“Exploring Factors Influencing Health Insurance Charges: A Comprehensive Analysis Using Statistical Techniques”

Name: Melrida Moras
ID: N1198965
Date: 15th March 2024
Professor: Dr Ayse Ulgen
Subject: MATH40031: Statistical Data Analysis & Vis 202324

INDEX

Introduction

Health insurance plays a crucial role in reducing the financial risks associated with healthcare expenses, providing individuals and families with access to necessary medical services without incurring overwhelming costs. Individuals may acquire coverage for a range of healthcare services, including as preventive care, treatment for illnesses and accidents, and specialised medical treatments, by pooling resources through insurance premiums. To establish strategies for pricing, coverage, and resource allocation, insurers, legislators, and healthcare providers must all have a thorough understanding of the factors driving health insurance costs. To contribute to a more comprehensive knowledge of healthcare funding and accessibility, this analysis of the health insurance dataset tries to identify the elements that influence medical insurance charges, such as lifestyle choices, geographic location, and personal traits.

Aim

In this report, we're diving into extensive health insurance dataset comprising personal attributes and geographic factors about people in the US. We need to understand what factors affect how the charges citizens pay for their medical insurance. Our goal is to clarify the key factors influencing insurance prices so that the healthcare industry may make well-informed policy decisions and use predictive modelling using comprehensive visualization, statistical analyses, and advanced modelling techniques. Our investigation begins with an exploration of the distributions and summary statistics of each variable, shedding light on the diverse characteristics of the population. The study uses multiple linear regression modelling to analyse insurance charges, CHARGE-split, and geographical dimension's influence on interval predictor variables. It employs statistical tests and advanced modelling techniques to reveal significant disparities among predictor variables.
This can help us make better decisions about healthcare policies and even predict future insurance costs.

Description of Dataset

The dataset comprises information on 1338 US citizens, detailing personal attributes (age, gender, BMI, family size, smoking habits), and geographic factors impacting medical insurance charges ($). It includes variables like age (years), gender (male/female), BMI (Body Mass Index), number of children covered (0-5), smoking status (yes/no), and region of coverage (southeast, southwest, northeast, northwest). This dataset enables the study of how these factors interplay with insurance costs, facilitating predictive modelling for healthcare expense estimation and informing strategies for personalized insurance plans and cost management.

Problem1: Visualization and Summary Statistics

Problem Statement:. The problem aims to develop an understanding of the distribution of each variable and derive Summary statistics. Furthermore, it aims to identify the relationships between the dataset attributes and insurance charges. Understanding the correlation between personal attributes, geographic factors, and medical insurance charges is crucial for developing personalized interventions, and cost-effective strategies, and promoting healthcare affordability and equity.

Methods: We are using Descriptive statistics to obtain a concise summary of the main characteristics of the data, helping to understand the measures of central tendency (mean, median, mode), measures of spread (standard deviation, variance, range) and distribution.
Visualization: To effectively visualise the distribution of variables, the R programming tool will be utilised. It will be used to plot boxplots for categorical data and histograms for continuous variables.

Descriptive Analysis of Dataset: We have used the R programming language to carry out our analysis. We examined for any missing values after importing the Health-Insurance-Data.csv file and handled them accordingly... In the given dataset age, bmi, children and charges can be classified as continuous variables and smoker, region, and sex as categorical variables. we calculated summary statistics including mean, median, minimum, maximum, and standard deviation for each variable. We visualized the distributions of age, BMI, children, and charges using histograms, and explored the relationship between insurance charges and categorical variables (smoker, region, sex) using boxplots. Which can be found in appendix 1. Additionally, we calculated the frequencies and percentages of smokers, sexes, and regions, and represented them using pie charts. The summary statistics is represented in a tabular format in Table 1. It is calculated by using the function summary ().

Table 1: Summary statistics of the dataset

Variable	min	1st Qu.	Median	Mean	3rd Qu	Max	Variance	Standard Deviation
age	18	27	39	39.21	51	64	197.40	14.04996
bmi	15.96	26.3	30.4	30.66	34.69	53.13	37.19	6.098187
children	0	0	1	1.095	2	5	1.45	1.205493
charges	1122	4740	9382	13270	16640	63770	146652380.00	12110.01124

Table 2: Frequency of categorical variables

Sex		Smoker		Region			
Female	Male	No	Yes	Northeast	Northwest	Southeast	Southwest
662	676	1064	274	324	325	364	325

Visualization and Interpretation: Appendix 1 presents data through histograms and boxplots, revealing the distribution of age, BMI, children, and insurance charges. Histograms show a normal distribution, while boxplots show a wide range of insurance charges. Smokers incur higher charges, while charges vary across geographic regions. Males have slightly higher median charges than females. The visualizations reveal that age, BMI, smoking status, geographic region, and sex influence insurance charges, emphasizing the need for considering multiple factors for effective healthcare cost management.

