1. **Given a table booking and a table shipment\_log, knowing that the shipment updates may occur and are logged into the shipment\_log table, write a SQL query that fetch all the bookings with their price, pickup\_location, delivery location, and the price variance.**

*Table descriptions:*

*booking:*

*booking\_id: Integer (primary key)*

*price: numeric*

*shipment\_log:*

*log\_id: Integer (primary key)*

*booking\_id: Integer (foreign key)*

*pickup\_location: String*

*delivery\_location: String*

*timestamp: Timestamp*

*price: numeric*

**Query:**

*Select*

*booking.booking\_id,*

*booking.price,*

*(shipment\_log.pickup\_city||" " ||shipment\_log.pickup\_state) as pickup\_location,*

*(shipment\_log.delivery\_city|| " " ||shipment\_log.delivery\_state) as delivery\_location,*

*(shipment\_log.price-booking.price) as variance*

*From*

*shipment\_log*

*Inner Join*

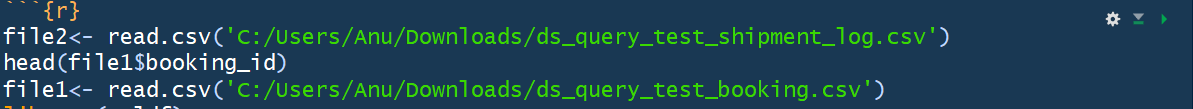
*booking*

*On*

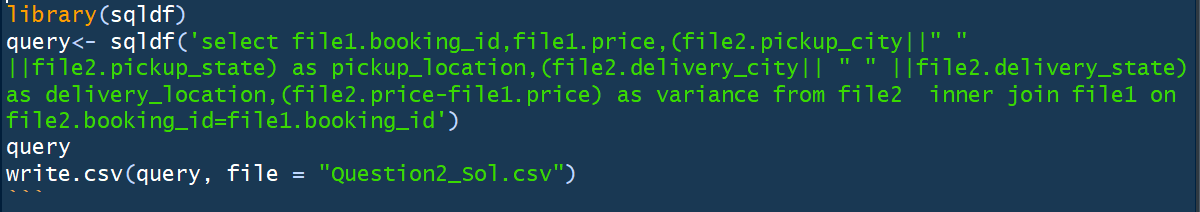
*shipment\_log.booking\_id=booking.book;*

1. **Based on the previous exercise, please download this booking and this shipment\_log csv files (using Google Spreadsheets, Excel, Numbers, or any other similar tool you are comfortable with), and create a new csv containing all the bookings with their price, pickup\_location, delivery location, and the price variance. This is basically the same exercise as the previous, but now the data is fetched inside a spreadsheet and returned as a csv.**

**R-Code**



*Reading and loading the CSV’s*



*Question2\_Sol.csv contains the output of the query in csv format.*

*Attached is the output in csv file.*



1. **For each of the following scenarios, first state whether you should model it as supervised learning or unsupervised learning, then state which model(s) you would choose and explain why. Possible models that you could choose from are linear regression, logistic regression, KNN, K‑Means Clustering, Hierarchical Clustering, etc. Argue the pros and cons of your selection.**

**Solution**

*Attached is the analysis on the situations provided:*



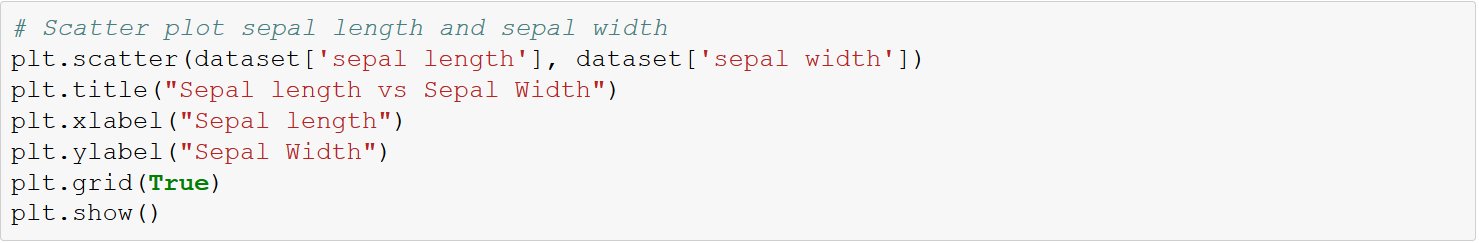
*Link to analysis*: <https://github.com/anupriya1519/Iris-Dataset/blob/master/Question3_Analysis.xlsx>

