AYOMIDE OLUWATOBI ADEKOYA

ANALYSIS OF HEALTH INSURANCE DATASET

N1223955

TABLE OF CONTENT

Introduction

The Health Insurance Dataset contains information on personal attributes, geographic factors, and medical insurance charges for 1338 US citizens. This dataset can be used to study how various features(preditor variable) such as age, gender, BMI, family size, smoking habits, and geographic region impact medical insurance costs(response variable).

Variables Description

Age: The insured person's age in years.
Sex: Gender of the insured (male(0) or female(1)).
BMI: Body Mass Index, a measure of body fat.
Children: Number of dependents covered.
Smoker: Whether the insured is a smoker(1) or non-smoker(0).
Region: Geographic area of coverage (southwest(1),southeast(2),northwest(3),northeast(4)).
Charges: Medical insurance costs incurred by the insured person in $.
additional variable
Age category:variable represents age intervals created from the "age" data. Ages are grouped into categories based on the intervals defined by the breakpoints: 15-24, 25-34, 35-44, 45-54, and 55-64. Each individual's age is assigned to one of these categories based on their age value.

Bmi category:variable categorizes BMI (Body Mass Index) data into specific intervals. BMIs are grouped into categories based on predefined breakpoints: 15-24, 25-34, 35-44, and 45-54. Each individual's BMI value is then assigned to one of these categories.

	age	sex	bmi	children	smoker	region	charges	Age_category	Bmi_category
1	19	1	27.900	0	1	1	16884.924	1	2
2	18	0	33.770	1	0	2	1725.552	1	2
3	28	0	33.000	3	0	2	4449.462	2	2
4	33	0	22.705	0	0	3	21984.471	2	1
5	32	0	28.880	0	0	3	3866.855	2	2

Objective

Utilize the Health Insurance Dataset (Health-Insurance-Dataset.csv) to analyze the correlation between personal attributes (age, gender, BMI, family size, smoking habits), geographic factors, and medical insurance charges ($) among 1338 US citizens. This dataset enables the exploration of how these variables affect insurance costs and facilitates the development of predictive models for estimating healthcare expenses.

Testing for Normality using Kolmogorov-Smirnov Test

In statistical analysis, assessing the normality of data is essential for making valid inferences and selecting appropriate statistical tests. The Kolmogorov-Smirnov test is a widely used method to determine whether a dataset follows a normal distribution. In this report, we will utilize the Kolmogorov-Smirnov test to evaluate the normality assumption for the variables in the Health Insurance Dataset.

Null Hypothesis (H0): The data follows a normal distribution.
Alternative Hypothesis (H1): The data does not follow a normal distribution.

By conducting the Kolmogorov-Smirnov test, we aim to statistically evaluate whether the distribution of each variable, such as age, BMI, number of children, and insurance charges, deviates significantly from a normal distribution. This test will provide valuable insights into the underlying distribution characteristics of the dataset and help guide the choice of appropriate statistical methods for further analysis.

In the sections that follow, we will examine the Kolmogorov-Smirnov test results as we explore the normalcy assumption and its implications for examining healthcare and insurance expenditures.

Visualization of data distribution

Appendix_1:

Plot(1) show the distribution of age data
Plot(2) show the distribution of region data
Plot(3) show the distribution of bmi data
Plot(4) show the distribution of charges data
Plot(5) show the distribution of sex data
Plot(6) show the distribution of smoker data
Plot(7) show the distribution of children data

we noticed that plot(1),Plot(2),Plot(4),Plot(5),Plot(6) and Plot(7) did not exhibit a bell curve shape typically associated with a normal distribution. This observation suggests that the data may not follow a normal distribution.

We intend to officially evaluate the normality of the datas by running a Kolmogorov-Smirnov test in order to support this assertion. Our goal in doing this statistical test is to assess statistically how well the observed datas distribution agrees with a hypothesized normal distribution. The age variable in the dataset's normalcy assumption will be quantitatively supported by the Kolmogorov-Smirnov test results.

Null Hypothesis (H0): The data follows a normal distribution.
Alternative Hypothesis (H1): The data does not follow a normal distribution.

