STATISTICAL DATA ANALYSIS & VISUALIZATION

DETERMINING FACTORS FOR HEALTH INSURANCE CHARGE

BY

Ene Joyce Ojaide

ABSTRACT

An important research topic in healthcare policy and economics is identifying factors that impact health insurance premiums. Policymakers, insurers, and healthcare providers must thoroughly understand the main elements that cause insurance premium differences to make educated decisions and create successful healthcare finance and access strategies. Using a large dataset including demographics, lifestyle variables, and medical history, this study intends to examine the factors that impact health insurance premiums. This study uses advanced statistical approaches like regression analysis and nonparametric statistics to explore the impact of key determinants of health insurance charges on premium pricing. The research findings offer useful insights for healthcare industry stakeholders looking to improve affordability, accessibility, and equity in healthcare coverage while also adding to our understanding of the complex mechanisms impacting health insurance prices.

Contents

Introduction

Today's complex healthcare funding structure requires health insurance to pay medical expenses. Health insurance charge issues concern economists, policymakers, insurers, and consumers. Sustainable healthcare financing and equitable healthcare access need an understanding of health insurance premium determinants. Company and individual health insurance premiums cover medical expenditures. Due to risk profiles, healthcare use, and health conditions, households, demographic groups, and people pay various amounts. Demographics, socioeconomic position, lifestyle, medical history, area, coverage options, and laws affect health insurance premiums.

Age, gender, marital status, and family size affect health insurance costs. Due to higher healthcare use, older people and households with more dependents pay higher premiums. Insurance expenses are higher for women due to healthcare use, reproductive health, and maternity care. Income, education, and jobs affect health insurance costs. Higher-income and more educated people may pay more for health coverage and benefits. Employee benefits packages in numerous sectors include cheaper group insurance. Smoking, drinking, and exercise affect health insurance costs. Smoking and unhealthy habits raise chronic illness risk and healthcare expenses. Preventive and healthy living insurance policies may offer wellness incentives and lower premiums. Medical history and pre-existing conditions affect health insurance costs.

Geographical location affects health insurance costs. Regional insurance premiums depend on healthcare expenses, provider networks, state laws, and competition. Rural or neglected areas with less care may have lower insurance rates than urban areas with higher healthcare expenses and provider concentration. Coverage and plan components affect health insurance costs. Individual, employer-sponsored, Medicare, Medicaid, and marketplace exchange coverage are available. This research studies demographic, socioeconomic, lifestyle, medical, regional, and policy characteristics to help policymakers, insurers, healthcare providers, and consumers understand insurance pricing dynamics and make evidence-based healthcare finance and policy decisions.

Statement of Problem

The study seeks to examine the characteristics that affect health insurance premiums for individuals, including age, BMI, smoking status, gender, number of children, and region. The problem statement aims to assess the degree to which these factors influence variations in insurance fees, providing insight into the discrepancies in healthcare expenses. The study aims to analyse the correlation between these variables and insurance charges to gain an improved grasp of the factors that influence the affordability and accessibility of healthcare. This will help inform policy decisions and interventions to reduce financial burdens and ensure fair access to healthcare services.

Description of Dataset

The health insurance dataset, which is named Health-Insurance-Dataset.csv, includes information on the relationship between personal characteristics (such as age, gender, body mass index, family size, and smoking habits), geographic location, and the impact that these factors have on medical insurance charges (in dollars) for a sample of 1338 people of the United States. It can be utilized to investigate how these characteristics impact insurance costs and to construct prediction models to estimate healthcare expenses. The Predictor variables, which entail age, gender (sex), body mass index (BMI), family size (children), smoking habits (smoker), and geographic location (region), are a mixture of numerical variables and categorical variables. Thus, to describe the dataset, the numerical variables need to be described using their summary statistics and the appropriate visualizations suitable for numerical variables (for example boxplots, histograms, etc). for the categorical variables, the frequency distribution table, which also includes proportion and percentage, would suffice as descriptive statistics of the data.

Data Visualization and Summary Statistics of Dataset

Table 1: Descriptive Statistics of Numerical Variables with the Response Variable.

