
GRADED PROJECT SMDM



greatlearning
Learning for Life

Submitted by : Nupur Sarkar
PGP DSBA Feb'23-24

Austo Motor
Godict Bank

Table of Contents

1	Problem 1 Austo Motor Company	4
1.1	What is the important technical information about the dataset that a database administrator would be interested in? (Hint: Information about the size of the dataset and the nature of the variables)	4
1.2	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
1.3	Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business.....	6
1.4	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.	9
1.5	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.	10
	E1) Steve Roger says “Men prefer SUV by a large margin, compared to the women”	10
	E2) Ned Stark believes that a salaried person is more likely to buy a Sedan.....	10
	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.	10
1.6	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.....	12
	F1) Gender F2) Personal_loan	12
1.7	From the current data set comment if having a working partner leads to the purchase of a higher- priced car.	13
1.8	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	14
2	Problem 2: GODIGT BANK.....	15
2.1	Analyse the dataset and list down the top 5 important variables, along with the business justifications.	15

Table of figures

Figure 1	4
Figure 2	5
Figure 3	6
Figure 4	7
Figure 5	8
Figure 6	9
Figure 7	11
Figure 8	11
Figure 9	12
Figure 10	12
Figure 11	13
Figure 12	14
Figure 13	19
Figure 14	19
Figure 15	20
Figure 16	20
Figure 17	21
Figure 18	22
Figure 19	22
Figure 20	23
Figure 21	23
Figure 22	24
Figure 23	25
Figure 24	26

1 Problem 1 Austo Motor Company

In its recent board meeting, concerns were raised by the members on the efficiency of the marketing campaign currently being used.

The board decides to rope in analytics professional to **improve the existing campaign**.

- 1.1 What is the important technical information about the dataset that a database administrator would be interested in? (Hint: Information about the size of the dataset and the nature of the variables)

This dataset(austo_automobile+(2)+(1).csv) is a dataset of Austo Motor Company , a leading car manufacturer specializing in SUV, Sedan, and Hatchback models. It has total 22134 observations which describes 1581 customer records. Each record consists of 14 features which describes various details of customers related to the past purchase of cars by customers. More detailed explanation about these 14 features is provided by the below Data Dictionary. Please refer the Data dictionary below for detailed explanation .

Data Dictionary:
Age: Shows the age of the customer
Gender : Shows the gender of the customer
Profession : Shows the type of occupation the customer belongs to ,in the above 5 rows we can see two of them i.e Business or salaried .but through further analysis we will find each label of this.
Marital_status : customer is Married or single
Education : Level of Education customer has
No_of_Dependents :No of members in customer's Family
Personal_loan : Is there any personal_loan running for customer
House_loan :Is there any house loan running for customer
Partner_working : Customer's partner is working or not
Salary : Salary the customer is earning
Partner_salary : Customer's partner's Salary
Total_salary :salary+Partner_Salary (Given in point 2 in 'Answer your doubts')
Price : Price of the car
Make : Type of car model , SUV shown in the above ,but we will find the other labels in the coming code .

Figure 1

- 1.2 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.

Preliminary Analysis of the Variables include following steps:

1. Detecting duplicates in the dataset and removing them:

There are no duplicates in this dataset

2. Check the datatype of each variable in dataset:

Total 14 variables out of which 8 variables (Categorical) & 6 variables (Numeric)

3. Statistical Analysis

a. Numerical Variables

From the description above let's see Numerical features first Individually :

Age 🗓️ Age of the customers in this data set varies from 22 to 54 and the average age is 31.922 whereas median age is 29 and since the mean > median here so, the data would be **right skewed** and that means majority of data would be towards the left side of the data with some outliers on the right. We will find these outliers further in visualization step

No_of_Dependents 👨‍👩‍👧 the no of family members customer can have is in the range of 0 to 4 , where again we see , mean > median, data will be **right skewed**, that shows majority of customers will have less than 2 No_of_Dependents. This column may have outliers

Salary 💰 Customer earns in the range of 30000 to 99300 but we see that the average salary of customers is 60392.220114 , which is less than the median salary i.e 59500.0 so, the data is again **right skewed** and so majority of the customers will have salary distribution to the left side of the data

Partner_Salary 💰 Customer's Partner's Salary varies from 0 to 80500 , where 0 says that the customer's partner is not earning, these are those cases where customers don't have working partners and that's the reason average Partner_salary is 19233.776091 and Median Partner_salary is 25100 , here median > mean , so the data is **left skewed** and majority of the datapoints will be to the right side of the distribution . This column may have outliers

Total_salary 🏠 Customer's Total_salary i.e Salary + Partner_salary varies from 30000 to 171000 , where mean(79625.996205) > median(78000.0) , so the data is **right skewed**

Price 🚗 Customer's buying capacity the cars ranges from 18000 to 70000. Here mean(35597.722960) > median(31000.0) So, Price distribution is also **right skewed**

Figure 2

b. Categorical Variables :

On the basis of labels

Binary: Variables with two labels

Profession, Marital_status, Education , Personal_loan, House_loan Partner_working

Multi-level: Variables with more than two labels

Make and Gender (But as we know that Gender can have either male or Female so lets check the Bad Data in this Variable in further Analysis)

4. Detecting bad values and missing Values

Columns 'Gender' and 'Partner_Salary' has null values. We can see that 106 customers probably do not have any Partner_salary values and 53 Gender column values are also null. Spelling of 2 'Female' values in the 'Gender' column, these were misspelled as 'Femal' and 'Femle' earlier.

5. Treating the missing Values and bad values

53 values in female Variable with the mode of this Variable since it's a non-numeric variable and corrected spelling of 2 'Female' values in the 'Gender' column. These were misspelled as 'Femal' and 'Femle' earlier.

Imputed all 106 missing Partner_Salary with 'Total_Salary – salary'

1.3 Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business.

Univariate Analysis: Analysis of one variable at a time Visually
Numerical Variables:

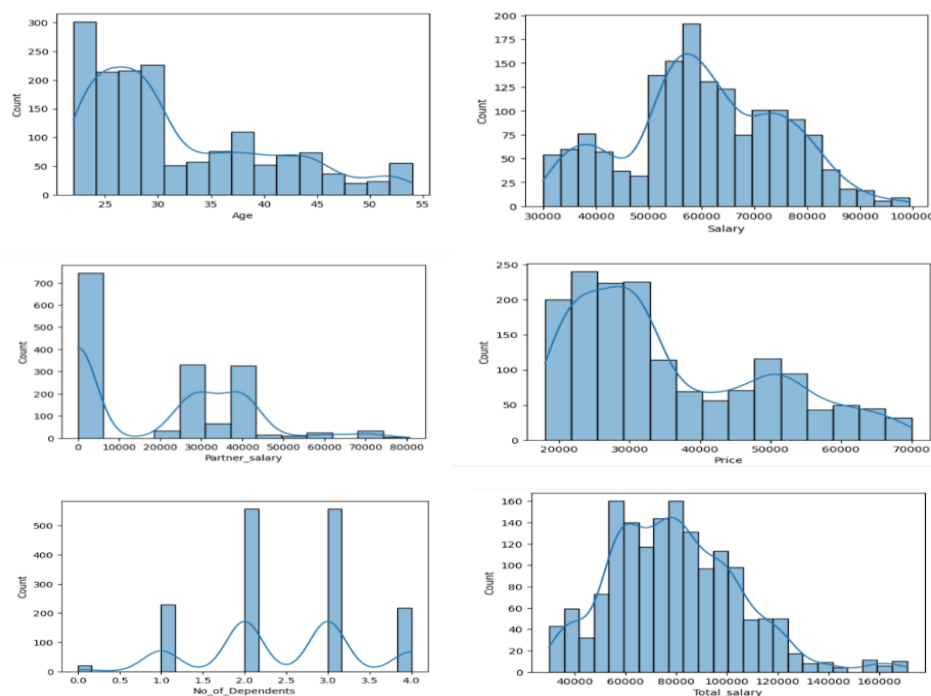


Figure 3

Numerical Variables Visual Insights

- ▶ Majority of Customers in this dataset falls in the age group of 25 to 38, this age group is where the selling has happened maximum.
- ▶ Customers who have 2-3 family members are the majority buyers in this dataset
- ▶ Majority of customers salary varies from 51900 to 71800
- ▶ Our customers have their partners earning from the range of 0 to 38300 and thats because many customers don't have working partners
- ▶ Selling is mostly happening for the cars ranging from 25000 to 47000

