T. Mohamed Nadhim
3122 21 3002 058
ECE A 2025

Ex 1: Data Visualization

In this experiment we will be visualizing a dataset in 11 different ways.

The dataset used for this experiment is: “Adult Income Dataset”

Dataset Description:

age	workclass	fnlwgt	education	educational-nu	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
25	Private	226802	11th	7	Never-married	Machine-op-ins	Own-child	Black	Male	0	0	40	United-States	<=50K
38	Private	89814	HS-grad	9	Married-civ-spc	Farming-fishing	Husband	White	Male	0	0	50	United-States	<=50K
28	Local-gov	336951	Assoc-acdm	12	Married-civ-spc	Protective-serv	Husband	White	Male	0	0	40	United-States	>50K
44	Private	160323	Some-college	10	Married-civ-spc	Machine-op-ins	Husband	Black	Male	7688	0	40	United-States	>50K
18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0	30	United-States	<=50K
34	Private	198693	10th	6	Never-married	Other-service	Not-in-family	White	Male	0	0	30	United-States	<=50K
29	?	227026	HS-grad	9	Never-married	?	Unmarried	Black	Male	0	0	40	United-States	<=50K
63	Self-emp-not-in	104626	Prof-school	15	Married-civ-spc	Prof-specialty	Husband	White	Male	3103	0	32	United-States	>50K
24	Private	369667	Some-college	10	Never-married	Other-service	Unmarried	White	Female	0	0	40	United-States	<=50K
55	Private	104996	7th-8th	4	Married-civ-spc	Craft-repair	Husband	White	Male	0	0	10	United-States	<=50K
65	Private	184454	HS-grad	9	Married-civ-spc	Machine-op-ins	Husband	White	Male	6418	0	40	United-States	>50K
36	Federal-gov	212465	Bachelors	13	Married-civ-spc	Adm-clerical	Husband	White	Male	0	0	40	United-States	<=50K
26	Private	82091	HS-grad	9	Never-married	Adm-clerical	Not-in-family	White	Female	0	0	39	United-States	<=50K
58	?	299831	HS-grad	9	Married-civ-spc	?	Husband	White	Male	0	0	35	United-States	<=50K
48	Private	279724	HS-grad	9	Married-civ-spc	Machine-op-ins	Husband	White	Male	3103	0	48	United-States	>50K
43	Private	346189	Masters	14	Married-civ-spc	Exec-manageria	Husband	White	Male	0	0	50	United-States	>50K
20	State-gov	444554	Some-college	10	Never-married	Other-service	Own-child	White	Male	0	0	25	United-States	<=50K
43	Private	128354	HS-grad	9	Married-civ-spc	Adm-clerical	Wife	White	Female	0	0	30	United-States	<=50K
37	Private	60548	HS-grad	9	Widowed	Machine-op-ins	Unmarried	White	Female	0	0	20	United-States	<=50K
40	Private	85019	Doctorate	16	Married-civ-spc	Prof-specialty	Husband	Asian-Pac-Islan	Male	0	0	45	?	>50K
34	Private	107914	Bachelors	13	Married-civ-spc	Tech-support	Husband	White	Male	0	0	47	United-States	>50K

An individual’s annual income results from various factors. Intuitively, it is influenced by the individual’s education level, age, gender, occupation, and etc.

Sure, let's briefly describe each attribute in the Adult Income dataset:

1. Age: Represents the age of an individual in years. It is a continuous numerical variable.

2. Workclass: Describes the type of employment or work arrangement of the individual, such as Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. It is a categorical variable.

3. Fnlwgt: Stands for "final weight" and represents the sampling weight associated with the observation. It is used to correct for biased sampling in the census data. It is a continuous numerical variable.

4. Education: Represents the highest level of education attained by the individual, such as Bachelors, Some-college, 11th, HS-grad, Prof-school, etc. It is a categorical variable.

5. Educational-num: Represents the numerical encoding of the education level. It is often redundant with the 'Education' attribute but encoded numerically. It is a continuous numerical variable.

6. Marital-status: Indicates the marital status of the individual, such as Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. It is a categorical variable.

7. Occupation: Describes the type of occupation or job role of the individual, such as Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, etc. It is a categorical variable

8. Relationship: Indicates the relationship status of the individual in the household, such as Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. It is a categorical variable.

9. Race: Represents the race of the individual, such as White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. It is a categorical variable.

10. Gender: Indicates the gender of the individual, typically 'Male' or 'Female'. It is a categorical variable.

T. Mohamed Nadhim
3122 21 3002 058
ECE A 2025

11. Capital-gain: Represents the capital gains of the individual from investments, stocks, or real estate. It is a continuous numerical variable.

12. Capital-loss: Represents the capital losses of the individual from investments, stocks, or real estate. It is a continuous numerical variable.