	Min	1st Quartile (Q1)	Median	Mean	3rd Quartile (Q3)	Max	Standard Deviation	Skew ness	Kurtosis
Age	18.00	27.00	39.00	39.21	51.00	64	14.05	0.06	1.76
BMI	15.96	26.30	30.40	30.66	34.69	53.13	6.10	0.28	2.94
Children	0.00	0.00	1.00	1.10	2.00	5.00	1.21	0.94	3.20
Charges	1122.0	4740.0	9382.0	13270.0	16640.0	63770.0	12110.01	1.51	4.60

The above shows the summary statistics of the numeric variables in the dataset. Their visualizations are in Appendix 2 to Appendix 6

Table 2: Description of Categorical Predictor Variables

Variables	Frequency (N)	Proportion	Percentage (%)
Sex			
Male	676	0.505	50.5
Female	662	0.495	49.5
Total	**1338**	**1.000**	**100.0**
Smoker			
Yes	274	0.205	20.5
No	1064	0.795	79.5
Total	**1338**	**1.000**	**100.0**
Region			
Northeast	324	0.242	24.2
Northwest	325	0.243	24.3
Southeast	364	0.272	27.2
Southwest	325	0.243	24.3
Total	**1338**	**1.000**	**100.0**

The visualization of Table 2 is in Appendix 1.

STATISTICAL ANALYSIS

To conduct a predictive analysis on our dataset, it is pertinent to test some assumptions to see if our dataset suits these assumptions for the predictive analysis. One model used in predictive analysis is the regression model.

For this report, the correlation matrix and correlation plots are used to check some of these assumptions for the numerical variables, while the Chi-square test would be used to test for the independence of the categorical variables. The visualization of the numerical variables is seen in Appendix 7 and Appendix 8. The correlation matrix is seen in Table 3, while the Chi-square test for the categorical variables is seen in Table 4. The formula for the Pearson moment correlation is given by,

$$\rho = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2(Y - \bar{Y})^2}}$$

Also, the Pearson Chi-Square test is given by,

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i^2}$$

Where,

o_i is the observed frequency for each category or cell in the contingency table,

e_i is the expected frequency for each category or cell in the contingency table under the null hypothesis of independence,

The sum is taken over all categories or cells in the contingency table.

The expected frequency e_i is calculated as: $e_i = \frac{row\ total \times column\ total}{grand\ total}$.

The regression model is given by

$$charges = \delta_0 + \delta_1 Age + \delta_2 BMI + \delta_3 children + \delta_4 Sex(female) + \delta_5 Smoker(no) + \delta_6 Region(northwest) + \delta_7 Region(northeast) + \delta_8 Region(southeast) + \varepsilon_{ij}.$$

Where male is the reference category for sex, yes is the reference category for smoker, and southwest is the reference category for region.

Table 3: Correlation Matrix of Numerical Variables

	Age	**BMI**	**Children**	**Charges**
Age	1.0000	0.1093	0.0425	0.2990
BMI	0.1093	1.0000	0.0128	0.1983
Children	0.0425	0.0128	1.0000	0.0680
Charges	0.2990	0.1983	0.0680	1.0000

Table 3 displays a little positive association between age and BMI, indicating that as age rises, BMI tends to see a slight increase. The connection between age and the number of children is quite poor, suggesting a negligible association between age and the number of children. Age and charges exhibit a moderate positive association. This implies that elderly adults tend to incur larger medical expenses. The correlation coefficient of 0.2990 signifies a modest linear relationship. The association between Body Mass Index (BMI) and the number of children exhibits a significant lack of strength. There is minimal correlation between BMI and the number of children. There exists a little positive association between Body Mass Index (BMI) and medical charges. These findings indicate that persons with a higher BMI tend to have slightly higher medical expenses, on average. The association between the number of children and charges is highly insignificant. There is little data to support a direct correlation between the number of children and medical expenses. Furthermore, the diagonal element signifies the absolute connection between a variable and itself.

The correlation matrix offers valuable information on the linear associations between different variables. It is crucial to acknowledge that correlation does not establish causality.

Table 4: Association Test for Categorical Predictor Variables

Association	**Pearson Chi-Square**	**P-value**
Sex vs Smoker	7.7959	0.0053
Sex vs Region	0.4351	0.9329

Region vs Smoker	7.3435	0.0617

Table 4 shows a statistically significant association between the variables "Sex" and "Smoker" at a significance level of 0.05. Put simply, the prevalence of smoking (differentiating between smokers and non-smokers) greatly differs among different genders. Furthermore, there is not a significant association between the variables "Sex" and "Region" at a significance level of 0.05. There is no substantial variation in the distribution of areas between the sexes. Furthermore, there is no statistically significant relationship between the variables "Sex" and "Region" at a significance level of 0.05. There is no substantial variation in the distribution of areas among different sexes. In addition, for the statistically significant one, this would inflate the variance of the model, leading to a problem of multicollinearity. This would not be handled in this term paper.

The regression model for the data under consideration is given in Table 5 below.

Table 5: Result of Regression Analysis

Parameters	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10818.631	1071.8148	10.09375	<.001
Sex_female	131.31436	332.94544	0.394402	0.693348
Region_southeast	-74.97106	470.63864	-0.1593	0.87346
Region_northeast	960.05099	477.93302	2.008756	0.044765
Region_northwest	607.08709	477.20391	1.272175	0.203533
bmi	339.19345	28.59947	11.86013	<.001
children	475.50055	137.80409	3.450555	0.000577
Smoker_no	-23848.53	413.15335	-57.7232	<.001
age	256.85635	11.898849	21.58666	<.001

Based on the result presented in Table 5, it is seen that region Northeast, BMI, smoking status, and age have significant influence on health insurance charges in the US, with smoking status: no having an inverse relationship with health insurance charges based on the data used for the analysis for the regression model.

Statistical Tests to Assess Differences in Central Tendencies Of Predictor Variables Between the Two Groups

In this case, the response variable charge was split into two using the median value as the threshold, where from zero to the median value was regarded as low and above the median value was regarded as high. To carry out test of independence on the predictor variables, the dataset for each group must be normally distributed. The result of the test for normality is seen in table 6 below. The normality test carried out is the Shapiro-Wilk's test. The formula is given by

$$W = \frac{\sum(a_i x_i)^2}{(x_i - \bar{x})^2}$$

The hypotheses to be tested are as follows:

(i) There is no significant difference between the low charges and high based on the BMI of the respondents.
(ii) There is no significant difference between the low charges and high based on the age of the respondents.

(iii) There is no significant difference between the low charges and high based on the number of children of the respondents.

Table 6: Normality Test

Predictors	Groups	P-value	Inference
Age	High Low	5.01×10^{-21} 1.73×10^{-15}	Non- Parametric test
BMI	High Low	0.00789 0.00188	Non-Parametric Test
Children	High Low	1.42×10^{-26} 1.10×10^{-26}	Non-Parametric Test

"Non-parametric test" is a statistical method that does not assume any specific probability distribution for the data being analysed. It is used when the data does not meet the assumptions of parametric tests, such as normality or homogeneity of variance. Non-parametric tests make fewer assumptions about the underlying population and are therefore more robust to outliers and non-normal data (Campbell, 2012). Since none of the data followed a normal distribution for the two groups, we use a non-parametric statistic called the Wilcoxon test. The formula is given by

$$W^{+} = \sum_{i=1}^{N} R_i^{+}$$

Where:

N is the number of pairs,

R_i^{+} is the rank of the i^{th} positive absolute difference.

The result of the non-parametric statistics is seen in Table 7

Table 7: Non-Parametric Test (Wilcoxon Sign Rank Test)

	W Statistic	P-value	Inference
BMI	246841	0.001102	Reject H_0 and conclude that the mean of the two groups are statistically significant
Age	355711	$2.2\times10^{--16}$	Reject H_0 and conclude that the mean of the two groups are statistically significant
Children	226225	0.7154	Reject H_0 and conclude that the mean of the two groups are statistically significant

The visualization of the differences is seen in Appendix 9.

Statistical Test to Assess Differences In Central Tendencies Of Predictor Variable Based On Geographical Location (Region)

For tests of more than two groups, it is important to test for the normality of the data among other assumptions. The assumptions include independence, normality, homogeneity of variance, randomization, and homogeneity of regression slopes. The test for normality is seen in Table 8 for BMI and age.

