Descriptive Statistics

May 1, 2022

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

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

1 Statistis

```
[3]: df.head()
```

```
[3]:    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
```

```
[4]: df.tail()
```

```
[4]:      CustomerID   Genre  Age  Annual Income (k$)  Spending Score (1-100)
     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
```

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

```
[5]: CustomerID                100.50
     Age                        38.85
     Annual Income (k$)         60.56
     Spending Score (1-100)     50.20
     dtype: float64
```

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

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

```
[7]: df.mode()
```

```
[7]:      CustomerID   Genre   Age  Annual Income (k$)  Spending Score (1-100)
     0             1  Female  32.0                54.0                    42.0
     1             2     NaN   NaN                78.0                     NaN
     2             3     NaN   NaN                 NaN                     NaN
     3             4     NaN   NaN                 NaN                     NaN
     4             5     NaN   NaN                 NaN                     NaN
     ..          …      …   …                  …                      …
     195         196     NaN   NaN                 NaN                     NaN
     196         197     NaN   NaN                 NaN                     NaN
     197         198     NaN   NaN                 NaN                     NaN
     198         199     NaN   NaN                 NaN                     NaN
     199         200     NaN   NaN                 NaN                     NaN

     [200 rows x 5 columns]
```

```
[8]: df.loc[:,'Age'].mode()
```

```
[8]: 0    32
     dtype: int64
```

```
[9]: df.min()
```

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

```
[10]: df.max()
```

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

```
[11]: df.std()
```

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

2 statistis of income grouped by age grouped

```
[12]: df.groupby(['Genre'])['Age'].mean()
```

```
[12]: Genre
      Female    38.098214
      Male      39.806818
      Name: Age, dtype: float64
```

```
[13]: df_u = df.rename(columns = {'Annual Income (k$)':'Income'}, inplace=False)
      df_u.groupby(['Genre']).Income.mean()
```

```
[13]: Genre
      Female    59.250000
      Male      62.227273
      Name: Income, dtype: float64
```

```
[14]: from sklearn import preprocessing
      one_hot_encoder = preprocessing.OneHotEncoder()
      encoding = pd.DataFrame(one_hot_encoder.fit_transform(df[['Genre']]).toarray())
      encoding
```

```
[14]:        0    1
      0    0.0  1.0
      1    0.0  1.0
      2    1.0  0.0
      3    1.0  0.0
      4    1.0  0.0
      ..   …   …
      195  1.0  0.0
      196  1.0  0.0
      197  0.0  1.0
      198  0.0  1.0
      199  0.0  1.0

      [200 rows x 2 columns]
```

```
[15]: df_encoding = df_u.join(encoding)
      df_encoding
```

```
[15]:      CustomerID   Genre  Age  Income  Spending Score (1-100)    0    1
      0             1    Male   19      15                      39  0.0  1.0
      1             2    Male   21      15                      81  0.0  1.0
      2             3  Female   20      16                       6  1.0  0.0
      3             4  Female   23      16                      77  1.0  0.0
      4             5  Female   31      17                      40  1.0  0.0
      ..          …      …   …    …                    …   …   …
      195         196  Female   35     120                      79  1.0  0.0
      196         197  Female   45     126                      28  1.0  0.0
      197         198    Male   32     126                      74  0.0  1.0
      198         199    Male   32     137                      18  0.0  1.0
      199         200    Male   30     137                      83  0.0  1.0

      [200 rows x 7 columns]
```

3 Statistical on iris dataset

```
[16]: iris = pd.read_csv("IRIS.csv")
```

```
[17]: iris.head()
```

```
[17]:    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
```

```
[18]: irisSet = (iris['species'] == 'Iris-setosa')
      print("Iris-virginica")
      print(iris[irisSet].describe())
```

```
Iris-virginica
       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
```

```
[19]: irisSet = (iris['species'] == 'Iris-versicolor')
      print("Iris-virginica")
      print(iris[irisSet].describe())
```

```
Iris-virginica
       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
```

[20]:
```
irisSet = (iris['species'] == 'Iris-virginica')
print("Iris-virginica")
print(iris[irisSet].describe())
```

```
Iris-virginica
       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
```