Name: Vedant Bothikar Roll No: 31115 Class: TE-1 Batch: K-1

DSDBAL 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-virginica' of iris.csv dataset. Provide the codes with outputs and explain everything that you do in this step.

In [123…
```
import numpy as np
import pandas as pd
```

# Part 1 -NBA Dataset

In [124…
```
#loading dataset

data = pd.read_csv("nba.csv")
df=pd.DataFrame(data)
```

In [125…
```
shape = df.shape
size = df.size
print("Dimenation of data frame: {}".format(shape))
print("Size of data frame: {}".format(size))
```

```
Dimenation of data frame: (458, 9)
Size of data frame: 4122
```

In [126…
```
print(df.columns)
```

```
Index(['Name', 'Team', 'Number', 'Position', 'Age', 'Height', 'Weight',
       'College', 'Salary'],
      dtype='object')
```

In [127…
```
df.head(5)
```

Out[127…

| | Name | Team | Number | Position | Age | Height | Weight | College | Salary |
|---|---|---|---|---|---|---|---|---|---|
| **0** | Avery Bradley | Boston Celtics | 0.0 | PG | 25.0 | 6-2 | 180.0 | Texas | 7730337.0 |
| **1** | Jae Crowder | Boston Celtics | 99.0 | SF | 25.0 | 6-6 | 235.0 | Marquette | 6796117.0 |
| **2** | John Holland | Boston Celtics | 30.0 | SG | 27.0 | 6-5 | 205.0 | Boston University | NaN |
| **3** | R.J. Hunter | Boston Celtics | 28.0 | SG | 22.0 | 6-5 | 185.0 | Georgia State | 1148640.0 |

| | Name | Team | Number | Position | Age | Height | Weight | College | Salary |
|---|---|---|---|---|---|---|---|---|---|
| **4** | Jonas Jerebko | Boston Celtics | 8.0 | PF | 29.0 | 6-10 | 231.0 | NaN | 5000000.0 |

In [128…
```
df.tail(5)
```

Out[128…

| | Name | Team | Number | Position | Age | Height | Weight | College | Salary |
|---|---|---|---|---|---|---|---|---|---|
| **453** | Shelvin Mack | Utah Jazz | 8.0 | PG | 26.0 | 6-3 | 203.0 | Butler | 2433333.0 |
| **454** | Raul Neto | Utah Jazz | 25.0 | PG | 24.0 | 6-1 | 179.0 | NaN | 900000.0 |
| **455** | Tibor Pleiss | Utah Jazz | 21.0 | C | 26.0 | 7-3 | 256.0 | NaN | 2900000.0 |
| **456** | Jeff Withey | Utah Jazz | 24.0 | C | 26.0 | 7-0 | 231.0 | Kansas | 947276.0 |
| **457** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

In [129…
```
df.describe()
```

Out[129…

| | Number | Age | Weight | Salary |
|---|---|---|---|---|
| **count** | 457.000000 | 457.000000 | 457.000000 | 4.460000e+02 |
| **mean** | 17.678337 | 26.938731 | 221.522976 | 4.842684e+06 |
| **std** | 15.966090 | 4.404016 | 26.368343 | 5.229238e+06 |
| **min** | 0.000000 | 19.000000 | 161.000000 | 3.088800e+04 |
| **25%** | 5.000000 | 24.000000 | 200.000000 | 1.044792e+06 |
| **50%** | 13.000000 | 26.000000 | 220.000000 | 2.839073e+06 |
| **75%** | 25.000000 | 30.000000 | 240.000000 | 6.500000e+06 |
| **max** | 99.000000 | 40.000000 | 307.000000 | 2.500000e+07 |

In [130…
```
df.dtypes
```

Out[130…
```
Name          object
Team          object
Number       float64
Position      object
Age          float64
Height        object
Weight       float64
College       object
Salary       float64
dtype: object
```

## checking for missing values

In [131…
```
df.isnull()
```

Out[131…

| | Name | Team | Number | Position | Age | Height | Weight | College | Salary |
|---|---|---|---|---|---|---|---|---|---|
| **0** | False | False | False | False | False | False | False | False | False |
| **1** | False | False | False | False | False | False | False | False | False |
| **2** | False | False | False | False | False | False | False | False | True |

| | Name | Team | Number | Position | Age | Height | Weight | College | Salary |
|---|---|---|---|---|---|---|---|---|---|
| **3** | False | False | False | False | False | False | False | False | False |
| **4** | False | False | False | False | False | False | False | True | False |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **453** | False | False | False | False | False | False | False | False | False |
| **454** | False | False | False | False | False | False | False | True | False |
| **455** | False | False | False | False | False | False | False | True | False |
| **456** | False | False | False | False | False | False | False | False | False |
| **457** | True | True | True | True | True | True | True | True | True |

458 rows × 9 columns

In [133…
```
df.isnull().sum()
```

Out[133…
```
Name          1
Team          1
Number        1
Position      1
Age           1
Height        1
Weight        1
College      85
Salary       12
dtype: int64
```

