Data Cleaning and Profiling

Task 2

Nathan Hefner

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

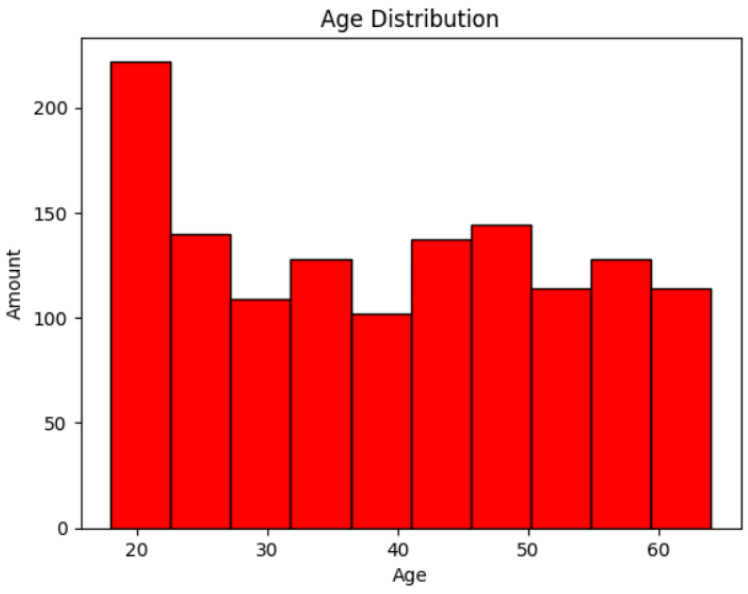
For my continuous variables, I chose Age and BMI. Below are my findings, which are both visualized in graphs as histograms, along with the stats for these two columns.

```
continous_variables = ['age', 'bmi']
stats = df[continous_variables].describe()
print(stats)
```

	age	bmi
count	1338.000000	1338.000000
mean	39.207025	30.663397
std	14.049960	6.098187
min	18.000000	15.960000
25%	27.000000	26.296250
50%	39.000000	30.400000
75%	51.000000	34.693750
max	64.000000	53.130000

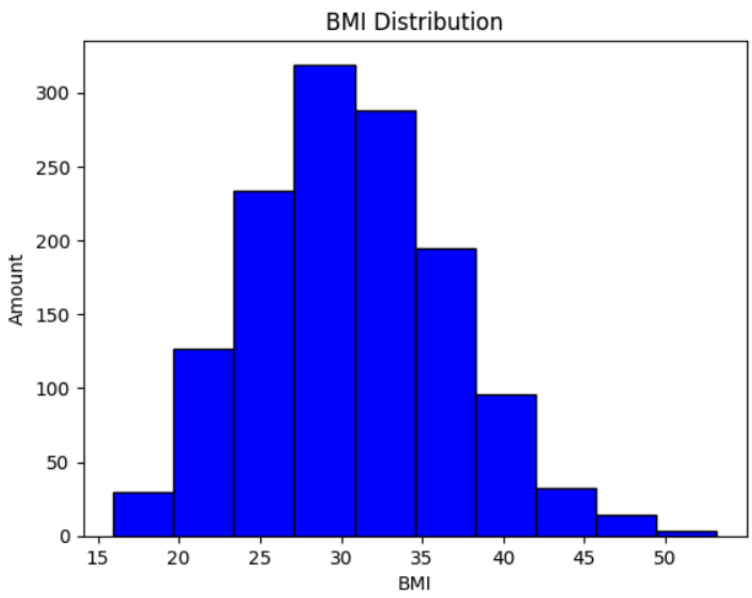
```
plt.hist(df['age'], color = 'red', edgecolor = 'black')
plt.xlabel('Age')
plt.ylabel('Amount')
plt.title('Age Distribution')
```

```
Text(0.5, 1.0, 'Age Distribution')
```




```
plt.hist(df['bmi'], color = 'blue', edgecolor = 'black')
plt.xlabel('BMI')
plt.ylabel('Amount')
plt.title('BMI Distribution')
```

```
Text(0.5, 1.0, 'BMI Distribution')
```



The distribution for age is skewed to the right, with a large majority of people being younger, and the distribution for BMI is slightly more right-skewed but is more evenly divided than Age.

