DATA ANALYSIS

PORTFOLIO

PREPARED BY
RAJAT PANWAN

PROFESSIONAL BACKGROUND

As a Regional Lead for a B2B eCommerce company, Udaan.com, I led a team of 14 market representatives in the apparel segment for Northern India. In this role, I executed and analyzed projects in coordination with marketing representatives, and regularly analyzed sales and operational data to support strategic decision making. I also coordinated multiple strategic pilot initiatives and assessed their utility for pan India execution.

Throughout my tenure, I successfully managed and mentored the 14-person team. I led daily team huddles, motivated the team, coached them in execution, trained them on system changes, and recognized exemplary performance. I understood business operations needs and accordingly set targets, communicated execution strategies to subordinates, and addressed problems faced by them in execution in the field. As a finance graduate, I pursued a Post Graduate diploma last year, and I am currently pursuing a Master of Computer Application from distance mode to enhance my technical knowledge. I have also completed various data analysis and data science courses.

With my experience in leading teams, analyzing sales data, and executing strategic initiatives, I am confident in my ability to excel in a data analyst role. My passion for data analysis, technical knowledge, and ability to lead and mentor teams make me an asset to any organization.

TABLE OF CONTENT

1. DATA ANALYTICS PROCESS

1.1 Description

Give the example(s) of such a real-life situation where we use Data Analytics and link it with the data analytics process. You can prepare a PPT/PDF on a real-life scenario explaining it with the process such as Plan, Prepare, Process, Analyze, Share and Act.

1.2 Problem

To get 30% of the total calls which is 35396 because 30% calls are received at night. Then 35396 is divided into different time bucket according to the ratio given.

1.3 Plan

Planning a trip is also a data analysis. Deciding the destination where we want to go, such as Beaches, Hill Stations, Towns and cities.

1.4 Prepare

Check for the budget of the holiday I am willing to spend on the holiday.

1.5 Process

Selecting the Destination. For example, if I decided to go on a hill station what it would be, like Himachal, North-East, J&K or foreign countries like Switzerland and Austria.

1.6 Analyze

Now I will check the weather conditions, travelling duration and political condition are favorable or not.

1.7 Share

Now communicate the idea to the travel agent, to find the best suitable destination under my budget.

1.8 Act

Finally buying the plan from the agent according to my decision.

1.9 Conclusion

In conclusion, data analysis has become an increasingly important aspect of our daily lives. From tracking our health and fitness goals to making informed financial decisions, data analysis can provide valuable insights that help us make better decisions. With the proliferation of data and the tools to analyze it, anyone can take advantage of the power of data analysis. By developing basic data analysis skills and using the right tools, we can gain a deeper understanding of the world around us and make more informed decisions in our personal and professional lives.

2. INSTAGRAM USER ANALYTICS

2.1 Description

In this project, MySQL commands are used to answer the questions asked by our investors and marketing team. The name of the database is "ig_clone" where I performed data manipulation on various tables suchas, comments, follows, likes, photo_tags, photos, tags and users. Joining data from multiple tables using Inner Join, Left Join, and Right Joins also performed.

2.2 Problem

User analysis is the process by which we track how users engage and interact with our digital product (software or mobile application) in an attempt to derive business insights for marketing, product & development teams.

These insights are then used by teams across the business to launch a new marketing campaign, decide on features to build for an app, track the success of the app by measuring user engagement and improve the experience altogether while helping the business grow. You are working with the product team of Instagram and the product manager has asked you to provide insights on the questions asked by the management team.

2.3 Approach

In this project I carefully read and understand the requirements and objective of the project. Then I go through the entire tables of the database to know tables attributes. In tables I use appropriate clauses such as Select, Order By, Group By and Where to extract the use

insights. Once I had formulated the final query, I tested it thoroughly to ensure that it was correctly returning the desired results.

2.4 Tech-Stack used

A Schema management system MySQL 8.0 is used to handle, store and modify and delete data and also store data in an organized way. In this process MySQL workbench is used which comes with MySQL.

2.5 Findings

2.5.1 Finding-1

```
1 • SELECT *
2   FROM users
3   ORDER BY created_at
4   LIMIT 5;
```

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

After analyzing the available data and sorting it by date created, we get the insight that Darby_Herzog is the oldest user followed by Emilio_Bernier52 and Elenor88. We can reword these loyal customers.

2.5.2 Finding-2

```
1 • SELECT *
2   FROM users
3   LEFT JOIN photos
4   ON users.id = photos.user_id
5   WHERE photos.user_id IS NULL;
```

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

id	username	created_at	id	image_url	user_id	created_dat
5	Aniya_Hackett	2016-12-07 01:04:39	NULL	NULL	NULL	NULL
7	Kasandra_Ho...	2016-12-12 06:50:08	NULL	NULL	NULL	NULL
14	Jadyn81	2017-02-06 23:29:16	NULL	NULL	NULL	NULL
21	Rocio33	2017-01-23 11:51:15	NULL	NULL	NULL	NULL
24	Maxwell.Halvo...	2017-04-18 02:32:44	NULL	NULL	NULL	NULL
25	Tierra.Trantow	2016-10-03 12:49:21	NULL	NULL	NULL	NULL
34	Pearl7	2016-07-08 21:42:01	NULL	NULL	NULL	NULL
36	Ollie_Ledner37	2016-08-04 15:42:20	NULL	NULL	NULL	NULL
41	Mckenna17	2016-07-17 17:25:45	NULL	NULL	NULL	NULL

Result 1

Based on the available data we can say that above query help us to identify the inactive users. I would suggest the company to remind these inactive users to start posting, by sending them promotional emails to post their first photo.

2.5.3 Finding-3

```
SELECT
    *
FROM
    (SELECT
        *
    FROM
        (SELECT
        MAX(likes_count) AS max_likes, photo_id
    FROM
        (SELECT
        COUNT(photo_id) likes_count, photo_id
    FROM
        likes
    GROUP BY photo_id
    ORDER BY likes_count DESC) AS photo_likes) AS max_liked_photo
    INNER JOIN photos ON photos.id = max_liked_photo.photo_id) AS most_liked_user_photo
        INNER JOIN
    users ON users.id = most liked user photo.user id;
```

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

max_likes	photo_id	id	image_url	user_id	created_dat	id	username	created_at
48	145	145	https://larret.name	52	2022-11-28 18:37:28	52	Zack Kemmer93	2017-01-01 05:58:22

Based on the available data we can say that photo with photo_id 145 get the most like on a single photo. We can reward the user to motivate other users to share more photos on the platform.

2.5.4 Finding-4

```
1  SELECT
2      COUNT(weekdays) weekday_counts, weekdays
3  FROM
4      (SELECT
5          DAYNAME(created_at) weekdays
6      FROM
7          users) AS week_table
8  GROUP BY weekdays
9  ORDER BY weekday_counts DESC;
```

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

weekday_counts	weekdays
16	Thursday
16	Sunday
15	Friday
14	Tuesday
14	Monday
13	Wednesday
12	Saturday

Result 1

Based on the available data and queries performed on the data we can say that Thursday is day of a week when most of the users register on. I want to suggest the company to launch ad campaign on Thursday would be more appropriate.

2.5.5 Finding-5

```
SELECT
    ((SELECT
            COUNT(id)
        FROM
            photos) / (SELECT
            COUNT(id)
        FROM
            users)) AS divide;
```

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

Average_Post_per_User
3.4730

Result 1

Based on the available data and queries performed on the data we can say that the average user post is around 3.5. I would suggest the company to remind users to start posting by sending them promotional emails to post so average would increase.

2.5.6 Finding-6

```
SELECT
    user_id, COUNT(user_id) AS like_count
FROM
    likes
GROUP BY user_id
HAVING COUNT(user_id) = (SELECT
        COUNT(DISTINCT (photo_id)) AS distinct_photo_id
    FROM
        likes);
```

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

user_id	like_count
5	257
14	257
21	257
24	257
36	257
41	257
54	257
57	257
66	257
71	257
75	257

Result 1

Based on the available data and queries performed on the data we can say that above list of the user likes every single photos on the site, so we can say that these are the fake accounts. We would suggest that Instagram should suspend these accounts.

2.6 Data Analysis

Here are some insights and knowledge that I gained while working on Instagram User Analytics project such as understating of the SQL language and how to use it to retrieve and manipulate data in a database Develop an ability to design and execute complex queries using a range of SQL clauses, functions and operators Skills in data analysis and problem solving as the process of creating an SQL query often involves identifying patterns in the data.

We get the insight that Darby_Herzog is the oldest user followed by Emilio_Bernier52 and Elenor88. We can reword these loyal customers. Identify the inactive users and suggest the company to remind these inactive users to start posting, by sending them promotional emails to post their first photo.

photo with photo_id 145 get the most like on a single photo. We can reward the user to motivate other users to share more photos on the platform.

2.7 Conclusion

In conclusion, I would like to tell that after doing a thorough analysis we were able to derive the insights from the data and was able to plot various graphs using that data. The data that once looked useless became useful and helped to find out the users who is active and the inactive users. Analyzing the data proved helpful in finding various issues among the users.

3. OPERATION ANALYTICS AND INVESTIGATING METRIC SPIKE

3.1 Description:

The project is about analyzing two different data sets related to the end to end operations of a company As a Data Analyst Lead at Microsoft, the aim is to derive insights from the given data sets and answer the questions asked by different departments of the company The first case study is related to job data and requires the calculation of various metrics, such as the number of jobs reviewed, throughput, percentage share of each language, and displaying duplicate rows The second case study is related to investigating metric spikes and requires the calculation of various metrics, such as weekly user engagement, user growth, weekly retention, weekly engagement per device, and email engagement metrics.

3.2 Problem:

Operation Analytics is the analysis done for the complete end to end operations of a company. With the help of this, the company then finds the areas on which it must improve upon. You work closely with the ops team, support team, marketing team, etc. and help them derive insights out of the data they collect.

Being one of the most important parts of a company, this kind of analysis is further used to predict the overall growth or decline of a company's fortune. It means better automation, better understanding between cross-functional teams, and more effective workflows.

Investigating metric spike is also an important part of operation analytics as being a Data Analyst you must be able to understand or make other teams understand questions like- Why is there a dip in daily engagement?

Why have sales taken a dip? Etc. Questions like these must be answered daily and for that its very important to investigate metric spike.

You are working for a company like Microsoft designated as Data Analyst Lead and is provided with different data sets, tables from which you must derive certain insights out of it and answer the questions asked by different departments.

3.3 Approach:

To complete the project, I go through the data provided in each table and understood the relationships between them Next, I created a database metric_spike and tables using SQL queries Once the tables are created, I started performing the analysis by writing SQL queries to answer the questions asked in each case study Finally, compiled the results and insights gained from the analysis.

3.4 Tech Stack Used:

A database management system MySQL 8.0 is used to handle, store and modify and delete data and also store data in an organized way In this process MySQL Workbench in used which comes with MySQL.

3.5 Findings:

3.5.1 Finding-1:

```
SELECT
    COUNT(*) / (24*30) AS job
FROM
    job_data
WHERE
    ds >= '2020-11-01' AND ds < '2020-12-01';
```

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

job
0.0111

Based on available data and queries performed on it we can say that jobs reviewed per hour per day for November 2020 is 0.0111.

3.5.2 Finding-2:

```
1  SELECT ds, AVG(event_per_second)
2  OVER (ORDER BY ds ROWS BETWEEN 6 PRECEDING AND CURRENT ROW) AS rolling_avg
3  FROM (
4  SELECT ds, COUNT(event) / SUM(time_spent) AS event_per_second
5  FROM job_data
6  GROUP BY ds
7  )rolling_avg_tbl;
```

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

ds	rolling_avg
2020-11-25	0.02220000
2020-11-26	0.02005000
2020-11-27	0.01656667
2020-11-28	0.02757500
2020-11-29	0.03206000
2020-11-30	0.03505000

Throughput is the no. of events happening per second. Based on the available data here is the 7 day rolling average of throughput. 7 day rolling average is used to smoothing out the fluctuations, end make it easier to identify long term trend.

3.5.3 Finding-3:

```
SELECT language, COUNT(job_id), 100 * COUNT(job_id) / SUM(COUNT(job_id))
OVER() AS language_share
FROM job_data
GROUP BY language;
```

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

language	COUNT(job_id)	language_share
English	1	12.5000
Arabic	1	12.5000
Persian	3	37.5000
Hindi	1	12.5000
French	1	12.5000
Italian	1	12.5000

Based on the available data we can say that maximum content is in Persian language which is 37.5% and all of the other five language has 12.5% share.

3.5.4 Finding-4:

```
SELECT *
FROM job_data
WHERE (ds, job_id, actor_id, event, language, time_spent, org) IN (
    SELECT ds, job_id, actor_id, event, language, time_spent, org
    FROM job_data
    GROUP BY ds, job_id, actor_id, event, language, time_spent, org
    HAVING COUNT(*) > 1
    )
ORDER BY ds DESC, job_id DESC;
```

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

ds	job_id	actor_id	event	language	time_spent	org

Based on the available data we can say that their isn't any duplicate rows in the data. Which is good for analysis.

3.5.5 Finding-5:

```
1  SELECT
2      user_id,
3      ((COUNT(*)) / (COUNT(DISTINCT (EXTRACT(WEEK FROM occurred_at))))) AS user_engagement_per_week
4  FROM
5      events
6  GROUP BY user_id
7  ORDER BY user_id;
```

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

user_id	user_engagement_per_week
4	10.3333
8	7.2000
11	31.5000
17	27.5000
19	14.2000
20	11.3333
22	44.6250
30	8.3750
49	5.0000
59	42.8333

Result 30

From the above query we can calculate the user engagement per week to measure the activeness of a user. By measuring user engagement per week we can say if the user finds quality in a product/service.

