Data Analysis

Portfolio

Created by: M Abhiram

Contents

Professional Background.

I hold a B.E. in Electronics and Communication Engineering, graduating in 2019 from VTU. With one year of professional experience in Informatica Power Center at CBSI, I have developed strong technical skills in data integration and ETL processes using Informatica Power Center, SQL, and Excel.

In addition to my experience, I have expanded my skill set to include advanced Excel, Power BI, Python, and MySQL. I am eager to apply these skills in a challenging environment where I can continue to grow and contribute to impactful projects. My background in data management and analytics, coupled with my adaptability and eagerness to learn, positions me well for new opportunities in the field.

Instagram User Analytics

Project Description:

The purpose of this project is to analyse the data within Instagram to provide valuable insights that aid in the business's growth.

The Problem

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

User analysis involves tracking how users engage with a digital product, such as a software application or a mobile app. The insights derived from this analysis can be used by various teams within the business. For example, the marketing team might use these insights to launch a new campaign, the product team might use them to decide on new features to build, and the development team might use them to improve the overall user experience.

Approach

To achieve this, I utilized my SQL skills to extract the necessary insights in alignment with the project's requirements. I employed SQL techniques ranging from basic syntax such as the SELECT statement to more advanced concepts like subqueries and joins in some of my queries.

Tech-Stack Used:

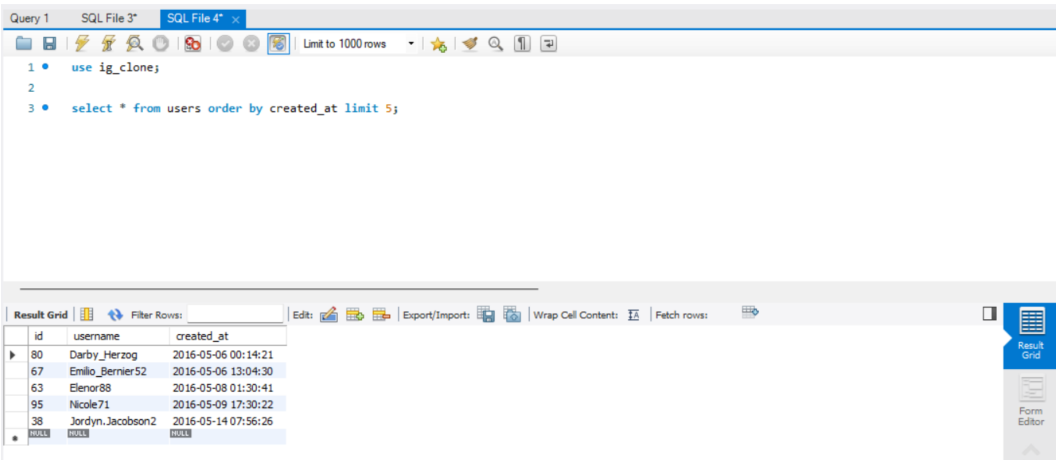
The technology I utilized for this project is MySQL Workbench, which provided a robust platform for data extraction, analysis, and visualization.

Findings

<u>Tasks:</u>

A. Marketing Analysis:

1. **Loyal User Reward**:
   To fulfil the business team's request for identifying the top 5 users who have been using the platform for the longest duration, I utilized a basic SELECT statement along with the LIMIT function to retrieve the desired information based on the account creation date.
   Query used:
   select * from users order by created_at limit 5;



   Result:
   From the above result we can see the oldest users on the platform.

2. **Inactive User Engagement**:
   To identify users who have not posted a single photo since their account creation, I utilized a JOIN operation to combine data from two tables based on a common column. We have a table named users containing user details and another table named photos containing post information, where the users table has a common column user_id with the posts table.

   Query used:
   select * from users
   left join photos

on users.id = photos.user_id
where image_url is null;

```
select u.id,u.username, p.id,p.image_url from users u
left join photos p
on u.id = p.user_id
where image_url is null;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Conte

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

Result:

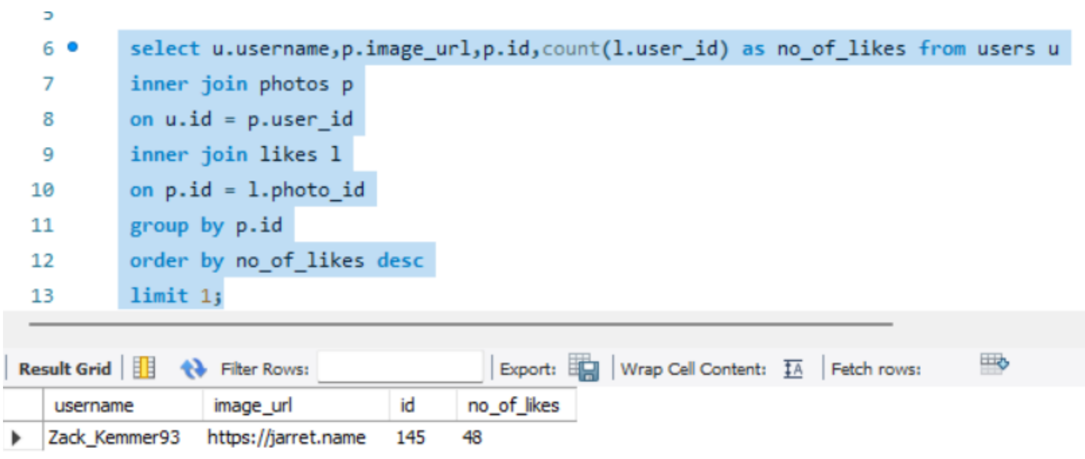
From the above result we found out that there are 26 users who didn't posted a single photo on their account.

3. **Contest Winner Declaration:**

To determine the user with the most likes for their posts, I joined three tables—users, photos, and likes—based on their common columns. I also used the aggregate function COUNT() to calculate the total number of likes for each user's posts

Query Used:

select u.username, p.image_url, p.id, count(l.user_id) as no_of_likes from users u
inner join photos p
on u.id = p.user_id
inner join likes l
on p.id = l.photo_id
group by p.id
order by no_of_likes desc
limit 1;



Result:

From the above result we can see Zack_kemmer93 is the winner of the contest because he got the 48 likes for one post.

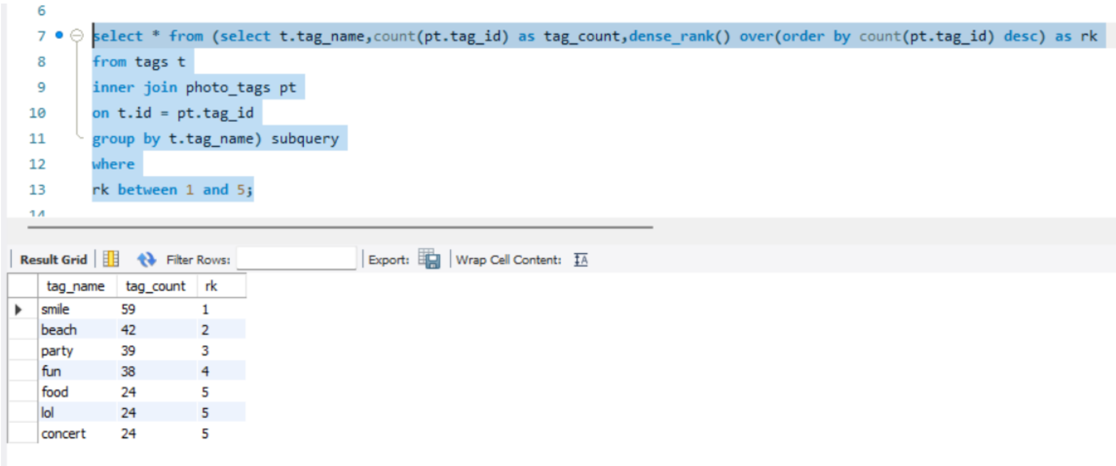
4. **Hashtag Research:**

   To identify the top 5 most used hashtags for the brand to promote their product, I combined the use of subqueries, joins, and the DENSE_RANK() function.

Query use:

select * from (select t.tag_name,count(pt.tag_id) as tag_count,dense_rank() over(order by count(pt.tag_id) desc) as rk
from tags t
inner join photo_tags pt
on t.id = pt.tag_id

group by t.tag_name) subquery
where
rk between 1 and 5;



```
select * from (select t.tag_name,count(pt.tag_id) as tag_count,dense_rank() over(order by count(pt.tag_id) desc) as rk
from tags t
inner join photo_tags pt
on t.id = pt.tag_id
group by t.tag_name) subquery
where
rk between 1 and 5;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

tag_name	tag_count	rk
smile	59	1
beach	42	2
party	39	3
fun	38	4
food	24	5
lol	24	5
concert	24	5

Result:

From above result I can say the Smile, Beach, Party, Fun, Food, LOL, Concert are the most used hastags.

5. **Ad Campaign Launch:**

To determine the best days to promote an ad campaign based on user registration patterns, I analyzed the data to find out on which days the most users registered. By using the DAYNAME function to convert the registration date into the day of the week and then counting the number of registrations for each day, I can identify the optimal days for ad promotion.

Query used:

select dayname(created_at) as day_name,
count(*) as registered_date
from users
group by dayname(created_at);

```sql
select dayname(created_at) as day_name,
count(*) as registered_date
from users
group by dayname(created_at);
```

Result Grid | Filter Rows: | Export: | Wrap Cell

day_name	registered_date
Thursday	16
Sunday	16
Tuesday	14
Saturday	12
Wednesday	13
Monday	14
Friday	15

Result:

From by above result we can do ad campaign on Thursday and Sunday.

B. Investor Metrics:

1. User Engagement:

To calculate the average number of posts per user on Instagram, and to provide the total number of photos divided by the total number of users

Query used:

```sql
WITH
photo_counts AS (
   SELECT user_id,COUNT(*) AS no_of_posts
   FROM photos
   GROUP BY
      user_id
),
total_counts AS (
   SELECT COUNT(*) AS total_photos,COUNT(DISTINCT user_id) AS
total_users
   FROM photos
)
SELECT
   (SELECT AVG(no_of_posts) FROM photo_counts) AS
avg_posts_per_user,
   (SELECT total_photos FROM total_counts) / (SELECT total_users
FROM total_counts) AS avg_photos_per_user;
```

```sql
WITH
photo_counts AS (
    SELECT user_id,COUNT(*) AS no_of_posts
    FROM photos
    GROUP BY
        user_id
),
total_counts AS (
    SELECT COUNT(*) AS total_photos,COUNT(DISTINCT user_id) AS total_users
    FROM photos
)
SELECT
    (SELECT AVG(no_of_posts) FROM photo_counts) AS avg_posts_per_user,
    (SELECT total_photos FROM total_counts) / (SELECT total_users FROM total_counts) AS avg_photos_per_user;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

avg_posts_per_user	avg_photos_per_user
3.4730	3.4730

Result:

From the above result we can see the average posts per user and average photos per users.

2. **Bots & Fake Accounts:**

   To identify potential fake or dummy accounts, I identified users who liked every photo on the platform. Assuming that any user who has liked every single photo is likely to be a fake or dummy account,

   Query used:

   select u.username,l.user_id,
   count(l.photo_id) as no_of_likes from users u
   inner join likes l
   on u.id = l.user_id
   group by user_id
   having no_of_likes = 257
   order by no_of_likes desc;

```sql
select  u.username,l.user_id, count(l.photo_id) as no_of_likes from users u
inner join likes l
on u.id = l.user_id
group by user_id
having no_of_likes = 257
order by no_of_likes desc;
```

esult Grid | Filter Rows: | Export: | Wrap Cell Content:

username	user_id	no_of_likes
Aniya_Hackett	5	257
Jaclyn81	14	257
Rocio33	21	257
Maxwell.Halvorson	24	257
Ollie_Ledner37	36	257
Mckenna17	41	257
Duane60	54	257
Julien_Schmidt	57	257
Mike.Auer39	66	257
Nia_Haag	71	257
Leslie67	75	257
Janelle.Nikolaus81	76	257
Bethany20	91	257

Result:

From the above result we found out the 13 dummy accounts who liked every single photo posted.

Operation Analytics and Investigating Metric Spike

Project Description:

The purpose of this project is to analyze the data within Operation Analytics and investigate metric spikes.

The Problem

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, you'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, you'll need to answer these questions daily, making it crucial to understand how to investigate these metric spikes.

Approach

To achieve this, I utilized my SQL skills to extract the necessary insights in alignment with the project's requirements. I employed advanced SQL techniques, including window functions, CASE statements, and CTEs (common table expressions).

Tech-Stack Used:

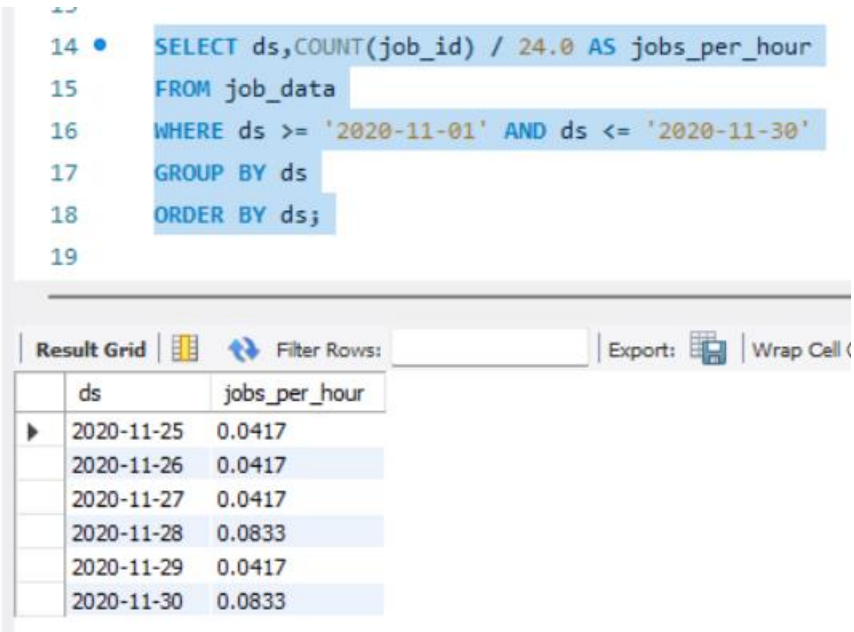
The technology I utilized for this project is MySQL Workbench, which provided a robust platform for data extraction, analysis, and visualization.

Findings

From doing this project, I gained confidence in using subqueries, joins, aggregate functions, and window functions. To achieve this, I first worked out the expected data in Excel, determining the kind of data I wanted to show. Once I confirmed the expected data in Excel, I applied the appropriate SQL concepts to get the actual results.

Case study 1: Job Data Analysis

1. Jobs Reviewed Over Time:
   In this, we need to find the jobs that were reviewed every day in November 2020.
   Query used: SELECT ds,COUNT(job_id) / 24.0 AS jobs_per_hour
   FROM job_data
   WHERE ds >= '2020-11-01' AND ds <= '2020-11-30'
   GROUP BY ds
   ORDER BY ds;



```
SELECT ds,COUNT(job_id) / 24.0 AS jobs_per_hour
FROM job_data
WHERE ds >= '2020-11-01' AND ds <= '2020-11-30'
GROUP BY ds
ORDER BY ds;
```

ds	jobs_per_hour
2020-11-25	0.0417
2020-11-26	0.0417
2020-11-27	0.0417
2020-11-28	0.0833
2020-11-29	0.0417
2020-11-30	0.0833

Result:

In the above result we can see the jobs reviewed per hour in that day

2. Throughput Analysis:

   In this we need to calculate the 7-day rolling average of throughput

   Query Used:

```
WITH daily_throughput AS (
   SELECT ds, COUNT(*) AS jobs_per_day
   FROM job_data
   GROUP BY ds
),
rolling_avg AS (
   SELECT ds, jobs_per_day,AVG(jobs_per_day) OVER (ORDER BY ds ROWS BETWEEN 6 PRECEDING AND CURRENT ROW) AS rolling_avg_7_days
   FROM daily_throughput
)
SELECT ds, jobs_per_day,rolling_avg_7_days
FROM rolling_avg
ORDER BY ds;
```

```
WITH daily_throughput AS (
    SELECT ds, COUNT(*) AS jobs_per_day
    FROM job_data
    GROUP BY ds
),
rolling_avg AS (
    SELECT ds, jobs_per_day,AVG(jobs_per_day) OVER (ORDER BY ds ROWS BETWEEN 6 PRECEDING AND CURRENT ROW) AS rolling_avg_7_days
    FROM daily_throughput
)
SELECT ds, jobs_per_day,rolling_avg_7_days
FROM rolling_avg
ORDER BY ds;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

ds	jobs_per_day	rolling_avg_7_days
2020-11-25	1	1.0000
2020-11-26	1	1.0000
2020-11-27	1	1.0000
2020-11-28	2	1.2500
2020-11-29	1	1.2000
2020-11-30	2	1.3333

Result:

I prefer using the 7-day rolling average for throughput because it provides a more stable and reliable metric. It smooths out the daily volatility and helps in identifying longer-term trends, which are crucial for strategic planning and decision-making. While the daily metric is useful for identifying specific events, the rolling average is better for understanding overall performance and trends over time.

3. Language Share Analysis:
   We need to find out the percentage share of the language in that month
   Query Used:

```
WITH last_30_days AS (
   SELECT *
   FROM job_data
),
language_counts AS (
   SELECT language, COUNT(*) AS language_count
   FROM last_30_days
   GROUP BY language
),
total_jobs AS (
   SELECT COUNT(*) AS total_count
   FROM last_30_days
)
SELECT lc.language, lc.language_count, (lc.language_count /
tj.total_count) * 100 AS percentage_share
FROM language_counts lc
CROSS JOIN total_jobs tj
ORDER BY percentage_share DESC;
```

```sql
WITH language_counts AS (
    SELECT language, COUNT(*) AS language_count
    FROM job_data
    GROUP BY language
),
total_jobs AS (
    SELECT COUNT(*) AS total_count
    FROM job_data
)
SELECT lc.language, lc.language_count, (lc.language_count / tj.total_count) * 100 AS percentage_share
FROM language_counts lc
CROSS JOIN total_jobs tj
ORDER BY percentage_share DESC;
```

sult Grid | Filter Rows: | Export: | Wrap Cell Content:

language	language_count	percentage_share
Persian	3	37.5000
English	1	12.5000
Arabic	1	12.5000
Hindi	1	12.5000
French	1	12.5000
Italian	1	12.5000

Result:

In the above result, we can see the language percentage for the month of November

4. Duplicate Rows Detection:

   In the given data, we need to determine whether there are any duplicate records.

