Abalone Age Prediction

Midterm Project

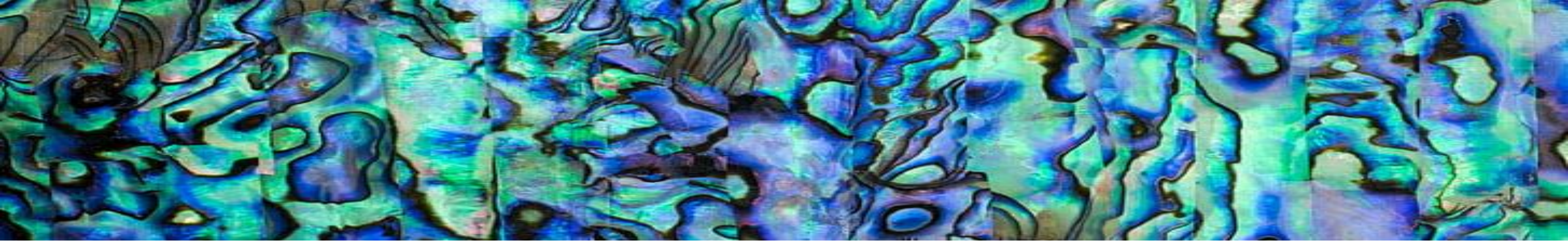
Christin, Christina

General Info about The Dataset

- The Dataset was collected in December 1995 about Abalones that grow in the southeast of Australian
- Abalones grow by adding layers to their shell
- Adding 1.5 to the number of Rings will give you the age of the Abalone.
- 8 Variables that are continuous, because they were scaled dividing it by 200
- One categorical variable
- Total of 4177 observations
- Data was cleaned, Nan were removed

	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	0.1500	15
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.0700	7
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.2100	9
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.1550	10
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.0550	7
...	...	...	...	...	...	...	...	...	...
4172	F	0.565	0.450	0.165	0.8870	0.3700	0.2390	0.2490	11
4173	M	0.590	0.440	0.135	0.9660	0.4390	0.2145	0.2605	10
4174	M	0.600	0.475	0.205	1.1760	0.5255	0.2875	0.3080	9
4175	F	0.625	0.485	0.150	1.0945	0.5310	0.2610	0.2960	10
4176	M	0.710	0.555	0.195	1.9485	0.9455	0.3765	0.4950	12

