

Clustering of Credit Card Users

Project Start Date:	June 13, 2024
Project Creators:	Isaiah Mariano, Jeremiah Regalario
University:	University of the Philippines Diliman
Project End Date:	August 6, 2024
Project Tags:	<i>k-means, pca, credit card</i>

Introduction

Banks in modern day Philippines typically tailor their services in order to meet the specific needs of their different customer groups. However, addressing the unique preferences of the multitude of characteristics of the customers in the country remains a huge challenge. Hence, this project was implemented in order to attempt to cluster BPI credit card users in using data provided by BPI. Nine datasets were provided, and exploratory data analysis was conducted on all of them in order to gain useful and meaningful insights and to determine which parts of the data to consider in the clustering process.

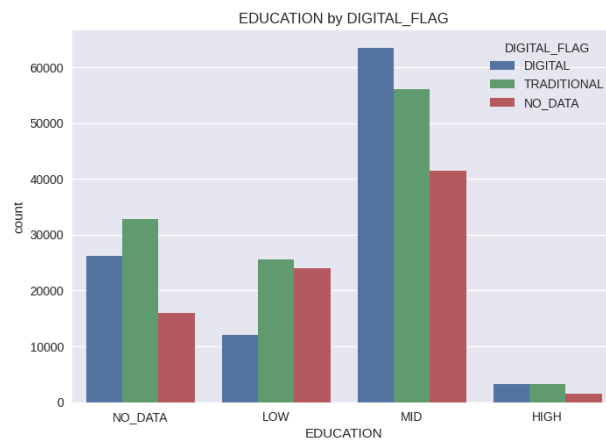
Methods

Prior to the data processing, the team prepared the dataset by extracting each dataset, applying the necessary cleaning and filtering techniques, such as removal of duplicate entries, renaming of columns, and null values removal.

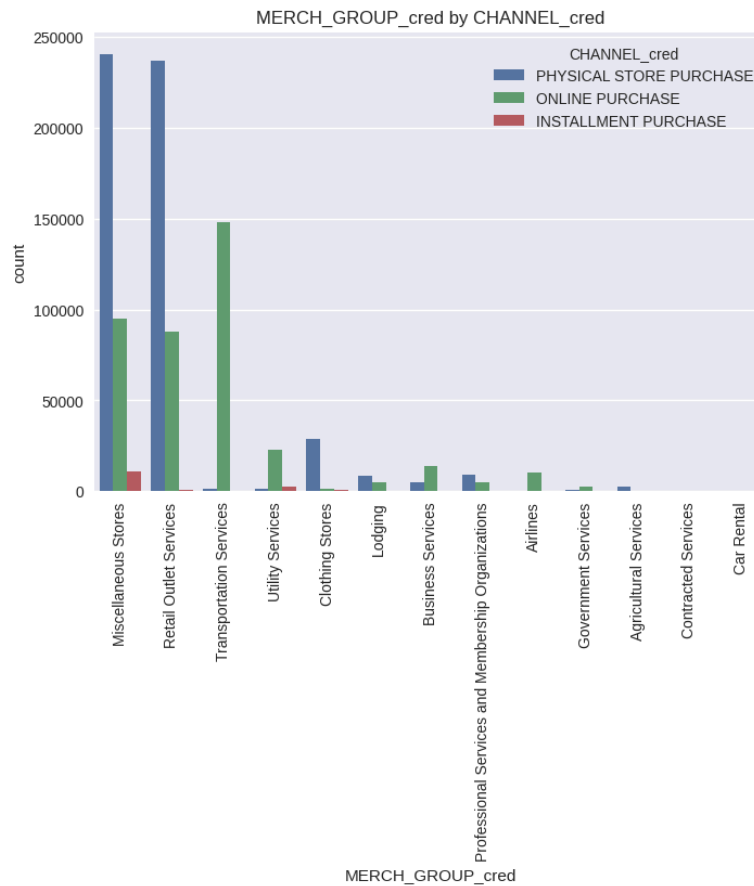
Rendering of univariate plots such as histograms, bar charts, and pie charts for the dataset became the next step of the team in order to have initial insights regarding the demographics and distributions of spending habits of the BPI credit card users.

Next, bivariate analysis was conducted where boxplots, scatterplots, and bar plots were used to present the relationships between two variables.

Customer Info

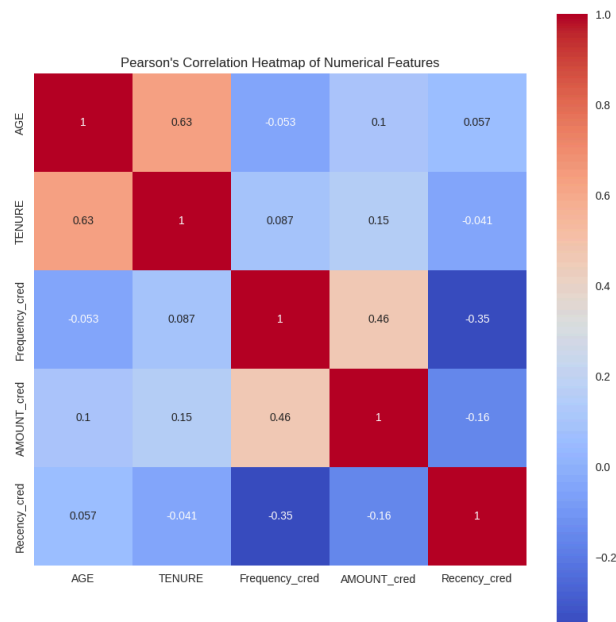


Credit Card Transactions

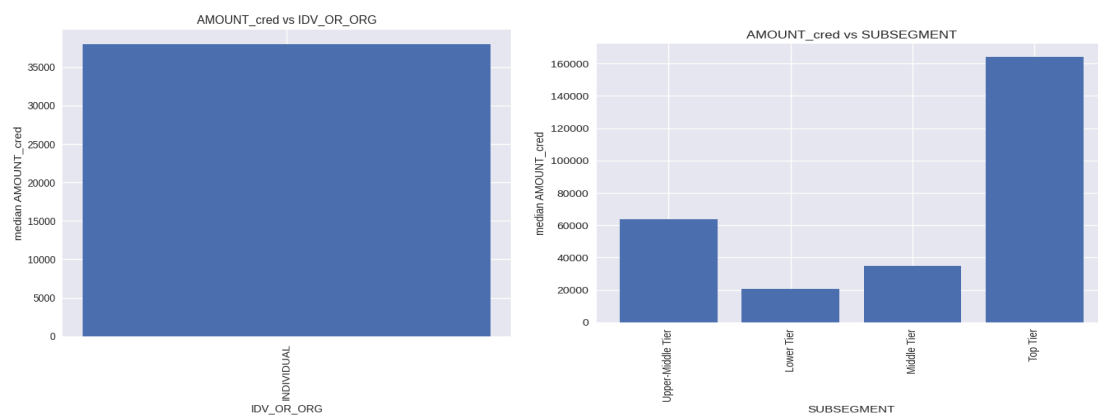


Afterward, feature engineering was implemented in order to add the Recency, Frequency, and Monetary (total amount of transactions) for RFM analysis by means of multivariate analyses and clustering.

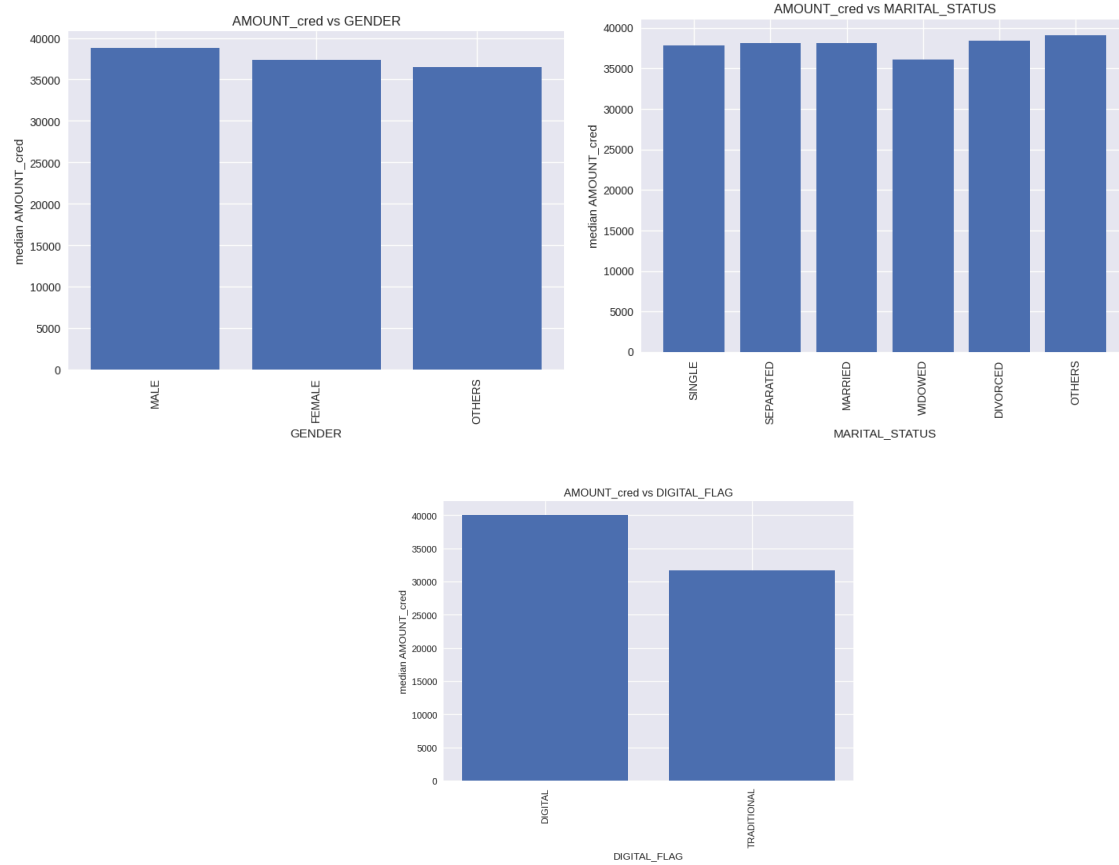
Following this are the multivariate analyses through visualizations like scatterplots, box plots, 3D plots and Pearson Correlation heat map. Null values were removed, and the mean and median plots were also included.



