

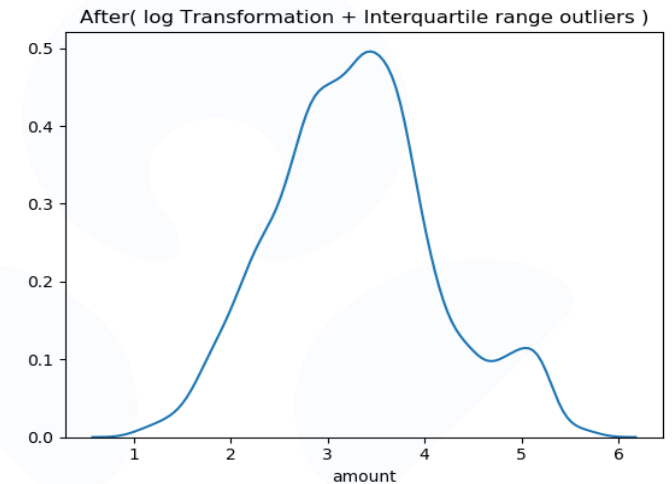
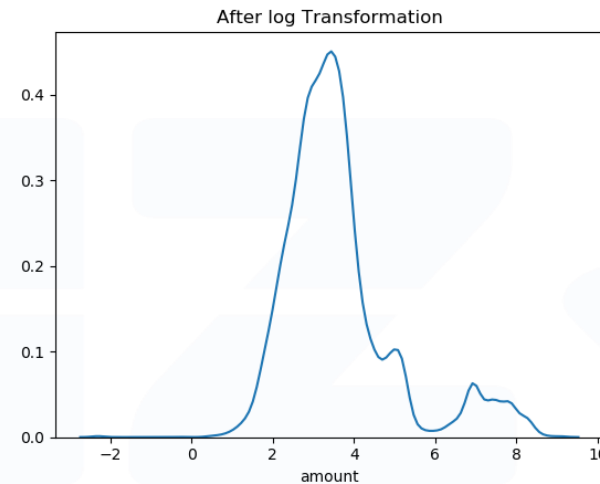
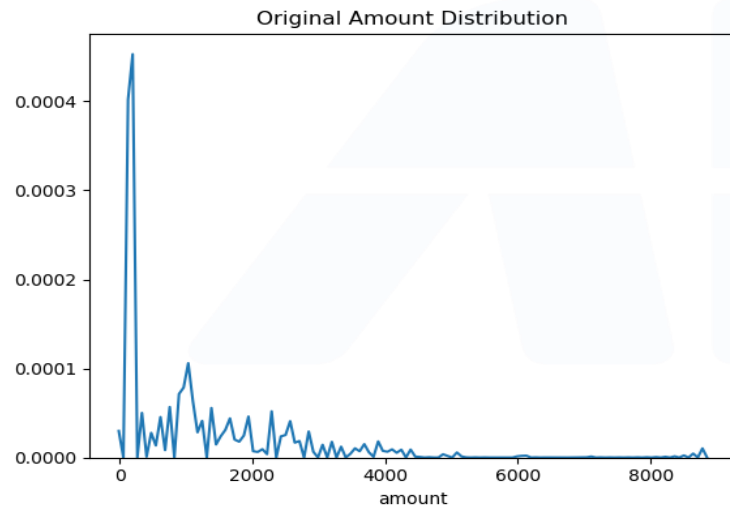
1.DATA QUALITY ASSESSMENT

1.Location Column of Dataset contains 577 as longitude is not a valid Value.

2.All of 4326 missing values cause there is no merchant involved, So replacing all missing values with 'non-merchant payment' would be better and other columns ['merchant_code','bpay_biller_code'] have more than 90 percent null values so it would be better if we drop these columns.

3.Imputing Null value of a row from its transaction_description would yield better result.

