Data Analysis

Portfolio

Prepared By :-

SANSKRUTI RANE

- Currently in my Third year pursuing B.Tech-IT. I have secured 8.91 CGPA (till 4TH sem) and have several skills including Python, sql, java, excel, c, c++, advance java.

- I have worked with several companies as an intern like , Internshala. I have also worked as a Project Manager with workskills where I have worked on their Data Analytics course from scratch and managed different teams.

- I have also published a research paper titled- "REVIEW OF E-VOTING SYSTEM" in a International journal of current science and have worked on micro projects like "Human resources and payroll management system".

- As I am a fresher it would be great to experience the real challenges of the corporate world and understand how things work. Being a fresher, I think I am very flexible and adaptive to learn new things. I have theoretical knowledge. But I am waiting to use my theoretical knowledge in a practical way. And I believe by putting significant efforts I will learn.

Table ofcontents:-

Project 1

Data Analytics Process

Application In Real Life Scenario Case Study

College admission process

- **Plan:-** We have to first decide, gather information and find the colleges and institutes which fits us as we always look for the environment where we can personally and academically thrive.

- **Prepare :-** Then we have to check for the amount(capital) available with us and how much fees we can pay. Based on that we can go for the colleges.

- **Process**:- After deciding the right college for us we have to go for the application process along with the application we also have to provide general information about us to the institute. Based on this decision of admission will then be made.

- **Analyze** :- We will check the facilities provided by the colleges we have chosen the facilities includes the Physical Infrastructure like College Building, Laboratories, Library, Sports Space, Equipment, Furnishings, Parking Space etc.

- **Share** :- Then we will finally visit the college and interact or communicate with the members at the inquiry office for rest of the doubts and queries clearance.

- **Act** :- And then we will finally take an admission by depositing a particular amount

Project 2

Instagram User Analytics

Instagram Analytics

❖ Project description :-

Here in this project I suppose that I am a data analyst working with the product team at Instagram. Where I walk-on part of analysing user interactions and engagement on Instagram app. My main task is to provide the insights such that it help out to multiply and grow the company. The information provided by me can help out or enables company to make data- driven decisions based on measurable and significant insights. In this project I have analyse the users behaviour by using different sql quries on provided dataset and based on my analysis the company will try to increase the focus on the parts with the lower engagement.

❖ Approach :-

My approach towards this project is to perform the analysis on the provided dataset by using different sql quries and concepts. I have perform this project in sql workbench. So by analysing the users engagement in different areas and using sql quries concepts like joints, keys, etc I performed the given task.

For eg: Why did the Marketing team wanted to know the most inactive users?

So, they can reach out to those users through mail and ask them why the are so away from instagram.

❖ Tech-stack used:-

- For this project I have used the MySQL Workbench 8.0 CE software.
- I used this application because MySQL Workbench contains many tools which allows database administrators to create physical database design models that suppose to be easily undergo into MySQL databases.

Instagram Analytics

- MySQL Workbench bonds all the objects such as tables, views, etc.

- Mysql workbench is a tool that can be used for the database development which includes modeling, maintaining, configuring, designing, creating. It also provides the mechanisms like object management, Connection Management, The Visual SQL Editor, Database Documentation, Server Administration, Data Modelling.

❖ Insights :-

- Engagement analysis:- this insight deals with the engagement of time of the users with Instagram as we done into our first task that is to find the most loyal i.e. the top 5 oldest users of Instagram and as well as in second task that is to find the users who have never posted a single photo on Instagram the most inactive users.

Rewarding Most Loyal Users: People who have been using the platform for the longest time. Remind Inactive Users to Start Posting: By sending them emails to post their photo.

- Post like analysis:- It deals with the insight like likes, shares, saves, etc as we find the user with the most likes on a single photo in our third task.

Declaring Contest Winner: the user who gets the most likes on a single photo will win the contest started by the team.

- Hashtag analysis:- this insight refers to the engagement of the hashtags in the particular post as we have suggested the top five most commonly used hashtags on the platform in our fourth task.

Instagram Analytics

- Account overview:- through this insight we get to know about the visits to the Instagram, impressions etc as we have done in our fifth task that is to Determine the day of the week when most users register on Instagram. So from our analysis we come to know that the team should launch their ads on the Thursday.

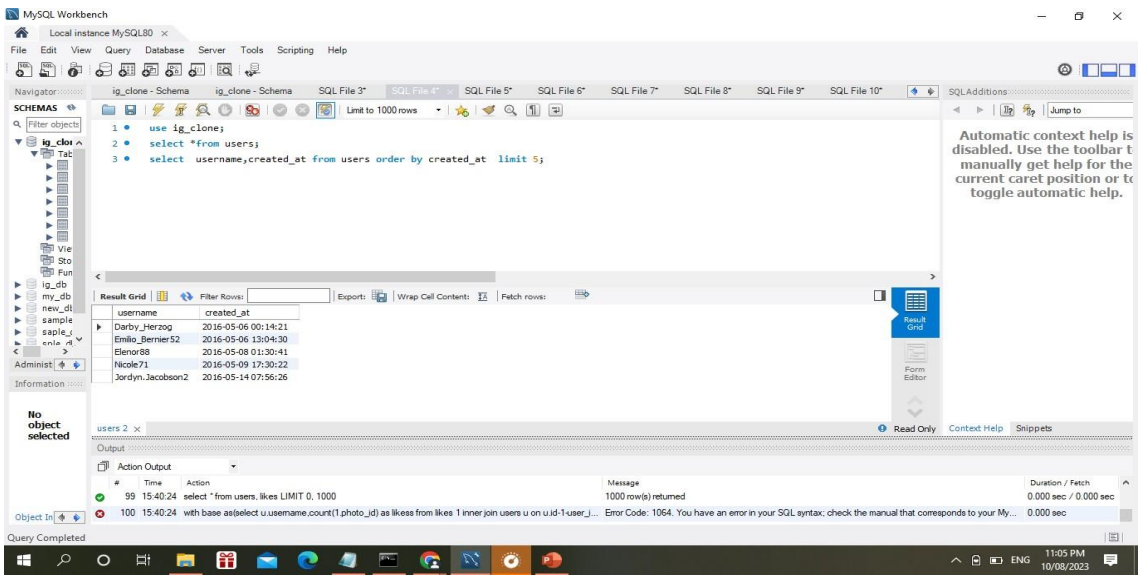
- Posting time analysis:- through this insight we come to know the active and inactive users the user engagement in case of posting on Instagram as we have Calculate the average number of posts per user on Instagram.

❖ Results:-

A) Marketing Analysis:-

1. **Loyal User Reward:-** we find the top 5 oldest users of Instagram. We will use the data from the users table. Then using the order by function we will order the desired output and after using the limit function, the output will be displayed for top 5 oldest Instagram users.

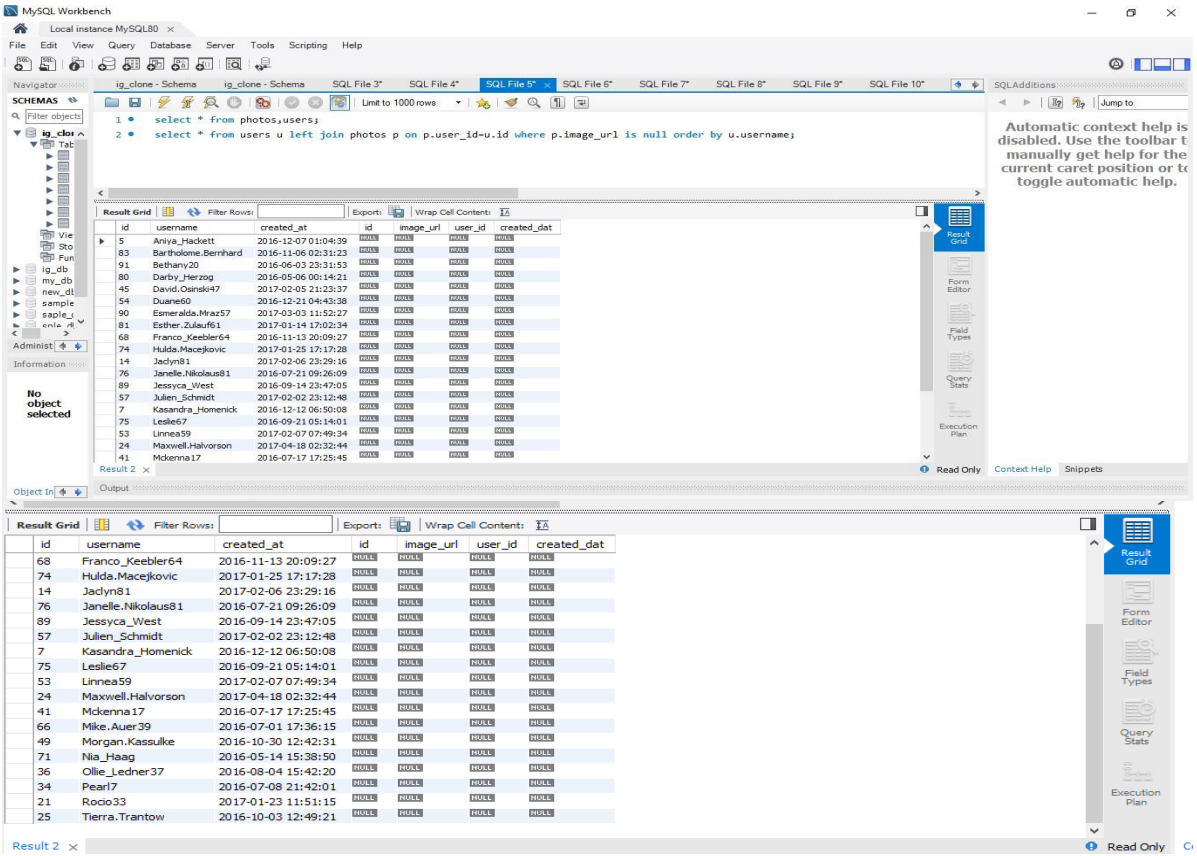
Output:-



Analytics Instagram

2. **Inactive User Engagement:-** we have Identified users who have never posted a single photo on Instagram. We will first select username column from the users table. Then we will left join photos table on the users table because both have common contents in them. Then we will find rows from the users table where the photos.id is null.

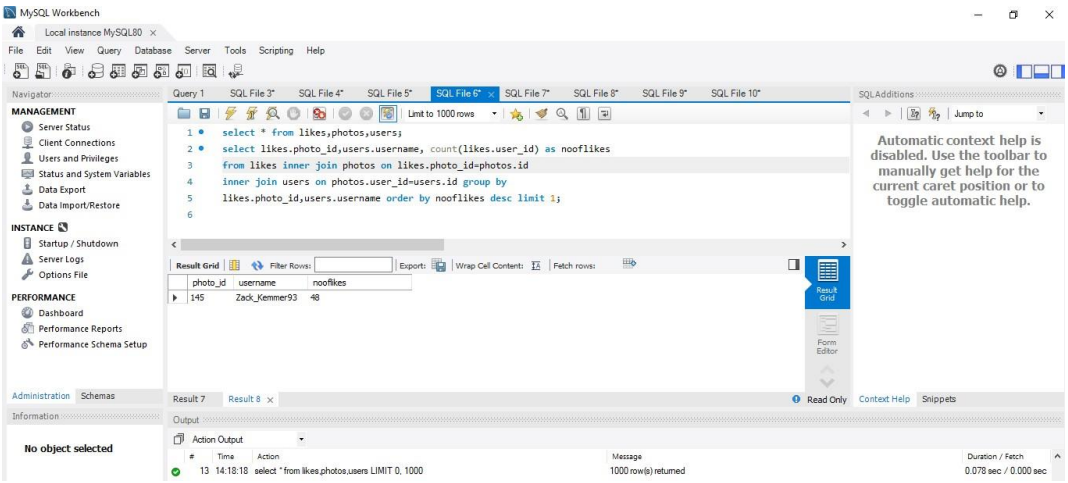
Output:-



3. **<u>Contest Winner Declaration</u>:** - In this task we have identified the user with the most likes on a single photo and provide their details to the team.First we will select the users.username, photos.id, photos.image_url and count(*) as total then, we will perform inner join on the three tables after that , by using group by function we will group the output. Then, using order by function we will sort the data on the

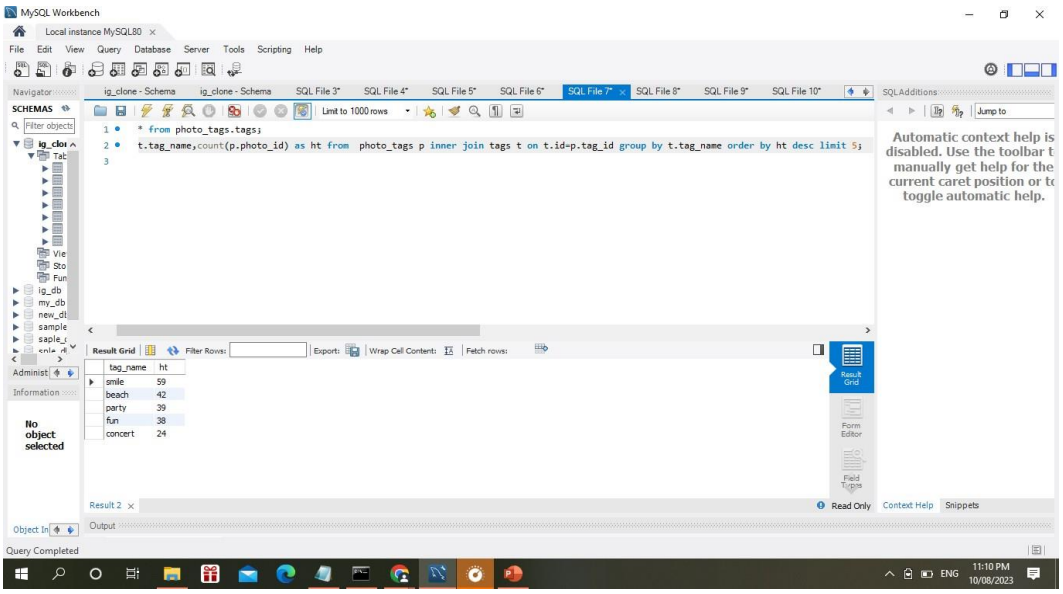
basis of the total (descending order). Then, we will use the <u>limit function.</u>

Output:-



4. **<u>Hashtag Research:-</u>** here we have Identified and suggest the top five most commonly used hashtags on the platform. We need to select the tag_name column from the tag table to count the number of tags used individually then, we need to join tags table and photo_tags table. After this by using the group by function we need to group the desired output. Then using the order by function we need to sort the output in descending order and at last we have to use limit 5 function.

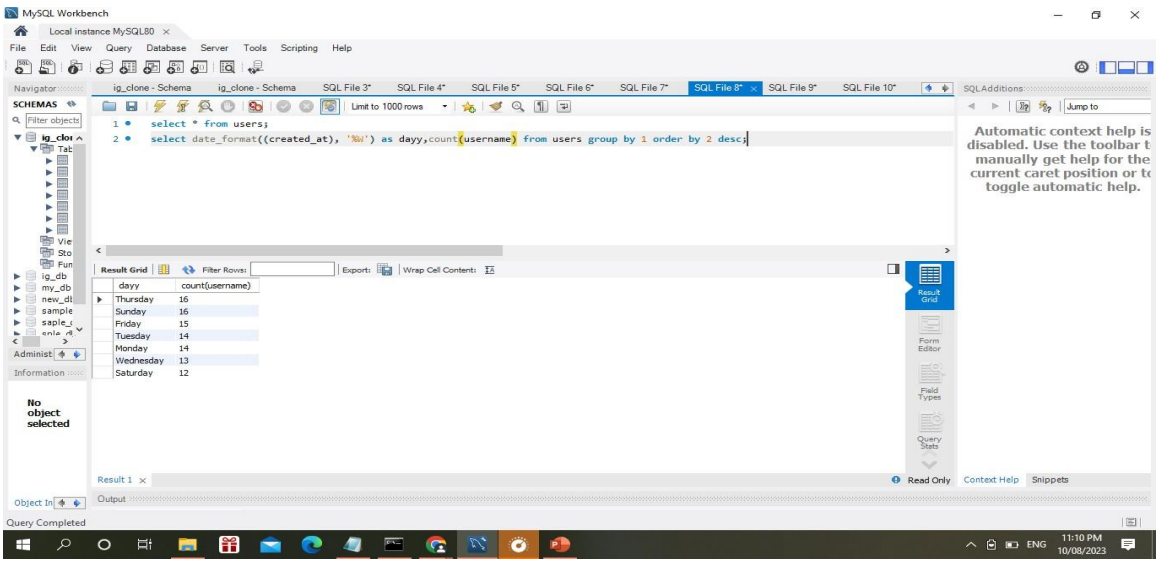
Output:-



Instagram Analytics

5. **Ad Campaign Launch**:- here we have found the best day of the week to launch ads. Now firstly we define the columns using select dayname(created_at) as day_of_week and count(*) as total_number_of_users_registered from the users table after that using the group by function we group the output table. Then using the order by function we sort the output table on the basis of total_number_of_users_registered in descending order

Output:-

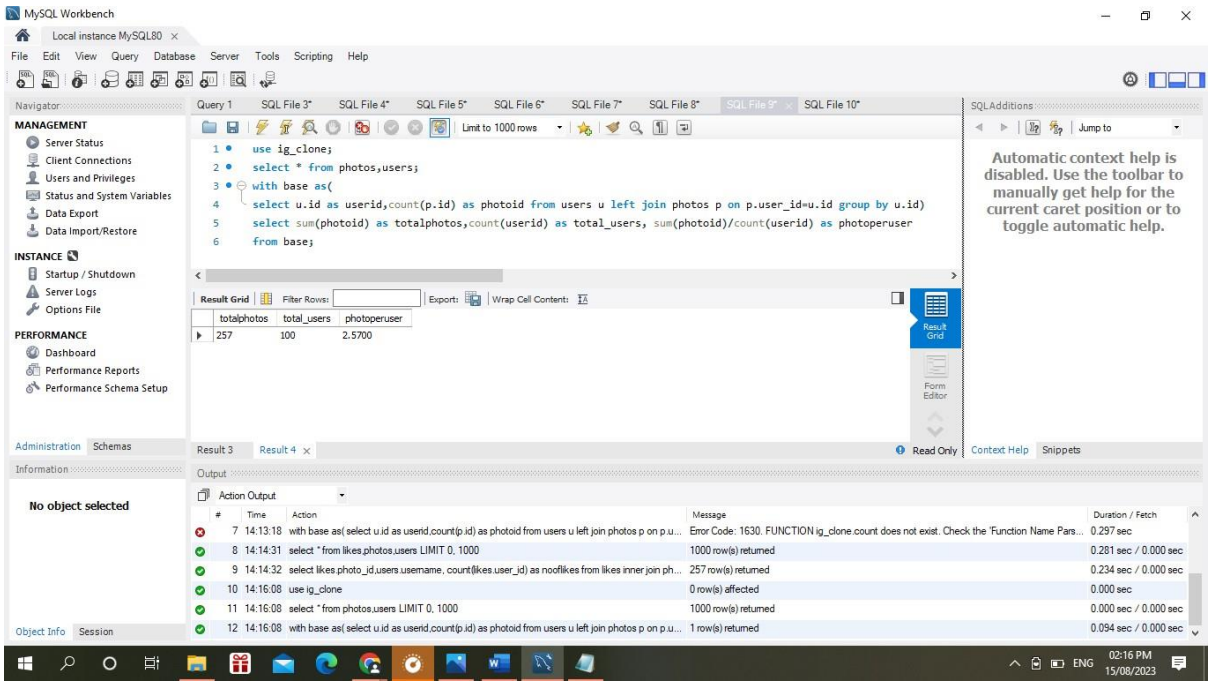


Instagram Analytics

B) Investor Metrics:-

1. **User Engagement:-** here we have Calculated the average number of posts per user on Instagram. First, we have to find the count number of posts that are present in the photos.id column. Similarly, we need to find the number of users that are present in the users.id column. After that we have to divide both the values and at last we need to find the total occurrences of each user_id in photos table.

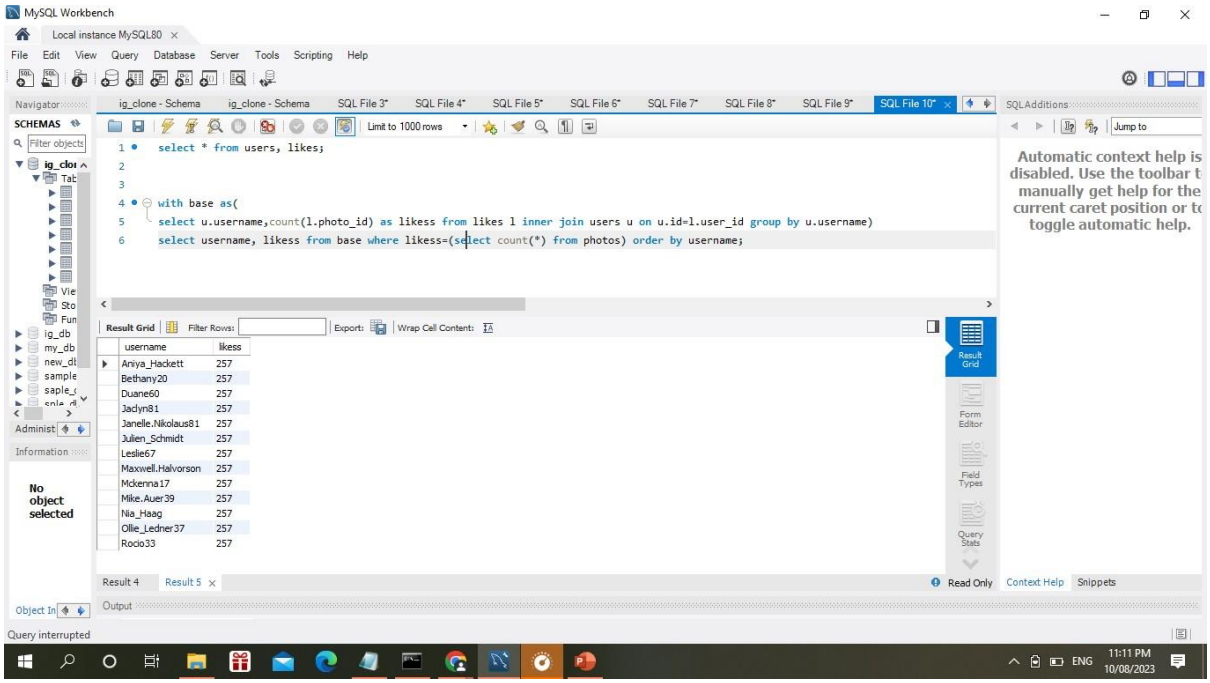
Output:-



Instagram Analytics

2. **Bots & Fake Accounts:-** here we have Identified the users (potential bots) who have liked every single photo on the site, as this is not typically possible for a normal user. First, we select the user_id column from the photos table and then we select the username column from the users table. After that we will use the count(*) function to count total number of likes from the likes table and we will perform inner join users and likes table Then by using the group by function we group the output table on the basis of likes. user_id . Then, we search for the values from the count(*) from photos having equal values with the total_likes_per_user.

Output:-



Instagram Analytics

❖ **Drive link:-**

- ☐ we find the top 5 oldest users of Instagram.
- ☐ we have Identified users who have never posted a single photo on Instagram.
- ☐ we have identified the user with the most likes on a single photo and provide their details to the team.
- ☐ we have Identified and suggest the top five most commonly used hashtags on the platform.
- ☐ we have found the best day of the week to launch ads.
- ☐ we have Calculated the average number of posts per user on Instagram.
- ☐ we have Identified the users (potential bots) who have liked every single photo on the site, as this is not typically possible for a normal user.

Project 3

Operation analytics and investigating metric spike

OPERATION & METRIC ANALYTICS

Project Description

- **Purpose:**

Operational analytics is crucial process that involves analyzing a company's end-to-end operations. It helps in increasing profits. Operational analytics or Operational data is all about individual entities at specific points in time. Using this analysis we can identify areas for improvement within the company. Here in this project I suppose that I am a lead data analyst working at the company like Microsoft . here by using advanced SQL skills I have to analyze the data and provide valuable insights that can help improve the company's operations and understand sudden changes in key metrics. One of the key aspects of Operational Analytics is investigating metric spikes which is all about understanding and explaining sudden changes in key metrics, such as a dip in daily user engagement or a drop in sales. Here as my role is of lead Data Analyst, I have to answer all these questions daily, making it crucial to understand how to investigate these metric spikes. And according to my role I will be working closely with various teams, such as operations, support, and marketing, and help them derive valuable insights from the collected data.

- **Approach:**

My approach towards this project is to perform the analysis on the provided dataset by using different sql quries and concepts. I have perform this project in sql workbench. So by analysing the job_data databae and by applying different sql quries concepts like joints, keys, etc I performed the given task. Here I have to work with two different case studies that are Investigating Metric Spike where I will be working with three tables users,events,email_events and the another is Job Data Analysis where I will be will be working with a table named job_data .

❖ Tech-stack used:-

- For this project I have used the MySQL Workbench 8.0 CE software.
- I used this application because MySQL Workbench contains many tools which allows database administrators to create physical database design models that suppose to be easily undergo into MySQL databases.
- MySQL Workbench bonds all the objects such as tables, views, etc.
- Mysql workbench is a tool that can be used for the database development which includes modeling, maintaining, configuring, designing, creating. It also provides the mechanisms like object management, Connection Management, The Visual SQL Editor, Database Documentation, Server Administration, Data Modelling.

