Data Analysis Portfolio

Prepared By:-

Shubhayan Sarkar

Table Of Contents

PROFESSIONAL BACKGROUND

I am an Associate Data Scientist at a fintech company. While pursuing my B.Sc. in Economics I got fascinated about data science and from there my journey began. I have also worked as intern in couple of companies. I have an eagerness to learn and have the ability to cope up with any environment.

DATA ANALYTICS PROJECT

Here I had to provide a scenario from my daily life where I can use analytics for more better outcome:

Daily journey PLAN : I have to go to the office.

PREPARE : I have to figure out by what time I should get ready so that I can have more time for going to office without any rush, and have more options for commute.

PROCESS : I should check which mode of transport will cost less but also take to my office comfortably and within time also.

ANALYZE : If I am able to reach bus stop by 8:30 am, there will be a AC bus which will take me directly to my office comfortably otherwise I can take the bus which will cost less but will arrive there 20 minutes later.

SHARE : I should consult with my managers if reaching there after 20 minutes would not be a problem for them.

ACT : After getting the feedback from them I should act accordingly

INSTAGRAM USER ANALYTICS

Project Description:

As we have a vast data of Instagram, we are providing some insights to our internal team using those data so that the developers could know our users a little better and could give them better experience while using Instagram.

Approach:

We have 7 tables in total named as comments, follows, likes, photo_tags, photos, tags and users. We have some business queries from our stakeholders so we have tried to provide the answers through those tables so that they can take more wise decisions to make users more interactive.

Tech-Stack Used:

We have used MySQL Workbench here.

Insights:

A1. Our oldest members,

username	created_at
Darby_Herzog	2016-05-06 00:14:21
Emilio_Bernier52	2016-05-06 13:04:30
Elenor88	2016-05-08 01:30:41
Nicole71	2016-05-09 17:30:22
Jordyn.Jacobson2	2016-05-14 07:56:26

A2. Inactive users,

username
Aniya_Hackett
Kasandra_Homenick
Jaclyn81
Rocio33
Maxwell.Halvorson
Tierra.Trantow
Pearl7
Ollie_Ledner37
Mckenna17
David.Osinski47
Morgan.Kassulke
Linnea59
Duane60
Julien_Schmidt
Mike.Auer39
Franco_Keebler64
Nia_Haag
Hulda.Macejkovic
Leslie67
Janelle.Nikolaus81
Darby_Herzog
Esther.Zulauf61
Bartholome.Bernhard
Jessyca_West
Esmeralda.Mraz57
Bethany20

A3. User with most liked photo,

username	photo_id
Zack_Kemmer93	145

A4. Top 5 commonly used hashtags,

tag_name	frequency
smile	59
beach	42
party	39
fun	38
concert	24

A5. Day name most users have registered on,

dayname(created_at)	frequency
Thursday	16
Sunday	16
Friday	15

B1. Average post by a user,

avg_post
3

B2. Bot accounts,

username
Aniya_Hackett
Jaclyn81
Rocio33
Maxwell.Halvorson
Ollie_Ledner37
Mckenna17
Duane60
Julien_Schmidt
Mike.Auer39
Nia_Haag
Leslie67
Janelle.Nikolaus81
Bethany20

Result: as we have found some insights from the data, we can make sure that,

1. **the oldest users, most liked picture and the user should get featured by us for more interactions,**
2. **we can recommend the top common hashtags to the other users for put them in their posts,**
3. **we can show our advertisements frequently on those days when most of our users have registered so there might be a high probability of getting new users,**
4. **and by those we might be able to increase the average post per user,**
5. **we should remove the bot accounts as they are carrying other's identity so it can lead the real person in danger.**

OPERATION AND METRIC ANALYTICS

Project Description:

As we have a vast data from Microsoft, we are providing some insights to our internal team using those data so that we could forecast our growth and the developers could know our users a little better and could give them better experience.

Approach:

We have 1 table for case study 1 and 3 tables for case study 2. We have some business queries from our stakeholders so we have tried to provide the answers through those tables so that they can take more wise decisions regarding future.

Tech-Stack Used:

We have used MySQL Workbench here.

Insights:

Case study 1:

1. Jobs reviewed per hour,

ds	job_count	hours
2020-11-30	2	0.0111
2020-11-29	1	0.0056
2020-11-28	2	0.0092
2020-11-27	1	0.0289
2020-11-26	1	0.0156
2020-11-25	1	0.0125

2. Throughput of 7 days,

weekly_rolling_avg
0.03

Throughput daily,

dates	daily_rolling_avg
2020-11-25	0.02
2020-11-26	0.02
2020-11-27	0.01
2020-11-28	0.06
2020-11-29	0.05
2020-11-30	0.05

As we have only 7 days data, we are preferring daily throughput so that we can have more details. Otherwise, we may take 7 days throughput.

3. Language share,

language	share
English	12.5000
Arabic	12.5000
Persian	37.5000
Hindi	12.5000
French	12.5000
Italian	12.5000

4. Duplicate rows,

   Result:

ds	job_id	actor_id	event	language	time_spent	org

We don’t have any duplicate rows with same values

Case study 2:

1. Weekly user engagement,

weekno	usercount
17	663
18	1068
19	1113
20	1154
21	1121
22	1186
23	1232
24	1275
25	1264
26	1302
27	1372
28	1365
29	1376
30	1467
31	1299
32	1225
33	1225
34	1204
35	104

2. User growth,

monthno	growthrate
1	NULL
2	-3.79
3	11.68
4	18.56
5	9.48
6	9.37
7	17.96
8	5.15
9	-75.50
10	18.18
11	2.31
12	21.80

3. Weekly retention,

weeknumbers	week0	week1	week2	week3	week4	week5	week6	week7	week8	week9	week10	week11	week12	week14	week15	week16	week17	week18
17	740	472	324	251	205	187	167	146	145	145	136	131	132	143	91	82	77	5
18	788	362	261	203	168	147	144	127	113	122	106	118	127	110	85	67	4	0
19	601	284	173	153	114	95	91	81	95	82	68	65	63	42	49	2	0	0
20	555	223	165	121	91	72	63	67	63	65	67	41	40	33	0	0	0	0
21	495	187	131	91	74	63	75	72	58	48	45	39	35	28	0	0	0	0
22	521	224	150	107	87	73	63	60	55	48	41	39	31	1	0	0	0	0
23	542	219	138	101	90	79	69	61	54	47	35	30	0	0	0	0	0	0
24	535	205	143	102	81	63	65	61	38	39	29	0	0	0	0	0	0	0
25	500	218	139	101	75	63	50	46	38	35	2	0	0	0	0	0	0	0
26	495	181	114	83	73	55	47	43	29	0	0	0	0	0	0	0	0	0
27	493	199	121	106	68	53	40	36	1	0	0	0	0	0	0	0	0	0
28	486	194	114	69	46	30	28	3	0	0	0	0	0	0	0	0	0	0
29	501	186	102	65	47	40	1	0	0	0	0	0	0	0	0	0	0	0
30	533	202	121	78	53	3	0	0	0	0	0	0	0	0	0	0	0	0
31	430	145	76	57	1	0	0	0	0	0	0	0	0	0	0	0	0	0
32	496	188	94	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	499	202	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	518	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

4. Weekly engagement per device,

weekno	device	device_count
17	acer aspire desktop	67
17	acer aspire notebook	206
17	amazon fire phone	83
17	asus chromebook	251
17	dell inspiron desktop	187
17	dell inspiron notebook	503
17	hp pavilion desktop	132
17	htc one	190
17	ipad air	330
17	ipad mini	205
17	iphone 4s	217
17	iphone 5	706
17	iphone 5s	473
17	kindle fire	57
17	lenovo thinkpad	793
17	mac mini	59
17	macbook air	490
17	macbook pro	1516
17	nexus 10	145
17	nexus 5	382
17	nexus 7	177
17	nokia lumia 635	128
17	samsumg galaxy tablet	70
17	samsung galaxy note	116
17	samsung galaxy s4	449
17	windows surface	87
18	acer aspire desktop	295
18	acer aspire notebook	363
18	amazon fire phone	177
18	asus chromebook	523
18	dell inspiron desktop	683
18	dell inspiron notebook	953
18	hp pavilion desktop	373
18	htc one	174
18	ipad air	520
18	ipad mini	309
18	iphone 4s	448
18	iphone 5	1328
18	iphone 5s	778
18	kindle fire	265
18	lenovo thinkpad	1732
18	mac mini	159
18	macbook air	1604
18	macbook pro	3301
18	nexus 10	370
18	nexus 5	938
18	nexus 7	252
18	nokia lumia 635	341
18	samsumg galaxy tablet	79
18	samsung galaxy note	139
18	samsung galaxy s4	1130
18	windows surface	107

[These are the first few results]

5. Email engagement,

weekno	email_open_rate	email_clickthrough_rate	weekly_digest_rate	reengagement_email_rate
17	21.28	11.39	62.32	5.01
18	22.24	10.49	63.45	3.83
19	22.67	11.13	62.16	4.04
20	22.64	11.43	61.62	4.31
21	22.82	9.97	63.52	3.69
22	21.56	10.66	63.59	4.19
23	22.34	11.18	62.39	4.09
24	22.92	10.99	61.61	4.48
25	21.79	10.54	63.77	3.90
26	22.22	10.61	62.99	4.18
27	22.49	11.37	62.24	3.90
28	22.48	10.77	62.92	3.83
29	21.71	10.51	63.98	3.79
30	23.24	10.59	62.29	3.88
31	23.25	7.66	65.27	3.82
32	22.85	7.14	66.59	3.42
33	23.10	7.91	64.73	4.26
34	23.91	7.67	64.33	4.08
35	32.28	29.92	0.00	37.80

Result: through this project I have learned many sql concepts like window function, lag lead function, sub query, and some business concepts like growth, engagement, throughput etc. which will definitely help me in future.

HIRING PROCESS ANALYTICS

Project Description:

As we have a vast data from Google, we are providing some insights to our internal team using those data so that we could know candidates a little better.

Approach:

We have some business queries from our stakeholders so we have tried to provide the answers using ms excel so that they can take more wise decisions regarding future.

Tech-Stack Used:

We have used MS Excel here.

Insights:

1. Males and females hired:

2563	Male Hired
1856	Female Hired

2. Average salary:

Average_Salary
49983

3. Class intervals:

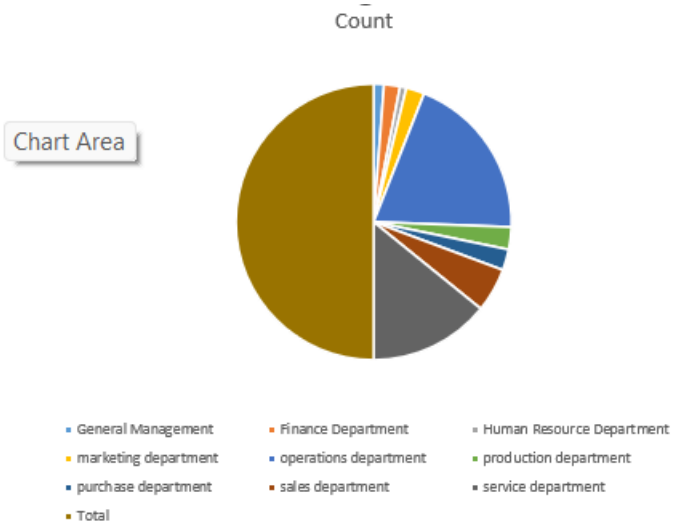
:

Salary_Range	Count
100-10000	678
10001-20000	732
20001-30000	711
30001-40000	710
40001-50000	781
50001-60000	750
60001-70000	698
70001-80000	734
80001-90000	711
90001-100000	659
100001-150000	0
150000-250000	1
250000-400000	2

4. Proportion of department:

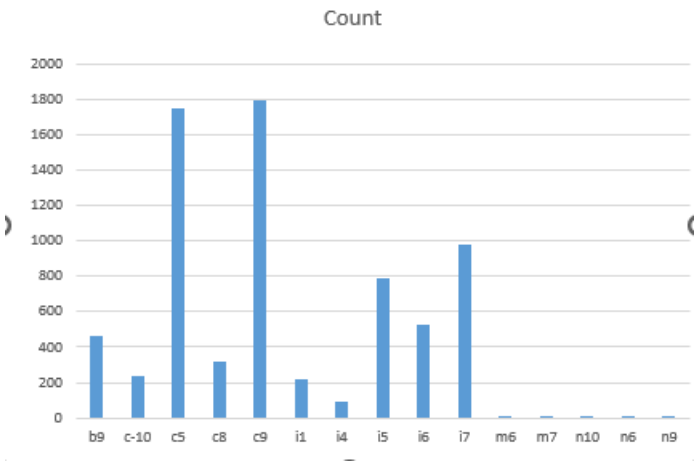
Result:

Deparment	Count	Fraction
General Management	113	2.4057909
Finance Department	176	3.7470726
Human Resource Departm	70	1.490313
marketing department	202	4.3006174
operations department	1843	39.237811
production department	246	5.2373856
purchase department	230	4.8967426
sales department	485	10.32574
service department	1332	28.358527
Total	4697	



5. Different post tiers:

Post Nam	Count
b9	463
c-10	232
c5	1747
c8	320
c9	1792
i1	222
i4	88
i5	787
i6	527
i7	982
m6	3
m7	1
n10	1
n6	1
n9	1



Result: through this project I have learned many MS Excel concepts which will definitely help me in future.

IMDB MOVIE ANALYSIS

Project Description:

As we have a vast data from IMDB, we are providing some insights to our internal team using those data so that we could know the market a little better.

Approach:

We have some business queries from our stakeholders so we have tried to provide the answers using ms excel so that they can take more wise decisions regarding future.

Tech-Stack Used:

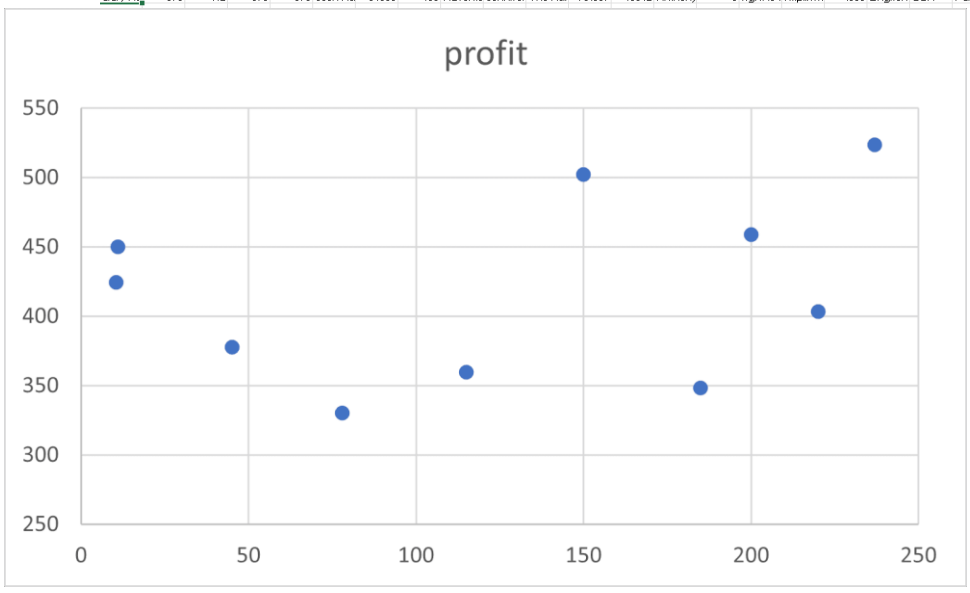
We have used MS Excel here.

Insights:

A. Data cleaning:
We had 2697 blank cells and 45 duplicate rows which we have removed. Besides that we have also removed some unwanted columns like color, duration, plot_keywords, cast_total_facebook_likes which would not be helpful for this analysis.

B. Highest profit:

director_	num_cri	duration	director_	actor_3_	actor_2_	actor_1_f	gross	genres	actor_1_f	movie_ti	num_vo	cast_tota	actor_3_f	facenum	plot_key	movie_ir	num_us	languag	country	content_	budget	title_yea	actor_2_	imdb_sc	aspect_r	movie_fa	profit
James C	723	178	0	855	Joel Dav	1000	760.5	Action|A	CCH Pot	AvatarA	886204	4834	Wes Stu	0	avatar|fi	http://ww	3054	English	USA	PG-13	237	2009	936	7.9	1.78	33000	523.5
Colin Tr	644	124	365	1000	Judy Gr	3000	652.2	Action|A	Bryce D	Jurassic	419214	9458	Omar Sy	0	dinosau	http://ww	1290	English	USA	PG-13	150	2015	2000	7	2	150000	502.2
James C	315	194	0	794	Kate Wi	29000	658.7	Drama|F	Leonard	TitanicA	793059	45223	Gloria S	0	artist|lov	http://ww	2528	English	USA	PG-13	200	1997	14000	7.7	2.35	26000	458.7
George I	282	125	0	504	Peter Cu	11000	460.9	Action|A	Harrison	Star Wa	911097	13485	Kenny E	1	death st	http://ww	1470	English	USA	PG	11	1977	1000	8.7	2.35	33000	449.9
Steven S	215	120	14000	548	Dee Wal	861	434.9	Family|S	Henry T	E.T. the	281842	2811	Peter Co	0	bicyclist	http://ww	515	English	USA	PG	10.5	1982	725	7.9	1.85	34000	424.4
Joss Wh	703	173	0	19000	Robert D	26000	623.3	Action|A	Chris He	The Ave	995415	87697	Scarlett	3	alien inv	http://ww	1722	English	USA	PG-13	220	2012	21000	8.1	1.85	123000	403.3
Roger A	186	73	28	847	Nathan L	2000	422.8	Adventu	Matthew	The Lior	644348	6458	Niketa C	0	king|pri	http://ww	656	English	USA	G	45	1994	886	8.5	1.66	17000	377.8
George I	320	136	0	1000	Liam Ne	20000	474.5	Action|A	Natalie F	Star Wa	534658	37723	Ian McD	1	alien|ch	http://ww	3597	English	USA	PG	115	1999	14000	6.5	2.35	13000	359.5
Christor	645	152	22000	11000	Heath Le	23000	533.3	Action|C	Christiar	The Dar	2E+06	57802	Morgan	0	based o	http://ww	4667	English	USA	PG-13	185	2008	13000	9	2.35	37000	348.3
Gary Ro	673	142	378	575	Josh Hu	34000	408	Adventu	Jennifer	The Hur	701607	49942	Anthony	0	fight to t	http://ww	1959	English	USA	PG-13	78	2012	14000	7.3	2.35	140000	330



C. Top250movies:

movie_title	imdb_sc	rank
The Shawshank RedemptionÂ	9.3	1
The GodfatherÂ	9.2	2
The Godfather: Part IIÂ	9	3
The Dark KnightÂ	9	4
Schindler's ListÂ	8.9	5
Pulp FictionÂ	8.9	6
The Good, the Bad and the UglyÂ	8.9	7
The Lord of the Rings: The Return of the KingÂ	8.9	8
Fight ClubÂ	8.8	9
InceptionÂ	8.8	10
Star Wars: Episode V - The Empire Strikes BackÂ	8.8	11
Forrest GumpÂ	8.8	12
The Lord of the Rings: The Fellowship of the RingÂ	8.8	13
Seven SamuraiÂ	8.7	14
The Lord of the Rings: The Two TowersÂ	8.7	15
City of GodÂ	8.7	16
The MatrixÂ	8.7	17
GoodfellasÂ	8.7	18
One Flew Over the Cuckoo's NestÂ	8.7	19
Star Wars: Episode IV - A New HopeÂ	8.7	20
The Silence of the LambsÂ	8.6	21
Se7enÂ	8.6	22
InterstellarÂ	8.6	23
Saving Private RyanÂ	8.6	24
The Usual SuspectsÂ	8.6	25
Spirited AwayÂ	8.6	26
American History XÂ	8.6	27
Modern TimesÂ	8.6	28

[These are the the top few values}

Non-English movies:

Top_Foreign_Lang_Film
The Good, the Bad and the UglyÂ
Seven SamuraiÂ
City of GodÂ
Spirited AwayÂ
The Lives of OthersÂ
Children of HeavenÂ
OldboyÂ
Princess MononokeÂ
A SeparationÂ
Das BootÂ
AmÃ©lieÂ
DownfallÂ
The HuntÂ
MetropolisÂ
Howl's Moving CastleÂ
IncendiesÂ
The Secret in Their EyesÂ
Pan's LabyrinthÂ
Elite SquadÂ
The CelebrationÂ
Amores PerrosÂ
AkiraÂ
Tae Guk Gi: The Brotherhood of WarÂ
The Sea InsideÂ
PersepolisÂ
A Fistful of DollarsÂ
Waltz with BashirÂ
My Name Is KhanÂ
Central StationÂ
The ChorusÂ
AmourÂ
Letters from Iwo JimaÂ
HeroÂ
4 Months, 3 Weeks and 2 DaysÂ

D. Top10directors:

top10director	imdb_score
Akira Kurosawa	8.7
Charles Chaplin	8.6
Tony Kaye	8.6
Alfred Hitchcock	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Ron Fricke	8.5
Sergio Leone	8.433333
Christopher Nola	8.425
Richard Marquan	8.4

E. Top genres:

genre_1	genre_2	gross
Drama	Romance	658.6723
Action	Adventure	594.2886
Action	Crime	533.3161
Family	Sci-Fi	434.9495
Adventure	Animation	422.7838
Adventure	Drama	407.9993

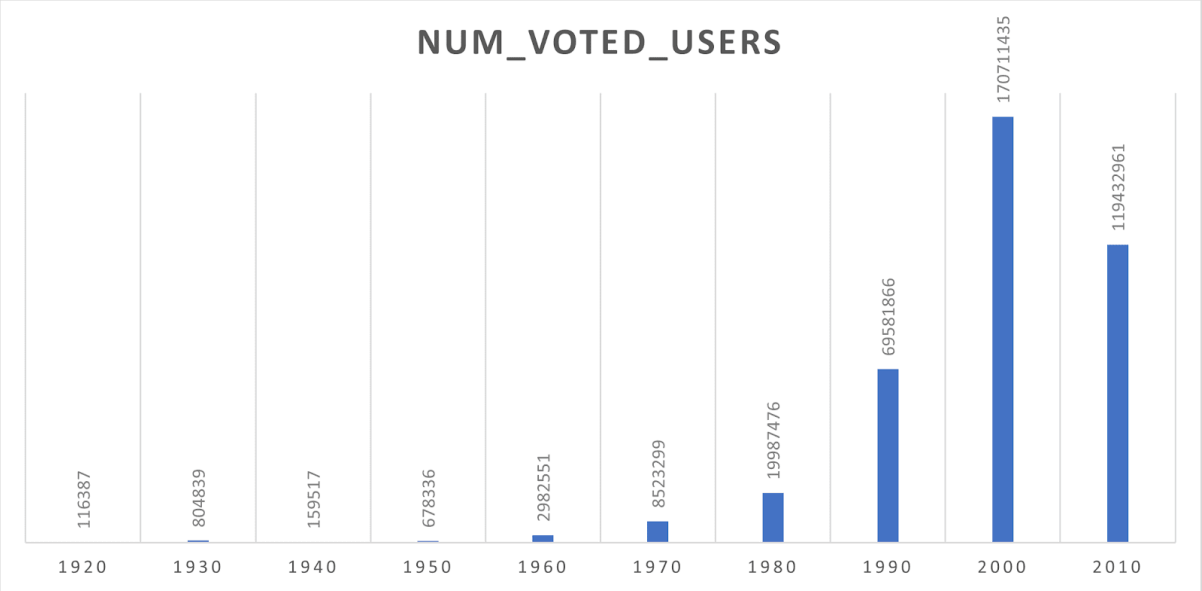
F. Critic and audience favourites:

actor_1_name	num_critic_for_reviews
Leonardo DiCapri	330.1905
Brad Pitt	245
Meryl Streep	181.4545

actor_1_name	num_user_for_reviews
Leonardo DiCaprio	1.20E+67
Brad Pitt	4.98E+49
Meryl Streep	5.74E+30

We can observe that for both Leonardo Dicaprio is the best actor

decades	num_voted_users
1920	116387
1930	804839
1940	159517
1950	678336
1960	2982551
1970	8523299
1980	19987476
1990	69581866
2000	1.71E+08
2010	1.19E+08



Result: through this project I have learned many MS Excel concepts which will definitely help me in future.

BANK LOAN CASE STUDY

Project Description:

As we have a vast data of customers, we are providing some insights to our internal team using those data so that we could classify the customers and take the decision wisely.

Approach:

We have current and previous application data. Based on the features we have tried to rectify the non-useful columns and also the null values and after some feature engineering we have provided some answers through those tables so that they can take more wise decisions regarding future.

Tech-Stack Used:

We have used jupyter notebook here.

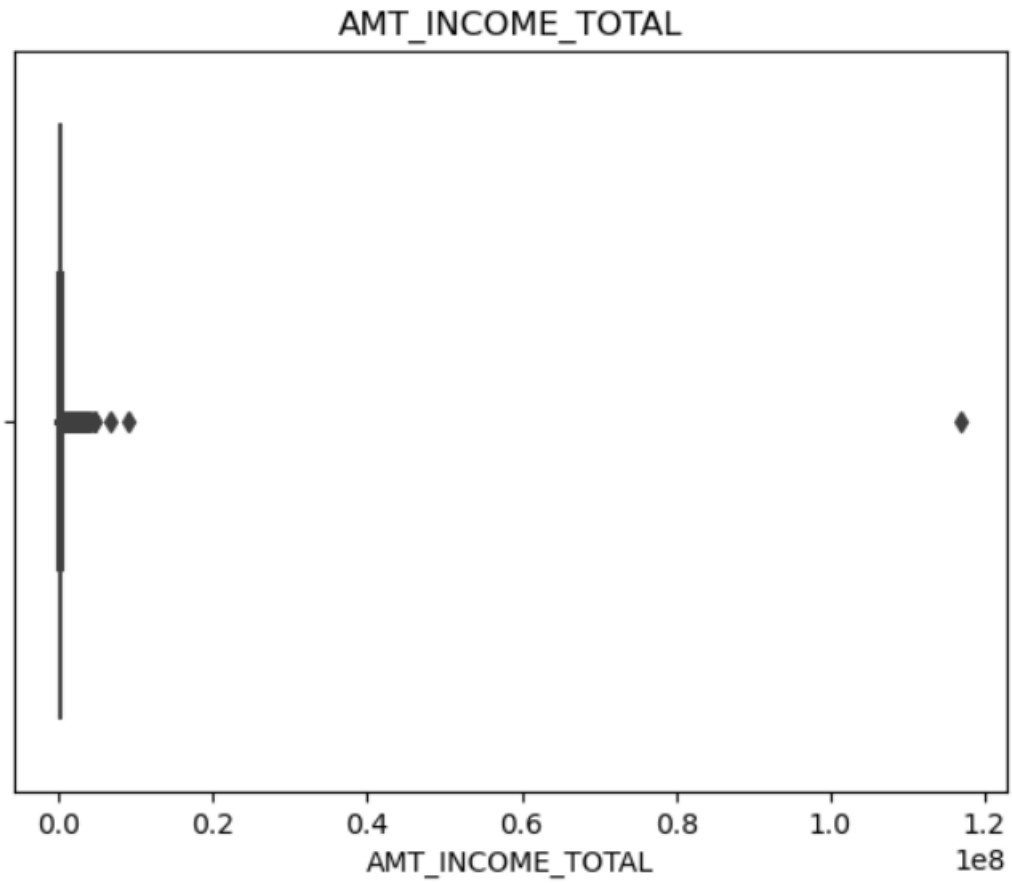
Insights:

1. Missing values:
   We have found that application table and previous table has 41,4 columns with more than 50% of missing values, so I have dropped those columns respectively.

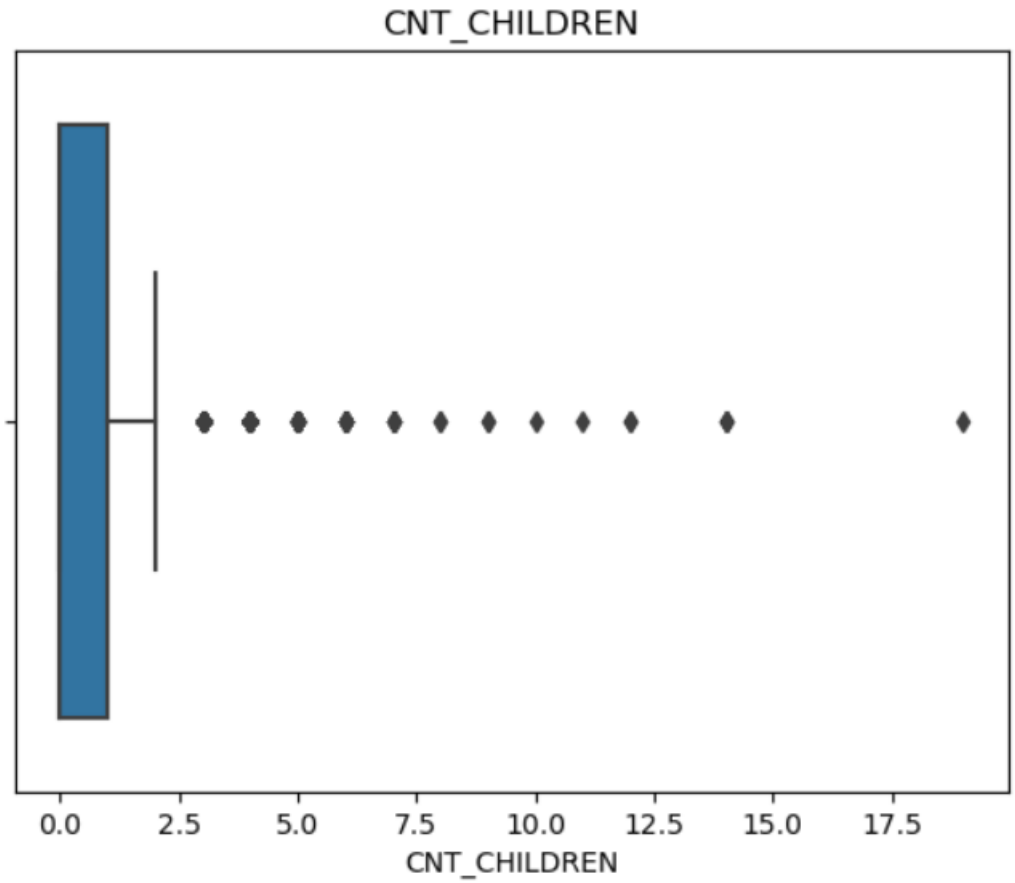
   For columns with less than 5% missing values I have dropped the rows respectively.

   For the rest of the columns, I have replaced the blanks with mean and mode for categorical columns.

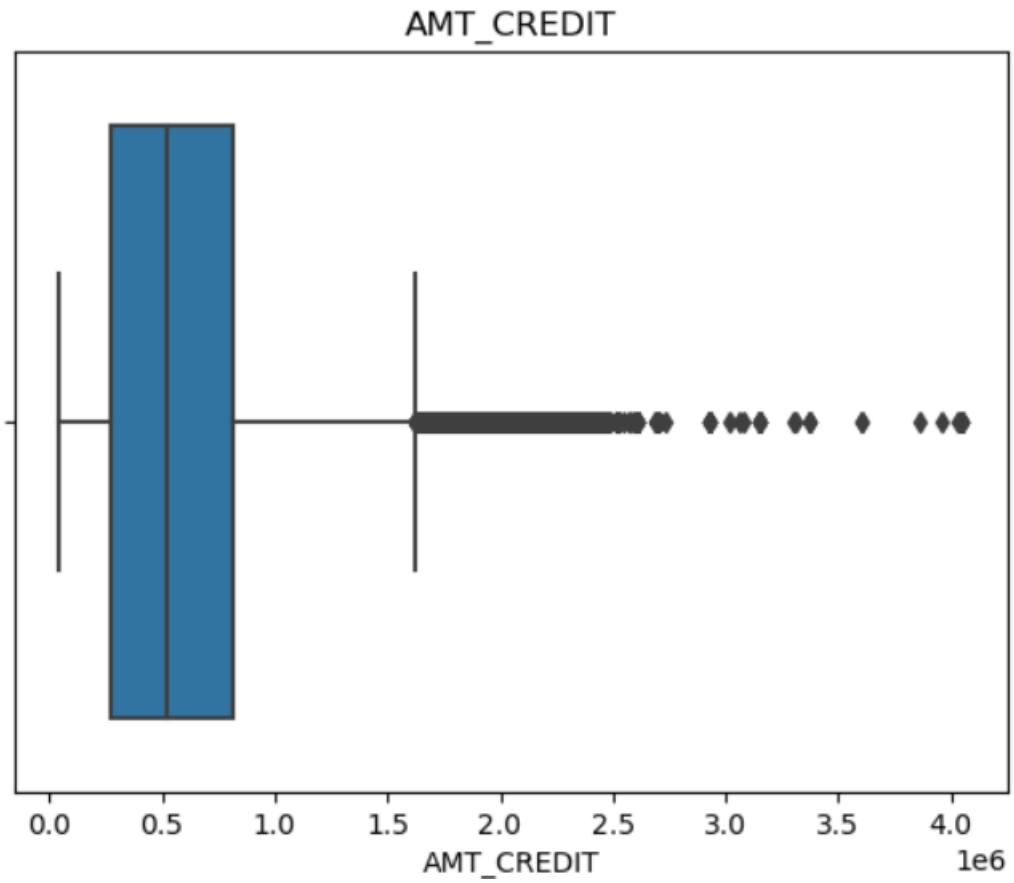
2. Outliers:
   I have found there are significant outliers in :
   1. 'AMT_INCOME_TOTAL'



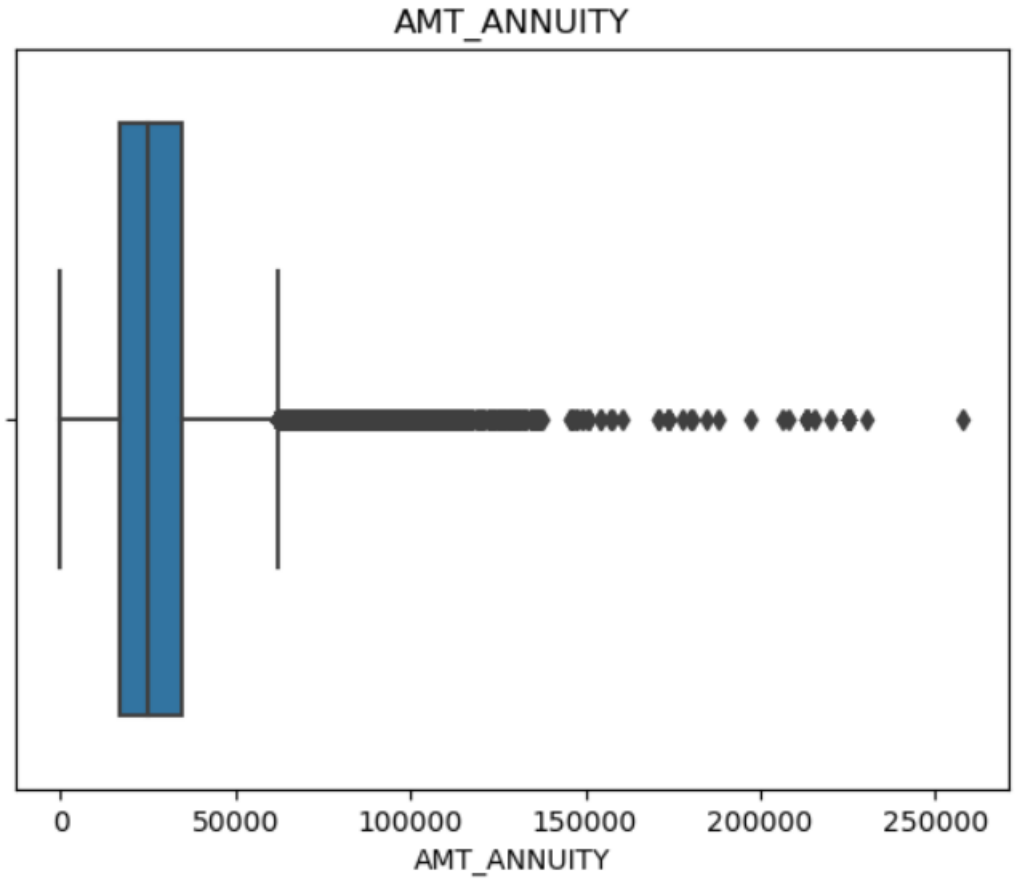
2. 'CNT_CHILDREN'



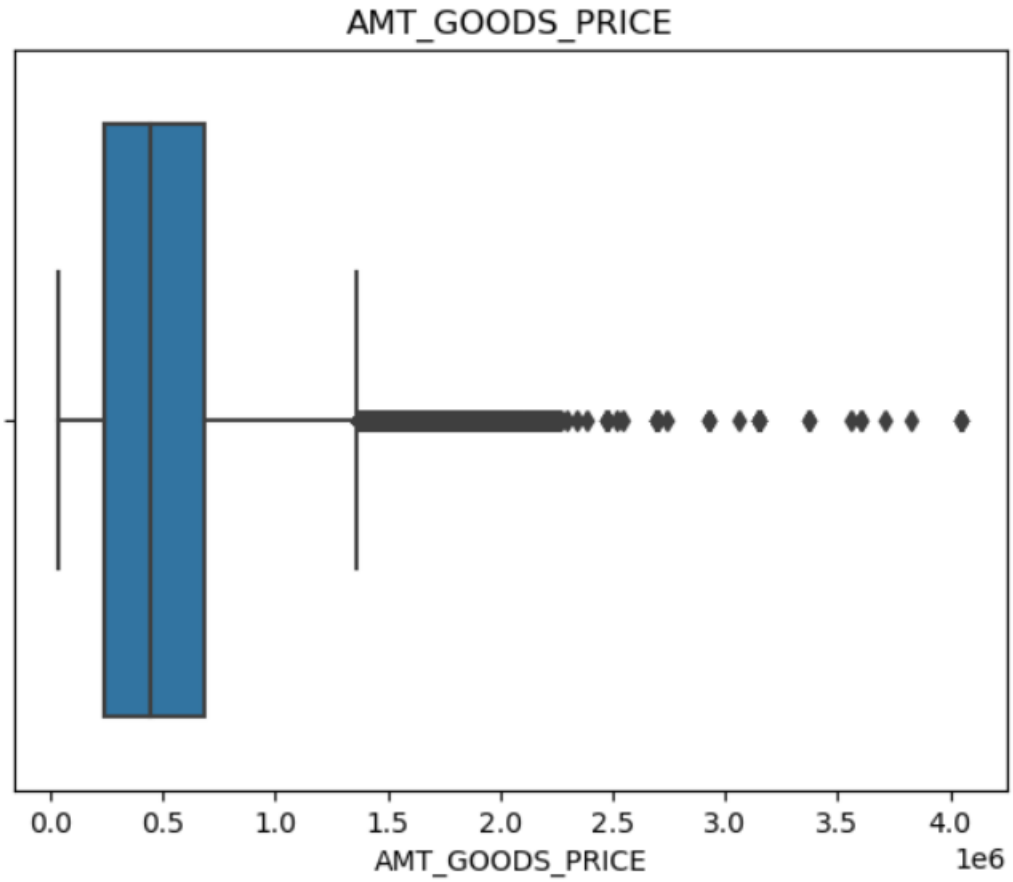
3. 'AMT_CREDIT'



