Assignment 1

Introduction to Data Analysis (DATA 1200)

Dheer Parikh

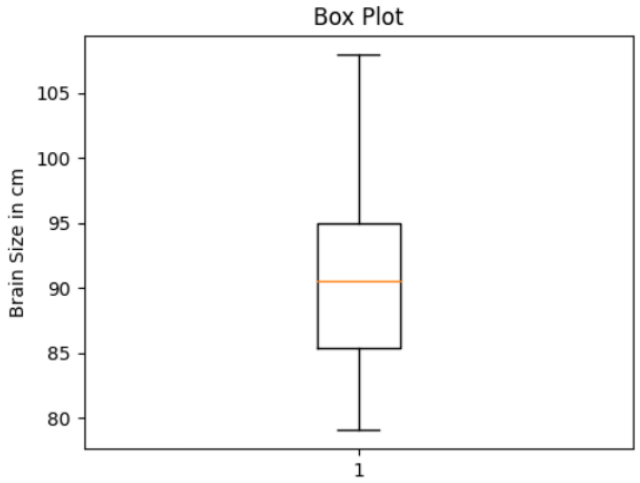
100919437

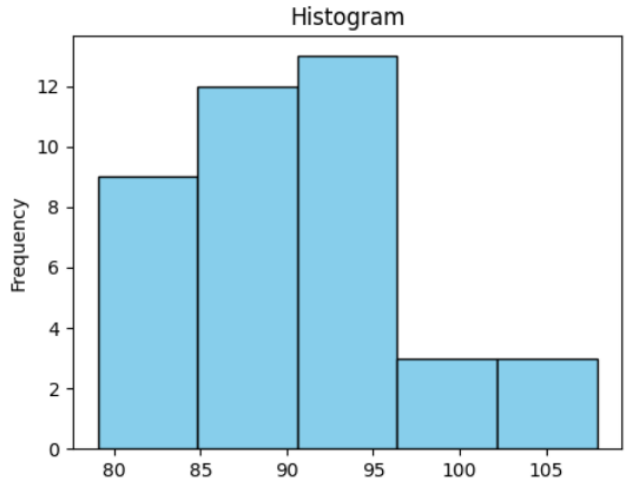
Co-Relation between dependent and independent variables:

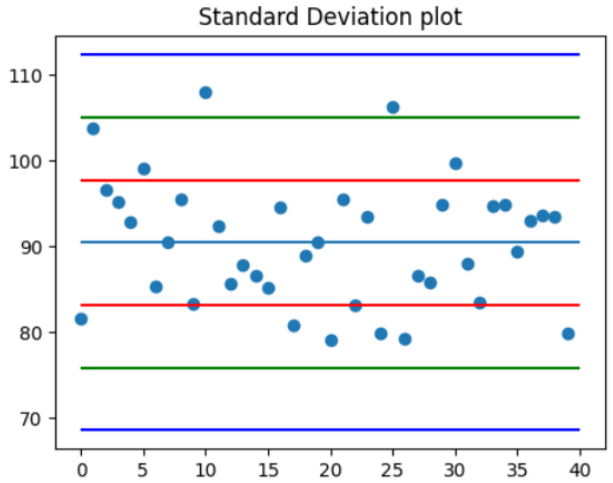
- The correlation map suggests that height and weight have minimum or no correlation with PIQ, whereas brain has a considerably low correlation with PIQ.
- It is also noticeable that height and weight have comparatively the highest correlation of 0.69.
- The correlation of brain size with height is 0.56, which is comparatively low with respect to the correlation between height and weight.
- Additionally, brain size with weight follows with a correlation of 0.50, which is slightly lower with respect to the correlation between brain size and height.