Findings and conclusion: Through descriptive analysis, we gained valuable insights into the factors affecting health insurance charges. By examining summary statistics and visualizing the distribution of variables, we identified patterns and relationships that can inform decision-making processes aimed at managing healthcare costs effectively. These insights are crucial for healthcare providers and policymakers in achieving targeted actions to optimize healthcare expenditure and improve affordability and access to healthcare services.

Problem 2: Testing Independence of Predictor Variables.

Problem Statement: This analysis's main goal is to identify the variables that affect health insurance costs. We want to Verify that the predictor variables are independent of one another. Examine the connections between personal traits and insurance costs. Examine correlations between categorical factors including gender, smoking status, and location.

Method: We will begin by conducting a descriptive analysis of the dataset to understand the distribution of variables and identify any outliers or missing values. Next, we will test the assumption of independence among predictor variables using correlation coefficients and tests of association such as Chi-square test for categorical variables and Pearson correlation for continuous variables. We will visualize the results using scatter plots to explore the relationships between variables and a correlation matrix to assess the strength and direction of associations.

Assumptions and Hypothesis: The Pearson correlation coefficient assumes a normal distribution for variables like age, BMI, and insurance charges, ensuring independence of observations and avoiding dependencies to ensure accurate correlation analysis.
Null Hypothesis (H0): There is no observable connection between geographic location, personal characteristics, and health insurance costs.
Alternative Hypothesis (H1): Medical insurance costs and at least one personal attributes are significantly correlated with one another and with regional variables.

To test the assumption of independence among predictor variables, we compute correlation coefficients using Pearson's method. The correlation matrix presented in Table 3 indicates the relationships between age, BMI, number of children, and insurance charges. The formula for calculating the Pearson correlation coefficient is

$$\boldsymbol{r} = \frac{\sum(x_{i-}\bar{x})(y_{i-}\bar{y})}{\sqrt{\sum(\,x_{i-}\bar{x})^2\;\sum(y_{i-}\bar{y})^2}}$$

x_i and y_i are the individual data points for variables
$\bar{x}$ and $\bar{y}$ are the means of variables *x* and *y* respectively.

By comparing observed frequencies in a contingency table, chi-square test is used to identify the significant association between categorical variables including gender, geographic region, and smoking behaviours. The frequency table or contingency table in which the data are arranged determines the formula for the chi-square test statistic We use contingency tables in our analysis table. Table 4 shows the chi-square test results. Table 5,6,7 depicts the Associations between Variables sex smoker region respectively.

Table 3: correlation coefficients (Pearson method)

	age	bmi	children	charges
age	**1**	**0.109272**	**0.042469**	**0.299008**
bmi	**0.109272**	**1**	**0.012759**	**0.198341**
children	**0.042469**	**0.012759**	**1**	**0.067998**
charges	**0.299008**	**0.198341**	**0.067998**	**1**

Table 4: Chi-square for categorical variables

Categorial Variables Pair	χ2	df	p-value
sex vs smoker	**7.3929**	**1**	**0.06548**
sex vs region	**0.43514**	**3**	**0.9329**
smoker vs region	**7.3435**	**3**	**0.06172**

Table 5: contingency table Sex Vs smoker

Sex	Smoker	
	no	yes
female	**547**	**115**
male	**517**	**159**

Table 6: contingency table Sex Vs Region

Sex	Region			
	northeast	northwest	southeast	southwest
female	**161**	**164**	**175**	**162**
male	**163**	**161**	**189**	**163**

Table 6: contingency table Region Vs Smoker

Region	Smoker	Frequency
northeast	**no**	**257**
northeast	**yes**	**67**
northwest	**no**	**267**
northwest	**yes**	**58**
southeast	**no**	**273**
southeast	**yes**	**91**
southwest	**no**	**267**
southwest	**yes**	**58**

Visualization Interpretation: Correlation coefficients close to 1 or -1 indicate strong positive or negative correlations, respectively. Scatter plots visually represent the relationships between pairs of variables. The correlation matrix provides a comprehensive summary of the correlations between all variables. Following computation, the chi-square test statistic is compared to the critical value obtained from the (r-1) × (c-1) degrees of freedom chi-square distribution. If the calculated chi-square value is higher than the critical value at a given significance level (often 0.05), we dismiss the null hypothesis of independence and come to the conclusion that there is a significant relationship between the variables.