By setting the confidence level at 95%, we aim to evaluate whether the observed data significantly deviates from a normal distribution pattern. The statistical test results will help us make an informed decision regarding the normality assumption of the data and its implications for further analysis.

variable	Test type	P-values	Decision on H0
age	Kolmogorov-Smirnov	1.143e-07	reject
bmi	Kolmogorov-Smirnov	0.3218	accept
sex	Kolmogorov-Smirnov	2.2e-16	reject
children	Kolmogorov-Smirnov	2.2e-16	reject
region	Kolmogorov-Smirnov	2.2e-16	reject
charges	Kolmogorov-Smirnov	2.2e-16	reject
smoker	Kolmogorov-Smirnov	2.2e-16	reject

From the table above, it shows that we don't have enough evidence to reject the null hypothesis for the bmi data, which implies that the bmi data is normally distributed. Therefore, we accept the null hypothesis for this data.

However, for the age, sex, children, region, charges, and smoker data, we reject the null hypothesis, indicating that there is enough evidence to suggest that these variables do not follow a normal distribution. This rejection implies that the assumption of normality may not be valid for these specific variables.

Summary of statistics of all variable

variable	Min	Max	Mean	1^{st} quartile	Median	3^{rd} quartile	SD	Mode
age	18.00	64.00	39.21	27.00	39.00	51.00	14.04996	18
bmi	15.96	53.13	30.66	26.30	30.40	34.69	6.098187	32.3
sex	0.00	1.00	0.4948	0.0000	0.0000	1.0000	0.5001596	0
children	0.00	5.00	1.0950	0.000	1.000	2.000	1.205493	0
region	1.00	4.00	2.4840	2.000	2.000	3.000	1.104885	2
smoker	0.00	2.00	0.2048	0.0000	0.0000	0.0000	0.403694	0
charges	1122	63770	13270	4740	9382	16640	12110.01	1639.563

The table above is summary of our dataset the analysis of the age data revealed an average age of 39.21 with a standard deviation of 14.04996. The mean age serves as the central tendency of the dataset, indicating that the average age among the sample population is approximately 39 years old. the standard deviation of

14.04996 suggests a moderate level of variability in ages within the dataset, indicating that ages are spread out to some extent around the average age of 39.21.

The analysis of the bmi data revealed a mean BMI of 30.66 with a standard deviation of 6.098187. The standard deviation of 6.098187 suggests a moderate level of variability in BMI within the dataset, indicating that there is some range in BMI values around the average of 30.66.

The analysis of the charges data revealed a median charge amount of 9382 with an inter-quartile range (IQR) of 11900. The median charge serves as a central measure of the middle value of the data-set, indicating that half of the charges fall below 9382 and half fall above this value. The IQR, which represents the range of values between the first and third quartiles, gives insight into the spread of data within the middle 50% of the dataset.

the analysis of the sex data revealed that the mode value for sex is male (0), with a frequency of 50.52% in the sample. indicating that the majority of individuals in the sample are male.

The analysis of the smoker data revealed that the mode value for the smoking status variable is "no" (0), with a percentage frequency of 79.52% in the sample. indicating that the majority of individuals in the sample are non-smokers.

The finding that non-smokers make up 79.52% of the sample suggests a significant presence of individuals who do not smoke in the dataset.

the analysis of the region data revealed that the mode value for the region variable is "southeast" (2), with a percentage frequency of 27.20% in the sample. indicating that a significant proportion of individuals in the sample come from the southeast region.

the analysis of the children data revealed that the individual with 0 number of children has a percentage frequency of 42.89% in the sample.

The finding that individuals with 0 children make up 42.89% of the sample suggests that a significant portion of the population represented in the dataset does not have any children

In summary An analysis of the dataset revealed a median charge amount of $9382 and an inter-quartile range of $11900, indicating a balanced distribution of charges. The average age was found to be 39years, with an average BMI of 30.66. The mode value for sex was male (0) at 50.52%, while non-smokers dominated the sample at 79.52%. Individuals from the southeast region accounted for 27.20% of the dataset. The most common number of children was 0, representing 42.89% of the sample. These findings provide insight into the demographic makeup, health behaviors, and general characteristics of the participants.

Test of Independence and association

A test of independence or association is used to determine if there is a relationship between two variables. This type of test is commonly used in statistics to analyze categorical data and can be applied to various fields such as social sciences, marketing, and healthcare.

