DATA ANALYSIS

PORTFOLIO

Prepared by:
Gnana Rangasai Nath . V

Professional Background

- Graduated with First Class Distinction (70.1%) in B.Tech – CSE. Gained knowledge on various skills which includes Python, SQL, ML, Regular Expressions, Excel, Tableau along with good analytical and presentation skills.

- Worked as a Team Lead and specialist in Movate (formerly as CSS Corp) where my work included in building JSON configs and manipulating queries with Regular Expressions to scrape the data as needed from websites and JSON files. Designed required Dashboards with Broadstreet Analytics (an application of Tableau).

- Joined as a Data Analyst Intern and worked on multiple Live Projects which also included datasets from Kaggle. Being experienced in professional sector and by handling workload along with my team effectively and consistently, I think I am adaptive and flexible in handling critical situations

Table of Contents

Data Analytics Process

Description:

- We use Data Analytics in everyday life without even knowing it.
- Your task is to give the example(s) of such a real-life situation where we use Data Analytics and link it with the data analytics process.

Design:

- Based on the six steps involved in Data Analytics, we used the following real-life scenario to describe them:
  - Choosing a Bachelor's Degree

Data Analytics Process

Choosing a Bachelor's Degree:

The six steps involved in the data analytics process are as follows:

Plan:

We will aim to get a Bachelor's Degree based on the education received on either Mathematical Sciences (Like picking Bachelor's in Engineering or Technology etc.) or Biological Sciences (E.g. Bachelor's in Medical Field like Pharmacy/Dental/Agriculture etc.)

Prepare:

After this, one would draw up the conditions such as choosing a good university and cost associated with the relative expenses.

Process:

Based on the factor of education, one has to select a particular specialization like CSE/ECE/Civil etc.. in Bachelors of Technology.

Analyze:

We would take many things into consideration for selecting a particular field of specialization based on interest, compatibility and future outcomes/expectations on that particular field.

Share:

We will then discuss the same with family and friends or consultants whether the selected University and course is in the individual's best interests.

Act:

After putting together all the pieces of information, one would act to obtain the knowledge on the chosen specialization by going through any entrance examinations that are needed etc.

Data Analytics Process

Conclusion:

- The listed example above can be considered as a frequent and a good real life scenario where data analysis is followed without knowing the actual process and vice-versa can be applied as well.
- There are many such scenarios in the real world where this process can be applied.
- Some of the other cases would be as such:
  - Buying a smartphone/Laptop
  - Developing a plot or buying a house
  - Purchasing automobiles

Instagram User Analytics

Description:

- User analysis is the process by which we track how users engage and interact with our digital product (software or mobile application) in an attempt to derive business insights for marketing, product & development teams.
- These insights are then used by teams across the business to launch a new marketing campaign, decide on features to build for an app, track the success of the app by measuring user engagement and improve the experience altogether while helping the business grow.
- You are working with the product team of Instagram and the product manager has asked you to provide insights on the questions asked by the management team.

Instagram User Analytics

Problem:

- There are two parts consisting 7 queries in total which are analysed from the given dataset.
  - Marketing Queries
    - Rewarding Most Loyal Users
    - Remind Inactive Users to start posting
    - Declaring Contest Winner
    - Hashtag Researching
    - Launching AD campaign
  - Investor Metric Queries
    - User Engagement
    - Bots & Fake Accounts

Design:

- Some of the steps listed in Data Analytic process are implemented in answering the queries.
- The tasks and sample Instagram Database are already provided.
- Have to provide the needed analysis with related SQL queries.
- I have used MySQL Workbench to load the data set and answer the queries as listed.

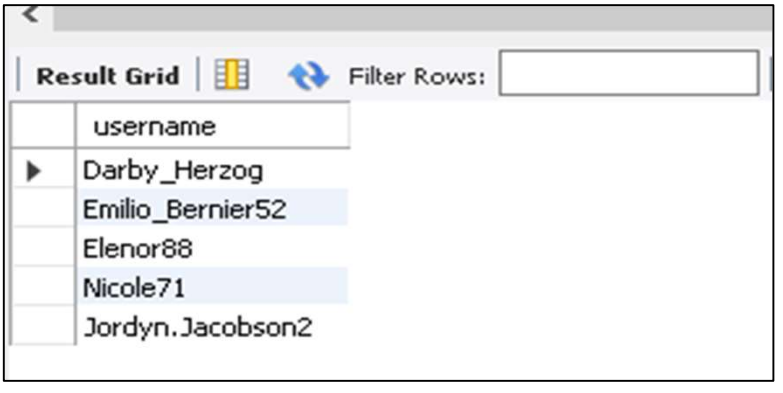
Instagram User Analytics

Findings:

Marketing Analysis:

Rewarding the Most Loyal Users

- This is to find the 5 oldest users in the given Instagram DB .
- I have used the ORDER BY and LIMIT keywords to sort the users based on their account creation and to limit the required number of users.

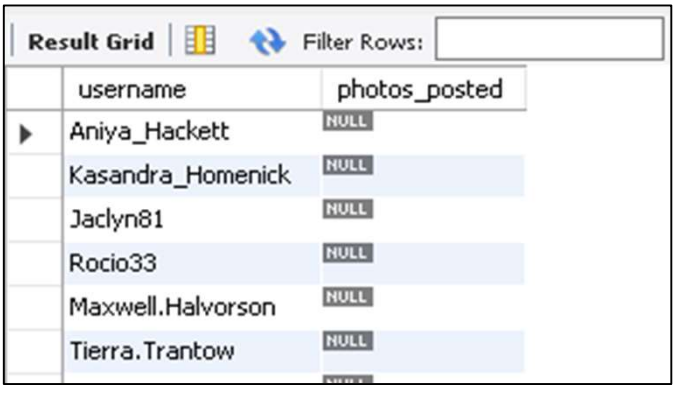


Result Grid | Filter Rows:

username
Darby_Herzog
Emilio_Bernier52
Elenor88
Nicole71
Jordyn.Jacobson2

Remind Inactive users to start posting

- This is to find the users who have not posted a single photo .
- I have joined “users” & “photos” tables using LEFT JOIN on users table to select all the values in users and added IS NULL condition on IDs in “photos” table such that the users who has no ID linked with any photo are selected.



Result Grid | Filter Rows:

username	photos_posted
Aniya_Hackett	NULL
Kasandra_Homenick	NULL
Jaclyn81	NULL
Rocio33	NULL
Maxwell.Halvorson	NULL
Tierra.Trantow	NULL

Instagram User Analytics

Remind Inactive users to start posting

- The table below shows 26 users who have not yet posted a single photo.

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

Declaring Contest Winner

- This is to find the user who got the most likes for the posted photo.
- This task requires combining three tables “users”, “photos” & “likes”.
- I have added all the likes for each photo in “photos” table and then linked the user with photo id as “users” and “photos” table is interlinked with each other.
- The below user won the contest for getting most likes on a single photo.

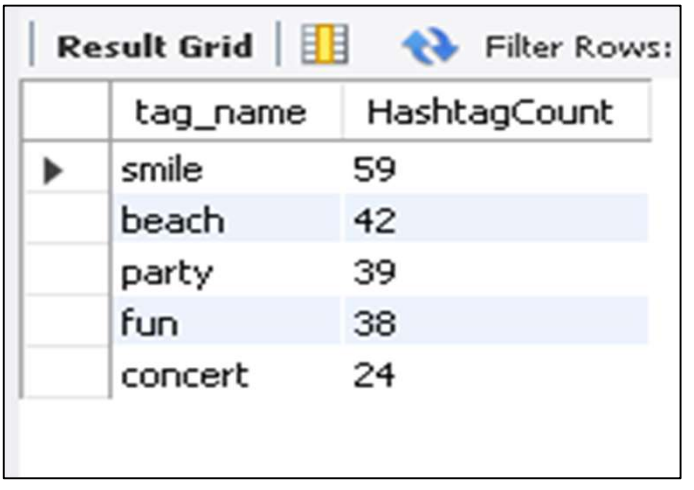
Result Grid | Filter Rows: | Export:

username	id	image_url	likes
Zack_Kemmer93	145	https://jarret.name	48

Instagram User Analytics

Hashtag Researching

- This is to identify the top 5 most used hashtags in the Instagram.
- Here, I have summed up the count of each hashtag used in all photos posted, then linked with "tags" table and used GROUP BY and ORDER BY keywords to separate each tag and to sort it as per the count.
- The following image shows the top 5 hashtags and the count of each hashtag used.



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

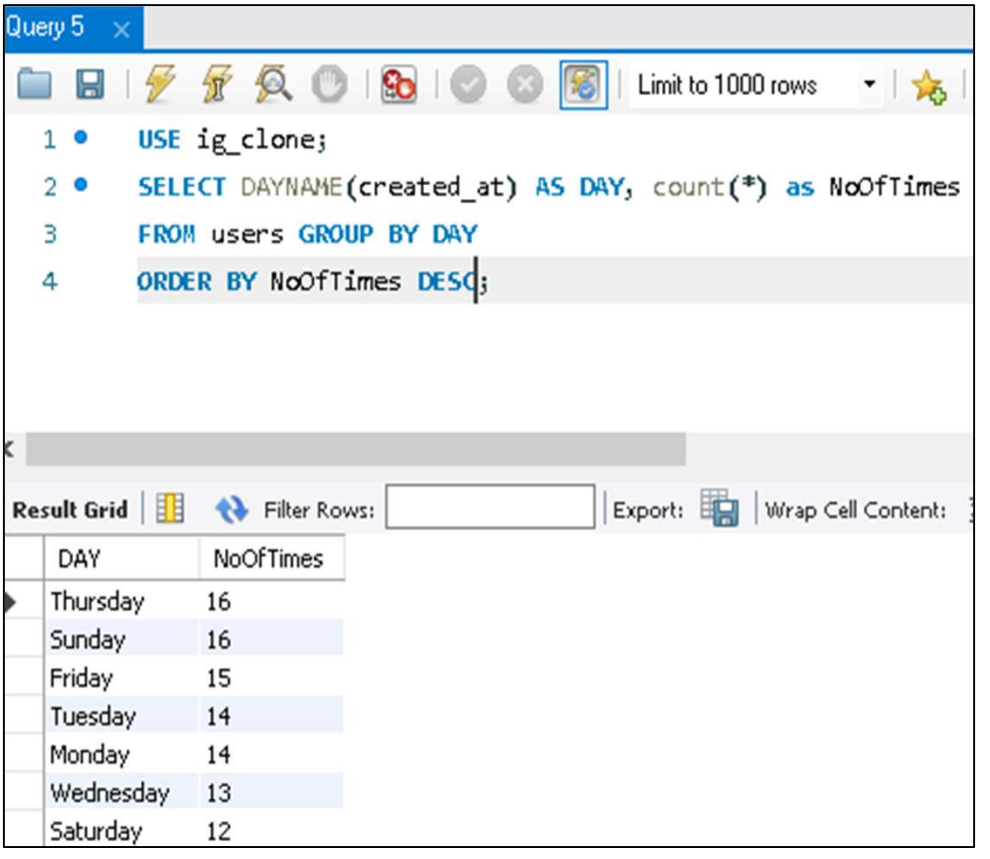
Launch AD Campaign

- This is to find the day on which most users registered on to schedule an AD campaign.
- Here, I used a built-in function called DAYNAME() to get the day from the created timestamp and grouped the total count of each day in "users" table, then sorted the top day where most registrations happened.

Instagram User Analytics

Launch AD Campaign

- When analysed this task, I got two days with matching count as showed below.
- Hence used LIMIT keyword to get the single day which was "Thursday".



Investor Metrics:

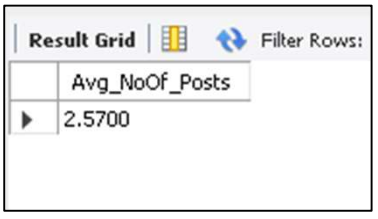
User Engagement

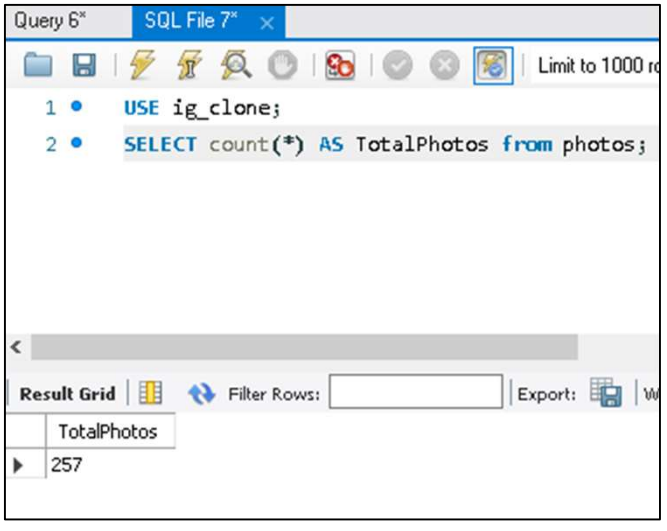
- This task is to find the average number of posts.
- I have used nested query concept from SQL documentation to select total count of photos and users and used arithmetic function to get the average.

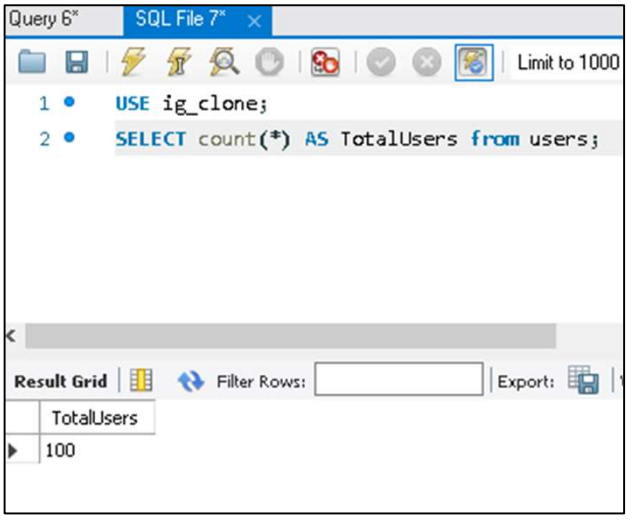
Instagram User Analytics

User Engagement

- The average was '2.57' as listed below
- The below queries showcase the total number of photos and users.







Bots & Fake Accounts

- This task is identify the bots/fake accounts by finding the users who liked every photo in the database.
- Have grouped each user with the total count of likes by linking "users" and "likes" tables.
- Then, matched it with the total count of photos and grouped users who matched the count.

Instagram User Analytics

Bots & Fake Accounts

- The listed picture shows the fake users or bots who have liked every single photo in the database

Result Grid | Filter Rows:

username	No_of_Likes
Aniya_Hackett	257
Jaclyn81	257
Rocio33	257
Maxwell.Halvorson	257
Ollie_Ledner37	257
Mckenna17	257
Duane60	257
Julien_Schmidt	257
Mike.Auer39	257
Nia_Haag	257
Leslie67	257
Janelle.Nikolaus81	257
Bethany20	257

Analysis

- Out of the 100 total users there are 26 users who are inactive and they have never posted any kind of stuff of Instagram may it be any photo, video or any type of text.
- So, the Marketing team of Instagram needs to remind such inactive users

Instagram User Analytics

Analysis:

- The user named Zack_Kemmer93 with user_id: 52 is the winner of the contest cause his photo with photo_id: 145 has the highest number of likes i.e. 48
- Top 5 most commonly used #hashtags along with the total count are smile(59), beach(42), party(39), fun(38) and concert(24)
- Most of the users registered on Thursday and Sunday i.e. 16 and it would prove beneficial to start AD Campaign on these two days
- So, there are in total 257 rows i.e. 257 photos in the photos table and 100 rows i.e. 100 ids in the users table which makes the desired output to be 257/100 = 2.57 (avg. users posts on Instagram)
- Out of the total user id's there are 13 such user id's who have liked each and every post on Instagram (which is not practically possible) and so such user id's are considered as BOTS and Fake Accounts

Conclusion:

- In conclusion, I would like to conclude that not only Instagram but other social media firms use such Analysis to find the insights from their customer data which in turn help the firms to find the customers who will be an Asset to the firm in the future and not some Liability.
- Such Analysis of customer data is done at weekly, monthly, quarterly or yearly basis as per the needs of the business firms so as to maximize their profits in future with minimal cost to the company

Operational Analytics & Investigating Metric Spike

Description:

Operation Analytics is the analysis done for the complete end to end operations of a company. With the help of this, the company then finds the areas on which it must improve upon. You work closely with the ops team, support team, marketing team, etc and help them derive insights out of the data they collect.

Being one of the most important parts of a company, this kind of analysis is further used to predict the overall growth or decline of a company's fortune. It means better automation, better understanding between cross-functional teams, and more effective workflows.

Investigating metric spike is also an important part of operation analytics as being a Data Analyst you must be able to understand or make other teams understand questions like- Why is there a dip in daily engagement? Why have sales taken a dip? Etc. Questions like these must be answered daily and for that its very important to investigate metric spike.

You are working for a company like Microsoft designated as Data Analyst Lead and is provided with different data sets, tables from which you must derive certain insights out of it and answer the questions asked by different departments.

Operational Analytics & Investigating Metric Spike

Problem:

Case 1 (Job Data):

A.Number of jobs reviewed: Amount of jobs reviewed over time.
Your task: Calculate the number of jobs reviewed per hour per day for November 2020?
B.Throughput: It is the no. of events happening per second.
Your task: Let's say the above metric is called throughput. Calculate 7 day rolling average of throughput? For throughput, do you prefer daily metric or 7-day rolling and why?
C.Percentage share of each language: Share of each language for different contents.
Your task: Calculate the percentage share of each language in the last 30 days?
D.Duplicate rows: Rows that have the same value present in them.
Your task: Let's say you see some duplicate rows in the data. How will you display duplicates from the table?

Case 2 (Investigating Metric Spike):

A.User Engagement: To measure the activeness of a user. Measuring if the user finds quality in a product/service.
Your task: Calculate the weekly user engagement?
B.User Growth: Amount of users growing over time for a product.
Your task: Calculate the user growth for product?

Operational Analytics &
Investigating Metric Spike

Case 2 (Investigating Metric Spike):

C. **Weekly Retention:** Users getting retained weekly after signing-up for a product.
Your task: Calculate the weekly retention of users-sign up cohort?

D. **Weekly Engagement:** To measure the activeness of a user. Measuring if the user finds quality in a product/service weekly.
Your task: Calculate the weekly engagement per device?

E. **Email Engagement:** Users engaging with the email service.
Your task: Calculate the email engagement metrics?

Design:

- The dataset contained a sample of 8 rows with columns as needed for analyzing the above listed queries.
- Hence, used the randbetween() and rand() functions in Microsoft Excel and generated 1000 random rows for the dataset.
- Then, loaded the dataset in My SQL and handled the queries.

Operational Analytics &
Investigating Metric Spike

Findings:

Number of Jobs Reviewed

- Here, we calculated the jobs reviewed per day per hours in the month of November.

ds	jobs_reviewed	ds	jobs_reviewed	ds	jobs_reviewed
01/11/2020	55	11/11/2020	48	21/11/2020	55
02/11/2020	46	12/11/2020	47	22/11/2020	50
03/11/2020	58	13/11/2020	62	23/11/2020	49
04/11/2020	52	14/11/2020	53	24/11/2020	61
05/11/2020	58	15/11/2020	56	25/11/2020	62
06/11/2020	61	16/11/2020	52	26/11/2020	56
07/11/2020	59	17/11/2020	50	27/11/2020	63
08/11/2020	52	18/11/2020	51	28/11/2020	60
09/11/2020	52	19/11/2020	47	29/11/2020	52
10/11/2020	65	20/11/2020	55	30/11/2020	51

Throughput

- We will be first taking the count of job_id(distinct and non-distinct) and ordering them w.r.t ds (date of interview)
- Then by using the ROW function we will be considering the rows between 6 preceding rows and the current row to get the 7-day rolling avg.

Operational Analytics &
Investigating Metric Spike

Throughput

- The below table shows the rolling avg calculated for each day in November.

datestamp	7-day_rolling_avg
01/11/2020	0.02
02/11/2020	0.01
03/11/2020	0.02
04/11/2020	0.01
05/11/2020	0.02
06/11/2020	0.02
07/11/2020	0.02
08/11/2020	0.01
09/11/2020	0.01
10/11/2020	0.02

datestamp	7-day_rolling_avg
11/11/2020	0.01
12/11/2020	0.01
13/11/2020	0.02
14/11/2020	0.01
15/11/2020	0.02
16/11/2020	0.01
17/11/2020	0.01
18/11/2020	0.01
19/11/2020	0.01
20/11/2020	0.02

datestamp	7-day_rolling_avg
21/11/2020	0.02
22/11/2020	0.01
23/11/2020	0.01
24/11/2020	0.02
25/11/2020	0.02
26/11/2020	0.02
27/11/2020	0.02
28/11/2020	0.02
29/11/2020	0.01
30/11/2020	0.01

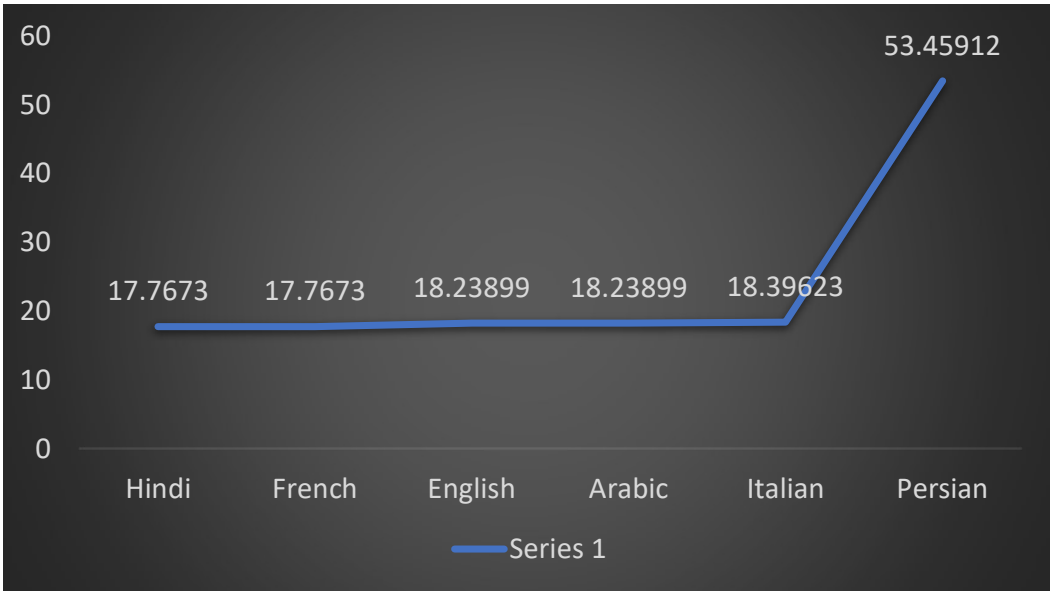
% Share of each Language

- We will first divide the total number of languages (distinct/non-distinct) by the total number of rows presents in the table
- Then we will do the grouping based on the languages.

Operational Analytics & Investigating Metric Spike

% Share of each Language

- Here, we have the percentage share of each language in the dataset based on the unique job ids



Duplicate Rows

- There are 365 duplicate rows from the dataset and I have displayed few of them here from the output

ds	job _id	actor _id	event	language	time_ spent	org	rownum
11-11-2020	12	1279	transfer	Persian	63	D	2
24-11-2020	17	1300	skip	English	87	D	2
09-11-2020	20	1917	transfer	Hindi	71	D	2
28-11-2020	21	1151	transfer	Italian	120	B	2
06-11-2020	21	1982	skip	Persian	21	B	3
15-11-2020	22	1845	decision	English	53	B	2

Operational Analytics &
Investigating Metric Spike

User Engagement

- We will extract the week from the occurred_at column of the events table using the EXTRACT function and WEEK function
- Then we will be counting the number of distinct user_id from the events table
- Now, we will use the GROUP BY function to group the output w.r.t week from occurred_at

week_num	active_users
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

Operational Analytics &
Investigating Metric Spike

User Growth

- Displayed the number of users created in each month and used mathematical functions to round it as percentage for user increase/decrease comparing previous month respectively.

month_num	No_of_users	Growth in %	month_num	No_of_users	Growth in %
1	712	NULL	7	1281	17.96
2	685	-3.79	8	1347	5.15
3	765	11.68	9	330	-75.5
4	907	18.56	10	390	18.18
5	993	9.48	11	399	2.31
6	1086	9.37	12	486	21.8

Weekly Retention

- Here, I have summed up the users and displayed the weeks for all months based on their event type "engagement" after completing signing in.
- The below table shows the result of users in each week.

users	week_1	week_2	week_3	week_4	week_5
615	114	73	49	37	26

Operational Analytics &
Investigating Metric Spike

Weekly Engagement

- Firstly we will extract the year and week from the occurred_at column of the events table using the extract, year and week function.
- Then we will select those rows where event_type = 'engagement' using the WHERE clause.
- Now, using Group By and Order By function we will group and order the result on the basis of year, week and device

year	week	device	num_users	year	week	device	num_users
2014	17	dell inspiron notebook	46	2014	17	nexus 10	16
2014	17	hp pavilion desktop	14	2014	17	nexus 5	40
2014	17	htc one	16	2014	17	nexus 7	18
2014	17	ipad air	27	2014	17	nokia lumia 635	17
2014	17	ipad mini	19	2014	17	samsumg galaxy tablet	8
2014	17	iphone 4s	21	2014	17	samsung galaxy note	7
2014	17	iphone 5	65	2014	17	samsung galaxy s4	52
2014	17	iphone 5s	42	2014	17	windows surface	10

Operational Analytics &
Investigating Metric Spike

Email Engagement

- The below table shows the email engagement metrics of users with respect to multiple events listed in “email_events” dataset.

week	weekly_digest_mails	reengagement_emails	email_opens	email_clickthroughs
17	908	73	310	166
18	2602	157	912	430
19	2665	173	972	477
20	2733	191	1004	507
21	2822	164	1014	443
22	2911	192	987	488
23	3003	197	1075	538
24	3105	226	1155	554
25	3207	196	1096	530
26	3302	219	1165	556
27	3399	213	1228	621
28	3499	213	1250	599
29	3592	213	1219	590
30	3706	231	1383	630
31	3793	222	1351	445
32	3897	200	1337	418
33	4012	264	1432	490
34	4111	261	1528	490
35	0	48	41	38

Operational Analytics & Investigating Metric Spike

Analysis:

Why is the weekly user engagement so less in the beginning and then got increased?

- It is a fact that for any new product or service launched, during it's initial phase it is less known and only some people use the product and based on their experience the product/service engagement increases or decreases depending on whether the consumer experience was good or bad.

Why is weekly retention so important?

- Weekly retention helps the firms to convince and help those visitors who just complete the sign-up or leave the sign-up process in between, such visitors may become customers in future if they are guided and convinced properly

Why weekly engagement per device plays an important role

- Based on the reviews from users weekly engagement per device helps the firms on which devices they must focus more and which devices need more improvements

Why Email Engagement plays an important role?

- Email Engagement helps the firms to decide the discounts and offers on specific products.

Conclusion:

- I would like to conclude that Operation Analytics and Investigating Metric Spike are very necessary and they must be done on daily, weekly, Monthly, Quarterly or Yearly basis based on the Business needs of the firm.
- Also, any firm/entity must focus on the Email Engagement with the customers; the firm must use catchy headings along with reasonable discounts and coupons so as to increase their existing customer base.

Hiring Process Analytics

Description:

Hiring process is the fundamental and the most important function of a company. Here, the MNCs get to know about the major underlying trends about the hiring process. Trends such as- number of rejections, number of interviews, types of jobs, vacancies etc. are important for a company to analyze before hiring freshers or any other individual. Thus, making an opportunity for a Data Analyst job here too!

Being a Data Analyst, your job is to go through these trends and draw insights out of it for hiring department to work upon.

You are working for a MNC such as Google as a lead Data Analyst and the company has provided with the data records of their previous hirings and have asked you to answer certain questions making sense out of that data.

Problem:

Below are project questions:

How many males and females are hired?
What is the average salary offered in the company?
Draw the class intervals for salary in the company?
Draw Pie Chart to show the proportion of people working in company?
Represent different post tiers through chart/graph?

