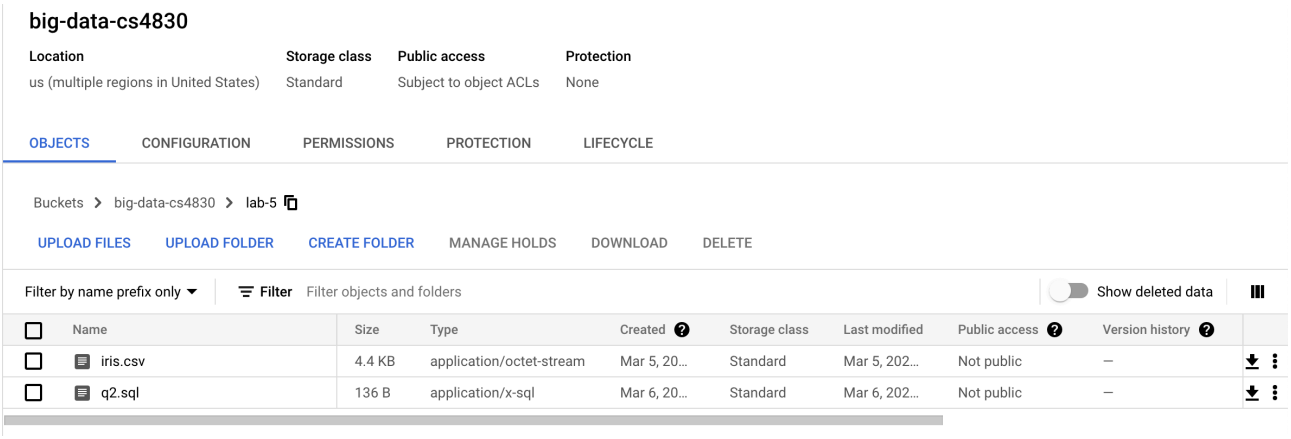
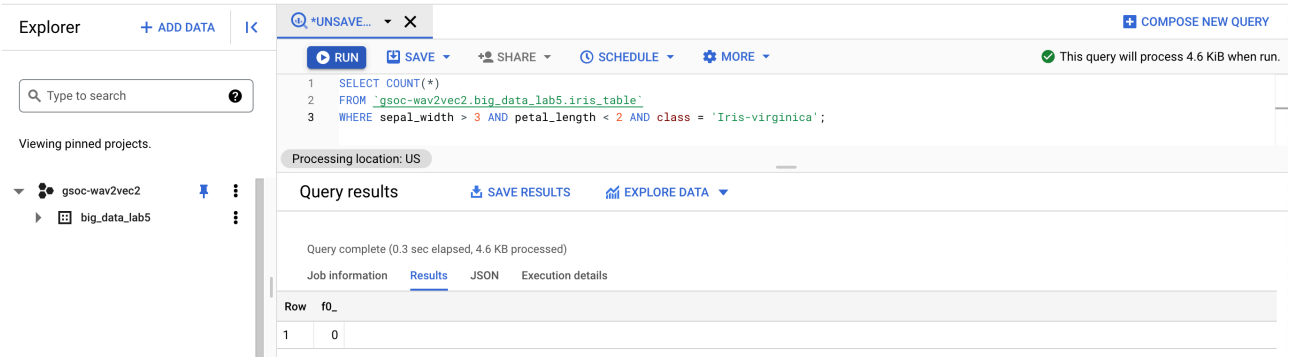
Big Data Lab - lab 5 Assignment

Vasudev Gupta (ME18B182)

1) Iris data is downloaded from the website, converted to CSV and uploaded to the GCS bucket.