3.5.6 Finding-6:

```sql
SELECT
  YEAR(created_at) AS year,
  MONTH(created_at) AS month,
  COUNT(DISTINCT user_id) AS user_counts,
  (COUNT(DISTINCT user_id)
    / LAG(COUNT(DISTINCT user_id)) OVER (ORDER BY YEAR(created_at), MONTH(created_at)) - 1)*100
    AS user_percentage_growth
FROM users
GROUP BY year, month;
```

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

year	month	user_counts	user_percentage_growth
2013	1	332	NULL
2013	2	328	-1.2048
2013	3	383	16.7683
2013	4	410	7.0496
2013	5	486	18.5366
2013	6	485	-0.2058
2013	7	608	25.3608
2013	8	636	4.6053
2013	9	699	9.9057
2013	10	826	18.1688

Result 9

This is how we can know what is the growth of the users for a particular product. Here user growth means amount of users growing over time for a product.

3.5.7 Finding-7:

```sql
SELECT EXTRACT(WEEK FROM occurred_at) AS weeks,
       COUNT(CASE WHEN e.event_type = 'signup_flow' THEN e.user_id ELSE NULL END) AS signup
FROM events e
GROUP BY weeks
ORDER BY weeks;
```

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

weeks	signup
17	385
18	901
19	954
20	955
21	961
22	1042
23	1065
24	1158
25	1075
26	1065

Result 8

Based on the available data we can say that weekly retention of users-sign up cohort is increasing every week. On week 24 there is highest value of the users-sign up cohort, then it start to decrease.

3.5.8 Finding-8:

```
SELECT
    device,
    (COUNT(event_name) /
    (COUNT(DISTINCT(EXTRACT(WEEK FROM occurred_at))))) AS weekly_avg_engagement_per_device
FROM
    events
GROUP BY device
ORDER BY weekly_avg_engagement_per_device DESC;
```

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

device	weekly_avg_engagement_per_device
macbook pro	3155.1579
lenovo thinkpad	2035.7368
macbook air	1479.1579
iphone 5	1428.1053
dell inspiron notebook	1077.6842
samsung galaxy s4	1031.2632
nexus 5	907.8421
iphone 5s	879.2105
dell inspiron desktop	556.2632
iphone 4s	531.4211

Result 13

Based on the above extract table we can say that weekly engagement per device is highest for "Macbook pro" followed by "Lenovo Thinkpad". That means user finds quality in of the service on the "Macbook Pro" as compare to any other devices.

3.5.9 Finding-9:

```
SELECT
  user_id,
  COUNT(*) AS email_events_count,
  SUM(CASE WHEN action = 'email_open' THEN 1 ELSE 0 END) AS email_opens_count,
  SUM(CASE WHEN action = 'email_clickthrough' THEN 1 ELSE 0 END) AS email_clickthrough_count,
  SUM(CASE WHEN action = 'sent_weekly_digest' THEN 1 ELSE 0 END) AS sent_weekly_digest_count,
  SUM(CASE WHEN action = 'sent_reengagement_email' THEN 1 ELSE 0 END) AS sent_reengagement_email_count
FROM email_events
GROUP BY user_id;
```

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

user_id	email_events_count	email_opens_count	email_clickthrough_count	sent_weekly_digest_count	sent_reengagement_email_count
0	22	5	0	17	0
4	26	5	4	17	0
8	21	3	1	17	0
11	24	5	2	17	0
17	22	4	1	17	0
19	23	5	1	17	0
20	28	8	3	17	0
22	27	7	3	17	0
30	25	6	1	18	0
49	23	5	1	17	0

Result 3

With help of the above email engagement matrix, we can say that most used email service is "email weekly digest" followed by "email opens". In the above matrix email event is the total of all the email services.

3.6 Data Analysis:

Here are some insights and knowledge that I gained while working on the project.

jobs reviewed per hour per day for November 2020 is 0.0111.

The maximum content is in Persian language which is 37.5% and all of the other five language has 12.5% share.

There aren't any duplicate rows in the data. Which is good for analysis.

3.7 Conclusion:

In conclusion, I would like to tell that after doing a thorough analysis we were able to derive the insights from the data and was able to plot various graphs using that data. By providing valuable insights to different departments, we can help improve cross-functional understanding, workflows, and overall company growth.

4. HIRING PROCESS ANALYTICS

4.1 Description:

This project is about performing exploratory data analysis on a dataset provided by a company for their hiring process. As a lead data analyst, the task is to analyze the data to draw insights and provide recommendations to the hiring department. The project will involve understanding the data columns, checking for missing data, clubbing columns with multiple categories, checking for outliers, removing outliers, and drawing data summaries. The analysis will be performed using Excel or Google Sheets.

4.2 Problem:

Hiring process is the fundamental and the most important function of a company. Here, the MNCs get to know about the major underlying trends about the hiring process. Trends such as- number of rejections, number of interviews, types of jobs, vacancies etc. are important for a company to analyse before hiring freshers or any other individual. Thus, making an opportunity for a Data Analyst job here too!
Being a Data Analyst, your job is to go through these trends and draw insights out of it for hiring department to work upon.

4.3 Approach:

The first step in the project was to download the dataset provided by the Trainity and understand the data columns and data. This involved checking the format of the data, identifying the variables, and understanding the range of values for each variable. The next step was to check for missing data. After that, I checked for outliers and

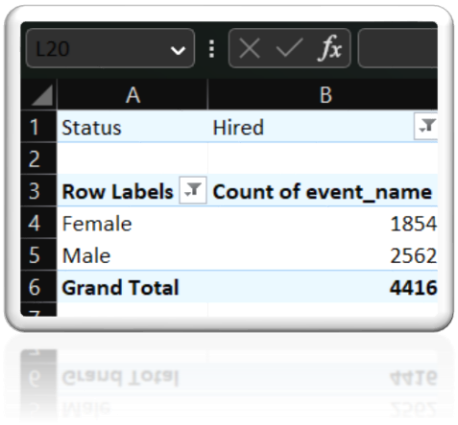
removed them using statistical methods. I then drew data summaries and analyzed the data to draw insights and provide recommendations.

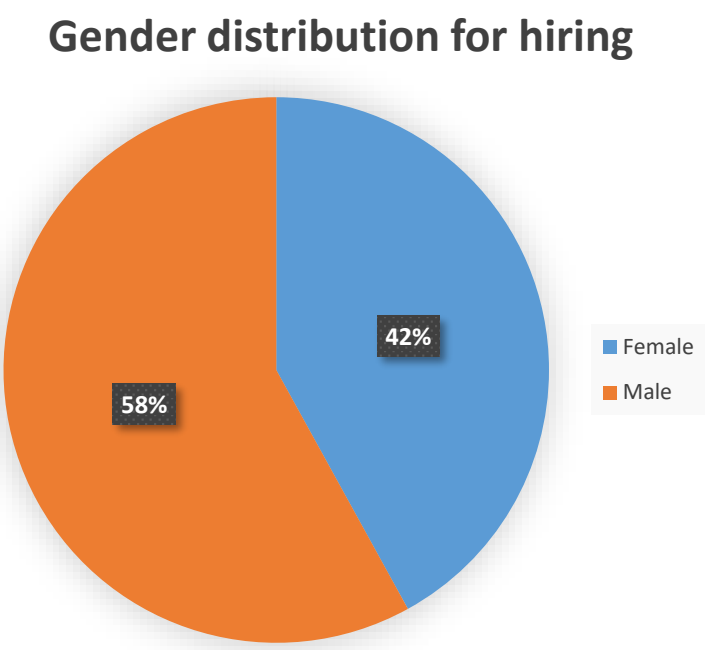
4.4 Tech-Stack Used:

The analysis was performed using Microsoft Excel. We used different Excel functions, such as COUNT, COUNTIFS, SUM, UNIQUE and AVERAGE to perform the analysis. We also used PIVOT TABLE, SLICER in Pivot Table, CHARTS and GRAPHS to visualize the data.

4.5 Findings:

4.5.1 Finding-1:





After plotting the pie chart for gender distribution, we can say that there is more males are applied for the job as compare to the females.

4.5.2 Finding-2:

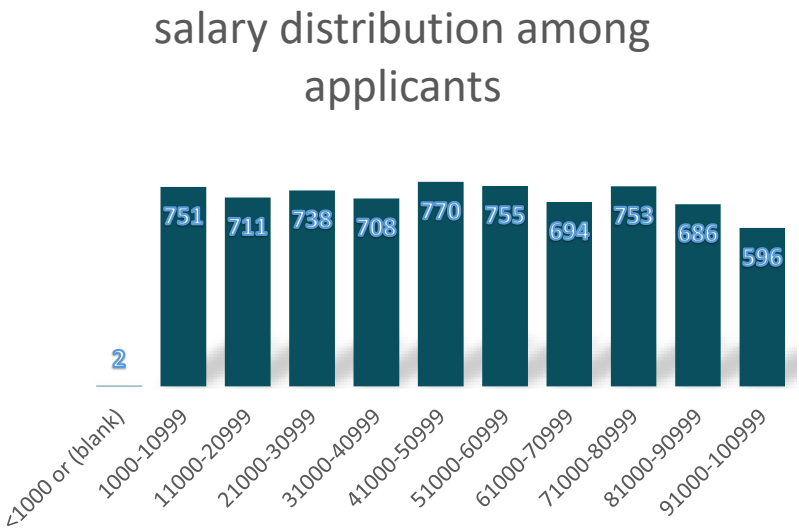
J7 =J2/J3

	I	J
1		
2	Total of the Salaries Offered =	357328269
3	Number of the applications (Counts) =	7163
4		
5		
6		
7	**Avg Salary After Removing Outliers =**	49885.281

With the help of above table and formula used in the table we can say that the average salary offered to the applicants is around 50,000. This is the salary after removing the outliers from the data.

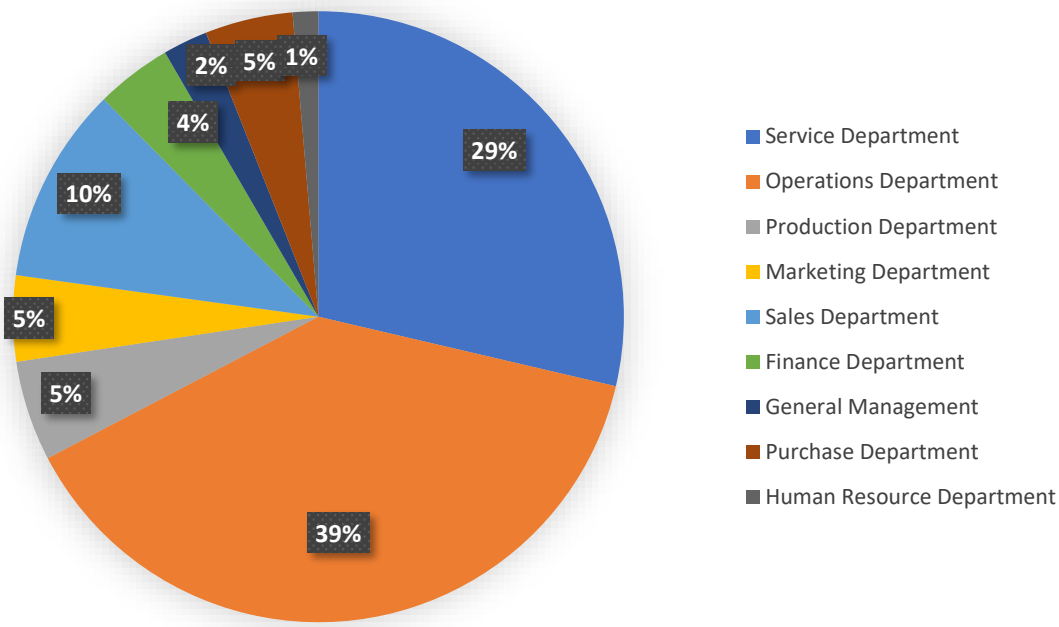
4.5.3 Finding-3:

Row Labels	Count of application_id
<1000 or (blank)	2
1000-10999	751
11000-20999	711
21000-30999	738
31000-40999	708
41000-50999	770
51000-60999	755
61000-70999	694
71000-80999	753
81000-90999	686
91000-100999	596
Grand Total	7164



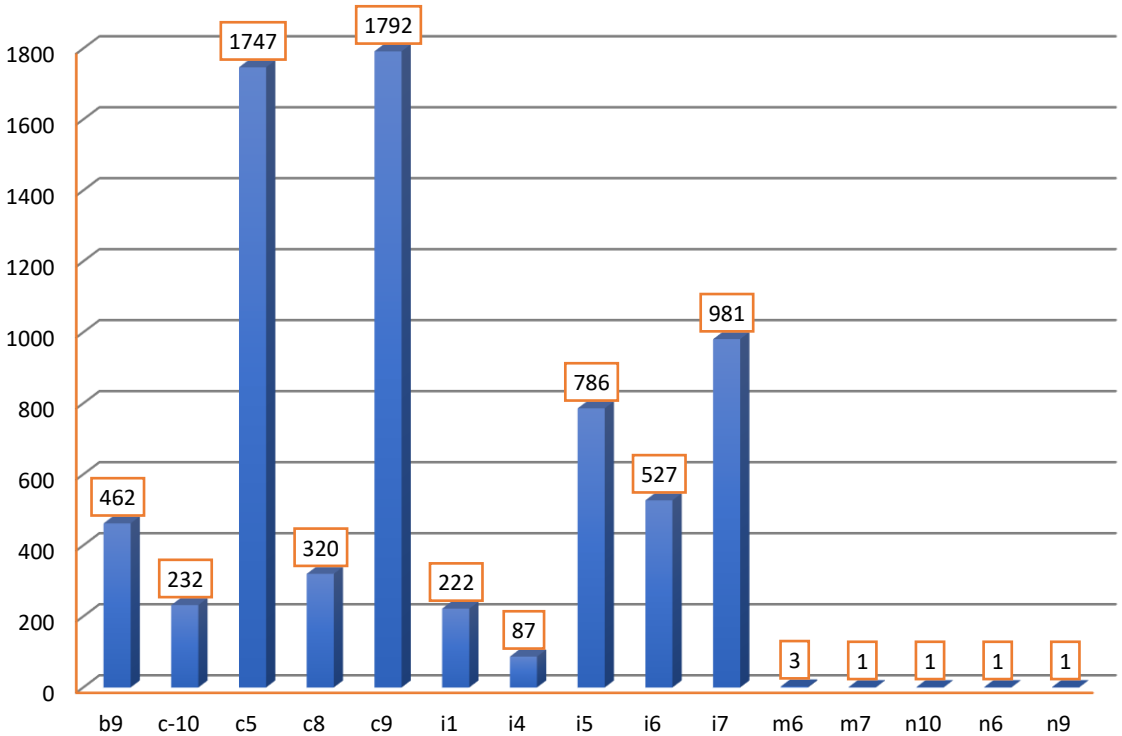
With the help of the above bar chart and table we can say that salary is distributed among the applicants equally, there is very less fluctuations in the number of applicants.

