Exploring Regional Variations and Predictive Factors in Health Insurance Charges: A Statistical Analysis Approach

Name: Neethushree Kumar

NTU_ID: N1229607

Module title: Statistical Data Analysis and Visualisation

Module Code: MATH40031

Table of Contents

INTRODUCTION

“The goal is to turn data into information, and information into insight.” (Carly Fiorina, et.al,2005)

The principle of "Prevention is better than cure" possesses a lot of importance when you consider the significance of data analysis is in the healthcare field, especially when you consider into the way health insurance trends affect policy along with how patients receive medical attention. The present research takes an in-depth look at a Health Insurance Dataset in order to demonstrate how insurance coverage, patient demographics, and healthcare use are all related in complicated ways. To start, we indicate a few key concepts that will assist us with our analysis: "Health insurance" is an approach to protect yourself financially from healthcare expenses. It includes different types of benefits and plans that make paying for healthcare easier for individuals. The data being examined at presents a statistical picture of insured people, including their age, gender, medical history, insurance costs, and other significant information.

Predictive modelling is a commonly used statistical technique to predict future behaviour.” Predictive modelling solutions are a form of data-mining technology that works by analysing historical and current data and generating a model to help predict future outcomes”. (Gartner et.al,2018).”The benefits of predictive modelling in healthcare are vast, including the ability to predict outcomes with greater accuracy, manage population health, enhance clinical decision-making, optimize traditional clinical processes, and improve risk scores and stratification. “(Audrey et.al,2024). Predictive modelling is essential in the health insurance market since it uses historical data to extract useful insights. The objective is to improve financial management, increase customer satisfaction, and encourage healthier behaviours among policyholders. Moreover, predictive modelling has been utilised in health insurance for a wide range of objectives, which include the following:

- Predictive modelling assesses risk by considering factors like age, BMI, smoking status, and the number of children to determine insurance costs.
- It enhances premium pricing accuracy by forecasting healthcare expenses, ensuring rates fairly reflect individual risk.
- Insurers use predictive modelling to customize policies and offer personalized advice or wellness programs based on key risk variables.
- Advanced models identify patterns suggesting fraudulent claims, helping maintain fair prices and reduce costs.
- Predictive modelling forecasts future healthcare needs, aiding in the efficient allocation of healthcare resources.

DESCRIPTION OF DATASET

The Health Insurance Dataset provides a complete statistical overview of individual insurance expenses, considering consideration of several demographic and health-related variables. The dataset consists of 1,338 individuals, ranging in age from 18 to 64 years variety of samples from multiple adulthood stages is indicated. The mean age of 39.2 years suggests a middle-aged population with different health insurance preferences and risks. The average BMI, which indicates overweight and obesity, is 30.66. This means the average person in this is overweight by BMI standards. Higher BMI levels increase health risks, which could raise insurance costs. The average number of children per person is 1.09, demonstrating that family planning may affect insurance costs. Insurance costs range from $1,121.87 to $63,770.43, averaging $13,270.42. “ it is challenging to select an appropriate method and find the most accurate predictive model for a given dataset due to many aspects and multiple factors involved in the modelling process”(Li, J et.al, 2019). Predictive modelling for health insurance costs can be challenged by models identifying correlations without causality, dataset biases that can affect prediction accuracy, and ethical concerns about equity and confidentiality:

- Potential data bias: The dataset's representativeness of the broader population might be skewed, affecting prediction accuracy.
- Risk of overfitting: Models might learn the noise instead of the signal, performing well on training data but poorly on new data.
- Complexity of interactions: The multifaceted nature of factors influencing insurance costs might not be fully captured.
- Lack of comprehensive features: Important predictors of insurance costs, like pre-existing conditions, may be absent from the dataset.
- Ethical concerns: Predictive modelling in healthcare can lead to decisions that might disadvantage certain groups or raise privacy issues.

Techniques include linear regression for a straightforward approach, decision trees for capturing non-linear relationships. It offers comprehensive packages for each of these methods, such as **lm()** for linear regression, **rpart** for decision trees, enabling nuanced analysis and predictive capabilities.

