Nottingham Trent University

Statistical Data Analysis and Visualization
Determining factors for health insurance charge

Dai Nam Vo N1237846

9th March 2024

Table of Contents

Introduction

With the development of technology and society, modern problems have also been created and need to be solved immediately. How can scientists give the answer that is the most accurate in the shortest time? "*The answers all rely on systematic empirical observation, followed by the application of statistical tests. These test results – the data analysis – help answer the research question. The better the execution of the analysis, the stronger the conclusions.*" (Lewis-Beck, 1995, p.01). From the statement above, we know that with a good sample and choosing the right analytic method, every answer in the world can be solved.

Nowadays, everyone in the world can access a regular health checkup or see a doctor whenever you have disease with minimum fees by using health insurance. It is not wrong to say that health insurance has become one of the most important things in a person's life. But not everyone will pay the same prices on health insurance as the others. The main goal of the report will study how elements (like age, sex…) can influence insurance costs and develop predictive models for estimating healthcare expenses.

Before going to analysis the dataset, we need to know basic information about the dataset:

- Dataset name: Health-Insurance-Dataset.csv
- Number of records: 1338 US citizens
- Number of data fields: seven
- Description of data fields:
    - Age: The insured person's age (in years).
    - Sex: Gender (male or female) of the insured.
    - BMI (Body Mass Index): A measure of body fat based on height and weight ranging. from 15.96 -53.13.
    - Children: The number of dependents covered ranges from 0 to 5.
    - Smoker: Whether the insured is a smoker (yes or no).
    - Region: The geographic area of coverage (southeast, southwest, northeast, northwest).
    - Charges: The medical insurance costs incurred by the insured person (in $).

Summary Statistics and Visualize distribution.

Analyzing dataset requires a series of step from cleaning data, dealing with missing data, processing data… So, is there any preprocessing step before doing these steps mentioned above? The answer is yes. Getting basic information of the dataset is not only getting description of the data but also statistical summary of that dataset. The summary will give you more insight into your dataset like Min, Max, Median, Mean, SD (Standard Deviation), Q1, Q3 of all existing data fields. The result of statistical summary of dataset is shown in **Table 1**.

Table 1: Summary Statistics of all Variables

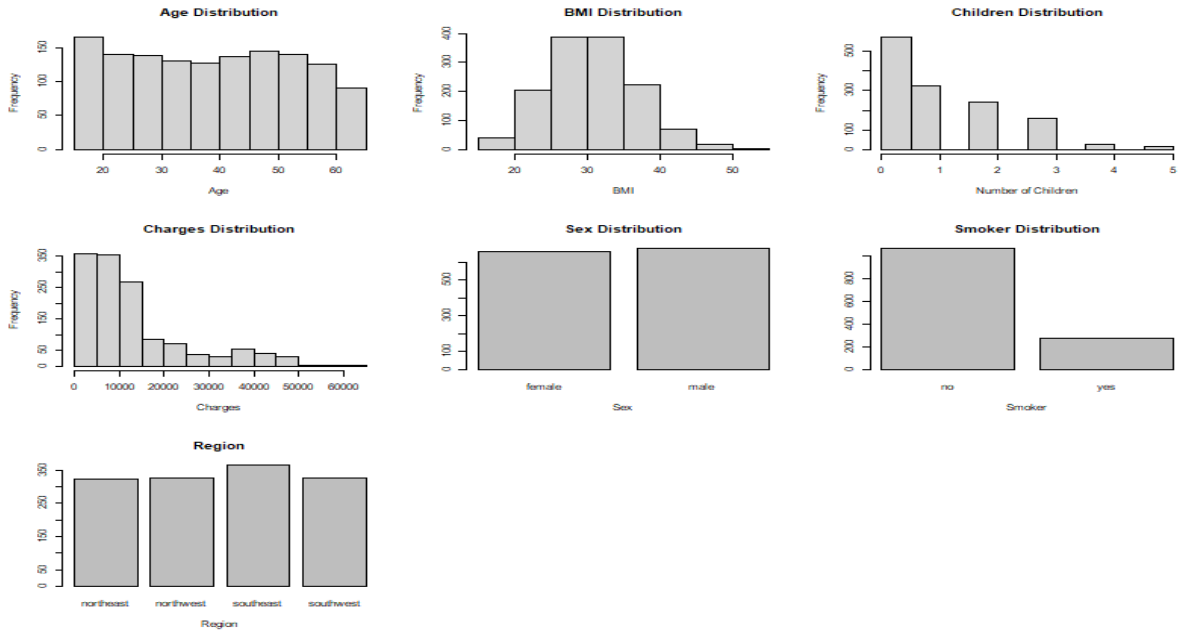
Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max	SD
age	18.00	27.00	39.00	39.21	51.00	64.00	14.04996
bmi	15.96	26.30	30.40	30.66	34.69	53.13	6.098187
children	0.000	0.000	1.000	1.095	2.000	5.000	1.205493
charges	1122	4740	9382	13270	16640	63770	12110.01
sex	Categorical variable	Categorical variable	Categorical variable	Categorical variable	Categorical variable	Categorical variable	Categorical variable
smoker	Categorical variable	Categorical variable	Categorical variable	Categorical variable	Categorical variable	Categorical variable	Categorical variable