Categorical Variables

For my categorical variables, I chose Sex and Region. Below are the graphs visualized with bar graphs.

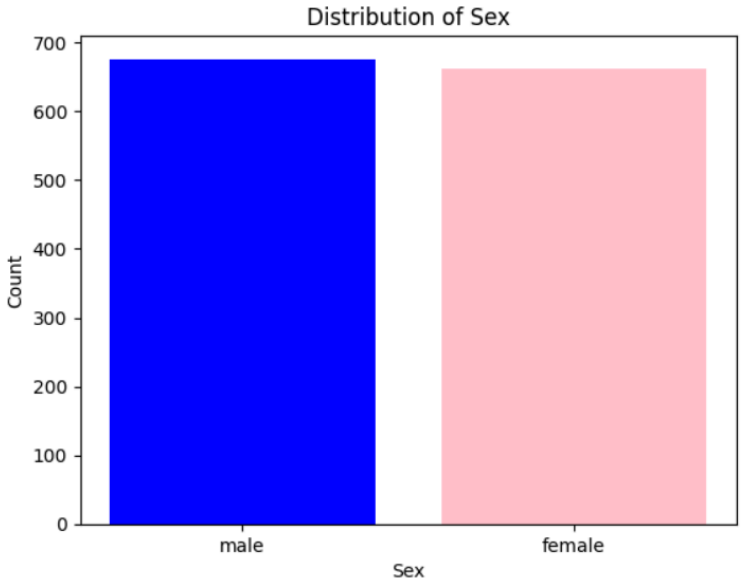
```
sex_counts = df["sex"].value_counts()

colors = ["pink" if sex == "female" else "blue" for sex in sex_counts.index]

plt.bar(sex_counts.index, sex_counts.values, color=colors)

plt.xlabel("Sex")
plt.ylabel("Count")
plt.title("Distribution of Sex")

plt.show()
```



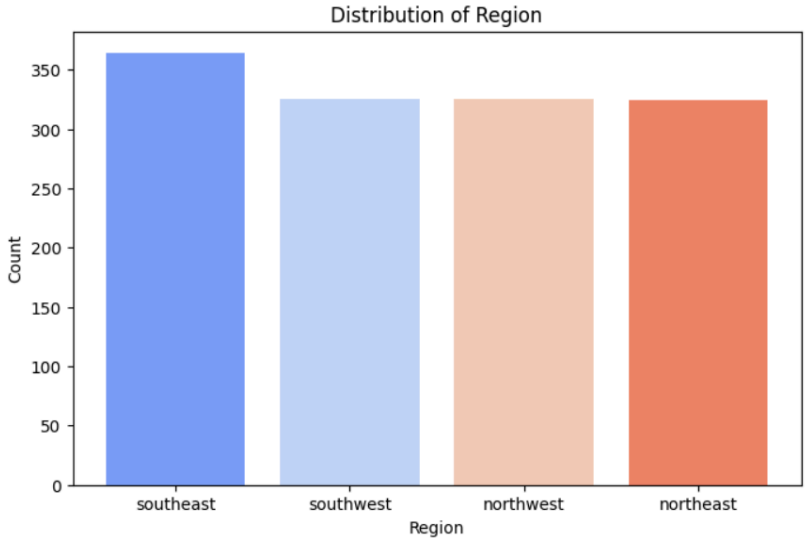

```
region_counts = df["region"].value_counts()

colors = sns.color_palette("coolwarm", len(region_counts))

plt.figure(figsize=(8, 5))
plt.bar(region_counts.index, region_counts.values, color=colors)

plt.xlabel("Region")
plt.ylabel("Count")
plt.title("Distribution of Region")

plt.show()
```



The distribution for sex is relatively even, and the distribution for Region is also even, but the southeast has a slightly higher count than the rest, which is about even as well.

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.

