Robert Pavlik

D599 Data Preparation and Exploration

TCN1 Task 2: Data Exploration

Part 1: Univariate and Bivariate Analysis and Visualization

A. Univariate Analysis of Continuous and Categorical Variables

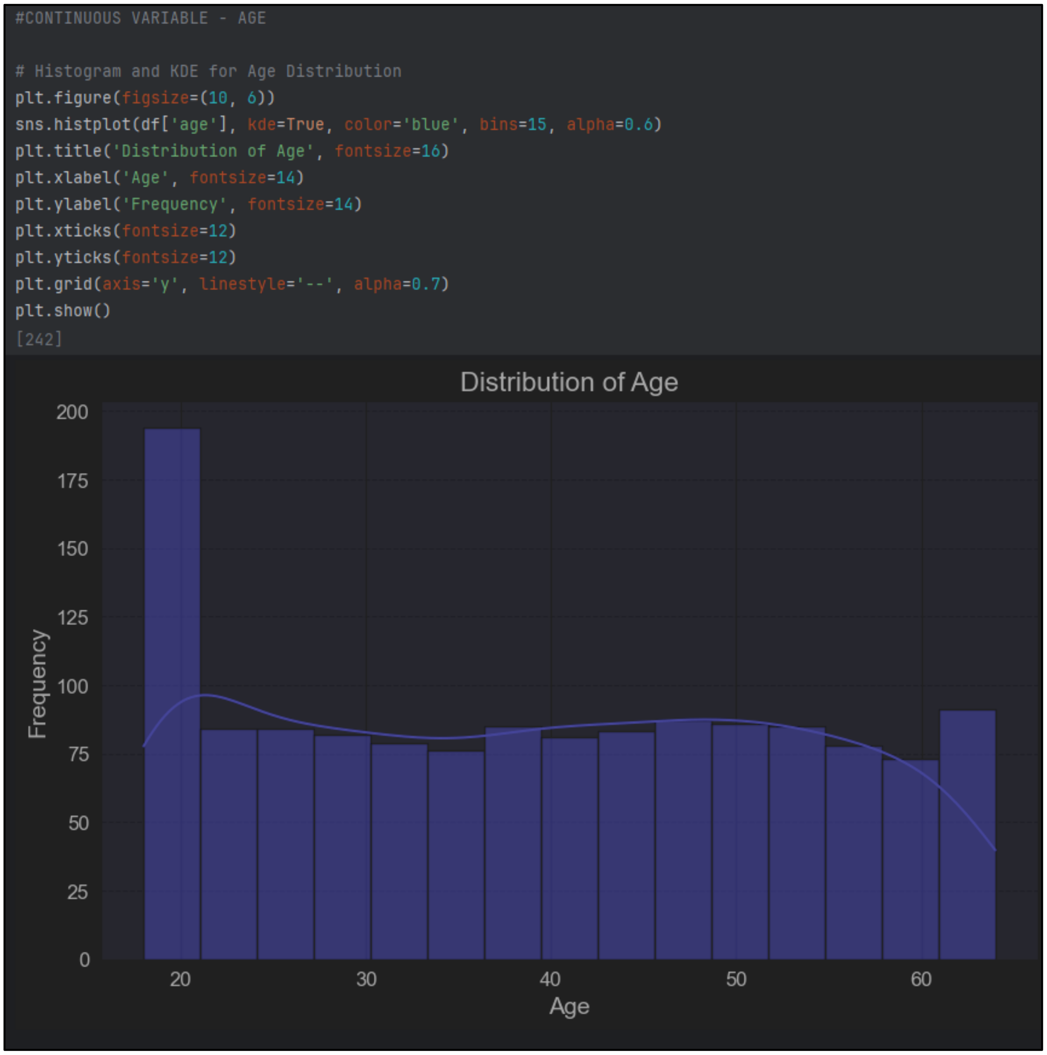
Continuous Variables

1. AGE Distribution:

The AGE variable's distribution is analyzed using a histogram and a boxplot. The histogram includes a Kernel Density Estimate curve, which provides insight into the distribution's shape. The skewness and median of the AGE variable are also calculated to provide additional information about the distribution.

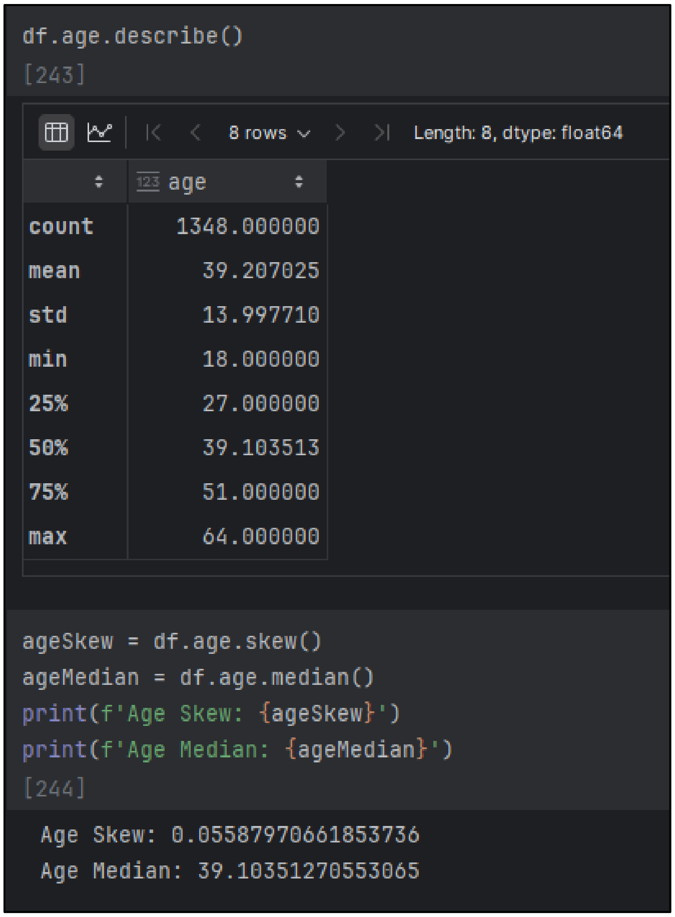
Skewness: The age distribution is slightly skewed

Median: The median age provides a central tendency measure consistent with the distribution shape.



```
#CONTINUOUS VARIABLE - AGE

# Histogram and KDE for Age Distribution
plt.figure(figsize=(10, 6))
sns.histplot(df['age'], kde=True, color='blue', bins=15, alpha=0.6)
plt.title('Distribution of Age', fontsize=16)
plt.xlabel('Age', fontsize=14)
plt.ylabel('Frequency', fontsize=14)
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
[242]
```

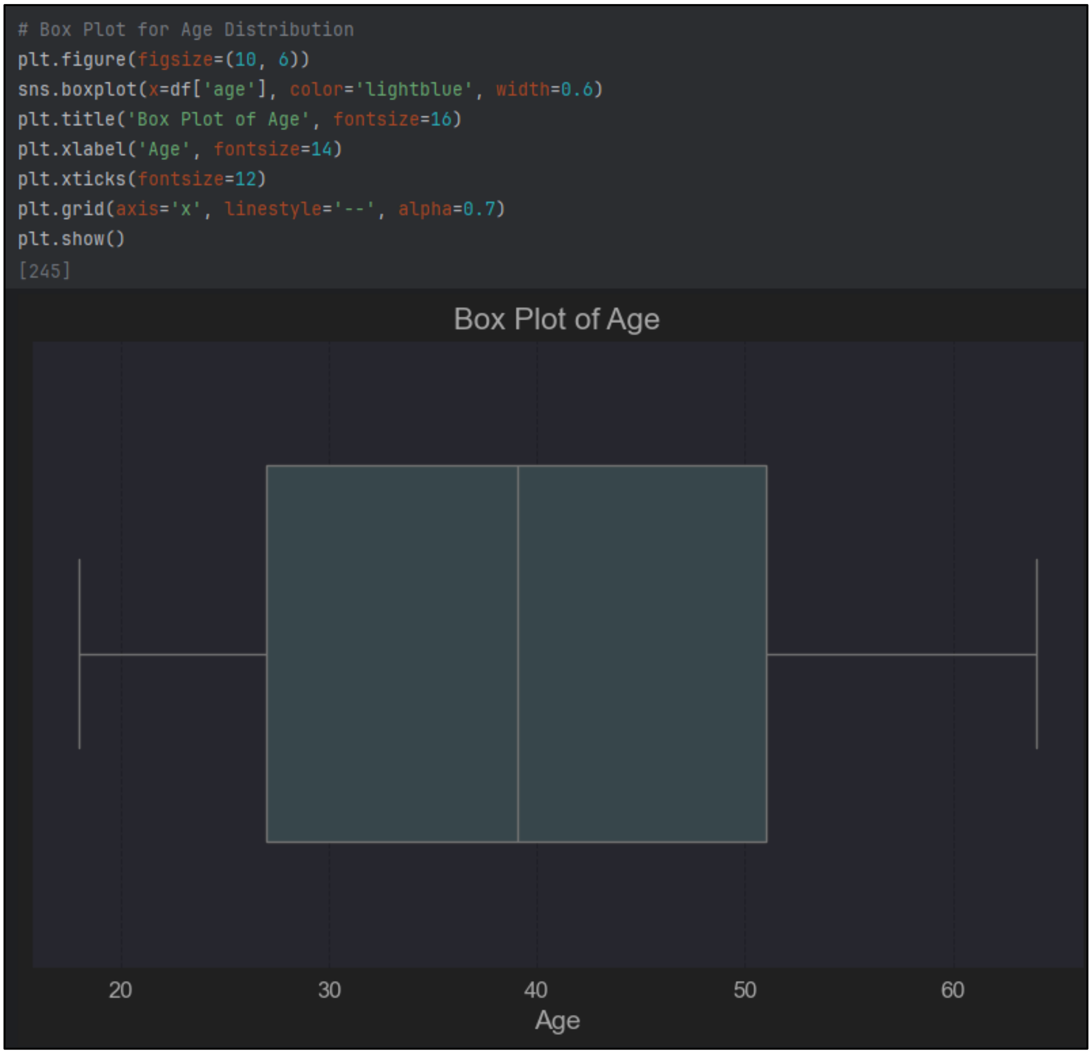


```
df.age.describe()
[243]
```

| | age |
|---|---|
| count | 1348.000000 |
| mean | 39.207025 |
| std | 13.997710 |
| min | 18.000000 |
| 25% | 27.000000 |
| 50% | 39.103513 |
| 75% | 51.000000 |
| max | 64.000000 |

```
ageSkew = df.age.skew()
ageMedian = df.age.median()
print(f'Age Skew: {ageSkew}')
print(f'Age Median: {ageMedian}')
[244]
```

```
Age Skew: 0.05587970661853736
Age Median: 39.10351270553065
```



