Data Analysis Portfolio

By – Ashok Reddy

Table of Contents

Professional Background

I recently graduated with a Bachelor of Computer Science and Artificial Intelligence from Indian Institute of Information Technology Lucknow with 7.6 CGPA

My undergraduate studies have provided me with a strong foundation in programming, which has fueled my enthusiasm for transitioning into the field of Data Analytics.

I possess skills in Python, Microsoft Excel, SQL, Power BI, Machine Learning, and Data Visualization.

As a fresher, I am eager to face the real challenges of the corporate world and gain a deeper understanding of industry operations.

I am adaptable and enthusiastic about learning new things. While I have solid theoretical knowledge, I am eager to apply it practically and am confident that my dedication and effort will facilitate my growth in this field.

i) Description

We use Data Analytics in everyday life without even knowing it.

Your task is to give the example(s) of such a real-life situation where we use Data Analytics and link it with the data analytics process.

ii)Design

Case 1: Learning a new programming language

- Plan: I will first decide the programming language I want to learn.

- Prepare: Then I will decide whether I want to learn it by taking an online course or by doing it myself through YouTube and documentations.

- Process: I will take short breaks between learning, and during that break, I will do some other chores thus doing proper time management.

- Analyze: I will see to it that I am being productive enough and check my understanding of the concept by making small projects/assignments.

- Share: Now I communicate my ideas to my parents for further confirmation.

- Act: I finally start learning the programming language!

Case 2: Deciding a college for post-graduation

- Plan: I will first decide the college in which I want to pursue post-graduation.

- Prepare: Then I need to check whether I can get admission in the college directly or by giving an entrance examination.

- Process: I need to check the admission process and the fee structure of the college. I will also check the eligibility criteria for taking admission in the college.

- Analyze: I will see the college infrastructure and check their website. I will also see the student reviews about the college to get a better understanding.

- Share: Now I communicate my ideas to my parents for further confirmation.

- Act: I finally decide to visit the college.


5010
1002
56

Instagram User Analytics

i) Description

You're a data analyst working with the product team at Instagram. Your role involves analyzing user interactions and engagement with the Instagram app to provide valuable insights that can help the business grow.

The insights derived from this analysis can be used by various teams within the business. For example, the marketing team might use these insights to launch a new campaign, the product team might use them to decide on new features to build, etc.

The goal of this project is to use your SQL skills to extract meaningful insights from the data. Your findings could potentially influence the future development of one of the world's most popular social media platforms.

ii) The Problem

A) Marketing Analysis

- Loyal User Reward: The marketing team wants to reward the most loyal users, i.e., those who have been using the platform for the longest time.

- Inactive User Engagement: The team wants to encourage inactive users to start posting by sending them promotional emails.

- Contest Winner Declaration: The team has organized a contest where the user with the most likes on a single photo wins.

- Hashtag Research: A partner brand wants to know the most popular hashtags to use in their posts to reach the most people.

- Ad Campaign Launch: The team wants to know the best day of the week to launch ads.

B) Investor Metrics

- User Engagement: Investors want to know if users are still active and posting on Instagram.

- Bots and Fake Accounts: Investors want to know if the platform is crowded with fake and dummy accounts.

iii) Design

Steps taken to load the database:

- Imported the provided dataset into MySQL Workbench using Table Data Import Wizard.

- Executed SQL queries with the help of various commands.

Software Used: MySQL Workbench

Version: 8.0 Community Edition

iv) Insights

A) Marketing Analysis

- Loyal User Award: The five oldest users who were using the platform were:

```
SELECT id, username, created_at as oldest_users
FROM ig_clone.users
ORDER BY oldest_users ASC LIMIT 5
```

id	username	oldest_users
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.Jacobson2	2016-05-14 07:56:26

- Inactive User Engagement: The inactive users who had never posted a single photo were:

```
SELECT u.id, username
FROM ig_clone.users u
LEFT JOIN ig_clone.photos p on u.id = p.user_id
WHERE image_url IS NULL
```

id	username
5	Aniya_Hackett
7	Kasandra_Homenick
14	Jaclyn81
21	Rocio33
24	Maxwell.Halvorson
25	Tierra.Trantow
34	Pearl7
36	Ollie_Ledner37
41	Mckenna17
45	David.Osinski47
49	Morgan.Kassulke
53	Linnea59
54	Duane60
57	Julien_Schmidt
66	Mike.Auer39
68	Franco_Keebler64
71	Nia_Haag
74	Hulda.Macejkovic
75	Leslie67
76	Janelle.Nikolaus81
80	Darby_Herzog
81	Esther.Zulauf61
83	Bartholome.Bernhard
89	Jessyca_West
90	Esmeralda.Mraz57
91	Bethany20

- Contest Winner Declaration: The user with the most likes on a single photo was:

```
SELECT u.id, u.username, u.created_at, COUNT(l.user_id) as likes
FROM ig_clone.users u
LEFT JOIN ig_clone.likes l on u.id = l.user_id
GROUP BY u.id, u.username, u.created_at
ORDER BY likes DESC LIMIT 1
```

id	username	created_at	likes
5	Aniya_Hackett	2016-12-07 01:04:39	257

- Hashtag Research: The top five commonly used hashtags on the platform were:

```
SELECT t.id, t.tag_name, t.created_at, COUNT(p.tag_id) as tag_count
FROM ig_clone.tags t
LEFT JOIN ig_clone.photo_tags p on t.id = p.tag_id
GROUP BY t.id, t.tag_name, t.created_at
ORDER BY tag_count DESC LIMIT 5
```

id	tag_name	created_at	tag_count
21	smile	2024-06-14 18:32:19	59
20	beach	2024-06-14 18:32:19	42
17	party	2024-06-14 18:32:19	39
13	fun	2024-06-14 18:32:19	38
18	concert	2024-06-14 18:32:19	24

- <u>Ad Campaign Launch:</u> The best day of the week to launch an ad campaign would definitely be Thursday as most number of users registered on that day.

```
SELECT DAYOFWEEK(created_at) as day_of_week, COUNT(created_at) as number_of_users_registered
FROM ig_clone.users
GROUP BY day_of_week
ORDER BY number_of_users_registered DESC LIMIT 1
```

day_of_week	number_of_users_registered
5	16

<u>NOTE:</u> For the function DAYOFWEEK(), the day starts from Sunday and represents 0.

B) Investor Metrics

- User Engagement: By checking the average posts per user and the average of total number of photos and users, we can infer that the users are still active.

```
WITH AveragePosts AS (
    SELECT ROUND(AVG(posts_count),2) AS average_posts_per_user,
    ROUND((SUM(posts_count) / COUNT(*)),2) AS average
    FROM(
        SELECT u.id, COUNT(p.image_url) as posts_count
        FROM ig_clone.users u
        LEFT JOIN ig_clone.photos p ON u.id = p.user_id
        GROUP BY u.id
    ) AS like_count
)
SELECT * FROM AveragePosts
```

average_posts_per_user	average
2.57	2.57

- Bots and Fake Account: The criteria to find bots or fake account was to find the users who liked every single post on the platform, as it was not practically possible.

```
WITH MaxLikeCount AS (
    SELECT MAX(like_count) AS max_like
    FROM(
        SELECT u.id, COUNT(l.user_id) AS like_count
        FROM ig_clone.users u
        LEFT JOIN ig_clone.likes l ON u.id = l.user_id
        GROUP BY u.id
    ) AS count_of_likes
),
UsersWithMaxLikes AS  (
    SELECT u.id, u.username, u.created_at, COUNT(l.user_id) AS like_count
    FROM ig_clone.users u
    LEFT JOIN ig_clone.likes l ON u.id = l.user_id
    GROUP BY u.id, u.username, u.created_at
    HAVING COUNT(l.user_id) = (SELECT max_like FROM MaxLikeCount)
)
SELECT * FROM UsersWithMaxLikes
```

id	username	created_at	like_count
5	Aniya_Hackett	2016-12-07 01:04:39	257
14	Jaclyn81	2017-02-06 23:29:16	257
21	Rocio33	2017-01-23 11:51:15	257
24	Maxwell.Halvorson	2017-04-18 02:32:44	257
36	Ollie_Ledner37	2016-08-04 15:42:20	257
41	Mckenna17	2016-07-17 17:25:45	257
54	Duane60	2016-12-21 04:43:38	257
57	Julien_Schmidt	2017-02-02 23:12:48	257
66	Mike.Auer39	2016-07-01 17:36:15	257
71	Nia_Haag	2016-05-14 15:38:50	257
75	Leslie67	2016-09-21 05:14:01	257
76	Janelle.Nikolaus81	2016-07-21 09:26:09	257
91	Bethany20	2016-06-03 23:31:53	257

v) Conclusion

By working on this project, I was able to master SQL at a very fast rate.

