Data Analysis Portfolio

By Ritika Chaudhary

Table Of Contents

+91-7303314962
ameliarc2105@gmail.com

RITIKA CHAUDHARY

Introduction

I am Ritika Chaudhary, a dedicated second year B Tech student pursuing computer science and engineering with specialization in artificial intelligence and machine learning.

Education

I am currently enrolled at Manav Rachna International Institute of research and studies, working towards The Bachelor of Technology (BTech) degree in computer science and engineering, with an expected graduation in 2026. During my academic journey, I have consistently maintained a strong academic record.

Skills

My technical skills include proficiency in:

- C language
- Basic knowledge of HTML
- Basic proficiency in C++
- Proficiency in probability and statistics

Research And Projects

One of my significant achievements includes working on a case study that was published as “RENEWABLE ENERGY RESOURCES OPTIMIZATION USING ARTIFICIAL INTELLIGENCE”. In this project, I collaborated with a team to develop and implement ai driven solutions for optimizing renewable energy sources. Our research findings have the potential to contribute to sustainable energy practices.

Career Goals

My ultimate career goal is to leverage my skills and knowledge in AI and ML to address real-world challenges and make a positive impact on society. I am committed to continuous learning and look forward to exploring opportunities in research and industry that align with the passion for AI and ML.

Instagram User Analytics

Project Description

As a data analyst at Instagram, my role is to analyze user interactions and engagement with the Instagram app. This involves using SQL and MySQL Workbench to extract valuable insights from the data. These insights are crucial for various teams within the company, including marketing, product development, and engineering.

The project aims to analyze user behavior and provide insights for the marketing and investor teams of Instagram. By leveraging the provide database, the analysis will address various aspects such as rewarding loyal users, encouraging inactive users, declaring contest winners, researching popular hashtags, determining he best day to launched campaigns, assessing user engagement, and identifying potential bot accounts.

The Problems

A) Marketing Analysis:

- Loyal User Reward: The marketing team wants to reward the most loyal users, i.e., those who have been using the platform for the longest time. Task: Identify the five oldest users on Instagram from the provided database.
- Inactive User Engagement: The team wants to encourage inactive users to start posting by sending them promotional emails. Task: Identify users who have never posted a single photo on Instagram.
- Contest Winner Declaration: The team has organized a contest where the user with the most likes on a single photo wins. Task: Determine the winner of the contest and provide their details to the team.
- Hashtag Research: A partner brand wants to know the most popular hashtags to use in their posts to reach the most people. Task: Identify and suggest the top five most commonly used hashtags on the platform.
- Ad Campaign Launch: The team wants to know the best day of the week to launch ads. Task: Determine the day of the week when most users register on Instagram. Provide insights on when to schedule an ad campaign.

.B) Investor Metrics:

- User Engagement: Investors want to know if users are still active and posting on Instagram or if they are making fewer posts. Task: Calculate the average number of posts per user on Instagram. Also, provide the total number of photos on Instagram divided by the total number of users.
- Bots & Fake Accounts: Investors want to know if the platform is crowded with fake and dummy accounts. Task: Identify users (potential bots) who have liked every single photo on the site, as this is not typically possible for a normal user.

Design

The analysis will be conducted using SQL queries on the Instagram database. The database contains tables for users, photos, comments, likes, follows, tags and more. By combining these tables and performing appropriate joins, aggregations and calculations, we can derive the necessary insights to answer the questions posed by the marketing and investor teams.

For this project, the SQL language will be used to interact with the provided Instagram database. Various SQL queries will be written to extract the required information from the database. The choice of the SQL database management system (DBMS) is left open-ended, as SQL is a standard language that can be executed on different platforms such as MySQL, PostgreSQL, or SQLite. This approach was executed by using MySQL software.

Findings

- # Rewarding most loyal users

By identifying the 5 oldest users on Instagram, the marketing team can reward them for their loyalty, potentially through special promotions or exclusive features.

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

```
select *from users
order by created_at
limit 5;
```

- # Remind inactive users to start posting

14	Jaclyn81
21	Rocio33
24	Maxwell.Halvorson
25	Tierra.Trantow
34	Pearl7
36	Ollie_Ledner37
41	Mckenna17
45	David.Osinski47
49	Morgan.Kassulke
53	Linnea59
54	Duane60
57	Julien_Schmidt
66	Mike.Auer39
68	Franco_Keebler64
71	Nia_Haag
74	Hulda.Macejkovic
75	Leslie67
76	Janelle.Nikolaus81
80	Darby_Herzog
81	Esther.Zulauf61
83	Bartholome.Bernhard
89	Jessyca_West
90	Esmeralda.Mraz57
91	Bethany20

Identifying users who have never posted a photo on Instagram allows the marketing team to target some big promotional emails or campaigns to encourage their engagement.

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

Findings

- # Declaring Contest Winner

By determining the user who received the most likes on a single photo, the contest winner can be declared, leading to increased engagement and excitement among users.

```
SELECT u.id, u.username, p.image_url, COUNT(l.user_id) AS like_count
FROM users u
JOIN photos p ON u.id = p.user_id
JOIN likes l ON p.id = l.photo_id
GROUP BY u.id, p.id
ORDER BY like_count DESC
LIMIT 1;
```

id	username	image_url	like_count
52	Zack_Kemmer93	https://jarret.name	48

- # Hashtag researching

Identifying the top five most commonly used hashtags on Instagram can help a partner brand optimise your post reach and engagement by using popular hashtags. The following hashtags are the most commonly used hastags on the platform.

```
# 5 most commonly used hashtags
SELECT t.tag_name, COUNT(*) AS count
FROM tags t
JOIN photo_tags pt ON t.id = pt.tag_id
GROUP BY t.tag_name
ORDER BY count DESC
LIMIT 5;
```

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

Findings

- # Launch AD Campaign

By analysing the registration patterns of a users, the team can determine the best day to schedule and AD campaign for maximum impact and reach.

```
# launch AD campaign
SELECT DAYNAME(created_at) AS registration_day, COUNT(*) AS total_registrations
FROM users
GROUP BY registration_day
ORDER BY total_registrations DESC
LIMIT 1;
```

registration_day	total_registrations
Thursday	16

- # User Engagement

Calculating the average number of posts for rape provides in fights into future activity levels on Instagram. Analysing the total number of photos in relation to the total number of users helps access user participation.

```
SELECT COUNT(*) / COUNT(DISTINCT user_id) AS average_posts_per_user
FROM photos;
```

average_posts_per_user
3.4730

```
#total number of photos on Instagram/total number of users
SELECT COUNT(*) / (SELECT COUNT(*) FROM users) AS photos_per_user_ratio
FROM photos;
```

photos_per_user_ratio
2.5700

Findings

- # Bots & Fake Accounts

The investors want to know if the platform is crowded with fake and dummy accounts. Identifying users who have liked every single photo on the site that helps detect potential bot & fake accounts exhibit abnormal behaviour.

```
SELECT u.id, u.username
FROM users u
INNER JOIN likes l ON u.id = l.user_id
GROUP BY u.id
HAVING COUNT(DISTINCT l.photo_id) = (SELECT COUNT(*) FROM photos);
```

id	username
5	Aniya_Hackett
14	Jaclyn81
21	Rocio33
24	Maxwell.Halvorson
36	Ollie_Ledner37
41	Mckenna17
54	Duane60
57	Julien_Schmidt
66	Mike.Auer39
71	Nia_Haag
75	Leslie67
76	Janelle.Nikolaus81
91	Bethany20

Results

This project has provided me with hands-on experience in data analysis, query optimization and insights generation.

- It has helped me enhance my technical skills and business acumen.
- While working on this project, I have achieved several things that have helped me in learning and understanding of data analysis and providing actionable insights.

https://drive.google.com/file/d/1ASHlvwB_EA0nde-ngeL3sFzpn7GA4C_/view?usp=drive_link

Operation & Metric Analytics

Project Description

Operational Analytics is a crucial process that involves analyzing a company's end-to-end operations. This analysis helps identify areas for improvement within the company. As a Data Analyst, we'll work closely with various teams, such as operations, support, and marketing, helping them derive valuable insights from the data they collect.

One of the key aspects of Operational Analytics is investigating metric spikes. This involves understanding and explaining sudden changes in key metrics, such as a dip in daily user engagement or a drop in sales. As a Data Analyst, we'll need to answer these questions daily, making it crucial to understand how to investigate these metric spikes.

In this project, we took the role of a Lead Data Analyst at a company like Microsoft. We were provided with various datasets and tables, and our task was to derive insights from this data to answer questions posed by different departments within the company. Our goal is to use our advanced SQL skills to analyze the data and provide valuable insights that can help improve the company's operations and understand sudden changes in key metrics.

The Problems

A) 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.

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.

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.

.B) Investigating Metric Spike:

A.Weekly User Engagement:

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

B.User Growth Analysis:

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

C.Weekly Retention Analysis:

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

D.Weekly Engagement Per Device:

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

E.Email Engagement Analysis:

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

Design

- Created a database and tables to store the provided data sets.
- Utilized SQL queries to perform analysis on the tables and derive the required insights.
- Answered specific questions related to the operations, such as the number of jobs reviewed, throughput, percentage share of each language, duplicate rows, user engagement, user growth, weekly retention, weekly engagement per device, and email engagement.
- Documented the findings and insights in a report format.

Findings

- # Jobs Reviewed Over Time

	date	jobs_reviewed
▶	30-11-2020	180
	29-11-2020	180
	28-11-2020	218
	27-11-2020	35
	26-11-2020	64
	25-11-2020	80

```
SELECT ds AS dates,
round((COUNT(job_id)/SUM(time_spent))*3600) AS
'Jobs Reviewed per Hour per day'
from job_data
where  ds between '2020-11-01' AND ' 2020-11-30'
group by ds;
```

- # Throughput Analysis

```
SELECT  ds AS date,
AVG(COUNT(*)) OVER
(ORDER BY ds ROWS
BETWEEN 6
PRECEDING AND
CURRENT ROW) AS
7_day_rolling_average
FROM ritika.job_data
GROUP BY  ds
ORDER BY  ds;
```

	date	7_day_rolling_average
▶	25-11-2020	1.0000
	26-11-2020	1.0000
	27-11-2020	1.0000
	28-11-2020	1.2500
	29-11-2020	1.2000
	30-11-2020	1.3333

```
SELECT ds,
COUNT(DISTINCT event)
AS event_per_day
FROM ritika.job_data
GROUP by ds;
```

ds	event_per_day
25-11-2020	1
26-11-2020	1
27-11-2020	1
28-11-2020	2
29-11-2020	1
30-11-2020	2

Using both metrics together can provide a comprehensive understanding of throughput performance, with the daily metric offering insights into short-term fluctuations and the rolling average revealing long-term trends.

Findings

- # Language Share Analysis

```
SELECT language, COUNT(*) AS content_count,
COUNT(*)*100.0/SUM(COUNT(*))OVER()AS
percentage_shareFROM job_dataGROUP BY language;
```

language	content_count	percentage_share
English	1	12.50000
Arabic	1	12.50000
Persian	3	37.50000
Hindi	1	12.50000
French	1	12.50000
Italian	1	12.50000

- # Throughput Analysis

```
SELECT *
FROM  ( SELECT *, ROW_NUMBER()OVER(PARTITION
BY job_id) AS rownum
FROM job_data )a WHERE rownum > 1 ;
```

ds	job_id	actor_id	event	language	time_spent	org	rownum
28-11-2020	23	1005	transfer	Persian	22	D	2
26-11-2020	23	1004	skip	Persian	56	A	3

Findings

- # Weekly User Engagement

Week Number	Weekly Active Users
17	85
18	194
19	208
20	195
21	208
22	230
23	224
24	252
25	245
26	123
27	14
28	9
29	5
30	1
31	2
32	1

```
SELECT
extract(WEEK FROM occurred_at) AS "Week Number",
COUNT(DISTINCT user_id) AS "Weekly Active Users“
FROM events
GROUP BY 1;
```

- # User Growth Analysis

months	users	Growth(%)
1	712	NULL
2	685	-3.79
3	765	11.68
4	907	18.56
5	993	9.48
6	1086	9.37
7	1281	17.96
8	1347	5.15
9	330	-75.50
10	390	18.18
11	399	2.31
12	486	21.80

```
SELECT months, users,
round(((users/LAG(users,1)
over (ORDER BY months) -
1)*100),2)AS "Growth(%)"
FROM ( select extract(month
FROM created_at)AS Months,
COUNT(activated_at)AS Users
 from users
where activated_at NOT IN("")
group by 1
order by 1
)sub ;
```

Findings

- # Weekly User Engagement

Week Number	Weekly Active Users
17	85
18	194
19	208
20	195
21	208
22	230
23	224
24	252
25	245
26	123
27	14
28	9
29	5
30	1
31	2
32	1

```
SELECT
extract(WEEK FROM occurred_at) AS "Week Number",
COUNT(DISTINCT user_id) AS "Weekly Active Users“
FROM events
GROUP BY 1;
```

- # User Growth Analysis

months	users	Growth(%)
1	712	NULL
2	685	-3.79
3	765	11.68
4	907	18.56
5	993	9.48
6	1086	9.37
7	1281	17.96
8	1347	5.15
9	330	-75.50
10	390	18.18
11	399	2.31
12	486	21.80

```
SELECT months, users,
round(((users/LAG(users,1)
over (ORDER BY months) -
1)*100),2)AS "Growth(%)"
FROM ( select extract(month
FROM created_at)AS Months,
COUNT(activated_at)AS Users
 from users
where activated_at NOT IN("")
group by 1
order by 1
)sub ;
```

Findings

- # Weekly Retention Analysis

```
SELECT COUNT(user_id)as users,
SUM(CASE WHEN retention_week = 1 THEN 1 ELSE 0 END ) AS  week_1,
SUM(CASE WHEN retention_week = 2 THEN 1 ELSE 0 END ) AS  week_2,
SUM(CASE WHEN retention_week = 3 THEN 1 ELSE 0 END ) AS  week_3,
SUM(CASE WHEN retention_week = 4 THEN 1 ELSE 0 END ) AS  week_4,
SUM(CASE WHEN retention_week = 5 THEN 1 ELSE 0 END ) AS  week_5
FROM (
SELECT a.user_id,
a.sign_up_week,
b.engagement_week,
b.engagement_week - a.sign_up_week as retention_week
 FROM ((select distinct user_id, extract(week from occurred_at) as
sign_up_week
from events
where event_type='signup_flow'and event_name='complete_signup' and
extract(week from occurred_at)=18)a
left JOIN
(select distinct user_id,extract(week from occurred_at) as
engagement_weekfrom events
where event_type='engagement')b
on a.user_id=b.user_id )
order by a.user_id )a
```

users	week_1	week_2	week_3	week_4	week_5
182	11	4	3	0	1

- # Email Engagement Analysis

```
SELECT    action AS email_engagement_type,
COUNT(DISTINCT user_id) AS engaged_users
FROM    ritika.email_events
GROUP BY    action ;
```

email_engagement_type	engaged_users
email_clickthrough	5277
email_open	5927
sent_reengagement_email	3653
sent_weekly_digest	4111