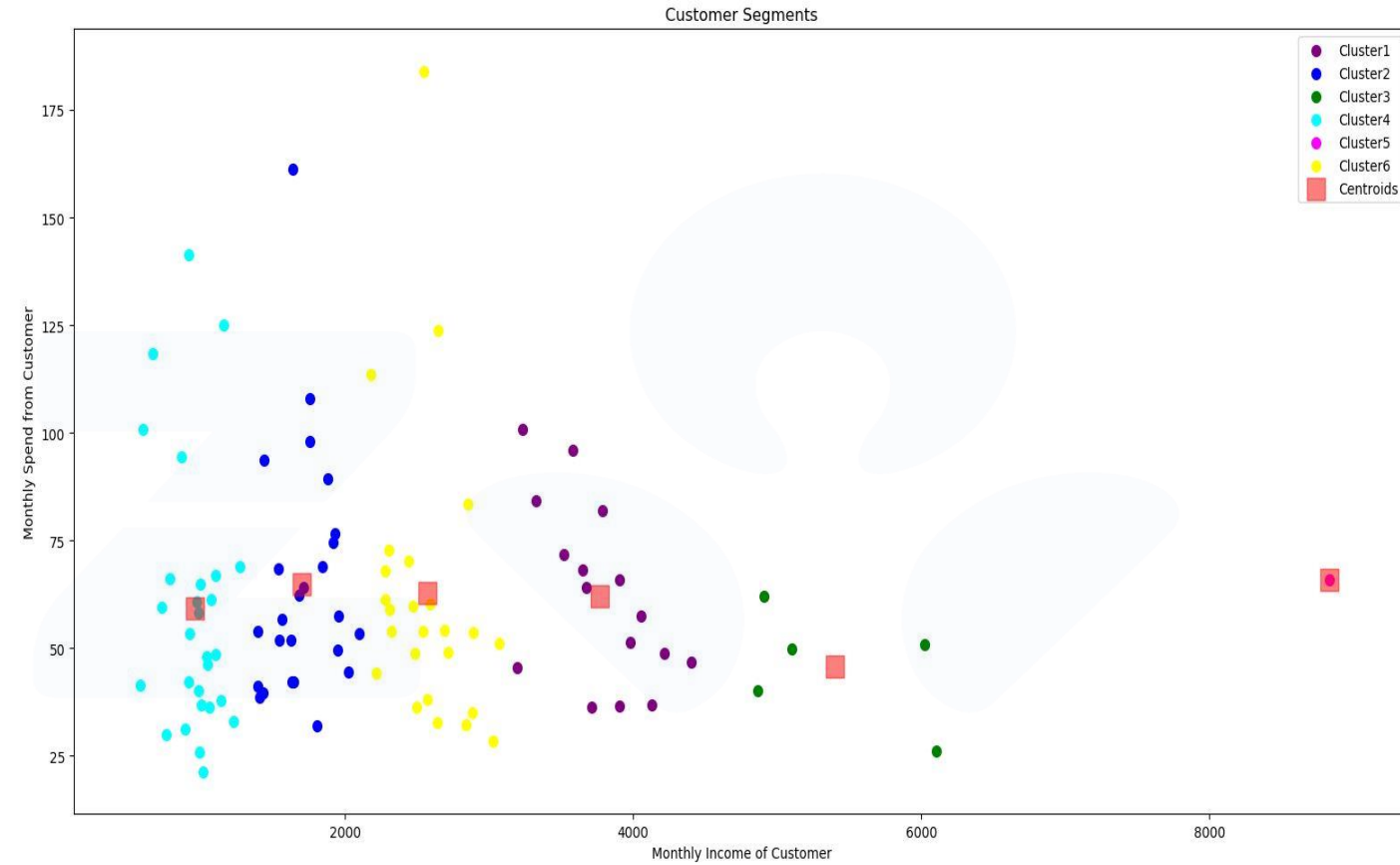
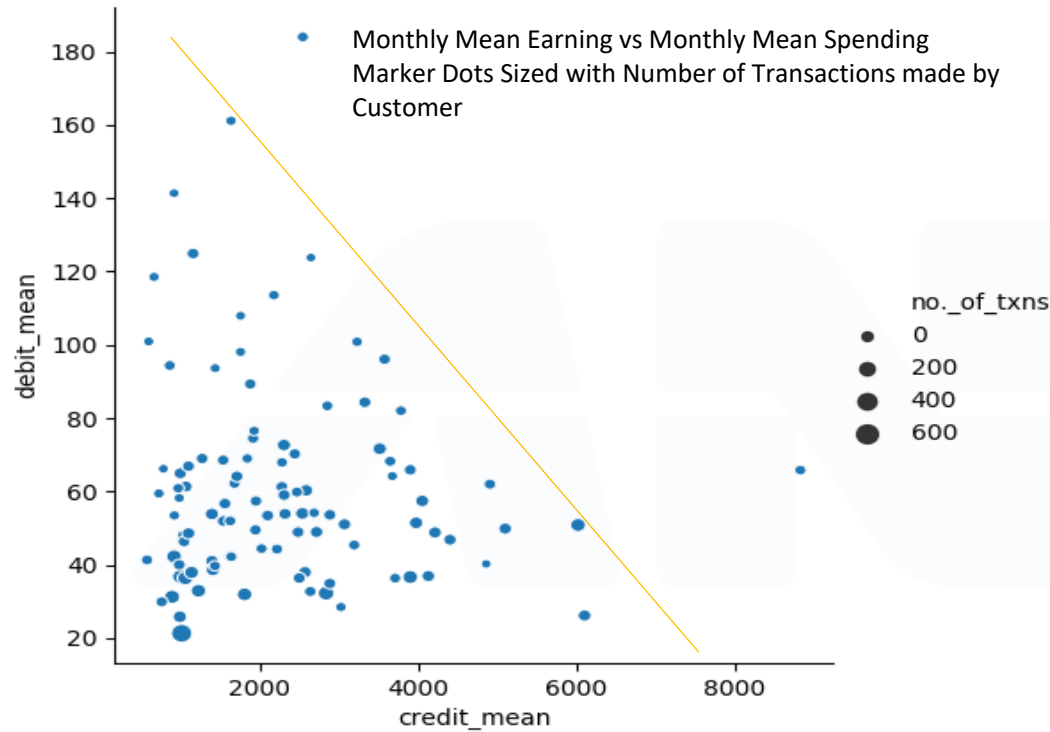
2.DEALING WITH OUTLIERS