13. Hours-per-week: Indicates the number of hours worked per week by the individual. It is a continuous numerical variable.

14. Native-country: Describes the country of origin or citizenship of the individual. It is a categorical variable.

15. income: Represents the annual income of the individual, categorized as either <=50K or >50K, indicating whether the individual earns less than or equal to $50,000 annually or more than $50,000 annually. It is the target variable for classification tasks.

Code:

#Importing the libraries and importing the dataset:

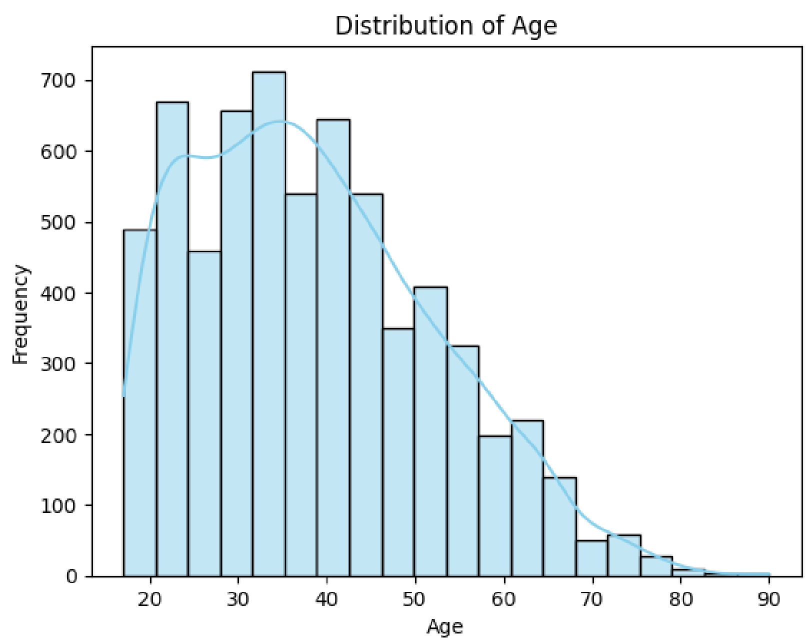
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

data = pd.read_csv('/content/adult_income.csv')
```

1. Histogram Visualization

Visualization of distribution of age in the dataset

```
sns.histplot(data['age'], bins=20, kde=True, color='skyblue')
plt.title('Distribution of Age')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```