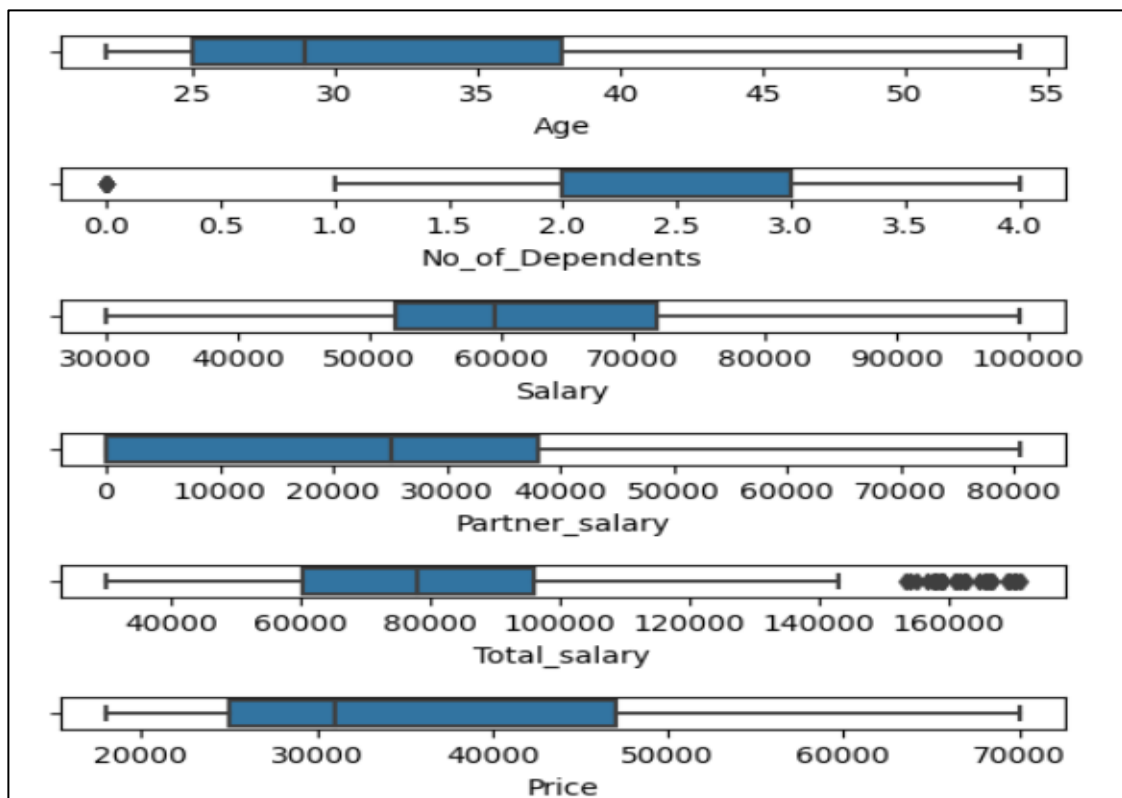


Figure 4

- ▶ Variables Total_salary and No_of_Dependents have outliers

Categorical Variables:

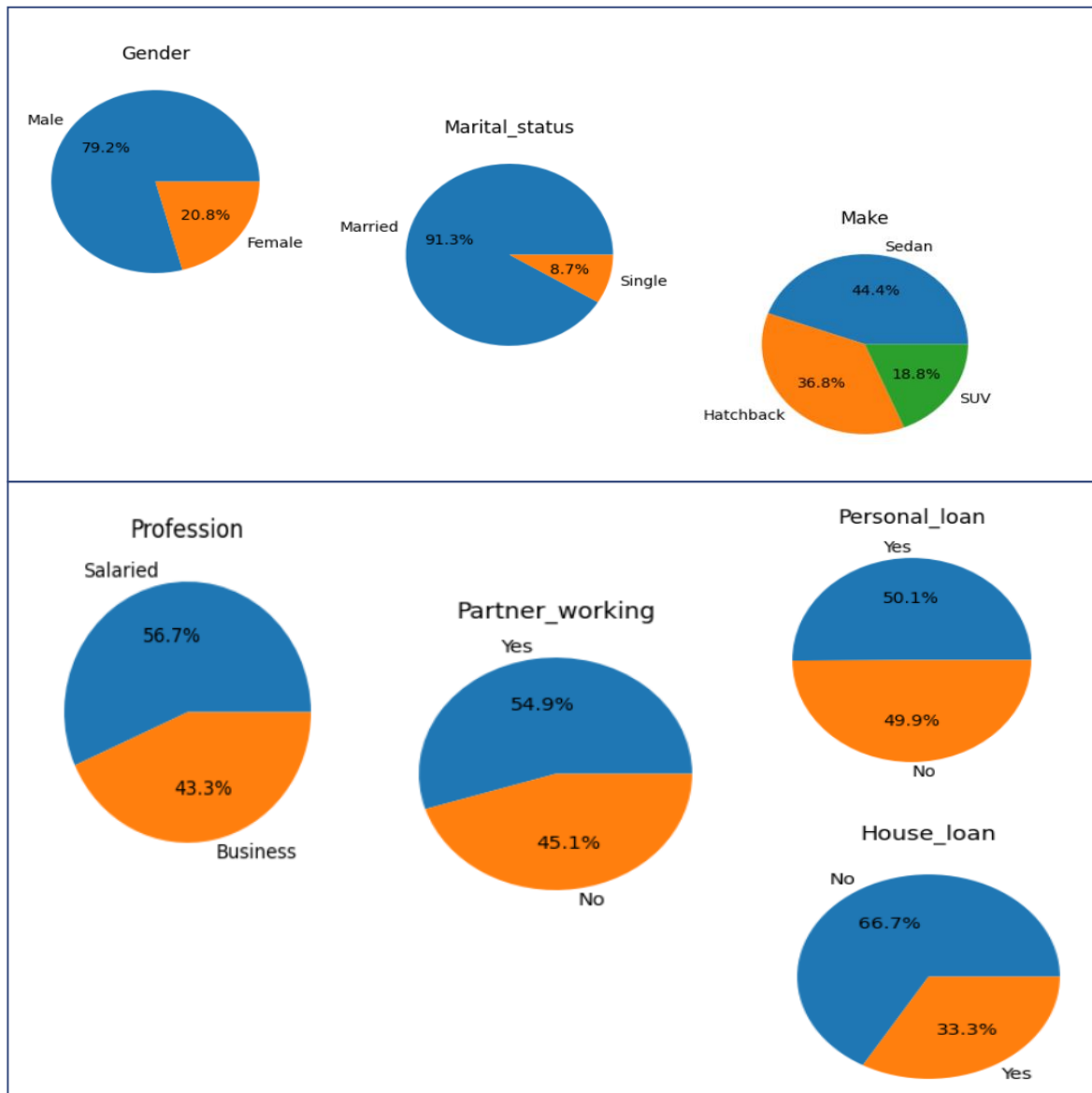


Figure 5

Categorical Variable Visualization Insights 🧐

- ▶ 79.2% of customers are Male and 20.8% are females, so we understand that Men are the major buyer customer segment in this dataset.
- ▶ More salaried customers are here as compared to customers doing Business.
- ▶ 50.1% customer who have loan running
- ▶ 66.7% customers don't have any house loan running.

- ▶ 91.3% customers are married
- ▶ 54.9% customers have working partners
- ▶ Buying preference of customers make wise is:
Sedan > Hatchback > SUV

1.4 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.



Figure 6

Numerical Bi- Variate Analysis Insights 🧐

+ve strong correlation or when increase in one variable will lead to increase of other and vice versa is observed in below variables:

- ▶ Age and Price So, As the age of the customers increase, the price of cars which they prefer buying also increases and vice versa
- ▶ Total_salary and Partner_salary as the salary of customer's partner increases, customer's salary also increases and vice versa
- ▶ Age and Salary as the age of the customers increase, the salary of customers also increases and vice versa

Bi Variate Analysis b/w 1 Numerical and 1 Categorical Variable:
Insights 🧐

- ▶ Middle 50% of the women have age range between 33 to 45 whereas men has 25-32
- ▶ Men has majority when it comes to more No_of_dependents than women
- ▶ Women has better salary range as compared to Men
- ▶ Women group has better combined salary with their partners
- ▶ Women prefer high priced cars when compared to men
- ▶ Salaried customers age group is on little higher end as compared to business customers
- ▶ Married people earn more than singles
- ▶ Married customers prefer high priced cars as compared to Singles
- ▶ Postgraduates earn more than Graduate with or without their partner's earning
- ▶ Customers with no house_loan liability go for high priced cars
- ▶ Price order of Make SUV > Sedan > Hatchback
- ▶ Hatchback is preferred by customers between age group 22 - 27
- ▶ Sedan is preferred by customers between age group 26-37
- ▶ SUV is preferred by customers between age group 39 - 50

1.5 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”

E2) Ned Stark believes that a 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.