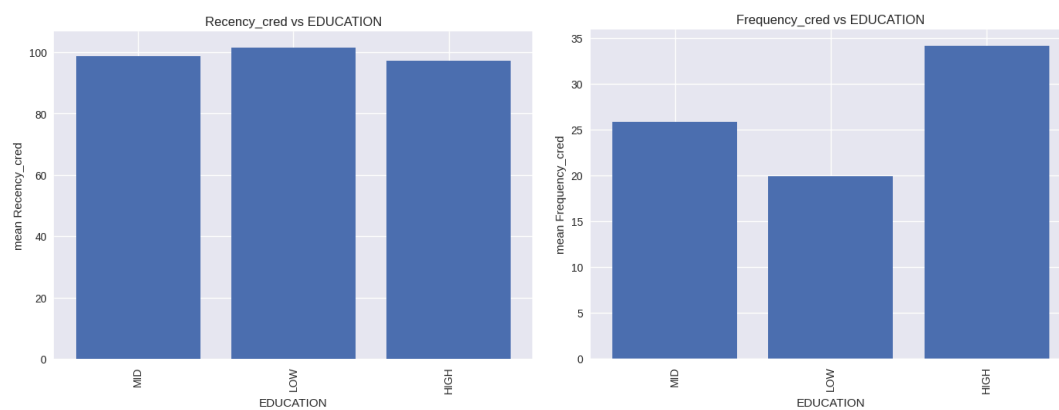
Median Plots

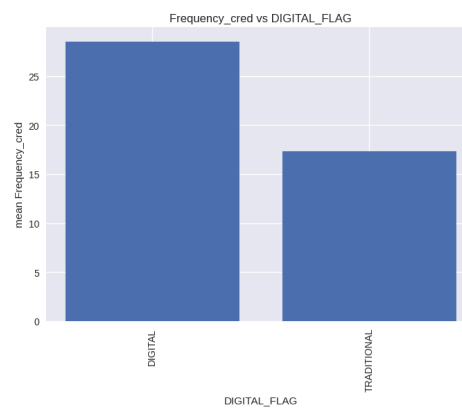
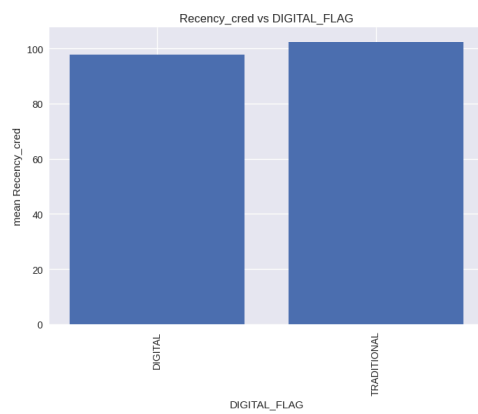
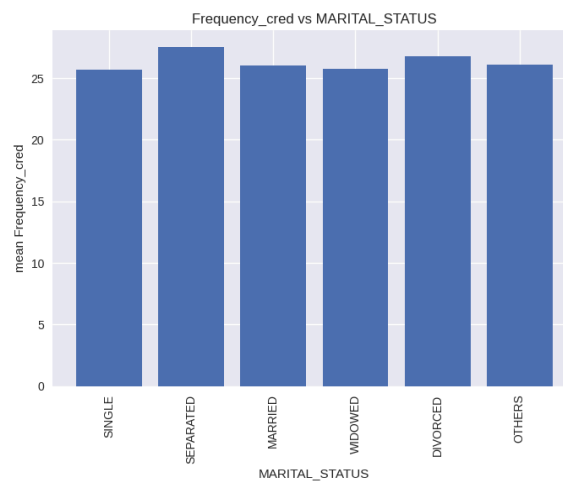
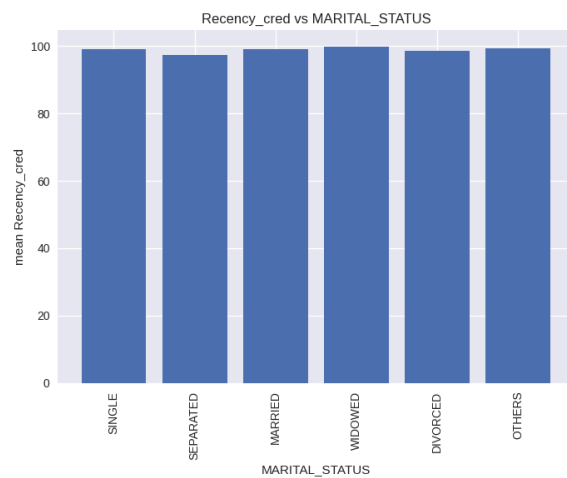
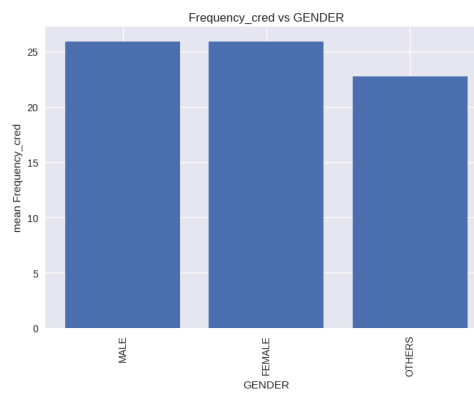
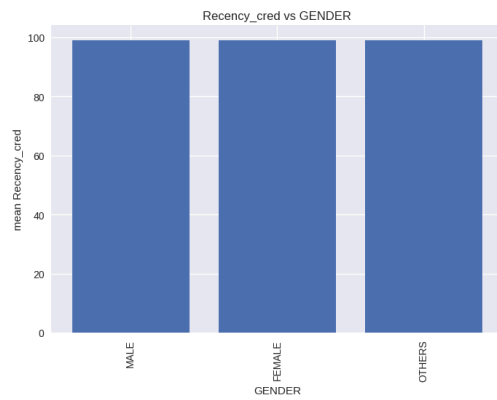