Correlation is a statistical test that measures the strength and direction of a relationship between two continuous variables. The correlation coefficient can range from -1 to 1, with 0 indicating no relationship, 1 indicating a perfect positive relationship, and -1 indicating a perfect negative relationship.

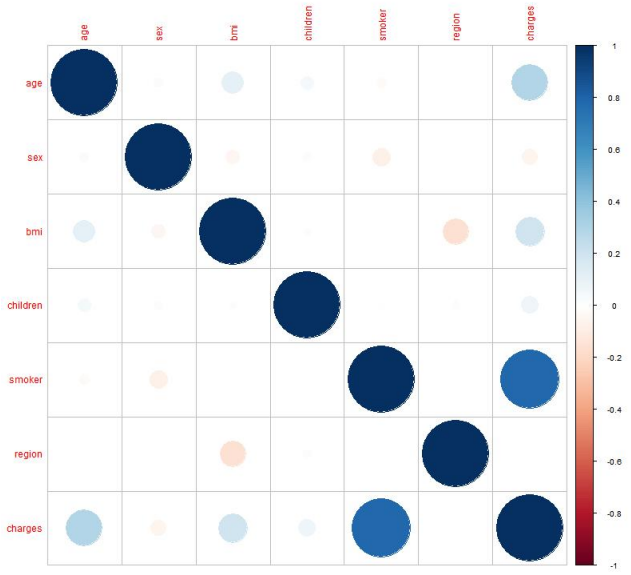
Chi-square test, on the other hand, is used to determine if there is a significant association between two categorical variables. It compares the observed frequencies of the data with the expected frequencies and calculates a chi-square statistic. The p-value of the test is used to determine if the variables are independent or if there is a significant association between them.

Null Hypothesis (H0): There is no association between the variable(I.e variable are independent). Alternative Hypothesis (H1): there is an association between variable (I.e variable are dependent)

variables	Type of test	X-squared	df	P-value	Decision on H0
Sex vs region	Chi-squared	0.43514	3	0.9329	accept
Sex vs age	Chi-squared	1.6405	46	1	accept
Sex vs bmi	Chi-squared	529.17	547	0.7001	accept
Sex vs children	Chi-squared	0.73521	5	0.981	accept
Sex vs smoker	Chi-squared	7.3929,	1	0.006548	reject
Children vs age	Chi-squared	450.98	230	2.2e-16	reject
Children vs bmi	Chi-squared	2771.4	2735	0.3091	accept
Children vs region	Chi-squared	13.773	15	0.5428	accept
Children vs smoker	Chi-squared	6.8877	5	0.2291	accept
Region vs age	Chi-squared	136.78	138	0.5133	accept
Region vs bmi	Chi-squared	2940.8	1641	2.2e-16	reject
Region vs smoker	Chi-squared	7.3435	3	0.06172	accept
Smoker vs age	Chi-squared	52.429	46	0.2388	accept
Smoker vs bmi	Chi-squared	571.72	547	0.2247	accept
Charges vs sex	Kruskal-Wallis	0.1204	1	0.7286	accept
Charges vs children	Kruskal-Wallis	29.487	5	1.86e-05	reject
Charges vs smoker	Kruskal-Wallis	588.52	1	< 2.2e-16	reject
Charges vs region	Kruskal-Wallis	4.7342	3	0.1923	accept

Null Hypothesis (H0): There is no significant relationship between the variable(I.e variable are independent). Alternative Hypothesis (H1): there is an significant relationship between variable (I.e variable are dependent)

variables	Type of test	r-value	P-value	Decision on H0
Age vs bmi	spearman	0.107736	7.859e-05	Reject
Charges vs bmi	spearman	0.1193959	1.193e-05	reject
Charges vs age	spearman	0.5343921	< 2.2e-16	reject



From the correlation matrix, it is observed that:

-There is a weak positive correlation between age and BMI, children and age, and children and charges.

- There is a weak negative correlation between sex and BMI, sex and smoker, and sex and charges.

- There is a moderate negative correlation between region and BMI.

- There is a moderate positive correlation between age and charges, and between BMI and charges.

- There is a strong positive correlation between smoker and charges.

