Data Analysis portfolio

- **By Dharmateja**

Professional Background

I am P.Dharmateja an undergraduate from National Institute of Technology, Calicut with specialization in Production Engineering with an overall CGPA of 7.87.My passion on data and analytics drive me towards data Analytic field.I am skilled at Python, SQL, Power BI and Excel. I have been working as a Data Analyst Trainee In Trainity.

During the program I have worked on several projects using excel and SQL. Being a fresher I am very flexible and adaptive to learn new things. I have good theoretical knowledge on data analytics and looking forward to put my knowledge in a practical way.

Table of Content

Project-1

Data Analytics Process

Cultivate paddy in the farming land

This project is a real world scenario that depicts the processes involved in data analytics.The real world scenario here is cultivation of paddy in the farming land.The steps involved in cultivating paddy that are same in data analytics are:

1) Plan

2) Prepare

3) Process

4) Analyze

5) Share

6) Act

PLAN

Decided to cultivate paddy crop in the farmland. In order to cultivate the paddy a detailed plan has been made which includes

- Choosing appropriate variety of paddy
- The month in which to cultivate it
- Estimating the required amount to buy seeds, Fertilizers, Manual work and machine work. Decided to cultivate paddy crop in the farmland. In order to cultivate the paddy a detailed plan has been made which includes

PREPARE

- Preparing the farming land by removing weeds, tilling of soil and adding organic mattersor fertilizers to enrich the nutrients in the soil.
- Preparing the seeds to sow in the soil
- Preparing proper irrigation channel to farming land.

PROCESS

- Cultivating the paddy at the right time (July). To enhance its growth spraying fertilizers and monitoring the crop carefully. Watering the crop on regular basis and removing weeds after few months of cultivating. Apply pesticides to protect the crop from pests.

ANALYZE

- Monitoring the crop and analysing the growth of paddy. Monitoring it enable to understand how is the growth of cropor what resisting the crop from proper growth or what kind ofdisease it has been affected if any and helps to take informed decision in this regard.

SHARE

- After yielding of the crop, harvesting it and making a pile.
- Sharing the information about the type of paddy, growth of crop and how much has been harvested with the local markets and vendors to know the demand, pricing and potential buyers of the crop.

ACT

- Protecting the paddy after harvesting and Packaging it so that it should be in an optimal condition while reaching to market.
- Based on the demand and pricing, acting accordingly. If the demand and pricing isless then storing it in a godown and waiting till increase in the price and selling itat right time for right price or else selling it at the moment of crop harvest.

Project 2
Instagram user Analytics

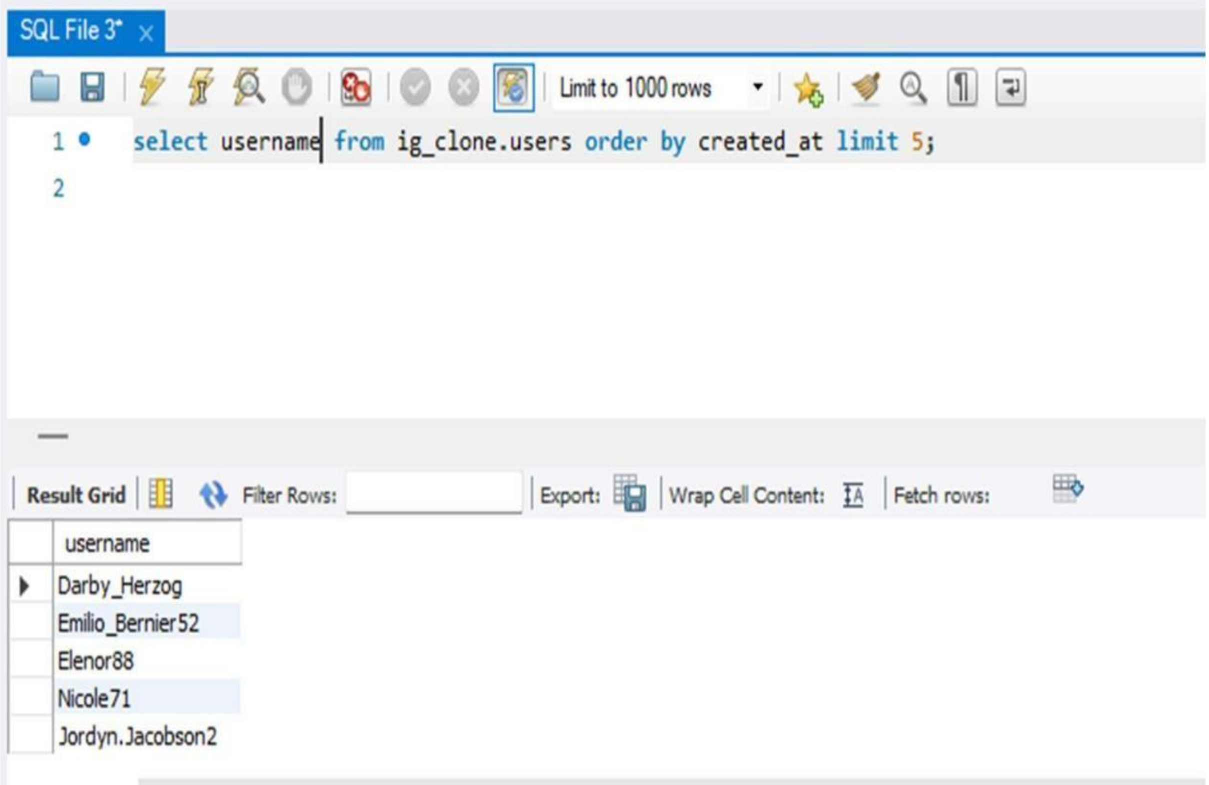
Problem statement:

Imagine you're a data analyst working with the product team at Instagram. Your role involves analyzing user interactions and engagement with the Instagram app to provide valuable insights that can help the business grow.

User analysis involves tracking how users engage with a digital product, such as a software application or a mobile app. The insights derived from this analysis can be used by various teams within the business.

A) Marketing Analysis

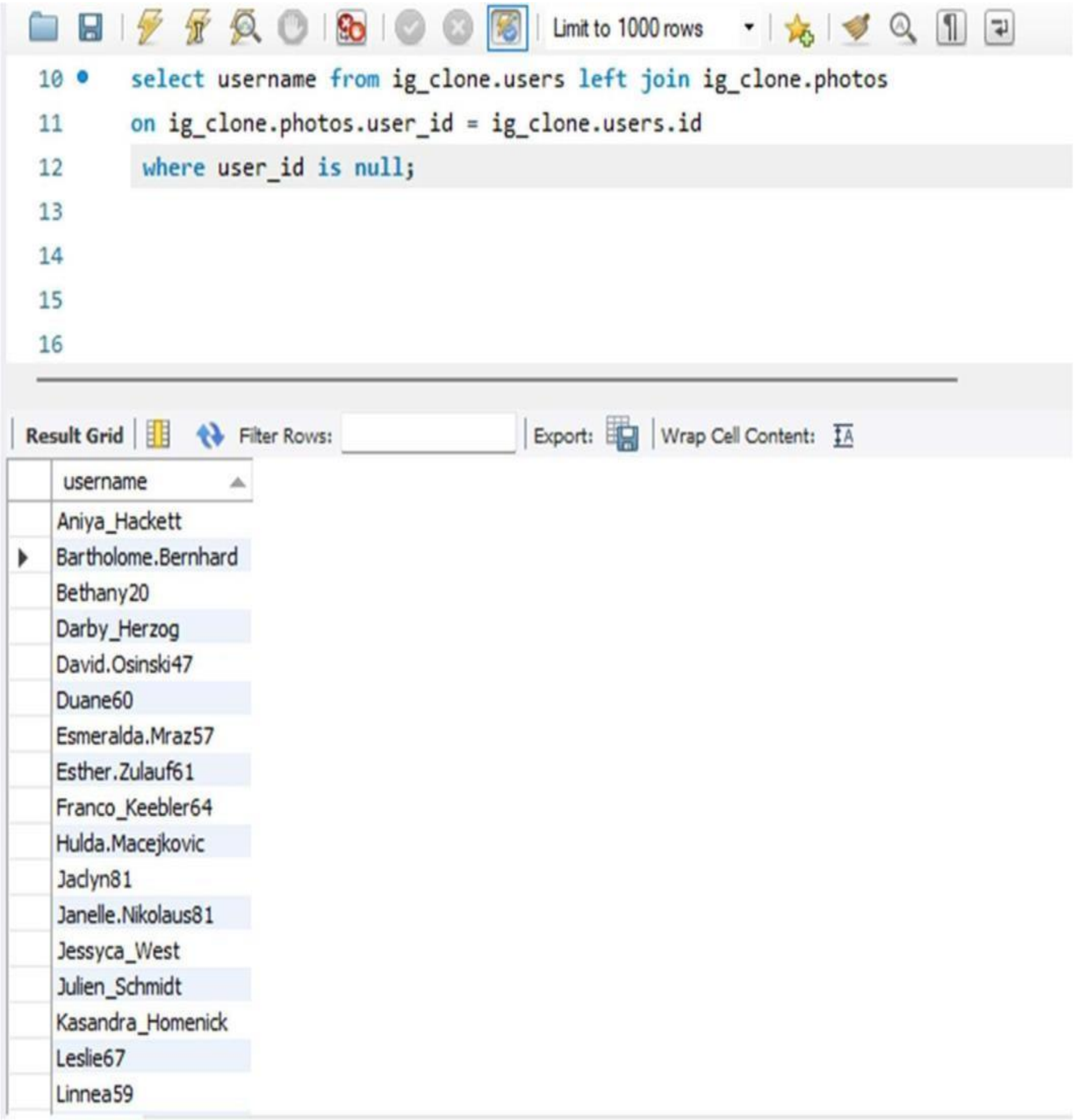
1) Task: Identify the five oldest users on Instagram from the provided database



```
select username from ig_clone.users order by created_at limit 5;
```

username
Darby_Herzog
Emilio_Bernier52
Elenor88
Nicole71
Jordyn.Jacobson2

2) Task: Identify users who have never posted a single photo on Instagram.



```
select username from ig_clone.users left join ig_clone.photos
on ig_clone.photos.user_id = ig_clone.users.id
 where user_id is null;
```

username
Aniya_Hackett
Bartholome.Bernhard
Bethany20
Darby_Herzog
David.Osinski47
Duane60
Esmeralda.Mraz57
Esther.Zulauf61
Franco_Keebler64
Hulda.Macejkovic
Jaclyn81
Janelle.Nikolaus81
Jessyca_West
Julien_Schmidt
Kasandra_Homenick
Leslie67
Linnea59

List of users who never posted a single photo in instagram

Aniya_Hackett

Kasandra_Homenick

Jaclyn81

Rocio33

Maxwell.Halvorson

Tierra.Trantow

Pearl7

Ollie_Ledner37

Mckenna17

David.Osinski47

Morgan.Kassulke

Linnea59

Duane60

Julien_Schmidt

Mike.Auer39

Franco_Keebler64

Nia_Haag

Hulda.Macejkovic

Leslie67

Janelle.Nikolaus81

Darby_Herzog

Esther.Zulauf61

Bartholome.Bernhard

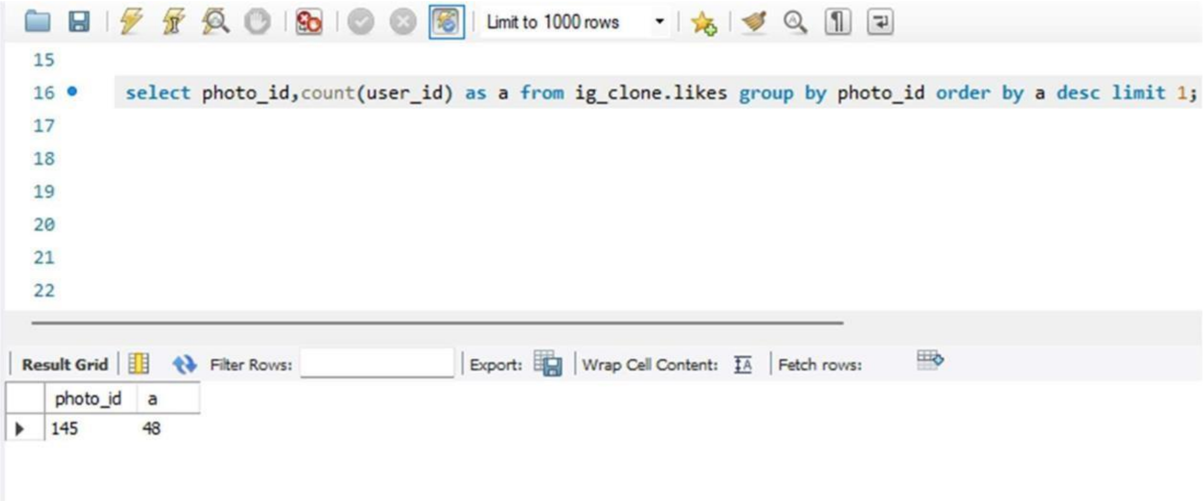
Jessyca_West

Esmeralda.Mraz57

Bethany20

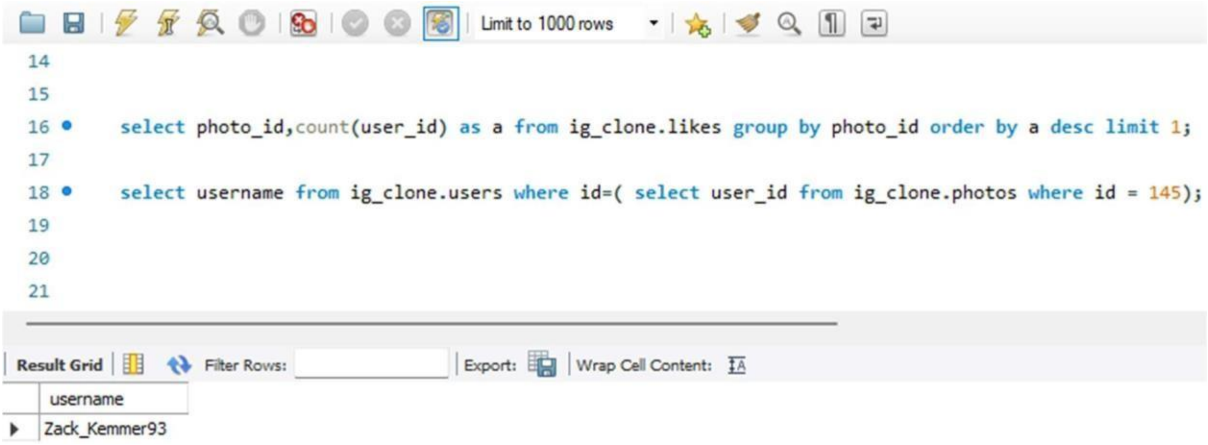
3) Task: Determine the winner of the contest and provide their details to the team.

The winner of the contest is the one who's photo has highest number of likes



```sql
select photo_id,count(user_id) as a from ig_clone.likes group by photo_id order by a desc limit 1;
```

photo_id	a
145	48



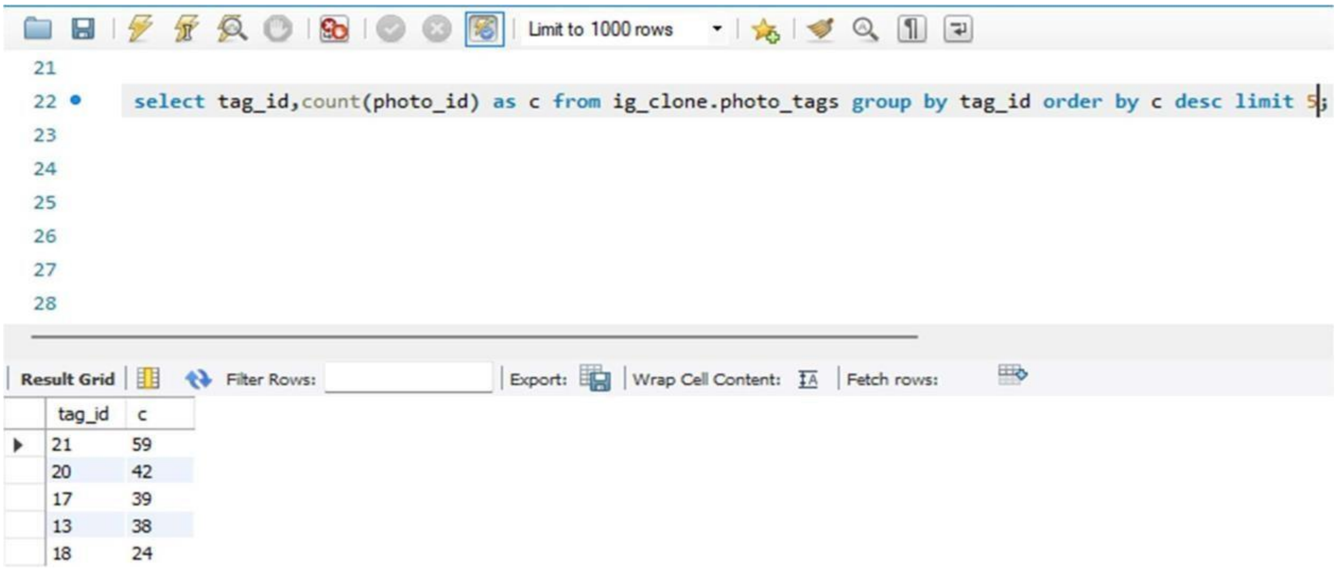
```sql
select photo_id,count(user_id) as a from ig_clone.likes group by photo_id order by a desc limit 1;

select username from ig_clone.users where id=( select user_id from ig_clone.photos where id = 145);
```

username
Zack_Kemmer93

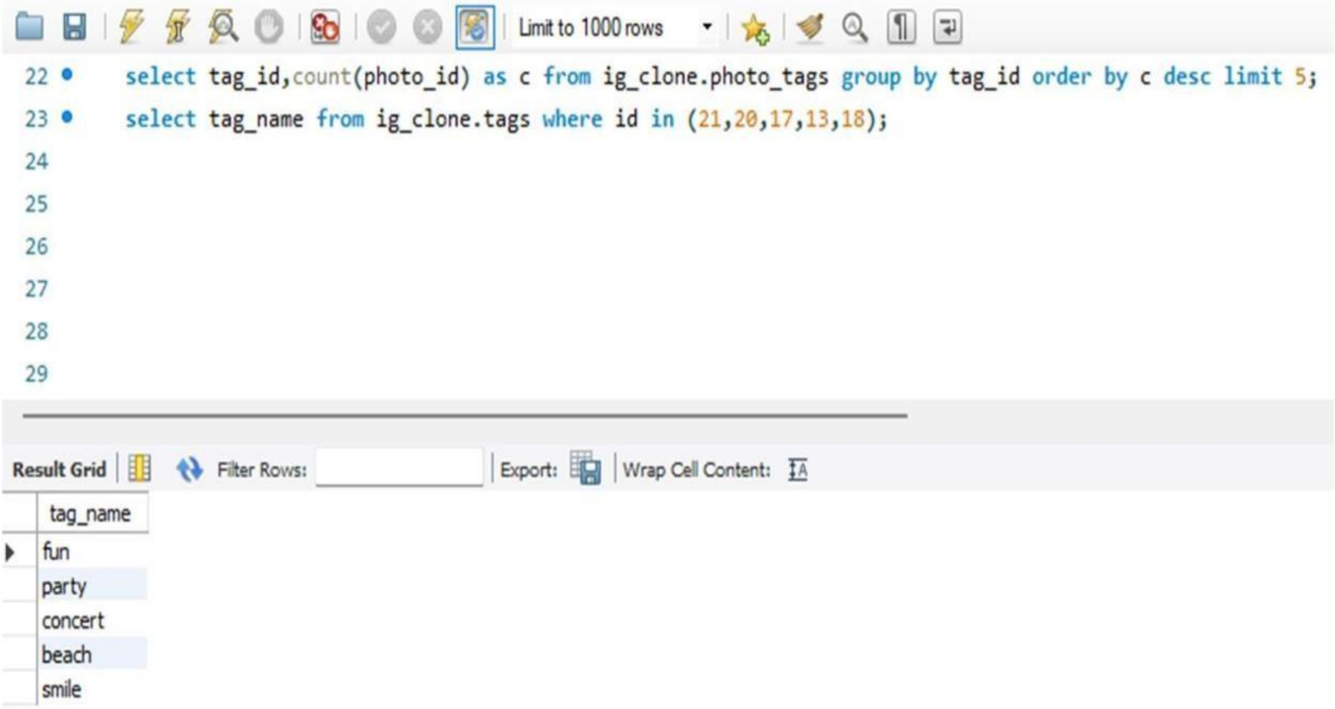
From the above queries we can see that the winner of the contest is with the username Zack_Kemmer93

4) Task: Identify and suggest the top five most commonly used hashtags on the platform.



```sql
select tag_id,count(photo_id) as c from ig_clone.photo_tags group by tag_id order by c desc limit 5;
```

tag_id	c
21	59
20	42
17	39
13	38
18	24

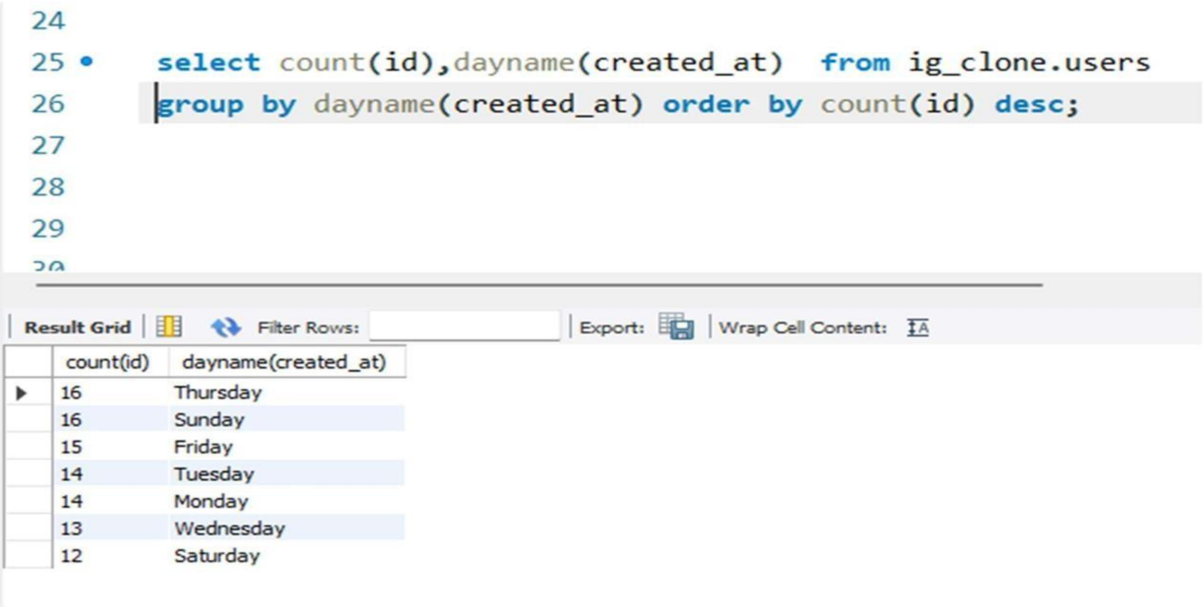


```sql
select tag_id,count(photo_id) as c from ig_clone.photo_tags group by tag_id order by c desc limit 5;
select tag_name from ig_clone.tags where id in (21,20,17,13,18);
```

tag_name
fun
party
concert
beach
smile

The commonly used tags are fun , party, concert, beach and smile

5) Task: Determine the day of the week when most users register on instagram. Provide insights on when to schedule an ad campaign.



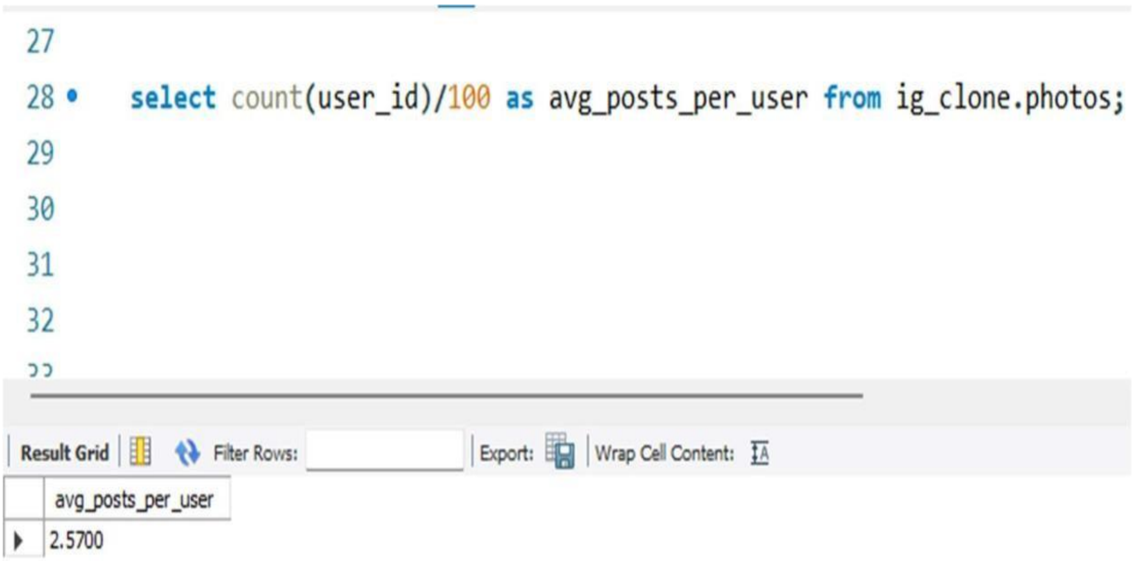
```
select count(id),dayname(created_at)  from ig_clone.users
group by dayname(created_at) order by count(id) desc;
```

count(id)	dayname(created_at)
16	Thursday
16	Sunday
15	Friday
14	Tuesday
14	Monday
13	Wednesday
12	Saturday

The best days in a week to post the ads are Thursday and Sunday because these are the days in which most of the users get registered.

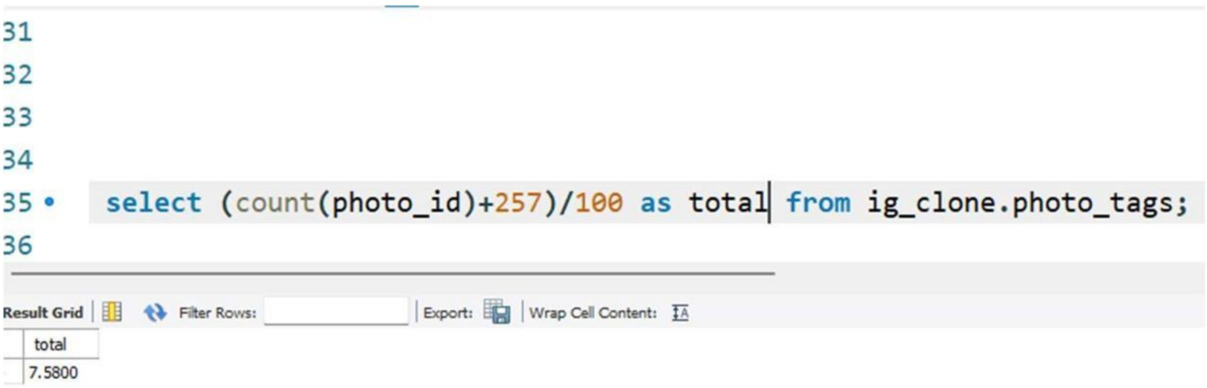
Investor Metrics:

1) Task: Calculate the average number of posts per user on Instagram. Also, provide the total number of photos on Instagram divided by the total number of users.



```
select count(user_id)/100 as avg_posts_per_user from ig_clone.photos;
```

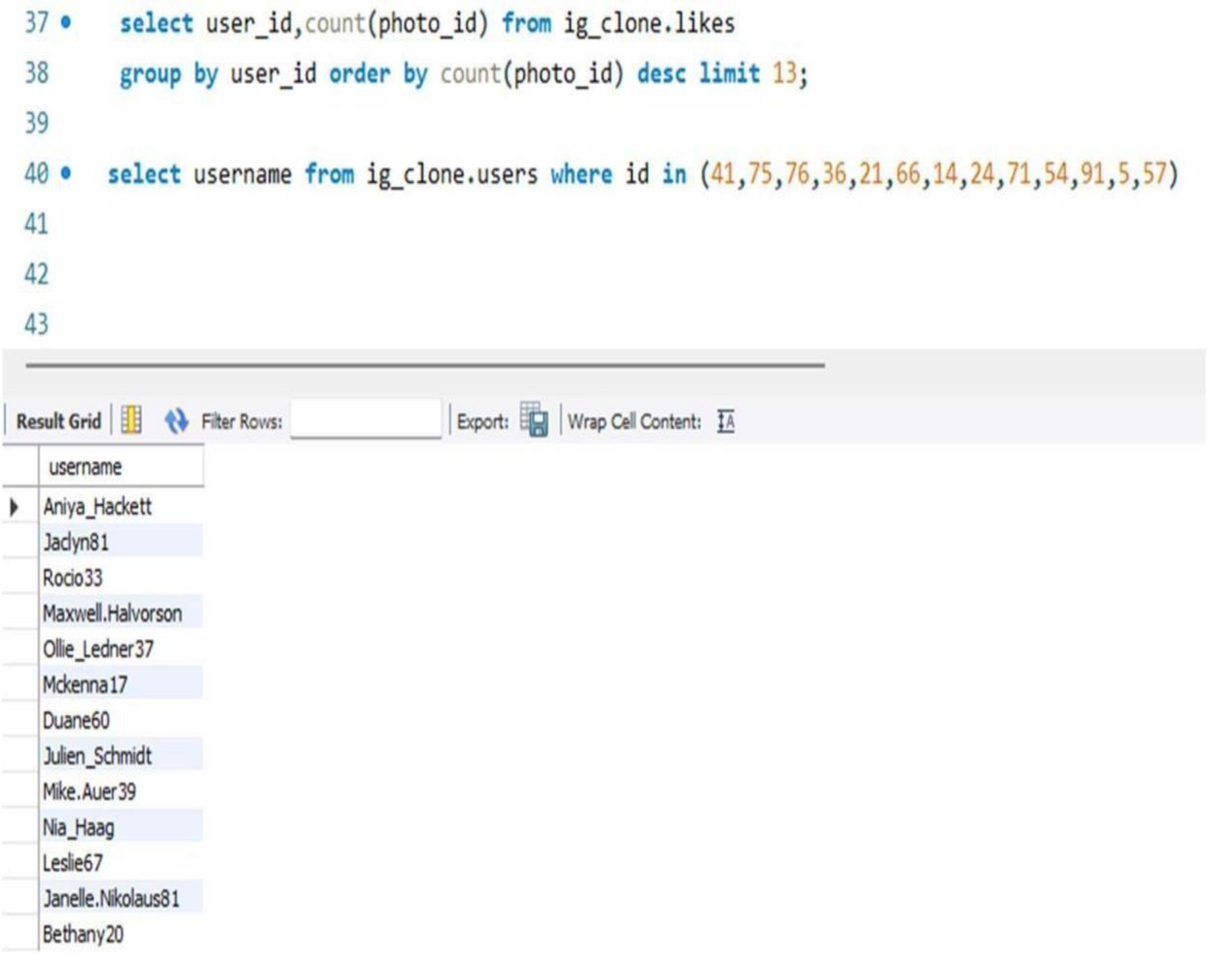
avg_posts_per_user
2.5700



```sql
select (count(photo_id)+257)/100 as total from ig_clone.photo_tags;
```

total
7.5800

2) Task: Identify users (potential bots) who have liked every single photo on the site, as this is not typically possible for a normal user



```sql
select user_id,count(photo_id) from ig_clone.likes
group by user_id order by count(photo_id) desc limit 13;

select username from ig_clone.users where id in (41,75,76,36,21,66,14,24,71,54,91,5,57)
```

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

These are the users who liked every single photo in the Instagram.

This project has been done to analyse the user interaction with the Instagram app. This helps in knowing the mindset of users so that new features, requirements with in the app can be made accessible to users.

Approach:

First of all I analysed the tables provided in the database in order to know the information present in each table.

The software I used to make this project is Mysql work bench

I successfully completed the project and got many useful insights which helps in making the Instagram app more user friendly. Idetified the more active users and inactive users, identified the days in which people used to get registered in the app, Identified most commonly used hastags andprovided awards to users who’s posts are mostly liked.

Project-3

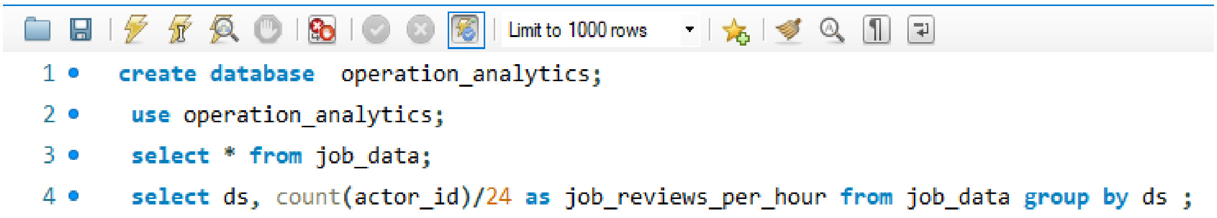
Operation Analytics and Investigating Metric Spike

Problem Statement:

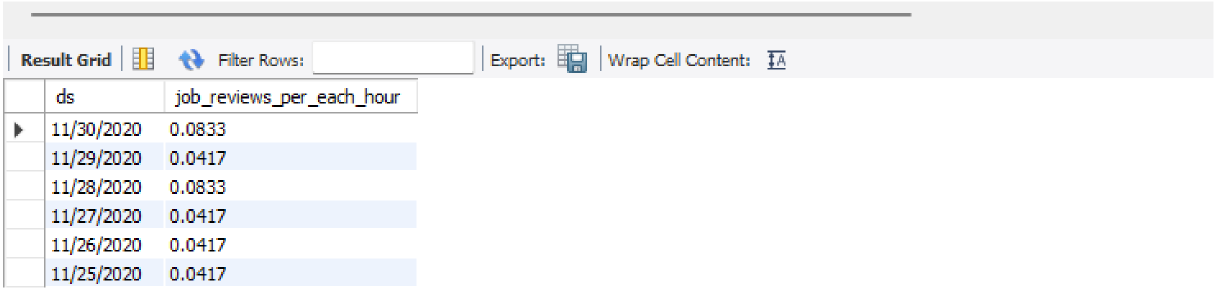
In this project, you'll take on the role of a Lead Data Analyst at a company like Microsoft. You'll be provided with various datasets and tables, and your task will be to derive insights from this data to answer questions posed by different departments within the company. Your goal is to use your advanced SQL skills to analyze the data and provide valuable insights that can help improve the company's operations and understand sudden changes in key metrics.

Case Study 1: Job Data Analysis

A) Task: Write an SQL query to calculate the number of jobs reviewed per hour for each day in November 2020.

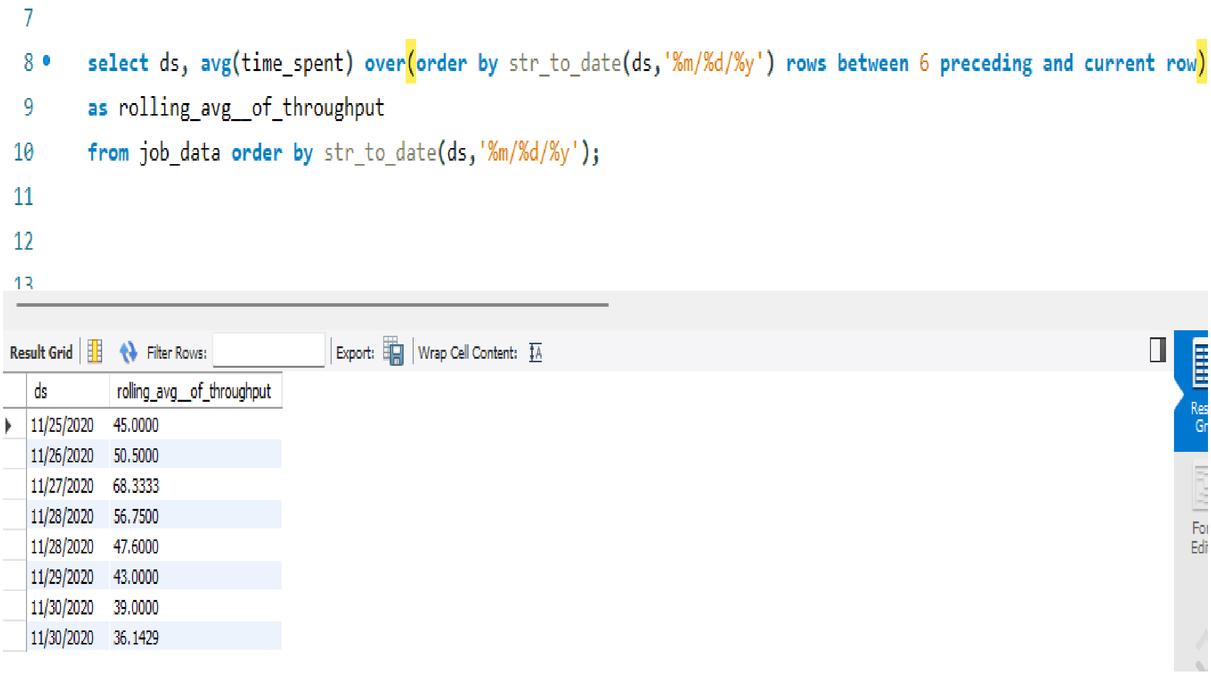


```sql
create database  operation_analytics;
 use operation_analytics;
 select * from job_data;
 select ds, count(actor_id)/24 as job_reviews_per_hour from job_data group by ds ;
```



ds	job_reviews_per_each_hour
11/30/2020	0.0833
11/29/2020	0.0417
11/28/2020	0.0833
11/27/2020	0.0417
11/26/2020	0.0417
11/25/2020	0.0417

B) Task: Write an SQL query to calculate the 7-day rolling average of throughput. Additionally, explain whether you prefer using the daily metric or the 7-day rolling average for throughput, and why.



```
select ds, avg(time_spent) over(order by str_to_date(ds,'%m/%d/%y') rows between 6 preceding and current row)
as rolling_avg__of_throughput
from job_data order by str_to_date(ds,'%m/%d/%y');
```

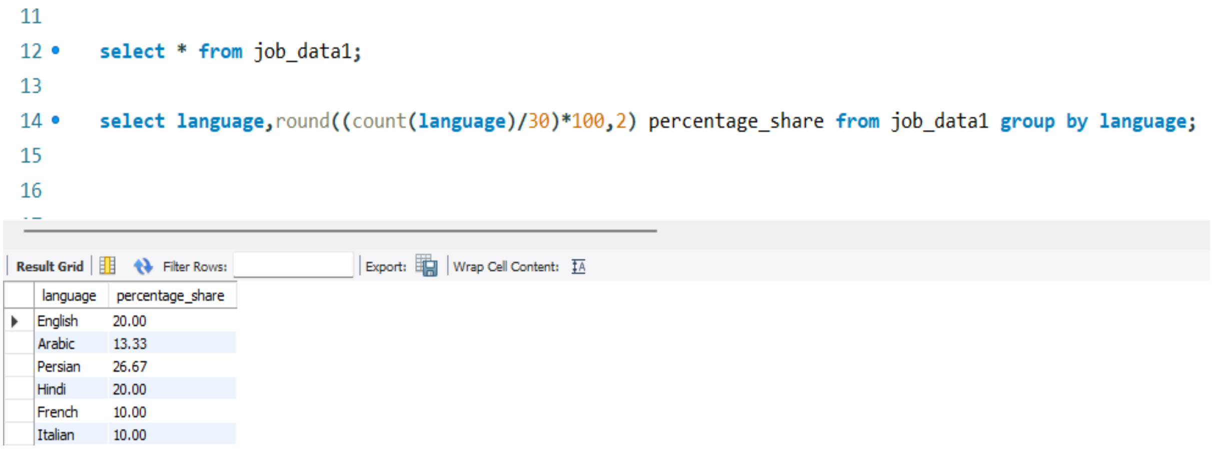
ds	rolling_avg__of_throughput
11/25/2020	45.0000
11/26/2020	50.5000
11/27/2020	68.3333
11/28/2020	56.7500
11/28/2020	47.6000
11/29/2020	43.0000
11/30/2020	39.0000
11/30/2020	36.1429

By using daily metrics we can able find the actual throughput for each day that is useful for identifying day to day variations. whereas 7 day rolling average nullifies the short term variation and helps to identify long term patterns.

Both metrics are useful in some or the other way. In order to find user daily engagementdaily metrics is quite useful.

C) Write an SQL query to calculate the percentage share of each language over the last 30 days.

Here I considered the table name as job_data1. This table is same as job_data table with additional 22 more rows as in the question it is mentioned to find for last 30 days.



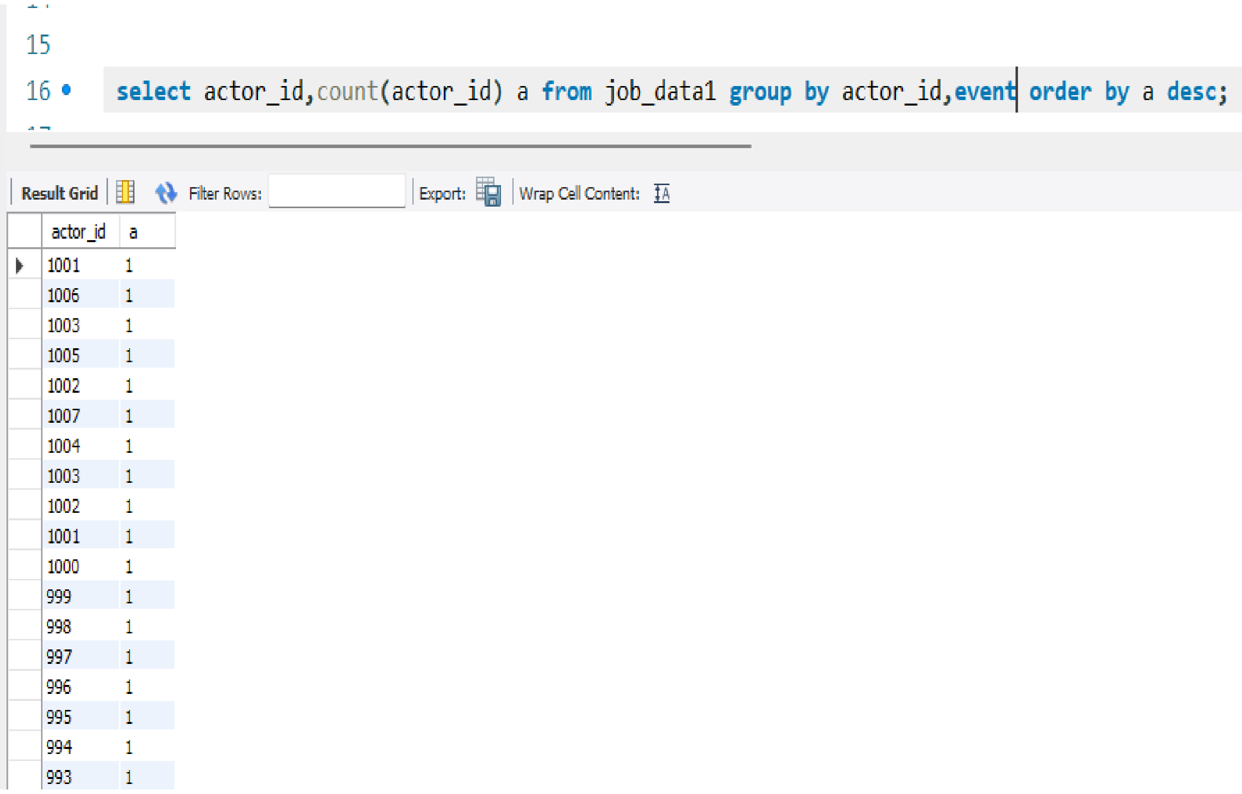
```
select * from job_data1;

select language,round((count(language)/30)*100,2) percentage_share from job_data1 group by language;
```

language	percentage_share
English	20.00
Arabic	13.33
Persian	26.67
Hindi	20.00
French	10.00
Italian	10.00

From the above results it can be seen that persian language share is more with an percentage share of 26.67 followed by English with 20%.

D) Write an SQL query to display duplicate rows from the job_data table.



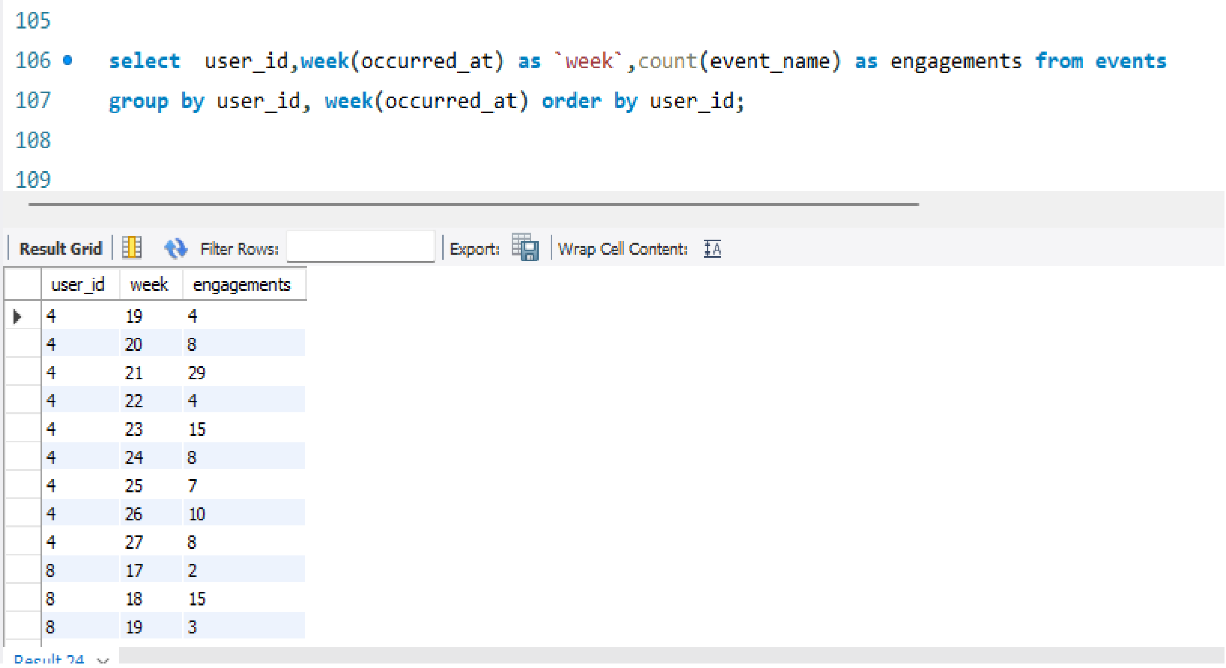
```
select actor_id,count(actor_id) a from job_data1 group by actor_id,event order by a desc;
```

actor_id	a
1001	1
1006	1
1003	1
1005	1
1002	1
1007	1
1004	1
1003	1
1002	1
1001	1
1000	1
999	1
998	1
997	1
996	1
995	1
994	1
993	1

From the above results we can see that there is no duplicate rows present in the table.

Case Study 2: Investigating Metric Spike

A) Write an SQL query to calculate the weekly user engagement.

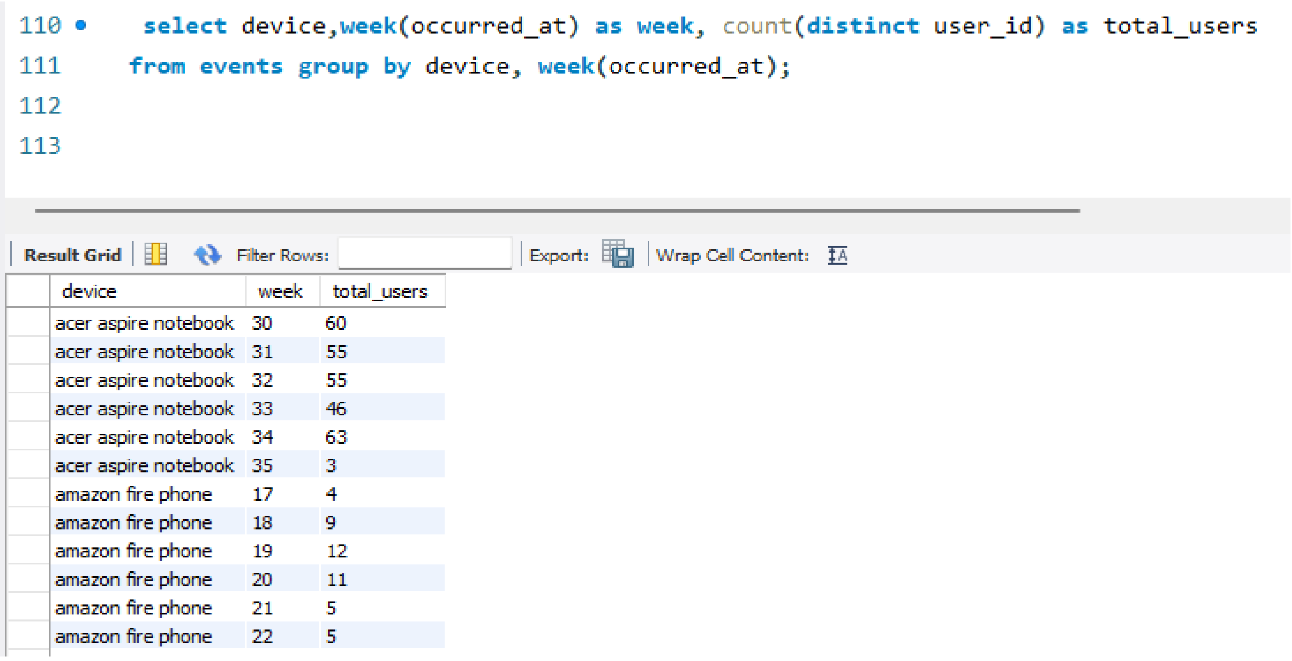


```sql
select  user_id,week(occurred_at) as `week`,count(event_name) as engagements from events
group by user_id, week(occurred_at) order by user_id;
```

user_id	week	engagements
4	19	4
4	20	8
4	21	29
4	22	4
4	23	15
4	24	8
4	25	7
4	26	10
4	27	8
8	17	2
8	18	15
8	19	3

Above query returns the number of times a user engaged in a particular week

B) Write an SQL query to calculate the weekly engagement per device.



```sql
select device,week(occurred_at) as week, count(distinct user_id) as total_users
from events group by device, week(occurred_at);
```

device	week	total_users
acer aspire notebook	30	60
acer aspire notebook	31	55
acer aspire notebook	32	55
acer aspire notebook	33	46
acer aspire notebook	34	63
acer aspire notebook	35	3
amazon fire phone	17	4
amazon fire phone	18	9
amazon fire phone	19	12
amazon fire phone	20	11
amazon fire phone	21	5
amazon fire phone	22	5

The above query returns the weekly engagement per device.

Project 4

Hiring Process Analytics

Problem statement:

Imagine you're a data analyst at a multinational company like Google. Your task is to analyze the company's hiring process data and draw meaningful insights from it. The hiring process is a crucial function of any company, and understanding trends such as the number of rejections, interviews, job types, and vacancies can provide valuable insights for the hiring department.

As a data analyst, you'll be given a dataset containing records of previous hires. Your job is to analyze this data and answer certain questions that can help the company improve its hiring process.

From the given data set

For the salary, first quartile(Q1) =25460.5

Third quartile (Q3) = 74438 Inter quartile range (IQR) = 48977.5

Upper bound =147904.25Lower bound = -48005.75

From the above information the identified outliers are with application_id’s 649039,795330, 874368.

So for obtaining accurate results removing these 3 outliers from the data.

There are 15 rows where gender was blank ,1 row with post name not specified. In order to consider these things into account random distribution of gender has been done even postname also filled. Zero salary considered for a blank in salary column

A) Determine the gender distribution of hires. How many males and females have beenhired by the company?

No of males hired= 2564 No of
females hired = 1862

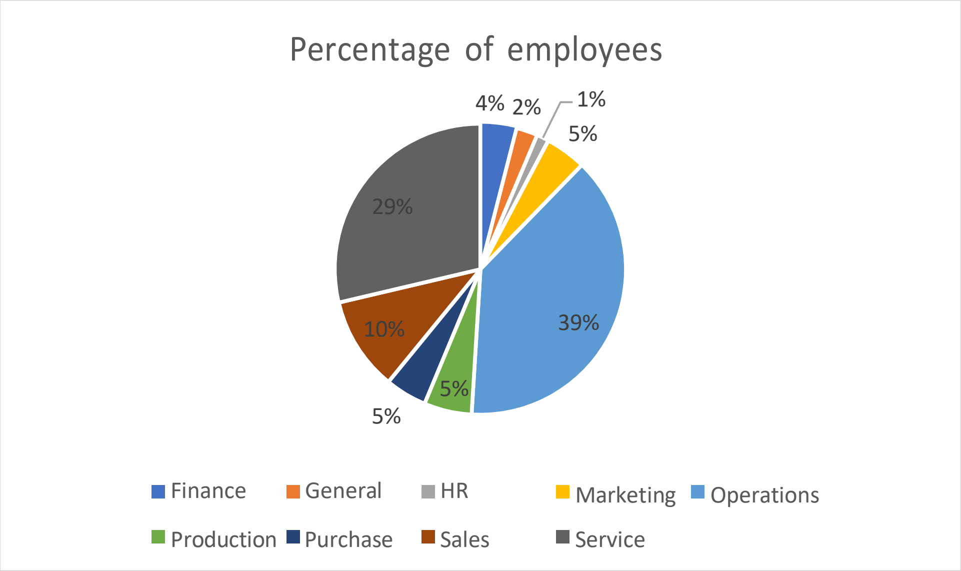
B) What is the average salary offered by this company? Use Excel functions to calculatethis.

Average salary offered by the company= 49871.37

C) Create class intervals for the salaries in the company. This will help you understandthe salary distribution.

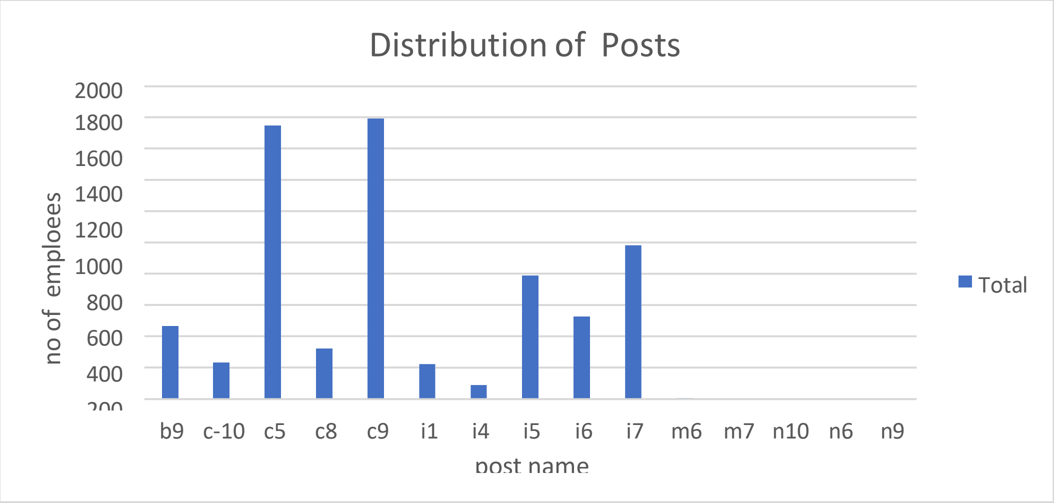
After removing the outliers the max salary is 99967 and min salary is 0.Therefore theclass intervals will be
0-10000
10001-20000
20001-30000
30001-40000
40001-50000
50001-60000
60001-70000
70001-80000
80001-90000
90001-100000

D) Use a pie chart, bar graph, or any other suitable visualization to show the proportionof people working in different departments.



The above pie chart shows the percentage of employees in each department.

E) Use a chart or graph to represent the different position tiers within the company.This will help you understand the distribution of positions across different tiers.



The above chart gives the information on distribution of posts in the company.

Project-5

IMDB Movie Analysis

Problem Statement:

The dataset provided is related to IMDB Movies. A potential problem to investigate could be: "What factors influence the success of a movie on IMDB?" Here, success can be defined by high IMDB ratings. The impact of this problem is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.

For the analysis duplicate rows and rows with blanks cells has been removed. After the cleaning, the count of movies in the data set are 3723.

A) Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

Genres	Count of movies for each genre
Action	951
Adventure	773
Fantasy	504
Sci-Fi	492
Thriller	1103
Romance	850
Comedy	1455
Family	440
Mystery	378
Drama	1876
Musical	96
Western	57
History	147
Documentary	45
Horror	386
Animation	196
Crime	704
Biography	238

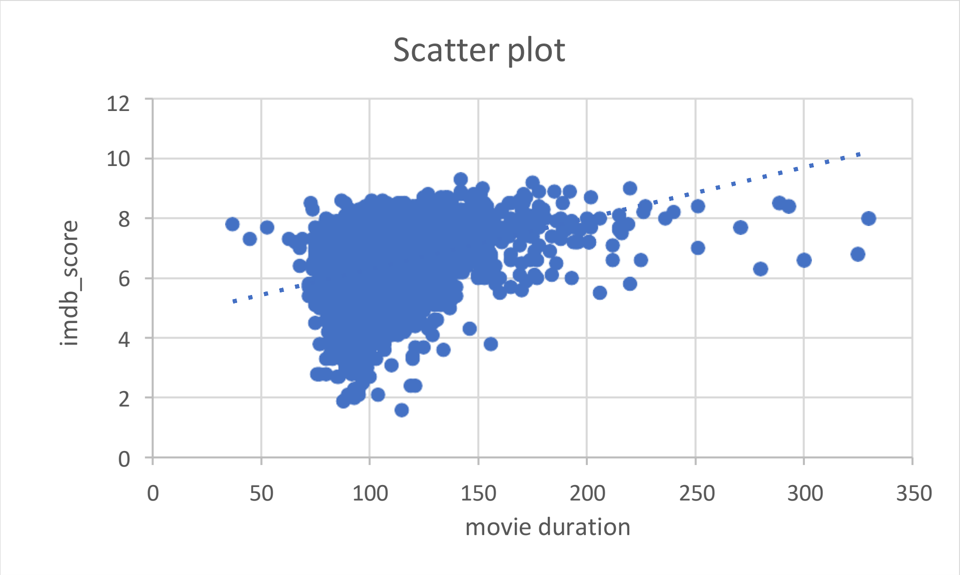
sport	147
war	150
film-noir	1
Music	149

The most common genres are Drama, comedy, thriller , action and romance.

Genre	**mean**	**std dev**	**Range**	**Variance**	**median**	**mode**
Action	6.289905	1.032168	6.9	1.065372	6.3	6.1
Adventure	6.452393	1.112731	6.6	1.238171	6.6	6.7
animation	6.702551	0.989506	5.8	0.979122	6.8	6.7
biography	7.157563	0.69252	4.4	0.479584	7.2	7
comedy	6.18811	1.032446	6.9	1.065945	6.3	6.7
crime	6.541903	0.981894	6.9	0.964116	6.6	6.6
documentary	6.988889	1.384693	6.9	1.917374	7.4	7.7
drama	6.792537	0.890811	7.2	0.793544	6.9	6.7
family	6.216364	1.163748	6.7	1.35431	6.3	6.7
fantansy	6.285317	1.127067	6.7	1.270281	6.4	6.7
film-noir	7.7	0	0	0	7.7	7.7
history	7.157823	0.667127	3.4	0.445058	7.2	7.7
horror	5.922539	0.997105	6.3	0.994218	6	5.9
music	6.336913	1.23208	6.9	1.51802	6.5	6.5
musical	6.596875	1.101908	6.4	1.214201	6.75	6.2
mystery	6.480688	1.003031	5.5	1.006072	6.5	6.6
romance	6.435059	0.956655	6.4	0.915189	6.5	6.5
scifi	6.325813	1.158096	6.9	1.341186	6.4	6.7
sport	6.589116	1.044039	6.3	1.090018	6.8	7.2
thriller	6.378422	0.967281	6.3	0.935632	6.4	6.5
war	7.062667	0.802308	4.3	0.643698	7.1	7.1
western	6.812281	0.941137	4.2	0.885739	6.8	6.8

B) Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

The avg movie duration is 110.2635
Median is 106
Standard deviation is 22.67832



From the above scatter plot one thing can be noticed from the trendline , there is positive relationship between imdb score and movie duration but this may not valid for long duration.

Highest imdb score is observed between above 100 and below 200.Lowest score is observed between 100 to 150.

From the graph the optimal duration is between 140 to 200.

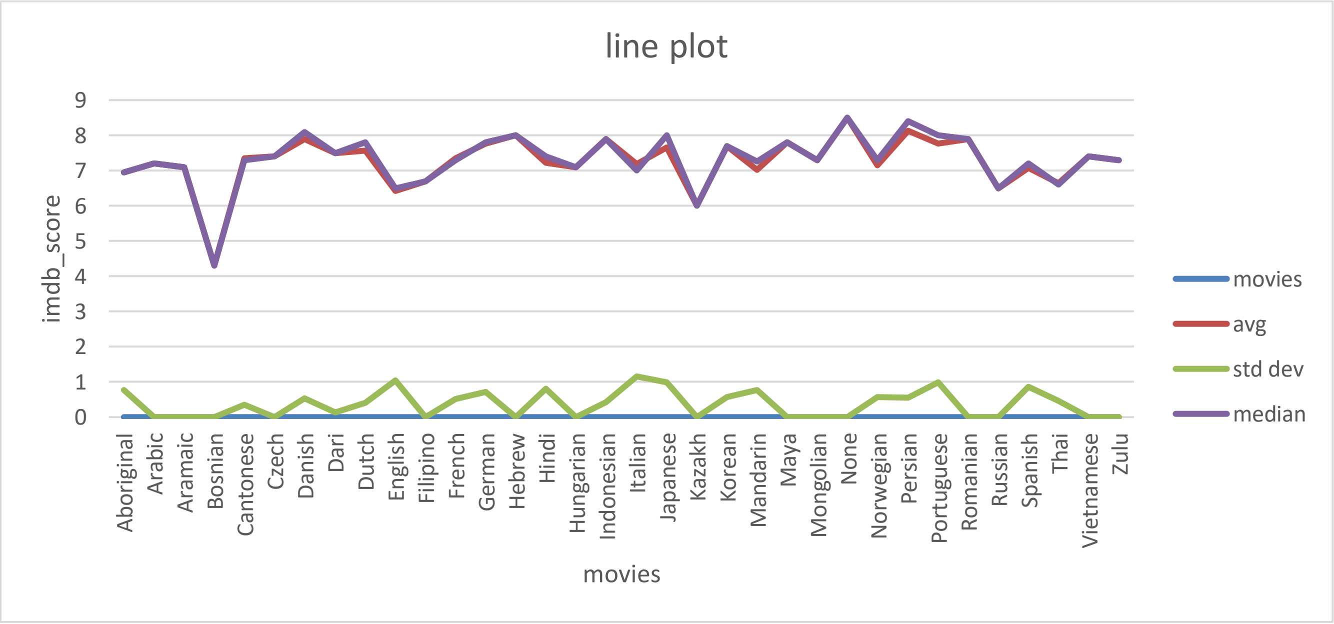
Considering only the movie duration for judging the imdb score will result in error.but by considering one parameter this thing can be noticed.

C) Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

Languages	Occurence of Languages	Average of imdb_score	StdDev of imdb_score	Median of imdb_score
Aboriginal	2	6.95	0.777817459	6.95
Arabic	1	7.2	0	7.2
Aramaic	1	7.1	0	7.1
Bosnian	1	4.3	0	4.3
Cantonese	7	7.34	0.35	7.3
Czech	1	7.4	0	7.4

Danish	3	7.9	0.529150262	8.1
Dari	2	7.5	0.141421356	7.5
Dutch	3	7.57	0.40	7.8
English	3566	6.43	1.05	6.5
Filipino	1	6.7	0	6.7
French	34	7.36	0.52	7.3
German	10	7.77	0.711883261	7.8
Hebrew	1	8	0	8
Hindi	5	7.22	0.801249025	7.4
Hungarian	1	7.1	0	7.1
Indonesian	2	7.9	0.424264069	7.9
Italian	7	7.19	1.16	7
Japanese	10	7.66	0.990173947	8
Kazakh	1	6	0	6
Korean	5	7.7	0.570087713	7.7
Mandarin	14	7.02	0.77	7.25
Maya	1	7.8	0	7.8
Mongolian	1	7.3	0	7.3
None	1	8.5	0	8.5
Norwegian	4	7.15	0.574456265	7.3
Persian	3	8.13	0.55	8.4
Portuguese	5	7.76	0.978774744	8
Romanian	1	7.9	0	7.9
Russian	1	6.5	0	6.5
Spanish	23	7.08	0.86	7.2
Thai	3	6.63	0.45	6.6
Vietnamese	1	7.4	0	7.4
Zulu	1	7.3	0	7.3

From the above table the most common language used is English. Even though most of the movies are in English their average imdb score is quite less compared to all others and standard deviation for English language is also high.



Higher the standard deviation signifies that there are many values which are deviating from themean values.

Considering language alone for knowing the effect on imdb score is not appropriate.

Mostly higher avg imdb score is observed for movies with lower standard deviation. As the count of movies of a particular language is increasing avg imdb score is decreasing.

D) Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculation.

The top 10 directors based on highest avg imdb score are

1) Akira Kurosawa
2) Charles Chaplin
3) Tony Kaye
4) Alfred Hitchcock
5) Damien Chazelle
6) Majid Majidi
7) Ron Fricke
8) Sergio Leone
9) Christopher Nolan
10) Asghar Farhadi

8.7 is the avg imdb score of Akira Kurosawa with 100th percentile which means there are 100% of values that are under this value.

percentile	Avg imdb score	Directors
100	8.7	Akira Kurosawa
99	8.2105	Elia Kazan
		George Roy Hill
		Joshua Oppenheimer
		Juan JosÃ© Campanella
		Quentin Tarantino
98	8	Ari Folman
		David Lean
		Michel Hazanavicius
		Stephen Chbosky
		Vincent Paronnaud
97	7.8426	Alejandro G. IÃ±Ã¡rritu
96	7.8	Stanley Kubrick
		Alfonso CuarÃ³n
		Bernardo Bertolucci
		Christian Carion
		Giuseppe Tornatore
		Henry Alex Rubin
		Jacques Perrin
		Jim Abrahams
		Josh Boone
		Mark Herman
		Mark Sandrich
		Mike van Diem

The above table shows the details of directors who are within the top 5 percentile. This reflects the consistency of imdb score of different movies of the directors.

E) Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

The correlation coefficient between budget and gross earnings is 0.0983181

Profit margin for individual movie is provided in the excel sheet attached along with the pdf.

Avatar is the movie with Max profit margin of 523505847.

Profit Margin	**Movie_title**
523505847	Avatar
502177271	Jurassic World
458672302	Titanic
449935665	Star Wars: Episode IV - A New Hope
424449459	E.T. the Extra-Terrestrial
403279547	The Avengers
377783777	The Lion King
359544677	Star Wars: Episode I - The Phantom Menace
348316061	The Dark Knight
329999255	The Hunger Games

The above table shows the list of top 10 movies with highest profit margin

Excel file link..\Desktop\trainity\IMDB_Movies(AutoRecovered).xlsx

PROJECT 6
Bank Loan Case Study

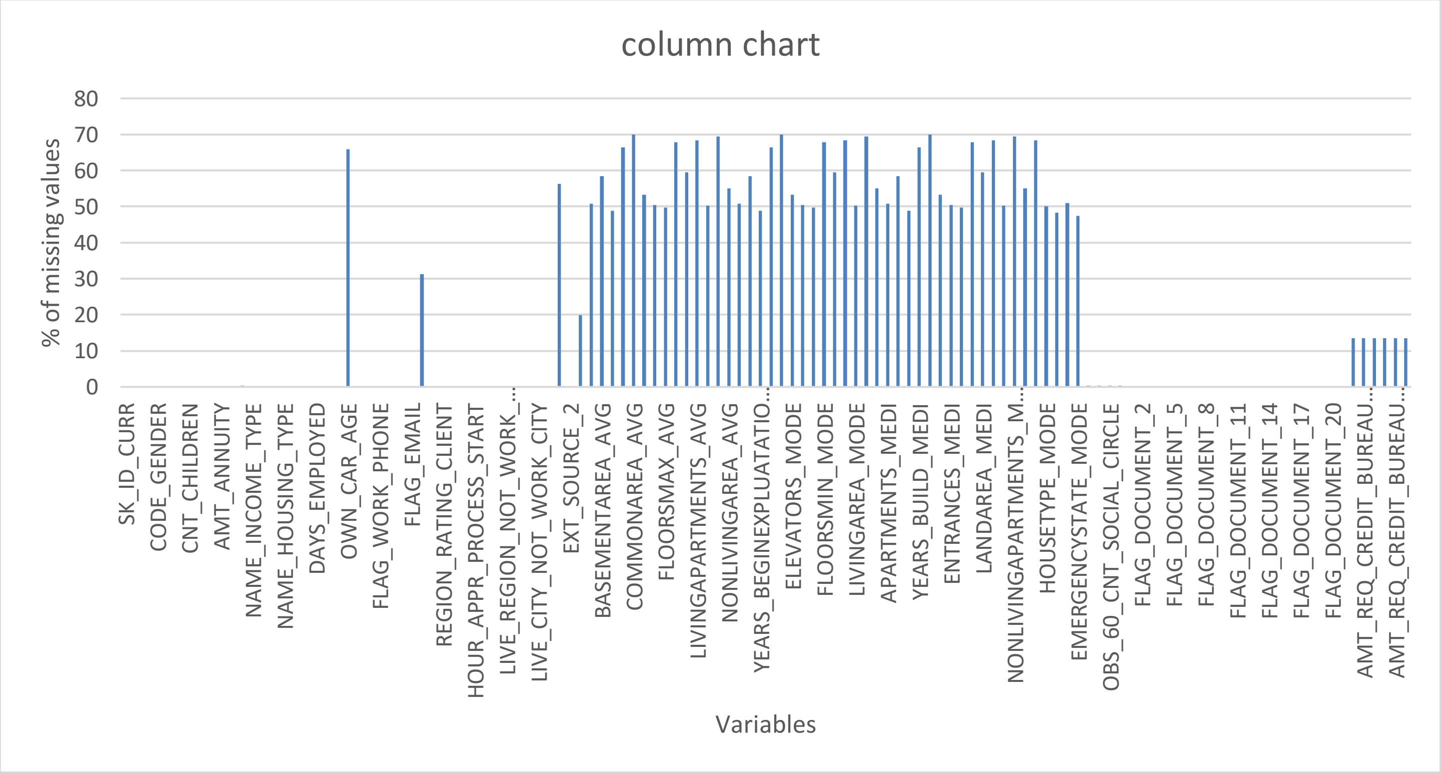
Description:

Imagine you're a data analyst at a finance company that specializes in lending various types of loans to urban customers. Your company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. Your task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

Business Objectives:

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.

A) Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.



The above bar chart represents the % of missing values of different variables.

In order to find the missing values countblank is used over each column. further calculated the blank cells percentage.

20 % of rows are around 9999

30% of rows are around 14999

After identifying blanks cell in each column, the columns with blanks above 20 % has been removed as it misleads the analysis

Initial columns are 123

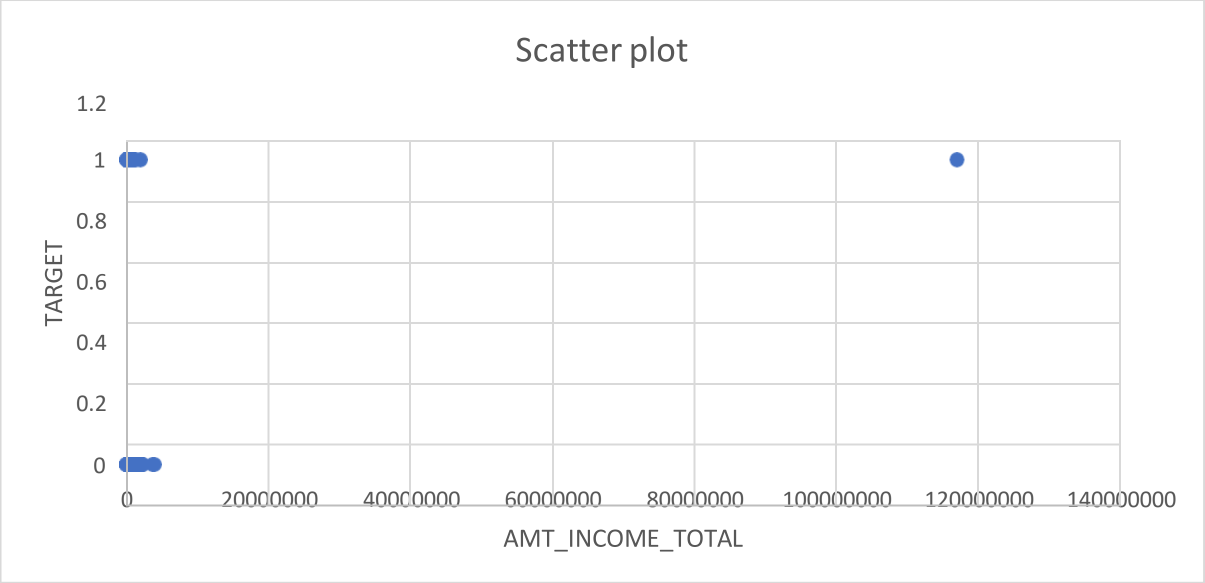
After removing few columns ,total columns are 72

Additional columns were also removed even though they are with less than 20 % blanks as there is no significance of it.

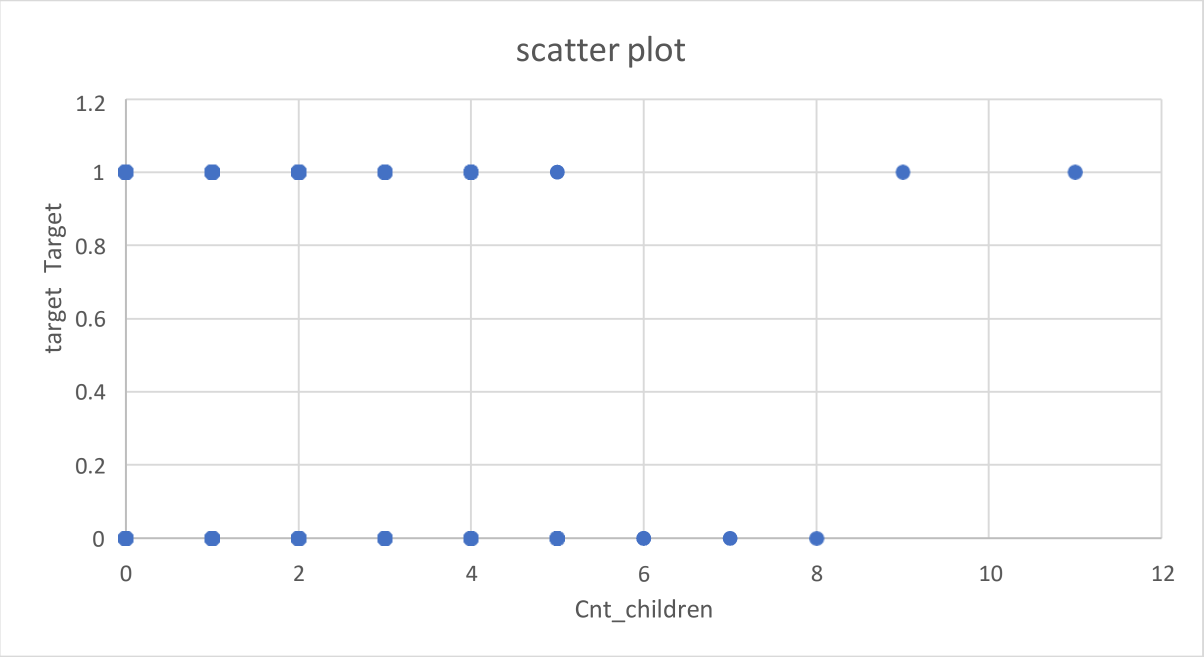
Median has been imputed in place of blanks for columns that have less than 20% of blanks.

B) Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

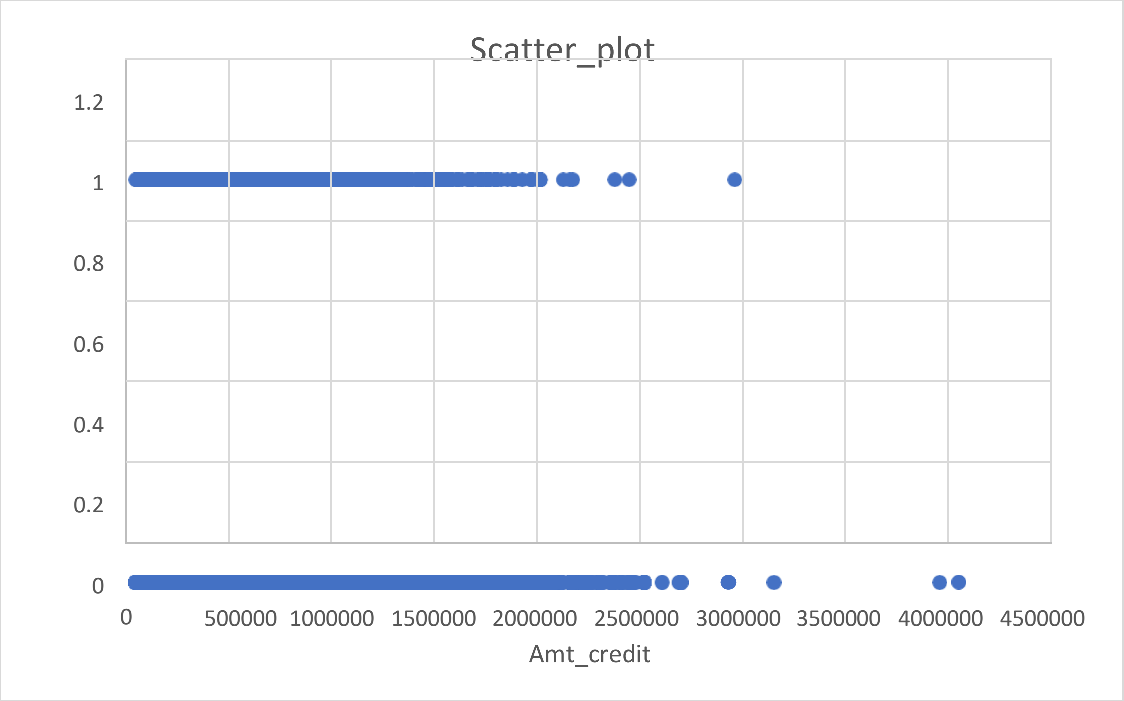
Outliers has been identified in the columns Amt_income_total, Cnt_children, days_employed.



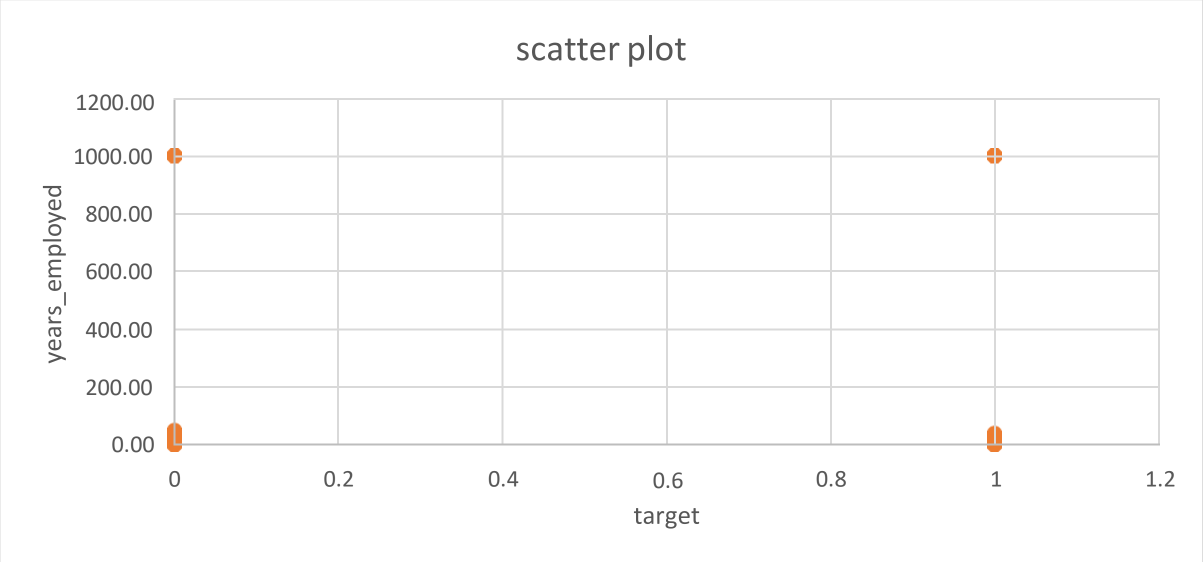
From the plot one outlier can be spotted which is exceptional income of the client. If it is the income the client may not come for loan.



8 can be considered as the maximum children.so other 2 are the outliers.

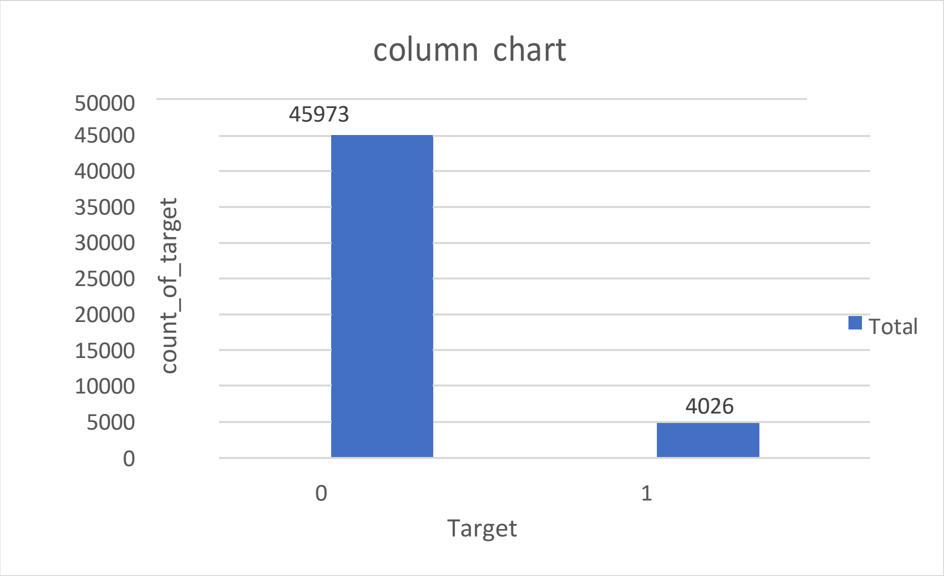


For the target 0 there are 2 points which are far away.this might not be considered as outlier as this can be possible credit amount for big businesses.



At target 1 we can find one outlier which shows that the employed work for 1000 years .xo this is an outlier.

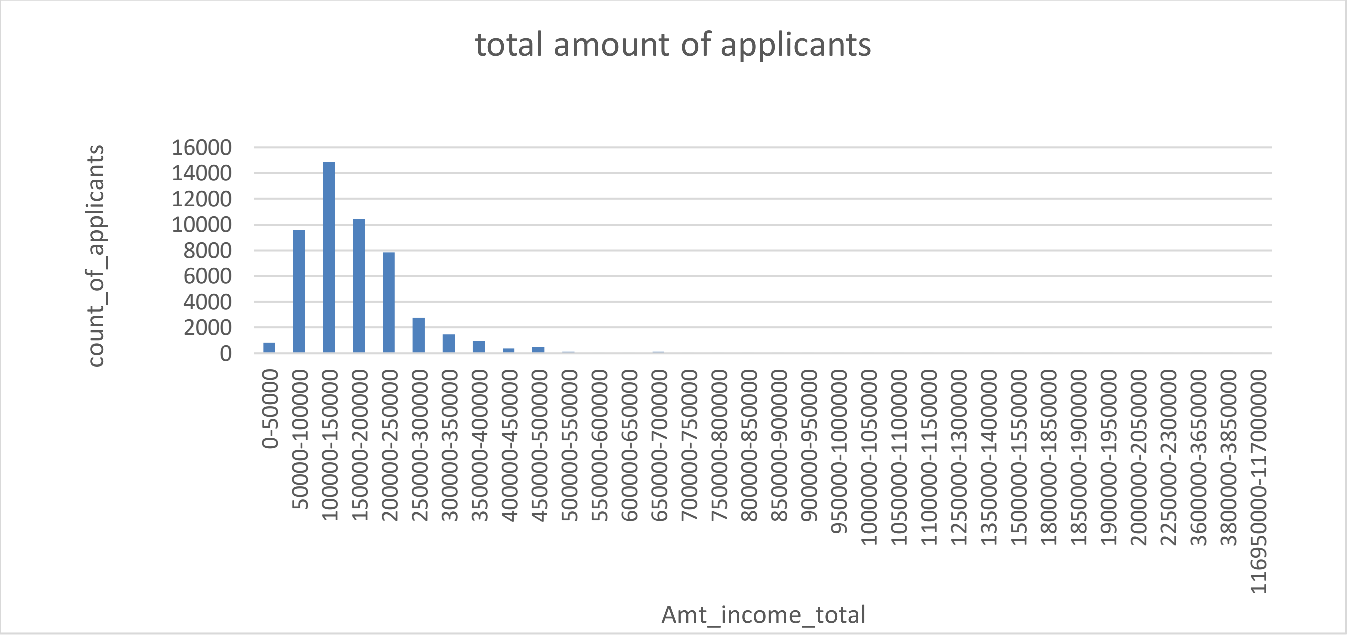
C) Determine if there is data imbalance in the loan application dataset and calculate theratio of data imbalance using Excel functions.



Target	ratio	percentage
0	11.42	91.95
1	1	8.05

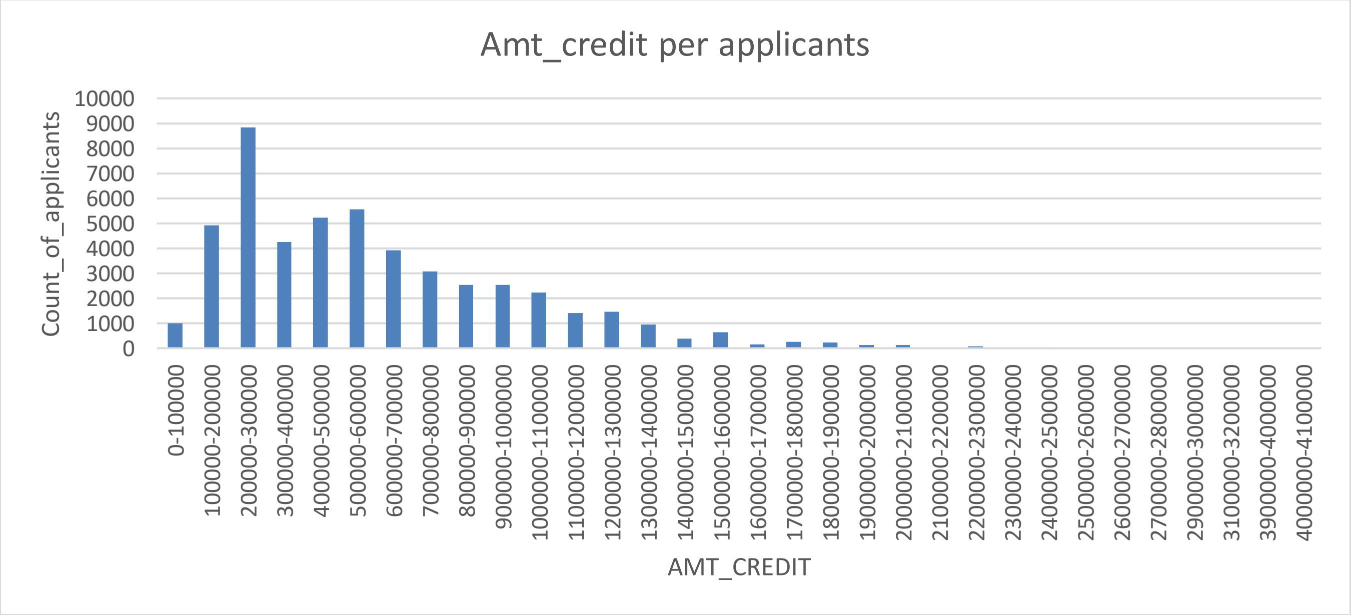
There is data imbalance between the target variables.

D) Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.



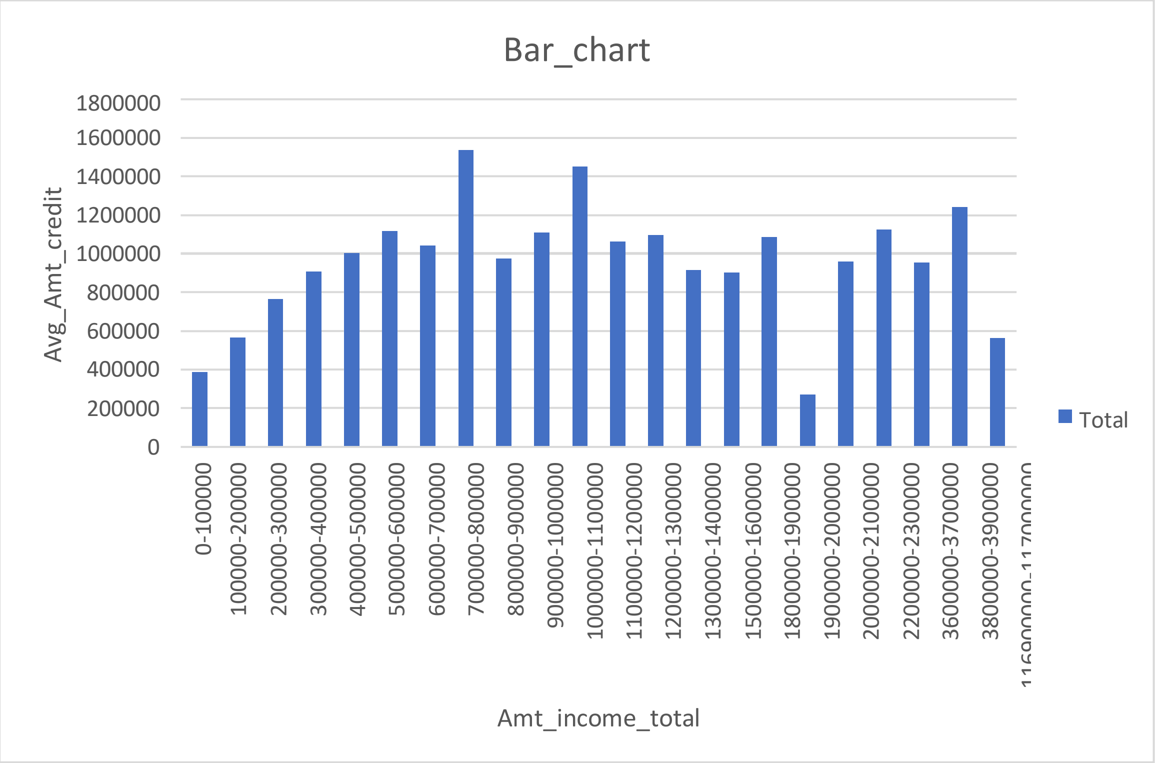
Univariate analysis

From the above graph one thing can be noted is highest count of applicants is observed for whom the income range is between 1-1.5 lakhs.



Univariate analysis

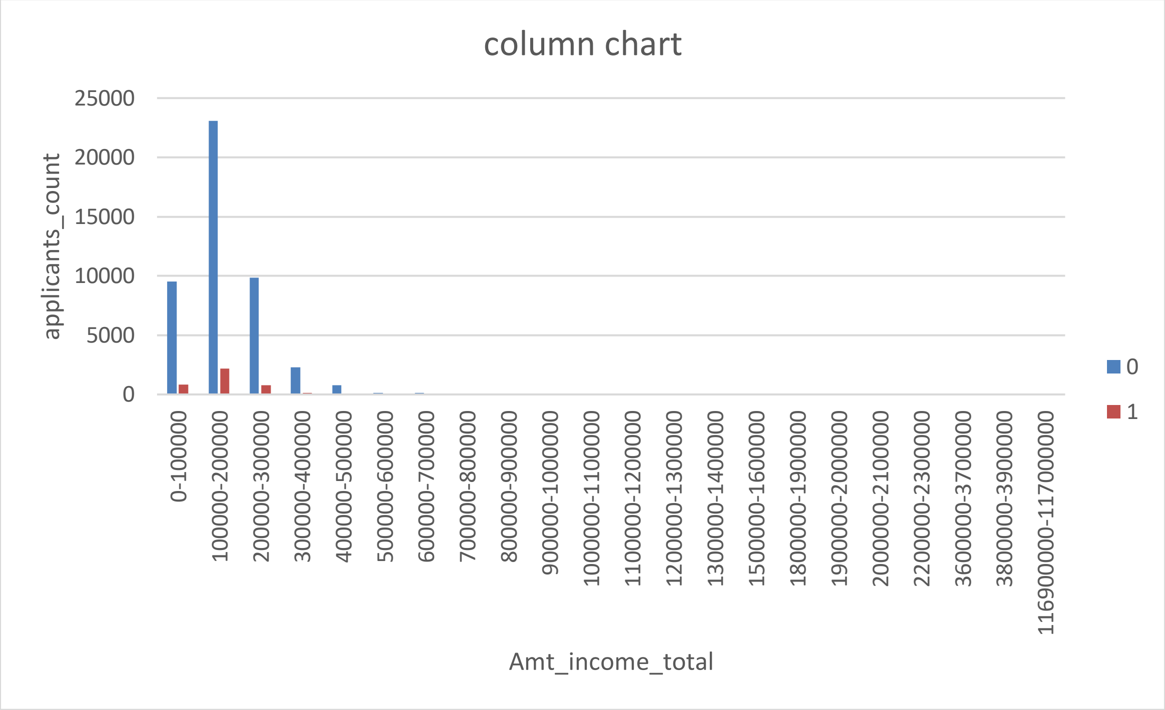
Highest number of count of applicants were there for whom the amount credited is between 2-3 lakhs.

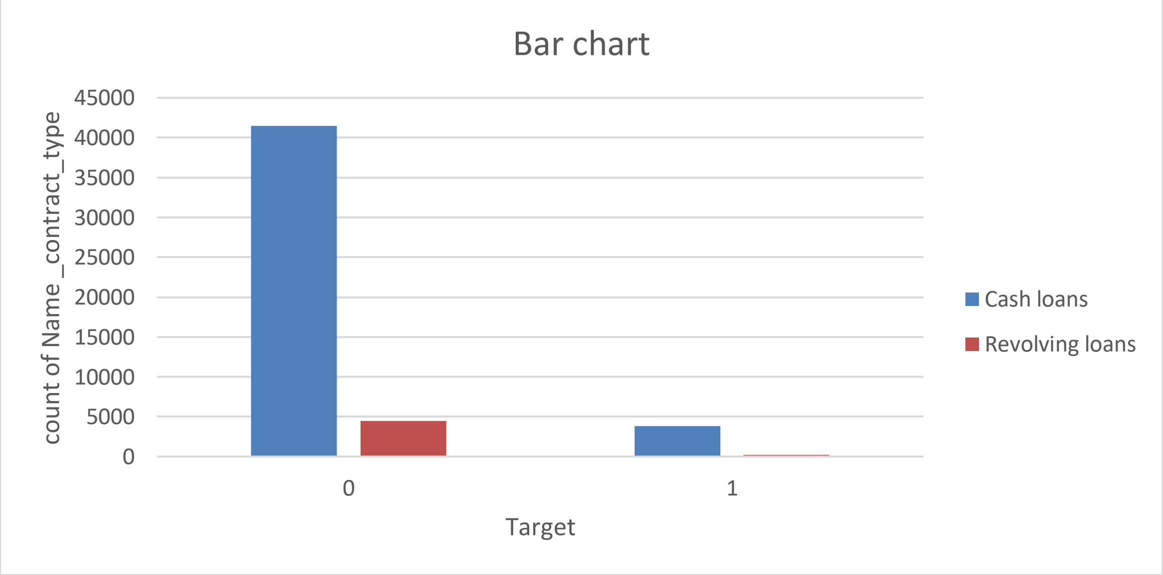


Bivariate Analysis

The above graph shows the bivariate analysis between total income v/s amount credited of applicants. Highest amount credited for the applicants whose income is between 7-8 lakhs and 10-11 lakhs.

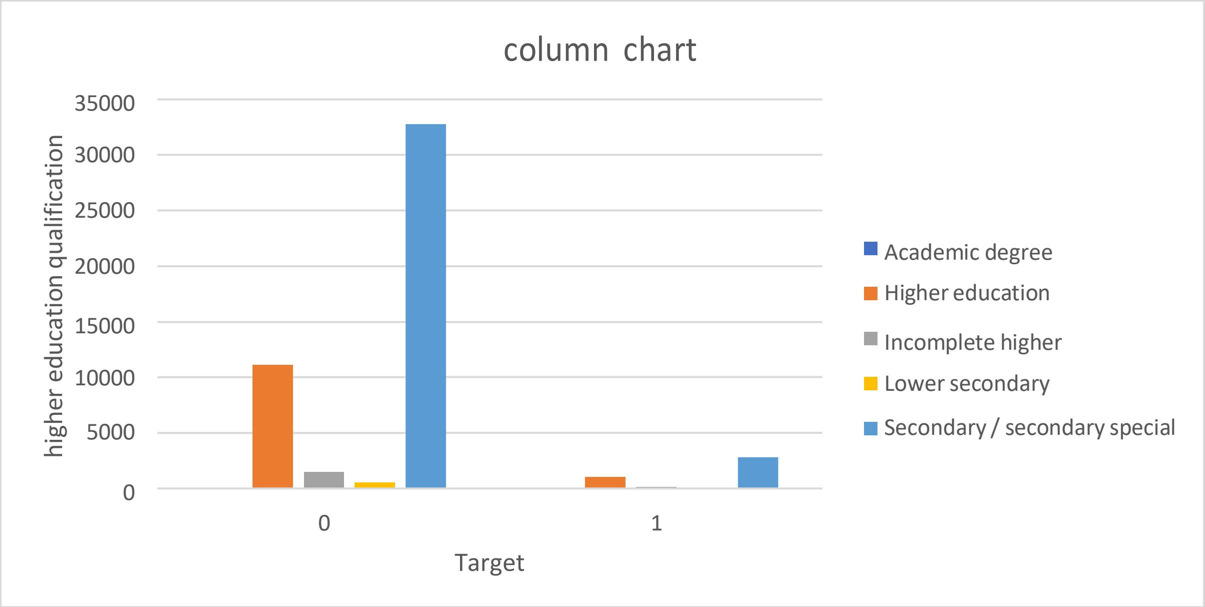
Segmented univariate analysis:

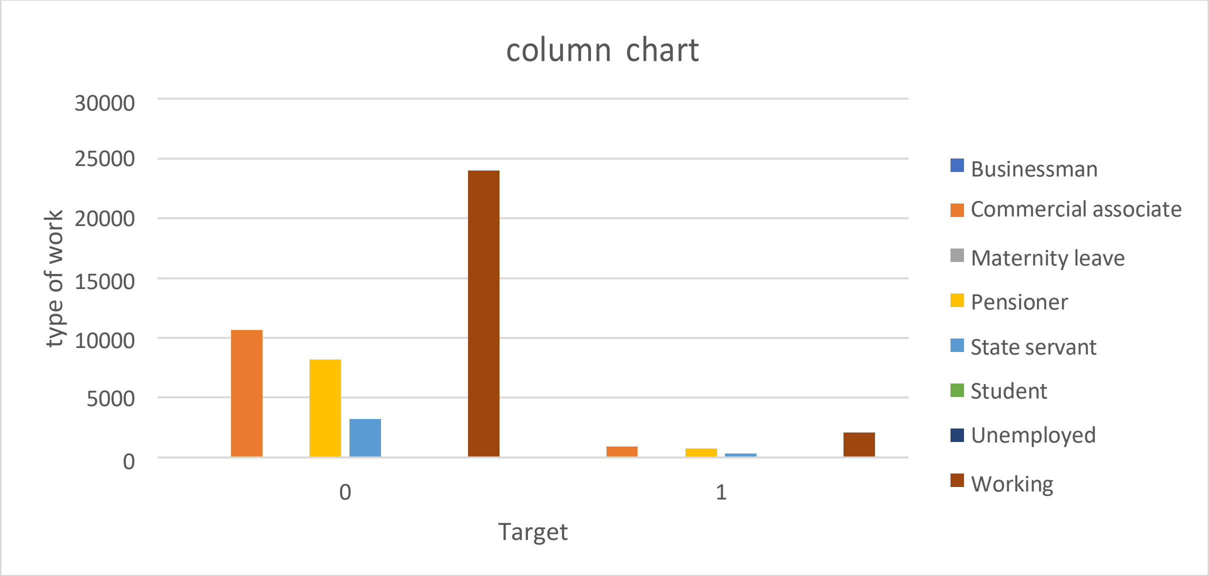


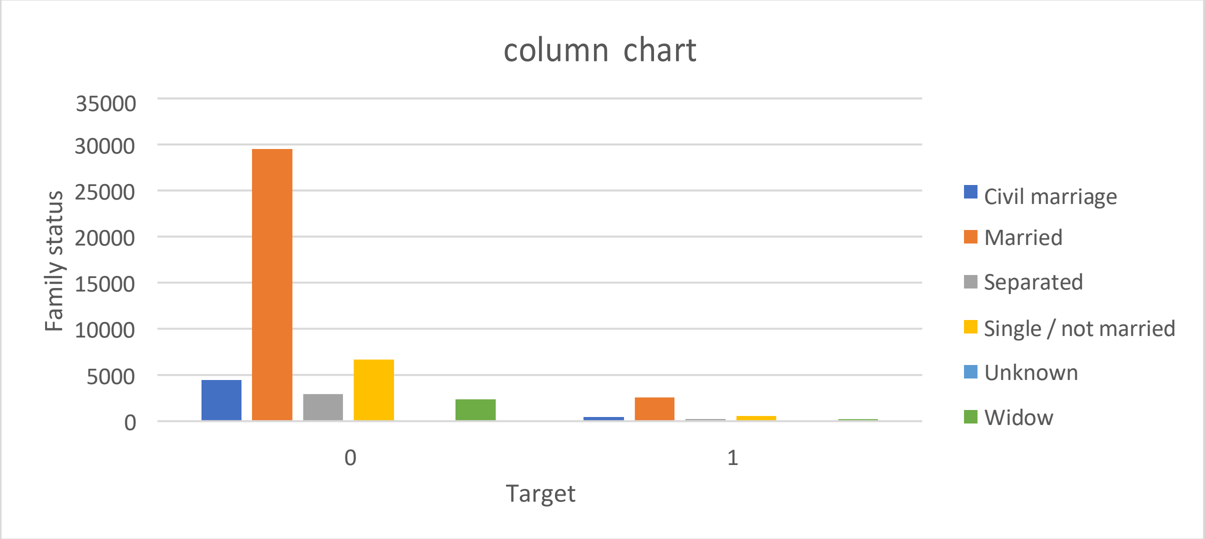


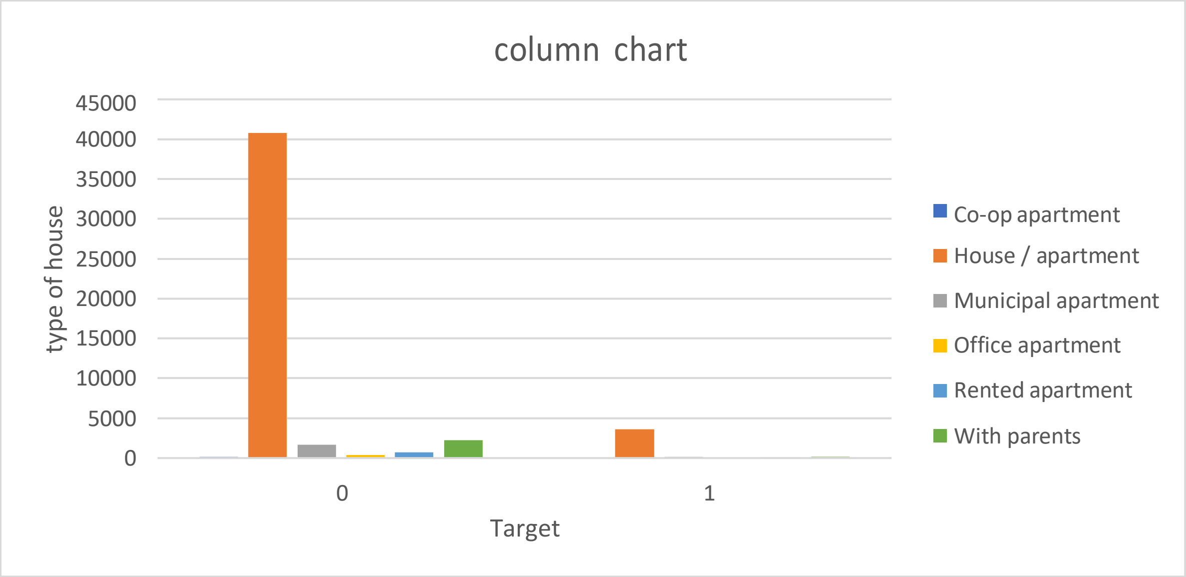
Segmented Univariate analysis

The above graph shows the what kind of loan most of the different target applicants has taken.


column chart
higher education qualification
35000
30000
25000
20000
15000
10000
5000
0
0
1
Target
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special


column chart
type of work
30000
25000
20000
15000
10000
5000
0
0
1
Target
Businessman
Commercial associate
Maternity leave
Pensioner
State servant
Student
Unemployed
Working


column chart
Family status
35000
30000
25000
20000
15000
10000
5000
0
0
1
Target
Civil marriage
Married
Separated
Single / not married
Unknown
Widow


column chart
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
type of house
0
1
Target
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents

E) Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

AMT_INCOME_TOTAL	1.00	0.38	0.04	-0.16	-0.07	0.04	-0.21	-0.07	0.18	0.45
AMT_CREDIT	0.38	1.00	0.01	-0.07	0.05	0.06	-0.10	-0.01	0.10	0.77
CNT_CHILDREN	0.04	0.01	1.00	-0.25	-0.34	0.88	0.02	-0.18	-0.02	0.03
Days_employed1	-0.16	-0.07	-0.25	1.00	0.62	-0.23	0.04	0.21	-0.01	-0.11
Days_birth1	-0.07	0.05	-0.34	0.62	1.00	-0.28	-0.01	0.34	0.03	-0.01
CNT_FAM_MEMBERS	0.04	0.06	0.88	-0.23	-0.28	1.00	0.02	-0.17	-0.02	0.08
REGION_RATING_CLIENT	-0.21	-0.10	0.02	0.04	-0.01	0.02	1.00	-0.08	-0.54	-0.13
DAYS_REGISTRATION1	-0.07	-0.01	-0.18	0.21	0.34	-0.17	-0.08	1.00	0.06	-0.03
Region_population_relative	0.18	0.10	-0.02	-0.01	0.03	-0.02	-0.54	0.06	1.00	0.12
AMT_Annuity	0.45	0.77	0.03	-0.11	-0.01	0.08	-0.13	-0.03	0.12	1.00
	AMT_INCOME_TOTAL	**AMT_CREDIT**	**CNT_CHILDREN**	**Days_employed1**	**Days_birth1**	**CNT_FAM_MEMBERS**	**REGION_RATING_CLIENT**	**DAYS_REGISTRATION1**	**Region_population_relative**	**AMT_Annuity**

Correlation matrix for target 0
Top positive correlations for target 0 are:

1) Client family members-client children
2) Amt credit-Amt Annuity
3) Days birth-days employed
4) Amt_income_total – Amt_annuity
5) Amt_income_total- Amt_credit
6) Region rating client-days birth
7) Days employed- days registration

AMT_INCOME_TOTAL	1.00	0.02	0.02	0.01	-0.01	-0.01	0.01	0.01	-0.01
AMT_CREDIT	0.02	1.00	0.75	0.01	0.02	0.14	0.06	0.04	0.07
AMT_Annuity	0.02	0.75	1.00	0.03	-0.08	0.01	0.08	-0.02	0.07
CNT_CHILDREN	0.01	0.01	0.03	1.00	-0.19	-0.25	0.89	-0.15	-0.02
Days_employed1	-0.01	0.02	-0.08	-0.19	1.00	0.59	-0.18	0.19	0.01
Days_birth1	-0.01	0.14	0.01	-0.25	0.59	1.00	-0.20	0.29	0.02
CNT_FAM_MEMBERS	0.01	0.06	0.08	0.89	-0.18	-0.20	1.00	-0.15	-0.02
DAYS_REGISTRATION1	0.01	0.04	-0.02	-0.15	0.19	0.29	-0.15	1.00	0.05
Region_population_relative	-0.01	0.07	0.07	-0.02	0.01	0.02	-0.02	0.05	1.00
	AMT_INCOME_TOTAL	**AMT_CREDIT**	**AMT_Annuity**	**CNT_CHILDREN**	**Days_employed1**	**Days_birth1**	**CNT_FAM_MEMBERS**	**DAYS_REGISTRATION1**	**Region_population_relative**

Correlation matrix for target 1

Top positive correlations for target 1 are:

1) Cnt_family_members- cnt_children
2) Amt_credit – Amt_Annuity
3) Days employed -days birth
4) Days registration – days birth
5) Days employed – days registration

Link of excel sheet: ..\Desktop\trainity\application_data.xlsx

Project 7

Analyzing the Impact of Car Features on Price and Profitability

Description:

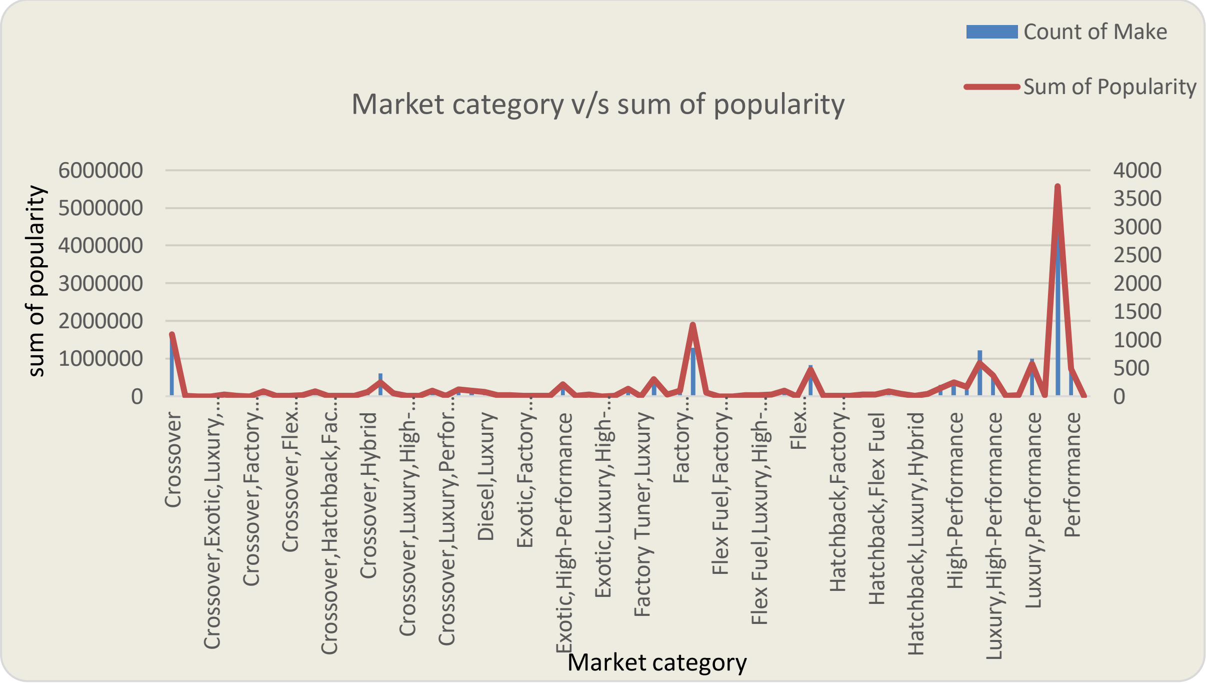
For the given dataset, as a Data Analyst, the client has asked How can a car manufacturer optimize pricing and product development decisions to maximize profitability while meeting consumer demand?
This problem could be approached by analyzing the relationship between a car's features, market category, 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.

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

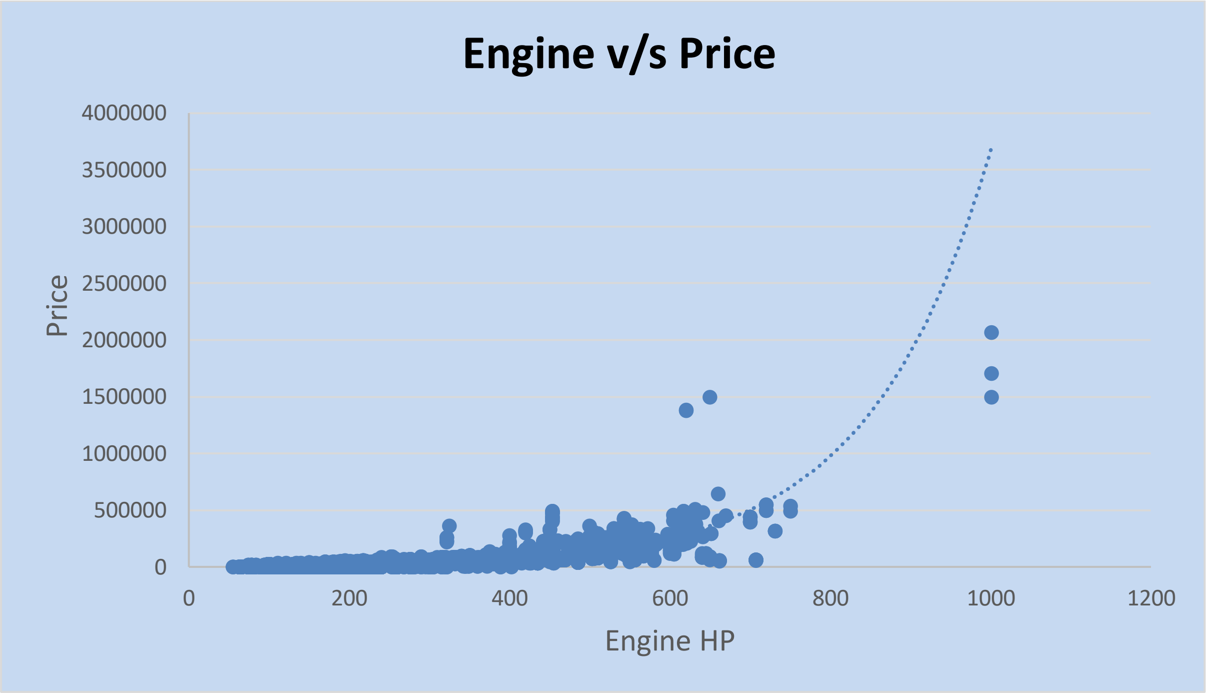
Market Category	Count of Make	Sum of Popularity
Crossover	1068	1644160
Crossover,Diesel	7	6111
Crossover,Exotic,Luxury,High-Performance	1	238
Crossover,Exotic,Luxury,Performance	1	238
Crossover,Factory Tuner,Luxury,High-Performance	26	47410
Crossover,Factory Tuner,Luxury,Performance	5	13037
Crossover,Factory Tuner,Performance	4	840
Crossover,Flex Fuel	64	132720
Crossover,Flex Fuel,Luxury	10	11732
Crossover,Flex Fuel,Luxury,Performance	6	9744
Crossover,Flex Fuel,Performance	6	33942
Crossover,Hatchback	72	120650
Crossover,Hatchback,Factory Tuner,Performance	6	12054
Crossover,Hatchback,Luxury	7	1428
Crossover,Hatchback,Performance	6	12054

The above table shows the sample data of how sum of popularity is varying along with themarket category.Full data can be found in the attached excel sheet.

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



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.



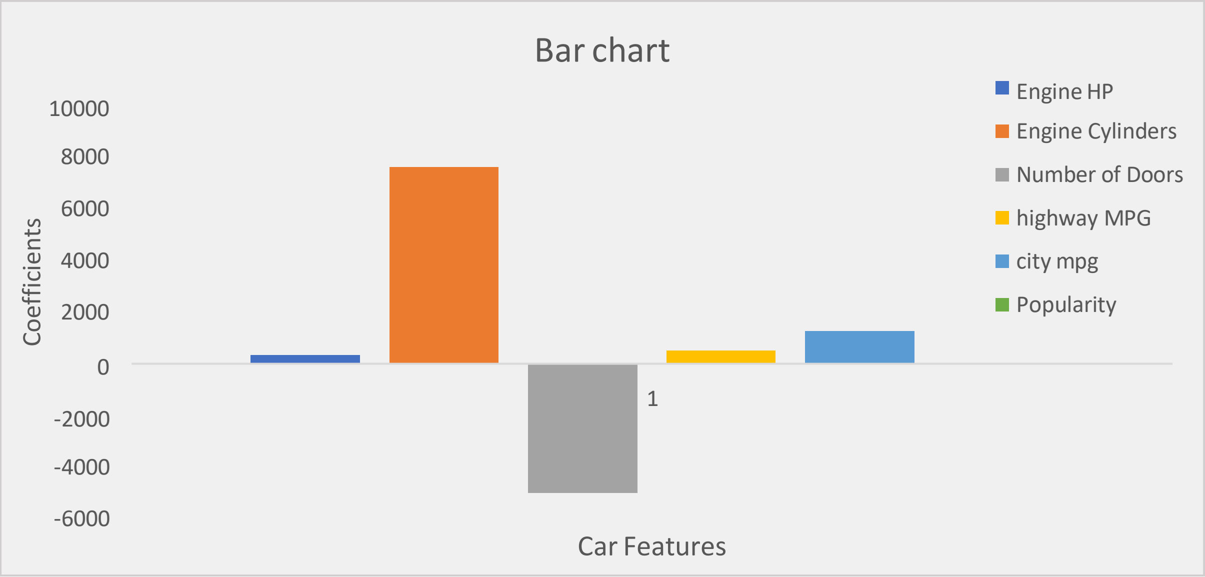
Task 3: Use regression analysis to identify the variables that have the strongest relationship with a car's price. Then create a bar chart that shows the coefficient values for each variable to visualize their relative importance.

