**BUSINESS DATA**

**ANALYSIS REPORT**

Contents

[Introduction 3](#_Toc128939125)

[Methodology: 4](#_Toc128939126)

[Problem1 5](#_Toc128939127)

[A. What is the important technical information about the dataset that a database administrator would be interested in? 5](#_Toc128939128)

[B. Take a critical look at the data and do a preliminary analysis of the variables. Do a quality check of the data so that the variables are consistent. Are there any discrepancies present in the data? If yes, perform preliminary treatment of data. 5](#_Toc128939129)

[C. Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business. 6](#_Toc128939130)

[D. Understanding the relationships among the variables in the dataset is crucial for every analytical project. Perform analysis on the data fields to gain deeper insights. Comment on your understanding of the data. 8](#_Toc128939131)

[E. Employees working on the existing marketing campaign have made the following remarks. Based on the data and your analysis state whether you agree or disagree with their observations. Justify your answer Based on the data available. 13](#_Toc128939132)

[E1) Steve Roger says “Men prefer SUV by a large margin, compared to the women” 13](#_Toc128939133)

[E2) Ned Stark believes that a salaried person is more likely to buy a Sedan. 13](#_Toc128939134)

[E3) Sheldon Cooper does not believe any of them; he claims that a salaried male is an easier target for a SUV sale over a Sedan Sale. 14](#_Toc128939135)

[F. From the given data, comment on the amount spent on purchasing automobiles across the following categories. Comment on how a business can utilize the results from this exercise. Give justification along with presenting metrics/charts used for arriving at the conclusions. 15](#_Toc128939136)

[From the Visualization and analysis, Following are the observations 15](#_Toc128939137)

[G. From the current data set comment if having a working partner leads to the purchase of a higher-priced car. 16](#_Toc128939138)

[From the Visualization and analysis, I disagree that working partner leads to the purchase of a higher-priced car. Though buyers with and without partners working are buying high priced cars, car buyers with partners working are spending slightly less on an average as compared to Car buyers with partners not working. 16](#_Toc128939139)

[H. The main objective of this analysis is to devise an improved marketing strategy to send targeted information to different groups of potential buyers present in the data. For the current analysis use the Gender and Marital\_status - fields to arrive at groups with similar purchase history. 17](#_Toc128939140)

[Problem 2 18](#_Toc128939141)

[Dataset Analysis 18](#_Toc128939142)

[Data Description 18](#_Toc128939143)

[Uni-Variate Analysis 19](#_Toc128939144)

[Bi-Variate Analysis 23](#_Toc128939145)

[Multi-Variate Analysis 33](#_Toc128939146)

[Results 35](#_Toc128939147)

Introduction**:**

**Austo Motor Company has concerns on the efficiency of the marketing campaign currently being used and wants to improve their marketing campaign.** In this report I aim to improve their existing marketing campaign using the dataset shared by the company. The way we will model this is by thoroughly analysing all the problems posed by the bank using various visualization techniques and providing solutions.

**GODIGT** Bank has a set of customers who were given credit cards based on risk policy and customer category class but due to huge competition in the credit card market, the bank is observing high attrition in credit card spending. In this report I aim to revisit their credit card policy using the dataset shared by the bank and provide recommendations on the right credit card to customers so that they show higher intent toward the recommended credit card. The way we will model this is by the predicting the top 5 important variables that impact the customer attrition.

# Methodology:

In this work, we will explore the following directions:

1)Analysing the raw data for any anomalies, duplicate records, missing values and null values, wrong datatypes

2)Processing the data by treating the missing/null values/anomalies by replacing values or imputing them, deleting duplicate records and assigning right datatypes to the variables

3) Univariate Analysis to understand the distribution of values for a single variable.

4) Bivariate Analysis to understand relationship between two sets of variables

5) Multi- Variate Analysis to understand relationship between more than two variables.

6)Drawing Inferences from the visualizations and suggesting solutions.

|  |
| --- |
| **Problem1**A. What is the important technical information about the dataset that a database administrator would be interested in? **From the Austo Motor Company Dataset we observe that there** 1581 records and 14 variables.  There are 8 object type variables,5 int type variables and 1 float type variable. B. Take a critical look at the data and do a preliminary analysis of the variables. Do a quality check of the data so that the variables are consistent. Are there any discrepancies present in the data? If yes, perform preliminary treatment of data. After a preliminary check on the categorical variables to check for anomalies, duplicates and missing values, we see that:  1. There are no duplicate records in the dataset.  2. Numerical variables are Age, No\_of\_Dependents, Salary, Partner\_salary, Total\_salary and Price.  Salary and Total Salary variables are integer type and Partner Salary is Float type variable. So first we convert Salary and Total Salary variables to float type.  All numerical variables except Partner\_salary have no null values and no negative values or anomalies.  Partners salary is checked for null values with Partner’s working and it is found that there are 16 records with Partners salary null but partner is working. Since Total Salary column is sum of Salary and Total Salary,  we can impute those values with difference of Total Salary and Salary. Rest of values can be imputed to 0 as Partner is not working, therefore partner salary will be 0  **3. Categorical variables Gender, Profession, Marital\_status, Education, Personal\_loan, House\_loan, Partner\_working and** Make do not have any anomalies or null values except Gender.**Categorical variable Gender has values Male, Female,** Femle, Femal and 53 null values. These have to be treated. So, values Femle, Femal is changed to Female. 53 null values are imputed to the mode value of **Gender which is Male.** |
| C. Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business.Visualization of Numerical Variables  Insights drawn from Visualizations 1.Price range of cars is $18000.00 to $70000.00. It’s a right skewed distribution.  2.Car buyers range from age 22 to 54. It’s a right skewed distribution.  3.Salary of the buyers ranges from $30000.00 to $99300.00. It’s a right skewed distribution.  4.Partner Salary of the buyers ranges from $100.00 to $80500.00, there are some outliers. It’s a right skewed distribution.  5.Total Salary of the buyers ranges from $32900.00 to $171000.00, there are some outliers. It’s a right skewed distribution.  6.Maximum number of buyers have either 2 or 3 dependents. Visualization of Categorical Variables  Insights drawn from Visualizations 1.There are more married buyers than single buyers  2.Salaried Car buyers are buying cars more than Business Car buyers  3.Post-Graduate buyers are more than Graduate buyers  4.Males outnumber Females in buying cars  5.Sedan buyers are more in comparison with SUV and Hatchback  6.There are more car buyers with partners working than those with partners not working  7.There are more car buyers with no house loan than those with house loan  8.The number of car buyers with or without personal loan are almost same |
| D. Understanding the relationships among the variables in the dataset is crucial for every analytical project. Perform analysis on the data fields to gain deeper insights. Comment on your understanding of the data.Visualization of relations between categorical variables  Insights drawn from Visualizations 1. Males prefer Hatchback and Females Prefer SUV. Approximately 47% of Males prefer Hatchback and 43% of Males prefer Sedan and mere 10% prefer SUV.52.5% of Females prefer Sedan and 43% of Females prefer SUV and mere 4.5% prefer Hatchback  2. Among Business Professionals 45% prefer Sedan,42% prefer Hatchback,13% prefer SUV  Among Salaried Professionals 44% prefer Sedan,33% prefer Hatchback,23% prefer SUV.  Sedan is most preferred car by both Business and Salaried professionals.  3.46% of Married Car buyers prefer Sedan 28% of Single Car buyers prefer Sedan.34.5% of Married Car buyers prefer Hatchback and 61% prefer Hatchback.19% of Married Car buyers prefer SUV and 11 % of Single Car buyers prefer SUV. To conclude Married buyers, prefer Sedan and Single buyers prefer hatchback.  4. Among Graduates 43% prefer Sedan,37% prefer Hatchback,20% prefer SUV  Among Post Graduates 45% prefer Sedan,37% prefer Hatchback,18% prefer SUV  Sedan is most preferred car by both Graduates and Post Graduates.  5. Among Car buyers who have taken personal loan 40% prefer Sedan,37% prefer Hatchback, 23% prefer SUV. Among Car buyers who have not taken personal loan 49% prefer Sedan,37% prefer Hatchback, 14% prefer SUV. Sedan is most preferred car by buyers with or without Personal loan.  6. Among Car buyers who have taken home loan 41% prefer Sedan,33% prefer Hatchback, 26% prefer SUV. Among Car buyers who have not taken home loan 51% prefer Sedan,45% prefer Hatchback, 4% prefer SUV. Sedan is most preferred car by buyers with or without House loan.  7. Among Car buyers who have Partners working 40% prefer Sedan,40% prefer Hatchback, 20% prefer SUV. Among Car buyers who don't have Partners working 48% prefer Sedan,35% prefer Hatchback, 17% prefer SUV. Sedan is most preferred car by buyers with or without Partner Working.  8. Of the car buyers with 0 dependents 75% prefer Hatchback 25% SUV.  Of the car buyers with 1 dependent 80% prefer Sedan, 15% SUV, 5 % Hatchback.  Of the car buyers with 2 dependents 52% prefer Sedan ,32% Hatchback,16% SUV.  Of the car buyers with 3 dependents 46% prefer Hatchback ,33% Sedan,21% SUV.  Of the car buyers with 4 dependents 55% prefer Hatchback ,25% SUV,20% Sedan.  Sedan is most preferred car by buyers with 1 or 2 dependents.  Hatchback is most preferred car by buyers with 3,4 or no dependents. Visualization of relations between categorical variables and Numerical VariablesInsights drawn from Visualizations 9. There are more buyers of hatchback in age range 20-30, 30-40 years age range prefer Sedan,40-60 years age prefer SUV.  10. Hatchback cars are the least expensive and SUVs are the most expensive cars.  11. On average Female buyers are spending more on cars than Male buyers.  12. On average Salaried buyers are spending more on cars than Business buyers.  13. On an average Graduate buyer are spending slightly more on cars than Post Graduate buyers.  14. On an average Married buyers are spending more on cars than Single buyers.  15. On an average Buyers with partners not working are spending slightly more on cars than Buyers with partners working.  16. Buyers with higher salary prefer SUV. Visualization of relations between Numerical Variables    Insights drawn from Visualizations 17.The correlation between Price of Cars and Total Salary is not very high.  18.There is a high correlation between Price of Cars and Age of Car Buyers  19.There is no correlation between Price of Cars and no of dependents of Buyers |
| E. Employees working on the existing marketing campaign have made the following remarks. Based on the data and your analysis state whether you agree or disagree with their observations. Justify your answer Based on the data available.E1) Steve Roger says “Men prefer SUV by a large margin, compared to the women” From the Visualization and analysis, I disagree that Men prefer SUV, in fact Men prefer Hatchback slightly more than Sedan, lastly followed be SUV   E2) Ned Stark believes that a salaried person is more likely to buy a Sedan. From the Visualization and analysis, I agree that salaried person is more likely to buy a Sedan.   E3) Sheldon Cooper does not believe any of them; he claims that a salaried male is an easier target for a SUV sale over a Sedan Sale.   From the Visualization and analysis, I disagree that salaried males are an easier target for a SUV sale over a Sedan Sale. In fact, salaried males prefer Sedan, followed by Hatchback and then SUV |
| F. From the given data, comment on the amount spent on purchasing automobiles across the following categories. Comment on how a business can utilize the results from this exercise. Give justification along with presenting metrics/charts used for arriving at the conclusions. Give justification along with presenting metrics/charts used for arriving at the conclusions.  F1) Gender  F2) Personal\_loan     From the Visualization and analysis, Following are the observations 1 Female buyer are buying more expensive cars  2.Male Buyers are buying car in range $18000.0 to $58000.0 with some going up to $70000.  3.Female Buyers are buying car in range $20000.0 to$ 69000.0.  4.Both Male and Female buyers with personal loan are spending slightly less on cars on an average as compared to buyers without personal loan |
| G. From the current data set comment if having a working partner leads to the purchase of a higher-priced car.  From the Visualization and analysis, I disagree that working partner leads to the purchase of a higher-priced car. Though buyers with and without partners working are buying high priced cars, car buyers with partners working are spending slightly less on an average as compared to Car buyers with partners not working. |
| H. The main objective of this analysis is to devise an improved marketing strategy to send targeted information to different groups of potential buyers present in the data. For the current analysis use the Gender and Marital\_status - fields to arrive at groups with similar purchase history.       1.Married Males prefer Sedan and Married females prefer SUV  2.Single Males prefer Hatchback and Single Females prefer Sedan  3.Married Males prefer cars in price range $18000 to $58000 with some outliers going up to $70000 and Married Females prefer cars in price range $22000 to $69000  4.Single Males prefer cars in price range $18000 to $46875.0 with some outliers going up to $68000 and Single Females prefer cars in price range $20000 to $66000.  5.Married Buyers are spending slightly more than Single Buyers and Females and spending more on cars than Males |

# **Problem 2**

**Analyse the dataset and list down the top 5 important variables, along with the business justifications.**

## **Dataset Analysis:**

### Data Description

The original dataset contains records of 8448 customers with 28 different variables. There is 1 datetime variable ,19 integer variables and 8 categorical variables. Variables User ID, Card No, Card Bin Number, Card\_source\_date is not of any significance for the data analysis, so we won’t be performing any analysis using these variables.

