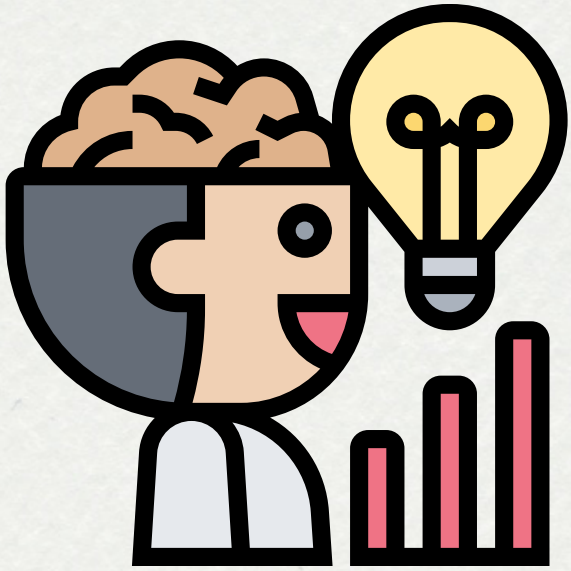


Portfolio

Aditya Palande

adityap.works@gmail.com

Professional Background

I am Aditya Palande, a third-year Information Technology engineering student at Thakur College of Engineering and Technology in Mumbai, with an expected graduation year of 2025 and an impressive cumulative grade point average of 9.5. I have consistently excelled academically, achieving outstanding marks of 95.20% in high school and 95% in intermediate studies.

With a strong passion for data analytics, I am proficient in utilizing tools like Excel for data manipulation and visualization, along with a keen aptitude for Python programming. My enthusiasm extends to competitive programming, where I showcase my problem-solving abilities and logical reasoning. I am deeply motivated to pursue a career in Artificial Intelligence, Machine Learning, and Data Science, aspiring to apply my technical skills and academic background to contribute to innovative projects in these fields. With a commitment to continuous learning and staying abreast of the latest advancements,

I aim to secure opportunities in esteemed organizations where I can leverage my expertise to solve real-world problems and drive business insights. Committed to lifelong learning and professional development, I am poised to excel in the dynamic realm of technology and data analytics.

Table of contents

Module 1: Data Analytics Process ---------- 5

- Project Description
- Report of the project

Module 2: Instagram User Analytics ---------- 6

- Project Description
- Report of the project

Module 3: Operation and Metric Analysis ---------- 13

- Project Description
- Report of the project

Module 4: Hiring Process Analytics ---------- 21

- Project Description
- Report of the project

Module 5: IMDB Movie Analysis ---------- 26

- Project Description
- Report of the project

Table of contents

Module 6: Bank Loan Case Study ---------- 33

- Project Description
- Report of the project

Module 7: Imapct of Car Features on Price and Profitability ---------- 40

- Project Description
- Report of the project

Module 8: ABC Call Volume Trend Analysis ------ 49

- Project Description
- Report of the project

Appendix ---------- 54

Module 1: Data Analytics Process

Project Description:

The process of data analytics includes planning, preparing, processing, analysing, sharing and implementing. We use data analytics in everyday life, The task was to give an example in accordance to the process.

Report:

Real life Example- Startup (Footwear Brand) Example

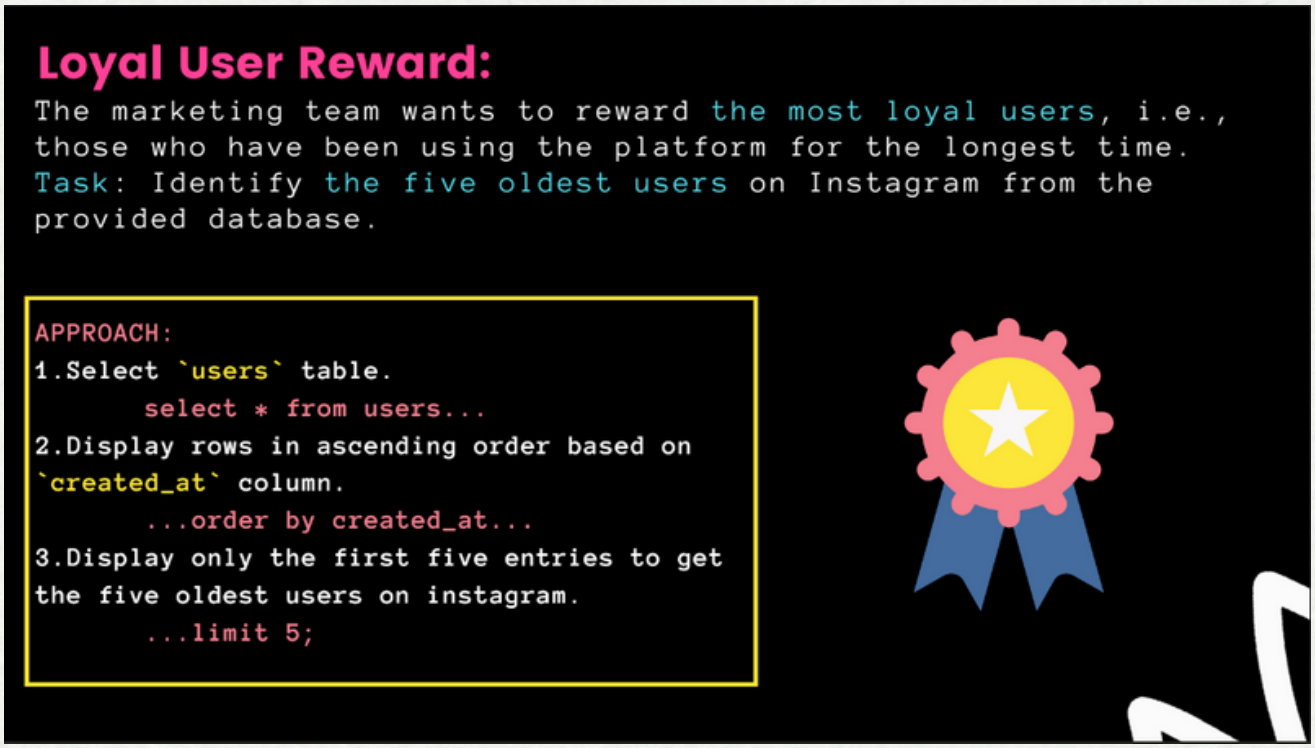
- **Plan**- We first decide what we must start our business in. For example, footwear
- **Prepare**- Next we must take a loan or look for investors willing to invest their money in the startup.
- **Process**- Contact manufacturers or partner with them (Sourcing of the material/products).
- **Analyse**- Target the right customers by analysing customer data. For example, if we are targeting the youth, we must style/price the footwear accordingly. Also, building stores/kiosk at the right location where the targeted customers are most active is important.
- **Share**- Communicate the idea to the investors/founder/cofounder/manufacturer so that execution can be done
- **Act**- Then we finally execute it!

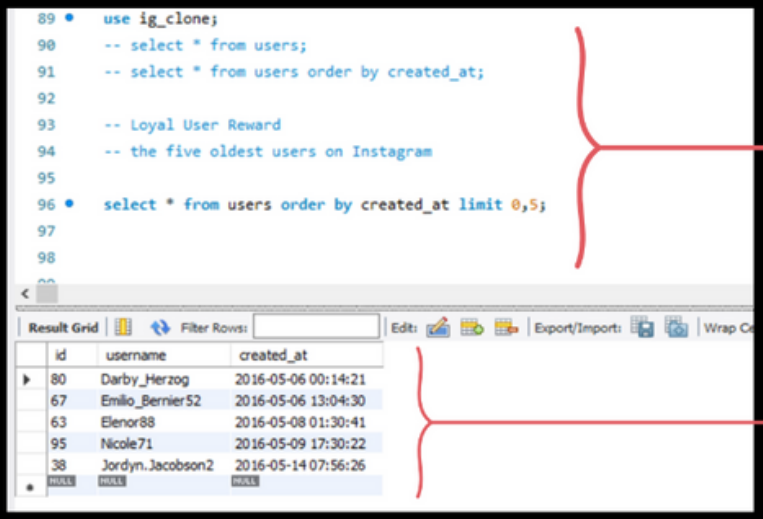
Module 2: Instagram User Analytics

Project Description:

This was an database related project. For the matter, I used SQL a relational database to write complex queries to get deeper insights into the dataset given. MySql workbench was used to interact with the batabase and write queries. The project had two parts : Marketing analysis and Investor metrics.

Report:

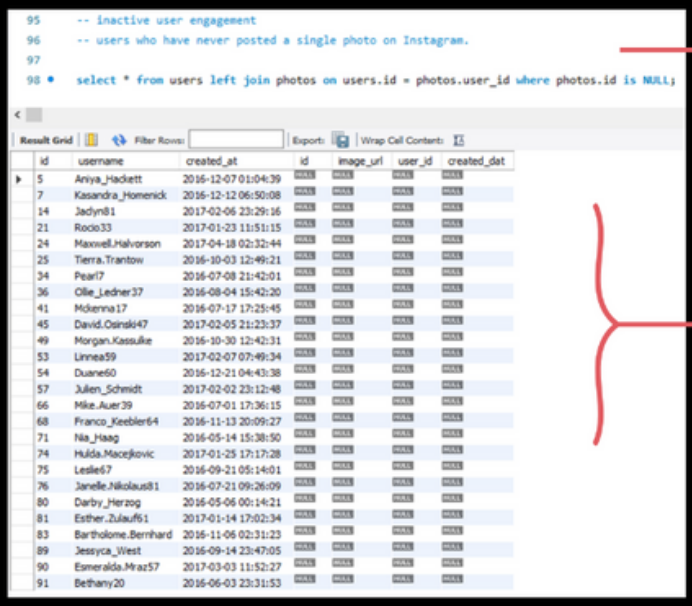




Inactive User Engagement:

The team wants to encourage inactive users to start posting by sending them promotional emails.
Task: Identify users who have never posted a single photo on Instagram.

APPROACH:
1.Select `users` table.
select * from users...
2.left join users table and photos table on users.id = photos.user_id
...left join photos on users.id = photos.user_id...
3.Use where clause to filter on those rows that are NULL.
...where photos.id is NULL;



Contest Winner Declaration:

The team has organized a contest where the user with the most likes on a single photo wins.
Task: Determine the winner of the contest and provide their details to the team.

APPROACH:

1.Select users.id, users.username, photos.id as photo_id,and count(*) as likes from `users` table.
2.inner join users, photos, likes tables on photos.id = likes.photo_id and photos.user_id = users.id
3.Group by photo_id and arrange the rows in descending order based on likes and finally select the first entry using limit.

Hashtag Research:

A partner brand wants to know the most popular hashtags to use in their posts to reach the most people.
Task: Identify and suggest the top five most commonly used hashtags on the platform.

APPROACH:
1.Select tag_name and count(*) as `used` from `tags` table.
select tag.tag_name, count(*) as used from tags...
2.inner join tags table and photo_tags table on photo_tags.tag_id = tags.id
...inner join photo_tags on photo_tags.tag_id = tags.id...
3.Group by tag_name, arrange entries in descending order based on `used` and finally select the first five entries using limit.
...group by tag_name order by used limit 5;

```
117     -- 5 most popular hashtags
118 •   select * from tags;
119 •   select
120         tags.tag_name, count(*) as used
121     from photo_tags inner join tags on photo_tags.tag_id = tags.id
122     group by tag_name
123     order by used desc
124     limit 5;
```

Query

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

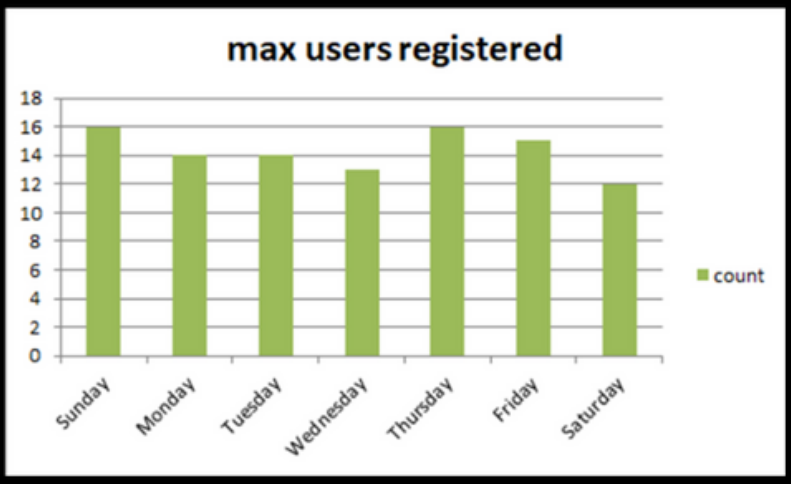
tag_name	used
smile	59
beach	42
party	39
fun	38
concert	24

Top five hashtags

Ad Campaign Launch:

The team wants to know the best day of the week to launch ads.
Task: Determine the day of the week when most users register on Instagram. Provide insights on when to schedule an ad campaign.

APPROACH:
1.Select dayname(created_at) as `day` and count(*) from `users` table.
select dayname(created_at) as `day`, count(*) as count from users...
2.Group by `day`
...group by `day`...
3.Arrange the rows in descending order of the count so that the desired day can be retrieved
...order by count desc;



The above graph depicts the best days to launch ad campaign.

User Engagement:

Investors want to know if users are still active and posting on Instagram or if they are making fewer posts.
Task: Calculate the average number of posts per user on Instagram. Also, provide the total number of photos on Instagram divided by the total number of users.

Query:

```
select user_id, count(*) from photos group by user_id;
```

```
-- average number of posts per user
select count(*) as total_users from users;
select count(*) as total_posts from photos;

select (select count(*) from photos) / (select count(*) from users) as average_posts_per_user;
```

average_posts_per_user
2.5700

Result

Bots & Fake Accounts:

Investors want to know if the platform is crowded with fake and dummy accounts.
Task: Identify users (potential bots) who have liked every single photo on the site, as this is not typically possible for a normal user.

APPROACH:
1.Select users.id, users.username, count(*) as total_liked from `users` table.
2.inner join users table and likes table on users.id = likes.user_id
3.Group by likes.user_id and using having keyword filter only those rows whose `total_liked` is same as that of total number of photos on instagram.

```
137    -- potential bots
138
139 •  select users.id, users.username, count(*) as total_liked from users inner join likes on users.id = likes.user_id
140    group by likes.user_id;
141
142 •  select
143        users.id, users.username, count(*) as total_liked
144    from users inner join likes on users.id = likes.user_id
145    group by likes.user_id
146    having total_liked = (select count(*) from photos);
```

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

id	username	total_liked
5	Aniya_Hackett	257
14	Jadyn81	257
21	Rocio33	257
24	Maxwell.Halvorson	257
36	Ollie_Ledner37	257
41	Mckenna17	257
54	Duane60	257
57	Julien_Schmidt	257
66	Mike.Auer39	257
71	Nia_Haag	257
75	Leslie67	257
76	Janelle.Nikolaus81	257
91	Bethany20	257

Query

Potential Bots

Module 3: Operation and Metric analysis

Project Description:

Two case studies were analysed and drawn insights from. In this project, the goal was to use your advanced SQL skills to analyze the data and provide valuable insights that can help improve the company's operations and understand sudden changes in key metrics.

Report:

Jobs Reviewed Over Time:

Task: Write an SQL query to calculate the number of jobs reviewed per hour for each day in November 2020.

QUERY:

```
SELECT
count(job_id)/sum(time_spent)/3600 as
no_of_jobs_reviewed_per_hour
FROM
job_data;
```

```
36 •  select count(job_id)/sum(time_spent)/3600 as no_of_jobs_reviewed_per_hour from job_data;
```

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

no_of_jobs_reviewed_per_hour
0.00000746

To find the number of jobs reviewed per hour, we make use of count() and sum() functions, divide them and further convert seconds to hours by dividing the quotient by 3600(60*60)

Throughput Analysis:

Task: Write an SQL query to calculate the 7-day rolling average of throughput. Additionally, explain whether you prefer using the daily metric or the 7-day rolling average for throughput, and why.

QUERY:

```
SELECT
`date`, count(job_id) as no_of_jobs,
sum(time_spent) as total_time_spent,
count(job_id)/sum(time_spent) as
throughput  FROM
job_data
GROUP BY `date`
ORDER BY date;
```

```sql
-- 2nd task
select
`date`, count(job_id) as no_of_jobs, sum(time_spent) as total_time_spent, count(job_id)/sum(time_spent) as throughput
from job_data group by `date` order by date;
```

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

date	no_of_jobs	total_time_spent	throughput
2020-11-25	1	45	0.0222
2020-11-26	1	56	0.0179
2020-11-27	1	104	0.0096
2020-11-28	2	33	0.0606
2020-11-29	1	20	0.0500
2020-11-30	2	40	0.0500

The throughput is calculated by divided the no_of_jobs by total_time_spent. This gives us the daily metric. For 7-day rolling average of throughput, we simply find the average of throughput for a duration of a week.

Language Share Analysis:

Task: Write an SQL query to calculate the percentage share of each language over the last 30 days.

QUERY:

```sql
SELECT
`language`, count(*) as count,
count(*)*100/(SELECT count(*) FROM
job_data) as percentage
FROM job_data
GROUP BY language;
```

```sql
-- 3nd task
-- Percentage share of each language
select `language`, count(*) as count, count(*)*100/(select count(*) from job_data) as percentage
from job_data
group by language;
```

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

language	count	percentage
English	1	12.5000
Arabic	1	12.5000
Persian	3	37.5000
Hindi	1	12.5000
French	1	12.5000
Italian	1	12.5000

The above table shows the percentage of each language. It can be seen that Persian language has the maximum percentage share

Duplicate Rows Detection:

Task: Write an SQL query to display duplicate rows FROM the job_data table.

QUERY:

```sql
SELECT
*, count(*) as count
 FROM job_data
GROUP BY date, job_id, actor_id, event,
language, time_spent, org
HAVING count > 1;
```

Weekly User Engagement:

Task: Write an SQL query to calculate the weekly user engagement.

QUERY:

```
SELECT
extract(year FROM activation_date) as
`year`, extract(week FROM
activation_date) as week_number,
count(*) as
weekly_measure_of_activeness
FROM users
 GROUP BY `year`,week_number;
```

```
125    -- 1st task
126 •  select extract(year from activation_date) as `year`, extract(week from activation_date) as week_number,
127    count(*) as weekly_measure_of_activeness
128    from users group by `year`,week_number;
```

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

year	week_number	weekly_measure_of_activeness
2013	0	23
2013	1	30
2013	2	48
2013	3	36
2013	4	30
2013	5	48
2013	6	38
2013	7	42
2013	8	34
2013	9	43
2013	10	32
2013	11	31
2013	12	33

User Growth Analysis:

Task: Write an SQL query to calculate the user growth for the product.

QUERY:

```
SELECT
extract(year FROM creation_date) as
year, extract(week FROM creation_date)
as `week`, count(state) as active_state
FROM users
GROUP BY `week`, `year`;
```

Weekly Retention Analysis:

Task: Write an SQL query to calculate the weekly retention of users based on their sign-up cohort.

OUTPUT:

user_id	no_of_user	per_week_retention
11768	1	0
11770	1	0
11775	1	0
11778	1	0
11779	1	0
11780	1	0
11785	1	0
11787	1	0
11791	1	0
11793	1	0
11795	1	0
11798	1	0
11799	1	0

Weekly Engagement Per Device:

Task: Write an SQL query to calculate the weekly engagement per device.

QUERY:

```
SELECT
extract(year FROM users.creation_date) as
year, extract(week FROM users.creation_date)
as `week`, device, count(distinct users.user_id)
as activity FROM users
INNER JOIN events ON users.user_id =
events.user_id
GROUP BY device, week, year
ORDER BY year,week,device;
```

```
141    -- 4th task
142    -- select * from events;
143 •  select extract(year from users.creation_date) as year, extract(week from users.creation_date) as `week`, device,
144    count(distinct users.user_id) as activity
145    from users inner join events on users.user_id = events.user_id
146    group by device, week, year order by year,week,device;
```

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

year	week	device	activity
2013	0	asus chromebook	1
2013	0	dell inspiron desktop	3
2013	0	hp pavilion desktop	1
2013	0	ipad air	1
2013	0	iphone 4s	1
2013	0	iphone 5	1
2013	0	iphone 5s	1
2013	0	kindle fire	1
2013	0	lenovo thinkpad	4
2013	0	macbook pro	4
2013	0	nexus 5	2
2013	0	nexus 7	3
2013	0	samsung galaxy s4	2
2013	1	acer aspire desktop	1
2013	1	acer aspire notebook	2
2013	1	amazon fire phone	1

Email Engagement Analysis

Task: Write an SQL query to calculate the email engagement metrics.

OUTPUT:

Result Grid | Filter Rows:

	opening_rate	clicking_rate
▶	33.5834	14.7899

QUERY:

```
SELECT
100*sum(CASE WHEN email_cat =
'email_opened' THEN 1 else 0
end)/sum(CASE WHEN
email_cat = 'email_sent' THEN 1 else 0
end) as opening_rate,
100*sum(CASE WHEN email_cat =
'email_clicked' THEN 1 else 0
end)/sum(CASE WHEN
email_cat = 'email_sent' THEN 1 else 0
end) as clicking_rate
FROM
(
SELECT
*,
CASE
WHEN action in
('sent_weekly_digest',
'sent_reengagement_email')
THEN 'email_sent'
WHEN action in ('email_open')
THEN 'email_opened'
WHEN action in ('email_clickthrough')
THEN 'email_clicked'
end as email_cat
FROM
job_data_analysis.email_events
) a;
```

Module 4: Hiring Process Analytics

Project Description:

As a data analyst, I was given a dataset containing records of previous hires. My job was to analyze this data and answer certain questions that can help the company improve its hiring process.

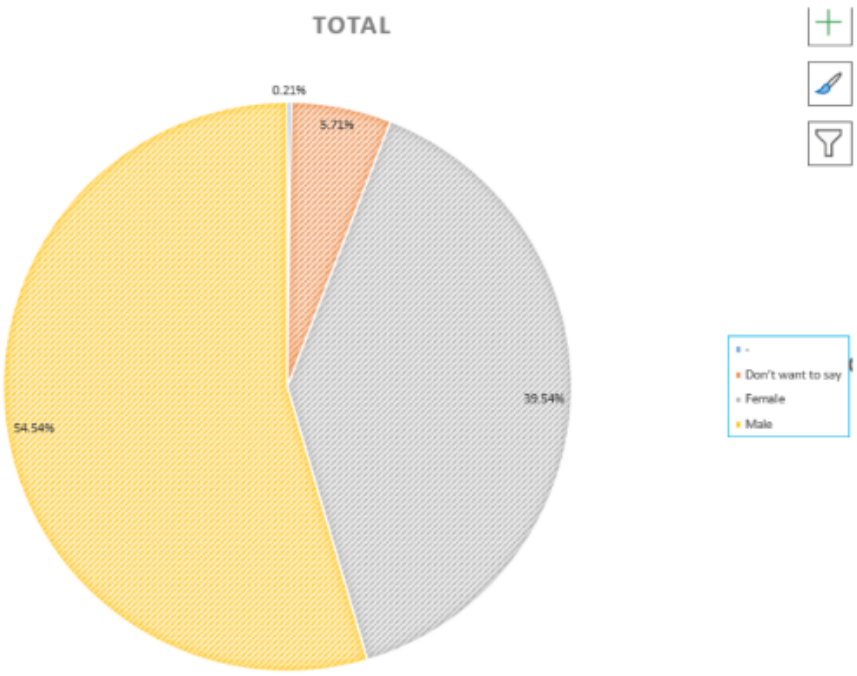
Report:

A. Hiring Analysis: The hiring process involves bringing new individuals into the organization for various roles.

Task: Determine the gender distribution of hires. How many males and females have been hired by the company?

TASK 1- Gender Distribution

Gender	Status	Count
Male	Hired	2552
Female	Hired	1850
-	Hired	10
Don't want to say	Hired	267



From the above pie chart it can be seen that, the number of male employees who got hired by the company is more than female employees and the employees who chose not to disclose their gender.

B. Salary Analysis: The average salary is calculated by adding up the salaries of a group of employees and then dividing the total by the number of employees.

Task: What is the average salary offered by this company? Use Excel functions to calculate this.

TASK 2- Average Salary
Average salary of all the employees is 50009.956302521

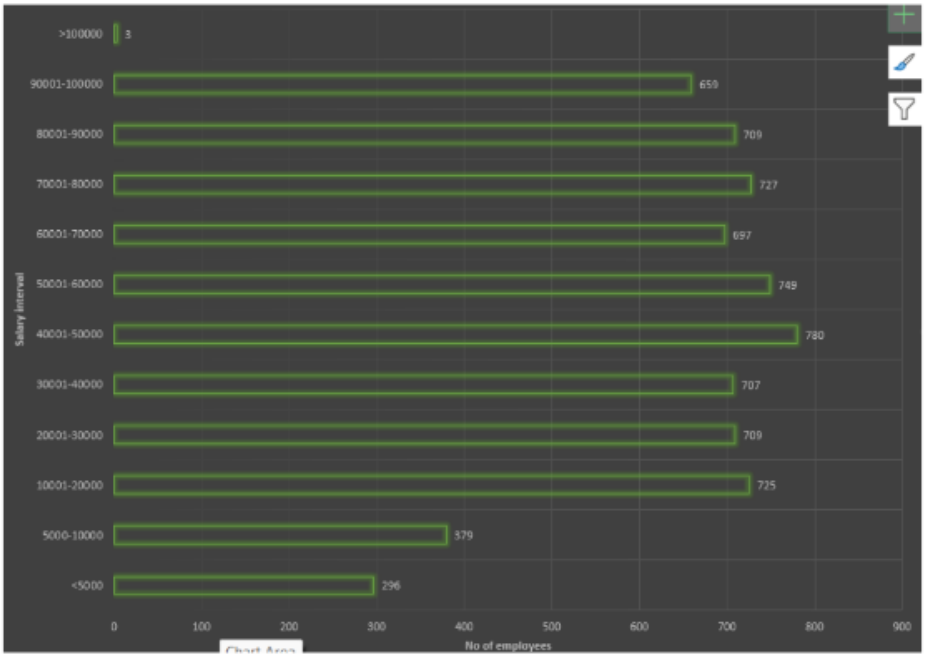
C. Salary Distribution: Class intervals represent ranges of values, in this case, salary ranges. The class interval is the difference between the upper and lower limits of a class.

Task: Create class intervals for the salaries in the company. This will help you understand the salary distribution.