region	Categorical variable	Categorical variable	Categorical variable	Categorical variable	Categorical variable	Categorical variable	Categorical variable

Another method to get more insight is through visualization. From the chart we can see if the record is normally distributed or not, is there any outlier existed. Visualization is a great method to see the actual viewing of the data.

According to the dataset we have four numerical variables and three categorical variables. So, the most suitable charts for the dataset are histogram (for numerical variables) and bar chart (categorical variables). **Figure 1** will illustrate the visualization of the distribution of the data.

Figure 1: Distribution of Variables



From **Figure 1,** we can easily see that numerical variables are not normally distributed. But we do not have enough evidence to say that the data is not normally distributed. In order to prove the hypothesis, we will execute the Shapiro-Wilk test to prove the normality.

The Shapiro-Wilk test.

- Null Hypothesis: The data is normally distributed.
- Significant Level: 0.05
- Result:

Table 2: The Kolmogorov-Smirnov test results.

Variables	P-Value	Conclusion
age	1.143e-07	P-Value < 0.05 ➔ Reject Null Hypothesis ➔ The data is not normally distributed
bmi	0.3218	P-Value > 0.05 ➔ Failed to reject Null Hypothesis ➔ Not have enough evidence but we can have assumption that the data is normally distributed
children	2.2e-16	P-Value < 0.05 ➔ Reject Null Hypothesis ➔ The data is not normally distributed
charges	2.2e-16	P-Value < 0.05 ➔ Reject Null Hypothesis ➔ The data is not normally distributed

Brief Conclusion:

- The dataset has four numerical variables and three categorical variables, but the numerical variables are not normally distributed.
- The age distribution has a slight positive incline.
- Both the number of children and medical charges distributions are positively inclined.
- The BMI distribution is symmetric.
- For the categorical variables, further research is needed based on their frequencies or proportions of the dataset.

Correlation coefficients

After having an overview of the dataset, the next step is doing a Correlation test and visualize the correlation to help us know more about the relationship between each variable. Understanding more about the variables will help us identify which corresponding method, model, calculation need to be used for the variables.

We have the assumption that "all predictor variables are independent." To assess this assumption, we execute a correlation test to create correlation matrix and then visualize it to a scatter plot. Before that we need to find which variables are the predictor variables. Performing correlation test requires numerical variables so which means all the categorical variables are not suitable for the test.

The Correlation test will evaluate the relationship between "Age" and "BMI", "Age" and "Children", "Age" and "Charges", "BMI" and "Children", "BMI" and "Charges", "Children" and "Charges". The correlation tests will follow two types of tests: Parametric test (Pearson) and Non-parametric test (Spearman). The test will be chosen based on the distribution of the variable, if there is a normally distributed variable, we will perform Parametric test and if there is a distribution-free variable, we will perform non-parametric test.