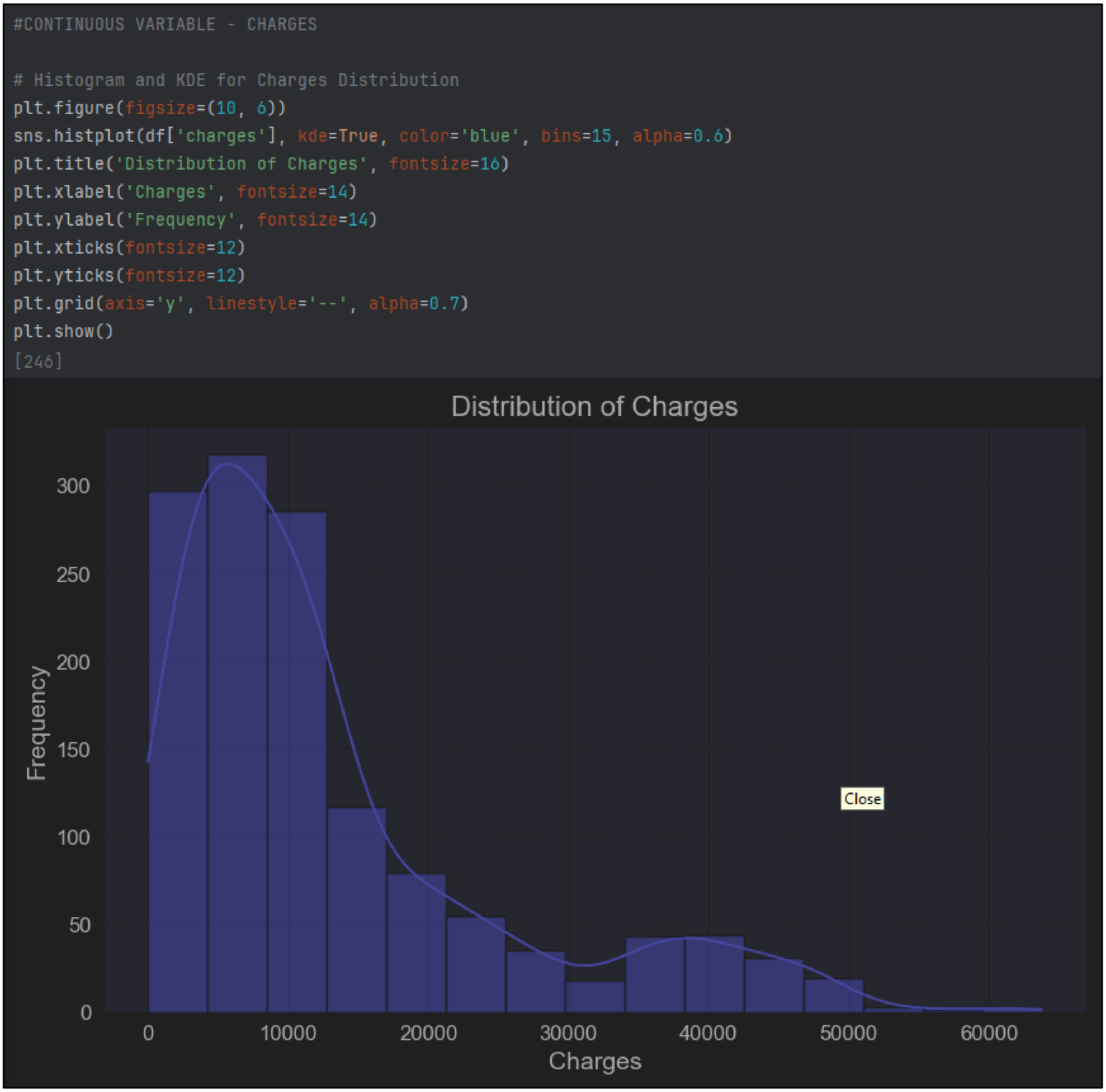
```
# Box Plot for Age Distribution
plt.figure(figsize=(10, 6))
sns.boxplot(x=df['age'], color='lightblue', width=0.6)
plt.title('Box Plot of Age', fontsize=16)
plt.xlabel('Age', fontsize=14)
plt.xticks(fontsize=12)
plt.grid(axis='x', linestyle='--', alpha=0.7)
plt.show()
[245]
```

2. CHARGES Distribution:

   Similarly, the CHARGES variable is examined with a histogram and KDE plot to understand the distribution. A boxplot is also used to detect outliers. Skewness and median values for CHARGES help summarize the distribution further.

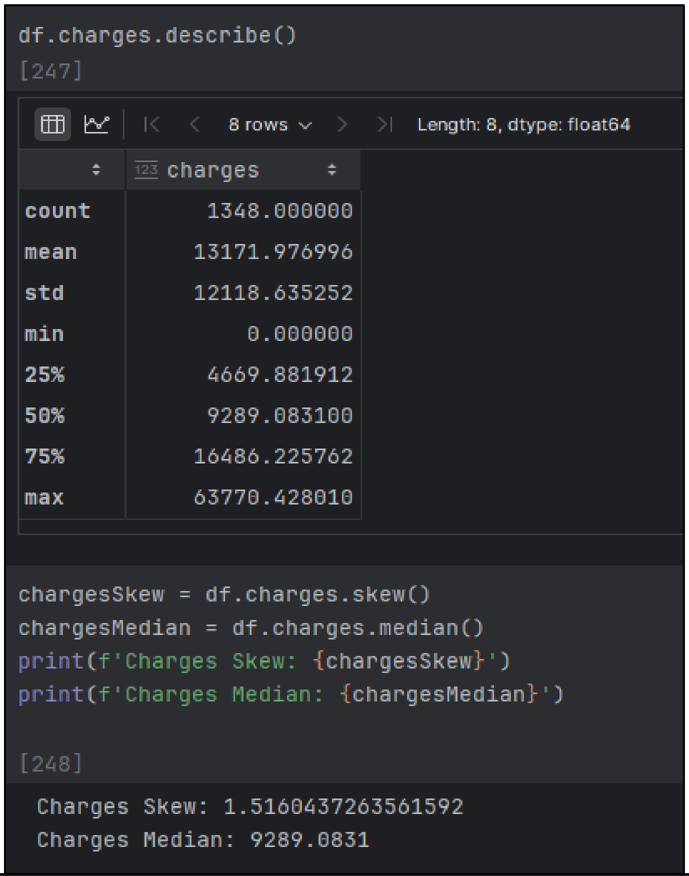
   Skewness: The charges distribution is right-skewed

   Median: The charges' median value helps indicate where most data points lie.



```
#CONTINUOUS VARIABLE - CHARGES

# Histogram and KDE for Charges Distribution
plt.figure(figsize=(10, 6))
sns.histplot(df['charges'], kde=True, color='blue', bins=15, alpha=0.6)
plt.title('Distribution of Charges', fontsize=16)
plt.xlabel('Charges', fontsize=14)
plt.ylabel('Frequency', fontsize=14)
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
[246]
```



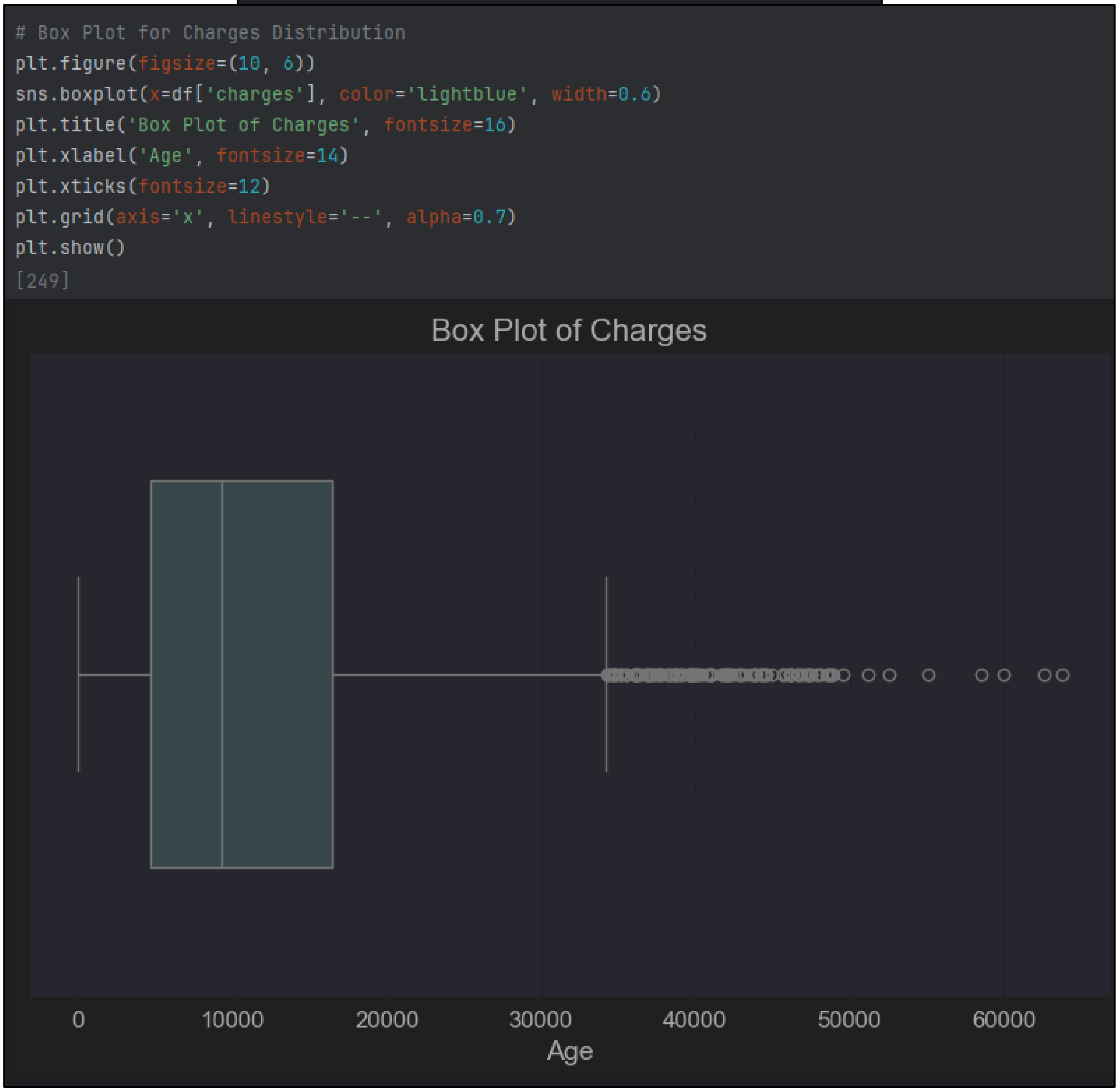
```
df.charges.describe()
[247]
```

```
        charges
count   1348.000000
mean    13171.976996
std     12118.635252
min     0.000000
25%     4669.881912
50%     9289.083100
75%     16486.225762
max     63770.428010
```

```
chargesSkew = df.charges.skew()
chargesMedian = df.charges.median()
print(f'Charges Skew: {chargesSkew}')
print(f'Charges Median: {chargesMedian}')

[248]
```

```
 Charges Skew: 1.5160437263561592
 Charges Median: 9289.0831
```

