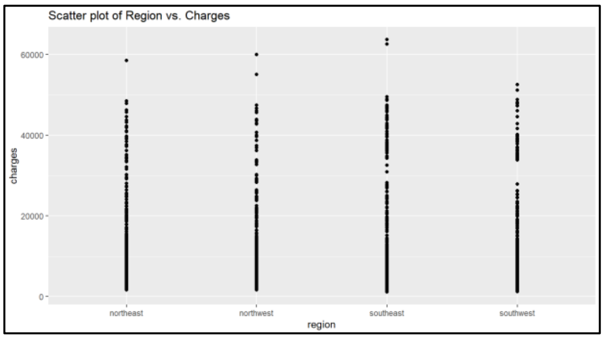

Scatter plot of Age vs. Charges
charges
age


Scatter plot of Sex vs. Charges
charges
female
male
sex


Scatter plot of BMI vs. Charges
charges
bmi


Scatter plot of Children vs. Charges
charges
children


Scatter plot of Smoker vs. Charges
charges
no
yes
smoker


Scatter plot of Region vs. Charges
charges
northeast
northwest
southeast
southwest
region