Assignment 3 - Descriptive Statistics (Data Science)

Atharva Taras (TE A - 73)

Perform the following operations on any open source dataset (e.g., data.csv)

Kaggle - Income Dataset (https://www.kaggle.com/datasets/mastmustu/income)

In [1]:

```
import pandas as pd
import seaborn as sns
```

In [2]:

```
data = pd.read_csv('Income_data.csv')
data.head()
```

Out[2]:

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race
0	39	Self-emp-not-inc	327120	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White
1	32	Private	123253	Assoc-acdm	12	Married-civ-spouse	Craft-repair	Husband	White
2	47	Private	232628	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	Black
3	19	Private	374262	12th	8	Never-married	Handlers-cleaners	Own-child	White
4	46	Self-emp-not-inc	311231	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White

1. Provide summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset (age, income etc.) with numeric variables grouped by one of the qualitative (categorical) variable. For example, if your categorical variable is age groups and quantitative variable is income, then provide summary statistics of income grouped by the age groups. Create a list that contains a numeric value for each response to the categorical variable.

In [3]:

```
1 data.describe()
```

Out[3]:

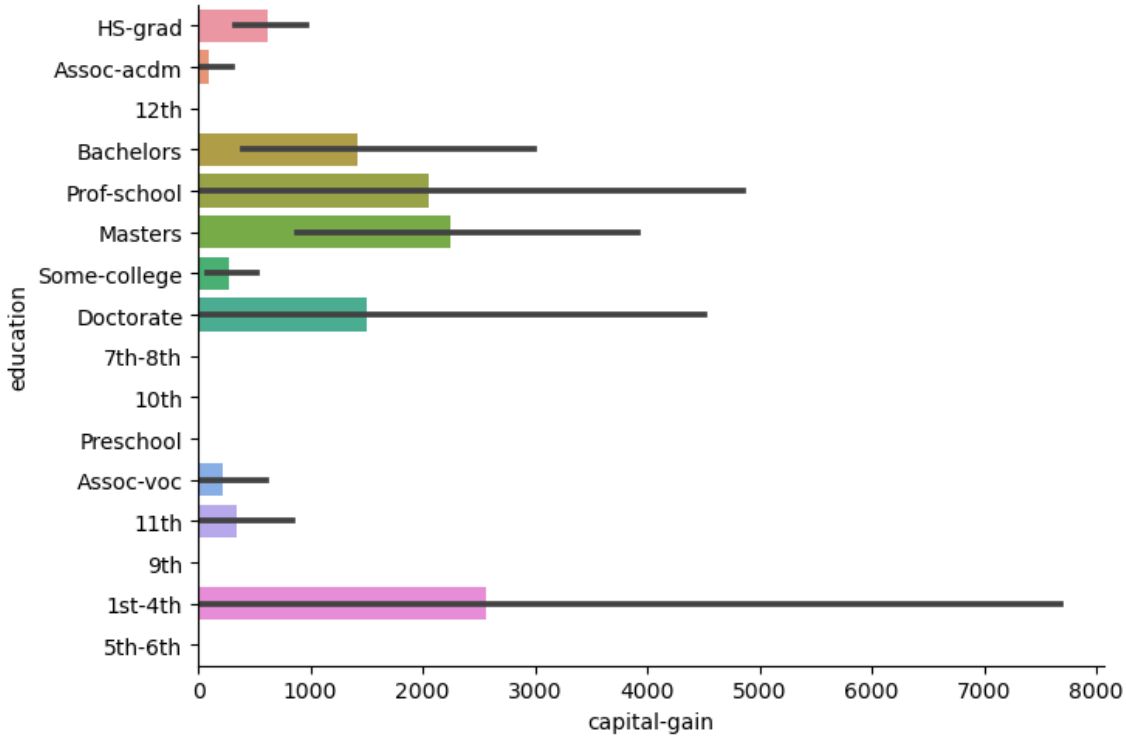
	age	fnlwgt	educational-num	capital-gain	capital-loss	hours-per-week
count	899.000000	899.000000	899.000000	899.000000	899.000000	899.000000
mean	38.576196	194150.017798	10.185762	728.913237	111.929922	41.121246
std	13.079061	104945.494349	2.477511	4355.969800	442.980441	12.397005
min	17.000000	21472.000000	1.000000	0.000000	0.000000	2.000000
25%	28.000000	120925.500000	9.000000	0.000000	0.000000	40.000000
50%	37.000000	181434.000000	10.000000	0.000000	0.000000	40.000000
75%	48.000000	243670.000000	13.000000	0.000000	0.000000	45.000000
max	90.000000	857532.000000	16.000000	99999.000000	2415.000000	99.000000

In [4]:

```
1 sns.catplot(data=data, y='education', x='capital-gain', aspect=1.5, kind='bar', orie
```

Out[4]:

```
<seaborn.axisgrid.FacetGrid at 0x1093cf62a30>
```



2. Write a Python program to display some basic statistical details like percentile, mean, standard deviation etc. of the species of 'Iris-setosa', 'Iris-versicolor' and 'Iris-versicolor' of iris.csv dataset.

Dataset - Iris (https://www.kaggle.com/datasets/uciml/iris)

In [5]:

```
df = pd.read_csv('Iris.csv')
df.head()
```

Out[5]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

In [6]:

```
tdf = df.groupby(by='Species')
tdf.first()
```

Out[6]:

Species	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
Iris-setosa	1	5.1	3.5	1.4	0.2
Iris-versicolor	51	7.0	3.2	4.7	1.4
Iris-virginica	101	6.3	3.3	6.0	2.5

In [7]:

```
tdf.get_group('Iris-setosa').describe()
```

Out[7]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	50.00000	50.00000	50.000000	50.000000	50.00000
mean	25.50000	5.00600	3.418000	1.464000	0.24400
std	14.57738	0.35249	0.381024	0.173511	0.10721
min	1.00000	4.30000	2.300000	1.000000	0.10000
25%	13.25000	4.80000	3.125000	1.400000	0.20000
50%	25.50000	5.00000	3.400000	1.500000	0.20000
75%	37.75000	5.20000	3.675000	1.575000	0.30000
max	50.00000	5.80000	4.400000	1.900000	0.60000

In [8]:

```
1 tdf.get_group('Iris-versicolor').describe()
```

Out[8]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	50.00000	50.000000	50.000000	50.000000	50.000000
mean	75.50000	5.936000	2.770000	4.260000	1.326000
std	14.57738	0.516171	0.313798	0.469911	0.197753
min	51.00000	4.900000	2.000000	3.000000	1.000000
25%	63.25000	5.600000	2.525000	4.000000	1.200000
50%	75.50000	5.900000	2.800000	4.350000	1.300000
75%	87.75000	6.300000	3.000000	4.600000	1.500000
max	100.00000	7.000000	3.400000	5.100000	1.800000

In [9]:

```
1 tdf.get_group('Iris-virginica').describe()
```

Out[9]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	50.00000	50.00000	50.000000	50.000000	50.00000
mean	125.50000	6.58800	2.974000	5.552000	2.02600
std	14.57738	0.63588	0.322497	0.551895	0.27465
min	101.00000	4.90000	2.200000	4.500000	1.40000
25%	113.25000	6.22500	2.800000	5.100000	1.80000
50%	125.50000	6.50000	3.000000	5.550000	2.00000
75%	137.75000	6.90000	3.175000	5.875000	2.30000
max	150.00000	7.90000	3.800000	6.900000	2.50000