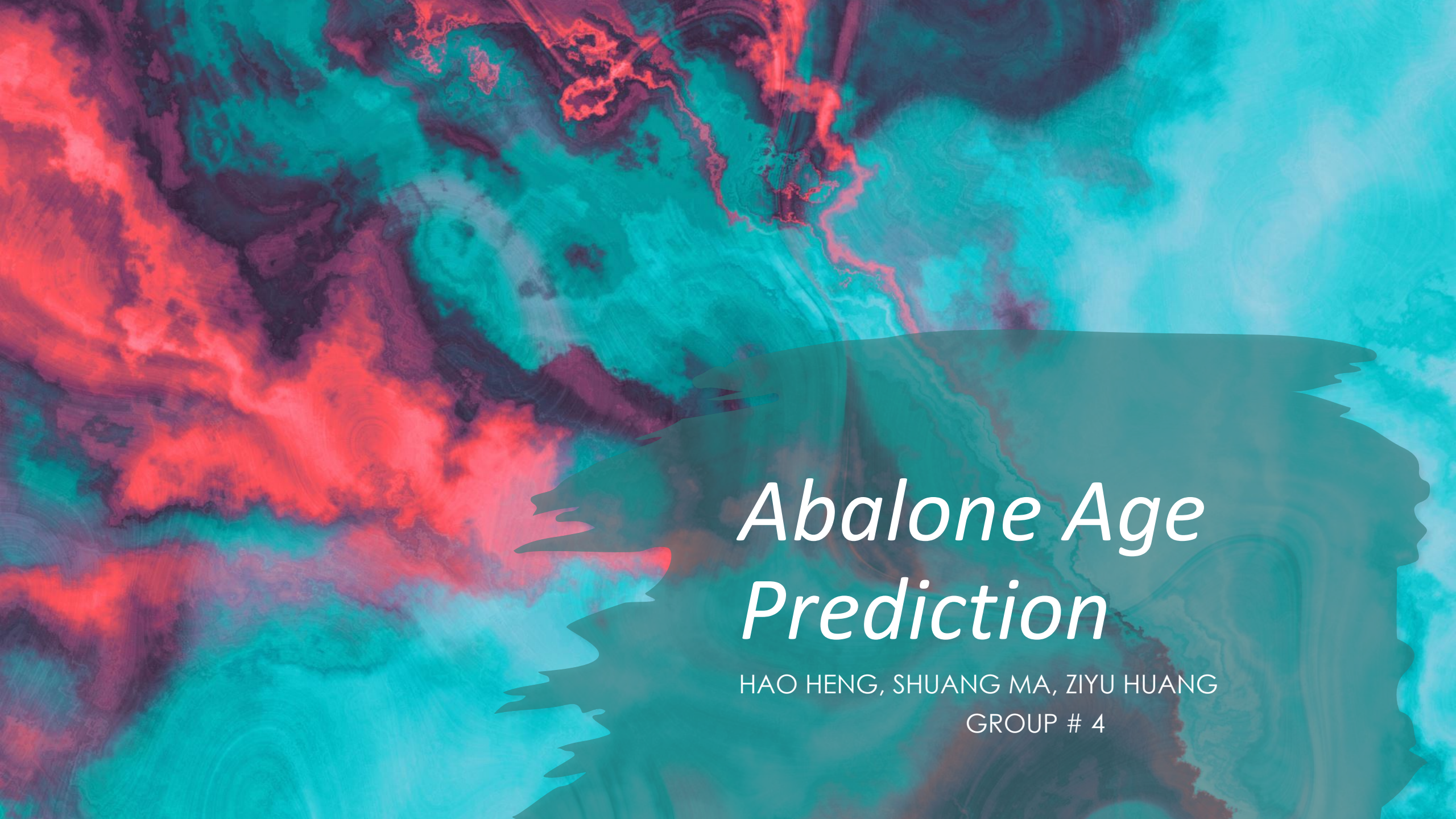


Abalone Age Prediction

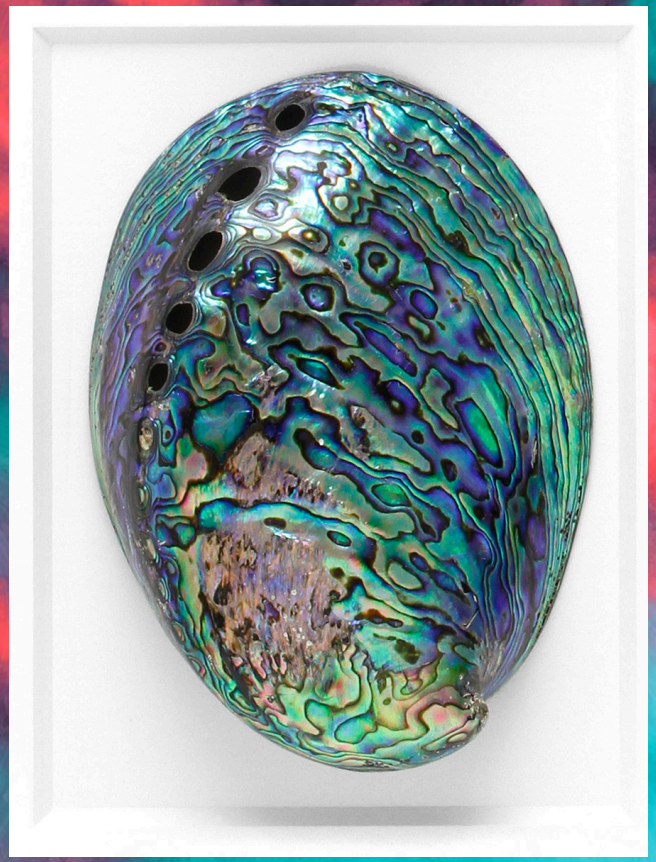
HAO HENG, SHUANG MA, ZIYU HUANG

GROUP # 4

Contents


Background

Background



Abalone is delicacy.

The economy value is closely correlated with the age.

Age of abalone is hard to determine.

Age = Rings + 1.5


Linear Regression
Random Forest
Logistic Regression
Multiple Layer Perceptron


Dataset Description

Variable Introduction

Sex(Categorical): Male, Female, Infant

Length(Continuous) : Longest Shell Measurement

Diameter (Continuous) : Perpendicular to Length

Height (Continuous) : With meat in shell

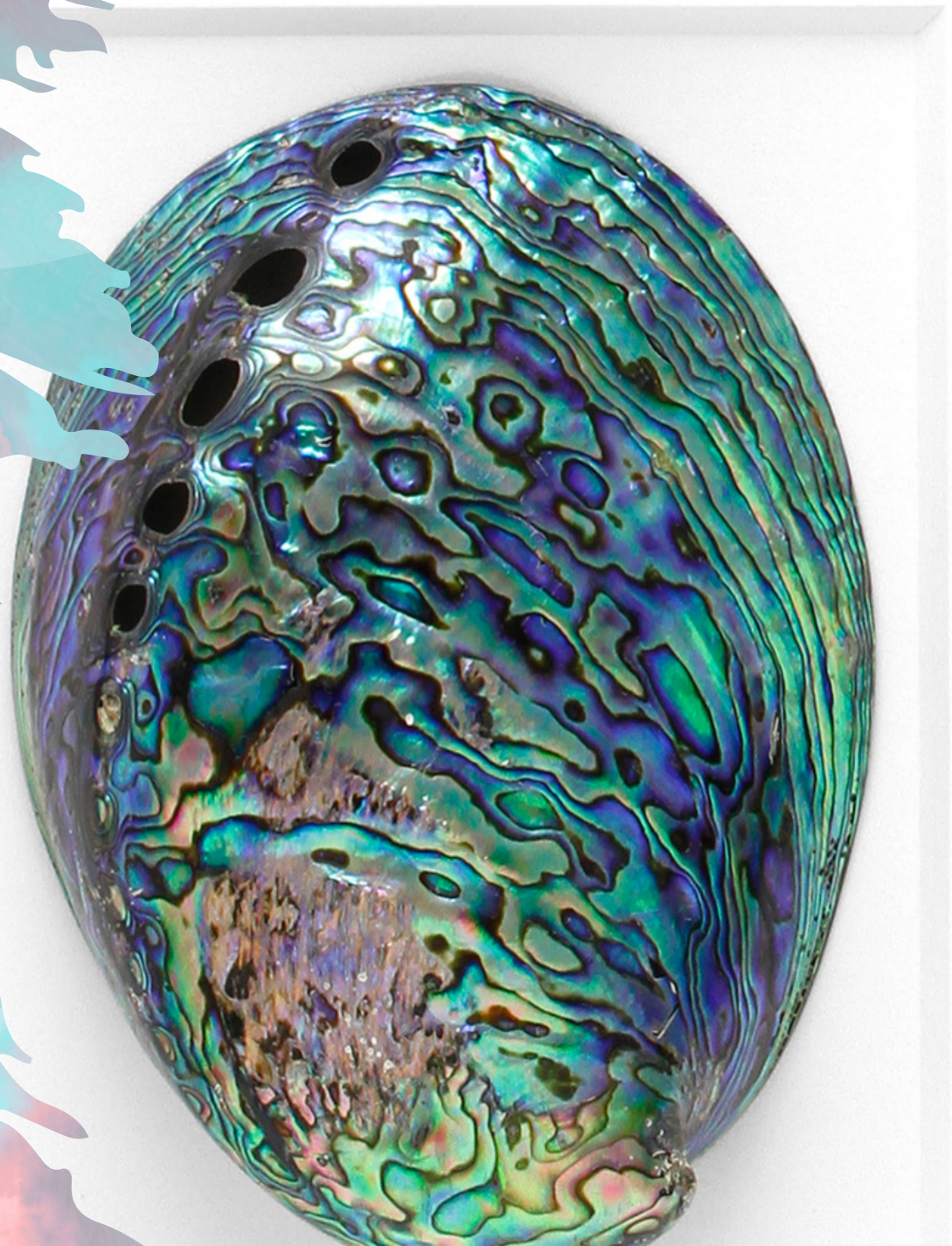
Whole Weight (Continuous) : Whole abalone

Shucked Weight (Continuous) : Weight of meat

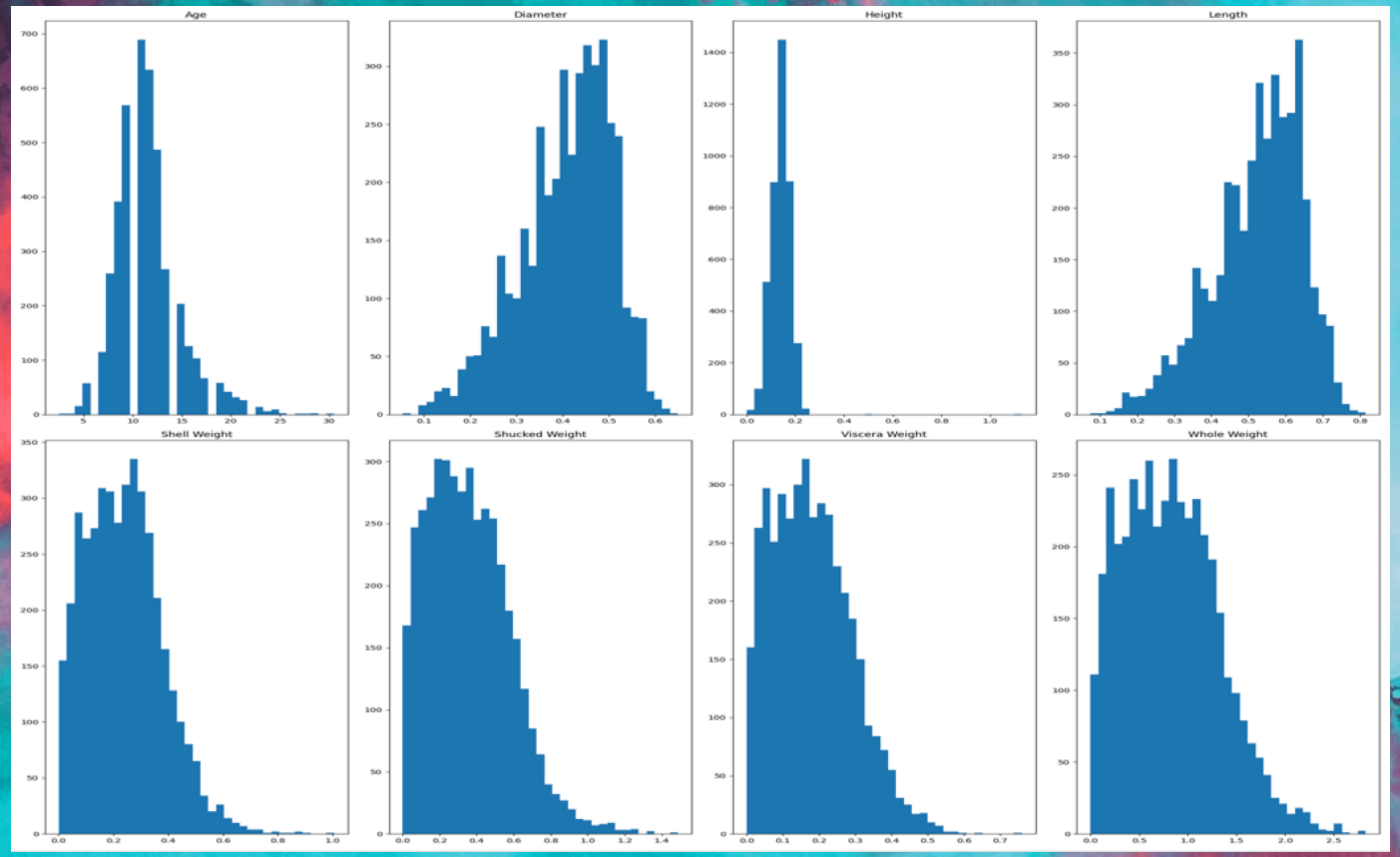
Viscera Weight (Continuous) : Gut weight after bleeding

Shell Weight (Continuous) : After being dried

Rings (Continuous) :+1.5 Gives the age in years



Data Distribution



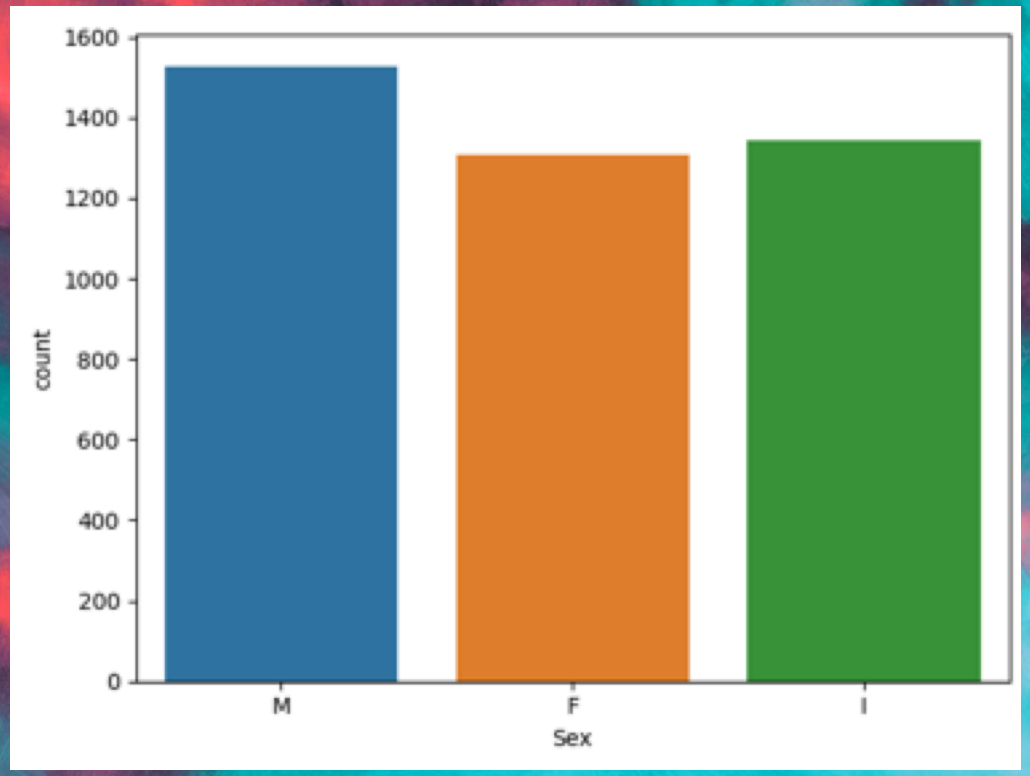
	Length	Diameter	Height	Whole Weight	Shucked Weight
count	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000
mean	0.523992	0.407881	0.139516	0.828742	0.359367
std	0.120093	0.099240	0.041827	0.490389	0.221963
min	0.075000	0.055000	0.000000	0.002000	0.001000
25%	0.450000	0.350000	0.115000	0.441500	0.186000
50%	0.545000	0.425000	0.140000	0.799500	0.336000
75%	0.615000	0.480000	0.165000	1.153000	0.502000
max	0.815000	0.650000	1.130000	2.825500	1.488000