Regression Statistics	
Multiple R	0.683387437
R Square	0.46701839
Adjusted R Square	0.466730032
Standard Error	45078.99614
Observations	11097

Anova

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	1.9747E+13	3.29117E+12	1619.578688	0
Residual	11090	2.25362E+13	2032115893		
Total	11096	4.22832E+13			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-97167.87274	3898.078612	-24.92711985	1.7136E-133	-104808.8004	-89526.94512	-104808.8004	-89526.94512
Engine HP	320.4942139	6.3774589	50.25421863	0	307.9932598	332.995168	307.9932598	332.995168
Engine Cylinders	7578.79133	461.2602827	16.43061762	5.88653E-60	6674.639109	8482.94355	6674.639109	8482.94355
Number of Doors	-4980.209981	496.4047724	-10.03255863	1.38198E-23	-5953.251655	-4007.168308	-5953.251655	-4007.168308
highway MPG	503.5834871	109.2773107	4.608307836	4.10488E-06	289.3805157	717.7864585	289.3805157	717.7864585
city mpg	1253.468123	125.6629389	9.974843293	2.46287E-23	1007.146405	1499.789841	1007.146405	1499.789841
Popularity	-3.553387511	0.297352947	-11.95006655	1.02989E-32	-4.136252193	-2.97052283	-4.136252193	-2.97052283

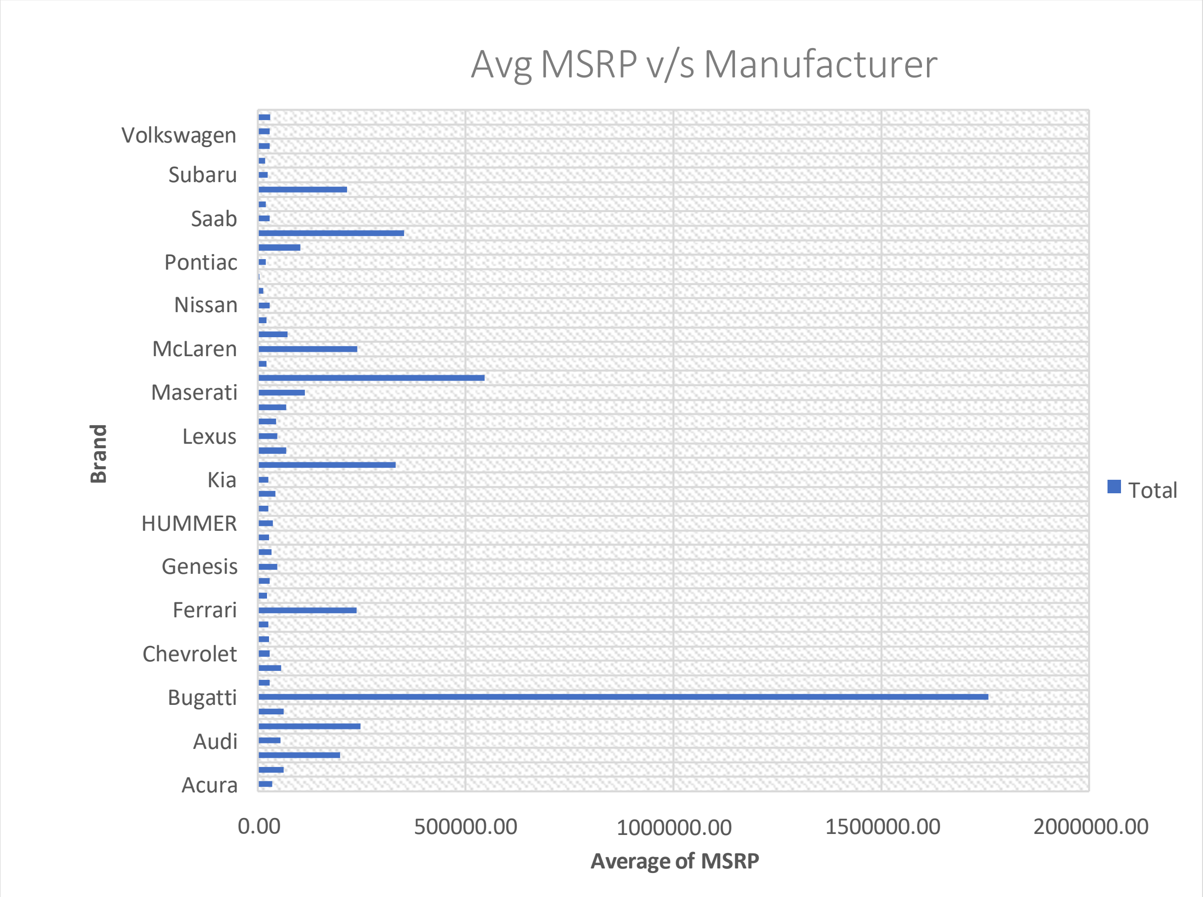


Task 4.A: Create a pivot table that shows the average price of cars for each manufacturer.

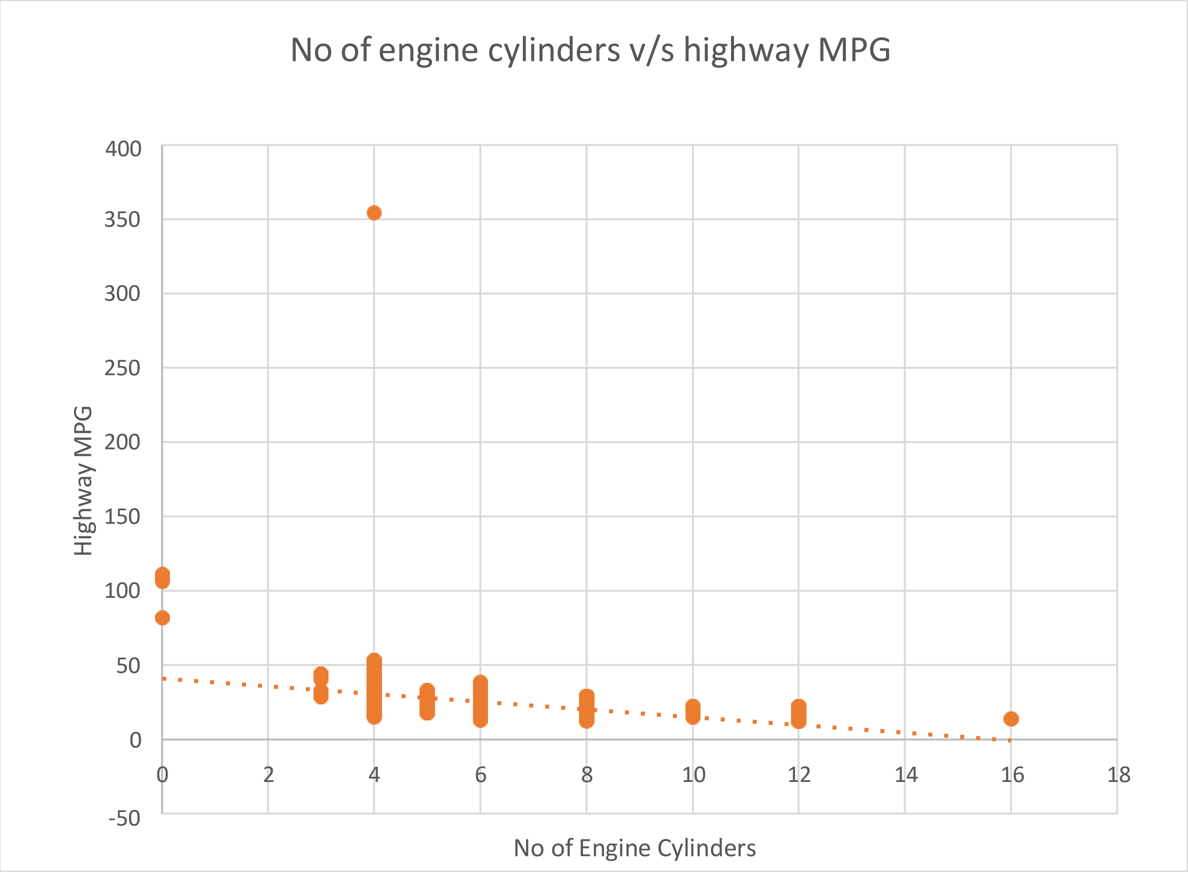
Brand	Average of MSRP
Acura	35087.49
Alfa Romeo	61600.00
Aston Martin	198123.46
Audi	54574.12
Bentley	247169.32
BMW	62162.56
Bugatti	1757223.67
Buick	29034.19
Cadillac	56368.27
Chevrolet	29000.22
Chrysler	26722.96
Dodge	24857.05

Ferrari	237383.82
FIAT	22206.02
Ford	28522.86
Genesis	46616.67
GMC	32444.09
Honda	26608.88
HUMMER	36464.41
Hyundai	24926.26
Infiniti	42640.27
Kia	25318.75
Lamborghini	331567.31
Land Rover	68067.09
Lexus	47549.07
Lincoln	43560.01
Lotus	68377.14
Maserati	113684.49
Maybach	546221.88
Mazda	20106.56
McLaren	239805.00
Mercedes-Benz	72135.03
Mitsubishi	21316.35
Nissan	28856.42
Oldsmobile	12843.80
Plymouth	3296.87
Pontiac	19800.04
Porsche	101622.40
Rolls-Royce	351130.65
Saab	27879.81
Scion	19932.50
Spyker	214990.00
Subaru	24240.67
Suzuki	18021.05
Toyota	28758.77
Volkswagen	28947.37
Volvo	29724.68
Grand Total	**41901.12**

Task 4.B: Create a bar chart or a horizontal stacked bar chart that visualizes the relationship between manufacturer and average price.



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.



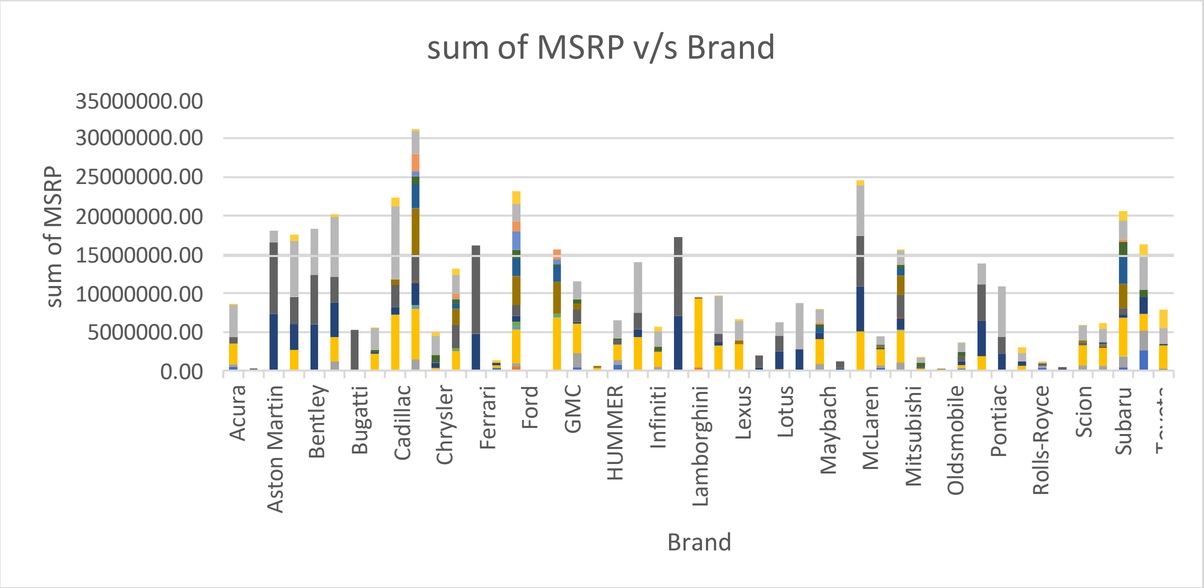
Task 5.B: Calculate the correlation coefficient between the number of cylinders and highwayMPG to quantify the strength and direction of the relationship.

The correlation coefficient between no of cylinders and highway MPG is -0.6147.

As there is negative sign so they have negative relation between them.

Dashboard Tasks

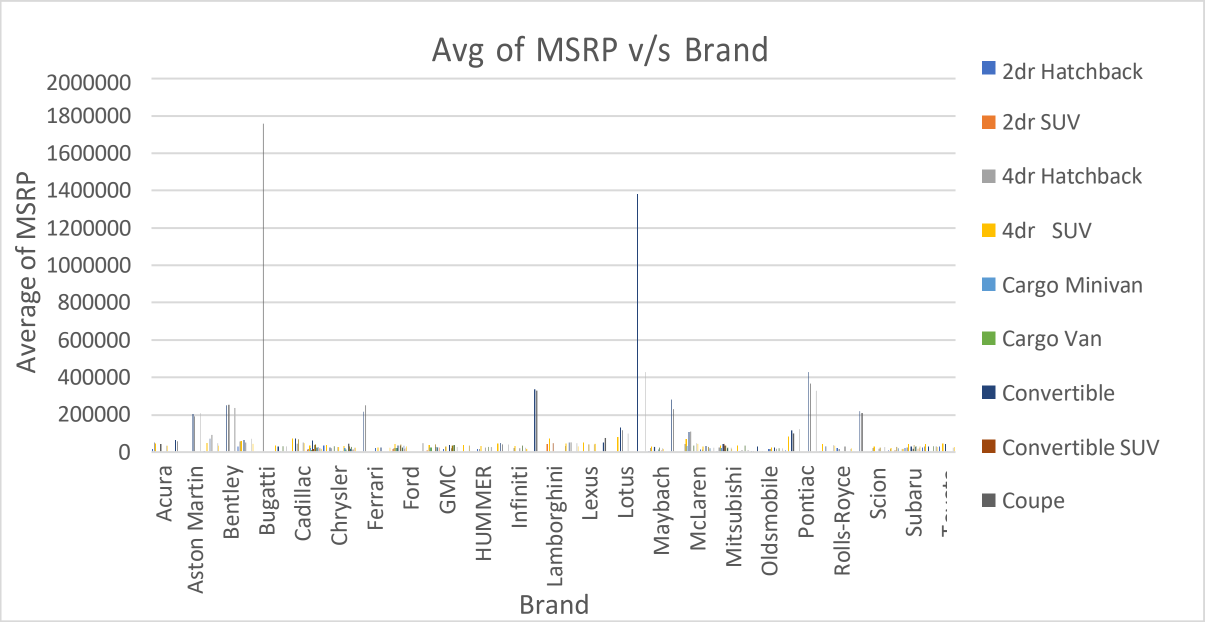
Task 1: How does the distribution of car prices vary by brand and body style?



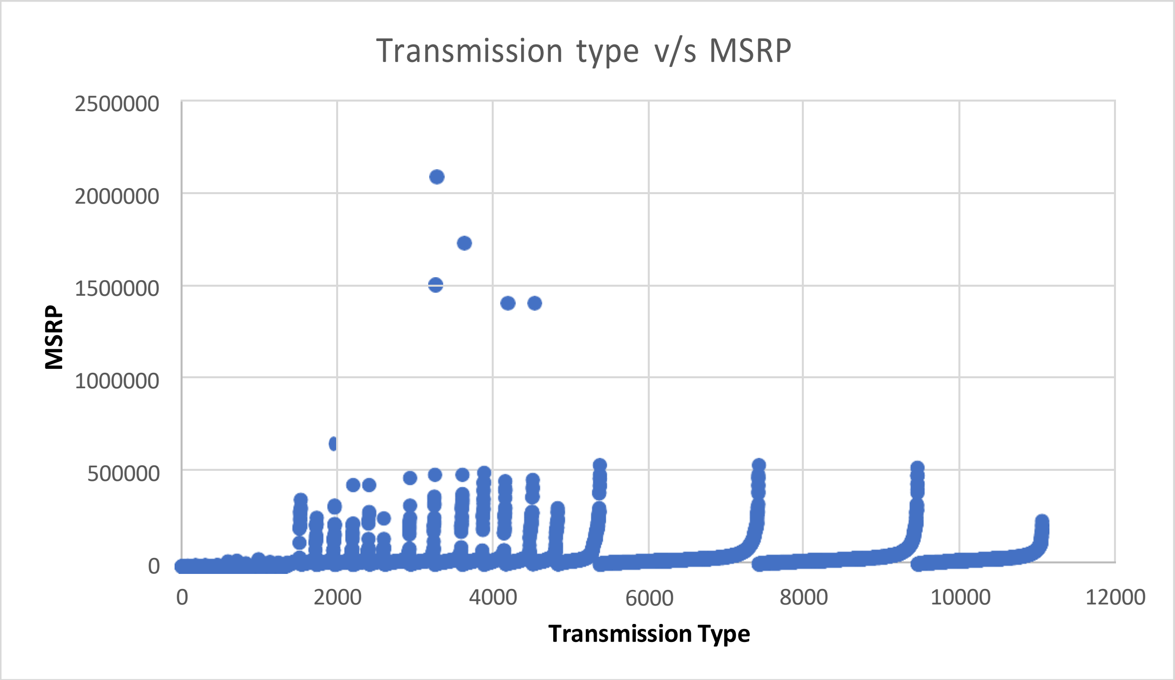
Task 2: Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?

