

VISHALSINGH RAVINDRANATHSINGH



Portfolio

DATA ANALYTICS



PROFESSIONAL BACKGROUND

CURRENTLY I AM IN THIRDED YEAR
YEAR PURSUING B.TECH-CSE. I HAVE
SECURED 7.5 CGPA (TILL 4TH SEM) AND
HAVE SEVERAL SKILLS
INCLUDING DATA ANALYSIS, PYTHON.

I AM GLAD TO TELL THAT I HAVE WORK
WITH TRAINITY COURSE AND LEARNA
ALL THE DATA ANALYSIS TOOL .I
REALLY APRICATE TRAINITY TO
PROVIDE SUCH CONTENT.

I HAD WORK WITH DIFFRENT
PROJECTS WHICH ARE GIVEN IN THE
TRANIITIY COURSE .

TABLE OF CONTENT

1. DATA ANALYTICS PROCESS
2. INSTAGRAM USER ANALYTICS
3. OPERATION ANALYTICS AND
INVESTIGATING METRIC SPIKE
4. HIRING PROCESS ANALYTICS
5. IMDB MOVIE ANALYSIS
6. BANK LOAN CASE STUDY
7. ANALYZING THE IMPACT OF CAR
FEATURES ON PRICE AND PROFITABILITY
8. ABC CALL VOLUME TREND ANALYSIS

1. DATA ANALYTICS PROCESS

THE PROBLEM

YOUR TASK IS TO GIVE THE EXAMPLE(S) OF SUCH A REAL-LIFE SITUATION WHERE WE USE DATA ANALYTICS AND LINK IT WITH THE DATA ANALYTICS PROCESS. YOU CAN PREPARE A PPT/PDF ON A REAL-LIFE SCENARIO EXPLAINING IT WITH THE ABOVE PROCESS (PLAN, PREPARE, PROCESS, ANALYZE, SHARE, ACT) AND SUBMIT IT AS PART OF THIS TASK.

SOLUTION

PLAN: I WANT TO BUY GROCERY FOR THE WHOLE MONTH.

PREPARE: I WILL BE LOOKING HOW MUCH IT IS GOING TO COST ME TO BUY THE GROCERY FOR THE ONE WHOLE MONTH. AND WHAT IS THE BUGGET FOR THE WHOLE MONTH.

PROCESS: I WILL GO KITCHEN AND LOOK FOR WHAT ARE THE NECESSARY THING THAT I WANT TO BUY AND WHAT ARE THE THINGS THAT I CAN SKIP TO BUY.

ANALYZE: OBVIOUSLY I WAN'T BUY THINGS THAT ARE NOT USED IN KITCHEN AND I WILL GOING TO LOOK FOR THE EXPIRE DATES OF THE PRODUCT SO IT CAN BE USED FOR THE LONG TIME .

SHARE: I WILL SHARE THE LIST OF THE ITEAMS THAT I NEED TO BUY TO THE SHOPEKIPPER OR IF I VISITE ANY GROCERY MALL/STORE I WILL WALK AROUND AND LOOK FOR THE ITEAMS ACCORDING TO THE LIST SO I DON'T SKIP ANY PRODUCT.

ACT: AFTER SHARING THE LIST OR WALKING IN THE MARKET AND COLLECTING THE ITEAM FINALLY I BUY IT WITHIN THE BUGET .

2.INSTAGRAM USER ANALYTICS

I have been provided the dataset which contains the usage of instagram with different users and their interaction with the instagram . I have to perform the analysis on mysql workbench.

THE PROBLEM

1.Loyal User Reward:

Your Task: Identify the five oldest users on Instagram from the provided database.

2.Inactive User Engagement:

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

3.Contest Winner Declaration:

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

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

5. Ad Campaign Launch: Your Task: Determine the day of the week when most users register on Instagram. Provide insights on when to schedule an ad campaign.

SOLUTION

FINDING 1

```
1      -- The oldest user of the instagram
2 •  select created_at , username
3   from users
4   order by created_at asc
5   limit 5;
6
7
```

Result Grid		Filter Rows:	Export:	Wrap Cell Co
	created_at	username		
▶	2016-05-06 00:14:21	Darby_Herzog		
	2016-05-06 13:04:30	Emilio_Bernier52		
	2016-05-08 01:30:41	Elenor88		
	2016-05-09 17:30:22	Nicole71		
	2016-05-14 07:56:26	Jordyn.Jacobson2		

WE CAN CLEARLY SEE THE 5 OLDEST USER OF THE INSTAGRAM

FINDING 2

```
select username  
from users  
where id not in (select user_id from photos);
```

username
Aniya_Hackett
Kassandra_Homenick
Jadyn81
Rocio33
Maxwell.Halvorson
Tierra.Trantow
Pearl7
Ollie_Ledner37
Mckenna17
David.Osinski47
Morgan.Kassulke
Linnea59
Duane60
Julien_Schmidt
Mike.Auer39
Franco_Keebler64
Nia_Haag
Hulda.Macejkovic
Leslie67
Janelle.Nikolaus81
Darby_Herzog
Esther.Zulauf61

Hulda.Macejkovic
Leslie67
Janelle.Nikolaus81
Darby_Herzog
Esther.Zulauf61
Bartholome.Bernhard
Jessyca_West
Esmeralda.Mraz57
Bethany20

WE CAN CLEARLY SEE THIS THE PERSON WHO HAVE NEVER POSTED ON INSTAGRAM

FINDING 3

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

	username	photo_id	like_count
▶	Zack_Kemmer93	145	48

FROM THE ALL MEMBER OF
THE INSTAGRAM
ZACK_KEMMER93 HAS GOT
THE MOST LIKED ON SINGLE
PHOTO AND WON THE
CONTEST.

FINDING 4

- ```
SELECT tags.tag_name, COUNT(photo_tags.tag_id) AS usage_count
FROM tags
JOIN photo_tags ON tags.id = photo_tags.tag_id
GROUP BY tags.tag_name
ORDER BY usage_count DESC
LIMIT 5;
```

|   | tag_name | usage_count |
|---|----------|-------------|
| ▶ | smile    | 59          |
|   | beach    | 42          |
|   | party    | 39          |
|   | fun      | 38          |
|   | concert  | 24          |

THE ABOVE ARE THE MOST USED HASHTAG ON THE INSTAGRAM

# FINDING 5

```
| ⚡ ⚡ 🔎 ⏴ | ✎ | ✓ ✖ | ⚡ | Limit to 50000 rows | ⭐ | 🖌️ 🔎 | 1 | ↻
SELECT DAYNAME(created_at) AS day_of_week, COUNT(*) AS registrations
FROM users
GROUP BY day_of_week
ORDER BY registrations DESC
LIMIT 2;
```

|   | day_of_week | registrations |
|---|-------------|---------------|
| ▶ | Thursday    | 16            |
|   | Sunday      | 16            |

**FROM THE WHOLE WEEK  
THURSDAY AND SUNDAY  
ARE THE DAY TO LAUNCH AD.**

## CONCLUSION

I WOULD HAVE ANALYSE THAT EACH TASK HAS ITS OWN VALUE WITHIT. THE MOST INSIGHT THAT COME OUT IS THAT SUNDAY AND THURSDAY ARE THE DAY TO LAUNCH ANY AD. AND IF WE DO PARTNERSHIP WITH ZACK\_KEMMER93 IT WILL BE BENEFICIAL AS IT GOT MOST LIKE ON THE SINGLE PHOTO. AND WE CAN TAKE HELP OF THE OLDEST USER TO KNOW MORE ABOUT THE INSTAGRAM AND FEATURES FROM EARLIER TILL KNOW.

### 3. OPERATION ANALYTICS AND INVESTIGATING METRIC SPIKE

#### IN THIS I HAVE BEEN PROVIDED 2 DATASET CASE STUDY 1: JOB DATA ANALYSIS

WORKING WITH FOLLOWING COLUMNS:

- JOB\_ID: UNIQUE IDENTIFIER OF JOBS
- ACTOR\_ID: UNIQUE IDENTIFIER OF ACTOR
- EVENT: THE TYPE OF EVENT (DECISION/SKIP/TRANSFER).
- LANGUAGE: THE LANGUAGE OF THE CONTENT
- TIME\_SPENT: TIME SPENT TO REVIEW THE JOB IN SECONDS.
- ORG: THE ORGANIZATION OF THE ACTOR
- DS: THE DATE IN THE FORMAT YYYY/MM/DD (STORED AS TEXT).

## **CASE STUDY 2: INVESTIGATING METRIC SPIKE**

WORKING WITH FOLLOWING COLUMNS

- 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.

**ALL THE ANALYSE HAVE BEEN DONE ON  
MYSQL WORK BENCH**

# **CASE STUDY 1: JOB DATA ANALYSIS**

## **TASK TO PERFORM**

- **JOBS REVIEWED OVER TIME:**
  - OBJECTIVE: CALCULATE THE NUMBER OF JOBS REVIEWED PER HOUR FOR EACH DAY IN NOVEMBER 2020.
  - YOUR TASK: WRITE AN SQL QUERY TO CALCULATE THE NUMBER OF JOBS REVIEWED PER HOUR FOR EACH DAY IN NOVEMBER 2020.
- **THROUGHPUT ANALYSIS:**
  - OBJECTIVE: CALCULATE THE 7-DAY ROLLING AVERAGE OF THROUGHPUT (NUMBER OF EVENTS PER SECOND).
  - YOUR TASK: WRITE AN SQL QUERY TO CALCULATE THE 7-DAY ROLLING AVERAGE OF THROUGHPUT. ADDITIONALLY, EXPLAIN WHETHER YOU PREFER USING THE DAILY METRIC OR THE 7-DAY ROLLING AVERAGE FOR THROUGHPUT, AND WHY.

- **LANGUAGE SHARE ANALYSIS:**
  - OBJECTIVE: CALCULATE THE PERCENTAGE SHARE OF EACH LANGUAGE IN THE LAST 30 DAYS.
  - YOUR TASK: WRITE AN SQL QUERY TO CALCULATE THE PERCENTAGE SHARE OF EACH LANGUAGE OVER THE LAST 30 DAYS.
- **DUPLICATE ROWS DETECTION:**
  - OBJECTIVE: IDENTIFY DUPLICATE ROWS IN THE DATA.
  - YOUR TASK: WRITE AN SQL QUERY TO DISPLAY DUPLICATE ROWS FROM THE JOB\_DATA TABLE.

# CASE STUDY 2: INVESTIGATING METRIC SPIKE TASK TO PERFOEM

- **WEEKLY USER ENGAGEMENT:**

- OBJECTIVE: MEASURE THE ACTIVENESS OF USERS ON A WEEKLY BASIS.
- YOUR TASK: WRITE AN SQL QUERY TO CALCULATE THE WEEKLY USER ENGAGEMENT.

- **USER GROWTH ANALYSIS:**

- OBJECTIVE: ANALYZE THE GROWTH OF USERS OVER TIME FOR A PRODUCT.
- YOUR TASK: WRITE AN SQL QUERY TO CALCULATE THE USER GROWTH FOR THE PRODUCT.

- **WEEKLY RETENTION ANALYSIS:**

- OBJECTIVE: ANALYZE THE RETENTION OF USERS ON A WEEKLY BASIS AFTER SIGNING UP FOR A PRODUCT.
- YOUR TASK: WRITE AN SQL QUERY TO CALCULATE THE WEEKLY RETENTION OF USERS BASED ON THEIR SIGN-UP COHORT.

- **WEEKLY ENGAGEMENT PER DEVICE:**
  - OBJECTIVE: MEASURE THE ACTIVENESS OF USERS ON A WEEKLY BASIS PER DEVICE.
  - YOUR TASK: WRITE AN SQL QUERY TO CALCULATE THE WEEKLY ENGAGEMENT PER DEVICE.
- **EMAIL ENGAGEMENT ANALYSIS:**
  - OBJECTIVE: ANALYZE HOW USERS ARE ENGAGING WITH THE EMAIL SERVICE.
  - YOUR TASK: WRITE AN SQL QUERY TO CALCULATE THE EMAIL ENGAGEMENT METRICS.

# CASE STUDY 1: JOB DATA ANALYSIS

## FINDING 1

**SELECT**

**ds AS review\_date,**  
**COUNT(job\_id) AS jobs\_reviewed**

**FROM**

**job\_data**

**WHERE**

**ds BETWEEN '11/01/2020' AND '11/30/2020'**

**GROUP BY**

**review\_date**

**ORDER BY**

**review\_date;**

|            |   |
|------------|---|
| 11/25/2020 | 1 |
| 11/26/2020 | 1 |
| 11/27/2020 | 1 |
| 11/28/2020 | 2 |
| 11/29/2020 | 1 |
| 11/30/2020 | 2 |

**WE CAN SEE THE  
JOB REVIEWED  
IN NOVEMBER**

## FINDING 2

```
WITH daily_counts AS (
 SELECT
 ds AS review_date,
 COUNT(job_id) AS event_count
 FROM
 job_data
 GROUP BY
 review_date
),
throughput AS (
 SELECT
 review_date,
 event_count,
 CAST(event_count AS FLOAT) / CAST(86400 AS FLOAT) AS throughput_per_sec
 FROM
 daily_counts
)
SELECT
 AVG(throughput_per_sec) OVER (
 ORDER BY review_date
 ROWS BETWEEN 6 PRECEDING AND CURRENT ROW
) AS rolling_avg_7_day_throughput
FROM
 throughput
ORDER BY
 review_date;
```

| review_date | throughput_per_sec      | rolling_avg_7_day_throughput |
|-------------|-------------------------|------------------------------|
| 11/25/2020  | 0.000011574074074074073 | 0.000011574074074074073      |
| 11/26/2020  | 0.000011574074074074073 | 0.000011574074074074073      |
| 11/27/2020  | 0.000011574074074074073 | 0.000011574074074074073      |
| 11/28/2020  | 0.000023148148148148147 | 0.000014467592592592593      |
| 11/29/2020  | 0.000011574074074074073 | 0.00001388888888888889       |
| 11/30/2020  | 0.000023148148148148147 | 0.0000154320987654321        |

- WE CAN CLEARLY SEE THE 7 DAYS ROLLING, I WOULD CHOOSE A 7-DAY ROLLING AVERAGE BECAUSE IT REDUCES DAILY VARIATIONS BY CALCULATING THE AVERAGE OVER A FULL WEEK. THIS APPROACH HELPS HIGHLIGHT OVERALL TRENDS MORE CLEARLY, MAKING IT EASIER TO SPOT CONSISTENT SHIFTS IN PERFORMANCE WITHOUT BEING MISLED BY UNUSUAL SINGLE-DAY SPIKES OR DROPS.

## FINDING 3

```
SELECT COUNT(*) as record_in_last_30days
FROM job_data
WHERE STR_TO_DATE(ds, '%m/%d/%Y') >= DATE_SUB(CURDATE(), INTERVAL 30 DAY);
```

|   | record_in_last_30days |
|---|-----------------------|
| ▶ | 0                     |

- SINCE THERE ARE NO RECORDS FROM THE PAST 30 DAYS, WE DON'T HAVE ANY DATA TO ANALYZE THE DISTRIBUTION OF LANGUAGES USED DURING THIS PERIOD. THIS LACK OF RECENT DATA PREVENTS US FROM ASSESSING LANGUAGE TRENDS OVER THE LAST MONTH.

## FINDING 4

```
SELECT ds, job_id, actor_id, `event`, `language`, time_spent, org, COUNT(*) as duplicate_count
FROM job_data
GROUP BY ds, job_id, actor_id, `event`, `language`, time_spent, org
HAVING COUNT(*) > 1;
```

| ds | job_id | actor_id | event | language | time_spent | org | duplicate_count |
|----|--------|----------|-------|----------|------------|-----|-----------------|
|    |        |          |       |          |            |     |                 |

**THERE ARE NO DUPLICATE ROWS IN THE GIVEN DATA SET**

# CASE STUDY 2: INVESTIGATING METRIC SPIKE

## TASK TO PERFOEM

### FINDING 1

```
WITH user_activity AS (
 SELECT
 user_id,
 YEARWEEK(occurred_at, 1) AS activity_week
 FROM
 events
 GROUP BY
 user_id,
 YEARWEEK(occurred_at, 1)
UNION ALL
 SELECT
 user_id,
 YEARWEEK(occurred_at, 1) AS activity_week
 FROM
 email_events
 GROUP BY
 user_id,
```

```
 user_id,
 YEARWEEK(occurred_at, 1) AS activity_week
FROM
 email_events
GROUP BY
 user_id,
 YEARWEEK(occurred_at, 1)
,
⑥ weekly_active_users AS (
 SELECT
 ua.activity_week,
 COUNT(DISTINCT ua.user_id) AS active_users
 FROM
 user_activity ua
 JOIN
 users u ON ua.user_id = u.user_id
 WHERE
 u.activated_at IS NOT NULL
```

```

 u.activated_at IS NOT NULL
 GROUP BY
 ua.activity_week
)
SELECT
 activity_week,
 active_users
FROM
 weekly_active_users
ORDER BY
 activity_week DESC;

```

| activity_week | active_users |
|---------------|--------------|
| 201435        | 4577         |
| 201434        | 4458         |
| 201433        | 4320         |
| 201432        | 4178         |
| 201431        | 4122         |
| 201430        | 3987         |
| 201429        | 3880         |
| 201428        | 3767         |
| 201427        | 3656         |
| 201426        | 3556         |
| 201425        | 3468         |
| 201424        | 3355         |
| 201423        | 3264         |
| 201422        | 3153         |
| 201421        | 3082         |
| 201420        | 3004         |
| 201419        | 2916         |
| 201418        | 1408         |

FROM THE ANALYSIS ABOVE, WE CAN CONCLUDE THAT THE 35TH WEEK OF 2014 HAD THE HIGHEST NUMBER OF ACTIVE USERS, WITH OTHER WEEKS SHOWING LOWER LEVELS OF USER ACTIVITY IN COMPARISON.

## FINDING 2

```
WITH MonthlyGrowth AS (
 SELECT
 DATE_FORMAT(created_at, '%Y-%m') AS month,
 COUNT(user_id) AS new_users
 FROM
 users
 GROUP BY
 month
),
GrowthCalculation AS (
 SELECT
 month,
 new_users,
 LAG(new_users) OVER (ORDER BY month) AS previous_month_users
 FROM
 MonthlyGrowth
)
SELECT
 month,
 new_users,
 LAG(new_users) OVER (ORDER BY month) AS previous_month_users
FROM
 MonthlyGrowth
)
SELECT
 month,
 new_users,
 previous_month_users,
 new_users - COALESCE(previous_month_users, 0) AS growth,
 CASE
 WHEN previous_month_users = 0 THEN NULL
 ELSE (new_users - previous_month_users) * 100.0 / previous_month_users
 END AS growth_percentage
FROM
 GrowthCalculation
ORDER BY
 month;
```

| month   | new_users | previous_month_users | growth | growth_percentage |
|---------|-----------|----------------------|--------|-------------------|
| 2013-01 | 160       | NULL                 | 160    | NULL              |
| 2013-02 | 160       | 160                  | 0      | 0.00000           |
| 2013-03 | 150       | 160                  | -10    | -6.25000          |
| 2013-04 | 181       | 150                  | 31     | 20.66667          |
| 2013-05 | 214       | 181                  | 33     | 18.23204          |
| 2013-06 | 213       | 214                  | -1     | -0.46729          |
| 2013-07 | 284       | 213                  | 71     | 33.33333          |
| 2013-08 | 316       | 284                  | 32     | 11.26761          |
| 2013-09 | 330       | 316                  | 14     | 4.43038           |
| 2013-10 | 390       | 330                  | 60     | 18.18182          |
| 2013-11 | 399       | 390                  | 9      | 2.30769           |
| 2013-12 | 486       | 399                  | 87     | 21.80451          |
| 2014-01 | 552       | 486                  | 66     | 13.58025          |
| 2014-02 | 525       | 552                  | -27    | -4.89130          |
| 2014-03 | 615       | 525                  | 90     | 17.14286          |
| 2014-04 | 726       | 615                  | 111    | 18.04878          |
| 2014-05 | 779       | 726                  | 53     | 7.30028           |
| 2014-06 | 873       | 779                  | 94     | 12.06675          |
| 2014-07 | 997       | 873                  | 124    | 14.20389          |
| 2014-08 | 1031      | 997                  | 34     | 3.41023           |

**THE ANALYSIS SHOWS THAT IN JANUARY 2013, WE GAINED 160 NEW USERS, WHICH WAS THE HIGHEST COUNT FOR THAT PERIOD. LATER, IN JULY 2014, WE SAW THE LARGEST GROWTH IN NEW USERS, WITH 124 ADDED, MARKING AN INCREASE OF 14.20%.**

# FINDING 3

```
COUNT(DISTINCT us.user_id) AS total_signups,
COUNT(DISTINCT wa.user_id) AS retained_users,
CASE
 WHEN COUNT(DISTINCT us.user_id) = 0 THEN 0
 ELSE COUNT(DISTINCT wa.user_id) * 100.0 / COUNT(DISTINCT us.user_id)
END AS retention_rate
FROM
 UserSignups us
LEFT JOIN
 WeeklyActivity wa ON us.user_id = wa.user_id
 AND wa.active_date >= us.signup_date
 AND wa.active_date < DATE_ADD(us.signup_date, INTERVAL 4 WEEK)
GROUP BY
 signup_week
ORDER BY
 signup_week;
```

```
WITH UserSignups AS (
 SELECT
 user_id,
 created_at AS signup_date
 FROM
 users
),
WeeklyActivity AS (
 SELECT
 user_id,
 created_at AS active_date
 FROM
 users
 WHERE
 created_at IS NOT NULL
)
SELECT
 DATE_FORMAT(us.signup_date, '%Y-%u') AS signup_week,
 COUNT(DISTINCT us.user_id) AS total_signups,
```

|         |     |     |           |         |     |     |           |
|---------|-----|-----|-----------|---------|-----|-----|-----------|
| 2014-13 | 151 | 151 | 100.00000 | 2013-44 | 92  | 92  | 100.00000 |
| 2014-14 | 161 | 161 | 100.00000 | 2013-45 | 97  | 97  | 100.00000 |
| 2014-15 | 166 | 166 | 100.00000 | 2013-46 | 94  | 94  | 100.00000 |
| 2014-16 | 165 | 165 | 100.00000 | 2013-47 | 82  | 82  | 100.00000 |
| 2014-17 | 176 | 176 | 100.00000 | 2013-48 | 103 | 103 | 100.00000 |
| 2014-18 | 172 | 172 | 100.00000 | 2013-49 | 96  | 96  | 100.00000 |
| 2014-19 | 160 | 160 | 100.00000 | 2013-50 | 117 | 117 | 100.00000 |
| 2014-20 | 186 | 186 | 100.00000 | 2013-51 | 123 | 123 | 100.00000 |
| 2014-21 | 177 | 177 | 100.00000 | 2013-52 | 104 | 104 | 100.00000 |
| 2014-22 | 186 | 186 | 100.00000 | 2013-53 | 41  | 41  | 100.00000 |
| 2014-23 | 197 | 197 | 100.00000 | 2014-01 | 91  | 91  | 100.00000 |
| 2014-24 | 198 | 198 | 100.00000 | 2014-02 | 122 | 122 | 100.00000 |
| 2014-25 | 222 | 222 | 100.00000 | 2014-03 | 112 | 112 | 100.00000 |
| 2014-26 | 210 | 210 | 100.00000 | 2014-04 | 113 | 113 | 100.00000 |
| 2014-27 | 199 | 199 | 100.00000 | 2014-05 | 130 | 130 | 100.00000 |
| 2014-28 | 223 | 223 | 100.00000 | 2014-06 | 132 | 132 | 100.00000 |
| 2014-29 | 215 | 215 | 100.00000 | 2014-07 | 135 | 135 | 100.00000 |
| 2014-30 | 228 | 228 | 100.00000 | 2014-08 | 127 | 127 | 100.00000 |
| 2014-31 | 234 | 234 | 100.00000 | 2014-09 | 127 | 127 | 100.00000 |
| 2014-32 | 189 | 189 | 100.00000 | 2014-10 | 135 | 135 | 100.00000 |
| 2014-33 | 250 | 250 | 100.00000 | 2014-11 | 152 | 152 | 100.00000 |
| 2014-34 | 259 | 259 | 100.00000 | 2014-12 | 132 | 132 | 100.00000 |
| 2014-35 | 266 | 266 | 100.00000 | 2014-13 | 151 | 151 | 100.00000 |

| signup_week | total_signups | retained_users | retention_rate |
|-------------|---------------|----------------|----------------|
| 2013-01     | 26            | 26             | 100.00000      |
| 2013-02     | 29            | 29             | 100.00000      |
| 2013-03     | 47            | 47             | 100.00000      |
| 2013-04     | 36            | 36             | 100.00000      |
| 2013-05     | 30            | 30             | 100.00000      |
| 2013-06     | 48            | 48             | 100.00000      |
| 2013-07     | 41            | 41             | 100.00000      |
| 2013-08     | 39            | 39             | 100.00000      |
| 2013-09     | 33            | 33             | 100.00000      |
| 2013-10     | 43            | 43             | 100.00000      |
| 2013-11     | 33            | 33             | 100.00000      |
| 2013-12     | 32            | 32             | 100.00000      |
| 2013-13     | 33            | 33             | 100.00000      |
| 2013-14     | 40            | 40             | 100.00000      |
| 2013-15     | 35            | 35             | 100.00000      |
| 2013-16     | 42            | 42             | 100.00000      |
| 2013-17     | 48            | 48             | 100.00000      |
| 2013-18     | 48            | 48             | 100.00000      |
| 2013-19     | 45            | 45             | 100.00000      |
| 2013-20     | 55            | 55             | 100.00000      |
| 2013-21     | 41            | 41             | 100.00000      |
| 2013-22     | 49            | 49             | 100.00000      |
| 2013-23     | 51            | 51             | 100.00000      |
| 2013-24     | 51            | 51             | 100.00000      |

|         |    |    |           |
|---------|----|----|-----------|
| 2013-24 | 51 | 51 | 100.00000 |
| 2013-25 | 46 | 46 | 100.00000 |
| 2013-26 | 57 | 57 | 100.00000 |
| 2013-27 | 57 | 57 | 100.00000 |
| 2013-28 | 52 | 52 | 100.00000 |
| 2013-29 | 71 | 71 | 100.00000 |
| 2013-30 | 66 | 66 | 100.00000 |
| 2013-31 | 69 | 69 | 100.00000 |
| 2013-32 | 66 | 66 | 100.00000 |
| 2013-33 | 73 | 73 | 100.00000 |
| 2013-34 | 71 | 71 | 100.00000 |
| 2013-35 | 79 | 79 | 100.00000 |
| 2013-36 | 65 | 65 | 100.00000 |
| 2013-37 | 71 | 71 | 100.00000 |
| 2013-38 | 84 | 84 | 100.00000 |
| 2013-39 | 92 | 92 | 100.00000 |
| 2013-40 | 81 | 81 | 100.00000 |
| 2013-41 | 88 | 88 | 100.00000 |
| 2013-42 | 74 | 74 | 100.00000 |
| 2013-43 | 97 | 97 | 100.00000 |
| 2013-44 | 92 | 92 | 100.00000 |
| 2013-45 | 97 | 97 | 100.00000 |

**THE ANALYSIS INDICATES THAT THE 35TH WEEK OF 2014 HAD THE HIGHEST NUMBER OF SIGNUPS, WHILE THE 1ST WEEK OF 2013 HAD THE FEWEST SIGNUPS.**

## FINDING 4

```
SELECT
 YEARWEEK(occurred_at) AS activity_week,
 device,
 COUNT(DISTINCT user_id) AS active_users
FROM
 events
WHERE
 occurred_at IS NOT NULL
GROUP BY
 activity_week, device
ORDER BY
 active_users desc;
```

|         |    |    |           |
|---------|----|----|-----------|
| 2013-31 | 69 | 69 | 100.00000 |
| 2013-32 | 66 | 66 | 100.00000 |
| 2013-33 | 73 | 73 | 100.00000 |
| 2013-34 | 71 | 71 | 100.00000 |
| 2013-35 | 79 | 79 | 100.00000 |

DUE TO MORE OUTPUT I HAVE MENTION FEW

BASED ON THE ANALYSIS, DURING THE 30TH WEEK OF 2014, THE DEVICE "MACBOOK PRO" HAD THE MOST ACTIVE USERS, WITH A TOTAL OF 322 USERS. ON THE OTHER HAND, IN THE 35TH WEEK OF THE SAME YEAR, THE "SAMSUNG GALAXY NOTE" HAD THE FEWEST ACTIVE USERS, WITH ONLY 1 USER RECORDED DURING THAT WEEK.

## FINDING 5

```
WITH WeeklyActivity AS (
 SELECT
 YEARWEEK(occurred_at, 1) AS activity_week,
 user_id,
 COUNT(*) AS activity_count
 FROM
 events
 GROUP BY
 activity_week, user_id
)
SELECT
 activity_week,
 COUNT(DISTINCT user_id) AS active_users,
 SUM(activity_count) AS total_activities
FROM
 WeeklyActivity
GROUP BY
 activity_week
ORDER BY
 activity_week desc;
```

|        |      |       |
|--------|------|-------|
| 201435 | 1194 | 16432 |
| 201434 | 1203 | 16417 |
| 201433 | 1215 | 16862 |
| 201432 | 1266 | 18530 |
| 201431 | 1443 | 21706 |
| 201430 | 1363 | 20360 |
| 201429 | 1345 | 20938 |
| 201428 | 1355 | 20188 |
| 201427 | 1271 | 19158 |
| 201426 | 1249 | 19069 |
| 201425 | 1263 | 19198 |
| 201424 | 1219 | 18431 |
| 201423 | 1173 | 18805 |

DURING THE 35TH WEEK OF 2014, THERE  
 WERE 1,194 ACTIVE USERS AND A TOTAL OF 16,432 ACTIVITIES,  
 WHICH IS THE HIGHEST NUMBER OF ACTIVITIES RECORDED  
 FOR THE ENTIRE PERIOD. DURING THE 35TH WEEK OF 2014,  
 THERE WERE 1,194 ACTIVE USERS AND A TOTAL OF 16,432 ACTIVITIES,  
 WHICH IS THE HIGHEST NUMBER OF ACTIVITIES RECORDED FOR THE ENTIRE  
 PERIOD.

## **CONCLUSION**

WITH ALL THE ANALYSIS I HAVE LEARN MAY THINGS DURING THIS DATASET SUCH AS ROLLING TOTAL , CTE ETC.

## **4,HIRING PROCESS ANALYTICS**

I AM WORKING WITH HIRING  
PROCRSS ANALYTICS DATASET . THIS  
TIME I HAVE TO PERFORM ANALYSIS  
USING EXCEL.

**D. DEPARTMENTAL ANALYSIS: VISUALIZING DATA THROUGH CHARTS AND PLOTS IS A CRUCIAL PART OF DATA ANALYSIS.**

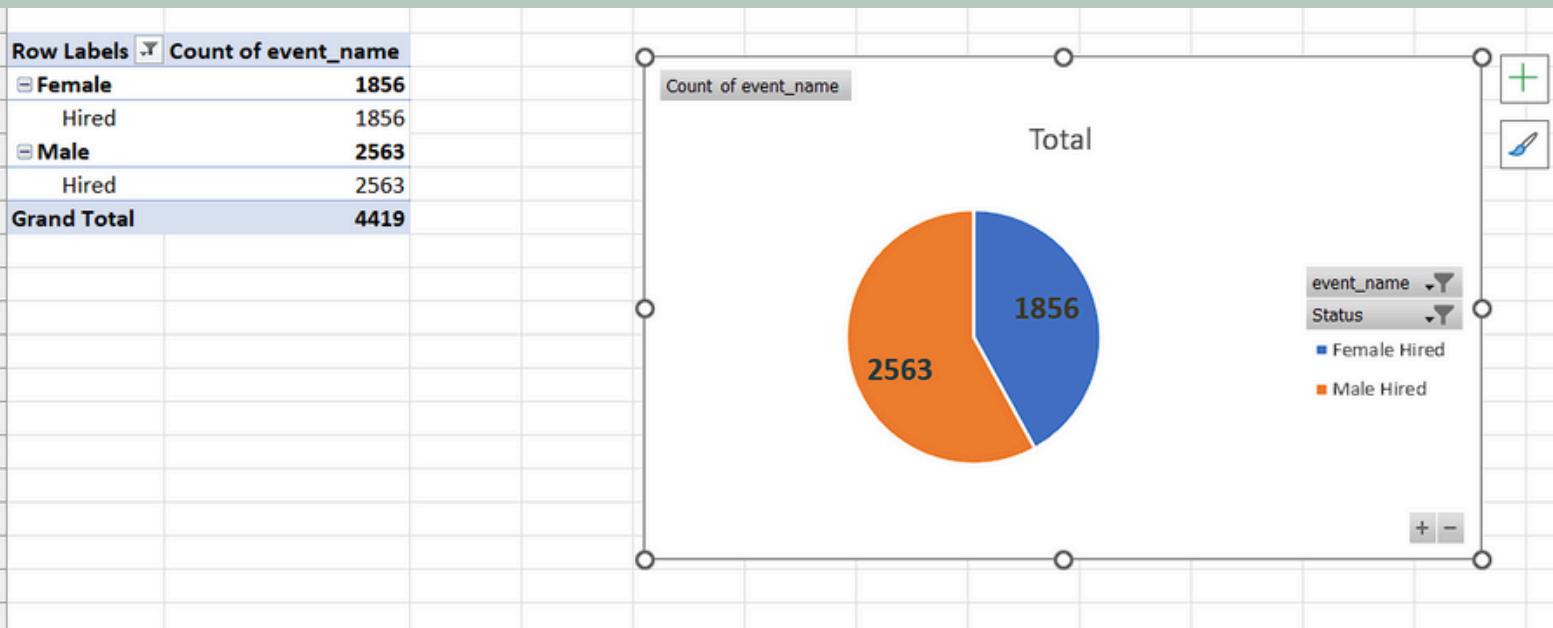
**YOUR TASK:** USE A PIE CHART, BAR GRAPH, OR ANY OTHER SUITABLE VISUALIZATION TO SHOW THE PROPORTION OF PEOPLE WORKING IN DIFFERENT DEPARTMENTS.

**E. POSITION TIER ANALYSIS: DIFFERENT POSITIONS WITHIN A COMPANY OFTEN HAVE DIFFERENT TIERS OR LEVELS.**

**YOUR TASK:** USE A CHART OR GRAPH TO REPRESENT THE DIFFERENT POSITION TIERS WITHIN THE COMPANY. THIS WILL HELP YOU UNDERSTAND THE DISTRIBUTION OF POSITIONS ACROSS DIFFERENT TIERS.

# SOLUTION

## FINDING 1



IN THE ABOVE FIGURE IT SHOWS THE MALE AND FEMALE EMPLOYEE HIRED.

## FINDING 2

| E       | F         | G              | H | I              |
|---------|-----------|----------------|---|----------------|
| t       | Post Name | Offered Salary |   | average_salary |
| artment | c8        | ₹ 56,553.00    |   | ₹ 49,983.03    |
| artment | c5        | ₹ 22,075.00    |   |                |

IN THE TASK WE HAVE TO FIND THE AVERAGE SALARY GIVEN BY THE COMPANY.

## FINDING 3



**SALARY DISTRIBUTION BAR CHART IS PLOTED ACROSS DIFFRENT DEPARTMENT**

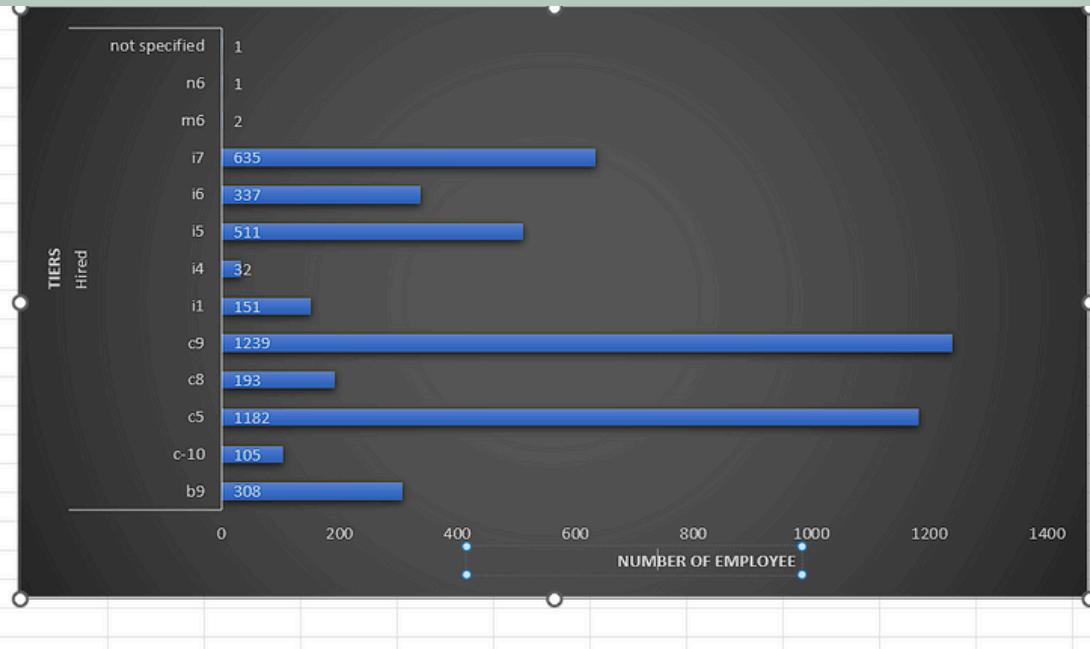
## FINDING 4



**TOTAL NUMBER OF MEMBERS HIRED AT DIFFRENT DEPARTMENT BAR CHART**

# FINDING 5

| Row Labels         | Count of Post Name |
|--------------------|--------------------|
| Hired              | 4697               |
| b9                 | 308                |
| c-10               | 105                |
| c5                 | 1182               |
| c8                 | 193                |
| c9                 | 1239               |
| i1                 | 151                |
| i4                 | 32                 |
| i5                 | 511                |
| i6                 | 337                |
| i7                 | 635                |
| m6                 | 2                  |
| n6                 | 1                  |
| not specified      | 1                  |
| <b>Grand Total</b> | <b>4697</b>        |



**TOTAL NUMBER OF MEMBERS HIRED AT DIFFRENT DEPARTMENT BRANCH BAR CHART.**

## **CONCLUSION**

**AS I HAVE GOT HABDS IN EXPRENCIE  
USING EXCEL AND PIVORT CHART  
WITH DIFFRENT DIFFERENT TYPE OF  
CHARTS.**

## **5. IMDB MOVIE ANALYSIS**

**IN THIS I AM WORKING WITH MOVIES  
DATASET TO ANALYSE DIFFRENT  
TYPE OF TRENDS .**

**STEPS INVOLED TO MAINTAIN DATA**

- **EXTRACT**
- **TRANSFORM**
- **LOAD**

# TASK TO PERFORM

**1. MOVIE GENRE ANALYSIS:** ANALYZE THE DISTRIBUTION OF MOVIE GENRES AND THEIR IMPACT ON THE IMDB SCORE.

**TASK:** DETERMINE THE MOST COMMON GENRES OF MOVIES IN THE DATASET. THEN, FOR EACH GENRE, CALCULATE DESCRIPTIVE STATISTICS (MEAN, MEDIAN, MODE, RANGE, VARIANCE, STANDARD DEVIATION) OF THE IMDB SCORES.

**2. MOVIE DURATION ANALYSIS:** ANALYZE THE DISTRIBUTION OF MOVIE DURATIONS AND ITS IMPACT ON THE IMDB SCORE.

**TASK:** ANALYZE THE DISTRIBUTION OF MOVIE DURATIONS AND IDENTIFY THE RELATIONSHIP BETWEEN MOVIE DURATION AND IMDB SCORE.

**3. LANGUAGE ANALYSIS:** SITUATION: EXAMINE THE DISTRIBUTION OF MOVIES BASED ON THEIR LANGUAGE.

**TASK:** DETERMINE THE MOST COMMON LANGUAGES USED IN MOVIES AND ANALYZE THEIR IMPACT ON THE IMDB SCORE USING DESCRIPTIVE STATISTICS.

**4. DIRECTOR ANALYSIS:** INFLUENCE OF DIRECTORS ON MOVIE RATINGS.

**TASK:** IDENTIFY THE TOP DIRECTORS BASED ON THEIR AVERAGE IMDB SCORE AND ANALYZE THEIR CONTRIBUTION TO THE SUCCESS OF MOVIES USING PERCENTILE CALCULATIONS.

**5. BUDGET ANALYSIS:** EXPLORE THE RELATIONSHIP BETWEEN MOVIE BUDGETS AND THEIR FINANCIAL SUCCESS.

**TASK:** ANALYZE THE CORRELATION BETWEEN MOVIE BUDGETS AND GROSS EARNINGS, AND IDENTIFY THE MOVIES WITH THE HIGHEST PROFIT MARGIN.

# SOLUTION

## FINDING 1

|                             |     |
|-----------------------------|-----|
| Drama                       | 236 |
| Comedy                      | 208 |
| Comedy Drama                | 189 |
| Comedy Drama Romance        | 186 |
| Comedy Romance              | 158 |
| Drama Romance               | 152 |
| Crime Drama Thriller        | 101 |
| Horror                      | 71  |
| Action Crime Drama Thriller | 68  |
| Action Crime Thriller       | 65  |

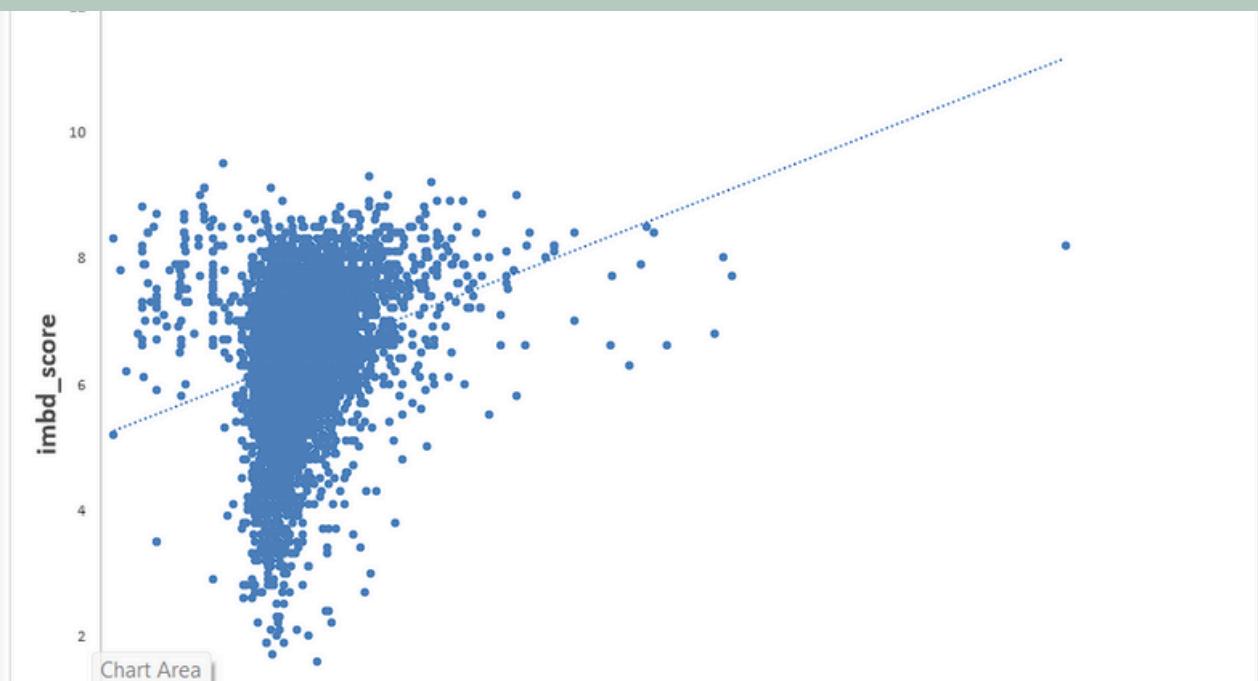
THIS ARE THE TOP 10 MOST LOVED MOVIE BY THE PEOPLE  
IN WHICH DRAMA IS LEADING AMONG ALL

|                    |          |
|--------------------|----------|
| mean               | 6.44159  |
| median             | 6.6      |
| mode               | 6.7      |
| range              | 7.9      |
| varience           | 1.2675   |
| Standard Deviation | 1.125833 |

BEACAUSE OF LARGE DATASET I HAVE TAKEN  
ALL GENRES IMBD MEAN, MEDIAN, MODE,  
RANGE, VARIENCE, STANDARD DEVIATION

# FINDING 2

| duration           |          |  |
|--------------------|----------|--|
| mean               | 107.2268 |  |
| median             | 103      |  |
| mode               | 90       |  |
| range              | 504      |  |
| varience           | 632.1356 |  |
| Standard Deviation | 25.14231 |  |



AS WE CAN SEE THAT THE LESS THE DURATION (BETWEEN 100-200MIN)IS THE MORE THE IMBD SCORE .

# FINDING 3

|   |          |      |
|---|----------|------|
| 1 | English  | 4703 |
| 2 | French   | 73   |
| 3 | Spanish  | 40   |
| 4 | Hindi    | 28   |
| 5 | Mandarin | 26   |

**TOP 5 MOST COMMON LANGUAGES.**

|   | language   | mean     | meedium | Standard Deviation |
|---|------------|----------|---------|--------------------|
| 1 | English    | 6.398873 | 6.5     | 1.12324006         |
| 2 | French     | 7.038356 | 7.2     | 0.726985812        |
| 3 | Spanish    | 6.9375   | 7.15    | 0.855056603        |
| 4 | Hindi      | 6.632143 | 6.95    | 1.398955582        |
| 5 | Mandarin   | 6.788462 | 7.05    | 1.042046802        |
| 6 | German     | 7.342105 | 7.6     | 0.954123093        |
| 7 | Japanese   | 7.394444 | 7.6     | 0.990823913        |
| 8 | Russian    | 6.363636 | 6.5     | 1.383671007        |
| 9 | Italian    | 7.227273 | 7.3     | 1.244259546        |
| 0 | Cantonese  | 6.954545 | 7.2     | 0.704788814        |
| 1 | Portuguese | 7.4875   | 7.7     | 0.883883476        |
| 2 | Korean     | 7.3875   | 7.5     | 0.825378701        |
| 3 | Swedish    | 7.44     | 7.6     | 0.756967635        |
| 4 | Hebrew     | 7.58     | 7.6     | 0.334664011        |
| 5 | Danish     | 7.5      | 8.1     | 1.077032961        |
| 6 | Arabic     | 7.38     | 7.4     | 0.884307639        |
| 7 | Persian    | 7.575    | 7.95    | 1.203813385        |
| 8 | Dutch      | 7.425    | 7.45    | 0.434932945        |
| 9 | Norwegian  | 7.15     | 7.3     | 0.574456265        |
| 0 | Polish     | 8.25     | 8.25    | 0.981495458        |
| 1 | Chinese    | 5.666667 | 5.7     | 0.550757055        |
| 2 | Thai       | 6.633333 | 6.6     | 0.450924975        |
| 3 | Dari       | 7.5      | 7.5     | 0.141421356        |

**TOP 25 LANGUAGES WITH MEAN , MEADIUM , SD.**

# FINDING 4

| directors               | average movie_imdb | percent rank | Column1      |
|-------------------------|--------------------|--------------|--------------|
| John Blanchard          | 9.5                | 1            | Top Director |
| Cary Bell               | 8.7                | 0.998        | Other        |
| Mitchell Altieri        | 8.7                | 0.998        | Other        |
| Sadyk Sher-Niyaz        | 8.7                | 0.998        | Other        |
| Charles Chaplin         | 8.6                | 0.997        | Other        |
| Mike Mayhall            | 8.6                | 0.997        | Other        |
| Damien Chazelle         | 8.5                | 0.996        | Other        |
| Majid Majidi            | 8.5                | 0.996        | Other        |
| Raja Menon              | 8.5                | 0.996        | Other        |
| Ron Fricke              | 8.5                | 0.996        | Other        |
| Christopher Nolan       | 8.425              | 0.995        | Other        |
| Sergio Leone            | 8.475              | 0.995        | Other        |
| Asghar Farhadi          | 8.4                | 0.991        | Other        |
| Bill Melendez           | 8.4                | 0.991        | Other        |
| Catherine Owens         | 8.4                | 0.991        | Other        |
| Jay Oliva               | 8.4                | 0.991        | Other        |
| Marius A. Markevicius   | 8.4                | 0.991        | Other        |
| Moustapha Akkad         | 8.4                | 0.991        | Other        |
| Rakeysh Omprakash Mehta | 8.4                | 0.991        | Other        |
| Richard Marquand        | 8.4                | 0.991        | Other        |
| Robert Mulligan         | 8.4                | 0.991        | Other        |
| S.S. Rajamouli          | 8.4                | 0.991        | Other        |
| Fritz Lang              | 8.3                | 0.988        | Other        |

**THE RESULT IS OF TOP 24  
BEST DIRECTOR AND  
AMONG THEM JOHN  
BLANCHARD IS AT TOP**

# FINDING 5

| genres                          | budget | gross     | movie_title              | Profit Margin |
|---------------------------------|--------|-----------|--------------------------|---------------|
| Crime Drama                     | 3250   | 48468408  | The Mongol King          | 14912.35610   |
| Horror                          | 15000  | 107917283 | Paranormal Activity      | 7193.48553    |
| Horror                          | 10000  | 48468408  | Tin Can Man              | 4845.84075    |
| Romance Short                   | 13000  | 48468408  | The Touch                | 3727.33904    |
| Drama                           | 15000  | 48468408  | Stories of Our Lives     | 3230.22716    |
| Drama Horror Thriller           | 17350  | 48468408  | The Ridges               | 2792.56815    |
| Biography Documentary           | 218    | 592014    | Tarnation                | 2714.66055    |
| Drama                           | 20000  | 48468408  | Flywheel                 | 2422.42037    |
| Fantasy Horror Sci-Fi           | 20000  | 48468408  | Eraserhead               | 2422.42037    |
| Family                          | 20000  | 48468408  | Dude, Where's My Dog?!   | 2422.42037    |
| Sci-Fi Thriller                 | 20000  | 48468408  | All Superheroes Must Die | 2422.42037    |
| Horror                          | 60000  | 140530114 | The Blair Witch Project  | 2341.16856    |
| Comedy Romance                  | 22000  | 48468408  | Dry Spell                | 2202.10943    |
| Drama Family                    | 24000  | 48468408  | Manito                   | 2018.51690    |
| Crime Drama Thriller            | 25000  | 48468408  | Dutch Kills              | 1937.73630    |
| Horror                          | 40000  | 48468408  | Raymond Did It           | 1210.71018    |
| Drama                           | 42000  | 48468408  | This Is Martin Bonner    | 1153.00970    |
| Comedy                          | 45000  | 48468408  | A True Story             | 1076.07572    |
| Documentary                     | 50000  | 48468408  | Counting                 | 968.368150    |
| Documentary                     | 50000  | 48468408  | Bending Steel            | 968.368150    |
| Biography Documentary           | 50000  | 48468408  | The Image Revolution     | 968.368150    |
| Horror Thriller                 | 50000  | 48468408  | Run, Hide, Die           | 968.368150    |
| Fantasy Horror Mystery Thriller | 50000  | 48468408  | The Call of Cthulhu      | 968.368150    |
| Crime Drama Horror              | 60000  | 48468408  | Hayride                  | 806.806792    |

Correlation

0.100967

THIS THE SOME MOST PROFITED GENRES.

# CONCLUSION

THE PROJECT PROVIDED VALUABLE INSIGHTS INTO THE FACTORS THAT INFLUENCE MOVIE SUCCESS

ON IMDB. KEY TAKEAWAYS INCLUDE:

- BUDGETS AND EARNINGS
- GENRES
- DIRECTOR IMPACT
- LANGUAGE AND DURATION
- PROFITABILITY

## **6. BANK LOAN CASE STUDY**

IN THIS DATASET WE ARE WORKING WITH BANK LOAN TO IDENTIFIY WHY THEIR IS HUGE AMOUNT OF REJECTION IN THE BANK LOAN APPLICATION.

**STEPS INVOLVED TO MAINTAIN DATA**

- EXTRACT
- TRANSFORM
- LOAD

# TASK TO PERFORM

**1. IDENTIFY MISSING DATA AND DEAL WITH IT APPROPRIATELY:** AS A DATA ANALYST, YOU COME ACROSS MISSING DATA IN THE LOAN APPLICATION DATASET. IT IS ESSENTIAL TO HANDLE MISSING DATA EFFECTIVELY TO ENSURE THE ACCURACY OF THE ANALYSIS.

**TASK:** IDENTIFY THE MISSING DATA IN THE DATASET AND DECIDE ON AN APPROPRIATE METHOD TO DEAL WITH IT USING EXCEL BUILT-IN FUNCTIONS AND FEATURES.

**2. IDENTIFY OUTLIERS IN THE DATASET:** OUTLIERS CAN SIGNIFICANTLY IMPACT THE ANALYSIS AND DISTORT THE RESULTS. YOU NEED TO IDENTIFY OUTLIERS IN THE LOAN APPLICATION DATASET.

**TASK:** DETECT AND IDENTIFY OUTLIERS IN THE DATASET USING EXCEL STATISTICAL FUNCTIONS AND FEATURES, FOCUSING ON NUMERICAL VARIABLES.

**3. ANALYZE DATA IMBALANCE:** DATA IMBALANCE CAN AFFECT THE ACCURACY OF THE ANALYSIS, ESPECIALLY FOR BINARY CLASSIFICATION PROBLEMS. UNDERSTANDING THE DATA DISTRIBUTION IS CRUCIAL FOR BUILDING RELIABLE MODELS.

**TASK:** DETERMINE IF THERE IS DATA IMBALANCE IN THE LOAN APPLICATION DATASET AND CALCULATE THE RATIO OF DATA IMBALANCE USING EXCEL FUNCTIONS.

**4. PERFORM UNIVARIATE, SEGMENTED UNIVARIATE, AND BIVARIATE ANALYSIS:** TO GAIN INSIGHTS INTO THE DRIVING FACTORS OF LOAN DEFAULT, IT IS IMPORTANT TO CONDUCT VARIOUS ANALYSES ON CONSUMER AND LOAN ATTRIBUTES.

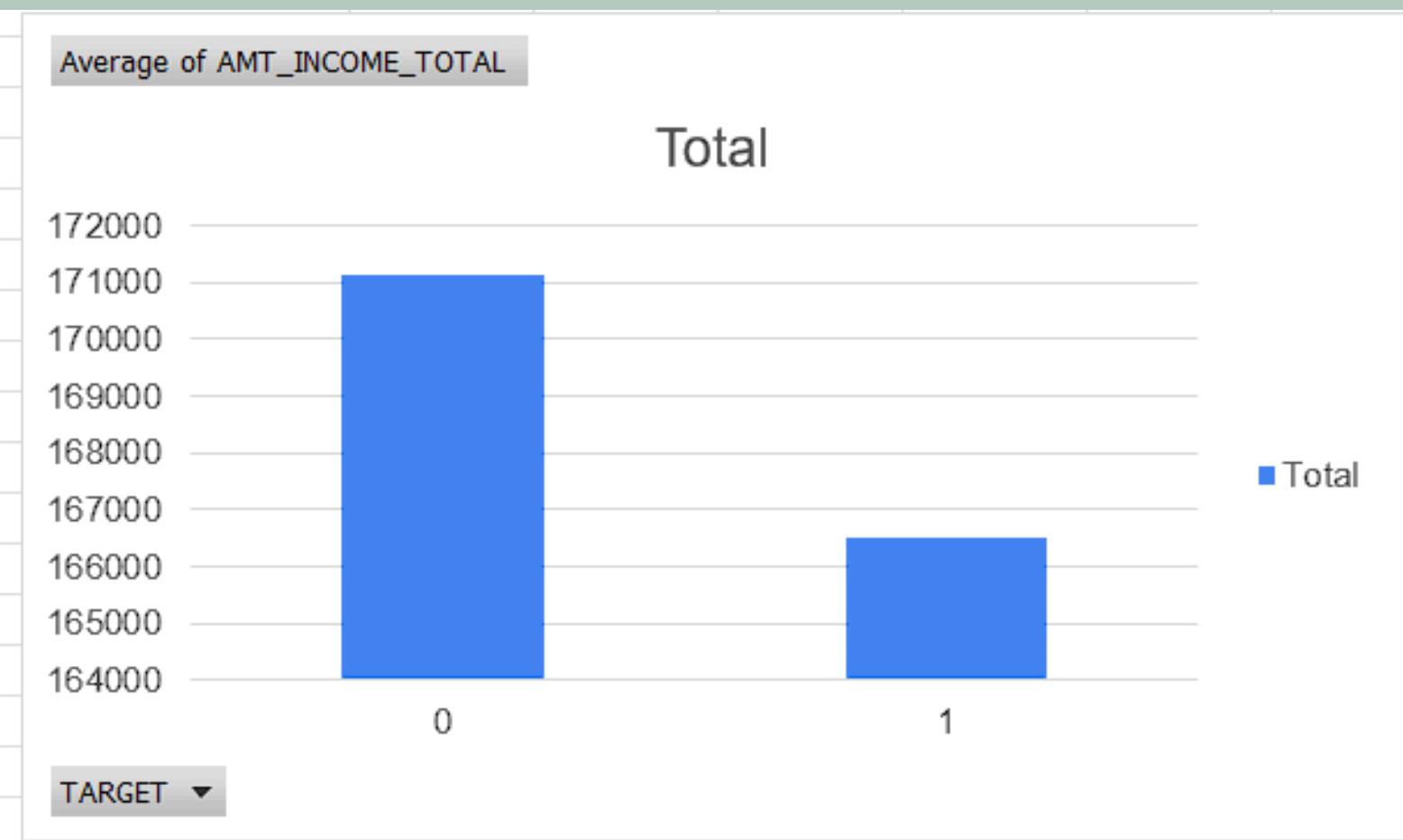
**TASK:** PERFORM UNIVARIATE ANALYSIS TO UNDERSTAND THE DISTRIBUTION OF INDIVIDUAL VARIABLES, SEGMENTED UNIVARIATE ANALYSIS TO COMPARE VARIABLE DISTRIBUTIONS FOR DIFFERENT SCENARIOS, AND BIVARIATE ANALYSIS TO EXPLORE RELATIONSHIPS BETWEEN VARIABLES AND THE TARGET VARIABLE USING EXCEL FUNCTIONS AND FEATURES.

**5. IDENTIFY TOP CORRELATIONS FOR DIFFERENT SCENARIOS:** UNDERSTANDING THE CORRELATION BETWEEN VARIABLES AND THE TARGET VARIABLE CAN PROVIDE INSIGHTS INTO STRONG INDICATORS OF LOAN DEFAULT.

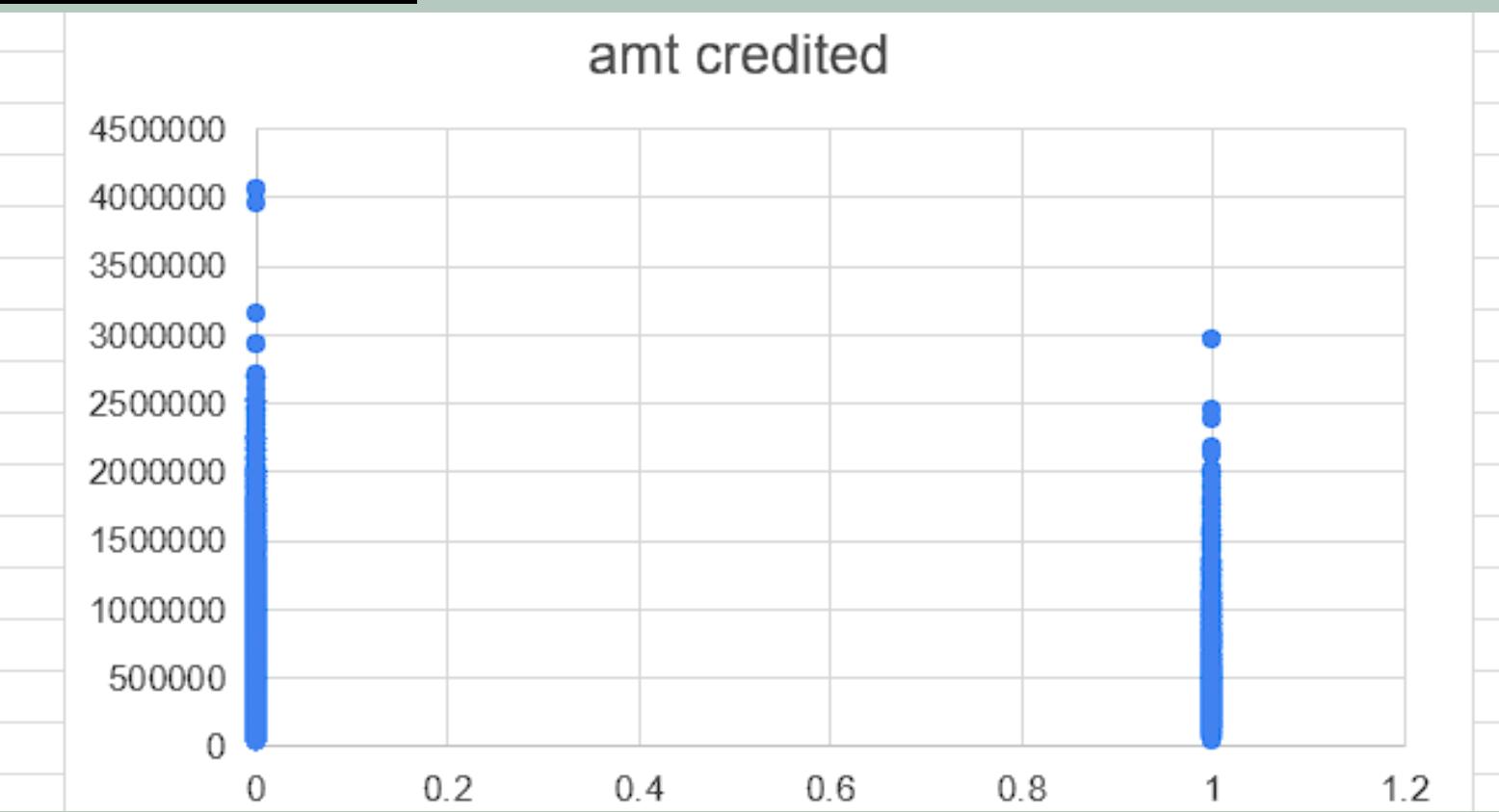
**TASK:** SEGMENT THE DATASET BASED ON DIFFERENT SCENARIOS (E.G., CLIENTS WITH PAYMENT DIFFICULTIES AND ALL OTHER CASES) AND IDENTIFY THE TOP CORRELATIONS FOR EACH SEGMENTED DATA USING EXCEL FUNCTIONS.

# SOLUTION

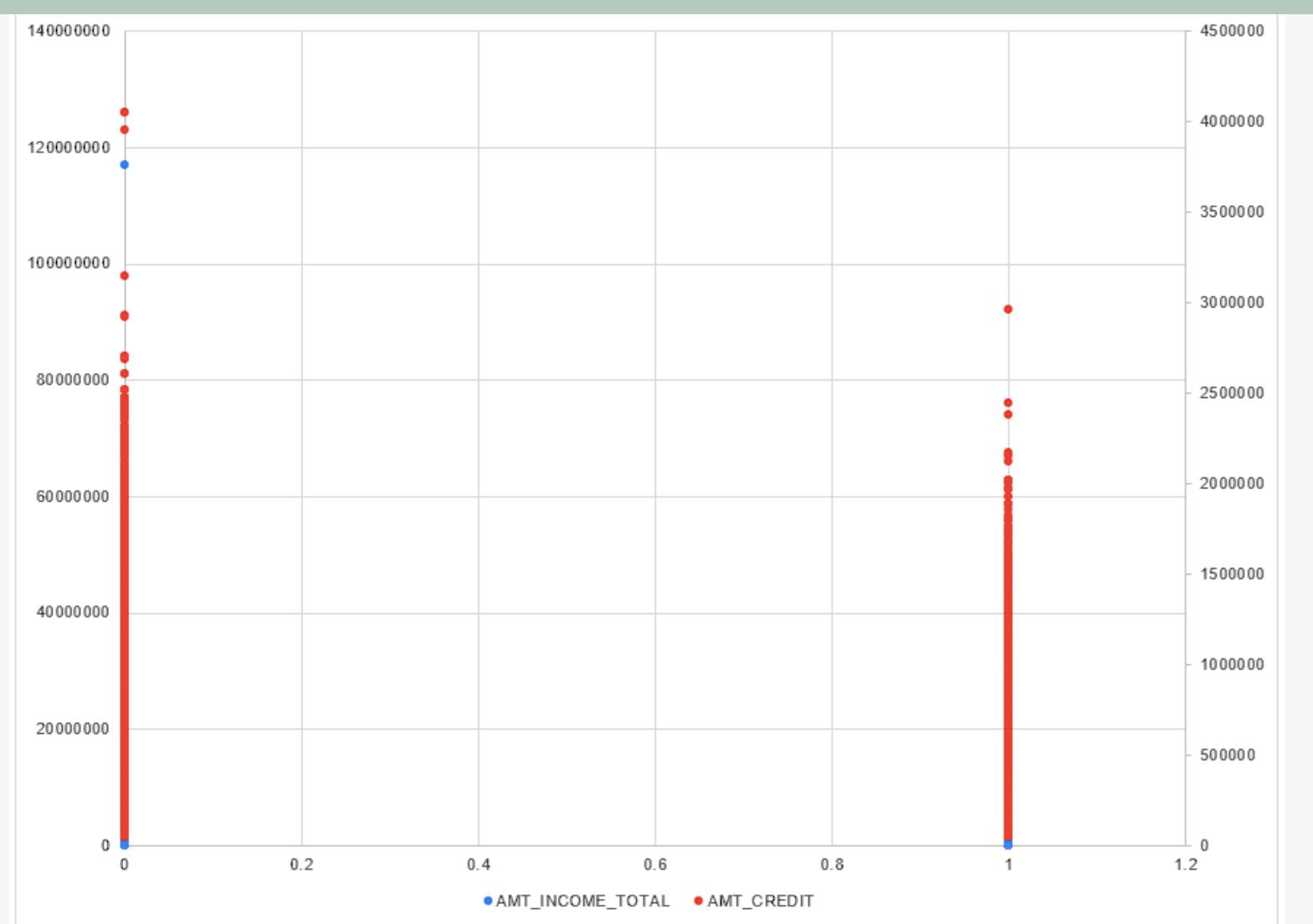
## FINDING 1



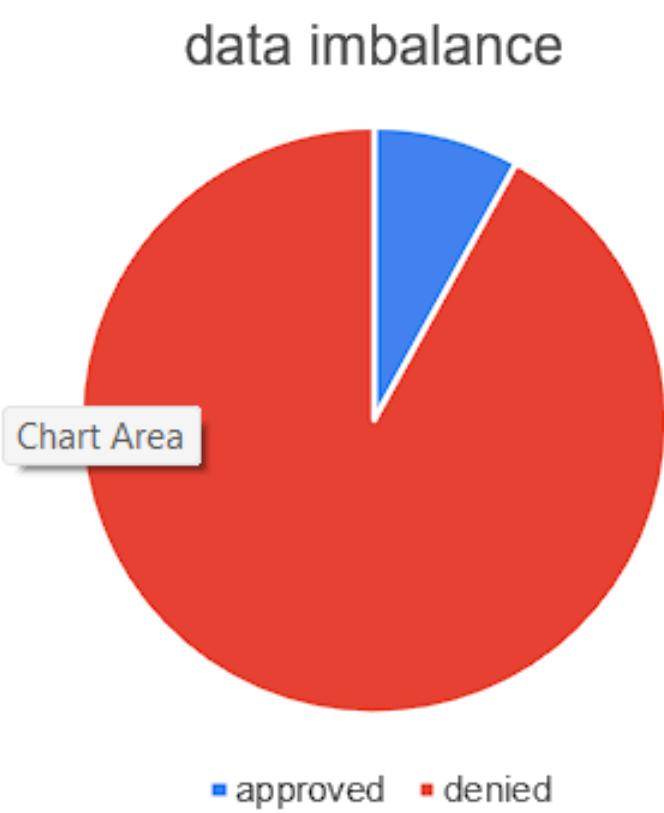
## FINDING 2



## FINDING 3

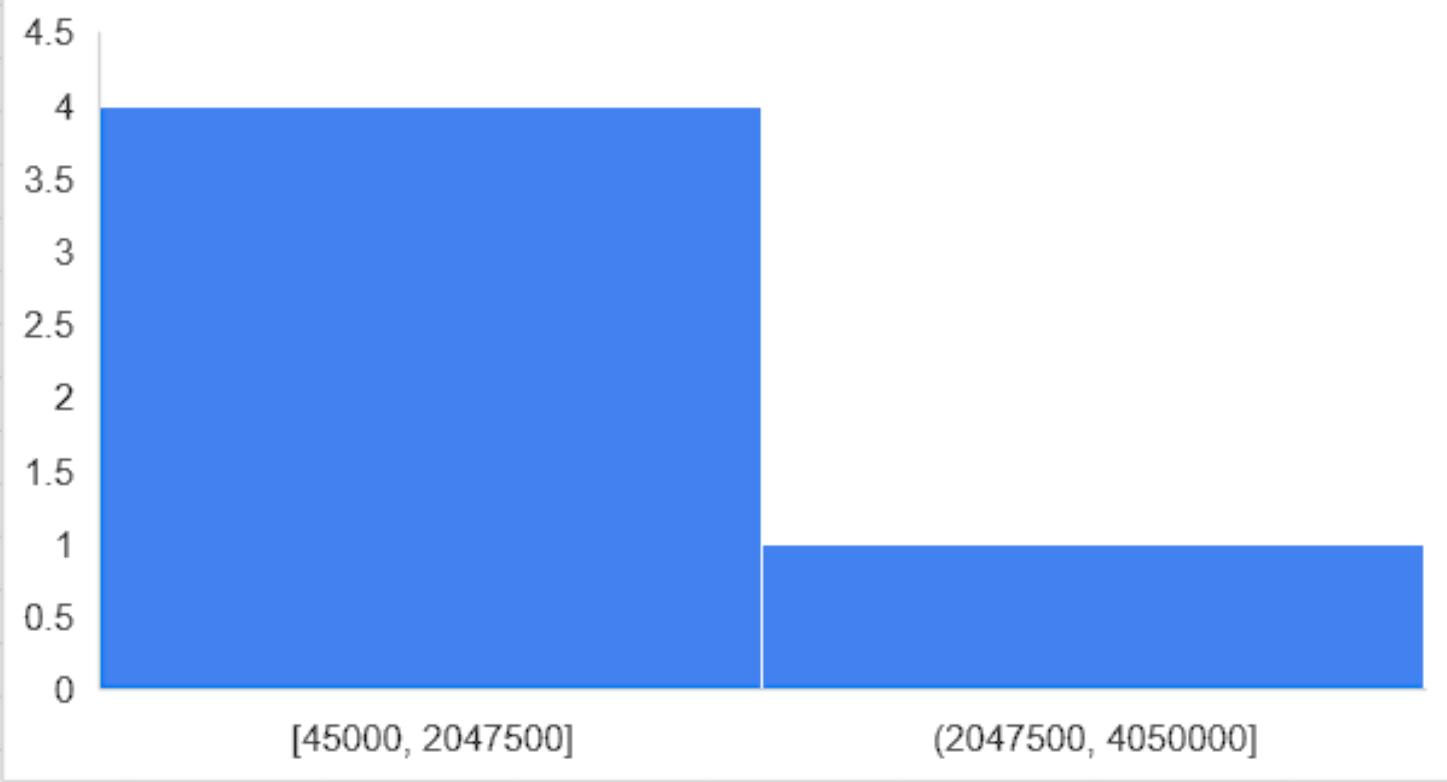


## FINDING 4



## FINDING 5

histogram



## CONCLUSION

- AS I HAVE SEEN THAT THE MOST OF THE LOAN DATA HAVE BEEN REJECTED BY THE BANK .
- DATA HAVE BEEN IMBALANCE IN HUGE AMOUNT

# 7. ANALYZING THE IMPACT OF CAR FEATURES ON PRICE AND PROFITABILITY

## STEPS INVOLVED TO MAINTAIN DATA

- EXTRACT
- TRANSFORM
- LOAD

# TASK TO PERFORM

1. HOW DOES THE POPULARITY OF A CAR MODEL VARY ACROSS DIFFERENT MARKET CATEGORIES?

- TASK 1.A: CREATE A PIVOT TABLE THAT SHOWS THE NUMBER OF CAR MODELS IN EACH MARKET CATEGORY AND THEIR CORRESPONDING POPULARITY SCORES.

- TASK 1.B: CREATE A COMBO CHART THAT VISUALIZES THE RELATIONSHIP BETWEEN MARKET CATEGORY AND POPULARITY.

2. WHAT IS THE RELATIONSHIP BETWEEN A CAR'S ENGINE POWER AND ITS PRICE?

TASK 2: CREATE A SCATTER CHART THAT PLOTS ENGINE POWER ON THE X-AXIS AND PRICE ON THE Y-AXIS. ADD A TRENDLINE TO THE CHART TO VISUALIZE THE RELATIONSHIP BETWEEN THESE VARIABLES.

3. WHICH CAR FEATURES ARE MOST IMPORTANT IN DETERMINING A CAR'S PRICE?

- TASK 3: USE REGRESSION ANALYSIS TO IDENTIFY THE VARIABLES THAT HAVE THE STRONGEST RELATIONSHIP WITH A CAR'S PRICE. THEN CREATE A BAR CHART THAT SHOWS THE COEFFICIENT VALUES FOR EACH VARIABLE TO VISUALIZE THEIR RELATIVE IMPORTANCE.

4. HOW DOES THE AVERAGE PRICE OF A CAR VARY ACROSS DIFFERENT MANUFACTURERS?

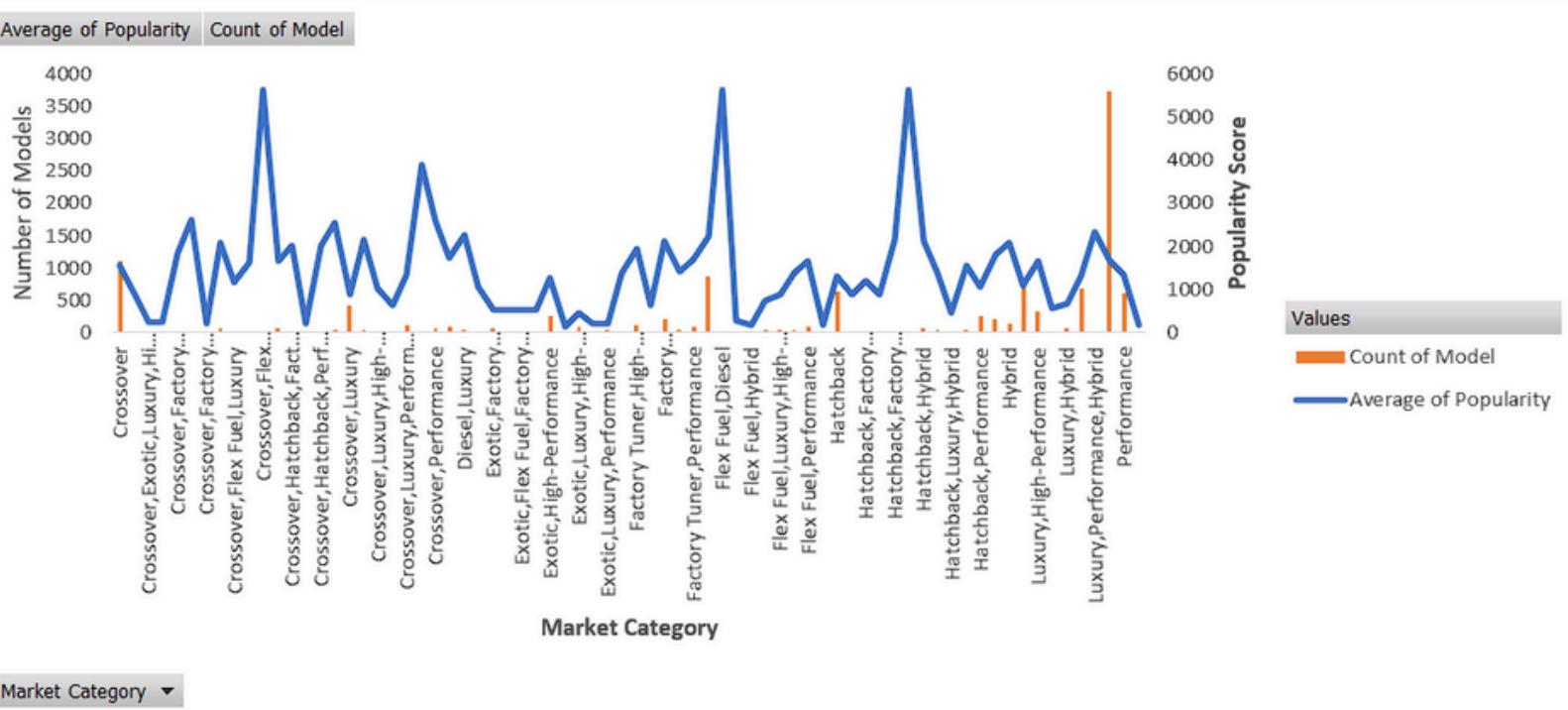
- TASK 4.A: CREATE A PIVOT TABLE THAT SHOWS THE AVERAGE PRICE OF CARS FOR EACH MANUFACTURER.
- TASK 4.B: CREATE A BAR CHART OR A HORIZONTAL STACKED BAR CHART THAT VISUALIZES THE RELATIONSHIP BETWEEN MANUFACTURER AND AVERAGE PRICE.

5. WHAT IS THE RELATIONSHIP BETWEEN FUEL EFFICIENCY AND THE NUMBER OF CYLINDERS IN A CAR'S ENGINE?

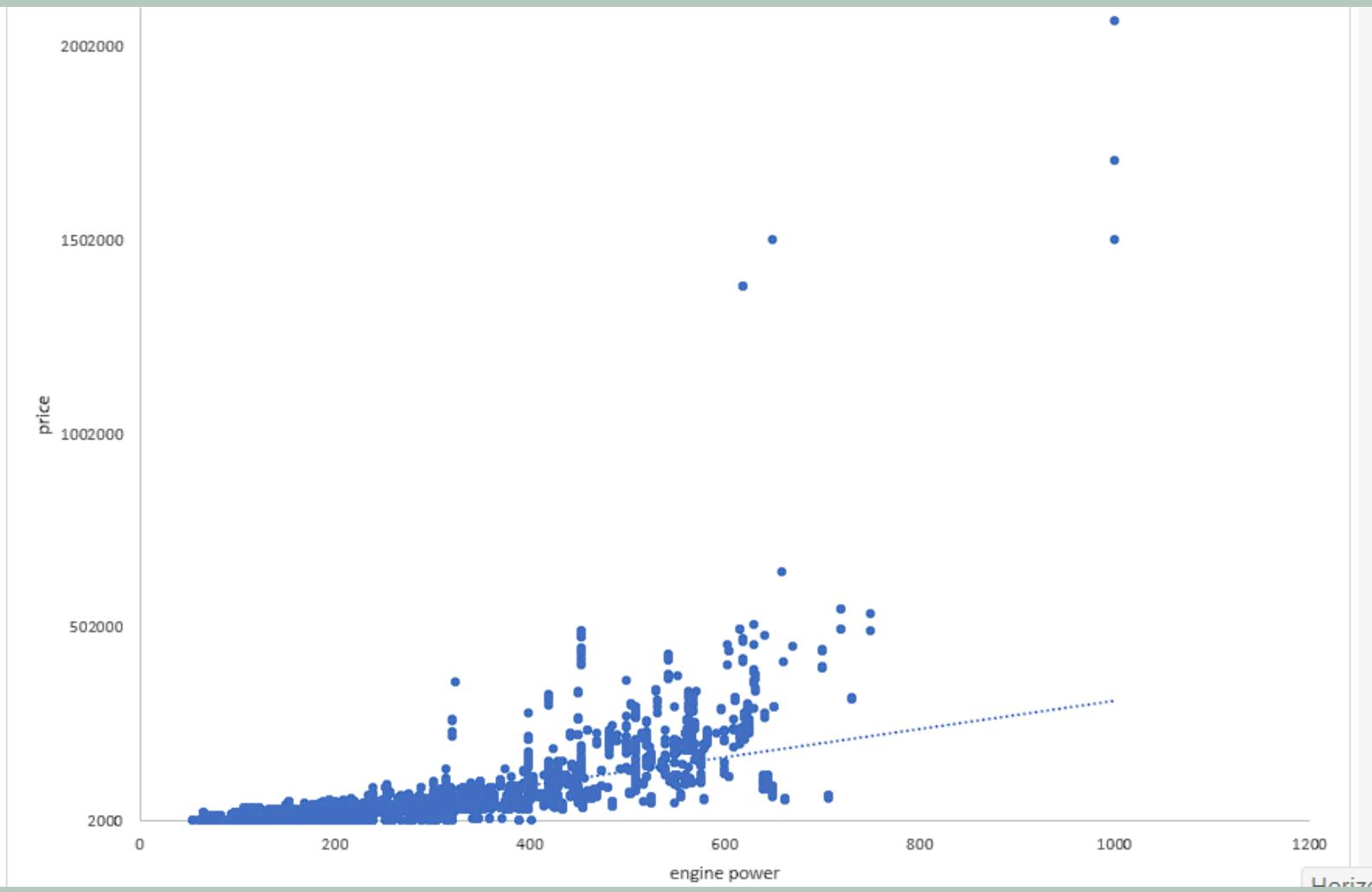
- TASK 5.A: CREATE A SCATTER PLOT WITH THE NUMBER OF CYLINDERS ON THE X-AXIS AND HIGHWAY MPG ON THE Y-AXIS. THEN CREATE A TRENDLINE ON THE SCATTER PLOT TO VISUALLY ESTIMATE THE SLOPE OF THE RELATIONSHIP AND ASSESS ITS SIGNIFICANCE.
- TASK 5.B: CALCULATE THE CORRELATION COEFFICIENT BETWEEN THE NUMBER OF CYLINDERS AND HIGHWAY MPG TO QUANTIFY THE STRENGTH AND DIRECTION OF THE RELATIONSHIP.

# SOLUTION

## FINDING 1

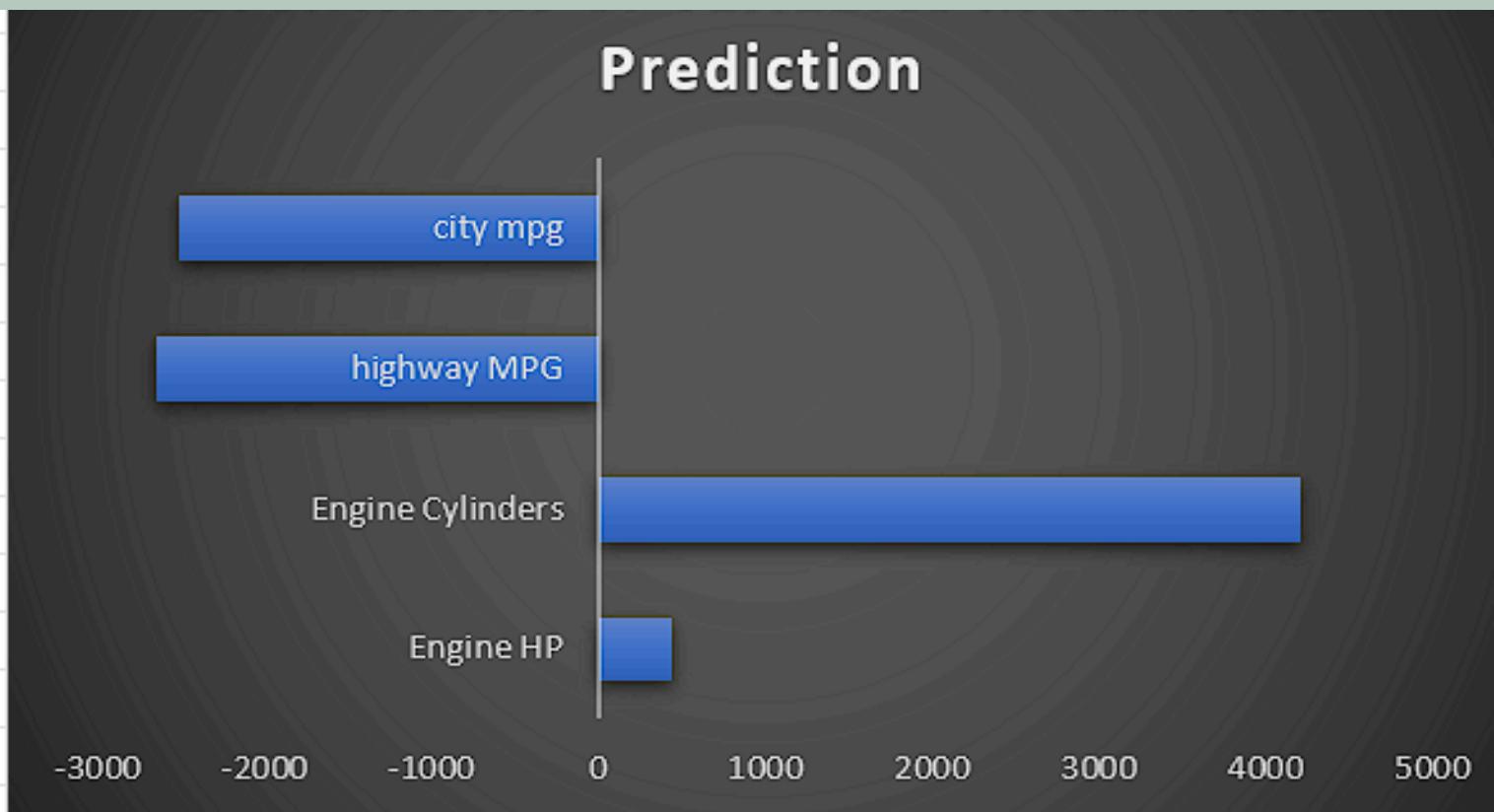


## FINDING 2



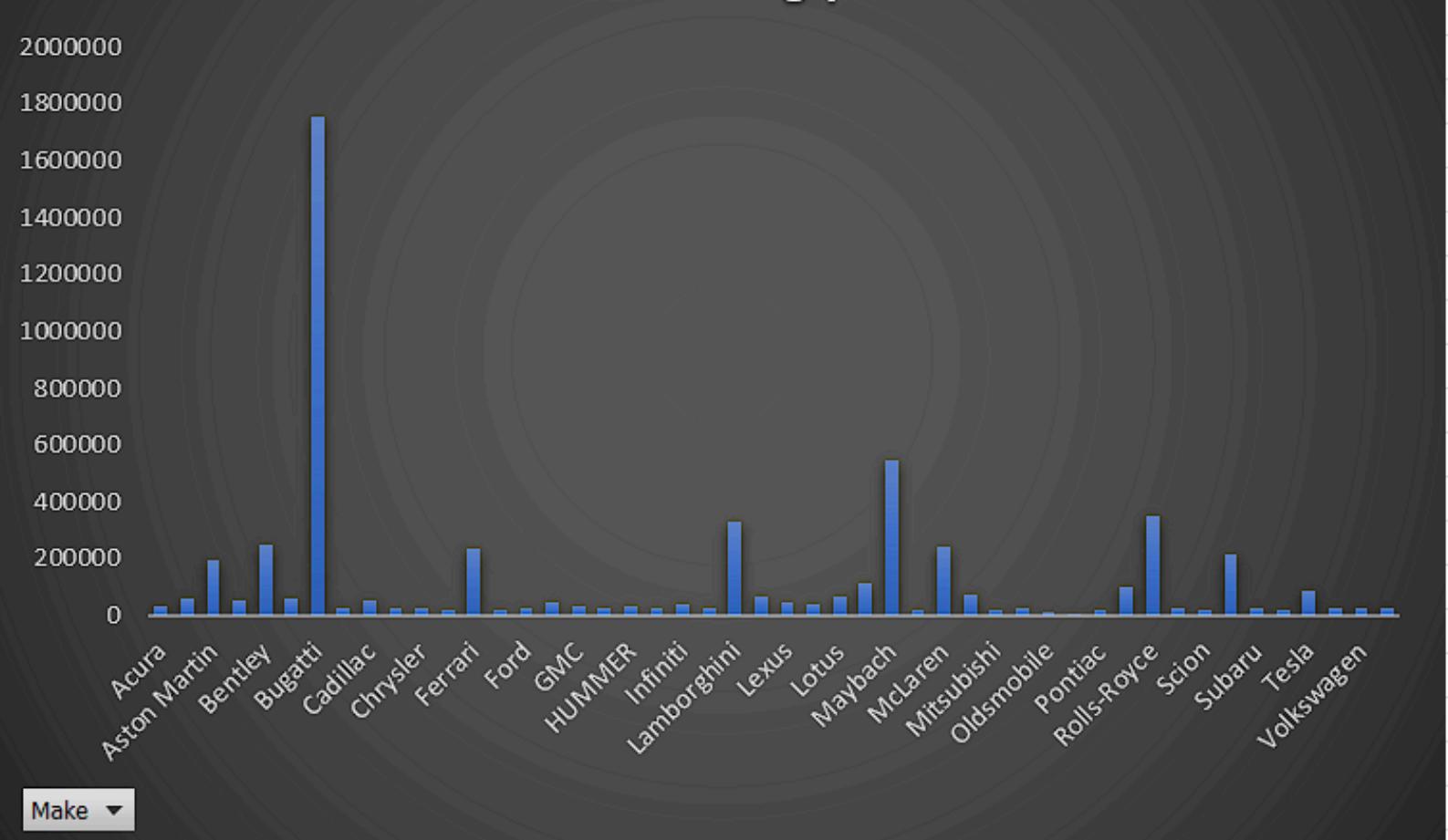
# FINDING 3

## Prediction

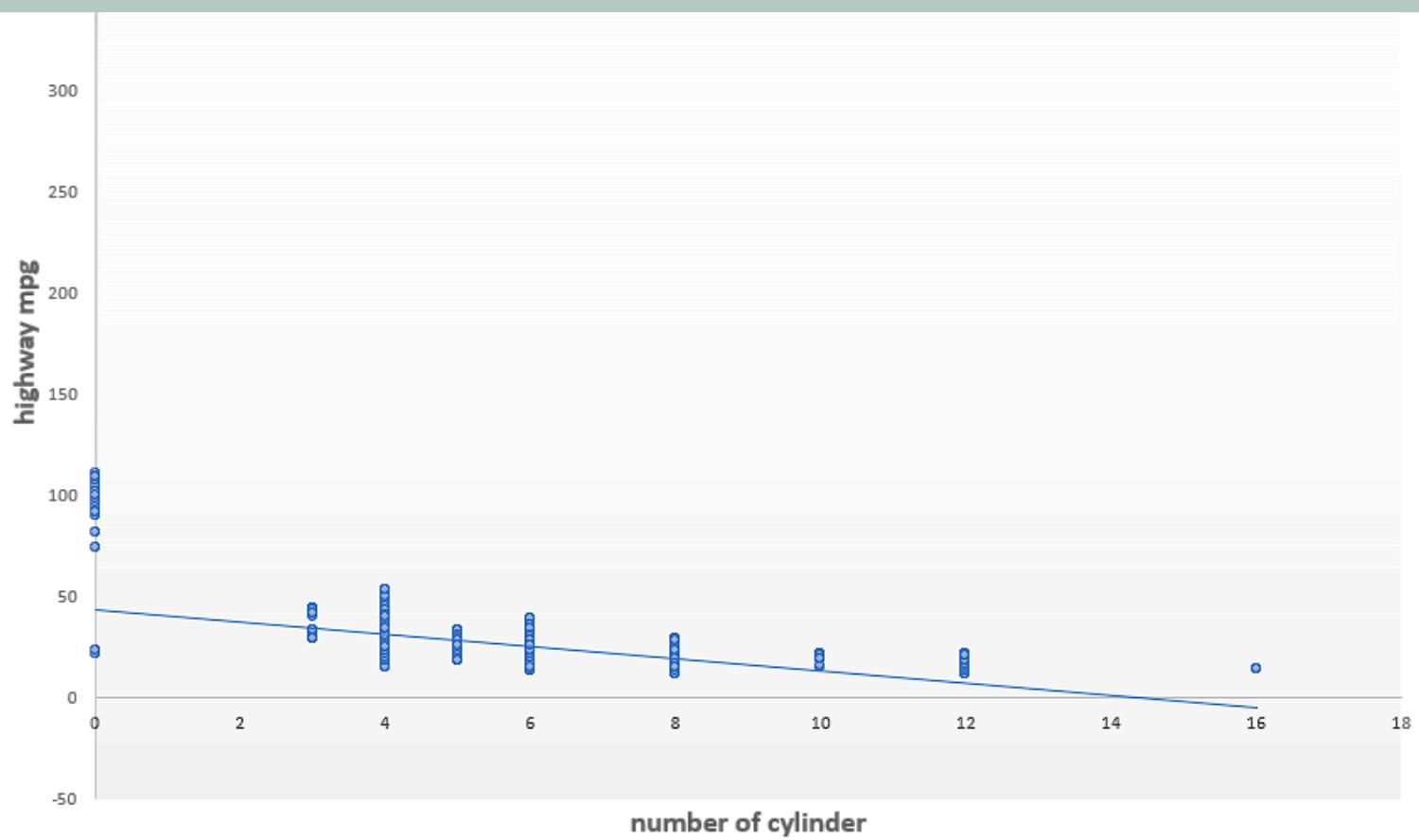


# FINDING 4

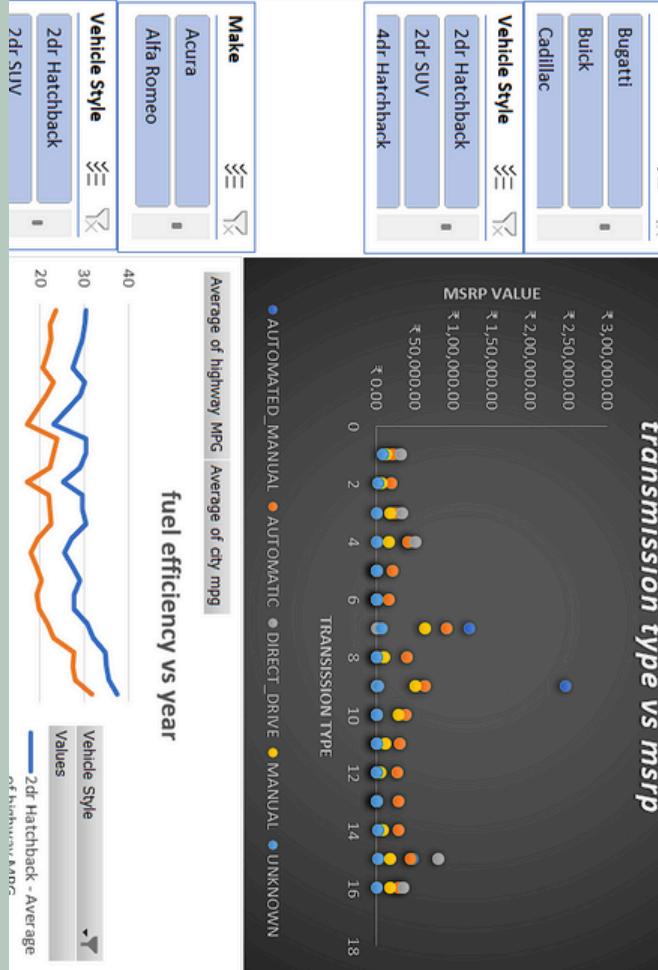
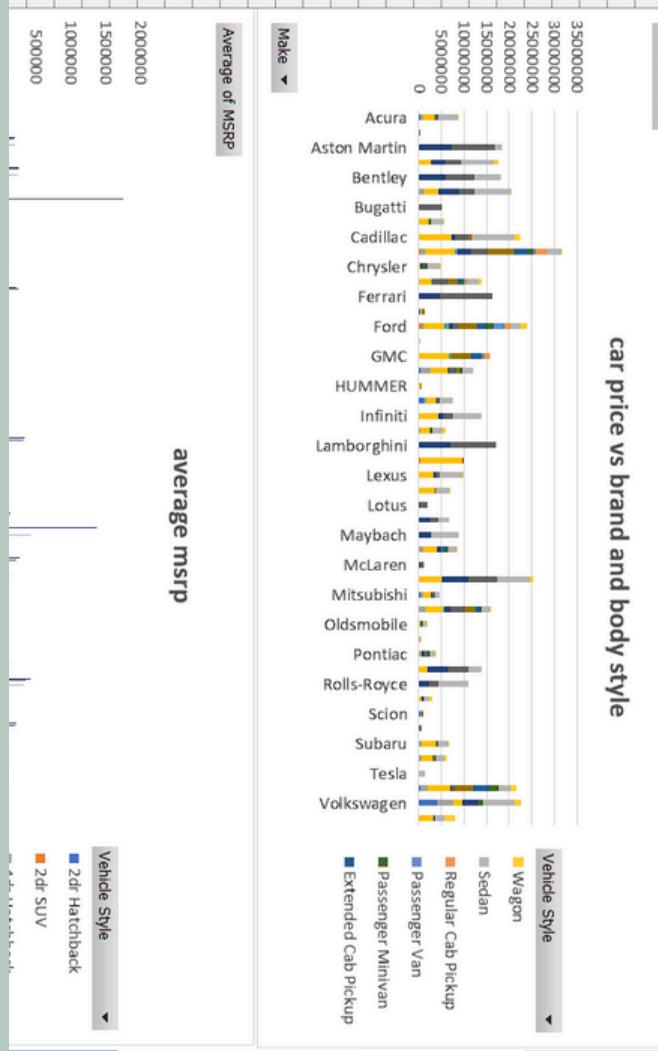
## model vs avg price

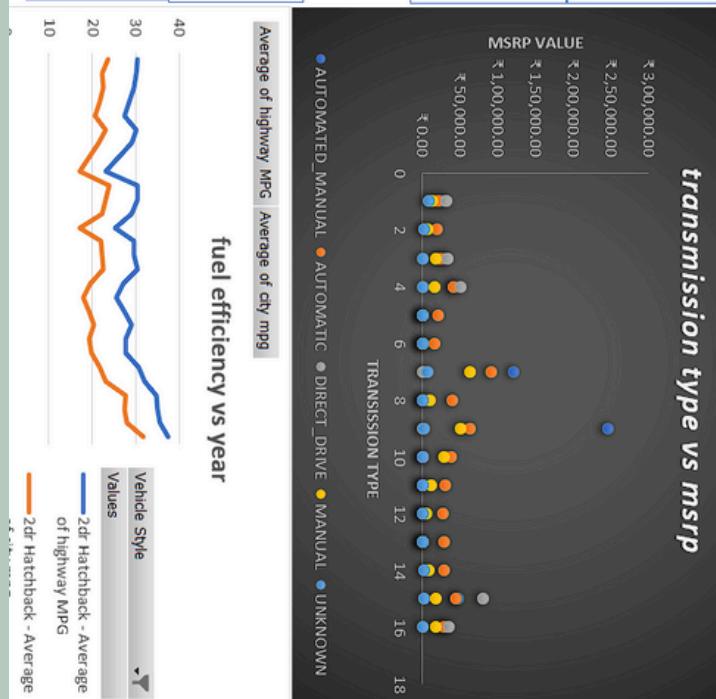
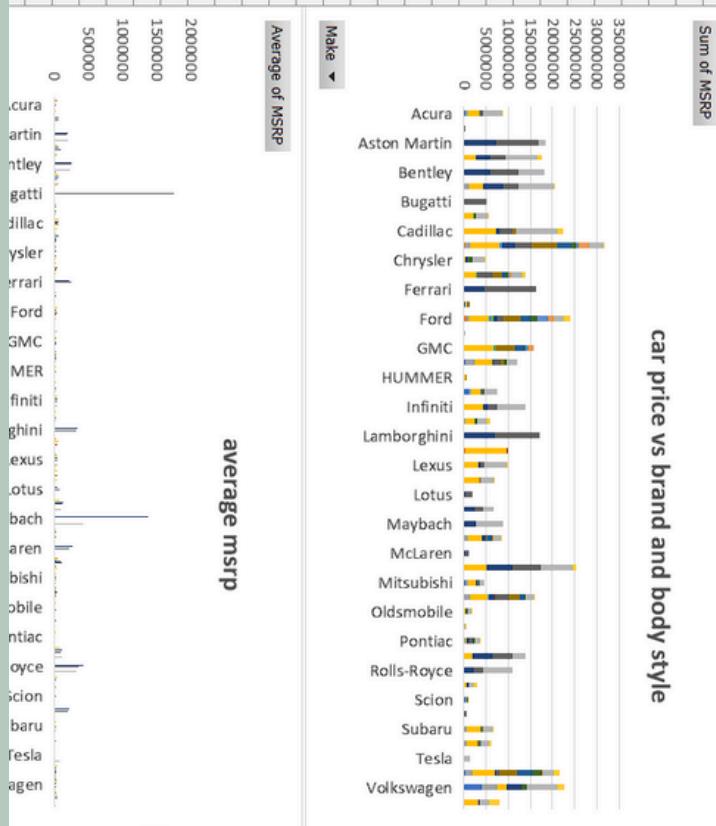


# FINDING 5



# FINDING 5





iii Columns **Avg(Engine HP)**  
ii Rows **Avg(highway MPG)**

Sheet 1



Avg(Msrp)  
3,123  
500,000  
1,000,000  
1,500,000  
1,757,222

Automatic  
Size  
Label  
Shape  
Toolip  
All

Make  
Avg(Msrp)  
Make

②



# **INSIGHTS**

**PRICING (MSRP) INSIGHTS:** VEHICLES WITH HIGHER HORSEPOWER AND LARGER ENGINE SIZES GENERALLY HAVE A HIGHER MSRP.

**TRANSMISSION TYPE AND BODY STYLE :** MANUAL TRANSMISSION CARS TEND TO HAVE LOWER MSRPS, OFTEN ASSOCIATED WITH COMPACT OR SPORTS MODELS.

**MARKET CATEGORY TRENDS:** LUXURY AND PERFORMANCE CATEGORIES EXHIBIT THE HIGHEST MSRPS, WHILE ECONOMY AND COMPACT CARS TARGET AFFORDABILITY.

# **CONCLUSION**

**TABLES AND PIVOT ANALYSIS:** PIVOT TABLE DISPLAYED THE AVERAGE MSRP, POPULARITY SCORES, AND MPG GROUPED BY TRANSMISSION TYPE AND BODY STYLE.

**GRAPHS AND CHARTS:** SCATTER PLOT SHOWED THE RELATIONSHIP BETWEEN HORSEPOWER AND MSRP, WITH DIFFERENT MARKERS FOR BODY STYLES AND TRANSMISSION TYPES.

**BAR CHART:** COMPARED THE NUMBER OF CAR MODELS IN EACH MARKET CATEGORY AND THEIR AVERAGE POPULARITY SCORES.

**DASHBOARD:** CREATED AN INTERACTIVE DASHBOARD IN EXCEL TO DYNAMICALLY FILTER AND VIEW INSIGHTS BASED ON BODY STYLE, MARKET CATEGORY, OR TRANSMISSION TYPE.

## 8.ABC CALL VOLUME TREND ANALYSIS

IN THIS DATASET WE  
ARE GOING TO WORK  
WITH ABC CALL  
VOLUME DATA TO  
FIND OUT VALUABLE  
INSIGHTS

# **TASK TO PERFORM**

1. **AVERAGE CALL DURATION:** DETERMINE THE AVERAGE DURATION OF ALL INCOMING CALLS RECEIVED BY AGENTS. THIS SHOULD BE CALCULATED FOR EACH TIME BUCKET.  
**YOUR TASK:** WHAT IS THE AVERAGE DURATION OF CALLS FOR EACH TIME BUCKET?
2. **CALL VOLUME ANALYSIS:** VISUALIZE THE TOTAL NUMBER OF CALLS RECEIVED. THIS SHOULD BE REPRESENTED AS A GRAPH OR CHART SHOWING THE NUMBER OF CALLS AGAINST TIME. TIME SHOULD BE REPRESENTED IN BUCKETS (E.G., 1-2, 2-3, ETC.).  
**YOUR TASK:** CAN YOU CREATE A CHART OR GRAPH THAT SHOWS THE NUMBER OF CALLS RECEIVED IN EACH TIME BUCKET?
3. **MANPOWER PLANNING:** THE CURRENT RATE OF ABANDONED CALLS IS APPROXIMATELY 30%. PROPOSE A PLAN FOR MANPOWER ALLOCATION DURING EACH TIME BUCKET (FROM 9 AM TO 9 PM) TO REDUCE THE ABANDON RATE TO 10%. IN OTHER WORDS, YOU NEED TO CALCULATE THE MINIMUM NUMBER OF AGENTS REQUIRED IN EACH TIME BUCKET TO ENSURE THAT AT LEAST 90 OUT OF 100 CALLS ARE ANSWERED.  
**YOUR TASK:** WHAT IS THE MINIMUM NUMBER OF AGENTS REQUIRED IN EACH TIME BUCKET TO REDUCE THE ABANDON RATE TO 10%?
4. **NIGHT SHIFT MANPOWER PLANNING:** CUSTOMERS ALSO CALL ABC INSURANCE COMPANY AT NIGHT BUT DON'T GET AN ANSWER BECAUSE THERE ARE NO AGENTS AVAILABLE. THIS CREATES A POOR CUSTOMER EXPERIENCE. ASSUME THAT FOR EVERY 100 CALLS THAT CUSTOMERS MAKE BETWEEN 9 AM AND 9 PM, THEY ALSO MAKE 30 CALLS AT NIGHT BETWEEN 9 PM AND 9 AM. THE DISTRIBUTION OF THESE 30 CALLS IS AS FOLLOWS:  
**YOUR TASK:** PROPOSE A MANPOWER PLAN FOR EACH TIME BUCKET THROUGHOUT THE DAY, KEEPING THE MAXIMUM ABANDON RATE AT 10%.

# STEPS INVOLVED TO MAINTAIN DATA

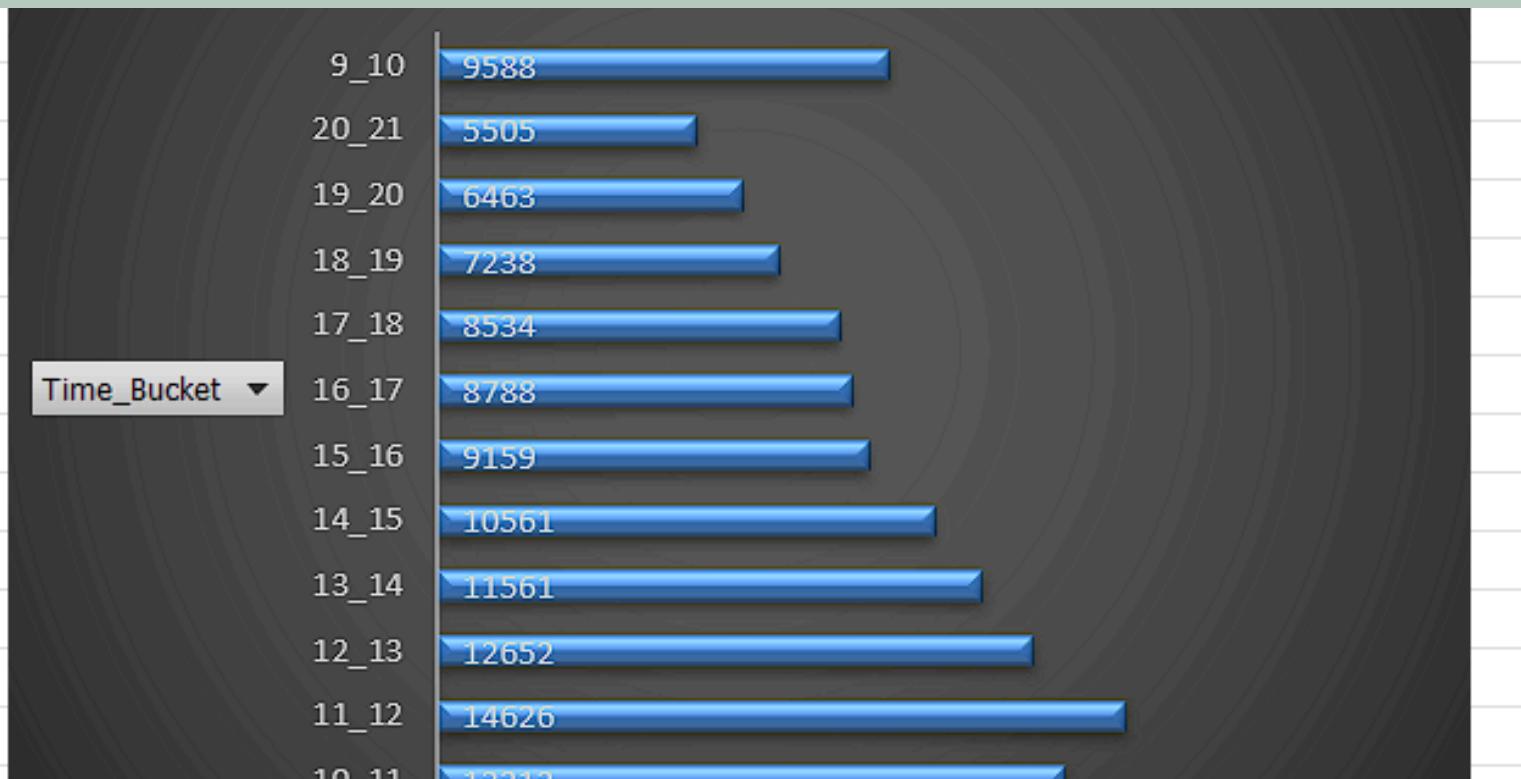
- EXTRACT
- TRANSFROM
- LOAD

# SOLUTION

## FINDING 1

| Row Labels         | Average of Call_Seconds (s) |
|--------------------|-----------------------------|
| 10_11              | 97.42402163                 |
| 11_12              | 116.7837413                 |
| 12_13              | 144.7250237                 |
| 13_14              | 149.5409567                 |
| 14_15              | 146.9693211                 |
| 15_16              | 169.8968228                 |
| 16_17              | 181.4393491                 |
| 17_18              | 179.7245137                 |
| 18_19              | 174.3246753                 |
| 19_20              | 144.5825468                 |
| 20_21              | 105.9491371                 |
| 9_10               | 92.01032541                 |
| <b>Grand Total</b> | <b>139.5321473</b>          |

## FINDING 2



## FINDING 3

| Time_Bucket | Count | Order |
|-------------|-------|-------|
| 12_13       | 12652 | 637   |
| 13_14       | 11561 | 582   |
| 14_15       | 10561 | 532   |
| 15_16       | 9159  | 461   |
| 16_17       | 8788  | 443   |
| 17_18       | 8534  | 430   |
| 18_19       | 7238  | 365   |
| 19_20       | 6463  | 326   |
| 20_21       | 5505  | 277   |
| 9_10        | 9588  | 483   |

# FINDING 4

| Time Slot     | Calls | Agents Required |
|---------------|-------|-----------------|
| 9 pm - 10 pm  | 3     | 1               |
| 10 pm - 11 pm | 3     | 1               |
| 11 pm - 12 am | 2     | 1               |
| 12 am - 1 am  | 2     | 1               |
| 1 am - 2 am   | 1     | 1               |
| 2 am - 3 am   | 1     | 1               |
| 3 am - 4 am   | 1     | 1               |
| 4 am - 5 am   | 1     | 1               |
| 5 am - 6 am   | 3     | 1               |
| 6 am - 7 am   | 4     | 1               |
| 7 am - 8 am   | 4     | 1               |
| 8 am - 9 am   | 5     | 1               |

**Working Hours:** 9 hours/day

**Breaks:** 1.5 hours/day

**Effective Working Hours:**  $9 - 1.5 = 7.5$  hours/day

**Time on Calls:** 60% of 7.5 hours =  $0.60 * 7.5 = 4.5$  hours/day

**Working Days per Month:** 6 days/week \* 4 weeks - 4 days (unplanned leaves) = 20 days/month

**Total Hours per Month:** 4.5 hours/day \* 20 days = 90 hours/month

each call takes 3 minutes on average

Calls handled per hour by one agent = 60 minutes / 3 minutes = 20 calls/hour

Calls handled per month by one agent = 20 calls/hour \* 90 hours = 1800 calls/month

Total night calls per month = 30 calls/night \* 30 days = 900 calls/month

Number of agents required = Total night calls / Calls handled per agent

# **CONCLUSION**

- 1. AVERAGE CALL DURATION**: IDENTIFYING TIME BUCKETS WITH LONGER AVERAGE CALL DURATIONS HELPS IN OPTIMIZING AGENT ALLOCATION AND ENSURING ADEQUATE COVERAGE DURING PEAK HOURS.
- 2. CALL VOLUME DISTRIBUTION**: VISUALIZING CALL VOLUMES HELPS IN UNDERSTANDING CUSTOMER CALLING BEHAVIOR, ALLOWING FOR BETTER PLANNING AND ALLOCATION OF AGENTS TO MEET DEMAND DURING PEAK PERIODS.
- 3. DAY SHIFT MANPOWER PLANNING**: BY ENSURING THAT THE NUMBER OF AGENTS IS SUFFICIENT TO HANDLE 90% OF THE CALLS DURING EACH TIME BUCKET, CUSTOMER SATISFACTION CAN BE SIGNIFICANTLY IMPROVED.
- 4. NIGHT SHIFT MANPOWER PLANNING**: BY ALLOCATING AGENTS BASED ON THE DISTRIBUTION OF NIGHT CALLS AND ENSURING COVERAGE THROUGHOUT THE NIGHT, WE CAN PROVIDE BETTER CUSTOMER SERVICE AND REDUCE THE ABANDON RATE.
- 5. AGENT PRODUCTIVITY**: UNDERSTANDING AGENT PRODUCTIVITY HELPS IN ACCURATELY CALCULATING THE NUMBER OF AGENTS REQUIRED TO HANDLE THE EXPECTED CALL VOLUME, ENSURING OPTIMAL MANPOWER ALLOCATION

# APPENDIX

## PROJECT 2:

**HTTPS://1DRV.MS/P/C/E76F6F243AB164ED/EAWJSAXXX-NNOY\_XCEK4SHUBM7FA-LYGCKH8KJLSRJoTIW?E=N5YOD4**

## PROJECT 3:

**HTTPS://1DRV.MS/P/C/E76F6F243AB164ED/EBKYKF4AE95DP9IT7RMREWABRSXQCT6D2SYPGG7SL4OM-A?E=JAJKL7**

## PROJECT 4:

**HTTPS://1DRV.MS/P/C/E76F6F243AB164ED/ESGLH2PKJH5JQWVLLJ8IY1CBMQBRM-C\_WBSFUX3BCADSoA?E=QFAH2I**

## PROJECT 5:

**HTTPS://1DRV.MS/P/C/E76F6F243AB164ED/ERODBPXIADHLQXEEUV9EB8GBUAIPTQKL\_LFHUUYFIWDCIA?E=QHEINB**

## PROJECT 6:

**HTTPS://1DRV.MS/P/C/E76F6F243AB164ED/ECW3SU\_TW-VJPUXN4\_3OESEBIA4OCSLUZ6GXFAHYBA6\_XG?E=HW6DJI**

## PROJECT 7:

**HTTPS://1DRV.MS/P/C/E76F6F243AB164ED/EXR28AHLCJLASNKS15DGRRoBGKZBE2LE5QQMC-RV56ZCGG?E=SZQMPA**

## PROJECT 8:

**HTTPS://1DRV.MS/P/C/E76F6F243AB164ED/EA65IASBN\_XMHVCJP2L3HGQBVOD7\_R7HoBWLYH5\_OIFWVA?E=XoPVAT**