Amount column is distributed very badly for our analysis, after log transformation extreme values tend towards mean.

For Filtering out outliers Inter Quantile Deviation is used instead of z-score as Z score is sensitive towards outliers

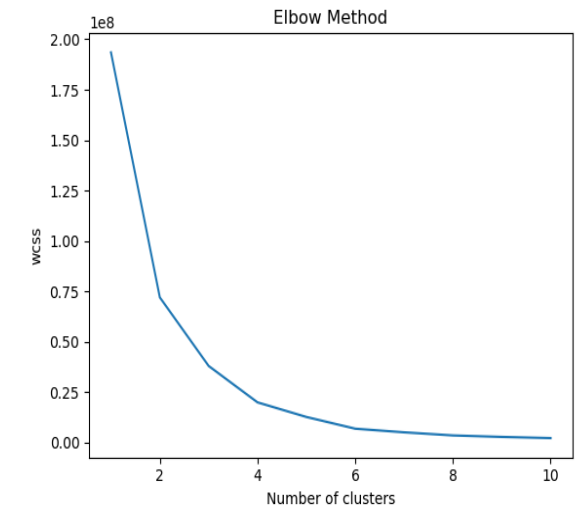
3.CUSTOMER SEGMENTATION



- More number of transactions are made by customers in the plot who are closer to the Origin and customers who are far away from origin made less number of transaction
- From Elbow Method optimal clusters for customer segmentation are 6.
- Customer Spending is very low compared to their Earnings
- Cluster 4,5 and 6 are more valued customers as they are Piling up the bank reserves.