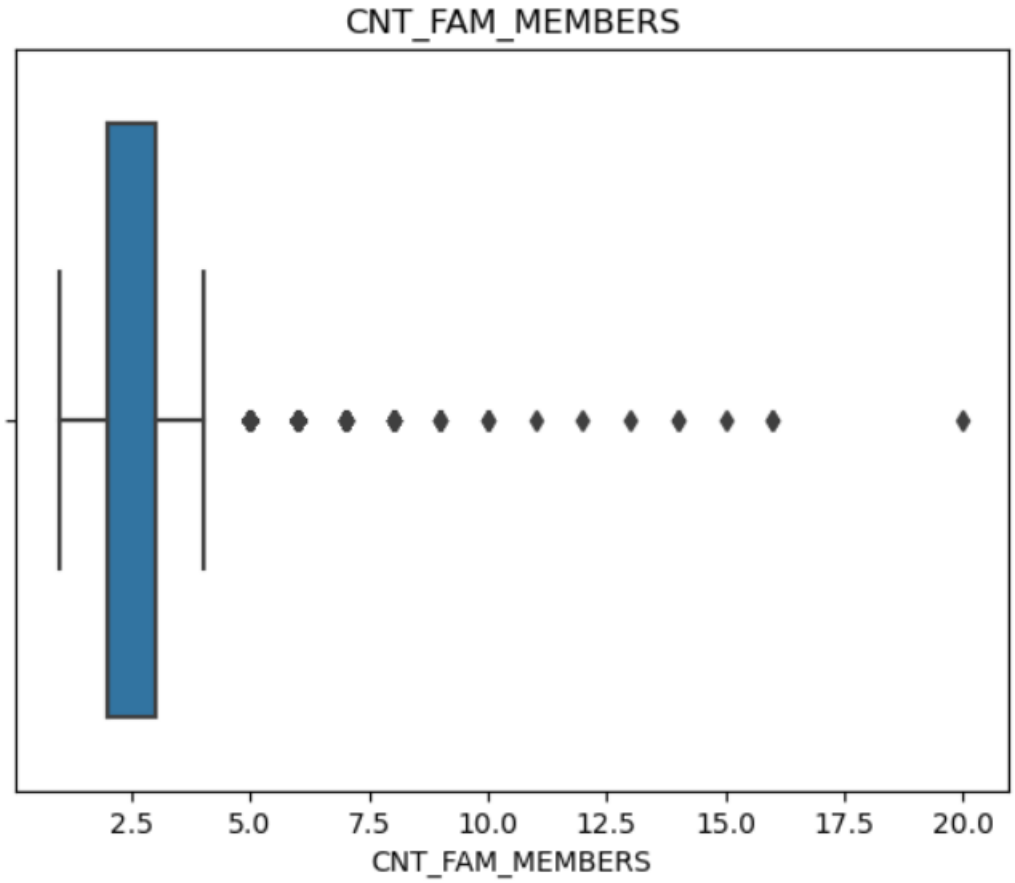
4. 'AMT_ANNUITY'



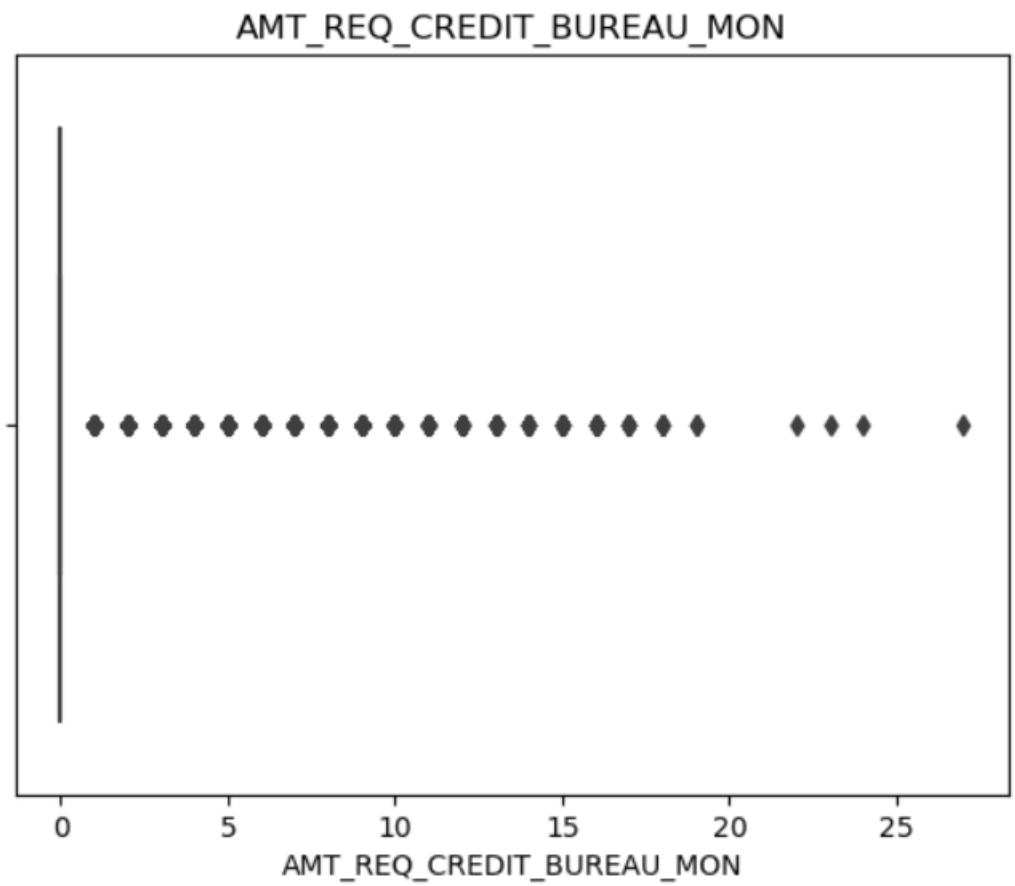
5. 'AMT_GOODS_PRICE'



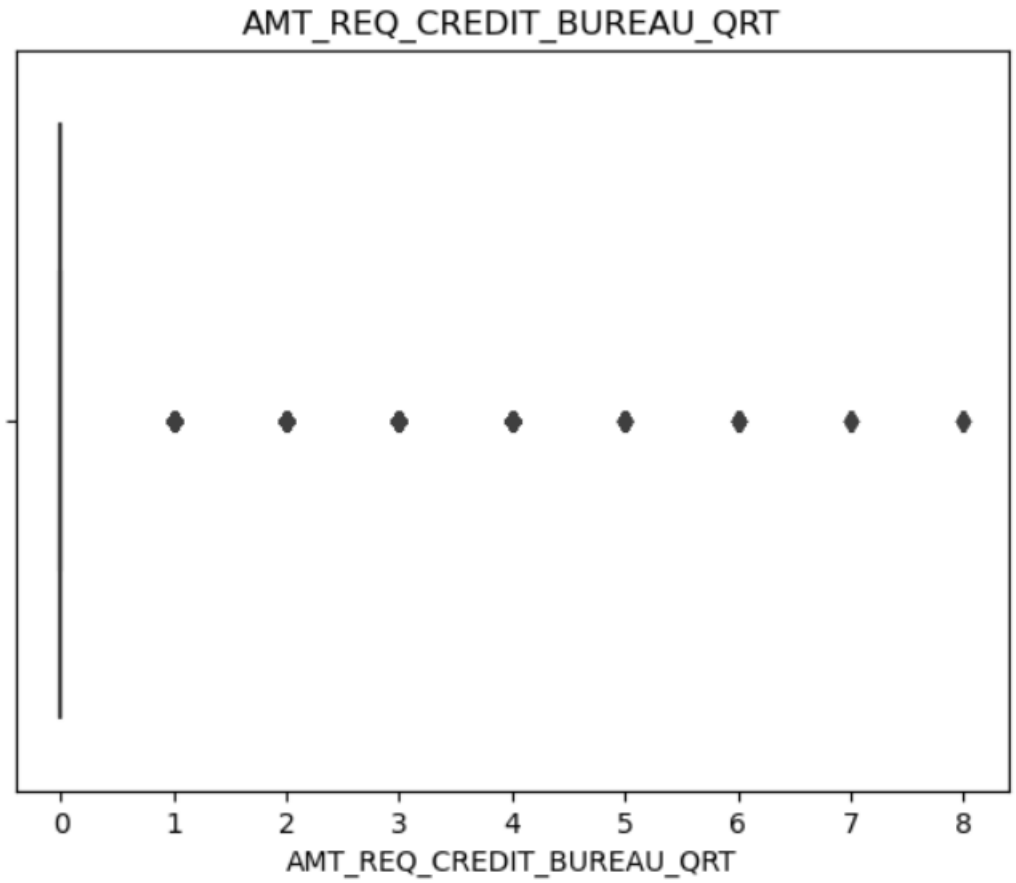
6. 'CNT_FAM_MEMBERS'



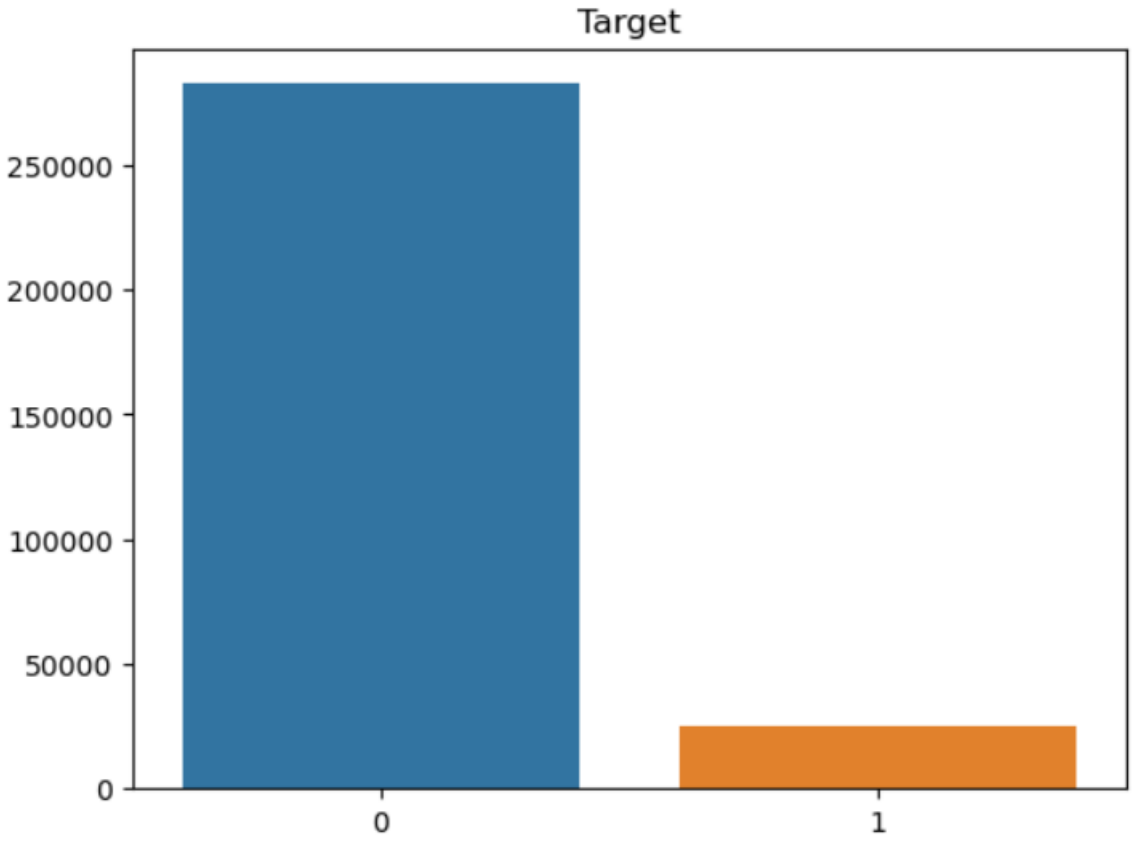
7. 'AMT_REQ_CREDIT_BUREAU_MON'


AMT_REQ_CREDIT_BUREAU_MON
0
5
10
15
20
25
AMT_REQ_CREDIT_BUREAU_MON

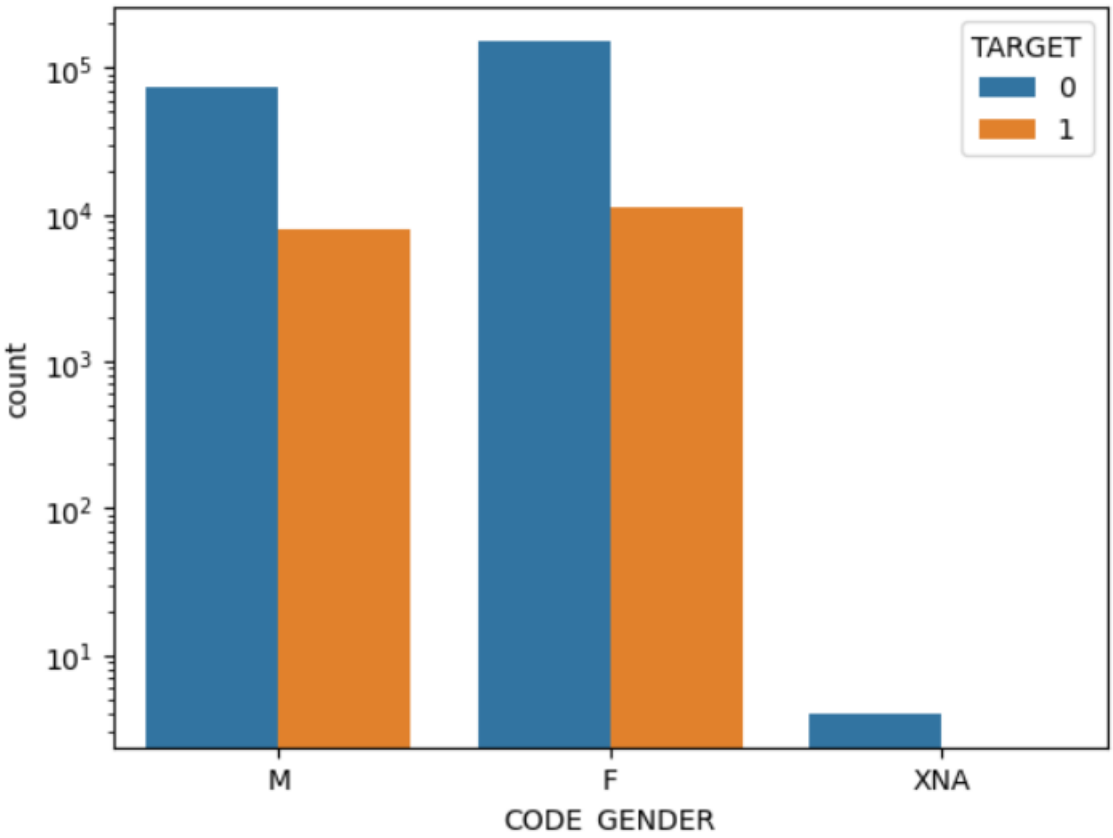
8. 'AMT_REQ_CREDIT_BUREAU_QRT'



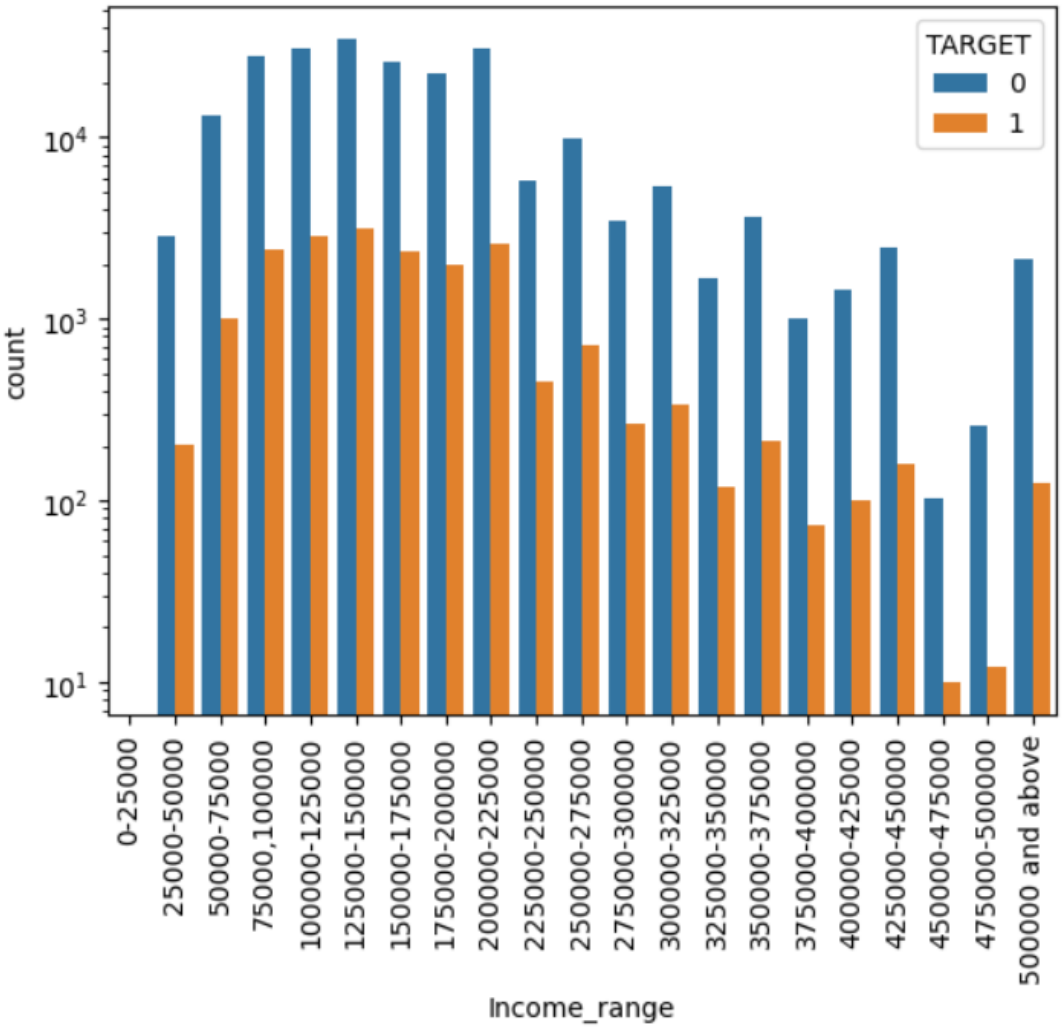
3. Data Imbalance:



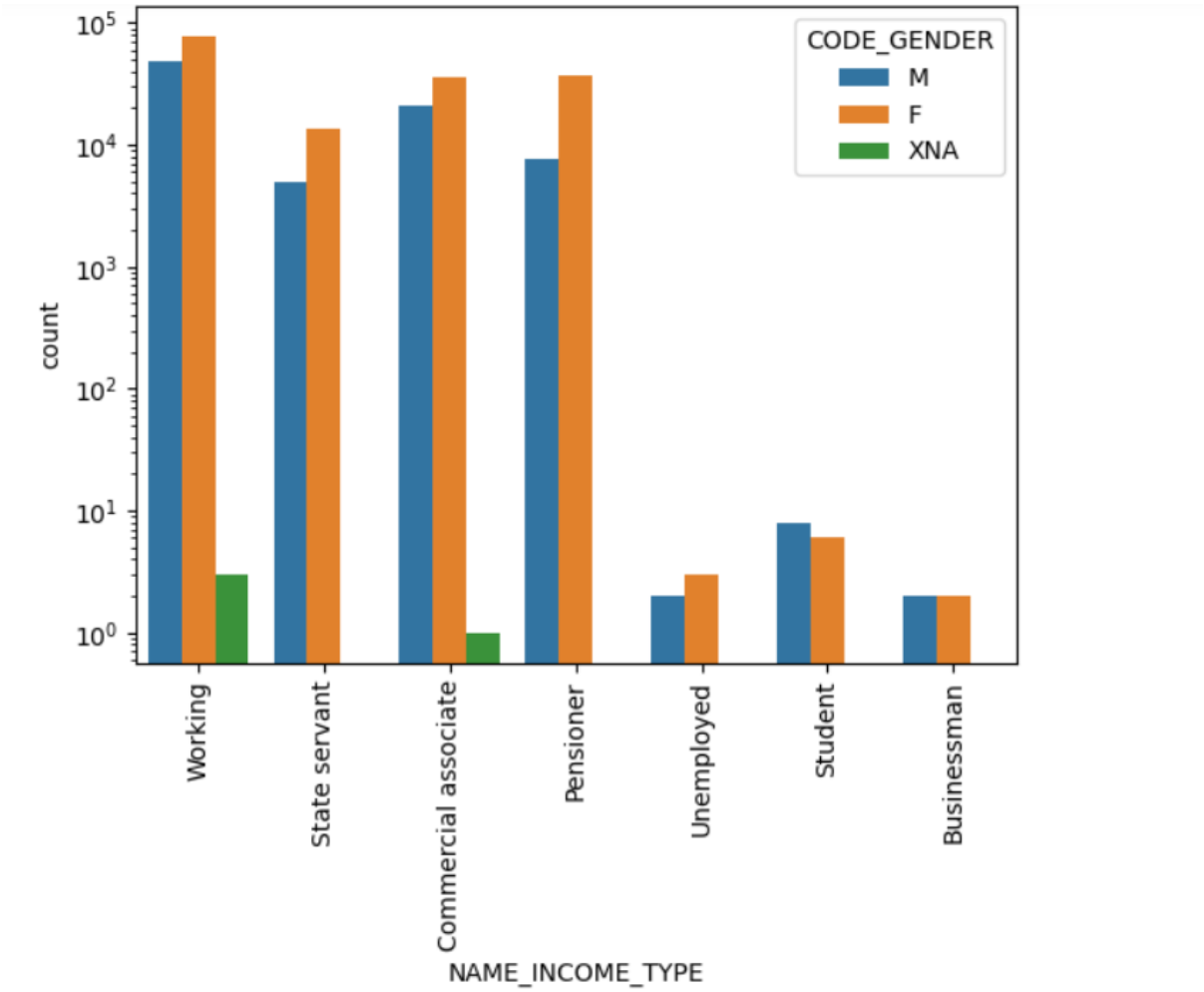
4. EDA:
   a. Univariate analysis:



