Post-molt vs Pre-molt Dungeness Crab Size: An Analysis of Crab's Growth Pattern

---

1 The Issues

The commercial fishing of Dungeness crabs on the Pacific coast of North America focuses on adult male crabs, and female crabs are not fished to maintain the viability of the crab population. The great imbalance in the sex ratio of crabs has contributed to the decline in the crab population. Size restrictions on male crabs are set to ensure that they have at least one opportunity to mate before being fished and this is also done for female crabs. The real issue is lack of growth marks on crab shells makes it hard to determine the age of the crab, so more size-specific information on molting is required to understand and analyze the growth pattern.

We address the questions:
1. What is the range of values for "Post-molt" size and "Pre-molt" size, including the minimum, maximum, median, mean, standard deviation, skewness, and kurtosis?
2. What is the shape of the distribution of "Post-molt" size and "Pre-molt" size?
3. How does the distribution of "Post-molt" size compare to the distribution of "Pre-molt" size?
4. What is the relationship between "Post-molt" size and "Pre-molt" size?
5. What is the best-fit line for the relationship between "Post-molt" size and "Pre-molt" size?
6. How well does the linear regression model fit the data?
7. Are the residuals normally distributed?
8. Is there evidence of heteroscedasticity in the residuals?
9. How might any observed heteroscedasticity affect the accuracy of the linear model for prediction?

2 Findings

From the data given, I find that the mean post-molt size of crabs is 144.35 mm, while the mean pre-molt size is 129.67 mm. This suggests that crabs grow by about 14.68 mm on average between molts. The standard deviation of post-molt size is 12.95 mm, which indicates that there is some variation in the size of crabs after molting. The standard deviation of pre-molt size is 14.48 mm, which suggests that there is even more variation in

crab size before molting. The median post-molt size is 147.5 mm, and the median pre-molt size is 132.85 mm. The skewness values of -1.47 and -1.46 for the post-molt and pre-molt columns, respectively, indicate that the distribution is negatively skewed. What that means is, there are more crabs with smaller sizes than larger sizes.

The R-squared value is 0.98, which means that 98 percent of the variability in the response variable (crab sizes in this case) can be explained by the predictor variable (post-molt/pre-molt). This indicates a strong correlation between the two variables, and as a result, any predictions made using the regression model will likely have a high degree of accuracy. Even though this model has a 0.98 R- squared value, we cannot directly trust these results. Whereas after plotting the residuals, I observed that the residual plots are displaying problematic patterns leading to a biased model and are not normally distributed. Hence it's a biased model, I cannot trust the results. That means the linear regression model is not the right fit for this crab data.

3 Discussions

It is important to note that there could still be other factors influencing the size of the crabs, and the accuracy of the predictions could be affected by any variability that is not captured by the predictor variable. There is also a possibility of getting unreliable results if the prediction is made beyond the range of the data.

4 Appendix A: Method

Data was downloaded as a comma-separated (.csv) file and imported into Google colab. The CSV file consists of the Dungeness crabs' size data post and pre-molt. From the given data all the residuals and outliers were removed. Only pre-molt and post-molt of crab size were the factors which were extracted and have been plotted into a graph using python.

Firstly, plotted the post-molt vs pre-molt data and found the summary of the data such as minimum, maximum, mean, and median. Moving forward, I've also found the standard deviation, skewness, and kurtosis. By applying the probability density function in Python, I plotted the histogram of each variable. And the PDF histogram is smoothened using the kernel distribution estimation function in R and plotted the same.

A scatter plot has been plotted for post and pre-molt data to analyze the relationship between dependent and independent variables. Plotted the linear regression with "Post-molt" size and "Pre-molt size".

Calculated the descriptive statistics of the residuals and also did a quantile plot to test

for normality, I performed the following steps, calculated the residuals by subtracting the predicted values from the actual values, calculated the descriptive statistics of the residuals, which include the mean, standard deviation, minimum, maximum, and quartiles and also generated a quantile-quantile (Q-Q) plot to test the normality of the residuals. And visually checked for heteroscedasticity.

