Customer Payment Analysis Report

This project examines the customer EMI Payments based on the following features

Features:

order_id : String

customer_id : String

merchant_id : String

order_amount : Decimal

checkout_started_at : Datetime

credit_decision_started_at : Datetime

approved_for_installments: Boolean

customer_credit_score: Integer

customer_age : Integer

customer_billing_zip : String

customer_shipping_zip : String

paid_installment_1 : Boolean

paid_installment_2 : Boolean

paid_installment_3 : Boolean

paid_installment_4 : Boolean

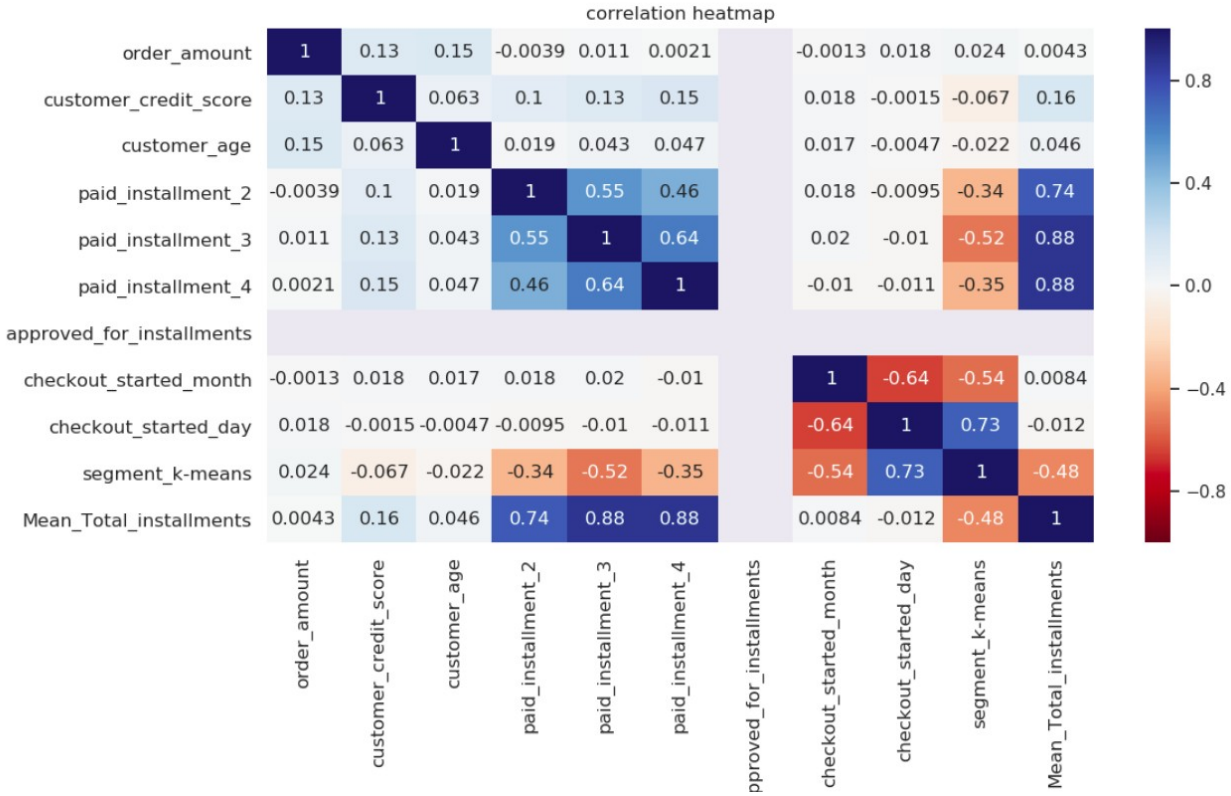
Approach:

- Python is used for data processing and visualization
- K-Means clustering technique is used for segmenting the custmoers based on the input features to find out the defaulters.
- A decision tree classification model is developed to classify the customers.

Answers from the Analysis :

1)Which features show strong correlation with a customer’s likelihood of paying back installments?

From the below graph the features such as, **Paid_installment_2, Paid_installment_3, Paid_installment_4, checkout_started_month, checkout_started_day and customer_credit score** are correlated mostly.



2) Which features should be discarded? Why?

- **Paid_installment_1** and **approved_for_installments** can be removed as they are always 1 in all cases
- **customer_id, merchent_id and order_id** can be removed as there are no insights we can get from the ID.
- **customer_billing_zip,customer_shipping_zip** can be removed as there are too many location details which can be analysed based on the geographic charts and convert into categorical variables.

3) What surprised you about the results/trends observed in the data?

The answer is given in the below preprocessing and visualization steps.

Data Preprocessing

1)Missing value treatment:

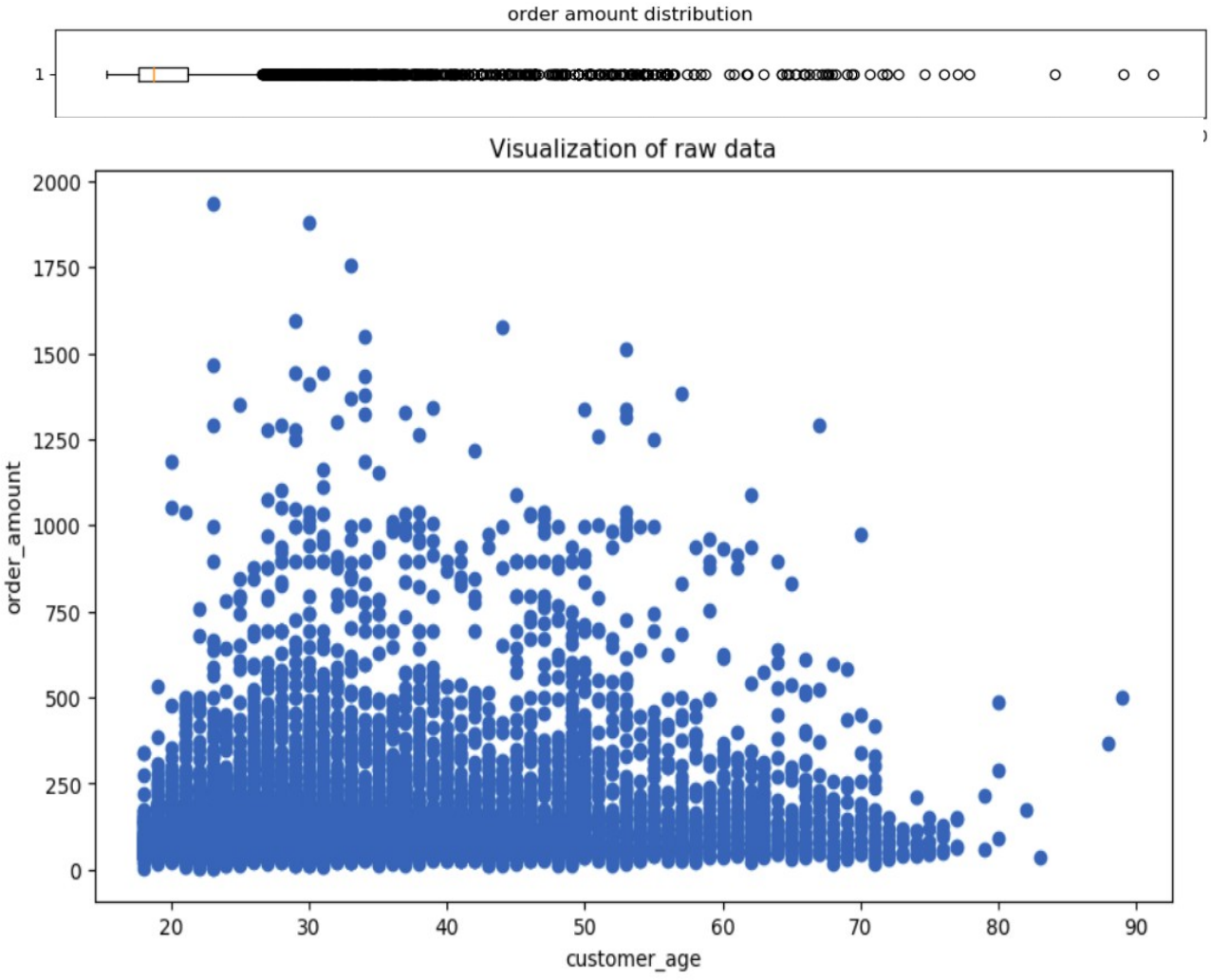
The columns which are having missing values are,