Table 1: Predictive modelling: an overview

Variable	Type	Description	Role in Predictive Modeling
Age	Numerical	The age of the insured individual.	A key predictor of insurance costs; older individuals may incur higher costs.
Sex	Categorical	The gender of the insured (male/female).	May influence costs due to gender-specific risk profiles.
BMI	Numerical	Body Mass Index, a measure of body fat based on height and	Used to assess health risk; higher BMI may lead to higher insurance charges.

		weight.	
Children	Numerical	The number of children/dependents covered by the insurance.	Can affect costs; policies covering more dependents may have higher premiums.
Smoker	Categorical	Whether the insured is a smoker (yes/no).	Strongly influences costs due to the health risks associated with smoking.
Region	Categorical	The insured's residential area in the US.	Regional health care costs and practices may influence insurance charges.
Charges	Numerical	The total insurance charges for the individual.	The outcome variable in predictive modeling, representing insurance costs.

Problem statement 1

In this problem statement, The objective is to study and offer health insurance data. Understanding numerical and categorical factors including age, BMI, charges, sex, smoking status, geography, and number of children (represented as a categorical variable for visualisation) is the key focus. These variables' summary statistics will show their central tendencies and variability, providing a complete picture of the dataset.

Statistical Analysis

The dataset, titled "Insurance_data," comprises a range of variables, including both numerical features (age, BMI, charges) and categorical attributes (sex, smoker status, area, number of children).” Data visualization in R offers a powerful way to communicate complex information clearly, allowing for an intuitive understanding of patterns, trends, and relationships within the data through the use of libraries like ggplot2 and base R plotting functions”(Ralph T et.al,2017). First, the dataset is examined to understand its structure and summary statistics. subsequently histograms are generated to analyse the distributions of numerical variables, while bar charts are used to understand the distribution of categorical variables.

Table 2: Summary statistics of numerical variable

Variable	**Min**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
Age	18.00	27.00	39.00	39.21	51.00	64.00
BMI	15.96	26.30	30.40	30.66	34.69	53.13
Children	0.000	0.000	1.000	1.095	2.000	5.000
Charges	1122	4740	9382	13270	16640	63770

The summary for numerical variables presents key statistics for each of the interval or ratio-scale variables in the dataset: Age, BMI, Children, and Charges. The dataset shows instantaneous insurance coverage for people aged 18 to 64, with an average age of 39. Individual BMIs vary, with an average of 30.66, indicating overweight. Insurance costs range from $1,122 to $63,770, averaging $13,270. The difference reflects consumers' diverse health conditions, coverage needs, and demographics.

Table 3: Summary statistics of categorical variable

Variable	**Length**	**Class**	**Mode**
Sex	1338	character	character
Smoker	1338	character	character
Region	1338	character	character

The categorical variables summary includes details for the Sex, Smoker, and Region variables have 1338 observations matching the dataset's total. The variables "Sex" and "Smoker" are categorical since their character types determine their qualities, allowing for gender and smoking status classifications. "Region" classifies people by their locations (southwest, southeast, northwest, northeast), adding spatial classification to the dataset.

Problem statement 2

The task involves analysing insurance data by separating it into separate factors, including age, BMI, number of children, and smoking status, which were initially assumed to be unrelated. By implementing statistical analytic techniques such as correlation coefficients and visual tools like scatter plots and correlation matrices, our objective is to identify any inherent relationships between these variables and insurance charges. This exploration is essential for examining our presumptions regarding the independence of variables and has the potential to identify complex patterns that influence the cost of insurance. Ultimately, this will assist in developing more precise insurance models.

Statistical analysis

The statistical studies provided use two fundamental methods: Pearson Correlation Coefficients and Chi-Square Tests, each fulfilling certain goals depending on the characteristics of the data being examined.

For Pearson Correlation Coefficient

It is a statistical measure that evaluates the linear relationship between two continuous, numerical variables. This coefficient, denoted as r. Correlation coefficient r ranges from -1 to 1. Close to 1 or -1 indicates a strong linear relationship, while close to 0 indicates no linear association. A positive r suggests a direct linear relationship, meaning one variable rises when the other rises." The PPC is adopted to determine whether the particles are close to the true states."(Zhou, H., Deng, Z., Xia, Y. and Fu, M. et.al, 2016) A negative r shows a reverse linear correlation, meaning an increase in one variable causes a reduction in the other. The correlation coefficient (r) is calculated by normalising two variables' covariance by their standard deviations.

Pearson correlation coefficient is r : $\frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2)][n\sum y^2-(\sum y)^2)]}}$

The resulting value quantifies the extent to which the variables change together.

For Fisher's Exact Tests

The Fisher's Exact Test is a statistical method used to determine the significance of the association between two categorical variables in a 2x2 contingency table. Unlike the Chi-Square Test, which relies on an approximation to the distribution of the test statistic under the null hypothesis of no association, Fisher's Exact Test calculates the exact probability of observing the data (or something more extreme) assuming no association between the variables.