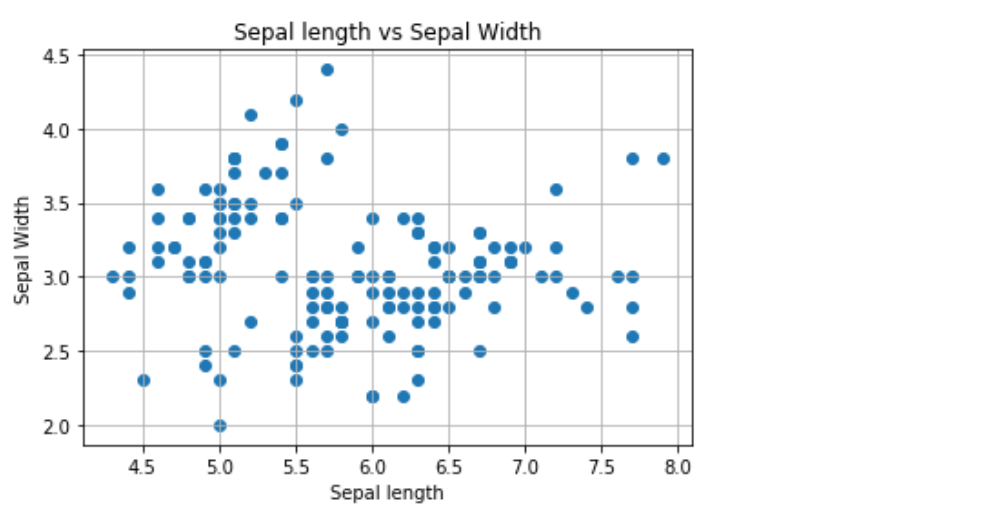
1. **You are given 150 samples of 3 different type of flowers (flowers.csv ). For each sample, there are 4 numeric attributes: sepal length, sepal width, petal length, petal width. You are asked to perform clustering algorithm to best assign the samples into 3 clusters. You can choose to do it**

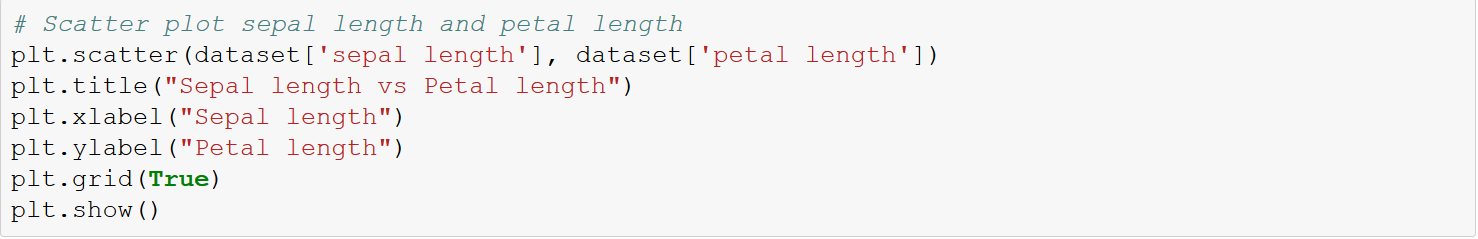
**in Python or R and you can use any build‑in functions or code it by yourself. If you choose iterative algorithms such as KMeans, please run at least 5 iterations. Please submit your code with the following items:**