   Query Used:

```
select
ds,job_id,actor_id,event,language,time_spent,org,count(*)
 from job_data
 group by ds,job_id,actor_id,event,language,time_spent,org
 having count(*) > 1;
```

```
select ds,job_id,actor_id,event,language,time_spent,org,count(*)
from job_data
group by ds,job_id,actor_id,event,language,time_spent,org
having count(*) > 1;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

ds	job_id	actor_id	event	language	time_spent	org	count(*)

Result:

From the above result we can find that there are no duplicate records present in the table.

<u>Case Study 2:</u>

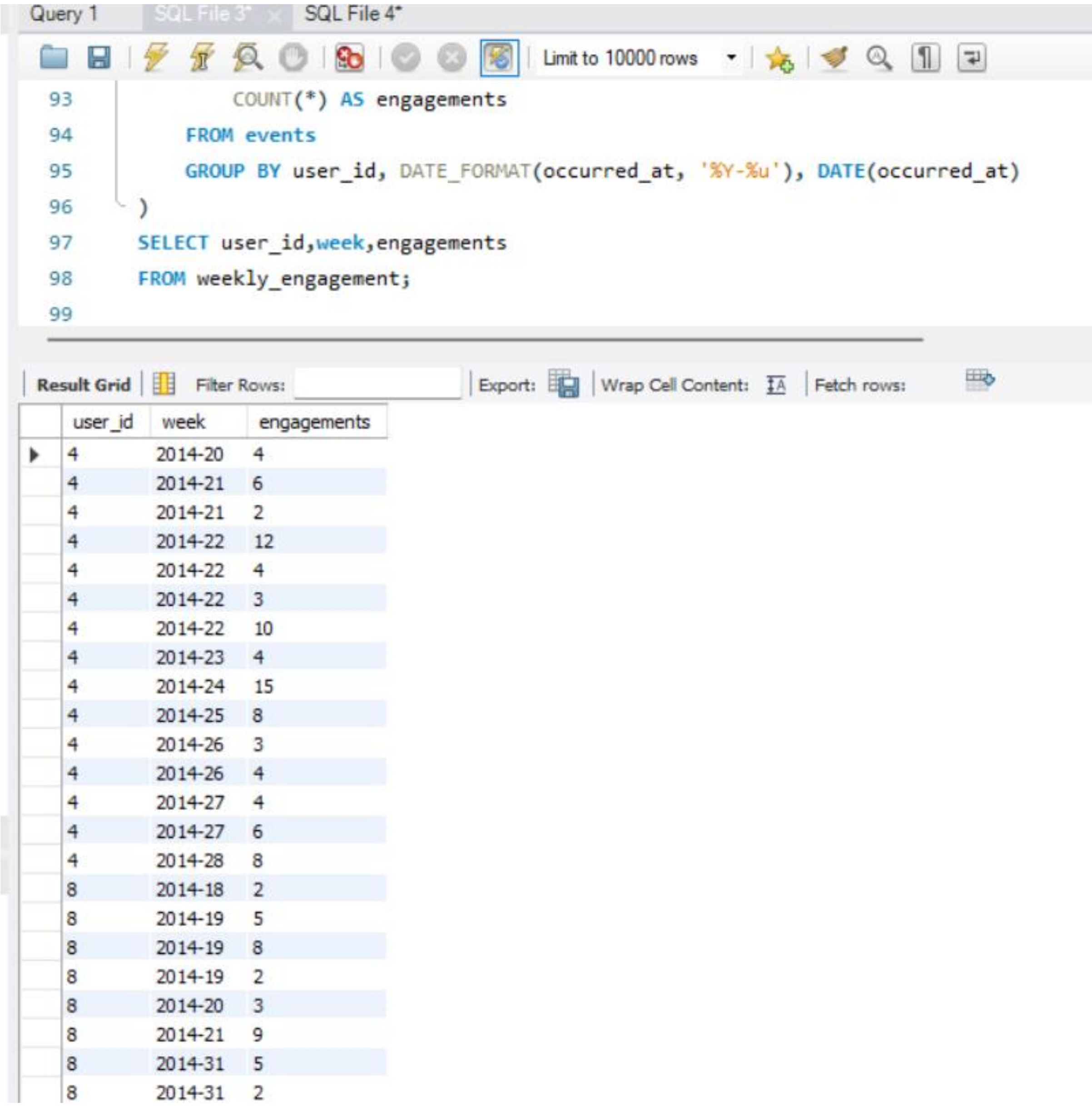
1. Weekly User Engagement:

   In the given table we need to find the no of user engagements in the week per user.

Query Used:

```
WITH weekly_engagement AS (
  SELECT
    user_id,
    DATE(occurred_at) AS date,
    DATE_FORMAT(occurred_at, '%Y-%u') AS week,
    COUNT(*) AS engagements
  FROM events
```

```
    GROUP BY user_id, DATE_FORMAT(occurred_at, '%Y-%u'),
DATE(occurred_at)
)
SELECT user_id,week,engagements
FROM weekly_engagement;
```



Result:

From the above result, we can see each user's engagement for every week of that year.

2. User Growth Analysis:

We need to find the growth analysis of the product in the engagement section per month

Query used:

```
WITH first_occurrence AS (
   SELECT user_id,
      MIN(DATE(occurred_at)) AS first_interaction
   FROM events
   GROUP BY user_id
),
monthly_new_users AS (
   SELECT DATE_FORMAT(first_interaction, '%Y-%m') AS y_m,
      COUNT(*) AS new_users
   FROM first_occurrence
   GROUP BY DATE_FORMAT(first_interaction, '%Y-%m')
   ORDER BY y_m
)
SELECT y_m,new_users,
   LAG(new_users) OVER (ORDER BY y_m) AS
previous_month_users,
   ((new_users - LAG(new_users) OVER (ORDER BY y_m)) /
LAG(new_users) OVER (ORDER BY y_m)) * 100 AS growth
FROM
   monthly_new_users;
```

```sql
WITH first_occurrence AS (
    SELECT user_id,
        MIN(DATE(occurred_at)) AS first_interaction
    FROM events
    GROUP BY user_id
),
monthly_new_users AS (
    SELECT DATE_FORMAT(first_interaction, '%Y-%m') AS y_m,
        COUNT(*) AS new_users
    FROM first_occurrence
    GROUP BY DATE_FORMAT(first_interaction, '%Y-%m')
    ORDER BY y_m
)
SELECT y_m,new_users,
    LAG(new_users) OVER (ORDER BY y_m) AS previous_month_users,
    ((new_users - LAG(new_users) OVER (ORDER BY y_m)) / LAG(new_users) OVER (ORDER BY y_m)) * 100 AS growth
FROM
    monthly_new_users;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

y_m	new_users	previous_month_users	growth
2014-05	2361	NULL	NULL
2014-06	1362	2361	-42.3126
2014-07	1292	1362	-5.1395
2014-08	1127	1292	-12.7709

Result:

From the above result we can see the grow of the prodect in each month and percentage of the growth when compared with previous month users.

3. Weekly Retention Analysis:

    We need to find the weekly retention of the users after signing up for the product

Query Used:

WITH user_cohorts AS (
  SELECT
    user_id,
    DATE(created_at) AS sign_up_date,
    DATE_SUB(created_at, INTERVAL (DAYOFWEEK(created_at) - 1) DAY) AS cohort_week
  FROM users
),
user_events AS (

```
    SELECT
        user_id,
        DATE(occurred_at) AS event_date,
        DATE_SUB(occurred_at, INTERVAL
(DAYOFWEEK(occurred_at) - 1) DAY) AS event_week
    FROM events
),
weekly_retention AS (
    SELECT
        uc.cohort_week,
        ue.event_week,
        COUNT(DISTINCT ue.user_id) AS retained_users
    FROM user_cohorts uc
    JOIN user_events ue ON uc.user_id = ue.user_id
    WHERE ue.event_week >= uc.cohort_week
    GROUP BY uc.cohort_week, ue.event_week
)
SELECT
    cohort_week,
    event_week,
    retained_users,
    TIMESTAMPDIFF(WEEK, cohort_week, event_week) AS
week_number
FROM weekly_retention
ORDER BY cohort_week, week_number;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

cohort_week	event_week	retained_users	week_number
2014-04-20 08:17:00	2014-04-27 03:16:00	1	0
2014-04-20 08:17:00	2014-04-27 03:17:00	1	0
2014-04-20 08:17:00	2014-04-27 03:18:00	1	0
2014-04-20 09:14:00	2014-04-27 08:50:00	1	0
2014-04-20 09:14:00	2014-04-27 08:51:00	1	0
2014-04-20 09:14:00	2014-04-27 08:52:00	1	0
2014-04-20 09:55:00	2014-04-27 09:06:00	1	0
2014-04-20 09:55:00	2014-04-27 09:07:00	1	0
2014-04-20 10:46:00	2014-04-27 10:15:00	1	0
2014-04-20 10:46:00	2014-04-27 10:16:00	1	0
2014-04-20 12:32:00	2014-04-27 11:13:00	1	0
2014-04-20 12:32:00	2014-04-27 11:14:00	1	0
2014-04-20 12:36:00	2014-04-27 12:00:00	1	0
2014-04-20 12:36:00	2014-04-27 12:01:00	1	0
2014-04-20 12:36:00	2014-04-27 12:02:00	1	0
2014-04-20 12:47:00	2014-04-27 07:32:00	1	0
2014-04-20 12:47:00	2014-04-27 07:33:00	1	0
2014-04-20 12:47:00	2014-04-27 07:34:00	1	0
2014-04-20 12:47:00	2014-04-27 07:35:00	1	0
2014-04-20 13:03:00	2014-04-27 08:49:00	1	0
2014-04-20 13:03:00	2014-04-27 08:50:00	1	0
2014-04-20 13:03:00	2014-04-27 08:51:00	1	0
2014-04-20 13:03:00	2014-04-27 08:52:00	1	0

Result:

From the above result we can see the weekly retention of the users after signing up the product

4. Weekly Engagement Per Device:

    We need to find the no of weekly engagement of the device in that month

Query used:

select device,date_format(occurred_at, '%Y-%u') as weekly,dayname(occurred_at) as week_name,count(device) as no_of_device

from events

group by device,date_format(occurred_at, '%Y-%u'),dayname(occurred_at)

order by device;

```
select device,date_format(occurred_at, '%Y-%u') as weekly,dayname(occurred_at) as week_name,count(device) as no_of_device
from  events
group by device,date_format(occurred_at, '%Y-%u'),dayname(occurred_at)
order by  device;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

device	weekly	week_name	no_of_device
acer aspire desktop	2014-18	Friday	23
acer aspire desktop	2014-18	Saturday	8
acer aspire desktop	2014-18	Sunday	2
acer aspire desktop	2014-18	Thursday	38
acer aspire notebook	2014-18	Friday	113
acer aspire notebook	2014-18	Saturday	27
acer aspire notebook	2014-18	Sunday	8
acer aspire notebook	2014-18	Thursday	67
amazon fire phone	2014-18	Friday	47
amazon fire phone	2014-18	Saturday	8
amazon fire phone	2014-18	Thursday	29
asus chromebook	2014-18	Friday	103
asus chromebook	2014-18	Saturday	56
asus chromebook	2014-18	Sunday	32
asus chromebook	2014-18	Thursday	95
dell inspiron desktop	2014-18	Friday	94
dell inspiron desktop	2014-18	Saturday	43
dell inspiron desktop	2014-18	Sunday	10
dell inspiron desktop	2014-18	Thursday	51
dell inspiron noteb...	2014-18	Friday	220
dell inspiron noteb...	2014-18	Saturday	65
dell inspiron noteb...	2014-18	Sunday	63
dell inspiron noteb...	2014-18	Thursday	221
hp pavilion desktop	2014-18	Friday	59
hp pavilion desktop	2014-18	Saturday	30

Result:

From the above result, we can find the number of devices that are engaged weekly in the month

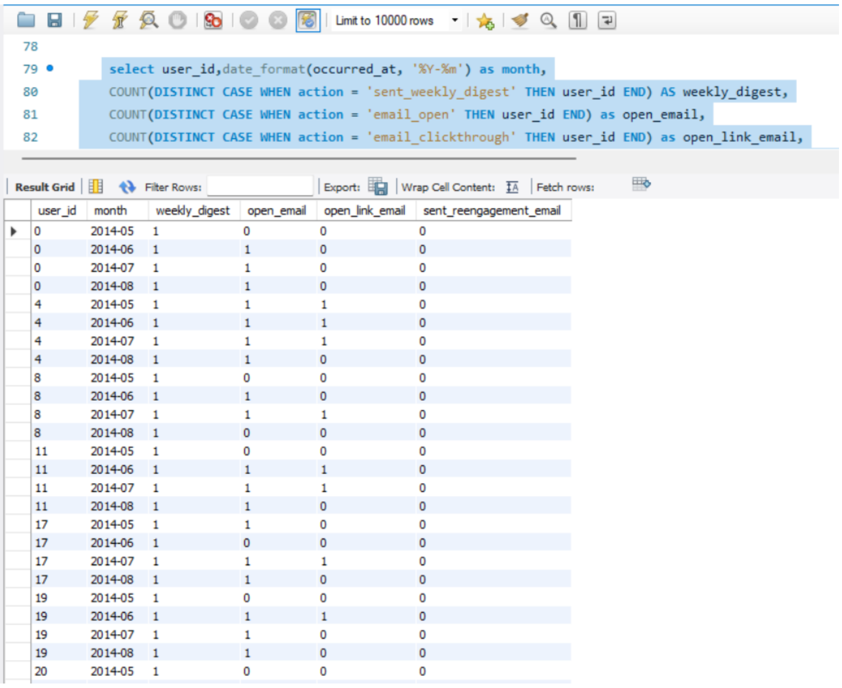
5. Email Engagement Analysis:

    We need to Analyze how users are engaging with the email service.

Query Used:

select user_id,date_format(occurred_at, '%Y-%m') as month,
COUNT(DISTINCT CASE WHEN action = 'sent_weekly_digest' THEN user_id END) AS weekly_digest,
COUNT(DISTINCT CASE WHEN action = 'email_open' THEN user_id END) as open_email,
COUNT(DISTINCT CASE WHEN action = 'email_clickthrough' THEN user_id END) as open_link_email,

COUNT(DISTINCT CASE WHEN action = 'sent_reengagement_email' THEN user_id END) as sent_reengagement_email
from email_events
group by user_id,date_format(occurred_at, '%Y-%m');



```
78
79 ●  select user_id,date_format(occurred_at, '%Y-%m') as month,
80     COUNT(DISTINCT CASE WHEN action = 'sent_weekly_digest' THEN user_id END) AS weekly_digest,
81     COUNT(DISTINCT CASE WHEN action = 'email_open' THEN user_id END) as open_email,
82     COUNT(DISTINCT CASE WHEN action = 'email_clickthrough' THEN user_id END) as open_link_email,
```

user_id	month	weekly_digest	open_email	open_link_email	sent_reengagement_email
0	2014-05	1	0	0	0
0	2014-06	1	1	0	0
0	2014-07	1	1	0	0
0	2014-08	1	1	0	0
4	2014-05	1	1	1	0
4	2014-06	1	1	1	0
4	2014-07	1	1	1	0
4	2014-08	1	1	0	0
8	2014-05	1	0	0	0
8	2014-06	1	1	0	0
8	2014-07	1	1	1	0
8	2014-08	1	0	0	0
11	2014-05	1	0	0	0
11	2014-06	1	1	1	0
11	2014-07	1	1	1	0
11	2014-08	1	1	0	0
17	2014-05	1	1	0	0
17	2014-06	1	0	0	0
17	2014-07	1	1	1	0
17	2014-08	1	1	0	0
19	2014-05	1	0	0	0
19	2014-06	1	1	1	0
19	2014-07	1	1	0	0
19	2014-08	1	1	0	0
20	2014-05	1	0	0	0

Result:

From the above result, we can see the users who are using email engagements per week in the month

Hiring Process Analytics

Project Description:

This analysis focuses on the hiring process, providing insights into the number of candidates selected, the number of candidates rejected, and the number of vacancies within the company.

The Problem

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

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

Approach

1. The first approach I took in this analysis was to filter out the rejected candidates in Excel, as the requirement was to analyze the data of only the hired candidates.
2. I discovered another approach where the post name column contained hyphens between the letters. Upon reviewing the other values in the column, I determined that this format was not useful. Therefore, I removed the hyphens and retained the post names in their original form.
3. In the same post name column, I found a blank value for one of the selected candidates. To fill this value, I compared two aspects: the date and time of the candidate and their salary. By identifying the salary and timing of the closest comparable candidate, I filled the blank value with the post name of a rejected candidate whose salary and timing were the nearest matches.
4. In the event name column, I found a few blank entries. I filled these with the value 'Don't want to say,' which had already been used for other candidates, making it a consistent choice.
5. In the salary column, there was a blank value. I checked the interview date and found that no one was selected for that post on that day, making the date unhelpful. Therefore, I filled the blank value with the nearest salary value from a selected candidate, as I was not considering rejected candidates for this analysis.

<u>Tech-Stack Used</u>:

The tool I used for this analysis was an Excel workbook.

Findings

A. Hiring Analysis:

For this analysis, I determined the number of candidates hired based on different genders.

Result: Male-2563, Female-1856, ohers-278 candidates are hired

Gender	Hired
Male	2563
Female	1856
Others	278

B. Salary Analysis:

In this analysis, we determined the average salary in each department using the AVERAGEIF function in Excel.

Result: From the results below, we can find the average salaries in each department.

Department	Average Salary
Finance Department	48748.28
General Management	60810.20
Human Resource Department	49014.40
Marketing Department	47843.40
Operations Department	48914.19
Production Department	49350.87
Purchase Department	52086.57
Sales Department	48539.55
Service Department	50549.52

C. Salary Distribution:

In this analysis, I first identified the maximum and minimum salaries in the given salary column using the MAX and MIN functions in Excel. Then, I calculated the difference between the two salaries and divided that value by an interval number of 10. In the next column, I created a range and determined the number of employees that fell within each range.

Result: From the results below, we can see the minimum and maximum salaries using the MIN and MAX functions, as well as the total number of employees within each interval.

Functons	calculations	intervals	interval lable	No.of.employess
Maximum	400000	800	800-40719	1910
Minimum	800	40720	40720-80639	1943
difference	399200	80640	80640-120559	841
intervals	39920	120560	120560-160479	0
		160480	160480-200399	1
		200400	200400-240319	0
		240320	240320-280239	0
		280240	280240-320159	1
		320160	320160-360079	0
		360080	360080-399999	0
		400000	400000--1	1

D. Departmental Analysis:

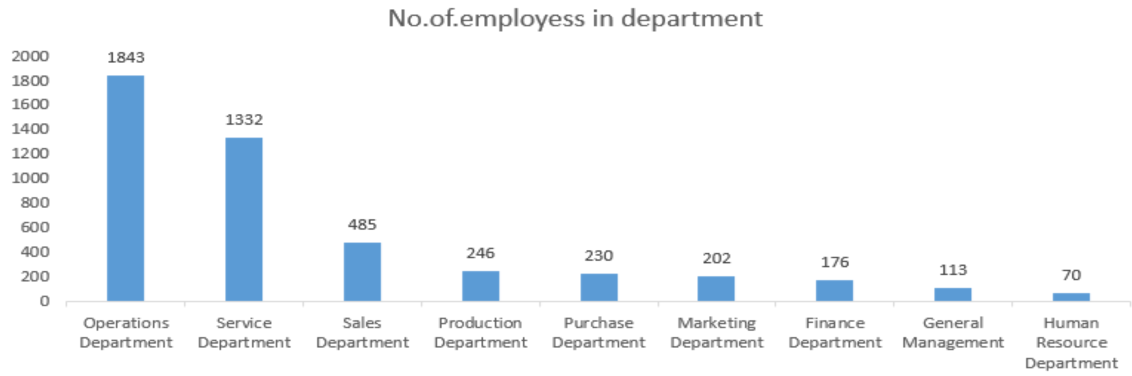
For this analysis, I used two Excel functions. First, I used the UNIQUE function to extract the unique values in the given column. Then, I used the COUNTIF function to get the count of candidates that met certain criteria. After that, I inserted a bar graph and applied a few formatting options, such as enabling data labels and disabling gridlines.

