Data Analysis Portfolio

Prepared By:- Aniket Ubale

Email:- aniketubale1433@gmail.com

Contact :- +919370272580

Address:- Chh. Sambhajinagar, Maharashtra

Linked In :- https://www.linkedin.com/in/aniket-ubale-0a4315219/

Professional Background

I am a final year BTech student with a current CGPA of 7.90 (till 5th sem) and a diverse range of skills including expertise in Artificial intelligence, Machine Learning, Data Analysis, Python, and Databases.

Throughout my academic career, I have gained practical experience in machine learning algorithm, python programming and data analysis .

I have also worked on several projects related to data science, Artificial intelligence, machine learning, predictive analytics, deep learning.

Driven by my passion for data analysis, I made a successful transition to my current role as a Data Analyst Trainee at Trainity. My creative mindset and analytical skills enable me to provide valuable insights that drive informed business decisions. I have quickly continued to develop my skills in this exciting field.

As a fresh graduate, I am eager to tackle the challenges of the corporate world and gain hands-on experience. I am highly flexible and adaptive, with a strong desire to learn and apply my theoretical knowledge practically. I am confident that with my dedication and hard work, I will be able to make valuable contributions to any team.

Index Of Content

1. What is Data Analysis?

Data analysis refers to the process of examining, cleaning, transforming, and modeling data in order to draw useful insights and conclusions. The aim of data analysis is to uncover patterns, trends, and relationships within data sets to inform decision-making.

Data analysis involves using various techniques and tools to organize, summarize, and interpret large sets of data. This can include statistical analysis, data mining, and machine learning. Some common methods used in data analysis include descriptive statistics, regression analysis, hypothesis testing, and clustering.

Data analysis is a crucial part of many fields, including business, finance, healthcare, and social sciences. It helps organizations and individuals make informed decisions by providing them with relevant information and insights. Effective data analysis requires a combination of technical skills, critical thinking, and domain knowledge.

2. Data Analysis Process

trainity

Data Analytics Process: Real World Application

Shopping & Use of 6 Step Data Analytics Process

Here is a real-life scenario in which data analytics can be applied using the "Plan, Prepare, Process, Analyze, Share, Act" process:

Scenario:

The real world example in fantasy games, where a fantasy games companies wants to reduce its customer churn rate by analyzing its customer data.

Plan:

In the planning stage, we define the problem we want to solve, set goals, what is output and identify the resources needed.

For this example, the fantasy games companies wants to reduce user churn rate by analyzing user data.

the objective is to identify the reasons for churn and the actions needed to retain user.

The resources required includes-

statistics, data analysts, data scientists, machine learning, python, database, visulazation tools like excel ,charts ,power BI etc

Prepare:

In the preparation stage, we collect, clean, and preprocess the data.

The fantasy games companies collects customer data.

Mostly data in the structured data format so that data can be easily analyze.

The data is then cleaned and preprocessed by removing duplicates, filling missing values, and standardizing data.

Handling a missing and null value is important in data analysis so that we can get better results that we want.

Process:

In the processing stage, we transform the data into a format suitable for analysis.

In this example, we transform the data by creating features, such as user tenure, transaction history, and user interactions. We also segment user based on their demographics and behavior.

Whether the use are playing daily or not ,how many games they are playing ,how many reward they win etc condition we analyze

Analyze:

In the analysis stage, we apply statistical ,machine learning and deep learning techniques to the data to uncover patterns and insights.

In this example, we can use techniques or algorithm such as regression analysis, decision trees, random forest to identify the factors that influence customer churn.

Mostly we analyze data that such that can we give some bonus so user can play games continuously so our business can increase.

We also use clustering algorithms to identify groups of user with similar characteristics and behavior.

Share:

In the sharing stage, we communicate the results of our analysis to stakeholders.

In this example, we share the insights gained from our analysis with managers and user service teams.

We can share our ideas to our collegues so that they can also understood what we need to do in future

We also create data visualizations, such as dashboards and reports, to communicate the insights effectively.

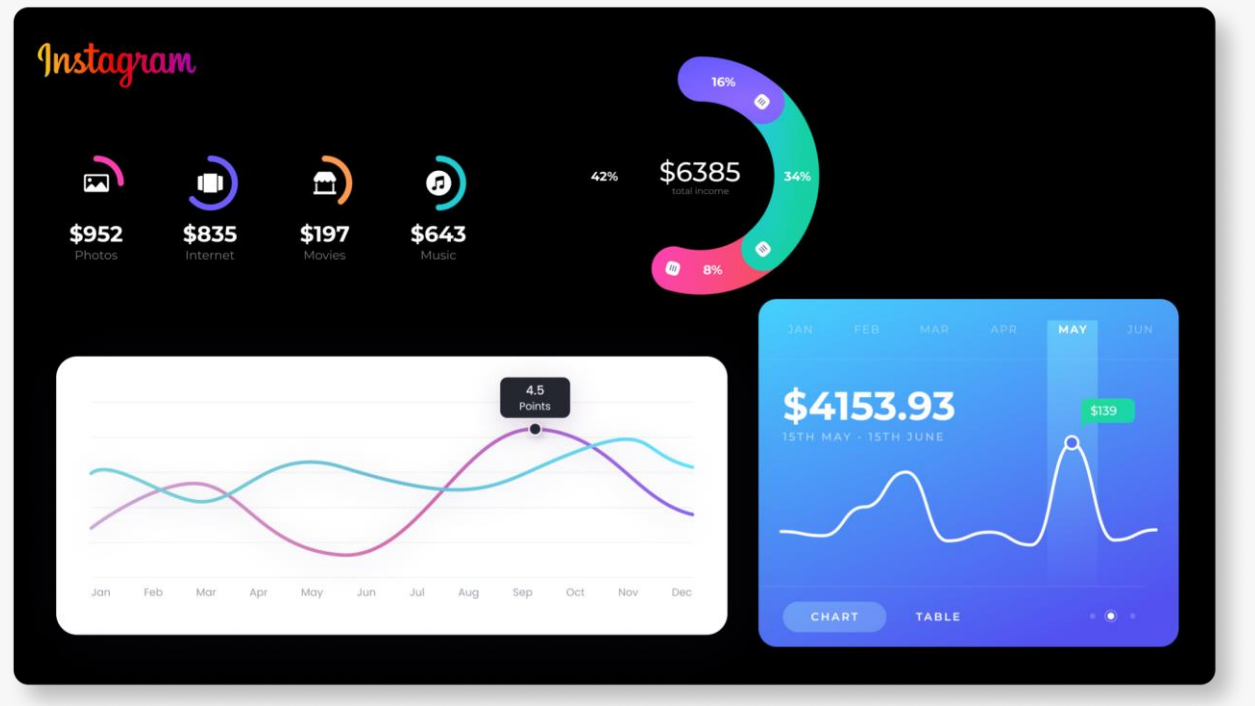
Act:

In the stage, we take action based on the insights gained from our analysis.

In this example, we implement targeted marketing campaigns to retain user identified as being at high risk of churning.

We also improve customer service by addressing the issues that led to customer churn, such as poor customer experience or unmet needs.

3. Instagram User Analytics



This project was undertaken to provide insights and data-driven recommendations to the Instagram product and marketing teams. The specific questions we sought to answer were:

- Who are the oldest users on the platform?
- Who has never posted a photo on Instagram?

- Who won the contest for the most likes on a single photo?
- What are the top 5 most commonly used hashtags on the platform?
- What day of the week do most users register on Instagram?
- What is the average number of posts per user on Instagram?
- How many bots or fake accounts are on the platform?

To answer these questions, we used SQL to perform queries on a database of Instagram users and their activity on the platform.

Approach:

Our approach to this project was as follows:

1. Set up the database and run the necessary SQL commands to create the tables and import the data.
2. Write and test SQL queries to extract the relevant data from the database.
3. Use statistical and visualization tools to analyze the data and identify trends and patterns.
4. Write a report to present our findings and recommendations to the leadership team.

Tech Stack Used:

The following tools and technologies were used in this project:

- SQL: We used SQL to query the database and extract the necessary data.
- Google Drive: We used Google Drive to store and share our report with the leadership team.

Insights

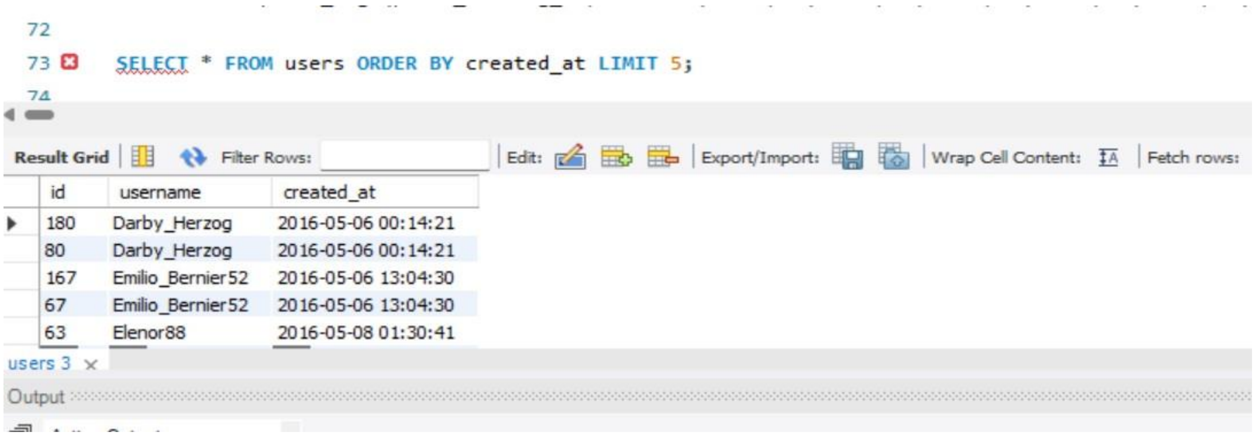
Some of the key insights we gained from our analysis include:

- The oldest users on the platform have been using Instagram for over 5 years.
- There are several users who have never posted a photo on Instagram.
- The winner of the contest for the most likes on a single photo is a user who has been on the platform for just over 4 years.
- The top 5 most commonly used hashtags on the platform are all related to party, and enjoyment.
- Most users register on Instagram on Thursday.
- The average number of posts per user on Instagram is just under 4.
- There are 13% of bots or fake accounts on the platform, as identified by users who have liked every single photo on the site.

Results

(A)Marketing: The marketing team wants to launch some campaigns, and they need your help with the following

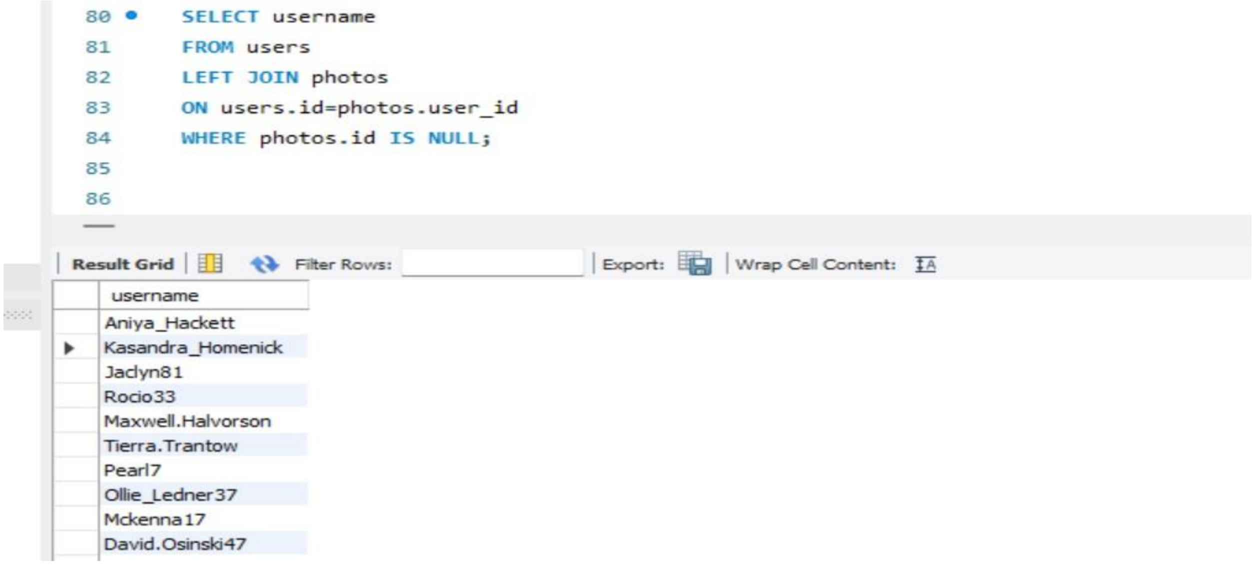
[1] Rewarding Most Loyal Users: People who have been using the platform for the longest time.
Your Task: Find the 5 oldest users of the Instagram from the database provided



```sql
SELECT * FROM users ORDER BY created_at LIMIT 5;
```

id	username	created_at
180	Darby_Herzog	2016-05-06 00:14:21
80	Darby_Herzog	2016-05-06 00:14:21
167	Emilio_Bernier52	2016-05-06 13:04:30
67	Emilio_Bernier52	2016-05-06 13:04:30
63	Elenor88	2016-05-08 01:30:41

[2] Remind Inactive Users to Start Posting: By sending them promotional emails to post their 1st photo.
Your Task: Find the users who have never posted a single photo on Instagram

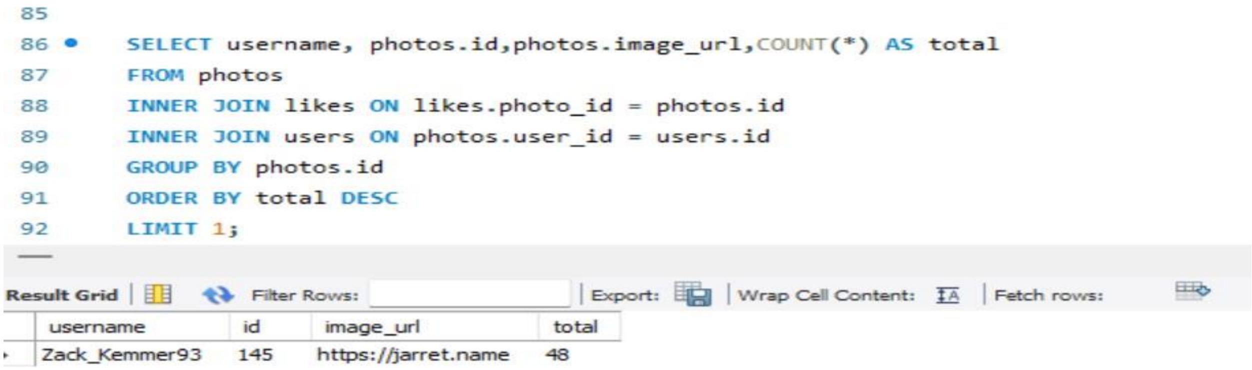


```sql
SELECT username
FROM users
LEFT JOIN photos
ON users.id=photos.user_id
WHERE photos.id IS NULL;
```

username
Aniya_Hackett
Kasandra_Homenick
Jaclyn81
Rocio33
Maxwell.Halvorson
Tierra.Trantow
Pearl7
Ollie_Ledner37
Mckenna17
David.Osinski47

[3] Declaring Contest Winner: The team started a contest and the user who gets the most likes on a single photo will win the contest now they wish to declare the winner.

Your Task: Identify the winner of the contest and provide their details to the team

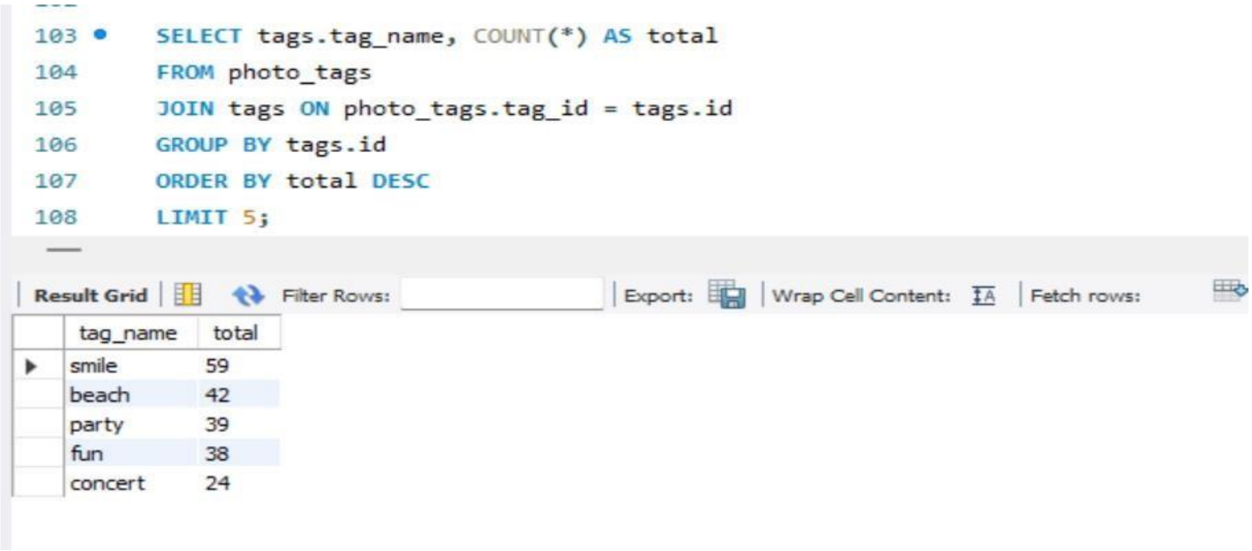


```sql
SELECT username, photos.id,photos.image_url,COUNT(*) AS total
FROM photos
INNER JOIN likes ON likes.photo_id = photos.id
INNER JOIN users ON photos.user_id = users.id
GROUP BY photos.id
ORDER BY total DESC
LIMIT 1;
```

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

[4] Hashtag Researching: A partner brand wants to know which hashtags to use in the post to reach the most people on the platform.

Your Task: Identify and suggest the top 5 most commonly used hashtags on the platform

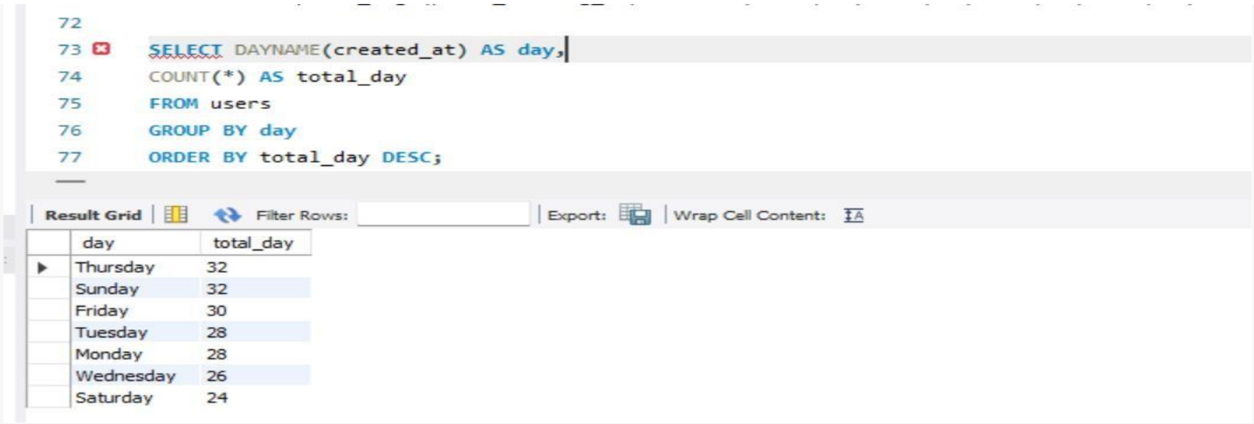


```sql
SELECT tags.tag_name, COUNT(*) AS total
FROM photo_tags
JOIN tags ON photo_tags.tag_id = tags.id
GROUP BY tags.id
ORDER BY total DESC
LIMIT 5;
```

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

[5] Launch AD Campaign: The team wants to know which day would be the best day to launch ADs.

Your Task: What day of the week do most users register on? Provide insights on when to schedule an ad campaign

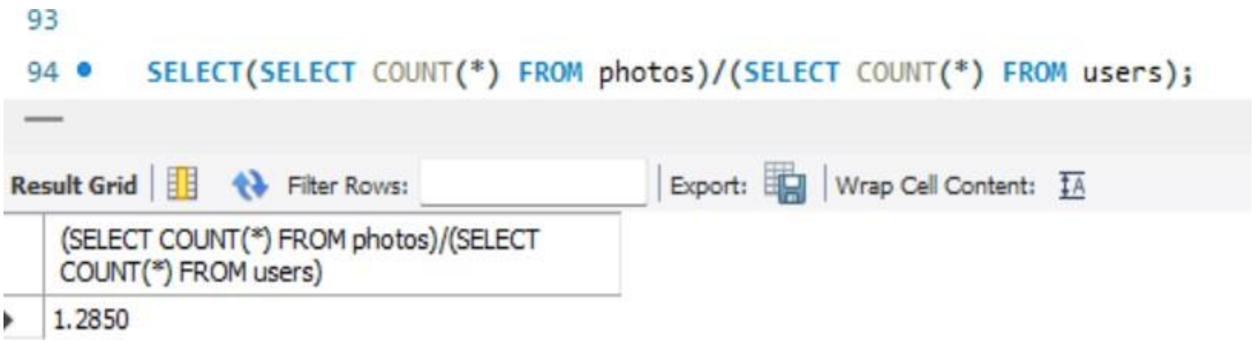


```
SELECT DAYNAME(created_at) AS day,
COUNT(*) AS total_day
FROM users
GROUP BY day
ORDER BY total_day DESC;
```

day	total_day
Thursday	32
Sunday	32
Friday	30
Tuesday	28
Monday	28
Wednesday	26
Saturday	24

(B)Investor Metrics: Our investors want to know if Instagram is performing well and is not becoming redundant like Facebook, they want to assess the app on the following grounds

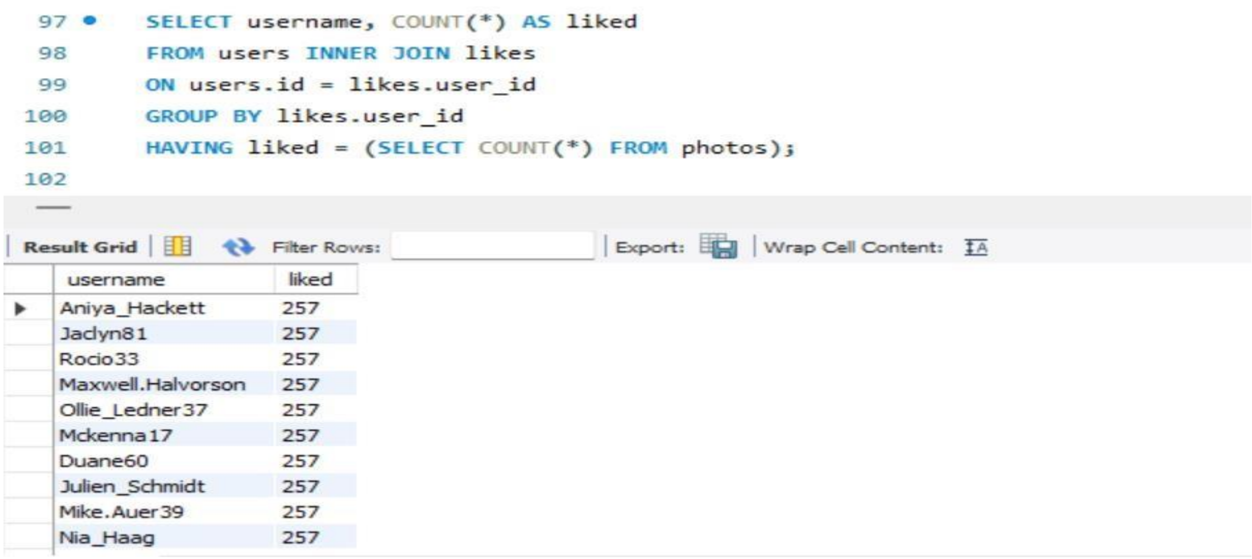
[1] User Engagement: Are users still as active and post on Instagram or they are making fewer posts
Your Task: Provide how many times an average user posts on Instagram. Also, provide the total number of photos on Instagram/total number of users



```
SELECT(SELECT COUNT(*) FROM photos)/(SELECT COUNT(*) FROM users);
```

(SELECT COUNT(*) FROM photos)/(SELECT COUNT(*) FROM users)
1.2850

[2]Bots & Fake Accounts: The investors want to know if the platform is crowded with fake and dummy accounts
Your Task: Provide data on users (bots) who have liked every single photo on the site (since any normal user would not be able to do this).



```
SELECT username, COUNT(*) AS liked
FROM users INNER JOIN likes
ON users.id = likes.user_id
GROUP BY likes.user_id
HAVING liked = (SELECT COUNT(*) FROM photos);
```

username	liked
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

Conclusion

- Overall, our analysis has provided valuable insights into the usage and engagement of Instagram users. These insights can be used by the product and marketing teams to inform business decisions and optimize the user experience on the platform. This knowledge has helped me better understand the sql language used in data analytics. I have gained knowledge of various data analytics techniques and methodologies, such as data preprocessing, exploratory data analysis.

4. Operation and Metric Analytics

trinity

Operation Analytics & Investagating metric spike case study

Project Description:

The project is focused on Operation Analytics and Investigating Metric Spike using advanced SQL. The project aims to analyze the complete end-to-end operations of a company, with the help of which the company can then find the areas that need improvement. The data collected is used to predict the overall growth or decline of a company's fortune, resulting in better automation, better understanding between cross-functional teams, and more effective workflows. The project also includes investigating metric spikes, which is an important part of operation analytics, as it helps to understand and answer questions like "Why is there a dip in daily engagement?" or "Why have sales taken a dip?"

Approach:

To execute the project, I firstly collected and analyzed the relevant data sets and tables provided by the company. I then applied advanced SQL queries to derive insights and answer the questions asked by different departments. I also used various data visualization techniques to better understand the data and draw meaningful conclusions.

Tech-Stack Used:

For completion of this project I used mysql workbench 8.0 CE version.

Insights:

Through this project, I gained insights into the importance of operation analytics in predicting a company's growth or decline. I also learned how to use advanced SQL queries to extract insights from large data sets and how to use data visualization techniques to better understand the data. Additionally, I learned how to investigate metric spikes and derive actionable insights from them.

Result:

As a result of this project, I was able to provide a detailed report for the two operations mentioned in the project description. I was able to calculate the number of jobs reviewed per hour per day for November 2020, the 7-day rolling average of throughput, the percentage share of each language in the last 30 days, and how to display duplicate rows from the table. Additionally, I was able to calculate the weekly user engagement, user growth for a product, weekly retention of users-sign up cohort, and weekly engagement of users.

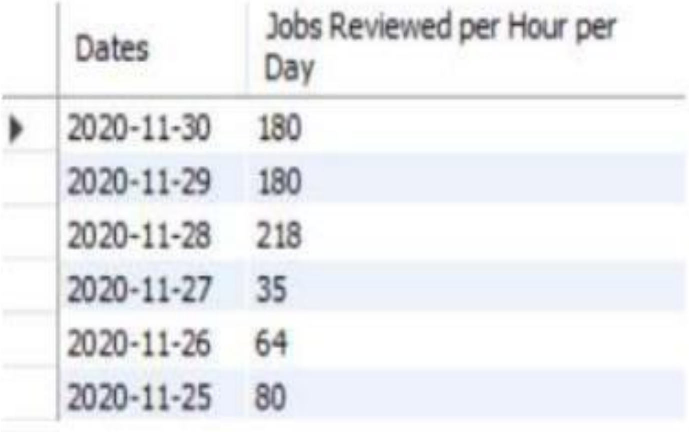
1.

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?

Query used-

```
SELECT ds AS Dates,  ROUND((COUNT(job_id)/SUM(time_spent))*3600) AS "Jobs Reviewed
per Hour per Day"FROM job data WHERE ds BETWEEN 2020-11-01'  AND '2020-11-30'
GROUP BY ds;
```



Dates	Jobs Reviewed per Hour per Day
2020-11-30	180
2020-11-29	180
2020-11-28	218
2020-11-27	35
2020-11-26	64
2020-11-25	80

2.

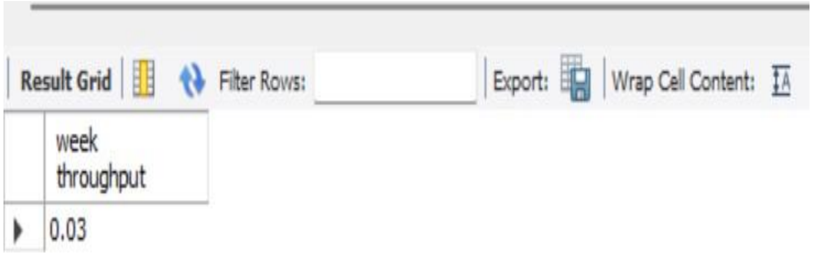
Throughput: It is the no. of events happening per second.

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?

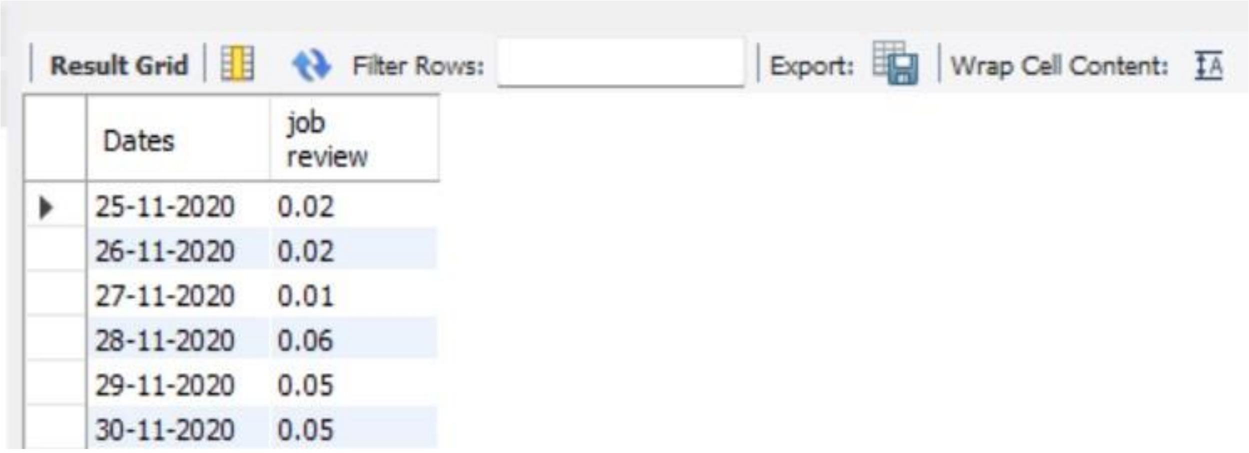
Query used-

```
SELECT ROUND(COUNT(event)/SUM(time_spent),2) AS "week      throughput" FROM
job_data;
```

```
SELECT ds AS Dates,ROUND(COUNT(event)/SUM(time_spent),2) AS  "job review"FROM
job_data GROUP BY ds ORDER BY ds;
```



week throughput
0.03



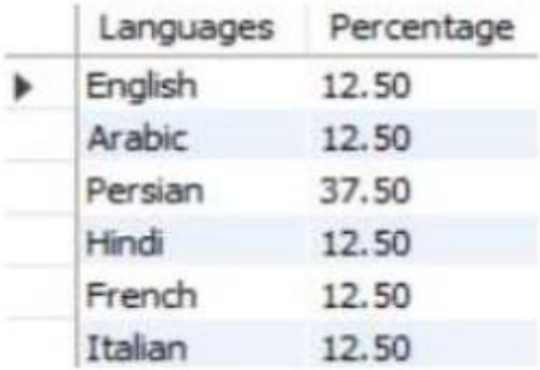
Dates	job review
25-11-2020	0.02
26-11-2020	0.02
27-11-2020	0.01
28-11-2020	0.06
29-11-2020	0.05
30-11-2020	0.05

3.

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?

Query used-

SELECT language AS language ,ROUND(100*COUNT(*)/total,2) AS percentage FROM job_data CROSS JOIN(SELECT COUNT(*) AS total FROM job_data)sub GROUP BY language;



Languages	Percentage
English	12.50
Arabic	12.50
Persian	37.50
Hindi	12.50
French	12.50
Italian	12.50

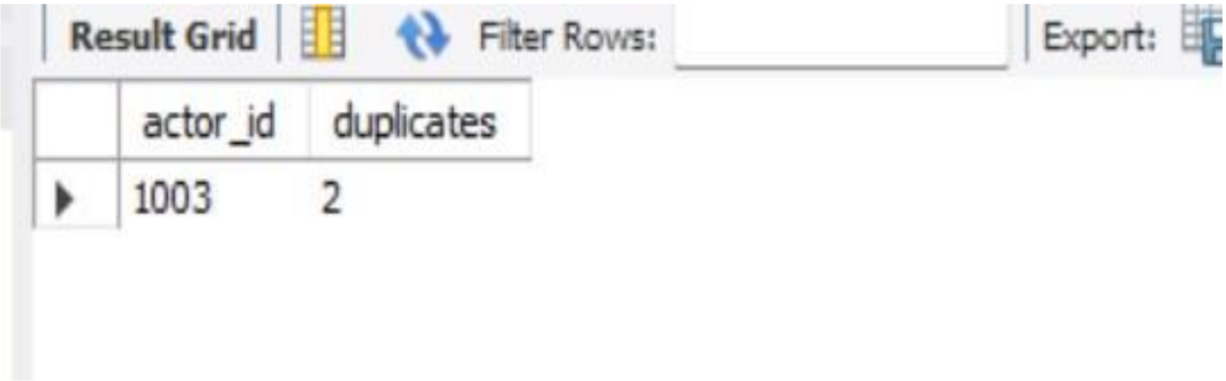
In this query i used , count , Round function to calculate the percentage

4.

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?

Query used-

SELECT actor_id,COUNT(*) AS duplicates FROM job_data GROUP BY actor_id HAVING COUNT(*)>1;



actor_id	duplicates
1003	2

From the above table we can conclude that actor_id 22 and 1003 have duplicate values

5. Hiring Process Analytics

Project Description:

- This project is about analyzing the data of a company's previous hirings in order to draw insights and make recommendations for the hiring department.
- The task is to use knowledge in statistics and Google Sheets to answer questions about the hiring process such as the number of males and females hired, the average salary offered, and the proportion of employees in different departments and post tiers.

Approach:

- First we will be performing our analysis on jupyter notebook using various in-built python libraries such as pandas ,numpy, matplotlib,seaborn etc
- We use EDA understanding columns and rows, identifying missing values,handling missing values, checking outliers ,removing outlier.
- we will understand the various columns, the data it contains and their characteristics. Then we will look for duplicates in the data set and if any we will remove it through remove duplicate inbuilt function of python.
- I have used **jupyter notebook** instead of excel or google sheets because I had already knowledge about datasets, python and jupyter notebook tool which I had learn from college.
- I have used excel also.

Tech-Stack Used:

- **python**
- **Jupyter notebook**
- **Microsoft excel**

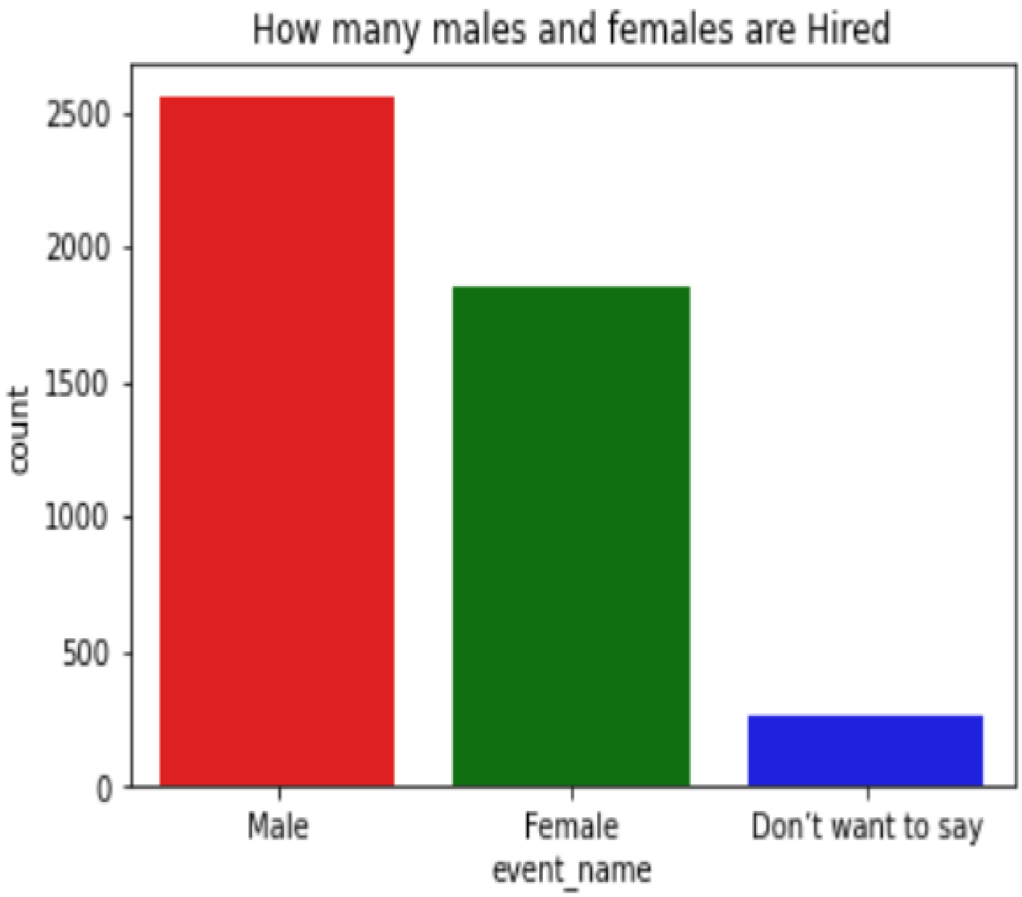
Insights and Results:

Q-1. Hiring: Process of intaking people into an organization for different kinds of positions.
Your task: How many males and females are Hired ?

Ans -> By Analyzing the data we have found that from the total 7168 people there are 2563 candidates are male and 1856 candidates are females which are Hired by the company.

Code used-

```
# Filtering males and females hired.
abc=data[data["Status"]=="Hired"] # only hired
plt.title("How many males and females are Hired") # title of graph  p=["Red","green","blue"] # for color
sns.countplot(x =abc.event_name, data = data,palette=p) # for count  plt.show()        # for showing graph
```



Q-2. Average Salary: Adding all the salaries for a select group of employees and then dividing the sum by the number of employees in the group.

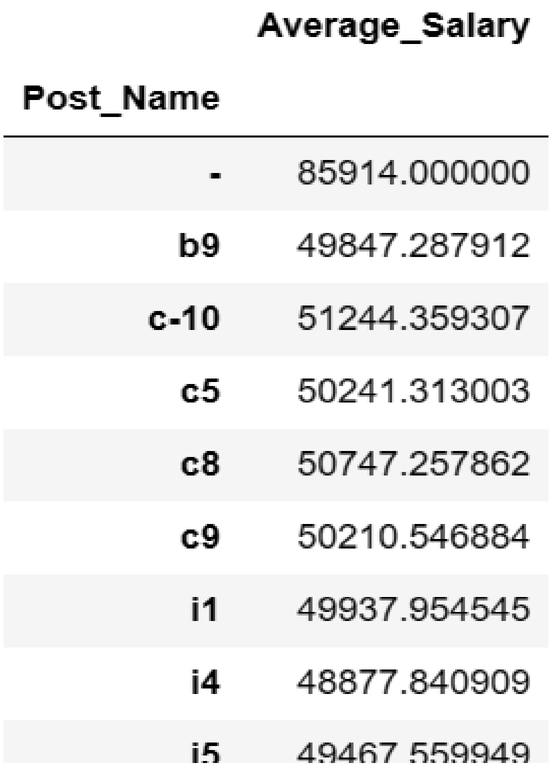
Your task: What is the average salary offered in this company ?

Ans -> The Average salary offered by the company is 85914 INR.

code used-

```
abc=data.groupby(["Post_Name"]).Offered_Salary.agg(["mean"])
abc.rename(columns={"mean":"Average_Salary"},inplace=True)
```

Abc

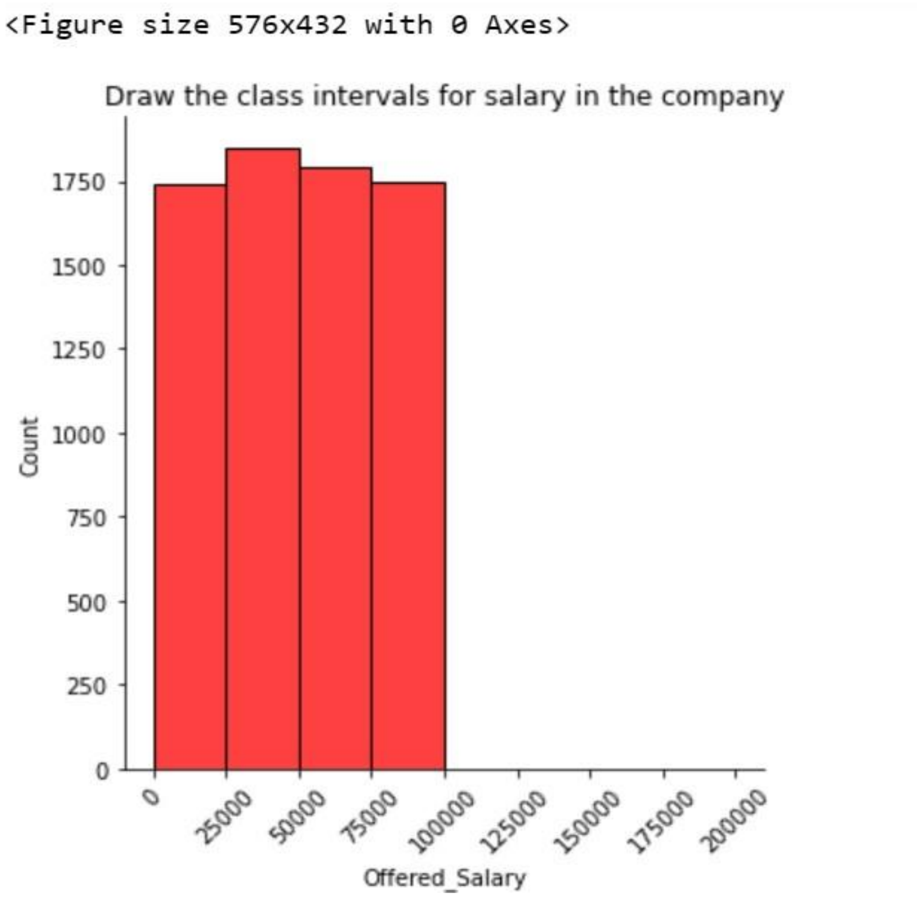


	Average_Salary
Post_Name	
-	85914.000000
b9	49847.287912
c-10	51244.359307
c5	50241.313003
c8	50747.257862
c9	50210.546884
i1	49937.954545
i4	48877.840909
i5	49467.559949

Q-3. Class Intervals: The class interval is the difference between the upper class limit and the lower class limit. Your task: Draw the class intervals for salary in the company ?

Ans -> code used-

```
plt.figure(figsize=(8,6))

sns.displot(data["Offered_Salary"],

bins=[0,25000,50000,75000,100000,125000,150000,175000,200000]
,color="red")
plt.title("Draw the class intervals for salary in the company")  locs,labels=plt.xticks()
plt.setp(labels, rotation=45)  plt.show()
```



```
<Figure size 576x432 with 0 Axes>
```

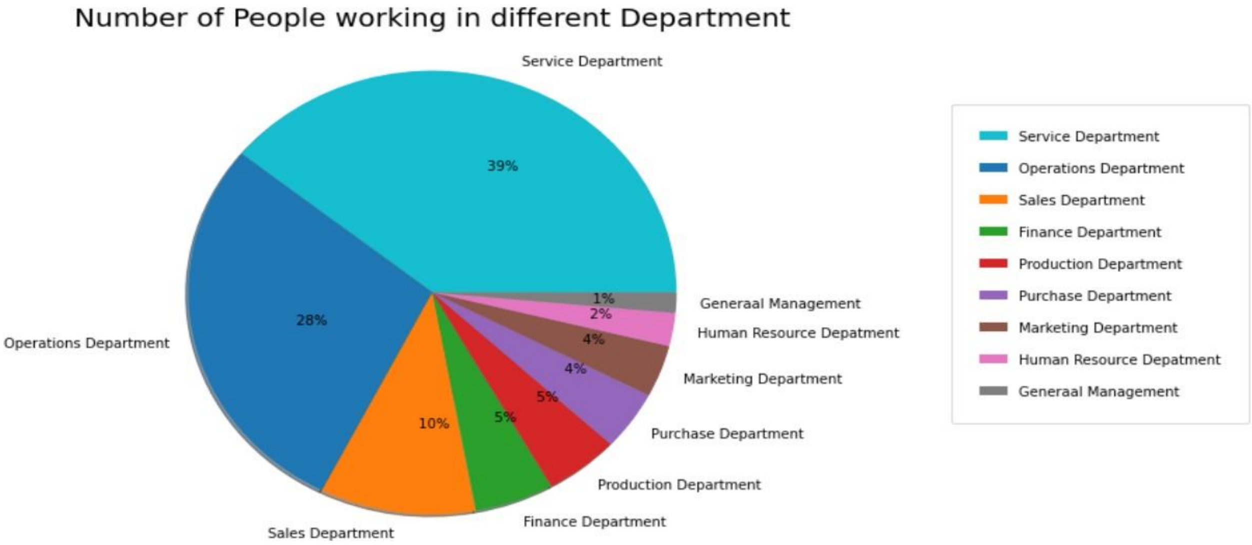
Q-4. **Charts and Plots: This is one of the most important parts of analysis to visualize the data.**

Your task: Draw Pie Chart / Bar Graph (or any other graph) to show the proportion of people working in different departments ?

Ans ->

Code used:-

```
Dep=["Service Department","Operations Department","Sales Department","Finance Department",

"Production Department","Purchase Department","Marketing Department", "Human Resource Depatment", "Generaal Management"]  total=c.Department.value_counts() explode=(0,0,0,0,0,0,0,0,0) plt.pie(total) explode=(0,0,0,0,0,0,0,0,0)

plt.pie(total,explode=explode,labels=Dep,autopct='%12.0f%%',shadow=True,radius = 2,startangle=360)  plt.title('Number of People working in different Department',fontsize = 20,pad=100.0)

plt.legend(loc="upper right",handlelength=2, borderpad=2, labelspacing=1.5,bbox_to_anchor=(3.2,1.2))  plt.show()
```

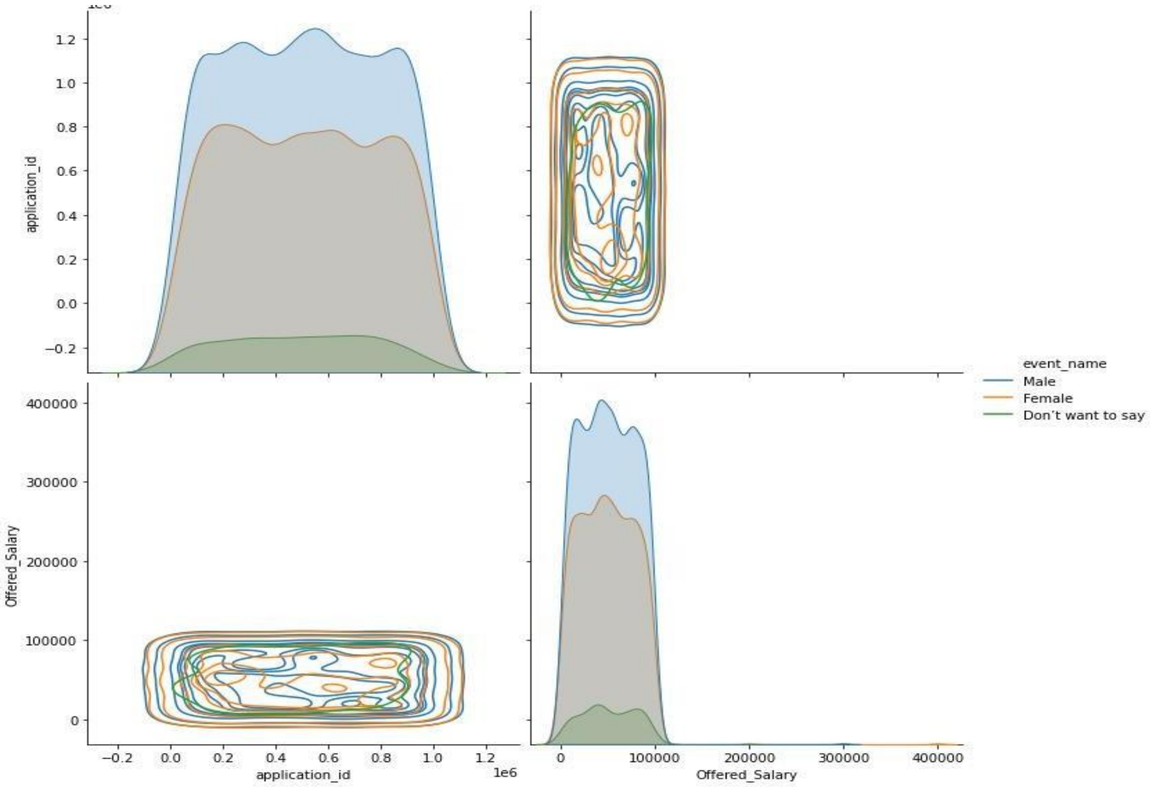


Q-5. Charts: Use different charts and graphs to perform the task representing the data.

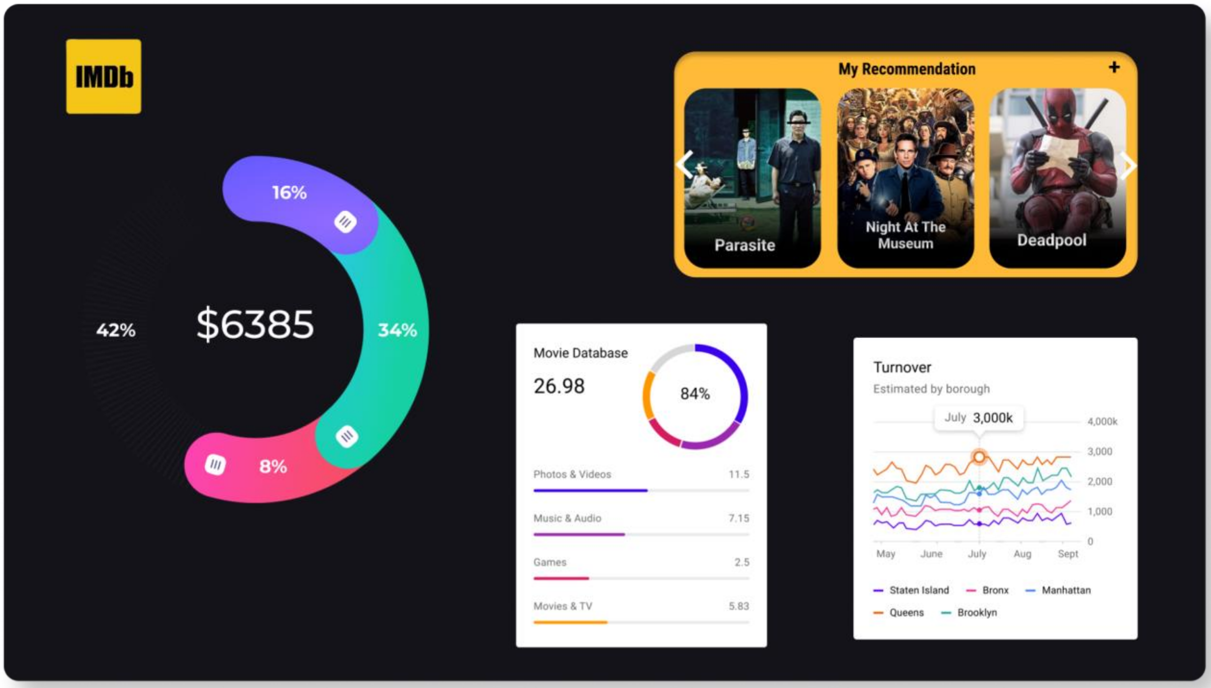
Your task: Represent different post tiers using a chart/graph?

Ans -> Code used:-

```
sns.pairplot(c,hue="event_name“
,kind="kde",height=5)  plt.show()
```



6. IMDB Movie Analysis



Project Description:

The purpose of this project is to analyze a data record of movies and gain insights into the various aspects of the film industry. This data record contains information such as movie title, director name, actors, IMDb score, budget, gross, and more.

Approach:

- **First we will be performing our analysis on jupyter notebook using various in-built python libraries such as pandas ,numpy, matplotlib,seaborn etc**
- **We use EDA understanding columns and rows, identifying missing values,handling missing values, checking outliers ,removing outlier.**
- **I have used jupyter notebook instead of excel or google sheets because I had already knowledge about datasets, python and jupyter notebook tool which I had learn from college.**
- **I have used excel also.**
- **Steps for doing project-**
- **Download the dataset > understanding the data > find duplicate and null > visualization> data insights**

Tech-Stack Used:

The software used in this project is a Python Jupyter notebook. Python was chosen for its ease of use and the ability to easily perform data analysis and manipulation.

Insights:

During the course of this project, the following insights were gained:

[1] The importance of cleaning the data before conducting any analysis.

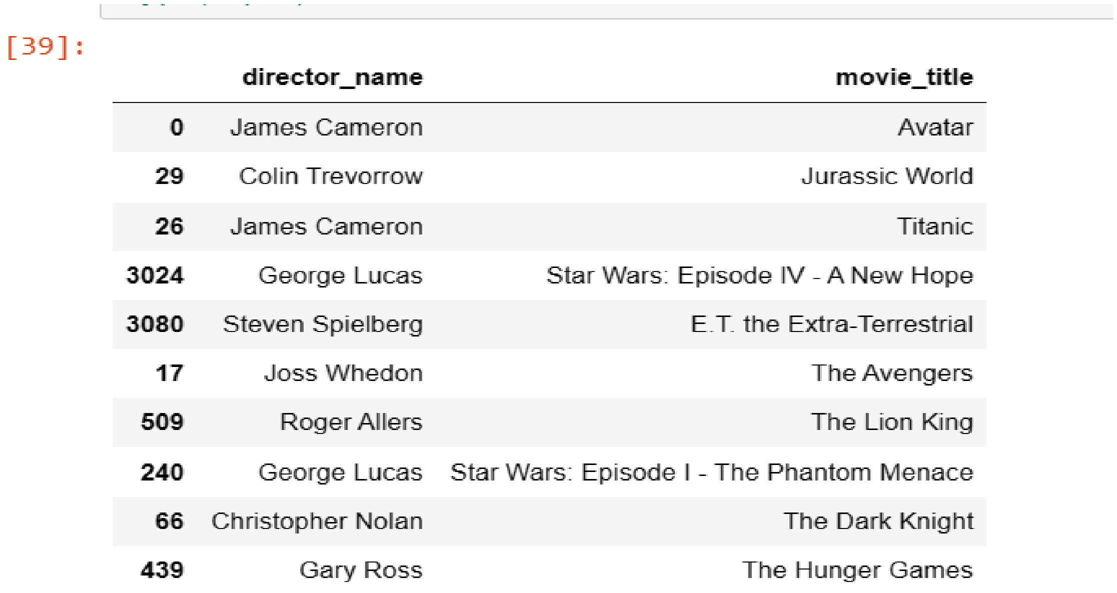
- Import the dataset
- Understood the dataset
- Remove irrelevant data
- Deal with null and missing data
- Fill the missing data

- Filter out data outlier
- Validate the data

[2] Find the movies with the highest profit?

Code used-

```
movies['profit']=movies['gross']-movies['budget']
movie=movies.sort_values(by=['profit'],ascending=False)
top10=movie[['director_name','movie_title']]
top10.head(10)
```



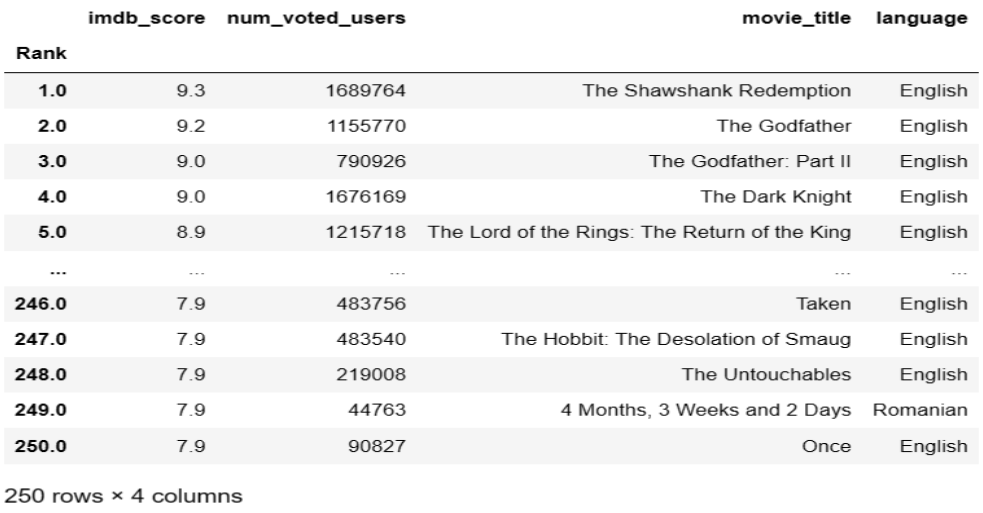
[39]:

	director_name	movie_title
0	James Cameron	Avatar
29	Colin Trevorrow	Jurassic World
26	James Cameron	Titanic
3024	George Lucas	Star Wars: Episode IV - A New Hope
3080	Steven Spielberg	E.T. the Extra-Terrestrial
17	Joss Whedon	The Avengers
509	Roger Allers	The Lion King
240	George Lucas	Star Wars: Episode I - The Phantom Menace
66	Christopher Nolan	The Dark Knight
439	Gary Ross	The Hunger Games

[3] The top 250 movies based on IMDb score and the popularity of these movies.

code used-

```
IMDb_Top_250 = IMDb_Top_250.set_index("Rank")
IMDb_Top_250.head(250)
```

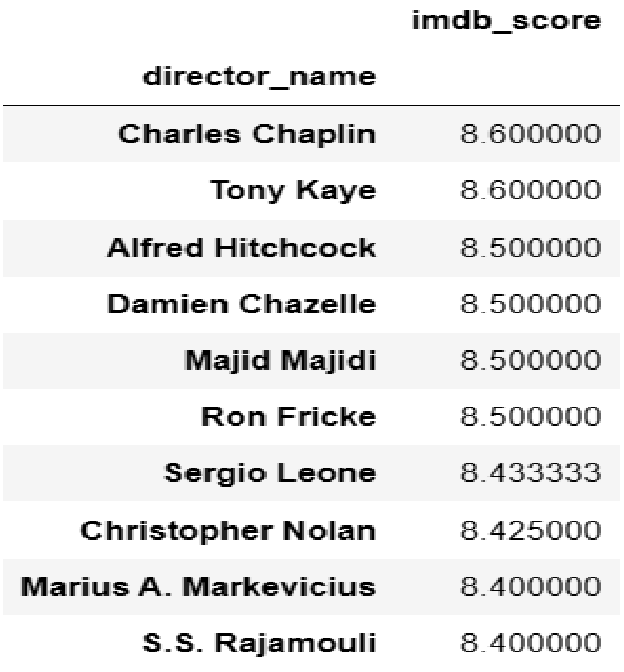


Rank	imdb_score	num_voted_users	movie_title	language
1.0	9.3	1689764	The Shawshank Redemption	English
2.0	9.2	1155770	The Godfather	English
3.0	9.0	790926	The Godfather: Part II	English
4.0	9.0	1676169	The Dark Knight	English
5.0	8.9	1215718	The Lord of the Rings: The Return of the King	English
...	...	...	...	...
246.0	7.9	483756	Taken	English
247.0	7.9	483540	The Hobbit: The Desolation of Smaug	English
248.0	7.9	219008	The Untouchables	English
249.0	7.9	44763	4 Months, 3 Weeks and 2 Days	Romanian
250.0	7.9	90827	Once	English

250 rows × 4 columns

[4] Find the best directors
code used-

```
mov=movies.groupby('director_name')

top10director=pd.DataFrame(mov['imdb_score'].mean().sort_values(ascending=False))

top10director=top10director.head(10)

top10director=top10director.sort_values(['imdb_score','director_name'],ascending=(False,True))  top10director
```

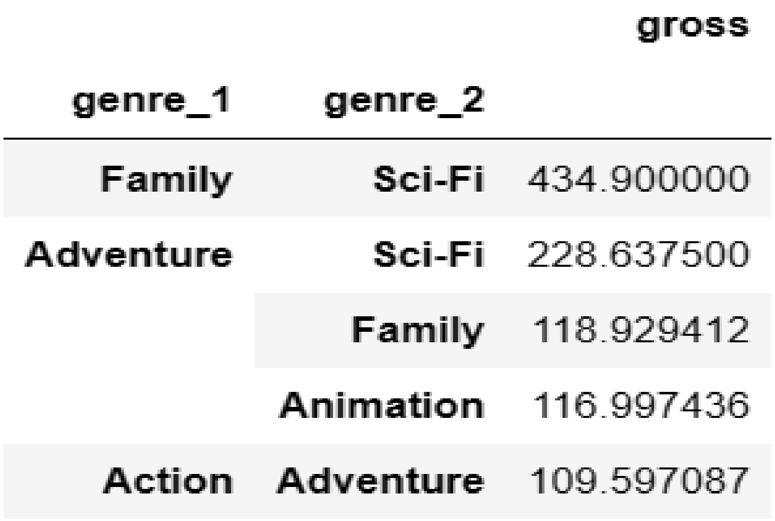


director_name	imdb_score
Charles Chaplin	8.600000
Tony Kaye	8.600000
Alfred Hitchcock	8.500000
Damien Chazelle	8.500000
Majid Majidi	8.500000
Ron Fricke	8.500000
Sergio Leone	8.433333
Christopher Nolan	8.425000
Marius A. Markevicius	8.400000
S.S. Rajamouli	8.400000

[5] Find popular genres

Code used:-

```
movies['genres']=movies.genres.str.split('|')

movies['genre_1']=movies['genres'].apply(lambda x: x[0])

movies['genre_2']=movies['genres'].apply(lambda x: x[1] if len(x)>1 else x[0])

movies.head()

..

..

PopGenre=pd.DataFrame(movies_by_segment.gross.

mean().sort_values(ascending=False) )

PopGenre[0:5]
```

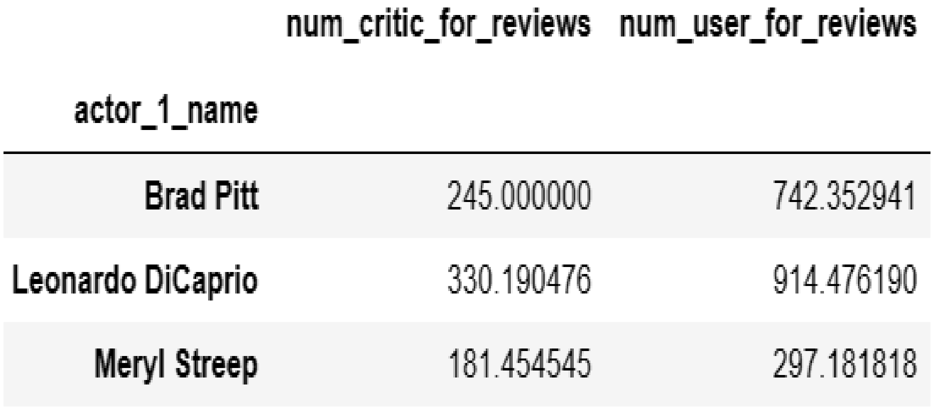


		gross
genre_1	genre_2	
Family	Sci-Fi	434.900000
Adventure	Sci-Fi	228.637500
	Family	118.929412
	Animation	116.997436
Action	Adventure	109.597087

[6] Find the critic-favorite and audience-favorite actors

Code used:-

```
Combined.groupby('actor_1_name').num_user_for_reviews.mean()

..

..

Combined.groupby('actor_1_name')[['num_critic_for_reviews','num_user_for_reviews']].mean()
```



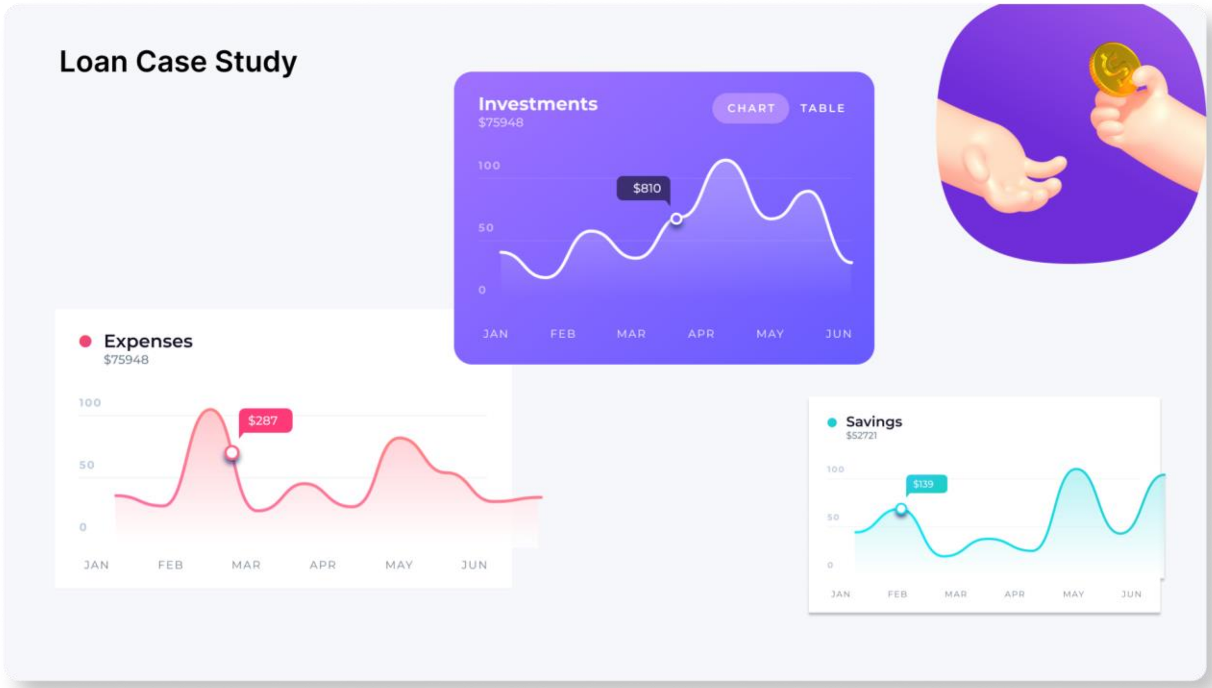
	num_critic_for_reviews	num_user_for_reviews
actor_1_name		
Brad Pitt	245.000000	742.352941
Leonardo DiCaprio	330.190476	914.476190
Meryl Streep	181.454545	297.181818

Result:

- I got a chance to work with real time dataset. Also I got chance to again work with python tools and libraries. I learned more techniques in jupyter notebbok.
- I have answered all the questions asked in the data set and tried to plot the required graphs and chat as per requirement and my understanding. This project has helped in getting a hands-on experience on real life data set and how we clean, manipulate, visualize, and draw insights from the data.
- Exploratory Data Analysis part has helped me to understand that before moving towards making further analytical treatment of data and making it fit for making models. We have to do the EDA to make the data error free and bias free so that the inference drawn from the data is a good a fit to the further statistical and analytical treatment.

- I understood how to derive insights from a dataset that we are given.

7. Bank Loan Case Study



Tech-Stack Used:

The software used in this project is a Python Jupyter notebook. Python was chosen for its ease of use and the ability to easily perform data analysis and manipulation.

Problem Statement:

- **This case study aims to identify patterns which indicate if an applicant has difficulty in paying his/her installments which may be used for taking**

actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected and the number of defaulters is also reduced.

- **Banks want to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The bank can utilize this knowledge for its portfolio and risk assessment.**

Datasets:

- ***1. 'application_data.csv'* contains all the information of the client at the time of application.**

- **The data is about whether an applicant has payment difficulties.**

- ***2. 'previous_application.csv'* contains information about the applicant's previous loan data. It contains the data whether the previous application had been Approved, Canceled, Refused or Unused offer.**

- ***3. 'columns_description.csv'* is a data dictionary which elaborates the meaning of the variables.**

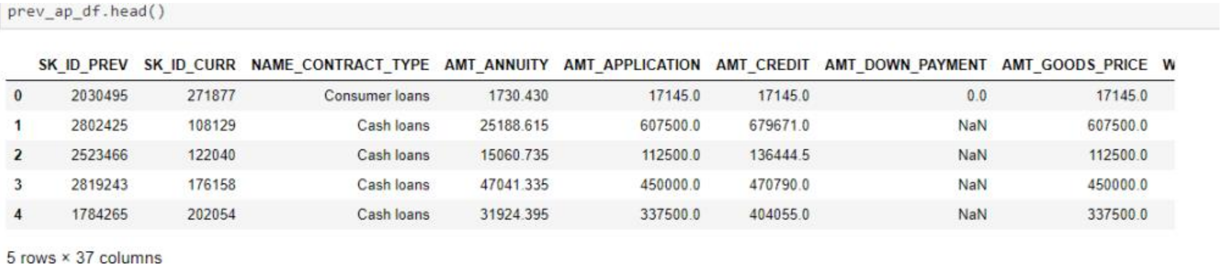
- **Previous Application Data Analysis:**

- **In this segment, I have mainly focused on analyzing previous_application.csv i.e. data about previous application of an applicant.**

Approach:

For the Exploratory data analysis, mentioned steps have been followed.

–> Import Modules, Read the dataset

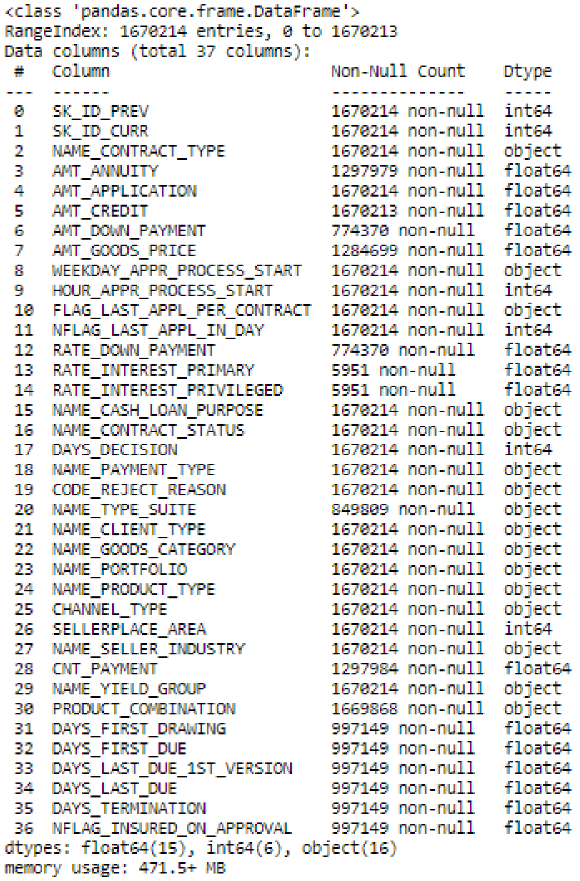


```
prev_ap_df.head()
```

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	W
0	2030495	271877	Consumer loans	1730.430	17145.0	17145.0	0.0	17145.0	
1	2802425	108129	Cash loans	25188.615	607500.0	679671.0	NaN	607500.0	
2	2523466	122040	Cash loans	15060.735	112500.0	136444.5	NaN	112500.0	
3	2819243	176158	Cash loans	47041.335	450000.0	470790.0	NaN	450000.0	
4	1784265	202054	Cash loans	31924.395	337500.0	404055.0	NaN	337500.0	