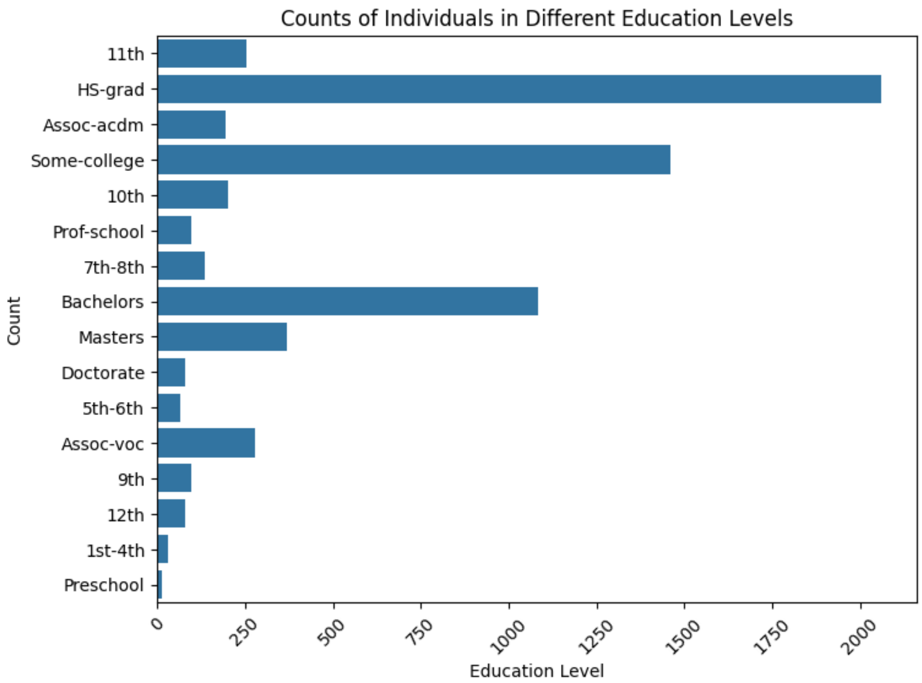


T. Mohamed Nadhim
3122 21 3002 058
ECE A 2025

2. Bar Chart

To display the number of individuals having the same educational qualification

```
plt.figure(figsize=(8, 6))
sns.countplot(data['education'])
plt.title('Counts of Individuals in Different Education Levels')
plt.xlabel('Education Level')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```

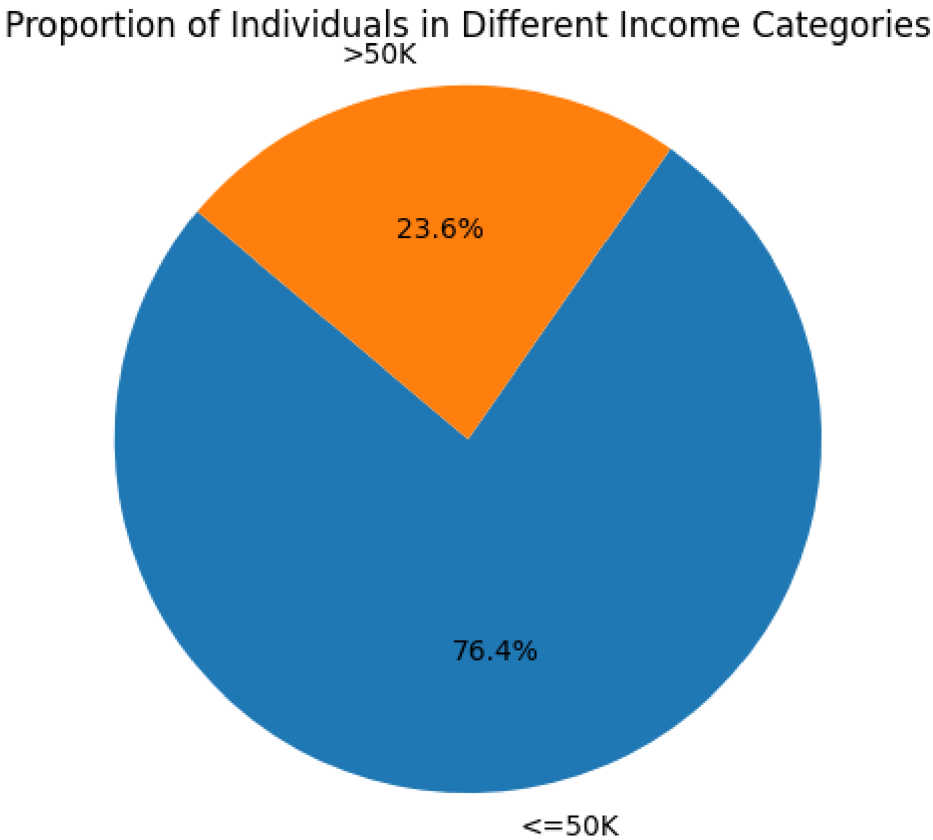


3. Pie chart:

To show the proportion of individuals in different income categories

```
plt.figure(figsize=(5, 5))
income_counts = data['income'].value_counts()
plt.pie(income_counts, labels=income_counts.index, startangle=140)
plt.title('Proportion of Individuals in Different Income Categories')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
plt.show()
```

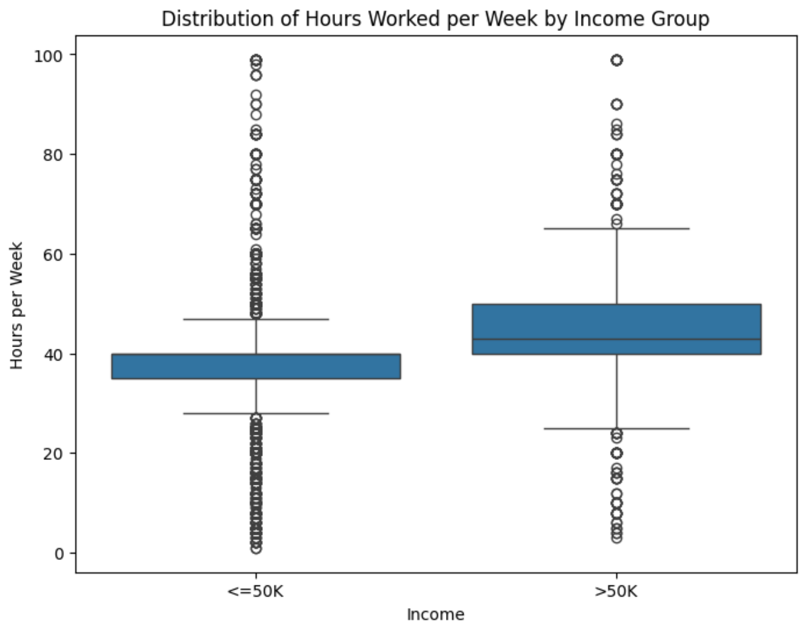
T. Mohamed Nadhim
3122 21 3002 058
ECE A 2025



4. Box plot:

Compare the distribution of hours worked per week between different income groups