**Issuer** is a categorical variable with 3 types of Issuers 'Visa' 'Mastercard' 'Amex'

**card type** is a categorical variable with 15 card types 'edge' 'prosperity' 'rewards' 'indianoil' 'cashback' 'shoprite',' chartered' 'aura' 'gold' 'smartearn' 'prime' 'pulse' 'platinum ‘centurion' 'elite'

**high\_networth** is a categorical variable with 5 unique values A, B, C, D, E

**Active\_30, Active\_60, Active\_90** are categorical variables with values 0 and 1

**CC\_active30, CC\_active60, CC\_active90** are categorical variables with values 0 and 1

**Hotlist\_flag** is a categorical variable with values Y and N

**widget\_products** is a categorical variable with 8 values 0,1,2,3,4,5,6,7

**engagement\_products** is a categorical variable with 9 values 0,1,2,3,4,5,6,7,8

**other\_bank\_cc\_holding** is a categorical variable with 2 values Y and N

**Transactor\_revolver** is a categorical variable with 2 values T and R and there are 38 null values

**Occupation\_at\_source** is a categorical variable with 5 unique values Self Employed,0, Student, Salaried, Retired, Housewife

**T+1\_month\_activity** is a categorical variable with 2 values 0 and 1

**T+2\_month\_activity** is a categorical variable with 2 values 0 and 1

**T+3\_month\_activity** is a categorical variable with 2 values 0 and 1

**T+6\_month\_activity** is a categorical variable with 2 values 0 and 1

**T+12\_month\_activity** is a categorical variable with 2 values 0 and 1

**annual\_income\_at\_source** is a numerical variable

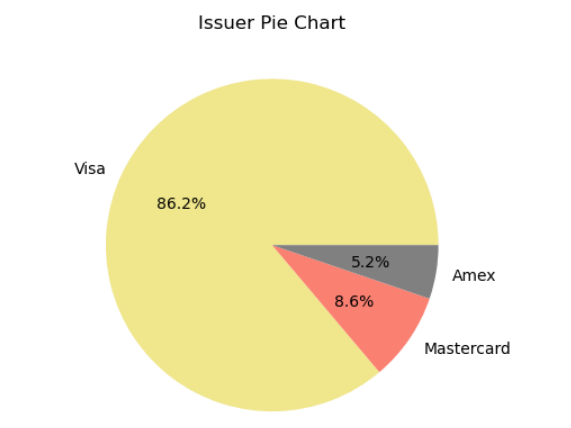
**avg\_spends\_l3m** is a numerical variable

**cc\_limit** is a numerical variable

**bank\_vintage** is a numerical variable

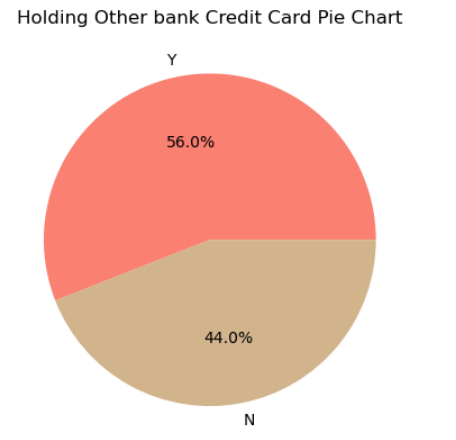
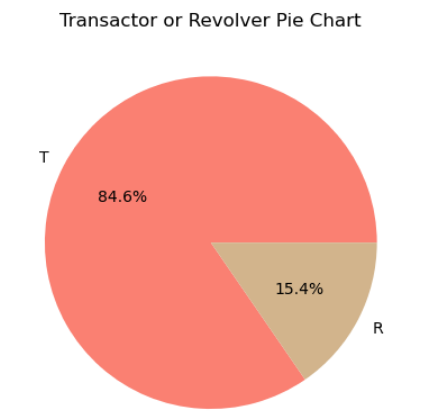
### Uni-Variate Analysis

Categorical Variables-There are 20 categorical variables, we can visualize the distribution using the count plot and pie chart. We can deduce the following aspects from the visualizations

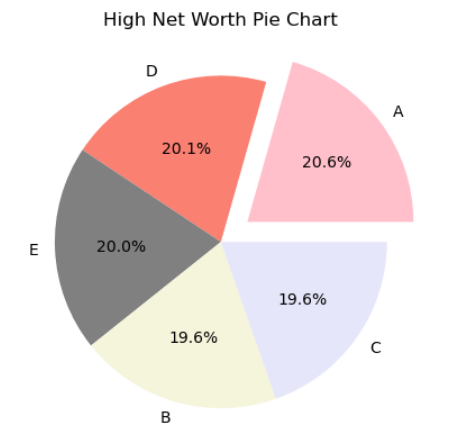
1.86% of Card holders use Visa card

2.There are more card holders with Rewards card.

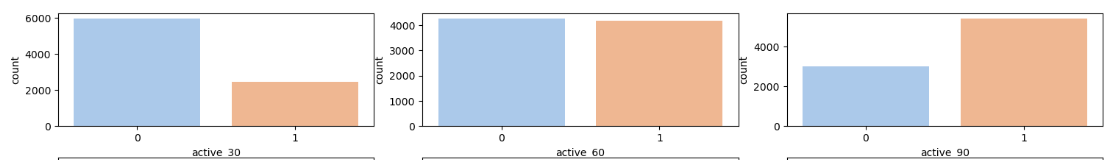


3. 84.6% of Card Holders are Transactors

4.56% of Card holders are having Credit Card of other banks



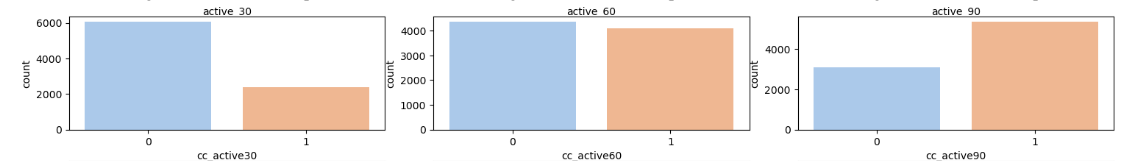
5.”A” type net worth card holders are more in number in comparison with other net worth kind



4.71% of Credit Card Holders did not have any account activity in last 30 days

5.51% of Credit Card Holders did not have any account activity in last 60 days

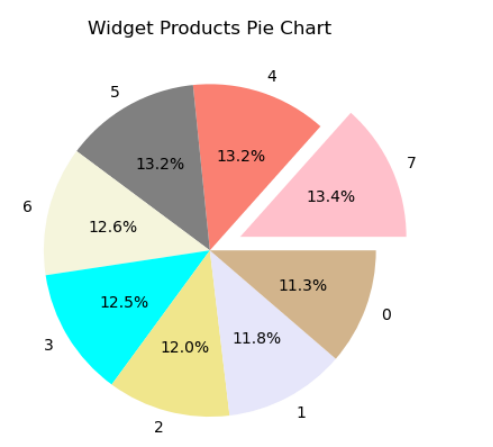
6.64% of Credit Card Holders had account activity in last 90 days



7.72% of Credit Card Holders were not active users of credit card in the last 30 days

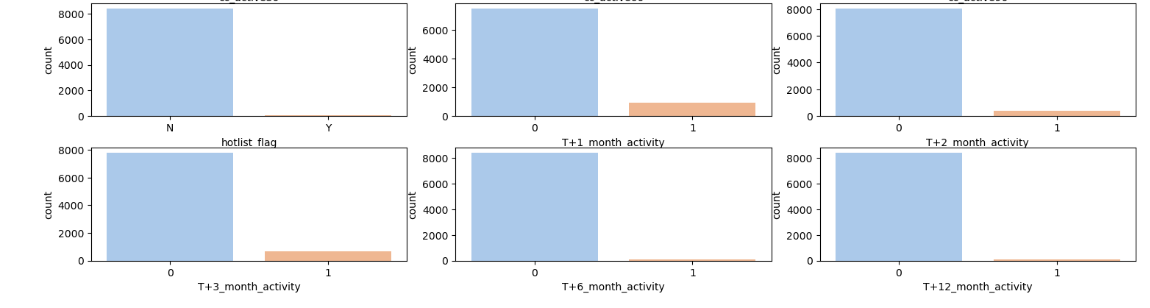
8. 52% of Credit Card Holders were not active users of credit card in the last 60 days

9. 63% of Credit Card Holders were active users of credit card in the last 90 days

10.Card holders having 7 widget products are slightly more in number.

