D599 - Data Preparation and Exploration - Task 2
By Eric Williams

Part I: Univariate and Bivariate Statistical Analysis and Visualization

A: UNIVARIATE STATISTICS

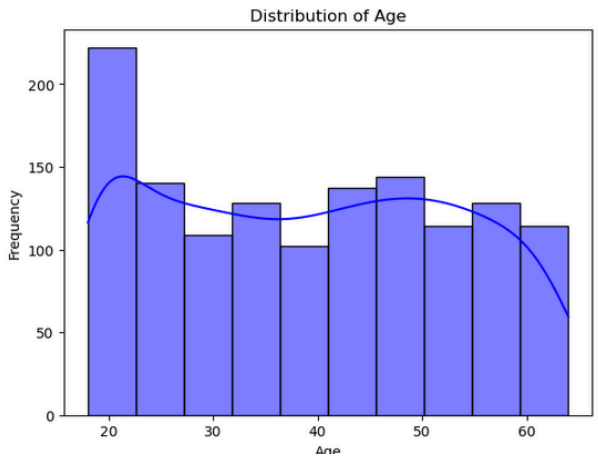
For my continuous variables, I chose Age and BMI. I plotted both distributions in a histogram. For my categorical tables, I chose Sex and Smoker to display in a countplot. I ran some statistics describing these variables and these were the results:

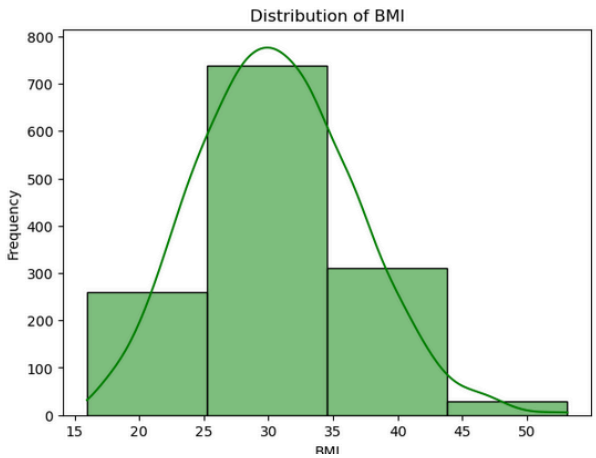
```
#Descriptive statistics for variables
variable_columns = ['age', 'bmi', 'sex', 'smoker']
stats = df_cleaned[variable_columns].describe(include='all')

# Display the results
print(stats)
```

	age	bmi	sex	smoker
count	1338.000000	1338.000000	1338	1338
unique	NaN	NaN	2	2
top	NaN	NaN	male	no
freq	NaN	NaN	676	1064
mean	39.207025	30.663397	NaN	NaN
std	14.049960	6.098187	NaN	NaN
min	18.000000	15.960000	NaN	NaN
25%	27.000000	26.296250	NaN	NaN
50%	39.000000	30.400000	NaN	NaN
75%	51.000000	34.693750	NaN	NaN
max	64.000000	53.130000	NaN	NaN

A1: VISUAL OF FINDINGS FROM PART A

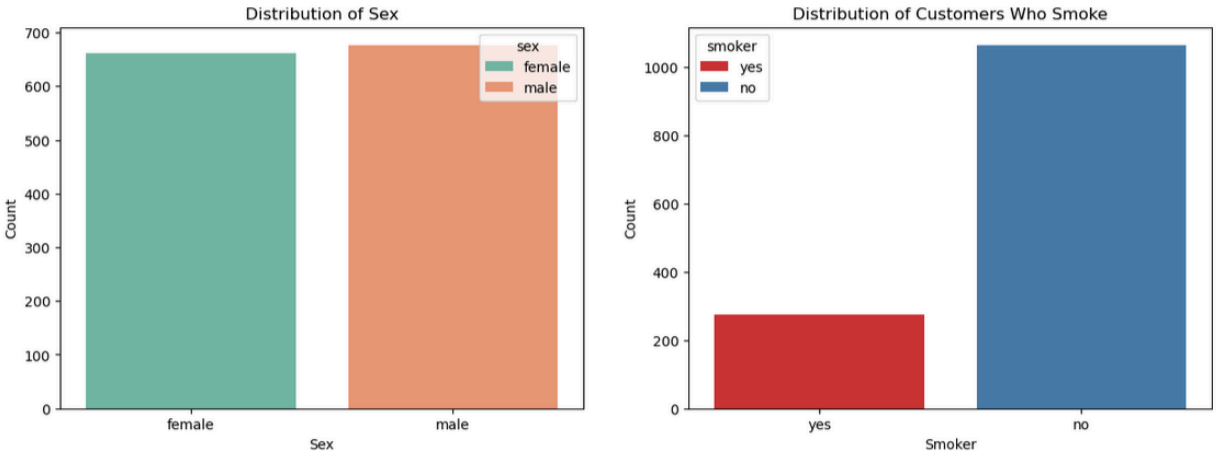
Two continuous variables, Age and BMI:





Distributions for both Age and BMI: As we can see above, the age distribution of Age heavily skewed right and is overrepresented by the younger population. BMI is much more normally distributed, but is slightly skewed by larger BMI, meaning the data is skewed slightly left.

Two categorical variables, Sex and Smoker:



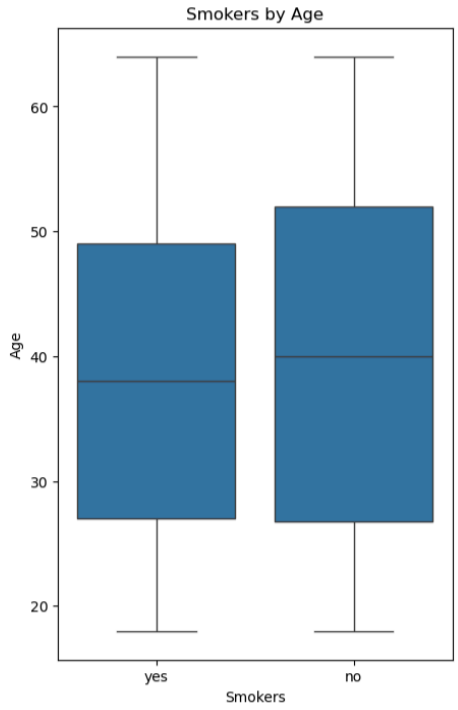
Distributions for both sex and smoking status: The distribution of male versus female is very close to even. However, in the distribution of smokers, we can see that nonsmokers are much more heavily represented in our data.

B: BIVARIATE STATISTICS

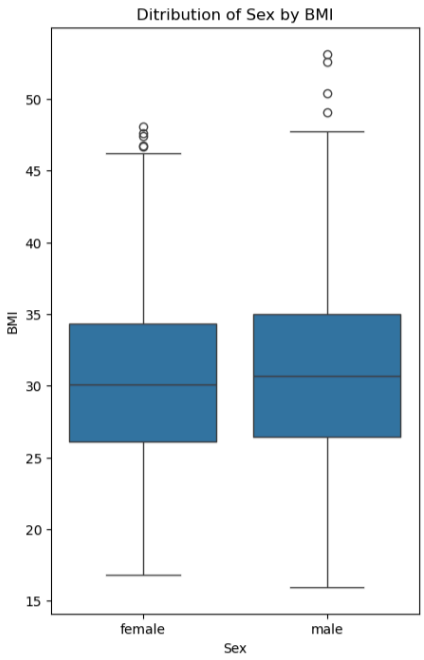
For my bivariate statistics, I chose to plot one continuous variable against a categorical variable. I did this twice, so there are two continuous variables and two categorical variables in the distributions below.

The first plot is a bivariate analysis on Smoker (a categorical variable) and Age (a continuous variable) in a box and whisker plot. The second plot is a bivariate analysis on Sex (a categorical variable) and BMI (a continuous variable) also in a box and whisker plot.