Hiring Process Analytics

Design:

- I have used Microsoft Excel for analyzing, cleaning and plotting for queries listed.
- Removed the duplicates in the dataset.
- I looked for blank spaces and NULL values if any for cleaning the given dataset.
- Then I imputed the numerical blanks and NULL cells with mean of the column(if no outliers existed for that particular column) and with median (if outliers existed for that column)
- Later, replaced outliers with the median of the particular column
- Then for blank cells of categorical variables, I had replaced with the variable with the highest count

Findings:

Hiring

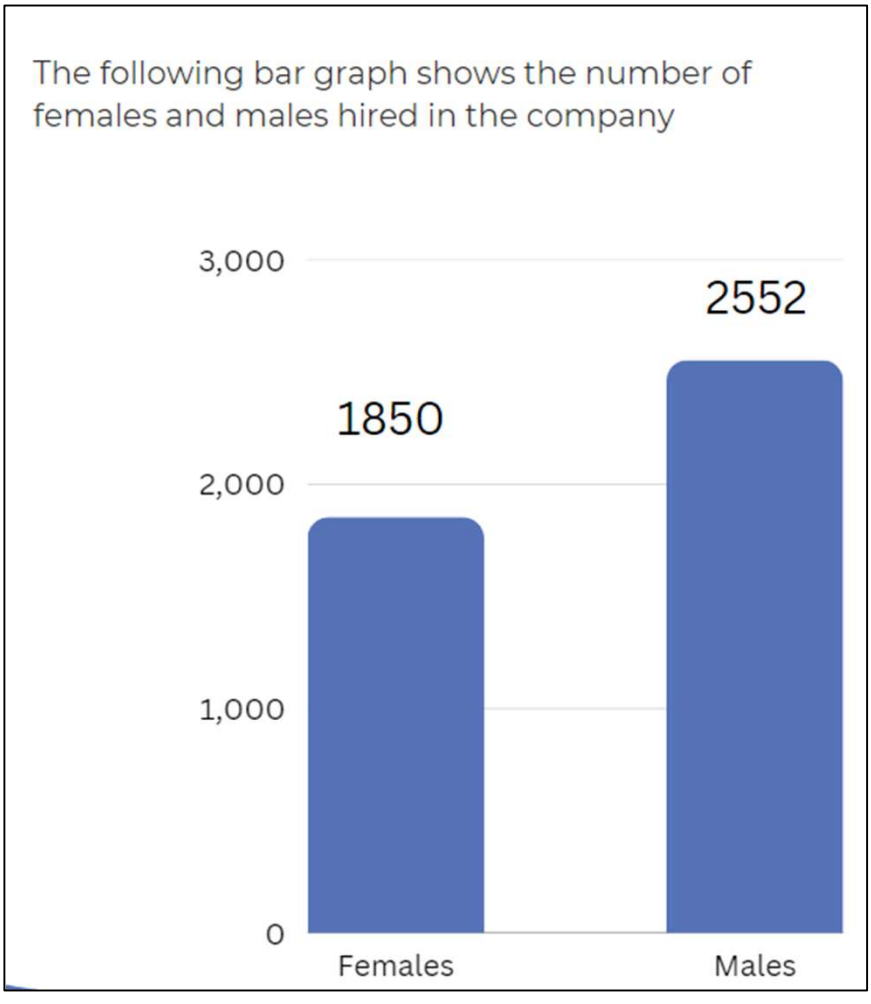
- Based on the pivot table result, male employees hired were 2,552 and Female were 1,850.

Count of Status	Column Labels		
Row Labels	Hired	Rejected	Grand Total
Don't want to say	267	125	392
Female	1850	814	2664
Male	2552	1518	4070
(blank)	10	5	15
Grand Total	**4679**	**2462**	**7141**

Event Name	Hired
Female	1850
Male	2552

Hiring Process Analytics

Hiring



Average Salary

Average Salary Offered

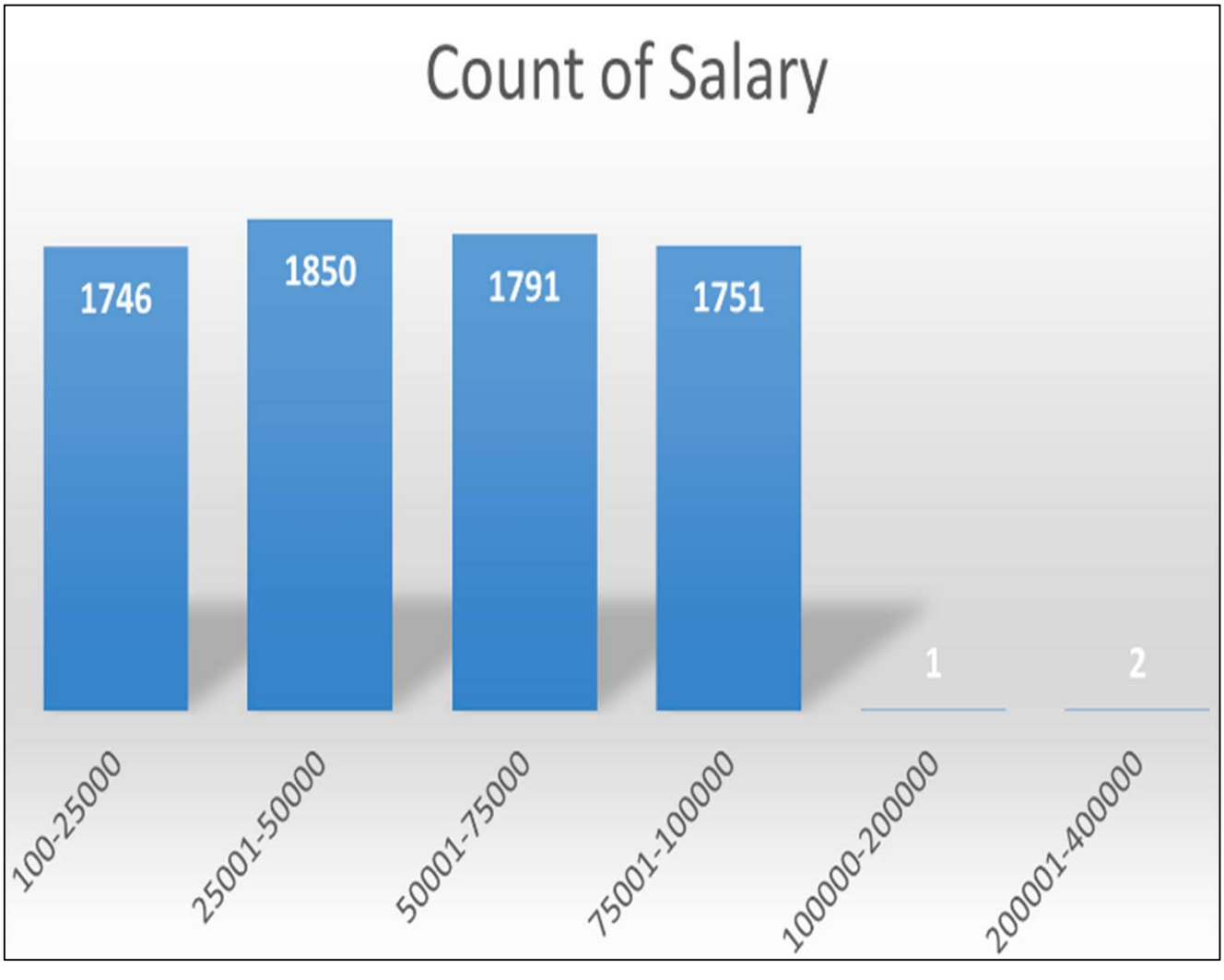
There was one blank record in the salary column which was added with the average of all data points and following table shows the average salary offered in the company.

Average Salary
50009.956

Hiring Process Analytics

Class Interval

- The following graph shows the count of salary in below class intervals.
- Most of the people are offered the salary ranging between 100-100000 rupees where highest count is in 25001-50000 interval.

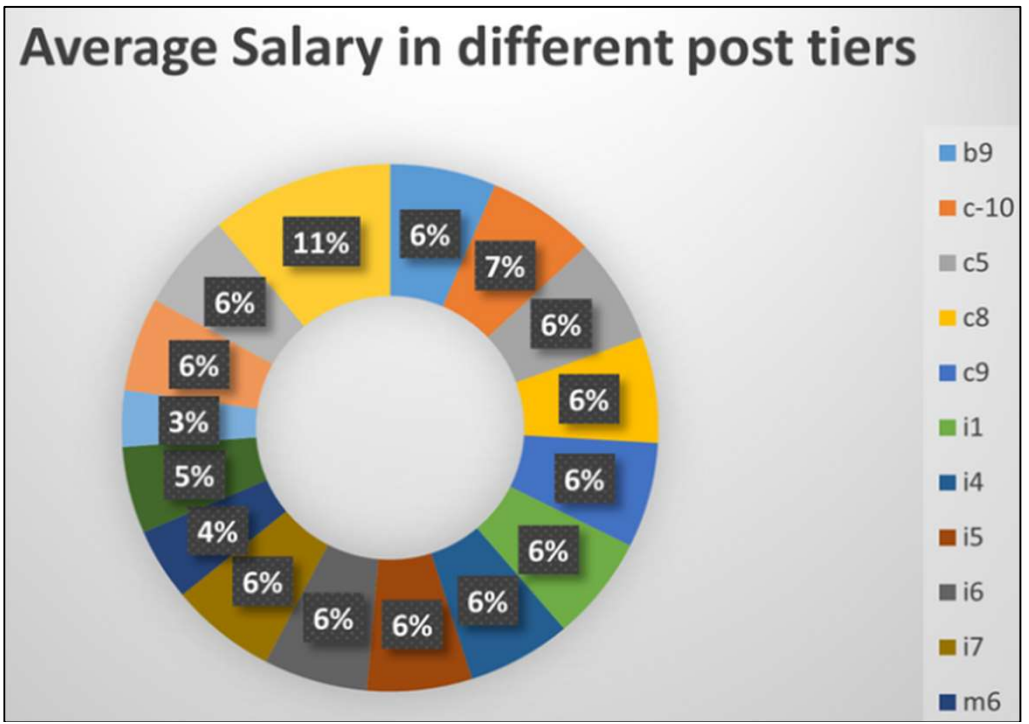


Hiring Process Analytics

Class Interval

- The following table and graph shows the average salary breakdown for different post tiers.
- The highest average which is 85914 was for one blank record in post tier column.

Post	Avg. Salary	Post	Avg. Salary
b9	49906.241	i6	48839.249
c-10	51134.621	i7	50106.005
c5	50264.477	m6	34521.333
c8	50627.266	m7	41402
c9	50209.250	n10	26990
i1	49943.937	n6	44700
i4	48877.841	n9	46219
i5	49347.296	blank	85914

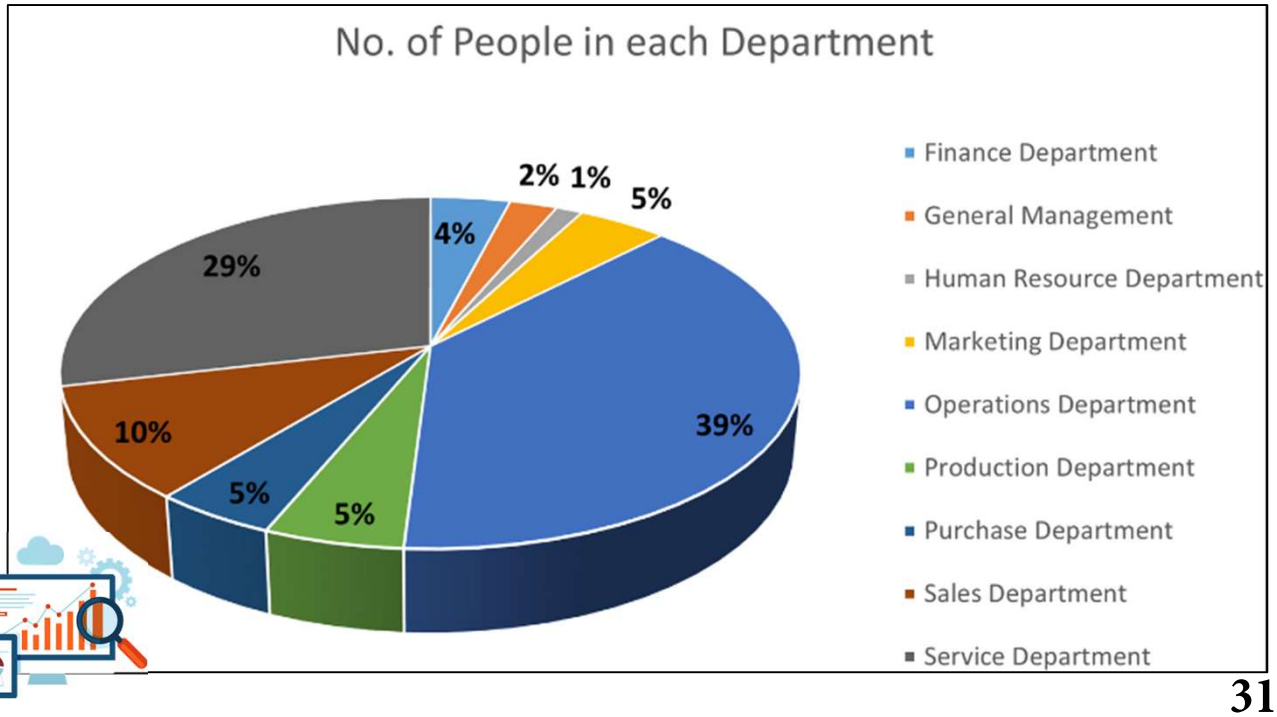


Hiring Process Analytics

Department Wise Pie chart

- The following table shows the proportion of people in each department.
- Most of them are in Operations department and less number of people are working in Human Resources.

Department	Count
Finance Department	287
General Management	171
Human Resource Department	96
Marketing Department	325
Operations Department	2762
Production Department	379
Purchase Department	332
Sales Department	744
Service Department	2045

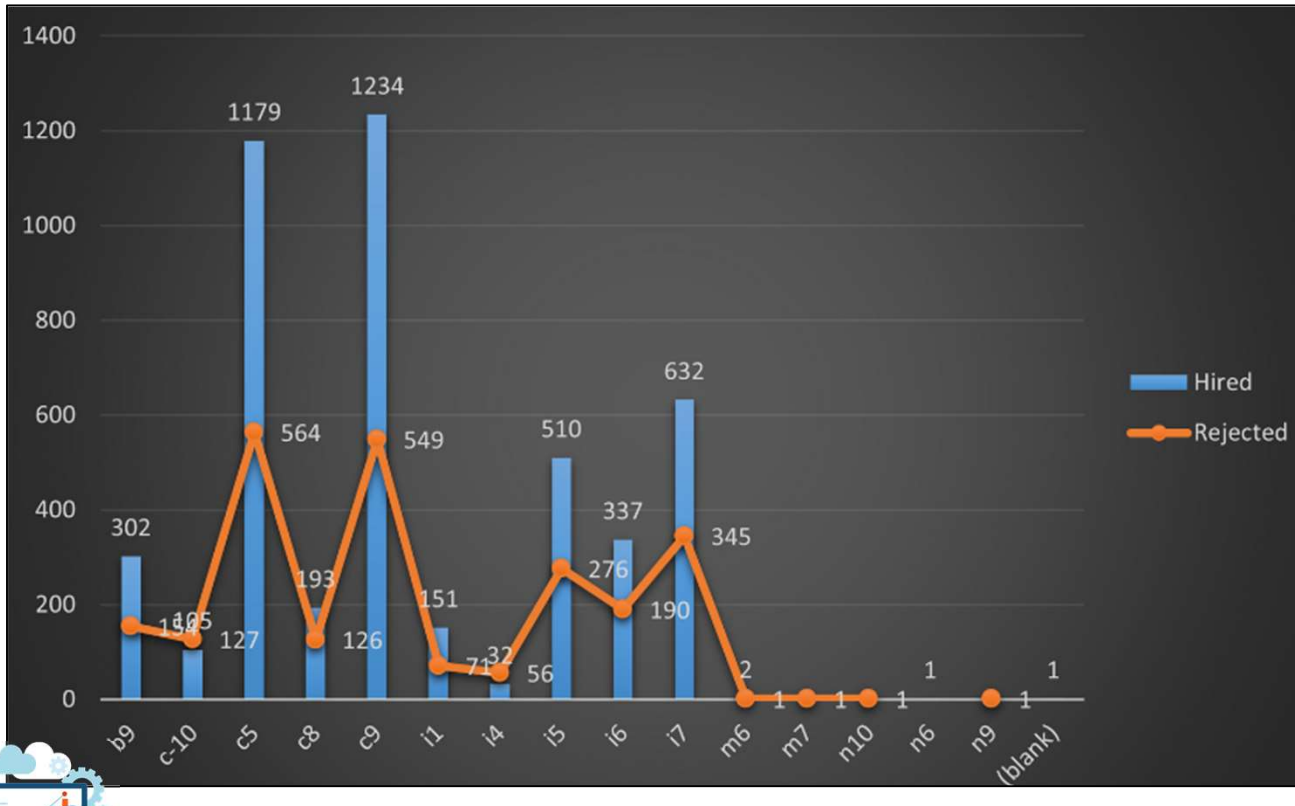


Hiring Process Analytics

Post Tiers Graphs

- The following table and graph shows hired and rejected count for different post tiers.

Post	Hired	Rejected	Post	Hired	Rejected
b9	302	154	i6	337	190
c-10	105	127	i7	632	345
c5	1179	564	m6	2	1
c8	193	126	m7		1
c9	1234	549	n10		1
i1	151	71	n6	1	
i4	32	56	n9		1
i5	510	276	(blank)	1	



Hiring Process Analytics

Analysis:

- When checked the post tier data, c9 and c5 are most sought out jobs due to the highest count and c-10 and c8 offered the highest salary.
- Operations Department has highest number of people as it works like a central hub for all other departments, all the execution tasks are carried out by this department.
- Operations department has the highest work load when compared to all other departments

Conclusion:

- I would like to conclude that Hiring Process Analytics plays an important part for all the companies and firms to decide the job openings for the near future.
- Hiring Process Analytics is done on monthly, quarterly or yearly basis as per the needs and policies of the organizations.
- For any company, there will be employees who have high salary packages compared to other employees, and this is due to the fact that they have extra skillset and years of experience in their particular field of work.
- Hiring Process Analytics helps the company
    - to decide the salaries for new freshers joining the company
    - also it tells requirement of workforce by each department
    - it also helps the company decide the appraisals for the current employees

IMDb Movie Analysis

Description:

In this project, we are provided with dataset having various columns of different IMDB Movies.

You are required to Frame the problem.
For this task, you will need to define a problem you want to shed some light on.

We can do this by asking 'What?' This is where you frame the problem i.e. What is the problem?

Once you have defined a problem, clean the data as necessary, and use your Data Analysis skills to explore the data set and derive insights.

Problem:

A.Cleaning the data:: This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)
Your task: Clean the data
B.Movies with highest profit: Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.
Your task: Find the movies with the highest profit?

IMDb Movie Analysis

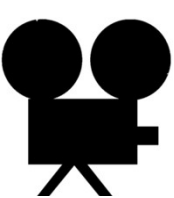
Problem:

C. **Top 250:** Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.
Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!
Your task: Find IMDB Top 250

D. **Best Directors:** Group the column using the director_name column.Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.
Your task: Find the best directors

E. **Popular Genres:** Perform this step using the knowledge gained while performing previous steps.
Your task: Find popular genres

F. **Charts:**
Group the combined column using the actor_1_name column. Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.
Your task: Find the critic-favorite and audience-favorite actors



IMDb Movie Analysis

Design:

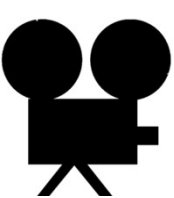
- For this Project, I have used the MS Excel for cleaning, analyzing the dataset and plotting charts and graphs.
- I have used Data Analytic steps while answering queries and used Pivot tables, formulae needed to consolidate and extract data etc..

Findings:

Cleaning the Data

- First and foremost thing I have checked in the dataset is for duplicate movies.
- Based on the unique movies, there were 127 duplicates in 5043 records, after cleaning, 4916 unique values are present.
- Later, for each query I have cleaned and removed unnecessary data for getting the desired output.

Original	5043
Unique	4916
Duplicate records	127



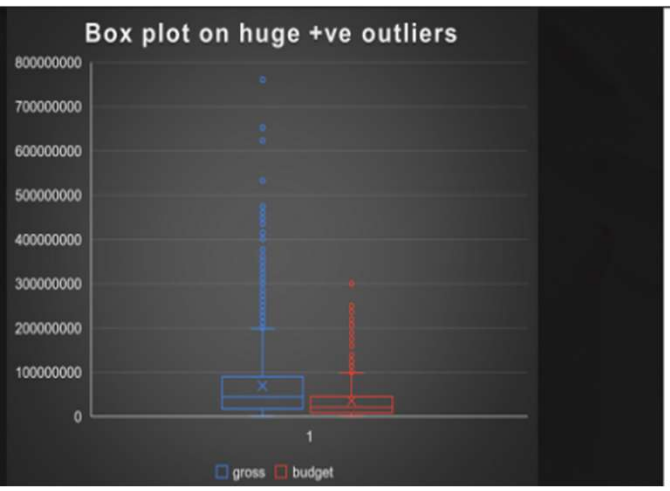
IMDb Movie Analysis

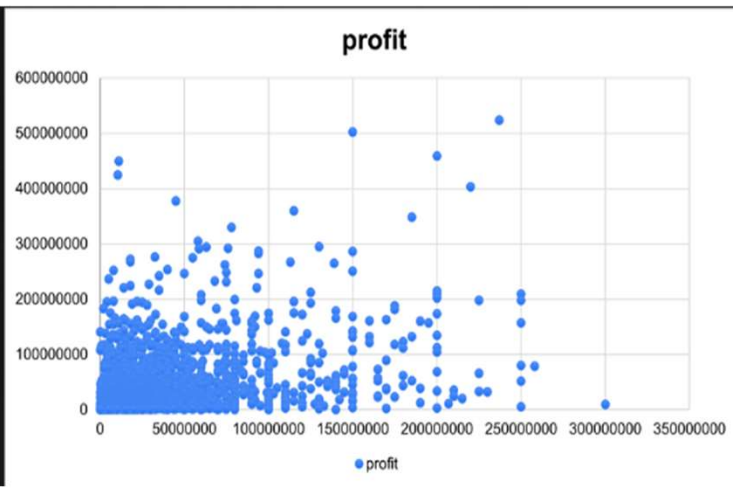
Movies with Highest Profit

- Here, the Inter Quartile Range is used to identify the outliers and also I have removed the negative values to plot.
- I have used the quartile functions to know the levels and understood the class intervals for profit data.
- The movie "**Avatar**" has the highest profit from the dataset which is "**523505847**"

Movie with highest Profit	Value
Avatar	523505847

gross	budget	profit	movie_title	Outliers
760505847	237000000	523505847	Avatar	TRUE
652177271	150000000	502177271	Jurassic World	TRUE
658672302	200000000	458672302	Titanic	TRUE
460935665	11000000	449935665	Star Wars: Episode IV - A New Hope	TRUE
434949459	10500000	424449459	E.T. the Extra-Terrestrial	TRUE





IMDb Movie Analysis

Top 250

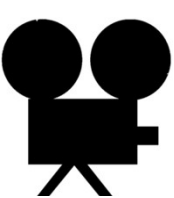
- For this query, I have extracted the movies with language and no. of voted users data from cleaned data.
- Then, separated the English movies and all other languages where 12 blank records under language column were deleted.
- Later, I sorted the data with highest IMDB score and num_voted_users and ranked them to get the required output.
- The below image shows the Top 10 Movies in English and Foreign Language category for above mentioned conditions.

IMDB_Top_250	num_voted_users	language	imdb_score	Rank
The Shawshank Redemption	1689764	English	9.3	1
The Godfather	1155770	English	9.2	2
The Dark Knight	1676169	English	9	3
The Godfather: Part II	790926	English	9	3
Fargo	170055	English	9	3
Pulp Fiction	1324680	English	8.9	6
The Lord of the Rings: The Return of the King	1215718	English	8.9	6
Schindler's List	865020	English	8.9	6
12 Angry Men	447785	English	8.9	6
Inception	1468200	English	8.8	10

English

IMDB_Top_250	num_voted_users	language	imdb_score	Rank
The Good, the Bad and the Ugly	503509	Italian	8.9	1
City of God	533200	Portuguese	8.7	2
Seven Samurai	229012	Japanese	8.7	2
Spirited Away	417971	Japanese	8.6	4
The Lives of Others	259379	German	8.5	5
Airlift	30977	Hindi	8.5	5
Children of Heaven	27882	Persian	8.5	5
Amélie	534262	French	8.4	8
Oldboy	356181	Korean	8.4	8
Princess Mononoke	221552	Japanese	8.4	8

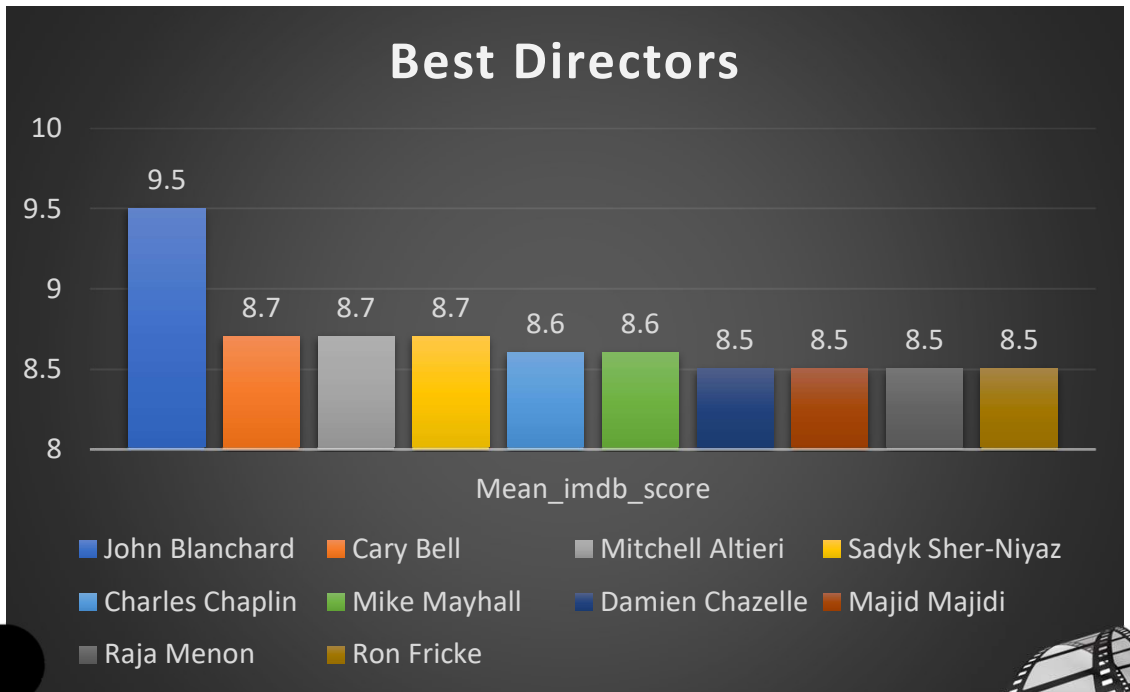
Foreign Languages



Best Directors

- For this query, I have extracted related variables from the cleaned data and 102 blank records in director_name column is removed.
- Here, I have consolidated the data in director_name and imdb_score columns with averaging function in order to group director with their mean_imdb_scores, then sorted the values with two conditions

Top10 Director	Mean_imdb_score
John Blanchard	9.5
Cary Bell	8.7
Mitchell Altieri	8.7
Sadyk Sher-Niyaz	8.7
Charles Chaplin	8.6
Mike Mayhall	8.6
Damien Chazelle	8.5
Majid Majidi	8.5
Raja Menon	8.5
Ron Fricke	8.5



IMDb Movie Analysis

Popular Genres

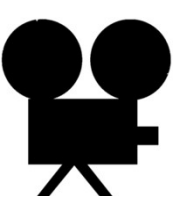
- For this query, I have consolidated all the genres into a group and filtered each group with three conditions:(Maximum profit for the genre, Average of its imdb_score and average of the users voted for the genre).
- Then, I have sorted the consolidated result table using the above three conditions(max profits, highest mean of imdb_score, Highest voted users).
- Top 10 genres are placed below as output.

genres	num_voted_users	imdb_score	Profit
Action|Adventure|Fantasy|Sci-Fi	451464	6.985	523505847
Action|Adventure|Sci-Fi|Thriller	203959.6765	6.406	502177271
Drama|Romance	66329.76667	6.933	458672302
Family|Sci-Fi	141025.5	5.650	424449459
Action|Adventure|Sci-Fi	300716.7917	6.623	403279547
Adventure|Animation|Drama|Family|Musical	644348	8.500	377783777
Action|Crime|Drama|Thriller	86263.56923	6.465	348316061
Adventure|Drama|Sci-Fi|Thriller	361775	7.350	329999255
Action|Adventure|Comedy|Romance|Sci-Fi	173494.3333	6.133	305024263
Adventure|Sci-Fi|Thriller	308063.2857	6.871	294645577

Combined data

- Here, as listed in query, the below table is the combined movies output of three actors from actor_1 column.