4177 rows × 9 columns



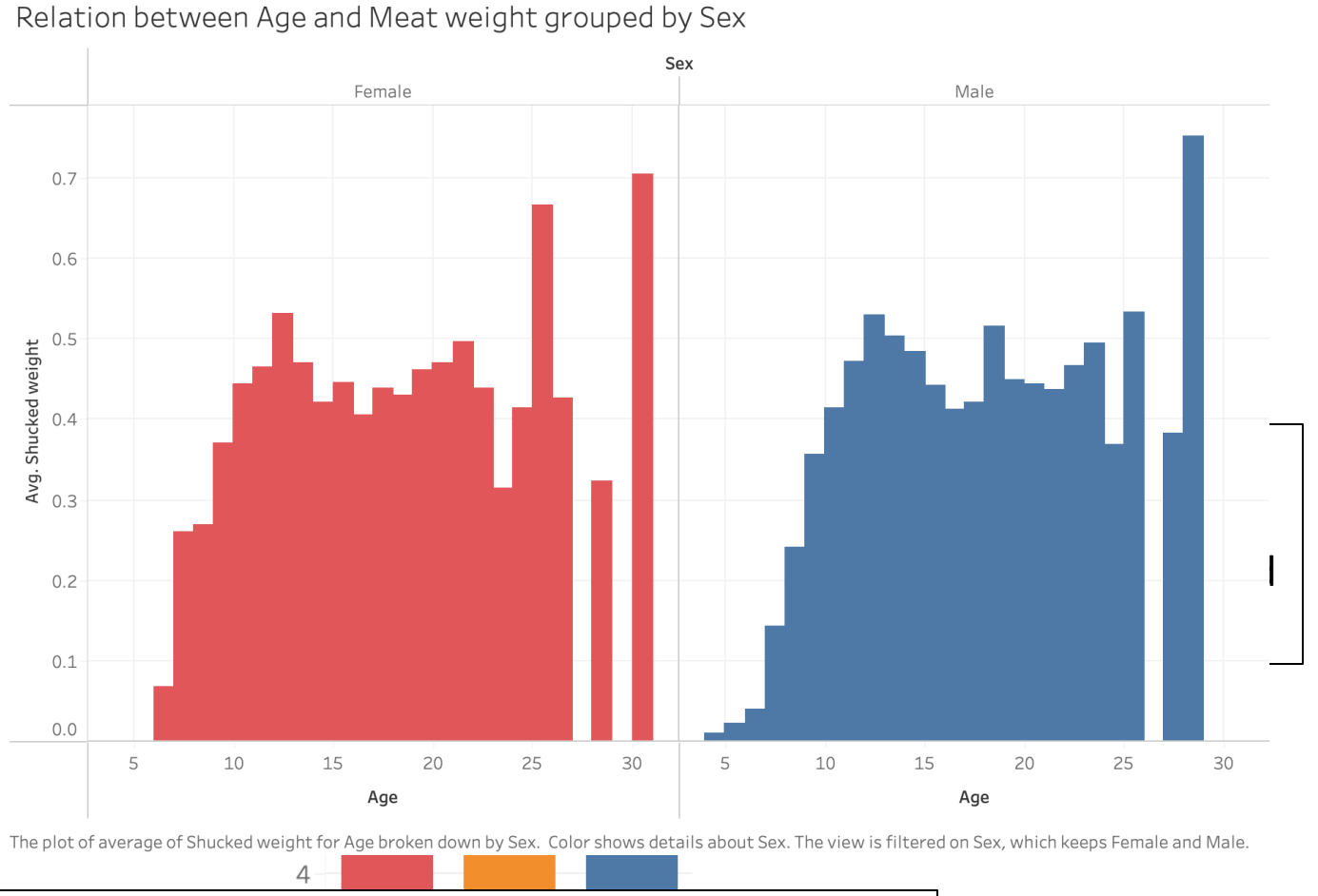
Central Research Question

Is there an easier way to successfully predict the age of abalones with physical measurements by given observations?