TASK 3- Salary Distribution

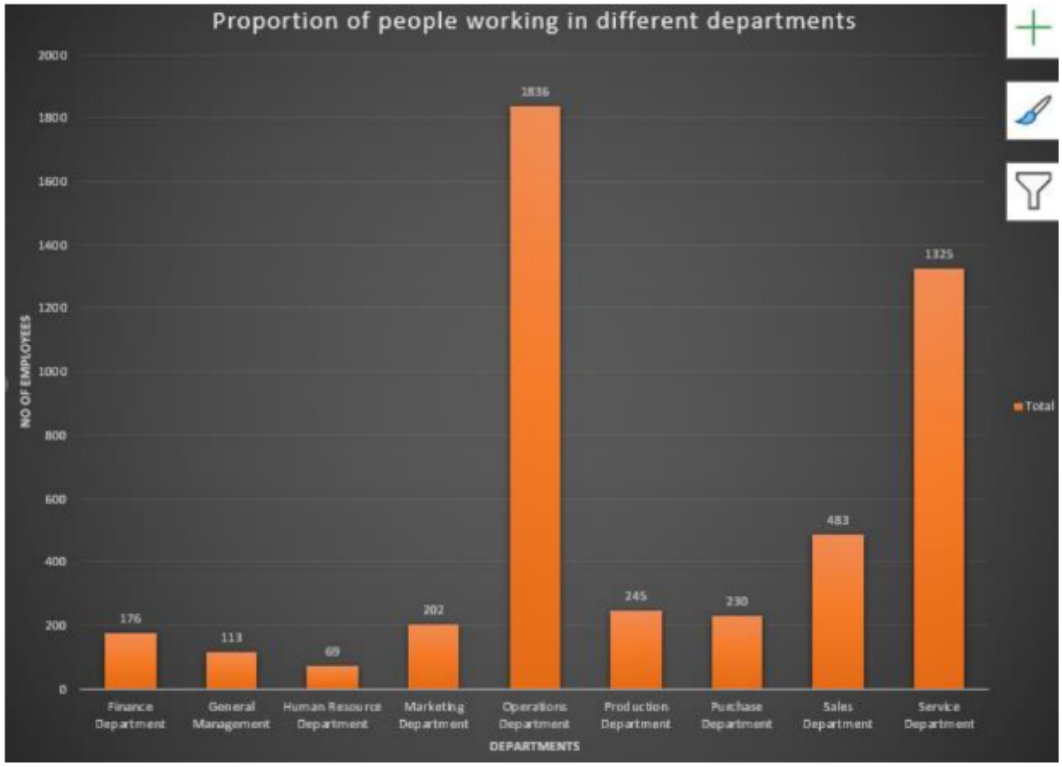
Salary interval	Count
<5000	296
5000-10000	379
10001-20000	725
20001-30000	709
30001-40000	707
40001-50000	780
50001-60000	749
60001-70000	697
70001-80000	727
80001-90000	709
90001-100000	659
>100000	3

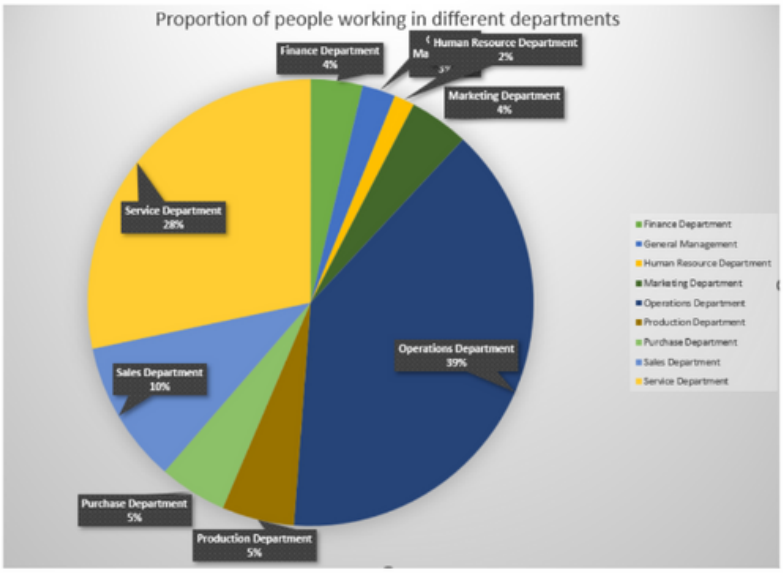
The above table depicts that maximum employees earn between 40,000 to 50,000 and only 3 employees earn above 1 lakh. Graphically, the above table is shown as below:



D. Departmental Analysis: Visualizing data through charts and plots is a crucial part of data analysis.

Task: Use a pie chart, bar graph, or any other suitable visualization to show the proportion of people working in different departments.

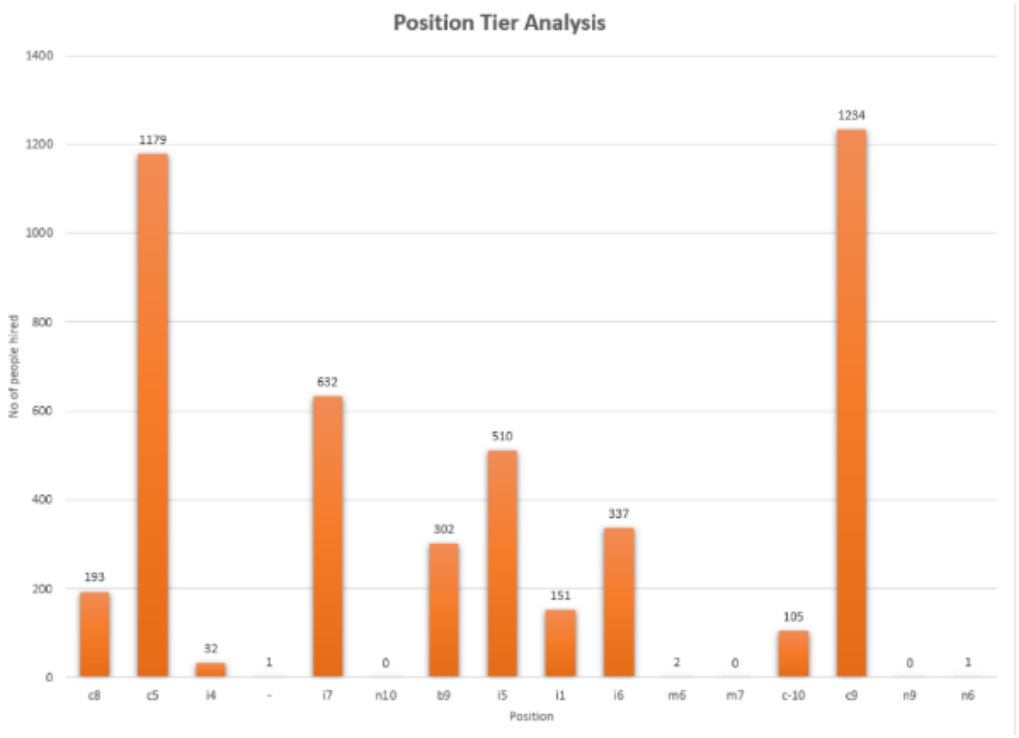




The above two charts conclude that the Operations department has the maximum number of employees; whereas the HR department has the least.

E. Position Tier Analysis: Different positions within a company often have different tiers or levels.

Task: Use a chart or graph to represent the different position tiers within the company. This will help you understand the distribution of positions across different tiers.



Conclusion:

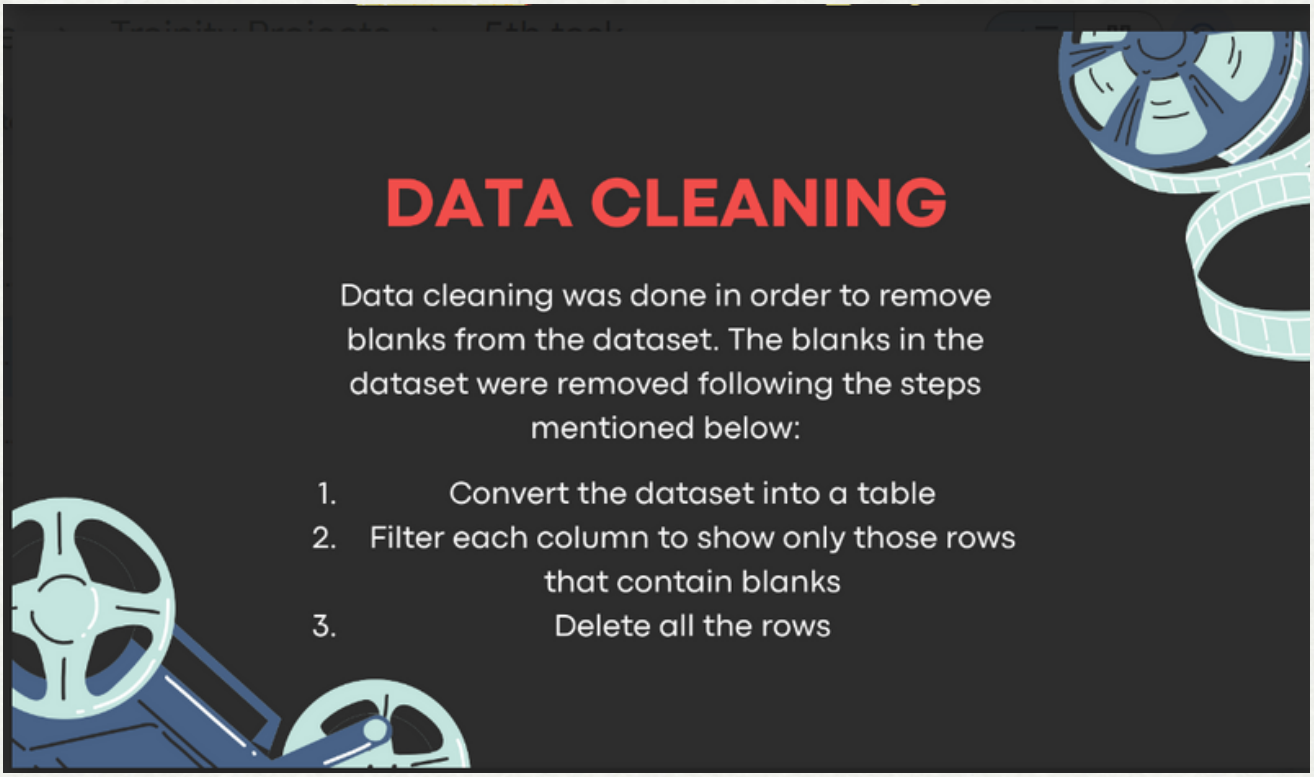
The Excel project focused on deriving valuable insights from the hiring process, shedding light on the dynamics of employee selection, salaries, positions, and departmental assignments. The findings provide a data-driven foundation for optimizing recruitment strategies and making informed decisions.

Module 5: IMDB Movie Analysis

Project Description:

The dataset provided was related to IMDB Movies.The impact of analysis done on the dataset would be significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.

Report:



MOVIE GENRE ANALYSIS

Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

Approach:

1. Manipulation of Genre Column:
used "Text to Column" (delimiter- "|")

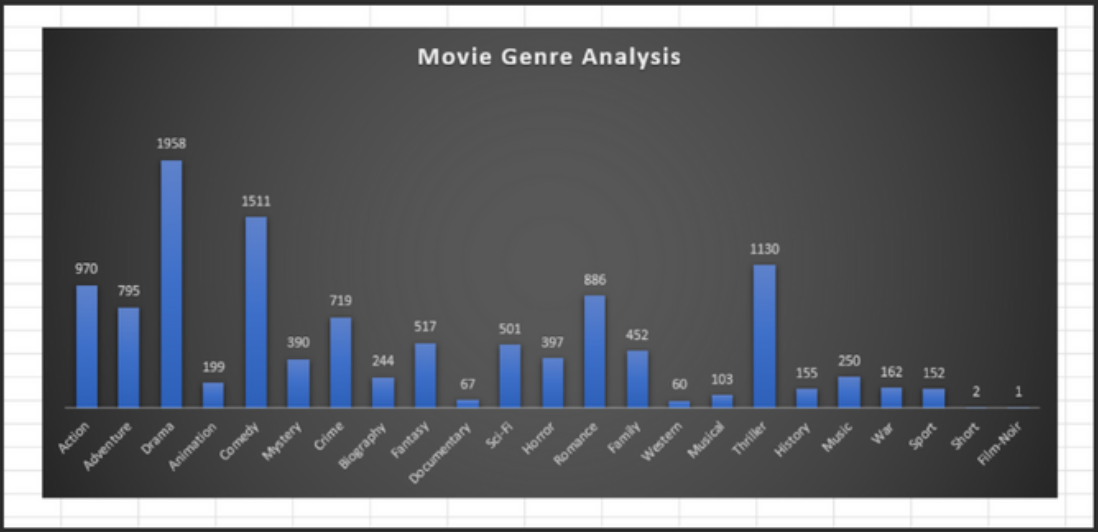
2. Creation of columns : Genre, No. of Movies, Measn, Median, Mode, Min, Max, Range, Variance and StdDev

3. Functions used : UNIQUE(), COUNTIF(), MEDIAN(), AVERAGEIF(), MODE.MULT(), MIN(), MAX(), VAR.P(), SQRT(), IF(), ISNUMBER(), SEARCH(), IFERROR()

Genre	No. of movies
Action	970
Adventure	795
Drama	1958
Animation	199
Comedy	1511
Mystery	390
Crime	719
Biography	244
Fantasy	517
Documentary	67
Sci-Fi	501
Horror	397
Romance	886
Family	452
Western	60
Musical	103
Thriller	1130
History	155
Music	250
War	162
Sport	152
Short	2
Film-Noir	1

Genre	Mean	Median	Mode	Min	Max	Range	Variance	StdDev
Action	6.290618557	6.35	6.6	2.1	9	6.9	1.076375906	1.037485376
Adventure	6.45572327	6.6	6.7	2.3	8.9	6.6	1.229335169	1.108753881
Drama	6.78299285	6.9	6.7	2.1	9.3	7.2	0.803791451	0.896544171
Animation	6.700502513	6.8	6.7	2.8	8.6	5.8	0.972411808	0.98610943
Comedy	6.184513567	6.3	6.7	1.9	8.8	6.9	1.082486841	1.040426279
Mystery	6.466410256	6.5	6.6	3.1	8.6	5.5	1.03438455	1.017046975
Crime	6.54798331	6.6	6.6	2.4	9.3	6.9	0.962746281	0.981196352
Biography	7.141803279	7.2	7	4.5	8.9	4.4	0.498580355	0.706102227
Fantasy	6.281431335	6.4	6.7	2.2	8.9	6.7	1.288011104	1.134905769
Documentary	7.011940299	7.2	6.6	1.6	8.5	6.9	1.418364892	1.190951255
Sci-Fi	6.323952096	6.4	6.7	1.9	8.8	6.9	1.340663822	1.157870382
Horror	5.927959698	6	5.9	2.3	8.6	6.3	0.994256039	0.997123883
Romance	6.431264108	6.5	6.5	2.1	8.5	6.4	0.929981923	0.964355704
Family	6.207743363	6.3	5.4	1.9	8.6	6.7	1.344431191	1.159496093
Western	6.748333333	6.75	6	4.1	8.9	4.8	0.957830556	0.978688181
Musical	6.559223301	6.7	7.1	2.1	8.5	6.4	1.289211047	1.135434299
Thriller	6.377699115	6.4	6.5	2.7	9	6.3	0.940865502	0.969982218
History	7.134193548	7.2	7.7	5.5	8.9	3.4	0.455798543	0.675128538
Music	6.4636	6.7	6.2	1.6	8.5	6.9	1.39647504	1.18172545
War	7.048148148	7.1	7.1	4.3	8.6	4.3	0.647681756	0.804786777
Sport	6.607236842	6.8	7.2	2	8.4	6.4	1.076592365	1.03758969
Short	6.8	6.8	-	6.5	7.1	0.6	0.09	0.3
Film-Noir	7.7	7.7	-	7.7	7.7	0	0	0

Table generated after performing the task



From the above chart it can be seen that movies made in "Drama" genre have the highest mean IMDB ratings followed by Comedy and Thriller. Film-Noir, Short have the least