The scatter plots show a positive correlation between age and insurance charges, suggesting older individuals have higher medical expenses in Appendix 2 Higher BMI may also result in higher charges. However, there is a weak relationship between the number of children and insurance charges. No significant associations were found between gender, geographic region, smoking habits, appendix 3 shows the Association of categorical variables through bar plots on Contingency tables of variables.

Conclusion: The analysis offers insightful information on the variables affecting health insurance costs. Insurance costs positively correlate with age and BMI, suggesting that these are important factors. Additionally, there is a correlation between gender and smoking behaviour that could affect insurance costs even more. However, there is no measurable correlation found between region and smoking behaviours or gender and geography.

Problem 3: Multiple Linear Regression Modelling for Medical Charges Prediction

Problem statement: Specifically, The task is to utilize linear regression modelling to identify and describe the variables that have an influence on medical insurance charges. The objective is to develop a predictive model that can accurately estimate insurance costs based on various predictor variables, such as age, BMI, smoking habits, geographic region, and other relevant factors. Additionally, the study aims to explore the possibility of using alternative models, beyond conventional linear regression.

Method: To investigate the association between the dependent variable (medical insurance costs) and the independent factors (age, BMI, children, smoking status, gender, and region), we use multiple linear regression analysis in this study. To learn more about the relationships between the variables, we also perform descriptive analyses using contingency tables and correlation analysis. Instead of using basic linear regression in our research, we choose to employ multiple linear regression to examine the effects of several independent factors on medical insurance costs. many linear regressions enables us to evaluate the cumulative effects of many predictors on the dependent variable, whereas simple linear regression can only handle one independent variable. This is particularly relevant to our research because it offers a clear understanding of the joint effects of individual characteristics and regional characteristics on health insurance premiums for US citizens, while accounting for the effects of other variables to enable a more thorough examination of the associations between the predictors and health insurance premiums.

Multiple Linear Regression: Table 8 shows the coefficients of Multiple Linear Regression Model
Multiple linear regression is a statistical method that models the relationship between a dependent variable and multiple independent variables, extending the concept of simple linear regression to situations where multiple predictors are involved, thereby enhancing the understanding of the relationship between the dependent and independent variables. The dependent variable is charges, while the independent variables include age, BMI, children, smokers, and region. In multiple linear regression, the relationship between the dependent variable Y and p independent variables $X1$, $X2$,..., Xp is represented by the following linear equation:

$$Y= \beta 0+ \beta 1X1+ \beta 2X2+...+\beta pXp+\varepsilon$$

Where Y is the dependent variable (outcome or response variable).$X1$, $X2$,..., Xp are the independent variables (predictors or explanatory variables).$\beta 0$ is the intercept term (the value of Y when all predictor variables are zero).$\beta 1$, $\beta 2$,...,βp are the coefficients (regression coefficients) that represent the change in Y for a one-unit change in each predictor variable, holding all other variables constant. ε is the error term (residuals), which represents the difference between the observed values of Y and the values predicted by the model. After fitting the model, the **summary()** function provides a summary of the multiple linear regression model, including coefficients, standard errors, t-statistics, p-values and other statistics.

Table 8: Multiple Linear regression model coefficients

Predictor	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	**-11938.539**	**987.819**	**-12.09**	**< 2e-16**
age	**256.856**	**11.899**	**21.59**	**< 2e-16**
bmi	**339.193**	**28.599**	**11.86**	**< 2e-16**
children	**475.501**	**137.804**	**3.45**	**0.000577**
Smoker yes	**23848.535**	**413.153**	**57.72**	**< 2e-16**
Sex male	**-131.314**	**332.945**	**-0.39**	**0.693348**
Region northwest	**-352.964**	**476.276**	**-0.74**	**0.458769**
Region southeast	**-1035.022**	**478.692**	**-2.16**	**0.030782**
Region southwest	**-960.051**	**477.933**	**-2.01**	**0.044765**