A significant result suggests that the distribution of one variable is dependent on the distribution of the other, indicating an association between them.

Table 4: Summary Table of Independence of Predictor Variables

Analysis Type	Variable Pair	Statistic	Value	p-value	Interpretation
Correlation	Age vs Charges	Pearson r	0.299	-	Moderate positive correlation
Correlation	BMI vs Charges	Pearson r	0.198	-	Weak positive correlation
Correlation	Children vs Charges	Pearson r	0.068	-	Very weak positive correlation
Fisher's Exact Test	Smoker vs Sex		7.3929	0.006548	Significant association
Fisher's Exact Test	Region vs Has Children		1.5332	0.6757	No significant association
Fisher's Exact Test	BMI Category vs Smoker		571.72	0.6934	No significant association due to large p-value, but note on approximation warning

From Table 4, The Pearson correlation analysis provides insight into how certain numerical variables linearly relate to insurance charges, with age showing the strongest positive correlation among the variables tested. Chi-squared tests discover categorical variable relationships. The table provides a complete insurance dataset predictor variable correlation analysis. According to moderate positive correlation studies (r = 0.299), insurance rates climb with age. A slight positive link (r=0.198) implies BMI increases insurance costs. Insurance costs are not linearly related to children (r=0.068). This study reveals that age, smoking status, and other demographic and health characteristics affect insurance premiums.

Problem statement 3

To explore the predictive power of various factors on insurance charges, this study employs linear regression modelling alongside decision tree and random forest algorithms within the R environment. The analysis aims to determine the impact of demographic and health-related characteristics, such as age, Body Mass Index (BMI), number of children, gender, smoking status, and geographical region, on insurance charges using a comprehensive dataset. The analysis utilises the rpart and randomForest libraries in R to build models, evaluate their significance, and measure the impact of these features on charge forecasts.

Statistical analysis

The statistical analysis provided is based on two different regression models: Random Forest and Linear Regression, along with insights from a Decision Tree model. Each model has contributed unique insights into the factors that influence health insurance charges.

Table 5: Statistical Analysis Summary Table

Model Type	Metric/Variable	Value	Note/Significance
Random Forest	% Var Explained	91.59%	Model explains a high percentage of variance in charges
	Mean Sq. Residuals	12,331,346	Average squared difference between observed and predicted charges
Linear	R-squared	0.7509	Model explains ~75.09% of variance in

Regression			charges
	Adjusted R-squared	0.7494	Adjusted for the number of predictors in the model
	MSE	36,501,893.0074154	Mean Squared Error
coefficient	Intercept	-11,938.5386	($p < 0.001$)
	Age	256.8564	($p < 0.001$)
	BMI	339.1935	($p < 0.001$)
	Children	475.5005	($p < 0.001$)
	Sex (Male)	-131.3144	ns (not significant)
	Smoker (Yes)	23,848.5345	($p < 0.001$)
	Region (Northwest)	-352.9639	ns (not significant)
	Region (Southeast)	-1,035.0220	($p < 0.05$)
	Region (Southwest)	-960.0510	($p < 0.05$)
Decision Tree	-	-	Model provides a visual representation of decision rules for predicting charges

For Random Forest Analysis

The Random Forest model consistently captures complicated predictor-health insurance bill relationships with 91.59% variation explained. The Mean Squared Residuals are 12,331,346, which demonstrates prediction error, yet this figure indicates great performance given the significant variability. "Random forest includes construction of decision trees of the given training data and matching the test data with these."(R, P.T. et.al, 2015).This model is good in predicting insurance costs because it can detect subtle trends that linear techniques miss.

For Linear Regression Analysis

Table 6: Linear Regression Model Summary

Variable	**Coefficient (Estimate)**	**Std. Error**	**t value**	**p-value**	**Significance**
Intercept	-11,938.5386	987.81918	-12.085753	< 0.001	0.001
Age	256.8564	11.89885	21.5866552	< 0.001	0.001
BMI	339.1935	28.59947	11.8601306	< 0.001	0.001
Children	475.5005	137.80409	3.4505546	0.000577	0.001
Sex(male)	-131.3144	332.94544	-0.394402	0.693348	ns
Smoker (yes)	23,848.5345	413.15335	57.723202	< 0.001	0.001
Region (Northwest)	-352.9639	476.27579	-0.7410914	0.458769	ns
Region (Southeast)	-1,035.0220	478.69221	-2.162187	0.030782	0.05
Region (Southwest)	-960.0510	477.93302	-2.0087563	0.044765	0.05

The Linear Regression model covers 75.09% of the variation in insurance charges, as indicated by a R^2 value of 0.7509. Additionally, the model has beneficial predictive accuracy and durability against overfitting caused by many predictors, as evidenced by an adjusted R^2 of 0.7494. "Enhancing the standard significance test approach to display information about structural changes in regression relationships and to assess their significance" (Zeileis, A., Leisch, F., Hornik, K. and Kleiber, C et.al,2007)The model's Mean Squared Error (MSE) of 36,501,893 demonstrates a respectable level of prediction error, confirming its effectiveness in accurately projecting charges.