```
plt.figure(figsize=(8, 6))
sns.boxplot(x='income', y='hours-per-week', data=data)
plt.title('Distribution of Hours Worked per Week by Income Group')
plt.xlabel('Income')
plt.ylabel('Hours per Week')
plt.show()
```

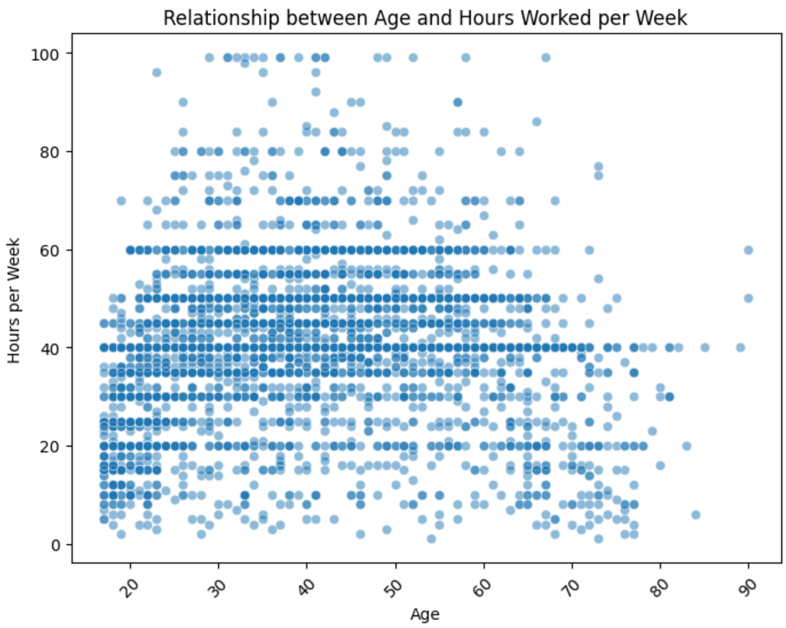


T. Mohamed Nadhim
3122 21 3002 058
ECE A 2025

5. Scatter plot:

Explore the relationship between age and hours worked per week

```
plt.figure(figsize=(8, 6))
sns.scatterplot(x='age', y='hours-per-week', data=data, alpha=0.5)
plt.title('Relationship between Age and Hours Worked per Week')
plt.xlabel('Age')
plt.ylabel('Hours per Week')
plt.xticks(rotation=45)
plt.show()
```

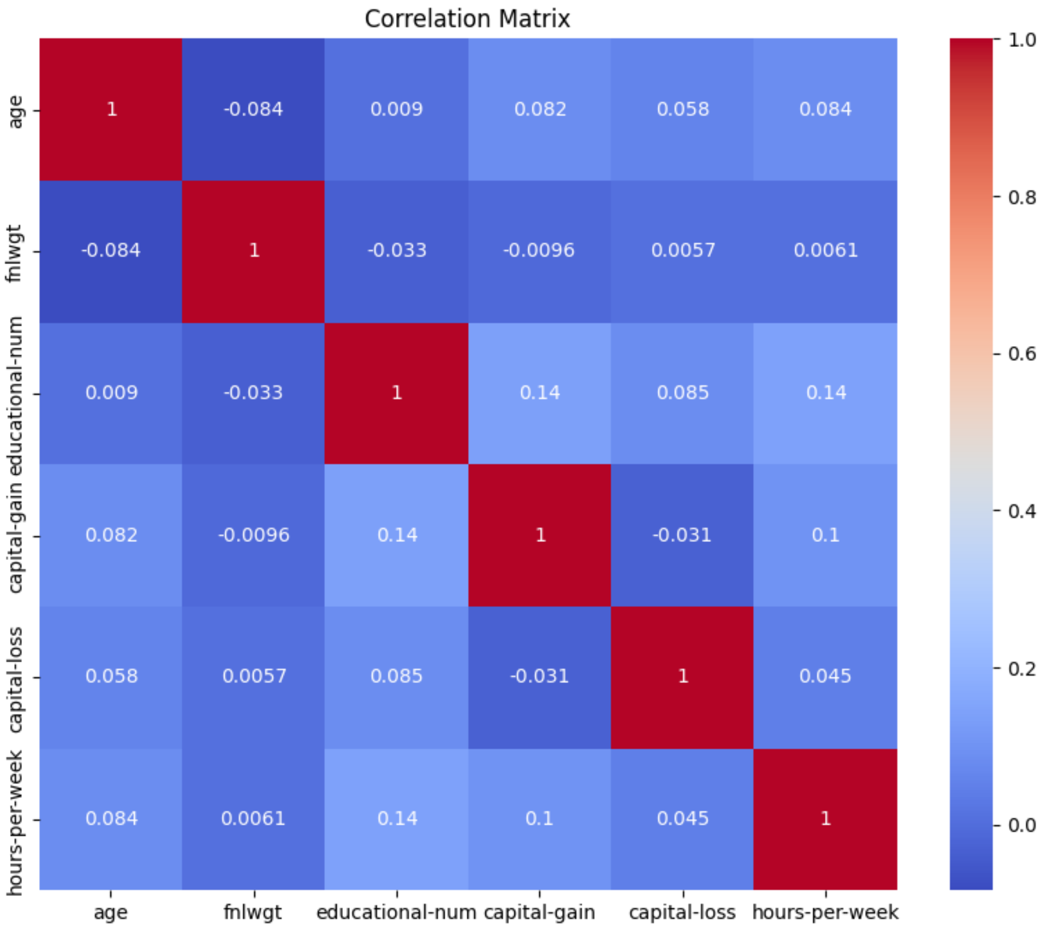


6. Heatmap:

Display a correlation matrix between numerical attributes