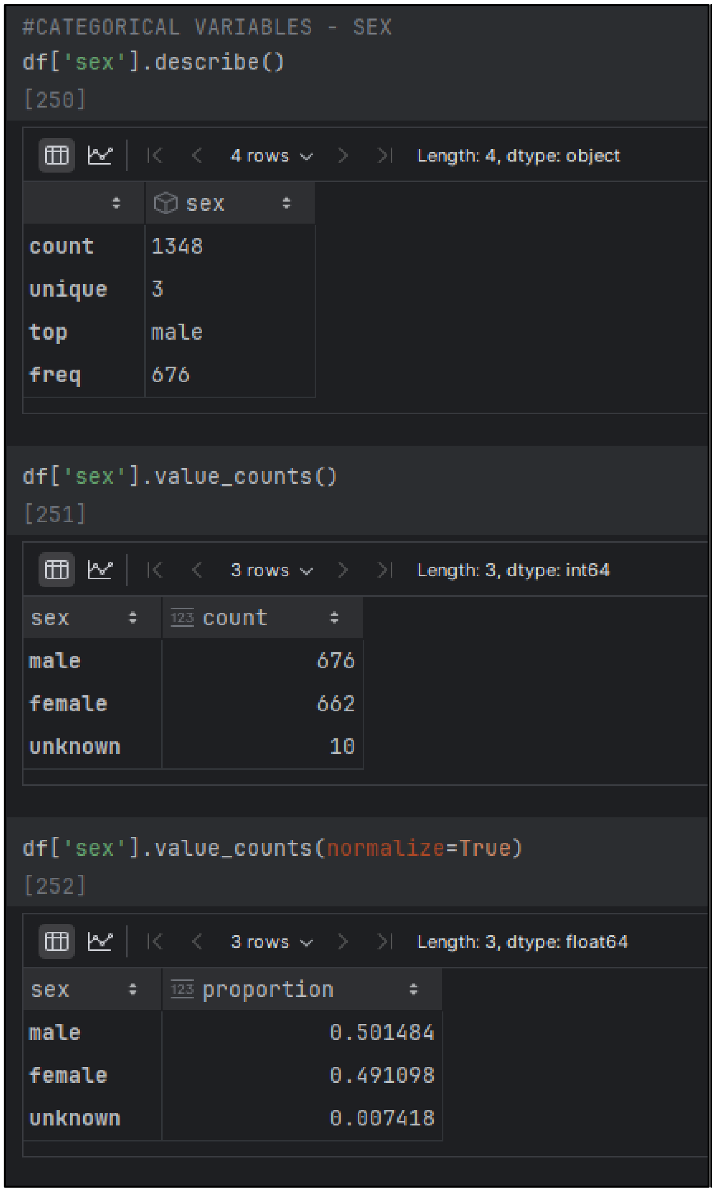


```
# Box Plot for Charges Distribution
plt.figure(figsize=(10, 6))
sns.boxplot(x=df['charges'], color='lightblue', width=0.6)
plt.title('Box Plot of Charges', fontsize=16)
plt.xlabel('Age', fontsize=14)
plt.xticks(fontsize=12)
plt.grid(axis='x', linestyle='--', alpha=0.7)
plt.show()
[249]
```

Categorical Variables

1. SEX Distribution:

   A countplot is used to display the distribution of males and females in the dataset along with proportions.



```
#CATEGORICAL VARIABLES - SEX
df['sex'].describe()
[250]
```

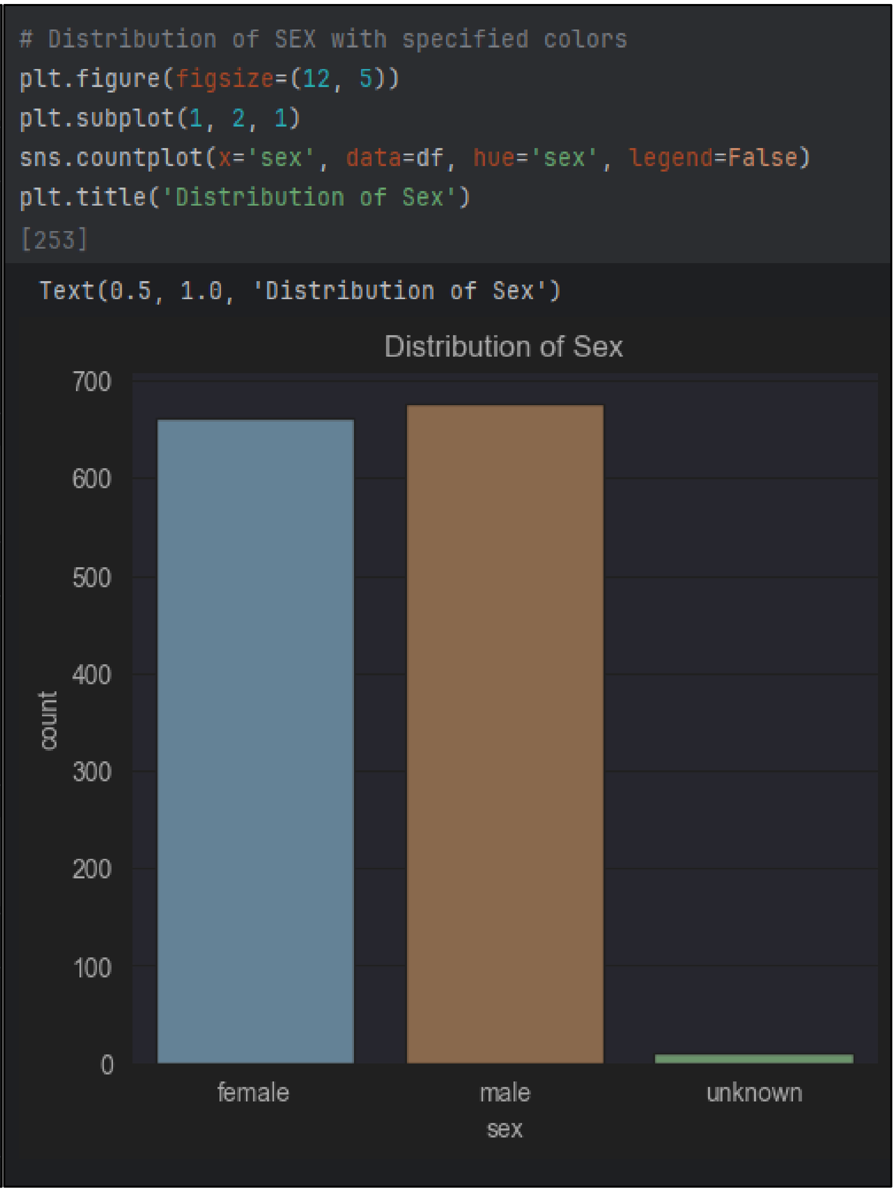
```
        sex
count   1348
unique  3
top     male
freq    676
```

```
df['sex'].value_counts()
[251]
```

```
sex      count
male       676
female     662
unknown     10
```

```
df['sex'].value_counts(normalize=True)
[252]
```

```
sex      proportion
male       0.501484
female     0.491098
unknown    0.007418
```

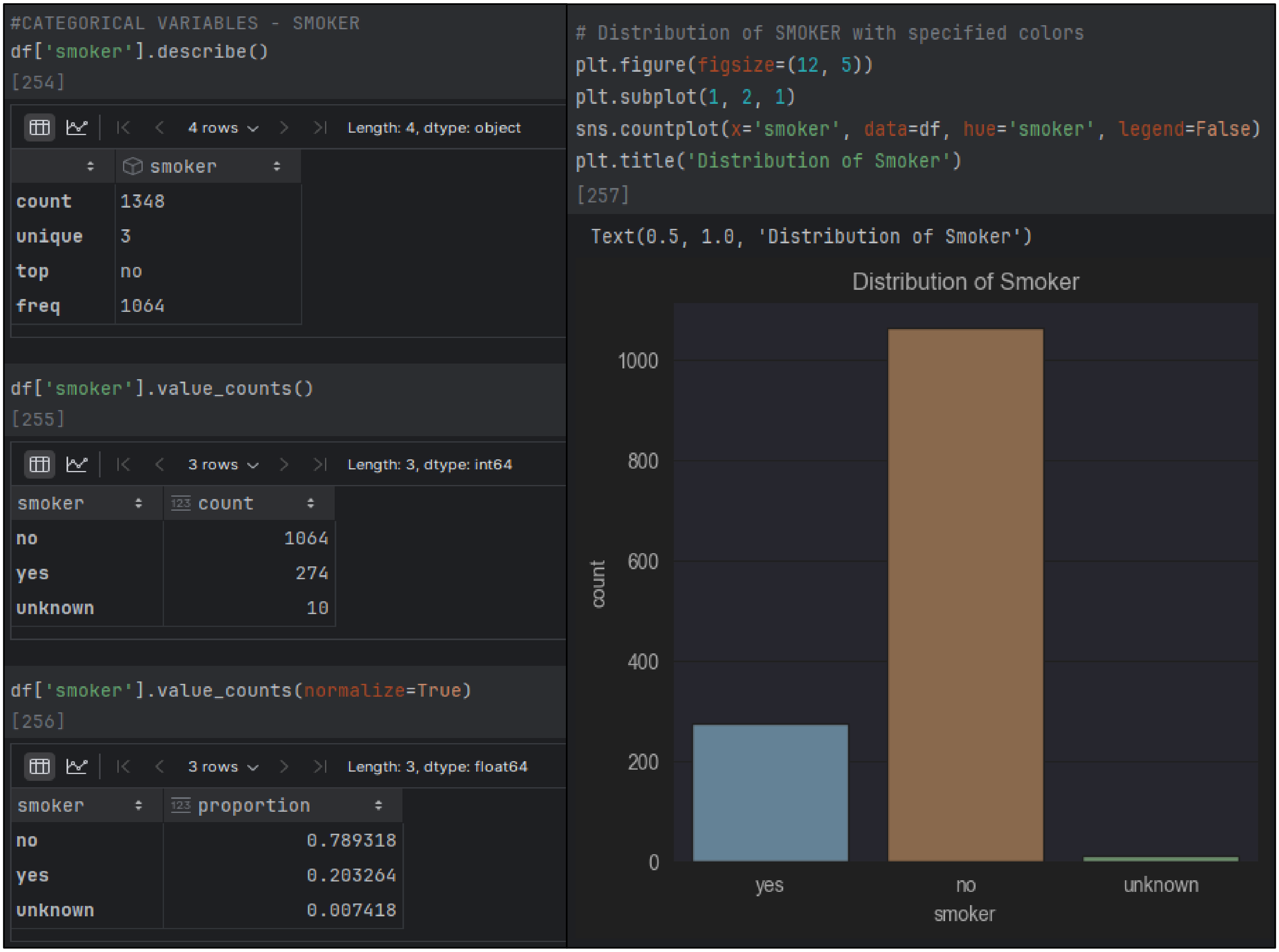


```
# Distribution of SEX with specified colors
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
sns.countplot(x='sex', data=df, hue='sex', legend=False)
plt.title('Distribution of Sex')
[253]
```

```
Text(0.5, 1.0, 'Distribution of Sex')
```

2. SMOKER Distribution:

   A countplot is also used to analyze the SMOKER variable to understand the prevalence of smoking in the dataset.



```
#CATEGORICAL VARIABLES - SMOKER
df['smoker'].describe()
[254]
```

4 rows Length: 4, dtype: object

```
          smoker
count     1348
unique    3
top       no
freq      1064
```

```
df['smoker'].value_counts()
[255]
```

3 rows Length: 3, dtype: int64

```
smoker        count
no             1064
yes             274
unknown          10
```

```
df['smoker'].value_counts(normalize=True)
[256]
```

3 rows Length: 3, dtype: float64

```
smoker        proportion
no              0.789318
yes             0.203264
unknown         0.007418
```

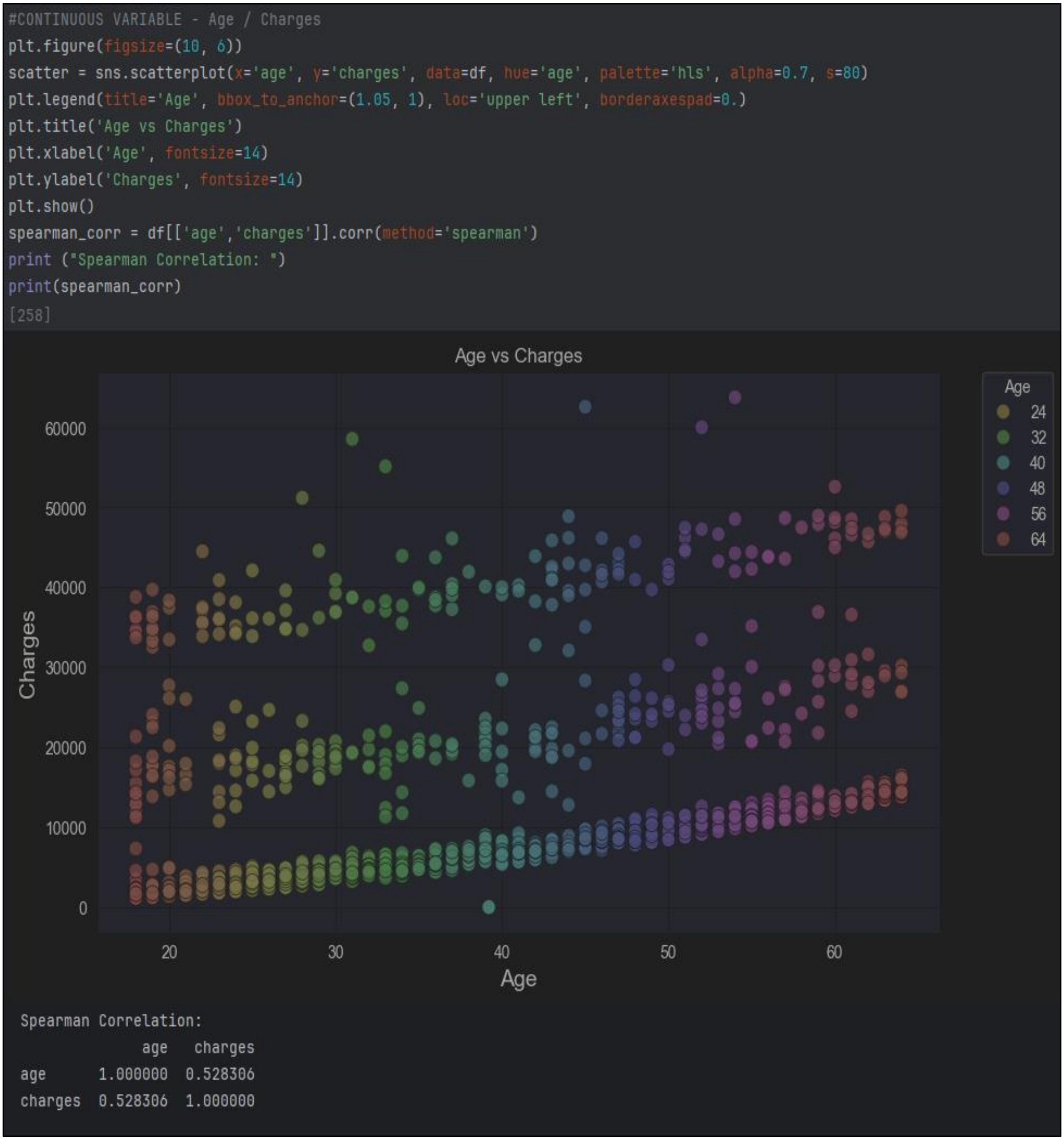
```
# Distribution of SMOKER with specified colors
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
sns.countplot(x='smoker', data=df, hue='smoker', legend=False)
plt.title('Distribution of Smoker')
[257]
```

```
Text(0.5, 1.0, 'Distribution of Smoker')
```

B. Bivariate Analysis of Continuous and Categorical Variables

1. Continuous vs Continuous: AGE and CHARGES

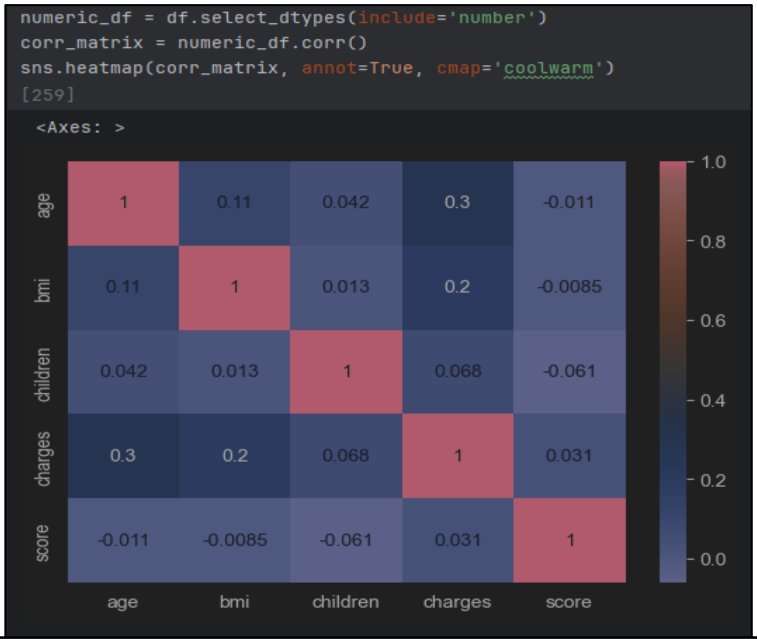
   A scatterplot is used to visualize the relationship between AGE and CHARGES, with color representing AGE to detect any trend. The ‘Spearman Correlation’ is used to quantify the relationship between the two variables, revealing a positive correlation.



```
#CONTINUOUS VARIABLE - Age / Charges
plt.figure(figsize=(10, 6))
scatter = sns.scatterplot(x='age', y='charges', data=df, hue='age', palette='hls', alpha=0.7, s=80)
plt.legend(title='Age', bbox_to_anchor=(1.05, 1), loc='upper left', borderaxespad=0.)
plt.title('Age vs Charges')
plt.xlabel('Age', fontsize=14)
plt.ylabel('Charges', fontsize=14)
plt.show()
spearman_corr = df[['age','charges']].corr(method='spearman')
print ("Spearman Correlation: ")
print(spearman_corr)
[258]
```

```
Spearman Correlation:
              age   charges
age      1.000000  0.528306
charges  0.528306  1.000000
```

