Demographic Data Analyzer

Project Description :

> In this project, a dataset was extracted from Census 1994. In order to give a picture of this data, some exploratory analysis will be provided. This project source is from freecodecamp.org

Questions :

1. How many people of each race are represented in this dataset? This should be a Pandas series with race names as the index labels. (race column)
2. What is the average age of men?
3. What is the percentage of people who have a Bachelor's degree?
4. What percentage of people with advanced education (Bachelors, Masters, or Doctorate) make more than 50K?
5. What percentage of people without advanced education make more than 50K?
6. What is the minimum number of hours a person works per week?
7. What percentage of the people who work the minimum number of hours per week have a salary of more than 50K?
8. What country has the highest percentage of people that earn >50K and what is that percentage?
9. Identify the most popular occupation for those who earn >50K in India.

Data Cleaning & Preparation

In [1]:

```
import pandas as pd

csv_file = "adult_data.csv"
df = pd.read_csv(csv_file)
df.head()
```

Out[1]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	sala
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=5
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=5
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=5
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=5
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=5

In [2]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   age             32561 non-null  int64 
 1   workclass       32561 non-null  object
 2   fnlwgt          32561 non-null  int64 
 3   education       32561 non-null  object
 4   education-num   32561 non-null  int64 
 5   marital-status  32561 non-null  object
 6   occupation      32561 non-null  object
 7   relationship    32561 non-null  object
 8   race            32561 non-null  object
 9   sex             32561 non-null  object
 10  capital-gain    32561 non-null  int64 
 11  capital-loss    32561 non-null  int64 
 12  hours-per-week  32561 non-null  int64 
 13  native-country  32561 non-null  object
 14  salary          32561 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

In [3]:

```
display(df.isnull().sum())
df = df.drop_duplicates()
display(df)
```

```
age               0
workclass         0
fnlwgt            0
education         0
education-num     0
marital-status    0
occupation        0
relationship      0
race              0
sex               0
capital-gain      0
capital-loss      0
hours-per-week    0
native-country    0
salary            0
dtype: int64
```

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32556	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States

32537 rows × 15 columns

> Since this dataset is a public data, most of them are already prepared and cleaned. So, it is approximately 90% ready for analysis.

Question 1

How many people of each race are represented in this dataset? This should be a Pandas series with race names as the index labels. (race column)

In [4]:

```
race = df.columns[8]
df[race].value_counts()
```

Out[4]:

```
race
White                 27795
Black                  3122
Asian-Pac-Islander     1038
Amer-Indian-Eskimo      311
Other                   271
Name: count, dtype: int64
```

Question 2

What is the average age of men?

In [5]:

```
# Groupping the dataframe by 'sex'
df.groupby(df.columns[9]).agg({df.columns[0]:'mean'}).reset_index()[1:]
```

Out[5]:

	sex	age
1	Male	39.436051

Question 3

What is the percentage of people who have a Bachelor's degree?

In [6]:

```
# Creating bachelors variable to store number of people who have bachelor's degree
bachelors = df['education'].value_counts().iloc[2]
```

```
# The total population in dataset
population = len(df.index)

# percentage calculation
bachelors_percentage = (bachelors/population)*100
print(f"Bachelor's Percentage : {round(bachelors_percentage, 3)}%")
```

```
Bachelor's Percentage : 16.452%
```

Question 4

What percentage of people with advanced education (Bachelors, Masters, or Doctorate) make more than 50K?

In [7]:

```
# create education filter to Bachelors, Masters, and Doctorate only
filter_1 = (df['education'] == 'Bachelors') | (df['education'] == 'Masters') | (df['education'] == 'Doctorate')

# create salary filter to more than 50K
filter_2 = df['salary'] == '>50K'

# count the number of people with the criterion
number_of_people_with_criterion = len(df[filter_1 & filter_2])

# Percentage calculation
percentage = (number_of_people_with_criterion/population)*100
print(f"Percentage of people with advanced education and make more than 50K : {round(percentage, 3)}%")
```

```
Percentage of people with advanced education and make more than 50K : 10.714%
```

Question 5

What percentage of people without advanced education make more than 50K?

