Unraveling the Mysteries of Crab Molt: A Statistical Analysis of Post-Molt and Pre-Molt Sizes

A Punchline report

The Issues:

The dataset used in this lab was collected as part of a study on female Dungeness crabs. The study was conducted by researchers Hankin, Diamond, Mohr, and Ianelli with help from the California Department of Fish and Game and commercial crab fishers from northern California and southern Oregon. The dataset includes two types of data. The first is the pre-molt and post-molt widths of the carapaces of 472 female Dungeness crabs, which were collected by scientists and commercial fisheries over three fishing seasons in 1981, 1982, and 1992.

We address the questions:

- Exploration of the relationship between post-molt size and pre-molt size using a scatterplot and simple linear regression analysis.
- Testing of the assumptions of linear regression by analyzing the residuals' distribution using quantile plots and the Shapiro-Walks test and heteroscedasticity.

Findings:

The EDA of the dataset showed some interesting insights in the datasets. There is a considerable difference in sizes of molts noticed when plotted smoothed histograms.

Also, after fitting the Ordinal Least Squares model, the residuals are missing halfway through the plot, it could indicate that there is a systematic pattern in the residuals that is related to the values of the dependent variable. This indicated the presence of heteroscedasticity,

which means that the variance of the errors is not constant across the range of the dependent variable. In a linear model, heteroscedasticity can arise when the spread of the residuals increases or decreases systematically as the predicted values of the dependent variable increase or decrease. Overall, these findings suggest that the linear regression model may not be the best fit for the crab data, as the residuals are not normally distributed and have heavy tails. Other models, such as a generalized linear model, may be more appropriate for this data.

Discussion:

The descriptive statistics of residuals show that the mean value of residuals is very close to zero. The variance of residuals is 4.547, which indicates that the residuals are spread out from the mean value. The skewness value of 0.950 shows that the residuals are slightly skewed to the right. The kurtosis value of 7.575 indicates that the residuals are very peaked and have heavy tails, which means that the residuals are not normally distributed.
The Shapiro-Wilk test statistic of 0.936 suggests that the residuals are not normally distributed, which is also supported by the high kurtosis value.
The Pearson's r^2 value of 0.983 indicates a strong correlation between the pre-molt and post-molt sizes, which supports the use of linear regression to model the relationship between these two variables.

Appendix A: Methods

Data collection: The crab molt data was collected by measuring the size of crabs in the laboratory both before and after molting.
Variable creation: The dataset contains only two variables Post-molt and Pre-molt which represent the size of the crabs before and after molting.
Analytic method:
The Excel sheet containing the data was imported into a dataframe. Only two variables, Pre-Molt and Post-Molt, were extracted and their descriptive statistics (minimum, maximum, median, mean, standard deviation, skewness, and kurtosis) were calculated. Histogram plots, including probability density functions, were created for both factors

separately and also overlaid on top of each other. A scatter plot was also created to show the relationship between PreMolt and PostMolt size. Least square linear regression was applied and Pearson (r^2) regression was calculated. Descriptive statistics were performed for the residuals obtained from the least square regression. A histogram plot with density lines was created to plot the residuals and checked for their normality using the Quantile plot test and Shapiros Walks test. Finally, to check for the presence of heteroscedasticity, the residuals were plotted against the dependent variable (PreMolt).

Appendix B: Results

The dataset contains 450 records and 2 columns. And the descriptive stats are as follows.

	Post-molt	Pre-molt
count	450.000000	450.000000
mean	144.140444	129.488000
std	14.900687	16.314499
min	38.800000	31.100000
25%	137.100000	121.100000
50%	147.500000	133.300000
75%	154.200000	140.400000
max	166.500000	155.100000

The kurtosis and skewness are calculated using standard dataframe methods and the values are as follows.