E1) Steve Roger says “Men prefer SUV by a large margin, compared to the women”

Solution ▶ But according to the Analysis in the below graph. We can see that female prefer SUVs by larger margin

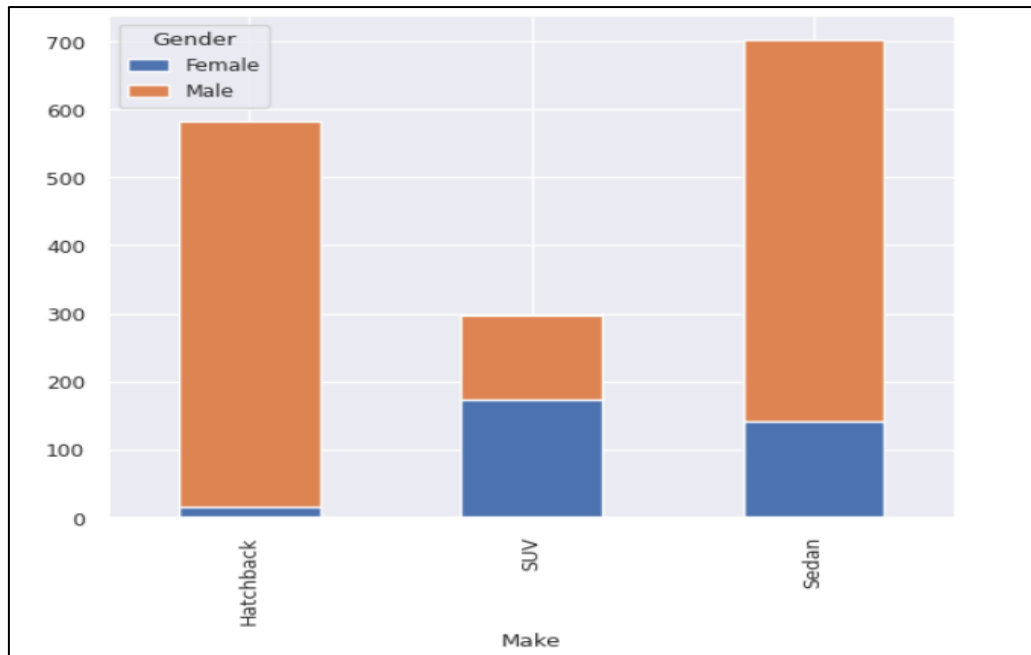


Figure 7

E2) Ned Stark believes that a salaried person is more likely to buy a Sedan

Solution ► Ned is right as we can see through the Bi Variate Analysis below graph.

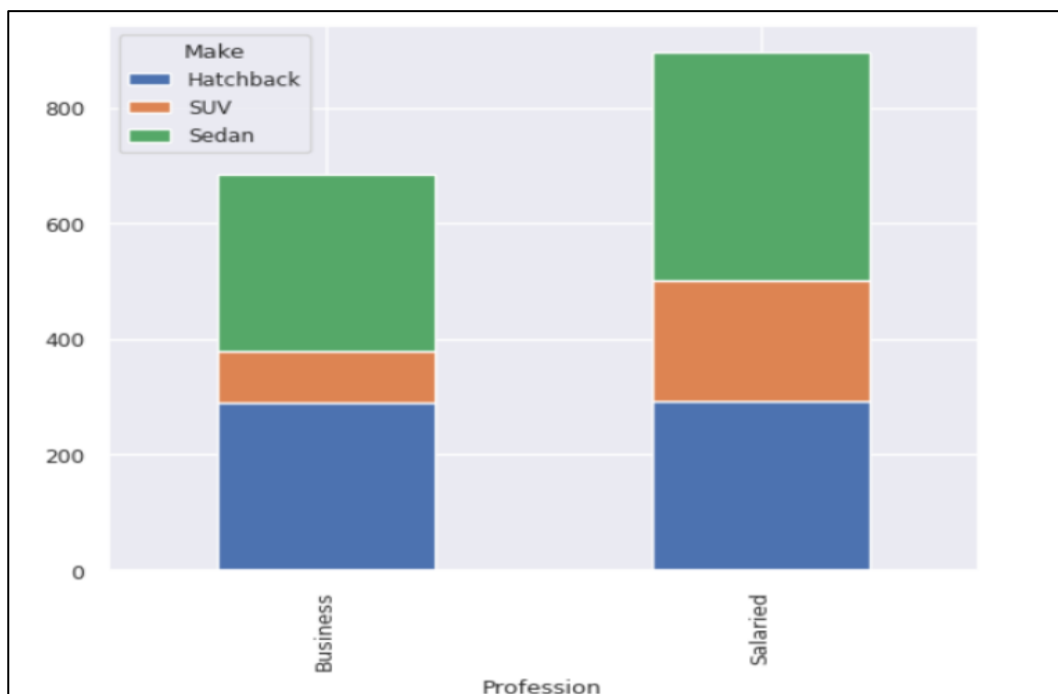


Figure 8

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

Solution► Sheldon's Claim is wrong. Salaried Male prefer Sedan more than SUV as shown in the below graph and Salaried Male are an easier Target for Sedan sale over SUV

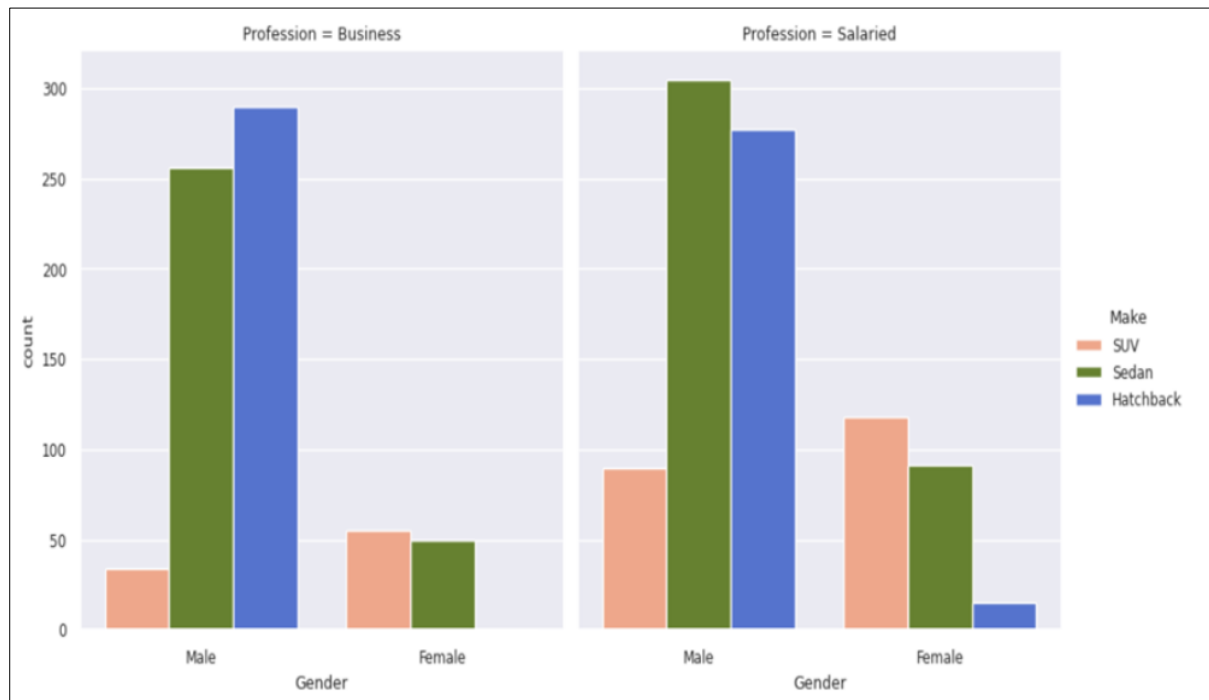


Figure 9

1.6 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.

F1) Gender F2) Personal_loan

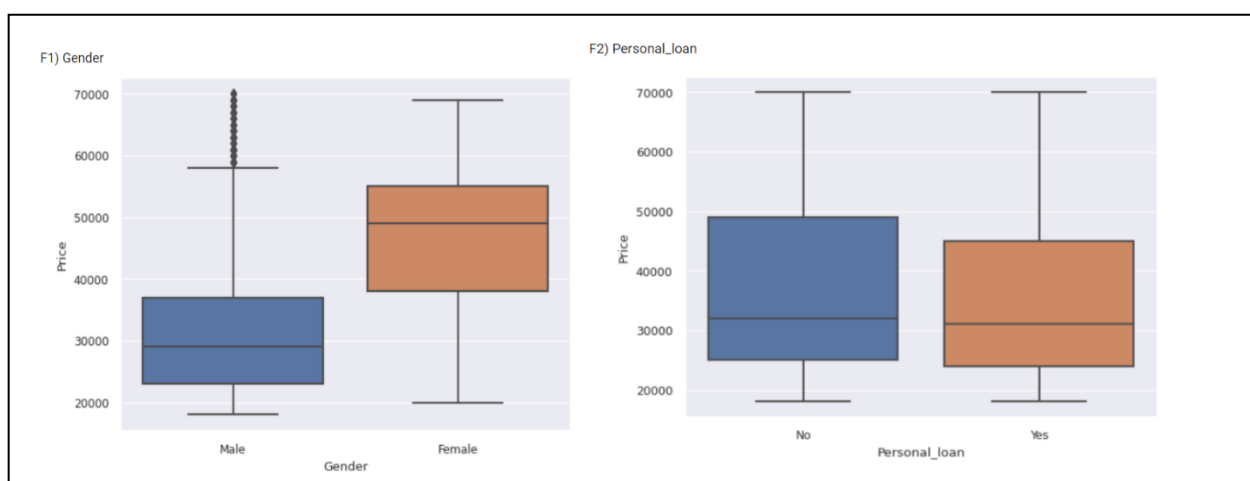


Figure 10

Insights 👁👁

- ▶ Our Target should be female customers for cars with higher price range
- ▶ We see those customers having no Personal_loan running prefers the high-priced vehicle, so we can target those customers who have no personal_loan running when comes to selling higher range vehicles

According to the above visual, Females are spending more on automobiles as compared to males and more of them have no personal loans running.