(organizations are not credit card users?)

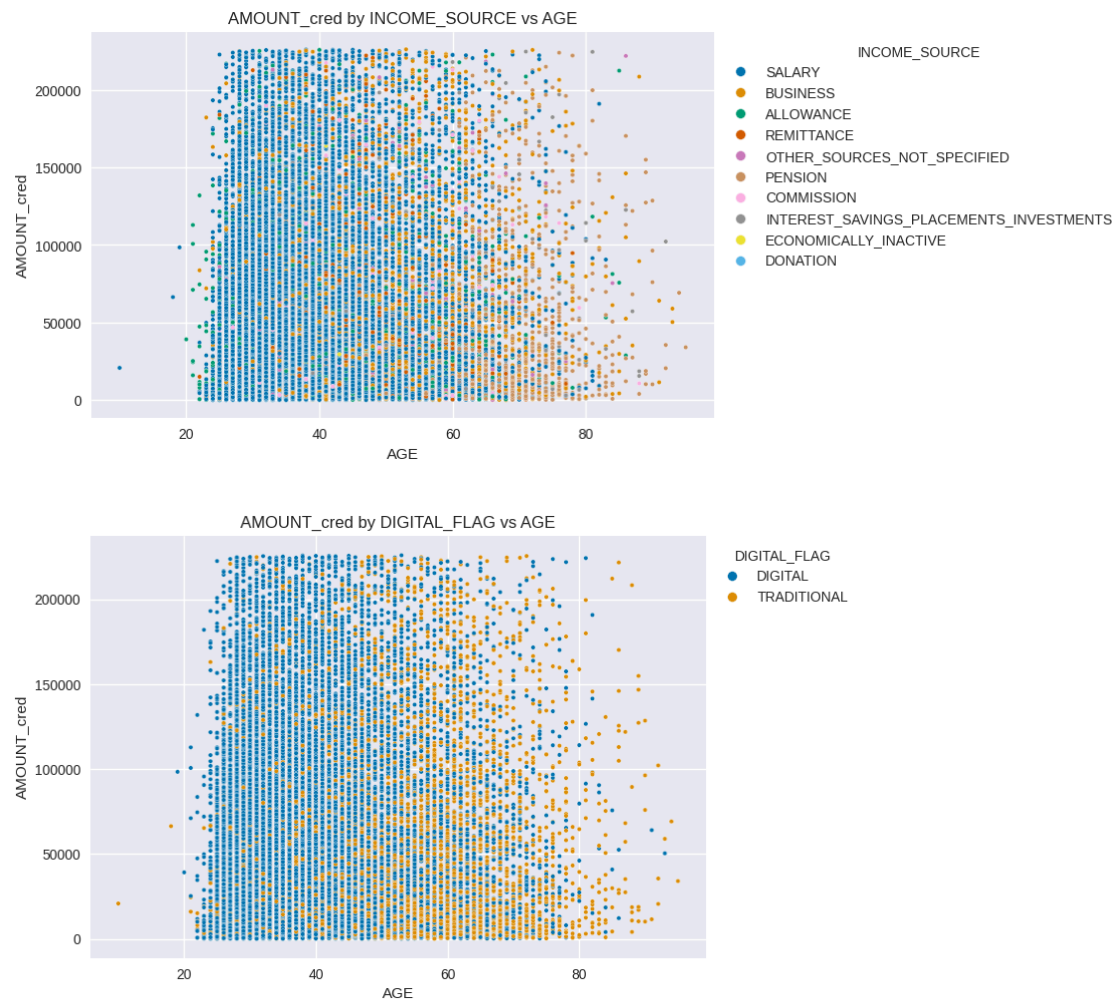


Mean Plots

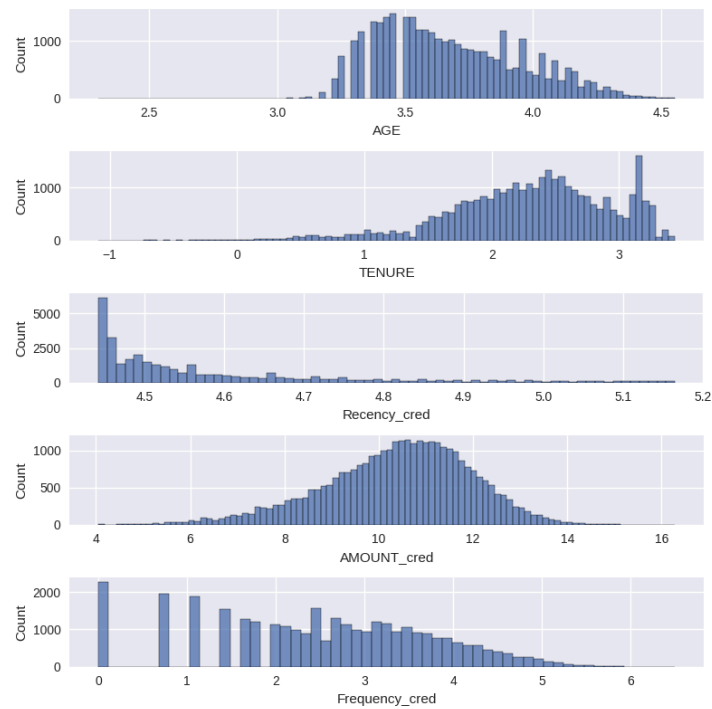




Multivariate Plots (relevant)



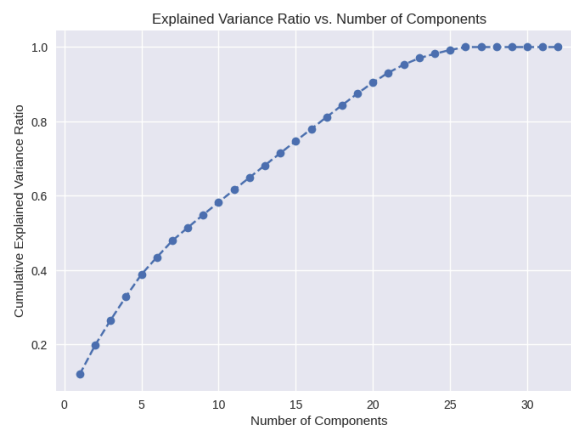
Prior to clustering, data processing was first performed. All the features were selected except for province of the credit card transactions were selected to avoid the problem of high dimensionality (high number of features). Next, the categorical features were converted to numerical features since the clustering algorithms that were chosen can only interpret numerical values: One-Hot Encoding for