```
plt.figure(figsize=(10, 8))
sns.heatmap(data.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

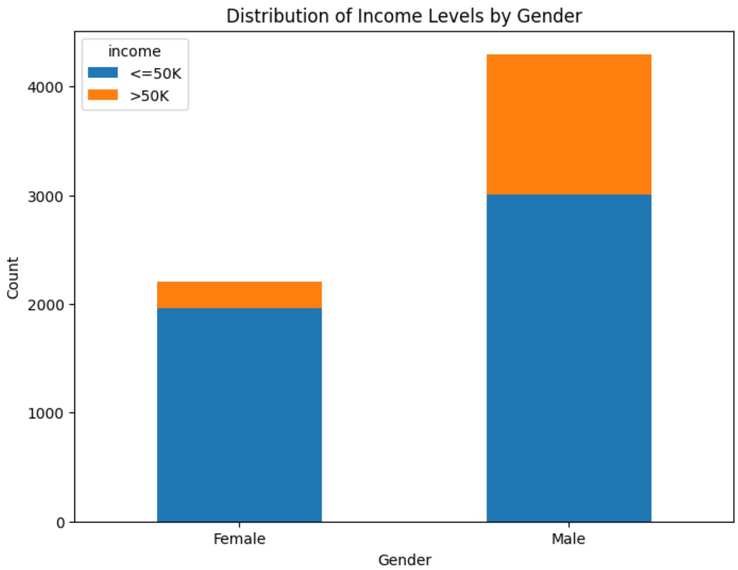
T. Mohamed Nadhim
3122 21 3002 058
ECE A 2025



7. Stacked bar chart:

Compare the distribution of income levels across different genders

```
income_by_gender = pd.crosstab(index=data['gender'], columns=data['income'])
income_by_gender.plot(kind='bar', stacked=True, figsize=(8, 6))
plt.title('Distribution of Income Levels by Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```

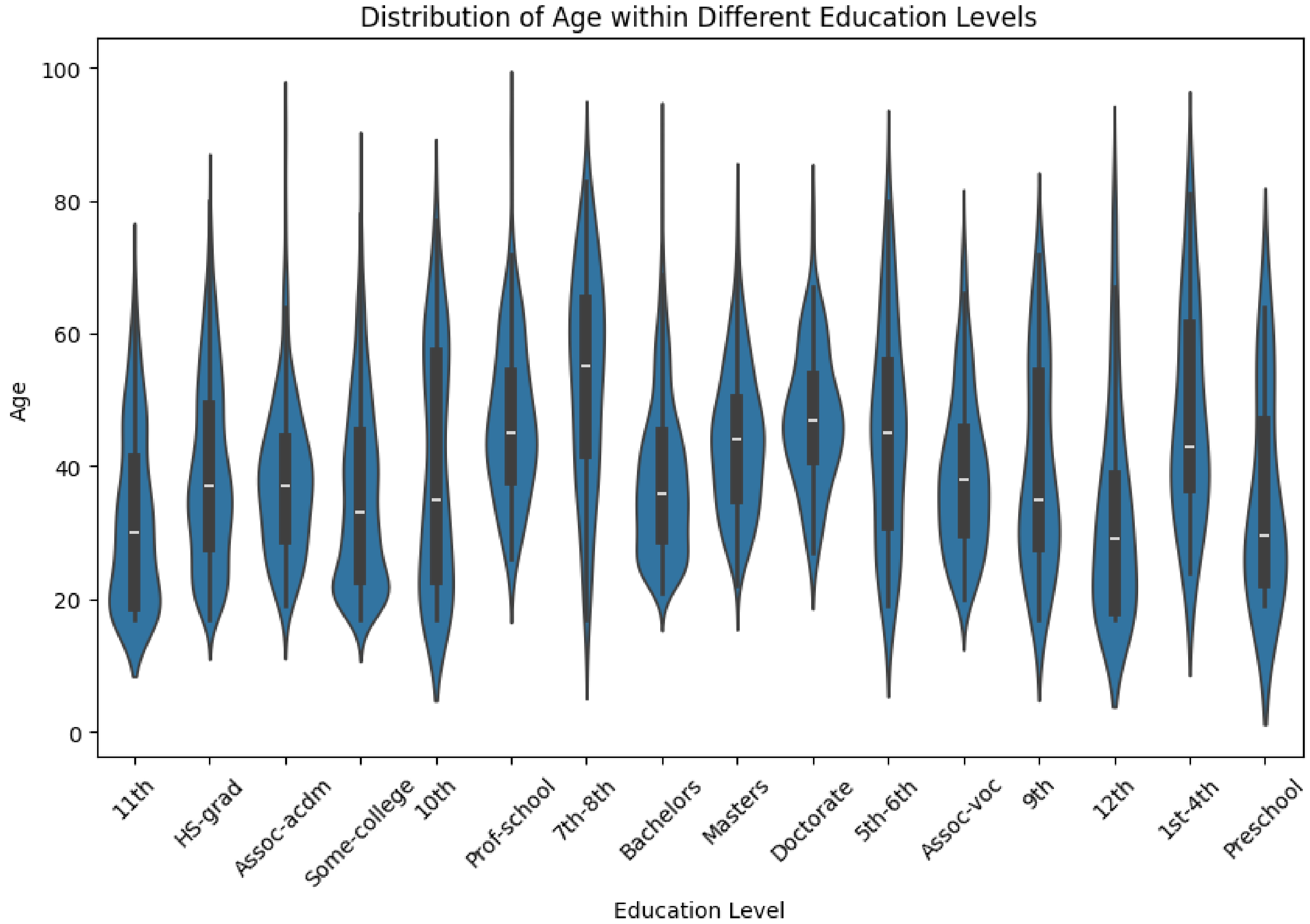


T. Mohamed Nadhim
3122 21 3002 058
ECE A 2025

8. Violin plot:

Visualize the distribution of age within different education levels

