

**Credit Card Customers Segmentation Project**

**Final Report**

A Github Project

Created by: Parth Patel

Email: pkpatelds@gmail.com

Sydney, Australia

# Executive summary:

The "Credit Card Customer Segmentation" project focuses on developing a data-driven marketing strategy for a financial institution by analysing the transaction behaviour of credit card holders. Over a six-month period, the behaviours of 8,500 active credit card users were meticulously studied to identify distinct customer segments based on their spending habits, credit utilization, and repayment patterns.

The project is divided into several phases:

1. **Data Preparation and Standardization:** The raw dataset, consisting of 17 features, was cleaned, processed, and standardized to ensure consistency across all variables. This step was essential for applying Principal Component Analysis (PCA) and clustering techniqueseffectively.
2. **Exploratory Data Analysis (EDA)**: A thorough analysis was conducted to uncover patterns, correlations, and outliers within the data, which informed the subsequent clustering process.
3. **Dimensionality Reduction using PCA:** To simplify the dataset and focus on the most significant features, PCA was applied, reducing the original 17 features to 5 principal components. These components captured the majority of the variance in the data, making them ideal for the subsequent clustering analysis.
4. **Clustering Analysis**: K-means clustering was employed to segment customers into distinct groups based on the KPIs. The optimal number of clusters was determined using the silhouette coefficient, leading to the identification of four unique customer segments.
5. **Insights and Strategic Recommendations**: Each cluster was analysed to understand its characteristics, enabling the development of tailored marketing strategies. For example, high-spending customers with good credit scores were identified as ideal candidates for loyalty programs and additional credit offers, while customers with poor credit scores were targeted for payment reminders and incentives to improve their repayment behaviour.
6. **Clusters behaviour on a Tableau Dashboard**: A tableau dashboard has also been prepared showing the customer behaviour in each of the 4 clusters. The metrics like Customer distribution by Clusters, Average credit card balance by clusters, Average credit card utilization, Spending and Payment behaviour has been presented on a Tableau Dashboard.

The project’s findings offer actionable insights that can significantly enhance the bank’s marketing strategies, allowing for more personalized and effective customer engagement.

Table of Contents

[Acknowledgement: 2](#_Toc174915578)

[Executive summary: 3](#_Toc174915579)

[Project Title: 5](#_Toc174915580)

[Reason for choosing this Subject: 6](#_Reason_for_choosing)

[Why we chose the clustering method & Data Source: 7](#_Why_we_chose)

[Overview of code and insights:](#_Overview_of_code) 8

1. [Importing Necessary Libraries: 8](#_1._Importing_Necessary)
2. [Data Loading: 8](#_2.Data_Loading)
3. [Data Cleaning: 8](#_3._Data_Cleaning)
4. [Missing Value Treatment: 9](#_4._Missing_Value)
5. [Adding New KPIs: 9](#_5._Adding_new)
6. [Exploratory Data Analysis (EDA): 10](#_6._Exploratory_Data)
7. [Feature Engineering and Principal component Analysis (PCA): 12](#_7._Feature_Engineering)
8. [Clustering (unsupervised learning): 13](#_8._Clustering_(Unsupervised)
9. [Interpretation & Business Insights: 15](#_9._Interpretation_and)
10. [Recommendations: 16](#_10._Recommendations:)

[Tableau Dashboard showcasing the clusters behaviour: 17](#_Tableau_Dashboard_showing)

# Project Title:

**Credit Card Customer Segmentation:**

**A Data-Driven Approach to Targeted Marketing**



In the competitive financial services industry, understanding customer behaviour is critical for designing effective marketing strategies and improving customer satisfaction.

This project involves the segmentation of credit card customers based on an analysis of their transaction behaviours over a six-month period. The dataset, initially containing 17 features, was standardized and reduced to 5 principal components using Principal Component Analysis (PCA).

These components were then used for clustering the customers into distinct segments via K-means clustering.

The project provides strategic insights that can help financial institutions tailor their marketing efforts, improve customer loyalty, and optimize resource allocation.

# Reason for choosing this Subject:

The subject of credit card customer segmentation was selected due to its significant impact on the banking and financial services industry. In today’s data-rich environment, the ability to analyse and segment customers based on their behaviour is a crucial asset for any financial institution. For credit card companies, understanding customer behaviours enables the creation of personalized products and services that meet individual needs.

This project’s focus on segmentation addresses a practical challenge faced by financial institutions: how to identify and target specific customer groups with tailored marketing strategies. By segmenting customers, banks can better allocate resources, enhance customer satisfaction, and increase profitability. The insights generated from this segmentation can also aid in risk management by identifying customers who may be at higher risk of default.