BUSINESS_OWNER, DIGITAL_FLAG, INCOME_SOURCE, IDV_OR_ORG, GENDER, and MARITAL_STATUS, and Label Encoding for EDUCATION, REGION, and SUBSEGMENT since it is ordinal data (Sethi, 2023). After this, logarithmic transformation was performed on the dataframe. This removes the right-skewness of the numerical features (since K-Means prefer normal distributions) and reduces the impact of outliers.



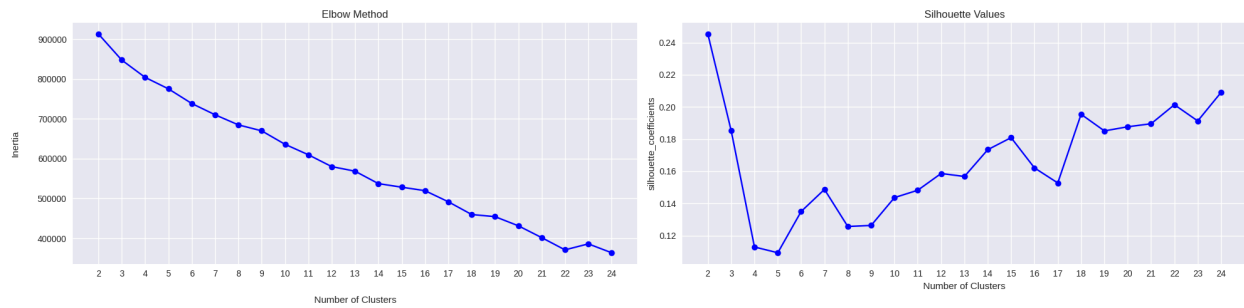
The features were also standardized using StandardScaler in order to make the features be in the same scale and for K-Means to consider each feature equally (Ryzhkov, 2020). The algorithm implemented was K-Means with PCA.

