Group A

Assignment 3

Descriptive Statistics - Measures of Central Tendency and variability

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

Import all the required Python Libraries.

```
import numpy as np
import pandas as pd
import seaborn as sns
```

Load the Dataset into pandas dataframe.

```
tip_df=sns.load_dataset('tips')
```

```
tip_df.head()
```

```
   total_bill   tip     sex smoker  day    time  size
0       16.99  1.01  Female     No  Sun  Dinner     2
1       10.34  1.66    Male     No  Sun  Dinner     3
2       21.01  3.50    Male     No  Sun  Dinner     3
3       23.68  3.31    Male     No  Sun  Dinner     2
4       24.59  3.61  Female     No  Sun  Dinner     4
```

```
tip_df.shape
```

```
(244, 7)
```

```
tip_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
```

```
 0   total_bill  244 non-null    float64
 1   tip         244 non-null    float64
 2   sex         244 non-null    category
 3   smoker      244 non-null    category
 4   day         244 non-null    category
 5   time        244 non-null    category
 6   size        244 non-null    int64
dtypes: category(4), float64(2), int64(1)
memory usage: 7.4 KB
```

```
tip_df.isnull().sum()
```

```
total_bill    0
tip           0
sex           0
smoker        0
day           0
time          0
size          0
dtype: int64
```

Provide summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset

```
tip_df.describe()
```

```
       total_bill         tip        size
count  244.000000  244.000000  244.000000
mean    19.785943    2.998279    2.569672
std      8.902412    1.383638    0.951100
min      3.070000    1.000000    1.000000
25%     13.347500    2.000000    2.000000
50%     17.795000    2.900000    2.000000
75%     24.127500    3.562500    3.000000
max     50.810000   10.000000    6.000000
```