Additionally, the application of PCA to reduce dimensionality before clustering ensures that the analysis is both efficient and focused on the most critical aspects of customer behaviour, making this subject particularly relevant and valuable

# Why we chose the clustering method & Data Source:

**Why we chose the Clustering Method?**

Clustering was chosen as the primary analytical method for this project because it is a powerful unsupervised learning technique that uncovers natural groupings within a dataset. Unlike classification methods, clustering does not require predefined labels, allowing it to identify patterns and structures inherent in the data.

In this project, K-means clustering was applied to the 5 principal components derived from PCA. PCA was employed to reduce the dimensionality of the data from 17 original features to 5 key components, which retained the majority of the variance. This step was crucial in simplifying the dataset, making it more manageable and focusing the clustering process on the most important aspects of customer behaviour.

K-means clustering was particularly suitable for this project because of its efficiency and effectiveness in handling large datasets. By clustering customers based on the principal components, the analysis was able to identify distinct segments with similar spending patterns, credit utilization, and repayment behaviours.

One of the key advantages of using clustering in this context is its ability to provide a more granular understanding of customer segments, which can be leveraged to design highly targeted marketing strategies. The silhouette coefficient was used to validate the number of clusters, ensuring that the segmentation was both meaningful and actionable.

**Data Source**

The data used in this project was obtained from a public repository on GitHub, a widely recognized platform for sharing and collaborating on data science projects. The dataset comprises anonymized transaction records for 9,000 active credit card holders over a six-month period.

This dataset includes **17 original features**, covering various aspects of credit card usage, such as purchase amounts, cash advances, credit limits, and payment behaviours. These features provided a rich source of information for analysing customer behaviour. Following standardization, the dataset was further refined using PCA, reducing the 17 features to 5 principal components that captured the majority of the variance in the data.

The use of this dataset allowed for a comprehensive analysis of credit card customer behaviour, leading to meaningful and actionable segmentation insights. The public nature of the data also ensures that the methods and findings of this project can be easily replicated and validated by others.

# Overview of code and insights:

## 1.Importing Necessary Libraries

Importing libraries that are essential for data manipulation, analysis, and visualization. Some of them are:

* pandas, which is used for handling data in DataFrames.
* numpy is used for numerical operations.
* matplotlib.pyplot and seaborn are used for plotting graphs and visualizations.

## 2.Data Loading

After importing the libraries, we start by loading the dataset into the environment for analysis and processing. The dataset is loaded using the pandas library's **read\_csv()** function.

This reads the data from a CSV file into a DataFrame called **creditcarddata**. This step is essential to bring the data into the environment where it can be manipulated and analysed.

## 3. Data Cleaning

Once we have a dataframe we start preparing the data for analysis by handling any inconsistencies or errors. The steps for that are:

- Removing Duplicates: Check for and remove any duplicate rows that could skew analysis.

- Handling Missing Values: Address any missing data points (detailed in the next section).

- Standardizing Data: Ensure that all data is in a consistent format (e.g., numeric values, date formats).

Data cleaning ensures that the dataset is free from errors and inconsistencies, making the subsequent analysis more reliable.

## 4. Missing Value Treatment

Identifying and handling the missing values in the dataset, is what we take into account next. The **isnull().sum()** function identifies missing values in each column, by the column we found missing values in CREDIT\_LIMIT and MINIMUM\_PAYMENTS.