❖ Insights :

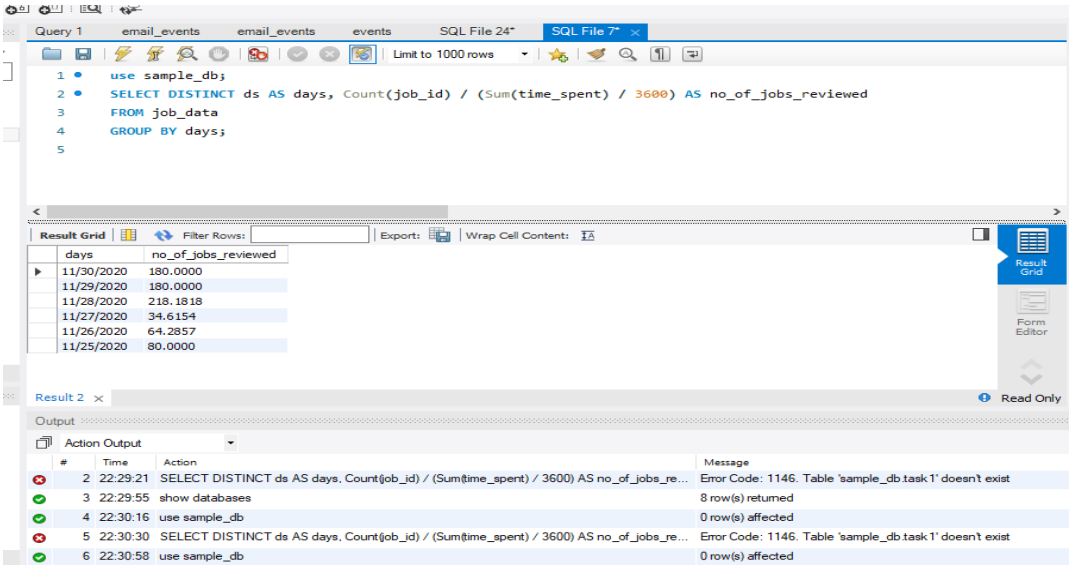
- **Engagement analysis:-** this insight deals with the engagement of time of the user as we perform in case study 2 that are **Weekly User Engagement, Email Engagement Analysis,etc.**
- **Identify the Metric Spikes:** here at this stage we have to identify the specific metrics that have significant spikes. These metrics could include website sales,traffic, customer inquiries, production output.
- **Throughput Analysis:** it determines the timeframe during which the spike occurred which can be a day, week, month, or even an hour.
- **Gather Contextual Information:** Collect additional information that might be relevant to the spike. This could include events, campaigns, promotions, external factors (like news or market trends), or internal changes (such as system updates or process modifications).
- **Lifecycle Analysis:** Analysing the entire lifecycle of a product or service can reveal areas for improvement in sustainability and cost efficiency.

❖ SQL Tasks :

Case Study 1: Job Data Analysis

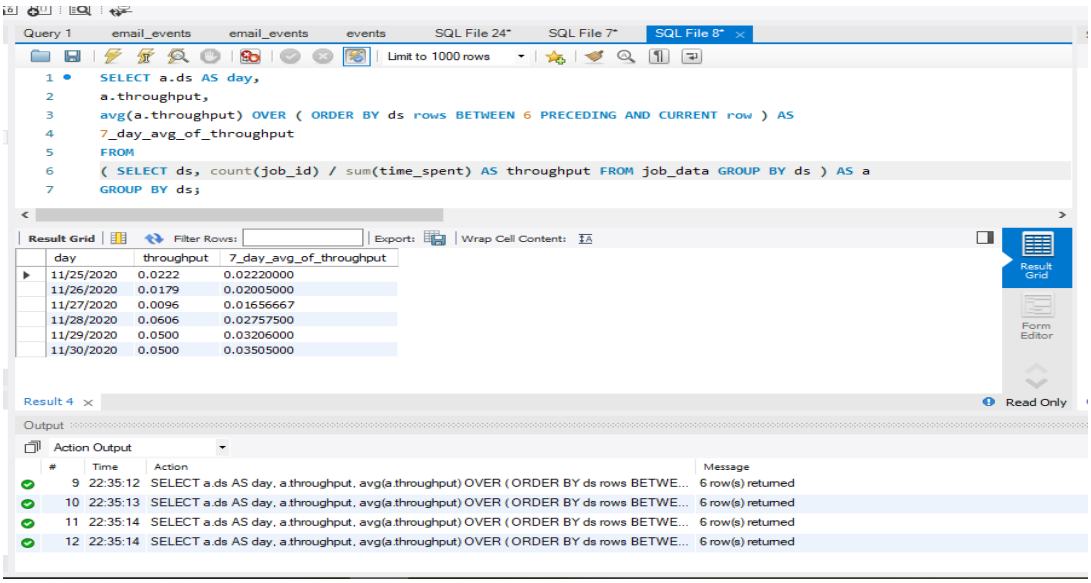
A. Jobs Reviewed Over Time

- Objective: Calculate the number of jobs reviewed per hour for each day in November 2020.
- Your Task: Write an SQL query to calculate the number of jobs reviewed per hour for each day in November 2020.



B. Throughput Analysis:

- Objective: Calculate the 7-day rolling average of throughput (number of events per second).
- Your 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.



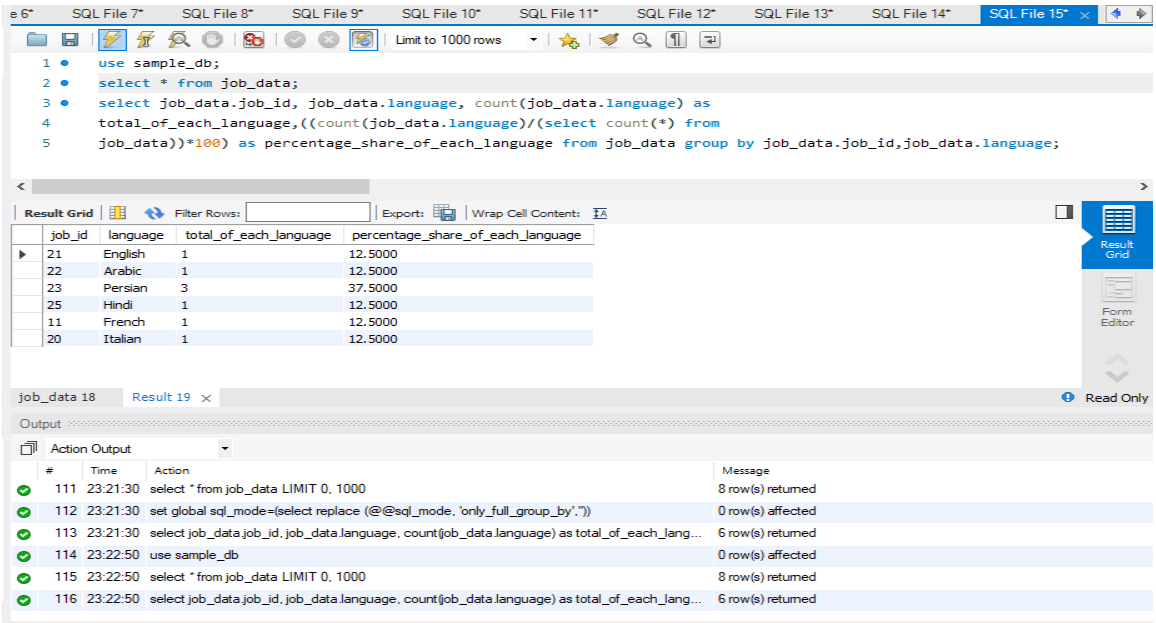
```
SELECT a.ds AS day,
a.throughput,
avg(a.throughput) OVER ( ORDER BY ds rows BETWEEN 6 PRECEDING AND CURRENT row ) AS
7_day_avg_of_throughput
FROM
( SELECT ds, count(job_id) / sum(time_spent) AS throughput FROM job_data GROUP BY ds ) AS a
GROUP BY ds;
```

day	throughput	7_day_avg_of_throughput
11/25/2020	0.0222	0.02220000
11/26/2020	0.0179	0.02005000
11/27/2020	0.0096	0.01656667
11/28/2020	0.0606	0.02757500
11/29/2020	0.0500	0.03206000
11/30/2020	0.0500	0.03505000

#	Time	Action	Message
9	22:35:12	SELECT a.ds AS day, a.throughput, avg(a.throughput) OVER (ORDER BY ds rows BETWE...	6 row(s) returned
10	22:35:13	SELECT a.ds AS day, a.throughput, avg(a.throughput) OVER (ORDER BY ds rows BETWE...	6 row(s) returned
11	22:35:14	SELECT a.ds AS day, a.throughput, avg(a.throughput) OVER (ORDER BY ds rows BETWE...	6 row(s) returned
12	22:35:14	SELECT a.ds AS day, a.throughput, avg(a.throughput) OVER (ORDER BY ds rows BETWE...	6 row(s) returned

C. Language Share Analysis:

- Objective: Calculate the percentage share of each language in the last 30 days.
- Your Task: Write an SQL query to calculate the percentage share of each language over the last 30 days.



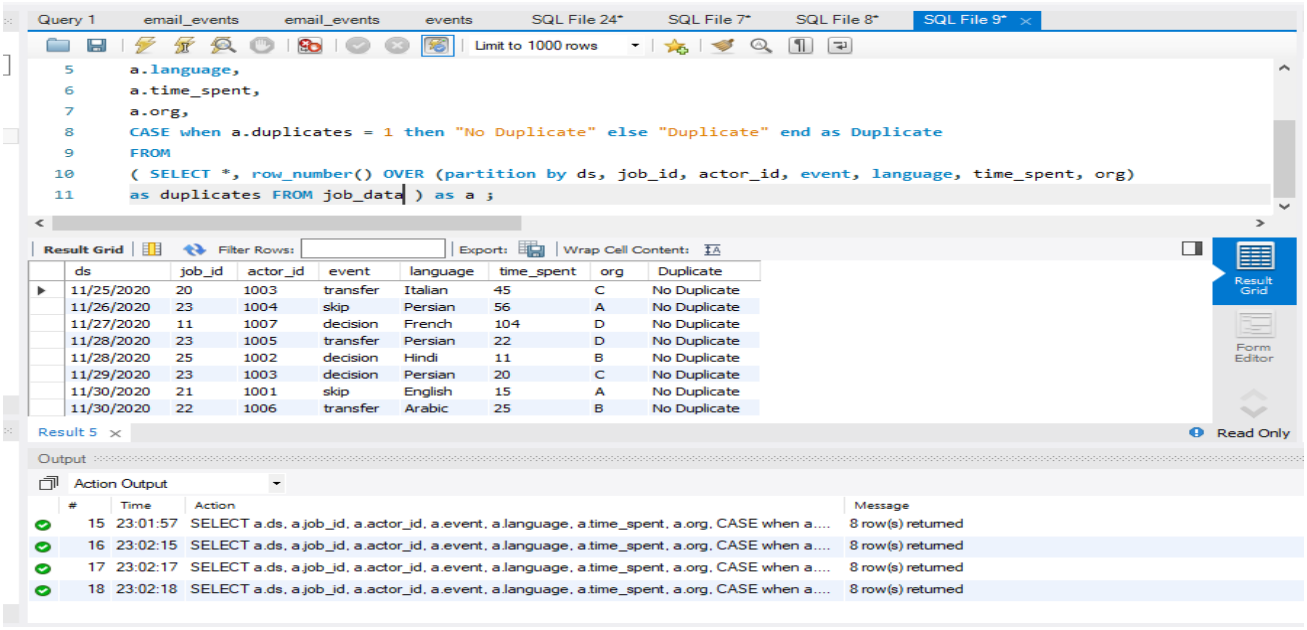
```
use sample_db;
select * from job_data;
select job_data.job_id, job_data.language, count(job_data.language) as
total_of_each_language,((count(job_data.language)/(select count(*) from
job_data))*100) as percentage_share_of_each_language from job_data group by job_data.job_id,job_data.language;
```

job_id	language	total_of_each_language	percentage_share_of_each_language
21	English	1	12.5000
22	Arabic	1	12.5000
23	Persian	3	37.5000
25	Hindi	1	12.5000
11	French	1	12.5000
20	Italian	1	12.5000

#	Time	Action	Message
111	23:21:30	select * from job_data LIMIT 0, 1000	8 row(s) returned
112	23:21:30	set global sql_mode=(select replace (@@sql_mode, 'only_full_group_by',''))	0 row(s) affected
113	23:21:30	select job_data.job_id, job_data.language, count(job_data.language) as total_of_each_lang...	6 row(s) returned
114	23:22:50	use sample_db	0 row(s) affected
115	23:22:50	select * from job_data LIMIT 0, 1000	8 row(s) returned
116	23:22:50	select job_data.job_id, job_data.language, count(job_data.language) as total_of_each_lang...	6 row(s) returned

D. Duplicate Rows Detection:

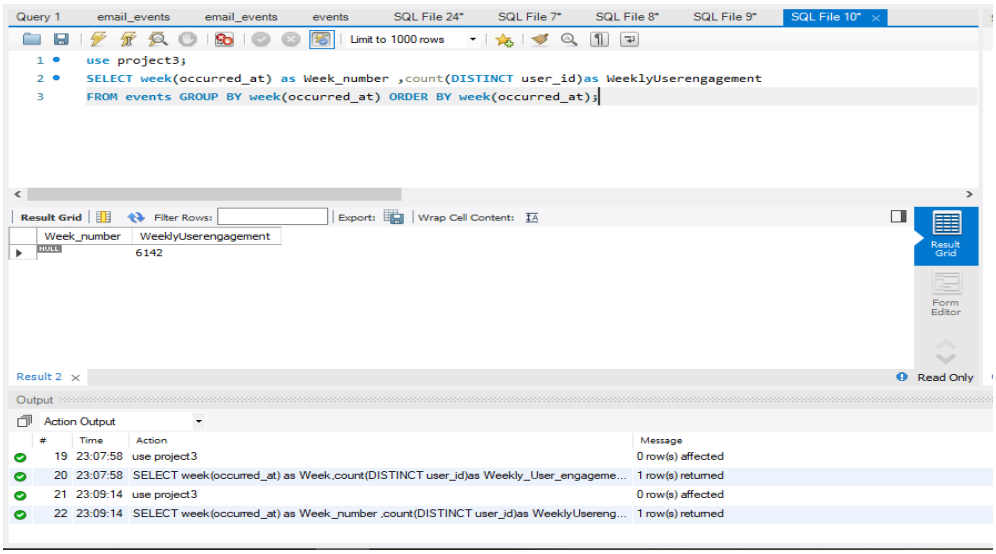
- Objective: Identify duplicate rows in the data.
- Your Task: Write an SQL query to display duplicate rows from the job_data table.



Case study 2:- Investigating Metric Spike

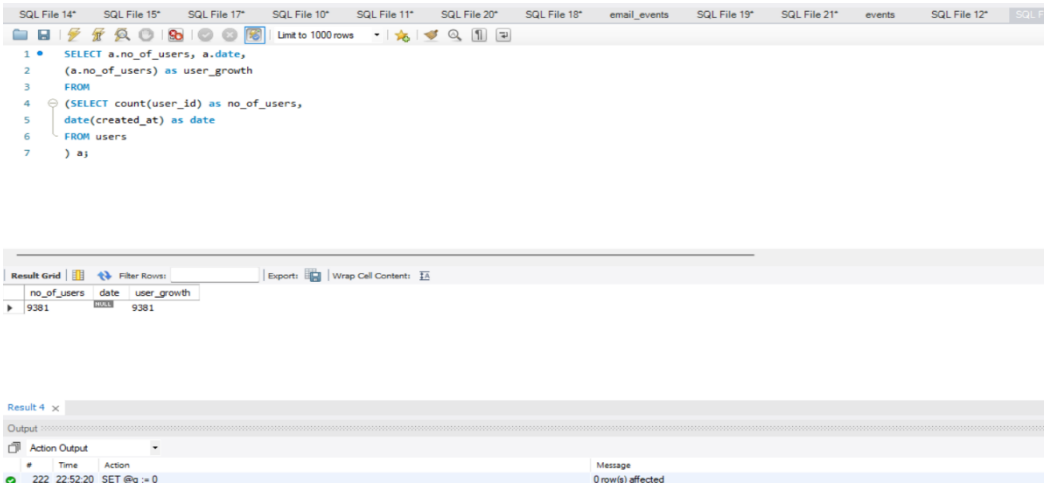
A. Weekly User Engagement:

- Objective: Measure the activeness of users on a weekly basis.
- Your Task: Write an SQL query to calculate the weekly user engagement.



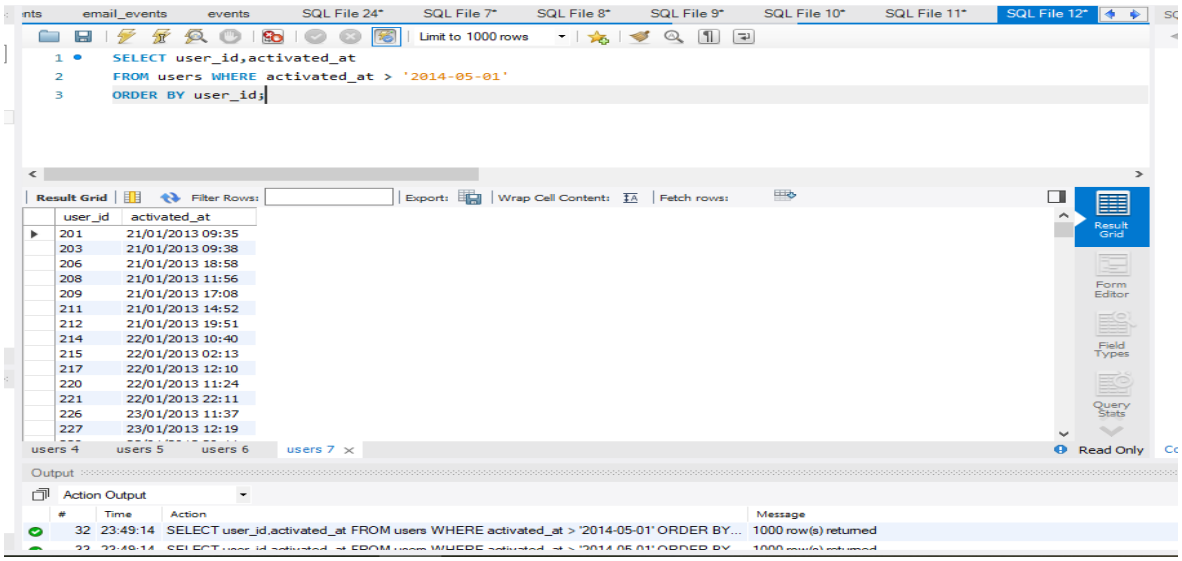
B. *User Growth Analysis:*

- *Objective: Analyze the growth of users over time for a product.*
- *Your Task: Write an SQL query to calculate the user growth for the product.*



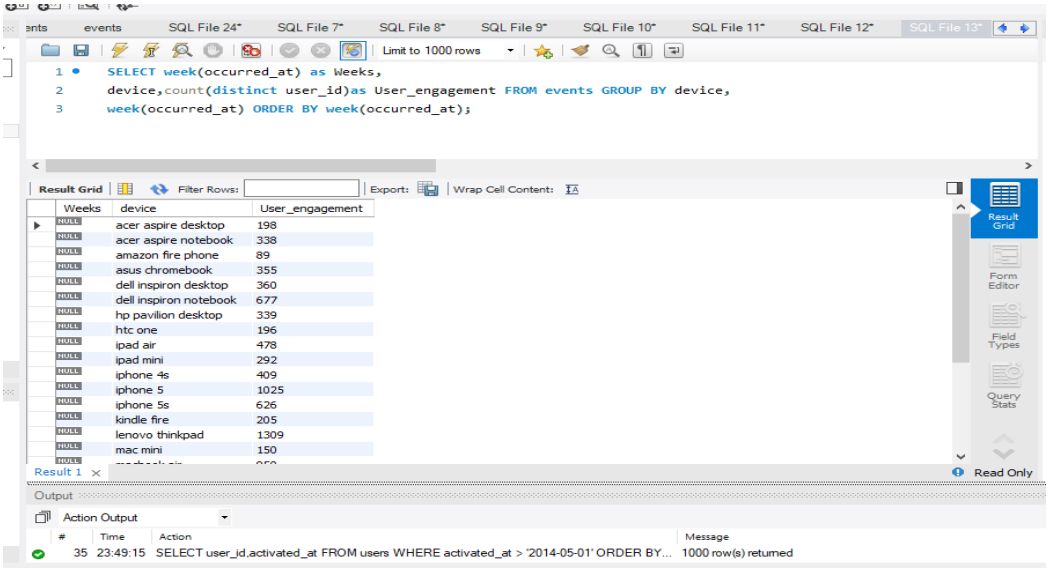
C. Weekly Retention Analysis:

- Objective: Analyze the retention of users on a weekly basis after signing up for a product.
- Your Task: Write an SQL query to calculate the weekly retention of users based on their sign-up cohort.



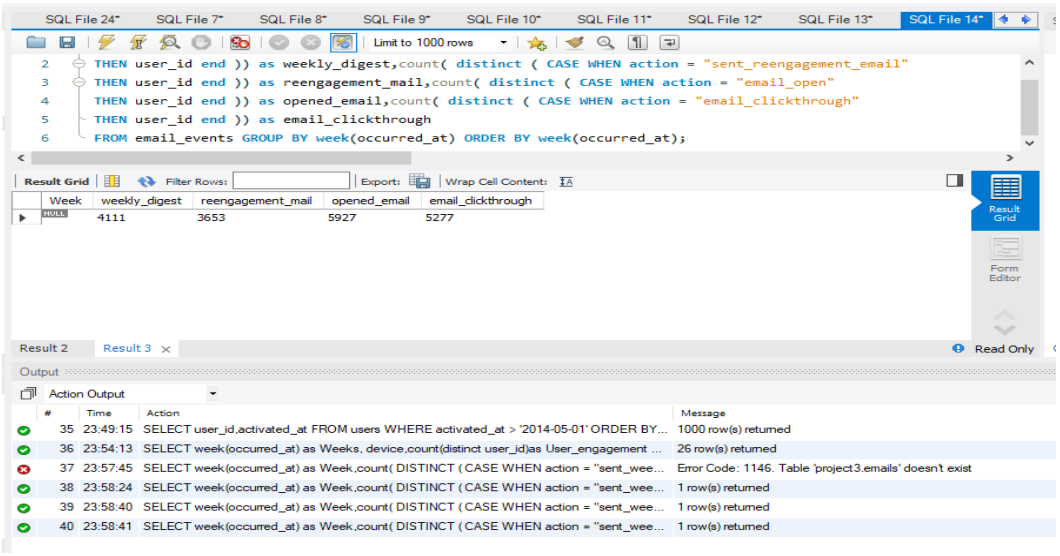
D. Weekly Engagement Per Device:

- Objective: Measure the activeness of users on a weekly basis per device.
- Your Task: Write an SQL query to calculate the weekly engagement per device



E. Email Engagement Analysis:

- Objective: Analyze how users are engaging with the email service.
- Your Task: Write an SQL query to calculate the email engagement metrics.



❖ **Drive link:-**

Case study 1:-

- We have calculated the number of jobs reviewed per hour for each day in November 2020.

- We have calculated 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.
- We have calculated the percentage share of each language over the last 30 days.
- We have displayed duplicate rows from the job_data table.

Case study 2:-

- We have calculated the weekly user engagement
- We have calculated the user growth for the product

- We have calculated the weekly retention of users based on their sign-up cohort.
- We have calculated the weekly engagement per device.
- We have calculated the email engagement metrics.

Project 4

Hiring Process Analytics

❖ PROJECT DESCRIPTION:-

- Hiring process analytics consist of the collection, analysis, interpretation of data related to the recruitment and hiring of employees within an organization. Basically the process of intaking people into an organization refers to the process of hiring.
- It aims to carry out the deep analysis ,and to optimize and improve the hiring process by providing insights and actionable information.

❖ APPROACH

My approach towards this project "Hiring Process Analytics" is perform the analysis on the provided dataset and perform different task using different SQL queries and concepts. I will be given a dataset containing records of previous hires.where I need to analyze this data and answer certain questions that can help the company improve its hiring process. Analyzing the hiring process through a data-driven analytics approach is essential for organizations to make informed decisions and continually improve their recruitment efforts. Centralize your data in a dedicated data repository or data warehouse. Integrating data from multiple sources can provide a holistic view of your hiring process.

❖ TECH - STACK USED:

1. For this project I have used the Microsoft Excel 2019.

2. I used this application because Microsoft Excel 2019contains many tools which format, organize and calculate data in a spreadsheet.

3. Microsoft Excel 2019 represent the data in from of different charts like bar graph, pie chart, histogram graph., etc.

4. Microsoft Excel is a widely used spreadsheet application developed by Microsoft. It is part of the Microsoft Office suite of productivity software. Excel is known for its powerful data manipulation and analysis features and is used for a wide range of tasks, including data entry, calculation, analysis, and visualization.

❖ INSIGHTS:-

- Data Analysis: this refers to data analysis techniques such as statistical analysis, data visualization, and machine learning to gain insights from hiring data.
- Communication and Reporting: this refers to the sharing of findings and progress with key stakeholders in the organization, hiring managers, and executives. Use data-driven insights to guide discussions .

- Data Collection: Collect data at each stage of the hiring process. This data can come from various sources. Ensure that data is integrated seamlessly into your analytics platform.
- Predictive Analytics: Implement predictive analytics models to forecast future hiring needs and predict candidate success in specific roles. Machine learning algorithms can assist in predicting which candidates are more likely to be a good fit for your organization.

❖ DATA ANALYTICS TASKS :-

A. Hiring Analysis:

The hiring process involves bringing new individuals into the organization for various roles.

Your Task: Determine the gender distribution of hires. How many males and females have been hired by the company.
Task 1:-

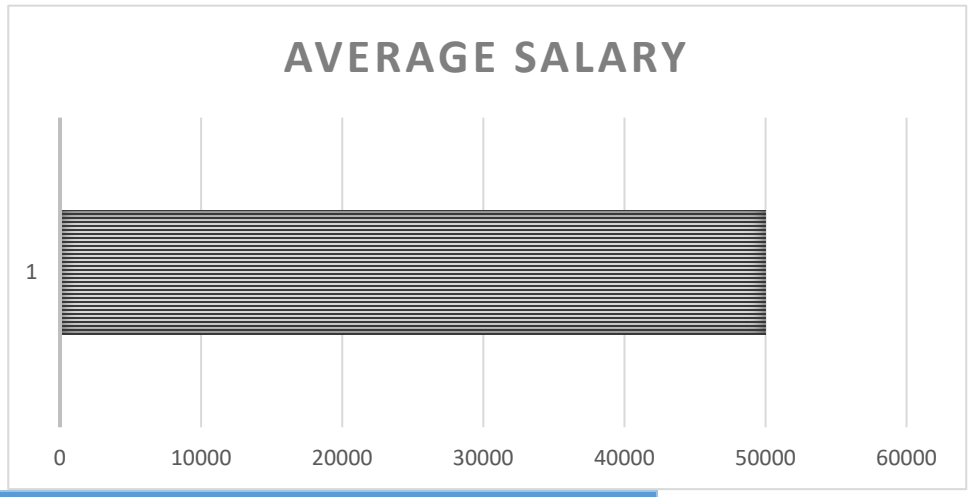
Gender	Status	Hiring Count
Female	Hired	1856
Male	Hired	2563

B. Salary Analysis:

The average salary is calculated by adding up the salaries of a group of employees and then dividing the total by the number of employees.

Your Task: What is the average salary offered by this company? Use Excel functions to calculate this.

Task 2:-



Average Salary	49983.029

C. Salary Distribution:

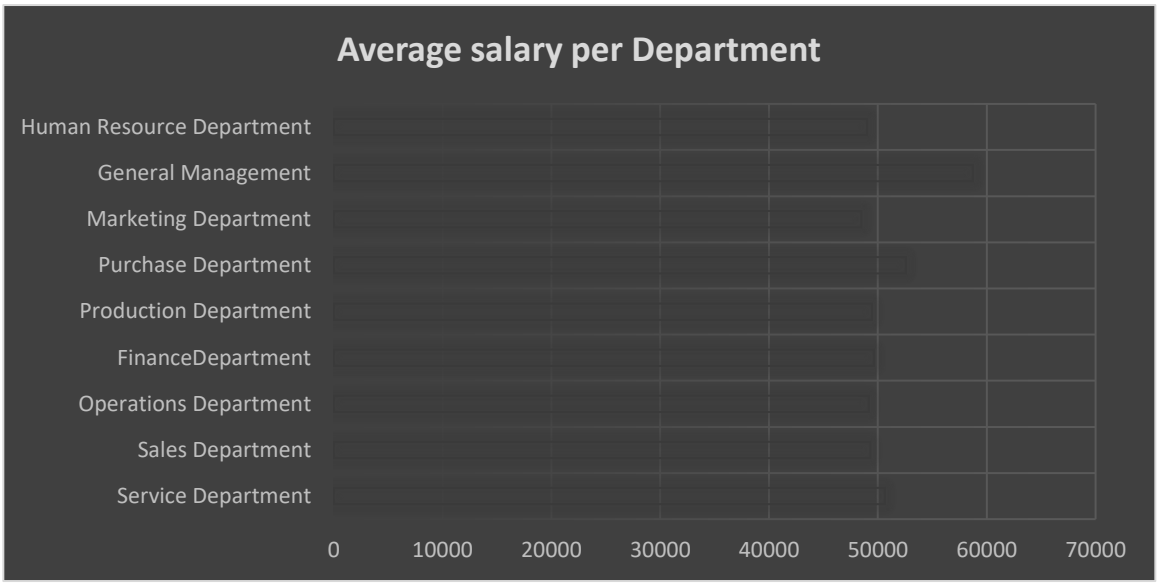
Class intervals represent ranges of values, in this case, salary ranges. The class interval is the difference between the upper and lower limits of a class.

Your Task: Create class intervals for the salaries in the company. This will help you understand the salary distribution.

Task 3 :-

Department	Average salary per Department
Service Department	50629.88418
Sales Department	49310.3807
Operations Department	49151.35438
Finance Department	49628.00694
Production Department	49448.48421
Purchase Department	52564.77477
Marketing Department	48489.93538
General Management	58722.09302

Human Resource Department	49002.27835



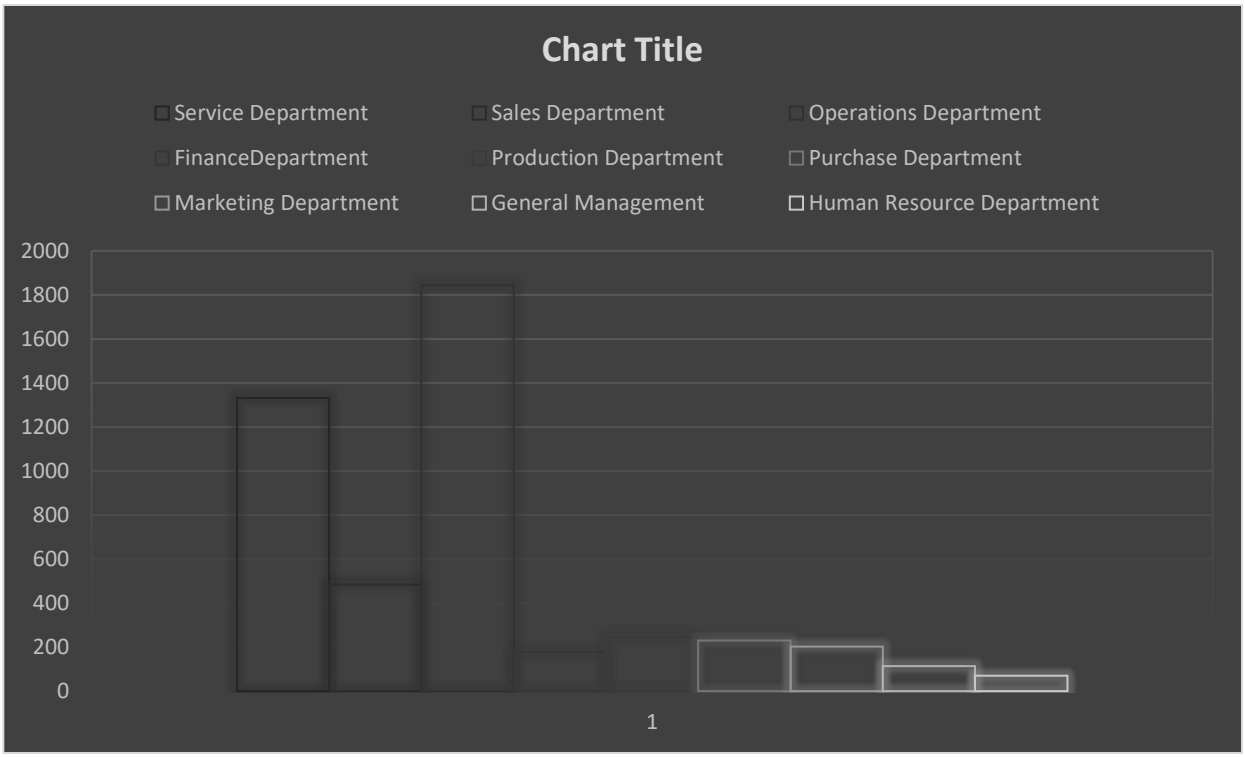
D. Departmental Analysis:

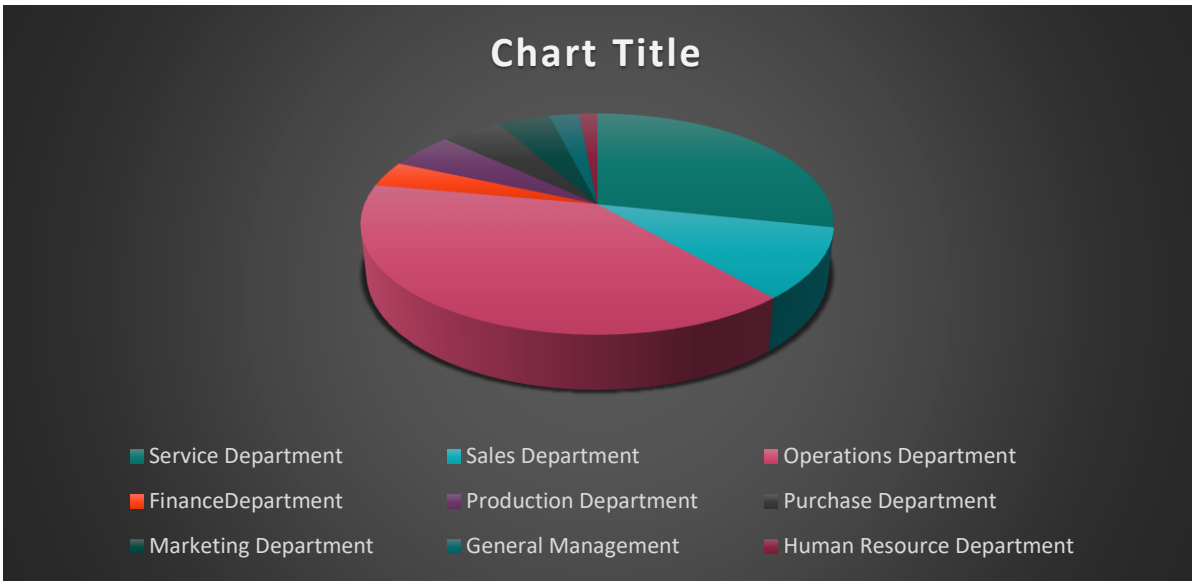
Visualizing data through charts and plots is a crucial part of data analysis.

Your Task: Use a pie chart, bar graph, or any other suitable visualization to show the proportion of people working in different departments.

Task 4 :-

Department	Count
Sales Department	485
Operations Department	1843
Finance Department	176
Production Department	246
Purchase Department	230
Marketing Department	202
General Management	113
Human Resource Department	70





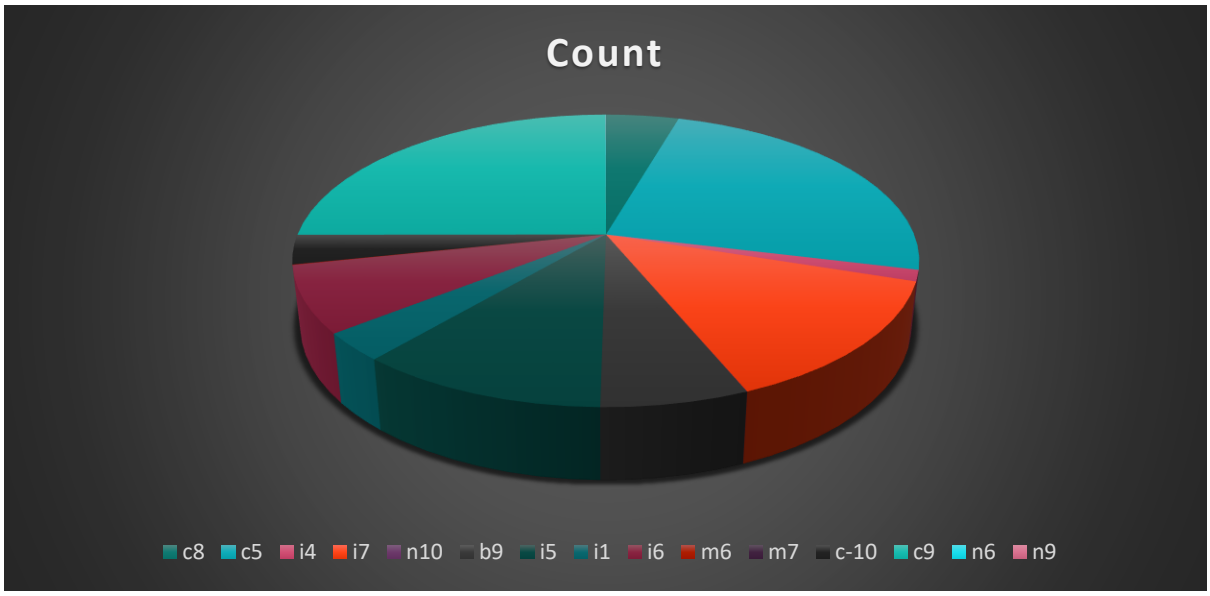
E. Position Tier Analysis:

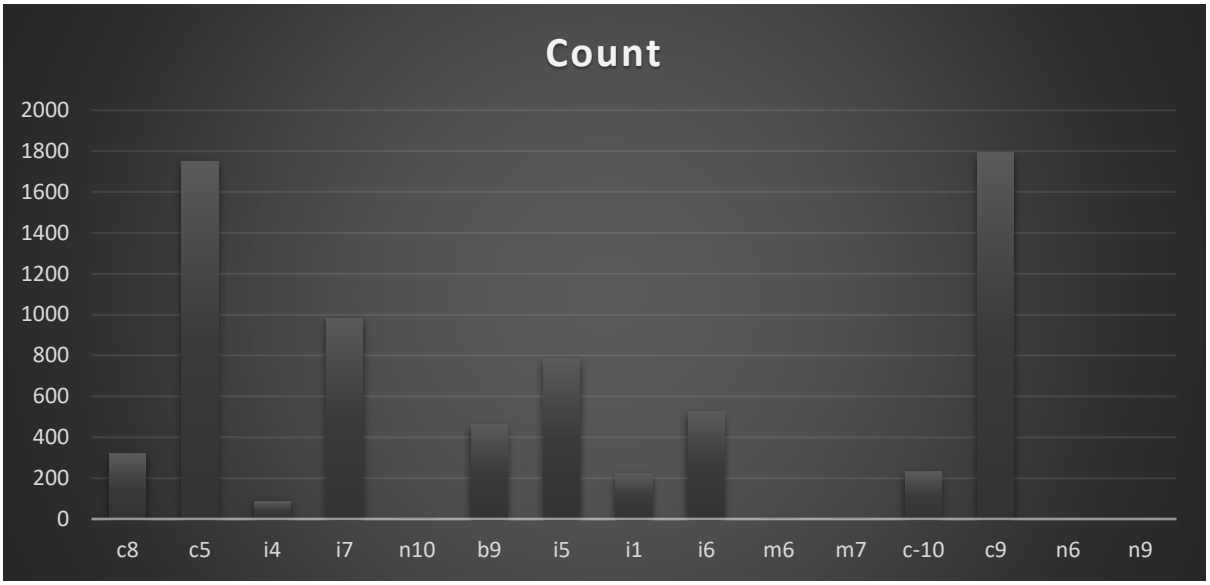
Different positions within a company often have different tiers or levels.

Your Task: 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.

Task 5 :

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




Count
2000
1800
1600
1400
1200
1000
800
600
400
200
0
c8
c5
i4
i7
n10
b9
i5
i1
i6
m6
m7
c-10
c9
n6
n9

❖ **Drive link: Data Analytics Tasks :**

• We have calculated the number of males and females have been hired by the company.
• We have calculated the average salary offered by this company.
• We have calculated the salary in different department.
• We have calculated the number of people working in different departments.
• We have calculated the number of distribution of positions across different tiers

Project 5

IMDB Movie Analysis

INTRODUCTION

❖ PROJECT DESCRIPTION

Internet Movie Database popularly called as " IMDb" is a online database of movies, TV shows, and celebrities. The IMDb consists of the information related to films, television series, podcasts, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews. IMDb actively gather information from and verify items with studios and filmmakers, the bulk of information is submitted by people in the industry and visitors . Analyzing IMDb data can offer insights into the popularity, critical reception, and cultural impact of movies.

❖ PROCESS APPROACH

My approach towards this project "IMDB Movie Analysis" is to explore the data to understand the relationships between different variables, correlation between movie ratings and other factors like genre, director, budget, etc. I also have to consider the year of release, the actors involved, and other relevant factors. And after exploring I have to perform the analysis on the provided dataset and perform different task using Excel functions and concepts. Analyzing the IMDB movie process through a data-driven analytics approach is essential for organizations to make informed decisions and continually improve their recruitment efforts. Centralize your data in a dedicated data repository or data warehouse. And after exploring I have to

❖ TECH-STACK USED:

- For this project I have used the Microsoft Excel 2019.
- I used this application because Microsoft Excel 2019contains many tools which format, organize and calculate data in a spreadsheet.
- Microsoft Excel 2019 represent the data in from of different charts like bar graph, pie chart, histogram graph., etc.
- Microsoft Excel is a widely used spreadsheet application developed by Microsoft. It is part of the Microsoft Office suite of productivity software. Excel is known for its powerful data manipulation and analysis features and is used for a wide range of tasks, including data entry, calculation, analysis, and visualization.

❖ INSIGHTS:

A. User Ratings:

IMDb provides user ratings on a scale of 1 to 10 for most movies. You can analyze these ratings to determine a movie's overall reception. Movies with higher average ratings are generally better received by the audience.

B. Popular Genres:

IMDb categorizes movies into various genres, such as action, drama, comedy, horror, etc. Analyzing the distribution of genres can help identify trends in movie preferences and the most popular, most common movie genres.

C. Movie duration:

Analyze the distribution of movie durations and its impact on the IMDB score. And it helps to identify the relationship between movie duration and IMDB score.

D. **Reviews:**

IMDb also contains user reviews and critic reviews. Analyzing these reviews can provide a deeper understanding of what viewers liked or disliked about a particular movie.

E. **Box Office Performance:**

IMDb may include information about a movie's box office earnings. Analyzing this data can provide insights into a movie's commercial success.

F. **Director and Cast:**

IMDb lists the directors and cast members for each movie. You can analyze whether certain directors or actors consistently produce highly-rated films.

☐ **DATA ANALYTICS TASKS:**

1.You are required to provide a detailed report for the below data record mentioning the answers of the questions that follows:

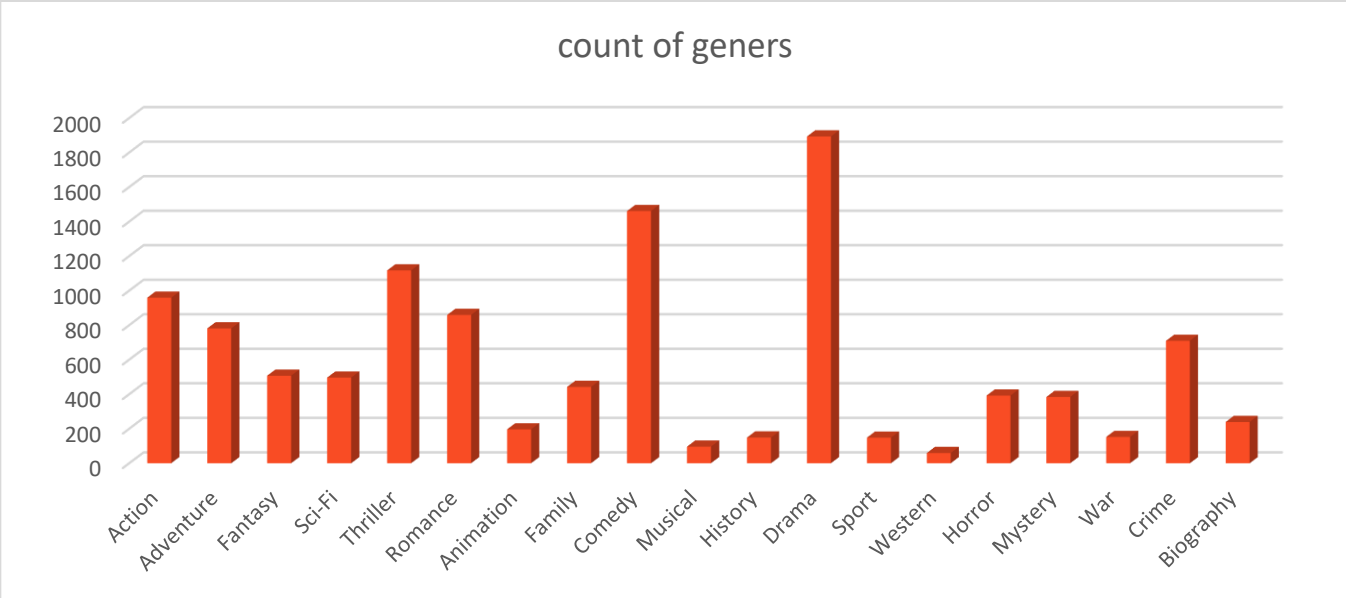
A. Movie Genre Analysis: Analyze the distribution of movie genres and their impact on the IMDB score.

Task: 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.

count of geners	average	median	mode	max	min	var	stdev
959	6.475157	6.6	6	9.3	1.6	1.115657	1.056247
781	6.421951	6.6	6	9.3	1.6	1.115657	1.056247

507	6.448649	6.6	6	9.3	1.6	1.115657	1.056247
496	7.071429	6.6	6	9.3	1.6	1.115657	1.056247
1117	7.4	6.6	6	9.3	1.6	1.115657	1.056247
859	7.9	6.6	6	9.3	1.6	1.115657	1.056247
196	6.413333	6.6	6	9.3	1.6	1.115657	1.056247
442	7.533333	6.6	6	9.3	1.6	1.115657	1.056247
1461	6.470576	6.6	6	9.3	1.6	1.115657	1.056247
96	6.65	6.6	6	9.3	1.6	1.115657	1.056247
149	#DIV/0!	6.6	6	9.3	1.6	1.115657	1.056247
1893	6.429192	6.6	6	9.3	1.6	1.115657	1.056247
148	#DIV/0!	6.6	6	9.3	1.6	1.115657	1.056247
59	7.65	6.6	6	9.3	1.6	1.115657	1.056247
392	6.582927	6.6	6	9.3	1.6	1.115657	1.056247
384	6.517391	6.6	6	9.3	1.6	1.115657	1.056247
152	#DIV/0!	6.6	6	9.3	1.6	1.115657	1.056247
709	6.412549	6.6	6	9.3	1.6	1.115657	1.056247
239	6.50878	6.6	6	9.3	1.6	1.115657	1.056247
Total:-11039							

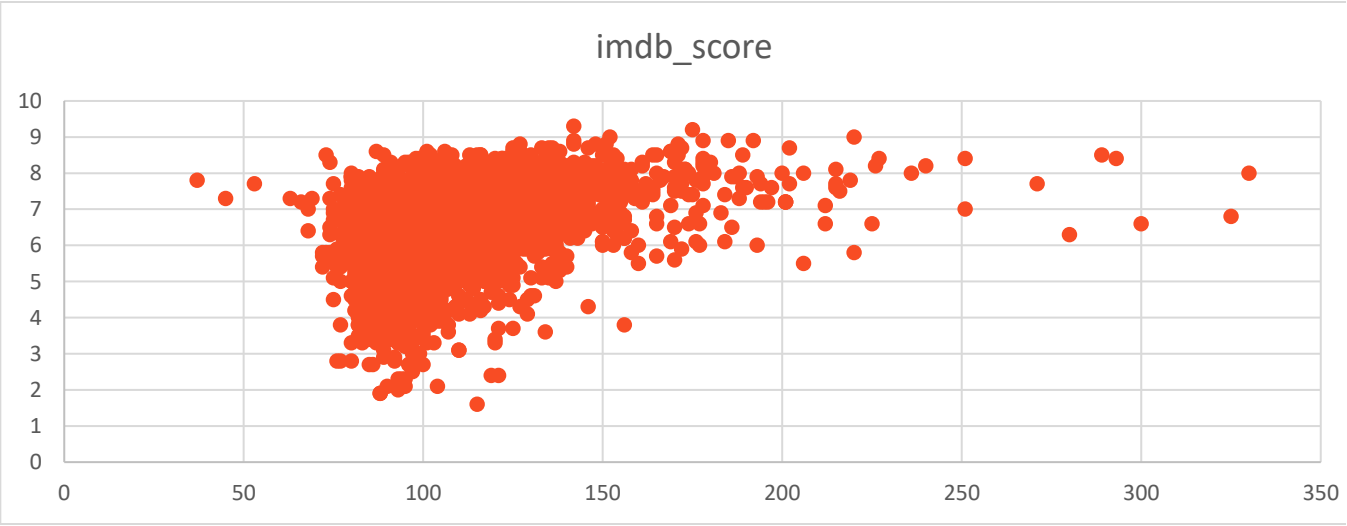
Average	581
Median	442
mode	11
Max	1893
min	59
var	249844.2222
STDEV	499.8441979



2. Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.

Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

MEAN	110.258
MEDIAN	106
Standard deviation	22.64672



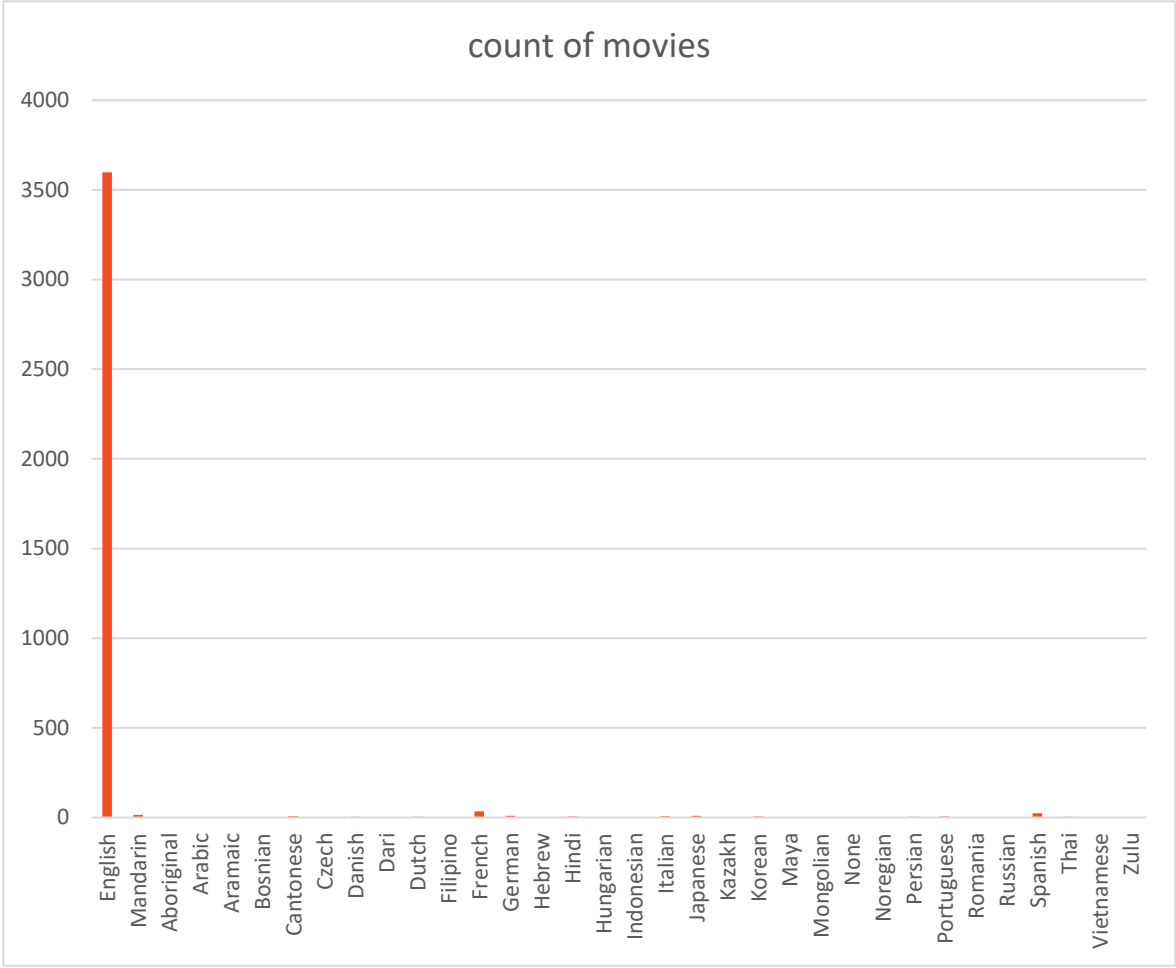
3.Language Analysis: Situation: Examine the distribution of movies based on their language.