1.7 From the current data set comment if having a working partner leads to the purchase of a higher- priced car.

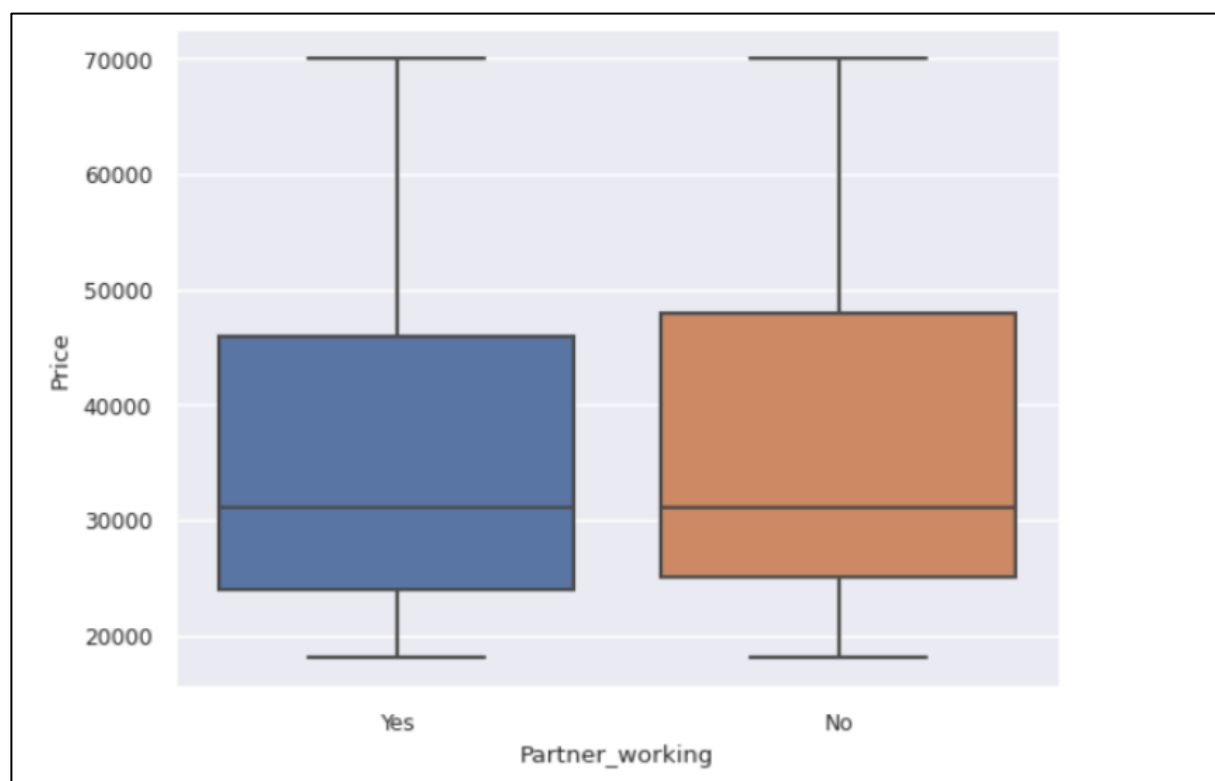


Figure 11

Insights

- ▶ No, having a working partner does not lead to the purchase of cars on higher price range by customers as we can see in the above graph.

1.8 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

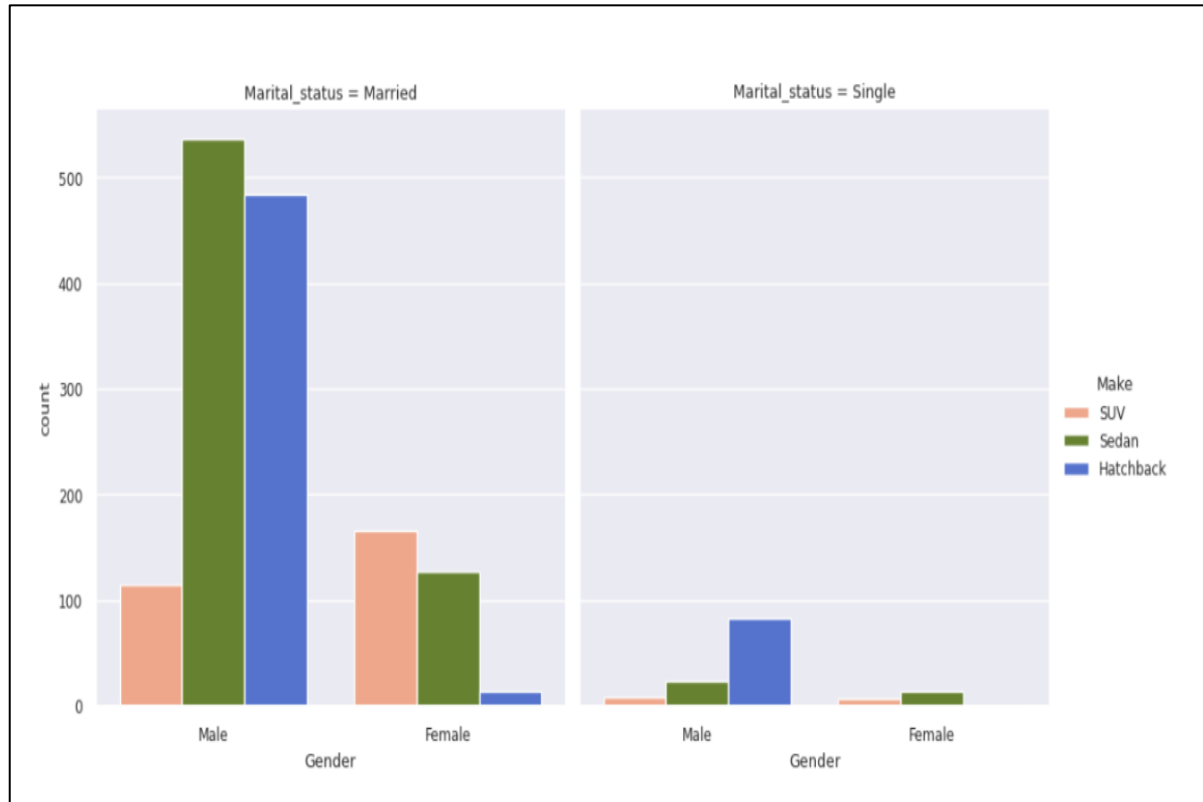


Figure 12

Suggestions for better Marketing Campaign with similar purchase history to target customers in the future marketing Campaigns 🚗

- ▶ Target married men customers for the make Sedan or Hatchback
- ▶ Target married women customers for the make SUV or Sedan
- ▶ Very few singles go for cars and majority of single men shows inclination towards Hatchback and single women, hardly purchase much and majority of them go for Sedan if they purchase. So, for single Men Hatchback must be the selling Targets and for single women Sedan must be the selling Targets

2 Problem 2: GODIGT BANK

2.1 ***Framing An Analytics Problem*** Analyse the dataset and list down the top 5 important variables, along with the business justifications.

Problem Statement

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. The bank makes money only if customers spend more on credit cards. Given the attrition, the Bank wants to revisit its credit card policy and make sure that the card given to the customer is the right credit card. The bank will make a profit only through the customers that show higher intent towards a recommended credit card. (Higher intent means consumers would want to use the card and hence not be attrite.)

Objective

Analyse the data using statistical analysis or by using plots. The bank wants **to make profits but decrease attrition rate**. List down top 5 features which are really important for the problem statement. Give detailed justification for selecting those 5 variables.

