Customer Segmentation for Marketing Strategies

Introduction/Objective

The main objective of this analysis is to identify patterns or clusters that can provide insights into customer behavior or credit card usage patterns to aid with marketing strategies.

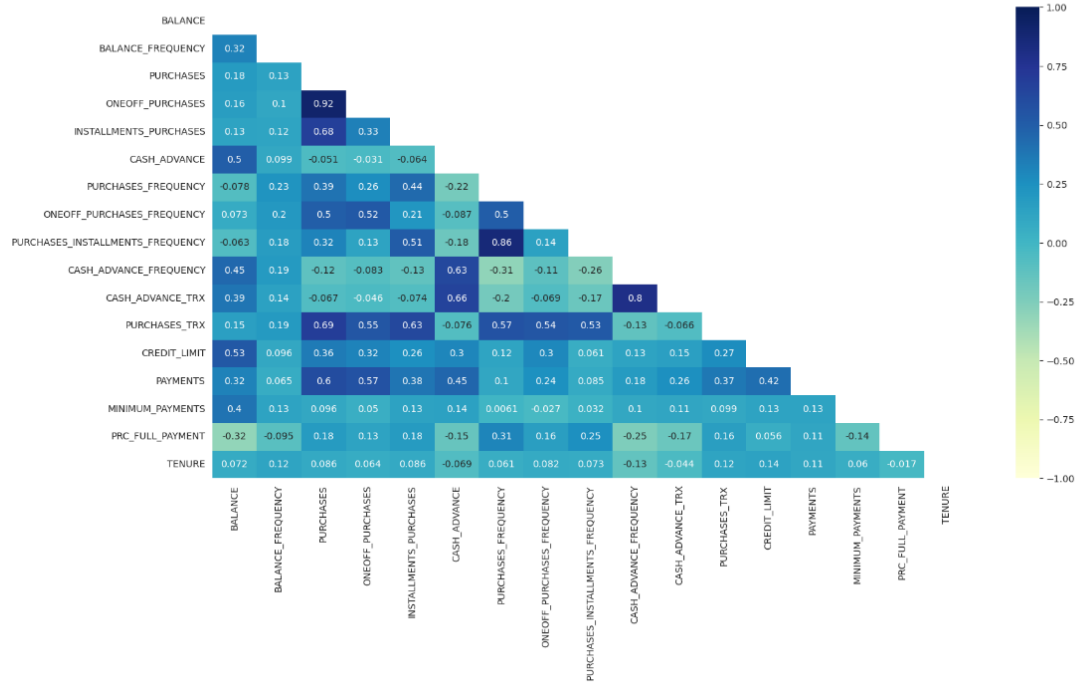
Background

The chosen dataset is the credit card data from Kaggle(Credit Card Dataset for Clustering | Kaggle), which includes information on the usage behavior of about 9000 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioral variables. The variables include behaviors like balance, balance frequency, cash advance, purchase installment frequency, credit limit, minimum payments, tenure, and more.

EDA

The data exploration phase involved analyzing the distribution of variables, checking for missing values, and scaling data. For data visualization, I included a histogram for each of the features, as well as a heat map to see the correlation between variables. Data cleaning and preprocessing steps included removing duplicate entries, handling missing data, and scaling data.

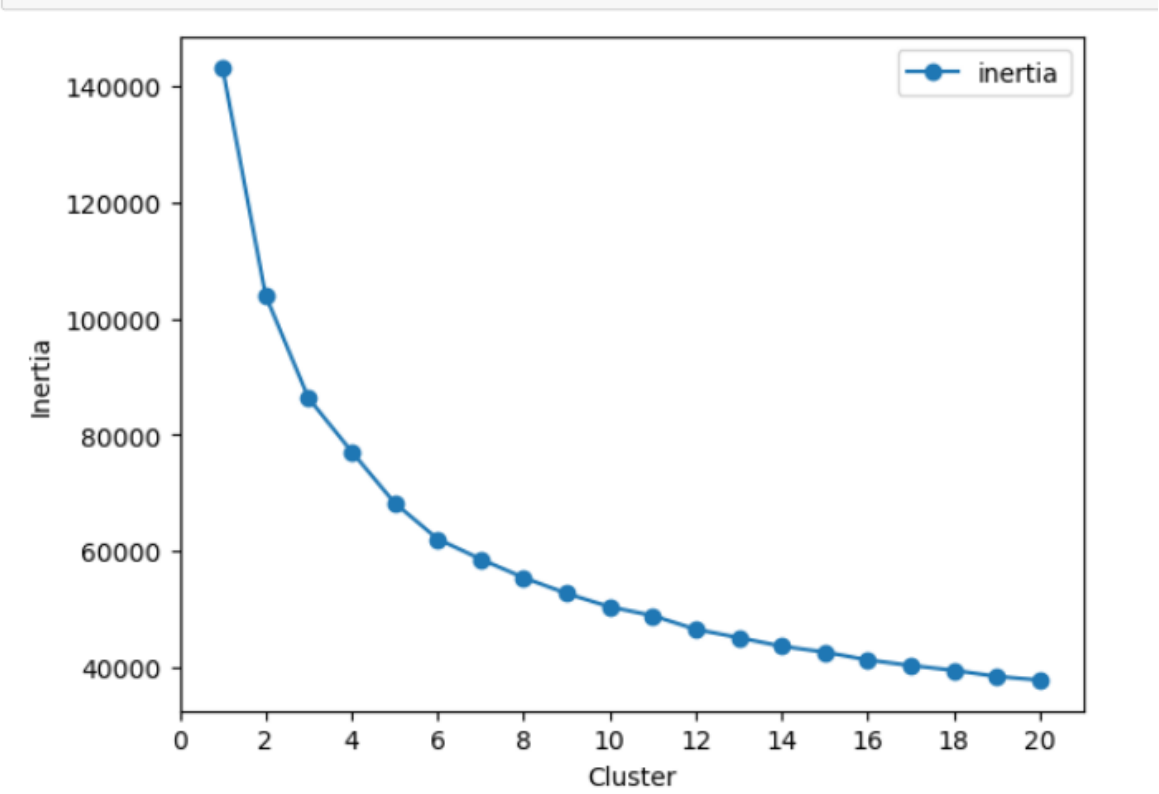
```
mask[np.triu_indices_from(mask)] = True

with sns.axes_style("white"):
    f, ax = plt.subplots(figsize=(18, 9))
    ax = sns.heatmap(corr , mask=mask, vmin = -1,vmax=1, annot = True, cmap="YlGnBu")
```



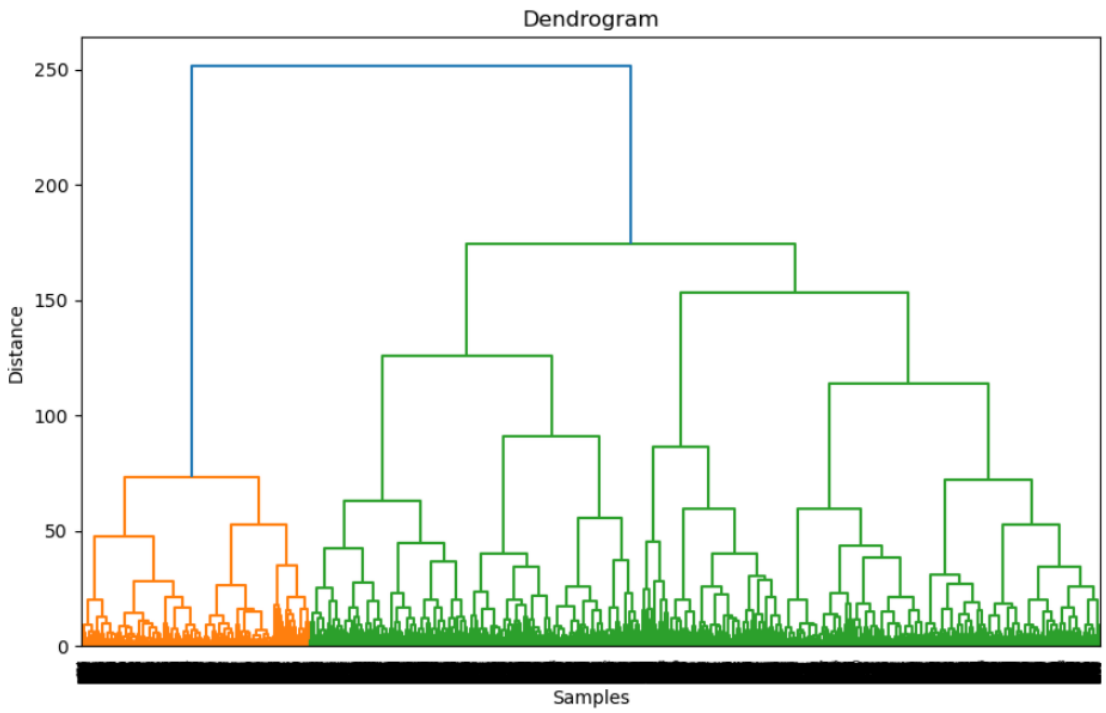
Models

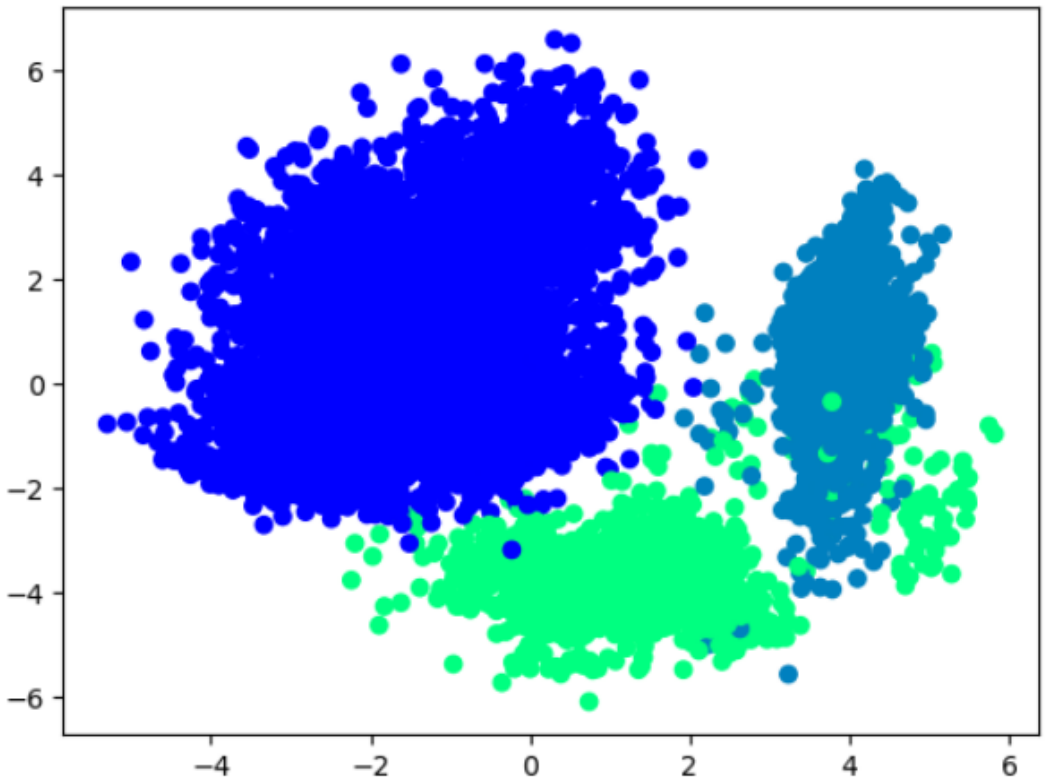
Three models were trained and evaluated: K-Means, Hierarchical Clustering, and DBSCAN. The models were trained using the same training and test splits, and performance metrics such as the silhouette score were calculated to assess their predictive capabilities. PCA Dimensionality Reduction was performed in order to help identify patterns in data based on the correlation between features. PCA helps us find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with fewer dimensions than the original one.

K-Means (Elbow method shows optimal clusters is between 4 and 7)



Hierarchal Clustering


Dendrogram
Distance
Samples
250
200
150
100
50
0


6
4
2
0
−2
−4
−6
−4
−2
0
2
4
6

```
Silhouette Score:
K-means Silhouette Score: 0.18731850642061654
Hierarchical Silhouette Score: 0.1838801569707745
DBScan Silhouette Score: -0.24781665582874784
```

Key Findings:

The key findings from the analysis of the Credit Card Dataset for Clustering include that the K-Means algorithm had the best silhouette score meaning that the clusters were the most clearly distinguished. Hierarchal Clustering had a similar silhouette score. DBSCAN require more time to find the optimal number of clusters since you have to keep changing the expected number of clusters. These findings can help credit card companies make informed decisions about customer segmentation, personalized marketing campaigns, credit limit adjustments, or product offerings tailored to specific customer groups.

Flaws and Future Plan:

Possible flaws in the analysis could arise from feature selection and model sensitivity. The chosen features may not capture all relevant aspects of credit card usage or customer behavior. Different clustering algorithms produce different results, and the choice of model parameters could impact the clustering outcome. To address these flaws, the analysis could be revisited with additional data sources, such as transaction history or external demographic data, to enrich the understanding of customer behavior. Exploring alternative modeling techniques and evaluating the robustness of the results through cross-validation can also help improve the accuracy and reliability of the findings.