MOVIE DURATION ANALYSIS

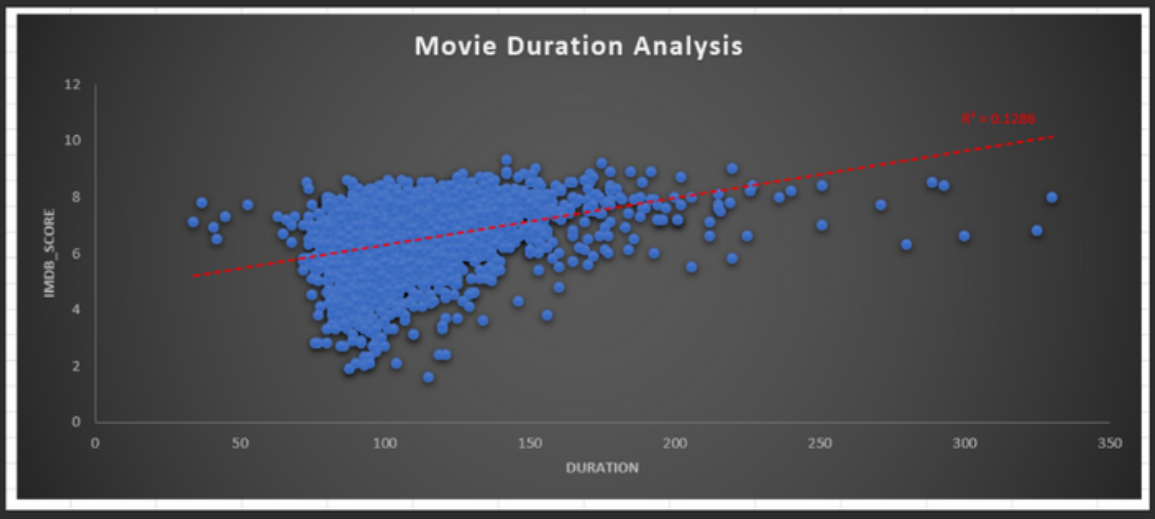
Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

Approach:

1. Duration analysis :

	Mean	Median	Mode	Min	Max	Range	Variance	StdDev
Duration	109.902	106	101	34	330	296	515.7277	22.70964

3. Functions used : MEDIAN(), AVERAGEIF(), MODE.MULT(), MIN(), MAX(), VAR.P(), SQRT(), IF(), ISNUMBER(), SEARCH(), IFERROR()



The above scatter plot chart gives us the relationship between Duration of movies and the IMDB scores. A trendline has been added whose R_squared values is around 0.13. Most movies made were around 70 to 150 mins long.

LANGUAGE ANALYSIS

Task: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

Approach:

1. Pivot table for analysis

2. Functions used : UNIQUE(), COUNTIF(), MEDIAN(), AVERAGEIF(), MODE.MULT(), MIN(), MAX(), VAR.P(), SQRT(), IF(), ISNUMBER(), SEARCH(), IFERROR()

Languages	No. of movies	Mean	Median	Mode	Min	Max	Range	Variance	StdDev
English	3706	6.424042094	6.5	6.7	1.6	9.3	7.7	1.104173732	1.050796713
Mandarin	15	7.08	7.4	7.9	5.6	7.9	2.3	0.556266667	0.745832868
Aboriginal	2	6.95	6.95	-	6.4	7.5	1.1	0.3025	0.55
Spanish	26	7.05	7.15	7.2	5.2	8.2	3	0.656346154	0.810151933
French	37	7.286486486	7.2	7.2	5.8	8.4	2.6	0.306574142	0.553691378
Filipino	1	6.7	6.7	-	6.7	6.7	0	0	0
Maya	1	7.8	7.8	-	7.8	7.8	0	0	0
Kazakh	1	6	6	-	6	6	0	0	0
Telugu	1	8.4	8.4	-	8.4	8.4	0	0	0
Cantonese	8	7.2375	7.3	7.3	6.5	7.8	1.3	0.16984375	0.412121038
Japanese	12	7.625	7.8	-	6	8.7	2.7	0.741875	0.861321659
Aramaic	1	7.1	7.1	-	7.1	7.1	0	0	0
Italian	7	7.185714286	7	-	5.3	8.9	3.6	1.144081633	1.069617517
Dutch	3	7.566666667	7.8	7.8	7.1	7.8	0.7	0.108888889	0.329983165
Dari	2	7.5	7.4	7.6, 7.9	5.6	7.9	2.3	0.510311419	0.714360846
German	13	7.692307692	7.7	7.4, 7.8, 8.3, 7.3, 7.7	6.1	8.5	2.4	0.379171598	0.615769111
Mongolian	1	7.3	7.3	-	7.3	7.3	0	0	0
Thai	3	6.633333333	6.6	-	6.2	7.1	0.9	0.135555556	0.368178701
Bosnian	1	4.3	4.3	-	4.3	4.3	0	0	0
Korean	5	7.7	7.7	-	7	8.4	1.4	0.26	0.509901951
Hungarian	1	7.1	7.1	-	7.1	7.1	0	0	0
Hindi	10	6.76	7.05	-	4.8	8	3.2	1.1124	1.05470375
Icelandic	1	6.9	6.9	-	6.9	6.9	0	0	0
Danish	3	7.9	8.1	-	7.3	8.3	1	0.186666667	0.43204938
Portuguese	5	7.76	8	-	6.1	8.7	2.6	0.7664	0.875442745
Norwegian	4	7.15	7.3	7.6	6.4	7.6	1.2	0.2475	0.497493719
Czech	1	7.4	7.4	-	7.4	7.4	0	0	0
Russian	1	6.5	6.5	-	6.5	6.5	0	0	0
None	1	8.5	8.5	-	8.5	8.5	0	0	0
Zulu	1	7.3	7.3	-	7.3	7.3	0	0	0
Hebrew	3	7.5	7.3	-	7.2	8	0.8	0.126666667	0.355902608
Dzongkha	1	7.5	7.5	-	7.5	7.5	0	0	0
Arabic	1	7.2	7.2	-	7.2	7.2	0	0	0
Vietnames	1	7.4	7.4	-	7.4	7.4	0	0	0
Indonesiar	2	7.9	7.9	-	7.6	8.2	0.6	0.09	0.3
Romanian	1	7.9	7.9	-	7.9	7.9	0	0	0
Persian	3	8.133333333	8.4	-	7.5	8.5	1	0.202222222	0.449691252
Swedish	1	7.6	7.6	-	7.6	7.6	0	0	0

The table alonside shows all the languages the movies were made in and their descriptive analysis. 3706 movies were made in English language. The movie that recieved the highest ratings was in English language. The second most popular language was seen to be French.

DIRECTOR ANALYSIS

Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

Approach:

1. Pivot table for analysis

2. Columns : Directors, Average of imdb_scores, percentile

3. Functions used : PERCENTRANK.INC()

Directors	Average of imdb_score	Percentile
Ã‰mile Gaudreault	6.7	60
Álex de la Iglesia	6.1	35
Aaron Schneider	7.1	77.2
Aaron Seltzer	2.7	0.2
Abel Ferrara	6.6	55.3
Adam Carolla	6.1	35
Adam Goldberg	5.4	15.6
Adam Marcus	4.3	4.5
Adam McKay	6.916666667	71.2
Adam Rapp	6.4	46.4
Adam Rifkin	6.8	63.9
Adam Shankman	5.9625	30.8
Adrian Lyne	6.4	46.4
Adrienne Shelly	7.1	77.2
Agnieszka Holland	6.8	63.9
Agnieszka Wojtowicz-Vosloo	5.9	27
Aki KaurismÃ¤ki	7.2	81.2
Akira Kurosawa	8.1	98.1
Akiva Goldsman	6.2	39.1
Akiva Schaffer	5.7	23
Alan Cohn	6	31
Alan J. Pakula	6.3	42.2
Alan Metter	3.3	1
Alan Parker	7.033333333	76.5

The table contains three columns viz, Directors, Average of imdb_scores, percentile. The most rated director of all comes out to be Charles Chaplin, Tony Kaye both having the average rating of 8.6

BUDGET ANALYSIS

Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

Approach:

1. Pivot table for analysis
2. Conditional Formatting : Profit margins have been formatted using color scaling for easy and quick insightful understanding.
3. Functions used : CORREL(), MAX(), INDEX(), MATCH()

Movie Title	Sum of gross	Sum of budget	net_profit
[Rec] 2Â	27024	5600000	-5572976
10 Cloverfield LaneÂ	71897215	15000000	56897215
10 Days in a MadhouseÂ	14616	12000000	-11985384
10 Things I Hate About YouÂ	38176108	16000000	22176108
102 DalmatiansÂ	66941559	85000000	-18058441
10th & WolfÂ	53481	8000000	-7946519
12 RoundsÂ	12232937	22000000	-9767063
12 Years a SlaveÂ	56667870	20000000	36667870
127 HoursÂ	18329466	18000000	329466
13 Going on 30Â	56044241	37000000	19044241
13 HoursÂ	52822418	50000000	2822418
1408Â	71975611	25000000	46975611
15 MinutesÂ	24375436	42000000	-17624564
16 BlocksÂ	36883539	52000000	-15116461
17 AgainÂ	64149837	20000000	44149837
1911Â	127437	18000000	-17872563
2 Fast 2 FuriousÂ	127083765	76000000	51083765
2 GunsÂ	75573300	61000000	14573300
20 DatesÂ	536767	60000	476767
20 Feet from StardomÂ	4946250	1000000	3946250
200 CigarettesÂ	6851636	6000000	851636
2001: A Space OdysseyÂ	56715371	12000000	44715371
2012Â	166112167	200000000	-33887833
2016: Obama's AmericaÂ	33349949	2500000	30849949

Conditional Formatting has been used to color scale the profit margin sa that losses and profits can be seen instantly.

correlation coefficient	0.127289984

movie with max profit

movie title	The AvengersÂ
gross	1246559094
budget	440000000
profit	806559094

The correlation coefficient was found using the CORREL() function and has a value of 0.128 approx. The movie that made the maximum profit was "The AvengersA" .

Module 6: Bank Loan Case Study

Project Description:

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

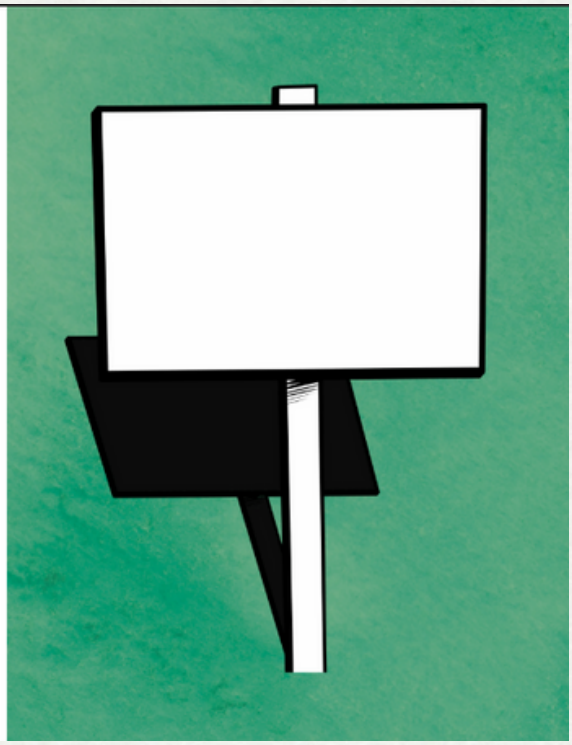
Report:

Task A: Identify Missing Data and Deal with it Appropriately.

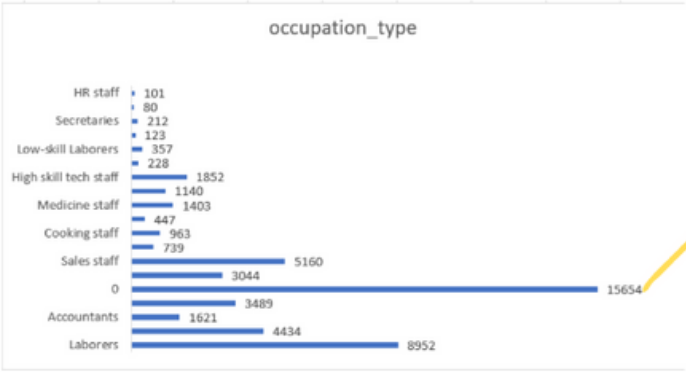
To identify blanks in each column/row, I used **COUNTA()** function to calculate the count of non-empty cells and subtracted the result from the total number of cells.

Missing values were handled using two approaches:

1. **Deleting** rows/columns with maximum blanks
2. **Replacing** blanks with appropriate values.



1. Columns/rows containing more blanks than actual data (i.e columns/rows with more than 50% blank cells) were deleted.
2. Remaining empty cells were replaced by either median or mode of the respective column.



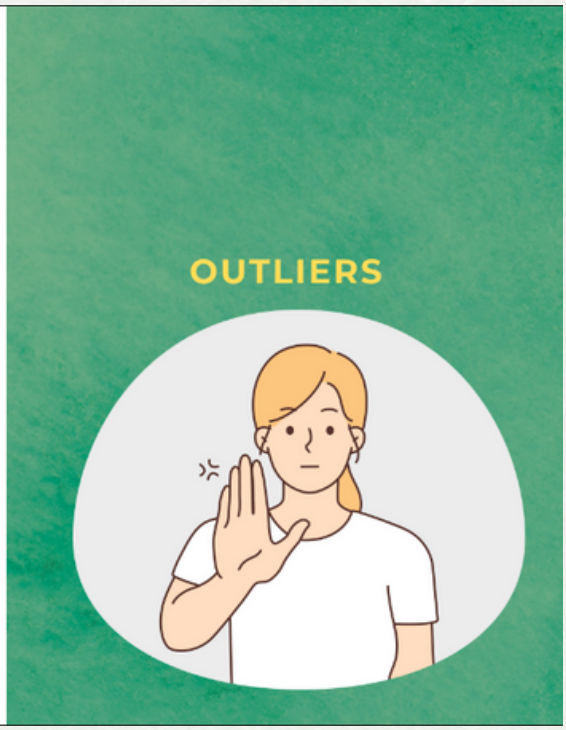
Link to my excel sheet

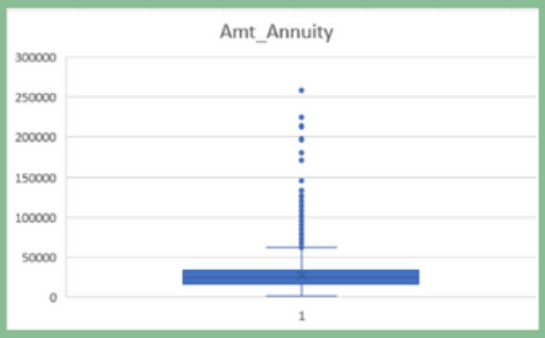
Occupation_type column in application dataset has 15654 blank cells contributing to 31% of blank cells

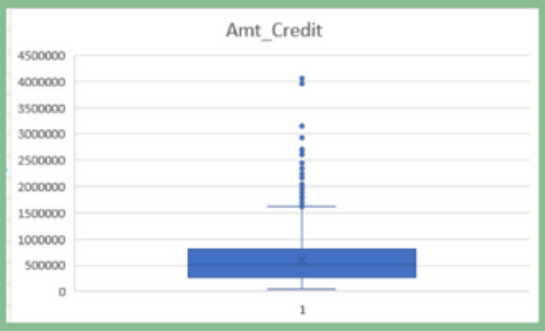
Task B: Identify Outliers in the Dataset

Ouliers were found using **QUARTILE(), IQR()** functions and plotting box plots for each variable. Every value greater than the maximum and lesser than the minimum were considered outliers.