Initial Thought before Solving the problem:

To determine the top 5 most important features in the given dataset for maximizing the profit of Godict Bank while decreasing attrition rate, we need to consider the following factors:

- **Correlation with Profit:** The features that have a higher correlation with the profit of the bank are likely to be more important.
- **Correlation with Attrition:** The features that have a higher correlation with the attrition rate are likely to be more important.
- **Actionability:** The features that can be directly controlled or influenced by the bank to decrease the attrition rate and increase the profit are more valuable.
- **Cost-effectiveness:** The features that can be modified or optimized with minimum cost and effort are more desirable

Data Information

Variable	Desc
userid	Unique bank customer id
card_no	Masked credit card number
card_bin_no	Credit card IIN number
Issuer	Card network issuer
card_type	Credit card type
card_source_date	Credit card sourcing date
high_networkth	Customer category basis their networkth value (A: High to E: Low)
active_30	Savings/Current/Salary etc account activity in last 30 days
active_60	Savings/Current/Salary etc account activity in last 60 days
active_90	Savings/Current/Salary etc account activity in last 90 days
cc_active30	CC activity in last 30 days
cc_active60	CC activity in last 60 days
cc_active90	CC activity in last 90 days
hotlist_flag	Whether card is hotlisted
widget_products	Number of convenient product customer holds (dc, cc, netbanking active, mobile banking active, wallet active etc)
engagement_products	Number of investment/loan product customer holds (FD, RD, Personal loan,
annual_income_at_source	Annual income recorded in credit card application
other_bank_cc_holding	Hold other bank credit card
bank_vintage	Vintage with the bank (in months) as on Tth month
T+1_month_activity	Customer spends next (T) month using credit card
T+2_month_activity	Customer spends in T+2 month using credit card
T+3_month_activity	Customer spends next month using credit card
T+6_month_activity	Customer spends next month using credit card
T+12_month_activity	Customer spends next month using credit card
Transactor_revolver	Revolver: Customer who carries balances over from one month to the next.
avg_spends_3m	Average credit card spends in last 3 months
Occupation_at_source	Occupation recorded at the time of credit card application
cc_limit	Current credit card limit

*All above data has been recorded as on Tth month excluding T+1_month_activity, T+2_month_activity, T+3_month_activity, T+6_month_activity, T+12_month_activity

look at the type of data that has been made available to us for analysing
Let's take a look at the datatypes of the variables

We have Total 8448 rows and 28 Variables in this Dataset (godigt_cc_data.xlsx)

Categorical Variables:

'card_no', 'Issuer', 'card_type', 'high_network', 'hotlist_flag', 'other_bank_cc_holding',
'Transactor_revolver', 'Occupation_at_source'

Continuous Variables:

'userid', 'card_bin_no', 'card_source_date', 'active_30', 'active_60', 'active_90', 'cc_active30',
'cc_active60', 'cc_active90', 'widget_products', 'engagement_products',
'annual_income_at_source', 'bank_vintage', 'T+1_month_activity', 'T+2_month_activity',
'T+3_month_activity', 'T+6_month_activity', 'T+12_month_activity', 'avg_spends_l3m',
'cc_limit'

Features that can have higher correlation with profit :

- Annual_income_at_source : More the creditworthiness of customer , more the bank can make profit by offering more credit limit and better interest rates and credit products to these customers.
- Occupation_at_source : The more the stable is occupation of customer , better the customer's creditworthiness and hence same as above point
- Transactor_revolver : Customers who have high income but carry forward their due to next month are actually a source of better revenue as these customers will be paying more interest rates due to late payments done by them.

Features that can have higher correlation with attrition:

- 'active_30', 'active_60', 'active_90', 'cc_active30', 'cc_active60', 'cc_active90',
- engagement_products
- cc_limit

Some Variables like 'card_no', 'card_bin_no', 'userid' will not have any effect on the dependent variable 'avg_spends_l3m'. Hence we drop these three variables as these have the least importance

Some questions that can be raised initially that can act as a starting point to analyse the dataset

- Among all the cards which has the highest customer numbers? Which card type is sold the most?
- Customers prefer which issuer the most in this dataset and which card type has the highest percentage of that issuer in it?
- Percentage of customers high net worth category wise? Also check in which occupation they are in and how much is the earning? On the basis of occupation distribute the customers's card type preference.
- Percentage of customers active on their credit card last 30 days, last 60 days and last 90 days and compare among these 3 buckets where customer activity is highest and where its lowest?
- How much percentage of customers are holding more no of engagement products? Because more the products they are involved in ... more possibility is there that they lose the interest in cc usage because they have other obligations also to pay ... We need to find insights from this variable.
- What percentage of customers have other bank credit cards; chances are there that they are using that card more ...?
- Percentage of customers who carry forward their balance from one month to next... More customers will generate more profit that way as these customers will tend to pay more interest to bank.
- Order of average spend last3 months network and occupation wise?
- Compare the cc_limit provided to customers and what is their average spend last3 months card wise?

Among all the cards which has the highest customer numbers? Which card type is sold the most?

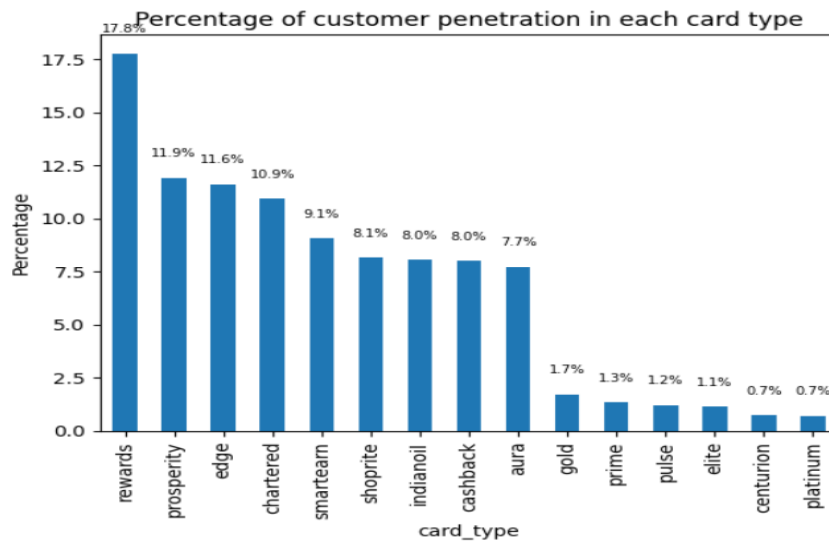


Figure 13

Insights

Rewards has the maximum customer penetration whereas platinum and centurion has the least penetration.

Customers prefer which issuer the most in this dataset and which card type has the highest percentage of that issuer in it?

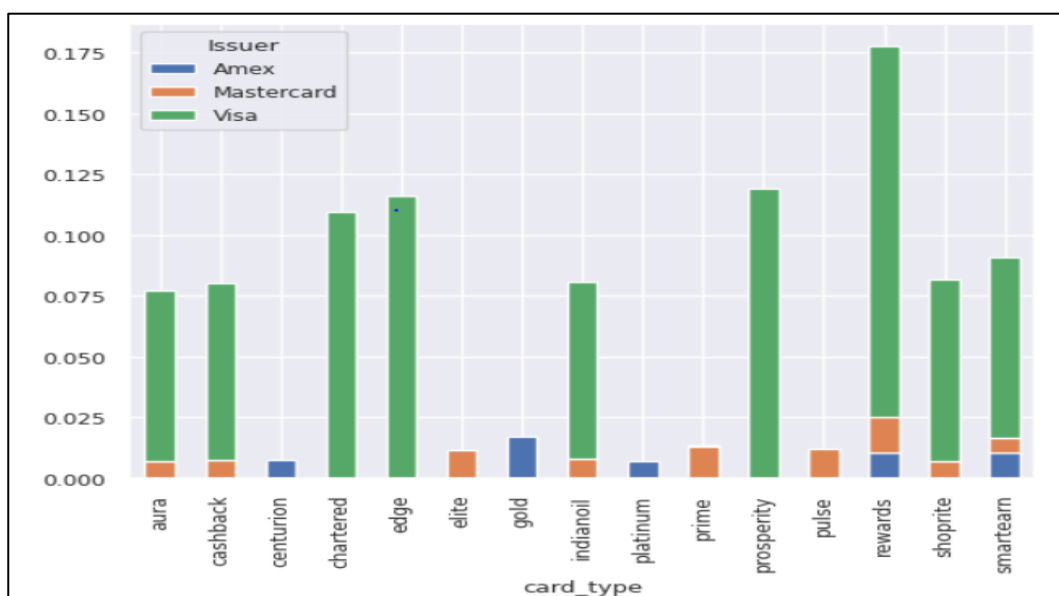


Figure 14

Insights Visa is been sourced the most or is the prominent issuer in majority of cards and prosperity card has highest penetration of it.

Percentage of customers high network category wise? Also check in which occupation they are in and how much is the earning?

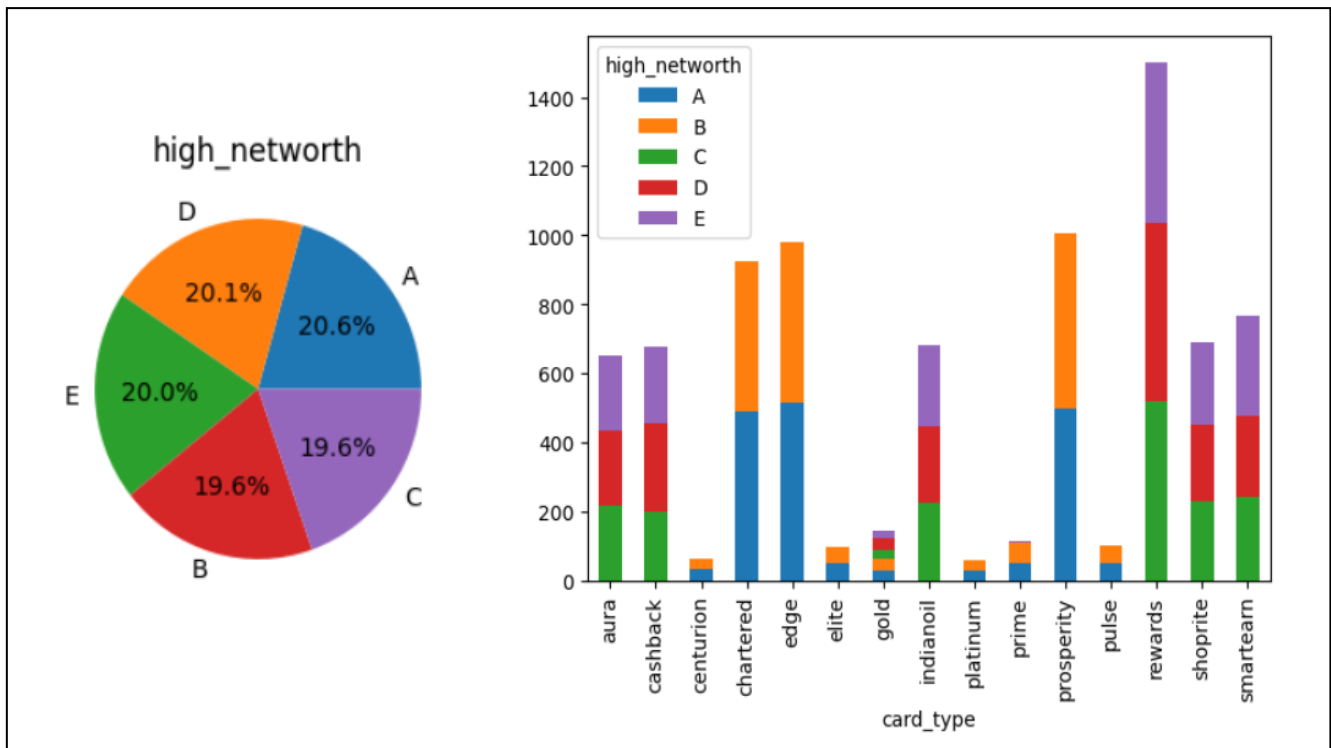


Figure 15

Insights Rewards card has the maximum customer penetration of C > D & E

Followed by prosperity card which has maximum customer penetration A > D edge which has maximum customer penetration of A and B

Followed by Chartered which has maximum customer penetration of A and B

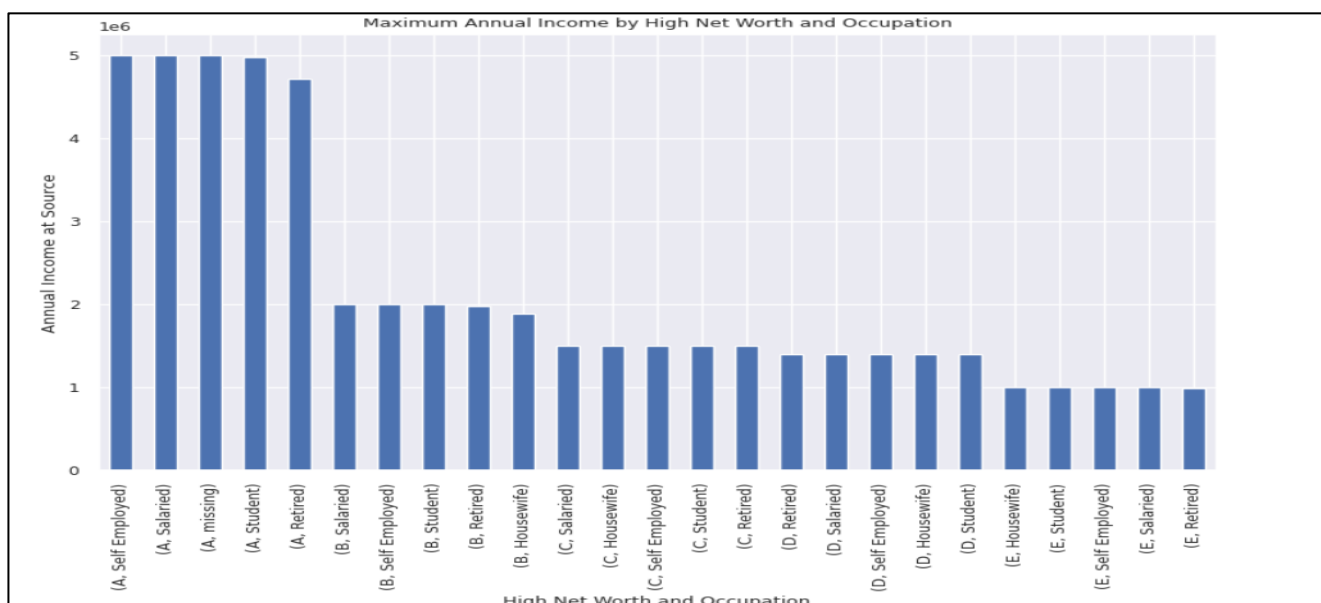



Figure 16

Insights  Earning Order of the customers

A(SE SAL MISSING STUDENT) > A(RETIRED)>B(SE SAL STUDENT RETIRED) > B(House wife)>C(SAL HOUSEWIFE SE STUDENT RETIRED) > D >E

So, annual_income_at_source is an important variable to understand the distribution of customers in dataset according to their salary bracket

Percentage of customers active on their credit cards last 30 days, last 60 days and last 90 days and compare among these 3 buckets where customer activity is highest and where its lowest?



Figure 17

Insights  cc activity is highest in last 90 days , these customers will be definitely the one .

These customers are the most active one and hence will generate the most revenue , so these customers must be thought of providing upgrades or other vouchers ,etc to stop them from Attrition or going to other bank.

How much percentage of customers are holding more no of engagement products? Because more the products they are involved in ... more possibility is there that they lose the interest in cc usage because they have other obligations also to pay ... We need to find insights from this variable.

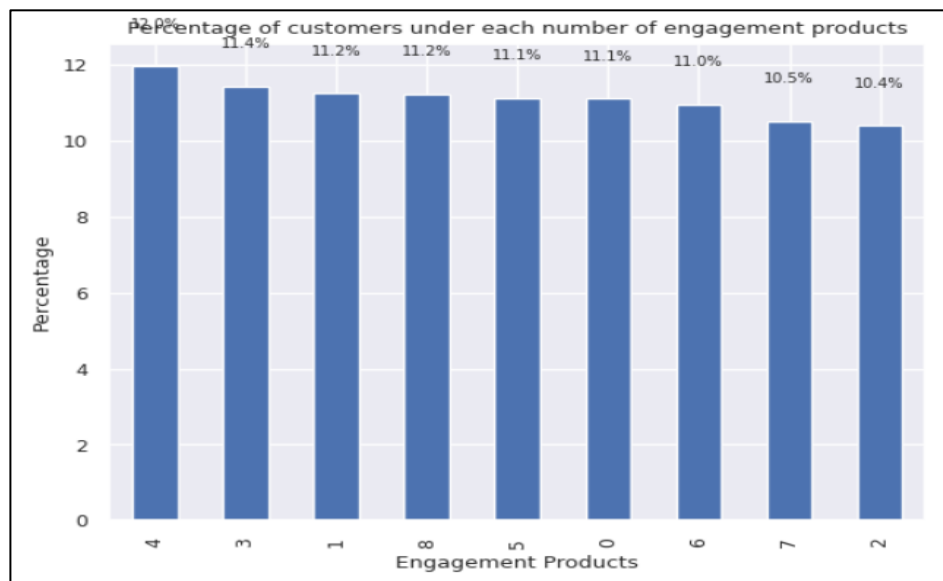


Figure 18

Insights 00 12% of customers have 4 more products engagement, that means loyalty with bank, means these customers must be offered certain offers, so that they don't attrite

What percentage of customers have other bank credit cards; chances are there that they are using that card more ...?

