**Task 2:** Data Exploration

Krescens Kok

Department of Technology, Western Governors University

**D599:** Data Preparation and Exploration

Keiona Middleton

2025 September 20

**Part I: Univariate and Bivariate Statistical Analysis and Visualization**

**A.  Using the provided dataset, do the following using R or Python:**

**1.  Select four variables (e.g., two quantitative/numeric variables and two qualitative/categorical variables) and provide univariate visualizations for each variable selected.**

**A comparison of a graph

AI-generated content may be incorrect.Variables:** BMI (quantitative), Height (quantitative), Sex (qualitative), Region (qualitative)

**A comparison of different colored bars

AI-generated content may be incorrect.**

**2.  Provide two bivariate visualizations for each variable selected from part A1.**

**A bar chart of different colored squares

AI-generated content may be incorrect.A chart with a box plot

AI-generated content may be incorrect.**

**A diagram of a box plot

AI-generated content may be incorrect.A bar chart with different colored bars

AI-generated content may be incorrect.**

**B.  Complete the following using the attached “Health Insurance Dataset” and R or Python:**

**1.  Provide the descriptive statistics (e.g., mean, median, range, standard deviation, variance, percentiles, quartiles) for all quantitative (i.e., numeric) variables selected in the dataset.**

**Mean, Median (50th percentile), Range (Min/Max), Standard Deviation, Percentiles/Quartiles:**

**A screenshot of a black screen

AI-generated content may be incorrect.**

**Variance:**

**A black background with white text

AI-generated content may be incorrect.**

**2.  Provide the descriptive statistics (e.g., frequency counts and percentages) for all qualitative (i.e., categorical) variables in the dataset.**

**A screenshot of a computer program

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**Part II: Parametric Statistical Testing**

**C.  Describe a real-world organizational situation or issue in the attached “Health Insurance Dataset” by doing the following:**

**1.  Create one research question that is relevant to the dataset and any organizational needs that can be answered through data analysis and is appropriate for parametric testing.**

Does a region affect the BMI of an individual?

Since the business is trying to identify risk in the health insurance industry, if the region impacts the BMI of an individual, this could also affect insurance premiums. Since BMI determines whether or not an individual is underweight, healthy, overweight, or obese, this can impact the insurance cost or plans.

**D.  Analyze the dataset by doing the following:**

**1.  Identify a parametric statistical test that is relevant to your research question from part C1.**

ANOVA would be best suited for this research question as it is a test that is used to determine the difference between 3 or more unrelated groups (*What Is ANOVA (Analysis of Variance) Testing? - Qualtrics*, 2024).

**2.  List the dataset variables relevant to answering your research question from part C1.**

The variables that will be used to answer the research question are BMI and Region.

**3.  Justify why you chose the statistical test identified in part D1 based on the variables.**

I chose ANOVA because the independent variable, region, is categorical with more than 2 groups: Southeast, Southwest, Northwest, and Northeast, while BMI, the dependent variable, is continuous and the BMI histogram is bell shaped. According to the course textbook, a t-test would not work due to the fact that there are more than 2 groups in the region variable.

**4.  Develop null and alternative hypotheses related to your chosen parametric test from part D1.**

**Null Hypothesis:** There is no significant relationship between the means of the BMI across all regions are equal

**Alternate Hypothesis:** There is a significant relationship between the means of the BMI across all regions are not equal

**5.  Write error-free code in either Python or R to run the parametric test and provide the output and the results of all calculations from the parametric statistical test you perform.**

**A screenshot of a computer program

AI-generated content may be incorrect.**In order to perform the ANOVA test, the package, scipy.stats needs to be installed and import the f\_oneway function. (GeeksforGeeks, 2025d)

**E.  Evaluate parametric test results by doing the following:**

**1.  Discuss the test results, including the decision to reject or fail to reject the null hypothesis from part D4.**

The null hypothesis states that there is no significant evidence to suggest that the means across all regions are equal. However, looking at the test results of the ANOVA test, the p-value is less than .05, the significance level, which means that we reject the null hypothesis and conclude that there is significant evidence to suggest that the means across all regions are not equal.

**2.  Create an answer to your research question from part C1 based on the decision to reject or fail to reject the null hypothesis.**

Does a region affect the BMI of an individual? Yes, based on the ANOVA test that was performed on the BMI and region, we can conclude that a region does in fact affect the BMI of an individual. We rejected the null hypothesis and stated that there is significant evidence to suggest that the means across all regions are not equal. Therefore, geographic region is a meaningful factor in explaining the variation in BMI among individuals.

**3.  Explain how stakeholders in the organization benefit from your choice of testing method.**

The stakeholders benefit from the testing method of using ANOVA because there are more than 2 groups in the region variable, making it a great test to use to identify whether or not there was a difference in BMI across the regions.

From the results of the ANOVA test, the stakeholders can make informed decisions on insurance premiums because there is significant evidence that the BMI is different across the regions. Furthermore, the company could justify the risk assessment if the BMI is too high or too low, which could cause the individual to be at risk, therefore, increasing the cost of the insurance.

**F.  Summarize the implications of your parametric statistical testing by doing the following:**

**1.  Recommend a course of action based on your findings.**

The company should create policies based on the regions and make more meaningful decisions based on each region. With evidence that the BMI differs across regions, the company could have different prices for insurance premiums based on whether the BMI is higher or lower than the other regions. Another recommendation that the company could make is to identify health programs for each region based on the average BMI. Lastly, the next step of analysis could be to identify more specific details as to why a region may have a higher BMI than another region.

