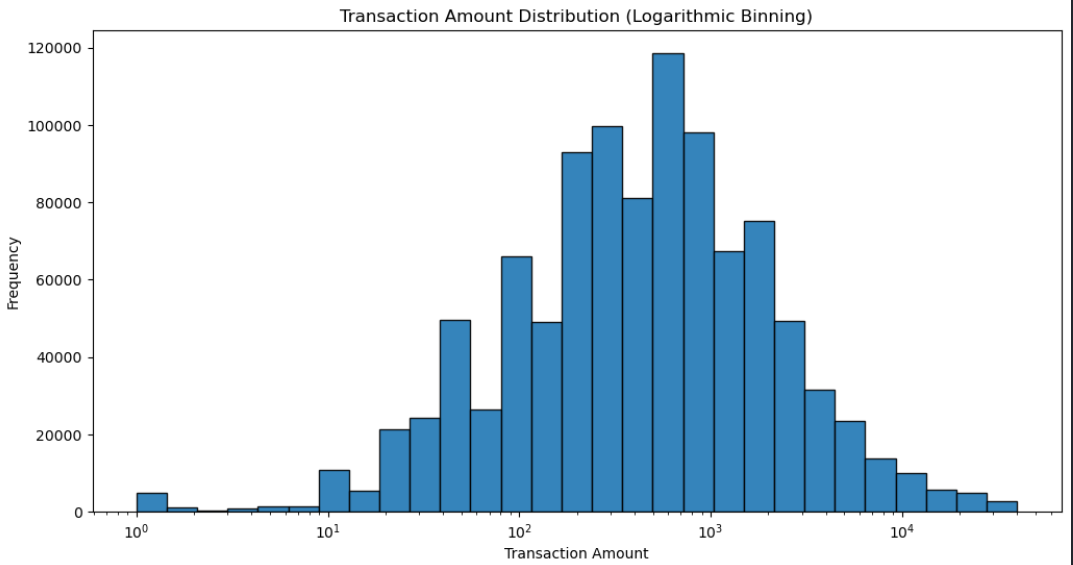
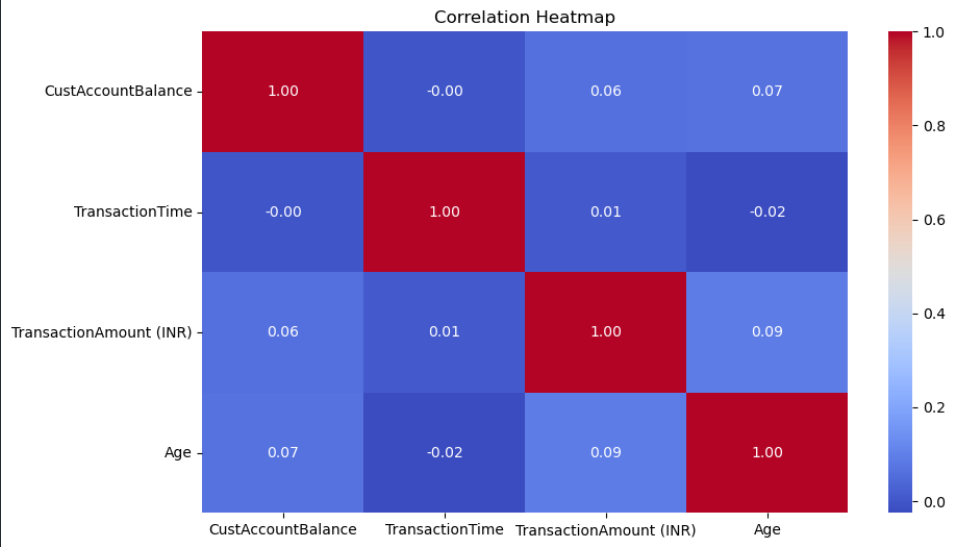
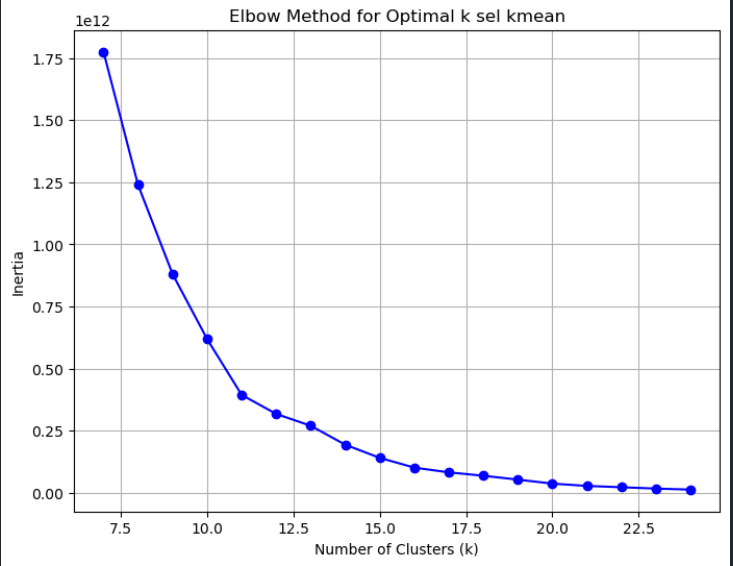
In this task, our goal was to explore clustering techniques for customer segmentation of a bank. Our data was from an Indian bank, which included just over 1 million samples where each row consisted of customer info such as gender, date of birth, location, customer id, customer account balance and transaction info like transaction date, transaction time and transaction amount in INR. We chose this data from Kaggle.com and the reason in choosing this specific dataset was that it had high number of examples so It was a bit challenging working with this dataset.