Result:

a. Tabular Format:

Department	No.of.employess
Operations Department	1843
Service Department	1332
Sales Department	485
Production Department	246
Purchase Department	230
Marketing Department	202
Finance Department	176
General Management	113
Human Resource Department	70

b. Chart Format: From the below bar chat we can see the no of candidates got hired into each department.



E. Position Tier Analysis:

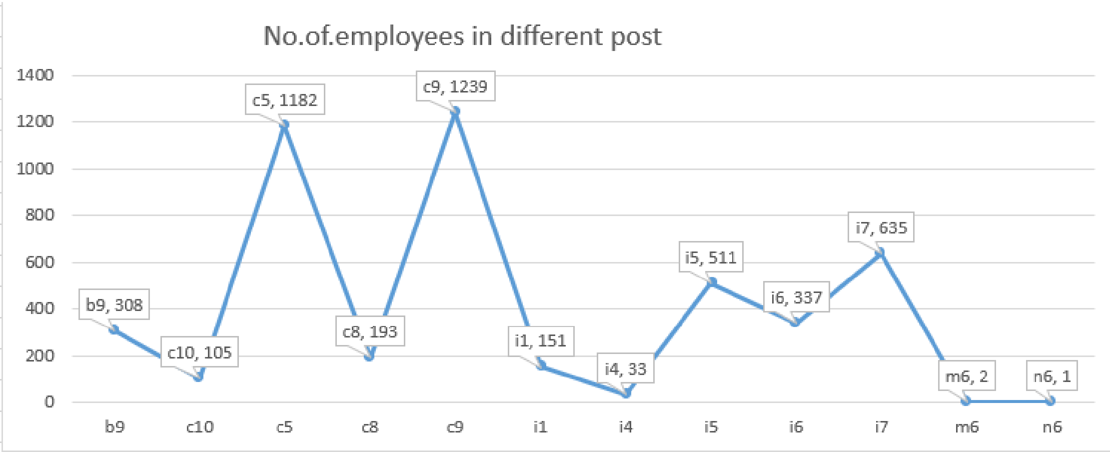
For this analysis, I used two Excel functions. First, I used the UNIQUE function to extract the unique values in the given column. Then, I used the COUNTIF function to get the count of candidates that met certain criteria. After that, I inserted a line graph and applied a few formatting options, such as enabling data labels and disabling gridlines

Result:

a. Tabular Format

Post Name	No.of.employees in different post
b9	308
c10	105
c5	1182
c8	193
c9	1239
i1	151
i4	33
i5	511
i6	337
i7	635
m6	2
n6	1

b. Chart Format: From the line graph below, we can observe the distribution of candidates across different positions in the company.



IMDB Movie Analysis

Project Description:

From the given dataset, which contains movie details and IMDb scores, we can analyze the factors that influence a movie's success on IMDb. This type of movie analysis will help producers, directors, and investors understand what makes a movie successful.

The Problem

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

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.

Approach

- To analyze the given dataset, we need to identify the types of data required for our analysis.
- After identifying the requirements, I performed several data cleaning steps.
- These steps involve:
    - Removing unnecessary columns
    - Correcting data in a few columns using the clean function in Excel
    - Removing rows with blank values, since filling a large amount of missing data can be time-consuming
    - Removing duplicates based on the movie_title column

Tech-Stack Used:

For this analysis, I used Excel for data cleaning, visualization, and analysis, and Microsoft PowerPoint for the presentation.

Insights:

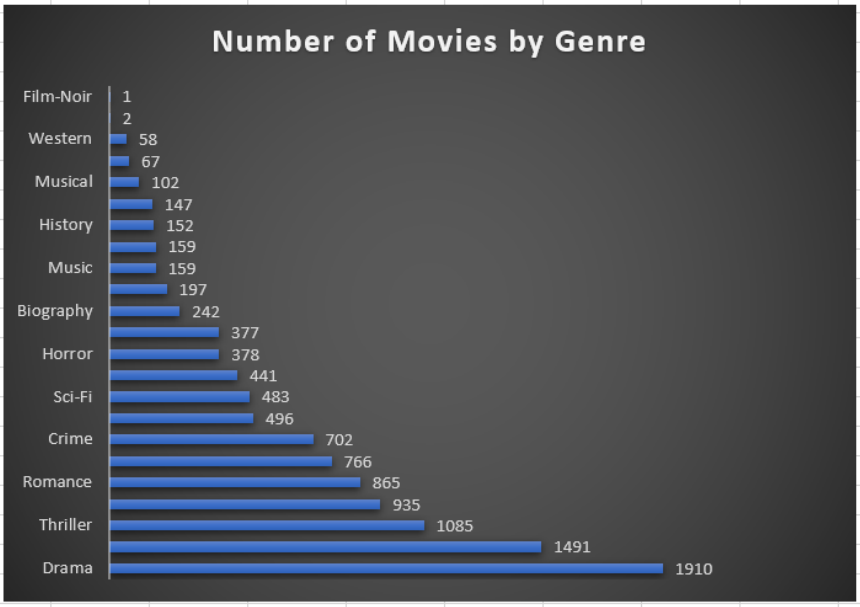
- Analyze the distribution of movie genres and their impact on IMDb scores.
- Investigating the relationship between movie duration and IMDb score using descriptive statistics and scatter plots in Excel.
- Analyzing the impact of these languages on IMDb scores using descriptive statistics.
- Analyzing their contribution to movie success using percentile calculations and visual representations.
- Identifying movies with the highest profit margin by calculating the profit margin (gross earnings - budget) and using Excel's MAX function.

Findings

1. Movie Genre Analysis:
    - Identify and Separate Genres:
    - Examined the 'genres' column to identify how genres are listed. In this movie have multiple genres listed in a single cell, separated them into individual genres.
    - Used Excel's Text to Columns feature to separate multiple genres into different columns or rows as needed.
    - Utilized Excel's COUNTIF function to count the occurrences of each genre. This can be done by creating a list of unique genres and using COUNTIF to tally how many times each genre appears in the dataset.
    - Used Excel's statistical functions to calculate the following descriptive statistics for the IMDb scores:
    - Mean (Average): =AVERAGE(range)
    - Median: =MEDIAN(range)
    - Mode: =MODE.SNGL(range)
    - Max: =MAX(range)
    - Min: =MIN(range)
    - Variance: =VAR.P(range)

Unique Ge	Count of M
Drama	1910
Comedy	1491
Thriller	1085
Action	935
Romance	865
Adventure	766
Crime	702
Fantasy	496
Sci-Fi	483
Family	441
Horror	378
Mystery	377
Biography	242
Animation	197
Music	159
War	159
History	152
Sport	147
Musical	102
Document	67
Western	58
Short	2
Film-Noir	1

Descriptive Satistics						
AVERAGE	MEDIAN	MODE	MAX	MIN	VAR	STDEV
6.461906	6.6	6.7	9.3	1.6	1.118351	1.057521

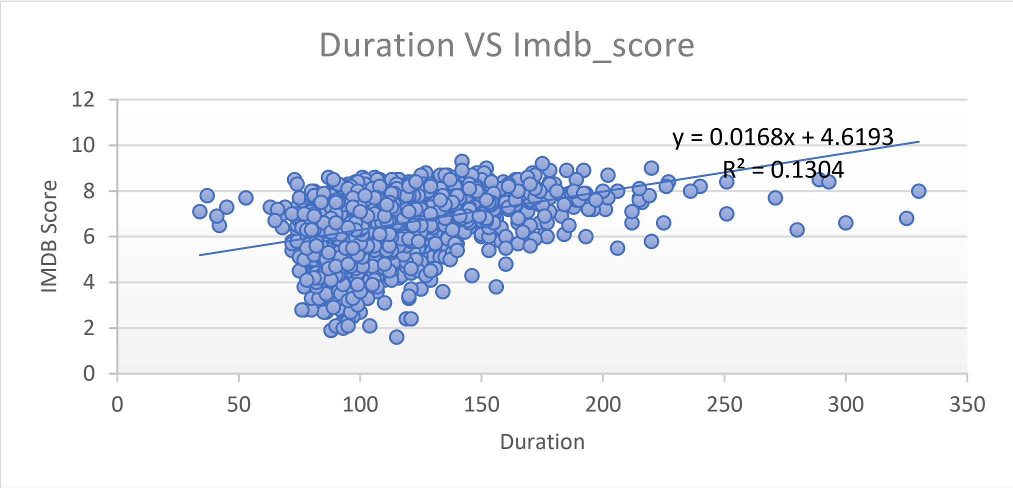


2. Movie Duration Analysis:
   - Created another sheet in excel to ensure the dataset includes columns for movie durations and their corresponding IMDb scores.
   - Calculated Descriptive Statistics
   - Mean Duration: Calculated the average movie duration using Excel's AVERAGE function.
   - Median Duration: Calculated the median movie duration using Excel's MEDIAN function
   - Standard Deviation of Durations: Calculated the standard deviation of movie durations using Excel's STDEV function.

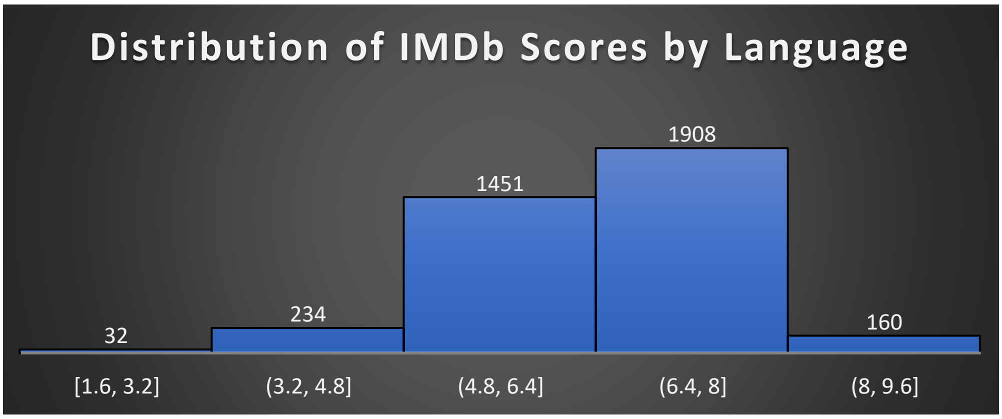
Descriptive statistics	
Average	109.8028
Median	105
STDEV	22.75721

   - To Create a Scatter Plot
   - Select Data: Highlight the columns containing movie durations and IMDb scores.
   - Insert Scatter Plot: Navigate to the Insert tab, click on Scatter, and choose the appropriate scatter plot type (e.g., scatter with markers only).

- Insert Trendline: Right-click on any data point within the scatter plot.
- Checkd the option to display the equation and R-squared value on the chart to assess the relationship's strength and direction.



3. Language Analysis:
   - Created another sheet in excel to ensure that dataset includes columns for movie languages and their corresponding IMDb scores.
   - Created a pivot table on the same sheet, selecting various ranges to calculate the count of languages, average IMDb score, median IMDb score, and standard deviation of IMDb score.
   - The below histogram chart represent the Distribution of IMDb Scores by Language.



   - Below is a table from the pivot table displaying counts and statistical calculations.

Row Labels	Count of language	Average of imdb_score	Median of IMDB Scores	StdDevp of imdb_score
Aboriginal	2	6.95	6.6	0.55

Arabic	1	7.2	6.6	0
Aramaic	1	7.1	6.6	0
Bosnian	1	4.3	6.6	0
Cantonese	8	7.2375	6.6	0.412121038
Czech	1	7.4	6.6	0
Danish	3	7.9	6.6	0.43204938
Dari	2	7.5	6.6	0.1
Dutch	3	7.566666667	6.6	0.329983165
Dzongkha	1	7.5	6.6	0
English	3606	6.421436495	6.6	1.052352956
Filipino	1	6.7	6.6	0
French	37	7.286486486	6.6	0.553691378
German	13	7.692307692	6.6	0.615769111
Hebrew	3	7.5	6.6	0.355902608
Hindi	10	6.76	6.6	1.05470375
Hungarian	1	7.1	6.6	0
Icelandic	1	6.9	6.6	0
Indonesian	2	7.9	6.6	0.3
Italian	7	7.185714286	6.6	1.069617517
Japanese	12	7.625	6.6	0.861321659
Kazakh	1	6	6.6	0
Korean	4	7.875	6.6	0.414578099
Mandarin	14	7.021428571	6.6	0.737930089
Maya	1	7.8	6.6	0
Mongolian	1	7.3	6.6	0
None	1	8.5	6.6	0
Norwegian	4	7.15	6.6	0.497493719
Persian	3	8.133333333	6.6	0.449691252
Portuguese	5	7.76	6.6	0.875442745
Romanian	1	7.9	6.6	0
Russian	1	6.5	6.6	0
Spanish	26	7.05	6.6	0.810151933
Swedish	1	7.6	6.6	0
Telugu	1	8.4	6.6	0
Thai	3	6.633333333	6.6	0.368178701
Vietnamese	1	7.4	6.6	0
Zulu	1	7.3	6.6	0
Grand Total	**3785**	**6.462324967**	**6.6**	**1.057472316**

4. Director Analysis:

- Created new sheet in excel to ensure the dataset includes columns for directors and their corresponding IMDb scores.
- Inserted a pivot table to find the average of imdb scores for each director
- Utilized the =PERCENTILE.INC function to determine the percentile within the IMDb score range.

- After determining the percentile value using the IF function to identify top directors by comparing with the average IMDb score of each director, filtered the column to display the top values, and inserted a chart for visual representation.
- Below is the sample screenshot from the pivot table showing average IMDb scores.

E	F
Row Labels	**Average of imdb_score**
Aaron Schneider	7.1
Aaron Seltzer	2.7
Abel Ferrara	6.6
Adam Carolla	6.1
Adam Goldberg	5.4
Adam Marcus	4.3
Adam McKay	6.916666667
Adam Rapp	6.4
Adam Rifkin	6.8
Adam Shankman	5.9625
Adrian Lyne	6.4
Adrienne Shelly	7.1
Agnieszka Holland	6.8
Agnieszka Wojtowicz-Vosloo	5.9
Aki Kaurismäki	7.2
Akira Kurosawa	8.1
Akiva Goldsman	6.2
Akiva Schaffer	5.7
Alan Cohn	6
Alan J. Pakula	6.3
Alan Metter	3.3
Alan Parker	7.033333333
Alan Poul	5.3
Alan Rudolph	4.6
Alan Shapiro	5.2
Alan Taylor	6.85

- Below is the table displaying statistical data and percentile values.

Average_of_imdb_scores	Standard deviation_of_imdb_scores	Percentail
6.462325	1.057472	8.2125

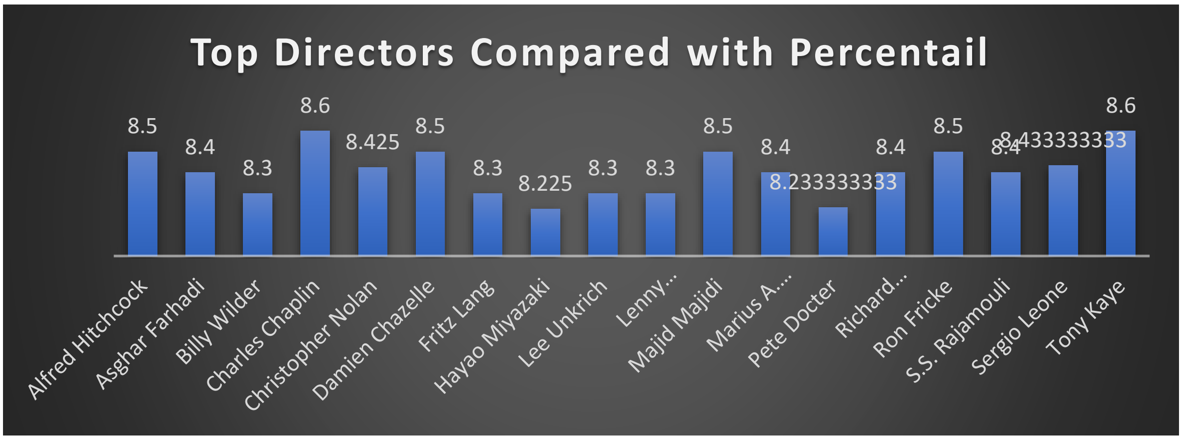
- Below is the table and chart displaying the top directors based on percentile and average IMDb score.

director_name	Average of imdb_score	Top_director
Alfred Hitchcock	8.5	Top
Asghar Farhadi	8.4	Top
Billy Wilder	8.3	Top
Charles Chaplin	8.6	Top
Christopher Nolan	8.425	Top

Table:

Damien Chazelle	8.5	Top
Fritz Lang	8.3	Top
Hayao Miyazaki	8.225	Top
Lee Unkrich	8.3	Top
Lenny Abrahamson	8.3	Top
Majid Majidi	8.5	Top
Marius A. Markevicius	8.4	Top
Pete Docter	8.233333	Top
Richard Marquand	8.4	Top
Ron Fricke	8.5	Top
S.S. Rajamouli	8.4	Top
Sergio Leone	8.433333	Top
Tony Kaye	8.6	Top

Chart:



5. Budget Analysis:
   - Created another sheet in the excel to ensure the dataset includes columns for movie budgets and gross earnings.

- Used Excel's CORREL function to calculate the correlation coefficient between movie budgets and gross earnings, below is the correlation value.

Correlation
0.223116

- Create a new column to calculate the profit margin for each movie.
- Below is the sample screenshot of the profit margin.

movie_title	budget	gross	Profit
Avatar	237000000	760505847	523505847
Pirates of the (	300000000	309404152	9404152
Spectre	245000000	200074175	-44925825
The Dark Knigh	250000000	448130642	198130642
John Carter	263700000	73058679	-190641321
Spider-Man 3	258000000	336530303	78530303
Tangled	260000000	200807262	-59192738
Avengers: Age	250000000	458991599	208991599
Harry Potter ar	250000000	301956980	51956980
Batman v Supe	250000000	330249062	80249062
Superman Retu	209000000	200069408	-8930592
Quantum of Sc	200000000	168368427	-31631573
Pirates of the (	225000000	423032628	198032628
The Lone Rang	215000000	89289910	-125710090
Man of Steel	225000000	291021565	66021565
The Chronicles	225000000	141614023	-83385977
The Avengers	220000000	623279547	403279547
Pirates of the (	250000000	241063875	-8936125
Men in Black 3	225000000	179020854	-45979146
The Hobbit: Th	250000000	255108370	5108370

- To identify the movie with the highest profit margin, we can use Excel's INDEX and MATCH functions, below is the value and formula to find the highest profit margin movie.
- Formula: =INDEX(A:A, MATCH(MAX(D:D), D:D, 0))

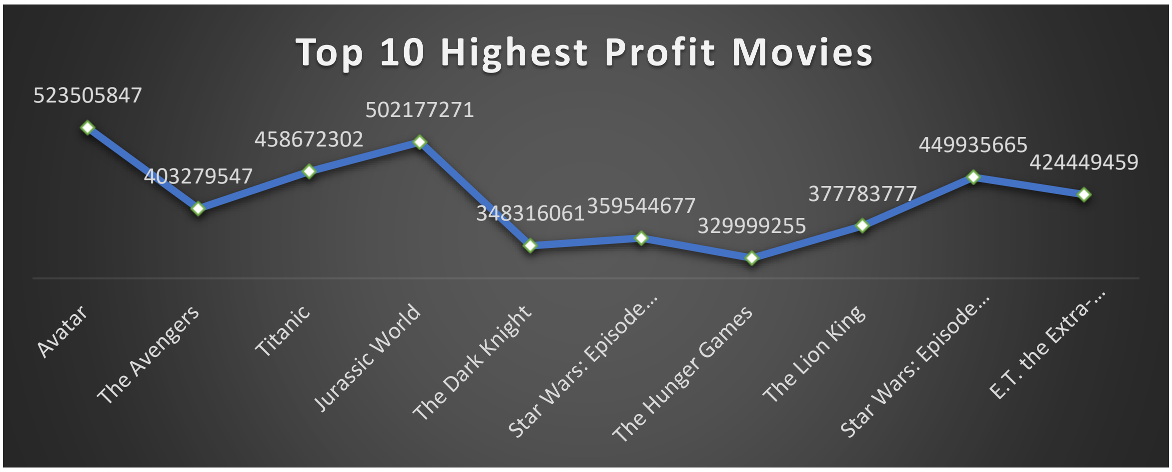
Highest Profit Movie
Avatar

- Also inserted a chart to display the top 10 highest profit margin movies according to the data.

