Main objective of the analysis that specifies whether your model will be focused on prediction or interpretation:

- The report will be focused on prediction.

Brief description of the data set you chose and a summary of its attributes:

- The dataset that I chose is the Iris dataset from Fisher, 1936. The dataset contains three classes with fifty instances each, with every class referring to an Iris plant. The dataset contains four features related to the sepal and petal lengths and widths, along with the class representing the type of Iris plant being studied.

Brief summary of data exploration and actions taken for data cleaning and feature engineering:

- Data variables, quantile information, summary statistics, and null value checks are performed to determine dataset information integrity. A plot of the sepal length and width, along with the petal length and width are displayed with their associated class of flower. Furthermore, a heatmap is used to depict the correlations between different columns of the dataset. Finally, a histogram is used to show that there are fifty samples for each flower in the associated dataset and that all one-hundred and fifty flowers are converted to integer representations for the linear regression step.

Summary of training at least three linear regression models which should be variations that cover using a simple linear regression as a baseline, adding polynomial effects, and using a regularization regression. Preferably, all use the same training and test splits, or the same cross-validation method:

- Polynomial regression is used with a random state of 72018 and a testing size of 0.3. Bias is not included, and three separate tests are run using degrees of 1, 2, and 3.
- With a degree of 2, we see that the R2 score is 0.9361737337310533.
- With a degree of 3, we see that the R2 score is 0.8433251863794418.
- With a degree of 1, we see that the R2 score is 0.932062765252852.
- When bias is included, we notice a change for tests with a degree of 3:
- With a degree of 3, we see that the R2 score is 0.533853671465681.

A paragraph explaining which of your regressions you recommend as a final model that best fits your needs in terms of accuracy and explainability:

- The best model to use is the one with a degree of two as it had the highest R-squared score. We know that one class is linearly separable from the other two classes due to it having a high petal length relative to its other dimensions and that the other two classes are not linearly separable from one another since they do not have strong features to differentiate each other from. Two degrees for polynomial regression can be used to determine the three separate classes of variables since one degree can be used for petal length to separate one class from the other two, and another degree can be used to separate the remaining two classes from each other.

Summary Key Findings and Insights, which walks your reader through the main drivers of your model and insights from your data derived from your linear regression model:

- Petal length and width are strongly correlated with class, whereas sepal width is negatively correlated with plant class.
- Linear regression can be used to create a strong fit with the dataset, meaning sepal and petal length and width can be used to predict the plant species class.
- Using two degrees without bias results in the strongest fit for linear regression.

Suggestions for the next steps in analyzing this data, which may include suggesting revisiting this model and adding specific data features to achieve a better explanation or a better prediction:

- Using models other than linear regression to analyze the data, such as polynomial regression, ridge regression, or lasso regression. Such methods can be used to help resolve bias in the data and make better predictions given that the length and width measures in the data are heavily correlated with one another.
- Adding more entries to the dataset, with a wider diversity of samples or species of flowers.
- Adding more measurements for the flowers being analyzed by the model.