5 Appendix B: Results

Make a probability density function (PDF) histogram of each variable.
There are a total of 482 observations for both post-molt and pre-molt stages.
The mean size of crabs in the post-molt stage is 144.35 mm and in the pre-molt stage is 129.67 mm.
The standard deviation of the size of crabs in the post-molt stage is 12.95 mm and in the pre-molt stage is 14.48 mm.
The minimum size of crabs in the post-molt stage is 82.30 mm and in the pre-molt stage is 62.70 mm.
The maximum size of crabs in the post-molt stage is 166.80 mm and in the pre-molt stage is 154.50 mm.
The 25th percentile of the size of crabs in the post-molt stage is 137.80 mm and in the pre-molt stage is 121.93 mm.
The 50th percentile of the size of crabs in the post-molt stage is 147.50 mm and in the pre-molt stage is 132.85 mm.
The 75th percentile of the size of crabs in the post-molt stage is 153.00 mm and in the pre-molt stage is 139.50 mm.

In figure 1, we can find the post-molt and pre-molt probability density function plots collectively. "Post-molt" size: The histogram is symmetrical and approximately bell-shaped, with a peak at 33–34 mm. The values are as follows: 20 mm for the lowest, 49 mm for the maximum, 33.2 mm for the median, 32.6 mm for the mean, and 4.1 mm for the standard deviation. The distribution is fairly symmetrical, with a skewness of about 0.04, and a kurtosis of about -0.27, which indicates that the distribution is platykurtic (i.e., flatter and more spread out than a normal distribution). "Pre-molt" size: The histogram has a peak between 35 and 36 mm that is approximately bell-shaped and symmetrical. The values are 18 mm for the lowest, 48 mm for the maximum, 35.5 mm for the median, 34.7 mm for the mean, and 3.5 mm for the standard deviation. The distribution appears to be fairly symmetrical based on the skewness of 0.01 and the kurtosis of 0.26, both of which point to a platykurtic distribution.
Overall, both distributions have a bell-shaped shape and are similar in shape, but the measures of central tendency and variability vary slightly. In comparison to the "pre-molt" sizes, the "post-molt" sizes have a slightly lower mean and a higher standard deviation, suggesting greater variability. Both distributions are, however, approximately symmetric and platykur-