Skewness → Post-molt: -1.966404
Pre-molt: -1.742212

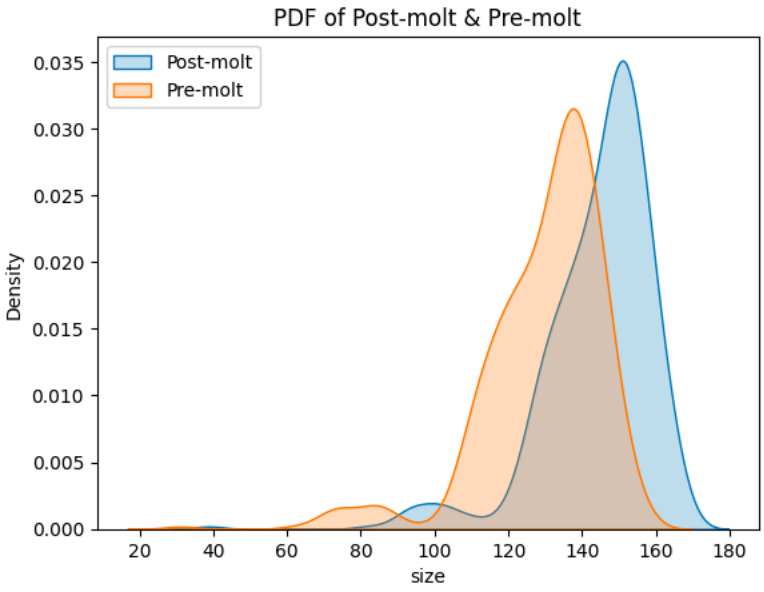
Kurtosis → Post-molt: 7.296487
Pre-molt: 4.984250

The "Post-molt" variable has a highly negatively skewed distribution, with a skewness value of -1.966404. This suggests that the distribution has a

long tail on the left side and that there may be a large number of very small values relative to the rest of the data. The high positive kurtosis value of 7.296487 suggests that the distribution is more peaked than a normal distribution, with a sharper peak and heavier tails.

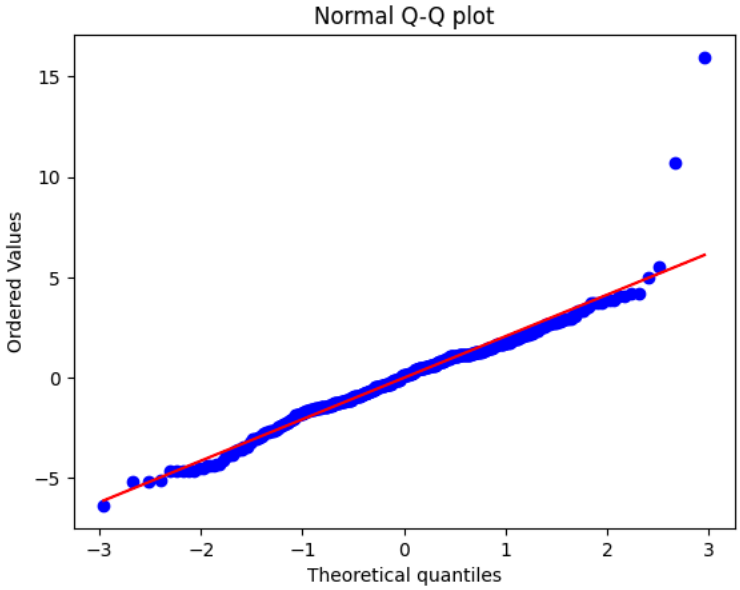
The "Pre-molt" variable also has a negatively skewed distribution, although less so than the "Post-molt" variable, with a skewness value of -1.742212. The kurtosis value of 4.984250 suggests that the distribution is still more peaked than a normal distribution, but not as much as the "Post-molt" variable.

The smooth histogram approximations to the size of disparity shows a visible difference between Pre & Post molt sizes.