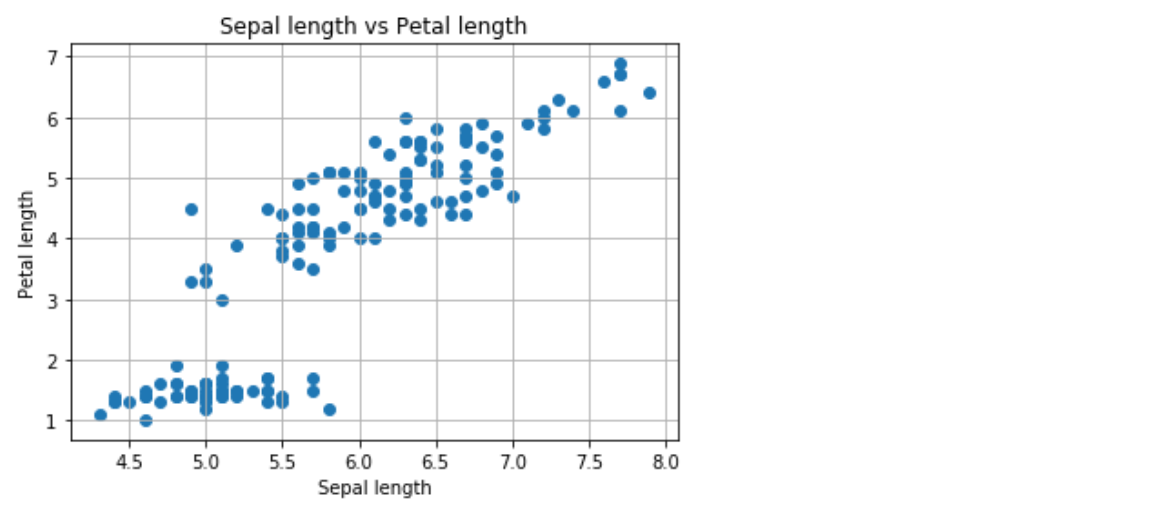
1. **3 scatter plots (of any two columns) of the original data 1**