In [8]:

```
# Using Question 4 variable to create the number of people without advanced education
without_advance_ed = len(df[~filter_1 & filter_2])

# Calculate the percentage
```

```
percentage_without_advance_ed = (without_advance_ed/population)*100
print(f"Percentage of people without advanced education make more than 50K : {round(percentage_without_advance_ed, 4)}%")
```

```
Percentage of people without advanced education make more than 50K : 13.3786%
```

Question 6

What is the minimum number of hours a person works per week?

In [9]:

```
# Minimum number of hours per person per week
minimum_hours = df[df.columns[12]].min()
print(f"{minimum_hours} hour(s)")
```

```
1 hour(s)
```

Question 7

What percentage of the people who work the minimum number of hours per week have a salary of more than 50K?

In [10]:

```
# Create the filter with minimum work hours
criterion_1 = df[df.columns[12]] == df[df.columns[12]].min()

# Create filter for salary more than 50K
criterion_2 = df[df.columns[14]] == '>50K'

# Number of people with filtered condition
people_min_hours_more_50k = len(df[criterion_1 & criterion_2])

# Calculate the percentage
people_min_hours_more_50k_percentage = (people_min_hours_more_50k/population)*100
print(f"Percentage : {round(people_min_hours_more_50k_percentage, 4)}%")
```

```
Percentage : 0.0061%
```

Question 8

What country has the highest percentage of people that earn >50K and what is that percentage?

In [11]:

```python
# Create dataframe to find number of people with 'salary' more than 50K each country
df_1 = df[df['salary'] == '>50K'].groupby('native-country').agg({'salary':'count'}).reset_index()

# Create dataframe to find number of people each country
df_2 = df.groupby('native-country').agg({'salary':'count'}).reset_index()

# merge 2 dataframe to ease the aggregation
merged_df = df_1.merge(df_2, how='left', left_on='native-country', right_on='native-country')

# calculate the percentage of people that earn more than 50K each country
merged_df['percentage'] = round((merged_df[merged_df.columns[1]]/merged_df[merged_df.columns[2]])*100, 4)
merged_df = merged_df.sort_values(by='percentage', ascending=False)
merged_df.iloc[:1]
```

Out[11]:

	native-country	salary_x	salary_y	percentage
19	Iran	18	43	41.8605

Question 9

Identify the most popular occupation for those who earn >50K in India

In [12]:

```python
# Create filter based on salary and native country
df_filter = (df['salary'] == '>50K') & (df['native-country'] == 'India')

# Apply filter to find the data and specify to occupation
occupation = df[df_filter]['occupation'].value_counts().sort_values(ascending=False)

# Create dataframe for more readable and usable variable
occupation_df = pd.DataFrame(occupation).reset_index().iloc[:1]
occupation_df
```

Out[12]:

	occupation	count
0	Prof-specialty	25

Conclusions

In [13]:

```python
print(f"""
1. Educational advancement doesn't define how high or low their salary. it's shown that people with advanced
education with more than 50K salary are only {round(percentage, 4)}%, compared to people without advanced education
with {round(percentage_without_advance_ed, 4)}%.\n
2. Iran has the highest percentage of people for more than 50K salary with {round(merged_df.iloc[:1, 3:].squeeze(), 4)}%\n
3. {occupation_df['occupation'].squeeze()} is the most popular with {occupation_df['count'].squeeze()} people
""")
```

```
1. Educational advancement doesn't define how high or low their salary. it's shown that people with advanced
education with more than 50K salary are only 10.714%, compared to people without advanced education
with 13.3786%.

2. Iran has the highest percentage of people for more than 50K salary with 41.8605%

3. Prof-specialty is the most popular with 25 people
```