**STATISTICAL ANALYSIS PROJECT  
CRAB AGE PREDICTION**

TEAM#16

Mariam Khaled Ahmed Mohamed Sobhy

Mennatullah Emad Abdelsamie Shehab

Hana Tarek Mahmoud Ahmed

**Table of contents:**

**Introduction**

**Graphical representations**

Bar chart

Pie chart

Histogram

Dot plot

**Shape of distribution**

**Dispersion of data**

Range

IQR and Box plot

outliers

variance

standard deviation

**Central tendency**

Mean

Mode

Median

**Correlation and Regression**

**Conclusion**

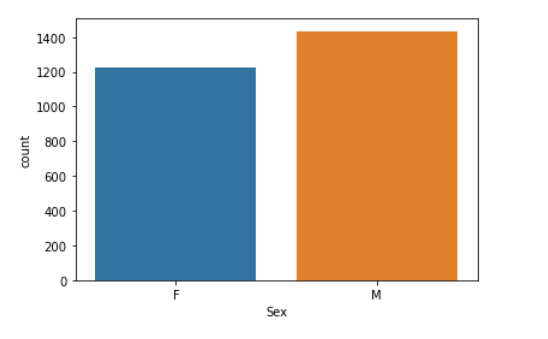
**References**

The dataset is about crabs, where it is used to estimate the age of the crab based on the physical attributes. The project is for the commercial crab farmers to know the right age of the crab helping them decide if and when to harvest the crabs. Beyond a certain age, there is negligible growth in crab's physical characteristics and hence, it is important to time the harvesting to reduce cost and increase profit.

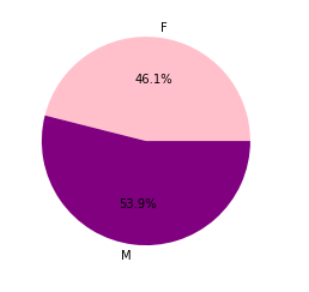
The dataset is analyzed through graphical representations and numerical summaries such as dispersion to describe spread of data, central tendency to show centers of data and correlation to describe the relation between a pair of variables in the dataset and finally regression to express the relationship in the form of an equation.

After removing the observations where the gender was unknown (I), we analyzed the data graphically (Bar Chart, Pie Chart, Histogram, Dot plot, Z-score graph)

Bar Chart (for nominal variables):

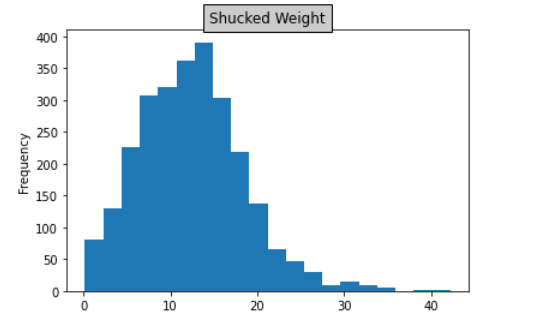


Pie Chart (for nominal variables):



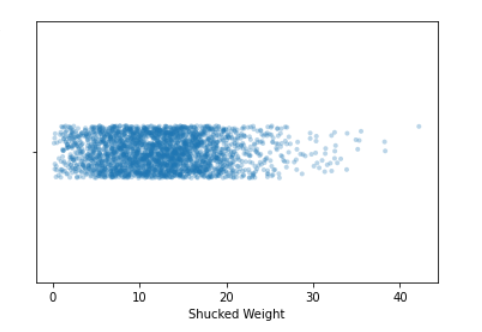
Histogram (for quantitative variables):

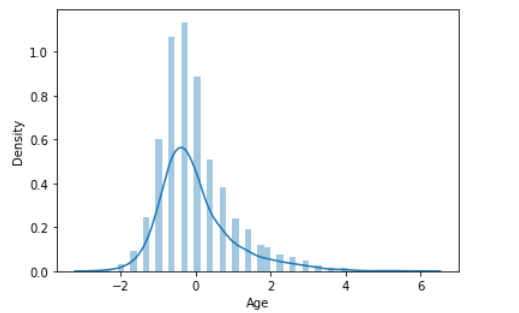
This is an example of the histogram graph, you can find the rest in the jupyter source attached file



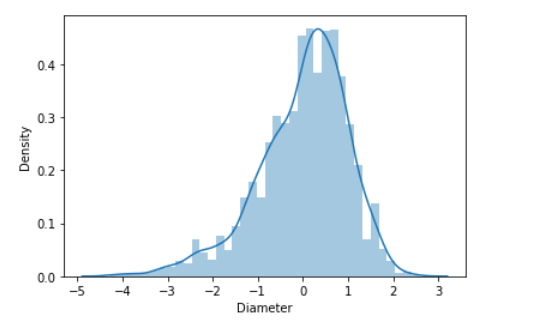
Dot Plot (for quantitative variables):

