Trainity

# Portfolio Submission-Data Analysis

**By:- Priti Varma**

## Educational Qualification:

- **Master's in business administration**
- **I am learning Adv Excel, Power Bi, Python, Machine Learning, Sql from Trainity and Data Trained Institute. I am also trying to gain knowledge on Generative AI from Metaleap University.**

## Professional Background:

I have 4.5 year of experience in Recruitment Life Cycle. Working for Banking, Software Companies (IT/Non-IT).I have worked with Ascentiant Business Solutions and Randstad Technology serving companies like John Deere,PWC, Wellsfargo. I have taken a career break from Oct 24, 2023, till date to upskill myself in Data Analysis.
As I am pursuing my education with Trainity, I got an opportunity to work on different types of project which has enhanced my skills on Advance Excel using Microsoft 365, Sql using Mysql Workbench, Python using Jupiter, Tableau and Statistics.

Learnings from my recent projects:

✔ **Gained technical skills in Python, Tableau, SQL, Machine Learning, and more**: These are essential for data manipulation, visualization, and creating predictive models.

✔ **Developed soft skills such as problem-solving and critical thinking**: Critical for analyzing complex data sets and deriving actionable insights.

✔ **Analyzed data to extract useful insights and provide recommendations for price optimization, product development, manpower planning, and loan decisions**: Demonstrating the application of data analysis to various business scenarios.

✔ **Found business insights for marketing campaigns, app features, and user engagement**: Helping to improve marketing strategies and user experiences.

✔ **Used Advanced SQL to answer questions and improve company operations**: Essential for querying databases and optimizing business processes.

✔ **Generated meaningful insights from the IMDB dataset to minimize risk and appeal to a global audience**:

✔ **Analyzed hiring trends and provided useful insights for the department**:

I am excited to continue growing and learning in the field of data analysis!

Trainity

# Table of Content……..

# Module 1: Data Analytics Process in Everyday Life.


## Meal Planning Data Analysis Process followed by me.


1. **Plan**: Identify dietary needs and preferences, and decide on daily meals based on nutritional balance, variety, and ingredient availability.
2. **Prepare**: List required ingredients and check the kitchen for availability and expiration dates.
3. **Process**: Order missing items online. Watch videos for new recipes to ensure efficient preparation and nutritional integrity.
4. **Analyze**: Use the Healthyfi app to monitor calorie intake. Enter meal data into the app to ensure balanced nutrition and adjust exercise as needed.
5. **Communicate**: Discuss meal plans with family and incorporate their feedback for adjustments.
6. **Action Taken**: Purchase ingredients and follow recipes. Adjust recipes to ensure quality and taste.

By following this data-driven approach, meals are nutritious, delicious, and tailored to family needs and preferences.

----------**********----------

# Module 2: Instagram Users Analytics

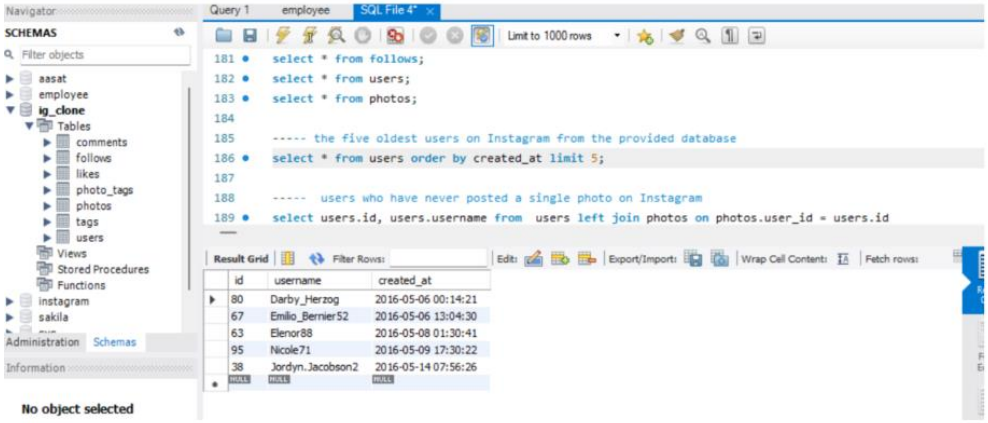

**Description:** The project aims to analyze user interactions and engagement within the Instagram app using SQL queries. Its purpose is to provide valuable insights to the management team, guiding strategic decisions for Instagram's future development. We'll focus on extracting insights from the data to understand user behavior, preferences, and trends.

**Tech-Stack Used:** Utilized MySQL Workbench.

## A) Marketing Analysis:

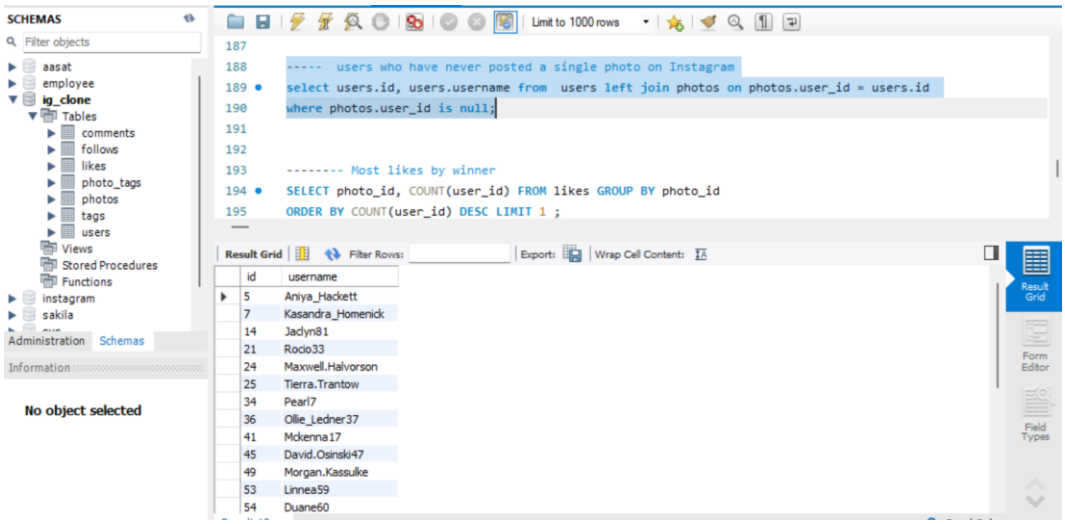
**1.Executed the sql statement to get the oldest 5 users from the given Database.**

**Select * from users order by created_at limit 5;**



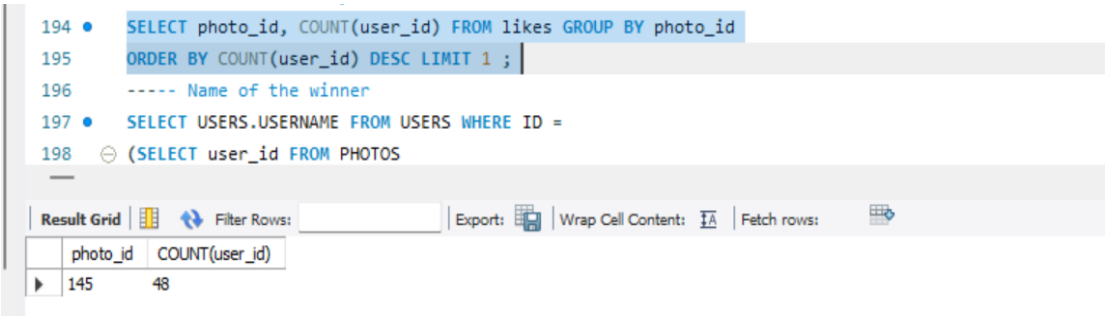
**2.Executed the Sql statement, to find users, who have never posted a single photo on Instagram. This will help the management to share them promotional E-mails**

select users.id, users.username from  users left join photos on photos.user_id = users.id




**3.The team has organized a contest where the user with the most likes on a single photo wins. Sql statement executed to find the Most lokes by winner**

SELECT photo_id, COUNT(user_id) FROM likes GROUP BY photo_id ORDER BY COUNT(user_id) DESC LIMIT 1 ;




**Name of the winner**

SELECT USERS.USERNAME FROM USERS WHERE ID =
(SELECT user_id FROM PHOTOS WHERE
ID = (SELECT  photo_id FROM likes GROUP BY photo_id ORDER BY COUNT(user_id) DESC LIMIT 1));

```
196     ----- Name of the winner
197 •   SELECT USERS.USERNAME FROM USERS WHERE ID =
198     (SELECT user_id FROM PHOTOS
199     WHERE
200     ID = (SELECT  photo_id FROM likes GROUP BY photo_id ORDER BY COUNT(user_id) DESC LIMIT 1));
201     -------- fetching the details of the winner
```

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

| USERNAME |
|---|
| Zack_Kemmer93 |

To find the details of the winner following is the SQL statement used.
WITH Mostlikespic AS ( SELECT photo_id, COUNT(user_id) AS total_likes FROM likes
GROUP BY photo_id ORDER BY COUNT(user_id) DESC LIMIT 1) SELECT username, user_id, photo_id, MostLikespic.total_likes FROM Mostlikespic
JOIN photos ON MostLikespic.photo_id = photos.id JOIN users ON photos.user_id = users.id;

```
202     -------- fetching the details of the winner
203 •   WITH Mostlikespic AS (
204         SELECT photo_id, COUNT(user_id) AS total_likes FROM likes
205         GROUP BY photo_id ORDER BY COUNT(user_id) DESC LIMIT 1)
206     SELECT username, user_id, photo_id, MostLikespic.total_likes FROM Mostlikespic
207     JOIN photos ON MostLikespic.photo_id = photos.id JOIN users ON photos.user_id = users.id;
208
209     -------- hatshtags
210 •   select * from tags;
```

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

| username | user_id | photo_id | total_likes |
|---|---|---|---|
| Zack_Kemmer93 | 52 | 145 | 48 |

3. Identifying and suggesting the top five most commonly used hashtags on the platform. Following is the SQL statement used to help the partner brand to know most popular hashtags.
   Select * from tags;
   with max_hashtags as (select tag_id from photo_tags group by tag_id order by count(tag_id) desc
   limit 5) select tag_name from max_hashtags join tags on max_hashtags.tag_id = tags.id;

```
209     -------- hatshtags
210 •   select * from tags;
211 •   with max_hashtags as (select tag_id from photo_tags  group by tag_id order by count(tag_id) desc
212     limit 5) select tag_name from max_hashtags join tags on max_hashtags.tag_id = tags.id;
213
214     -------- Day of the week when most users register on Instagram
215 •   select dayname(created_at) as weekday, count(*) as 'numnber of registrations' from users
216     group by dayname(created_at) order by 'numnber of registrations' desc;
```

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

| tag_name |
|---|
| smile |
| beach |
| party |
| fun |
| concert |

4. The team wants to know the best day of the week to launch ads. Following is the SQl statement used to find out the Day of the week when most of the users register on Instagram. As per the outcome **Thursdays and Sundays** are the best days to promote campaign followed by **Tuesdays.**
   Select dayname(created_at) as weekday, count(*) as ‘numnber of registrations’ from users
   group by dayname(created_at) order by ‘numnber of registrations’ desc;

```
213
214     -------- Day of the week when most users register on Instagram
215     select dayname(created_at) as weekday, count(*) as 'numnber of registrations' from users
216     group by dayname(created_at) order by 'numnber of registrations' desc;
217
218     SELECT COUNT(id) AS 'total instagram users' from users; ----- total no.of users
219     SELECT COUNT(*) AS total_photos from photos; ----- total no.of photos
```

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

| weekday | numnber of registrations |
|---|---|
| Thursday | 16 |
| Sunday | 16 |
| Tuesday | 14 |
| Saturday | 12 |
| Wednesday | 13 |
| Monday | 14 |
| Friday | 15 |

## B). Investor Metrics:

1. Investors want to know if users are still active and posting on Instagram or if they are making fewer posts. Following are the SQL queries to find the requirements.
   SELECT COUNT(id) AS ‘total nstagram users’ from users; ----- total no.of users
   SELECT COUNT(*) AS total_photos from photos; ----- total no.of photos
   ----- total photos/total users
   SELECT (SELECT Count(*) from photos) / (SELECT Count(*) FROM users) AS avg ;

```
218     SELECT COUNT(id) AS 'total instagram users' from users; ----- total no.of users
219     SELECT COUNT(*) AS total_photos from photos; ----- total no.of photos
220     ----- total photos/total users
221     SELECT (SELECT Count(*) from photos) / (SELECT Count(*) FROM users) AS avg ;
222
223     ----- Average posts per user
```

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

| avg |
|---|
| 2.5700 |

So total Photos per nstagram users are 257/100=2.57

The average number of posts per user on Instagram is found using the following SQL query.
Select avg(total_post_count) as 'Average Post Count' from (select user_id, count(*) as total_post_count from photos group by user_id order by total_post_count desc) as posts_by_users;

```
222
223     ---- Average posts per user
224 •   select avg(total_post_count) as 'Average Post Count' from (select user_id, count(*) as total_post_count
225     photos group by user_id order by total_post_count desc) as posts_by_users;
226
```

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

| Average Post Count |
|---|
| 3.4730 |

**2.Investors want to know if the platform is crowded with fake and dummy accounts. Following is the SQL executed to find out Bots or Dummy/Fake accounts.**

Select username, count(*) as number_of_likes from users join likes on users.id = likes.user_id group by likes.user_id having number_of_likes = (select count(*) from photos);

```
227 •   select username, count(*) as number_of_likes
228     from users join likes on users.id = likes.user_id
229     group by likes.user_id having number_of_likes = (select count(*) from photos);
230
```

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

| username | number_of_likes |
|---|---|
| Aniya_Hackett | 257 |
| Jaclyn81 | 257 |
| Rocio33 | 257 |
| Maxwell.Halvorson | 257 |
| Ollie_Ledner37 | 257 |
| Mckenna17 | 257 |
| Duane60 | 257 |
| Julien_Schmidt | 257 |
| Mike.Auer39 | 257 |
| Nia_Haag | 257 |
| Leslie67 | 257 |
| Janelle.Nikolaus81 | 257 |
| Bethany20 | 257 |

Result 23

## Conclusion:

**From the above provided SQL queries the Management should get the insights or help onto when to promote campaigns, offers, choosing winners and promotional mails to inactive users.**

----------**********----------

# Module3:Operation-Analytics-and-Investigating-Metric-Spike


## Description:

This project involves analyzing datasets provided by the company to derive insights and answer questions posed by various departments such as operations, support, and marketing. The goal is to use data analysis to predict the overall growth or decline of the company's fortunes, improve automation, enhance cross-functional team understanding, and optimize workflows.

**Case Study 1: Job Data Analysis** - We'll work with the job_data table, containing information such as job IDs, actor IDs, event types, time spent on tasks, organization details, and dates to gain insights into job-related activities and performance metrics.

**Case Study 2: Investigating Metric Spikes** - We'll analyze the users, events, and email_events tables to investigate sudden changes or spikes in key metrics such as user engagement and email open rates.

**Tech-Stack Used:**MySQL Workbench

## Case study 1 (Operational Analytics)

SELECT ds AS dates,(COUNT(job_id)/SUM(time_spent))*3600 AS 'jobs reviewed per Hour per Day' FROM job_data where month(ds)=11 GROUP BY 1;

```
19    # Q1. Calculating the number of jobs reviewed per hour for each day in November 2020
20    SELECT ds AS dates,(COUNT(job_id)/SUM(time_spent))*3600 AS 'jobs reviewed per Hour per Day'
21    FROM job_data where month(ds)=11 GROUP BY 1;
22    #  Maximum number of jobs reviewed is 218 On 2020-11-28
23    # The average number of jobs reviewed per hour per day is for november 2020 is 126.
24    # The average number of jobs reviewed per hour per day is for november 2020 is 35 on Nov 27.
25
```

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

| dates | jobs reviewed per Hour per Day |
|---|---|
| 2020-11-30 | 180.0000 |
| 2020-11-29 | 180.0000 |
| 2020-11-28 | 218.1818 |
| 2020-11-27 | 34.6154 |
| 2020-11-26 | 64.2857 |
| 2020-11-25 | 80.0000 |

**Insights:**
Maximum number of jobs reviewed is 218 On 2020-11-28
The average number of jobs reviewed per hour per day is for November 2020 is 126.
The average number of jobs reviewed per hour per day is for November 2020 is 35 on Nov 27.

**Q2.Syntax used: To find weekly throughput**
**Select round(count(event)/sum(time_spent),2) as "7 days throughput" from job_data;**

```
67    # Q2.Calculate the 7-day rolling average of throughput (number of events per second).
68    # Task: Write an SQL query to calculate the 7-day rolling average of throughput.
69    # Additionally, explain whether you prefer using the daily metric or the 7-day rolling average for throug
70 •  Select round(count(event)/sum(time_spent),2) as "7 days throughput" from job_data;
```

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

| 7 days throughput |
|---|
| 0.03 |

**Insights: 7 day throughput is 0.03**

**Syntax used: to find throughput per day**
select ds as dates, round(count(event)/sum(time_spent),2) as "Throughput per day"
FROM job_data group by ds order by ds;

```
72 •  select ds as dates, round(count(event)/sum(time_spent),2) as "Throughput per day"
73    FROM job_data group by ds order by ds; # The throughput is highest 0.06 on 28 Nov 2020.
74
```

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

| dates | Throughput per day |
|---|---|
| 2020-11-25 | 0.02 |
| 2020-11-26 | 0.02 |
| 2020-11-27 | 0.01 |
| 2020-11-28 | 0.06 |
| 2020-11-29 | 0.05 |
| 2020-11-30 | 0.05 |

**Insights: The throughput is highest 0.06 on 28 Nov 2020**

**Q3.Syntax used:**
SELECT language, ROUND (COUNT(*)/sum*100, 2) as 'Percentage share of each language' from job_data
CROSS JOIN(SELECT count(*) as sum from job_data) as sum_jobdata group by language, sum ;

```
77 • SELECT language, ROUND (COUNT(*)/sum*100, 2) as 'Percentage share of each language' from job_data
78   CROSS JOIN(SELECT count(*) as sum from job_data) as sum_jobdata group by language, sum ;
79   #Persian language is highest with 37.5% total.
80
```

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

| language | Percentage share of each language |
|---|---|
| English | 12.50 |
| Arabic | 12.50 |
| Persian | 37.50 |
| Hindi | 12.50 |
| French | 12.50 |
| Italian | 12.50 |

**Insights: Persian language is highest with 37.5% total.**

Q4. **Syntax used:**
SELECT language, ROUND (COUNT(*)/sum*100, 2) as 'Percentage share of each language' from job_data
CROSS JOIN(SELECT count(*) as sum from job_data) as sum_jobdata group by language, sum ;

```
77 • SELECT language, ROUND (COUNT(*)/sum*100, 2) as 'Percentage share of each language' from job_data
78   CROSS JOIN(SELECT count(*) as sum from job_data) as sum_jobdata group by language, sum ;
79   #Persian language is highest with 37.5% total.
80
```

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

| language | Percentage share of each language |
|---|---|
| English | 12.50 |
| Arabic | 12.50 |
| Persian | 37.50 |
| Hindi | 12.50 |
| French | 12.50 |
| Italian | 12.50 |

**Q4.Syntax used:**
SELECT actor_id, COUNT(*) AS Duplicate_rows FROM job_data
GROUP BY actor_id HAVING COUNT(*) > 1;

```
83 •  SELECT actor_id, COUNT(*) AS Duplicate_rows FROM job_data
84    GROUP BY actor_id HAVING COUNT(*) > 1;
85    # Actor_ID 1003 has duplicate rows.
```

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

| actor_id | Duplicate_rows |
|---|---|
| 1003 | 2 |

**Insights: Actor_ID 1003 has duplicate rows.**

**<u>Case Study 2 (Investigating Metric Spike)</u>**
**Q5. Syntax used:**
SELECT EXTRACT(WEEK FROM occured_at) AS 'Number of week', COUNT(DISTINCT user_id) AS Weekly_Active_Users
FROM events WHERE event_type='engagement' GROUP BY 1;

```
115 •  SELECT EXTRACT(WEEK FROM occured_at) AS 'Number of week', COUNT(DISTINCT user_id) AS Weekly_Active_Users
116    FROM events WHERE event_type='engagement' GROUP BY 1;
117    # The highest week is 30th with 1467 users and the lowest week is 35th with 104 users.
118
```

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

| Number of week | Weekly_Active_Users |
|---|---|
| 17 | 663 |
| 18 | 1068 |
| 19 | 1113 |
| 20 | 1154 |
| 21 | 1121 |
| 22 | 1186 |
| 23 | 1232 |
| 24 | 1275 |
| 25 | 1264 |
| 26 | 1302 |
| 27 | 1372 |
| 28 | 1365 |
| 29 | 1376 |
| 30 | 1467 |
| 31 | 1299 |
| 32 | 1225 |
| 33 | 1225 |
| 34 | 1204 |
| 35 | 104 |

**Insights: The highest week is 30th with 1467 users and the lowest week is 35th with 104 users.**

**Q6.Syntax used:**

```
SELECT
   Months,
   User_count,
   ((User_count / LAG(User_count, 1) OVER (ORDER BY Months)) - 1) * 100 AS Growth_percentage
FROM
   (SELECT EXTRACT(MONTH FROM created_at) AS Months,
         COUNT(*) AS User_count
    FROM users
    WHERE activated_at IS NOT NULL
    GROUP BY 1
    ORDER BY 1) as subquery;
```

```
121 •  SELECT
122        Months,
123        User_count,
124        ((User_count / LAG(User_count, 1) OVER (ORDER BY Months)) - 1) * 100 AS Growth_percentage
125    FROM
126        (SELECT EXTRACT(MONTH FROM created_at) AS Months,
127                COUNT(*) AS User_count
128         FROM users
129         WHERE activated_at IS NOT NULL
130         GROUP BY 1
131         ORDER BY 1) as subquery;
```

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

| Months | User_count | Growth_percentage |
|---|---|---|
| 1 | 712 | NULL |
| 2 | 685 | -3.7921 |
| 3 | 765 | 11.6788 |
| 4 | 907 | 18.5621 |
| 5 | 993 | 9.4818 |
| 6 | 1086 | 9.3656 |
| 7 | 1281 | 17.9558 |
| 8 | 1347 | 5.1522 |
| 9 | 330 | -75.5011 |
| 10 | 390 | 18.1818 |
| 11 | 399 | 2.3077 |
| 12 | 486 | 21.8045 |

**Insights: There was a positive increase in the percentage growth in the users from JAN TO APRIL and then fluctuating.**

**Q7.Syntax used:**

```
SELECT first AS 'Number of weeks',
   SUM(CASE WHEN week_number = 0 THEN 1 ELSE 0 END) AS 'Week 0',
   SUM(CASE WHEN week_number = 1 THEN 1 ELSE 0 END) AS 'Week 1',
   SUM(CASE WHEN week_number = 2 THEN 1 ELSE 0 END) AS 'Week 2',
   SUM(CASE WHEN week_number = 3 THEN 1 ELSE 0 END) AS 'Week 3',
   SUM(CASE WHEN week_number = 4 THEN 1 ELSE 0 END) AS 'Week 4',
   SUM(CASE WHEN week_number = 5 THEN 1 ELSE 0 END) AS 'Week 5',
   SUM(CASE WHEN week_number = 6 THEN 1 ELSE 0 END) AS 'Week 6',
   SUM(CASE WHEN week_number = 7 THEN 1 ELSE 0 END) AS 'Week 7',
   SUM(CASE WHEN week_number = 8 THEN 1 ELSE 0 END) AS 'Week 8',
   SUM(CASE WHEN week_number = 9 THEN 1 ELSE 0 END) AS 'Week 9',
   SUM(CASE WHEN week_number = 10 THEN 1 ELSE 0 END) AS 'Week 10',
   SUM(CASE WHEN week_number = 11 THEN 1 ELSE 0 END) AS 'Week 11',
```

```
   SUM(CASE WHEN week_number = 12 THEN 1 ELSE 0 END) AS 'Week 12',
   SUM(CASE WHEN week_number = 13 THEN 1 ELSE 0 END) AS 'Week 13',
   SUM(CASE WHEN week_number = 14 THEN 1 ELSE 0 END) AS 'Week 14',
   SUM(CASE WHEN week_number = 15 THEN 1 ELSE 0 END) AS 'Week 15',
   SUM(CASE WHEN week_number = 16 THEN 1 ELSE 0 END) AS 'Week 16',
   SUM(CASE WHEN week_number = 17 THEN 1 ELSE 0 END) AS 'Week 17',
   SUM(CASE WHEN week_number = 18 THEN 1 ELSE 0 END) AS 'Week 18'
FROM
   (SELECT a.user_id, a.week_initial, b.first, a.week_initial- b.first AS week_number
    FROM
       (SELECT user_id, EXTRACT(WEEK FROM occured_at) AS week_initial
        FROM events
        GROUP BY 1, 2) a,
       (SELECT user_id, MIN(EXTRACT(WEEK FROM occured_at)) AS first
        FROM events
        GROUP BY 1) b
    WHERE a.user_id = b.user_id) as subquery
GROUP BY first
ORDER BY first;
```

```
135 •   SELECT first AS 'Number of weeks',
136         SUM(CASE WHEN week_number = 0 THEN 1 ELSE 0 END) AS 'Week 0',
137         SUM(CASE WHEN week_number = 1 THEN 1 ELSE 0 END) AS 'Week 1',
138         SUM(CASE WHEN week_number = 2 THEN 1 ELSE 0 END) AS 'Week 2',
139         SUM(CASE WHEN week_number = 3 THEN 1 ELSE 0 END) AS 'Week 3',
140         SUM(CASE WHEN week_number = 4 THEN 1 ELSE 0 END) AS 'Week 4',
141         SUM(CASE WHEN week_number = 5 THEN 1 ELSE 0 END) AS 'Week 5',
142         SUM(CASE WHEN week_number = 6 THEN 1 ELSE 0 END) AS 'Week 6',
143         SUM(CASE WHEN week_number = 7 THEN 1 ELSE 0 END) AS 'Week 7',
144         SUM(CASE WHEN week_number = 8 THEN 1 ELSE 0 END) AS 'Week 8',
145         SUM(CASE WHEN week_number = 9 THEN 1 ELSE 0 END) AS 'Week 9',
146         SUM(CASE WHEN week_number = 10 THEN 1 ELSE 0 END) AS 'Week 10',
147         SUM(CASE WHEN week_number = 11 THEN 1 ELSE 0 END) AS 'Week 11',
148         SUM(CASE WHEN week_number = 12 THEN 1 ELSE 0 END) AS 'Week 12',
149         SUM(CASE WHEN week_number = 13 THEN 1 ELSE 0 END) AS 'Week 13',
150         SUM(CASE WHEN week_number = 14 THEN 1 ELSE 0 END) AS 'Week 14',
151         SUM(CASE WHEN week_number = 15 THEN 1 ELSE 0 END) AS 'Week 15',
152         SUM(CASE WHEN week_number = 16 THEN 1 ELSE 0 END) AS 'Week 16',
153         SUM(CASE WHEN week_number = 17 THEN 1 ELSE 0 END) AS 'Week 17',
154         SUM(CASE WHEN week_number = 18 THEN 1 ELSE 0 END) AS 'Week 18'
155     FROM
156         (SELECT a.user_id, a.week_initial, b.first, a.week_initial- b.first AS week_number
157          FROM
158              (SELECT user_id, EXTRACT(WEEK FROM occured_at) AS week_initial
159               FROM events
160               GROUP BY 1, 2) a,
161              (SELECT user_id, MIN(EXTRACT(WEEK FROM occured_at)) AS first
162               FROM events
163               GROUP BY 1) b
164          WHERE a.user_id = b.user_id) as subquery
165     GROUP BY first
166     ORDER BY first;
```

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

| Number of weeks | Week 0 | Week 1 | Week 2 | Week 3 | Week 4 | Week 5 | Week 6 | Week 7 | Week 8 | Week 9 | Week 10 | Week 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 17 | 663 | 472 | 324 | 251 | 205 | 187 | 167 | 146 | 145 | 145 | 136 | 131 |
| 18 | 596 | 362 | 261 | 203 | 168 | 147 | 144 | 127 | 113 | 122 | 106 | 118 |
| 19 | 427 | 284 | 173 | 153 | 114 | 95 | 91 | 81 | 95 | 82 | 68 | 65 |
| 20 | 358 | 223 | 165 | 121 | 91 | 72 | 63 | 67 | 63 | 65 | 67 | 41 |
| 21 | 317 | 187 | 131 | 91 | 74 | 63 | 75 | 72 | 58 | 48 | 45 | 39 |
| 22 | 326 | 224 | 150 | 107 | 87 | 73 | 63 | 60 | 55 | 48 | 41 | 39 |
| 23 | 328 | 219 | 138 | 101 | 90 | 79 | 69 | 61 | 54 | 47 | 35 | 30 |
| 24 | 339 | 205 | 143 | 102 | 81 | 63 | 65 | 61 | 38 | 39 | 29 | 0 |
| 25 | 305 | 218 | 139 | 101 | 75 | 63 | 50 | 46 | 38 | 35 | 2 | 0 |
| 26 | 288 | 181 | 114 | 83 | 73 | 55 | 47 | 43 | 29 | 0 | 0 | 0 |
| 27 | 292 | 199 | 121 | 106 | 68 | 53 | 40 | 36 | 1 | 0 | 0 | 0 |
| 28 | 274 | 194 | 114 | 69 | 46 | 30 | 28 | 3 | 0 | 0 | 0 | 0 |
| 29 | 270 | 186 | 102 | 65 | 47 | 40 | 1 | 0 | 0 | 0 | 0 | 0 |
| 30 | 294 | 202 | 121 | 78 | 53 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 31 | 215 | 145 | 76 | 57 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 32 | 267 | 188 | 94 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 33 | 286 | 202 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 34 | 279 | 44 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 35 | 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Result Grid, Form Editor, Field Types, Query Stats

| Week 12 | Week 13 | Week 14 | Week 15 | Week 16 | Week 17 | Week 18 |
|---|---|---|---|---|---|---|
| 132 | 143 | 116 | 91 | 82 | 77 | 5 |
| 127 | 110 | 97 | 85 | 67 | 4 | 0 |
| 63 | 42 | 51 | 49 | 2 | 0 | 0 |
| 40 | 33 | 40 | 0 | 0 | 0 | 0 |
| 35 | 28 | 2 | 0 | 0 | 0 | 0 |
| 31 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |

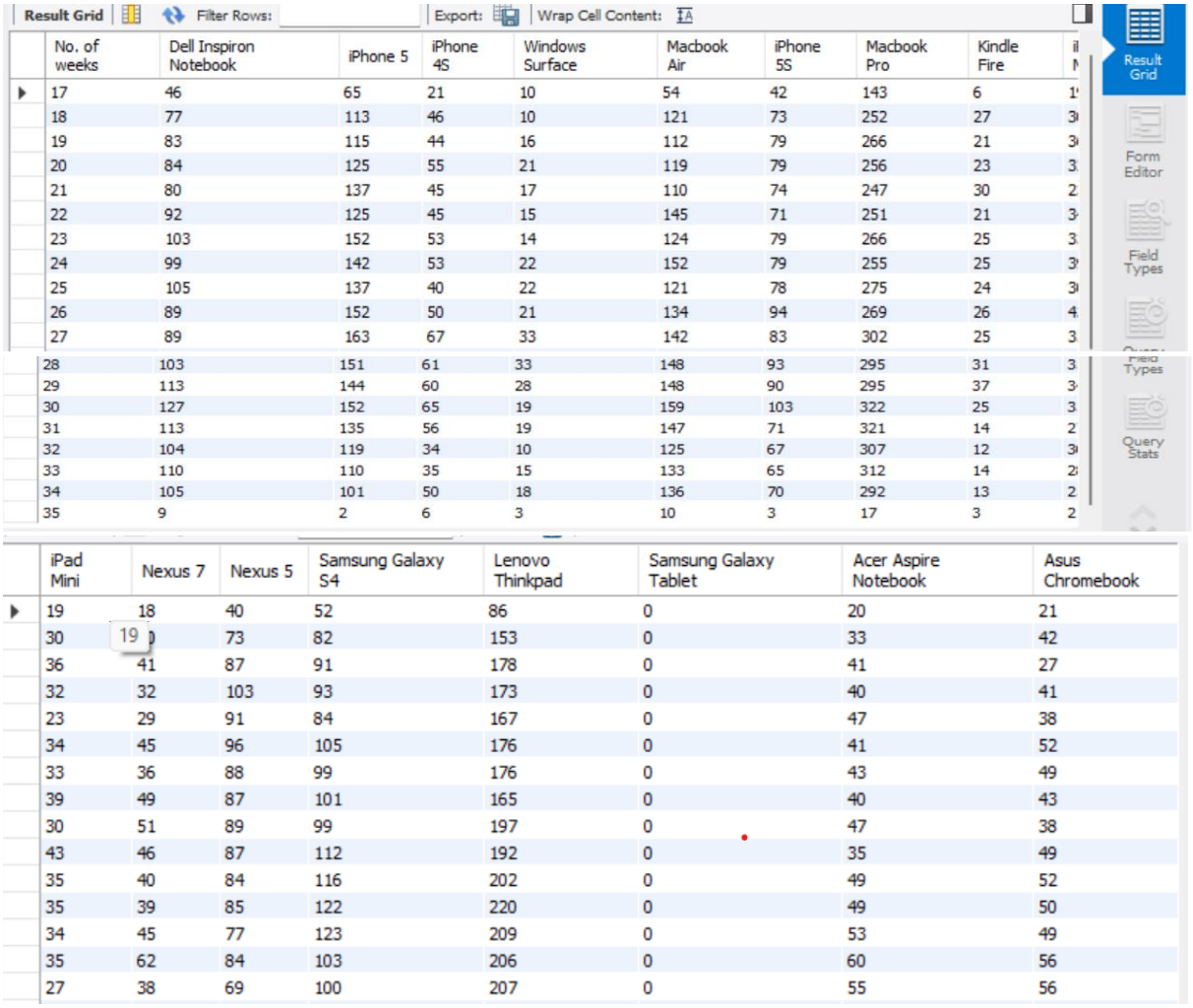
Result Grid, Form Editor, Field Types, Query Stats

**Insights: It is observed that once the customers sign-up there is a drastic drop in the weekly retention of customers. Necessary and effective strategies should be adopted to keep to customers engaged.**

**Q8.Syntax used:**

```
Select EXTRACT(WEEK FROM occured_at) AS "No. of weeks",
   COUNT(DISTINCT CASE WHEN device IN ('dell inspiron notebook') THEN user_id ELSE NULL END) AS "Dell Inspiron Notebook",
   COUNT(DISTINCT CASE WHEN device IN ('iphone 5') THEN user_id ELSE NULL END) AS "iPhone 5",
   COUNT(DISTINCT CASE WHEN device IN ('iphone 4s') THEN user_id ELSE NULL END) AS "iPhone 4S",
   COUNT(DISTINCT CASE WHEN device IN ('windows surface') THEN user_id ELSE NULL END) AS "Windows Surface",
   COUNT(DISTINCT CASE WHEN device IN ('macbook air') THEN user_id ELSE NULL END) AS "Macbook Air",
   COUNT(DISTINCT CASE WHEN device IN ('iphone 5s') THEN user_id ELSE NULL END) AS "iPhone 5S",
   COUNT(DISTINCT CASE WHEN device IN ('macbook pro') THEN user_id ELSE NULL END) AS "Macbook Pro",
   COUNT(DISTINCT CASE WHEN device IN ('kindle fire') THEN user_id ELSE NULL END) AS "Kindle Fire",
   COUNT(DISTINCT CASE WHEN device IN ('ipad mini') THEN user_id ELSE NULL END) AS "iPad Mini",
```

```
   COUNT(DISTINCT CASE WHEN device IN ('nexus 7') THEN user_id ELSE NULL END)
AS "Nexus 7",
   COUNT(DISTINCT CASE WHEN device IN ('nexus 5') THEN user_id ELSE NULL END)
AS "Nexus 5",
   COUNT(DISTINCT CASE WHEN device IN ('samsung galaxy s4') THEN user_id ELSE
NULL END) AS "Samsung Galaxy S4",
   COUNT(DISTINCT CASE WHEN device IN ('lenovo thinkpad') THEN user_id ELSE NULL
END) AS "Lenovo Thinkpad",
   COUNT(DISTINCT CASE WHEN device IN ('samsung galaxy tablet') THEN user_id ELSE
NULL END) AS "Samsung Galaxy Tablet",
   COUNT(DISTINCT CASE WHEN device IN ('acer aspire notebook') THEN user_id ELSE
NULL END) AS "Acer Aspire Notebook",
   COUNT(DISTINCT CASE WHEN device IN ('asus chromebook') THEN user_id ELSE
NULL END) AS "Asus Chromebook",
   COUNT(DISTINCT CASE WHEN device IN ('htc one') THEN user_id ELSE NULL END) AS
"HTC One",
   COUNT(DISTINCT CASE WHEN device IN ('nokia lumia 635') THEN user_id ELSE NULL
END) AS "Nokia Lumia 635",
   COUNT(DISTINCT CASE WHEN device IN ('samsung galaxy note') THEN user_id ELSE
NULL END) AS "Samsung Galaxy Note",
   COUNT(DISTINCT CASE WHEN device IN ('acer aspire desktop') THEN user_id ELSE
NULL END) AS "Acer Aspire Desktop",
   COUNT(DISTINCT CASE WHEN device IN ('mac mini') THEN user_id ELSE NULL END)
AS "Mac Mini",
   COUNT(DISTINCT CASE WHEN device IN ('hp pavilion desktop') THEN user_id ELSE
NULL END) AS "HP Pavilion Desktop",
   COUNT(DISTINCT CASE WHEN device IN ('dell inspiron desktop') THEN user_id ELSE
NULL END) AS "Dell Inspiron Desktop",
   COUNT(DISTINCT CASE WHEN device IN ('ipad air') THEN user_id ELSE NULL END) AS
"iPad Air",
   COUNT(DISTINCT CASE WHEN device IN ('amazon fire phone') THEN user_id ELSE
NULL END) AS "Amazon Fire Phone",
   COUNT(DISTINCT CASE WHEN device IN ('nexus 10') THEN user_id ELSE NULL END)
AS "Nexus 10",
   7
FROM events
WHERE event_type = 'engagement'
GROUP BY 1
ORDER BY 1;
```



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

| No. of weeks | Dell Inspiron Notebook | iPhone 5 | iPhone 4S | Windows Surface | Macbook Air | iPhone 5S | Macbook Pro | Kindle Fire | il |
|---|---|---|---|---|---|---|---|---|---|
| 17 | 46 | 65 | 21 | 10 | 54 | 42 | 143 | 6 | 1 |
| 18 | 77 | 113 | 46 | 10 | 121 | 73 | 252 | 27 | 3 |
| 19 | 83 | 115 | 44 | 16 | 112 | 79 | 266 | 21 | 3 |
| 20 | 84 | 125 | 55 | 21 | 119 | 79 | 256 | 23 | 3 |
| 21 | 80 | 137 | 45 | 17 | 110 | 74 | 247 | 30 | 2 |
| 22 | 92 | 125 | 45 | 15 | 145 | 71 | 251 | 21 | 3 |
| 23 | 103 | 152 | 53 | 14 | 124 | 79 | 266 | 25 | 3 |
| 24 | 99 | 142 | 53 | 22 | 152 | 79 | 255 | 25 | 3 |
| 25 | 105 | 137 | 40 | 22 | 121 | 78 | 275 | 24 | 3 |
| 26 | 89 | 152 | 50 | 21 | 134 | 94 | 269 | 26 | 4 |
| 27 | 89 | 163 | 67 | 33 | 142 | 83 | 302 | 25 | 3 |
| 28 | 103 | 151 | 61 | 33 | 148 | 93 | 295 | 31 | 3 |
| 29 | 113 | 144 | 60 | 28 | 148 | 90 | 295 | 37 | 3 |
| 30 | 127 | 152 | 65 | 19 | 159 | 103 | 322 | 25 | 3 |
| 31 | 113 | 135 | 56 | 19 | 147 | 71 | 321 | 14 | 2 |
| 32 | 104 | 119 | 34 | 10 | 125 | 67 | 307 | 12 | 3 |
| 33 | 110 | 110 | 35 | 15 | 133 | 65 | 312 | 14 | 2 |
| 34 | 105 | 101 | 50 | 18 | 136 | 70 | 292 | 13 | 2 |
| 35 | 9 | 2 | 6 | 3 | 10 | 3 | 17 | 3 | 2 |

Result Grid | Form Editor | Field Types | Query Stats

| iPad Mini | Nexus 7 | Nexus 5 | Samsung Galaxy S4 | Lenovo Thinkpad | Samsung Galaxy Tablet | Acer Aspire Notebook | Asus Chromebook |
|---|---|---|---|---|---|---|---|
| 19 | 18 | 40 | 52 | 86 | 0 | 20 | 21 |
| 30 | 19 | 73 | 82 | 153 | 0 | 33 | 42 |
| 36 | 41 | 87 | 91 | 178 | 0 | 41 | 27 |
| 32 | 32 | 103 | 93 | 173 | 0 | 40 | 41 |
| 23 | 29 | 91 | 84 | 167 | 0 | 47 | 38 |
| 34 | 45 | 96 | 105 | 176 | 0 | 41 | 52 |
| 33 | 36 | 88 | 99 | 176 | 0 | 43 | 49 |
| 39 | 49 | 87 | 101 | 165 | 0 | 40 | 43 |
| 30 | 51 | 89 | 99 | 197 | 0 | 47 | 38 |
| 43 | 46 | 87 | 112 | 192 | 0 | 35 | 49 |
| 35 | 40 | 84 | 116 | 202 | 0 | 49 | 52 |
| 35 | 39 | 85 | 122 | 220 | 0 | 49 | 50 |
| 34 | 45 | 77 | 123 | 209 | 0 | 53 | 49 |
| 35 | 62 | 84 | 103 | 206 | 0 | 60 | 56 |
| 27 | 38 | 69 | 100 | 207 | 0 | 55 | 56 |

**Insights:**

**We can observe that the most widely used device for engagement on weekly basis is Macbook Pro followed by Lenovo thinkpad and Macbook Air.**

**Q9. Syntax used:**

```
SELECT Week,
   ROUND(( email_opens / total * 100), 2) AS 'Email Opening Rate',
   ROUND((weekly_digest / total * 100), 2) AS 'Weekly Digest Rate',
   ROUND((email_clickthroughs / total * 100), 2) AS 'Email Clicking Rate',
   ROUND((reengagement_emails / total * 100), 2) AS 'Reengaging Email Rate'
FROM
   (SELECT EXTRACT(WEEK FROM occured_at) AS Week,
         COUNT(CASE WHEN action = 'email_open' THEN user_id END) AS email_opens,
         COUNT(CASE WHEN action = 'sent_weekly_digest' THEN user_id END) AS
weekly_digest,
          COUNT(CASE WHEN action = 'email_clickthrough' THEN user_id END) AS
email_clickthroughs,
```

COUNT(CASE WHEN action = 'sent_reengagement_email' THEN user_id END) AS reengagement_emails,
COUNT(user_id) AS total
FROM email_events
GROUP BY 1) as subquery
GROUP BY 1
ORDER BY 1;

```
SELECT Week,
    ROUND(( email_opens / total * 100), 2) AS 'Email Opening Rate',
    ROUND((weekly_digest / total * 100), 2) AS 'Weekly Digest Rate',
    ROUND((email_clickthroughs / total * 100), 2) AS 'Email Clicking Rate',
    ROUND((reengagement_emails / total * 100), 2) AS 'Reengaging Email Rate'
FROM
    (SELECT EXTRACT(WEEK FROM occured_at) AS Week,
            COUNT(CASE WHEN action = 'email_open' THEN user_id END) AS email_opens,
            COUNT(CASE WHEN action = 'sent_weekly_digest' THEN user_id END) AS weekly_digest,
            COUNT(CASE WHEN action = 'email_clickthrough' THEN user_id END) AS email_clickthroughs,
            COUNT(CASE WHEN action = 'sent_reengagement_email' THEN user_id END) AS reengagement_emails,
            COUNT(user_id) AS total
     FROM email_events
     GROUP BY 1) as subquery
GROUP BY 1
ORDER BY 1;
```

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

| Week | Email Opening Rate | Weekly Digest Rate | Email Clicking Rate | Reengaging Email Rate |
|---|---|---|---|---|
| 17 | 21.28 | 62.32 | 11.39 | 5.01 |
| 18 | 22.24 | 63.45 | 10.49 | 3.83 |
| 19 | 22.67 | 62.16 | 11.13 | 4.04 |
| 20 | 22.64 | 61.62 | 11.43 | 4.31 |
| 21 | 22.82 | 63.52 | 9.97 | 3.69 |
| 22 | 21.56 | 63.59 | 10.66 | 4.19 |
| 23 | 22.34 | 62.39 | 11.18 | 4.09 |
| 24 | 22.92 | 61.61 | 10.99 | 4.48 |
| 25 | 21.79 | 63.77 | 10.54 | 3.90 |
| 26 | 22.22 | 62.99 | 10.61 | 4.18 |
| 27 | 22.49 | 62.24 | 11.37 | 3.90 |
| 28 | 22.48 | 62.92 | 10.77 | 3.83 |
| 29 | 21.71 | 63.98 | 10.51 | 3.79 |
| 30 | 23.24 | 62.29 | 10.59 | 3.88 |
| 31 | 23.25 | 65.27 | 7.66 | 3.82 |
| 32 | 22.85 | 66.59 | 7.14 | 3.42 |
| 33 | 23.10 | 64.73 | 7.91 | 4.26 |
| 34 | 23.91 | 64.33 | 7.67 | 4.08 |
| 35 | 32.28 | 0.00 | 29.92 | 37.80 |

Result 57 | Read Only

**Insights: The email opening rate is around 21.82%, email clicking rate is around 11.15%. The customers are continuously engaged with email services.**

## Conclusion:

Through the project, I have gained a deeper understanding of operational analytics and the importance of investigating metric spikes in identifying areas for improvement within a company. The analysis has provided valuable insights that can help optimize operations, enhance user experiences, and drive business growth.

----------**********----------

# Module4: Hiring Process Analytics


**Description:**

Assuming the role of a lead Data Analyst at a prominent MNC like Google, the company has entrusted you with the task of dissecting the dataset encompassing their previous recruitment endeavours. Your objective is to elucidate meaningful patterns and draw informed conclusions about the efficacy of the company's recruitment process. This necessitates employing statistical analysis techniques, with Excel.

**Tech Stack used**:- Microsoft Excel 365

**Exploratory Data Analysis:-**

The process approach utilized in analysing the given dataset entails several key steps:

**Data Cleaning**: : The initial step involves cleaning the data to eliminate any duplicate or irrelevant entries. Ensuring proper formatting of the data is essential at this stage.

**Outlier Identification**: Next, the dataset is scrutinized to identify any outliers that could significantly influence the analysis. These outliers are then removed to ensure the accuracy of the results.

**Duplicate Rows:**27 rows

## A. Hiring Analysis:


**Insights:-**It is observed that most of the hired candidates hired are males. It is suggested to balance to diversity as the observation can affect the organization negatively. We should try to maintain the GENDER RATIO.

## B. Salary Analysis:


**Insights:-**

- Average salary offered is 49,987
- Average salary offered to the hired candidates is 49,753
- **The Average Salary of Hired Candidates is almost same as that of Offered Salary. This shows that the hiring team is recruiting candidates as per the pre-determined salary bands of the organization.**

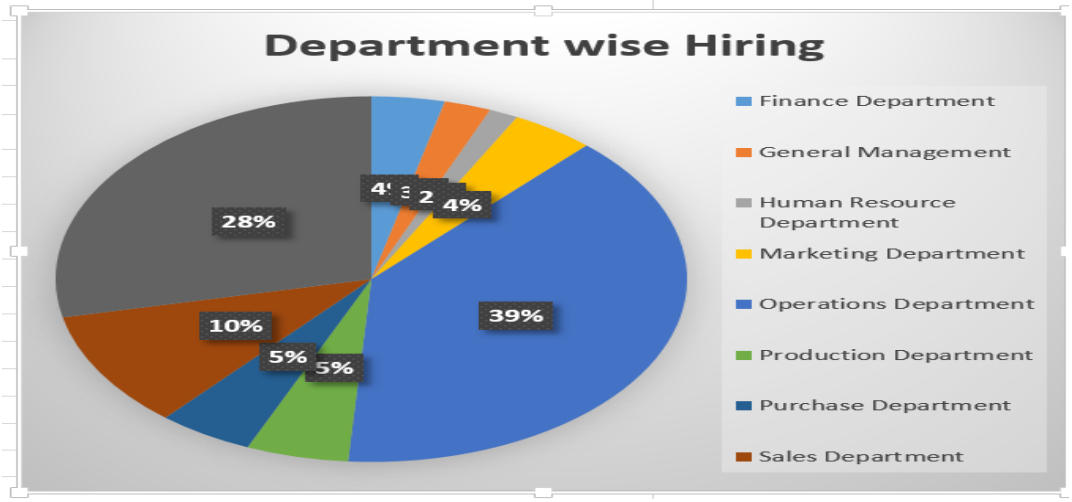
## C. Salary Distribution:

| Salary Range | Salary offered |
|---|---|
| 100-10099 | 686 |
| 10100-20099 | 728 |
| 20100-30099 | 711 |
| 30100-40099 | 713 |
| 40100-50099 | 777 |
| 50100-60099 | 754 |
| 60100-70099 | 698 |
| 70100-80099 | 733 |
| 80100-90099 | 716 |
| 90100-100099 | 649 |
| 190100-200099 | 1 |
| 290100-300099 | 1 |
| 390100-400000 | 1 |
| Total | 7168 |

| Salary Range | Salary offered and hired |
|---|---|
| 100-10099 | 444 |
| 10100-20099 | 487 |
| 20100-30099 | 457 |
| 30100-40099 | 488 |
| 40100-50099 | 523 |
| 50100-60099 | 496 |
| 60100-70099 | 450 |
| 70100-80099 | 479 |
| 80100-90099 | 462 |
| 90100-100099 | 408 |
| 190100-200099 | 1 |
| 290100-300099 | 1 |
| 390100-400000 | 1 |
| Total | 4697 |

**Insights:-**We see that most of 777 number of people fall into the salary range of 40100-50099 and 523 candidates are hired having salary between 40100-50099.

## D. Departmental Analysis:



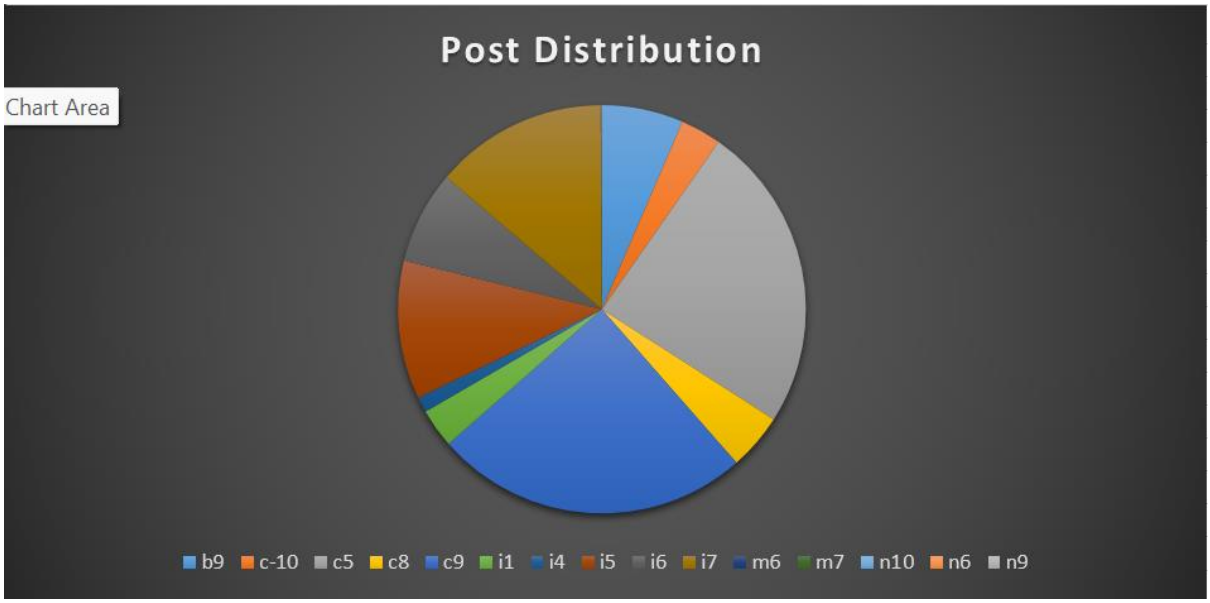

**Insights:-**

From the above pie chart, we can observe that most candidates are hired in Operations Department followed by Services Department and Sales Department and the least candidates are hired in Human Resource Department. We see that Female candidates are majorly hired in Finance and General Department.

E. **Position Tier Analysis:**

| Department/Post Name | b9 | c-10 | c5 | c8 | c9 | i1 | i4 | i5 | i6 | i7 | m6 | m7 | n10 | n6 | n9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Finance Department | 13 | 4 | 68 | 4 | 107 | 9 | 3 | 41 | 12 | 27 | 0 | 0 | 0 | 0 | 0 |
| General Management | 2 | 10 | 29 | 7 | 39 | 1 | 1 | 31 | 9 | 43 | 0 | 0 | 0 | 0 | 0 |
| Human Resource Department | 2 | 2 | 21 | 6 | 7 | 2 | 0 | 42 | 6 | 9 | 0 | 0 | 0 | 0 | 0 |
| Marketing Department | 28 | 18 | 74 | 26 | 70 | 13 | 1 | 30 | 15 | 50 | 0 | 0 | 0 | 0 | 0 |
| Operations Department | 158 | 99 | 671 | 98 | 711 | 94 | 38 | 272 | 278 | 351 | 1 | 0 | 0 | 0 | 0 |
| Production Department | 40 | 8 | 79 | 8 | 87 | 28 | 3 | 37 | 26 | 64 | 0 | 0 | 0 | 0 | 0 |
| Purchase Department | 22 | 5 | 107 | 4 | 74 | 2 | 3 | 36 | 23 | 55 | 0 | 0 | 0 | 1 | 1 |
| Sales Department | 28 | 23 | 216 | 48 | 176 | 2 | 10 | 88 | 43 | 113 | 0 | 0 | 0 | 0 | 0 |
| Service Department | 170 | 63 | 482 | 119 | 522 | 71 | 29 | 210 | 115 | 270 | 2 | 1 | 1 | 0 | 0 |
| **Total** | **463** | **232** | **1747** | **320** | **1793** | **222** | **88** | **787** | **527** | **982** | **3** | **1** | **1** | **1** | **1** |

| Post Name | |
|---|---|
| **b9** | 463 |
| **c-10** | 232 |
| **c5** | 1747 |
| **c8** | 320 |
| **c9** | 1793 |
| **i1** | 222 |
| **i4** | 88 |
| **i5** | 787 |
| **i6** | 527 |
| **i7** | 982 |
| **m6** | 3 |
| **m7** | 1 |
| **n10** | 1 |
| **n6** | 1 |
| **n9** | 1 |
| **Total** | **7168** |




## Insight:-

Here, we can observe that the organization has hired most candidates for post tier c9 followed by c5 and then i7.

### Conclusion:-

I have completed the analysis of the provided dataset in accordance with the questions posed, offering the necessary insights and creating relevant charts and graphs as per the requirements and my interpretation. This project has been instrumental in enhancing my understanding of the Exploratory Data Analysis (EDA) process.

----------**********----------

# Module 5: IMDB Movie Analysis


❖ **Description:**

The main goal of this project is to analyze a movies dataset and identify the factors that contribute to a movie's success. This analysis aims to uncover how various factors such as genre, duration, budget, etc., influence IMDb ratings. The insights gained from this analysis will assist investors, producers, and directors in making data-driven decisions.

❖ **Tech Stack Used:**

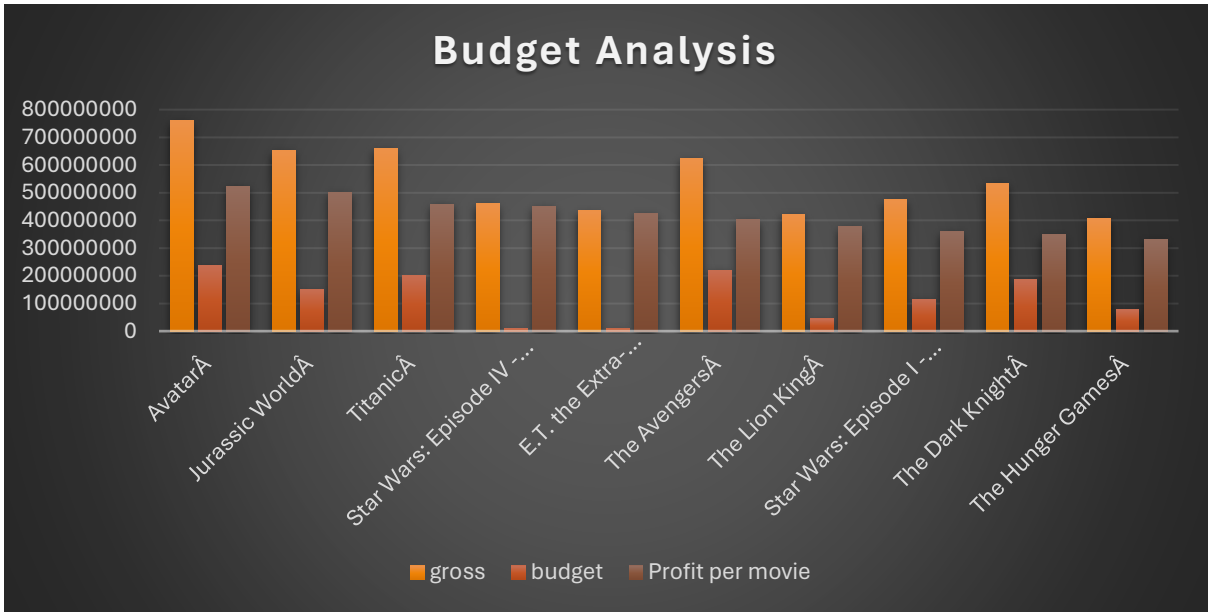
- Microsoft Excel (2019): Used for data analysis and visualization.
- Microsoft Word: Used to create a presentation report summarizing the findings of the analysis.

**Exploratory Data Analysis:**

- Deleting unwanted rows
- 35 Duplicates were removed 3825 rows remains
- Excel sheet is sorted using sort function in Data tab making the sheet in descending order of Profit per movie.
- Languages for 3 movies were missing. Because the country is USA, the blank spaces were replaced by English Language.