I learned some advanced SQL queries which I didn't knew before and also implemented them with the help of this project.

I got a glimpse of how data is analyzed by a data analyst with the help of MySQL Workbench.

I was able to generate actionable insights and was able to aid the team.

SQL queries may seem difficult at first, but with proper practice, they become very easy to execute!

Operation Analytics and Investigating Metric Spikes

i) Description

Operational Analytics is a crucial process that involves analyzing a company's end-to-end operations. This analysis helps identify areas for improvement within the company. As a Data Analyst, you'll work closely with various teams, such as operations, support, and marketing, helping them derive valuable insights from the data they collect.

One of the key aspects of Operational Analytics is investigating metric spikes. This involves understanding and explaining sudden changes in key metrics, such as a dip in daily user engagement or a drop in sales.

Your goal is 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.

ii) The Problem

A) Case Study 1: Job Data Analysis

- Jobs Reviewed Over Time: Calculate the number of jobs reviewed per hour for each day in November 2020.

- Throughput Analysis: Calculate the 7-day rolling average of throughput (number of events per second).

- Language Share Analysis: Calculate the percentage share of each language in the last 30 days.

- Duplicate Rows Detection: Identify duplicate rows in the data.

B) Case Study 2: Investigating Metric Spike

- Weekly User Engagement: Measure the activeness of users on a weekly basis.

- User Growth Analysis: Analyze the growth of users over time for a product.

- Weekly Retention Analysis: Analyze the retention of users on a weekly basis after signing up for a product.

- Weekly Engagement Per Device: Measure the activeness of users on a weekly basis per device.

- Email Engagement Analysis: Analyze how users are engaging with the email service.

iii) Design

Steps taken to load the database:

- Imported the provided dataset into MySQL Workbench using Table Data Import Wizard.

- Executed SQL queries with the help of various commands.

Software Used: MySQL Workbench

Version: 8.0 Community Edition

iv) Insights

A) Case Study 1: Job Data Analysis

- Jobs Reviewed Over Time: The number of jobs reviewed per hour for each day in November 2020 were:

```
SELECT
    ds,
    COUNT(job_id) AS number_of_jobs_reviewed,
    ROUND(SUM(time_spent) / 3600, 3) AS time_per_hour
FROM
    job_data
WHERE
    ds LIKE '11/__/2020'
GROUP BY ds
ORDER BY ds ASC
```

ds	number_of_jobs_reviewed	time_per_hour
11/25/2020	1	0.013
11/26/2020	1	0.016
11/27/2020	1	0.029
11/28/2020	2	0.009
11/29/2020	1	0.006
11/30/2020	2	0.011

- <u>Throughput Analysis:</u> We can say that using the 7-day rolling average method is more preferrable as it has a constant growth compared to the daily metrics.

```
SELECT
    ds, ROUND(AVG(throughput) OVER(ORDER BY ds ROWS BETWEEN 6 PRECEDING AND CURRENT ROW), 2) as moving_average,
    throughput as daily_metric
    FROM(
        SELECT ds, ROUND(COUNT(event) / SUM(time_spent), 2) as throughput
        FROM job_data
        GROUP BY ds
    )AS daily_throughput
ORDER BY ds
```

ds	moving_average	daily_metric
11/25/2020	0.02	0.02
11/26/2020	0.02	0.02
11/27/2020	0.02	0.01
11/28/2020	0.03	0.06
11/29/2020	0.03	0.05
11/30/2020	0.04	0.05

➢ <u>Language Share Analysis:</u> The percentage share of each language for the last 30 days were:

```
SELECT
    language,
    ROUND(100 * COUNT(*) / (SELECT
                    COUNT(*)
                FROM
                    job_data),
            2) AS percentage_share
FROM
    job_data
GROUP BY language
```

language	percentage_share
English	12.50
Arabic	12.50
Persian	37.50
Hindi	12.50
French	12.50
Italian	12.50

- Duplicate Rows Detection: We can clearly see that there were no duplicate rows.

```sql
SELECT
    ds,
    job_id,
    actor_id,
    event,
    language,
    time_spent,
    org,
    COUNT(*) AS count
FROM
    job_data
GROUP BY ds , job_id , actor_id , event , language , time_spent , org
HAVING COUNT(*) > 1
```

ds	job_id	actor_id	event	language	time_spent	org	count

B) Case Study 2: Investigating Metric Spikes

- Weekly User Engagement: The activeness of users on a weekly basis was:

```
WITH WeeklyUserEngagement AS (
    SELECT WEEK(STR_TO_DATE(activated_at, '%d-%m-%Y')) AS weeknum,
    COUNT(user_id) AS users_count
    FROM users_case2
    WHERE state = 'active'
    GROUP BY weeknum
)
SELECT * FROM WeeklyUserEngagement
```

weeknum	users_count
0	106
1	156
2	157
3	149
4	160
5	181
6	173
7	167
8	163
9	176
10	186
11	161
12	181
13	206
14	197
15	207
16	225
17	219
18	207
19	242
20	215
21	232
22	250
23	246
24	274

➢ <u>User Growth Analysis:</u> The growth of users over time for a product was:

```
SELECT
    e.device, COUNT(u.user_id) AS user_count
FROM
    users_case2 u
        LEFT JOIN
    events e ON u.user_id = e.user_id
WHERE
    e.device IS NOT NULL
GROUP BY e.device
```

device	user_count
dell inspiron notebook	19669
iphone 5	25883
iphone 4s	9615
windows surface	3451
macbook air	26786
iphone 5s	15929
kindle fire	4090
macbook pro	57295
ipad mini	5591
nexus 7	6540
nexus 5	16502
samsung galaxy s4	18653
lenovo thinkpad	36978
samsumg galaxy tablet	1811
acer aspire notebook	8930
asus chromebook	9542
samsung galaxy note	2677
mac mini	4454
hp pavilion desktop	8881
ipad air	9469
htc one	4276
dell inspiron desktop	10141
amazon fire phone	2168
acer aspire desktop	5173
nokia lumia 635	5612

➢ <u>Weekly Retention Analysis:</u> The retention of users on a weekly basis based on their sign-up cohort was:

```
SELECT
    WEEK(STR_TO_DATE(u.activated_at, '%d-%m-%Y')) AS weeknum,
    COUNT(u.user_id) AS weekly_user_count
FROM
    users_case2 u
        LEFT JOIN
    events e ON u.user_id = e.user_id
WHERE
    event_type = 'engagement'
GROUP BY weeknum
ORDER BY weeknum
```

weeknum	weekly_user_count
0	2361
1	3922
2	5166
3	4952
4	5481
5	6286
6	5334
7	6477
8	3659
9	5778
10	5325
11	6471
12	6083
13	7063
14	7626
15	6773
16	9369
17	11855
18	10394
19	10544
20	10996
21	9429
22	11255
23	10849
24	10616

➢ <u>Weekly Engagement Per Device:</u> The activeness of users on a weekly basis per device was:

```
SELECT
    weekly_count.device AS device,
    SUM(weekly_count.weekly_user_count) AS weekly_user_count
FROM
    (SELECT
        WEEK(STR_TO_DATE(u.activated_at, '%d-%m-%Y')) AS weeknum,
            e.device AS device,
            COUNT(u.user_id) AS weekly_user_count
    FROM
        users_case2 u
    LEFT JOIN events e ON u.user_id = e.user_id
    WHERE
        e.device IS NOT NULL
    GROUP BY weeknum , e.device) AS weekly_count
GROUP BY weekly_count.device
ORDER BY weekly_count.device ASC
```

device	weekly_user_count
acer aspire desktop	5173
acer aspire notebook	8930
amazon fire phone	2168
asus chromebook	9542
dell inspiron desktop	10141
dell inspiron notebook	19669
hp pavilion desktop	8881
htc one	4276
ipad air	9469
ipad mini	5591
iphone 4s	9615
iphone 5	25883
iphone 5s	15929
kindle fire	4090
lenovo thinkpad	36978
mac mini	4454
macbook air	26786
macbook pro	57295
nexus 10	5139
nexus 5	16502
nexus 7	6540
nokia lumia 635	5612
samsumg galaxy tablet	1811
samsung galaxy note	2677
samsung galaxy s4	18653

- Email Engagement Analysis: The users engaging with the email service were:

```
SELECT action, COUNT(u.user_id) AS user_count
FROM users_case2 u
LEFT JOIN email_events e ON u.user_id = e.user_id
WHERE action IS NOT NULL
GROUP BY action
```

action	user_count
sent_weekly_digest	57267
email_open	20459
email_clickthrough	9010
sent_reengagement_email	3653

v) Conclusion

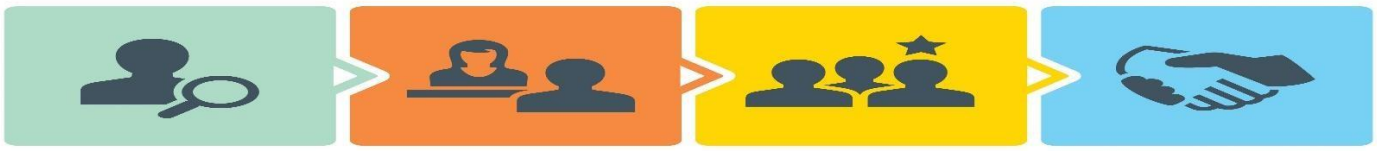
By working on this project, I made myself acquainted with MySQL and various business concepts.

I was able to learn some advanced SQL queries like window functions and some technical indicators like moving/rolling average and throughput which helped me gain domain knowledge.

I got hands-on experience on the role of a data analyst with the help of this project.

I was able to generate valuable insights and was able to answer the questions posed by the team.

This project significantly enhanced my analytical abilities and helped me to prepare for future challenges in data analysis.



Hiring Process Analytics

i) Description

Imagine you're a data analyst at a multinational company like Google. Your task is to analyze the company's hiring process data and draw meaningful insights from it.

The hiring process is a crucial function of any company, and understanding trends such as the number of rejections, interviews, job types, and vacancies can provide valuable insights for the hiring department.

As a data analyst, you'll be given a dataset containing records of previous hires. Your job is to analyze this data and answer certain questions that can help the company improve its hiring process.

The goal of this project is to use your knowledge of statistics and Excel to draw meaningful conclusions about the company's hiring process

ii) The Problem

Data Analytics Tasks

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

- 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.

- 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.

- Departmental Analysis: Visualizing data through charts and plots is a crucial part of data analysis. Show the proportion of people working in different departments.

- Position Tier Analysis: Different positions within a company often have different tiers or levels. Represent the different position tiers within the company.

iii) Design

Steps taken to clean the data

- First downloaded the dataset and opened it using Microsoft Excel.
- Converted the raw data into a table and removed duplicate rows.
- Removed special characters for better understandability.
- Removed the missing values in categorical columns with its mode.
- Checked for potential outliers and removed them.

Software Used: Microsoft Excel Professional

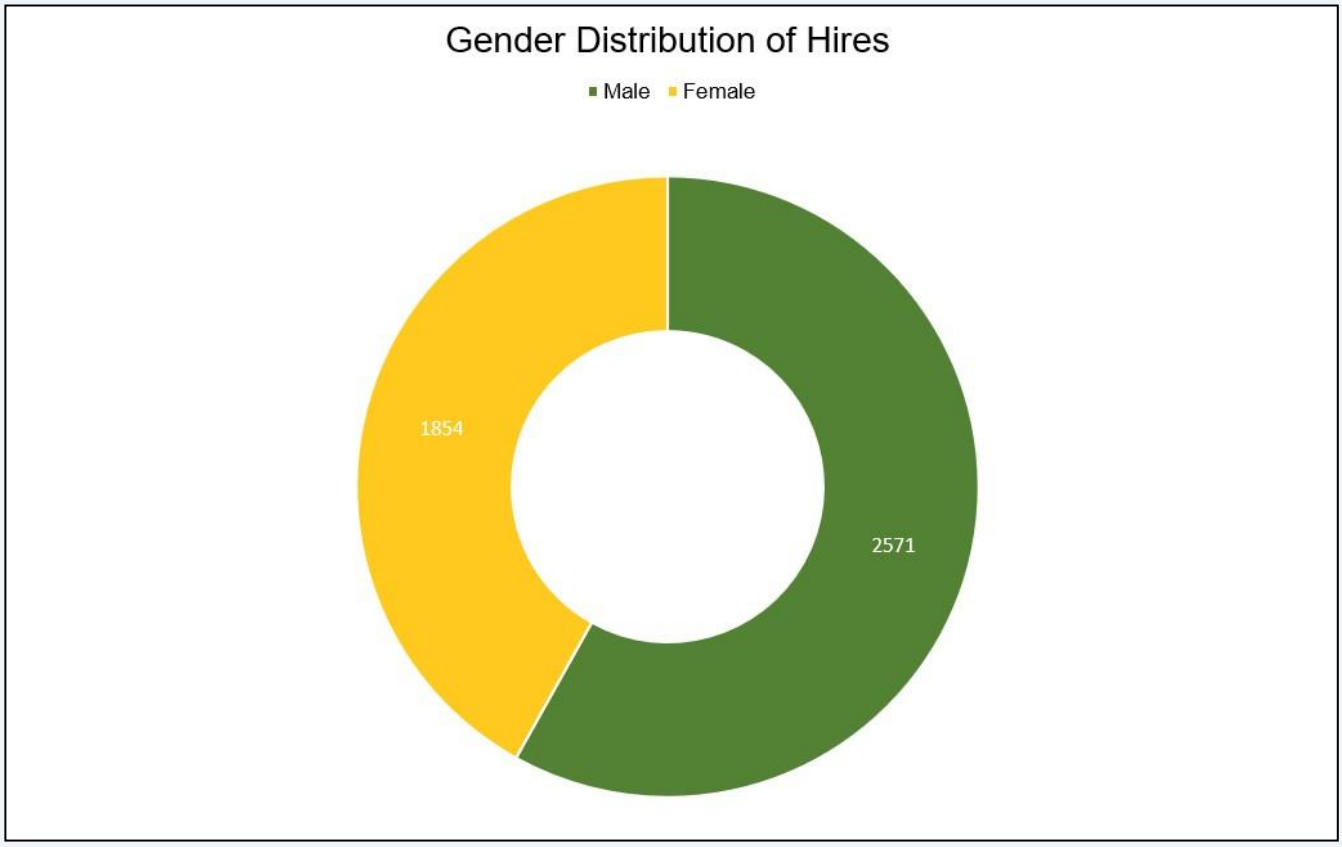
Version: 2021

iv) Insights

Data Analytics Tasks

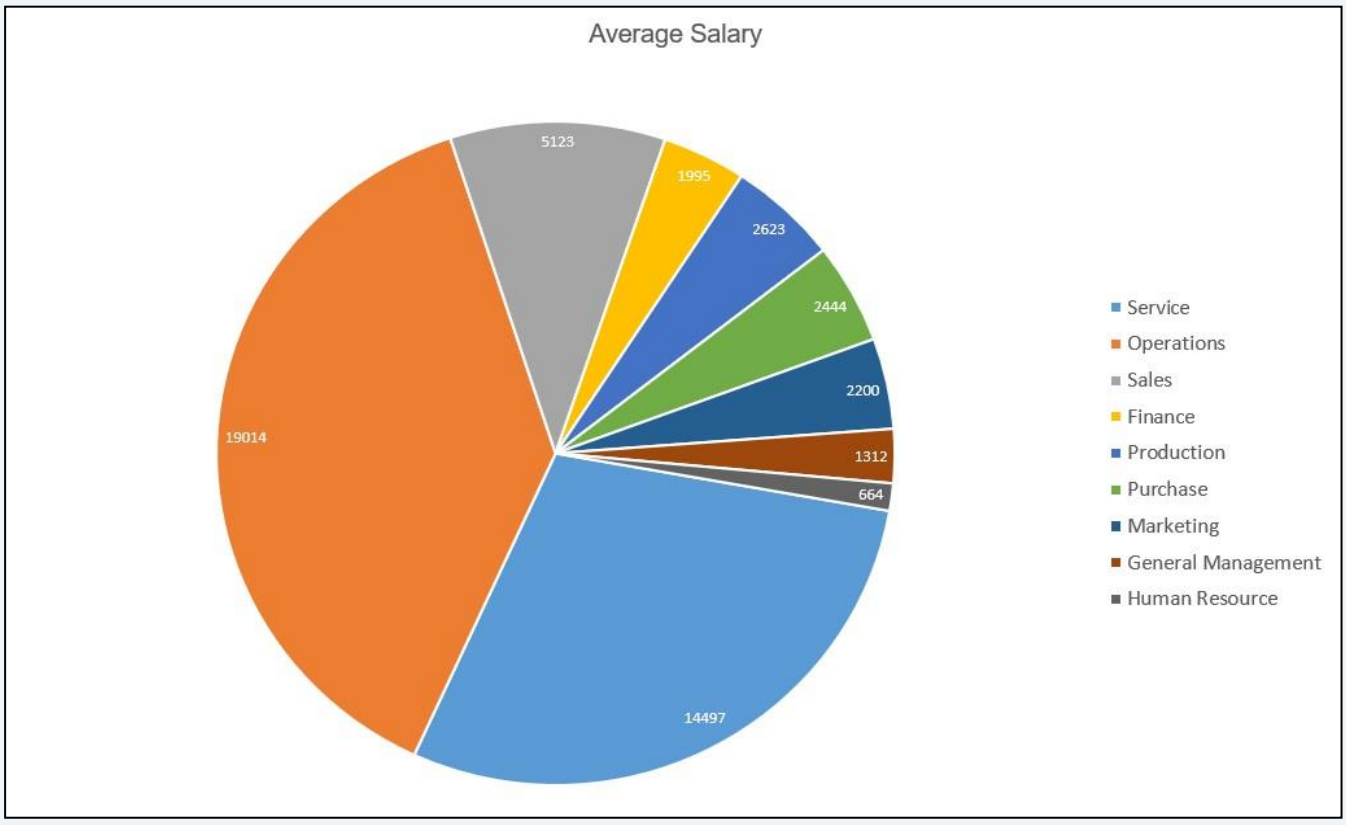
- <u>Hiring Analysis:</u> The gender distribution of hires i.e. the number of males and females hired by the company were:

Male	Female
2571	1854



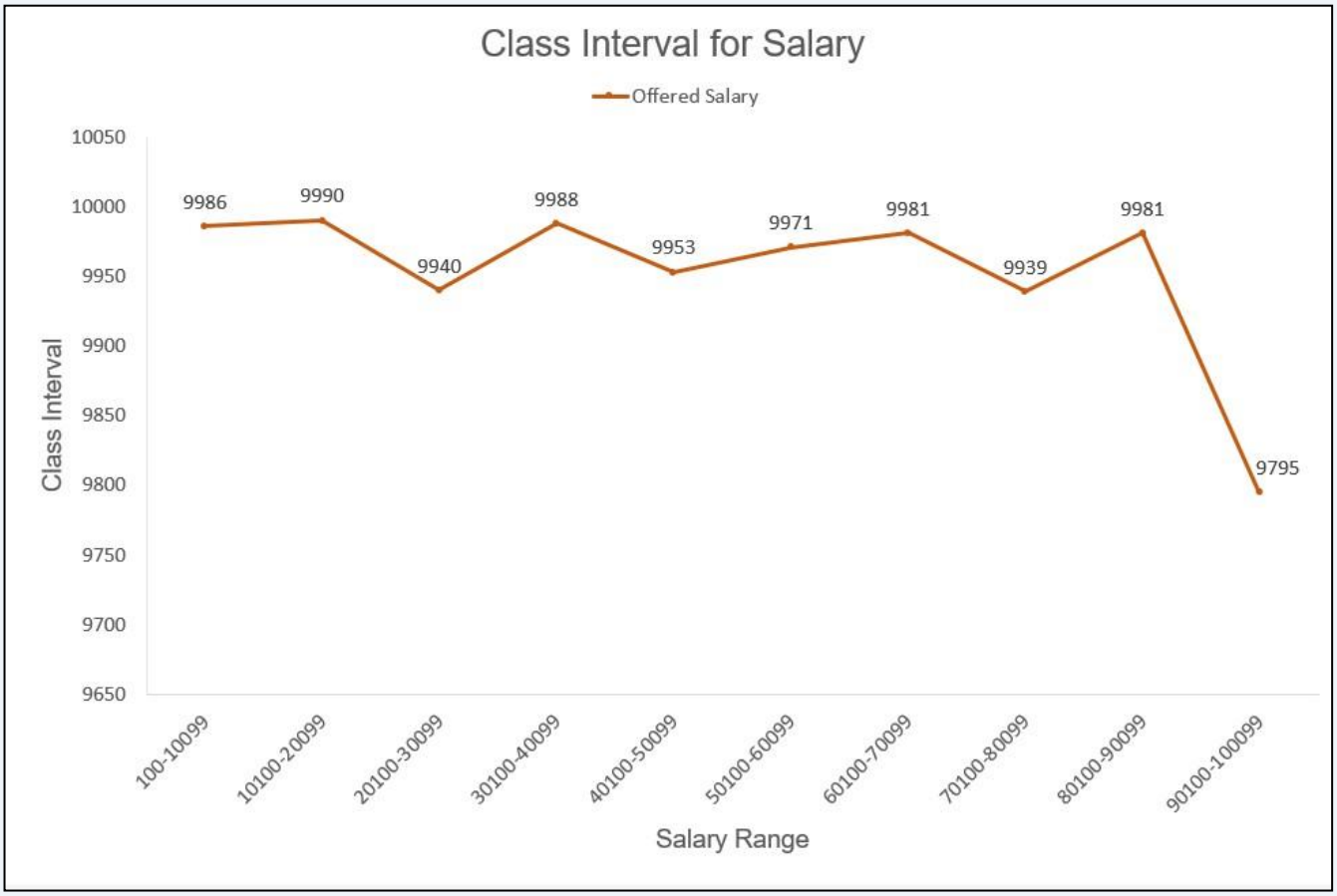
➢ <u>Salary Analysis:</u> The average salaries for a group of employees were:

Department	Average
Service	14497
Operations	19014
Sales	5123
Finance	1995
Production	2623
Purchase	2444
Marketing	2200
General Management	1312
Human Resource	664
Total	**49873**



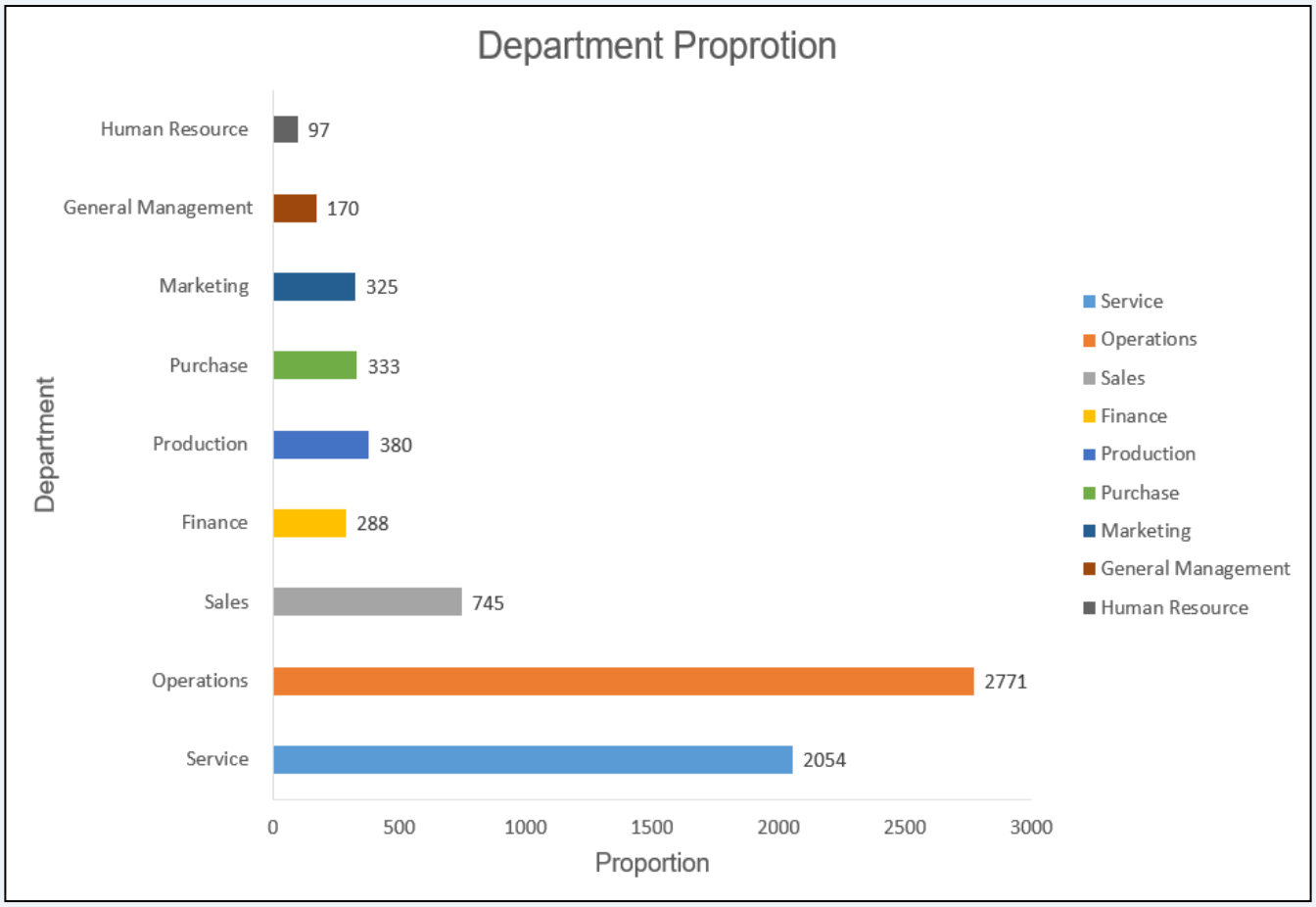
➢ <u>Salary Distribution:</u> The class intervals for the salaries in the company were:

Salary Range	Max	Min	Class Interval
100-10099	10086	100	9986
10100-20099	20095	10105	9990
20100-30099	30070	20130	9940
30100-40099	40088	30100	9988
40100-50099	50062	40109	9953
50100-60099	60096	50125	9971
60100-70099	70096	60115	9981
70100-80099	80076	70137	9939
80100-90099	90092	80111	9981
90100-100099	99967	90172	9795



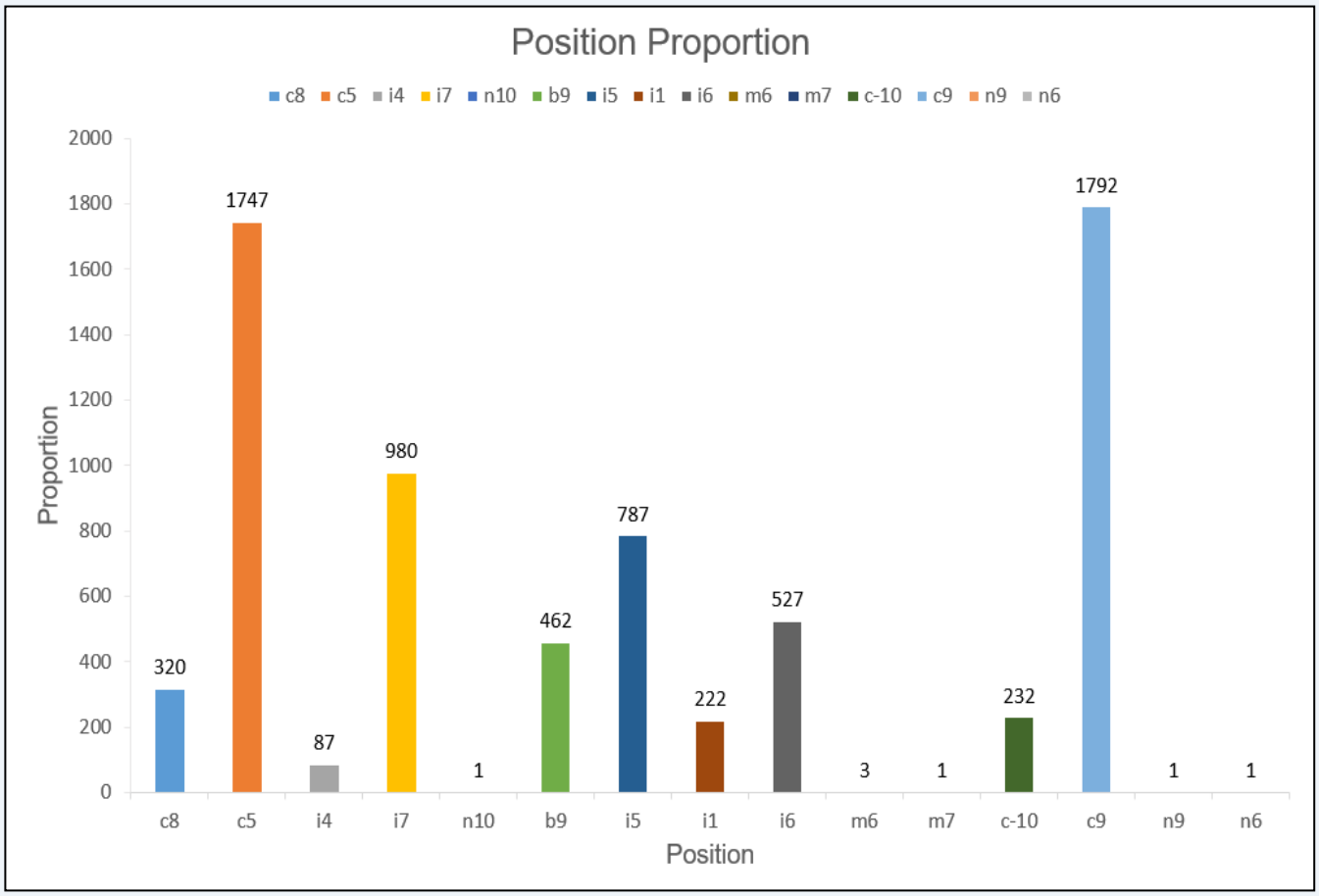
➢ <u>Departmental Analysis:</u> The proportion of people working in different departments was:

Department	Proportion
Service	2054
Operations	2771
Sales	745
Finance	288
Production	380
Purchase	333
Marketing	325
General Management	170
Human Resource	97



➢ <u>Position Tier Analysis:</u> The proportion of different position tiers within the company is:

Position	Proportion
c8	320
c5	1747
i4	87
i7	980
n10	1
b9	462
i5	787
i1	222
i6	527
m6	3
m7	1
c-10	232
c9	1792
n9	1
n6	1



v) Conclusion

By working on this project, I understood the importance of Microsoft Excel as a data analyst.

I was able to learn some advanced Excel commands like lookup functions, date and time functions, text functions, etc.

I also learned to visualize the data using various charts provided by MS Excel.

Additionally, I learned the entire process of transforming the raw data into meaningful information, also known as ETL (Extract Transform Load).

I got a brief overview of how hiring process actually happens for a company and what are the key factors for it.

Overall, this project allowed me to enhance my Excel proficiency and gain practical experience in generating meaningful insights from data.

Dataset Link: Hiring Process Analytics

IMDb Movie Analysis

i) Description

The dataset provided is related to IMDB Movies. A potential problem to investigate could be: "What factors influence the success of a movie on IMDB?" Here, success can be defined by high IMDB ratings.

The impact of this problem is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future project.

Remember, as a data analyst, your goal is not just to answer questions but to provide insights that can drive decision-making. Your analysis should aim to provide actionable insights that can help stakeholders make informed decisions.

ii) The Problem

Data Analytics Tasks

- Movie Genre Analysis: Analyze the distribution of movie genres and their impact on the IMDB score.
- Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.
- Language Analysis: Examine the distribution of movies based on their language.
- Director Analysis: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies.
- Budget Analysis: Explore the relationship between movie budgets and their financial success.

iii) Design

Steps taken to clean the data

- First downloaded the dataset and opened it using Microsoft Excel.
- Converted the raw data into a table and removed duplicate rows.
- Removed the missing values in categorical columns with its mode and numerical columns with its median with the help of histograms.
- Removed special characters for better understandability.
- Created heatmap to understand the correlation between the variables.