4.5.4 Finding-4:



Based on the available data and pie chart drawn from that data, we can say that most of the applicants apply for the Operations Department followed by Service Department and sales Department. Least applied department is Human Resource Department.

4.5.5 Finding-5:



Based on the data given we can say that C9 is the most applied post tier followed by C5 and then i7. The least applied post tier is m7, n10, n6, n9 which as only one applicant each.

4.6 Data Analysis:

The analysis of the dataset provided insights into the hiring process of the company.

I found that the company had hired more males than females. The average salary offered by the company was found to be ₹49885.

I drew the class intervals for salary and found that the salary is equally distributed between ₹1000 to ₹100000.

The pie chart graph showed that most of the employees were working in the Operations department, followed by the Service department.

We also represented different post tiers using a chart and found that most of the employees were in C5 and C9 position

4.7 Conclusion:

In conclusion, I would like to tell that after doing a thorough analysis we were able to derive the insights from the data and was able to plot various graphs using that data. The data that once looked useless became useful and helped to find out to whom be should hire more and in which department we should hire more employees. Analyzing the data proved helpful in finding various issues among the employees and their hiring process.

5. IMDB MOVIE ANALYSIS

5.1 Description:

The aim of this project is to perform an analysis on a dataset containing information on various movies from IMDB. The dataset includes columns such as the director name, gross, genres, movie title, num voted users, plot keywords, num user for reviews, language, rating, budget, IMDB score etc. The main objective is to extract useful insights from the data and identify any trends or patterns that can be useful for decision-making.

5.2 Problem:

Company is providing you with dataset having various columns of different IMDB Movies. You are required to Frame the problem. For this task, you will need to define a problem you want to shed some light on.

This is where you frame the problem i.e. What is the problem?

These questions to guide your thinking:

1. What do you see happening?
2. What is your hypothesis for the cause of the problem? (this will be broadly based on intuition initially)
3. What is the impact of the problem on stakeholders?
4. What is the impact of the problem not being solved?

Answering these questions will help you define a problem you are trying to solve and will allow you to find the right data to solve it.

5.3 Approach:

The project involved several steps including data cleaning, data visualization, and statistical analysis. Initially, the dataset was explored to identify any missing values, outliers, or errors. The data was then cleaned

using various techniques such as removing duplicates and correcting errors, after deleting rows with duplicates and blanks only 3956 rows are left.. Data visualization tools were used to create charts, graphs, and histograms to analyse the data.

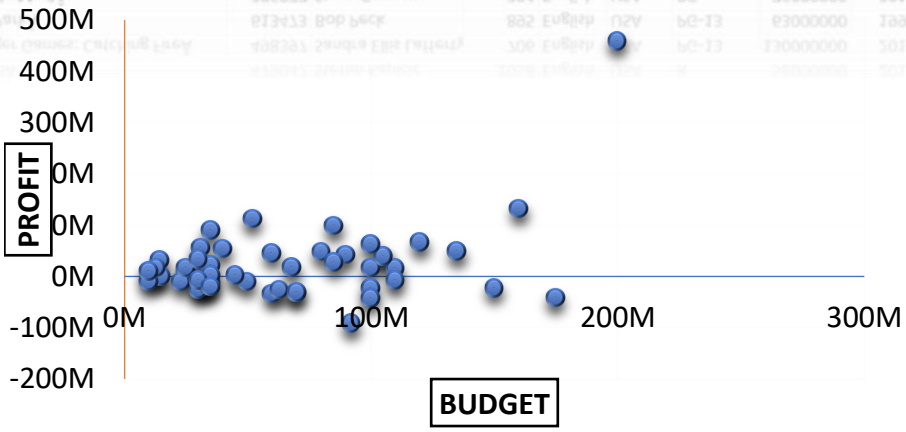
5.4 Tech-Stack Used:

The project was performed using Excel version 2021. Excel was chosen for its powerful data analysis and visualization capabilities. It is also widely used in the industry and provides a familiar environment for users.

5.5 Findings:

5.5.1 Finding-1:

	G	H	I	J	K	L	M	N	O	P	Q
1	movie_title	num_vote	actor_3_name	num_use	language	country	content_ra	budget	title_yea	imdb_s	Profit
2	Jurassic WorldÂ	418214	Omar Sy	1290	English	USA	PG-13	150000000	2015	7	502177271
3	TitanicÂ	793059	Gloria Stuart	2528	English	USA	PG-13	200000000	1997	7.7	458672302
4	Star Wars: Episode IV - A New HopeÂ	911097	Kenny Baker	1470	English	USA	PG	11000000	1977	8.7	449935665
5	E.T. the Extra-TerrestrialÂ	281842	Peter Coyote	515	English	USA	PG	10500000	1982	7.9	424449459
6	The AvengersÂ	995415	Scarlett Johansson	1722	English	USA	PG-13	220000000	2012	8.1	403279547
7	The Lion KingÂ	644348	Niketa Calame	656	English	USA	G	45000000	1994	8.5	377783777
8	Star Wars: Episode I - The Phantom Me	534658	Ian McDiarmid	3597	English	USA	PG	115000000	1999	6.5	359544677
9	The Dark KnightÂ	1676169	Morgan Freeman	4667	English	USA	PG-13	185000000	2008	9	348316061
10	The Hunger GamesÂ	701607	Anthony Reynolds	1959	English	USA	PG-13	78000000	2012	7.3	329999255
11	DeadpoolÂ	479047	Stefan Kapicic	1058	English	USA	R	58000000	2016	8.1	305024263
12	The Hunger Games: Catching FireÂ	498397	Sandra Ellis Lafferty	706	English	USA	PG-13	130000000	2013	7.6	294645577
13	Jurassic ParkÂ	613473	Bob Peck	895	English	USA	PG-13	63000000	1993	8.1	293784000
14	Despicable Me 2Â	286877	Steve Coogan	284	English	USA	PG	76000000	2013	7.5	292049628



Based on the data available we can say that "Jurassic World" has made the maximum profit. "The host" is the outlier with budget 12,215 million and loss of 12,213 million. The host should be removed while performing the

data analysis on the data because it'll impact the statistical measures such as the mean and standard deviation. This can lead to incorrect interpretations of the data and inaccurate conclusions.

5.5.2 Finding-2:

IMDb_Top_250	Rank
The Shawshank RedemptionÂ	1
The GodfatherÂ	2
The Dark KnightÂ	3
The Godfather: Part IIÂ	4
The Lord of the Rings: The Return of the KingÂ	5
Pulp FictionÂ	6
Schindler's ListÂ	7
The Good, the Bad and the UglyÂ	8
Forrest GumpÂ	9
Star Wars: Episode V - The Empire Strikes BackÂ	10
The Lord of the Rings: The Fellowship of the RingÂ	11
InceptionÂ	12
Fight ClubÂ	13
Star Wars: Episode IV - A New HopeÂ	14
The Lord of the Rings: The Two TowersÂ	15
The MatrixÂ	16
One Flew Over the Cuckoo's NestÂ	17
GoodfellasÂ	18
City of GodÂ	19
Seven SamuraiÂ	20
Saving Private RyanÂ	21
The Silence of the LambsÂ	22
Se7enÂ	23
InterstellarÂ	24
The Usual SuspectsÂ	25
American History XÂ	26
Modern TimesÂ	27

With the help of the given data, we can say that "The Shawshank Redemption" is get highest rating in the movies number of voted user is greater than 25,000 and get the rank 1. Number of voted user is taken greater than 25,000 because it ensures the consistent sample size and not influenced by outliers or irregularities in the voting pattern.

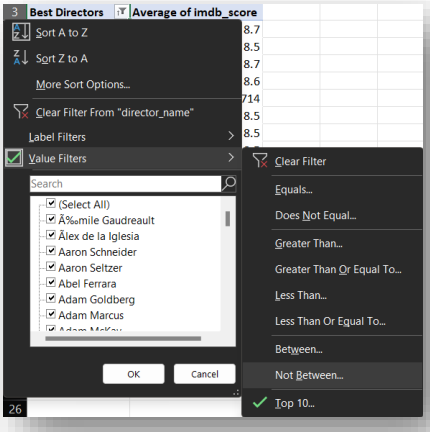
5.5.3 Finding-3:

	R	S	T
1	IMDb_Top_250	Rank	Top_Foreign_Lang_Film_
2	The Shawshank RedemptionÂ	1	
3	The GodfatherÂ	2	
4	The Dark KnightÂ	3	
5	The Godfather: Part IIÂ	4	
6	The Lord of the Rings: The Return of the KingÂ	5	
7	Pulp FictionÂ	6	
8	Schindler's ListÂ	7	
9	The Good, the Bad and the UglyÂ	8	The Good, the Bad and the UglyÂ
10	Forrest GumpÂ	9	
11	Star Wars: Episode V - The Empire Strikes BackÂ	10	
12	The Lord of the Rings: The Fellowship of the RingÂ	11	
13	InceptionÂ	12	
14	Fight ClubÂ	13	
15	Star Wars: Episode IV - A New HopeÂ	14	
16	The Lord of the Rings: The Two TowersÂ	15	
17	The MatrixÂ	16	
18	One Flew Over the Cuckoo's NestÂ	17	
19	GoodfellasÂ	18	
20	City of GodÂ	19	City of GodÂ
21	Seven SamuraiÂ	20	Seven SamuraiÂ
22	Saving Private RyanÂ	21	
23	The Silence of the LambsÂ	22	
24	Se7enÂ	23	
25	InterstellarÂ	24	
26	The Usual SuspectsÂ	25	
27	American History XÂ	26	
28	Modern TimesÂ	27	

In the above figure column ‘T’ is including the name of the movies which are not in ‘English’ out of the top rated 250 movies which has number of voted users more than 25000.

5.5.4 Finding-4:

	Best Directors	Average of imdb_score
3		
4	Akira Kurosawa	8.7
5	Alfred Hitchcock	8.5
6	Cary Bell	8.7
7	Charles Chaplin	8.6
8	Christopher Nolan	8.414285714
9	Damien Chazelle	8.5
10	Majid Majidi	8.5
11	Ron Fricke	8.5
12	Sergio Leone	8.433333333
13	Tony Kaye	8.6
14	**Grand Total**	**8.488888889**



In the above figure pivot table function is used to calculate the top 10 directors name with highest Average value of IMDb score. “Akira Kurosawa” and “Cary Bell” are the highest rated directors.

5.5.5 Finding-5:

genres1	Count
Crime	550
Action	778
Biography	134
Western	40
Comedy	1252
Drama	1338
Adventure	554
Animation	125
Horror	376
Mystery	295
Sci-Fi	390
Document	57
Family	337
Fantasy	394
Musical	67
Romance	678
Thriller	902
0	0
War	93
Music	126
History	86
Sport	115
Short	1
News	1
Film-Noir	1

From the given data genre is counted using Count() function of Excel to get the popular genre. From the above table can say that 'Drama' is the most popular genre followed by comedy and then thriller.