5 rows × 37 columns

Previous_application_data



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1670214 entries, 0 to 1670213
Data columns (total 37 columns):
 #   Column                       Non-Null Count    Dtype  
---  ------                       --------------    -----  
 0   SK_ID_PREV                   1670214 non-null  int64  
 1   SK_ID_CURR                   1670214 non-null  int64  
 2   NAME_CONTRACT_TYPE           1670214 non-null  object 
 3   AMT_ANNUITY                  1297979 non-null  float64
 4   AMT_APPLICATION              1670214 non-null  float64
 5   AMT_CREDIT                   1670213 non-null  float64
 6   AMT_DOWN_PAYMENT             774370 non-null   float64
 7   AMT_GOODS_PRICE              1284699 non-null  float64
 8   WEEKDAY_APPR_PROCESS_START   1670214 non-null  object 
 9   HOUR_APPR_PROCESS_START      1670214 non-null  int64  
 10  FLAG_LAST_APPL_PER_CONTRACT  1670214 non-null  object 
 11  NFLAG_LAST_APPL_IN_DAY       1670214 non-null  int64  
 12  RATE_DOWN_PAYMENT            774370 non-null   float64
 13  RATE_INTEREST_PRIMARY        5951 non-null     float64
 14  RATE_INTEREST_PRIVILEGED     5951 non-null     float64
 15  NAME_CASH_LOAN_PURPOSE       1670214 non-null  object 
 16  NAME_CONTRACT_STATUS         1670214 non-null  object 
 17  DAYS_DECISION                1670214 non-null  int64  
 18  NAME_PAYMENT_TYPE            1670214 non-null  object 
 19  CODE_REJECT_REASON           1670214 non-null  object 
 20  NAME_TYPE_SUITE              849809 non-null   object 
 21  NAME_CLIENT_TYPE             1670214 non-null  object 
 22  NAME_GOODS_CATEGORY          1670214 non-null  object 
 23  NAME_PORTFOLIO               1670214 non-null  object 
 24  NAME_PRODUCT_TYPE            1670214 non-null  object 
 25  CHANNEL_TYPE                 1670214 non-null  object 
 26  SELLERPLACE_AREA             1670214 non-null  int64  
 27  NAME_SELLER_INDUSTRY         1670214 non-null  object 
 28  CNT_PAYMENT                  1297984 non-null  float64
 29  NAME_YIELD_GROUP             1670214 non-null  object 
 30  PRODUCT_COMBINATION          1669868 non-null  object 
 31  DAYS_FIRST_DRAWING           997149 non-null   float64
 32  DAYS_FIRST_DUE               997149 non-null   float64
 33  DAYS_LAST_DUE_1ST_VERSION    997149 non-null   float64
 34  DAYS_LAST_DUE                997149 non-null   float64
 35  DAYS_TERMINATION             997149 non-null   float64
 36  NFLAG_INSURED_ON_APPROVAL    997149 non-null   float64
dtypes: float64(15), int64(6), object(16)
memory usage: 471.5+ MB
```

- **prev_ap_df contains 37 features and 1670214 rows(Out of which 15 features are float64, 6 features are integer, 16 features are object data type)**
- **Following are the common features among application data and previous application data [‘SK_ID_CURR’, ‘NAME_CONTRACT_TYPE’, ‘AMT_CREDIT’, ‘AMT_ANNUITY’,**

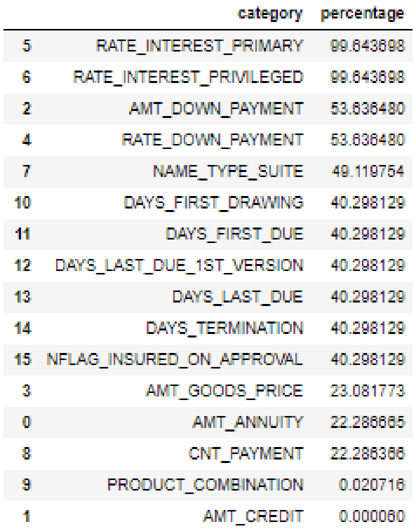
‘AMT_GOODS_PRICE’, ‘NAME_TYPE_SUITE’, ‘WEEKDAY_APPR_PROCESS_START’, ‘HOUR_APPR_PROCESS_START’]

- SK_ID_CURR is an unique identifier, which will be used to merge the relevant columns of 2 dataframes (application data and previous application data).

–> Data Cleaning

Missing value handling, Type Casting, Fixing Rows and Columns – removing unnecessary rows/columns (through missing value handling and correlation), Handling Outliers.

First I have calculated the missing value percentages for each feature in previous application data.



	category	percentage
5	RATE_INTEREST_PRIMARY	99.643698
6	RATE_INTEREST_PRIVILEGED	99.643698
2	AMT_DOWN_PAYMENT	53.636480
4	RATE_DOWN_PAYMENT	53.636480
7	NAME_TYPE_SUITE	49.119754
10	DAYS_FIRST_DRAWING	40.298129
11	DAYS_FIRST_DUE	40.298129
12	DAYS_LAST_DUE_1ST_VERSION	40.298129
13	DAYS_LAST_DUE	40.298129
14	DAYS_TERMINATION	40.298129
15	NFLAG_INSURED_ON_APPROVAL	40.298129
3	AMT_GOODS_PRICE	23.081773
0	AMT_ANNUITY	22.286665
8	CNT_PAYMENT	22.286366
9	PRODUCT_COMBINATION	0.020716
1	AMT_CREDIT	0.000060

Missing value percentages for features.

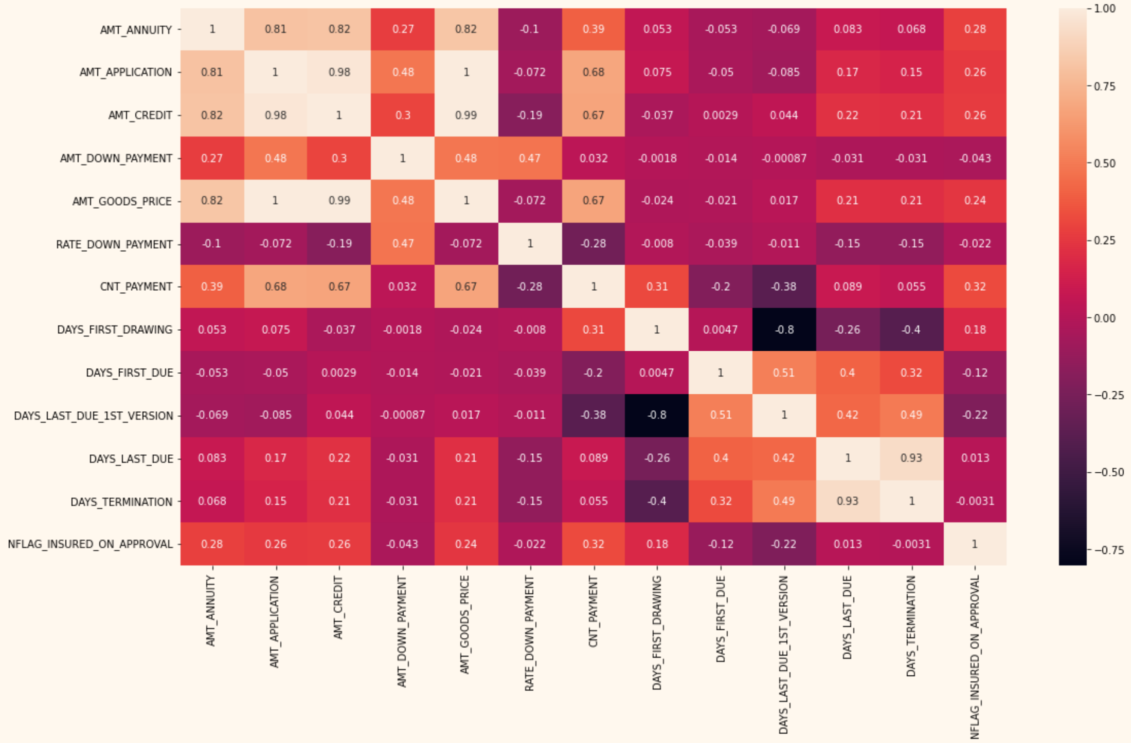
There are 16 features in prev_app_df that have missing values.

- **Permanently dropping the features (RATE_INTEREST_PRIMARY and RATE_INTEREST_PRIVILEGED) as 99% data is missing.**

- **Dropping rows containing missing values for the features(AMT_CREDIT and PRODUCT_COMBINATION) for very low % of missing data. Dropping entries would not impact the analysis as the percentage of missing value is very low (~2%).**

–> Univariate Analysis, Bivariate and Multivariate Analysis

First, I extracted the numerical variables in the dataset and checked out the correlation coefficients with the help of a heatmap.



	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	RATE_DOWN_PAYMENT	CNT_PAYMENT	DAYS_FIRST_DRAWING	DAYS_FIRST_DUE	DAYS_LAST_DUE_1ST_VERSION	DAYS_LAST_DUE	DAYS_TERMINATION	NFLAG_INSURED_ON_APPROVAL
AMT_ANNUITY	1	0.81	0.82	0.27	0.82	-0.1	0.39	0.053	-0.053	-0.069	0.083	0.068	0.28
AMT_APPLICATION	0.81	1	0.98	0.48	1	-0.072	0.68	0.075	-0.05	-0.085	0.17	0.15	0.26
AMT_CREDIT	0.82	0.98	1	0.3	0.99	-0.19	0.67	-0.037	0.0029	0.044	0.22	0.21	0.26
AMT_DOWN_PAYMENT	0.27	0.48	0.3	1	0.48	0.47	0.032	-0.0018	-0.014	-0.00087	-0.031	-0.031	-0.043
AMT_GOODS_PRICE	0.82	1	0.99	0.48	1	-0.072	0.67	-0.024	-0.021	0.017	0.21	0.21	0.24
RATE_DOWN_PAYMENT	-0.1	-0.072	-0.19	0.47	-0.072	1	-0.28	-0.008	-0.039	-0.011	-0.15	-0.15	-0.022
CNT_PAYMENT	0.39	0.68	0.67	0.032	0.67	-0.28	1	0.31	-0.2	-0.38	0.089	0.055	0.32
DAYS_FIRST_DRAWING	0.053	0.075	-0.037	-0.0018	-0.024	-0.008	0.31	1	0.0047	-0.8	-0.26	-0.4	0.18
DAYS_FIRST_DUE	-0.053	-0.05	0.0029	-0.014	-0.021	-0.039	-0.2	0.0047	1	0.51	0.4	0.32	-0.12
DAYS_LAST_DUE_1ST_VERSION	-0.069	-0.085	0.044	-0.00087	0.017	-0.011	-0.38	-0.8	0.51	1	0.42	0.49	-0.22
DAYS_LAST_DUE	0.083	0.17	0.22	-0.031	0.21	-0.15	0.089	-0.26	0.4	0.42	1	0.93	0.013
DAYS_TERMINATION	0.068	0.15	0.21	-0.031	0.21	-0.15	0.055	-0.4	0.32	0.49	0.93	1	-0.0031
NFLAG_INSURED_ON_APPROVAL	0.28	0.26	0.26	-0.043	0.24	-0.022	0.32	0.18	-0.12	-0.22	0.013	-0.0031	1

Heatmap of correlation coefficients

- **‘DAYS_LAST_DUE’ and ‘DAYS_TERMINATION’ are highly correlated**
- **‘DAYS_FIRST_DRAWING’ and ‘DAYS_LAST_DUE_1st_VERSION’ have high negative correlation**

- **‘AMT_ANNUITY’,’AMT_APPLICATION’,’AMT_CREDIT’,’AMT_GOODS_PRICE’ are highly correlated**

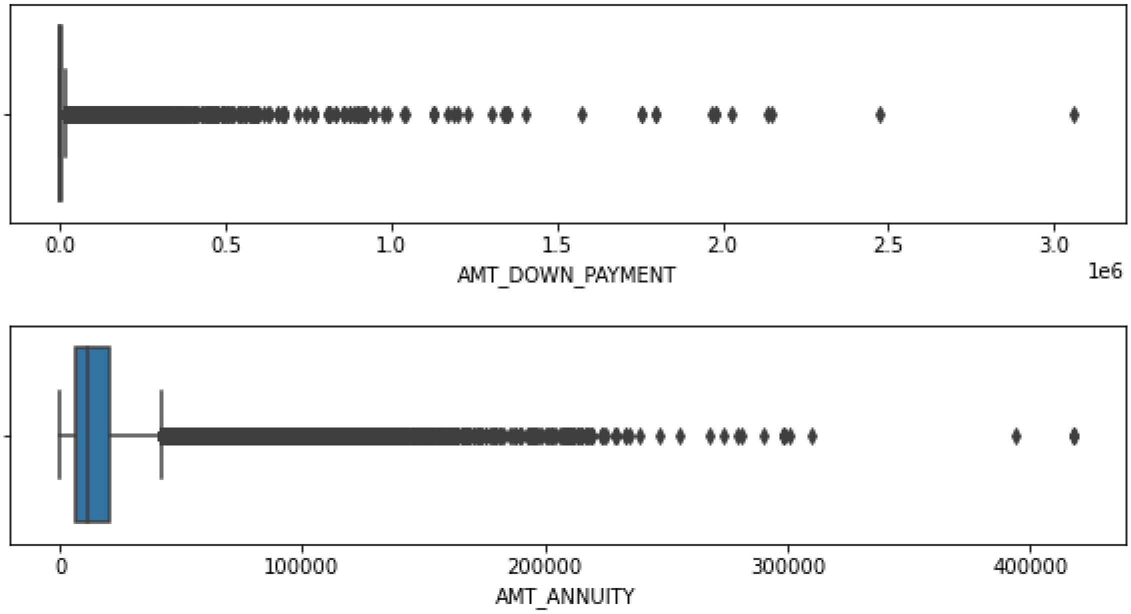
The features can be removed before modeling this data, as they would cause collinearity ‘DAYS_TERMINATION’,’DAYS_LAST_DUE_1st_VERSION’,’AMT_APPLICATION’,’AMT_CREDIT’,’AMT_GOODS_PRICE’ For EDA purpose we are not removing them.

- **‘SK_ID_PREV’ column is not required for analysis.**
- **Filling missing value as ‘Unaccompanied’ as most common value.**

For merging the 2 dataframes (application data and previous application data), I have take remaining features of previous application data and only 2 columns from application data (‘SK_ID_CURR’, ‘TARGET’)

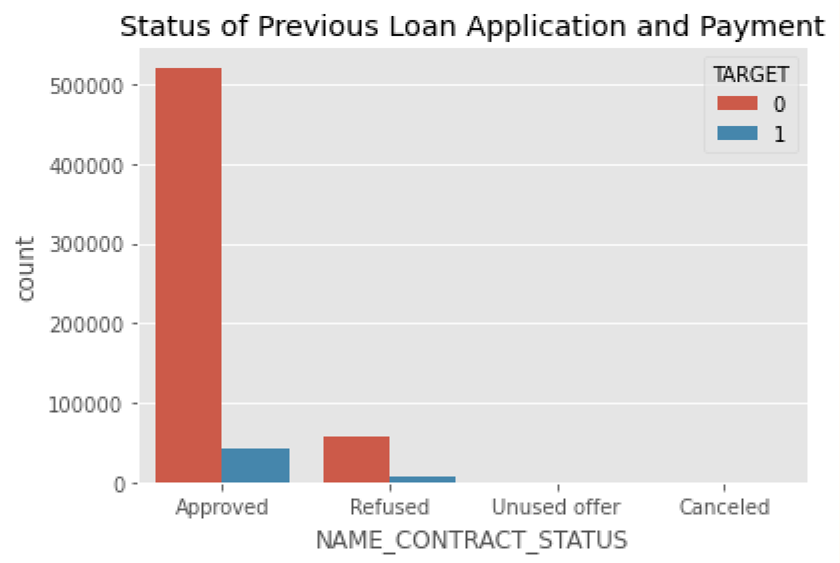
Handling Outliers

Boxplot of Amount Annuity && Boxplot of Amount of Down Payment

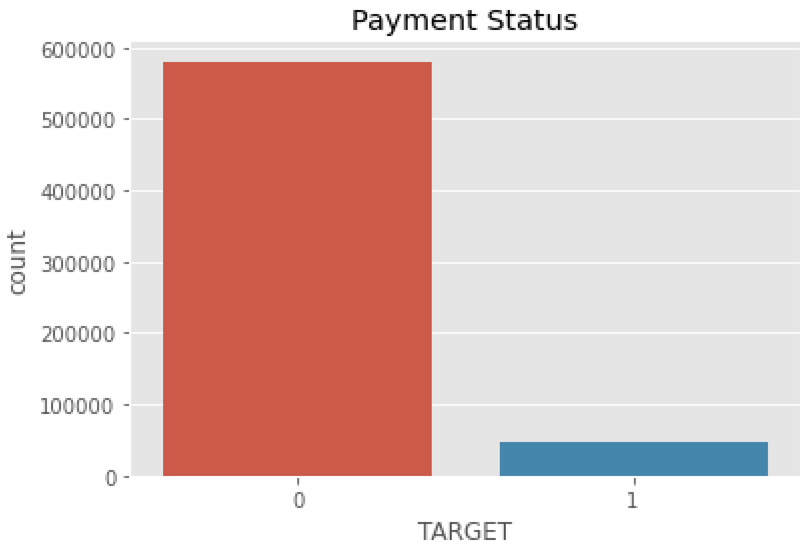


Excluding values outside 99 percentile for AMT_ANNUITY and AMT_DOWN_PAYMENT

Checking Data Imbalance in Previous Application Data

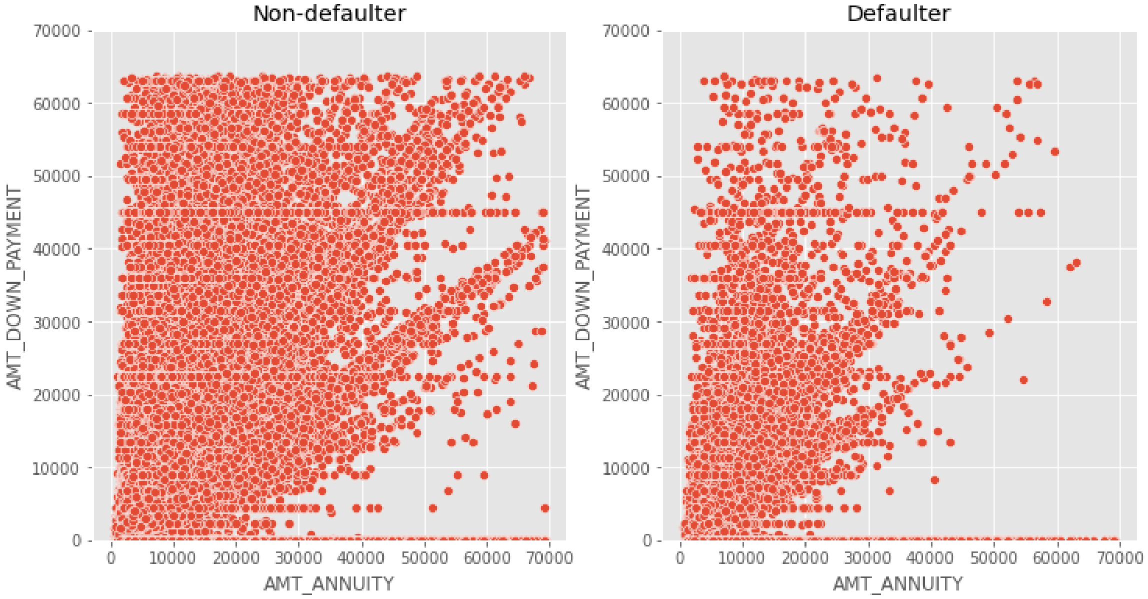


