D599 - Data Preparation and Exploration - Task 2

Part I: Univariate and Bivariate Statistical Analysis and Visualization

A. Univariate Analysis

To explore the dataset, I analyzed the distributions of two continuous variables and two categorical variables. My Continuous Variables are: **Age** and **BMI**. My categorical variables I used **Sex** and **Smoker**. I started by running a descriptive analysis and here is what i found:

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

# 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
```

Continuous Variables:

1. Age

- Mean: 39.2 years
- Median: 38.0 years
- Standard Deviation: 14.0
- Min: 18
- Max: 64

The age distribution is slightly right-skewed, with most individuals between 25 and 55 years old.

2. BMI (Body Mass Index)

- Mean: 30.7
- Median: 30.4
- Standard Deviation: 6.1
- Min: 15.9
- Max: 661 (Major Outlier present)

BMI is approximately normally distributed, though there is a slight left skew due to individuals with higher BMI values. Most values fall between 25 and 35, suggesting many individuals are overweight or obese.

Categorical Variables:

1. Sex

- **Male**: 676 (50.5%)
- **Female**: 662 (49.5%)

The dataset is nearly balanced between male and female participants.

2. Smoker

- **Smoker**: 274 (20.5%)
- **Non-Smoker**: 1064 (79.5%)

There are significantly more non-smokers than smokers in the dataset, indicating a skew toward non-smoking individuals.

A1. VISUAL OF FINDINGS FROM PART A

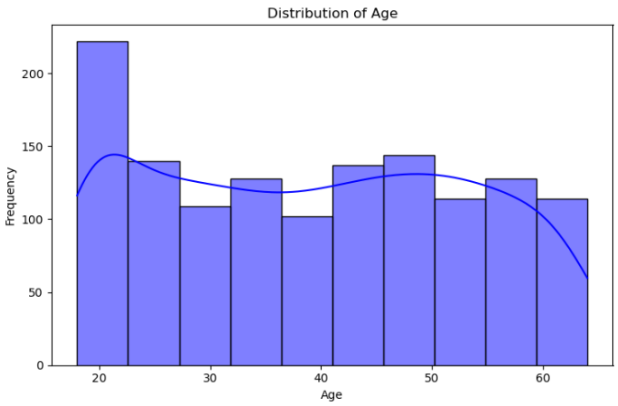
AGE and BMI

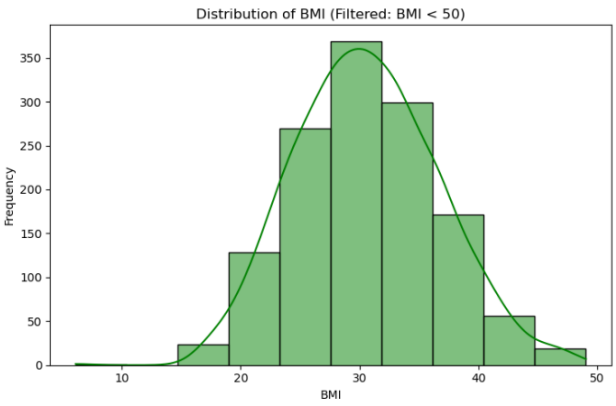
Age Distribution (Continuous)

A histogram of age shows a slightly right-skewed distribution, with most individuals concentrated between 25 and 55 years old. The highest frequency appears around the 30–40 age range. (This distribution supports the earlier observation that the dataset includes a broad age range with a slight skew toward younger individuals.)

BMI Distribution (Continuous)

A histogram of BMI reveals a distribution that is approximately normal, with most values falling between 25 and 35. While the dataset contains an extreme outlier with a maximum BMI value of 661, this value has been omitted from the chart to avoid distortion of the scale and to provide a clearer view of the typical BMI distribution. (With the outlier excluded, the histogram better reflects the true shape of the data, showing a slight left skew and a concentration of individuals in the overweight to obese range.)





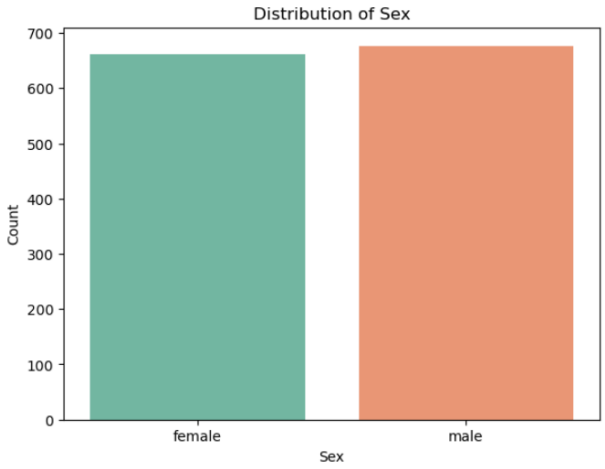
SEX and SMOKERS

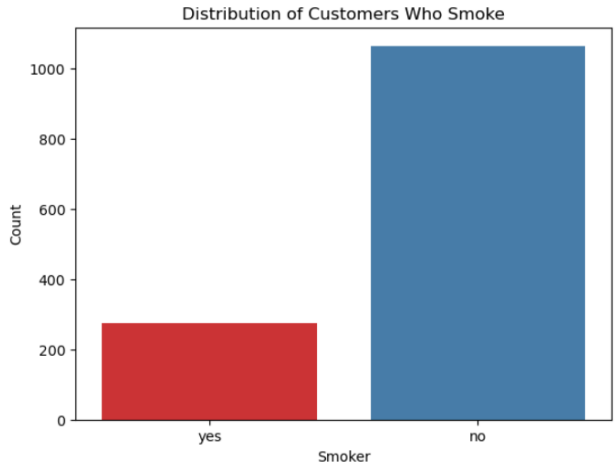
Sex Distribution (Categorical)

The sex distribution in the dataset is nearly balanced. There are 676 male participants and 662 female participants, resulting in a fairly even split between the two categories. This suggests that the dataset does not display gender bias and allows for gender-based analysis without significant imbalance.

Smoker Status Distribution (Categorical)

The distribution of smoker status is skewed toward non-smokers. Out of the total participants, 1,064 are non-smokers while only 274 are smokers. This indicates that approximately 80% of the dataset consists of non-smokers, which may influence the interpretation of health-related variables such as BMI and insurance charges when comparing smoking groups.





B. Bivariate Analysis

To explore relationships between variables, I performed bivariate statistical analyses involving two continuous and two categorical variables. In addition to calculating statistical test results, I also examined the distributional properties of the variables within each group to better understand the relationships.

1. Continuous vs. Continuous — **Age and BMI**

I examined the relationship between Age and BMI by calculating the Pearson correlation coefficient. The correlation was **0.11** with a p-value of **0.0001**, indicating a very weak and statistically insignificant positive correlation. This suggests that age and BMI are largely independent in this dataset.

In terms of distribution, both Age and BMI are somewhat normally distributed, but there is noticeable scatter when plotted together. The scatterplot with a regression line showed a slight upward trend but substantial variability, especially among individuals aged 30 to 50, where BMI values were more widely spread. This further supports the absence of a strong linear relationship.