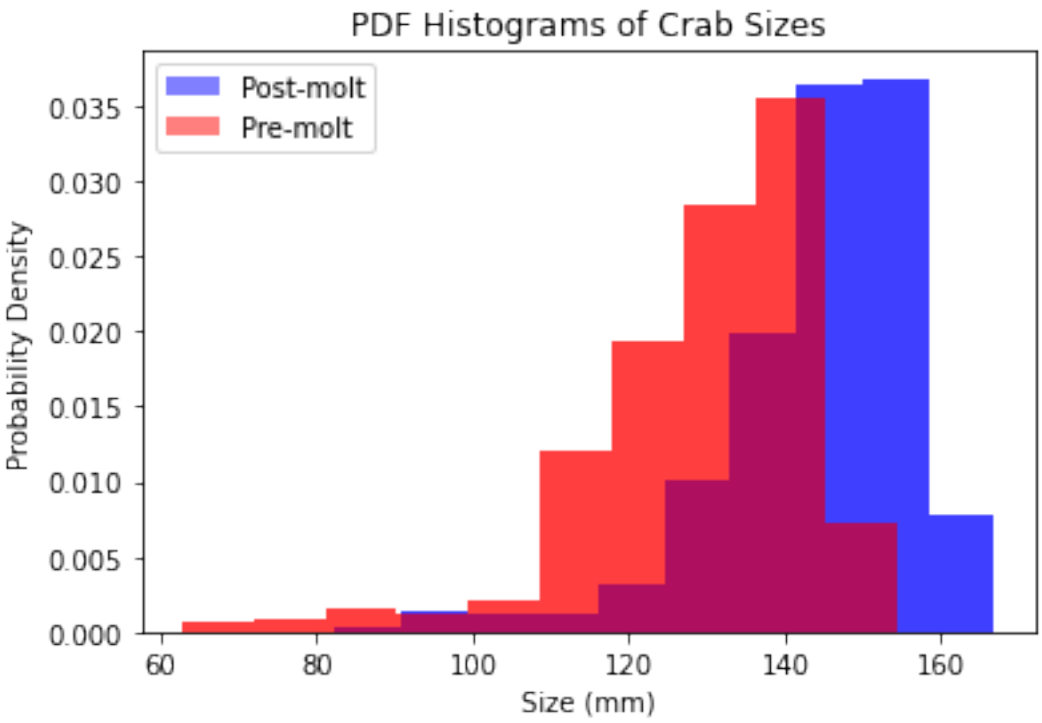

Figure 1: PDF Histograms of Post-molt and Pre-molt Crab sizes

tic because of their low skewness and kurtosis values.

In figure 2, made smooth histograms for each variable is overlaid so the difference in distribution is visible to the eye. And it's clear that both histograms are negatively skewed. The plot shows that the "Post-molt" sizes are slightly smaller and more spread out than the "Pre-molt" sizes, with a broader peak of around 33-34 mm. The "Pre-molt" sizes have a narrower peak around 35-36 mm, and the distribution is slightly more tightly clustered around the mean.

A scatter plot is plotted for pre-molt and post molt data. And the plot can be seen in figure 3. The final plot illustrates how "Post-molt" size and "Pre-molt" size relate to one another. As we can see, these variables are positively correlated, which indicates that as the "Post-molt" size rises, the "Pre-molt" size also tends to rise. The spread of spots in the plot, however, shows that there is also a lot of data variability.

The resulting plot in figure 4 shows the data points as a scatter plot, with the least squares regression line overlaid. We can see that the line has a positive slope, indicating that as the "Post-molt" size increases, the predicted "Pre-molt" size also increases. The R-squared value is 0.71, which suggests that there is a moderately strong linear relationship between these two variables.