Ouliers in each column are represented by <light red filled cells with dark red font>. This was done by conditional formatting.







AMT_CREDIT	AMT_ANNUITY
254700	25321.5
876078	49050
2250000	83515.5
1040985	30568.5
684054	77494.5
1256400	36864
720000	28552.5
270000	13500
1024740	55719
1102500	32364
544491	16047
405000	20250
912240	30276
679671	28926
180000	9000
364896	19926
312768	20353.5
1288350	37800
781920	50148
1483231.5	51687
590337	28530
101880	10827
161730	13833
447768	35505
495216	26995.5
402939	19381.5
808650	23643
76410	4513.5
90000	6529.5
270000	14134.5

APPLICATION DATASET

AMT_GOODS_PRICE	AMT_INCOME_TOTAL
351000	202500
1129500	270000
135000	67500
297000	135000
513000	121500
454500	99000
1395000	171000
1530000	360000
913500	112500
405000	135000
652500	112500
135000	38419.155
67500	67500
697500	225000
679500	189000
247500	157500
387000	108000
270000	81000
157500	112500
454500	90000
427500	135000
927000	202500
450000	450000
225000	83250
247500	135000

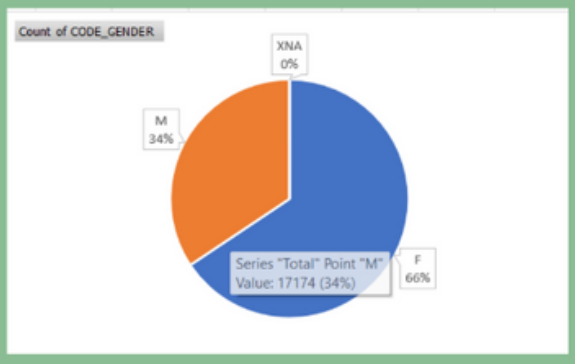
APPLICATION DATASET

Task C: Analyze Data Imbalance

An **imbalanced dataset** refers to a situation in which the distribution of classes or categories within the dataset is uneven. This **data imbalance ratio** will give you an indication of the data imbalance between the two classes. If the ratio is close to 1, it indicates a balanced dataset.

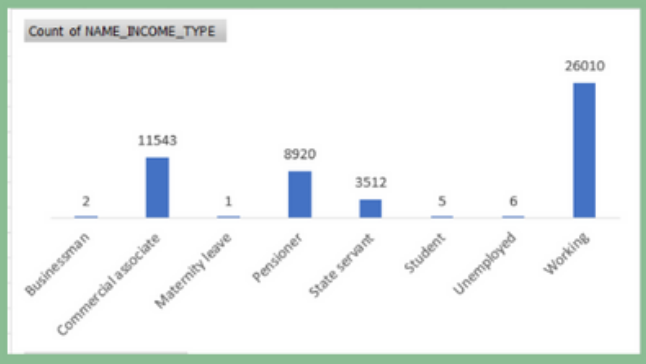
For example, if the counts of "Class 1" and "Class 2" are in cells C2 and C3 respectively, you would use the formula =**C2/C3** to calculate the ratio.

Row Labels	Count of CODE_GENDER
F	32823
M	17174
XNA	2
Grand Total	**49999**



Data imbalance ratio
0.5232

Row Labels	Count of NAME_INCOME_TYPE
Businessman	2
Commercial associate	11543
Maternity leave	1
Pensioner	8920
State servant	3512
Student	5
Unemployed	6
Working	26010



Data imbalance ratio
0.00004

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

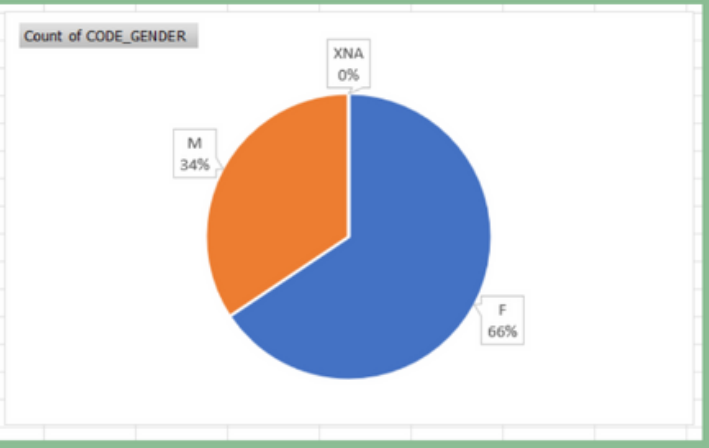
Univariate analysis focuses on understanding individual variables. - **Bivariate** analysis examines relationships between two variables

Histograms, bar charts, or box plots have been created to visualize the distributions of variables.

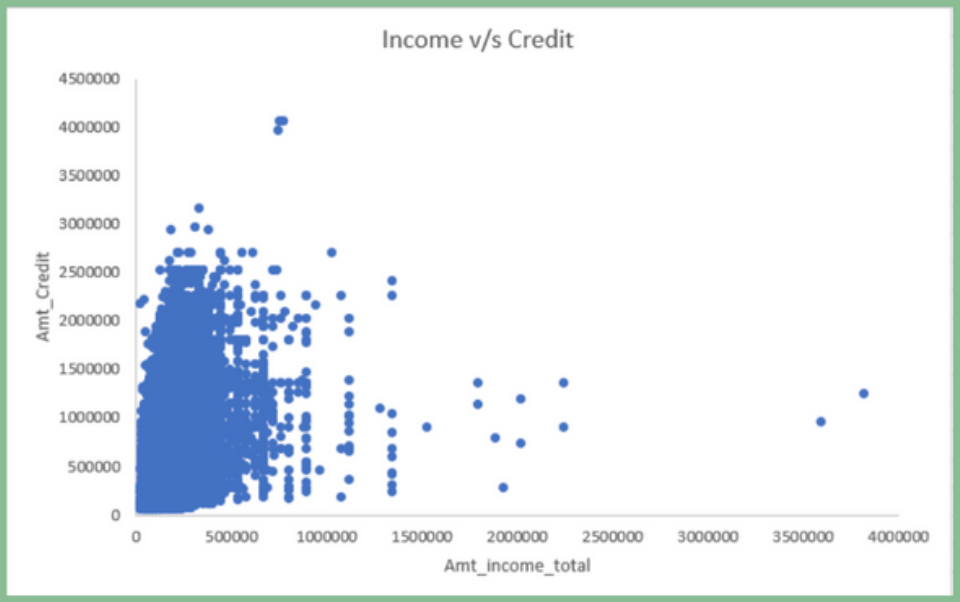
Univariate Analysis

Row Labels	Count of CODE_GENDER
F	32823
M	17174
XNA	2
Grand Total	**49999**

Univariate analysis- There are more female applicants than male applicants



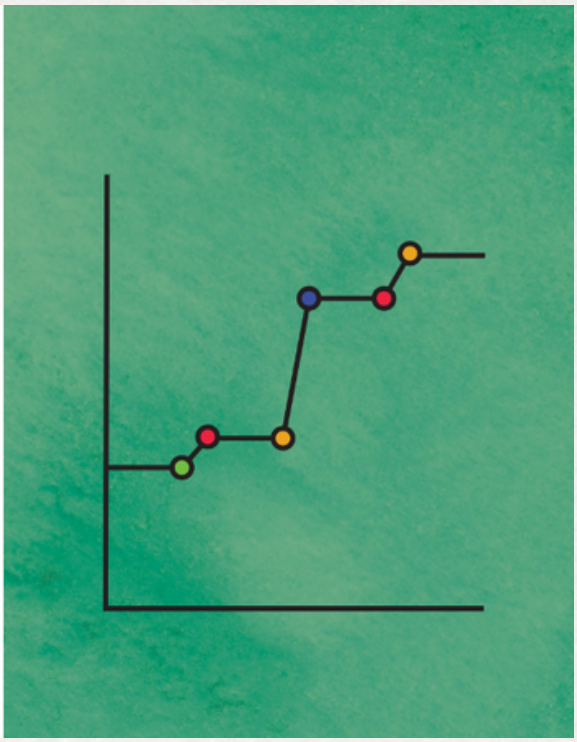
Bivariate Analysis



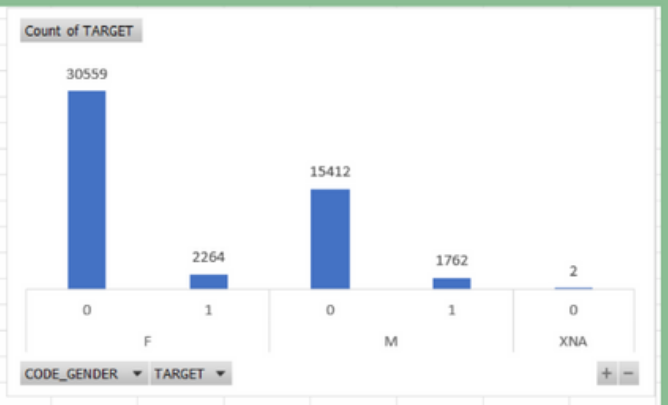
Task E: Identify Top Correlations for Different Scenarios

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

Heatmaps have been created to visualize the correlations between variables within each segment. The top correlated variables for each scenario have been highlighted using different colors.



Row Labels	Count of TARGET	
F	32823	
0	30559	
1	2264	
M	17174	
0	15412	
1	1762	
XNA	2	
0	2	
Grand Total	49999	
ratio of female defaulters		0.068976
ratio of male defaulters		0.102597



Ratio of male defaulters is more than that of female defaulters.

Income_type / defaulter	0	1	Total	Ratio of defaulters
Working	23549	2461	26010	0.094617455
State servant	3314	198	3512	0.056378132
Commercial associate	10679	864	11543	0.074850559
Pensioner	8419	501	8920	0.056165919
Unemployed	4	2	6	0.333333333
Student	5	0	5	0
Businessman	2	0	2	0
Maternity leave	1	0	1	0

Module 7: Impact of Car Features on Price and Profitability

Project Description:

The dataset contains information about car features such as the model name, brand name, number of doors, efficiency of fuels, prices etc. on which analysis have to be performed to help understand the market and make business decisions.

Report:

- **APPROACH**:
  We approach the project by analysing the dataset by cleaning it, finding the blanks and missing values, and imputing the missing values with the appropriate method(mean, median ,mode).We then find outliers and handle them to make data more efficient. Once this is done, the dataset becomes ready to be analysed and visualized.

- **TECH -STACK USED:**
  Microsoft excel was used for doing the tasks. Microsoft word was used for the report of the same.

- **INSIGHTS**:
  Following insights were drawn based on understanding and capabilities:

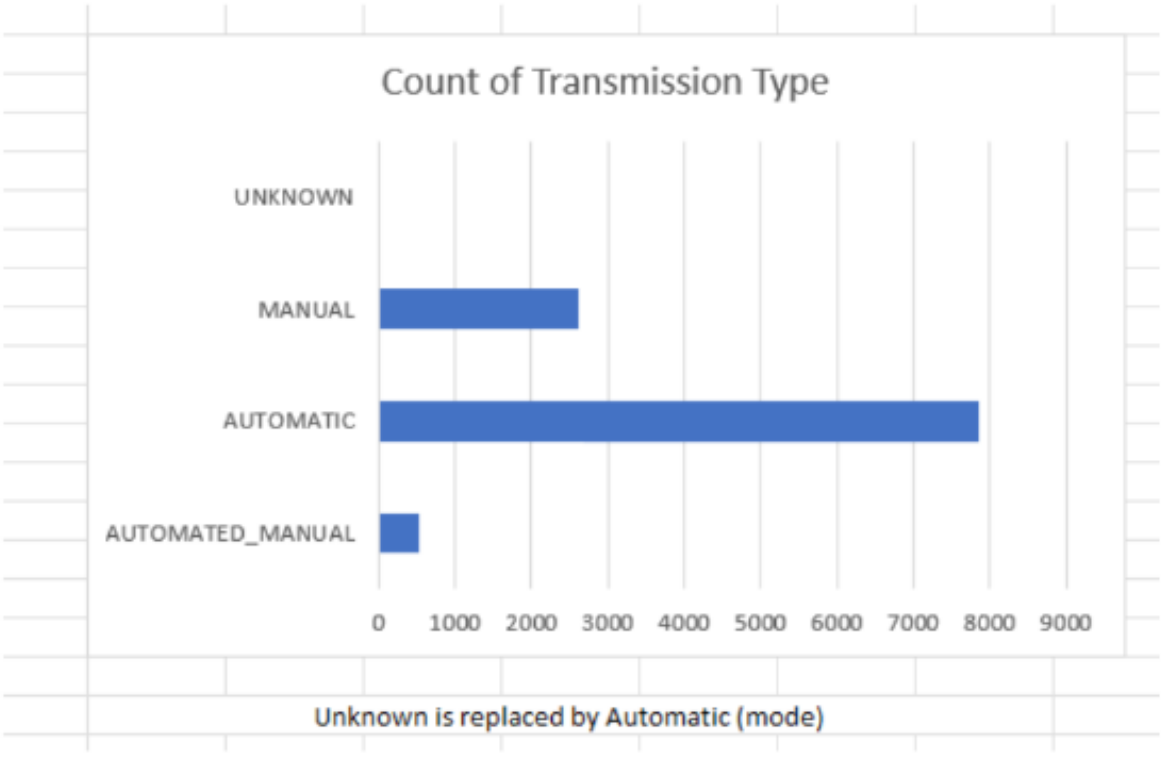
  1. As the power of engine increases, the price of cars also increases.
  2. The engine power and number of cylinders highly affect the price of cars.
  3. Luxury, sports and exotic cars are the most expensive.
  4. As the number of cylinders increases, fuel efficiency decreases.

- **RESULTS**:

  Following results were obtained while doing the project:

  1. importing dataset in excel:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Make	Model	Year	Engine Fue	Engine HP	Engine Cyli	Transmissi	Driven_Wh	Number of	Market Cat	Vehicle Siz	Vehicle Sty	highway M	city mpg	Popularity	MSRP	
2	BMW	1 Series M	2011	premium u	335	6	MANUAL	rear wheel	2	Factory Tur	Compact	Coupe	26	19	3916	46135	
3	BMW	1 Series	2011	premium u	300	6	MANUAL	rear wheel	2	Luxury,Perf	Compact	Convertible	28	19	3916	40650	
4	BMW	1 Series	2011	premium u	300	6	MANUAL	rear wheel	2	Luxury,Higl	Compact	Coupe	28	20	3916	36350	
5	BMW	1 Series	2011	premium u	230	6	MANUAL	rear wheel	2	Luxury,Perf	Compact	Coupe	28	18	3916	29450	
6	BMW	1 Series	2011	premium u	230	6	MANUAL	rear wheel	2	Luxury	Compact	Convertible	28	18	3916	34500	
7	BMW	1 Series	2012	premium u	230	6	MANUAL	rear wheel	2	Luxury,Perf	Compact	Coupe	28	18	3916	31200	
8	BMW	1 Series	2012	premium u	300	6	MANUAL	rear wheel	2	Luxury,Perf	Compact	Convertible	26	17	3916	44100	
9	BMW	1 Series	2012	premium u	300	6	MANUAL	rear wheel	2	Luxury,Higl	Compact	Coupe	28	20	3916	39300	
10	BMW	1 Series	2012	premium u	230	6	MANUAL	rear wheel	2	Luxury	Compact	Convertible	28	18	3916	36900	
11	BMW	1 Series	2013	premium u	230	6	MANUAL	rear wheel	2	Luxury	Compact	Convertible	27	18	3916	37200	
12	BMW	1 Series	2013	premium u	300	6	MANUAL	rear wheel	2	Luxury,Higl	Compact	Coupe	28	20	3916	39600	
13	BMW	1 Series	2013	premium u	230	6	MANUAL	rear wheel	2	Luxury,Perf	Compact	Coupe	28	19	3916	31500	
14	BMW	1 Series	2013	premium u	300	6	MANUAL	rear wheel	2	Luxury,Perf	Compact	Convertible	28	19	3916	44400	
15	BMW	1 Series	2013	premium u	230	6	MANUAL	rear wheel	2	Luxury	Compact	Convertible	28	19	3916	37200	
16	BMW	1 Series	2013	premium u	230	6	MANUAL	rear wheel	2	Luxury,Perf	Compact	Coupe	28	19	3916	31500	
17	BMW	1 Series	2013	premium u	320	6	MANUAL	rear wheel	2	Luxury,Higl	Compact	Convertible	25	18	3916	48250	
18	BMW	1 Series	2013	premium u	320	6	MANUAL	rear wheel	2	Luxury,Higl	Compact	Coupe	28	20	3916	43550	
19	Audi	100	1992	regular unl	172	6	MANUAL	front whee	4	Luxury	Midsize	Sedan	24	17	3105	2000	
20	Audi	100	1992	regular unl	172	6	MANUAL	front whee	4	Luxury	Midsize	Sedan	24	17	3105	2000	
21	Audi	100	1992	regular unl	172	6	AUTOMATI	all wheel d	4	Luxury	Midsize	Wagon	20	16	3105	2000	
22	Audi	100	1992	regular unl	172	6	MANUAL	front whee	4	Luxury	Midsize	Sedan	24	17	3105	2000	
23	Audi	100	1992	regular unl	172	6	MANUAL	all wheel d	4	Luxury	Midsize	Sedan	21	16	3105	2000	
24	Audi	100	1993	regular unl	172	6	MANUAL	front whee	4	Luxury	Midsize	Sedan	24	17	3105	2000	
25	Audi	100	1993	regular unl	172	6	AUTOMATI	all wheel d	4	Luxury	Midsize	Wagon	20	16	3105	2000	
26	Audi	100	1993	regular unl	172	6	MANUAL	front whee	4	Luxury	Midsize	Sedan	24	17	3105	2000	
27	Audi	100	1993	regular unl	172	6	MANUAL	front whee	4	Luxury	Midsize	Sedan	24	17	3105	2000	
28	Audi	100	1993	regular unl	172	6	MANUAL	all wheel d	4	Luxury	Midsize	Sedan	21	16	3105	2000	
29	Audi	100	1994	regular unl	172	6	AUTOMATI	front whee	4	Luxury	Midsize	Wagon	21	16	3105	2000	
30	Audi	100	1994	regular unl	172	6	MANUAL	all wheel d	4	Luxury	Midsize	Sedan	22	16	3105	2000	
31	Audi	100	1994	regular unl	172	6	MANUAL	front whee	4	Luxury	Midsize	Sedan	22	17	3105	2000	



2. Handling missing values:

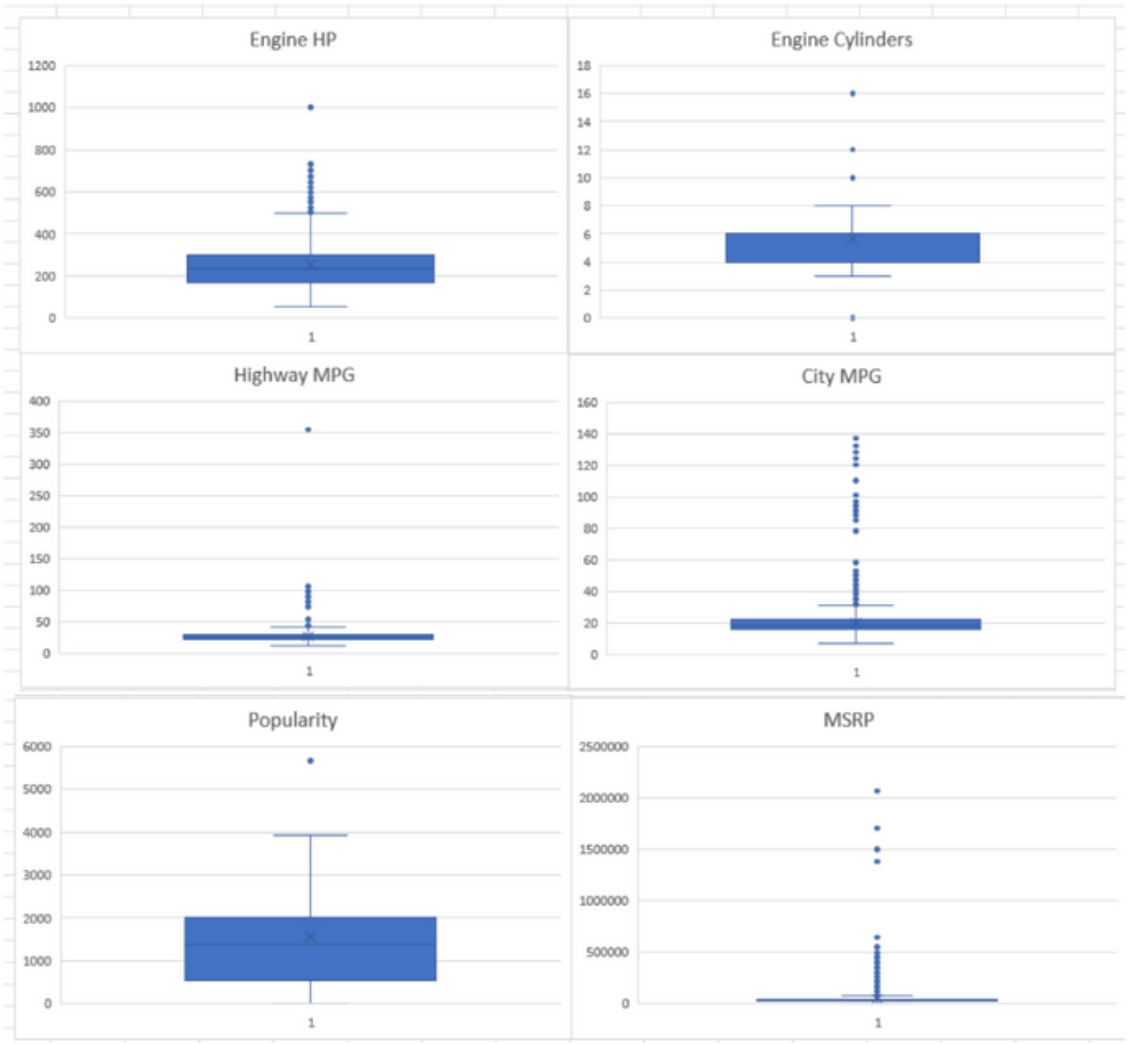
Missing values were found in four columns, engine HP, engine fuel type, number of doors and engine cylinders. The blanks in these columns, however, were not more than 50%, hence were not dropped. The blanks were replaced by appropriate values of mean, median, mode.

Row Labels	Count of Engine Fuel Type	Average of Number of Doors
Acura	**246**	**4**
premium unleaded (recommended)	146	4
premium unleaded (required)	40	3
regular unleaded	60	3
Alfa Romeo	**5**	**2**
premium unleaded (required)	5	2
Aston Martin	**91**	**2**
premium unleaded (required)	91	2
Audi	**321**	**3**
diesel	28	4
flex-fuel (premium unleaded recommended/E85)	5	3
premium unleaded (recommended)	94	3
premium unleaded (required)	149	3
regular unleaded	45	4
Bentley	**74**	**3**
flex-fuel (premium unleaded required/E85)	24	3
premium unleaded (required)	50	3
BMW	**324**	**3**
diesel	20	4
electric	4	4
premium unleaded (recommended)	27	3
premium unleaded (required)	263	3
regular unleaded	10	2
Bugatti	**3**	**2**
premium unleaded (required)	3	2
Buick	**190**	**4**

Row Labels	Average of Engine HP	Average of Engine Cylinders
diesel	184	5
electric	145	0
flex-fuel (premium unleaded recommended/E85)	283	5
flex-fuel (premium unleaded required/E85)	515	9
flex-fuel (unleaded/E85)	286	7
flex-fuel (unleaded/natural gas)		6
natural gas	110	4
premium unleaded (recommended)	277	5
premium unleaded (required)	376	7
regular unleaded	208	5
(blank)	155	6
Grand Total	**253**	**6**

Transmission Channel column had unknown values and were replaced by the most occurring category, which came out to be “Automatic” type.

3. Outliers:

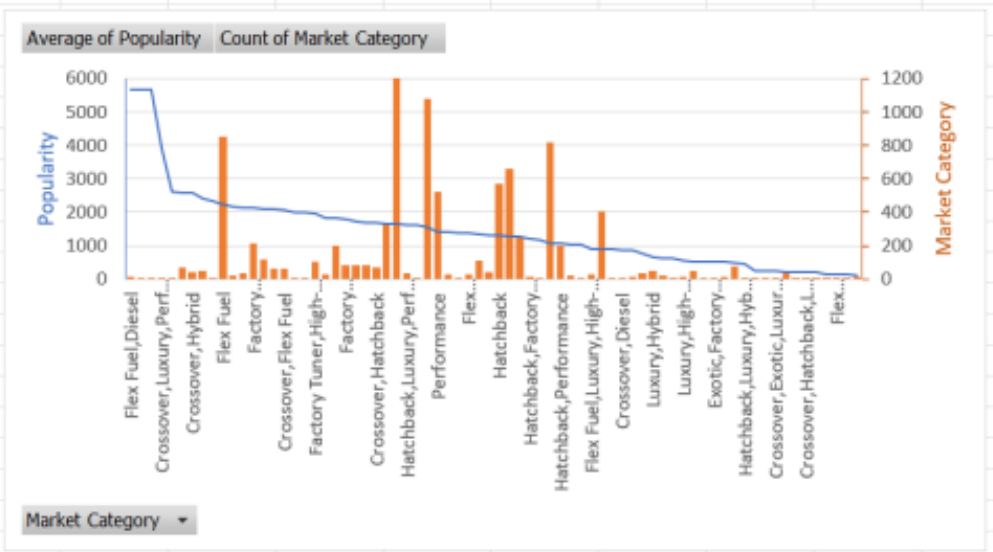


The outliers in the dataset belong to a specific category in the dataset.
For example, outliers in 'Engine HP' represented cars that require maximum engine power, that is, all cars that in sports/luxury category.
Hence, these outliers were not removed. The only outlier in 'Popularity' column was removed.

4. Analysis:

a) Understanding relation between market category and its popularity:

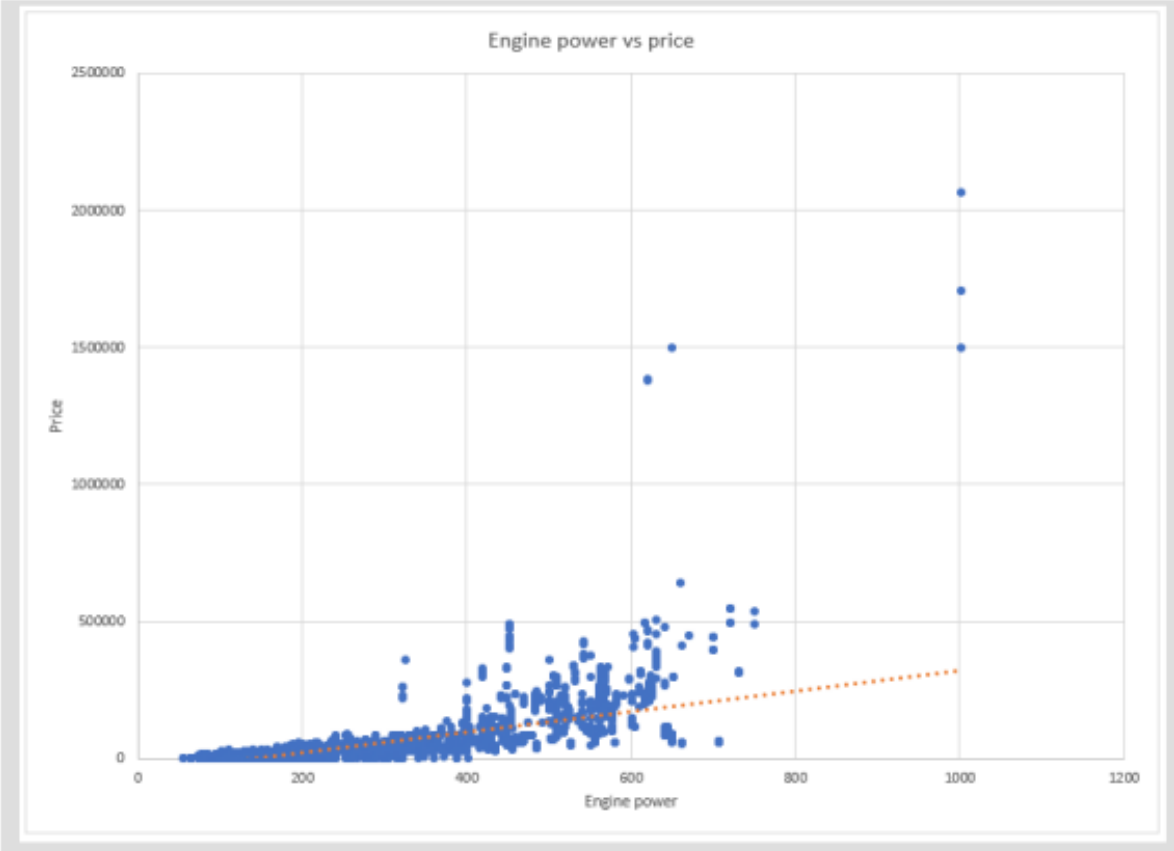
market category	Average of Popularity	Count of Market Category
Flex Fuel,Diesel	5657	16
Hatchback,Flex Fuel	5657	7
Crossover,Flex Fuel,Performance	5657	6
Crossover,Luxury,Performance,Hybrid	3916	2
Crossover,Factory Tuner,Luxury,Performance	2607	5
Crossover,Performance	2586	69
Crossover,Hybrid	2563	42
Diesel,Luxury	2416	47
Luxury,Performance,Hybrid	2333	11
Flex Fuel	2226	855
Hatchback,Factory Tuner,Performance	2174	21
Crossover,Luxury,Diesel	2149	34
Factory Tuner,Luxury,High-Performance	2133	215
Hybrid	2117	121
Hatchback,Hybrid	2111	64
Crossover,Flex Fuel	2074	64
Crossover,Hatchback,Factory Tuner,Performance	2009	6
Crossover,Hatchback,Performance	2009	6
Factory Tuner,High-Performance	1966	104
Crossover,Factory Tuner,Luxury,High-Performance	1823	26
High-Performance	1823	198
Factory Tuner,Performance	1774	84
Diesel	1731	84
Flex Fuel,Performance	1680	87



Most Popular Cars: Flex Fuel, Diesel

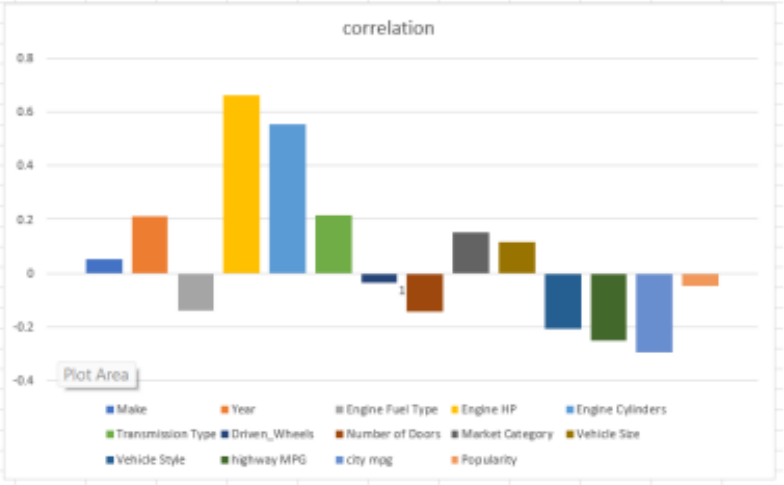
Least Popular Cars: Exotic, luxury

b) Understanding relation between engine power and price of cars:



c) Correlation of each car feature and the price of cars:

correlation with MSRP	
Make	0.051879
Year	0.211529
Engine Fuel Type	-0.138295
Engine HP	0.660886
Engine Cylinders	0.555365
Transmission Type	0.214419
Driven_Wheels	-0.035394
Number of Doors	-0.144914
Market Category	0.150428
Vehicle Size	0.116215
Vehicle Style	-0.207681
highway MPG	-0.252349
city mpg	-0.293097
Popularity	-0.048294



Engine HP and Engine Cylinders have the highest positive correlation coefficient

d) Understanding relation between car manufacturers and price of cars:

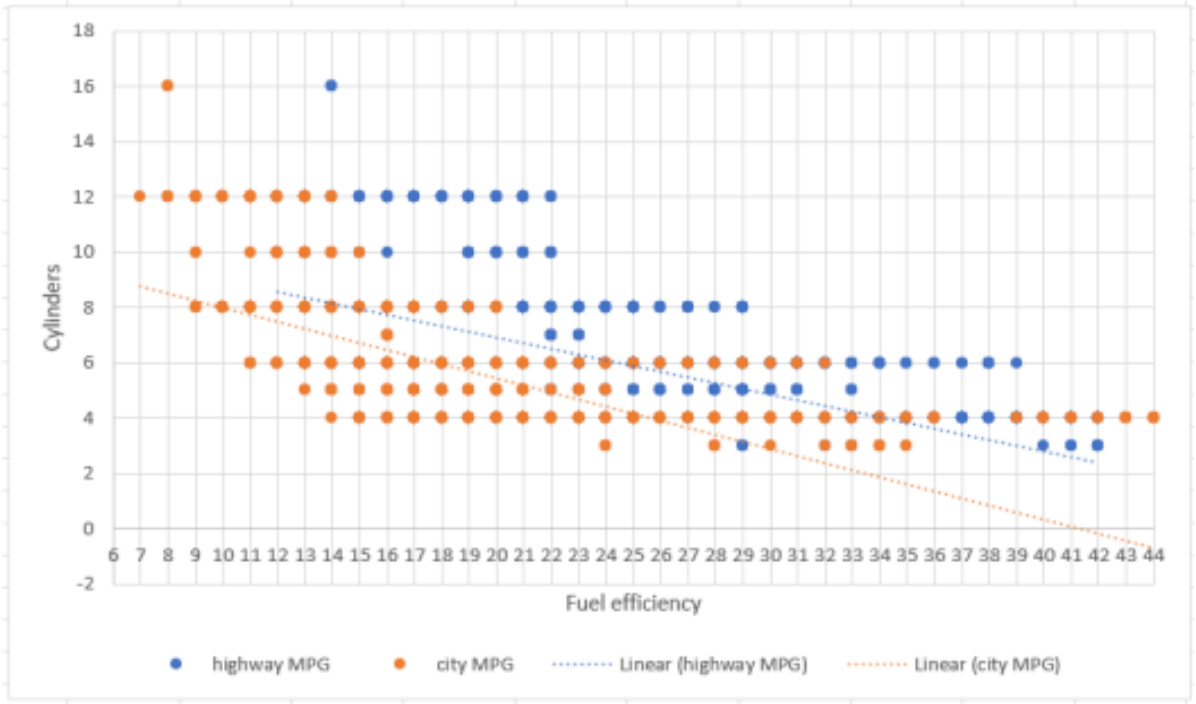
Row Labels	Average of MSRP
Bugatti	1757223.667
Maybach	546221.875
Rolls-Royce	351130.6452
Lamborghini	331567.3077
Bentley	247169.3243
McLaren	239805
Ferrari	238218.8406
Spyker	214990
Aston Martin	198123.4615
Maserati	113684.4909
Porsche	101622.3971
Tesla	85255.55556
Mercedes-Benz	72069.52786

Cars like Bugatti,Maybach, Rolls-Royce are the most expensive cars as they provide luxury

Cars like suzuki, oldsmobile, plymouth are everyday cars and hence are least priced

e) Fuel efficiency v/s Number of cylinders:

highway MPG	city mpg	Engine Cylinders
26	19	6
28	19	6
28	20	6
28	18	6
28	18	6
28	18	6
26	17	6
28	20	6
28	18	6
27	18	6
28	20	6



It can be seen that as efficiency and number of cylinders are inversely proportional.

f) Dashboard:

The dashboard is created using slicers and pivot charts in excel.

- **Conclusion**: In conclusion, this project underscores the importance of understanding the impact of car features on pricing in the automotive industry. By leveraging Excel for data analysis, we have equipped stakeholders with actionable insights to navigate the competitive landscape and drive business success.

Module 8: ABC Call Volume Trend Analysis

Project Description:

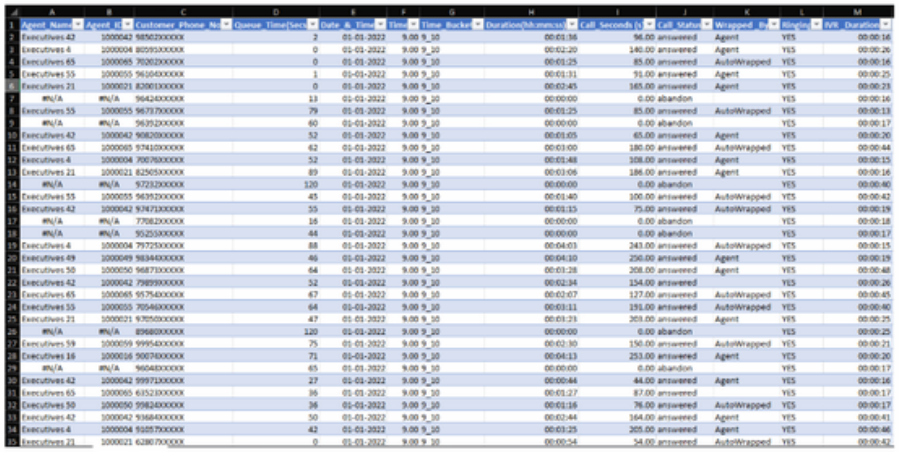
In this project, we will be diving into the world of Customer Experience (CX) analytics, specifically focusing on the inbound calling team of a company. The dataset spans 23 days and includes various details such as the agent's name and ID, the queue time (how long a customer had to wait before connecting with an agent), the time of the call, the duration of the call, and the call status (whether it was abandoned, answered, or transferred).

Report:

- **Approach:**
  The dataset is first processed to find duplicate values and missing values. After processing the data, the four major tasks were performed, the details of which are included in the report.

- **Tech stack used:**
  Microsoft excel was used for doing the tasks. Microsoft word was used for the report of the same.

- **Results:**
  1. **Importing dataset in excel:**



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Agent_Name	Agent_ID	Customer_Phone_No	Queue_Time(Secs)	Date_&_Time	Time	Time_Bucket	Duration(hh:mm:ss)	Call_Seconds (s)	Call_Status	Wrapped_By	Ringing	IVR_Duration
2	Executives 42	1000042	98502200XXX	2	01-01-2022	9.00 9_10		00:01:36	96.00	answered	Agent	YES	00:00:16
3	Executives 4	1000004	80595XXXXX	0	01-01-2022	9.00 9_10		00:02:20	140.00	answered	Agent	YES	00:00:26
4	Executives 65	1000065	70262XXXXX	0	01-01-2022	9.00 9_10		00:01:25	85.00	answered	AutoWrapped	YES	00:00:16
5	Executives 55	1000055	96160XXXXX	1	01-01-2022	9.00 9_10		00:01:31	91.00	answered	Agent	YES	00:00:25
6	Executives 21	1000021	82001XXXXX	0	01-01-2022	9.00 9_10		00:02:45	165.00	answered	Agent	YES	00:00:23
7	#N/A	#N/A	96424XXXXX	13	01-01-2022	9.00 9_10		00:00:00	0.00	abandon		YES	00:00:16
8	Executives 55	1000055	96737XXXXX	79	01-01-2022	9.00 9_10		00:01:25	85.00	answered	AutoWrapped	YES	00:00:13
9	#N/A	#N/A	96352XXXXX	60	01-01-2022	9.00 9_10		00:00:00	0.00	abandon		YES	00:00:17
10	Executives 42	1000042	90820XXXXX	52	01-01-2022	9.00 9_10		00:01:05	65.00	answered	Agent	YES	00:00:20
11	Executives 65	1000065	97410XXXXX	62	01-01-2022	9.00 9_10		00:03:00	180.00	answered	AutoWrapped	YES	00:00:44
12	Executives 4	1000004	90078XXXXX	52	01-01-2022	9.00 9_10		00:01:48	108.00	answered	Agent	YES	00:00:15
13	Executives 21	1000021	82505XXXXX	89	01-01-2022	9.00 9_10		00:03:06	186.00	answered	Agent	YES	00:00:16
14	#N/A	#N/A	97232XXXXX	120	01-01-2022	9.00 9_10		00:00:00	0.00	abandon		YES	00:00:40
15	Executives 55	1000055	96352XXXXX	45	01-01-2022	9.00 9_10		00:01:40	100.00	answered	AutoWrapped	YES	00:00:42
16	Executives 42	1000042	97471XXXXX	55	01-01-2022	9.00 9_10		00:01:15	75.00	answered	AutoWrapped	YES	00:00:19
17	#N/A	#N/A	77082XXXXX	16	01-01-2022	9.00 9_10		00:00:00	0.00	abandon		YES	00:00:18
18	#N/A	#N/A	95253XXXXX	44	01-01-2022	9.00 9_10		00:00:00	0.00	abandon		YES	00:00:17
19	Executives 4	1000004	79725XXXXX	88	01-01-2022	9.00 9_10		00:04:03	243.00	answered	AutoWrapped	YES	00:00:15
20	Executives 49	1000049	98344XXXXX	46	01-01-2022	9.00 9_10		00:04:10	250.00	answered	Agent	YES	00:00:19
21	Executives 50	1000050	96873XXXXX	64	01-01-2022	9.00 9_10		00:01:28	208.00	answered	Agent	YES	00:00:48
22	Executives 42	1000042	79899XXXXX	52	01-01-2022	9.00 9_10		00:02:34	154.00	answered		YES	00:00:26
23	Executives 65	1000065	95754XXXXX	67	01-01-2022	9.00 9_10		00:02:07	127.00	answered	AutoWrapped	YES	00:00:45
24	Executives 55	1000055	70546XXXXX	64	01-01-2022	9.00 9_10		00:03:11	191.00	answered	AutoWrapped	YES	00:00:40
25	Executives 21	1000021	97055XXXXX	47	01-01-2022	9.00 9_10		00:03:23	203.00	answered	Agent	YES	00:00:25
26	#N/A	#N/A	89680XXXXX	120	01-01-2022	9.00 9_10		00:00:00	0.00	abandon		YES	00:00:25
27	Executives 59	1000059	99954XXXXX	75	01-01-2022	9.00 9_10		00:02:30	150.00	answered	AutoWrapped	YES	00:00:21
28	Executives 16	1000016	90078XXXXX	71	01-01-2022	9.00 9_10		00:04:13	253.00	answered	Agent	YES	00:00:20
29	#N/A	#N/A	96048XXXXX	65	01-01-2022	9.00 9_10		00:00:00	0.00	abandon		YES	00:00:17
30	Executives 42	1000042	99971XXXXX	27	01-01-2022	9.00 9_10		00:00:44	44.00	answered	Agent	YES	00:00:16
31	Executives 65	1000065	61521XXXXX	36	01-01-2022	9.00 9_10		00:01:27	87.00	answered		YES	00:00:17
32	Executives 50	1000050	99824XXXXX	36	01-01-2022	9.00 9_10		00:01:16	76.00	answered	AutoWrapped	YES	00:00:17
33	Executives 42	1000042	93684XXXXX	50	01-01-2022	9.00 9_10		00:02:44	164.00	answered	Agent	YES	00:00:41
34	Executives 4	1000004	91057XXXXX	42	01-01-2022	9.00 9_10		00:03:25	205.00	answered	Agent	YES	00:00:46
35	Executives 21	1000021	62807XXXXX	0	01-01-2022	9.00 9_10		00:00:54	54.00	answered	AutoWrapped	YES	00:00:42