	Viscera Weight	Shell Weight	Age
count	4177.000000	4177.000000	4177.000000
mean	0.180594	0.238831	11.433684
std	0.109614	0.139203	3.224169
min	0.000500	0.001500	2.500000
25%	0.093500	0.130000	9.500000
50%	0.171000	0.234000	10.500000
75%	0.253000	0.329000	12.500000
max	0.760000	1.005000	30.500000

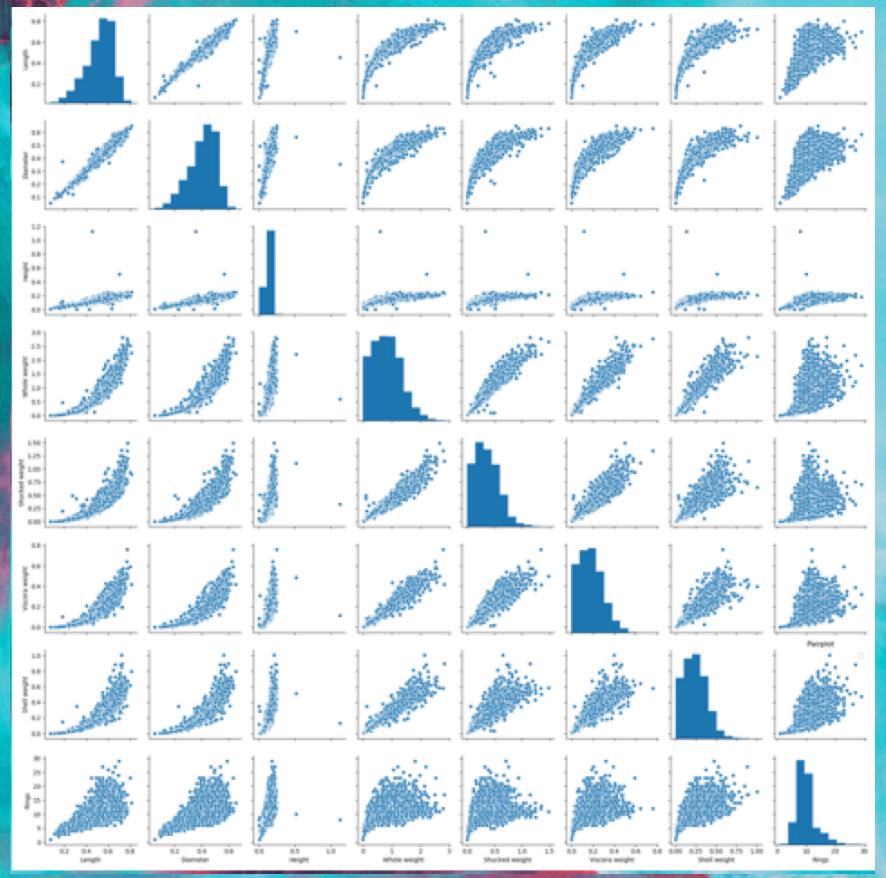
Check missing value

	Missing value	% Missing
Rings	0	0.0
Shell weight	0	0.0
Viscera weight	0	0.0
Shucked weight	0	0.0
Whole weight	0	0.0
Height	0	0.0
Diameter	0	0.0
Length	0	0.0
Sex	0	0.0