```
plt.figure(figsize=(10, 6))
sns.violinplot(x='education', y='age', data=data)
plt.title('Distribution of Age within Different Education Levels')
plt.xlabel('Education Level')
plt.ylabel('Age')
plt.xticks(rotation=45)
plt.show()
```

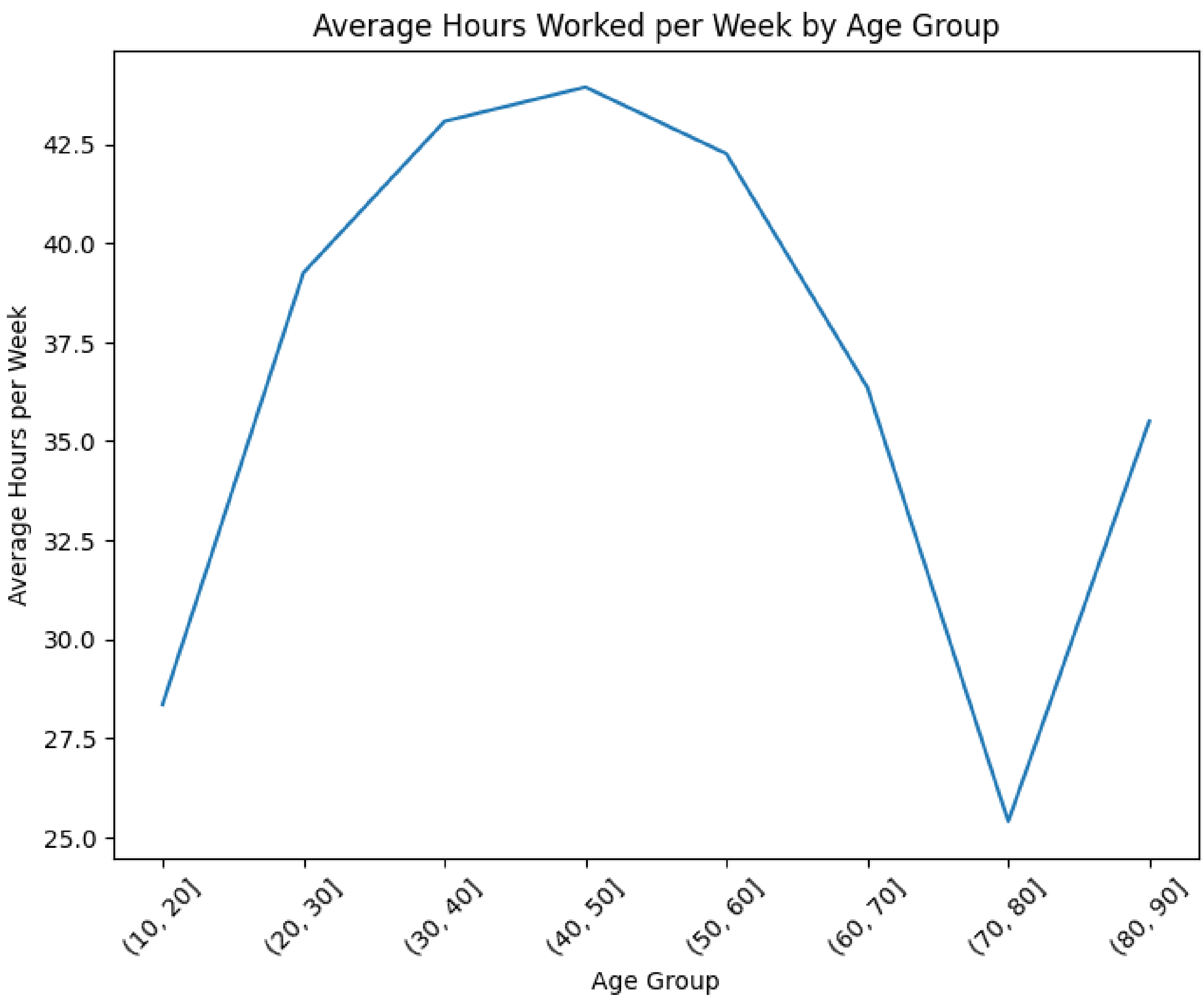


9. Line chart:

Track the changes in average hours worked per week over different age groups