2. Removing Duplicates and handling missing values:

No duplicates were found in the dataset. The first two columns, viz, Agent_name and Agent_id contained #N/A values, however, these were not removed as they indicated that the calls were abandoned.

There were blanks in the wrapped_by column. These blanks were handled as follow:

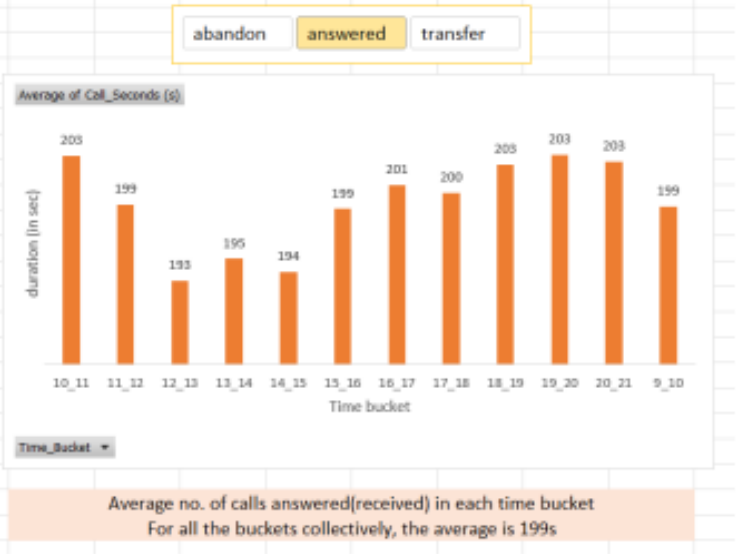
- For the blanks whose corresponding Agent_name / Agent_id were "#N/A", the values was set as "Abandoned calls".
- Rest of the blanks were replaced by the mode of the column, which came out to be "Agent".

Agent_Name	Agent_ID	Customer_Phone_No	Queue_Time(Secs)	Date_&_Time	Time	Time_Bucket	Duration(hh:mm:ss)	Call_Seconds (s)	Call_Status	Wrapped_By	Ringing	IVR_Duration
#N/A	#N/A	9642400000	13	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	Abandoned calls	YES	00:00:16
#N/A	#N/A	9639200000	60	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	Abandoned calls	YES	00:00:17
#N/A	#N/A	9723200000	120	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	Abandoned calls	YES	00:00:40
#N/A	#N/A	7708200000	16	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	Abandoned calls	YES	00:00:18
#N/A	#N/A	9525500000	44	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	Abandoned calls	YES	00:00:17
#N/A	#N/A	8968000000	120	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	Abandoned calls	YES	00:00:25
#N/A	#N/A	9604800000	65	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	Abandoned calls	YES	00:00:17
#N/A	#N/A	8778200000	16	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	Abandoned calls	YES	00:00:17
#N/A	#N/A	8778200000	16	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	Abandoned calls	YES	00:00:16
#N/A	#N/A	8778200000	16	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	Abandoned calls	YES	00:00:16
#N/A	#N/A	8778200000	7	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	Abandoned calls	YES	00:00:26
#N/A	#N/A	7622900000	44	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	Abandoned calls	YES	00:00:44
#N/A	#N/A	8553200000	50	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	Abandoned calls	YES	00:00:15
#N/A	#N/A	7007600000	45	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	Abandoned calls	YES	00:00:18

3. Analysis:

a) Average Call Duration:

Time bucket	Average of Call_Seconds (s)
10_11	203
11_12	199
12_13	193
13_14	195
14_15	194
15_16	199
16_17	201
17_18	200
18_19	203
19_20	203
20_21	203
9_10	199
Grand Total	**199**

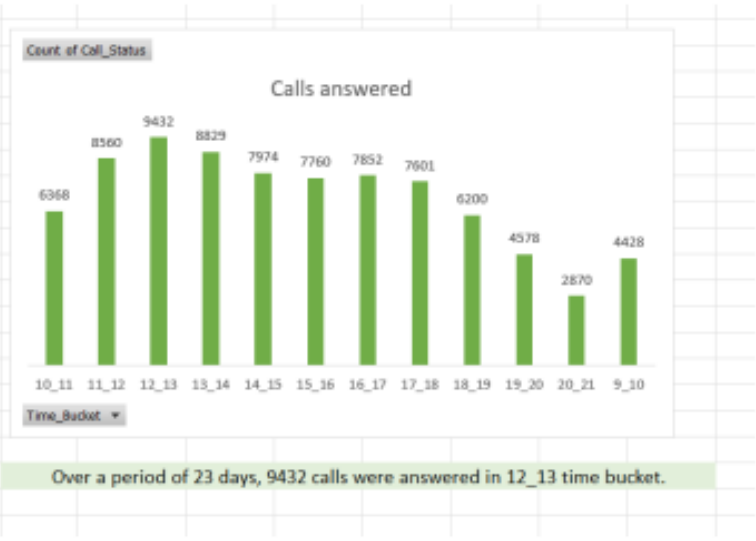


Average no. of calls answered(received) in each time bucket
For all the buckets collectively, the average is 199s

The average duration of calls for 23 days is 199 secs

b) Call Volume Analysis:

Count of Call_Status	Column Labels
Row Labels	answered
10_11	6368
11_12	8560
12_13	9432
13_14	8829
14_15	7974
15_16	7760
16_17	7852
17_18	7601
18_19	6200
19_20	4578
20_21	2870
9_10	4428



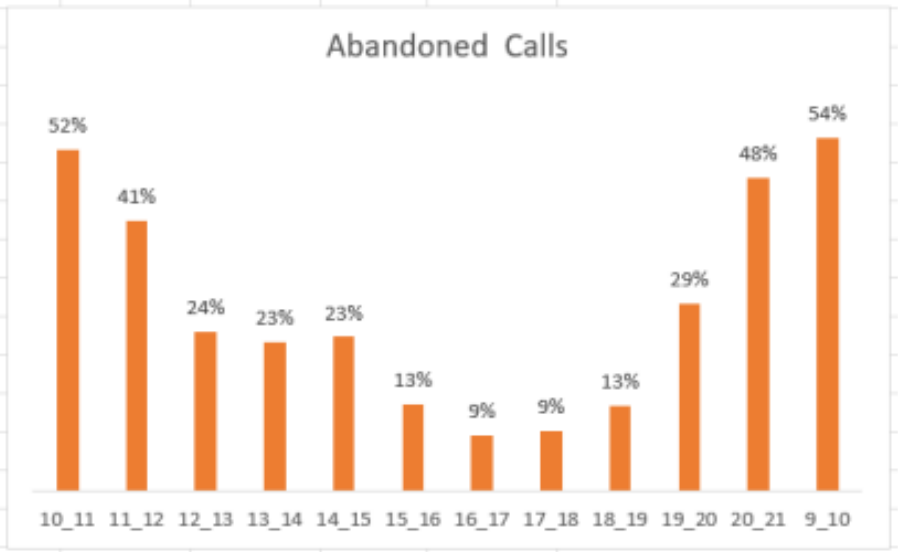
The above column chart shows the number of calls answered in 23 days for each time bucket.

c) Manpower Planning:

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

per day per agent	
Total working hours	9
time spent on break	1.5
actual working hours	7.5
avg. hours spent on calls	4.5
for 23 days per agent	
Total working hours	207
time spent on break	34.5
actual working hours	172.5
avg. hours spent on calls	103.5
avg duration of a call (in s)	199
Avg. no. of calls answered by an agent in 23 days	1872

Assumptions based insights



It can be seen that most calls (>50%) were abandoned on the beginning or in the end of the day.

for 23 days						
Time bucket	abandoned calls	Total calls	abandoned rate	% of calls to be answered	abandoned calls if rate is dropped to 10%	Man power required
10_11	6911	13313	52%	11981.7	1331.3	6
11_12	6028	14626	41%	13163.4	1462.6	7
12_13	3073	12652	24%	11386.8	1265.2	6
13_14	2617	11561	23%	10404.9	1156.1	6
14_15	2475	10561	23%	9504.9	1056.1	5
15_16	1214	9159	13%	8243.1	915.9	4
16_17	747	8788	9%	7909.2	878.8	4
17_18	783	8534	9%	7680.6	853.4	4
18_19	933	7238	13%	6514.2	723.8	3
19_20	1848	6463	29%	5816.7	646.3	3
20_21	2625	5505	48%	4954.5	550.5	3
9_10	5149	9588	54%	8629.2	958.8	5

The "Manpower required" column represents the number of additional agents needed to answer abandoned calls so that abandon rate drops down to 10%.

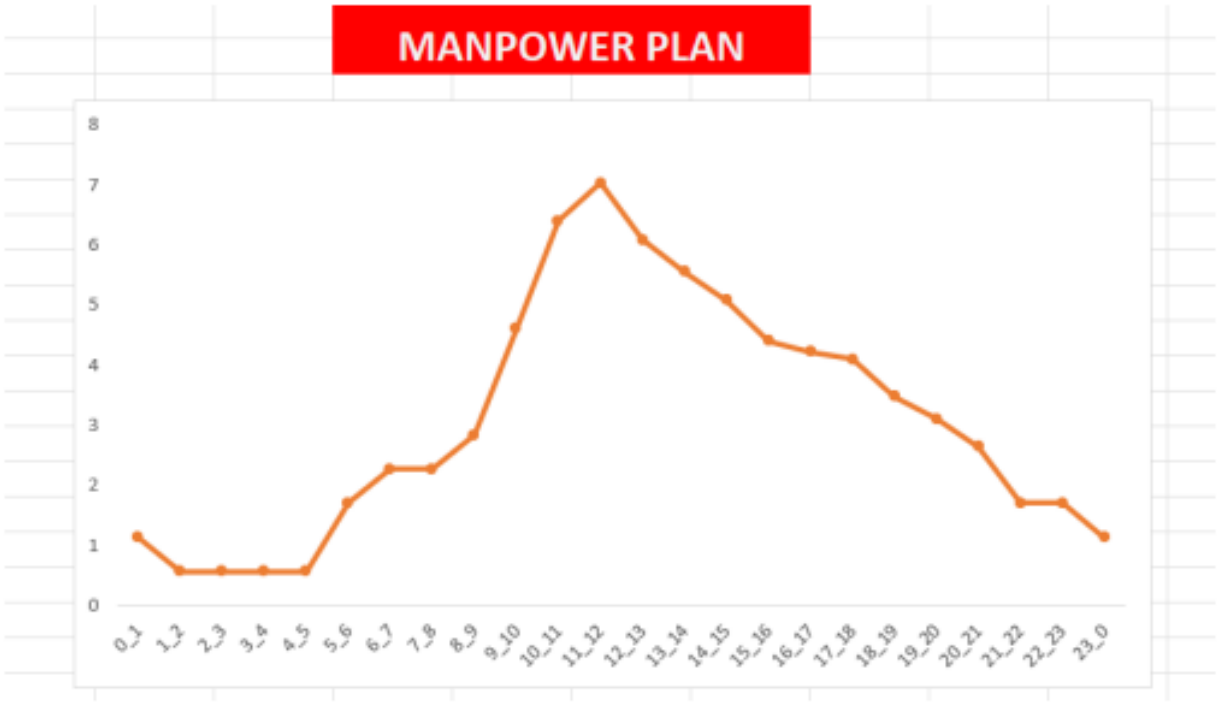
d) Night Shift Manpower Planning:

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

Distribution of calls between 9pm to 9am

Total calls coming in day	117988
Total calls coming at night	35396.4
no. of calls answered by an agent	1872

Time bucket	% of incoming calls (night)	Total incoming calls (night)	abandoned calls	Calls answered	Manpower required
22_23	10%	3540	354	3186	2
23_0	7%	2360	236	2124	1
0_1	7%	2360	236	2124	1
1_2	3%	1180	118	1062	1
2_3	3%	1180	118	1062	1
3_4	3%	1180	118	1062	1
4_5	3%	1180	118	1062	1
5_6	10%	3540	354	3186	2
6_7	13%	4720	472	4248	2
7_8	13%	4720	472	4248	2
8_9	17%	5899	590	5309	3
21_22	10%	3540	354	3186	2



Above is the manpower plan (day + night)

- Conclusion:
  The project helped me solve complex problems on man power needed to do the task. It was challenging and helped me strengthen my skills.

Appendix

Find all my data analytics project here

Projects

View my linkedin profile here

Linkedin

View my instagram profile here

instagram