***CAPSTONE 3 PROJECT REPORT***

**Intro and Problem Statement:**

* Customer segmentation is the practice of dividing a company’s customers into groups that reflect similarity among customers in each group.
* The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business.
* Customer segmentation has the potential to allow marketers to address each customer in the most effective way.
* One approach to customer segmentation is leveraging machine learning algorithms to discover new segments.

**Data fields:**

order\_status

order\_purchase\_timestamp

order\_approved\_at

order\_delivered\_carrier\_date

order\_delivered\_customer\_date

order\_estimated\_delivery\_date

customer\_zip\_code\_prefix

customer\_city

customer\_state

shipping\_limit\_date

price

freight\_value

payment\_sequential

payment\_type

payment\_installments

payment\_value

product\_name\_lenght

product\_description\_lenght

product\_photos\_qty

product\_weight\_g

product\_length\_cm

product\_height\_cm

product\_width\_cm

seller\_zip\_code\_prefix

seller\_city

seller\_state

product\_category\_name\_english

**Exploratory Data Analysis:**

* The product and orders datasets have NaNs in them. While the other datasets(customers, items, payments, sellers and categories) don’t
* A total of 3K sellers are serving the 95K customers present in the dataset
* 98K orders were placed to purchase 33K different products
* Majority of the customers are from Sao Paulo state
* While very few from Roraima state

Chart

Description automatically generated

* Most of the orders were placed from the bed & bath category
* While very few orders were from Automobiles

Chart, bar chart

Description automatically generated

* Most of the orders are placed stating from 8 AM till 11 PM. While 4 PM being the most active hour

Chart, histogram

Description automatically generated

**Data Preprocessing:**

* The customers, items, payments, sellers and categories datasets are merged
* The rows with NaNs are dropped from the dataset
* Timestamp columns i.e.,‘order\_approved\_at','order\_delivered\_carrier\_date','order\_delivered\_customer\_date','order\_estimated\_delivery\_date','shipping\_limit\_date’ are split into year, month, day, hour, is\_weekday or is\_weekend
* Customer\_city and seller\_city columns are dropped since the dimensionality will increase
* Numerical columns are standardized
* Categorical columns are one-hot encoded and merged with standardized numerical columns

**Modeling:**

**K-Means**

* After the data was one-hot encoded and scaled, K-Means clustering algorithm was applied, and the optimal number of clusters were found to be 5 based on the elbow method.

**Chart, line chart

Description automatically generated**

**DBScan**

* DBScan clustering algorithm was applied after PCA with epsilon value of 1 and 9 as minimum number of samples
* The silhouette coefficient is 0.457 with 117 records classified as noise

**Chart, scatter chart

Description automatically generated**