Combined
It's Complicated Titanic The Curious Case of Benjamin Button
The River Wild The Great Gatsby Troy
Julie & Julia Inception Ocean's Twelve
The Devil Wears Prada The Revenant Mr. & Mrs. Smith
Lions for Lambs The Aviator Spy Game
Out of Africa Django Unchained Ocean's Eleven
Hope Springs Blood Diamond Fury
One True Thing The Wolf of Wall Street Seven Years in Tibet
Florence Foster Jenkins Gangs of New York Fight Club
The Hours The Departed Sinbad: Legend of the Seven Seas
The Iron Lady Shutter Island Interview with the Vampire: The Vampire Chronicles
A Prairie Home Companion Body of Lies The Tree of Life
Julia Catch Me If You Can The Assassination of Jesse James by the Coward Robert Ford
The Beach Babel
Revolutionary Road By the Sea
The Man in the Iron Mask Killing Them Softly
J. Edgar True Romance
The Quick and the Dead Johnny Suede
Marvin's Room
Romeo + Juliet



IMDb Movie Analysis

Critic Favorite & Audience Favorite Actors

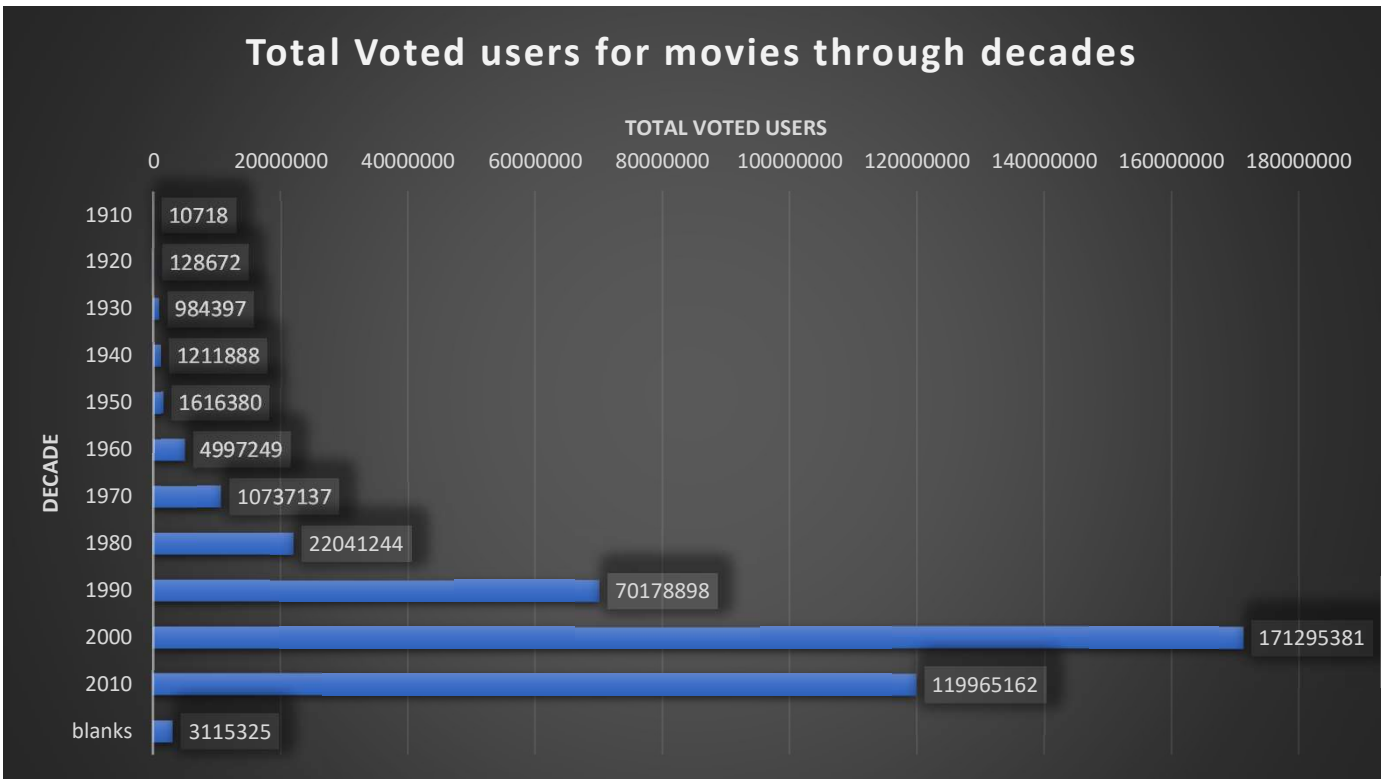
- For this query, I have consolidated the actor_1 data with mean of critic_reviews and num_user_reviews separately and found the highest mean in both cases and displayed the output in the below.

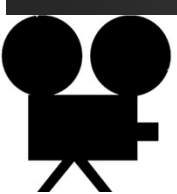
critic_favorite_actor	mean_critic_reviews
Phaldut Sharma	738

audience_favorite_actor	mean_user_reviews
Heather Donahue	3400

Movies with Voted users through each decade

- For this query, I have extracted the movies along with their release year and sorted using year column and grouped the movies by decades based on the year and totaled the number of voted users and plotted a bar chart which is showcased below.





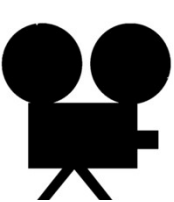
IMDb Movie Analysis

Analysis:

- I have observed while answering critic and audience favorite actors that some of them have been casted into more movies due to which mean score got affected when compared with the actors who have only one record with high number.
- The dataset contained many columns which played critical role for each query answered and some columns were unused, ex: no. of Facebook likes , plot keywords where interesting questions can be framed using these as well.
- Action|Adventure|Fantasy|Sci-Fi mixed genre has the highest popularity based on given dataset.

Conclusion:

- I would like to conclude that IMDB Movie Analysis or any such analysis is done not only by Movie makers before movie production, but it is also done by various investors, stakeholders, theatre outlet owners.
- Such analysis plays an crucial part during the pre-production phase of the movies and also during the post-production phase
- Also, it is not necessary that the movie with the highest IMDB rating will have the highest profit.
- Profit is calculated truly on the basis on the number of tickets sold by theatres all over the world.
    - So, directors and production team must keep these kind of factors in mind and shall do the pre-production analysis before the commencement of filming



Bank Loan Case Study

Description:

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.

Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Problem:

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Bank Loan Case Study

Problem:

•Present the overall approach of the **analysis**. Mention the problem statement and the analysis approach briefly

•**Identify** the missing data and use appropriate method to deal with it. Identify if there are **outliers** in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.

•Identify if there is data imbalance in the data. Find the ratio of data imbalance.

•Explain the **results of univariate, segmented univariate, bivariate analysis, etc.** in business terms.

•Find the top 10 **correlation** for the Client with payment difficulties and all other cases (Target variable). Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.

•**Include visualizations** and **summarize** the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases

Bank Loan Case Study

Design:

- I have used Python(with Pandas, Numpy, Matplotlib & Seaborn) in Jupyter Notebook for Handling, analysing and cleaning data along with required Visualizations.

- I have used and implemented the EDA steps .

- We need to handle two datasets given:
  - application_data
  - Previous_application_data

- These two datasets has the information on clients loan data and current data by which we have to determine whether loan should be given or not with keeping the risks in mind.

- The steps used for analysing the dataset was:
  - Reading and getting an overall idea about the dataset
  - Data cleaning(Handling the Nan Values)
  - Missing Value Imputation – appending the missing values for columns as needed.
  - Finding Outliers – Identifying the abnormal values in the listed columns
  - Data Analysis – Performing the three analysis steps as needed to find insights
- Drawing the conclusions

Bank Loan Case Study

Findings:

Reading & Understanding Data

- After, loading the datasets using Pandas read, we had a general understanding on the shape of data, information on the columns and their datatypes and descriptive statistics.

```
#checking the information of the dataset

app_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

- **We have 122 columns and 307511 rows in the application.csv dataset with multiple data types**

Application Dataset

- Application dataset has 122 columns with 307511 rows whereas prev_application dataset has 37 columns with 1670214 rows.

```
prev_app.shape
```

```
(1670214, 37)
```

- **There are 37 columns having various data types and 1670214 rows.**

```
prev_app.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1670214 entries, 0 to 1670213
Data columns (total 37 columns):
```

Previous Application Dataset

Bank Loan Case Study

Data Cleaning & Manipulation

- We have identified the missing value percentages in the datasets and dropped the columns with high percentage of null values.
- For application_data , I have analysed the flag columns and two unnecessary columns through their correlation with the 'TARGET" variable and dropped them as well.
- Same process was done with previous application dataset, where columns with high null value percentage and unnecessary columns were dropped.

2.1.1 Dealing with Null Values more than 45%

```
# created a new df to store the variables with high null values and dropped them

missing_col = missing_data[missing_data['missing_pcnt'] >= 45]['col_name'].to_list()
app_missing_del = app_data.drop(labels = missing_col,axis=1)
app_missing_del.shape
```

```
(307511, 73)
```

- **Removed 49 columns which had high null values percentage**

- After removing unnecessary columns the columns were reduced to 43 in application_data.

2.2 Checking EXT_SOURCE_3 , EXT_SOURCE_2 as they have normalised values

```
round(app_flag_del[['EXT_SOURCE_2','EXT_SOURCE_3','TARGET']].corr(),2)
```

	EXT_SOURCE_2	EXT_SOURCE_3	TARGET
EXT_SOURCE_2	1.00	0.11	-0.16
EXT_SOURCE_3	0.11	1.00	-0.18
TARGET	-0.16	-0.18	1.00

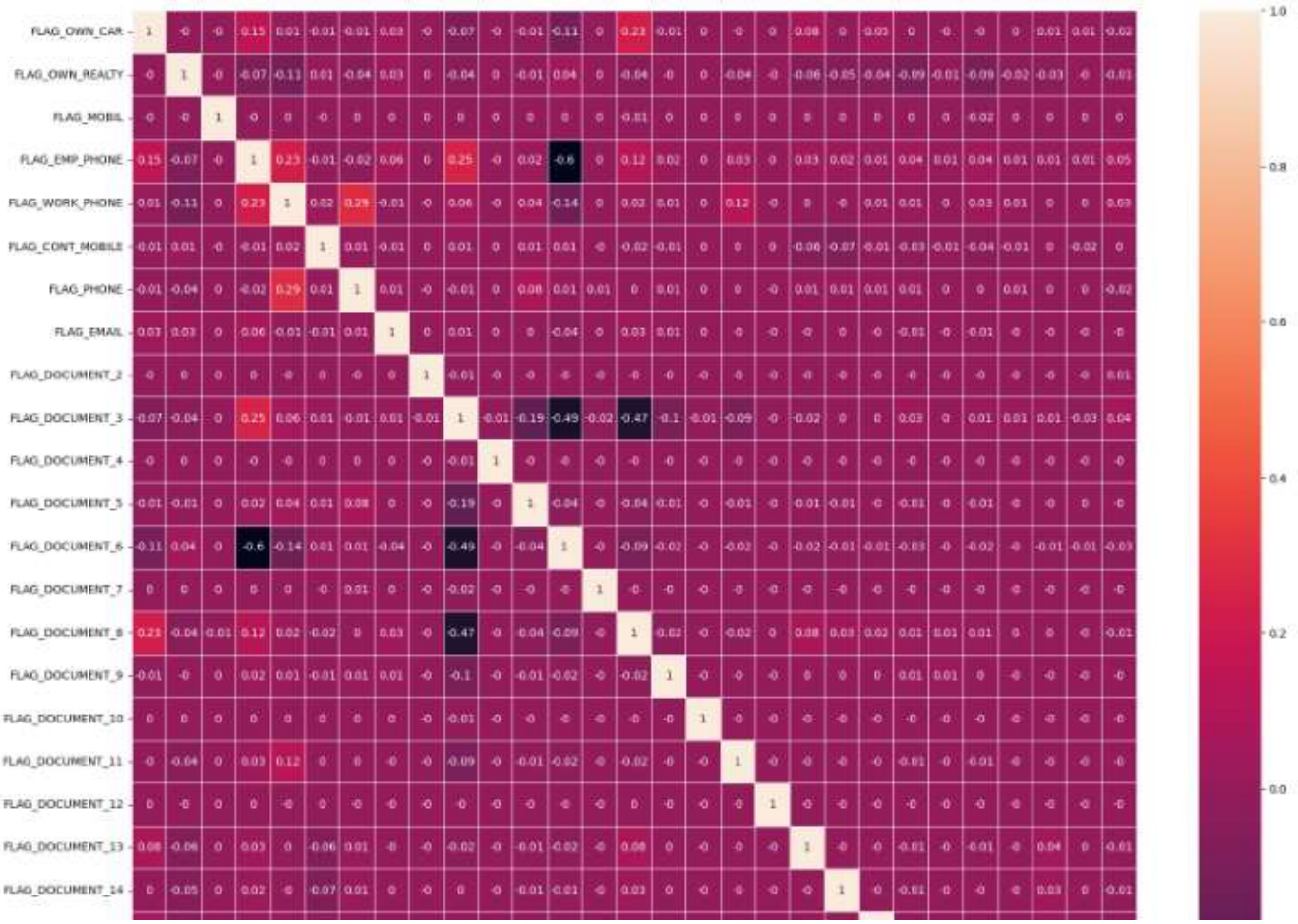
```
app_clean_data = app_flag_del.drop(['EXT_SOURCE_2','EXT_SOURCE_3'],axis=1)
app_clean_data.shape
```

```
(307511, 43)
```

- There seems to be no linear correlation
- After Data Cleaning and dropping unnecessary columns we are left with 43 columns

Bank Loan Case Study

- This image is the correlation matrix plotted with FLAG columns and TARGET variable.



- These FLAG columns or variables are also listed in the removed 43 columns from the previous slide .
- As no variable is correlated with the TARGET variable.

Bank Loan Case Study

Missing Value Imputation

- Here, we have used the three imputations{mean, median, mode} as needed for relevant columns to fill the missing values as needed for analysis.
- This was done for both application and previous application datasets
- Here, in the application_data, I have filled the values with Unknown for OCCUPATION_TYPE, as the percentage of NaN was high to use the imputation technique
- Later, for the rest six columns used the median imputation based on the descriptive statistics .

```
# imputing null values with "Unknown"

app_clean_data['OCCUPATION_TYPE'] = app_clean_data['OCCUPATION_TYPE'].fillna('Unknown')
```

- As we have more than 30% of null values for this categorical column, we filled with Unknown or NA values or we could have filled with most repeated value "Laborers" as well but the graphs would have been too deviated towards this value.

```
amt_req_credit =['AMT_REQ_CREDIT_BUREAU_HOUR',
 'AMT_REQ_CREDIT_BUREAU_DAY',
 'AMT_REQ_CREDIT_BUREAU_WEEK',
 'AMT_REQ_CREDIT_BUREAU_MON',
 'AMT_REQ_CREDIT_BUREAU_QRT',
 'AMT_REQ_CREDIT_BUREAU_YEAR']
```

```
#filling missing values with median values

app_clean_data.fillna(app_clean_data[amt_req_credit].median(), inplace = True)
```

```
nan_val(app_clean_data).head()
```

```
NAME_TYPE_SUITE             0.42
DEF_60_CNT_SOCIAL_CIRCLE    0.33
OBS_60_CNT_SOCIAL_CIRCLE    0.33
DEF_30_CNT_SOCIAL_CIRCLE    0.33
OBS_30_CNT_SOCIAL_CIRCLE    0.33
dtype: float64
```

- Still there some null value present columns but imputing them is not needed as the % is very low.

Bank Loan Case Study

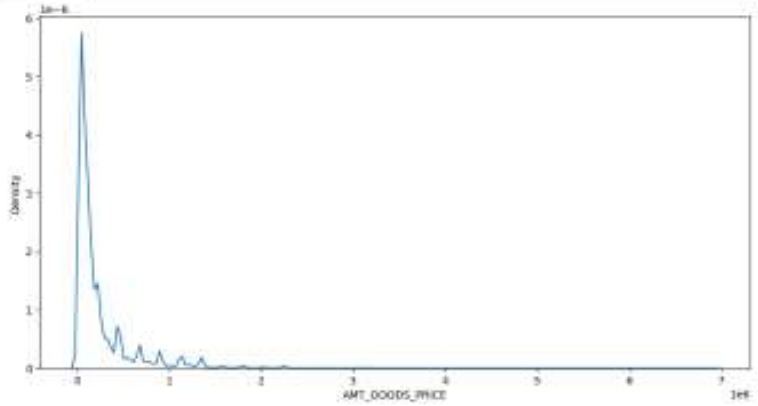
Missing Value Imputation

- In the previous_application_data, I have implemented median and mode imputation as needed

- To impute null values in continuous variables, I plotted distribution of columns

```
# Plotting kde plot for "AMT_GOODS_PRICE" to understand the distribution

plt.figure(figsize=(12,6))
sns.kdeplot(prev_app['AMT_GOODS_PRICE'])
plt.show()
```

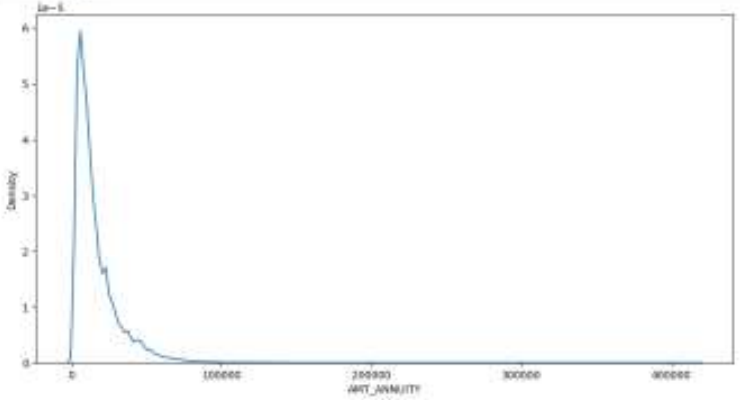


- There are several peaks along the distribution. Imputing the mode as checked

```
In [82]:
prev_app['AMT_GOODS_PRICE'].fillna(prev_app['AMT_GOODS_PRICE'].mode()[0], inplace=True)
```

```
#plotting a kdeplot to understand distribution of "AMT_ANNUITY"

plt.figure(figsize=(12,6))
sns.kdeplot(prev_app['AMT_ANNUITY'])
plt.show()
```



Insight:

- There is a single peak at the left side of the distribution and it indicates the presence of outliers and hence imputing with median.

```
In [80]:
prev_app['AMT_ANNUITY'].fillna(prev_app['AMT_ANNUITY'].median(),inplace = True)
```

- **Median** if the distribution is skewed.
- **Mode** if the distribution pattern is preserved.

Bank Loan Case Study

Standardizing Values

- There were some columns with high range of values and for this binning values was needed.
- Here, in the application_data, I have handled five columns by creating bins as needed and the DAYS columns had some negative values which was found in descriptive stats which was changed as well.

	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	DAYS_LAST_PHONE_CHANGE
count	307511.000000	307511.000000	307511.000000	307511.000000	307510.000000
mean	-16036.995067	63815.045904	-4986.120328	-2994.202373	-962.858788
std	4363.988632	141275.766519	3522.886321	1509.450419	826.808487
min	-25229.000000	-17912.000000	-24672.000000	-7197.000000	-4292.000000
25%	-19682.000000	-2760.000000	-7479.500000	-4299.000000	-1570.000000
50%	-15750.000000	-1213.000000	-4504.000000	-3254.000000	-757.000000
75%	-12413.000000	-289.000000	-2010.000000	-1720.000000	-274.000000
max	-7489.000000	365243.000000	0.000000	0.000000	0.000000

```
app_clean_data[days_list] = abs(app_clean_data[days_list]) #using abs() function to correct the days values
```

```
app_clean_data[days_list].describe()
```

4.1 Handling columns: AMT_INCOME_TOTAL, AMT_CREDIT, AMT_GOODS_PRICE

```
# Binning Numerical Columns to create a categorical column

# Creating bins for income amount in term of Lakhs

app_clean_data['AMT_INCOME_TOTAL'] = app_clean_data['AMT_INCOME_TOTAL']/100000

bins = [0,1,2,3,4,5,6,7,8,9,10,11]
slot = ['0k-100k','100k-200k','200k-300k','300k-400k','400k-500k','500k-600k','600k-700k','700k-800k','800k-900k','900k-1M','1M-