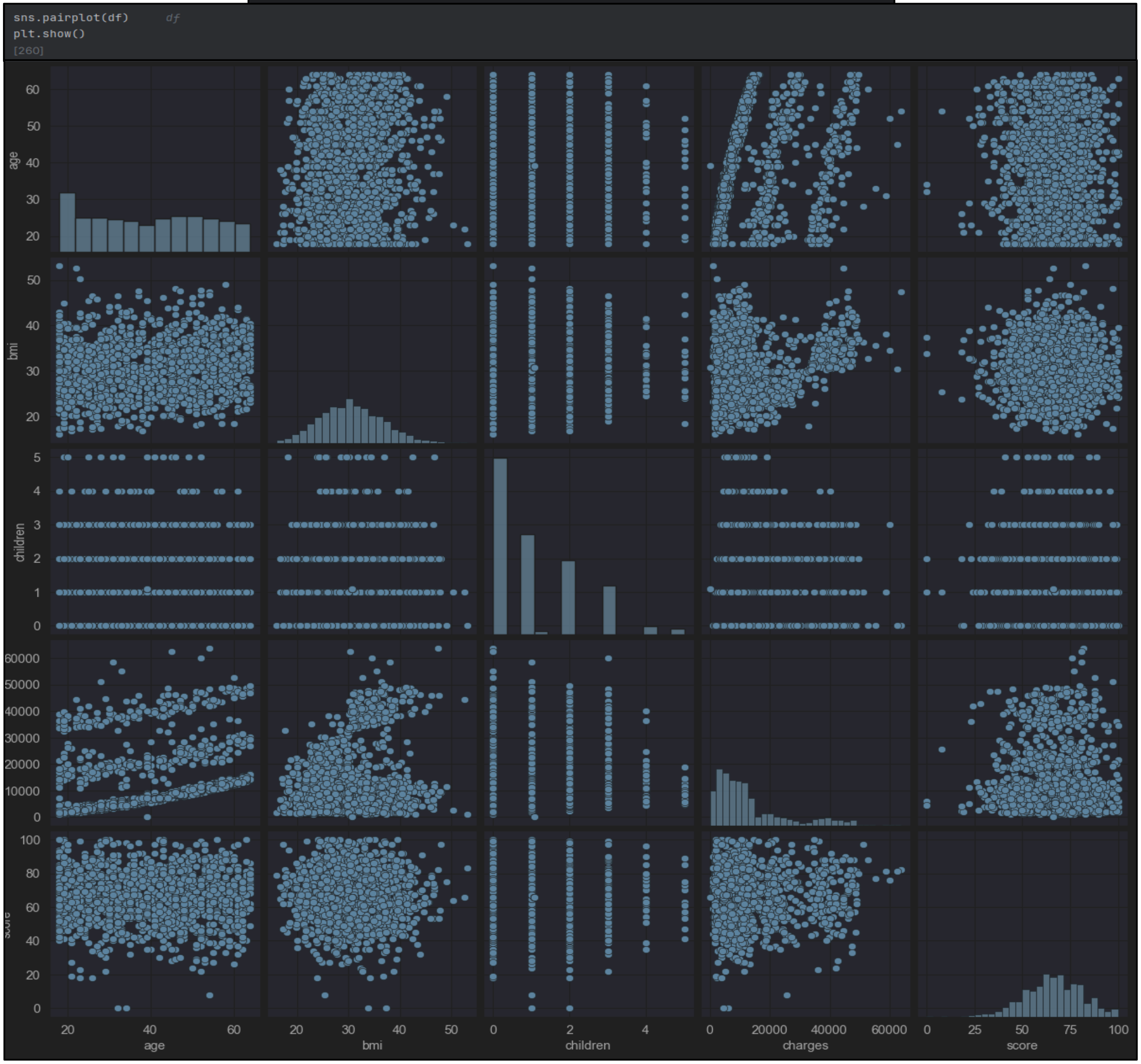
(Other looks at Bivariate Statistics among Continuous Variables)



```
numeric_df = df.select_dtypes(include='number')
corr_matrix = numeric_df.corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
[259]
```

```
<Axes: >
```

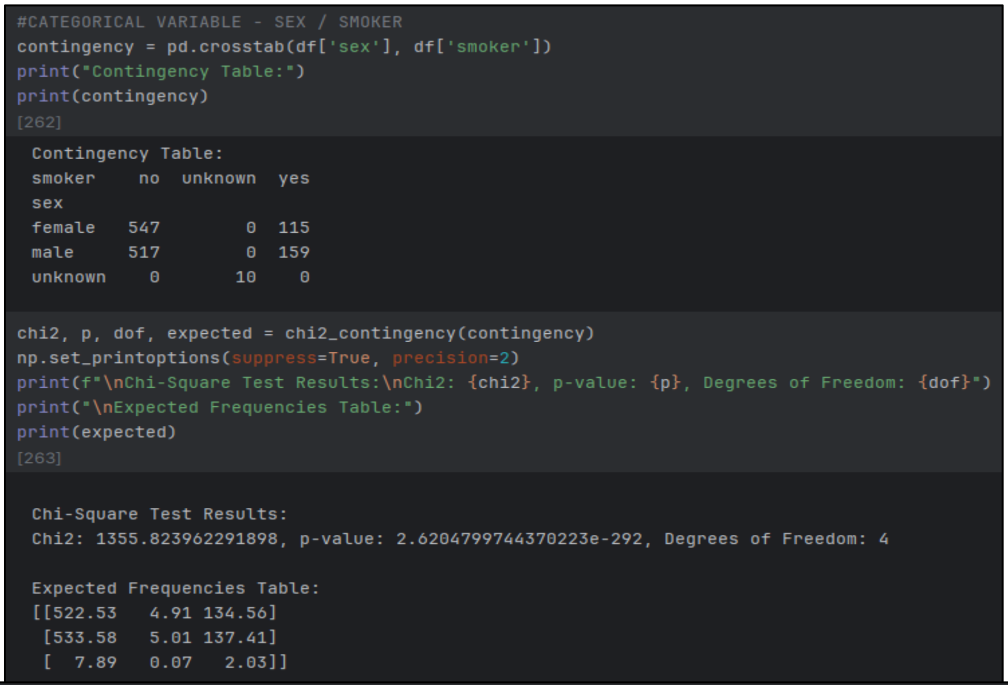
| | age | bmi | children | charges | score |
|---|---|---|---|---|---|
| age | 1 | 0.11 | 0.042 | 0.3 | -0.011 |
| bmi | 0.11 | 1 | 0.013 | 0.2 | -0.0085 |
| children | 0.042 | 0.013 | 1 | 0.068 | -0.061 |
| charges | 0.3 | 0.2 | 0.068 | 1 | 0.031 |
| score | -0.011 | -0.0085 | -0.061 | 0.031 | 1 |



```
sns.pairplot(df)    df
plt.show()
[260]
```

2. Categorical vs Categorical: SEX and SMOKER

   A contingency table and Chi-square test are used to analyze the relationship between SEX and SMOKER. The results, along with a heatmap, show whether there is a significant association between the two categorical variables.



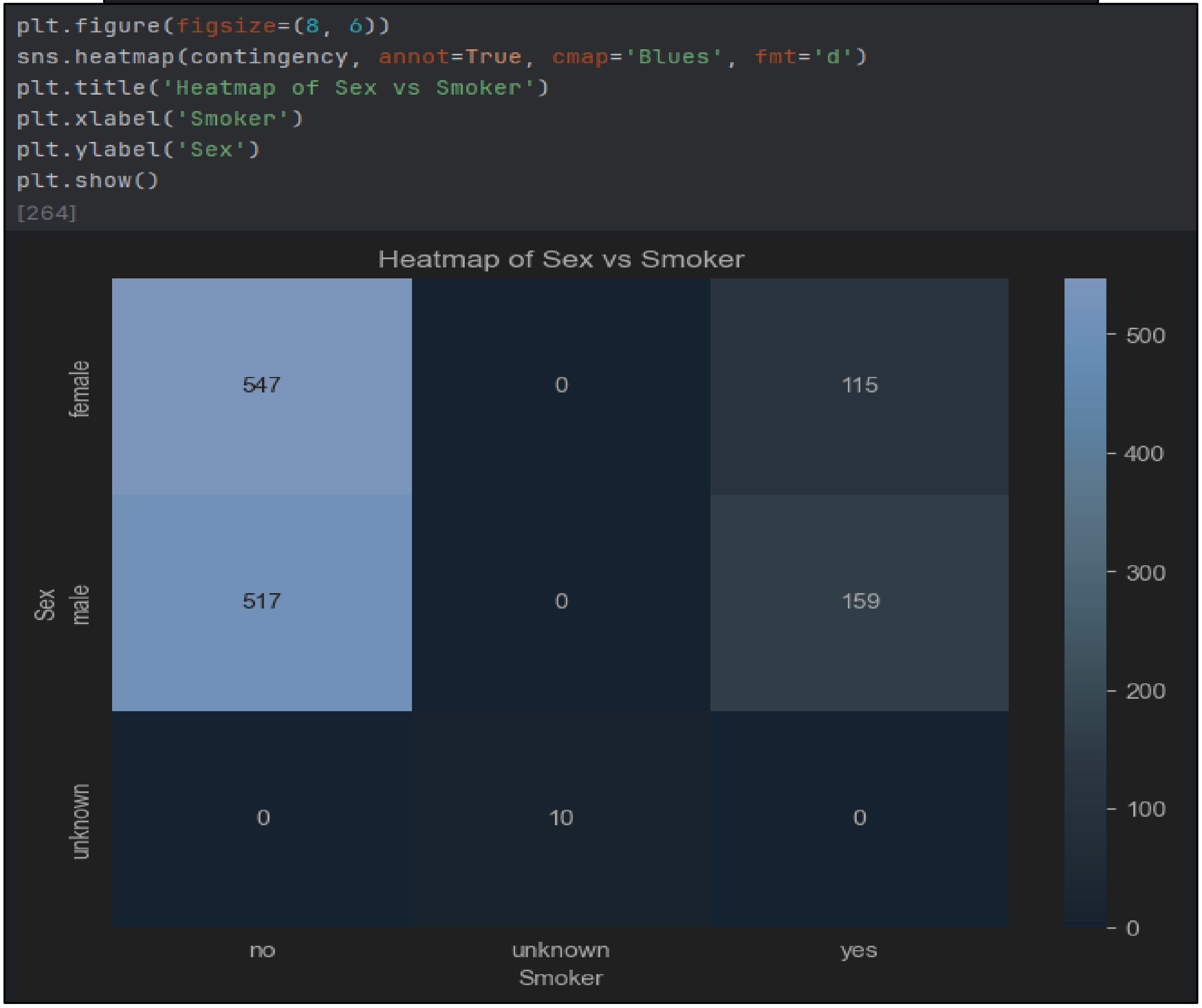
```
#CATEGORICAL VARIABLE - SEX / SMOKER
contingency = pd.crosstab(df['sex'], df['smoker'])
print("Contingency Table:")
print(contingency)
[262]
```

```
Contingency Table:
smoker    no  unknown  yes
sex
female   547        0  115
male     517        0  159
unknown    0       10    0
```

```
chi2, p, dof, expected = chi2_contingency(contingency)
np.set_printoptions(suppress=True, precision=2)
print(f"\nChi-Square Test Results:\nChi2: {chi2}, p-value: {p}, Degrees of Freedom: {dof}")
print("\nExpected Frequencies Table:")
print(expected)
[263]
```

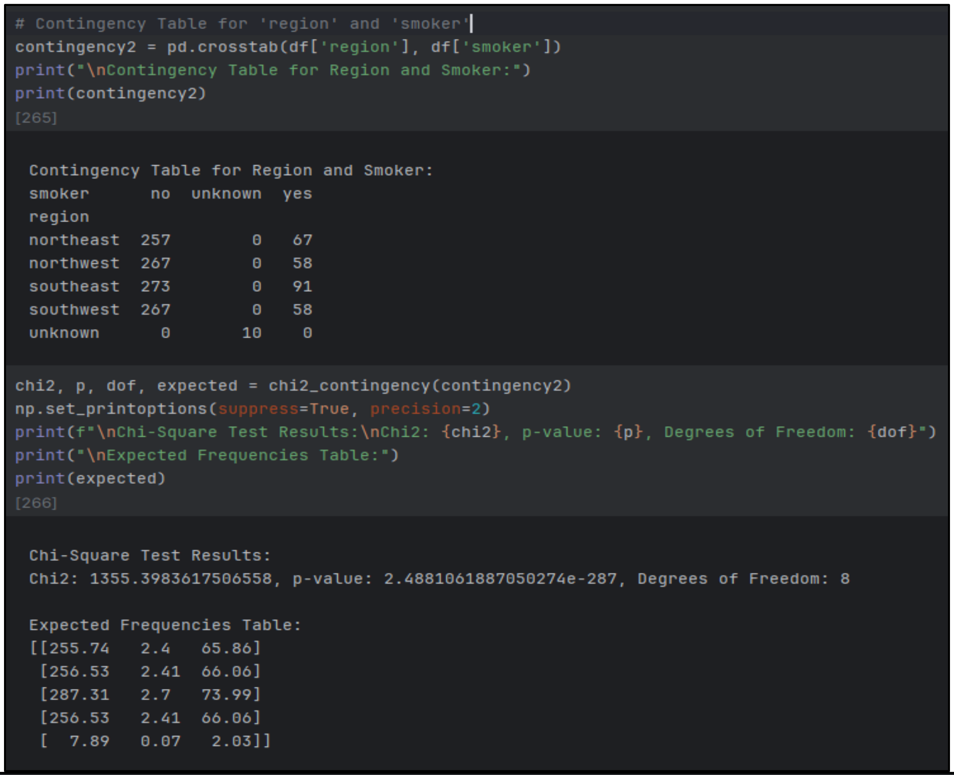
```
Chi-Square Test Results:
Chi2: 1355.823962291898, p-value: 2.6204799744370223e-292, Degrees of Freedom: 4

Expected Frequencies Table:
[[522.53   4.91 134.56]
 [533.58   5.01 137.41]
 [  7.89   0.07   2.03]]
```



```
plt.figure(figsize=(8, 6))
sns.heatmap(contingency, annot=True, cmap='Blues', fmt='d')
plt.title('Heatmap of Sex vs Smoker')
plt.xlabel('Smoker')
plt.ylabel('Sex')
plt.show()
[264]
```

(Another Look at Bivariate Statistics between two Categorical Variables – REGION & SMOKER)



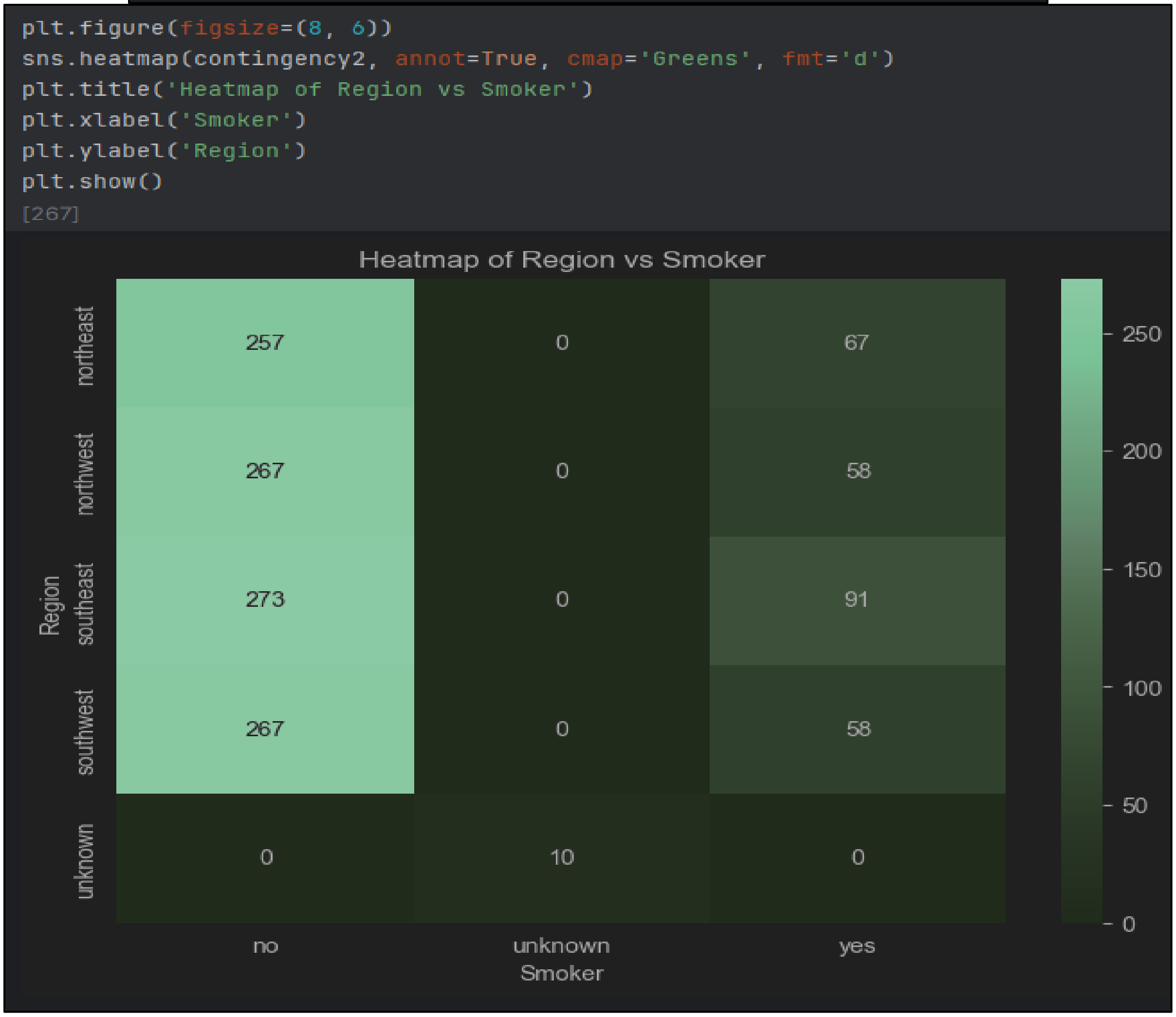
```
# Contingency Table for 'region' and 'smoker'
contingency2 = pd.crosstab(df['region'], df['smoker'])
print("\nContingency Table for Region and Smoker:")
print(contingency2)
[265]
```

```
Contingency Table for Region and Smoker:
smoker      no  unknown  yes
region
northeast  257        0   67
northwest  267        0   58
southeast  273        0   91
southwest  267        0   58
unknown      0       10    0
```

```
chi2, p, dof, expected = chi2_contingency(contingency2)
np.set_printoptions(suppress=True, precision=2)
print(f"\nChi-Square Test Results:\nChi2: {chi2}, p-value: {p}, Degrees of Freedom: {dof}")
print("\nExpected Frequencies Table:")
print(expected)
[266]
```

```
Chi-Square Test Results:
Chi2: 1355.3983617506558, p-value: 2.4881061887050274e-287, Degrees of Freedom: 8

Expected Frequencies Table:
[[255.74   2.4   65.86]
 [256.53   2.41  66.06]
 [287.31   2.7   73.99]
 [256.53   2.41  66.06]
 [  7.89   0.07   2.03]]
```



```
plt.figure(figsize=(8, 6))
sns.heatmap(contingency2, annot=True, cmap='Greens', fmt='d')
plt.title('Heatmap of Region vs Smoker')
plt.xlabel('Smoker')
plt.ylabel('Region')
plt.show()
[267]
```

## Part 2: Parametric Statistical Testing

### C. Research Question

“Does smoking status influence BMI ?”

Relevant Variables:

Smokers: A categorical variable indicating whether the individual is a smoker (yes/no)

BMI: A continuous variable representing the body mass index of an individual.

### D. Parametric Test

The independent t-test is chosen to compare the DMI between two independent groups: smokers and non-smokers. It is a parametric test suitable for comparing the means of two groups when the dependent variable (BMI) is continuous.

The null and alternative hypotheses are:

Null Hypothesis: Smoking has no effect on BMI. There is no significant difference in BMI between smokers and non-smokers.

Alternative Hypothesis: Smoking has an effect on BMI. There is a significant difference in BMI between smokers and non-smokers.

### E. Evaluation of Parametric Test

Justification:

The independent t-test was selected because it compares the means of two independent groups: smokers and non-smokers. The dependent variable, BMI, is continuous, making it suitable for parametric testing. The test assumes that the BMI data for each group follows a normal distribution, and the sample sizes are large enough for the t-test to be robust to slight deviations from normality.

Results:

T-statistic: 0.1335
P-value: 0.8938

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

Stakeholder Benefit:

While the analysis indicates no significant difference in BMI between smokers and non-smokers, this finding can still inform wellness and health initiatives. Knowing that BMI does not significantly differ by smoking status allows stakeholders to focus on other factors, such as diet or physical activity when addressing BMI-related health concerns. Further analysis incorporating additional variables may provide a more comprehensive understanding of BMI determinants.

F. Summarization of the Implications of the Parametric Test

Answer to the question:

Based on the results of the independent t-test, smoking status does not significantly affect BMI. There is no meaningful difference in BMI between smokers and non-smokers in this dataset.

Limitations of the Data Analysis:

Assumptions of Normality: Although the t-test is robust to minor deviations for normality, the assumption of normality in BMI data was checked.

Potential Confounding Variables:

Other factors, such as age, gender, physical activity level, diet, or pre-existing conditions may influence BMI and were not considered in this analysis, Including these variables would help strengthen the findings and provide a more comprehensive view.

Recommendations:

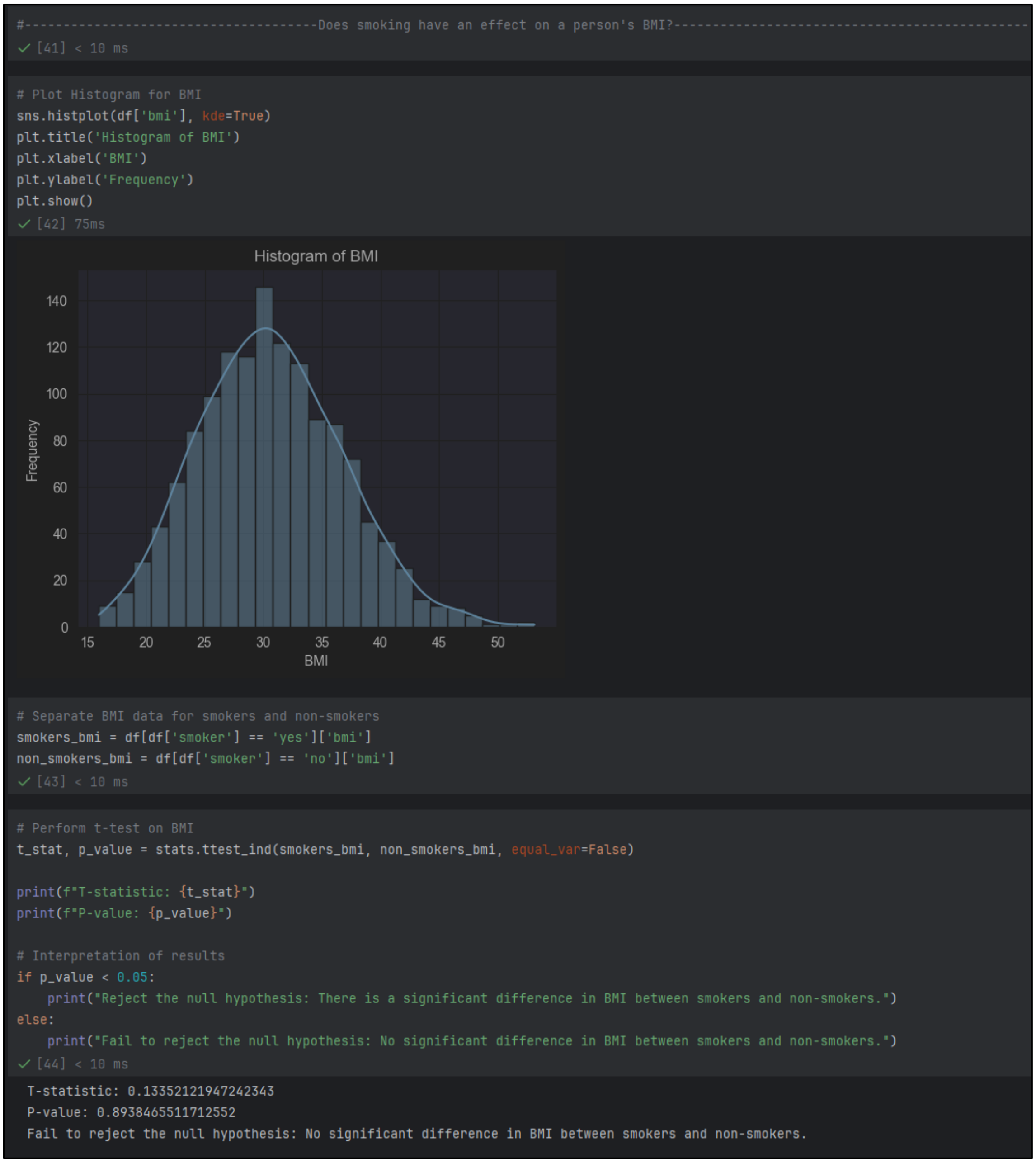
Based on the findings, it is recommended that the organization consider adjusting premiums for smokers to reflect the higher medical costs associated with their lifestyles. Additionally, implementing smoking cessation programs could help reduce overall healthcare costs in the long run. Targeted health interventions for smokers may also help reduce the burden of smoking-related diseases.

Recommendations:

Consider analyzing additional variables that may impact BMI, such as physical activity or diet.

Conduct further research to explore other factors influencing BMI to better guide wellness programs.

If the goal is to understand health impacts associated with smoking, focus on other health outcomes (such as medical charges, and incidence of chronic diseases) that may show a clearer relationship with smoking.



```
#--------------------------------------Does smoking have an effect on a person's BMI?---------------------------------------------
✓ [41] < 10 ms
```

```
# Plot Histogram for BMI
sns.histplot(df['bmi'], kde=True)
plt.title('Histogram of BMI')
plt.xlabel('BMI')
plt.ylabel('Frequency')
plt.show()
✓ [42] 75ms
```

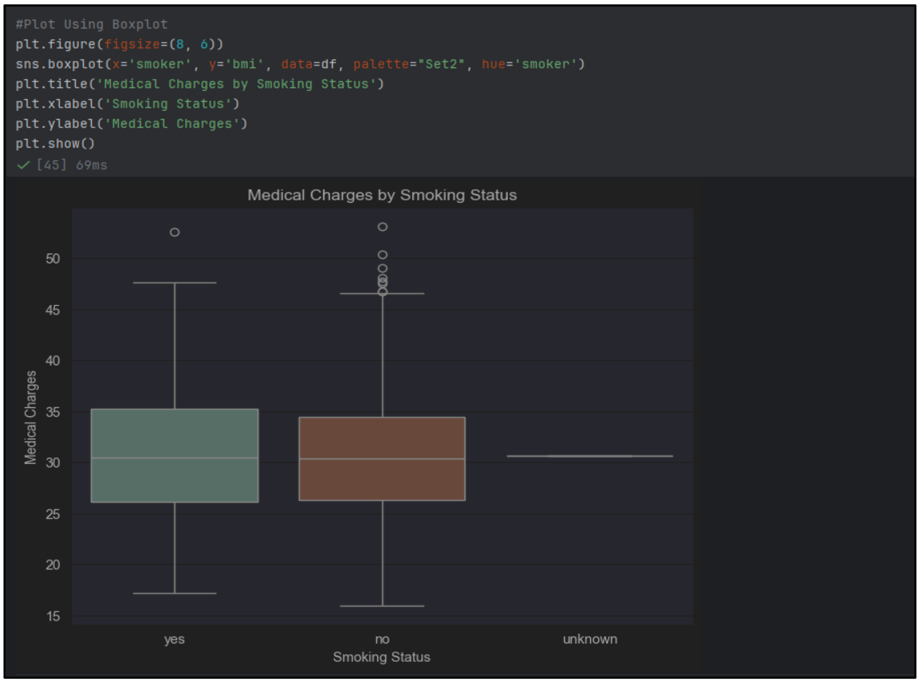
```
# Separate BMI data for smokers and non-smokers
smokers_bmi = df[df['smoker'] == 'yes']['bmi']
non_smokers_bmi = df[df['smoker'] == 'no']['bmi']
✓ [43] < 10 ms
```

```
# Perform t-test on BMI
t_stat, p_value = stats.ttest_ind(smokers_bmi, non_smokers_bmi, equal_var=False)

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

# Interpretation of results
if p_value < 0.05:
    print("Reject the null hypothesis: There is a significant difference in BMI between smokers and non-smokers.")
else:
    print("Fail to reject the null hypothesis: No significant difference in BMI between smokers and non-smokers.")
✓ [44] < 10 ms
```

```
T-statistic: 0.13352121947242343
P-value: 0.8938465511712552
Fail to reject the null hypothesis: No significant difference in BMI between smokers and non-smokers.
```



```
#Plot Using Boxplot
plt.figure(figsize=(8, 6))
sns.boxplot(x='smoker', y='bmi', data=df, palette="Set2", hue='smoker')
plt.title('Medical Charges by Smoking Status')
plt.xlabel('Smoking Status')
plt.ylabel('Medical Charges')
plt.show()
✓ [45] 69ms
```

## Part 3: Nonparametric Statistical Testing

### G. Research Question

“Is there a significant difference between CHARGES based on GENDER and BMI categories?”

Relevant Variables:

SEX: A categorical variable (Male, Female, Unknown)

BMI Category: A categorical variable derived from the BMI variable

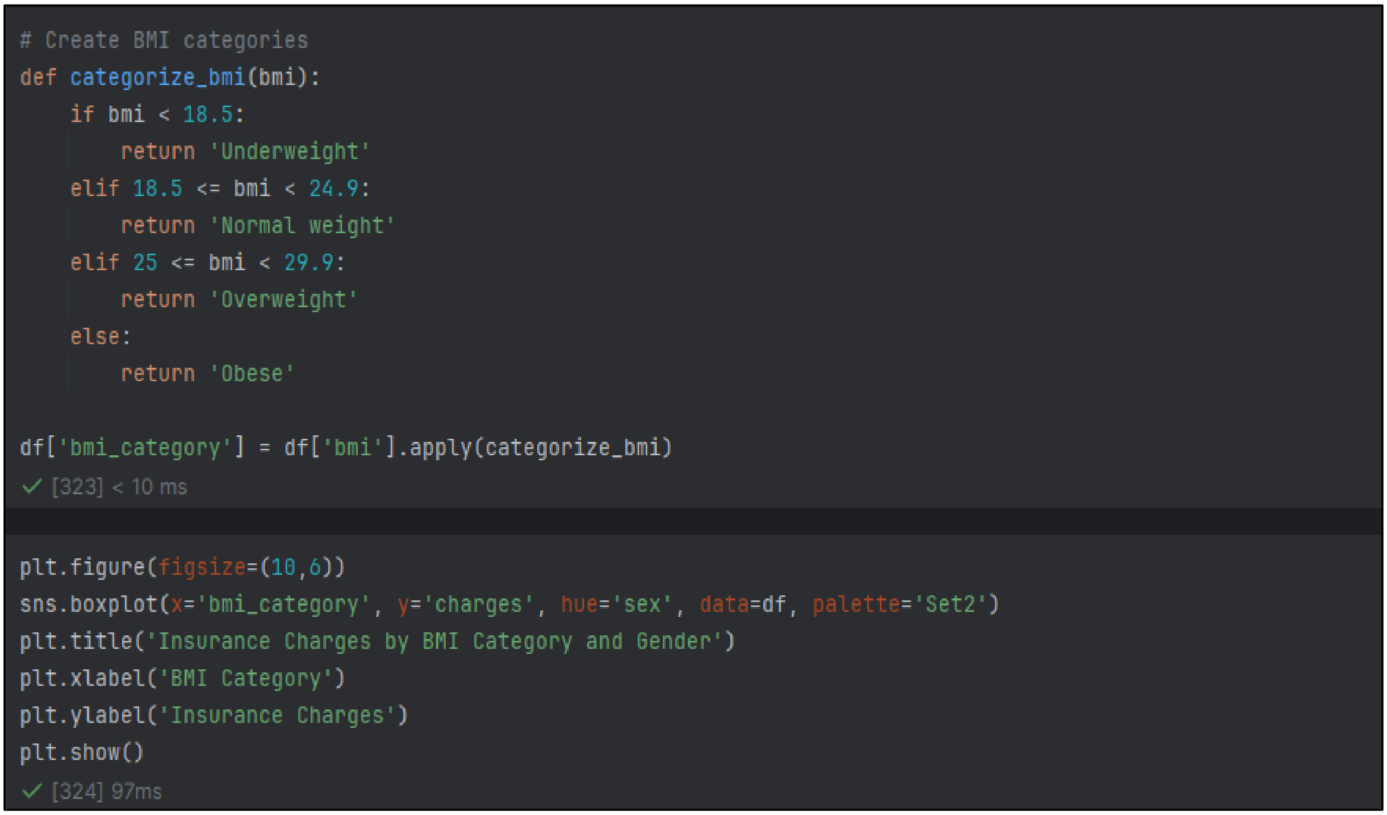
### H. Nonparametric Test

The Kruskal-Wallis test is selected because it is a nonparametric test that compares the medians of multiple independent groups, which is suitable for analyzing differences in insurance charges across BMI categories for males and females.

The null and alternative hypotheses are:

Null Hypothesis: There is no significant difference in insurance charges between the different BMI categories for males and females. That is, BMI categories do not influence charges for either gender.

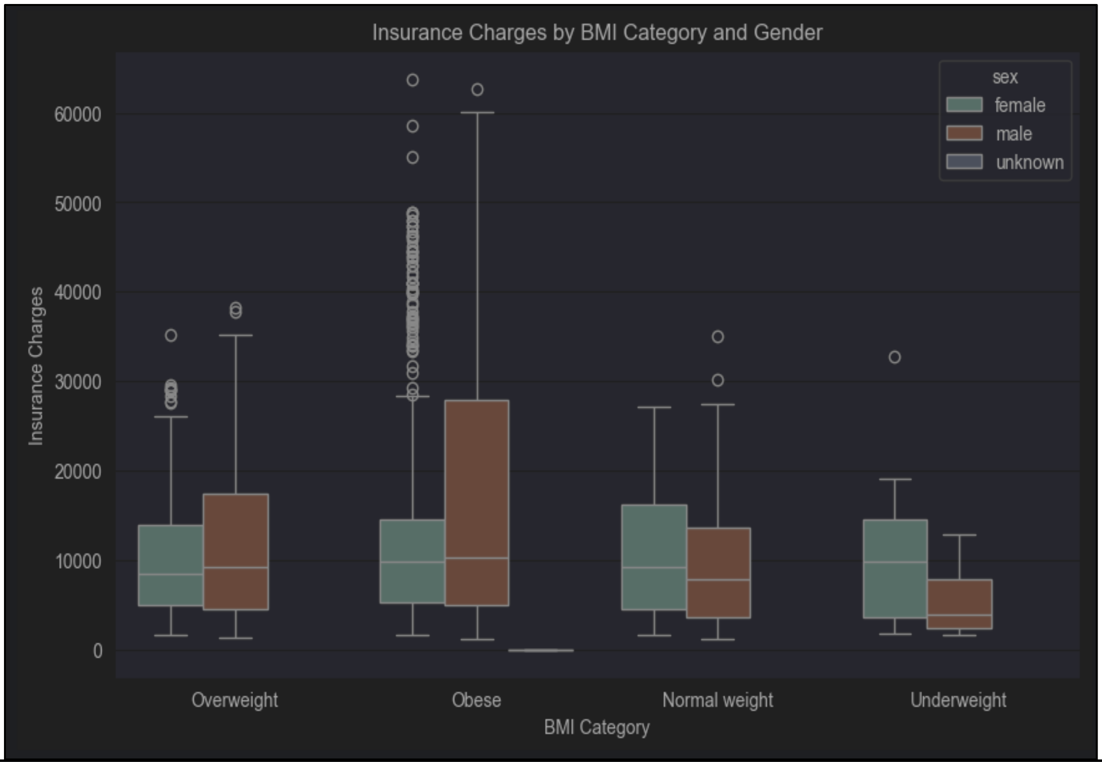
Alternative Hypothesis: There is a significant difference in insurance charges between the different BMI categories for males and females. That is, BMI categories do influence insurance charges for either gender.

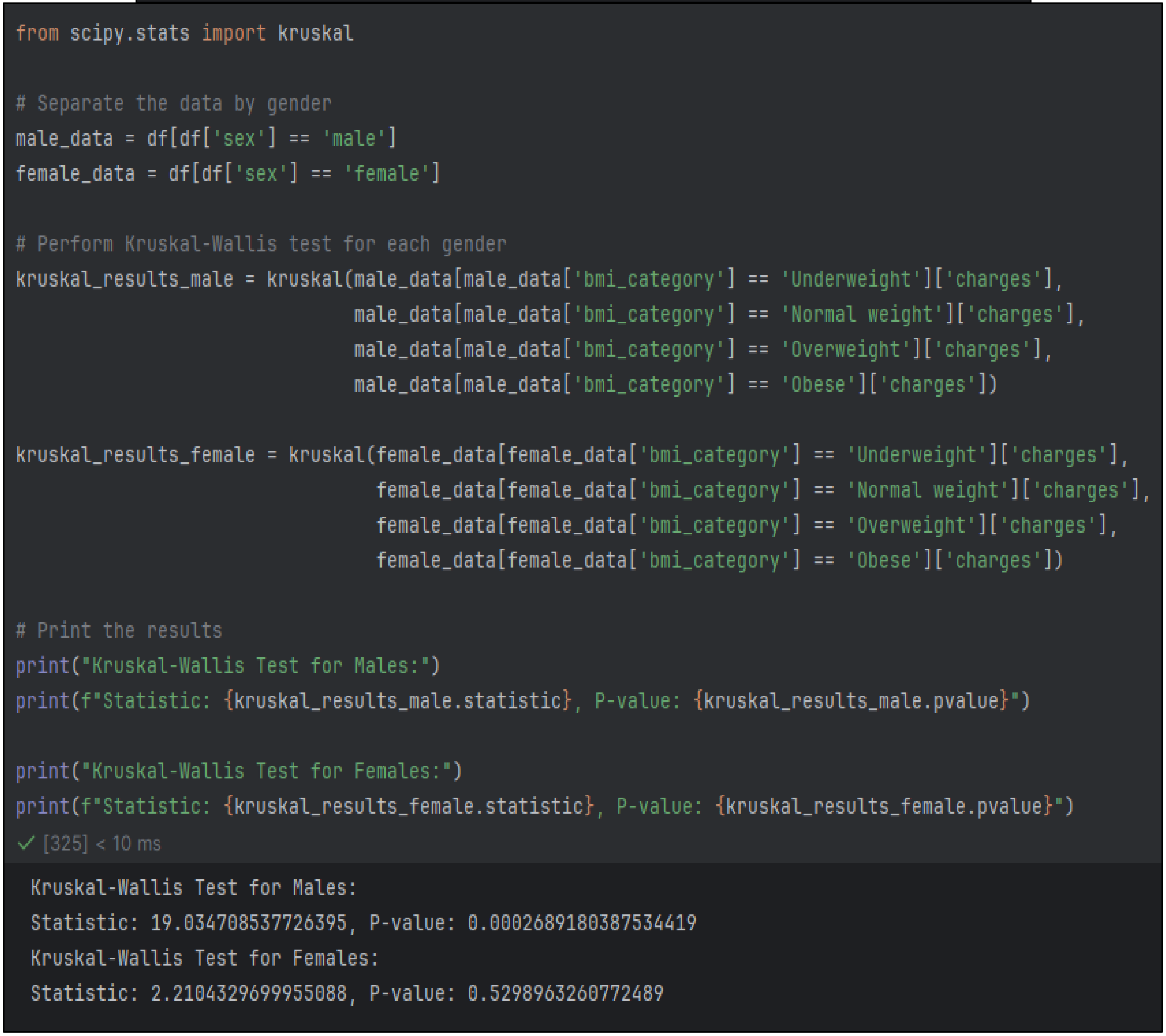


```
# Create BMI categories
def categorize_bmi(bmi):
    if bmi < 18.5:
        return 'Underweight'
    elif 18.5 <= bmi < 24.9:
        return 'Normal weight'
    elif 25 <= bmi < 29.9:
        return 'Overweight'
    else:
        return 'Obese'

df['bmi_category'] = df['bmi'].apply(categorize_bmi)
✓ [323] < 10 ms
```

```
plt.figure(figsize=(10,6))
sns.boxplot(x='bmi_category', y='charges', hue='sex', data=df, palette='Set2')
plt.title('Insurance Charges by BMI Category and Gender')
plt.xlabel('BMI Category')
plt.ylabel('Insurance Charges')
plt.show()
✓ [324] 97ms
```





```python
from scipy.stats import kruskal

# Separate the data by gender
male_data = df[df['sex'] == 'male']
female_data = df[df['sex'] == 'female']

# Perform Kruskal-Wallis test for each gender
kruskal_results_male = kruskal(male_data[male_data['bmi_category'] == 'Underweight']['charges'],
                               male_data[male_data['bmi_category'] == 'Normal weight']['charges'],
                               male_data[male_data['bmi_category'] == 'Overweight']['charges'],
                               male_data[male_data['bmi_category'] == 'Obese']['charges'])

kruskal_results_female = kruskal(female_data[female_data['bmi_category'] == 'Underweight']['charges'],
                                 female_data[female_data['bmi_category'] == 'Normal weight']['charges'],
                                 female_data[female_data['bmi_category'] == 'Overweight']['charges'],
                                 female_data[female_data['bmi_category'] == 'Obese']['charges'])

# Print the results
print("Kruskal-Wallis Test for Males:")
print(f"Statistic: {kruskal_results_male.statistic}, P-value: {kruskal_results_male.pvalue}")

print("Kruskal-Wallis Test for Females:")
print(f"Statistic: {kruskal_results_female.statistic}, P-value: {kruskal_results_female.pvalue}")
```

✓ [325] < 10 ms

```
Kruskal-Wallis Test for Males:
Statistic: 19.034708537726395, P-value: 0.0002689180387534419
Kruskal-Wallis Test for Females:
Statistic: 2.2104329699955088, P-value: 0.5298963260772489
```

Kruskal-Wallis Test Results:

For Males:

Statistic: 19.03

P-value: 0.00027 (significant)

For Females:

Statistic: 2.21

P-value: 0.53 (not significant)

I. Evaluation of Nonparametric Test Results

Test Justification:

The Kruskal-Wallis test was chosen because it does not assume normal distribution and can be used to compare the median charges across multiple BMI categories, making it appropriate for this analysis.

Test Results:

For males, there is a significant difference in insurance charges across BMI categories.

For females, no significant difference was found.

Stakeholder Benefit:

This insight could help insurance companies segment customers more effectively based on gender and BMI, particularly in male customers where BMI categories significantly impact charges.

J. Summary of Nonparametric Statistical Test

Answer to Research Question:

There is a significant difference in insurance charges for males based on BMI categories, but this is not the case for females.

Limitations:

The analysis may not fully capture other important factors affecting charges, such as lifestyle or health conditions.

Recommendation:

Insurance companies should consider gender and BMI in their pricing models, particularly for male customers.