In [11]:
```
df['Salary'].fillna(df['Salary'].mean(),inplace=True)
```

In [134…
```
mode = df['College'].mode()[0]
df['College'].fillna(mode,inplace=True)
```

In [135…
```
df['Age'].fillna(df['Age'].median(),inplace=True)
df['Weight'].fillna(df['Weight'].median(),inplace=True)
```

In [136…
```
df.dropna(inplace=True)
```

In [137…
```
df.isnull().sum()
```

Out[137…
```
Name         0
Team         0
Number       0
Position     0
Age          0
Height       0
Weight       0
College      0
Salary       0
dtype: int64
```

In [138…
```
df.shape
```

Out[138… (446, 9)

## Statistical Analysis

In [139…
```
df['Age'].describe()
```

Out[139…
```
count    446.000000
mean      26.919283
std        4.398951
min       19.000000
25%       24.000000
50%       26.000000
75%       30.000000
max       40.000000
Name: Age, dtype: float64
```

In [140…
```
df['Age'].value_counts()
```

Out[140…
```
24.0    45
25.0    44
27.0    40
23.0    39
26.0    36
28.0    30
30.0    30
29.0    27
22.0    26
31.0    22
20.0    19
21.0    19
33.0    14
32.0    13
34.0    10
36.0    10
35.0     8
38.0     4
37.0     3
40.0     3
39.0     2
19.0     2
Name: Age, dtype: int64
```

In [141…
```
bins= [11,21,31,41]
labels = ['11 to 20','21 to 30','31 to 40']
df['ageGroup'] = pd.cut(df['Age'], bins=bins,labels=labels, right=False)
```

In [142…
```
ageGroup_categories=list(df['ageGroup'].value_counts().index)
ageGroup_categories
```

Out[142… ['21 to 30', '31 to 40', '11 to 20']

In [143…
```
df['ageGroup'].value_counts()
```

Out[143…
```
21 to 30    336
31 to 40     89
11 to 20     21
Name: ageGroup, dtype: int64
```

In [144…
```
df.head(5)
```

Out[144…

| | Name | Team | Number | Position | Age | Height | Weight | College | Salary | ageGroup |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Avery Bradley | Boston Celtics | 0.0 | PG | 25.0 | 6-2 | 180.0 | Texas | 7730337.0 | 21 to 30 |
| **1** | Jae Crowder | Boston Celtics | 99.0 | SF | 25.0 | 6-6 | 235.0 | Marquette | 6796117.0 | 21 to 30 |
| **3** | R.J. Hunter | Boston Celtics | 28.0 | SG | 22.0 | 6-5 | 185.0 | Georgia State | 1148640.0 | 21 to 30 |
| **4** | Jonas Jerebko | Boston Celtics | 8.0 | PF | 29.0 | 6-10 | 231.0 | Kentucky | 5000000.0 | 21 to 30 |
| **5** | Amir Johnson | Boston Celtics | 90.0 | PF | 29.0 | 6-9 | 240.0 | Kentucky | 12000000.0 | 21 to 30 |

In [145…
```
df.groupby(df['ageGroup']).get_group('21 to 30')
```

Out[145…

| | Name | Team | Number | Position | Age | Height | Weight | College | Salary | ageGroup |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Avery Bradley | Boston Celtics | 0.0 | PG | 25.0 | 6-2 | 180.0 | Texas | 7730337.0 | 21 to 30 |
| **1** | Jae Crowder | Boston Celtics | 99.0 | SF | 25.0 | 6-6 | 235.0 | Marquette | 6796117.0 | 21 to 30 |
| **3** | R.J. Hunter | Boston Celtics | 28.0 | SG | 22.0 | 6-5 | 185.0 | Georgia State | 1148640.0 | 21 to 30 |
| **4** | Jonas Jerebko | Boston Celtics | 8.0 | PF | 29.0 | 6-10 | 231.0 | Kentucky | 5000000.0 | 21 to 30 |
| **5** | Amir Johnson | Boston Celtics | 90.0 | PF | 29.0 | 6-9 | 240.0 | Kentucky | 12000000.0 | 21 to 30 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **451** | Chris Johnson | Utah Jazz | 23.0 | SF | 26.0 | 6-6 | 206.0 | Dayton | 981348.0 | 21 to 30 |
| **453** | Shelvin Mack | Utah Jazz | 8.0 | PG | 26.0 | 6-3 | 203.0 | Butler | 2433333.0 | 21 to 30 |
| **454** | Raul Neto | Utah Jazz | 25.0 | PG | 24.0 | 6-1 | 179.0 | Kentucky | 900000.0 | 21 to 30 |
| **455** | Tibor Pleiss | Utah Jazz | 21.0 | C | 26.0 | 7-3 | 256.0 | Kentucky | 2900000.0 | 21 to 30 |
| **456** | Jeff Withey | Utah Jazz | 24.0 | C | 26.0 | 7-0 | 231.0 | Kansas | 947276.0 | 21 to 30 |