Software Used: Microsoft Excel Professional

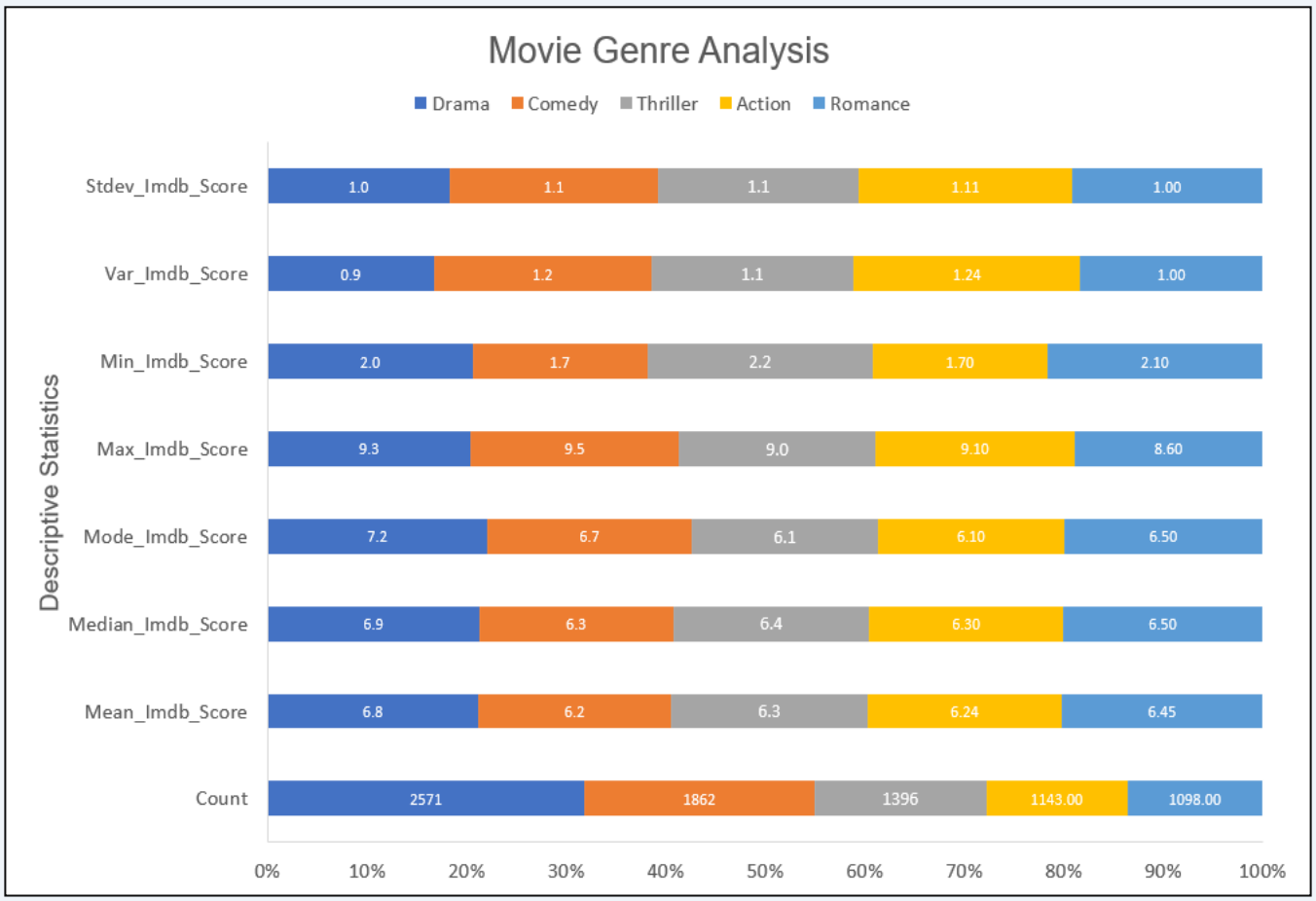
Version: 2021

iv) Insights

Data Analytics Tasks

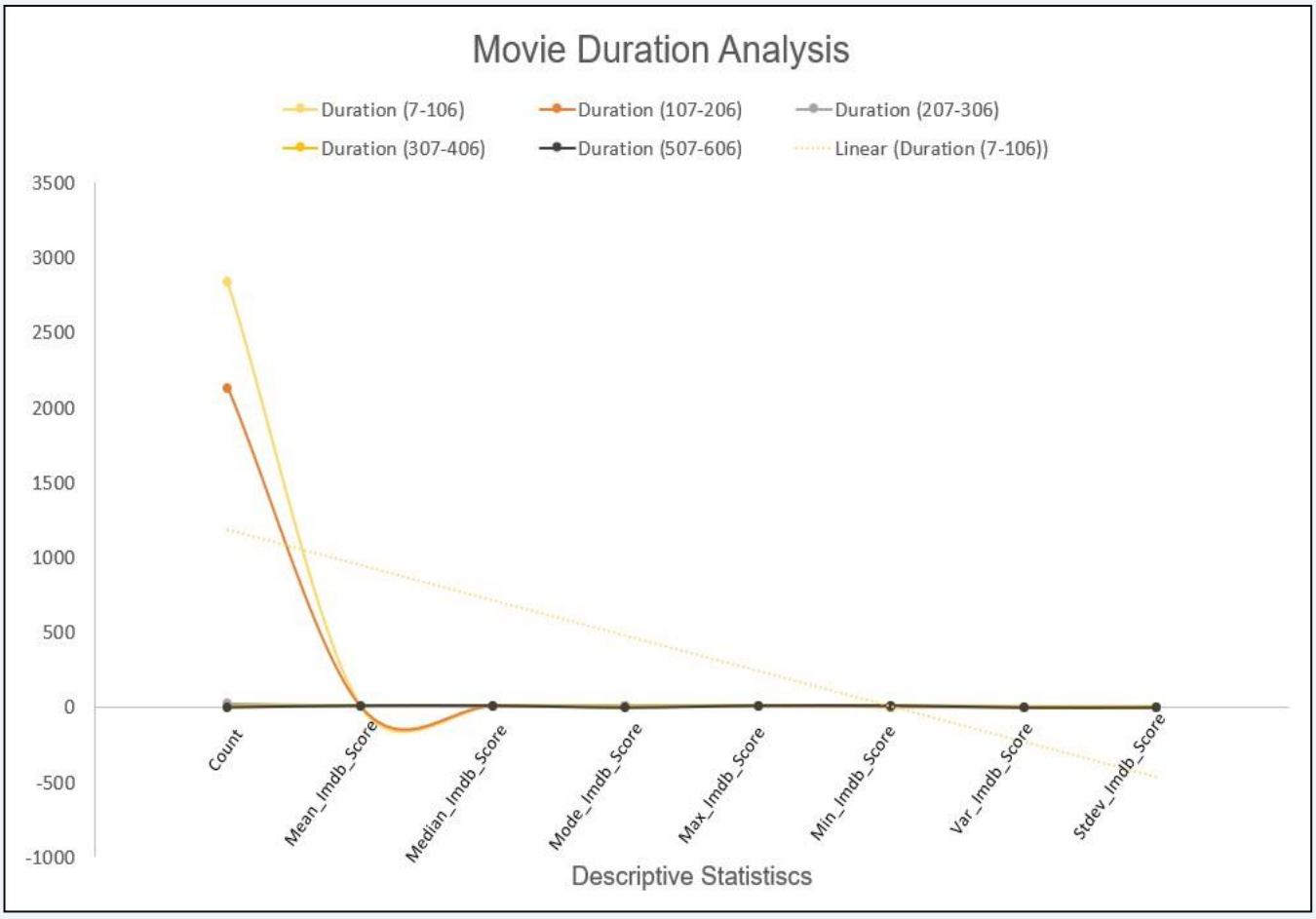
- Movie Genre Analysis: The distribution of movie genres and their impact on the IMDB score was:

	Drama	Comedy	Thriller	Action	Romance
Count	2571	1862	1396	1143.00	1098.00
Mean_Imdb_Score	6.8	6.2	6.3	6.24	6.45
Median_Imdb_Score	6.9	6.3	6.4	6.30	6.50
Mode_Imdb_Score	7.2	6.7	6.1	6.10	6.50
Max_Imdb_Score	9.3	9.5	9.0	9.10	8.60
Min_Imdb_Score	2.0	1.7	2.2	1.70	2.10
Var_Imdb_Score	0.9	1.2	1.1	1.24	1.00
Stdev_Imdb_Score	1.0	1.1	1.1	1.11	1.00



