CSE CSE3020

Data visualisation

Lab

Lab experiment 10

TOPIC: Tableau and R integration

Name: Makesh Srinivasan
Registration number: 19BCE1717
Slot: F1 + TF1
Date: 09-April-2022-Saturday
Faculty: Prof. Parvathi

NOTE: I have decided to submit the lab report for lab 10 as a document rather than a dashboard or a PDF of the dashboard from Tableau software. This is because I wanted to show the steps I took to perform both the experiments (linear regression and K means clustering) along with explanations. This would not be possible to perform neatly on a dashboard

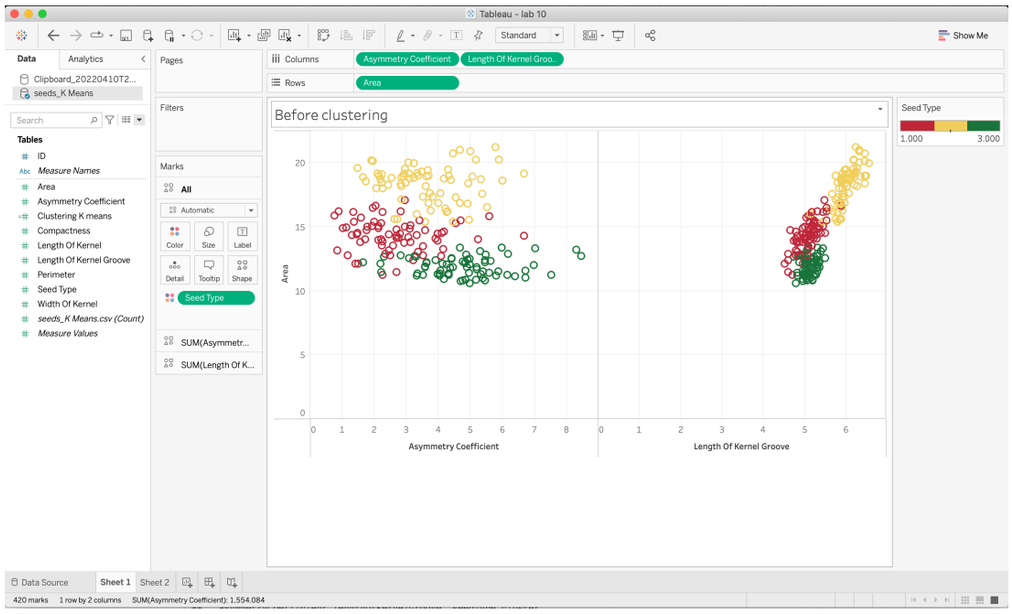
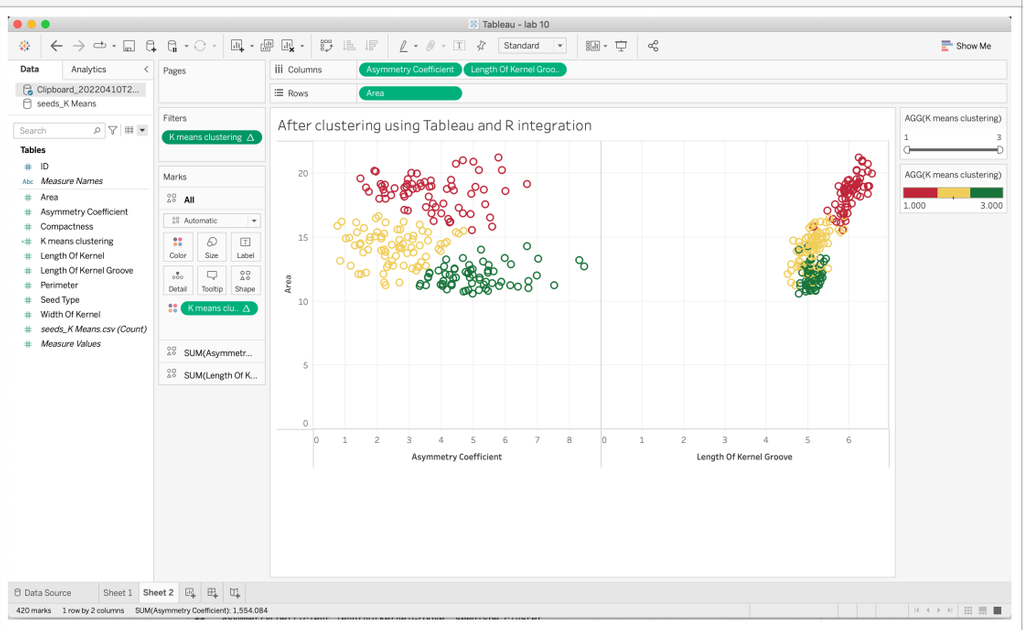
AIM:

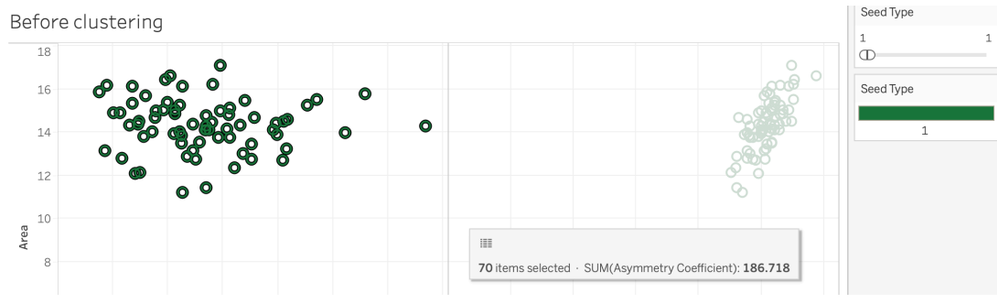
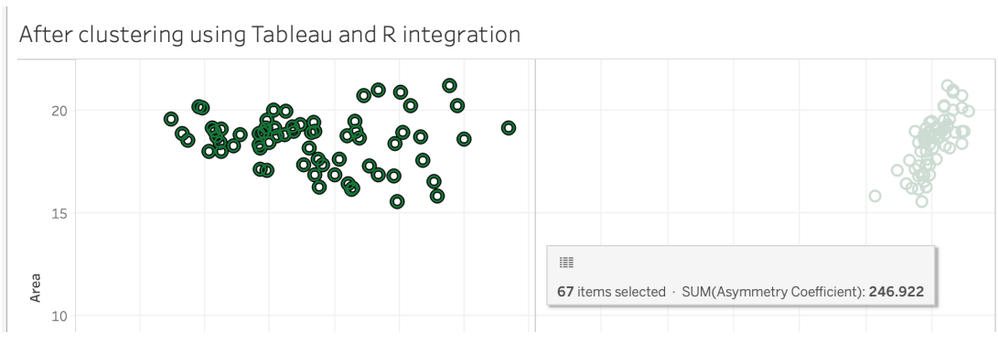
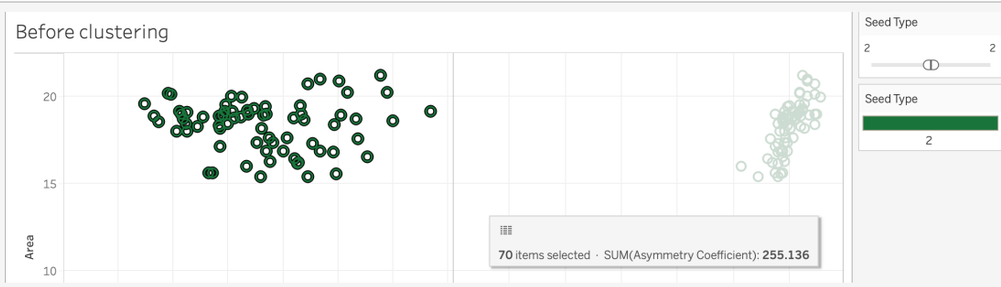
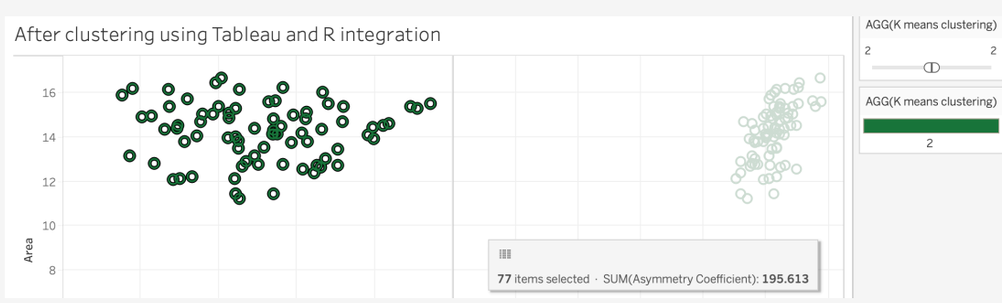
1) K means clustering using Tableau and R integration
2) Linear regression using Tableau and R integration

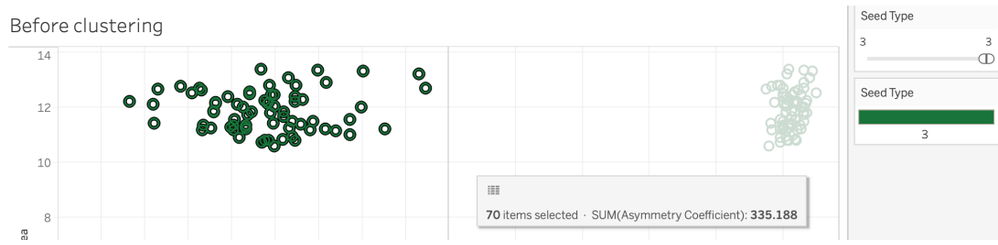
Dataset used: From the link provided on Moodle, I chose to use seeds dataset for this experiment

Experiment 1: K-means Clustering

Procedure:

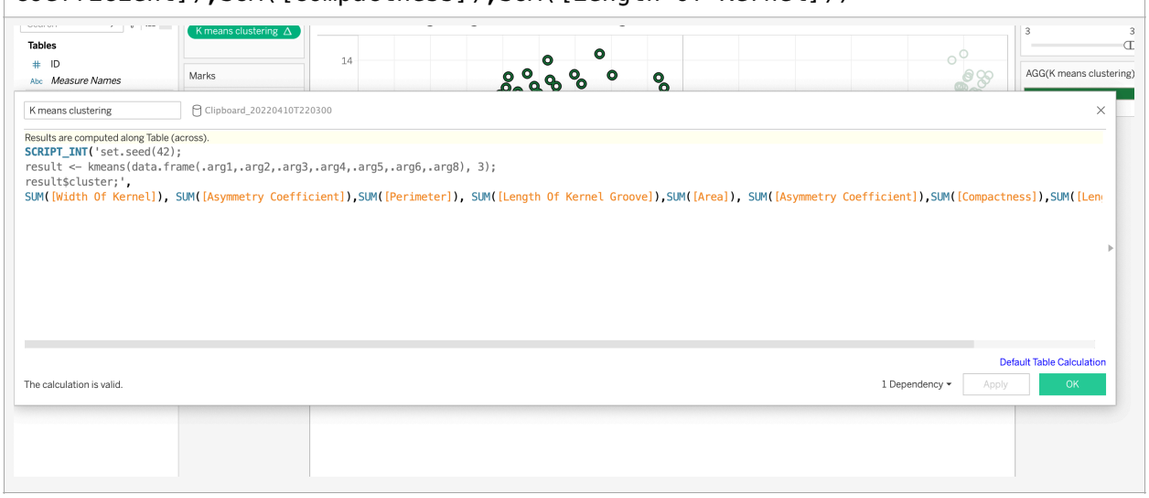
Steps	Description
1	
	Before I clustered using R code, i wanted to see the types of seeds. I knew there are 3 in total from the dataset description. Hence I chose clustering k = 3 in the code. But before I performed the clustering I wanted to observe the spread of points in asymmetry coefficient vs area plot and length of kernel grove vs area plot as they showed the most distinction between the types of seeds. As we see above, the spread is in different colours - green, red and gold - across the plane. Now, let us see the spread after clustering
2	
	As we can observe there is quite a difference between the ground truth values of the seed types given by the column “seed type” and the clusters I formed. Note, the colours encoded for the clusters or the seed types are not the same and cannot be controlled. They happen randomly in the backend (R code) and are labelled 1, 2 and 3 for each cluster. There are multiple misclassifications as well. They are explained below

Steps	Description
	  The number of type = 1 seed is 70 but the clustering shows 67. Hence 3 are misclassified and missing here
	  The number of type = 2 seed is 70 but the clustering shows 77. Hence 7 misclassifications

Steps	Description
	  The number of type = 3 seed is 70 but the clustering shows 66. Hence 4 are misclassified and also missing here.
	Hence, we see the clustering algorithm is fairly accurate but there is still room for improvement as there are some misclassifications too.

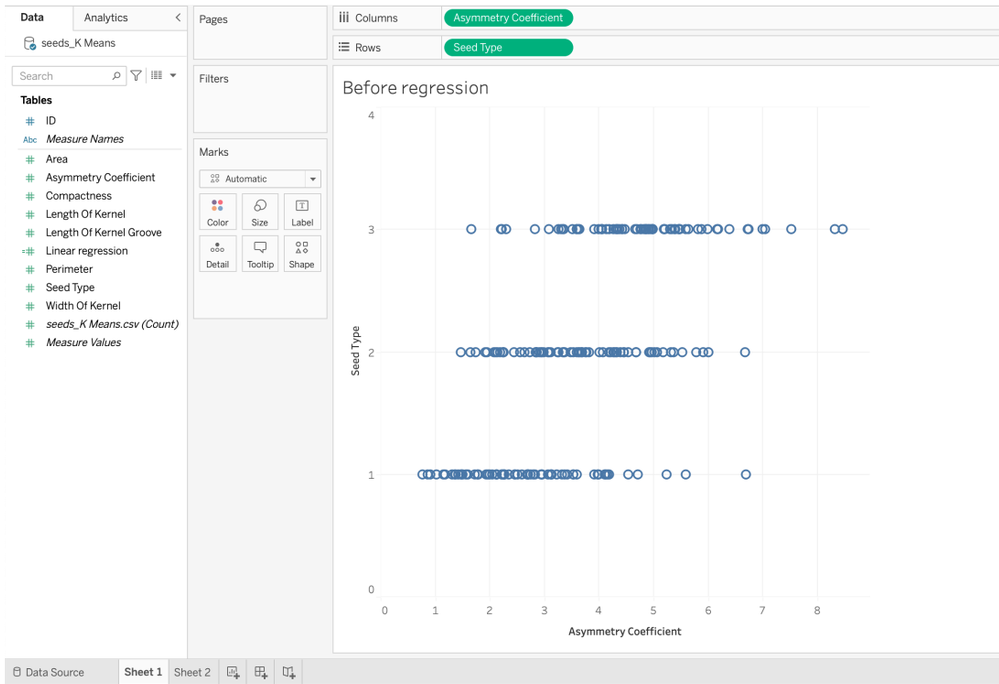
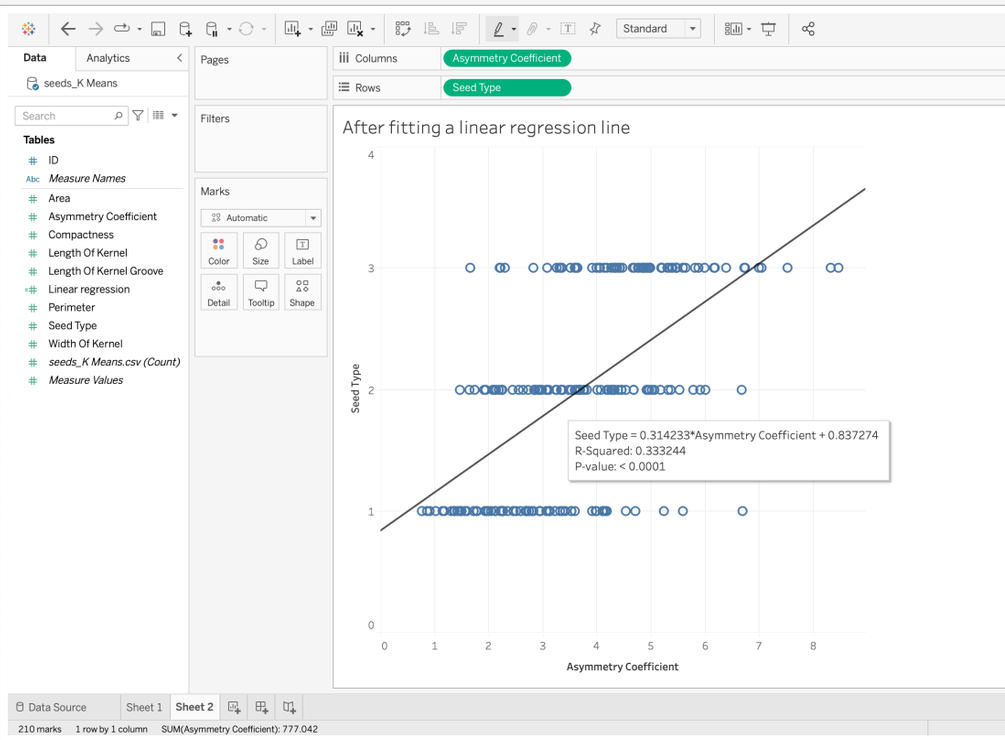
Code snippet

```
SCRIPT_INT('set.seed(42);
result <- kmeans(data.frame(.arg1,.arg2,.arg3,.arg4,.arg5,.arg6,.arg8), 3);
result$cluster;',
SUM([Width Of Kernel]), SUM([Asymmetry Coefficient]),SUM([Perimeter]),
SUM([Length Of Kernel Groove]),SUM([Area]), SUM([Asymmetry
Coefficient]),SUM([Compactness]),SUM([Length Of Kernel]))
```



Experiment 2: Linear regression

Procedure:

Steps	Description
1	
	Before we perform linear regression let us see the dependent variable of asymmetry coefficient against the label attribute - seed type. We see that it can perform fairly well using linear regression. Now let us implement regression here.
2	
	After plotting the regression line, we see the fit on the plane above. The model is also provided in the screenshot with the weight is also given above. The r squared value is found to be 0.33 which is not as accurate as only 33% of the data is described by the model above.

APPENDIX

```
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

Rserv started in daemon mode.
> library(cluster)
> Rserve()
```

Proof of Rserve running in the background