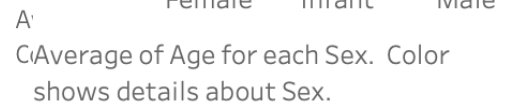
Outline

1. Bivariate Distributions
2. Correlations
3. Linearity
4. OLS Regression
5. Conclusion and further Research

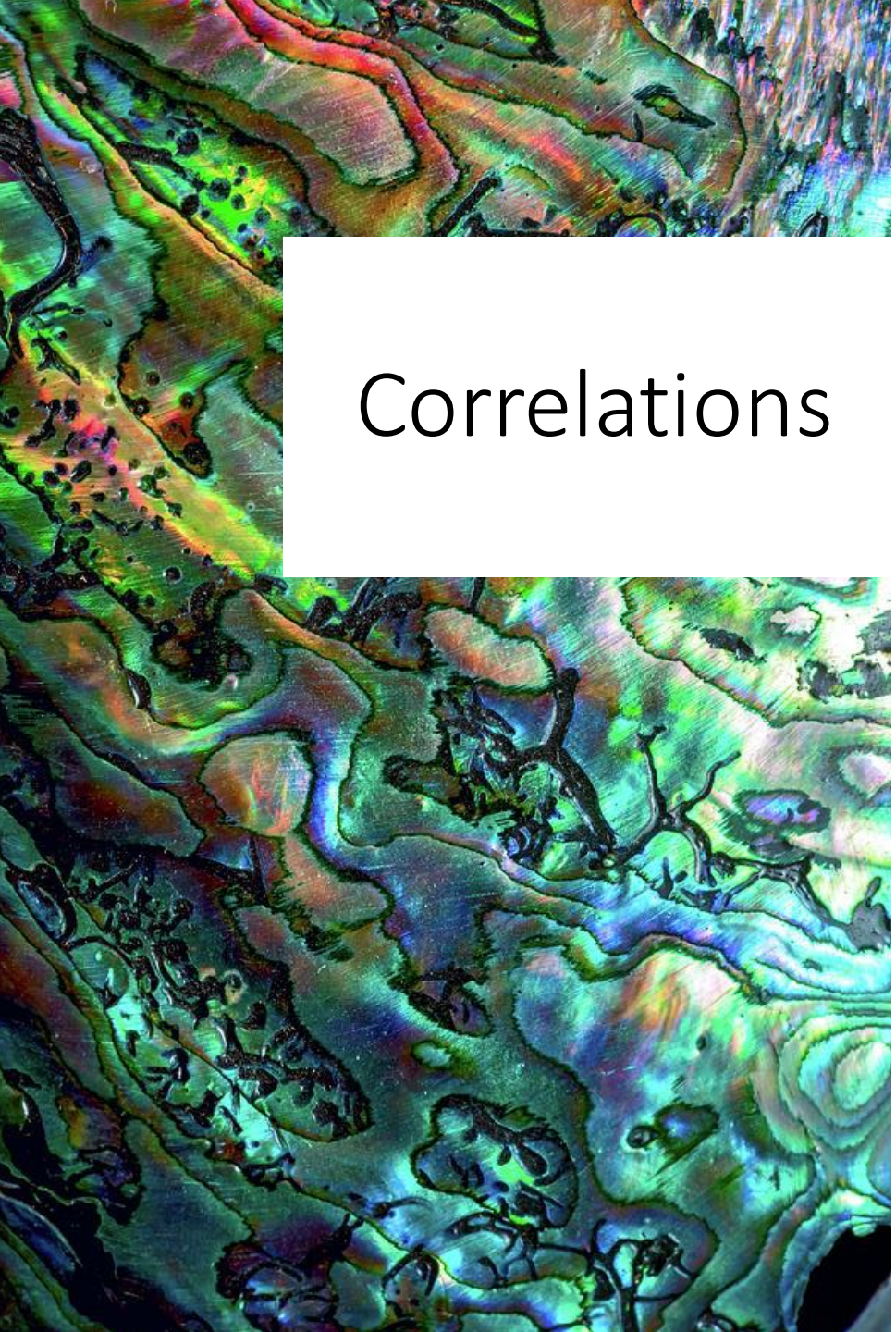
Bivariate Distribution



Average Meat Weight by Age
There are no shells older than 31 years old, it is an indication of survival and overharvesting
Meat weight of Males and Females are similar based on age