Visualisation and Interpretation: Appendix 3 depicts a scatter plot used to visualise a multiple linear regression model, revealing the direction of the relationship between the predictor and outcome variables. The model reveals that 74.94% of medical insurance charges can be explained by factors such as age, BMI, and smoking. Age is a significant influencer, with each additional year leading to a substantial increase in charges. Higher BMI levels are also associated with elevated medical expenses, highlighting the interplay between lifestyle factors and healthcare costs. The stark contrast in charges between smokers and non-smokers highlights the financial ramifications of smoking, with smokers facing significantly higher charges compared to non-smokers. This underscores the detrimental health effects of smoking and the economic burden it places on healthcare systems. Interestingly, while certain regions show significant disparities in charges, other factors like sex do not significantly influence medical insurance costs. These findings highlight the importance of targeted interventions and policy measures to address regional disparities while recognizing the limited impact of certain demographic factors

Conclusion: Our findings highlight the importance of various personal and geographic factors in determining medical insurance charges. Older individuals and smokers tend to incur higher medical insurance costs, emphasizing the need for targeted active strategies to promote healthier behaviours and lifestyles. The regional differences observed in medical insurance charges suggest potential differences in healthcare access and utilization across different areas, requiring further research.

Problem 4: Comparison of Predictor Variables Across Charge Groups.

Problem Statement: Understanding the factors influencing medical insurance charges is crucial for insurance companies and policymakers. This report aims to investigate how personal attributes, geographic factors, and health-related habits affect medical insurance costs. We utilize a dataset containing information on 1338 By splitting the medical insurance charge data into two groups, the goal is to produce a new categorical variable known as "CHARGE-split." Statistical tests will be performed to assess variations in the central tendency of predictor variables with respect to the newly created categorical variable, assuming that all predictor variables are independent of each other. p between these variables and develop insights for predictive modelling**.**

Method: Data Preprocessing: The dataset containing information on personal attributes, geographic factors, and insurance charges is pre-processed for analysis.
Variable Splitting: Insurance charges are divided into two categories, "High" and "Low," based on the median charge value.
Statistical Analysis: To evaluate variations in the mean of predictor variables with regard to the newly created categorical variable, CHARGE-split, we run statistical tests, such as t-tests and chi-square tests.
Visualisation: To aid with interpretation, the results are displayed graphically using boxplots and bar graphs.
Assumption: We assume that predictor variables are not dependent of one another.
Hypothesis:
Null Hypothesis (H0): The central tendency of the predictor variables in the high and low medical insurance charge groups do not differ significantly.
Alternative Hypothesis (H1): There are significant differences in central tendencies of predictor variables between high and low medical insurance charge groups.
We conduct t-tests for continuous variables (age, BMI, and number of children) and chi-square tests for categorical variables (gender, smoking habits, and region) to analyse differences in central tendencies between high and low charge groups. Table 9 show t tests for continuous variables.

The study compares the mean BMI between high and low insurance charge groups using a t-test. The results show a significant difference in BMI between the two groups, with the high insurance charge group having a higher mean BMI. However, the number of children between the two groups is not statistically significant. The high insurance charge group has 1.118087 children, while the low insurance charge group has 1.071749. The results suggest that individuals in the high insurance charge group tend to have a higher BMI.

Table 9: T-test for Continuous Variables

Continuous Variables	t	df	p-value	Confidence Interval		Mean	
						Group High	Group Low
Age by CHARGE_split	**21.822**	**1206.3**	**< 2.2e-16**	**13.10454**	**15.69367**	**46.40658**	**32.00747**
bmi by CHARGE_split	**3.2992**	**1336**	**0.0009952**	**0.4443057**	**1.7476972**	**31.2114**	**30.1154**
Children by CHARGE_split	**0.70289**	**1333.4**	**0.4822**	**-0.8298978**	**1.7476972**	**1.118087**	**1.071749**

The chi-square tests are used to analyse the relationships between categorical variables such as gender, smoking habits, and geographic region, and insurance charge groups. Table 10 shows Chi-Square test for Categorical Variables The results show no significant association between gender and insurance charge groups, suggesting that the distribution of males and females does not differ significantly between high and low insurance charge groups. Smoking habits are strongly associated with insurance charge groups, with smokers more likely to be in high insurance charge groups. Geographic region is not significantly associated with insurance charge groups, suggesting that the distribution of individuals across different regions does not differ significantly between high and low insurance charge groups.