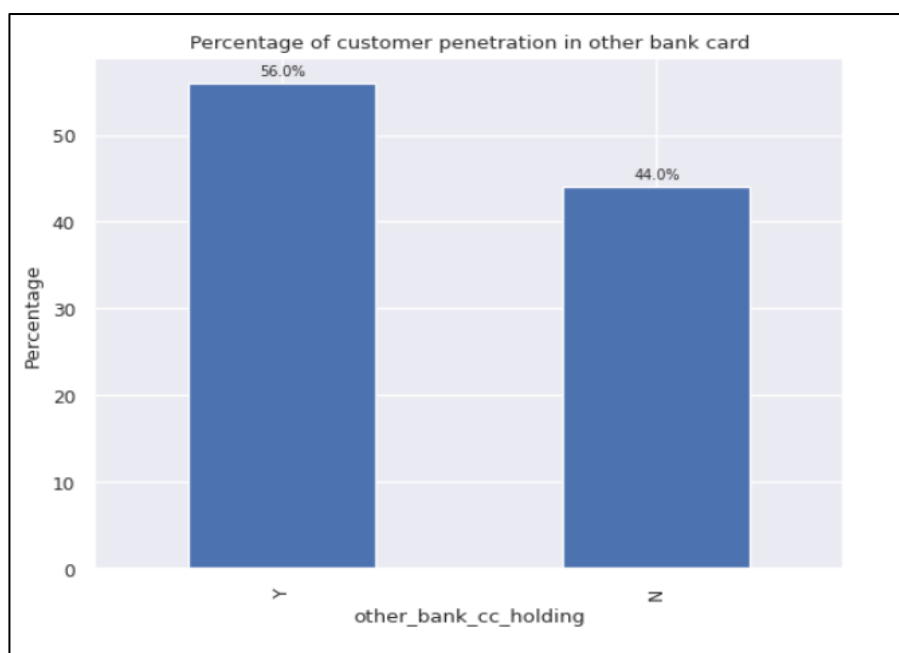


Figure 19

Insights 56.0% customers prefer credit cards from other banks as well

Bi-Variate Analysis

Correlation between variables:

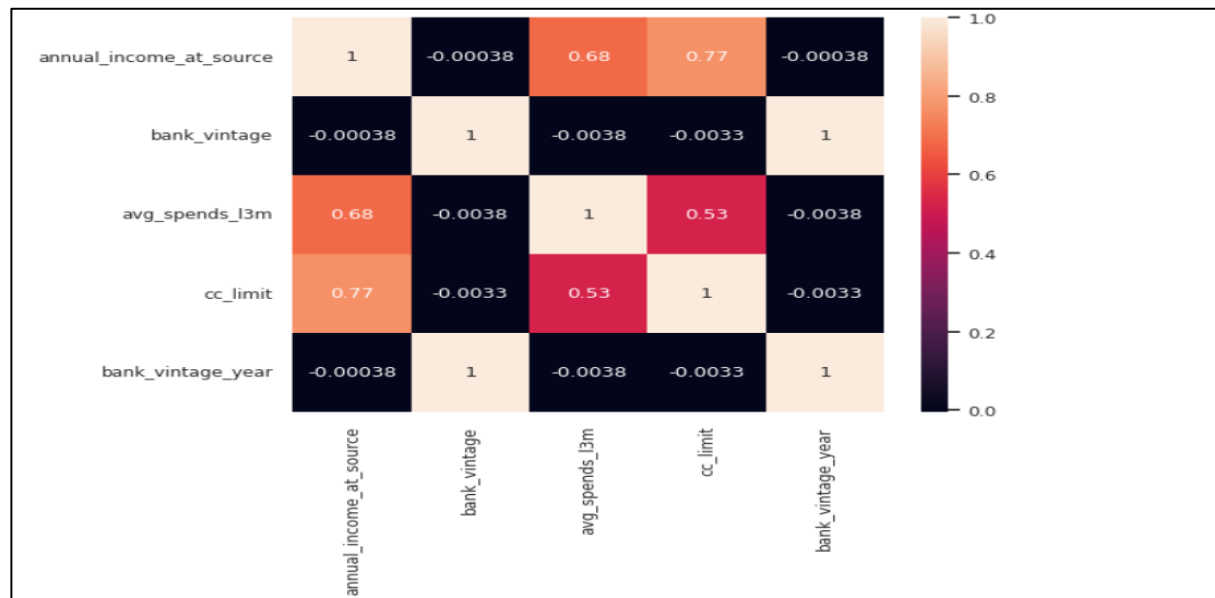


Figure 20

Observation

- Strong positive correlation between annual_income_at_source and avg_spends_l3m
- Strong positive correlation between annual_income_at_source and cc_limit
- medium positive correlation between cc_limit and avg_spends_l3m

Concentration of customers who carry forward their balance from one month to next. More customers will generate more profit that way as these customers will tend to pay more interest to bank

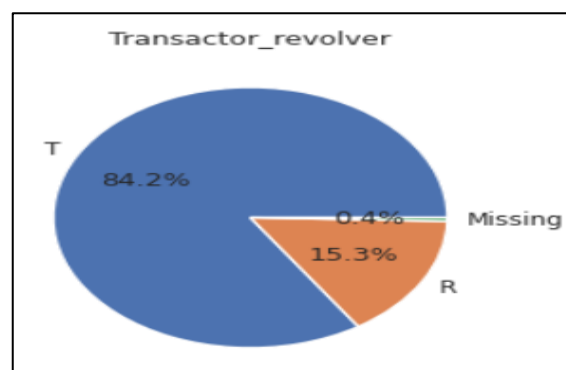


Figure 21

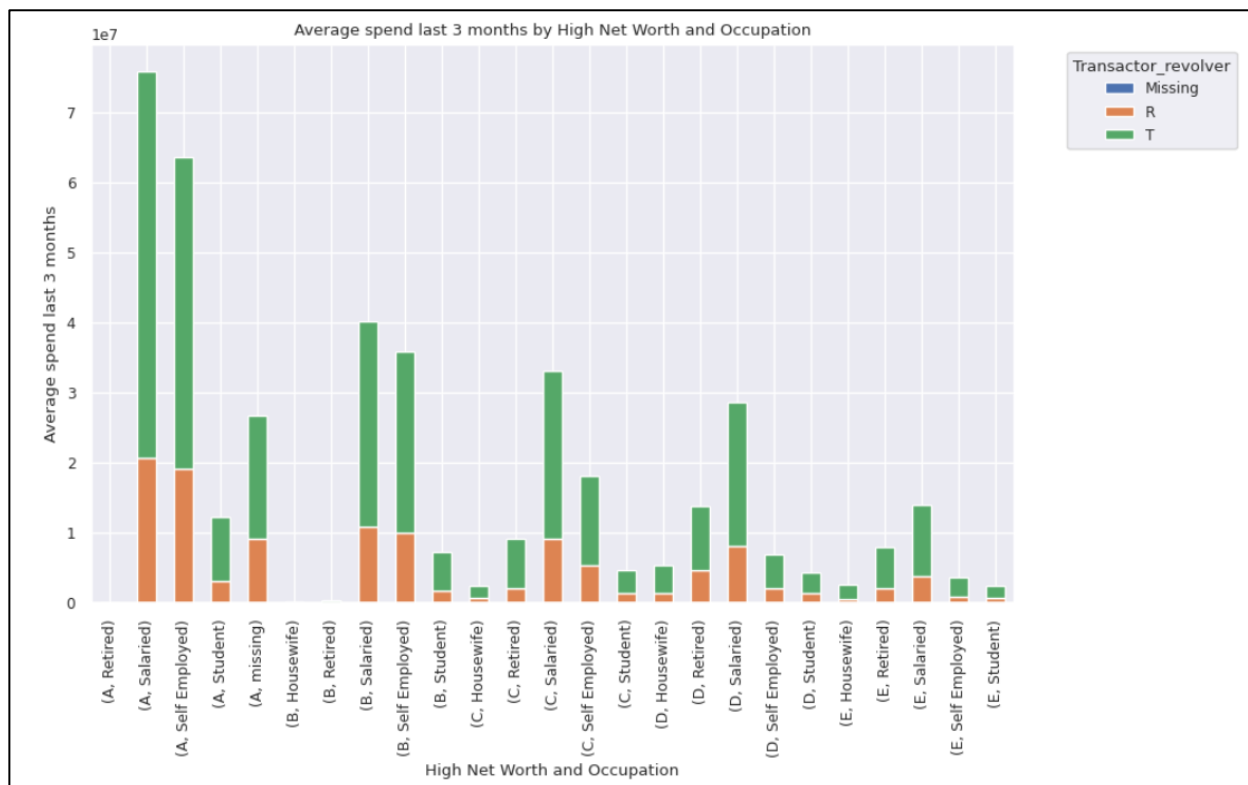


Figure 22

Insights

- ▶ 15.3% customers are revolvers
 - ▶ A(Salaried and Self Employed) customer has good average spend in last 3 months and also has good percentage of revolvers in them.
- Since revolvers are the customers who carry forward their bill to next month and hence are those segments through which bank can earn better interest so will help in revenue generation hence finding which all segments have more revolvers can also be a good segmenting. So, 'Transactor_revolver' is a important variable which can help in better revenue generation.

Order of average spend last3 months network and occupation wise?

Variables avg_spends_l3m, Occupation_at_source and high_network are three very important variables as we can see from above graph that its giving a clear segmentation of customers on the basis of the revenue they generate?