Linear regression model

Linear regression is a commonly used statistical technique that aims to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. In a simple linear regression model with one independent variable, the relationship between the dependent variable

Linear regression analysis can be extended to multiple linear regression, where the model includes more than one independent variable. This allows for the examination of the combined effect of multiple predictors on the dependent variable. Linear regression is widely used in various fields, such as economics, social sciences, healthcare, and engineering, to analyze and predict relationships between variables.

$$Y = C0 + C1X1 + C2X2 + C3X3 + C4X4 + C5X5 + C6X6$$

- Where:
  Y is the dependent variable (medical insurance charges $).
  X1 represents age.
  X2 represents gender (coded as 0 for female, 1 for male).
  X3 represents BMI (body mass index).
  X4 represents family size.
  X5 represents smoking habits (coded as 0 for non-smoker, 1 for smoker).
  X6 represents geographic factors.
C0 is the intercept of the line.
C1,C2,C3,C4,C5,C6 are the coefficients for each respective independent variable.
  ε is the error term capturing unexplained variability.

In multilinear regression, the objective remains the same to estimate coefficients (C0,C1,C2,C3,C4,C5,C6.) that minimize the sum of squared differences between observed data points and predicted values from the regression equation. This approach quantifies relationships between multiple independent variables and the dependent variable, enabling predictions based on these variables

	Estimate	Standard error	t value	Pr(>|t|)
Intercept	-13361.12	1087.67	-12.284	< 2e-16
age	257.29	11.89	21.647	< 2e-16
sex	131.11	332.81	0.394	0.693681
bmi	332.57	27.72	11.997	< 2e-16
children	479.37	137.64	3.483	0.000513
smoker	23820.43	411.84	57.839	< 2e-16
region	353.64	151.93	2.328	0.020077

Based on the regression summary provided, the estimated regression model can be written as:

Charges = -13361.12 + 257.29(age) + 131.11(sex) + 332.57(bmi) + 479.37(children) + 23820.43(smoker) +

353.64(region)

Interpretation of the coefficients:
- tThe intercept term (-13361.12) represents the estimated charges when all other variables are zero (age, sex, BMI, children, smoker status, and region).
- The coefficient for age (257.29) suggests that for each additional year of age, charges are expected to increase by $257.29, holding all other variables constant.
- The coefficient for sex (131.11) is not statistically significant (p-value > 0.05), indicating that gender may not have a significant influence on charges.
- The coefficient for BMI (332.57) indicates that for each one-unit increase in BMI, charges are expected to increase by $332.57, holding all other variables constant.
- The coefficient for children (479.37) suggests that for each additional child covered, charges are expected to increase by $479.37, holding all other variables constant.
- The coefficient for smoker (23820.43) is highly significant, indicating that smokers are expected to have significantly higher charges compared to non-smokers.
- The coefficient for region (353.64) is statistically significant (p-value < 0.05), suggesting that charges vary based on the geographic region of coverage.

Overall, this regression model suggests that age, BMI, number of children, smoker status, and region have a significant influence on medical insurance charges, with smoker status being the most influential variable in predicting charges.

Charge split

The Kruskal-Wallis rank sum test was used to assess the differences in central tendencies of various predictor variables with respect to the newly formed categorical variable CHARGE-split.

Den_graph(3),Den_graph(5) and Den_graph(6)
The densities for these categories overlap significantly. This suggests that the distributions of charges across these categories might be similar, and the means may not be significantly different.
The shapes of the distributions are identical for these categories. This further strengthens the possibility of no significant difference in means across these groups ,further testing will be carried out using kruskal-wallis test.

Hypotheses:
Null Hypothesis (H0): There is no significant difference in central tendencies of the predictor across charge-split
Alternative Hypothesis (H1): There is a significant difference in central tendencies of the predictor across charge-split

The results of the Kruskal-Wallis rank sum tests between charge-split are as follows

	Kruskal-Wallis chi-squared	df	p-value	Decision on H0
age	348.75	1	< 2.2e-16	reject
bmi	10.649	1	0.001102	reject
sex	0.011951	1	0.9129	accept
children	0.13303	1	0.7153	accept
region	344.3	1	0.08619	accept
smoker	2.9441	1	< 2.2e-16	reject

Overall, the results suggest that age, BMI, and smoking status have a significant impact on the medical insurance costs incurred by the insured individuals, while sex, number of children, and geographic region do not show significant differences across the CHARGE-split categories.