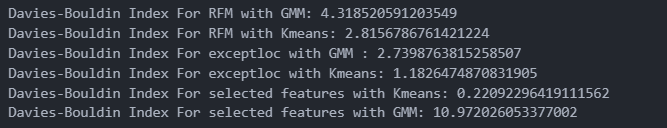
Our first step was data preprocessing. In this step, we first deleted rows that contained null values, and then we transformed our date of birth column to age. Next, we did a quick exploratory data analysis (EDA). We looked into age distribution, transaction amount distribution and correlation heat map between features. After EDA, we preformed feature engineering. Using columns transaction date and transaction amount we created three features: recency, frequency and monetary. These three features is widely used in in business and marketing analysis with the name of RFM. It is a behavioral targeting method often used to divide and score customers such as high value or low value customers. Then via the help of transaction time column, we determined in what time of day the transaction was made and named this new feature time of day. In addition, we binned our age values into four categories. Finally, we did one hot encoding for our categorical values age group, customer gender and tome of day. We skipped customer location in one hot encoding because there was too many distinct values here and it increased dimensionality significantly so we used frequency encoding instead.



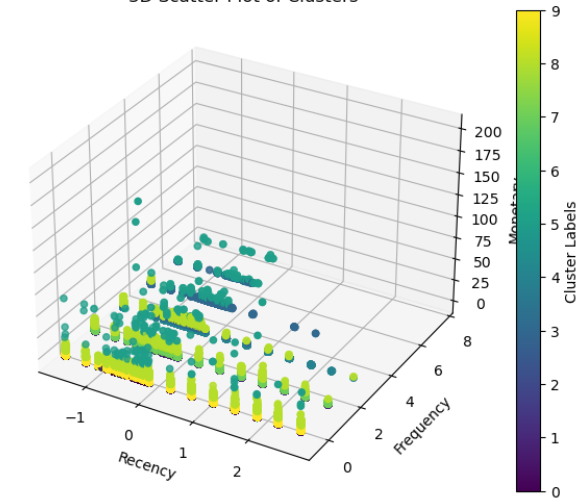
For training, we tried different clustering methods with different types of input data. Several methods were tried but because our data was, too large we could not use DBSCAN, Hierarchical clustering or Affinity propagation. Methods that we used was K-means and Gaussian Mixture Models (GMM). We trained these models with different input data so that we can figure out which was the best. Our data inputs was RFM, Scaled RFM, Selected Features that consisted of one hot encoded age gender and time of day , scaled RFM and scaled customer account balance and frequency encoded customer location, the last input type was selected features without location. We trained each of these features on our models K-means and GMM. In K-means used elbow method to inertia values in order to determine optimal value for number of clusters.

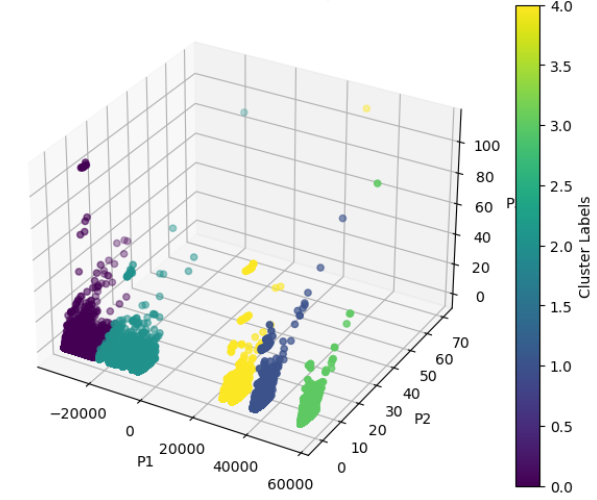


For evaluating our models and scoring them, we used two approaches, first Davies-Bouldin Index that is based on that clusters should have low within-cluster variation and high between-cluster separation. For this index, the lower is the value the better clustering there is. So best one according to this index was selected features with k-means method.



Moreover, the second one was to evaluate clusters via 3D plotting where three dimensions was the highest impactful values and the colors was based on cluster labels. Below you in the first picture, you can find visualization for scaled RFM with K-means and in the second one, you can find our best model selected features with K-means.





In conclusion, we saw that K-means was a better fit for our data than GMM. In addition, as input features selected features that consisted of several types of inputs like financial and personal gave better results.