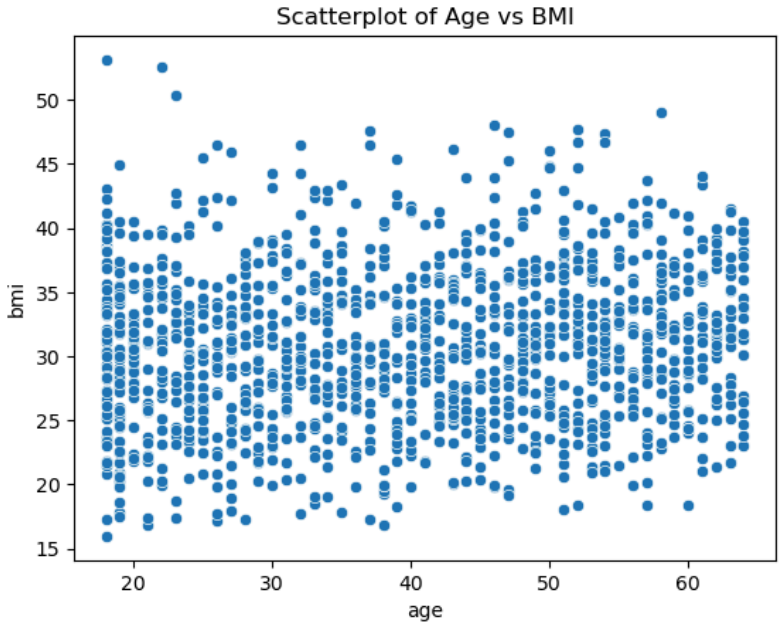
B1: VISUAL OF FINDINGS FROM PART B



Distribution for both age and smoking status: As you can see above, smoking and nonsmoking by age are fairly similar, however, there are a few differences. The median age of nonsmokers is higher and the 3rd quartile trends older, even though the minimums and maximums of Q1 are identical. To establish this connection further, we'll need to do some parametric test analysis below.



Distribution for both sex and BMI: As you can see above, BMI for men and women are fairly similar, however, there are a few differences. The median BMI is slightly higher in men, and men generally have more extreme values (Q1 and Q3 values are more extreme and the outliers are higher). Q1 and the minimum values are distributed fairly similarly for both men and women.



Distribution for age and bmi: The distribution of BMI by age appears to be mostly random without well defined trends. As you can see above, there is a wide variety of BMI for every age group and the two do not seem to be well connected.

Outside of the parametric testing I'll use below to answer the business question, I decided to do a quick t-test to look for the bivariate relationship between BMI and age to make sure I fulfill the requirement to provide bivariate statistics.

```
#T-test for BMI between males and females
male_bmi = df_cleaned[df_cleaned['sex'] == 'male']['bmi']
female_bmi = df_cleaned[df_cleaned['sex'] == 'female']['bmi']

#The T-test
t_stat, p_value = stats.ttest_ind(male_bmi, female_bmi)
print(f"T-test for BMI between Males and Females: t_stat = {t_stat}, p_value = {p_value}")
```

```
T-test for BMI between Males and Females: t_stat = 1.696752635752224, p_value = 0.08997637178984932
```

Another bivariate statistic is given below for my parametric test when analyzing if BMI has an impact on smoking.

Part II: Parametric Statistical Testing

C1:RESEARCH QUESTION

Does smoking impact BMI? Because the business is trying to identify trends in the data and connections between variables, both BMI and smoking could potentially contribute to higher costs. It might be helpful to try to separate the two, or find a relationship between the two to see if they are impacting each other.

C2:VARIABLE IDENTIFICATION

The variables I will use in this analysis are BMI and Smoking.

D1:PARAMETRIC TEST METHOD

I will run a T-test, one of the parametric tests from the course materials.

D2:DEVELOP PARAMETRIC HYPOTHESES

Null Hypothesis: There is no significant relationship between people who smoke and their BMI.

Alternative Hypothesis: There is a significant relationship 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 a t-test because it's a powerful and efficient tool for comparing the mean between two independent groups. The recommendations for the test as per the course materials were generally met: BMI served as a continuous variable, the sample size is somewhat large, and we want to compare the mean of two datasets. 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 listed above are:

T-statistic: 0.13708403310827058
P-value: 0.8909850280013041

The null hypothesis stated that there is no significant difference in BMI and when comparing smokers and nonsmokers. Because the p-value did not reach the level of statistical significance (p=0.05), we cannot reject the null hypothesis. We cannot reasonably determine that there is a significant difference in the BMI of smokers and nonsmokers. The two factors do not seem to trend together or predict each other.