```
age_groups = data.groupby(pd.cut(data['age'], bins=range(10, 100, 10)))
avg_hours_per_week = age_groups['hours-per-week'].mean()
avg_hours_per_week.plot(kind='line', figsize=(8, 6))
plt.title('Average Hours Worked per Week by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Average Hours per Week')
plt.xticks(rotation=45)
plt.show()
```

T. Mohamed Nadhim
3122 21 3002 058
ECE A 2025

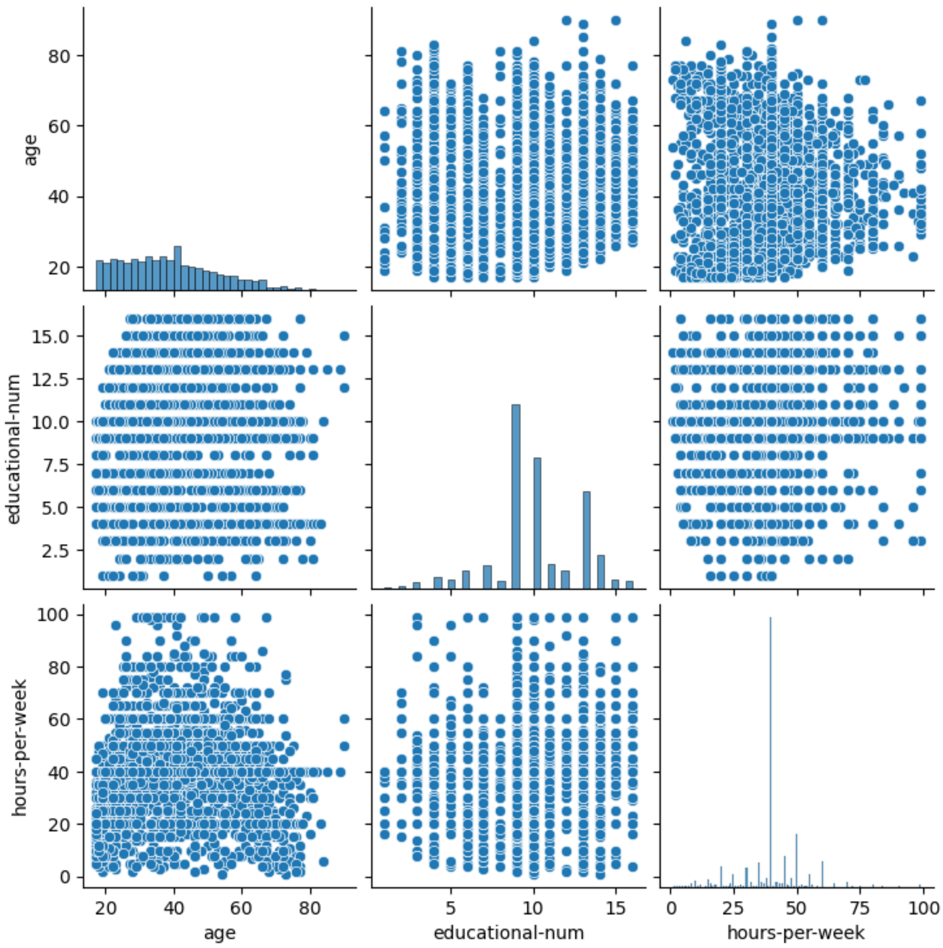


10. Pair plot:

Explore pairwise relationships between multiple numerical attributes

