1.2 Exercise. DSC630 - Jennifer Barrera Conde

June 4, 2024

1 DSC630

2 1.2 Exercise

3 Jennifer Barrera Conde

3.1 The dataset:

Was collected from Kaggle and is called the “Adult Income dataset”. https://www.kaggle.com/datasets/rabailanees/adult-income-dataset According to the description: The “Adult Income Dataset” contains demographic and employment details for individuals. Attributes include age, work class, education, marital status, occupation, race, gender, work hours, and country of origin. Each record indicates whether the individual’s income exceeds $50K per year. This dataset is used for income prediction and demographic analysis.

I made some changes to the raw excel file to make handling of information easier.

3.1.1 First Question:

Question 1: Does experience lead to higher income? First I need to get more acquainted with my data

[1]:

```
import pandas as pd

# Load the Excel file
file_path = 'adult-income.xlsx'
df = pd.read_excel(file_path)

# Display the first few rows of the dataframe to understand its structure
df.head()
```

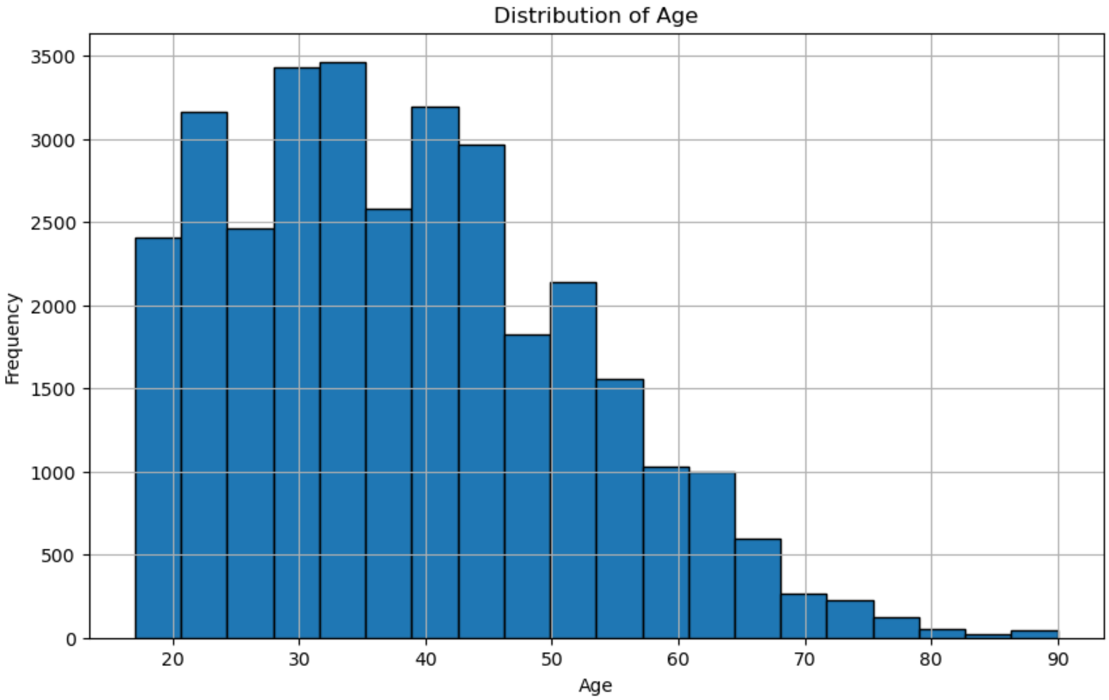
[1]:

```
   Age    Employment type  Yearly Income  Education       Marital status  \
0   39          State-gov          77516  Bachelors        Never-married
1   50   Self-emp-not-inc          83311  Bachelors   Married-civ-spouse
2   38            Private         215646    HS-grad             Divorced
3   53            Private         234721       11th   Married-civ-spouse
4   28            Private         338409  Bachelors   Married-civ-spouse

          Occupation    Race   Gender  Work Hours Weekly Country of origin
```

```
0        Adm-clerical   White     Male                 40     United-States
1     Exec-managerial   White     Male                 13     United-States
2   Handlers-cleaners   White     Male                 40     United-States
3   Handlers-cleaners   Black     Male                 40     United-States
4      Prof-specialty   Black   Female                 40              Cuba
```