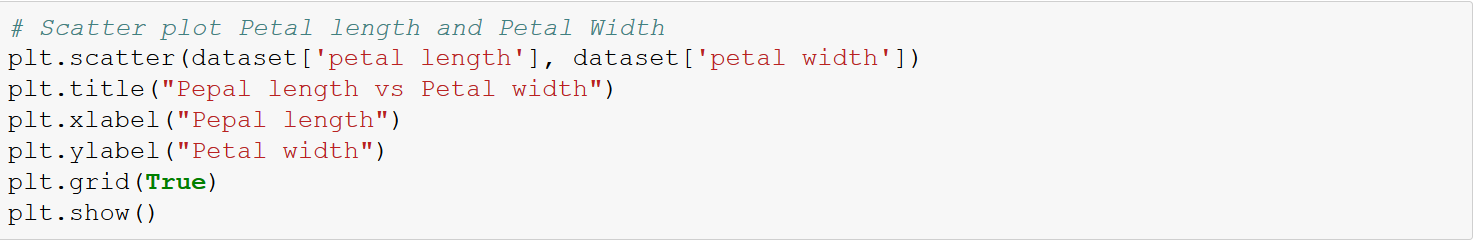
**Python Code and Output**

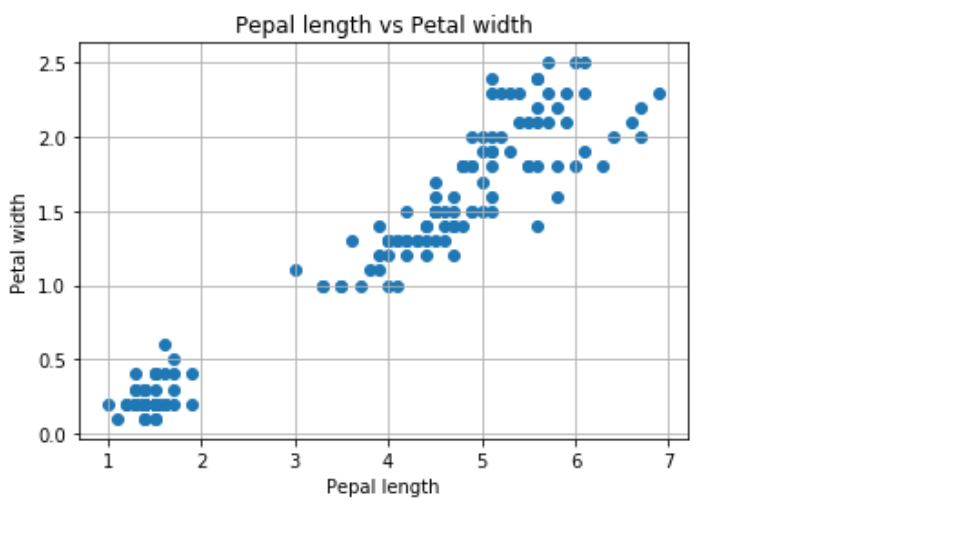






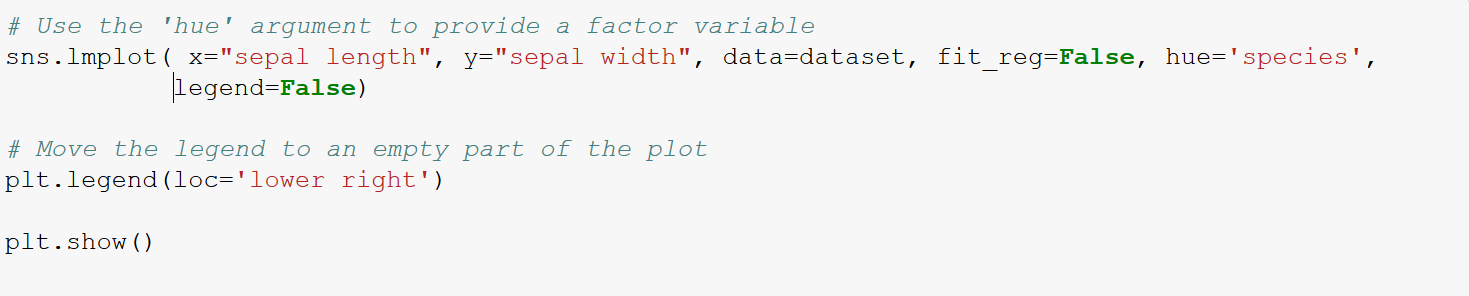


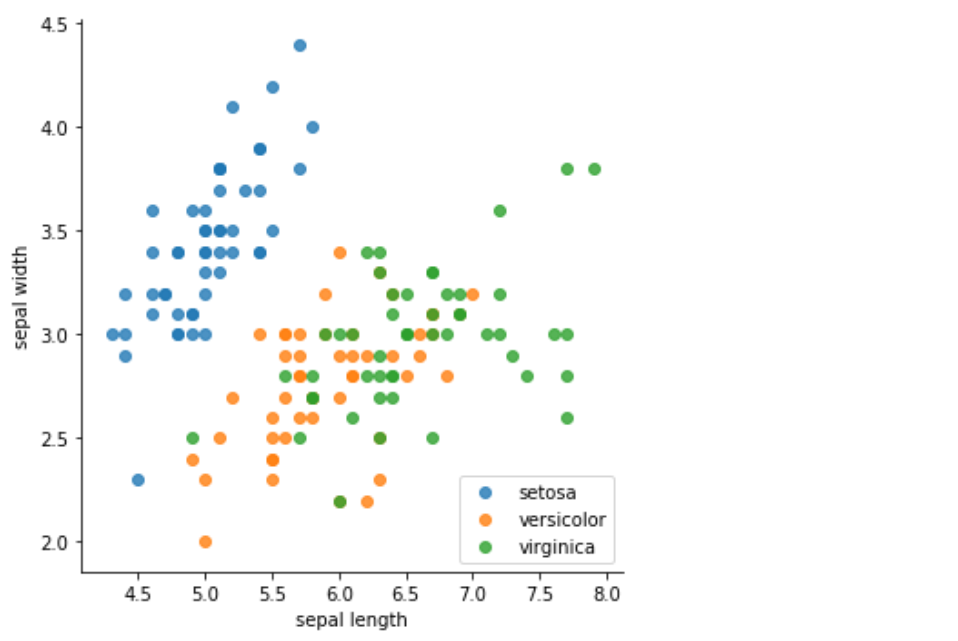


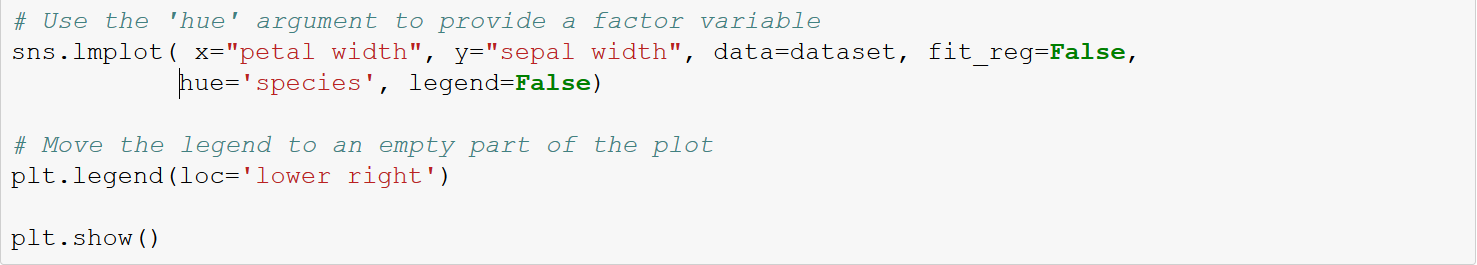


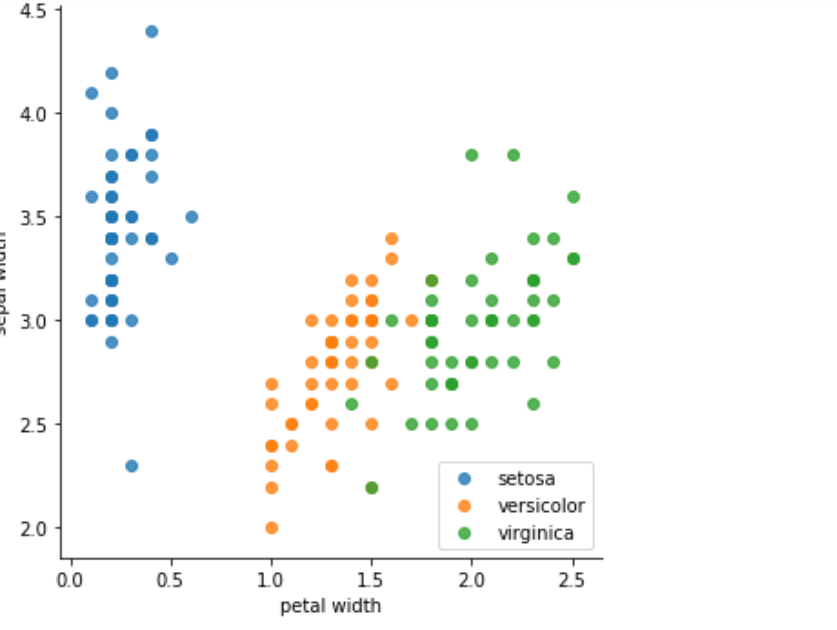
1. **3 scatter plots (of any two columns) of the data color coded with cluster membership**

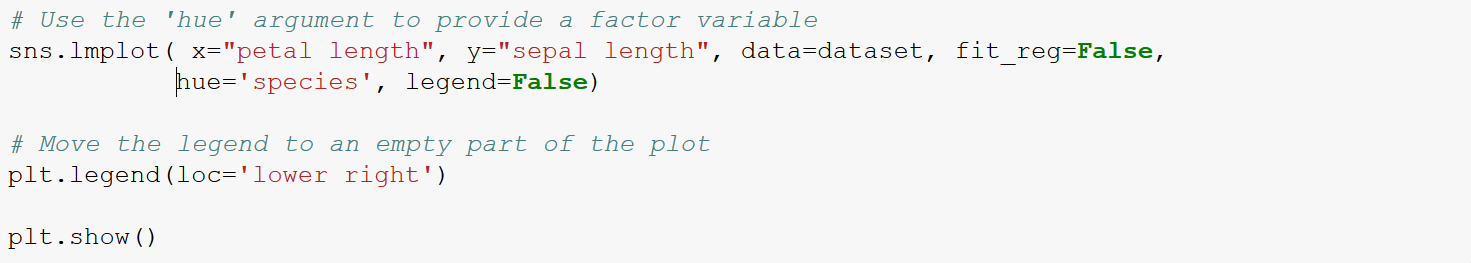
**Python Code and Output**

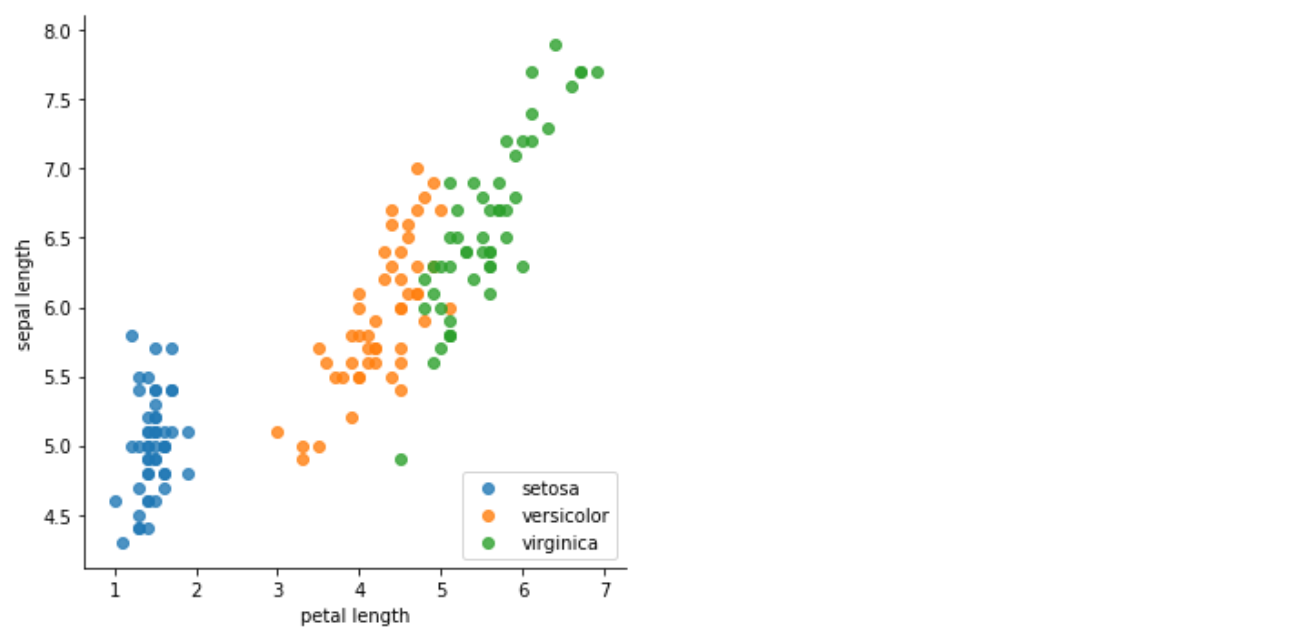












1. **It is known that sample 1‑50 belong to flower A, 51‑100 belong to flower B and 101‑150 belong to flower C. Compare your answer with the ground truth and explain why the algorithm worked/failed.**

**Solution**

*To solve this problem, I followed two approaches:*

**First**: *Since* *the number of clusters and labels were know, implemented classification using Naïve Bayes and Linear Discriminant Analysis. LDA outperforms with an accuracy of 97%.*

**Second**: *Implemented K-means clustering on the dataset since dataset is relatively small with few outliers. Running time complexity is linear i.e. O(n). To find the optimum number of ‘k’ used elbow method iteratively.*

*I have attached the entire code as PDF as another attachment in the mail.*

*Following is the link to the code*: <https://github.com/anupriya1519/Iris-Dataset/blob/master/Iris%20dataset.ipynb>