Provide summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset with numeric variables grouped by one of the qualitative (categorical) variable.

```
col_list=tip_df.select_dtypes('category').columns

for i in col_list:
    print(tip_df[i].value_counts(),end='\n\n')
```

```
sex
Male      157
Female     87
Name: count, dtype: int64
```

```
smoker
No     151
Yes     93
Name: count, dtype: int64

day
Sat     87
Sun     76
Thur    62
Fri     19
Name: count, dtype: int64

time
Dinner    176
Lunch      68
Name: count, dtype: int64


tip_df.groupby('sex').describe()

       total_bill
\
            count       mean       std   min    25%    50%    75%
max
sex

Male        157.0  20.744076  9.246469  7.25  14.00  18.35  24.71
50.81
Female       87.0  18.056897  8.009209  3.07  12.75  16.40  21.52
44.30

          tip             ...                size
\
        count      mean  ...   75%   max  count      mean       std
min  25%
sex                     ...

Male    157.0  3.089618  ...  3.76  10.0  157.0  2.630573  0.955997
1.0  2.0
Female   87.0  2.833448  ...  3.50   6.5   87.0  2.459770  0.937644
1.0  2.0


        50%  75%  max
sex
Male    2.0  3.0  6.0
Female  2.0  3.0  6.0

[2 rows x 24 columns]
```

```
tip_df.groupby('sex')['total_bill'].describe()
```

```
        count       mean       std   min    25%    50%    75%    max
sex
Male    157.0  20.744076  9.246469  7.25  14.00  18.35  24.71  50.81
Female   87.0  18.056897  8.009209  3.07  12.75  16.40  21.52  44.30
```

```
tip_df.groupby('sex')['tip'].describe()
```

```
        count      mean       std  min  25%   50%   75%   max
sex
Male    157.0  3.089618  1.489102  1.0  2.0  3.00  3.76  10.0
Female   87.0  2.833448  1.159495  1.0  2.0  2.75  3.50   6.5
```

```
tip_df.groupby('smoker')['tip'].describe()
```

```
        count      mean       std  min  25%   50%    75%   max
smoker
Yes      93.0  3.008710  1.401468  1.0  2.0  3.00  3.680  10.0
No      151.0  2.991854  1.377190  1.0  2.0  2.74  3.505   9.0
```

```
tip_df.groupby('day')['total_bill'].describe()
```

```
      count       mean       std   min      25%    50%      75%    max
day
Thur   62.0  17.682742  7.886170  7.51  12.4425  16.20  20.1550  43.11
Fri    19.0  17.151579  8.302660  5.75  12.0950  15.38  21.7500  40.17
Sat    87.0  20.441379  9.480419  3.07  13.9050  18.24  24.7400  50.81
Sun    76.0  21.410000  8.832122  7.25  14.9875  19.63  25.5975  48.17
```

```
tip_df.groupby('day')['tip'].describe()
```

```
      count      mean       std   min     25%    50%     75%    max
day
Thur   62.0  2.771452  1.240223  1.25  2.0000  2.305  3.3625   6.70
Fri    19.0  2.734737  1.019577  1.00  1.9600  3.000  3.3650   4.73
Sat    87.0  2.993103  1.631014  1.00  2.0000  2.750  3.3700  10.00
Sun    76.0  3.255132  1.234880  1.01  2.0375  3.150  4.0000   6.50
```

```
tip_df.groupby('time')['total_bill'].describe()
```

```
        count       mean       std   min      25%     50%      75%
max
time

Lunch    68.0  17.168676  7.713882  7.51  12.2350  15.965  19.5325
43.11
Dinner  176.0  20.797159  9.142029  3.07  14.4375  18.390  25.2825
50.81
```

```
tip_df.groupby('time')[['tip','size']].describe()
```

```
          tip                                                          
size  \
        count      mean       std   min  25%   50%     75%   max      
count
time                                                                   

Lunch    68.0  2.728088  1.205345  1.25  2.0  2.25  3.2875   6.7      
68.0
Dinner  176.0  3.102670  1.436243  1.00  2.0  3.00  3.6875  10.0      
176.0


            mean       std  min  25%  50%  75%  max
time
Lunch   2.411765  1.040024  1.0  2.0  2.0  2.0  6.0
Dinner  2.630682  0.910241  1.0  2.0  2.0  3.0  6.0
```

```
tip_df.groupby(['day','sex'])['total_bill'].describe()
```

```
              count       mean        std   min      25%     50%     
75%  \
day  sex

Thur Male     30.0  18.714667   8.019728  7.51  13.6975  16.975     
22.3600
     Female   32.0  16.715312   7.759764  8.35  12.1625  13.785     
18.6750
Fri  Male     10.0  19.857000  10.015847  8.58  12.2350  17.215     
26.0825
     Female    9.0  14.145556   4.788547  5.75  11.3500  15.380     
16.2700
Sat  Male     59.0  20.802542   9.836306  7.74  13.9050  18.240     
24.1650
     Female   28.0  19.680357   8.806470  3.07  14.0500  18.360     
25.5625
Sun  Male     58.0  21.887241   9.129142  7.25  15.1350  20.725     
26.5500
     Female   18.0  19.872222   7.837513  9.60  15.1750  17.410     
24.8975

               max
day  sex
Thur Male    41.19
     Female  43.11
Fri  Male    40.17
     Female  22.75
Sat  Male    50.81
     Female  44.30
Sun  Male    48.17
     Female  35.26
```

```
tip_df.groupby(['sex','smoker'])['total_bill'].describe()
```

```
                count       mean       std   min      25%    50%
75%  \
sex    smoker
Male   Yes      60.0  22.284500  9.911845  7.25  15.2725  20.39
28.5725
       No       97.0  19.791237  8.726566  7.51  13.8100  18.24