Parametric test (Pearson):

- Pearson's correlation coefficient: $r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}$
- Null Hypothesis: There is no linear correlation between the two predictive variables.
- Significant level: 0.05

Non-parametric test (Spearman):

- Spearman's correlation coefficient: $r_s = \frac{1}{n-1} \frac{\sum(R_i^x - u_{Rx})(R_i^Y - u_{Ry})}{\sigma_{Rx}\, \sigma_{Ry}}$
- Null Hypothesis: There is no relationship between the two predictive variables.
- Significant level: 0.05

The result for the test is shown in **Table 3 and Table 4.**

Table 3: Spearman's correlation coefficient test.

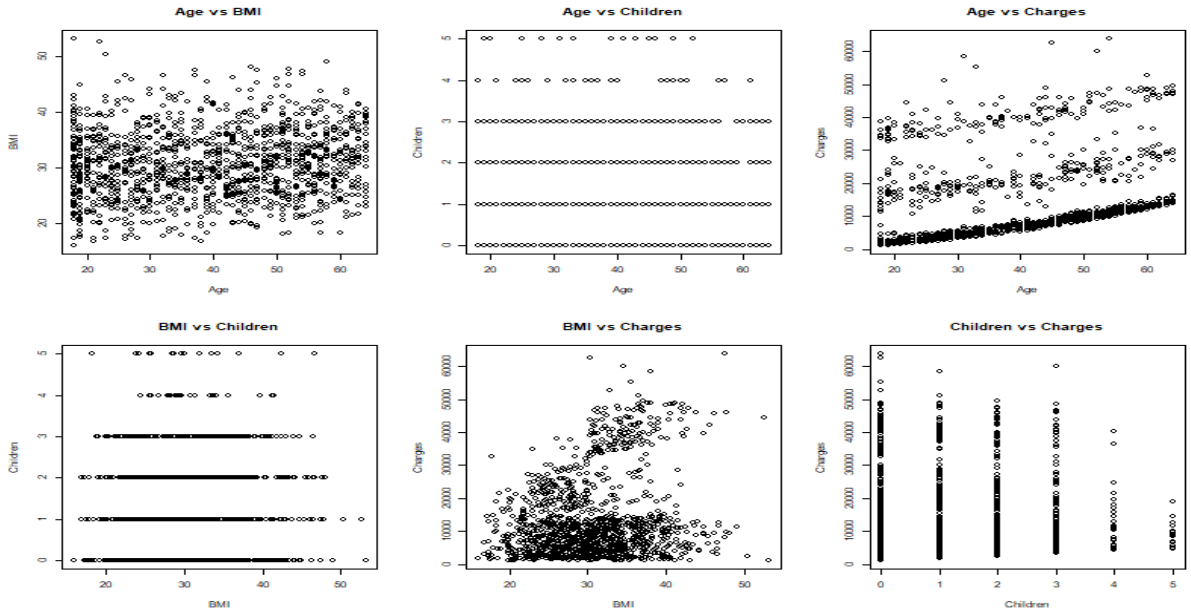
Samples	P-value	Name of the test	Conclusion
Age and BMI	6.194e-05	Pearson correlation test	p-value is smaller than any significant level ➔ Reject null hypothesis ➔ We have compelling evidence that there is a significant linear correlation between Age and BMI
Age and Children	0.03712	Spearman correlation test	p-value is smaller than 0.05 ➔ Reject null hypothesis ➔ There is compelling evidence to conclude that there is significant relationship between two variables

Age and Charges	< 2.2e-16	Spearman correlation test	p-value is smaller than any significant level ➔ Reject null hypothesis ➔ There is compelling evidence to conclude that there is significant relationship between two variables
BMI and Children	0.641	Pearson correlation test	p-value > 0.05 ➔ Failed to reject null hypothesis ➔ We do not have enough evidence to conclude that there is a significant linear correlation between the two variables.
BMI and Charges	2.459e-13	Pearson correlation test	p-value is smaller than any significant level ➔ Reject null hypothesis ➔ We have compelling evidence that there is a significant linear correlation between two variables.
Children and Charges	9.847e-07	Spearman correlation test	p-value is smaller than any significant level ➔ Reject null hypothesis ➔ There is compelling evidence to conclude that there is significant relationship between two variables

Table 4: Correlation coefficients matrix

	Age	BMI	Children	Charges
Age	1	0.1092719	0.05699222	0.5343921
BMI	0.1092719	1	0.0127589	0.198341
Children	0.05699222	0.0127589	1	0.1333389
Charges	0.5343921	0.198341	0.1333389	1

Figure 2: Scatter plot for variables.