This is an example of the Dot Plot graph, you can find the rest in the jupyter source attached file

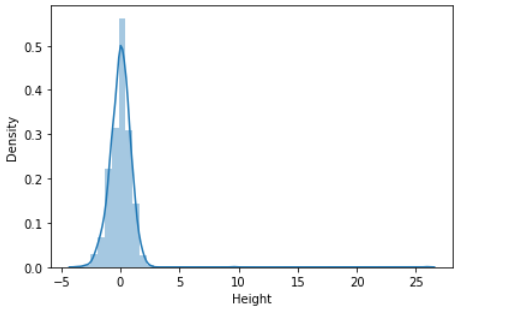


Z-score graphs (for quantitative variables):

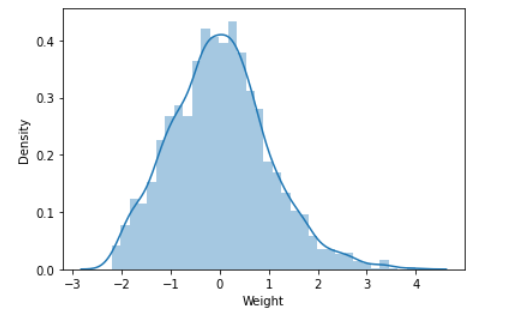
The distribution of Age is right skewed.



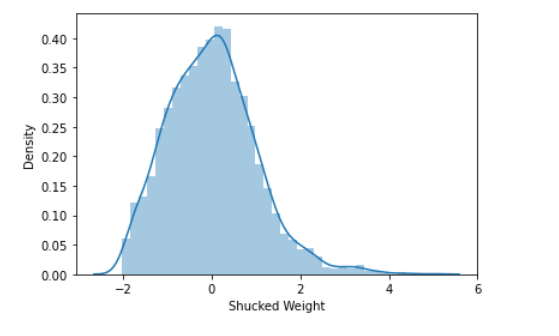
The distribution of Diameter is left skewed.



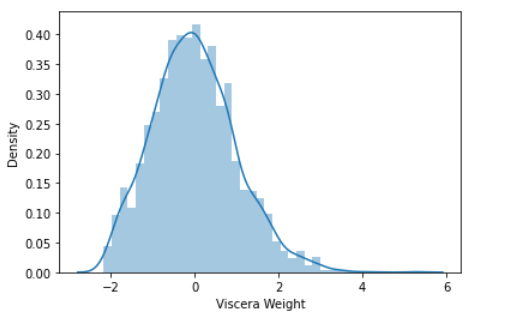
The distribution of Height is left skewed.



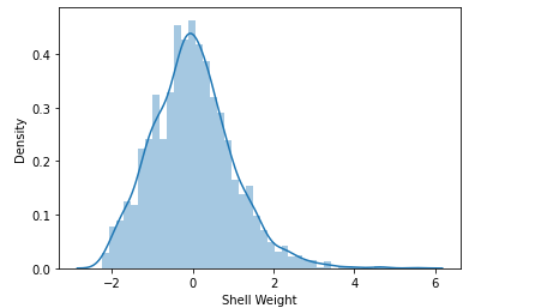
The distribution of Weight is right skewed.



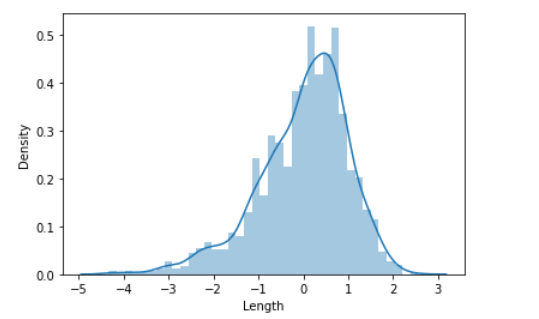
The distribution of Shucked Weight is right skewed.



The distribution of Viscera Weight is right skewed.



The distribution of Shell Weight is right skewed.



The distribution of Length is left skewed.