- The applicants whose previous loans were approved are more likely to pay the current loan in time, than the applicants whose previous loans were rejected.
- 7% of the previously approved loan applicants that defaulted in current loan
- 90 % of the previously refused loan applicants that were able to pay current loan

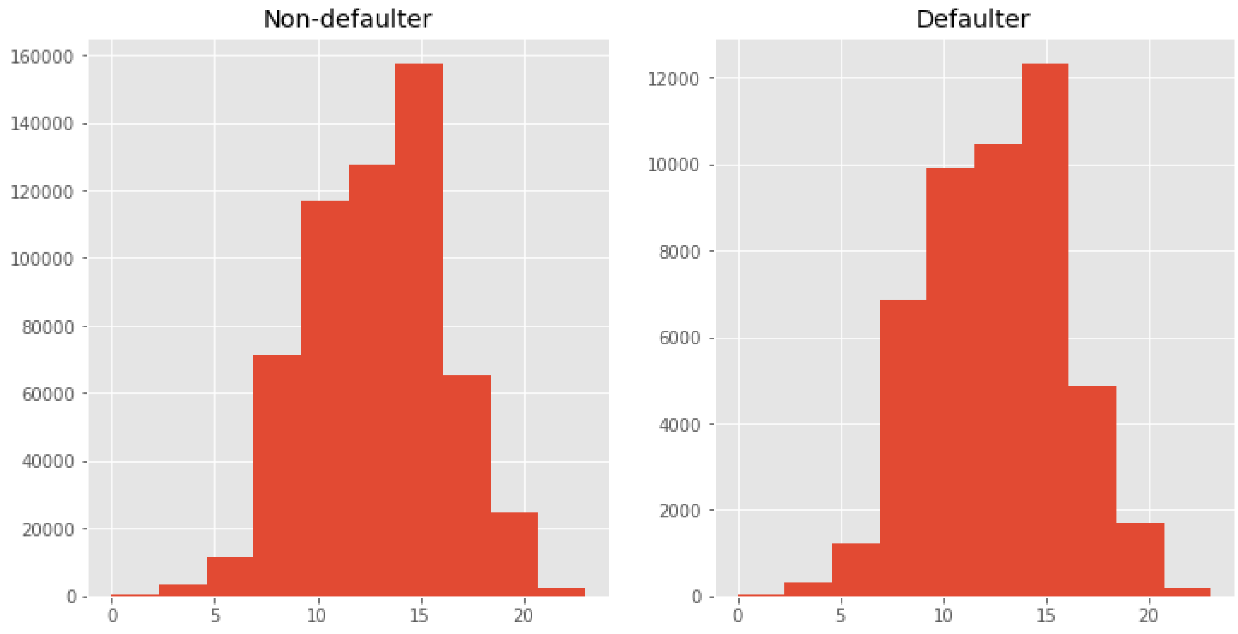


- This data is highly imbalanced as the number of defaulters is very less in total population.
- ‘FLAG_LAST_APPL_PER_CONTRACT’ can be dropped for having fixed value in all entries.
- ‘NFLAG_LAST_APPL_IN_DAY’ can be dropped for having highly imbalanced data.

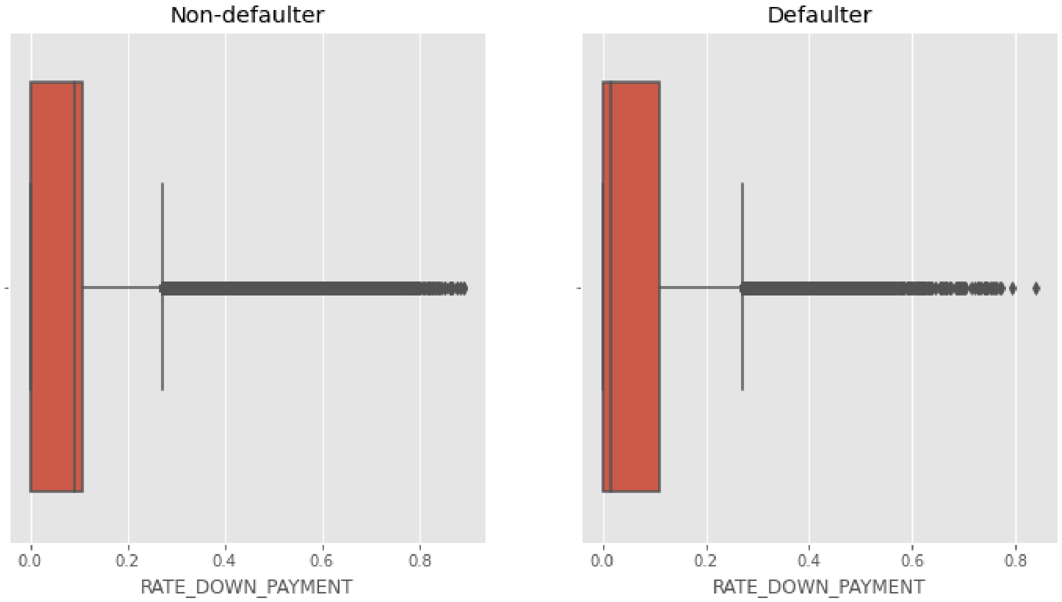
Analysis of Numeric Features of Previous Application Data



- Number of defaulters is less for the larger amount of annuity of the previous application.
- For higher down payment, defaulter cases are less.

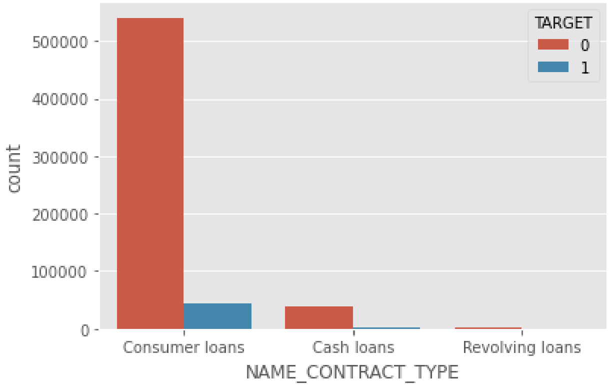


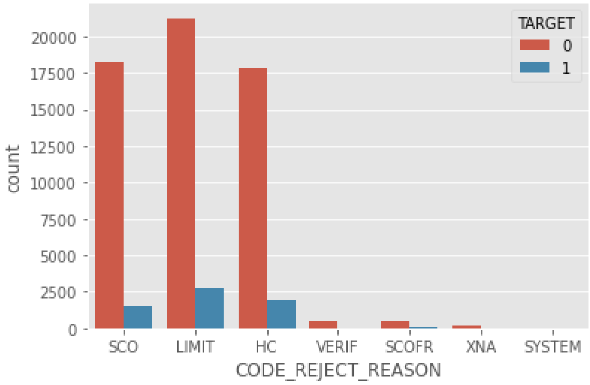
Most of the loans are applied around 15:00 hours. This feature is does not have visible impact on TARGET variable

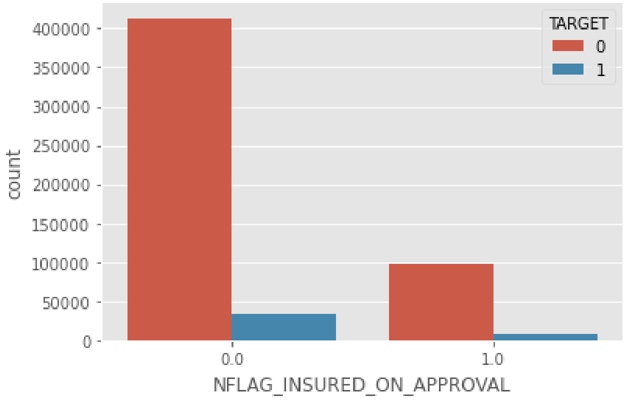


For those who had lower rates of down payment in previous applications, cases of default are higher.

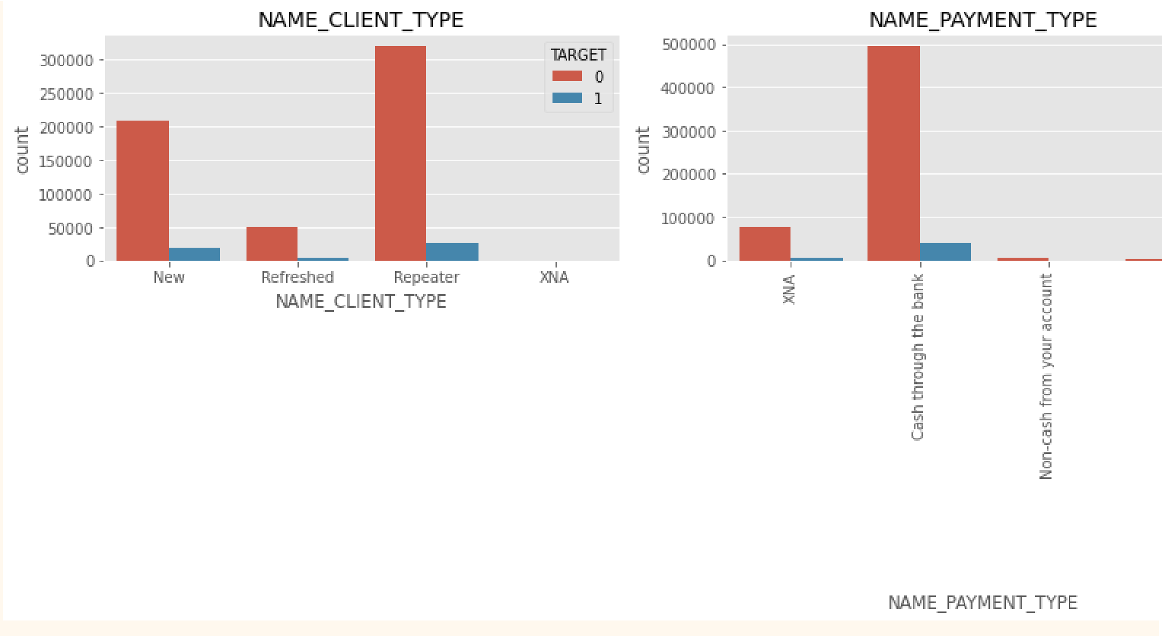
Analysis of Categorical Features of Previous Application Data





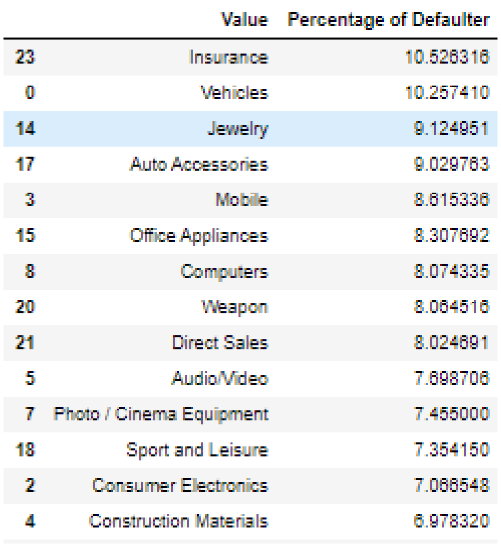
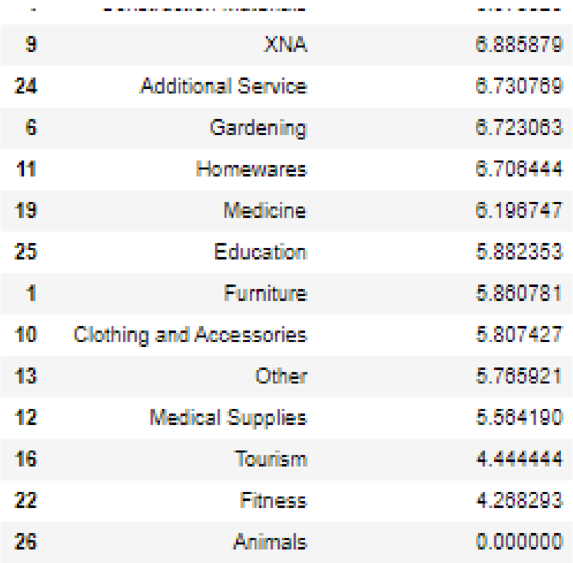


- Highest number of loans are applied for Consumer Loans
- As seen in the above plot, ‘SCO’, ‘LIMIT’ and ‘HC’ are the most common reasons for rejection.
- Most of the people did not request insurance during the previous loan application.



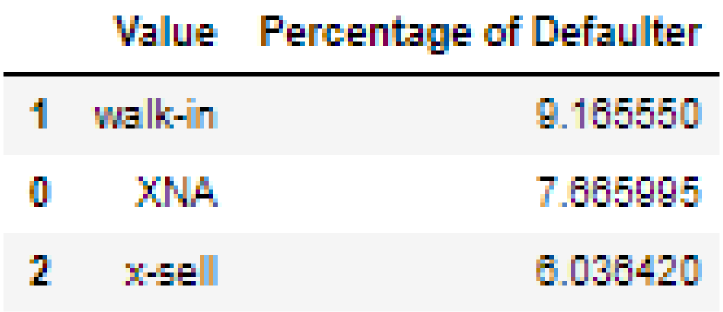
- Most of the applicants are repeaters.
- ‘Cash through the bank’ is the most frequently used payment method

Then for a given categorical feature, I obtained a percentage of defaulters.

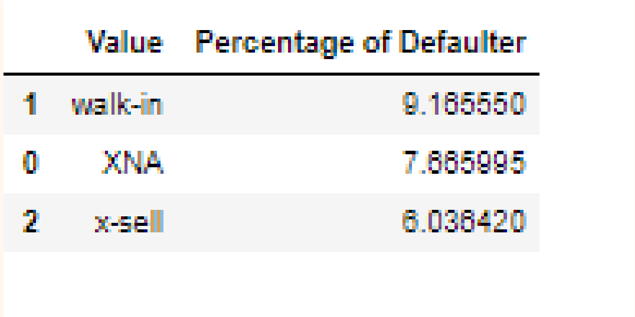
	Value	Percentage of Defaulter
23	Insurance	10.526316
0	Vehicles	10.257410
14	Jewelry	9.124951
17	Auto Accessories	9.029763
3	Mobile	8.615336
15	Office Appliances	8.307692
8	Computers	8.074335
20	Weapon	8.064516
21	Direct Sales	8.024691
5	Audio/Video	7.698706
7	Photo / Cinema Equipment	7.455000
18	Sport and Leisure	7.354150
2	Consumer Electronics	7.066548
4	Construction Materials	6.978320
9	XNA	6.885879
24	Additional Service	6.730769
6	Gardening	6.723063
11	Homewares	6.706444
19	Medicine	6.196747
25	Education	5.882353
1	Furniture	5.860781
10	Clothing and Accessories	5.807427
13	Other	5.765921
12	Medical Supplies	5.584190
16	Tourism	4.444444
22	Fitness	4.268293
26	Animals	0.000000

Highest percentage of default cases are for the applicants who previously applied for Insurance and Vehicles.



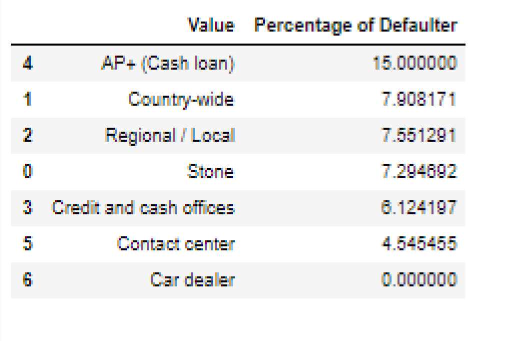
	Value	Percentage of Defaulter
1	walk-in	9.165550
0	XNA	7.665995
2	x-sell	6.038420

For Cards the default rate is highest.



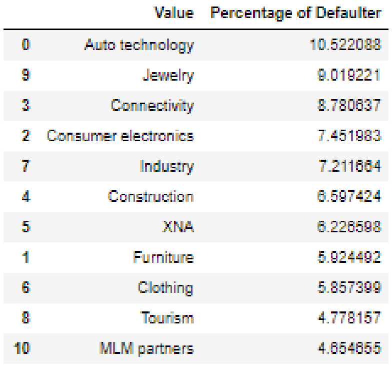
	Value	Percentage of Defaulter
1	walk-in	9.165550
0	XNA	7.665995
2	x-sell	6.038420

From all the walk-in applicants 9% defaulted in the current loan.



	Value	Percentage of Defaulter
4	AP+ (Cash loan)	15.000000
1	Country-wide	7.908171
2	Regional / Local	7.551291
0	Stone	7.294692
3	Credit and cash offices	6.124197
5	Contact center	4.545455
6	Car dealer	0.000000

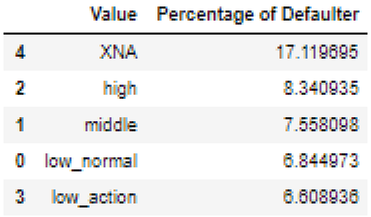
15% loan applicants defaulted for AP+ (Cash Loan)



	Value	Percentage of Defaulter
0	Auto technology	10.522088
9	Jewelry	9.019221
3	Connectivity	8.780637
2	Consumer electronics	7.451983
7	Industry	7.211664
4	Construction	6.597424
5	XNA	6.226598
1	Furniture	5.924492
6	Clothing	5.857399
8	Tourism	4.778157
10	MLM partners	4.654655

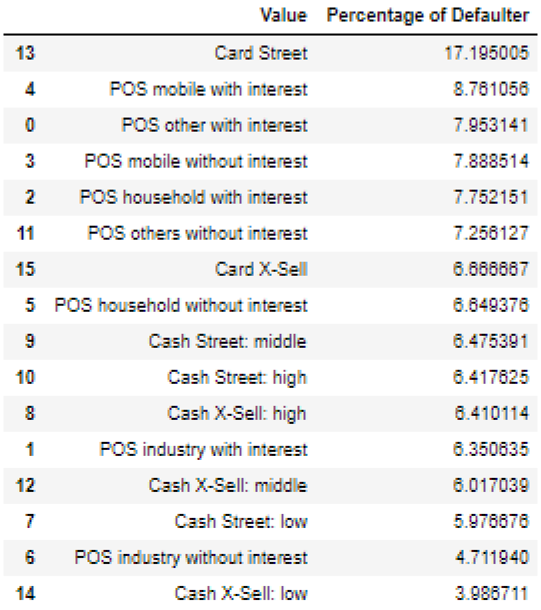
-In seller Industry “Auto technology” has highest rate of defaulter

-MLM partners has lowest number of defaulters



	Value	Percentage of Defaulter
4	XNA	17.119695
2	high	8.340935
1	middle	7.558098
0	low_normal	6.844973
3	low_action	6.608936

Defaulter percentage is highest where NAME_YIELD_GROUP is not known.



	Value	Percentage of Defaulter
13	Card Street	17.195005
4	POS mobile with interest	8.761056
0	POS other with interest	7.953141
3	POS mobile without interest	7.888514
2	POS household with interest	7.752151
11	POS others without interest	7.256127
15	Card X-Sell	6.666667
5	POS household without interest	6.649376
9	Cash Street: middle	6.475391
10	Cash Street: high	6.417625
8	Cash X-Sell: high	6.410114
1	POS industry with interest	6.350635
12	Cash X-Sell: middle	6.017039
7	Cash Street: low	5.976676
6	POS industry without interest	4.711940
14	Cash X-Sell: low	3.986711

Highest percentage of default cases is for Card Street.

Summary of Previous Application Data :

1. **There are feature columns in the dataset that are highly correlated to each other. Which means both will have similar impact on the target value. Those features can be removed before feeding this data to a model to avoid collinearity.**

2. **Feature columns with 50% or more missing data can be dropped.**

Following columns should be converted to integer.
DAYS_FIRST_DRAWING float64 DAYS_FIRST_DUE float64
DAYS_LAST_DUE_1ST_VERSION float64
DAYS_LAST_DUE float64 DAYS_TERMINATION float64

This categorical column has only 0 and 1 and hence can be converted into integer columns.
NFLAG_INSURED_ON_APPROVAL float64

1. **This dataset is highly imbalanced**
2. **The applicants whose previous loans were approved are more likely to pay the current loan in time, than the applicants whose previous loans were rejected. NAME_CONTRACT_STATUS is an important feature.**

- **7% of the previously approved loan applicants that defaulted in current loan**
- **90 % of the previously refused loan applicants that were able to pay current loan**

1. **‘SCO’, ‘LIMIT’ and ‘HC’ are the most common reasons for rejection.**
2. **Most of the people did not request insurance during the previous loan application.**

3. **For “Cards” the default percentage is highest (17%). ‘NAME_PORTFOLIO’ is an important feature for analyzing the ‘TARGET’ variable.**
4. **15% loan application defaulted for AP+ (Cash Loan). ‘CHANNEL_TYPE’ is an important feature for analyzing the ‘TARGET’ variable.**
5. **Highest percentage (17%) of default cases is for ‘Card Street’. ‘PRODUCT_COMBINATION’ is an important driving factor.**

Current Application Data Analysis:

For analyzing current application data, I have taken a different approach than that of previous application data. By taking a close look at the features, I could identify the features of different aspects of the loan applicant.

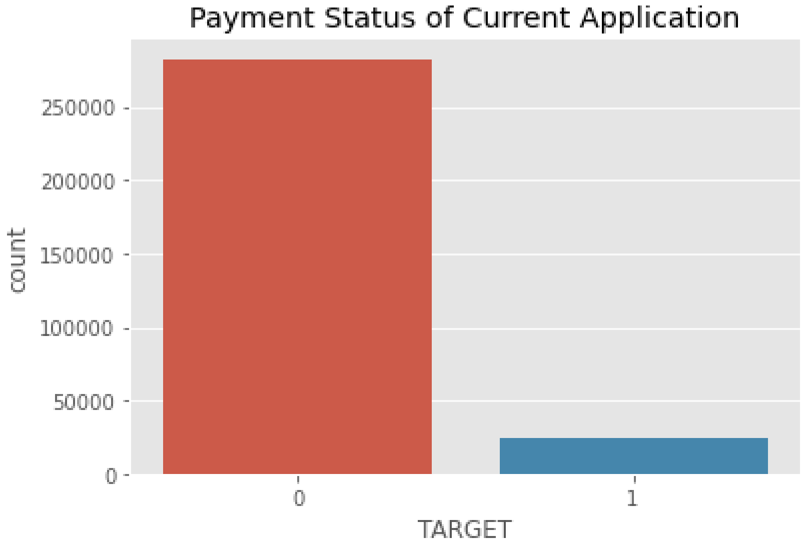
That’s why I have divided the features into small segments and analyzed them segment-wise using a smaller data-frame containing only relevant categories.

Data Cleaning, Missing Data Handling, Type casting are done segment-wise.

Plots and percentage wise Defaulter calculation are done segment-wise as well.

- **prev_ap_df (i.e. Previous application data) contains 37 features and 1670214 rows(Out of which 15 features are float64, 6 features are integer, 16 features are object data type)**
- **application_df (i.e. Current Application data) contains 121 features, 1 target variable, and 307511 rows(Out of which 65 features are float64, 41 features are integer, 16 features are object data type)**

