Practical No:3

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) variables. 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.

[1]:
```
import pandas as pd
import numpy as np
from sklearn import preprocessing
```

[2]:
```
df = pd.read_csv("Mall_Customers.csv")
```

[3]:
```
df
```

[3]:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

[4]:
```
df.columns
```

[4]:
```
Index(['CustomerID', 'Gender', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)'],
      dtype='object')
```

[5]:
```
df.describe()
```

[5]:

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

[6]:
```
df.isnull().sum()
```

[6]:
```
CustomerID              0
Gender                  0
Age                     0
Annual Income (k$)      0
Spending Score (1-100)  0
dtype: int64
```

[7]:
```
# Compute summary statistics for Annual Income grouped by Gender

summary_stats = df.groupby('Gender')['Annual Income (k$)'].describe()[['mean', '50%', 'min', 'max', 'std']]
summary_stats.rename(columns={'50%': 'median'}, inplace=True)
print(summary_stats)
```

[7]:
```
# Compute summary statistics for Annual Income grouped by Gender

summary_stats = df.groupby('Gender')['Annual Income (k$)'].describe()[['mean', '50%', 'min', 'max', 'std']]
summary_stats.rename(columns={'50%': 'median'}, inplace=True)
print(summary_stats)
```

```
            mean  median   min    max        std
Gender
Female  59.250000    60.0  16.0  126.0  26.011952
Male    62.227273    62.5  15.0  137.0  26.638373
```

[8]:
```
# Create a dictionary with numeric values for each gender category
income_list = df.groupby('Gender')['Annual Income (k$)'].apply(list).to_dict()
print(income_list)
```

```
{'Female': [16, 16, 17, 17, 18, 18, 19, 19, 20, 20, 21, 23, 25, 28, 28, 29, 29, 30, 33, 33, 34, 34, 37, 37, 38, 39, 39, 39, 40, 40, 40, 40, 42, 43, 43, 4
4, 46, 47, 47, 48, 48, 48, 49, 50, 50, 54, 54, 54, 54, 54, 57, 57, 58, 58, 59, 60, 60, 60, 60, 62, 62, 62, 63, 63, 64, 65, 65, 65, 65, 67, 67, 67, 69, 7
0, 70, 72, 72, 73, 73, 74, 75, 76, 76, 77, 78, 78, 78, 78, 78, 78, 78, 79, 79, 81, 85, 86, 87, 88, 88, 97, 97, 98, 99, 101, 103, 103, 103, 103, 113, 120,
120, 126], 'Male': [15, 15, 19, 19, 20, 20, 21, 23, 24, 24, 25, 28, 28, 30, 33, 33, 38, 39, 42, 43, 43, 44, 46, 46, 46, 48, 48, 48, 49, 54, 54, 54, 54, 5
4, 54, 54, 59, 60, 60, 61, 61, 62, 62, 62, 63, 63, 63, 63, 64, 67, 69, 71, 71, 71, 71, 71, 71, 73, 73, 74, 75, 77, 77, 77, 78, 78, 78, 78, 78, 81, 85, 8
6, 87, 87, 87, 87, 87, 88, 88, 93, 93, 98, 99, 101, 113, 126, 137, 137]}
```

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.

[9]:
```
df = pd.read_csv("iris.csv")
```

[10]:
```
print("Dataset Preview:\n", df.head())
```

```
Dataset Preview:
    sepal_length  sepal_width  petal_length  petal_width species
0           5.1          3.5           1.4          0.2  setosa
1           4.9          3.0           1.4          0.2  setosa
2           4.7          3.2           1.3          0.2  setosa
3           4.6          3.1           1.5          0.2  setosa
4           5.0          3.6           1.4          0.2  setosa
```

[11]:
```
# Group by Species and Provide Summary Statistics for Numeric Columns
summary = df.groupby('species').agg(['mean', 'median', 'min', 'max', 'std'])
print("\nSummary Statistics by Grouping Categorical Variable:\n")
print(summary)
```

```
Summary Statistics by Grouping Categorical Variable:

           sepal_length                            sepal_width             \
                   mean median  min  max       std        mean median  min
species
setosa            5.006    5.0  4.3  5.8  0.352490       3.428    3.4  2.3
versicolor        5.936    5.9  4.9  7.0  0.516171       2.770    2.8  2.0
virginica         6.588    6.5  4.9  7.9  0.635880       2.974    3.0  2.2

                        petal_length                            petal_width  \
            max       std         mean median  min  max       std        mean
species
setosa      4.4  0.379064        1.462   1.50  1.0  1.9  0.173664       0.246
versicolor  3.4  0.313798        4.260   4.35  3.0  5.1  0.469911       1.326
virginica   3.8  0.322497        5.552   5.55  4.5  6.9  0.551895       2.026


           median  min  max       std
species
setosa        0.2  0.1  0.6  0.105386
versicolor    1.3  1.0  1.8  0.197753
virginica     2.0  1.4  2.5  0.274650
```

[12]:
```
# Display Statistical Details for Each Species
print("\nStatistical Details for Iris-setosa:")
print(df[df['species'] == 'Iris-setosa'].describe())

print("\nStatistical Details for Iris-versicolor:")
print(df[df['species'] == 'Iris-versicolor'].describe())

print("\nStatistical Details for Iris-virginica:")
print(df[df['species'] == 'Iris-virginica'].describe())
```

```
Statistical Details for Iris-setosa:
       sepal_length  sepal_width  petal_length  petal_width
count           0.0          0.0           0.0          0.0
mean            NaN          NaN           NaN          NaN
std             NaN          NaN           NaN          NaN
min             NaN          NaN           NaN          NaN
25%             NaN          NaN           NaN          NaN
50%             NaN          NaN           NaN          NaN
75%             NaN          NaN           NaN          NaN
max             NaN          NaN           NaN          NaN

Statistical Details for Iris-versicolor:
       sepal_length  sepal_width  petal_length  petal_width
count           0.0          0.0           0.0          0.0
mean            NaN          NaN           NaN          NaN
std             NaN          NaN           NaN          NaN
min             NaN          NaN           NaN          NaN
25%             NaN          NaN           NaN          NaN
50%             NaN          NaN           NaN          NaN
75%             NaN          NaN           NaN          NaN
max             NaN          NaN           NaN          NaN

Statistical Details for Iris-virginica:
       sepal_length  sepal_width  petal_length  petal_width
count           0.0          0.0           0.0          0.0
mean            NaN          NaN           NaN          NaN
std             NaN          NaN           NaN          NaN
min             NaN          NaN           NaN          NaN
25%             NaN          NaN           NaN          NaN
50%             NaN          NaN           NaN          NaN
75%             NaN          NaN           NaN          NaN
max             NaN          NaN           NaN          NaN
```