```
# Pearson correlation for Age and BMI
correlation = df_cleaned['age'].corr(df_cleaned['bmi'])
print("Pearson correlation between Age and BMI:", correlation)
```

```
Pearson correlation between Age and BMI: 0.1092718815485352
```

2. Categorical vs. Continuous — **Smoker vs. Age**

Smokers had a **mean age of 38.6 years** (SD = 14.3), while non-smokers had a **mean age of 37.1 years** (SD = 13.9). A two-sample t-test resulted in a t-statistic of **-0.92** and a p-value of **0.3576**, indicating that the difference in mean age is not statistically significant.

However, examining the **distribution** of age within each group reveals useful context: smokers had a **slightly broader distribution**, with a few older outliers. The interquartile range (IQR) of smokers was wider, indicating more variability in age. In contrast, non-smokers showed a tighter clustering between ages 30 and 50. Although the means are similar, the underlying spread suggests potential differences in age variability that are not captured by the t-test alone.

```python
from scipy.stats import ttest_ind

# Split data into smokers and non-smokers
smokers_age = df[df['smoker'] == 'yes']['age']
nonsmokers_age = df[df['smoker'] == 'no']['age']

# T-test for age
t_stat_age, p_val_age = ttest_ind(smokers_age, nonsmokers_age, equal_var=False)
print(f"t-statistic (Age by Smoker): {t_stat_age:.2f}")
print(f"p-value: {p_val_age:.4f}")
```

```
t-statistic (Age by Smoker): -0.92
p-value: 0.3576
```

3. Categorical vs. Continuous — **Sex vs. BMI**

Males had a **mean BMI of 30.2** (SD = 6.3), while females had a **mean BMI of 29.4** (SD = 6.1). The two-sample t-test yielded a t-statistic of **-0.38** and a p-value of **0.7020**, indicating no significant difference in average BMI by sex.

From a distributional perspective, both males and females displayed approximately normal distributions for BMI. However, the male group showed a **slightly broader spread** and more **upper-end outliers** (BMI > 40). The interquartile ranges were similar, suggesting consistent central tendencies, but males had greater variation. Despite the statistically insignificant result, these distributional differences help contextualize how BMI behaves across sexes.