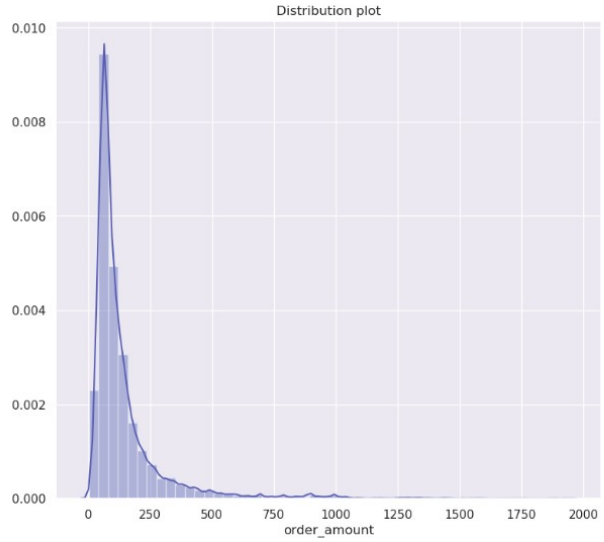
```
column  customer_billing_zip having  0.001 % missing values
column  customer_shipping_zip having  0.002 % missing values
column  paid_installment_1 having  0.0 % missing values
column  paid_installment_3 having  0.0 % missing values
column  paid_installment_4 having  0.058 % missing values
```

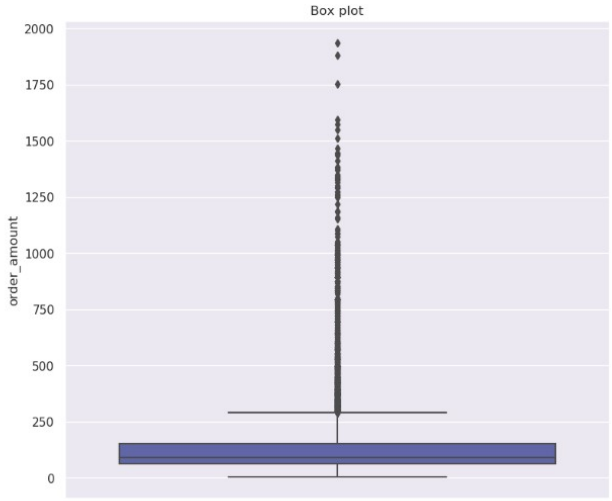
As this dataset is having missing values less than 5%, they are dropped.

2) Distribution of the features:

i) order amount distribution

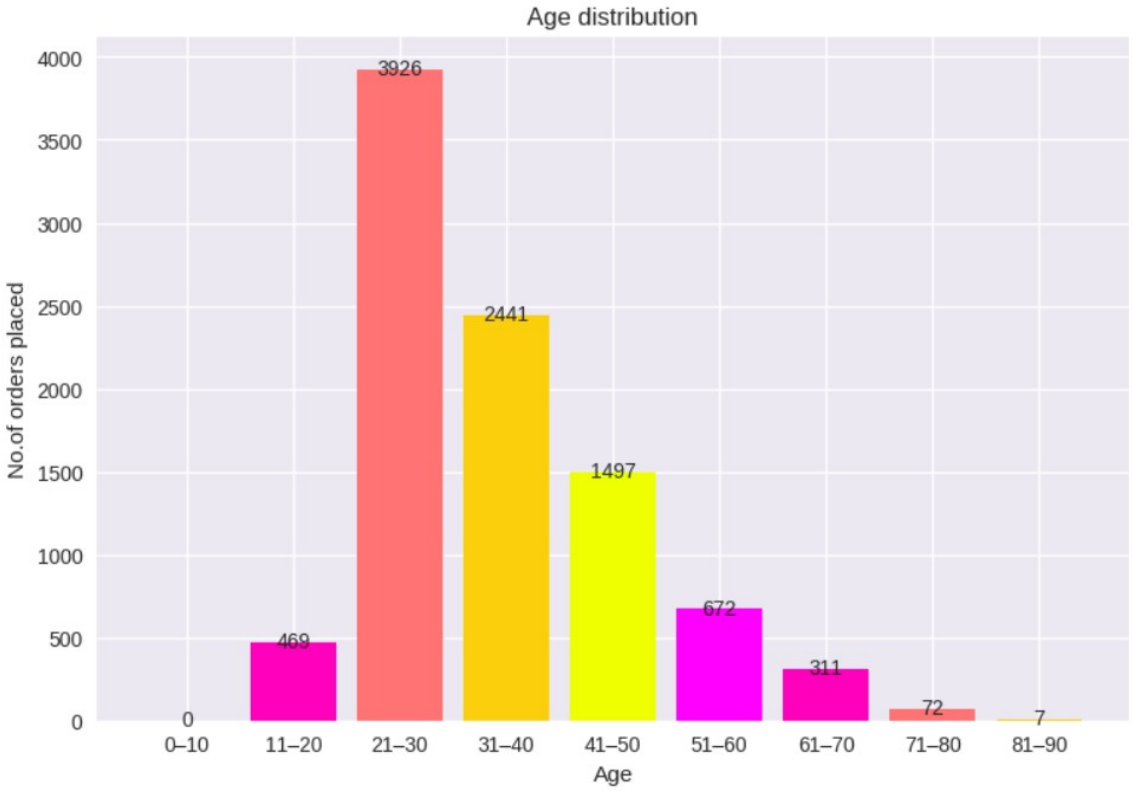


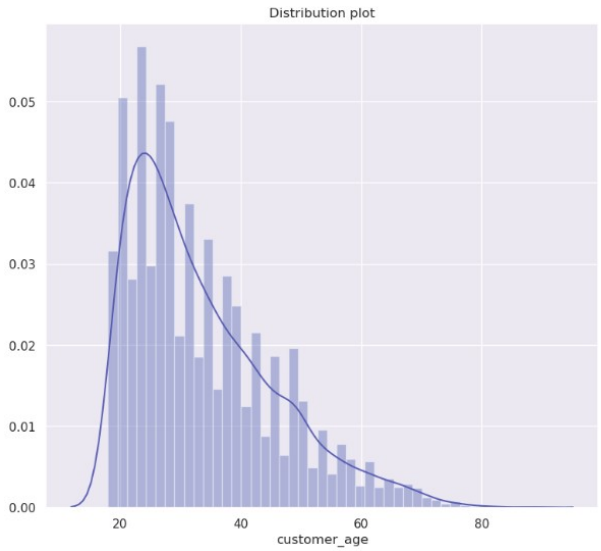


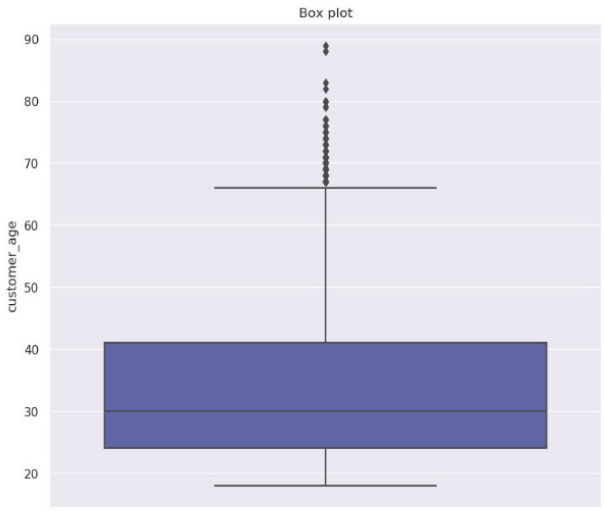


Most of the order amount is spread between 0 to 500 and there are few outliers too. So outlier detection is performed and got the results as 245 orders are not between +/-3 standard deviation.