Table 10: Chi-Square test for Categorical Variables

Categorial Variables	χ2	df	p-value
Sex and CHARGE_split	**0.0029899**	**1**	**0.9564**
Smoker vs CHARGE_split	**342.05**	**1**	**< 2.2e-16**
Region vs CHARGE_split	**4.466**	**3**	**0.2153**

Visualization: Appendix 4 shows the Boxplots to Visualize the distribution of age, BMI, and number of children across high and low charge groups, showing differences in central tendencies and Bar Plots to Display the association between gender, smoking habits, region, and charge groups, highlighting any significant differences in distribution.

Conclusion and Findings: Statistical tests reveal significant differences in age and BMI between high and low charge groups, indicating their potential influence on medical insurance costs. Smoking habits exhibit a strong association with medical charges, as evidenced by the chi-square test. Gender and region do not show significant differences in distribution between charge groups. Overall, age, BMI, and smoking habits play crucial roles in determining medical insurance charges, highlighting the importance of lifestyle and health factors in healthcare expenses. This analysis provides insights for insurance companies and policymakers to better understand the factors driving medical insurance costs and develop targeted strategies for cost management and risk assessment.

Problem 5: Assessment of Central Tendencies of Predictor Variables by Geography

Problem Statement: The study aims to analyse variances in BMI, age, and other interval predictor variables across geographic regions, using ANOVA and Kruskal-Wallis tests. The aim is to identify significant differences in central tendencies across locations, providing insights for healthcare policy and resource allocation decision-making.

Method: The dependent variable (Age or BMI) should be approximately normally distributed within each group defined by the independent variable (Region).ze assumptions and hypotheses for the ANOVA test:

Independence: Observations are independent of each other.
Normality: The dependent variable (Age or BMI) should be approximately normally distributed within each group defined by the independent variable (Region).

Homogeneity of variances: The variances of the dependent variable should be approximately equal across all groups defined by the independent variable Hypotheses:
Null Hypothesis (H0): There is no significant difference in the means of the interval predictor variable (e.g., Age, BMI) among different regions.
Alternative Hypothesis (H1): There is a significant difference in the means of the interval predictor variable among different regions.

Statistical tests to check the assumptions.
Shapiro-Wilk normality test: The Shapiro-Wilk normality test is a statistical test used to assess whether a data sample comes from a normally distributed population. In the results you provided, there are two tests conducted, one for the variable "age" and the other for the variable "bmi."

Table1: Shapiro-Wilk normality test

Variables	w	p-value
age	**0.9447**	**< 2.2e-16**
bmi	**0.99389**	**2.61E-05**
charges	**0.81469**	**< 2.2e-16**

Interpretation: The p-value associated with the Shapiro-Wilk test for both Age and BMI is much less than the significance level (typically 0.05). This indicates strong evidence against the null hypothesis (the data is normally distributed).
For Age: Since the p-value is practically zero, we reject the null hypothesis and conclude that the data for Age is not normally distributed.
For BMI: Although the p-value is small, it is not as small as for Age. Still, it is smaller than the typical significance level of 0.05, indicating that we reject the null hypothesis and conclude that the data for BMI is not normally distributed.

Bartlett's test
Bartlett's test of homogeneity of variances is used to assess the homogeneity of variances across multiple groups or samples. In the context of statistical analysis, it is important to ensure that the variances of the residuals (or errors) are approximately equal across different groups or conditions. This is an assumption underlying many statistical tests, including ANOVA (Analysis of Variance) and regression analysis. Table 12 displays the results of test.

Table 12 : Bartlett's test of homogeneity of variances

Variables	Bartlett's K squared	df	p-value
Age	**0.073134**	**3**	**0.9564**
bmi	**18.742**	**3**	**0.000309**

Interpretation:
For Age: Since the p-value is greater than the significance level, we fail to reject the null hypothesis. Bartlett's test suggests that the assumption of homogeneity of variances is met for age, meaning that the variances of age are approximately equal across different regions.
Regarding BMI: We deny the idea of a null hypothesis since the value of p is smaller than the significance level. According to Bartlett's test, the BMI variances are not similar over different regions, which implies that the premise of homogeneity of variances is broken.
In conclusion, since the assumptions of homogeneity of variances and normality are both broken, it would be more appropriate to use a non-parametric alternative, such as the Kruskal-Wallis test, which is robust to homogeneity of variance violations and does not rely on the assumption of normality.