## Outlier Found

## Top Movies




**Findings:-**

Correlation Coefficient for Top 10 movies is 0.86

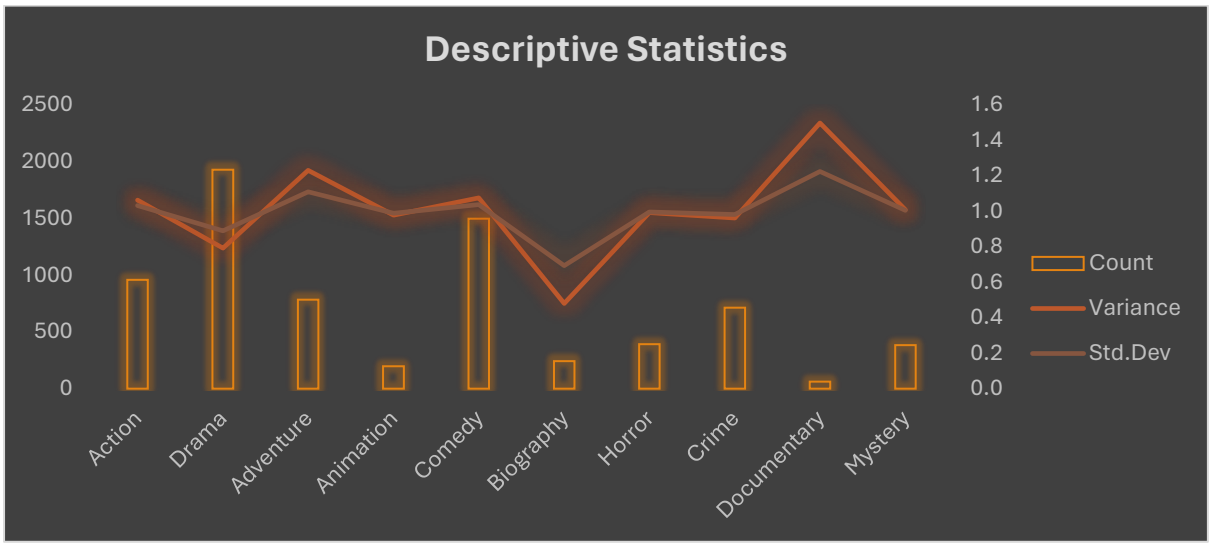
A correlation coefficient of 0.86 indicates a strong positive linear relationship between Gross and Budget

Correlation Coefficient for all the Movies is 0.100

This depicts weak linear relationship between Gross and Budget.

Avatar movie has the highest profit. Star wars and Extra Terrestrial had lowest Budget but made huge profits.

### ❖ Genre Analysis:-




## Findings:-

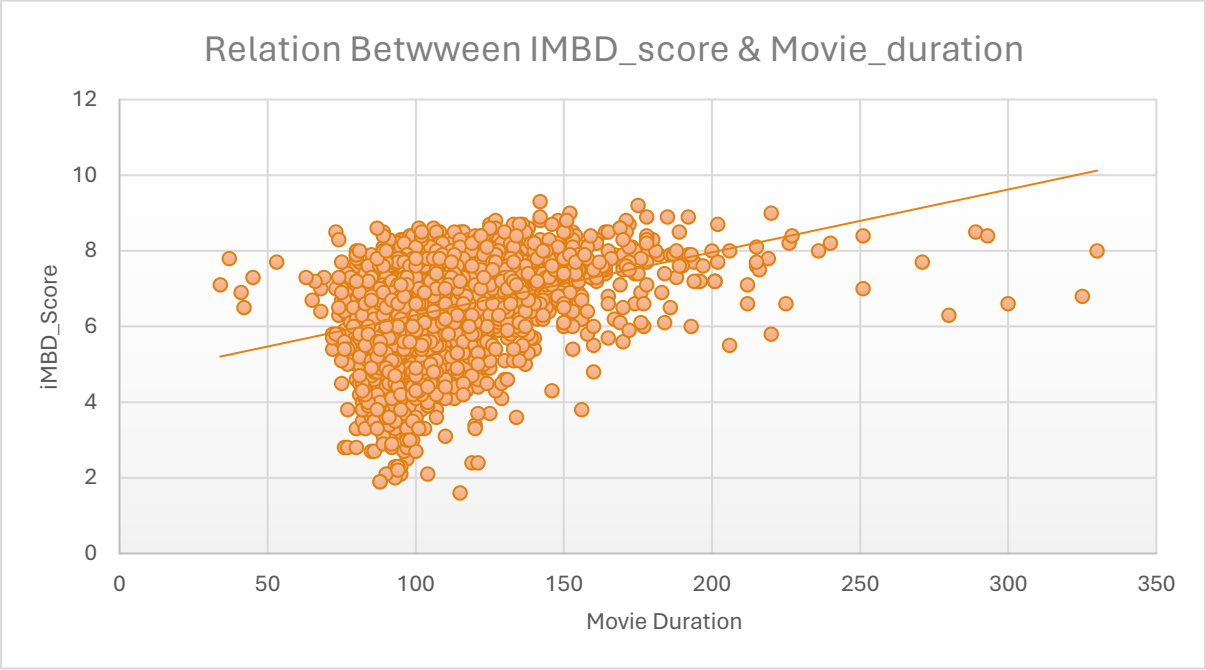
Drama is one of the most popular genre that has appeared 1928 movies with an average imdb rating of 6.8 followed by Comedy with an average imdb rating of 6.2.

## ❖ Movie Duration Analysis:-

| Movie Duration | Values |
|---|---|
| Avg Duration | 109.99 |
| Median | 106.00 |
| Std.Dev | 22.77 |
| Mode | 101.00 |
| Variance | 518.57 |
| Short Movie Duration | 34 |
| Long Movie Duration | 330 |

## Findings:-

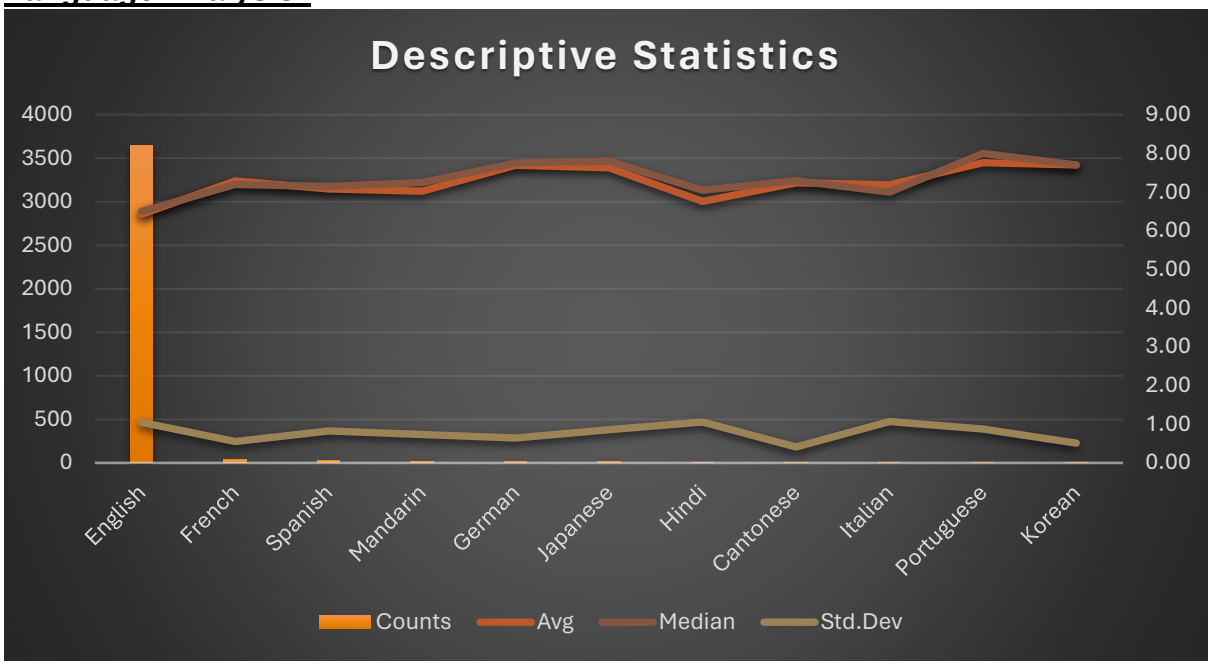
Correlation Between Imbd_Score and Movie Duration close to 0 indicate a weak or no linear relationship.




## Findings:-

Duration of 80 to 130 has got the maximum films and the imdb score lies between 4.5 –8.5. Shortest movie duration is of 34 min whereas longest movie duration is of 330 min.

## ❖ Language Analysis:




## Findings:-

English is the language used in maximum of movies. The avg imbd_score is 6.43.
Persian and Telugu are languages with more than 8 imbd_score.
From the top 10 movies French, Spanish, mandarin, German, Japanese, Cantonese, Italian, Portuguese, Korean have higher imbd_score than English.
This is due to consistent audience because of fewer movies in these languages.
The Average of movie ratings are consistent across languages ranging from 6.4 –7.7

## ❖ Director Analysis:

| Rank | Ranking Directors as per their IMBD_Score | | 90th Percentile |
|---|---|---|---|
| 1 | Tony Kaye | 8.60 | 7.5 |
| 1 | Charles Chaplin | 8.60 | 7.5 |
| 3 | Alfred Hitchcock | 8.50 | 7.5 |
| 3 | Damien Chazelle | 8.50 | 7.5 |
| 3 | Majid Majidi | 8.50 | 7.5 |
| 3 | Ron Fricke | 8.50 | 7.5 |
| 7 | Sergio Leone | 8.43 | 7.5 |
| 8 | Christopher Nolan | 8.43 | 7.5 |
| 9 | Richard Marquand | 8.40 | 7.5 |
| 9 | Asghar Farhadi | 8.40 | 7.5 |

| Directors | Count of Movies | Avg IMBD_Score |
|---|---|---|
| Steven Spielberg | 25 | 7.54 |
| Clint Eastwood | 19 | 7.21 |
| Woody Allen | 19 | 7.00 |
| Ridley Scott | 17 | 7.07 |
| Tim Burton | 16 | 6.93 |
| Steven Soderbergh | 16 | 6.71 |
| Martin Scorsese | 16 | 7.68 |
| Renny Harlin | 15 | 5.75 |
| Spike Lee | 15 | 6.73 |

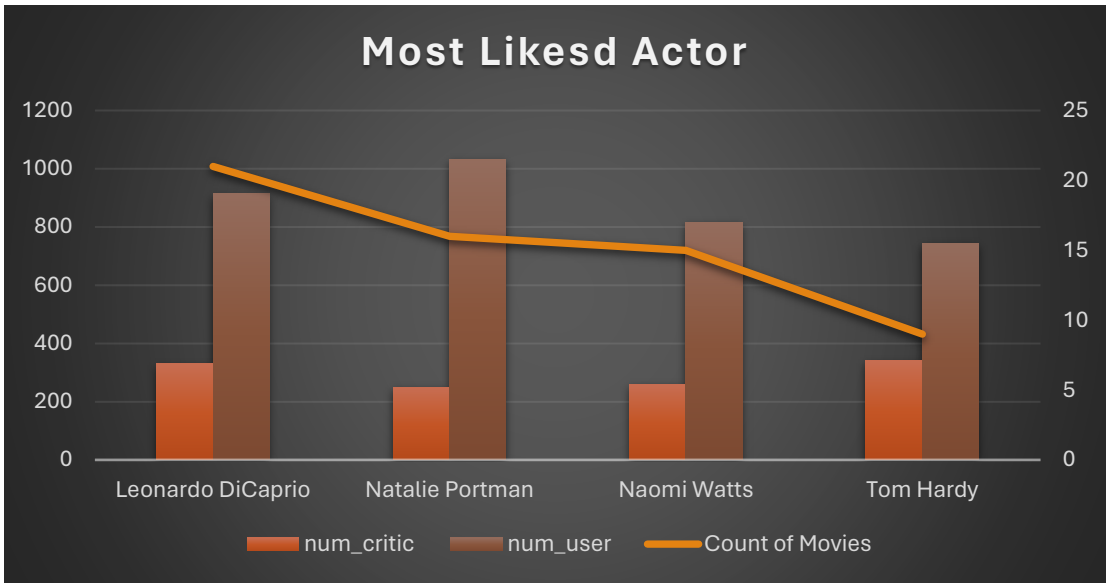
**Findings:-**

Charles Chaplin and Tony Kaye have the highest average IMDb score of 8.60, with only 1 movie.

Steven Spielberg has the highest average imdb ratings of 7.54 for a total of 25 movies indicating a consistent record.

Percentile : Each director's average IMDb score is compared against a common benchmark of 90$^{th}$ percentile, to know their relative position in dataset.

## ❖ Actor_1_Analysis:-




## Findings:-

From the Actor_1 Analysis we can understand that **Leonardo DiCaprio** is Critics and Users Favourite Actor.

## ❖ Conclusion:-

In this IMDb Movie Analysis project, I've developed various logical, statistical, and technical skills to derive meaningful insights from the dataset. Concepts such as calculating averages, creating frequency tables, and identifying outliers have enabled me to deepen my understanding of the data and enhance my ability to analyze it effectively.

----------**********----------

# Module 6: Bank Loan Case Study


## ❖ Description:-

This project aims to analyse the risk appetite of banks when deciding whether to approve loan applications based on applicant profiles. There are two primary risks associated with these decisions:

1. **Risk of Not Approving a Loan to a Creditworthy Applicant**:
   If a bank rejects a loan application from an applicant who would have repaid the loan successfully, the bank loses potential business.
2. **Risk of Approving a Loan to a Default-Prone Applicant**:
   If a bank approves a loan for an applicant who is likely to default, it could lead to financial losses for the bank.

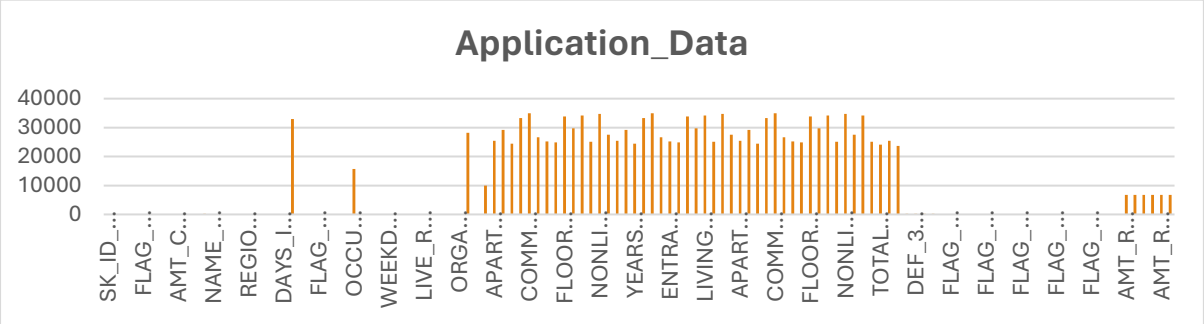
## ❖ Tech Stack Used:- Microsoft Excel 365

## ❖ Exploratory Data Analysis:-

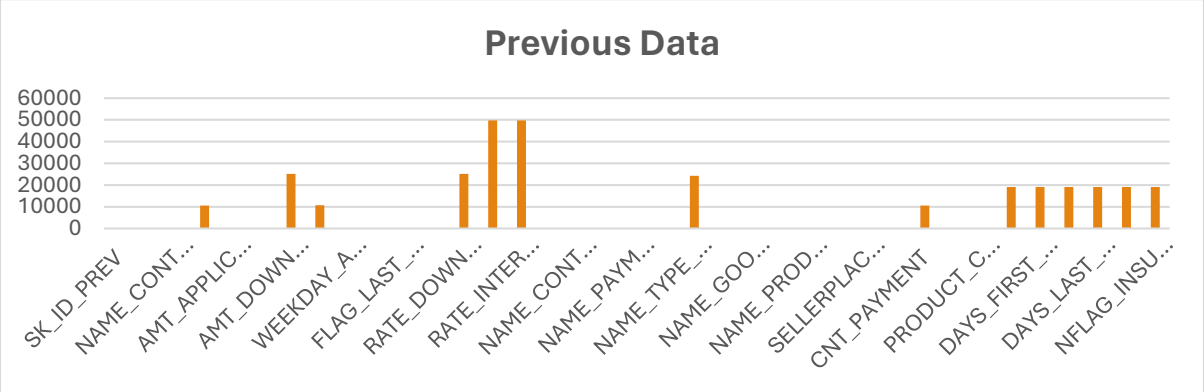
1. **Data Preprocessing**:
   - Counted total rows in each column using **COUNTA** function.
   - Calculated null value percentages for each column.
   - Removed columns with null value percentages exceeding 35%.
2. **Handling Missing Values**:
   - Imputed missing values (less than 35% null) using mean, median, or mode based on column characteristics.
3. **Outlier Detection**:
   - Identified outliers using interquartile range (IQR) method and BOX_PLOT for relevant columns.
4. **Data Transformation**:
   - Converted columns with day values into years by dividing days by 365.

## Task A:

**Application_Data:**

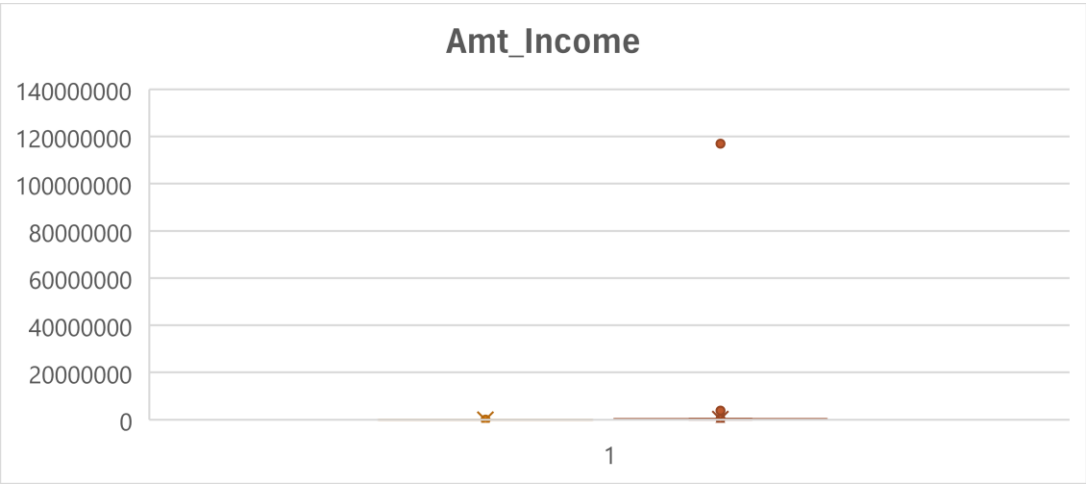


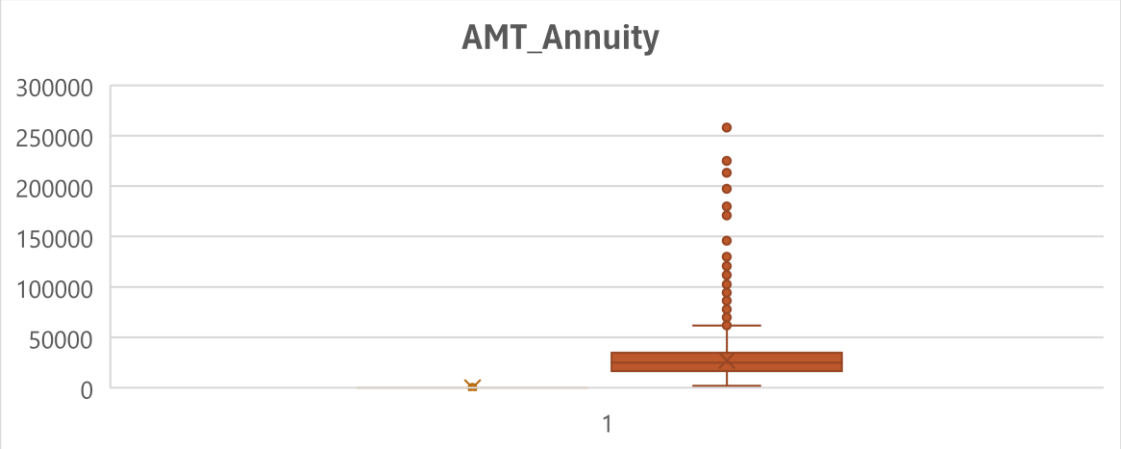

**Previous_Data:**

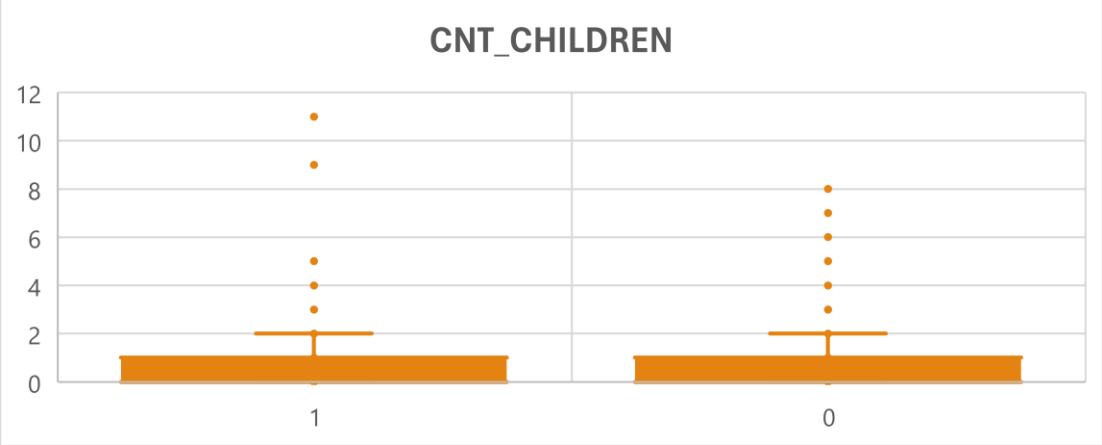


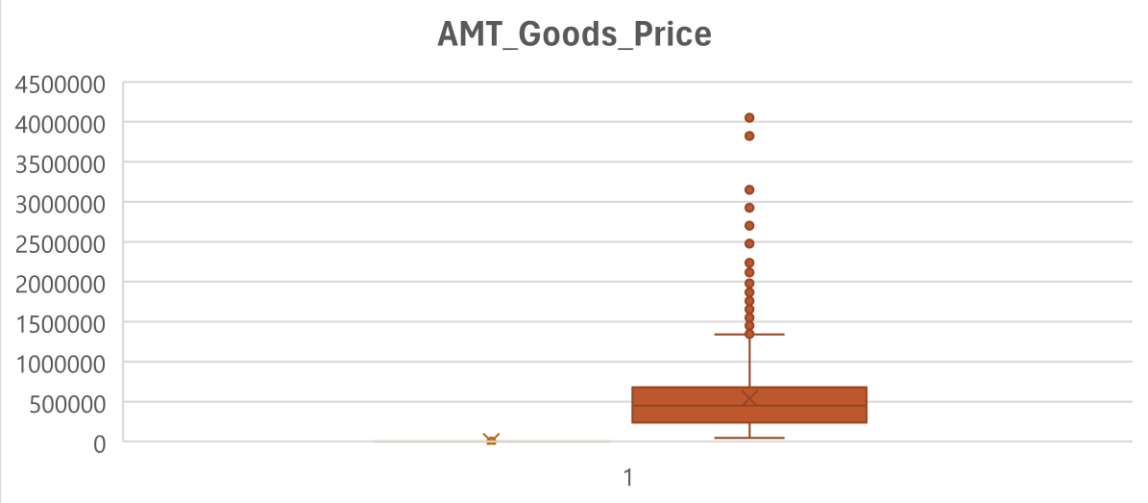

## Task B:

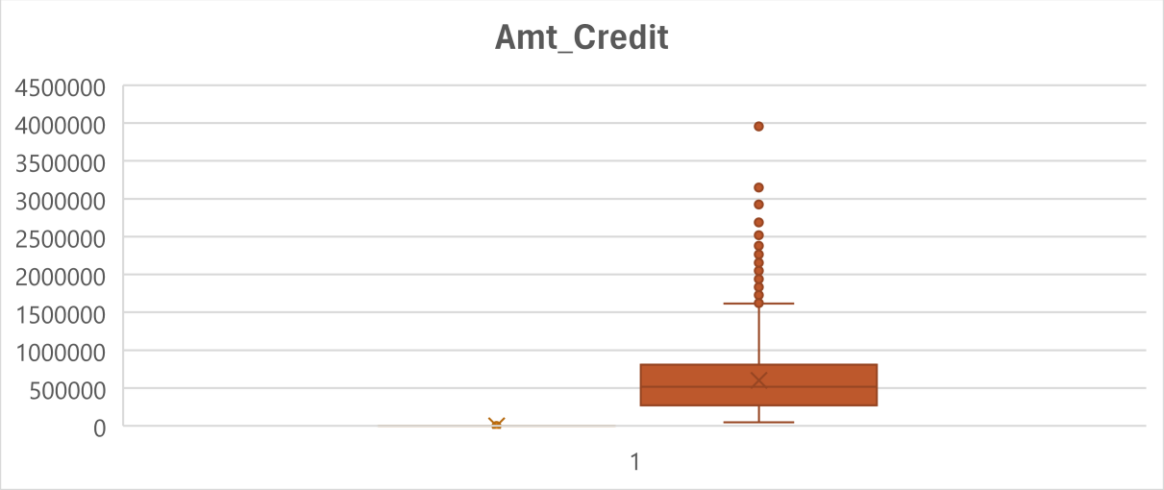
**Identify Outliers in the Dataset:**

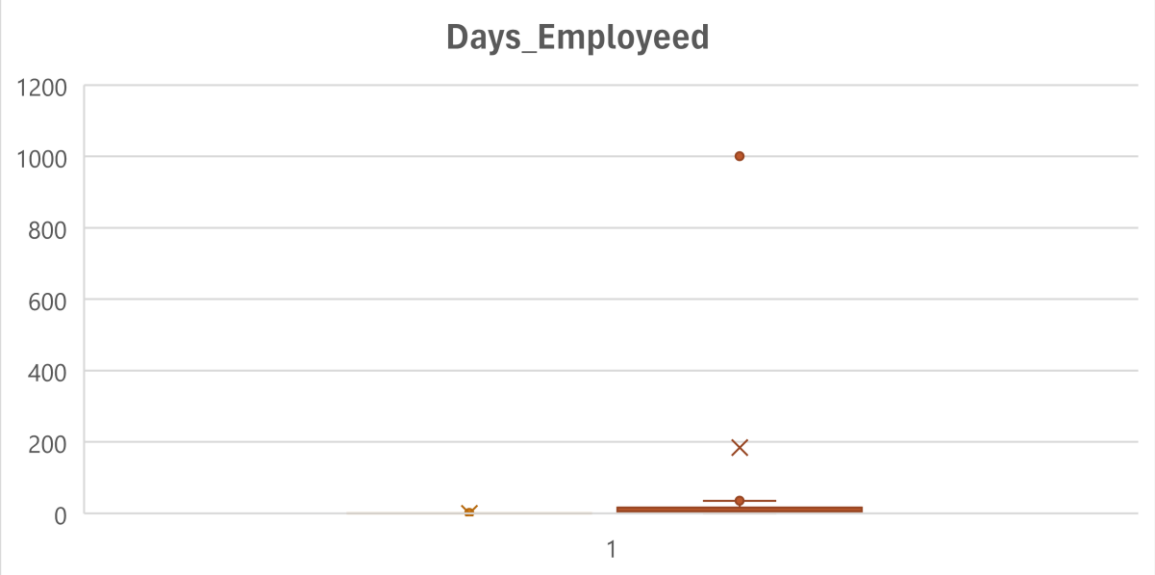


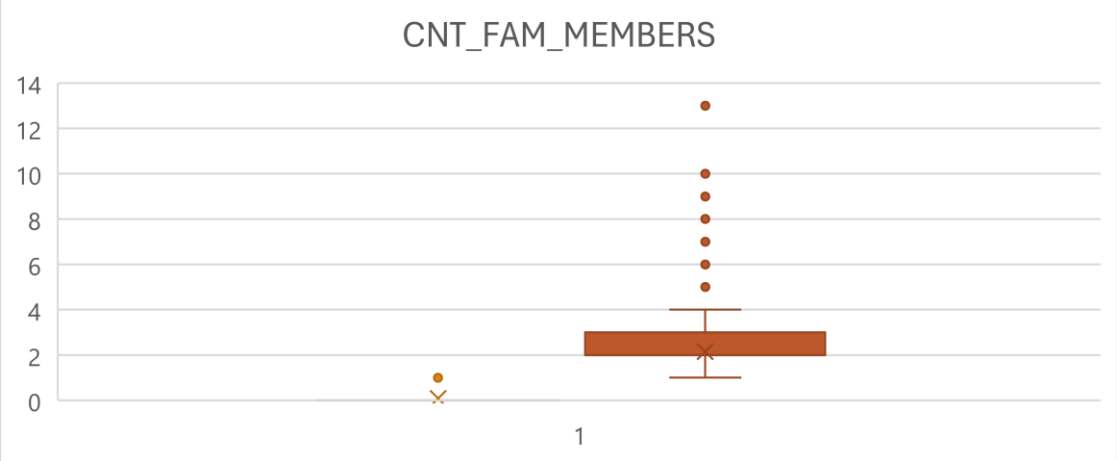

AMT_Annuity
300000
250000
200000
150000
100000
50000
0
1


CNT_CHILDREN
12
10
8
6
4
2
0
1
0


AMT_Goods_Price
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1


Amt_Credit
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1


Days_Employeed
1200
1000
800
600
400
200
0
1


CNT_FAM_MEMBERS
14
12
10
8
6
4
2
0
1

## Task C:

| Target | Count of Males | Count of females | Count of XNA | Total |
|---|---|---|---|---|
| 1 | 1762 | 2264 | 0 | 4026 |
| 0 | 15412 | 30559 | 2 | 45973 |
| Total | 17174 | 32823 | 2 | 49999 |

- Count of Female Candidates is more than the male candidates

| Target | Flag own cars(Y) | Flag own Cars(N) | Total |
|---|---|---|---|
| 1 | 1253 | 2773 | 4026 |
| 0 | 15797 | 30176 | 45973 |
| Total | 17050 | 32949 | 49999 |

- Clients applied for loans do not Own Cars.

