

Data Collection and Preprocessing Phase Report

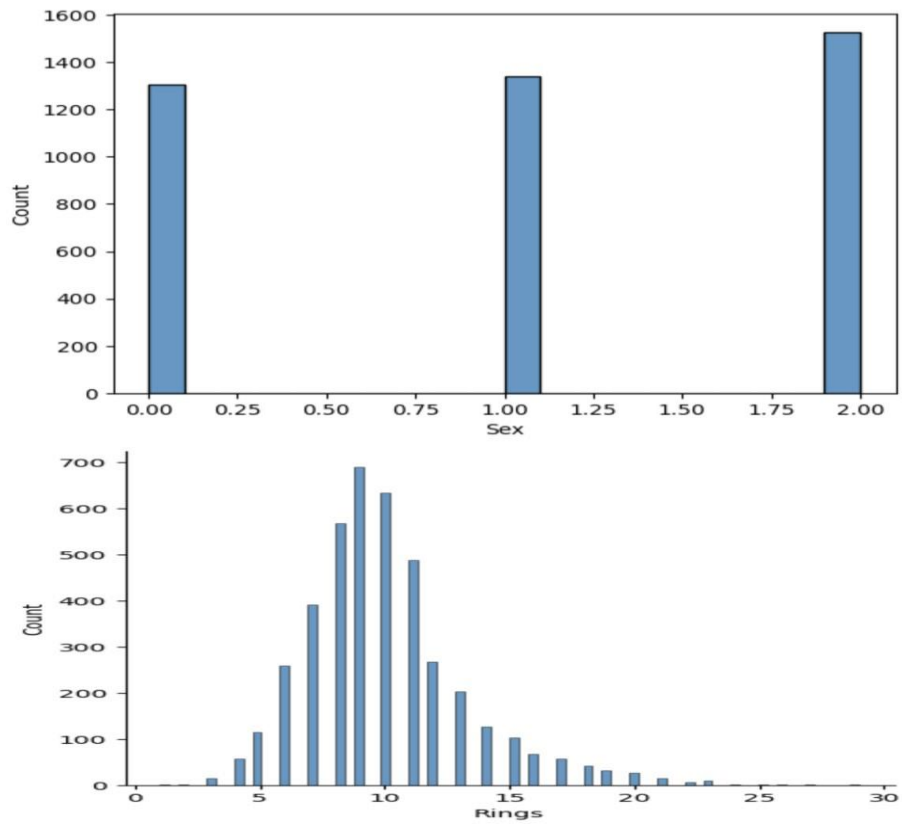
Date	12 July 2024
Team ID	Team - 739960
Project Title	Abalone Age Prediction
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report (6 Marks):

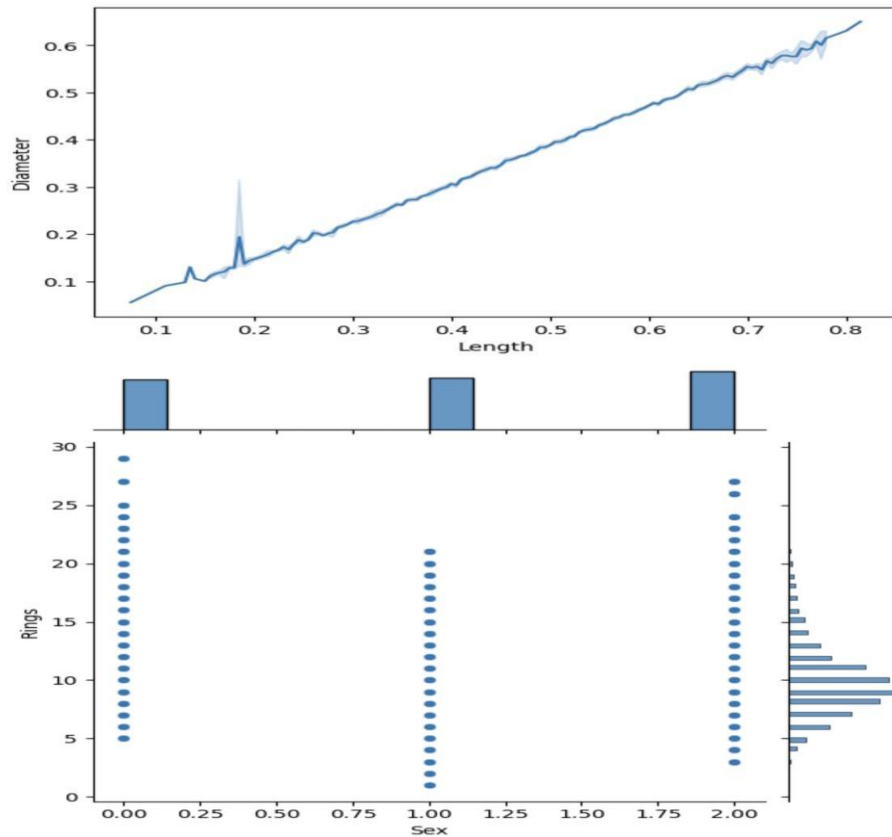
Dataset variables will be statistically analysed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modelling, and forming a strong foundation for insights and predictions.