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

Language	count of movies	average	median	stdev
English	3598	6.4270428	6.6	1.056247
Mandarin	15	7.08	6.6	1.056247
Aboriginal	2	6.95	6.6	1.056247
Arabic	1	7.2	6.6	1.056247
Aramaic	1	7.1	6.6	1.056247
Bosnian	1	4.3	6.6	1.056247
Cantonese	7	7.34285714	6.6	1.056247
Czech	1	7.4	6.6	1.056247
Danish	3	7.9	6.6	1.056247
Dari	2	7.5	6.6	1.056247
Dutch	3	7.56666667	6.6	1.056247
Filipino	1	6.7	6.6	1.056247
French	34	7.35588235	6.6	1.056247
German	10	7.77	6.6	1.056247
Hebrew	1	8	6.6	1.056247
Hindi	5	7.22	6.6	1.056247
Hungarian	1	7.1	6.6	1.056247
Indonesian	2	7.9	6.6	1.056247
Italian	7	7.18571429	6.6	1.056247
Japanese	10	7.66	6.6	1.056247
Kazakh	1	6	6.6	1.056247
Korean	5	7.7	6.6	1.056247
Maya	1	7.8	6.6	1.056247

Mongolian	1	7.3	6.6	1.056247
None	1	8.5	6.6	1.056247
Noregian	0	#DIV/0!	6.6	1.056247
Persian	3	8.13333333	6.6	1.056247
Portuguese	5	7.76	6.6	1.056247
Romania	0	#DIV/0!	6.6	1.056247
Russian	1	6.5	6.6	1.056247
Spanish	23	7.0826087	6.6	1.056247
Thai	3	6.63333333	6.6	1.056247
Vietnamese	1	7.4	6.6	1.056247
Zulu	1	7.3	6.6	1.056247

Average	110.3235294
Median	2
standard deviation	616.2969926


count of movies
4000
3500
3000
2500
2000
1500
1000
500
0
English
Mandarin
Aboriginal
Arabic
Aramaic
Bosnian
Cantonese
Czech
Danish
Dari
Dutch
Filipino
French
German
Hebrew
Hindi
Hungarian
Indonesian
Italian
Japanese
Kazakh
Korean
Maya
Mongolian
None
Noregian
Persian
Portuguese
Romania
Russian
Spanish
Thai
Vietnamese
Zulu

4. **Director Analysis:** Influence of directors on movie ratings.

Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

Movies	Dircetor	IMDB_score
The Shawshank RedemptionÂ	Frank Darabont	9.3
The GodfatherÂ	Francis Ford Coppola	9.2
The Dark KnightÂ	Christopher Nolan	9
The Godfather: Part IIÂ	Francis Ford Coppola	9
The Lord of the Rings: The Return of the KingÂ	Peter Jackson	8.9
Schindler's ListÂ	Steven Spielberg	8.9
Pulp FictionÂ	Quentin Tarantino	8.9
The Good, the Bad and the UglyÂ	Sergio Leone	8.9
InceptionÂ	Christopher Nolan	8.8
The Lord of the Rings: The Fellowship of the RingÂ	Peter Jackson	8.8
Fight ClubÂ	David Fincher	8.8
Forrest GumpÂ	Robert Zemeckis	8.8
Star Wars: Episode V - The Empire Strikes BackÂ	Irvin Kershner	8.8

LARGE	9.3
PERCENTRANK	1
PERCENTILE	9.3

5.Budget Analysis: Explore the relationship between movie budgets and their financial success.

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

Top 10 movies	Bugeet	Profit
Avatarâ	237000000	523505847
The avengersâ	220000000	403279547
Titanicâ	200000000	458672302
Jurassic worldâ	150000000	502177271
The dark knightâ	185000000	348316061
Star wars: episode i - the phantom menaceâ	115000000	359544677
The lion kingâ	45000000	377783777
The avengersâ	220000000	403279547
Star wars: episode iv - a new hopeâ	11000000	449935665
E.t. the extra-terrestrialâ	10500000	424449459

Average	4.25E+08
Max	5.24E+08
Min	3.48E+08
STDEV	58455408
Median	4.14E+08

correlation	high profit margin
0.099496423	523505847

❖ **DRIVE LINK:**

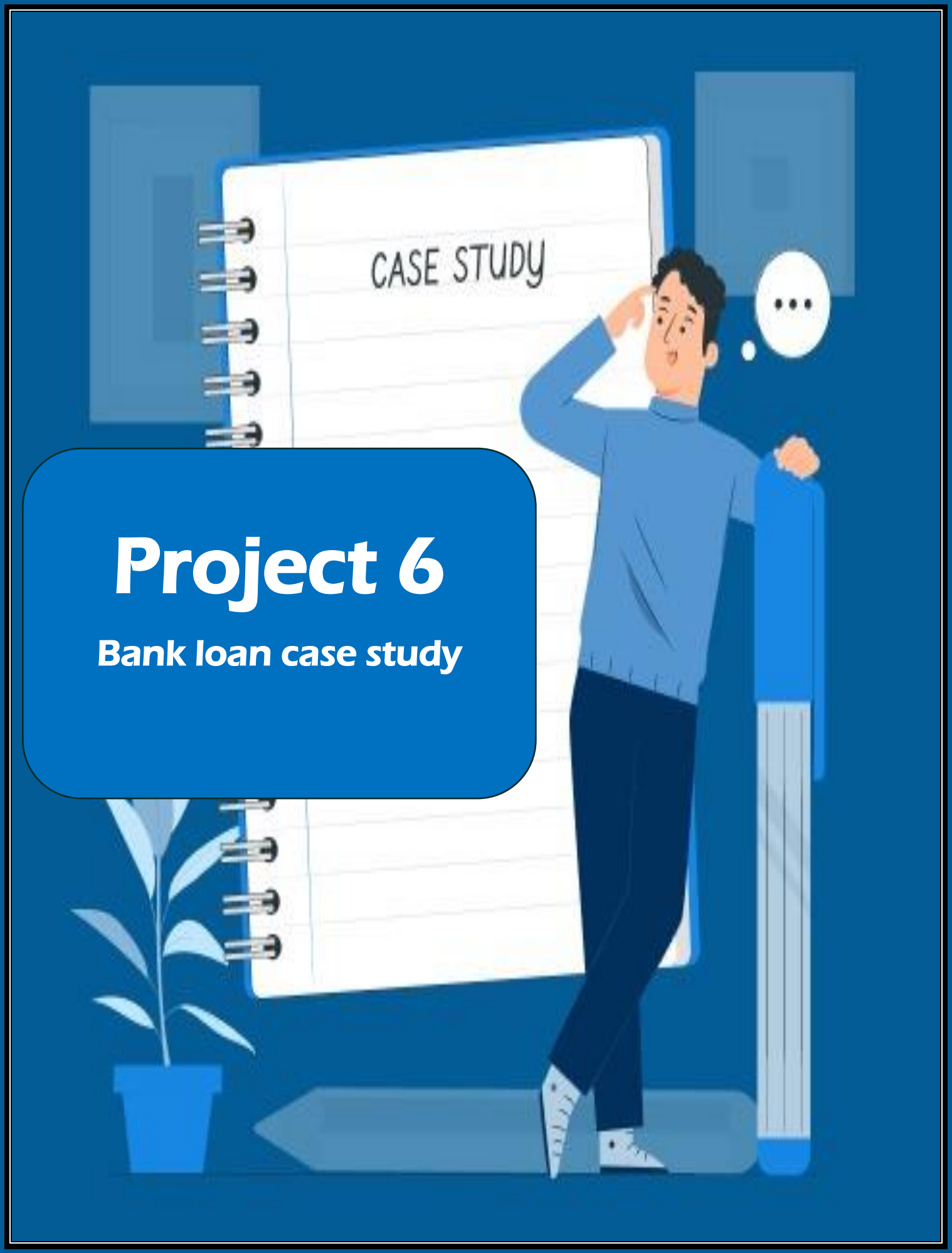
- **Task 1:-** We have make use of Use Excel's COUNTIF function to count the number of movies for each genre. You might need to manipulate the 'genres' column to separate multiple genres for a single movie. Use Excel's functions like AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV to calculate descriptive statistics. Compare the statistics to understand the impact of genre on movie ratings.

- **Task2:-** we have used Excel's functions like AVERAGE, MEDIAN, and STDEV. Create a scatter plot to visualize the relationship between movie duration and IMDB score. Add a trendline to assess the direction and strength of the relationship.

- **Task3:-**we make use of Excel's COUNTIF function to count the number of movies for each language. Calculate the mean, median, and standard deviation of the IMDB scores for each language. Compare the statistics to understand the impact of language on movie ratings.

- **Task4:-** here we make use of Excel's PERCENTILE function to identify the directors with the highest scores. Compare the scores of these directors to the overall distribution of scores.

- **Task5:-**we have used Excel's CORREL function to Calculate the profit margin (gross earnings - budget) for each movie and identify the movies with the highest profit margin using Excel's MAX function.



Project 6

Bank loan case study

INTRODUCTION

❖ PROJECT DESCRIPTION

- In this project I am suppose to be a data analyst at finance company that specializes in lending various types of loans to urban customers. And my company is facing lots of challenges like some customers who don't have a sufficient credit history take advantage of this and default on their loans.
- When a customer applies for a loan, my company faces some risks like If the applicant can repay the loan but is not approved, the company loses business. If the applicant cannot repay the loan and is approved, the company faces a financial loss. So here I am working as a data analyst where I am suppose to use the EDA the is Exploratory Data Analysis to analyze patterns in the data and ensure that capable applicants are not rejected.

❖ PROCESS APPROACH

- My approach towards Bank Loan Case Study project is to explore the data to understand the relationships between different variables, correlation.
- The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their instalments. 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. My approach towards this project is to use EDA to understand how customer attributes and loan attributes influence the likelihood of default.

❖ TECH-STACK USED

- For this project I have used the Microsoft Excel 2019.
- I used this application because Microsoft Excel 2019contains many tools which format, organize and calculate data in a spreadsheet.
- Microsoft Excel 2019 represent the data in from of different charts like bar graph, pie chart, histogram graph., etc.
- Microsoft Excel is a widely used spreadsheet application developed by Microsoft. It is part of the Microsoft Office suite of productivity software. Excel is known for its powerful data manipulation and analysis features and is used for a wide range of tasks, including data entry, calculation, analysis, and visualization.

❖ INSIGHTS

- **Financial Health:-** This insight determines the conditions like borrowers income ,availability of cash, manageable level of existing debt.
- **Loan Purpose:-** This determines the purpose of the loan whether it is for productive , business, less productive ,etc.
- **Loan Terms:-** This determines and examines the proposed loans terms which could involve loan duration, interest rate, etc.
- **Creditworthiness factors:-** determines and identifies the factors like credit score, debt to income ratio ,etc that influence the approval of loan.
- **Explore loan characteristics**:- This determines and defines the characteristics of loan such as the capital rate of interest ,etc
- **Collateral:** This determines the collateral, assess its value and the

adequacy of collateral to cover the loan amount. Insights may include whether the collateral provides adequate security for the bank.

- **Risk Assessment:** This Analyzes the risks associated with the loan. Identify potential risks such as economic conditions, industry-specific risks, or borrower-specific risks. It includes whether there are significant risk factors that need to be mitigated.

❖ DATA ANALYTICS TASKS:-

A. Identify Missing Data and Deal with it Appropriately:

As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

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

Column 1	Row	COUNT_NULL	Null_Percentage
A	SK_ID_CURR	0	0
B	TARGET	0	0
C	NAME_CONTRACT_TYPE	0	0
D	CODE_GENDER	0	0
E	FLAG_OWN_CAR	0	0
F	FLAG_OWN_REALTY	0	0
G	CNT_CHILDREN	0	0
H	AMT_INCOME_TOTAL	0	0
I	AMT_CREDIT	0	0
J	AMT_ANNUITY	1	0.002
K	AMT_GOODS_PRICE	38	0.076
L	NAME_TYPE_SUITE	192	0.384
M	NAME_INCOME_TYPE	0	0
N	NAME_EDUCATION_TYPE	0	0
O	NAME_FAMILY_STATUS	0	0
P	NAME_HOUSING_TYPE	0	0
Q	REGION_POPULATION_RELATIVE	0	0
R	DAYS_BIRTH	0	0
S	DAYS_EMPLOYED	0	0
T	DAYS_REGISTRATION	0	0
U	DAYS_ID_PUBLISH	0	0

V	OWN_CAR_AGE	32950	65.9
W	FLAG_MOBIL	0	0
X	FLAG_EMP_PHONE	0	0
Y	FLAG_WORK_PHONE	0	0
Z	FLAG_CONT_MOBILE	0	0
AA	FLAG_PHONE	0	0
AB	FLAG_EMAIL	0	0
AC	OCCUPATION_TYPE	15654	31.308
AD	CNT_FAM_MEMBERS	1	0.002
AE	REGION_RATING_CLIENT	0	0
AF	REGION_RATING_CLIENT_W_CITY	0	0
AG	WEEKDAY_APPR_PROCESS_START	0	0
AH	HOUR_APPR_PROCESS_START	0	0
AI	REG_REGION_NOT_LIVE_REGION	0	0
AJ	REG_REGION_NOT_WORK_REGION	0	0
AK	LIVE_REGION_NOT_WORK_REGION	0	0
AL	REG_CITY_NOT_LIVE_CITY	0	0
AM	REG_CITY_NOT_WORK_CITY	0	0
AN	LIVE_CITY_NOT_WORK_CITY	0	0
AO	ORGANIZATION_TYPE	0	0
AP	EXT_SOURCE_1	28172	56.344
AQ	EXT_SOURCE_2	126	0.252
AR	EXT_SOURCE_3	9944	19.888
AS	APARTMENTS_AVG	25385	50.77
AT	BASEMENTAREA_AVG	29199	58.398
AU	YEARS_BEGINEXPLUATATION_AVG	24394	48.788
AV	YEARS_BUILD_AVG	33239	66.478
AW	COMMONAREA_AVG	34960	69.92
AX	ELEVATORS_AVG	26651	53.302
AY	ENTRANCES_AVG	25195	50.39
AZ	FLOORSMAX_AVG	24875	49.75
BA	FLOORSMIN_AVG	33894	67.788
BB	LANDAREA_AVG	29721	59.442
BC	LIVINGAPARTMENTS_AVG	34226	68.452
BD	LIVINGAREA_AVG	25137	50.274
BE	NONLIVINGAPARTMENTS_AVG	34714	69.428
BF	NONLIVINGAREA_AVG	27572	55.144
BG	APARTMENTS_MODE	25385	50.77
BH	BASEMENTAREA_MODE	29199	58.398
BI	YEARS_BEGINEXPLUATATION_MODE	24394	48.788

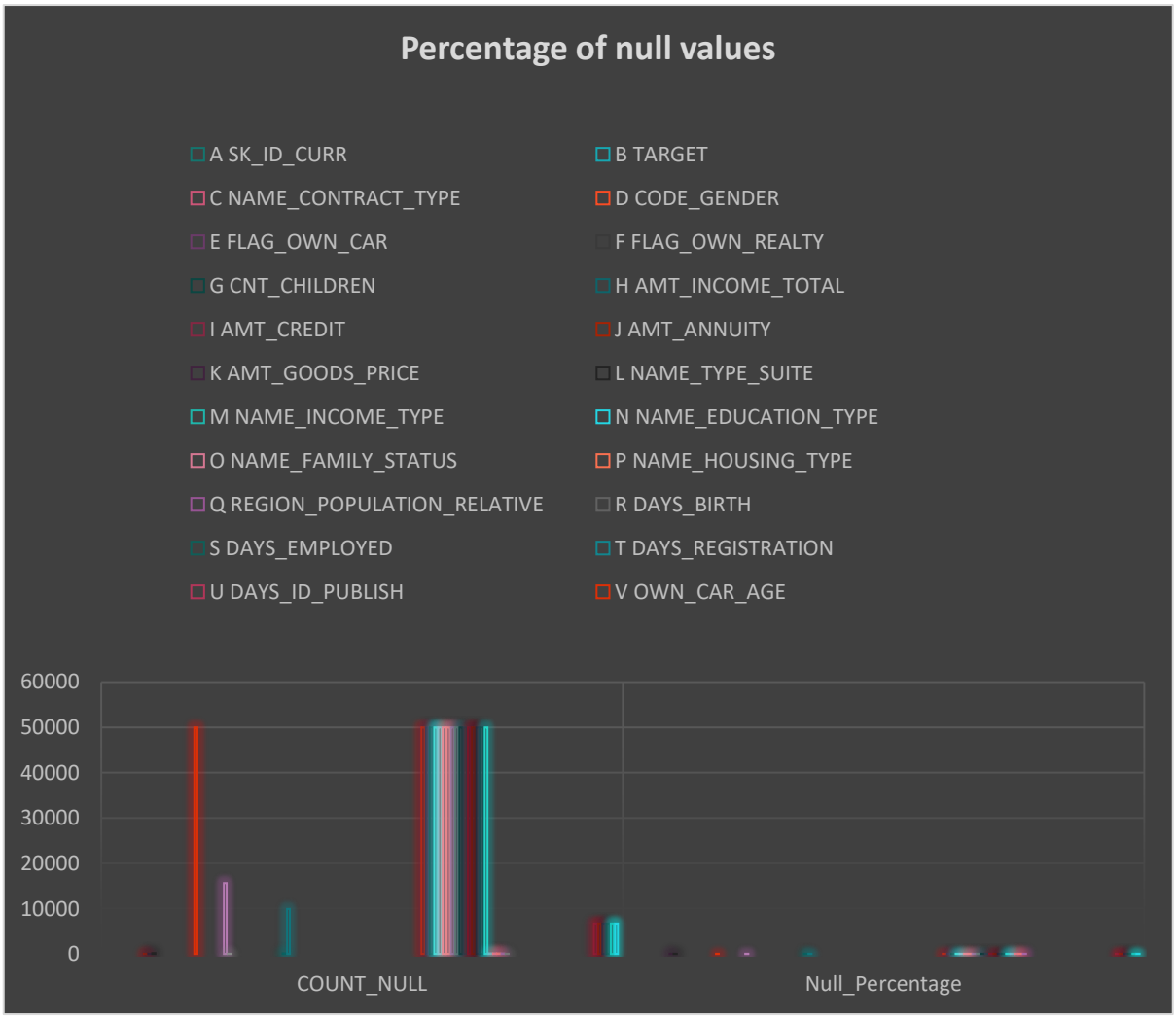
BJ	YEARS_BUILD_MODE	33239	66.478
BK	COMMONAREA_MODE	34960	69.92
BL	ELEVATORS_MODE	26651	53.302
BM	ENTRANCES_MODE	25195	50.39
BN	FLOORSMAX_MODE	24875	49.75
BO	FLOORSMIN_MODE	33894	67.788
BP	LANDAREA_MODE	29721	59.442
BQ	LIVINGAPARTMENTS_MODE	34226	68.452
BR	LIVINGAREA_MODE	25137	50.274
BS	NONLIVINGAPARTMENTS_MODE	34714	69.428
BT	NONLIVINGAREA_MODE	27572	55.144
BU	APARTMENTS_MEDI	25385	50.77
BV	BASEMENTAREA_MEDI	29199	58.398
BW	YEARS_BEGINEXPLUATATION_MEDI	24394	48.788
BX	YEARS_BUILD_MEDI	33239	66.478
BY	COMMONAREA_MEDI	34960	69.92
BZ	ELEVATORS_MEDI	26651	53.302
CA	ENTRANCES_MEDI	25195	50.39
CB	FLOORSMAX_MEDI	24875	49.75
CC	FLOORSMIN_MEDI	33894	67.788
CD	LANDAREA_MEDI	29721	59.442
CE	LIVINGAPARTMENTS_MEDI	34226	68.452
CF	LIVINGAREA_MEDI	25137	50.274
CG	NONLIVINGAPARTMENTS_MEDI	34714	69.428
CH	NONLIVINGAREA_MEDI	27572	55.144
CI	FONDKAPREMONT_MODE	34191	68.382
CJ	HOUSETYPE_MODE	25075	50.15
CK	TOTALAREA_MODE	24148	48.296
CL	WALLSMATERIAL_MODE	25459	50.918
CM	EMERGENCYSTATE_MODE	23698	47.396
CN	OBS_30_CNT_SOCIAL_CIRCLE	168	0.336
CO	DEF_30_CNT_SOCIAL_CIRCLE	168	0.336
CP	OBS_60_CNT_SOCIAL_CIRCLE	168	0.336
CQ	DEF_60_CNT_SOCIAL_CIRCLE	168	0.336
CR	DAYS_LAST_PHONE_CHANGE	1	0.002
CS	FLAG_DOCUMENT_2	0	0
CT	FLAG_DOCUMENT_3	0	0
CU	FLAG_DOCUMENT_4	0	0
CV	FLAG_DOCUMENT_5	0	0
CW	FLAG_DOCUMENT_6	0	0
CX	FLAG_DOCUMENT_7	0	0
CY	FLAG_DOCUMENT_8	0	0
CZ	FLAG_DOCUMENT_9	0	0
DA	FLAG_DOCUMENT_10	0	0

DB	FLAG_DOCUMENT_11	0	0
DC	FLAG_DOCUMENT_12	0	0
DD	FLAG_DOCUMENT_13	0	0
DE	FLAG_DOCUMENT_14	0	0
DF	FLAG_DOCUMENT_15	0	0
DG	FLAG_DOCUMENT_16	0	0
DH	FLAG_DOCUMENT_17	0	0
DI	FLAG_DOCUMENT_18	0	0
DJ	FLAG_DOCUMENT_19	0	0
DK	FLAG_DOCUMENT_20	0	0
DL	FLAG_DOCUMENT_21	0	0
DM	AMT_REQ_CREDIT_BUREAU_HOUR	6734	13.468
DN	AMT_REQ_CREDIT_BUREAU_DAY	6734	13.468
DO	AMT_REQ_CREDIT_BUREAU_WEEK	6735	13.47
DP	AMT_REQ_CREDIT_BUREAU_MON	6736	13.472
DQ	AMT_REQ_CREDIT_BUREAU_QRT	6737	13.474
DR	AMT_REQ_CREDIT_BUREAU_YEAR	6738	13.476

- The final sheet after removing the 47 columns with null percentage more than 40%

Column1	Row	COUNT_NULL	Null_Percentage
V	OWN_CAR_AGE	32950	65.9
AP	EXT_SOURCE_1	28172	56.344
AS	APARTMENTS_AVG	25385	50.77
AT	BASEMENTAREA_AVG	29199	58.398
AU	YEARS_BEGINEXPLUATATION_AVG	24394	48.788
AV	YEARS_BUILD_AVG	33239	66.478
AW	COMMONAREA_AVG	34960	69.92
AX	ELEVATORS_AVG	26651	53.302
AY	ENTRANCES_AVG	25195	50.39
AZ	FLOORSMAX_AVG	24875	49.75
BA	FLOORSMIN_AVG	33894	67.788
BB	LANDAREA_AVG	29721	59.442
BC	LIVINGAPARTMENTS_AVG	34226	68.452
BD	LIVINGAREA_AVG	25137	50.274
BE	NONLIVINGAPARTMENTS_AVG	34714	69.428
BF	NONLIVINGAREA_AVG	27572	55.144
BG	APARTMENTS_MODE	25385	50.77
BH	BASEMENTAREA_MODE	29199	58.398
BI	YEARS_BEGINEXPLUATATION_MODE	24394	48.788

BJ	YEARS_BUILD_MODE	33239	66.478
BK	COMMONAREA_MODE	34960	69.92
BL	ELEVATORS_MODE	26651	53.302
BM	ENTRANCES_MODE	25195	50.39
BN	FLOORSMAX_MODE	24875	49.75
BO	FLOORSMIN_MODE	33894	67.788
BP	LANDAREA_MODE	29721	59.442
BQ	LIVINGAPARTMENTS_MODE	34226	68.452
BR	LIVINGAREA_MODE	25137	50.274
BS	NONLIVINGAPARTMENTS_MODE	34714	69.428
BT	NONLIVINGAREA_MODE	27572	55.144
BU	APARTMENTS_MEDI	25385	50.77
BV	BASEMENTAREA_MEDI	29199	58.398
BW	YEARS_BEGINEXPLUATATION_MEDI	24394	48.788
BX	YEARS_BUILD_MEDI	33239	66.478
BY	COMMONAREA_MEDI	34960	69.92
BZ	ELEVATORS_MEDI	26651	53.302
CA	ENTRANCES_MEDI	25195	50.39
CB	FLOORSMAX_MEDI	24875	49.75
CC	FLOORSMIN_MEDI	33894	67.788
CD	LANDAREA_MEDI	29721	59.442
CE	LIVINGAPARTMENTS_MEDI	34226	68.452
CF	LIVINGAREA_MEDI	25137	50.274
CG	NONLIVINGAPARTMENTS_MEDI	34714	69.428
CH	NONLIVINGAREA_MEDI	27572	55.144
CI	FONDKAPREMONT_MODE	34191	68.382
CJ	HOUSETYPE_MODE	25075	50.15
CK	TOTALAREA_MODE	24148	48.296
CL	WALLSMATERIAL_MODE	25459	50.918
CM	EMERGENCYSTATE_MODE	23698	47.396


Percentage of null values
A SK_ID_CURR
B TARGET
C NAME_CONTRACT_TYPE
D CODE_GENDER
E FLAG_OWN_CAR
F FLAG_OWN_REALTY
G CNT_CHILDREN
H AMT_INCOME_TOTAL
I AMT_CREDIT
J AMT_ANNUITY
K AMT_GOODS_PRICE
L NAME_TYPE_SUITE
M NAME_INCOME_TYPE
N NAME_EDUCATION_TYPE
O NAME_FAMILY_STATUS
P NAME_HOUSING_TYPE
Q REGION_POPULATION_RELATIVE
R DAYS_BIRTH
S DAYS_EMPLOYED
T DAYS_REGISTRATION
U DAYS_ID_PUBLISH
V OWN_CAR_AGE
60000
50000
40000
30000
20000
10000
0
COUNT_NULL
Null_Percentage