Checking Data Imbalance:

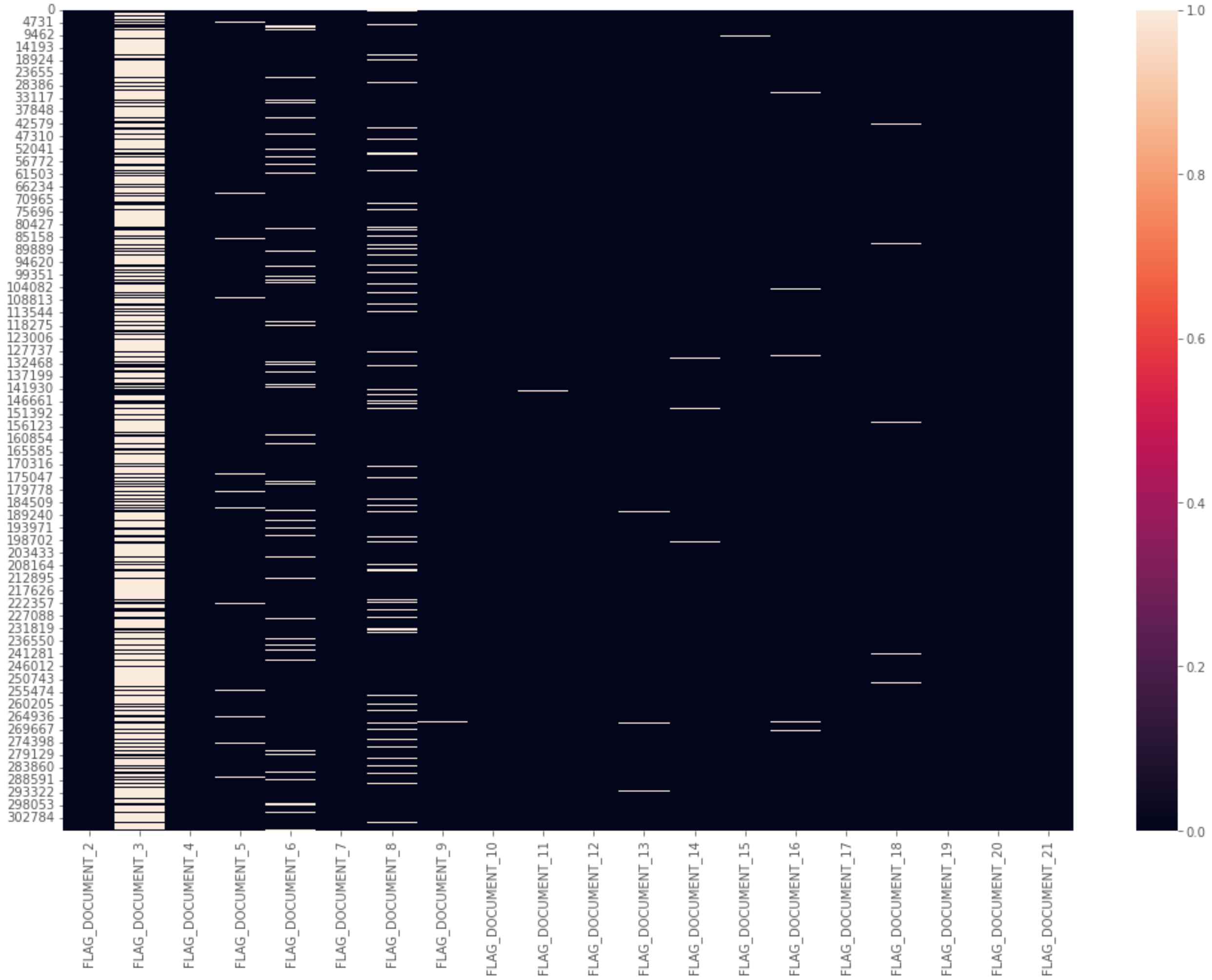


This data is highly imbalanced as the number of defaulters is very less in total population. Data Imbalance Ratio

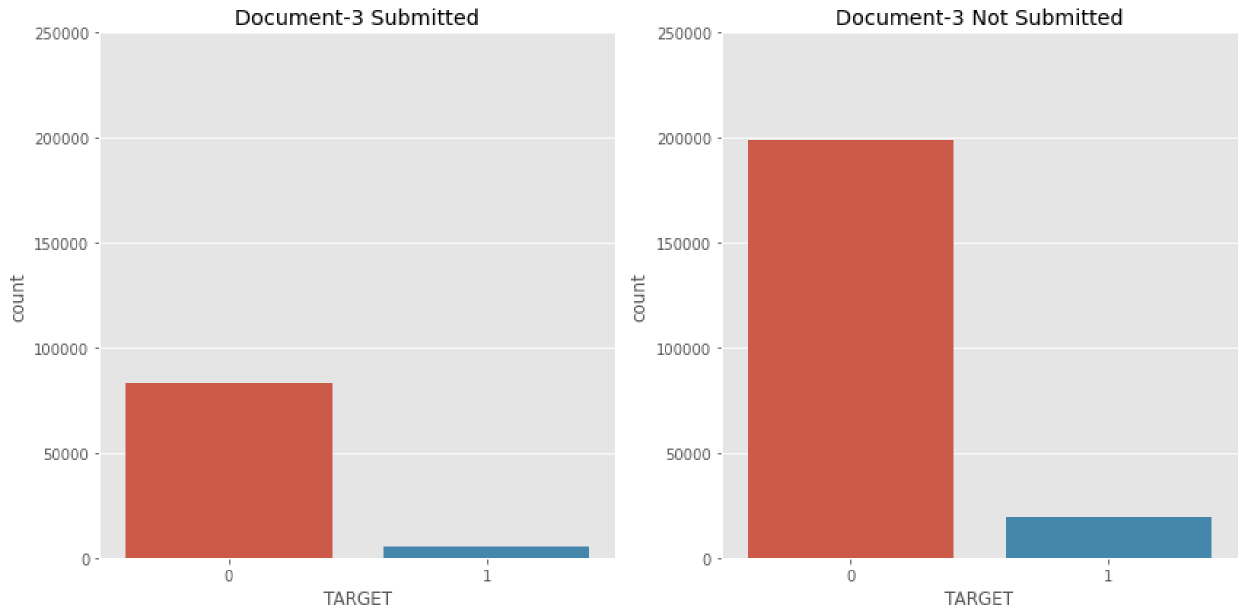
Defaulter : Non-Defaulter = 8 : 92 = 2 : 23

Segment 1: Documents Submitted by Applicant

Here we are analyzing ‘FLAG_DOCUMENT_2′,’FLAG_DOCUMENT_3′,…,’FLAG_DOCUMENT_21’ columns. Our goal is to understand the trend of document submission and identify the impact on the TARGET variable(if any).

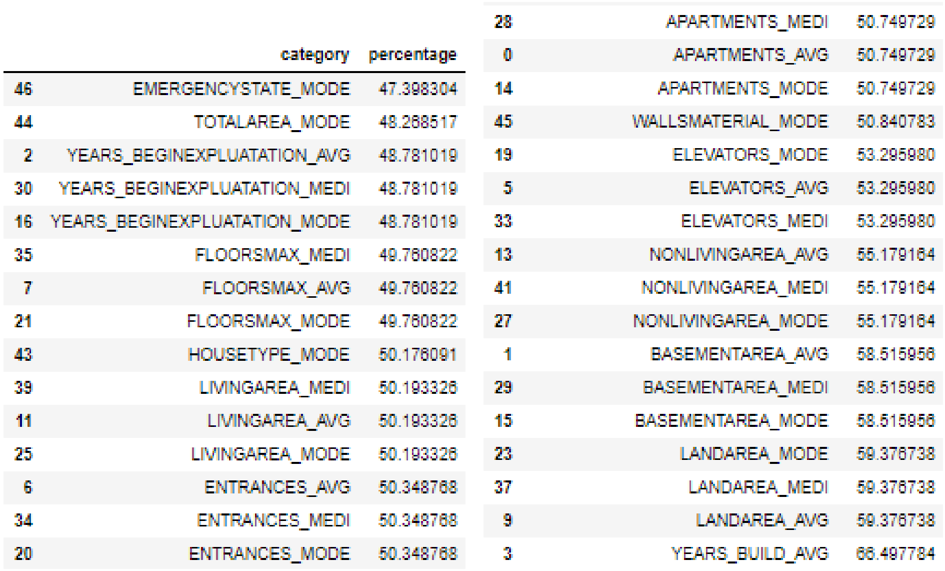


- **The heatmap suggests that all of the documents except Document 3 were not provided by applicants in the majority of the cases.**
- **Hence we can assume all the documents (except document 3) will not contribute towards analyzing the data. Hence all these columns can be dropped.**



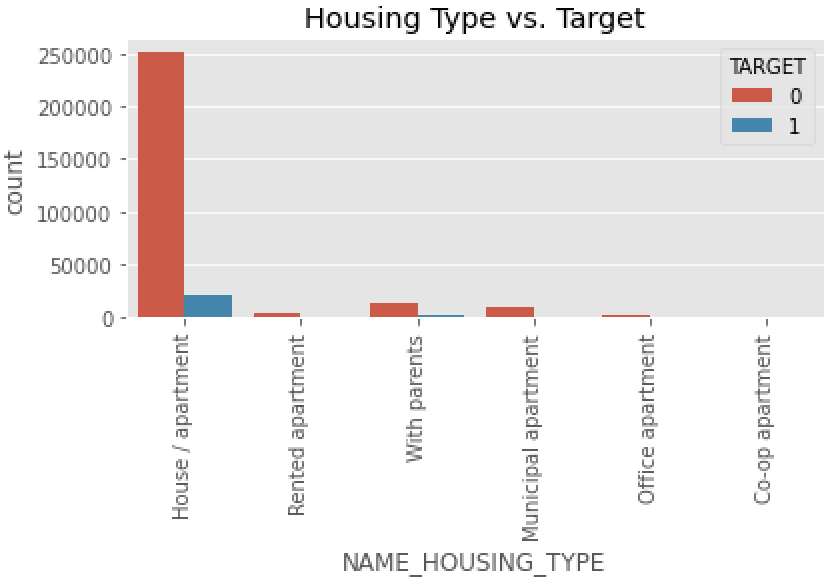
- **FLAG_DOCUMENT_3 is showing a similar trend for both non-defaulters and defaulters.**
- **Hence, this column can be dropped.**

Segment 2 : Housing Information of Applicant



	category	percentage
46	EMERGENCYSTATE_MODE	47.398304
44	TOTALAREA_MODE	48.268517
2	YEARS_BEGINEXPLUATATION_AVG	48.781019
30	YEARS_BEGINEXPLUATATION_MEDI	48.781019
16	YEARS_BEGINEXPLUATATION_MODE	48.781019
35	FLOORSMAX_MEDI	49.760822
7	FLOORSMAX_AVG	49.760822
21	FLOORSMAX_MODE	49.760822
43	HOUSETYPE_MODE	50.176091
39	LIVINGAREA_MEDI	50.193326
11	LIVINGAREA_AVG	50.193326
25	LIVINGAREA_MODE	50.193326
6	ENTRANCES_AVG	50.348768
34	ENTRANCES_MEDI	50.348768
20	ENTRANCES_MODE	50.348768
28	APARTMENTS_MEDI	50.749729
0	APARTMENTS_AVG	50.749729
14	APARTMENTS_MODE	50.749729
45	WALLSMATERIAL_MODE	50.840783
19	ELEVATORS_MODE	53.295980
5	ELEVATORS_AVG	53.295980
33	ELEVATORS_MEDI	53.295980
13	NONLIVINGAREA_AVG	55.179164
41	NONLIVINGAREA_MEDI	55.179164
27	NONLIVINGAREA_MODE	55.179164
1	BASEMENTAREA_AVG	58.515956
29	BASEMENTAREA_MEDI	58.515956
15	BASEMENTAREA_MODE	58.515956
23	LANDAREA_MODE	59.376738
37	LANDAREA_MEDI	59.376738
9	LANDAREA_AVG	59.376738
3	YEARS_BUILD_AVG	66.497784

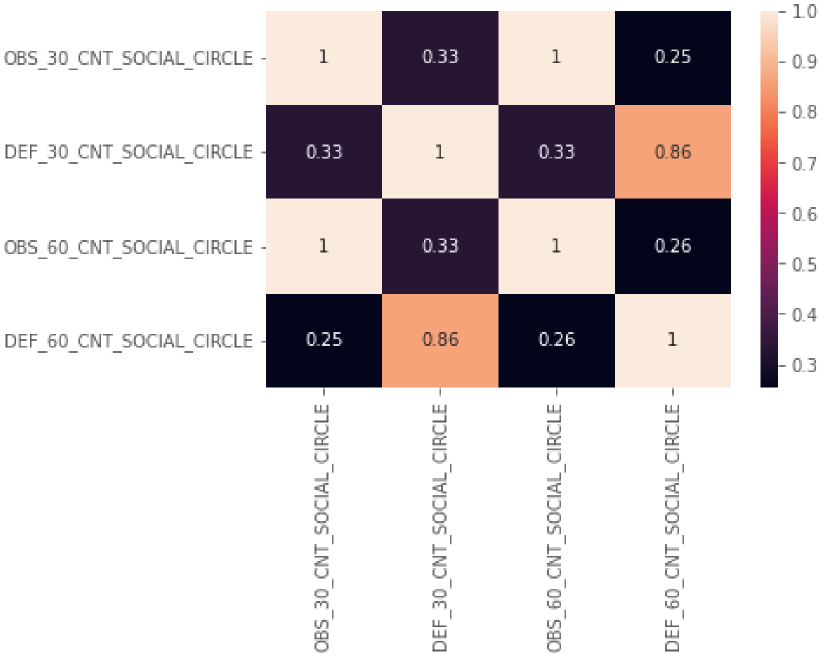
All of the features have a very high (47-70%) missing data percentage. Hence all these features can be dropped.



	Value	Percentage of Defaulter
1	Rented apartment	12.313051
2	With parents	11.698113
3	Municipal apartment	8.539748
5	Co-op apartment	7.932264
0	House / apartment	7.795711
4	Office apartment	6.572411

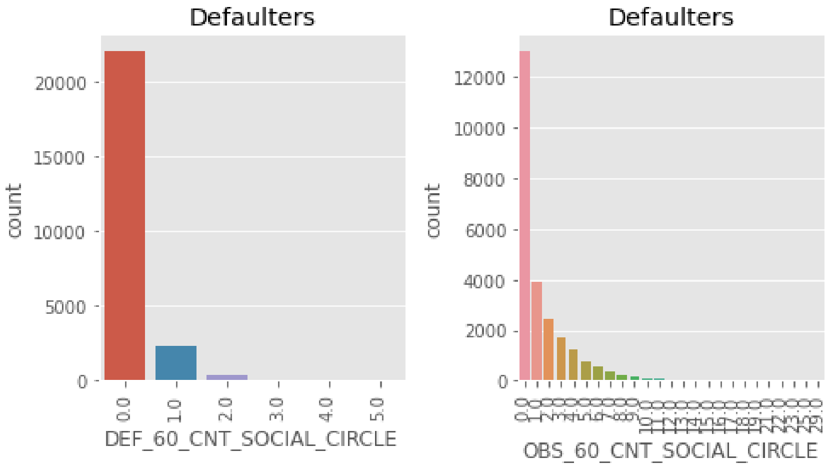
- Most of the applicants live in House/Apartment
- Applicants living with their parents or in rented apartment have a higher rate of default.

Segment 3 :Social Circle Info



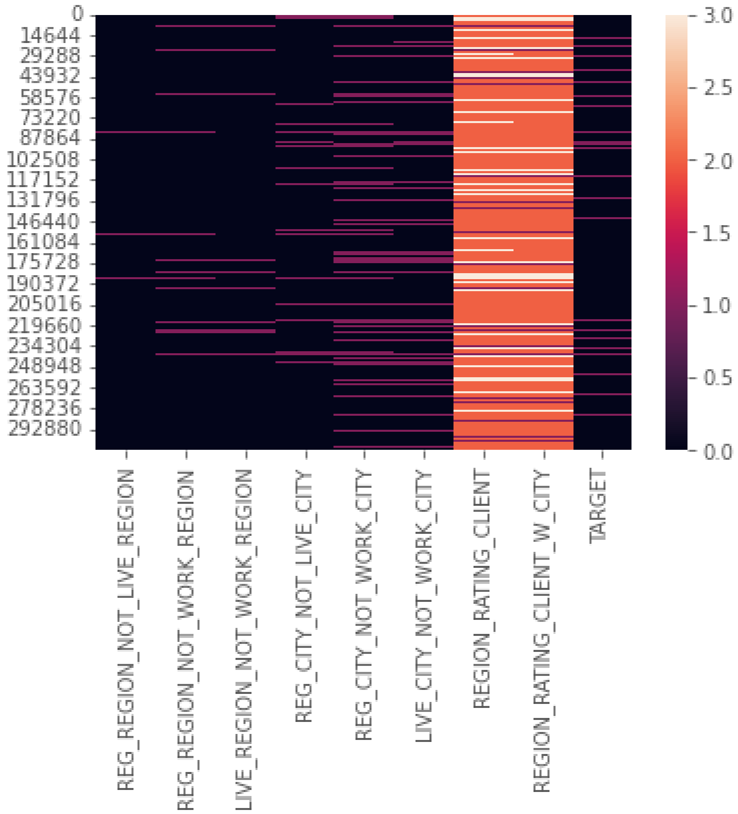
- **DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE are highly correlated**

- **OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE are identical columns**

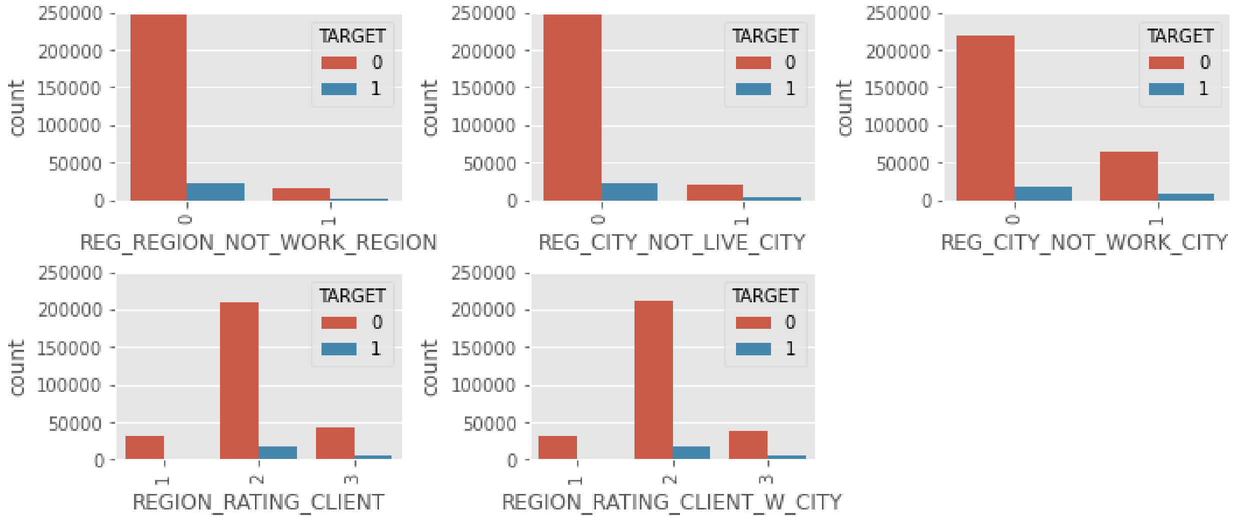


For defaulter and non-defaulter ‘DEF_60_CNT_SOCIAL_CIRCLE’,’OBS_60_CNT_SOCIAL_CIRCLE’ features show a similar trend.

Segment 4 : Region related data



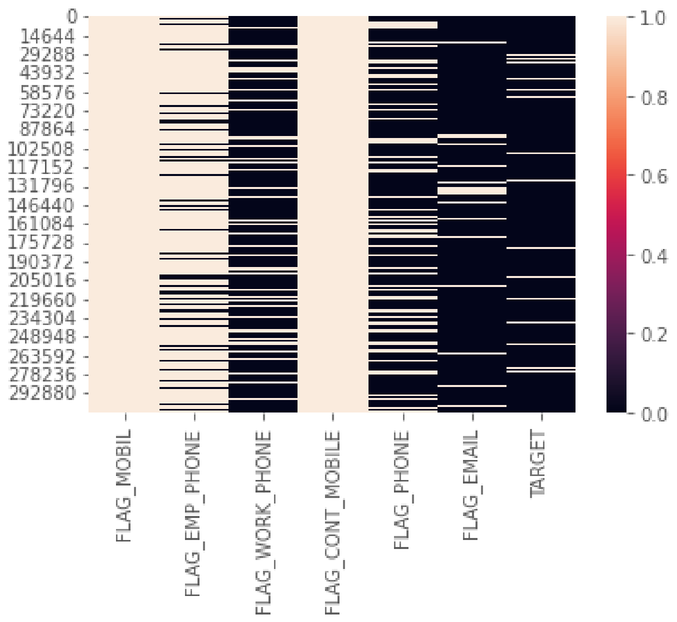
- **All the features are labeled as 0 and 1**
- **REG_REGION_NOT_LIVE_REGION mostly contains 0, hence can be removed**
- **REG_REGION_NOT_WORK_REGION, LIVE_REGION_NOT_WORK_REGION columns are identical, hence one of them can be removed**
- **REG_CITY_NOT_WORK_CITY, LIVE_CITY_NOT_WORK_CITY columns are identical, hence one of them can be removed**



- **Defaulter rate is highest when REG_REGION_NOT_WORK_REGION=0 i.e. permanent address and working address is same**

- **Highest Applicants have Region rating of 2**

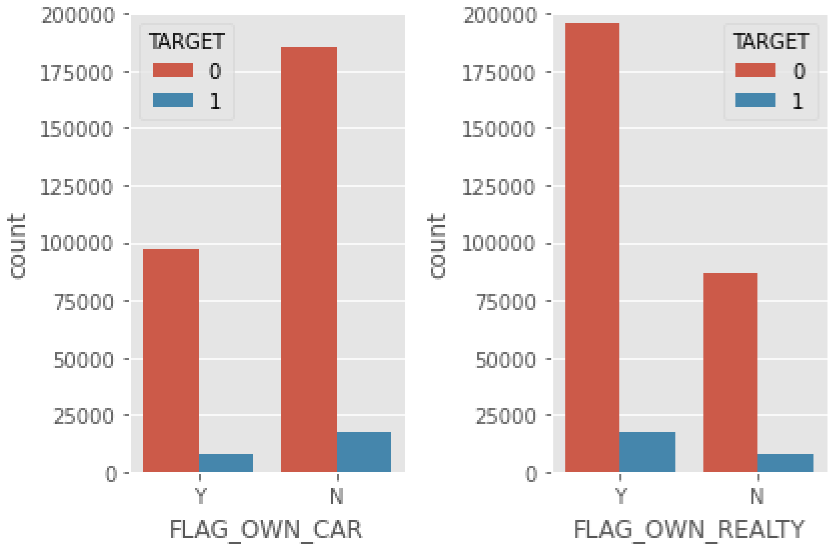
Segment 5: Contact Related Info



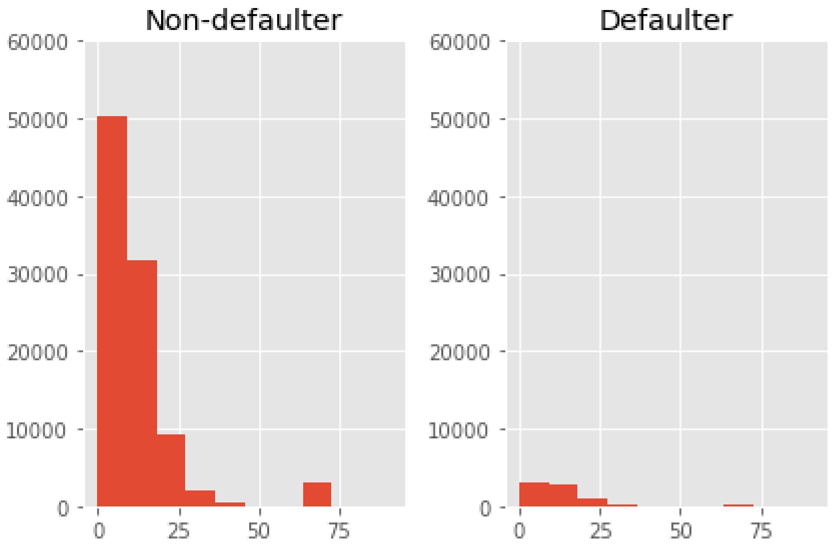
- **All the features in contact_df are categorical (0 and 1)**

- **As there is no similarity of patterns of TARGET value with the features, we are assuming the features are not useful for analysis.**
- **Hence all of the features can be removed**

Segment 6: Asset Details

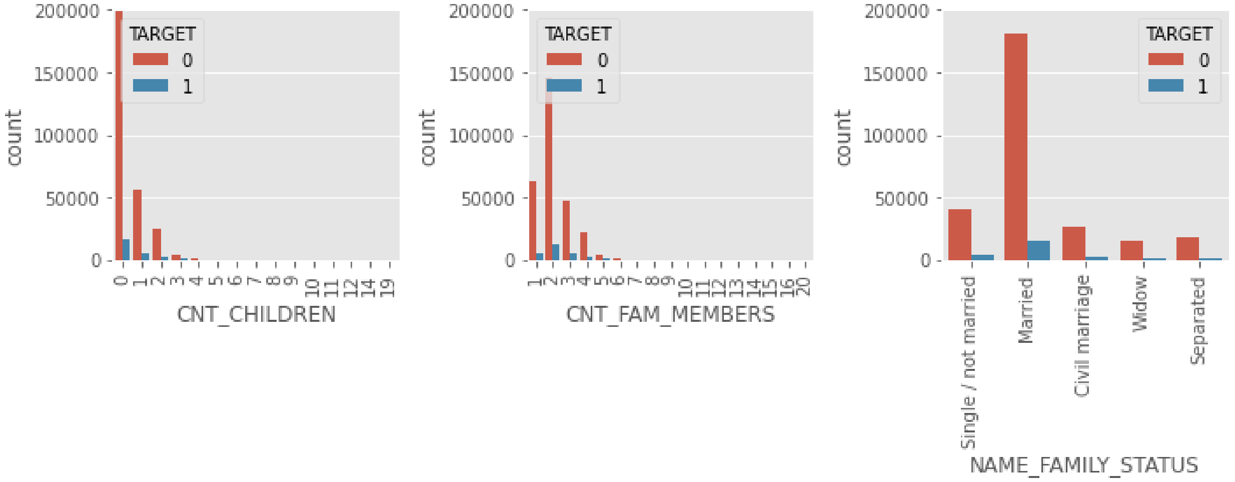


- Most of the applicants own realty
- Most of the applicants do not own cars
- People not owning reality and car and have a slightly higher default rate than the people who own reality and car



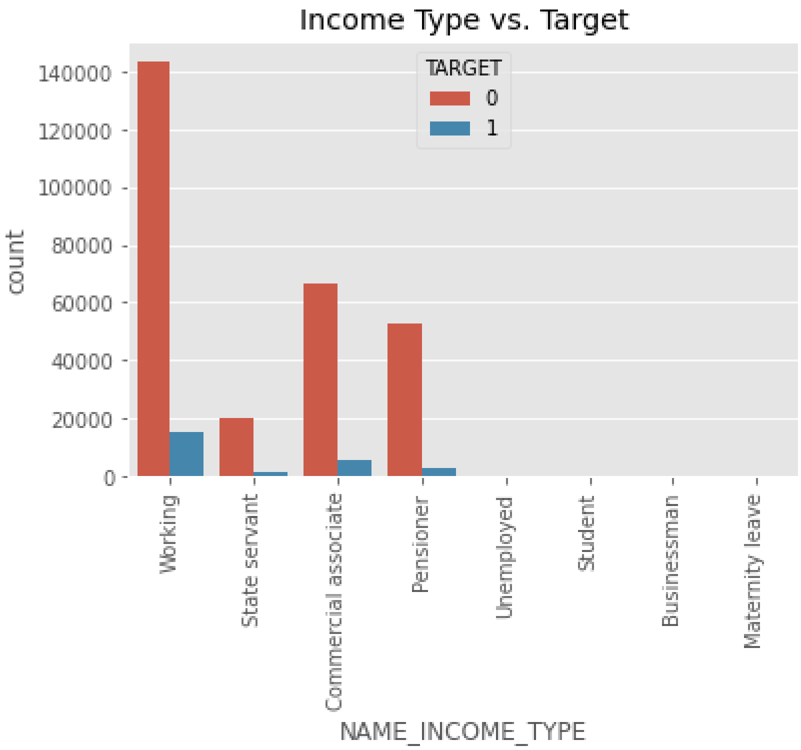
- Default or not, most applicants have a car age between 0-25 years.
- Since for both target values, trend is similar, this feature can dropped.

Segment 7: Family Related Info



- Default rate is highest for Civil Marriage and Single applicants
- Most of the applicants are married (and/or) no children (and/or) 2 family members.
- Applicants with relatively more children (and/or) family members have a higher default percentage.
- For some of the cases where the count of children/family members is high, and the default rate is very high or very low. These cases cannot be taken as a conclusion as the number of applicants having a large family is very low.

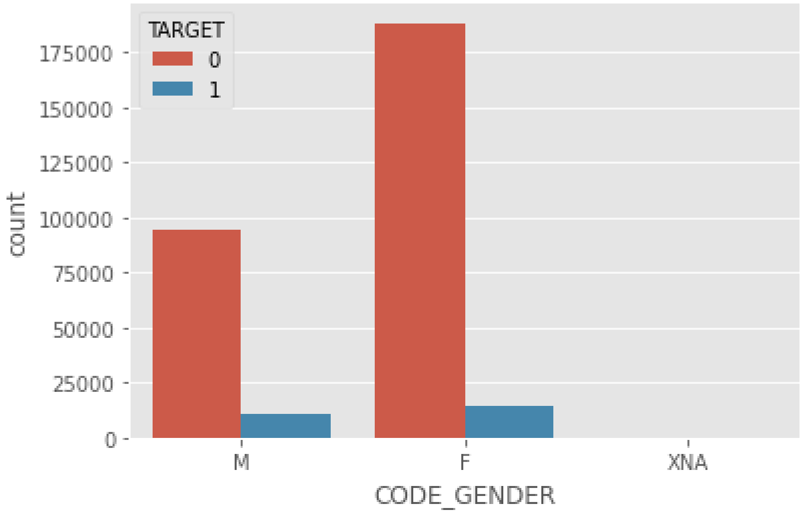
Segment 8: Education and Occupation Info



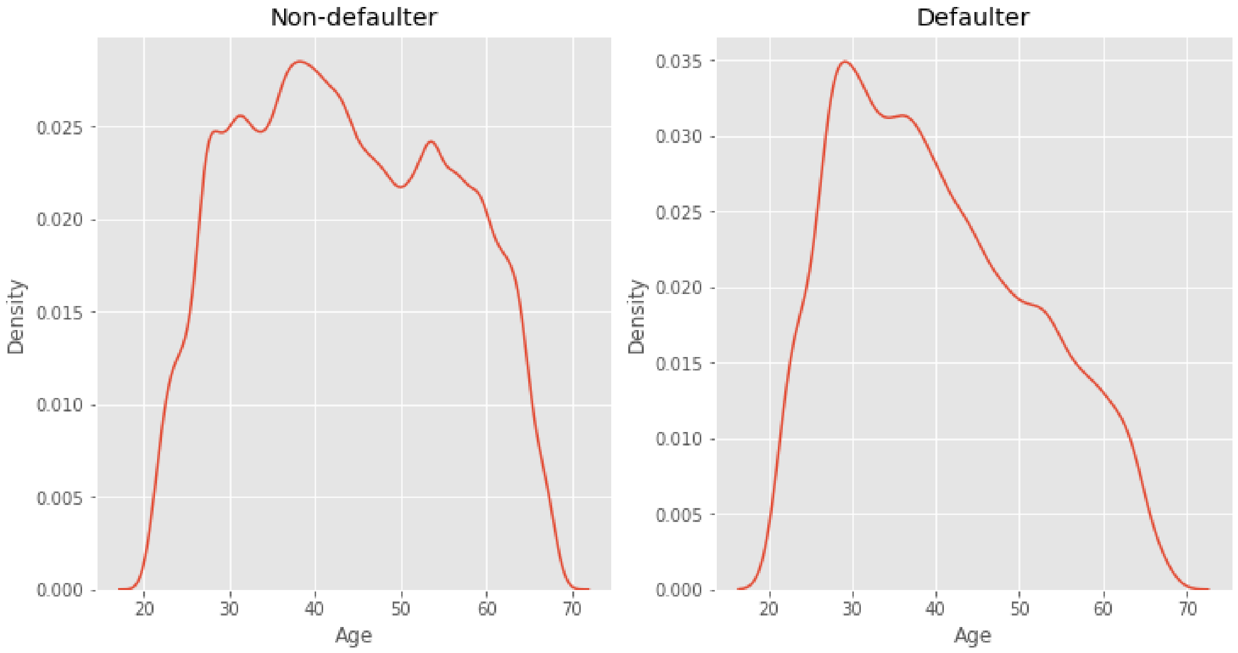
Value	Percentage of Defaulter

7	Maternity leave	40.0
4	Unemployed	36.363636
0	Working	9.588472
2	Commercial associate	7.484466
1	State servant	5.754965
3	Pensioner	5.386366
5	Student	0.0
6	Businessman	0.0

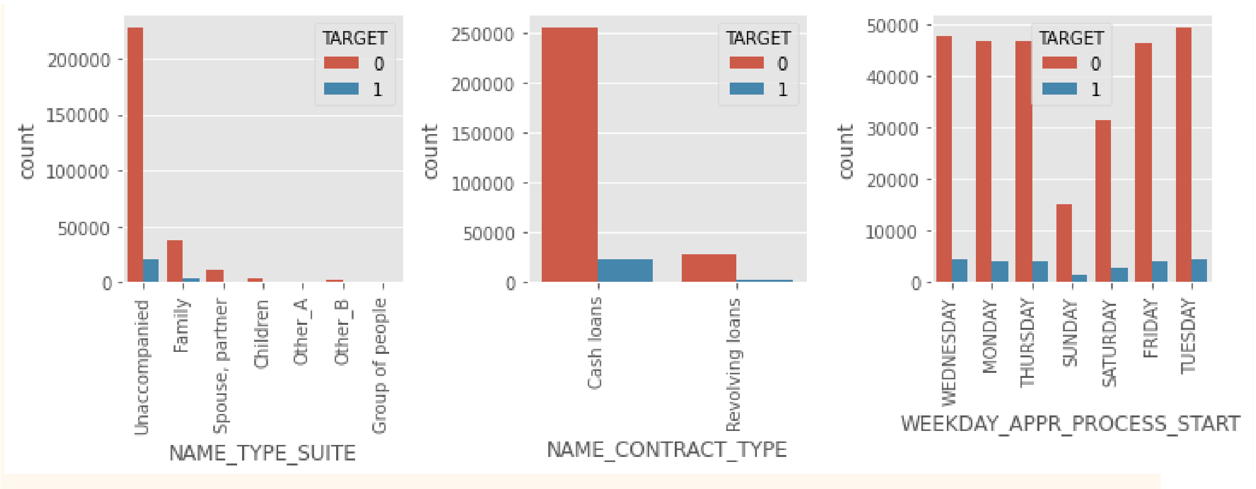
-Most of the applicants are working.
-Applicants on Maternity Leave and Unemployed has highest percentage of Defaulter
-Businessmen have the lowest (0) percentage of Default. However, applicants of income type(‘Unemployed’, ‘Student’, ‘Businessman’, ‘Maternity leave’) are very few in the dataset to contribute to the analysis.
-Applicants having “Lower secondary” education have the highest percentage of Default.
-Low skilled laborers have very high rate of defaulters in comparison to other occupations



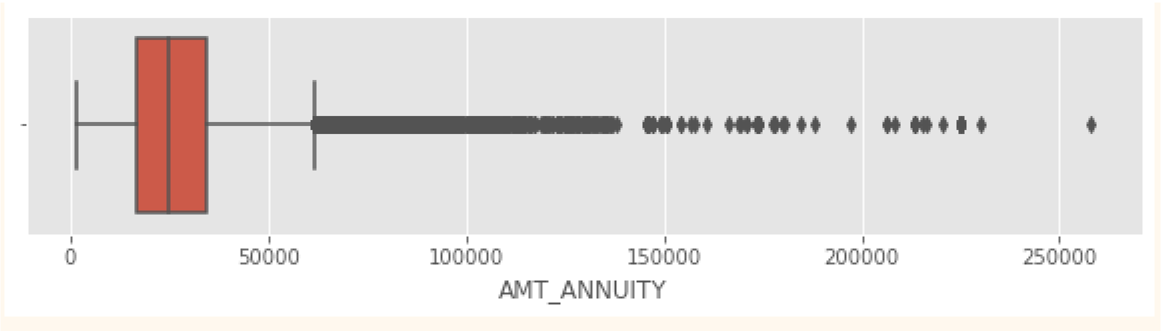
- Female applicants are more than male applicants
- Defaulter percentage is higher for male applicants

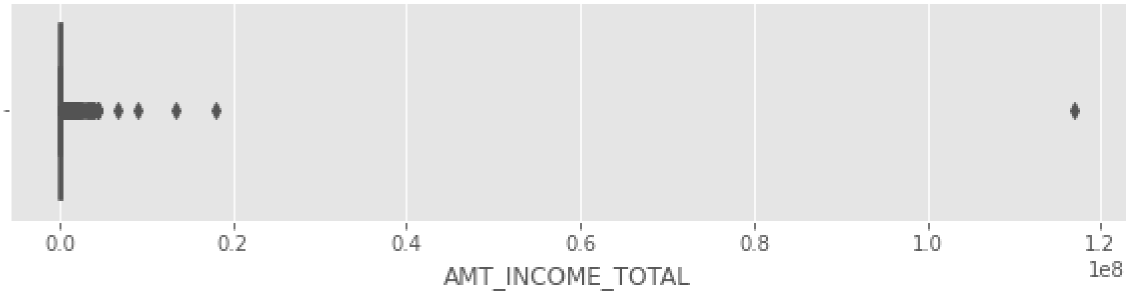


- People of age 30 have higher default rate
- Default cases are less for applicants more than 40 years old.



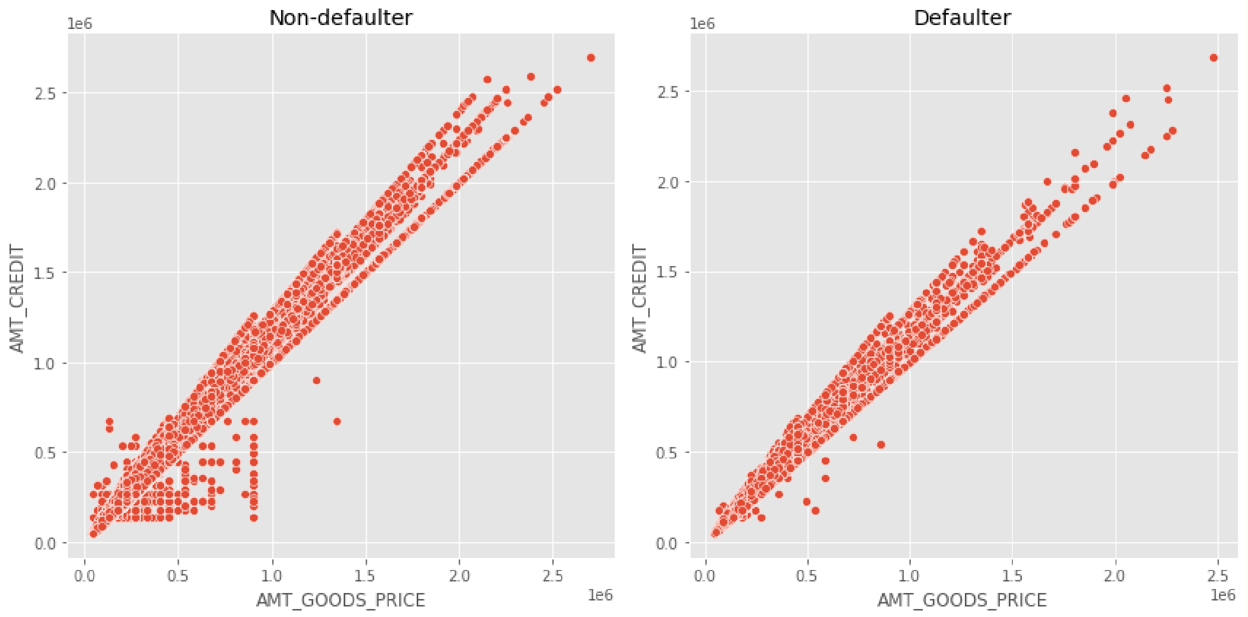
- Most of the applicants are unaccompanied while applying for loan
- Number Cash loans is quite higher than Revolving Loans
- All weekdays have similar number of applicants than weekend(Saturday and Sunday)



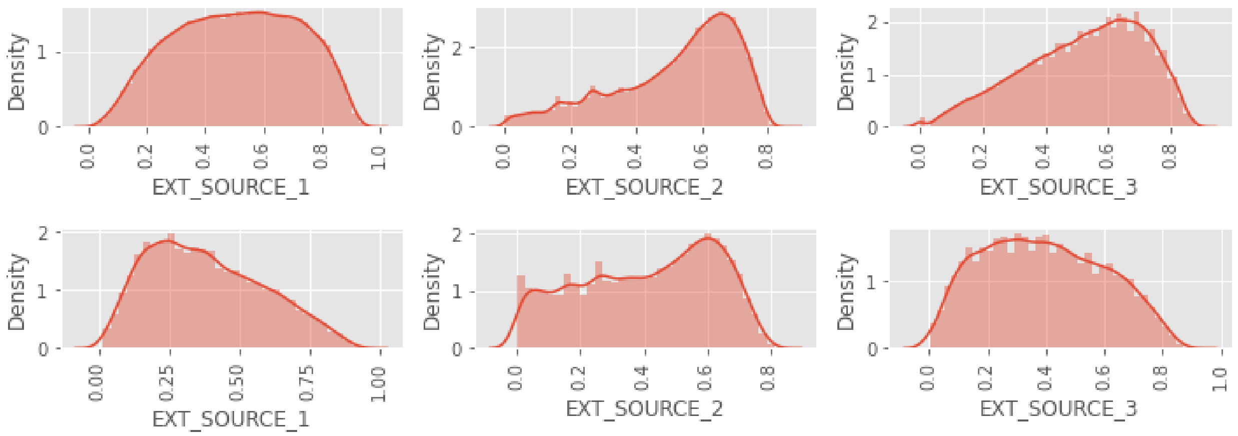


-Boxplot is showing the outliers for income and annuity, there are few entries having very large annuity and income than others.

-Considering these entries will mislead the average income of the entire population and further analysis. Excluding values outside 99 percentile for AMT_ANNUITY and AMT_INCOMRE_TOTAL.

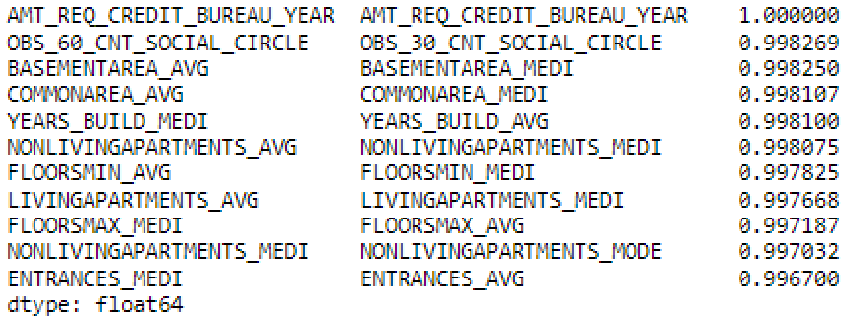


- AMT_CREDIT and AMT_GOODS_PRICE have linear relations.
- For lower range of AMT_CREDIT and AMT_GOODS_PRICE, amount of defaulters is less than that of non-defaulters



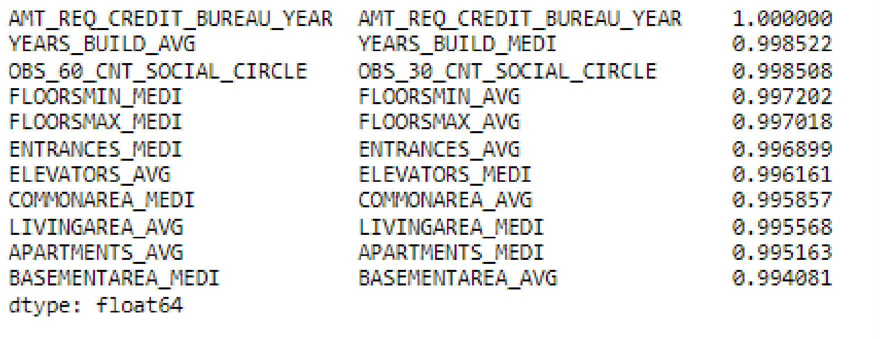
‘EXT_SOURCE_1’ and ‘EXT_SOURCE_3’ have very different distributions for defaulters and non-defaulters. It might be useful to know what these variables actually mean to derive important information about applicants.

Top 10 correlation for Defaulters



