Task 2:

Data Exploration

Keniyah Chestnut

Data Preparation and Exploration — D599

SID:012601305

Part I: Univariate and Bivariate Statistical Analysis and Visualization

A: Univariate Statistics

For my continuous variables, I chose Age and BMI. I plotted both distributions using histograms.

For categorical variables, I selected Sex and Smoker. These were displayed using countplots.

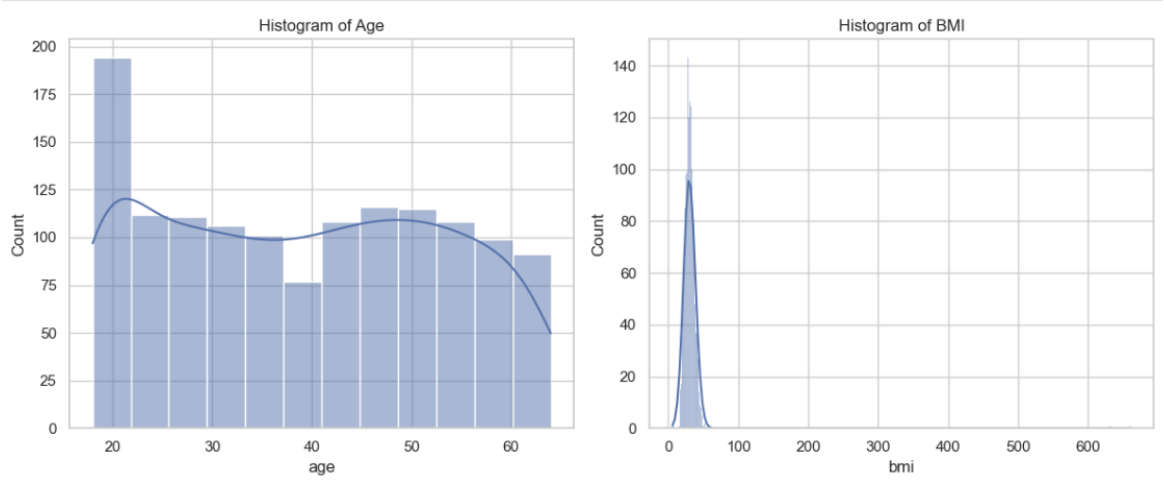
[10]:
```
# Descriptive statistics for variables
variable_columns = ['age', 'bmi', 'sex', 'smoker']
stats = df[variable_columns].describe(include='all')

# Display the results
print(stats)
```

```
                age          bmi   sex smoker
count   1338.000000  1342.000000  1342   1338
unique          NaN          NaN     5      2
top             NaN          NaN  male     no
freq            NaN          NaN   676   1064
mean      39.207025    31.562136   NaN    NaN
std       14.049960    24.530915   NaN    NaN
min       18.000000     6.098187   NaN    NaN
25%       27.000000    26.296250   NaN    NaN
50%       39.000000    30.400000   NaN    NaN
75%       51.000000    34.700000   NaN    NaN
max       64.000000   661.000000   NaN    NaN
```

A1: Visual of Findings from Part A

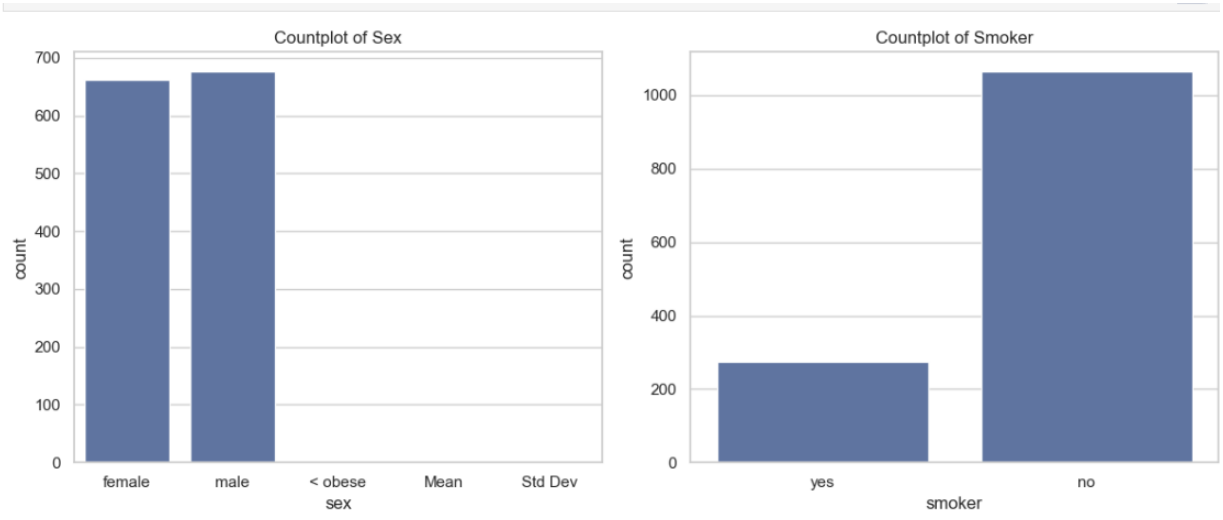
Below are the visual representations of the distributions of Age, BMI, Sex, and Smoker:



Distributions for Age and BMI:

The histogram for Age shows a right-skewed distribution, with a larger concentration of younger individuals between ages 18 and 30.

The histogram for BMI appears mostly normal but is slightly right-skewed, meaning higher BMI values in the upper range slightly pull the distribution to the right.



Distributions for Sex and Smoking Status:

The distribution of male versus female is very close to even, with similar numbers for each group.

However, the distribution of smokers shows that nonsmokers are heavily overrepresented, with significantly more nonsmokers in the dataset.

B: Bivariate Statistics

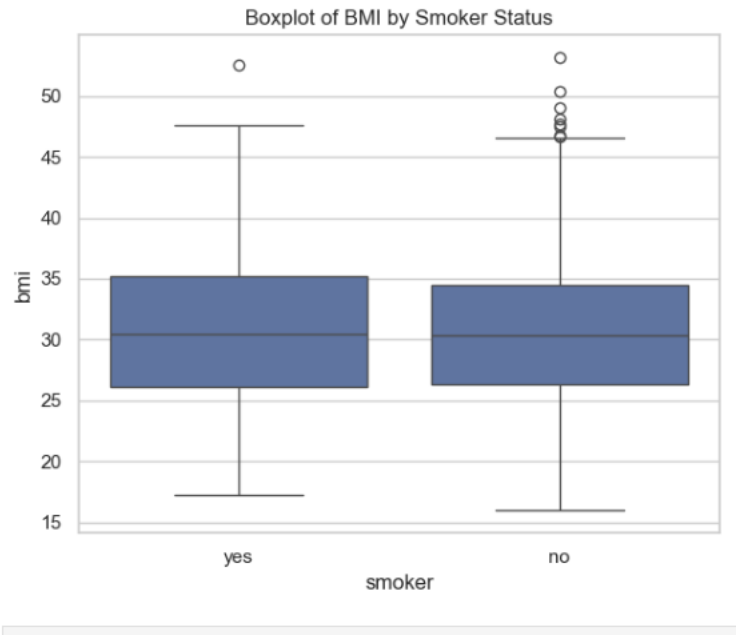
For my bivariate analysis, I chose to compare categorical variables against continuous variables to examine differences between groups.

The first plot examines Smoker (categorical) and BMI (continuous) using a boxplot.

The second plot examines Sex (categorical) and Charges (continuous) using a boxplot.

These visualizations allow us to observe how the continuous variables are distributed across the groups defined by the categorical variables.

B1: Visual of Findings from Part B

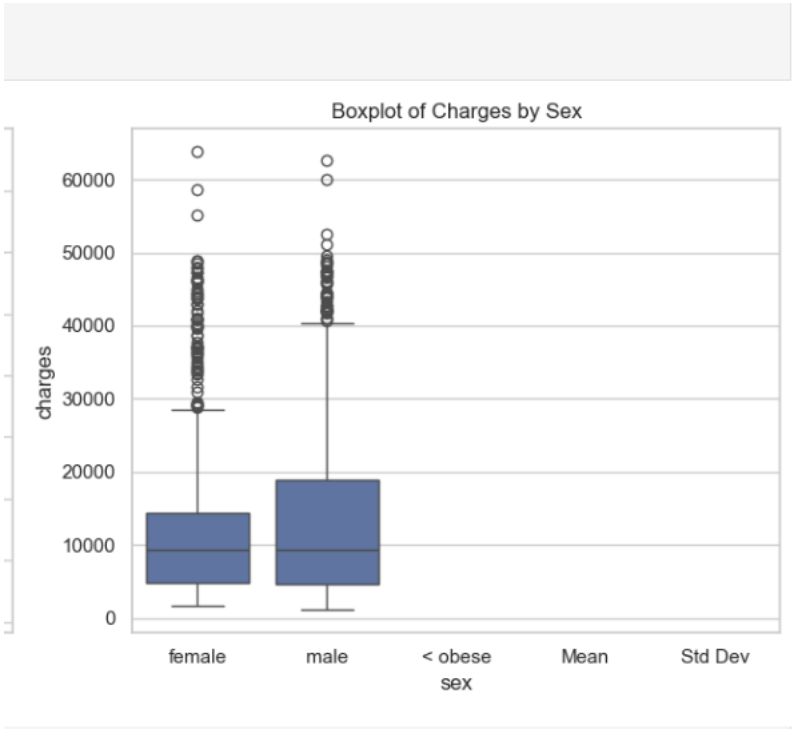


Distribution of BMI by Smoking Status:

The boxplot above shows that the BMI distributions for smokers and nonsmokers are fairly similar.