336 rows × 10 columns

In [146…
```
list_of_salaries_by_ageGroup = list(df.groupby('ageGroup')['Salary'])
list_of_salaries_by_ageGroup
```

Out[146…
```
[('11 to 20',
  13     1749840.0
  40     4131720.0
  56     4582680.0
  60      525093.0
  62     1524000.0
  85     1131960.0
  116    5103120.0
```

```
 122    2127840.0
 192    2841960.0
 208    2357760.0
 226    1733040.0
 352    2481720.0
 356    4171680.0
 377    1920240.0
 393    3102240.0
 401    1282080.0
 410    5703600.0
 427     525093.0
 441    2637720.0
 445    3777720.0
 452    2239800.0
 Name: Salary, dtype: float64),
('21 to 30',
 0       7730337.0
 1       6796117.0
 3       1148640.0
 4       5000000.0
 5      12000000.0
           ...
 451     981348.0
 453    2433333.0
 454     900000.0
 455    2900000.0
 456     947276.0
 Name: Salary, Length: 336, dtype: float64),
('31 to 40',
 19      6300000.0
 31      1635476.0
 33     22875000.0
 34      7402812.0
 43       947276.0
           ...
 406      947276.0
 413     3750000.0
 415     3135000.0
 420      222888.0
 434     5016000.0
 Name: Salary, Length: 89, dtype: float64)]
```

In [147…
```
df.groupby('ageGroup')['Salary'].mean()
```

Out[147…
```
ageGroup
11 to 20    2.650043e+06
21 to 30    4.674760e+06
31 to 40    5.994010e+06
Name: Salary, dtype: float64
```

In [148…
```
df.groupby('ageGroup')['Salary'].median()
```

Out[148…
```
ageGroup
11 to 20    2357760.0
21 to 30    2502860.0
31 to 40    4000000.0
Name: Salary, dtype: float64
```

In [149…
```
df.groupby('ageGroup')['Salary'].describe()
```

Out[149…

| | count | mean | std | min | 25% | 50% | 75% |
|---|---|---|---|---|---|---|---|

| ageGroup | count | mean | std | min | 25% | 50% | 75% | |
|---|---|---|---|---|---|---|---|---|
| **ageGroup** | | | | | | | | |
| **11 to 20** | 21.0 | 2.650043e+06 | 1.454546e+06 | 525093.0 | 1733040.0 | 2357760.0 | 3777720.00 | 5703 |
| **21 to 30** | 336.0 | 4.674760e+06 | 5.063389e+06 | 30888.0 | 1007026.0 | 2502860.0 | 6649029.25 | 22359 |
| **31 to 40** | 89.0 | 5.994010e+06 | 6.132133e+06 | 200600.0 | 1449187.0 | 4000000.0 | 7500000.00 | 25000 |

In [150…

```
list_of_colleges_by_ageGroup = list(df.groupby('ageGroup')['College'])
list_of_colleges_by_ageGroup
```

Out[150…

```
[('11 to 20',
  13       Kentucky
  40       Kentucky
  56           Duke
  60           UNLV
  62       Kentucky
  85           UCLA
  116    Ohio State
  122      Kentucky
  192       Arizona
  208         Texas
  226          UNLV
  352          Duke
  356       Arizona
  377        Kansas
  393      Kentucky
  401          Duke
  410      Kentucky
  427        Kansas
  441       Indiana
  445      Kentucky
  452      Kentucky
  Name: College, dtype: object),
 ('21 to 30',
  0              Texas
  1          Marquette
  3      Georgia State
  4           Kentucky
  5           Kentucky
             ...
  451           Dayton
  453           Butler
  454         Kentucky
  455         Kentucky
  456           Kansas
  Name: College, Length: 336, dtype: object),
 ('31 to 40',
  19         Georgia Tech
  31                 UNLV
  33             Syracuse
  34             Kentucky
  43             Kentucky
               ...
  406            Kentucky
  413              Kansas
  415           Villanova
  420            Kentucky
  434    Central Michigan
  Name: College, Length: 89, dtype: object)]
```

In [152…
```
df['Height'].value_counts()
```

Out[152…
```
6-9     57
6-10    45
6-7     44
6-8     43
6-6     41
6-11    40
6-5     31
6-3     31
6-4     29
7-0     27
6-1     16
6-2     14
6-0     10
7-1      7
7-3      4
5-11     3
7-2      3
5-9      1
Name: Height, dtype: int64
```

In [153…
```
df.groupby(df['Height']).get_group('6-10')
```

Out[153…

| | Name | Team | Number | Position | Age | Height | Weight | College | Salary |
|---|---|---|---|---|---|---|---|---|---|
| **4** | Jonas Jerebko | Boston Celtics | 8.0 | PF | 29.0 | 6-10 | 231.0 | Kentucky | 5000000.0 |
| **25** | Willie Reed | Brooklyn Nets | 33.0 | PF | 26.0 | 6-10 | 220.0 | Saint Louis | 947276.0 |
| **26** | Thomas Robinson | Brooklyn Nets | 41.0 | PF | 25.0 | 6-10 | 237.0 | Kansas | 981348.0 |
| **27** | Henry Sims | Brooklyn Nets | 14.0 | C | 26.0 | 6-10 | 248.0 | Georgetown | 947276.0 |
| **39** | Kyle O'Quinn | New York Knicks | 9.0 | PF | 26.0 | 6-10 | 250.0 | Norfolk State | 3750000.0 |
| **41** | Kevin Seraphin | New York Knicks | 1.0 | C | 26.0 | 6-10 | 278.0 | Kentucky | 2814000.0 |
| **51** | Richaun Holmes | Philadelphia 76ers | 22.0 | PF | 22.0 | 6-10 | 245.0 | Bowling Green | 1074169.0 |
| **88** | Marreese Speights | Golden State Warriors | 5.0 | C | 28.0 | 6-10 | 255.0 | Florida | 3815000.0 |
| **96** | Blake Griffin | Los Angeles Clippers | 32.0 | PF | 27.0 | 6-10 | 251.0 | Oklahoma | 18907726.0 |
| **129** | Jon Leuer | Phoenix Suns | 30.0 | PF | 27.0 | 6-10 | 228.0 | Wisconsin | 1035000.0 |
| **149** | Eric Moreland | Sacramento Kings | 25.0 | PF | 24.0 | 6-10 | 238.0 | Oregon State | 845059.0 |
| **155** | Cristiano Felicio | Chicago Bulls | 6.0 | PF | 23.0 | 6-10 | 275.0 | Kentucky | 525093.0 |
| **160** | Nikola Mirotic | Chicago Bulls | 44.0 | PF | 25.0 | 6-10 | 220.0 | Kentucky | 5543725.0 |
| **174** | Kevin Love | Cleveland Cavaliers | 0.0 | PF | 27.0 | 6-10 | 251.0 | UCLA | 19689000.0 |

| | Name | Team | Number | Position | Age | Height | Weight | College | Salary |
|---|---|---|---|---|---|---|---|---|---|
| **182** | Aron Baynes | Detroit Pistons | 12.0 | C | 29.0 | 6-10 | 260.0 | Washington State | 6500000.0 |
| **201** | Jordan Hill | Indiana Pacers | 27.0 | C | 28.0 | 6-10 | 235.0 | Arizona | 4000000.0 |
| **221** | Steve Novak | Milwaukee Bucks | 6.0 | SF | 32.0 | 6-10 | 225.0 | Marquette | 295327.0 |
| **238** | Chandler Parsons | Dallas Mavericks | 25.0 | SF | 27.0 | 6-10 | 230.0 | Florida | 15361500.0 |
| **243** | Michael Beasley | Houston Rockets | 8.0 | SF | 27.0 | 6-10 | 235.0 | Kansas State | 306527.0 |
| **246** | Clint Capela | Houston Rockets | 15.0 | PF | 22.0 | 6-10 | 240.0 | Kentucky | 1242720.0 |
| **259** | Chris Andersen | Memphis Grizzlies | 7.0 | PF | 37.0 | 6-10 | 245.0 | Blinn College | 5000000.0 |
| **268** | Jarell Martin | Memphis Grizzlies | 10.0 | PF | 22.0 | 6-10 | 239.0 | LSU | 1230840.0 |
| **274** | Brandan Wright | Memphis Grizzlies | 34.0 | PF | 28.0 | 6-10 | 210.0 | North Carolina | 5464000.0 |
| **276** | Ryan Anderson | New Orleans Pelicans | 33.0 | PF | 28.0 | 6-10 | 240.0 | California | 8500000.0 |
| **281** | Anthony Davis | New Orleans Pelicans | 23.0 | PF | 23.0 | 6-10 | 253.0 | Kentucky | 7070730.0 |
| **292** | Kendrick Perkins | New Orleans Pelicans | 5.0 | C | 31.0 | 6-10 | 270.0 | Kentucky | 947276.0 |
| **296** | Matt Bonner | San Antonio Spurs | 15.0 | C | 36.0 | 6-10 | 235.0 | Florida | 947276.0 |
| **312** | Al Horford | Atlanta Hawks | 15.0 | C | 30.0 | 6-10 | 245.0 | Florida | 12000000.0 |
| **330** | Al Jefferson | Charlotte Hornets | 25.0 | C | 31.0 | 6-10 | 289.0 | Kentucky | 13500000.0 |
| **346** | Josh McRoberts | Miami Heat | 4.0 | PF | 29.0 | 6-10 | 240.0 | Duke | 5543725.0 |
| **348** | Amar'e Stoudemire | Miami Heat | 5.0 | PF | 33.0 | 6-10 | 245.0 | Kentucky | 947276.0 |
| **358** | Ersan Ilyasova | Orlando Magic | 7.0 | PF | 29.0 | 6-10 | 235.0 | Kentucky | 7900000.0 |
| **372** | Drew Gooden | Washington Wizards | 90.0 | PF | 34.0 | 6-10 | 250.0 | Kansas | 3300000.0 |
| **376** | Markieff Morris | Washington Wizards | 5.0 | PF | 26.0 | 6-10 | 245.0 | Kansas | 8000000.0 |
| **388** | Danilo Gallinari | Denver Nuggets | 8.0 | SF | 27.0 | 6-10 | 225.0 | Kentucky | 14000000.0 |
| **390** | Nikola Jokic | Denver Nuggets | 15.0 | C | 21.0 | 6-10 | 250.0 | Kentucky | 1300000.0 |
| **398** | Nemanja Bjelica | Minnesota Timberwolves | 88.0 | PF | 28.0 | 6-10 | 240.0 | Kentucky | 3950001.0 |
| **404** | Adreian Payne | Minnesota Timberwolves | 33.0 | PF | 25.0 | 6-10 | 237.0 | Michigan State | 1938840.0 |
| **413** | Nick Collison | Oklahoma City Thunder | 4.0 | PF | 35.0 | 6-10 | 255.0 | Kansas | 3750000.0 |