E3:BENEFIT OF PARAMETRIC TESTING and F3:RECOMMENDED COURSE OF ACTION

If we had been able to reject the null hypothesis, I could have recommended the company accounts for BMI when analyzing smoker status. This would have been useful information when determining if it is the BMI or the smoking that is raising costs. However, given that we were unable to find a connection, there is no specific recommendation for stakeholders regarding BMI

pay and smoking status. However, this analysis did provide the benefit of ruling out a significant relationship between two variables that might impact cost. My recommendation is to do several more tests and to run more analysis on different variables to find one that trends the most with rising costs.

F1:ANSWER TO PARAMETRIC RESEARCH QUESTION

My research question for my parametric test was: Does smoking impact BMI? The answer is that we cannot make a direct connection between smoking status and BMI. Given the results of the parametric test, a conclusion regarding correlation between the two variables cannot be determined. In short, a rise in BMI might be related to other variables.

F2:LIMITATIONS OF PARAMETRIC DATA ANALYSIS

One limitation of the t-test is that it assumes we have a normally distributed dataset. If our BMI data was skewed at all, our t-test may have been affected and been less valid. Looking at the BMI distribution visuals above, the distribution is only somewhat normal and could be considered slightly skewed left (meaning higher BMI might be overrepresented relative to a normal curve). Also, it could be true that smoking in conjunction with other factors (age and sex, for example) could be a predictor for increased BMI, even though smoking itself is not a determining factor. But this does not mean we can say definitively that smoking does not play a role at all. A simple t-test is unable to tell us if this is the case. Further analysis would be needed to be certain.

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

My two variables for the non-parametric test are smoker and charges.

H1:NONPARAMETRIC TEST METHOD

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

H2:DEVELOP NONPARAMETRIC HYPOTHESES

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:NONPARAMETRIC TEST CODE

```
#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.")
```

H4:NONPARAMETRIC TEST OUTPUT

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

According to course materials, non-parametric tests are useful when data is not normally distributed and for comparing medians of two variables. This means that since our data is not exactly normally distributed, this test should be more reliable and accurate than the parametric test. Also, the Mann-Whitney U Test is helpful when we have a variable (smoker) that is ordinal and not continuous. This test will determine if high charges can be attributed to smoker status.

I2:NONPARAMETRIC HYPOTHESIS SUPPORT

Because the output of the Mann-Whitney U Test was p=5.27e-130, we know that there is a

statistical difference in charges between smokers and nonsmokers. Because our p-value is very nearly 0, we must reject the null hypothesis.

I3:BENEFIT OF NONPARAMETRIC DATA ANALYSIS

This result provided a definitive answer to stakeholders that smoking determines higher charges. As such, when determining insurance premiums, the company should ask about and account for smoking status, with the knowledge that smokers will cost the company more money. Smokers should be charged a higher premium, since it would be unethical to deny smokers coverage outright.

J1:ANSWER TO NONPARAMETRIC RESEARCH QUESTION

The research question was: Is there a significant difference in charges for smokers and nonsmokers? The answer is yes. The analysis says there is a statistically significant cost in covering smokers.

J2:LIMITATIONS OF NONPARAMETRIC DATA ANALYSIS

A few things limit this from being a perfect analysis. Firstly, we need to examine how the smoking data was collected. If it was collected by survey, our data would be skewed to overrepresent people who are honest about their smoking status versus those who are not. It also does not account for nuance or the amount of smoking, as the data was a binary yes/no entry for smoking. With this data analysis, we cannot determine if heavy smoking raises costs, or if heavy smoking incurs costs more than light smoking. It also does not account for the number of years someone has been smoking--a young, new smoker might not register as an active smoker while not yet damaging their health enough to increase charges. In short, if the data doesn't accurately represent the customers, then the test will not be as accurate. Also, non-parametric tests aren't as sensitive to the small correlations and connections that parametric tests can find because parametric tests infer certain conditions.

J3:RECOMMENDED COURSE OF ACTION

As mentioned above, this result provided a definitive answer to stakeholders that smoking determines higher charges. As such, when determining insurance premiums, the company should ask about and account for smoking status, with the knowledge that smokers will cost the company more money. Smokers should be charged a higher premium, since it would be unethical to deny smokers coverage outright and it would be unfair to charge nonsmokers more money because smokers are driving up costs.

Sources

No sources were used besides the WGU course materials.