| Target | Flag Own Realty(Y) | Flag Own Realty(N) | Total |
|---|---|---|---|
| 1 | 2752 | 1274 | 4026 |
| 0 | 31939 | 14034 | 45973 |
| Total | 34691 | 15308 | 49999 |

- We observe that Clients owning Properties/Realty/Estate has applied for Loans

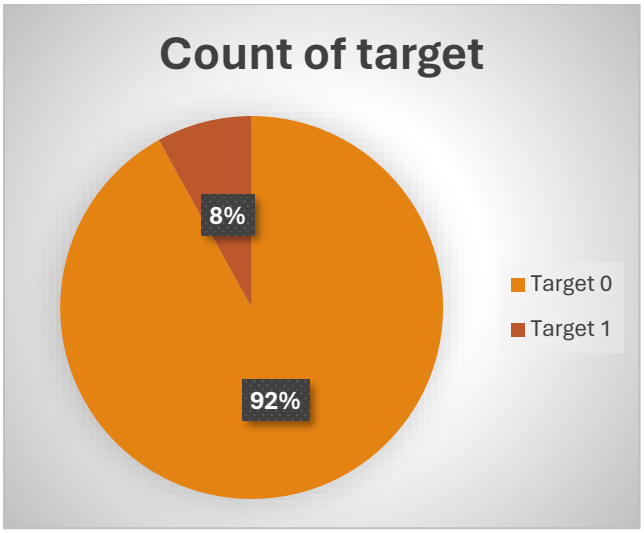
| Target | Cash Loans | Revolving Loans | Total |
|---|---|---|---|
| Target | Cash Loans | Revolving Loans | Total |
| 1 | 3792 | 234 | 4026 |
| 0 | 41484 | 4489 | 45973 |
| Total | 45276 | 4723 | 49999 |

- Clients have applied for Cash loans more than revolving loans

| Target | Count of Children |
|---|---|
| 1 | 4026 |
| 0 | 45973 |
| Total | 49999 |

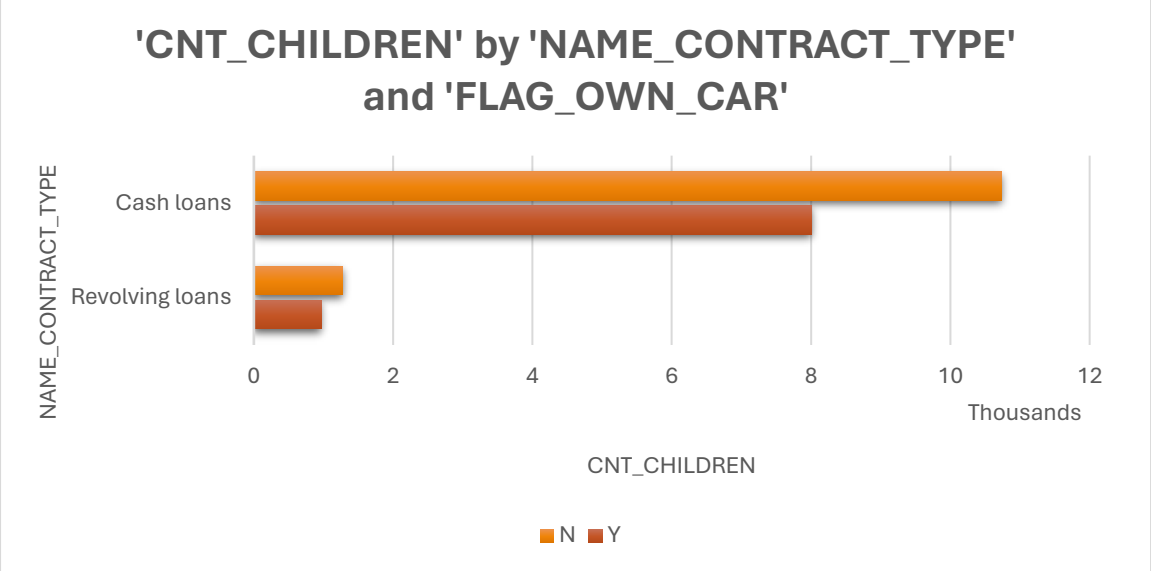
- Count of children of Clients with payment difficulties are lesser than normal Payers.

| Target | Count of target |
|---|---|
| Target 0 | 45973 |
| Target 1 | 4026 |
| Total | 49999 |




Ratio of Data Imbalance = 45973:4026≈11.42

| Sum of CNT_CHILDREN | FLAG_OWN_CAR | | |
|---|---|---|---|
| NAME_CONTRACT_TYPE | N | Y | Grand Total |
| Cash loans | 10740 | 8006 | 18746 |
| Revolving loans | 1274 | 972 | 2246 |
| **Grand Total** | **12014** | **8978** | **20992** |



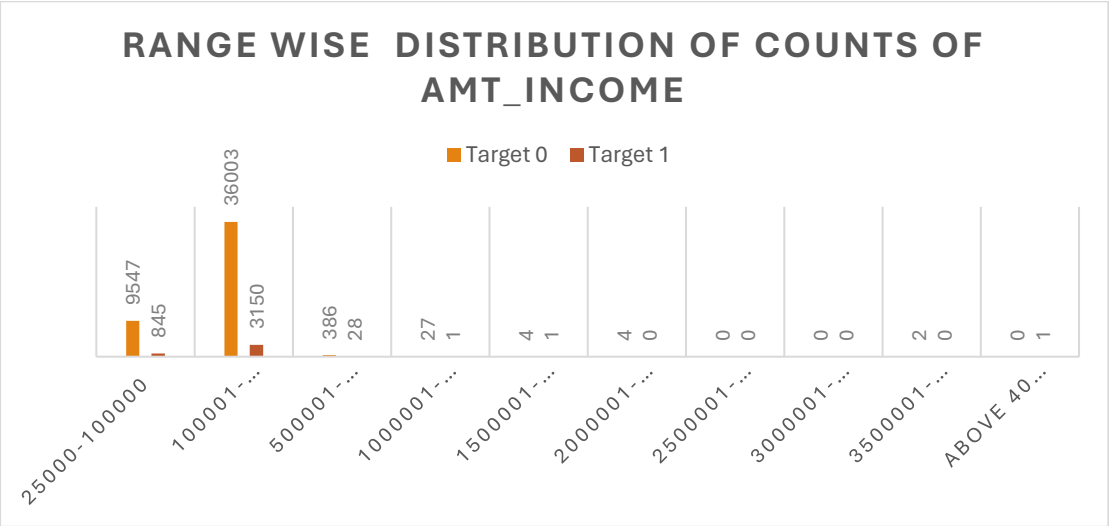

- The Data seems to be very Imbalanced and is more inclined towards the target Variable 0. Cash Loans are disbursed more and More in demand than Revolving loans.

## Task D:

### Application_data

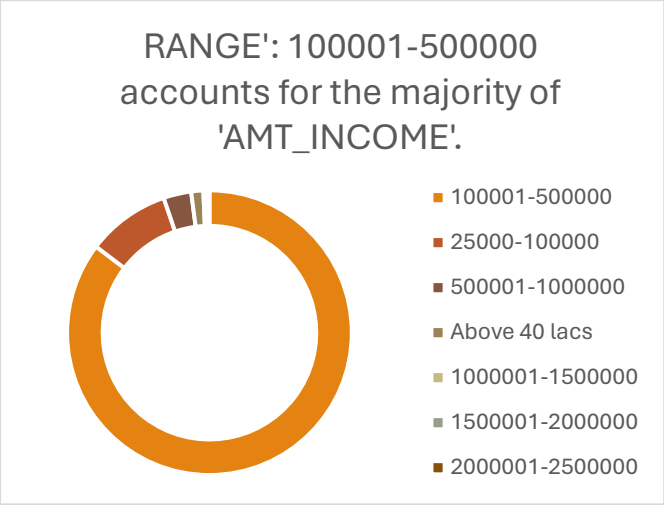
Count of target 0 and Target 1 as per the Income.

| Amount_Income_Total | Target 0 | Target 1 |
|---|---|---|
| 25000-100000 | 9547 | 845 |
| 100001-500000 | 36003 | 3150 |
| 500001-1000000 | 386 | 28 |
| 1000001-1500000 | 27 | 1 |
| 1500001-2000000 | 4 | 1 |
| 2000001-2500000 | 4 | 0 |
| 2500001-3000000 | 0 | 0 |
| 3000001-3500000 | 0 | 0 |
| 3500001-4000000 | 2 | 0 |
| Above 40 lacs | 0 | 1 |

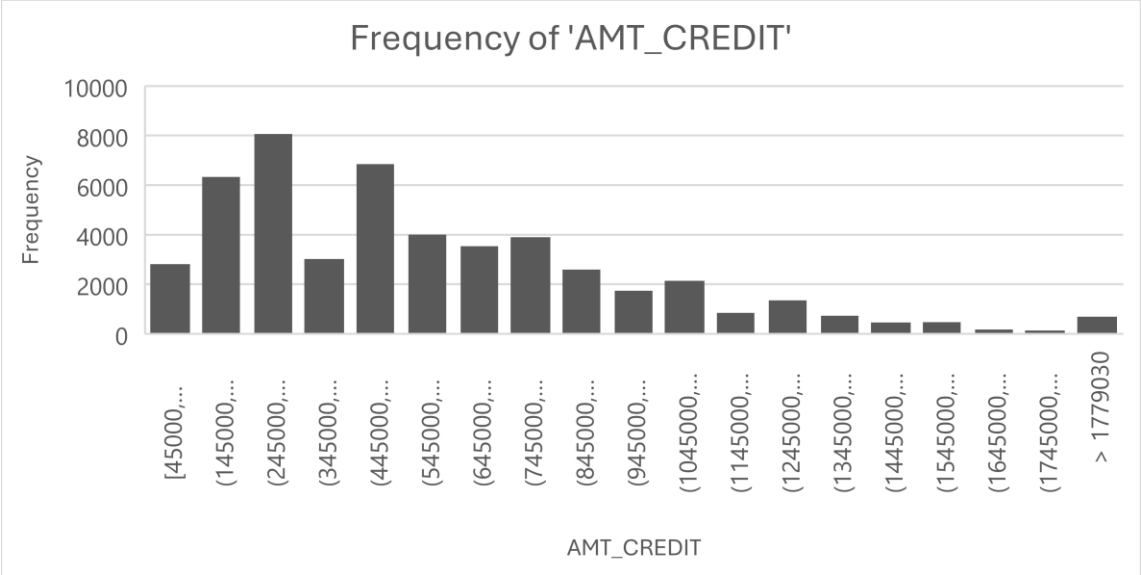



We can see that Clients who apply for Loans fall into the category of 100001-500000, and their loans get approval with target 0.

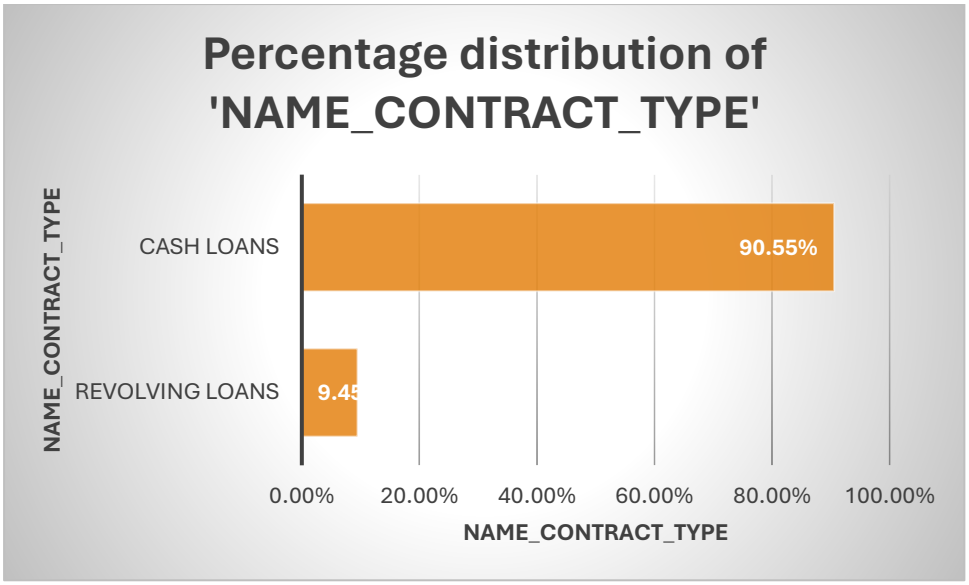
| RANGE | Sum of AMT_INCOME |
|---|---|
| 100001-500000 | 7286100131 |
| 25000-100000 | 807168449.7 |
| 500001-1000000 | 269327677.5 |
| Above 40 lacs | 117000000 |
| 1000001-1500000 | 33682500 |
| 1500001-2000000 | 8955000 |
| 2000001-2500000 | 8550000 |
| 3500001-4000000 | 7425000 |
| **Grand Total** | **8538208758** |



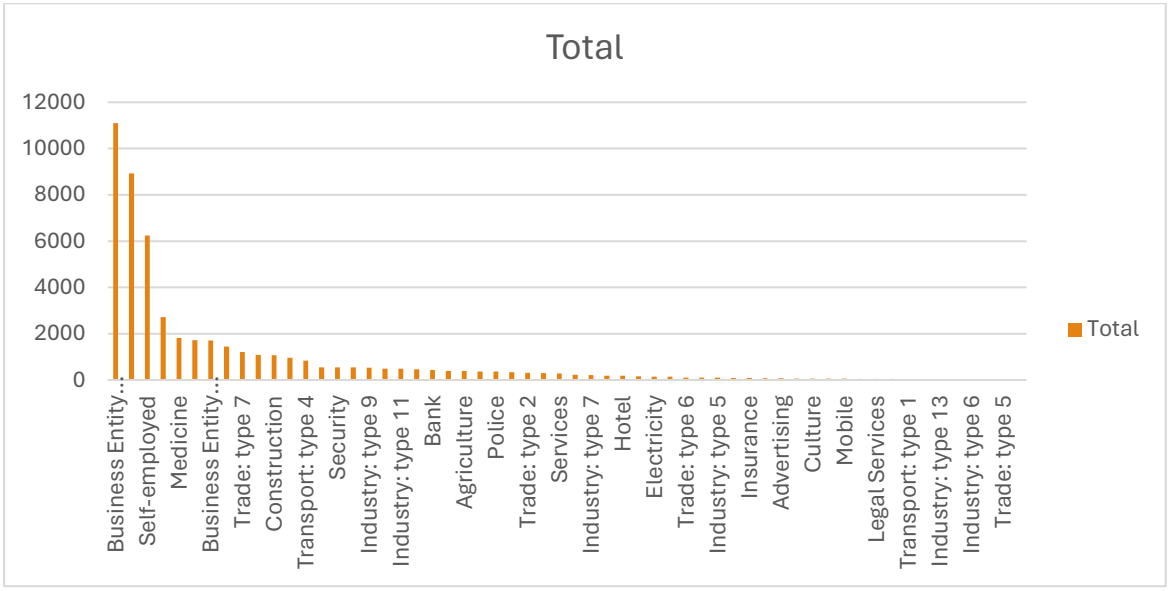

- Clients having Income Between 100001-500000 has applied for Loans. But only 36003 got the Loans approved.




- Clients have received the max amount credit between 245000-345000. As their AMT_Income falls in criteria 100001-500000.

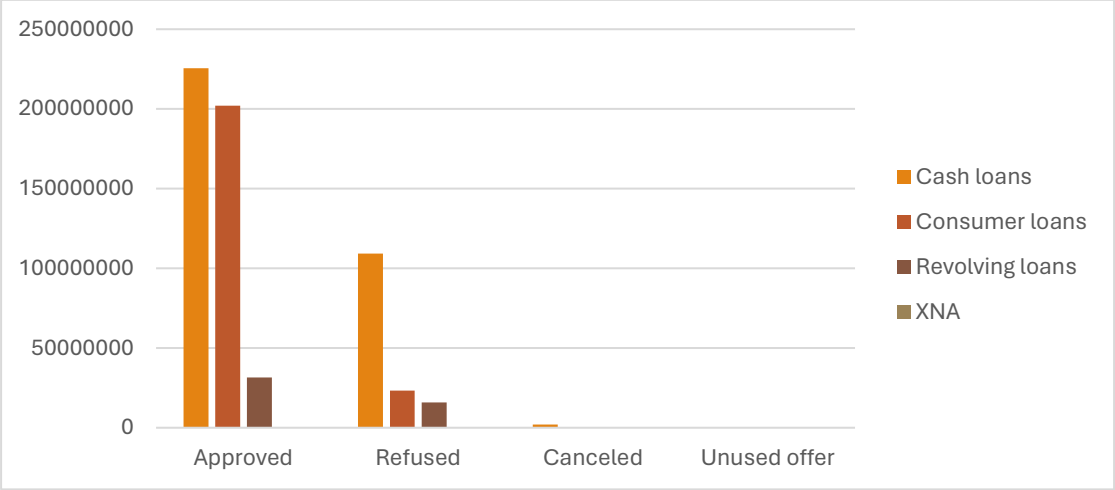



- 90.55% Loans applied for are Cash loans.

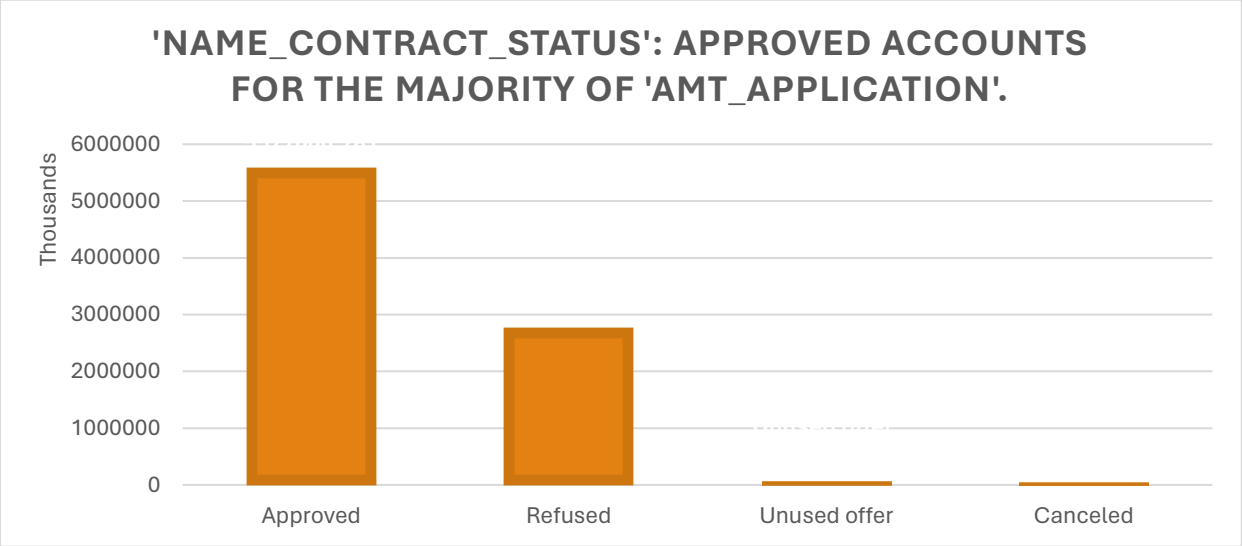


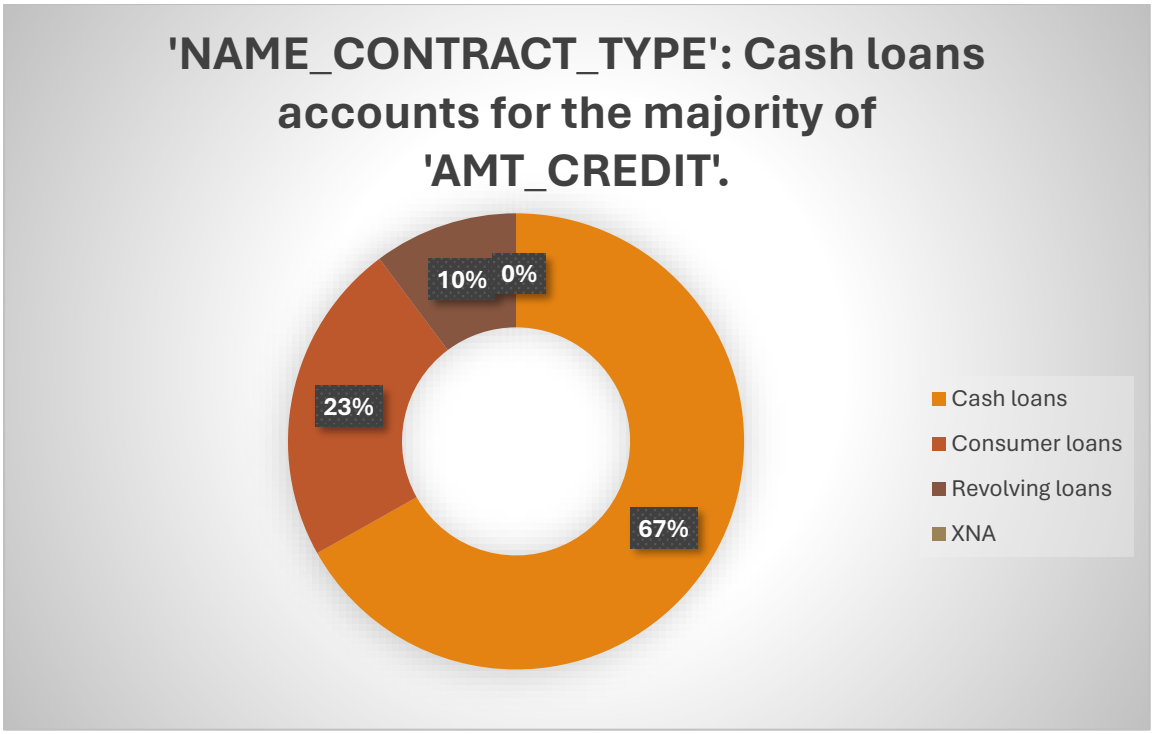

- Clients applying more for Loans belong to business Entity Type 3.

**Previous data:-**

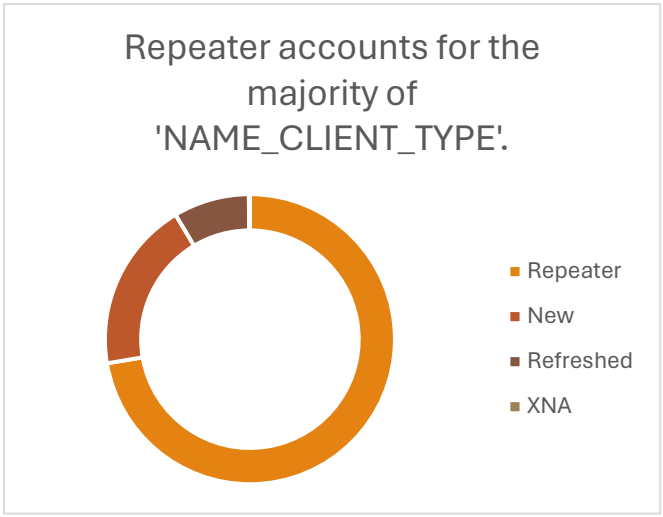



- The amount of loans which is approved is a Cash Loan and Consumer Loans.



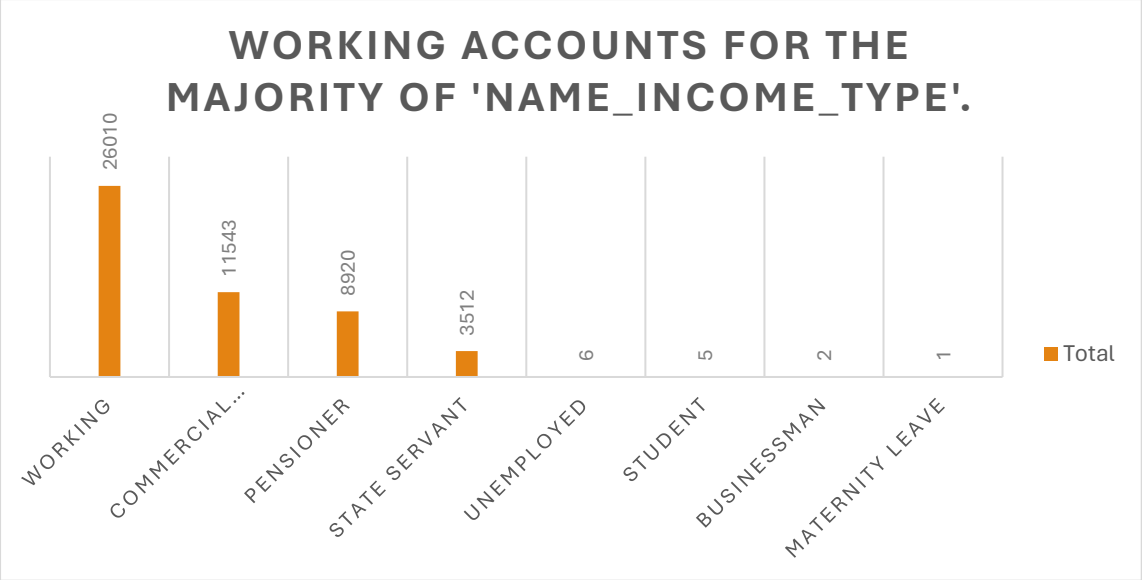



- Cash Loans are majorly approved than other Loans in AMT_CREDIT

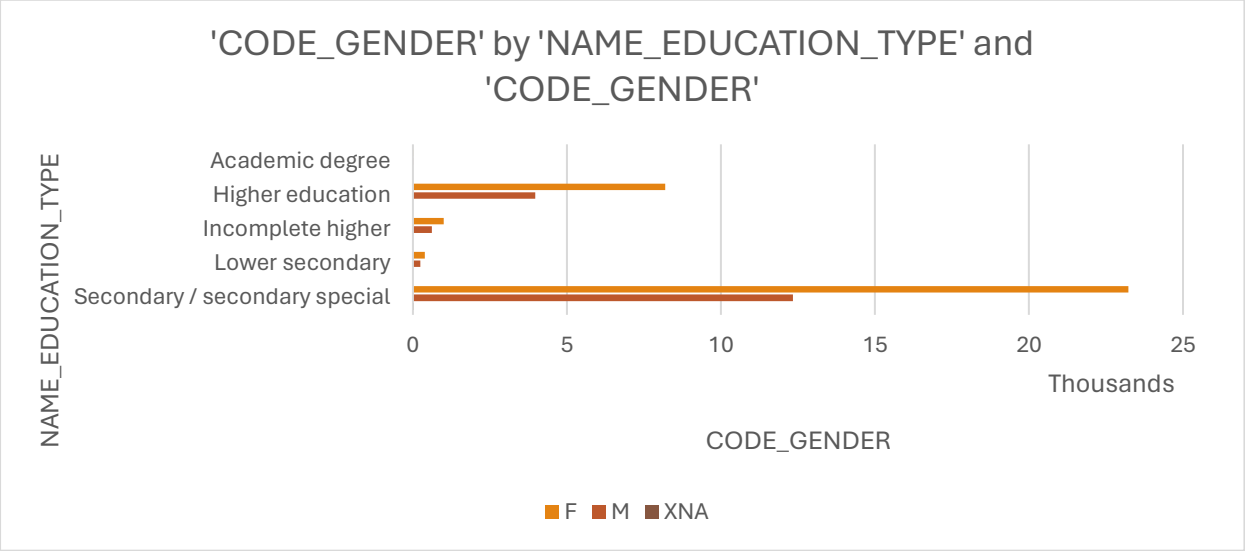



- Majority of the Clients who have already taken loans earlier(Repeater) have again applied for Loans, followed by new ones.

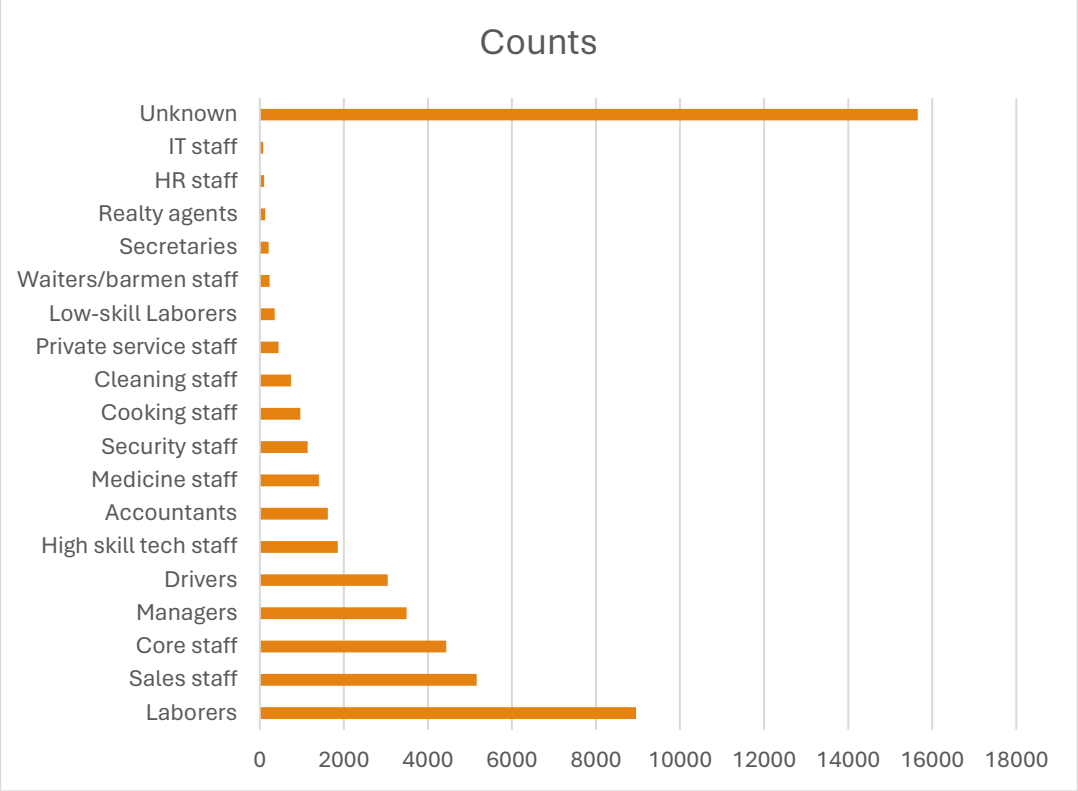
## Multi-Variate Analysis




- 26010 Working Professional are the one applying more for cash loans followed by Commercials and then Pensioners.




- Females who have complete their secondary education have applied for most of the cash loans. Followed by males completed their secondary education. Clients completing higher education are on rank 2.




- It is observed that clients whose occupation is unknown have the major counts for applying loans followed by Laborers. So we interpret that the clients who did not mention their Occupation or Laborers are either completed their Secondary or Higher Education.

Descriptive Statistics:-

| Descriptive Statistics | AMT_INCOME_TOTAL | AMT_CREDIT | CNT_Children | AMT_ANNUITY | AMT_GOODS_PRICE | Days_Employed |
|---|---|---|---|---|---|---|
| Mean | 170767.5905 | 599700.5815 | 0.419848397 | 27107.37739 | 538992.3491 | 184.0008887 |
| Median | 145800 | 514777.5 | 0 | 24939 | 450000 | 6.071232877 |
| Mode | 135000 | 450000 | 0 | 9000 | 450000 | 1000.665753 |
| Standard Deviation | 531813.7768 | 402411.4096 | 0.724031307 | 14562.65317 | 369717.1252 | 380.7027603 |
| Variance | 282825893198.07 | 161934942605.41 | 0.524221333 | 212070867.3 | 136690752694.68 | 144934.5917 |
| Minimum | 25650 | 45000 | 0 | 2052 | 45000 | 0 |
| Maximum | 117000000 | 4050000 | 11 | 258025.5 | 4050000 | 1000.665753 |
| Range | 116974350 | 4005000 | 11 | 255973.5 | 4005000 | 1000.665753 |
| Sum | 8538208758 | 29984429376 | 20992 | 1355341762 | 26949078465 | 9199860.436 |
| Count | 49999 | 49999 | 49999 | 49999 | 49999 | 49999 |
| Quartile 1 | 112500 | 270000 | 0 | 16456.5 | 238500 | 2.556164384 |
| Quartile 3 | 202500 | 808650 | 1 | 34596 | 679500 | 15.66575342 |
| IQR | 90000 | 538650 | 1 | 18139.5 | 441000 | 13.10958904 |

- It is observed that, in most of the columns Mean is greater than Median which means the data is Right skewed or positively skewed. It is also observed that there is a huge difference between Quartile3(75%) and Maximum, which means there are Outliers are present in the data.

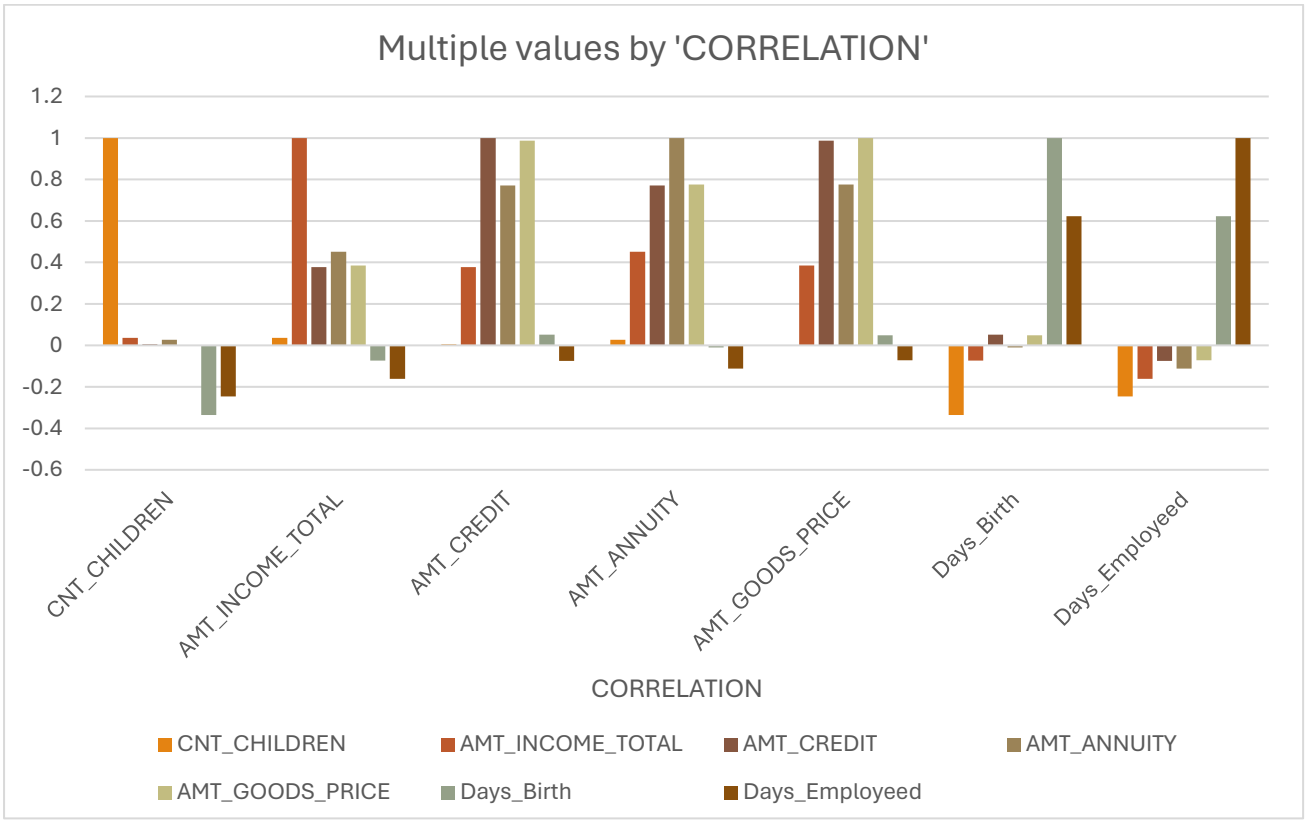
## Task E.

**Identify Top Correlations for Different Scenarios:**

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

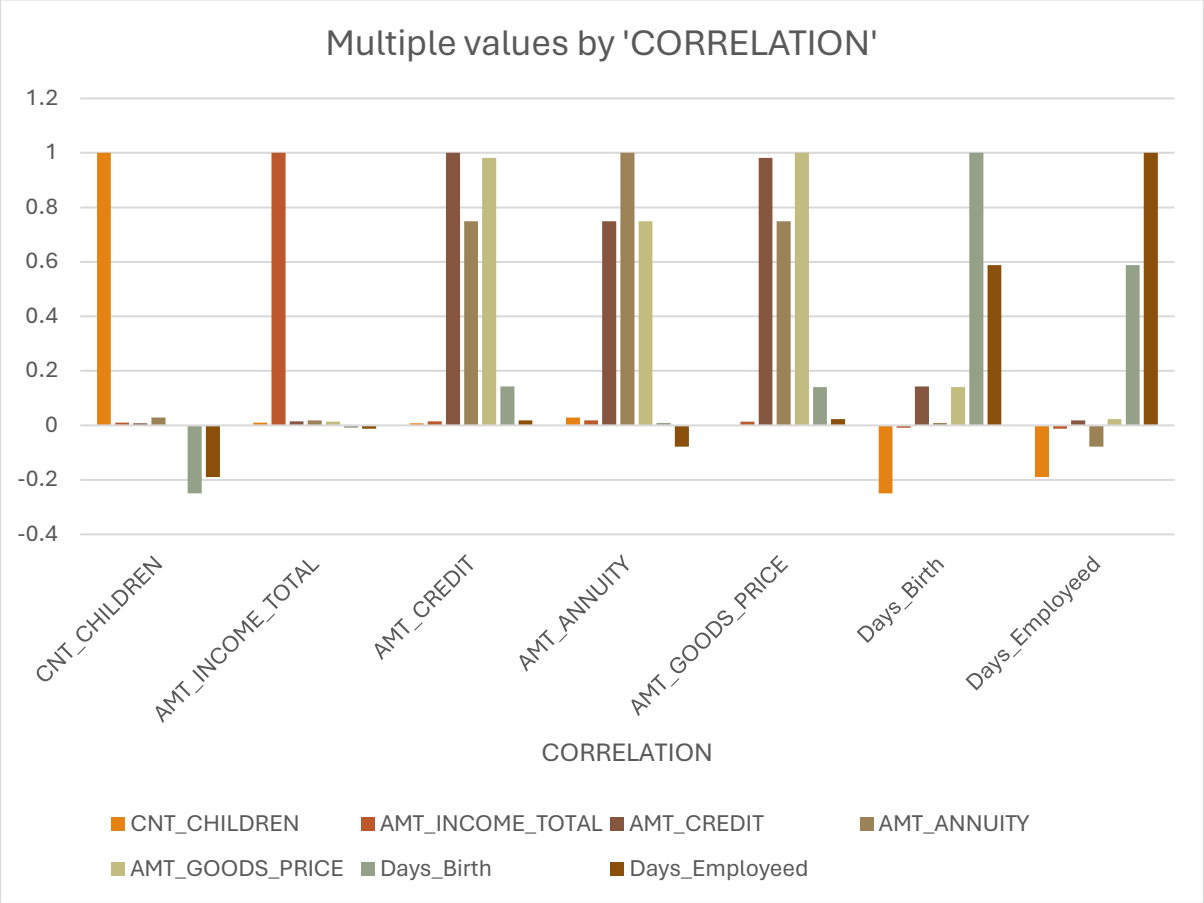
- Correlation of all the features for the Target 0:-

| CORRELATION | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | Days_Birth | Days_Employeed |
|---|---|---|---|---|---|---|---|
| CNT_CHILDREN | 1 | 0.0363197 | 0.0057055 | 0.0263821 | 0.0015181 | -0.3358763 | -0.2455215 |
| AMT_INCOME_TOTAL | 0.0363197 | 1 | 0.3779658 | 0.4511356 | 0.3845759 | -0.0737694 | -0.1616809 |
| AMT_CREDIT | 0.0057055 | 0.3779658 | 1 | 0.7707718 | 0.9869998 | 0.0510842 | -0.0747334 |
| AMT_ANNUITY | 0.0263821 | 0.4511356 | 0.7707718 | 1 | 0.7758346 | -0.0099154 | -0.1112939 |
| AMT_GOODS_PRICE | 0.0015181 | 0.3845759 | 0.9869998 | 0.7758346 | 1 | 0.0487733 | -0.0724465 |
| Days_Birth | -0.3358763 | -0.0737694 | 0.0510842 | -0.0099154 | 0.0487733 | 1 | 0.6234747 |
| Days_Employeed | -0.2455215 | -0.1616809 | -0.0747334 | -0.1112939 | -0.0724465 | 0.6234747 | 1 |



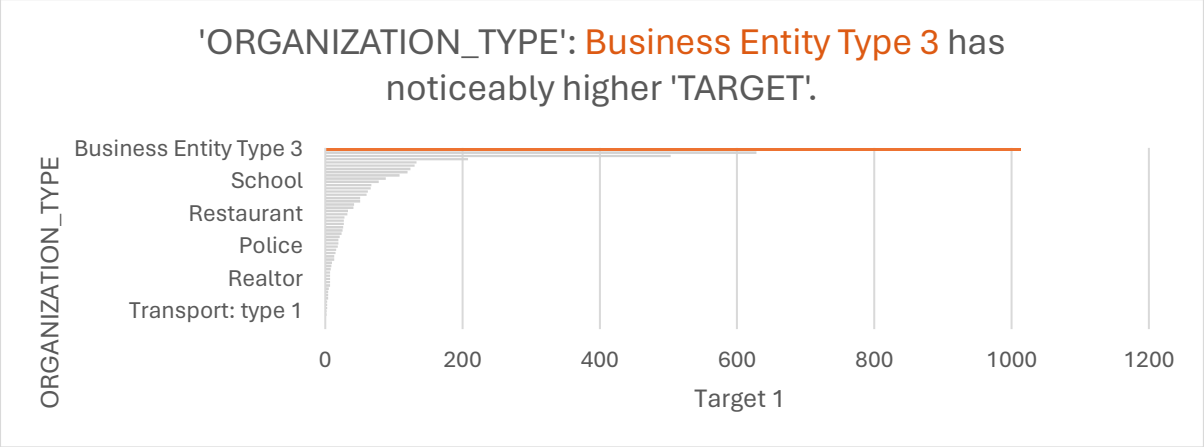
## ➢ Correlation of all the features for the Target 1:

| CORRELATION | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | Days_Birth | Days_Employeed |
|---|---|---|---|---|---|---|---|
| CNT_CHILDREN | 1 | 0.0101102 | 0.0076019 | 0.029173 | -0.0010797 | -0.2496732 | -0.1897732 |
| AMT_INCOME_TOTAL | 0.0101102 | 1 | 0.0152714 | 0.0180046 | 0.0132695 | -0.0090337 | -0.0117587 |
| AMT_CREDIT | 0.0076019 | 0.0152714 | 1 | 0.7496652 | 0.982268 | 0.142506 | 0.0187822 |
| AMT_ANNUITY | 0.029173 | 0.0180046 | 0.7496652 | 1 | 0.749504 | 0.0087517 | -0.0781139 |
| AMT_GOODS_PRICE | -0.0010797 | 0.0132695 | 0.982268 | 0.749504 | 1 | 0.1410059 | 0.0231816 |
| Days_Birth | -0.2496732 | -0.0090337 | 0.142506 | 0.0087517 | 0.1410059 | 1 | 0.5882428 |
| Days_Employeed | -0.1897732 | -0.0117587 | 0.0187822 | -0.0781139 | 0.0231816 | 0.5882428 | 1 |

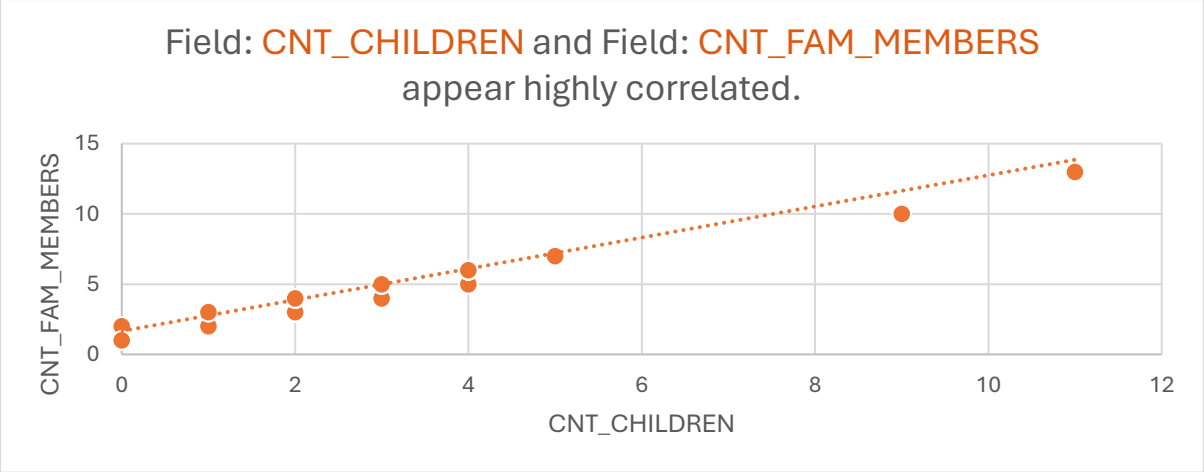



Findings:- AMT_CREDIT is is highly correlated to AMT_ANNUITY, AMT_GOODS.

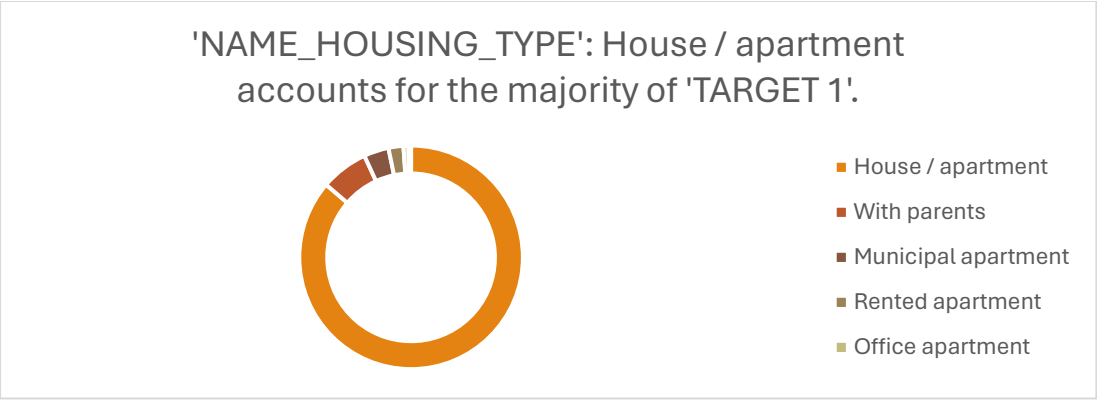
# Other Analysis and Interpretations of Defaulters:-



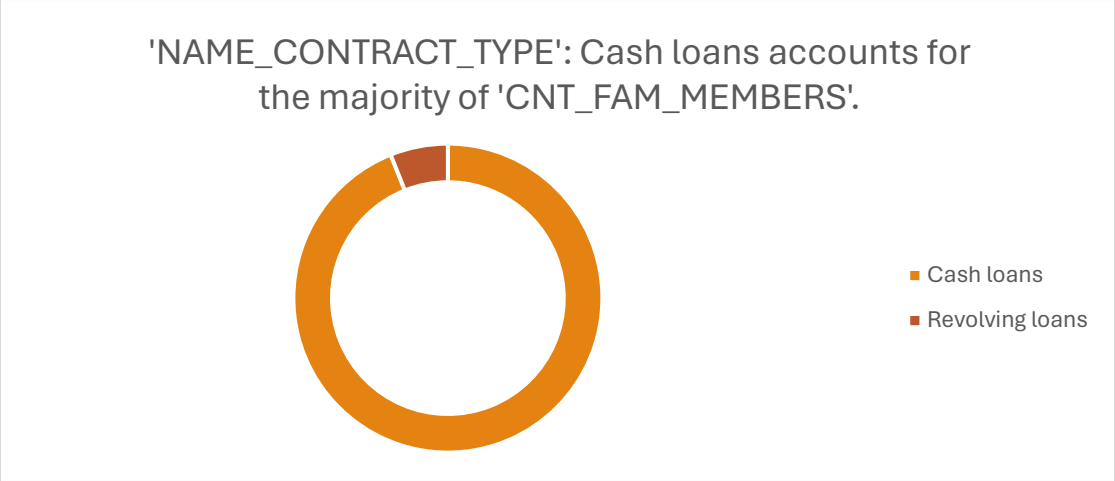

- So the clients belonging to Business Entity Type 3 have more number of children , which can be one of the reason for defaulting.



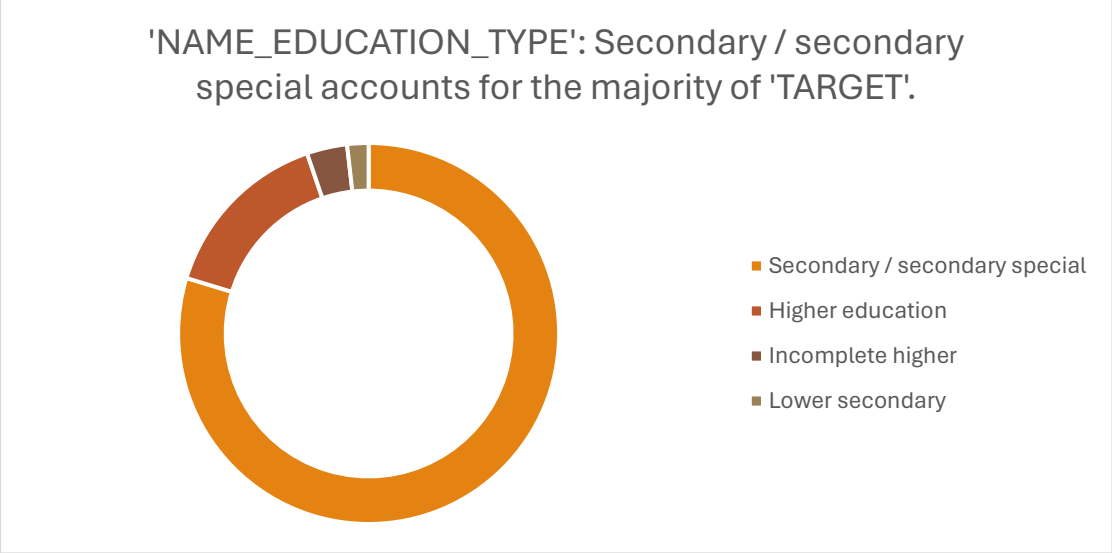

- CNT_Children and CNT_FAM_Members are highly correlated. This means the more number of children or Family members , higher is the chance of defaulting




- It is observed that clients belong to House /Apartments have defaulted a lot as they can stay on rent or have to pay huge EMIs.




- Clients have higher number of of CNT_FAM_Members have taken Cash Loans and defaulted.




## Conclusion:

This project demonstrates effective techniques for handling large datasets, particularly through the application of exploratory data analysis (EDA). When dealing with extensive datasets, it's crucial to streamline the analysis by selecting only the most relevant columns. Exploring correlations between columns can significantly aid in this process, optimizing time and resources by identifying key variables for analysis.

----------**********----------

# Module7: Analyzing the Impact of Car Features on Price and Profitability

**Description :-**

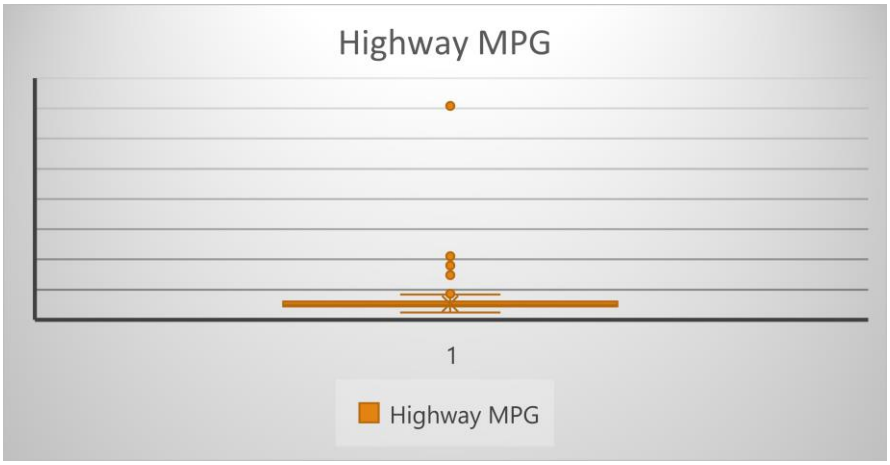
The automotive industry has seen significant changes in recent years, driven by a stronger emphasis on fuel efficiency, sustainability, and technological advancements. Manufacturers face intense competition and must navigate a shifting consumer landscape, making it crucial to understand what influences car buyers.

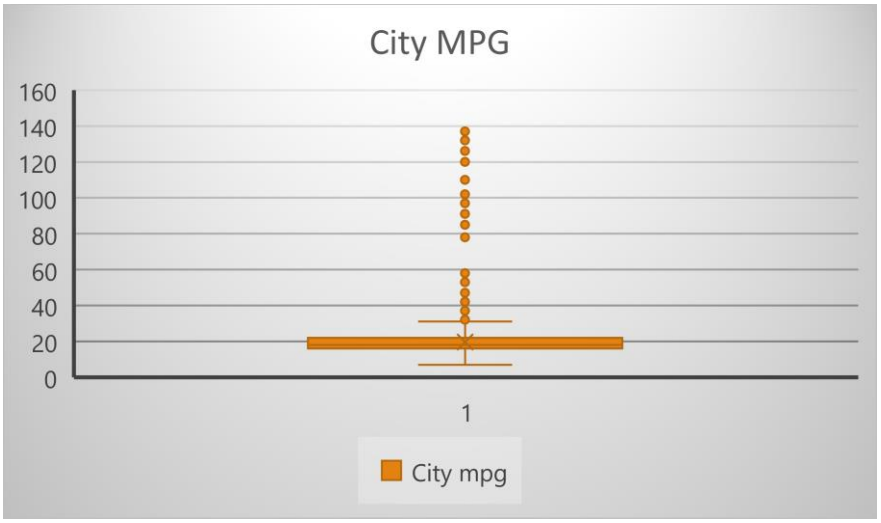
**Tech Stack Used:-**Microsoft Excel 365

## Exploratory Data Analysis:-

1. Engine Fuel Type in Suzuki Car had 3 blank space filled with the mode that is regular unleaded
2. Engine HP of 69 cars were not available which were filled using data from the website www.edmunds.com
3. 714 duplicates found in the data which have been removed.
4. There are no Engine cylinders in electric vehicles , so all blank columns are filled with 0.Mazda cars are filled with 2 and 4 cylinders respectively using the data www.autoevolution.com
5. Number of Doors in cars are filled using data from Wikipedia.com
6. 715 Duplicates found , hence the rows remaining for Analysis are 11199

❖ **Outliers:-**



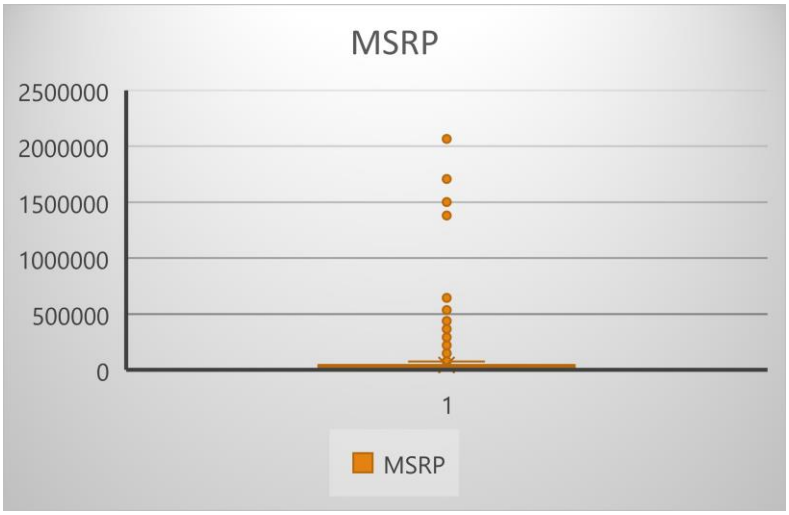

**MSRP**

| | |
|---|---|
| **Quartile Q1** | 21599 |
| **Quartile Q3** | 43035 |
| **IQR** | 21436 |

| | | |
|---|---|---|
| **Upper Bound** | Q3+1.5*IQR | 75189 |
| **Lower Bound** | Q1-1.5*IQR | -10555 |

| | |
|---|---|
| **Count of Outliers** | 960 |
| **Total Counts** | 11199 |
| **% of Outliers** | 8.57% |

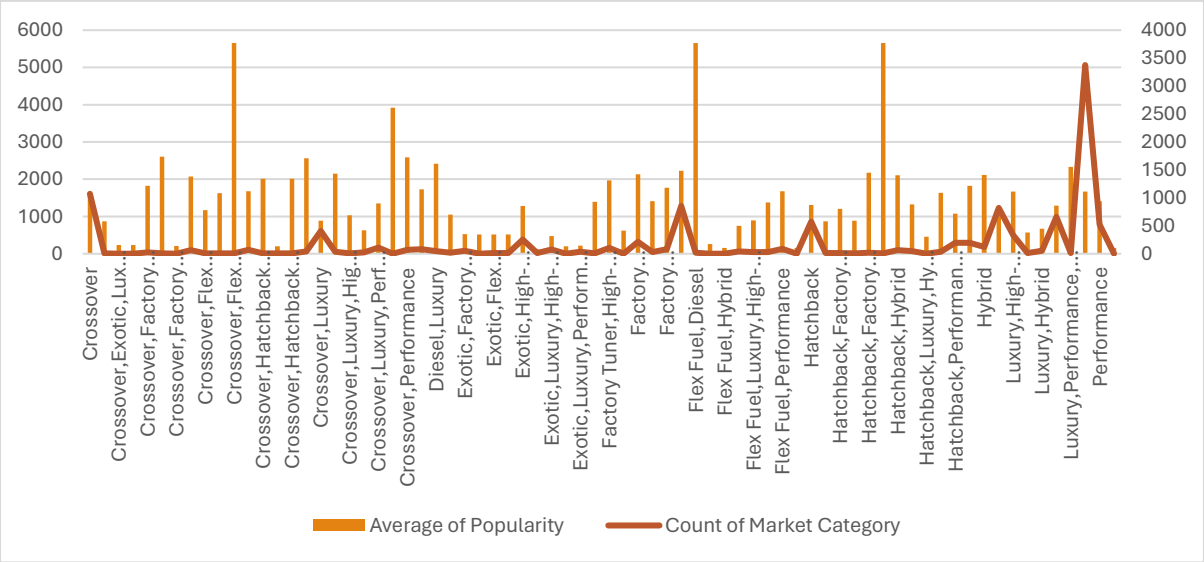
7.




**Task 1.A:** The entire Pivot table is created in Excel sheet.

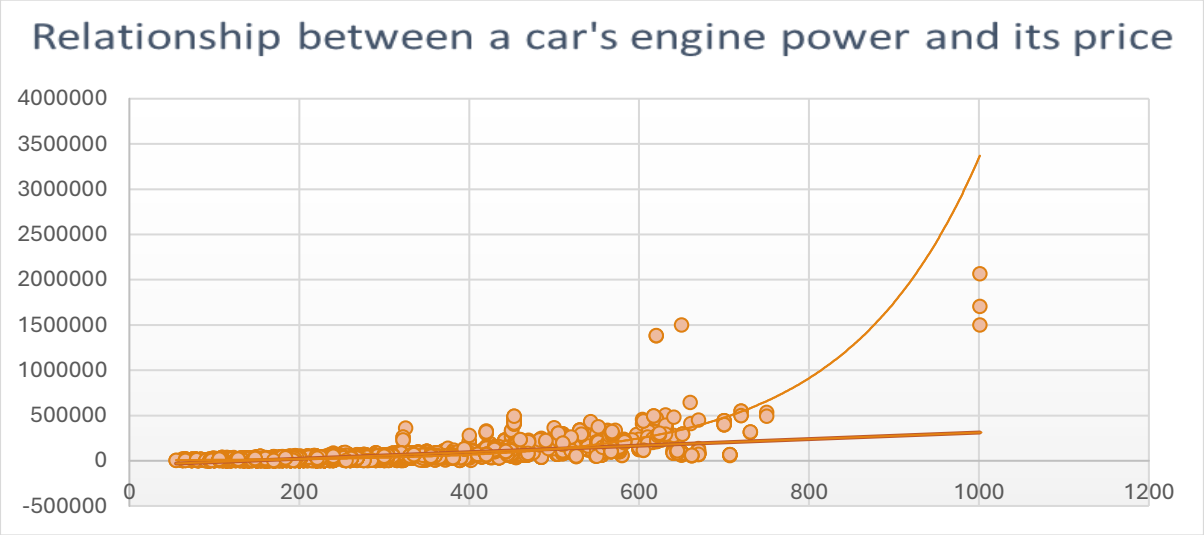
| Make, Model and their Market Category | Sum of Popularity |
|---|---|
| **Acura** | **50184** |
| **CL** | **1836** |
| Factory Tuner,Luxury,Performan | |
| Luxury | |
| **ILX** | **3264** |
| Luxury | 3060 |
| Luxury,Performance | 204 |
| **ILX Hybrid** | **408** |
| Luxury,Hybrid | 408 |
| **Integra** | **4896** |
| Hatchback,Factory Tuner,Luxury,Performance | 408 |
| Hatchback,Luxury | 1632 |
| Hatchback,Luxury,Performance | 612 |
| Luxury | 1632 |
| Luxury,Performance | 612 |
| **Legend** | **3264** |
| Luxury | 1632 |
| Luxury,Performance | 1632 |
| **MDX** | **6936** |
| Crossover,Luxury | 6936 |
| **NSX** | **1020** |

## Task 1.B:

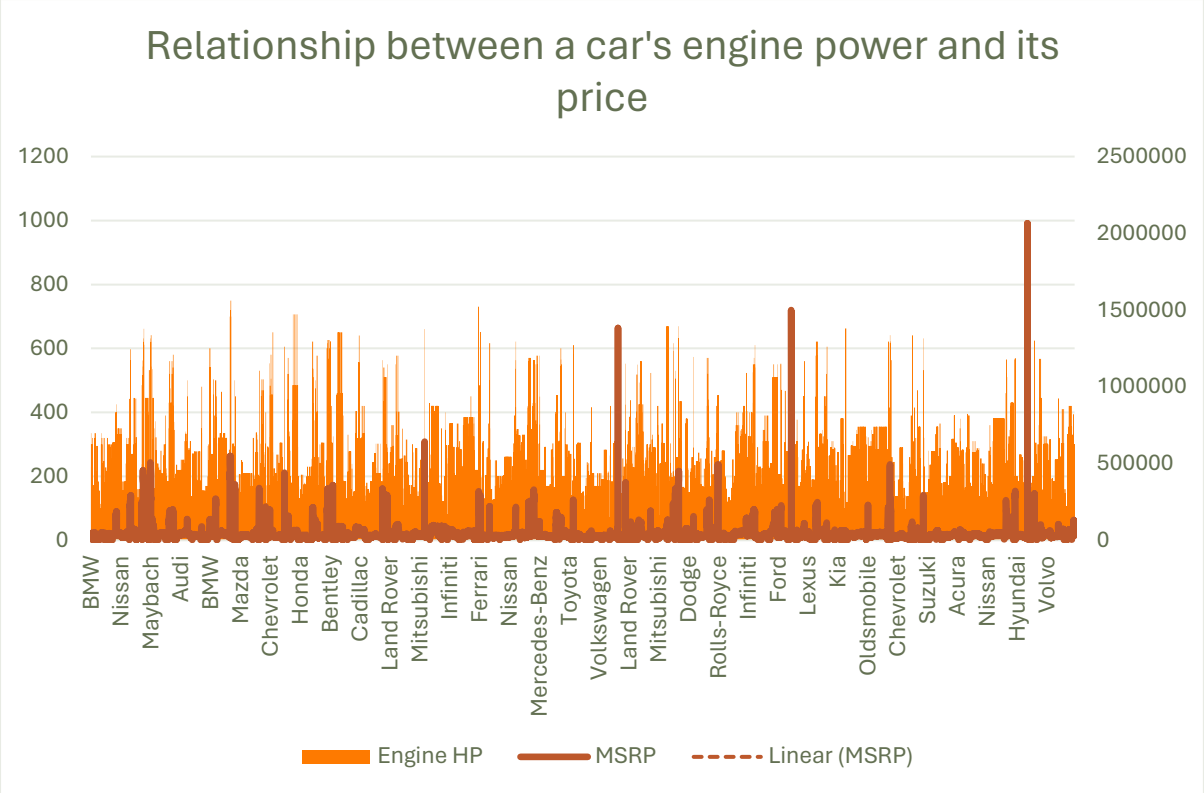


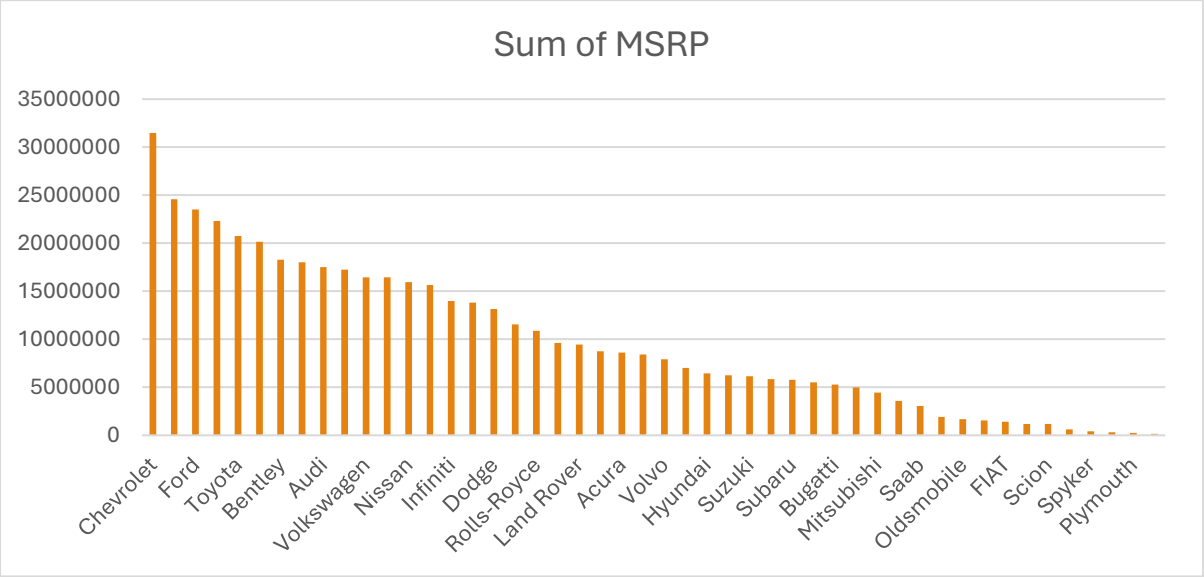

**Insights:-** The Crossover market category has the highest number of Car Models with the average Count of 1556 and average popularity of 1075. Some Market category like 1.Crossover,Flex Fuel,Performance 2.Flex Fuel,Diesel 3.Hatchback,Flex Fuel has highest average of Popularity but their count is very less.

## Task 2:




**Insights:-** Car having Engine Power of 1001 has the highest MSRP. So MSRP is directly proportional to Engine HP. With increase in engine Hp ,the MSRP increases. If someone prefers to buy a car with Better car engine, he will end up paying more price.



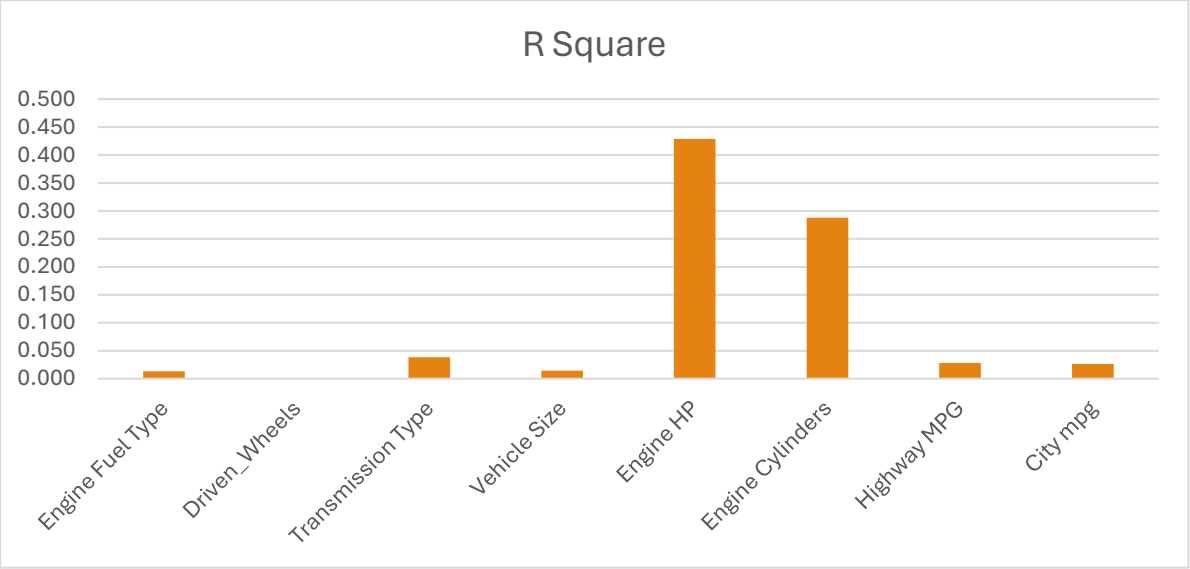

**Insights :-** Chevrolet Brand has the highest Sum of MSRP.

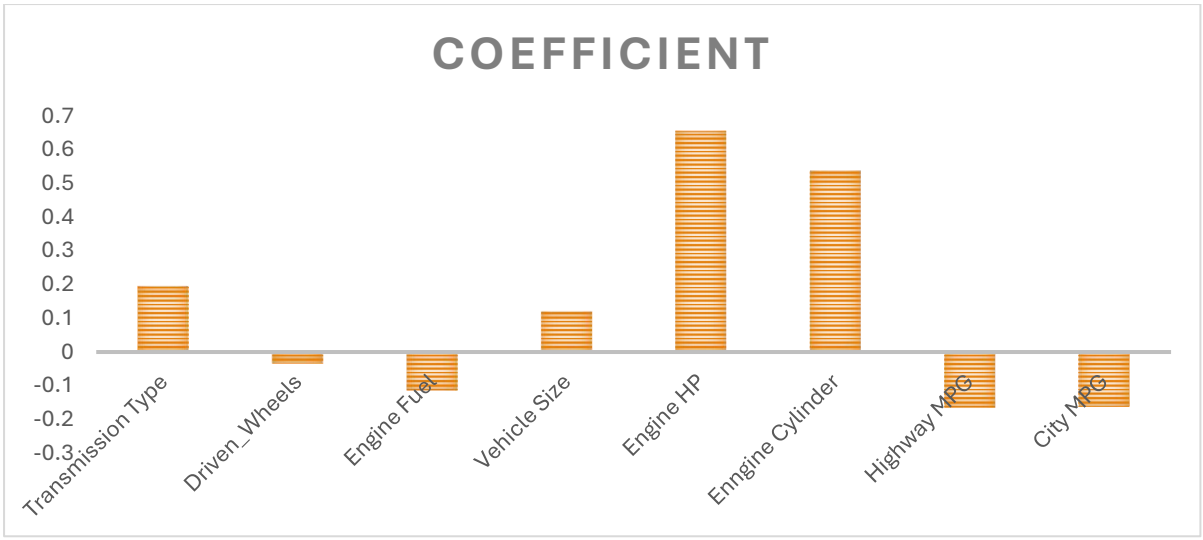
# Task 3:

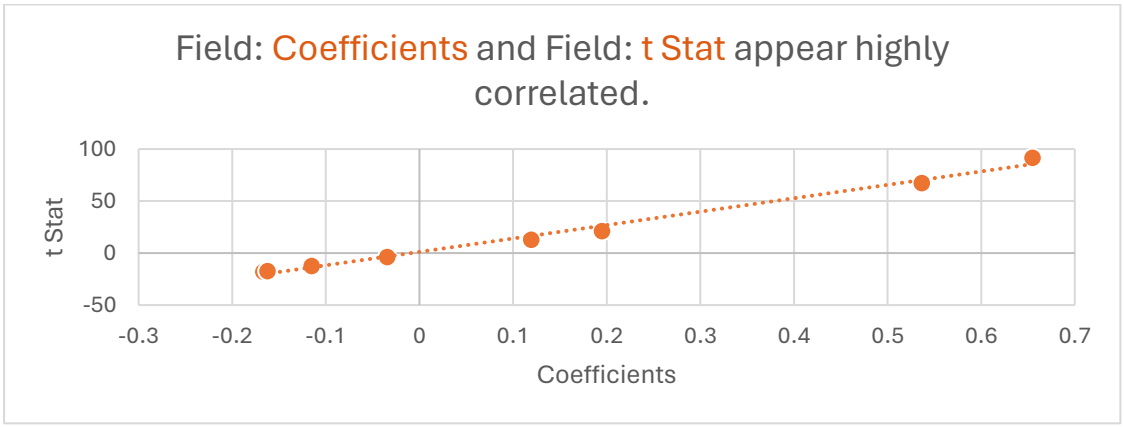
To find the regression analysis all the categorical data is changed to Numerical data **Reference: Encoding categorical values**. And then using the Z-score Method Normalization was done **Reference:Normalization**. Regression analysis of all the features with target MSRP was done using Excel 365 Data analysis. **Reference:Task3Regression Analysis**

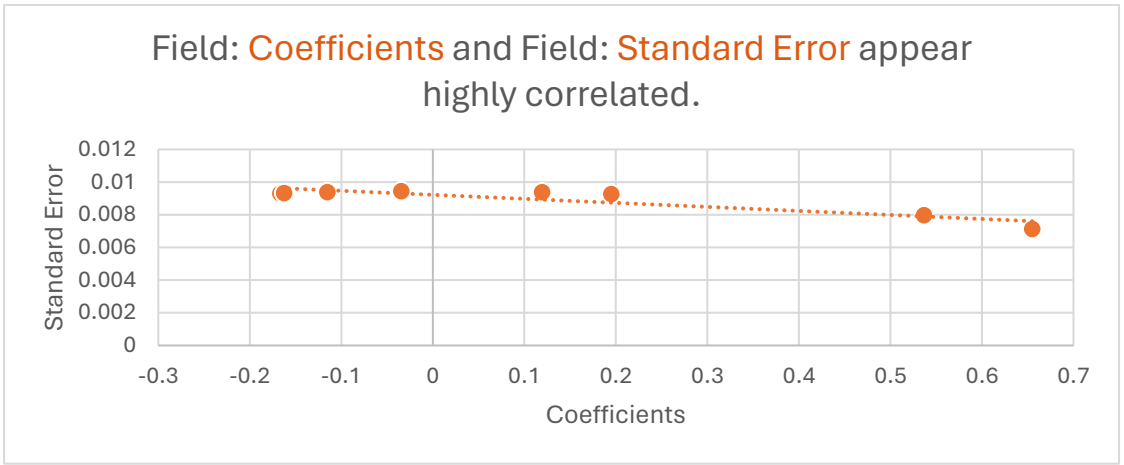
| Regression | Engine Fuel Type | Driven_Wheels | Transmission Type | Vehicle Size | Engine HP | Engine Cylinders | Highway MPG | City mpg |
|---|---|---|---|---|---|---|---|---|
| **Multiple R** | 0.115 | 0.034 | 0.195 | 0.119 | 0.655 | 0.537 | 0.167 | 0.162 |
| **R Square** | 0.013 | 0.001 | 0.038 | 0.014 | 0.429 | 0.288 | 0.028 | 0.026 |
| **Adjusted R Square** | 0.013 | 0.001 | 0.038 | 0.014 | 0.429 | 0.288 | 0.028 | 0.026 |
| **Standard Error** | 0.993 | 0.999 | 0.981 | 0.993 | 0.756 | 0.844 | 0.986 | 0.987 |
| **Observations** | 11198 | 11198 | 11198 | 11198 | 11198 | 11198 | 11198 | 11198 |

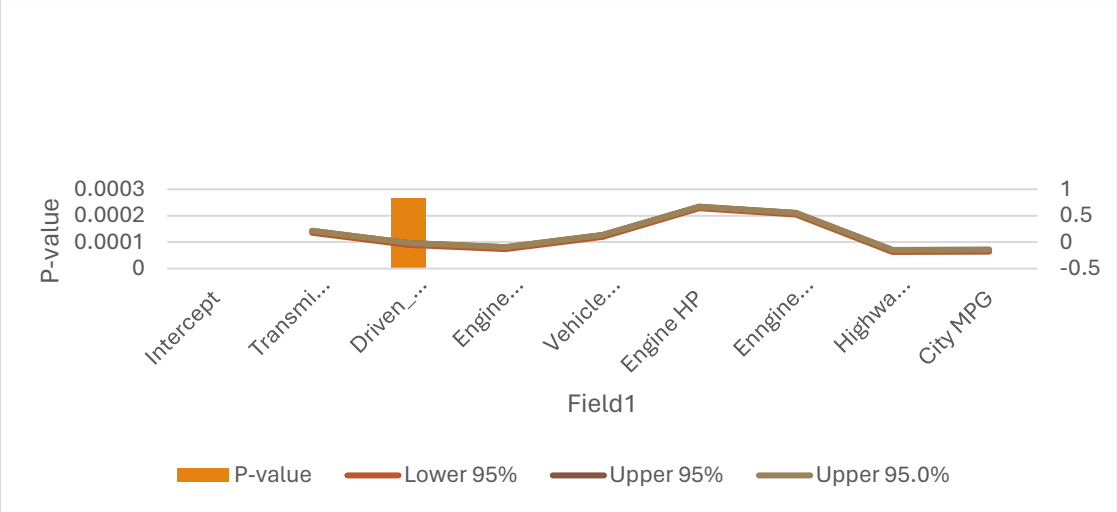
| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| **Intercept** | 0 | #N/A | #N/A | #N/A | #N/A | #N/A | #N/A | #N/A |
| **Transmission Type** | 0.194988518 | 0.009270003 | 21.03435241 | 2.2939E-96 | 0.176818 | 0.213159355 | 0.176817681 | 0.21316 |
| **Driven_Wheels** | -0.03444242 | 0.009445368 | -3.646488089 | 0.00026705 | -0.052957 | -0.015927839 | -0.052957002 | -0.0159 |
| **Engine Fuel** | -0.11520303 | 0.009388025 | -12.27127435 | 2.1483E-34 | -0.133605 | -0.09680085 | -0.13360521 | -0.0968 |
| **Vehicle Size** | 0.11950598 | 0.00938315 | 12.73623192 | 6.7032E-37 | 0.101113 | 0.137898605 | 0.101113355 | 0.1379 |
| **Engine HP** | 0.654855107 | 0.00714246 | 91.68481705 | 0 | 0.640855 | 0.668855584 | 0.64085463 | 0.66886 |
| **Enngine Cylinder** | 0.536629131 | 0.007974416 | 67.29384376 | 0 | 0.520998 | 0.55226039 | 0.520997873 | 0.55226 |
| **Highway MPG** | -0.166631148 | 0.009318255 | -17.88222734 | 1.5462E-70 | -0.184897 | -0.14836573 | -0.184896566 | -0.1484 |
| **City MPG** | -0.16234251 | 0.009325014 | -17.40935829 | 5.3245E-67 | -0.180621 | -0.144063843 | -0.180621177 | -0.1441 |




COEFFICIENT
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
Transmission Type
Driven_Wheels
Engine Fuel
Vehicle Size
Engine HP
Enngine Cylinder
Highway MPG
City MPG


Field: Coefficients and Field: t Stat appear highly correlated.
t Stat
100
50
0
-50
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Coefficients


Field: Coefficients and Field: Standard Error appear highly correlated.
Standard Error
0.012
0.01
0.008
0.006
0.004
0.002
0
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Coefficients



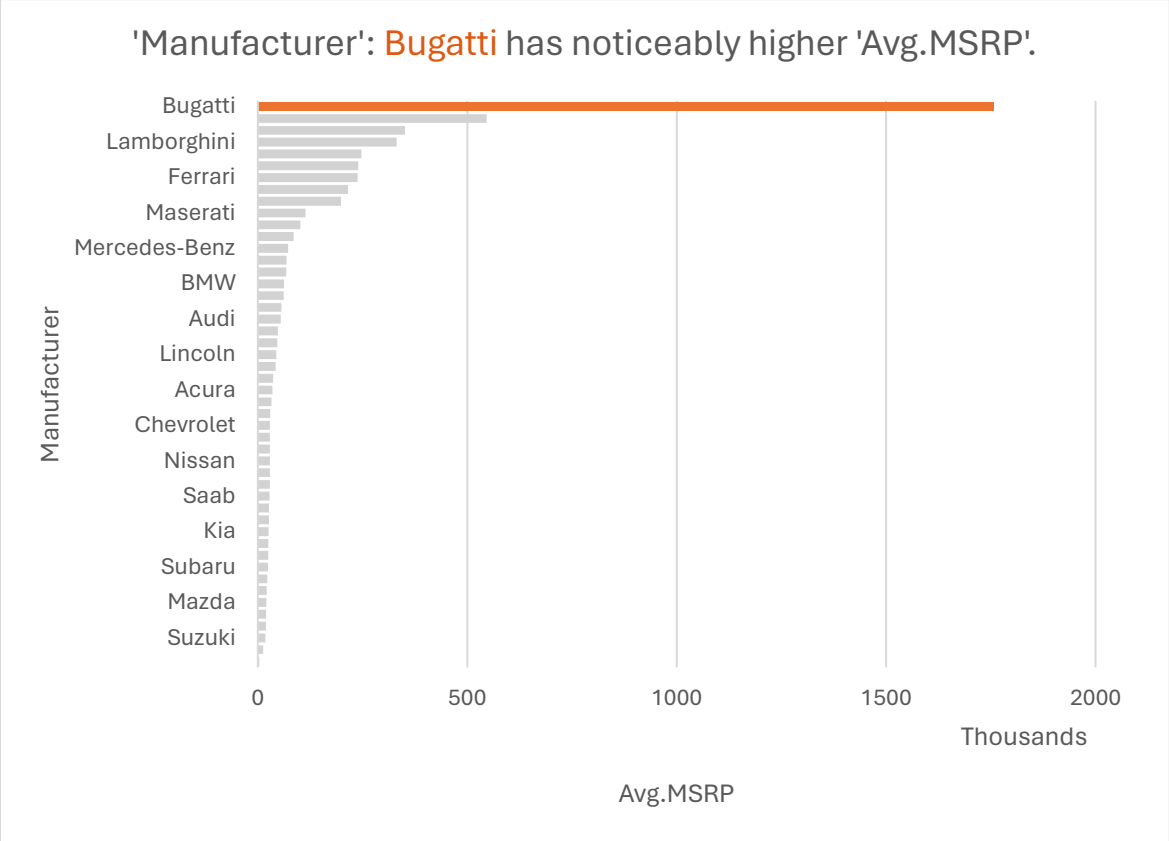

**Insights:-** To find the best suitable Feature we have to take into account R-square and Coefficient values. As per the observations engine HP has the strongest relationship with Price (target MSRP) followed by Engine cylinder. Hence we can say that Engine HP followed by Engine Cylinder are the most importance features to decide the Car price(MSRP).Coefficients and Standard Error appear highly correlated. Coefficients and t Stat appear highly correlated.

## Task4.A

| Manufacturer | Average of MSRP |
|---|---|
| Acura | 35087.49 |
| Alfa Romeo | 61600.00 |
| Aston Martin | 198123.46 |
| Audi | 54574.12 |
| Bentley | 247169.32 |
| BMW | 62162.56 |
| Bugatti | 1757223.67 |
| Buick | 29034.19 |
| Cadillac | 56368.27 |
| Chevrolet | 29074.73 |
| Chrysler | 26722.96 |
| Dodge | 24857.05 |
| Ferrari | 238218.84 |
| FIAT | 22670.24 |
| Ford | 28511.31 |
| Genesis | 46616.67 |
| GMC | 32444.09 |
| Honda | 26655.15 |
| HUMMER | 36464.41 |
| Hyundai | 24926.26 |
| Infiniti | 42640.27 |
| Kia | 25513.76 |
| Lamborghini | 331567.31 |
| Land Rover | 68067.09 |
| Lexus | 47549.07 |

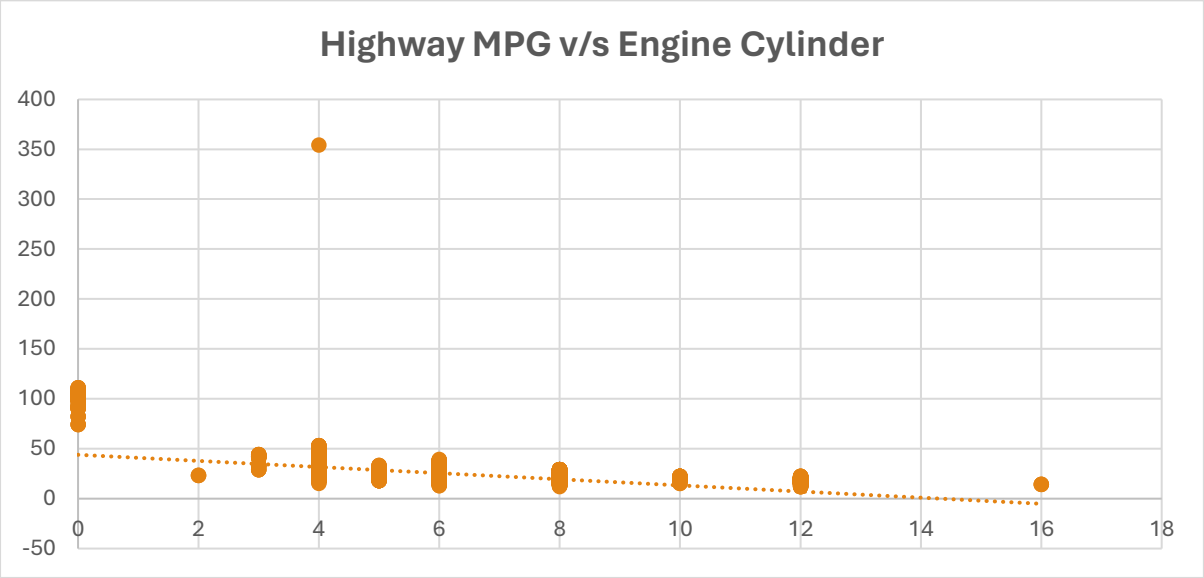
| | |
|---|---|
| Lincoln | 43860.83 |
| Lotus | 68377.14 |
| Maserati | 113684.49 |
| Maybach | 546221.88 |
| Mazda | 20416.62 |
| McLaren | 239805.00 |
| Mercedes-Benz | 72069.53 |
| Mitsubishi | 21340.56 |
| Nissan | 28921.15 |
| Oldsmobile | 12843.80 |
| Plymouth | 3296.87 |
| Pontiac | 19800.04 |
| Porsche | 101622.40 |
| Rolls-Royce | 351130.65 |
| Saab | 27879.81 |
| Scion | 19932.50 |
| Spyker | 214990.00 |
| Subaru | 24240.67 |
| Suzuki | 18026.42 |
| Tesla | 85255.56 |
| Toyota | 28846.56 |
| Volkswagen | 28978.52 |
| Volvo | 29724.68 |
| **Grand Total** | **41925.92714** |

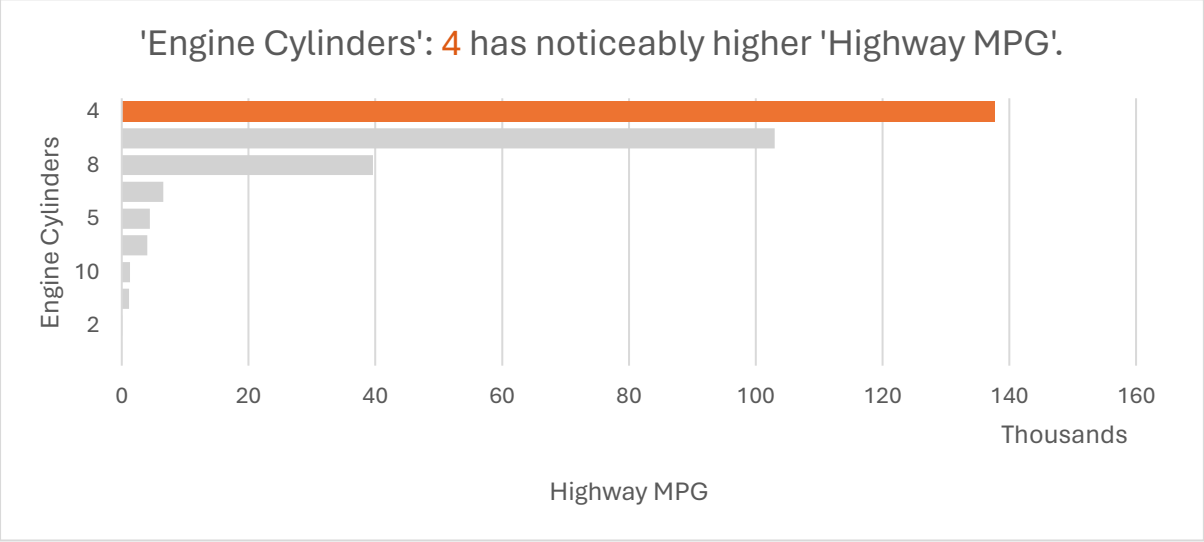
**Task 4.B:**



**Insights:-** From the above Pivot Chart and Graph it is clear that Buggati has highest average price while Plymouth has the Lowest Price .

## Task 5.A:

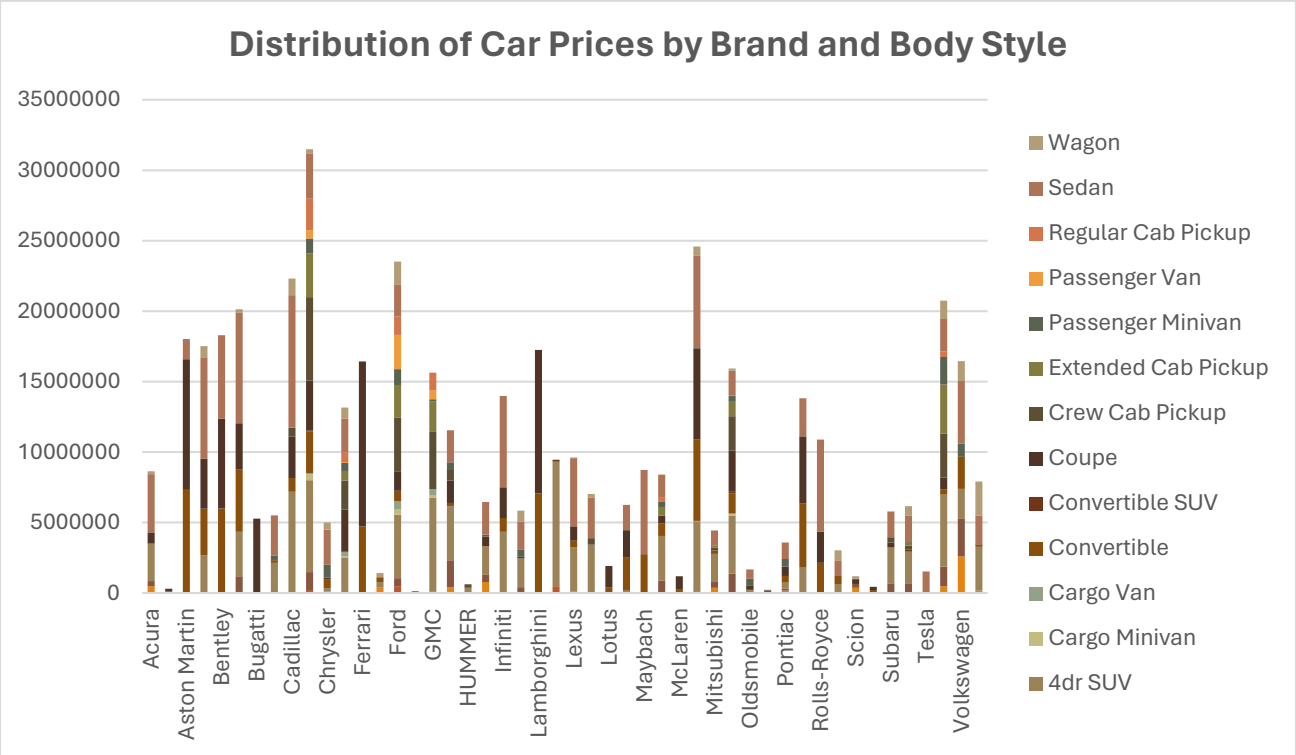



**Insights:-** We can observe that the plot between highway MPG and Engine Cylinders has a negative slope. The correlation coefficient is also Negative with a value of -0.6166.This is logical because as number of Engine Cylinders increases, the amount of fuel to be burnt also increases, thus decreasing the mileage (highway MPG).

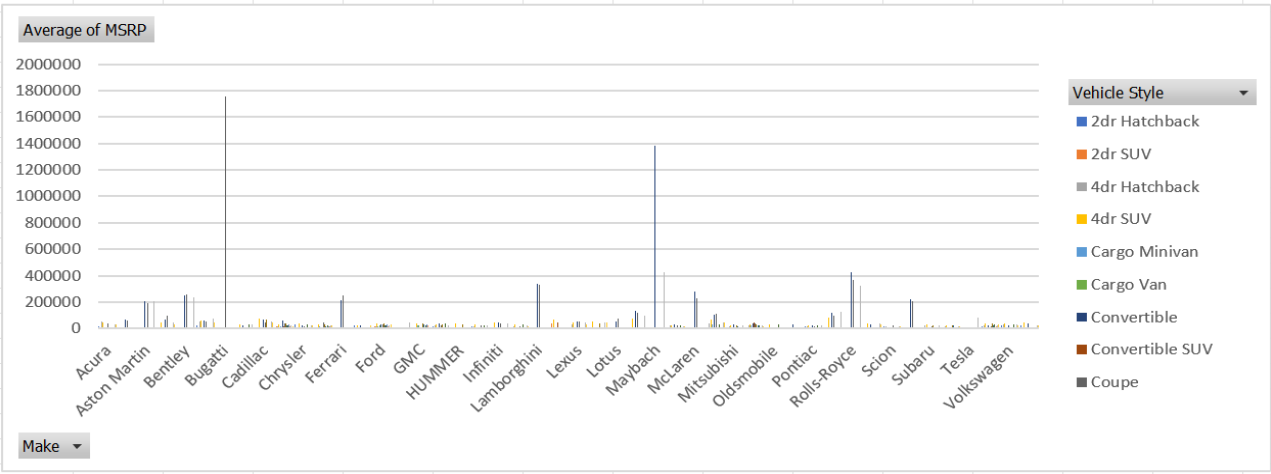
# Building the Dashboard

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

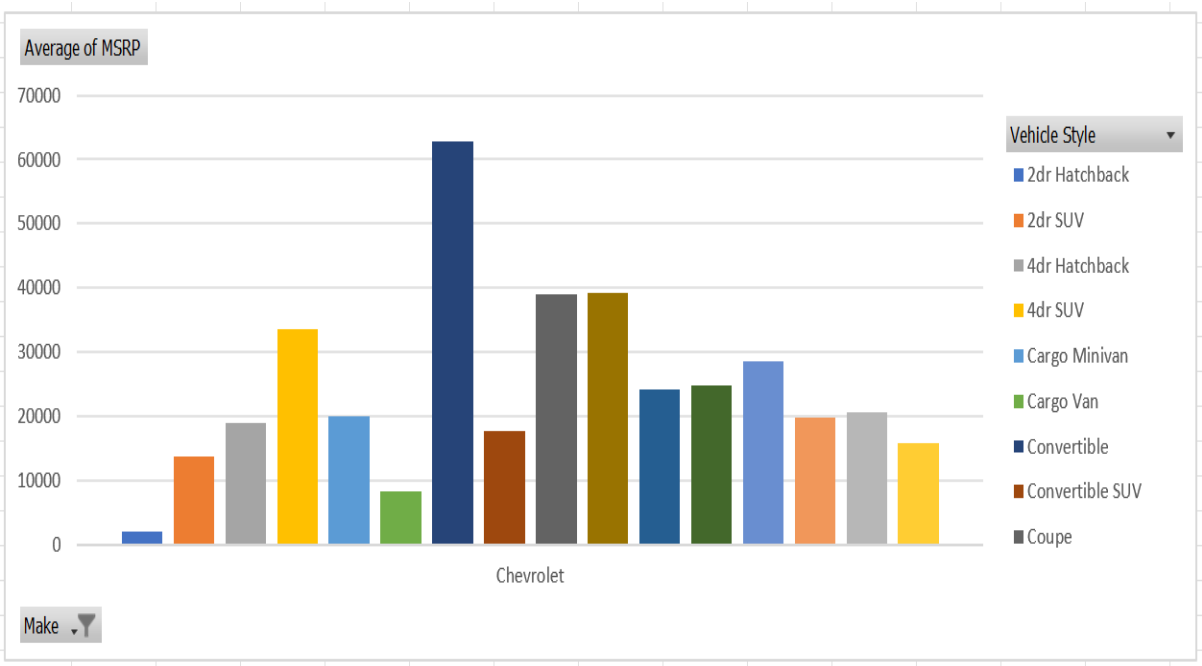



**Insights:** Pivot Table and stacked column chart is used to find out Distribution of Car Prices by Brand and Body Style. It is observed that Chevrolet has the highest sum of MSRP followed by Mercedes-Benz.

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

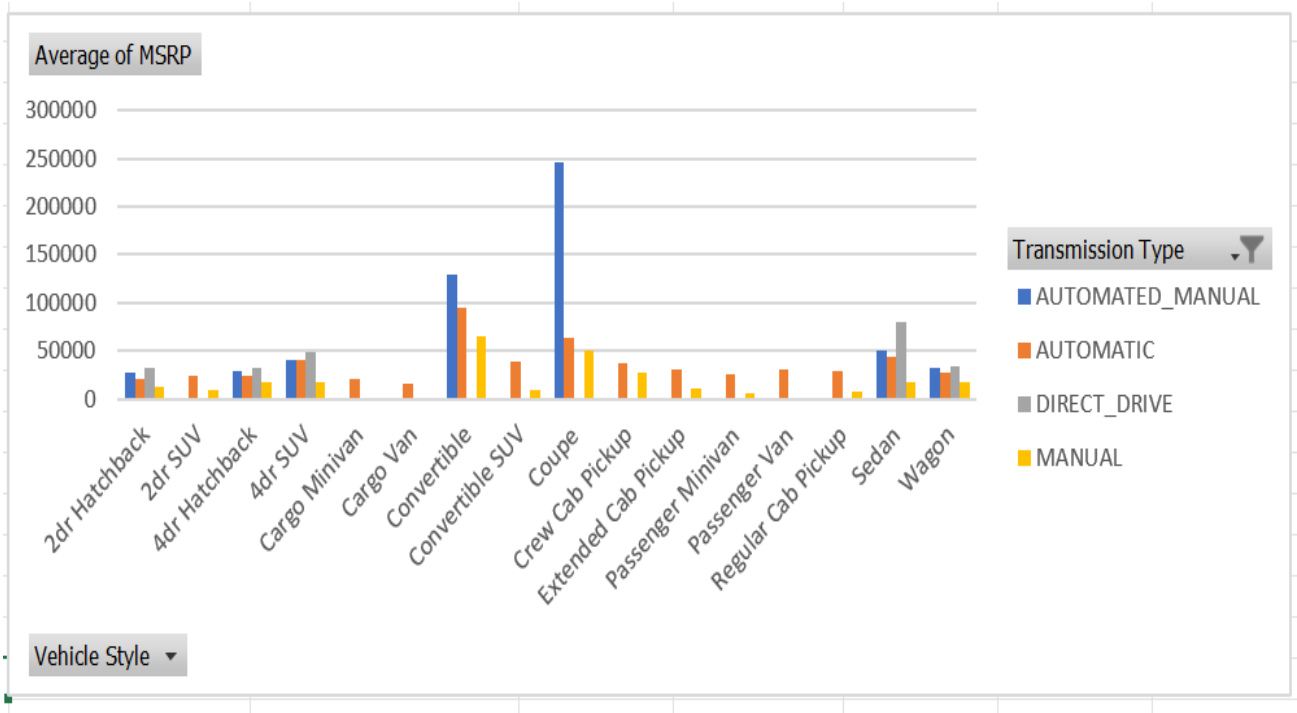


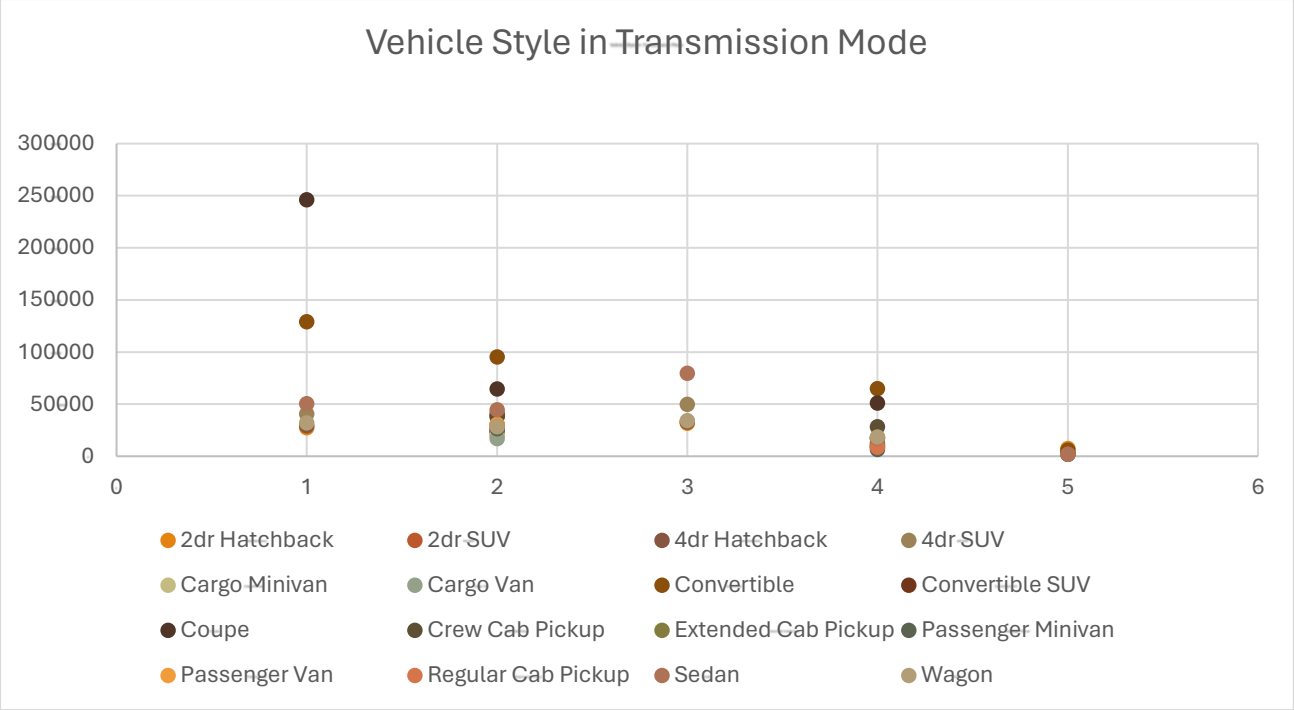

**Insights:-** Buggati has Highest Average MSRP of 1757223.667 and Plymouth has lowest Average MSRP 3296.87.

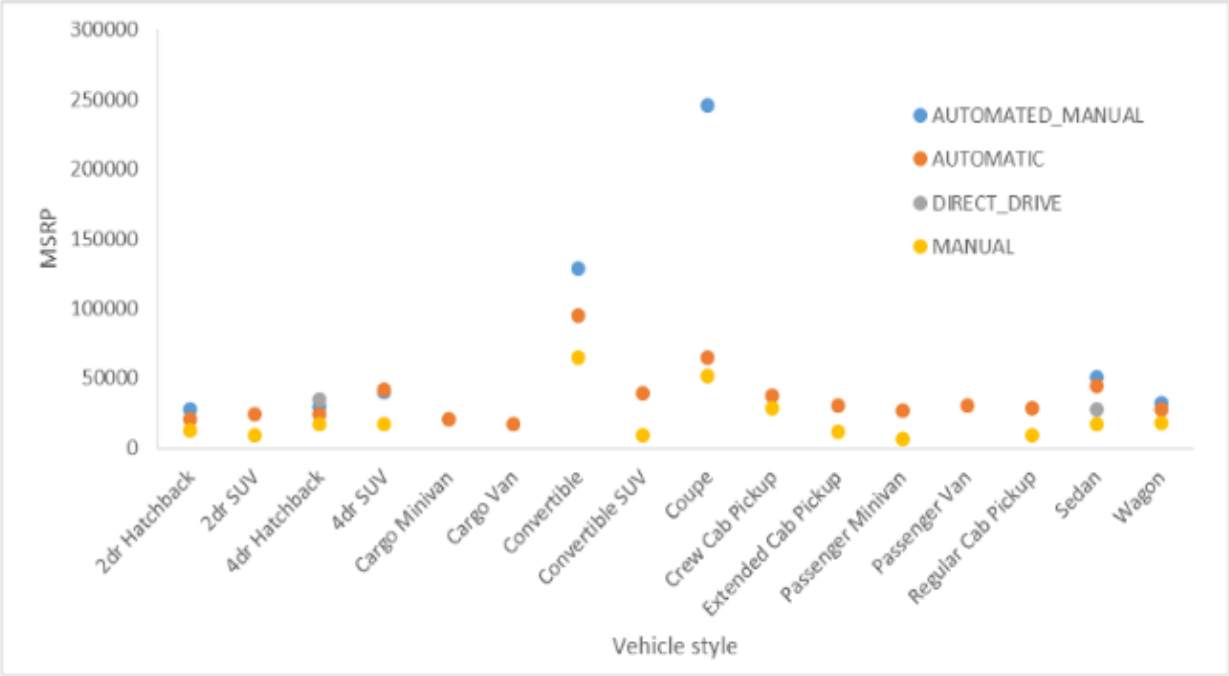



**Task 3:** How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?

Data for Unknown Transmission has been filtered out to get a proper analysis.

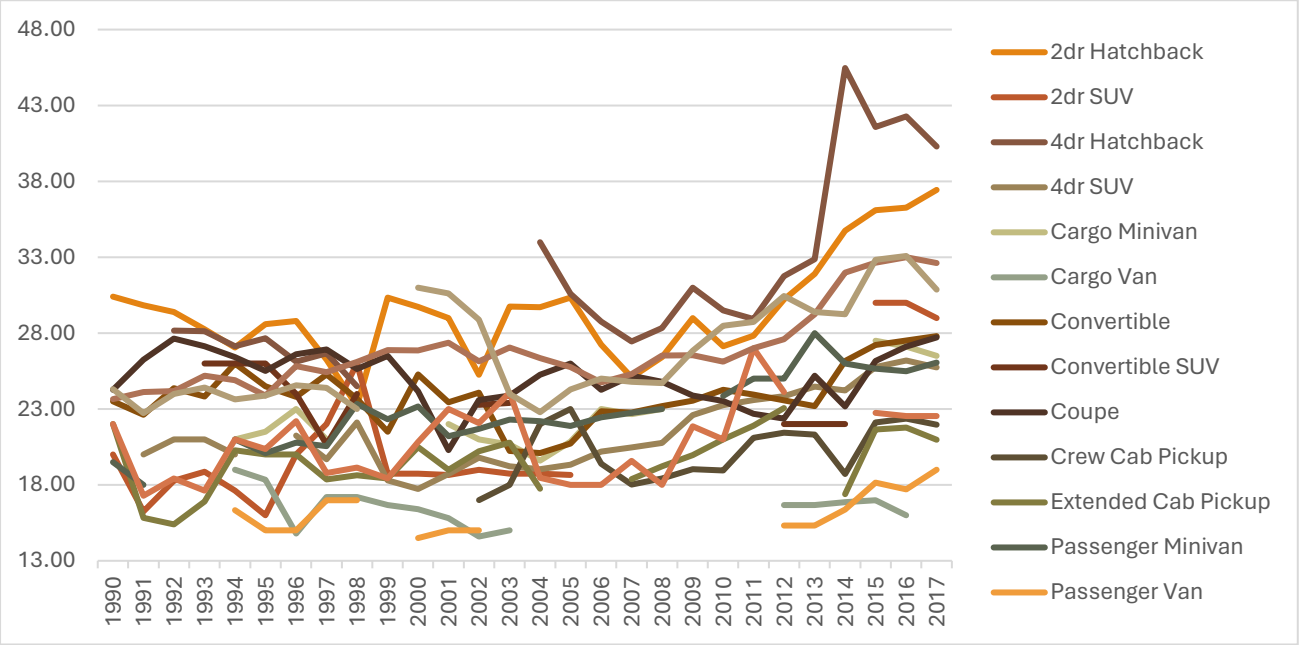


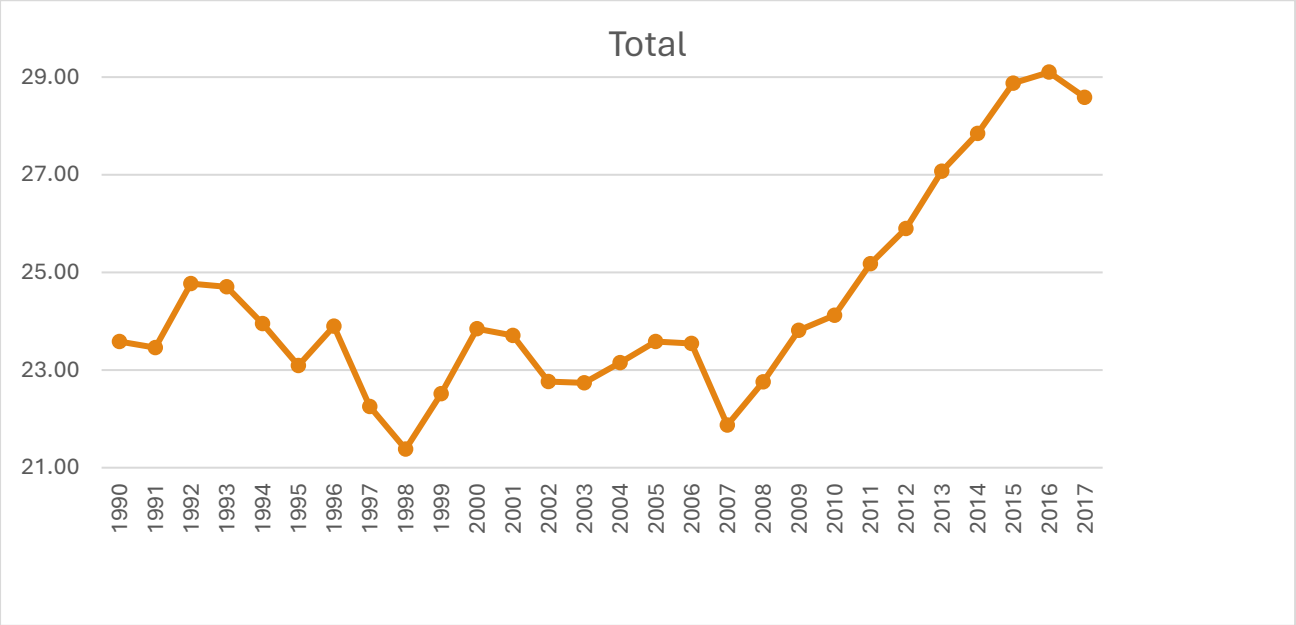



**Insights:-** Automated_Manual is the most saleable and MSRP making Transmission Mode. Coupe is the Vehicle Style which is mostly sold in Automated Manual Transmission.

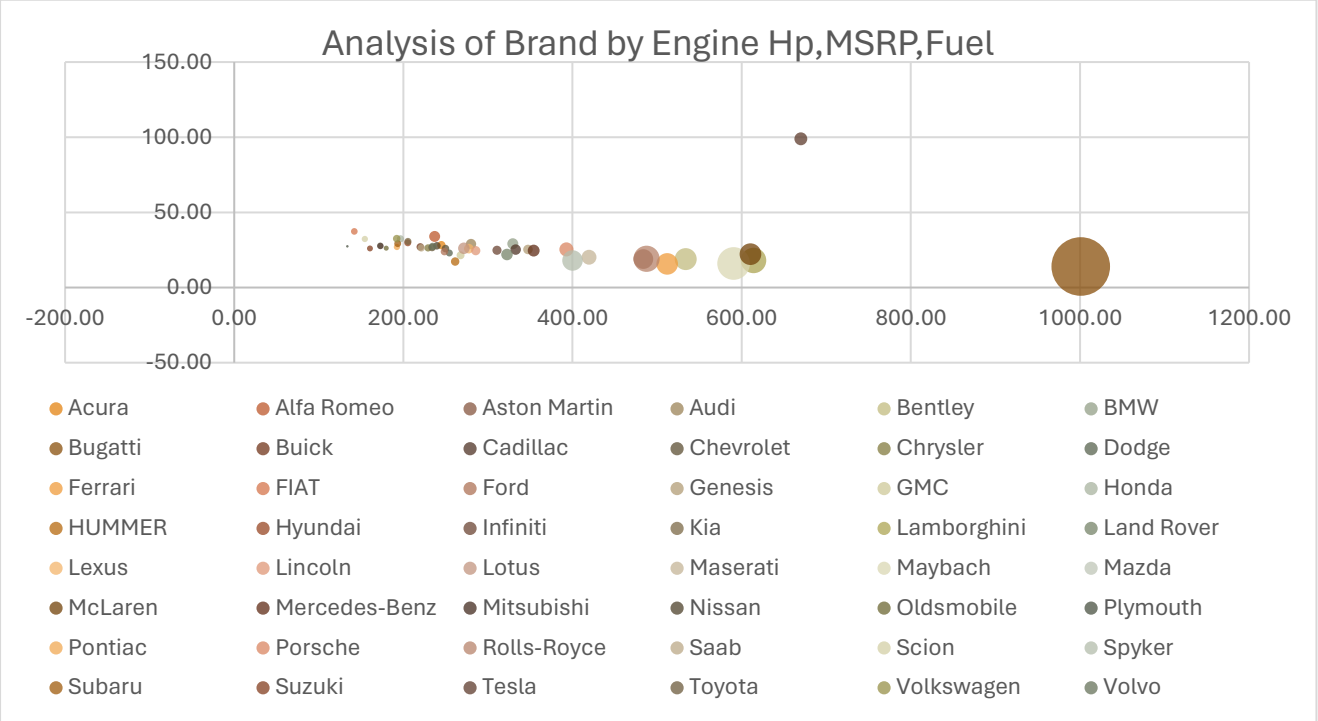
**Task 4:** How does the fuel efficiency of cars vary across different body styles and model years?