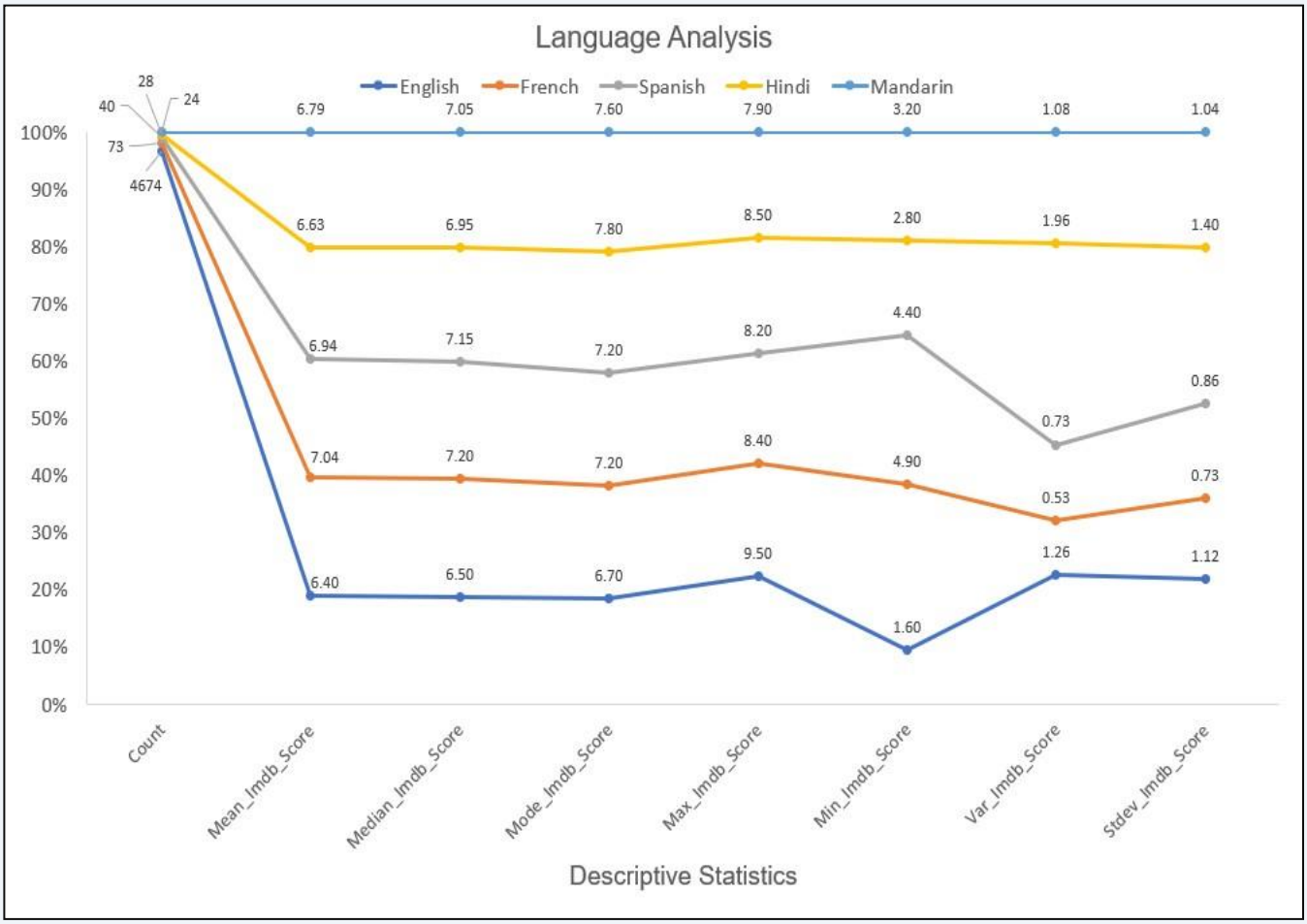
➢ <u>Movie Duration Analysis:</u> The distribution of movie durations and its impact on the IMDB score was:

	Duration (7-106)	Duration (107-206)	Duration (207-306)	Duration (307-406)	Duration (507-606)
Count	2837	2132	25	3	1
Mean_Imdb_Score	6.15	6.82	7.61	7.50	8.20
Median_Imdb_Score	6.30	6.90	7.80	7.70	8.20
Mode_Imdb_Score	6.30	6.70	6.60	0.00	0.00
Max_Imdb_Score	9.50	9.30	9.00	8.00	8.20
Min_Imdb_Score	1.70	1.60	5.80	6.80	8.20
Var_Imdb_Score	1.37	0.86	0.67	0.39	0.01
Stdev_Imdb_Score	1.17	0.92	0.82	0.62	0.12



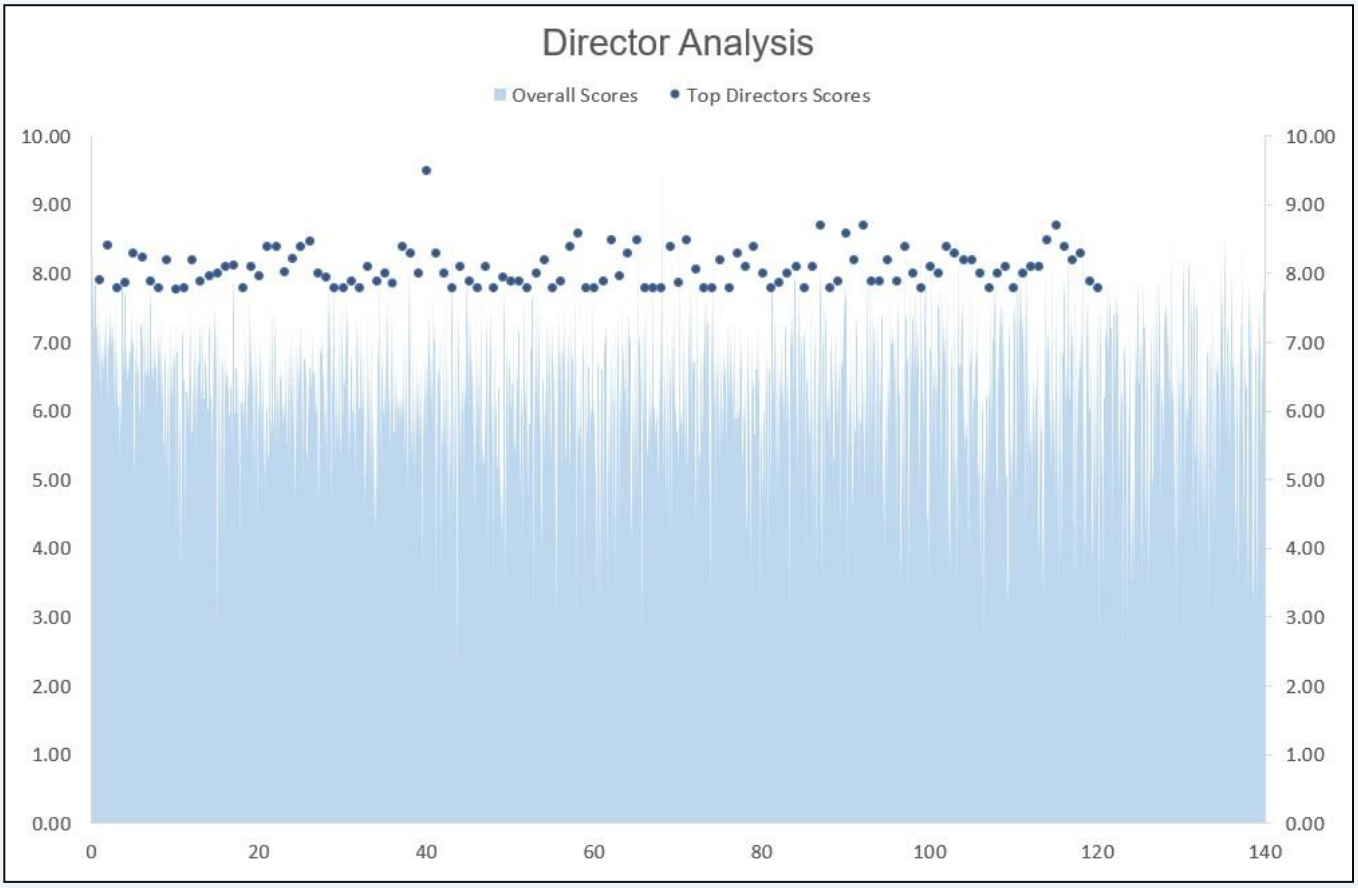
➢ <u>Language Analysis:</u> The distribution of movies based on their language was:

	English	French	Spanish	Hindi	Mandarin
Count	4674	73	40	28	24
Mean_Imdb_Score	6.40	7.04	6.94	6.63	6.79
Median_Imdb_Score	6.50	7.20	7.15	6.95	7.05
Mode_Imdb_Score	6.70	7.20	7.20	7.80	7.60
Max_Imdb_Score	9.50	8.40	8.20	8.50	7.90
Min_Imdb_Score	1.60	4.90	4.40	2.80	3.20
Var_Imdb_Score	1.26	0.53	0.73	1.96	1.08
Stdev_Imdb_Score	1.12	0.73	0.86	1.40	1.04