| | Name | Team | Number | Position | Age | Height | Weight | College | Salary |
|---|---|---|---|---|---|---|---|---|---|
| **417** | Serge Ibaka | Oklahoma City Thunder | 9.0 | PF | 26.0 | 6-10 | 245.0 | Kentucky | 12250000.0 |
| **419** | Mitch McGary | Oklahoma City Thunder | 33.0 | PF | 24.0 | 6-10 | 255.0 | Michigan | 1463040.0 |
| **420** | Nazr Mohammed | Oklahoma City Thunder | 13.0 | C | 38.0 | 6-10 | 250.0 | Kentucky | 222888.0 |
| **431** | Ed Davis | Portland Trail Blazers | 17.0 | C | 27.0 | 6-10 | 240.0 | North Carolina | 6980802.0 |
| **446** | Derrick Favors | Utah Jazz | 15.0 | PF | 24.0 | 6-10 | 265.0 | Georgia Tech | 12000000.0 |
| **452** | Trey Lyles | Utah Jazz | 41.0 | PF | 20.0 | 6-10 | 234.0 | Kentucky | 2239800.0 |

In [154…

```
list_of_salaries_by_HeightGroup = list(df.groupby('Height')['Salary'])
list_of_salaries_by_HeightGroup
```

Out[154…

```
[('5-11',
  22     1500000.0
  130      55722.0
  203     211744.0
  Name: Salary, dtype: float64),
 ('5-9',
  11    6912869.0
  Name: Salary, dtype: float64),
 ('6-0',
  47       947276.0
  57       947276.0
  67     12000000.0
  100    21468695.0
  142     5013559.0
  152     2250000.0
  228     4290000.0
  305     3578947.0
  384     3000000.0
  394     4345000.0
  Name: Salary, dtype: float64),
 ('6-1',
  118     7000000.0
  121    13500000.0
  150     9500000.0
  180     2100000.0
  230     3950313.0
  244     6486486.0
  262     9588426.0
  263      700902.0
  286      845059.0
  318     1763400.0
  336    12000000.0
  359     8344497.0
  361     1294440.0
  440     2854940.0
  443     2658240.0
  454      900000.0
  Name: Salary, dtype: float64),
 ('6-10',
  4       5000000.0
  25       947276.0
  26       981348.0
```

```
 27       947276.0
 39      3750000.0
 41      2814000.0
 51      1074169.0
 88      3815000.0
 96     18907726.0
 129     1035000.0
 149      845059.0
 155      525093.0
 160     5543725.0
 174    19689000.0
 182     6500000.0
 201     4000000.0
 221      295327.0
 238    15361500.0
 243      306527.0
 246     1242720.0
 259     5000000.0
 268     1230840.0
 274     5464000.0
 276     8500000.0
 281     7070730.0
 292      947276.0
 296      947276.0
 312    12000000.0
 330    13500000.0
 346     5543725.0
 348      947276.0
 358     7900000.0
 372     3300000.0
 376     8000000.0
 388    14000000.0
 390     1300000.0
 398     3950001.0
 404     1938840.0
 413     3750000.0
 417    12250000.0
 419     1463040.0
 420      222888.0
 431     6980802.0
 446    12000000.0
 452     2239800.0
 Name: Salary, dtype: float64),
('6-11',
 24      1140240.0
 55      3457800.0
 56      4582680.0
 60       525093.0
 73       245177.0
 81      2008748.0
 90       289755.0
 91      1100602.0
 98     19689000.0
 113     1724250.0
 143    15851950.0
 162    13400000.0
 163     1391160.0
 167     8193029.0
 173     1276000.0
 188     3272091.0
 204     4000000.0
 208     2357760.0
 209      845059.0
 211     1953960.0
```

```
 216     2943221.0
 220    16407500.0
 224     2109294.0
 237     5200000.0
 239      845059.0
 240      947276.0
 251    22359364.0
 294    19689000.0
 298     5250000.0
 316      947276.0
 321     9756250.0
 339    22192730.0
 373    11217391.0
 375    13000000.0
 391     1709719.0
 399     1474440.0
 400     8500000.0
 405    12100000.0
 418    16407500.0
 439     1415520.0
 Name: Salary, dtype: float64),
('6-2',