22.8200
Female Yes      33.0  17.977879  9.189751  3.07  12.7600  16.27
22.1200
       No       54.0  18.105185  7.286455  7.25  12.6500  16.69
20.8625

                  max
sex    smoker
Male   Yes     50.81
       No      48.33
Female Yes     44.30
       No      35.83
```

```
tip_df.groupby(['sex','smoker','day'])['total_bill'].describe()
```

```
                      count       mean        std    min      25%
50%  \
sex    smoker day
Male   Yes    Thur    10.0  19.171000   6.757421  10.34  15.6100
17.645
              Fri      8.0  20.452500  10.943815   8.58  12.1275
17.215
              Sat     27.0  21.837778   9.988045   7.74  15.1850
20.290
              Sun     15.0  26.141333  10.693824   7.25  17.3550
23.330
       No     Thur    20.0  18.486500   8.739134   7.51  12.6950
16.975
              Fri      2.0  17.475000   7.092281  12.46  14.9675
17.475
              Sat     32.0  19.929063   9.779061   9.55  13.3475
17.870
              Sun     43.0  20.403256   8.140559   8.77  14.4250
19.490
Female Yes    Thur     7.0  19.218571  10.847137  12.74  13.0000
16.400
              Fri      7.0  12.654286   3.883138   5.75  10.7200
13.420
              Sat     15.0  20.266667  10.485703   3.07  12.8300
22.120
```

```
                Sun     4.0  16.540000   4.854764   9.60  15.5325
17.830
       No       Thur   25.0  16.014400   6.783939   8.35  11.3800
13.420
                Fri     2.0  19.365000   4.787113  15.98  17.6725
19.365
                Sat    13.0  19.003846   6.730219   7.25  15.7700
17.070
                Sun    14.0  20.824286   8.396159  10.29  15.1750
17.150

                           75%    max
sex    smoker day
Male   Yes    Thur   20.2575  32.68
              Fri    27.7025  40.17
              Sat    26.2400  50.81
              Sun    33.7650  45.35
       No     Thur   22.7750  41.19
              Fri    19.9825  22.49
              Sat    20.5000  48.33
              Sun    24.5350  48.17
Female Yes    Thur   18.1400  43.11
              Fri    15.8250  16.32
              Sat    27.0200  44.30
              Sun    18.8375  20.90
       No     Thur   18.6400  34.83
              Fri    21.0575  22.75
              Sat    20.6900  35.83
              Sun    25.5325  35.26
```

```
tip_df.groupby(['sex','smoker','time','day'])['total_bill'].describe()
```

```
                              count       mean        std    min      25%
\
sex    smoker time   day

Male   Yes    Lunch  Thur   10.0  19.171000   6.757421  10.34  15.6100

                     Fri     3.0  11.386667   2.510963   8.58  10.3700

              Dinner Fri     5.0  25.892000  10.383290  12.03  21.0100

                     Sat    27.0  21.837778   9.988045   7.74  15.1850

                     Sun    15.0  26.141333  10.693824   7.25  17.3550

       No     Lunch  Thur   20.0  18.486500   8.739134   7.51  12.6950

              Dinner Fri     2.0  17.475000   7.092281  12.46  14.9675

                     Sat    32.0  19.929063   9.779061   9.55  13.3475
```

```
                       Sun    43.0  20.403256   8.140559   8.77  14.4250
Female Yes    Lunch    Thur    7.0  19.218571  10.847137  12.74  13.0000
                       Fri     3.0  13.260000   3.093105  10.09  11.7550
              Dinner   Fri     4.0  12.200000   4.810121   5.75   9.9500
                       Sat    15.0  20.266667  10.485703   3.07  12.8300
                       Sun     4.0  16.540000   4.854764   9.60  15.5325
       No     Lunch    Thur   24.0  15.899167   6.904808   8.35  11.3275
                       Fri     1.0  15.980000        NaN  15.98  15.9800
              Dinner   Thur    1.0  18.780000        NaN  18.78  18.7800
                       Fri     1.0  22.750000        NaN  22.75  22.7500
                       Sat    13.0  19.003846   6.730219   7.25  15.7700
                       Sun    14.0  20.824286   8.396159  10.29  15.1750


                                    50%      75%    max
sex    smoker time   day
Male   Yes    Lunch  Thur  17.645  20.2575  32.68
                     Fri   12.160  12.7900  13.42
              Dinner Fri   27.280  28.9700  40.17
                     Sat   20.290  26.2400  50.81
                     Sun   23.330  33.7650  45.35
       No     Lunch  Thur  16.975  22.7750  41.19
              Dinner Fri   17.475  19.9825  22.49
                     Sat   17.870  20.5000  48.33
                     Sun   19.490  24.5350  48.17
Female Yes    Lunch  Thur  16.400  18.1400  43.11
                     Fri   13.420  14.8450  16.27
              Dinner Fri   13.365  15.6150  16.32
                     Sat   22.120  27.0200  44.30
                     Sun   17.830  18.8375  20.90
       No     Lunch  Thur  13.290  18.3550  34.83
                     Fri   15.980  15.9800  15.98
              Dinner Thur  18.780  18.7800  18.78
                     Fri   22.750  22.7500  22.75
                     Sat   17.070  20.6900  35.83
                     Sun   17.150  25.5325  35.26
```