11.Card holders having 4 Engagement products are slightly more in number.



13.89% of card holders are predicted to have no activity in T+1st month

14.95% of card holders are predicted to have no activity in T+2nd month

15.92% of card holders are predicted to have no activity in T+3rd month

16.99% of card holders are predicted to have no activity in T+6th month

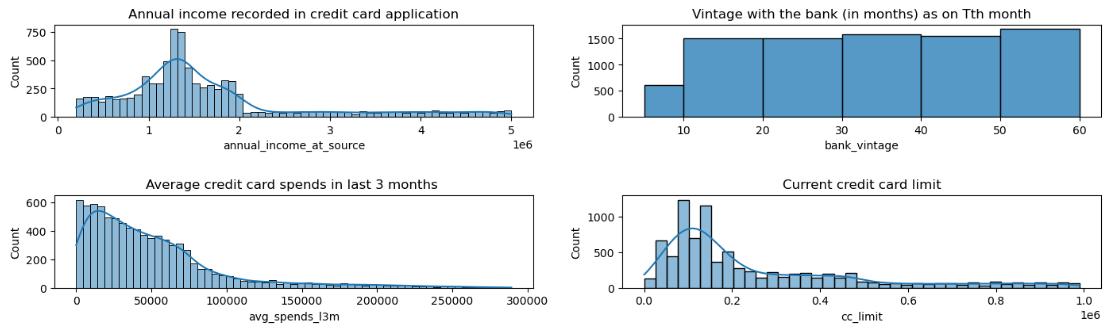
17.99% of card holders are predicted to have no activity in T+12th month

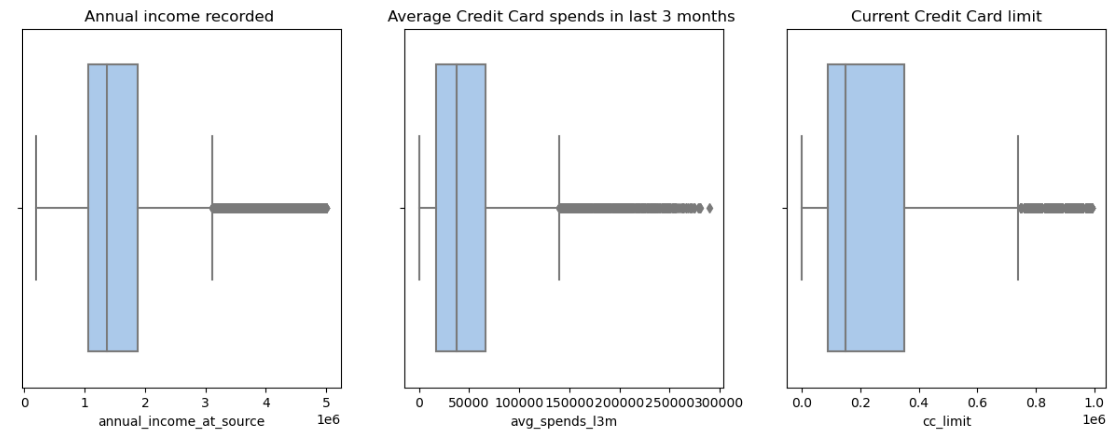
18.Only 0.4% of Card holders cards is hot listed card



19.46.4% of card holders are salaried professionals

Numerical Variables-There are 4 numerical variables, we can visualize the distribution using the Histogram and Boxplot





From the above distribution we can deduce that

Annual Income is right skewed distribution with outliers. There are more CC holders in income range 13-14 lakhs. Min salary is ₹2,00,095 and Maximum salary is ₹49,99,508

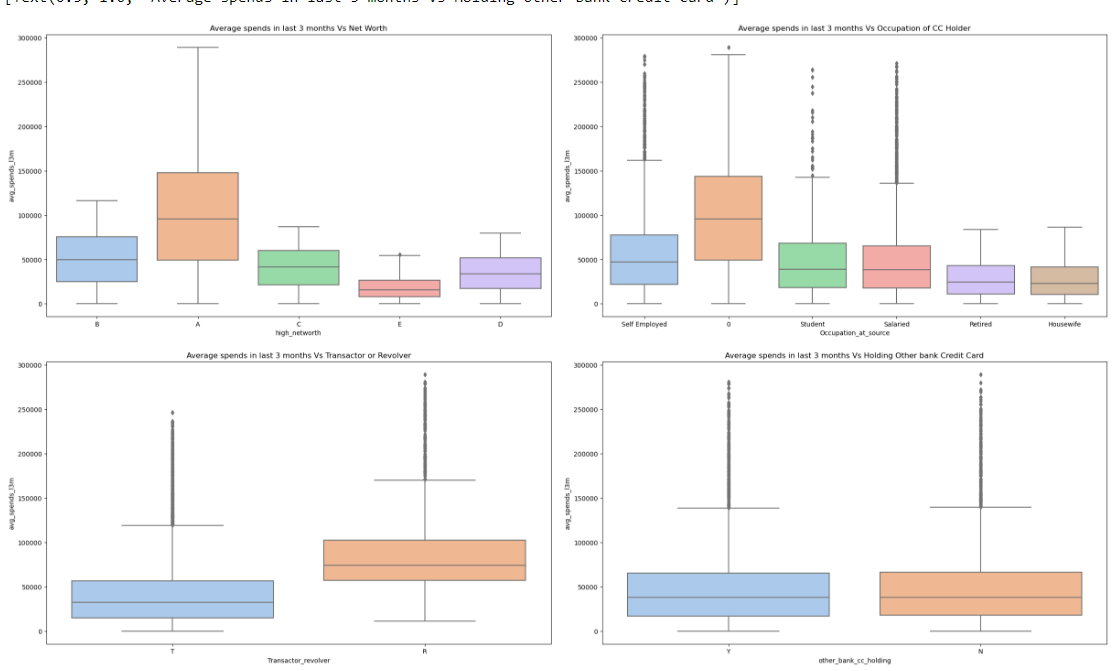
Average credit card spends in last 3 months is right skewed distribution with outliers. Min amount spent is 0 and maximum is ₹289292.00