Use a bar graph to visualize the age distribution from the data

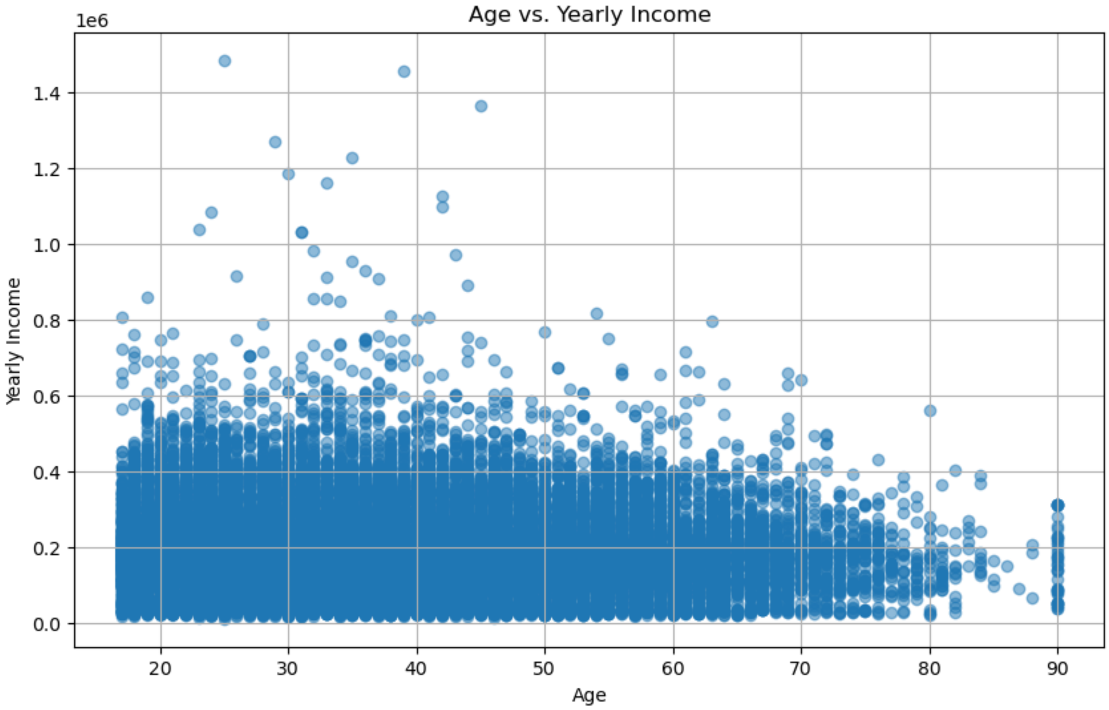
[3]:
```python
import matplotlib.pyplot as plt

# Create a bar graph of Age
plt.figure(figsize=(10, 6))
plt.hist(df['Age'], bins=20, edgecolor='black')
plt.title('Distribution of Age')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()
```



I did not know what a bivariate plot is, from my findings it is basically a scater plot, I decided to use this to find a relation between age and Yearly income, assuming that age may count as “experience”, this could answer my question.

[2]:
```python
import matplotlib.pyplot as plt
```

```
# Scatter plot of Age vs. Yearly Income
plt.figure(figsize=(10, 6))
plt.scatter(df['Age'], df['Yearly Income'], alpha=0.5)
plt.title('Age vs. Yearly Income')
plt.xlabel('Age')
plt.ylabel('Yearly Income')
plt.grid(True)
plt.show()
```



Conclusion: There seems to actually be a decrease in income that becomes more prominent after the age reaches 50. There is also a small increase in income that occurs around late 20's to the mid 40's.

Leading me to believe that there is some increase of income around those ages, but there is so much data and it is so much, without groups and sections to see if it is based on occupation or such that could be another factor as to why an income may increase or decrease.

