Peyton Bailey

March 15th, 2025

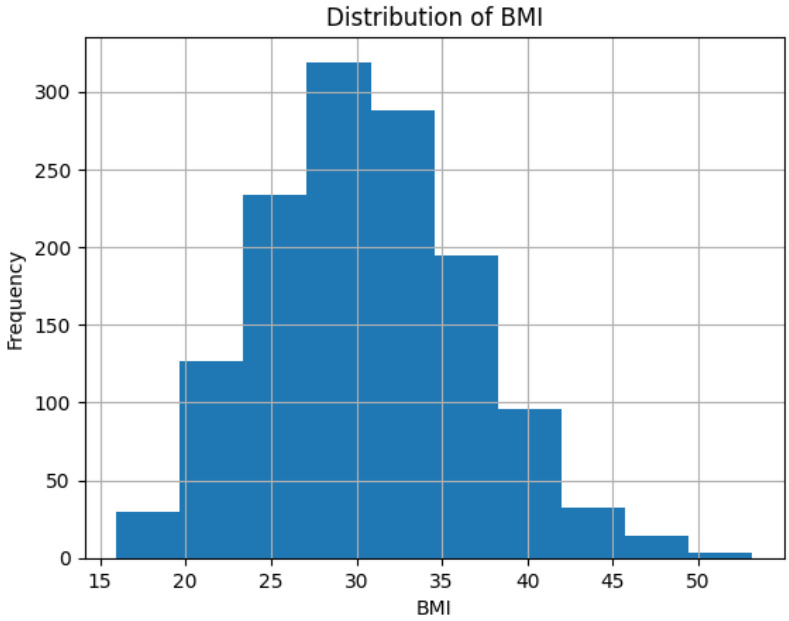
D599 – Task 2

A. Identify the distribution of two **continuous variables and** two **categorical variables using univariate statistics from the dataset.**

1. Represent your findings from part A visually as part of your submission.

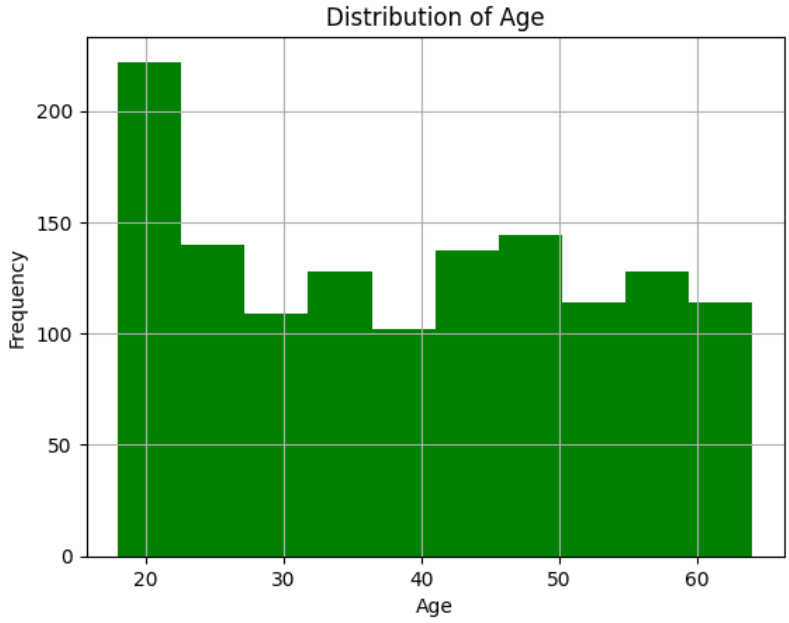
Continuous Variables

i. BMI



The BMI category is a continuous variable with a normal distribution. It has a mean of 30.66 with a minimum and maximum of 15.96 and 53.13, respectively. The standard deviation is 6.1, and 99.7% of the data falls within three standard deviations of the mean. Four data points are outliers. However, they do not invalidate the normality of the data.