B. Identify Outliers in the Dataset:

Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

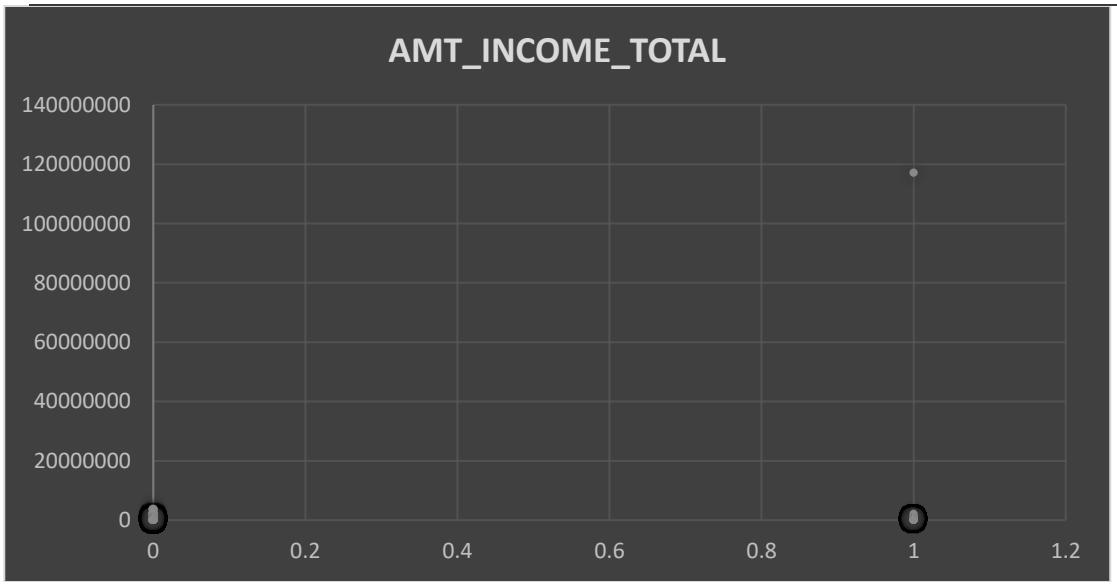
Task: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

(I) Outlier in AMT_INCOME_TOTAL :-

Quartiles 3	Quartiles 1	Inter Quartile Range
202500	112500	90000

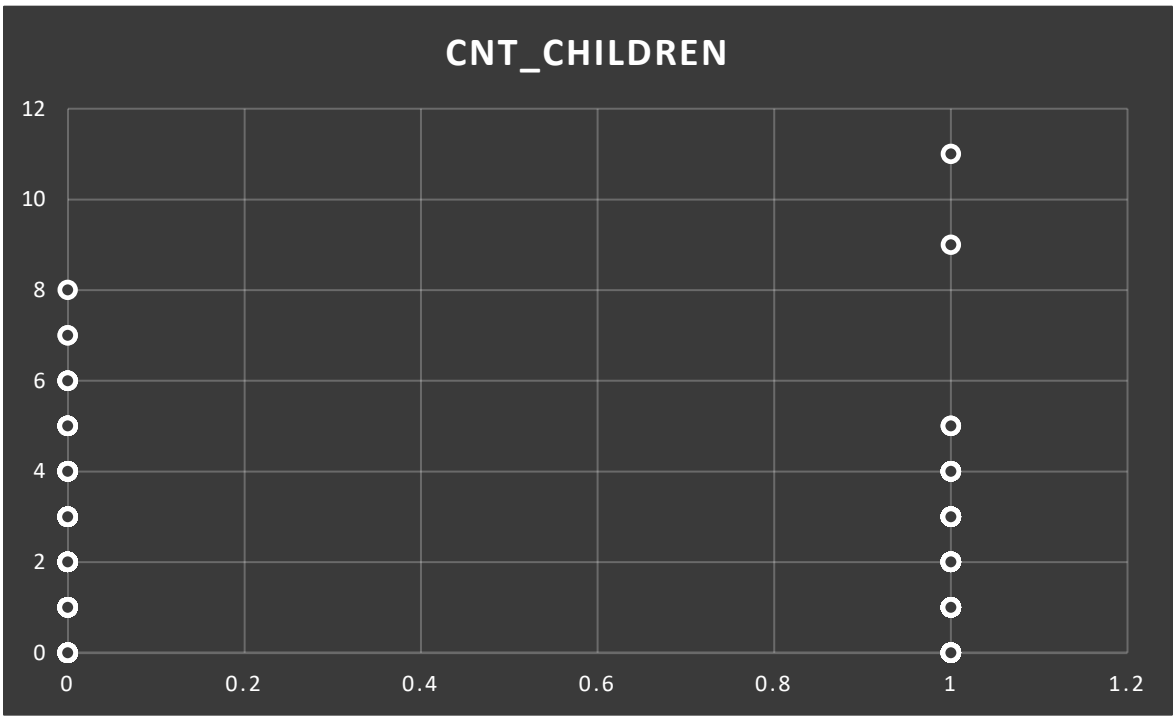
Upper Limit	Lower limit
337500	-22500

AMT_INCOME_TOTAL	
Mean	170767.5905
Standard Error	531813.7768
Median	145800
Mode	135000
Standard Deviation	531819.0951
Sample Variance	2.82832E+11
Range	116974350
Minimum	25650
Maximum	117000000
Sum	8538208758
Count	49999



(II) Outlier CNT_CHILDERN :-

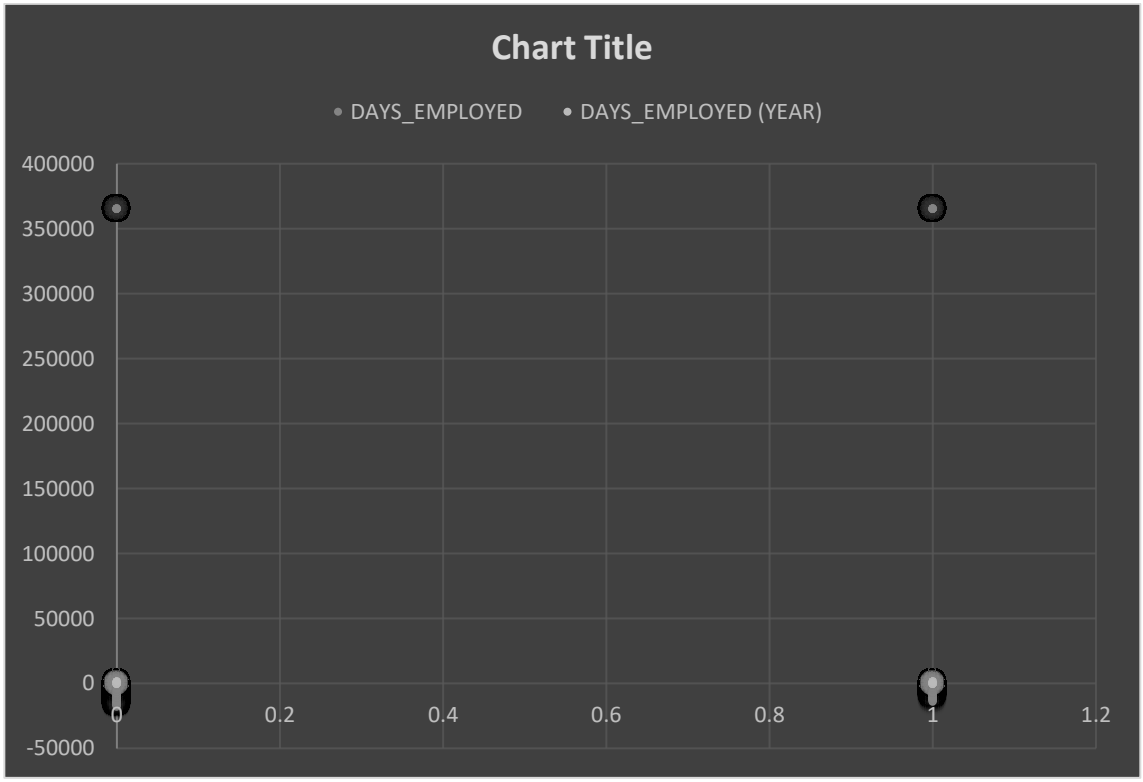
Mean	0.419848
Standard Error	0.724031
Median	0
Mode	0
Standard Deviation	0.724039
Sample Variance	0.524232
Range	11
Minimum	0
Maximum	11
Sum	20992
Count	0



(III) Outlier DAYS_EMPLOYED(YEAR):-

Quartiles 1	Quartiles 3	Inter Quartile Range	Upper Limit	Lower Limit
2.556164384	15.665753	13.10958904	35.33013699	22.22054795

Mean	63219.42449
Standard Error	140793.1977
Median	-1221
Standard Deviation	140794.6057
Sample Variance	19823120985
Range	382774
Minimum	-17531
Maximum	365243
Sum	3160908005
Count	49999



C. Analyze Data Imbalance:

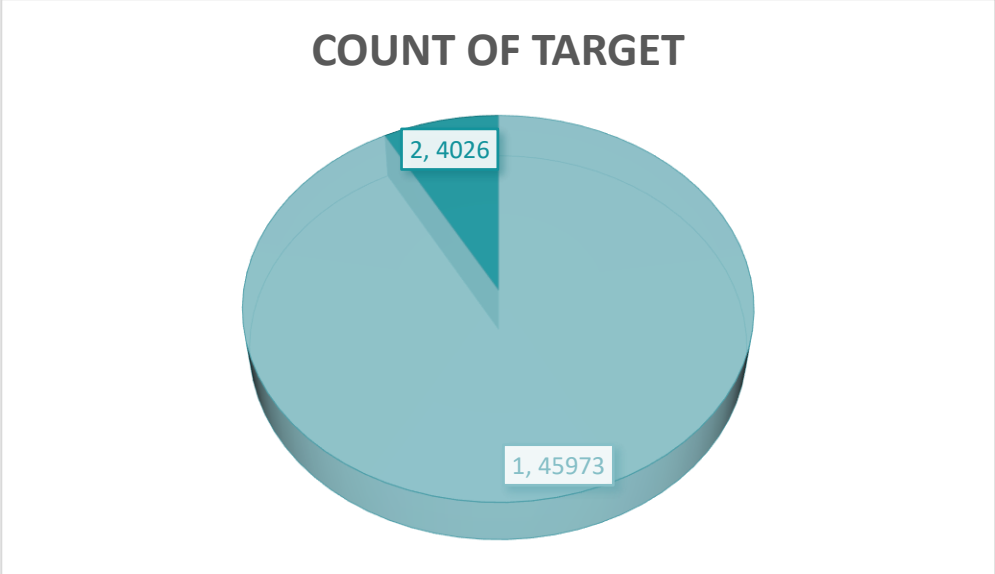
Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

Target	Count of Target
0	45973
1	4026
Grand Total	49999

Target	Contribution
0	91.94783896
1	8.052161043

Target	count of 0's and1's	ratio
0	45973	11.41903
1	4026	



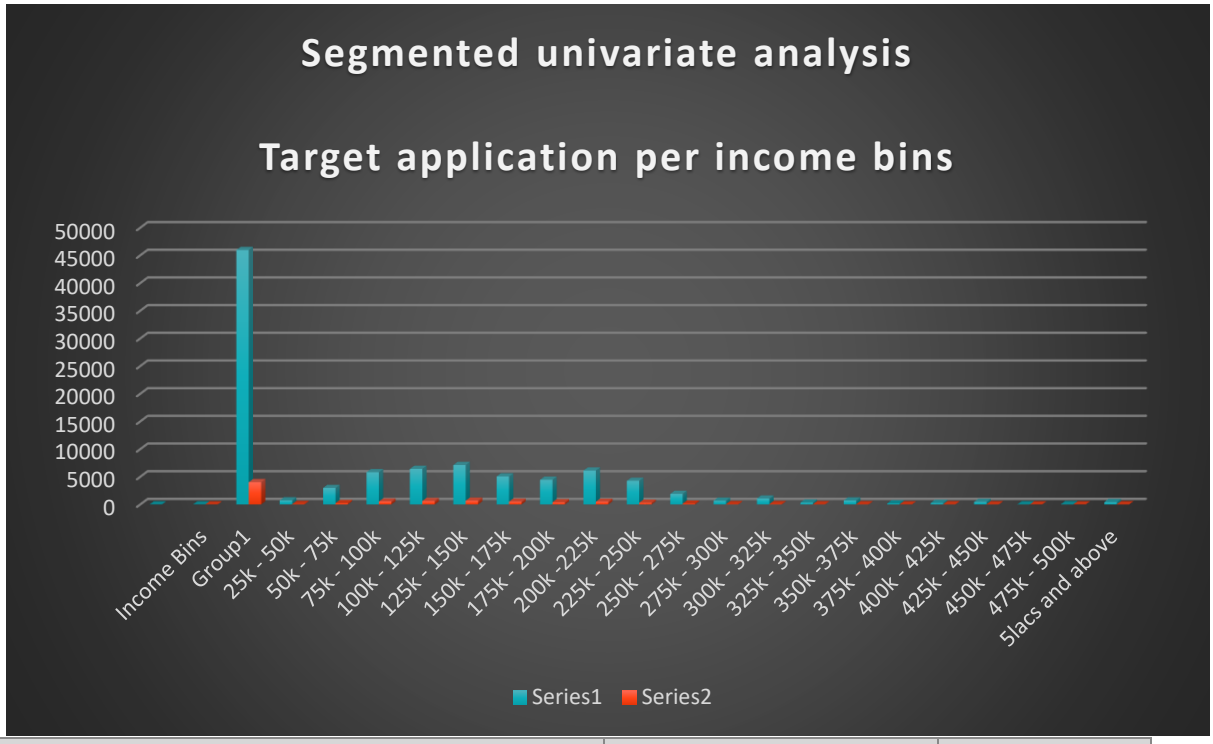
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis:

To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

Task: 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.

(I) Target applicants per income bins

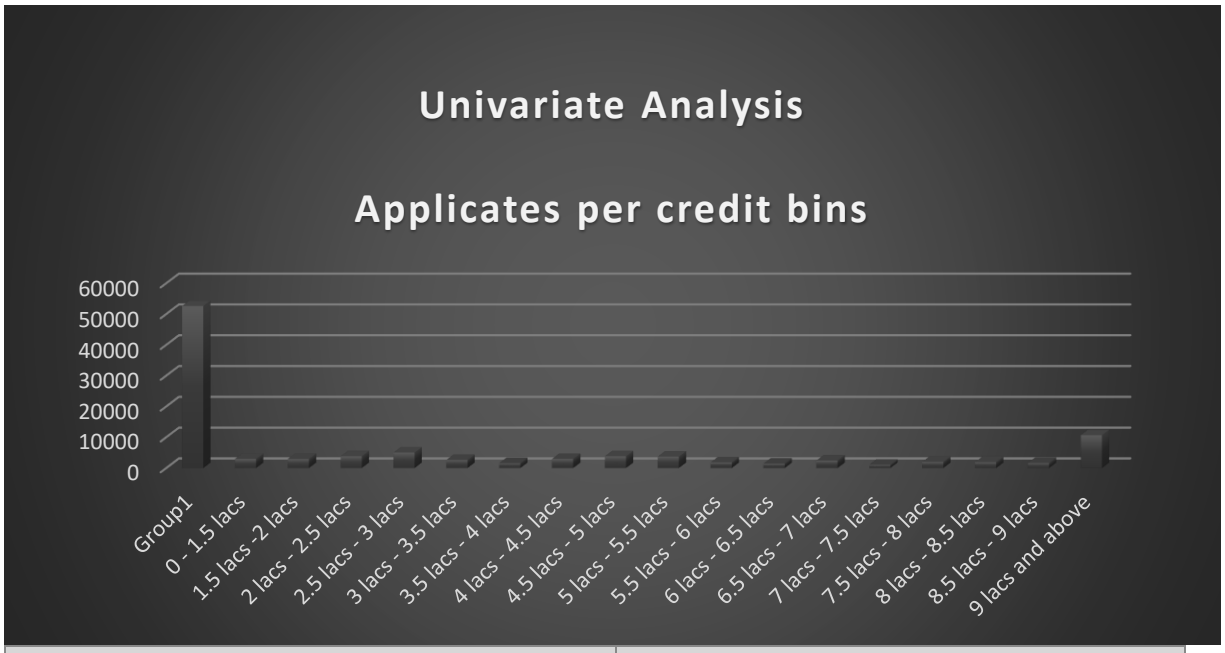
Target		
Income Bins	0	1
Group1	45973	4026
25k - 50k	741	63
50k - 75k	2980	246
75k - 100k	5826	536
100k - 125k	6428	620
125k - 150k	7126	678
150k - 175k	5060	501
175k - 200k	4458	389
200k -225k	6121	491
225k - 250k	4279	304
250k - 275k	1919	143
275k - 300k	681	45
300k - 325k	1076	59
325k - 350k	322	24
350k -375k	723	34
375k - 400k	186	14
400k - 425k	263	26
425k - 450k	456	36
450k - 475k	19	2



475k - 500k	**44**	**3**
5lacs and above	**423**	**31**

(II) Applicants per credit bins

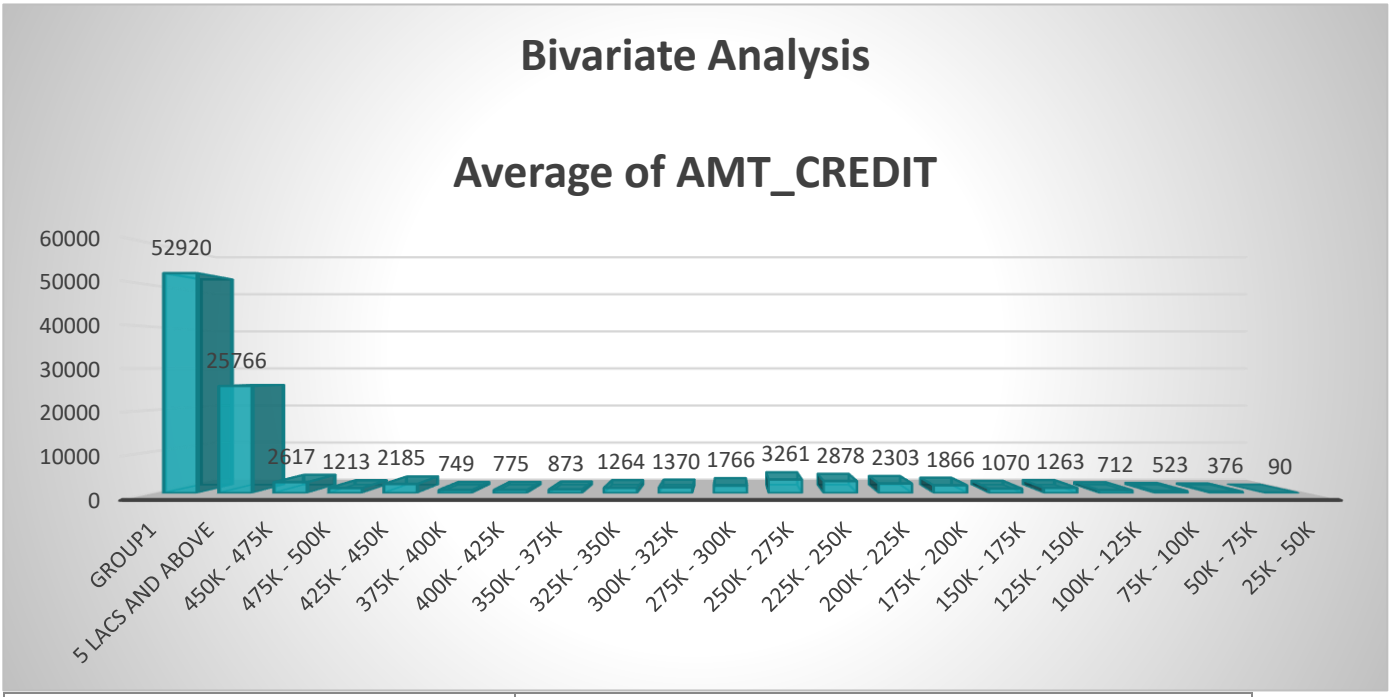
Credit Bins	Applicates
Group1	**52585**
0 - 1.5 lacs	**2964**
1.5 lacs -2 lacs	**2936**
2 lacs - 2.5 lacs	**3822**
2.5 lacs - 3 lacs	**5027**
3 lacs - 3.5 lacs	**2634**
3.5 lacs - 4 lacs	**1622**
4 lacs - 4.5 lacs	**2960**
4.5 lacs - 5 lacs	**3830**
5 lacs - 5.5 lacs	**3708**
5.5 lacs - 6 lacs	**1846**
6 lacs - 6.5 lacs	**1464**



6.5 lacs - 7 lacs	**2445**
7 lacs - 7.5 lacs	**1066**
7.5 lacs - 8 lacs	**1996**
8 lacs - 8.5 lacs	**1911**
8.5 lacs - 9 lacs	**1660**
9 lacs and above	**10694**

(III) AVERAGE CREDIT AMOUNT PER INCOME BIN

Income Bins	Average of AMT_CREDIT
Group1	**52920**
5 Lacs and above	**25766**
450k - 475k	**2617**
475k - 500k	**1213**
425k - 450k	**2185**
375k - 400k	**749**
400k - 425k	**775**



350k - 375k	**873**
325k - 350k	**1264**
300k - 325k	**1370**
275k - 300k	**1766**
250k - 275k	**3261**
225k - 250k	**2878**
200k - 225k	**2303**
175k - 200k	**1866**
150k - 175k	**1070**
125k - 150k	**1263**
100k - 125k	**712**
75k - 100k	**523**
50k - 75k	**376**
25k - 50k	**90**

E. Identify Top Correlations for Different Scenarios:

Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

Task: 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.

Column1	CNT_CHILDREN	AMT_INCOME_TOTAL
CNT_CHILDREN	1	0.036319722
AMT_INCOME_TOTAL	0.036319722	1
AMT_CREDIT	0.005705458	0.377965752
REGION_POPULATION_RELATIVE	-0.024912809	0.181941261
DAYS_BIRTH	0.335876269	0.073769425
DAYS_EMPLOYED (YEAR)	-0.245521512	-0.161680938
DAYS_ID_PUBLISH (YEAR)	0.032537221	-0.032286356
REGION_RATING_CLIENT	0.021288992	-0.205031899

AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH
0.005705458	-0.024912809	0.335876269
0.377965752	0.181941261	0.073769425
1	0.095539444	-0.051084182
0.095539444	1	-0.030435419
-0.051084182	-0.030435419	1
-0.074733443	-0.006767142	-0.623474675
0.008290189	0.002236288	-0.270073313
-0.102556478	-0.539333113	0.00902485

DAYS_EMPLOYED (YEAR)	DAYS_ID_PUBLISH (YEAR)	REGION_RATING_CLIENT
-0.245521512	0.032537221	0.021288992
-0.161680938	-0.032286356	-0.205031899

-0.074733443	0.008290189	-0.102556478
-0.006767142	0.002236288	-0.539333113
-0.623474675	-0.270073313	0.00902485
1	0.274516224	0.040937165
0.274516224	1	0.008097427
0.040937165	0.008097427	1

- **DRIVE LINK:-**

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

B TASK:- Here we have Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

C TASK:- Here we Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

D TASK:- Here we have to 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.

E TASK:- Herer we have identify the top correlations for each segmented data using Excel functions.

Project 7

Impact Of Car Features

INTRODUCTION

❖ PROJECT DESCRIPTION :

- Working 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.

- The provided dataset contains information on various car models and their specifications. The dataset contains information on over 11,000 car models and their specifications, including details on the car's make, model, year, fuel type, engine power, transmission, wheels, number of doors, market category, size, style, estimated miles per gallon, popularity, and manufacturer's suggested retail price (MSRP). This dataset could be use to gain insights into various aspects of the automotive industry.

- Before working with the dataset it is important to perform the process of data cleaning on the dataset so that the analysis process goes smoothly.

❖ PROCESS APPROACH :

- By using the provided dataset we have to perform some tasks and have to provide the related insights and to perform the task we have to use the analytical methods such as descriptive statistics, visualization, machine learning, or optimization.

- In the first, second and fourth task we have to use the data visualization technique to visualizes the relationship between market category and popularity, to visualize the relationship between price and engine power, to visualize the relationship between manufacturer and average price.

- In the third task we have to find use the regression analysis to identify the variables that have the strongest relationship with a car's price.

- In the fifth task we have calculated the correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.

- In the next part of the project where we have to create an interactive dashboard there we have use the functions like SUMIF, AVERAGEIFS, etc. we have also make the use of pivot table and also the slicers to make our dashboard interactive.

- To provide the accurate result for the particular task we have use the functions like optimization, visualization and also use the functions like sumif, averageifs, etc.

❖ TECH-STACK USED :

- For this project I have used the Microsoft Excel 2019.
- I used this application because Microsoft Excel 2019contains many tools which format, organize and calculate data in a spreadsheet.
- Microsoft Excel 2019 represent the data in from of different charts like bar graph, pie chart, histogram graph., etc.
- Microsoft Excel is a widely used spreadsheet application developed by Microsoft. It is part of the Microsoft Office suite of productivity software. Excel is known for its powerful data manipulation and analysis features and is used for a wide range of tasks, including data entry, calculation, analysis, and visualization.

❖ INSIGHTS :

- **Popularity :-** the popularity of the car model vary across different market categories which can be given as the car models in each market category and their corresponding popularity scores.
- **Relationship:-** the relation between the car's engine power and its price is derived using the data visualization technique.
- **Important features:-** car features which are most important in determining a car's price is given through regression analysis by identify the variables that have the strongest relationship with a car's price.
- **Average price vary:-** the average price of a car vary across different manufacturers is identified using data visualization technique by deriving the relationship between manufacturer and average price.
- **Strength and direction of the relationship:-** relationship between fuel

efficiency and the number of cylinders in a car's engine is Calculated by giving the correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.

