1. ***Optional Question***
2. **Schools in Portland Public School District are affected by churn. In this case, churn is defined as students switching schools midyear. There are reasons why this occurs. Investigate this issue.**
3. **Describe how you would go about developing a churn prediction model and discuss how you would derive input variables.**
4. **Provide a preliminary model (conceptual) in the form**

**Churn\_Value = M(x1, x2, … Xn)**

**Where M is a model. Here you want to suggest the variables for the model.**

1. **Create your own sample data (fictitious) to determine the churn Value. Keep it small. You only want to provide an example. Provide a sample of the data.**

**Solution:**

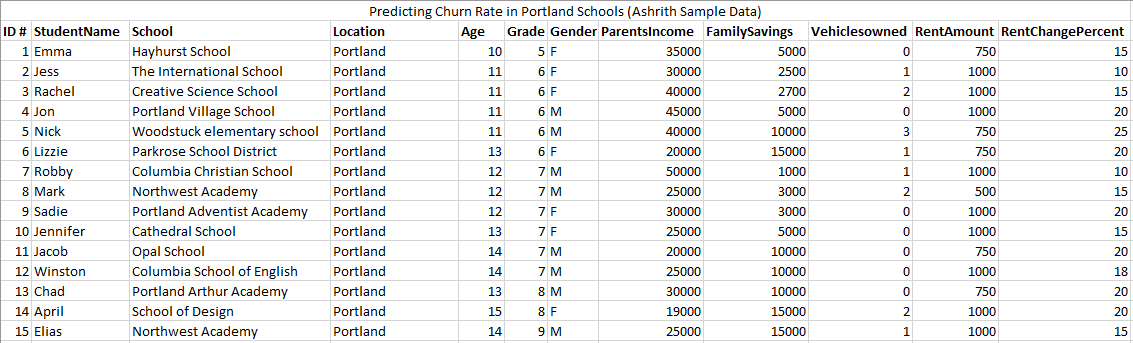
1. The links provided in the question are articles that talk about the housing crisis in Portland. It mentions that Schools have no control over rents to regulate this crisis and are unable to help people from getting evicted which is resulting in students switching schools. Understanding the housing data and the impact it would have on students would enable us to carry a churn prediction model. In order to do that, we must first analyze the problem correctly. The following steps help in the analysis:

* *Develop an understanding of the purpose of the data mining project:* In this step we use classification modeling to predict possibility of churn and predict the dependent variable after deriving all variables necessary. We could also use linear regression model to predict overall number of churns.
* *Obtain the dataset to be used for analysis:* The data obtained must have enough data for analysis of churn prediction. Multiple Datasets should be combined into one dataset and sampled for the purpose of analysis. Sampling enables us to get balanced dependent variables and improve accuracy of the model.
* *Explore, clean and Preprocess the data:* data should be checked for missing information, relevancy of data, reasonable range and outliers. Consistency in names of all fields should be checked for if multiple datasets are merged into one.
* *Reduce data dimension (if necessary):* In this step, we should make sensible decisions such as eliminating unnecessary variables, transforming variables and creating new variables for the purpose of the project.
* *Determine the data mining task:* We must choose which methods are applicable for this project(Clustering, Classification, prediction, etc.)
* *Partitioning the data:* If the project uses classification or prediction techniques, the dataset must be partitioned into 3 parts for training, validating and testing.
* *Choose data mining techniques to be used:* ‘Linear Regression’ can be used to predict the number of churns happening every year and ‘Classification’ for predicting dependent variable.
* *Use algorithms to perform the task:* In this step we would perform multiple variants of the same algorithms or multiple variants of different algorithms to understand how well our models work. For Example, Neural Network Regression, Linear Regression, Boosted Decision Tree Regression, etc.
* *Interpret the results of the model:* for this, we will interpret results of all models and make a choice for which model to deploy in this project. Testing gives and idea how well it will perform.
* *Deploy the Model:* The final model will be used in the project and is used to run it on real records for making decisions or taking actions.

1. *Suggestions of Variables for the Model:*

Churn\_Value = M(ID number, StudentName, School, Location, Age, Grade, Gender, ParentsIncome, FamilySavings, VehiclesOwned, RentAmount, RentChangePercent)

1. The Following Table is a *fictitious* dataset that I would use to determine the Churn\_Value



1. **We have used several classification costs besides accuracy. Below is the r-code for deriving a specific measure – F1-measure (also known as F score for the given table.**

**df #data frame with contingency table**

**Prediction One Zero**

**1 One 37 43**

**2 Zero 19 131**

*# Precision: tp/(tp+fp):*

*df[1,1]/sum(df[1,1:2])*

*[1] 0.4625*

*# Recall: tp/(tp + fn):*

*df[1,1]/sum(df[1:2,1])*

*[1] 0.6607143*

*# F-Score: 2 \* precision \* recall /(precision + recall):*

*2 \* 0.4625 \* 0.6607143 / (0.4625 + 0.6607143)*

*[1] 0.5441177*

**A. Review the results for two models below. Calculate the accuracy (% correct) precision, recall, and F-Score and enter the results here.**

***Model 1***

*PREDICTED.E PREDICTED.A  
TRUE.E         512       488  
TRUE.A          11         899*

***Model 2***

*PREDICTED.E PREDICTED.A  
TRUE.E         495         505  
TRUE.A        1203       98797*

1. **Discuss the differences in the scores between them. Which model is better? Why? Which measure(s) are more useful and why?**

**Solution:**

1. **Model 1 :**

**Accuracy**=[ (True.E\*Predicted.E)+(True.A\*Predicted.A)/(Total Observations)]\*100%

= [(512+899) **/** (512+899+11+488)]\* 100 %

= 0.7387 \* 100 % = **73.87%**

**Inaccuracy** =[(True.E\*Predicted.A)+(True.A\*Predicted.E)/(TotalObservations)]\*100%

= [(488+11) / (512+899+11+488)]\* 100 %

= 0.2612 \* 100 % = **26.12%**

**Recall on A** = [(True.A\*Predicted.A) / total True.A’s] \* 100%

= [899/(899+11)] \* 100 %

= 0.9870 \* 100 = **98.7%**

**Recall on E** = [(True.E\*Predicted.E) / total True.E’s] \* 100%

= [512/(512+488)] \* 100 %

= 0.5120 \* 100 = **51.2%**

**Precision on A**= [(True.A\*Predicted.A)\*Predicted A’s \* 100 %

= 899/(899+488) \* 100%

= 0.6481 \* 100% = **64.81%**

**Precision on E**= [(True.E\*Predicted.E)\*Predicted E’s \* 100 %

= 512/(512+11) \* 100%

= 0.9789 \* 100% = **97.89%**

**Fi Score** = Harmonic Mean of Precision and Recall

= 2PR/(P+R) = **0.782**

**Model 2:**

**Accuracy**=[ (True.E\*Predicted.E)+(True.A\*Predicted.A)/(Total Observations)]\*100%

= [(495+98797) **/** (495+98797+1203+505)]\* 100 %

= 0.9831 \* 100 % = **98.31%**

**Inaccuracy** =[(True.E\*Predicted.A)+(True.A\*Predicted.E)/(TotalObservations)]\*100%

= [(1203+505) / (495+98797+1203+505)]\* 100 %

= 0.0169 \* 100 % = **1.69%**

**Recall on A** = [(True.A\*Predicted.A) / total True.A’s] \* 100%

= [98797/ (98797+1203)] \* 100 %

= 0.9879 \* 100 = **98.79%**

**Recall on E** = [(True.E\*Predicted.E) / total True.E’s] \* 100%

= [495/(495+505)] \* 100 %

= 0.495 \* 100 = **49.5%**

**Precision on A**= [(True.A\*Predicted.A)\*Predicted A’s \* 100 %

= 98797/ (98797+505) \* 100%

= 0.9949 \* 100% = **99.49%**

**Precision on E**= [(True.E\*Predicted.E)\*Predicted E’s \* 100 %

= 495/(495+1203) \* 100%

= 0.2915 \* 100% = **29.15%**

**Fi Score** = Harmonic Mean of Precision and Recall

= 2PR/(P+R) = **0.991**

1. Model 2 is a better Model. My reason for choosing model 2 is because it scores higher than Model 1 in terms of Accuracy, Recall, Precision and Fi Score. It is also a better model because Inaccuracy percentage is lesser compared to Model 1.

Predictions on E was lower in Model 2 than in Model 1.

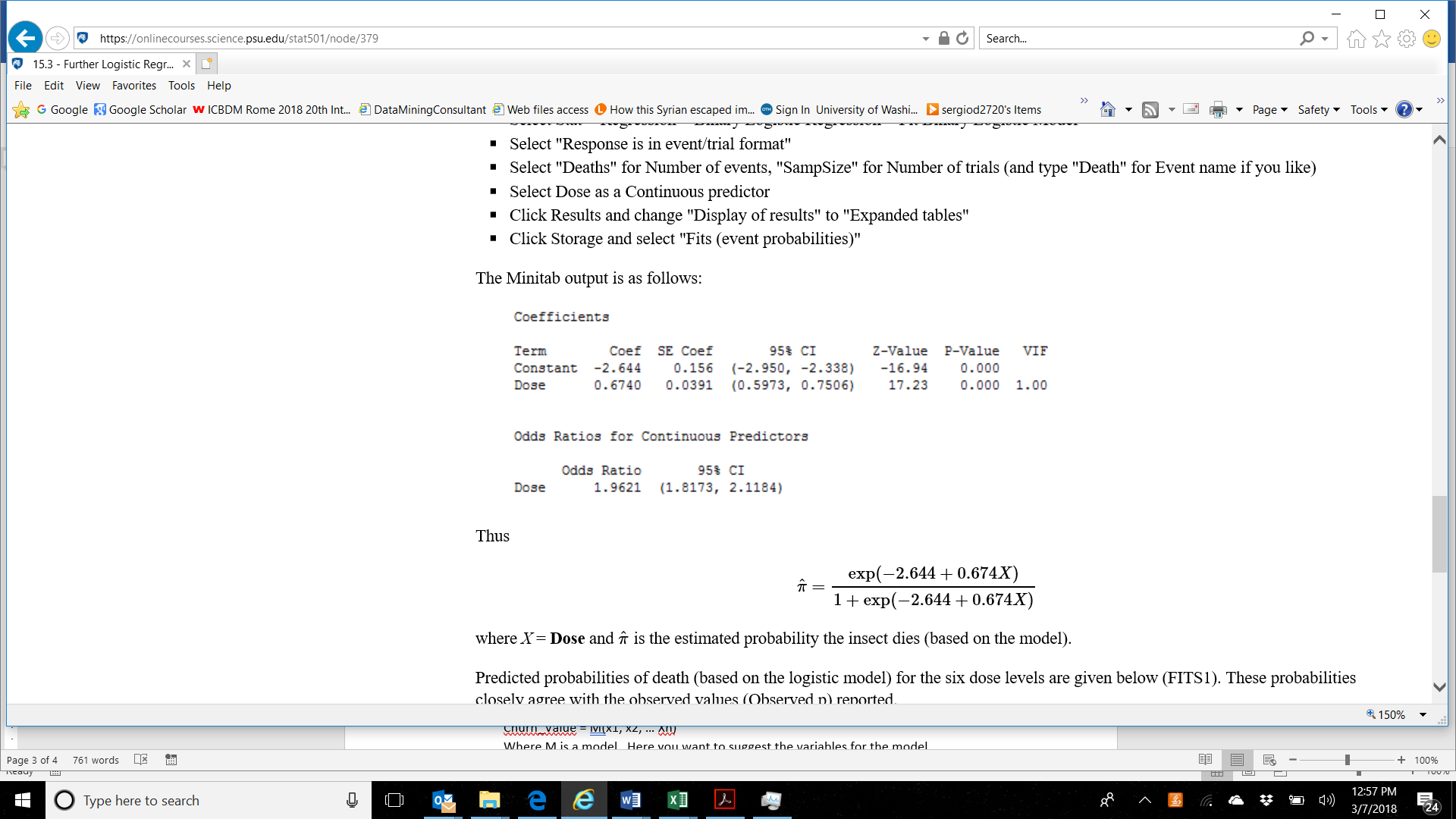
1. **An experiment is performed to test the effect of a toxic substance on insects. The data is from the textbook, Applied Linear Statistical Models by Kutner, Nachtsheim, Neter, & Li.**

**At each of six dose levels, 250 insects are exposed to the substance and the number of insects that die is counted.**

**Below is a summary table of the data:**

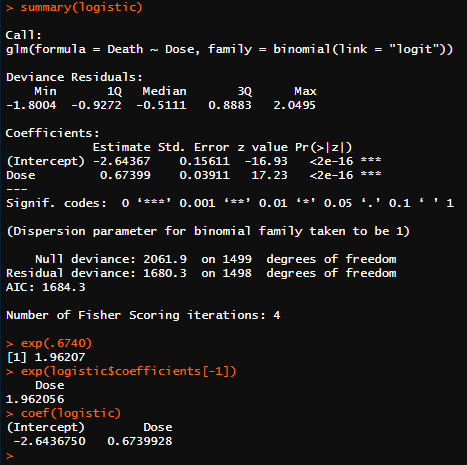
|  |  |  |
| --- | --- | --- |
| **Dose** | **SampSize** | **Deaths** |
| **1** | **250** | **28** |
| **2** | **250** | **53** |
| **3** | **250** | **93** |
| **4** | **250** | **126** |
| **5** | **250** | **172** |
| **6** | **250** | **197** |

**Below is the output from the program**

****

* 1. **Interpret the output.**
  2. **Use the data set called LRTest.xls associated with this test to conduct logistic regression. Confirm that you get the same results as the output shown. Show your output here. Note: you can use any software.**
  3. **Calculate the observed probabilities as the number of observed deaths out of 250 for each dose level.**
  4. **Calculate the predicted probabilities**

**Solution:**

1. **Output Interpretation:** This Logarithmic model predicts that the odds of death increase by 0.67 with every increase in dose. The odds of Dying increases by 96% as dose increases by a factor of 1
2. ****
3. **Observed Probabilities as the number of deaths out of 250 for each dose level:**

**Dose 1 =** 28/250 = 0.112

**Dose 2 =** 53/250 = 0.212

**Dose 3 =** 93/250 = 0.372

**Dose 4 =** 126/250 = 0.504

**Dose 5 =** 172/250 = 0.688

**Dose 6 =** 197/250 = 0.788

1. **Predicted Probabilities:**

**Dose 1 =** 0.1224

**Dose 2 =** 0.2148

**Dose 3 =** 0.3493

**Dose 4 =** 0.5130

**Dose 5 =** 0.6739

**Dose 6 =** 0.8022

1. **Describe an application of association rules method. Do not base your discussion on purchases such as grocery store transactions. You want to apply it in another area – marketing, healthcare, finance, law enforcement, fraud, etc. You want to include in the discussion of the benefit, the specific aspects of the application, and the potential itemset content.**

**Solution:** *Association Rules Mining* is the method of analytics where it aims to discover which groups of products tend to be purchased together. In this method, we aim to observe frequently occurring patterns or associations from datasets found in repositories like relational databases or transactional databases. Due to this analysis, the two products are recommended to be bought together online or displayed together or even offered a discounted rate for the latter one. It consists of 2 stages. To name, Rule Generation and choosing a subset through assessment of rule strength. *Association Rules Method* is an unsupervised learning method. Affinity Analysis and Market Basket Analysis are some of the other names given to this method.

*Examples:* **Census Data** obtained by governments are used in a democratic society to plan efficient public services and help public businesses as well. Another example, **Amazon** recommends users another item while checking out with one item. Say, if a customer orders a phone, a recommendation to buy Screen protector and phone case is shown to the buyer.

*Parameters to observe relationships:* **Support** and **Confidence** are 2 parameters that are used to thoroughly analyze data and find frequent patterns. Support indicates the frequency of a relationship appearing in the database and Confidence determines the number of cases where the said relationships are true.

*Advantages:*

Association Rules Method is beneficial in health care industries in terms of positive and negative association rule mining. Positive and Negative associations can help a medical practitioner confirm if the patient has a presence or an absence of certain diseases. Negative Association helps maximize the possibility of early disease identification by developing a decision support system and minimize errors while diagnosing.

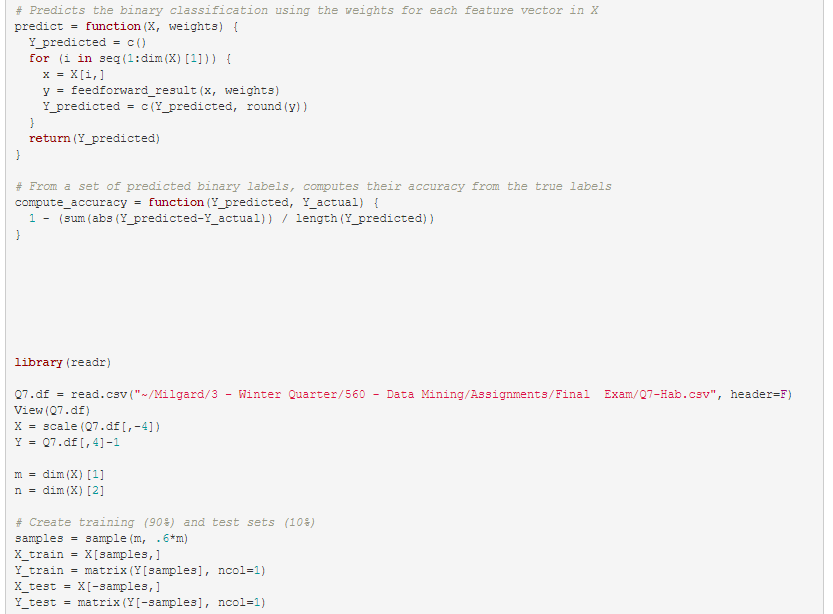
1. **OPTIONAL QUESTION**
2. **Conduct 8 experiments using any data set and any neural network implementation (does not have to be R ). You will vary the learning rate (2 different ones) as well as the number of hidden layers (2 different ones) as well nodes per layers (2 different ones).**
3. **Compare the results and discuss the outcomes**

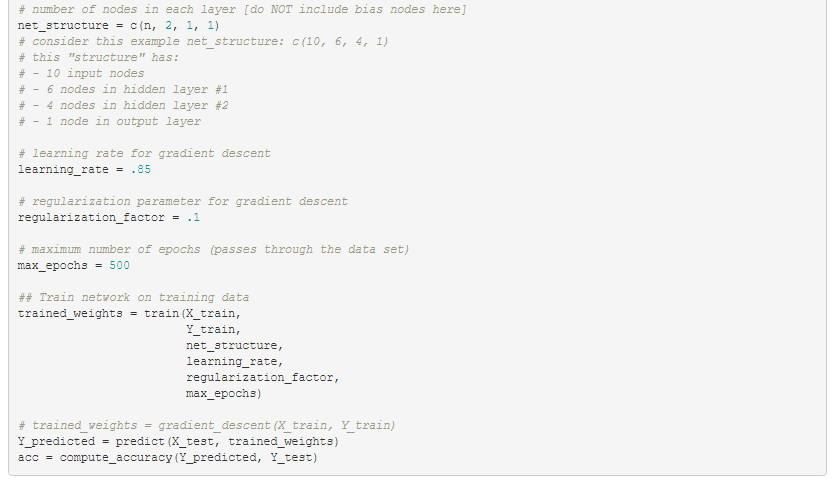
**Solutions:**

1. **Screenshots of RMD Report for experiment**

****

****

****

****

1. **OUTPUTS:**
2. **60-40**

**regularization\_factor = .1**

**# - 3 input nodes**

**# - 3 nodes in hidden layer**

**# - 1 node in output layer**

**Learning rate = .3**

**75.3 77.23 75.6**

**Learning rate = 0.85**

**77.23 78.4 78.8 70.6**

**Learning rate = 0.01**

**70.07 76.17 69.1**

**Learning rate = .95**

**76.4 78.1 67.4**

**=======================**

**60-40**

**Regularization\_factor = .1**

**# - 3 input nodes**

**# - 2 nodes in hidden layer #1**

**# - 1 node in hidden layer #2**

**# - 1 node in output layer**

**Learning rate = .3**

**72.3 75.6 73.1**

**Learning rate = 0.85**

**77.23 74.8 69.1**

**Learning rate = 0.01**

**73.1 74.7 69.9**

**Learning rate = .95**

**71.5 72.3 69.1**

The tunable parameters in this neural network are Learning Rate, Nodes per hidden layer, number of output classes and stopping criterion. More data leads to more stable training, more nodes lead to a better fitting model. Increased regularization led to increased training loss and a slow-moving training curve was due to smaller learning rate.

1. **Customer segmentation identifies customers into distinct groups.  You want to examine a set of customers and identify those that behave similarly to tailor marketing.  You want to group them based on purchases of four products.**

**Here is the data:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Customer#** | **ProductA** | **ProductB** | **ProductC** | **ProductD** |
| **1** | **0** | **0** | **1** | **1** |
| **2** | **0** | **1** | **0** | **1** |
| **3** | **1** | **1** | **0** | **0** |
| **4** | **1** | **1** | **0** | **1** |
| **5** | **1** | **0** | **0** | **0** |
| **6** | **0** | **0** | **1** | **0** |
| **7** | **1** | **0** | **1** | **1** |
| **8** | **1** | **1** | **0** | **0** |
| **9** | **1** | **0** | **0** | **0** |
| **10** | **0** | **0** | **1** | **1** |
| **11** | **0** | **0** | **1** | **1** |
| **12** | **1** | **1** | **0** | **0** |
| **13** | **1** | **0** | **1** | **0** |
| **14** | **0** | **1** | **0** | **0** |
| **15** | **0** | **1** | **0** | **1** |
| **16** | **0** | **0** | **1** | **1** |
| **17** | **1** | **0** | **0** | **0** |
| **18** | **0** | **0** | **0** | **1** |
| **19** | **0** | **1** | **1** | **1** |
| **20** | **0** | **1** | **0** | **1** |

**Data Exploration: 20 customers.  4 products + customer ID.  Binary where 1 = purchased.  0 = no purchase.**

* **9 / 20 (45%) have purchased product A**
* **9 / 20 (45%) purchased product B**
* **8 / 20 (40%) purchased product C**
* **11/ 20 (55%) purchased product D**

**PROCEDURE**

**Applying K Means: Pick random K centroids and then iteratively assign all other points to one of the K clusters by looking for the smallest distance to the centroids.  In this case use K = 3 and determine the customers in the three clusters.**

* 1. **Start by calculating the distance of each customer from the three starting centroids.**

**Use  Euclidean Distance to measure how far apart the  customers are from the centroids.**

**Starting with Cluster # 1 and determine which cluster the customers belong to.**

**2, Update the centroids.**

* + 1. **Repeat steps 1 and 2 for three iterations**

1. **Group the individuals into three clusters using K-Means cluster method and the Euclidean distance measure. Show your work. Define the initial cluster centroids as:**

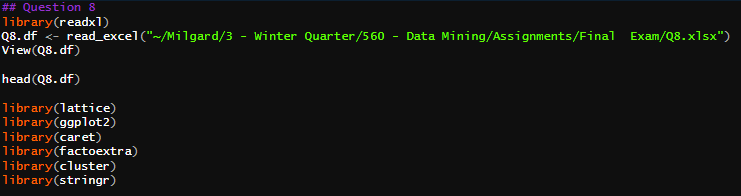
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Clusterr#** | **ProductA** | **ProductB** | **ProductC** | **ProductD** |
| **1** | **1** | **1** | **0** | **1** |
| **2** | **0** | **1** | **1** | **1** |
| **3** | **0** | **1** | **0** | **1** |

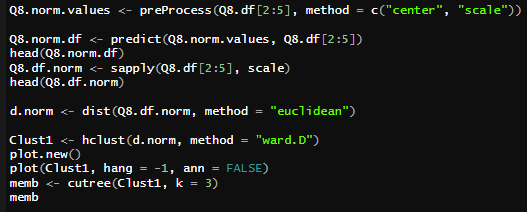
**Show each step in how records are added to a cluster. This will be based on the mean vector. Remember the mean vector is recalculated each time a new member is added to the cluster. At the end you will want to make sure that each record is assigned to the closest mean. Determine the clusters for the customers after the end of three iterations.**

**Show your work for the K-Means clustering steps using the three starting centroids.**

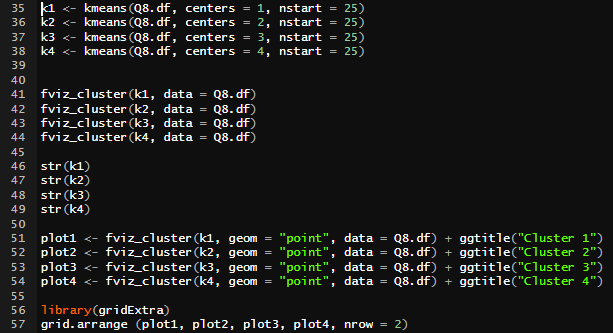
**Solution:**

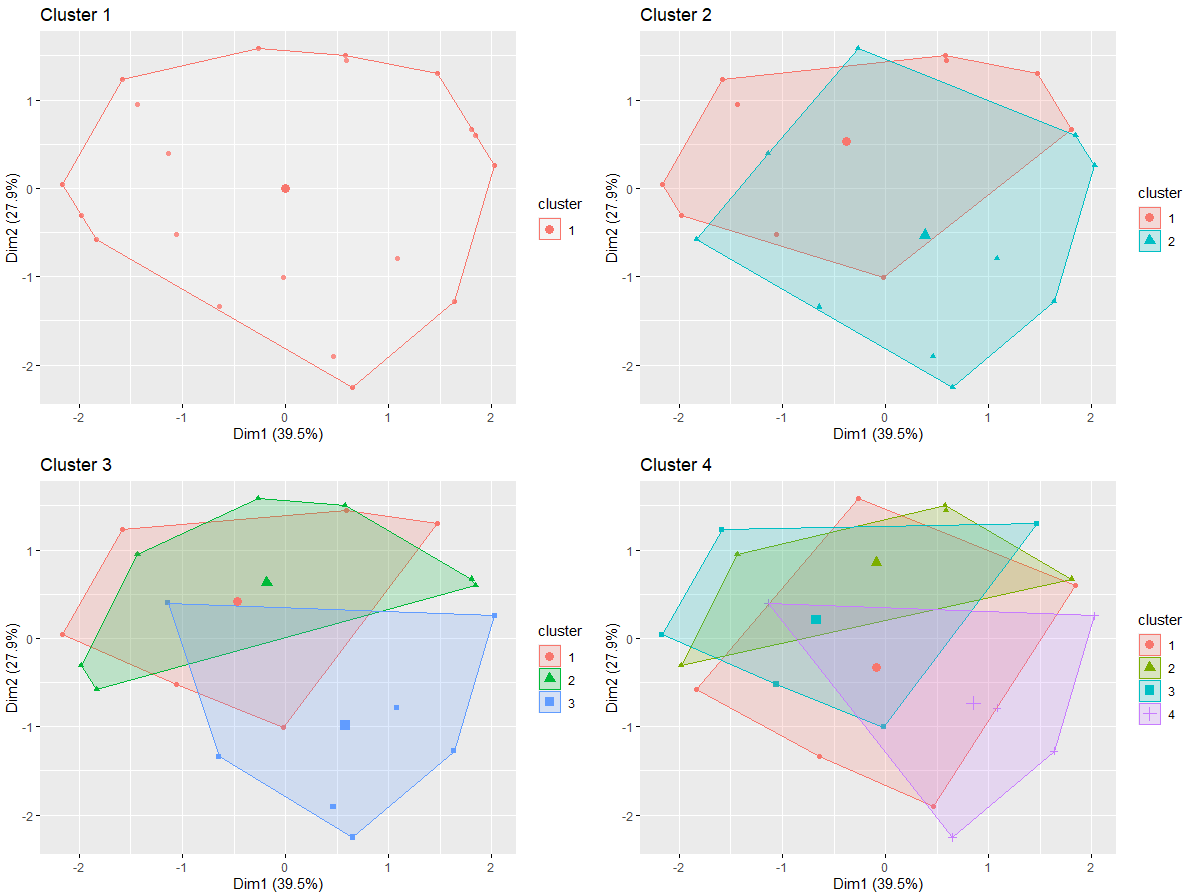
**Step 1:**

****

****

**Code used for Plotting: Library(factoextra) was installed to utilize its subsidiary packages “fviz\_cluster” to allow plotting of the clusters.**

****

****