Table 8: Normality Test for BMI and Age

Predictors	W-Statistic	P-value	Inference
BMI	0.9939	2.605×10^{-5}	Non-Parametric
Age	0.9447	2.2×10^{-16}	Non-Parametric

Since the dataset is non-normal, a Kruskal-Walis test which is the non-parametric version of analysis of variance test would be carried out. The results are shown in Table 9.

Table 9: Non-Parametric Test for more than Two Groups

Predictors	χ^2 Value	P-value	Inference
BMI	94.689	2.2×10^{-16}	Reject H_0 and conclude that the means are different across the regions/geographical location for BMI
Age	0.4138	0.9374	Do not reject H_0 and conclude that there is no substantial evidence to say that the means differ across regions/geographical location for age
Children	2.3754	0.4982	Do not reject H_0 and conclude that there is no substantial evidence to say that the means differ across regions/geographical location for number of children

The Dunn's test is used for the post-hoc test and this would only be carried out on the result that is statistically significant. In this case for BMI. This is seen in Table 10. The visualizations for the group are shown in Appendix 10 to Appendix 12

Table 10: Post-Hoc Test (Dunn's test) for BMI by Region

Region	northeast	northwest	southeast
northwest	-0.1398 >.999		
southeast	-8.6137 <.001	-8.2767 <.001	

southwest	-3.0369 0.0072*	-2.8994 0.0112*	5.2964 <.001*

*Represent p-values that are significant at .05

The post-hoc test revealed that only northwest-northeast is not significant. All other pairs are statistically significant.

Conclusion

Based on the comprehensive analysis carried out, it is evident that several factors significantly influence health insurance charges in the United States. The regression analysis revealed that age, BMI, smoking status, and region (specifically the northeast region) have significant impacts on insurance charges. For instance, the regression coefficients for age, BMI, and smoking status were found to be statistically significant (Table 5). Additionally, non-parametric tests demonstrated significant differences in central tendencies of predictor variables between low and high insurance charge groups, particularly in age and BMI (Table 6).

The study highlights the complex interplay between various demographic, socioeconomic, lifestyle, and regional factors in determining health insurance premiums. For example, the correlation matrix revealed moderate positive associations between age and insurance charges, indicating that elderly adults tend to incur larger medical expenses (Table 3). Similarly, the association between BMI and charges was found to be positive, suggesting that individuals with higher BMIs tend to have slightly higher medical expenses (Table 3).

These findings provide valuable insights for policymakers, insurers, healthcare providers, and consumers to understand insurance pricing dynamics and make evidence-based decisions to improve affordability, accessibility, and equity in healthcare coverage. By considering multiple variables in setting health insurance premiums, stakeholders can better address disparities in healthcare expenses and ensure fair access to healthcare services.

In conclusion, this report underscores the importance of leveraging statistical analyses to identify significant factors influencing health insurance charges. Further research and policy initiatives are warranted to enhance the fairness and efficiency of healthcare financing and access, ultimately improving health outcomes for all individuals.

References

Campbell, D. (2012): Nonparametric Methods in the Medical Sciences.

Cutler, D. M., & Zeckhauser, R. J. (1998). Adverse selection in health insurance. National Bureau of Economic Research.

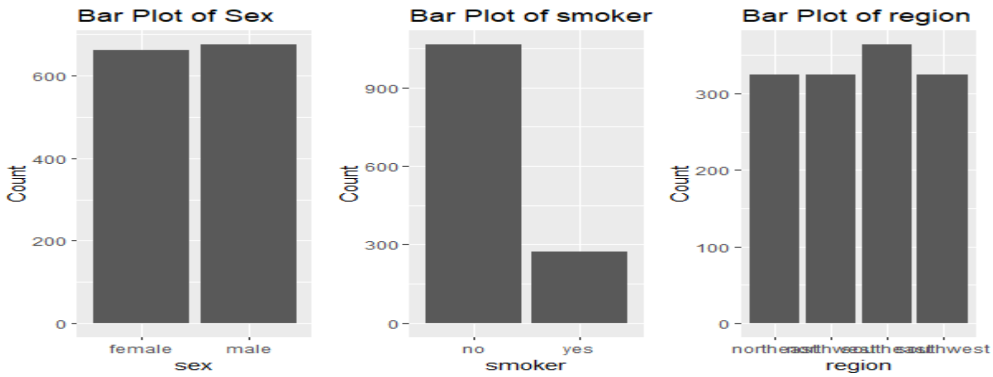
Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., ... & Baicker, K. (2012). The Oregon health insurance experiment: evidence from the first year. The Quarterly Journal of Economics, 127(3), 1057-1106.