Both groups have comparable medians and interquartile ranges, with slightly more outliers among nonsmokers.

Overall, there does not appear to be a major difference in BMI based on smoking status, but a parametric statistical test will be used to confirm this observation.



Distribution of Charges by Sex:

The boxplot shows that males generally have slightly higher charges compared to females.

The median charges for males are higher, and the spread of charges (IQR) is also wider for males, suggesting greater variability in male charges.

Both groups display a large number of outliers, particularly at the higher end, indicating that some individuals incur significantly higher charges regardless of sex.

Distribution of BMI and Charges:

The scatterplot of BMI versus Charges suggests a weak to moderate positive relationship between the two variables.

While most individuals with lower BMI tend to have lower charges, the spread of charges increases with higher BMI values.

This indicates that higher BMI is generally associated with higher charges, although the relationship is not perfectly linear.

This observation will be further supported and confirmed by the Pearson correlation coefficient.

Part II: Parametric Statistical Testing

C1: Research Question

Does smoking impact BMI?

C2: Variable Identification

The variables selected are Smoker (categorical) and BMI (continuous).

D1: Parametric Test Method

I will use a T-test to compare the means of BMI for smokers and non-smokers.

D2: Develop Parametric Hypotheses

Null Hypothesis (H0): Smoking does not impact BMI.

Alternative Hypothesis (H1): Smoking impacts BMI.

D3 & D4: Parametric Test Code & Output

```
[24]: # Parametric Test - T-Test for BMI based on Smoking Status
smoker_bmi = df[df['smoker'] == 'yes']['bmi']
nonsmoker_bmi = df[df['smoker'] == 'no']['bmi']

t_stat, p_value = stats.ttest_ind(smoker_bmi, nonsmoker_bmi)

print("Parametric Test - T-Test Results")
print("T-Statistic:", t_stat)
print("P-Value:", p_value)
```

```
Parametric Test - T-Test Results
T-Statistic: 0.13708403310827058
P-Value: 0.8909850280013041
```

T-Statistic: 0.137

P-Value: 0.891

E1: Justification for Parametric Test

A t-test was selected because it is a powerful and widely used statistical method for comparing the means of two independent groups. In this case, BMI is a continuous variable, and the groups (smokers and nonsmokers) are independent and categorical. The sample size is sufficiently large to meet the assumptions of the t-test, and the objective is to determine whether there is a statistically significant difference in the average BMI between the two groups. Therefore, a t-test was appropriate to evaluate if smoking status impacts BMI.

E2: Parametric Hypothesis Support

The t-test results yielded a t-statistic of 0.137 and a p-value of 0.891. The null hypothesis stated that there is no significant difference in BMI between smokers and nonsmokers.

Because the p-value is substantially higher than the standard significance threshold of 0.05, we fail to reject the null hypothesis. This indicates that there is no statistically significant difference in BMI between the two groups.

Based on these results, smoking status does not appear to be associated with BMI, and the two variables do not show a meaningful relationship.

E3 & F3: Benefit of Parametric Testing and Recommended Course of Action

Running the parametric test gave stakeholders valuable information. Although it would have been helpful to find a connection between BMI and smoking status, the test showed that there isn't a meaningful difference in BMI between smokers and nonsmokers. This result is still useful because it helps rule out BMI as a factor in this case, allowing the organization to shift focus toward other factors that may have a bigger impact on costs.

Since the analysis did not find a connection between BMI and smoking status, there is no immediate need to make decisions or policy changes based on this relationship. However, it would be a good idea to continue analyzing other variables to see which ones might better explain rising charges. Finding those key factors will help the organization make smarter decisions about managing costs moving forward.

F1: Answer to Parametric Research Question

My research question for the parametric test was: Does smoking impact BMI?

Based on the results, we cannot make a direct connection between smoking status and BMI. The t-test showed no significant difference between the BMI of smokers and nonsmokers, so we are unable to conclude that smoking has an impact on BMI. In short, any rise in BMI is likely related to other factors outside of smoking status.