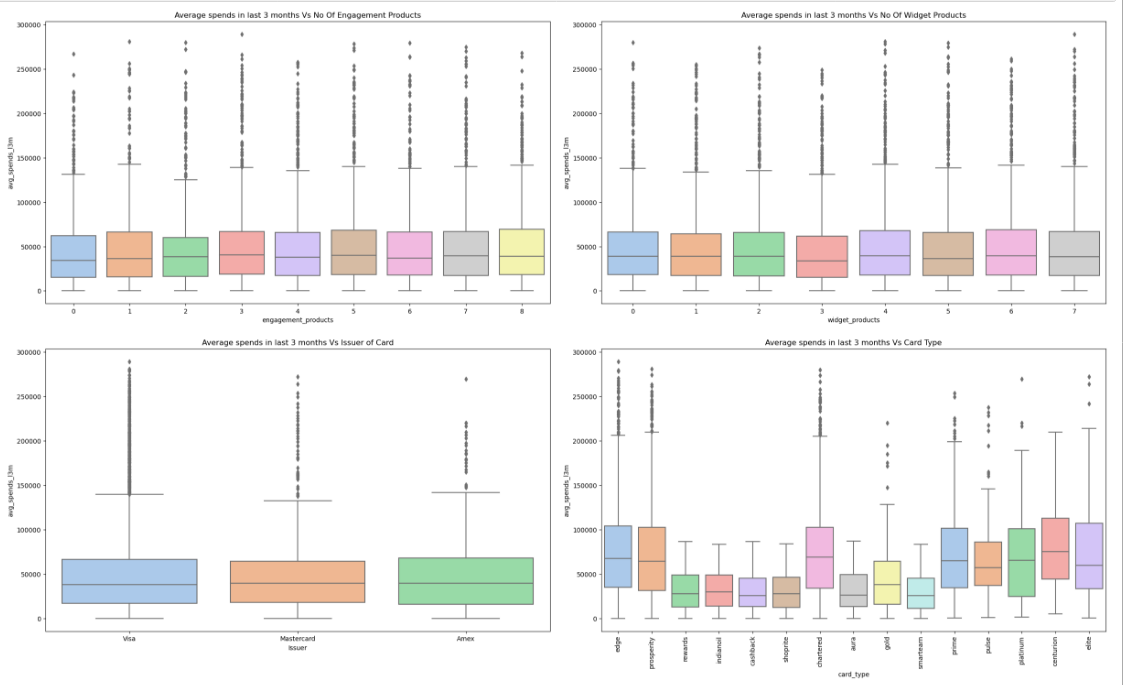
Current credit card limit is right skewed distribution with outliers. Min limit spent is ₹20000 and maximum limit is ₹990000.00. Though the distribution shows 0 as CC limit, its only for one record that has been hot listed

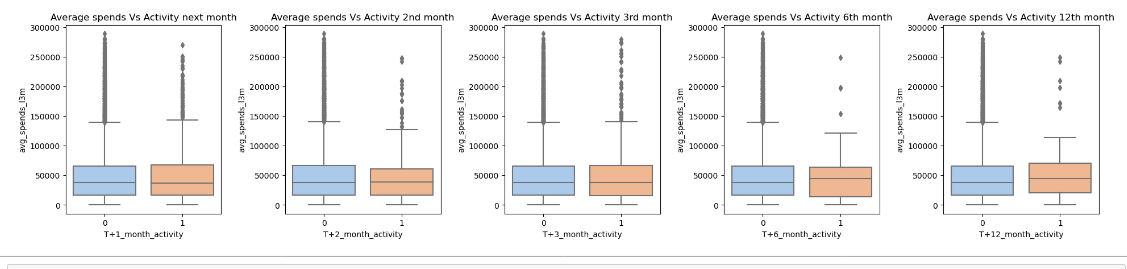
Vintage with Bank ranges from 6 months to 60 months

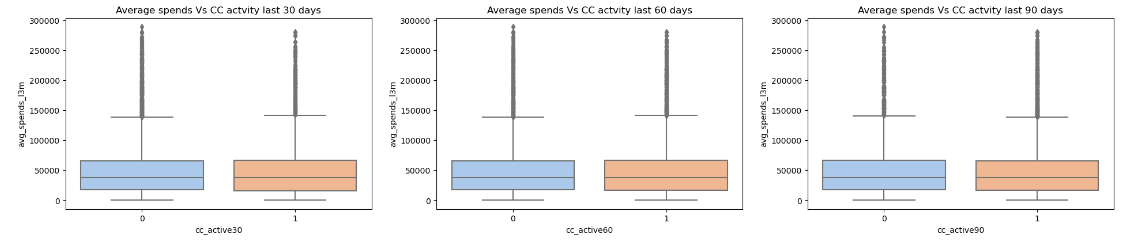
### Bi-Variate Analysis

#### Analysing Average spending on credit Card in last 3 months against different categorical variables:









1.People with net worth type A are spending the most on Credit Card. Type C and E are spending the least.

2.House wife and Retired people are spending least on Credit Card.

3.Revolvers are spending more on Credit Card than Transactors.

4.Number of Widget products the credit card holder has, no bearing on Amount spent on Credit Card

5. Number of Engagement products the credit card holder has, no bearing on Amount spent on Credit Card

6.Average Amount spent on Credit Card is not affected by holding other banks Credit cards

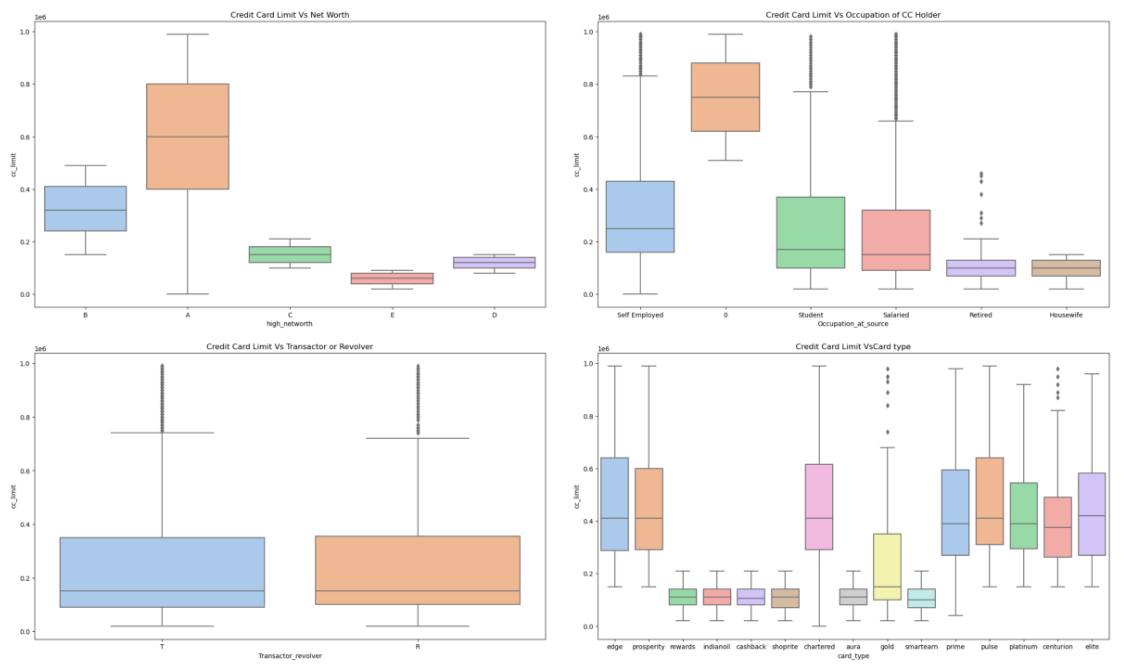
7.Card Holders with card type Edge, Prosperity, Centurian, chartered, Platinum, Prime, Prosperity, and Elite spending the most on CC