**Then we analyzed the dispersion of data (Range, IQR, Variance, Standard Deviation)**

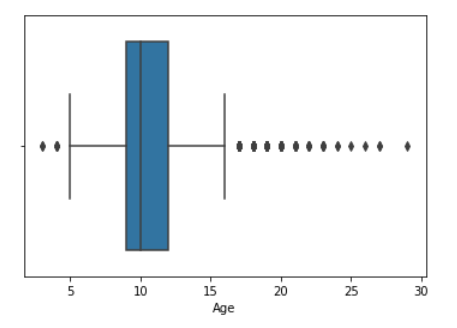
**Ranges:**

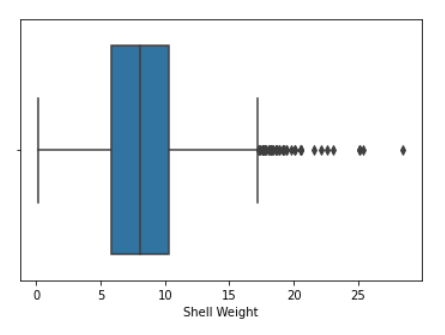
* Range of Age = 26
* Range of Shell Weight = 28.3495
* Range of Viscera Weight = 21.4605715
* Range of Shucked Weight = 41.99978425
* Range of Weight = 79.662095
* Range of Height = 2.7875
* Range of Diameter = 1.35
* Range of Length = 1.6500000000000001

**IQR (with respective box plot and outliers):**

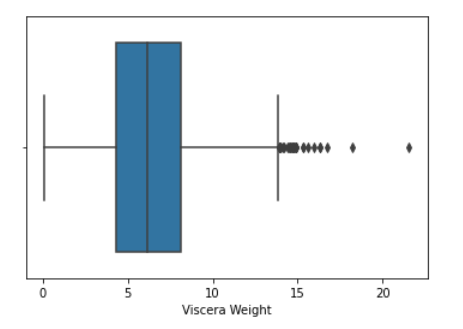
From the IQR We can construct a box plot using the 25th, 50th and 75th outliers or (Q1, Q2, Q3). And from there we can use the Formulas:

(Upper bound>Q3+1.5\*IQR) and (Lower bound<Q1-1.5\*IQR) to get the outliers.

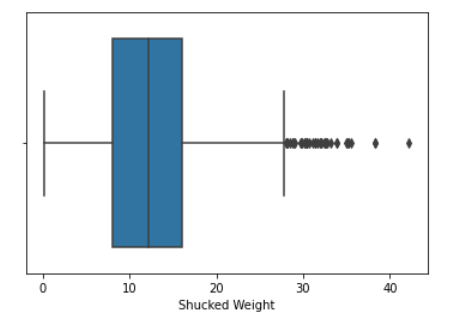
* Age IQR = 3
* Outliers of Age =
* Values<4.5
* Values>16.5



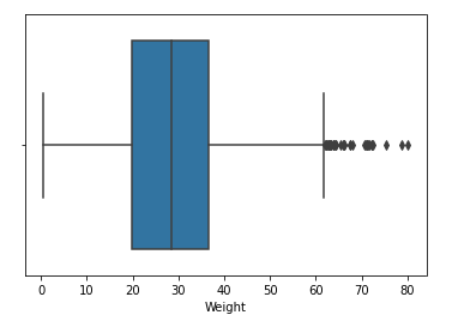
* IQR of Shell Weight=4.543007375
* Outliers of Shell Weight =
* Values>17.156



* IQR of Viscera Weight= 3.8130077499999997
* Outliers of Viscera Weight
* Values>13.855



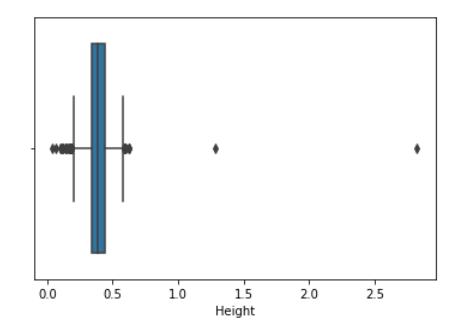
* IQR of Shucked Weight= 7.973296874999999
* Outliers of Shucked Weight=
* Values>28.018



* IQR of Weight=

16.712030249999998

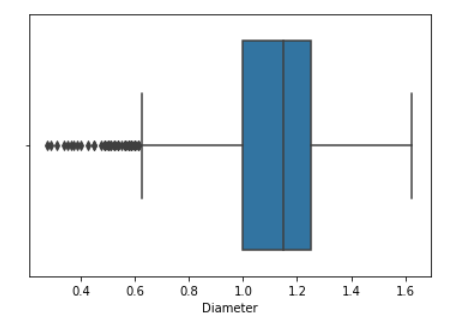
* Outliers of Weight=
* Values>61.617



* IQR of Height=

0.09999999999999998

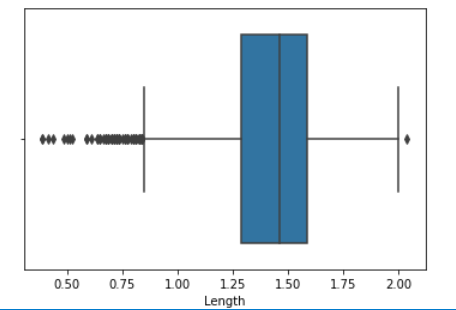
* Outliers of Height=
* Values<0.1875
* Values>0.587



