Assignment 3

Part A Perform the following operations on any open source dataset (eg. data.csv) 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.

Commonly used Measures

1. Measure of Central Tendency
2. Measure of Dispertion

Measure of Central Tendency

1. Mean
2. Mode
3. Medain
4. Std Devaition
5. Minimum
6. Maxmimum

In [1]:
```
import pandas as pd
import numpy as np
```

In [63]:
```
df=pd.read_csv("C:\\Users\\Admin\\Desktop\\Mall_Customers.csv")
df
```

Out[63]:

	CustomerID	Genre	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
5	6	Female	22	17	76
6	7	Female	35	18	6
7	8	Female	23	18	94
8	9	Male	64	19	3
9	10	Female	30	19	72
10	11	Male	67	19	14
11	12	Female	35	19	99
12	13	Female	58	20	15
13	14	Female	24	20	77
14	15	Male	37	20	13
15	16	Male	22	20	79
16	17	Female	35	21	35
17	18	Male	20	21	66
18	19	Male	52	23	29
19	20	Female	35	23	98
20	21	Male	35	24	35
21	22	Male	25	24	73
22	23	Female	46	25	5
23	24	Male	31	25	73
24	25	Female	54	28	14
25	26	Male	29	28	82
26	27	Female	45	28	32
27	28	Male	35	28	61
28	29	Female	40	29	31
29	30	Female	23	29	87
...	...	...	...	...	...
170	171	Male	40	87	13
171	172	Male	28	87	75
172	173	Male	36	87	10
173	174	Male	36	87	92
174	175	Female	52	88	13
175	176	Female	30	88	86
176	177	Male	58	88	15
177	178	Male	27	88	69
178	179	Male	59	93	14
179	180	Male	35	93	90
180	181	Female	37	97	32
181	182	Female	32	97	86
182	183	Male	46	98	15
183	184	Female	29	98	88
184	185	Female	41	99	39
185	186	Male	30	99	97
186	187	Female	54	101	24
187	188	Male	28	101	68
188	189	Female	41	103	17
189	190	Female	36	103	85
190	191	Female	34	103	23
191	192	Female	32	103	69
192	193	Male	33	113	8
193	194	Female	38	113	91
194	195	Female	47	120	16
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18

	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
199	200	Male	30	137	83

200 rows × 5 columns

In [4]:
```
df.mean()
```

Out[4]:
```
CustomerID                200.50
Age                        38.85
Annual Income (k$)         60.56
Spending Score (1-100)     50.20
dtype: float64
```

In [6]:
```
df.median()
```

Out[6]:
```
CustomerID                100.5
Age                        36.0
Annual Income (k$)         61.5
Spending Score (1-100)     50.0
dtype: float64
```

In [7]:
```
df.std()
```

Out[7]:
```
CustomerID                57.879185
Age                       13.969007
Annual Income (k$)        26.264721
Spending Score (1-100)    25.823522
dtype: float64
```

In [8]:
```
df.min()
```

Out[8]:
```
CustomerID                     1
Genre                     Female
Age                           18
Annual Income (k$)            15
Spending Score (1-100)         1
dtype: object
```

In [9]:
```
df.max()
```

Out[9]:
```
CustomerID                 200
Genre                     Male
Age                         70
Annual Income (k$)         137
Spending Score (1-100)      99
dtype: object
```

In [10]:
```
df["Age"].mean()
```

Out[10]:
```
38.85
```

In [11]:
```
df["Age"].mode()
```

Out[11]:
```
0    32
dtype: int64
```

In [12]:
```
df["Age"].median()
```

Out[12]:
```
36.0
```

In [13]:
```
df["Age"].std()
```

Out[13]:
```
13.969007331558883
```

In [15]:
```
gk=df.groupby(["Genre"])
```

In [17]:
```
gk.first()
```

Out[17]:

Genre	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
Female	3	20	16	6
Male	1	19	15	39

part B

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. Provide the codes with outputs and explain everything that you do in this step.

In [38]:
```
csv_url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'
```

In [39]:
```
df_iris = pd.read_csv(csv_url, header = None)
```

In [40]:
```
col_names = ['Sepal_Length','Sepal_Width','Petal_Length','Petal_Width','Species']
```

In [41]:
```
df_iris = pd.read_csv(csv_url, names = col_names)
```

In [43]:
```
df_iris
```

Out[43]:

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa
10	5.4	3.7	1.5	0.2	Iris-setosa
11	4.8	3.4	1.6	0.2	Iris-setosa
12	4.8	3.0	1.4	0.1	Iris-setosa
13	4.3	3.0	1.1	0.1	Iris-setosa
14	5.8	4.0	1.2	0.2	Iris-setosa
15	5.7	4.4	1.5	0.4	Iris-setosa
16	5.4	3.9	1.3	0.4	Iris-setosa
17	5.1	3.5	1.4	0.3	Iris-setosa
18	5.7	3.8	1.7	0.3	Iris-setosa
19	5.1	3.8	1.5	0.3	Iris-setosa
20	5.4	3.4	1.7	0.2	Iris-setosa
21	5.1	3.7	1.5	0.4	Iris-setosa
22	4.6	3.6	1.0	0.2	Iris-setosa
23	5.1	3.3	1.7	0.5	Iris-setosa
24	4.8	3.4	1.9	0.2	Iris-setosa
25	5.0	3.0	1.6	0.2	Iris-setosa
26	5.0	3.4	1.6	0.4	Iris-setosa
27	5.2	3.5	1.5	0.2	Iris-setosa
28	5.2	3.4	1.4	0.2	Iris-setosa
29	4.7	3.2	1.6	0.2	Iris-setosa
...	...	...	...	...	...
120	6.9	3.2	5.7	2.3	Iris-virginica
121	5.6	2.8	4.9	2.0	Iris-virginica
122	7.7	2.8	6.7	2.0	Iris-virginica
123	6.3	2.7	4.9	1.8	Iris-virginica
124	6.7	3.3	5.7	2.1	Iris-virginica
125	7.2	3.2	6.0	1.8	Iris-virginica
126	6.2	2.8	4.8	1.8	Iris-virginica
127	6.1	3.0	4.9	1.8	Iris-virginica
128	6.4	2.8	5.6	2.1	Iris-virginica
129	7.2	3.0	5.8	1.6	Iris-virginica
130	7.4	2.8	6.1	1.9	Iris-virginica
131	7.9	3.8	6.4	2.0	Iris-virginica
132	6.4	2.8	5.6	2.2	Iris-virginica
133	6.3	2.8	5.1	1.5	Iris-virginica
134	6.1	2.6	5.6	1.4	Iris-virginica
135	7.7	3.0	6.1	2.3	Iris-virginica
136	6.3	3.4	5.6	2.4	Iris-virginica
137	6.4	3.1	5.5	1.8	Iris-virginica
138	6.0	3.0	4.8	1.8	Iris-virginica
139	6.9	3.1	5.4	2.1	Iris-virginica
140	6.7	3.1	5.6	2.4	Iris-virginica
141	6.9	3.1	5.1	2.3	Iris-virginica
142	5.8	2.7	5.1	1.9	Iris-virginica
143	6.8	3.2	5.9	2.3	Iris-virginica
144	6.7	3.3	5.7	2.5	Iris-virginica
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Species
149	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 5 columns

In [62]:
```
# Iris Species are of three types 1. Iris-setosa, 2. Iris-versicolor,3.Iris-virginica

gk=df_iris.groupby('Species')
```

In [52]:
```
gk.first()
```

Out[52]:

Species	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width
Iris-setosa	5.1	3.5	1.4	0.2
Iris-versicolor	7.0	3.2	4.7	1.4
Iris-virginica	6.3	3.3	6.0	2.5

In [53]:
```
gk.describe()
```

Out[53]:

	Petal_Length								Petal_Width		...	Sepal_Length		Sepal_Width					
	count	mean	std	min	25%	50%	75%	max	count	mean	...	75%	max	count	mean	std	min	25%	50%	75%	max
Species																					
Iris-setosa	50.0	1.464	0.173511	1.0	1.4	1.50	1.575	1.9	50.0	0.244	...	5.2	5.8	50.0	3.418	0.381024	2.3	3.125	3.4	3.675	4.4
Iris-versicolor	50.0	4.260	0.469911	3.0	4.0	4.35	4.600	5.1	50.0	1.326	...	6.3	7.0	50.0	2.770	0.313798	2.0	2.525	2.8	3.000	3.4
Iris-virginica	50.0	5.552	0.551895	4.5	5.1	5.55	5.875	6.9	50.0	2.026	...	6.9	7.9	50.0	2.974	0.322497	2.2	2.800	3.0	3.175	3.8

3 rows × 32 columns

In [56]:
```
#load all rows of Iris-setosa into iris_Set
iris_Set=(df_iris['Species'] == "Iris-setosa")
```

In [57]:
```
#To display basic statistical details like percentile,mean,std deviation etc for Iris-setosa using describe()
print("Iris-setosa")
```

```
Iris-setosa
```

In [58]:
```
print(df_iris[iris_Set].describe())
```

```
       Sepal_Length  Sepal_Width  Petal_Length  Petal_Width
count      50.00000    50.000000     50.000000     50.00000
mean        5.00600     3.418000      1.464000      0.24400
std         0.35249     0.381024      0.173511      0.10721
min         4.30000     2.300000      1.000000      0.10000
25%         4.80000     3.125000      1.400000      0.20000
50%         5.00000     3.400000      1.500000      0.20000
75%         5.20000     3.675000      1.575000      0.30000
max         5.80000     4.400000      1.900000      0.60000
```

In [60]:
```
iris_Vir=(df_iris['Species'] == "Iris-virginica")
print(df_iris[iris_Vir].describe())
```

```
       Sepal_Length  Sepal_Width  Petal_Length  Petal_Width
count      50.00000    50.000000     50.000000     50.00000
mean        6.58800     2.974000      5.552000      2.02600
std         0.63588     0.322497      0.551895      0.27465
min         4.90000     2.200000      4.500000      1.40000
25%         6.22500     2.800000      5.100000      1.80000
50%         6.50000     3.000000      5.550000      2.00000
75%         6.90000     3.175000      5.875000      2.30000
max         7.90000     3.800000      6.900000      2.50000
```

In [61]:
```
iris_Ver=(df_iris['Species'] == "Iris-versicolor")
print(df_iris[iris_Ver].describe())
```

```
       Sepal_Length  Sepal_Width  Petal_Length  Petal_Width
count     50.000000    50.000000     50.000000    50.000000
mean       5.936000     2.770000      4.260000     1.326000
std        0.516171     0.313798      0.469911     0.197753
min        4.900000     2.000000      3.000000     1.000000
25%        5.600000     2.525000      4.000000     1.200000
50%        5.900000     2.800000      4.350000     1.300000
75%        6.300000     3.000000      4.600000     1.500000
max        7.000000     3.400000      5.100000     1.800000
```

In [ ]:
```

```