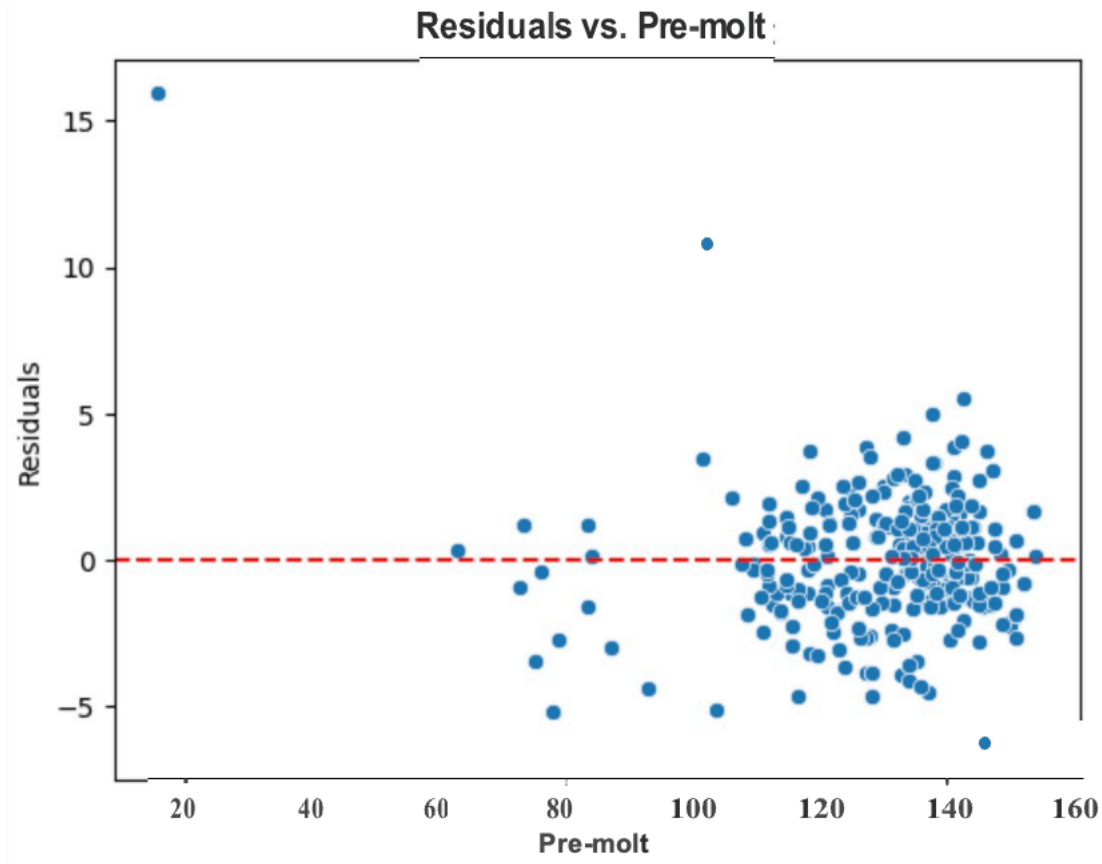


The scatter plot of residuals with regression line is as shown in below figure

A Quantile plot is drawn for the residuals to check for normatily of the residuals and test statistic value says in shapiro test the test statistic of 0.9356176853179932 is relatively close to 1, which suggests that the data may be somewhat close to normally distributed. However, the p-value of 0.000000000000499 is extremely small, indicating that there is very strong evidence against the null hypothesis of normality.



Finally, plotted the residuals against the dependent variable to check for heteroscedasticity.



Appendix C: Code

The statistical analysis is performed using following code and a linear model is trained using sklearn package of python.

a. Importing required libraries and loading the dataset into pandas dataframe

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

df = pd.read_excel("./Datasets/crab_molt_data_namala_durga.xls")
```

b. Dataframe describe method is used to generate descriptive stats of all the numerical columns.

```
df.describe()
```

✓ 0.0s

	Post-molt	Pre-molt
count	450.000000	450.000000
mean	144.140444	129.488000
std	14.900687	16.314499
min	38.800000	31.100000
25%	137.100000	121.100000
50%	147.500000	133.300000
75%	154.200000	140.400000
max	166.500000	155.100000

c. To computed skewness & kurtosis pandas offers skew, kurtosis methods.