Count plot

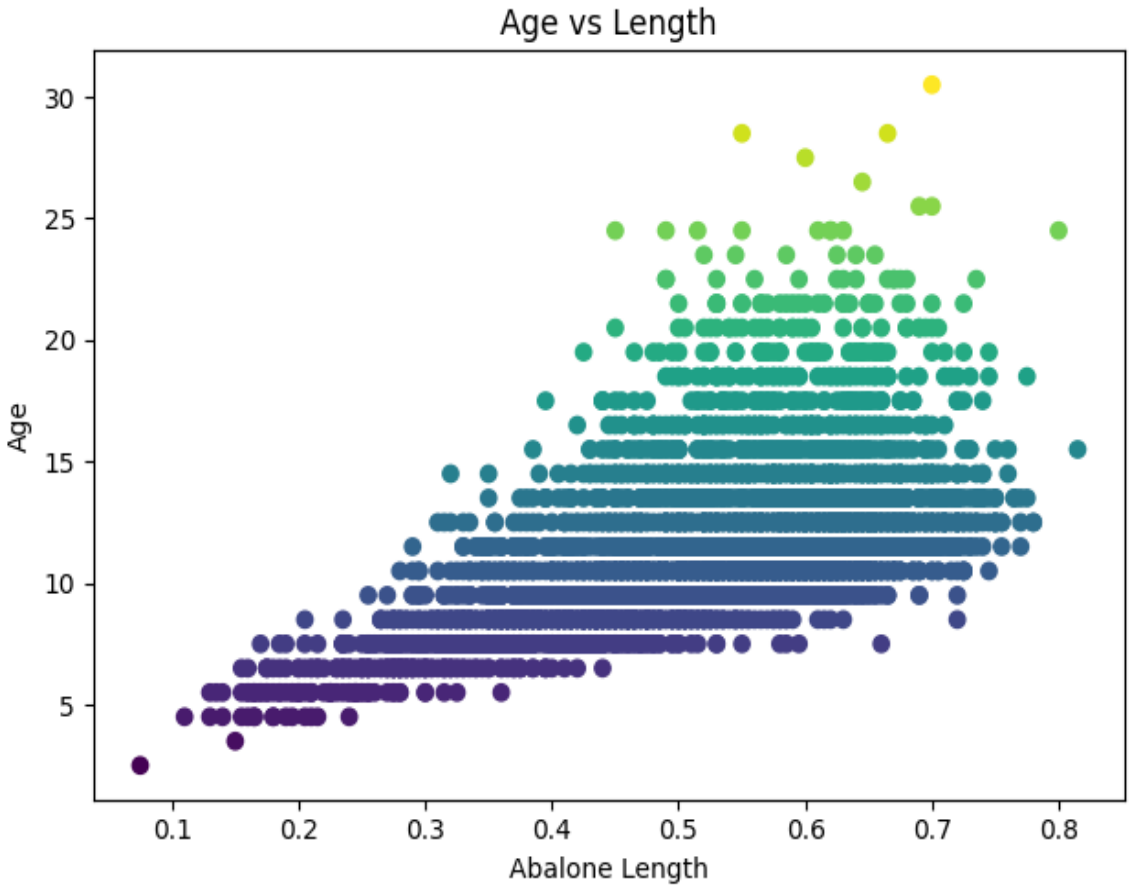


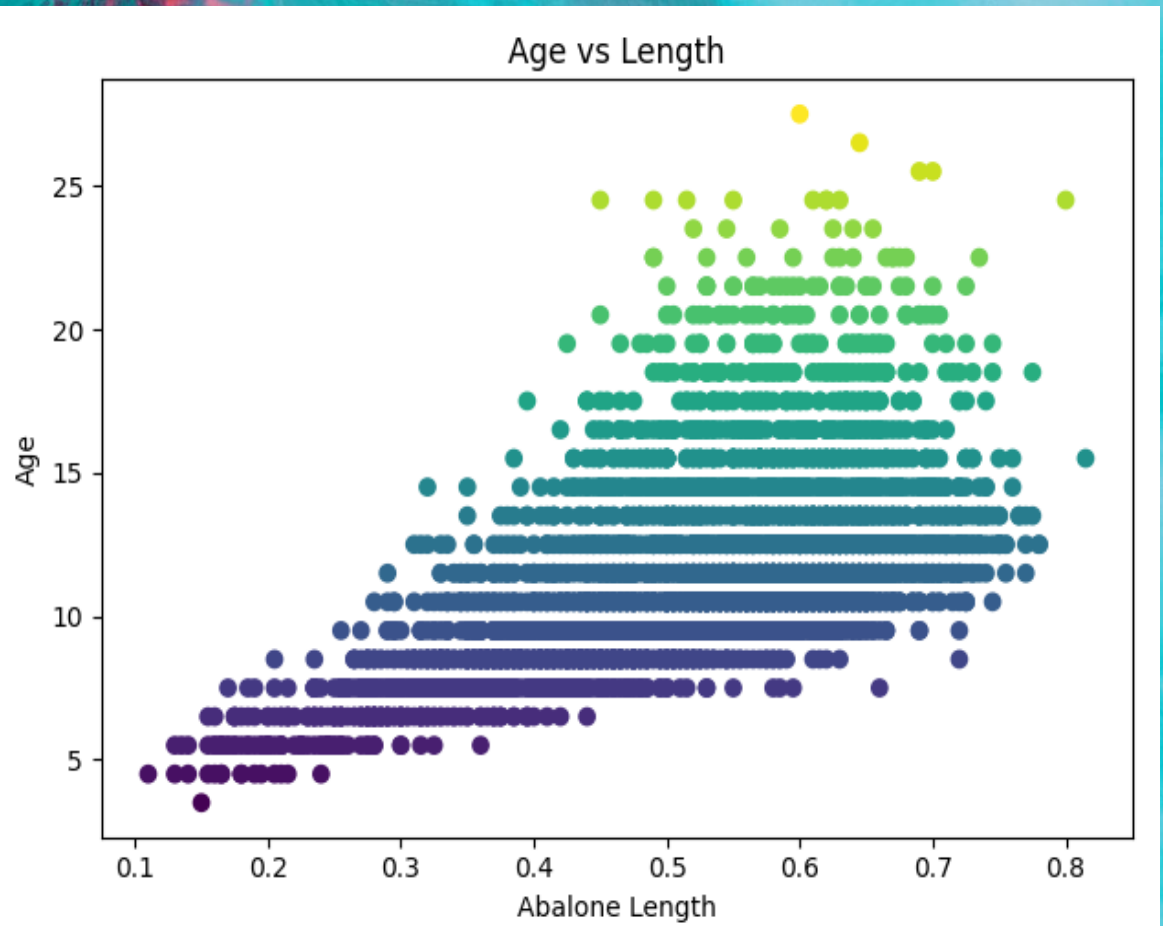
Pair plot



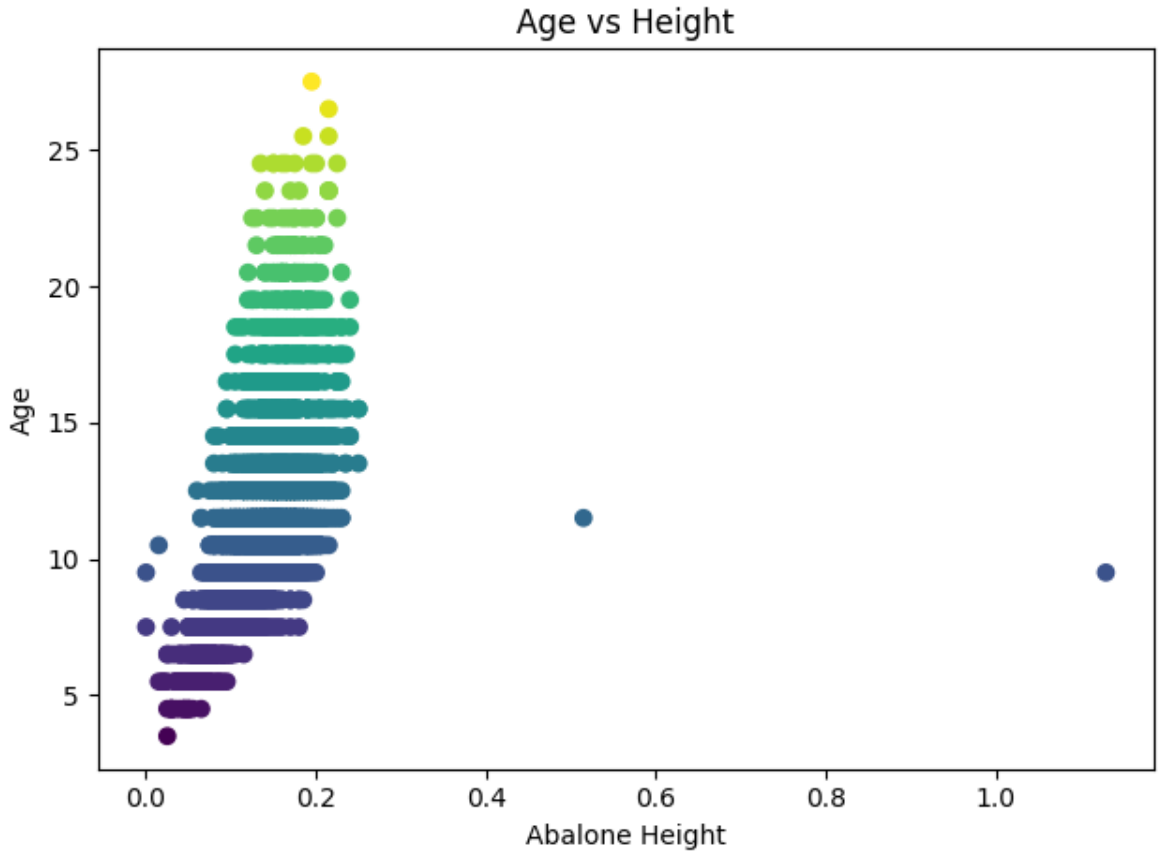

Preprocessing

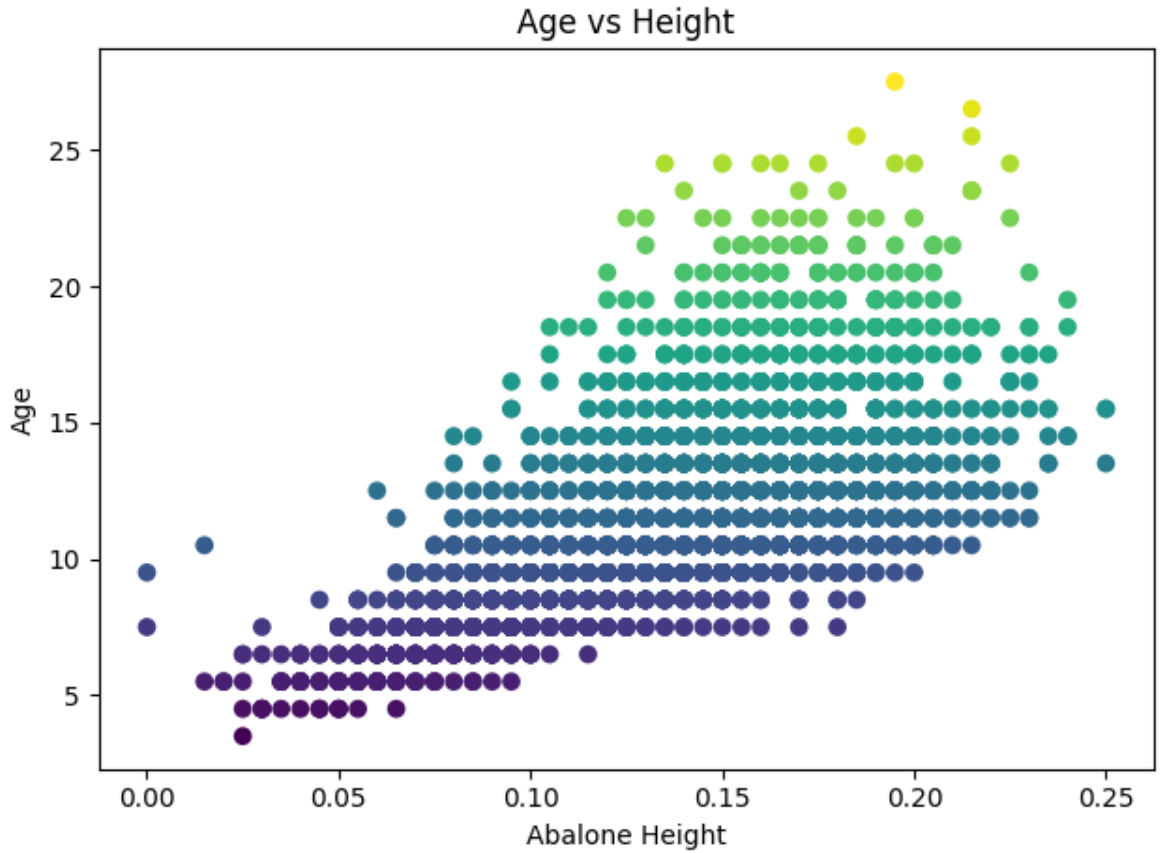
Age vs Length



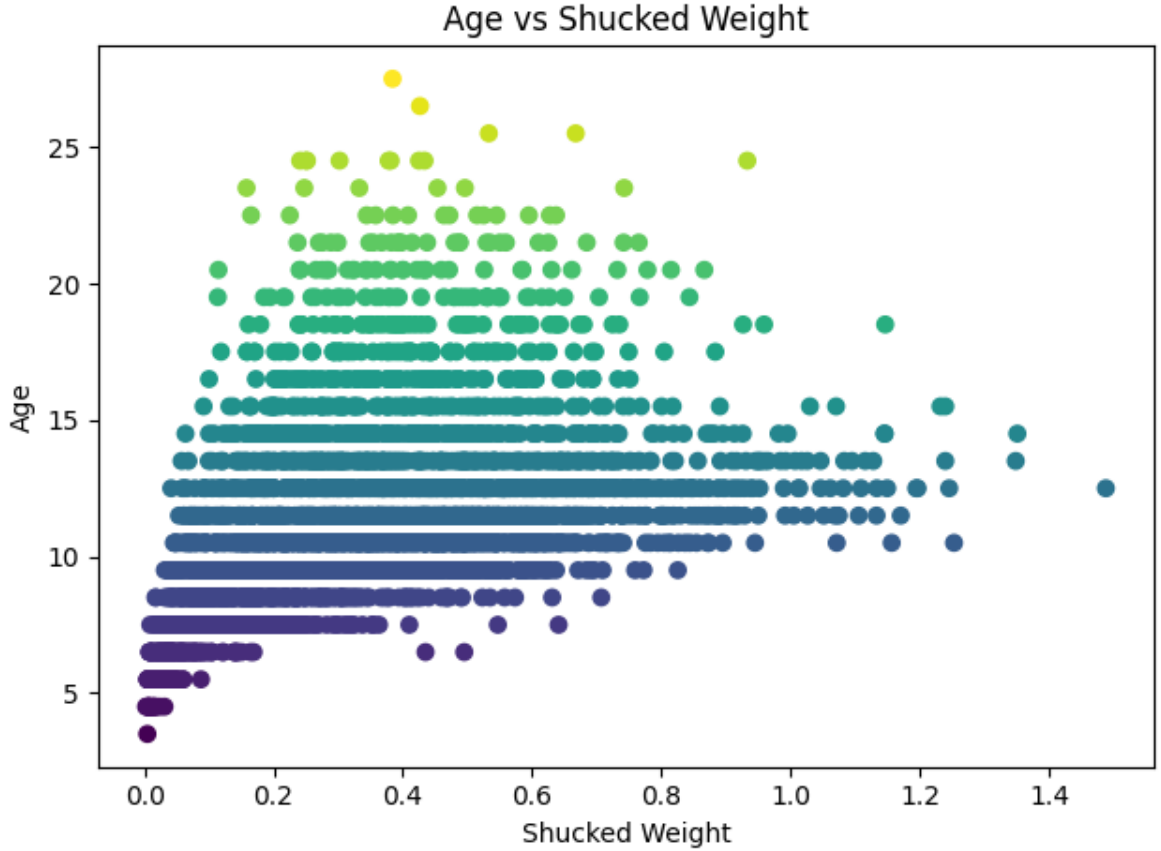


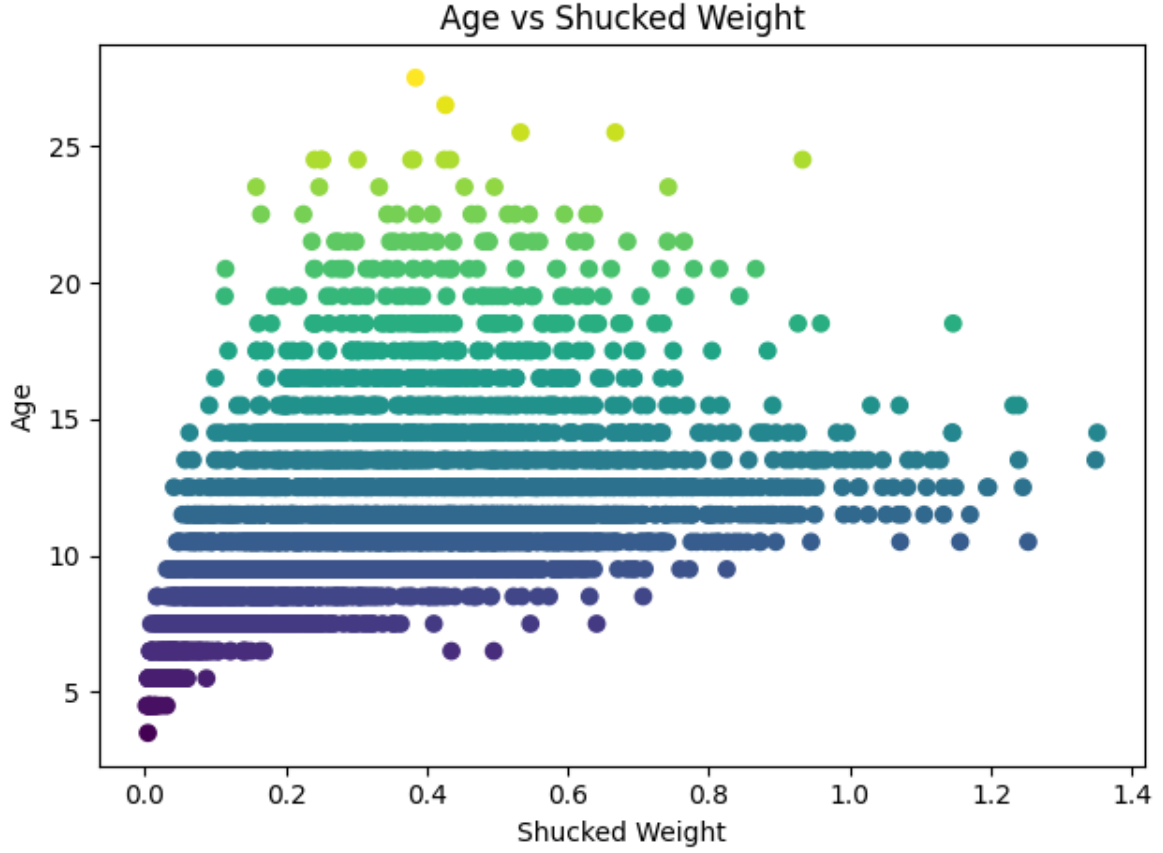
Age vs Height



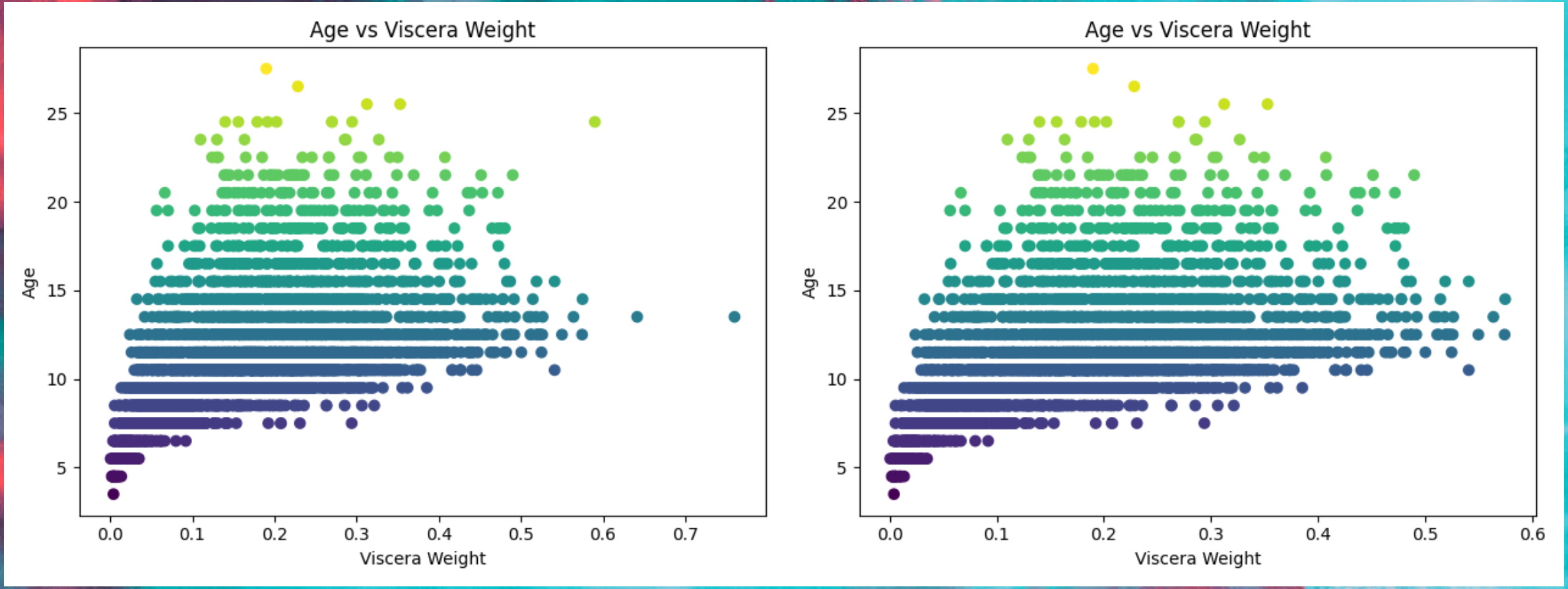


Age vs Shucked Weight