Section	Description								
Data Overview		Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
	count	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000
	mean	0.523992	0.407881	0.139516	0.828742	0.359367	0.180594	0.238831	9.933684
	std	0.120093	0.099240	0.041827	0.490389	0.221963	0.109614	0.139203	3.224169
	min	0.075000	0.055000	0.000000	0.002000	0.001000	0.000500	0.001500	1.000000
	25%	0.450000	0.350000	0.115000	0.441500	0.186000	0.093500	0.130000	8.000000
	50%	0.545000	0.425000	0.140000	0.799500	0.336000	0.171000	0.234000	9.000000
	75%	0.615000	0.480000	0.165000	1.153000	0.502000	0.253000	0.329000	11.000000
	max	0.815000	0.650000	1.130000	2.825500	1.488000	0.760000	1.005000	29.000000

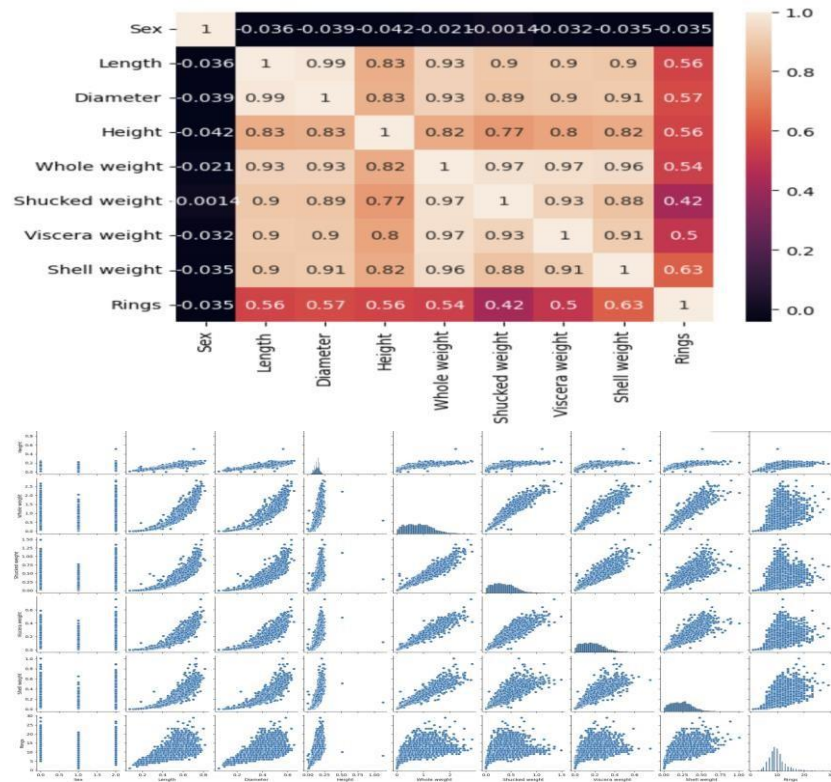
Univariate Analysis
(Hist plot, Dis plot)



Bivariate Analysis
(Line plot, Joint plot)



Multivariate Analysis
(Heatmap, Pair plot)



Outliers and Anomalies

-

Data Preprocessing Code Screenshots

Loading Data

```
#importing the dataset which is in csv file
df=pd.read_csv('/content/abalone.csv')
df.head()
```

	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
0	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7

Handling Missing Data	<pre>df.isnull().sum() Sex 0 Length 0 Diameter 0 Height 0 Whole weight 0 Shucked weight 0 Viscera weight 0 Shell weight 0 Rings 0 dtype: int64</pre>
Data Transformation	<pre>from sklearn.preprocessing import StandardScaler sc=StandardScaler() x_train_scaled=sc.fit_transform(x_train) x_test_scaled=sc.fit_transform(x_test) x_train_scaled array([[-1.26661948, -0.04375418, 0.16375944, ..., 0.16461909, 0.40936642, 0.58511393], [1.1549975 , 0.71476099, 0.77489631, ..., 0.78012036, 0.28950211, 0.01613635], [-1.26661948, 1.34685698, 1.23324896, ..., 1.72040642, 1.58495863, 0.96564034], ..., [-0.05581099, -0.46515151, -0.39644936, ..., -0.49857784, -0.60487 , -0.55284124], [-1.26661948, -0.12803365, -0.34552129, ..., -0.3327786 , -0.57720901, -0.66156307], [1.1549975 , -0.21231311, -0.34552129, ..., -0.38955916, -0.13463312, -0.65793901]]) x_test_scaled array([[-1.33946926e+00, -4.71700742e-01, -2.29814532e-01, ..., -2.52459826e-01, -2.46013428e-01, -5.35361844e-01], [-1.33946926e+00, 5.64153706e-01, 4.48387505e-01, ..., 8.75136245e-04, -2.15328603e-01, 5.41461461e-01], [-1.33946926e+00, -9.49787410e-01, -6.65801555e-01, ..., -6.43381881e-01, -5.26560391e-01, -8.13251729e-01], ..., [-1.33946926e+00, -1.18883074e+00, -1.24711759e+00, ..., -1.10855729e+00, -1.11395560e+00, -1.16061409e+00], [-1.33946926e+00, 1.12192149e+00, 1.27191855e+00, ..., 1.54272413e+00, 1.15672139e+00, 1.23618617e+00], [-1.33946926e+00, 1.24144315e+00, 1.22347555e+00, ..., 1.39640135e+00, 1.31014551e+00, 1.05555775e+00]])</pre>
Feature Engineering	Attached codes in final submission.
Save Processed Data	-