```
df.skew()
```

✓ 0.0s

```
Post-molt   -1.966404
Pre-molt    -1.742212
dtype: float64
```

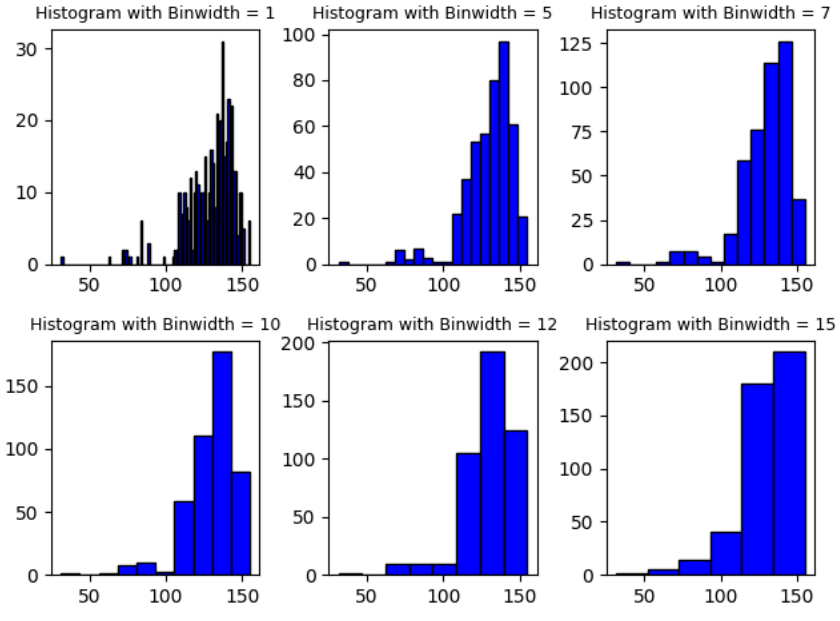
```
df.kurtosis()
```

✓ 0.0s