```
AMT_REQ_CREDIT_BUREAU_YEAR   AMT_REQ_CREDIT_BUREAU_YEAR     1.000000
OBS_60_CNT_SOCIAL_CIRCLE     OBS_30_CNT_SOCIAL_CIRCLE       0.998269
BASEMENTAREA_AVG             BASEMENTAREA_MEDI              0.998250
COMMONAREA_AVG               COMMONAREA_MEDI                0.998107
YEARS_BUILD_MEDI             YEARS_BUILD_AVG                0.998100
NONLIVINGAPARTMENTS_AVG      NONLIVINGAPARTMENTS_MEDI       0.998075
FLOORSMIN_AVG                FLOORSMIN_MEDI                 0.997825
LIVINGAPARTMENTS_AVG         LIVINGAPARTMENTS_MEDI          0.997668
FLOORSMAX_MEDI               FLOORSMAX_AVG                  0.997187
NONLIVINGAPARTMENTS_MEDI     NONLIVINGAPARTMENTS_MODE       0.997032
ENTRANCES_MEDI               ENTRANCES_AVG                  0.996700
dtype: float64
```

Top 10 Correlation for Non-defaulters



```
AMT_REQ_CREDIT_BUREAU_YEAR   AMT_REQ_CREDIT_BUREAU_YEAR     1.000000
YEARS_BUILD_AVG              YEARS_BUILD_MEDI               0.998522
OBS_60_CNT_SOCIAL_CIRCLE     OBS_30_CNT_SOCIAL_CIRCLE       0.998508
FLOORSMIN_MEDI               FLOORSMIN_AVG                  0.997202
FLOORSMAX_MEDI               FLOORSMAX_AVG                  0.997018
ENTRANCES_MEDI               ENTRANCES_AVG                  0.996899
ELEVATORS_AVG                ELEVATORS_MEDI                 0.996161
COMMONAREA_MEDI              COMMONAREA_AVG                 0.995857
LIVINGAREA_AVG               LIVINGAREA_MEDI                0.995568
APARTMENTS_AVG               APARTMENTS_MEDI                0.995163
BASEMENTAREA_MEDI            BASEMENTAREA_AVG               0.994081
dtype: float64
```

Top 5 important columns

-Family Info: (Important driving features : **‘CNT_FAM_MEMBERS’, ‘CNT_CHILDREN’**) i. Most of the applicants are married (and/or) no children (and/or) 2 family members. ii. Applicants with relatively more children (and/or) family members have a higher default percentage. (For some of the cases where the count of children/family members is high, and the default rate is very high or very low. These cases cannot be considered for analysis as the number of applicants having a large family is very low.)

- Education and Occupation Info: (Important driving features :**‘NAME_INCOME_TYPE’, ‘OCCUPATION_TYPE’**)
- Most of the applicants are working.
- Applicants on Maternity Leave and Unemployed has highest percentage of Defaulter
- Businessmen have the lowest (0) percentage of Default. However, applicants of income type(‘Unemployed’, ‘Student’, ‘Businessman’, ‘Maternity leave’) are very few in the dataset to contribute to the analysis.

CODE_GENDER

- Female applicants are more than male applicants
- Defaulter percentage is higher for male applicants
- XNA values can be replaced with “Unknown”

DAYS_BIRTH

- A derived column ‘Age’ from this gave useful information.
- People of age 25-35 have higher default rate
- Default cases are less for applicants more than 40 years old.

‘EXT_SOURCE_1’ and ‘EXT_SOURCE_3’ have very different distributions for defaulters and non-defaulters. These can be important features.

Summary :

1. This data is highly imbalanced as the number of defaulters is very less in total population.

‘CNT_FAM_MEMBERS’, ‘CNT_CHILDREN’,’NAME_INCOME_TYPE’, ‘OCCUPATION_TYPE’,CODE_GENDER, ‘EXT_SOURCE_1’ and ‘EXT_SOURCE_3’ are some of the important driving factors.

1. Documents : Considered features ‘FLAG_DOCUMENT_2′,’FLAG_DOCUMENT_3′,…,’FLAG_DOCUMENT_21’ for this segment. Majority of the applicants did not submit any documents apart from DOCUMENT_3. FLAG_DOCUMENT_3 has a similar

impact on defaulters and non-defaulters. Hence these columns can be dropped.

2. Housing: All of the features considered have a very high (47-70%) missing data percentage. Hence all these features can be dropped. Plot of ‘NAME_HOUSING_TYPE’ vs ‘TARGET’ shows that

i. Most of the applicants live in House/Apartment ii. Applicants living with their parents or in rented apartments have a higher rate of default.

1. Social Circle Info: The features show similar trend for defaulters and non-defaulters, can be dropped.
2. Regional Info: Defaulter rate is highest when REG_REGION_NOT_WORK_REGION=0 i.e. permanent address and working address is same
3. Contact Info : Considered ‘FLAG_MOBIL’,’FLAG_EMP_PHONE’ etc. for this segment. No impact on Target, features can be dropped.
4. Asset Info : i. Most of the applicants own realty ii. Most of the applicants do not own cars iii. People not owning reality and car and have a slightly higher default rate than the people who own reality and car

8. Analyzing impact of car features on price and profitability

METRICS OVERVIEW LAST 7 DAYS

KPIS	Last 7 Days
Reach	1.9M
Frequency	0.9
Social impressions	29,823
Actions	281,083
Clicks	39,190
CTR	1.80%
Total spent	$ 4,251.10
CPA	$ 0.01
CPC	$ 0.11

Project Description:

The automotive industry is evolving rapidly with a focus on fuel efficiency, sustainability, and innovation. It's essential to understand the factors that drive consumer demand for cars. The trend towards electric and hybrid vehicles exists, but traditional gasoline-powered cars still dominate the market.

To optimize pricing and product development decisions to maximize profitability while meeting consumer demand. By analyzing a car's features, market category, and pricing, manufacturers can develop a pricing strategy that balances consumer demand with profitability. Identifying popular features and categories to focus on for future product development efforts, improve competitiveness, and increase profitability over time.

The dataset contains the information of various cars information

- Number of observations: 11,159

- **Number of variables: 16**
- **File type: CSV (Comma Separated Values)**

Approach:

Analysis Approach: The analysis will include identifying missing data and using appropriate methods to handle it, identifying outliers and data imbalances. For analysis I have used descriptive statistics.firstly I clean the dataset and then finding the insights from data.

Missing Data: Missing data will be identified and replaced with an appropriate value or removed, depending on the context. Outliers: Outliers will be identified and explained in business terms, but will not necessarily be removed.

Visualization: The analysis will include visualizations to summarize important results, and insights will explain why variables are important for differentiating clients with payment difficulties from all other cases.

Tech-Stack Used:

➢ Microsoft excel 365

I have used Microsoft excel to do this project, because excel provide a all features that are used for data analysis and finding insights from data.

Excel provides a better visualization graph.

➢ Microsoft powerpoint

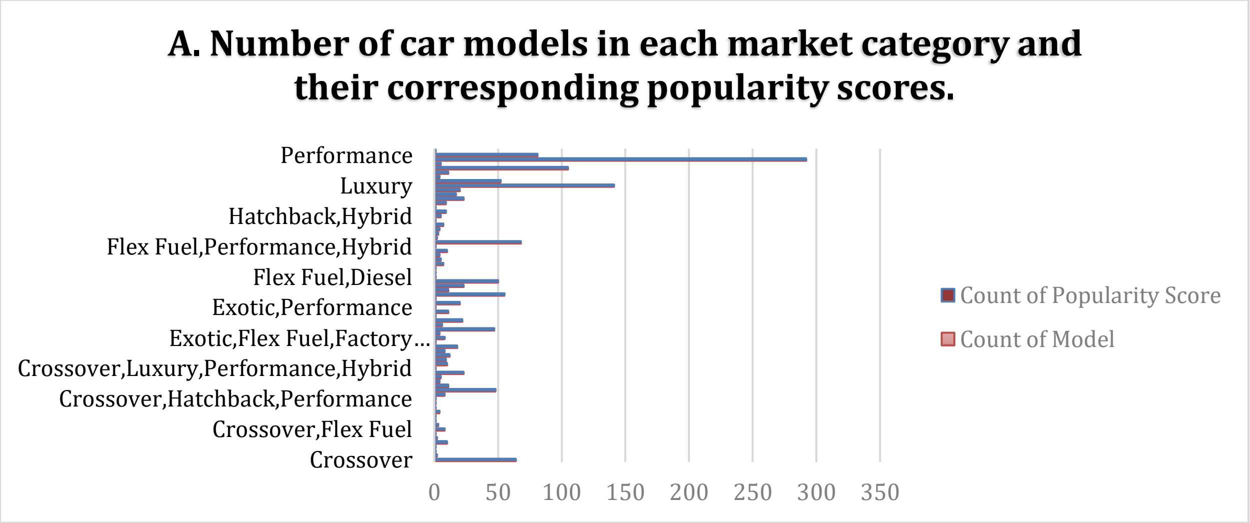
For making the project report I have used Microsoft powerpoint.

Insights:

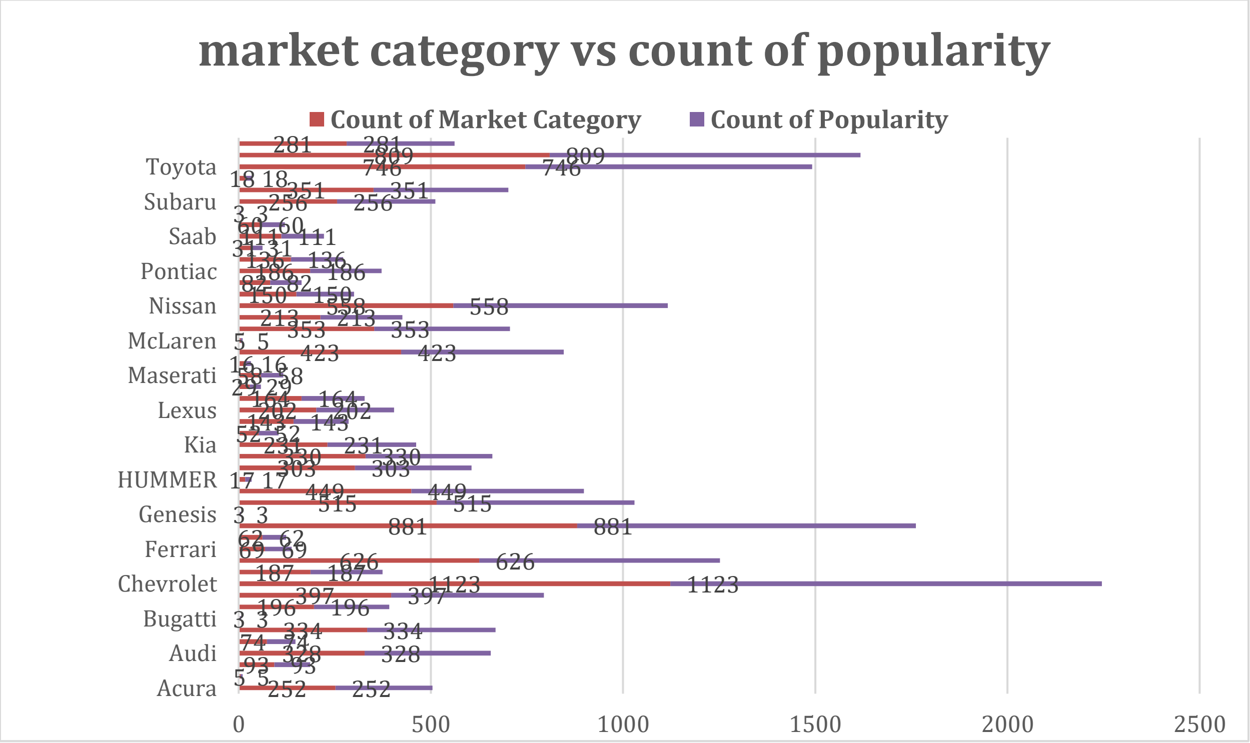
Insight Required: How does the popularity of a car model vary across different market categories?

- **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.

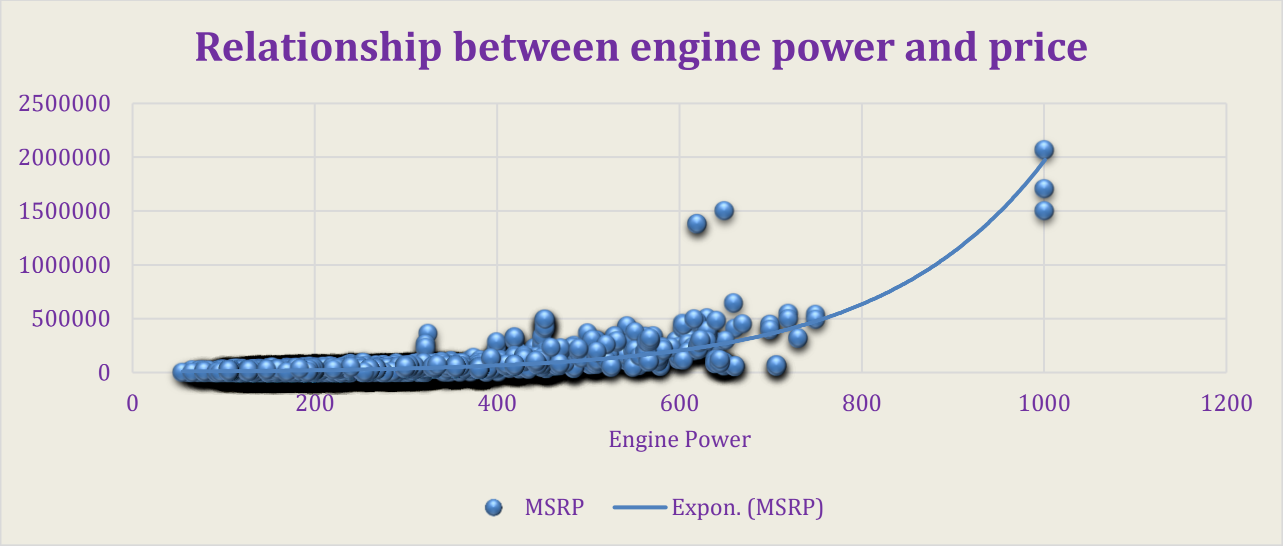


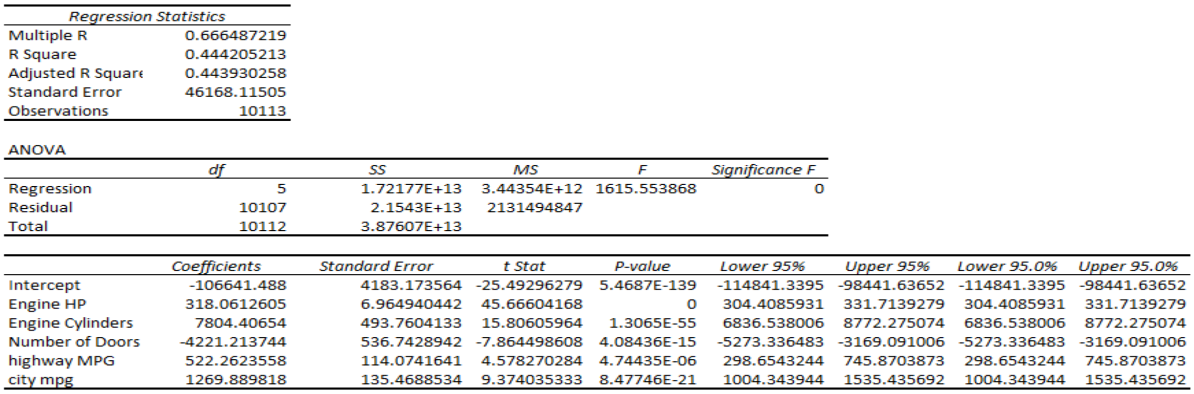
A. Number of car models in each market category and their corresponding popularity scores.



market category vs count of 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.



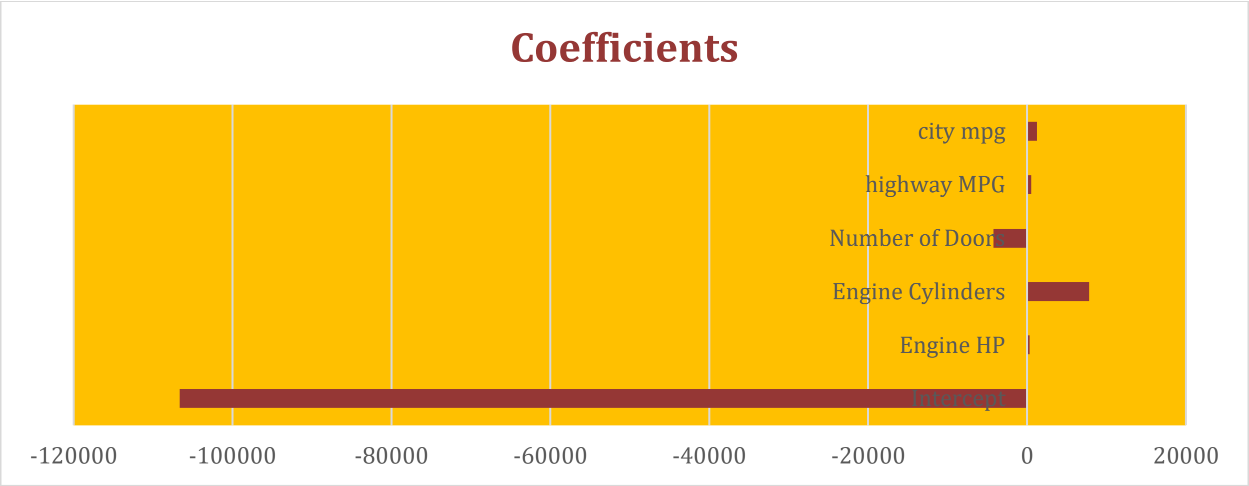


Regression Statistics	
Multiple R	0.666487219
R Square	0.444205213
Adjusted R Square	0.443930258
Standard Error	46168.11505
Observations	10113

ANOVA

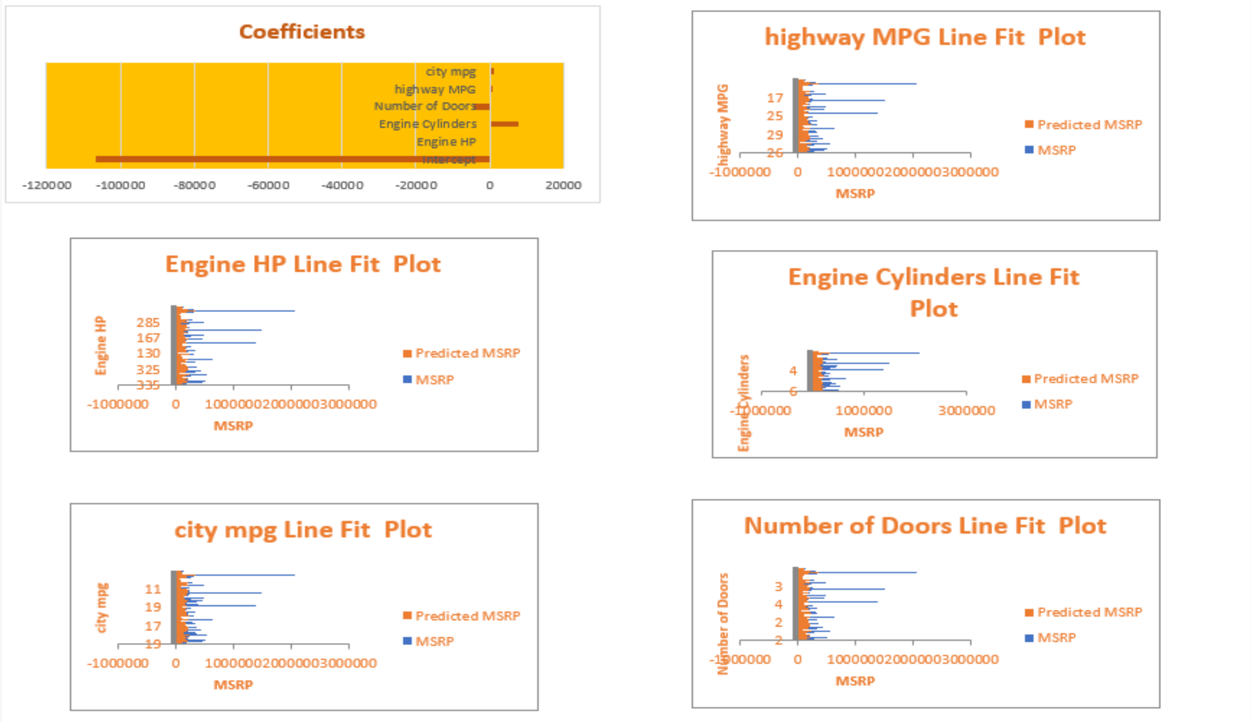
	df	SS	MS	F	Significance F
Regression	5	1.72177E+13	3.44354E+12	1615.553868	0
Residual	10107	2.1543E+13	2131494847		
Total	10112	3.87607E+13			

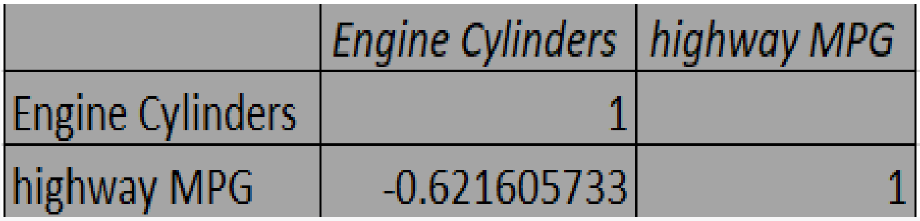
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-106641.488	4183.173564	-25.49296279	5.4687E-139	-114841.3395	-98441.63652	-114841.3395	-98441.63652
Engine HP	318.0612605	6.964940442	45.66604168	0	304.4085931	331.7139279	304.4085931	331.7139279
Engine Cylinders	7804.40654	493.7604133	15.80605964	1.3065E-55	6836.538006	8772.275074	6836.538006	8772.275074
Number of Doors	-4221.213744	536.7428942	-7.864498608	4.08436E-15	-5273.336483	-3169.091006	-5273.336483	-3169.091006
highway MPG	522.2623558	114.0741641	4.578270284	4.74435E-06	298.6543244	745.8703873	298.6543244	745.8703873
city mpg	1269.889818	135.4688534	9.374035333	8.47746E-21	1004.343944	1535.435692	1004.343944	1535.435692



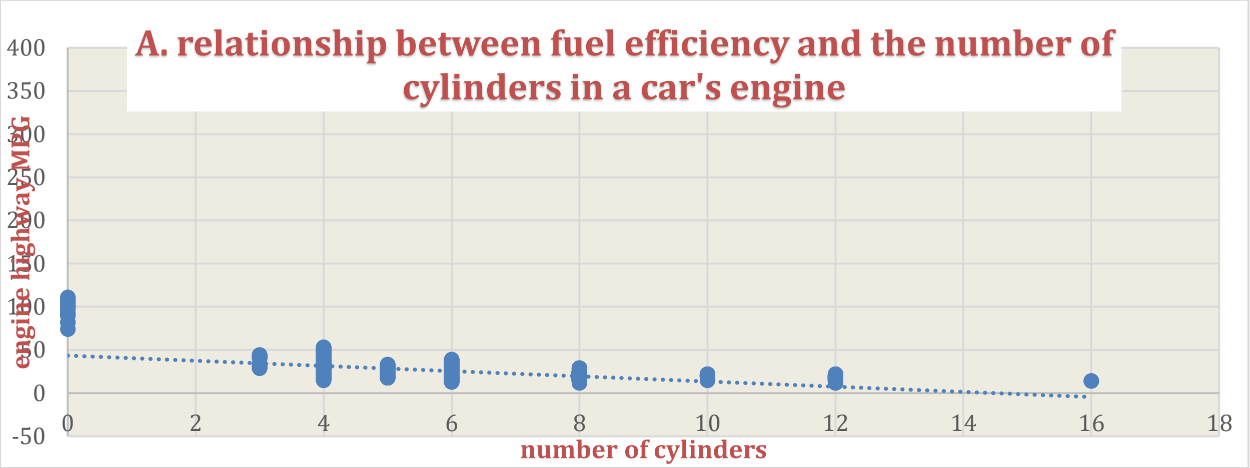
Insight Required: Which car features are most important in determining a car's price?

- ➢ **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.





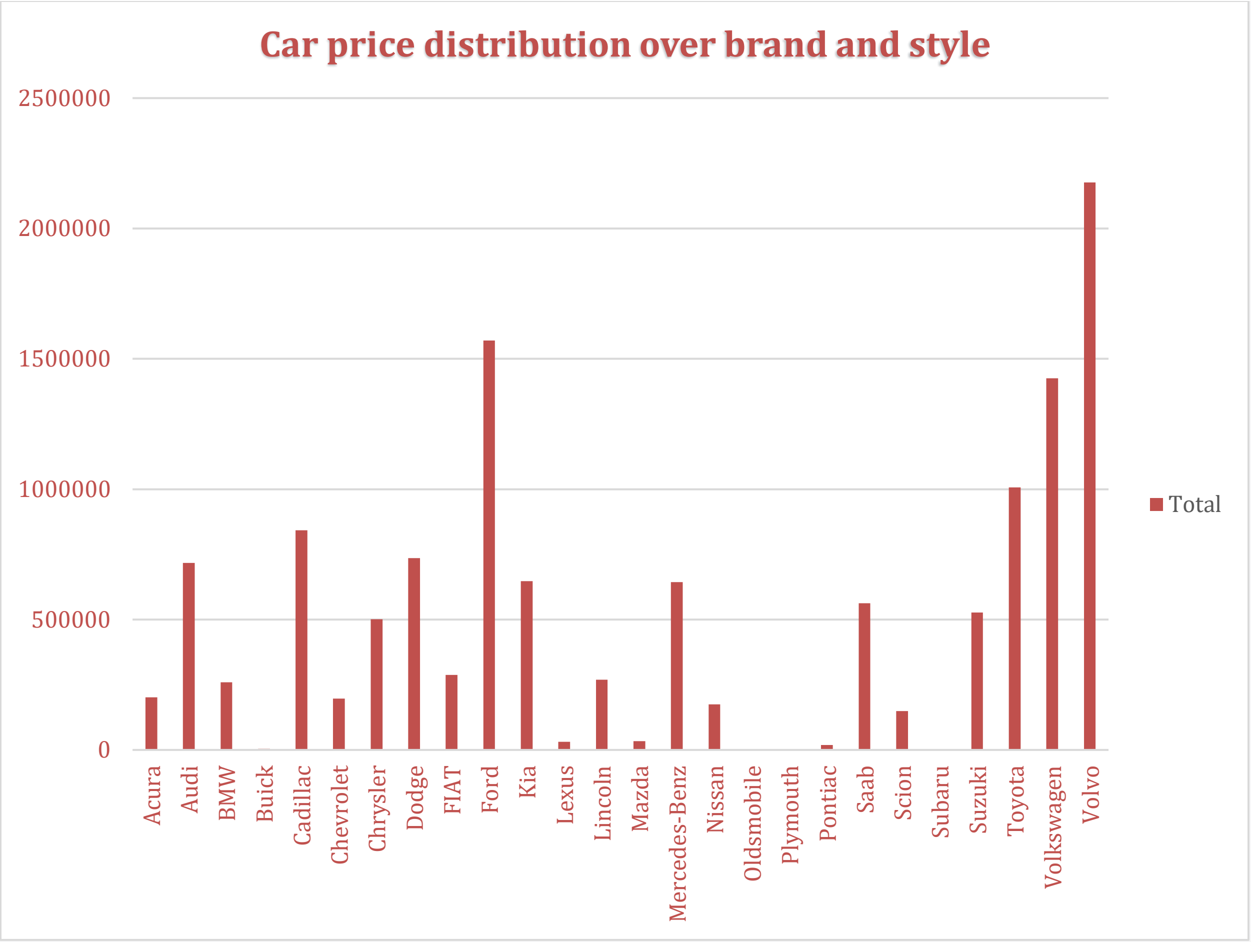
	Engine Cylinders	highway MPG
Engine Cylinders	1	
highway MPG	-0.621605733	1

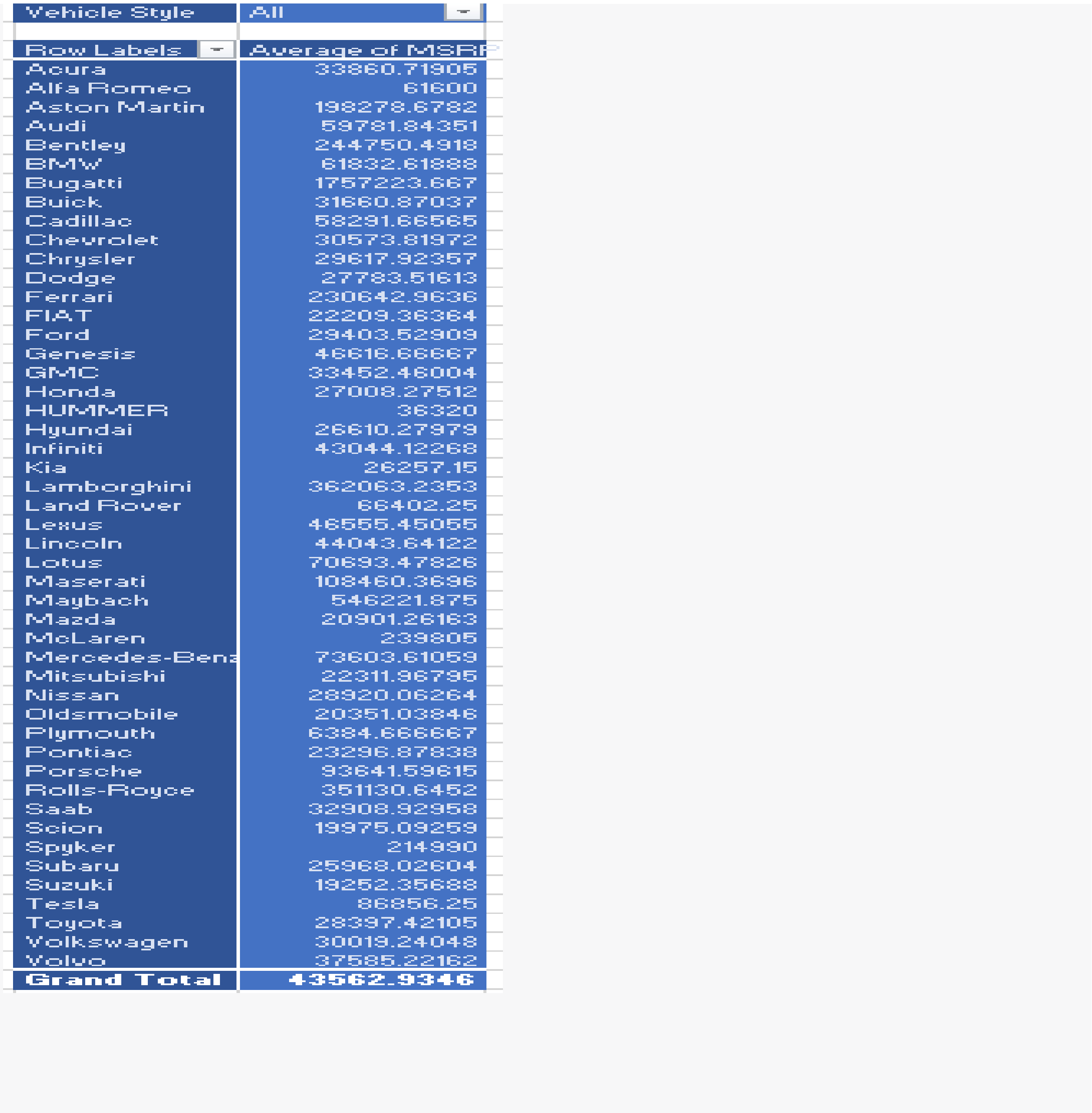


Insight Required: How does the average price of a car vary across different manufacturers?

- **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.


Car price distribution over brand and style
2500000
2000000
1500000
1000000
500000
0
Acura
Audi
BMW
Buick
Cadillac
Chevrolet
Chrysler
Dodge
FIAT
Ford
Kia
Lexus
Lincoln
Mazda
Mercedes-Benz
Nissan
Oldsmobile
Plymouth
Pontiac
Saab
Scion
Subaru
Suzuki
Toyota
Volkswagen
Volvo
Total



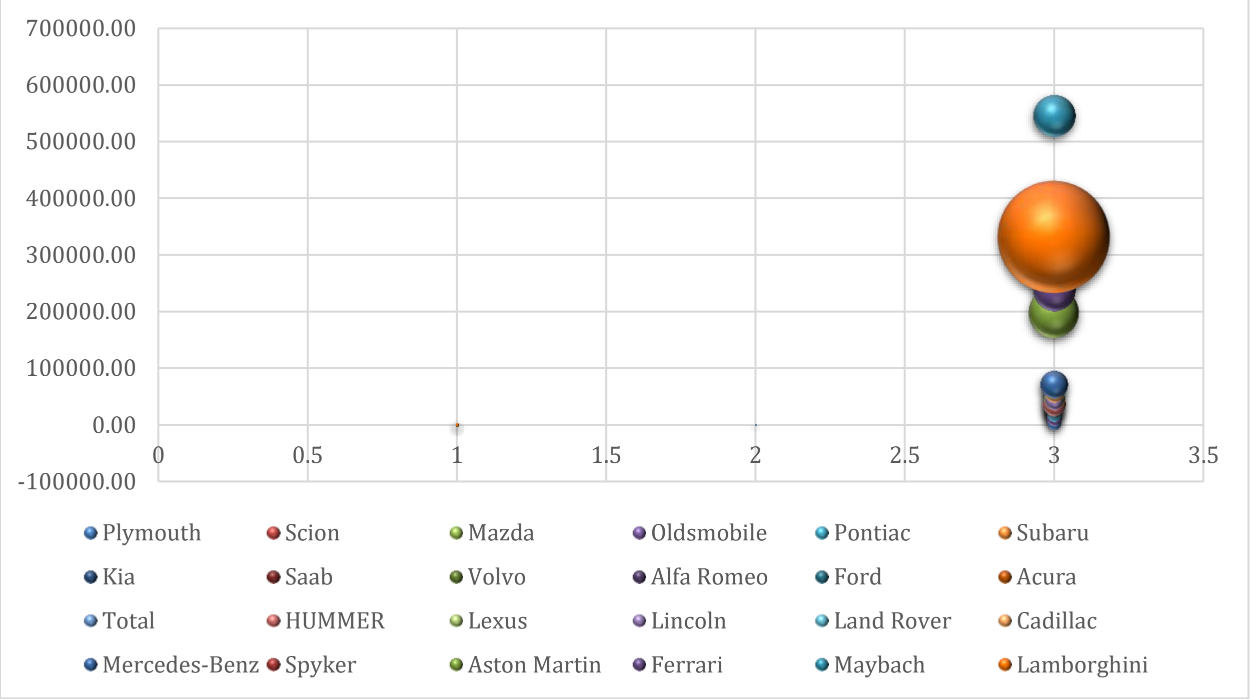
Vehicle Style	All

Row Labels	Average of MSRP
Acura	33860.71905
Alfa Romeo	61600
Aston Martin	198278.6782
Audi	59781.84351
Bentley	244750.4918
BMW	61832.61888