Results

For K-Means with PCA, the Explained Variance Ratio has been used, and it has been decided that 25 features are to be retained.



The elbow and silhouette methods were used. The silhouette scores were not good (~ 0.11 - ~ 0.24) and the elbow method suggests that a large number of clusters is required.



Discussion

References

- Chandra, M. G. (2021). *9 Clustering Techniques for Customer Segmentation*. Kaggle.
<https://www.kaggle.com/code/gaganmaahi224/9-clustering-techniques-for-customer-segmentation/notebook#Data-visualisation>.
- Connectif. (July 18, 2022). *What Are RFM Scores and How to Calculate Them*.
<https://connectif.ai/en/blog/what-are-rfm-scores-and-how-to-calculate-them/>
- DataReportal. (February 9, 2023). *Digital 2023: Philippines*.
<https://datareportal.com/reports/digital-2023-philippines>
- Google Developers. (n.d.). *Advantages and Disadvantages of Clustering Algorithms*.
<https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages>
- Kar, S. (2021). *Customer Segmentation - EDA & Clustering (K-Means)*. Kaggle.
https://www.kaggle.com/code/sonerkar/customer-segmentation-eda-clustering-kmeans?fbclid=IwAR2Ich1pmauPuUNVgg387VjY-vk11wBTKYzY1afN6SthUXAgCH3nzxE9_Do
- Kwiatkowska, P. (2019). *Bank Customer Segmentation*. Kaggle.
https://www.kaggle.com/code/paulinan/bank-customer-segmentation?fbclid=IwAR0qt7ofRkp-bNjkUYNq35oCZtum51ljZ_h3fnRwlCD75V__7XVVqKi54qo
- LinkedIn. (n.d.). *How Do You Validate and Compare Clustering Results with Missing Data?*
<https://www.linkedin.com/advice/3/how-do-you-validate-compare-clustering-results-missing>
- Mohanty, S. K. (2021). *Customer Segmentation - K-Means, DBSCAN, GMM, Agglo*. Kaggle.
<https://www.kaggle.com/code/sidharth178/customer-segmentation-kmeans-dbscan-gmm-agglo>
- MoneyMax Philippines. (December 27, 2017). *Savings Account vs. Current Account*.
<https://www.moneymax.ph/personal-finance/articles/savings-account-vs-current-account>
- Ryzhkov, E. (July 23, 2020). *5 Stages of Data Preprocessing for K-Means Clustering*.
<https://medium.com/@evgen.ryzhkov/5-stages-of-data-preprocessing-for-k-means-clustering-b755426f9932>
- Sethi, A. (July 15, 2023). *One-Hot Encoding vs. Label Encoding using scikit-learn*. Analytics Vidhya.
<https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/>

Sharma, A. (2020, September 7). *How to Master the Popular DBSCAN Clustering Algorithm for Machine Learning*. Analytics Vidhya.

<https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/>

Shawky Elgendy (n.d.). *Customer Segmentation - EDA, K-Means, PCA*. Kaggle.

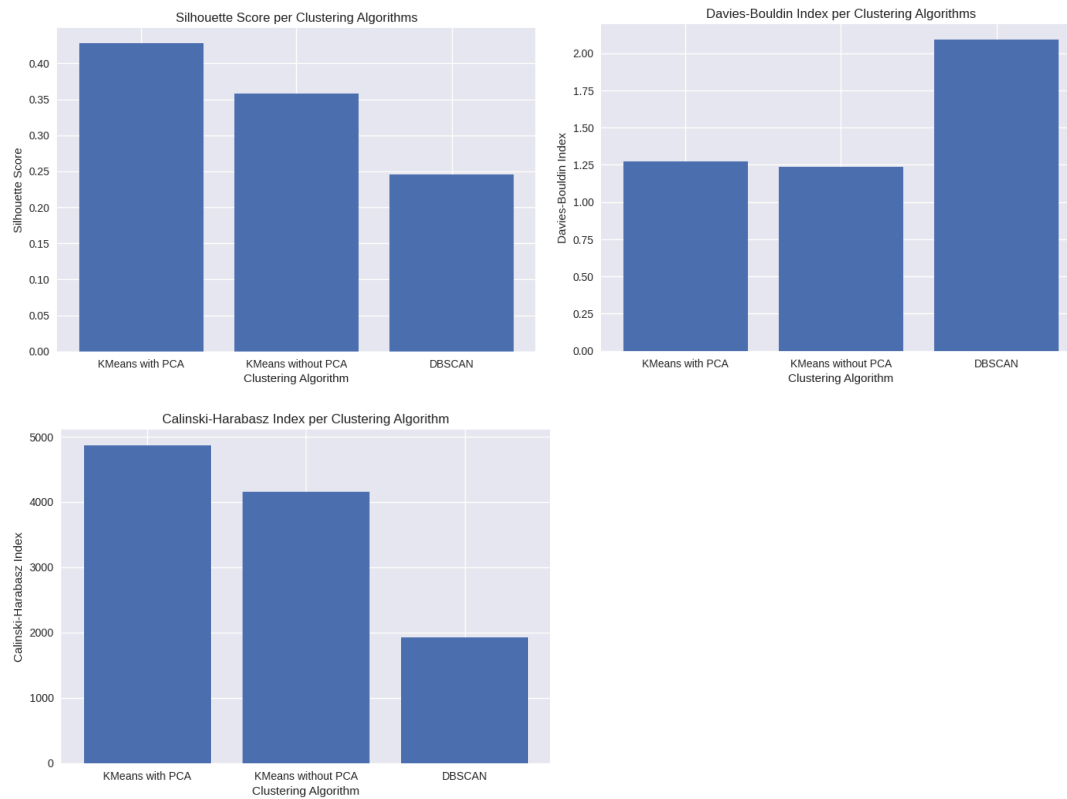
<https://www.kaggle.com/code/shawkyelgendy/customer-segmentation-eda-k-means-pca/notebook?fbclid=IwAR1cig7xnspaDnVLqZ28klS7Qm9PPesx12oIVu5gcfK--E4hqBMPL3ps0gg>