8. Average Amount spent on Credit Card is not affected by Issuer of credit card.

9.Future activity in next 12 months cannot be predicted based on Average spends in last 3 months

10.Credit card activity in the past 30,60 and 90 days is not dependent on Average spends in last 3 months

#### Analysing credit Card limit against different categorical variables:



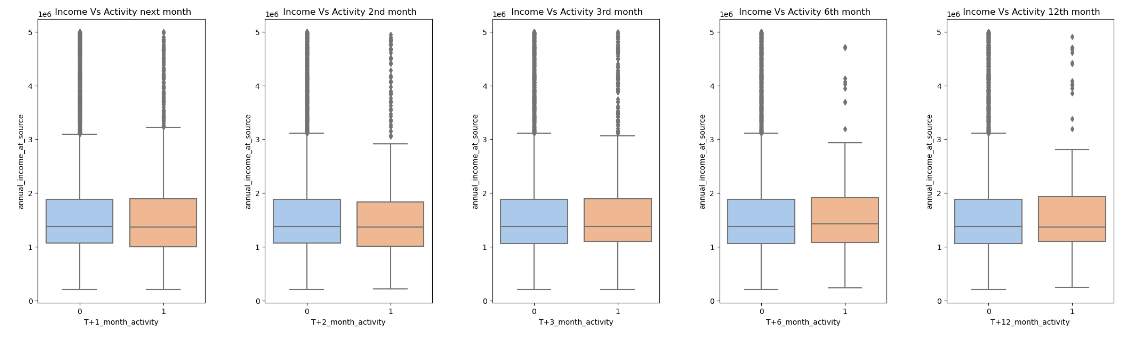
1. People with net worth type A are having the highest credit card Limit

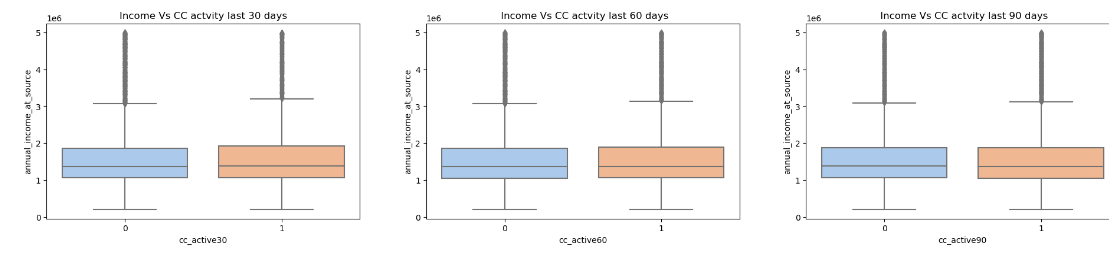
2.House wife and Retired people are having the least credit card Limit

3.Revolvers and Transactors have almost same credit limit.

4. Card Holders with card type rewards, Indian oil, cashback, shoprite, aura and smartearn have least credit limit **.**

#### Analysing Annual Income against Credit card activity and Future activity

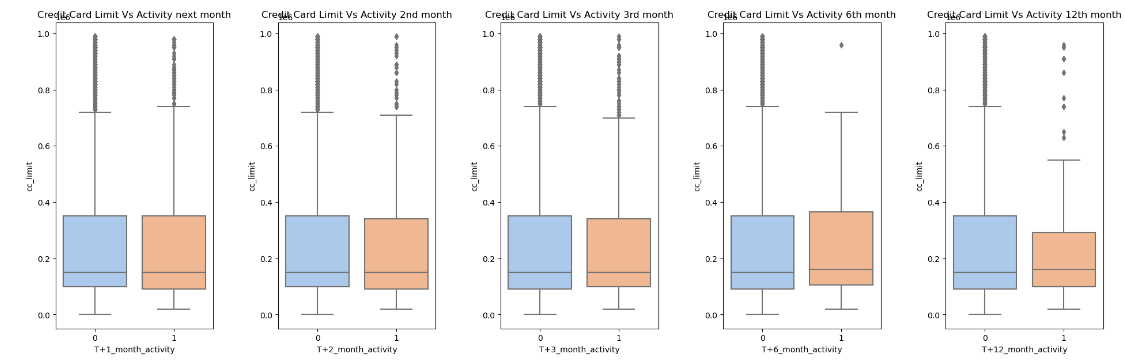


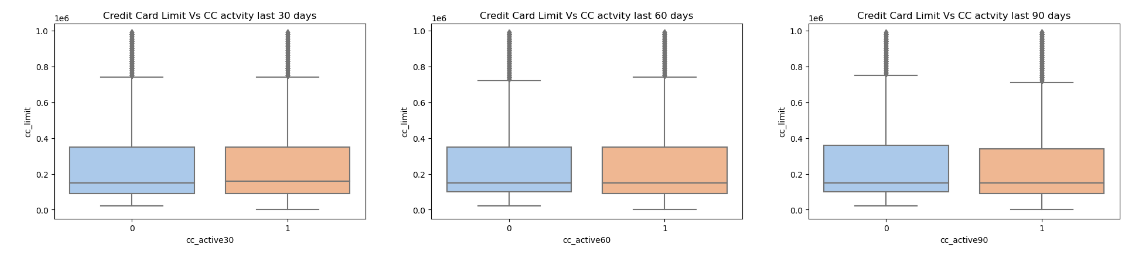
****

1.Future activity in next 12 months cannot be predicted based on Annual Income on the credit card holder

2.Credit card activity in the past 30,60 and 90 days is not dependent Annual Income on the credit card holder

#### Analysing Credit Card Limit against Credit card activity and Future activity

****

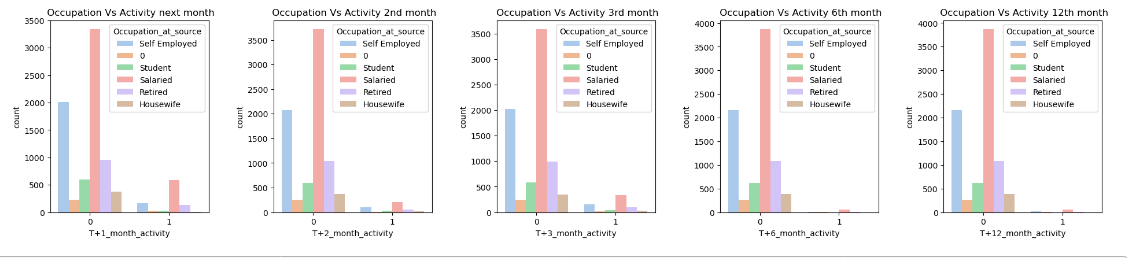
****

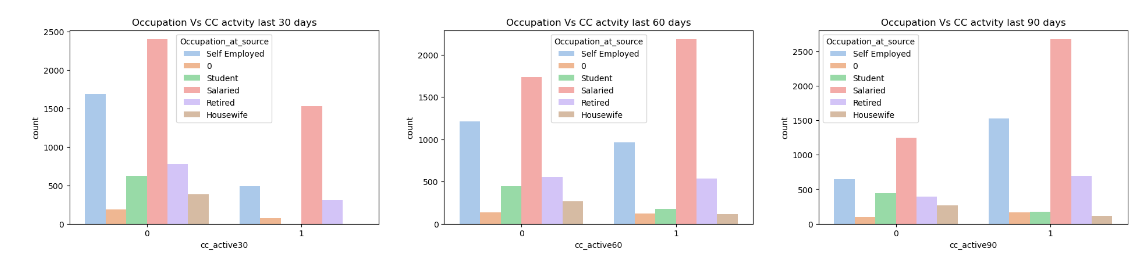
#### Insights drawn from Visualizations

1.Future activity in next 12 months cannot be predicted based on Credit card limit of the credit card holder

2.Credit card activity in the past 30,60 and 90 days is not dependent on Credit card limit of the credit card holder

#### Analysing Occupation against Credit card activity and Future activity

****

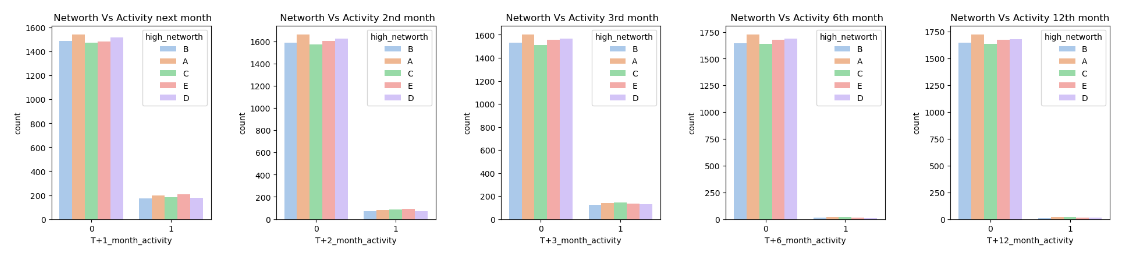
****

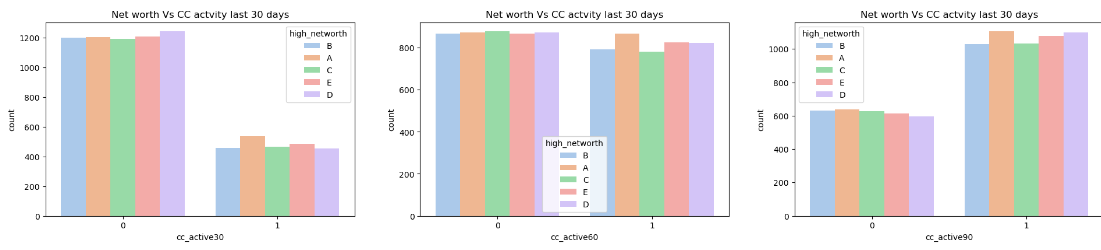
#### Insights drawn from Visualizations

1.Future activity in next 12 months is mostly done by Salaried professionals

2.Credit card activity in the past 30,60 and 90 days is mostly done by Salaried professionals

#### Analysing Net Worth against Credit card activity and Future activity

****

****

#### 

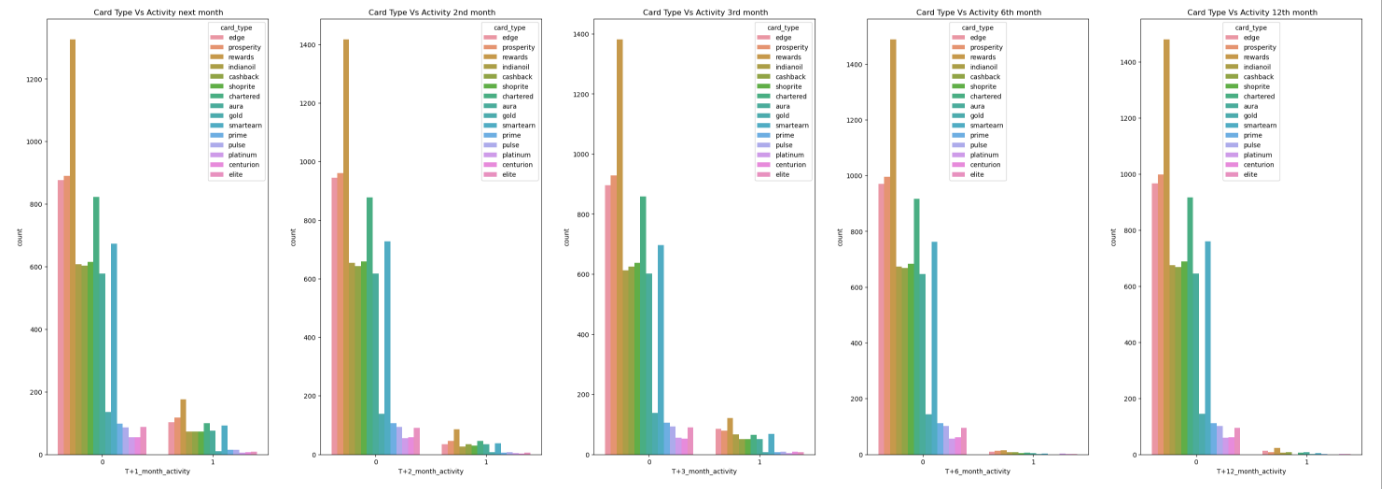
#### Insights drawn from Visualizations

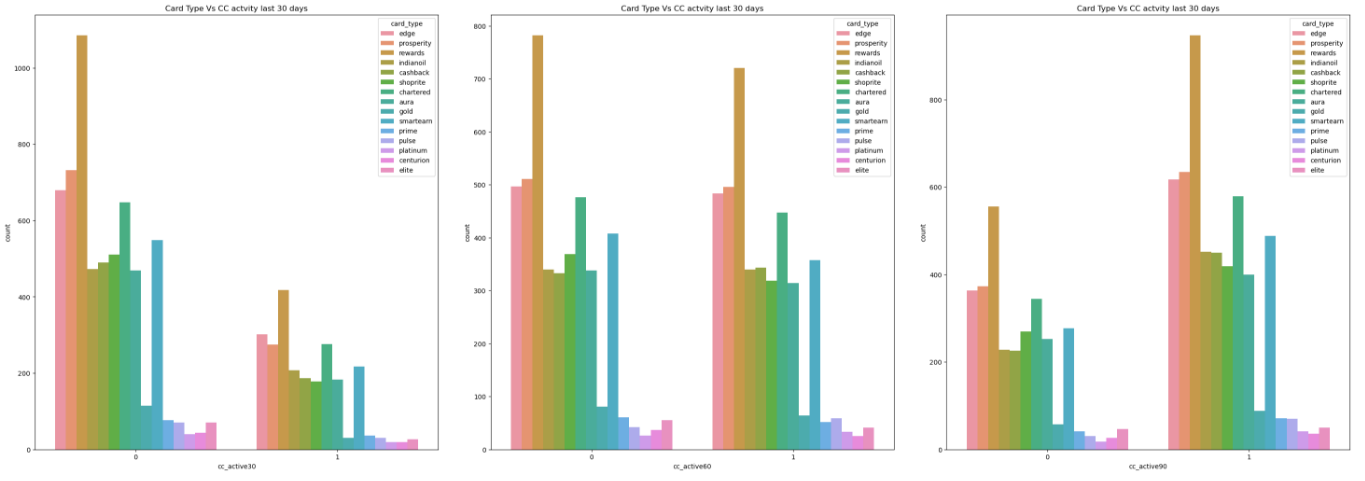
1.Networth of Card Holders is not impacting Future activity in next 12 months