Interval predictor across region variable

The Kruskal-Wallis rank sum test was used to assess differences in central tendencies of interval predictor variables, specifically age and BMI, with respect to different geographic regions. Additionally, post-hoc Bonferroni tests were conducted to further analyze the differences between regions.

Hypotheses:
Null Hypothesis (H0): There is no significant difference in central tendencies of age and BMI across the geographic regions.
Alternative Hypothesis (H1): There is a significant difference in central tendencies of age and BMI across the geographic regions.

data: age and region
Kruskal-Wallis chi-squared = 0.4138, df = 3, p-value = 0.9374

The Kruskal-Wallis test results for age by region showed a non-significant p-value of 0.9374, indicating that there were no significant differences in central tendencies of age across the geographic regions. The post-hoc Bonferroni tests also confirmed that there were no significant differences in age between any pair of regions.

	northeast	northwest	southeast
northwest	-0.063679 (1.0000)		
southeast	0.323533 (1.0000)	0.389305 (1.0000)	
southwest	-0.302681 (1.0000)	-0.239185 (1.0000)	-0.635167 (1.0000)

p-values adjusted using Bonferroni correction.

data: bmi and region
Kruskal-Wallis chi-squared = 94.6886, df = 3, p-value =< 2.2e-16

The Kruskal-Wallis test results for BMI by region revealed a highly significant p-value of < 2.2e-16, indicating significant differences in central tendencies of BMI across the geographic regions. The post-hoc Bonferroni tests also confirmed that there is a significant differences in bmi between some pair of regions.

	northeast	northwest	southeast
northwest	-0.139771 (1.0000)		
southeast	-8.413665 (0.0000*)	-8.276738 (0.0000*)	
southwest	-3.036924 (0.0072*)	-2.899387 (0.0112*)	5.296421 (0.0000*)

Values in parentheses represent the p-values.
Significant p-values (below the significance level, typically 0.05) are marked with an asterisk (*). "northeast", "northwest", "southeast", and "southwest" represent different regions being compared.

Overall, the analysis indicates that there are significant variations in BMI levels across different geographic regions. Specifically, all pairs of regions, except for the northwest-northeast pair, show statistically significant differences in mean BMI values. Additionally, while there were no significant differences observed in age across regions.

Conclusion

Based on the correlation analysis, it is evident that age and BMI exhibit a positive moderate correlation, while smoking status shows a strong positive correlation with insurance charges. Furthermore, significant differences were observed in charges between individuals with high and low charges for age, BMI, and smoking status.