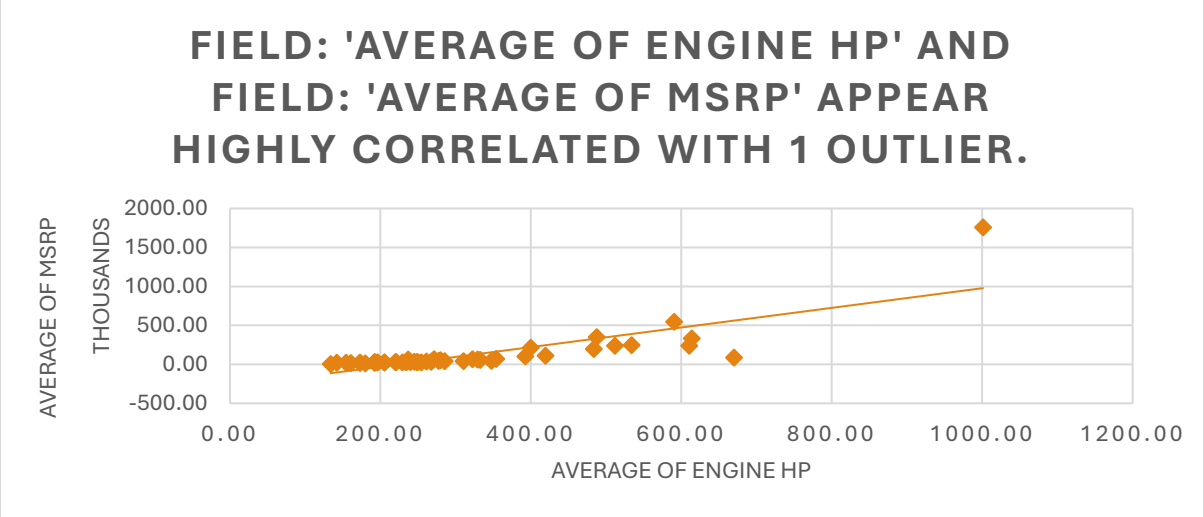


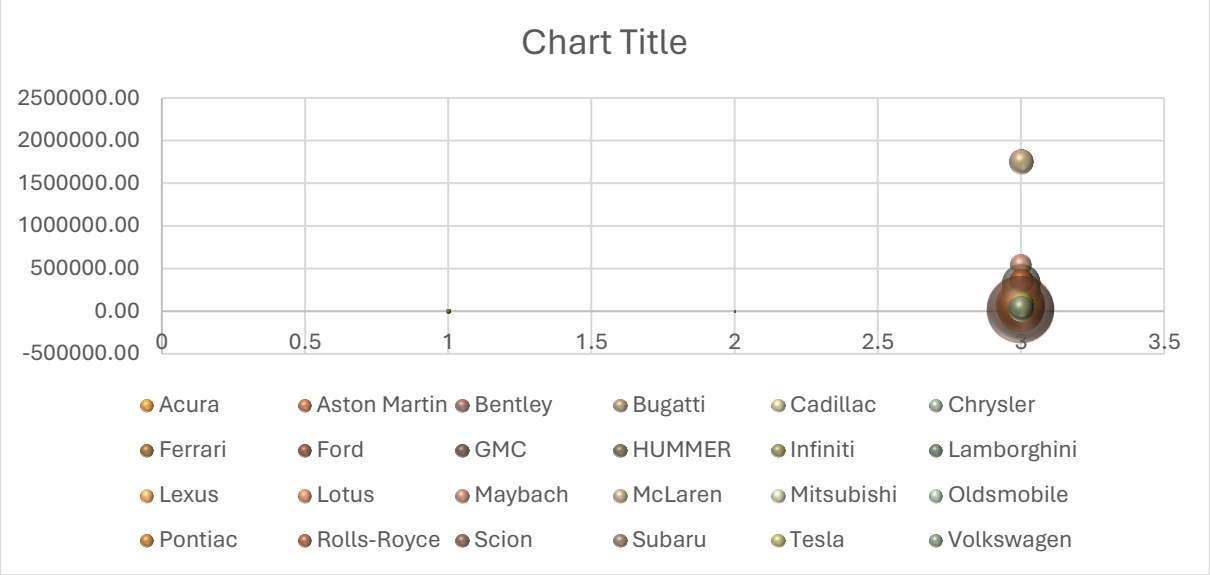



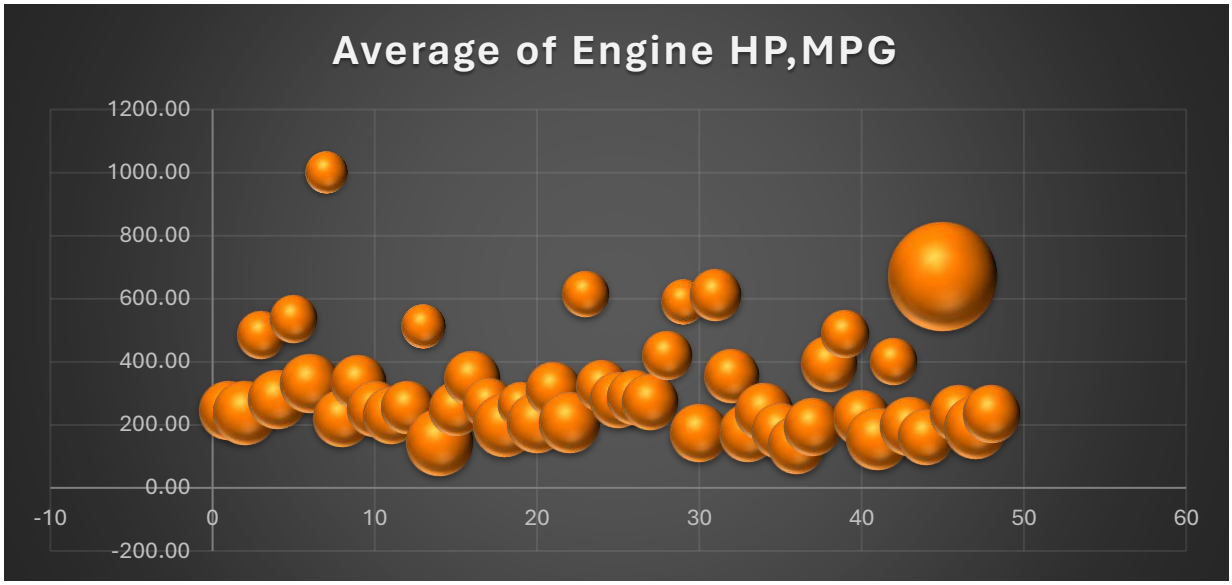

**Insights:-** We can understand from the graph that fuel efficiency has increased over the period of years for the Brands and Body Styles.

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

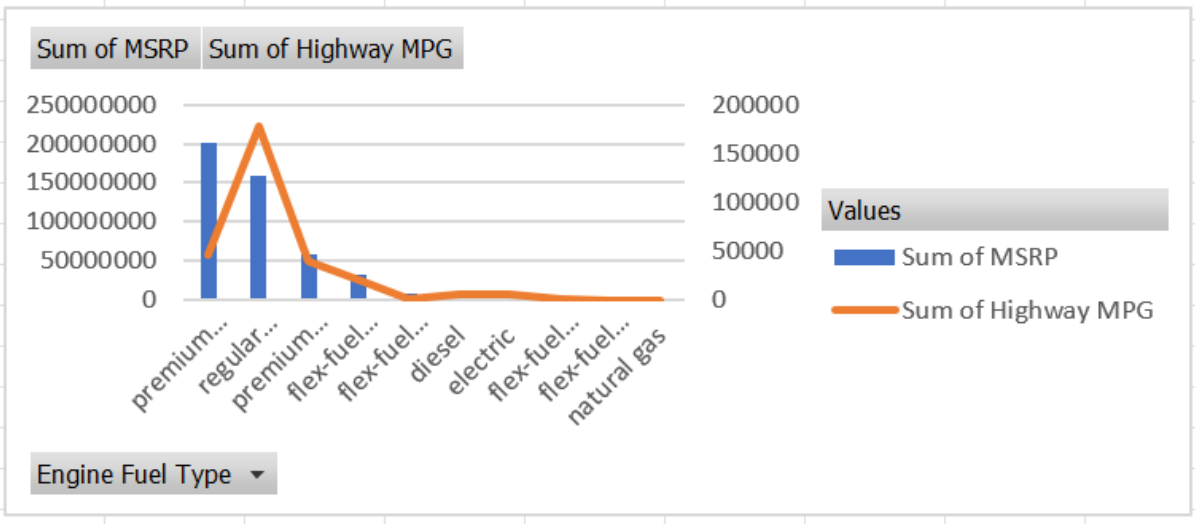


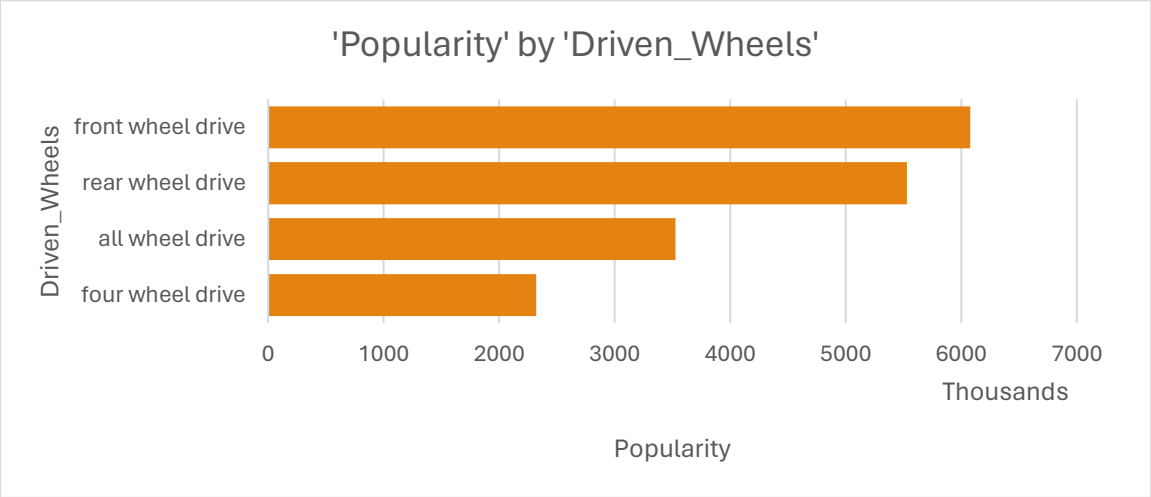


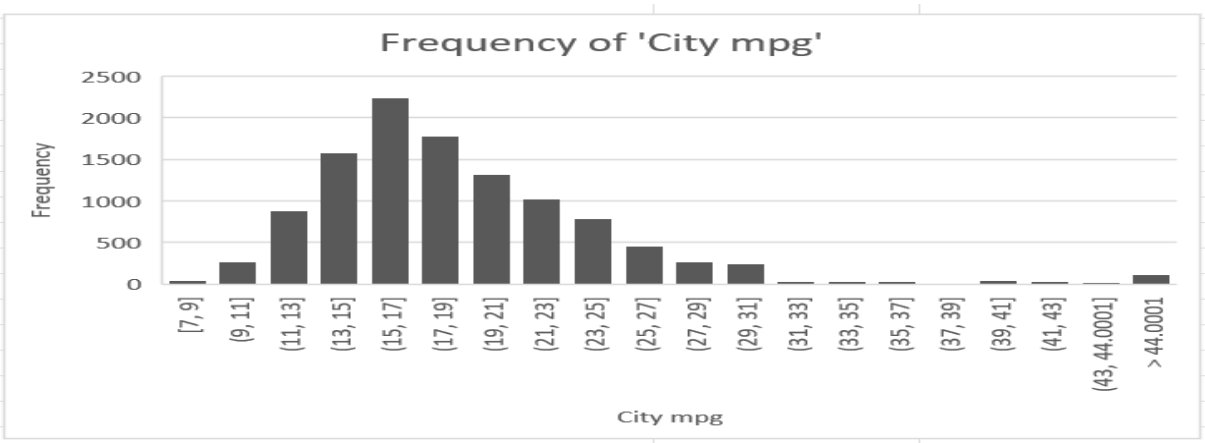

## Some More Analysis:-



Regular Unleaded Engine Fuel Type has less MSRP than Premium Unleaded(required) but still its Sum of Highway MPG is more than other Engine Fuel type.




Cars with Front wheel Drive are more Popular than other Cars.




Majority of the cars bought gives average off 15-17 MPG.

## Conclusion:

Analyzing the impact of car features on pricing and profitability reveals that advanced features, cutting-edge technology, and high-quality materials significantly enhance a car's market value, resulting in higher pricing. These features attract consumers who are willing to pay a premium for enhanced performance, comfort, and safety. However, the inclusion of these advanced features also increases production costs, which can affect profitability.

----------**********----------

# Module 8: ABC Call Volume Trend Analysis

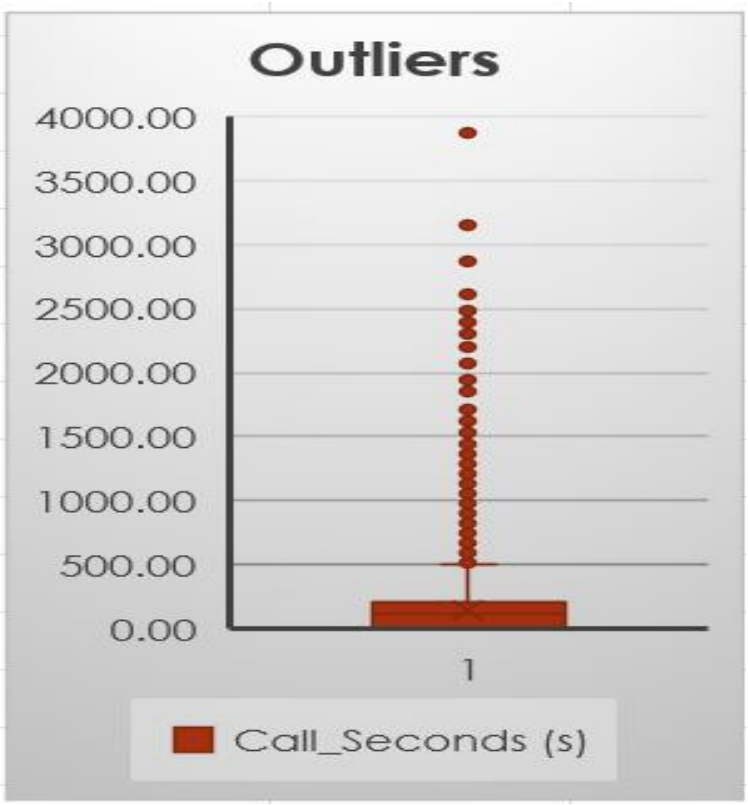

## Description:-

A Customer Experience (CX) team is vital to a company. They review customer feedback and data to generate insights, which they then communicate to the rest of the organization. This team oversees a variety of tasks such as managing customer experience programs, internal communications, mapping customer journeys, and handling customer data.

**Tech Stack Used:-**Microsoft Excel 365

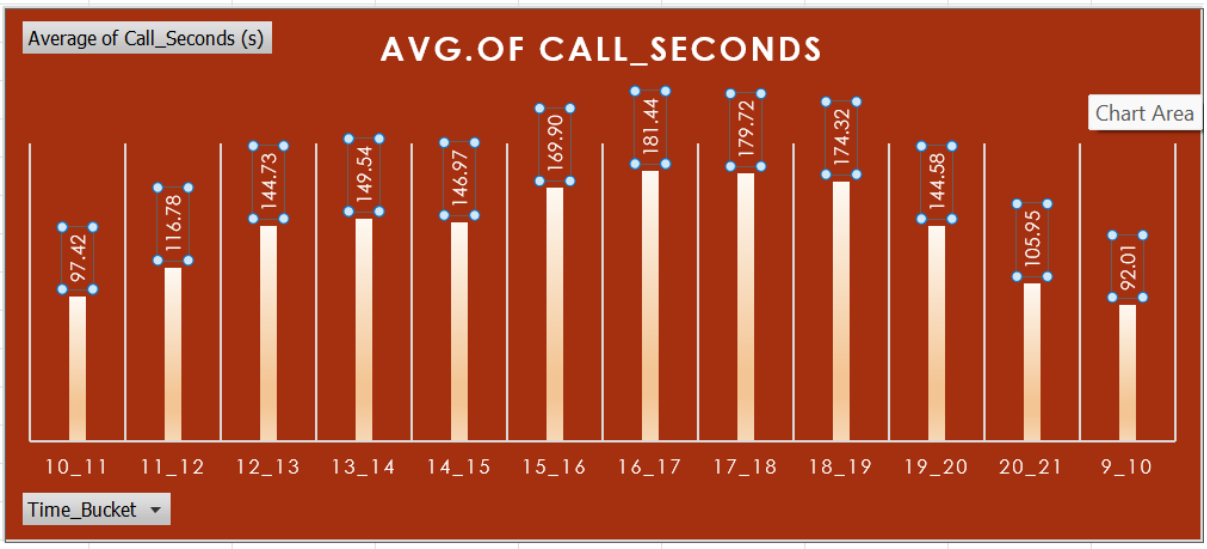
## Exploratory Data Analysis:-

1. It is observed that all the rows where Agent's Name is not mentioned the calls have been abandoned , call duration and wrapped by is also not mentioned. So replacing all the Agent's name from #N/A to Not Mentioned(Abandon).
2. Agent ID is also replaced by Not Mentioned(Abandon)
3. Some values in Wrapped by are missing even when agents have answered the calls. So the Blank cells have been replaced by Agents.
4. 8 Outliers found using Box Plot using threshold of 2500 seconds.




On further investigation, didn't find anything unusual. It just means that in a particular bucket Customers tend to talk more and ask for the I formation. So nothing was changed.

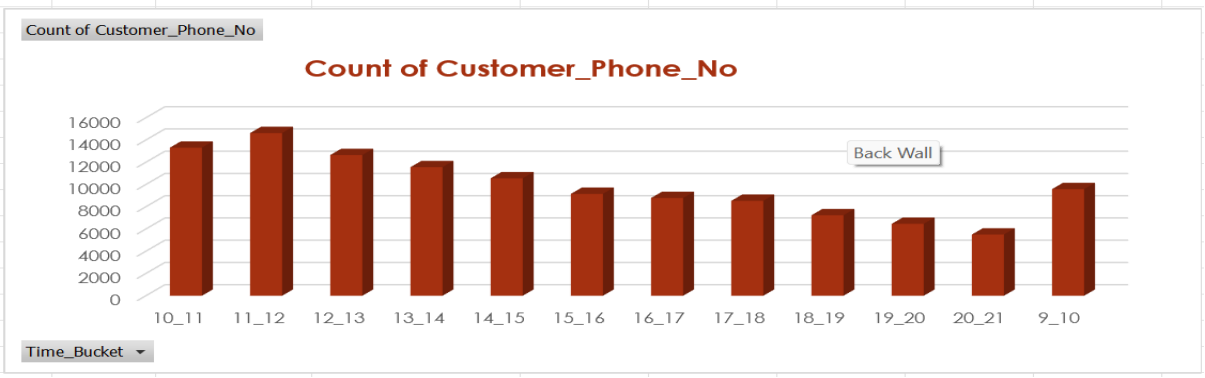
## Task 1

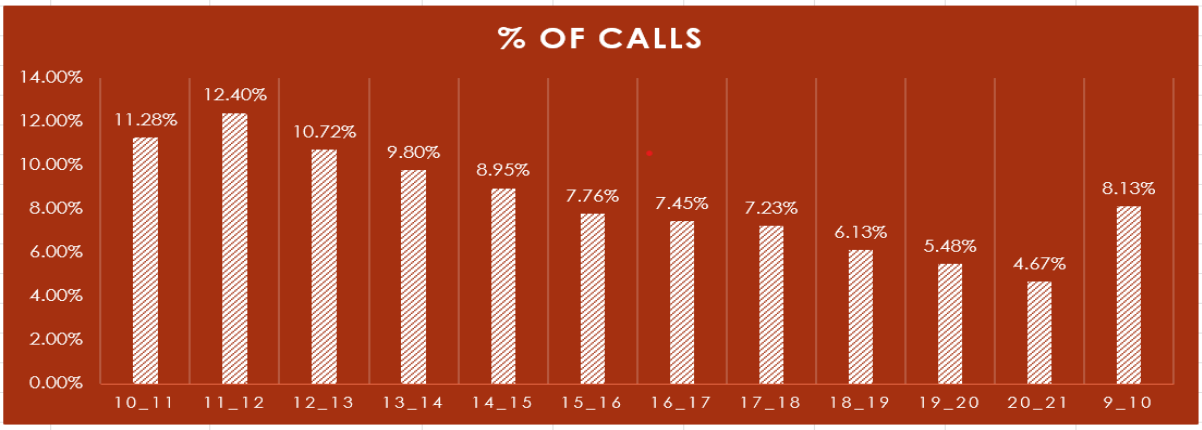



## Insights:-

The graph above displays the average call duration for all incoming calls received by agents during each time bucket. It indicates that agents tend to have longer conversations during the 17:00-18:00 and 18:00-19:00 time buckets.

## Task 2:-



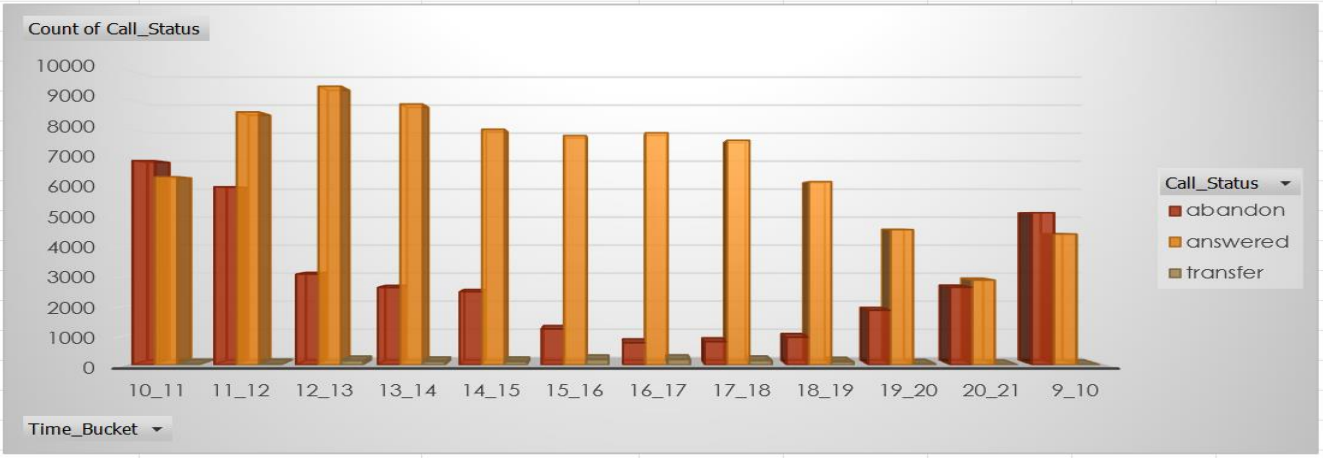




## Insights:-

Analysing call volume data helps identify peak times during the day when call centre activity is highest. In this case, the data reveals that call volume is significantly higher during the late morning hours, specifically between 10:00 AM and 12:00 PM.

## Task 3:-

What is the minimum number of agents required in each time bucket to reduce the abandon rate to 10%?Using Pivot Table and column chart we found the calls abandoned and answered.



As it is assumed that a month has 30 days, Agents work for 6 days a week and take 4 unplanned holidays. So the Agent is employed for 22 days in a month.

| | |
|---|---|
| With assumption total days in a month | 30 |
| 4 weekends | 4 |
| 4 unplanned leaves | 4 |
| Total working days in a month | 22 |

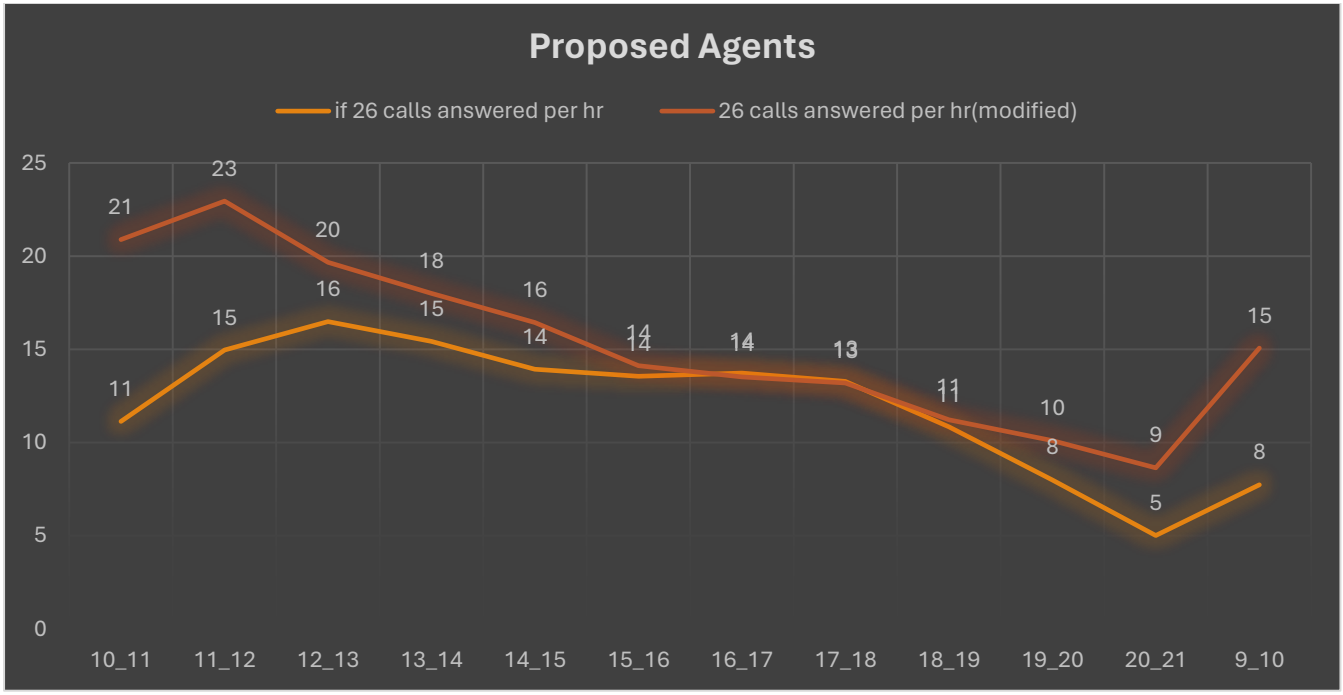
| Time-Bucket | abandon calls for 22 days | answered calls for 22 days | abandon calls for one day | answered calls for one day | Per day calls | Percentage of Abandoned calls | 90 % calls if answered |
|---|---|---|---|---|---|---|---|
| 10_11 | 6911 | 6368 | 314 | 289 | 604 | 52% | 543 |
| 11_12 | 6028 | 8560 | 274 | 389 | 663 | 41% | 597 |
| 12_13 | 3073 | 9432 | 140 | 429 | 568 | 25% | 512 |
| 13_14 | 2617 | 8829 | 119 | 401 | 520 | 23% | 468 |
| 14_15 | 2475 | 7974 | 113 | 362 | 475 | 24% | 427 |
| 15_16 | 1214 | 7760 | 55 | 353 | 408 | 14% | 367 |
| 16_17 | 747 | 7852 | 34 | 357 | 391 | 9% | 352 |
| 17_18 | 783 | 7601 | 36 | 346 | 381 | 9% | 343 |
| 18_19 | 933 | 6200 | 42 | 282 | 324 | 13% | 292 |
| 19_20 | 1848 | 4578 | 84 | 208 | 292 | 29% | 263 |
| 20_21 | 2625 | 2870 | 119 | 130 | 250 | 48% | 225 |
| 9_10 | 5149 | 4428 | 234 | 201 | 435 | 54% | 392 |
| **Grand Total** | **34403** | **82452** | **1564** | **3748** | **5312** | **29%** | **4780** |

Further call made per hour is calculated.

| | |
|---|---|
| **Working hrs per day:** | 9 hrs |
| **Break taken:** | 1.5 hrs |
| **Total hrs to be worked** | 7.5 hrs |
| **60 % of hrs worked** | 4.5 hrs |
| **Converted into seconds** | 16200 |
| **Avg of Call_seconds** | 139.53 |
| **Call that can be made per day** | 116 |
| **Call that can be made per hr** | 26 |

Later the call answered per day was divided by 26 to get Agents working per hour in the company. To optimize the call to 90% ,the proposed 90% answered calls were divided by 26 to get the required number of agents in the company. The data is as follows.

| Time-Bucket | answered calls for one day | if 26 calls answered per hr | 90 % calls if answered | 26 calls answered per hr(modified) |
|---|---|---|---|---|
| 10_11 | 289 | 11 | 543 | 21 |
| 11_12 | 389 | 15 | 597 | 23 |
| 12_13 | 429 | 16 | 512 | 20 |
| 13_14 | 401 | 15 | 468 | 18 |
| 14_15 | 362 | 14 | 427 | 16 |
| 15_16 | 353 | 14 | 367 | 14 |
| 16_17 | 357 | 14 | 352 | 14 |
| 17_18 | 346 | 13 | 343 | 13 |
| 18_19 | 282 | 11 | 292 | 11 |
| 19_20 | 208 | 8 | 263 | 10 |
| 20_21 | 130 | 5 | 225 | 9 |
| 9_10 | 201 | 8 | 392 | 15 |
| **Grand Total** | **3748** | **144** | **4780** | **184** |




## Insights:-

To ensure the call abandonment rate stays below 10%, the call center needs to adjust agent availability based on call volume patterns.

## Task 4:-

Assuming that for every 100 calls in morning 30 calls are received at night. So calculated 30% of the calls answered.

| Time-Bucket | 90 % calls if answered |
|---|---|
| 10_11 | 543 |
| 11_12 | 597 |
| 12_13 | 512 |
| 13_14 | 468 |
| 14_15 | 427 |
| 15_16 | 367 |
| 16_17 | 352 |
| 17_18 | 343 |
| 18_19 | 292 |
| 19_20 | 263 |
| 20_21 | 225 |
| 9_10 | 392 |
| **Grand Total** | 4780 |
| 30% of call at night | 1434 |

From the data given Assuming 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 made. And proposed calls with agents are found.

| Time_Slot | Distribution of 30 calls | % of Distribution of calls | Distribution of calls | 26 calls answered per hr(modified) |
|---|---|---|---|---|
| 9pm-10pm | 3 | 10% | 143 | 6 |
| 10pm-11pm | 3 | 10% | 143 | 6 |
| 11pm-12am | 2 | 7% | 96 | 4 |
| 12am-01am | 2 | 7% | 96 | 4 |
| 01am-02am | 1 | 3% | 48 | 2 |
| 02am-03am | 1 | 3% | 48 | 2 |
| 03am-04am | 1 | 3% | 48 | 2 |
| 04am-05am | 1 | 3% | 48 | 2 |
| 05-am06am | 3 | 10% | 143 | 6 |
| 06am-07am | 4 | 13% | 191 | 7 |
| 07am-08am | 4 | 13% | 191 | 7 |
| 08am-09am | 5 | 17% | 239 | 9 |
| Total | 30 | 100% | 1434 | 55 |

## Insights:-

To effectively manage call center operations and maintain a low call abandonment rate, it is essential to optimize agent availability based on call volume patterns. The highest demand for agents is observed during the morning hours from 9 A.M to 1 P.M, requiring the most number of agents. Conversely, the lowest demand occurs during the night hours from 12 A.M to 5 A.M, requiring the least number of agents.

## Conclusion:-

The ABC Call Volume Trend Analysis project focuses on understanding and enhancing customer experience by analyzing call volume trends and customer feedback. Key components of this project include:

1. **Customer Experience Team**: A dedicated team of professionals who analyze customer feedback and data, sharing valuable insights with the organization to improve customer satisfaction.
2. **Analyzing Customer Relationships**: The project emphasizes the importance of solving customer problems to maintain and enhance the relationship between the business and its customers. This helps in fostering customer loyalty and business growth.
3. **Skill Development**: Working on this project enhances various skills, including data analytical skills, where one learns to interpret and derive insights from data; and visualization skills, which involve creating clear and impactful charts and graphs to represent data trends.

----------**********----------

Dear Trainity Team,

I am writing to express my heartfelt gratitude for developing such an exceptional online portal for teaching data analysis. The learning experience was incredibly smooth and detailed, making complex concepts easy to understand.

The projects you assigned were particularly beneficial, providing hands-on experience that significantly enhanced my skills in Advanced Excel, Python, Statistics, SQL, and Tableau. These practical exercises not only reinforced my understanding but also boosted my confidence in applying these tools and techniques in real-world scenarios.

Thank you once again for your dedication and effort in creating a comprehensive and user-friendly learning platform. Your commitment to quality education has made a profound impact on my professional development.

Sincerely,

Priti Varma