From **Table 3**, **Table 4,** and **Figure 2**, we reach to conclusions:

- Age and BMI: The p-value from the test shows that the variables have relationship with each other but the with the correlation coefficient is only 0.1092719, we can say that this is a very weak positive correlation between Age and BMI.
- Age and Children: The p-value from the test shows that the variables have relationship with each other but the with the correlation coefficient is only 0.05699222, we can say that this is a very weak positive correlation between Age and Children.
- Age and Charges: The p-value from the test shows that the variables have relationship with each other but the with the correlation coefficient is 0.5343921, we can say that this is a very strong positive correlation between Age and Charges ➔ The amount of fee for health insurance will likely higher when you are older.
- BMI and Children: The p-value from the test shows that the variables have no relationship with each other but the with the correlation coefficient is only 0.0127589, we can say that this is a no correlation between BMI and Children ➔ Having children will not affect your BMI.
- BMI and Charges: The p-value from the test shows that the variables have a relationship with each other but the with the correlation coefficient is 0.198341, we can say that this is a weak positive correlation between Charges and BMI.
- Children and Charges: The p-value from the test shows that the variables have relationship with each other but the with the correlation coefficient is only 0.1333389, we can say that this is a very weak positive correlation between Age and BMI.

Regression Analysis

The concept of linear regression is a method used when you want to describe the linear association between two continuous variables, and to be able to predict the value of one variable give the value of the other.

Linear regression formular: $y = bx + a$ where a is the intercept and b is the gradient of the slope of the line.

Getting back to the dataset, from the previous sections, we know information about the data stored in the dataset like relationship between numerical variables, which variable is dependent/independent and statistic of the dataset. Finally, we can begin analyzing the elements that will affect the charges of health insurance.

The linear regression model will cover all variables except the target variable “Charges”. The null hypothesis for the model is “The X variable has no influence on charge.” With the significance level of 0.05.

Table 5 below will show the result of the model.

Table 5: Summary of results of the Linear Regression Model.

Coefficients	Estimate	Std.Error	t-value	Pr(>|t|)	Conclusion
Age	261.8	186.9	1.401	0.178	p-value > 0.05 ➔ Reject null hypothesis ➔ Age has statistical relationship with Charge
Region	Northwest: 9934 Southeast: 8232 Southwest: 8018	Northwest: 8914 Southeast: 7902 Southwest: 8457	Northwest: 1.114 Southeast: 1.042 Southwest: 0.948	Northwest: 0.282 Southeast: 0.313 Southwest: 0.357	p-value > 0.05 ➔ Reject null hypothesis ➔ Customer from the Region (Northwest, Southeast, Southwest) has statistical relationship with Charge
BMI	500.9	486.0	1.031	0.316	p-value > 0.05 ➔ Reject null hypothesis ➔ BMI has statistical relationship with Charge

Children	-4193	2695	-1.556	0.1372	p-value > 0.05 ➔ Reject null hypothesis ➔ Children has statistical relationship with Charge
Smoker	Yes: 22292	Yes: 4562	Yes: 4.887	Yes: 0.0001	p-value < 0.05 ➔ Failed to reject null hypothesis ➔ Smoker has no statistical relationship with Charge
Sex	Male: -3159	Male: 5632	Male: -0.561	Male: 0.58178	p-value > 0.05 ➔ Reject null hypothesis ➔ Sex has statistical relationship with Charge

Overview Evaluation from **Table 5:**

- It seems that for this dataset, whether you are a smoker or not you will not have any increase in Charge.
- Sex is the most unexpected variable that has influence on Charge. From this finding, we can also see that for this dataset, the reason male can have influence on Charge is because there are other variables (such as BMI, Age, Children…) have enough associated with Charge to the point Sex can have influence with Charge.
- For the remaining variables like Age, Region, BMI, Children, we can easily see that these variables will have heavy influence on Charge.
- Conclusion: Although linear regression model can prove to us that there are expected variables (BMI, age, region, children) that influence on Charge, but further testing is required because unexpected element has occurred (smoker has no influence on charge, sex has influence on charge).

Statistical tests

Coming to statistical testing, we want to assess differences in central tendencies of the dataset. First, we need to split the dataset into two groups according to their gender (Male and Female). The normality test needs to be executed again because of the change in the distribution in each group to find out the most suitable statistical tests for these two groups. The visualization of these two groups is drawn using qqplot and probabilistic test for normality.

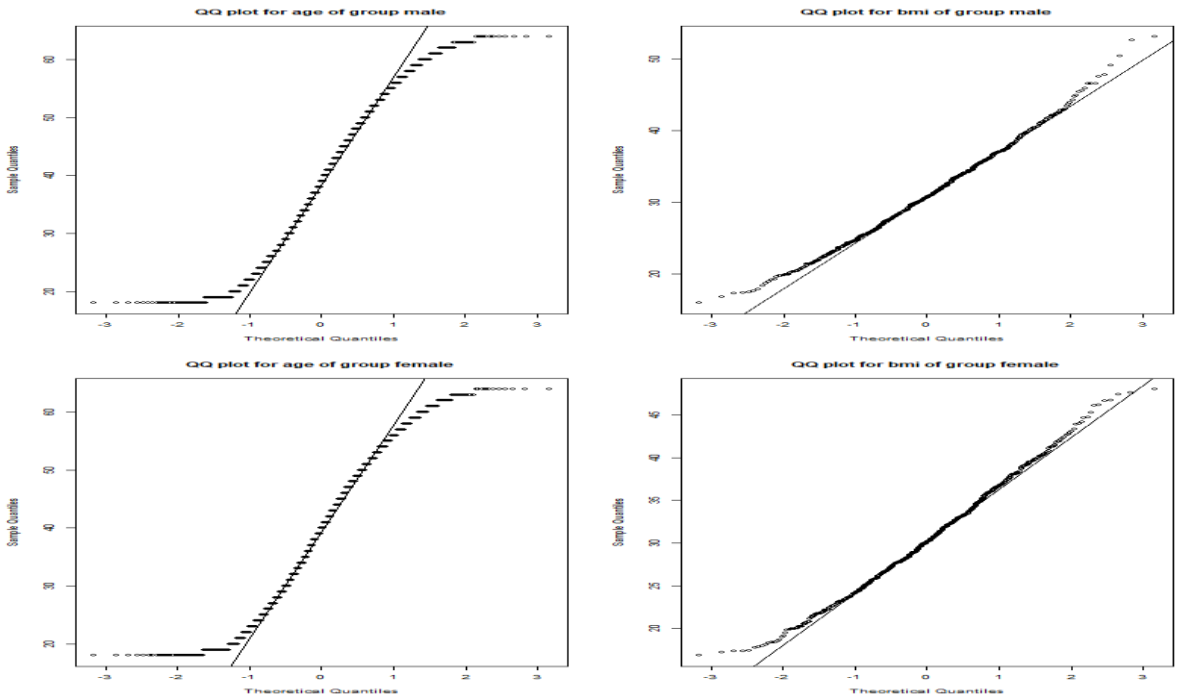
The normality test Kolmogorov-Smirnov test:

- Null Hypothesis: The group is from the dataset that is normally distributed with predictive variables.
- Alternative Hypothesis: The group is not from the dataset that is normally distributed with predictive variables.
- The significant level is 0.05.
- The result of normality test is shown in **Table 6 and Figure 3**

Table 6: Normality test results

Variable	Group	P-value	Type of statistical test
Age	Group male. Group female	0.0003315 0.0007117	Non-parametric test
BMI	Group male. Group female	0.7245 0.4237	Parametric test

Figure 3: QQ plots for independent numerical variables



From **Table 6** and **Figure 3,** for the male and female group, the null hypothesis was rejected for Age variables in both sample since the respective p-values less than significant level (0.05) and we can conclude that the records from the samples are from a distribution-free dataset. On the other hand, the null hypothesis was not rejected for BMI variables in both sample since the respective p-values more than significant level (0.05) but because of not have enough evidence so we can only have an assumption that the variable is from a normally distributed dataset.

To evaluating the differences between two groups, for numerical variables we will execute the test according to the following information below:

The parametric test:

- The null hypothesis: The distributions of the two groups are equal.
- The alternative hypothesis: There is a difference between the distributions of the two groups.
- The significant level is 0.05.
- The test statistic formular: $t = \frac{\overline{X_1} - \overline{X_2}}{\delta_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where $\delta_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$

The non-parametric test:

- The null hypothesis: The means of the two groups are equal.
- The alternative hypothesis: There is a difference between the means of the two groups.
- The significant level is 0.05.
- The test statistic formular: $U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum_{i=n_1+1}^{n_2} R_i$

The Chi-Square test:

- The null hypothesis: The variable does not affect the frequency of the group.
- The alternative hypothesis: The variable affects the frequency of the group.
- The significant level is 0.05.
- The test statistic formular: $X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

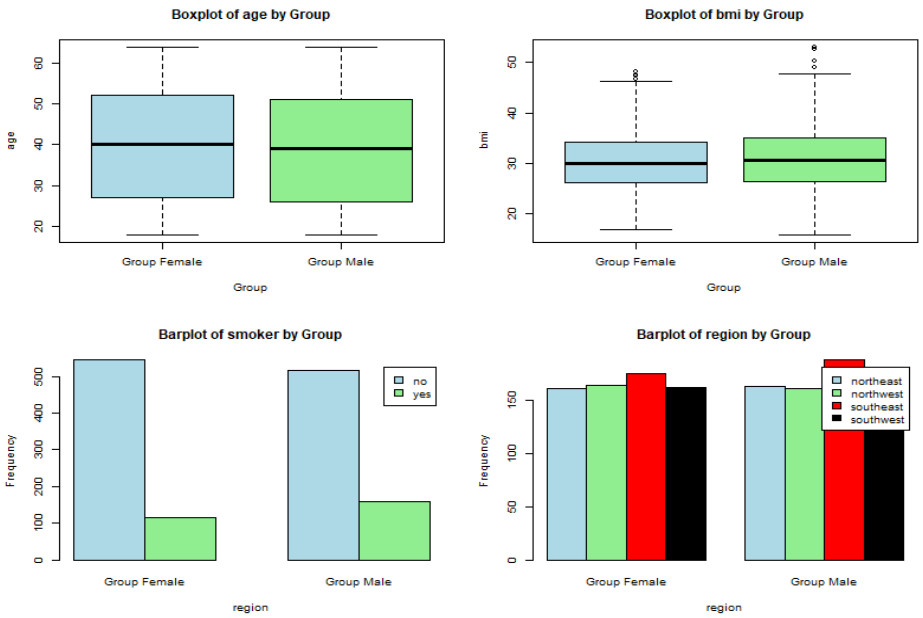
The results of the tests are shown in **Table 7**:

Table 7: Statistical test's Results

Variable	Name of the test	p-value	Assumption
Age	Mann-Whitney U test	0.4468	We have p-value > 0.05 ➔ Failed to reject null hypothesis. We can have an assumption that we do not have evidence to conclude that there is a statistically significant difference between the distributions of the two groups.
BMI	Two independent samples t-test	0.08992	We have p-value > 0.05 ➔ Failed to reject null hypothesis. We can have an assumption that we do not have evidence to conclude that there is a statistically significant difference in BMI between the means of two groups based on the provided data.
Smoker	Chi-square test	0.006548	We have p-value < 0.05 ➔ Reject null hypothesis. We can have an assumption that being a smoker has affected to the frequency of the provided data
Region	Chi-square test	0.9329	We have p-value > 0.05 ➔ Failed to reject null hypothesis. We can have an assumption that we do not have evidence to conclude that region has affected to the frequency of the provided data

The visualizations of these variables are shown in **Figure 4**:

Figure 4: Visualizations of predictive variables in two groups



ANOVA, Kruskal-Wallis's testing and post-hoc tests.