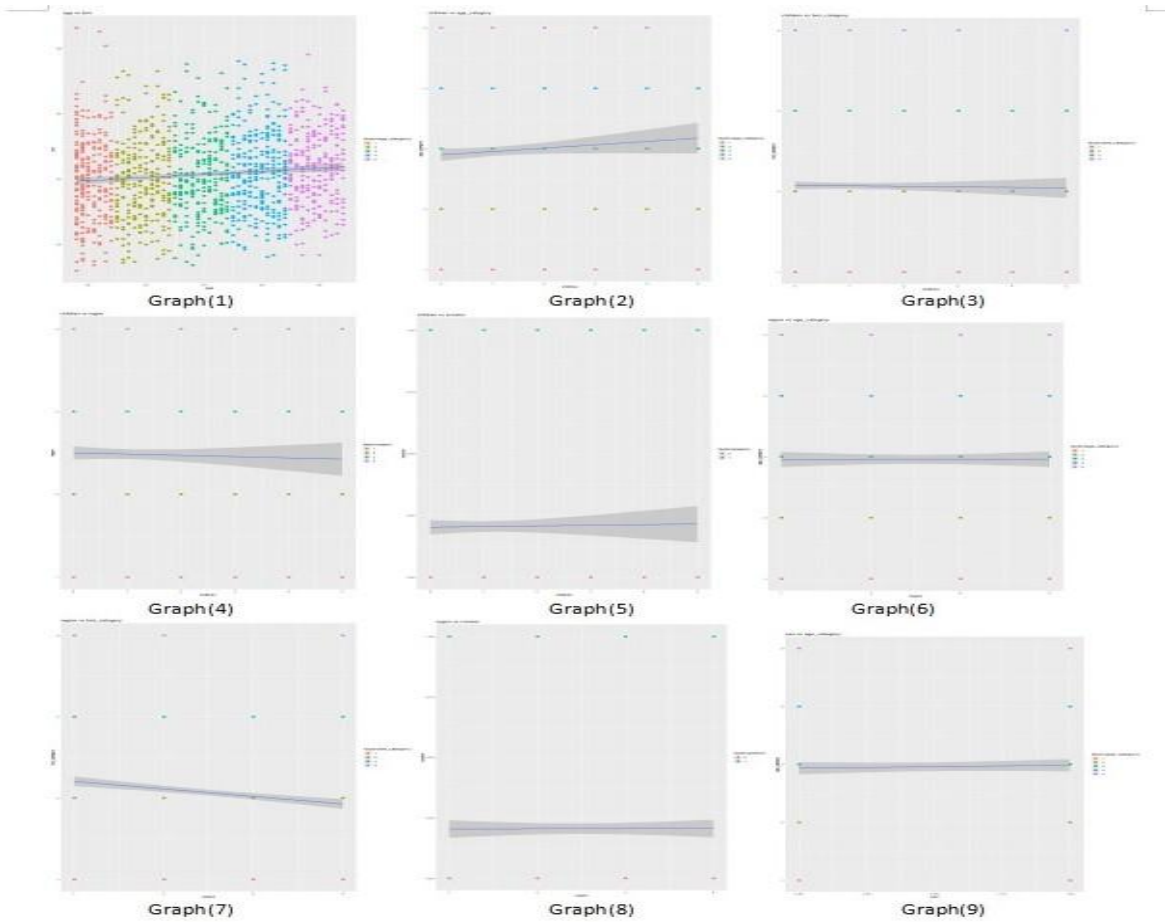
In our regression analysis, it is apparent that smokers tend to pay higher insurance fees compared to non-smokers, along with older individuals and those with higher BMI values. These variables smoking status, age, and BMI emerged as significant predictors of insurance charges, indicating their substantial influence on the costs incurred.

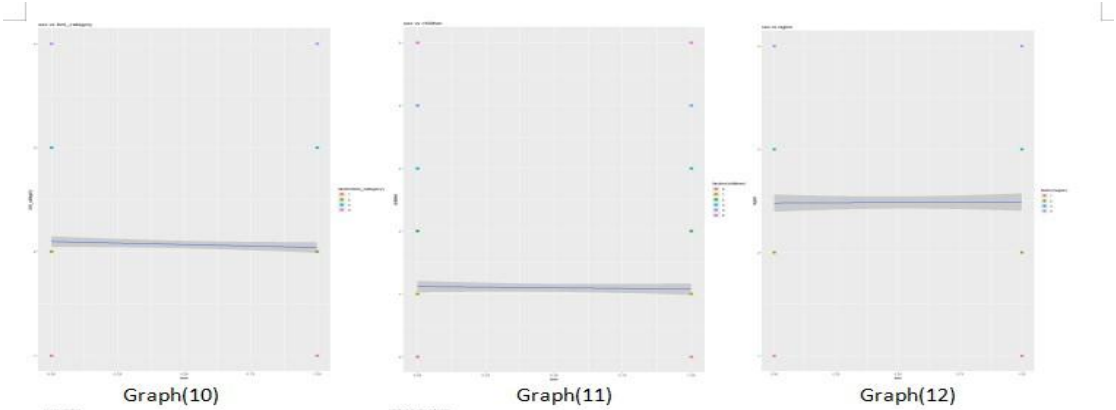
From the comprehensive analysis, it is clear that older individuals, smokers, and those with higher BMI tend to incur higher insurance charges. However, smoking emerges as a particularly significant factor influencing insurance costs, as evidenced by its strong correlation with charges and its significant predictive power in the regression model.