 0       7730337.0
 8       1824360.0
 36       845059.0
 54       525093.0
 131      947276.0
 144      947276.0
 210     1007026.0
 256      947276.0
 279     3036927.0
 283     1164858.0
 306    13437500.0
 323     8000000.0
 367     5000000.0
 401     1282080.0
 Name: Salary, dtype: float64),
('6-3',
 16       845059.0
 19      6300000.0
 28       947276.0
 34      7402812.0
 66      7000000.0
 76      2500000.0
 79       947276.0
 80     11370786.0
 102      947726.0
 112      525093.0
 127    13500000.0
 164    20093064.0
 168    16407501.0
 183     2170465.0
 191    13913044.0
 198    10300000.0
 200     8000000.0
 212     3000000.0
 215     1662360.0
 231     4053446.0
 241     5378974.0
 248      200600.0
 304      250750.0
 326      189455.0
 335     2139000.0
 341    14783000.0
```

```
 379     2170465.0
 422     2021520.0
 426    16744218.0
 436     4236287.0
 453     2433333.0
 Name: Salary, dtype: float64),
('6-4',
 9       3431040.0
 17      1500000.0
 21       134215.0
 37      1572360.0
 53      2144772.0
 70       650000.0
 103     7085000.0
 104     3110796.0
 126      981348.0
 161     1015421.0
 166     1147276.0
 193     6270000.0
 214      947276.0
 258     5158539.0
 288    15514031.0
 290    10595507.0
 311     2854940.0
 325      947276.0
 345      845059.0
 349    20000000.0
 363     5192520.0
 364     2505720.0
 381      200600.0
 382    15851950.0
 389     1584480.0
 407    12700000.0
 415     3135000.0
 425     5138430.0
 437     2525160.0
 Name: Salary, dtype: float64),
('6-5',
 3       1148640.0
 30      8000000.0
 75      1509360.0
 93      5675000.0
 105     1159680.0
 110      845059.0
 116     5103120.0
 125     1160160.0
 138     6060606.0
 148     3156600.0
 175      111196.0
 177     8988765.0
 184      111444.0
 186     2891760.0
 207     7000000.0
 218     8000000.0
 233    16407500.0
 249    15756438.0
 257     1404600.0
 272     9000000.0
 291       55722.0
 309     2000000.0
 317      525093.0
 333     3034356.0
 334     5675000.0
 369     5694674.0
```

```
 393     3102240.0
 402     2148360.0
 421     3344000.0
 429      625093.0
 433     6000000.0
 Name: Salary, dtype: float64),
('6-6',
 1       6796117.0
 13      1749840.0
 45       167406.0
 58      2869440.0
 83     11710456.0
 87      1270964.0
 94       525093.0
 109    25000000.0
 122     2127840.0
 133     5500000.0
 137     1015421.0
 158      947276.0
 178     5000000.0
 187      845059.0
 190      600000.0
 205     4394225.0
 213     2399040.0
 225     6600000.0
 226     1733040.0
 227     1449000.0
 253     3189794.0
 261     4088019.0
 267     1201440.0
 282      169883.0
 285    10734586.0
 287     1320000.0
 299     2814000.0
 300    10000000.0
 307      525093.0
 310     1304520.0
 328      525093.0
 347      525093.0
 360      845059.0
 368     4000000.0
 380     1100602.0
 385     3533333.0
 403     2056920.0
 430      947276.0
 444     9463484.0
 445     3777720.0
 451      981348.0
 Name: Salary, dtype: float64),
('6-7',
 12      3425510.0
 18      1335480.0
 20      1599840.0
 32        30888.0
 43       947276.0
 64     10050000.0
 71      3553917.0
 82     14260870.0
 84      5543725.0
 89     15501000.0
 97      1100602.0
 101     3376000.0
 108      700000.0
 119      947276.0
```

```
 120     5219169.0
 123      206192.0
 136      981348.0
 139     1449187.0
 153    16407500.0
 165     1535880.0