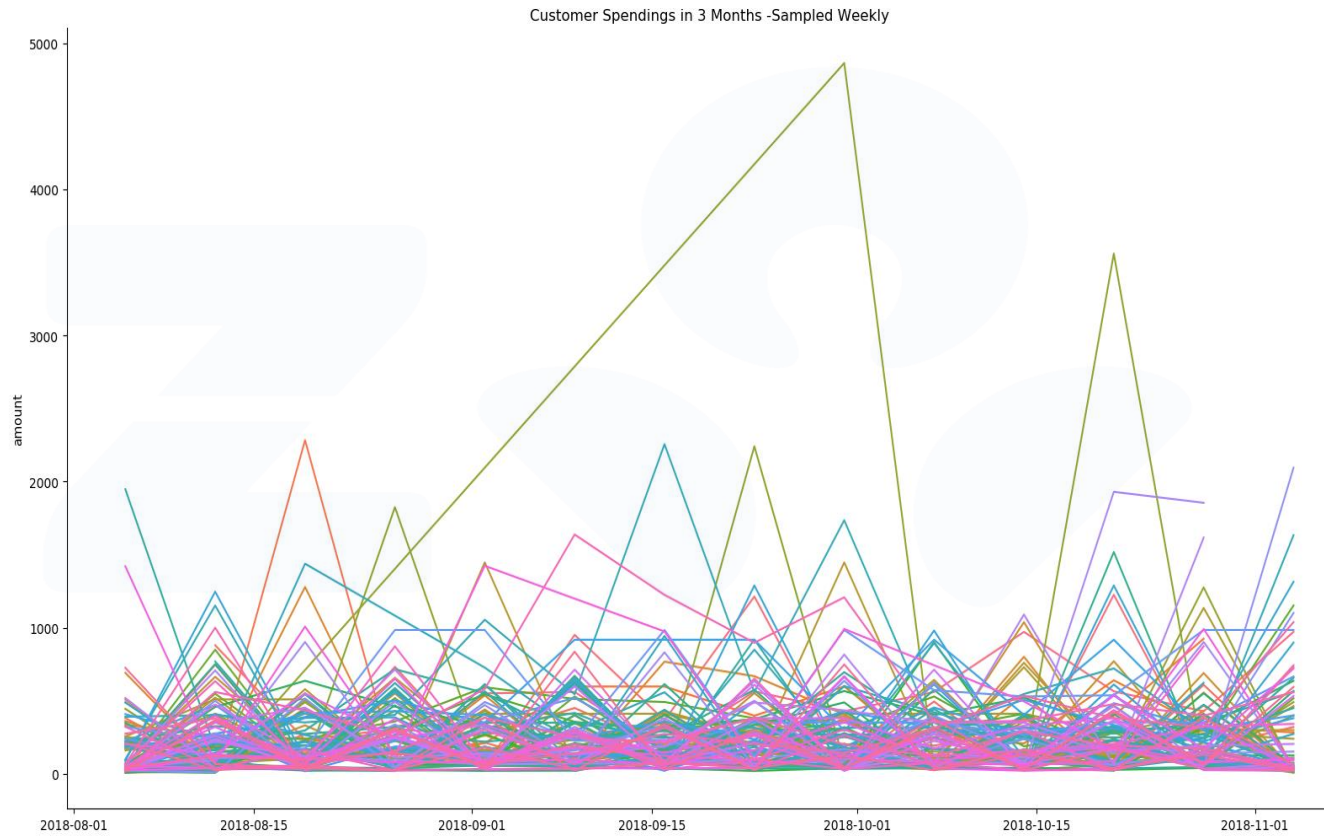
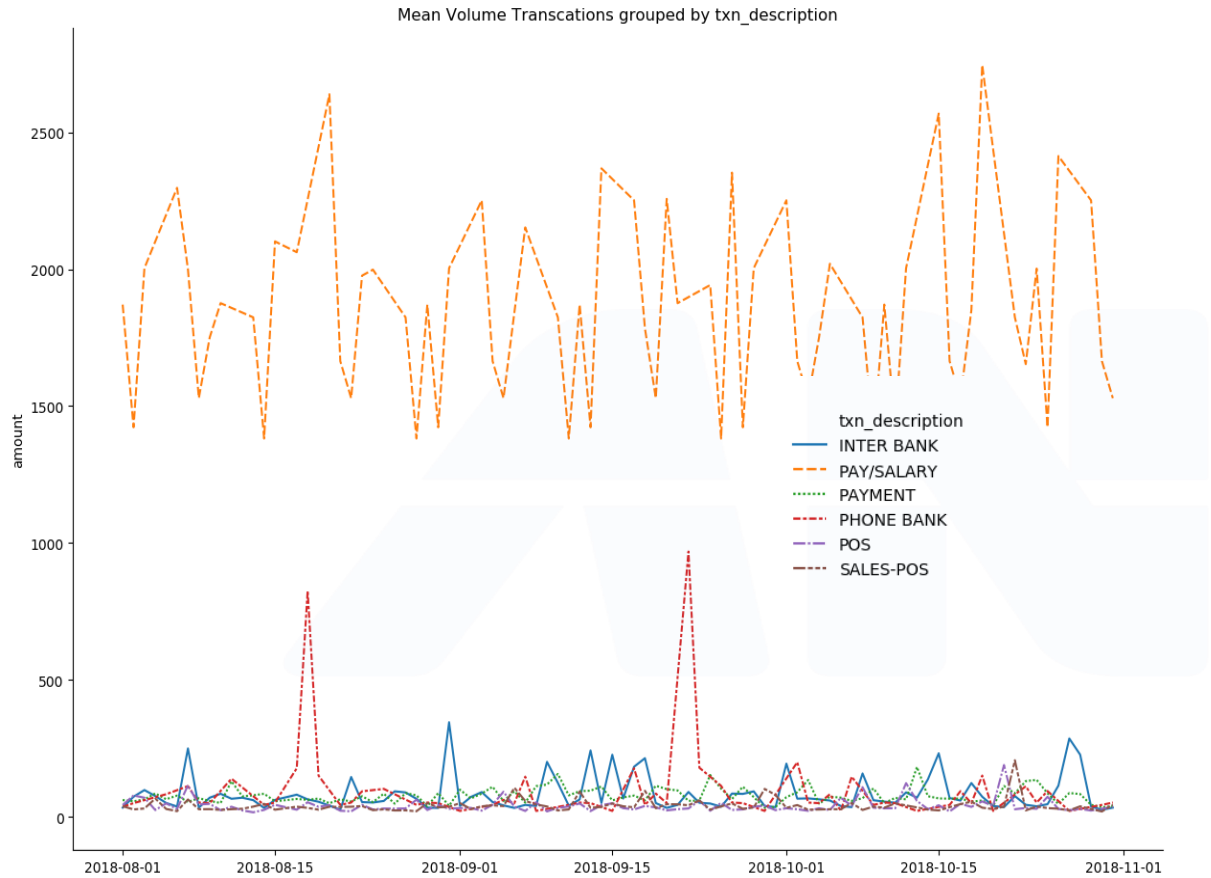
OUTLIERS USING SCATTER PLOT

- The Orange intercept plotted is the upper bound which is other way to remove Outliers from Visual Scatter Plot
- 'Monthly Spend Customers' who are greater than 150 can be consider as OUTLIERS



4.TRANSACTION VOLUME OVER TIME

Colored by Customers



- 1. Pay/Salary has highest transaction volumes and Phone bank has highest 2nd highest Transaction volume.
- 2. Transaction volumes follow cyclic order.
- 3. Customers spending habit follows sinusoidal order which can interpreted

by
DINESH Y
IIT DHARWAD