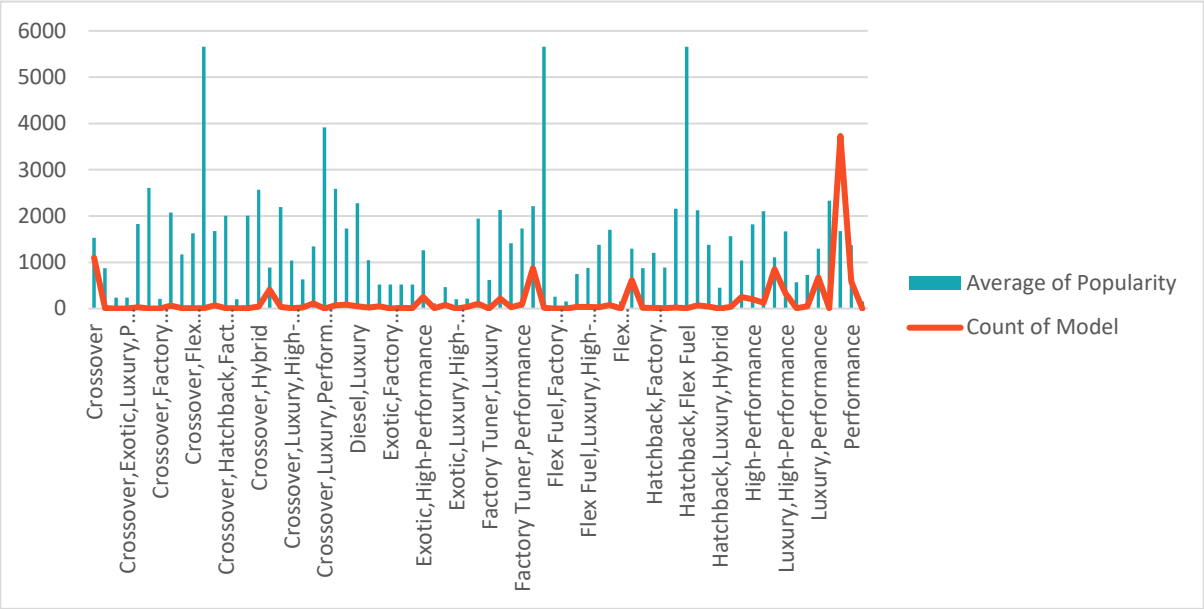
❖ TASK-ANALYSIS:

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

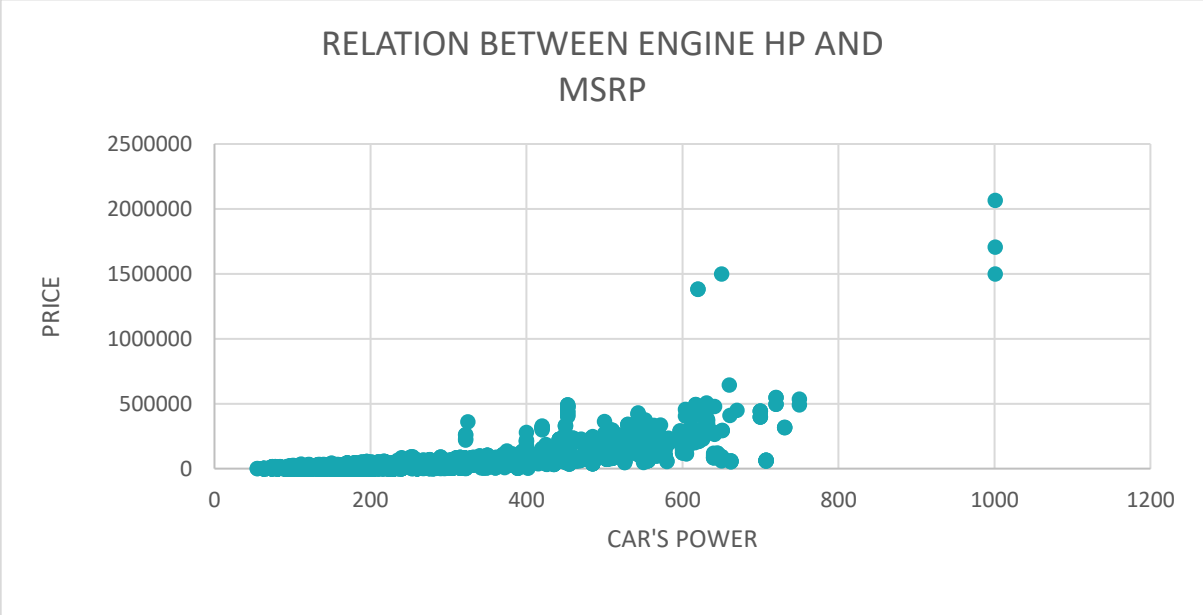
- **Task 1.A:** Create a pivot table that shows the number of car models in each market category and their corresponding popularity scores.
- **Task 1.B:** Create a combo chart that visualizes the relationship between market category and popularity.

MARKET CATEGORY	Average of Popularity	Count of Model
Crossover	1529.030825	1103
Crossover,Diesel	873	7
Crossover,Exotic,Luxury,High-Performance	238	1
Crossover,Exotic,Luxury,Performance	238	1
Crossover,Factory Tuner,Luxury,High-Performance	1823.461538	26
Crossover,Factory Tuner,Luxury,Performance	2607.4	5
Crossover,Factory Tuner,Performance	210	4
Crossover,Flex Fuel	2073.75	64
Crossover,Flex Fuel,Luxury	1173.2	10
Crossover,Flex Fuel,Luxury,Performance	1624	6
Crossover,Flex Fuel,Performance	5657	6
Crossover,Hatchback	1675.694444	72
Crossover,Hatchback,Factory Tuner,Performance	2009	6
Crossover,Hatchback,Luxury	204	7
Crossover,Hatchback,Performance	2009	6
Crossover,Hybrid	2563.380952	42
Crossover,Luxury	884.5487805	410
Crossover,Luxury,Diesel	2195.848485	33
Crossover,Luxury,High-Performance	1037.222222	9
Crossover,Luxury,Hybrid	630.9166667	24
Crossover,Luxury,Performance	1344.849558	113
Crossover,Luxury,Performance,Hybrid	3916	2
Crossover,Performance	2585.956522	69
Diesel	1730.904762	84
Diesel,Luxury	2275	51
Exotic,Factory Tuner,High-Performance	1046.380952	21
Exotic,Factory Tuner,Luxury,High-Performance	517.5384615	52

Exotic,Factory Tuner,Luxury,Performance	520	3
Exotic,Flex Fuel,Factory Tuner,Luxury,High-Performance	520	13
Exotic,Flex Fuel,Luxury,High-Performance	520	11
Exotic,High-Performance	1261.571429	252
Exotic,Luxury	112.6666667	12
Exotic,Luxury,High-Performance	467.0759494	79
Exotic,Luxury,High-Performance,Hybrid	204	1
Exotic,Luxury,Performance	217.0277778	36
Factory Tuner,High-Performance	1941.415094	106
Factory Tuner,Luxury	617	2
Factory Tuner,Luxury,High-Performance	2133.367442	215
Factory Tuner,Luxury,Performance	1413.419355	31
Factory Tuner,Performance	1733.101124	89
Flex Fuel	2217.302752	872
Flex Fuel,Diesel	5657	16
Flex Fuel,Factory Tuner,Luxury,High-Performance	258	1
Flex Fuel,Hybrid	155	2
Flex Fuel,Luxury	746.5384615	39
Flex Fuel,Luxury,High-Performance	878.9090909	33
Flex Fuel,Luxury,Performance	1380.071429	28
Flex Fuel,Performance	1702.358025	81
Flex Fuel,Performance,Hybrid	155	2
Hatchback	1292.998371	614
Hatchback,Diesel	873	14
Hatchback,Factory Tuner,High-Performance	1205.153846	13
Hatchback,Factory Tuner,Luxury,Performance	886.8888889	9
Hatchback,Factory Tuner,Performance	2159.045455	22
Hatchback,Flex Fuel	5657	7
Hatchback,Hybrid	2121.25	72
Hatchback,Luxury	1379.5	46
Hatchback,Luxury,Hybrid	454	3
Hatchback,Luxury,Performance	1566.131579	38
Hatchback,Performance	1039.646825	252
High-Performance	1821.447236	199
Hybrid	2105.569106	123
Luxury	1107.553467	851
Luxury,High-Performance	1668.017964	334
Luxury,High-Performance,Hybrid	568.8333333	12
Luxury,Hybrid	724.6875	48
Luxury,Performance	1292.615156	673
Luxury,Performance,Hybrid	2333.181818	11
N/A	1671.388144	3728
Performance	1371.080479	584
Performance,Hybrid	155	1
Grand Total	**1553.679902**	**11812**



❖ **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.



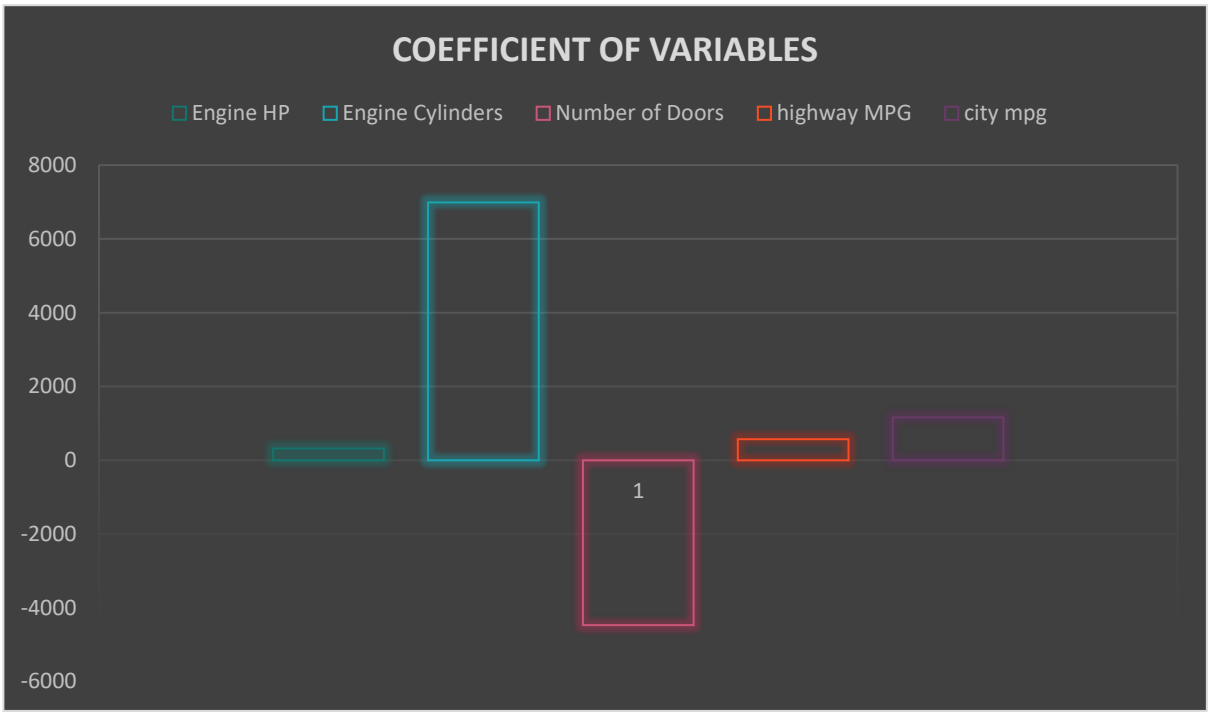
❖ **Which car features are most important in determining a car's price?**
Task 3: Use regression analysis to identify the variables that have the strongest relationship with a car's price. Then create a bar chart that shows the coefficient values for each variable to visualize their relative importance.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.680708139
R Square	0.46336357
Adjusted R Square	0.463136297
Standard Error	44170.77827
Observations	11812

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	1.99E+13	3.98E+12	2038.799	0
Residual	11806	2.3E+13	1.95E+09		
Total	11811	4.29E+13			

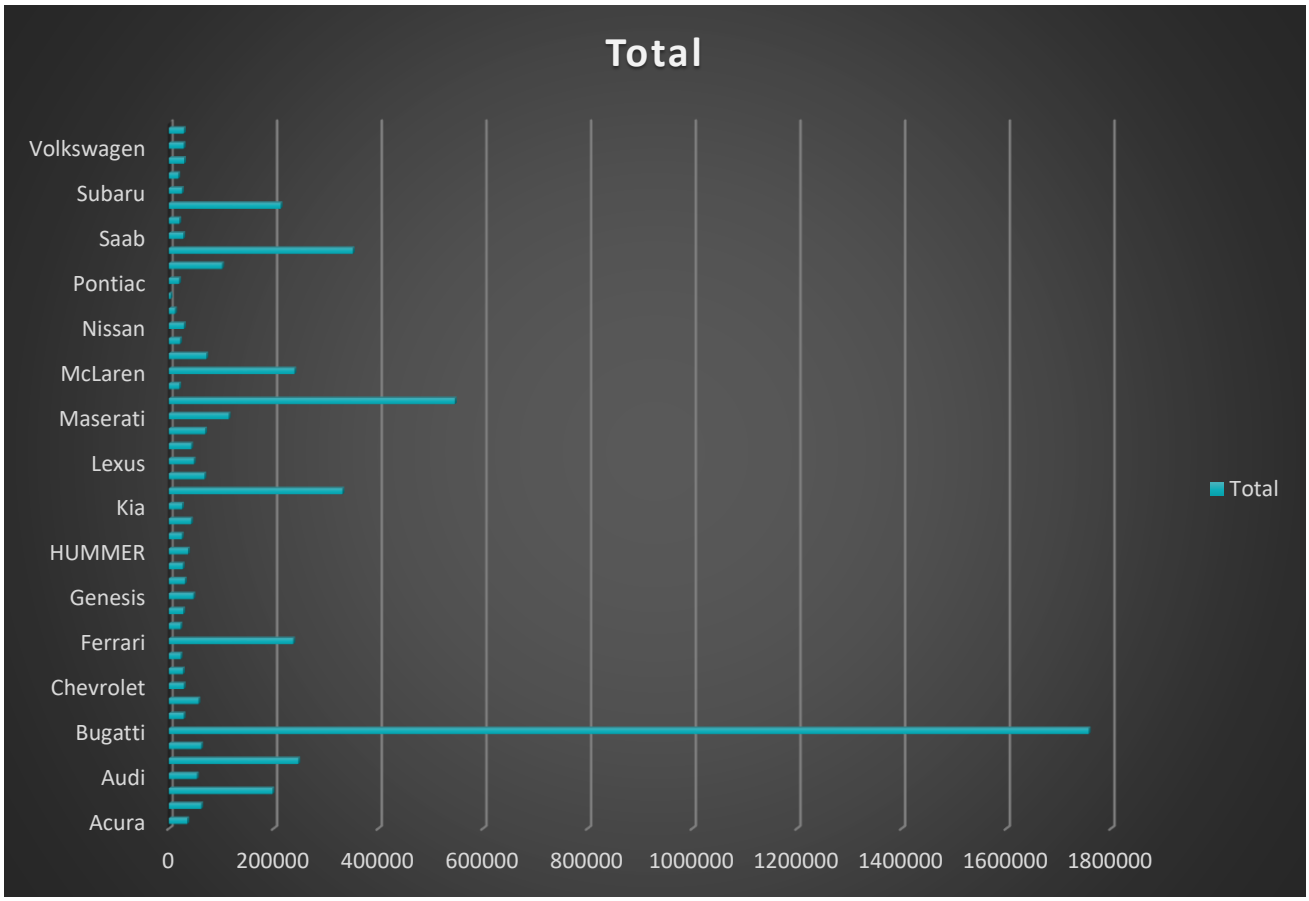
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-101601.736	3684.352	-27.5766	2.8E-162	-108824	-94379.8	-108824	-94379.8
Engine HP	322.7465574	6.017674	53.63311	0	310.9509	334.5422	310.9509	334.5422
Engine Cylinders	6989.177662	439.645	15.89732	2.54E-56	6127.401	7850.954	6127.401	7850.954
Number of Doors	-4472.158125	465.7181	-9.60272	9.35E-22	-5385.04	-3559.27	-5385.04	-3559.27
highway MPG	570.1808088	105.784	5.390049	7.18E-08	362.8268	777.5349	362.8268	777.5349
city mpg	1163.755457	121.9978	9.53915	1.72E-21	924.6196	1402.891	924.6196	1402.891



- 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.

BRAND	MSRP	Average of
Acura		34887.5873
Alfa Romeo		61600
Aston Martin		197910.3763
Audi		53452.1128
Bentley		247169.3243
BMW		61546.76347
Bugatti		1757223.667
Buick		28206.61224
Cadillac		56231.31738
Chevrolet		28273.35695

Chrysler	26722.96257
Dodge	22390.05911
Ferrari	237383.8235
FIAT	22206.01695
Ford	27393.42051
Genesis	46616.66667
GMC	30493.29903
Honda	26629.81879
HUMMER	36464.41176
Hyundai	24597.0363
Infiniti	42394.21212
Kia	25112.38938
Lamborghini	331567.3077
Land Rover	67823.21678
Lexus	47549.06931
Lincoln	42494.37179
Lotus	69188.27586
Maserati	114207.7069
Maybach	546221.875
Mazda	19719.05707
McLaren	239805
Mercedes-Benz	71537.80966
Mitsubishi	21215.47143
Nissan	28513.36679
Oldsmobile	11542.54
Plymouth	3122.902439
Pontiac	19321.54839
Porsche	101622.3971
Rolls-Royce	351130.6452
Saab	27413.5045
Scion	19932.5
Spyker	213323.3333
Subaru	24827.50391
Suzuki	17900.9569
Toyota	28946.15343
Volkswagen	28076.2
Volvo	28541.16014
Grand Total	**40559.93532**

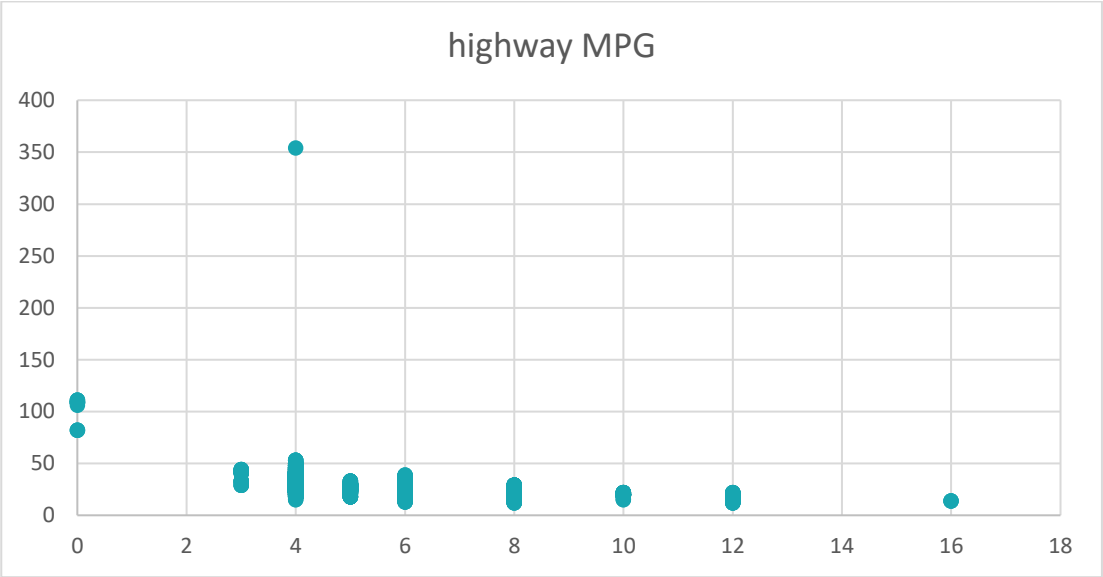


- What is the relationship between fuel efficiency and the number of cylinders in a car's engine?

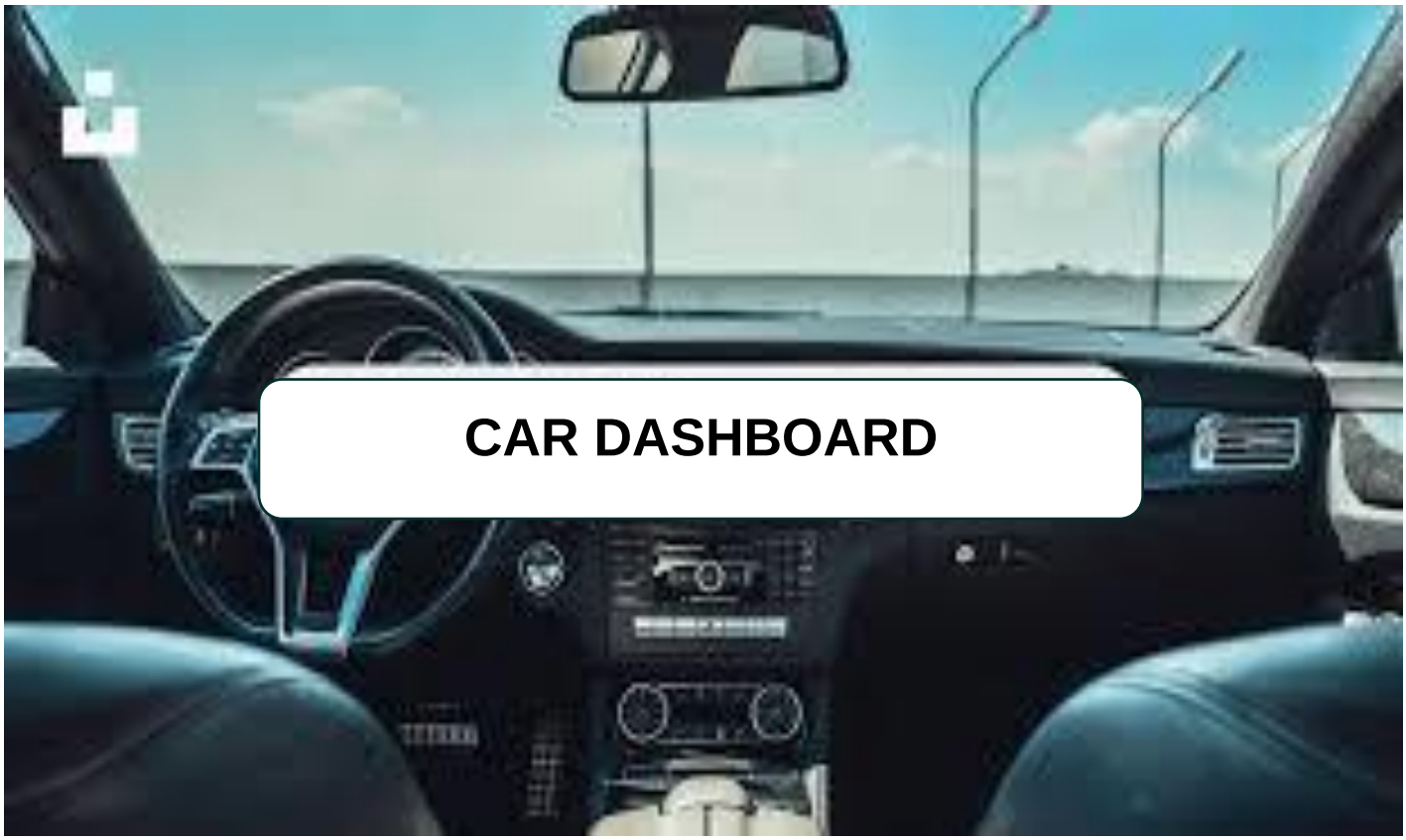
  Task 5.A: Create a scatter plot with the number of cylinders on the x-axis and highway MPG on the y-axis. Then create a trendline on the scatter plot to visually estimate the slope of the relationship and assess its significance.