Note: data collected was measured in mm and gr and scaled by 200

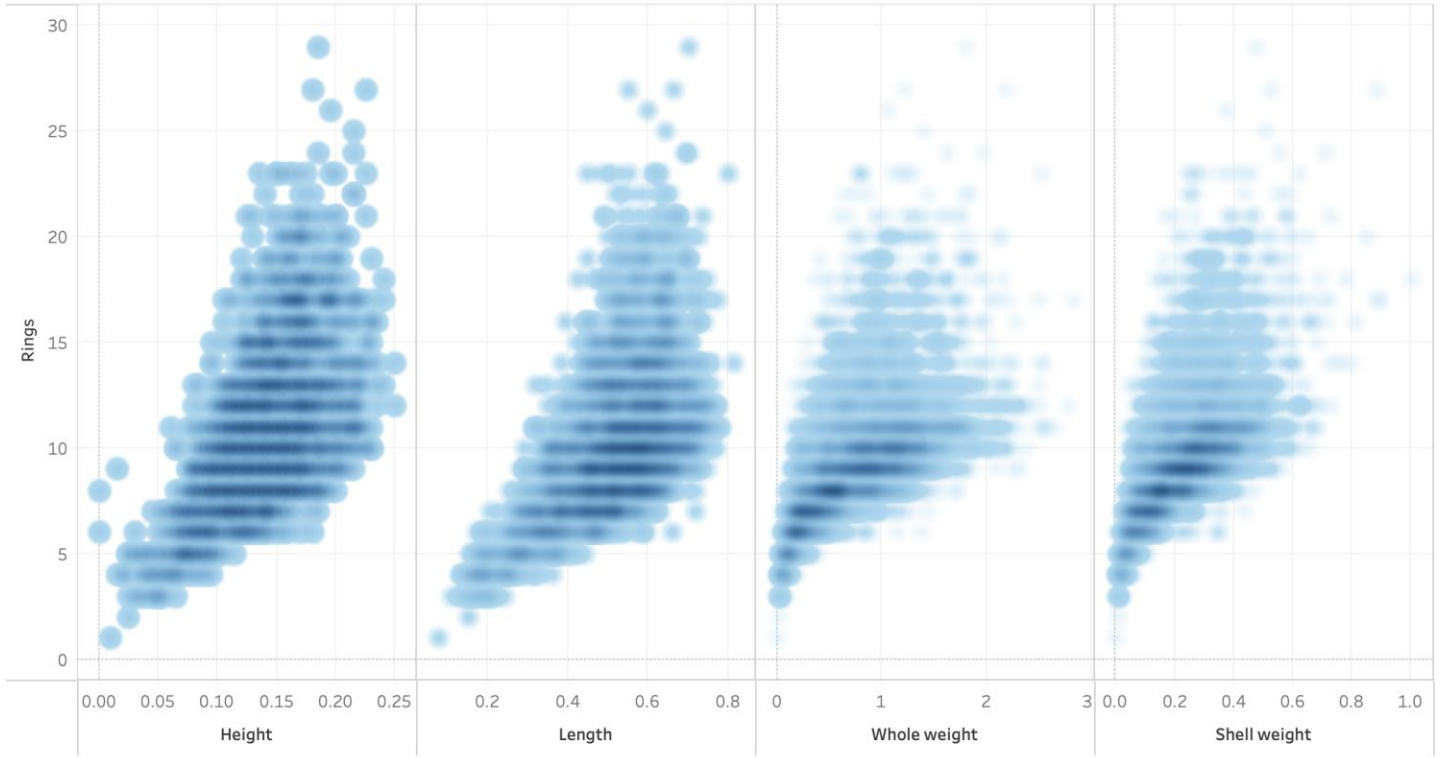


Correlations

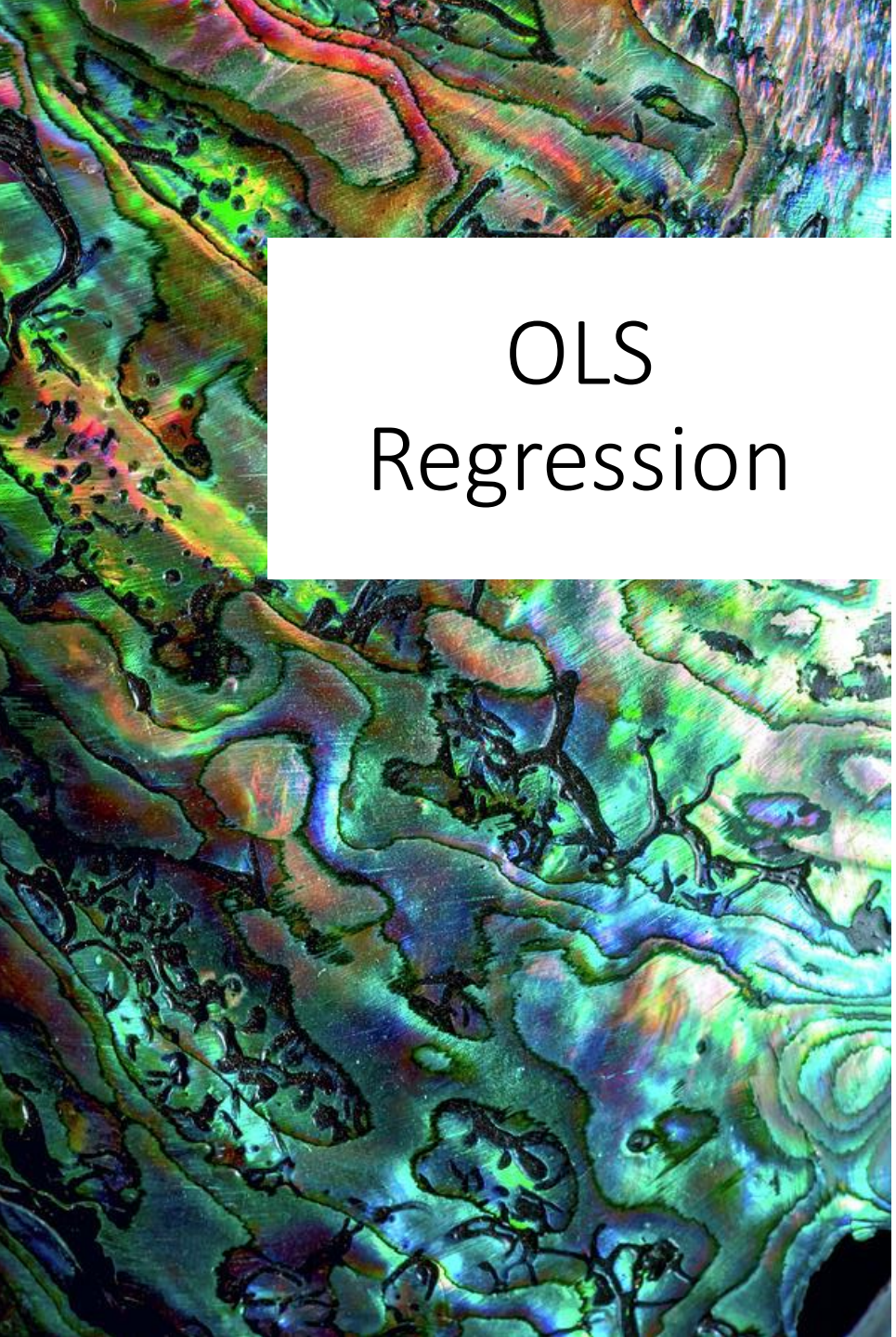
	length	diameter	height	whole_weight	meat_weight	gut_weight	shell_weight	age	F	I	M
length											
diameter	0.99										
height	0.83	0.83									
whole_weight	0.93	0.93	0.82								
meat_weight	0.9	0.89	0.77	0.97							
gut_weight	0.9	0.9	0.8	0.97	0.93						
shell_weight	0.9	0.91	0.82	0.96	0.88	0.91					
age	0.56	0.57	0.56	0.54	0.42	0.5	0.63				
F	0.31	0.32	0.3	0.3	0.26	0.31	0.31	0.25			
I	-0.55	-0.56	-0.52	-0.56	-0.52	-0.56	-0.55	-0.44	-0.46		
M	0.24	0.24	0.22	0.25	0.25	0.24	0.24	0.18	-0.51	-0.52	