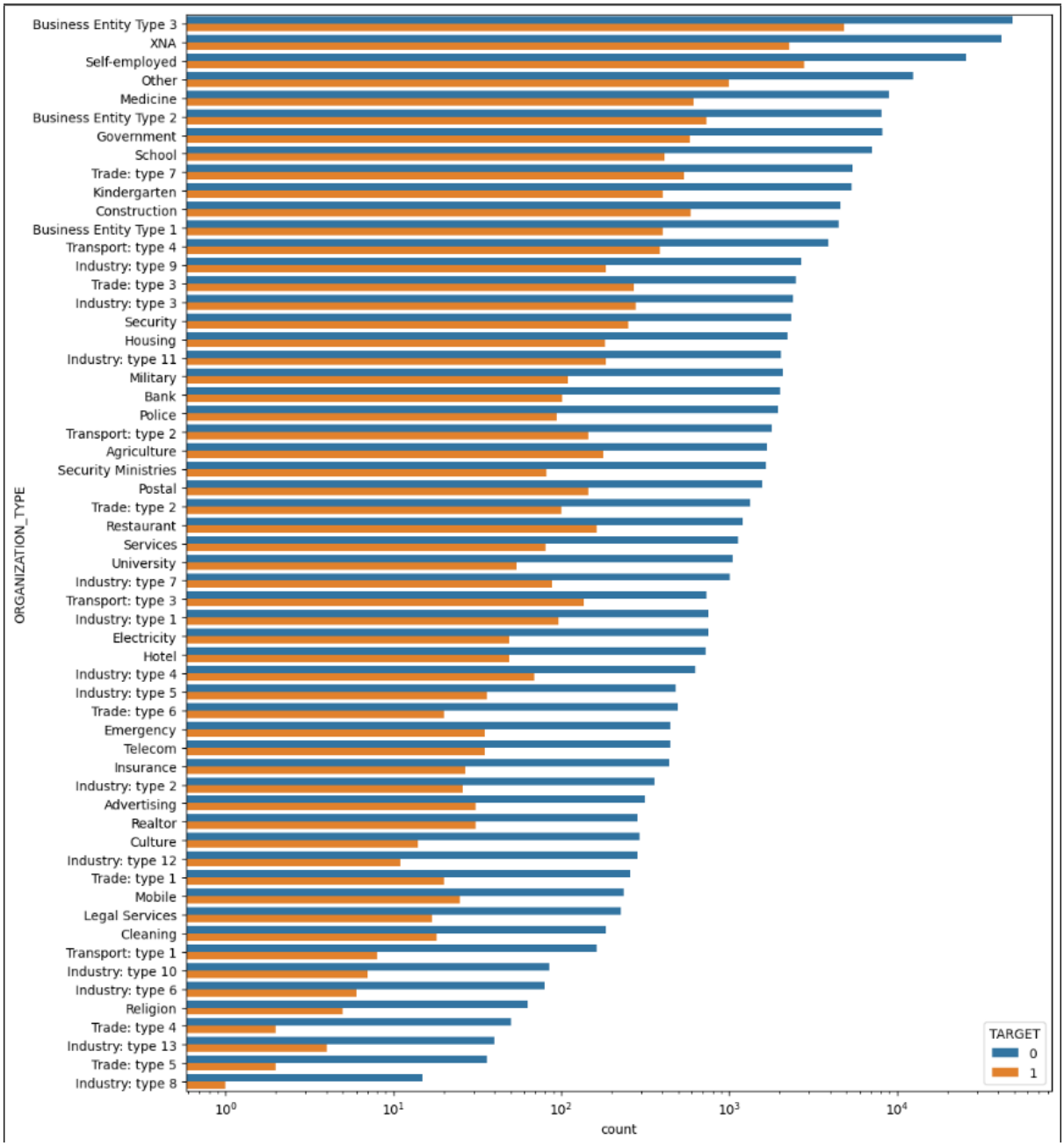
It can be seen that both number of males and females are almost same for having payment difficulties It can also be seen that number of females is more than males for having payment difficulties.



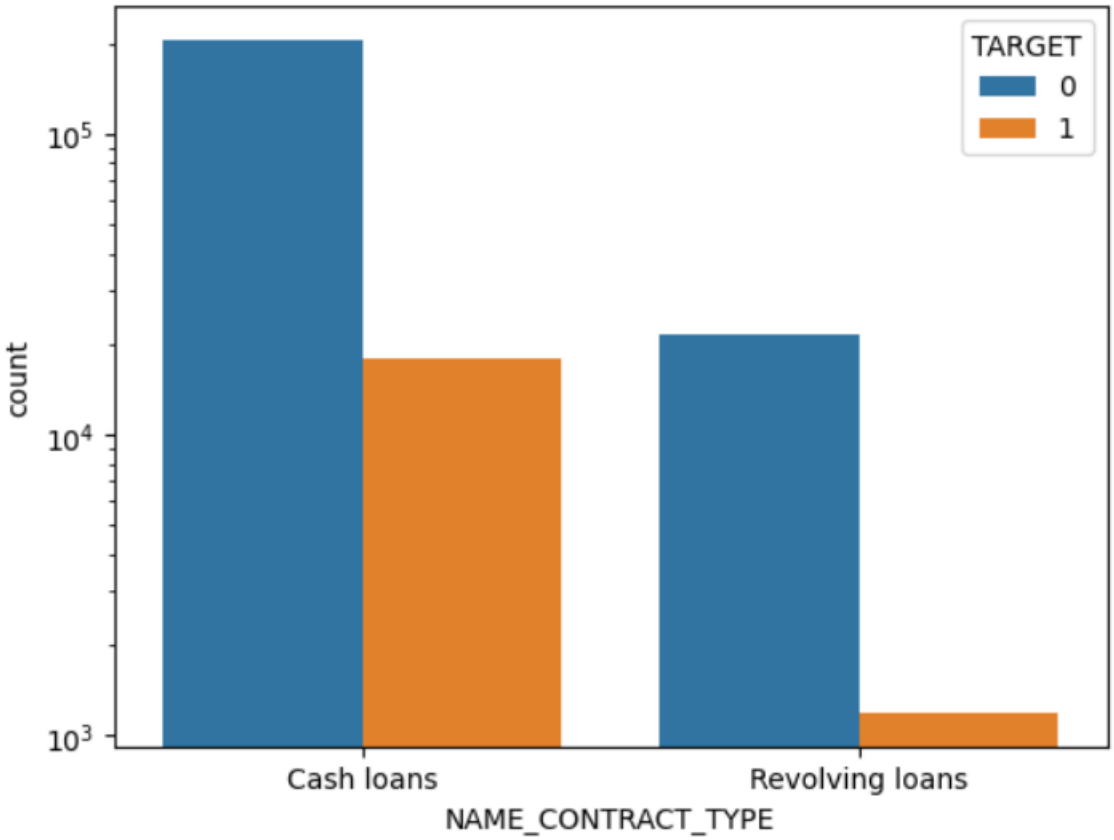
From the above plot we can infer that the maximum number of clients with payment difficulties lie in the income range 75k-2.25 lakhs and clients in the range of 4.5l - 4.75l are the without payment difficulties



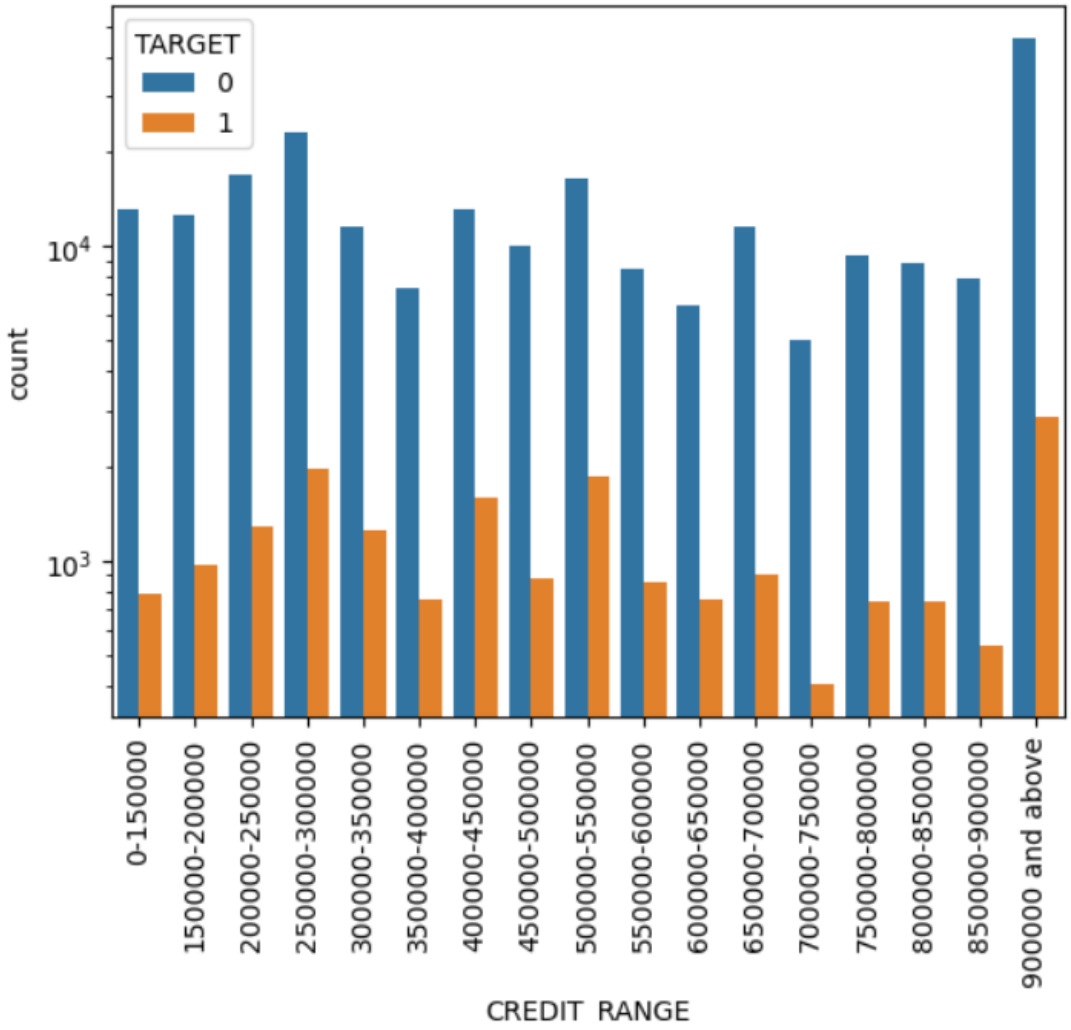
We can conclude that most clients who fall in working type income category are applying for loans more



It can be seen that Industry: type 8, organisation type have least count of payment difficulty clients Most clients with payment difficulties lie in organisation type named Business Entity Type 3

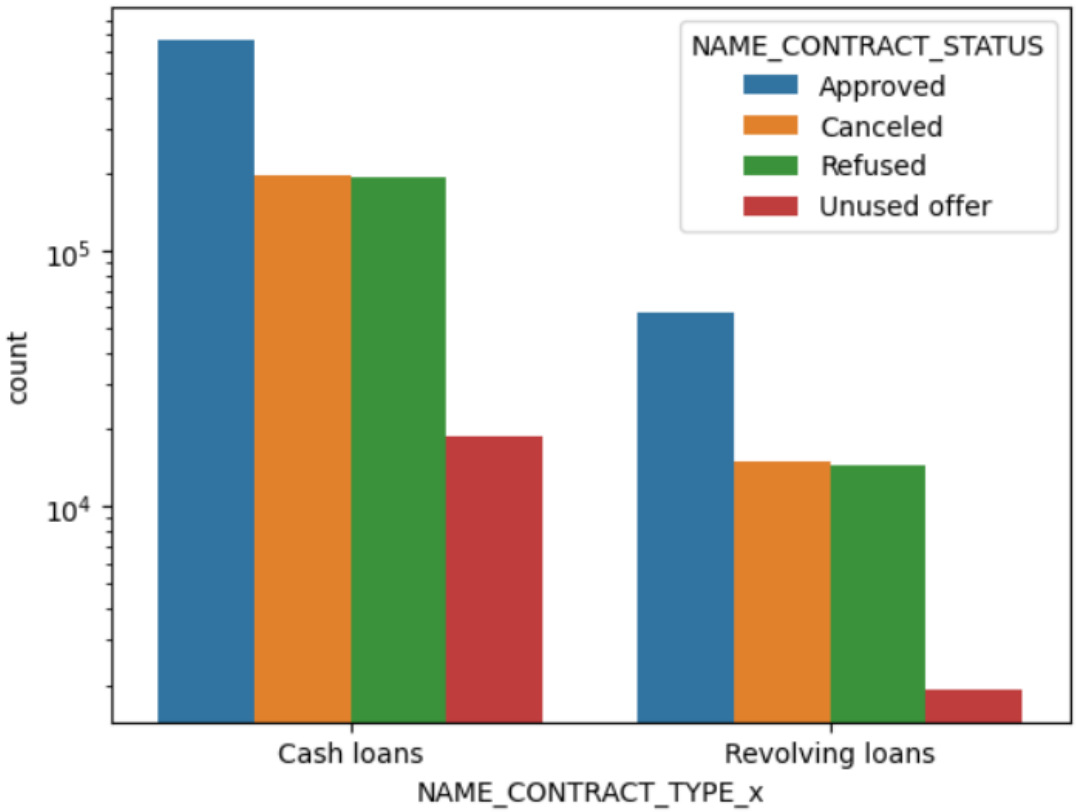


cash loans applicants have more payment difficulties

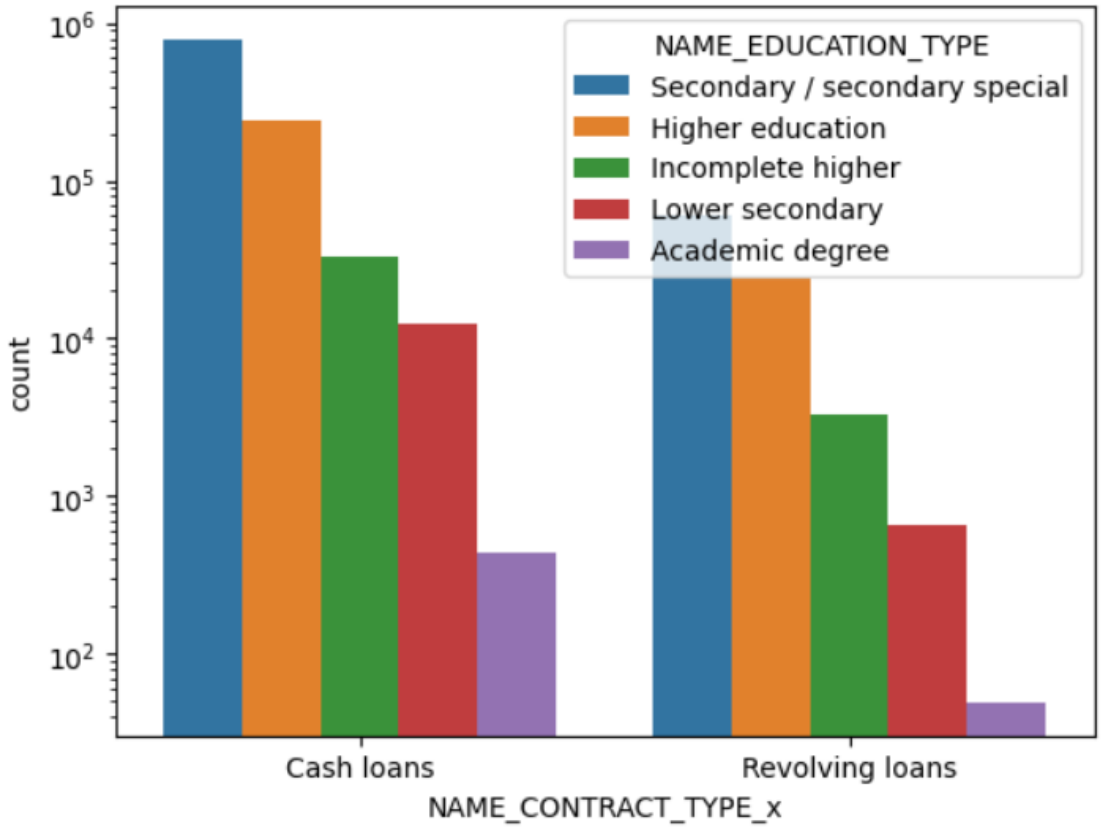


Clients with credit range lying in 900000 and above are the ones who deafualt the most and least number of clients lying in income range 7lac- 7.5 lac are capable of paying

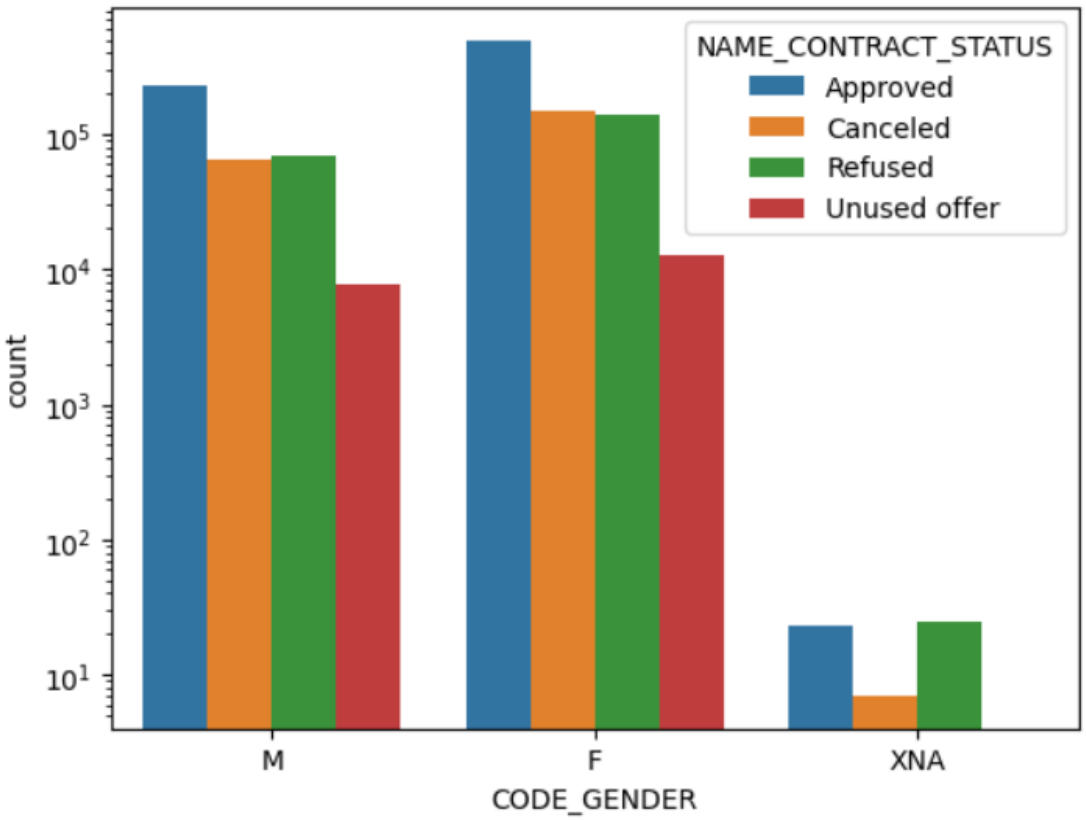
Prev.