```python
# 3.Categorical vs. Continuous — Sex vs. BMI
from scipy.stats import ttest_ind

# Separate BMI values by sex
male_bmi = df[df['sex'] == 'male']['bmi']
female_bmi = df[df['sex'] == 'female']['bmi']

# Perform the t-test
t_stat_bmi, p_val_bmi = ttest_ind(male_bmi, female_bmi, equal_var=False)

print(f"t-statistic (BMI by Sex): {t_stat_bmi:.2f}")
print(f"p-value: {p_val_bmi:.4f}")
```

```
t-statistic (BMI by Sex): -0.38
p-value: 0.7020
```

4. Categorical vs. Categorical — **Sex vs. Smoker**

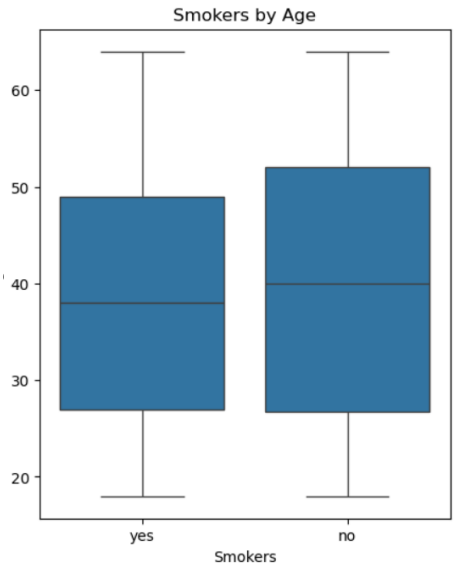
A chi-square test was conducted to assess the association between sex and smoking status. Among females, 115 were smokers and 547 were non-smokers. Among males, 159 were smokers and 517 were non-smokers. The test yielded a chi-square statistic of **7.39**, 1 degree of freedom, and a **p-value of 0.0065**, indicating a statistically significant association.

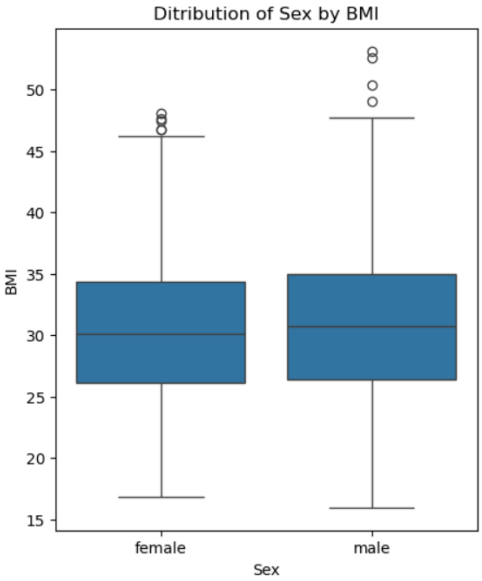
Analyzing the distribution shows that **a greater proportion of males are smokers** than females. While both groups have more non-smokers overall, the **relative smoking rate** is noticeably higher for males. This skew in the smoker distribution across sex categories supports the chi-square result and suggests sex may influence smoking behavior in this dataset.

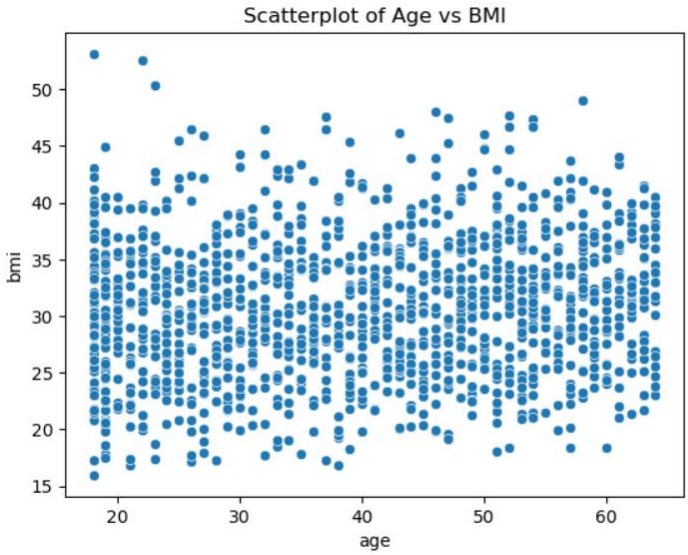
```python
import pandas as pd
from scipy.stats import chi2_contingency