5.5.6 Finding-6:

	Meryl_Streep	Leo_Caprio	Brad_Pitt	actor_1_name	combined
14			Interview with the Vampire: The Vampire ChroniclesÂ	Brad Pitt	Interview with the Vampire: The Vampire ChroniclesÂ
15			FuryÂ	Brad Pitt	FuryÂ
16			Fight ClubÂ	Brad Pitt	Fight ClubÂ
17			By the SeaÂ	Brad Pitt	By the SeaÂ
18			BabelÂ	Brad Pitt	BabelÂ
19		TitanicÂ		Leonardo DiCaprio	TitanicÂ
20		The Wolf of Wall StreetÂ		Leonardo DiCaprio	The Wolf of Wall StreetÂ
21		The RevenantÂ		Leonardo DiCaprio	The RevenantÂ
22		The Quick and the DeadÂ		Leonardo DiCaprio	The Quick and the DeadÂ
23		The Man in the Iron MaskÂ		Leonardo DiCaprio	The Man in the Iron MaskÂ
24		The Great GatsbyÂ		Leonardo DiCaprio	The Great GatsbyÂ
25		The Great GatsbyÂ		Leonardo DiCaprio	The Great GatsbyÂ
26		The DepartedÂ		Leonardo DiCaprio	The DepartedÂ
27		The BeachÂ		Leonardo DiCaprio	The BeachÂ
28		The AviatorÂ		Leonardo DiCaprio	The AviatorÂ
29		Shutter IslandÂ		Leonardo DiCaprio	Shutter IslandÂ
30		Romeo + JulietÂ		Leonardo DiCaprio	Romeo + JulietÂ
31		Revolutionary RoadÂ		Leonardo DiCaprio	Revolutionary RoadÂ
32		Marvin's RoomÂ		Leonardo DiCaprio	Marvin's RoomÂ
33		J. EdgarÂ		Leonardo DiCaprio	J. EdgarÂ
34		InceptionÂ		Leonardo DiCaprio	InceptionÂ
35		Gangs of New YorkÂ		Leonardo DiCaprio	Gangs of New YorkÂ
36		Django UnchainedÂ		Leonardo DiCaprio	Django UnchainedÂ
37		Catch Me If You CanÂ		Leonardo DiCaprio	Catch Me If You CanÂ
38		Body of LiesÂ		Leonardo DiCaprio	Body of LiesÂ
39		Blood DiamondÂ		Leonardo DiCaprio	Blood DiamondÂ
40	The River WildÂ			Meryl Streep	The River WildÂ
41	The Iron LadyÂ			Meryl Streep	The Iron LadyÂ
42	The HoursÂ			Meryl Streep	The HoursÂ
43	The Devil Wears PradaÂ			Meryl Streep	The Devil Wears PradaÂ
44	Out of AfricaÂ			Meryl Streep	Out of AfricaÂ
45	One True ThingÂ			Meryl Streep	One True ThingÂ
46	Lions for LambsÂ			Meryl Streep	Lions for LambsÂ
47	Julie & JuliaÂ			Meryl Streep	Julie & JuliaÂ

Above table is created to extract the movies of the 'Meryl Steep', 'Leonardo DiCaprio' and 'Brad Pitt'. Their combined list is also displayed here. To get the name of the movies I here used IF() function of MS Excel.

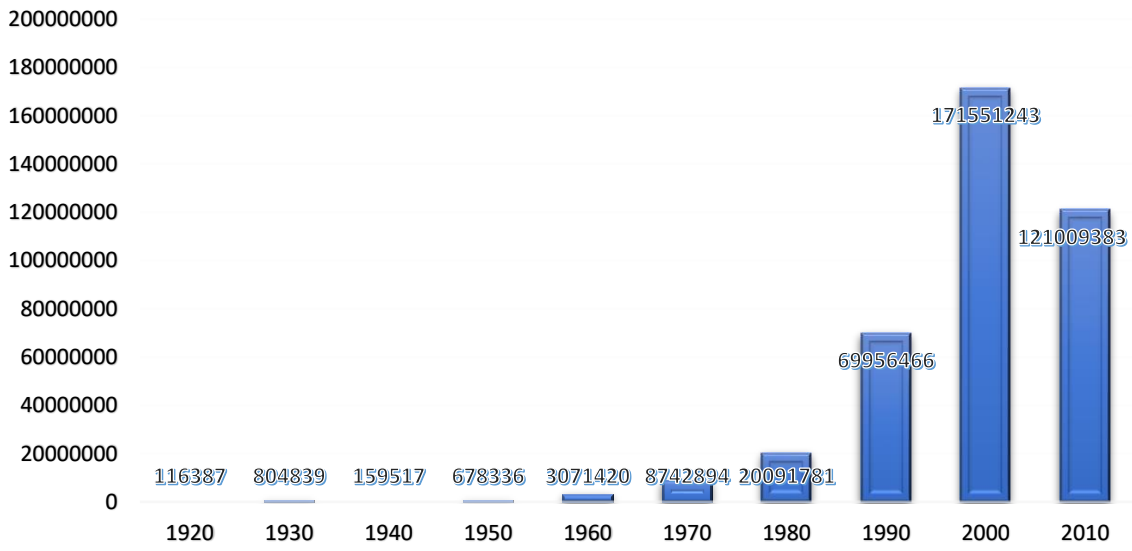
5.5.7 Finding-7:

Row Labels	Average of num_critic_for_reviews
Phaldut Sharma	738
Peter Capaldi	654
Craig Stark	596
BÃ©rÃ©nice Bejo	576
Suraj Sharma	552
Ellar Coltrane	548
Mike Howard	546
Lou Taylor Pucci	543
Maika Monroe	533
Tim Holmes	525
Albert Finney	510
Elina Alminas	489
Kurt Fuller	487
Iko Uwais	481
QuvenzhanÃ© Wallis	478.6666667
Edgar Arreola	478
Sharlto Copley	472
Cory Hardrict	452
Elizabeth McGovern	447
Aidan Turner	447
Wood Harris	432
Anil Kapoor	418
Jessica Barden	417
Chris Hemsworth	411.7333333
Danielle Kotch	411

Row Labels	Average of num_user_for_reviews
Heather Donahue	3400
Christo Jivkov	2814
Steve Bastoni	2789
Phaldut Sharma	1885
Keir Dullea	1736
Chen Chang	1641
Nick Stahl	1562
Kevin Rankin	1445
Noah Huntley	1441
Osama bin Laden	1416
Seychelle Gabriel	1382
Mathieu Kassovitz	1314
Eva Green	1290
Essie Davis	1285.5
Sharlto Copley	1262
Giancarlo Giannini	1243
Orlando Bloom	1242.333333
Luenell	1198
Micah Sloat	1189
Fionnula Flanagan	1109
Jim Meskimen	1107
Ivana Baquero	1083
Henry Cavill	1066.857143
Mhairi Calvey	1065
Talulah Riley	1058

The above lists is extracted using the pivot table function of the Excel. Based on the above data we can say that actor who has the highest mean of the 'number critics for review' is Actor 'Phaldut Sharma'. And actor who has the highest mean of 'number user for reviews' is 'Heather Donahue'.

5.5.8 Finding-8:



I extract the decade from the 'title year' column and calculate the number of voted users sorted in decades. From the above bar graph, we can say that 2000 is the decade where number of voted users is maximum.

5.6 Data Analysis:

Several useful insights were obtained from the analysis. For example, Jurassic World has made the highest profit . The analysis also showed that movies Shawshank Redemption is highest rated movie on IMDb. Another interesting finding was that movies directed by Cary Bell and Akira Kurosawa tended to have higher ratings than others. The analysis also identified certain genres that were more popular than others.

5.7 Conclusion:

In conclusion, I would like to tell that after doing a thorough analysis we were able to derive the insights from the data and was able to plot various graphs using that data. The data that once looked useless became useful and helped to find out the top rated movies and top rate directors. Also find the movies in other then in English language. Analyzing the data proved helpful in finding insight.

6. BANK LOAN CASE STUDY

6.1 Description:

The project aims to use exploratory data analysis (EDA) to analyze loan application data and identify patterns that indicate if a client has difficulty paying their installments. The project aims to understand the driving factors behind loan default, i.e. the variables which are strong indicators of default, and utilize this knowledge for the company's portfolio and risk assessment.

6.2 Problem:

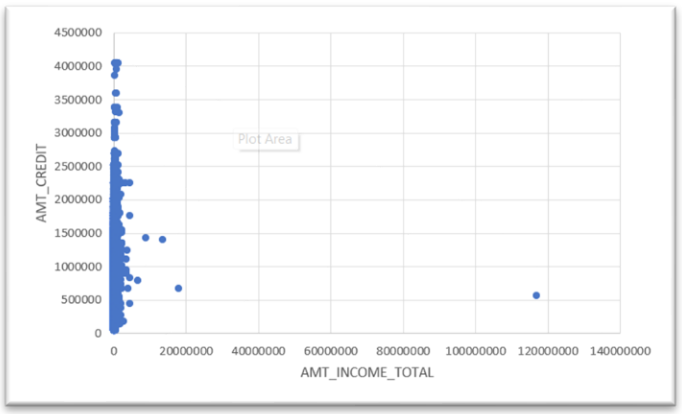
The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

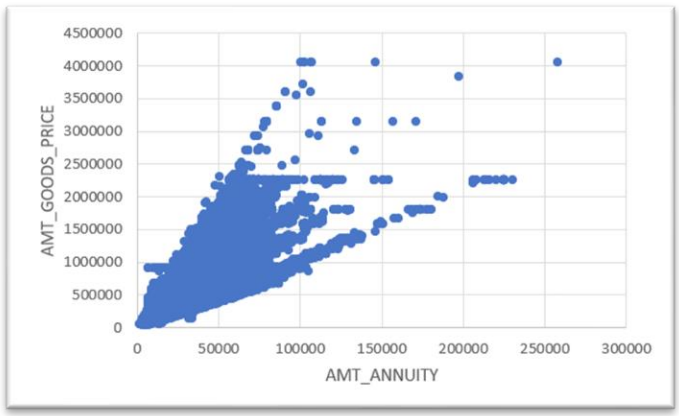
6.3 Tech-Stack Used:

The project was performed using Excel version 2021 Excel was chosen for its powerful data analysis and visualization capabilities It is also widely used in the industry and provides a familiar environment for users.

6.4 Findings:

6.4.1 Finding-1:





SK_ID_CURR 114967 is the outlier because his credit is very less as compare to his income. In second graph SK_ID_CURR 120926 is the outlier with respect to amount goods price and amount annuity. outliers can have a disproportionate impact on statistical measures such as the mean and standard deviation. This can lead to incorrect interpretations of the data and inaccurate conclusions so we should remove the outliers.

6.4.2 Finding-2:

Row Labels	Count of NAME_INCOME_TYPE	Ratio
Working	158774	31754.8
Commercial associate	71617	14323.4
Pensioner	55362	11072.4
State servant	21703	4340.6
Unemployed	22	4.4
Student	18	3.6
Businessman	10	2.0
Maternity leave	5	1.0
Grand Total	**307511**	

Row Labels	Count of HOUSETYPE_MODE	Ratio
block of flats	150503	124.2
specific housing	1499	1.2
terraced house	1212	1.0
Grand Total	**153214**	

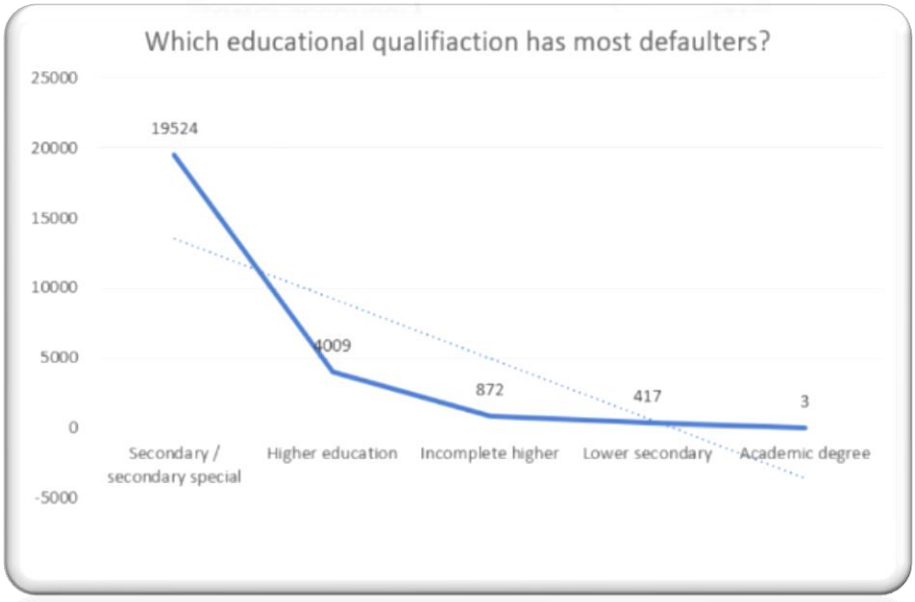
Row Labels	Count of NAME_EDUCATION_TYPE	Ratio
Secondary / secondary special	218391	1332
Higher education	74863	456
Incomplete higher	10277	63
Lower secondary	3816	23
Academic degree	164	1
Grand Total	**307511**	

Row Labels	Count of CODE_GENDER
F	202448
M	105059
XNA	4
Grand Total	**307511**

Above figures is used to find the data imbalance. I would suggest that we should remove the or use the specific data for the specific data. For example, to know the behavior we should analyze the male and the female differently so that female data isn't affect the analysis because females are 2 times more than the male.

6.4.3 Finding-3:

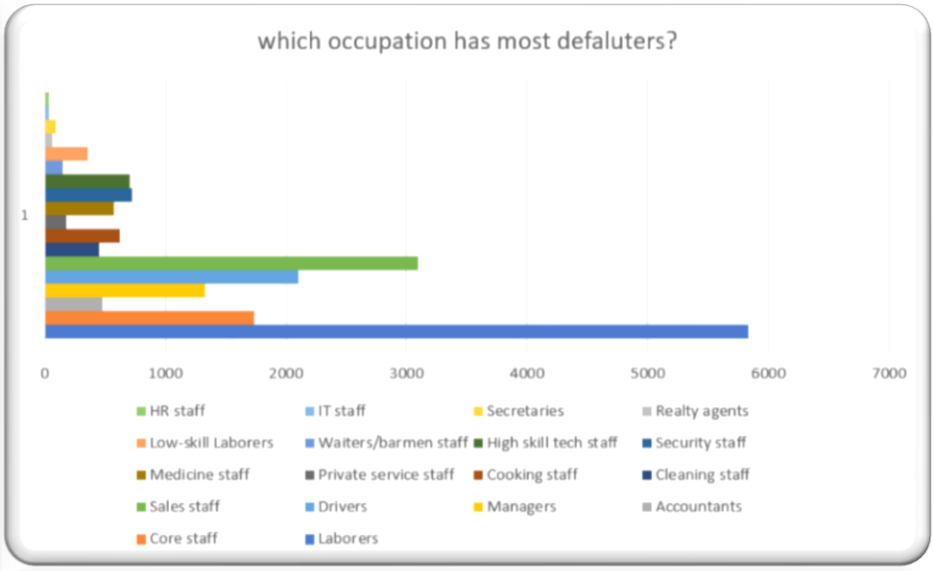
Educational Qualifiaction	1
Secondary / secondary special	19524
Higher education	4009
Incomplete higher	872
Lower secondary	417
Academic degree	3



On the basis of the data given we can say that applicants with 'secondary education' have the most defaulters. I would suggest the bank to reduce the loan approval for the applicants who has 'secondary education'.