Database Attacks

F2: Limitations of Parametric Data Analysis

One limitation of the t-test is that it assumes the data is normally distributed. If the BMI data is even slightly skewed, this could affect the validity of the results. Looking at the BMI distribution, it appears mostly normal but slightly skewed, which could influence the test.

Additionally, smoking may not act alone in affecting BMI. Other factors, such as age or sex, could combine with smoking to influence BMI, but a simple t-test does not account for multiple variables at once. Therefore, while this analysis suggests smoking does not significantly impact BMI on its own, we cannot completely rule out its role without further, more complex analysis.

Part III: Nonparametric Statistical Testing

G1: Research Question

Is there a significant difference in charges for smokers and nonsmokers?

G2: Variable Identification

The variables selected are Smoker (categorical) and Charges (continuous but not normally distributed).

H1: Nonparametric Test Method

I used the Mann-Whitney U Test to compare charges between smokers and nonsmokers.

H2: Develop Nonparametric Hypotheses

Null Hypothesis (H0): There is no difference in charges between smokers and nonsmokers.

Alternative Hypothesis (H1): There is a difference.

H3 & H4: Nonparametric Test Code & Output

```
[26]: # Nonparametric Test - Mann-Whitney U Test for Charges based on Smoking Status
smoker_charges = df[df['smoker'] == 'yes']['charges']
nonsmoker_charges = df[df['smoker'] == 'no']['charges']

u_stat, mann_p_value = stats.mannwhitneyu(smoker_charges, nonsmoker_charges)

print("\nNonparametric Test - Mann-Whitney U Test Results")
print("U-Statistic:", u_stat)
print("P-Value:", mann_p_value)
```

```
Nonparametric Test - Mann-Whitney U Test Results
U-Statistic: 284133.0
P-Value: 5.270233444503571e-130
```

U-Statistic: 284133

P-Value: 5.270e-130

I1: Justification for Nonparametric Test

According to course materials, nonparametric tests are useful when data is not normally distributed and when comparing medians between two groups. Since the distribution of charges is skewed, a nonparametric test, such as the Mann-Whitney U Test, is a more reliable and accurate choice than a parametric test in this case.

Additionally, the Mann-Whitney U Test is designed for comparing groups when one variable (smoker status) is categorical and the other (charges) is continuous but not normally distributed. This makes it an appropriate method to determine if smoker status is linked to higher charges.

I2: Nonparametric Hypothesis Support

The Mann-Whitney U Test produced a p-value of 5.27e-130, which is extremely close to zero. This indicates that there is a statistically significant difference in charges between smokers and nonsmokers. Since the p-value is well below the standard significance level of 0.05, we reject the null hypothesis. This means that smoker status is significantly associated with differences in charges.

I3 & J3: Benefit of Nonparametric Testing and Recommended Course of Action

This result gives a clear answer for stakeholders. Smoking is linked to higher charges. Knowing this, the company can use smoking status as a factor when setting insurance premiums. Since smokers are more likely to have higher medical costs, it makes sense for them to pay higher premiums. At the same time, it would not be ethical to deny them coverage completely. Instead, adjusting premiums helps balance fairness and cost while making sure the company can manage expenses responsibly.

J1: Answer to Nonparametric Research Question

The research question was: Is there a significant difference in charges for smokers and nonsmokers? Based on the analysis, the answer is yes. The results show there is a statistically significant difference, with smokers having higher charges. This suggests that covering smokers does come with higher costs.

J2: Limitations of Nonparametric Data Analysis

There are a few factors that limit this analysis. First, it is important to consider how the smoking data was collected. If the information came from self-reported surveys, the data could be skewed depending on how honest people were about their smoking habits. Additionally, the data only captured smoking status as a simple yes or no, which does not reflect the amount or frequency of smoking. For example, heavy smokers and light smokers were treated the same in this analysis,

and the number of years someone has been smoking was not considered. A new smoker and a long-term smoker may have very different impacts on charges.

Finally, nonparametric tests are not as sensitive as parametric tests when it comes to detecting smaller patterns or relationships. Because of this, some subtle connections may not have been captured through this analysis.

J3: Recommended Course of Action

As shown in the results, smoking is clearly linked to higher charges. When setting insurance premiums, the company should take smoking status into account, as smokers are likely to cost the company more. To manage these higher costs fairly, smokers should be charged higher premiums. This approach allows the company to avoid shifting the burden onto nonsmokers while still providing coverage to everyone. Denying coverage to smokers would be unethical, so adjusting premiums is a reasonable and fair solution.

References

No external sources were used beyond WGU course materials.