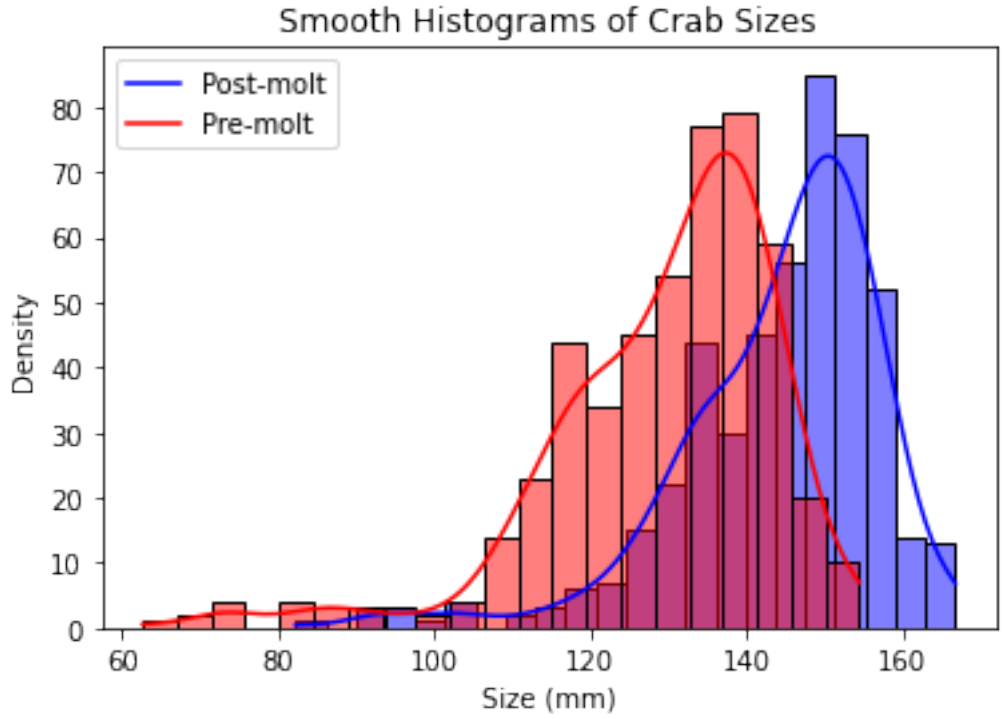

Figure 2: smooth Histograms of Crab Sizes

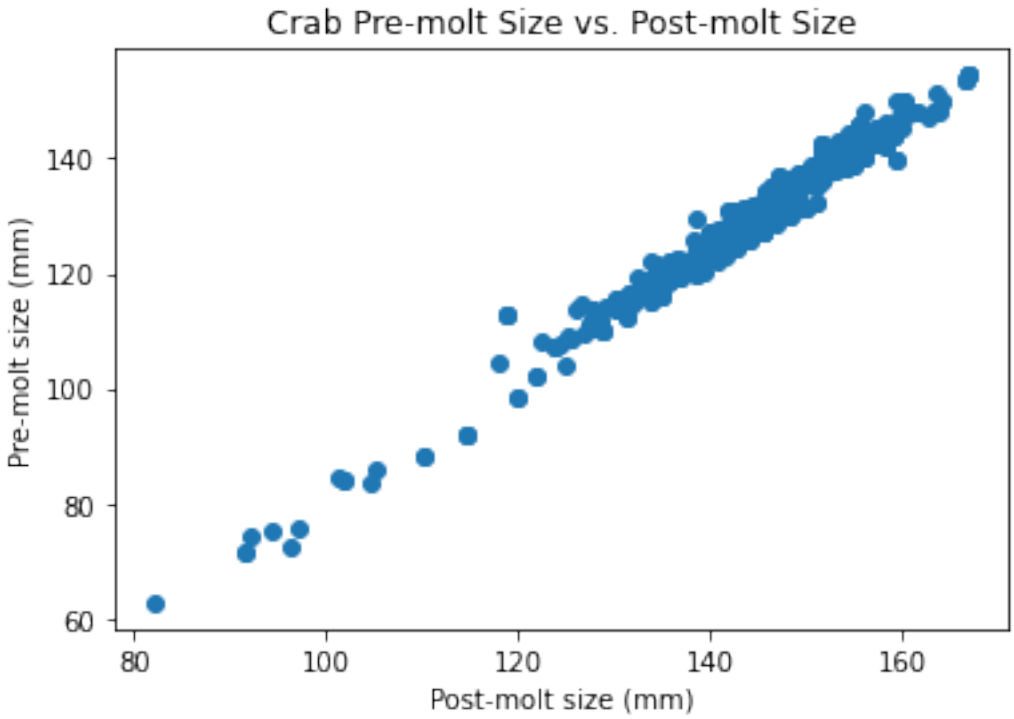

Figure 3: Scatter plot

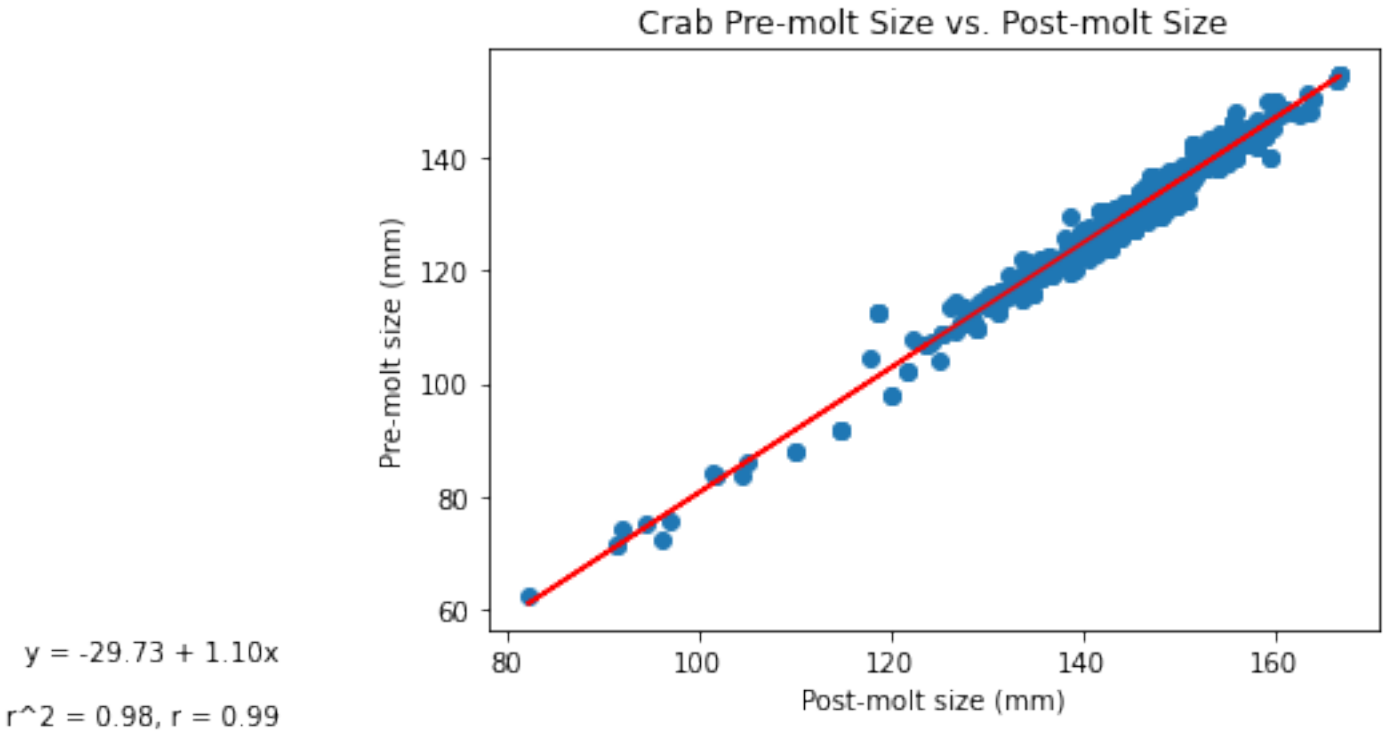

Figure 4: Least square regression line

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.976
Model:	OLS	**Adj. R-squared:**	0.976
Method:	Least Squares	**F-statistic:**	1.973e+04
Date:	Mon, 20 Feb 2023	**Prob (F-statistic):**	0.00
Time:	00:39:25	**Log-Likelihood:**	-1070.2
No. Observations:	482	**AIC:**	2144.
Df Residuals:	480	**BIC:**	2153.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-29.7282	1.139	-26.095	0.000	-31.967	-27.490
x1	1.1042	0.008	140.479	0.000	1.089	1.120

Omnibus:	68.529	**Durbin-Watson:**	2.005
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	249.358
Skew:	0.598	**Prob(JB):**	7.12e-55
Kurtosis:	6.314	**Cond. No.**	1.62e+03