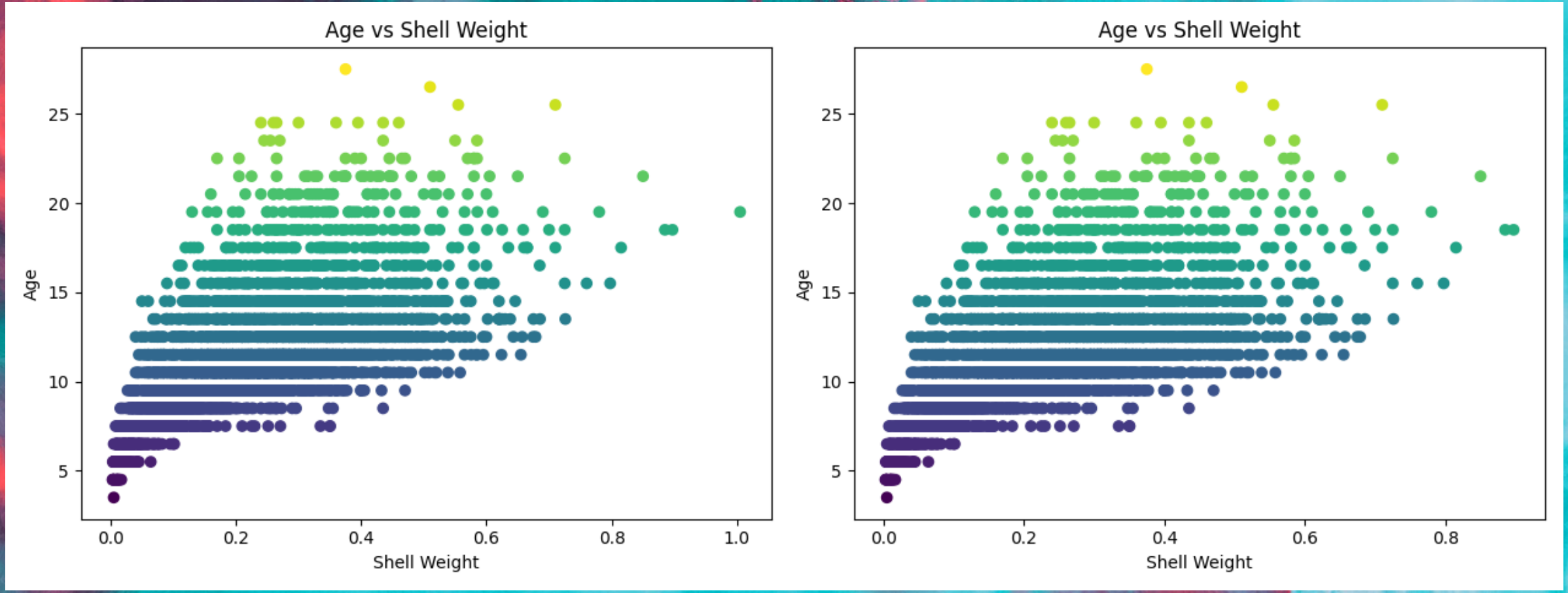




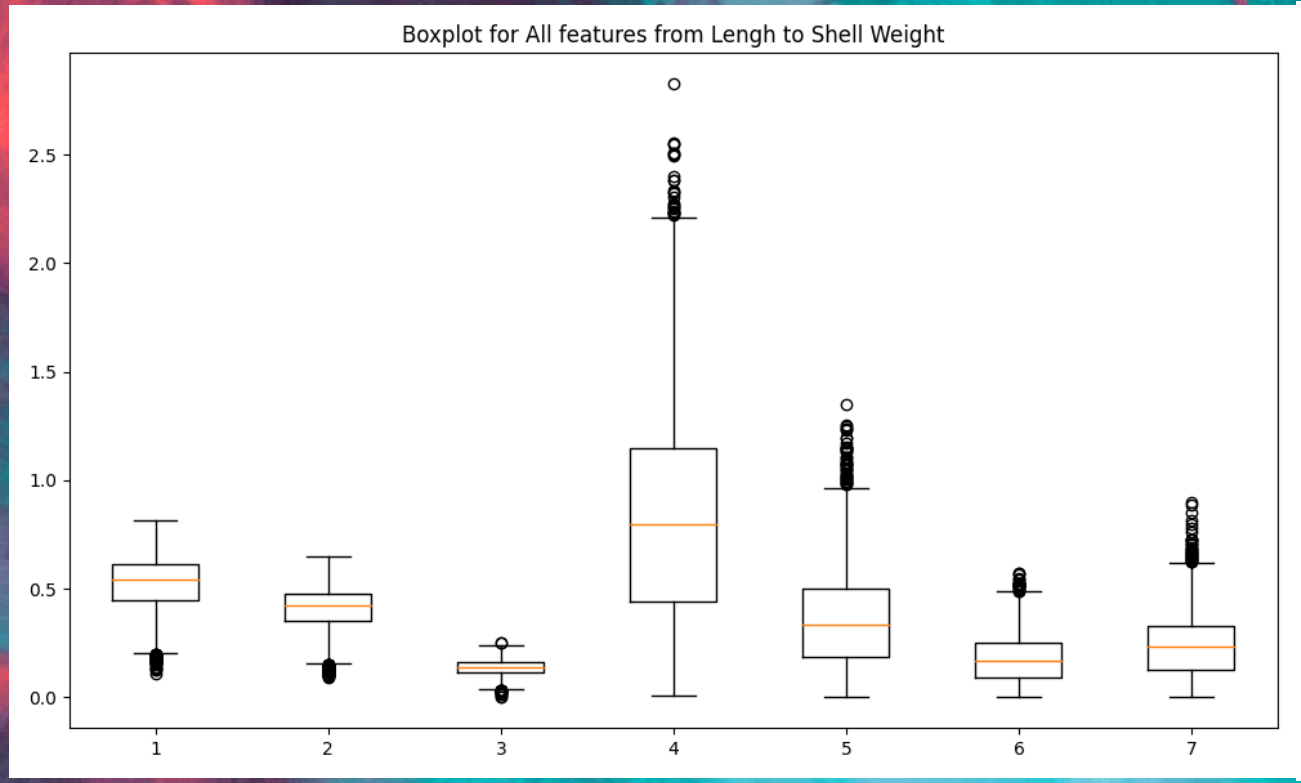
Age vs Viscera Weight

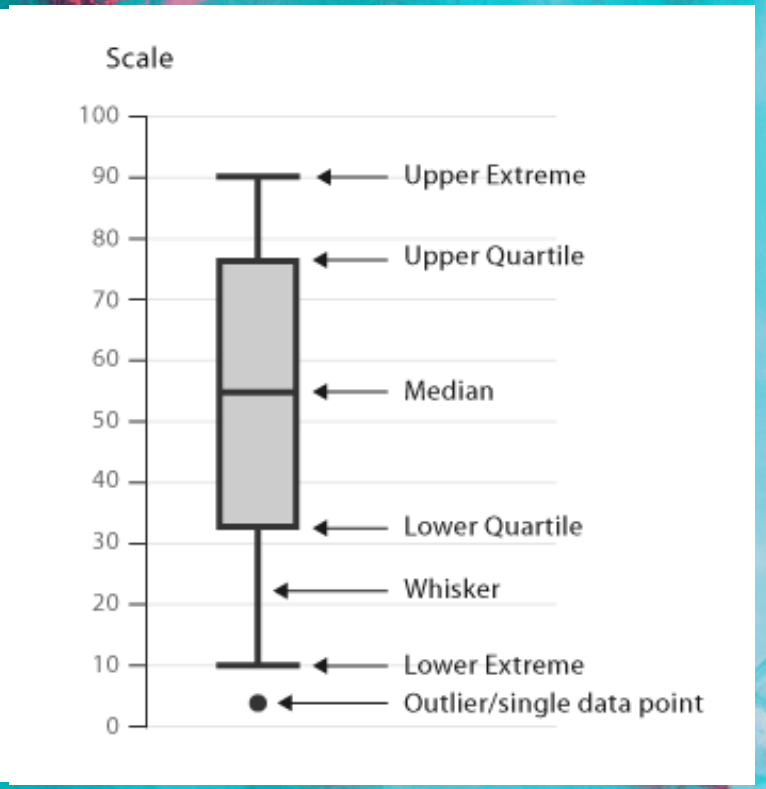


Age vs Shell Weight



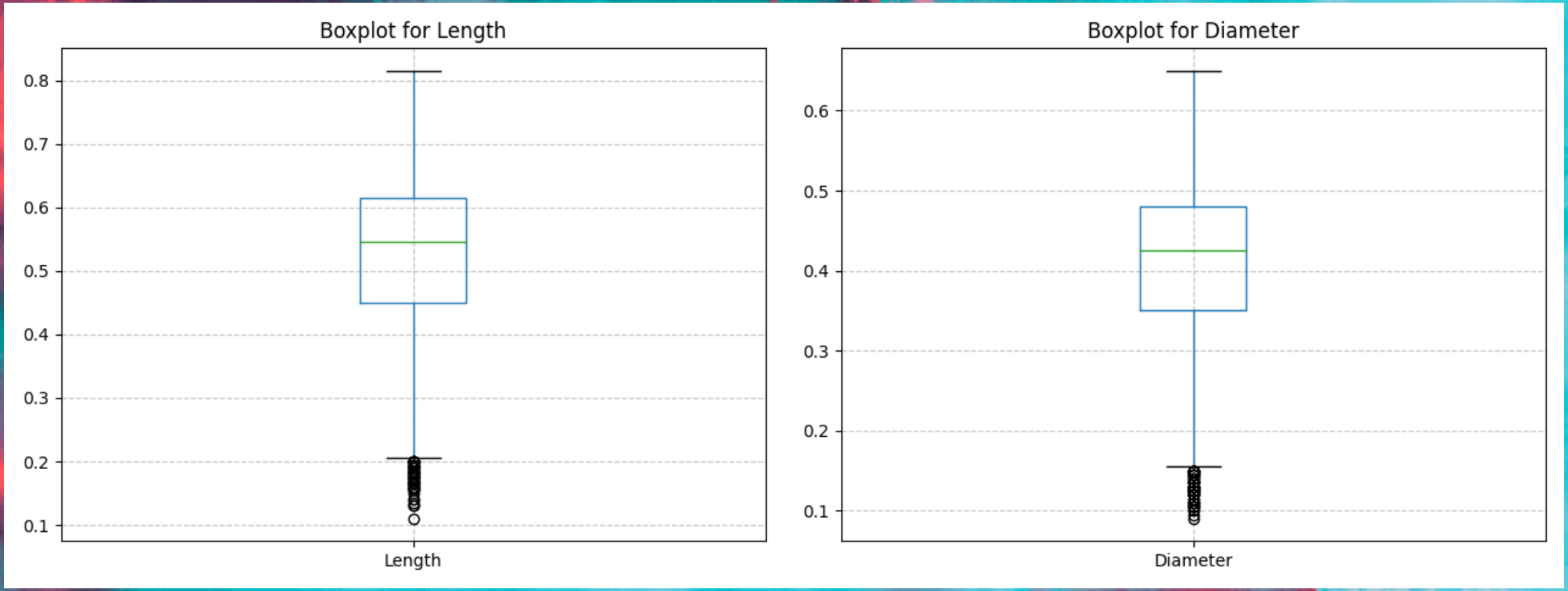
Check Each Feature: Boxplot



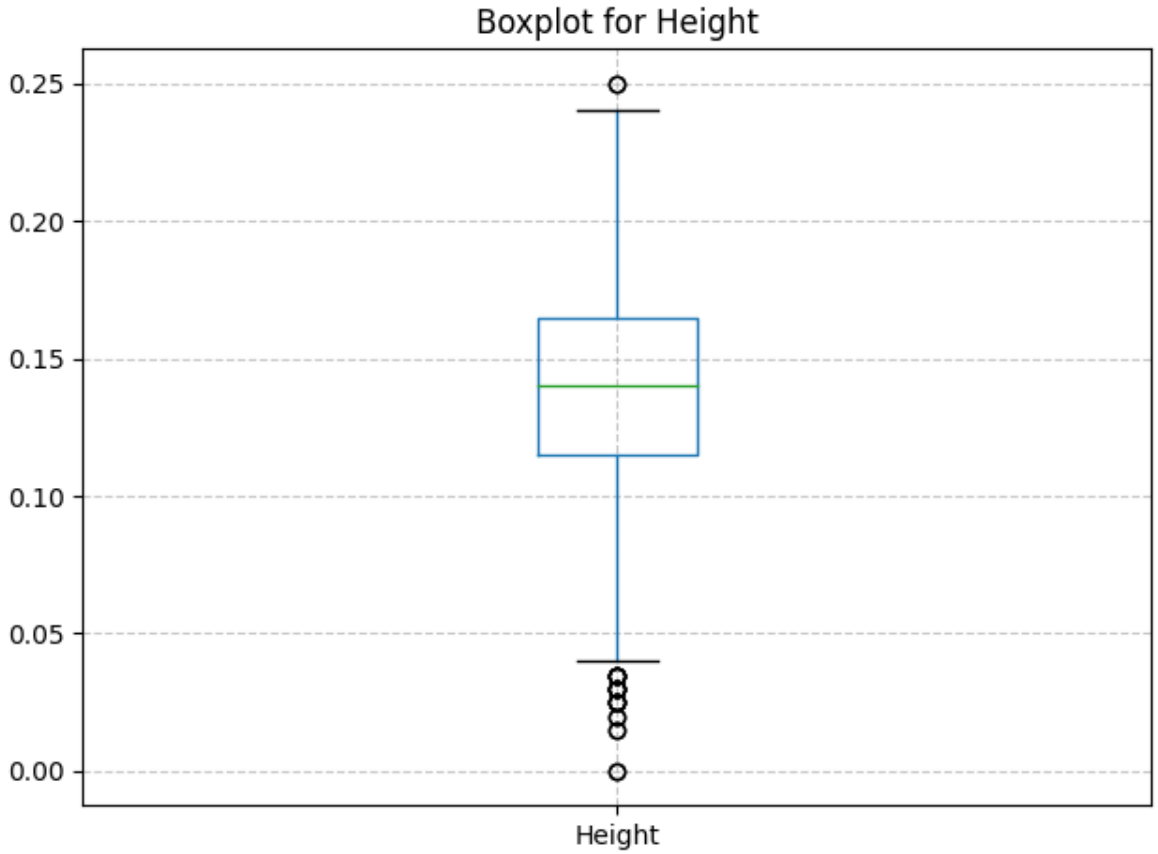


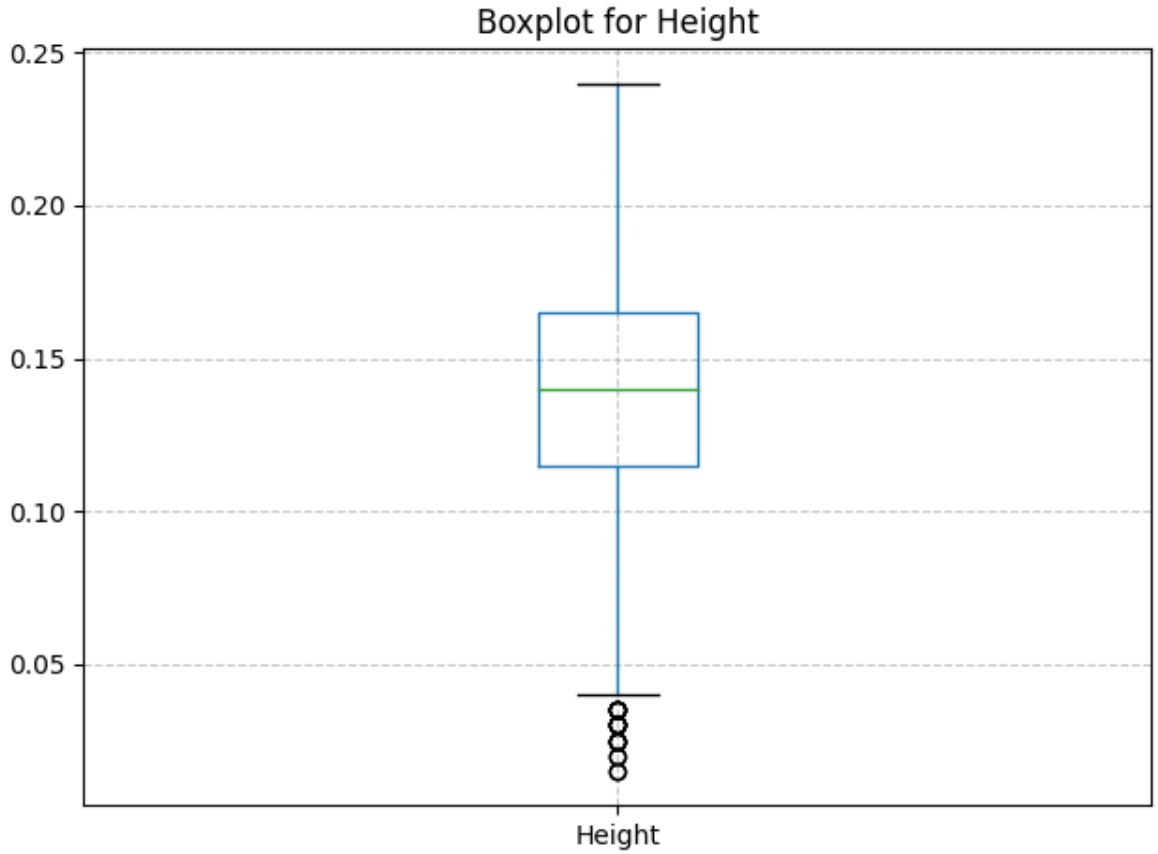
A boxplot gives a good indication of how the values in the data are spread out

Boxplot: Length & Diameter

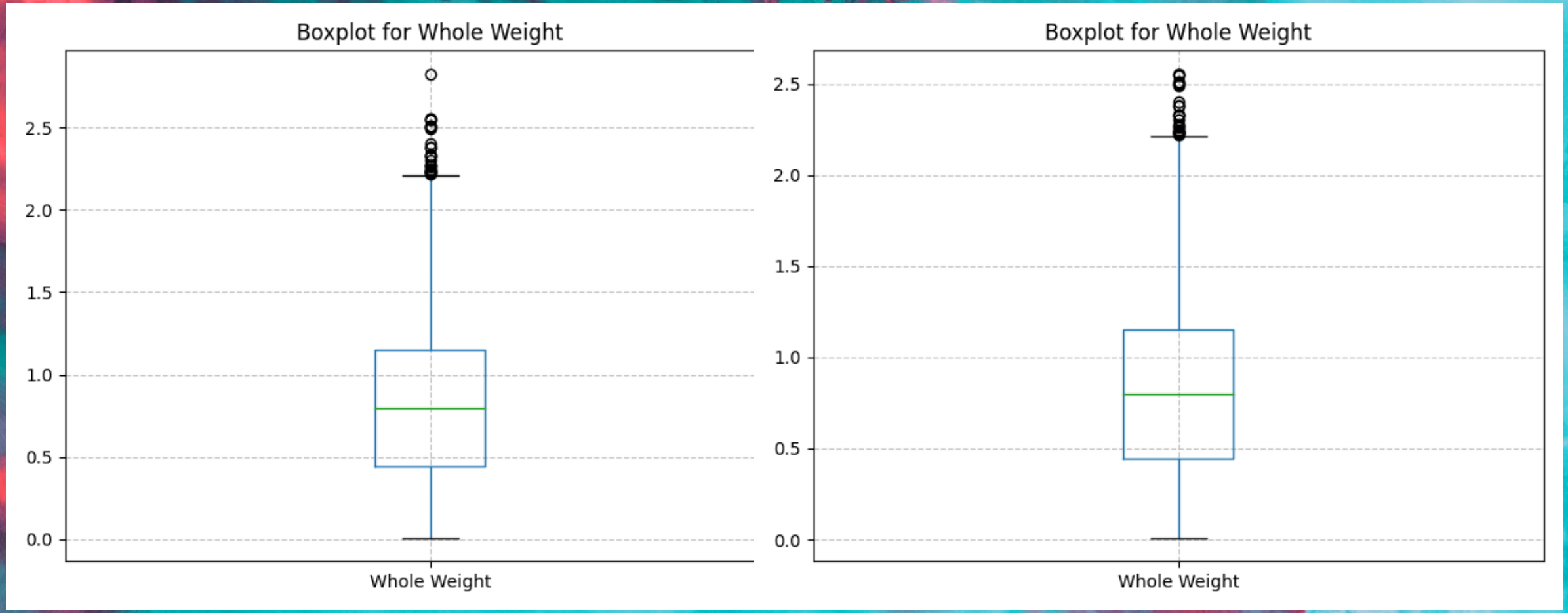


Boxplot: Height

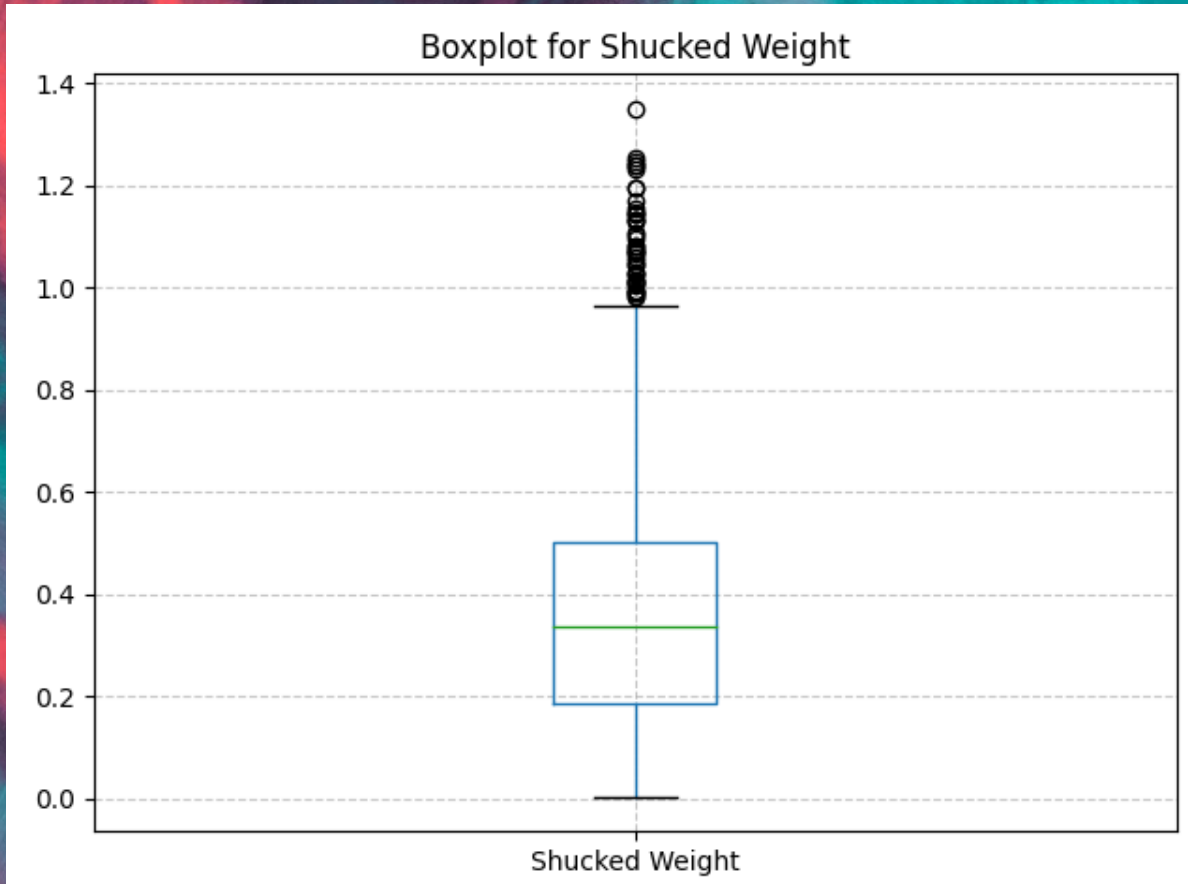


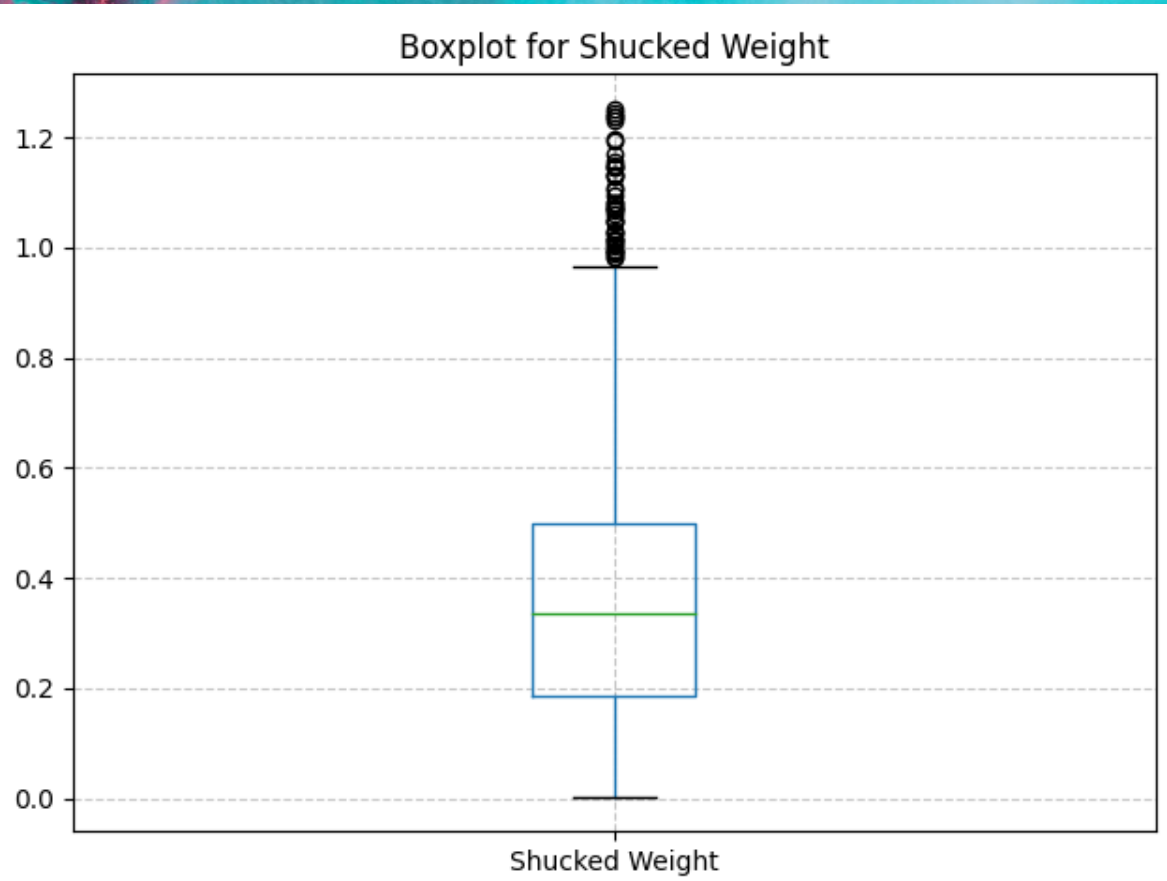


Boxplot: Whole Weight

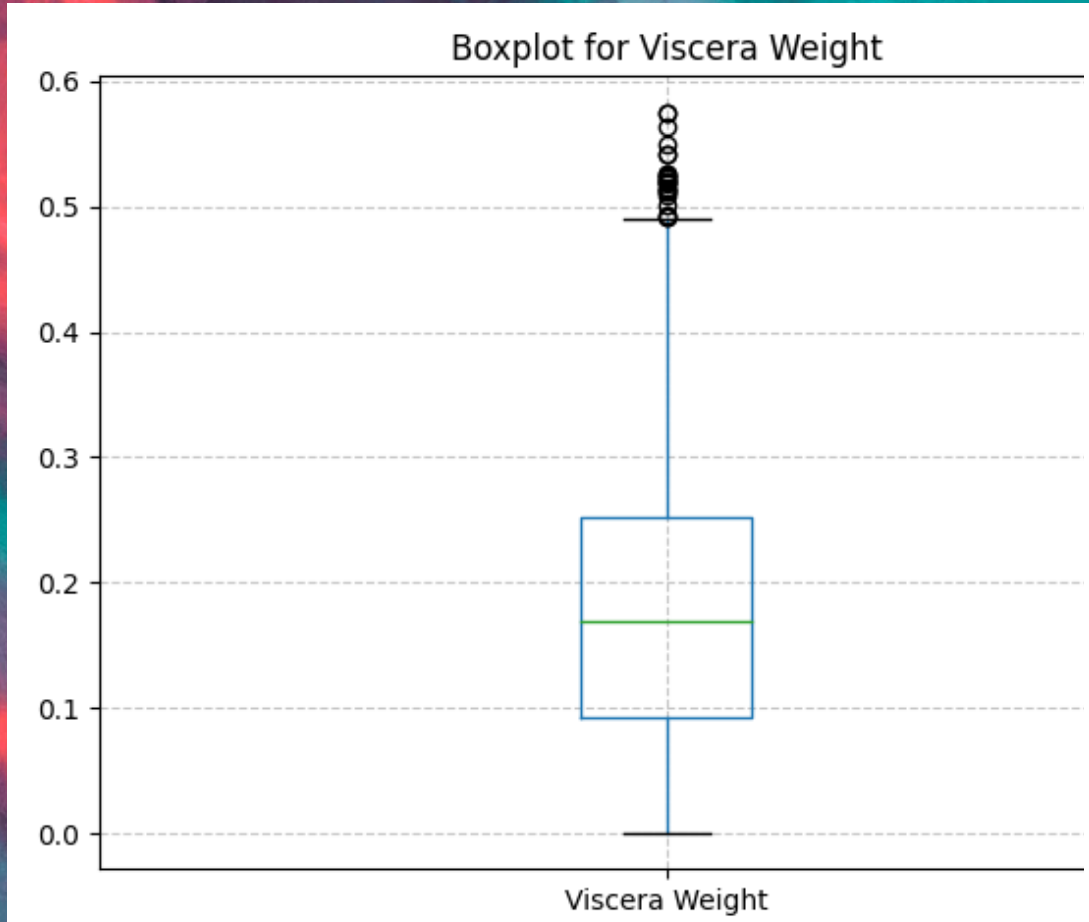


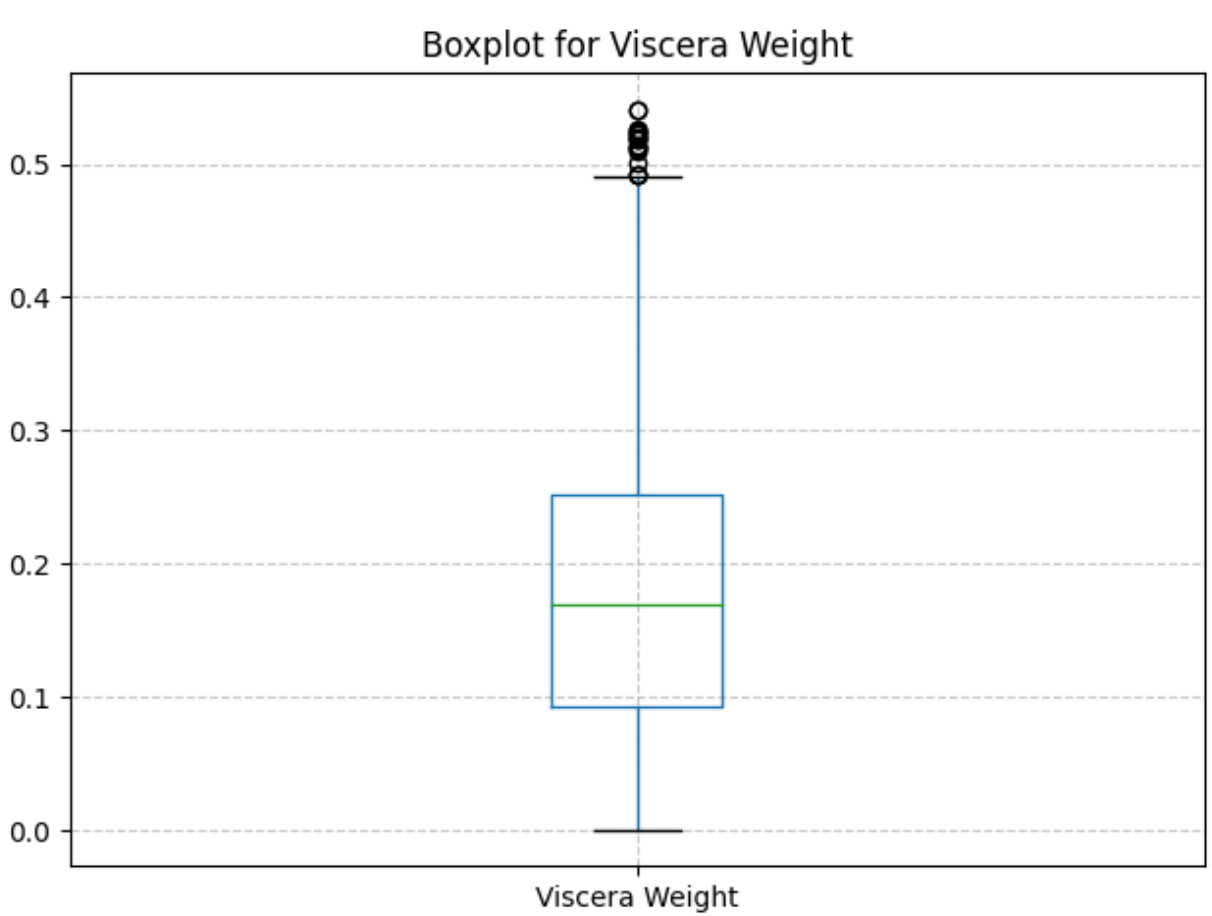
Boxplot: Shucked Weight



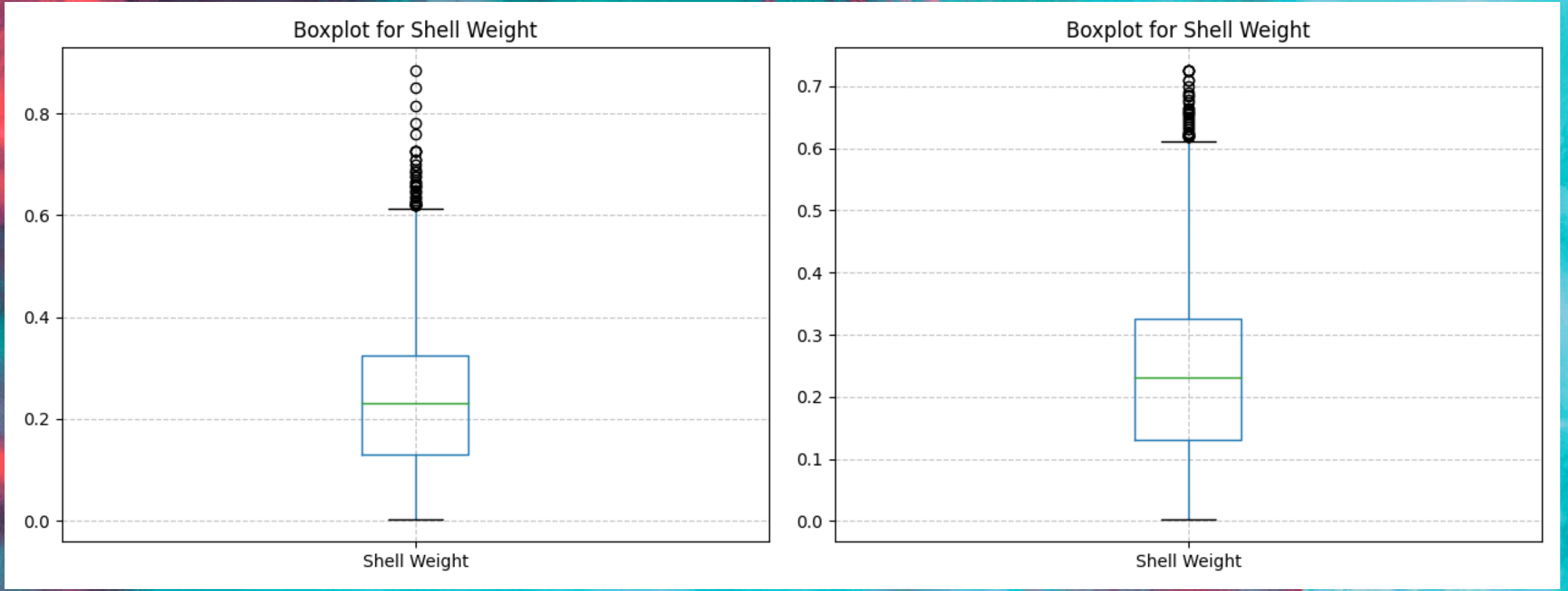


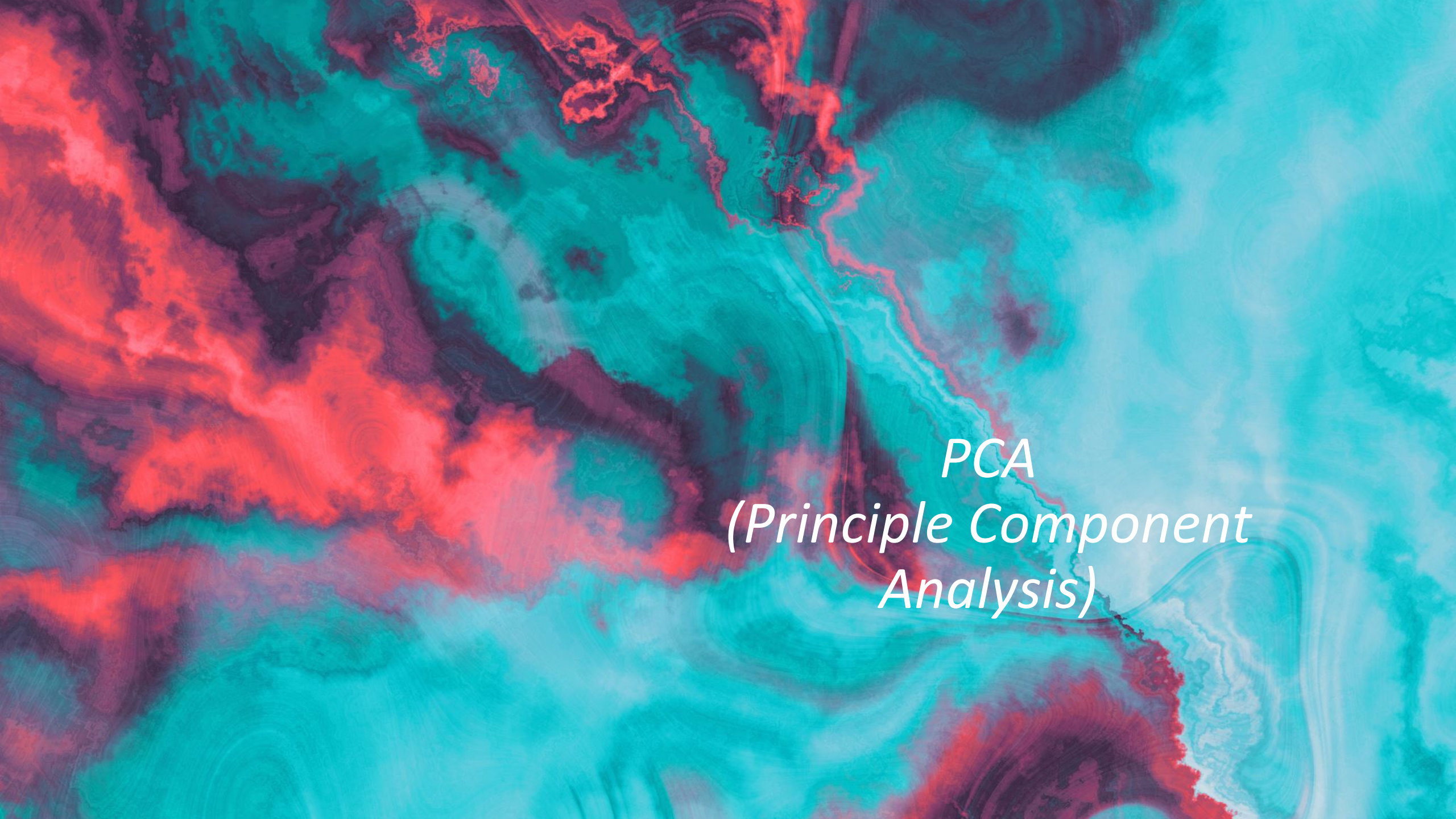
Boxplot: Viscera Weight