2.Credit card activity in the past 30,60 and 90 days is not very dependent on Net worth of Card Holders

#### 

#### Analysing Card Type against Credit card activity and Future activity

****

****

#### Insights drawn from Visualizations

1.Card holders with Rewards card are showing more Future activity in next 12 months

2.Credit card activity in the past 30,60 and 90 days are mostly done by card holders with Rewards card

#### Analysing Transactor or Revolver against Credit card activity and Future activity

#### 

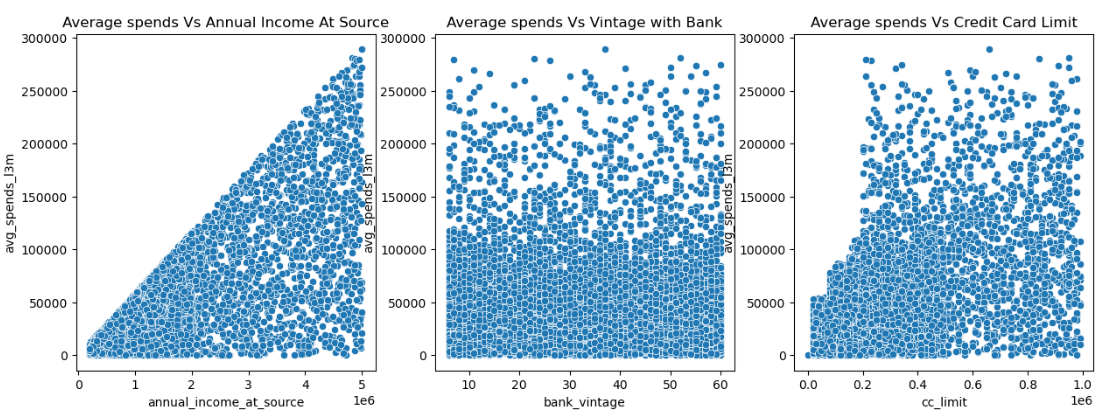
#### 

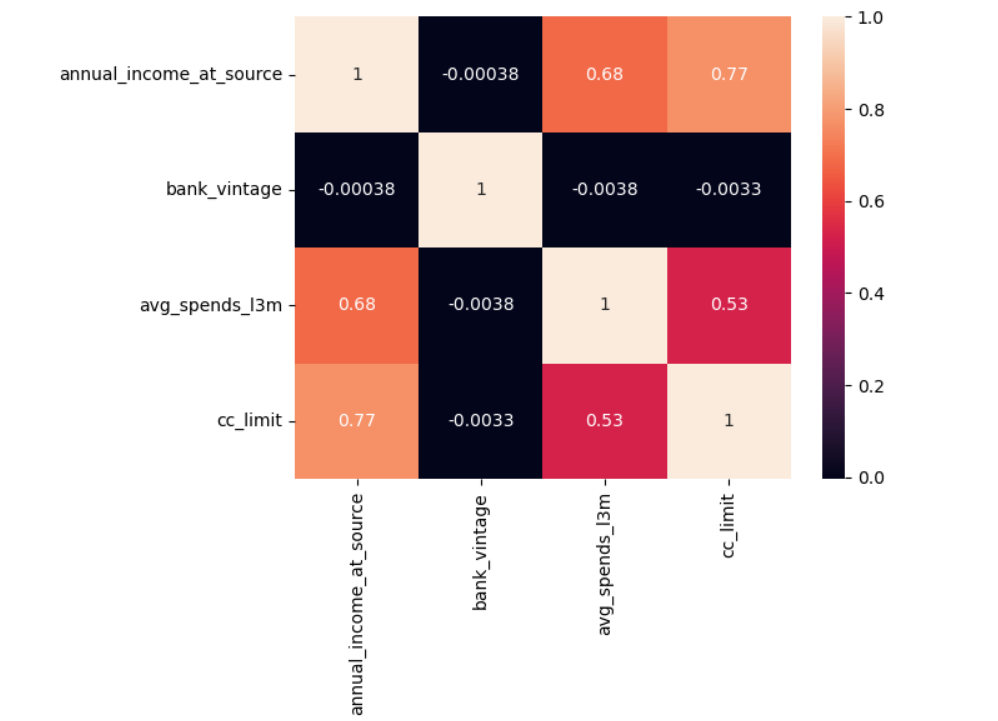
#### Insights drawn from Visualizations

1.Transactors are showing more Future activity in next 12 months

2.Credit Card activity is reduced drastically among Transactors in last 30 days as compared to Revolvers

#### Analysing Average spending on credit Card in last 3 months against different numerical variables

****

****

#### Insights drawn from Visualizations

1.Annual Income at source and Average Spending in Last 3 months has a positive correlation

2.Vintage with bank has no bearing on average spending on credit Card in last 3 months

3. Average spend in last 3 months and CC limit also show a positive correlation

### Multi-Variate Analysis

#### Analysing Average spending on credit Card in last 3 months against different numerical variables and categorical variables

#### 

#### 

#### Insights drawn from Visualizations

1.Retired and Housewives are having least income hence spending the least amount on credit card

2.Revolvers are spending more on credit card than transactors

3. Card Holders with card type Edge, Prosperity, Centurian, chartered, Platinum, Prime, Prosperity, and Elite are having higher income and spending the most on CC **.**

3. Card Holders with card type Edge, Prosperity, Centurian, chartered, Platinum, Prime, Prosperity, and Elite are having higher CC Limit and spending the most on CC **.**

## **Results**

From the various visualizations we can deduce that the top 5 variables to consider before recommending a card to a customer are **Occupation\_at\_source,**

**avg\_spends\_l3m, card \_type, annual\_income\_at\_source and Transactor\_revolver.** Other important variable could be **cc\_limit** .

1. Occupation is a very important variable because we see that Card holders who are retired and housewives are spending less on the credit card , whereas card holders who are salaried, self-employed ,students and Unknown profession are using the Credit card more

2. Card Type is an important variable to consider because as we can see from the visualizations, Rewards card is the most used card, but Card holders with card type Edge, Prosperity, Centurian, chartered, Platinum, Prime, Prosperity, and Elite spending the most on CC.

3. Annual Income at source is very important variable because, we get to see which income group spends the most . Card Holders in income group ₹13-16 lakhs seem to be using credit card more.

4. Revolvers are spending more on credit card than Transactors. This implies that revolvers are utilizing the credit card facility but also not paying the due amount . This may impact the business in future.

5.There is a positive correlation between credit card limit and average spending, so increasing credit limit may result in increase in average spending