	Brain	Height	Weight	PIQ
Brain	1.000000	0.560079	0.508834	0.391733
Height	0.560079	1.000000	0.689973	-0.050893
Weight	0.508834	0.689973	1.000000	0.015946
PIQ	0.391733	-0.050893	0.015946	1.000000

Detail description of Brain size Variable:



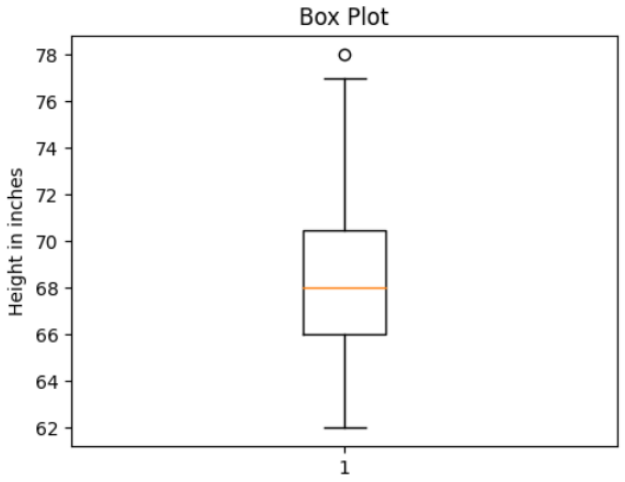


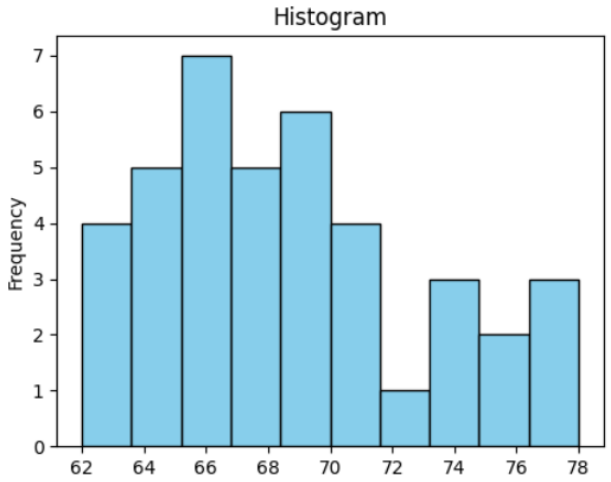


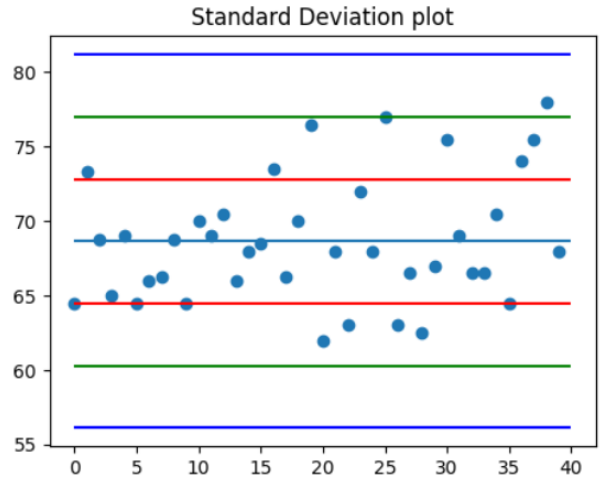

```
<Figure size 640x480 with 0 Axes>
count     40.000000
mean      90.477250
std        7.288643
min       79.060000
25%       85.377500
50%       90.540000
75%       94.945000
max      107.950000
Name: Brain, dtype: float64
```

- From the histogram, it can be seen that the data is left-skewed. This observation is easily verifiable from the box plot and the standard deviation plot as well, indicating that the majority of the data lies in the first quarter. The bins are arranged in groups of five to better understand the skewness of the data points. Examining the box plot, it can be deduced that the interquartile range (IQR) for brain size is between 85 to 95 cm in circumference. Additionally, the difference between the mean and the median is not significant. The smallest size of the brain is 79.06 cm, while the largest size is 107.95 cm. According to the statistics of the data, the standard deviation is 7.29 cm, implying that the spread of data from the mean is 7.29 cm. Around 50% of the studied individuals have a brain size of nearly 90.94 cm. For this variable, there isn't much difference between the mean and the median, so either one can be used to describe this variable. However, for accuracy, it's better to use the mean as there are no outliers present.