ii)Customer Age Distribution

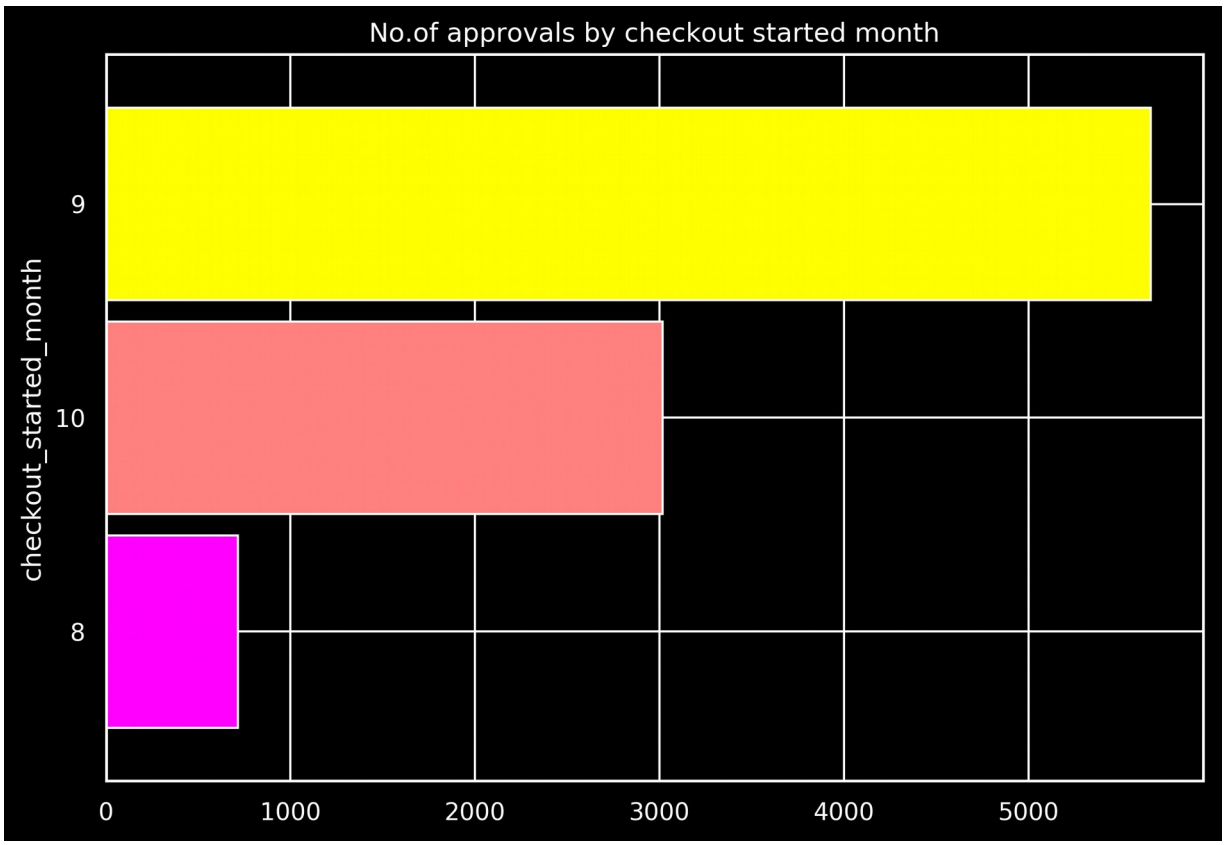




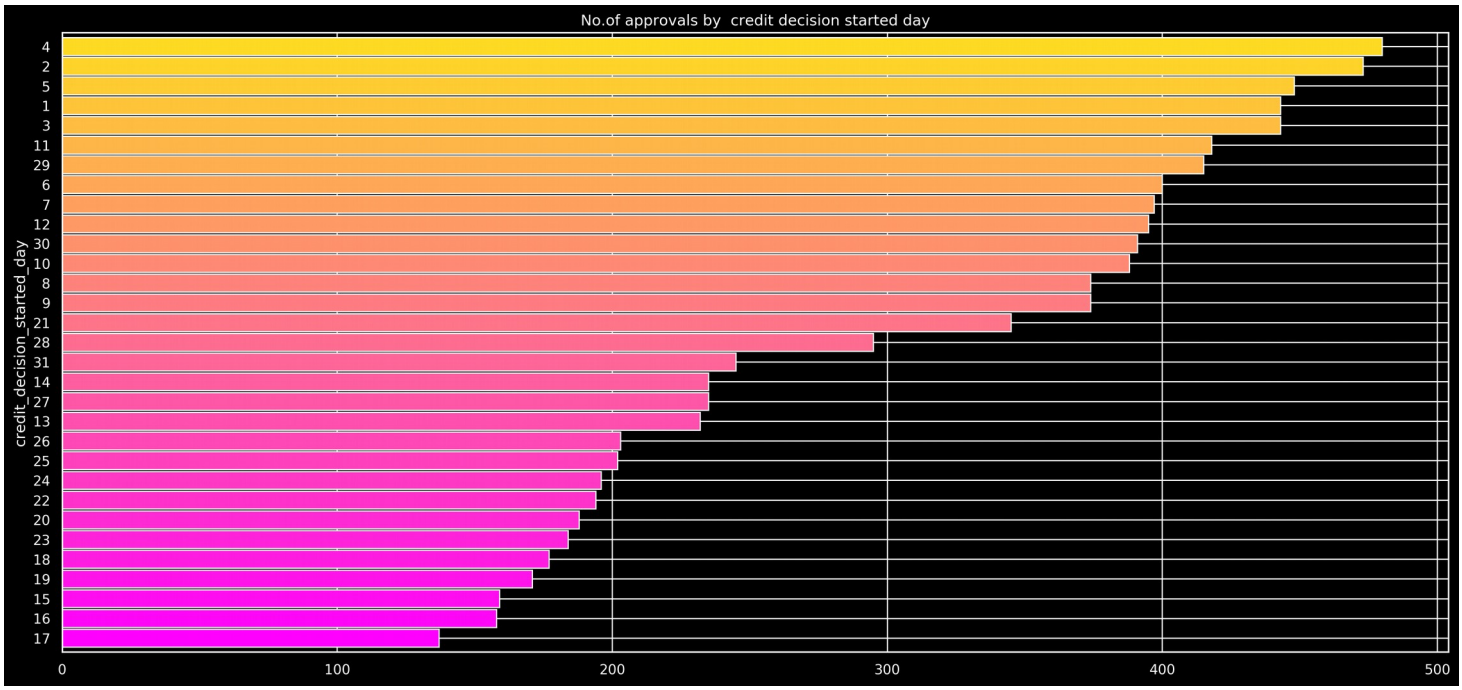


Most of the oders are placed by the customers who are between the age of 21-30 and second most orders are from the age group 31-40.

iii)Month and Day Wise Distribution:



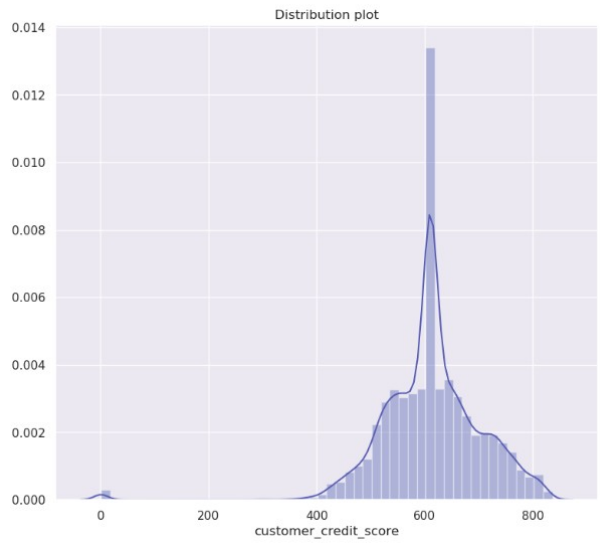
iv) Day Wise Distribution:

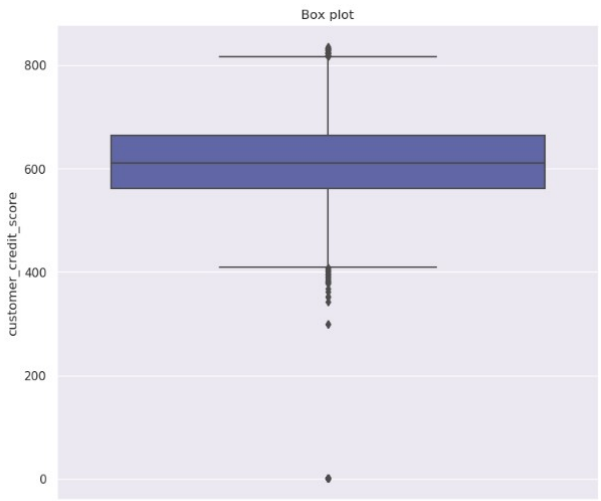


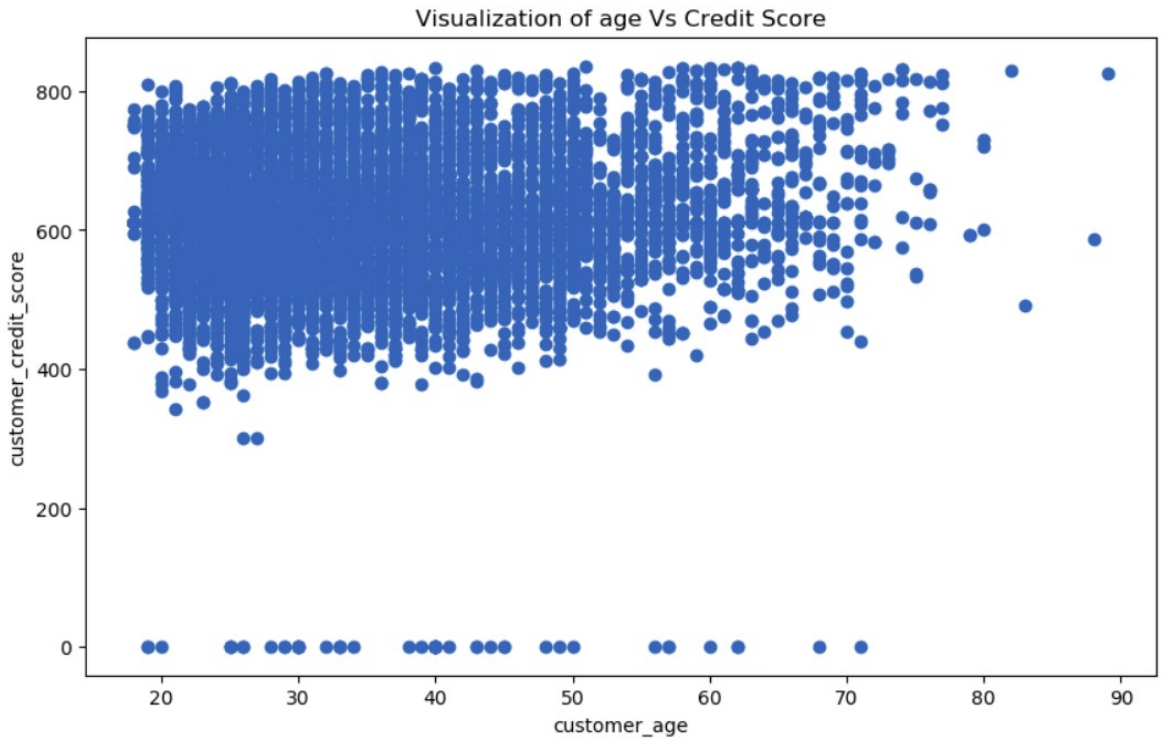
From the above charts we can conclude ,

- Most of the orders are placed in September,2018.
- Maximum orders are getting placed in the first 5 days of every month.

