Report Topic 01 - Regression

This report details the research and application of linear regression compared to logistic regression. The objective is to understand and learn these methods and how to apply them for other purposes. The data used for this analysis are taken from:

Nash, Warwick, et al. "Abalone." UCI Machine Learning Repository, 1994, https://doi.org/10.24432/C55C7W.

```
Dataset head:
     Length  Diameter  Height  Whole_weight  Shucked_weight  Viscera_weight
Sex
M     0.455    0.365   0.095        0.5140          0.2245          0.1010
M     0.350    0.265   0.090        0.2255          0.0995          0.0485
F     0.530    0.420   0.135        0.6770          0.2565          0.1415
M     0.440    0.365   0.125        0.5160          0.2155          0.1140
I     0.330    0.255   0.080        0.2050          0.0895          0.0395

     Shell_weight  target(Age)
Sex
M            0.150           15
M            0.070            7
F            0.210            9
M            0.155           10
I            0.055            7
```

Fig 1. Dataset of abalone

The train-test split ratio is 0.8
- Training set size: 3341
- Testing set size: 836

```
The model Mean Absolute Percentage Error: 0.17
The model Mean Squared Error: 5.73
The model r-squared: 0.47
```

Fig 2. Evaluation using sk-learn

```
                        OLS Regression Results
==============================================================================
Dep. Variable:        target(Age)   R-squared:                     0.535
Model:                        OLS   Adj. R-squared:                0.534
Method:             Least Squares   F-statistic:                   547.5
Date:            Sun, 06 Apr 2025   Prob (F-statistic):             0.00
Time:                    10:13:02   Log-Likelihood:               -7357.4
No. Observations:            3341   AIC:                        1.473e+04
Df Residuals:                3333   BIC:                        1.478e+04
Df Model:                       7
Covariance Type:        nonrobust
==============================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const            2.9431      0.298      9.874      0.000       2.359       3.528
Length          -3.3102      2.078     -1.593      0.111      -7.385       0.765
Diameter        12.2094      2.551      4.787      0.000       7.208      17.210
Height          23.0027      2.378      9.675      0.000      18.341      27.664
Whole_weight     9.5181      0.834     11.412      0.000       7.883      11.153
Shucked_weight -19.7104      0.927    -21.271      0.000     -21.527     -17.894
Viscera_weight -10.2366      1.447     -7.075      0.000     -13.074      -7.400
Shell_weight     6.7166      1.296      5.184      0.000       4.176       9.257
==============================================================================
Omnibus:                    748.154   Durbin-Watson:                 1.964
Prob(Omnibus):                0.000   Jarque-Bera (JB):           2001.276
Skew:                         1.188   Prob(JB):                       0.00
Kurtosis:                     5.954   Cond. No.                       136.
==============================================================================
```

Fig 3. Evaluation using OLS

| Sex	Length	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	target(Age)	predicted	MSE	SST	SSR
F	0.450	0.335	0.105	0.4250	0.1865	0.0910	0.1150	9	8.238814	0.579405	0.888464	0.579405
F	0.615	0.520	0.150	1.3435	0.6290	0.2605	0.3450	10	10.881268	0.776633	0.003297	0.776633
M	0.590	0.470	0.145	0.9235	0.4545	0.1730	0.2540	9	9.900945	0.811703	0.888464	0.811703
F	0.375	0.290	0.115	0.2705	0.0930	0.0660	0.0885	10	8.321296	2.818047	0.003297	2.818047
M	0.300	0.240	0.090	0.1610	0.0725	0.0390	0.0500	6	6.818367	0.669724	15.543966	0.669724

Fig 4. Summary table of all calculations
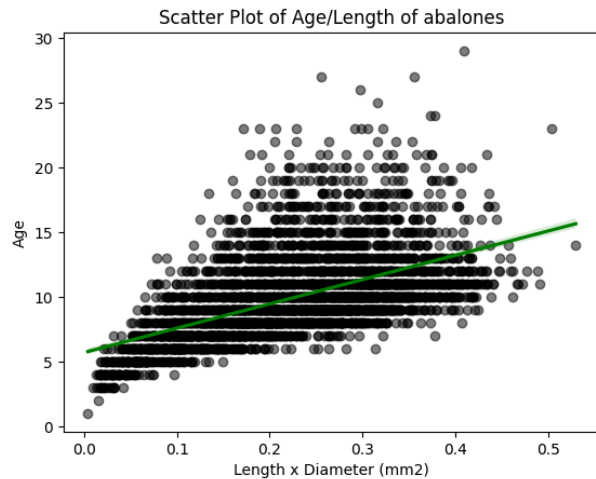
Fig 5. Age of abalone to size

From here is where logistic regression is applied to the data. The sigmoid curve is not present in this plot when I applied the code. This is most likely errors from me but could also be this method is not appropriate for this dataset.
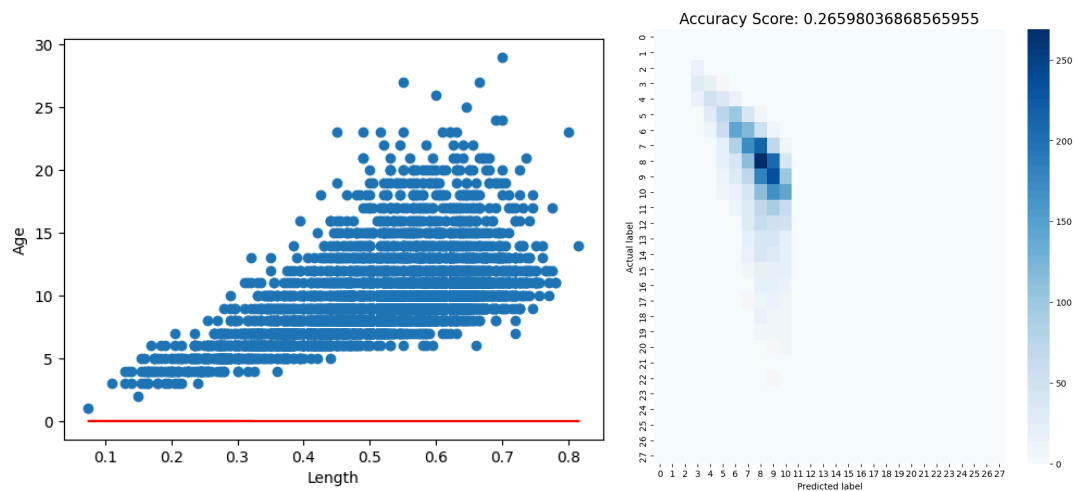

Fig 6&7. Scatter plot using logistic regression; heatmap of plot including accuracy score