Yadav, N. (n.d.). *Customer Segmentation using K-Means, HC, DBSCAN*. Kaggle.

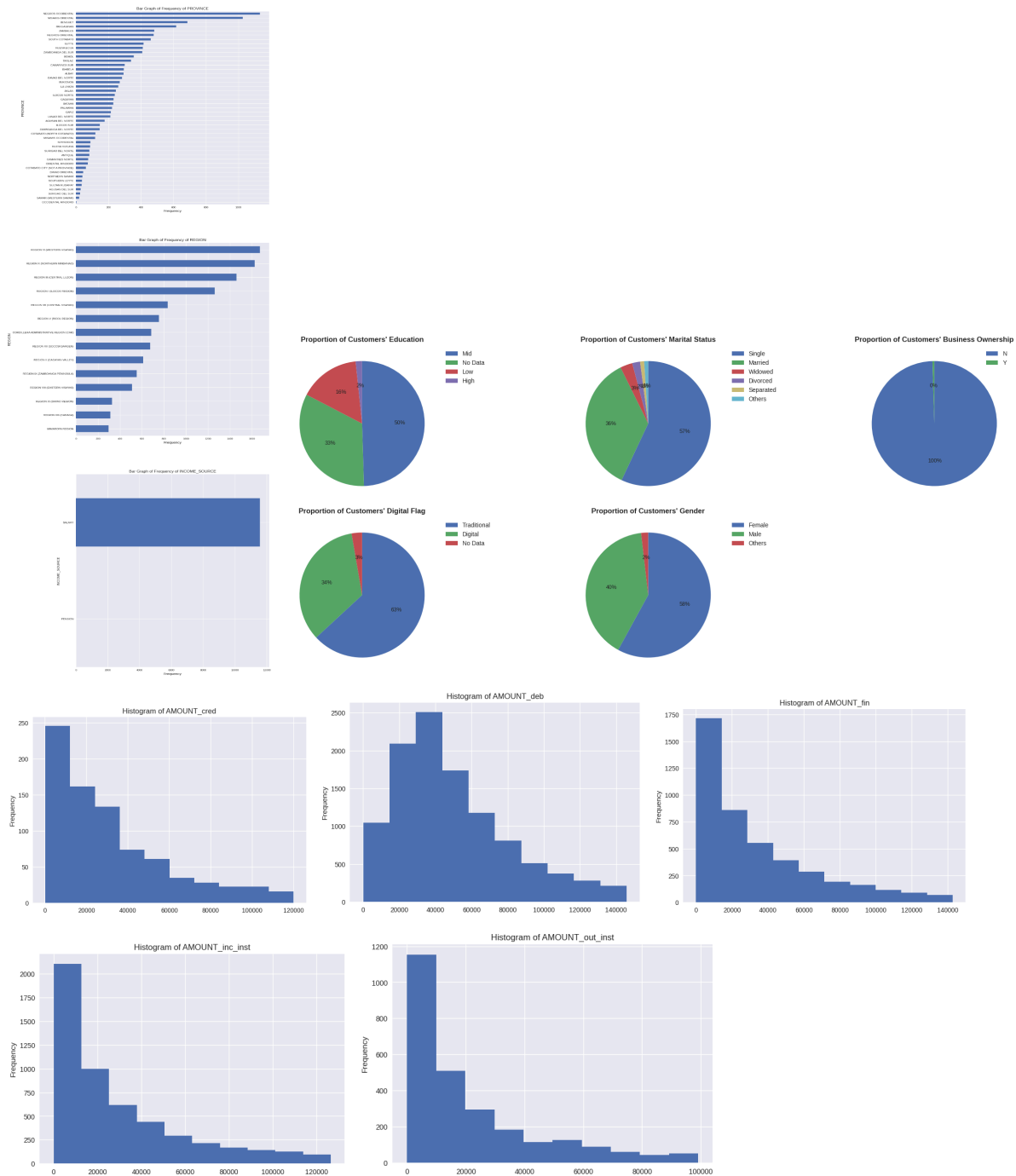
<https://www.kaggle.com/code/niteshyadav3103/customer-segmentation-using-kmeans-hc-dbscan>

Appendix

Changes Made



Cluster Validation Techniques were added.