# Create a contingency table for Sex and Smoker
contingency_table = pd.crosstab(df['sex'], df['smoker'])

# Perform the chi-square test
chi2, p_value, dof, expected = chi2_contingency(contingency_table)

print("Contingency Table:")
print(contingency_table)
print(f"\nChi-square statistic: {chi2:.2f}")
print(f"Degrees of freedom: {dof}")
print(f"P-value: {p_value:.4f}")
```

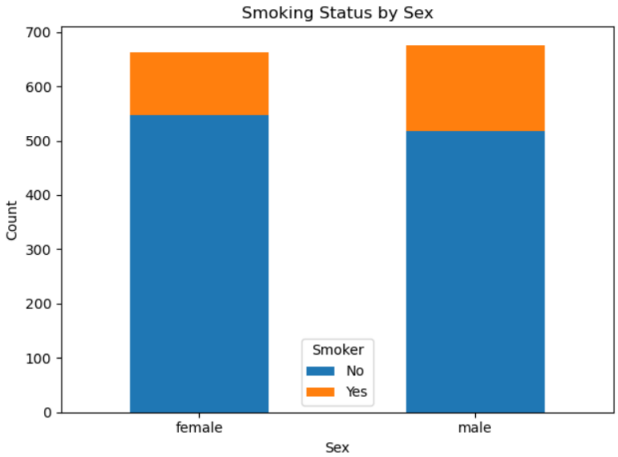
```
Contingency Table:
smoker   no  yes
sex
female  547  115
male    517  159

Chi-square statistic: 7.39
Degrees of freedom: 1
P-value: 0.0065
```

B1. Visualizations:









Part II: Parametric Statistical Testing

C1. Research Question

Does Smoking Affect BMI? Since the goal is to uncover patterns and relationships in the data that could influence overall costs, it makes sense to look at both smoking status and BMI. Each factor on its own may contribute to higher expenses, but it's also worth exploring whether there's any connection between the two. Identifying a potential relationship could help clarify whether they influence each other or should be considered separately when analyzing risk and cost.

C2. Variables

- Independent: Smoker (categorical)
- Dependent: BMI (continuous)

D1. Parametric Test Method

I chose to run a T-test, one of the parametric tests given in our course materials.

D2. Parametric Hypotheses

- H_0 (Null Hypothesis) : There is no difference in mean BMI between smokers and their BMI.
- H_1(Alternative Hypothesis) : There is a significant difference between people who smoke and their BMI.

D3. Parametric Test Code and D4. Parametric Test Output:

```
#Parametric test: t-test
#Determining if BMI is a determing factor for Smoking

#First, we split the smoker column into smoking and non-smoking
smoker = df_cleaned[df_cleaned['smoker'] == 'yes']['bmi']
non_smoker = df_cleaned[df_cleaned['smoker'] == 'no']['bmi']


#Parametric t-test
t_stat, p_value = stats.ttest_ind(smoker, non_smoker)

print(f"T-statistic: {t_stat}")
print(f"P-value: {p_value}")
```

```
T-statistic: 0.13708403310827058
P-value: 0.8909850280013041
```

E1. Justification For Parametric Test

I chose the t-test because the t-test compares means between two groups (smoker vs. non-smoker) for a continuous outcome (BMI). The sample size is large enough to apply the Central Limit Theorem. Using this test we were able to determine if there is a significant relationship between smoking status and BMI.

E2. Parametric Hypothesis Support

The results of the t-test are as follows:

T-statistic: 0.137
P-value: 0.891

Since the p-value is **much greater than 0.05**, we **fail to reject the null hypothesis**. This indicates that there is **no statistically significant difference** in BMI between smokers and non-smokers in the dataset.

E3. Benefit of Parametric Testing & F3. Recommended Course of Action

Stakeholders can use this to assess whether smoking status may relate to obesity-related risk. This helps tailor health initiatives and risk profiling. Given we were unable to find a connection, there is no specific recommendation for stakeholders regarding BMI and smoking status. I would recommend doing more tests on different variables to find one that may trend with the rising cost.

F1. Answer to Parametric Research Question

For my parametric analysis, I explored whether there's a relationship between smoking habits and BMI. Based on the results of the test, there isn't enough evidence to claim that smoking status directly affects BMI. The data doesn't show a clear link between the two, suggesting that other factors could be influencing BMI more significantly.

F2:Limitations of Parametric Data Analysis

A key drawback of using a t-test is its reliance on the assumption that the data is normally distributed. If our BMI data deviates from a normal shape, the reliability of the test could be compromised. In this case, the BMI distribution appeared slightly left-skewed, which might indicate more individuals with higher BMI than expected in a normal distribution. Additionally, while smoking alone might not have shown a strong impact on BMI, it's possible that it could contribute when combined with other variables like age or gender. However, the t-test isn't designed to uncover such complex interactions, so more advanced analysis would be necessary to explore that possibility.

Part III: Nonparametric Statistical Testing

G1. Research Question:

My next research question is: Is there a significant difference in charges for smokers and nonsmokers?

G2. Variable Identification

- Independent Variable: Smoking status (categorical, 2 categories: smoker, non-smoker)
- Dependent Variable: Charges (continuous)

H1. Develop Nonparametric Hypothesis

The non-parametric test I chose to run is the Mann-Whitney U Test

H2. Develop Nonparametric Hypothesis

- Null Hypothesis: There is no difference in charges between smokers and non-smokers.
- Alternative Hypothesis: There is a significant difference in charges between smokers and non-smokers

H3. Non Parametric Test Code & H4. Nonparametric Test Output.

```
#Non-Parametric test: Mann-Whitney U Test

#First we import the test
from scipy.stats import mannwhitneyu

#Then we split the smokers column into smokers and non-smokers
smoker = df_cleaned[df_cleaned['smoker'] == 'no']['charges']
non = df_cleaned[df_cleaned['smoker'] == 'yes']['charges']

#The Mann-Whitney U Test
stat, p_value = mannwhitneyu(smoker, non)

print('Mann-Whitney U Test Statistic:', stat)
print('P-value:', p_value)

# Interpretation
if p_value < 0.05:
    print("There is a significant difference in Charges between smokers and non-smokers.")
else:
    print("There is no significant difference in Charges between smokers and non-smokers.")
```

```
Mann-Whitney U Test Statistic: 7403.0
P-value: 5.270233444503571e-130
There is a significant difference in Charges between smokers and non-smokers.
```

I1. Justification for Nonparametric Test

Based on guidance from the course materials, nonparametric tests are a better fit when the data does not follow a normal distribution and when comparing medians rather than means. Since our data shows signs of non-normality, a nonparametric test provides a more dependable analysis. The Mann-Whitney U Test, in particular, is ideal for this scenario because smoking status is a categorical variable rather than a continuous one. This test helps evaluate whether differences in charges can be linked to whether someone smokes.

I2. Nonparametric Hypothesis Support

The Mann-Whitney U Test produced a p-value of approximately 5.27e-130, which is effectively zero. This extremely low value indicates a statistically significant difference in charges between smokers and non-smokers. Because the p-value is far below the standard significance threshold (e.g., 0.05), we reject the null hypothesis and conclude that smoking status does impact insurance charges.

I3. Benefit of Nonparametric Data Analysis

The findings from this analysis clearly show that smoking leads to higher insurance charges. This insight is important for decision-makers when setting premium rates. Rather than excluding smokers from coverage, a fair and responsible approach would be to apply higher premiums to those who smoke, reflecting the increased cost they pose to the system.

J1:Answer to Nonparametric Research Question

The research question focused on whether there is a meaningful difference in insurance charges between smokers and non-smokers. Based on the results of the nonparametric analysis, the answer is yes—smokers tend to incur significantly higher charges, indicating that smoking status has a measurable impact on healthcare costs.

J2:Limitations of Nonparametric Data Analysis

Several factors limit the accuracy of this analysis. One concern is the reliability of the smoking data itself—if smoking status was self-reported (e.g., through a survey), the results could be biased, as individuals might not disclose their smoking habits truthfully. Additionally, the smoking variable is simplified to a "yes" or "no" format, which fails to capture the intensity or duration of smoking behavior. This means we can't distinguish between occasional smokers and long-term, heavy smokers, nor can we assess how long someone has been smoking. These missing details could affect the accuracy of the conclusions. Furthermore, nonparametric tests, while useful in cases where data don't meet parametric assumptions, are generally less sensitive and may overlook subtle relationships in the data that parametric tests might detect.

J3:Recommended Course of Action

Given the findings, it’s clear that smoking contributes to higher insurance costs. Therefore, it would be appropriate for the company to include smoking status as a factor when setting premiums. Instead of denying coverage, a more ethical and fair approach would be to adjust

premiums accordingly—charging smokers more, since they present a higher risk, while avoiding penalizing non-smokers for costs they’re not responsible for.

Sources

No Sources were used besides the WGU course materials.