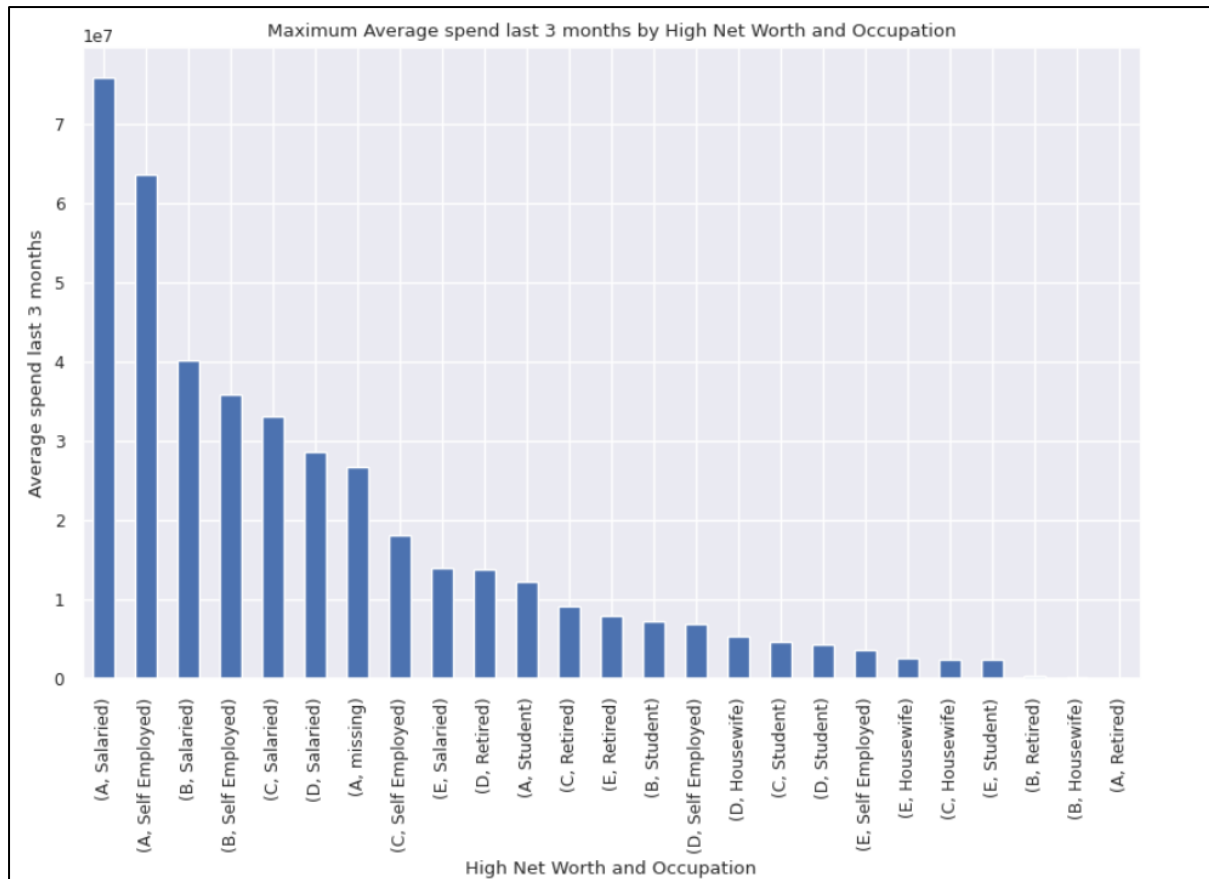


Figure 23

Insights

- Major Revenue generator segments of customers
A(salaried) > A(SE) > B(Sal) > B(SE) > C(Sal) > D(Sal)
- low spending segments of customers
D(SE) > E(SE)
- Segments where cards must not be issued as these customers are not spending at all.
E(Housewife) C(Housewife) E(student) and B(Retired Housewife) A(Retired)

We, must give more cc_limits to the Major revenue generator basket customers
And reduce the cc_limit for low spending segment customers.



Compare the cc_limit provided to customers and what is their average spend last3months card wise?

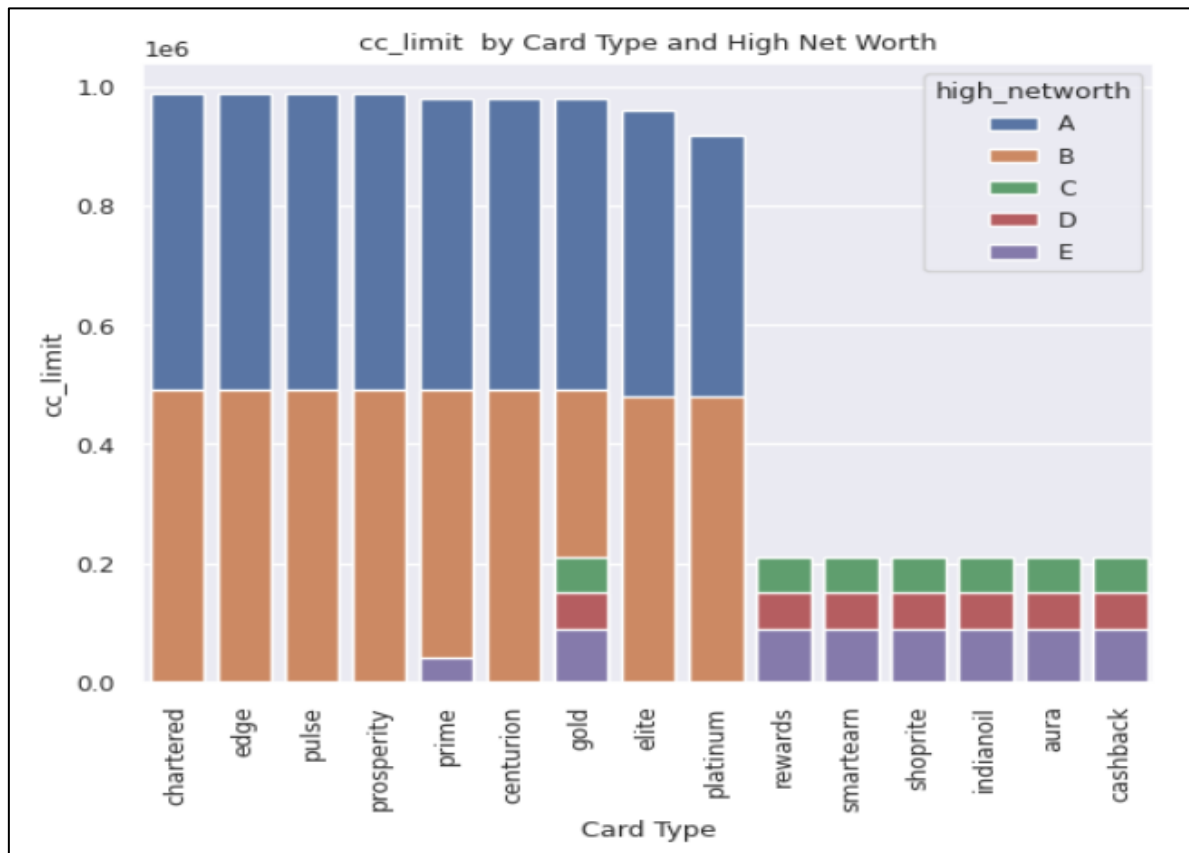


Figure 24

Insights 🧐

So , from above graph we can see that high_networth customers i.e. A and B are majorlgiven cards with high credit limits , when compared with C D and E but we have seen before than in C (Salaried)are also good in annual income , So, these customers must be checked and given better limits in future as we can see that many of them are revolvers.



Summary and Recommendations

Top 5 Important features which are really important for this dataset are:

1. **annual_Income_at_source**: This feature is likely to have a high correlation on the profit of the bank(average spend by customers), as it determines the ability of the customer to pay back and the amount of credit limit that can be offered to them. By offering more attractive credit products to customers with higher annual incomes, the bank can increase its revenue from interest and fees, while also reducing the attrition rate.

2. **cc_limit**: This feature is directly controllable by the bank and can be optimized to increase the revenue from interest and fees. By offering appropriate credit limits based on the creditworthiness and spending behaviour of the customer, the bank can reduce the attrition rate caused by customers seeking higher credit limits elsewhere. Here the average spend of C (salaried) are more and also they have stable income also , so why not increase a limit for these customers than before.

3. **Transactor_Revolver**: This feature indicates whether the customer tends to pay off their credit card balance in full every month or carries a balance and pays interest. By targeting customers who carry a balance and pay interest, the bank can increase its revenue from interest, while also reducing the attrition rate caused by customers seeking better credit terms elsewhere. We see that 15.3% of customers are revolvers out of which maximum penetration is in segment A(SAL SE) > B(SAL SE) and C(SAL SE) , so concentrate on charging interest from them but in a way that better credit terms are maintained with them.

4. **engagement_products**: This feature indicates the number of engagement products the customer has with the bank. By offering more engagement products, such as savings accounts, insurance products, or loyalty programs, the bank can increase customer loyalty and reduce the attrition rate.

5. **cc_active90**: This feature indicates the number of days the customer has been active on their credit card account in the last 90 days. By targeting customers who are highly active, the bank can increase its revenue from fees and interest, while also reducing the attrition rate caused by customers who are not actively using their credit card