ii. Age



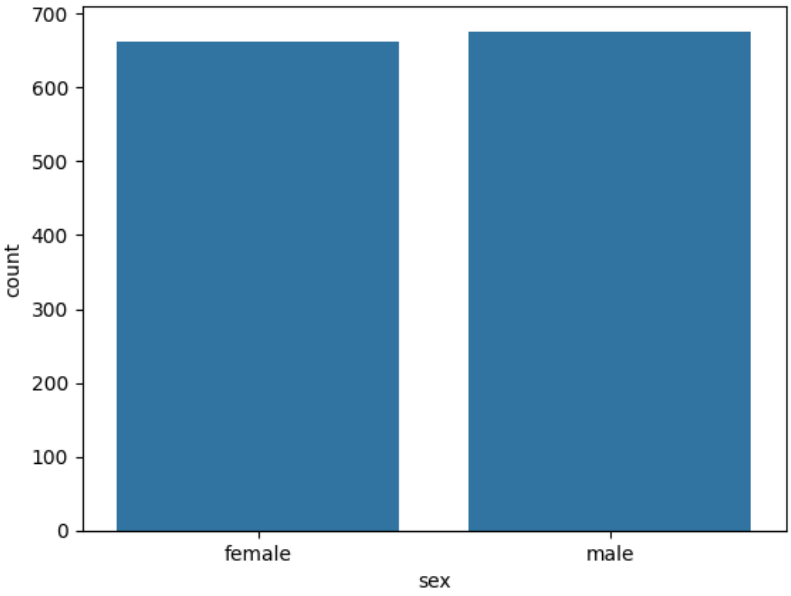
Age is another continuous variable; however, it does not have a normal distribution. Instead, a right-skewed distribution can be visually seen on a histogram. The data shows a high peak at

lower ages, around 20 years, and a gradual decline across higher ages. The minimum age is 18, and the maximum is 64. The median is 39, and the standard deviation is 14.05. The IQR of 24 spans from 27 to 51, which captures moderate variability within the age range.

Categorical Variables

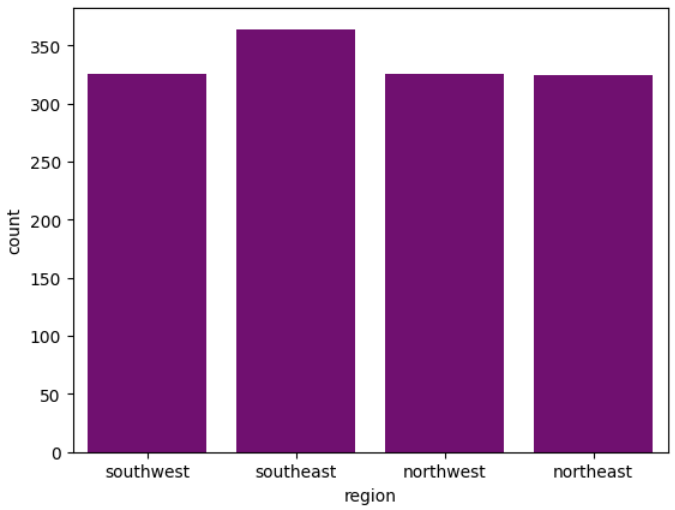
iii. Sex

Sex is a nominal categorical variable with two categories: male and female. There are 676 males and 662 females within the dataset. This is nearly a 50/50 split, but there are slightly more males than females (~1%). The bar plot on the right visualizes the proximity of the numbers.