* IQR of Diameter =

0.25

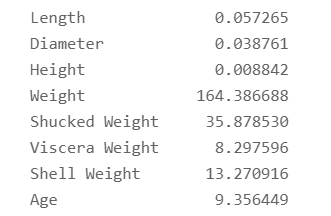
* Outliers of Diameter=
* Values<0.625
* Values>1.625



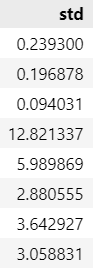
* IQR of Length =

1.6500000000000001

* Outliers of Length =
* Values<0.8
* Values>2.03
* **Variance:**



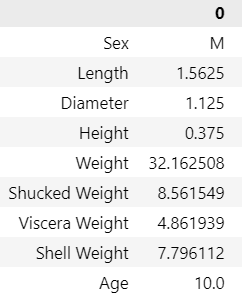
* **Standard Deviation:**

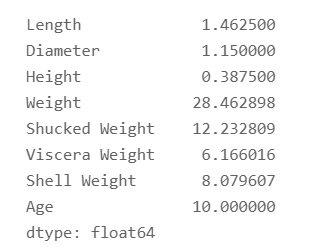


**After that we analyzed the central tendency of the data (Mean, Mode, Median)**

* **Mean:**

****From the outliers above we can conclude that the mean is biased as its affected by the outliers.

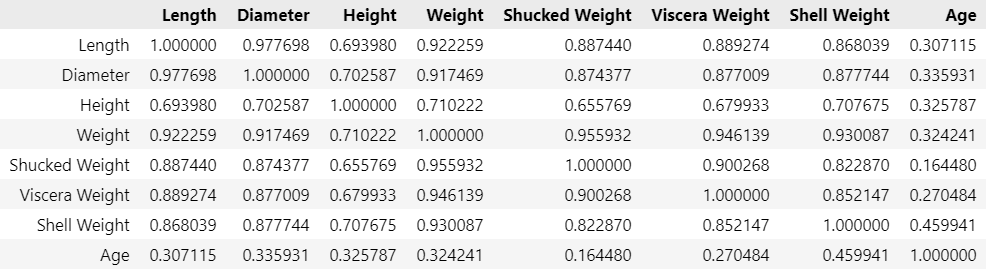
* **Mode:**
* **Median:**

****

**Next, we analyzed the correlation:**

From the correlation matrix below, we can conclude that the Shell Weight has the strongest relation with Age

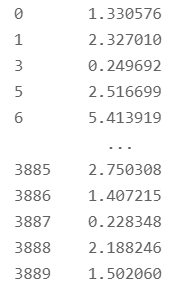
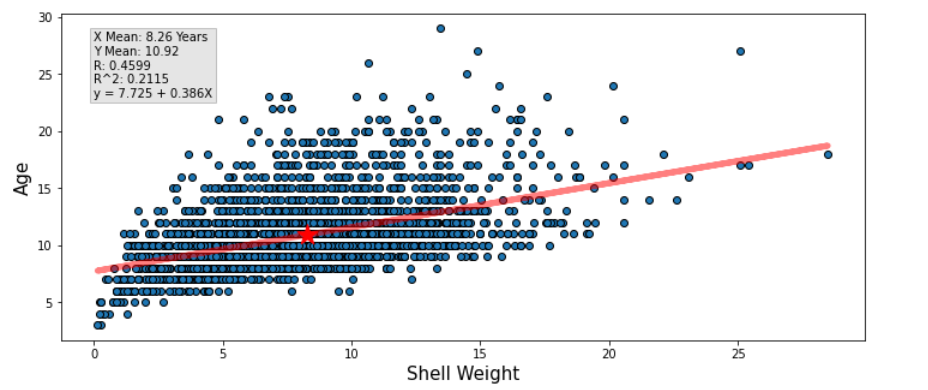
R= 0.4599 which s a positive moderate relation

****

**And finally, we analyzed the regression and made our Scatter plot and predictions:**

* Equation of Y:

y = 7.724845227392523 + 0.386x

* Equation of Y Hat:
* Error (|Y - (Y-Hat) |):
* **Scatter plot**

**References:**

[**https://towardsdatascience.com/simple-linear-regression-in-python-numpy-only-130a988c0212**](https://towardsdatascience.com/simple-linear-regression-in-python-numpy-only-130a988c0212)

[**https://www.kaggle.com/sidhus/crab-age-prediction**](https://www.kaggle.com/sidhus/crab-age-prediction)

**https://www.w3schools.com/python/python\_ml\_mean\_median\_mode.asp**