Findings

- # Weekly Retention Analysis

```
Select
extract(year from occurred_at)as year,
extract(week from occurred_at)as week,
device,
COUNT(distinct user_id) as num_users
FROM events
WHEREevent_type='engagement‘
group by 1,2,3
order BY1,2,3
```

year	week	device	num_users
2014	17	samsung galaxy tablet	2
2014	17	samsung galaxy note	1
2014	17	samsung galaxy s4	7
2014	18	acer aspire desktop	4
2014	18	acer aspire notebook	4
2014	18	amazon fire phone	2
2014	18	asus chromebook	4
2014	18	dell inspiron desktop	3
2014	18	dell inspiron notebook	12
2014	18	hp pavilion desktop	6
2014	18	htc one	2
2014	18	ipad air	8
2014	18	ipad mini	7
2014	18	iphone 4s	4
2014	18	iphone 5	7
2014	18	iphone 5s	8
2014	18	kindle fire	5
2014	18	lenovo thinkpad	28
2014	18	mac mini	1
2014	18	macbook air	18

year	week	device	num_users
2014	17	acer aspire desktop	2
2014	17	acer aspire notebook	2
2014	17	amazon fire phone	1
2014	17	asus chromebook	3
2014	17	dell inspiron desktop	1
2014	17	dell inspiron notebook	4
2014	17	hp pavilion desktop	2
2014	17	htc one	2
2014	17	ipad air	1
2014	17	ipad mini	3
2014	17	iphone 4s	3
2014	17	iphone 5	11
2014	17	iphone 5s	5
2014	17	lenovo thinkpad	8
2014	17	mac mini	1
2014	17	macbook air	4
2014	17	macbook pro	13
2014	17	nexus 5	4
2014	17	nexus 7	4
2014	17	nokia lumia 635	2

year	week	device	num_users
2014	18	macbook pro	41
2014	18	nexus 10	2
2014	18	nexus 5	9
2014	18	nexus 7	5
2014	18	nokia lumia 635	4
2014	18	samsung galaxy note	4
2014	18	samsung galaxy s4	12
2014	18	windows surface	1
2014	19	acer aspire notebook	8
2014	19	amazon fire phone	4
2014	19	asus chromebook	3
2014	19	dell inspiron desktop	1
2014	19	dell inspiron notebook	7
2014	19	hp pavilion desktop	5
2014	19	htc one	6
2014	19	ipad air	10
2014	19	ipad mini	1
2014	19	iphone 4s	7
2014	19	iphone 5	19
2014	19	iphone 5s	9

year	week	device	num_users
2014	19	kindle fire	4
2014	19	lenovo thinkpad	21
2014	19	mac mini	2
2014	19	macbook air	23
2014	19	macbook pro	37
2014	19	nexus 10	3
2014	19	nexus 5	15
2014	19	nexus 7	7
2014	19	nokia lumia 635	2
2014	19	samsung galaxy note	4
2014	19	samsung galaxy s4	14
2014	19	windows surface	3
2014	20	acer aspire desktop	2
2014	20	acer aspire notebook	4
2014	20	amazon fire phone	1
2014	20	asus chromebook	6
2014	20	dell inspiron desktop	11
2014	20	dell inspiron notebook	9
2014	20	hp pavilion desktop	2
2014	20	htc one	3

Results

The project was instrumental in improving my SQL proficiency and provided practical insights into operational aspects, user engagement, and the product's growth trajectory. I successfully presented these findings to the leadership team, showcasing my ability to deliver valuable insights that can positively impact business strategies. Overall, this experience has boosted my confidence and independence in handling data analysis tasks in real-world scenarios.

https://drive.google.com/file/d/1f4iKeA8VBxteL3vVn6ajrCE1gEk1QgJt/view?usp=sharing

Hiring Process Analytics

Project Description

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

As a data analyst, we were given a dataset containing records of previous hires. Our job was to analyze this data and answer certain questions that can help the company improve its hiring process.
The objective of this project is to analyze the hiring process data of company and derive meaningful insights to improve the hiring process. We worked with a dataset containing records of previous hires, including information on gender, salary, department, position, and other relevant factors. By performing data analysis using Excel, we aim to understand trends in hiring, salary distribution, departmental proportions, and position tiers.

The Problems

A. Hiring Analysis: The hiring process involves bringing new individuals into the organization for various roles.
Task: Determine the gender distribution of hires. How many males and females have been hired by the company?
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.
Task: What is the average salary offered by this company? Use Excel functions to calculate this.
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.
Task: Create class intervals for the salaries in the company. This will help you understand the salary distribution.
D. Departmental Analysis: Visualizing data through charts and plots is a crucial part of data analysis.
Task: Use a pie chart, bar graph, or any other suitable visualization to show the proportion of people working in different departments.
E. Position Tier Analysis: Different positions within a company often have different tiers or levels.
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.

Design

1.Handling Missing Data: Check if there are any missing values in the dataset. If there are, decide on the best strategy to handle them.

2.Clubbing Columns: If there are columns with multiple categories that can be combined, do so to simplify your analysis.

3.Outlier Detection: Check for outliers in the dataset that may skew your analysis.

4.Removing Outliers: Decide on the best strategy to handle outliers. This could be removing them, replacing them, or leaving them as is, depending on the situation.

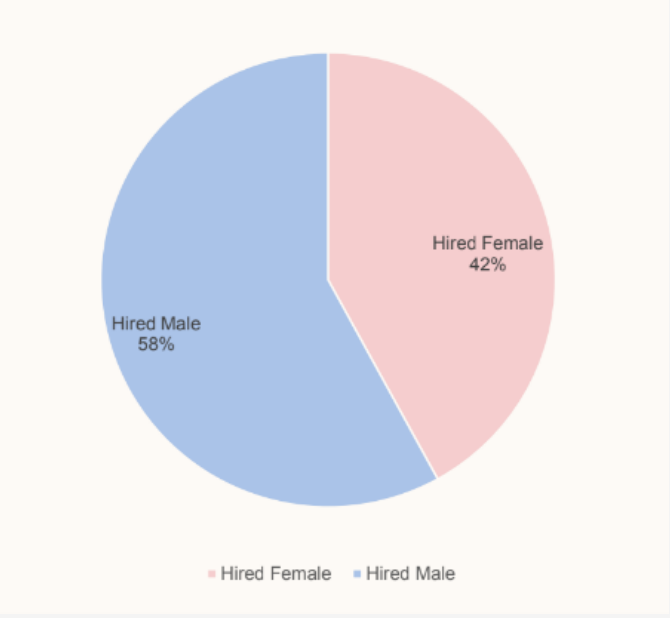
5.Data Summary: After cleaning and preparing your data, summarize your findings. This could involve calculating averages, medians, or other statistical measures. It could also involve creating visualizations to better understand the data.

Findings

- # Hiring Analysis

Used Excel's COUNTIF or PIVOT TABLE functionality to count the number of males and females hired by the company.

Gender	Proportion of Hired people
Female	1850
Male	2551
Total	**4401**



- # Salary Analysis

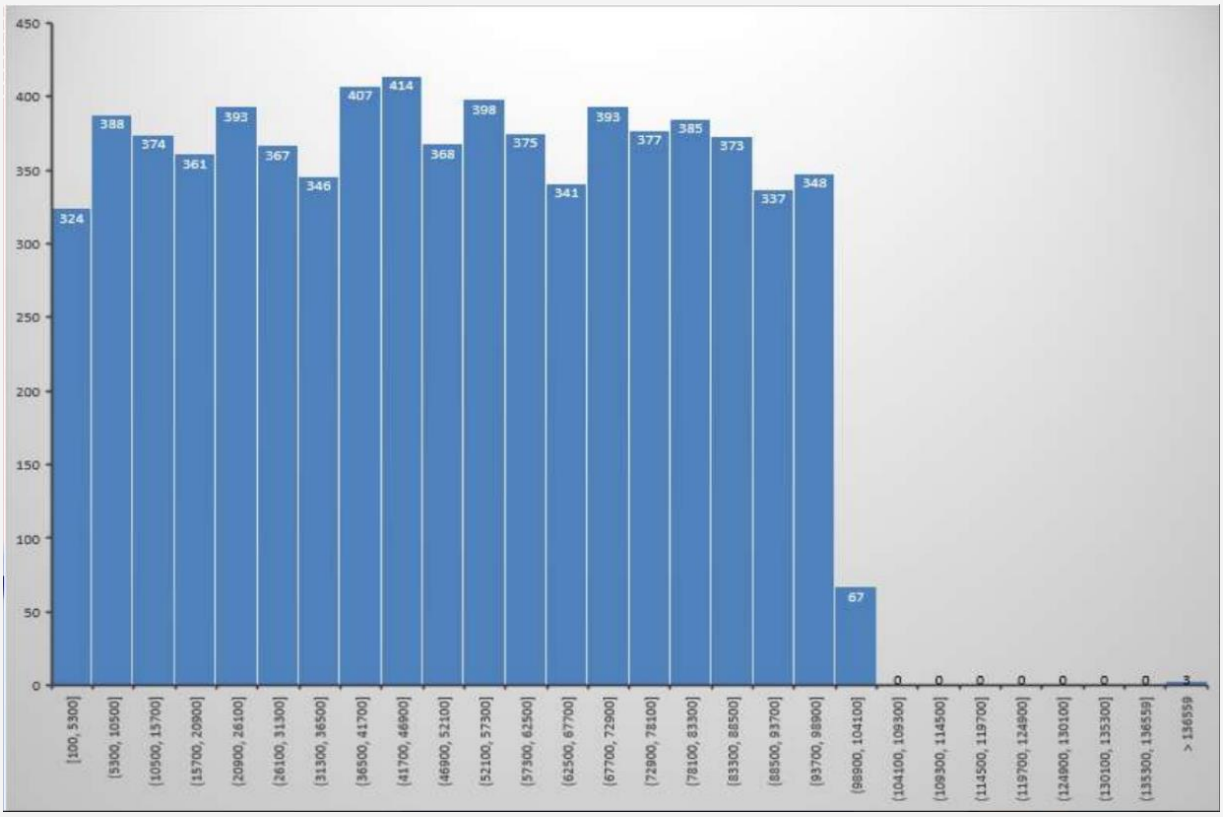
Using the AVERAGE function in excel to calculate the average salary of all hires.

Average Salary	50004.93
Max Salary	400000
Min Salary	100
Median	49636
Standard Deviation	28851.28

Findings

- # Salary Distribution

By using histogram we can create class intervals for salaries, which gives a clear understanding of the salary distribution, identifying any salary disparities or outliers.



- # Salary Analysis

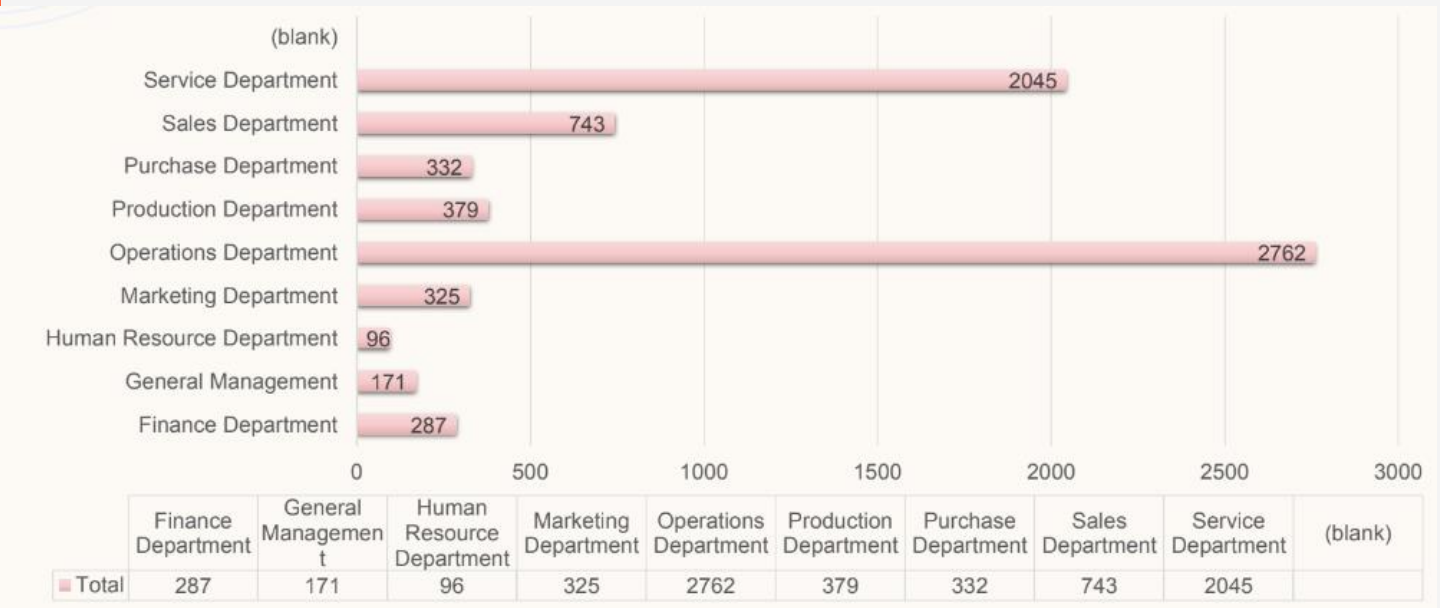
Using the AVERAGE function in excel to calculate the average salary of all hires.

Findings

- # Departmental Analysis

Visualizing departmental proportions allowed us to assess the distribution of employees across different departments and understand the hiring needs of each.

Department	Proportion of people
Finance Department	287
General Management	171
Human Resource Department	96
Marketing Department	325
Operations Department	2762
Production Department	379
Purchase Department	332
Sales Department	743
Service Department	2045
Total	7140



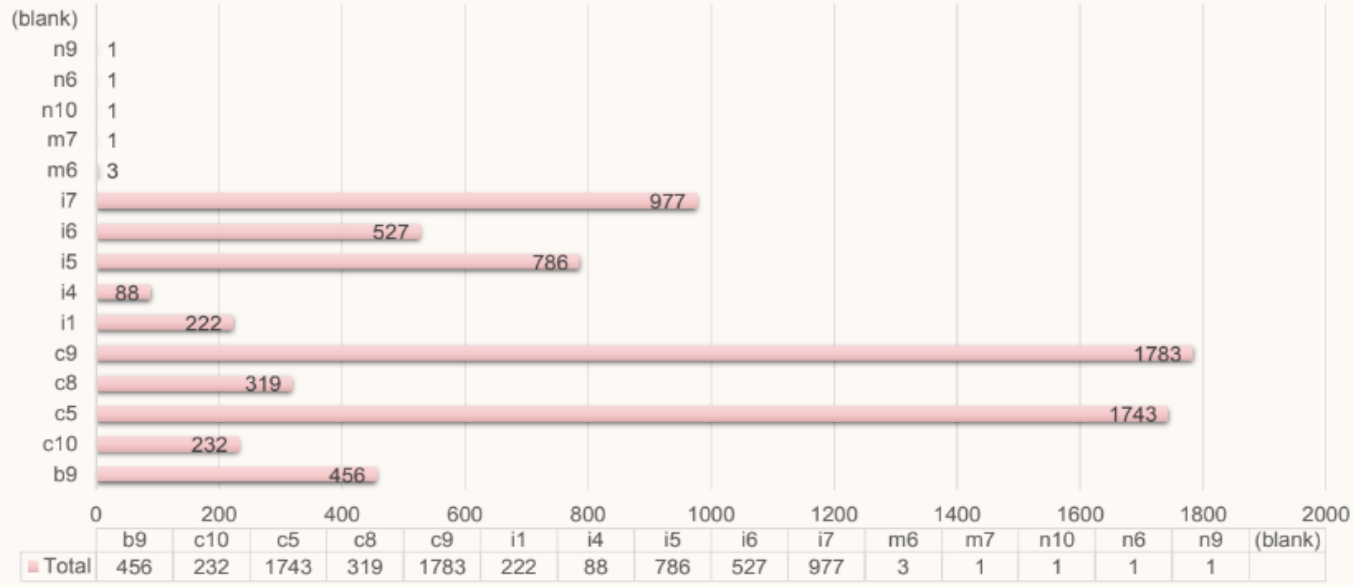
	Finance Department	General Management	Human Resource Department	Marketing Department	Operations Department	Production Department	Purchase Department	Sales Department	Service Department	(blank)
Total	287	171	96	325	2762	379	332	743	2045	

Findings

- # Position Tier Analysis

The position tier analysis will help to identify the distribution of positions across different levels within the company, providing insights into the organization's structure.

Position Tier	Distribution
-	1
b9	456
c10	232
c5	1743
c8	319
c9	1783
i1	222
i4	88
i5	786
i6	527
i7	977
m6	3
m7	1
n10	1
n6	1
n9	1
Grand Total	7140



Results

I have done the analysis on the provided dataset/data as per the questions asked and provided the necessary insights and tried to plot the necessary charts/graphs as per the requirement and my understanding. This project helped me in better understanding the process of Exploratory Data Analysis (EDA).
Exploratory Data Analysis (EDA) is the process of understanding and analyzing datasets to resource dataset which helped me to understand the steps involved in EDA better. EDA process can be carried out by using tools such as Excel, Google Sheets or WPS spreadsheet etc.

https://drive.google.com/file/d/19LFxlwvbjd77GbVSgFSOiT1eDCz-dxM8/view?usp=sharing

IMDB Movie Analysis

Project Description

The objective of this project is to analyze the IMDB movie dataset and identify factors that influence the success of movie, understand the relationship between budget and ratings. The analysis will focus on movie genre distribution, duration, language, directors with high IMDB scores and relation between profit and budget and how these all impact on IMDB scores.

The analysis presented here aims to explore various factors that influence the success of a movie on IMDB with success defined by high IMDB ratings this analysis is valuable for movie producers directors and investors who seek to make informed decisions in their future projects. The data analysis is divided into five tasks, each focusing on different aspects of the dataset.

The Problems

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.
B. 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.
C. 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.
D. 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.
E. 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.

Design

Data Cleaning: This step involves preprocessing the data to make it suitable for analysis. It includes handling missing values, removing duplicates, converting data types if necessary, and possibly feature engineering.

Data Analysis: Here, you'll explore the data to understand the relationships between different variables. You might look at the correlation between movie ratings and other factors like genre, director, budget, etc. You might also want to consider the year of release, the actors involved, and other relevant factors.

Five 'Whys' Approach: This technique will help you dig deeper into the problem. For instance, if you find that movies with higher budgets tend to have higher ratings, you can ask "Why?" repeatedly to uncover the root cause. Here's an example:

•Q: "Why do movies with higher budgets tend to have higher ratings?"
•A: They can afford better production quality.
•Q: "Why does better production quality lead to higher ratings?"
•A: It enhances the viewer's experience.
•Q: "Why does an enhanced viewer experience lead to higher ratings?"
•A: Viewers are more likely to rate a movie highly if they enjoyed watching it.
•Q: "Why are viewers more likely to rate a movie highly if they enjoyed watching it?"
•A: Positive experiences lead to positive reviews.
•Q: "Why do positive reviews matter?"
•A: They influence other viewers' decisions to watch the movie, increasing its popularity and success.

Findings

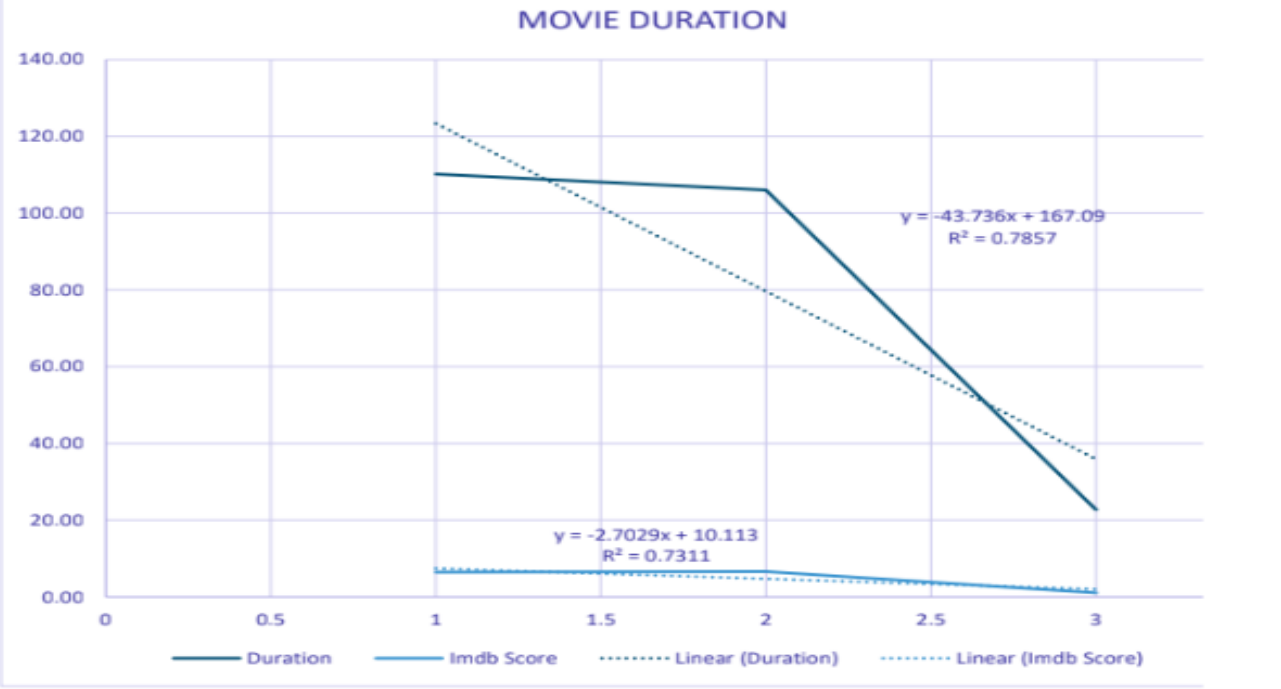
- # Movie Genre Analysis

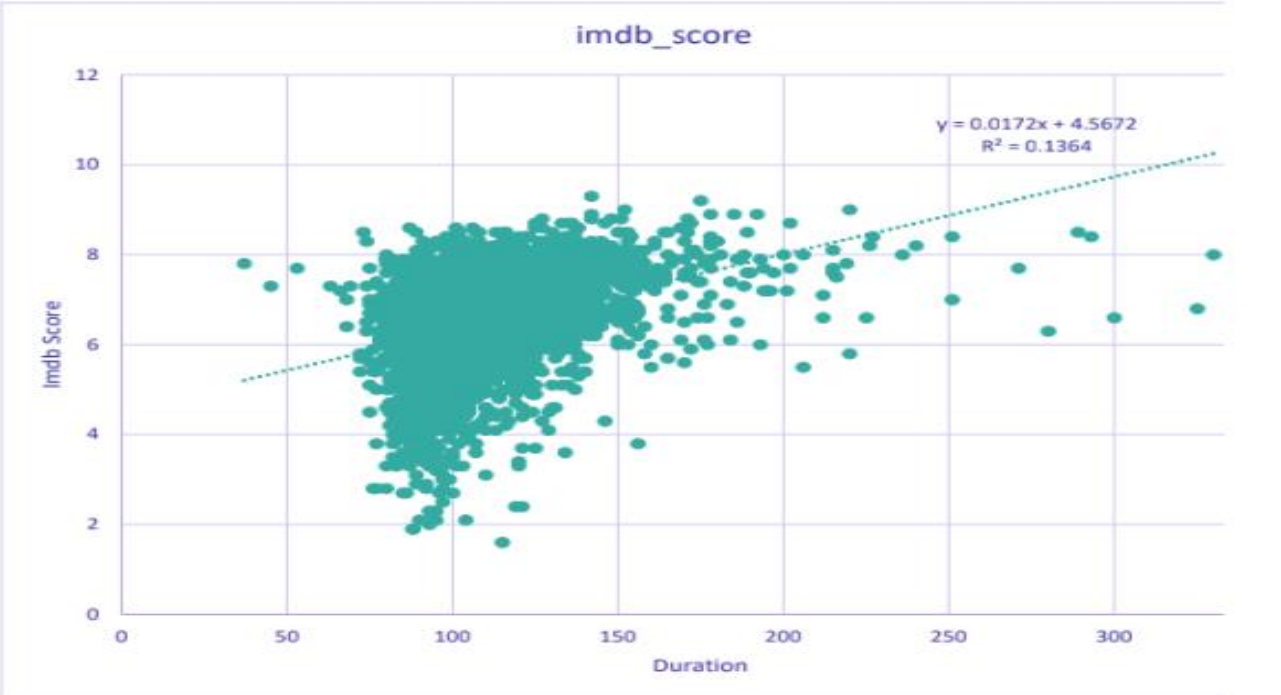
  We separated the multiple genres for a single movie from a single column to multiple columns by using text to column function.
  The following table represents the no. of movies for each genre.
  Formulas used for getting above insights:
  =AVERAGE(IF($F$2:$M$3657=A3662, $U$2:$U$3657))
  =MEDIAN(IF($F$2:$M$3657=A3662, $U$2:$U$3657))
  =MAX(IF($F$2:$M$3657=A3662, $U$2:$U$3657))
  =MIN(IF($F$2:$M$3657=A3662, $U$2:$U$3657))
  =VAR.P(IF($F$2:$M$3657=A3662, $U$2:$U$3657))
  =STDEV.S(IF($F$2:$M$3657=A3662, $U$2:$U$3657))

genres	No. of Movies	average	Median	MAX	MIN	VAR	STDEV
Action	924	6.29	6.3	9	2.1	1.08	1.04
Adventure	752	6.45	6.6	8.9	2.3	1.26	1.12
Animation	194	6.70	6.8	8.6	2.8	0.98	0.99
Biography	237	7.16	7.2	8.9	4.5	0.48	0.69
Comedy	1443	6.19	6.3	8.8	1.9	1.07	1.03
Crime	692	6.55	6.6	9.3	2.4	0.97	0.99
Documentary	45	6.99	7.4	8.5	1.6	1.87	1.38
Drama	1846	6.80	6.9	9.3	2.1	0.79	0.89
Family	431	6.21	6.3	8.6	1.9	1.37	1.17
Fantasy	486	6.28	6.4	8.9	2.2	1.29	1.14
Film-Noir	1	7.70	7.7	7.7	7.7	0.00	0.00
History	146	7.15	7.2	8.9	5.5	0.44	0.67
Horror	374	5.90	5.9	8.6	2.3	0.98	0.99
Music	149	6.34	6.5	8.5	1.6	1.51	1.23
Musical	95	6.59	6.7	8.5	2.1	1.21	1.10
Mystery	371	6.48	6.5	8.6	3.1	1.01	1.00
Romance	839	6.43	6.5	8.5	2.1	0.92	0.96
Sci-Fi	479	6.33	6.4	8.8	1.9	1.36	1.17
Sport	143	6.59	6.8	8.3	2	1.09	1.05
Thriller	1074	6.37	6.4	9	2.7	0.94	0.97
War	149	7.06	7.1	8.6	4.3	0.64	0.80
Western	57	6.81	6.8	8.9	4.7	0.87	0.94

Findings

Movie Duration Analysis





The slope is increasing as the movie duration increases IMDB scores tend to increase.