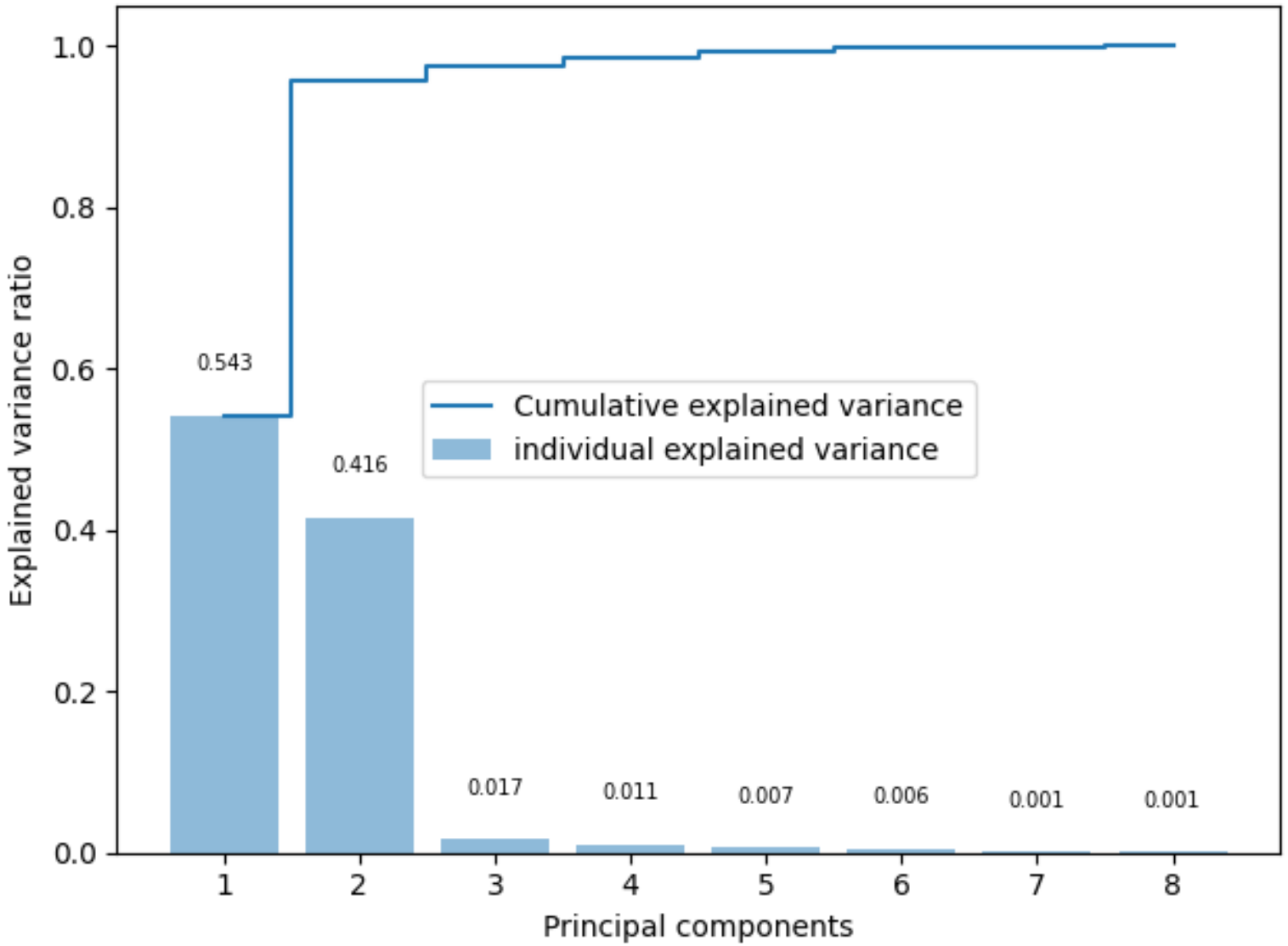




Boxplot: Shell Weight




PCA
(Principle Component
Analysis)

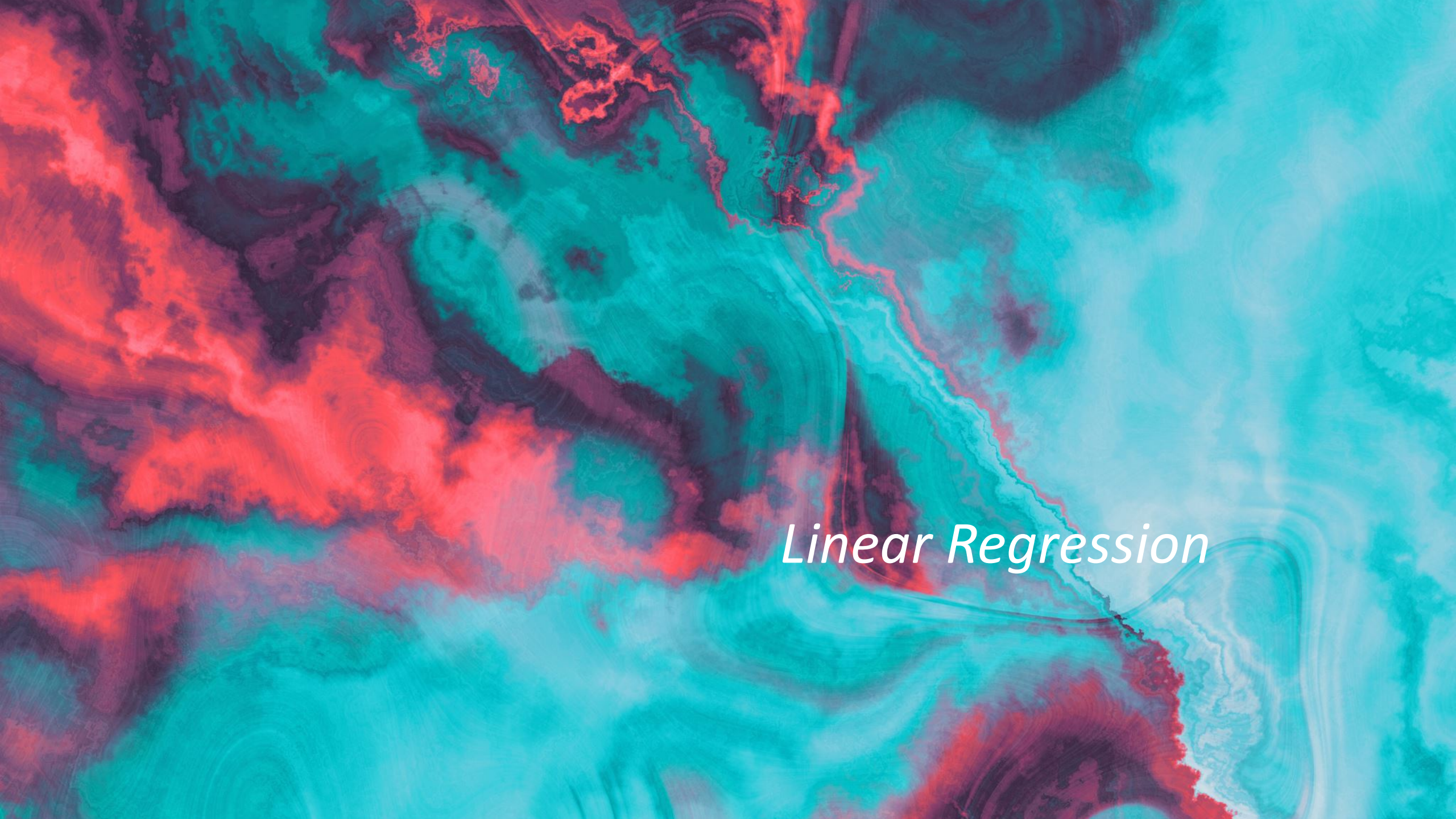


EXPLAINED VARIANCE

```
[0.54270354 0.41555799 0.01672544 0.01071558]
```

PCA.COMPONENTS

```
[[0.126 0.369 0.379 0.335 0.395 0.352 0.409 0.383]
 [0.992 0.044 0.044 0.038 0.054 0.056 0.05  0.047]
 [0.01  0.366 0.412 0.481 0.311 0.438 0.42  0.007]
 [0.004 0.451 0.394 0.501 0.075 0.301 0.045 0.54 ]]
```