6.4.4 Finding-4:

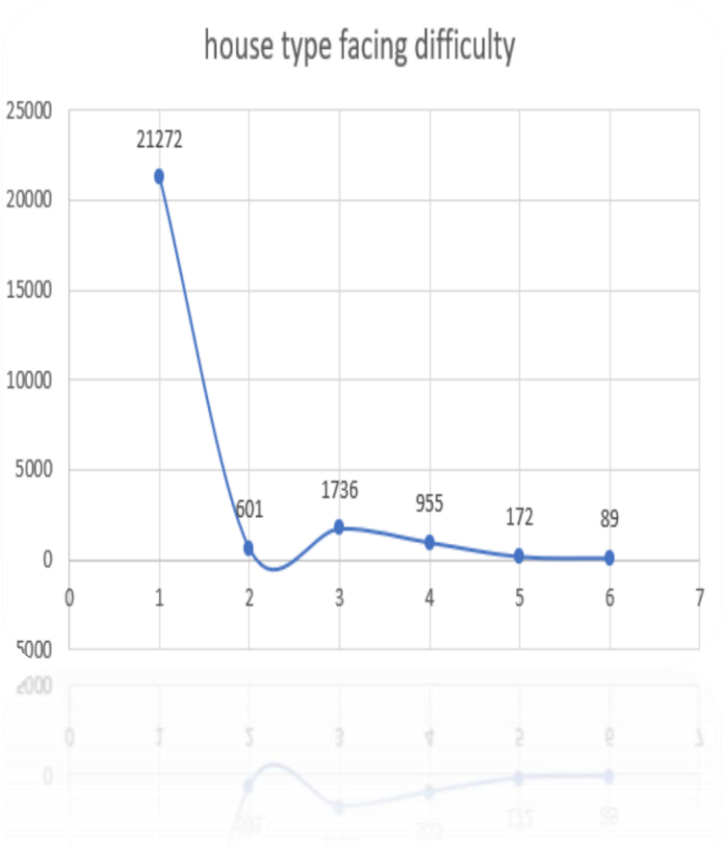
occupation types	1
Laborers	5838
Core staff	1738
Accountants	474
Managers	1328
Drivers	2107
Sales staff	3092
Cleaning staff	447
Cooking staff	621
Private service staff	175
Medicine staff	572
Security staff	722
High skill tech staff	701
Waiters/barmen staff	152
Low-skill Laborers	359
Realty agents	59
Secretaries	92
IT staff	34
HR staff	36



On basis of the data available we can say that occupation 'Laborers' has the most defaulters followed by the 'sales staff' and then drivers. 'IT staff' and 'HR staff' have the least number of defaulters. I would suggest that bank should give more loans to the applicants works into 'IT staff' and 'HR staff'.

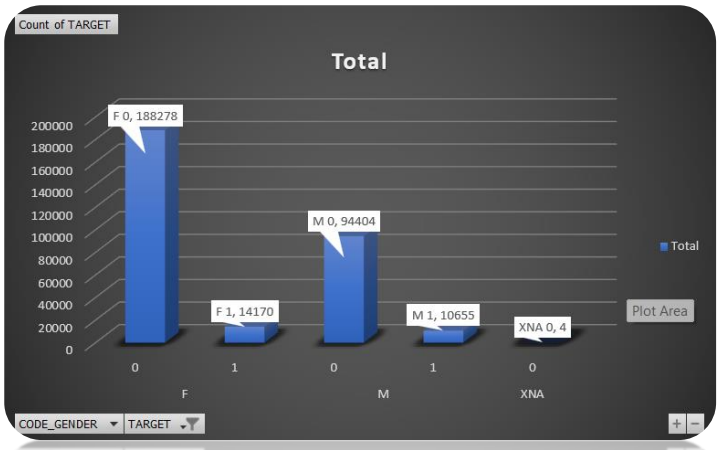
6.4.5 Finding-5:

ıouse type	1
ıouse / apartment	21272
?ented apartment	601
Vith parents	1736
Vunicipal apartment	955
)ffice apartment	172
:o-op apartment	89



On the basis of the given data we can say that house / apartment house type has the most defaulters. And Co-Op Apartments has least defaulters, I would suggest we should give more loan to the residents of the Co-op Apartment.

6.4.6 Finding-6:



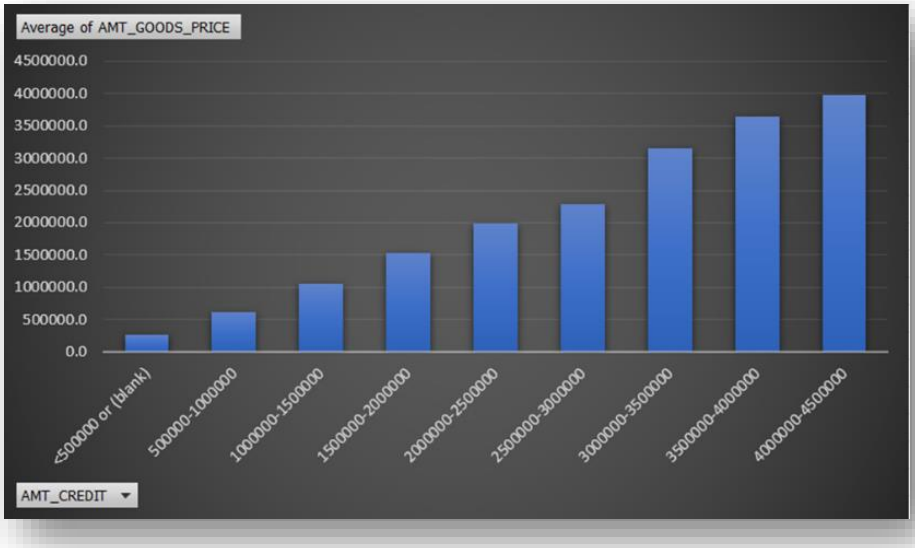
According to the above graph it is shown that although Females are more difficulties with payment but percentage wise it is male who face difficulties in payments. I would suggest the bank to give more loan to the females they default lesser as compare to the males.

6.4.7 Finding-7:

In above bar chart it can be seen that client with payment difficulties has lesser number of days employed. I would suggest that Average days for

number on employee should be more while deciding whether we should give loan or not.

6.4.8 Finding-8:



The adjacent bar graph show that Credit amount is increases with the increase of Good Price.

6.5 Data Analysis:

The mean income for the income is around 160000. There is something wrong in the data the mean of the days_employeed in not acceptable. 8 percent people have difficulties in payments. It shows that Amt_Goods_Price and Amt_Credit are the most correlated variable followed by Amt_Annuity & Amt_Goods_price and Amt_Credit & Amt_Annuity. Although Females are more difficulties with payment but percentage wise it is male who face difficulties in payments. Here are more females than males. The ratio of male and female in around 1:2. The average of the income for male is more as compared to female. Academic degree holder applies for the loan most. Academic degree holder asks for more credit as compare to other educational background. Male ask for greater money for loan as compare to the female.

6.6 Conclusion:

In conclusion, I would like to tell that after doing a thorough analysis we were able to derive the insights from the data and was able to plot various graphs using that data. The data that once looked useless became useful and helped to find out applicants who have difficulties in loan repayment. Analyzing the data proved helpful in finding various issues among the loan applicants.

7. XYZ ADS AIRING REPORT ANALYSIS

7.1 Description:

The project is based on a dataset of TV Ad Airings for some brands in the Automobile category. The aim of the project is to analyze the data and answer the given questions. The dataset includes various features like the network through which Ads are airing, the type of network like Cable/Broadcast, show name on which Ads got aired, Dayparts, Time zone, the time & date at which Ads got aired, Pod Position, duration for which Ads aired on screen, equivalent sales &, total amount spent on Ads aired. We will be analyzing this data to understand the different brands' share in TV airings, the change in brand share from Q1 to Q4 in 2021, conducting competitive analysis for the brands, and suggesting advertisement strategies for the brands.

7.2 Problem:

Advertising is a way of marketing your business in order to increase sales or make your audience aware of your products or services. Until a customer deals with you directly and actually buys your products or services, your advertising may help to form their first impressions of your business. Target audience for businesses could be local, regional, national or international or a mixture. So, they use different ways for advertisement. Some of the types of advertisement are: Internet/online directories, Trade and technical press, Radio, Cinema, Outdoor advertising, National papers, magazines and TV. Advertising business is very competitive as a lot of players bid a lot of money in a single segment of business to target the same audience. Here comes the analytical skills of the company to target those audiences from those types of media platforms where they convert them to their customers at a low cost.

7.3 Approach:

The approach to this project is to first clean the data and check for missing values. I then explored the dataset to understand the variables and their distribution. After that, I answered the given questions using different statistical techniques and visualizations. For example, I used descriptive statistics to analyze the variables and their relationship with each other. We will also use visualization techniques like bar graphs, pie charts, scatter plots, and heat maps to present our findings.

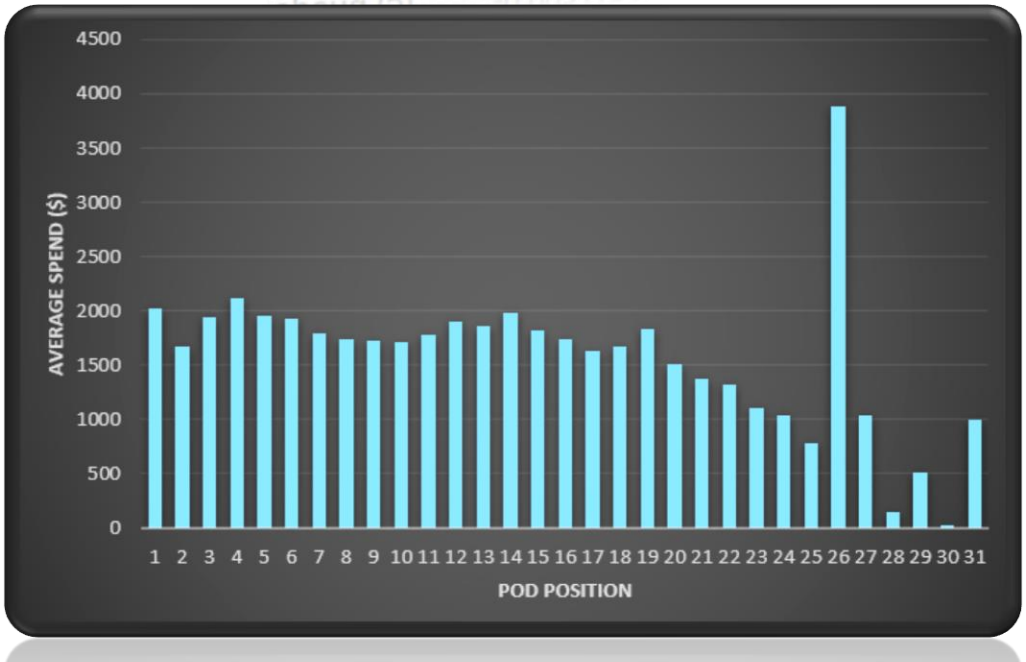
7.4 Tech-Stack Used:

The analysis was performed using Microsoft Excel. We used different Excel functions, such as COUNT, COUNTIFS, SUM, UNIQUE and AVERAGE to perform the analysis. We also used PIVOT TABLE, SLICER in Pivot Table, CHARTS and GRAPHS to visualize the data.

7.5 Findings:

7.5.1 Finding-1:

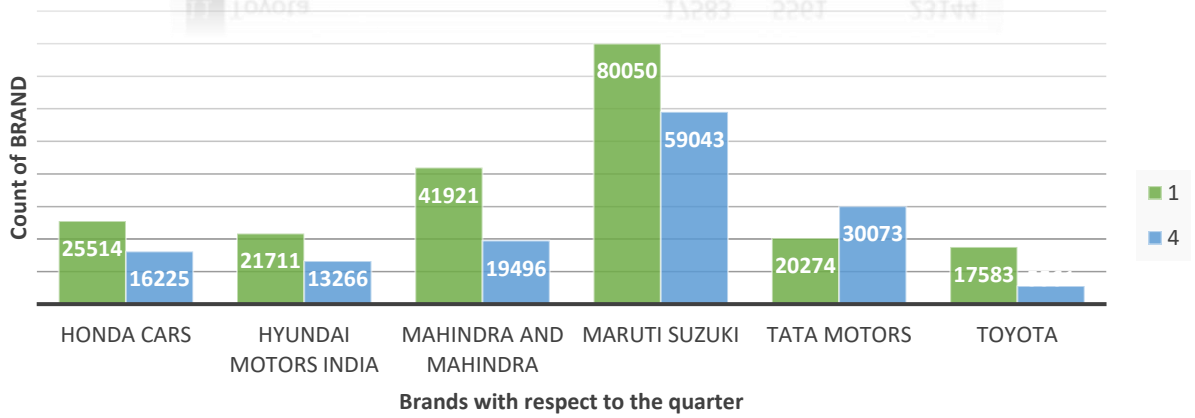
	Pod Position
Spend ($)	-0.0057157



with the help of adjacent bar graph we can say that there isn't any correlation between Average Spend and Pod Position. Although there is some trend in later parts, it seems decreasing but the value on 26 pod position is very high.

7.5.2 Finding-2:

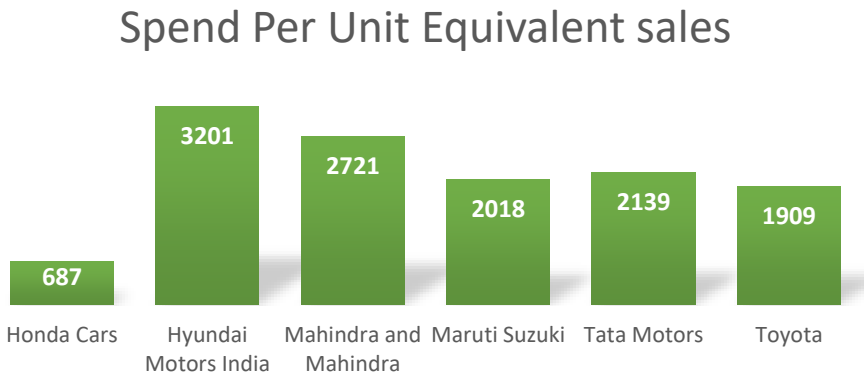
Count of Brand	Column Labels		
Row Labels	1	4	Grand Total
Honda Cars	25514	16225	41739
Hyundai Motors India	21711	13266	34977
Mahindra and Mahindra	41921	19496	61417
Maruti Suzuki	80050	59043	139093
Tata Motors	20274	30073	50347
Toyota	17583	5561	23144
Grand Total	**207053**	**143664**	**350717**



Based on available data, It seems that Maruti Suzuki is the most showed brand in TV airings followed by Mahindra & Mahindra and Honda. The least is Toyota. I would suggest that Toyota, Hyundai and Tata should increase their TV airing numbers. The share of various brands in TV airings changed from Q1 to Q4 in 2021. Tata Motors increased its share, while brands like Maruti, Honda and Toyota decreased their share.

7.5.3 Finding-3:

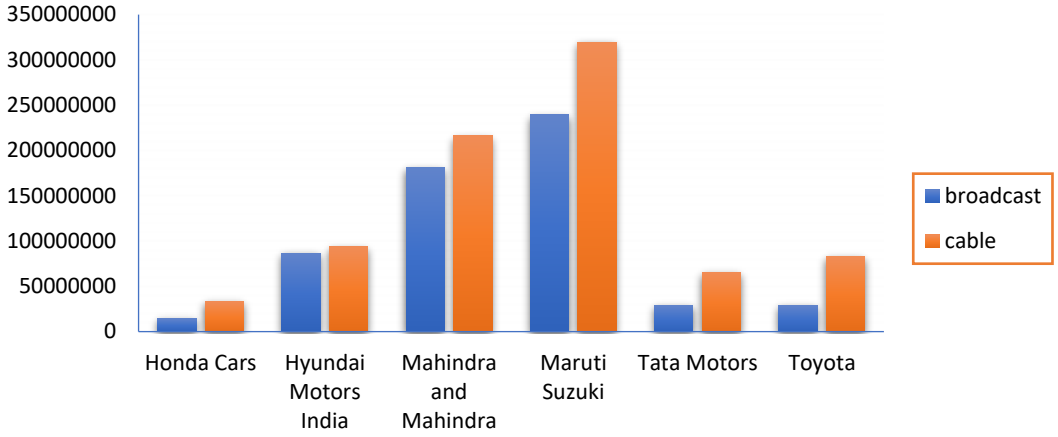
Row Labels	Sum of Spend ($)	Sum of EQ Units	Spend ($) on a unit of EQ sales
Honda Cars	48258340	70260.05	687
Hyundai Motors India	180808756	56481	3201
Mahindra and Mahindra	397305655	146036.18	2721
Maruti Suzuki	558646472	276874.46	2018
Tata Motors	94790227	44310.16	2139
Toyota	112653112	59016.87	1909
Grand Total	**1392462562**	**652978.72**	**2132.48**



Based on available data, It seems that brand Honda spends the least for sales of a car on average. Hyundai, Mahindra & Mahindra and Tata Motors are the worst performing ad campaign because they spends the most to selling a car. Hyundai, Mahindra & Mahindra and Tata should try different strategy and combination to make it more effective.

7.5.4 Finding-4:

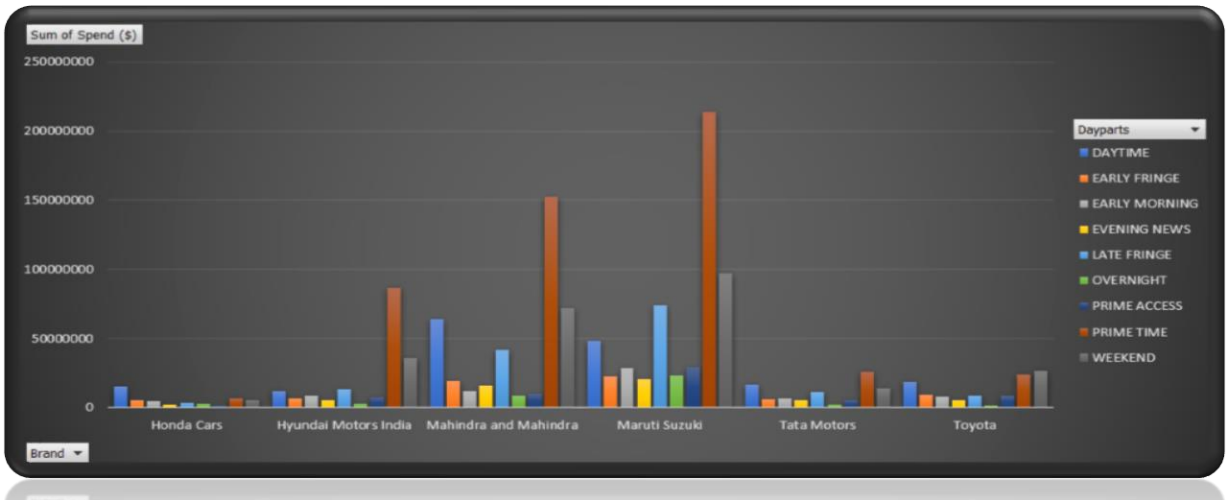
Sum of Spend ($)	Column Labels		
Row Labels	broadcast	cable	Grand Total
Honda Cars	14835303	33423037	48258340
Hyundai Motors India	86701728	94107028	180808756
Mahindra and Mahindra	181166689	216138966	397305655
Maruti Suzuki	239190273	319456199	558646472
Tata Motors	29303349	65486878	94790227
Toyota	29330783	83322329	112653112
Grand Total	**580528125**	**811934437**	**1392462562**



Based on available data, It seems that Honda runs the most successful ad campaign as it spend the less for a Equivalent sales. Honda's maximum part of campaign is in the cable network type. I would suggest to other brand such as Hyundai and Mahindra & Mahindra to increase their cable budget.

7.5.5 Finding-5:

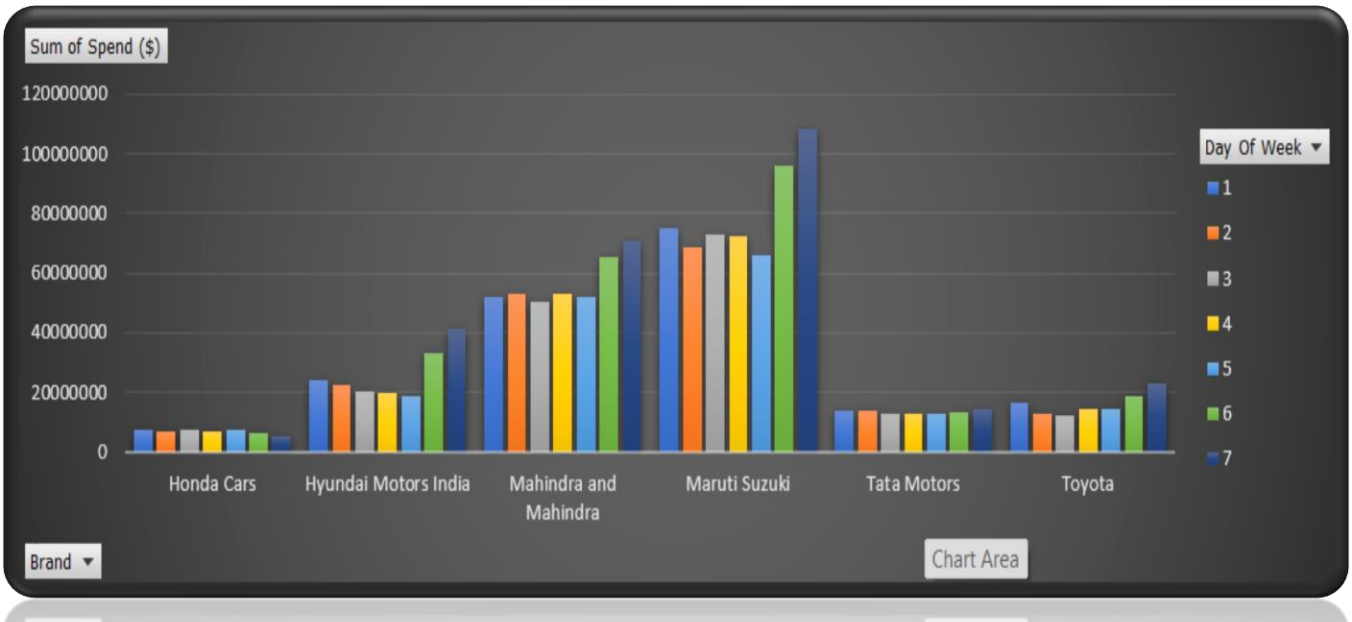
Sum of Spend ($)	Column Labels									
Row Labels	DAYTIME	EARLY FRINGE	EARLY MORNIN	EVENING NEWS	LATE FRINGE	OVERNIGH	PRIME ACC	PRIME TIMI	WEEKEND	Grand Total
Honda Cars	15106799	5763471	5190376	2105762	3421197	2820096	1352961	7002902	5494776	48258340
Hyundai Motors India	12360920	7156835	8708318	5364194	13648569	3181379	7711727	86737738	35939076	180808756
Mahindra and Mahindra	64154402	19204408	12119383	16018235	41781609	8597788	10299276	152713257	72417297	397305655
Maruti Suzuki	48678486	22745305	28920899	20776891	74069950	23614157	29021227	213609797	97209760	558646472
Tata Motors	16513542	6058611	7110565	5829272	11161135	2552537	5799904	25652452	14112209	94790227
Toyota	18560894	9744570	8294798	5409054	8863807	1716154	8979945	24146575	26937315	112653112
Grand Total	**175375043**	**70673200**	**70344339**	**55503408**	**152946267**	**42482111**	**63165040**	**509862721**	**252110433**	**1392462562**



Based on available data, It seems that brand Honda spends the least for sales of a car on average and it's maximum percentage of spend is on Daytime dayparts. I would suggest that Hyundai, Mahindra & Mahindra and Tata Motors should increase their Daytime daypart to increase their effectiveness.

7.5.6 Finding-6:

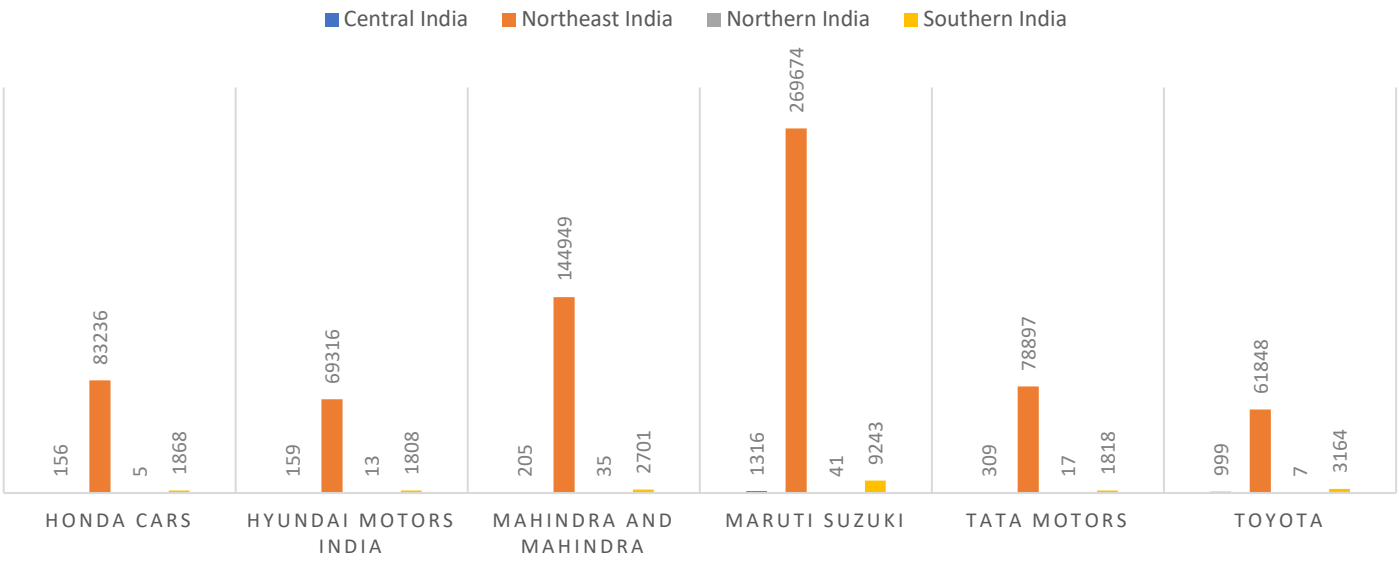
Sum of Spend ($)	Column Labels							
Row Labels	1	2	3	4	5	6	7	Grand Total
Honda Cars	7295660	7272297	7430315	7218942	7606955	6251541	5182630	48258340
Hyundai Motors In	23948566	22423170	20515865	19880982	18929031	33527438	41583704	180808756
Mahindra and Ma	51962016	53028844	50694312	53323646	51835345	65723919	70737573	397305655
Maruti Suzuki	74950307	68745503	72893206	72270606	65798163	95767180	108221507	558646472
Tata Motors	13842790	13780519	13029067	13164347	12916368	13478145	14578991	94790227
Toyota	16557975	13019177	12204227	14304986	14362454	19013743	23190550	112653112
Grand Total	**188557314**	**178269510**	**176766992**	**180163509**	**171448316**	**233761966**	**263494955**	**1392462562**



Based on available data, It seems that brand Honda spends the least for sales of a car on average and Honda's spend throughout the week is same. I would suggest that Hyundai, Mahindra & Mahindra and Maruti should reduce their weekend spends.

7.5.7 Finding-7:

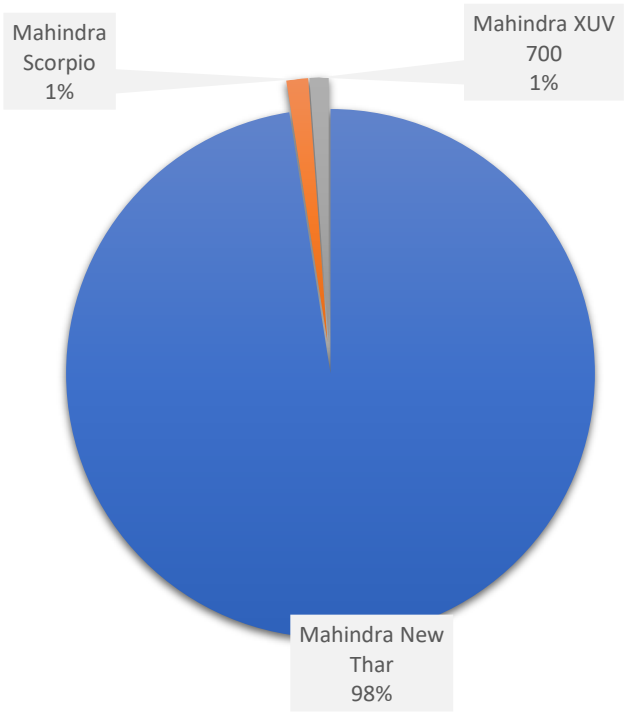
Count of Brand	Column Labels				
Row Labels	Central India	Northeast India	Northern India	Southern India	Grand Total
Honda Cars	156	83236	5	1868	85265
Hyundai Moto	159	69316	13	1808	71296
Mahindra and	205	144949	35	2701	147890
Maruti Suzuki	1316	269674	41	9243	280274
Tata Motors	309	78897	17	1818	81041
Toyota	999	61848	7	3164	66018
Grand Total	**3144**	**707920**	**118**	**20602**	**731784**



Based on the available data, Northeast India seems to be the target market for the ads. Therefore, I would suggest Mahindra and Mahindra to target this region in the digital ad campaign as well.

7.5.8 Finding-8:

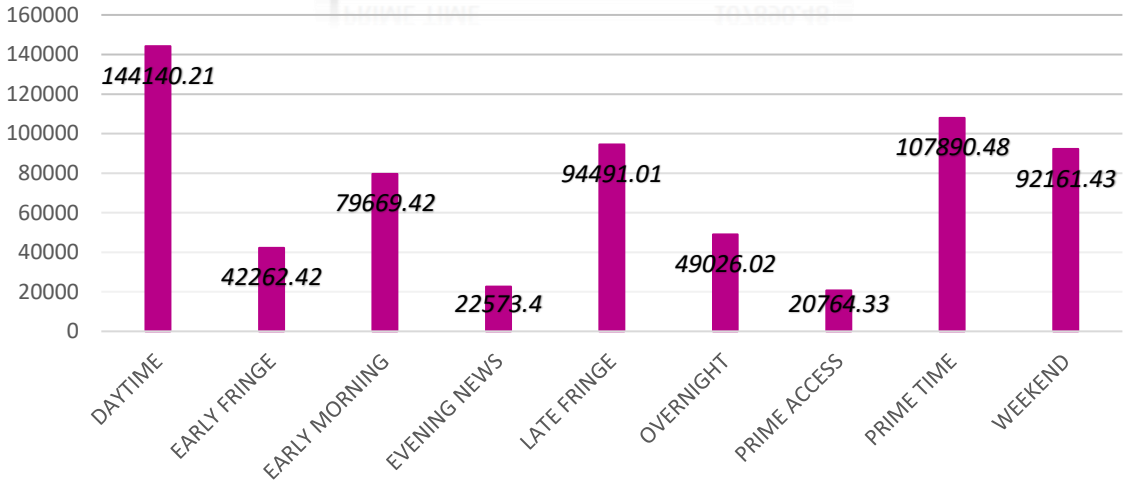
Brand	Mahindra and Mahindra
Row Labels	**Sum of EQ Units**
Mahindra New	142399
Mahindra Scor	1939
Mahindra XUV	1698
Grand Total	**146036**



Based on the available data, Mahindra New Thar seems to be the most popular car in sales. Therefore, I would suggest Mahindra and Mahindra to focus on Mahindra New Thar in the digital ad campaign as well.

Finding-9:

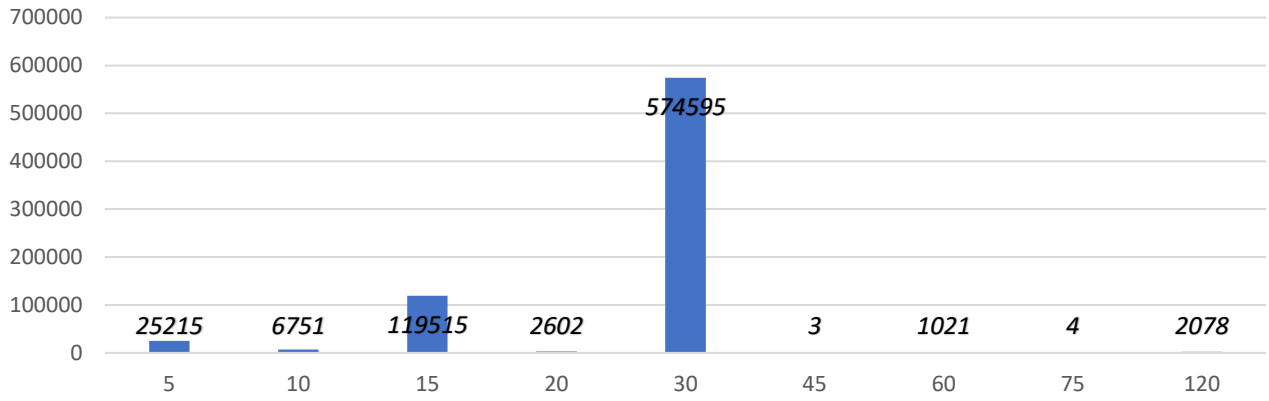
Row Labels	Sum of EQ Units
DAYTIME	144140.21
EARLY FRINGE	42262.42
EARLY MORNII	79669.42
EVENING NEW	22573.4
LATE FRINGE	94491.01
OVERNIGHT	49026.02
PRIME ACCESS	20764.33
PRIME TIME	107890.48
WEEKEND	92161.43
Grand Total	**652978.72**



Based on the available data, the most effective dayparts for the TV ads are Prime Time, Daytime, and Late Fringe. Therefore, I would suggest targeting these dayparts for the digital ad campaign as well.

7.5.9 Finding-10:

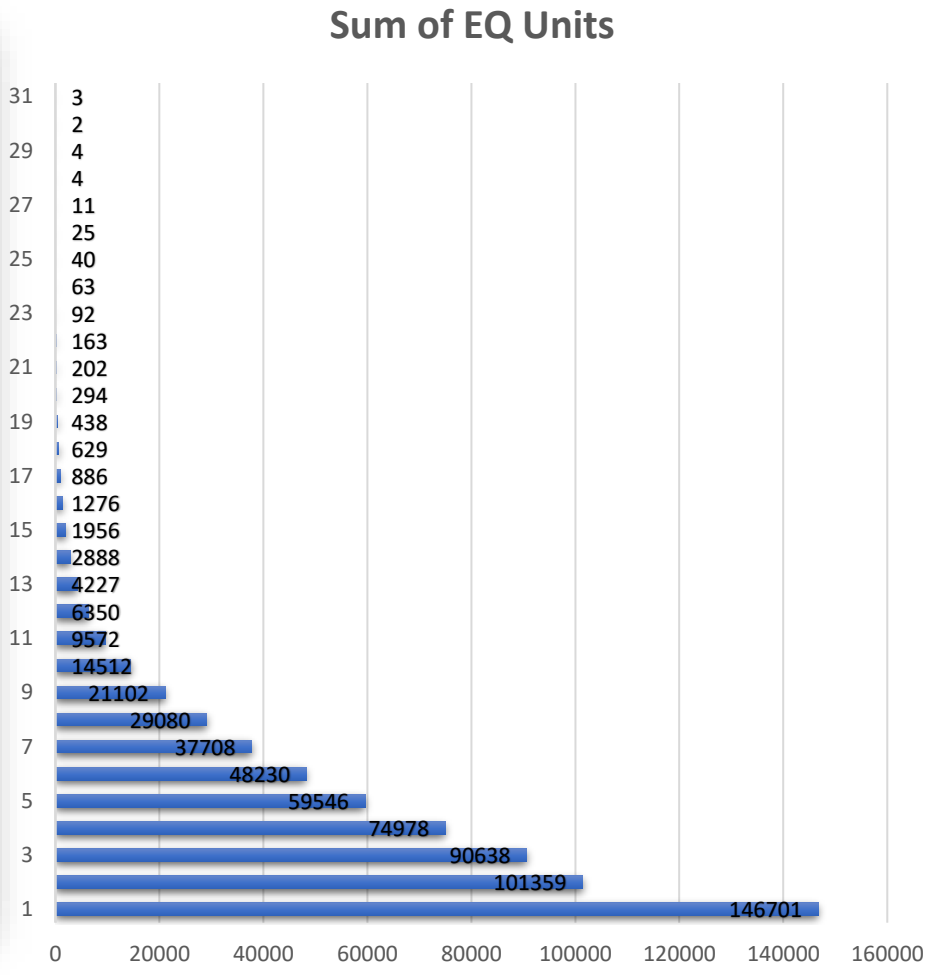
Row Labels	Sum of EQ Units
DAYTIME	144140.21
EARLY FRINGE	42262.42
EARLY MORNII	79669.42
EVENING NEW	22573.4
LATE FRINGE	94491.01
OVERNIGHT	49026.02
PRIME ACCESS	20764.33
PRIME TIME	107890.48
WEEKEND	92161.43
Grand Total	**652978.72**



Based on the available data, the most popular ad duration for the TV ads is 30 seconds. Therefore, I would suggest Mahindra and Mahindra to target this duration for the digital ad campaign as well.

7.5.10 Finding-11:

Row Labels	Sum of EQ Units
1	146701
2	101359
3	90638
4	74978
5	59546
6	48230
7	37708
8	29080
9	21102
10	14512
11	9572
12	6350
13	4227
14	2888
15	1956
16	1276
17	886
18	629
19	438
20	294
21	202
22	163
23	92
24	63
25	40
26	25
27	11
28	4
29	4
30	2
31	3
Grand Total	652979



In terms of ad placement, the most effective pod position for the TV ads is 1. Therefore, I would suggest targeting this pod position for the digital ad campaign as well.

7.6 Data Analysis:

Some of the insights we gained from the data are:

Based on available data, It seems that Maruti Suzuki is the most showed brand in TV airings followed by Mahindra & Mahindra and Honda. The least is Toyota. I would suggest that Toyota, Hyundai and Tata should increase their TV airing numbers.

The share of various brands in TV airings changed from Q1 to Q4 in 2021. Tata Motors increased its share, while brands like Maruti, Honda and Toyota decreased their share.

Based on available data, It seems that brand Honda spends the least for sales of a car on average..

Based on the available data, Mahindra New Thar seems to be the most popular car in sales. Therefore, I would suggest Mahindra and Mahindra to focus on Mahindra New Thar in the digital ad campaign as well.

In terms of ad placement, the most effective pod position for the TV ads is 1. Therefore, I would suggest targeting this pod position for the digital ad campaign as well.

7.7 Conclusion:

In conclusion, I would like to tell that after doing a thorough analysis we were able to derive the insights from the data and was able to plot various graphs using that data. The data that once looked useless became useful and helped to find out the average number of airing of the ad of a brand and suggestion based on the data. Analyzing the data proved helpful in finding various issues among the brands.

8. ABC CALL VOLUME TREND ANALYSIS

8.1 Description:

The project is about analysing the Customer Experience (CX) Inbound calling team dataset for 23 days. The dataset includes various parameters such as Agent_Name, Agent_ID, Queue_Time, Time, Time_Bucket, Duration, Call_Seconds, and Call status. The goal of the project is to analyze the data and find insights that can help improve the customer experience and optimize the performance of the inbound calling team.

8.2 Problem:

Inbound customer support is defined as the call centre which is responsible for handling inbound calls of customers. Inbound calls are the incoming voice calls of the existing customers or prospective customers for your business which are attended by customer care representatives. Inbound customer service is the methodology of attracting, engaging, and delighting your customers to turn them into your business' loyal advocates. By solving your customers' problems and helping them achieve success using your product or service, you can delight your customers and turn them into a growth engine for your business.