Graphs that show the characteristics of the target segment were added.

AI Models Used

ChatGPT. (September 21, 2023). *Customer Segmentation Guide*.

<https://chat.openai.com/share/124cd1ab-0564-45a4-9997-9bd13cb0f1e5>

ChatGPT. (October 13, 2023). *Insurance Requirements by Employment Type*.

<https://chat.openai.com/share/ba9d0e8b-a6c0-475a-9c0f-ba0c603af40d>

Claude. (October 13, 2023). Key elements of a health insurance business model.

hi claude, i am proposing a health insurance project in my company. who are the key suppliers in this business model?

Here are some of the key suppliers that would be involved in a health insurance business model:

- Healthcare providers - Doctors, hospitals, pharmacies, labs, and other medical facilities that provide services to insured members. They supply the actual healthcare services.
- Insurance clearinghouses - Process and facilitate medical claims between healthcare providers and insurers. Key for claims data exchange.
- Pharmacy benefit managers (PBMs) - Manage prescription drug benefits on behalf of health plans. Supply drug coverage services.
- Medical device manufacturers - Companies that produce health equipment like pacemakers, crutches, blood pressure cuffs, etc. They supply medical devices.
- Health information technology vendors - Provide software, analytics, population health tools, and digital infrastructure for insurance operations and care delivery.
- Brokerages and agencies - Enroll members into health plans and serve as a distribution network. Supply member acquisition and servicing.
- Actuarial and underwriting partners - Help analyze risk, set premium pricing, and structure health plan benefits. Supply actuarial capabilities.
- Third party administrators - Help administer health plan operations like claims processing. Can supply administrative services.
- Data providers - Supply data and analytics for risk assessment, fraud detection.

my business model template asks me "what key activities do our value propositions require?", what should i say?

Here are some potential key activities to consider for a health insurance business model:

- Providing member services - Activities like enrollment, education, claim support, handling inquiries to deliver excellent member experience.
- Administering benefits - Key tasks like claims processing, administering copays/deductibles, managing pharmacy benefits, coordinating care, ensuring accurate payments.
- Managing provider networks - Contracting with and maintaining networks of doctors, hospitals, pharmacies to provide adequate access for members.
- Controlling healthcare costs - Utilization review, care coordination, auditing, provider negotiation to optimize spending on member care.
- Managing risk - Underwriting, pricing, analyzing actuarial data, detecting fraud to effectively evaluate and manage financial risk.
- Managing operations - Overseeing technology, infrastructure, reporting, regulatory compliance, security, workflows to enable efficient operations.
- Acquiring members - Marketing, selling through brokers/agents, managing distribution partnerships to attract and retain members.
- Developing health plans - Designing a portfolio of health insurance products to meet diverse member needs and segments.
- Population health management - Aggregating data to understand needs,

Before i move on, i want you to take note that this project aims to provide accessible and affordable health insurance for the far-flung areas of the Philippines. The target customer segment characteristics are the following:

now for the value proposition section, my business model asks me these questions: "what values do we deliver to the customer?", "which one of our customers problem are we trying to solve?" "which customers need are we satisfying?" "what products and services are we offering to our customer segments."

Given your aim to provide accessible and affordable health insurance to underserved areas in the Philippines, here are some potential value propositions to consider:

- Accessibility - Offering health insurance that is convenient and easy to obtain for populations in remote areas through community outreach, mobile enrollment, telehealth services.
- Affordability - Providing coverage at low premium costs with discounted rates, government subsidies or sponsored programs to make insurance affordable for lower-income populations.
- Health protection - Giving peace of mind by protecting individuals and families from financial hardship in case of illness, injury or hospitalization through insurance coverage.
- Health education - Educating members about prevention, wellness, managing chronic conditions to improve health literacy and outcomes.
- Localized care - Building networks of nearby rural clinics, midwives, pharmacies and

which key resources and activities are the most expensive?

For a health insurance company focused on rural populations, some of the most expensive key resources and activities would likely include:

Most Expensive Resources:

- Provider networks - Contracting with and maintaining adequate networks of doctors, hospitals, clinics and pharmacies in rural areas is very costly.
- Health information technology - Robust IT systems for claims, billing, reporting and analytics require major investment.
- Call centers - Setting up rural call centers for localized member support carries high fixed costs.
- Sales force - Building an in-person sales team to do community-based outreach in remote areas is expensive.

Most Expensive Activities:

- Claims administration - Processing and paying medical claims is labor-intensive and carries high variable costs.
- Product development - Designing plans and pricing premiums requires actuarial analysis and regulatory approval.
- Member acquisition - In-person marketing and lead generation in rural areas costs more than urban campaigns.