Linearity

Relations between the important Variables



Height, Length, Whole weight and Shell weight vs. Rings. Details are shown for Sex. The view is filtered on Exclusions (Height,Rings,Sex), which keeps 875 members.



OLS Regression

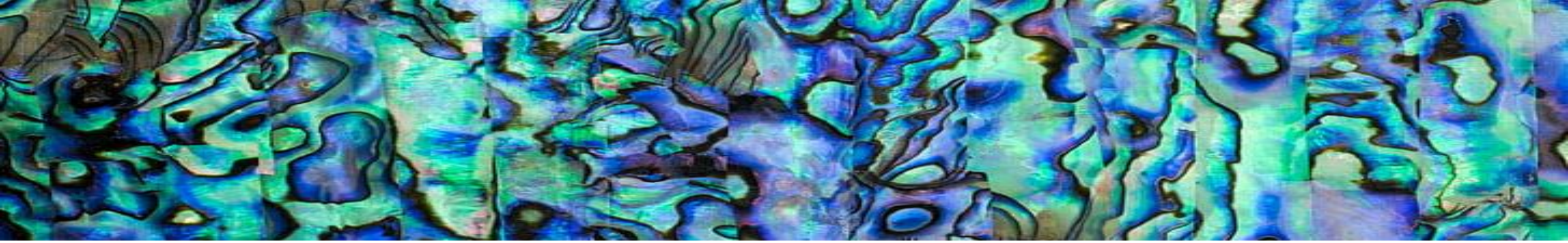
OLS Regression Results

Dep. Variable:	rings	R-squared:	0.559
Model:	OLS	Adj. R-squared:	0.558
Method:	Least Squares	F-statistic:	460.5
Date:	Wed, 20 Apr 2022	Prob (F-statistic):	0.00
Time:	15:02:37	Log-Likelihood:	-6351.7
No. Observations:	2918	AIC:	1.272e+04
Df Residuals:	2909	BIC:	1.278e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.8989	0.040	250.241	0.000	9.821	9.976
diameter	0.8419	0.120	7.001	0.000	0.606	1.078
height	0.8059	0.101	7.983	0.000	0.608	1.004
whole_weight	4.1218	0.424	9.720	0.000	3.290	4.953
shucked_weight_meat	-4.2360	0.213	-19.878	0.000	-4.654	-3.818
gut_weight	-1.0713	0.164	-6.530	0.000	-1.393	-0.750
shell_weight	1.1752	0.191	6.168	0.000	0.802	1.549
F	0.1043	0.029	3.582	0.000	0.047	0.161
I	-0.2267	0.033	-6.932	0.000	-0.291	-0.163
M	0.1196	0.027	4.436	0.000	0.067	0.172

Omnibus:	616.368	Durbin-Watson:	1.937
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1544.289
Skew:	1.146	Prob(JB):	0.00
Kurtosis:	5.730	Cond. No.	6.09e+15

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 4.81e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.



Conclusion
and Further Research

- **Adjusted R2** shows a max. **55% explained variance** to predict the age of an abalone
- Variables like **nutrition, location, weather conditions** etc. are left out, that might could helped **improve the model**
- **Sex** could be a **better predictor** for age, if **unidentified** abalones were **better splitted**
- **Multicollinearity** allows no valid and logical information about the variable that has the biggest impact

! There are more variables needed to fully predict the age of abalones !