Table:

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

Chart:



Bank Loan Case Study

Project Description:

As a data analyst at a finance company that specializes in providing various types of loans to urban customers, you encounter a significant challenge. Some customers with insufficient credit history take advantage of the system and default on their loans. To address this issue, your task is to employ Exploratory Data Analysis (EDA) to uncover patterns in the data and ensure that eligible applicants are not unjustly rejected.

The Problem

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

Approach

For this case study, I followed a structured approach: understanding the requirements, cleaning the data, and selecting the necessary columns in the Excel sheet, and presenting the results in both table and visual formats.

Note: For this analysis, we used two datasets: application_data and previous_application. Based on the requirements and provided hints, I focused my analysis on the application_data dataset.

Tech-Stack Used:

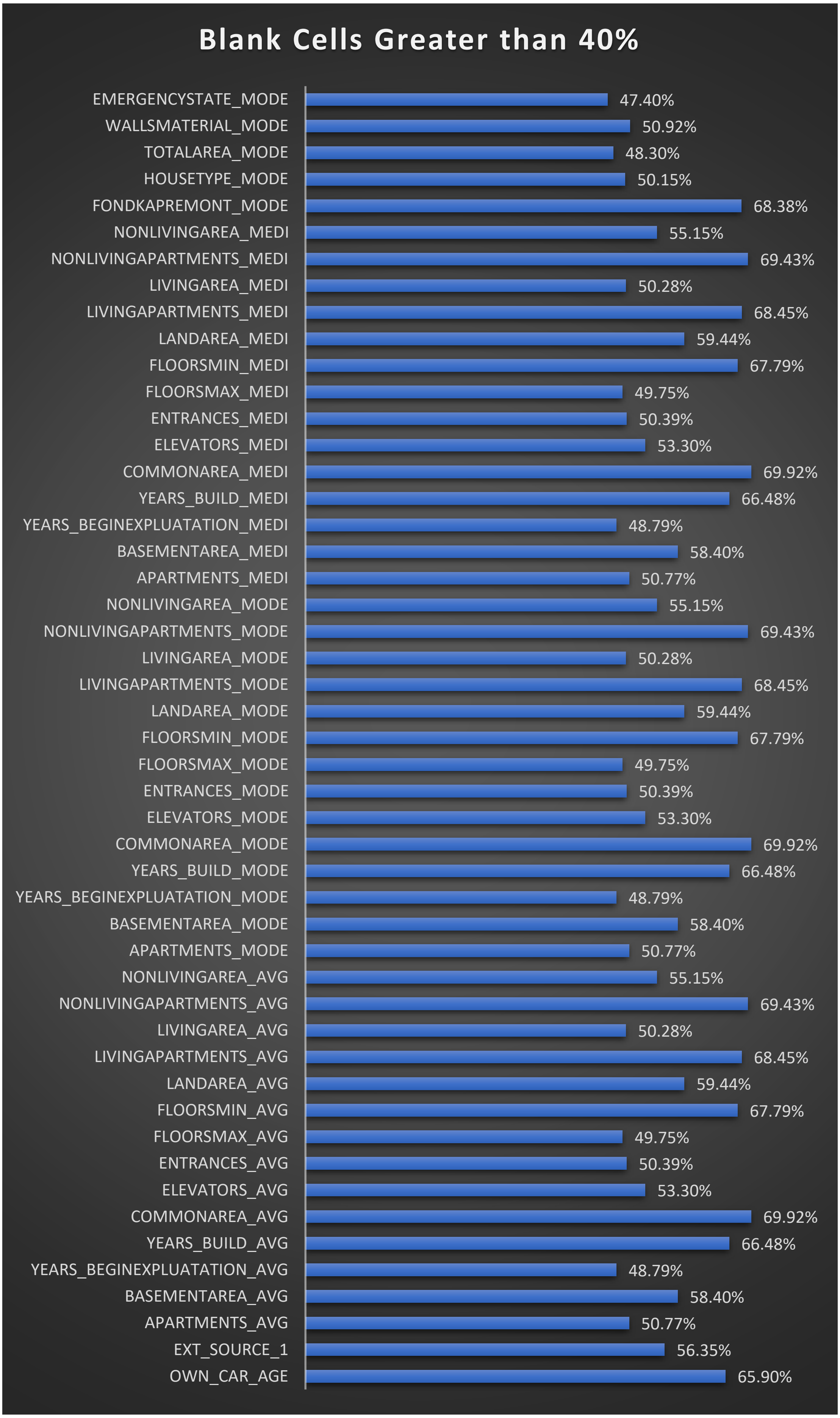
For this analysis, I used Excel for data cleaning and visualizations.

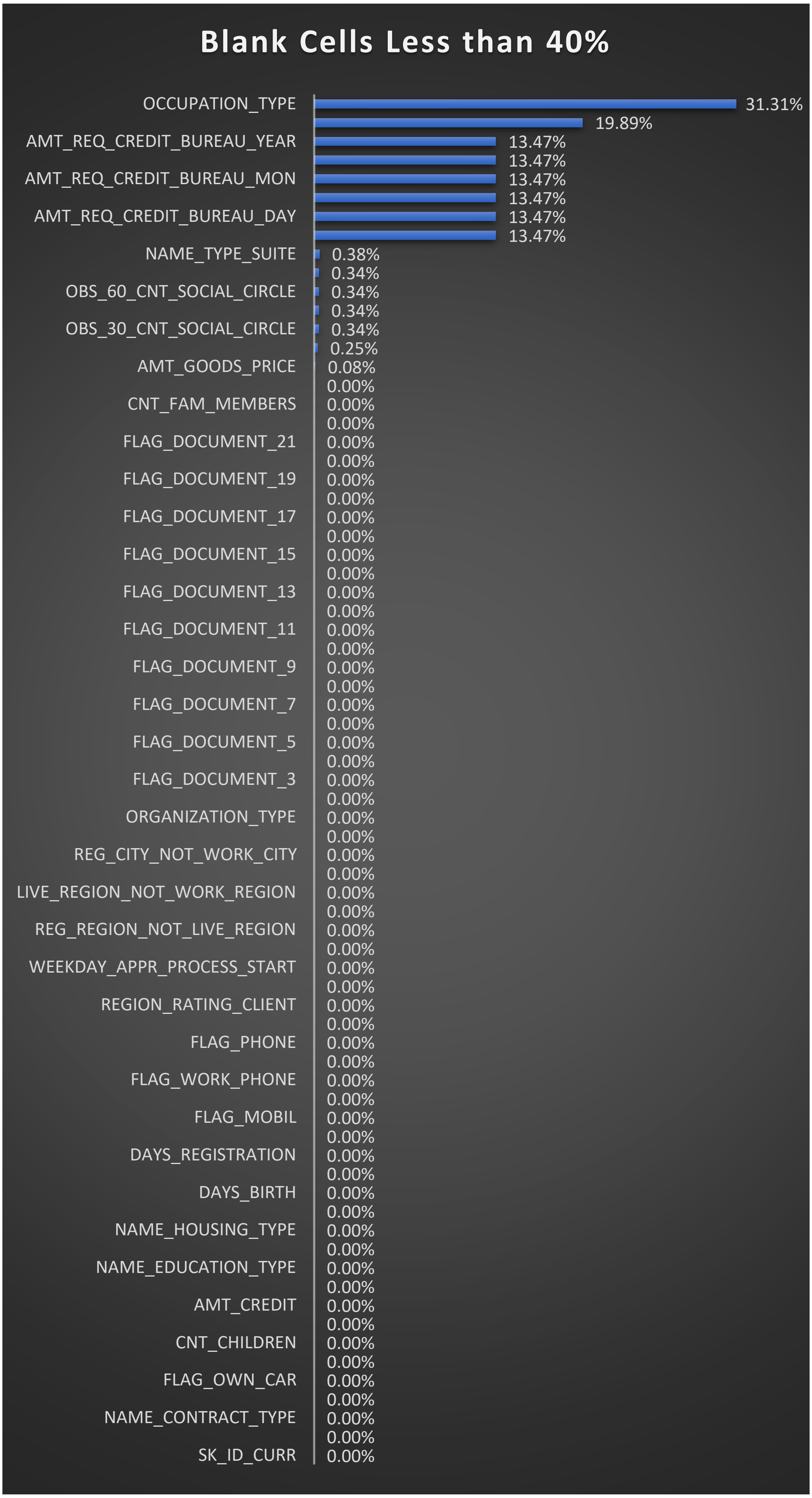
Findings

A. Identify Missing Data and Deal with it Appropriately:

The first task for any given dataset is to identify duplicates and missing data. For this task, we used Excel's COUNTIF function to find missing data and calculated the percentage of blank values in each column. If a column had more than 40% blank values, we dropped it. For columns with less than 40% blank values, we filled the missing data using mean and median values. The visual representation of the blank data cells is shown below.

Charts:


Blank Cells Greater than 40%
EMERGENCYSTATE_MODE 47.40%
WALLSMATERIAL_MODE 50.92%
TOTALAREA_MODE 48.30%
HOUSETYPE_MODE 50.15%
FONDKAPREMONT_MODE 68.38%
NONLIVINGAREA_MEDI 55.15%
NONLIVINGAPARTMENTS_MEDI 69.43%
LIVINGAREA_MEDI 50.28%
LIVINGAPARTMENTS_MEDI 68.45%
LANDAREA_MEDI 59.44%
FLOORSMIN_MEDI 67.79%
FLOORSMAX_MEDI 49.75%
ENTRANCES_MEDI 50.39%
ELEVATORS_MEDI 53.30%
COMMONAREA_MEDI 69.92%
YEARS_BUILD_MEDI 66.48%
YEARS_BEGINEXPLUATATION_MEDI 48.79%
BASEMENTAREA_MEDI 58.40%
APARTMENTS_MEDI 50.77%
NONLIVINGAREA_MODE 55.15%
NONLIVINGAPARTMENTS_MODE 69.43%
LIVINGAREA_MODE 50.28%
LIVINGAPARTMENTS_MODE 68.45%
LANDAREA_MODE 59.44%
FLOORSMIN_MODE 67.79%
FLOORSMAX_MODE 49.75%
ENTRANCES_MODE 50.39%
ELEVATORS_MODE 53.30%
COMMONAREA_MODE 69.92%
YEARS_BUILD_MODE 66.48%
YEARS_BEGINEXPLUATATION_MODE 48.79%
BASEMENTAREA_MODE 58.40%
APARTMENTS_MODE 50.77%
NONLIVINGAREA_AVG 55.15%
NONLIVINGAPARTMENTS_AVG 69.43%
LIVINGAREA_AVG 50.28%
LIVINGAPARTMENTS_AVG 68.45%
LANDAREA_AVG 59.44%
FLOORSMIN_AVG 67.79%
FLOORSMAX_AVG 49.75%
ENTRANCES_AVG 50.39%
ELEVATORS_AVG 53.30%
COMMONAREA_AVG 69.92%
YEARS_BUILD_AVG 66.48%
YEARS_BEGINEXPLUATATION_AVG 48.79%
BASEMENTAREA_AVG 58.40%
APARTMENTS_AVG 50.77%
EXT_SOURCE_1 56.35%
OWN_CAR_AGE 65.90%


Blank Cells Less than 40%
OCCUPATION_TYPE
AMT_REQ_CREDIT_BUREAU_YEAR
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_DAY
NAME_TYPE_SUITE
OBS_60_CNT_SOCIAL_CIRCLE
OBS_30_CNT_SOCIAL_CIRCLE
AMT_GOODS_PRICE
CNT_FAM_MEMBERS
FLAG_DOCUMENT_21
FLAG_DOCUMENT_19
FLAG_DOCUMENT_17
FLAG_DOCUMENT_15
FLAG_DOCUMENT_13
FLAG_DOCUMENT_11
FLAG_DOCUMENT_9
FLAG_DOCUMENT_7
FLAG_DOCUMENT_5
FLAG_DOCUMENT_3
ORGANIZATION_TYPE
REG_CITY_NOT_WORK_CITY
LIVE_REGION_NOT_WORK_REGION
REG_REGION_NOT_LIVE_REGION
WEEKDAY_APPR_PROCESS_START
REGION_RATING_CLIENT
FLAG_PHONE
FLAG_WORK_PHONE
FLAG_MOBIL
DAYS_REGISTRATION
DAYS_BIRTH
NAME_HOUSING_TYPE
NAME_EDUCATION_TYPE
AMT_CREDIT
CNT_CHILDREN
FLAG_OWN_CAR
NAME_CONTRACT_TYPE
SK_ID_CURR
31.31%
19.89%
13.47%
13.47%
13.47%
13.47%
13.47%
13.47%
0.38%
0.34%
0.34%
0.34%
0.34%
0.25%
0.08%
0.00%

Mean & Median				
Column_names	**Count_blank**	**Count_blank_percentage**	**Mean**	**Median**
AMT_GOODS_PRICE	38	0.08%	539060.04	450000
NAME_TYPE_SUITE	192	0.38%	Filled with Unaccompanied	
OCCUPATION_TYPE	15654	31.31%	Filled with unknow	
EXT_SOURCE_2	126	0.25%	0.5138236	0.565585
EXT_SOURCE_3	9944	19.89%	0.5118814	0.535276
OBS_30_CNT_SOCIAL_CIRCLE	168	0.34%	1.4207822	0
DEF_30_CNT_SOCIAL_CIRCLE	168	0.34%	0.1418193	0
OBS_60_CNT_SOCIAL_CIRCLE	168	0.34%	1.4036644	0
DEF_60_CNT_SOCIAL_CIRCLE	168	0.34%	0.0983324	0
AMT_REQ_CREDIT_BUREAU_HOUR	6734	13.47%	0.0070958	0
AMT_REQ_CREDIT_BUREAU_DAY	6734	13.47%	0.0075118	0
AMT_REQ_CREDIT_BUREAU_WEEK	6734	13.47%	18390.647	0
AMT_REQ_CREDIT_BUREAU_MON	6734	13.47%	0.2702878	0
AMT_REQ_CREDIT_BUREAU_QRT	6734	13.47%	0.2609731	0
AMT_REQ_CREDIT_BUREAU_YEAR	6734	13.47%	1.8810355	1

The above data displays the mean and median values used to fill in the missing data for the respective columns. Additionally, we removed three rows with fewer blank cells by filtering on the columns AMT_ANNUITY, CNT_FAM_MEMBERS, and DAYS_LAST_PHONE_CHANGE.

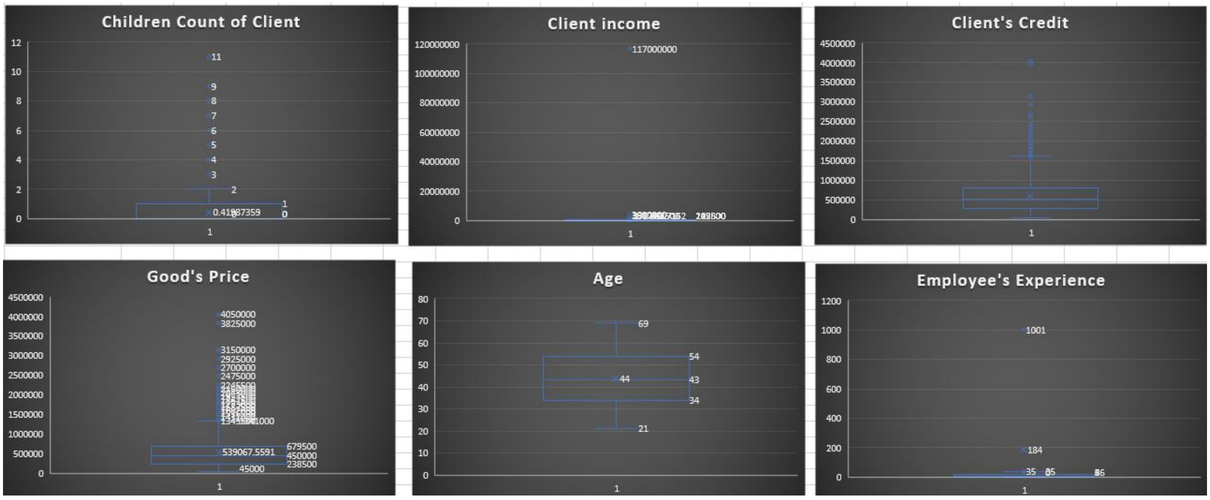
B. Identify Outliers in the Dataset:

We need to identify outliers because they can significantly affect the results of our analysis and may indicate errors or unique cases that need special attention.

To identify outliers in the dataset, I examined six columns, which are:

Column_name	Q1	Q3	IQR	Lower Bound	Upper Bound
CNT_CHILDREN	0	1	1	-1.5	2.5
AMT_INCOME_TOTAL	112500	202500	90000	-22500	337500
AMT_CREDIT	270000	808650	538650	-537975	1616625
AMT_GOODS_PRICE	238500	679500	441000	-423000	1341000
DAYS_BIRTH	34	54	20	4	84
DAYS_EMPLOYED	3	16	13	-17	35

Visual Representation of Outliers:



- Among the identified outliers, one client has 11 children, which seems implausible, so I classified this as an outlier.
- One client has an income of 1 crore, but their occupation is listed as a labourer, which makes such a high income unlikely. Therefore, I identified this as an outlier.
- Another outlier shows an employee with 1000 years of experience, which is clearly unrealistic.

Considering the above outliers, I removed those rows and prepared a new dataset for the next analysis.

C. Analyze Data Imbalance:

Analysing data imbalance is important because it helps ensure that our model or analysis does not become biased towards the more prevalent class. Imbalanced data can lead to inaccurate results, as the model may perform well on the majority class while neglecting the minority class, which might be of greater interest or importance. By addressing data imbalance, we can improve the fairness and accuracy of our analysis or predictive models.