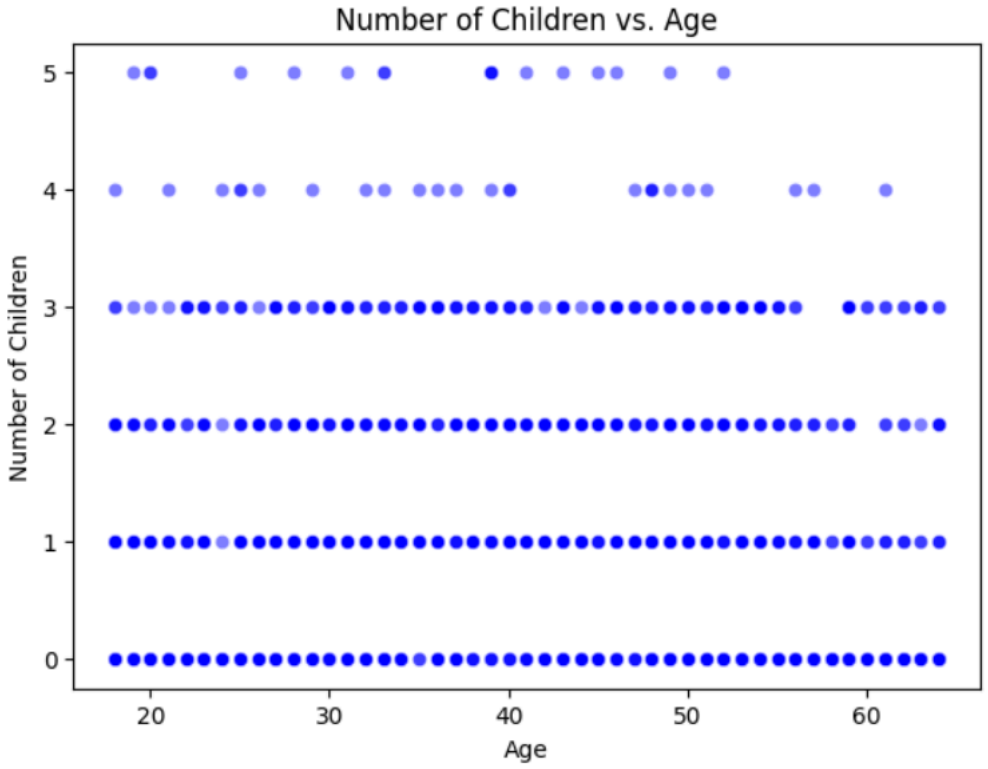
Continuous Variables

My two continuous variables are Age and Children. I will be comparing them using a scatterplot.

```
plt.figure(figsize=(7, 5))
sns.scatterplot(x=df["age"], y=df["children"], color="blue", alpha=0.5)

plt.xlabel("Age")
plt.ylabel("Number of Children")
plt.title("Number of Children vs. Age")

plt.show()
```



The data appears to be non-monotonic, as there is no correlation between these two variables.

Categorical Variables

My two variables are Region and Smoker, visualized as a comparative bar chart.