iv. Region

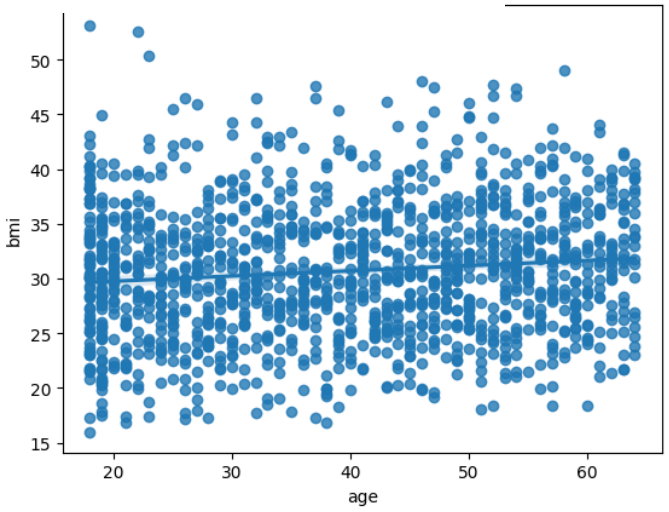
Region is another nominal categorical variable with four categories: northwest, northeast, southwest, and southeast. This is also an equal split; however, the southeast is in the majority, representing 27% of the data, while the remaining three are roughly 24% each.



B. Identify the distribution of two **continuous variables and** two **categorical variables using bivariate statistics from the dataset.**

1. **Represent your findings from part B visually as part of your submission.**

Age vs. BMI

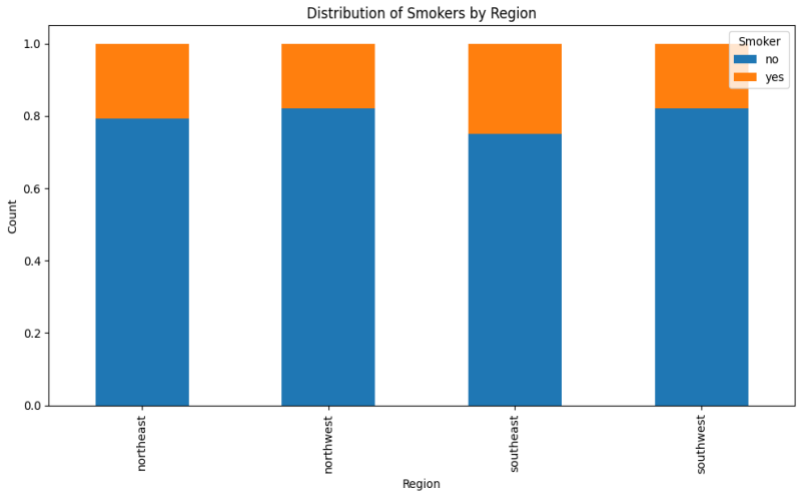


The continuous variables of age and BMI do not have a strong relationship. The BMI data points are widely scattered across all ages, meaning that the scatterplot does not suggest a clear upward or downward trend. There is a high concentration of BMI values around 25-35 across all ages. A fitted regression line resulted in a y-intercept of 28.8 and a slope of 0.05. The r-squared value of this regression was 0.012, which suggests that it is a very weak fit.

smoker	no	yes
region		
northeast	0.793210	0.206790
northwest	0.821538	0.178462
southeast	0.750000	0.250000
southwest	0.821538	0.178462

Region vs. Smoker

The insurance data was grouped by region, and values were counted for smokers vs. non-smokers. The following contingency table shows the distribution. Each region has a majority of non-smokers. The northwest and southwest regions have significantly more non-smokers (82%) than the northeast and southeast regions, which have 79% and 75% non-smokers, respectively. The chi-square test of independence on the two variables resulted in a p-value of 0.06, which means there is not enough evidence to reject the claim that smoking is independent of the region.



Part II: Parametric Statistical Testing

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

1. **Provide one research question relevant to the dataset and *any* organizational needs that can be answered through data analysis.**

Is there a relationship between smoking and BMI score?

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

Explanatory Variable – Smoker

Response Variable – BMI

D. Analyze the dataset by doing the following:

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

Independent t-test

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

H_O - There is no significant difference in the average BMI between smokers and non-smokers

H_A – There is a significant difference in the average BMI between smokers and non-smokers.