 170      947276.0
 185     1252440.0
 192     2841960.0
 202     1358880.0
 206     1100000.0
 260     3542500.0
 284      845059.0
 289     1015421.0
 293     3382023.0
 301    16407500.0
 303      200600.0
 314     5746479.0
 320     4000000.0
 332     6331404.0
 342      947276.0
 344      261894.0
 352     2481720.0
 355     2288205.0
 370     4375000.0
 371      561716.0
 377     1920240.0
 416     1140240.0
 423     1210800.0
 438      525093.0
 Name: Salary, dtype: float64),
('6-8',
 6       1170960.0
 15      3425510.0
 29     11235955.0
 33     22875000.0
 35       845059.0
 42      1636842.0
 44      4000000.0
 50       845059.0
 59       947276.0
 63     13600000.0
 77      3873398.0
 99       947276.0
 106     3000000.0
 134     2041080.0
 135       83397.0
 146    12403101.0
 159     2380440.0
 169    22970500.0
 172      947276.0
 195     3000000.0
 217      855000.0
 219    14700000.0
 223     5152440.0
 242     8193030.0
 250     1000000.0
 280     2850000.0
 297     7500000.0
 315    18671659.0
 319     3333333.0
 324    13125306.0
 343     2854940.0
 357     3741480.0
```

```
 378     4662960.0
 386    10449438.0
 387    11235955.0
 392      947276.0
 411     5758680.0
 424     4500000.0
 427      525093.0
 442     4775000.0
 448    15409570.0
 449     1348440.0
 450     2050000.0
 Name: Salary, dtype: float64),
('6-9',
 5      12000000.0
 10      2569260.0
 31      1635476.0
 48      1000000.0
 52      6500000.0
 61      2814000.0
 62      1524000.0
 65      2500000.0
 69      6268675.0
 72      2900000.0
 85      1131960.0
 86       845059.0
 92       111444.0
 95      9650000.0
 107      845059.0
 114     1155600.0
 115     3132240.0
 132     5500000.0
 140     2836186.0
 145      525093.0
 151      845059.0
 154     4500000.0
 157     8500000.0
 179    14260870.0
 181     2500000.0
 189    16000000.0
 194     5000000.0
 196     4050000.0
 197     1007026.0
 199    17120106.0
 222      845059.0
 229     1100602.0
 232     2085671.0
 245     8229375.0
 247     1646400.0
 252     2489530.0
 255      947276.0
 266      845059.0
 271     9638555.0
 278     1100602.0
 295     1142880.0
 308     1499187.0
 313     1000000.0
 327      947276.0
 337     7000000.0
 340    10151612.0
 356     4171680.0
 362     2380593.0
 374      273038.0
 383     2814000.0
 396      258489.0
```

```
406      947276.0
408     1149500.0
414    20158622.0
428     8042895.0
432     2894059.0
441     2637720.0
Name: Salary, dtype: float64),
('7-0',
7       2165160.0
14      2616975.0
23     19689000.0
38     12650000.0
49      4626960.0
68      1842000.0
74      4660482.0
78     13800000.0
117      981348.0
141     3398280.0
147     7700000.0
156     7448760.0
234     1270964.0
236     8333334.0
254     2288205.0
277     9213483.0
331     2612520.0
338     4204200.0
351      981348.0
354      947276.0
365     4300000.0
366    11250000.0
395     1842000.0
410     5703600.0
412     2279040.0
434     5016000.0
456      947276.0
Name: Salary, dtype: float64),
('7-1',
124    13000000.0
128     3807120.0
176     4950000.0
265    19688000.0
329     6110034.0
435     3075880.0
447     1175880.0
Name: Salary, dtype: float64),
('7-2',
111    15592217.0
235      525093.0
275     4389607.0
Name: Salary, dtype: float64),
('7-3',
40     4131720.0
302    1200000.0
322    1000000.0
455    2900000.0
Name: Salary, dtype: float64)]
```