app_clean_data['AMT_INCOME_RANGE'] = pd.cut(app_clean_data['AMT_INCOME_TOTAL'], bins, labels = slot)
```

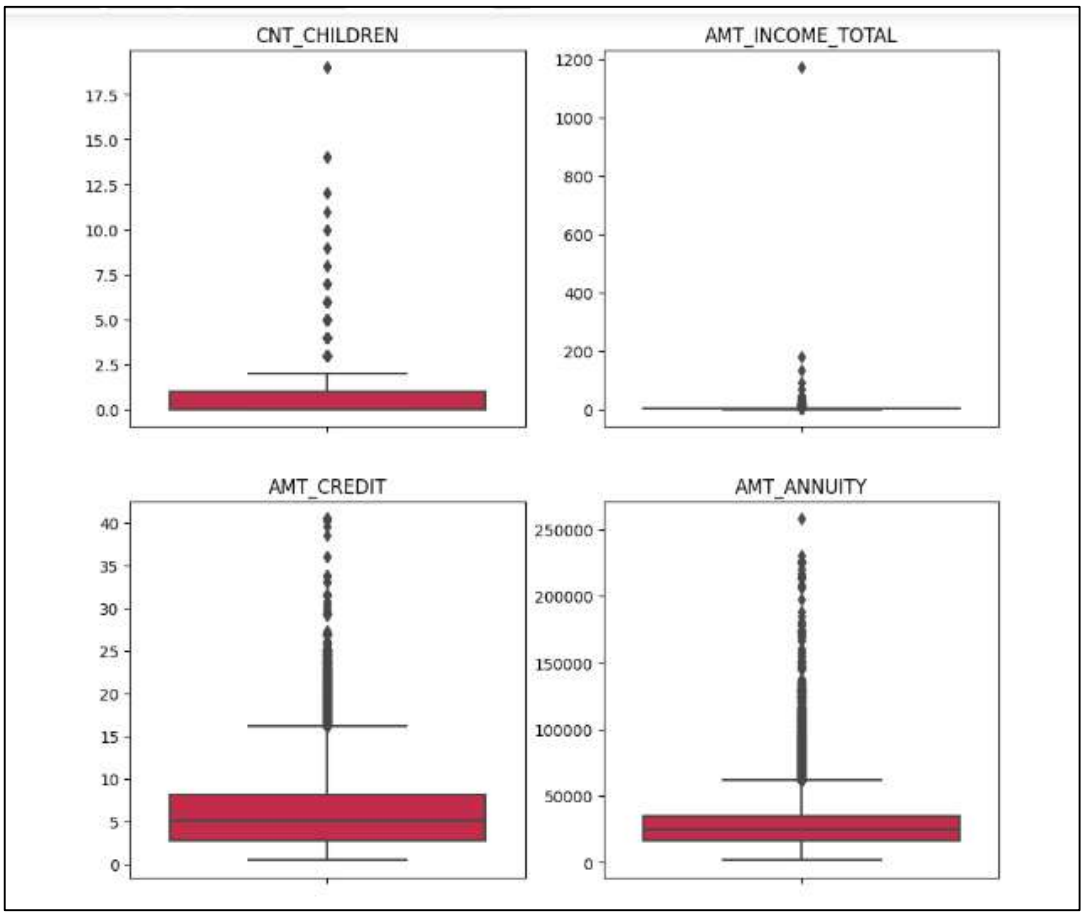
```
round((app_clean_data['AMT_INCOME_RANGE'].value_counts(normalize = True)*100),2)
```

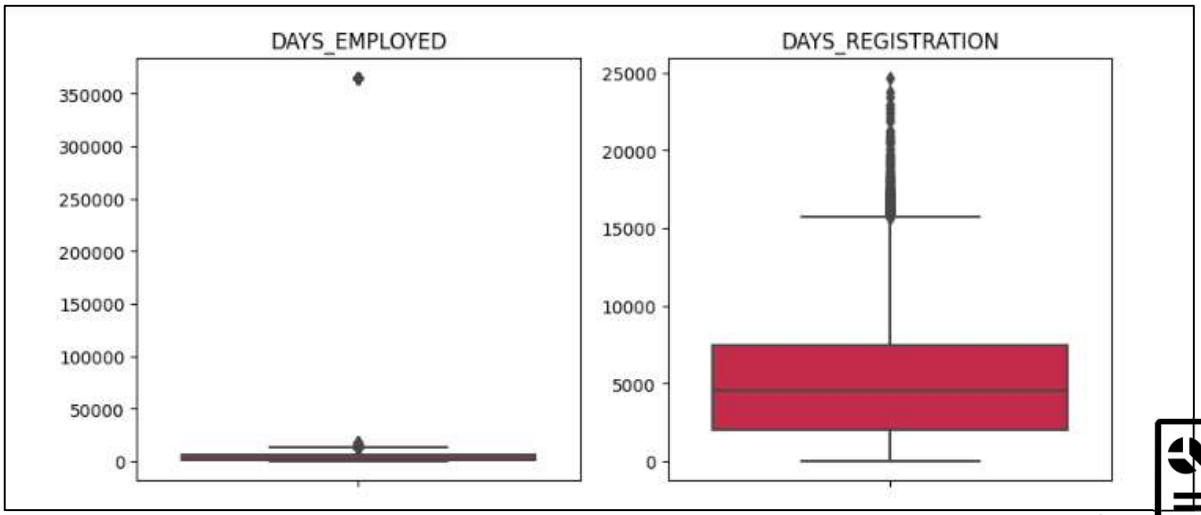
```
AMT_INCOME_RANGE
100k-200k    50.73
200k-300k    21.21
0k-100k      20.73
300k-400k     4.78
400k-500k     1.74
500k-600k     0.36
600k-700k     0.28
800k-900k     0.10
700k-800k     0.05
900k-1M       0.01
1M-Above      0.01
Name: proportion, dtype: float64
```

Outlier Detection

- Here, We have checked for six columns where four of them had high outliers as checked with the box plots.

Application Dataset





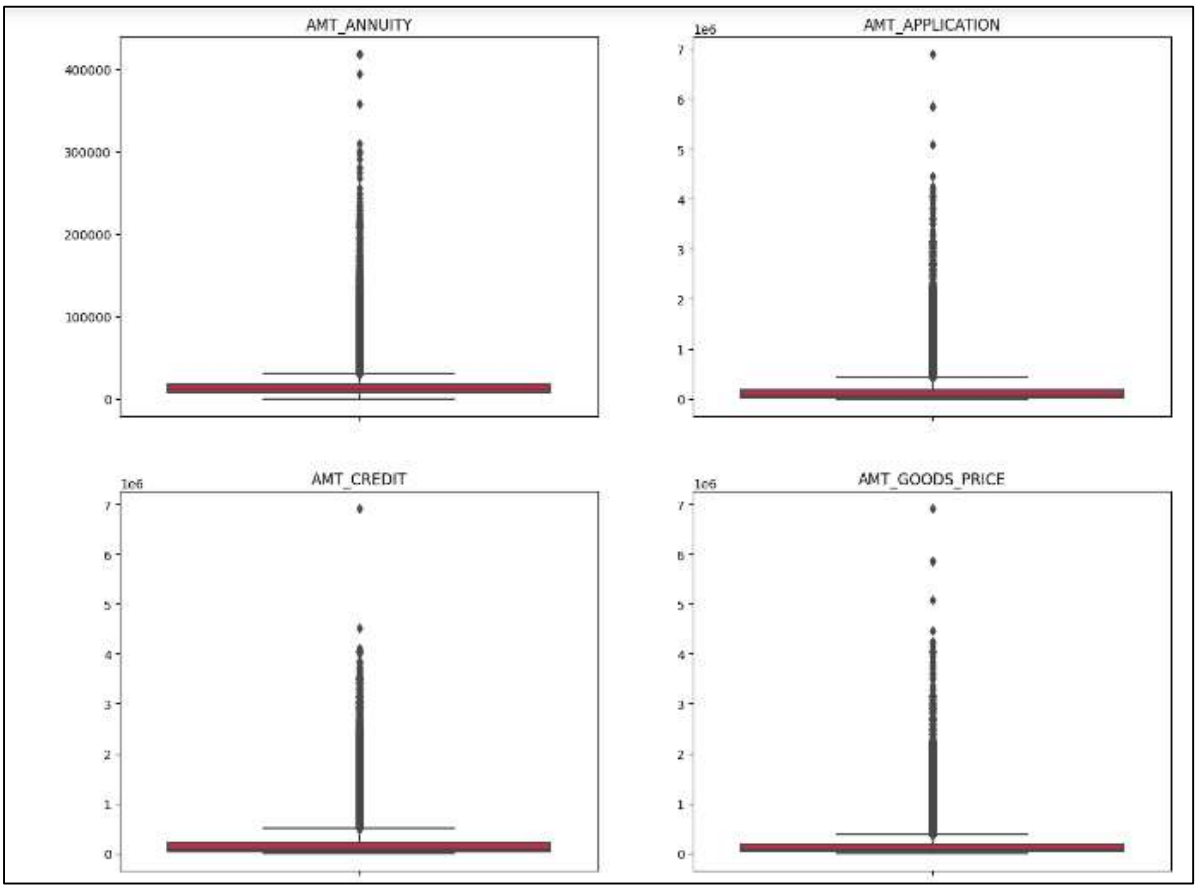
Bank Loan Case Study

- Insights:
- We can see that:
- CNT_CHILDREN, AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE have some number of outliers.
- AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.
- DAYS_BIRTH has no outliers which means the data available is reliable.
- DAYS_EMPLOYED has outlier values around 360000(days) which is around 986 years which is not possible.

Outlier Detection

- Here, We have checked for seven columns where five of them had high outliers as checked with the box plots.

Previous Application Dataset



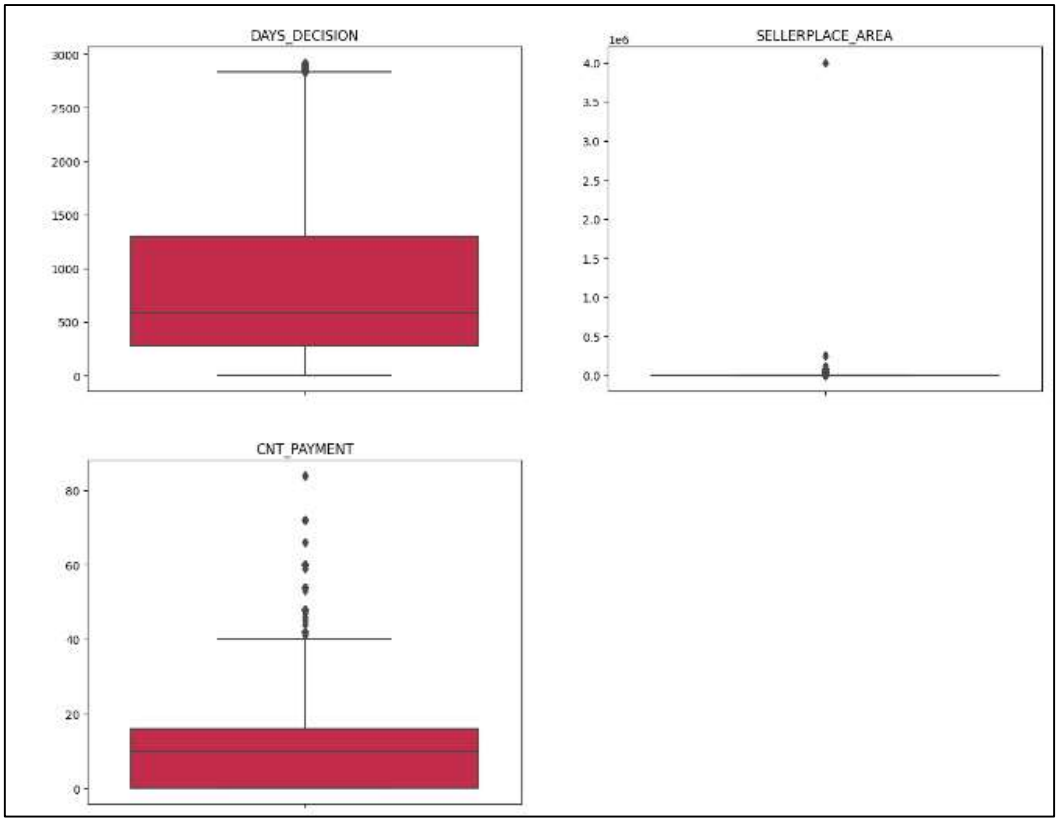
Insight:

It can be seen that in previous application data that:

- AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
- CNT_PAYMENT has few outlier values.
- DAYS_DECISION has little number of outliers indicating that these previous applications decisions were taken long back.

Bank Loan Case Study

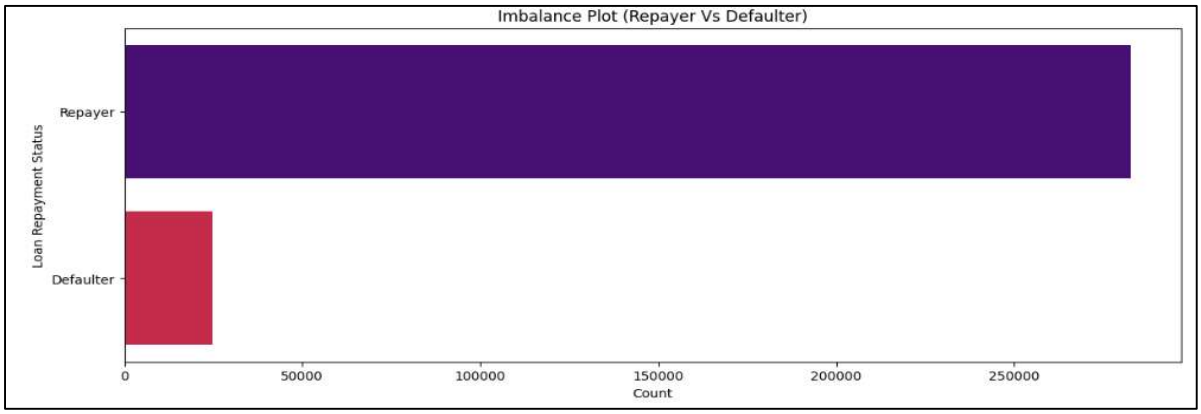
Previous Application Dataset



Data Imbalance

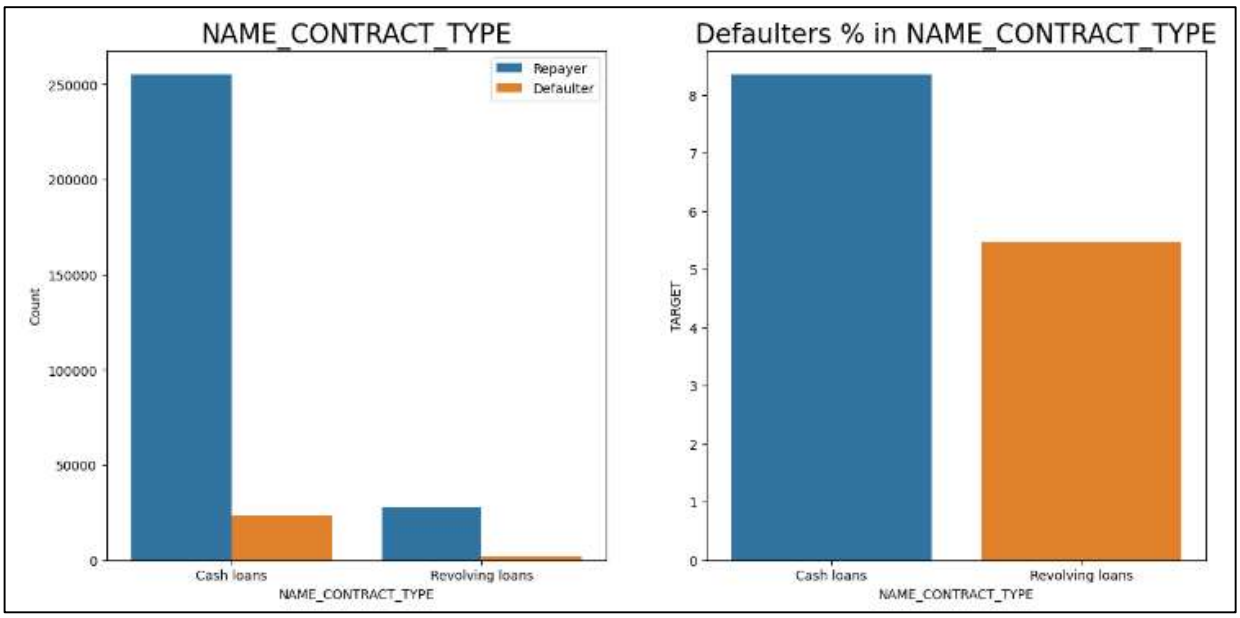
- Following figure shows the horizontal bar plot to check the repayer and defaulter ratio which shows the imbalance.

- The Imbalance ratio of Repayer to Defaulter is around (11.39:1).



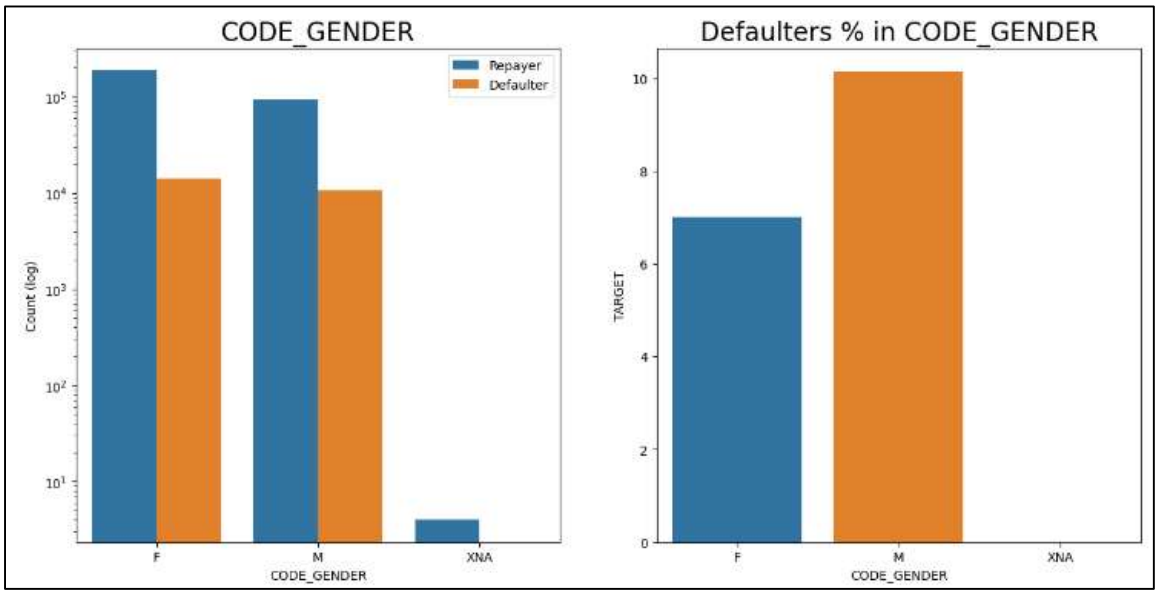
Bank Loan Case Study

Univariate Analysis on Categorical Variables



Insights: Contract type

- Revolving loans are just a small fraction (10%) from the total number of loans
- Around 8-9% Cash loan applicants and 5-6% Revolving loan applicants are in defaulters



Bank Loan Case Study

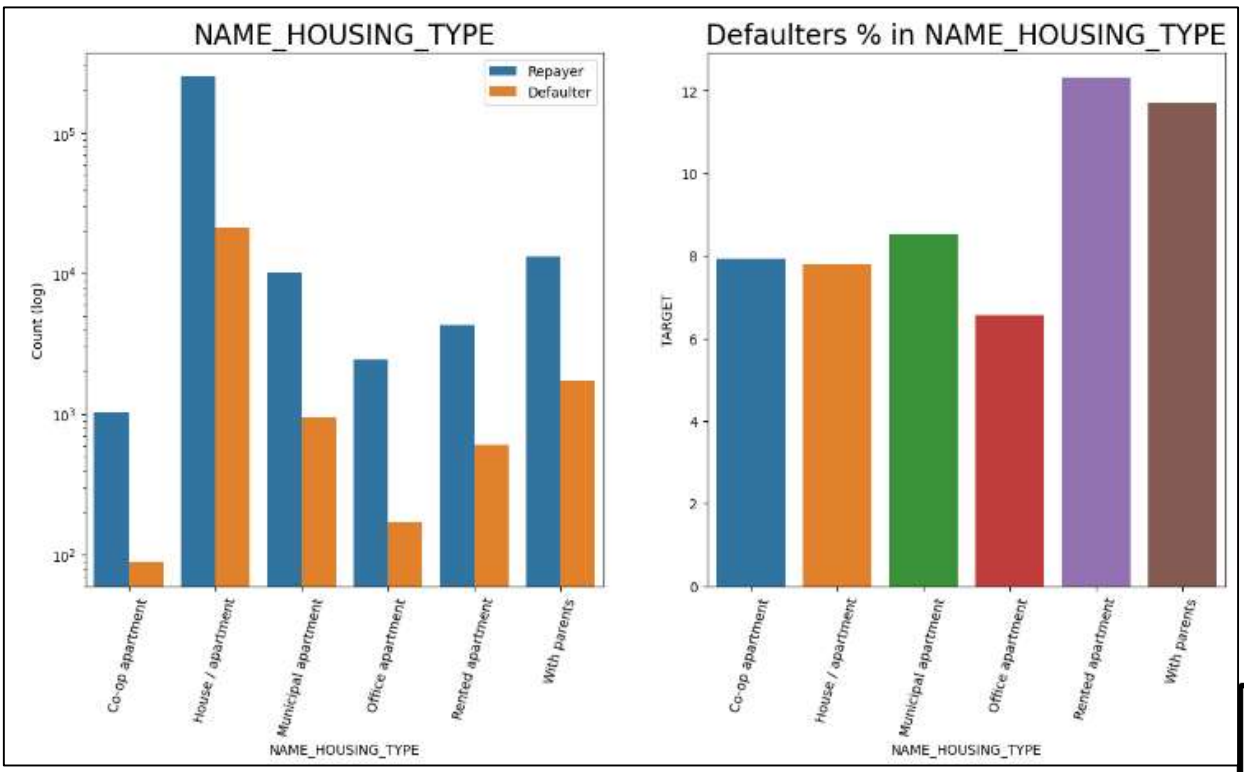
Univariate Analysis on Categorical Variables

Insights: Gender Type

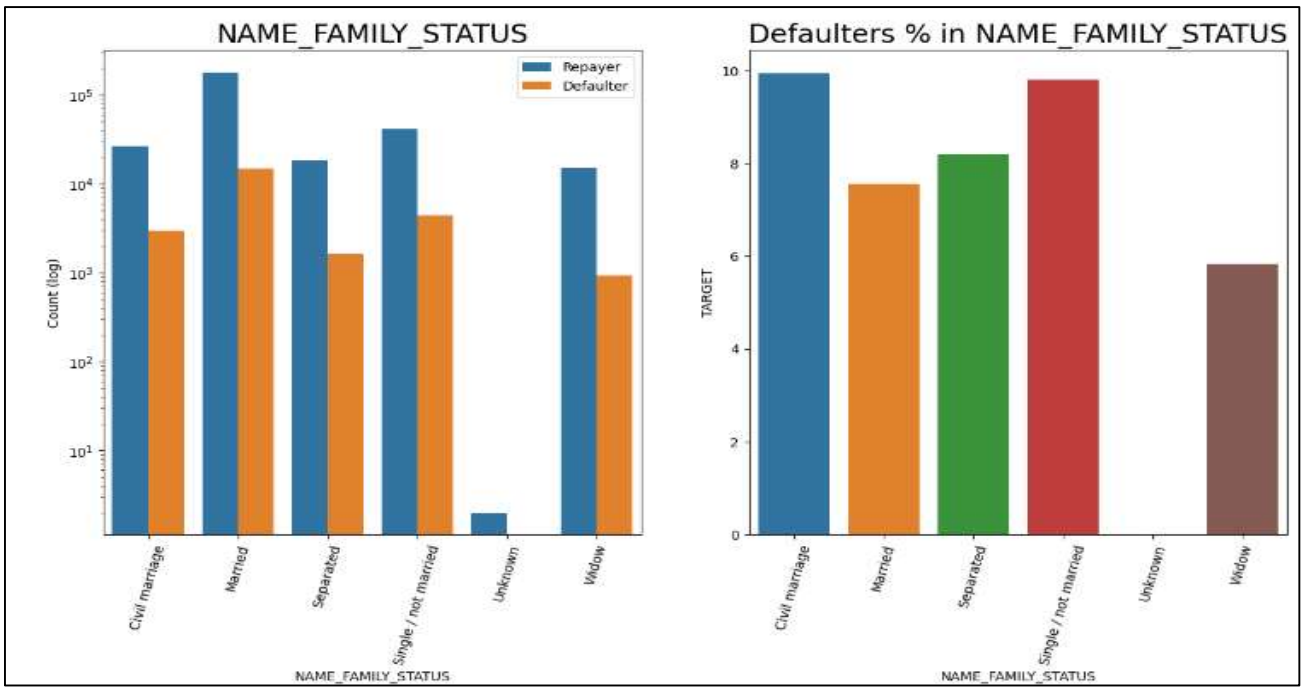
- The number of female clients is more than the number of male clients.
- Based on the percentage of defaulted credits, males have a higher chance of not returning their loans about 10%, comparing with women which is about 7%

Insights: House type

- Majority of people live in House/apartment
- People living in office apartments have lowest default rate
- People living in rented apartments(>12%) and living with parents (>11%) have higher probability of defaulting



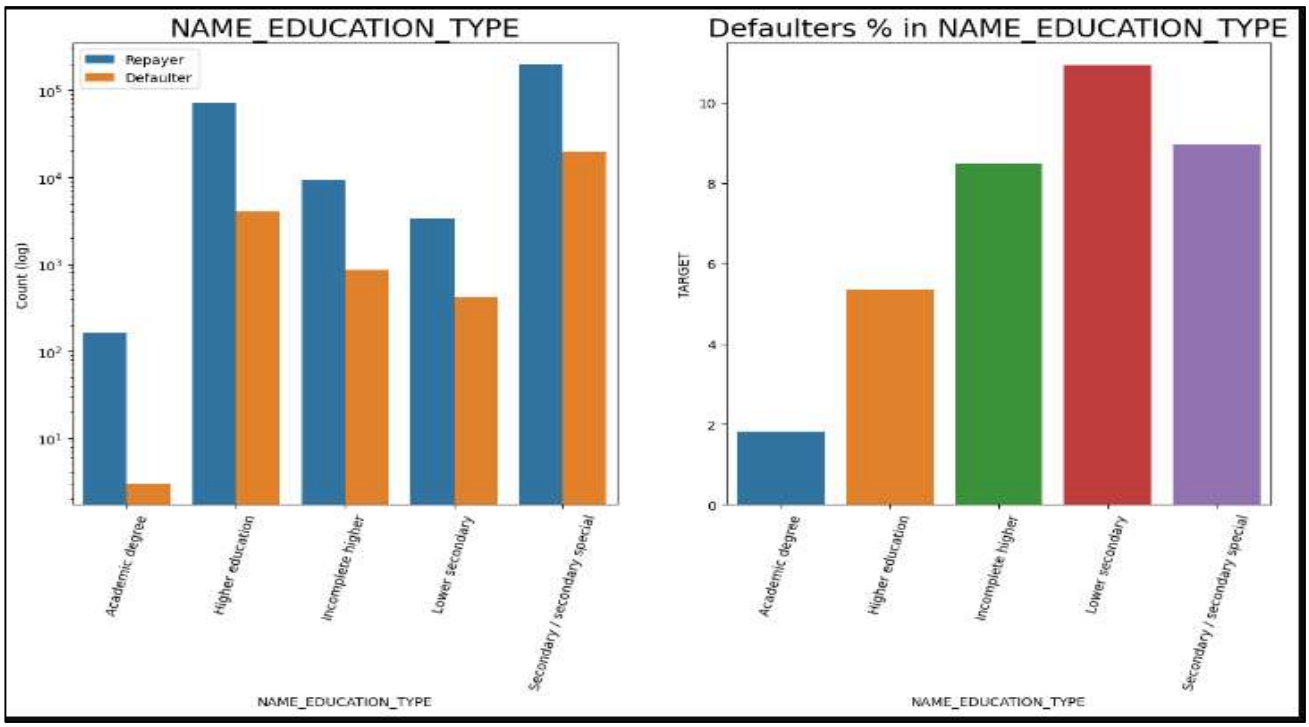
Univariate Analysis on Categorical Variables



Insights:

- Most of the people who have taken loan are married, followed by Single/not married and civil marriage
- In Percentage of defaulters, Civil marriage has the highest percent around (10%) and widow has the lowest around 6% (exception : Unknown).

Univariate Analysis on Categorical Variables

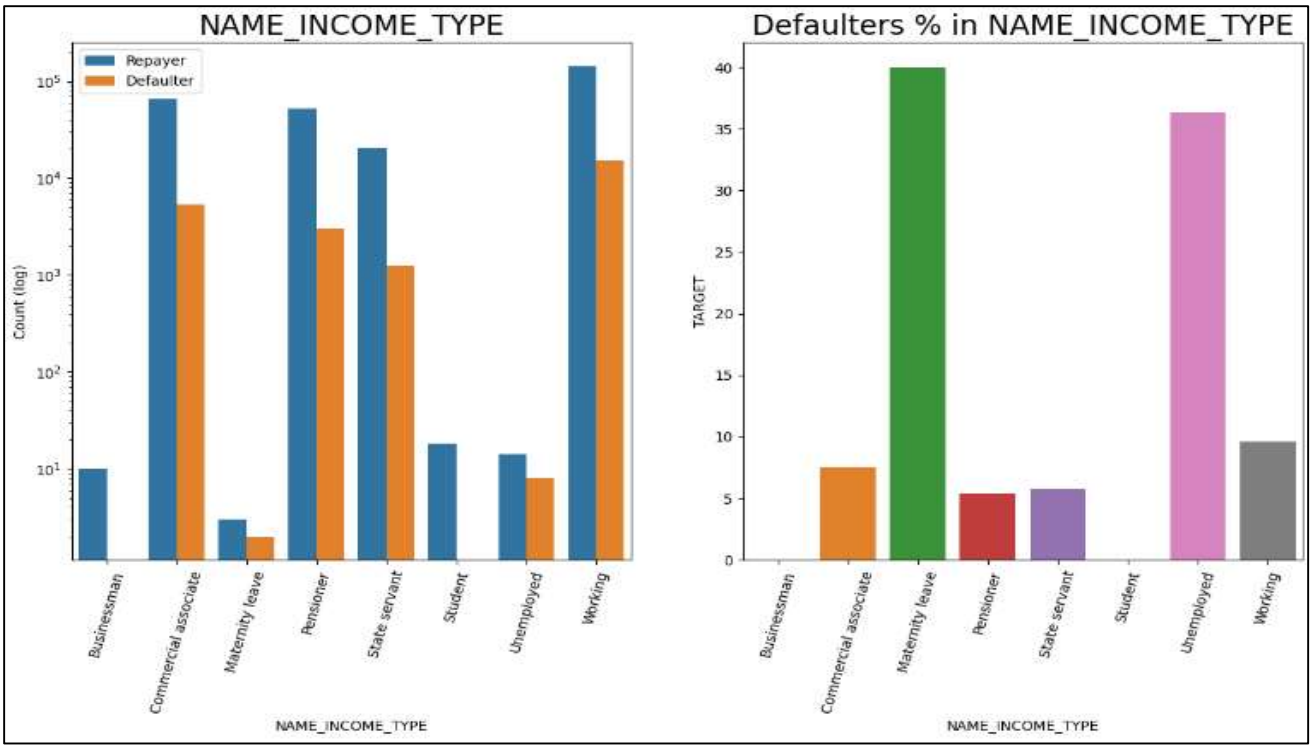


Insights: Education type

- Majority of clients have Secondary/secondary special education, followed by clients with Higher education.
- Very few clients have an academic degree
- Lower secondary category have highest rate of defaulting around 11%.
- People with Academic degree are least likely to default.

Bank Loan Case Study

Univariate Analysis on Categorical Variables

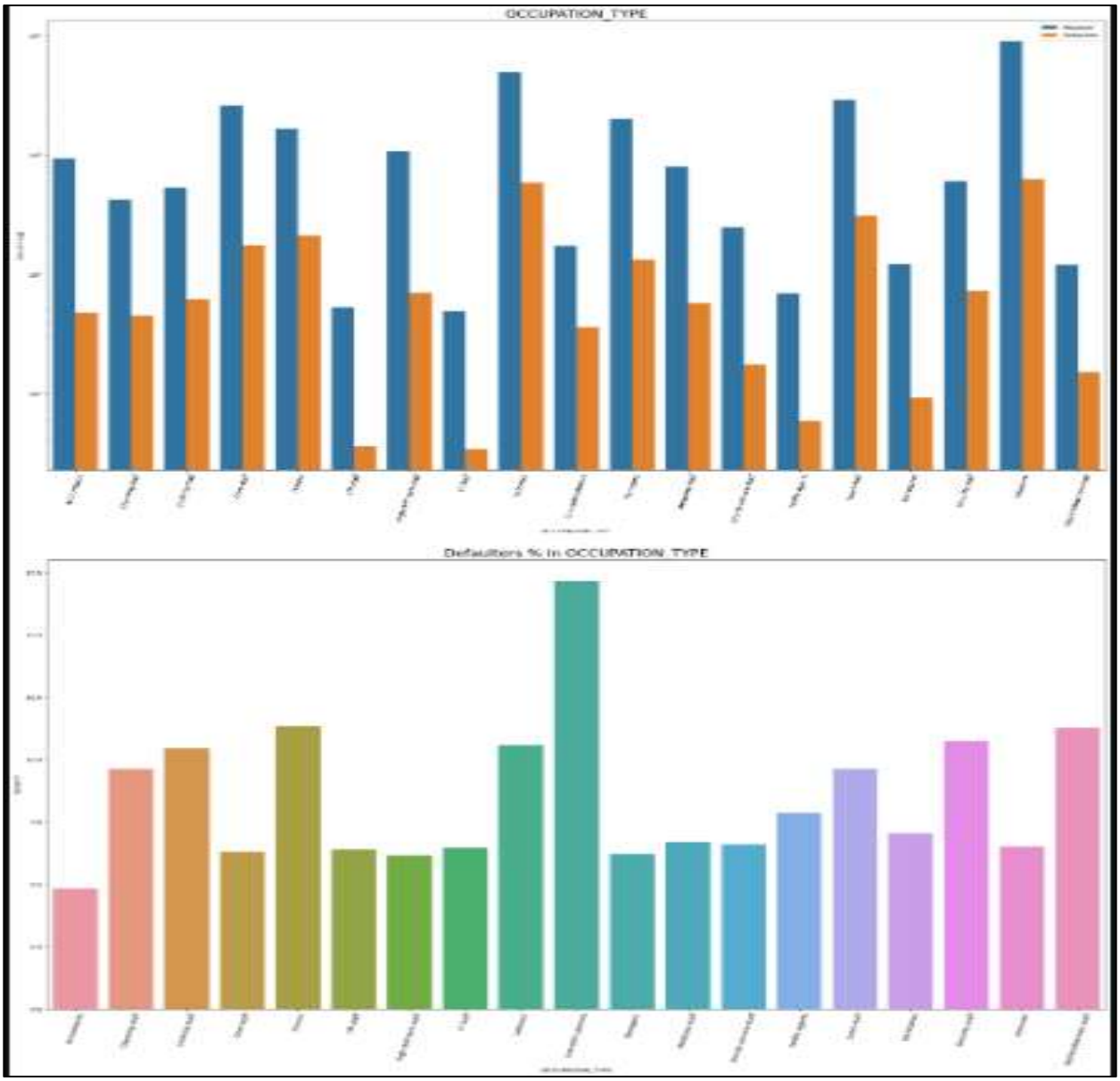


Insights: Income Type

- Most of applicants for loans income type is Working, followed by Commercial associate, Pensioner and State servant.
- The applicants who are on Maternity leave have defaulting percentage of 40% which is the highest, followed by Unemployed (~37%). The rest under average around 10% defaulters.
- Student and Businessmen though less in numbers, do not have default record. Safest two categories for providing loan

Bank Loan Case Study

Univariate Analysis on Categorical Variables

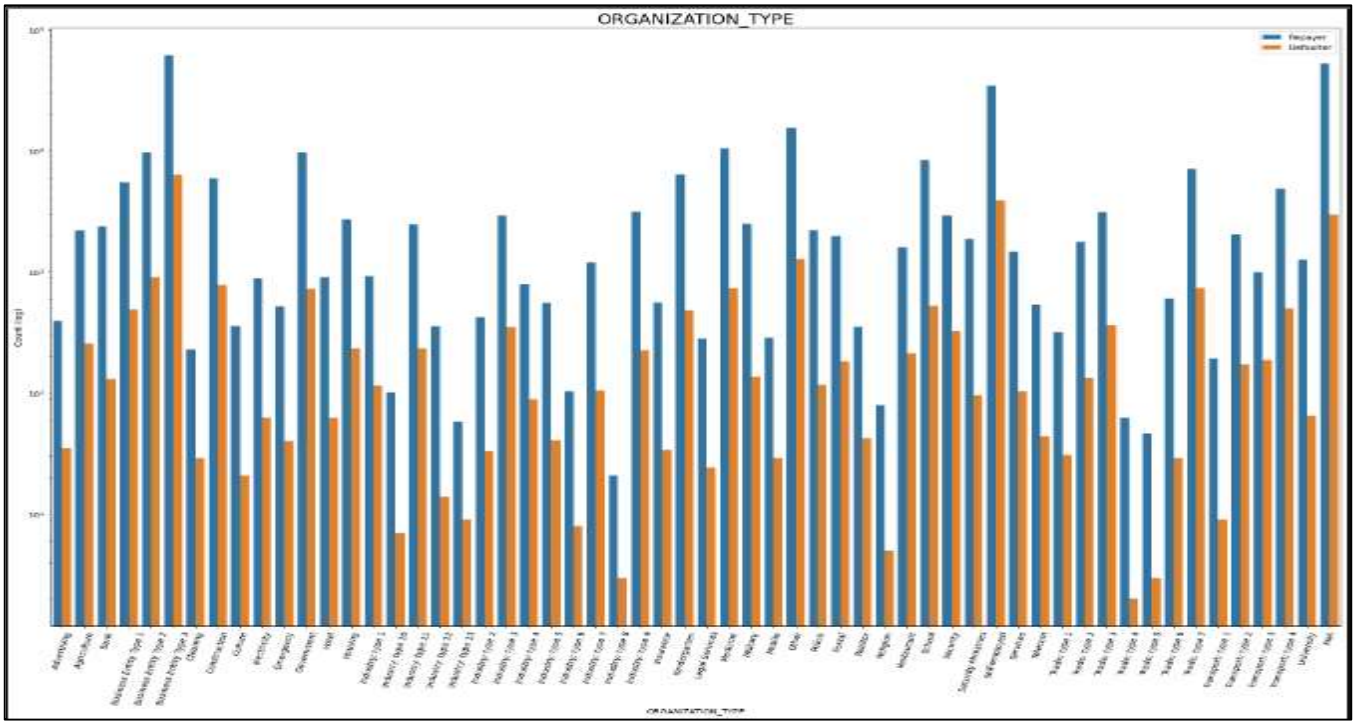


Insights: Occupation type

- Most of the loans are taken by Laborers, followed by Sales staff.(exception : Unknown)
- IT staff are less likely to apply for Loan.
- Category with highest percent of defaulters are Low-skill Laborers (above 17%), followed by Drivers and Waiters/barmen staff, Security staff

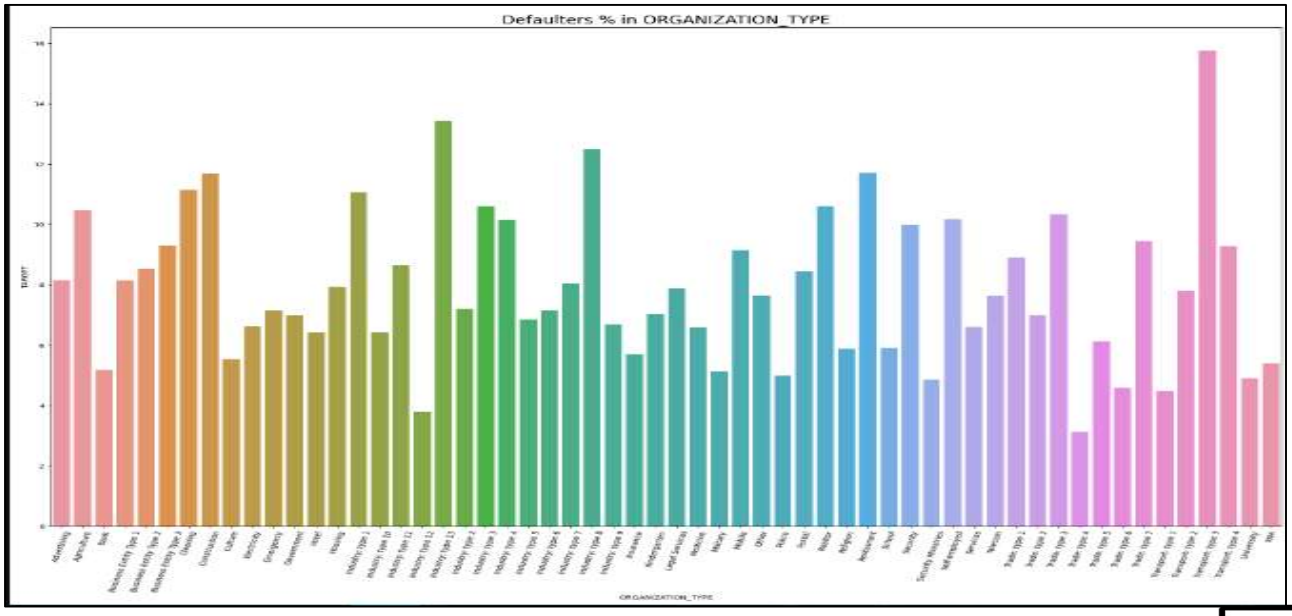
Bank Loan Case Study

Univariate Analysis on Categorical Variables



Insights: Organization Type:

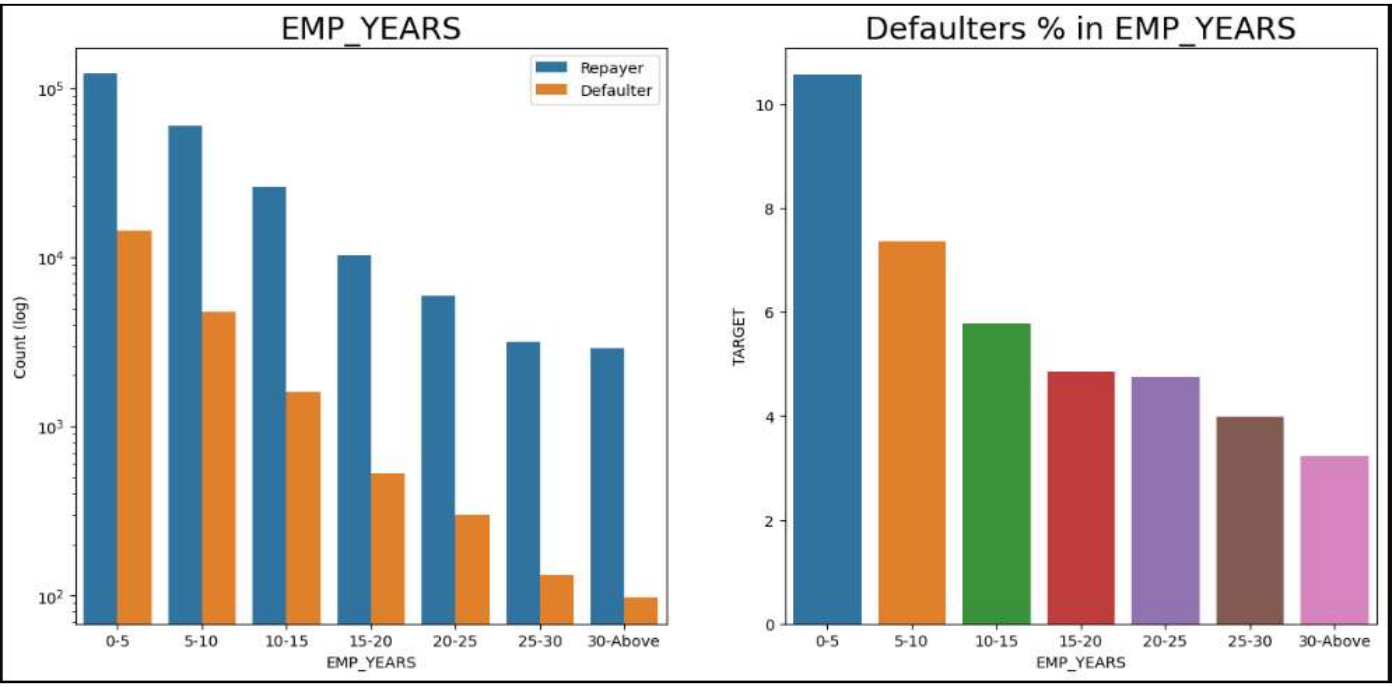
- **Organizations with highest percent of defaulters are Transport: type 3 (~16%), Industry: type 13 (~13.5%), Industry: type 8 (>12%) and Restaurant (less than 12%).**



- **Most of the people are from Business Entity Type 3.**
- **Trade Type 4 and 5, Industry type 8 has lesser defaulters.**

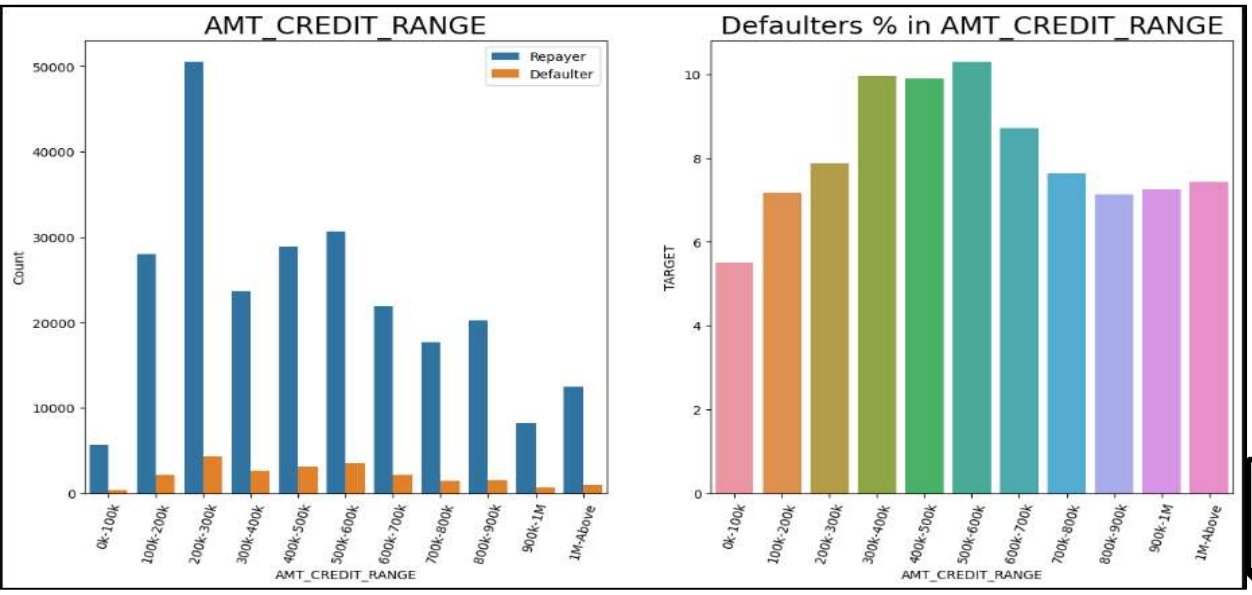
Bank Loan Case Study

Univariate Analysis on Categorical Variables



Insights: EMP in Years

- Majority of the applicants having working experience between 0-5 years are defaulters. The defaulting rating of this group is also the highest which is above 10%
- With increase of employment year, defaulting rate is gradually decreasing.

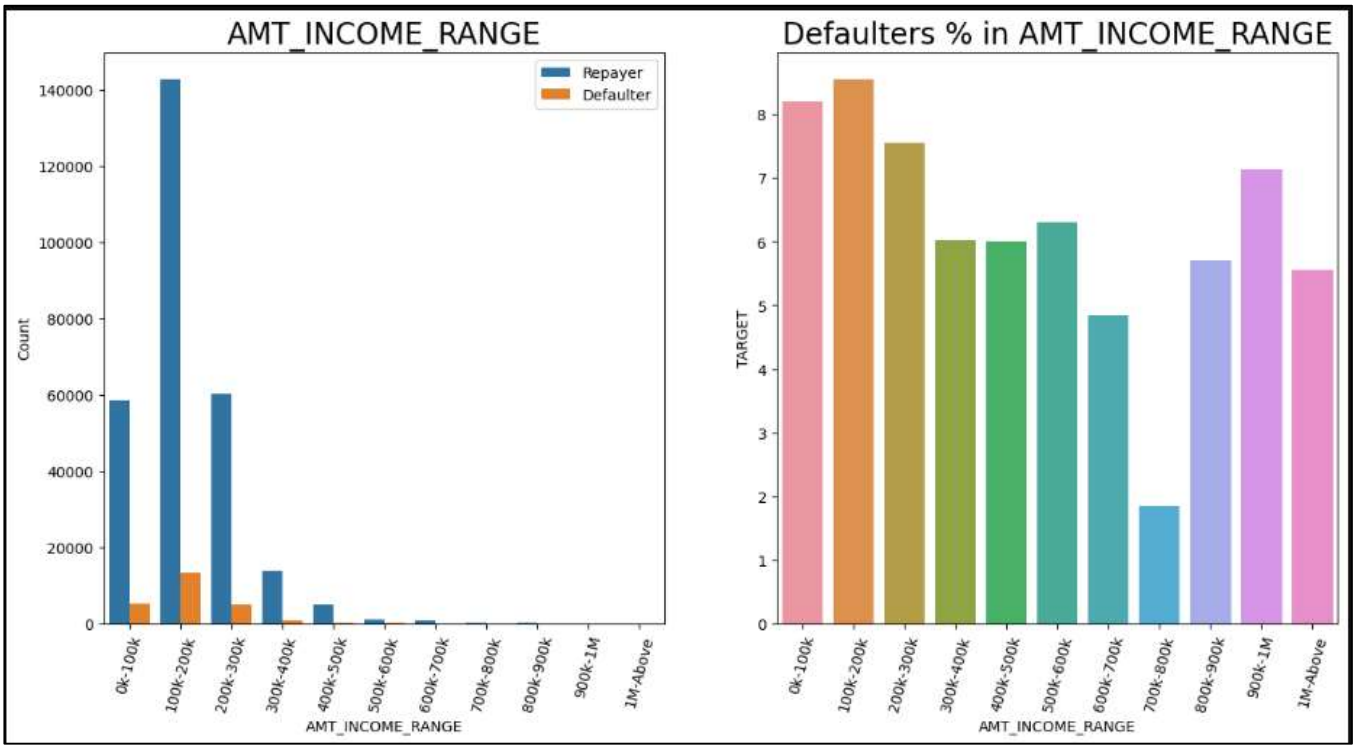


Bank Loan Case Study

Univariate Analysis on Categorical Variables

Insights: Credit Amount

There are high number of applicants have loan in range of 2-3 Lakhs followed by 5-6 Lakh range
People who get loan for 500k-600k have most number of defaulters than other credit range.

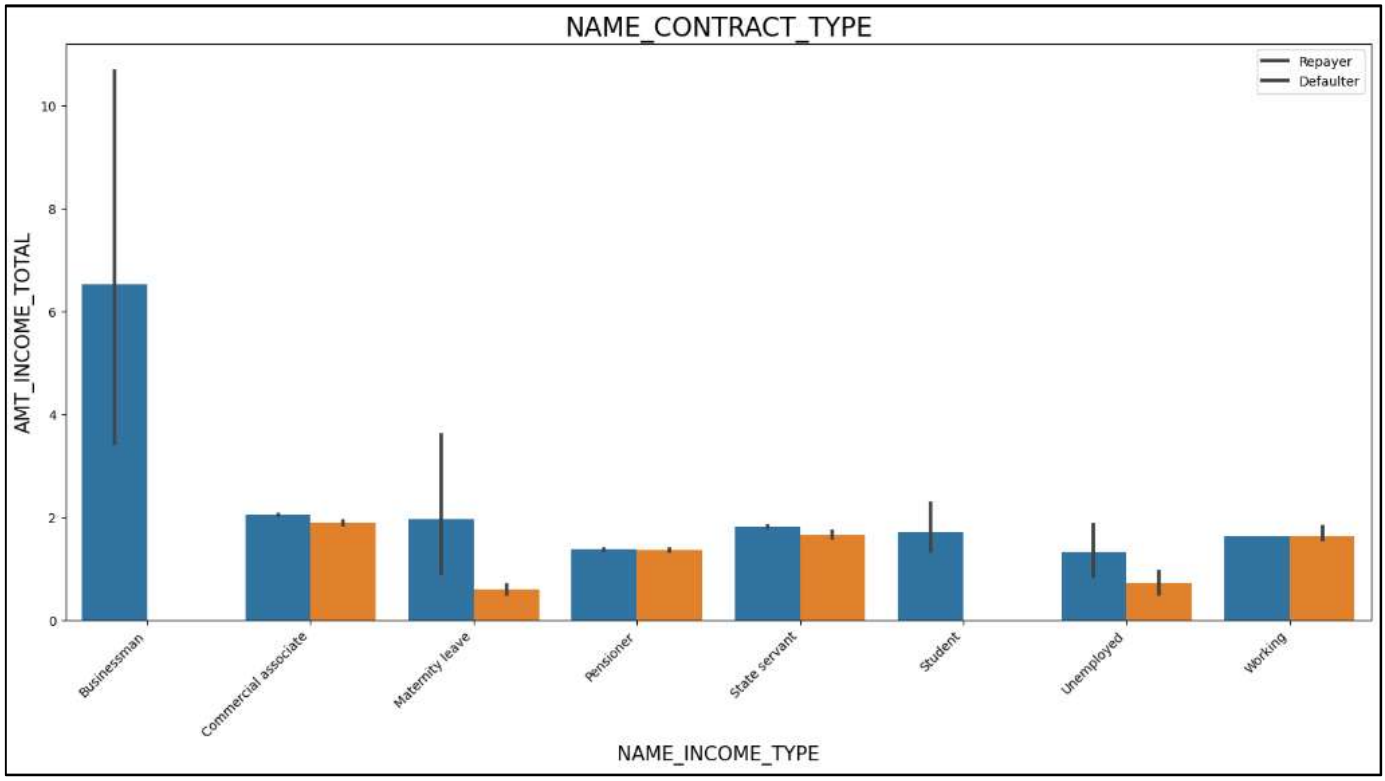


Insights: Income Range

- **Highest Defaulters are in 1-2 lakh range which is above 8% while the least is in 7-8 lakh range which is around 2%**

Bank Loan Case Study

Categorical Bivariate Analysis



- Here, I have taken the CONTRACT TYPE column with INCOME TYPE and got the descriptive stats to check the type of business which was having highest income
- As per the plot, businessman has the highest income.

Bank Loan Case Study

Top 10 correlation between variables w.r.t TARGET

- Here, I have created two dataframes to store repayers for all columns and defaulters to one.
- Then, using the correlation function between each variable listed the top 10 below.

Repayer (TARGET 0)

	VAR1	VAR2	Correlation
262	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508
64	AMT_GOODS_PRICE	AMT_CREDIT	0.987250
284	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859332
65	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
43	AMT_ANNUITY	AMT_CREDIT	0.771309
131	DAYS_EMPLOYED	DAYS_BIRTH	0.626114
42	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953
63	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349462
21	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799
152	DAYS_REGISTRATION	DAYS_BIRTH	0.333151

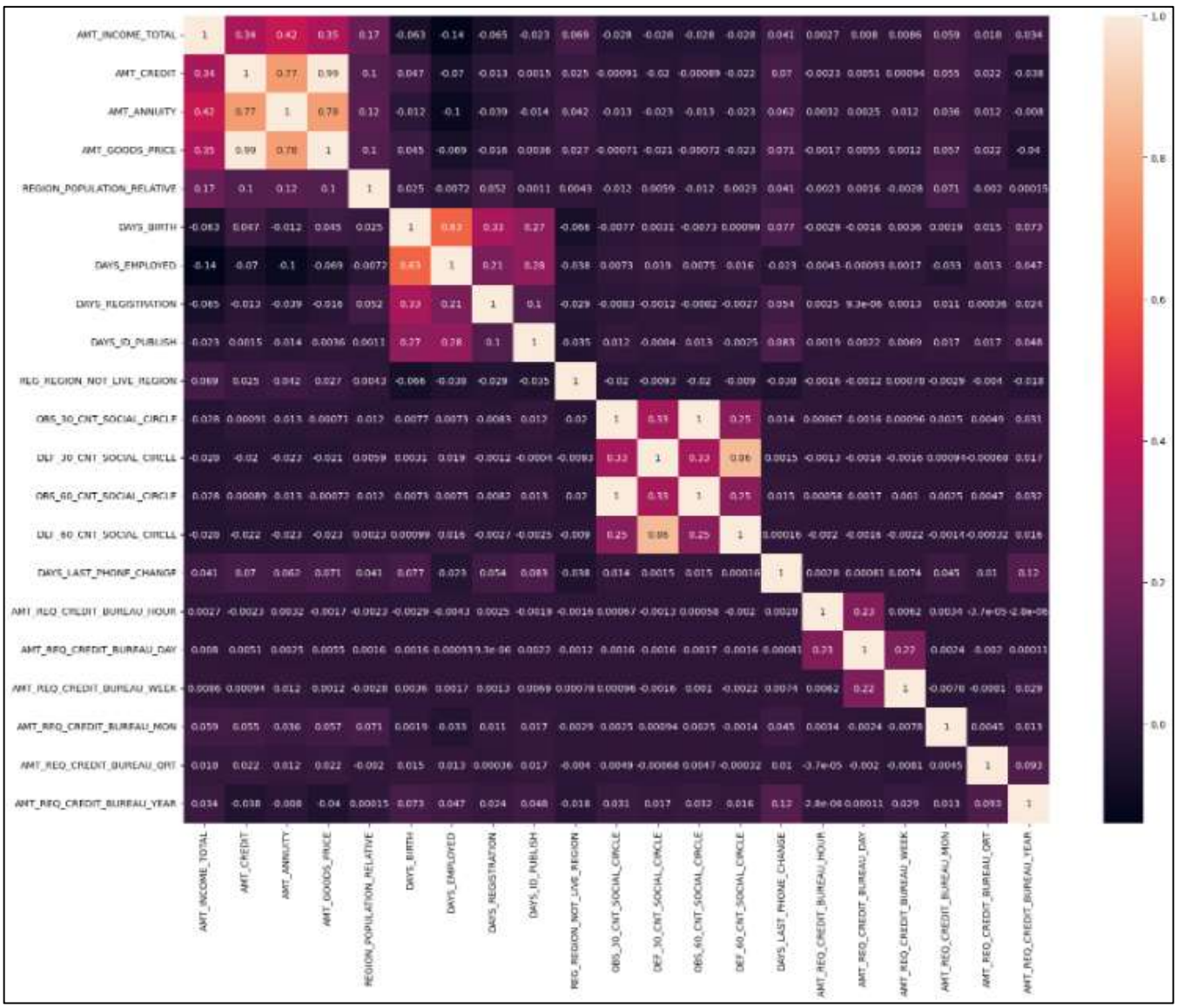
Defaulter (TARGET 1)

	VAR1	VAR2	Correlation
262	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269
64	AMT_GOODS_PRICE	AMT_CREDIT	0.983103
284	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994
65	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
43	AMT_ANNUITY	AMT_CREDIT	0.752195
131	DAYS_EMPLOYED	DAYS_BIRTH	0.582185
263	OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.337181
241	DEF_30_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.333825
152	DAYS_REGISTRATION	DAYS_BIRTH	0.289114
285	DEF_60_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.264159

Bank Loan Case Study

Correlation Matrix

The following heatmap shows the linear correlation of variables among repayers



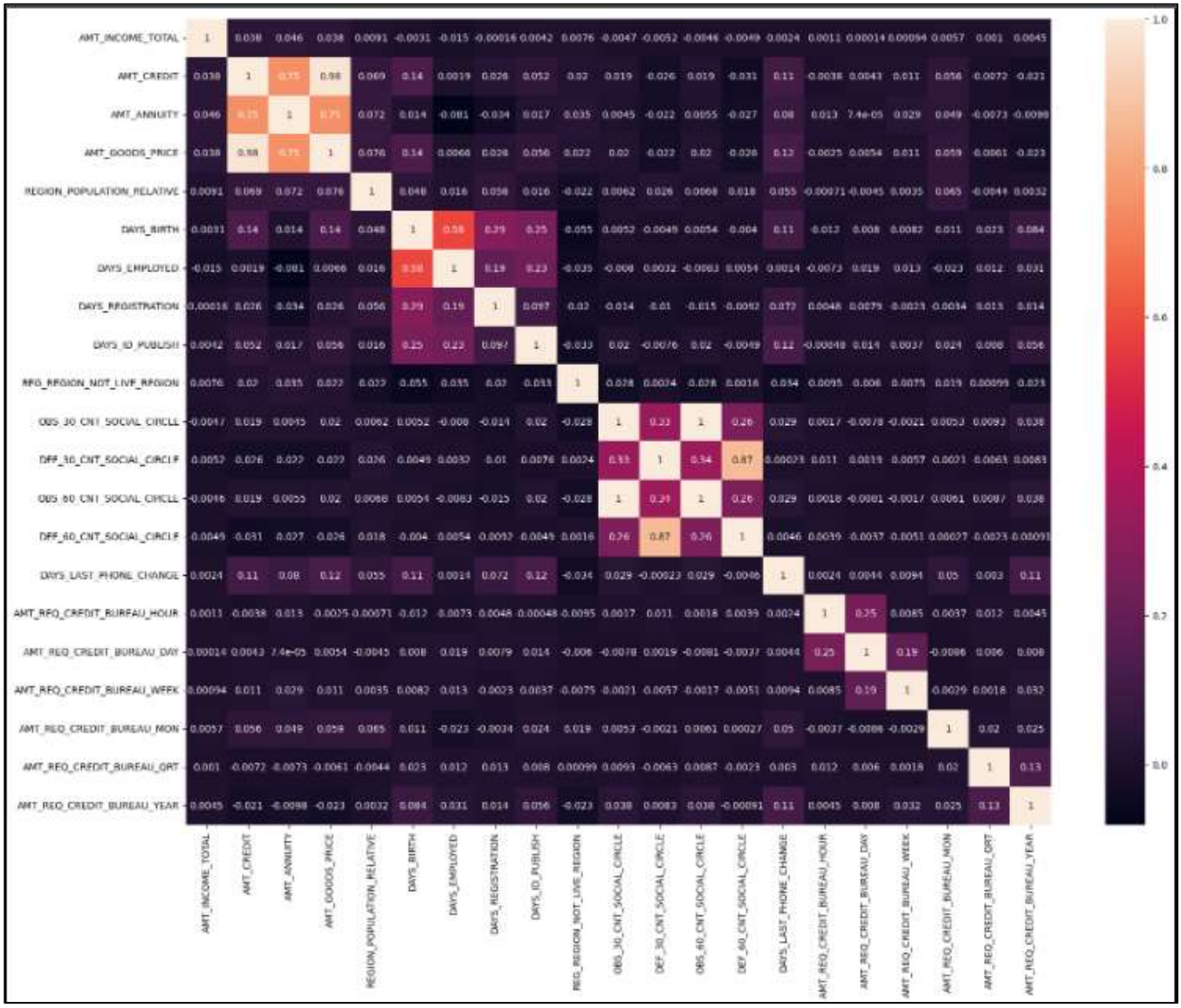
Insights: Correlating factors amongst repayers

- **Credit amount is highly correlated with:**
  - Goods Price Amount
  - Loan Annuity
  - Total Income

Bank Loan Case Study

Correlation Matrix

The following heatmap shows the linear correlation of variables among defaulters

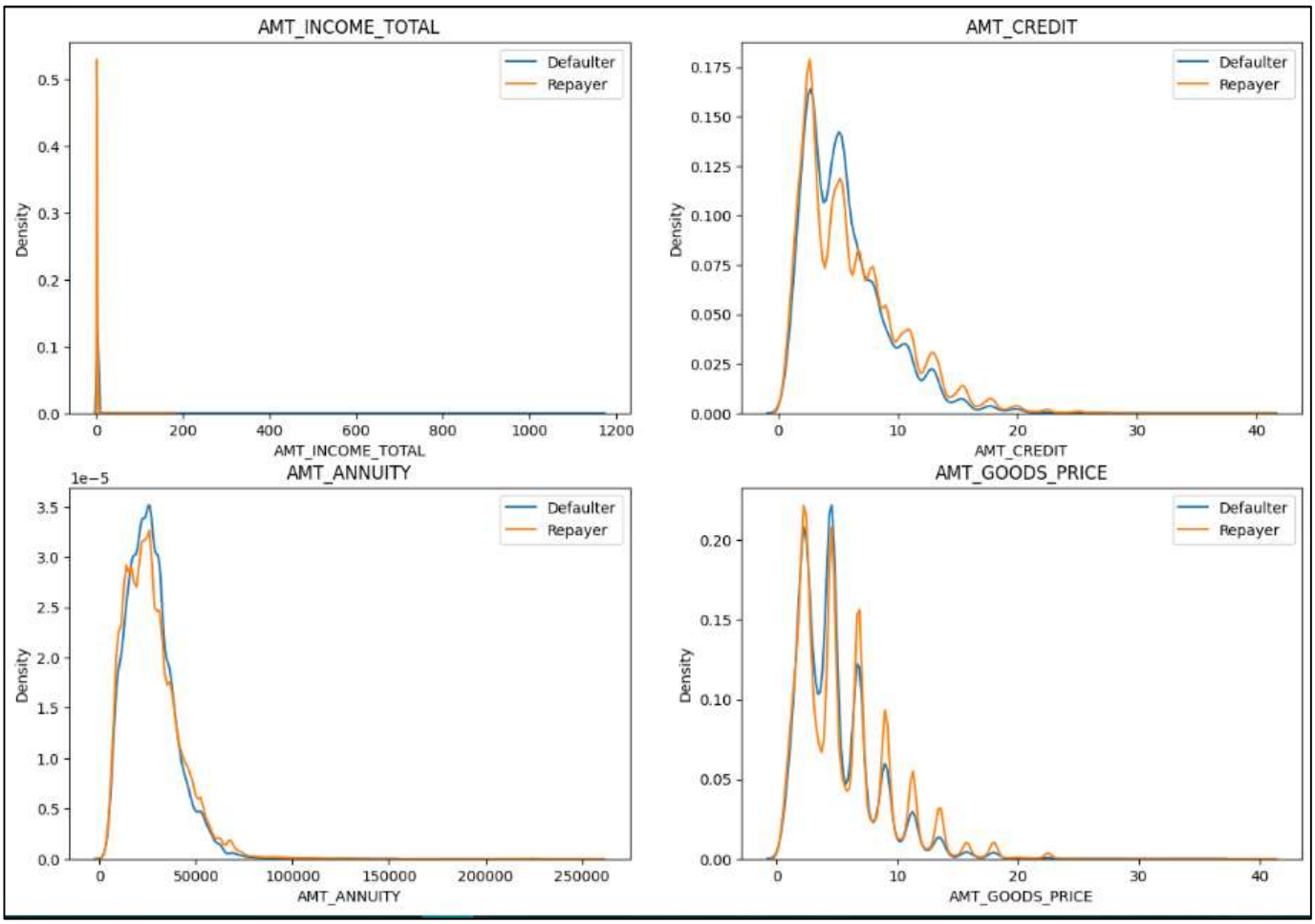


Insights: Correlating factors amongst defaulters

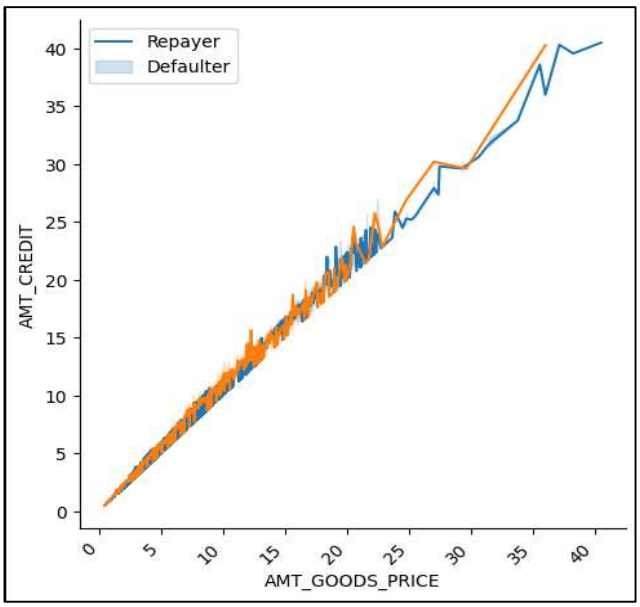
- Credit amount is highly correlated with good price amount which is same as repayers.
- Loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayers(0.77)
- There is a severe drop in the correlation between total income of the client and the credit amount(0.038) amongst defaulters whereas it is 0.34 among repayers.

Bank Loan Case Study

Numerical Univariate Analysis



- Most no of loans are given for goods price below 10 lakhs
- Most people pay annuity below 50K for the credit loan
- Credit amount of the loan is mostly less then 10 lakhs

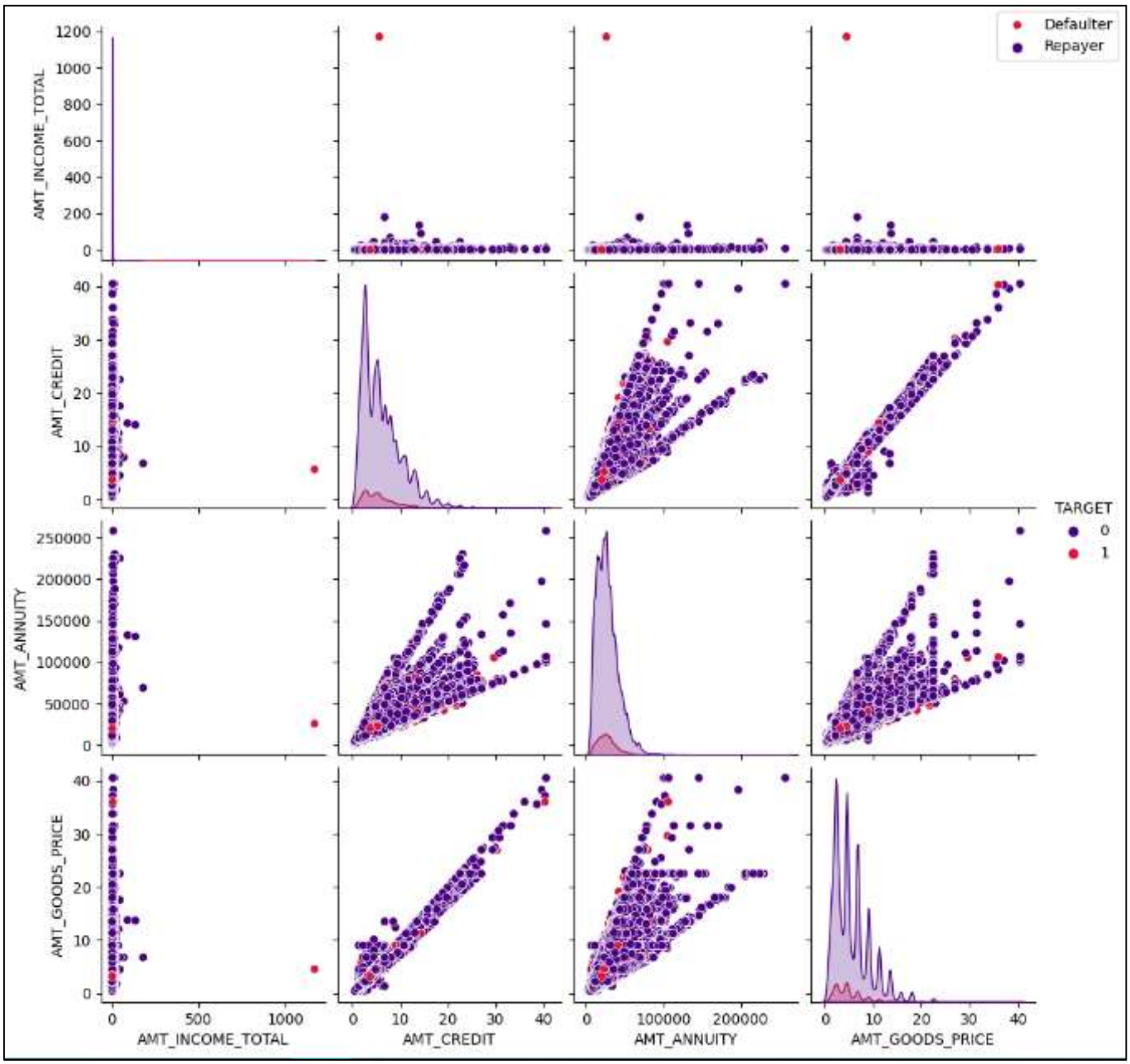


Numerical Bivariate Analysis

- The following graph is the relationship b/w Goods price, credit and comparing with loan repayment status
- Based on this, the defaulter percent was increasing as the credit amount and goods_price increased and this continued more after 30 lakhs

Bank Loan Case Study

Numerical Multivariate Analysis



- When Annuity Amount > 15K and Good Price Amount > 20 Lakhs, there is a lesser chance of defaulters
- Loan Amount(AMT_CREDIT) and Goods price(AMT_GOODS_PRICE) are highly correlated as based on the scatterplot where most of the data are consolidated in form of a line
- There are very less defaulters for AMT_CREDIT >20 Lakhs

Bank Loan Case Study

Merged Data Analysis

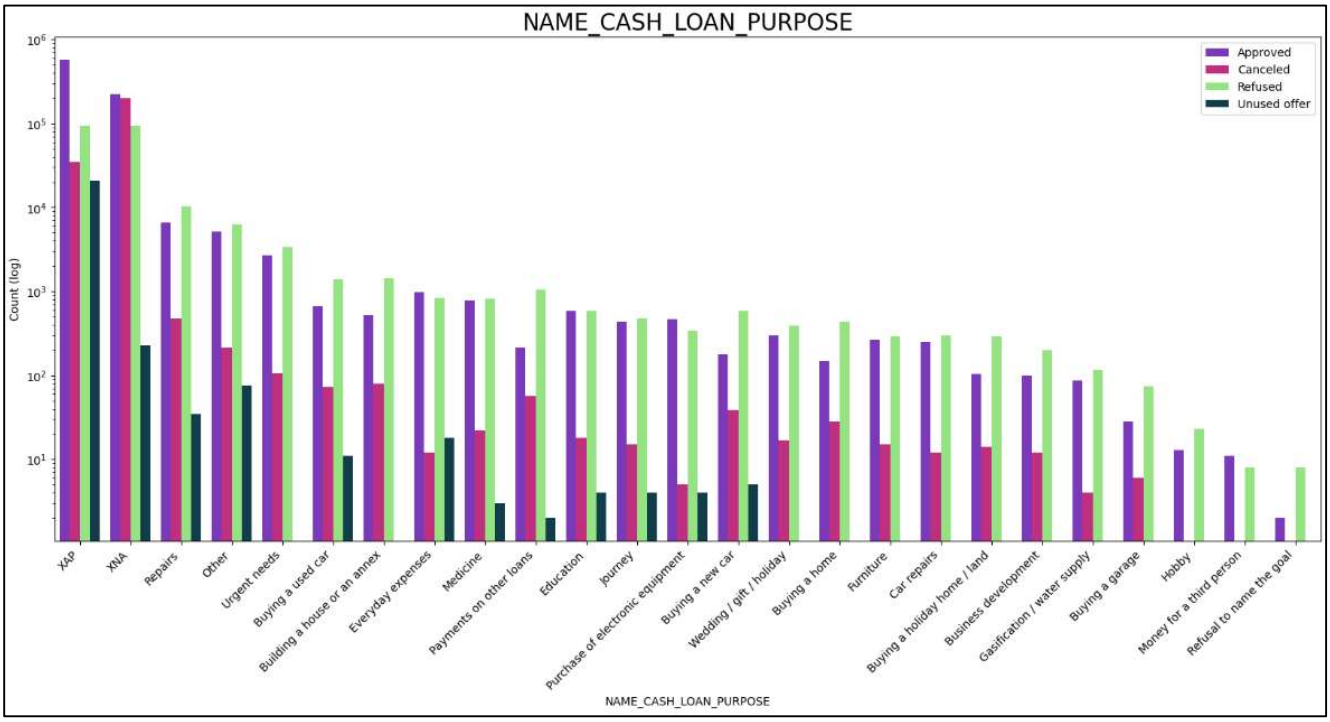
- Here, the next analysis shows on the data frame formed from merging both the given datasets and this was possible by using "SK_ID_CURR" variable