v) Customer credit score distribution:

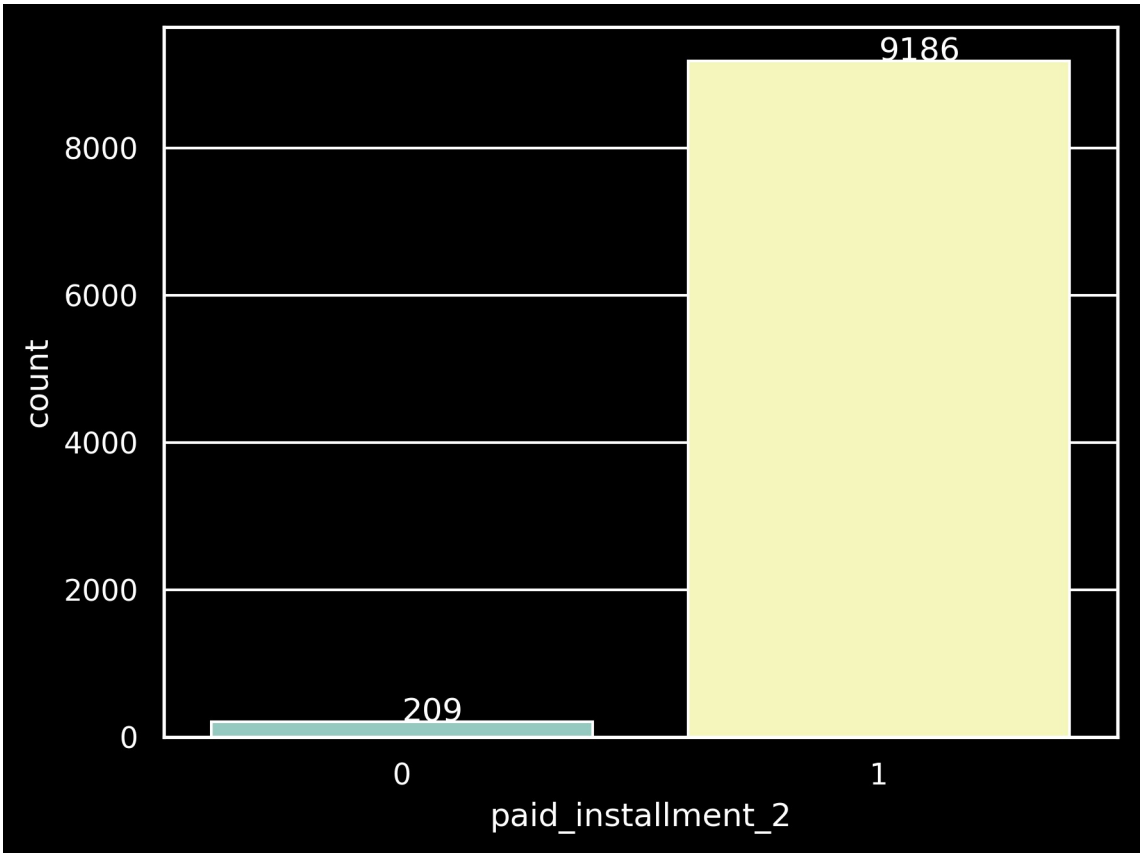


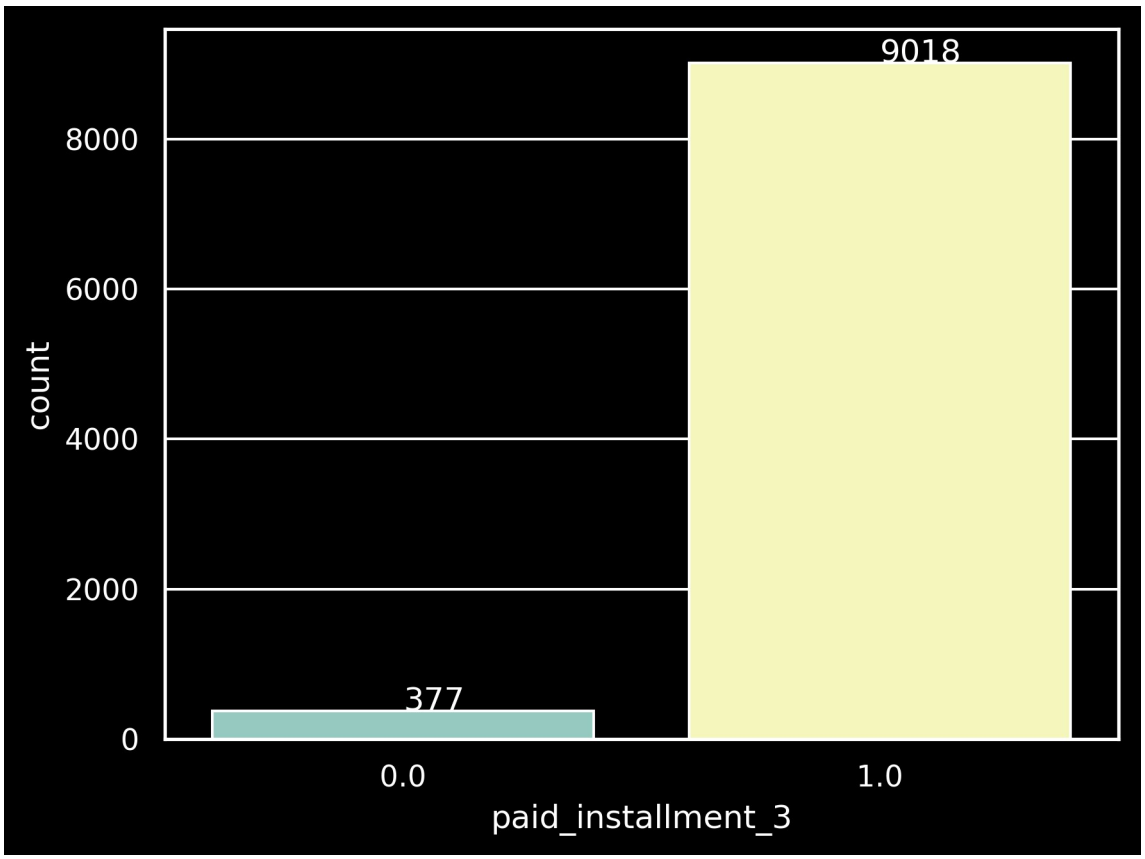


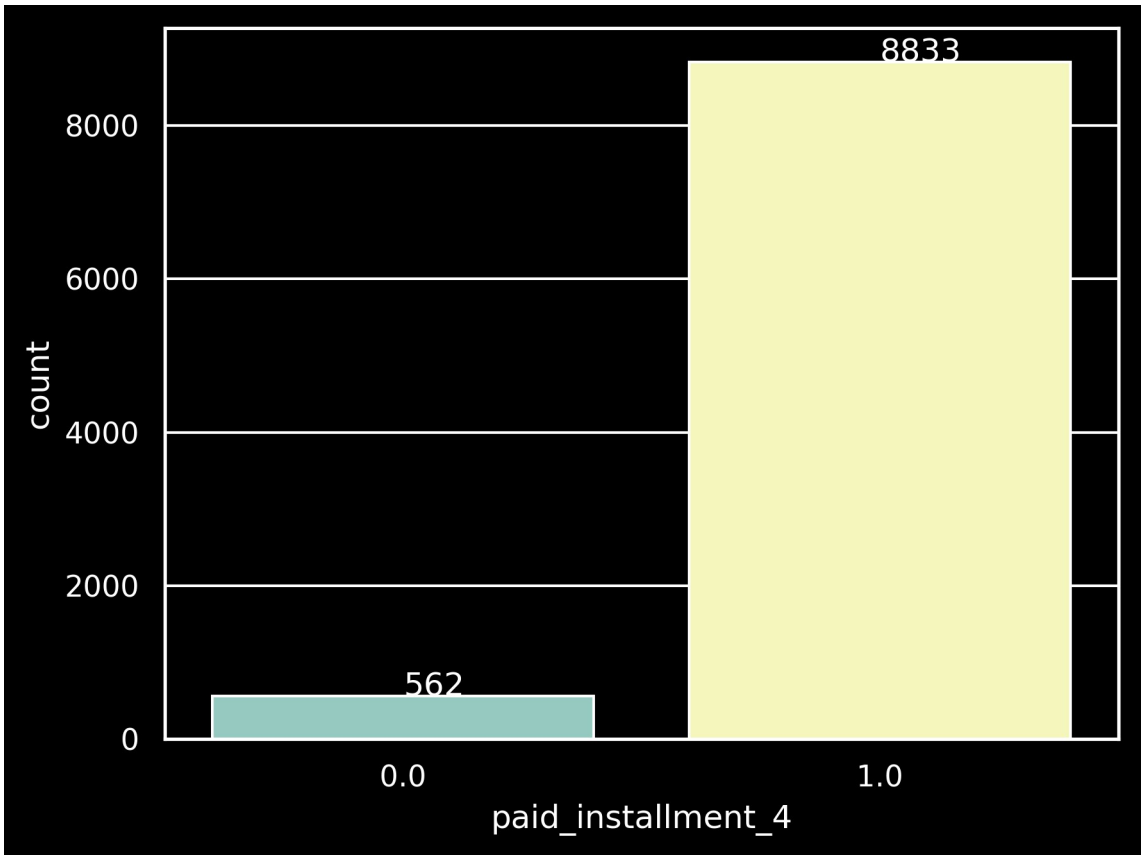


Customer credit scores are in the range between 400 to 800 which indicates the defaulters are less in this dataset.

vi) Installment paid by the customers:




9018
377
8000
6000
4000
2000
0
count