A screenshot of a computer

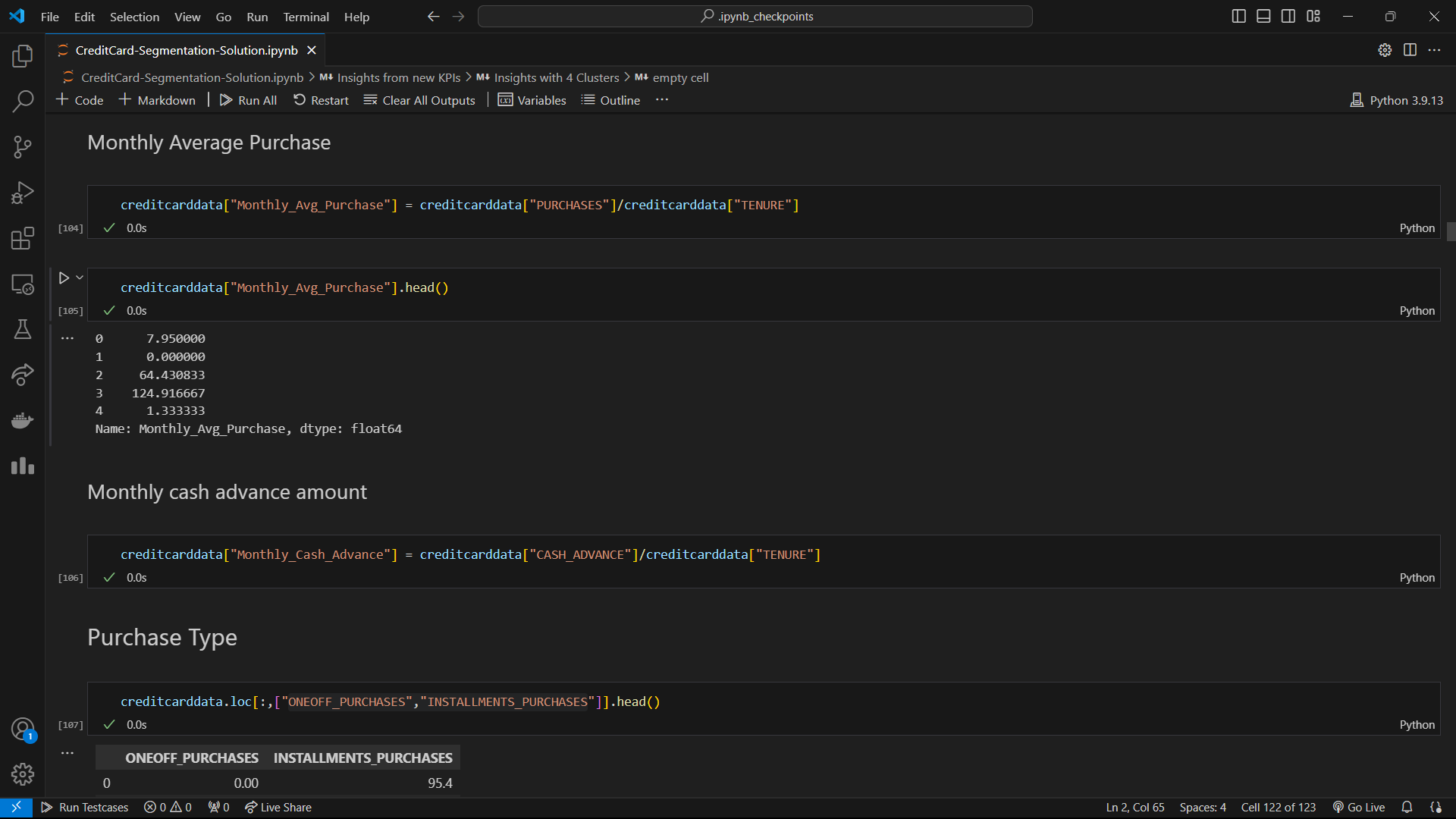
Description automatically generated A screenshot of a computer

Description automatically generated

- From the boxplot above we know the data is skewed, so we will use median to impute the missing data

## 5. Adding new KPIs

-Monthly Average Purchase:



-Monthly Cash Advance:

A screenshot of a computer

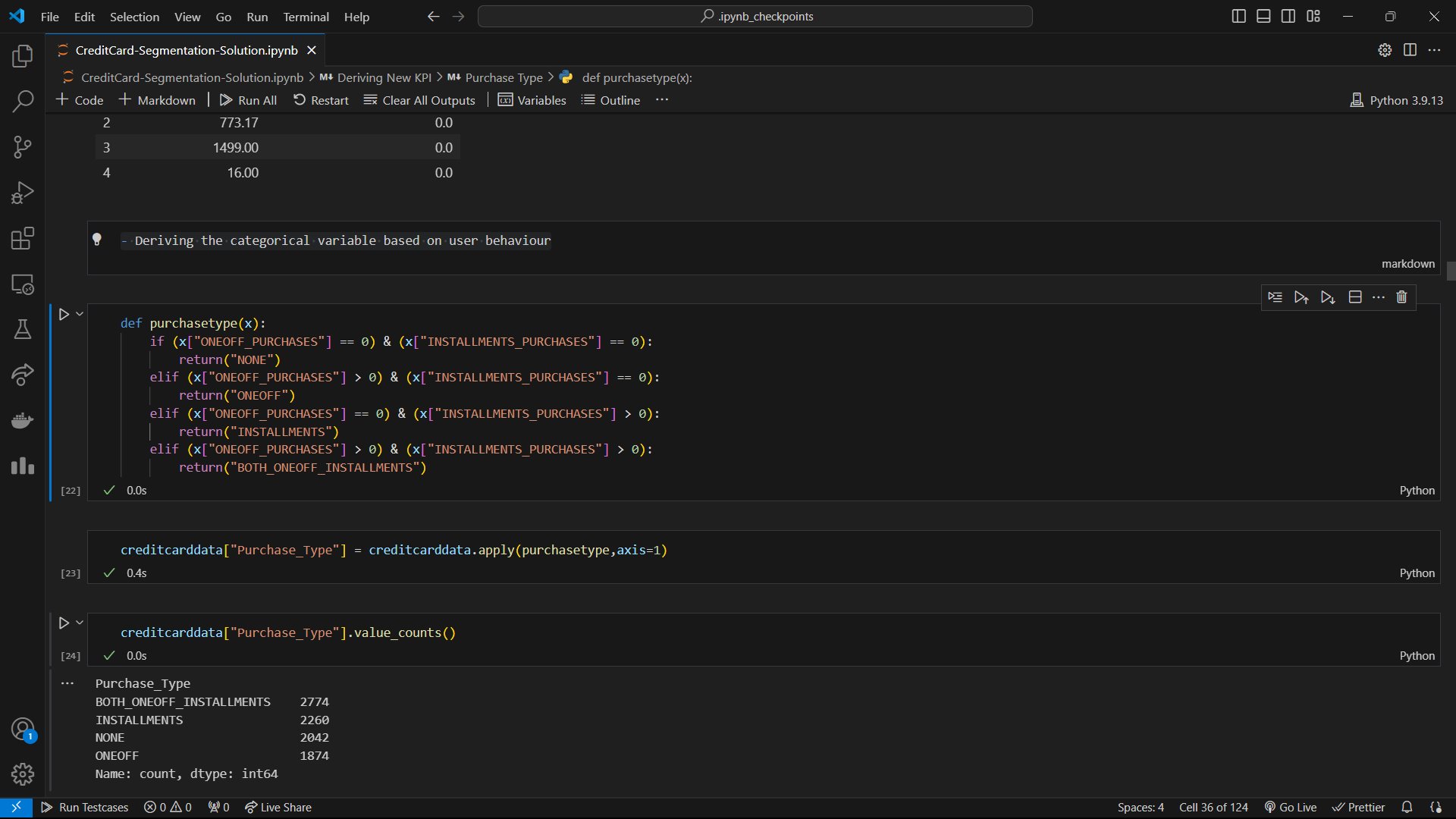
Description automatically generated

-Based on ONEOFF\_PURCHASES and INSTALLMENTS\_PURCHASES we described the user Purchase\_Type behaviour.

A screenshot of a computer

Description automatically generated

Deriving the categorical variable based on user behaviour:



## 6. Exploratory Data Analysis (EDA)

By performing EDA we gain insights into the data through visualizations and summary statistics. The describe() function provides summary statistics, helping to identify the central tendency, spread, and shape of the data’s distribution.

A screenshot of a computer

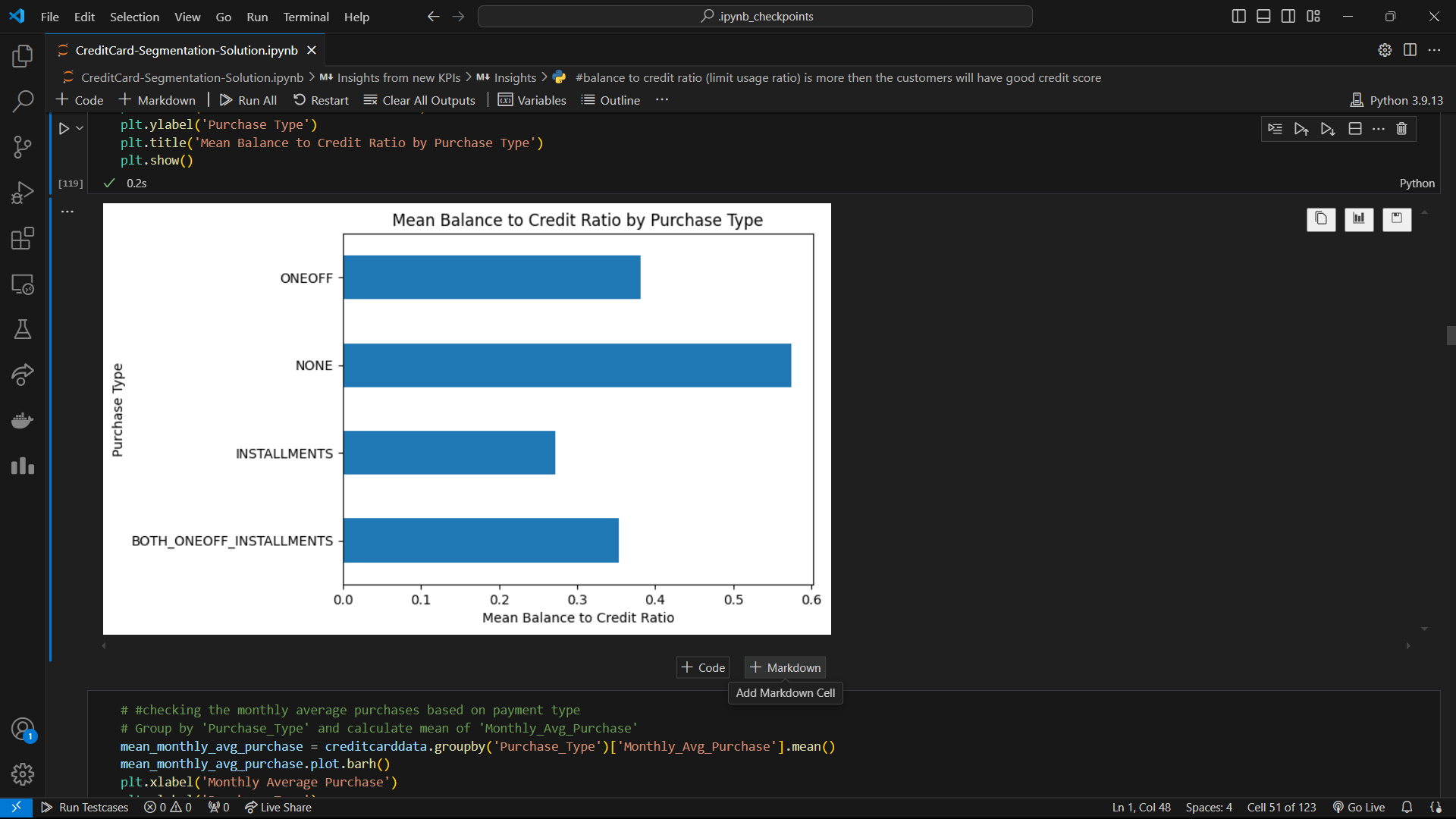
Description automatically generated

total payment ratio based on purchase type

A screenshot of a computer

Description automatically generated

Monthly cash advance based on purchase type



Mean Balance to Credit ration by Purchase Type

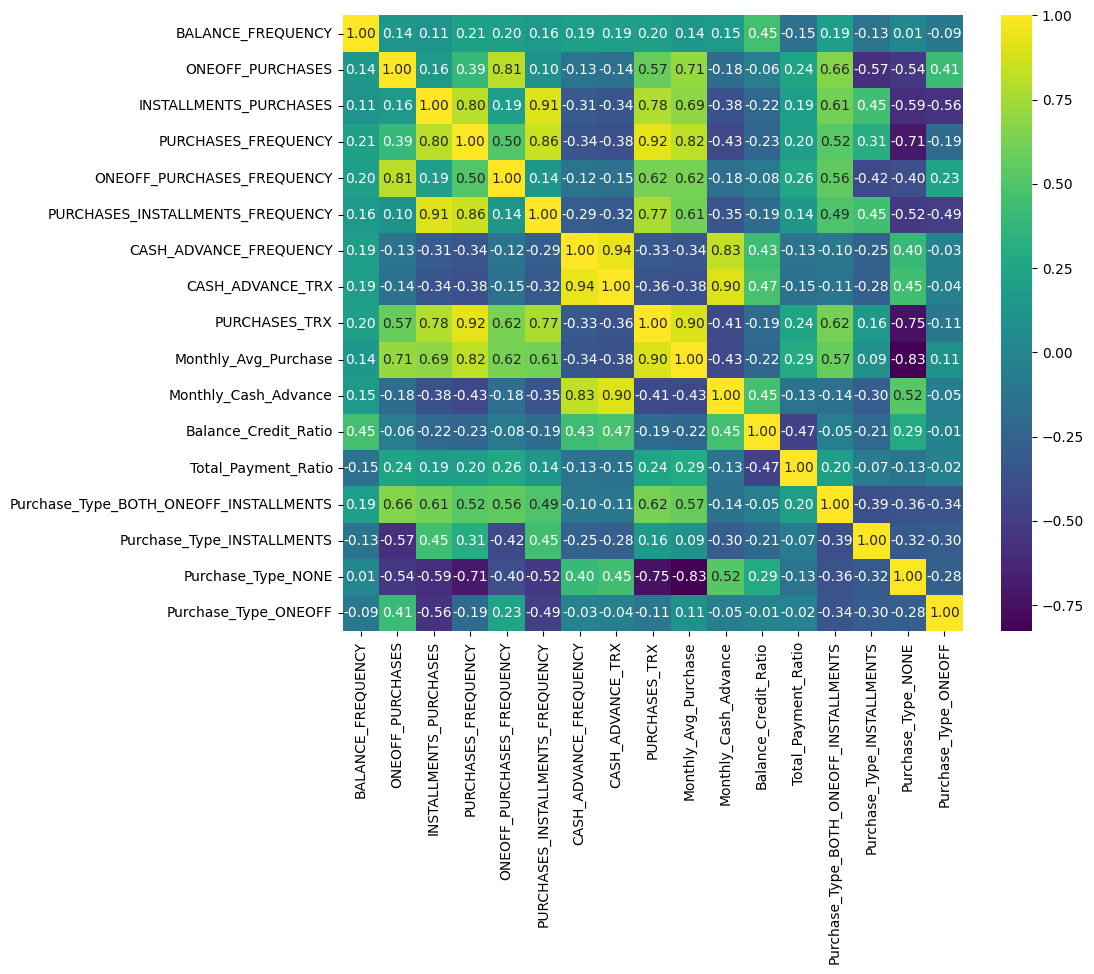
A screenshot of a computer

Description automatically generated

Monthly Average Purchase by Purchase type

## 7. Feature Engineering and Principal Component Analysis (PCA)

Reducing the dimensionality of the data while retaining as much variance as possible.



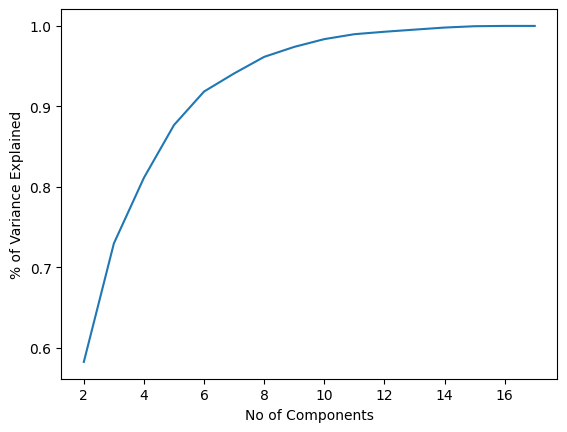
Heat map shows that many features are co-related so applying dimension reduction will help negating multi-collinearity in data. But before applying the PCA we will standardize data to avoid effect of scale on the result.

Scaling will make all features with equal weight.

**Steps:-**

**Standardization**: Before applying PCA, the data should be standardized so that each feature contributes equally to the analysis. StandardScaler() is used to standardize the data, meaning each feature will have a mean of 0 and a standard deviation of 1.

- Applying PCA: Reduce the number of features while retaining the most important information.



-Since 5 Components are explaining about 88% of the variance we select 5 components.

## 8. Clustering (Unsupervised Learning)

Segment the customers into distinct groups based on their credit card usage patterns.

- Steps Involved:

- Choosing the Number of Clusters: Determine the optimal number of clusters using the silhouette coefficient.

A graph with a line

Description automatically generated

From the SC score we can see that the number of optimal clusters are 4.

- Applying K-Means Clustering: Segment the data into clusters

- Visualizing Clusters: Plotting the clusters to visualize the segmentation.

A colorful dots on a white background

Description automatically generated

## 9. Interpretation and Business Insights

- Objective: Derive actionable insights from the clustered data.

A screenshot of a computer program

Description automatically generated

Percentage of customers from each cluster.

- Each cluster represents a different segment of customers.

A graph of different colored bars

Description automatically generated

which can be analysed to tailor marketing strategies, personalize offers, or enhance customer experience.

**Insights with 4 Clusters:**

**Cluster 0** customers have lowest balance to credit ratio (good credit score) and are paying dues also doing maximum instalment purchases. **This group is about 25% of the total customer base.**

**Cluster 1** customers are taking maximum advance cash and has high debt to credit ratio (poor credit rating) & not doing any purchase transaction. **This group is about 23% of the total customer base.**

**Cluster 2** is the group of customers who have the highest Monthly Average Purchases and are not doing instalment or one-off purchases and are only taking cash advances. **This group is about 31% of the total customer base.**

**Cluster 3** customers are doing maximum one-off payments and least total payment ratio. **This group is about 21% of the total customer base.**

## 10. Recommendations:

**Cluster 0:** Offer them loyalty points and special discounts on one-off purchases. They could also be targeted with reminders for instalment payments and given additional credit card offers to enhance their purchasing power.

**Cluster 1:** Implement a payment reminder system and offer incentives for early or full payments to encourage better financial behaviour. This group should be monitored closely to prevent further debt accumulation.

**Cluster 2:** These are ideal customers, so they should be offered loyalty points and exclusive offers. Offering them add-on credit cards can boost their spending further. Regular engagement with personalized offers can help maintain their loyalty.

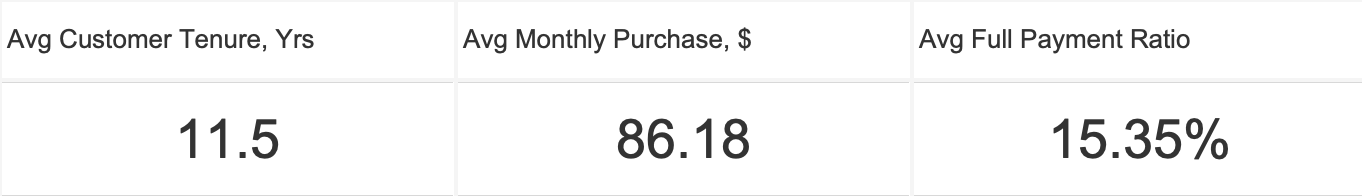
**Cluster 3:** Encourage them to explore instalment plans by offering special deals on EMI purchases. Providing discounts on full payments for future transactions could also incentivize them to pay more than the minimum due, reducing their debt and improving their credit standing.