The Kruskal-Wallis test: The Kruskal-Wallis test is a non-parametric statistical test used to compare the central tendencies of three or more independent groups, it is employed as an alternative to the one-way analysis of variance (ANOVA). The Kruskal-Wallis test can be used for both continuous and ordinal data, making it versatile in various research settings.
Interpretation: If the p-value from the Kruskal-Wallis test is less than a chosen significance level (e.g., 0.05), it indicates that there are significant differences in the central tendencies of the groups. If the p-value is greater than the significance level, there is no evidence to suggest significant differences among groups. The Table 13 displays the results of the test.

Variables	chi-squared	df	p-value
Age	0.41382	3	0.9374
bmi	94.689	3	< 2.2e-16

Table 13: Kruskal-Wallis test

Age: The p-value associated with the Kruskal-Wallis test with respect to Age is 0.9374, which is much higher than the significance level (e.g., 0.05). Therefore, we fail to reject the null hypothesis. This suggests that there is no significant difference in the central tendencies of Age among different regions.

BMI: The p-value associated with the Kruskal-Wallis test for BMI is practically zero, which is less than any reasonable significance level (e.g., 0.05). Therefore, we reject the null hypothesis and conclude that there is a significant difference in the central tendencies of BMI among different regions.

Since the Kruskal-Wallis test for BMI yielded a significant result, we can proceed with post-hoc pairwise comparisons to determine which specific regions differ significantly from each other in terms of BMI.

One of the commonly used post-hoc tests for Kruskal-Wallis is the Dunn's test with Bonferroni correction. The results of D**unn'** test are displayed in table 14.

Table14: Dunn's test

Comparison	Z_Statistic	Raw P-Value	Adjusted P-value
northeast - northwest	-0.1397714	4.44E-01	1.00E+00
northeast - southeast	-8.413665	1.99E-17	1.19E-16
northwest - southeast	-8.2767381	6.33E-17	3.80E-16
northeast - southwest	-3.0369246	1.20E-03	7.17E-03
northwest - southwest	-2.8993878	1.87E-03	1.12E-02
southeast - southwest	5.2964216	5.90E-08	3.54E-07

Interpretation:. The southeast and northwest regions have significantly lower mean BMIs than the northeast region. Conversely, the mean BMI of the southwest region is notably greater than that of the northeast region. Ultimately, our findings demonstrate that there are notable geographical variations in both the mean BMI values and the central trends of BMI.

Visualization: appendix 5 displays the distribution of BMI for each region, with whiskers extending to the minimum and maximum values within 1.5 times the IQR. Outliers are represented by individual points.

Conclusion and findings: Considering the analysis that was done We discovered evidence to disprove the BMI null hypothesis, showing that the central tendencies of BMI vary significantly between locations. The fact that we were unable to reject the null hypothesis for age, however, indicates that the central tendencies of age do not significantly differ between areas. In order to effectively address regional inequities and enhance overall health outcomes, healthcare planning should take geographic location into account when calculating body mass index (BMI).

Conclusion

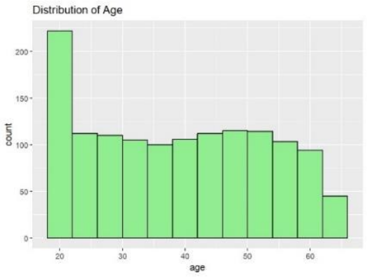
All things considered, this investigation offers insightful information about the complex relationships among individual characteristics, regional circumstances, and medical insurance costs, which improves knowledge about healthcare finance and accessibility in the US. In order to enhance healthcare outcomes and reduce inequities, these insights can help insurance companies, lawmakers, and healthcare professionals make well-informed decisions.

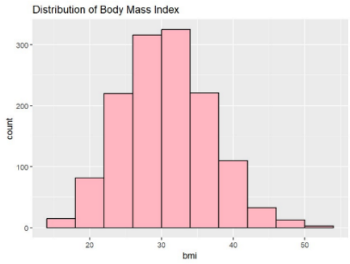
References

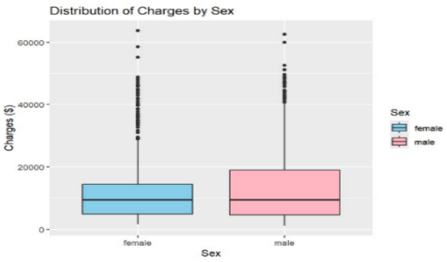
1. Akalin, A. (n.d.). 2.8 Plotting in R with ggplot2 | Computational Genomics with R. [online] compgenomr.github.io. Available at: https://compgenomr.github.io/book/plotting-in-r-with-ggplot2.html [Accessed 5 Mar. 2024].