Linear Regression

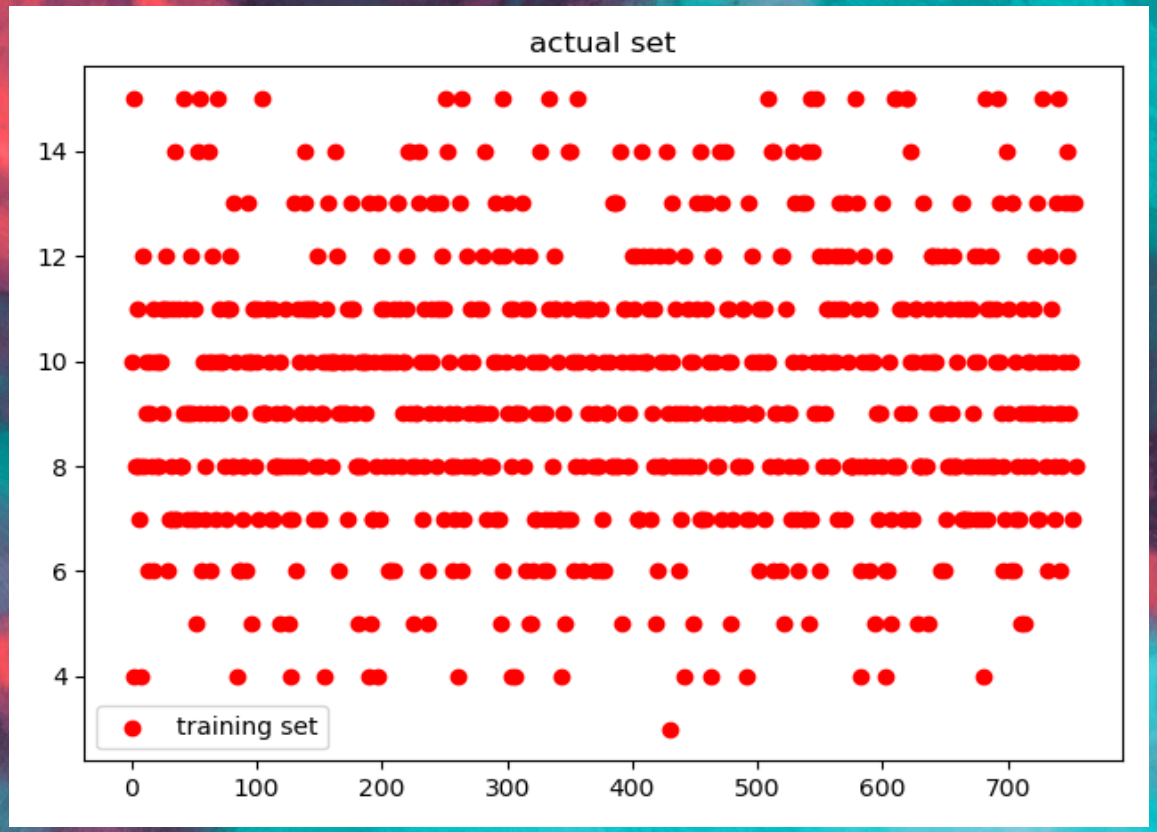
Standardization

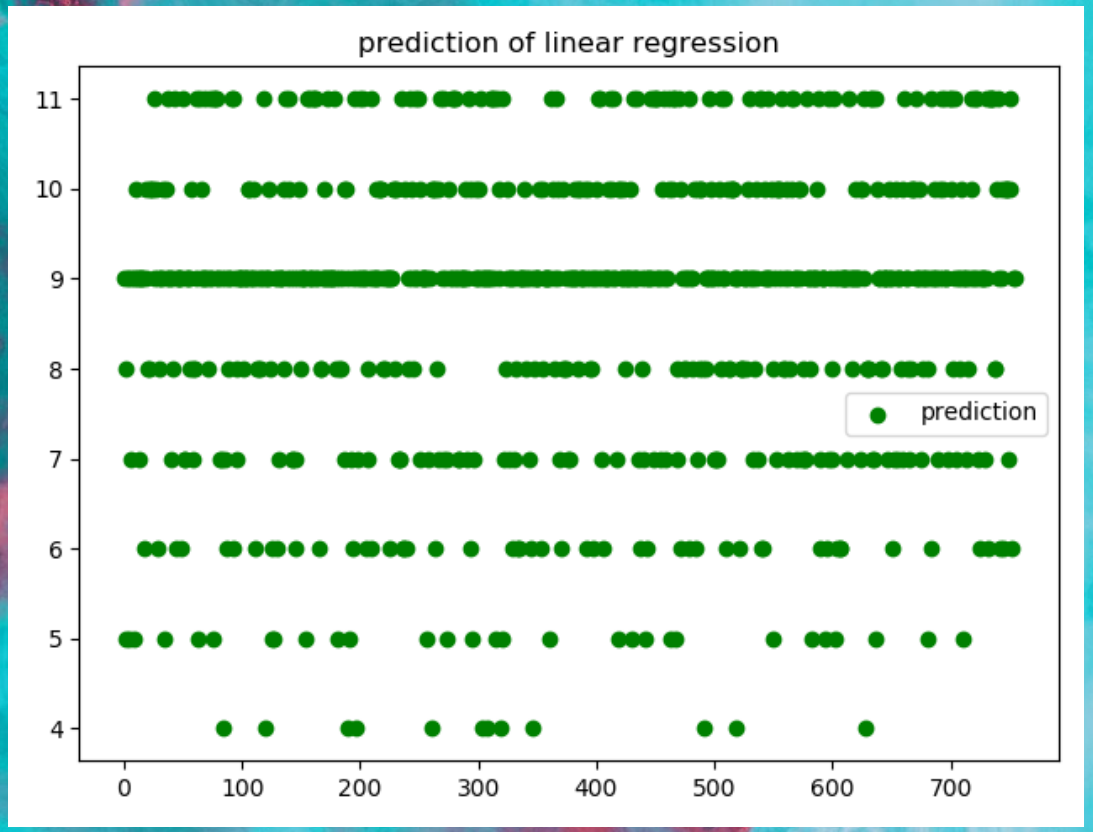
```
StandardScaler_train [[ 0.21935428  1.16602996]
 [ 1.09013478 -0.58906117]
 [-1.35726558 -0.66875823]
 ...
 [ 1.37548018  0.86538308]
 [ 0.19219773  1.02758631]
 [ 0.91266412 -1.49236626]]
StandardScaler_test [[ 0.17114712  0.92426609]
 [ 1.2477301   0.21140746]
 [ 1.38917454  0.93870123]
 ...
 [ 0.9406967  -1.34881909]
 [ 1.03244846 -0.88226441]
 [ 0.26098533  1.38059099]]
```

Result

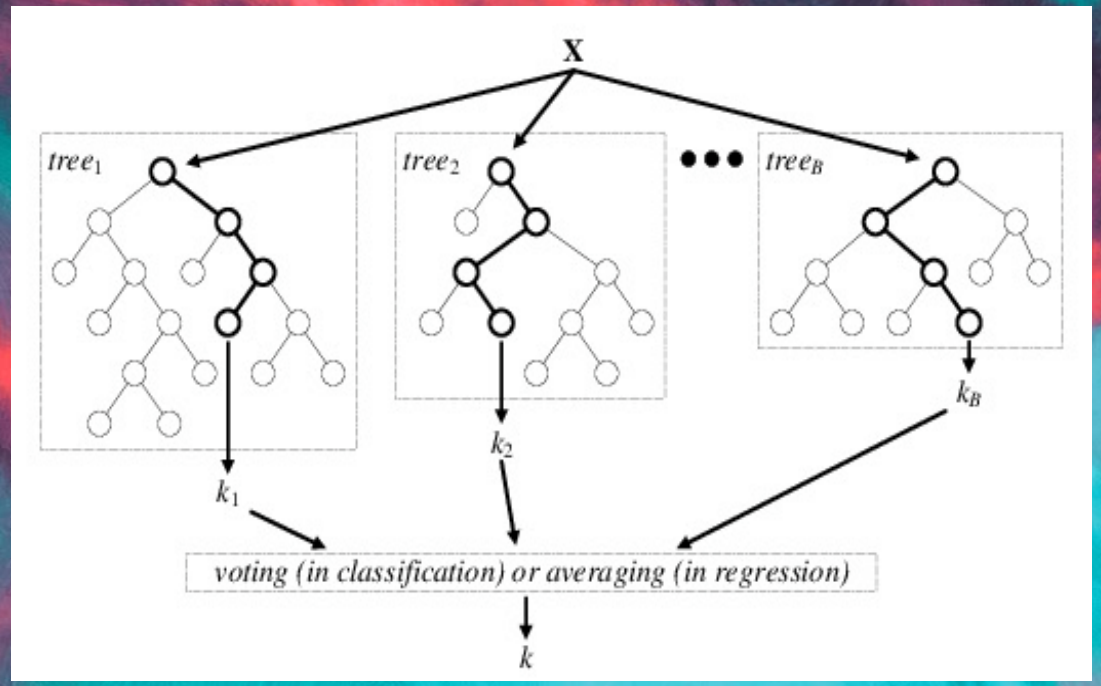
```
print result of lr:
score:  0.3019867549668874
mse:  2.6494631577077326
rms:  1.6277171614588735
r^2:  0.5483131246344208
```

Plot of linear regression



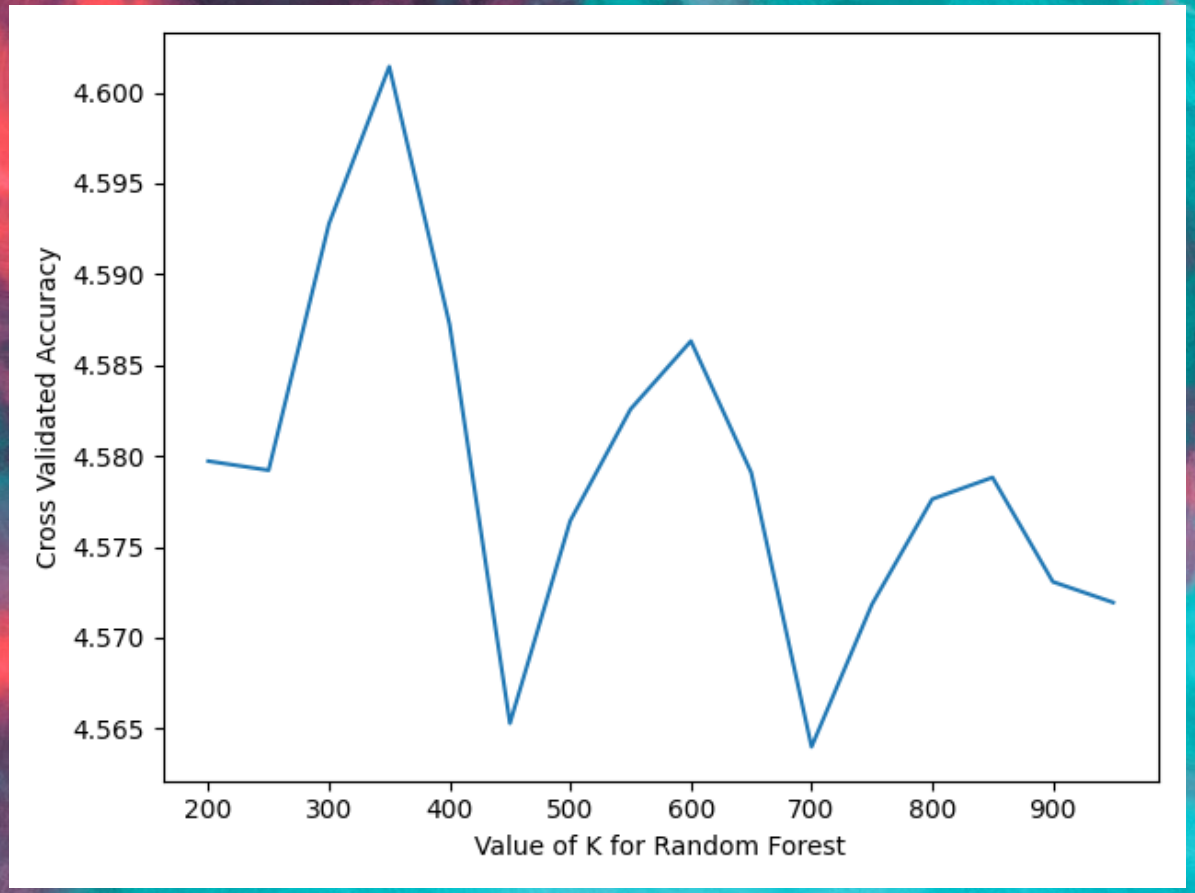


Random forests is an ensemble learning Method for classification , regression and other tasks that operate by constructing a multitude of decision trees.



