Gabriela Masak

D599 – Data Preparation and Exploration

Task 2: Data Exploration

9/21/2024

Data Exploration Report

**Part 1: Univariate and Bivariate Statistical Analysis and Visualization**

Univariate Statistics

Statistical Analysis of Continuous Variable #1: BMI

The distribution of data in the BMI column is normal, represented by the bell curve in Figure 2 with its peak, the median, only very slightly to the left of the mean, demarked by the red dashed line. The distribution is skewed slightly to the right, with mostly mesokurtic tailedness.

A screenshot of a computer code

Description automatically generated

Figure 1: Screenshot of Statistical Measures of BMI

A diagram of a normal distribution

Description automatically generated

Figure 2: Distribution of Values of BMI

Statistical Analysis of Continuous Variable #2: Charges

The distribution of data in the BMI column is extremely skewed to the right, represented by the bell curve in Figure 4 with its peak, the median, to the extreme left of the mean, demarked by the red dashed line. The distribution also shows a leptokurtic distribution. While there are a greater quantity of lower values, the higher values in the dataset are much higher than the average.

A white background with black text

Description automatically generated

Figure 3: Screenshot of Statistical Measures of Charges

A graph of a graph

Description automatically generated with medium confidence

Figure 4: Distribution of Charges Values

Statistical Analysis of Categorical Variable #1: Sex

The gender data is close to perfectly even distribution, with both gender types approximately 0.5% from representing their respect half of the data.

A computer code with numbers and symbols

Description automatically generated

Figure 5: Screenshot of Statistical Measures of Sex

A blue and pink squares

Description automatically generated

Figure 6: Distribution of Sex Data

Statistical Analysis of Categorical Variable #2: Region

The region data is relatively close to even distribution, with the four region types approximately 1-2% from representing their respect quarter of the data. Southwest is the region type that is most represented, representing 27.2% of the data population; Northeast is the least, with 24.2%.

A close-up of a computer screen

Description automatically generated

Figure 7: Screenshot of Statistical Measures of Regions

A graph of different colored rectangular shapes

Description automatically generated

Figure 8: Distribution of Region Data

Bivariate Statistics

Statistical Analysis of Continuous Variable #1: BMI vs. Age

A graph of age versus bmi

Description automatically generated

Figure 9: Scatterplot of Age vs. BMI

As seen in Figure 9, there is a slightly positive, weak relationship between age and BMI, indicated by the thick and wide band on the scatterplot.

A screenshot of a computer code

Description automatically generated

Figure 10: Screenshot of Statistical Analysis Values in R

Statistical Analysis of Continuous Variable #2: Charges vs. Score

A graph of numbers and a number of blue dots

Description automatically generated

Figure 11: Scatterplot of Charges vs. Score

As seen in Figure 11, the relationship between amount charged and score is hard to accurately define, as the data points cluster approximately between scores of 40 and 100 and amounts charged between 1200 and 14000. Data is more sparse outside of this cluster without any real pattern present visually other than a slight positive trend.

A screenshot of a computer code

Description automatically generated

Figure 12: Screenshot of Statistical Analysis Values in R

Statistical Analysis of Categorical Variable #3: Smoker vs. Sex

A collage of a mosaic plot

Description automatically generated

Figure 13: Mosaic Plot of Smoker and Sex Variables

As seen in Figure 14, between the two genders, men are more likely to smoke, although the difference is rather mild, represented by the chi-squared value in Figure 16.

A graph of a person and person

Description automatically generated

Figure 14: Box Plots of Smoker and Gender Data

A close up of text

Description automatically generated

Figure 15: Screenshot of Statistical Analysis Values in R

A screenshot of a computer code

Description automatically generated

Figure 16: Second Screenshot of Statistical Values

Statistical Analysis of Categorical Variable #4: Region vs. Level

**A mosaic plot with text

Description automatically generated**

Figure 17: Mosaic Plot of Region vs. Level

As seen in Figures 17 and 18, the regions follow similar patterns in that the levels for each region in order from most likely to least likely are C, D, B, E, A, with the exception of the southwest, where there are more B levels than C levels. Otherwise, the variations in the distributions between regions are only slight.

**A graph of different colored bars

Description automatically generated**

**A screenshot of a computer code

Description automatically generated**

**Part 2: Parametric Statistical Analysis**

Health is usually an afterthought until an issue arises. While many individuals understand how certain factors affect their health, they may not be aware of how much of a correlation exists between certain lifestyles, choices, and health outcomes. These correlations are of great interest to health insurance companies, as they look for meaningful proactive health programs to encourage and develop healthy habits in their clients and ultimately save costs in the long run. One area of research they may focus on is the correlation of BMI and age. If BMI increases as one ages, there are benefits to targeting younger clients to teach healthier eating and exercise habits, who will likely maintain those habits in later years. From the dataset provided, “BMI” and “age” variables will undergo parametric statistical testing to test the correlation prediction, ultimately leading future business actions and research, acknowledged by the hypothesis below.

HO: No correlation exists between age and BMI (p = 0).

HA: A correlation exists between age and BMI (p ≠ 0).

Using Python, it can be calculated to determine whether the Pearson correlation coefficient is statistically significant between two numeric variables, justified by the use of age and BMI for this analysis, using the “pearsonr” function in Python. While correlation does not mean causation, it is an important feature that can be used for predictive practices or to guide business or research decisions. The function also displays the two-tailed p-value, as displayed below Figure 18.

A screenshot of a computer program

Description automatically generated

Figure 18: Screenshot of Python Code

The output of the Python code in Figure 18 is:

Pearson correlation coefficient: 0.109272

p-value: 0.000062

  As a result of this output, it can be determined that the null hypothesis is rejected with 99% confidence and there is a correlation. With a confidence level of 99% (alpha is 0.01), 1336 degrees of freedom, the critical value is 0.07045, less than the calculated value of approximately 0.11. Stakeholders at the insurance company benefit from this analysis because this information will encourage them to explore age-based health programs to lower BMI preemptively.

Despite ultimately determining that the mull hypothesis should be rejected several limitations of this type of analysis should be addressed. Firstly, correlation does not mean causation. Also, establishing that an association exists between age and BMI does not mean that associations cannot exist between age, BMI, and other variables. A recommended course of action is to continue to analyze the various relationships, or lack thereof, between different variables to define a more comprehensive list of factors.

**Part 3: Nonparametric Statistical Analysis**

Healthcare can be expensive. For some families, a diagnosis could mean both physical and financial ruin. If they have health insurance, these bills could be offset for a fraction of the original price, leaving the insurance companies on the hook for bills sometimes in the hundreds of thousands. Insurance companies look for meaningful insights into health and predictive analysis that could mitigate future expensive bills. A research question could be whether smoking is aligned with higher health care costs. Using statistical analysis, the variables “smoker” and “charges” will offer insight.

Nonparametric statistical analysis is reserved for datasets that lack normal distribution. The “charges” variable is extremely skewed, as observed in Figure 4. Setting up a hypothesis testing allows for statistical analysis.

HO: The median value of charges is equal between smokers and non-smokers.

HA: The median value of charges is not equal between smokers and non-smokers.

Due to the binary categorical nature of the smoker variable and the numerical variable of the charges variable, the Mann-Whitney U test is appropriate to determine whether the variables are independent. This can be completed in Python using the “mannwhitneyu” function, as seen in Figure 19 below, after separating the charges column into respective smoker and non-smoker arrays.

A computer screen shot of a computer code

Description automatically generated

Figure 19: Python Code for Mann Whitney U

The output of the Python code in Figure 19 is:

Mann-Whitney U statistic: 7403

p-value: 5.270233444503571e-130

  As a result of this output, it can be determined that the null hypothesis is rejected and there is difference between the charges of smokers and non-smokers in the data with 95% confidence. Stakeholders at the insurance company benefit from this analysis because this information by focusing on efforts and programs that will ultimately decrease smoking within their clientele.

While the Mann-Whitney U test is ideal for variables that can be grouped and show skewed distributions of data, making it an effective statistical test for the charges and smoker data, it has its limitations. Data can only be grouped once, meaning that only two groups can be compared. Imbalanced sample sizes are also of concern, as large variance in the groupings of the data can negatively affect the test’s power. Further testing to uncover the relationships between other variables is highly recommended.

**Part 4: Panopto Video Submission**

An audiovisual discussion of these data exploration techniques and results can be found following the link to this Panopto video recording:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=ed11cc52-b333-4db6-aca7-b1fd001dffc2