0.0
1.0
paid_installment_3


8833
562
8000
6000
4000
2000
0
count
0.0
1.0
paid_installment_4

As the first installment is mandatory, the other installment details are analysed in the above charts.

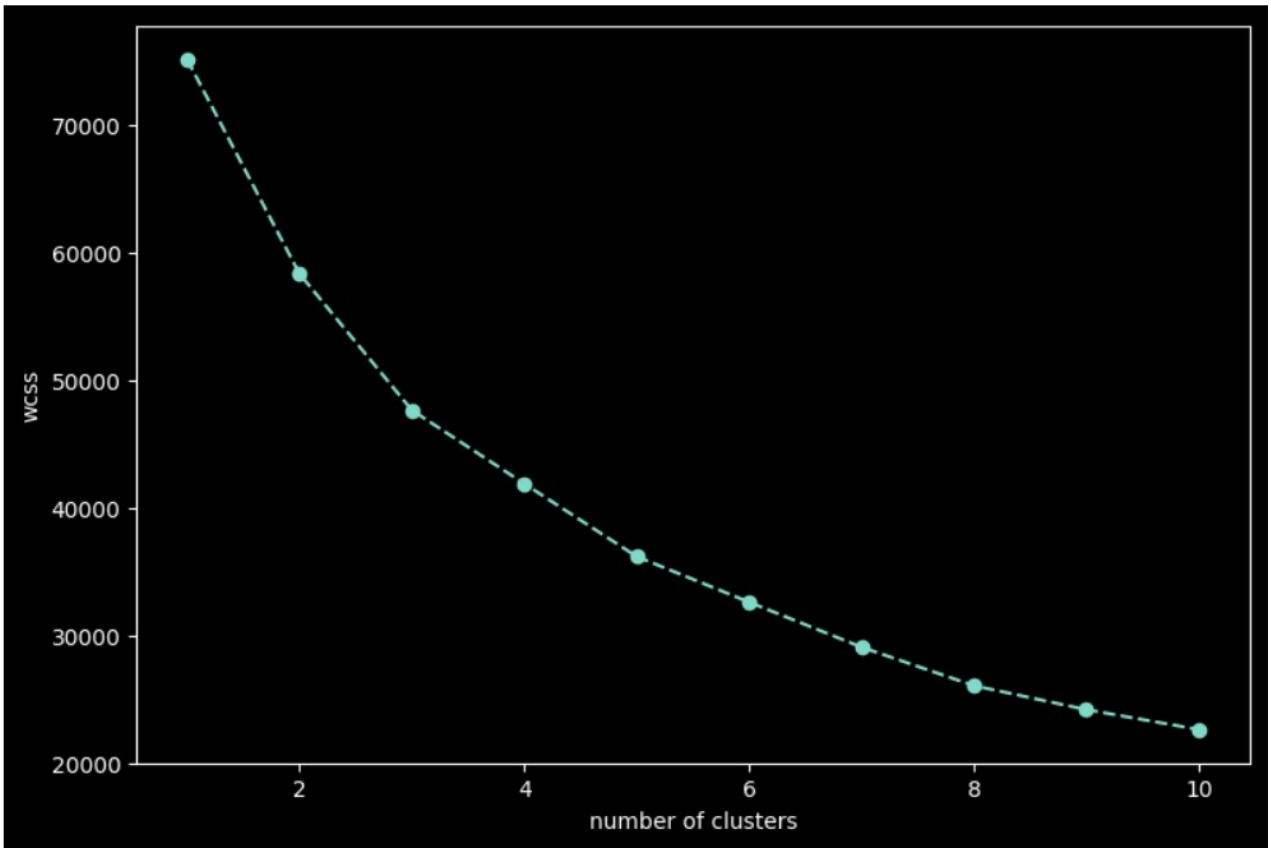
Compared to the second, third and fourth installments most of the customers are having pending payment in fourth installment.