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.

Load iris.csv dataset Dataset into pandas dataframe.

```
iris_df=sns.load_dataset('iris')
```

```
iris_df.head()
```

```
   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
```

```
iris_df.shape
```

```
(150, 5)
```

```
iris_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   sepal_length  150 non-null    float64
 1   sepal_width   150 non-null    float64
 2   petal_length  150 non-null    float64
 3   petal_width   150 non-null    float64
 4   species       150 non-null    object 
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

```
iris_df.isnull().sum()
```

```
sepal_length    0
sepal_width     0
petal_length    0
petal_width     0
species         0
dtype: int64
```

```
iris_df['species'].value_counts()
```

```
species
setosa        50
versicolor    50
```

```
virginica     50
Name: count, dtype: int64
```

Display some basic statistical details like percentile, mean, standard deviation etc. of the species of 'Iris-setosa', 'Iris-versicolor' and 'Iris-verginica' of iris.csv dataset.

```
iris_df.describe()
```

```
       sepal_length  sepal_width  petal_length  petal_width
count    150.000000   150.000000    150.000000   150.000000
mean       5.843333     3.057333      3.758000     1.199333
std        0.828066     0.435866      1.765298     0.762238
min        4.300000     2.000000      1.000000     0.100000
25%        5.100000     2.800000      1.600000     0.300000
50%        5.800000     3.000000      4.350000     1.300000
75%        6.400000     3.300000      5.100000     1.800000
max        7.900000     4.400000      6.900000     2.500000
```

```
iris_df.groupby(['species'])['sepal_length'].describe()
```

```
            count   mean       std  min    25%  50%  75%  max
species
setosa       50.0  5.006  0.352490  4.3  4.800  5.0  5.2  5.8
versicolor   50.0  5.936  0.516171  4.9  5.600  5.9  6.3  7.0
virginica    50.0  6.588  0.635880  4.9  6.225  6.5  6.9  7.9
```

```
iris_df.groupby(['species'])['sepal_width'].describe()
```

```
            count   mean       std  min    25%  50%    75%  max
species
setosa       50.0  3.428  0.379064  2.3  3.200  3.4  3.675  4.4
versicolor   50.0  2.770  0.313798  2.0  2.525  2.8  3.000  3.4
virginica    50.0  2.974  0.322497  2.2  2.800  3.0  3.175  3.8
```

```
iris_df.groupby(['species'])
[['petal_length','petal_width']].describe()
```

```
petal_length                                                    \
                  count   mean       std  min  25%   50%    75%  max

species

setosa             50.0  1.462  0.173664  1.0  1.4  1.50  1.575  1.9

versicolor         50.0  4.260  0.469911  3.0  4.0  4.35  4.600  5.1

virginica          50.0  5.552  0.551895  4.5  5.1  5.55  5.875  6.9


           petal_width
                 count   mean       std  min  25%  50%  75%  max
```

```
species
setosa            50.0  0.246  0.105386  0.1  0.2  0.2  0.3  0.6
versicolor        50.0  1.326  0.197753  1.0  1.2  1.3  1.5  1.8
virginica         50.0  2.026  0.274650  1.4  1.8  2.0  2.3  2.5
```