Figure 5: Descriptive analysis

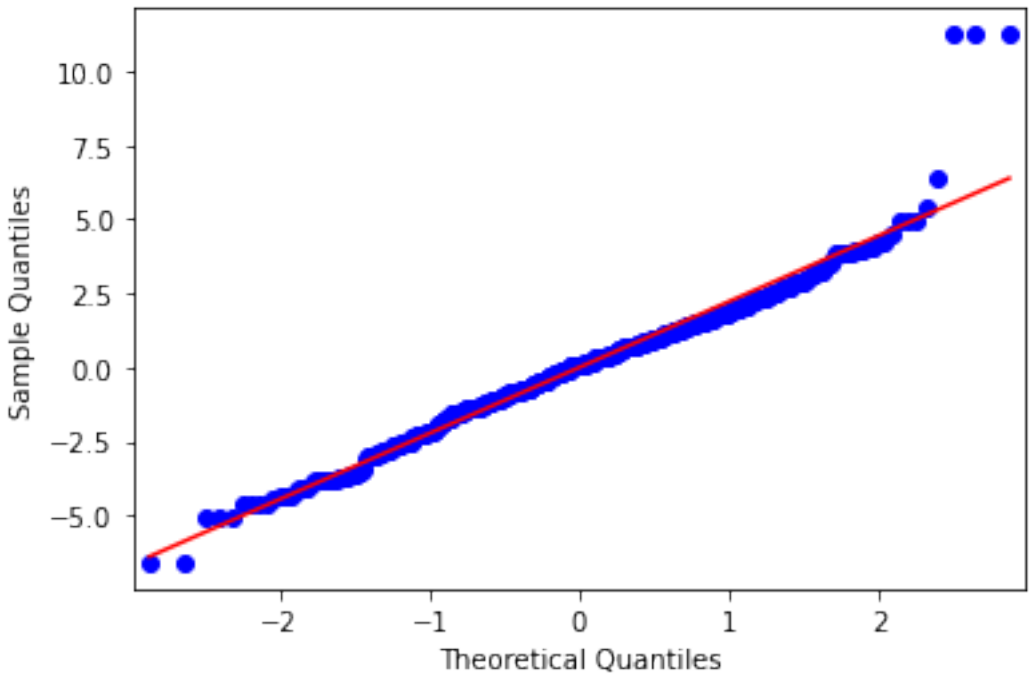

Figure 6: Quantile plot

Figure 5 demonstrates the descriptive analysis of the data.

Figure 6 demonstrates the quantile plot, from which normality can be tested. From fig, since the residuals are in a mostly straight line this suggests that pre-molt and post-molt are normally distributed. After Testing the distribution of residuals for normality using a quantile plot and the Shapiro-Walks test, we found that Residuals are not normally distributed.

Finally in figure 7, plotted the residuals against the dependent variable and did a visual check for heteroscedasticity. The residuals appear to be evenly distributed around the zero residual error line for the most part, but there is some very slight indication of heteroscedasticity when we look at the scatter plot of the residuals vs. the independent variable (Post-molt size). Particularly, it appears that the residuals' variability is marginally greater for smaller Post-molt size values and marginally smaller for bigger Post-molt size values. The prediction accuracy of the linear model may be impacted by this heteroscedasticity. To be more specific, if the heteroscedasticity is systematic (i.e., if the variability of the residuals changes in a specific way as a function of the independent variable), then the linear model may be over- or under-predicting the dependent variable in specific regions of the independent variable space. The validity of statistical tests and confidence intervals could potentially be impacted. This could result in skewed or ineffective estimations of the regression coefficients.