```
Post-molt    7.296487
Pre-molt     4.984250
dtype: float64
```

d. The following code used for finding out the visually best bin width for histograms.

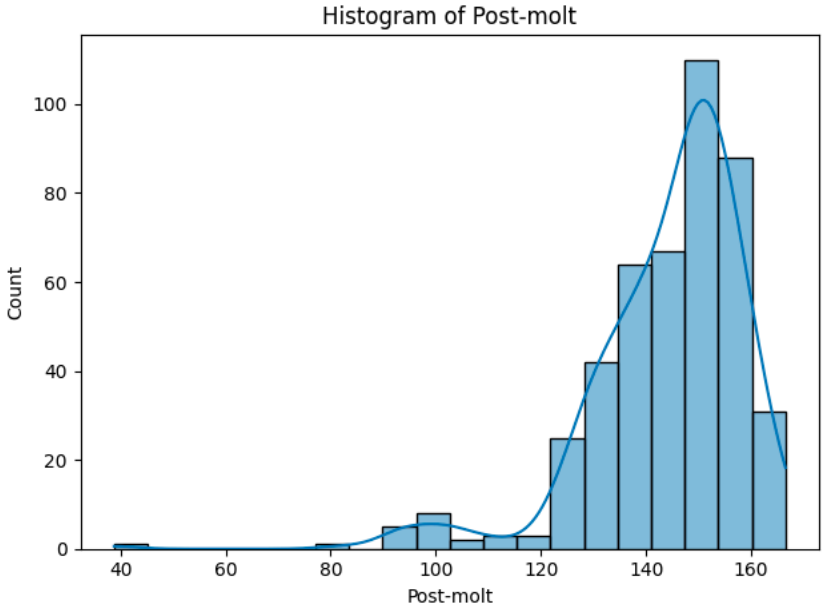
```
for i, binwidth in enumerate([1, 5, 7, 10, 12, 15]):
    ax = plt.subplot(2, 3, i + 1)
    ax.hist(df['Pre-molt'], bins = int(100/binwidth), color = 'blue', edgecolor = 'black')
    ax.set_title('Histogram with Binwidth = %d' % binwidth, size = 9)
plt.tight_layout()
plt.show()
```

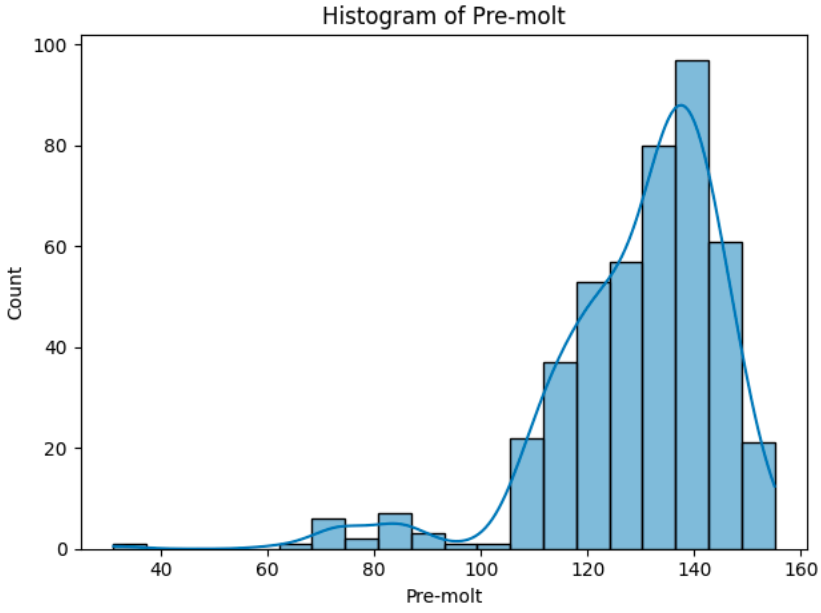


e. To draw histograms for each columns the following code is used.

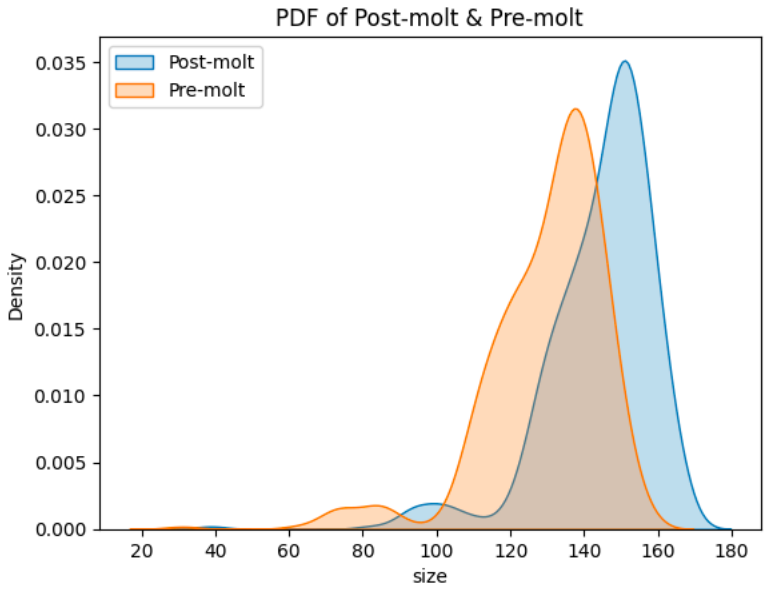
```
for idx, col in enumerate(df.columns):
    fig, ax = plt.subplots()
    sns.histplot(df[col],kde=True, bins=int(100/5), ax=ax)
    ax.set_title(f"Histogram of {col}")
    plt.tight_layout()
```

✓ 0.4s



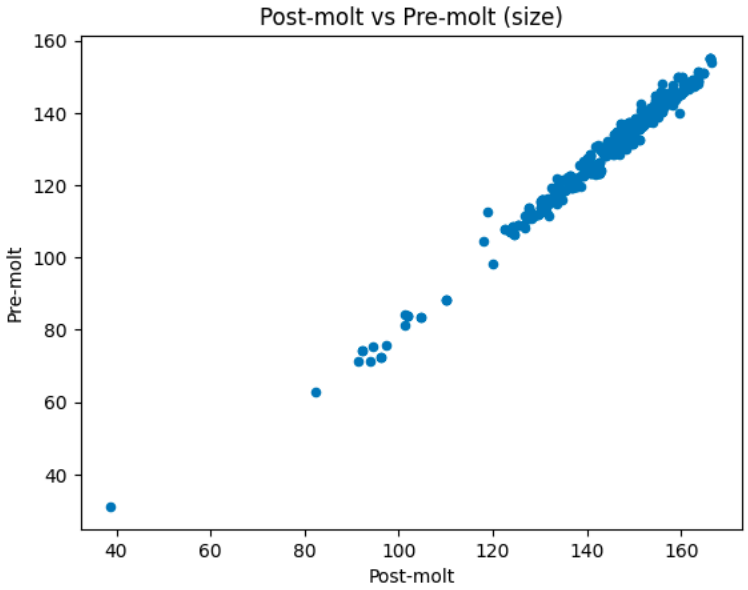


f. Here is the code to compute overlayed smooth histogram(PDF)'s.



g. A scatter plot of Pre-molt vs Post-molt is generated using below code.

```
ax = df.plot.scatter(x='Post-molt', y='Pre-molt')
ax.set_xlabel("Post-molt")
ax.set_ylabel("Pre-molt")
ax.set_title("Post-molt vs Pre-molt (size)")
```



h. The linear model is trained used sklearn's Linear Regression class as shown in the below code.

```
from sklearn.linear_model import LinearRegression

X, y = df[['Post-molt']], df['Pre-molt']
reg = LinearRegression().fit(X, y)
print(f"Pearson's r^2: {reg.score(X, y)}")
```

✓ 0.0s

```
Pearson's r^2: 0.982916643995486
```

- The intercept and coefficient are stored in the following variables by the class and can be accessed as shown below.

```
reg.intercept_, reg.coef_
```

✓ 0.0s

```
(-26.97500009674272, array([1.08548992]))
```

- The residuals are generated using below code and store in a variable named residuals.

```
y_pred = reg.predict(X)
residuals = y - y_pred
```

✓ 0.0s

- Generated a scatter plot with pre-molt vs post-molt with regression line.

```
ax = df.plot.scatter(x='Post-molt', y='Pre-molt', marker='o')
ax.plot(X, y_pred, color='black')
for x, y_actual, y_ in zip(df['Post-molt'], y, y_pred):
    ax.plot((x, x), (y_actual, y_), color='red', linestyle='dashed')
ax.set_title("Scatter plot of Post-molt vs Pre-molt with regression line")
plt.tight_layout()
plt.show()
```

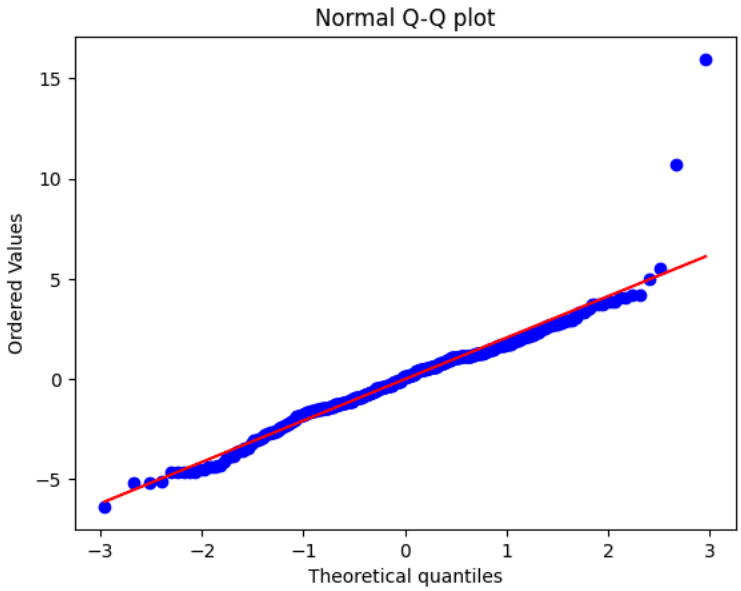
i. The descriptive stats of residuals are computed and Q-Q plot is generated using following code.

```
from scipy import stats

print(f"Descriptive statistics of residuals:")
for k,v in stats.describe(residuals)._asdict().items():
    print(f"{k:9}: {v:}")
fig = plt.figure()
stats.probplot(residuals, dist="norm", plot=plt)
plt.title("Normal Q-Q plot")
```

✓ 0.1s

```
Descriptive statistics of residuals:
nobs     : 450
minmax   : (-6.360642793890349, 15.957991042469601)
mean     : -3.0821764893415904e-14
variance : 4.546955313365687
skewness : 0.9495635966715045
kurtosis : 7.575141014399341
```



j. The t-statstic and p-value are computed using scipy package using the below code.

```
shapiro_stat, p_value = stats.shapiro(residuals)
print(f"Shapiro-Wilk test: statistic={shapiro_stat:.5f}, p-value={p_value:.15f}")
```

✓ 0.0s

```
Shapiro-Wilk test: statistic=0.93562, p-value=0.000000000000499
```

k. To visually check for the heteroscedasticity, plotted residuals against pre-molt data.

```
fig, ax = plt.subplots()
sns.scatterplot(x=y_pred, y=residuals, ax= ax)
ax.axhline(y=0, color="red", linestyle="--")
ax.set_title("Residuals vs. Pre-molt")
ax.set_xlabel("Pre-molt")
ax.set_ylabel("Residuals")
```

✓ 0.1s