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

CORRELATION COEFFICIENT:-	**-0.6203**



BUILDING DASHBOARD:-



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

Sum of MSRP BRAND	VEHICL STYLE 2dr Hatchback	2dr SUV	4dr Hatchback	4dr SUV	Cargo Minivan
Acura	480917		357440	2663505	
Alfa Romeo					
Aston Martin					
Audi	4000			2674900	
Bentley					
BMW	80097		1144950	3160950	
Bugatti					
Buick				2141770	
Cadillac				7182555	
Chevrolet	8000	213310	1209735	6569568	420150
Chrysler	98805			250545	
Dodge	48000	44000	18000	2572405	60520
Ferrari					
FIAT	325315			369305	
Ford	36000	479873	480155	4370871	680770
Genesis					
GMC		144319		6641919	142750
Honda	413200		2015270	3953209	
HUMMER				377490	
Hyundai	1038050		528880	2128890	
Infiniti				4340200	
Kia			406960	2049645	
Lamborghini					
Land Rover		476394		9076595	
Lexus			94700	3152974	
Lincoln				3422570	
Lotus					
Maserati				155000	
Maybach					
Mazda	22000	24000	853180	3222525	
McLaren					
Mercedes-Benz			122800	4924810	28950
Mitsubishi	394868		338850	2066505	2000

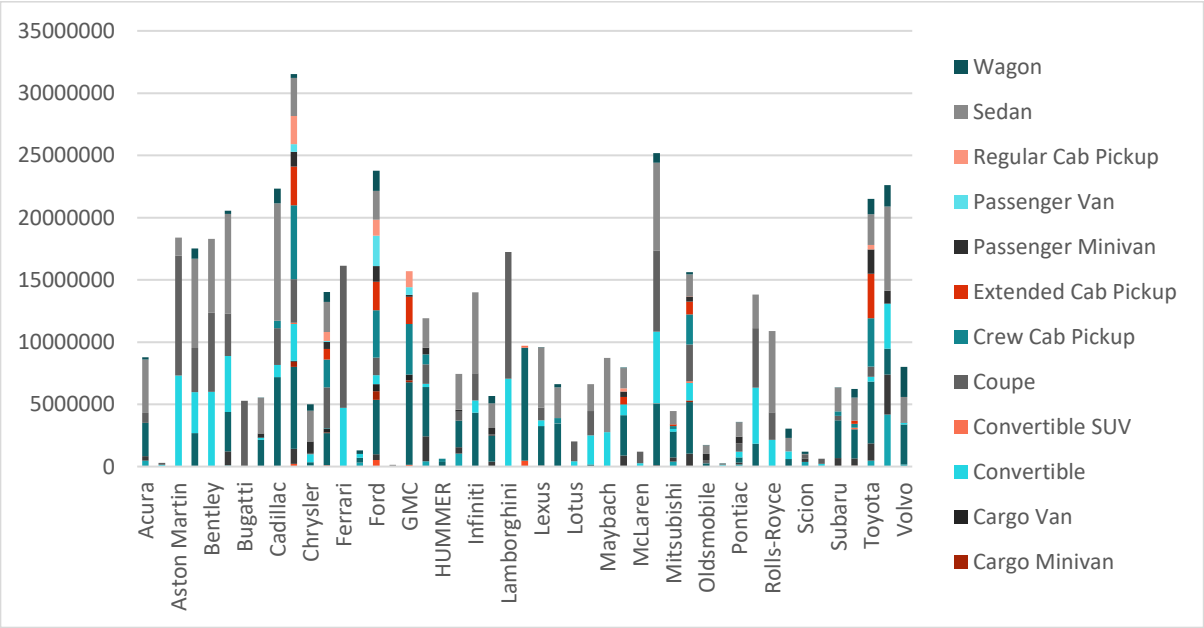
Nissan	14683		1023090	4149630	128620
Oldsmobile				238150	
Plymouth	42000		16000		
Pontiac	163505		162975	401550	
Porsche	28827			1815200	
Rolls-Royce					
Saab	14000		36586	541905	
Scion	366325		282470		
Spyker					
Subaru	12000		678060	3020230	
Suzuki	46496	14000	584387	2362141	
Toyota	473750		1397750	4957050	
Volkswagen	4171275		3222275	2084955	
Volvo	157550			3219000	
Grand Total	**8439663**	**1395896**	**14974513**	**100258517**	**1463760**

Cargo Van	Convertible	Convertible SUV	Coupe	Crew Cab Pickup	Extended Cab Pickup
			793748		
	129800		178200		
	7321655		9635275		
	3291405		3556290		
	6012870		6356760		
	4502671		3419051		
			5271671		
	179325		18534		
	985607		2953574	599150	
78688	2953245	106300	3504525	5927617	3117951
	630105		114510		
338497	12000		3264627	2235775	864172
	4723811		11418289		
	327965				
566351	730007		1398144	3812353	2285584
468085				4062482	2183866
	252135		1588705	787720	
				242405	
			724070		
	980050		2175750		
			142630		
	7064450		10177050		
		145731			
	472065		1016472		
			25342	453260	

	413260		1593200		
	2342963		1972284		
	2762750				
	870505		14000		580033
	280225		918800		
	5753964		6473107		
	209893			240210	134360
	1406552	131075	2943632	2422300	1026379
	2000		286015		
	85631		14000		
	473481		667715		
	4504586		4758533		
	2141365		2204675		
	632628				
			330210		
	219990		419980		
			356476	365975	
		122194		304131	259659
	386668		811995	3893760	3558504
	3612631		8000		
	121600		6000		
1451621	**66789858**	**505300**	**91511839**	**25347138**	**14010508**

Passenger Minivan	Passenger Van	Regular Cab Pickup	Sedan	Wagon	Grand Total
			4294702	201360	8791672
					308000
			1448735		18405665
			7158348	847350	17532293
			5920900		18290530
			7989300	259600	20556619
					5271671
330065			2850590	8212	5528496
			9418847	1184100	22323833
1178515	607670	2260032	3068812	300675	31524793
922295			2479859	501075	4997194
557425	70708	719408	2417585	793055	14016177
					16142100
				287570	1310155
1271330	2431898	1299240	2299348	1635565	23777489
			139850		139850
150630	603670	1306328			15704049
553185			2340105		11903529
					619895

133075			2899937		7452902
			6494090		13990090
494650			1980360	601155	5675400
					17241500
					9698720
			4837596	31105	9604912
			2458245	269705	6629122
					2006460
			2153800		6624047
			5976800		8739550
443130		265486	1618571	33350	7946780
					1199025
32500			7080243	764935	25181309
2000		8000	1058563		4455249
413320		21914	1769130	175000	15625325
492055			691161	22000	1731381
33688			46759	18000	256078
541192			1160535	22855	3593808
			2713500		13820646
			6539010		10885050
			1066500	751280	3042899
			32500	184445	1195950
					639970
			1913100	10000	6355841
			1850818	685707	6229533
1956518		373446	2459596	1237955	21506992
1038130			6760050	1704025	22601341
			2086945	2428971	8020066
10543703	**3713946**	**6253854**	**117474790**	**14959050**	**479093956**



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

BRAND	2dr Hatchback	2dr SUV	4dr Hatchback	4dr SUV	Cargo Minivan
Acura	17175.60714		51062.85714	42959.75806	
Alfa Romeo					
Aston Martin					
Audi	2000			48634.54545	
Bentley					
BMW	26699		54521.42857	58536.11111	
Bugatti					
Buick				33996.34921	
Cadillac				72551.06061	
Chevrolet	2000	8887.916667	18329.31818	32046.67317	20007.14286
Chrysler	32935			35792.14286	
Dodge	2000	2000	2000	30992.83133	20173.33333
Ferrari					
FIAT	19136.17647			24620.33333	
Ford	2000	13710.65714	18467.5	42027.60577	21274.0625
Genesis					
GMC		5550.730769		36695.68508	23791.66667
Honda	17216.66667		25836.79487	28855.54015	
HUMMER				37749	
Hyundai	18536.60714		17629.33333	30412.71429	
Infiniti				45686.31579	
Kia			19379.04762	31533	
Lamborghini					
Land Rover		39699.5		70910.89844	
Lexus			31566.66667	45042.48571	
Lincoln				50331.91176	
Lotus					
Maserati				77500	
Maybach					
Mazda	2000	2000	20809.26829	27080.04202	
McLaren					
Mercedes-Benz			40933.33333	68400.13889	28950
Mitsubishi	13162.26667		12101.78571	26158.29114	2000
Nissan	2097.571429		22241.08696	34294.46281	21436.66667

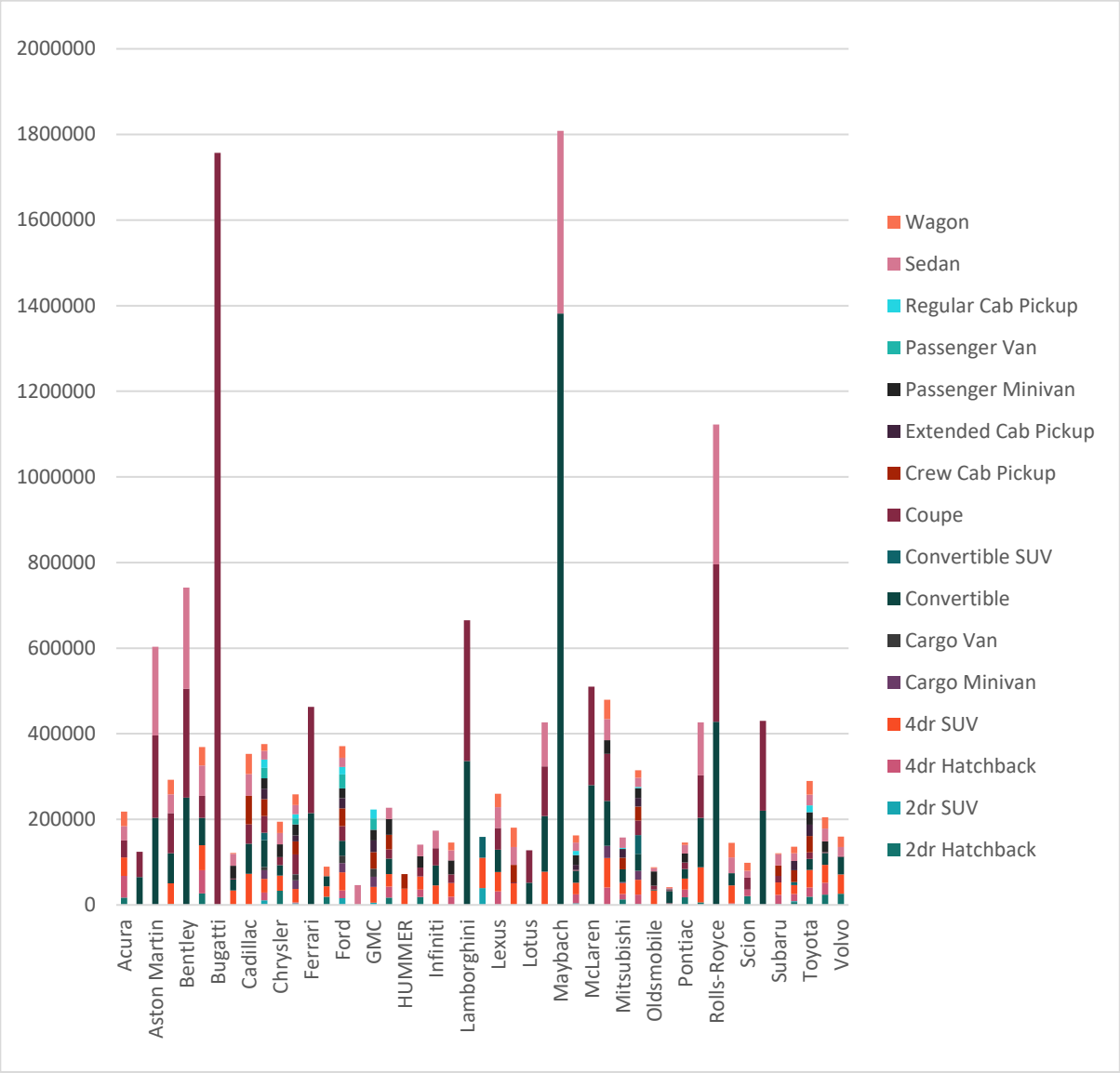
Oldsmobile				34021.42857	
Plymouth	2000		2000		
Pontiac	18167.22222		18108.33333	25096.875	
Porsche	5765.4			82509.09091	
Rolls-Royce					
Saab	2000		2032.555556	41685	
Scion	20351.38889		15692.77778		
Spyker					
Subaru	2000		21189.375	29322.62136	
Suzuki	6642.285714	2000	16696.77143	21090.54464	
Toyota	18950		22186.50794	40631.55738	
Volkswagen	24251.59884		27778.23276	41699.1	
Volvo	26258.33333			45338.02817	
Grand Total	**16778.65408**	**10115.18841**	**22086.30236**	**40426.82137**	**20910.85714**

Cargo Van	Convertible	Convertible SUV	Coupe	Crew Cab Pickup	Extended Cab Pickup
			39687.4		
	64900		59400		
	203379.3056		192705.5		
	70029.89362		93586.57895		
	250536.25		254270.4		
	63417.90141		51803.80303		
			1757223.667		
	25617.85714		2059.333333		
	70400.5		45439.6	66572.22222	
7153.454545	62835	17716.66667	38939.16667	39255.74172	24170.16279
	24234.80769		19085		
12536.92593	2000		45980.66197	31052.43056	13938.25806
	214718.6818		248223.6739		
	23426.07143				
17698.46875	34762.2381		34101.07317	41438.61957	23808.16667
18723.4				39062.32692	26632.5122
	36019.28571		21763.08219	34248.69565	
				34629.28571	
			20687.71429		
	46669.04762		40291.66667		
			20375.71429		
	336402.381		328291.9355		
		48577			
	52451.66667		50823.6		
			2111.833333	41205.45455	
	51657.5		75866.66667		

130164.6111		116016.7059			
1381375					
28080.80645		2000			11600.66
280225		229700			
104617.5273		109713.678			
29984.71429			26690		19194.28571
39070.88889	43691.66667	34228.27907	32733.78378		20527.58
2000		9226.290323			
28543.66667		2000			
22546.71429		15528.25581			
115502.2051		99136.10417			
428273		367445.8333			
28755.81818					
		27517.5			
219990		209990			
		15498.95652	24398.33333		
	7187.882353		27648.27273		21638.25
25777.86667		15615.28846	37803.49515		26359.28889
27789.46923		2000			
40533.33333		2000			
15280.22105	**84224.28499**	**17424.13793**	**76900.70504**	**37220.46696**	**22488.77689**

Passenger Minivan	Passenger Van	Regular Cab Pickup	Sedan	Wagon	Grand Total
			33292.26357	33560	34887.5873
					61600
			206962.1429		197910.3763
			44461.78882	33894	53452.1128
			236836		247169.3243
			70701.76991	43266.66667	61546.76347
					1757223.667
30005.90909			27946.96078	2053	28206.61224
			50912.68649	47364	56231.31738
24552.39583	24306.8	19824.84211	19798.7871	15825	28273.35695
29751.45161			26103.77895	26372.36842	26722.96257
25337.5	14141.6	9342.961039	21780.04505	24782.96875	22390.05911
					237383.8235
				22120.76923	22206.01695
23115.09091	32425.30667	17797.80822	21290.25926	27259.41667	27393.42051
			46616.66667		46616.66667
25105	26246.52174	21069.80645			30493.29903
36879			26001.16667		26629.81879
					36464.41176
26615			27102.21495		24597.0363

			40588.0625		42394.21212
32976.66667			23298.35294	18216.81818	25112.38938
					331567.3077
					67823.21678
			48864.60606	31105	47549.06931
			41665.16949	44950.83333	42494.37179
					69188.27586
			102561.9048		114207.7069
			426914.2857		546221.875
23322.63158		9154.689655	19738.67073	16675	19719.05707
					239805
32500			49168.35417	44996.17647	71537.80966
2000		2000	24058.25		21215.47143
22962.22222		2191.4	21841.11111	17500	28513.36679
32803.66667			8131.305882	2000	11542.54
2105.5			2597.722222	2000	3122.902439
20815.07692			20009.22414	5713.75	19321.54839
			123340.9091		101622.3971
			326950.5		351130.6452
			36775.86207	34149.09091	27413.5045
			16250	18444.5	19932.5
					213323.3333
			26570.83333	2000	24827.50391
			18145.27451	15237.93333	17900.9569
29201.76119		16236.78261	24844.40404	31742.4359	28946.15343
25320.2439			29911.72566	25818.56061	28076.2
			20869.45	24785.41837	28541.16014
25591.51214	**29015.20313**	**15953.70918**	**38989.30966**	**25483.90119**	**40559.93532**

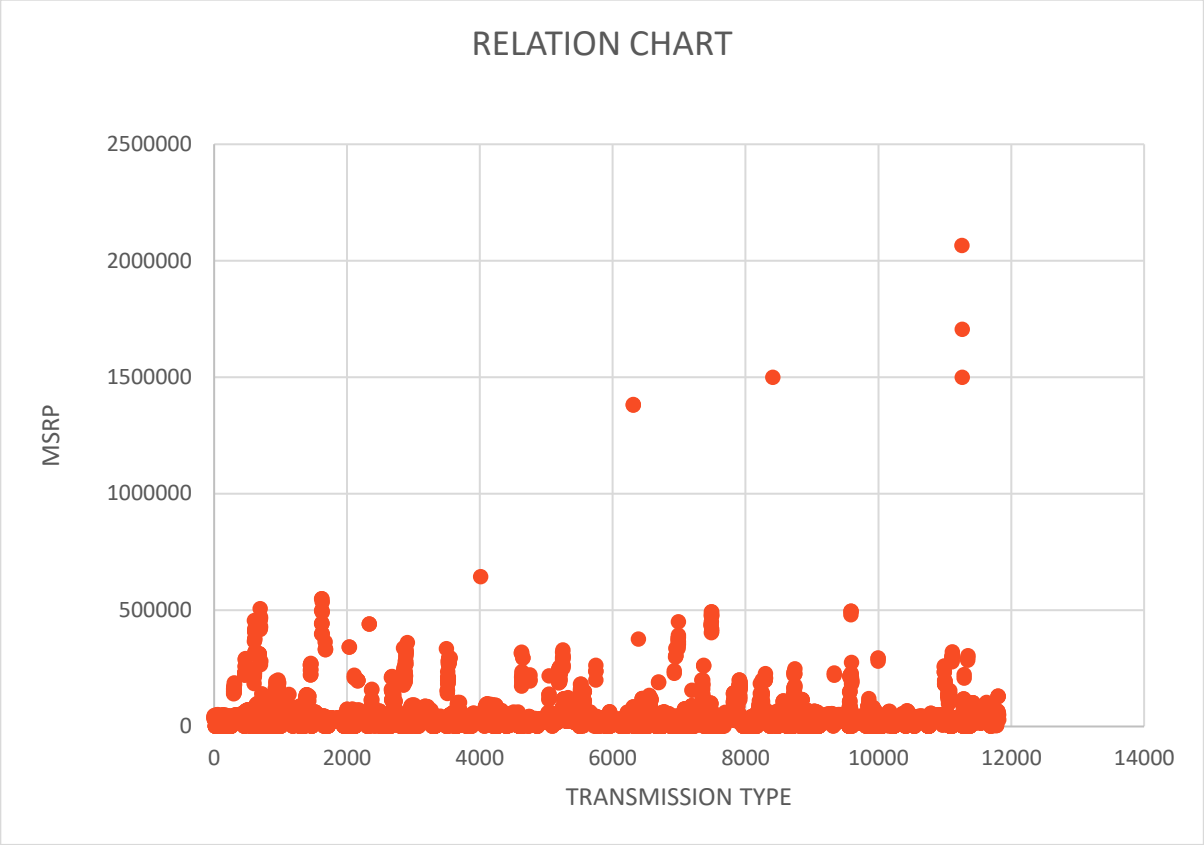

2000000
1800000
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
Acura
Aston Martin
Bentley
Bugatti
Cadillac
Chrysler
Ferrari
Ford
GMC
HUMMER
Infiniti
Lamborghini
Lexus
Lotus
Maybach
McLaren
Mitsubishi
Oldsmobile
Pontiac
Rolls-Royce
Scion
Subaru
Toyota
Volvo
Wagon
Sedan
Regular Cab Pickup
Passenger Van
Passenger Minivan
Extended Cab Pickup
Crew Cab Pickup
Coupe
Convertible SUV
Convertible
Cargo Van
Cargo Minivan
4dr SUV
4dr Hatchback
2dr SUV
2dr Hatchback

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

Average of MSRP TRANSMISSION TYPE	VEHICLE STYLE 2dr Hatchback	2dr SUV	4dr Hatchback	4dr SUV	Cargo Minivan
AUTOMATED_MANUAL	27180.96491		29249.07407	40451.15385	
AUTOMATIC	20926.464	18615.20455	23833.67898	41555.18825	20910.85714
DIRECT_DRIVE			34511.92308		
MANUAL	13353.65831	6303.811111	17594.41313	15426.46226	
UNKNOWN	7361.5	2371			
Grand Total	**16778.65408**	**10115.18841**	**22086.30236**	**40426.82137**	**20910.85714**

Cargo Van	Convertible	Convertible SUV	Coupe	Crew Cab Pickup	Extended Cab Pickup
	121256.6444		245588.3571		
15280.22105	90637.3869	38925.5	63852.00808	37744.07154	30637.34973
	62357.75625	9233.142857	51070.47972	28360.52632	10884.19455
	5783.5		2000		
15280.22105	**84224.28499**	**17424.13793**	**76900.70504**	**37220.46696**	**22488.77689**

Passenger Minivan	Passenger Van	Regular Cab Pickup	Sedan	Wagon	Grand Total
			47498.70813	31985.27778	99195.584
26391.99748	29015.20313	28536.8239	43794.38665	27613.19169	41137.264
			27822.5		33620
4405.333333		7557.773333	17119.23374	17844.13971	26671.39699
		2000	2000		3040.736842
25591.51214	**29015.20313**	**15953.70918**	**38989.30966**	**25483.90119**	**40559.93532**



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

Sum of highway MPG YEAR VEHICLE STYLE	2dr Hatchback	2dr SUV	4dr Hatchback	4dr SUV	Cargo Minivan
1990	152	280	62		20
1991	451	195		116	
1992	980	297	227	64	
1993	856	351	273	126	
1994	547	258	190	120	21
1995	211	64	83		43
1996	174	100	209	108	46
1997	235	66	265	197	21
1998	116	78	49	199	
1999	273	75		183	
2000	365	75		266	
2001	406	112		412	22
2002	101	152		673	21
2003	119	225		673	124
2004	416	75	68	933	98
2005	273	56	153	812	124
2006	327		345	916	69
2007	332		358	1210	68
2008	601		228	1414	23
2009	232		308	2101	
2010	361		472	1322	
2011	167		839	1698	
2012	707		1766	999	
2013	607		2131	1321	
2014	937		1630	2857	
2015	3018	90	5307	14093	336
2016	2102	60	4574	15511	360
2017	496	29	4374	12247	320
Grand Total	**15562**	**2638**	**23911**	**60571**	**1716**

Cargo Van	Convertible	Convertible SUV	Coupe	Crew Cab Pickup	Extended Cab Pickup
	94		343		132
	181		497		190
	306		382		156
	318	52	740		234
116	104	26	632		142
95	98	26	773		220
131	119	48	882		160
137	177	62	789		514
172	71	72	394		596
50	43		248		479
82	177		145		164
79	375		345		133
73	337	163	472	51	364
75	263	117	788	36	374
	201		758	132	142
	228		624	115	
	160		655	252	
	569		756	541	1085
	1058		942	849	384
	998		769	705	477
	517		429	739	395
	431		523	422	219
100	613	22	817	643	369
100	510	22	1095	469	
236	2345	22	1809	283	87
68	4420		5035	2972	1971
64	3624		4955	3116	1940
	2113	56	4077	3015	1618
1578	**20450**	**688**	**30674**	**14340**	**12545**

Passenger Minivan	Passenger Van	Regular Cab Pickup	Sedan	Wagon	Grand Total
264		289	840	362	2838
72		356	993	316	3367

		304	1177	364	4257
		300	1469	318	5037
21	164	130	1110	286	3867
241	60	212	746	241	3113
187	45	111	566	222	3108
185	51	526	557	122	3904
234	51	632	678	23	3365
134		258	1014		2757
139	29	125	1208	62	2837
212	30	161	1451	245	3983
347	15	331	1307	260	4667
825		289	1407	96	5411
444		240	1744	114	5365
483		18	1571	437	4894
495		18	990	575	4802
410		509	987	744	7569
161		162	1336	865	8023
		306	2104	886	8886
242		231	1720	655	7083
175		54	2104	431	7063
175	92	193	2514	1161	10171
112	92		2261	1243	9963
312	262		3377	1410	15567
1418	508	796	19311	2010	61353
1410	496	766	20218	1736	60932
1019	304	766	15145	1142	46721
9717	**2199**	**8083**	**89905**	**16326**	**310903**

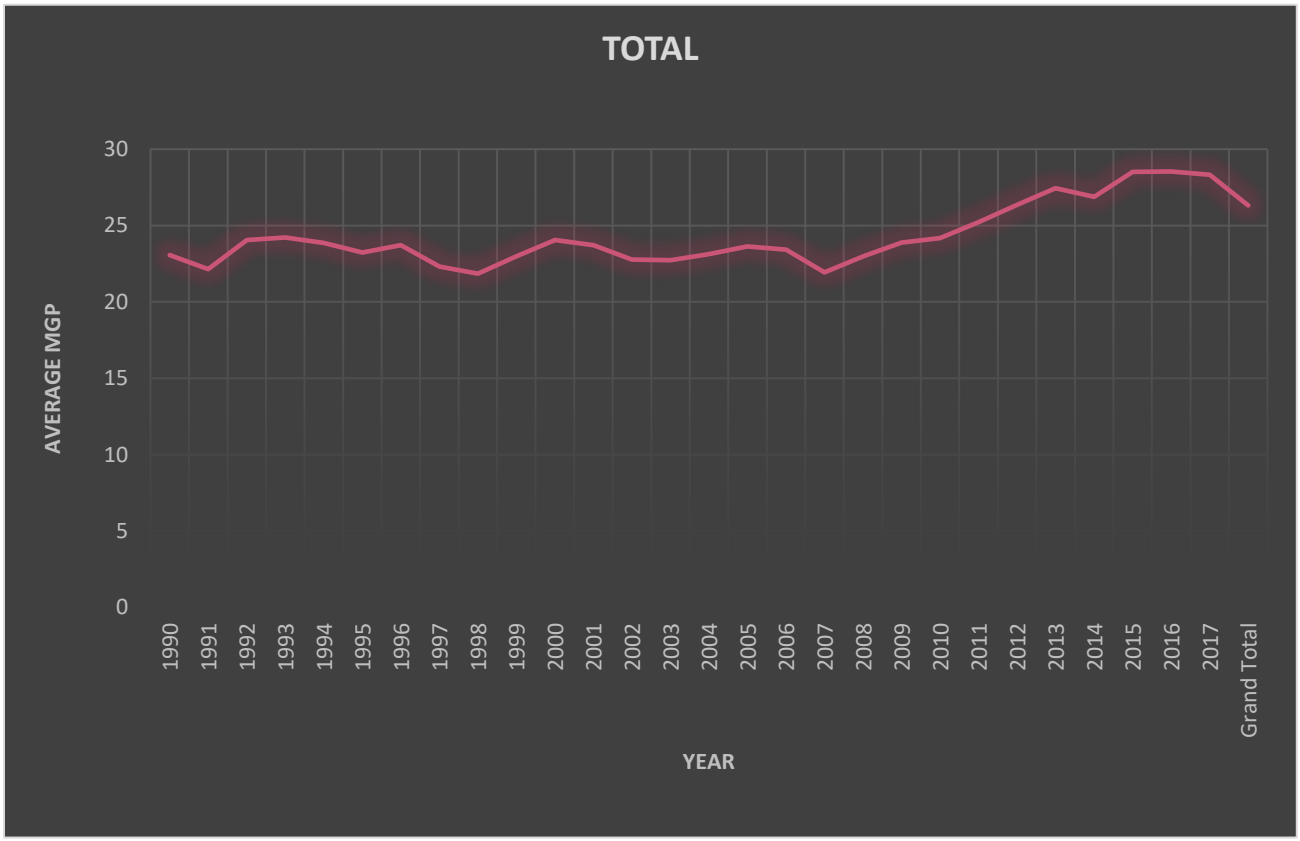
YEAR	highway MPG	Average of
1990		23.07317073
1991		22.15131579
1992		24.05084746
1993		24.21634615
1994		23.87037037
1995		23.23134328
1996		23.72519084
1997		22.30857143
1998		21.85064935
1999		22.975
2000		24.04237288