In [155…

```
df.groupby('Height')['Salary'].describe()
```

Out[155…

| Height | count | mean | std | min | 25% | 50% | 75% |
|---|---|---|---|---|---|---|---|

| Height | count | mean | std | min | 25% | 50% | 75% | |
|---|---|---|---|---|---|---|---|---|
| **5-11** | 3.0 | 5.891553e+05 | 7.926627e+05 | 55722.0 | 133733.0 | 211744.0 | 855872.00 | 15000 |
| **5-9** | 1.0 | 6.912869e+06 | NaN | 6912869.0 | 6912869.0 | 6912869.0 | 6912869.00 | 69128 |
| **6-0** | 10.0 | 5.784075e+06 | 6.337144e+06 | 947276.0 | 2437500.0 | 3934473.5 | 4846419.25 | 214686 |
| **6-1** | 16.0 | 5.217919e+06 | 4.286013e+06 | 700902.0 | 1646160.0 | 3402626.5 | 8633372.75 | 135000 |
| **6-10** | 45.0 | 5.200605e+06 | 5.176373e+06 | 222888.0 | 1035000.0 | 3750000.0 | 7070730.00 | 196890 |
| **6-11** | 40.0 | 6.544397e+06 | 6.906416e+06 | 245177.0 | 1362370.0 | 3107656.0 | 11438043.25 | 223593 |
| **6-2** | 14.0 | 3.335362e+06 | 3.861320e+06 | 525093.0 | 947276.0 | 1223469.0 | 4509231.75 | 134375 |
| **6-3** | 31.0 | 5.884952e+06 | 5.848304e+06 | 189455.0 | 1305043.0 | 3000000.0 | 9150000.00 | 200930 |
| **6-4** | 29.0 | 4.646163e+06 | 5.275308e+06 | 134215.0 | 1015421.0 | 2525160.0 | 5192520.00 | 200000 |
| **6-5** | 31.0 | 4.377241e+06 | 4.181469e+06 | 55722.0 | 1159920.0 | 3102240.0 | 6030303.00 | 164075 |
| **6-6** | 41.0 | 3.556182e+06 | 4.570697e+06 | 167406.0 | 947276.0 | 1749840.0 | 4088019.00 | 250000 |
| **6-7** | 44.0 | 3.473986e+06 | 4.383151e+06 | 30888.0 | 947276.0 | 1492533.5 | 3665437.75 | 164075 |
| **6-8** | 43.0 | 5.950412e+06 | 6.133934e+06 | 83397.0 | 1259700.0 | 3425510.0 | 9321234.00 | 229705 |
| **6-9** | 57.0 | 4.133756e+06 | 4.595223e+06 | 111444.0 | 1007026.0 | 2500000.0 | 5500000.00 | 201586 |
| **7-0** | 27.0 | 5.287712e+06 | 4.675298e+06 | 947276.0 | 2003580.0 | 4204200.0 | 7574380.00 | 196890 |
| **7-1** | 7.0 | 7.400988e+06 | 6.587462e+06 | 1175880.0 | 3441500.0 | 4950000.0 | 9555017.00 | 196880 |
| **7-2** | 3.0 | 6.835639e+06 | 7.825718e+06 | 525093.0 | 2457350.0 | 4389607.0 | 9990912.00 | 155922 |
| **7-3** | 4.0 | 2.307930e+06 | 1.484918e+06 | 1000000.0 | 1150000.0 | 2050000.0 | 3207930.00 | 41317 |

# Part2 - IRIS Dataset

In [156…
```
#loading dataset

data1 = pd.read_csv("iris.csv")
df1=pd.DataFrame(data1)
```

In [157…
```
shape = df1.shape
size = df1.size
print("Dimenation of data frame: {}".format(shape))
print("Size of data frame: {}".format(size))
```

```
Dimenation of data frame: (150, 6)
Size of data frame: 900
```

In [158…
```
print(df1.columns)
```

```
Index(['Id', 'SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm', 'PetalWidthC
m',
       'Species'],
      dtype='object')
```

In [159…
```
df1.head(5)
```

Out[159…

| | 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 [160…
```
df1.describe()
```

Out[160…

| | Id | SepalLengthCm | SepalWidthCm | PetalLengthCm | PetalWidthCm |
|---|---|---|---|---|---|
| **count** | 150.000000 | 150.000000 | 150.000000 | 150.000000 | 150.000000 |
| **mean** | 75.500000 | 5.843333 | 3.054000 | 3.758667 | 1.198667 |
| **std** | 43.445368 | 0.828066 | 0.433594 | 1.764420 | 0.763161 |
| **min** | 1.000000 | 4.300000 | 2.000000 | 1.000000 | 0.100000 |
| **25%** | 38.250000 | 5.100000 | 2.800000 | 1.600000 | 0.300000 |
| **50%** | 75.500000 | 5.800000 | 3.000000 | 4.350000 | 1.300000 |
| **75%** | 112.750000 | 6.400000 | 3.300000 | 5.100000 | 1.800000 |
| **max** | 150.000000 | 7.900000 | 4.400000 | 6.900000 | 2.500000 |

In [161…
```
df1.dtypes
```

Out[161…
```
Id                int64
SepalLengthCm    float64
SepalWidthCm     float64
PetalLengthCm    float64
PetalWidthCm     float64
Species           object
dtype: object
```

In [162…
```
df1.isnull().sum()
```

Out[162…
```
Id               0
SepalLengthCm    0
SepalWidthCm     0
PetalLengthCm    0
PetalWidthCm     0
Species          0
dtype: int64
```

## Statistical Analysis

In [163…
```
df1['Species'].unique()
```

Out[163…
```
array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

In [164…
```
print('Iris-setosa')
Iris_setos=(df1['Species']=='Iris-setosa')
df1[Iris_setos].describe()
```

Iris-setosa

Out[164…

| | 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 [165…

```
print('Iris-versicolor')
Iris_versicolor=(df1['Species']=='Iris-versicolor')
df1[Iris_versicolor].describe()
```

Iris-versicolor

Out[165…

| | 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 [166…

```
print('Iris-virginica')
Iris_virginica=(df1['Species']=='Iris-virginica')
df1[Iris_virginica].describe()
```

Iris-virginica

Out[166…

| | 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 |

In [ ]:

In [ ]: