# MODULE 4 UNIT 3

## Activity submission

Learning outcomes:

LO4: Analyse the results of a cluster analysis in a business context.

LO5: Reflect on the use of cluster analysis.

### Name: Matlotlo Magasa

#### 1. Instructions and guidelines (Read carefully)

##### Instructions

1. Insert your name and surname in the space provided above, as well as in the **file name.** Save the file as: **First name Surname M4U3 Activity Submission** – **e.g. Lilly Smith M4U3 Activity Submission.** **NB:** *Please ensure that you use the name that appears in your student profile on the Online Campus.*

2. Write all your answers in this document. There is an instruction that says, “Start writing here” under each question. Please type your answer there.

3. Submit your assignment in **Microsoft Word only**. No other file types will be accepted.

4. You will be required to include visualisations that you have made in Tableau in this activity. To do this, export them as an image file and paste them into this document. Tableau has a page that details how to export your Tableau view as an image file.

5. **Do not delete the plagiarism declaration** or the **assignment instructions and guidelines**. They must remain in your assignment when you submit.

**PLEASE NOTE:** **Plagiarism cases will be investigated in line with the Terms and Conditions for Students.**

**IMPORTANT NOTICE:** Please ensure that you have checked your course calendar for the due date for this assignment.

##### Guidelines

1. There are 7 pages and 4 questions in this assignment.

2. This activity submission relies heavily on work done in the IDE notebook in the previous unit. If you have not done so already, return to the IDE and complete this activity before attempting to answer the questions.

3. You will be required to download a data file, titled LOANS.sav, in order to interact with the data in Tableau. Return to the Module 4 downloads folder and save a copy of the data file before attempting to answer the questions.

4. Make sure that you have carefully read and fully understood the questions before answering them. Answer the questions fully, but concisely, and as directly as possible. Follow all specific instructions for individual questions (e.g. “list”, “in point form”).

5. Answer all questions in your own words. Do not copy any text from the notes, readings, or other sources. **The assignment must be your own work only.**

|  |
| --- |
| **Plagiarism declaration:** |
| **1. I know that plagiarism is wrong. Plagiarism is to use another’s work and pretend that it is one’s own.**  **2. This assignment is my own work.**  **3. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.**  **4. I acknowledge that copying someone else’s assignment (or part of it) is wrong and declare that my assignments are my own work.** |

#### 2. Mark allocation

Each question receives a mark allocation. However, you will only receive a final percentage mark and will not be given individual marks for each question. The mark allocation is there to show you the weighting and length of each question.

Question 1 5

Question 2 10

Question 3 10

Question 4 15

**TOTAL 40**

#### 3. Questions

The management team at Speedy Loans is interested in better understanding their customer base and have asked you, as the new data analyst, to provide some simple visualisations and analyses to communicate these insights. You conducted a cluster analysis to group customers based on similar characteristics, which was completed in the previous component. After reviewing the results, the management team asked you to answer specific questions related to the outputs you provided.

##### Question 1

The management team would like to separate their customer base according to whether –after applying for a loan – the bank approved or denied a customer’s request. However, the current data set categorises those who were denied for a loan as “0” and those who were approved for a loan as “1”, which is not intuitive to understand. As a result, the management team has asked you to use calculated fields to assign “Approved” and “Denied” to each category to allow the data to be visualised in Tableau. They have also asked you to plot this new variable against credit card debt and other debt.

To do this, create a calculated field using a logical “IF ELSE” command, stating that, if customers are categorised as “1”, they should be grouped into the “Approved” category. In contrast, if they are categorised as “0”, they should be grouped into the “Denied” category. Label this new variable “Approval status” and drag it into the columns field. Plot this variable, using bar charts, against “Cred\_debt” and “Other\_debt” simultaneously. Remember to change the measure for these variables from “SUM” to averages. Change the colour of one of the graphs to make the distinction between the two variables more pronounced and rename the worksheet “Loan approval status”.

The board has now asked you to provide a brief summary of what the results show and what possible insights might be extracted from the data.

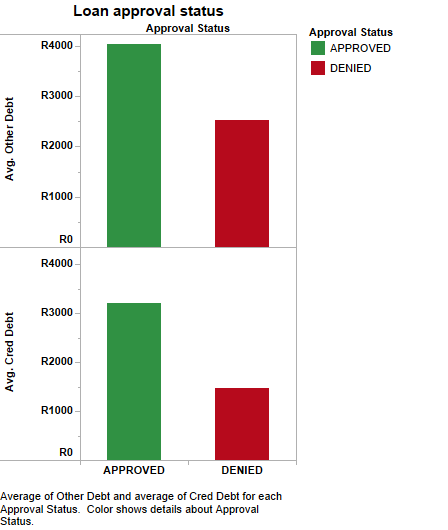
(Max. 150 words)

Start writing here:

From the graph we see that people are denied a loan when their average ‘other debt’ is at R2 500 and at R1 500 for credit debt. The average debt where one gets an approval of a loan is when their debt is R4 000 and R3 300 for ‘other debt’ and credit debt respectively.

This means that potential customers are currently being approved for A loan at a generally high debt. Customers are likely to get approved at an average debt of R3 300 If that debt is allocated to credit. If not, then based on his other debt, the customer will get approved at R4 000. If the customer does not have a history of substantial debt, he is likely to be denied, or other factors may first be considered.

Paste your Tableau visualisation here:



##### Question 2

The first method of cluster analysis you performed used a single-linkage clustering strategy with different distance metrics, including Euclidean, Manhattan, and correlation distance. The management team wants to know whether each method of clustering was effective in separating their customer base into well-defined clusters. In your answer, you should give special consideration to the following:

* Are any of the distance metrics more effective at separating the data observations than others?
* Do the observation allocations indicate that there is effective separation of the data?
* Finally, based on the previous answers, would you recommend single-linkage clustering for this data set? Give reasons for your recommendation.

(Max. 250 words)

Start writing here:

The single linkage method has not produced visually effective dendrograms, more especially for the Euclidean and Manhattan metrics, because the ‘tree branches’ are spatial squashed making cut-off points a little harder to establish. The correlation metric in this regard produced a visually better dendrogram.

For all metrics used with single linkage method, the observations have not been equally distributed in the different clusters whether you increase or decrease the number of clusters. This means that as a business analyst, you are not able to classify the company’s customer into sensible groups/ clusters.

I have also observed using a cluster specification of 3, that all the metrics in single linkage produce the same distribution results. Due to a great majority (about 98% for 3 clusters) of the observation being clustered in only one cluster, this is not an effective method f clustering for the Speedy Loans use case and I would not recommend this clustering method.

##### Question 3

In providing your answer to the previous question, you informed the management team that the clustering solution could be more effective when using complete-linkage, as opposed to single-linkage, clustering. With their interest piqued, the team has now asked you to explain the results, giving particular consideration to the following:

* How does complete-linkage clustering group the customers when using different distance metrics?
* Do the observation allocations indicate that there is better separation of the data compared to single-linkage clustering?
* Finally, based on these answers, would you consider complete linkage clustering an effective solution for the data set? Substantiate your answer.

(Max. 300 words)

Start writing here:

Complete linkage clustering is a clustering method that is similar to the single linkage in that it’s hierarchal, and sees each object being in its own cluster at the beginning of the analysis. However, merging happens when observations with the farthest distance from each other are clustered together until all observations belong to the same cluster.

The complete linkage method produces visually better dendrograms than the single linkage method. The allocations distribution has also improved from the single linkage allocations making it complete linkage the better clustering method compared to the single linkage method.

Decreasing the number of clusters below three provided an even poorer distribution in the cluster for complete linkage. Increasing the clusters to more than 4 seemed to make no improvement in distribution, it just added one or two observation in the new clusters that are increased. As a result, I investigated the complete linkage using a cluster of 4.

From the three-distance metrics, the Euclidean method produced the better allocation results:



Euclidean



Manhattan



Correlation

The Manhattan and correlation measures only distributes the majority of observation in the first 2 clusters leaving the other two with a very small number of observations, while the Euclidean distribution is majority in the first 3 clusters of the 4 clusters. Based on this observation, the Euclidian method is better of the three measures.

I don’t think that the final results of the complete linkage is effective because of the distribution/ allocations in the clusters. It cannot be that we have a potential range of 1% to 5% of observations in only one of the clusters while the other clusters carry a great majority of the observations. I would first explore other clustering methods farther e.g. centroid method or k-means method.

##### Question 4

The final method you applied was *k*-means clustering. However, given that this method is non-hierarchical, the management team is unsure how to interpret the results without a data visual, such as a dendrogram. As a result, they have asked you to compile a brief final report profiling the *k*-means clustering solution, providing insight into the following:

* The frequency of observations within each cluster.
* How the clusters might have been formed, based on the means of the different variables that make up each cluster.

To answer this question, consider and characterise each cluster individually.

(Max. 300 words)

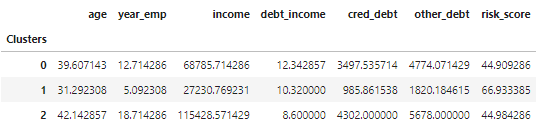
Start writing here:

The k-means clustering method is an iterative method and non-hierarchal and will therefore not produce a dendogram.



The allocations of the observations have been distributed relatively better than the hierarchal methods. The third cluster has the lowest observation, which is what this method has in common with the hierarchal methods.

Table 1: Cluster Profiling



From the cluster means graph it can be seen that the age, income and year employed are the observations that seem to have the highest variation with their means, while the risk score has cluster 1 and 2 with little variance means between them. From face value, the risk score seems like the observation that may not contribute in differentiating between the clusters. However, it is important to investigate the statistical significance of all the observations to determine this.

From the ANOVA test, the top three observation with the highest F-value is the ‘income’, ‘year employed’ and ‘other debt’. These also correspond to the lowest p-value respectively. This means these are the observations that have statistical significance between other observations and contribute strongly to differentiating between the three clusters.

The clusters are classifying the Speedy loan customers generally by their income. From the result from the Question 1, we can conclude that on average, the younger clients are most likely to be declined a loan due to their lower credit and ‘other’ debt.

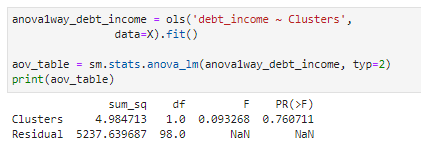
The clusters can be seen as:

Cluster 0: Average-Earning Customers

Cluster 1: Lower Earning Customers

Cluster 2: High-Earning Customers

It is also important to note that the ‘debt-to-income’ ratio has a very low F-value. This means that this observation, has a high variation within the different clusters. The P-value is also much higher than the significance level of 0.05. Debt-to-income ratio can therefore be removed in the clustering of the observations.



#### 4. Rubric

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **No Submission** | **Poor** | **Good** | **Excellent** |
| **Question 1**  *The image provided contains the correct graphs and uses the correct variables, adhering to all instructions. The interpretation of the results is adequate.* | No submission or attempt. (0) | The image provided contains the incorrect graphs, or the incorrect variables have been used. The interpretation of the results is poor. (1) | The image provided contains the correct graphs and uses the correct variables, but not all the instructions are adhered to. The interpretation of the results is acceptable. (3) | The image provided contains the correct graphs and uses the correct variables, and all instructions are adhered to. The interpretation of the results is exceptional. (5) |
| **Question 2**  *The student correctly describes the effectiveness of single-linkage clustering using different distance metrics. Observation allocations are used to support this answer, and a final recommendation is provided.* | No submission or attempt. (0) | The student attempts to describe the effectiveness of single-linkage clustering, but fails to reference some of the different distance metrics or observation allocations. (4) | The student correctly describes the effectiveness of single-linkage clustering, but fails to reference some of the different distance metrics or observation allocations (6) | The student correctly describes the effectiveness of single-linkage clustering and references the different distance metrics and observation allocations to support their answer. A final recommendation is provided. (10) |
| **Question 3**  *The student correctly describes the effectiveness of complete-linkage clustering using different distance metrics. Observation allocations are used to support this answer, and a final recommendation is provided.* | No submission or attempt. (0) | The student attempts to describe the effectiveness of complete-linkage clustering, but fails to reference some of the different distance metrics or observation allocations. (4) | The student correctly describes the effectiveness of complete-linkage clustering, but fails to reference some of the different distance metrics or observation allocations. (6) | The student correctly describes the effectiveness of complete-linkage clustering, and makes reference to the different distance metrics and observation allocations to support their answer. A final recommendation is provided. (10) |
| **Question 4**  *The student successfully profiles each cluster derived from the* k*-means clustering method. Observation allocations to each cluster are mentioned.* | No submission or attempt. (0) | The student attempts to profile each cluster, but fails to mention observation allocations in each. (5) | The student provides a satisfactory profile of each cluster, but fails to mention observation allocations in each. (10) | The student provides an exceptional and detailed profile of each cluster, and provides the correct observation allocations in each. (15) |

**Total:** 40 marks