```
# merge both the dataframe on SK_ID_CURR with Inner Joins
merge_df = pd.merge(app_clean_data, prev_app, how='inner', on='SK_ID_CURR')
merge_df.head()
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE_x	CODE_GENDER	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT_x	AMT_ANNUITY_x	AMT_GOODS_
0	100002	1	Cash loans	M	0	2.025	4.065975	24700.5	
1	100003	0	Cash loans	F	0	2.700	12.935025	35698.5	
2	100003	0	Cash loans	F	0	2.700	12.935025	35698.5	
3	100003	0	Cash loans	F	0	2.700	12.935025	35698.5	
4	100004	0	Revolving loans	M	0	0.675	1.350000	6750.0	

```
merge_df.shape
```

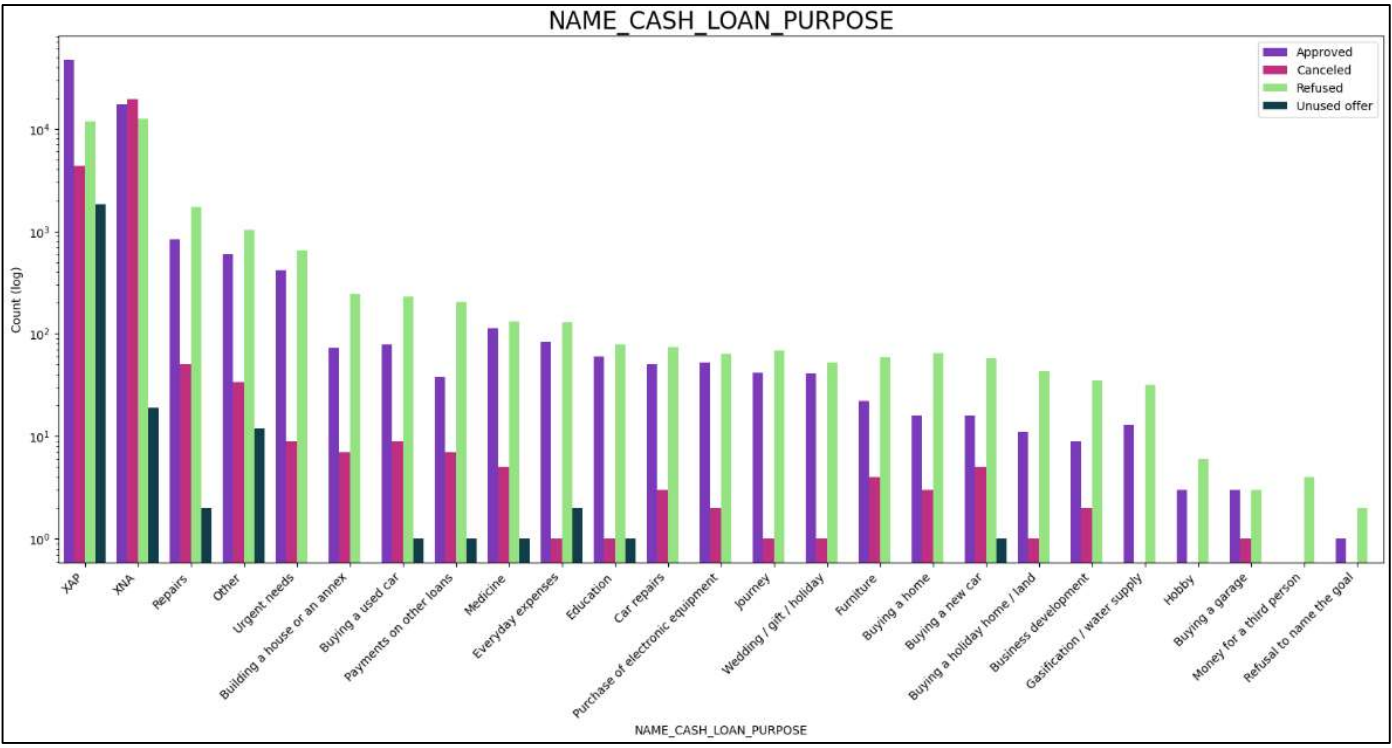
```
(1413701, 82)
```



- Here, I have created two dataframes which is the split of merged dataframe where one holds the repayer data with merged data and other one contains defaulter data.

Bank Loan Case Study

Merged Data Analysis



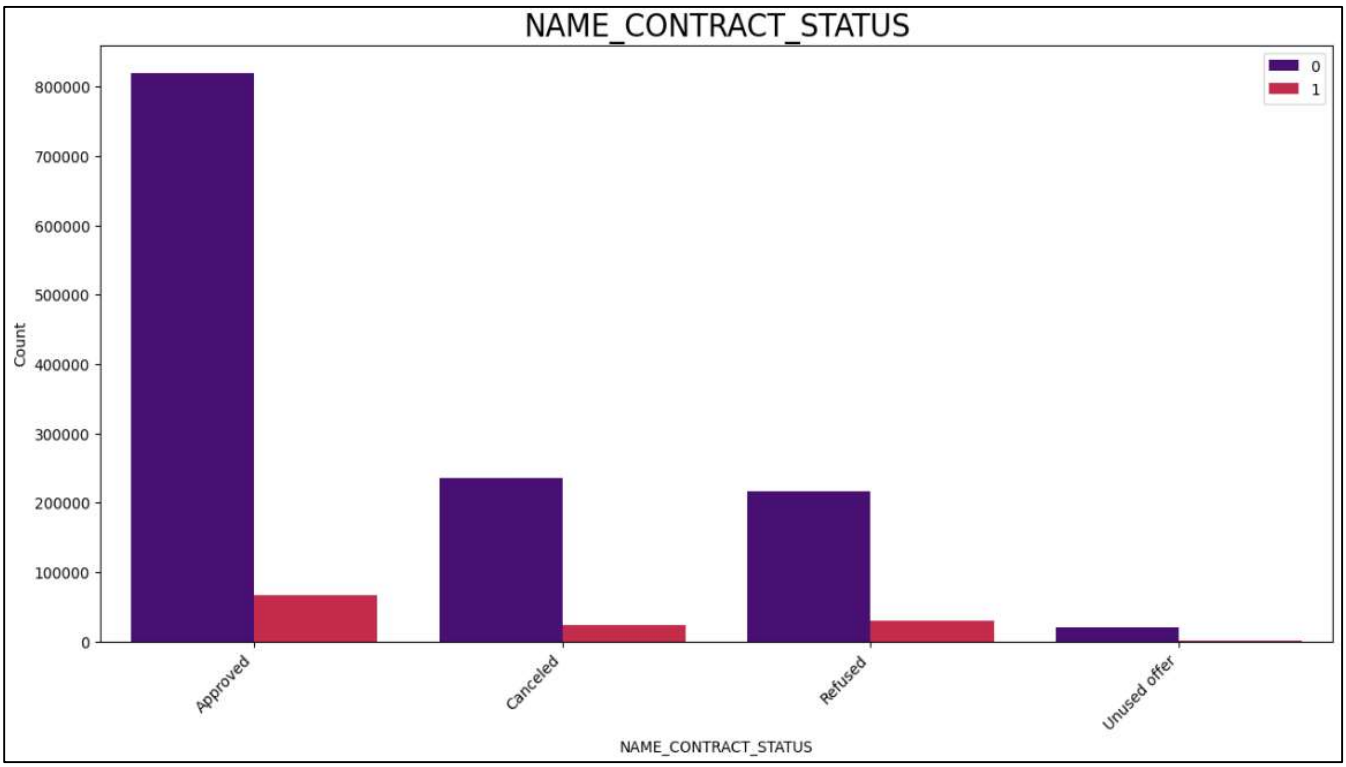
- Loan purpose has high number of unknown values (XAP, XNA)
- Loan taken for the purpose of Repairs looks to have highest default rate
- Huge number application have been rejected by bank or refused by client which are applied for Repair or Other.

NAME_CONTRACT_STATUS	TARGET	Counts	Percentage
Approved	0	818856	92.41%
	1	67243	7.59%
Canceled	0	235641	90.83%
	1	23800	9.17%
Refused	0	215952	88.0%
	1	29438	12.0%
Unused offer	0	20892	91.75%
	1	1879	8.25%

Bank Loan Case Study

Merged Data Analysis

- This plot was to check the Contract Status based on loan repayment status whether there is any business loss or financial loss



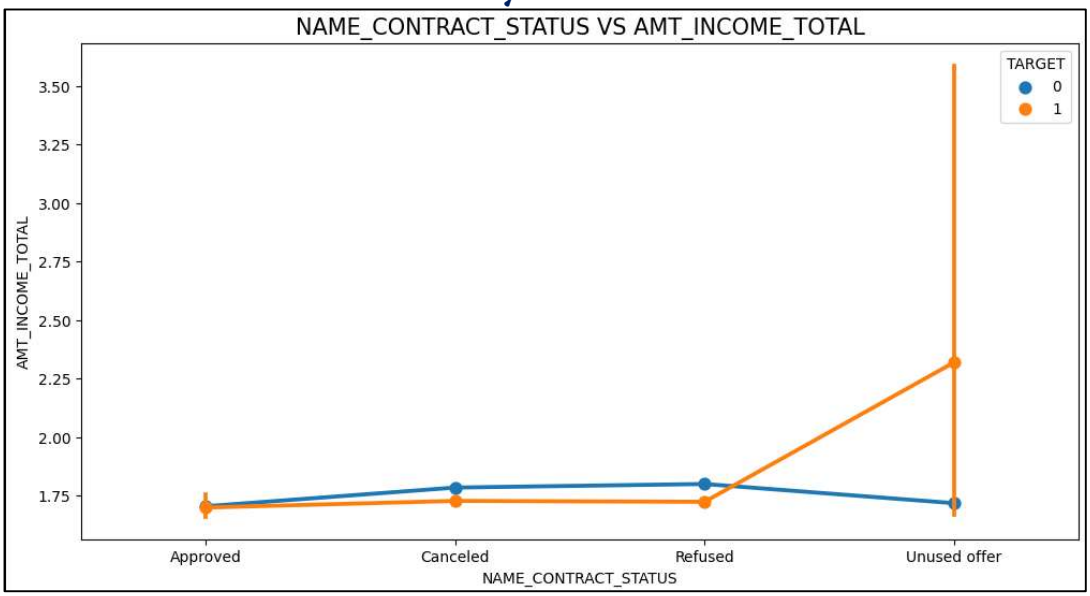
Insights:

- More than 92% of the previously canceled client have repaid the loan. Revising the interest rates would increase business opportunity for these clients
- 88% of the clients who have been previously refused a loan has payed back the loan in present case.
- Refusal reason should be recorded for further analysis as these clients could turn into potential repaying customer.

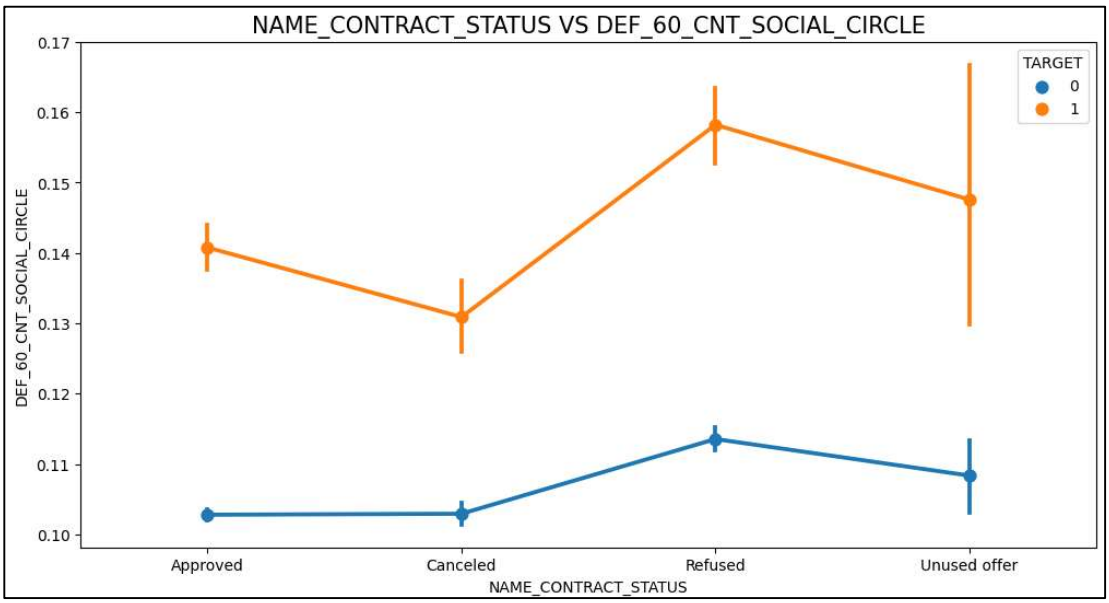
Bank Loan Case Study

Merged Data Analysis

- The graphs shows the relationship between income total and contract status as well as between contract status and on people who were defaulters for 60 days



- The point plot show that the people who have not used offer earlier have defaulted even when there average income is higher than others.
- Clients who have average of ~0.13 or higher score tend to default more and thus analyzing client's social circle could help in disbursement of the loan.



Bank Loan Case Study

Analysis:

- **After analyzing the datasets, there are some factors from which bank can make its decision on granting loan.**
- **The below listed are those factors:**
- **Factors whether an applicant will be Repayer:**
  - NAME_EDUCATION_TYPE: Academic degree has less defaults.
  - NAME_INCOME_TYPE: Student and Businessmen have no defaults.
  - REGION_RATING_CLIENT: RATING 1 is safer.
  - ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%
  - DAYS_BIRTH: People above age of 50 have low probability of defaulting
  - DAYS_EMPLOYED: Clients with 40+ year experience having less than 1% default rate
  - AMT_INCOME_TOTAL: Applicant with Income more than 700,000 are less likely to default
- **Factors whether an applicant will be Defaulter:**
  - CODE_GENDER: Men are at relatively higher default rate
  - NAME_FAMILY_STATUS : People who have civil marriage or who are single default a lot.
  - NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education

Bank Loan Case Study

- **Factors whether an applicant will be Defaulter:**

  - NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot.
  - REGION_RATING_CLIENT: People who live in Rating 3 has highest defaults.
  - OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as their default rate is huge.
  - ORGANIZATION_TYPE: Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (~13.5%), Industry: type 8 (>12%) and Restaurant (less than 12%).
  - DAYS_BIRTH: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting
  - DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.
  - AMT_GOODS_PRICE: When the credit amount goes beyond 5lakhs, there is an increase in defaulters.

- **Factors that Loan can be given on Condition of High Interest rate to mitigate any default risk leading to business loss:**

  - NAME_HOUSING_TYPE: High number of loan applications are from the category of people who live in Rented apartments & living with parents and hence offering the loan would mitigate the loss if any of those default.
  - AMT_CREDIT: People who get loan for 3-6 Lakhs tend to default more than others and hence having higher interest specifically for this credit range would be ideal.

Bank Loan Case Study

- **Factors that Loan can be given on Condition of High Interest rate to mitigate any default risk leading to business loss:**

  - AMT_INCOME: Since 90% of the applications have Income total less than 3Lakhs and they have high probability of defaulting, they could be offered loan with higher interest compared to other income category.
  - NAME_CASH_LOAN_PURPOSE: Loan taken for the purpose of Repairs seems to have highest default rate. A very high number applications have been rejected by bank or refused by client in previous applications as well which has purpose as repair or other.

Conclusion:

- **More than 92% of the previously canceled client have repaid the loan. Revising the interest rates would increase business opportunity for these clients**
- **88% of the clients who have been previously refused a loan has payed back the loan in present case.**
- **Refusal reason should be recorded for further analysis as these clients could turn into potential repaying customer.**

Analyzing the Impact of Car Features on Price & profitability

Description:

The automotive industry has been rapidly evolving over the past few decades, with a growing focus on fuel efficiency, environmental sustainability, and technological innovation.

With increasing competition among manufacturers and a changing consumer landscape, it has become more important than ever to understand the factors that drive consumer demand for cars.
In recent years, there has been a growing trend towards electric and hybrid vehicles and increased interest in alternative fuel sources such as hydrogen and natural gas. At the same time, traditional gasoline-powered cars remain dominant in the market, with varying fuel types and grades available to consumers.
For the given dataset, as a Data Analyst, the client has asked How can a car manufacturer optimize pricing and product development decisions to maximize profitability while meeting consumer demand?
This problem could be approached by analyzing the relationship between a car's features, market category, and pricing, and identifying which features and categories are most popular among consumers and most profitable for the manufacturer. By using data analysis techniques such as regression analysis and market segmentation, the manufacturer could develop a pricing strategy that balances consumer demand with profitability, and identify which product features to focus on in future product development efforts. This could help the manufacturer improve its competitiveness in the market and increase its profitability over time.

Analyzing the Impact of Car Features on Price & profitability

Problem:

Tasks: Analysis

Before diving into the analysis of the given dataset, it is important to perform thorough data cleaning to ensure accurate and reliable results. You need to build an interactive dashboard in Excel from the tasks given below:

Insight Required: How does the popularity of a car model vary across different market categories?

•**Task 1.A:** Create a pivot table that shows the number of car models in each market category and their corresponding popularity scores.

•**Task 1.B:** Create a combo chart that visualizes the relationship between market category and popularity.

Insight Required: What is the relationship between a car's engine power and its price?

•**Task 2:** Create a scatter chart that plots engine power on the x-axis and price on the y-axis. Add a trendline to the chart to visualize the relationship between these variables.

Insight Required: Which car features are most important in determining a car's price?

•**Task 3:** Use regression analysis to identify the variables that have the strongest relationship with a car's price. Then create a bar chart that shows the coefficient values for each variable to visualize their relative importance.

Analyzing the Impact of Car Features on Price & profitability

Tasks: Analysis

Insight Required: How does the average price of a car vary across different manufacturers?

- **Task 4.A:** Create a pivot table that shows the average price of cars for each manufacturer.
- **Task 4.B:** Create a bar chart or a horizontal stacked bar chart that visualizes the relationship between manufacturer and average price.

Insight Required: What is the relationship between fuel efficiency and the number of cylinders in a car's engine?

- **Task 5.A:** Create a scatter plot with the number of cylinders on the x-axis and highway MPG on the y-axis. Then create a trendline on the scatter plot to visually estimate the slope of the relationship and assess its significance.
- **Task 5.B:** Calculate the correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.

Tasks: Building Dashboard

After analyzing the above tasks , need to build a dashboard which includes certain visualizations to be used.

Analyzing the Impact of Car Features on Price & profitability

Design:

Here is a brief overview of the dataset:

•

Number of observations: 11,159
•**Number of variables:** 16
•**File type:** CSV (Comma Separated Values)

A data analyst could use this dataset to gain insights into various aspects of the automotive industry, such as:

•

- **Analyzing trends in car features and pricing over time**
- **Comparing the fuel efficiency of different types of cars**
- **Investigating the relationship between a car's features and its popularity**
- **Predicting the price of a car based on its features and market category**

I have used Microsoft Excel for data cleaning, analysing and creating visualizations as needed.

For the given dataset, I have applied the EDA steps as well as used Regression analysis for predicting the manufacturer's price based on the independent variables listed.

Regression and correlation is calculated by using the Data Analysis Add-In pack in Excel which is more efficient and complex than using the LINEST function.

Analyzing the Impact of Car Features on Price & profitability

Findings:

Data Cleaning

- In the given dataset, there were duplicate rows as well as missing values.
- As listed in the below picture, the missing value percentage for the variables is less than 1% which can be ignored.
- After using the remove duplicates function in Data, 891 rows were removed.

A11916 =COUNTIF(A1:A11915,"")

	A	B	C	D	E	F	G	H	I	J
11911	Acura	ZDX	2012	premium ur	300	6	AUTOMATI(	all wheel d	4	Crossover,
11912	Acura	ZDX	2012	premium ur	300	6	AUTOMATI(	all wheel d	4	Crossover,
11913	Acura	ZDX	2012	premium ur	300	6	AUTOMATI(	all wheel d	4	Crossover,
11914	Acura	ZDX	2013	premium ur	300	6	AUTOMATI(	all wheel d	4	Crossover,
11915	Lincoln	Zephyr	2006	regular unl	221	6	AUTOMATI(	front wheel	4	Luxury
11916	0	0	0	3	69	30	0	0	6	0
11917	0	0	0	0.025178	0.579102	0.251783	0	0	0.050357	0
11918										
11919										

Descriptive Statistics

Descriptive Stats

Year		Engine HP		Engine Cylinders		Number of Doors		highway MPG		city mpg		Popularity		MSRP	
Mean	2010.384	Mean	249.3861	Mean	5.628829	Mean	3.436093	Mean	26.63749	Mean	19.73325	Mean	1554.911	Mean	40594.74
Standard Er	0.069443	Standard Er	1.003281	Standard Er	0.016333	Standard Er	0.008076	Standard Er	0.081199	Standard Er	0.082343	Standard Er	13.2097	Standard Er	550.6954
Median	2015	Median	227	Median	6	Median	4	Median	26	Median	18	Median	1385	Median	29995
Mode	2015	Mode	200	Mode	4	Mode	4	Mode	24	Mode	17	Mode	1385	Mode	2000
Standard D	7.57974	Standard D	109.1919	Standard D	1.780559	Standard D	0.881315	Standard D	8.863001	Standard D	8.987798	Standard D	1441.855	Standard D	60109.1
Sample Var	57.45246	Sample Var	11922.86	Sample Var	3.170392	Sample Var	0.776717	Sample Var	78.55278	Sample Var	80.78052	Sample Var	2078947	Sample Var	3.61E+09
Kurtosis	0.299838	Kurtosis	2.323884	Kurtosis	1.974316	Kurtosis	-1.00911	Kurtosis	181.6274	Kurtosis	74.62349	Kurtosis	2.27709	Kurtosis	268.9263
Skewness	-1.22198	Skewness	1.29123	Skewness	0.964542	Skewness	-0.96867	Skewness	7.573931	Skewness	7.106681	Skewness	1.653443	Skewness	11.77199
Range	27	Range	946	Range	16	Range	2	Range	342	Range	130	Range	5655	Range	2063902
Minimum	1990	Minimum	55	Minimum	0	Minimum	2	Minimum	12	Minimum	7	Minimum	2	Minimum	2000
Maximum	2017	Maximum	1001	Maximum	16	Maximum	4	Maximum	354	Maximum	137	Maximum	5657	Maximum	2065902
Sum	23951719	Sum	2953978	Sum	66893	Sum	40917	Sum	317359	Sum	235102	Sum	18525212	Sum	4.84E+08
Count	11914	Count	11845	Count	11884	Count	11908	Count	11914	Count	11914	Count	11914	Count	11914
Confidence	0.136119	Confidence	1.966596	Confidence	0.032016	Confidence	0.015831	Confidence	0.159164	Confidence	0.161405	Confidence	25.89316	Confidence	1079.453

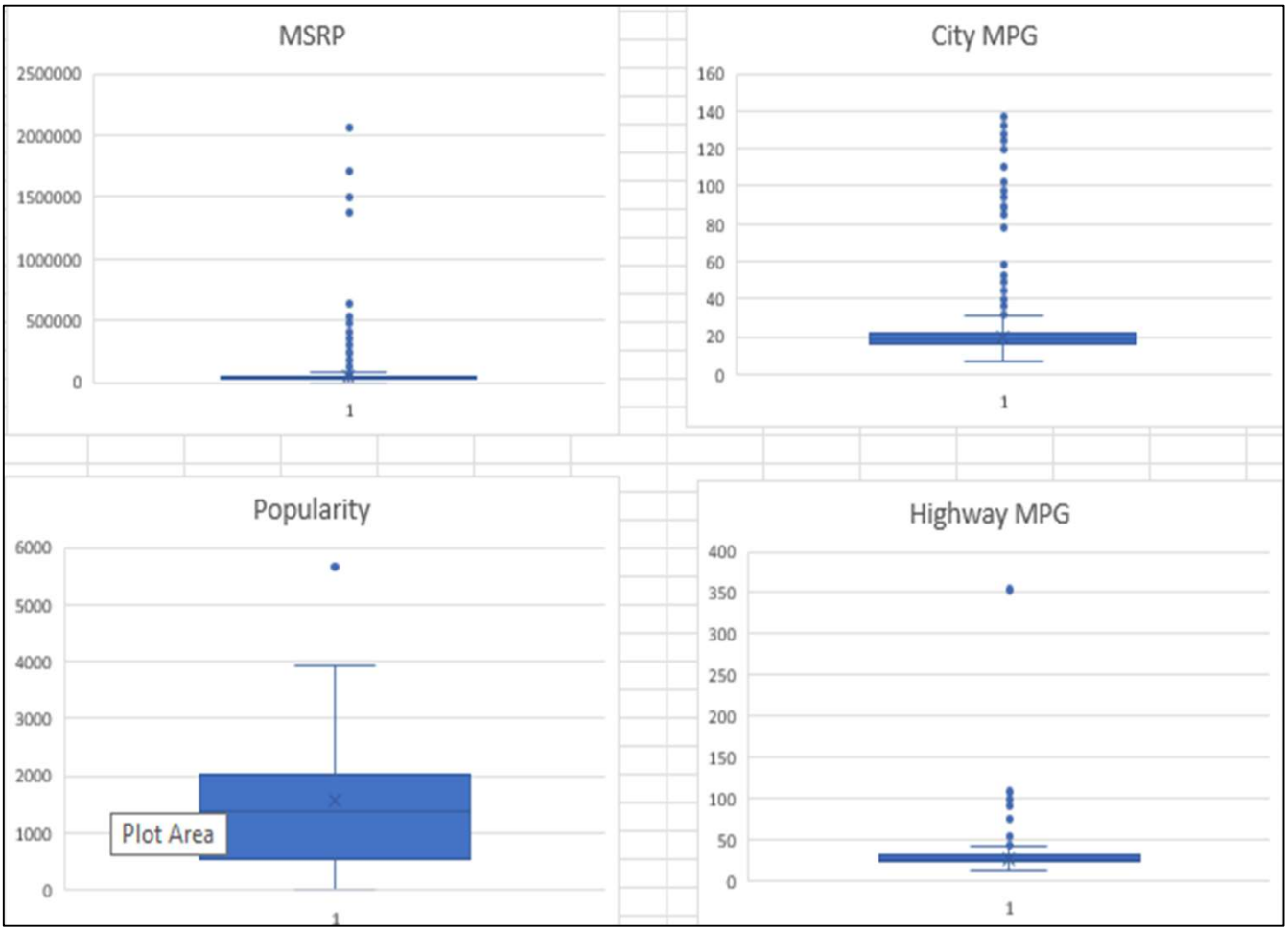
- In the given dataset, there were eight numeric variables which also includes the year of each model released up to 2017.

- Based on the descriptive statistics, there were some huge outliers in some of the variables which can be seen from the min and max values.

Analyzing the Impact of Car Features on Price & profitability

Outliers & Missing Imputation

- As checked in descriptive stats and based on the box plots, there were outliers listed
- For MSRP, through median imputation handled the outliers which is inserting the average value to outliers found.
- As the outlier percentage was low for highway and city MPG, removed the outliers in the both columns as it was less than 3%.



Analyzing the Impact of Car Features on Price & profitability

Task 1

Result: The below table shows the top 5 and bottom 5 market categories based on popularity.

- Hatchback, Flex Fuel category is the top market category whereas Exotic , Luxury category is the least preferred.

Row Labels	Count of Model	Average of Popularity
Hatchback,Flex Fuel	7	5657.00
Flex Fuel,Diesel	16	5657.00
Crossover,Flex Fuel,Performance	6	5657.00
Crossover,Luxury,Performance, Hybrid	2	3916.00
Crossover,Factory Tuner,Luxury,Performance	5	2607.40

Row Labels	Count of Model	Average of Popularity
Exotic, Luxury, High-Performance , Hybrid	1	204.00
Performance, Hybrid	1	155.00
Flex Fuel, Performance, Hybrid	2	155.00
Flex Fuel, Hybrid	2	155.00
Exotic, Luxury	12	112.67

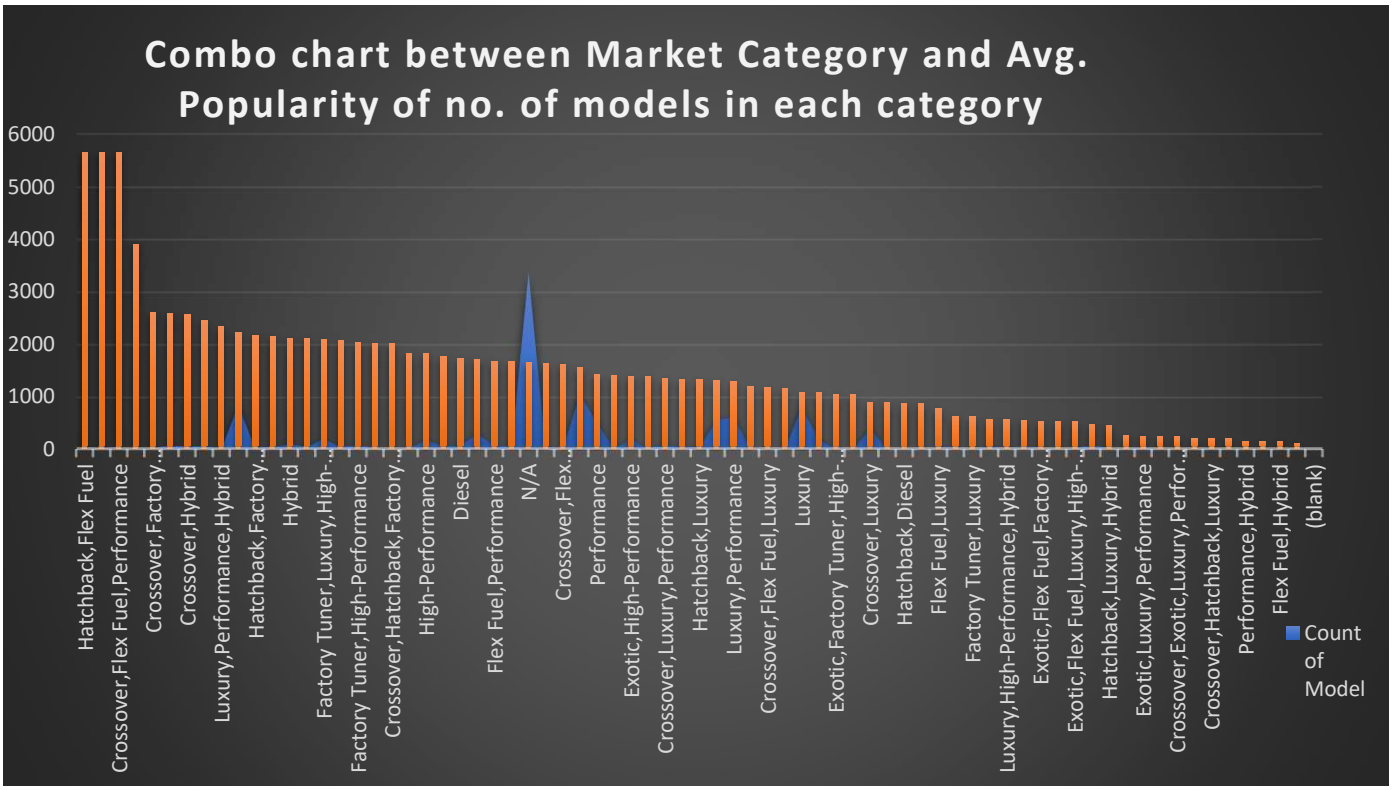


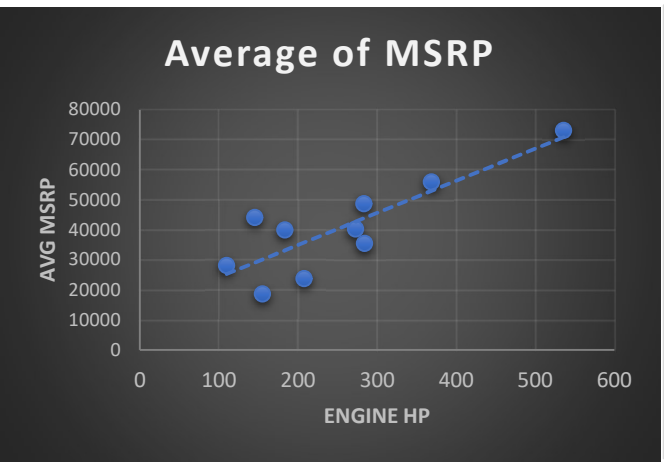
Chart: The following combo chart is between the market category and average popularity with the count of models in each category where the chart is filtered on average popularity in descending order.

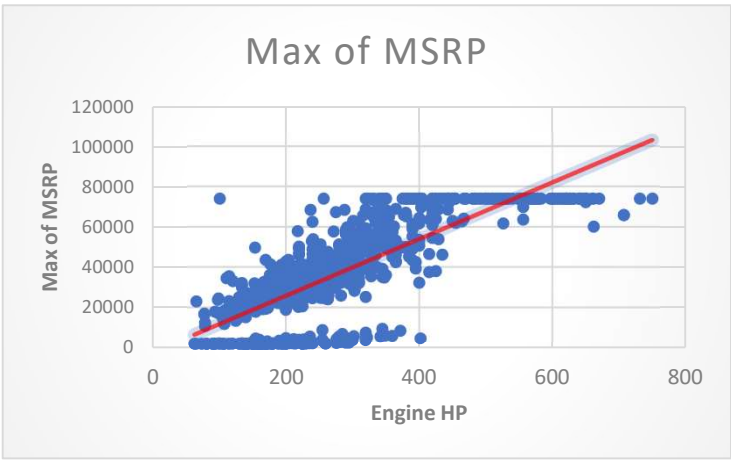
Analyzing the Impact of Car Features on Price & profitability

Task 2

Result: The below scatter plots shows the positive trend for both average and Max. of price with the engine power.

- That means, the price increases when the engine power is increased.





Task 3

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.843506044
R Square	0.711502446
Adjusted R Square	0.711370169
Standard Error	10279.0472
Observations	10911

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	2.84162E+12	5.68324E+11	5378.856135	0
Residual	10905	1.15221E+12	105658811.2		
Total	10910	3.99383E+12			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-19129.59422	949.4177092	-20.1487649	1.12206E-88	-20990.62529	-17268.56314	-20990.62529	-17268.56314
Engine HP	185.0193794	1.512804117	122.3022712	0	182.0540087	187.9847502	182.0540087	187.9847502
Engine Cylinders	-1580.989678	110.2792958	-14.33623299	3.41252E-46	-1797.157119	-1364.822237	-1797.157119	-1364.822237
highway MPG	307.5325096	41.4762121	7.414672026	1.3097E-13	226.231604	388.8334153	226.231604	388.8334153
city mpg	351.5519122	46.05175389	7.633844152	2.46888E-14	261.2821139	441.8217104	261.2821139	441.8217104
Popularity	-0.258280793	0.068126995	-3.791166688	0.00015075	-0.391822071	-0.124739515	-0.391822071	-0.124739515

Analyzing the Impact of Car Features on Price & profitability

Task 3

Result:

- This is the result of regression analysis for MSRP with Engine HP, Cylinders, highway & city MPG and Popularity as independent variables.
- Before using the regression in Data analysis pack, there were some blank values in few variables which was less than 0.5%.
- Hence, removed those rows and performed the analysis.
- Here, we can see that the engine HP , Highway and city MPG variables effect in increasing the manufacturer's price

The following bar chart shows the effect of price based on each variables both in positive and negative trend . (Engine HP, highway & city MPG in positive)

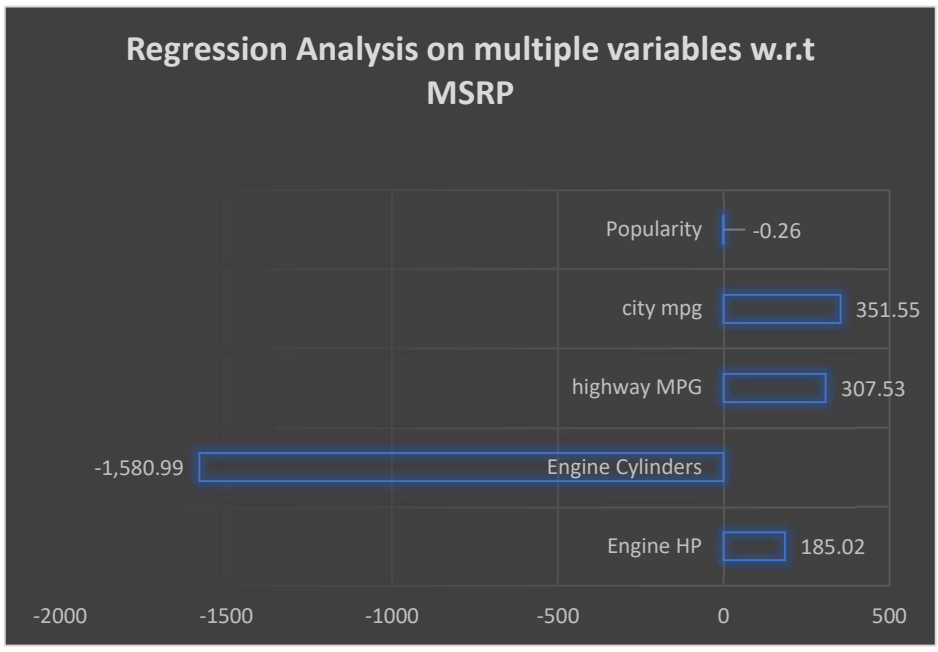
Regression Eq.:

$$Y = a0+a1*X1+a2*X2+.....+an*Xn$$

Engine HP	Engine Cylinders	highway MPG	city mpg	Popularity	MSRP
340	8	18	10	600	40025.213

- The below table shows the predicted MSRP from the above equation.

Adjusted R Square: 0.711370169

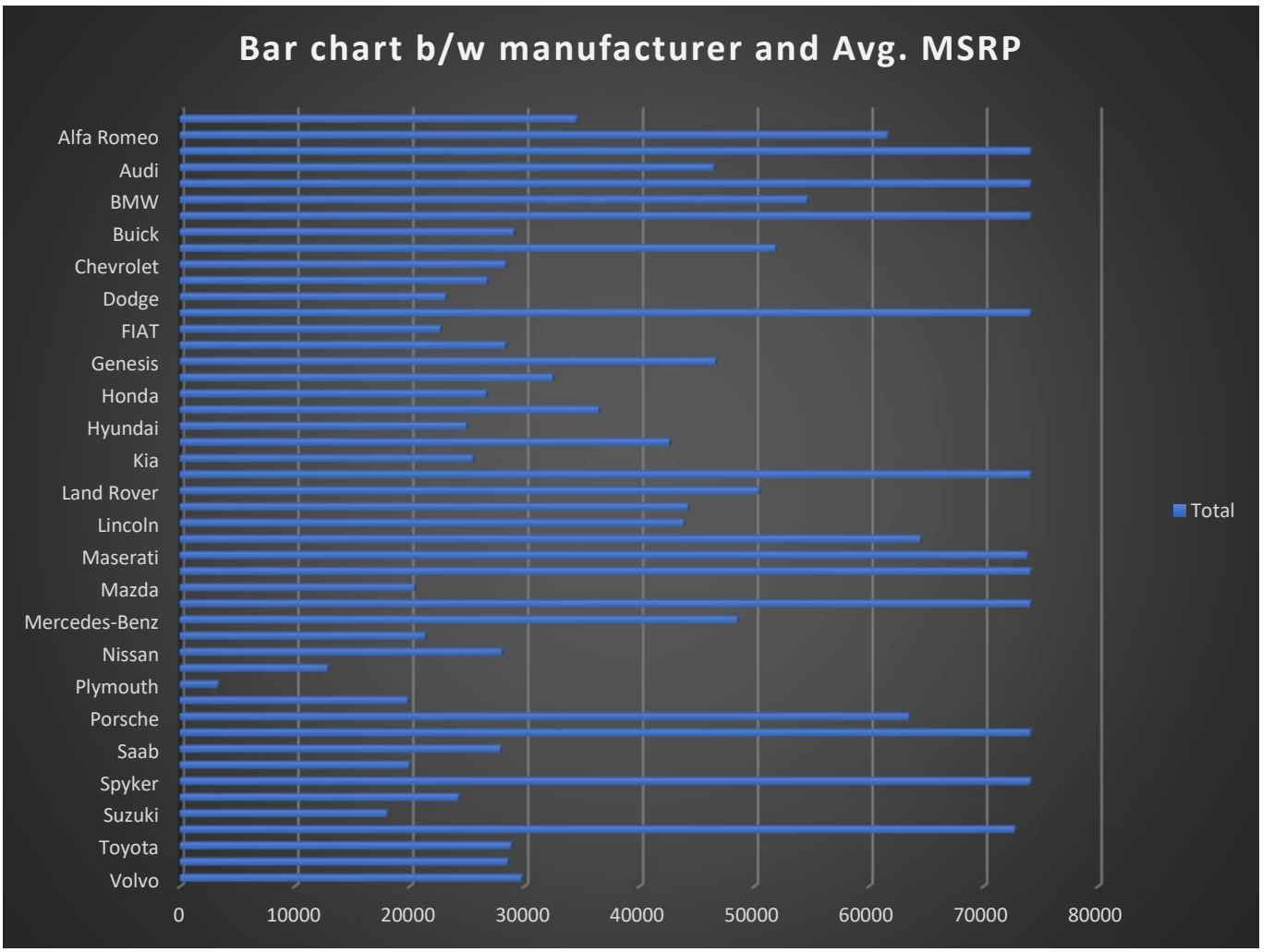


Analyzing the Impact of Car Features on Price & profitability

Task 4

Result: The bar chart shows the relationship b/w make and average MSRP.

Top 5 Manufacturers are**: Bugatti, Lamborghini, Aston Martin, Spyker and Ferrari** where as bottom 5 are **Scion, Pontiac, Suzuki, Oldsmobile and Plymouth.**



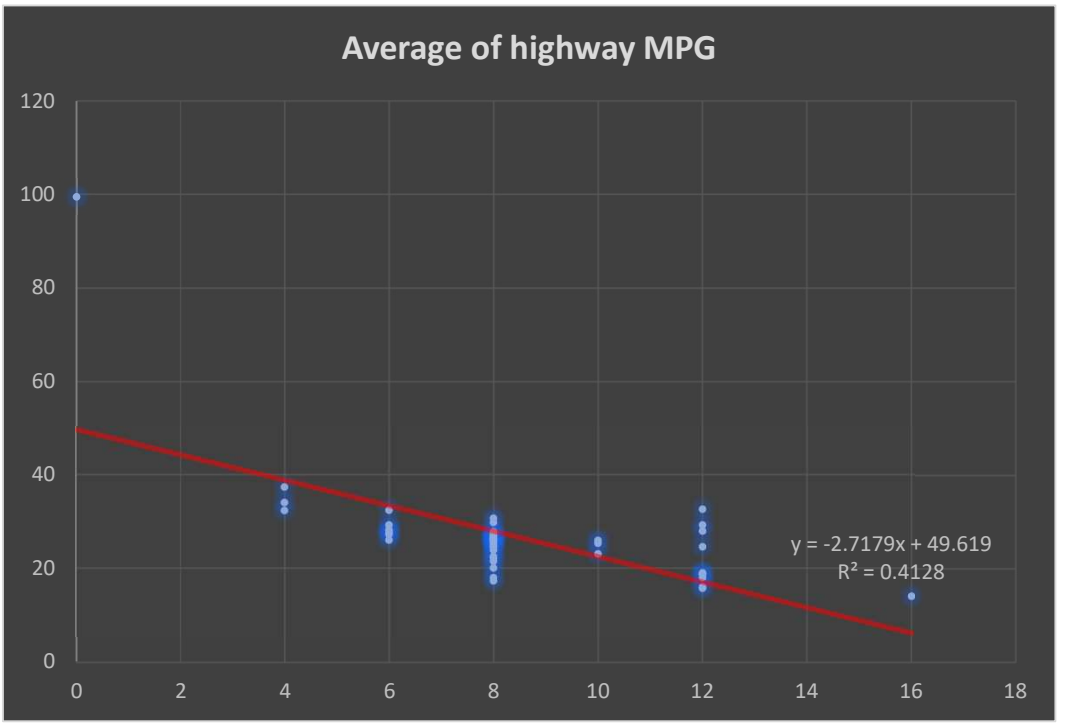
Analyzing the Impact of Car Features on Price & profitability

Task 5

Result:

The scatter plot shows the negative trend for highway MPG with no. of cylinders and the slope is -2.7179.