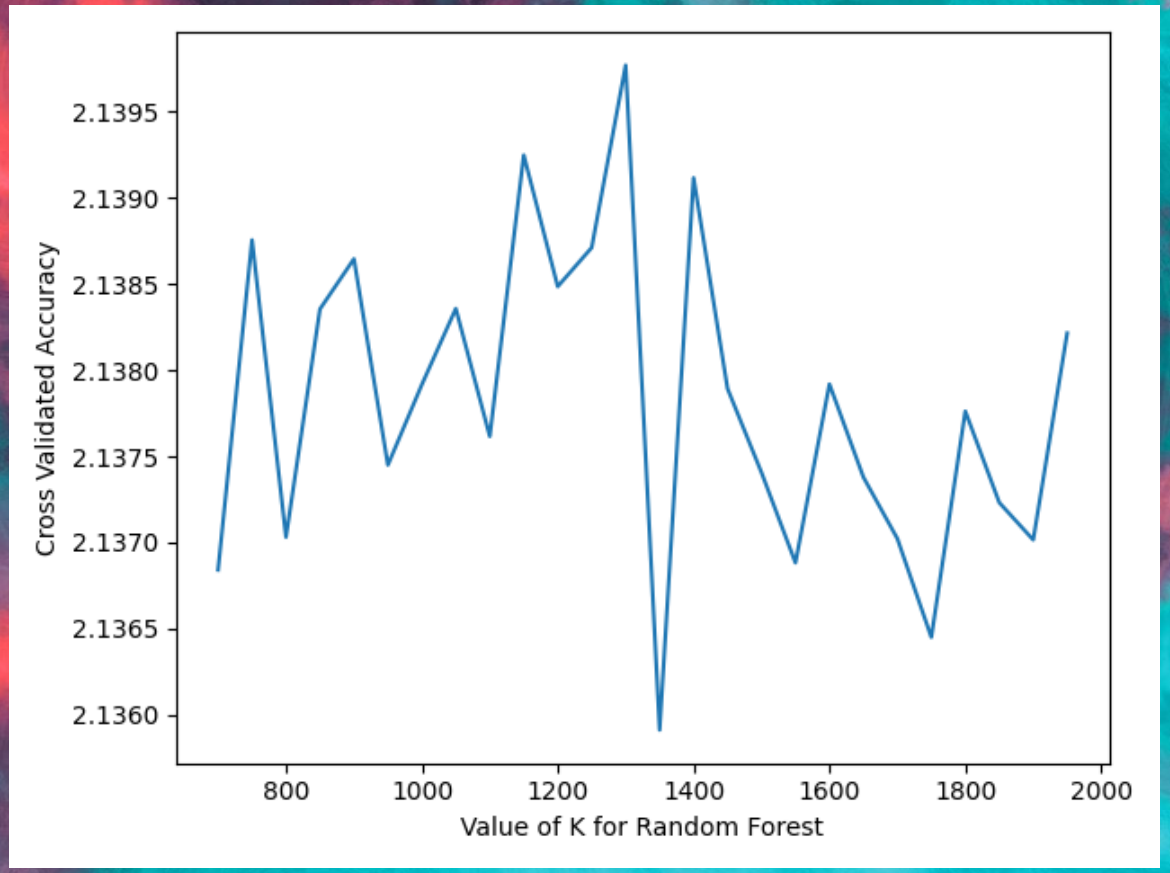
Random Forest

Model Training & Parameter Optimization



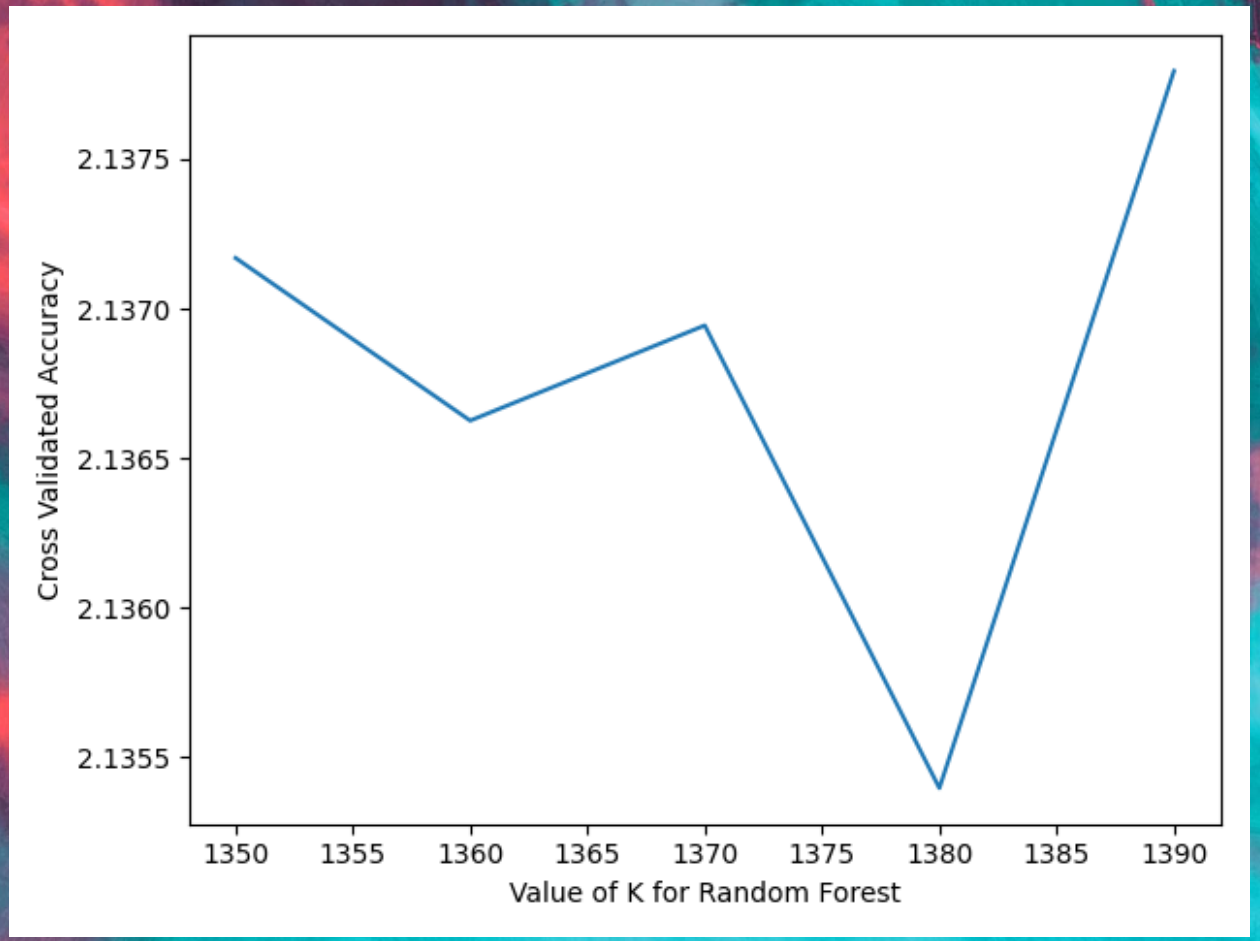
200 – 900 Number of Trees

Model Training & Parameter Optimization



500– 2000 Number of Trees with step size 50

Model Training & Parameter Optimization



1300– 1400
Number of Trees
with step size 10

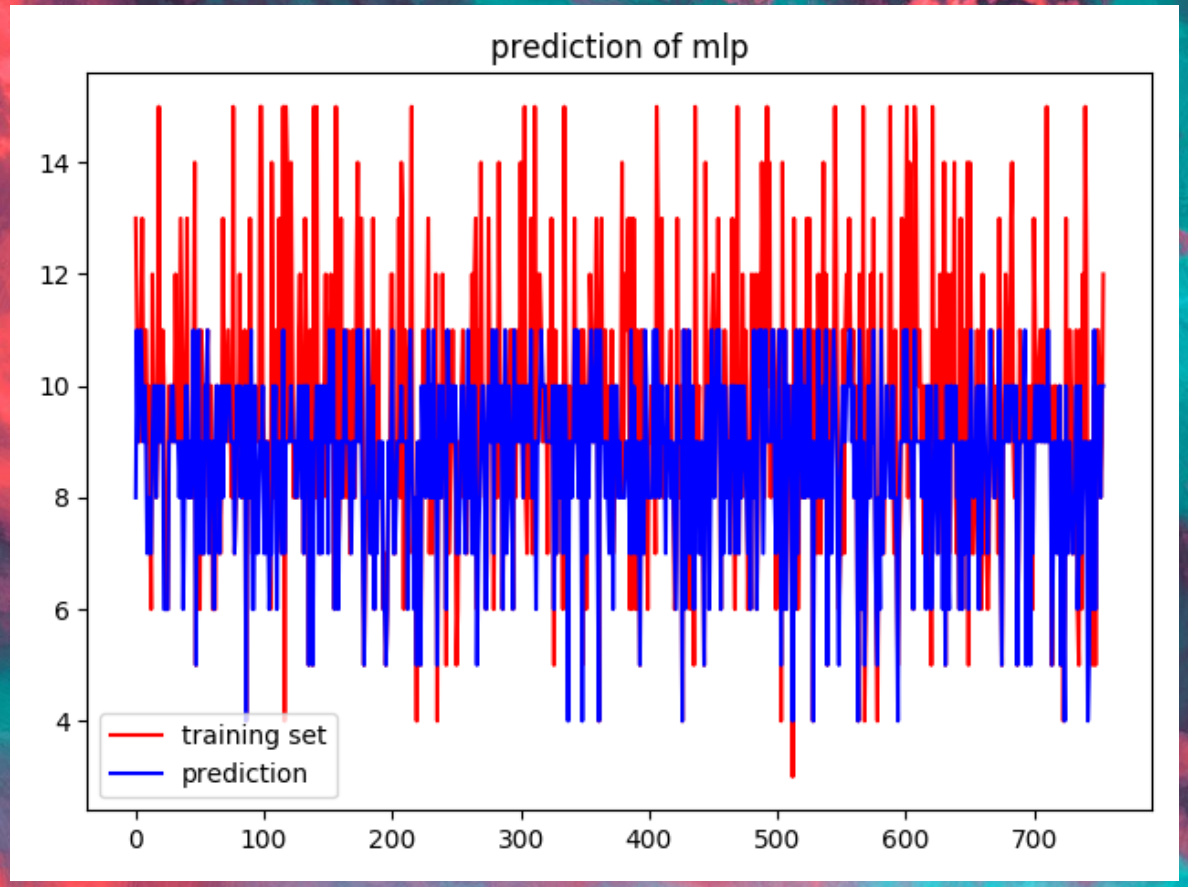
Model Training & Parameter Optimization

Number of Trees in Forest: 1380

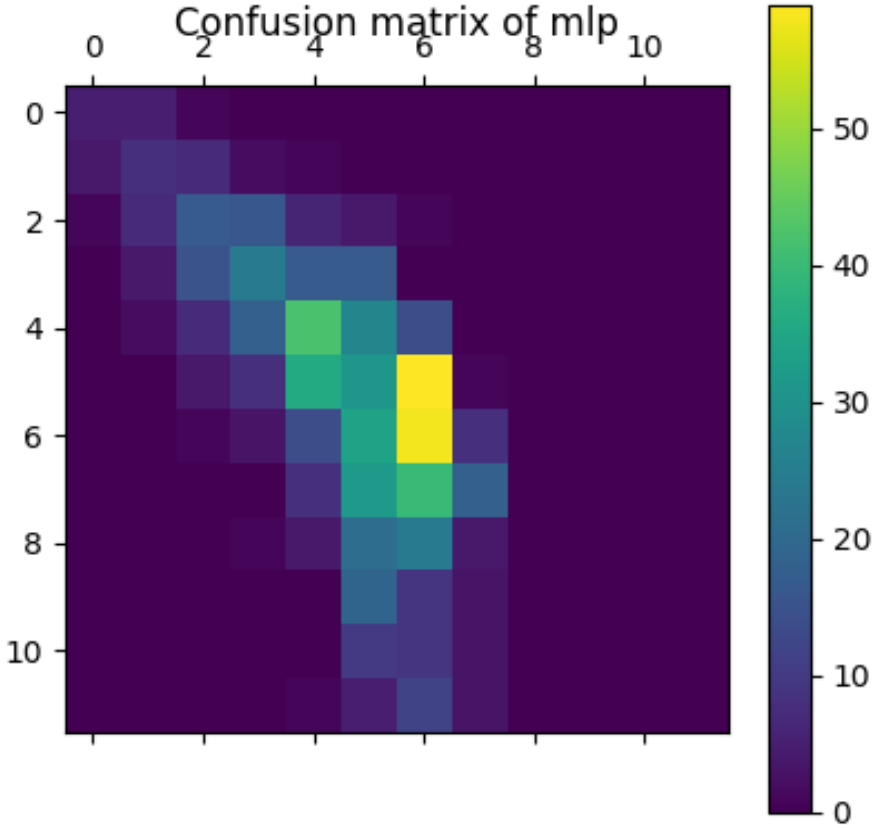
RMSE: 2.1029581

R2 : 0.5353836


Neuron Network

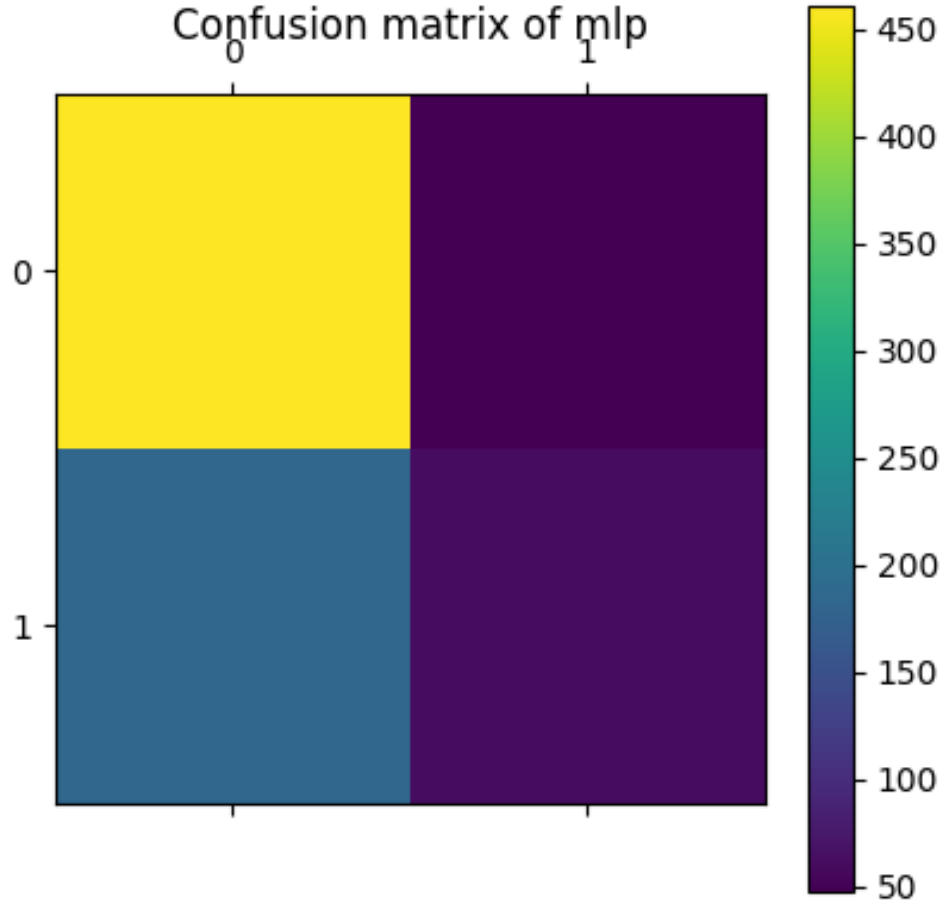


Plot of training set and prediction



Confusion matrix of NN

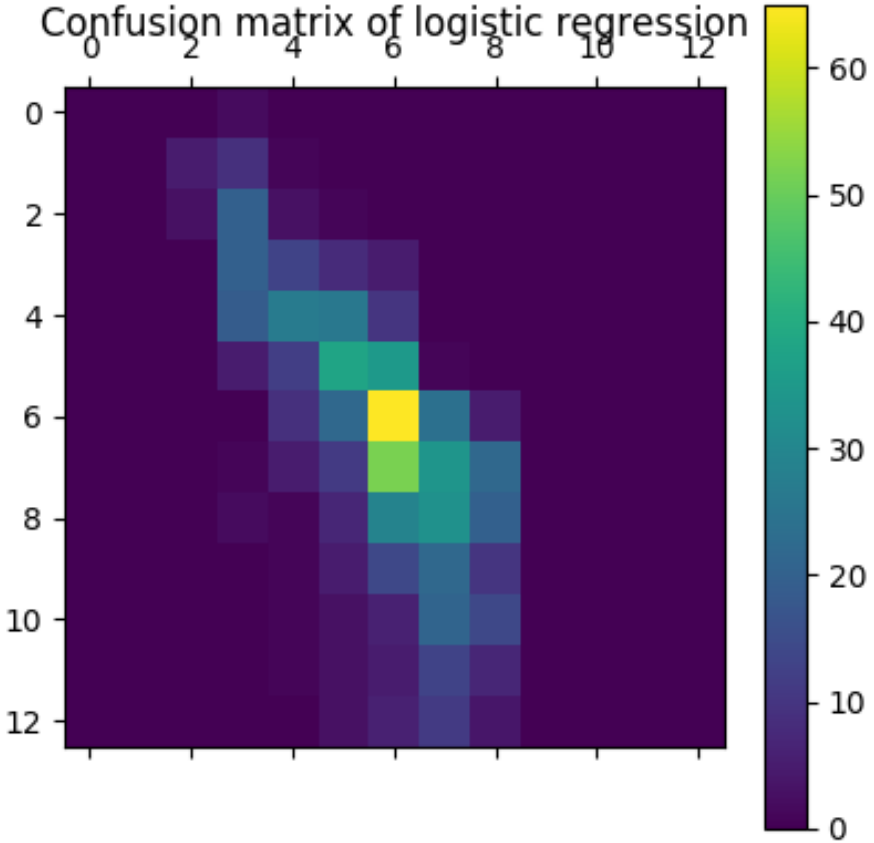
CV: 0.233-0.278



Confusion matrix of NN

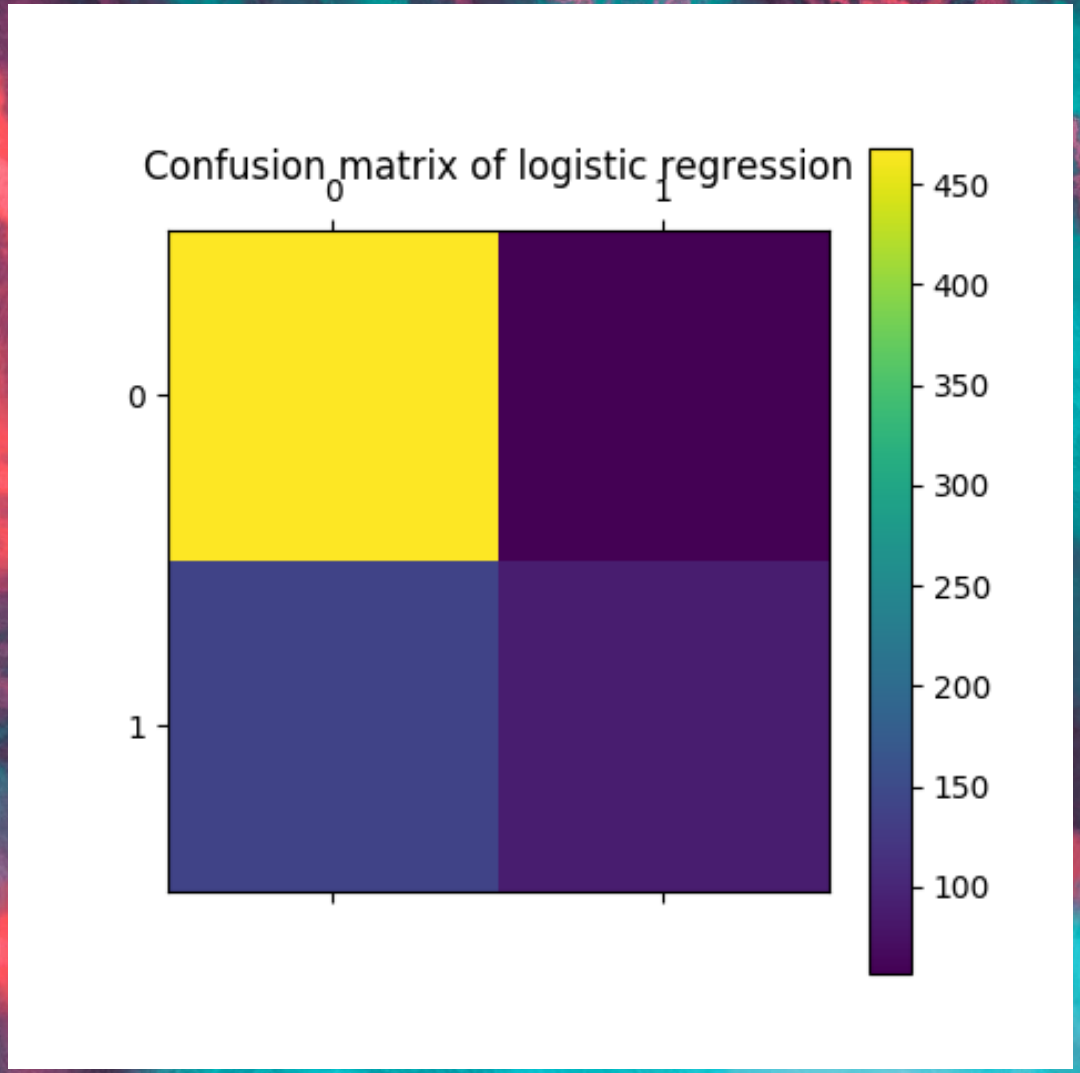
CV: 0.70


Logistic Regression



Confusion matrix of Logistic Regression

CV:0.263



Confusion matrix of Logistic Regression

CV:0.68


Results

Comparison between models

Regression algorithms		
Model	RMSE	R-Squared
Linear regression	1.628	0.548
Random Forest	2.104	0.535

The better one

Classification algorithms			
Model	RMSE	R-Squared	Accuracy score
Multiple layer perceptron	0.55	-0.39	0.70
Logistic regression	0.46	-0.004	0.72

In our report, we compare our classification modes with more indexes in classification report.

The better one

The model still needs to be improved.


Summary & Conclusion

Summary & Conclusion

Regression algorithm

The lower RMSE and higher R-squared, the better the model.

Linear regression is better

Classification algorithm

Observing confusion matrix and accuracy score would decide the effect of the model.
The higher accuracy score, the better the model.

Logistic regression is better


Limitation & Proposed Solution

The performance of our model is not as good as what we have expected, probably due to the imbalanced data set.

The solution to this problem is reducing the number of class into binary class. The model can predict general range of age of abalone.
Other solutions involve increasing more instance and features by provide some feature engineering.

Thank You