2) The count for the given question is 0



Code for question-2 lies in **q2.sql**

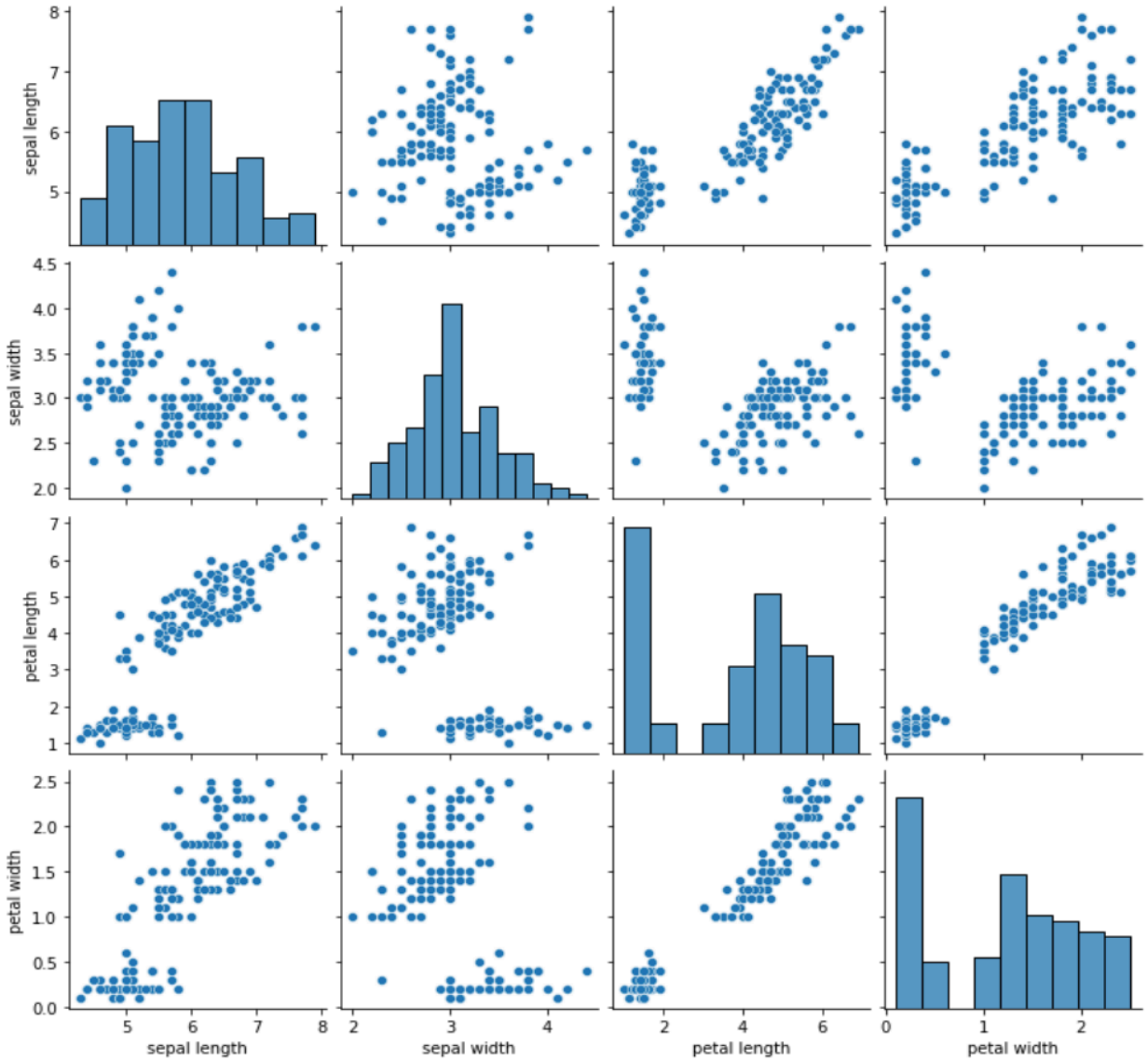
3) Dataset has 4 features and task is to predict 1 target variable. Models like Logistic Regression, Decision Trees, Random Forest Classifier and One-vs-Rest model are chosen for experimentation. The data pipeline is kept simple considering the complexity of the dataset and mix-max scalar was used for Logistic Regression & One-vs-Rest model.

Results obtained using these models are reported in the following table:

	Logistic Regression	Decision Trees	Random Forest	One-vs-Rest
Train Accuracy	85.3%	99.1%	99.13%	100%
Test Accuracy	88.2%	97.05%	97.1%	97.06%

The dataset is split into 80:20 for validating the models properly. The hyperparameters for the model can be inferred from the code directly. Code for question-3 lies in **q3.py** and logging from dataproc cluster lies in **dataproc_job.txt**

The pair plot for the the dataset can be found below:



Screen shot of the dataproc job can be found below: