DATA ANALYTICS PORTFOLIO

Shreya Gunjal

ASPIRING DATA ANALYST

A portfolio showcasing projects where I've applied data-driven techniques to solve real-world problems and uncover insights. With a strong focus on continuous learning, I am dedicated to refining my analytical skills and expanding my expertise as I pursue a career in data analytics.

PROFESSIONAL BACKGROUND

I hold a Master's degree in Biochemistry and previously worked as a Junior Scientist in a diagnostic laboratory, where I was responsible for genetic testing and the analysis of clinical data. My experience in the healthcare and life sciences sector has not only deepened my interest in healthcare analytics but also cultivated a strong foundation in analytical thinking, precision, and problem-solving—skills that are highly transferable to data analytics in any industry.

In the lab, I regularly worked with complex lab techniques under strict regulatory standards, which demanded attention to detail, data integrity, and clear documentation. These responsibilities have naturally translated into proficiency in analytical thinking, data cleaning, exploratory analysis, and communicating insights effectively. Additionally, my background required collaboration across multidisciplinary teams, sharpening my communication, adaptability, and time management skills.

As I transition into data analytics, I bring a blend of domain knowledge, curiosity, and a solid technical skill set, including SQL, Excel, Python and data visualization tools. While I am particularly interested in healthcare-related analytics, I am equally enthusiastic about applying my analytical abilities to diverse fields and uncovering meaningful insights that drive decisions.

TABLE OF CONTENTS

INSTAGRAM USER ANALYTICS

DESCRIPTION:

As a data analyst collaborating with Instagram's product team, tasked with examining user interactions to deliver insights that drive strategic growth. Understanding user behavior helps teams across the company, including marketing, product, and development, make informed decisions such as launching targeted campaigns, designing impactful features, and enhancing user experience.

In this project, you'll leverage SQL and MySQL Workbench to analyze Instagram user data, providing valuable answers to key business questions. Your insights will empower the product team to make data-driven decisions that could shape the platform's future success.

THE PROBLEM:

A) **Marketing Analysis:** The marketing team seeks data-driven insights to optimize user engagement, reward loyalty, and improve campaign effectiveness on Instagram.

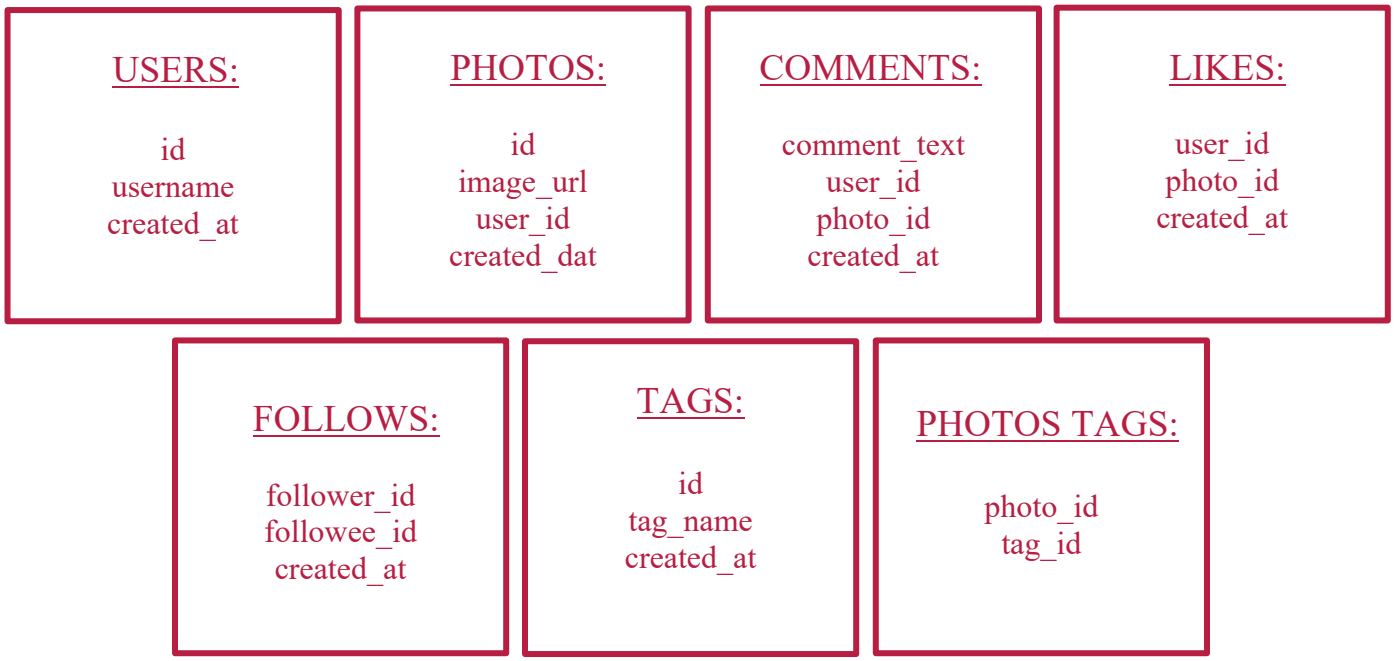
Objective	Task
a. Loyal User Reward: The marketing team wants to reward the most loyal users, i.e., those who have been using the platform for the longest time.	Identify the five oldest users on Instagram from the provided database.
b. Inactive User Engagement: The team wants to encourage inactive users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

B) **Investor Metrics Overview:** Investor metrics help assess a platform's health, user activity, and authenticity, providing insights into its value and growth potential.

Objective	Task
1. User Engagement: Investors want to know if users are still active and posting on Instagram or if they are making fewer posts.	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.
2. Bots & Fake Accounts: Investors want to know if the platform is crowded with fake and dummy accounts.	Identify users (potential bots) who have liked every single photo on the site, as this is not typically possible for a normal user.

DESIGN:

1. MySQL Workbench 8.0 was used to create a database and then solve the problems.
2. Using CREATE DATABASE and CREATE TABLE commands in SQL Workbook, we create the database 'ig_clone' having 7 tables and the following columns in it:



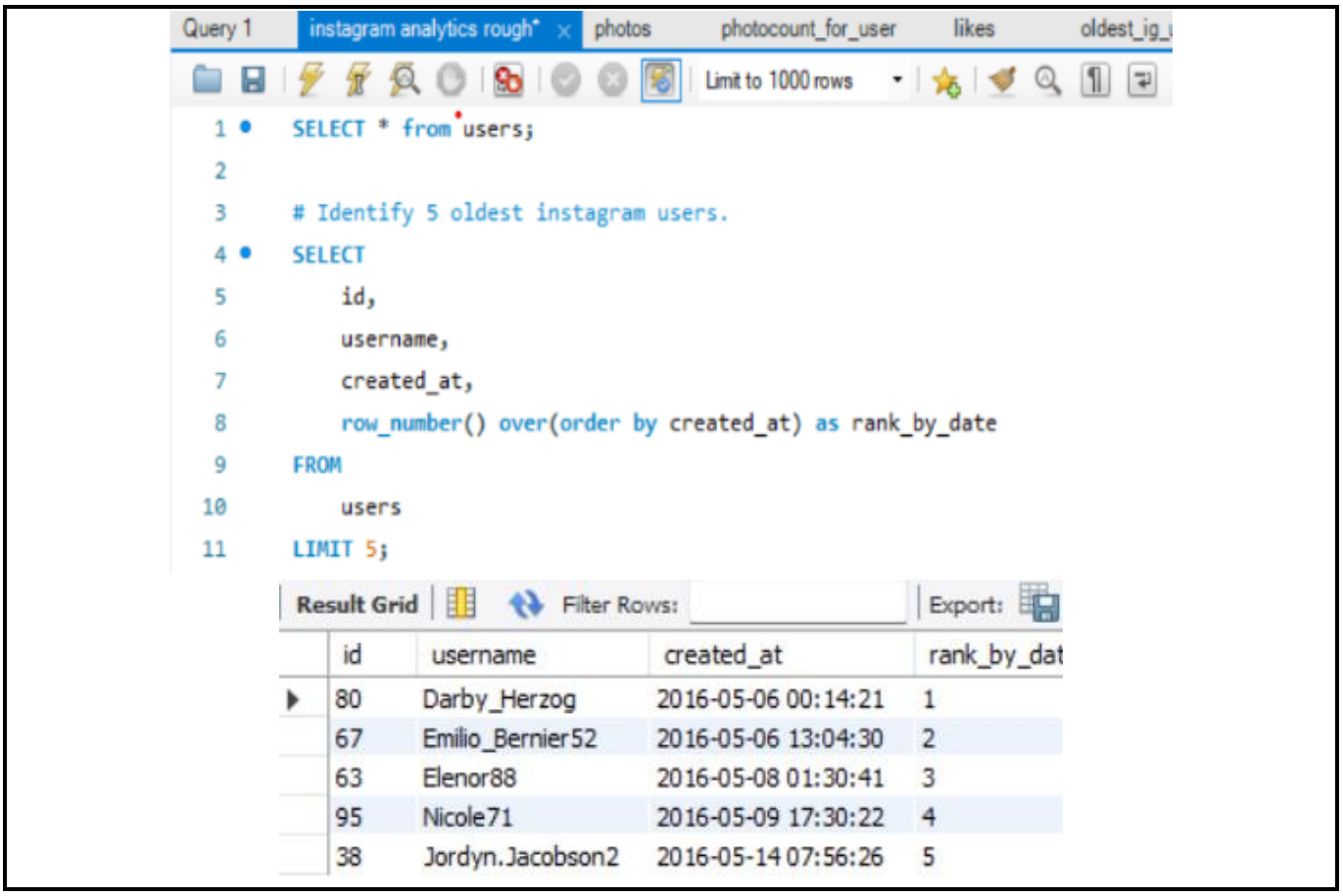
3. Using INSERT INTO command the tables and columns were given values. PRIMARY KEY and FOREIGN KEY were used to link multiple tables with each other for ease.

FINDINGS:

A) **Marketing Analysis:**

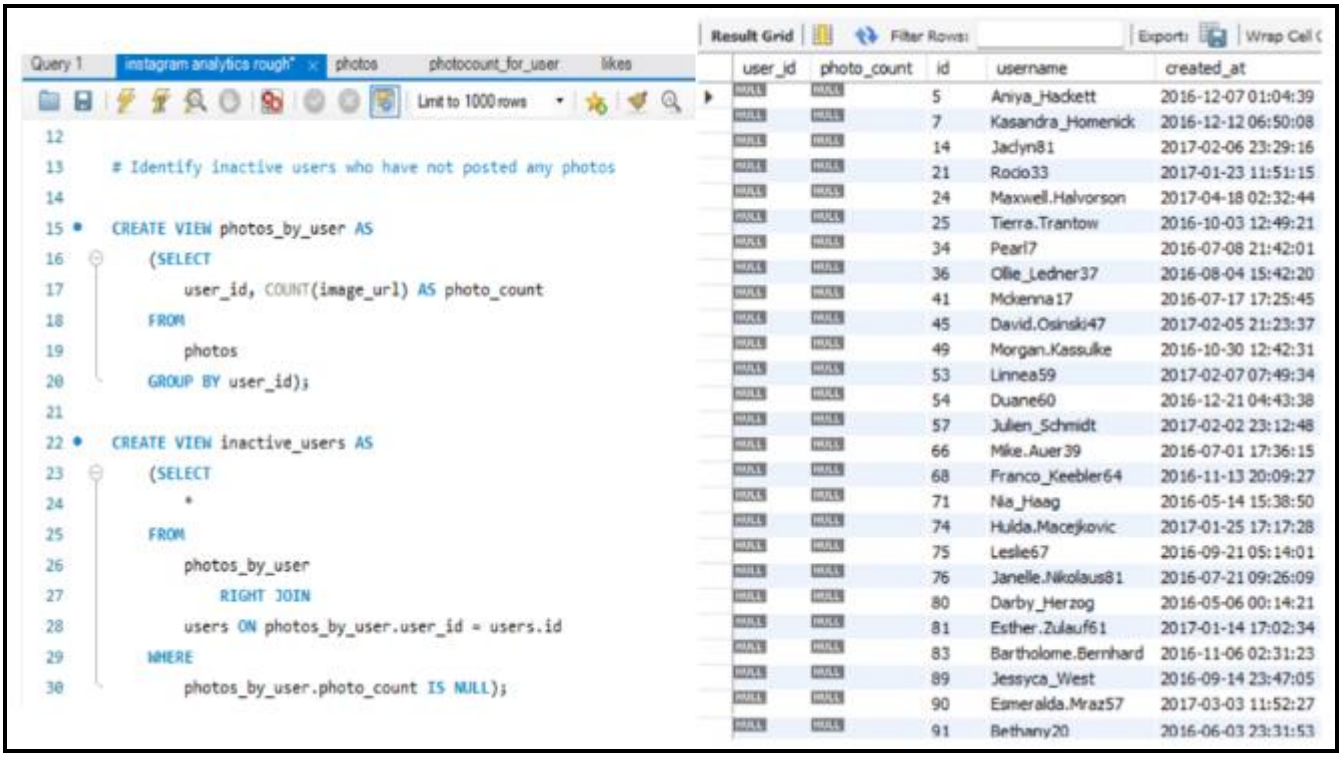
1. Loyal User Reward:

a. The appropriate columns were selected from the **USERS** table alongwith a new column named 'rank_by_date'. This new column was created by using **ORDER BY** function and ranked using **ROW_NUMBER** function. **LIMIT** was used to select top 5 results based on 'rank_by_date'.



2. Inactive User Engagement:

a. Using **CREATE VIEW**, a view called '**PHOTOS_BY_USER**' having 'user_id' and 'photo_count' columns from the table **PHOTOS** was made and grouped by user_id using **GROUP BY** function.
b. The new column 'photo_count' was created by using **COUNT** function on 'image_url' column.
c. A second view '**INACTIVE_USERS'** was made by taking '**PHOTOS_BY_USER**' table and joining it with **USERS** table by the 'user_id' column using the **RIGHT JOIN** function.
d. Using **WHERE** alongwith **IS NULL** functions only user ids with no photos would be displayed.

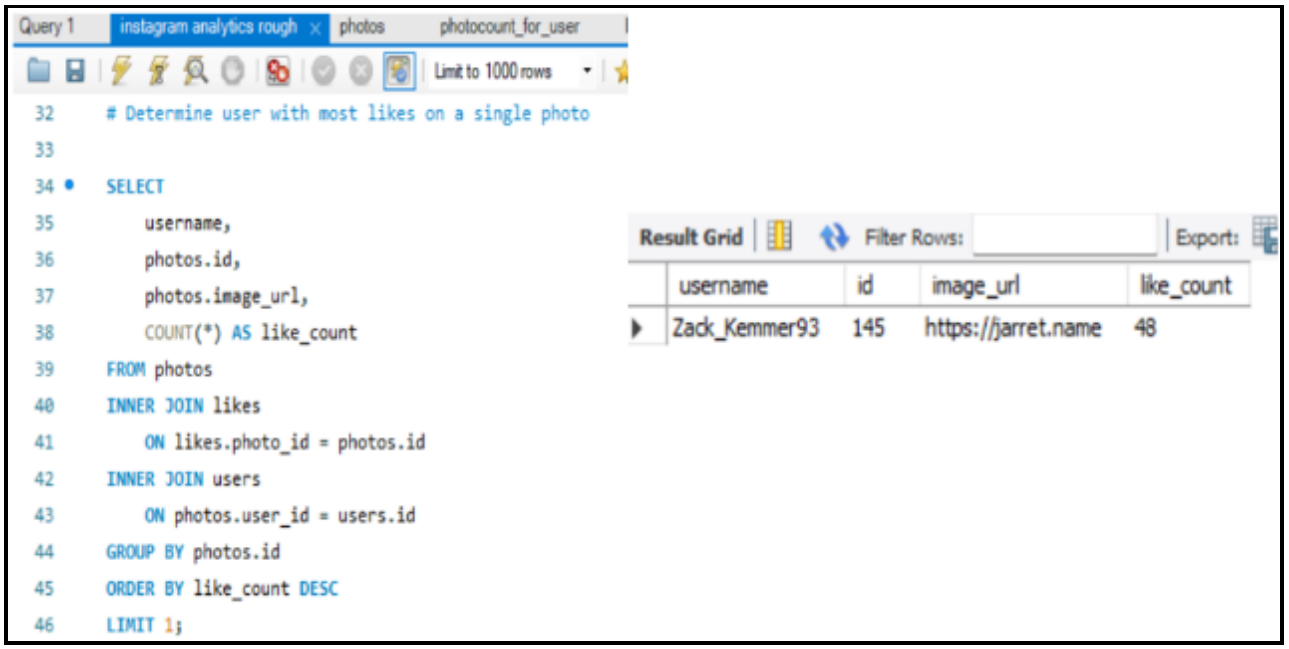


```
# Identify inactive users who have not posted any photos

CREATE VIEW photos_by_user AS
    (SELECT
        user_id, COUNT(image_url) AS photo_count
    FROM
        photos
    GROUP BY user_id);

CREATE VIEW inactive_users AS
    (SELECT
        *
    FROM
        photos_by_user
            RIGHT JOIN
        users ON photos_by_user.user_id = users.id
    WHERE
        photos_by_user.photo_count IS NULL);
```

user_id	photo_count	id	username	created_at
NULL	NULL	5	Aniya_Hackett	2016-12-07 01:04:39
NULL	NULL	7	Kasandra_Homenick	2016-12-12 06:50:08
NULL	NULL	14	Jaclyn81	2017-02-06 23:29:16
NULL	NULL	21	Rocio33	2017-01-23 11:51:15
NULL	NULL	24	Maxwell.Halvorson	2017-04-18 02:32:44
NULL	NULL	25	Tierra.Trantow	2016-10-03 12:49:21
NULL	NULL	34	Pearl7	2016-07-08 21:42:01
NULL	NULL	36	Ollie_Ledner37	2016-08-04 15:42:20
NULL	NULL	41	Mckenna17	2016-07-17 17:25:45
NULL	NULL	45	David.Osinski47	2017-02-05 21:23:37
NULL	NULL	49	Morgan.Kassulke	2016-10-30 12:42:31
NULL	NULL	53	Linnea59	2017-02-07 07:49:34
NULL	NULL	54	Duane60	2016-12-21 04:43:38
NULL	NULL	57	Julien_Schmidt	2017-02-02 23:12:48
NULL	NULL	66	Mike.Auer39	2016-07-01 17:36:15
NULL	NULL	68	Franco_Keebler64	2016-11-13 20:09:27
NULL	NULL	71	Nia_Haag	2016-05-14 15:38:50
NULL	NULL	74	Hulda.Macejkovic	2017-01-25 17:17:28
NULL	NULL	75	Leslie67	2016-09-21 05:14:01
NULL	NULL	76	Janelle.Nikolaus81	2016-07-21 09:26:09
NULL	NULL	80	Darby_Herzog	2016-05-06 00:14:21
NULL	NULL	81	Esther.Zulauf61	2017-01-14 17:02:34
NULL	NULL	83	Bartholome.Bernhard	2016-11-06 02:31:23
NULL	NULL	89	Jessyca_West	2016-09-14 23:47:05
NULL	NULL	90	Esmeralda.Mraz57	2017-03-03 11:52:27
NULL	NULL	91	Bethany20	2016-06-03 23:31:53

3. Contest Winner Declaration:

a. Columns ‘username’, ‘id’, ‘image_url’ and ‘like_count’ (new column) were selected from the table **PHOTOS**. This new column was created using the **COUNT** function.
b. Using the **INNER JOIN** function, the **PHOTOS** table was joined by the ‘id’ column to the **LIKES** table’s ‘photo_id’ column.
c. Then, **USERS** table was joined by the ‘id’ column to **PHOTOS** table’s ‘user_id’ column using the same function. These were grouped by ‘id’ values of the **PHOTOS** table.
d. These were arranged using **ORDER BY** on the ‘like_count’ column and the **LIMIT** was set to 1.

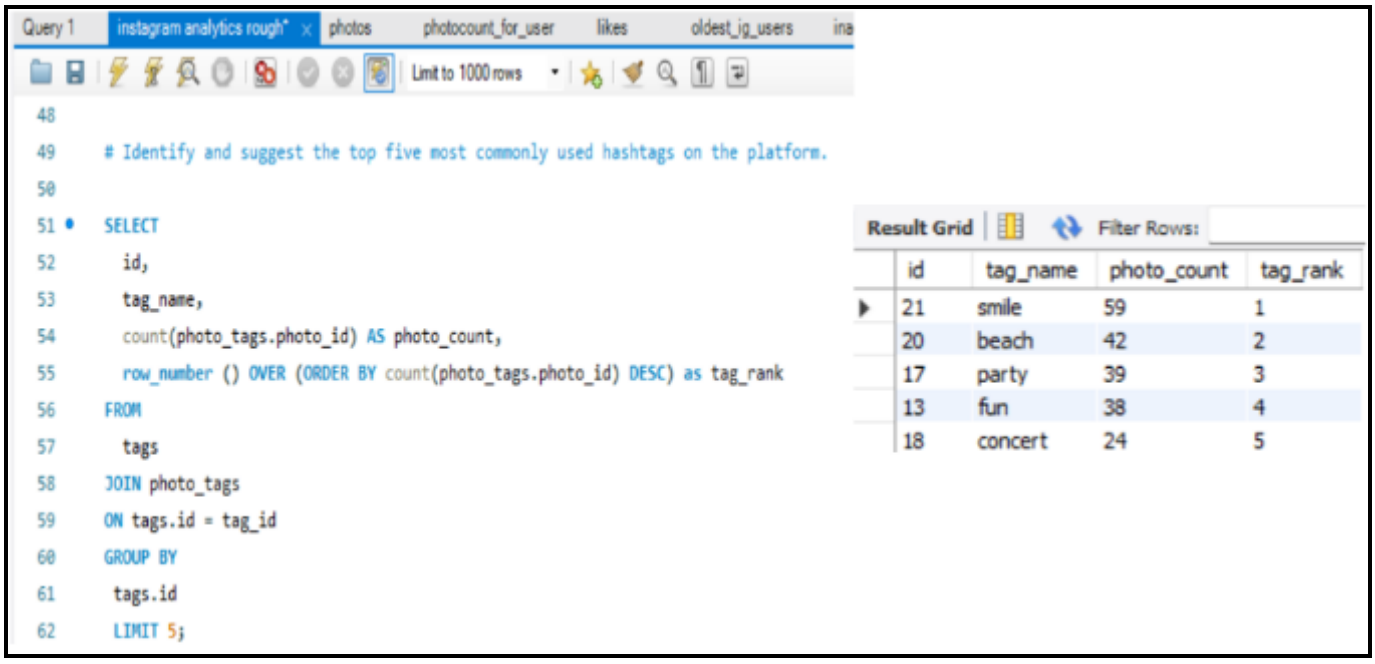


```
# Determine user with most likes on a single photo

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

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

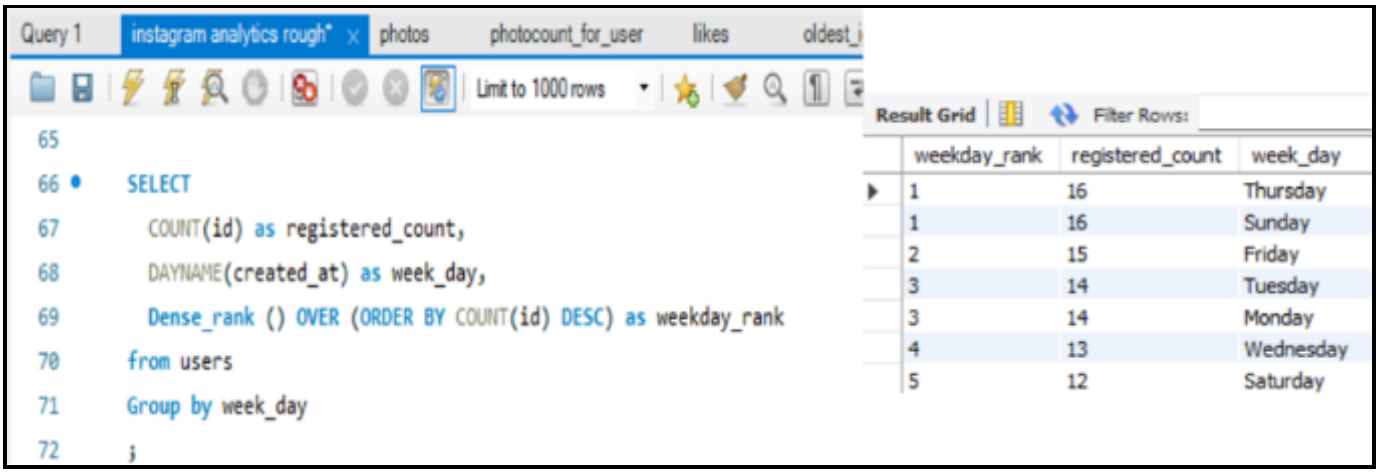
4. Hashtag Research:

a. Columns like ‘id’, ‘tag_name’ and ‘photo_count’ were selected from the **TAGS** table. ‘photo_count’ was created by using **COUNT** function on ‘photo_id’ from **PHOTO_TAGS** table.
b. A new column called ‘tag_rank’ was created using **ORDER BY** function having ‘photo_count’ column arranged using **ROW_NUMBER ()** function in a descending order (**DESC** function).
c. Using the **JOIN** function, the **PHOTO_TAGS** table was joined by the ‘tag_id’ column to the **TAGS** table (id’ column). After grouping by ‘id’ values from **TAGS** table using **GROUP BY** function the **LIMIT** was set to 5.



5. Ad Campaign Launch:

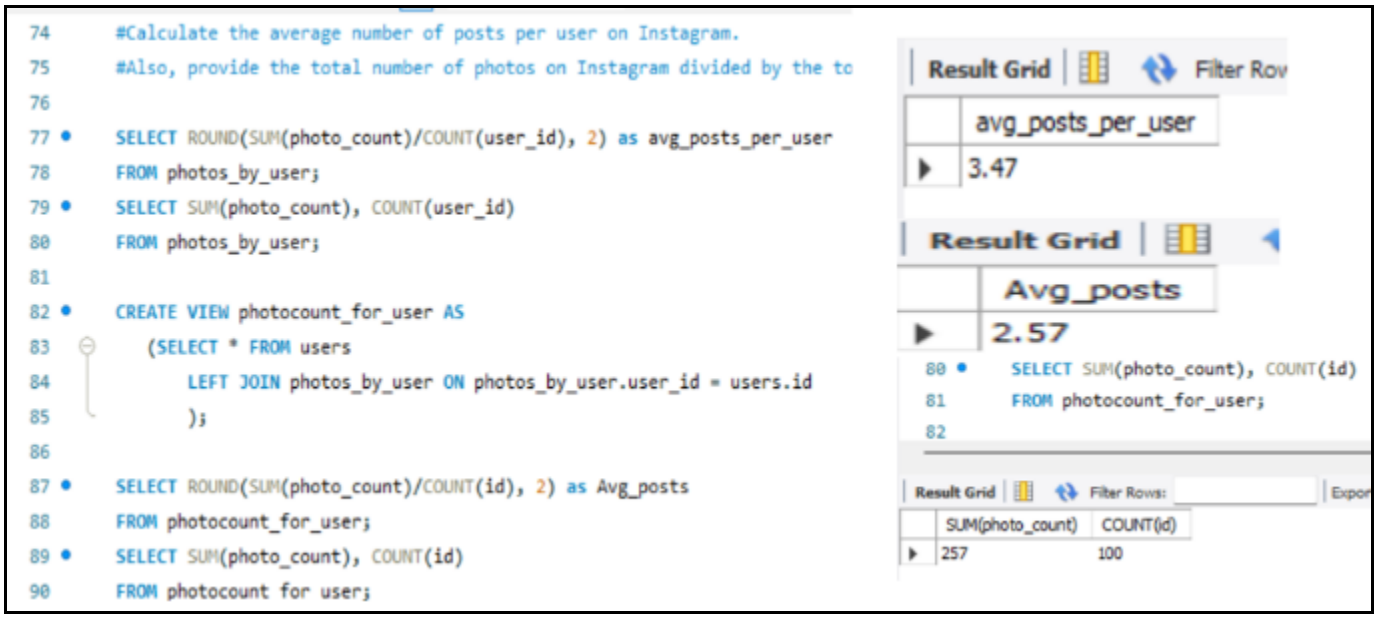
a. From **USERS** table, ‘registered_count’ was created using **COUNT** function on ‘id’ column, ‘week_day’ using **DAYNAME** function on ‘created_at’ column and ‘weekday_rank’ was created by arranging ‘registered_count’ column by **DENSE_RANK**, **ORDER BY** and **DESC** functions.
b. ‘weekday_rank’ was made to rank the week days in a descending order of ‘registered_count’ on the weekday. The results were grouped by ‘week_day’ values.



B. **Investor Metrics Overview:**

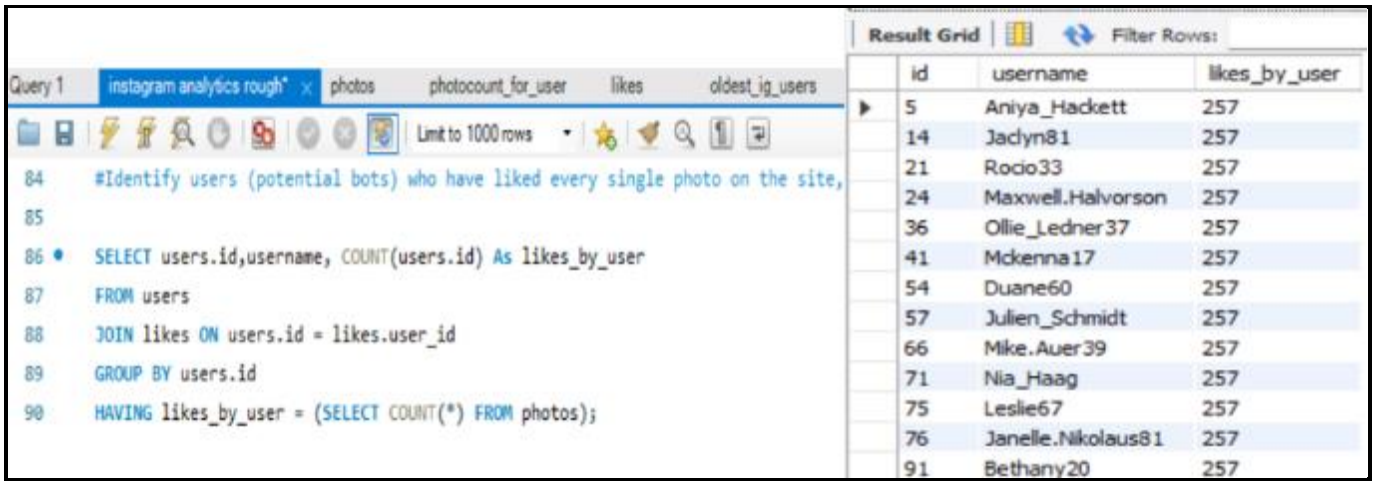
a) User Engagement:

a. Using the view '**PHOTOS_BY_USER**', the **SUM** of 'photo_count' was divided by the **COUNT** of 'user_id') and represented as 'avg_posts_per_user'. This is the average no. of posts made active users.
b. A new view '**PHOTOCOUNT_FOR_USER'** was made by joining the **USERS** table by the 'id' column to the '**PHOTOS_BY_USER**' view using the **LEFT JOIN** function**.**
c. Using the new view made, the **SUM** of 'photo_count' was divided by the **COUNT** of 'id' and represented as 'Avg_posts'. This is the average no. of posts made by all users (active and inactive).



2. Bots & Fake Accounts:

a. Columns were selected from the **USERS** table alongwith a new column named 'likes_by_user' created using **COUNT** function on 'id' column.
b. The table **LIKES** was joined by the 'user_id' column to the **USERS** table and grouped by 'id'.
c. A condition was placed by **HAVING** function to only show users where the 'likes_by_user' value was equal to **COUNT** of entries from **PHOTOS** table.



ANALYSIS:

1. As a marketing strategy it is important that we launch contests that help users feel appreciated for their loyalty and encourage active participation on the platform. The five oldest users on the platform are: Darby_Herzog, Emilio_Bernier52, Elenor88, Nicole71 and Jordyn.Jacobson2.

2. There are 26 inactive users who have never posted on the platform. They are listed previously in the findings. It may help to send them promotional emails to help increase activity on the platform.

3. Another strategy for marketing was to reward the user with most liked post on the platform. The reward goes to Zack_Kemmer93 with 48 likes, user id 145 for image URL https://jarret.name .

4. To better company's market brand and get brand deals, seems like the following 5 hashtags are the most popular ones and can be incorporated to promote brands: #smile, #beach, #party, #fun and #concert in that order.

5. It seems like Sunday and Thursday are the two most active days on the platform when it comes to account creation. It might benefit to send promotional emails on these days.

6. There is a total of 257 posts made on the platform and a total of 100 users, making user engagement an average of 2.57 posts per user. However, the user engagement for active users is 3.47 posts per active user. It might help to encourage all users to post more.

7. 13 accounts have been identified as bots or fake accounts as they have liked every single photo. Eliminating these will help make the platform more authentic.

CONCLUSION:

This project helped me realize that all data is important. Something as simple as the data on likes can be used to arrive to significant conclusions such as identifying bots/ fake accounts. Another insight is how the user engagement affects brand deals for every company. Popular trends and hashtags can be used to market products and gain confidence of other companies and this can be used to negotiate collaborations to push our own business further.

While we realize that ad campaigns are important for every company, its important that we save resources all the while be more efficient by targeting campaigns on days when people are more likely to respond to them. Such an approach is better than targeting everyone at all times.

Representation of data is definitely as important as running queries and getting results. We have to present the data in a report rather than simply answering questions. It is no surprise that all companies require data analysts for their businesses to push marketing teams, investment and finance teams and other teams to make better decisions using analysis provided by data analysts.

OPERATION ANALYTICS AND INVESTIGATING METRIC SPIKE

DESCRIPTION:

Operational Analytics involves examining a company's complete operations to identify areas for improvement. As a Data Analyst, you'll collaborate with teams across operations, support, and marketing to extract valuable insights from collected data. A critical aspect of this role is investigating sudden changes in key metrics, such as dips in user engagement or sales. This requires a strong understanding of how to analyze data patterns and fluctuations effectively.

In this project, you'll step into the role of a Lead Data Analyst at a company similar to Microsoft. Using advanced SQL, you'll analyze various datasets to answer questions from different departments, providing insights that can help enhance operations and explain metric variations.

THE PROBLEM:

A. Case Study 1: Job Data Analysis

Objective	Task
1. **Jobs Reviewed Over Time:** Calculate the number of jobs reviewed per hour for each day in November 2020.	Write an SQL query to calculate the number of jobs reviewed per hour for each day in November 2020.
2. **Throughput Analysis:** Calculate the 7-day rolling average of throughput (number of events per second).	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.
3. **Language Share Analysis:** Calculate the percentage share of each language in the last 30 days.	Write an SQL query to calculate the percentage share of each language over the last 30 days.
4. **Duplicate Rows Detection:** Identify duplicate rows in the data	Write an SQL query to display duplicate rows from the job_data table.

B. Case Study 2: Investigating Metric Spike

Objective	Task
1. **Weekly User Engagement:** Measure the activeness of users on a weekly basis.	Write an SQL query to calculate the weekly user engagement.
2. **User Growth Analysis:** Analyze the growth of users over time for a product.	Write an SQL query to calculate the user growth for the product.
3. **Weekly Retention Analysis:** Analyze the retention of users on a weekly basis after signing up for a product.	Write an SQL query to calculate the weekly retention of users based on their sign-up cohort.
4. **Weekly Engagement Per Device:** Measure the activeness of users on a weekly basis per device.	Write an SQL query to calculate the weekly engagement per device.
5. **Email Engagement Analysis:** Analyze how users are engaging with the email service	Write an SQL query to calculate the email engagement metrics.

DESIGN:

1. MYSQL Workbench 8.0 was used to import the database, make tables and run queries on the database while Microsoft Excel was used to visualize the results and represent them.
2. Four tables were to be used. For the first case study the table "**JOB_DATA**" was used with the following columns:

3. The second case study involved using following three tables:
   - **users:** Contains one row per user, with descriptive information about that user's account.
   - **events:** Contains one row per event, where an event is an action that a user has taken (e.g., login, messaging, search).
   - **email_events:** Contains events specific to the sending of emails

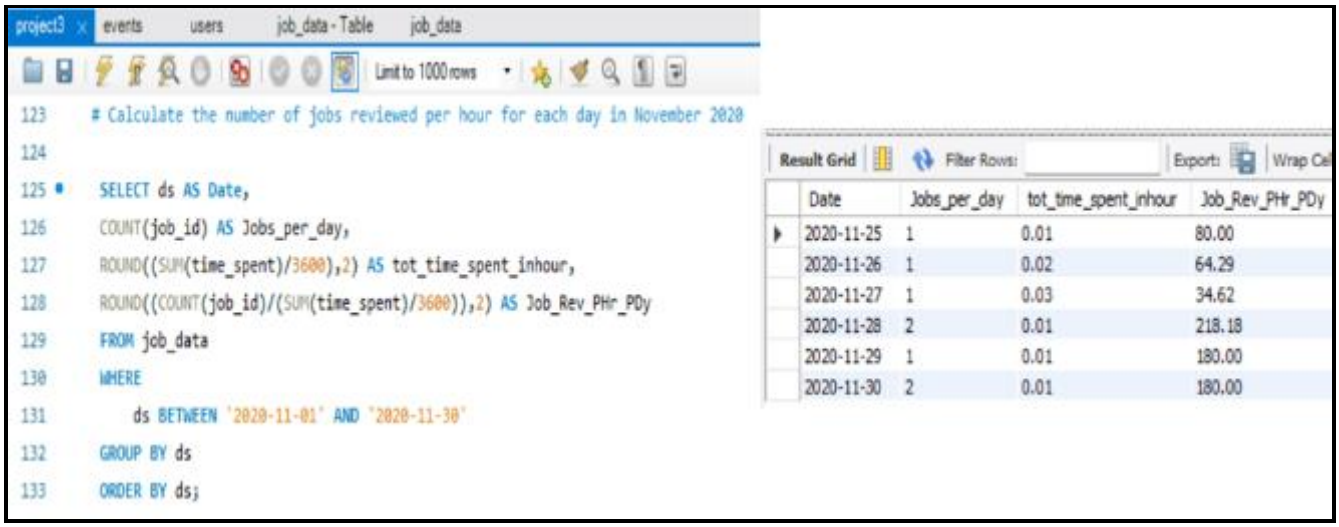
4. Using **CREATE DATABASE** and **CREATE TABLE** commands in SQL Workbook, we create the database 'project_3' having 4 tables. Using **INSERT INTO** command the table "job_data" was given values.
5. The other three tables were also created and CSV files were imported to insert the data into the tables. The date and time columns were altered to appropriate formats using **ALTER TABLE**, **DAYTIME** and **STR_TO_DATE** functions.

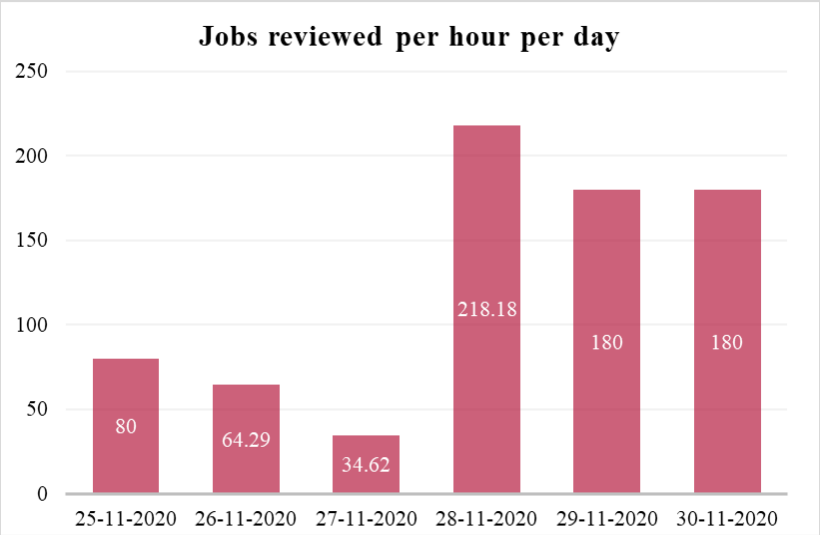
FINDINGS:

Case Study 1: Job Data Analysis

1. Jobs Reviewed Over Time:

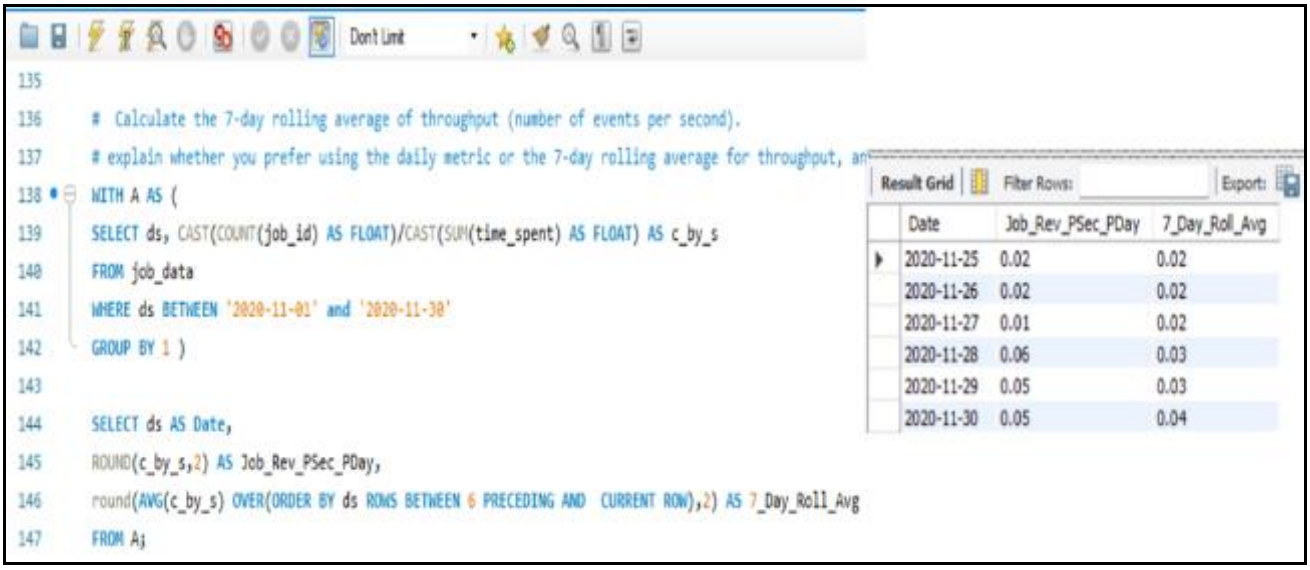
a. From the table **JOB_DATA**, 'ds' column was selected alongwith **SUM** of 'time_spent' divided by 3600 (as 'tot_time_spent_inhour') and **COUNT** of 'job_id' divided by 'tot_time_spent_inhour' (as 'Job_Rev_PHr_PDy').
b. Using **WHERE** function, only the results that had 'ds' values **BETWEEN** '2020-01-11' and '2020-01-30' were displayed.
c. Also, these were arranged and grouped by their 'ds; values using **ORDER BY** and **GROUP BY** functions.





2. Throughput Analysis:

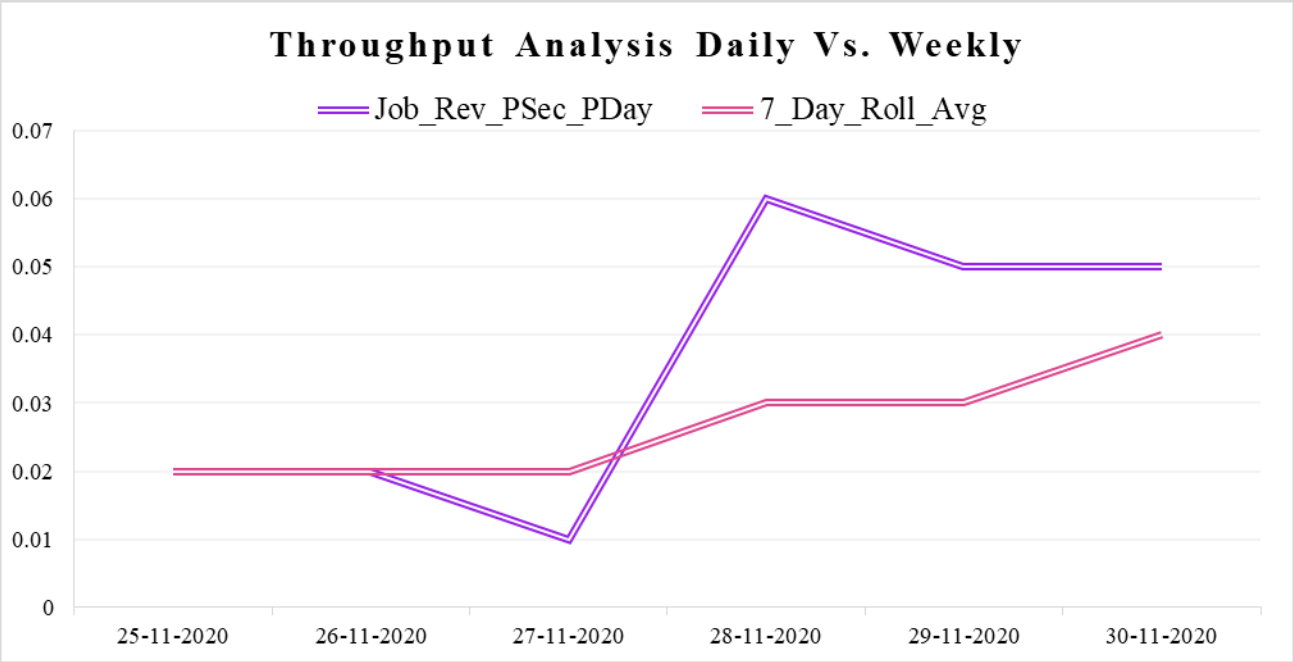
a. A temporary table **A** was created with column ds, c_by_s from table **JOB_DATA**. A condition was placed using **WHERE** to only show results where ds value was between 2020-11-01 to 2020-11-30.
b. The column c_by_s was calculated by dividing **COUNT** of job_id by **SUM** of job_id using the **CAST** and **AS FLOAT** functions. The data was grouped by ds using **GROUP BY** function.
c. From **A**, columns ds, 'Job_Rev_PSec_PDay' (**ROUND** value of c_by_s) and '7_Day_Roll_Avg' were selected. Average of c_by_s was taken and **OVER ()** clause was used to define how this average is computed across rows.
d. The clause **ROWS BETWEEN 6 PRECEDING AND CURRENT ROW** specifies a rolling window of seven days. For each row, it takes the current row's c_by_s value and the previous six rows' values and then averages them.
e. The function **ORDER BY** ensures that the calculation follows the chronological order of dates.



```
135
136     # Calculate the 7-day rolling average of throughput (number of events per second).
137     # explain whether you prefer using the daily metric or the 7-day rolling average for throughput, an
138     WITH A AS (
139     SELECT ds, CAST(COUNT(job_id) AS FLOAT)/CAST(SUM(time_spent) AS FLOAT) AS c_by_s
140     FROM job_data
141     WHERE ds BETWEEN '2020-11-01' and '2020-11-30'
142     GROUP BY 1 )
143
144     SELECT ds AS Date,
145     ROUND(c_by_s,2) AS Job_Rev_PSec_PDay,
146     round(AVG(c_by_s) OVER(ORDER BY ds ROWS BETWEEN 6 PRECEDING AND  CURRENT ROW),2) AS 7_Day_Roll_Avg
147     FROM A;
```

Result Grid | Filter Rows: | Export:

Date	Job_Rev_PSec_PDay	7_Day_Roll_Avg
2020-11-25	0.02	0.02
2020-11-26	0.02	0.02
2020-11-27	0.01	0.02
2020-11-28	0.06	0.03
2020-11-29	0.05	0.03
2020-11-30	0.05	0.04



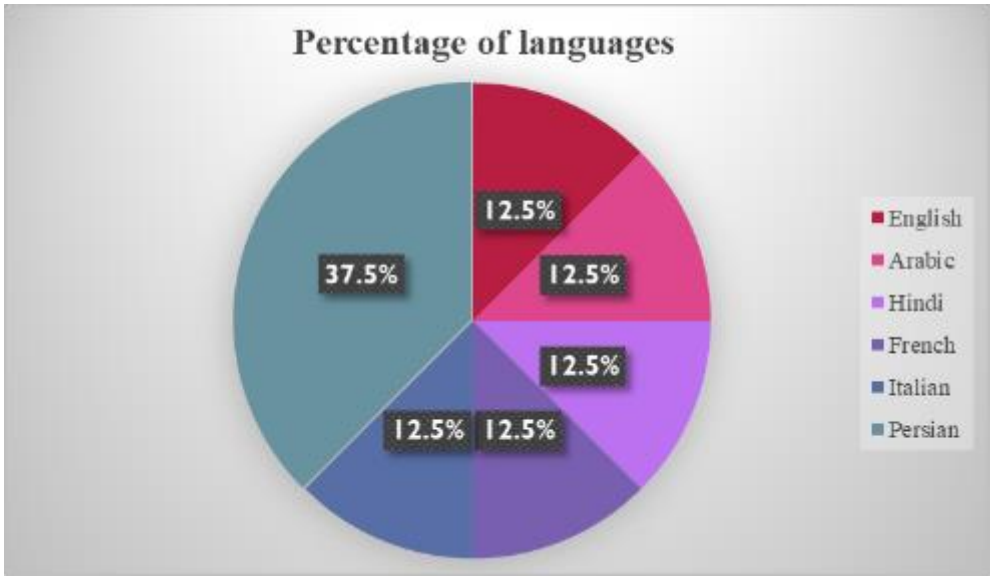
3. Language Share Analysis:

a. A temporary table **L** with language and 'lang_count' (**COUNT** of language) was created. The data was grouped by language using **GROUP BY**.
b. A condition was placed using **WHERE** to count occurrences of each language in **JOB_DATA** where ds value was **BETWEEN** 2020-11-01 and 2020-11-30. to
c. From **L**, the final query retrieves the language, lang_count, and the percentage contribution of each language to the total.
d. The percentage (Perc_Lang) is calculated by dividing each lang_count by the **SUM** of count of all languages and multiplying by 100. The results are sorted in descending order of Perc_Lang.

```
148
149     #  calculate the percentage share of each language over the last 30 days.
150
151 •   WITH L AS (
152     SELECT language, COUNT(language) AS lang_count
153     FROM job_data
154     WHERE ds BETWEEN '2020-11-01' AND '2020-11-30'
155     GROUP BY language )
156
157     SELECT language AS Lang, lang_count,
158     ROUND((100*lang_count/SUM(lang_count) OVER()), 2) AS Perc_Lang
159     FROM L
160     ORDER BY Perc_Lang;
```

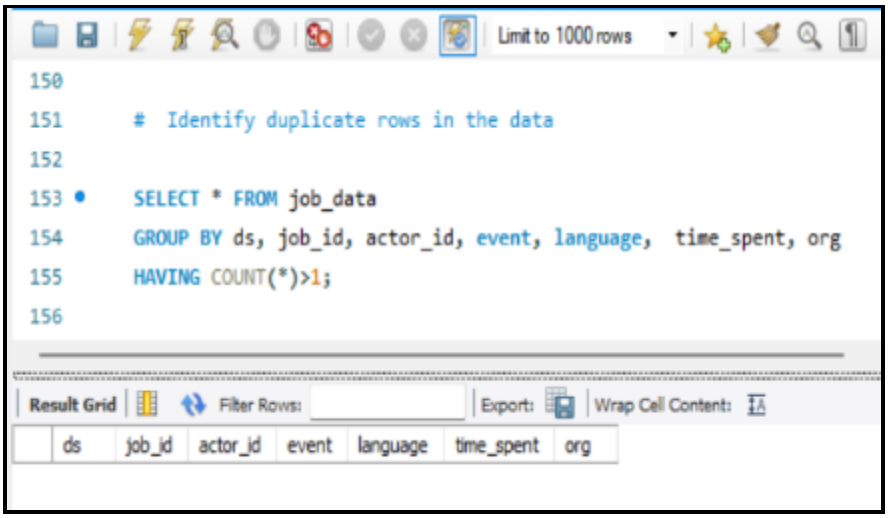
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Lang	lang_count	Perc_Lang
English	1	12.50
Arabic	1	12.50
Hindi	1	12.50
French	1	12.50
Italian	1	12.50
Persian	3	37.50



4. Duplicate Rows Detection:

a. From table **JOB_DATA**, the data was grouped by all columns of the table using **GROUP BY**. Using **HAVING** function a condition was placed where only data that had **COUNT** of all rows as more than 1 was shown.



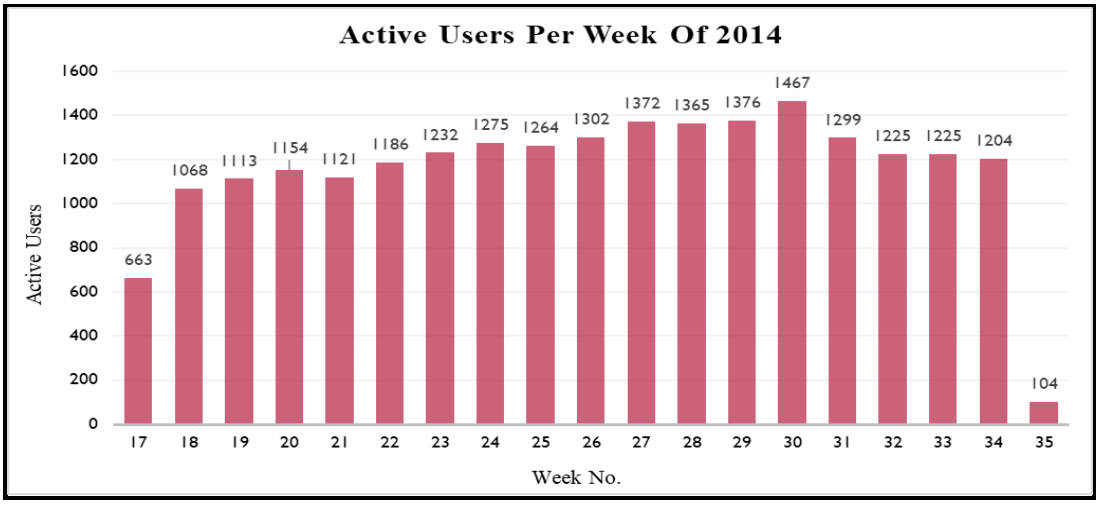
Case Study 2: Investigating Metric Spike

1. Weekly User Engagement:

a. A new view, '**USERS_IN_WK**' was created including new columns 'week_of_year', 'active_user' (**DISTINCT** user_id values) and 'week_number' from **EVENTS** table. Using **EXTRACT** and **WEEK FROM** functions week_of_year was created from occurred_at.
b. The column week_number was created by arranging week_of_year column in descending order using **DENSE_RANK ()**, **ORDER BY** and **DESC** functions. The view was grouped by 'week_of_year'.

c. Also, the view was filtered for rows **WHERE** 'event_type' was 'engagement'. From '**USERS_IN_WK'**, **SUM** of 'active_user' was divided by **COUNT** of 'week_of_year' and displayed as 'Avg_Users_PWeek' to get average users per week.
d. A temporary table called **W1** was created by **WITH** function from **EVENTS** columns 'user_id', **COUNT** of 'user_id' as 'Cnt' and 'week_of_year'. These were arranged by 'user_id' (**ORDER BY** function) and grouped by 'user_id' and 'week_of_year' using **GROUP BY** function.
e. Finally, weekly engagements per user was displayed as 'Weekly_Eng_PUser' by averaging and rounding up 'Cnt' from **W1** using **AVG** and **ROUND** functions.

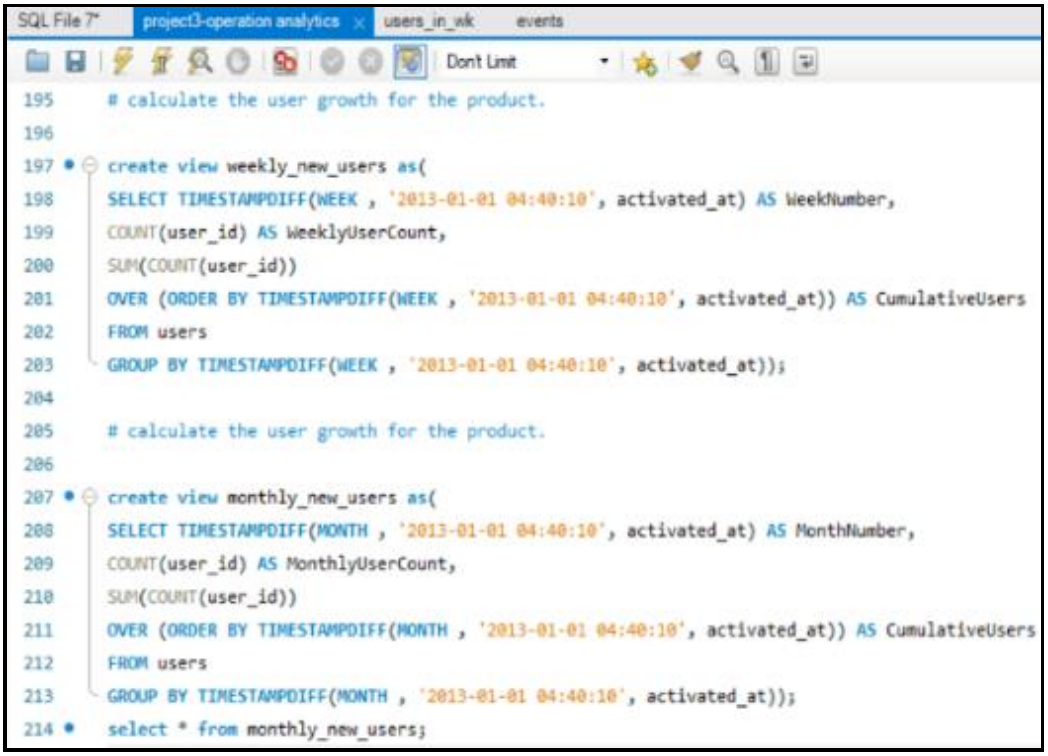
```
SQL File 7*   project3-operation analytics*   users_in_wk   events
                                                Don't Limit
167   # Write an SQL query to calculate the weekly user engagement.
168   create view users_in_wk as
169   (select EXTRACT(WEEK FROM occurred_at) AS week_of_year,
170   count(distinct user_id) as active_user,
171   dense_rank() over(order by EXTRACT(WEEK FROM occurred_at)) as week_number
172   from events
173   where event_type = 'engagement'
174   group by week_of_year
175   );
176
177   # average users per week
178   select ROUND(SUM(active_user)/COUNT(week_of_year),2) as Avg_Users_PWeek
179   from users_in_wk;
180
181   # average weekly user engagement
182   With W1 as (SELECT user_id,
183   EXTRACT(WEEK FROM occurred_at) AS week_of_year,
184   COUNT(user_id) AS Cnt
185    FROM events
186    GROUP BY user_id, week_of_year
187    ORDER BY user_id)
188   SELECT ROUND(AVG(Cnt),2) AS Weekly_Eng_PUser
189    FROM W1;
```

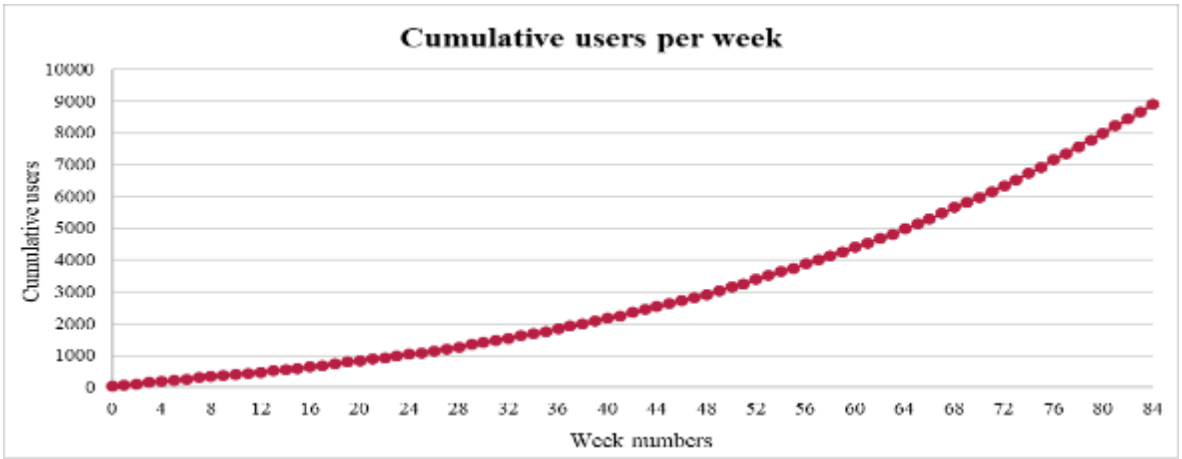


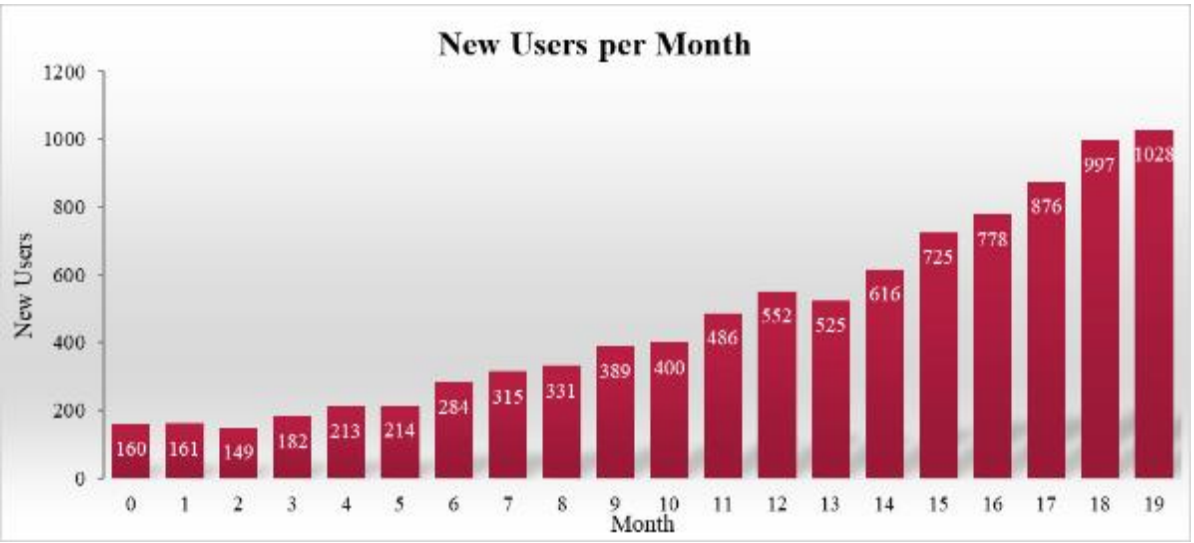
Users Engaging Per Week:	Engagements Per User Per Week:
1158.68	14.77

2. User Growth Analysis:

a) A view **WEEKLY_NEW_USERS** was made from **USERS** with 'WeekNumber', **SUM** of WeekNumber 'WeeklyUserCount' and 'CumulativeUsers'. The view was grouped by WeekNumber using **GROUP BY.**
b) WeeklyUserCount is derived from **COUNT** of user_id and WeekNumber was extracted from activated_at by converting it into week number starting from first account activation date using **TIMESTAMPDIFF** and **WEEK** functions.
c) CumulativeUsers is the running total (cumulative sum) of WeeklyUserCount, ordered by WeekNumber using **ORDER BY** inside a window function.
d) **MONTHLY_NEW_USERS** was created from **USERS** using the same query but for months instead of weeks.







3. Weekly Retention Analysis:

a) This SQL query calculates weekly user retention based on their sign-up cohort and stores it in a view **Q5**. It breaks users into weekly cohorts (groups based on sign up weeks) and measures how many of them remain engaged in later weeks.
b) From **USERS** table, using **WITH** clause, temporary table **COHORTS** was created with columns 'total_users' which is derived from **COUNT** of rows and 'cohort_start_week' which was extracted from activated_at by converting it into week number starting from first account activation date using **TIMESTAMPDIFF** and **WEEK** functions.
c) The clause **GROUP BY 1,** groups the data by the computed **cohort_start_week**, ensuring all users who activated in the same week belong to the same cohort.
d) A second temporary table called **WEEKLY_STATS** was made from table **USERS** and **EVENTS**. Column 'active_users' was made from **COUNT** of **DISTINCT** user_id values from table **EVENTS.**
e) Using **EVENTS** table 'engagement_week' was extracted from occurred_at by converting it into week number starting from first event date using **TIMESTAMPDIFF** and **WEEK** functions.
f) Table **EVENTS** or '**E**' was joined to **USERS** '**U**' by their user_id columns using **JOIN**. A condition was placed using **WHERE** function to only show results with event_type value is 'engagement'.
g) **WEEKLY_STATS** was then grouped by cohort_start_week and engagement_week.

h) Next, columns cohort_start_week, total_users (from table **COHORTS)** engagement_week, and active_users (from table **WEEKLY_STATS**) were selected alongwith 'retention_rate' which was derived from dividing active_users by total_users and multiplying it by 100.

i) Finally, table **WEEKLY_STATS** was joined to **COHORTS** by the column cohort_start_week. This table was arranged by cohort_start_week and engagement_week using **ORDER BY** function.

```sql
# calculate the weekly retention of users based on their sign-up cohort.
CREATE VIEW Q5 AS (
    WITH cohorts AS (SELECT
        TIMESTAMPDIFF(WEEK , '2013-01-01 04:40:10', activated_at) AS cohort_start_week,
            COUNT(*) AS total_users
        FROM users
        GROUP BY 1
    ),
    weekly_stats AS (SELECT
            TIMESTAMPDIFF(WEEK , '2013-01-01 04:40:10', activated_at) AS cohort_start_week,
            TIMESTAMPDIFF(WEEK , '2013-01-01 04:40:10', occurred_at) AS engagement_week,
            COUNT(DISTINCT e.user_id) AS active_users
        FROM users u
        JOIN events e ON u.user_id = e.user_id
        WHERE e.event_type = 'engagement'
        GROUP BY 1, 2
    )
    SELECT
        cohorts.cohort_start_week,
        weekly_stats.engagement_week,
        weekly_stats.active_users,
        cohorts.total_users AS total_users,
        ROUND(weekly_stats.active_users / cohorts.total_users * 100, 2) AS retention_rate
    FROM cohorts
    JOIN weekly_stats ON cohorts.cohort_start_week = weekly_stats.cohort_start_week
    ORDER BY cohort_start_week, engagement_week);
```

Engagement based Week number starting from 2013

Signup Cohort Week of Year 2013	Week no.	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86
	0	10	13.33	10	10	13.33	13.33	13.33	26.67	13.33	10	10	6.67	3.33	13.33	10	13.33	10	6.67
	1		5.88	11.76	8.82	11.76	14.71	8.82	14.71	17.65	5.88		8.82	11.76	14.71	11.76	5.88	14.71	8.82
	2	6.52	13.04	13.04	15.22	15.22	19.57	15.22	15.22	4.35	2.17	8.7	10.87	4.35	4.35	2.17		8.7	8.7
	3	11.43	17.14	17.14	20	22.86	14.29	11.43	14.29	17.14	14.29	14.29	17.14	20	11.43	5.71	2.86	2.86	8.57
	4	5.71	22.86	14.29	17.14	17.14	11.43	14.29	20	20	17.14	20	20	25.71	22.86	22.86	20	8.57	
	5	12.5	10	17.5	10	10	17.5	15	15	17.5	15	15	12.5	17.5	12.5	7.5	7.5	12.5	10
	6	6.82	18.18	13.64	13.64	15.91	18.18	11.36	11.36	13.64	15.91	13.64	20.45	13.64	15.91	6.82	2.27		6.82
	7	7.69	17.95	10.26	15.38	10.26	10.26	15.38	20.51	20.51	15.38	17.95	20.51	17.95	12.82	12.82	10.26	5.13	7.69
	8	8.57	5.71	14.29	20	11.43	11.43	11.43	14.29	11.43	14.29	2.86	17.14	14.29	14.29	8.57	5.71	5.71	5.71
	9	11.9	19.05	16.67	23.81	11.9	16.67	23.81	19.05	19.05	26.19	14.29	19.05	19.05	16.67	9.52	7.14	4.76	7.14
	10	16.13	6.45	9.68	9.68	12.9	22.58	12.9	6.45	9.68	16.13	12.9	16.13	6.45	12.9	6.45	6.45	6.45	
	11	11.76	17.65	17.65	14.71	20.59	11.76	14.71	8.82	11.76	17.65	11.76	8.82	14.71	23.53	20.59	17.65	14.71	14.71
	12	12.5	6.25	6.25	6.25	12.5	18.75	9.38	9.38	15.63	15.63	6.25	21.88	12.5	15.63	9.38	6.25	6.25	6.25
	13	7.14	9.52	9.52	11.9	9.52	7.14	7.14	7.14	7.14	7.14	19.05	14.29	14.29	23.81	21.43	9.52	9.52	4.76
	14	6.25	9.38	12.5	12.5	6.25	15.63	18.75	21.88	12.5	6.25	9.38	12.5	15.63	15.63	9.38	15.63	9.38	6.25
	15	13.64	15.91	13.64	15.91	13.64	15.91	11.36	18.18	11.36	13.64	15.91	13.64	22.73	22.73	9.09	20.45	13.64	6.82
	16	14	18	22	20	32	20	16	22	20	26	22	24	22	16	14	18	12	14
	17	6.98	11.63	18.6	18.6	9.3	6.98	16.28	13.95	9.3	16.28	11.63	13.95	20.93	18.6	9.3	11.63	18.6	13.95
	18	13.21	15.09	16.98	20.75	16.98	18.87	22.64	15.09	15.09	13.21	24.53	15.09	16.98	13.21	15.09	13.21	7.55	3.77
	19	8	16	22	16	12	2	14	12	8	14	16	16	10	10	12	12	6	8
	20	7.5	5	5	2.5	5	7.5	2.5	7.5	7.5	12.5	7.5	10	7.5	2.5	2.5	5	2.5	2.5
	21	5.45	10.91	7.27	9.09	7.27	5.45	7.27	3.64	5.45	3.64	9.09	5.45	1.82	5.45	5.45	3.64	3.64	7.27
	22	12	10	14	10	4	6	10	10	6	16	14	20	18	22	12	8	10	10
	23	14	8	6	10	12	12	10	10	10	10	12	8	10	4	4	8	6	4
	24	8	4	8	6	8	4	6	8	12	14	16	16	12	6	6	2	10	8
	25	10	14	18	10	10	6	8	6	12	10	16	10	10	12	14	6	8	12
	26	5.56	7.41	18.52	14.81	14.81	9.26	9.26	14.81	9.26	7.41	11.11	11.11	9.26	7.41	7.41	5.56	7.41	7.41
	27	6.78	13.56	13.56	10.17	8.47	5.08	8.47	6.78	8.47	6.78	10.17	10.17	13.56	16.95	8.47	8.47	3.39	5.08
	28	12	14.67	14.67	10.67	13.33	14.67	10.67	10.67	14.67	12	13.33	12	9.33	5.33	9.33	8	6.67	6.67

Engagement based Week number starting from 2013

Signup Cohort Week of Year 2013

Week no.	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86
29	15.15	15.15	18.18	15.15	13.64	21.21	21.21	18.18	13.64	15.15	15.15	9.09	13.64	10.61	6.06	7.58	4.55	7.58
30	15.87	15.87	12.7	11.11	7.94	14.29	12.7	12.7	14.29	15.87	12.7	11.11	12.7	12.7	9.52	12.7	7.94	4.76
31	12.33	12.33	13.7	15.07	8.22	8.22	16.44	17.81	12.33	8.22	2.74	5.48	5.48	5.48	8.22	8.22	12.33	8.22
32	15.15	22.73	15.15	13.64	10.61	4.55	10.61	13.64	13.64	13.64	12.12	10.61	10.61	7.58	6.06	7.58	4.55	7.58
33	7.89	13.16	14.47	14.47	13.16	13.16	10.53	15.79	14.47	22.37	17.11	13.16	18.42	21.05	15.79	11.84	10.53	10.53
34	9.21	9.21	10.53	6.58	9.21	7.89	11.84	14.47	11.84	10.53	19.74	10.53	13.16	11.84	9.21	6.58	3.95	2.63
35	8.82	13.24	14.71	16.18	14.71	11.76	13.24	10.29	11.76	17.65	14.71	13.24	10.29	13.24	13.24	11.76	14.71	10.29
36	6.94	11.11	13.89	12.5	11.11	13.89	12.5	13.89	12.5	9.72	12.5	13.89	13.89	13.89	12.5	11.11	6.94	5.56
37	8.14	8.14	8.14	8.14	9.3	12.79	16.28	15.12	12.79	6.98	8.14	5.81	6.98	4.65	3.49	2.33	6.98	3.49
38	11.24	14.61	11.24	11.24	11.24	10.11	16.85	8.99	12.36	7.87	13.48	13.48	8.99	11.24	10.11	7.87	4.49	5.62
39	6.17	8.64	11.11	14.81	9.88	13.58	11.11	8.64	7.41	11.11	9.88	11.11	11.11	8.64	4.94	6.17	6.17	6.17
40	8.33	13.1	9.52	5.95	9.52	7.14	13.1	15.48	16.67	11.9	5.95	11.9	13.1	19.05	13.1	11.9	8.33	7.14
41	8.86	11.39	15.19	11.39	10.13	8.86	8.86	8.86	5.06	7.59	8.86	11.39	8.86	7.59	6.33	2.53	2.53	3.8
42	8.33	10.42	9.38	12.5	10.42	11.46	14.58	13.54	11.46	8.33	6.25	5.21	11.46	11.46	13.54	11.46	11.46	7.29
43	10.64	12.77	14.89	13.83	15.96	15.96	18.09	10.64	10.64	9.57	9.57	8.51	8.51	10.64	6.38	2.13	5.32	6.38
44	11.7	11.7	9.57	6.38	7.45	10.64	14.89	13.83	9.57	13.83	13.83	11.7	9.57	11.7	10.64	11.7	7.45	3.19
45	7.37	13.68	21.05	14.74	9.47	15.79	14.74	16.84	11.58	17.89	15.79	12.63	9.47	11.58	8.42	7.37	15.79	11.58
46	5.88	8.24	7.06	3.53	8.24	5.88	10.59	8.24	10.59	8.24	11.76	10.59	7.06	10.59	10.59	5.88	11.76	5.88
47	5.83	6.8	6.8	10.68	11.65	11.65	14.56	9.71	10.68	15.53	11.65	9.71	13.59	10.68	6.8	5.83	5.83	3.88
48	3.88	7.77	8.74	9.71	10.68	9.71	14.56	14.56	9.71	11.65	9.71	11.65	15.53	11.65	7.77	5.83	5.83	1.94
49	7.08	7.96	9.73	13.27	9.73	11.5	10.62	10.62	11.5	7.96	9.73	8.85	11.5	11.5	6.19	7.96	2.65	6.19
50	5.65	13.71	14.52	14.52	13.71	15.32	12.9	16.13	11.29	8.87	7.26	8.87	5.65	5.65	9.68	5.65	0.81	1.61
51	4.95	13.86	7.92	9.9	7.92	7.92	7.92	7.92	9.9	11.88	5.94	6.93	5.94	6.93	3.96	6.93	4.95	5.94
52	5.93	11.11	17.04	14.07	12.59	11.85	9.63	14.07	11.85	13.33	9.63	10.37	9.63	8.89	8.89	10.37	5.19	4.44
53	5.04	9.24	7.56	10.08	10.92	6.72	8.4	10.08	7.56	9.24	11.76	8.4	6.72	5.88	9.24	5.88	4.2	1.68
54	6.61	13.22	15.7	13.22	12.4	17.36	16.53	12.4	12.4	15.7	12.4	9.09	12.4	9.92	12.4	9.92	11.57	9.92
55	7.77	7.77	17.48	17.48	18.45	13.59	11.65	8.74	6.8	10.68	12.62	9.71	8.74	10.68	8.74	7.77	10.68	7.77
56	11.35	14.89	11.35	9.93	7.8	14.18	12.77	13.48	12.77	14.89	12.06	9.22	9.22	14.18	11.35	10.64	9.22	6.38

Engagement based Week number starting from 2013

Signup Cohort Week of Year 2013

Week no.	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86
57	14.06	16.41	16.41	10.94	9.38	10.16	12.5	13.28	13.28	12.5	10.16	9.38	10.16	11.72	10.94	7.81	12.5	8.59
58	11.28	9.02	9.02	9.77	12.78	14.29	14.29	12.03	9.02	9.77	11.28	9.77	8.27	8.27	9.02	9.02	7.52	5.26
59	8.4	12.21	7.63	11.45	9.92	11.45	10.69	9.92	13.74	9.92	14.5	12.98	15.27	11.45	13.74	8.4	6.87	9.16
60	9.6	12	10.4	9.6	9.6	12.8	12	9.6	12.8	9.6	7.2	9.6	12.8	9.6	10.4	8.8	11.2	10.4
61	11.89	16.08	11.89	9.79	11.89	9.79	10.49	8.39	6.99	13.29	16.78	13.29	13.29	9.09	7.69	5.59	6.99	5.59
62	8.39	11.89	14.69	10.49	9.79	8.39	8.39	9.79	8.39	6.29	7.69	7.69	9.09	9.79	9.09	7.69	4.9	4.9
63	12.5	16.18	17.65	17.65	17.65	14.71	13.97	13.97	15.44	13.97	13.24	14.71	14.71	16.18	15.44	12.5	10.29	5.88
64	18.35	17.72	14.56	17.72	14.56	15.82	10.76	11.39	11.39	10.76	10.76	12.66	13.29	12.66	12.66	9.49	10.13	8.86
65	24.53	16.35	15.72	12.58	11.32	11.95	14.47	12.58	11.32	10.06	12.58	14.47	11.95	12.58	11.32	7.55	8.18	7.55
66	17.06	20.59	17.06	18.82	15.88	14.71	11.18	11.18	10	11.18	11.76	11.76	12.94	14.71	10.59	8.82	10	7.65
67	39.39	36.97	30.3	21.21	15.76	15.15	12.12	10.3	12.73	10.91	9.7	10.91	12.73	10.91	9.7	7.88	6.06	4.24
68	44.2	38.67	28.18	19.34	14.36	9.39	7.18	8.29	10.5	11.6	9.94	8.29	7.73	8.29	9.39	6.08	6.08	3.87
69	94.84	65.81	35.48	27.74	17.42	18.06	14.84	12.26	10.32	8.39	9.68	10.32	9.68	9.68	7.74	9.68	5.81	5.16
70		100	65.88	43.53	30.59	22.94	12.35	10	13.53	10	12.94	10	7.65	7.65	5.88	6.47	5.29	1.76
71			100	68.82	41.4	33.87	22.04	13.44	10.22	10.22	10.75	9.68	8.6	8.06	6.99	7.53	6.45	3.23
72				100	67.07	44.91	28.74	23.95	14.97	11.98	17.37	13.77	13.77	13.17	11.98	7.78	3.59	5.39
73					100	60	42.56	30.26	16.92	15.38	17.44	16.41	11.28	10.26	7.69	10.26	6.67	4.62
74						100	66.5	44	34.5	22	19	13.5	12.5	9.5	8	7	3.5	2
75							100	70.24	41.95	29.76	24.88	22.44	18.05	15.61	10.73	11.22	7.32	3.9
76								100	61.26	42.79	27.03	18.47	15.77	13.51	9.46	5.86	6.31	4.95
77									100	65.69	43.14	28.92	21.08	15.2	14.22	10.78	8.33	7.84
78										100	67.91	45.12	32.56	24.65	21.4	15.35	14.88	6.98
79											100	69.08	44.44	35.75	21.74	16.43	9.66	8.21
80												100	64.22	40.83	23.39	17.43	9.63	9.63
81													100	65.52	36.64	27.59	18.53	13.79
82														100	59.91	35.02	27.65	19.82
83															100	66.83	33.17	22.6
84																100	65.75	36.61
85																	100	65.9
86																		100

4. Weekly Engagement Per Device:

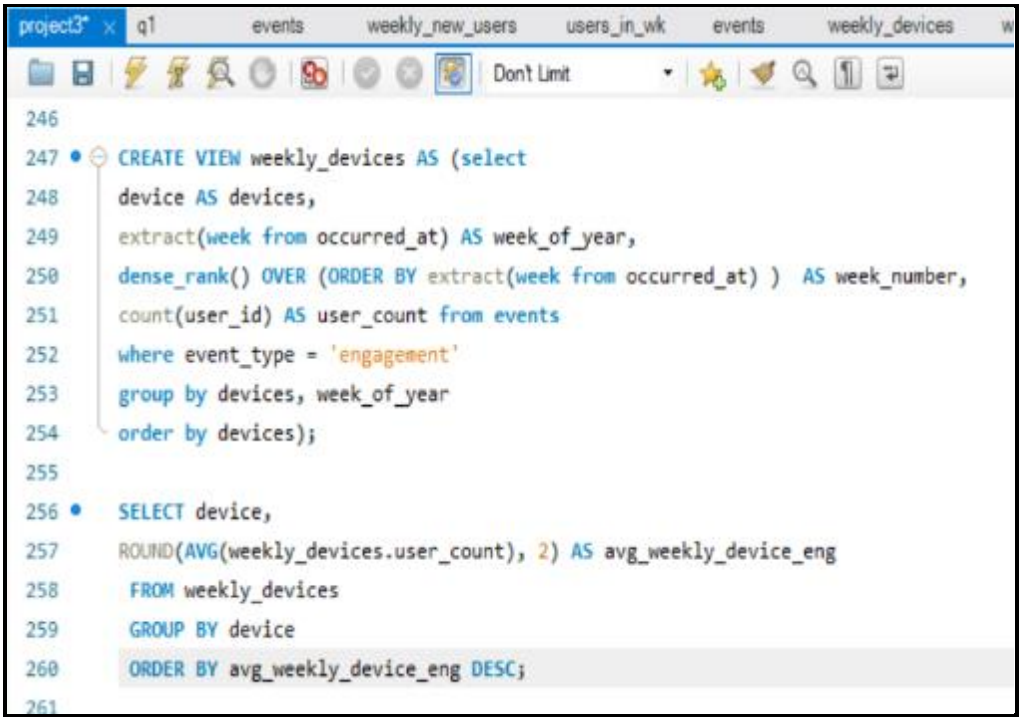
a) A view **WEEKLY_DEVICES** was created including 'devices', 'week_of_year', 'user_count' (**DISTINCT** user_id values) and 'week_number' columns derived from **EVENTS** table.

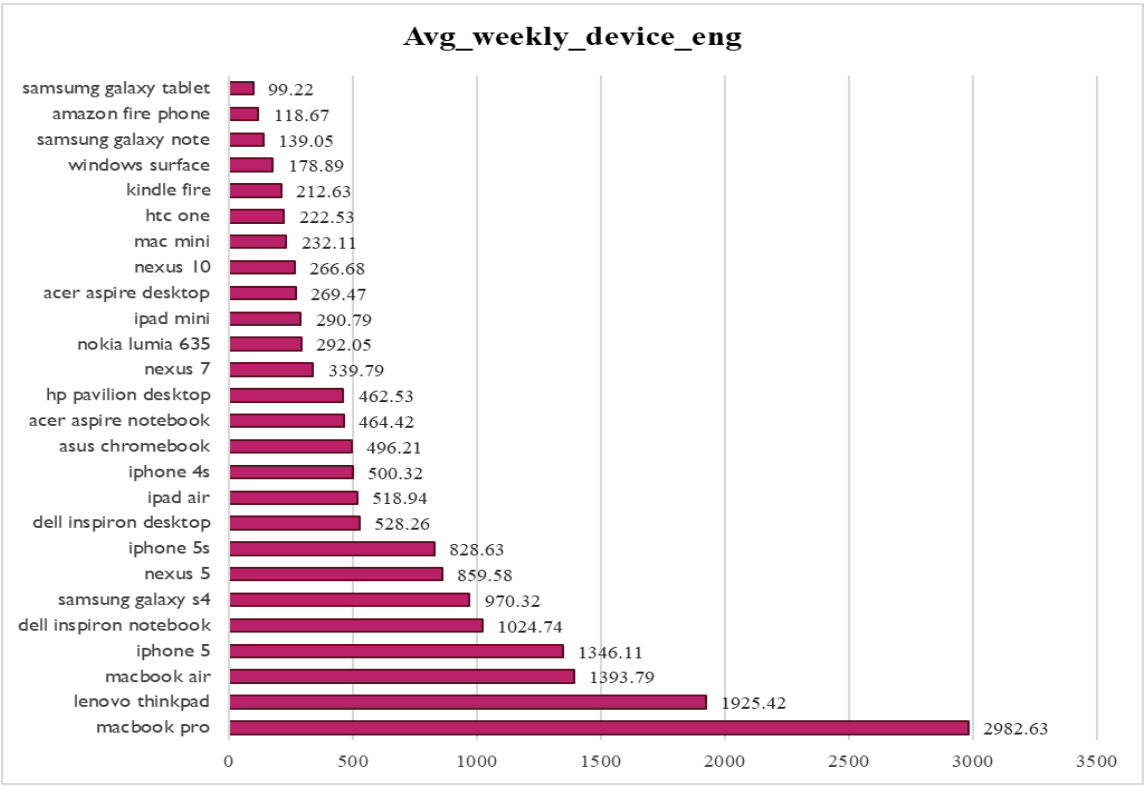
b) Using **EXTRACT** and **WEEK FROM** functions week_of_year was created from occurred_at. The column week_number was created by arranging week_of_year column using **DENSE_RANK ()**, **ORDER BY** and **DESC** functions. The view was grouped by 'week_of_year'.

c) The view was filtered for rows **WHERE** event_type was 'engagement'. These were arranged by devices (**ORDER BY** function) and grouped by 'devices' and 'week_of_year' using **GROUP BY** function.

d) From **WEEKLY_DEVICES,** columns device and 'avg_weekly_device_eng' were selected. This column, avg_weekly_device_eng, was created from calculating the average of user_count from the view using the function **AVG**.

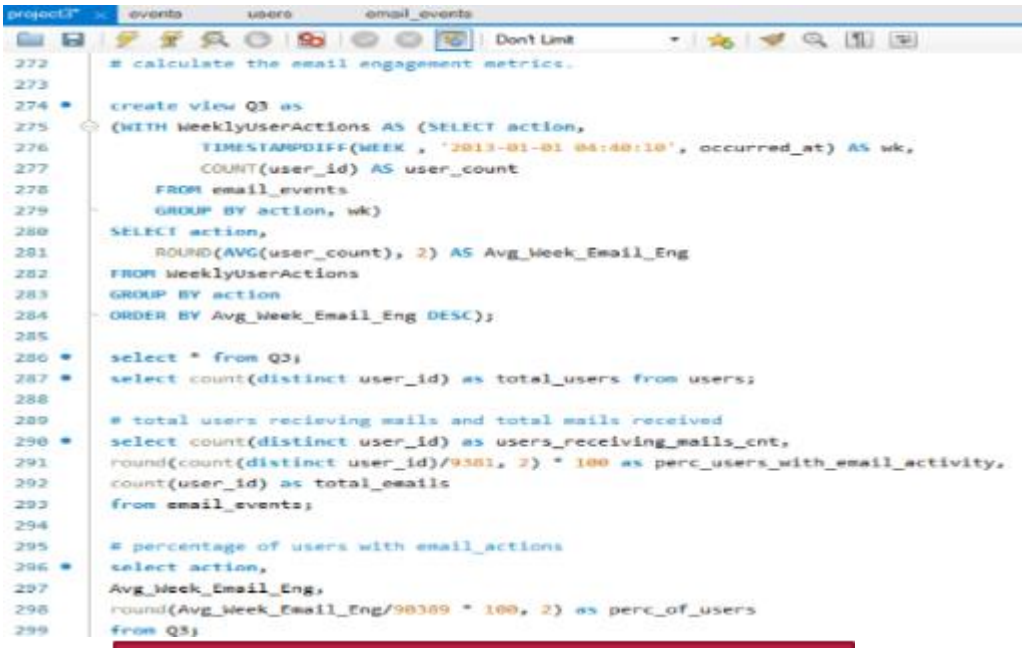
e) They were grouped by device and arranged in descending order by avg_weekly_device_eng using the functions **GROUP BY** and **ORDER BY**.



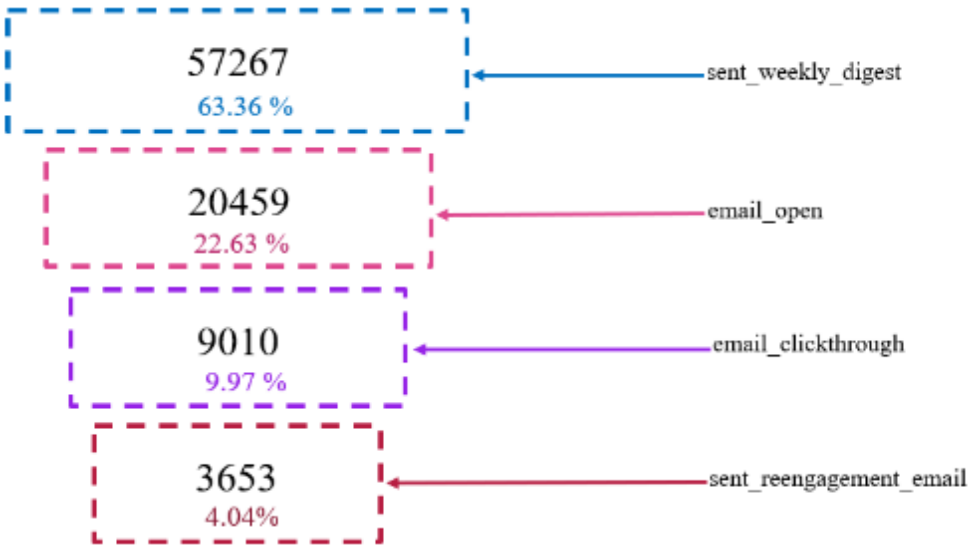


5. Email Engagement Analysis:

a. This SQL query calculates **email engagement metrics** and stores it in a view **Q3**. It also calculates the **average weekly email actions** and the **percentage of users with email activity.**
b. From **EMAIL_EVENTS** table, using **WITH** clause, temporary table **WeeklyUserActions** was created having action, 'user_count' (**COUNT** of user_id) and 'wk' which was extracted from occurred_at by converting it into week number starting from first email activity date using **TIMESTAMPDIFF** and **WEEK** functions. This temporary table was grouped by action and wk using **GROUP BY**.
c. From **WeeklyUserActions**, columns action and 'Avg_Week_Email_Eng' are selected and grouped by action using **GROUP BY**. They are arranged in a descending order by Avg_Week_Email_Eng using **ORDER BY** and **DESC** functions.
d. Avg_Week_Email_Eng is derived from taking an average of user_count rounded to 2 decimal places
e. To calculate total no. of users (total_users), **DISTINCT** user_id values were calculated from **USERS** table.
f. The next query calculates total users receiving mails and total mails received from the columns total_emails, 'users_receiving_mails_cnt' and 'perc_users_with_email_activity' derived from the table **EMAIL_EVENTS**.
g. The **COUNT** of **DISTINCT** user_id values was named users_receiving_mails_cnt and it was further used to derive perc_users_with_email_activity by dividing the users_receiving_mails_cnt by '9381'which was the Avg_Week_Email_Eng value.
h. Finally, to calculate percentage of users with email actions using the view **Q3,** columns 'perc_of_users' and Avg_Week_Email_Eng are selected. This new column is derived from dividing Avg_Week_Email_Eng with '90389' (users_receiving_mails_cnt). and then multiplying it by 100.



```
# calculate the email engagement metrics.

create view Q3 as
(WITH WeeklyUserActions AS (SELECT action,
        TIMESTAMPDIFF(WEEK , '2013-01-01 04:40:10', occurred_at) AS wk,
        COUNT(user_id) AS user_count
    FROM email_events
    GROUP BY action, wk)
SELECT action,
    ROUND(AVG(user_count), 2) AS Avg_Week_Email_Eng
FROM WeeklyUserActions
GROUP BY action
ORDER BY Avg_Week_Email_Eng DESC);

select * from Q3;
select count(distinct user_id) as total_users from users;

# total users recieving mails and total mails received
select count(distinct user_id) as users_receiving_mails_cnt,
round(count(distinct user_id)/9381, 2) * 100 as perc_users_with_email_activity,
count(user_id) as total_emails
from email_events;

# percentage of users with email_actions
select action,
Avg_Week_Email_Eng,
round(Avg_Week_Email_Eng/90389 * 100, 2) as perc_of_users
from Q3;
```



ANALYSIS:

1) Due to limited data, jobs reviewed per hour per day are only available for the last five days. The highest hourly review rate occurred on the last three days, especially Saturday, 28th Nov 2020, while the lowest was on 27th Nov. Most job reviews in November 2020 took between 0.01 and 0.03 hours daily.

2) The daily line plot shows no clear trend, with job reviews per second ranging from 0.01 to 0.06. The 7-day rolling average plot shows either a steady trend or an increase in the final two days.

3) The 7-day rolling average is more reliable than daily metrics, which fluctuate unpredictably. Relying on daily values may be misleading due to their inconsistency.

4) The most used language is Persian, with the remaining five languages appearing only once. There are no duplicate rows—only repeated values like language and dates, which don't affect data validity.

5) The data starts from week 17 of 2014 (1st May). Week 17 had only three days and week 35 was a Sunday, explaining fewer active users. The highest number of active users occurred in July.

6) Active users increased consistently, peaking in week 30 (27th July–2nd Aug), then decreased slightly with no strong trend. On average, users have 14.77 engagements per week. Out of 9381 sign-ups, about 1158.68 engage weekly.

7) From week 0 (1st Jan 2013) to week 84 (31st Aug 2014), new users increased monthly from 160 (Jan 2013) to 1028 (Aug 2014), with a dip in Jan 2014. There’s clear growth in user base, beneficial for the company.

8) Signup week cohorts span from 1st Jan 2013 (week 0) to 31st Aug 2014 (week 86), while engagement weeks span from 1st May 2014 (week 69) to 31st Aug 2014. Retention is below 10% in week 69, likely due to late engagement post-signup.

9) Retention drops significantly in weeks 84–86, with week 86 being just one day (31st Aug). Older users show declining interest, requiring intervention through email or ad campaigns. New user retention is highest (60–100%).

10) New users engage more initially, but 30–40% drop off the following week between weeks 69–86. Middle-week users show consistent 15–25% retention. Overall low retention needs to be addressed.

11) MacBook Pro is the most used device, followed by Lenovo ThinkPad, MacBook Air, iPhone 5, and Dell Inspiron Notebook. Most top-used devices are laptops, common in corporate settings.

12) Out of 9381 users, 6179 interacted with or were sent emails. 63.36% of these were weekly digests; 22.63% opened the emails. Expanding email outreach is essential.

CONCLUSION:

This project has helped me dive deeper into the world of operation analytics and metric studies. It was interesting to see how we can answer so many questions regarding customer behavior towards the product and get insights on underlying patterns observed. The important insights gathered here were the need to address weekly user retention rate and why it's low. It would be beneficial to address this issue via ad-campaigns targeting inactive users. Also, targeting emails to users that haven't received an email before might be beneficial. However, a great insight observed is the consistent increase in users.

HIRING PROCESS ANALYTICS

DESCRIPTION:

As a data analyst working for a global corporation like Google your role involves examining data related to the company's hiring procedures to uncover valuable insights. Since recruitment plays a vital role in any organization's success, understanding patterns such as rejection rates, interview stages, job categories, and open positions can offer important guidance for the hiring team.

In this role, you'll receive a dataset containing information on past hires. Your objective is to explore this data and address key questions that can assist the company in refining its recruitment strategies.

THE PROBLEM:

The purpose of this project is to apply your expertise in statistics and Excel to extract valuable insights from the company's hiring data. Your findings could play a significant role in enhancing the hiring process and supporting better decision-making for future recruitment efforts.

A. Data pre-processing

Handling Missing Data:

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

Clubbing Columns:

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

Outlier Detection:

Check for outliers in the dataset that may skew your analysis.

Removing Outliers:

Decide on the best strategy to handle outliers. (removing them, replacing them, or leaving them as is).

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.

B. Data Analytics Tasks:

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

A) **Dataset Summary: "Statistics"**

- **Number of Observations**: 7168
- **Number of Variables**: 7
- **Tech-Stack Used**: Microsoft Excel 365

Names of Columns	Significance
application_id	Id no. of the application
Interview Taken on	Date and time of interview of candidates
Status	Hiring status of candidate
event_name	Gender of the candidate/employee
Department	Name of the department
Post Name	The post candidate applied for
Offered Salary	The offered salary to the candidate

B) Data Cleaning, rectification and outlier detection:

1. There are 54 duplicate application id values with different details in other columns. These values need to be replaced or corrected. The missing data was handled by replacing "-" in the "event_name" (gender) column with "Don't want to say" as they mean the same thing. A list unique post names revealed a post name "c-10" which was replaced with "c10".

2. There is also a missing value in the "Post Name" column. The department name is "Sales Department" and the offered salary is "85914". The counts of offered salary between 85000-90000 for Sales department showed "c5" having the highest count. The null value was replaced with the same.

3. There is a missing value in the "Offered salary" column. The department name is "Sales Department" and the post name is "i7". The median of offered salaries for both of those criteria came out to be 44449. The value was replaced with the same.

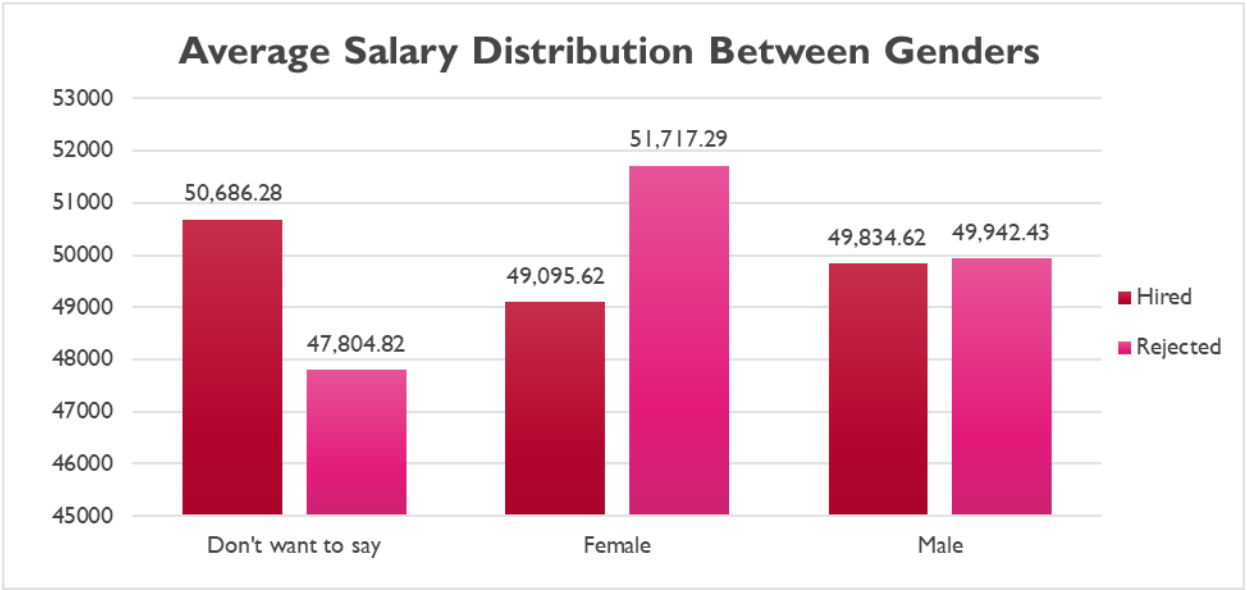
4. Three outliers were detected in the "Offered Salary" column using box-whiskers plot. These outliers (200000, 300000 and 400000) were replaced with median salary of the corresponding department name and post name.

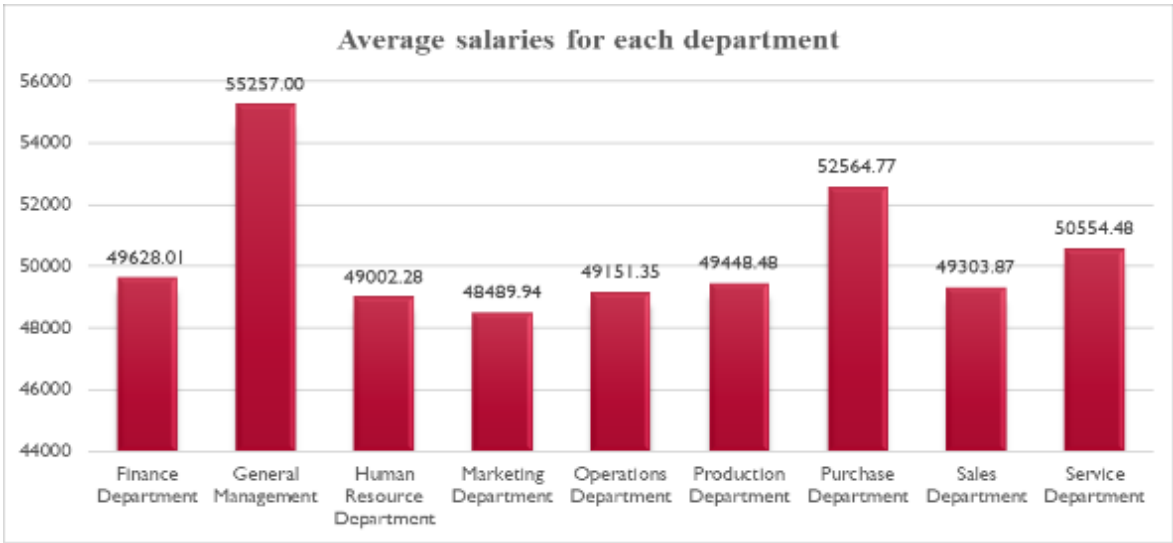
5. Since there was only one entry with department "general management" and post name "i4", Count of entries for both the department name and post name were taken separately. As there are more entries with post name "i4", median of offered salary with i4 post name was taken to replace 400000.

FINDINGS:

A) Data Summary:

a) A number of pivot tables were created to obtain average salaries for genders, maximum and minimum salaries for each post and average salaries, etc.
b) Using these pivot tables, various bar and column charts were plotted all centering salaries offered to different groups of people.

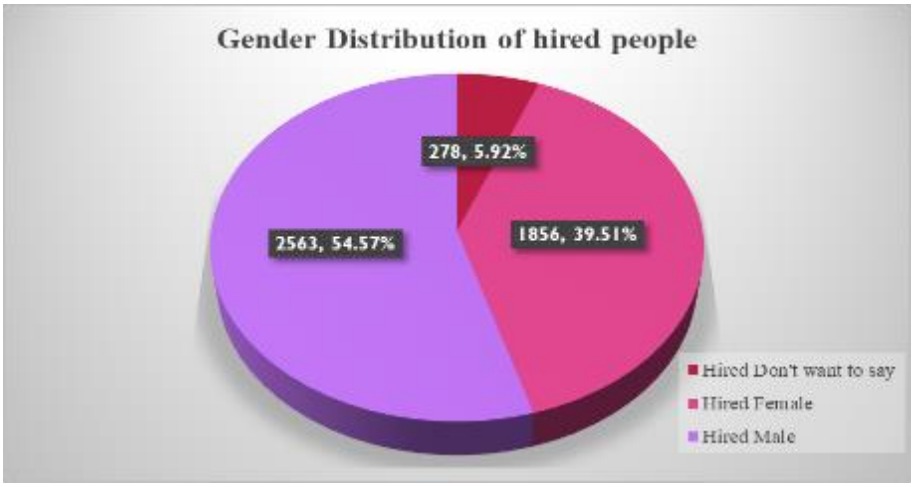




B) Data Analytics Tasks:

1. Hiring Analysis:

a) A pivot table was created to obtain average salaries for genders. A filter was applied to only show hired People. The table was represented as a pie chart showing ‘Gender Distribution of Hired People’.



2. Salary Analysis:

a. The function AVERAGE was used to calculate the average value for the column 'Offered Salary'. Also, the function AVERAGEIF was used to place a condition to only calculate average offered salary for candidates that had value of 'Status' as 'Hired'.

49593.02

Average Salary For Hired Employees

49877.49

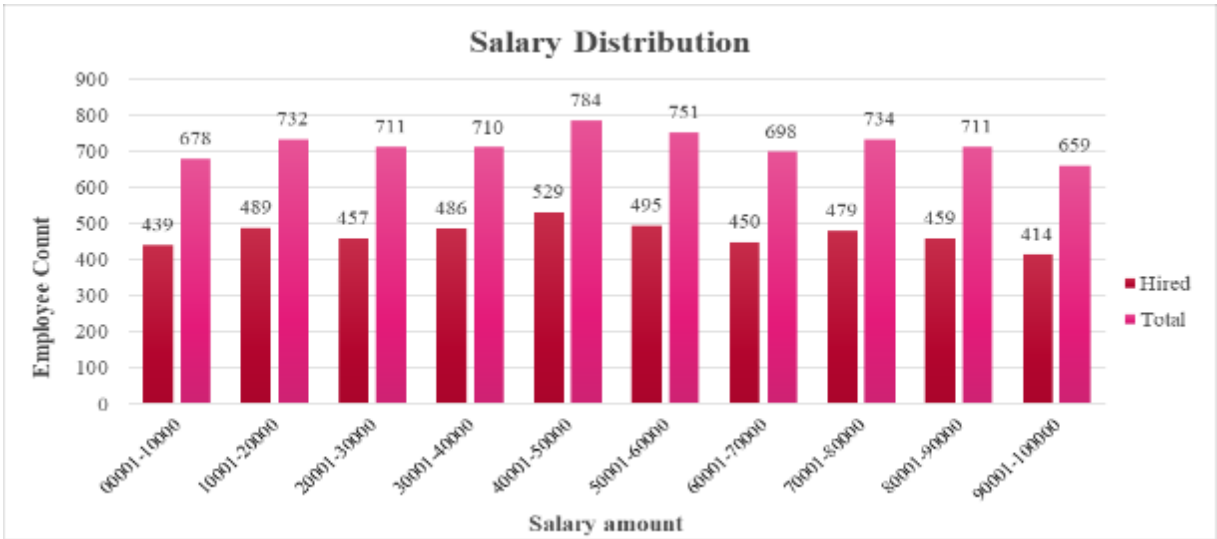
Average Salary For All Employees

```
=AVERAGEIF(C2:C7169,"Hired",G2:G7169)
```

N	O
49593.01597	avg salary for hired employees
49877.49177	avg salary for all employees

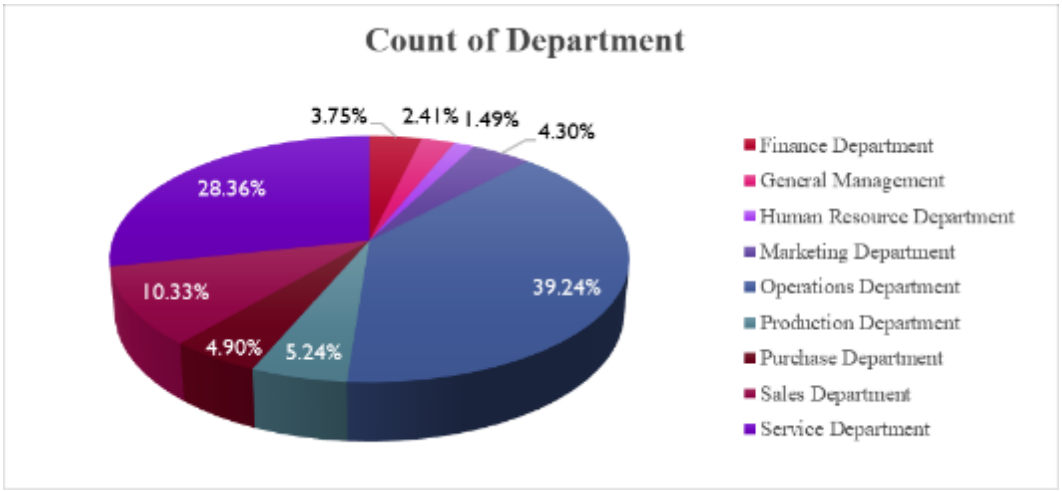
3. Salary Distribution:

a) A table was created to analyze the distribution of employees—both overall and newly hired—across different salary brackets. The salary brackets were defined in increments of 10,000 (e.g., 1-10,000, 10,001-20,000, 20,001-30,000, and so on up to 100,000).
b) The function **COUNTIFS** was used to place conditions to only calculate count of employees who are paid salaries within the salary bracket and that had value of 'Status' as 'Hired'. Values in 'Status' were not considered while considering overall salary distribution.
c) The table was represented as a clustered column chart showing 'Salary Distribution'.



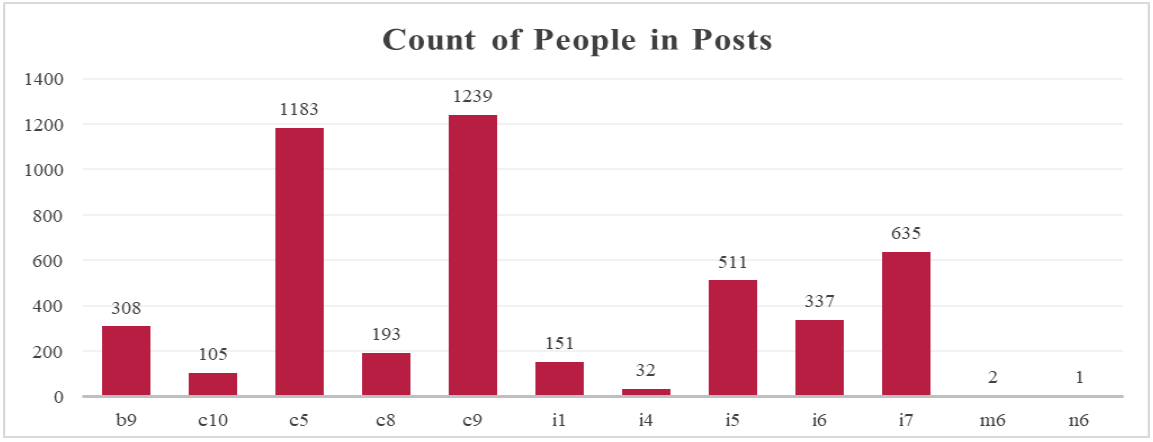
4. Departmental Analysis:

a) A pivot table was created to obtain count of employees in all departments. A filter was applied to only show hired people. The table was represented as a pie chart showing 'Count of Department'.



5. Position Tier Analysis:

a) A pivot table was created to obtain count of employees in all posts. A filter was applied to only show hired people. The table was represented as a column chart showing 'Count of People in Posts'.



ANALYSIS:

1. The graph “Average Salary Distribution between Genders (Hired vs. Rejected)” shows a slight decrease in offered salary for rejected people with the gender “Don’t want to say”. Whereas for females it is slightly high. The salary offered for other cases is almost similar.

2. The graph “Minimum vs. Maximum Salaries for each post” shows consistency in first 4 posts as these only have one employee each. There is a significant difference in minimum and maximum salaries of other posts. This data maybe important to plan incentives or salary increments for future.

3. The graph “Average salaries for each department” displays that the highest average salary belongs to the general management department followed by purchase and service departments. The average salaries for other departments are consistent.

4. Out of total 4697 hired employees, majority of hired employees are males, 54.57% which is slightly more than half the employees. The percentage of females in the company is 39.51%. A small minority of employees that don’t want to reveal their gender make up 5.92% of the company. A gender distribution that’s close to 50% female is good for company image.

5. The average salary offered by the company to hired employees vs. all the employees (hired and rejected) is almost the same, around 50,000. Other factors may contribute to candidates rejecting the offer or being rejected for a position.

6. Out of total 4697 hired employees, majority of hired employees, a little more than a third (39.24%) of total employees work in the operations department. This is justifiable as the production of goods or a service should be the center of focus for every company.

7. A little over a quarter of total employees (28.36%) work in the service department. The distribution of employees for general management (2.41%) and human resources (1.41%) is the least. This also makes sense as both of these departments focus on managing other departments.

8. The highest number of employees share the post ‘c9’ with 1239 employees followed by post ‘c5’ with 1183 employees. There is only one employee working in the post ‘n6’ and two in ‘m6’ post. The post ‘i4’ also has very less number (32) of employees.

CONCLUSION:

This project helped me understand the importance of analyzing statistics related to hiring committee. The patterns observed during calculating the number of employees in different groups of people as well as their salaries helped me in getting insights on company’s work culture (gender vs. salary distribution) as well as importance of each department. These patterns can help the company is strategizing for future by making changes to suit their company image as well as help them in hiring employees according to their needs.

IMDB MOVIE ANALYSIS

DESCRIPTION:

The dataset provided is related to IMDB movies, with the objective of investigating factors influencing a movie's success on IMDB, defined by high ratings. This analysis can benefit producers, directors, and investors seeking insights for future projects. The goal is to deliver data-driven insights that empower stakeholders to make informed choices, not just answer questions. Key steps to be taken:

1. **Data Cleaning:** Prepare the data by handling missing values, removing duplicates, adjusting data types, and performing feature engineering.
2. **Data Analysis:** Explore relationships between variables such as genre, director, budget, actors, and year of release in correlation with movie ratings.
3. **Five Whys Approach:** A root cause analysis technique to dig deeper into findings. For example, if high-budget movies tend to have higher ratings, explore reasons such as better production quality and viewer experience.
4. **Report & Storytelling:** Develop a report summarizing the findings, insights, and their significance. Use visualizations to enhance clarity and focus on providing actionable insights for decision-making.

THE PROBLEM:

Provide a detailed report for the below data record mentioning the answers of the questions that follows:

Objective	Task
1. **Movie Genre Analysis:** Analyze the distribution of movie genres and their impact on the IMDB score.	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.
2. **Movie Duration Analysis:** Analyze the distribution of movie durations and its impact on the IMDB score.	Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.
3. **Language Analysis:** Situation: Examine the distribution of movies based on their language.	Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.
4. **Director Analysis:** Influence of directors on movie ratings.	Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

5. **Budget Analysis:** Explore the relationship between movie budgets and their financial success.	Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

DESIGN:

A) Dataset Summary for "IMDB_Movies":

- **Number of Observations**: 5043
- **Number of Variables**: 28
- **File Type**: Excel file

Name of Columns:	
color	movie_title
director_name	facenumber_in_poster
director_facebook_likes	plot_keywords
duration	movie_imdb_link
num_critic_for_reviews	num_user_for_reviews
actor_3_facebook_likes	language
actor_2_facebook_likes	country
actor_1_facebook_likes	content_rating
gross	budget
genres	title_year
actor_1_name	num_voted_users
actor_2_name	imdb_score
actor_3_name	aspect_ratio
cast_total_facebook_likes	movie_facebook_likes

B) Data pre-processing, cleaning and error rectification:

1. Out of 5043 rows and 28 columns, 126 rows with duplicate data (filtered by movie name) were removed. Rows with 8 or more empty column entries were deleted. 4909 rows remained. Out of these rows, three rows containing video games were removed. We are now left with 4906 rows.
2. Columns like ‘color’, ‘aspect ratio’, ‘facenumber_in_poster’ and ‘plot keywords’ were deleted as those didn’t seem important.
3. In these rows blank spaces in columns, ‘director name’, ‘duration’, actor_1_name, actor_2_name, actor_3_name, ‘num_user_for_reviews’, ‘language’, ‘country’ and ‘content_rating’ were filled with correct values after referring to their websites.
4. Blank values in budget, gross and ‘director_facebook_likes’ were replaced with correct values if available or replaced with NA. Budget values of some movies were given in currencies other than

dollars. Those were converted into dollars. Outliers were recognized for duration column and replaced with median values.

5. Data was cleaned by removing random characters or replacing with correct letters in the director and movie columns. The entries in language column "None" were also replaced with right values.
6. Some entries in the country column were given as "Official Site". They were replaced with the correct country name. "West Germany" in country column was replaced with "Germany".
7. A new column was added "no. of genres" signifying the number of genres shown in genre column.
8. Replaced "Musical" genre with "Music" genre as it means the same.
9. Further 1025 rows were deleted but only for the 'Budget Analysis' part as there was no information available on these 1025 movies for the missing budget and/or gross value.

FINDINGS:

1. Movie Genre Analysis:

a) Using the **COUNT, COUNTIF** and **COUNTIFS** formulas in excel along with wildcard characters ('*' and '&') the Genre count was determined along with combinations of genres wherever required.
b) To obtain statistics of these genres, **AVERAGEIFS, MINIFS, MAXIFS** formulas were used for calculating averages, minimums and maximums with conditions.
c) In order to calculate standard deviations, variances, mode, maximums and minimums with single or multiple conditions, **STDEV.S, VAR.S** and **MODE.SNGL** were used in combination with **IF**, **ISNUMBER** and **SEARCH** formulas. These were then represented in form of bar charts

Fig 1: Genre Distribution

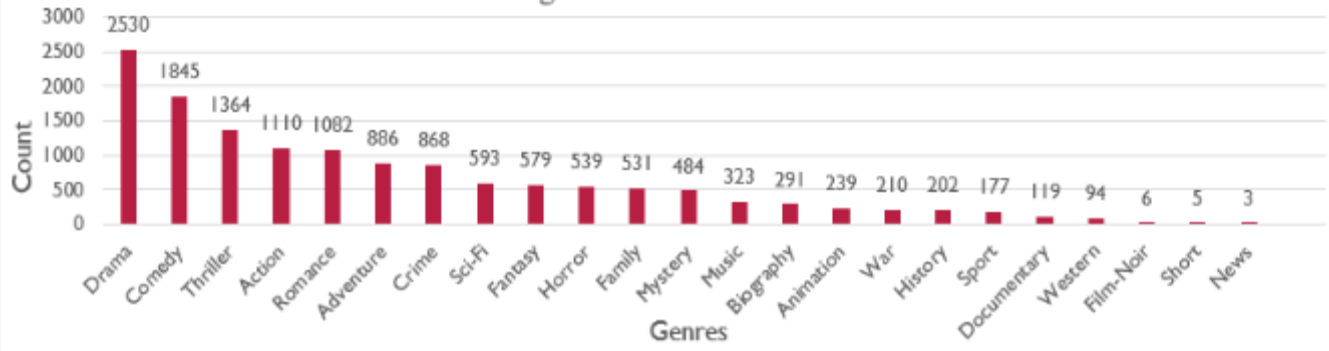


Fig 2: Stats of top genres

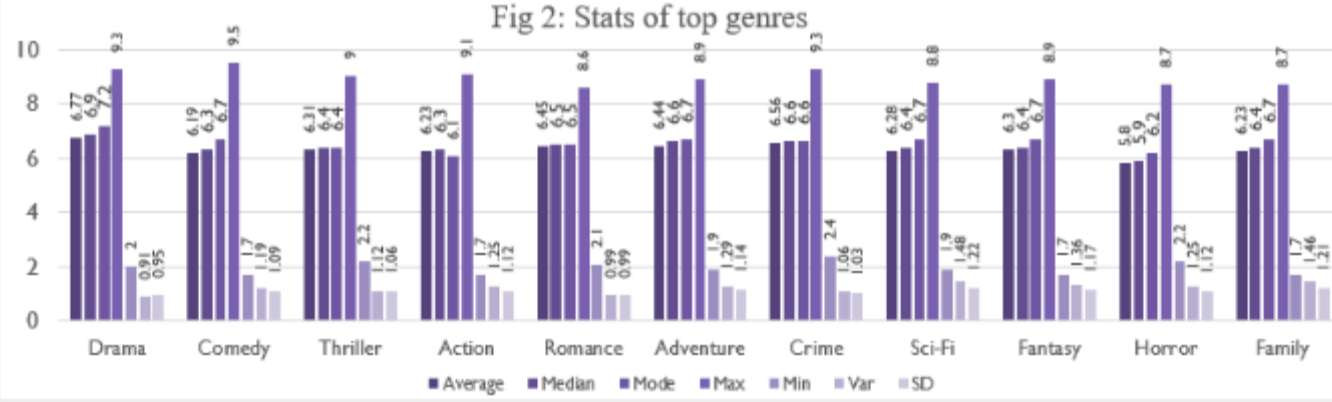


Fig 3: Genres common with Drama

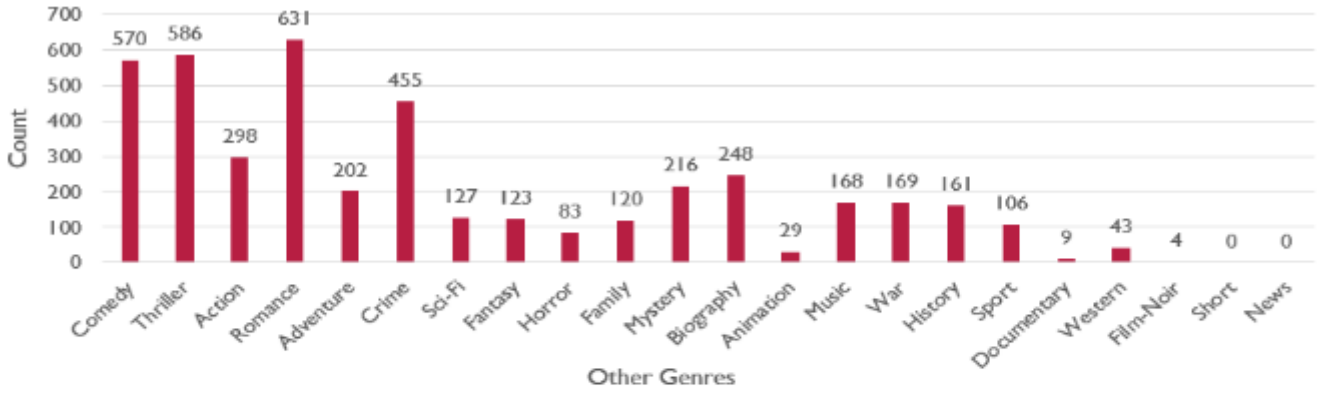
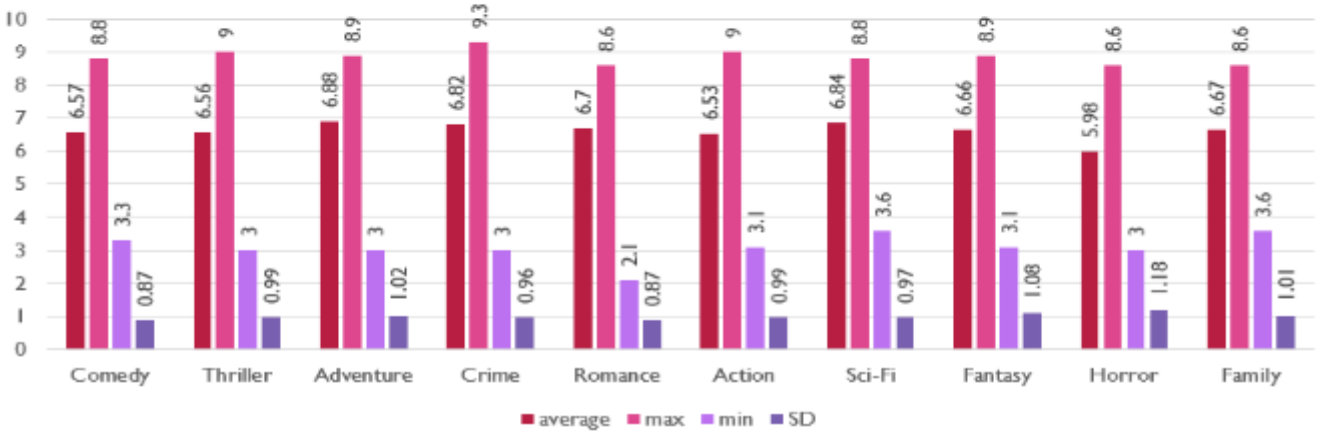


Fig 4: Stats of Genre common with Drama



2. Movie Duration Analysis:

d) A number of pivot tables were created to obtain count of movie durations, average of IMDB scores for each duration value, average budget for each duration value, etc.
e) Using these pivot tables, various combination charts were plotted all centering movie durations and related variables to it.
f) IMDB Score vs. duration chart was plotted directly by selecting the two columns and applying scatter plot chart to them.
g) For the duration count and normal distribution chart, the **NORM.DIST** formula was used to compute the probability density function (PDF) for each duration value. The mean and standard deviation (SD), essential for the calculation, were derived using the **AVERAGE** and **STDEV** functions.
h) Additionally, SD bands were determined alongside the PDF values. Finally, a table was created, compiling duration values, their counts, PDF values, and SD bands, which was then used to plot the combination chart.

Fig 5: IMDB_score vs. duration

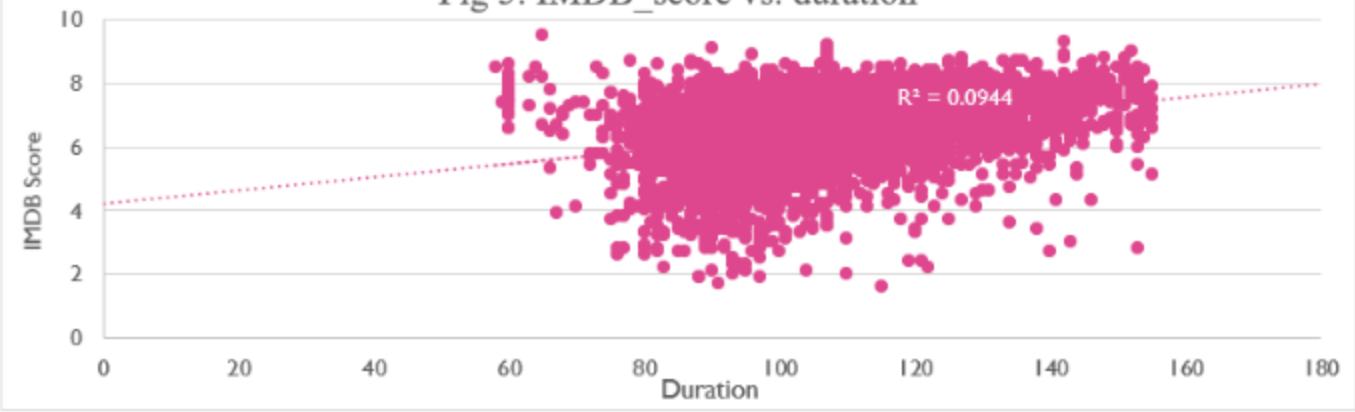


Fig 6: Duration count and normal distribution

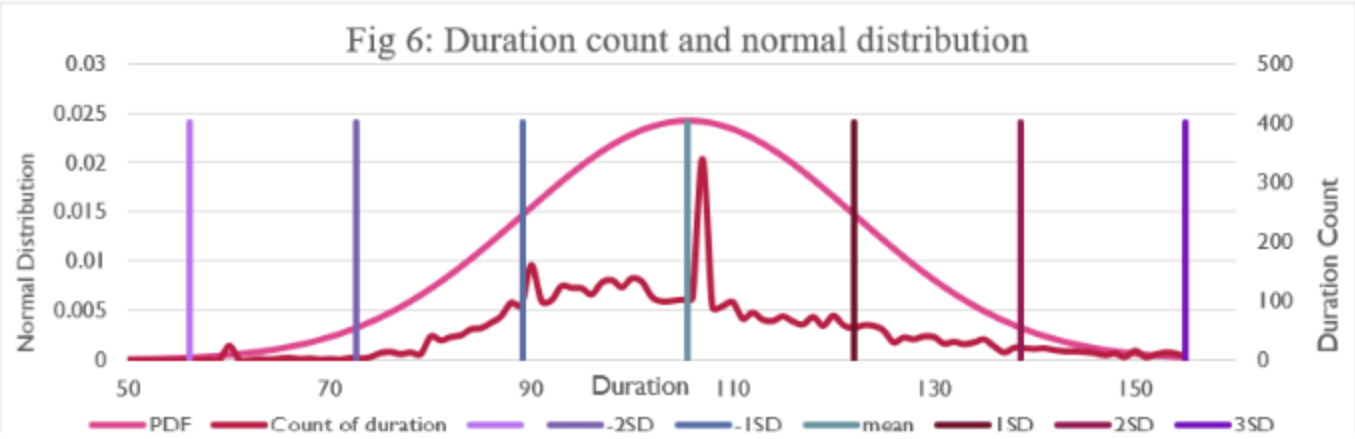


Fig 7: Duration vs. average budget and count

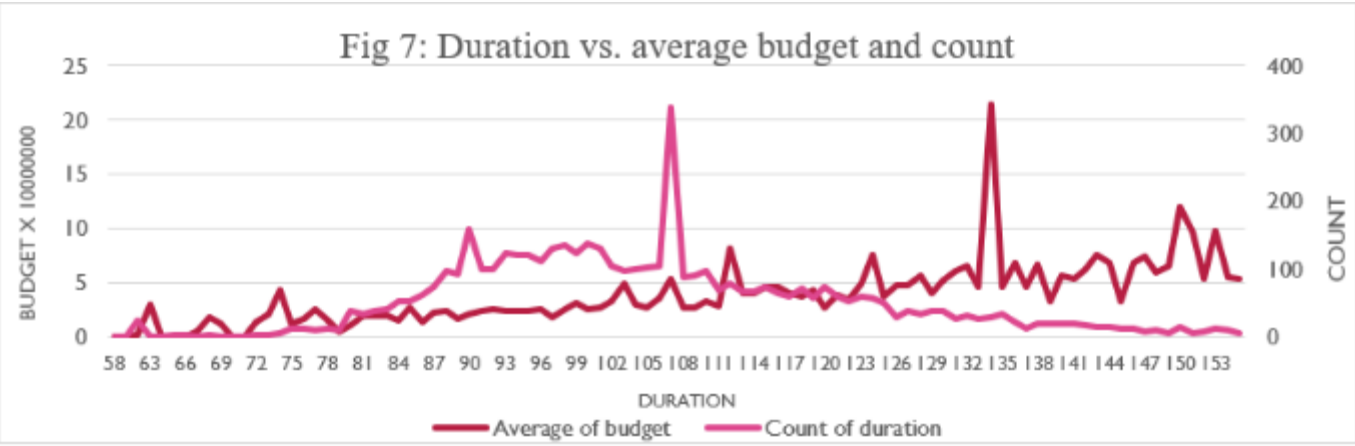
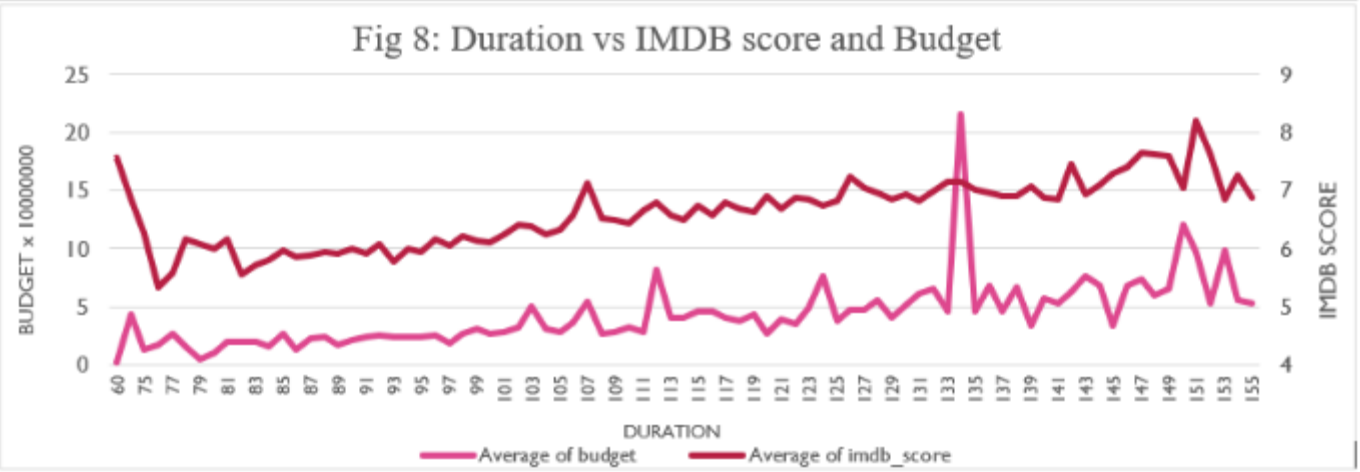


Fig 8: Duration vs IMDB score and Budget



3. Language Analysis:

a) A simple bar graph for count of language obtained from pivot table showing count for each language shows English to be the language with highest movie count by a mile. Then, IMDB score statistics for top 7 languages were calculated and represented in a clustered column chart.
b) For the IMDB Score distribution and PDF for English movies, the **NORM.DIST** formula was used to compute the probability density function (PDF) for each IMDB score for English movies. The mean and SD, essential for the calculation, were derived using the **AVERAGE** and **STDEV** functions.
c) SD bands were determined alongside the PDF values. Then, a table was created to compile IMDB score values, counts, PDF values, and SD bands, which was then used to plot the combination chart.
d) Pivot tables were created to obtain count of English and other movies for different countries, average gross & budget made by English vs. other movies for countries, cast likes for English vs. other movies in each country, etc. for further analysis.
e) Using these pivot tables, various line charts and stacked column charts were plotted all comparing success of English movies with other movies in different countries.

Fig 9A: Language Distribution

Fig 9B: Count of language

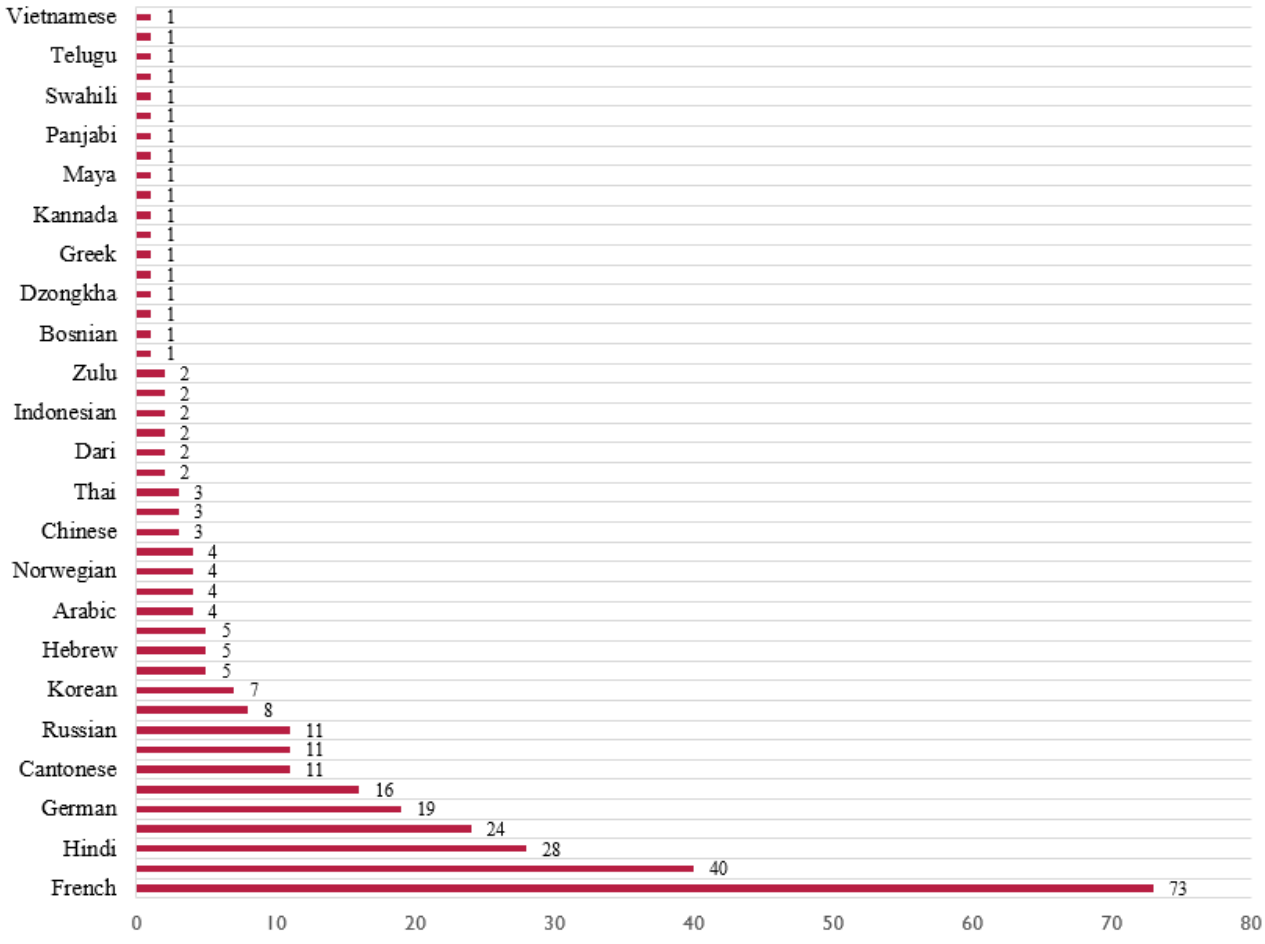


Fig 10: Language statistics

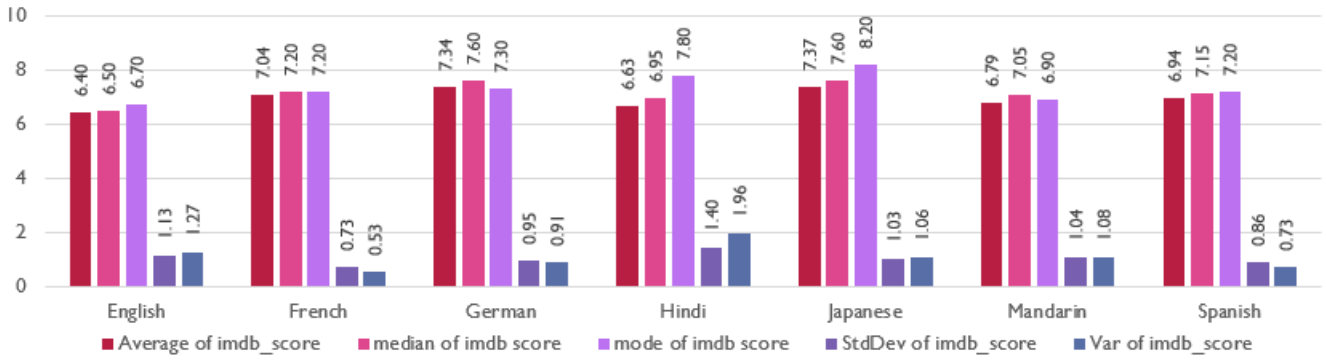


Fig 11: IMDB Score distribution and PDF for English movies

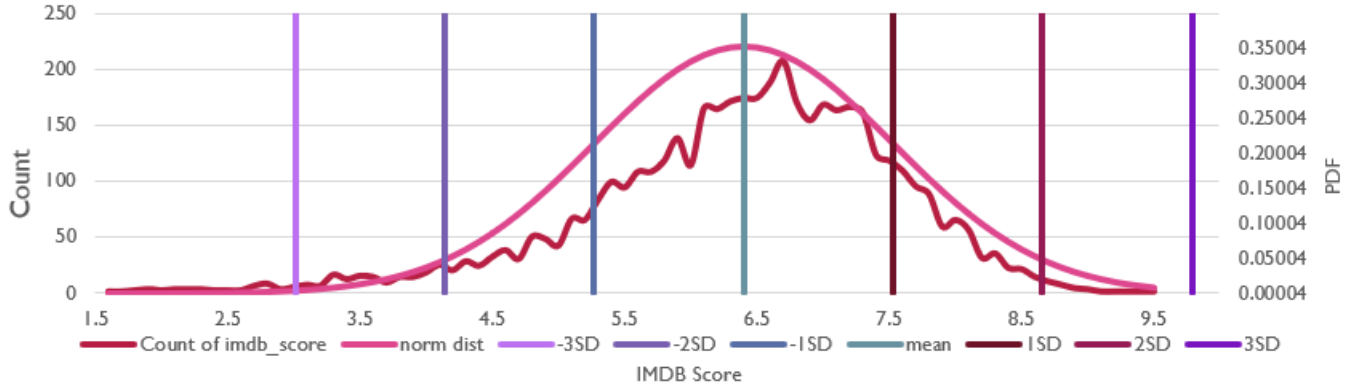


Fig 12: Countries with movies in English (and other languages)

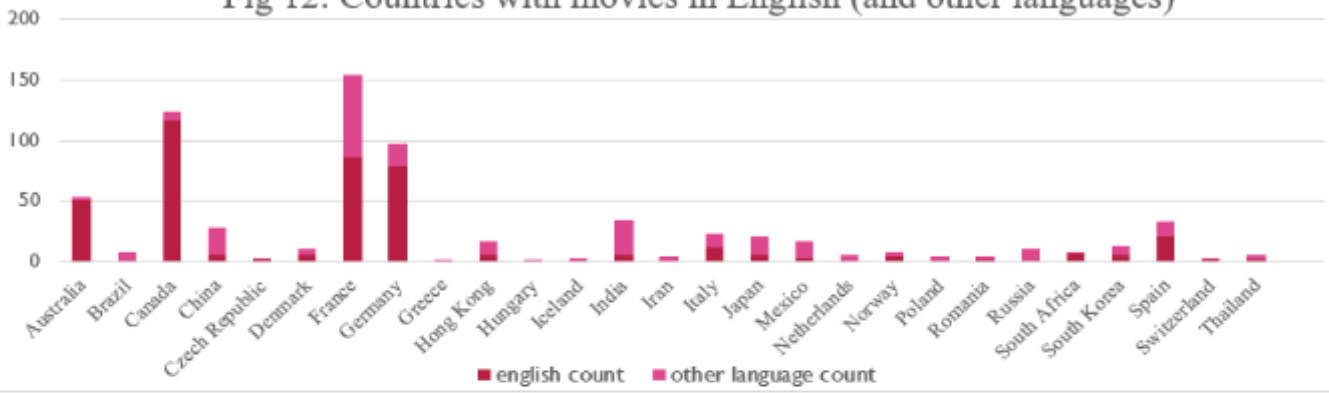
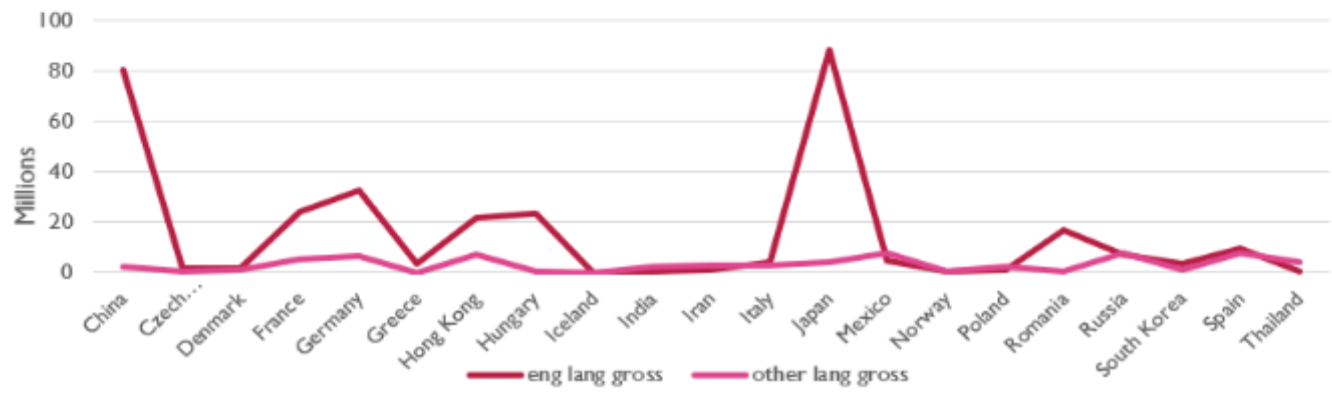
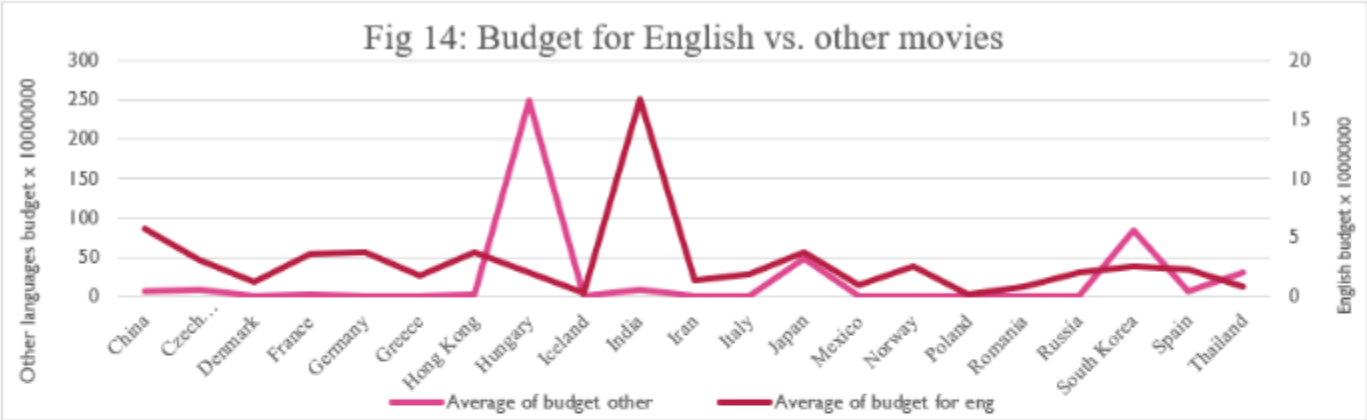
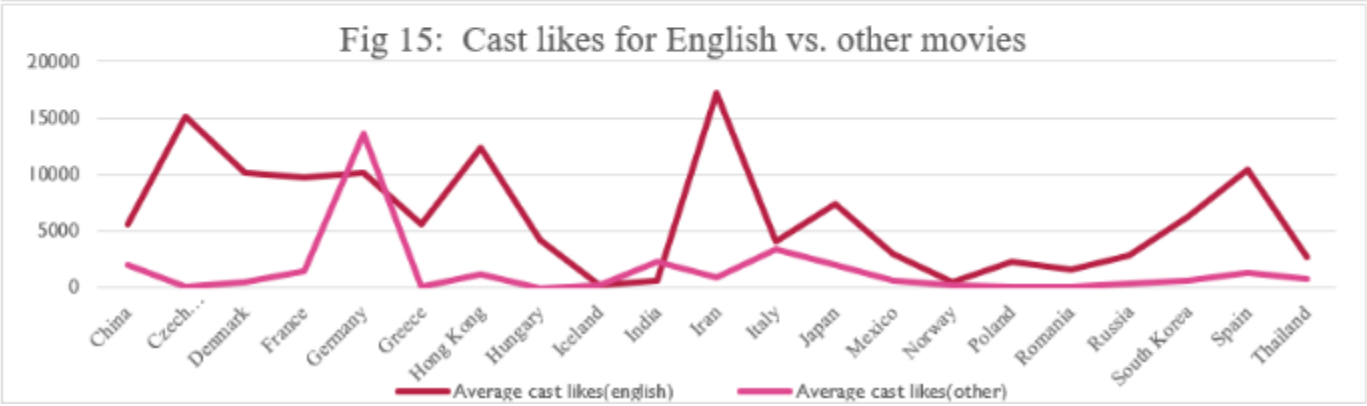


Fig 13: Country-wise Gross for English movies vs. other movies

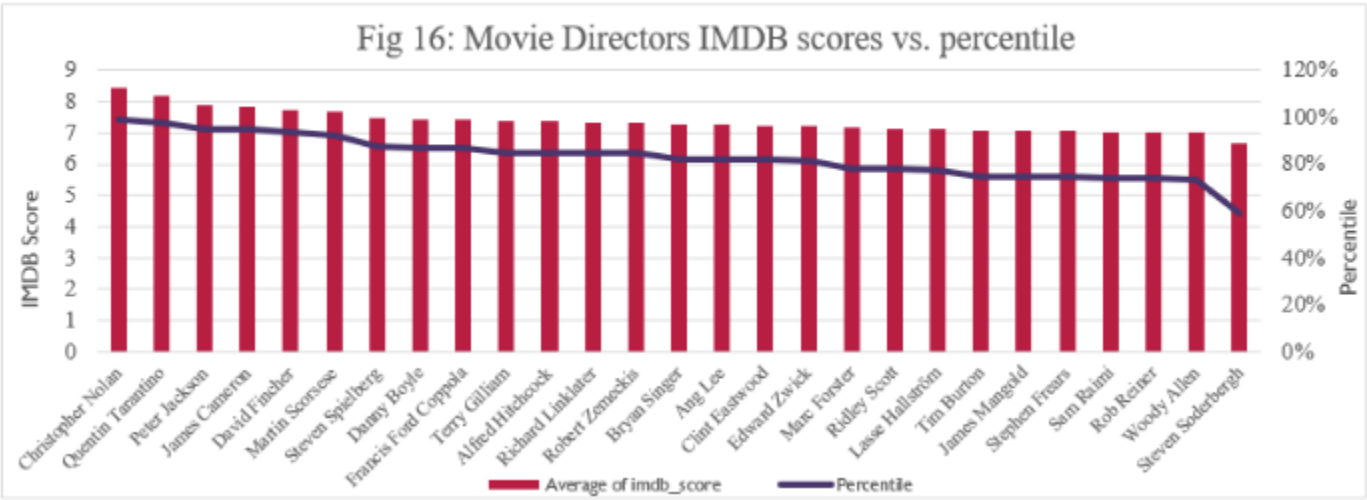


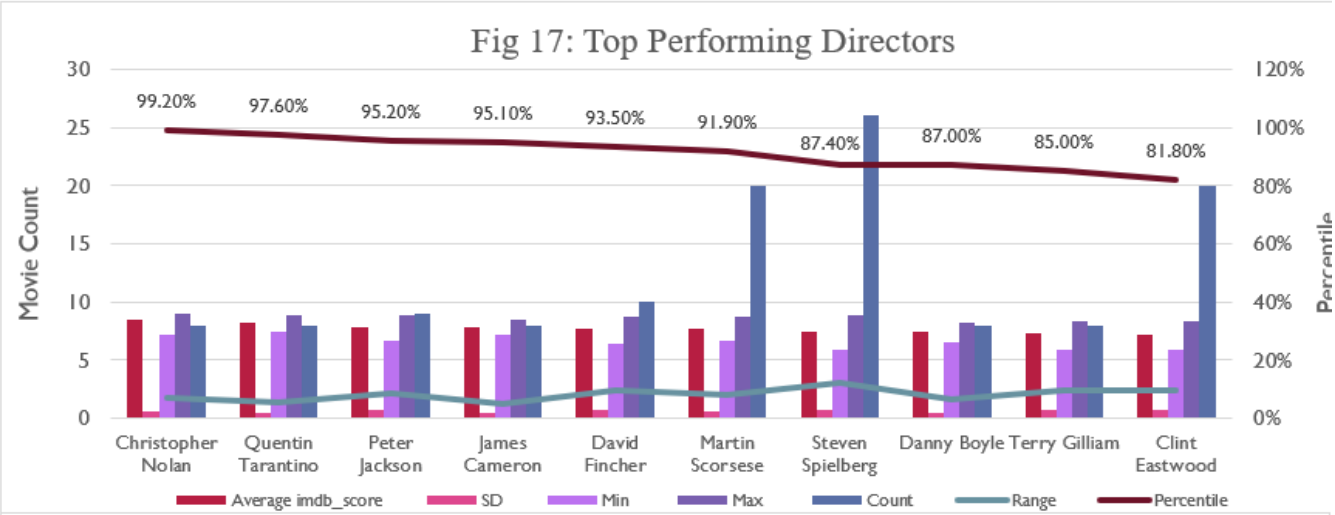


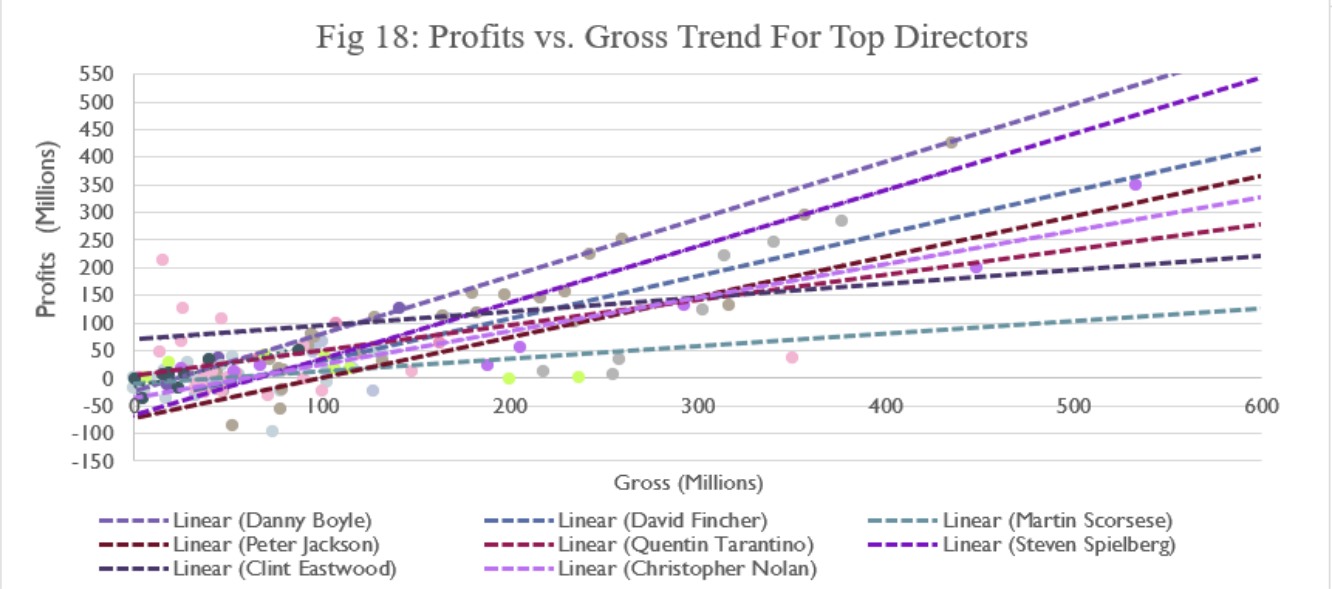


4. Director Analysis:

a) Pivot tables were used to analyze directors' IMDB statistics (count, average, SD, min, max), average/total gross and budget, and average IMDB per movie. Directors with at least eight movies and in the top 50th percentile were shortlisted. Further filters (70%+ percentile, score range ≤3, average ≥7) narrowed this to ten top performers.
b) IMDB score percentiles were calculated using Excel's percentile function, and directors above the 80th percentile were visualized in a combination chart. The top eight meeting all criteria were also charted.
c) Another pivot table calculated profits from total gross and budget. The top 10 directors were detailed in a table (movies, scores, gross, profit) and analyzed via scatter plots with trendlines and R^2 values.







5. Budget Analysis:

a) Using the 3881 observations in budget and gross columns in the table, a profits column was created to determine the top 300 and worst 300 movies based on profits gathered. Also, a table was created for the top 20 movies.

b) Three scatter plots were made, budget vs. gross, budget vs. profit and IMDB score count (Top 300 and worst 300 movies). The third graph utilized only IMDB scores of the 300 best and worst movies (profit-wise). R-squared values were displayed for the trendlines for profit and loss incurring movies.

c) Using the **COUNTIF** and **AVERAGEIF** formulas, the count of each genre was determined alongwith their average IMDB scores for both top 300 and worst 300 movies. A combination chart was made for genre count vs. average IMDB score.

d) Similarly, the genre-wise average gross, budget, profit and loss made was determined for both best and worst 300 movies. Using this information, two graphs were made: 'Genre wise avg. budget and avg. profits' and 'Genre vs. Avg profit and loss made'

Fig 19: Budget Vs. Gross

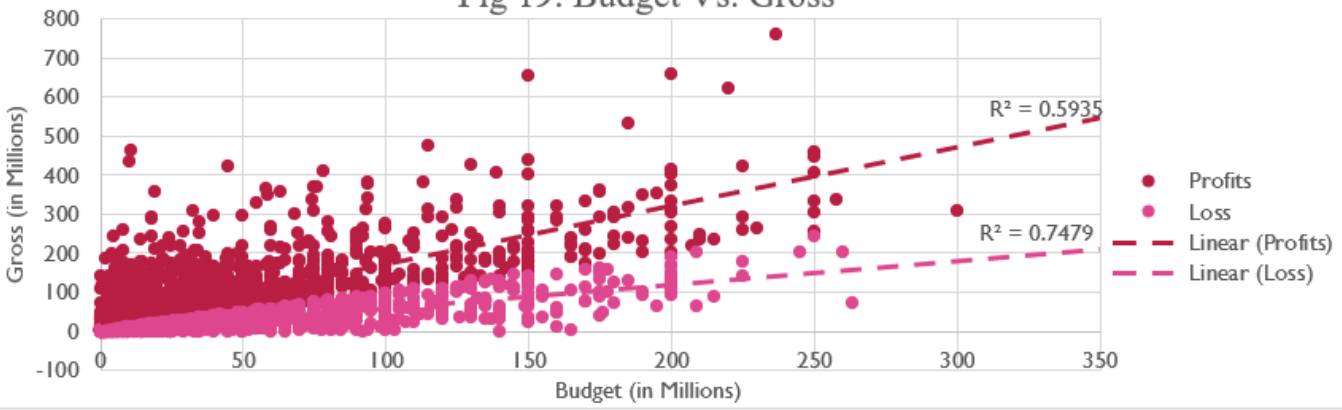


Fig 20: Budget vs. Profit

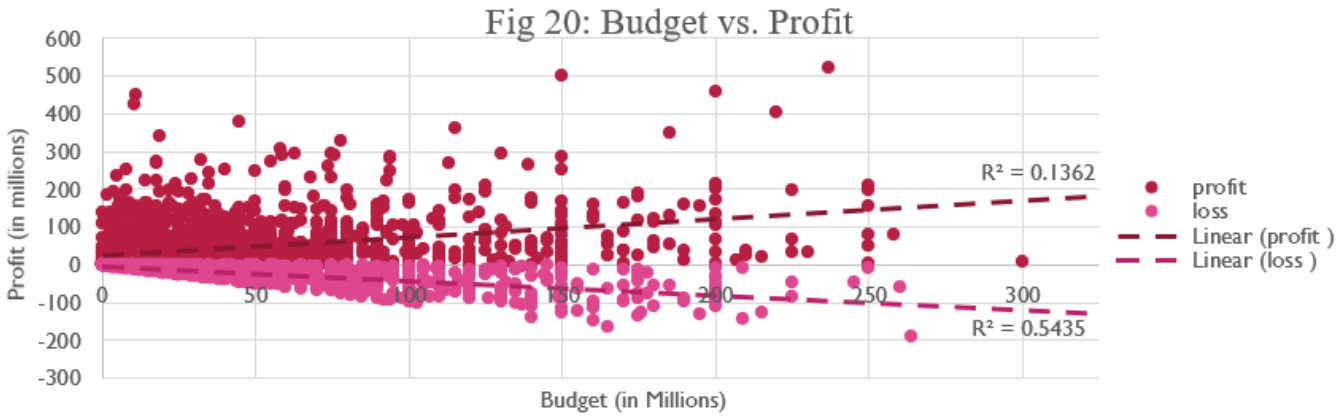


Fig 21: IMDB_Score Count (Top 300 Vs. Worst 300 Movies)

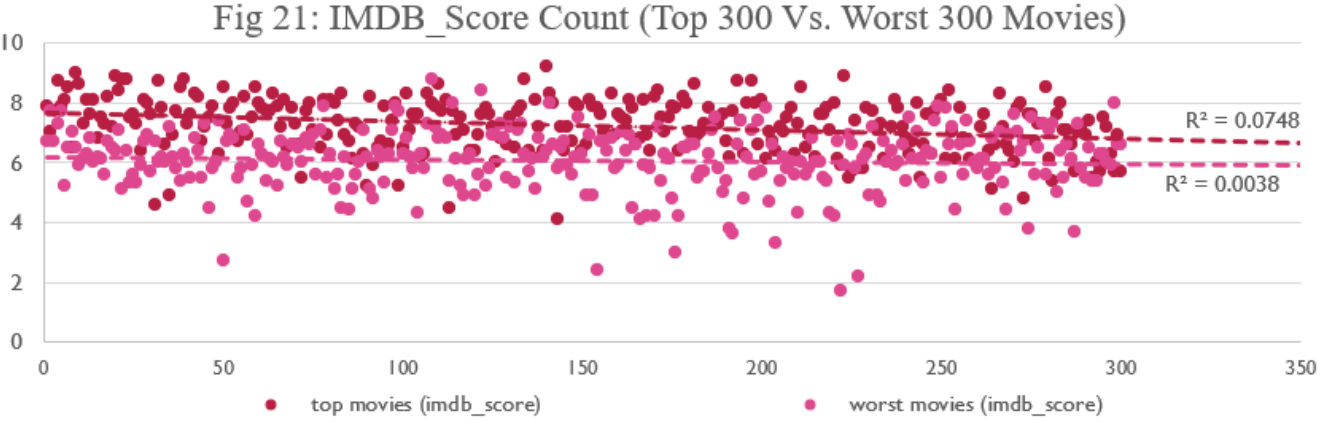


Fig 22: Genre count vs. avg imdb score

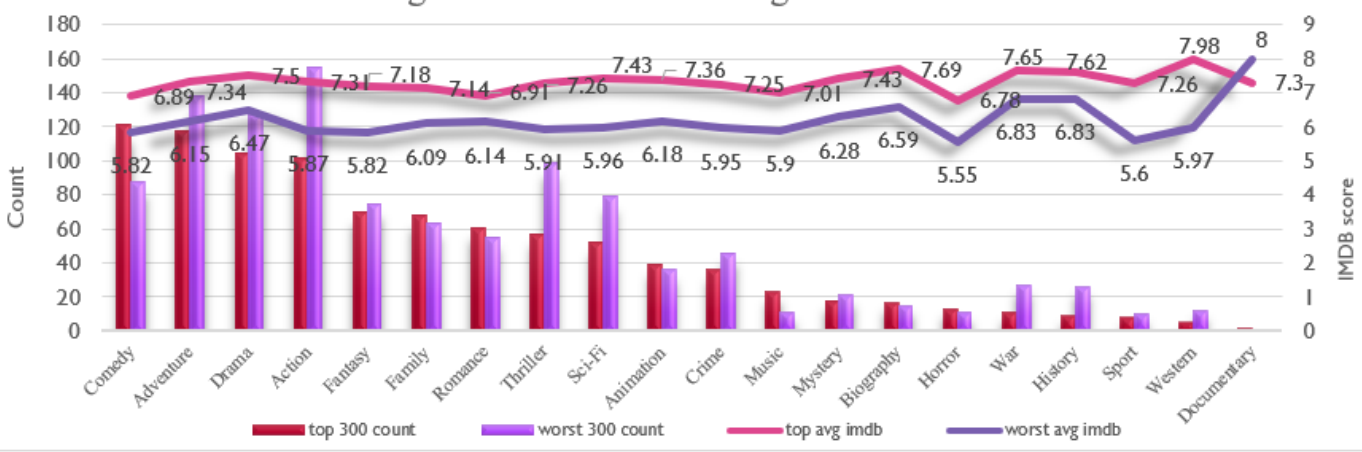


Fig 23: Genre wise avg. budget and avg. profits

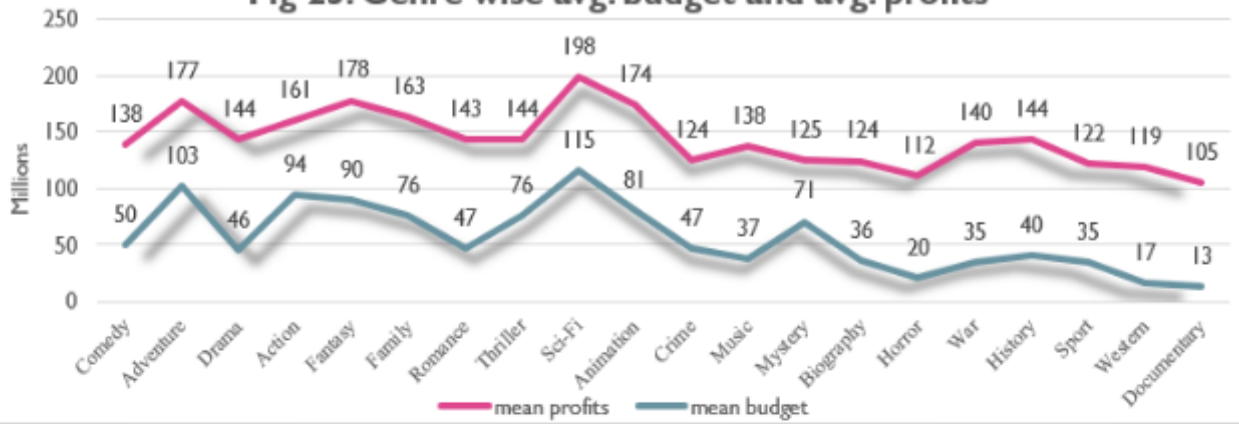


Fig 24: Genre vs. Avg profit and loss made

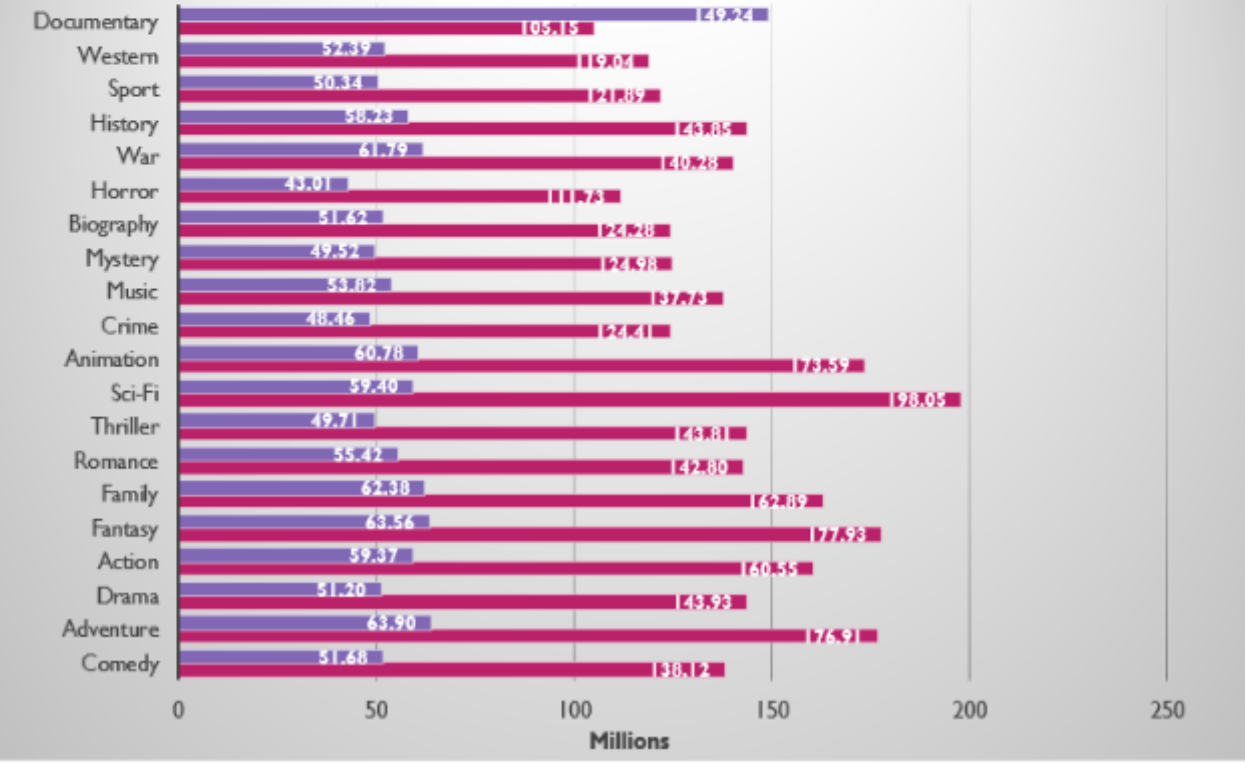
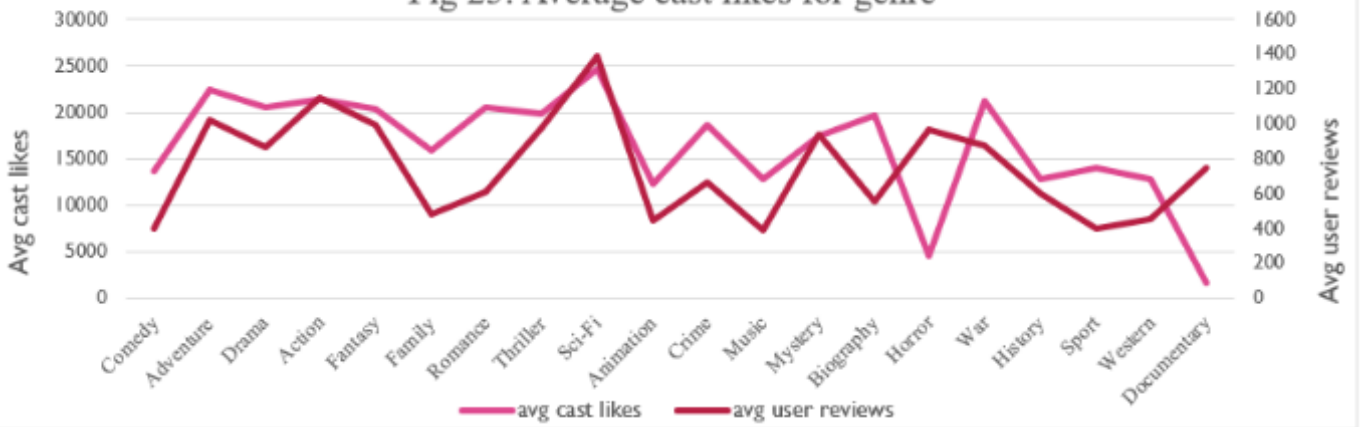


Fig 25: Average cast likes for genre



Movie Title	Genres
Avatar	Action Adventure Fantasy Sci-Fi
Jurassic World	Action Adventure Sci-Fi Thriller
Titanic	Drama Romance
Star Wars: Episode IV - A New Hope	Action Adventure Fantasy Sci-Fi
E.T. the Extra-Terrestrial	Family Sci-Fi
The Avengers	Action Adventure Sci-Fi
The Lion King	Adventure Animation Drama Family Music
Star Wars: Episode I - The Phantom Menace	Action Adventure Fantasy Sci-Fi
The Dark Knight	Action Crime Drama Thriller
Spirited Away	Adventure Animation Family Fantasy
The Hunger Games	Adventure Drama Sci-Fi Thriller
Deadpool	Action Adventure Comedy Romance Sci-Fi
The Hunger Games: Catching Fire	Adventure Sci-Fi Thriller
Jurassic Park	Adventure Sci-Fi Thriller
The Secret Life of Pets	Animation Comedy Family
Despicable Me 2	Animation Comedy Family Sci-Fi
American Sniper	Action Biography Drama History Thriller War
Finding Nemo	Adventure Animation Comedy Family
Shrek 2	Adventure Animation Comedy Family Fantasy Romance
The Lord of the Rings: The Return of the King	Action Adventure Drama Fantasy

ANALYSIS:

1. Movie Genre Analysis:

The count of each genre is given in Fig 1. Some movies have a single genre and some have multiple genres. The top genres are: Drama, Comedy, Thriller, Action, Romance, Adventure, Crime, Sci-Fi, Fantasy, Horror and Family. The statistics related to each genre is shown in Fig 2.

Q1. Why does Drama have the highest count in genre?

It may be due to the fact that most people want movies to invoke strong emotions in them which leads to the audience feeling a connection to the movie. Also, drama to a large degree, entails some other forms of genres or is used as a secondary genre to other genres. This may help in attracting a larger audience.

Q2. Why do a lot of movies with drama genre have other genres in combination with Drama?

Drama genre has the highest count of movies however it's mostly in conjunction with other genres. These genres in association with the drama genre, create stories and movie scripts that invoke a variety of emotions in the audience and attracting a larger audience. According to Fig 3. Comedy, Romance, Thriller and Crime are the genres used most in combination with the Drama genre. The related statistics can be seen in Fig 4.

2. Movie Duration Analysis:

Fig 5. displays the relationship between movie duration and IMDB score. The trendline seems to suggest a positive correlation between duration and IMDB scores.

Q1. Why is the correlation between duration and IMDB scores weak?

The low R-squared value indicates no strong relationship between duration and IMDB scores. This may be due to the fact that most datapoints lie between 80 to 120 minutes and the sample points are less towards either ends. Fig 6. shows a normal distribution for duration, count and IMDB scores. Therefore, 67% of datapoints lie between 90 to 120 minutes.

Q2. Why do most movies have a duration between 90 to 120 minutes?

According to Fig 7. most movies of a higher duration require a higher budget. There is a moderate correlation between duration of a movie with its budget suggesting a positive correlation.

Q3. Why do movies with higher duration require higher budget?

This may be understood by Fig 8. there is a positive correlation between movie durations and the average IMDB scores. Also, the average budget increases along with duration of the show. This may be due to the fact that higher budget leads to higher production quality and better resources leading to better ratings. Since there is a moderate relationship between duration and budget, the same can be said about duration and IMDB scores.

---

3. Language Analysis:

Fig 9. shows that English language dominates the list of movies, followed by French, Spanish, Hindi, Mandarin, German and Japanese. The statistics for these languages are given in Fig 10 suggesting English having lowest median and high SD.

Q1. Why does English have overall the lowest median and comparatively higher standard deviation?

English has the highest datapoints which also follow the normal distribution curve in Fig 11. Most movies (around 67%) fall into the IMDB Score range 5.5-7.5 which lowers the median IMDB Score. The presence of datapoints on extreme ends of the scale contributes to a slightly higher SD.

Q2. Why is English the most popular language for movies?

English is the most popular language spoken across the world. Coupling that with the success of Hollywood, it makes sense that most movies on the list are produced in English. A few other countries outside USA have also produced English movies (Fig 12).

Q3. Why do countries that don't speak English as their primary language make English movies?

Fig 13 shows that for most countries, the average gross made by English movies is higher than the average gross made by other movies. Another reason could be Hollywood production members taking an interest in other countries and collaborating with those to produce English movies as special projects.

Q4. Why is the gross revenue collected higher for English movies than gross revenues of movies in regional languages?

According to the charts (Fig 14 & 15), even though the budget for most movies is similar regardless of the language (except for few countries), the gross revenue is higher for English movies. This could be due to the production using better resources and casting popular actors for their cast in English movies.

4. Director Analysis:

Ten directors were determined to be the top performing directors according to previously mentioned criteria and shortlisted from the list of directors in Fig 16:

Top Directors	Average IMDB score	SD	Min	Max	Count	Range	Percentile
Christopher Nolan	8.43	0.54	7.2	9	8	1.8	99.20%
Quentin Tarantino	8.20	0.42	7.5	8.9	8	1.4	97.60%
Peter Jackson	7.89	0.77	6.7	8.9	9	2.2	95.20%
James Cameron	7.85	0.46	7.2	8.5	8	1.3	95.10%
David Fincher	7.75	0.72	6.4	8.8	10	2.4	93.50%
Martin Scorsese	7.66	0.60	6.7	8.7	20	2	91.90%
Steven Spielberg	7.48	0.74	5.9	8.9	26	3	87.40%
Danny Boyle	7.44	0.52	6.6	8.2	8	1.6	87.00%
Terry Gilliam	7.36	0.77	5.9	8.3	8	2.4	85.00%
Clint Eastwood	7.23	0.70	5.9	8.3	20	2.4	81.80%

Q1. Why are these directors the top performing directors?

The above mentioned 10 directors all have produced a minimum of 8 movies as well as have an average IMDB score in the >80% range, low standard deviations for the IMDB scores, movies with average IMDB score > 7 and a low range showing good consistency (Fig 17).

Q2. Why is it important to determine top-performing directors?

Stronger statistics of these directors is a result of positive response from the audience leading to a higher box-office revenue. Fig 18 shows trendlines for gross revenue vs. profits made by the movies of mentioned directors. The directors that increase in margin with increase in gross revenue are:

Top performing directors	R-squared values
Christopher Nolan	0.8885
Quentin Tarantino	0.5498
Peter Jackson	0.6086
David Fincher	0.512
Martin Scorsese	0.0899
Steven Spielberg	0.8405
Danny Boyle	0.9145
Terry Gilliam	0.6581

5. Budget Analysis:

A comparison between movie budgets vs. movie gross revenue (Fig 19) and movie budgets vs. profits made (Fig 20) shows that most movies were profitable in number. However, the R-squared value was higher for movies that incurred a loss than for those that generated a profit.

Q1. Why are the R-squared values bigger for movies that incurred a loss?

A bigger R-squared value means that the higher the budget of a movie is, the more likely it is to incur a loss instead of making a profit. However, R-squared values for both profitable movies and loss-making movies are more than 0.5 in the first chart indicating a strong relationship for both scenarios. Therefore, more analysis is required to determine what makes a movie a financial success or failure.

Q2. Why does higher budget not guarantee higher profits?

Fig 20. shows a moderate relationship between loss incurred and budget meanwhile a weaker relationship for budget vs. profitable movies. Further analysis of budget and other factors like genre, cast, director, and marketing across a sample of profitable and unprofitable movies may provide deeper insights. Fig 21 shows almost no correlation between profit/loss incurred movies and the IMDB scores. However, most profitable movies have the IMDB score between 6 and 9.

Q3. Why is there almost no correlation between IMDB scores and the profitability of movies?

Even though most profitable movies have IMDB scores between 6 and 9, the correlation is very weak. This suggests that while having an above-average IMDB score might be beneficial, it is not the sole determinant of a movie's financial success. Therefore, it may be more insightful to analyze the correlation between genre, budget/profit, and average IMDB scores to understand the broader picture.

The charts (Fig 22, 23 & 24) show that average profits rise with budget. Sci-Fi, Action, Adventure, Fantasy, and Animation yield high profits but also have more losses. Comedy, Drama, Adventure, and Action have the most films, with similar counts of profitable and unprofitable movies. Overall, the top-performing genres are **Sci-Fi, Adventure, Fantasy, Animation, Family,** and **Action.**

Q4. Why are genres like Sci-Fi, Action, Adventure, Fantasy, and Animation more profitable to exploit than others?

According to the list for top 20 most profitable movies, most movies belong to a combination of these genres. This may be due to a few reasons. These genres typically offer high levels of excitement, visual effects, and imaginative storytelling, which can captivate audiences and encourage repeat viewings, franchises & sequels and merchandising deals which add to their profitability.

Q5. Why do these genres captivate audiences and encourage repeat viewings as well as franchises?

Despite their fantastical elements, these genres often explore universal themes such as heroism, good vs. evil, and personal growth. These themes resonate deeply with viewers, making the stories memorable and impactful. Audiences foster strong fan communities as they become emotionally invested in the characters, which encourages repeat viewings and anticipation for sequels. Fig 25 depicts the universal appeal of such genres based on their average user reviews and cast Facebook likes.

According to this chart movies with genres Sci-Fi, Action, Adventure, Fantasy seems to get the highest number of reviews from viewers and higher amount of likes for actors on Facebook. The genre animation has considerably low reviews and cast likes.

Q6. Why does the Animation genre have fewer user reviews and cast likes?

For most animated movies, the cast, even though they may include famous actors for voice-acting, is not the central focus since the characters are animated. Additionally, most animated movies target children as their primary audience that may not be familiar with reviews and may not care to leave one. However, they are still a good target for merchandise sales, similar to other genres that garner a fanbase.

CONCLUSION:

This project was very crucial in helping me understand the complexities in the world of movie production as well as the analysis that goes into predicting what makes a successful movie in terms of IMDB scores and gross revenue. A big realization during this project was that there's not a single component determining the success of a movie but rather a combination of factors. A genre on its own does not determine the success of a movie, but with a combination of factors such as the director's vision, the budget, the cast's performance, the screenplay, and even the marketing strategy, a movie can achieve significant success.

Another important aspect uncovered was the significance of targeting niche groups of consumers. By understanding and exploiting the interests of these specific audiences, movies can achieve substantial success even without broad appeal.

Furthermore, the project highlighted the financial benefits of movie franchises. Franchises tend to bring in more money due to their established fan base, brand recognition, and the ability to create a series of interconnected stories that keep audiences engaged over multiple installments.

BANK LOAN CASE STUDY

DESCRIPTION:

This project involves a dataset from a bank for calculating risky loan applications and deciding which applicants are more trustworthy than others. The bank is dealing with an issue: some customers with limited credit histories are defaulting on their loans. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. Using Exploratory Data Analysis (EDA) analyze patterns in the data and gain insights so that capable applicants are not rejected. When a customer applies for a loan, the company faces two key risks:

a) If a qualified applicant is denied, the company misses out on potential business.
b) If an unqualified applicant is approved and cannot repay the loan, the company incurs a financial loss.

The dataset analyzed contains information about loan applications, specifically in two scenarios:

a) Customers with repayment issues: These individuals were late by more than X days on at least one of the first Y installments.
b) All other cases: These represent customers who made their payments on time.

THE PROBLEM:

Objective	Task
1. **Identify Missing Data and Deal with it Appropriately**	Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.
2. **Identify Outliers in the Dataset**	Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.
3. **Analyze Data Imbalance**	Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.
4. **Perform Univariate, Segmented Univariate, and Bivariate Analysis**	Use 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.
5. **Identify Top Correlations 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.

DESIGN:

Dataset Summary: "application_data"

- **Number of Observations**: 49,999
- **Number of Variables**: 124
- **File Type**: Excel file
- **Names of Columns**:

SK_ID_CURR	REGION_RATING_CLIENT_W_CITY	COMMONAREA_MODE	OBS_60_CNT_SOCIAL_CIRCLE
TARGET	WEEKDAY_APPR_PROCESS_START	ELEVATORS_MODE	DEF_60_CNT_SOCIAL_CIRCLE
NAME_CONTRACT_TYPE	HOUR_APPR_PROCESS_START	ENTRANCES_MODE	DAYS_LAST_PHONE_CHANGE
CODE_GENDER	REG_REGION_NOT_LIVE_REGION	FLOORSMAX_MODE	FLAG_DOCUMENT_2
FLAG_OWN_CAR	REG_REGION_NOT_WORK_REGION	FLOORSMIN_MODE	FLAG_DOCUMENT_3
FLAG_OWN_REALTY	LIVE_REGION_NOT_WORK_REGION	LANDAREA_MODE	FLAG_DOCUMENT_4
CNT_CHILDREN	REG_CITY_NOT_LIVE_CITY	LIVINGAPARTMENTS_MODE	FLAG_DOCUMENT_5
AMT_INCOME_TOTAL	REG_CITY_NOT_WORK_CITY	LIVINGAREA_MODE	FLAG_DOCUMENT_6
AMT_CREDIT	LIVE_CITY_NOT_WORK_CITY	NONLIVINGAPARTMENTS_MODE	FLAG_DOCUMENT_7
AMT_ANNUITY	ORGANIZATION_TYPE	NONLIVINGAREA_MODE	FLAG_DOCUMENT_8
AMT_GOODS_PRICE	EXT_SOURCE_1	APARTMENTS_MEDI	FLAG_DOCUMENT_9
NAME_TYPE_SUITE	EXT_SOURCE_2	BASEMENTAREA_MEDI	FLAG_DOCUMENT_10
NAME_INCOME_TYPE	EXT_SOURCE_3	YEARS_BEGINEXPLUATATION_MEDI	FLAG_DOCUMENT_11
NAME_EDUCATION_TYPE	APARTMENTS_AVG	YEARS_BUILD_MEDI	FLAG_DOCUMENT_12
NAME_FAMILY_STATUS	BASEMENTAREA_AVG	COMMONAREA_MEDI	FLAG_DOCUMENT_13
NAME_HOUSING_TYPE	YEARS_BEGINEXPLUATATION_AVG	ELEVATORS_MEDI	FLAG_DOCUMENT_14
REGION_POPULATION_RELATIVE	YEARS_BUILD_AVG	ENTRANCES_MEDI	FLAG_DOCUMENT_15
DAYS_BIRTH	COMMONAREA_AVG	FLOORSMAX_MEDI	FLAG_DOCUMENT_16
DAYS_EMPLOYED	ELEVATORS_AVG	FLOORSMIN_MEDI	FLAG_DOCUMENT_17
DAYS_REGISTRATION	ENTRANCES_AVG	LANDAREA_MEDI	FLAG_DOCUMENT_18
DAYS_ID_PUBLISH	FLOORSMAX_AVG	LIVINGAPARTMENTS_MEDI	FLAG_DOCUMENT_19
OWN_CAR_AGE	FLOORSMIN_AVG	LIVINGAREA_MEDI	FLAG_DOCUMENT_20
FLAG_MOBIL	LANDAREA_AVG	NONLIVINGAPARTMENTS_MEDI	FLAG_DOCUMENT_21
FLAG_EMP_PHONE	LIVINGAPARTMENTS_AVG	NONLIVINGAREA_MEDI	AMT_REQ_CREDIT_BUREAU_HOUR
FLAG_WORK_PHONE	LIVINGAREA_AVG	FONDKAPREMONT_MODE	AMT_REQ_CREDIT_BUREAU_DAY
FLAG_CONT_MOBILE	NONLIVINGAPARTMENTS_AVG	HOUSETYPE_MODE	AMT_REQ_CREDIT_BUREAU_WEEK
FLAG_PHONE	NONLIVINGAREA_AVG	TOTALAREA_MODE	AMT_REQ_CREDIT_BUREAU_MON
FLAG_EMAIL	APARTMENTS_MODE	WALLSMATERIAL_MODE	AMT_REQ_CREDIT_BUREAU_QRT
OCCUPATION_TYPE	BASEMENTAREA_MODE	EMERGENCYSTATE_MODE	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_FAM_MEMBERS	YEARS_BEGINEXPLUATATION_MODE	OBS_30_CNT_SOCIAL_CIRCLE	
REGION_RATING_CLIENT	YEARS_BUILD_MODE	DEF_30_CNT_SOCIAL_CIRCLE	

B) Data pre-processing, cleaning and error rectification:

1. Columns with >40% null values were deleted to avoid skewed data. Other unimportant columns were also deleted. In total, 75 columns were dropped.

2. New columns like no. of parents, concat (income type + family) Years employed, FLAG IF EMP>AGE, registered for years, FLAG IF REG>AGE, id publish years, concat income + edu, CREDIT: INC, ANNUITY: INC, ANNUITY: CREDIT and CREDIT: GOODS. This was done to effectively calculate missing values in some columns as well as deal with outliers.

FINDINGS:

1. Identify Missing Data and Deal with it Appropriately

a) Missing values for AMT_GOODS_PRICE were filled with the corresponding rows' values for AMT_CREDIT.
b) Missing values in CNT_FAM_MEMBERS were replaced with mode value 2. All values with 'XNA' in CODE_GENDER were replaced with 'F' as most values were 'F'. Median values were used to fill null data in other numerical columns which didn't include amount of money or a period of time.
c) All values with 'XNA' in ORGANIZATION_TYPE have pensioners and unemployed people in NAME_INCOME_TYPE and therefore were replaced with "Unemployed".
d) Missing value in AMT_ANNUITY was replaced with median value of AMT_ANNUITY for the corresponding values from AMT_INCOME and AMT_CREDIT columns.
e) For missing values in OCCUPATION_TYPE, 'concat income + edu' column was evaluated corresponding to null values in OCCUPATION_TYPE. Then, the most commonly occurring OCCUPATION_TYPE for each 'concat income + edu' value was used to fill the null values in OCCUPATION_TYPE.
f) One 'Unknown' value in NAME_FAMILY_STATUS column was replaced with 'Married' considering the row's values in CNT_CHILDREN, CNT_FAM_MEMBERS, no. of parents in family and NAME_HOUSING_TYPE.

2. Identify Outliers in the Dataset

a) The outliers were made for certain numerical value columns using box and whiskers plot.
b) CNT_CHILDREN column shows one of the values as '11'. Even though this number is too high, the outliers in CNT_FAM_MEMBERS show the outliers suggesting '13' family members in a family which means it was a correct entry. A new column called 'no. of parents' was made to see if any value wasn't '1' or '2'. No cases were seen therefore outliers were kept as it is.
c) In the outlier plot for AMT_INCOME, values over 4000000 were taken as outliers. There was one value 117000000 which was too high and therefore replaced with median value 145800.
d) Both plots for AMT_CREDIT and AMT_GOODS_PRICE had outliers. To deal with this, new columns were made to look at the ratio between credit and income (CREDIT:INC) and credit and goods price (CREDIT: GOODS).
e) Plot for AMT_ANNUITY had outliers. To deal with this, new columns were made to look at the ratio between annuity and income (ANNUITY:INC) and annuity and credit (ANNUITY: CREDIT).
f) Any values in CREDIT: INC higher than 20 were looked into. Those that were unemployed and pensioners were removed from the list. Any values in CREDIT: GOODS higher than 2 were looked into. These values of credit were replaced with their respective goods prices.
g) Any values in ANNUITY: INC higher than 0.65 were looked into as annuity amounts higher than that would prove to be difficult to repay. Those that were unemployed and pensioners were removed.
h) Values in ANNUITY: CREDIT were looked into and no changes were made as the values were smaller.

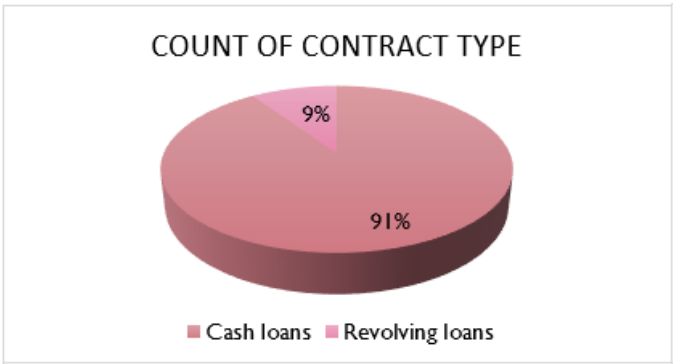
i) No values removed for registration and id publish years columns as all values were less that Age of applicant column values. No values removed for last phone years column as the highest value was around 10.
j) No outliers were seen in Age in years plot. However, the Years employed column had an outlier of 1000.67 years. These were replaced with median value 6.07.
k) Finally, 49937 rows and 61 columns of data remained.

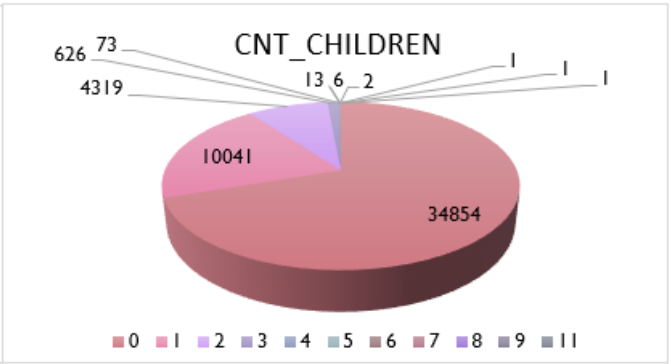
3. Analyze Data Imbalance

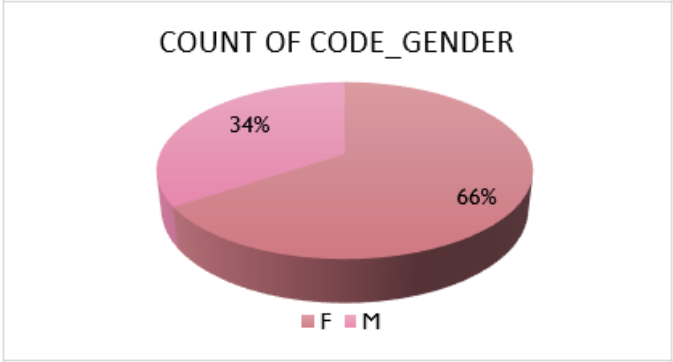
a) The number of applicants with no payment difficulties (0) significantly outnumbers the applicants with payment difficulties (1) within a dataset. This imbalance can lead to challenges in classification problems and studying reasons for payment difficulties.

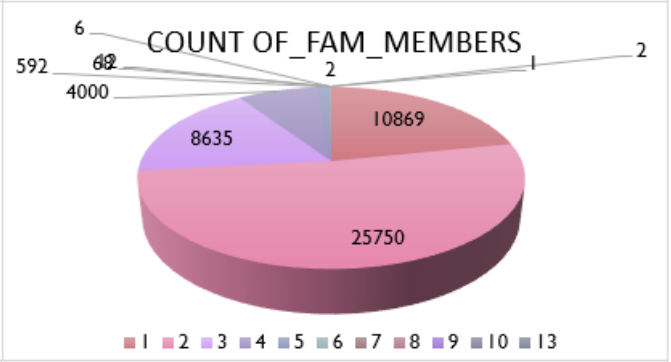
4. Perform Univariate, Segmented Univariate, and Bivariate Analysis

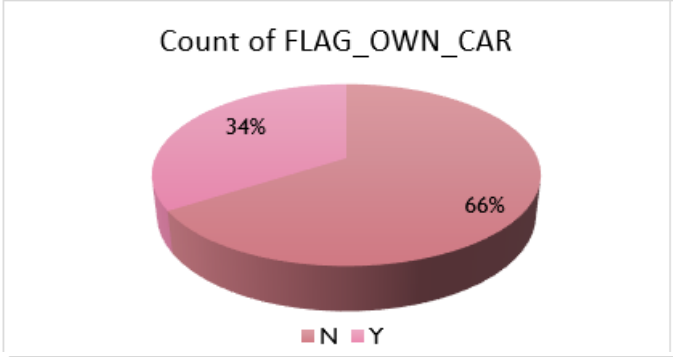
a) For univariate analysis, various pivot tables were made to get the distribution of count of gender, contract type, children, family members, ownership of car, house and region rating client. These were then used to generate pie charts.
b) Column charts were made for distribution of income type, education type, family status, housing type, occupation type and organization type. Histograms were made for distribution of age, employment years, income, annuity and credit amounts.
c) For segmented univariate analysis, clustered and stacked column charts were made for target (difficulty or no difficulty in paying off loans) was calculated against income types, gender, family count, parenthood, education type, region type, housing type, ownership of car and property.
d) Histograms were also made for age distribution, employed years and target.
e) For bivariate analysis, clustered column charts were made for average of income and credit vs. income type, education type, family status and housing type.
f) Finally, scatter plots were made for contract types and income vs. annuity, income vs. credit and credit vs. annuity.

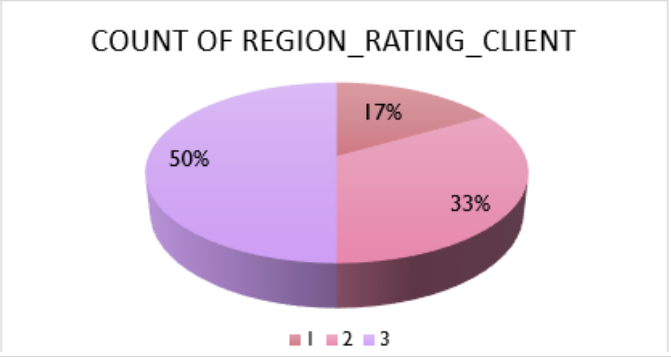

COUNT OF CONTRACT TYPE
9%
91%
Cash loans
Revolving loans

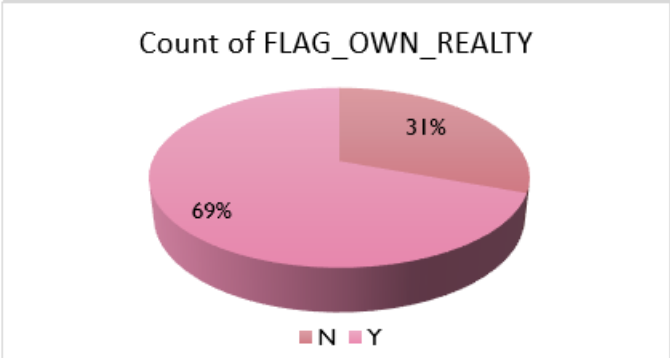

CNT_CHILDREN
626
73
4319
13
6
2
1
1
1
10041
34854
0
1
2
3
4
5
6
7
8
9
11

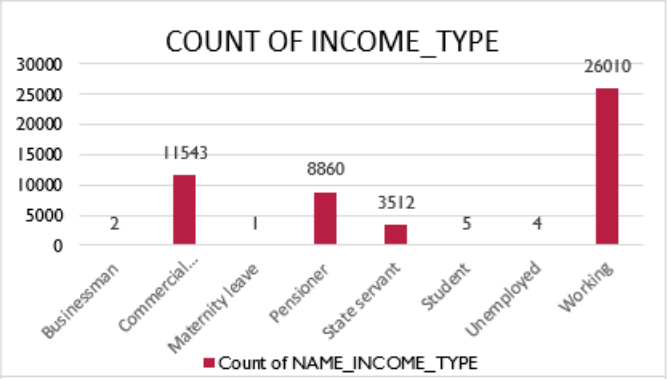

COUNT OF CODE_GENDER
34%
66%
F
M

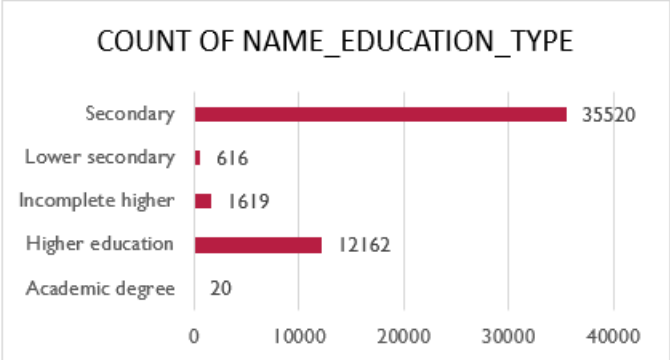

COUNT OF_FAM_MEMBERS
6
592
4000
2
1
2
8635
10869
25750
1
2
3
4
5
6
7
8
9
10
13

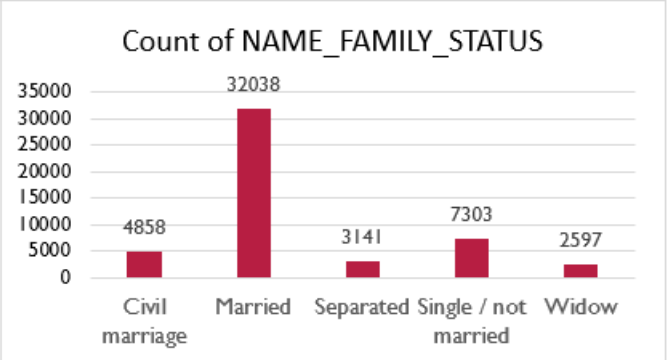

Count of FLAG_OWN_CAR
34%
66%
N
Y

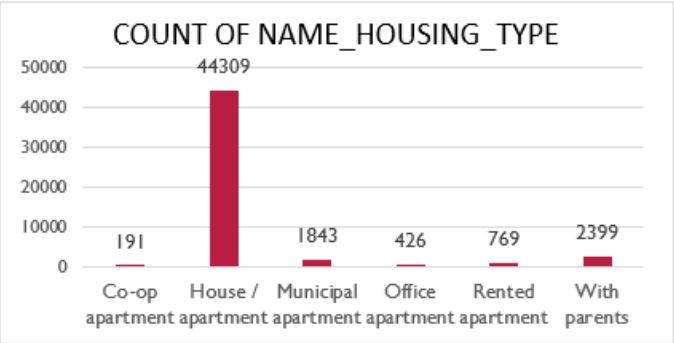

COUNT OF REGION_RATING_CLIENT
17%
50%
33%
1
2
3

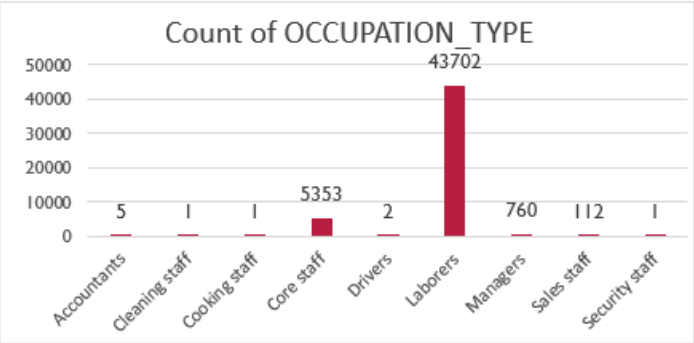

Count of FLAG_OWN_REALTY
31%
69%
N
Y

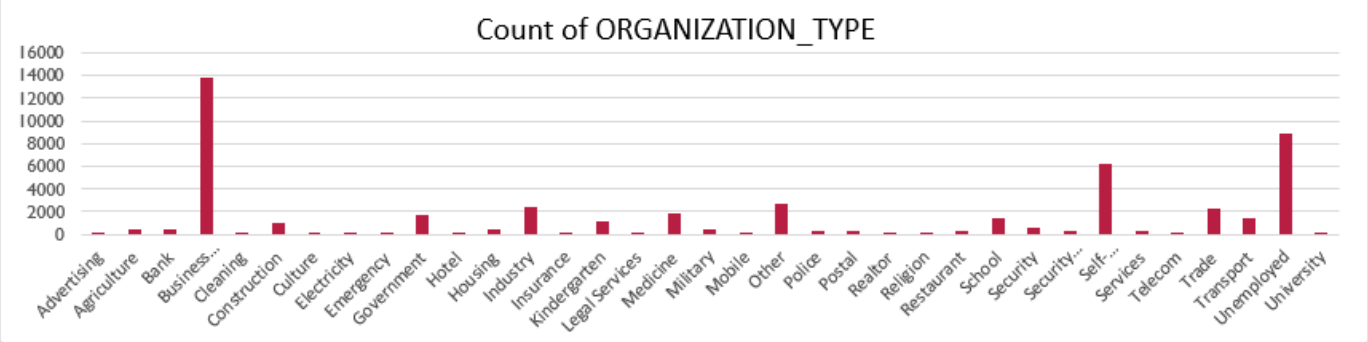

COUNT OF INCOME_TYPE
30000
25000
20000
15000
10000
5000
0
2
11543
1
8860
3512
5
4
26010
Businessman
Commercial...
Maternity leave
Pensioner
State servant
Student
Unemployed
Working
Count of NAME_INCOME_TYPE

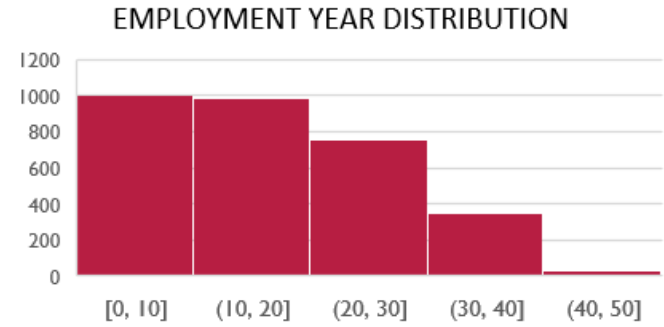

COUNT OF NAME_EDUCATION_TYPE
Secondary
35520
Lower secondary
616
Incomplete higher
1619
Higher education
12162
Academic degree
20
0
10000
20000
30000
40000

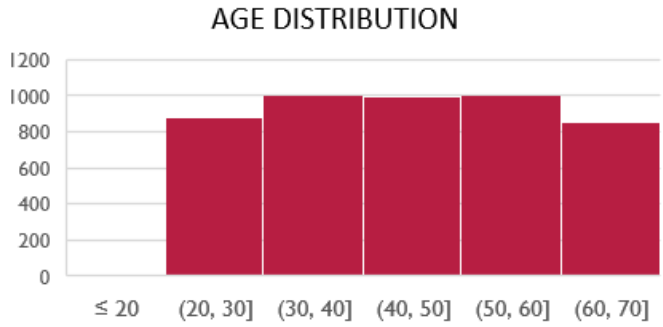

Count of NAME_FAMILY_STATUS
35000
30000
25000
20000
15000
10000
5000
0
4858
32038
3141
7303
2597
Civil marriage
Married
Separated
Single / not married
Widow

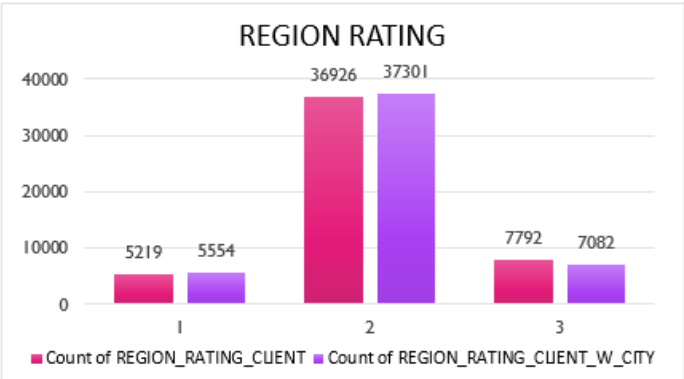

COUNT OF NAME_HOUSING_TYPE
50000
40000
30000
20000
10000
0
191
44309
1843
426
769
2399
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents

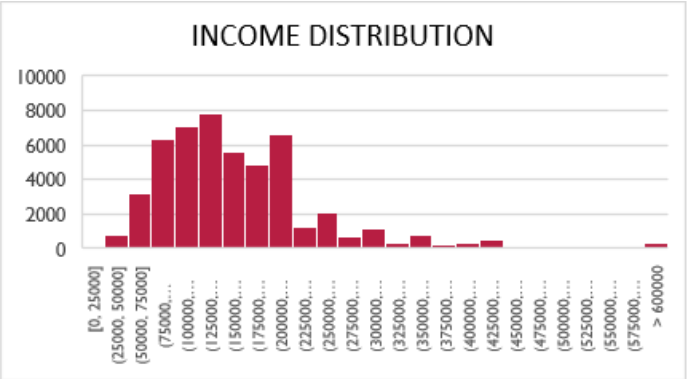

Count of OCCUPATION_TYPE
50000
40000
30000
20000
10000
0
5
1
1
5353
2
43702
760
112
1
Accountants
Cleaning staff
Cooking staff
Core staff
Drivers
Laborers
Managers
Sales staff
Security staff

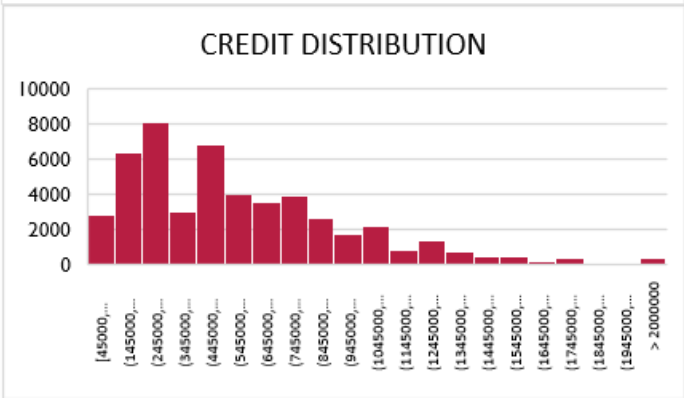

Count of ORGANIZATION_TYPE
16000
14000
12000
10000
8000
6000
4000
2000
0
Advertising
Agriculture
Bank
Business...
Cleaning
Construction
Culture
Electricity
Emergency
Government
Hotel
Housing
Industry
Insurance
Kindergarten
Legal Services
Medicine
Military
Mobile
Other
Police
Postal
Realtor
Religion
Restaurant
School
Security
Security...
Self-...
Services
Telecom
Trade
Transport
Unemployed
University

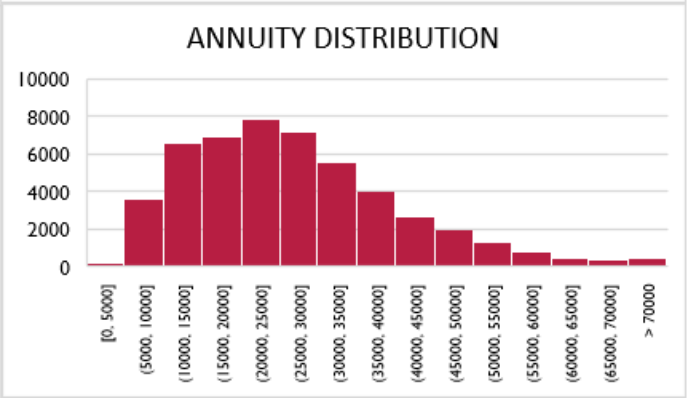

EMPLOYMENT YEAR DISTRIBUTION
1200
1000
800
600
400
200
0
[0, 10]
(10, 20]
(20, 30]
(30, 40]
(40, 50]

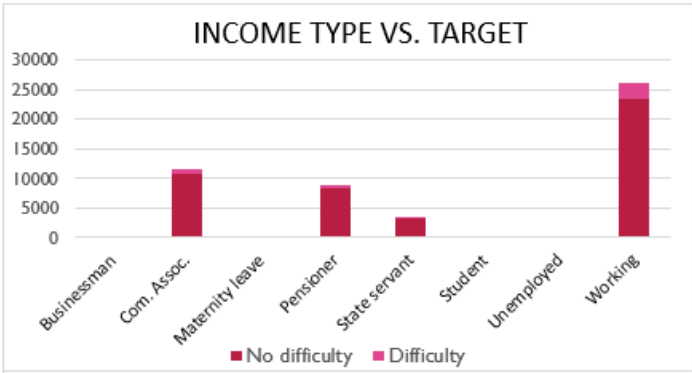

AGE DISTRIBUTION
1200
1000
800
600
400
200
0
≤ 20
(20, 30]
(30, 40]
(40, 50]
(50, 60]
(60, 70]

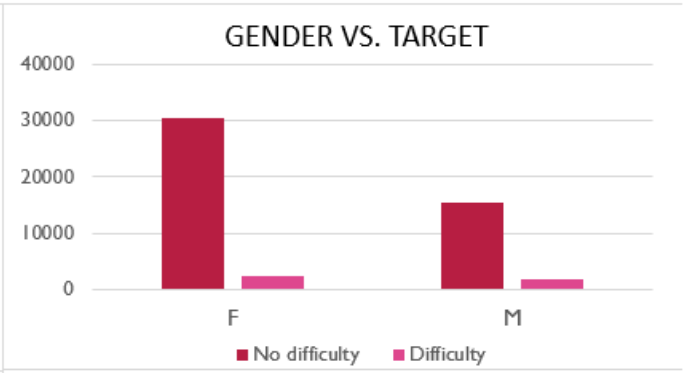

REGION RATING
40000
30000
20000
10000
0
5219
5554
36926
37301
7792
7082
1
2
3
Count of REGION_RATING_CLIENT
Count of REGION_RATING_CLIENT_W_CITY

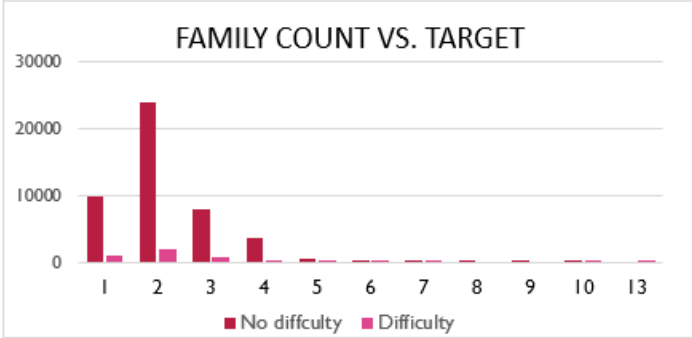

INCOME DISTRIBUTION
10000
8000
6000
4000
2000
0
[0, 25000]
(25000, 50000]
(50000, 75000]
> 600000

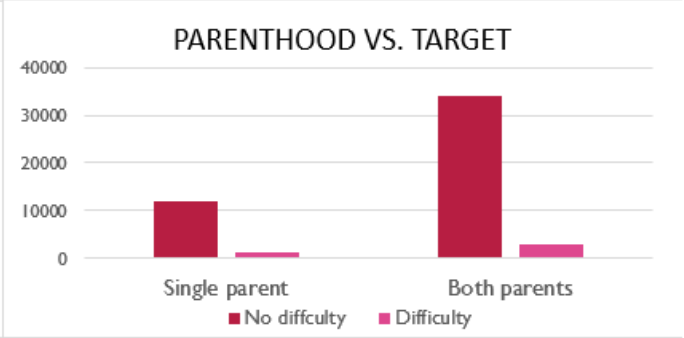

CREDIT DISTRIBUTION
10000
8000
6000
4000
2000
0
> 2000000

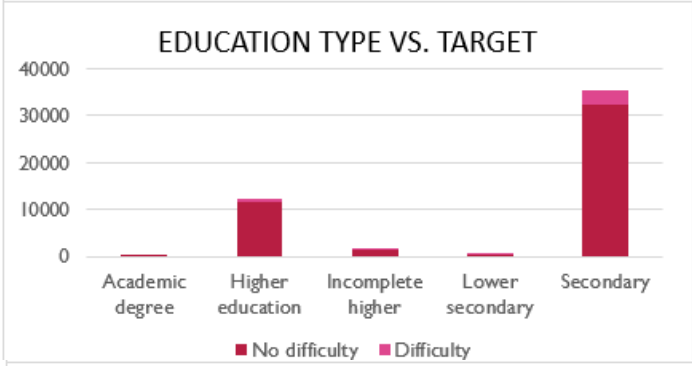

ANNUITY DISTRIBUTION
10000
8000
6000
4000
2000
0
[0, 5000]
(5000, 10000]
(10000, 15000]
(15000, 20000]
(20000, 25000]
(25000, 30000]
(30000, 35000]
(35000, 40000]
(40000, 45000]
(45000, 50000]
(50000, 55000]
(55000, 60000]
(60000, 65000]
(65000, 70000]
> 70000

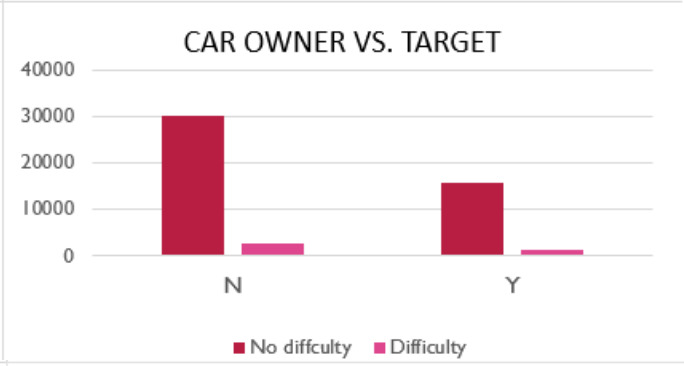

INCOME TYPE VS. TARGET
30000
25000
20000
15000
10000
5000
0
Businessman
Com. Assoc.
Maternity leave
Pensioner
State servant
Student
Unemployed
Working
No difficulty
Difficulty

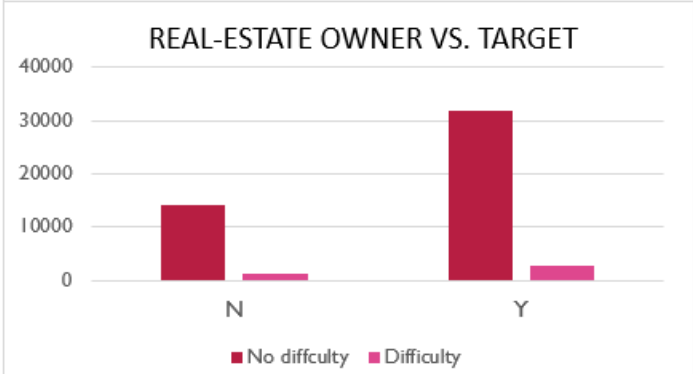

GENDER VS. TARGET
40000
30000
20000
10000
0
F
M
No difficulty
Difficulty

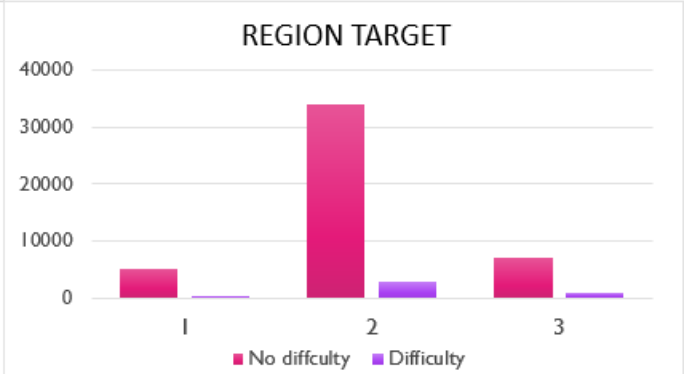

FAMILY COUNT VS. TARGET
30000
20000
10000
0
1
2
3
4
5
6
7
8
9
10
13
No diffculty
Difficulty

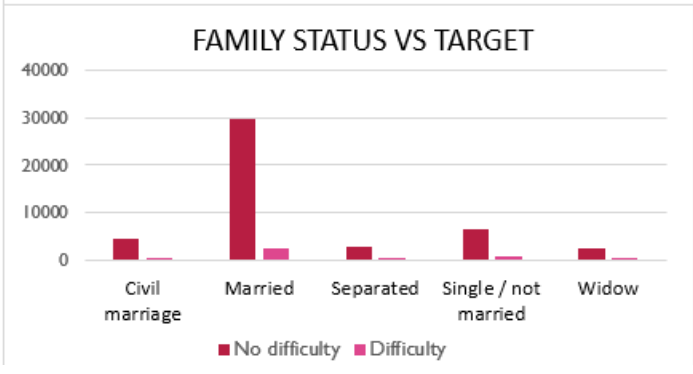

PARENTHOOD VS. TARGET
40000
30000
20000
10000
0
Single parent
Both parents
No diffculty
Difficulty

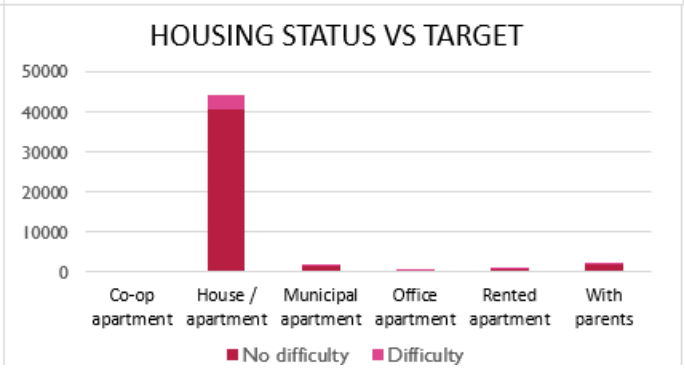

EDUCATION TYPE VS. TARGET
40000
30000
20000
10000
0
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary
No difficulty
Difficulty

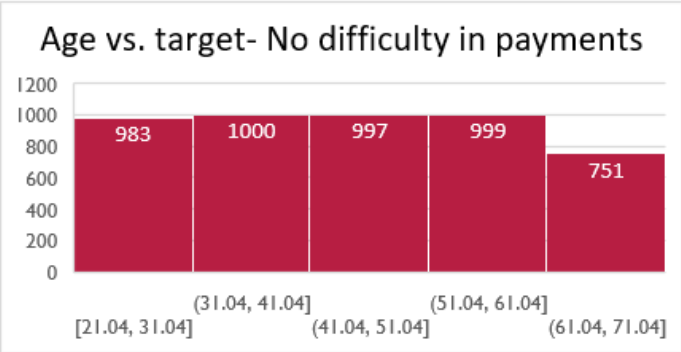

CAR OWNER VS. TARGET
40000
30000
20000
10000
0
N
Y
No diffculty
Difficulty

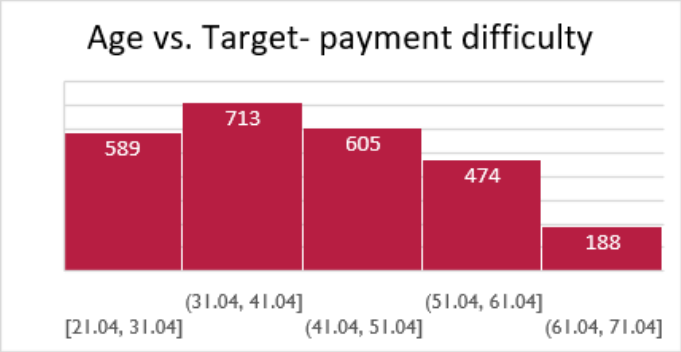

REAL-ESTATE OWNER VS. TARGET
40000
30000
20000
10000
0
N
Y
No diffculty
Difficulty

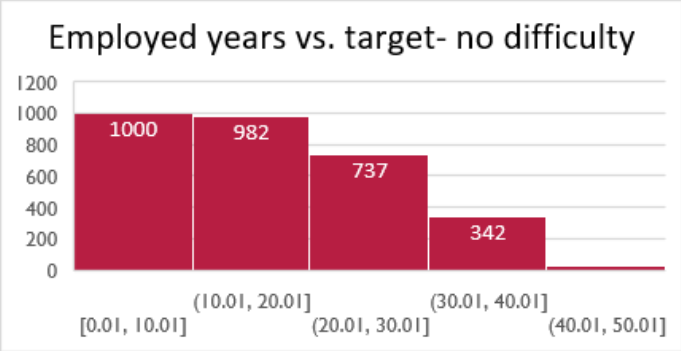

REGION TARGET
40000
30000
20000
10000
0
1
2
3
No diffculty
Difficulty

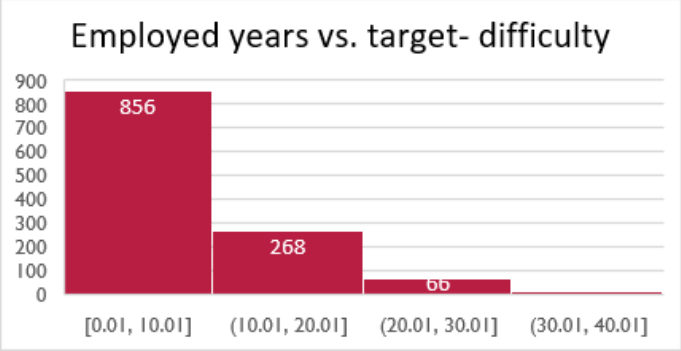

FAMILY STATUS VS TARGET
40000
30000
20000
10000
0
Civil marriage
Married
Separated
Single / not married
Widow
No difficulty
Difficulty

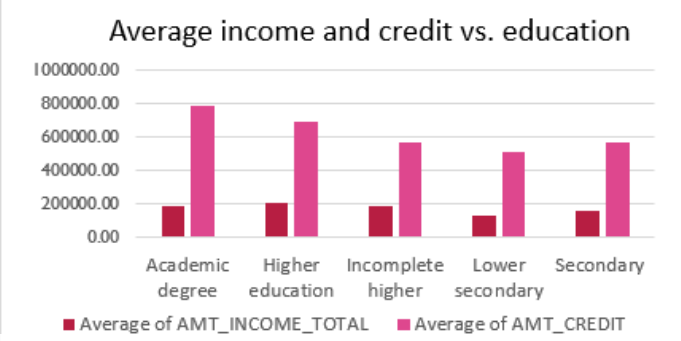

HOUSING STATUS VS TARGET
50000
40000
30000
20000
10000
0
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents
No difficulty
Difficulty

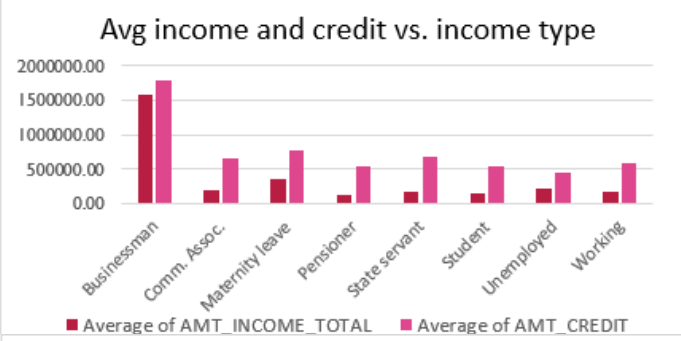

Age vs. target- No difficulty in payments
1200
1000
800
600
400
200
0
983
1000
997
999
751
(31.04, 41.04]
(51.04, 61.04]
[21.04, 31.04]
(41.04, 51.04]
(61.04, 71.04]

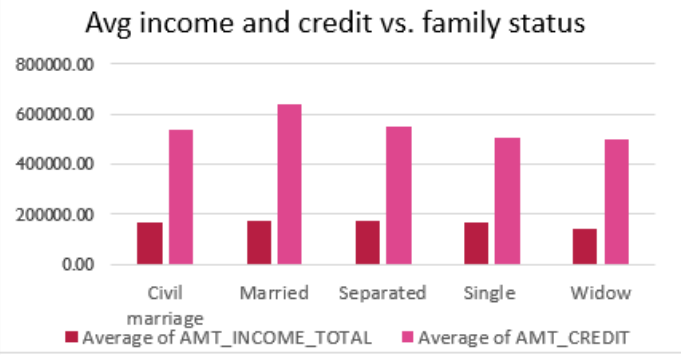

Age vs. Target- payment difficulty
589
713
605
474
188
(31.04, 41.04]
(51.04, 61.04]
[21.04, 31.04]
(41.04, 51.04]
(61.04, 71.04]

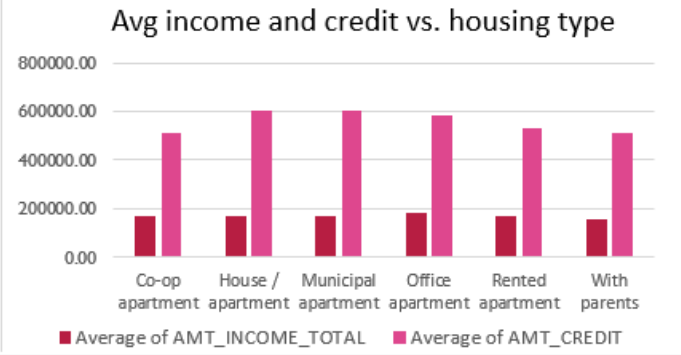

Employed years vs. target- no difficulty
1200
1000
800
600
400
200
0
1000
982
737
342
(10.01, 20.01]
(30.01, 40.01]
[0.01, 10.01]
(20.01, 30.01]
(40.01, 50.01]

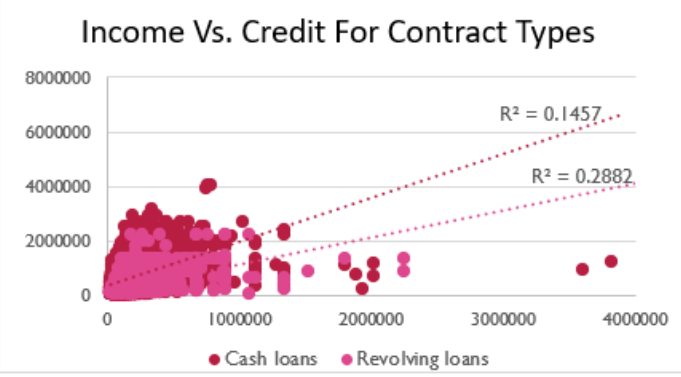

Employed years vs. target- difficulty
900
800
700
600
500
400
300
200
100
0
856
268
66
[0.01, 10.01]
(10.01, 20.01]
(20.01, 30.01]
(30.01, 40.01]

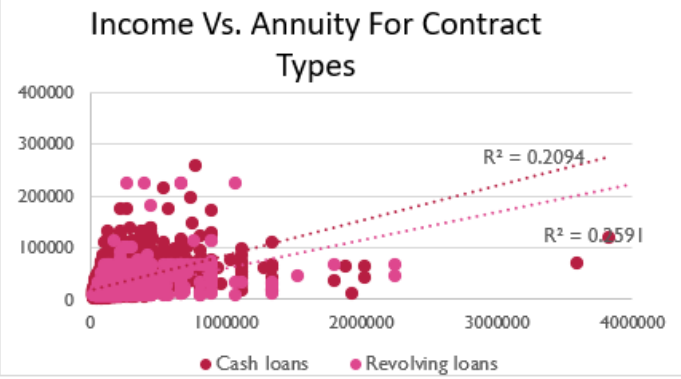

Average income and credit vs. education
1000000.00
800000.00
600000.00
400000.00
200000.00
0.00
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary
Average of AMT_INCOME_TOTAL
Average of AMT_CREDIT

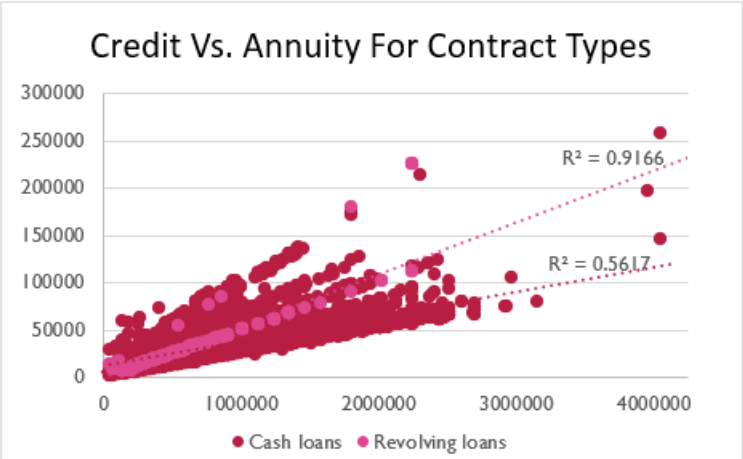

Avg income and credit vs. income type
2000000.00
1500000.00
1000000.00
500000.00
0.00
Businessman
Comm. Assoc.
Maternity leave
Pensioner
State servant
Student
Unemployed
Working
Average of AMT_INCOME_TOTAL
Average of AMT_CREDIT


Avg income and credit vs. family status
800000.00
600000.00
400000.00
200000.00
0.00
Civil marriage
Married
Separated
Single
Widow
Average of AMT_INCOME_TOTAL
Average of AMT_CREDIT


Avg income and credit vs. housing type
800000.00
600000.00
400000.00
200000.00
0.00
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents
Average of AMT_INCOME_TOTAL
Average of AMT_CREDIT


Income Vs. Credit For Contract Types
8000000
6000000
4000000
2000000
0
R² = 0.1457
R² = 0.2882
0
1000000
2000000
3000000
4000000
Cash loans
Revolving loans


Income Vs. Annuity For Contract Types
400000
300000
200000
100000
0
R² = 0.2094
R² = 0.3591
0
1000000
2000000
3000000
4000000
Cash loans
Revolving loans



5. Identify Top Correlations for Different Scenarios

a) Correlations between the following categories: gender, ownership of car, property, count of children, age, employed years, registered for years, amount of income and credit were made using an excel function called **CORREL.**

b) Heat maps were made for two types of loan takers, defaulters and non-defaulters to get insights on how to predict the type of people that can pay off their loans.

NON-DEFAULTERS	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOT	AMT_CREDIT	Age in years	Years employed	registered for years
CODE_GENDER	1.000	-0.354	0.040	-0.061	-0.189	-0.030	0.148	0.076	0.071
FLAG_OWN_CAR	-0.354	1.000	0.004	0.111	0.206	0.110	-0.132	-0.017	-0.085
FLAG_OWN_REALTY	0.040	0.004	1.000	-0.002	-0.002	-0.043	0.118	0.030	0.022
CNT_CHILDREN	-0.061	0.111	-0.002	1.000	0.036	0.006	-0.335	-0.053	-0.183
AMT_INCOME_TOTAL	-0.189	0.206	-0.002	0.036	1.000	0.379	-0.072	0.040	-0.068
AMT_CREDIT	-0.030	0.110	-0.043	0.006	0.379	1.000	0.050	0.089	-0.009
Age in years	0.148	-0.132	0.118	-0.335	-0.072	0.050	1.000	0.242	0.335
Years employed	0.076	-0.017	0.030	-0.053	0.040	0.089	0.242	1.000	0.138
registered for years	0.071	-0.085	0.022	-0.183	-0.068	-0.009	0.335	0.138	1.000

DEFAULTERS	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOT	AMT_CREDIT	Age in years	Years employed	registered for years
CODE_GENDER	1.000	-0.313	0.005	0.010	-0.168	0.007	0.092	0.089	0.057
FLAG_OWN_CAR	-0.313	1.000	0.019	0.054	0.195	0.085	-0.072	-0.013	-0.045
FLAG_OWN_REALTY	0.005	0.019	1.000	0.001	0.025	-0.023	0.103	0.020	0.002
CNT_CHILDREN	0.010	0.054	0.001	1.000	-0.004	0.011	-0.249	-0.030	-0.152
AMT_INCOME_TOTAL	-0.168	0.195	0.025	-0.004	1.000	0.314	-0.007	0.012	-0.029
AMT_CREDIT	0.007	0.085	-0.023	0.011	0.314	1.000	0.137	0.098	0.039
Age in years	0.092	-0.072	0.103	-0.249	-0.007	0.137	1.000	0.285	0.287
Years employed	0.089	-0.013	0.020	-0.030	0.012	0.098	0.285	1.000	0.146
registered for years	0.057	-0.045	0.002	-0.152	-0.029	0.039	0.287	0.146	1.000

ANALYSIS:

1) Loan Distribution & Demographics

    a) Cash loans dominate (91%), while revolving loans account for 9%.
    b) Females make up 66% of applicants, males 34%.
    c) Most applicants are single, and the largest group has no children (34,354), but over 4,000 applicants have eight children, showing a notable presence of large families. Applicants with more than eight children are rare.

2) Assets & Employment

    a) Less than 60% of applicants own a house or car; 30-35% own neither.
    b) The majority of applicants are **working individuals** (26,000+), followed by **commercial associates** (11,543), **pensioners** (8,860), and **state servants** (2,985).

3) Education & Loan Eligibility

    a) **Secondary school graduates** are the largest applicant group (21,831).
    b) Higher education correlates with **higher income and better loan eligibility**, while those with only lower secondary education (355 applicants) are the least likely to apply.

4) Marital & Family Status

    a) **Married applicants** dominate, likely due to family-related financial needs.
    b) **Single individuals** also take loans, while **widows, divorcees, and civil-married individuals** have the fewest applications.
    c) House/apartment ownership is common among applicants, while fewer live in municipal/rented housing.

5) Occupation & Industry

    a) **Business entities** have the highest number of applicants, followed by **unemployed and self-employed individuals** (potentially due to business funding or upskilling needs).
    b) **Labourers** apply for loans more than any other occupation, followed by **core staff**.

6) Age & Employment Trends

    a) **Middle-aged individuals (30-60 years old)** apply for the most loans.
    b) **Younger (20-30) and older (60-70) applicants** are less likely to seek loans.
    c) Those with **0-10 years of employment** file the most applications, suggesting financial instability early in careers.

7) Income & Credit Behavior

   a) A large portion of applicants come from **lower income brackets**, making them a key market for loan products.
   b) Most applicants seek **smaller credit amounts** and **lower annuities**, indicating cautious borrowing or affordability concerns.
   c) **Businessmen** have the highest average income and credit amounts, while **students and unemployed individuals** struggle the most to secure loans.
   d) **Higher education levels** correlate with **higher income and larger loan approvals**.

8) Loan Defaults & Financial Stability

   a) **Working individuals** have the highest number of both **defaulters and non-defaulters**, indicating that while many manage loans well, a significant portion struggle.
   b) **Females have higher non-defaulter rates** compared to males, meaning they repay loans on time more often.
   c) **House/apartment owners** report fewer payment difficulties, suggesting **real estate ownership correlates with financial stability**.
   d) **Car owners also show fewer difficulties**, possibly indicating better financial health.
   e) **Dual-parent households** have higher numbers of both **defaulters and non-defaulters**, suggesting greater financial responsibility.
   f) **Regions rated "2"** have higher loan difficulties and more applicants, possibly due to economic conditions and cost of living.
   g) **Middle-aged applicants (30-40 years)** are the most likely to take loans and have both the highest **non-defaulter** and **defaulter** rates.
   h) **Applicants aged 51-70** have similar loan counts but are **less likely to default**, possibly due to financial stability in later years.

9) Loan Type Insights

   a) **Revolving loans** have a stronger correlation between **income, credit amount, and annuity**, implying stricter lender considerations.
   b) **Cash loans** have more **variability in loan size and repayment terms**, suggesting that factors beyond income influence approvals.
   c) **Revolving loans tend to have more consistent payment structures**, leading to a more predictable relationship between credit and income.

RESULT:

The dataset required cleaning to handle inconsistencies, missing values, and outliers before analysis. Key preprocessing steps included addressing missing data in employment duration and income fields, standardizing categorical variables (e.g., education levels, marital status), and removing extreme outliers in loan amounts and annuities.

After cleaning, exploratory data analysis (EDA) revealed significant trends in loan distribution, applicant demographics, and repayment behavior. Summary statistics and visualizations helped identify key borrower segments, such as working individuals and middle-aged applicants being the most active loan seekers. Feature correlations highlighted relationships between income, credit amount, and default risk, showing that financial stability and asset ownership influence repayment success.

Further analysis included segmentation based on employment type, family structure, and loan purpose to identify high-risk groups. Loan default patterns were examined using grouping techniques and comparisons across demographics. Insights from regional ratings helped assess economic impact on loan performance.

Overall, the data cleaning and analysis process provided a structured approach to understanding loan trends, default risks, and borrower behavior, offering valuable insights for financial decision-making and risk assessment.

IMPACT OF CAR FEATURES ON PRICE & PROFITABILITY

DESCRIPTION:

The automotive industry has undergone significant changes driven by advancements in fuel efficiency, sustainability, and technology. As competition grows and consumer preferences shift, understanding factors that influence car buying decisions is crucial for success. A Data Analyst plays a key role in helping manufacturers optimize pricing and product development to maximize profitability while meeting consumer demand.

This involves analyzing the relationship between a car's features, market segments, and pricing to identify the most desirable and profitable attributes. By applying techniques like regression analysis and market segmentation, manufacturers can create pricing strategies that balance consumer expectations with revenue goals. Identifying key product features for future development further strengthens competitiveness, ensuring long-term profitability and responsiveness to market trends.

THE PROBLEM

A) Data Analysis Tasks: All tasks aim to clean and analyze data thoroughly for accurate, actionable insights.

Objective	Task
1. Car Model Popularity Across Market Categories	Create a pivot table showing car models by market category and popularity scores. Use a combo chart to visualize their relationship between market categories and popularity
2. Engine Power vs. Price Relationship	Create a scatter chart with engine power (x-axis) and price (y-axis), including a trendline to illustrate the relationship.
3. Key Features Influencing Price	Perform regression analysis to identify variables strongly related to car price. Visualize variable importance using a bar chart of coefficient values.
4. Average Car Price by Manufacturer	Create a pivot table for average car prices by manufacturer. Visualize the data using a bar chart or horizontal stacked bar chart.
5. Fuel Efficiency vs. Cylinders	Create a scatter plot showing the number of cylinders (x-axis) and highway MPG (y-axis), adding a trendline to assess the relationship. Calculate the correlation coefficient to quantify the strength and direction of the relationship.

B) Building the Dashboard: The next part of the project involves creating an **Interactive Dashboard** using filters and slicers to answer the following client questions:

Objective	Task
1. Car Prices by Brand and Body Style	Use a stacked column chart to show the distribution of car prices. Calculate total MSRP for each brand and body style using SUMIF or Pivot Tables.
2. Highest and Lowest Average MSRPs by Brand	Use a clustered column chart to compare average MSRPs by brand and body style. Calculate averages using AVERAGEIF or Pivot Tables.
3. Transmission Type's Impact on MSRP	Use a scatter plot to show the relationship between MSRP and transmission type, with symbols for body styles. Calculate averages using AVERAGEIFS or Pivot Tables.
4. Fuel Efficiency by Body Style and Model Year	Use a line chart to display MPG trends over time for body styles. Calculate averages using AVERAGEIFS or Pivot Tables.
5. Horsepower, MPG, and Price by Brand	Use a bubble chart to illustrate the relationship between horsepower, MPG, and price by brand. Color-code brands and label bubbles with car models. Calculate averages using AVERAGEIFS or Pivot Tables.

THE DESIGN:

Dataset Summary: "Car Features and MSRP"

- **Source**: Collected by Cooper Union, a private college in New York City, and hosted on Kaggle.
- **Number of Observations**: 11,914
- **Number of Variables**: 16
- **File Type**: CSV (Comma Separated Values)

Names of Columns	
Make	Number of Doors
Model	Market Category
Year	Vehicle Size
Engine Fuel Type	Vehicle Style
Engine HP	Highway MPG
Engine Cylinders	City MPG
Transmission Type	Popularity
Driven_Wheels	MSRP

B) Data pre-processing, cleaning and error rectification:

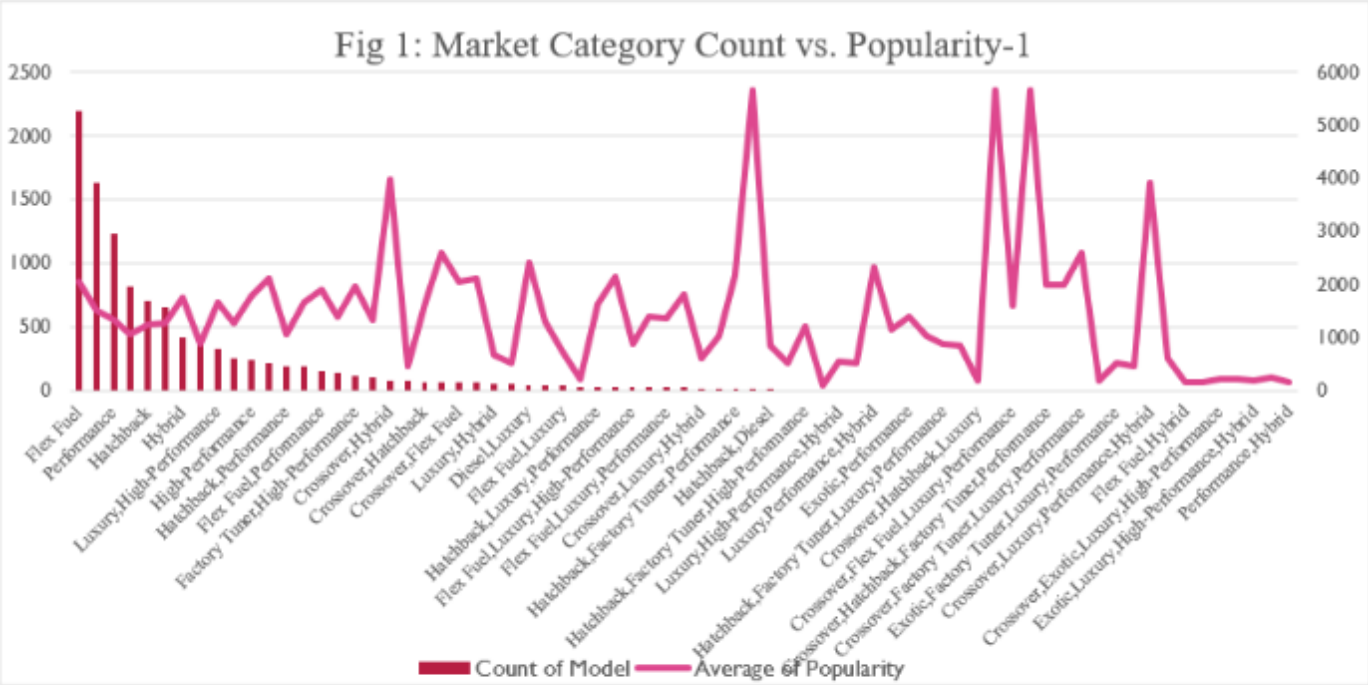
1. Out of 11,914 rows and 16 columns, 715 rows with duplicate data (filtered by movie name) were removed. 11199 rows remained. Out of all columns, Engine HP, Engine Cylinders, Number of Doors, Transmission Type and Market Category had empty or unknown values.
2. Blank values in the 'Number of Doors' column were found in rows with the same Make, Model, and Year values (Tesla Model S, 2016). Other matching rows had '4' in this column, so the blanks were replaced with 4.
3. Similarly, blank values in the same where Make, Model, Vehicle style and Vehicle Size values (Ferrari FF, Large Coupe). Other matching rows had '2' in this column, so the blanks were replaced with 2.
4. Blank values in the 'Engine Cylinders' column were found in where the 'Engine Fuel Type' value was 'electric'. Since electric cars don't have engine cylinders, the blanks were replaced with 0. The other blank values in 'Engine Cylinders' column were filled with the help of the internet by verifying their car make, model and year values.
5. The blank or unknown values in 'Engine HP' and 'Transmission Type' columns were filled with the help of the internet by verifying their car make, model and year values.
6. To fill missing 'Market Category' values, the most common market category was used from rows with the same Make, Vehicle Style, and Vehicle Size values. However, some values remained missing. These remaining values were filled by finding the market category from rows with the same Make, Engine Fuel & Vehicle Style or the same Make, Engine Fuel & Vehicle Size.
7. This was done by using the **IF, MODE, INDEX**, and **MATCH** functions. The **IF** function checked if the cell in the 'Market Category' column was empty. If it was, the **INDEX** function retrieved a value from the 'Market Category' column. The **MODE** function identified the most frequently occurring value from rows where the 'Make, Vehicle Style & Vehicle Size' column matched and 'Market Category' was not empty. The **MATCH** function helped locate these matching values. If the cell was not empty, it remained unchanged.
8. Outliers in Engine HP, Engine Cylinders and MSRP all have market categories as Exotic, Luxury and High-Performance. For popularity, only Ford cars had 5000+ likes. So, these values were unchanged.
9. New column Highway MPG-city MPG was mad to check and correct negative values. Negative values occurred with Hybrid and Electric type cars so they values were kept as it is.
10. One value with the difference of 330 had wrong highway mpg value so checked on internet and replaced with 34.

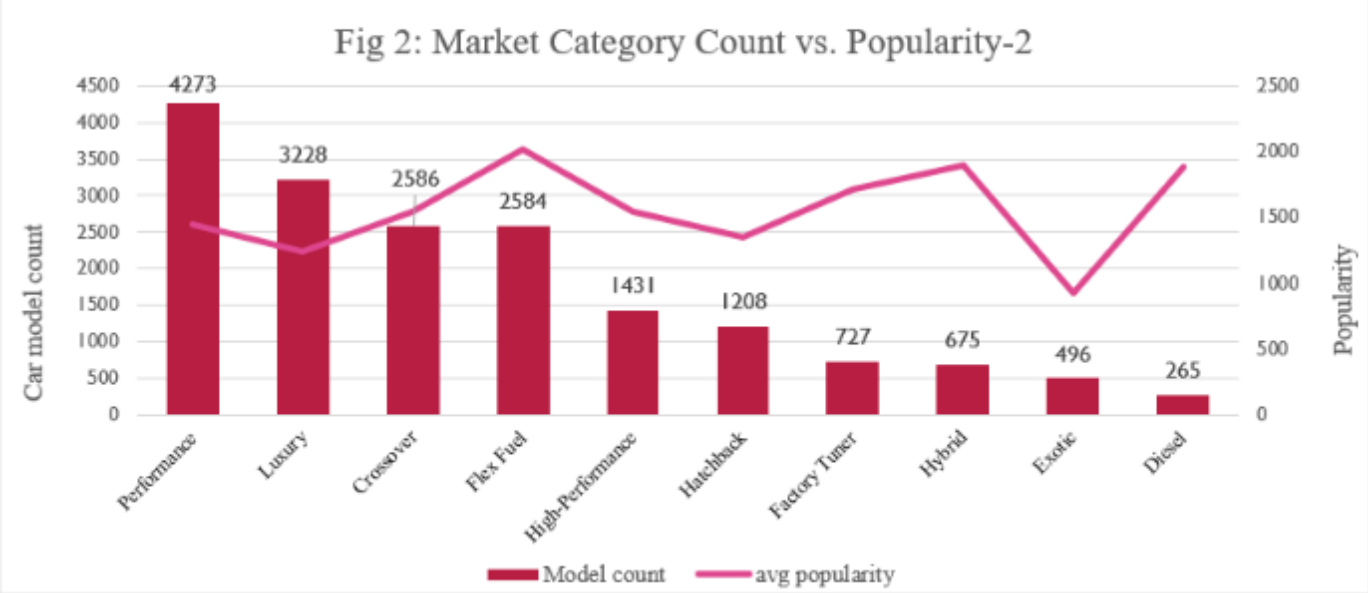
FINDINGS:

A) Data Analysis Tasks:

1. Car Model Popularity Across Market Categories

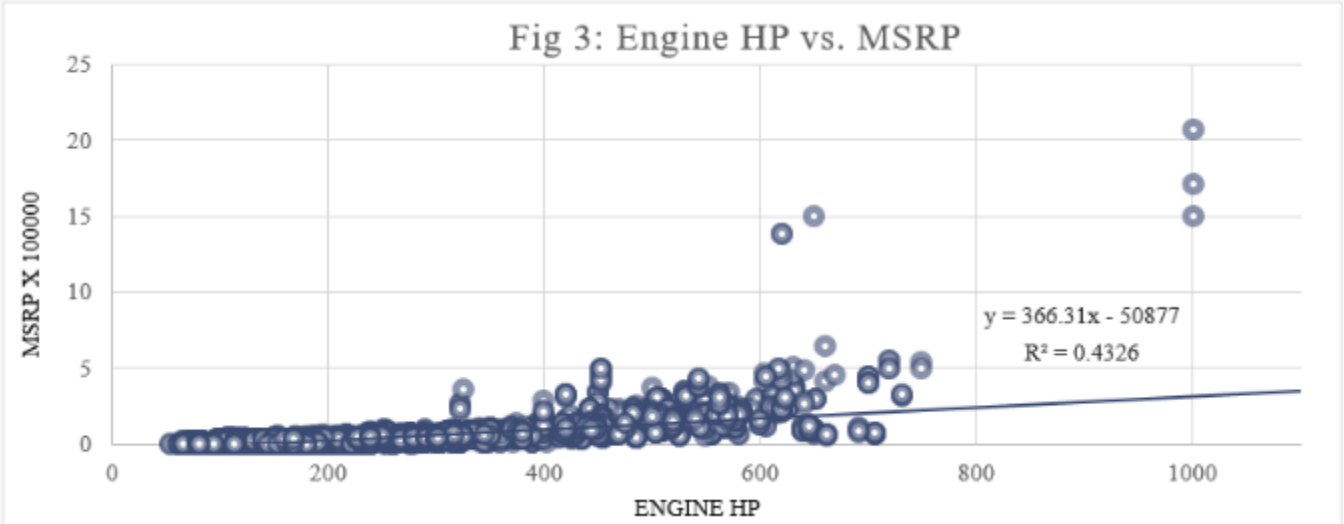
a) A pivot table was created to obtain count and average popularity for market categories. The table was represented as a combination chart showing 'Market Category Count vs. Popularity-1'.
b) The functions **COUNTIF** and **AVERAGEIF** were used to place conditions to only calculate count of car models under each individual market category as well as calculate the average popularity for those market categories.
c) The table was represented as a combination chart showing 'Market Category Count vs. Popularity-2'.





2. Engine Power vs. Price Relationship

a) Engine HP vs. MSRP chart was plotted directly by selecting the two columns and applying scatter plot chart to them. The y-intercept and R-squared value were shown.



3. Key Features Influencing Price

a) Multiple regression analysis was carried on the dependent variable 'MSRP' and numerical independent variables Year, Highway MPG, City MPG, Engine HP/Highway MPG, Popularity and MSRP/Engine HP using Excel's 'Regression' function under the 'Data Analysis' Add-in ToolPak.

b) For non-numerical independent variables (Market Category, Engine Fuel Type and Vehicle Style), 'dummy variables' were used to convert each unique category into a numerical variable. For each of the three variables, one category was eliminated from multiple regression analysis as the reference category.

c) This would give the coefficient values in comparison to the reference categories. The presence of a category in a variable was labeled as 1 and the absence as 0, using the **IF, ISNUMBER** and **SEARCH** functions. For example, under the variable, 'Market Category', 'Crossover' was used as the reference category and others were compared to 'Crossover' after being assigned values (1 or 0).

=IF(ISNUMBER(SEARCH("Hatchback", AS2)), 1, 0)

R	S	T	U	V	W	X	Y	Z	AA
Hatchb	Luxury	Exotic	Factory	Flex Fue	Diesel	Perforn	Hybrid	High-Pe	MSRP
0	1	0	0	0	0	0	1	0	28900
0	1	0	0	0	0	0	1	0	34600
0	1	1	0	0	0	1	1	1	156000
0	1	0	0	0	0	0	0	0	2397
0	1	0	0	0	0	0	0	0	2488

d) The reference category for Vehicle Style was 'Cargo Minivan' and for Engine Fuel Type was 'regular unleaded'. Multiple regression analysis was carried out on each variable.

e) The coefficients were used to create a column chart to compare the effect of variables on MSRP. The coefficients for reference categories were kept as 0.

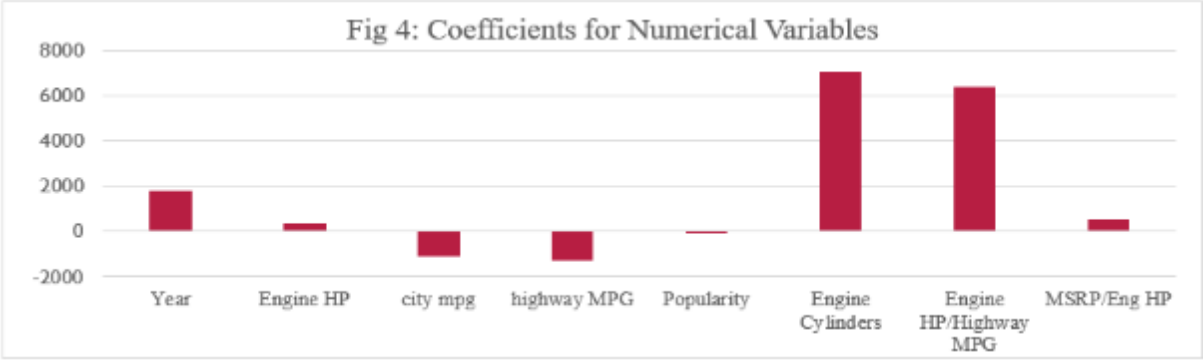


Fig 4: Coefficients for Numerical Variables
8000
6000
4000
2000
0
-2000
Year
Engine HP
city mpg
highway MPG
Popularity
Engine Cylinders
Engine HP/Highway MPG
MSRP/Eng HP


Fig 5: Coefficients for Market Categories
200000
160000
120000
80000
40000
0
-40000
Crossover
Hatchback
Luxury
Exotic
Factory Tuner
Flex Fuel
Diesel
Performance
Hybrid
High-Performance

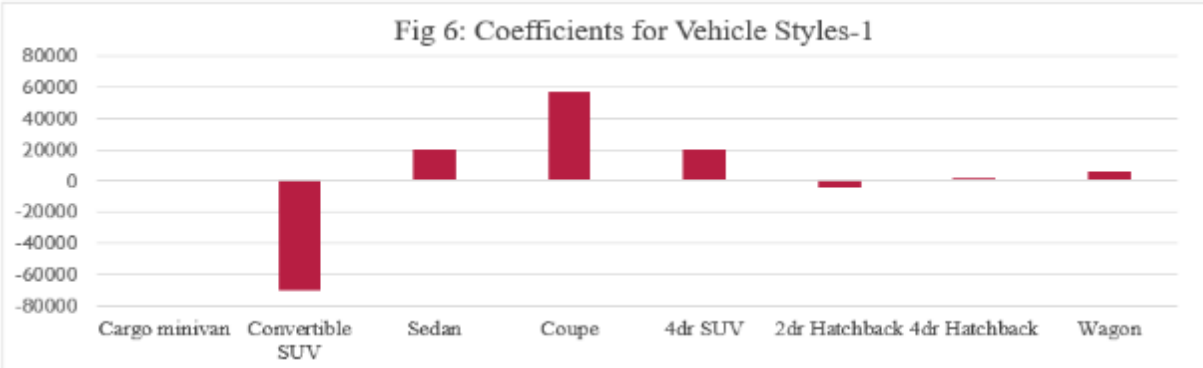


Fig 6: Coefficients for Vehicle Styles-1
80000
60000
40000
20000
0
-20000
-40000
-60000
-80000
Cargo minivan
Convertible SUV
Sedan
Coupe
4dr SUV
2dr Hatchback
4dr Hatchback
Wagon

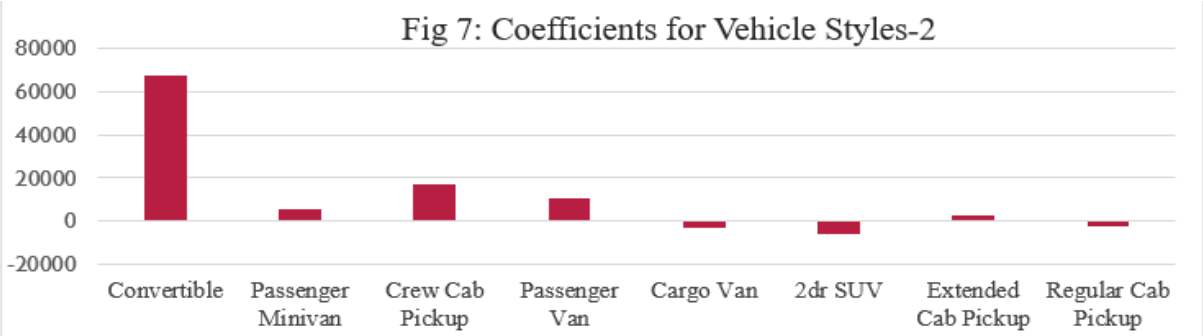


Fig 7: Coefficients for Vehicle Styles-2
80000
60000
40000
20000
0
-20000
Convertible
Passenger Minivan
Crew Cab Pickup
Passenger Van
Cargo Van
2dr SUV
Extended Cab Pickup
Regular Cab Pickup

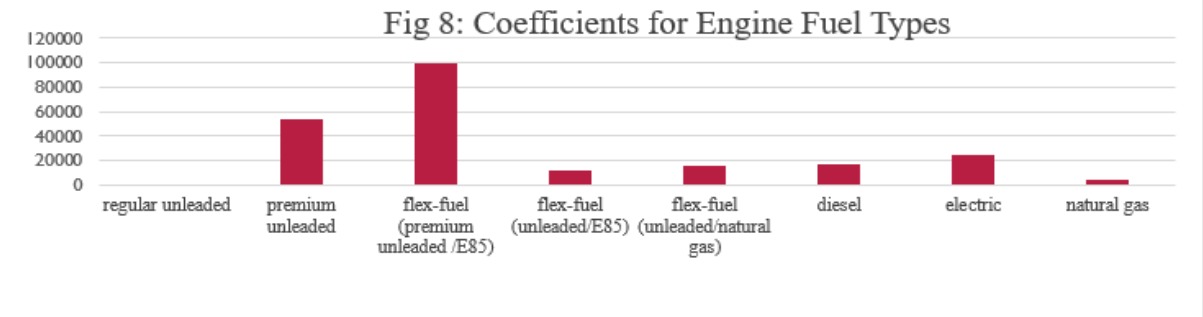
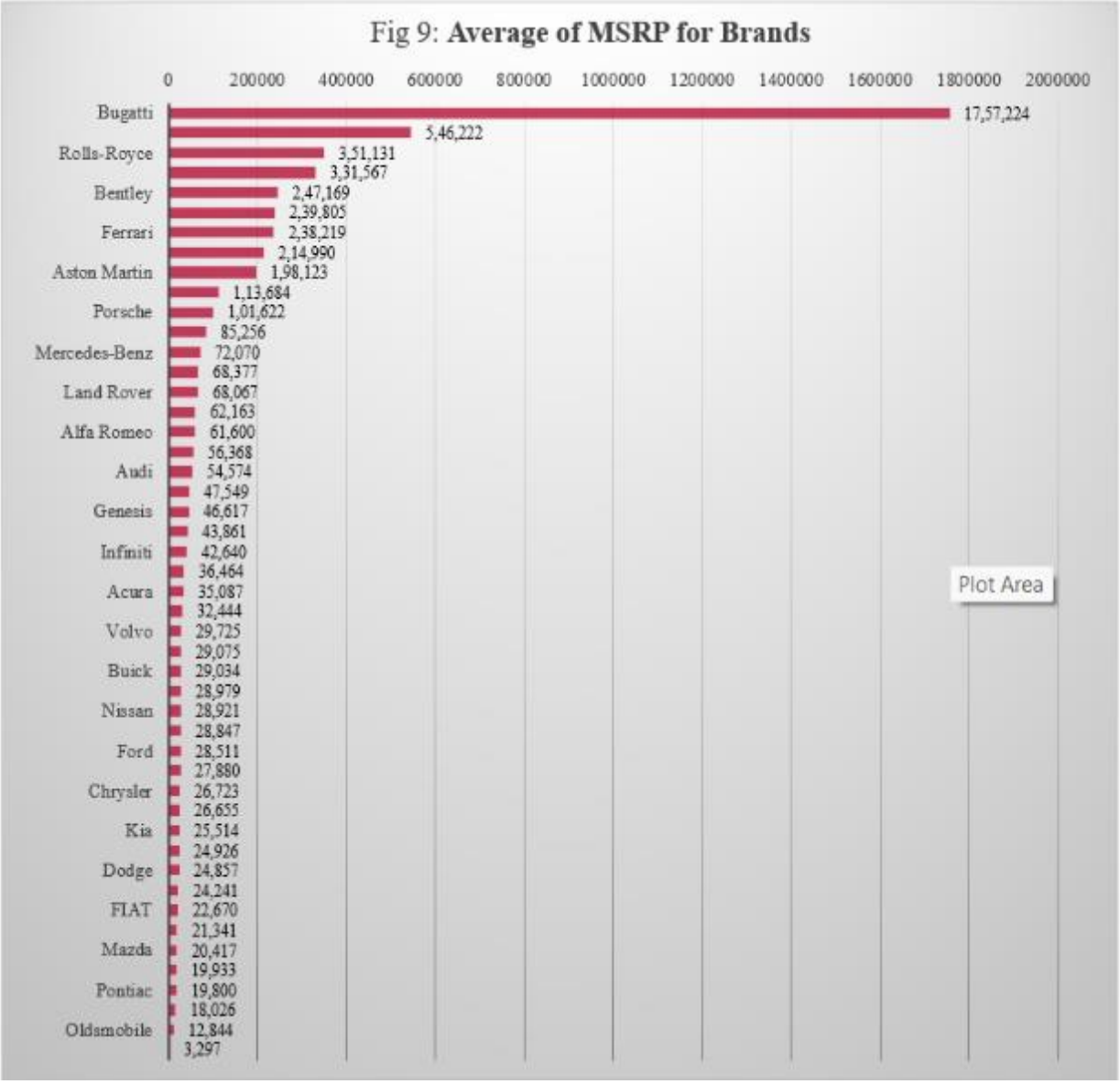


Fig 8: Coefficients for Engine Fuel Types
120000
100000
80000
60000
40000
20000
0
regular unleaded
premium unleaded
flex-fuel (premium unleaded /E85)
flex-fuel (unleaded/E85)
flex-fuel (unleaded/natural gas)
diesel
electric
natural gas

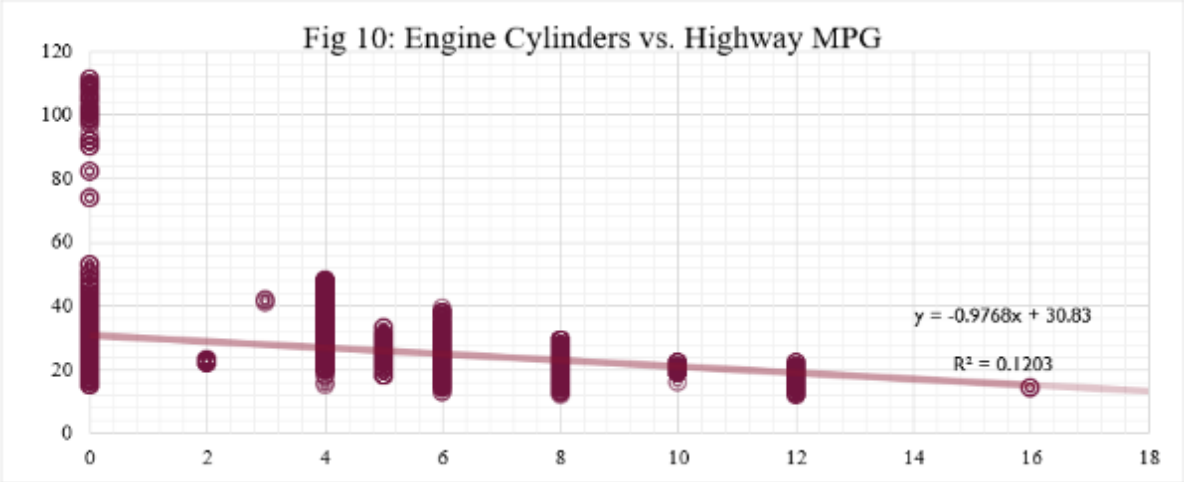
4. Average Car Price by Manufacturer

a) A pivot table was created to obtain count and average MSRP for car brands. The table was represented as a bar chart showing ‘Average of MSRP for Brands’.



5. Fuel Efficiency vs. Cylinders

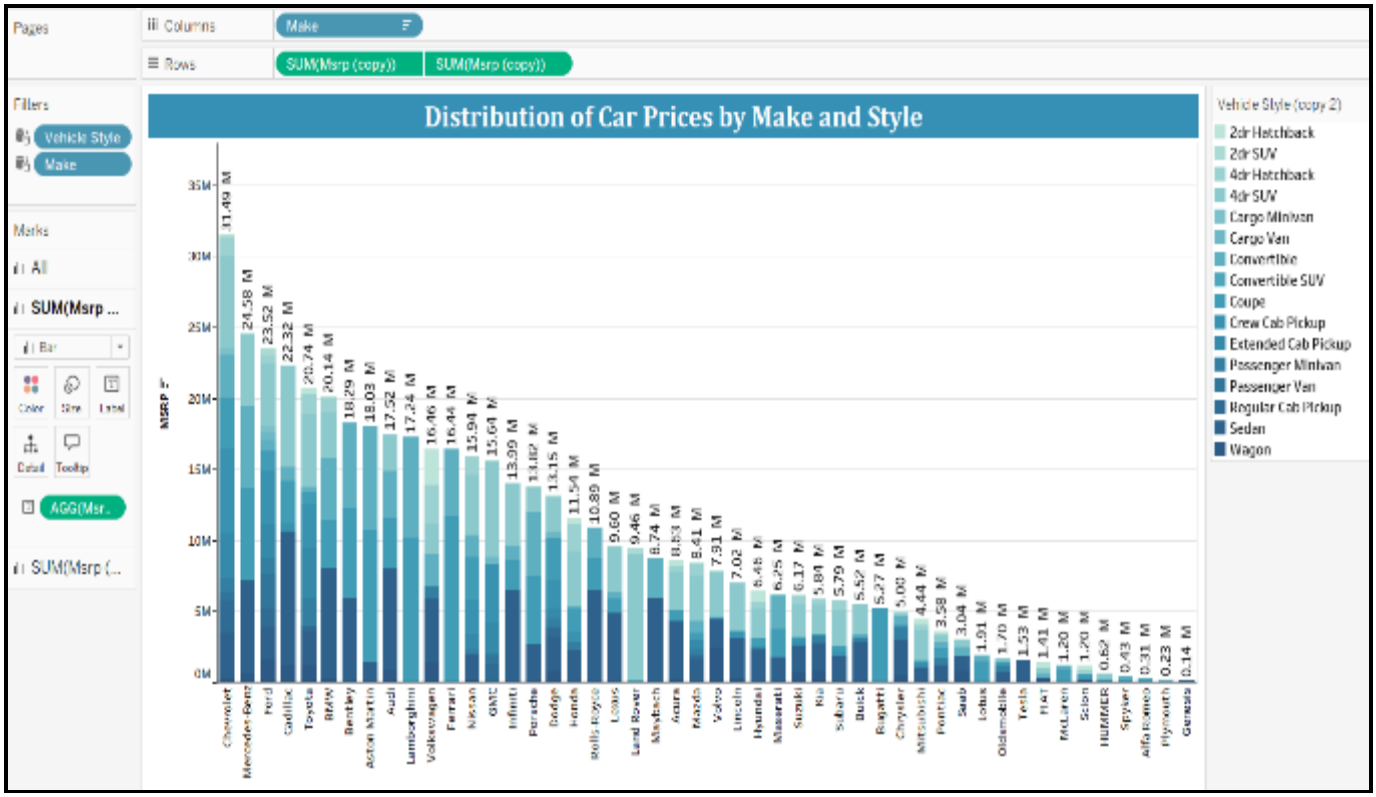
b) Engine Cylinders vs. Highway MPG chart was plotted directly by selecting the two columns and applying scatter plot chart to them. The y-intercept and R-squared value were shown.



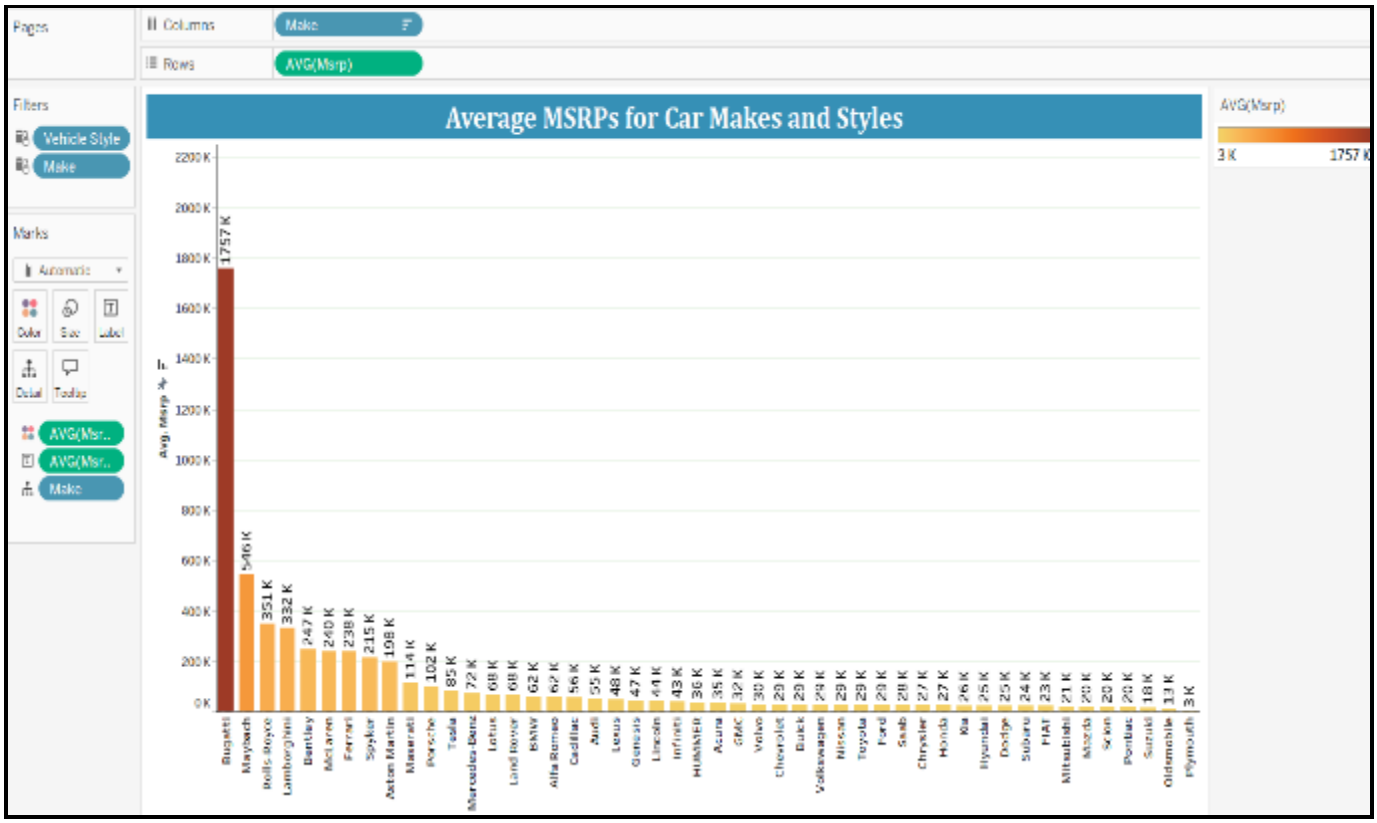
B) Building the Dashboard:

a) The following steps were used to create various charts in Tableau. While the process remains largely the same, a specific example is provided for clarity.
b) The Excel file **'Impact on Car Features Final'** was opened in Tableau Public under the 'Connect' option.
c) Under 'Sheets', the sheet **'Outlier Free Data'** was dragged into the workspace to view the tables (dimensions) in the sheet.
d) Multiple sheets containing graphs were created using the sorted dimensions. Depending on the specified problem, a dimension (categorical data) was placed in the Columns shelf, and a measure (numerical data) was placed in the Rows shelf.
e) Tableau automatically generated a default chart (e.g., bar chart, line chart). The chart type was modified under the 'Show Me' panel when needed.
f) The charts were further customized by dragging fields into the Color**,** Label**,** Tooltip**,** or Filters sections in the Marks card. The axis names, labels, and colors were formatted accordingly.
g) For example, the chart **'Highway MPG vs. Engine HP and Avg. MSRPs'** was created by placing the **average of Engine HP** in the Rows shelf and the **average of Highway MPG** in the Columns shelf.
h) Under the 'Marks' section, the 'Circle' option was selected instead of 'Bar'. The **average of MSRP** was dragged into the Size section, and **Make** was dragged into the Color section.
i) This transformed the chart into a **bubble graph**, where each car make was represented by a colored bubble, and the size of the bubble indicated the **average MSRP** of each car make. Filters were added for **Car Make** and **Vehicle Style**.
j) After creating all the charts, a new dashboard was added by selecting the dashboard tab. The dashboard size was set to 1920 x 1080 pixels. All five charts were arranged within the dashboard, with common filters for **Vehicle Style and Make**. The dashboard was named **'Impact of Car Features on Prices'**.

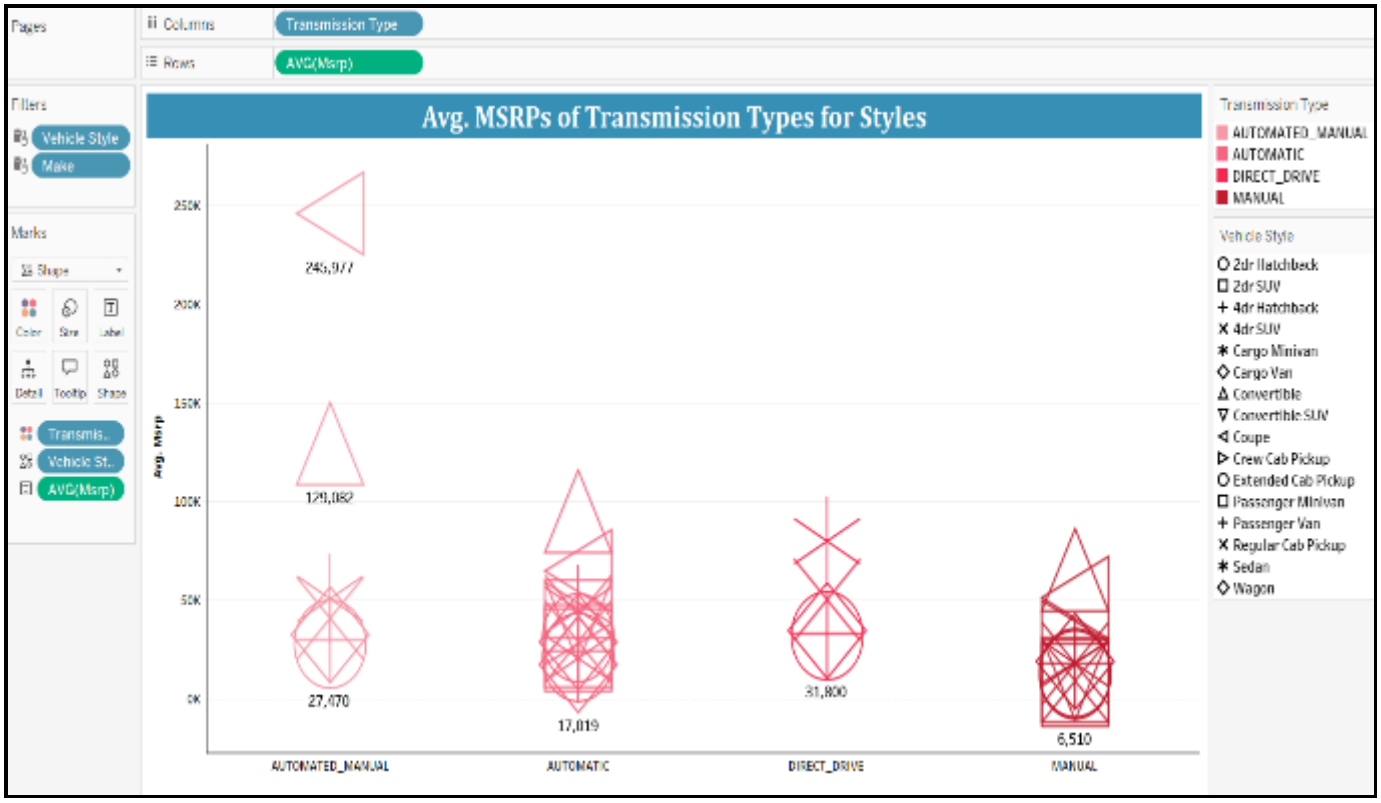
1. Car Prices by Brand and Body Style (Fig 11):



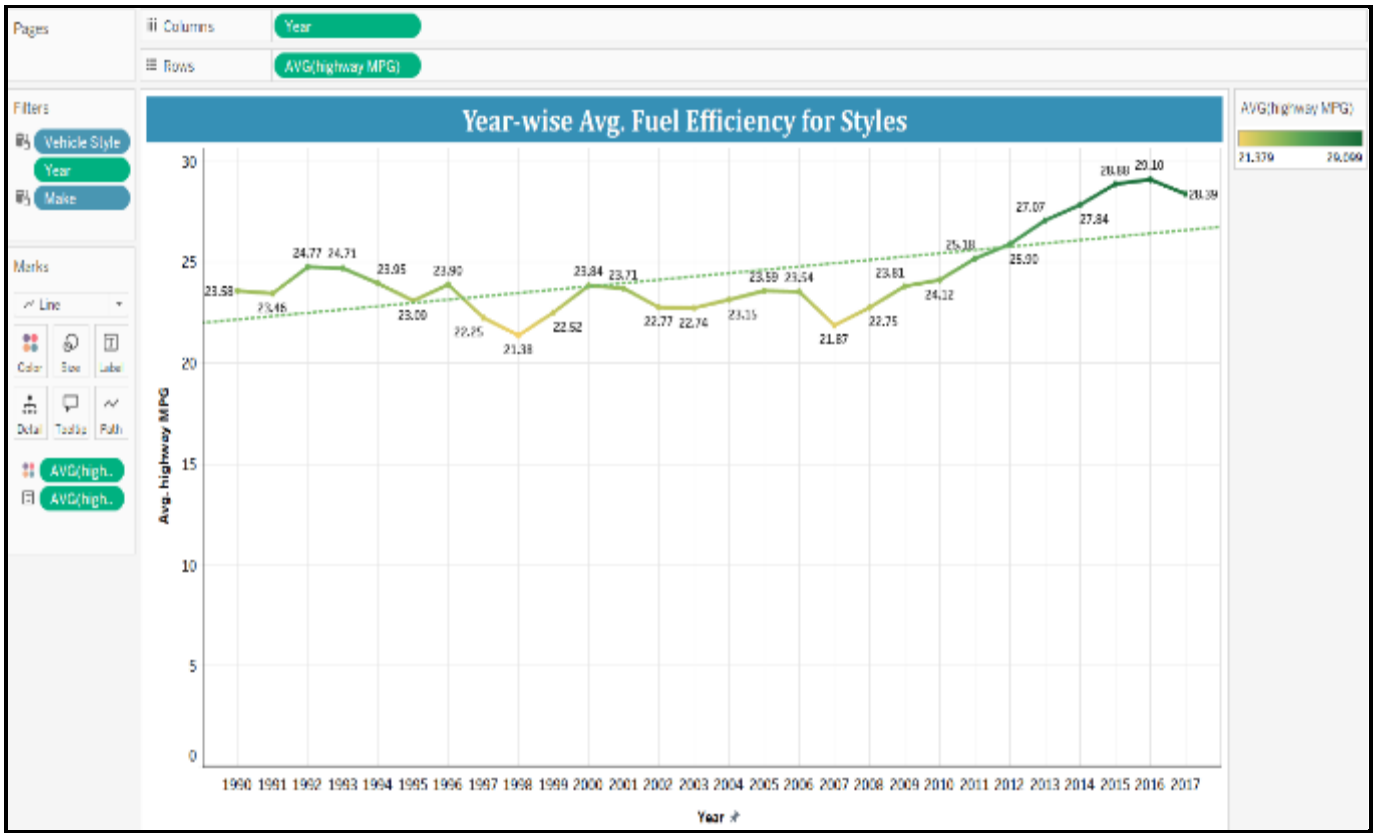
2. Highest and Lowest Average MSRPs by Brand (Fig 12):



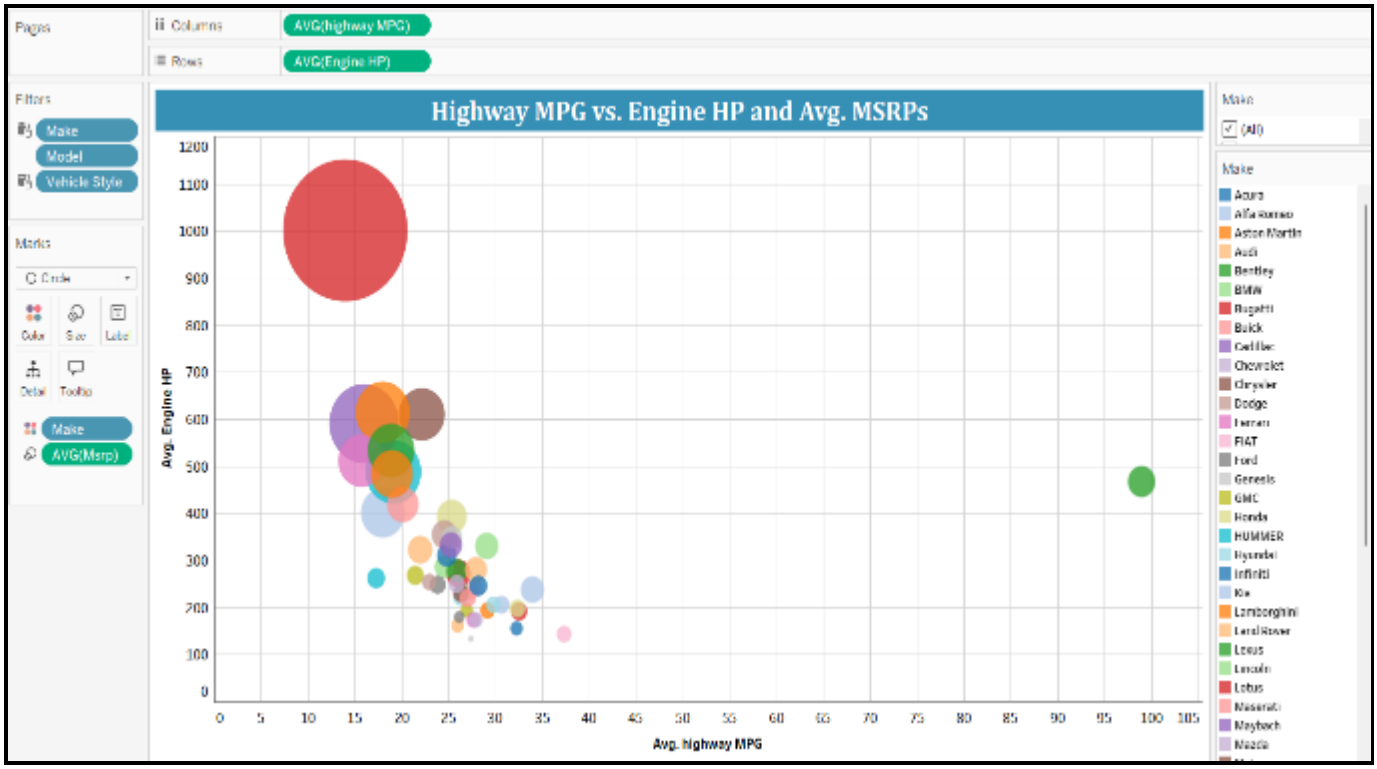
3. Transmission Type's Impact on MSRP (Fig 13):



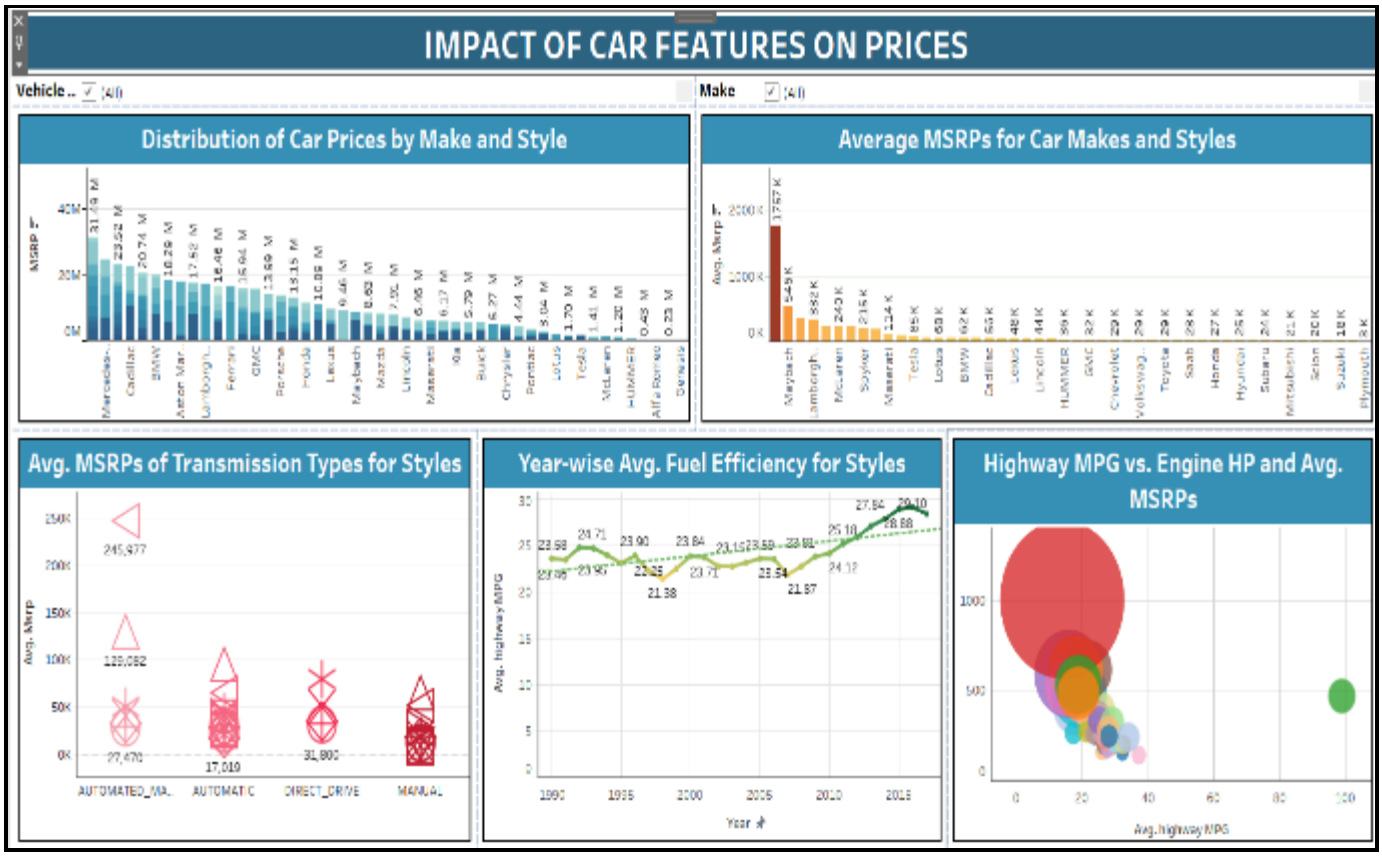
4. Fuel Efficiency by Body Style and Model Year (Fig 14):



5. Fuel Efficiency vs. Cylinders (Fig 15):



The Dashboard:



ANALYSIS:

1. The 5 highest market categories in popularity overall are: 1) Flex Fuel, diesel 2) Hatchback, Flex Fuel 3) Crossover, Flex Fuel, Performance 4) Crossover, Luxury, Performance, Hybrid and 5) Crossover, Hybrid according to Fig 1. Meanwhile Fig 2 with individual market categories shows that Flex Fuel, Diesel and Hybrid are most popular.

2. According to Fig 1, market categories with highest car model count are Flex Fuel, Crossover, Performance, Luxury and Hatchback. From Fig 2, its Luxury, Exotic, Hatchback and Crossover.

3. There is a positive correlation in Fig 3 between engine horsepower (HP) and MSRP with the R^2 value = 0.4326 indicating a moderate correlation. Above 500 HP, MSRPs rise steeply, suggesting that higher horsepower models command premium prices (likely exotic cars or supercars).

4. Most of the data points are clustered between 100 HP to 400 HP, with MSRP values mainly between $0 to $200,000 covering mainstream vehicles like sedans, hatchbacks, and SUVs. In summary, while HP does influence MSRP, the relationship is more significant for high-performance and luxury cars, whereas regular vehicles show a more consistent pricing.

5. The analysis in Fig 5 shows that market categories Luxury, Exotic and High-Performance categories have the highest positive impact on car price, especially Exotic cars, highlighting a strong price premium.

6. Key numerical features (Fig 4) with positive effects include Engine Cylinders, Engine HP/Highway MPG and Year. Highway MPG and City MPG have negative association with MSRP which makes sense as high MPG is associated with economy cars and not luxury/exotic cars.

7. Body styles such as Convertible, 4-door SUV, Sedan, and Coupe also contribute positively, while Wagon, Passenger Van & Minivan, Crew Cab Pickup show moderate price premiums. Notably, Convertible SUVs have a significant negative impact on price (Fig 6 & 7).

8. Alternative Fuel categories have positive coefficients especially Flex-Fuel (premium unleaded/E85) and Premium unleaded as well as Electric cars, reflecting that hybrid or alternative fuel cars are generally priced higher, likely due to the added technology and fuel efficiency benefits (Fig 8).

9. Based on Fig 9 & 12, the car manufacturers have been divided into 4 categories based on their MSRPs, Top-Tier Luxury/ Exotic Brands, Luxury Brands, Mid-Range/Entry-Level Brands and Mainstream/ Budget Brands.

10. **Top-Tier Luxury/ Exotic Brands**: Bugatti has by far the highest average MSRP at over $1.7 million, reflecting its position as an ultra-luxury and high-performance brand. Maybach and Rolls-Royce follow. Lamborghini, Bentley, McLaren, Ferrari, and Spyker also feature high average MSRPs.

11. **Luxury Brands:** Brands like Aston Martin, Maserati, Porsche, Tesla and Mercedes-Benz show high average MSRPs, typically between $70,000 and $200,000.

12. **Mid-Range/ Entry-Level Brands:** Brands like BMW, Land Rover, Audi, Cadillac, Lexus, and Volvo have more moderate average MSRPs, generally around $40,000 to $70,000. These brands provide a balance between luxury and affordability, catering to a wider audience.

13. **Mainstream/Budget Brands:** Mainstream brands such as Chevrolet, Ford, Toyota, Honda, Hyundai, Nissan, and Kia show lower average MSRPs, mostly under $30,000. Suzuki, Oldsmobile, and Plymouth have some of the lowest average MSRPs, which aligns with their reputation for economical and practical vehicles.

14. The trend line in Fig 10, has a slight negative slope, indicating a weak negative correlation (R^2 value is 0.1203) between the number of engine cylinders and highway miles per gallon (MPG). This suggests that vehicles with more cylinders tend to have slightly lower fuel efficiency on highways.

15. Fig 13 shows transmission types alongwith vehicle styles and their average MSRPs. **Automated Manual Transmission** vehicles have the highest average MSRP and possibly high-end models possibly, with prominent contributions from coupes, convertibles, and sedans.

16. **Automatic Transmission** vehicles show versatile price range that appeals to both budget and high-end vehicles. They are available across a wide variety of vehicle styles, including hatchbacks, sedans, SUVs, and trucks, making them a common choice.

17. **Direct Drive** vehicles are primarily associated with sedans and coupes, this transmission type has a moderate average MSRP. May reflect newer or electric vehicle models where direct drive systems are common.

18. **Manual Transmissions** tend to have the lowest average MSRPs. Still see notable contributions from coupes and convertibles, indicating some performance-oriented or enthusiast models.

19. Fig 14 shows vehicle styles and their year-wise average fuel efficiency. **Sedans** show increased fuel efficiency steadily, especially from the mid-2000s onward peaking at ~33 MPG in 2016.

20. For **Coupes** it grew moderately, stabilizing at 25-27 MPG by 2017. **SUVs** started with low MPG but improved significantly from the 2000s, reaching 29 MPG by 2017. **4dr SUVs** experienced variability, with a dip in the late 1990s but steady growth to 27 MPG by 2017 with advancements in fuel efficiency for larger vehicle types.

21. **2-dr Hatchbacks** consistently achieve high fuel efficiency, averaging 30-40 MPG in 2010-2018, with a peak of 45.47 MPG in 2016. **4-dr Hatchbacks** show steady improvement, becoming one of the most fuel-efficient styles by 2018.

22. **Convertibles** have the lowest fuel efficiency, averaging 20-25 MPG, with slight recent improvements. **Wagons** show stable fuel efficiency, mostly between 25-30 MPG, with improvements from 2009 and a peak in 2015. They become more efficient in the 2010s, trailing behind hatchbacks.

23. Fig 15 shows a bubble chart of highway MPG vs. Engine HP for car manufacturers alongwith their average MSRPs. Depending on their performances, the car makes can be divided into three categories.

24. **Low MPG, High HP (Performance vs. Efficiency):** High-performance cars, luxury cars and sports car, are linked to high horsepower, low MPG and higher MSRPs such as Bugatti, Lamborghini, McLaren, Ferrari, Rolls-Royce, Bentley, Aston Martin coupes, convertibles and sedans.

25. **High MPG, Low HP:** Economy-focused vehicles (40-60 MPG, up to 200 HP) from smaller, economy-focused cars such as sedans, convertibles, wagons and hatchbacks of companies like FIAT, Nissan, Chevrolet, Scion, Toyota, etc.

26. **Mid-Range:** Sedans, coupes, convertibles and SUVs (20-40 MPG, 200-400 HP) from brands like Porsche, Cadillac, Volvo, Volkswagen, Audi, BMW, Ford, Subaru, Dodge, Honda, Toyota, etc.

27. Notable Outliers are: Bugatti Coupes with highest horsepower (~1000 HP) and lowest MPG (~15). They're extremely powerful, they aren't the most fuel-efficient. Mercedes-Benz Hatchback are extremely high fuel efficiency (~85 MPG) with balanced power (~200 HP) that is ideal for everyday commuting and family use. Tesla Sedans have high horsepower (~450 HP) and exceptional MPG (~100) due to its electric powertrain.

RESULT:

The conclusions that have been derived in the analysis can be summarized as follows:

Car Feature	Impact
1. Market Categories	Most popular market categories: Flex Fuel, Hybrid and Diesel. Most car model count: Flex Fuel, Crossover, Performance, Luxury, Exotic and Hatchback.
2. High-Performance and Luxury Segments vs. Mainstream Brands	Higher MSRPs: Luxury/exotic segments like Bugatti, Maybach, Rolls-Royce, Lamborghini, and McLaren. Lower average MSRPs brands: Chevrolet, Ford, Toyota, Honda. Lowest average MSRPs (economical and practical vehicles): Suzuki, Oldsmobile, and Plymouth.
3. Fuel Type and Performance	High-performance cars with low MPG and high HP (Bugatti, Lamborghini) as well as alternative fuel categories correlate with higher MSRPs; economy cars with low HP and mid-low range MPG have lower MSRPs.
4. Engine Horsepower (HP)	Higher HP generally increases MSRP, in high-performance and luxury vehicles. Regular cars show stable pricing with small HP changes.
5. Transmission Types	Automated Manual Transmissions have the highest MSRP, followed by Automatic (broad price range), while Manual Transmissions have the lowest MSRPs.
6. Vehicle Type and Fuel Efficiency	Fuel efficiency improved overall since the 1980s; hatchbacks and sedans are most fuel-efficient. Convertibles have the lowest fuel efficiency.
7. Engine Cylinders vs. highway MPG	Weak negative correlation; more cylinders slightly reduce fuel efficiency.

Through this project, I learned how to analyze car market trends, pricing, and performance using Excel and Tableau. By exploring market category popularity and car model distribution, I identified how specific

consumer needs drive demand while posing challenges for manufacturers in balancing cost and performance.

My analysis revealed a moderate correlation between horsepower and MSRP, with higher-performance vehicles commanding premium prices, though other factors like luxury features and alternative fuel types also play a significant role. Regression analysis showed that features such as engine power, fuel type, and body style strongly impact pricing, while utility-focused vehicles tend to have lower price associations.

Using Tableau, I created an interactive dashboard to visualize how various car features influence prices, helping to identify key trends. Additionally, I examined fuel efficiency patterns over time, noting steady improvements, especially in sedans, hatchbacks, and SUVs. Overall, this project strengthened my data analysis skills, particularly in using Excel for calculations and Tableau for insightful data visualization.

ABC CALL VOLUME TREND ANALYSIS

DESCRIPTION:

This project delves into Customer Experience (CX) analytics, focusing on the inbound calling team of a company. You'll work with a 23-day dataset containing details like agent IDs, queue times, call durations, and statuses (answered, abandoned, or transferred).

CX teams play a vital role in analyzing customer feedback, deriving insights, and improving customer interactions. They use tools like Interactive Voice Response (IVR), Robotic Process Automation (RPA), and Predictive Analytics to enhance customer support. Inbound support, the project's focus, handles incoming customer calls, aiming to engage and retain customers, ultimately turning them into loyal advocates for the business.

THE PROBLEM:

You are tasked with analyzing inbound call data from ABC, an insurance company, to address four key issues:

Objective	Task
1. Average Call Duration	Determine the average duration of all incoming calls received by agents. This should be calculated 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.).
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.
4. Night Shift Manpower Planning	Customers also call ABC Insurance Company at night but there are no agents available. 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 given. Propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%.

DESIGN:

A) Dataset Summary for "Call_Volume_Trend_Analysis":

- **Number of Observations**: 117989
- **Number of Variables**: 14
- **File Type**: Excel file

Name of Columns:	
Agent_Name	Time_Bucket
Agent_ID	Duration (hh:mm:ss)
Customer_Phone_No	Call_Seconds (s)
Queue_Time (Secs)	Call_Status
Date	Wrapped _By
Time (hh:mm:ss)	Ringing
Hour start	IVR _Duration

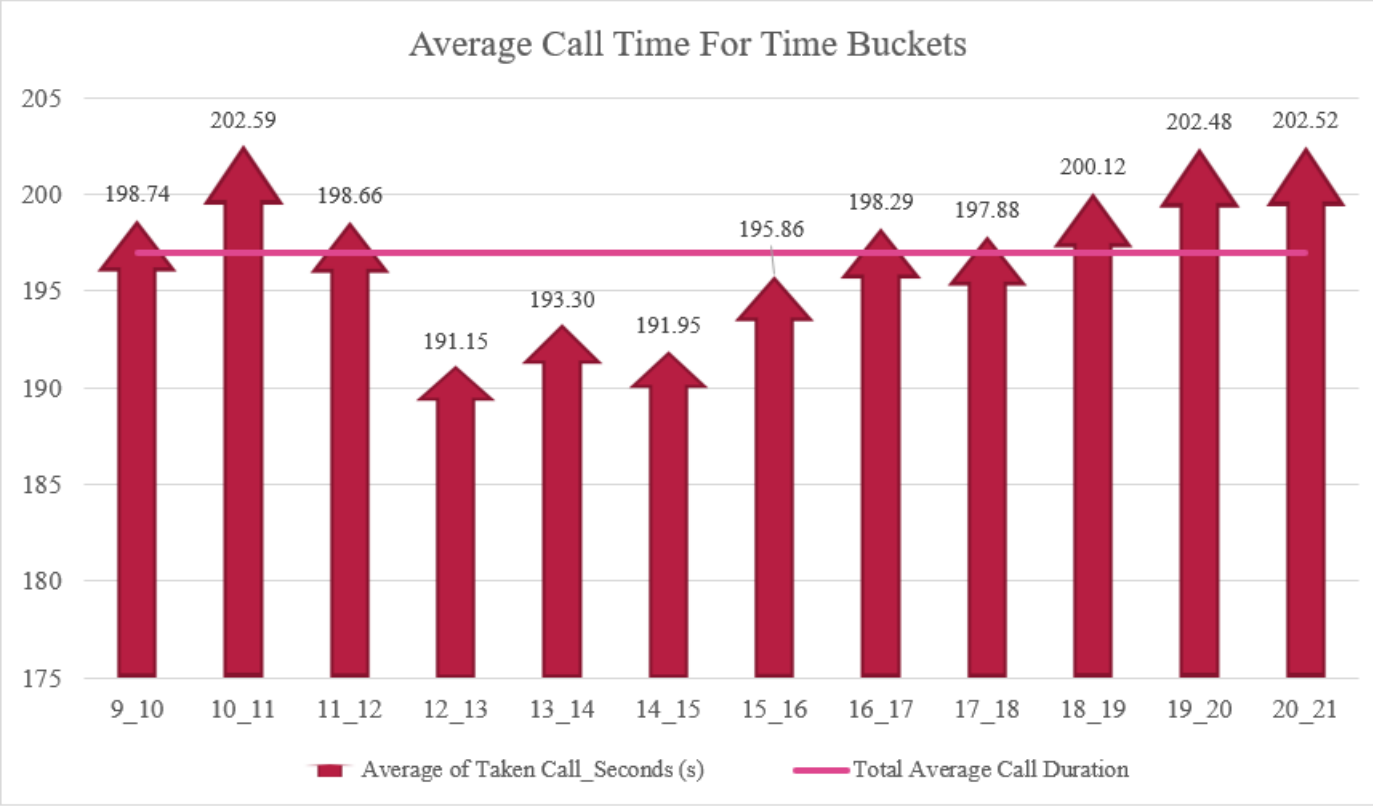
B) Data pre-processing, cleaning and error rectification:

1. The original data has 13 columns and 117989 rows. There were no duplicate rows in the table so all rows were kept. The '#N/A' was replaced with 'N/A' in columns 'Agent_Name' and 'Agent_ID'.
2. The 'Date_&_Time' column has both date and time of calls in the number format. This was resolved by converting the number into date format and generating another column for time of calls in hh:mm:ss format. It was observed that the call data starts from 01-01-2022 to 23-01-2022.
3. The column 'IVR_Duration' is also in number format and is converted to time format (hh:mm:ss). The column 'Wrapped_By' has some blank entries. The other entries in the column are 'Agent' and 'AutoWrapped'.
4. The blank entries in 'Wrapped_By' column that had call status 'abandon' were replaced with 'NA' and those that had call status 'answered' and 'transfer' were replaced with 'Agent'.
5. Following this, pivot tables were made to analyze data and make visualizations to represent the results. These visualizations were then used to gather insights and plan an effective system for manpower allocation for the company.

FINDINGS:

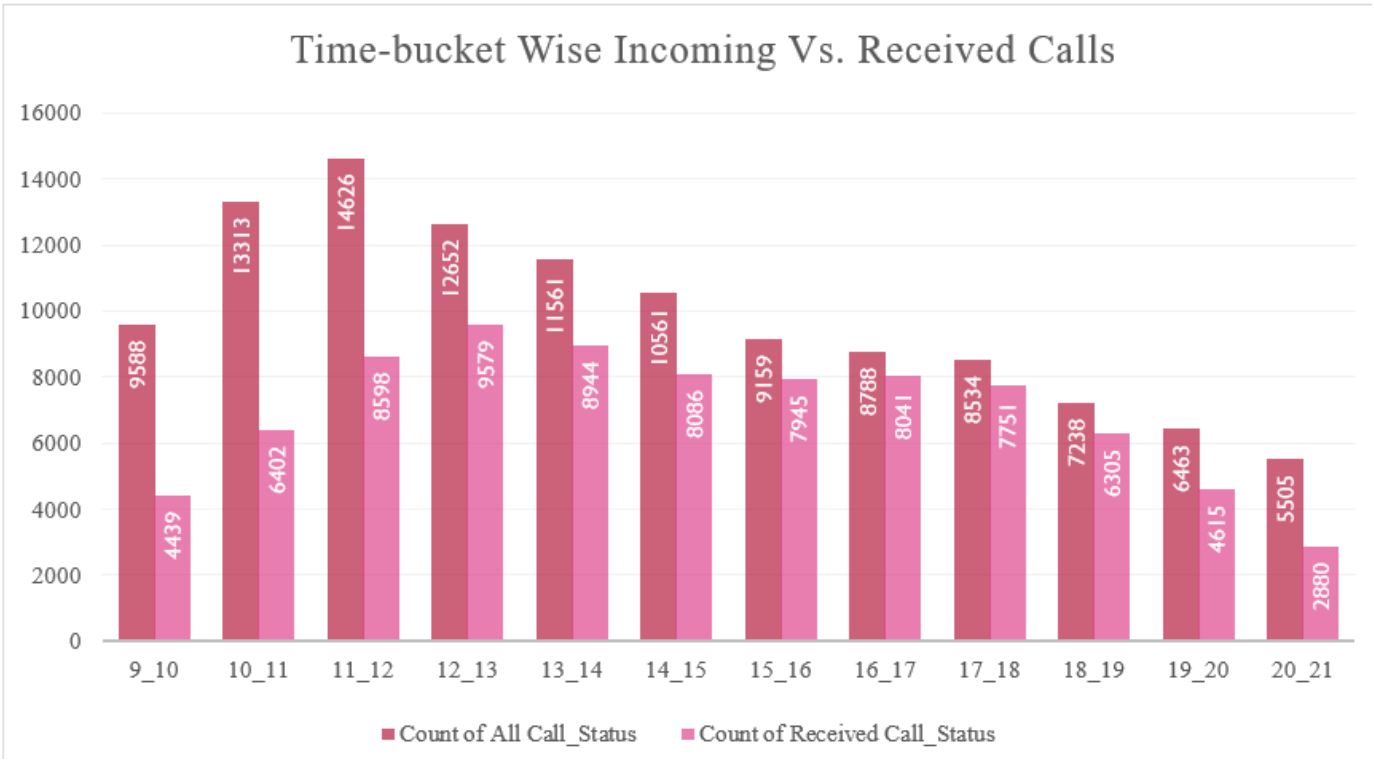
1. Average Call Duration

a) A pivot table was created to obtain average call time for each time bucket. The table was represented as a combination chart showing '**Average Call Time For Time Buckets**'.
b) The overall average was also obtained and displayed in the combination chart as a straight line.



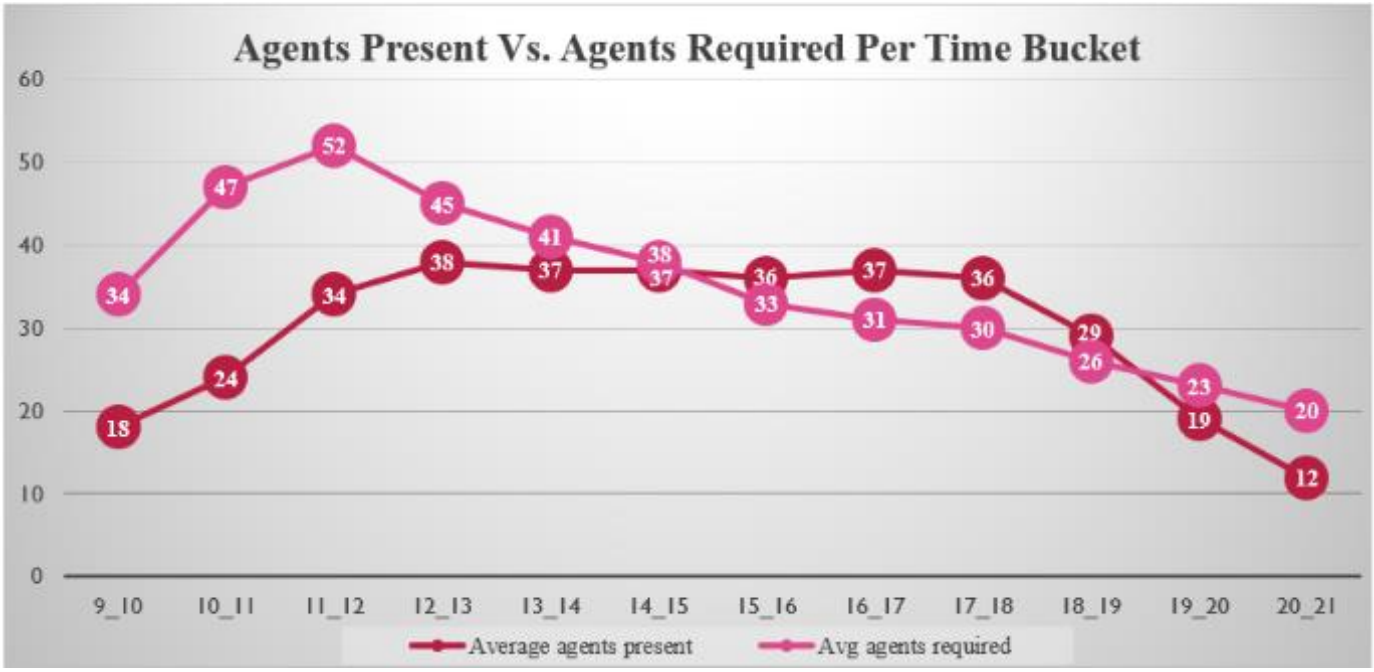
2. Call Volume Analysis

a) A pivot table was created to obtain counts of incoming and received calls for each time bucket. The table was filtered by 'Call_Status' values. The chart was represented as a clustered column chart called **'Time-bucket Wise Incoming vs. Received Calls'**.



3. Manpower Planning

a) Using a pivot table, the count of all calls per time bucket for each day was obtained. Also, a filter was applied on 'Call_Status' to only show count of received calls (answered and transferred calls) per time bucket for each day.
b) These two tables were used to get average calls overall as well as average calls taken for each time bucket. Another table was formed for calculating the number of unique agents that received calls per time bucket for each day.
c) The number of agents was calculated using a combination of **COUNTA**, **UNIQUE** and **FILTER** functions to filter rows that had 'Call_Status' values "answered" and "transfer". Within those rows unique agent IDs in each time bucket were counted for all days.
d) The table was used to get average number of agents present that attend calls per time bucket using **AVERAGE** function.
e) Next, the average calls taken per agent per time bucket was calculated by dividing average calls received by the average no. of agents working per time bucket. A mode of this was taken using **MODE** function which came out to be 11 calls per agent.
f) The average calls received per time bucket was divided by 11 to get the no. of agents required per time bucket. This was taken alongwith previously calculated no. of agents present per time bucket and a combination line chart was derived.



4. Night Shift Manpower Planning

a) **Assumptions for night-shift manpower planning:** An agent works for 6 days a week; On average, each agent takes 4 unplanned leaves per month; An agent's total working hours are 9 hours, out of which 1.5 hours are spent on lunch and snacks in the office. On average, an agent spends 60% of their total actual working hours (i.e., 60% of 7.5 hours) on calls with customers/users. The total number of days in a month is 30.

Distribution of 30 calls coming in night for every 100 calls coming in between 9am - 9pm (i.e. 12 hrs slot)											
9pm- 10pm	10pm - 11pm	11pm- 12am	12am- 1am	1am - 2am	2am - 3am	3am - 4am	4am - 5am	5am - 6am	6am - 7am	7am - 8am	8am - 9am
3	3	2	2	1	1	1	1	3	4	4	5

d) By referring to the above table of the distribution of calls during night shift, the number for each time bucket was multiplied by the sum of all calls for that day and then multiplied by 100. This was done for sum of calls for each to obtain no. of calls during night shift for all days.

e) To count the average no. of calls received per time bucket per weekday, the dates were converted to week days using the function **TEXT** to convert it into "dddd" format. Then, using the **AVERAGEIF** function, the average no. of calls was calculated for each weekday.

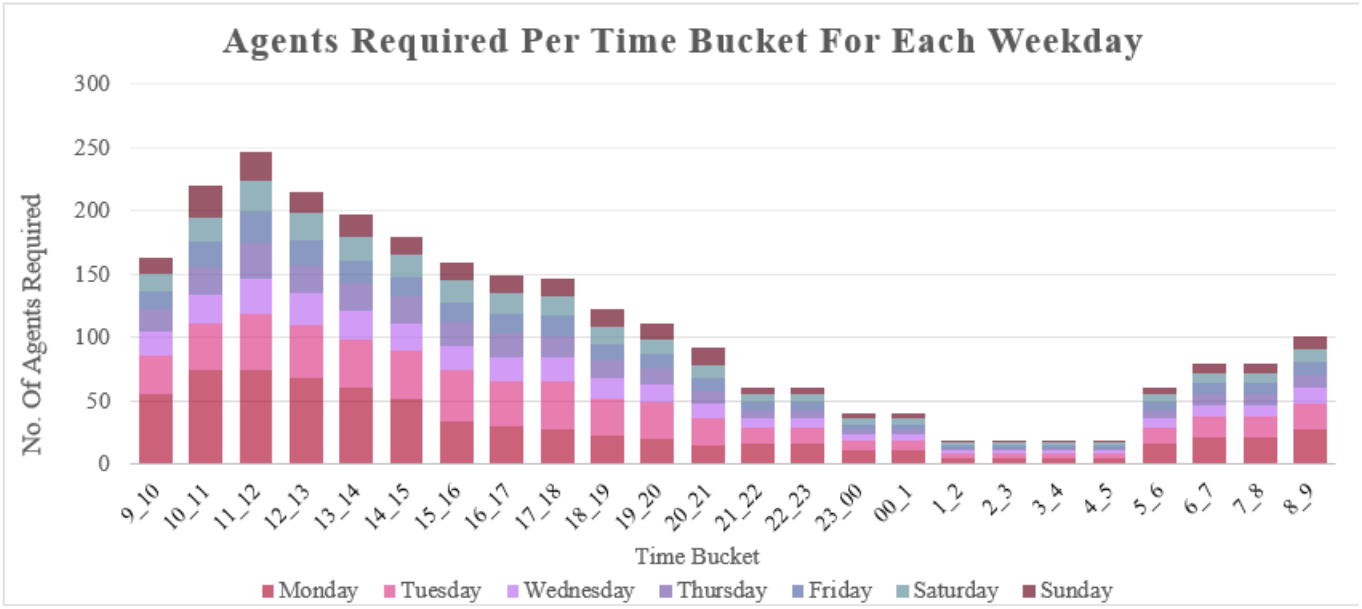
f) This was then multiplied by 90 and divided by 100 to calculate 90% of the calls each weekday. Further the average of calls was also calculated for Monday to Thursday, Friday to Sunday as well as average calls overall in each time bucket.

g) The above-mentioned conditions were taken to calculate the most no. of calls an agent can attend. The result came out to be 18 calls per time bucket.

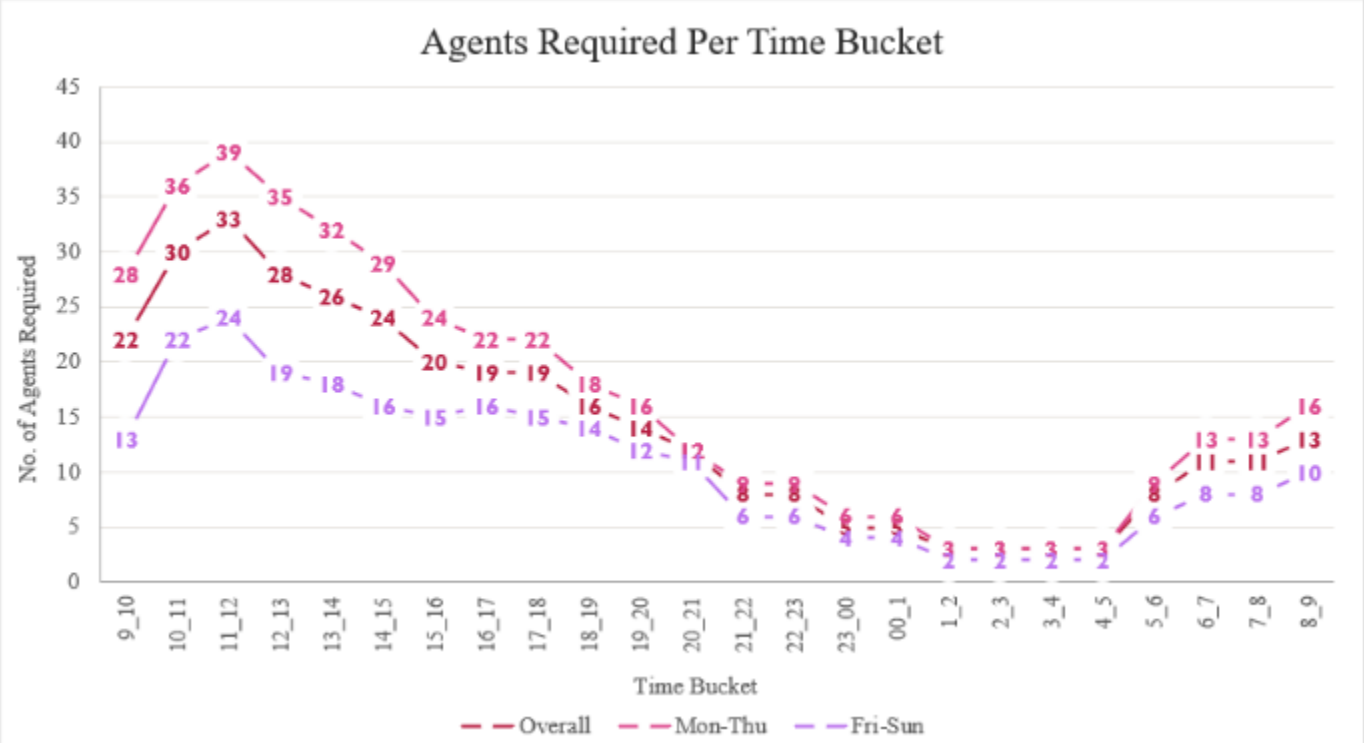
Avg. time spent by agent attending calls in 9 hour shift (i.e., 60% of 7.5 hours)	4.5 hours
Seconds in 4.5 hours (3600 x 4.5)	16200 secs
As earlier calculated, average call time	197 secs
So, average calls an agent can receive per shift (16200/197)	82.2335 calls (approx)
At maximum an agent can take 82.2335/4.5 calls in a time bucket	18.27 or 18 calls

h) Now, the previously calculated table of no. of calls per time bucket per week day was divided by 18 to obtain agents required per time bucket per week day. The average no. of agents was also calculated for Monday to Thursday, Friday to Sunday as well as average agents overall in each time bucket.

i) Finally, the following charts were prepared: a heatmap with agents required per time bucket per week day, a stacked column chart representing the same and a line chart representing no. of agents required (overall, weekdays and weekends).



Time Bucket	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
9_10	55	31	19	18	14	13	13
10_11	75	36	23	22	20	18	26
11_12	75	44	28	27	26	24	22
12_13	68	42	25	22	20	21	17
13_14	61	38	22	22	18	18	18
14_15	51	39	21	21	16	18	14
15_16	34	40	19	20	15	17	14
16_17	30	36	18	19	16	16	14
17_18	28	38	18	17	16	16	14
18_19	23	29	16	14	13	14	14
19_20	20	29	14	13	11	11	13
20_21	15	22	11	10	10	10	14
21_22	16	13	7	7	6	6	6
22_23	16	13	7	7	6	6	6
23_00	11	8	5	4	4	4	4
00_1	11	8	5	4	4	4	4
1_2	5	4	2	2	2	2	2
2_3	5	4	2	2	2	2	2
3_4	5	4	2	2	2	2	2
4_5	5	4	2	2	2	2	2
5_6	16	13	7	7	6	6	6
6_7	21	17	9	9	8	8	8
7_8	21	17	9	9	8	8	8
8_9	27	21	12	11	10	10	10



ANALYSIS:

1. The highest average call time is at 10–11 AM (202.59s), with evening peaks (7–9 PM) around 202.5s. The lowest occurs at 12–1 PM (191.15s) and 2–3 PM (191.95s). Incoming calls peak between 11 AM–1 PM, with 10 AM–12 PM being the busiest (13,313 to 14,626 calls) before declining.

2. Call duration starts high in the morning, drops from 12–3 PM due to simpler calls or higher volumes, then rises in the evening as customers address issues after work. The overall average call time is ~197s, with peak complexity in morning and evening hours.
3. Many incoming calls go unanswered, with significant gaps (5,000–6,000+) from 9 AM–12 PM. Received call trends mirror incoming calls but are lower in the morning.

4. Peak unanswered calls occur at 9 AM–12 PM, suggesting staffing shortages or system limits. Afternoon gaps (12–4 PM) are smaller, indicating a more manageable workload.

5. Late evening still has unanswered calls, but total volume is lower. The 4–5 PM gap is minimal (747 calls), while 8–9 PM has 2,625 unanswered out of 5,505.

6. Agents Required peak from 11 AM–1 PM, then decline. Agents Present peak at 12–1 PM, stabilize in the afternoon, then drop after 5 PM. A major shortage occurs from 9–11 AM. The largest staffing gap is at 10–11 AM (-23 agents). Availability improves at 2–4 PM, sometimes exceeding demand.

7. Surplus agents exist from 4–6 PM (+6 agents), causing inefficiency, while shortages return from 7–9 PM (4–8 agents).

8. Monday–Thursday has the highest agent demand, peaking at 10–11 AM (39 agents). Monday and Tuesday require the most staffing throughout the day. Friday–Sunday sees lower demand, peaking at 11–12 PM (24 agents). Demand drops steadily after 2 PM, reaching a low at 2–5 AM (2–6 agents).

9. Agent demand rises from 5–6 AM, spiking at 8–9 AM (16 agents). Weekdays need more staff, especially from 10 AM–1 PM.

10. Minimal demand (2–6 agents) from 2–5 AM suggests a chance to reduce staffing or automate tasks. Lower weekend demand allows for reallocation to Monday-Tuesday peaks. Flexible or rotating shifts can help manage early morning demand surges (8–10 AM).

RESULT:

1. This project provided me with hands-on experience in analyzing structured datasets related to customer support operations and deriving actionable insights to enhance customer engagement and satisfaction.
2. I identified patterns in customer behavior, such as peak busy hours, and call-handling efficiencies, which contributed to developing insights aimed at improving customer engagement and satisfaction.
3. Additionally, this project taught me the importance of effective call management for customer retention and loyalty. I worked on strategies to reduce queue times and abandoned calls while proposing data-driven solutions to improve call center efficiency and customer satisfaction.
4. Overall, this project helped me practice presenting findings in a way that aligns with business goals and supports better decision-making.

SUMMARY

Through my journey in data analytics, I've developed the ability to extract meaningful insights from data to support smarter, evidence-based decision-making. I've learned how to approach data with a critical eye—identifying patterns, uncovering trends, and solving problems using a mix of analytical thinking and practical tools.

Using **Excel**, I can organize and explore data efficiently. With **SQL**, I retrieve and manipulate information from structured databases. **Power BI** enables me to create compelling dashboards and visual reports that communicate findings clearly. Through **Python**, I've gained the ability to automate processes, perform statistical analyses, and build predictive models.

I've applied these tools across a range of projects that simulate real-world business challenges. I analyzed **social media data** to identify trends in public interest and engagement, explored **which features impact a car's MSRP**, and used **bank loan data to determine factors influencing loan eligibility**. I also worked with **call center performance data** to uncover areas for operational improvement, and studied **HR hiring data** to understand patterns and potential biases in recruitment processes.

Each project has deepened my understanding of how data can guide smarter strategies and decisions, whether in marketing, finance, operations, or human resources. These experiences have prepared me to bring both technical expertise and a problem-solving mindset to any data-driven role.