Pauly, M. V., & Kunreuther, H. (2013). Guaranteed renewability in insurance. Journal of risk and uncertainty, 46(1), 51-72.

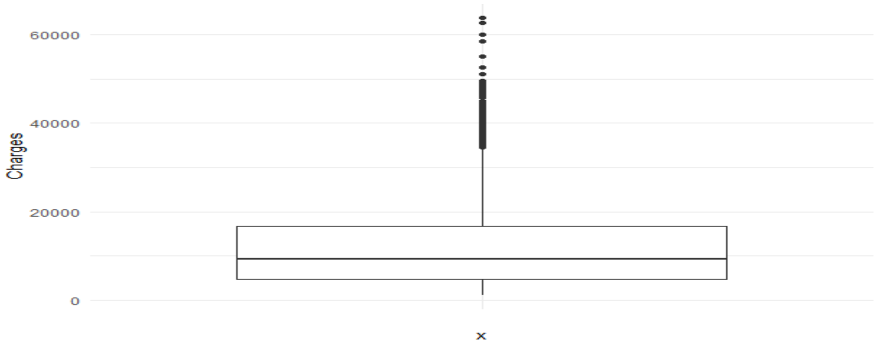
Zweifel, P., & Manning, W. G. (2000). Moral hazard and consumer incentives in health care. Handbook of health economics, 1, 409-459.