**2.  Discuss the limitations of your data analysis.**

A limitation to the data analysis is that ANOVA tests are designed to use data from a normal distribution. Therefore, it may not produce exact p-values when the data is skewed (Discovery, n.d.)**.**

Another limitation to using the ANOVA test is that the test can identify whether or not there is a difference in the mean for at least one group, but it cannot identify which group(s) are different. Therefore, it may require more analysis to figure out which group(s) are different than the rest.

**Part III: Nonparametric Statistical Testing**

**G.  Describe a real-world organizational situation or issue in the provided dataset by doing the following:**

**1.  Create one research question that is relevant to the dataset and any organizational needs that can be answered through data analysis and is appropriate for nonparametric testing.**

Do the charges of individual health insurance claims for medical expenses change based on gender?

This is important to stakeholders as it will help them assess whether they have fair charges based on gender.

**H.  Analyze the dataset further by doing the following:**

**1.  Identify a nonparametric statistical test that is relevant to your question from part G1.**

Mann-Whitney U test

**2.  List the dataset variables relevant to answering your research question from part G1.**

Charges and Sex will be used to answer the research question.

**3.  Justify why you chose the statistical test identified in part H1 based on variables.**

Looking at the distribution of Charges, the data is skewed right, therefore, it does

**A screenshot of a graph

AI-generated content may be incorrect.**not have a normal distribution. Non-parametric testing is typically used for distributions that are not normal. The Mann-Whitney U test is useful for assessing the difference between two independent groups (male vs female), where the data (charges) is continuous (McClenaghan & McClenaghan, 2024).

**4.  Develop null and alternative hypotheses related to your chosen nonparametric test from part H1.**

**Null Hypothesis:** The distribution between male vs female is the same

**Alternate Hypothesis:** The distribution between male vs female is not the same

**5.  Write error-free code in either Python or R to run the nonparametric test and provide a screenshot of the output and the results of all calculations from the nonparametric statistical test you performed.**

**A screen shot of a computer

AI-generated content may be incorrect.**In order to perform the Mann-Whitney u test, the package, scipy.stats needs to be installed and import the mannwhitneyu function (*Mannwhitneyu — SciPy v1.16.2 Manual*, n.d.).

**I.  Evaluate nonparametric test results by doing the following:**

**1.  Discuss the test results, including the decision to reject or fail to reject the null hypothesis from part H4.**

Since the p-value is greater than .05, the significance level, I do not reject the null hypothesis and conclude that there is not enough statistical evidence to suggest a difference in charges between males and females.

**2.  Create an answer to your research question from part G1 based on the decision to reject or fail to reject the null hypothesis.**

Do the charges of individual health insurance claims for medical expenses change based on gender? No, based on the Mann-Whitney u test that was performed on Sex and Charges, we can conclude that there is not enough statistical evidence to suggest a difference in charges between males and females; therefore, we can say that the charges do not change based on gender.

**3.  Explain how stakeholders in the organization benefit from your choice of testing method.**

Stakeholders in the organization benefit from the testing method because the Mann–Whitney U test was an appropriate choice, given that the data was not normally distributed. The results showed no evidence that charges differ based on gender. This provides valuable insights to stakeholders, as it indicates that gender is not a risk factor contributing to how much a customer pays for medical expenses. This helps the stakeholders focus on other variables that may better explain differences in charges, leading to more accurate risk assessments and fairer pricing strategies.

**J.  Summarize the implications of your nonparametric statistical testing by doing the following:**

**1.  Recommend a course of action based on your findings.**

A recommended course of action would be not to change the prices based on gender, as it would create unfair pricing strategies. In addition, the next best course of action could be to focus on other features in the dataset that could actually contribute to the charges in order for the stakeholders to understand the differences in pricing for their customers.

**2.  Discuss the limitations of your data analysis.**

**A screenshot of a graph

AI-generated content may be incorrect.** A limitation of this analysis is that even though the Mann-Whitney u test is good for skewed data, the results can still be influenced by outliers in the data. Looking at the boxplot of charges, it appears that there are many outliers in the data. Another limitation of the Mann-Whitney u test is that we cannot conclude that the distributions are the same, there just isn’t enough statistical evidence to say that they are different. Lastly, since this statistical test only compares charges by gender, more analysis must be done to understand the differences in charges by looking at the other variable

**References**

Discovery, J. S. (n.d.). *One-Way ANOVA*. https://www.jmp.com/en/statistics-knowledge-portal/one-way-anova

GeeksforGeeks. (2025d, July 23). *How to perform a OneWay ANOVA in Python*. GeeksforGeeks. <https://www.geeksforgeeks.org/python/how-to-perform-a-one-way-anova-in-python/>

*mannwhitneyu — SciPy v1.16.2 Manual*. (n.d.). https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

McClenaghan, E., & McClenaghan, E. (2024, March 25). *Mann-Whitney U Test: Assumptions and example*. Informatics From Technology Networks. https://www.technologynetworks.com/informatics/articles/mann-whitney-u-test-assumptions-and-example-363425

*What is ANOVA (Analysis Of Variance) Testing? - Qualtrics*. (2024, September 26). Qualtrics. https://www.qualtrics.com/experience-management/research/anova/