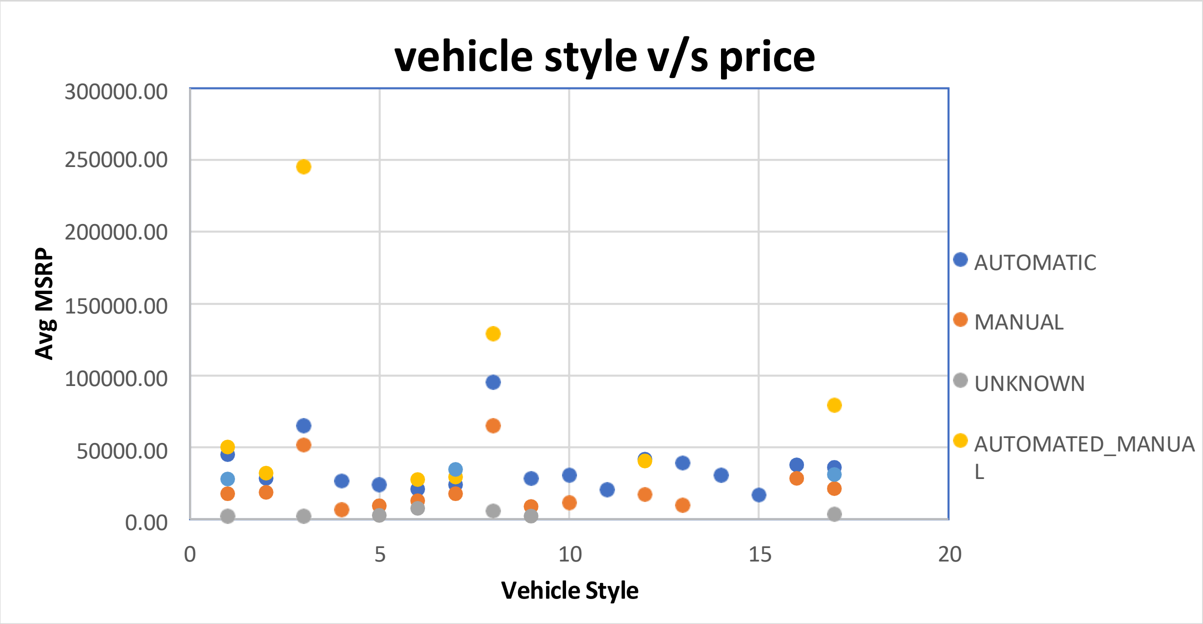
Highest average MSRP can be found from the graph is Bugatti

Lowest average MSRP is for Plymouth



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

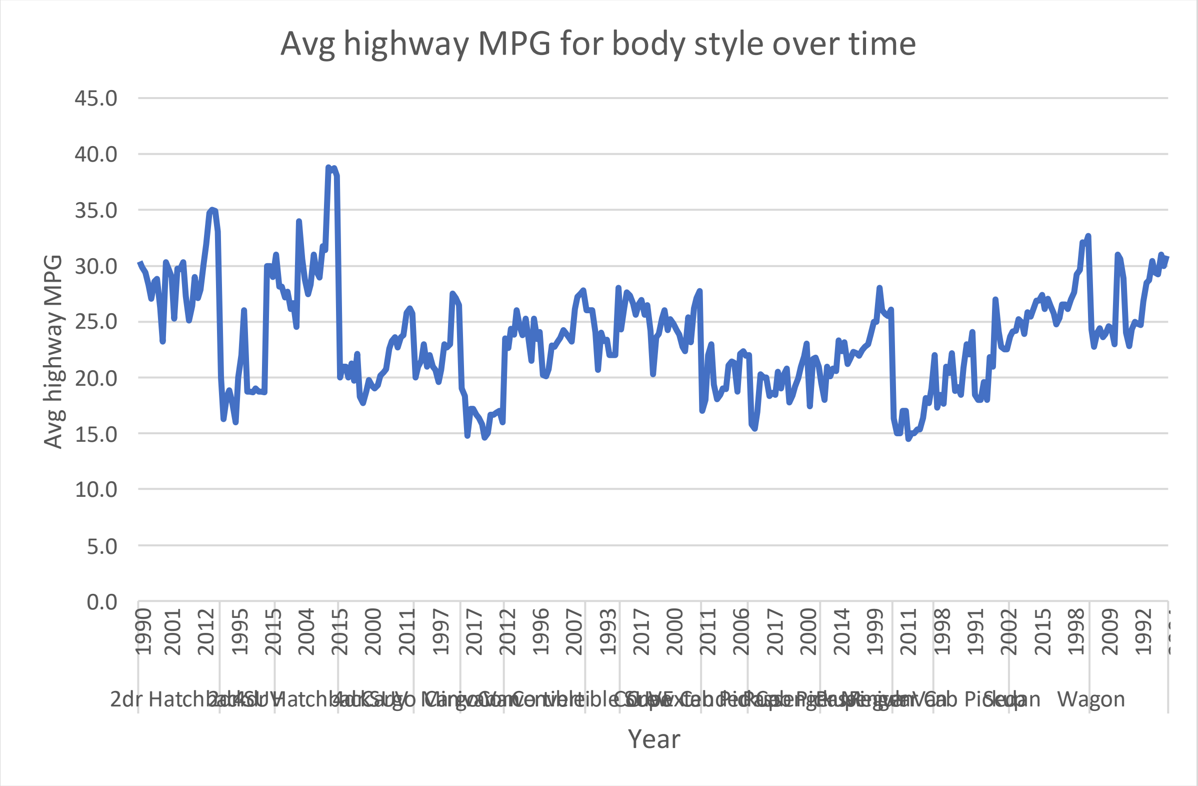




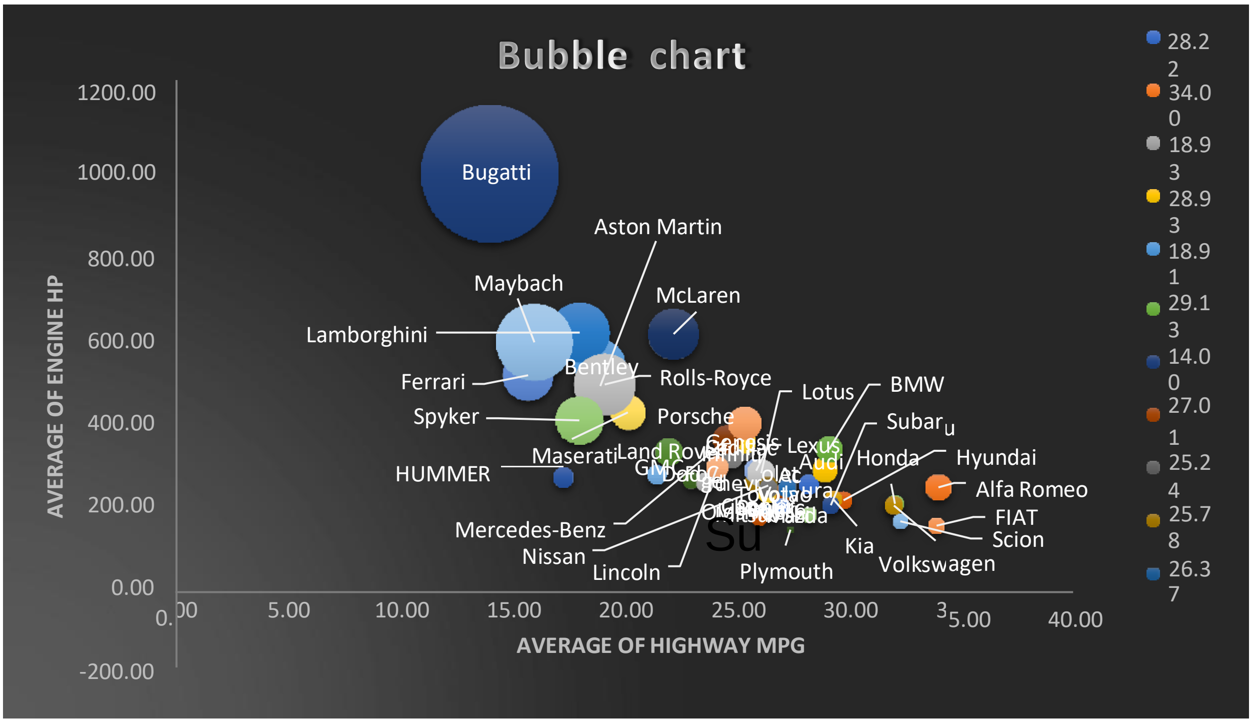
	AUTOMATIC	MANUAL	UNKNOWN	AUTOMATED_MANUAL	DIRECT_DRIVE
Sedan	44705.13	17557.26	2000.00	50385.39	27822.50
Wagon	28219.46	18398.58		31985.28	
Coupe	65031.19	51524.64	2000.00	245588.36	
Passenger Minivan	26570.02	6510.00			
2dr SUV	24153.61	9173.02	2371.00		
2dr Hatchback	20784.10	12840.66	7361.50	27470.42	
4dr Hatchback	23888.74	17500.36		29347.05	34511.92
Convertible	95153.31	64794.34	5783.50	129082.23	
Regular Cab Pickup	28536.82	8759.45	2000.00		
Extended Cab Pickup	30711.45	11553.30			
Cargo Minivan	20292.93				
4dr SUV	41658.40	17422.09		40451.15	
Convertible SUV	38925.50	9594.80			
Passenger Van	30578.07				
Cargo Van	17019.30				
Crew Cab Pickup	37718.95	28233.11			
total	35871.69	21066.28	3586.00	79187.13	31167.21

The above table shows the average of MSRP for for different transmission type and style

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



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



Brand	Average of highway MPG	Average of MSRP	Average of Engine HP
Acura	28.22	35087.49	244.96
Alfa Romeo	34.00	61600.00	237.00
Aston Martin	18.93	198123.46	483.76
Audi	28.93	54574.12	280.00
Bentley	18.91	247169.32	533.85
BMW	29.13	62162.56	329.62
Bugatti	14.00	1757223.67	1001.00
Buick	27.01	29034.19	220.01
Cadillac	25.24	56368.27	332.80
Chevrolet	25.78	29000.22	249.58
Chrysler	26.37	26722.96	229.14
Dodge	22.99	24857.05	254.35
Ferrari	15.72	237383.82	509.91
FIAT	33.92	22206.02	143.56

The above table shows the sample of averages of different parameters across the brand.Detailed information is provided in the excel sheet as it is lengthy.

Excel sheet link: C:\Users\Dharmateja\Desktop\trainity\Car_data(AutoRecovered).xlsx

PROJECT – 8
ABC Call Volume Trend

Description:

In this project, you'll be diving into the world of Customer Experience (CX) analytics, specifically focusing on the inbound calling team of a company. You'll be provided with a dataset that spans 23 days and includes various details such as the agent's name and ID, the queue time (how long a customer had to wait before connecting with an agent), the time of the call, the duration of the call, and the call status (whether it was abandoned, answered, or transferred).

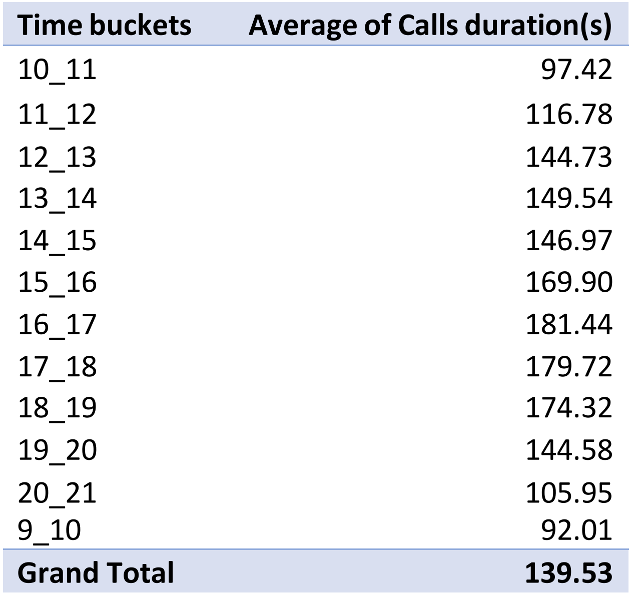
A Customer Experience (CX) team plays a crucial role in a company. They analyze customer feedback and data, derive insights from it, and share these insights with the rest of the organization. This team is responsible for a wide range of tasks, including managing customer experience programs, handling internal communications, mapping customer journeys, and managing customer data, among others.

Inbound customer support, which is the focus of this project, involves handling incoming calls from existing or prospective customers. The goal is to attract, engage, and delight customers, turning them into loyal advocates for the business.

Objective:

In this project, you'll be using your analytical skills to understand the trends in the call volume of the CX team and derive valuable insights from it.

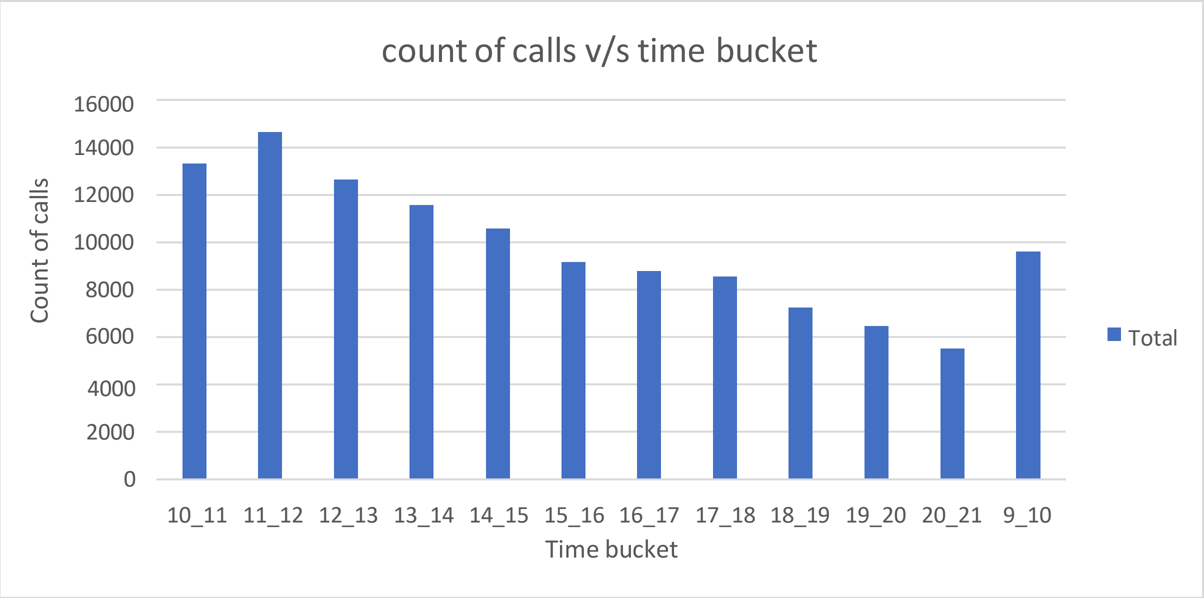
1) What is the average duration of calls for each time bucket?



Time buckets	Average of Calls duration(s)
10_11	97.42
11_12	116.78
12_13	144.73
13_14	149.54
14_15	146.97
15_16	169.90
16_17	181.44
17_18	179.72
18_19	174.32
19_20	144.58
20_21	105.95
9_10	92.01
Grand Total	**139.53**

2) Can you create a chart or graph that shows the number of calls received in each time bucket?

Time bucket	Count of calls
10_11	13313
11_12	14626
12_13	12652
13_14	11561
14_15	10561
15_16	9159
16_17	8788
17_18	8534
18_19	7238
19_20	6463
20_21	5505
9_10	9588
Grand Total	**117988**



3) What is the minimum number of agents required in each time bucket to reduce the abandon rate to 10%?

From the data given,

Total no of agents = 65

Total no of calls received =117988

No of abandoned calls = 35396

Calls answered or transferred = 82592

time bucket	Sum of Call_Seconds (s)
10_11	1297006
11_12	1708079
12_13	1831061
13_14	1728843
14_15	1552143
15_16	1556085
16_17	1594489
17_18	1533769
18_19	1261762
19_20	934437
20_21	583250
9_10	882195
Grand Total	**16463119**

From the above table total time spent by all agents for the duration of 23 days is16463119 secs.

Time buckets	Average of Calls duration(s)
10_11	97.42
11_12	116.78
12_13	144.73
13_14	149.54
14_15	146.97
15_16	169.90
16_17	181.44
17_18	179.72
18_19	174.32
19_20	144.58
20_21	105.95
9_10	92.01
Grand Total	**139.53**

Assuming the possible duration for a abandoned call can take , an average call duration.

So average duration of abandoned call from the table =139.5 secs

To reduce the abandoned calls percentage to 10 %

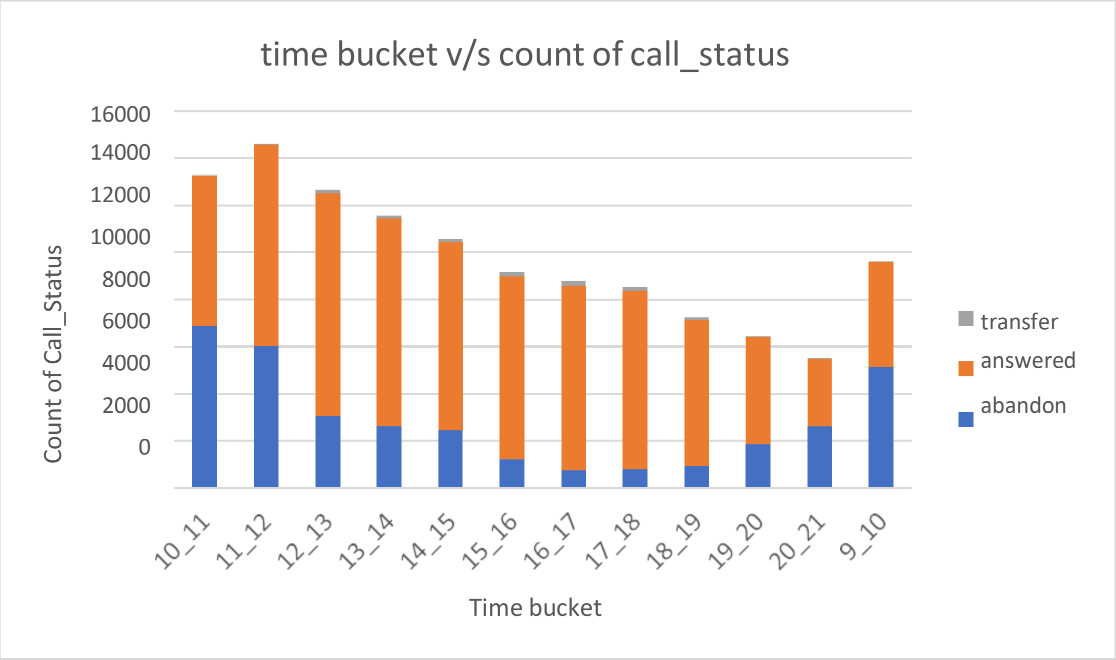
20% of 117988 is 23598

Total duration =16463119+ (23598*139.5) = 19755040 secs

So, for the duration of 19755040 secs , that is to reduce the abandoned % to 10 % the required no of agents are 78.

The distribution of agents along the time buckets is as follows

Count of Call_Status	**Column Labels**					
Time bucket	**abandon**	**answered**	**transfer**	**Grand Total**	**% of call received**	**agents distribution**
10_11	6911	6368	34	13313	11	9
11_12	6028	8560	38	14626	12	9
12_13	3073	9432	147	12652	11	9
13_14	2617	8829	115	11561	10	8
14_15	2475	7974	112	10561	9	7
15_16	1214	7760	185	9159	8	6
16_17	747	7852	189	8788	7	5
17_18	783	7601	150	8534	7	5
18_19	933	6200	105	7238	6	5
19_20	1848	4578	37	6463	5	4
20_21	2625	2870	10	5505	5	4
9_10	5149	4428	11	9588	8	6
Grand Total	**34403**	**82452**	**1133**	**117988**		



The plotted graph is based on the above table. The plot helped us to determine the required no of agents In each time bucket .

4) Propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%.

Given,

An agent will work around 4.5hrs per day.

Total no of calls =117988

30% 0f 117988 =35396

That implies for the duration of 23 days in night shift ,total no of calls that were received are 35,396

For night shift per day no of calls received =35,396/23

=1539.

Avg duration of a call =139.5 secs

As mentioned in question the abandoned rate is 10 %

Considering all this

Total time for the calls =(1539*139.5*0.9)/3600

= 54hrs

Therefore a total of 54 hrs agents need to spend.

Total no of agents required =54/4.5=12

time buckets	given_calls_distribution	% of calls distribution	agents_distribution	Agents_required(rounded)
21_22	3	10	1.2	1
22_23	3	10	1.2	1
23_24	2	7	0.84	1
24_1	2	7	0.84	1
1_2	1	3	0.36	0
2_3	1	3	0.36	0
3_4	1	4	0.48	0
4_5	1	3	0.36	0
5_6	3	10	1.2	1
6_7	4	13	1.56	2
7_8	4	13	1.56	2
8_9	5	17	2.04	2

Based on the requirement , the above table shows the distribution of agents.

As rounding the values turned to 0 which is not the possible case. That need to replaced by a single agent.so a total of 16 agents are required.

Excel link: ..\Desktop\trainity\Call_Volume_Trend_Analysis_Project_9.xlsx