Appendices

Appendix 1: Plot for Non-Numeric Variables

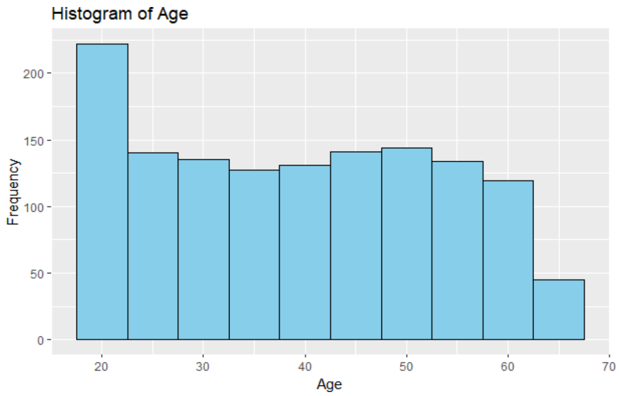


Appendix 2

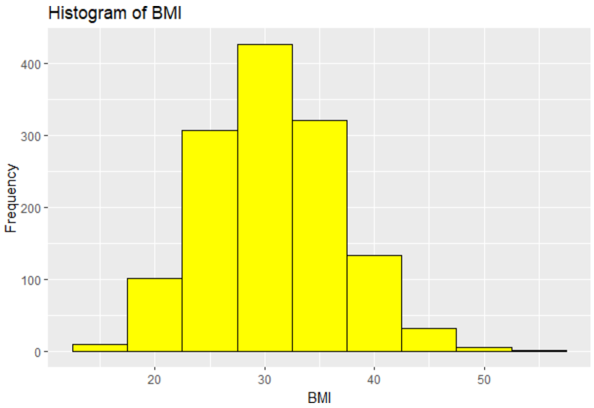


Histogram of numeric variables

Appendix 3

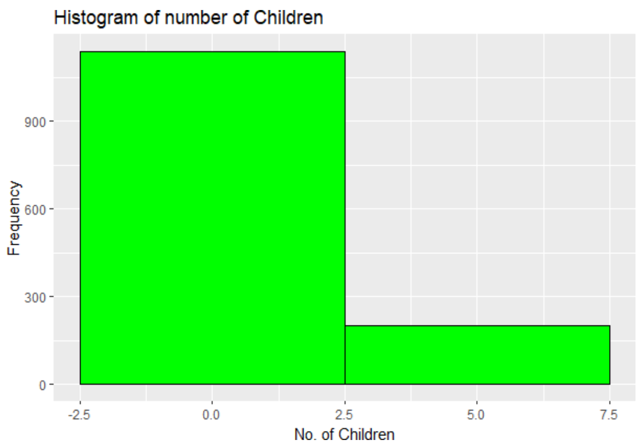


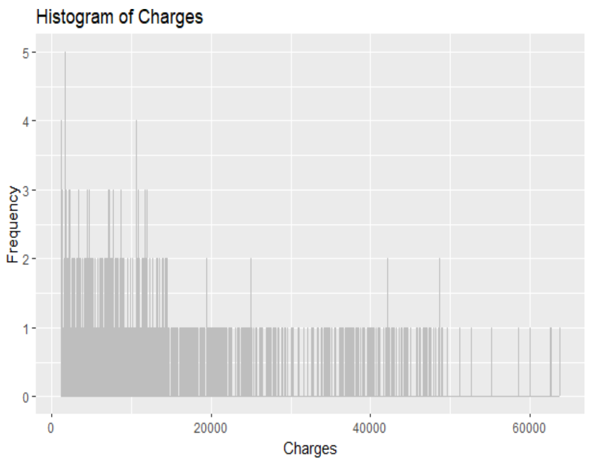
Appendix 4



Appendix 5

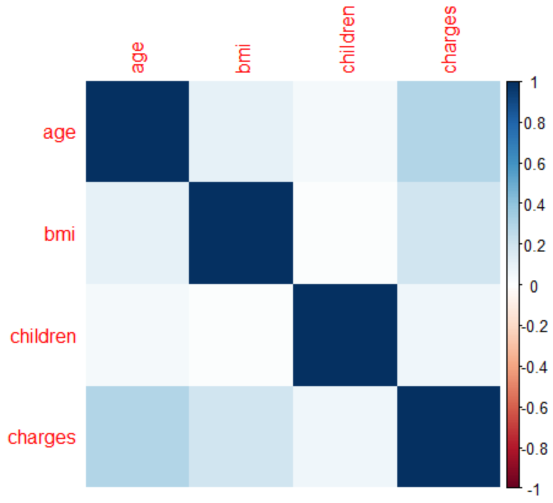
Appendix 6


Histogram of number of Children

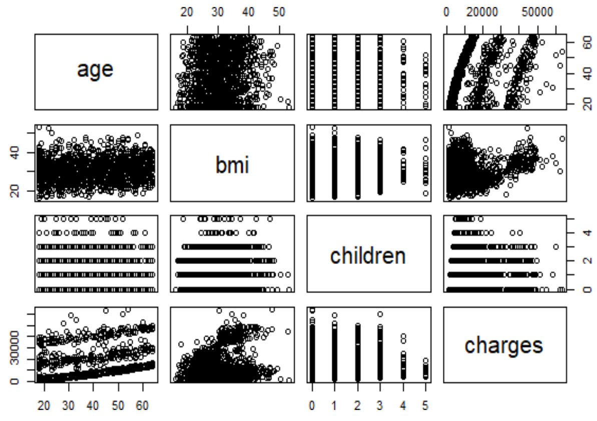

Histogram of Charges

Correlation plots.