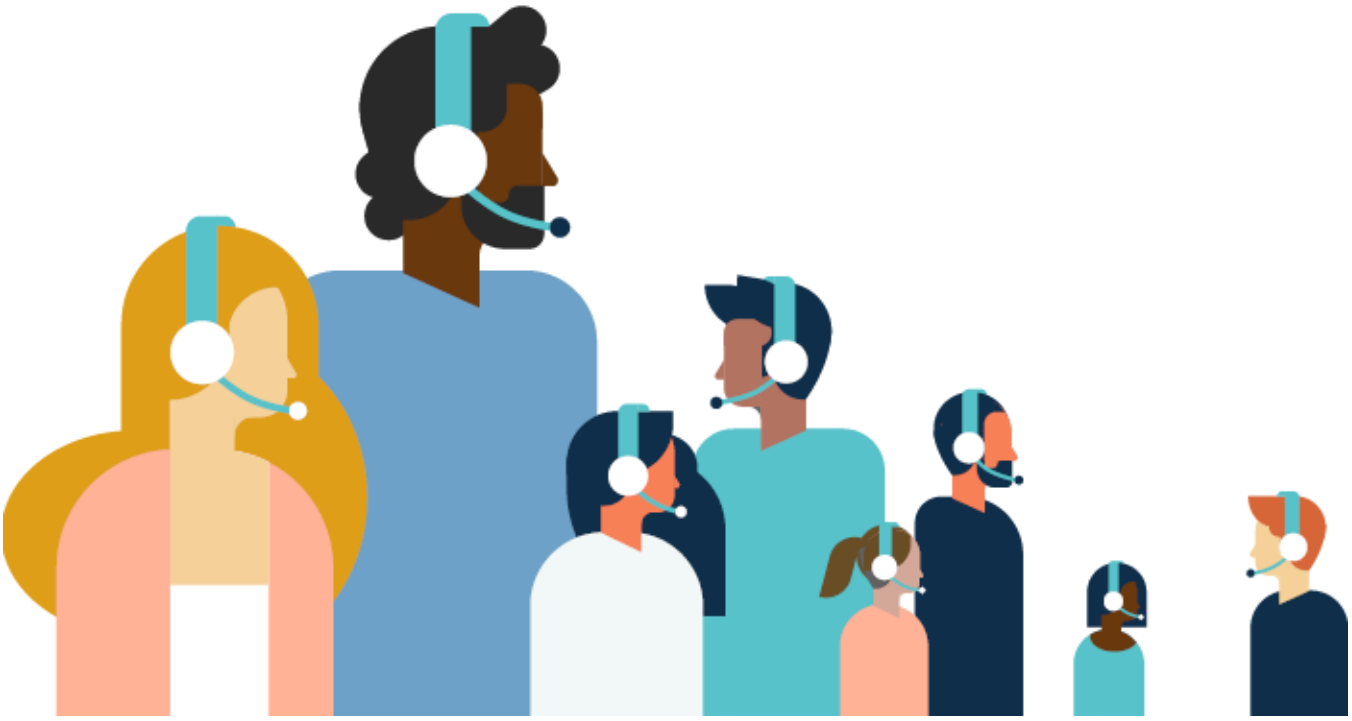
2001	23.70833333
2002	22.76585366
2003	22.73529412
2004	23.125
2005	23.64251208
2006	23.42439024
2007	21.93913043
2008	22.98853868
2009	23.88709677
2010	24.17406143
2011	25.225
2012	26.34974093
2013	27.44628099
2014	26.88601036
2015	28.50975836
2016	28.53957845
2017	28.33292905
Grand Total	**26.3209448**

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

Row Labels	Average of Engine HP	Average of highway MPG	Average of MSRP
Acura	244.797619	28.11111111	34887.5873
Alfa Romeo	237	34	61600
Aston Martin	484.3225806	18.89247312	197910.3763
Audi	277.695122	28.82317073	53452.1128
Bentley	533.8513514	18.90540541	247169.3243
BMW	326.9071856	29.24550898	61546.76347
Bugatti	1001	14	1757223.667
Buick	219.244898	26.94897959	28206.61224
Cadillac	332.3098237	25.23677582	56231.31738
Chevrolet	247.0565022	25.6690583	28273.35695
Chrysler	229.1390374	26.36898396	26722.96257
Dodge	244.4153355	22.34504792	22390.05911
Ferrari	509.9117647	15.72058824	237383.8235
FIAT	143.559322	33.91525424	22206.01695
Ford	243.0979263	23.74078341	27393.42051
Genesis	347.3333333	25.33333333	46616.66667

GMC	259.8446602	21.4038835	30493.29903
Honda	195.7494407	32.25055928	26629.81879
HUMMER	261.2352941	17.29411765	36464.41176
Hyundai	201.9174917	30.39273927	24597.0363
Infiniti	310.0666667	24.77878788	42394.21212
Kia	206.8274336	29.29646018	25112.38938
Lamborghini	614.0769231	18.01923077	331567.3077
Land Rover	322.0979021	22.12587413	67823.21678
Lexus	277.4158416	25.87623762	47549.06931
Lincoln	284.9102564	24.1025641	42494.37179
Lotus	275.9655172	26.55172414	69188.27586
Maserati	420.7931034	20.29310345	114207.7069
Maybach	590.5	16	546221.875
Mazda	169.191067	28.11662531	19719.05707
McLaren	610.4	22.2	239805
Mercedes-Benz	350.1818182	24.81818182	71537.80966
Mitsubishi	174.452381	26.50952381	21215.47143
Nissan	239.9215328	26.46350365	28513.36679
Oldsmobile	177.4666667	26.23333333	11542.54
Plymouth	131.5609756	27.96341463	3122.902439
Pontiac	190.2956989	27.06989247	19321.54839
Porsche	392.7941176	25.36764706	101622.3971
Rolls-Royce	487.5483871	19.12903226	351130.6452
Saab	220.5225225	26.35135135	27413.5045
Scion	154.4333333	32.3	19932.5
Spyker	400	18	213323.3333
Subaru	197.3085938	28.68359375	24827.50391
Suzuki	160.3333333	26.04310345	17900.9569
Toyota	236.2584118	26.26110363	28946.15343
Volkswagen	190.1291925	31.76645963	28076.2
Volvo	230.9715302	27.20284698	28541.16014
Grand Total	**249.504487**	**26.3209448**	**40559.93532**


TOTAL
AVERAGE MGP
30
25
20
15
10
5
0
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Grand Total
YEAR



Project 8

ABC Call Volume Trend Analysis

INTRODUCTION:-

PROJECT DESCRIPTION

- Analyzing the call volume trend for an organization, such as "ABC," is a valuable exercise to understand the dynamics of incoming and outgoing calls over a specific period. This analysis can provide insights into customer behavior, operational efficiency, and help in making data-driven decisions.

- In this project, we will be diving into the world of Customer Experience (CX) analytics, specifically focusing on the inbound calling team of a company. we will 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).

PROCESS APPROACH

- My approach towards this project "ABC CALL VOLUME TREND" is perform the analysis on the provided dataset and perform different task using Excel functions and concepts. diving into the world of Customer Experience analytics, specifically focusing on the inbound calling team of a company.

- 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.

TECH-STACK USED:

- For this project I have used the Microsoft Excel 2019.
- I used this application because Microsoft Excel 2019contains many tools which format, organize and calculate data in a spreadsheet.

- Microsoft Excel 2019 represent the data in from of different charts like bar graph, pie chart, histogram graph., etc.
- Microsoft Excel is a widely used spreadsheet application developed by Microsoft. It is part of the Microsoft Office suite of productivity software. Excel is known for its powerful data manipulation and analysis features and is used for a wide range of tasks, including data entry, calculation, analysis, and visualization.

INSIGHTS:

- Data Cleaning: Clean and preprocess the data to ensure accuracy. Remove duplicates, handle missing values, and standardize the data format.

- Time Period Analysis: here we will be Analyzing call volume trends over different time periods, such as daily, weekly, monthly, or seasonally. Look for patterns, seasonality, and any significant changes.

- Comparison Analysis: Compare call volume trends with other relevant data, such as marketing campaigns, customer feedback, or operational changes, to find correlations between them

- Data Collection: we gathered historical call data for the desired time frame. The data should include information on the number of calls received, calls made ,call durations, and timestamps for each call.

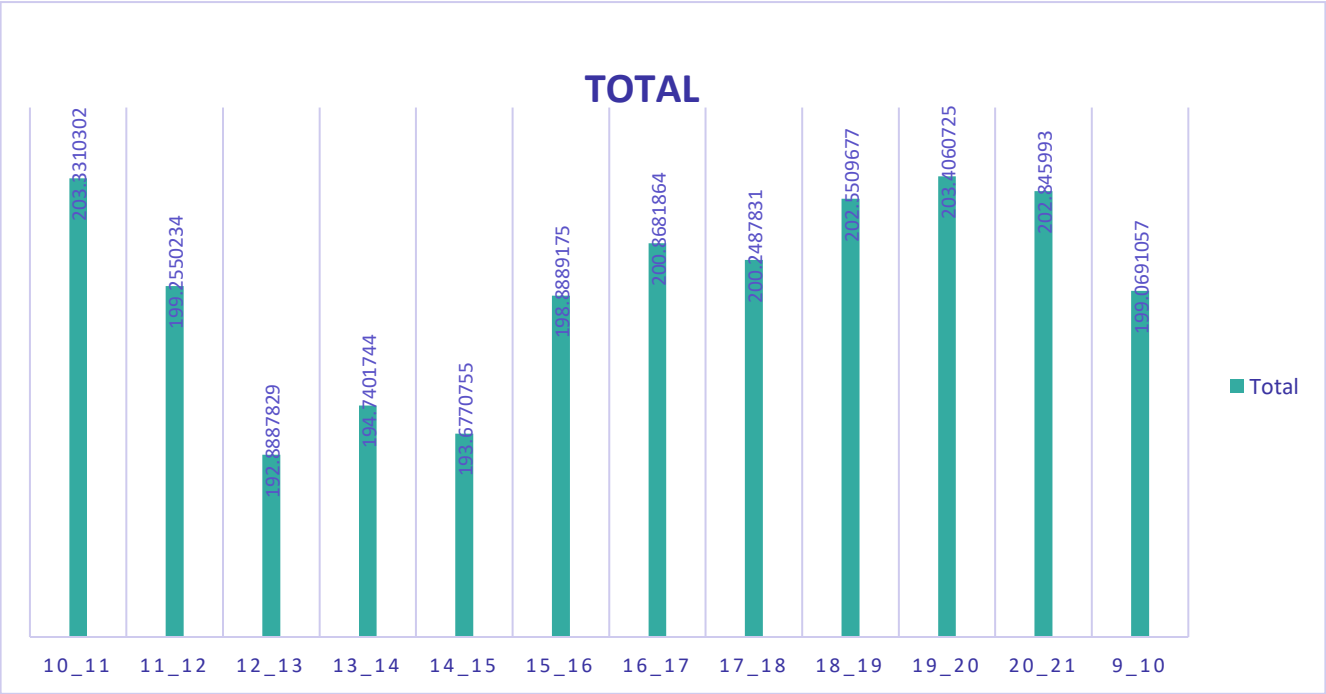
- ☐ Forecasting: Use historical data to build forecasts for future call volumes. This can help in resource allocation and planning.

- ☐ Actionable Insights: Based on analysis we have performed, actionable insights can be given as, For example, if you find that call volumes tend to spike on certain days, consider increasing staff during those times to reduce wait times.

Data Analytics Tasks:

Task 1 : Average Call Duration: Determine the average duration of all incoming calls received by agents. This should be calculated for each time bucket.
Your Task: What is the average duration of calls for each time bucket?

Call_Status	answered
Row Labels	Average of Call_Seconds (s)
10_11	203.3310302
11_12	199.2550234
12_13	192.8887829
13_14	194.7401744
14_15	193.6770755
15_16	198.8889175
16_17	200.8681864
17_18	200.2487831
18_19	202.5509677
19_20	203.4060725
20_21	202.845993
9_10	199.0691057
Grand Total	198.6227745



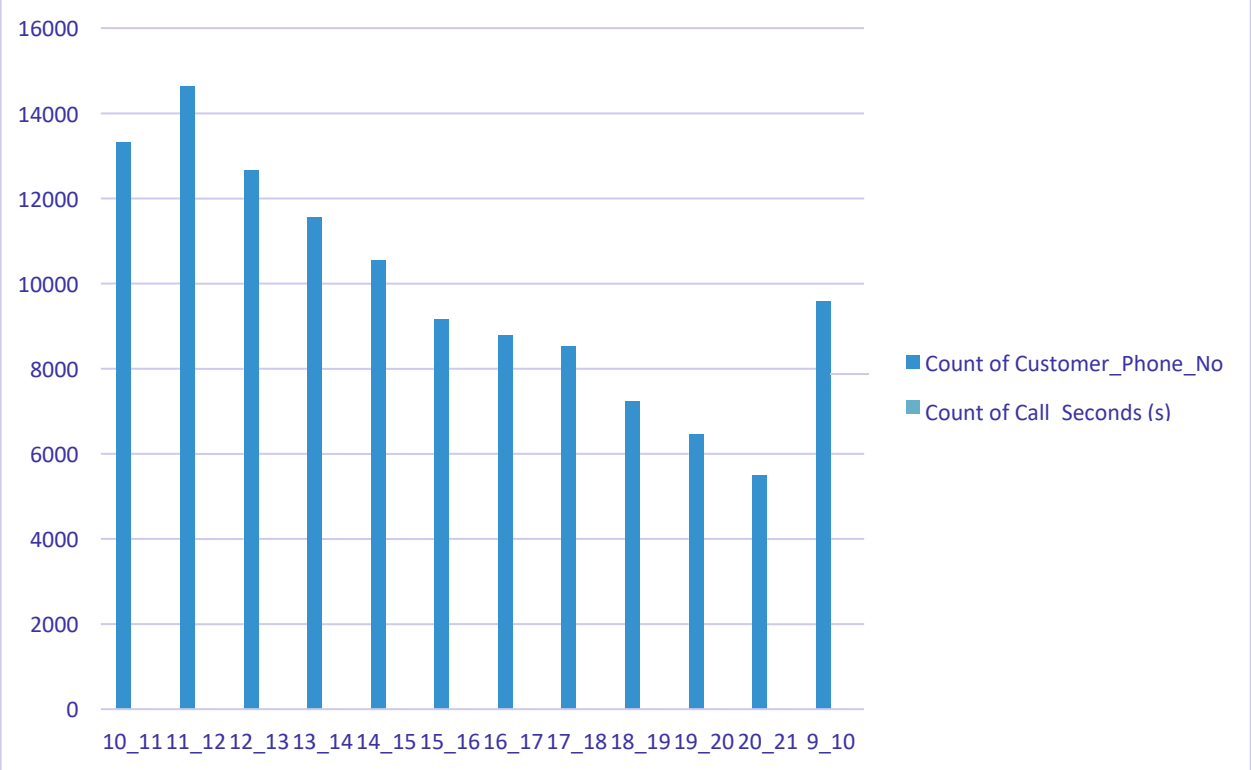
Task 2: Call Volume Analysis: Visualize the total number of calls received. This should be represented as a graph or chart showing the number of calls against time. Time should be represented in buckets (e.g., 1-2, 2-3, etc.).

Your Task: Can you create a chart or graph that shows the number of calls received in each time bucket?

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%
9_10	9588	8.13%
Grand Total	117988	100.00%

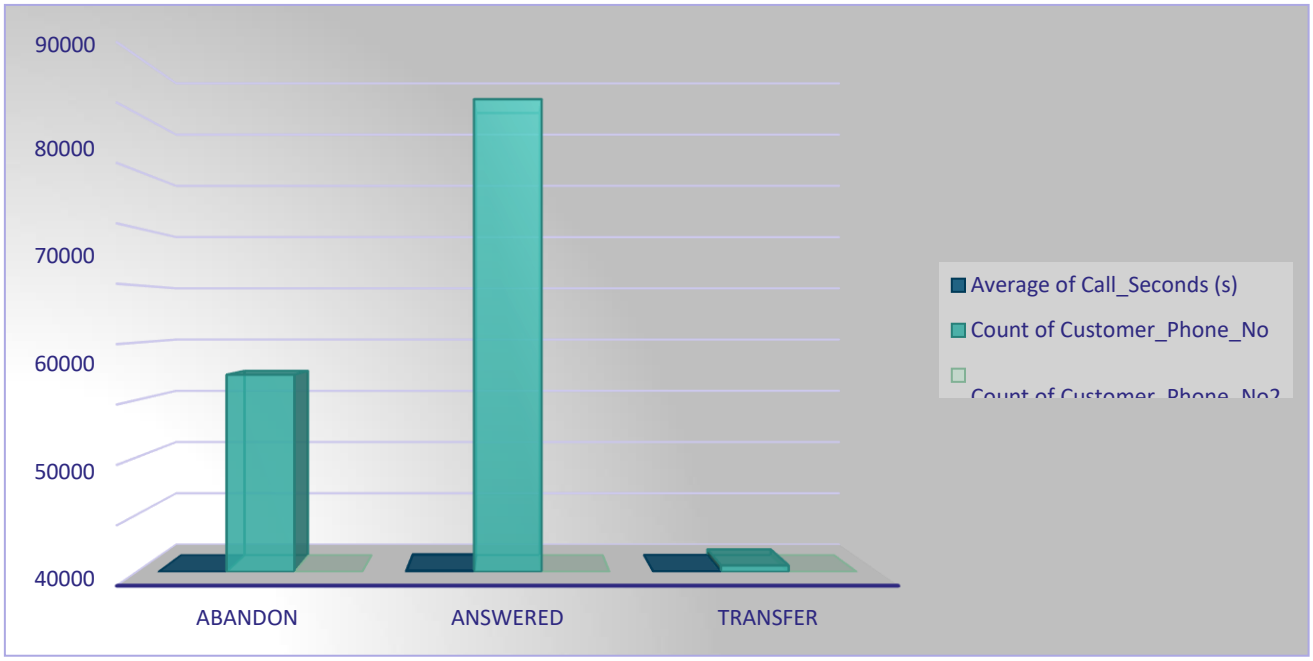
Row Labels	Count of Customer_Phone_No	Count of Time
10_11	13313	11%
11_12	14626	12%
12_13	12652	11%
13_14	11561	10%
14_15	10561	9%
15_16	9159	8%
16_17	8788	7%
17_18	8534	7%
18_19	7238	6%
19_20	6463	5%
20_21	5505	5%
9_10	9588	8%



Task 3: Manpower Planning: The current rate of abandoned calls is approximately 30%. Propose a plan for manpower allocation during each time bucket (from 9 am to 9 pm) to reduce the abandon rate to 10%. In other words, you need to calculate the minimum number of agents required in each time bucket to ensure that at least 90 out of 100 calls are answered.

Your Task: What is the minimum number of agents required in each time bucket to reduce the abandon rate to 10%?

Call_Status	Average of Call_Seconds (s)	Count of Customer_Phone_No	Count of Customer_Phone_No 2
abandon	0	34403	29.16%
answered	198.6227745	82452	69.88%
transfer	76.14651368	1133	0.96%
Grand Total	139.5321473	117988	100.00%



Row Labels	Sum of Call_Seconds (s)
09 AM	35313
10 AM	53087
11 AM	67751
12 PM	72680
01 PM	59693
02 PM	76137
03 PM	65689
04 PM	59464
05 PM	68155
06 PM	53096
07 PM	40141
08 PM	25281
09 PM	177
Grand Total	676664

01-Jan	sum of all call second	sum of hour
	676664	187.96
	total agent for 60%	37.6

	agent required for 90%	56

Row Labels	Count of Call_Seconds (s)	Count of Call_Seconds (s)2	Agent required
10_11	11.28%	11.28%	6
11_12	12.40%	12.40%	7
12_13	10.72%	10.72%	6
13_14	9.80%	9.80%	5
14_15	8.95%	8.95%	5
15_16	7.76%	7.76%	4
16_17	7.45%	7.45%	4
17_18	7.23%	7.23%	4
18_19	6.13%	6.13%	3
19_20	5.48%	5.48%	3
20_21	4.67%	4.67%	3
9_10	8.13%	8.13%	5
Grand Total	100.00%	100.00%	56

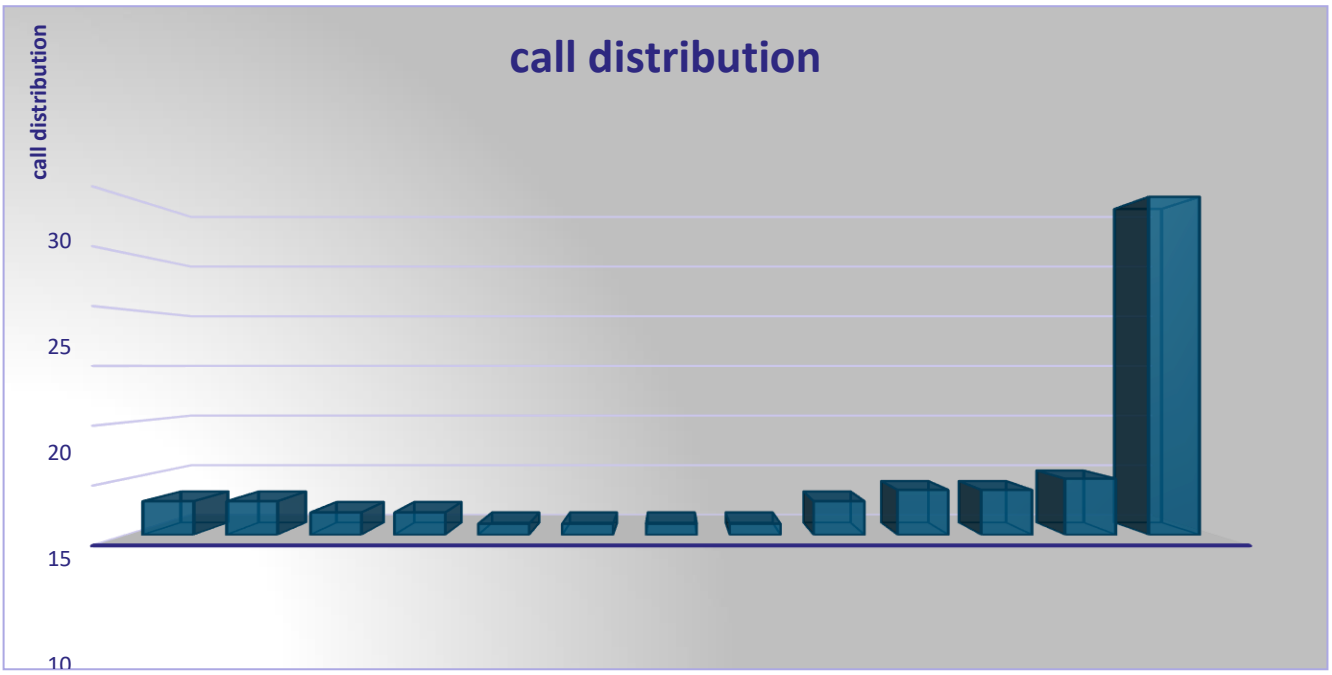
Task 4: Night Shift Manpower Planning: Customers also call ABC Insurance

Company at night but don't get an answer because there are no agents available. This creates a poor customer experience. Assume that for every 100 calls that customers make between 9 am and 9 pm, they also make 30 calls at night between 9 pm and 9 am. The distribution of these 30 calls is as follows:

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

9pm - 9am	call distribution	time distribution	agent req	Column2
9_10	3	10.00	1.50	2
10_11	3	10.00	1.50	2
11_12	2	15.00	1.00	1

12_1	2	15.00	1.00	1
1_2	1	30.00	0.50	1
2_3	1	30.00	0.50	1
3_4	1	30.00	0.50	1
4_5	1	30.00	0.50	1
5_6	3	10.00	1.50	2
6_7	4	7.50	2.00	2
7_8	4	7.50	2.00	2
8_9	5	6.00	2.50	3
TOTAL	30	1.00	15.00	15



Count of	Call Status				
date	abandon	answered	transfer	(blank)	Grand Total
01-01-2022	684	3883	77	4644	
02-01-2022	356	2935	60	3351	
03-01-2022	599	4079	111	4789	
04-01-2022	595	4404	114	5113	
05-01-2022	536	4140	114	4790	
06-01-2022	991	3875	85	4951	
07-01-2022	1319	3587	42	4948	

08-01-2022	1103	3519	50	4672
09-01-2022	962	2628	62	3652
10-01-2022	1212	3699	72	4983
11-01-2022	856	3695	86	4637
12-01-2022	1299	3297	47	4643
13-01-2022	738	3326	59	4123
14-01-2022	291	2832	32	3155
15-01-2022	304	2730	24	3058
16-01-2022	1191	3910	41	5142
17-01-2022	16636	5706	5	22347
18-01-2022	1738	4024	12	5774
19-01-2022	974	3717	12	4703
20-01-2022	833	3485	4	4322
21-01-2022	566	3104	5	3675
22-01-2022	239	3045	7	3291
23-01-2022	381	2832	12	3225
(blank)				
Grand Total	34403 5129.913	82452	1133	117988

Average call daily for night 09am-9pm	5130 1539
additional hour req	76
additional agent req	15

Learning from these projects

• **Machine Learning Modelling:** Machine learning models are used in many data analysis projects to create predictions or categorize data. This includes choosing appropriate algorithms, training models, and assessing their performance. Data visualization techniques are used to show the study' findings in a clear and comprehensible manner. Creating charts, graphs, and interactive dashboards is one example.

• **Gaining Insights and Making Data-Driven Recommendations**: The ultimate purpose of data analysis initiatives is to gather insights and generate data-driven recommendations. These insights may help businesses make more informed decisions, optimize operations, and improve performance.

• **Continuous Improvement:** Data analysis initiatives are frequently iterative, with findings and suggestions being updated in response to feedback and new data. This contributes to the precision and efficacy of the analysis over time.

• **Data Cleaning and Preprocessing:** Cleaning and preprocessing data is one of the first processes in data analysis projects. This includes resolving missing numbers, eliminating outliers, and converting the data into an analysis-ready format.

• **EDA (Exploratory Data Analysis):** EDA is the process of evaluating and visualizing data in order to obtain insights and comprehend its features. It aids in the discovery of patterns, trends, and correlations between variables. Feature engineering is the process of producing new features or changing existing features in order to improve the performance of machine learning models. This procedure may involve feature scaling, one-time encoding, and the creation of interaction variables.

• **Statistical Analysis:** Techniques of statistical analysis are used to assess data and make relevant findings. Hypothesis testing, regression analysis, and correlation analysis are all examples of this.

CONCLUSION:

Here I conclude that, I would like to tell that after doing a thorough analysis we were able to derive the insights . from the data and was able to plot various graphs using that data. The ultimate purpose of data analysis initiatives is to gather insights and generate data-driven recommendations. These insights may help businesses make more informed decisions, optimize operations, and improve performance. The data that once looked useless became useful and helped to find out the courses that were a burden for Trainity to continue providing. Analysing the data proved helpful in finding various issues among the courses.