Project 2: Credit K-Means Clustering

Manuel Duran

DSC 680

Bellevue University

- **Overview**
    - Credit is an important aspect that many individuals are not able to understand within the world today. In fact, approximately, 16 percent of Americans have incredibly poor credit, while only 1.2 percent of Americans have perfect credit to work with (**Investopedia)**. Individuals within the world today take a line of credit from our banks each day. The banks need to take in several determining factors to decide which individuals should qualify for a line of credit. This included age, amount within their savings\checking account, duration, and purpose of taking the credit. With all this information that the bank must process, must understand how to decide on who to provide credit lines too as well as how to market to them. This is where clustering\segmentation comes into to play to assist with those decisions **(Medium).** By using this method, we can determine how to market each line of credit to each customer and understand each situation. Understanding how to use this method properly we can create clusters from the bank data to discover new insights about the group of customers.

- **Data Understanding**
  - The dataset for my project was acquired from Kaggle.com. The dataset featured different numerical and categorical variables to work with when utilizing K-Means clustering. Reviewing the data, I began to try and understand which variables would be a good or bad indicator when it comes to a bank trying to decide to provide a credit to an individual.
- **Data Preparation**
  - The dataset that I utilized needed some corrections when it came to data preparation. First, I needed to adjust the dataset to not have any null values so it would not affect my clustering analysis. There were several missing values within the checking and savings account columns, so I decided to use imputation on the average amount. Additionally, when performing data preparation, I needed to remove symbols and certain spelling of variable columns. This is because when it came to performing my clustering analysis, certain spellings of the columns would not take in the analysis. Lastly, I added a job directory to include the skills of each individual to allow me to create different visualizations for EDA.

- **EDA Insights**
    - To learn more about the data I was working with I created several data visualizations to attempt to discover actionable insights. The first Insight was discovered when creating histograms of the age variables. I was able to see that many young individuals were attempting to get credit from the bank. This makes sense because this could be due to young individuals attempting to build their credit in their life (Appendix A). Following the age distribution, I geneated several horizaontal bar graphs of the dataset. One graph that stood out was the savings account section where there were many indviduals that were under that category for little for their savings. It makes sense because people well off would not need to credit compared to people who have little (Appendix B). The next visualization that presented me with some more insight was the corrleation matrix. When created I was able to discover the corelation between credit total and duration of the credit (Appendix C). The last visualizaton that provided actionable insight during my eda process was a Seaborn Regplot I created between Credit Amount and Duration. Reviewing the plot, we see the trend that was previously discussed from the correlation matrix in action. The trend discovered is that with the increase of credit

amounts, the increase of duration. This makes sense many inviduals who take the credit will not be paying it back as soon as they accept it (Appendix D).

- **Method and Interpretation**
  - The method of choice for the project was K-Means Clustering. When performing my code on the 1000 record dataset I ended up with a total of 4 unique clusters (Appendix E). Additionally, I generated the results of each cluster within a table format to include Age, Job, Credit Amt and Duration (Appendix F). Lastly, to gain more insight of the clusters I generated Bar Graphs for each cluster and several Box Plots on the cluster analysis. Examples of this can be seen in Appendix G; however, for the total analysis it can be seen within my notebook (Appendix G).

- **Interpretation and Conclusion Plan**
  - Overall, I believe we have gained much value from each cluster discovered from the project to provide the bank with a proper action plan on how to market to their customers. The bank will need to market to customers between the ages of 20-35. This was discovered by looking at the average age of each cluster, and many of the young

individuals are looking into developing their credit at an early age. Additionally, if the bank is looking for faster turnaround on their credit process being paid back, then they would like to investigate Clusters 2 and 3, due to shorter duration periods. If the bank would like to market to both men and women, then they would need to investigate Clusters 0 and 1. Cluster 1 has all females within its cluster which cluster 0 has the most men of all the clusters. The only liability when performing that action would be the duration period of each credit. Lastly, when reviewing the main purposes of why the individuals are attempting to get a credit from the bank, it is revealed that the top categories include Car, Radio\TV, and Furniture\Equipment. By knowing this the bank could team up with a local dealership or furniture location and develop a marketing plan if a customer were to take a credit from the bank. I am not sure if this would be possible, but the bank could create a rewards program where if customers pay their credit earlier in a small duration, they can earn rewards discounts where a percentage of the their payment gets deducted. It is something to tinker with; however, I think a rewards strategy would work very well. Lastly, the only limitation I would see is the dataset itself is the lack of female that are accounted for within the dataset. We can see from our cluster amounts that it is slightly skewed since some clusters do not have females at all. If more data is collected over time it would lead to greater results.

References

1. Customer Segmentation: Definition, Models. (2020, May 25). Optimove.

2. Rahmi, F. (2021, February 7). Credit Card Customers Segmentation - Analytics Vidhya. Medium. https://medium.com/analytics-vidhya/credit-card-customers-segmentation-bc3c5c87ddc

3. D. (2021, March 4). What is Market Segmentation? 4 Types & 5 Benefits. Lotame. https://www.lotame.com/what-is-market-segmentation/

4. Chand, S. (2014, February 24). Market Segmentation: 7 Bases for Market Segmentation | Marketing Management. Your Article Library. https://www.yourarticlelibrary.com/marketing/marketing-management/market-segmentation-7-bases-for-market-segmentation-marketing-management/27959

5. Ehrens, T. (2019, April 4). customer segmentation. SearchCustomerExperience. https://searchcustomerexperience.techtarget.com/definition/customer-segmentation

6. What Is Considered Bad Credit? (n.d.). Investopedia. Retrieved October 4, 2021, from https://www.investopedia.com/terms/b/bad-credit.asp#:%7E:text=A%20person%20is%20considered%20to,or%20obtain%20a%20credit%20card.

7. Good Credit. (n.d.). Investopedia. Retrieved October 4, 2021, from https://www.investopedia.com/terms/g/good-credit.asp

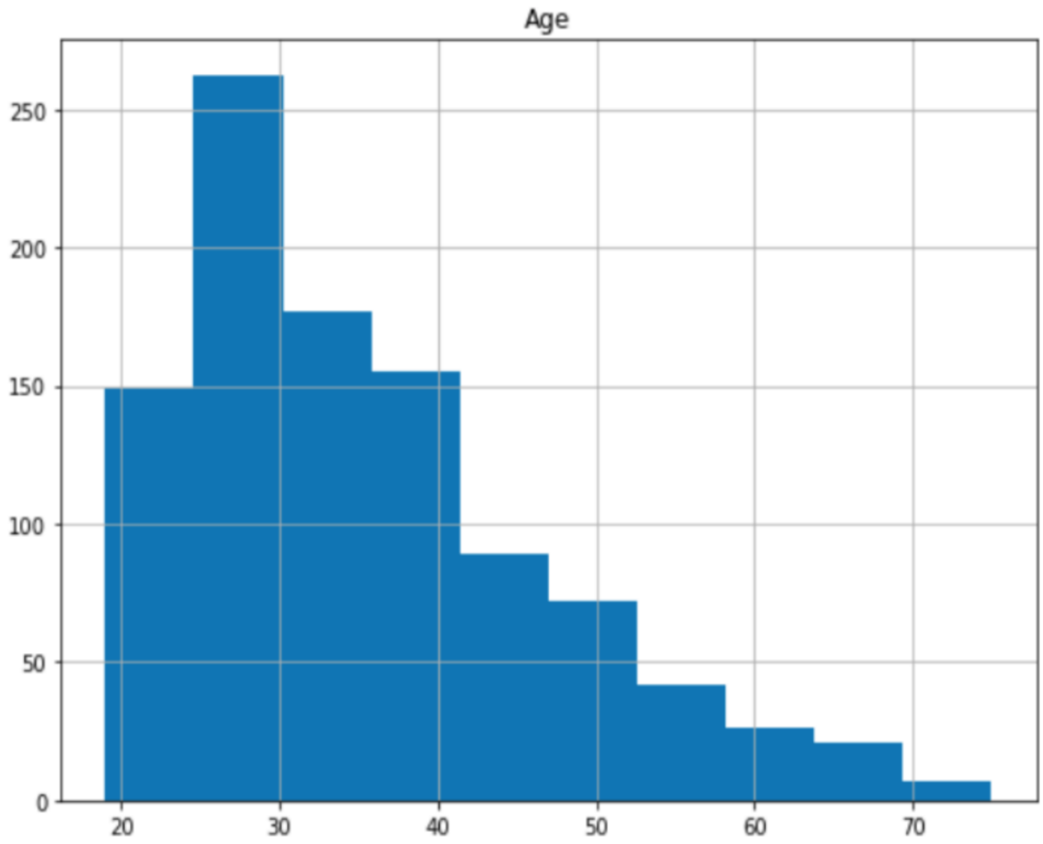
8. Khalid, I. A. (2020, June 1). Customer Segmentation in Python - Towards Data Science. Medium. https://towardsdatascience.com/customer-segmentation-in-python-9c15acf6f945

9. Vickery, R. (2019, September 6). Segmenting Credit Card Customers with Machine Learning. Medium. https://towardsdatascience.com/segmenting-credit-card-customers-with-machine-learning-ed4dbcea009c

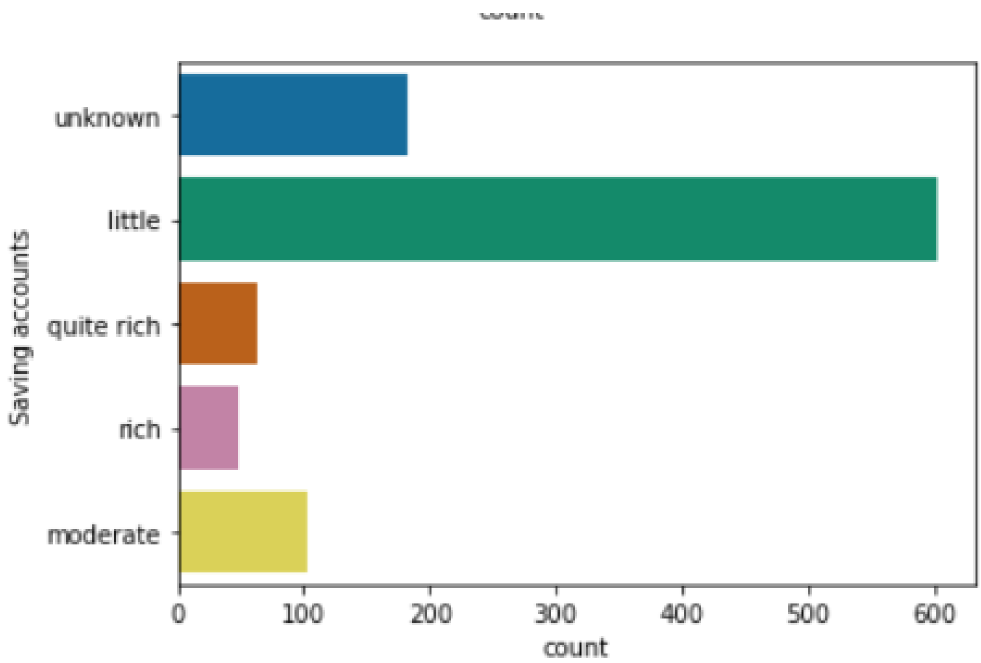
10. **Brownlee, J. (2021, January 5). Develop a Model for the Imbalanced Classification of Good and Bad Credit. Machine Learning Mastery. https://machinelearningmastery.com/imbalanced-classification-of-good-and-bad-credit/**

Appendices

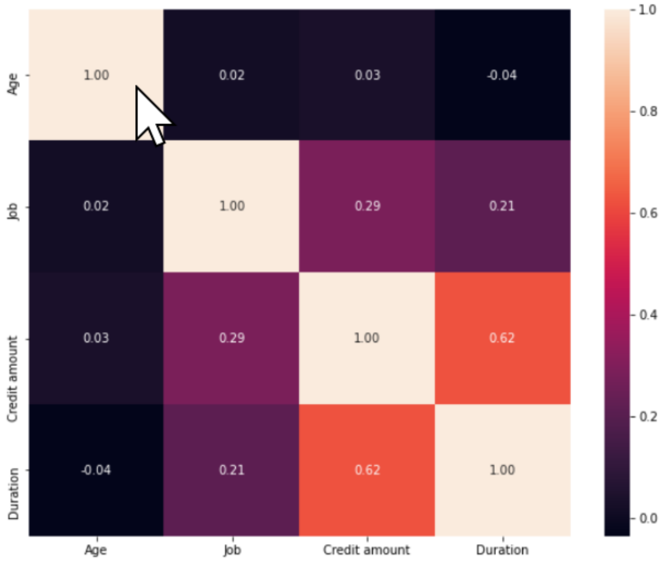
Appendix A



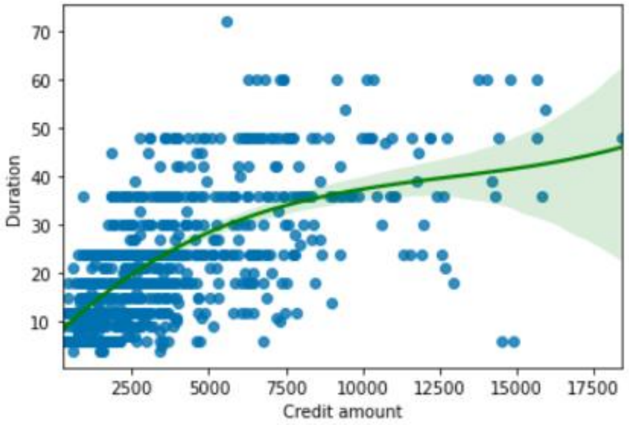
Appendix B



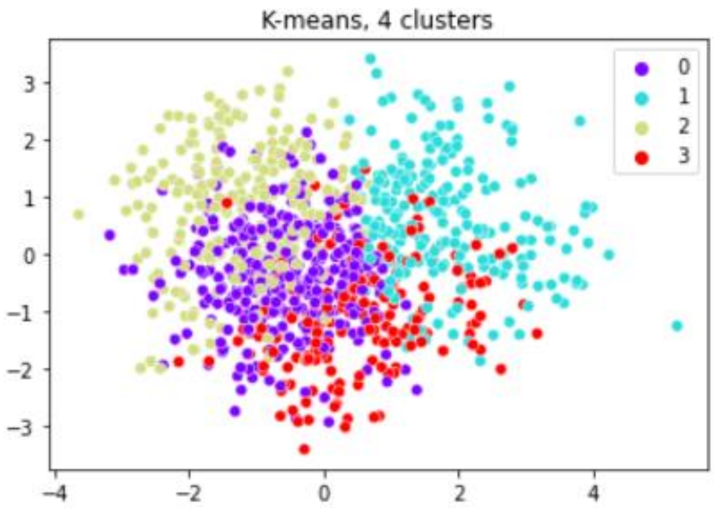
Appendix C



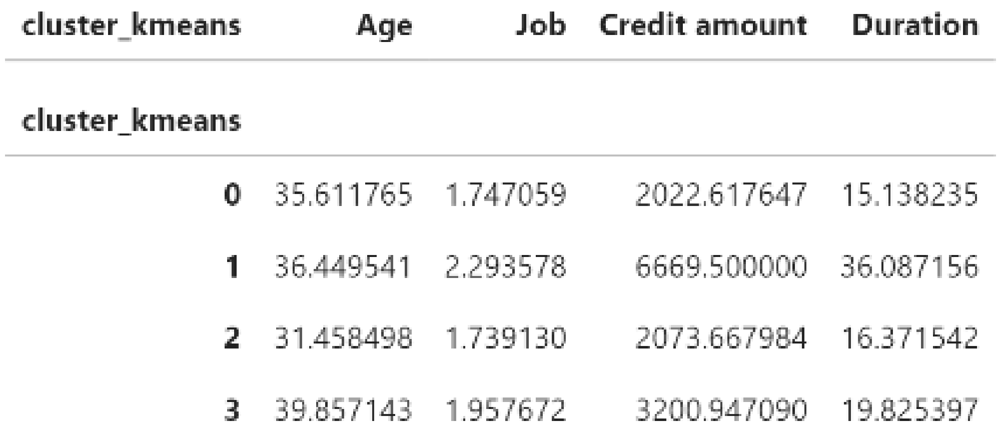
Appendix D



Appendix E

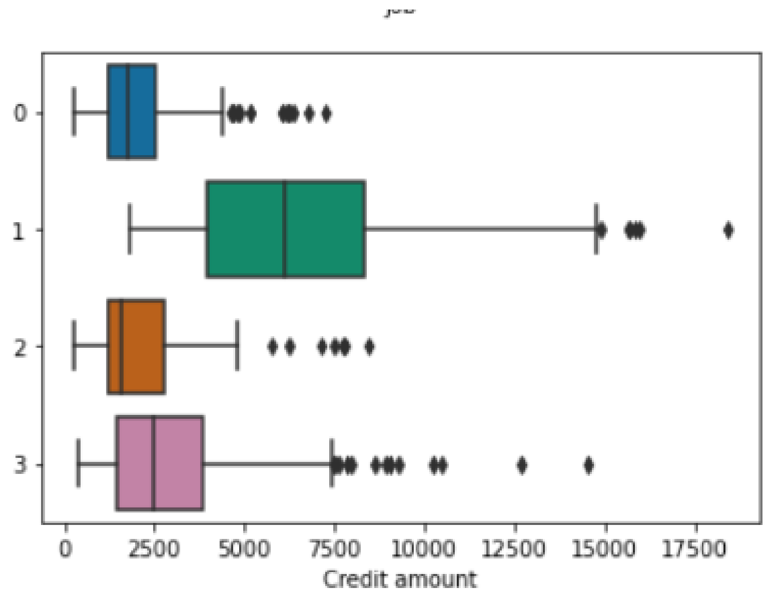


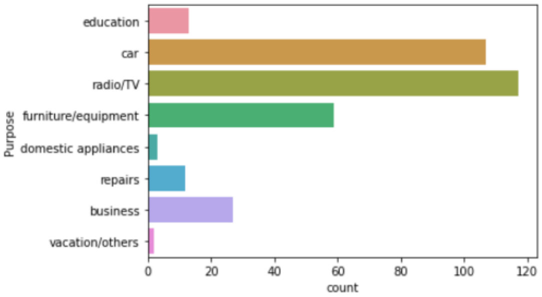
Appendix F



cluster_kmeans	Age	Job	Credit amount	Duration
cluster_kmeans				
0	35.611765	1.747059	2022.617647	15.138235
1	36.449541	2.293578	6669.500000	36.087156
2	31.458498	1.739130	2073.667984	16.371542
3	39.857143	1.957672	3200.947090	19.825397

Appendix G


Job
0
1
2
3
0
2500
5000
7500
10000
12500
15000
17500
Credit amount


education
car
radio/TV
furniture/equipment
domestic appliances
repairs
business
vacation/others
Purpose
0
20
40
60
80
100
120
count