Clustering the customers:

The total amount of unique customers are 8428. As there are too many unique customers details, its better to segment them based on similar properties and then analyse.

Before segmentation, elbow method is used to find out the optimal value for K.

As per the below chart, there is a sharp decline in the error at k=3 and k=5 and i choose k=3 for clustering.



After segmentation, the mean of all the columns for each clusters is calculated as follows.

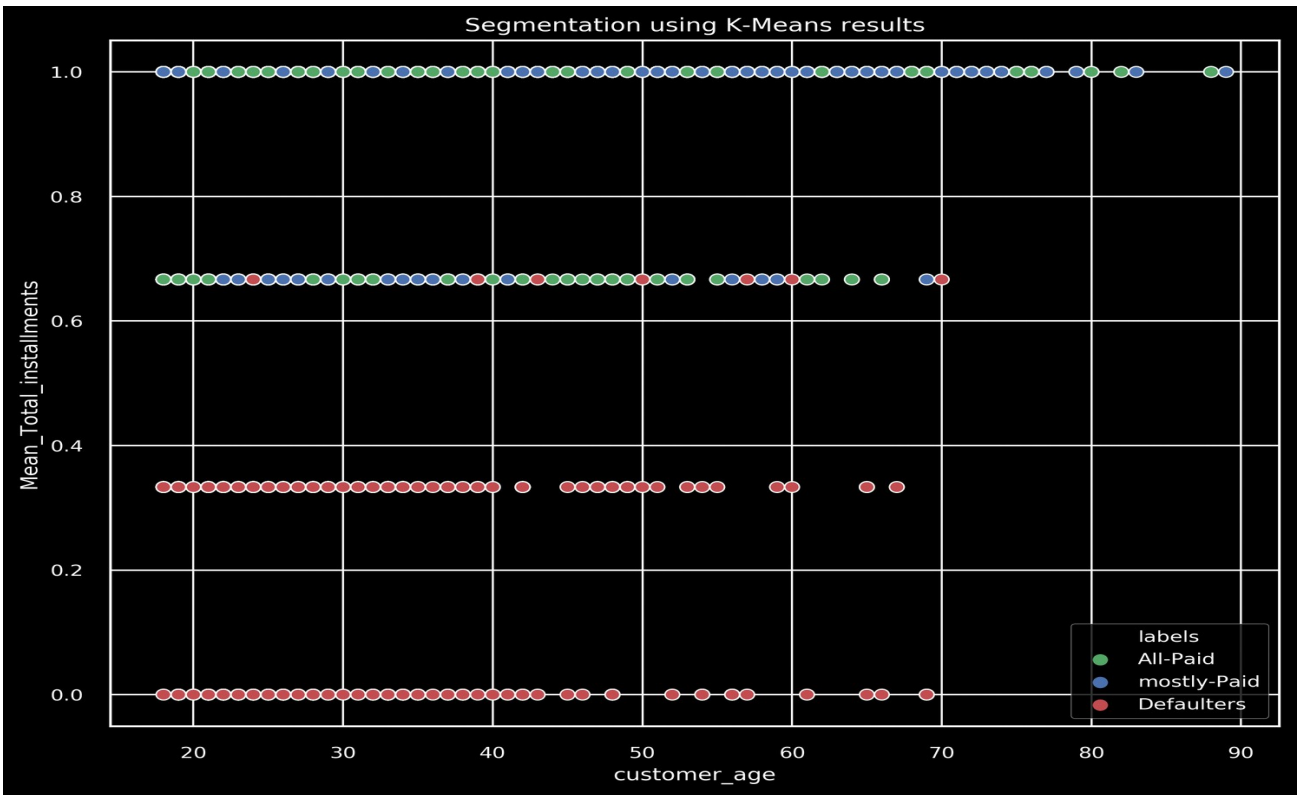
egment_k-means	order_amount	customer_credit_score	customer_age	paid_installment_2	paid_installment_3	paid_installment_4	N obeservations	proportion_of_obs
All-Paid	140.578397	615.260552	33.668831	0.998377	1.000000	0.970982	4928	0.524534
mostly-Paid	152.319607	616.825221	33.726893	0.995329	0.999754	0.976647	4068	0.432996
Defaulters	138.082531	552.411028	31.070175	0.543860	0.057644	0.187970	399	0.042469

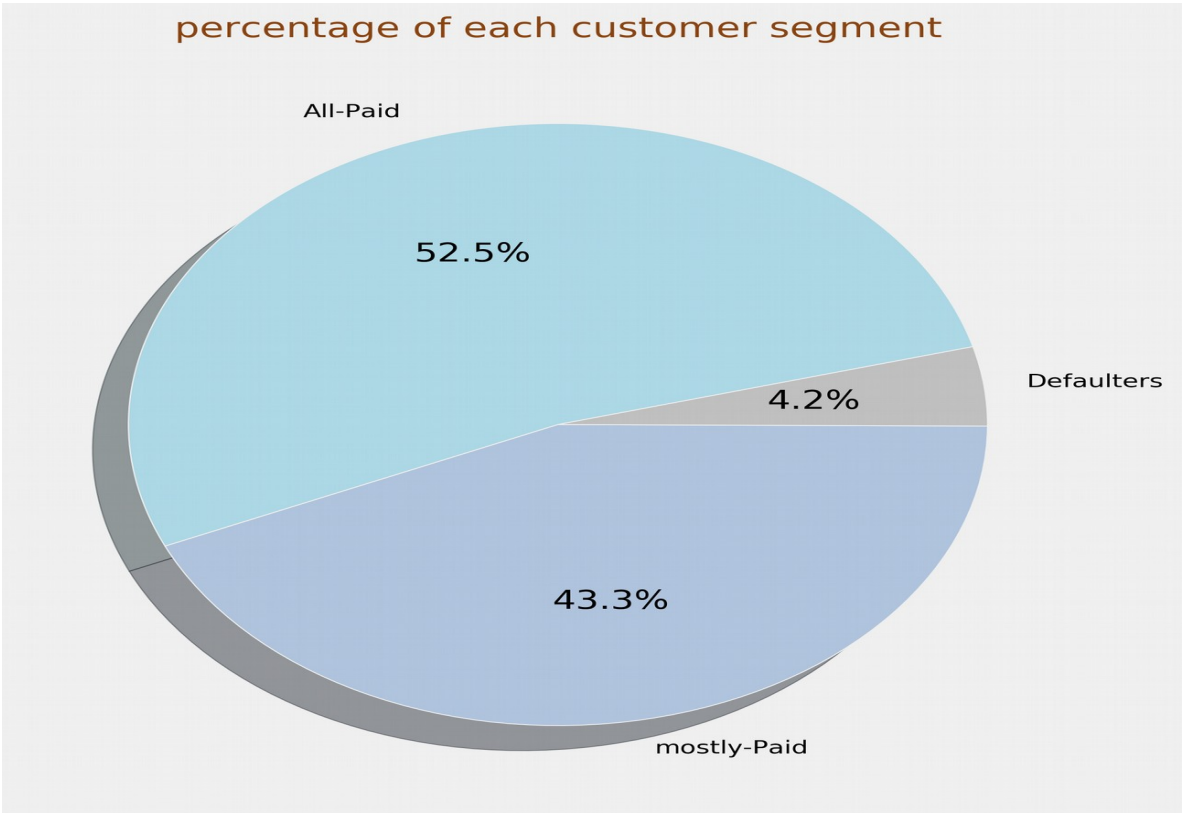
1) As per the above image, the Defaulters/ customers who are having pending for the installments are 399 in total and they are in the proportion of 0.04. The mean of their pending installment details are 0.5,0.5 and 0.1 respectively.