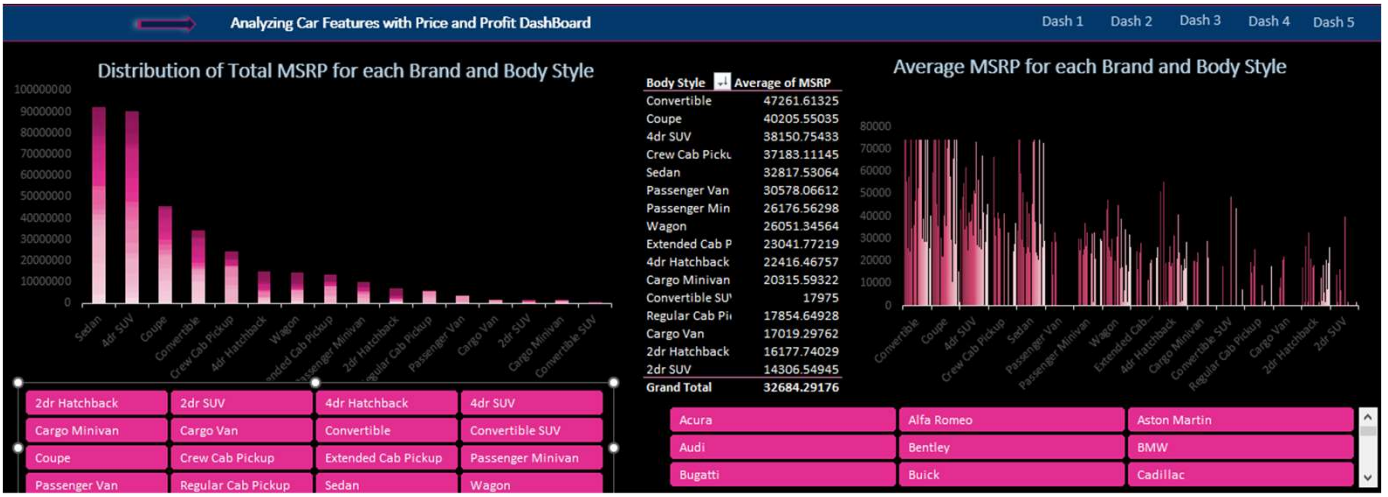
- That means, as the no. of cylinders increases the fuel efficiency on highway decreases.
- The below table is the correlation between no. of cylinders and highway MPG which is negative 0.65.
- That means, both are strongly correlated in negative trend which is also shown in scatter plot.



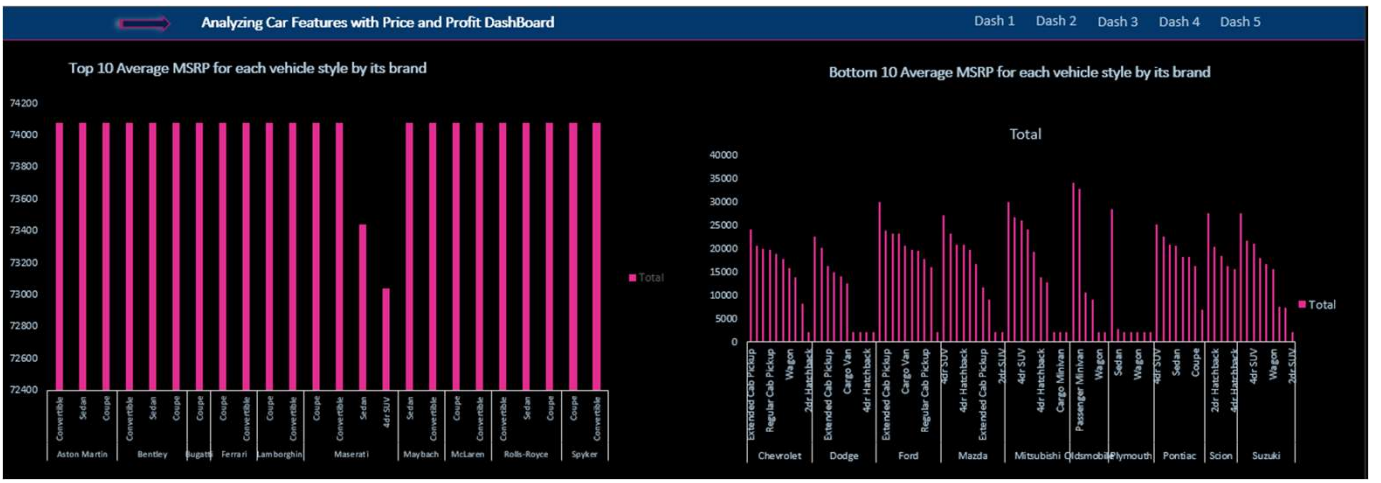
	Engine Cylinders	highway MPG
Engine Cylinders	1	-0.65851821
highway MPG	-0.65851821	1

Analyzing the Impact of Car Features on Price & profitability

Building Dashboard:

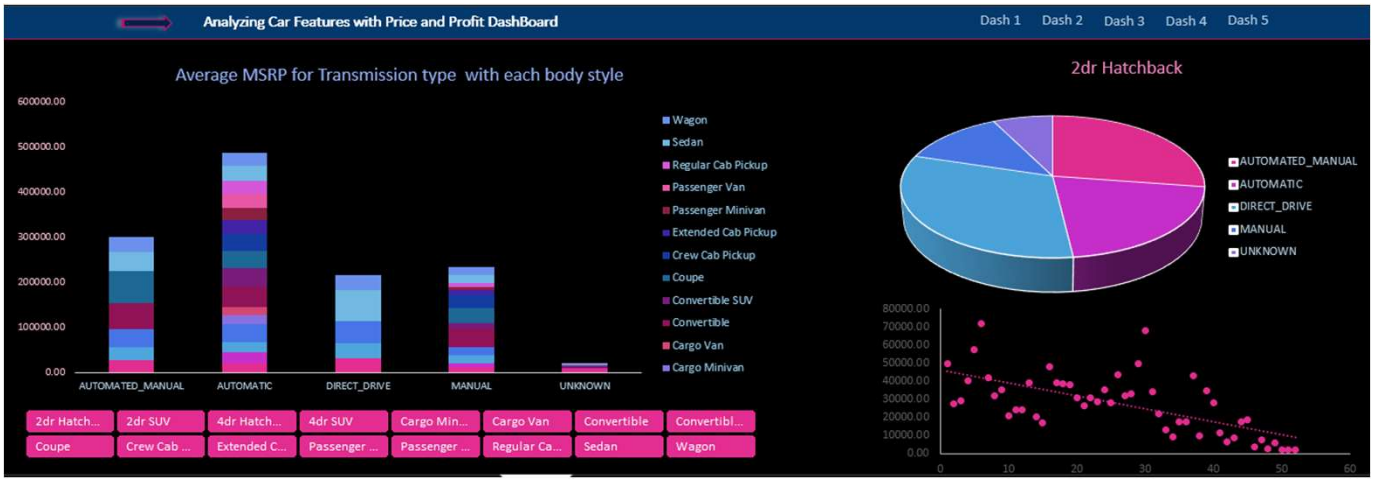


- The above dashboard shows the Total MRSP and Sum MRSP with each brand and body style and is interactive with the help of slicers inserted.
- The top brand with high total MRSP is **Chevrolet** whereas lowest is **Genesis**.

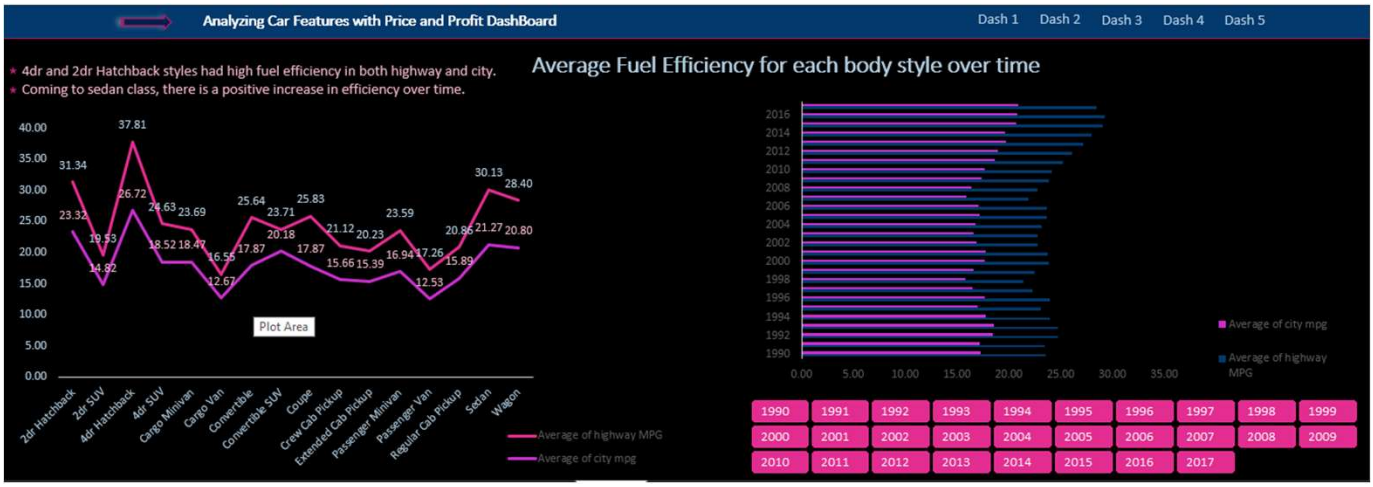


- The above dashboard shows the top 10 and bottom 10 brands list in each body style with average MSRP.
- The Top 5 brands are **Aston Martin, Bentley, Bugatti, Ferrari and Lamborghini** whereas bottom 5 are **Chevrolet, Dodge, Ford, Mazda and Mitsubishi.**

Analyzing the Impact of Car Features on Price & profitability



- The **Automatic** transmission type has the highest average MRSP and the pie chart varies based on the body style selected to show the transmission type ratio.



- **4dr** and **2dr Hatchback** had high fuel efficiency in both highway and city MPG.

- **Sedan** style has a positive increase in efficiency over time from 1990-2017.

Analyzing the Impact of Car Features on Price & profitability

- **Aston Martin**, **Bentley**, **Ferrari**, **Bugatti** and **Lamborghini** have high average MSRP and **Lamborghini** has the highest average engine horsepower.
- **FIAT** has the highest average city MPG where **Tesla** has the highest in highway MPG.

Analysis:

- Have gained adequate knowledge in regression & variable relations along with pivot table understanding by analysing this project.
- Based on the data, the adjusted R^2 was not that high but might improve if accurate data is used for the independent variables.

Conclusion:

- We have used regression analysis and have predicted the selling price of the car based on various features of the cars, including the present price of the cars.
- Depending on other Features and other classification techniques the predicted price may vary a bit but will not deviate too much from the regression analysis used.

ABC Call Volume Trend Analysis

Description:

A customer experience (CX) team consists of professionals who analyze customer feedback and data, and share insights with the rest of the organization. Typically, these teams fulfil various roles and responsibilities such as: Customer experience programs (CX programs), Digital customer experience, Design and processes, Internal communications, Voice of the customer (VoC), User experiences, Customer experience management, Journey mapping, Nurturing customer interactions, Customer success, Customer support, Handling customer data, Learning about the customer journey.

Let's look at some of the most impactful AI-empowered customer experience tools you can use today: Interactive Voice Response (IVR), Robotic Process Automation (RPA), Predictive Analytics, Intelligent Routing In a Customer Experience team there is a huge employment opportunities for Customer service representatives. Some of the roles for them include: Email support, Inbound support, Outbound support, social media support.

Inbound customer support is defined as the call center which is responsible for handling inbound calls of customers. Inbound calls are the incoming voice calls of the existing customers or prospective customers for your business which are attended by customer care representatives. Inbound customer service is the methodology of attracting, engaging, and delighting your customers to turn them into your business' loyal advocates. By solving your customers' problems and helping them achieve success using your product or service, you can delight your customers and turn them into a growth engine for your business.

ABC Call Volume Trend Analysis

Problem:

a.Calculate the average call time duration for all incoming calls received by agents (in each Time Bucket).
b.Show the total volume/ number of calls coming in via charts/ graphs [Number of calls v/s Time].
c.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%. 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%.

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.

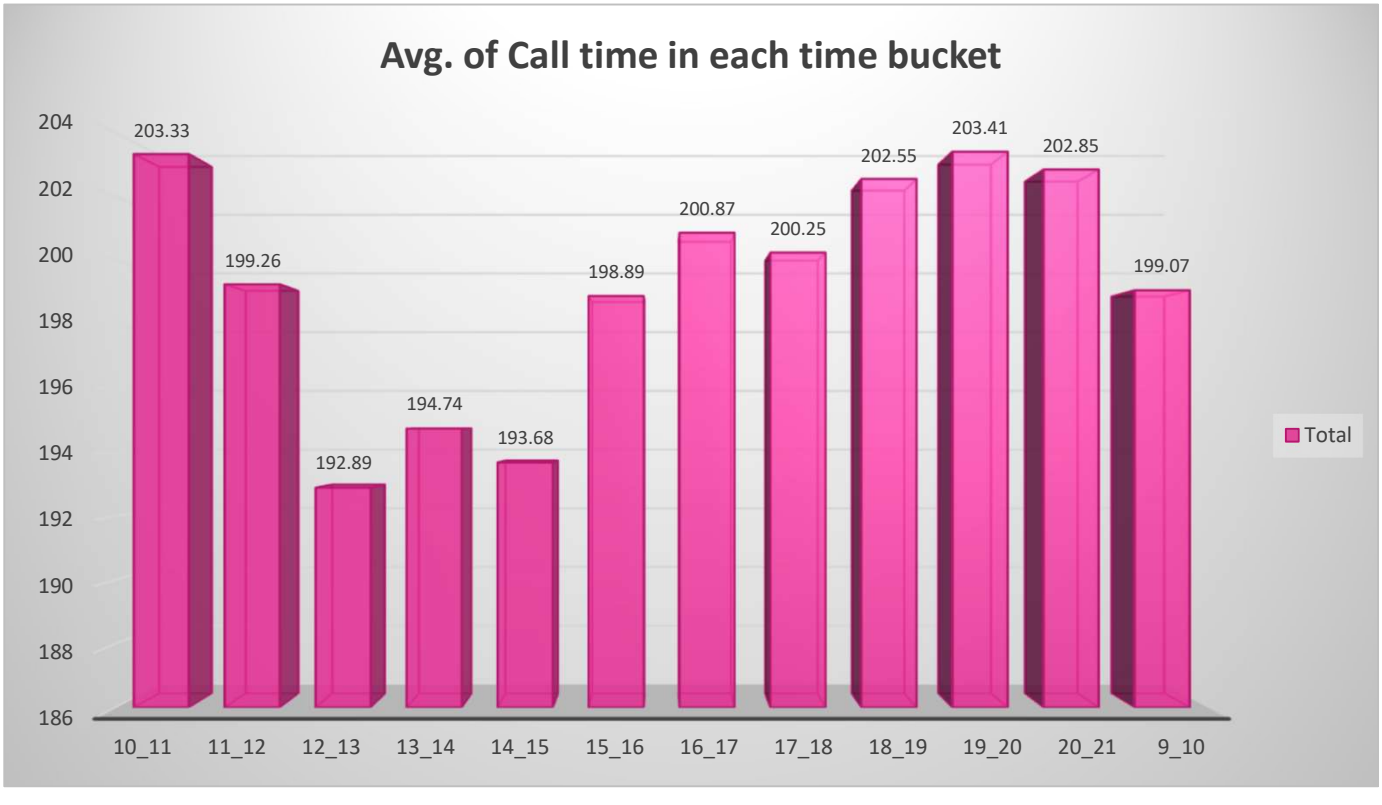
ABC Call Volume Trend Analysis

Design:

- For this project, I have used Microsoft Excel for analyzing and answering the queries.
- Here, using excel was helpful and easy as there were no complex functions to be handled and the queries were answered in simple and clear manner with required functions in excel.
- Used pivot Tables to handle the queries and while performing the Data cleaning, only one variable has around 40% of missing values which is not an important variable. Hence, ignored.

Findings:

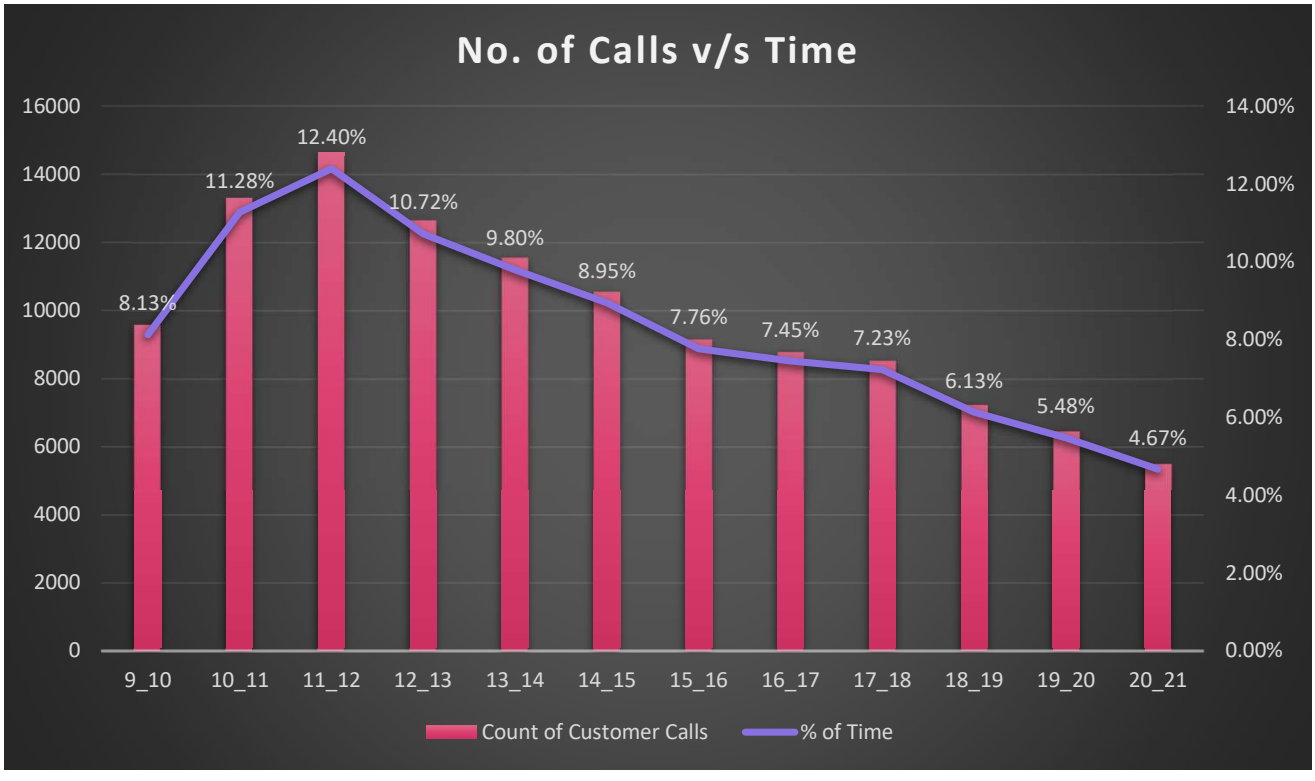
Query 1



- Here, the total average call time which was answered by the agents is 198.62 seconds.
- Average call time is highest between 7pm and 8pm while the least is between 12pm and 1pm

ABC Call Volume Trend Analysis

Query 2



- The above combo chart showcases the total volume of calls in each time bucket with percentage of calls in each time bucket.
- The highest customer calls is recorded in between 11 am and 12 pm whereas the least flow was seen in 8 pm and 9pm time range.

Assumption

Total days	30
No. of weeks	4.29
Work Days of agent in week	6
Total agent work days in month	25.71429
Total unplanned leaves per month	4
Total work days excluding leaves	21.71429
Total working Hrs per day	9
Lunch & Snacks	1.5
Call time per day (60% of rem.)	4.5

ABC Call Volume Trend Analysis

Query 3

Time Bucket	% of Time	Req. Agents
9_10	8.13%	5
10_11	11.28%	6
11_12	12.40%	7
12_13	10.72%	6
13_14	9.80%	6
14_15	8.95%	5
15_16	7.76%	4
16_17	7.45%	4
17_18	7.23%	4
18_19	6.13%	3
19_20	5.48%	3
20_21	4.67%	3
Total	100.00%	57

The above table shows the required agents in each time bucket to get 90% answered status.

From the pivot table we can see the percentage of each call status where abandon rate is nearly 30% as listed in query and 1% is transferred and 70% is answered.

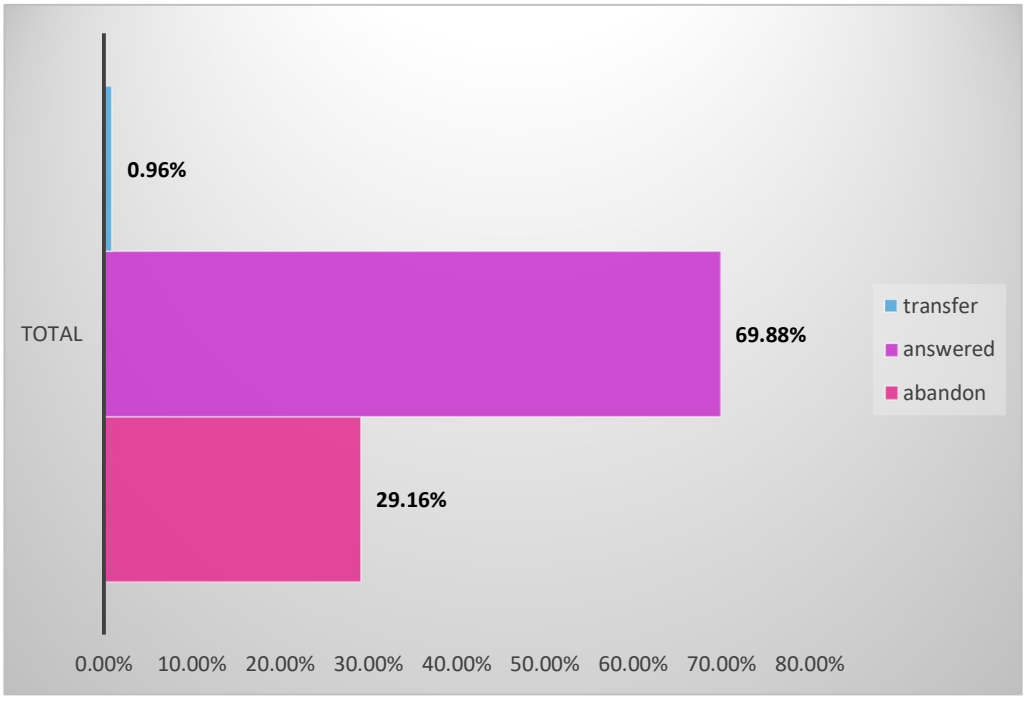
Date	abandon	answered	transfer	Grand Total
01-Jan	684	3883	77	4644
02-Jan	356	2935	60	3351
03-Jan	599	4079	111	4789
04-Jan	595	4404	114	5113
05-Jan	536	4140	114	4790
06-Jan	991	3875	85	4951
07-Jan	1319	3587	42	4948
08-Jan	1103	3519	50	4672
09-Jan	962	2628	62	3652
10-Jan	1212	3699	72	4983
11-Jan	856	3695	86	4637
12-Jan	1299	3297	47	4643
13-Jan	738	3326	59	4123
14-Jan	291	2832	32	3155
15-Jan	304	2730	24	3058
16-Jan	1191	3910	41	5142
17-Jan	16636	5706	5	22347
18-Jan	1738	4024	12	5774
19-Jan	974	3717	12	4703
20-Jan	833	3485	4	4322
21-Jan	566	3104	5	3675
22-Jan	239	3045	7	3291
23-Jan	381	2832	12	3225
	1495.783	3584.87	49.26087	5129.91304
	29%	70%	1%	

ABC Call Volume Trend Analysis

Query 3

Time taken to answer call(avg.)	198.62	
Time required for answering 90% calls (secs)	254.725832	
Total agents required per day	57	(call time 4.5 hrs)

- As discussed in the previous slide, the required agents were distributed in each time bucket based on the percentage of calls received in the respective hours.
- For getting the total number of agents required to get 90% of answered rate, I have calculated the total call hours per day with 90% answered average which in turn was obtained by taking the total average of call data and multiplying it with 0.9(90%) and dividing with the seconds in hour(3600secs).
- The bar chart on the below shows the percentage of each call status.



ABC Call Volume Trend Analysis

Query 4

Calls made during day(9am-9pm)	5130
30% support during night (9pm-9am)	1539
Additional hrs needed	76.419045
Additional agents per day	17
Total HC	74

Night Time Bucket	Calls Dist.	Time Dist.	Req. Agents
21_22	3	10.00%	2
22_23	3	10.00%	2
23_00	2	6.67%	1
00_1	2	6.67%	1
1_2	1	3.33%	1
2_3	1	3.33%	1
3_4	1	3.33%	1
4_5	1	3.33%	1
5_6	3	10.00%	2
6_7	4	13.33%	2
7_8	4	13.33%	2
8_9	5	16.67%	3
Total	30	100.00%	17

- First, I calculated the total average of calls in night time(9pm-9am) with 30% as listed by which obtained the additional hours needed.
- With this, the no. of agents required was 17(got by dividing with 4.5 hrs(call time of each agent)).
- Based on the given call distribution in each time bucket, we got the time distribution percentage and by multiplying this with the additional agent count(i.e. 17), the required agents in each time bucket is obtained.

ABC Call Volume Trend Analysis

Analysis:

Why is it that the time bucket 11_12 has the highest number of incoming calls but it does not have the highest number of average answered calls?

- Maybe there were more number of incoming calls in the time bucket 11_12 but there were not enough personnel to handle most of the queries.

Why is proportion if the monthly transfer rate is less than compared to monthly answered and abandon rate?

- In most of the customer service centers they have the dedicated toll free number to be used by customer, also there are skilled people at the call center who are well versed with the problems they come across while handling and guiding thousands of customers on daily basis.

Conclusion:

- From the above query results, we can see that customer call percentage reduces in the evening.
- The additional agents required in the night time zone is 17(while considering the abandon rate is 10%).
- If the call time for each agent can be increased by half an hour or one hour in their total working time(7.5 hrs), then the number of agents can be reduced a bit as well as the call status percentage for answering can be increased.
- I got to know how analysts play crucial role in the customer service industry.
- As the time bucket column was added to the dataset, pulling data for queries made easy.

Appendix

Data Analytic Process

- **Google Drive Link for the PDF file**

  https://drive.google.com/file/d/1_M1Pp_B6sMsrvCO7KNMevbD4UUFzNqWs

Instagram User Analysis

- **Google Drive Link for the PDF file**

  https://drive.google.com/file/d/1bm1zgN7FvUHPW14IX2UTA8nvmUdDHz5F

- **Google Drive Link for the SQL file**

  https://drive.google.com/drive/folders/1y1myd34xJhaySS6OgRYx1BRPwDz8NK-P

Operational Analysis & Investigating Metric Spike

- **Google Drive Link for the PDF file**

  https://drive.google.com/file/d/1J6DnRUiSMc3xHmjwbrb6PNGlPbxA2FbI

- **Google Drive Link for the SQL file**

  https://drive.google.com/drive/folders/1pSb0M9dsyeUOMFIylwE5orwO_O1aErnb

Hiring Process Analytics

- **Google Drive Link for the Excel and PDF file**

  https://drive.google.com/drive/folders/1GkhRKUOJb9UfFrZCLQWJ-k-5svJqFqvt

Appendix

IMDB Movie Analysis

- **Google Drive Link for the Excel and PDF file**

  https://drive.google.com/drive/folders/1C1FX0QMy9oEsbEJzKrVQOYMi5p444T3A

Bank Loan Case Study

- **Google Drive Link for the Python(ipynb) & PDF file**

  https://drive.google.com/drive/folders/1qI6pyyuvafviZLn6bCkiCwmFaIH-vDPX

Impact of Car Features on Price & Profitability

- **Google Drive Link for the PDF file**

  https://drive.google.com/drive/folders/1zjVK_Xlh35gaHwrM8-mVx3IsbIIhSw4V

- **Google Drive Link for the Excel files**

  https://drive.google.com/drive/folders/1YnKNHkHC5up54A58-3lJk7BTWcncMCL_

ABC Call Volume Trend Analysis

- **Google Drive Link for the Excel and PDF file**

  https://drive.google.com/drive/folders/1ET2m94e4FEW4aXZxlHYg-lb1hKuaVXBz