## Tableau Dashboard showing the clusters behaviour:

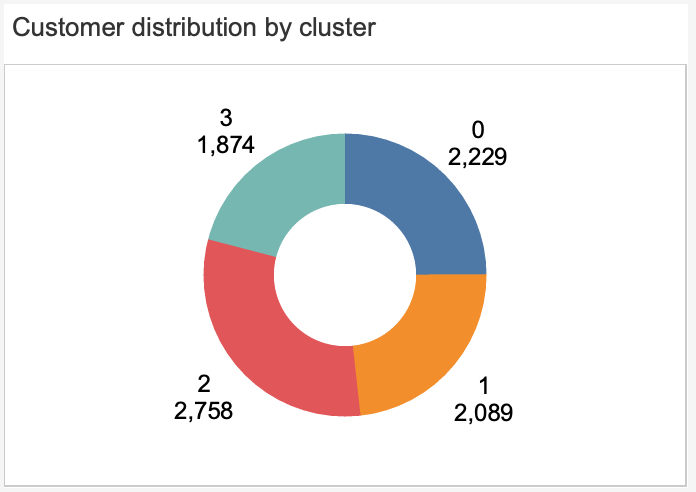
A Tableau dashboard has also been prepared showing the following metrics:

1. Exploratory Data Analysis: Few high level metrics explaining the data has been presented at the top of the Tableau Dashboard (as shown in the image below).

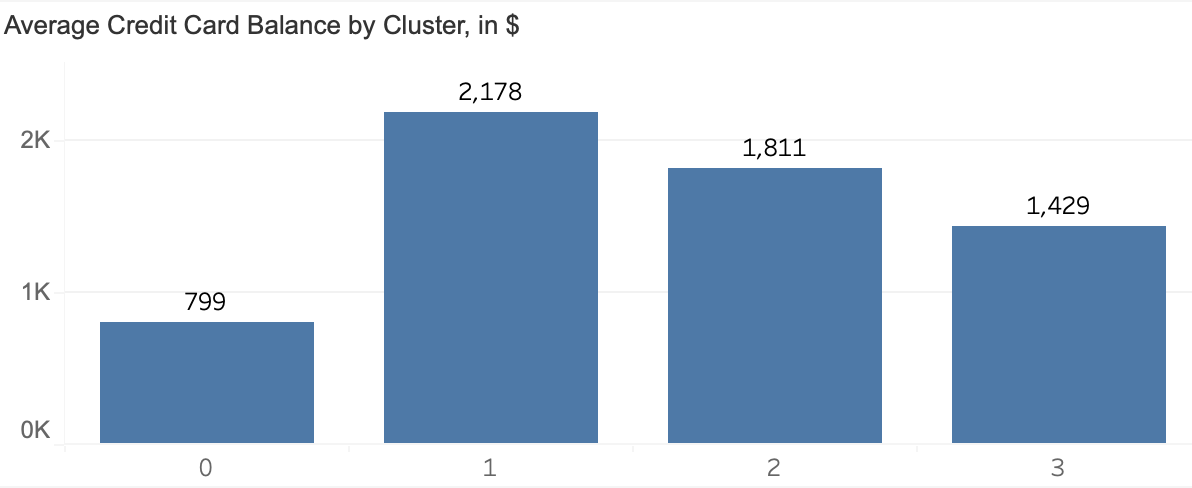




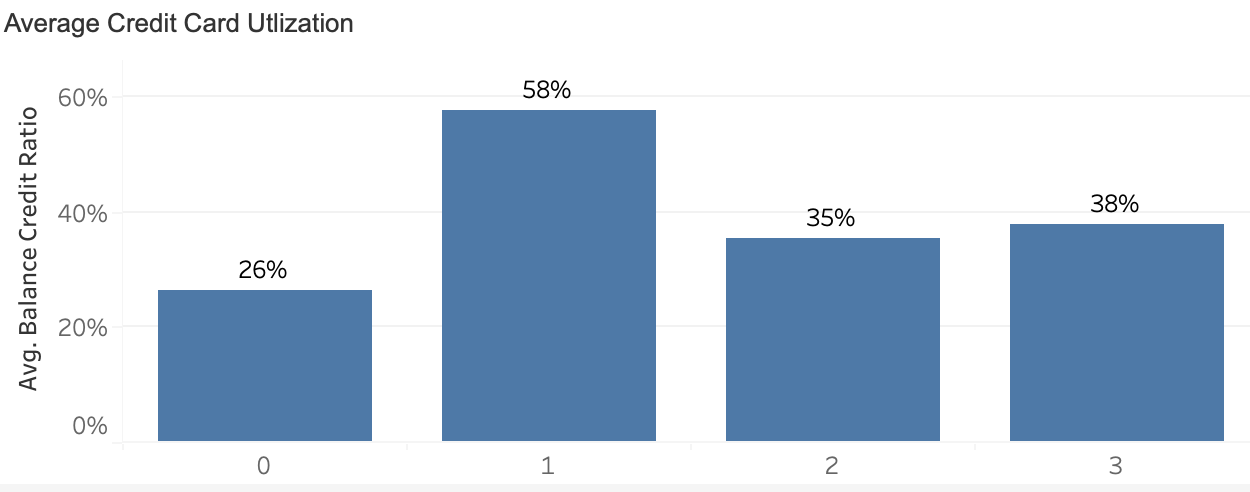
1. Clusters Behaviour: Few customer behaviour related metrics for each of the clusters has been presented on the Tableau Dashboard throwing more light on the clusters. The following are the snapshots of the metrices covered:
   1. Customer Distribution by cluster:



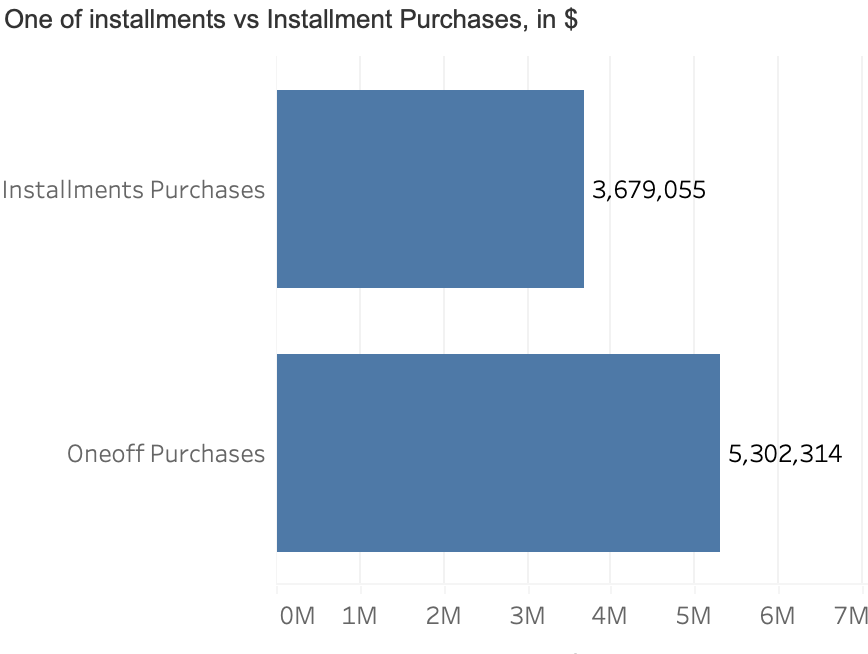
* 1. Avg. Credit card balance by clusters (in $)



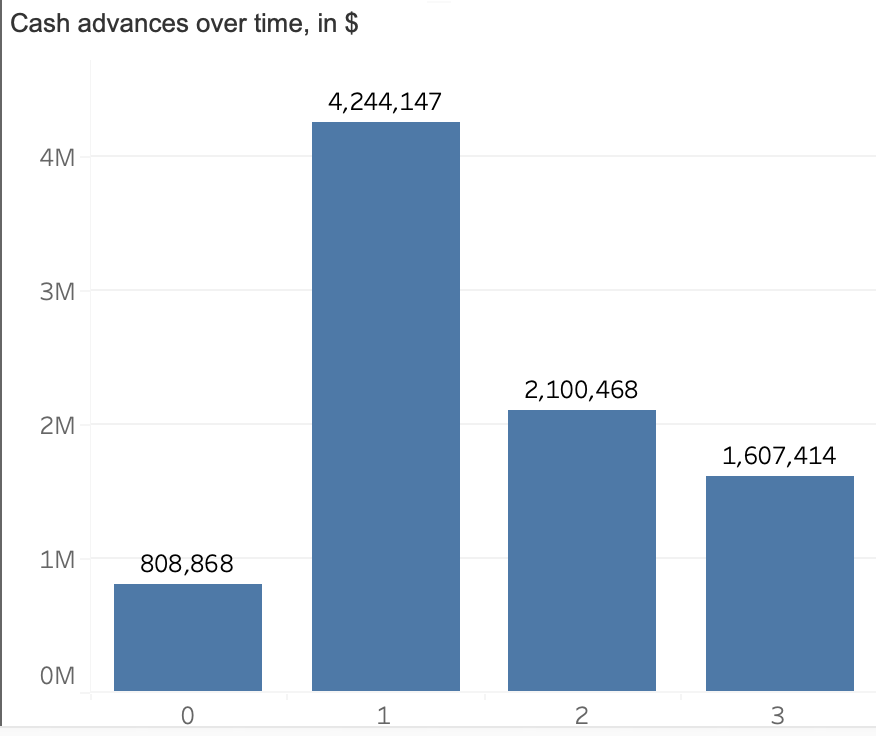
* 1. Avg. credit card utilization



* 1. Spending Behaviour



* 1. Cash Advances over time by clusters



* 1. Payment Behaviour