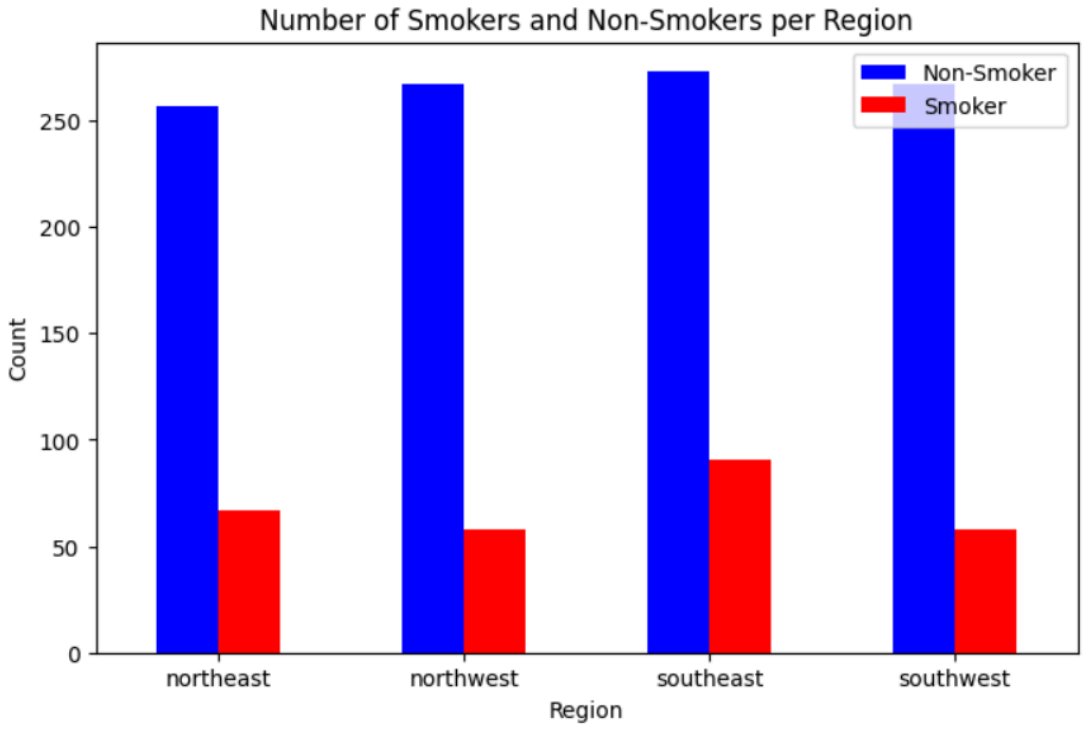
```
smoker_region_counts = df.groupby(["region", "smoker"]).size().unstack()

smoker_region_counts.plot(kind="bar", figsize=(8, 5), color=["blue", "red"])

plt.xlabel("Region")
plt.ylabel("Count")
plt.title("Number of Smokers and Non-Smokers per Region")
plt.legend(labels=["Non-Smoker", "Smoker"])

plt.xticks(rotation=0)

plt.show()
```



This graph shows that the number of non-smokers in the data set outweighs the number of smokers, and the number of smokers and non-smokers per region is relatively the same except for the southeast, which has slightly more smokers than the other regions.

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.

A research question I propose is: Is there a significant difference in mean BMI levels across different regions?

2. Identify the variables in the dataset that are relevant to answering your research question from part C1.

The variables I will be using are BMI and Region.

D. Analyze the dataset by doing the following:

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

ANOVA will be our best test as we are comparing means across multiple groups.

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

Null Hypothesis - There is no significant difference between the means of the different regions.

Alternative Hypothesis - There is a significant difference between the means of the different regions.

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

```
import scipy.stats as stats

df_clean = df.dropna(subset=["region", "bmi"])

groups = [df_clean["bmi"][df_clean["region"] == category] for category in df_clean["region"].unique()]
anova_result = stats.f_oneway(*groups)

print("ANOVA F-statistic:", anova_result.statistic)
print("ANOVA p-value:", anova_result.pvalue)

if anova_result.pvalue < 0.05:
    print("Alternative Hypothesis")
else:
    print("Null Hypothesis")
```

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

Anova F-statistic: 39.49505720170283

ANOVA p-value: 1.881838913929143e-24

Since the ANOVA p-value is far below 0.05 there is a significant difference between the means across all the regions.

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 ANOVA test because we want to compare the mean of BMI values across the four different regions. ANOVA excels at testing the mean across various groups, which is perfect for our situation.

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

The ANOVA p-value was about 2e-24, which is far below the statistical significance amount of p=0.05. This means we can reject the null hypothesis and accept the alternative hypothesis, which is that there is a significant difference between the means across the four regions. This will then help us conclude that the region you are from may play a pivotal role in what your BMI is.

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

Stakeholders benefit from this choice of testing because we can compare the four different regions' BMI means at once, which can help pinpoint where BMI is at its highest or lowest. This can then help calculate the new risk of implementing the insurance in certain regions if it is found that having a certain BMI level incurs more or less charges.