```
sns.pairplot(data[['age', 'educational-num', 'hours-per-week']])
plt.show()
```

T. Mohamed Nadhim
3122 21 3002 058
ECE A 2025

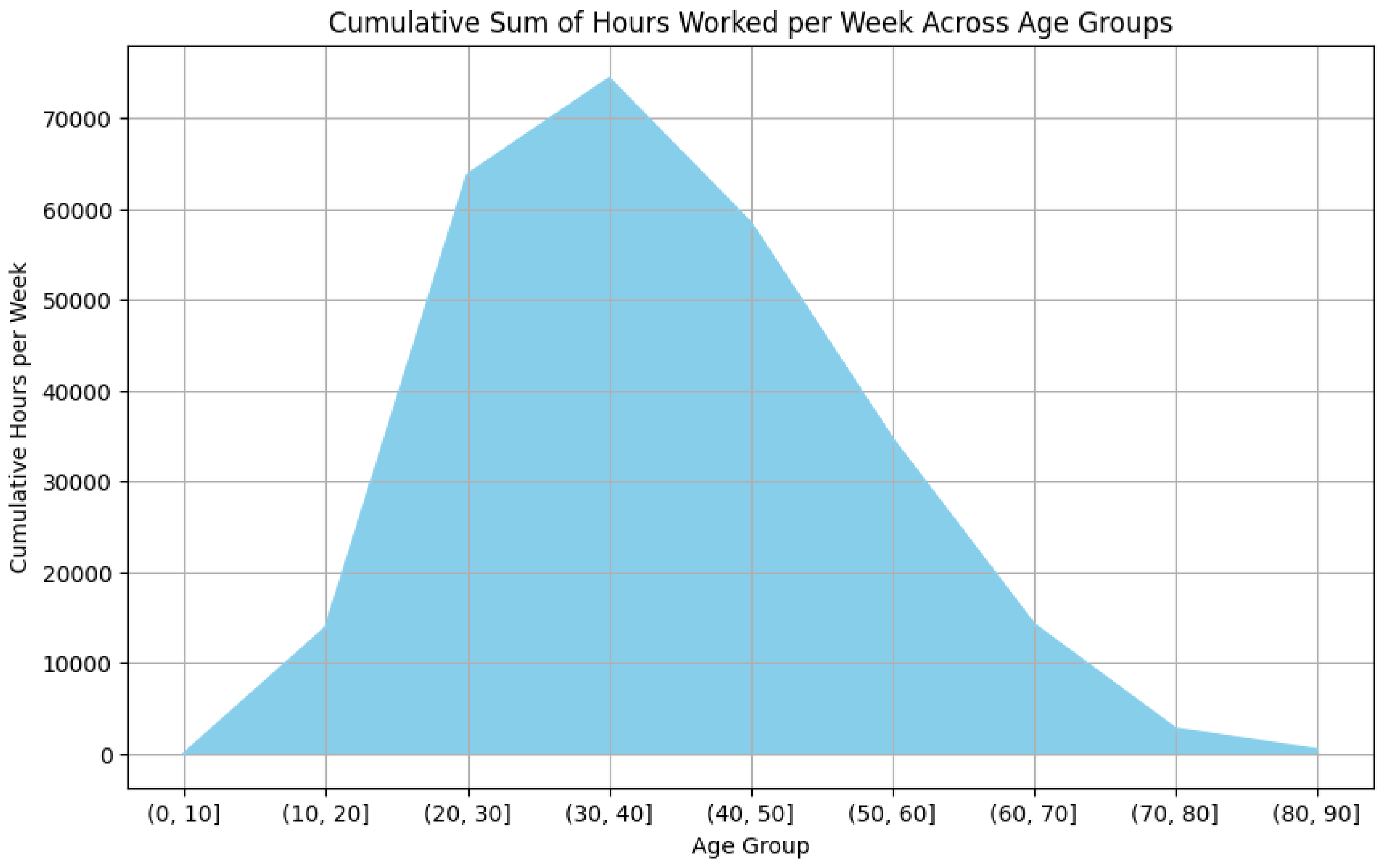


11. Area Chart:

To show the cumulative hours worked by different age groups

```
age_groups = data.groupby(pd.cut(data['age'], bins=range(0, 100, 10)))['hours-per-
week'].sum()
plt.figure(figsize=(10, 6))
age_groups.plot(kind='area', color='skyblue')
plt.title('Cumulative Sum of Hours Worked per Week Across Age Groups')
plt.xlabel('Age Group')
plt.ylabel('Cumulative Hours per Week')
plt.grid(True)
plt.show()
```

T. Mohamed Nadhim
3122 21 3002 058
ECE A 2025



Inference:

Different data columns were visualized revealing interesting correlations between the various factors involved in income ratio. 11 Different types of data visualization was explored.

Project Link:

https://github.com/Nadhim/ML-Lab/tree/main/Experiment_0%20-%20Data%20Visualisation