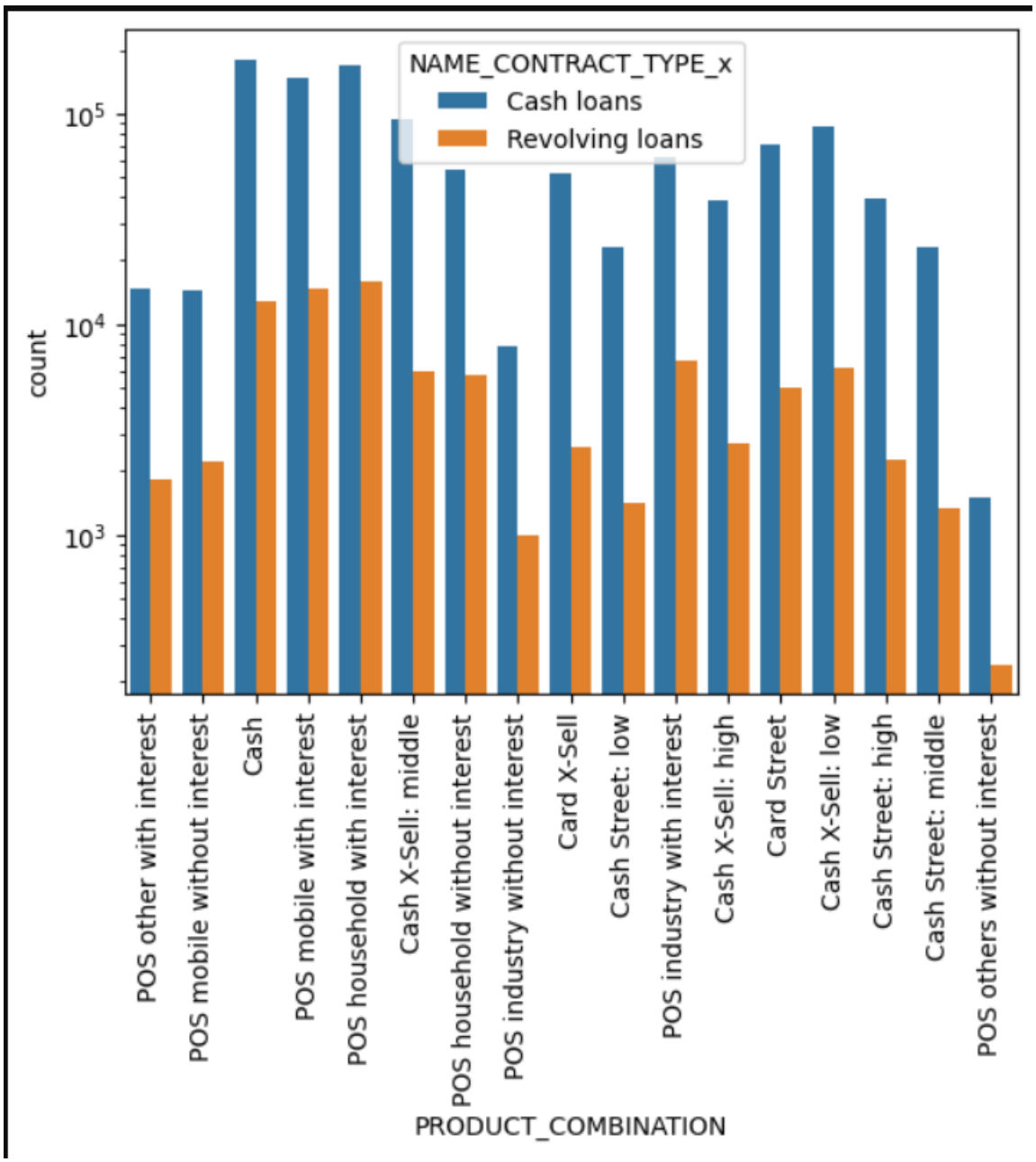
Contract type cash loans are maximum in number where all kinds of contract statuses are more than contract type revolving loans



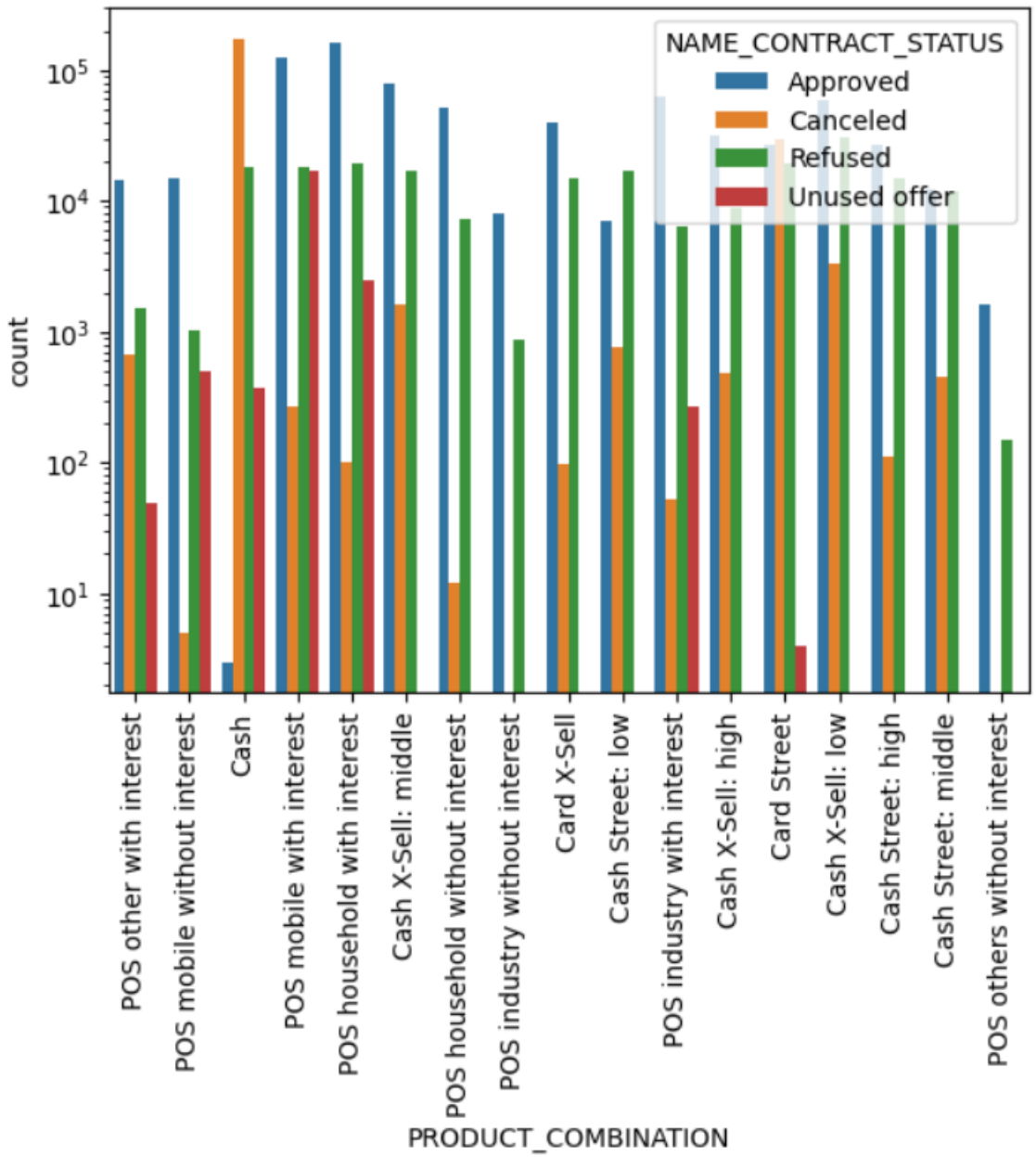
Most clients with all kinds of education types apply mostly for cash loans rather than revolving loans



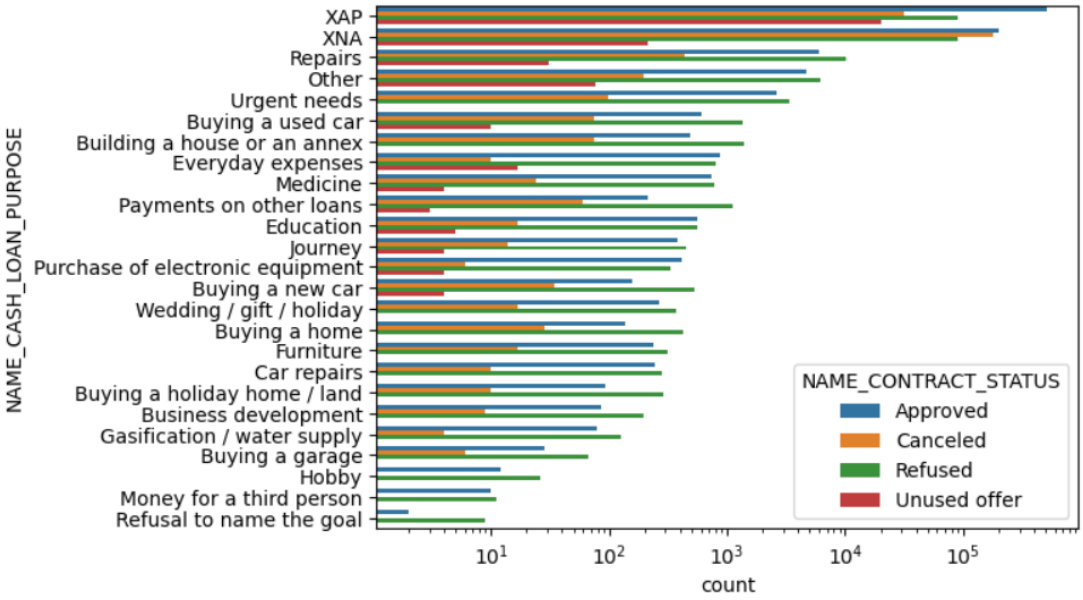
Maximum approved loans are for female clients It can also be observed that male clients use most of the offers of loans as unused offers are less for male clients than that of female clients



Maximum number of contract types are of Product combination of POS household with interest followed by Cash and then followed by POS mobile with interest Least number of clients opt for product combination of POS others without interest

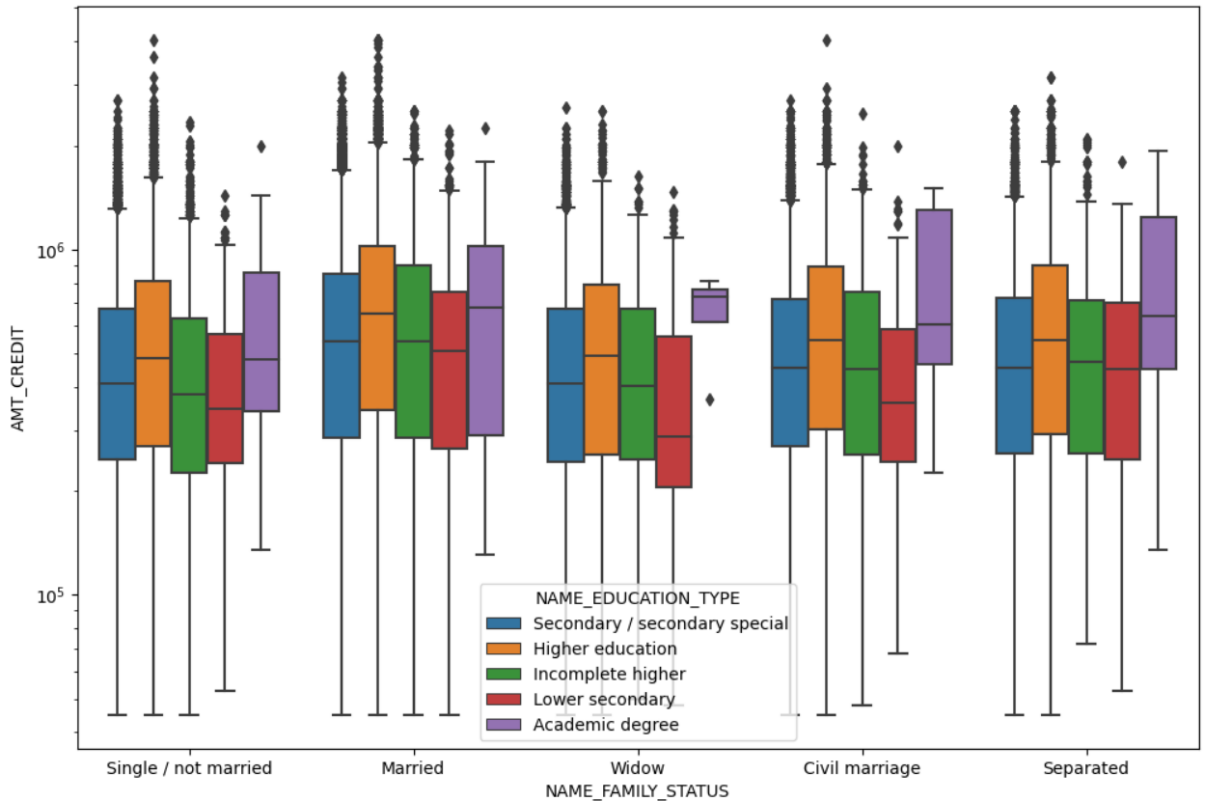


Most canceled laons are of the product combination, Cash. Most refused loans are of product combination Cash X-Sell: low There are almost nill unused offers in product combinations listed below Cash X-Sell: low Card X-Sell Card X-Sell: high Cash Street: high POS household with interest Cash Street: middle Cash X-Sell: middle Cash Street: low Some product combiations are nill incase of unused offers as well as cancelled loans POS industry without interest POS others without interest

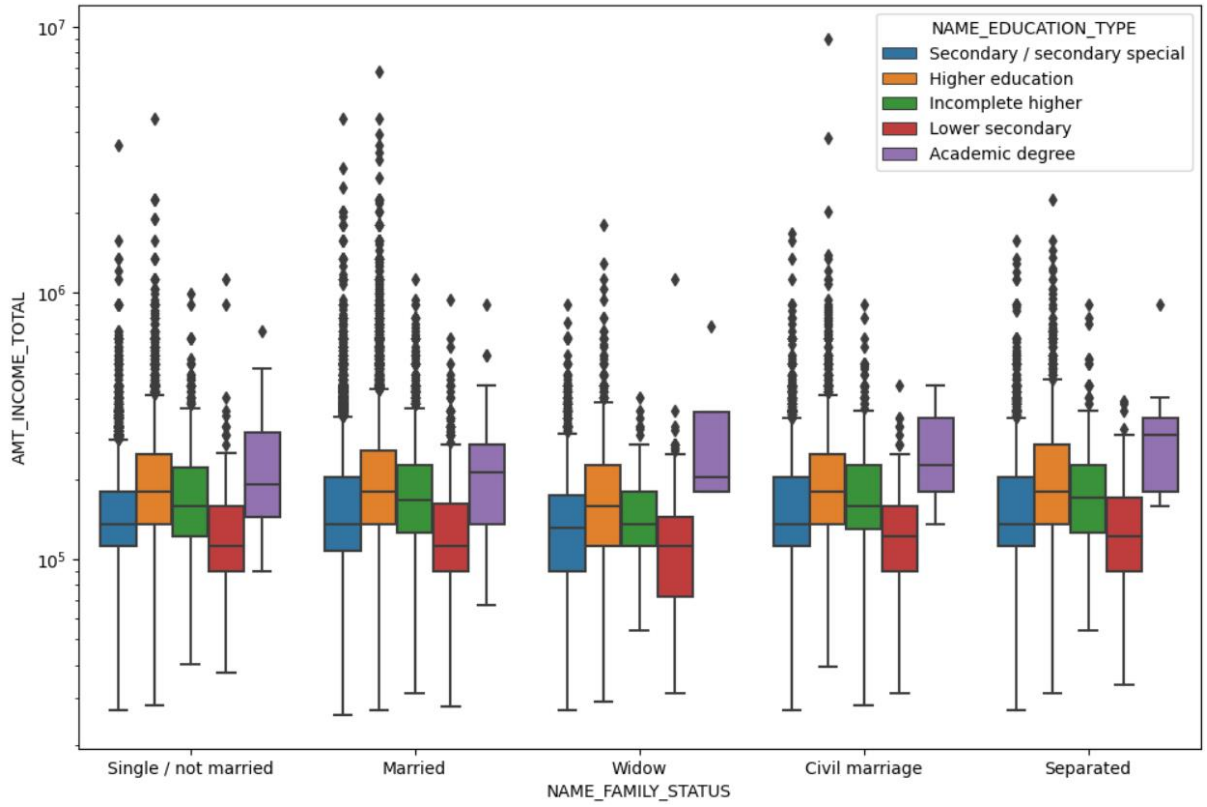


Most rejection of loans came from purpose 'repairs'. For education purposes we have equal number of approves Rejection for Paying other loans and buying a new car are having significant higher rejection than approves.