Decision Tree Analysis

The Decision Tree model offers a visual representation of the decision rules used to forecast charges, showcasing the logical significance of variables. "While growing a single tree is subject to small changes in the training data, random forests procedure is introduced to address this problem."(Zhang Z et.al, 2016)Although the summary lacks specific statistical data, this model is highly beneficial for interpreting complex relationships and nonlinear connections in a user-friendly.

Problem statement 4

The objective of this study is to analyse the relationship between predictor factors (age, BMI, children, smoking status, sex, and area) and a newly classified variable, CHARGE_split (Low, Average, High), in a health insurance dataset. Utilizing statistical tests like Wilcoxon and Chi-squared, along with visualizations, we assess differences in central tendencies and explore associations between these predictors and charge categories. The analysis is grounded on the assumption that all predictor variables are independent.

Statistical analysis

The goal of this analysis is to examine how charges are influenced by various factors, such as age, BMI (Body Mass Index), the number of children, smoking status, sex, and geographical region.

ASSUMPTIONS

In our analysis, we rely on certain assumptions tailored to our data and statistical methods. The Wilcoxon Rank Sum Test, used for comparing ordinal or continuous data, assumes similar distribution shapes between groups and serves as a non-parametric alternative when data don't follow a normal distribution. The Fisher's Exact Test, applied to categorical data, suits any sample size and requires data representation in a contingency table. These foundational assumptions ensure the integrity of our statistical findings.

HYPOTHESIS

Wilcoxon Rank Sum Test

- The null hypothesis (H0) : There is no difference in the distribution of the dependent variable between the two groups.
- Alternative Hypothesis (H1): There exists a disparity in the distribution of the dependent variable between the two groups.

Fisher's Exact Test:

- The null hypothesis H0: There is no association between the categorical variables being compared (e.g., smoker status and CHARGE_split categories).
- Alternative Hypothesis HA: There is a significant association between the variables.

The assumptions and hypotheses serve a vital role in accurately understanding the statistical test results and generating valid inferences about the relationships between variables in the data

For Wilcoxon Rank Sum Tests

The Wilcoxon rank sum test, also referred to as the Mann-Whitney U test when comparing two independent samples, is applied to determine whether two populations exhibit the same central tendency.

Test Statics is $U = n_1 n_2 + \frac{n_1 \quad (n_1+1)}{2} - R_1$

It is particularly helpful for non-parametric data.

Table 7: Summary of Wilcoxon Rank Sum Test Results

Variable	Comparison Groups	Test Name	W-Statistic	p-value	Conclusion
Age	Low vs. Average	Wilcoxon rank sum test	9936	< 0.001	Significant difference
Age	Low vs. High	Wilcoxon rank sum test	1755	< 0.001	Significant difference
Age	Average vs. High	Wilcoxon rank sum test	94900	0.00061	Significant difference
BMI	Low vs. Average	Wilcoxon rank sum test	2865	0.02378	Significant difference
BMI	Low vs. High	Wilcoxon rank sum test	813.5	0.6613	No significant difference
BMI	Average vs. High	Wilcoxon rank sum test	70007	< 0.001	Significant difference
Children	Low vs. Average	Wilcoxon rank sum test	7996.5	0.001817	Significant difference
Children	Low vs. High	Wilcoxon rank sum test	1422	0.001449	Significant difference
Children	Average vs. High	Wilcoxon rank sum test	108912	0.518	No significant difference
Charges	Low vs. Average	Wilcoxon rank sum test	10179	< 0.001	Significant difference
Charges	Low vs. High	Wilcoxon rank sum test	1782	< 0.001	Significant difference
Charges	Average vs. High	Wilcoxon rank sum test	0	< 0.001	Significant difference
BMI	vs Smoker	Fisher's Exact Test	$\chi^2 = 0.957$	0.6934	No significant difference
Smoker	No vs. Yes	Fisher's Exact Test	$\chi^2 = 582.55$	< 2.2e-16	Significant difference
Sex	vs. CHARGE_split	Fisher's Exact Test		6.839e-05	Significant difference

The table provides an overview of the statistical tests performed on both numerical and categorical variables within insurance data. Wilcoxon rank sum tests were employed for numerical variables such as age, BMI, children, and charges. Statistically significant variations were seen in age and charges among all comparison groups ($p < 0.001$). Additionally, BMI exhibited significant differences in the Low vs. Average and Average vs. High comparisons ($p = 0.02378$ and $p < 0.001$, respectively).

Problem statement 5

The goal is to examine the variation of interval predictor variables such as BMI, age, children, and insurance charges across various regions. It will be done by conducting Kruskal-Wallis tests and following Dunn's tests for post-hoc analysis. Furthermore, the correlation between categorical variables (smoker and sex) and region is evaluated using chi-square tests. Visualisations like as histograms, density plots, and boxplots are used to illustrate distributions and differences, providing visual support for statistical conclusions.

Statistical analysis

The statistical analysis uses Shapiro-Wilk, Kruskal-Wallis, Dunn's post-hoc, and Chi-squared tests to examine the distribution and variability of health-related variables across various geographic regions. The study revealed significant regional variations in BMI, with smoker status and sex being unrelated to geographical location, showcasing a focused exploration of geographical disparities in health metrics.

ASSUMPTIONS

In our analysis, we employ several statistical tests each with specific assumptions. The Shapiro-Wilk test checks for normal distribution, assuming a significant p-value indicates non-normality. The Kruskal-Wallis test, suitable for non-normal distributions, requires similar distribution shapes across groups without assuming normality, focusing on median differences. Both tests assume sample independence. While homogeneity of variances is a common assumption for ANOVA-like tests, the Kruskal-Wallis test is less affected by this assumption due to its non-parametric nature. Lastly, the Chi-Square test assumes categorical data and independence of observations, aiming to detect associations between categorical variables. These foundational assumptions ensure the tests are correctly applied and interpreted.

HYPOTHESIS

Shapiro-Wilk Test:

- H0: The distribution of charges is normally distributed.
- HA: The distribution of charges deviates from normality.

Kruskal-Wallis Test:

- H0: There is no difference in the median values of BMI, age, children, and charges across regions.
- HA: At least one region has a different median value than the others.

Dunn's Test with Bonferroni Correction (following significant Kruskal-Wallis results):

- H0: There are no differences between any pair of groups.
- HA: There is a significant difference between at least one pair of groups.

Chi-Square Test of Independence (for smoker and sex by region):

- H0: There is no association between smoker status and region or between sex and region.
- HA: There is an association between smoker status and region or between sex and region.

For Shapiro-Wilk Test for Normality

The Shapiro-Wilk test applied to insurance charges indicated non-normality of the data distribution ($W=0.81469$, $p<2.2e-16$), suggesting the charges are not distributed as per a normal distribution. This outcome necessitates the use of non-parametric methods for further statistical analysis of the charges data.

The Shapiro-Wilk statistic is $W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_{i-\bar{x}})^2}$

The results show that the charges distribution is not normal ($W=0.81469$, $p<2.2e{-}16$, $^{0}<<2.2$).

For Kruskal-Wallis Test

It is a statistical technique used to determine if several samples are drawn from the same fundamental probability distribution, without making any assumptions about the shape or parameters of the distribution. It is employed when the conditions required for ANOVA are not satisfied, such as having a non-normal distribution.

The Kruskal-Wallis statistic is $H = \frac{12}{N(N+1)}\sum_{i=1}^{g}\frac{R_i^2}{n_i} - 3(N+1)$

It performed the examination to assess BMI, age, number of children, and charges based on geographical location. There were significant variations in BMI across different regions ($H=94.689$, $p<2.2e{-}16$).

For Dunn's Test for Multiple Comparisons with Bonferroni Correction

The Dunn's test is applied after identifying significant differences with the Kruskal-Wallis test to do pairwise comparisons between groups, while correcting for Type I error using Bonferroni correction.
The Adjusted P-value (Bonferroni Correction) is: $p_{adjusted=p*\frac{m(m-1)}{2}}$

Statistically significant changes were observed between several combinations of regions after modification.

Table 8: Summary of Results from Combined Statistical Tests

Test Type	Variable/Comparison	Statistic or Z-Value	Degrees of Freedom (df)	P-value	Adjusted P-value	Significant?
Kruskal-Wallis Test	BMI	94.689	3	< 2.2e-16	-	Yes
Kruskal-Wallis Test	Age	0.41382	3	0.9374	-	No
Kruskal-Wallis Test	Children	2.3754	3	0.4982	-	No
Kruskal-Wallis Test	Charges	4.7342	3	0.1923	-	No
Dunn's Post-Hoc Test	BMI: Northeast - Southeast	-8.413665	-	-	1.192184e-16	Yes
Dunn's Post-Hoc Test	BMI: Northwest - Southeast	-8.2767381	-	-	3.797856e-16	Yes
Dunn's Post-Hoc Test	BMI: Northeast - Southwest	-3.0369246	-	-	7.170156e-03	Yes
Dunn's Post-Hoc Test	BMI: Northwest - Southwest	-2.8993878	-	-	1.121676e-02	Yes
Dunn's Post-Hoc Test	BMI: Southeast - Southwest	5.2964216	-	-	3.542820e-07	Yes
Chi-Squared Test	Smoker by Region	7.3435	3	0.06172	-	No
Chi-Squared Test	Sex by Region	0.43514	3	0.9329	-	No

The table displays the results of statistical studies conducted on health data by area. These analyses utilised Kruskal-Wallis and Dunn's post-hoc tests for continuous variables, as well as Chi-squared testing for categorical variables. The differences in BMI between regions were very significant (p < 2.2e-16), particularly between the Northeast and Southeast regions (p = 1.192184e-16) as indicated by Dunn's test. According to the Kruskal-Wallis results, there were no significant differences in age, number of children, and charges between regions. Furthermore, the Chi-squared tests revealed that there was no notable regional correlation between smoking status and sex.

CONCLUSION

The analysis of health insurance data revealed significant regional variations in BMI but not in age, number of children, or insurance charges. Significant insights were gained into the non-normal distribution of insurance charges and the independence of smoking status and sex from geographical location. These findings underscore the potential of predictive modelling to enhance premium pricing, policy customization, and healthcare management, highlighting the intricate web of factors that influence health insurance premiums.

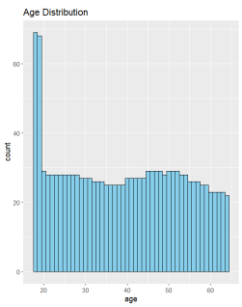
REFERENCE

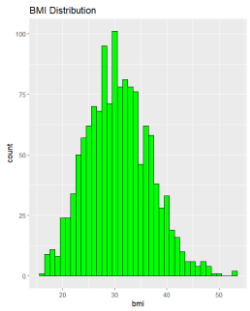
1. AnaCarolAsics (2022). *The Most Brilliant Statistical & Data Analysis Quotes*. [online] LEARN STATISTICS EASILY. Available at: https://statisticseasily.com/2022/05/17/statistics-quotes/.
2. Gartner (n.d.). *Definition of Predictive Modeling - Gartner Information Technology Glossary*. [online] Gartner. Available at: https://www.gartner.com/en/information-technology/glossary/predictive-modeling#:~:text=Predictive%20modeling%20is%20a%20commonly.
3. www.carepatron.com. (n.d.). *Predictive Modeling in Healthcare*. [online] Available at: https://www.carepatron.com/guides/predictive-modeling-in-healthcare [Accessed 15 Mar. 2024].
4. Li, J. (2019). A Critical Review of Spatial Predictive Modeling Process in Environmental Sciences with Reproducible Examples in R. *Applied Sciences*, 9(10), p.2048. doi:https://doi.org/10.3390/app9102048.
5. Rahlf, T. (2017). *Data Visualisation with R*. Springer.
6. PennState: Statistics Online Courses. (n.d.). *1.6 - (Pearson) Correlation Coefficient, \(r\) | STAT 501*. [online] Available at: https://online.stat.psu.edu/stat501/lesson/1/1.6#:~:text=for%20r%2C%20namely%3A-.

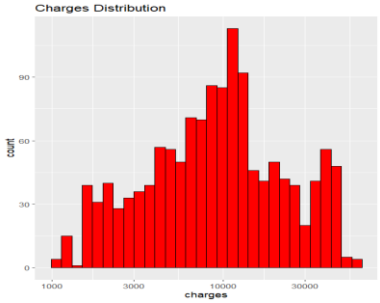
7. Taiyun Wei [cre, aut] (2021) Corrplot: Visualization of a correlation matrix version 0.92 from cran, version 0.92 from CRAN. Available at: https://rdrr.io/cran/corrplot/ (Accessed: 2nd March 2024).
8. Zeileis, A., Leisch, F., Hornik, K. and Kleiber, C. (2002). strucchange: AnRPackage for Testing for Structural Change in Linear Regression Models. *Journal of Statistical Software*, [online] 7(2). doi:https://doi.org/10.18637/jss.v007.i02.
9. Simulator, A. and Acumen SimulatorAcumen Simulator 2311 silver badge77 bronze badges (1963) Splitting into two identical standard deviations, Cross Validated. Available at: https://stats.stackexchange.com/questions/327363/splitting-into-two-identical-standard-deviations (Accessed: 06 March 2024).
10. Zhang, Z. (2016). Decision tree modeling using R. *Annals of Translational Medicine*, 4(15), pp.275–275. doi:https://doi.org/10.21037/atm.2016.05.14.
11. Zhou, H., Deng, Z., Xia, Y. and Fu, M. (2016). A new sampling method in particle filter based on Pearson correlation coefficient. *Neurocomputing*, 216, pp.208–215. doi:https://doi.org/10.1016/j.neucom.2016.07.036.
12. Technik, D. (2019). *Shapiro-Wilk Test for Normality in R | R-bloggers*. [online] Available at: https://www.r-bloggers.com/2019/08/shapiro-wilk-test-for-normality-in-r/.
13. R, P.T. (2015). A Comparative Study on Decision Tree and Random Forest Using R Tool. *IJARCCE*, 4(1), pp.196–199. doi:https://doi.org/10.17148/ijarcce.2015.4142.
14. LaMorte, W. (2017). *Mann Whitney U Test (Wilcoxon Rank Sum Test)*. [online] sphweb.bumc.bu.edu. Available at: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Nonparametric/BS704_Nonparametric4.html.
15. www.statskingdom.com. (n.d.). *Shapiro Wilk Test*. [online] Available at: https://www.statskingdom.com/doc_shapiro_wilk.
16. www.sciencedirect.com. (n.d.). *Kruskal Wallis Test - an overview | ScienceDirect Topics*. [online] Available at: https://www.sciencedirect.com/topics/medicine-and-dentistry/kruskal-wallis
17. Real-statistics.com. (2024). Available at: https://real-statistics.com/one-way-analysis-of-variance-anova/kruskal-wallis-test/dunns-test-after-kw].
18. Zach (2022) How to perform a Bonferroni correction in R, Statology. Available at: https://www.statology.org/bonferroni-correction-in-r/ (Accessed: 07 March 2024).

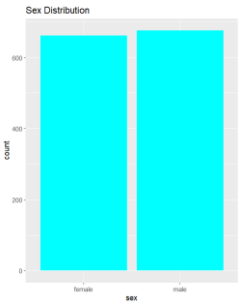
APPENDIX

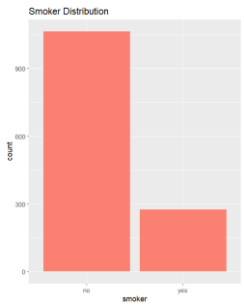
Problem 1:

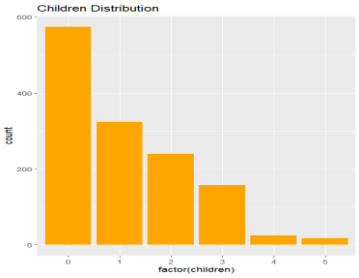




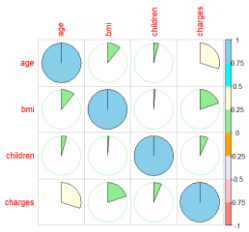


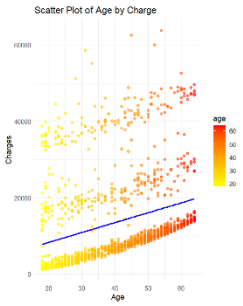


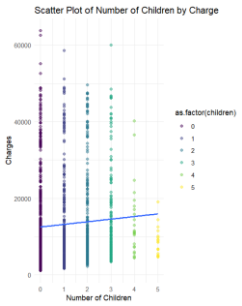


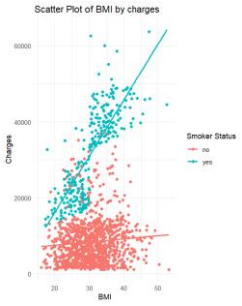


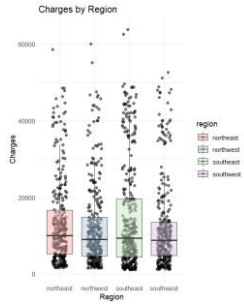
Problem 2:

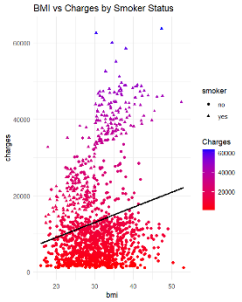


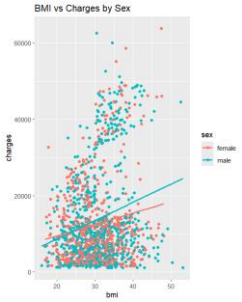


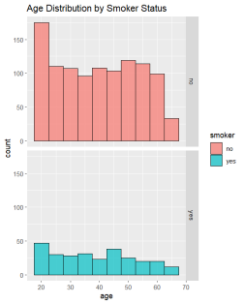


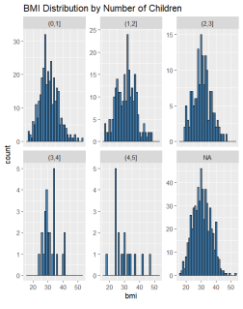


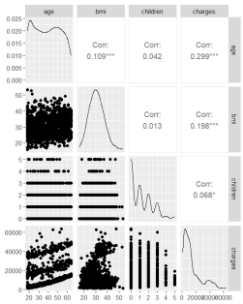


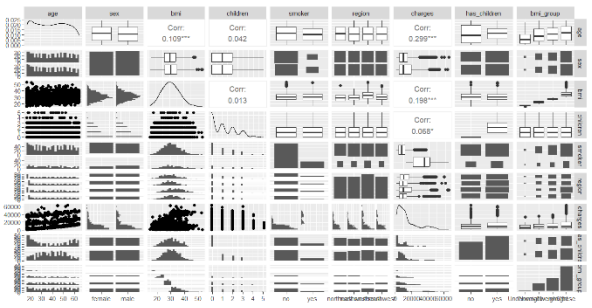




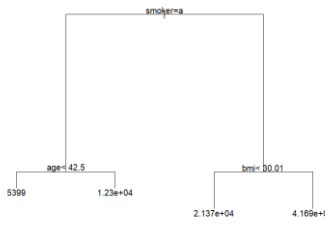




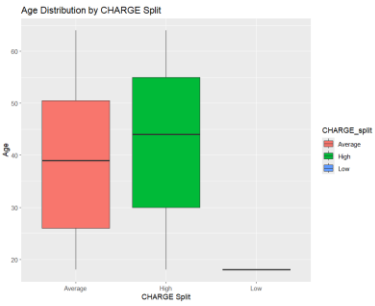


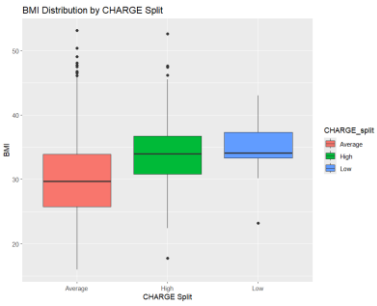


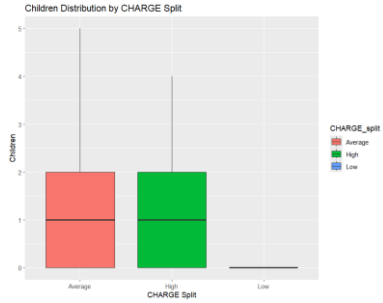
Problem 3

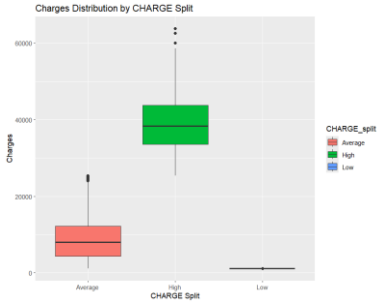


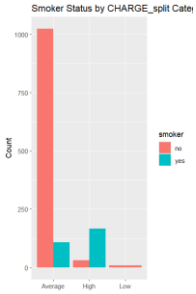
Problem 4

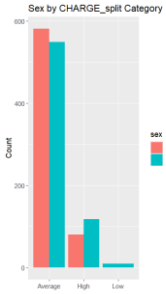


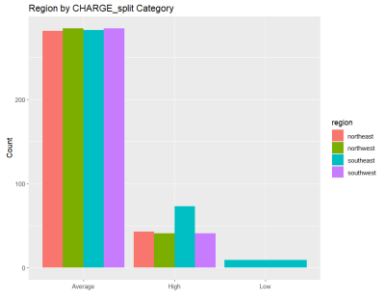












Problem 5