Findings

Language Analysis

Created a list of unique languages in a separate column.
=UNIQUE(AB2:AB3724)
Then, statistics (Mean, Median and Standard Deviation)
=AVERAGEIFS(AI2:AI3725, AB2:AB3724, "Language")
=MEDIAN(IF(AB2:AB3724= "Language", AI2:AI3725))
=STDEV.S(IF(AB2:AB3724= "Language", AI2:AI3725))

language	Mean	Median	STD
English	6.43	6.50	1.05
Mandarin	7.02	7.25	0.77
French	7.36	7.30	0.52
Spanish	7.08	7.20	0.86
Cantonese	7.34	7.30	0.35
Maya	7.80	7.80	0.00
German	7.77	7.80	0.71
Hindi	7.22	7.40	0.80
Persian	8.13	8.40	0.55
Italian	7.19	7.00	1.16
Portuguese	7.76	8.00	0.98
Indonesian	7.90	7.90	0.42
Dari	7.50	7.50	0.14
Hebrew	8.00	8.00	0.00
Romanian	7.90	7.90	0.00
Danish	7.90	8.10	0.53
Aboriginal	6.95	6.95	0.78
Zulu	7.30	7.30	0.00
Arabic	7.20	7.20	0.00
Norwegian	7.15	7.30	0.57
Korean	7.70	7.70	0.57
Vietnamese	7.40	7.40	0.00
Japanese	7.66	8.00	0.99
Russian	6.50	6.50	0.00
Dutch	7.57	7.80	0.40
Bosnian	4.30	4.30	0.00
Mongolian	7.30	7.30	0.00
Kazakh	6.00	6.00	0.00
Aramaic	7.10	7.10	0.00
Filipino	6.70	6.70	0.00
Czech	7.40	7.40	0.00
Thai	6.63	6.60	0.45
Hungarian	7.10	7.10	0.00

Findings

Director Analysis

=UNIQUE(B2:B3724)

=AVERAGEIFS($U$2:$U$3725, $B$2:$B$3724, "Director_Name")

Directors	Movie Ratings
Akira Kurosawa	8.70
Tony Kaye	8.60
Charles Chaplin	8.60
Ron Fricke	8.50
Majid Majidi	8.50
Damien Chazelle	8.50
Alfred Hitchcock	8.50
Sergio Leone	8.43
Christopher Nolan	8.43
Richard Marquand	8.40
Asghar Farhadi	8.40

Budget Analysis

Using the following function

=CORREL($R$2:$R$3724,$E$2:$E$3724)

Profit = Gross – Budget

Top 10 movies with highest profit margins

Profit	Movie Titles
$523.506	AvatarÂ
$502.177	Jurassic WorldÂ
$458.672	TitanicÂ
$449.936	Star Wars: Episode IV - A New HopeÂ
$424.449	E.T. the Extra-TerrestrialÂ
$403.280	The AvengersÂ
$377.784	The Lion KingÂ
$359.545	Star Wars: Episode I - The Phantom MenaceÂ
$348.316	The Dark KnightÂ
$329.999	The Hunger GamesÂ

Results

This project fosters critical thinking, problem-solving and the ability to draw meaningful conclusions from data, all of which are crucial for a future career in data analytics. Additionally, the project's structured approach and use of tools like Excel enhance technical proficiency and provide a strong foundation for more advanced data analysis tasks. This project helped to develop skills in data preprocessing, visualization and statistical analysis.

SHEET LINK

https://docs.google.com/spreadsheets/d/18H-jwOuunc_4UzYbG1_Wlc2U2PQJJdul/edit?usp=drive_link&ouid=102610263604488941402&rtpof=true&sd=true

Bank Loan Case Study

Project Description

As a data analyst at a finance company specializing in urban lending, I undertook a comprehensive Exploratory Data Analysis (EDA) project to address a significant challenge faced by the organization. The challenge was rooted in the fact that some customers, particularly those lacking sufficient credit history, were exploiting the system by defaulting on loans. This issue posed a dual risk to the company: losing potential business from capable applicants and incurring financial losses from loan defaults.

The dataset I worked with encompassed detailed information on loan applications and comprised two distinct scenarios:

1.Customers with Payment Difficulties: These were applicants who exhibited late payments of more than X days on at least one of the first Y installments of the loan.

2. All Other Cases: This category included cases where payments were made promptly.

The project's primary objective was to leverage EDA techniques to unearth underlying patterns that could predict the likelihood of a customer facing payment difficulties. This predictive insight, once extracted, could inform crucial business decisions such as loan approval, loan amount adjustments, or the imposition of higher interest rates on risky applicants. By discerning key factors contributing to loan default, our company aimed to enhance its decision-making process regarding loan approvals.

This project contributed to a deeper understanding of risk analytics in the banking and financial services sector. It encompassed the exploration of various customer and loan attributes to unveil patterns indicative of potential defaults. The insights gained from this analysis could be instrumental in improving the company's lending practices, mitigating risks, and ensuring that capable applicants were not unfairly rejected.

The Problems

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.
•**Graph suggestion:** Create a bar chart or column chart to visualize the proportion of missing values for each variable.

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.
•**Graph suggestion:** Create box plots or scatter plots to visualize the distribution of numerical variables and highlight the outliers.

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.
•**Graph suggestion:** Create a pie chart or bar chart to visualize the distribution of the target variable and highlight the class imbalance.

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.
•**Graph suggestion:** Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

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.
•**Graph suggestion:** Create correlation matrices or heatmaps to visualize the correlations between variables within each segment. Highlight the top correlated variables for each scenario using different colors or shading.

Design

- Data collection: Download the data set.

- Data cleaning: Handling duplicates by removing rows. Deleting the columns with more than 50% blank cells.

- Outlier detection: Identify outliers in the data set that might skew analysis.

- Data summary: Calculates statistical measures such as averages, medians, quartiles and range. Create visualization like box and whisker, column charts, heatmaps and histograms for better understanding of data.

- Documented the findings and insights in a report format.

Findings

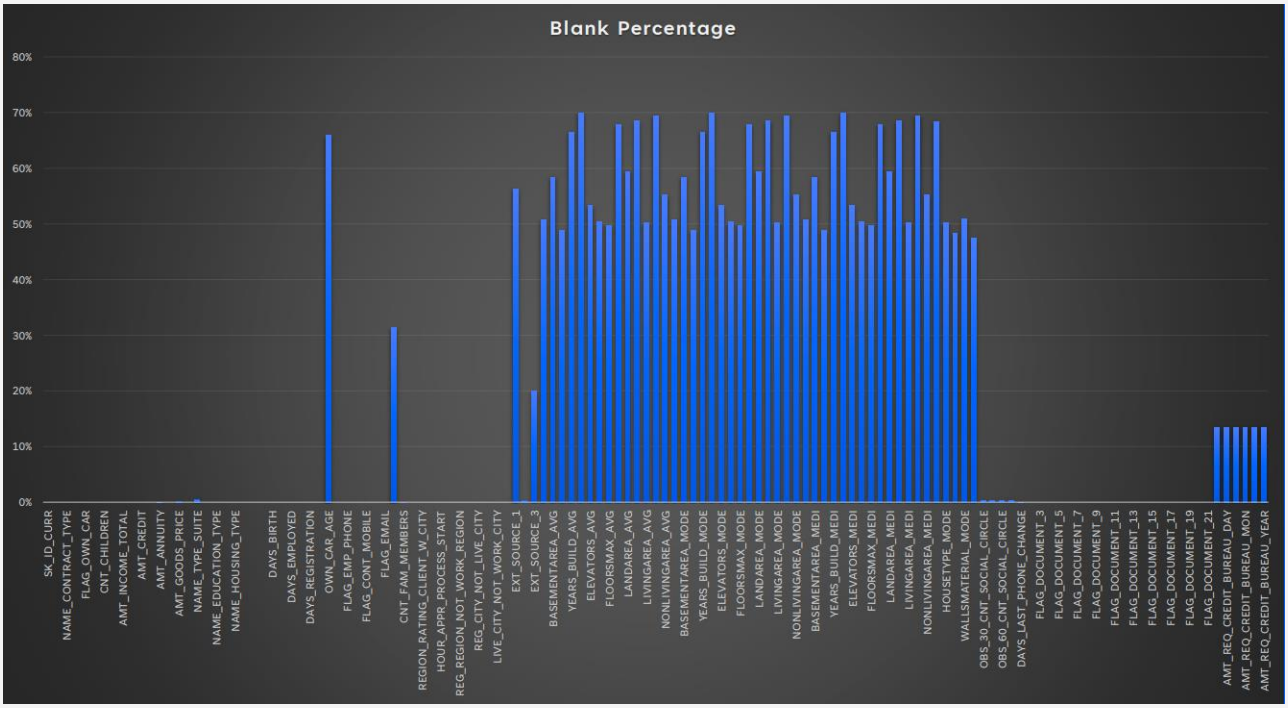
Identify Missing Data and Deal with it Appropriately

Utilize Excel functions like COUNT, ISBLANK, and IF to identify missing data. Consider using functions like AVERAGE or MEDIAN for imputation or other appropriate methods available in Excel.

The percentage of missing values varies across columns, with some columns having no missing data while others exhibit significant levels of missingness, up to 70%. This insight highlights the importance of addressing missing data effectively to ensure the accuracy of our analysis. In subsequent tasks, we will employ appropriate methods such as imputation to handle these missing values, ensuring that our analysis is based on complete and reliable task.

Columns having more than 50 % blanks were deleted. Following is the list of deleted columns:

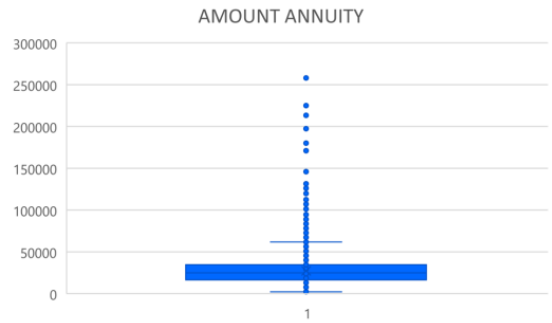
Deleted Columns			
OWN_CAR_AGE	BASEMENTAREA_AVG	NONLIVINGAPARTMENTS_AVG	LANDAREA_MEDI
EXT_SOURCE_1	YEARS_BUILD_AVG	NONLIVINGAREA_AVG	LIVINGAPARTMENTS_MEDI
APARTMENTS_AVG	COMMONAREA_AVG	APARTMENTS_MODE	LIVINGAREA_MEDI
ELEVATORS_AVG	ENTRANCES_AVG	BASEMENTAREA_MODE	NONLIVINGAPARTMENTS_MEDI
FLOORSMIN_AVG	LANDAREA_AVG	FLOORSMIN_MODE	NONLIVINGAREA_MEDI
LIVINGAPARTMENTS_AVG	LIVINGAREA_AVG	LANDAREA_MODE	FONDKAPREMONT_MODE
NONLIVINGAPARTMENTS_AVG	NONLIVINGAREA_AVG	LIVINGAPARTMENTS_MODE	HOUSETYPE_MODE
APARTMENTS_MODE	BASEMENTAREA_MODE	LIVINGAREA_MODE	WALLSMATERIAL_MODE
NONLIVINGAPARTMENTS_MODE	NONLIVINGAREA_MODE	APARTMENTS_MEDI	
BASEMENTAREA_MEDI	YEARS_BUILD_MEDI	COMMONAREA_MEDI	
ELEVATORS_MEDI	ENTRANCES_MEDI	FLOORSMIN_MEDI	



Findings

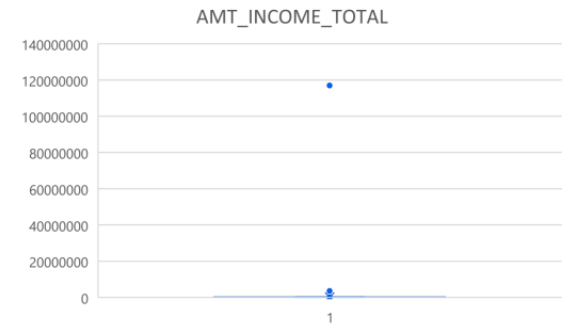
Identify Outliers In Dataset

Utilize Excel functions like QUARTILE, IQR, and conditional formatting to identify potential outliers. Consider applying thresholds or business rules to determine if the outliers are valid data points or require further investigation.

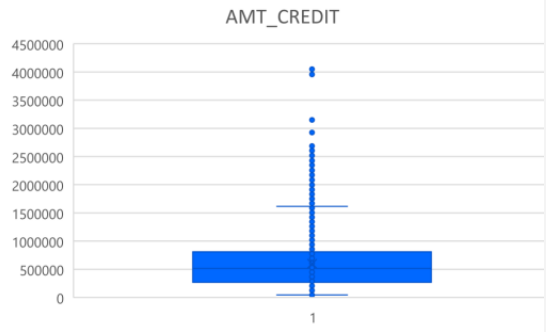


Amt annuity has outliers ranging from 50,000 to 1,32,000

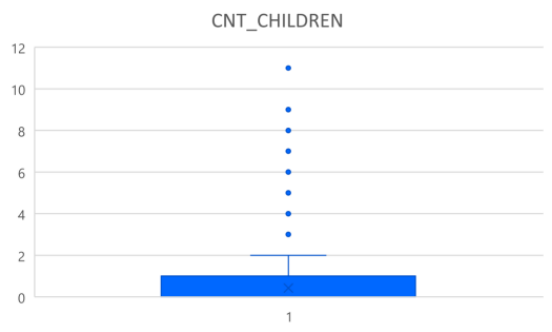
Amount goods price has outliers ranging from 3L to 23L.



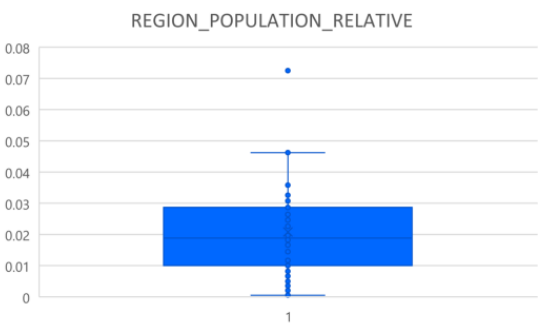
Amount income has outliers ranging from 0 to 36.00.000.



Amount credit has most outliers ranging from 1.3L to 27L.



Count of children has most outliers ranging from 0 to 2.



Region population relative has most outliers ranging from 0 to 0.046.