8.3 Approach:

My approach towards the project was to first understand the dataset and identify the key metrics that can provide insights into the performance of the inbound calling team. I then used various data analysis and visualization techniques to analyze the data and find patterns and trends. Finally, I summarized my findings and provided recommendations to improve the customer experience.

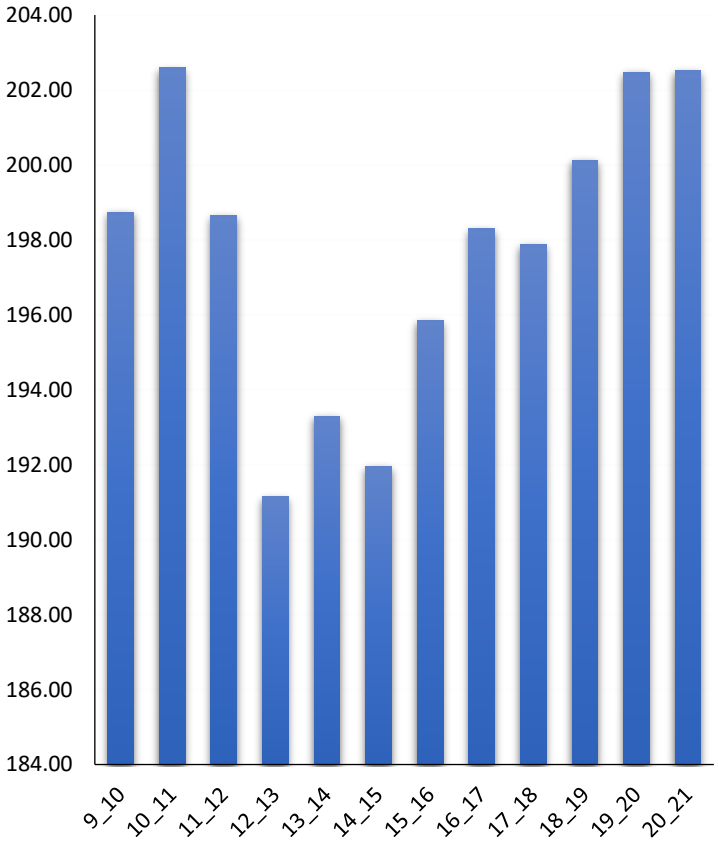
8.4 Tech-Stack Used:

The analysis was performed using Microsoft Excel. I used different Excel functions, such as COUNT, COUNTIFS, SUM, UNIQUE and AVERAGE to perform the analysis. I also used PIVOT TABLE, SLICER in Pivot Table, CHARTS and GRAPHS to visualize the data.

8.5 Findings:

8.5.1 Finding-1:

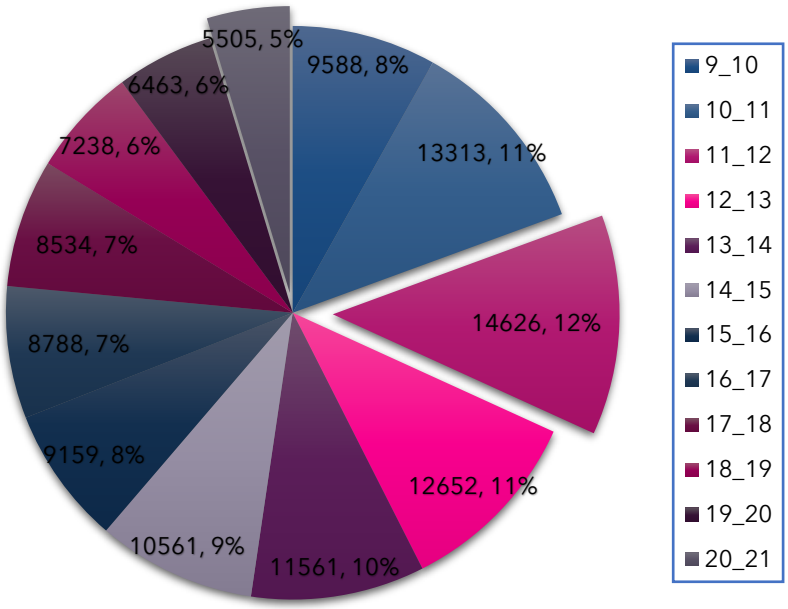
3	Row Labels	Average of Call_Seconds (s)
4	9_10	198.74
5	10_11	202.59
6	11_12	198.66
7	12_13	191.15
8	13_14	193.30
9	14_15	191.95
10	15_16	195.86
11	16_17	198.29
12	17_18	197.88
13	18_19	200.12
14	19_20	202.48
15	20_21	202.52
16	Grand Total	196.9626009



This pivot table includes the counts of the abandon calls also to filter it out "Call_Status" is used as filter. With the help of the given data we can say that calling time in morning and in the evening is high, company should train the agents to reduce the calling time. So less number of agents are required.

8.5.2 Finding-2:

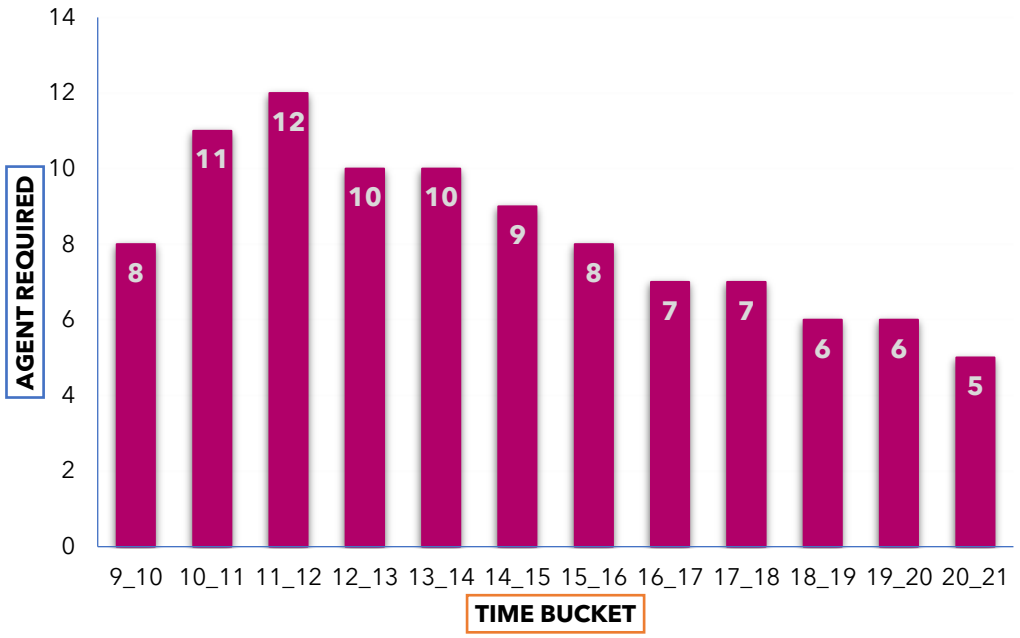
Row Labels	Count of Call_Status
9_10	9588
10_11	13313
11_12	14626
12_13	12652
13_14	11561
14_15	10561
15_16	9159
16_17	8788
17_18	8534
18_19	7238
19_20	6463
20_21	5505
Grand Total	**117988**



Based on the available data we can say that 12-13 hours is the busiest time. Between 10 am to 3 pm more than 50% of the calls are received. I would suggest that almost 50 % of the work force should be hired for the 10 am to 3 pm.

8.5.3 Finding-3:

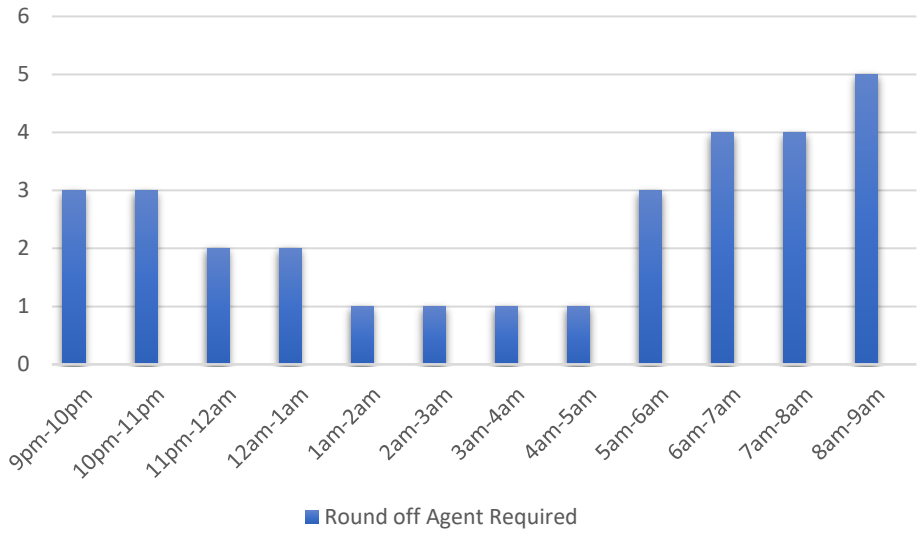
Count of Call_Status	Column Labels						
Row Labels	abandon	answered	transfer	Grand Total	Target Calls	Target Call Duration	Agent Rquired
9_10	5149	4428	11	9588	8629	2026237	8
10_11	6911	6368	34	13313	11982	2813444	11
11_12	6028	8560	38	14626	13163	3090921	12
12_13	3073	9432	147	12652	11387	2673754	10
13_14	2617	8829	115	11561	10405	2443192	10
14_15	2475	7974	112	10561	9505	2231862	9
15_16	1214	7760	185	9159	8243	1935576	8
16_17	747	7852	189	8788	7909	1857173	7
17_18	783	7601	150	8534	7681	1803495	7
18_19	933	6200	105	7238	6514	1529610	6
19_20	1848	4578	37	6463	5817	1365829	6
20_21	2625	2870	10	5505	4955	1163375	5
Grand Total	**34403**	**82452**	**1133**	**117988**	106189	24934468	93



Based on the data available we can say that at least 12 agents are required in the time bracket 11am to 12 pm. If the agents are working 5 hours a week and take 4 leaves in a month also because data given here is for 23 days so for 30 days' time span 93 agents a required.

8.5.4 Finding:4

Time Bucket Night	Ratio	Call_Counts	Target_calls	Average_call_Duration	Round off Agent Required
9pm-10pm	3	3539.64	3185.68	748034.05	3
10pm-11pm	3	3539.64	3185.68	748034.05	3
11pm-12am	2	2359.76	2123.78	498689.36	2
12am-1am	2	2359.76	2123.78	498689.36	2
1am-2am	1	1179.88	1061.89	249344.68	1
2am-3am	1	1179.88	1061.89	249344.68	1
3am-4am	1	1179.88	1061.89	249344.68	1
4am-5am	1	1179.88	1061.89	249344.68	1
5am-6am	3	3539.64	3185.68	748034.05	3
6am-7am	4	4719.52	4247.57	997378.73	4
7am-8am	4	4719.52	4247.57	997378.73	4
8am-9am	5	5899.40	5309.46	1246723.41	5
Grand Total	**30**	**35396.4**	**31856.76**	**7480340.467**	**30**



To get 30% of the total calls which is 35396 because 30% calls are received at night. Then 35396 is divided into different time bucket according to the ratio given. On the basis of the given data we can say that if we want to answer 90% of the calls we should hire 30 agents to start night calling service.

8.6 Data Analysis:

Call duration in the morning and also evening is higher as compare to the noon. I would suggest that company should try to decrease the call duration by giving agents proper training. This is how we can decrease the call abundance.

12 pm -1 pm is the busiest time. Infect between 10 am to 3 pm more than 50% of the call volume is received. For this reason, we should hire agents for the 10 am to 3 pm.

If the agents are working 5 hours a week and take 4 leaves in a month, 93 agents a required.

A total of 30 agents are required for the night calls to answer the 90% of the calls.

8.7 Conclusion:

In conclusion, I would like to tell that after doing a thorough analysis we were able to derive the insights from the data and was able to plot various graphs using that data. The data that once looked useless became useful and helped to find out requirement of the agents to answer the calls and also to number of agent required to answer night calling. Analyzing the data proved helpful in finding various issues among the calling agent.

APPENDIX

Link for Data Analytics Process Project:

https://drive.google.com/file/d/1VtfT-84OC51lD1DSb2oieSUHrKIwk7_Y/view?usp=share_link

Link for Instagram User Analytics Project:

https://drive.google.com/file/d/1sfCCvmvw4s5sRGnHpEv7AvBs1t9oojYp/view?usp=share_link

Link for Operation & Metric Analytics Project:

https://drive.google.com/file/d/1E-PcuC-InAaIAEX3_Ft59ZCo-GoBaPXb/view?usp=share_link

Link for Hiring Process Analytics Project:

https://drive.google.com/file/d/16x5NeKBkpdGgEaREGnkUks7X1i78kjxZ/view?usp=share_link

Link for IMDB Movie Analysis Project:

https://drive.google.com/file/d/1ToB5G-vDyyen4h82jooTZhoJnAnwABdN/view?usp=share_link

Link for Bank Loan Case Study Project:

https://drive.google.com/file/d/1BZCVamLw04Me6dxNygb71NXX2Seq4w3D/view?usp=share_link

Link for XYZ Ads Airing Report Project:

https://docs.google.com/presentation/d/10TP7_x_jOuTQpzmIvsuON4pV5Is7lRPt/edit?usp=share_link&ouid=111299441641572124265&rtpof=true&sd=true

Link for ABC Call Volume Trend Project:

https://docs.google.com/presentation/d/1DR8E1A5SMfw4lSVvbd03Foobs1Y_HtRJ/edit?usp=share_link&ouid=111299441641572124265&rtpof=true&sd=true