3. **Write code (in either Python or R) to run the parametric test.**

```
from scipy.stats import ttest_ind

smoker_bmi = insurance_data[insurance_data['smoker'] == 'yes']['bmi']
non_smoker_bmi = insurance_data[insurance_data['smoker'] == 'no']['bmi']

# Perform independent t-test
t_stat, p_value = ttest_ind(smoker_bmi, non_smoker_bmi)


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

4. **Provide the output and the results of *any* calculations from the parametric statistical test you performed.**

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

E. Evaluate parametric test results by doing the following:

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

I chose the independent t-test because the explanatory variable is categorical and has two groups (smoker and non-smoker), and the BMI variable is continuous with a normal distribution.

2. Discuss the test results, including the decision to reject or fail to reject the null hypothesis from part D2.

The p-value was significantly above the alpha level of 0.05, which failed to reject the null hypothesis. Ultimately, there is not enough evidence to support the claim that smoking has an impact on the average BMI.

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

Stakeholders can understand that smoking does not affect BMI. However, it still leaves them with many unanswered questions. In this case, we should do further analysis.

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

1. Discuss the answer to your question from part C1.

The Independent t-test results did not produce enough evidence to support the claim that smoking status affects BMI. This failed to reject the null hypothesis and conclude that there is no relationship between the two variables.

2. Discuss the limitations of your data analysis.

This test does not consider other factors affecting BMI score, such as demographics, body type, exercise, eating habits, etc. Certain factors coupled with smoking habits could contribute to a higher BMI within individuals.

3. Recommend a course of action based on your findings.

Since the test results did not provide enough evidence supporting a relationship between smoking and BMI, I cannot recommend a course of action regarding an interaction between these two variables. However, I recommend performing more tests to discover other potential patterns in the data.

Part III: Nonparametric Statistical Testing

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

1. Provide one research question relevant to the dataset and any organizational needs that can be answered through data analysis.

Do smokers have a higher average insurance charge than non-smokers?

2. Identify the variables in the dataset that are relevant to answering your research question from part G1.

Response variable – Charges

Explanatory variable – Smoker

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

H_O = The mean insurance charge of smokers is equal to that of non-smokers.

H_A = The mean insurance charge of smokers is greater than that of non-smokers.

3. Write code (in either Python or R) to run the nonparametric test.

```
# Separate charges by smoker status
smoker_charges = insurance_data[insurance_data['smoker'] == 'yes']['charges']
non_smoker_charges = insurance_data[insurance_data['smoker'] == 'no']['charges']

# Perform Mann-Whitney U-Test
stat, p_value = mannwhitneyu(smoker_charges, non_smoker_charges, alternative='greater')

# Print results
print(f"Mann-Whitney U Statistic: {stat}")
print(f"P-value: {p_value}")
```

✓ 0.0s

4. Provide the output and the results of *any* calculations from the nonparametric statistical test you performed.

```
Mann-Whitney U Statistic: 284133.0
P-value: 2.6351167222517853e-130
```

I. Evaluate nonparametric test results by doing the following:

1. Justify why you chose the statistical test identified in part G1 based on variables.

I chose the Mann-Whitney U Test because my explanatory variable is categorical with two groups, and my response variable is a continuous variable with no normal distribution. Therefore, this would be the appropriate non-parametric test based on these parameters.

2. Discuss test results, including the decision to reject or fail to reject the null hypothesis from part H2.

The test resulted in an extremely low p-value, which allowed me to reject the null hypothesis and accept the alternative hypothesis, indicating that the observed distribution of charges between

smokers and non-smokers is statistically significant. In conclusion, smokers tend to have higher insurance charges than non-smokers.

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

Stakeholders benefit from this knowledge by incorporating it during an initial risk assessment of their policyholders. By having clear evidence of smoking's impact on charges, they can justify the higher premiums.

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

1. Discuss the answer to your question from part G1.

The Mann-Whitney U test results show that smokers have significantly higher insurance charges than non-smokers.

2. Discuss the limitations of your data analysis.

The data does not consider pre-existing conditions that may affect the insurance charges. Additionally, the data may not represent all demographics and regions equally.

3. Recommend a course of action based on your findings.

I recommend that this insurance company implement wellness programs promoting healthier lifestyles and smoking alternatives. They should also conduct further analysis to include other factors such as preexisting conditions, exercise habits, family history, etc.