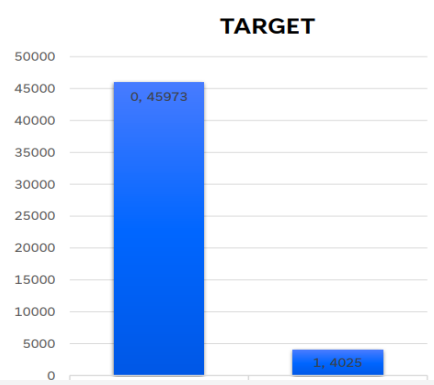
Findings

Analyze Data Imbalance

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

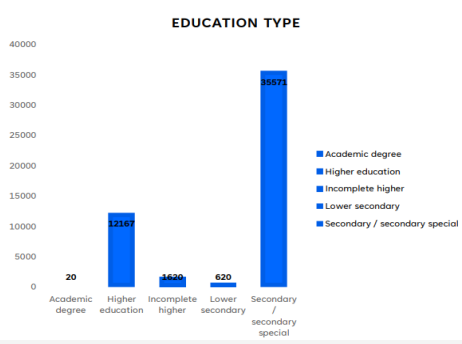
The target variable analysis reveals a class in balance issue with significantly high number of instances of “0” class compared to the “1” class

TARGET	Count
0	45973
1	4025
Grand Total	49998



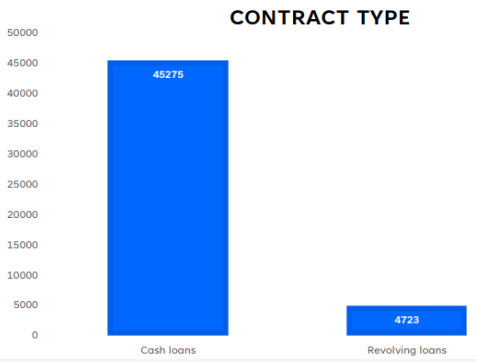
The majority of loan applicants have completed Secondary/ secondary special education, followed by higher education, indicating a diverse educational background among applicants.

Education type	Count of Secondary / secondary special
Academic degree	20
Higher education	12167
Incomplete higher	1620
Lower secondary	620
Secondary / secondary special	35571
Grand Total	**49998**



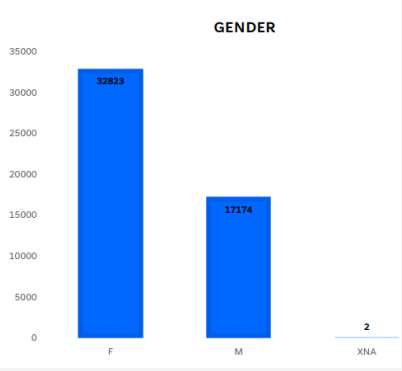
The majority of loans are Cash loans (45,275), while a smaller proportion are Revolving loans (4723).

CONTRACT TYPE	Count
Cash loans	45275
Revolving loans	4723
Grand Total	49998



The CODE_GENDER analysis reveals a gender distribution among loan applicants with a higher number of females(32,823) compared to males(17,174).

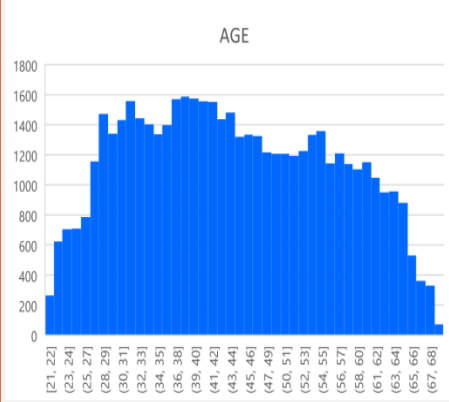
GENDER	Count
F	32823
M	17174
XNA	2
Grand Total	49999

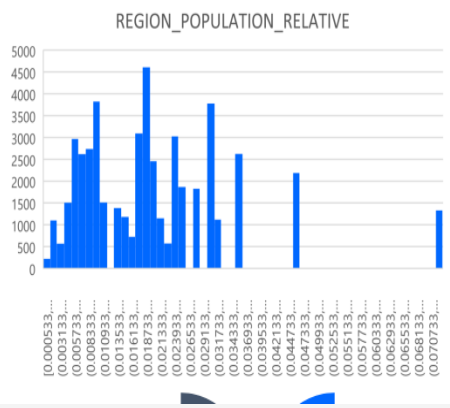


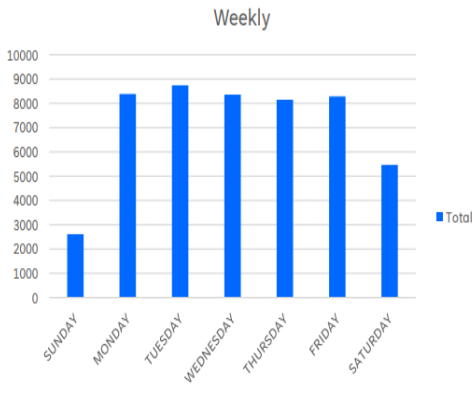
Findings

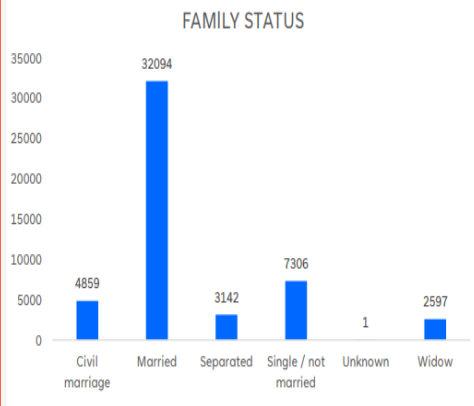
Perform univariate/segmented univariate and Bivariate analysis

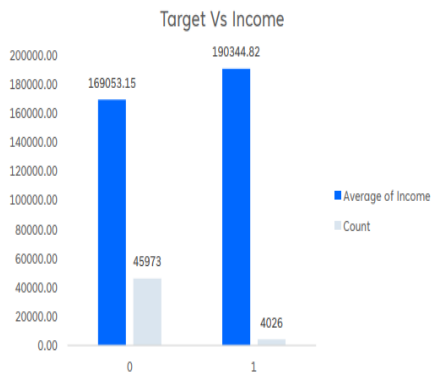
Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios and bivariate analysis to explore relationships between variables and the target value using Excel functions and features.

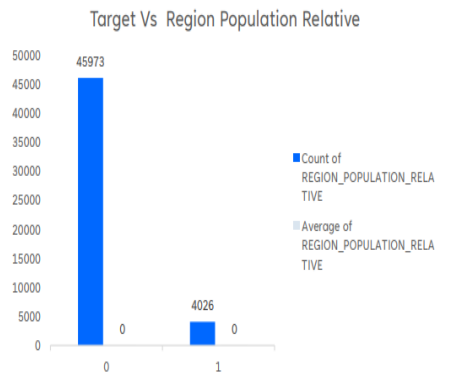


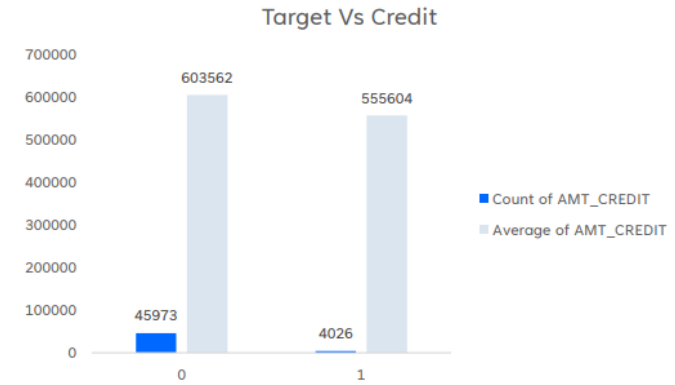


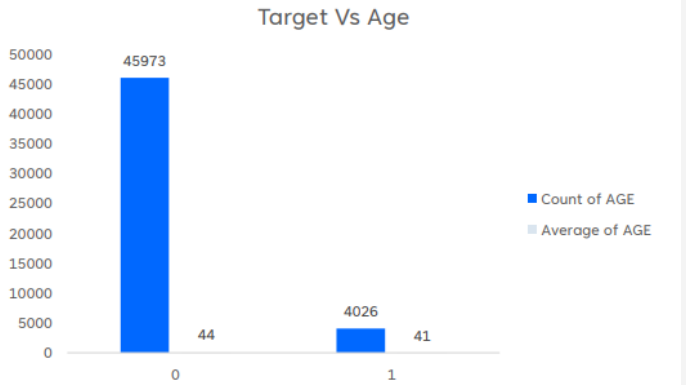












Findings

Identify Top Correlation For Different Scenarios

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.

Columns	Corrrelation Coefficient
AMT_ANNUITY	-0.012399094
AMT_CREDIT	-0.032428347
AMT_GOODS_PRICE	-0.041306523
AMT_INCOME_TOTAL	0.010893745
CNT_CHILDREN	0.010893745
DAYS_BIRTH	0.076787685
DAYS_EMPLOYED	-0.040294905
EXT_SOURCE_1	-0.156806293
EXT_SOURCE_2	-0.158424274
EXT_SOURCE_3	-0.181275965
REGION_POPULATION_RELATIVE	-0.040799172
REGION_RATING_CLIENT	0.066130148
REGION_RATING_CLIENT_W_CITY	0.067079294

AMT_ANNUITY	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	AMT_CREDIT	TARGET	REGION_POPULATION_RELATIVE	CNT_CHILDREN	DAYS_EMPLOYED	DAYS_BIRTH	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3
24700.5	351000	202500	406597.5	1	0.018801	0	-637	-9461	2	2	0.083036967	0.262948593	0.139375
35698.5	1129500	270000	1293502.5	0	0.003541	0	-1188	-16765	1	1	0.311267311	0.622245775	
6750	135000	67500	135000	0	0.010032	0	-225	-19046	2	2		0.555912083	0.729566691
29686.5	297000	135000	312682.5	0	0.008019	0	-3039	-19005	2	2		0.65044169	
21865.5	513000	121500	513000	0	0.028663	0	-3038	-19932	2	2		0.322738287	
27517.5	454500	99000	490495.5	0	0.035792	0	-1588	-16941	2	2		0.354224732	0.621226338
41301	1395000	171000	1560726	0	0.035792	1	-3130	-13778	2	2	0.774761413	0.723999852	0.492060094
42075	1530000	360000	1530000	0	0.003122	0	-449	-18850	3	3		0.714279286	0.54065445
33826.5	913500	112500	1019610	0	0.018634	0	365243	-20099	2	2	0.587334047	0.205747288	0.751723715
20250	405000	135000	405000	0	0.019689	0	-2019	-14469	2	2		0.746643629	
21177	652500	112500	652500	0	0.0228	1	-679	-10197	2	2	0.319760172	0.651862333	0.363945239
10678.5	135000	38419.155	148365	0	0.015221	0	365243	-20417	2	2	0.72204445	0.555183162	0.652896552
5881.5	67500	67500	80865	0	0.031329	0	-2717	-13439	2	2	0.464831117	0.715041819	0.176652579
28966.5	697500	225000	918468	0	0.016612	1	-3028	-14086	2	2		0.566906613	0.77008707
32778	679500	189000	773680.5	0	0.010006	0	-203	-14583	2	1	0.721939769	0.642656205	
20160	247500	157500	299772	0	0.020713	0	-1157	-8728	3	3	0.115634337	0.346633981	0.678567689
26149.5	387000	108000	509602.5	0	0.018634	0	-1317	-12931	2	2		0.23637784	0.062103036
13500	270000	81000	270000	0	0.010966	1	-191	-9776	2	2		0.683513346	
7875	157500	112500	157500	0	0.04622	0	-7804	-17718	1	1		0.706428403	0.556727426
17563.5	454500	90000	544491	0	0.015221	1	-2038	-11348	2	2		0.58661714	0.477649155
21375	427500	135000	427500	0	0.015221	0	-4286	-18252	2	2	0.565654882	0.113374513	
37561.5	927000	202500	1132573.5	0	0.025164	1	-1652	-14815	2	2	0.43770902	0.233766958	0.542445144
32521.5	450000	450000	497520	0	0.020713	1	-4306	-11146	3	2		0.457142972	0.358951229
23850	225000	83250	239850	0	0.006296	0	365243	-24827	3	3		0.624304737	0.669056695
12703.5	247500	135000	247500	0	0.026392	2	-746	-11286	2	2		0.786179309	0.565607981
11074.5	225000	90000	225000	0	0.028663	0	-3494	-19334	2	2	0.561948409	0.651405637	0.461482391
27076.5	702000	112500	979992	1	0.018029	0	-2628	-18724	3	2		0.54847716	0.190705948
23827.5	270000	112500	327024	0	0.019101	1	-1234	-15948	2	2		0.541123702	0.659405532
57676.5	675000	270000	790830	0	0.04622	0	-1796	-9994	1	1	0.600395905	0.68501099	0.524496446
9000	180000	90000	180000	0	0.030755	0	-1010	-10341	2	2	0.297913509	0.502779038	
24592.5	477000	292500	665892	0	0.025164	0	-2668	-15280	2	2		0.479987342	0.410102573
25033.5	360000	112500	512064	0	0.008575	0	-1104	-11144	2	2	0.274422372	0.627300401	
20893.5	180000	90000	199008	0	0.010032	0	-4404	-12974	2	2		0.559466792	0.798137231
39069	679500	360000	733315.5	0	0.015221	1	-2060	-11694	2	2		0.321744896	0.411848556
32895	1125000	135000	1125000	0	0.019689	0	-4585	-15997	2	2		0.172497508	
44509.5	450000	112500	450000	0	0.008575	0	-1275	-12158	2	2		0.663158076	0.678567689
23157	553500	198000	641173.5	0	0.01885	2	-768	-17199	2	2	0.842763466	0.681698803	0.754406173
15151.5	454500	121500	454500	0	0.030755	0	-1288	-21077	2	2	0.804586121	0.719798537	0.722392895

Results

This project on exploratory data analysis (EDA) of a loan application data set has significantly advanced my skills and knowledge it enabled me to master critical data handling and cleaning techniques, statistical analysis and data visualization using Excel. Moreover, I developed valuable problem-solving skills, particularly in addressing data quality issues such as missing data, outliers, data imbalances. This project has enhanced my ability to think analytically and communicate complex data insights effectively. It has strengthened my portfolio as a budding data scientist and analyst.

Impact of Car Features on Price and Profitability

Project Description

The automotive industry has been rapidly evolving over the past few decades, with a growing focus on fuel efficiency, environmental sustainability, and technological innovation. With increasing competition among manufacturers and a changing consumer landscape, it has become more important than ever to understand the factors that drive consumer demand for cars.

In recent years, there has been a growing trend towards electric and hybrid vehicles and increased interest in alternative fuel sources such as hydrogen and natural gas. At the same time, traditional gasoline-powered cars remain dominant in the market, with varying fuel types and grades available to consumers.

For the given dataset, as a Data Analyst, the client has asked How can a car manufacturer optimize pricing and product development decisions to maximize profitability while meeting consumer demand?