[ ]:

```
```

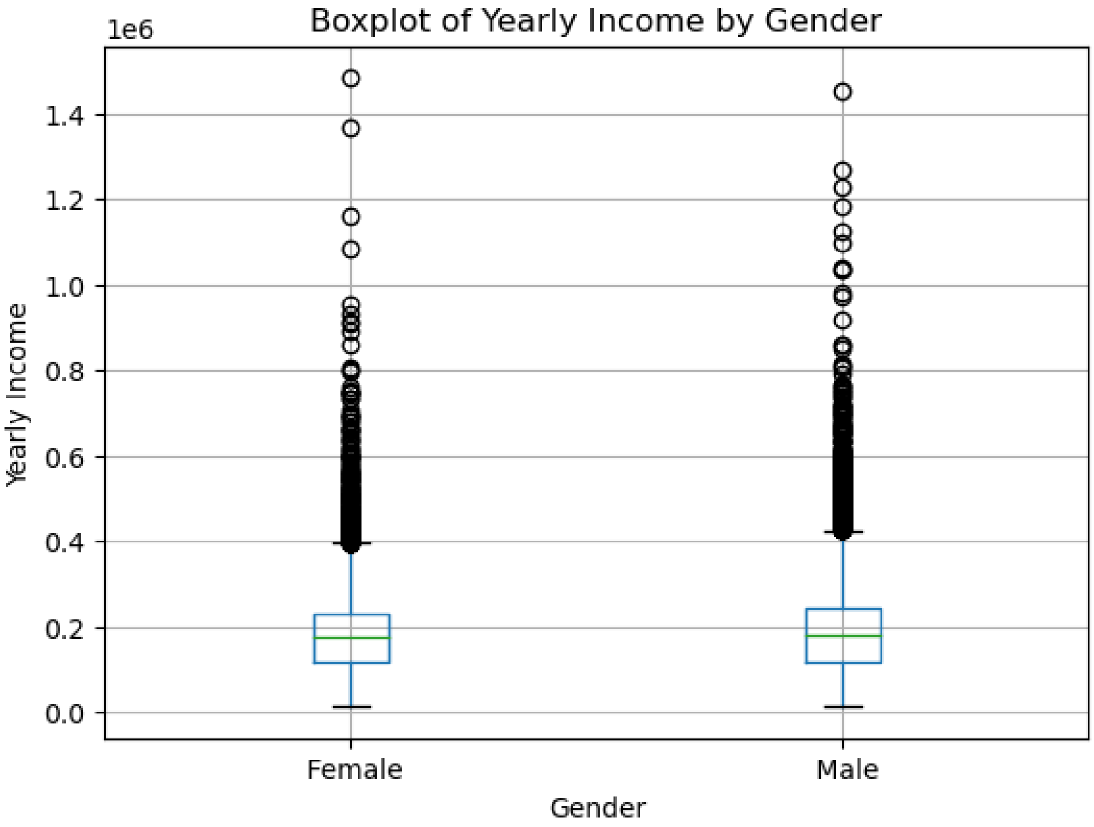
3.1.2 Second Question:

I would like to create a comparison of Yearly income based on Gender and have it side by side to find any discrepancies.

[4]:

```
import matplotlib.pyplot as plt

# Create a boxplot for Gender and Yearly Income
plt.figure(figsize=(10, 6))
df.boxplot(column='Yearly Income', by='Gender')
plt.title('Boxplot of Yearly Income by Gender')
plt.suptitle('')
plt.xlabel('Gender')
plt.ylabel('Yearly Income')
plt.grid(True)
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```



The results tell me that although the highest yearly income goes to a woman, there are a lot more spread in the higher paying jobs for men which leads to the idea that man may have the most higher paying jobs. Could there be an uneven number of participants which leads to skewed results?

[5]:

```
# Count the number of males and females in the dataset
gender_counts = df['Gender'].value_counts()
print(gender_counts)
```

```
Gender
 Male      21790
 Female    10771
Name: count, dtype: int64
```

Conclusion: There are double the participants from male to female, so one third of participants were female. From this I could assume or leverage the results with the idea that if there had been an even number of participants may be the results would hav been more even, or maybe there would have been a lot more females making a higher yearly income than the male counterpart.

[ ]:

```
```