2) Compared to other 2 categories the mean credit scores for the Defaluters are low.

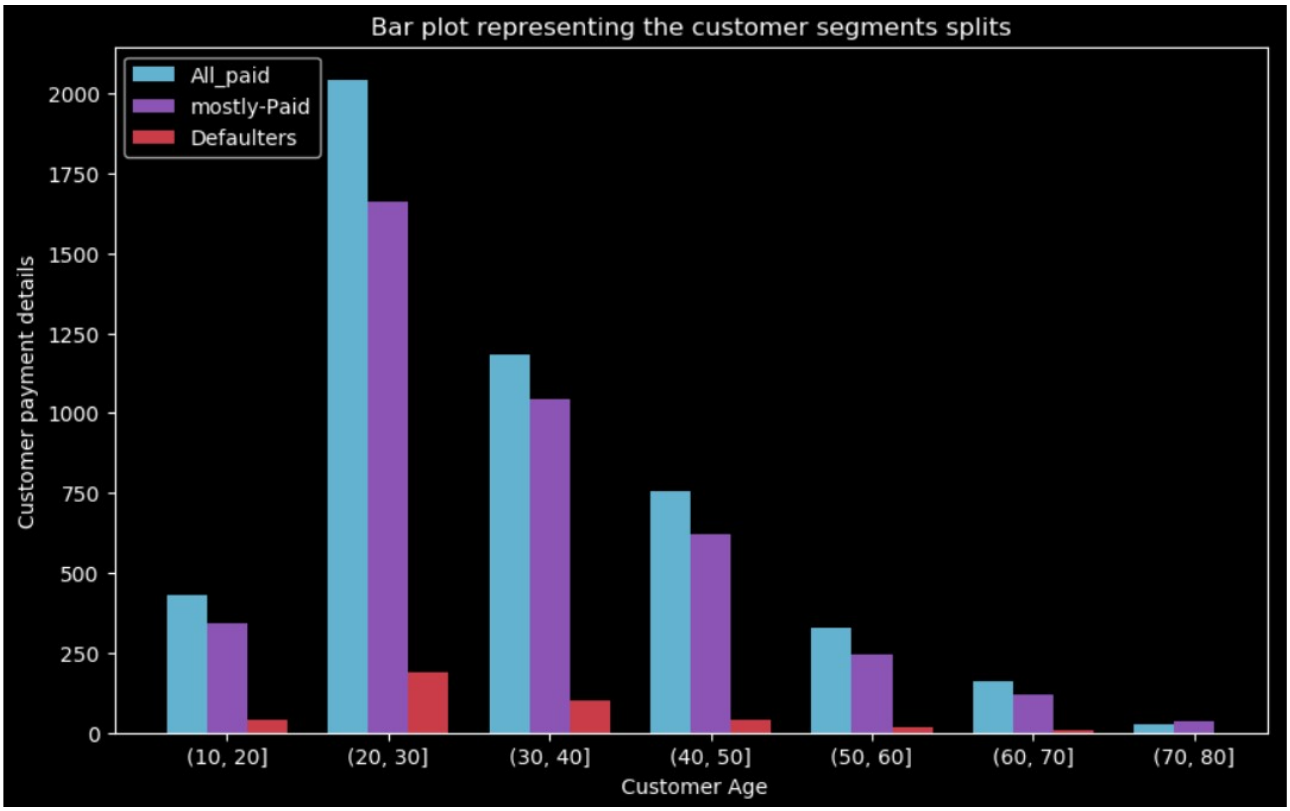
The below charts represents the customers in three different groups and thier proportions of split after clustering.

There is a very less variance between All-paid and mostly-paid customers, but the defaulters are well-splitted from other two.


Segmentation using K-Means results
Mean_Total_installments
customer_age
1.0
0.8
0.6
0.4
0.2
0.0
20
30
40
50
60
70
80
90
labels
All-Paid
mostly-Paid
Defaulters


percentage of each customer segment
All-Paid
52.5%
Defaulters
4.2%
43.3%
mostly-Paid

Age wise distributiom of clusters:



The defaulters are more in the age group of 20 to 30.

The customers are splitted into 3 different groups and labelled them accrodingly for the supervised classification.

4) What additional data would you like to see that might help build a better installment- approval classifier?

- If the feature **approved_for_installments** is given for both the approval and non-approval, it would be helpful to analyse the customers well.

- If the **Customer Income details** are provided, based on the income level how many customers are paying correctly can be analysed

- If the **product details** for all orders are provided, the installment-approval can be analysed in terms of products prespective also.

- If **payment date and time** is provided , there is an option to anayse the customers for on time payment

5) What would be your next steps to train/build a model that we could use to make real time customer approval decisions?

As per the analysis, some columns are having outliers. So instead of removing them from the dataset, Decision tree classifier is chosen for modeling which is not affected by the outliers mostly.

The accuracy from the model is 0.5508 and the confusion matrix is shown below.

	All-Paid	Defaulters	mostly-Paid
All-Paid	576	0	444
Defaulters	0	80	0
mostly-Paid	400	0	379

There are misclassifications between All-Paid and mostly paid customers as there is a very slight deviation between them but the defaulters are classified correctly.