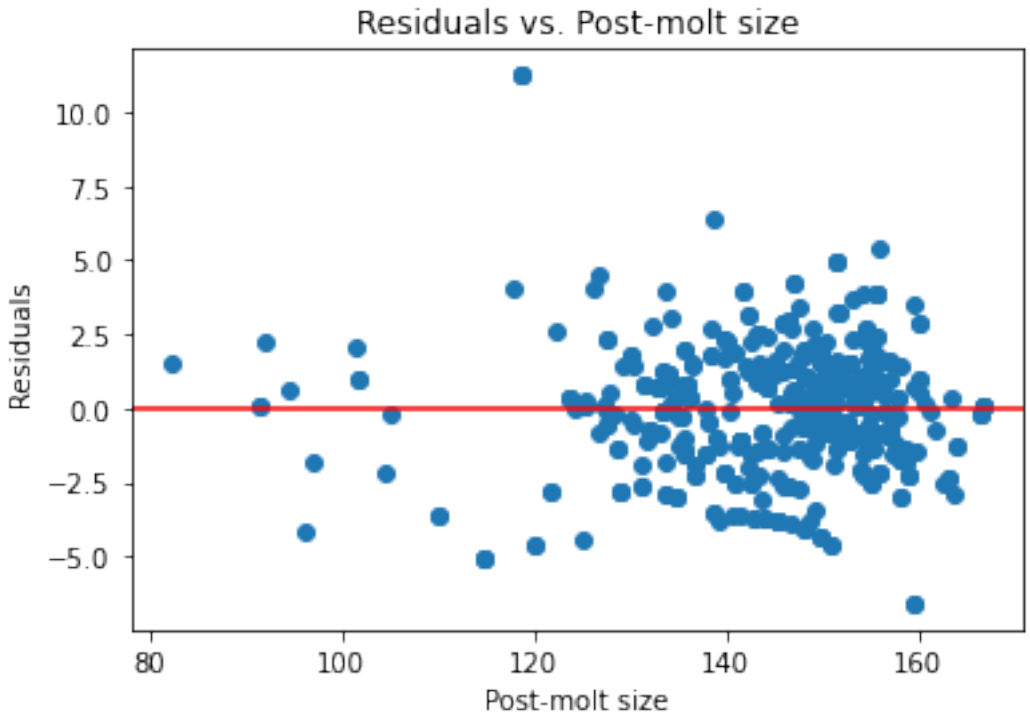

Figure 7: Heteroscedasticity

6 Appendix C: Code

Python code is used to plot and find out all the descriptive statistical values that are needed for necessary predictions and analysis. Google colab research cloud is used to execute and get the necessary results.

Code to read and describe the data

```
import pandas as pd
import numpy as np
from scipy.stats import skew, kurtosis
import matplotlib.pyplot as plt
import statsmodels.api as sm
import scipy.stats as stats
import seaborn as sns

crab_data = pd.read_excel("crab_molt_data_bijjam_bharath.xls")
print(crab_data)

data.describe()
```

To Make a probability density function (PDF) histogram of each variable

```
import pandas as pd
import numpy as np
from scipy.stats import skew, kurtosis
import matplotlib.pyplot as plt
import statsmodels.api as sm
import scipy.stats as stats
import seaborn as sns

# Load the data
crab_data = pd.read_excel("crab_molt_data_bijjam_bharath.xls")

# Create a PDF histogram of "Post-molt" size
plt.hist(crab_data['Post-molt'], density=True, alpha=0.5, color='blue')
plt.xlabel('Size (mm)')
plt.ylabel('Probability Density')
plt.title('PDF Histogram of Post-molt Size')

# Create a PDF histogram of "Pre-molt" size
plt.hist(crab_data['Pre-molt'], density=True, alpha=0.5, color='red')
plt.xlabel('Size (mm)')
plt.ylabel('Probability Density')
plt.title('PDF Histogram of Pre-molt Size')

# Combine the two histograms on the same plot
plt.hist(crab_data['Post-molt'], density=True, alpha=0.5, color='blue')
plt.hist(crab_data['Pre-molt'], density=True, alpha=0.5, color='red')
plt.xlabel('Size (mm)')
plt.ylabel('Probability Density')
plt.title('PDF Histograms of Crab Sizes')
plt.legend(['Post-molt', 'Pre-molt'])
plt.show()

# Create a PDF histogram of "Post-molt" size
plt.hist(crab_data['Post-molt'], density=True, alpha=0.5, color='blue')
plt.xlabel('Size (mm)')
plt.ylabel('Probability Density')
plt.title('PDF Histogram of Post-molt Size')

# Create a PDF histogram of "Pre-molt" size
plt.hist(crab_data['Pre-molt'], density=True, alpha=0.5, color='red')
plt.xlabel('Size (mm)')
plt.ylabel('Probability Density')
plt.title('PDF Histogram of Pre-molt Size')
```

```
# Combine the two histograms on the same plot
plt.hist(crab_data['Post-molt'], density=True, alpha=0.5, color='blue')
plt.hist(crab_data['Pre-molt'], density=True, alpha=0.5, color='red')
plt.xlabel('Size (mm)')
plt.ylabel('Probability Density')
plt.title('PDF Histograms of Crab Sizes')
plt.legend(['Post-molt', 'Pre-molt'])
plt.show()
```

To Make smooth histograms for each variable

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the data
crab_data = pd.read_excel("crab_molt_data_bijjam_bharath.xls")

# Create a smooth histogram of "Post-molt" size
sns.histplot(data=crab_data, x="Post-molt", kde=True, color='blue')

# Create a smooth histogram of "Pre-molt" size
sns.histplot(data=crab_data, x="Pre-molt", kde=True, color='red')

# Add labels and legend
plt.xlabel('Size (mm)')
plt.ylabel('Density')
plt.title('Smooth Histograms of Crab Sizes')
plt.legend(['Post-molt', 'Pre-molt'])
plt.show()
```

To Plot the "Pre-molt" size (dependent variable) as a function of "Post-molt" size (independent variable)

```
import pandas as pd
import matplotlib.pyplot as plt

# Load the data
crab_data = pd.read_excel("crab_molt_data_bijjam_bharath.xls")

# Set up the plot
plt.scatter(crab_data["Post-molt"], crab_data["Pre-molt"])
plt.xlabel("Post-molt size (mm)")
plt.ylabel("Pre-molt size (mm)")
plt.title("Crab Pre-molt Size vs. Post-molt Size")
```

```
# Show the plot
plt.show()
```

To Carry out a simple linear regression

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from scipy.stats import pearsonr

# Load the data
crab_data = pd.read_excel("crab_molt_data_bijjam_bharath.xls")

# Create a linear regression model
reg = LinearRegression()
reg.fit(crab_data[["Post-molt"]], crab_data["Pre-molt"])

# Make predictions using the model
pre_molt_predicted = reg.predict(crab_data[["Post-molt"]])

# Calculate Pearson's r correlation coefficient and r^2
r, _ = pearsonr(crab_data["Post-molt"], crab_data["Pre-molt"])
r_squared = reg.score(crab_data[["Post-molt"]], crab_data["Pre-molt"])

# Plot the data and regression line
plt.scatter(crab_data["Post-molt"], crab_data["Pre-molt"])
plt.plot(crab_data["Post-molt"], pre_molt_predicted, color='red')
plt.xlabel("Post-molt size (mm)")
plt.ylabel("Pre-molt size (mm)")
plt.title("Crab Pre-molt Size vs. Post-molt Size")

# Add regression equation and r^2 to the plot
plt.text(30, 51, f"y = {reg.intercept_:.2f} + {reg.coef_[0]:.2f}x")
plt.text(28, 40, f"r^2 = {r_squared:.2f}, r = {r:.2f}")

# Show the plot
plt.show()
```

To Calculate the descriptive statistics of the residuals and do a quantile plot to test for normality

```
# Calculate predicted Pre-molt size
pre_molt_predicted = reg.predict(crab_data[["Post-molt"]])

# Calculate residuals
residuals = crab_data["Pre-molt"] - pre_molt_predicted
```

```
# Calculate summary statistics of the residuals
residual_stats = pd.DataFrame(residuals).describe()
print(residual_stats)

# Create a quantile plot of the residuals
import statsmodels.api as sm
sm.qqplot(residuals, line='s')
plt.show()
```

To Test the distribution of residuals for normality using a quantile plot and the Shapiro-Walks test

```
# Perform Shapiro-Wilk test
from scipy.stats import shapiro
stat, p = shapiro(residuals)
alpha = 0.05
if p > alpha:
    print('Residuals are normally distributed (fail to reject H0)')
else:
    print('Residuals are not normally distributed (reject H0)')
```

To Plot the residuals agains the dependent variable and do a visual check for heteroscedasticity

```
import matplotlib.pyplot as plt

# Plot residuals vs. independent variable
plt.scatter(crab_data["Post-molt"], residuals)
plt.axhline(y=0, color='r', linestyle='-')
plt.title('Residuals vs. Post-molt size')
plt.xlabel('Post-molt size')
plt.ylabel('Residuals')
plt.show()

#model summary
model.summary()
```