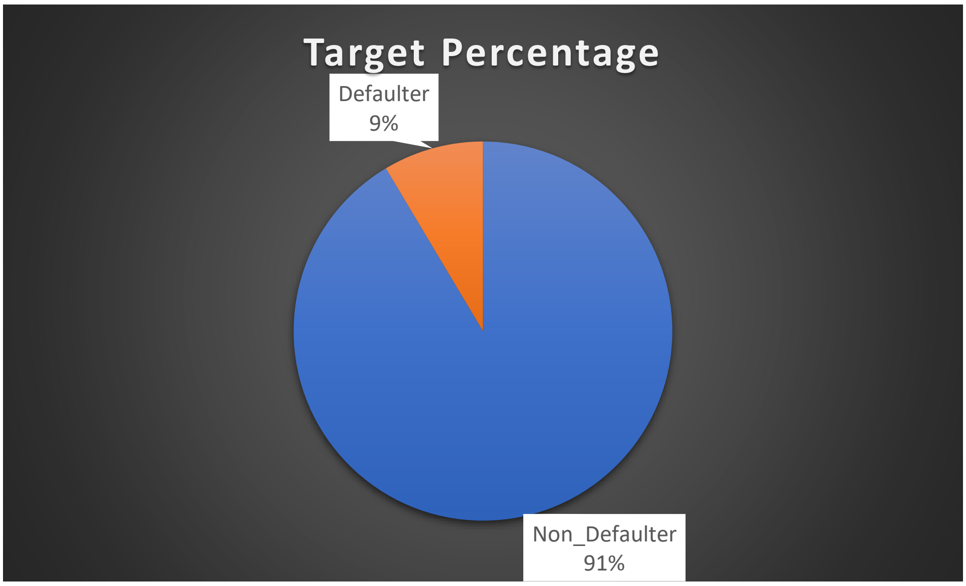
To identify data imbalance, I analysed two columns: Target and Contract Type. The Target column indicates whether loans are paid on time or are delayed, while the Contract Type column reveals the most common types of loans provided.

Chart & Table:

Target Column

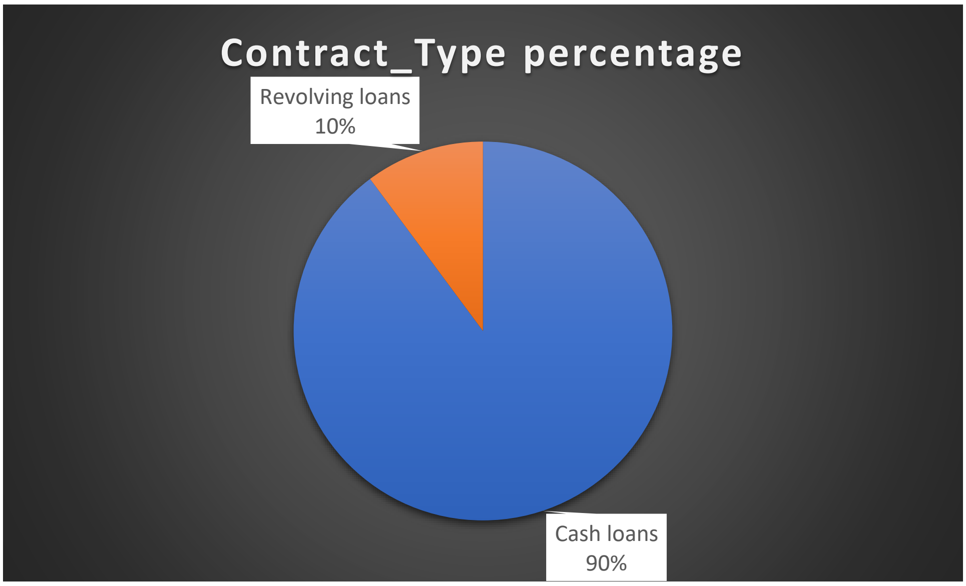
Target	Count_of_target	Count_percentage
Non_Defaulter	37548	91%
Defaulter	3520	9%

0	Non_Defaulter(Who is payin on time)
1	Defaulter(Who is not payin ontime)



Contract Type:

Contract_Type	Count_of	Count_Percentage
Cash loans	36892	90%
Revolving loans	4176	10%



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

Univariate analysis involves examining one variable at a time. The purpose is to understand the distribution, central tendency, and dispersion of the variable.

Segmented univariate analysis involves analysing a single variable within different segments or groups of data. This helps to compare the behaviour of the variable across different subsets.

Bivariate analysis examines the relationship between two variables. The goal is to determine if there is an association or correlation between them.

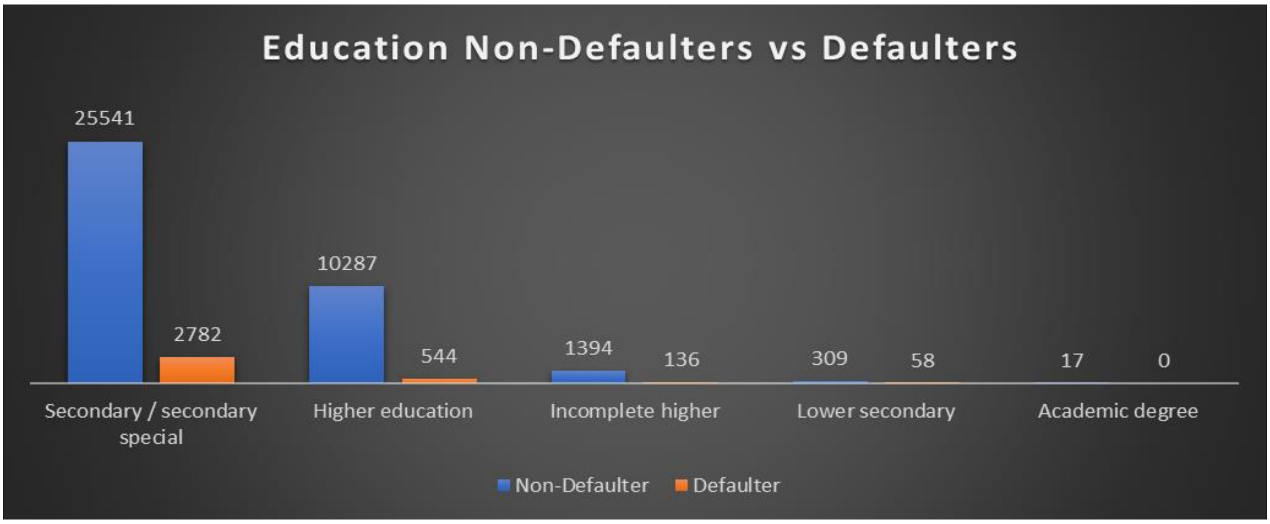
Univariate analysis:

To do the Univariate analysis considered two columns which are NAME_EDUCATION_TYPE and Target:

Table:

NAME_EDUCATION_TYPE	Secondary / secondary special	Higher education	Incomplete higher	Lower secondary	Academic degree
Total_Clients	28323	10831	1530	367	17
Non-Defaulter	25541	10287	1394	309	17
Defaulter	2782	544	136	58	0

Chart:



- From the above analysis, we can gain insights into loan repayment patterns based on education level, identifying who is paying on time versus those who are delayed.

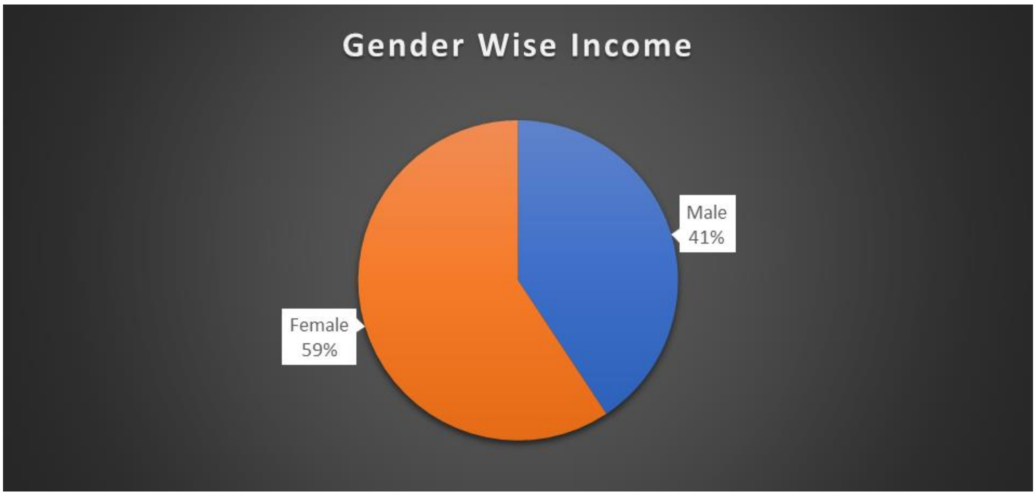
Segmented univariate:

To do segmented univariate analysis I considered two columns gender and income type columns.

Table:

Gender	Income
Male	3090016265
Female	4504070686

Chart:



- From the above results, we observe that female clients have higher incomes compared to male clients.

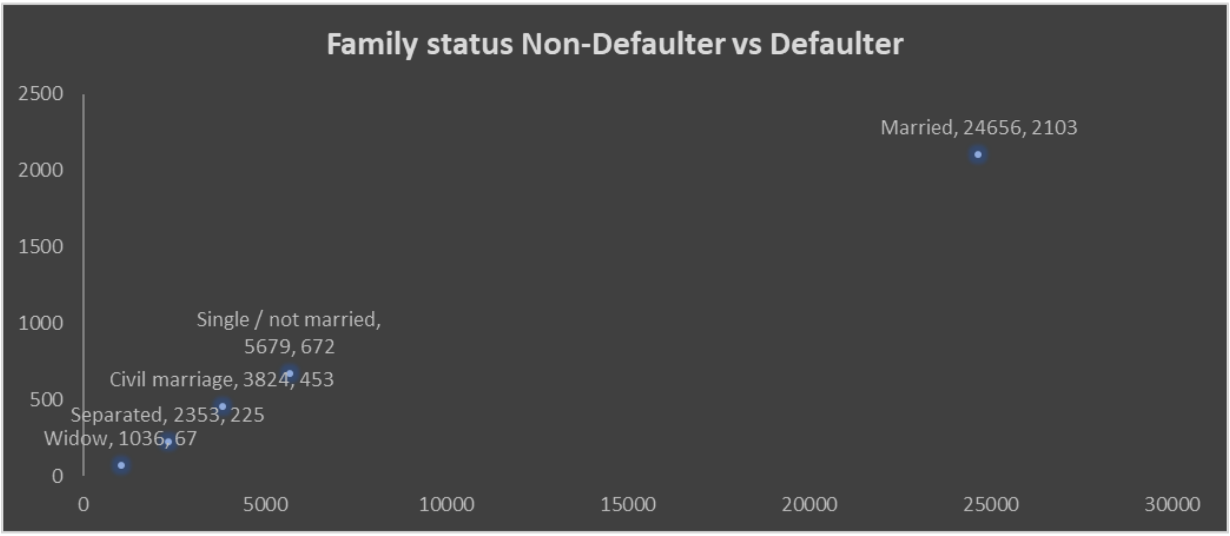
Bivariate analysis

To do the analysis I considered family type and target columns and also found the correlation between non-defaulter and defaulter.

Table:

NAME_FAMILY_STATUS	Non-Defaulter	Defaulter	Correlation
Single / not married	5679	672	0.994240921
Married	24656	2103	
Civil marriage	3824	453	
Widow	1036	67	
Separated	2353	225	

Chart:



- From the above results, we can determine which families are paying their loans on time and which are delayed.

E. Identify Top Correlations for Different Scenarios:

Identifying top correlations for different scenarios is important because it helps uncover relationships between variables that might impact outcomes. By understanding these correlations, we can:

I. Improve Predictions
II. Optimize Strategies
III. Spot Anomalies
IV. Focus Efforts.

To do the correlations for different scenarios I divided the columns based on different variables. They are:

Financial Variables

Income vs. Credit Amount:

AMT_INCOME_TOTAL and AMT_CREDIT

Income vs. Annuity:

AMT_INCOME_TOTAL and AMT_ANNUITY

Income vs. Goods Price:

AMT_INCOME_TOTAL and AMT_GOODS_PRICE

Annuity vs. Credit Amount:

AMT_ANNUITY and AMT_CREDIT

Annuity vs. Goods Price:

AMT_ANNUITY and AMT_GOODS_PRICE

Credit Amount vs. Goods Price:

AMT_CREDIT and AMT_GOODS_PRICE

Employment and Demographic Variables

Age vs. Employment Duration:

DAYS_BIRTH and DAYS_EMPLOYED

Age vs. Registration Days:

DAYS_BIRTH and DAYS_REGISTRATION

Employment Duration vs. Registration Days:

DAYS_EMPLOYED and DAYS_REGISTRATION

External Scores

EXT_SOURCE_2 and EXT_SOURCE_3

Social and Document Variables

OBS_30_CNT_SOCIAL_CIRCLE and DEF_30_CNT_SOCIAL_CIRCLE

OBS_60_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE

Credit Bureau Requests

Credit Bureau Requests in Hour vs. Day:

AMT_REQ_CREDIT_BUREAU_HOUR and AMT_REQ_CREDIT_BUREAU_DAY

Credit Bureau Requests in Day vs. Week:

AMT_REQ_CREDIT_BUREAU_DAY and AMT_REQ_CREDIT_BUREAU_WEEK

Credit Bureau Requests in Month vs. Quarter:

AMT_REQ_CREDIT_BUREAU_MON and AMT_REQ_CREDIT_BUREAU_QRT

Credit Bureau Requests in Quarter vs. Year:

AMT_REQ_CREDIT_BUREAU_QRT and AMT_REQ_CREDIT_BUREAU_YEAR

Table of Correlations with ranking and heatmap:

Correlation Column Names	Correlation Value	Rank	Heat Map
AMT_CREDIT and AMT_GOODS_PRICE	0.986038231	1	
AMT_ANNUITY and AMT_GOODS_PRICE	0.76519272	2	
AMT_ANNUITY and AMT_CREDIT	0.761024747	3	
AMT_INCOME_TOTAL and AMT_ANNUITY	0.42961991	4	
AMT_INCOME_TOTAL and AMT_GOODS_PRICE	0.367532161	5	
AMT_INCOME_TOTAL and AMT_CREDIT	0.359975889	6	
DAYS_BIRTH and DAYS_EMPLOYED	0.351442617	7	
OBS_30_CNT_SOCIAL_CIRCLE and DEF_30_CNT_SOCIAL_CIRCLE	0.313301568	8	
DAYS_BIRTH and DAYS_REGISTRATION	0.300963129	9	
AMT_REQ_CREDIT_BUREAU_DAY and AMT_REQ_CREDIT_BUREAU_WEEK	0.254894056	10	
OBS_60_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE	0.241601227	11	
AMT_REQ_CREDIT_BUREAU_HOUR and AMT_REQ_CREDIT_BUREAU_DAY	0.219599394	12	
DAYS_EMPLOYED and DAYS_REGISTRATION	0.175848444	13	
AMT_REQ_CREDIT_BUREAU_QRT and AMT_REQ_CREDIT_BUREAU_YEAR	0.118156653	14	
EXT_SOURCE_2 and EXT_SOURCE_3	0.102952775	15	
AMT_REQ_CREDIT_BUREAU_MON and AMT_REQ_CREDIT_BUREAU_QRT	0.011141088	16	

- From the above results, we can see that financial variables show the highest correlations.

Analyzing the Impact of Car Features on Price and Profitability

Project Description:

The automotive industry is rapidly evolving with a focus on fuel efficiency, sustainability, and technological innovation. Increasing competition and changing consumer preferences highlight the need to understand consumer demand for cars.

Recent trends show a rise in electric and hybrid vehicles and interest in alternative fuels like hydrogen and natural gas, while traditional gasoline-powered cars still dominate the market.

The Problem

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.

Approach

For this, I followed a descriptive analysis method. To perform descriptive analysis on a given dataset, you can follow these steps:

Understand the Data

Data Cleaning

Data Preparation

Data Visualization

Interpretation and Reporting

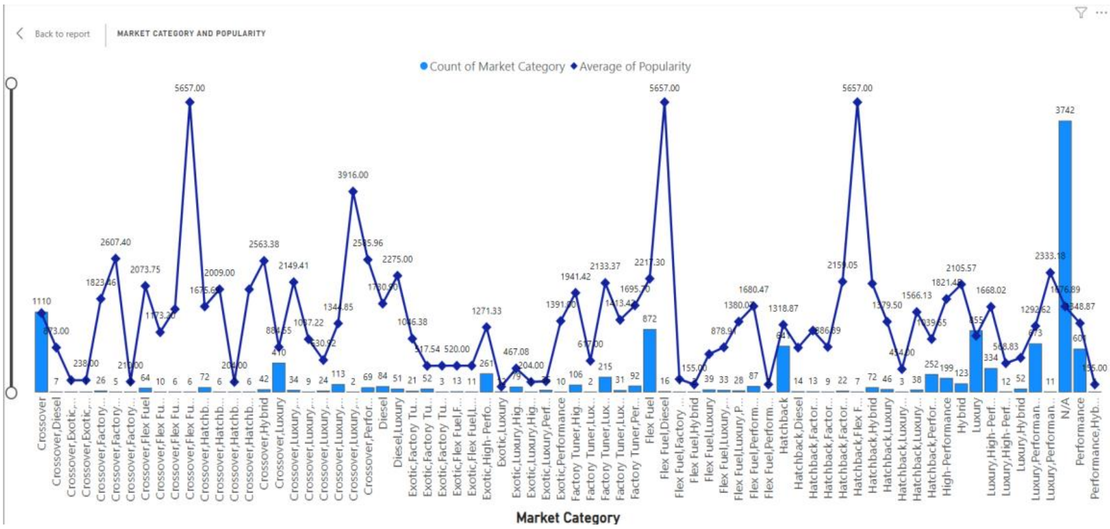
Documentation

Tech-Stack Used:

For this analysis, I used Microsoft Excel for data cleaning and statistical calculations, Power BI for creating interactive visualizations and dashboards, and PowerPoint to present the insights and findings.

Findings

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

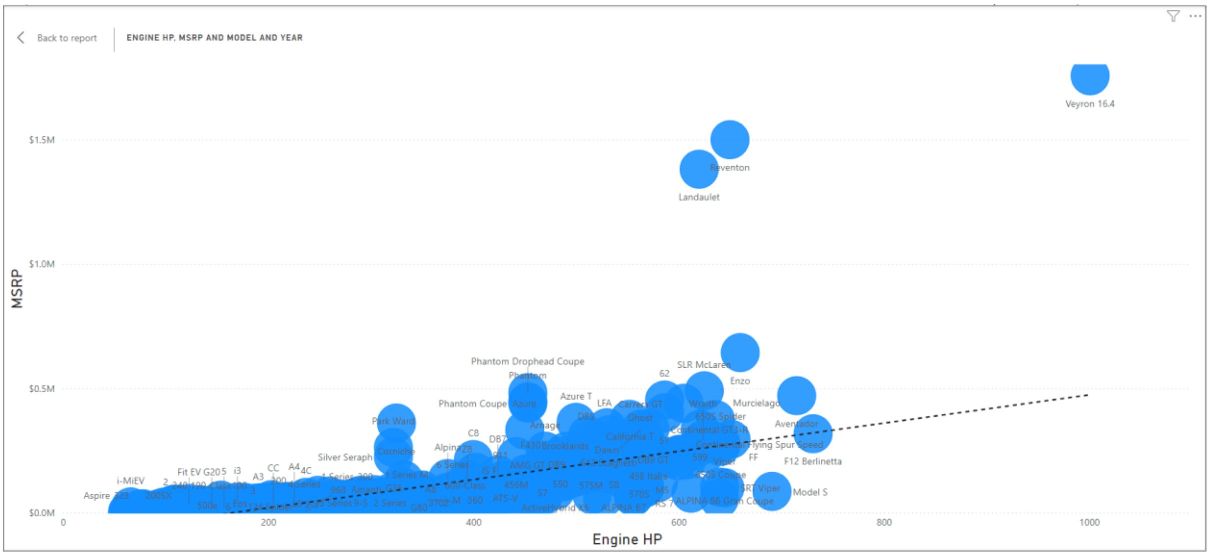


Insights:

Categories such as "Crossover Exotic" and "Luxury/Performance" have low counts but high average popularity. Introducing more items in these categories could lead to significant growth, given their positive reception.
"Crossover" and "Hatchback" categories have high counts but lower average popularity. This may suggest oversaturation or misalignment with consumer preferences, indicating a need for differentiation or enhancement.
The "Luxury/Hybrid" category balances count and popularity, showing a well-managed product line. Continuing this balance can satisfy market demand effectively while avoiding oversaturation.
"Performance/Hybrid" has high counts and popularity, signaling a strong market trend. This category could be a strategic focus for expansion and marketing to capitalize on its current momentum.

Categories like "Exotic/Factory" with low counts but decent popularity represent niche markets. Targeted marketing and increased inventory could uncover underserved segments with high growth potential.

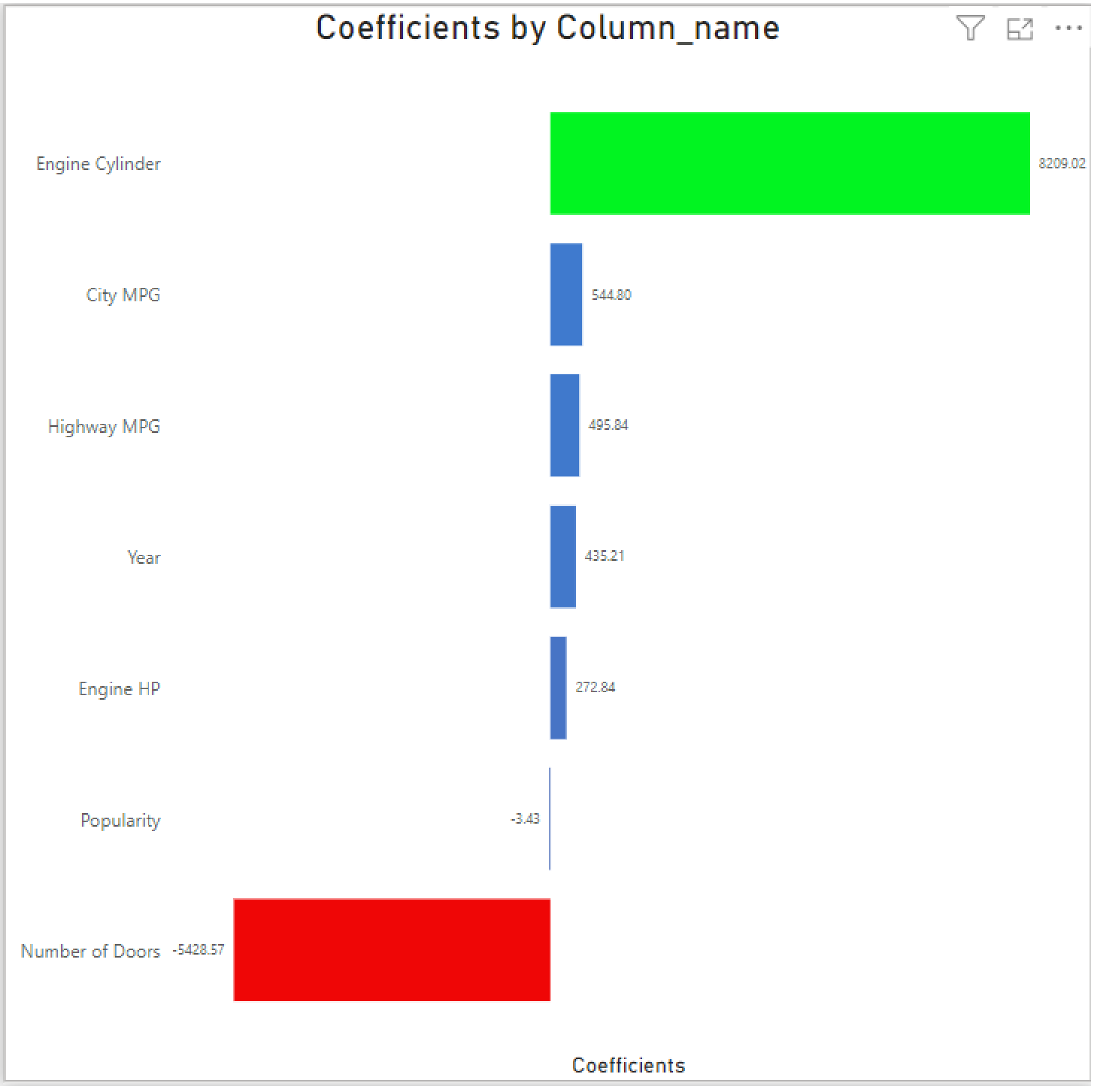
2. What is the relationship between a car's engine power and its price?



Insights:
Positive Correlation: There is an upward trend between engine horsepower and MSRP, indicating a positive correlation. As engine horsepower increases, the MSRP tends to rise, suggesting that higher-performance cars are more expensive.
Outliers: The Bugatti Veyron 16.4 is a notable outlier with high horsepower and MSRP, highlighting its status as a luxury performance vehicle. Other models like the Reventon and Landaulet also fall into the high-cost, high-performance category.
Clustered Data: Most car models are clustered in the lower range of both horsepower and MSRP. This indicates that most vehicles are moderately priced and perform well, appealing to a broad consumer base.
Market Segmentation: Cars with lower horsepower and MSRP, such as the "Aspire 323" and "i-MiEV," target the budget and economy segment. In contrast, vehicles like the "Aventador" and "Ferrari models" are in the sports and luxury segments, appealing to affluent consumers.
Strategic Opportunities: Manufacturers might explore increasing horsepower for models in the lower MSRP range to enhance market

appeal without significantly raising prices. Conversely, enhancing features and luxury in high-horsepower models could further attract premium buyers.

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

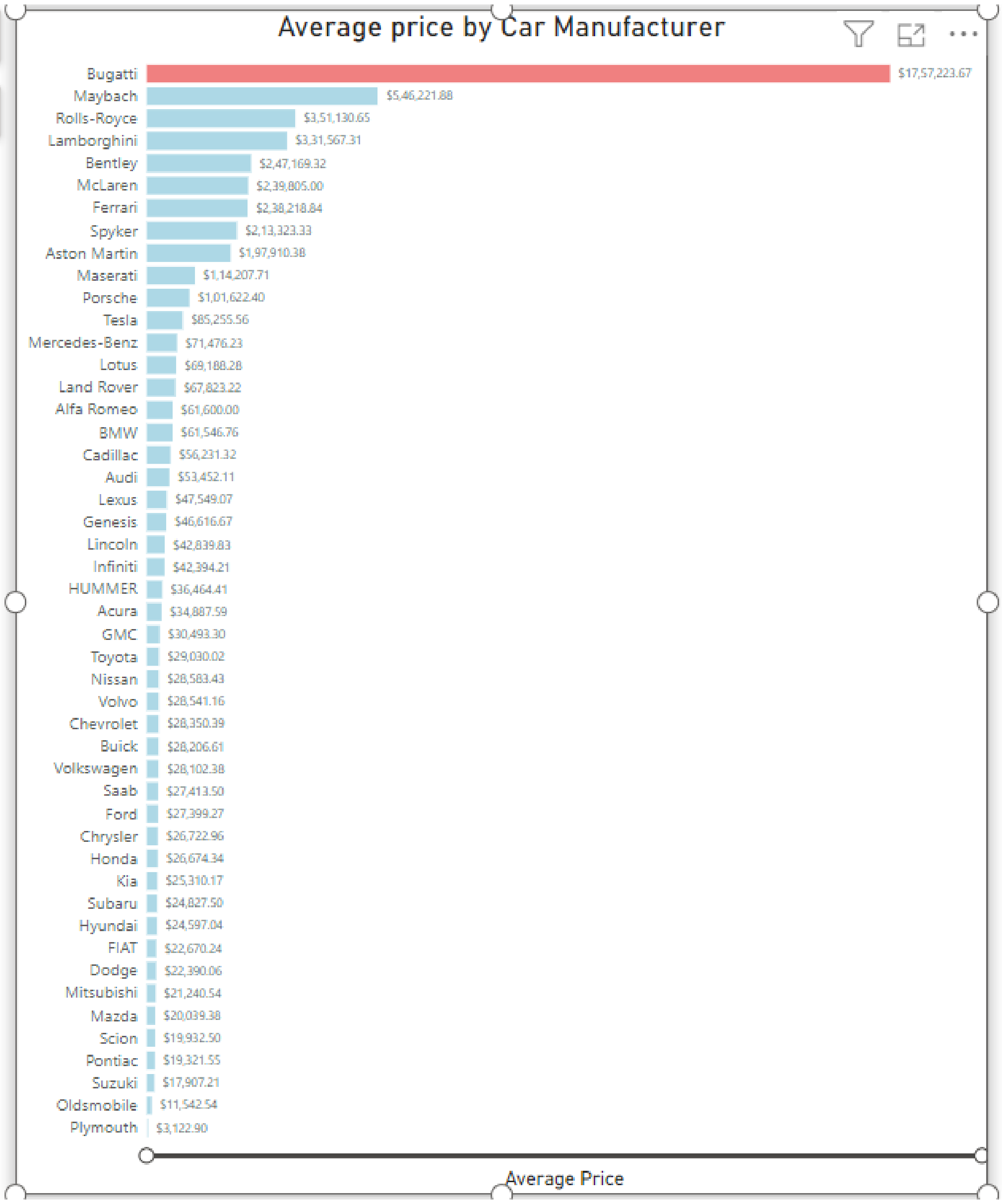


Insights:
Engine Cylinder (8209.02)
Impact: Positive and large.
Insight: This feature has the most significant positive impact on the target variable, suggesting that vehicles with more engine cylinders tend to have a higher target value (e.g., price).
City MPG (544.80)
Impact: Positive.

Insight: Higher city miles per gallon (MPG) contributes positively to the target variable, indicating a preference for fuel efficiency in city driving conditions.

Highway MPG (495.84)
Impact: Positive.
Insight: Like city MPG, better highway fuel efficiency also positively influences the target variable, though slightly less than city MPG.
Year (435.21)
Impact: Positive.
Insight: Newer models positively impact the target variable, likely reflecting trends of modern features or technology being more valued.
Engine HP (272.84)
Impact: Positive.
Insight: More horsepower contributes positively, aligning with the trend where performance is a desirable feature.
Popularity (-3.43)
Impact: Negative, but minimal.
Insight: Popularity has a slightly negative impact, possibly indicating that being overly common might reduce perceived uniqueness or value.
Number of Doors (-5428.57)
Impact: Negative and significant.
Insight: Vehicles with more doors tend to have a lower target value. This might suggest a preference or premium for sportier, two-door models in the dataset being analysed.

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



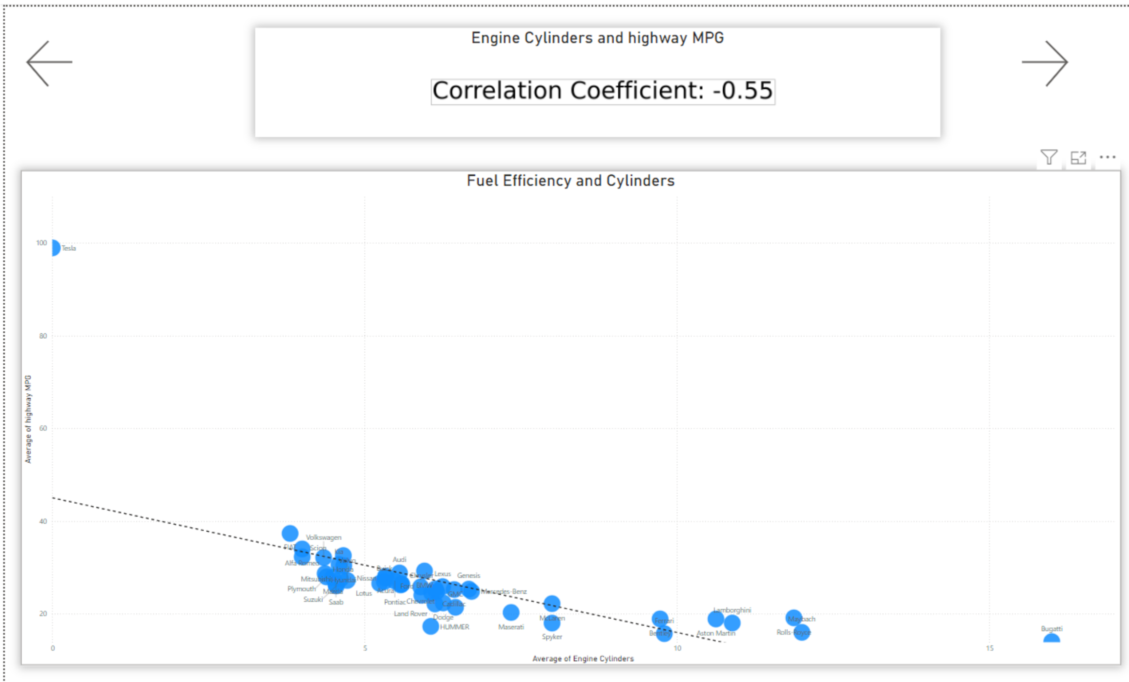
Insights:

Luxury Brands Dominate: The chart is dominated by high-end luxury and exotic car manufacturers. Bugatti and Maybach are at the top with the highest average prices of $17,57,222.67 and $5,46,221.88 respectively, indicating their market positioning as ultra-luxury brands.

Price Variation: There's a significant price variation between manufacturers. For example, Bugatti's average price is more than 500 times higher than that of Plymouth, which has the lowest average price of $3,132.90.

Mid-Range and Budget Manufacturers: Brands like Toyota, Nissan, and Ford are positioned towards the middle to lower end of the chart, reflecting their market strategy of offering more affordable and mass-market vehicles.
Market Segmentation: The chart effectively highlights the segmentation in the automotive market, with luxury brands at the top and budget brands at the bottom. This segmentation is crucial for understanding market dynamics and consumer purchasing power.
Outliers: Bugatti and Maybach are significant outliers with exceptionally high average prices, suggesting a niche market for ultra-luxurious vehicles.

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

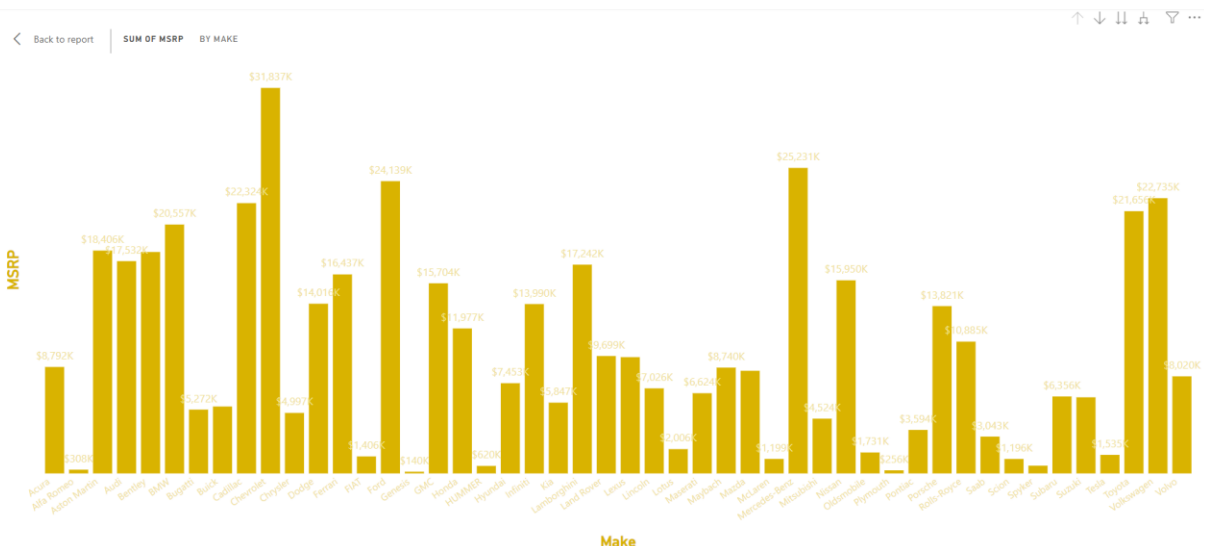


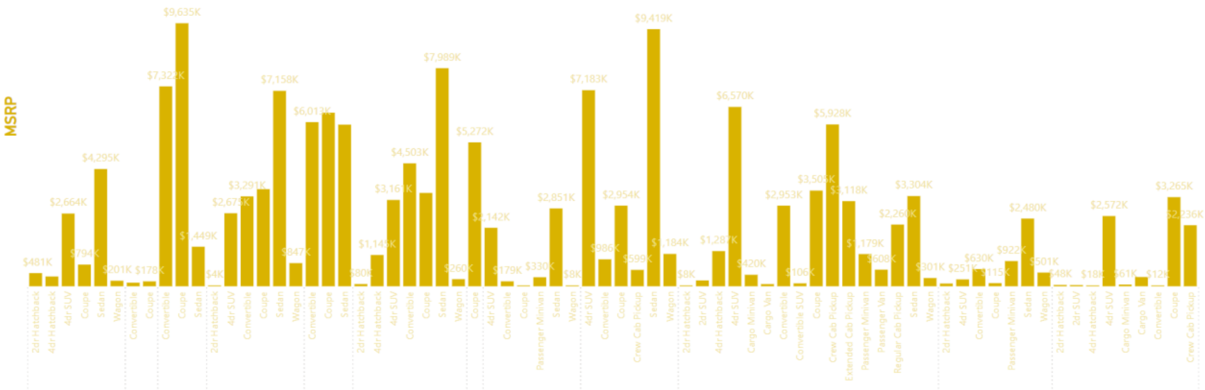
Insights:
Negative Correlation: The correlation coefficient of -0.55 indicates a moderate negative correlation between the number of engine cylinders and highway MPG. This means that as the number of cylinders increases, the fuel efficiency tends to decrease.
Tesla as an Outlier: Tesla stands out significantly with a high highway MPG (around 100). This is expected as Tesla vehicles are electric and not

dependent on traditional engine cylinders, leading to much higher fuel efficiency.
Luxury and High-Performance Brands: Brands like Bugatti, Rolls-Royce, Maybach, Lamborghini, and Aston Martin are clustered on the right side of the plot with a higher number of engine cylinders but lower highway MPG, indicating their focus on performance over fuel efficiency.
Economy and Mid-Range Brands: Brands such as Volkswagen, Scion, Honda, and Nissan are positioned on the left side of the plot with fewer engine cylinders and higher highway MPG, indicating a focus on fuel efficiency.
Main Cluster: Most brands are clustered in the middle, indicating a balance between engine performance and fuel efficiency, with an average of around 4-6 cylinders and highway MPG around 30-40.

Building the Dashboard

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





Insights:

High-End Brands:Bugatti has the highest total MSRP at $31,837K, indicating a significant presence of high-priced models.Ferrari, Porsche, and Mercedes-Benz also show high total MSRPs, reflecting their luxury and performance-oriented vehicles.

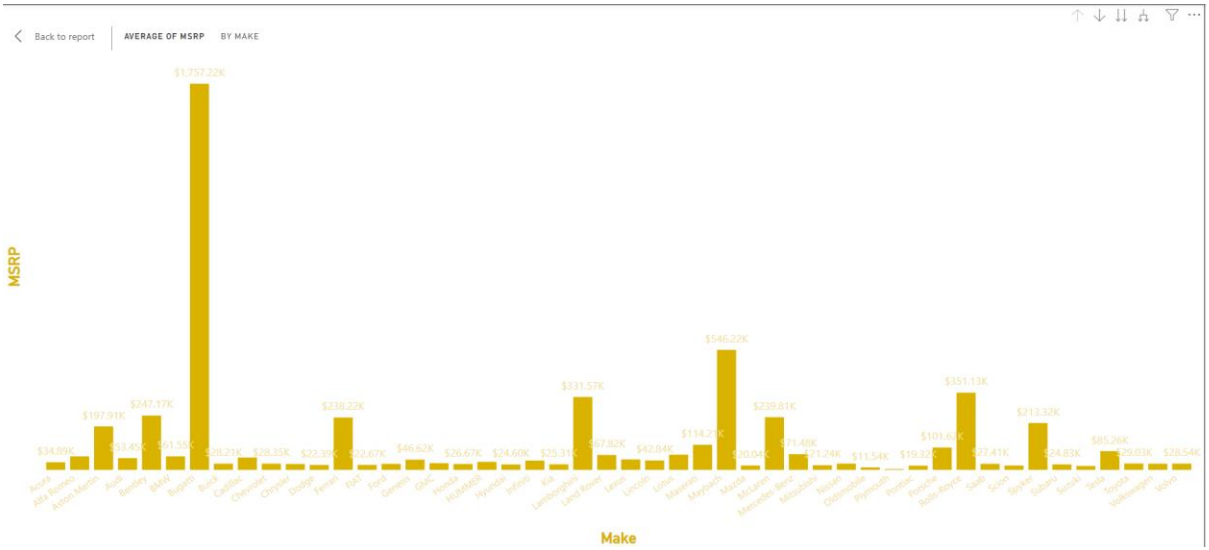
Mid-Range Brands:BMW and Audi have substantial total MSRPs of $20,557K and $18,406K respectively, indicating a mix of high-end and mid-range models.Cadillac, Chevrolet, and Ford also have notable MSRPs, showing their diverse range of vehicles from economy to luxury.

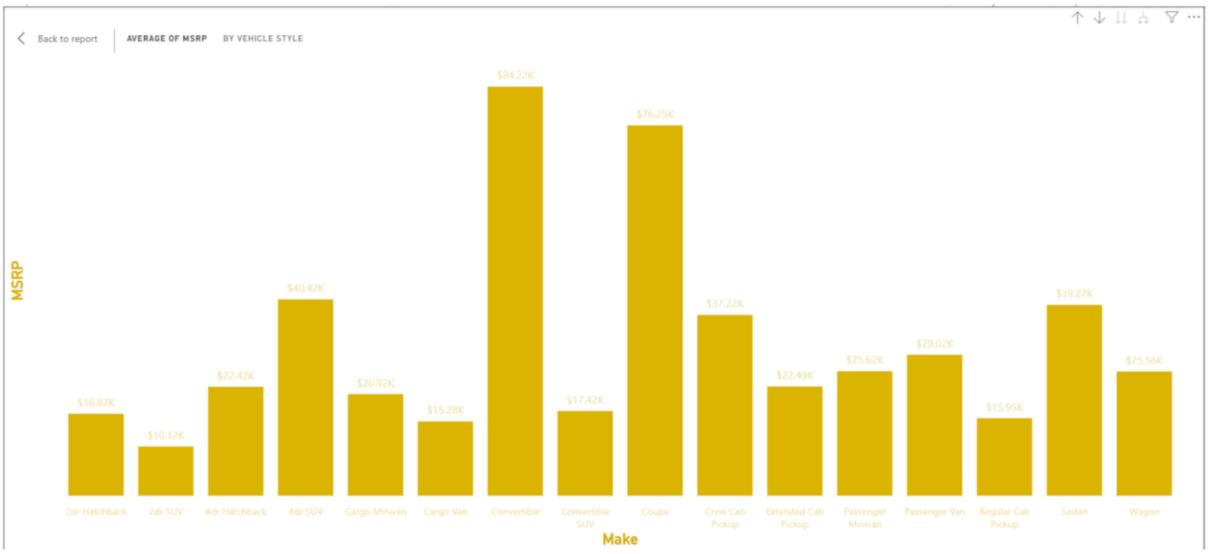
Economy Brands:Brands like Hyundai, Kia, and Toyota show lower total MSRPs, reflecting their focus on affordable, mass-market vehicles.Honda and Nissan are also in this category with moderate total MSRPs.

Low Presence or Niche Brands:Brands like Aston Martin and Lotus show relatively lower total MSRPs, indicating a smaller market presence or a niche focus.FIAT and Mitsubishi have lower total MSRPs, reflecting a focus on more affordable vehicles.

Notable Variations: Volkswagen and Volvo have substantial MSRPs, indicating a mix of economy and premium vehicles.Tesla has a moderate total MSRP, reflecting its position as a high-tech, premium electric vehicle manufacturer.

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

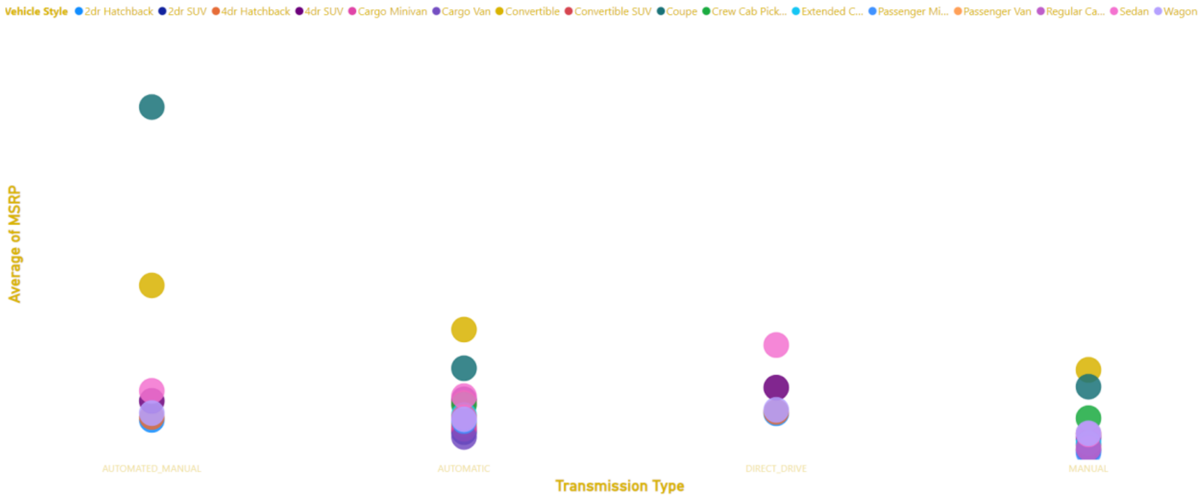




Insights:
Luxury vs. Economy: There is a clear distinction between luxury car manufacturers (Bugatti, Ferrari, Rolls-Royce) and economy brands (Geo, Yugo, Daewoo).
Convertibles and Coupes are typically luxury vehicles, which is reflected in their high average MSRPs.
Vehicle Style Trends: More practical vehicle styles like SUVs, Hatchbacks, and Pickups have lower average MSRPs compared to luxury styles. Sedans, though not as high as Convertibles and Coupes, still hold a significant average MSRP, indicating a mix of luxury and practicality.

Market Positioning: High-end brands position themselves at a premium price point, likely offering advanced features, better materials, and superior performance.
Lower-end brands focus on affordability and practicality, catering to a broader audience.

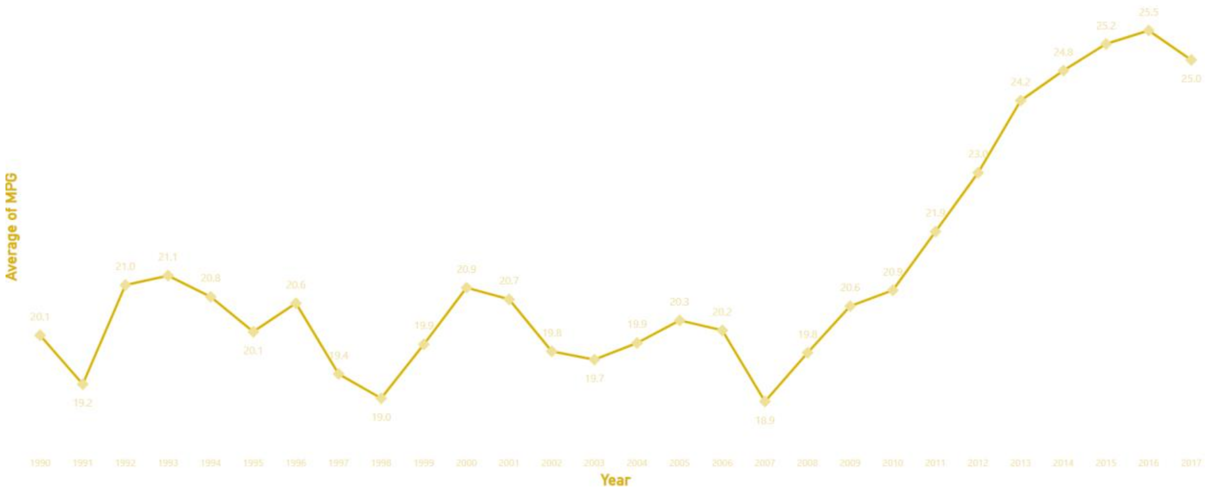
8. How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?

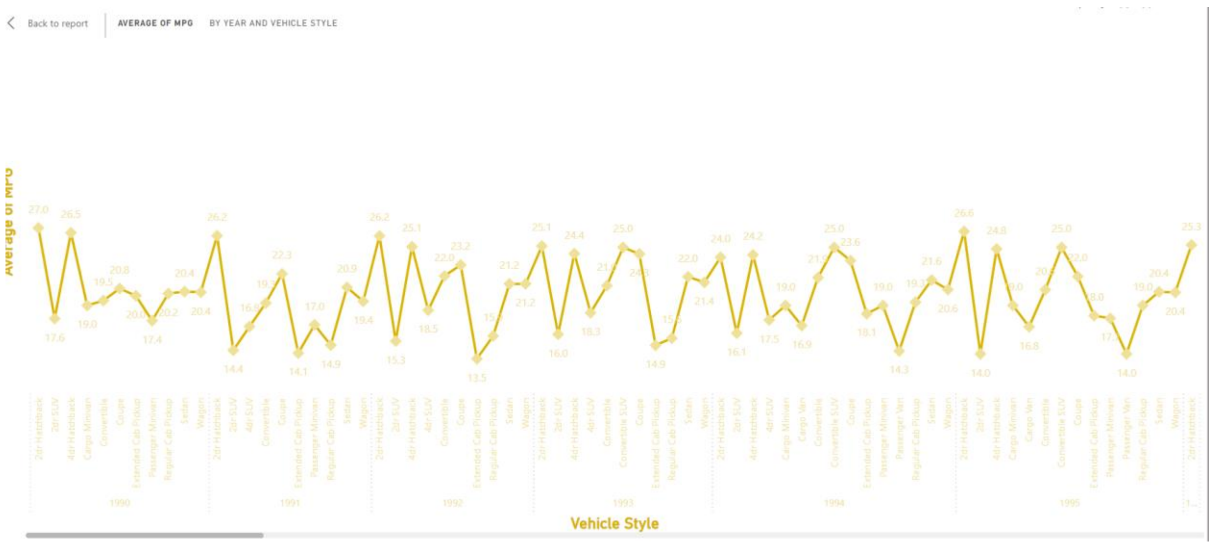


Insights:
Transmission Types:
Automated Manual:
Features fewer vehicle styles, predominantly represented by a single yellow bubble (Convertible) with a higher MSRP.
Automatic: The most diverse transmission type, showing a variety of vehicle styles with a wide range of MSRP values.
Direct Drive: Limited vehicle styles, with bubbles for Cargo Minivan, Cargo Van, and Regular Cab Pick-up Truck.
Manual: Several vehicle styles are present, with smaller bubbles indicating lower MSRP values on average compared to Automated Manual and Automatic transmission types.
Vehicle Styles:
Convertible: Consistently appears with a high MSRP across different transmission types.
Coupes, Sedans, SUVs:
Frequently appear under automatic and manual transmission types with varied MSRP values.

Cargo and Passenger Vans/Minivans:
Predominantly appear under Direct Drive and Automatic transmission types, indicating their typical use in commercial and family settings.
Price Trends:
Higher MSRP:
Automated Manual and Automatic transmissions for certain vehicle styles (like Convertibles) tend to have higher MSRP values.
Lower MSRP:
Manual transmissions generally show lower MSRP values, indicating cost-effective options.
Conclusion:
Automated Manual and Automatic transmissions are associated with higher MSRPs, especially for luxury or premium vehicle styles like Convertibles.
Manual transmissions offer more economical options with lower MSRPs across various vehicle styles.
Certain vehicle styles like Cargo Vans and Minivans are more common in Direct Drive and Automatic transmissions, reflecting their use cases and target markets.

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





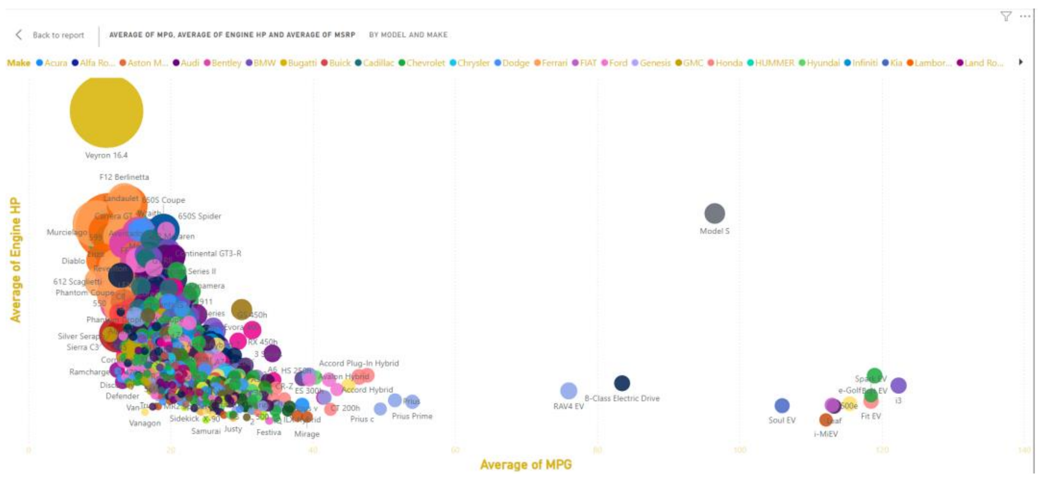
Insights:

Initial Stability and Fluctuations (1990-2004): During this period, the average MPG remains relatively stable, indicating consistent fuel efficiency standards and technologies.

Significant Improvement (Post-2009): The sharp increase in MPG from 2009 onwards could be attributed to technological advancements, increased adoption of hybrid and electric vehicles, and stricter environmental regulations pushing for better fuel efficiency.

Vehicle Style Variations: Different vehicle styles show varying MPG, highlighting those certain styles (e.g., smaller cars, hybrids) are more fuel-efficient compared to others (e.g., SUVs, trucks).

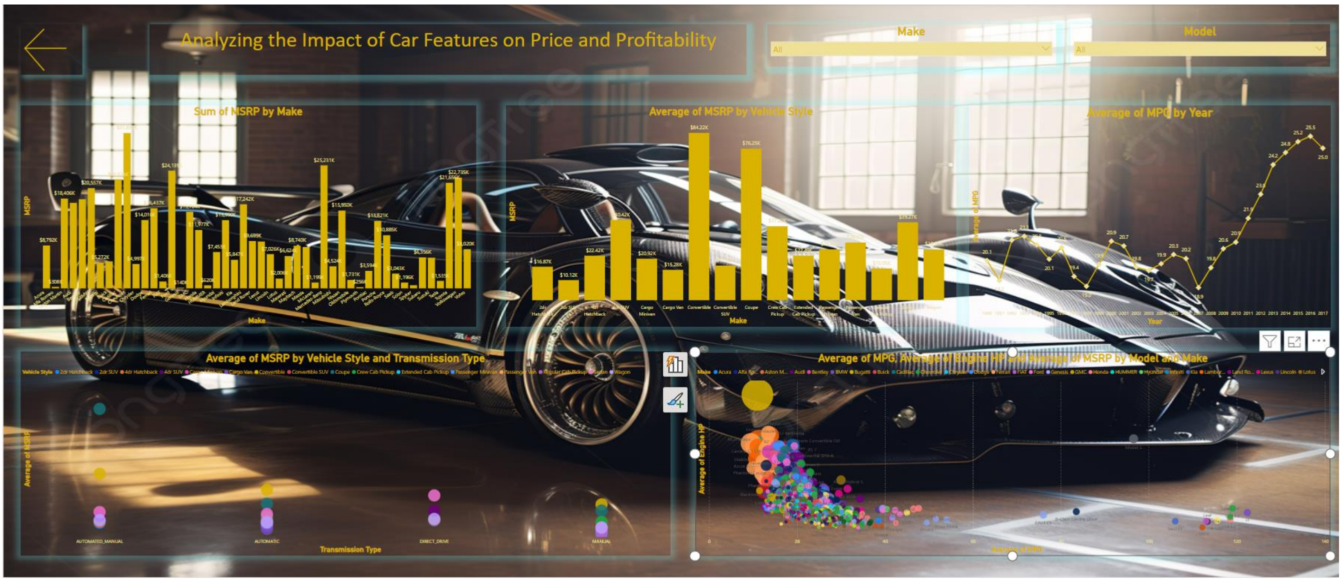
10. How does the car's horsepower, MPG, and price vary across different Brands?



Insights:

Fuel Efficiency vs. Power: The visualization highlights the common trade-off between fuel efficiency and engine power. Consumers often need to make choices based on their priorities.
Market Segmentation: The plot reveals distinct segments in the car market, with different combinations of MPG and HP catering to various consumer preferences.
Price as a Factor: The bubble size emphasizes how price can influence both fuel efficiency and engine power. Higher-priced vehicles tend to have more powerful engines but might compromise on fuel economy.

Dashboard In Power bi:



ABC Call Volume Trend Analysis

Project Description:

This project focuses on analyzing the Customer Experience (CX) of the inbound calling team at a company. The provided dataset spans 23 days, including details such as agent names and IDs, queue times, call times, call durations, and call statuses (abandoned, answered, or transferred).

The CX team is crucial in a company, responsible for analyzing customer feedback, deriving insights, and managing customer experience programs. Key tasks include handling internal communications, mapping customer journeys, and managing customer data.

Modern CX enhancement utilizes AI-powered tools like Interactive Voice Response (IVR), Robotic Process Automation (RPA), Predictive Analytics, and Intelligent Routing.

Customer service representatives, or call center agents, are vital in CX teams, providing various types of support, including email, inbound, outbound, and social media support. This project specifically focuses on inbound customer support, aiming to attract, engage, and delight customers, ultimately turning them into loyal advocates for the business.

The Problem

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

In the current era, several AI-powered tools are being used to enhance customer experience. These include Interactive Voice Response (IVR), Robotic Process Automation (RPA), Predictive Analytics, and Intelligent Routing.

One of the key roles in a CX team is that of the customer service representative, also known as a call center agent. These agents handle various types of support, including email, inbound, outbound, and social media support.

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.

Approach

For this, I followed a descriptive analysis method. To perform descriptive analysis on a given dataset, you can follow these steps:

Understand the Data

Data Cleaning

Data Preparation

Data Visualization

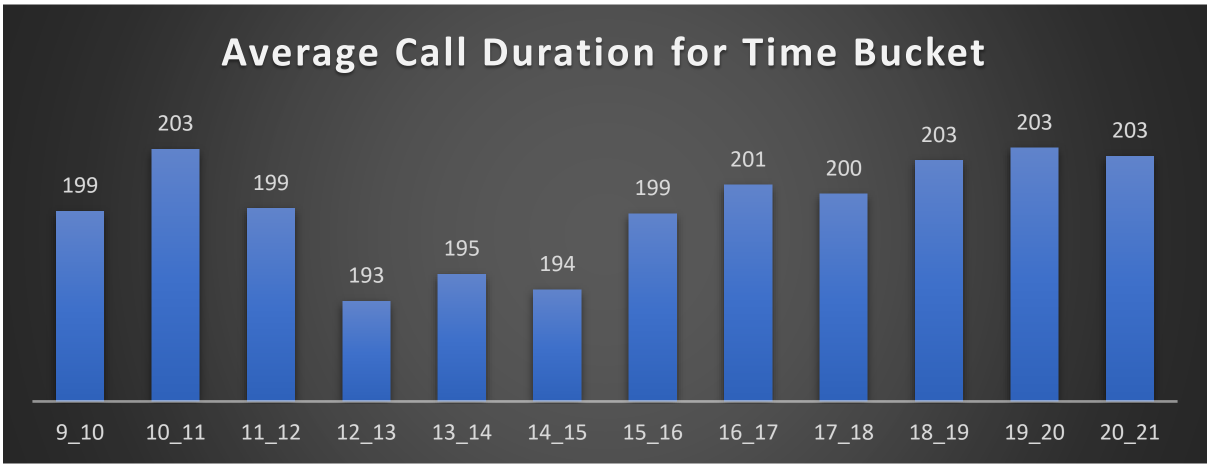
Documentation

Tech-Stack Used

For this analysis, I used Microsoft Excel for data cleaning, statistical calculations, and creating interactive visualizations. PowerPoint was used to present the insights and findings.

Findings

1. Average Call Duration



Insights:

Peak Average Call Durations:

The highest average call durations are observed during the time buckets of 10-11 AM, 19-20 PM, and 20-21 PM, each with an average duration of 203 seconds. This suggests that these time slots may experience more complex or lengthier customer interactions.

Consistently High Durations in Evening Hours:

The evening hours from 18-21 PM have consistently high average call durations, all around 200-203 seconds. This could indicate that customers calling during these times may have more detailed queries or that the nature of interactions in the evening tends to be more time-consuming.

Lower Average Call Durations in Midday:

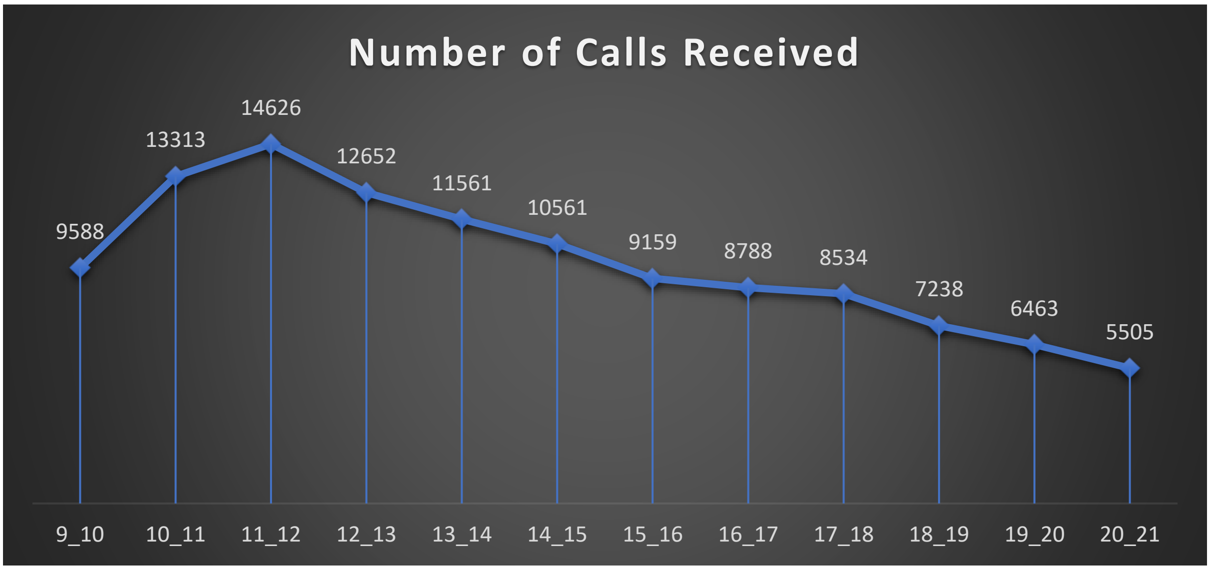
The time buckets from 12-14 PM show the lowest average call durations, specifically 193 and 194 seconds. This might suggest that customer interactions are quicker during these hours, possibly due to less complex issues being addressed or fewer calls being handled.

Steady Increase in Call Duration Towards the Evening:

There is a noticeable increase in average call duration from 14-21 PM, peaking at 203 seconds in the later hours. This trend could indicate that as the day progresses, the nature of customer inquiries might become more intricate, requiring longer handling times.

Potential Staffing Adjustments:
Given the higher average call durations in the evening, it might be beneficial to ensure adequate staffing during these peak hours to manage the increased workload and maintain customer satisfaction.

2. Call Volume Analysis



Insights:
Peak Call Volume in the Late Morning:
The highest number of calls is received between 11-12 AM, with a total of 14,626 calls. This indicates that late morning is the busiest time for inbound calls.
Morning Surge:
There is a significant increase in call volume from 9-10 AM (9,588 calls) to 10-11 AM (13,313 calls), continuing to the peak at 11-12 AM. This suggests a high demand for customer support services as the day starts.

Steady Decline Throughout the Day:
After reaching the peak at 11-12 AM, the number of calls steadily declines throughout the day, with notable drops during 12-13 PM (12,652 calls) and further to 13-14 PM (11,561 calls).
Lower Evening Call Volume:

The call volume continues to decrease in the evening, reaching its lowest point between 20-21 PM with 5,505 calls. This suggests that fewer customers are reaching out for support during the later hours.

Correlation with Call Duration:

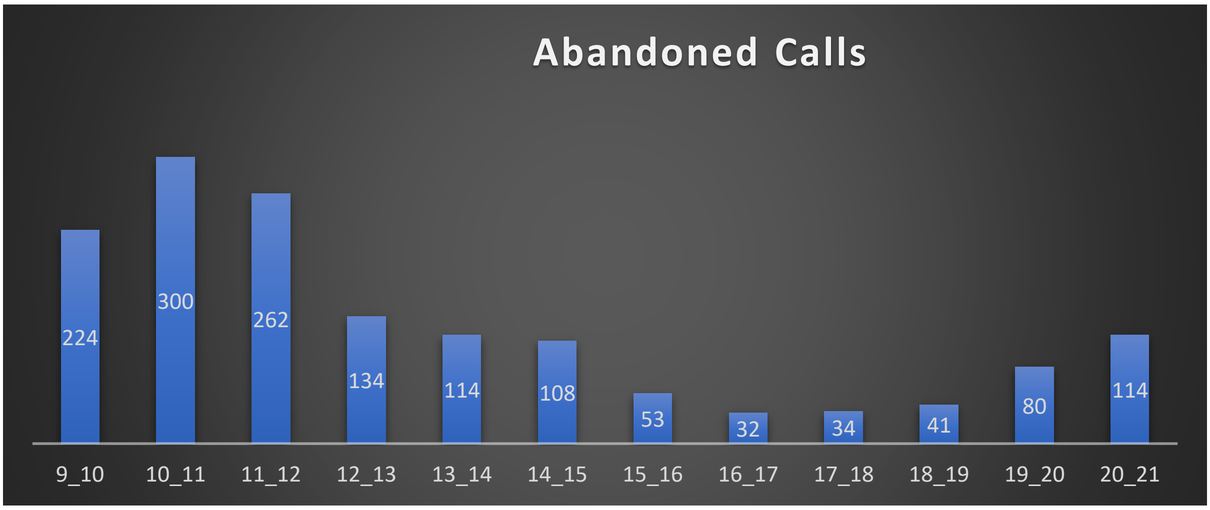
Comparing this chart with the previous one on average call duration, there appears to be an inverse relationship: as the number of calls decreases, the average call duration tends to be higher. This might indicate that agents can spend more time per call when the call volume is lower, particularly in the evening hours.

Implications for Staffing and Resource Allocation:

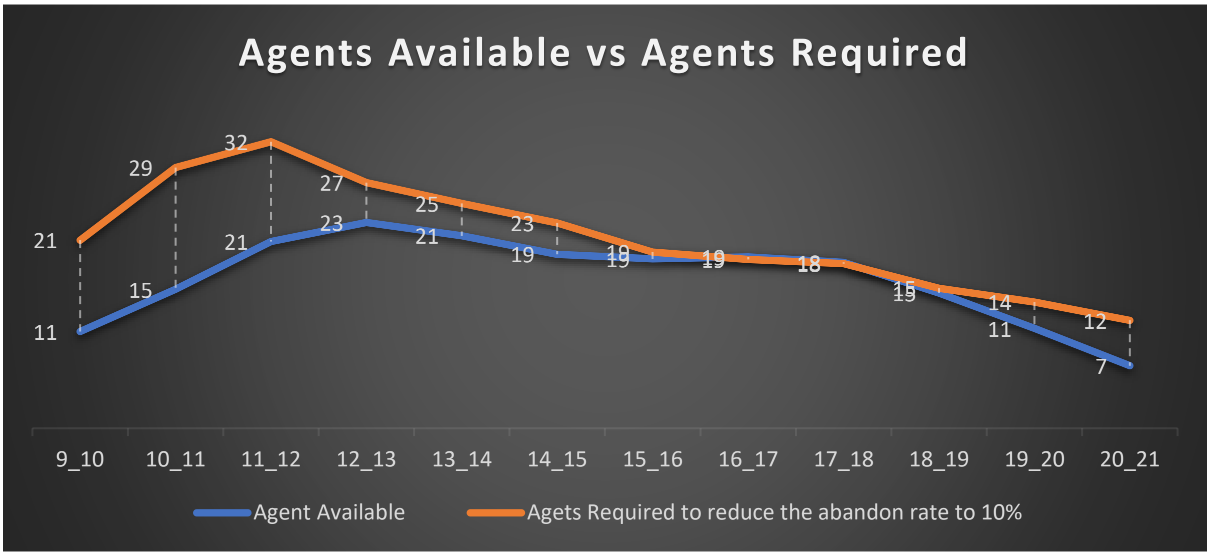
Given the high call volumes in the late morning, it would be beneficial to allocate more staff during this time to handle the increased workload. Conversely, fewer staff might be needed during the evening when call volumes are lower, but it's important to consider the higher average call durations during these times.

3. Manpower Planning

Abandoned calls:



Agents available vs Agents required:



Insights:

Peak Call Abandonment and Agent Availability:

The highest number of abandoned calls occurred between 10:00 and 11:00, with 300 calls abandoned. During this time, there were only 15 agents available, while 29 were required to reduce the abandonment rate to 10%. This indicates a significant gap in agent availability during peak call times, leading to a high abandonment rate.

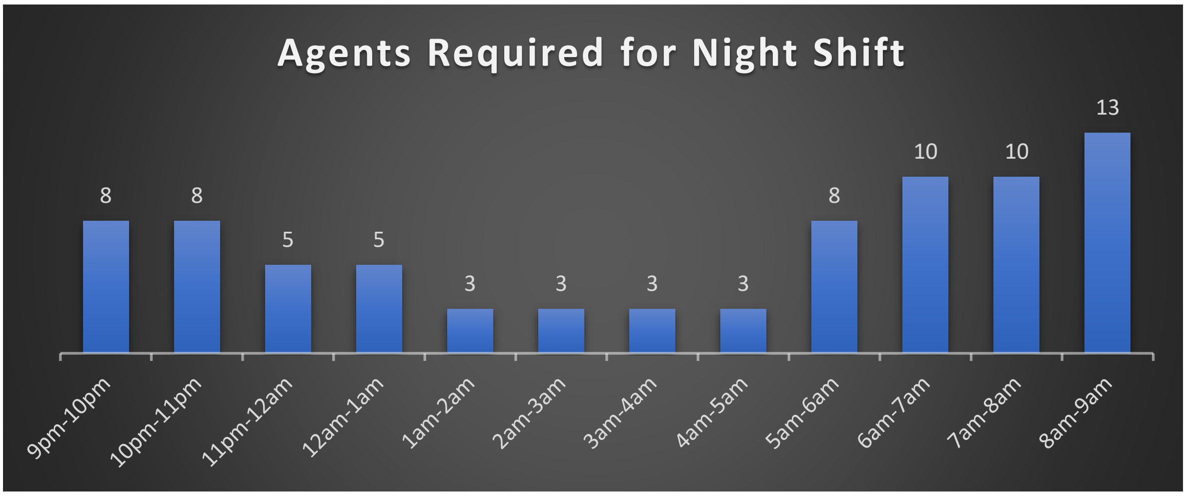
Agent Demand vs. Supply:

Between 11:00 and 12:00, the number of abandoned calls was 262, and the available agents increased to 21. However, the required number of agents was 32, showing that while the availability improved, there was still a shortfall of 11 agents. This period still had a high abandonment rate due to insufficient staffing compared to demand.

Improvement in Afternoon:

In the afternoon hours, particularly from 15:00 to 17:00, the number of abandoned calls dropped significantly to 53 and then further to 32. During this time, the gap between available agents and those required to meet the target abandonment rate narrowed. This suggests that call volume and the need for agents decreased, aligning more closely with the available workforce.

4. **Night Shift Manpower Planning**



Insights:

Peak Call Volume Periods:

The highest call volumes occur between 8am-9am, with 229 calls, requiring 13 agents. This indicates that the period just before the shift ends sees the most activity, which necessitates a larger number of agents to handle the workload.

Low Call Volume Periods:

The lowest call volumes are between 1am-4am, with only 46 calls each hour, requiring 3 agents per hour. This suggests that during these hours, fewer agents are needed, potentially allowing for a reduced workforce or redistribution of agents to busier times.

Steady Workload Distribution:

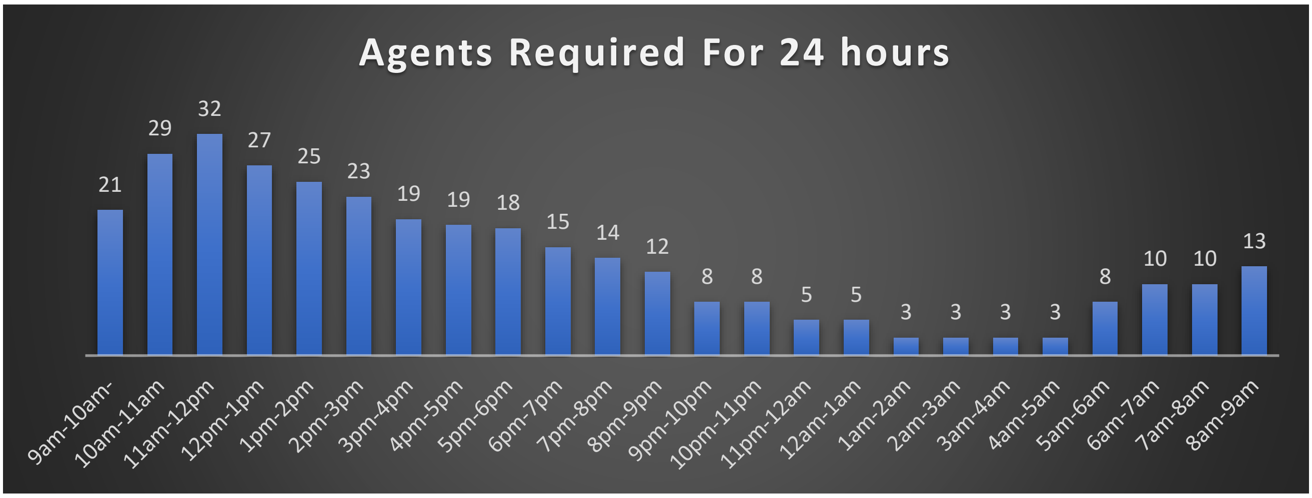
The periods of 9pm-10pm, 10pm-11pm, and 5am-6am have similar call volumes of 137 calls each, requiring 8 agents for each period. This consistency indicates that these hours experience a moderate but steady flow of calls, allowing for a more balanced distribution of agent workload.

Gradual Increase in Call Volume:

From 5am to 9am, there is a noticeable increase in call volume. The call volume starts at 137 calls (5am-6am) and peaks at 229 calls (8am-9am). This indicates a trend where call volumes gradually increase as the morning approaches, requiring a corresponding increase in the number of agents to maintain service levels.

Evening Call Volume Consistency:
The call volumes from 9pm to 12am are relatively consistent, with 137 calls in the first two hours and 91 calls in the third hour. This suggests that the early part of the night shift (9pm-12am) maintains a steady demand for agent support, which can be efficiently managed with a consistent number of agents during this time.

Agents Required for whole 24hrs:



Appendix

Link for Instagram user Analytics:

https://drive.google.com/file/d/1W_-twpqlZxLXAvwVc20IRDq7ie63lM3o/view?usp=drive_link

Link for Operation & Metric Analytics:

https://drive.google.com/file/d/1zTFqqjc5YLFMLiLiYf_f2W4j724p28NO/view?usp=drive_link

Link for Hiring Process Analytics:

https://docs.google.com/spreadsheets/d/1HWLzpwENUvt9UE4K9BwySGB0sat6FHQF/edit?usp=drive_link&ouid=104841036983256583952&rtpof=true&sd=true

Link for IMDB Movie Analysis:

https://docs.google.com/spreadsheets/d/1jG2SMuPEstb4Ik0kX6z4dOKYEVmr5iFZ/edit?usp=drive_link&ouid=104841036983256583952&rtpof=true&sd=true

Link for Bank Loan Case Study:

https://docs.google.com/spreadsheets/d/10N_oAw0DOry3BypZ4QsZvec9L_mMzl1m/edit?usp=drive_link&ouid=104841036983256583952&rtpof=true&sd=true

Link for Impact of Car Features:

https://docs.google.com/spreadsheets/d/1UpRs-pMi9ANcaNF1OWtMibmGb8LtjYjU/edit?usp=drive_link&ouid=104841036983256583952&rtpof=true&sd=true

Link for ABC Call Volume Trend:

https://docs.google.com/spreadsheets/d/1xcHCGmIOv_-1QPBddIJRx1eYD4pzuYv5/edit?usp=drive_link&ouid=104841036983256583952&rtpof=true&sd=true