To find differences in central tendencies of the internal predictor variables (Age, BMI) with respect to the geography. First, we need to identify which evaluate we will use. From previous sections we have proved that Age is from distribution-free sample and BMI is from normally distributed sample so for Age we will use Kruskal-Wallis's testing and for BMI we will use ANOVA. **Figure 4** shows us that in "Region" variables there are four different areas which are "northeast", "northwest", "southeast", "southwest" so the tests will also surround these following areas. After the testing, we will perform additional post-hoc tests to have more evidence to support the final conclusion.

Parametric Test – ANOVA

- Null hypothesis: There is no significant difference in the mean BMI across different areas.
- Alternative hypothesis: There is a significant difference in the mean BMI across different areas.
- The significant level is 0.05.

Post-hoc test – Tukey test

- Null hypothesis: There is no significant difference in the mean BMI between pair of areas.
- Alternative hypothesis: There is a significant difference in the mean BMI between at least one pair of areas.
- The significant level is 0.05.

The result for ANOVA and Tukey testing is shown in **Table 8**.

Table 8: ANOVA and Tukey testing results.

Variable	Name of the test	P-Value	Assumption
BMI	ANOVA	<2e-16	p-value is smaller than any significant level ➔ Reject null hypothesis ➔We have convincing evidence that there is a significant difference in the mean of BMI across different areas.
BMI	Tukey	northwest-northeast: 0.9999328 southeast-northeast: 0 southwest-northeast: 0.0106965 southeast-northwest: 0 southwest-northwest: 0.0127393 southwest southeast: 0	The p-value for pair of areas southeast-northeast, southwest-northeast, southeast-northwest, southwest-northwest, southwest-southeast are smaller than any significant level ➔ Reject null hypothesis ➔ We have convincing evidence that there is a significant difference in the mean BMI between pair of areas mentioned above. In the other hand, the p-value for pair of areas northwest-northeast > 0.05 ➔ Failed to reject null hypothesis ➔ We do not have enough evidence to conclude that there is a significant difference in the mean BMI between at least one pair of area northwest-northeast

Non-parametric Test - Kruskal-Wallis

- Null hypothesis: There is no significant difference in the central tendencies of age across different areas.
- Alternative hypothesis: There is a significant difference in the central tendencies of age across different areas.
- The significant level is 0.05.

Post-hoc test – Dunn test

- Null hypothesis: There is no significant difference in the central tendencies Age between pair of areas.
- Alternative hypothesis: There is a significant difference in the central tendencies Age between at least one pair of areas.
- The significant level is 0.05.

Table 9: Kruskal-Wallis and Dunn testing results.

Variable	Name of the test	P-Value	Assumption

Age	Kruskal-Wallis	0.9374	p-value is > 0.05 ➔Failed to reject null hypothesis ➔We do not have enough evidence that there is a significant difference in the central tendencies of age across different areas.
BMI	Tukey	northwest-northeast: 0.4746126 southeast-northeast: 0.3731455 southwest-northeast: 0.3810663 southeast-northwest: 0.3485251 southwest-northwest: 0.4054807 southwest southeast: 0.2626596	the p-value for all pair of areas > 0.05 ➔ Failed to reject null hypothesis ➔ We do not have enough evidence to conclude that there is a significant difference in the central tendencies between at least one pair of area.

Conclusion from **Table 8** and **Table 9**

- BMI and region: There is a significant difference in the mean of BMI with areas that listed in the Region except the pair area northwest and northeast is the only pair that we do not have enough evidence to say that there is a significant difference in this pair of area.
- Age and region: Further research needs to be done because we do not have enough evidence to conclude that there is a significant difference in the central tendencies between all areas or pair of areas.

sConclusion

From all the tests above, we can draw conclusion that the charge of insurance is affected by customer's age, BMI, number of children, place to live. In unusual circumstances, whether you are a smoker or not can also affect the charge of insurance.

References

Lewis-Beck, M., 1995. *Data analysis: An introduction* (No. 103). Sage.

Additional Reading:

R-bloggers, 2021. How to Perform Tukey HSD Test in R. Available at: https://www.r-bloggers.com/2021/08/how-to-perform-tukey-hsd-test-in-r/ [Accessed: 05/03/2024].

Statology, Year. Dunnett's Test in R. Available at: https://www.statology.org/dunns-test-in-r/ [Accessed: 05/03/2024].