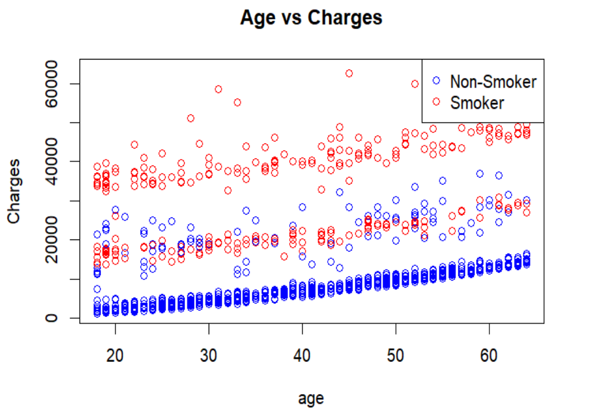
Appendix 7



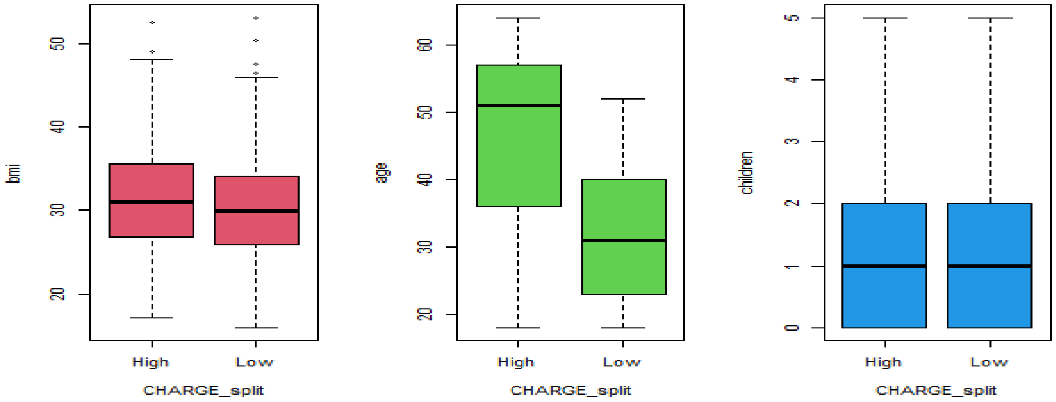
Appendix 8: Scatter plots for continuous variables



Plot for continuous variable vs categorical variable.

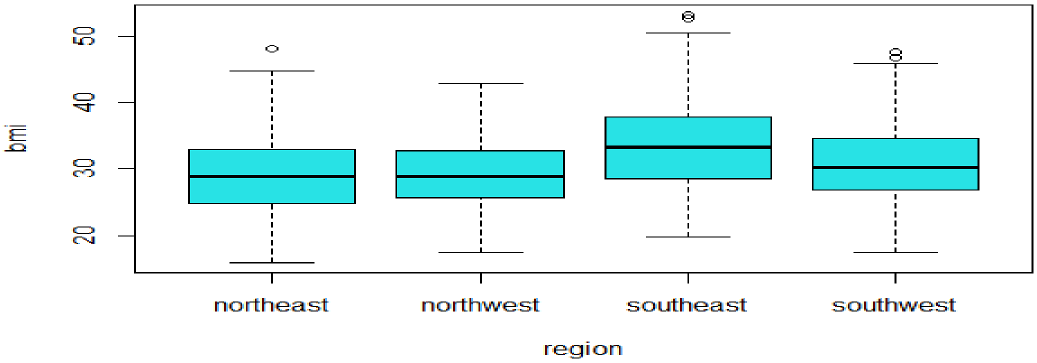

Age vs Charges

Appendix 9: Boxplots for Non-Parametric Test

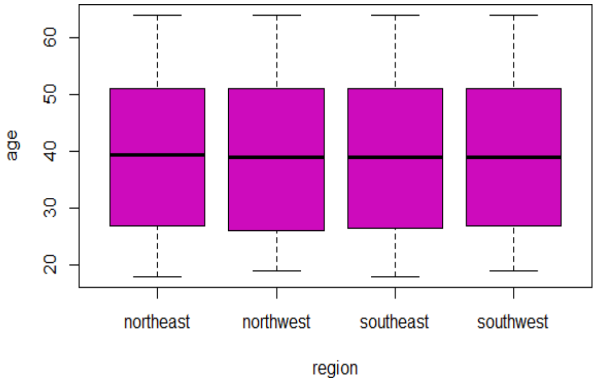


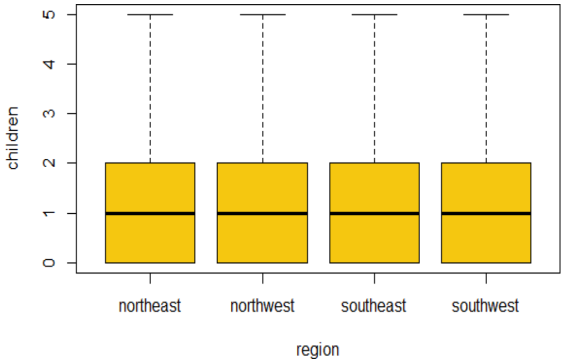
Differences in central tendencies of the interval predictor variables with respect to the geography (Region)

Appendix 10



Appendix 11 **Appendix 12**


age
northeast
northwest
southeast
southwest
region


children
northeast
northwest
southeast
southwest
region