Bugatti	1757223.667
Buick	31660.87037
Cadillac	58291.66565
Chevrolet	30573.81972
Chrysler	29617.92357
Dodge	27783.51613
Ferrari	230642.9636
FIAT	22209.36364
Ford	29403.52909
Genesis	46616.66667
GMC	33452.46004
Honda	27008.27512
HUMMER	36320
Hyundai	26610.27979
Infiniti	43044.12268
Kia	26257.15
Lamborghini	362063.2353
Land Rover	66402.25
Lexus	46555.45055
Lincoln	44043.64122
Lotus	70693.47826
Maserati	108460.3696
Maybach	546221.875
Mazda	20901.26163
McLaren	239805
Mercedes-Benz	73603.61059
Mitsubishi	22311.96795
Nissan	28920.06264
Oldsmobile	20351.03846
Plymouth	6384.666667
Pontiac	23296.87838
Porsche	93641.59615
Rolls-Royce	351130.6452
Saab	32908.92958
Scion	19975.09259
Spyker	214990
Subaru	25968.02604
Suzuki	19252.35688
Tesla	86856.25
Toyota	28397.42105
Volkswagen	30019.24048
Volvo	37585.22162
Grand Total	**43562.9346**

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.

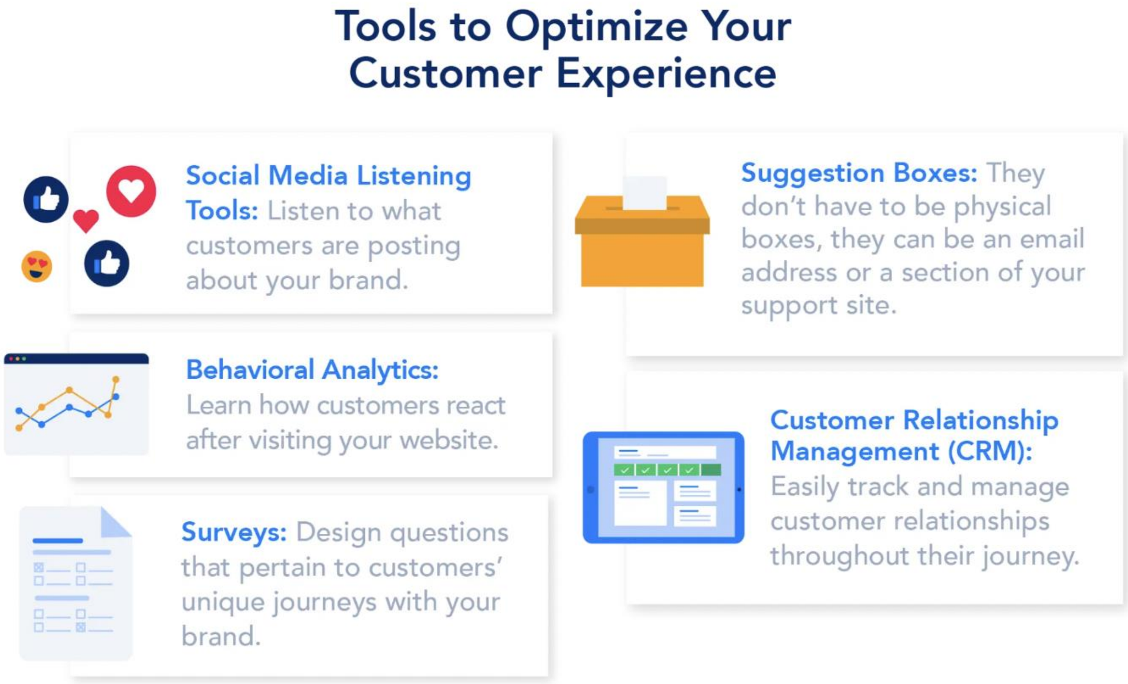
Task 5.B: Calculate the correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.



Conclusion

- **From this project I have learned how exploratory data analysis(EDA) is being performed on real time dataset.**
- **I have learned how to visualize the results and gain meaningful insights from datasets.**
- **Also I got a chance to work with real time dataset.**
- **This project will be helpful for me to find correlation, insights and conclusion from any datasets.**

9. ABC Call Volume Trend Analysis



Tools to Optimize Your Customer Experience

Social Media Listening Tools: Listen to what customers are posting about your brand.

Behavioral Analytics: Learn how customers react after visiting your website.

Surveys: Design questions that pertain to customers' unique journeys with your brand.

Suggestion Boxes: They don't have to be physical boxes, they can be an email address or a section of your support site.

Customer Relationship Management (CRM): Easily track and manage customer relationships throughout their journey.

Project Description:

- The project is about analyzing the call volume trend and call handling of a Customer Experience (CX) Inbound calling team of ABC Company for 23 days. The data includes various parameters such as Agent_Name, Agent_ID, Queue_Time, Time, Time_Bucket, Duration, Call_Seconds, and call status. The objective of this project is to calculate the average call time duration for all incoming calls received by agents, show the total volume/ number of calls coming in via charts/ graphs [Number of calls v/s Time], propose a manpower plan required during each time bucket [between 9 am to 9 pm] to reduce the abandon rate to 10%, and propose a manpower plan required during each time bucket in a day for answering the night calls. The project aims to provide insights into the call handling of the CX team and help in improving the overall customer experience.

Approach:

- Data Importing
- Data Cleaning
- Data Exploration
- Data Modeling
- Data Visualization

Tech-Stack Used:

Microsoft Excel has been used to perform the entire analysis.Microsoft powerpoint used for project presentation.

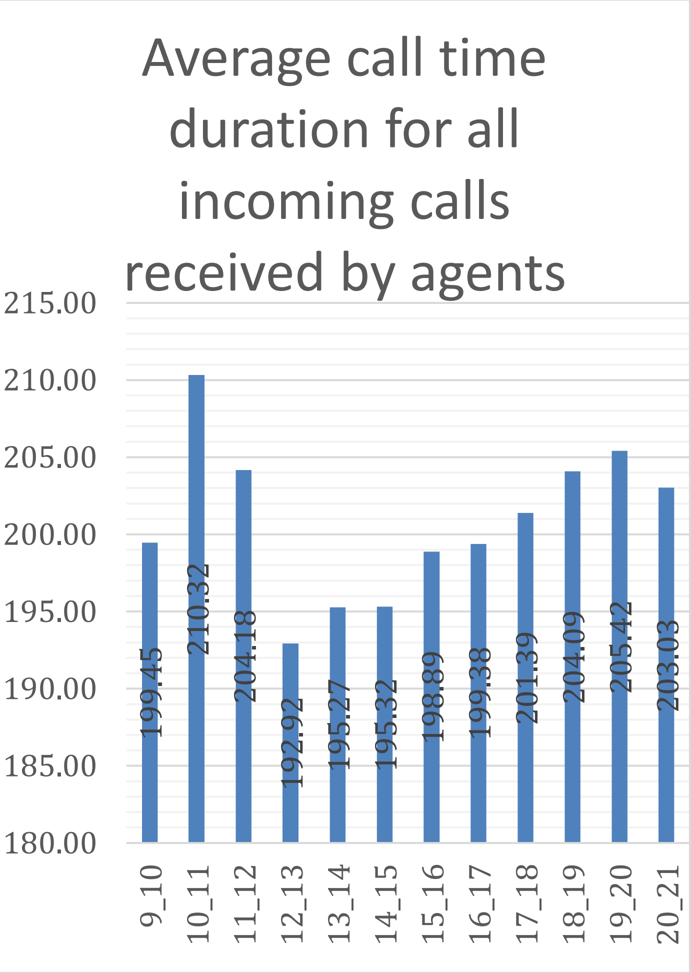
Insights:

The insights gained from the analysis are:

Q1. Calculate the average call time duration for all incoming calls received by agents (in each Time_Bucket).

Ans. The average call time duration for all incoming calls received by agents in each Time_Bucket ranges from 192.9 seconds to 203.4 seconds.

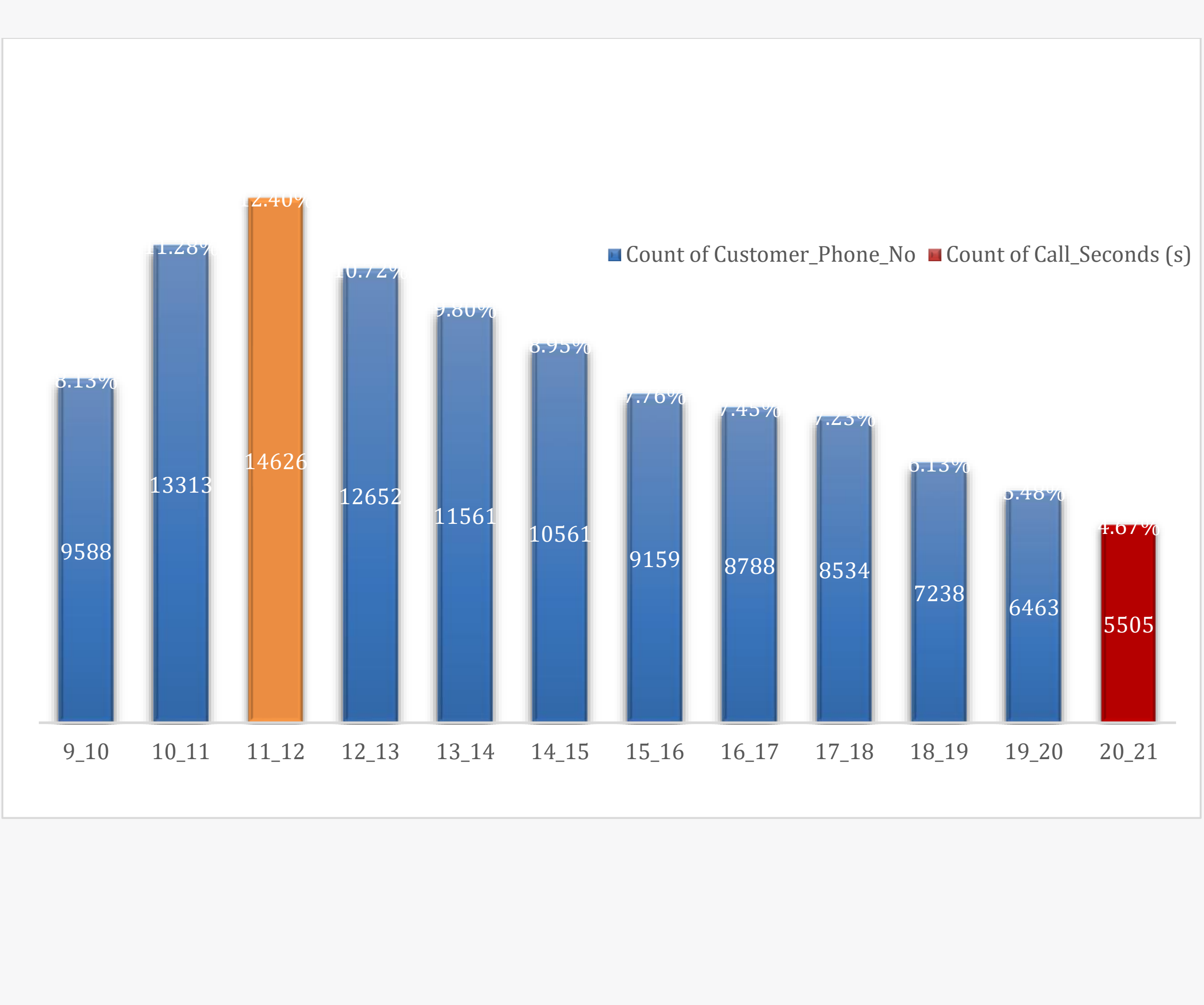
Call_Status	answered
Wrapped _By	Agent
Time_bucket	**Average of Call_Seconds (s)**
10_11	210.32
11_12	204.18
12_13	192.92
13_14	195.27
14_15	195.32
15_16	198.89
16_17	199.38
17_18	201.39
18_19	204.09
19_20	205.42
20_21	203.03
9_10	199.45
Grand Total	**199.81**



Q2. Show the total volume/ number of calls coming in via charts/ graphs [Number of calls v/s Time]. You can select time in a bucket form (i.e. 1-2, 2-3, …..)

Ans. The total volume/number of calls coming in is highest during the time bucket 11-12 and lowest during the time bucket 20-21.

Time_bucket	Count of Customer_Phone_No	Count % of Call_Seconds (s)
9_10	9588	8.13%
10_11	13313	11.28%
11_12	14626	12.40%
12_13	12652	10.72%
13_14	11561	9.80%
14_15	10561	8.95%
15_16	9159	7.76%
16_17	8788	7.45%
17_18	8534	7.23%
18_19	7238	6.13%
19_20	6463	5.48%
20_21	5505	4.67%
Grand Total	**117988**	**100.00%**

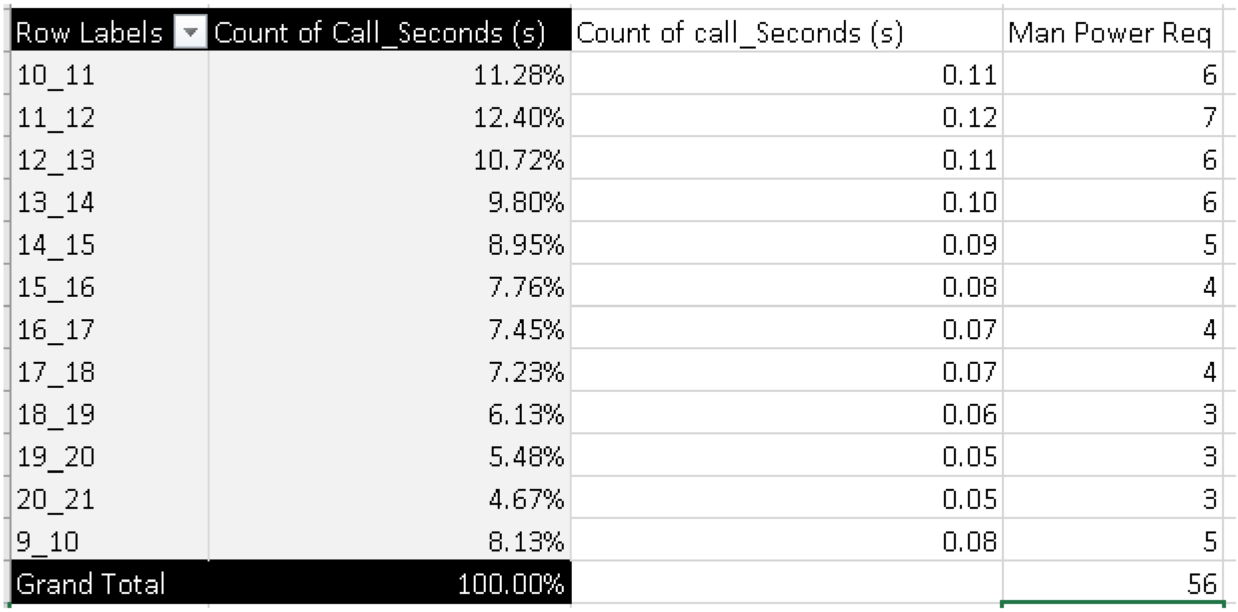


The trend is increasing till noon then continuous decrease as the day progresses.

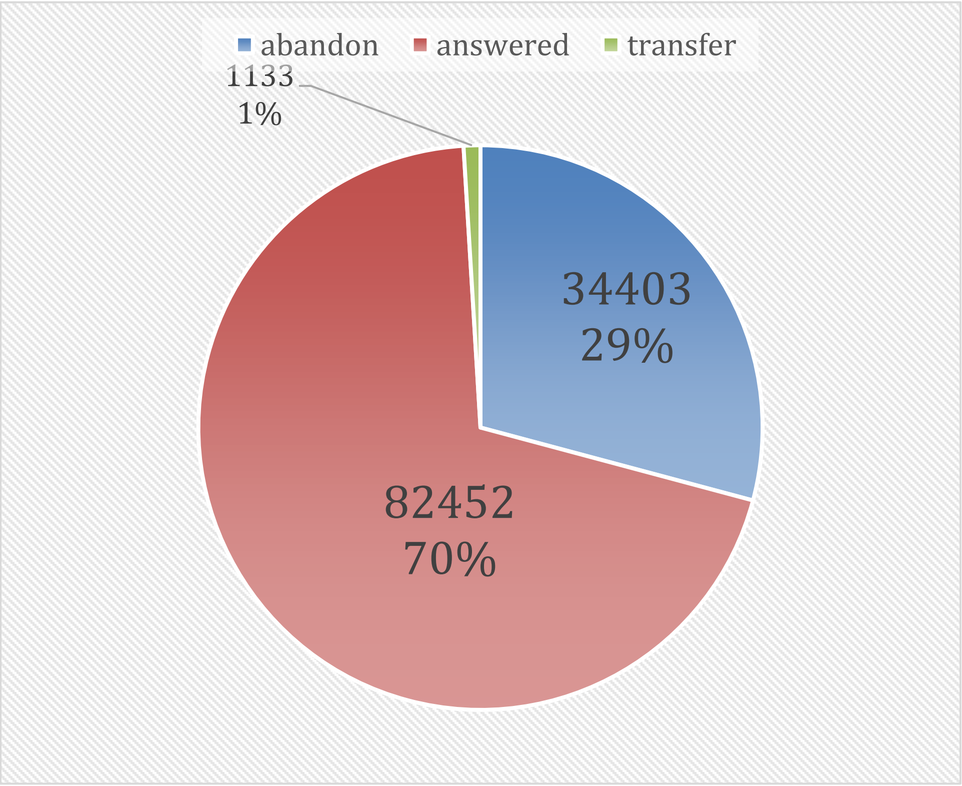
Q3. As you can see, the current abandon rate is approximately 30%. Propose a manpower plan required during each time bucket [between 9am to 9pm] to reduce the abandon rate to 10%. (i.e. You have to calculate the minimum number of agents required in each time bucket so that at least 90 calls should be answered out of 100.)

Ans. To reduce the abandon rate to 10%, the manpower plan required during each time bucket has been proposed in the report. The number of agents required varies from 4 to 7 during different time buckets.

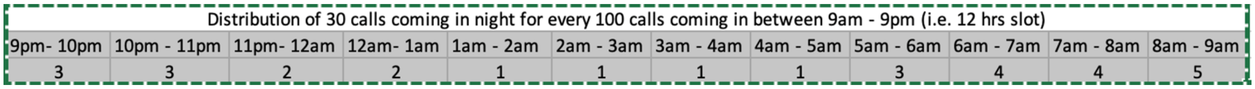
We need 56 agents to reduce it at 10%.



Row Labels	Count of Call_Seconds (s)	Count of call_Seconds (s)	Man Power Req
10_11	11.28%	0.11	6
11_12	12.40%	0.12	7
12_13	10.72%	0.11	6
13_14	9.80%	0.10	6
14_15	8.95%	0.09	5
15_16	7.76%	0.08	4
16_17	7.45%	0.07	4
17_18	7.23%	0.07	4
18_19	6.13%	0.06	3
19_20	5.48%	0.05	3
20_21	4.67%	0.05	3
9_10	8.13%	0.08	5
Grand Total	100.00%		56



Q4. Let’s say customers also call this ABC insurance company at night but don't get an 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:

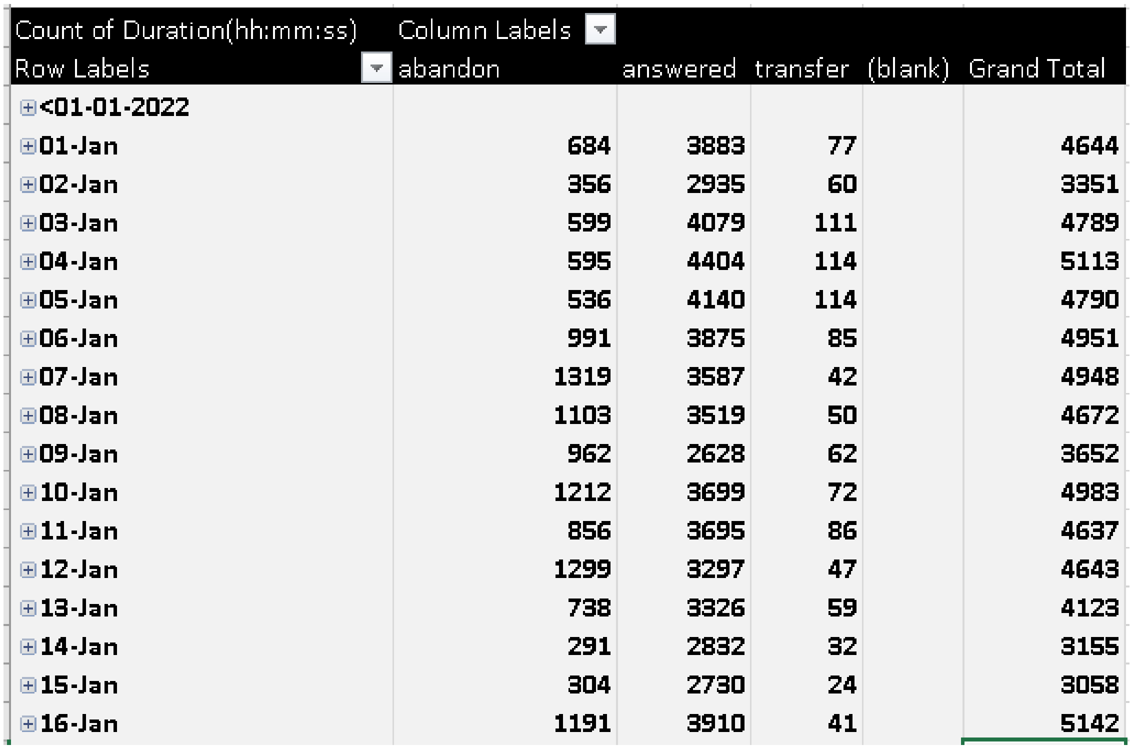


Distribution of 30 calls coming in night for every 100 calls coming in between 9am - 9pm (i.e. 12 hrs slot)

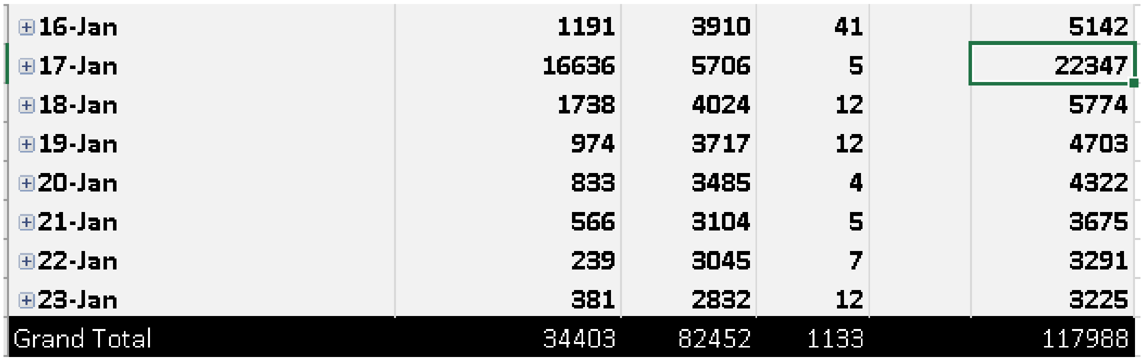
9pm- 10pm	10pm - 11pm	11pm- 12am	12am- 1am	1am - 2am	2am - 3am	3am - 4am	4am - 5am	5am - 6am	6am - 7am	7am - 8am	8am - 9am
3	3	2	2	1	1	1	1	3	4	4	5

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

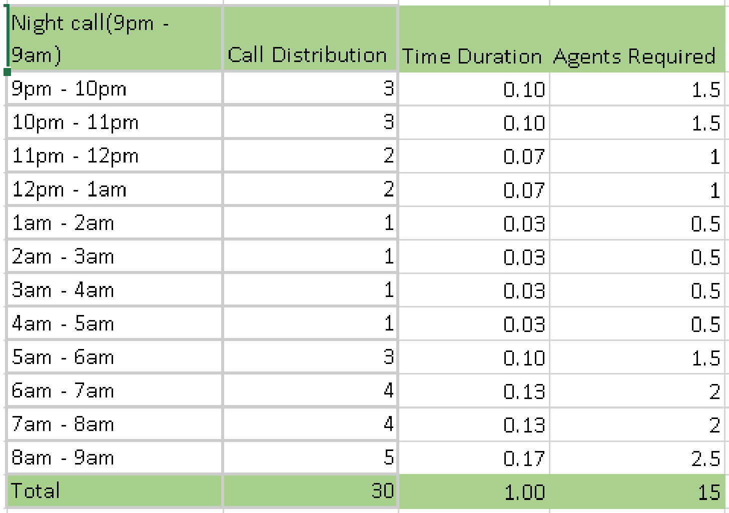
Ans. To answer the night calls, the manpower plan required during each time bucket in a day has been proposed in the report. The number of agents required varies from 1 to 6 during different time buckets.



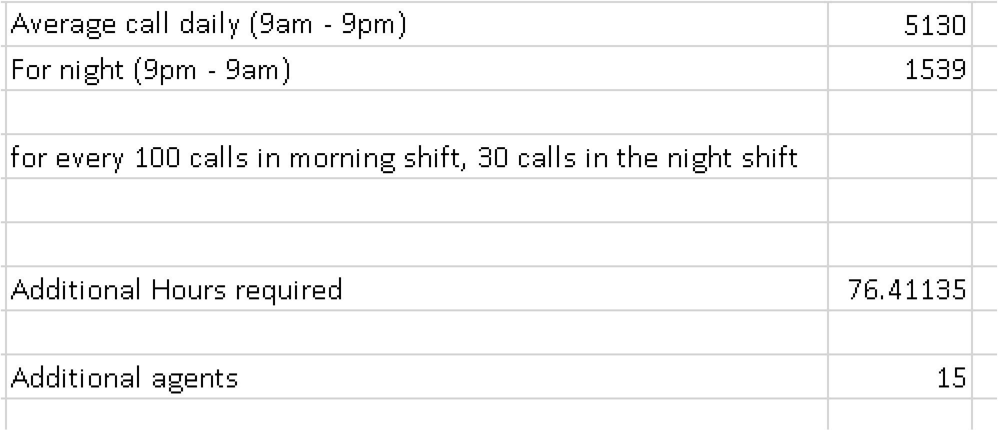
Count of Duration(hh:mm:ss)	Column Labels				
Row Labels	abandon	answered	transfer	(blank)	Grand Total
<01-01-2022					
01-Jan	684	3883	77		4644
02-Jan	356	2935	60		3351
03-Jan	599	4079	111		4789
04-Jan	595	4404	114		5113
05-Jan	536	4140	114		4790
06-Jan	991	3875	85		4951
07-Jan	1319	3587	42		4948
08-Jan	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



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
Grand Total	34403	82452	1133		117988



Night call(9pm - 9am)	Call Distribution	Time Duration	Agents Required
9pm - 10pm	3	0.10	1.5
10pm - 11pm	3	0.10	1.5
11pm - 12pm	2	0.07	1
12pm - 1am	2	0.07	1
1am - 2am	1	0.03	0.5
2am - 3am	1	0.03	0.5
3am - 4am	1	0.03	0.5
4am - 5am	1	0.03	0.5
5am - 6am	3	0.10	1.5
6am - 7am	4	0.13	2
7am - 8am	4	0.13	2
8am - 9am	5	0.17	2.5
Total	30	1.00	15



Average call daily (9am - 9pm)	5130
For night (9pm - 9am)	1539
for every 100 calls in morning shift, 30 calls in the night shift	
Additional Hours required	76.41135
Additional agents	15

While doing this ABC Call Volume Trend Analysis, I have first of all learned about how data is stored and utilized in any Call center for observing the customer experience or company's growth. I have observed that the company utilizes this data to check its requirements for more workers, to check their customers get satisfied. To draw out any conclusions from the dataset, first we have to think how it is possible to do this and how this can be achieved. I have used various mathematical calculations while doing the project.

Results:

During the project, I have learned about various techniques and possible ways to solve or deal with any problem. Also, how minute error leads to wrong calculations and thus wrong result, therefore, one should be very vigilant while doing any type of analysis.

Throughout this project, I have gained valuable insights into the impact of an analyst in the customer service department.

Additionally, I have delved into the realm of behavioral analytics, which involves studying customer behavior patterns to identify trends, preferences, and opportunities for enhancing the overall customer experience.

Conclusion:-

As a data analyst, I have completed several projects that involved analyzing data to gain insights and make informed decisions. These projects covered a range of topics, including social media user behavior, operational performance, hiring processes, movie success factors, loan default rates, advertising campaign effectiveness, and call center trends. As a data analyst, your role is to extract insights from data and use those insights to make informed recommendations. To be successful, you need to be able to work with a variety of data types, apply different analytical techniques, and communicate your findings effectively to stakeholders.

Appendix:-

All The Projects:- Data Analytics Projects

1. Data analytics process link

2. Instagram User Analytics link

3. Operation & Metric Analytics link

4. Hiring Process Analytics link

5. IMDB Movie Analysis link

6. Bank Loan Case Study link

7. Impact of car features link

8. ABC Call Volume Trend Analysis link

Thank you..!