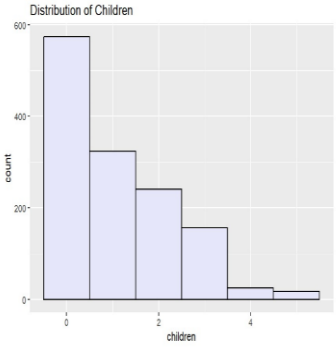
2. Mishra, P., Pandey, C., Singh, U., Sahu, C., Keshri, A. and Gupta, A. (2019). Descriptive Statistics and Normality Tests for Statistical Data. *Annals of Cardiac Anaesthesia*, [online] 22(1), pp.67–72. doi:https://doi.org/10.4103%2Faca.ACA_157_18.
3. Wei, T. and Simko, V. (2021). An Introduction to corrplot Package. [online] cran.r-project.org. Available at: https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html.
4. GeeksforGeeks. (2020). Bartlett's Test in R Programming. [online] Available at: https://www.geeksforgeeks.org/bartletts-test-in-r-programming/.
5. Zach (2020). How to Perform Dunn's Test in R. [online] Statology. Available at: https://www.statology.org/dunns-test-in-r/.
6. Saed Jama Abdi (2023). A Comprehensive Guide for Selecting Appropriate Statistical Tests: Understanding When to Use Parametric and Nonparametric Tests. Open Journal of Statistics, 13(04), pp.464–474. doi:https://doi.org/10.4236/ojs.2023.134023.
7. Marill, K.A., 2004. Advanced statistics: linear regression, part II: multiple linear regression. Academic emergency medicine, 11(1), pp.94-102.

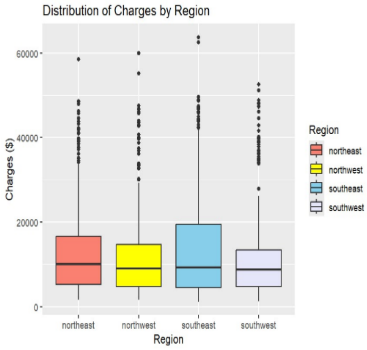
Appendix 1: Visualisation of distribution of variables