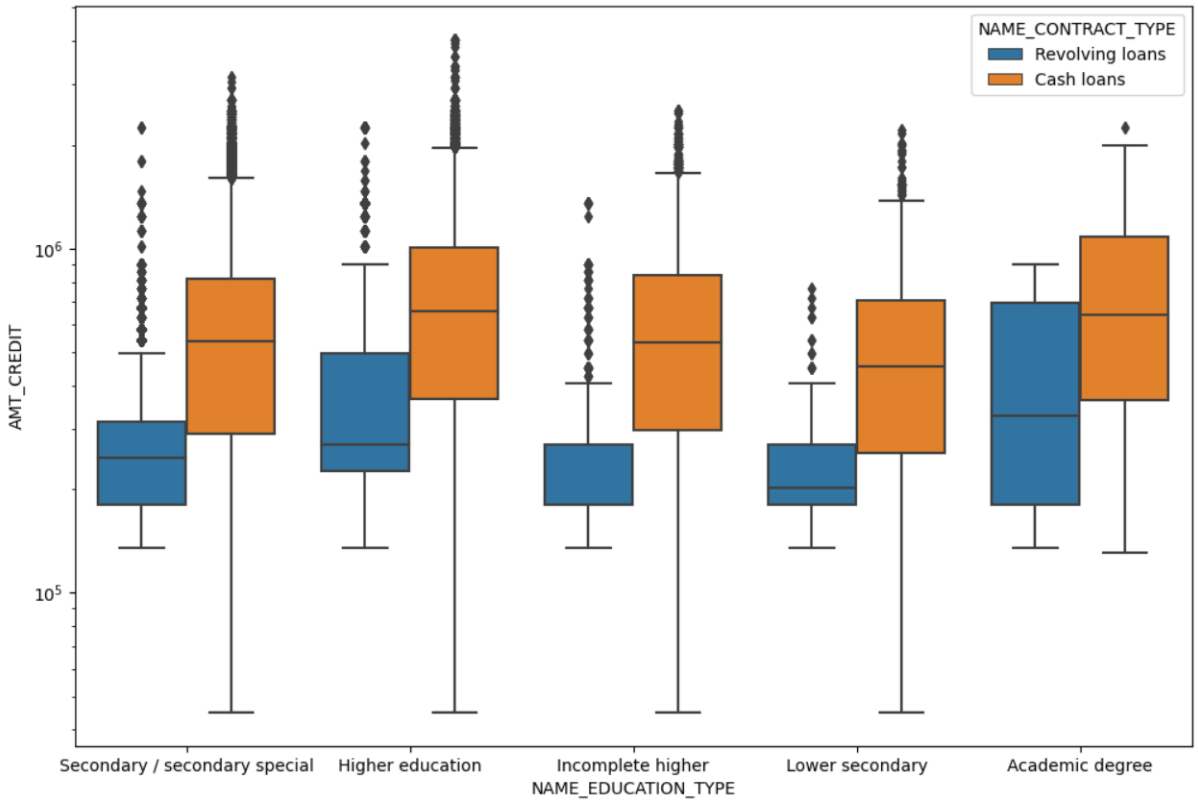
b. Bivariate analysis;



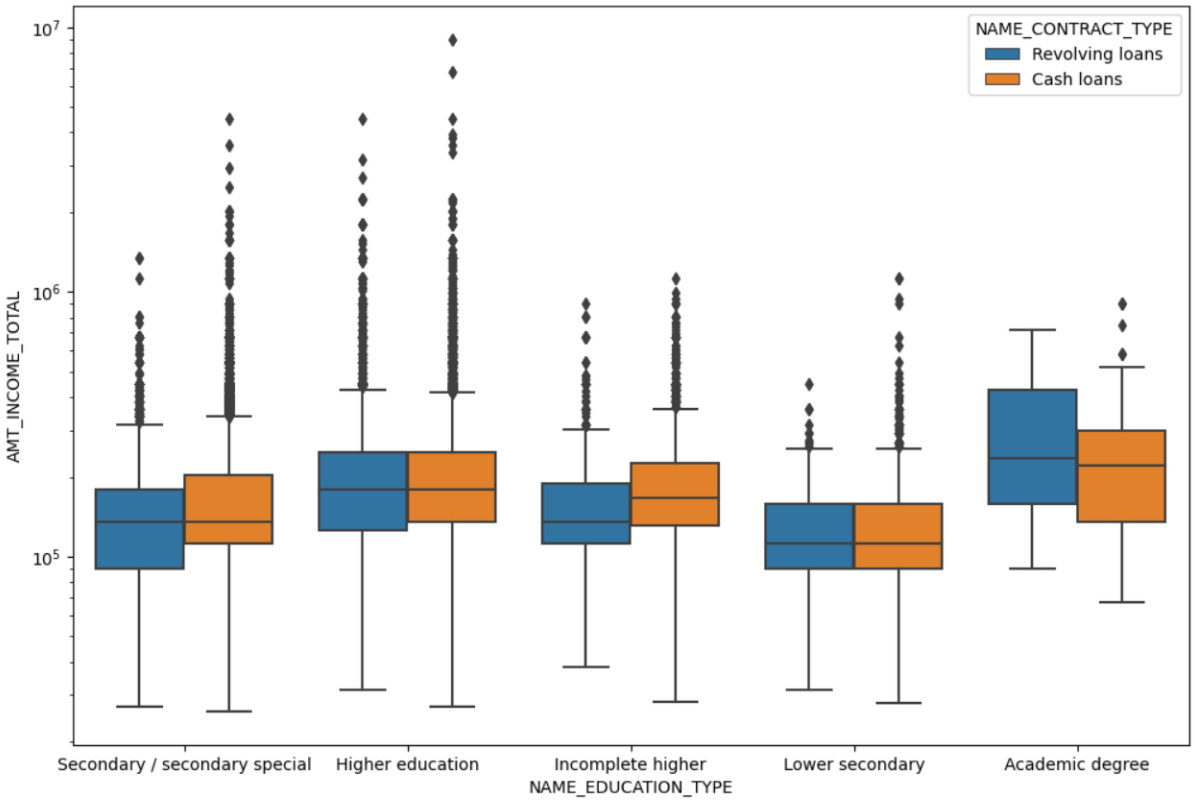
Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others. Also, higher education of family status of 'marriage', 'single or not' and 'civil marriage' are having more outliers. Civil marriage for Academic degree is having most of the credits in the third quartile.



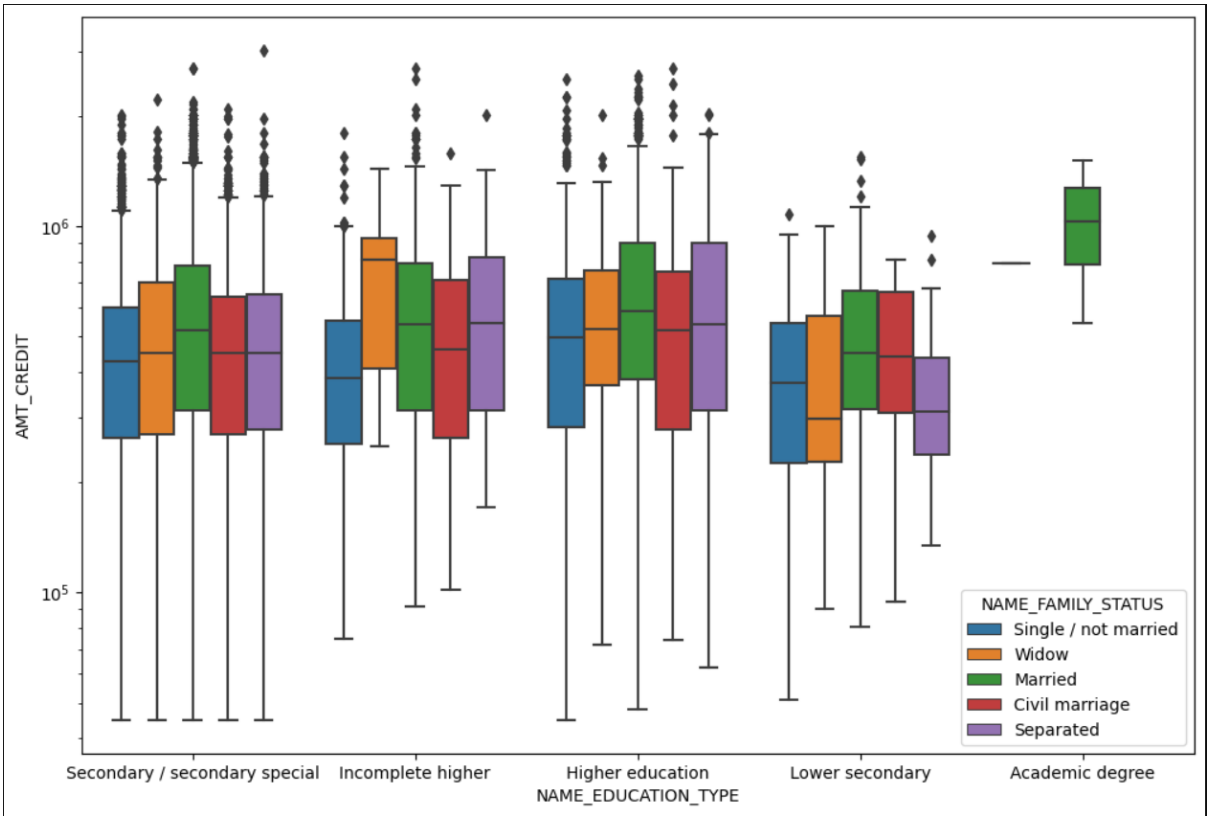
Family status of 'civil marriage', 'marriage' and 'separated' of Higher education are having higher number of income than others. Also, higher education and secondary/second special education statuses with family status of 'marriage', 'single or not' and 'civil marriage' are having more outliers. Married for Higher education is having most of the incomes in the lower bound.



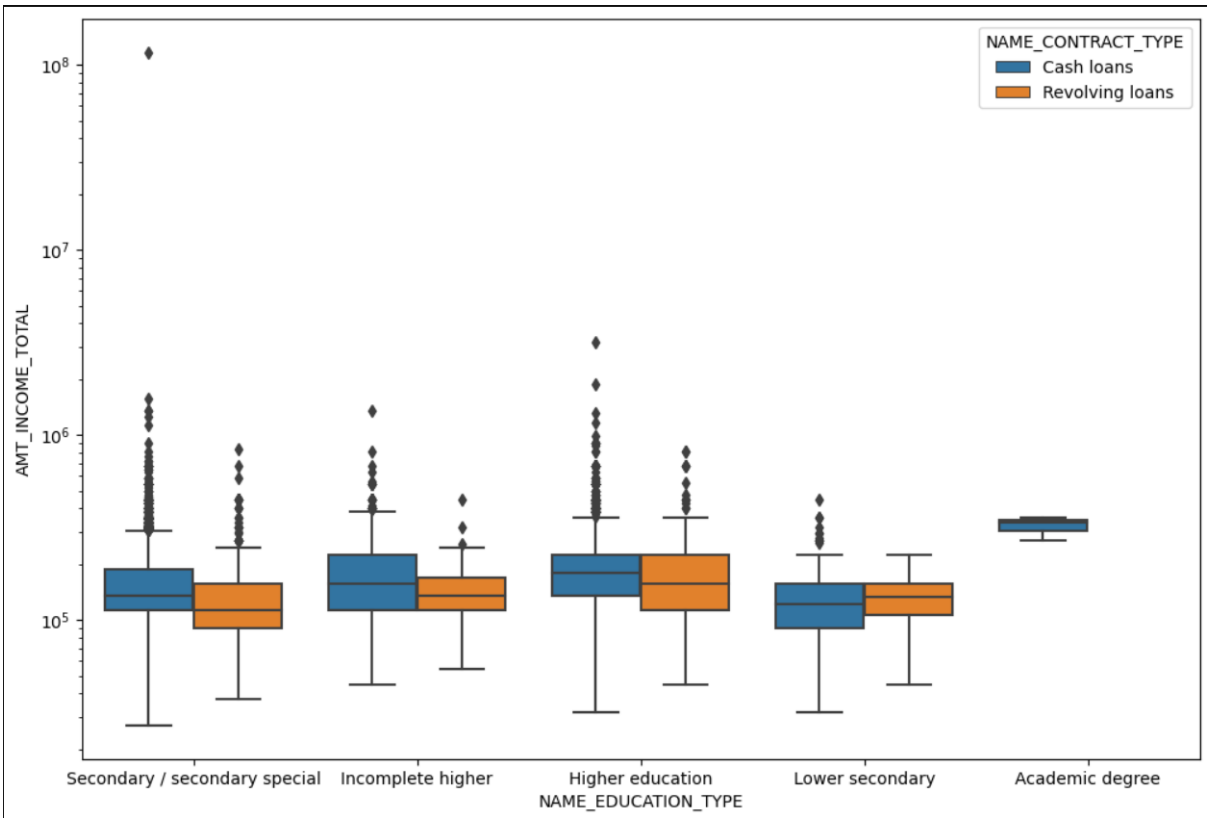
Education status of higher education and secondary/secondary special have most clients for contract type cash loans Most number of clients applying for revolving loans are in education status Acadamic degree and higher education



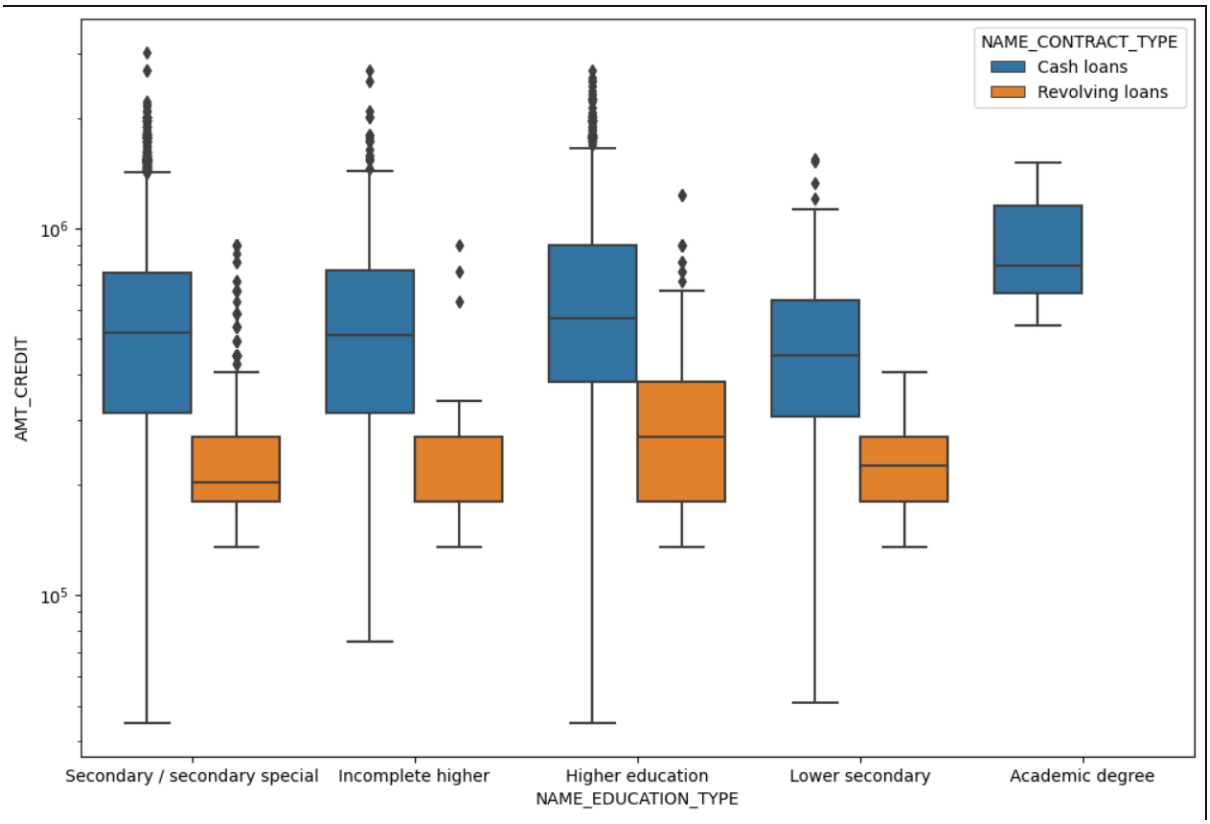
It can be seen that the clients with education status of Higher education are the maximum credit seekers with highest in terms of contract type of cash loans of contract type It can also be observed that the contract type revolving loans issued maximum credit amount holders education status Higher education The hightest credit amount in cash loans is given to a client with education level secondary/secondary special basically education level or type is not playing much role as of who gets what amount of credit



Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having a smaller number of credits than others. Most of the outliers are from Education type 'Higher education' and 'Secondary'. Most number of all types of education as well as family lie in lower bound

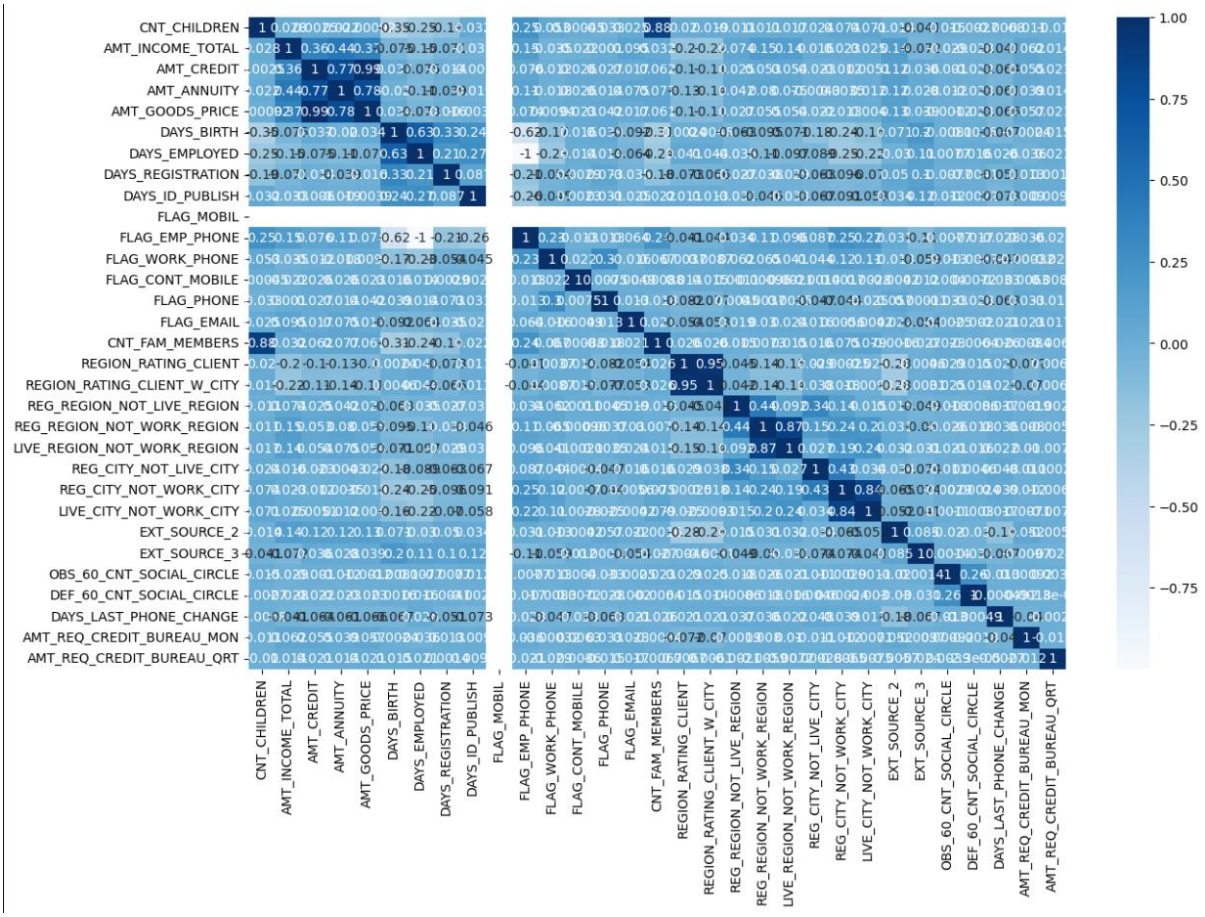


It can be seen that the maximum income amount holders are with education levels secondary and higher educatin levels its strange that, there are no revolving amount loans are demanded by clients with education level academic degree

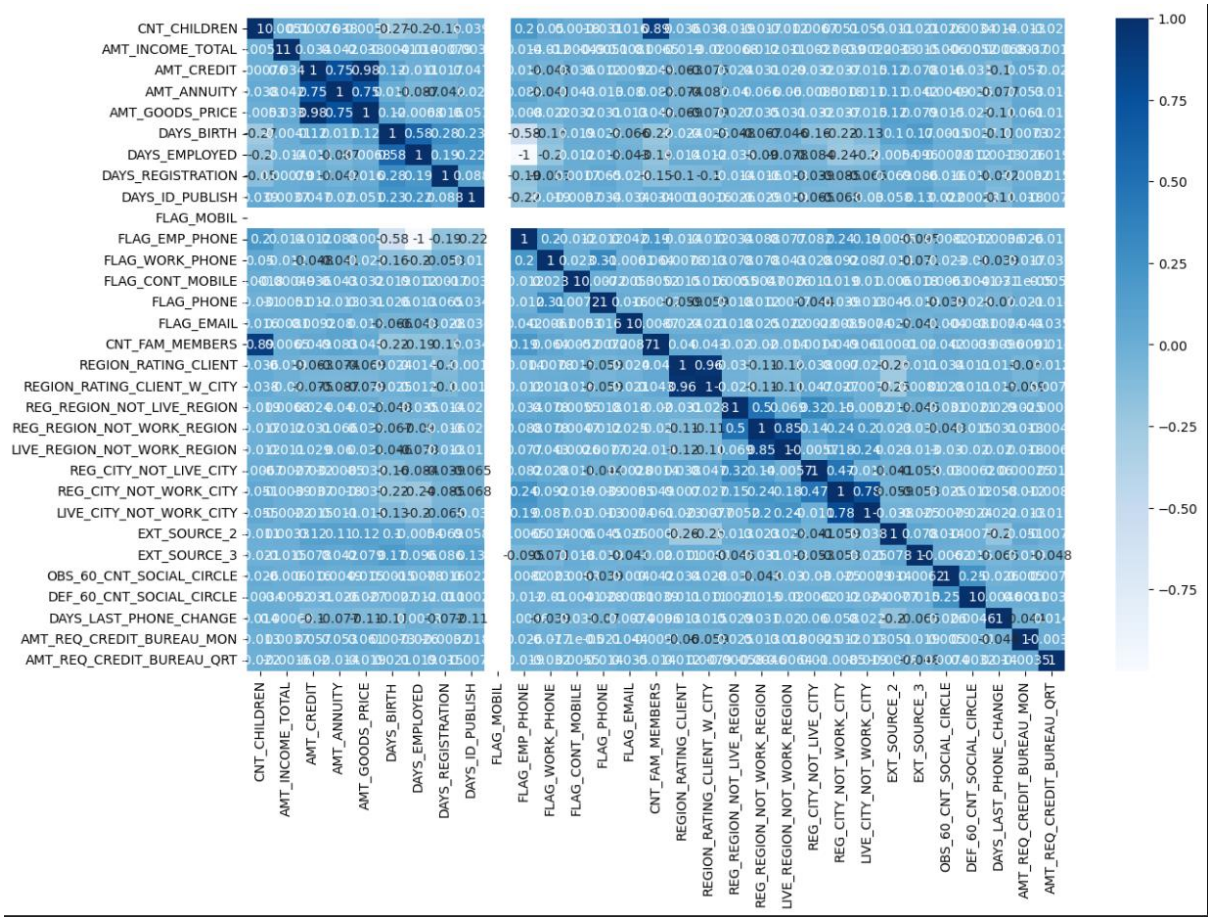


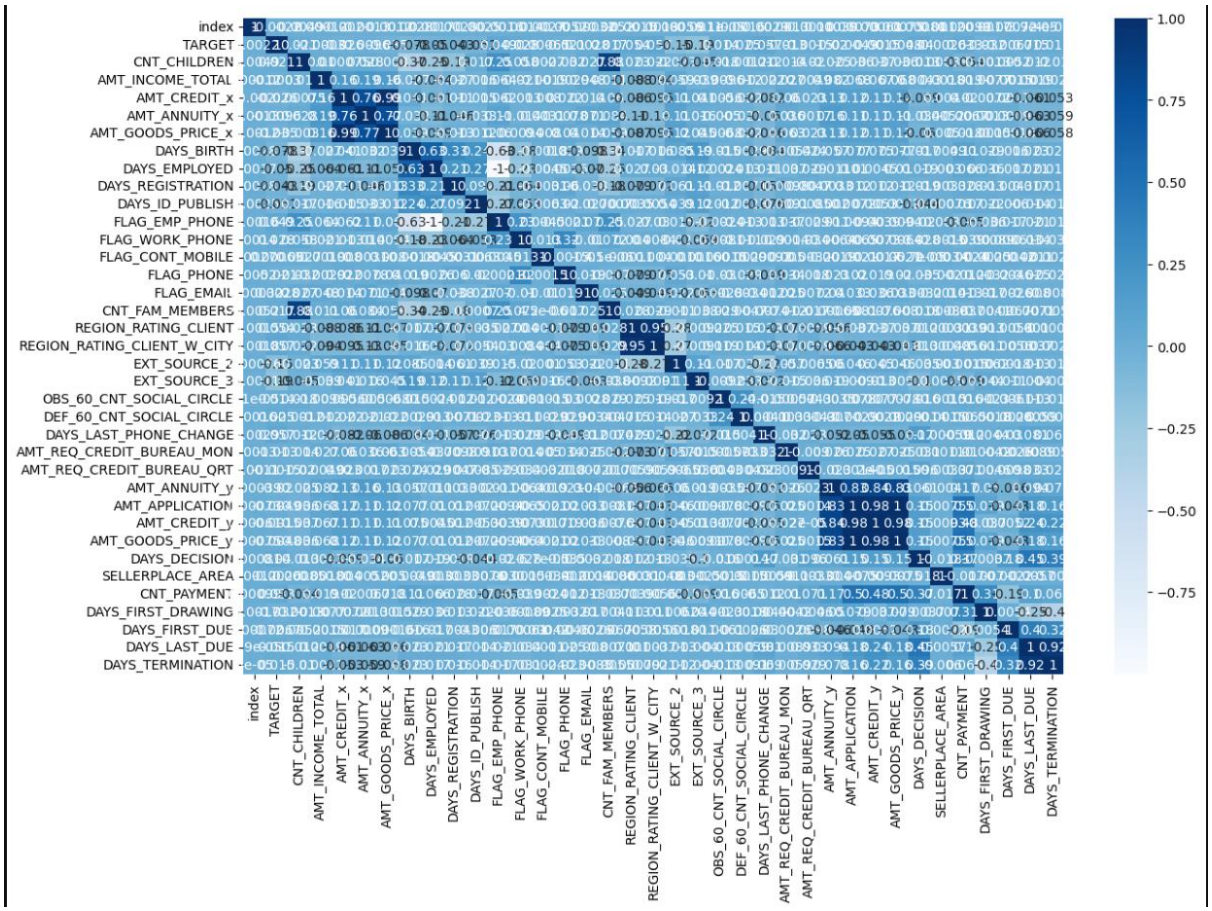
Even though people have difficulties with payments, loan application are still pouring in with similar credit amount demands from the clients

Univariate analysis:



It is observed that the maximum correlation is between the variables listed below in pairs AMT_CREDIT to AMT_ANNUITY AMT_CREDIT to AMT_TOTAL_INCOME AMT_ANNUITY to AMT_INCOME_TOTAL REG_CITY_NOT_WORK_CITY to REG_CITY_NOT_LIVE_CITY DAYS_EMPLOYED to DAYS_BIRTH DAYS_BIRTH to DAYS_REGISTRATION Least or negative correlation is between the variables listed below in pairs DAYS_BIRTH to CNT_CHILDREN


CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
EXT_SOURCE_2
EXT_SOURCE_3
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


index
TARGET
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT_x
AMT_ANNUITY_x
AMT_GOODS_PRICE_x
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
EXT_SOURCE_2
EXT_SOURCE_3
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT
AMT_ANNUITY_y
AMT_APPLICATION
AMT_CREDIT_y
AMT_GOODS_PRICE_y
DAYS_DECISION
SELLERPLACE_AREA
CNT_PAYMENT
DAYS_FIRST_DRAWING
DAYS_FIRST_DUE
DAYS_LAST_DUE
DAYS_TERMINATION
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

We can observe correlations of different variales plotted in the above graph and draw almost similar inference as above, with few extra variables

5. Top 10 correlation:

DAYS_EMPLOYED	FLAG_EMP_PHONE	-0.999762
FLAG_EMP_PHONE	DAYS_EMPLOYED	-0.999762
DAYS_BIRTH	FLAG_EMP_PHONE	-0.621050
FLAG_EMP_PHONE	DAYS_BIRTH	-0.621050
CNT_CHILDREN	DAYS_BIRTH	-0.352385
DAYS_BIRTH	CNT_CHILDREN	-0.352385
CNT_FAM_MEMBERS	DAYS_BIRTH	-0.305681
DAYS_BIRTH	CNT_FAM_MEMBERS	-0.305681
EXT_SOURCE_2	REGION_RATING_CLIENT	-0.284835
REGION_RATING_CLIENT	EXT_SOURCE_2	-0.284835

DAYS_EMPLOYED	FLAG_EMP_PHONE	-0.999641
FLAG_EMP_PHONE	DAYS_EMPLOYED	-0.999641
	DAYS_BIRTH	-0.579718
DAYS_BIRTH	FLAG_EMP_PHONE	-0.579718
	CNT_CHILDREN	-0.272562
CNT_CHILDREN	DAYS_BIRTH	-0.272562
REGION_RATING_CLIENT	EXT_SOURCE_2	-0.255014
EXT_SOURCE_2	REGION_RATING_CLIENT	-0.255014
REGION_RATING_CLIENT_W_CITY	EXT_SOURCE_2	-0.254780
EXT_SOURCE_2	REGION_RATING_CLIENT_W_CITY	-0.254780

DAYS_EMPLOYED	FLAG_EMP_PHONE	-0.999780
FLAG_EMP_PHONE	DAYS_EMPLOYED	-0.999780
DAYS_BIRTH	FLAG_EMP_PHONE	-0.628749
FLAG_EMP_PHONE	DAYS_BIRTH	-0.628749
DAYS_TERMINATION	DAYS_FIRST_DRAWING	-0.400150
DAYS_FIRST_DRAWING	DAYS_TERMINATION	-0.400150
CNT_CHILDREN	DAYS_BIRTH	-0.373633
DAYS_BIRTH	CNT_CHILDREN	-0.373633
	CNT_FAM_MEMBERS	-0.340112
CNT_FAM_MEMBERS	DAYS_BIRTH	-0.340112

Result:

Banks should focus more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ and 'office apartment' for successful payments. Banks should focus less on income types maternity leave and working as they have the greatest number of unsuccessful payments in loan purpose ‘Repairs’:

Although having higher number of rejections in loan purposes with 'Repairs' we can observe difficulties in payment. There are few places where loan payment diffuculty is significantly high.

Bank should continue to be cautious while giving loan for this purpose. Bank can focus mostly on housing type with parents, House or apartment and municipal apartment with purpuse of education, buying land, buying a garage, purchase of electronic equipment and some other purposes with target0 significantly more than target1 for successful payments. Banks can offer more offers to clients who are students and pensioners as they take all offers and are more likely to pay back

IMPACT OF CAR FEATURES

Project Description:

As we have a vast data of various car features from various brands, we are providing some insights to our internal team using those data so that we could classify the customers and take the decision wisely.

Approach:

I have the dataset with various features like transmission type, engine hp, highway mpg for different models by brands for every year. So after imputing the null values using some domain knowledge and statistics based on the features I have tried to provide some answers so that they can take more wise decisions regarding future.

Tech-Stack Used:

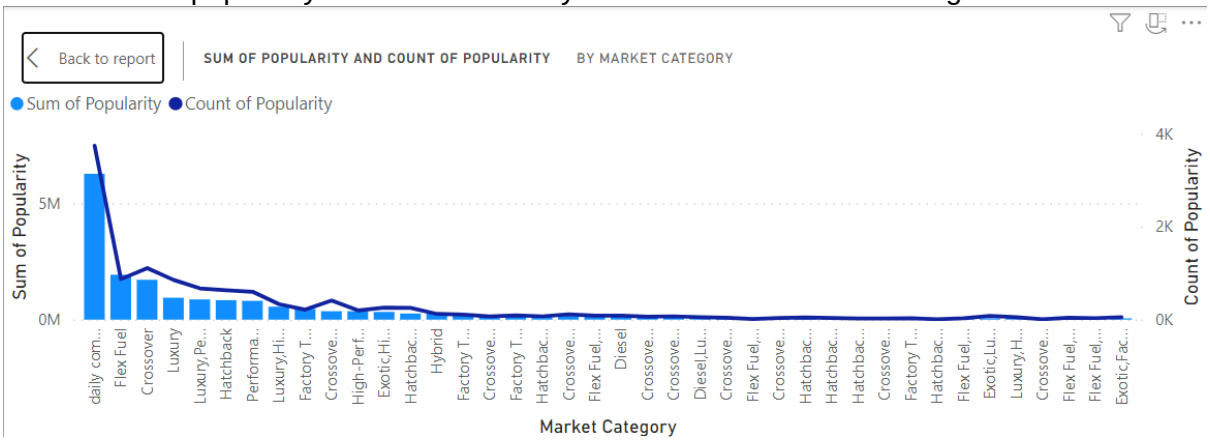
I have used Jupyter notebook and Power BI here.

Insights:

(FOR MORE INSIGHTS PLEASE VISIT THE DASHBOARDS)

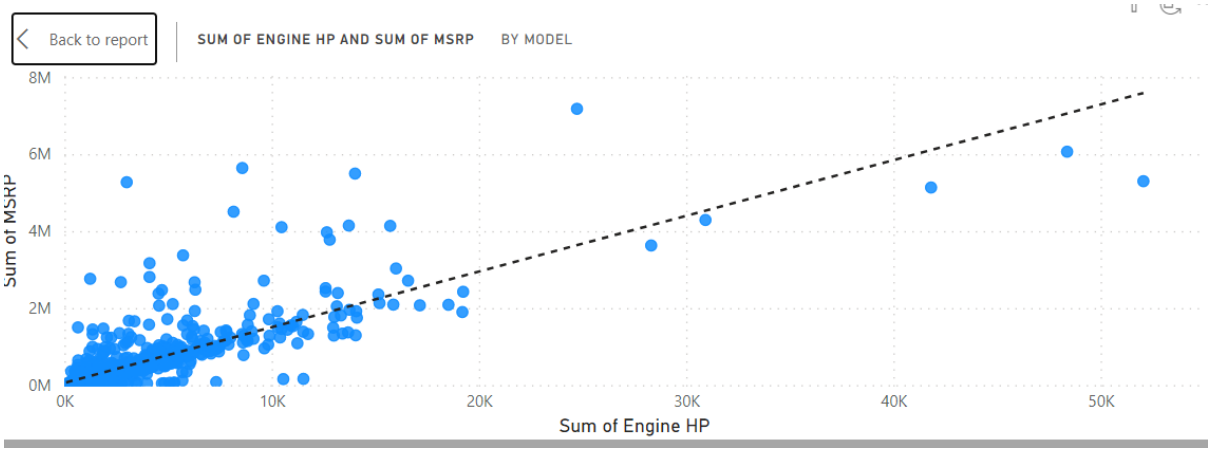
Dashboard 1:

1. How does the popularity of a car model vary across different market categories?



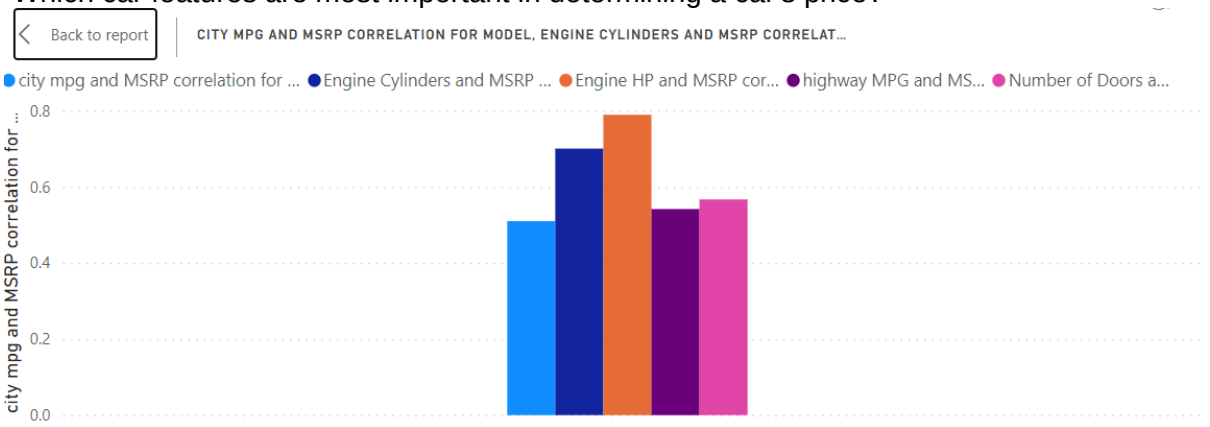
I can say by looking into this graph that daily commuting vehicles have the highest popularity among the market categories
(Pivot table is already in power bi dashboard)

2. What is the relationship between a car's engine power and its price?



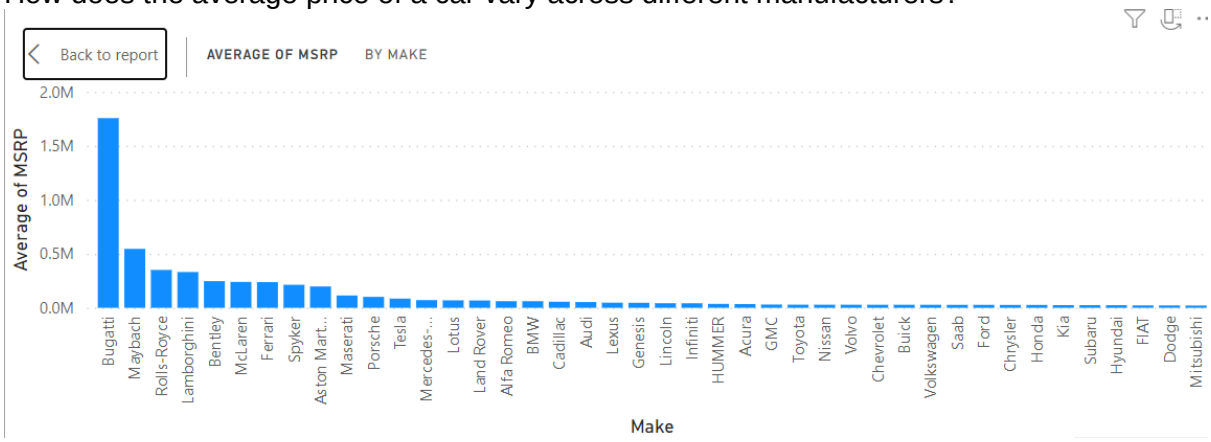
So, If engine power increases price also increases, so they have positive relation

3. Which car features are most important in determining a car's price?



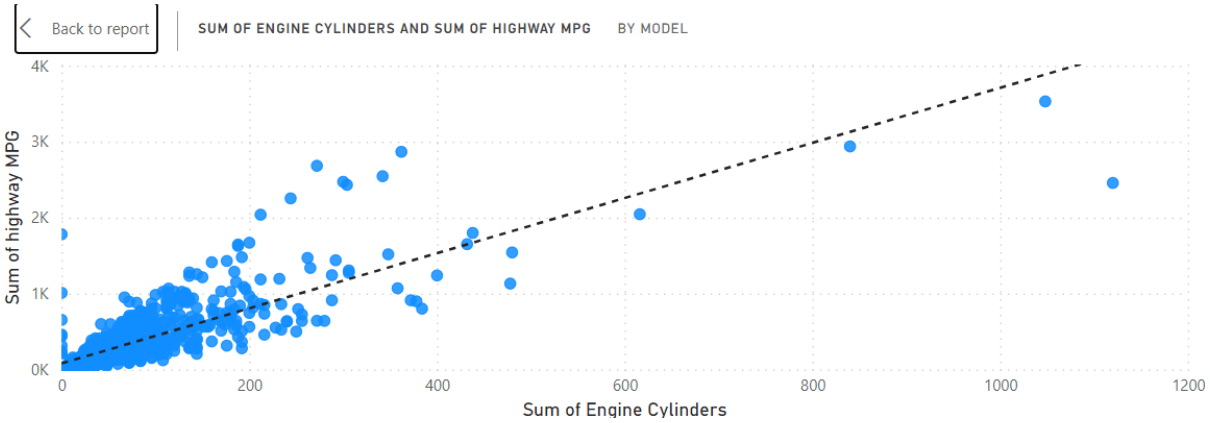
Based on the correlation coefficient computed using power bi quick measure I found out that these are the five features reasonable for price change

4. How does the average price of a car vary across different manufacturers?



I can say by looking into this graph that Bugatti has the highest average price (Pivot table is already in power bi dashboard)

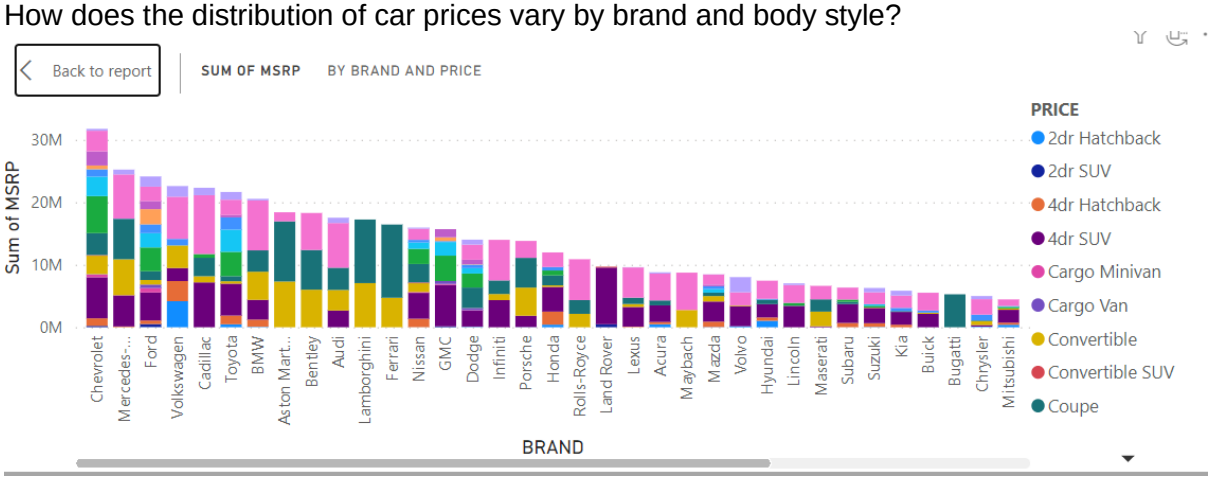
5. What is the relationship between fuel efficiency and the number of cylinders in a car's engine?



So, if number of cylinders increases highway mileage also increases as they are positively correlated and value is 0.81

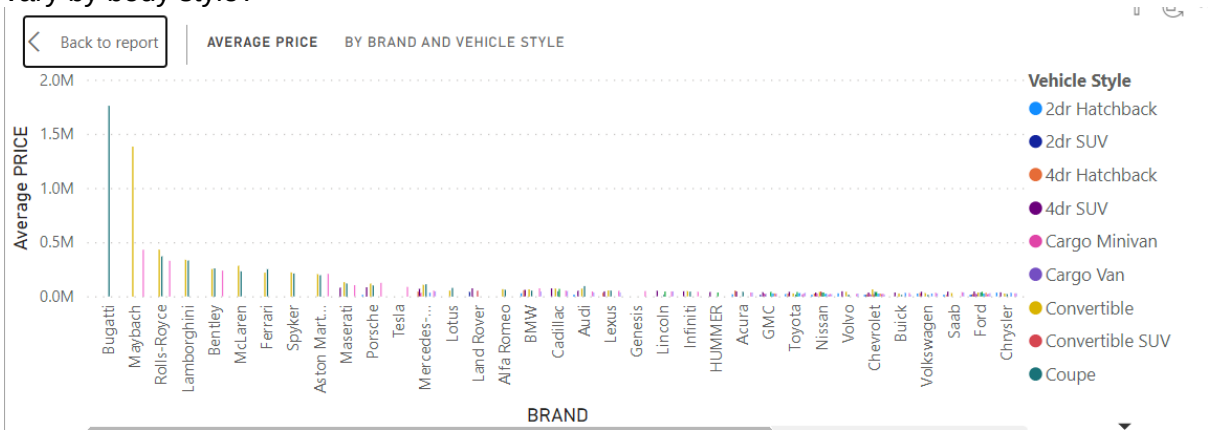
Dashboard 2:

1. How does the distribution of car prices vary by brand and body style?



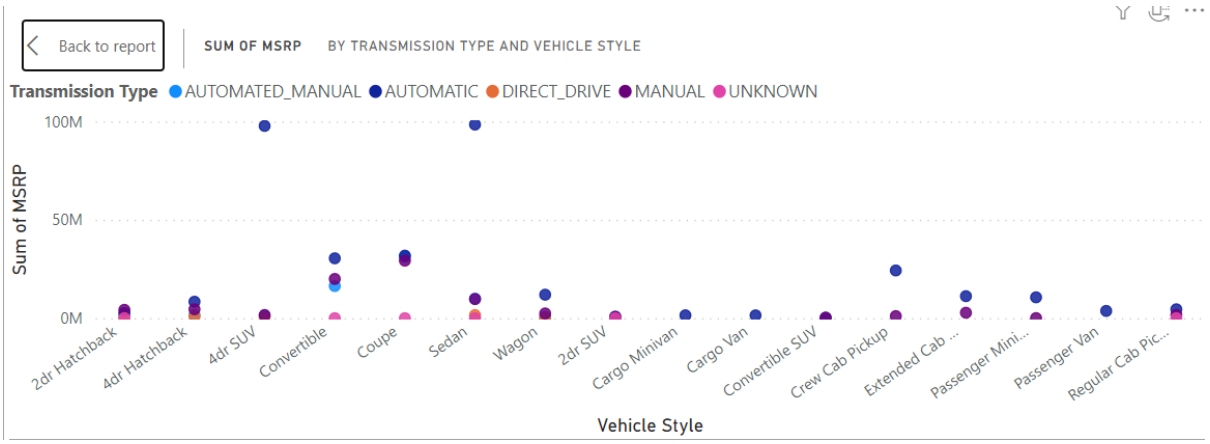
Chevrolet is the costliest brand and Genesis is the most budget friendly brand

2. Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?



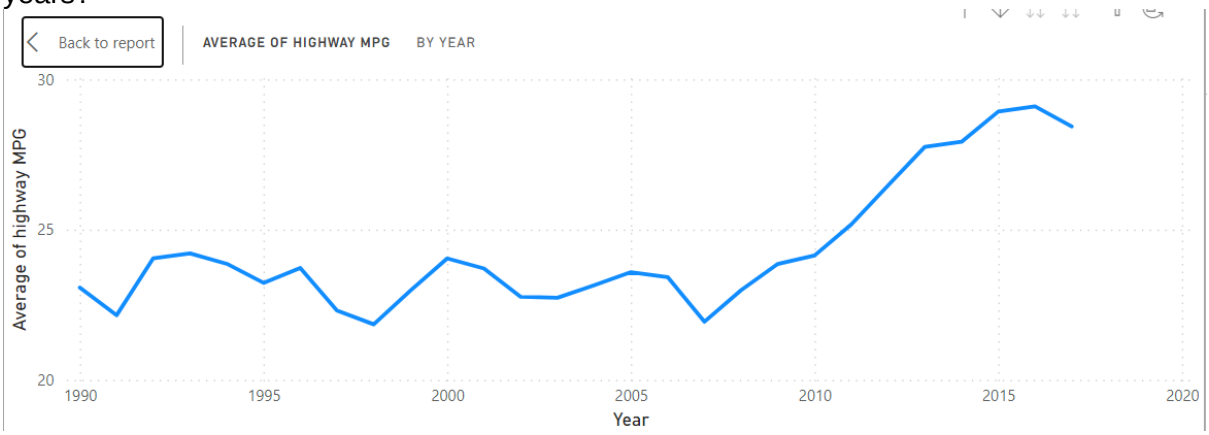
Bugatti has the highest average price and Plymouth has the lowest average price

3. How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?



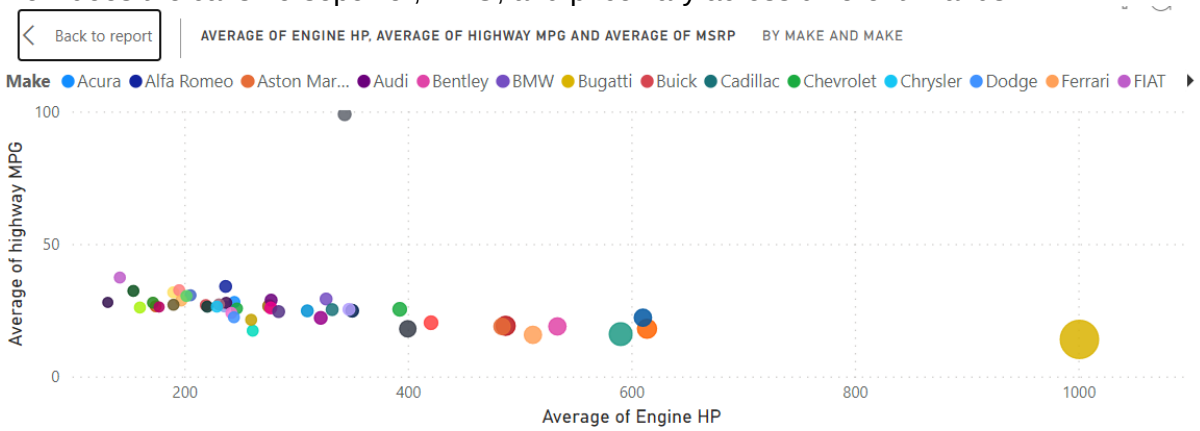
Automatic 4dr SUVs and Sedans are most expensive cars

4. How does the fuel efficiency of cars vary across different body styles and model years?



Because of technological advancement mileage is started increasing after 2007

5. How does the car's horsepower, MPG, and price vary across different Brands?



I can say that with the increase in engine power there is a decrease in mileage

Result:
So as daily commuting vehicles are more popular than expensive ones we can focus to develop more features in the daily commute segment with a slight increase in price as consumers are willing to pay more for better upgrades in terms of features and mileage.

ABC CALL VOLUME TREND

Project Description:

As we have a vast data of customer experience of an inbound call centre, we are providing some insights to our internal team using those data so that we could rectify what is their optimal employee requirements and take decisions wisely.

Approach:

I have the dataset with various features like Agent_Name, Agent_ID, Queue_Time [duration for which customer have to wait before they get connected to an agent], Time [time at which call was made by customer in a day], Time_Bucket [for easiness we have also provided you with the time bucket], Duration [duration for which a customer and executives are on call, Call_Seconds [for simplicity we have also converted those time into seconds], call status (Abandon, answered, transferred). So, after imputing the null values using some domain knowledge and some tweaks, I have tried to provide some answers so that they can take more wise decisions regarding future.

Tech-Stack Used:

I have used Jupyter Notebook and Microsoft Excel here.

Insights:

(FOR MORE DETAILS, PLEASE VISIT THE EXCEL FILE)

Assumption:

An agent work for 6 days a week; On an average total unplanned leaves per agent is 4 days a month; An agent total working hrs is 9 Hrs out of which 1.5 Hrs goes into lunch and snacks in the office. On average an agent occupied for 60% of his total actual working Hrs (i.e 60% of 7.5 Hrs) on call with customers/ users. Total days in a month is 30 days.

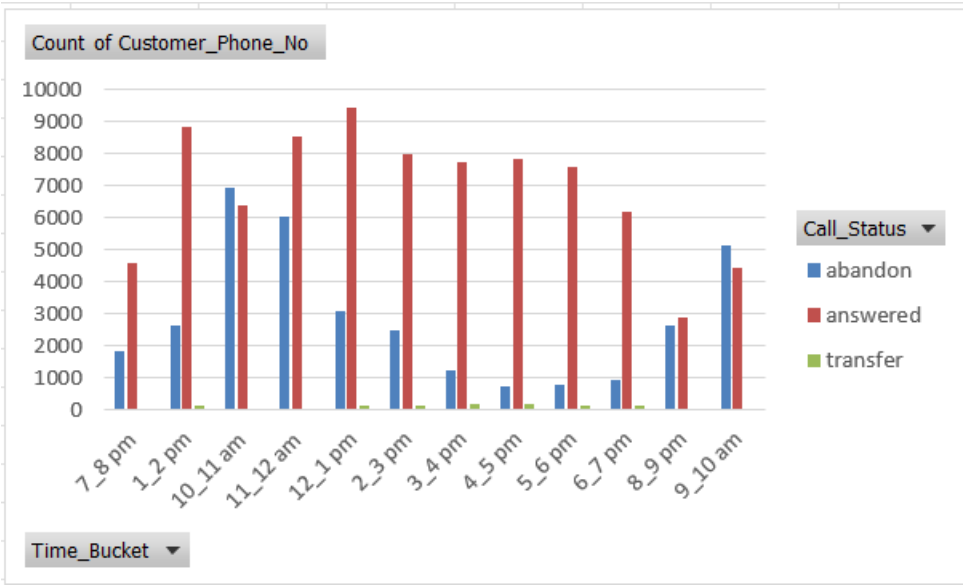
1. Calculate the average call time duration for all incoming calls received by agents (in each Time_Bucket).

Call_Status	answered
Row Labels	**Average of Call_Seconds (s)**
7_8 pm	203.4060725
1_2 pm	194.7401744
10_11 am	203.3310302
11_12 am	199.2550234
12_1 pm	192.8887829
2_3 pm	193.6770755
3_4 pm	198.8889175
4_5 pm	200.8681864
5_6 pm	200.2487831
6_7 pm	202.5509677
8_9 pm	202.845993
9_10 am	199.0691057
Grand Total	**198.6227745**

From here I can say that average answered call duration during day are almost the same for every time bucket

2. Show the total volume/ number of calls coming in via charts/ graphs [Number of calls v/s Time]

Count of Customer_Phone_No	**Column Labels**			
Row Labels	**abandon**	**answered**	**transfer**	**Grand Total**
7_8 pm	1848	4578	37	6463
1_2 pm	2617	8829	115	11561
10_11 am	6911	6368	34	13313
11_12 am	6028	8560	38	14626
12_1 pm	3073	9432	147	12652
2_3 pm	2475	7974	112	10561
3_4 pm	1214	7760	185	9159
4_5 pm	747	7852	189	8788
5_6 pm	783	7601	150	8534
6_7 pm	933	6200	105	7238
8_9 pm	2625	2870	10	5505
9_10 am	5149	4428	11	9588
Grand Total	**34403**	**82452**	**1133**	**117988**



From this chat I can say,

1. 12_1 pm time bucket has the highest answered call volume
2. 8_9 pm time bucket has the lowest answered call volume
3. Most of the abandoned calls coming from 9_10 am,10_11 am, 11_12 am

3. As you can see current abandon rate is approximately 30%. Propose a manpower plan required during each time bucket [between 9am to 9pm] to reduce the abandon rate to 10%.

Row Labels	Count of Customer_Phone_No	Count of Customer_Phone_No2
abandon	34403	29.16%
answered	82452	69.88%
transfer	1133	0.96%
Grand Total	**117988**	**100.00%**

I can see 29% of calls are getting abandoned and 1 % are getting transferred

Row Labels	Sum of Call_Seconds (s)	hours
1/1/2022	676664	187.9622222
1/10/2022	778739	216.3163889
1/11/2022	785717	218.2547222
1/12/2022	709934	197.2038889
1/13/2022	691320	192.0333333
1/14/2022	564227	156.7297222
1/15/2022	556267	154.5186111
1/16/2022	674394	187.3316667
1/17/2022	945615	262.6708333
1/18/2022	796768	221.3244444
1/19/2022	750270	208.4083333
1/2/2022	574003	159.4452778
1/20/2022	759613	211.0036111
1/21/2022	639855	177.7375
1/22/2022	621577	172.6602778
1/23/2022	553899	153.8608333
1/3/2022	812863	225.7952778
1/4/2022	861946	239.4294444
1/5/2022	846798	235.2216667
1/6/2022	829040	230.2888889
1/7/2022	757019	210.2830556
1/8/2022	735444	204.29
1/9/2022	541147	150.3186111

From here I can calculate that on an average how much hours required to answer all the calls

I.e., 198.8299396 hours

Daily work hours by employees are (9-1.5) x 60/100= 4.5 hours [based on the assumptions]

So, actual employee strength which is making sure that 60% calls are getting answered is,

198.8299396/4.5 ≈ 44

So, now required employee strength for having 90% of the calls answered is,

90 x 44 x 60 = 66

The distribution of optimal employee strength for each time bucket:

Row Labels	Sum of Call_Seconds (s)	hours	ratio	required employee strength
7_8 pm	934437	259.5658333	0.056759415	4
1_2 pm	1728843	480.2341667	0.10501309	7
10_11 am	1297006	360.2794444	0.07878252	5
11_12 am	1708079	474.4663889	0.103751847	7
12_1 pm	1831061	508.6280556	0.111221999	7
2_3 pm	1552143	431.1508333	0.094280009	6
3_4 pm	1556085	432.2458333	0.094519453	6
4_5 pm	1594489	442.9136111	0.096852182	6
5_6 pm	1533769	426.0469444	0.093163938	6
6_7 pm	1261762	350.4894444	0.076641735	5
8_9 pm	583250	162.0138889	0.035427673	2
9_10 am	882195	245.0541667	0.05358614	4
Grand Total	**16463119**	**4573.088611**	**1**	**66**

4. Let's say customers also call this ABC insurance company in night but didn't get answer as there are no agents to answer, this creates a bad customer experience for this Insurance company. Suppose every 100 calls that customer made during 9 Am to 9 Pm, customer also made 30 calls in night between interval [9 Pm to 9 Am] and distribution of those 30 calls are as follows:

Distribution of 30 calls coming in night for every 100 calls coming in between 9am - 9pm (i.e. 12 hrs slot)											
9pm- 10pm	10pm - 11pm	11pm- 12am	12am- 1am	1am - 2am	2am - 3am	3am - 4am	4am - 5am	5am - 6am	6am - 7am	7am - 8am	8am - 9am
3	3	2	2	1	1	1	1	3	4	4	5

Now propose a manpower plan required during each time bucket in a day. Maximum Abandon rate assumption would be same 10%.

Count of Customer_Phone_No	Column Labels			
Row Labels	abandon	answered	transfer	Grand Total
1/1/2022	684	3883	77	4644
1/10/2022	1212	3699	72	4983
1/11/2022	856	3695	86	4637
1/12/2022	1299	3297	47	4643
1/13/2022	738	3326	59	4123
1/14/2022	291	2832	32	3155
1/15/2022	304	2730	24	3058
1/16/2022	1191	3910	41	5142
1/17/2022	16636	5706	5	22347
1/18/2022	1738	4024	12	5774
1/19/2022	974	3717	12	4703
1/2/2022	356	2935	60	3351
1/20/2022	833	3485	4	4322
1/21/2022	566	3104	5	3675
1/22/2022	239	3045	7	3291
1/23/2022	381	2832	12	3225
1/3/2022	599	4079	111	4789
1/4/2022	595	4404	114	5113
1/5/2022	536	4140	114	4790
1/6/2022	991	3875	85	4951
1/7/2022	1319	3587	42	4948
1/8/2022	1103	3519	50	4672
1/9/2022	962	2628	62	3652
AVERAGE	**1496**	**3585**	**49**	**5130**

so, the average calls at day is around 5130

so, the average calls at night would be around 5130*30/10 ≈ 1539

required extra time is 1539 x 198.82 x 90 /100/3600 = 76.495995 hours

therefore, required extra man power 76.495995/4.5 =17

Time_Bucket	Call_Distribution	Ratio	required employee strength
7_8 am	3	0.1	2
1_2 am	3	0.1	2
10_11 pm	2	0.066666667	1
11_12 pm	2	0.066666667	1
12_1 am	1	0.033333333	1
2_3 am	1	0.033333333	1
3_4 am	1	0.033333333	1
4_5 am	1	0.033333333	1
5_6 am	3	0.1	2
6_7 am	4	0.133333333	2
8_9 am	4	0.133333333	2
9_10 pm	5	0.166666667	3
Total	**30**	**1**	**17**

Results:

So, as we can see at day in the first half there are more abandoned rate so while increasing the employee strength there, we can also investigate what causing the problem and may be after reaching to the final verdict lower employee strength can also attend all the calls.

LEARNINGS

These projects have forced me to learn technologies in a better way to cope up with data industry and have made me confident enough to accept any challenges for future endeavours.

APPENDIX

DATA ANALYTICS PROJECT:
https://drive.google.com/drive/folders/1zC2uYvuxVIS9B66KkEdzAa55TKEOxzlS?usp=drive_link

INSTAGRAM USER ANALYTICS:

https://drive.google.com/drive/folders/13nIEjPQJOqi7osNn8flrSP-Eqy9BAQOR?usp=drive_link

OPERATION AND METRIC ANALYTICS:
https://drive.google.com/drive/folders/101gNbSbeHCpN5YktRtm_sA0BKQW3Ph6r?usp=sharing

HIRING PROCESS ANALYTICS:
https://drive.google.com/drive/folders/10_7CU8AEiAyR2gTzZNM1s63t_2apzBT4?usp=sharing

IMDB MOVIE ANALYSIS:
https://drive.google.com/drive/folders/1qaa26tK532W4d8po5LMYg6W1Qqp9c1FD?usp=sharing

BANK LOAN CASE STUDY:
https://drive.google.com/drive/folders/1zgAoqx-sVU2GS25jskBs-NuYULCPYwpa?usp=sharing

IMPACT OF CAR FEATURES:
https://drive.google.com/drive/folders/1S4euoHdIEW_F3pV8bluZLP7NBEKynqjU?usp=sharing

ABC CALL VOLUME TREND:
https://drive.google.com/drive/folders/1voOrjS5A-_Ye0f-sgNMS0Y4ewVtfE4sA?usp=sharing