This problem could be approached by analyzing the relationship between a car's features, market category, and pricing, and identifying which features and categories are most popular among consumers and most profitable for the manufacturer. By using data analysis techniques such as regression analysis and market segmentation, the manufacturer could develop a pricing strategy that balances consumer demand with profitability, and identify which product features to focus on in future product development efforts. This could help the manufacturer improve its competitiveness in the market and increase its profitability over time.

The Problems

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.

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.

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.

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.

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

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

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

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

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

Design

■ **DATA CLEANING:** Removal of duplicates, rows containing blank cells.

■ **DESCRIPTIVE STATISTICS:** Descriptive statistics served as our starting point, providing a comprehensive overview of the dataset. Measures such as means, median and standard deviations allowed us to grasp the central tendencies and dispersions of key variables.

■ **DATA VISUALIZATION:** Visualization techniques including bar charts, scatter plots and pivot tables, were instrumental in presenting complex data in an interpretable format. These visualizations enhanced our ability to uncover patterns, trends and relationships within the data.

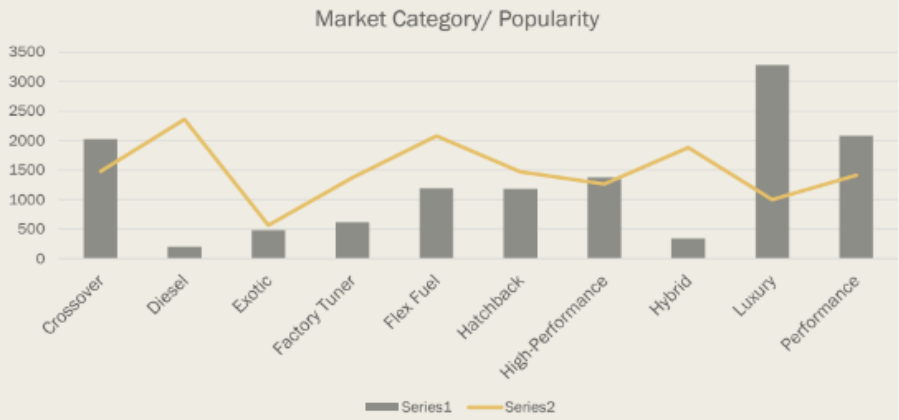
■ **REGRESSION ANALYSIS:** Regression analysis emerged as a powerful tool to explore the relationship between variables. By conducting regression analysis we identified the factors most strongly associated with the car prices, offering invaluable insights into pricing determinants.

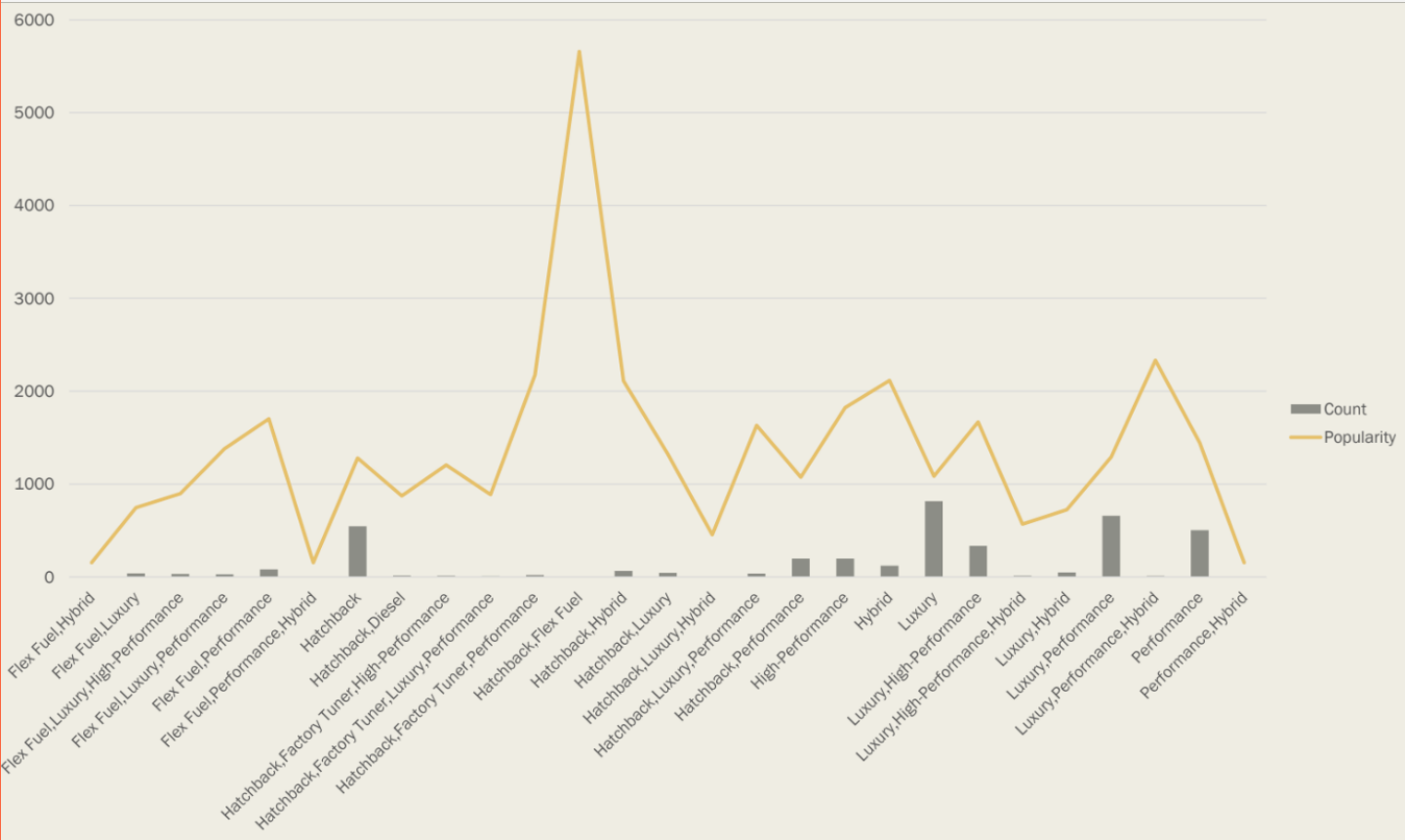
Findings

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

Multiple categories in a single column so after splitting got the following table and chart.

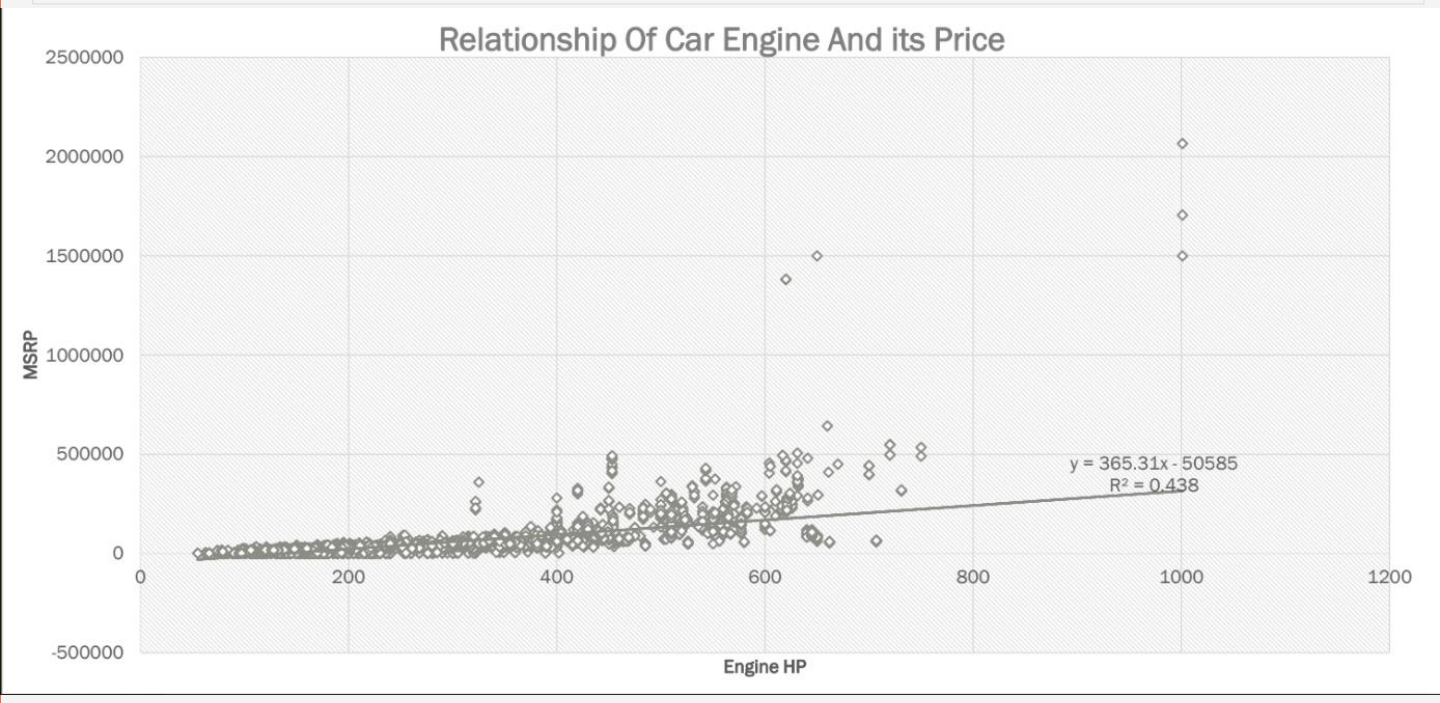
Market Category	Count	Popularity
Crossover	2026	1477.90
Diesel	205	2358.02
Exotic	482	563.91
Factory Tuner	618	1371.33
Flex Fuel	1191	2078.49
Hatchback	1181	1471.07
High-Performance	1378	1262.51
Hybrid	343	1881.04
Luxury	3279	997.50
Performance	2078	1413.01





Findings

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

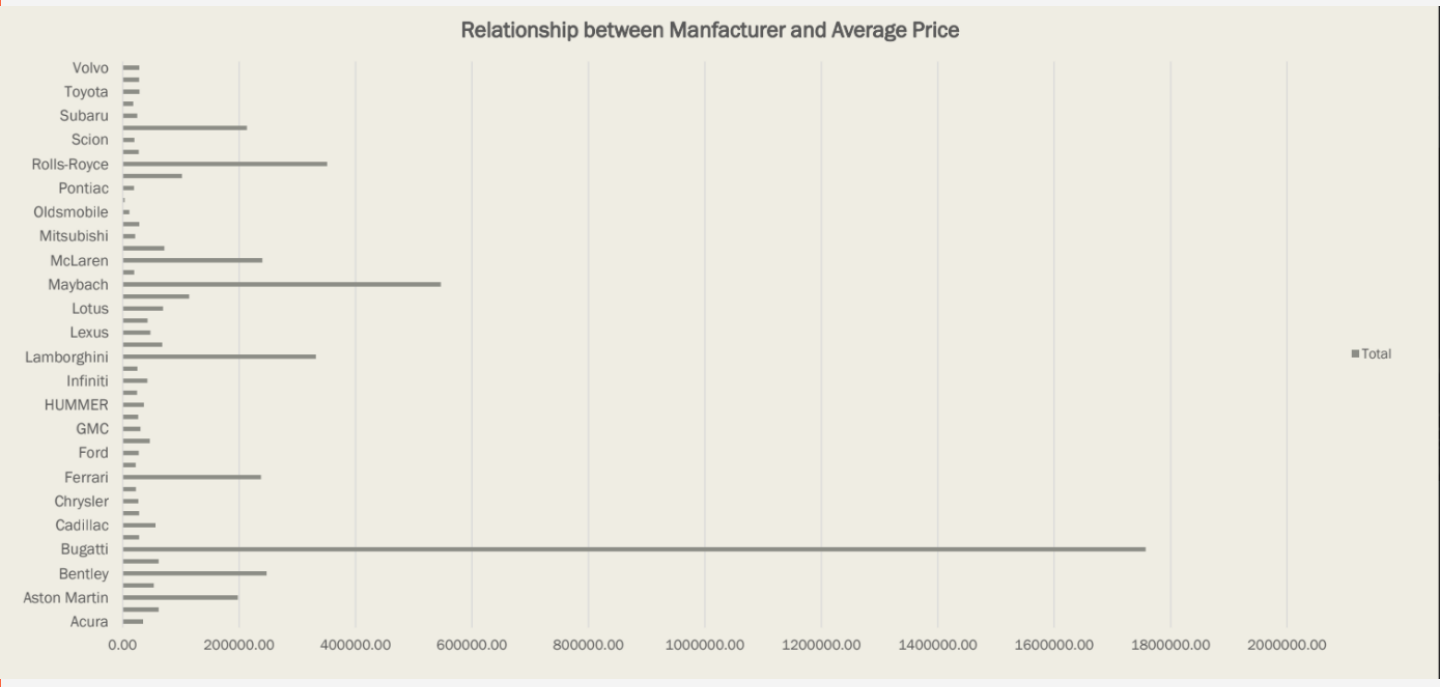


Which car features are most important in determining a car's price?

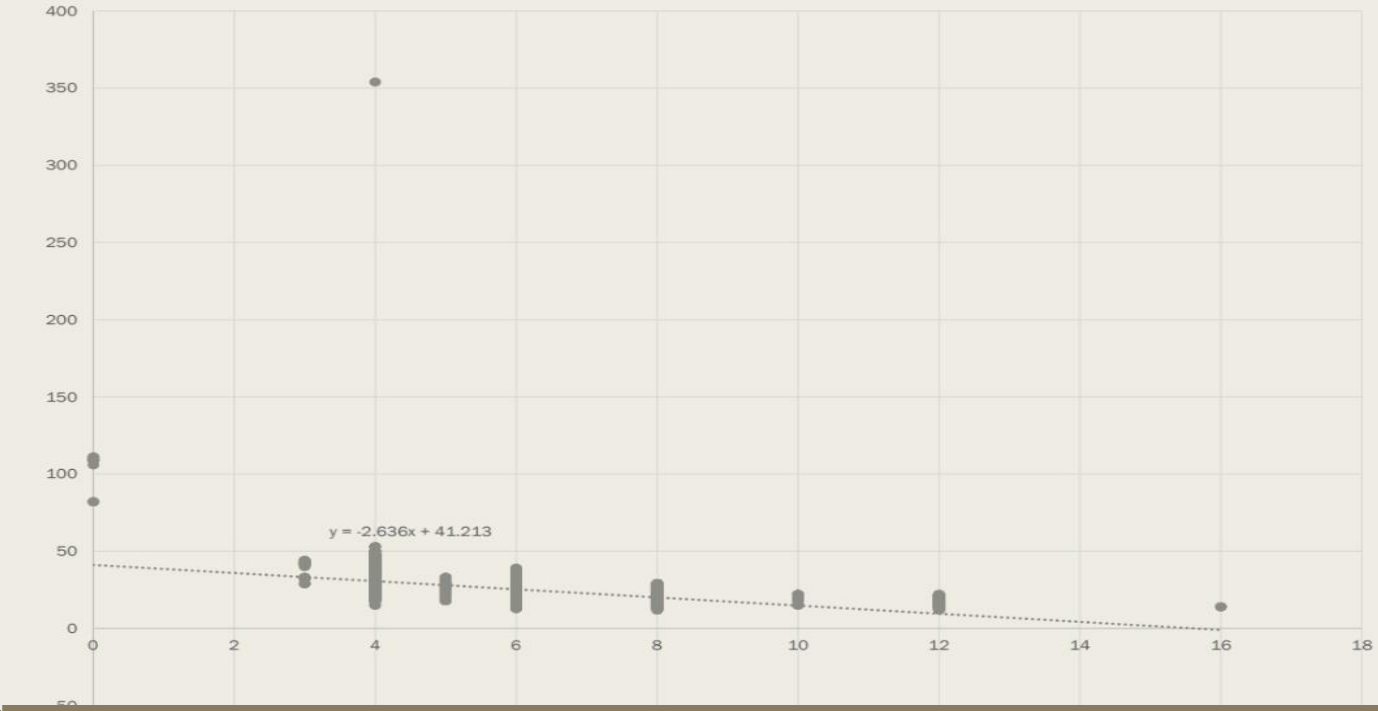
HIGHWAY MPG								
Regression Statistics								
Multiple R	0.198941632							
R Square	0.039577773							
Adjusted R Square	0.039496471							
Standard Error	59075.12953							
Observations	11815							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	1.69886E+12	1.69886E+12	486.7986353	0			
Residual	11813	4.12258E+13	3489870929					
Total	11814	4.29247E+13						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	82962.37196	1997.447559	41.53419278	0	79047.04563	86877.69829	79047.04563	86877.69829
highway MPG	-1611.209062	73.02595074	-22.06351367	9.531E-106	-1754.351958	-1468.066166	-1754.351958	-1468.066166

Findings

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



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



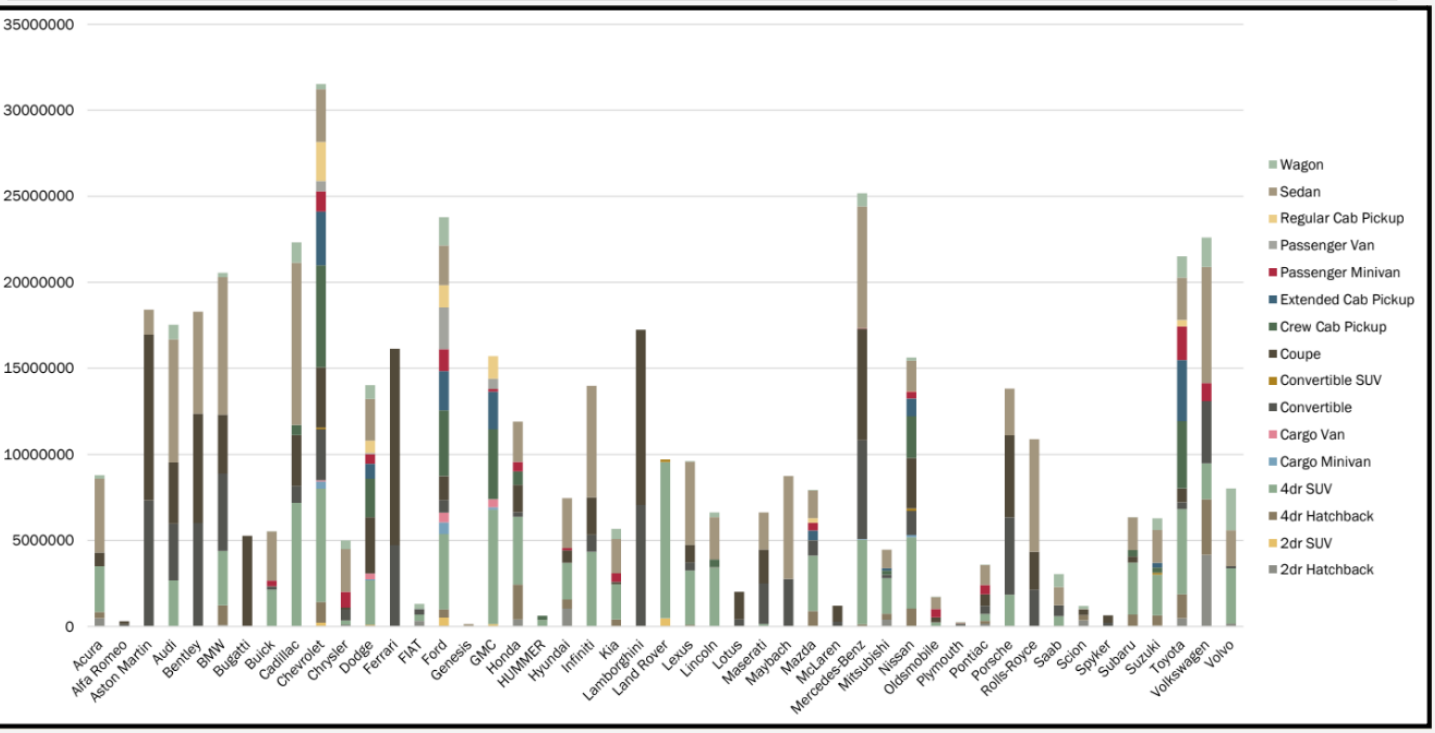
correlation coefficient between the number of cylinders and highway MPG

-0.620315935

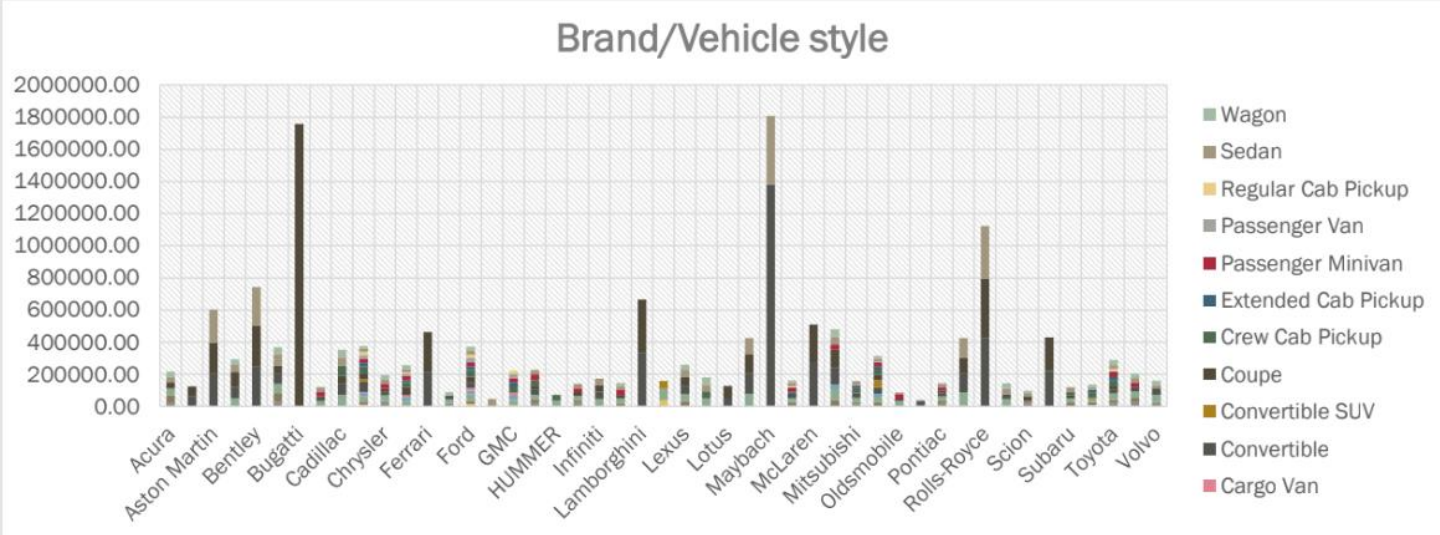
Findings

Building the Dashboard:

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



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

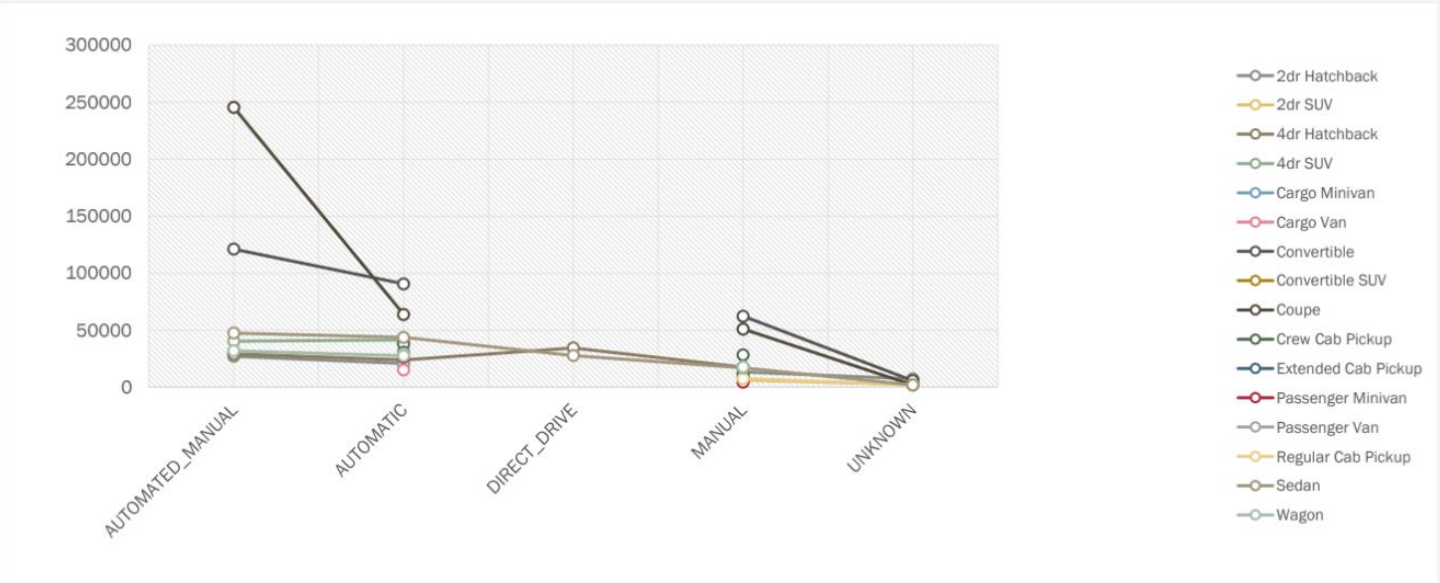


Findings

Building the Dashboard:

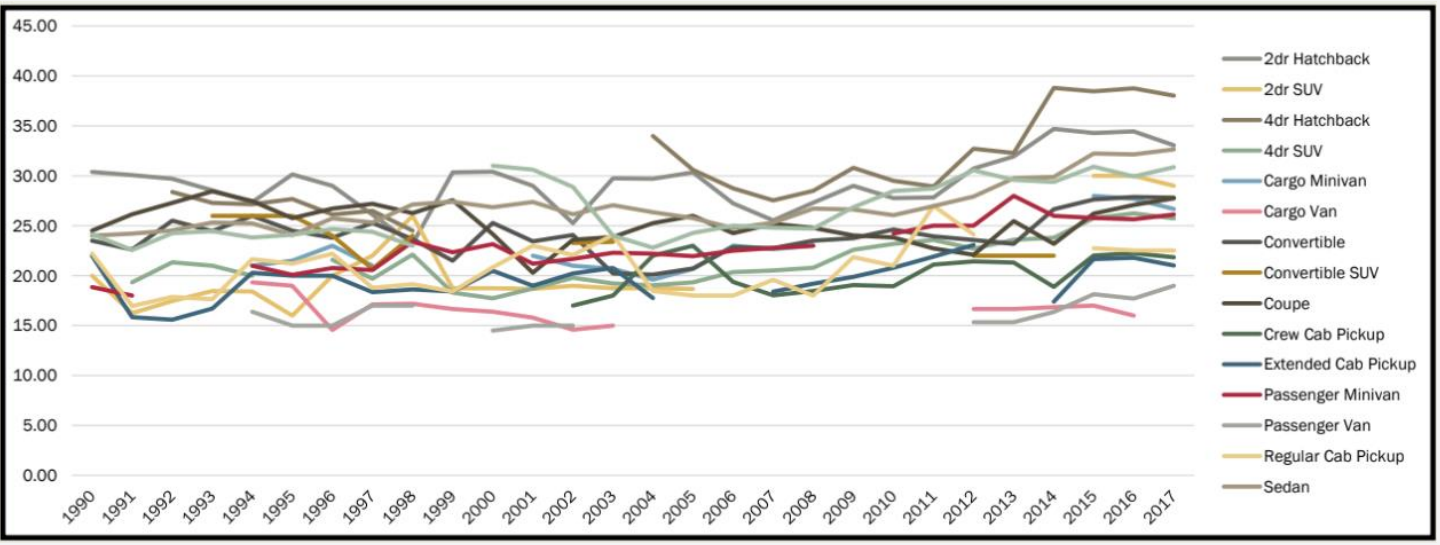
How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?

Scatter plot chart to visualize the relationship between MSRP and transmission type, with different symbols for each body style. Calculate the average MSRP for each combination of transmission type and body style using AVERAGEIFS or Pivot Tables.



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

Line chart to show the trend of fuel efficiency (MPG) over time for each body style. Calculate the average MPG for each combination of body style and model year using AVERAGEIFS or Pivot Tables.

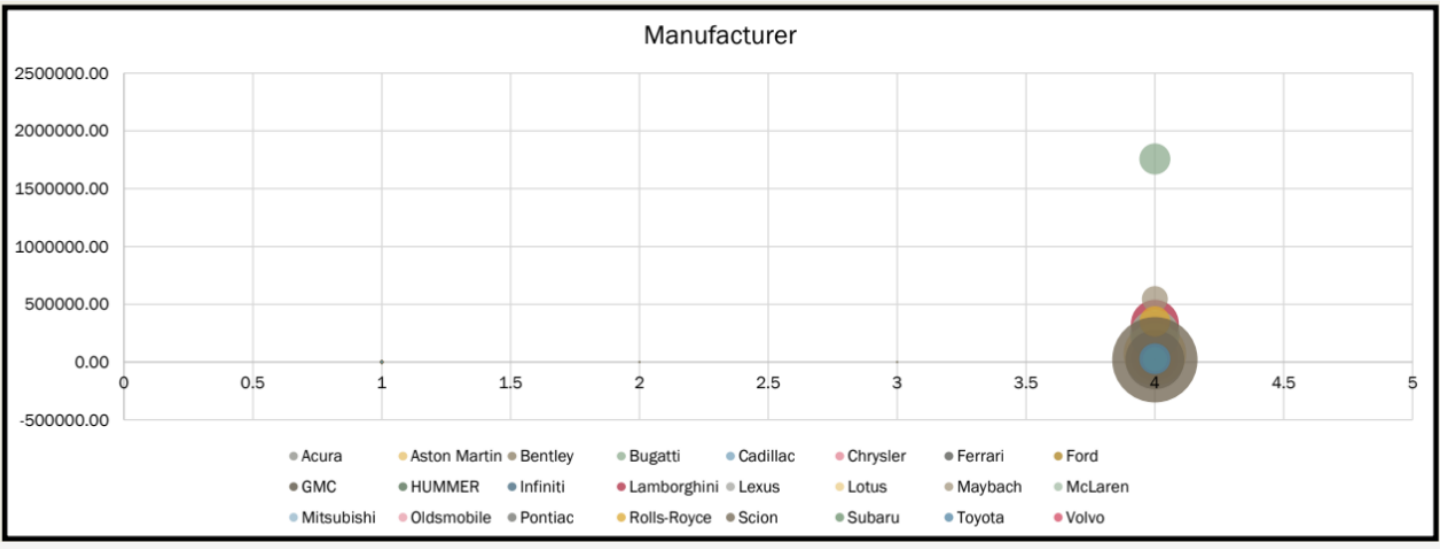


Findings

Building the Dashboard:

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

Bubble chart to visualize the relationship between horsepower, MPG, and price across different car brands. Assign different colors to each brand and label the bubbles with the car model name. Calculate the average horsepower, MPG, and MSRP for each car brand using AVERAGEIFS or Pivot Tables.



Results

■ This project has greatly benefited me by providing practical hands on experience in data analysis. It equipped me with essential skills in data processing, regression analysis and data visualization using tools like Excel. Moreover, the project enhanced problem-solving and communication skills as to present actionable insight. These acquired skills and the successful completion of a real world analysis project have not only bolstered my confidence but also improved the employability and prepared me for the future data-related roles.

■ For the business this project has been instrumental in informed decision making it allowed the business to optimize its pricing and product development strategies based on data driven insights by understanding the factors influencing car prices the business can make adjustment that enhance the competitiveness reduce cost and increase the profitability. The project demonstrates the value of data analysis in guiding market efforts, risk medication, and improving customer satisfaction, reinforcing the importance of data-driven approaches in the business's future endeavors.

https://drive.google.com/file/d/1IAuA80TlOOelgjd9KQZOt6Mj92as47az/view?usp=drive_link

ABC Call Volume Trend

Project Description

This project centers on optimizing ABC Company's inbound calling team by harnessing da ta analytics. With a 23-day dataset encompassing agent in formation queue times call data and statuses, we aim to improve consumer satisfaction and operational efficiency. We'll explore AI tools like IBR, RPA, Predictive Analytics and Intelligent Routing and examine the role of customer service representatives in delivering exceptional support. Our primary focus is on inbound customer support, where we'll strive to captivate and delight customers, turning them in to loyal advocates for our business.

Effective advertising is vital for any business, as it drives sales and shapes initial perceptions. Business have diverse target audiences, ranging from local to international. They employ various advertising channels, from online directories to traditional media like radios, cinema and print. In the fiercely competitive advertising landscape, analytics plays a pivotal role. The objective is to identify cost effective media platforms that can efficiently convert audiences in to customers. This project leverages analytic skills on cover call volume trends with the CX team yielding valuable insights for the business.

The Problems

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

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

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

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

Design

Descriptive Statistics: Descriptive statistics techniques were likely used to summarize the main characteristics of the call volume data. These statistics could include measures like average call duration, total call volume, maximum and minimum call times, and call status distribution (e.g., the percentage of abandoned calls).

Data Visualization: Excel's charting and graphing capabilities were likely utilized to create visual representations of the data. Charts and graphs can help you visually identify trends and patterns in call volume over time, by time bucket, or in relation to other factors.

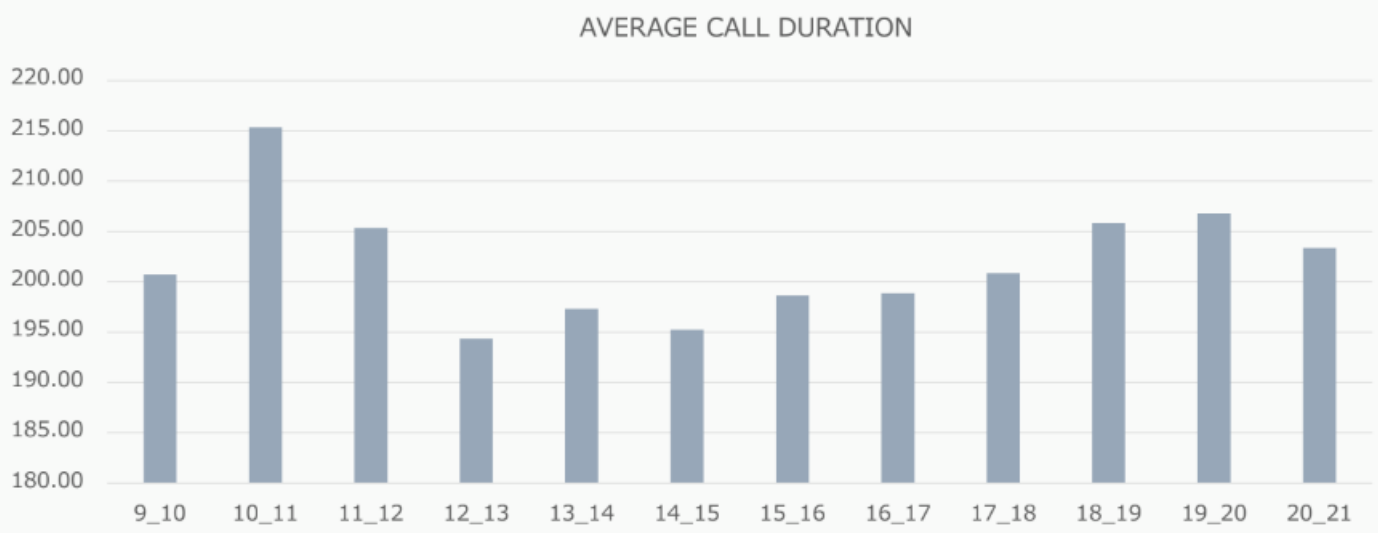
Data Filtering and Sorting: Excel's data filtering and sorting capabilities can be handy for isolating specific subsets of data, such as calls of a certain status (abandoned, answered) or calls during particular time periods.

Pivot Tables: Pivot tables in Excel are useful for summarizing and aggregating data, making it easier to analyze call volume metrics from different perspectives and dimensions.

Report Generation: Excel's capabilities for creating reports, tables, and charts were likely used to present your findings and insights in a clear and structured manner.

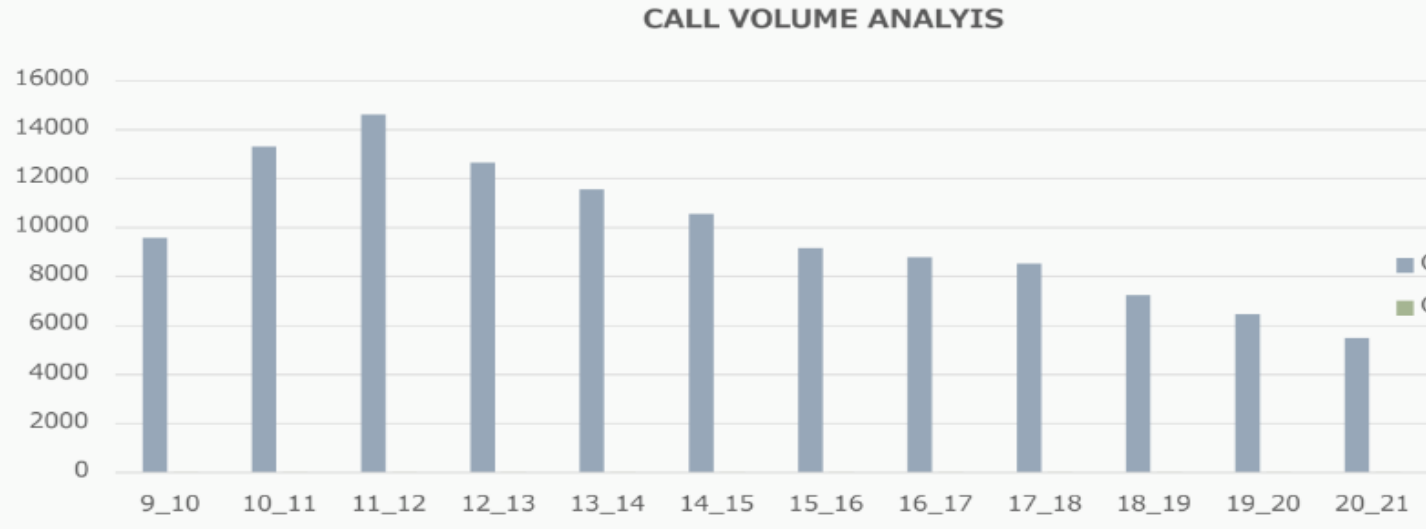
Findings

Average Call Duration



Call Volume Analysis

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



Findings

Manpower Planning

time bucket	Count of Call_Seconds (s)	Count of Call_Seconds (s)2	Agent Required
9_10	8.13%	0.08	5
10_11	11.28%	0.11	7
11_12	12.40%	0.12	8
12_13	10.72%	0.11	7
13_14	9.80%	0.10	6
14_15	8.95%	0.09	6
15_16	7.76%	0.08	5
16_17	7.45%	0.07	5
17_18	7.23%	0.07	5
18_19	6.13%	0.06	4
19_20	5.48%	0.05	3
20_21	4.67%	0.05	3
Grand Total	**100.00%**	**1.00**	**63**

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 customer makes between 9 AM to 9 PM they also make 30 calls at nights between 9 PM to 9 AM the distribution of these calls is follows:

average calls per day	5130
Night calls (9PM-9AM)	1539
hours	77
agents	17

Time bucket	Distribution of calls	time distribution	Agents Required
9PM-10PM	3	10	2
10PM-11PM	3	10	2
11PM-12AM	2	15	1
12AM-1AM	2	15	1
1AM-2AM	1	30	1
2AM-3AM	1	30	1
3AM-4AM	1	30	1
4AM-5AM	1	30	1
5AM-6AM	3	10	2
6AM-7AM	4	7.5	2
7AM-8AM	4	7.5	2
8AM-9AM	5	6	3
TOTAL	**30**	**201**	**17**

Results

Throughout this project, I have gained valuable insights in to ABC call volume trend analysis. I now understand the nuances of call duration patterns across different time buckets, allowing for precise resource allocation and enhance customer satisfaction. Visualizing call volume trends has provided a clarity on when peak call times occur, enabling effective staffing strategies. Additionally, by calculating the minimum required agents for each time bucket, we have a concrete plan to reduce the abundant rate to 10% , ultimately leading to the better customer service. Addressing night time calls with a dedicated manpower plan ensures the round- the-clock support. This project simplified the practicality of excel for data analysis, showcasing its role in data-driven decision making in the sum, these insights contribute to improve custom experience and operational efficiency within the inbound calling team at ABC company.

https://drive.google.com/file/d/1QVq75dCUOxWLVJIzO_thP3mQuEp7jppq/view?usp=drive_link

Conclusion

These projects require the use of various technical skills, including data analysis, data cleaning, data visualization, and the use of tools like Excel and SQL. The student will gain practical experience in using these tools and techniques, which are highly valuable in the field of computer science and data analysis.

Each project presents specific challenges and questions that need to be answered. Working through these problems enhances the student's problem-solving abilities, a crucial skill in computer science.

Data is at the core of these projects. Learning to work with data, clean it, analyze it, and draw insights from it is a fundamental skill in data science and computer science in general.

Some projects involve statistical analysis, which can help the student understand how to use statistics to draw conclusions from data. This knowledge is not only useful in data analysis but also in research and other areas of computer science.

Depending on the project, the student may gain insights into various domains such as customer experience, marketing, finance, or the automotive industry. This exposure to different domains broadens their understanding and can be applied in future projects or job roles.

Many of these projects involve presenting findings and insights to others, which enhances communication skills. This is valuable for explaining complex technical concepts to non-technical stakeholders.

Completing these projects provides concrete examples of work that can be included in a portfolio or resume. This is crucial when applying for internships or jobs in the future.

Analyzing data and drawing insights require critical thinking skills. Working on these projects helps students develop the ability to think critically and make informed decisions based on data.

Managing a project from start to finish involves time management skills. Students learn how to set deadlines, allocate time to different tasks, and ensure that the project is completed on time.

These projects simulate real-world scenarios, helping students understand how data analysis is applied in various industries and situations. This practical experience is invaluable.

Appendix

1	INSTAGRAM USER ANALYTICS
2	OPERATIONAL ANALYTICS AND INVESTIGATING METRIC SPIKE
3	HIRING PROCESS ANALYSIS

4	IMDB MOVIE ANALYSIS
5	BANK LOAN CASE STUDY
6	IMPACT OF CAR FEATURES ON PRICE AND PROFITABILITY
7	ABC CALL VOLUME ANALYSIS