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

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

The answer to my previously stated question is yes, there is a significant difference in BMI levels across the four different regions.

2. Discuss the limitations of your data analysis.

We should consider a few limitations: parametric testing requires normal distribution, but the BMI scores were slightly skewed right, meaning there may be overexposure to lower BMI scores. Also, we did not consider other variables in our analysis, nor did we check for an even amount of sex in the data. We also did not test if smoking played a factor in BMI level or if the number of children did. Not checking these variables could mean our data analysis isn't telling the whole story, and therefore we cannot come to an entirely conclusive answer.

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

We have a significant lead in the question of Region affecting BMI levels. To further prove our theory, we need to test different variables further and to see if we get a null hypothesis or alternative hypothesis, along with making sure there is an even split of sex, to not skew our data the wrong way. If we test other variables and find there is no significant difference in BMI levels we can then safely conclude that region does play a part in overall BMI levels for the patients that live there. After that, we could calculate how certain BMI levels affect accrued charges. The insurance company can then calculate a new risk/reward theorem and decide on how they want to operate in each region, whether it be different because of higher/lower BMI levels or maintain their current operations.

Non-Parametric 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.

My research question is: Do smokers have a higher overall charge amount than nonsmokers? This question can help us learn and identify what types of health habits cost the most for the insurance agencies.

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

The two variables I will be using are Charges and 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.

The test I will use is the Mann-Whitney U Test, which is the best when comparing two independent variables.

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

Null Hypothesis - There is no correlation between the amount charged between smokers and non-smokers.

Alternative Hypothesis - There is a correlation between the amount charged between smokers and non-smokers.

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

```
from scipy.stats import mannwhitneyu

smoker = df[df['smoker'] == 'no']['charges']
non_smoker = df[df['smoker'] == 'yes']['charges']

smoker = pd.to_numeric(smoker, errors='coerce')
non_smoker = pd.to_numeric(non_smoker, errors='coerce')

smoker = smoker.dropna()
non_smoker = non_smoker.dropna()

stat, p_value = mannwhitneyu(smoker, non_smoker)

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

if p_value < 0.05:
    print("Alternative Hypothesis")
else:
    print("Null Hypothesis")
```

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

P-Value - 5.270233444503571e-130

Mann-Whitney U Test Statistic: 740.3

Since the P-Value is far below 0.05, there is a correlation between the amount charged between smokers and non-smokers.

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 our two variables are numeric and categorical, and this specific test is best when comparing the differences between two independent groups, regardless of whether they're numeric or categorical. The Chi-Square Test is meant for two categorical variables, and the Kruskal-Wallis Test is an alternative to ANOVA, which compares means across different groups, neither of which we are dealing with at this time.

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

The P-Value is approximately 5e-130, which is far below the P-Value significant value of 0.05, which means the Alternative Hypothesis is not rejected. Henceforth, we find that there is a correlation between the amount charged between Smokers and Non-Smokers.

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

Stakeholders benefit from my choice of testing method because we can accurately and confidently say that there is a different amount of charges between Smokers and Non-Smokers. We can now use this information when finding new potential customers to calculate their risk level, depending on how smoking affects charges.

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

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

The answer to my question from earlier is yes; there is a strong correlation between the difference in the amount charged by Smokers and Non-Smokers.

2. Discuss the limitations of your data analysis.

A limitation of my data analysis is that we do not know what the difference of charges is between Smokers and Non-Smokers. It could be either possibility either smokers incur low charges and non-smokers incur high charges, or vice versa. We also don't know if smoking is the only variable that affects charges, there could be other reasons such as age, BMI, or amount of changes. Without these crucial pieces of information, we would not know the full story.

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

My recommendation is with this strong lead into different charge amounts, I would first see how the charges are different, and figure out if smokers have higher or lower charges. Then, I would test other variables to pinpoint what exactly is causing high charges. This could then help calculate risk level of potential new clients.

Sources

No sources were used except for WGU Material