Detail description of Height Variable:



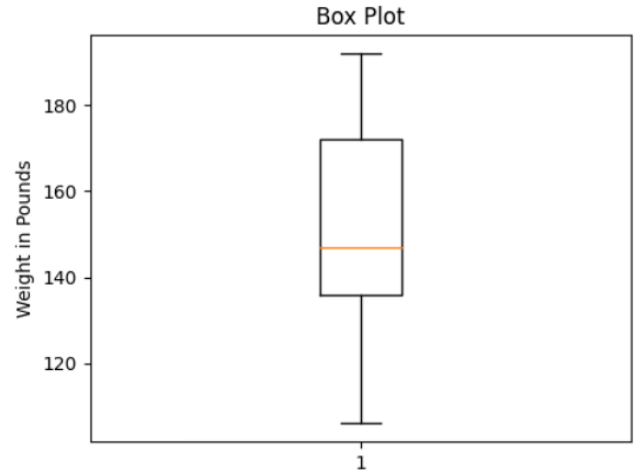


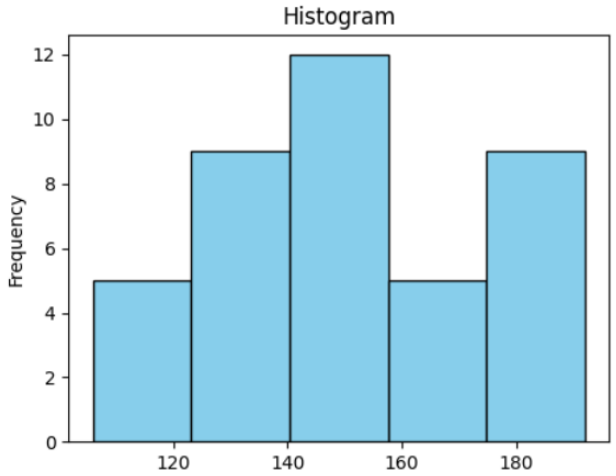


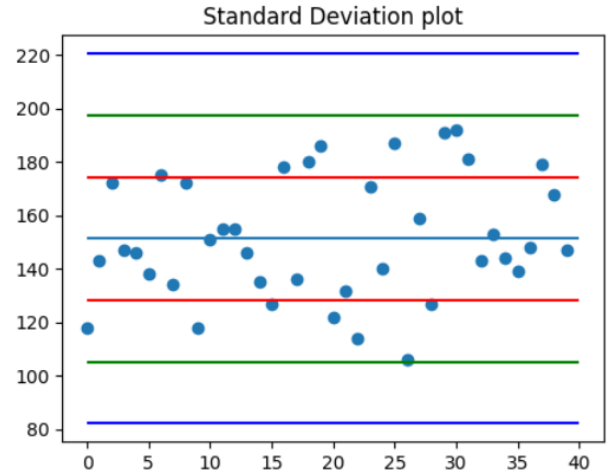

```
<Figure size 640x480 with 0 Axes>
count    40.000000
mean     68.650000
std       4.175631
min      62.000000
25%      66.000000
50%      68.000000
75%      70.500000
max      78.000000
Name: Height, dtype: float64
```

- From the box plot, it can be seen that there is an outlier. It is also noticeable that there is not much difference between the mean and median, but due to the outlier in this case, it is advisable to consider the median to describe the variable. Moreover, the histogram indicates that the variable has left-skewed data points. The majority of the data points lie within the 1st quadrant, and from the box plot, it can be deduced that 50% of data points lie within the range of 66 to 72 inches. There is only one data point present in the 3rd quadrant in the standard deviation plot, and this is considered an outlier in the box plot. Also, it is noticeable that the spread of data is very small, as the range is considerably short.

Detail description of Weight Variable:








```
<Figure size 640x480 with 0 Axes>
count     40.000000
mean     151.375000
std       23.035856
min      106.000000
25%      135.750000
50%      147.000000
75%      172.000000
max      192.000000
Name: Weight, dtype: float64
```

- From the box plot, it can be seen that there are no outliers, which can be verified from the standard deviation plot where no data points can be observed in the 3rd quadrant. In addition to that, 50% of data points lie within the range of 135 to 170 lbs. However, the range is very large, starting from 106 to 192 lbs. It is noticeable that there is not much difference between the mean and median, so either of them can be considered for describing the variable. Since there are no outliers, it is advisable to use the mean to describe the variable. Furthermore, the histogram illustrates that the data is slightly right-skewed.