Appendix

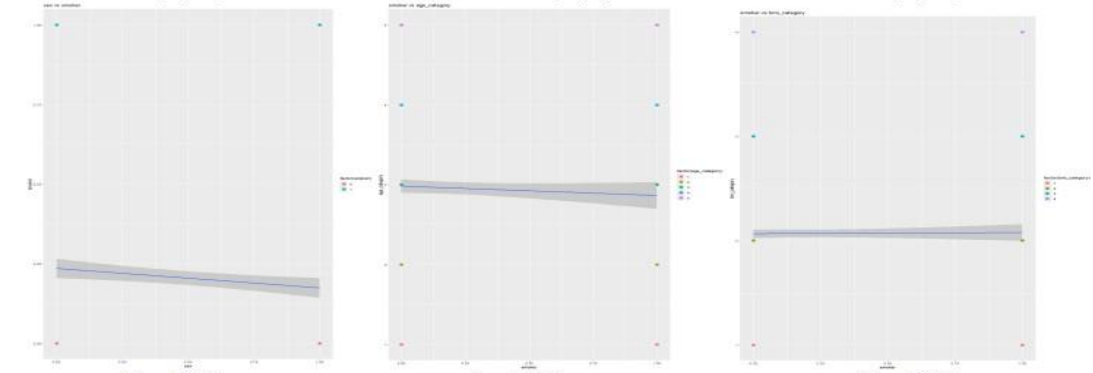
Appendix_i

Appendix_ii

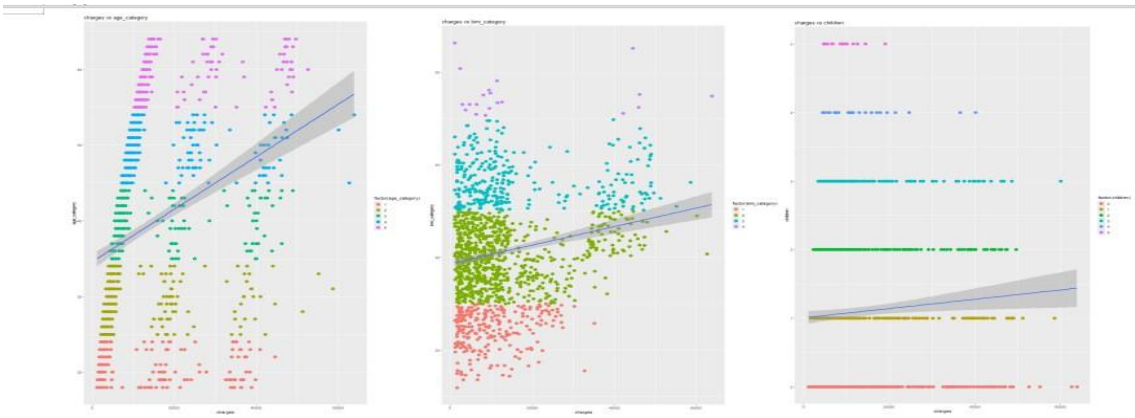




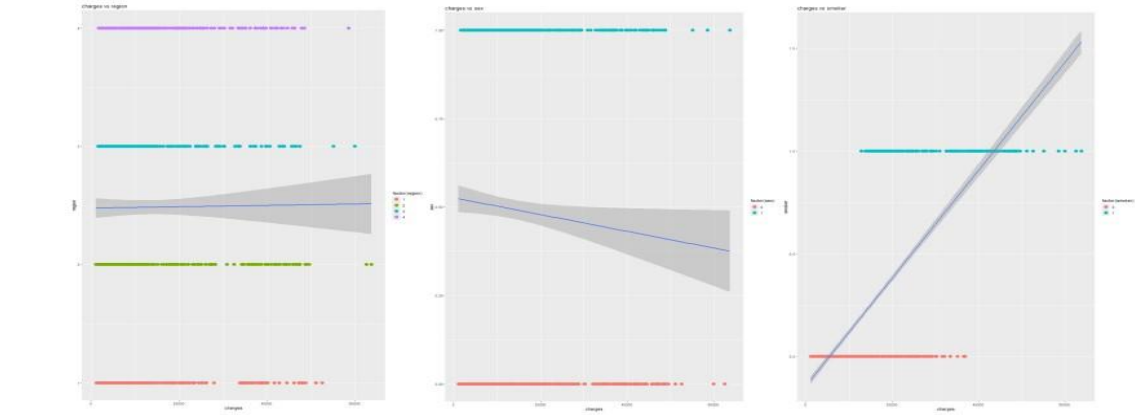
Graph(10) Graph(11) Graph(12)



Graph(13) Graph(14) Graph(15)

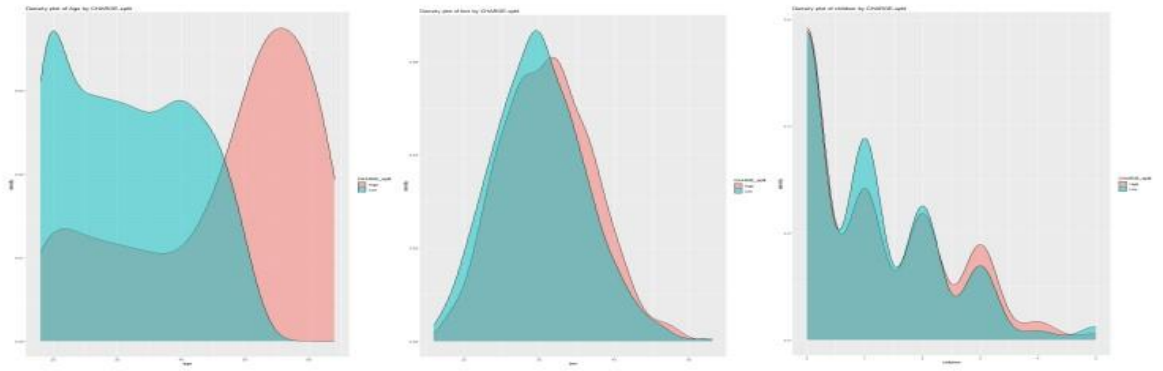


Graph(16) Graph(17) Graph(18)



Graph(19) Graph(20) Graph(21)

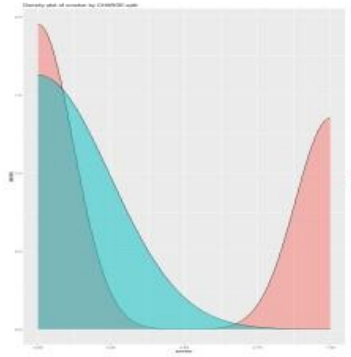
Appendix iii



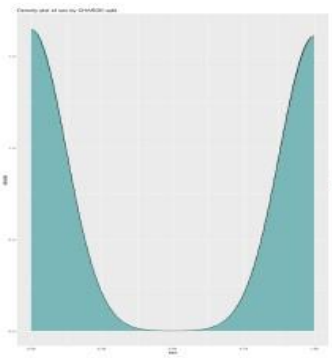
Den_graph(1)

Den_graph(2)

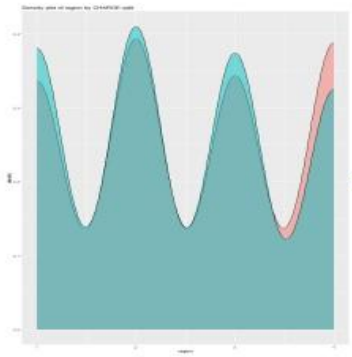
Den_graph(3)



Den_graph(4)



Den_graph(5)



Den_graph(6)

Refrence

[1]Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. (Journal not specified), February. Accessed March 7, 2024..

[2]Sun, Y., Wang, X., Zhang, C., & Zuo, M. (2023). Multiple Regression: Methodology and Applications. Highlights in Science Engineering and Technology, 49*(5), 542-548. DOI: 10.54097/hset.v49i.8611. Accessed March 8, 2024.

[3]Tutorialspoint. (n.d.). How to Perform Post Hoc Test for Kruskal-Wallis in R. Retrieved from https://www.tutorialspoint.com/how-to-perform-post-hoc-test-for-kruskal-wallis-in-r. Accessed March 11, 2024.

[4]Udemy. (n.d.). Applied Statistical Modeling for Data Analysis in R. Retrieved from https://www.udemy.com/course/applied-statistical-modeling-for-data-analysis-in-r/learn/lecture/7090304. Accessed March 4, 2024.