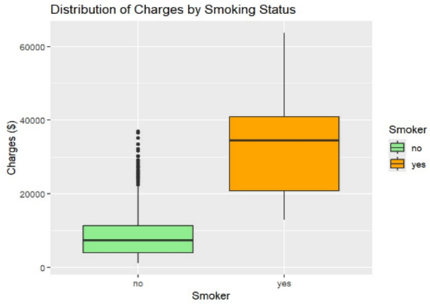


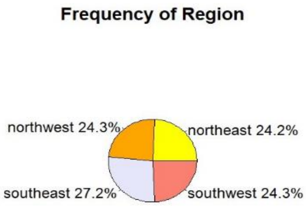


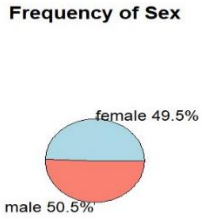


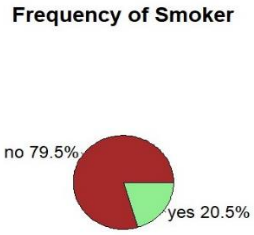




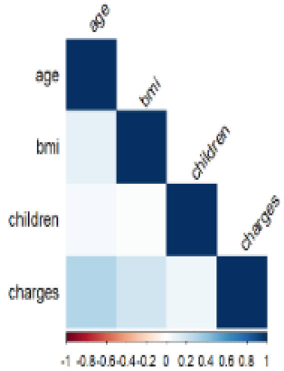


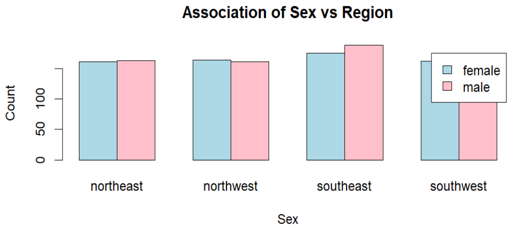


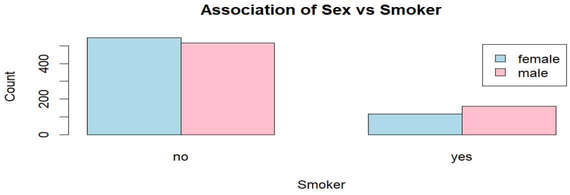


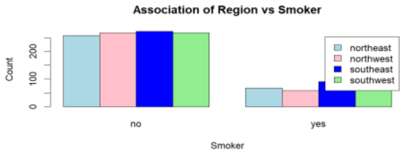


Appendix 2: Correlation Matrix and tests of associations

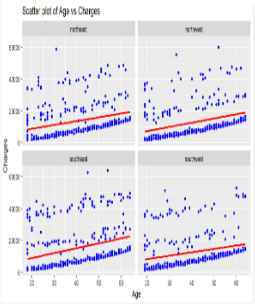


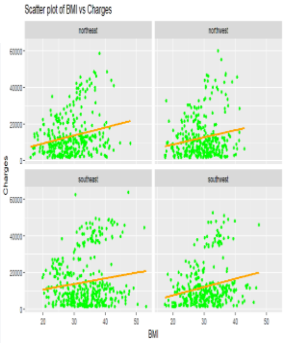


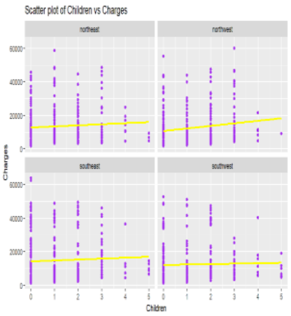


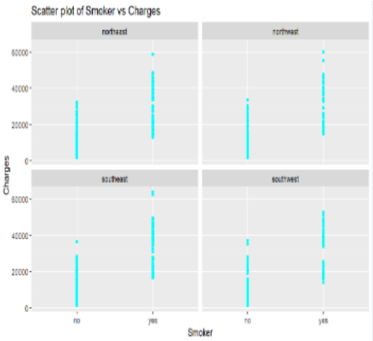


Appendix 3: Multiple Linear Regression scatter Plot (continuous variable Vs Charge)

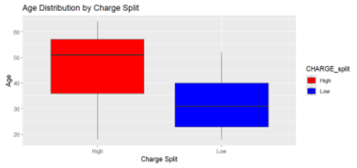


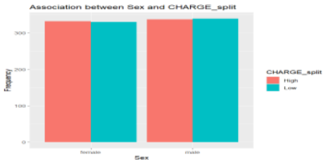


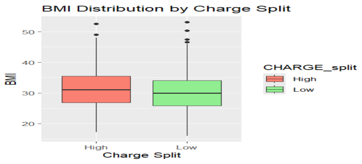


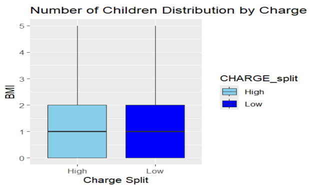


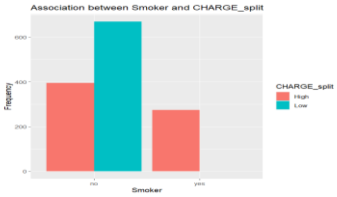
Appendix 4: Distribution of variables by charge split ,association between charge split and variables

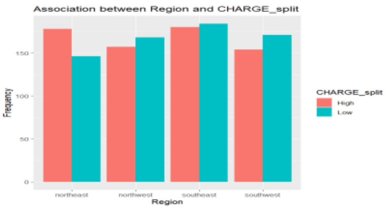

Age Distribution by Charge Split
Age
High
Low
Charge Split
CHARGE_split


Association between Sex and CHARGE_split
Frequency
female
male
Sex
CHARGE_split
High
Low


BMI Distribution by Charge Split
BMI
High
Low
Charge Split
CHARGE_split
High
Low


Number of Children Distribution by Charge
BMI
High
Low
Charge Split
CHARGE_split
High
Low


Association between Smoker and CHARGE_split
Frequency
no
yes
Smoker
CHARGE_split
High
Low


Association between Region and CHARGE_split
Frequency
northeast
northwest
southeast
southwest
Region
CHARGE_split
High
Low