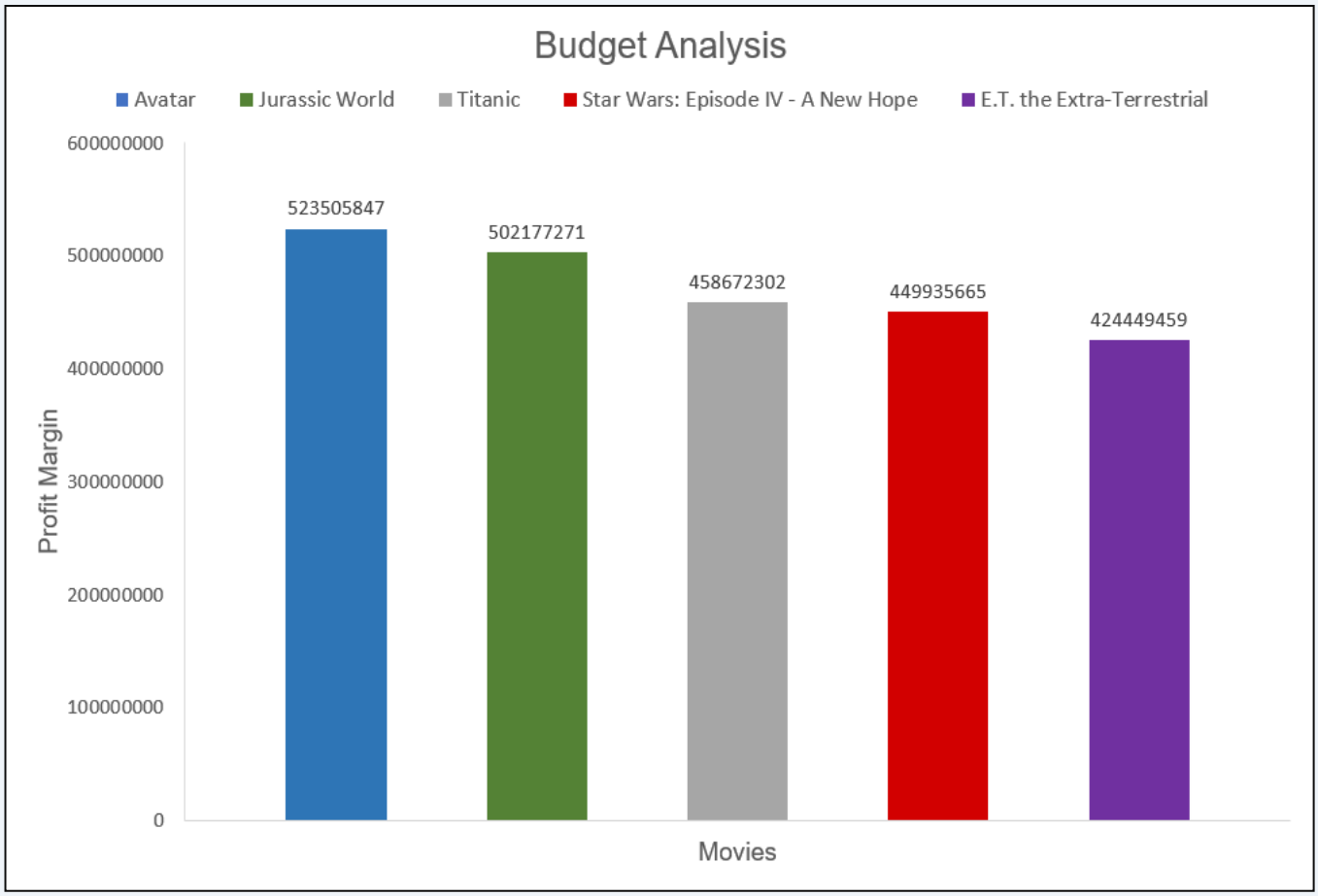
- Director Analysis: These were the top directors based on their average IMDB score and their contribution to the success of movies.

Director Name	Average_Imdb_Score	Percentile Threshold	Top Directors	Top Directors Name	Top Directors Score
James Cameron	7.91	7.77	Top	James Cameron	7.91
Gore Verbinski	6.99		Not top	Christopher Nolan	8.43
Sam Mendes	7.50		Not top	Nathan Greno	7.80
Christopher Nolan	8.43		Top	Joss Whedon	7.87
Doug Walker	7.10		Not top	Lee Unkrich	8.30
Andrew Stanton	7.73		Not top	Pete Docter	8.23
Sam Raimi	6.91		Not top	Don Hall	7.90
Nathan Greno	7.80		Top	Rich Moore	7.80
Joss Whedon	7.87		Top	Hideaki Anno	8.20
David Yates	7.20		Not top	Alejandro G. IÃ±Ã¡rritu	7.78
Zack Snyder	7.18		Not top	Alfonso CuarÃ³n	7.80
Bryan Singer	7.29		Not top	Quentin Tarantino	8.20
Marc Forster	7.15		Not top	Jacques Perrin	7.90
Andrew Adamson	7.08		Not top	Frank Darabont	7.98
Rob Marshall	6.60		Not top	Stanley Kubrick	8.00
Barry Sonnenfeld	6.46		Not top	Tim Miller	8.10
Peter Jackson	7.68		Not top	Milos Forman	8.13
Marc Webb	7.13		Not top	Deepa Mehta	7.80
Ridley Scott	7.07		Not top	Andrei Tarkovsky	8.10
Chris Weitz	6.08		Not top	Denis Villeneuve	7.97
Anthony Russo	7.00		Not top	S.S. Rajamouli	8.40
Peter Berg	6.67		Not top	Moustapha Akkad	8.40
Colin Trevorrow	7.00		Not top	Tony Kaye	8.03
Shane Black	7.40		Not top	Hayao Miyazaki	8.23
Tim Burton	6.93		Not top	Richard Marquand	8.40
Brett Ratner	6.46		Not top	Sergio Leone	8.48
Dan Scanlon	7.30		Not top	David Lean	8.00
Michael Bay	6.64		Not top	Bernardo Bertolucci	7.95
Joseph Kosinski	7.18		Not top	Giuseppe Tornatore	7.80



➢ <u>Budget Analysis:</u> The relationship between movie budgets and their financial success was:

Movie	Movie Budget	Movie Gross	Correlation	Profit Margin	Top 5 Movies Profit Margin	Top 5 Highest Earning Movies
Avatar	237000000	760505847	0.239286948	523505847	523505847	Avatar
Pirates of the Caribbean: At World's End	300000000	309404152		9404152	502177271	Jurassic World
Spectre	245000000	200074175		-44925825	458672302	Titanic
The Dark Knight Rises	250000000	448130642		198130642	449935665	Star Wars: Episode IV - A New Hope
Star Wars: Episode VII - The Force Awakens	20000000	25445749		5445749	424449459	E.T. the Extra-Terrestrial
John Carter	263700000	73058679		-190641321		
Spider-Man 3	258000000	336530303		78530303		
Tangled	260000000	200807262		-59192738		
Avengers: Age of Ultron	250000000	458991599		208991599		
Harry Potter and the Half-Blood Prince	250000000	301956980		51956980		
Batman v Superman: Dawn of Justice	250000000	330249062		80249062		
Superman Returns	209000000	200069408		-8930592		
Quantum of Solace	200000000	168368427		-31631573		
Pirates of the Caribbean: Dead Man's Chest	225000000	423032628		198032628		
The Lone Ranger	215000000	89289910		-125710090		
Man of Steel	225000000	291021565		66021565		
The Chronicles of Narnia: Prince Caspian	225000000	141614023		-83385977		
The Avengers	220000000	623279547		403279547		
Pirates of the Caribbean: On Stranger Tides	250000000	241063875		-8936125		
Men in Black 3	225000000	179020854		-45979146		
The Hobbit: The Battle of the Five Armies	250000000	255108370		5108370		
The Amazing Spider-Man	230000000	262030663		32030663		
Robin Hood	200000000	105219735		-94780265		
The Hobbit: The Desolation of Smaug	225000000	258355354		33355354		
The Golden Compass	180000000	70083519		-109916481		
King Kong	207000000	218051260		11051260		
Titanic	200000000	658672302		458672302		
Captain America: Civil War	250000000	407197282		157197282		
Battleship	209000000	65173160		-143826840		



v) Conclusion

I was able to learn some advanced Excel commands like percentile, Correl, index matching, filter, etc.

I also learned to visualize the data using various charts provided by MS Excel.

Additionally, I learned the entire process of transforming the raw data into meaningful information, also known as ETL (Extract Transform Load).

I got an idea of how IMDB rates the movies and what are the key factors for its success.

I also understood the importance of descriptive statistics for analyzing the data.

Dataset Link: IMDb Movie Analysis

Bank Loan Case Study

i) Description

Imagine you're a data analyst at a finance company that specializes in lending various types of loans to urban customers. Your company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans.

The dataset you'll be working with contains information about loan applications. It includes two types of scenarios:

1) Customers with payment difficulties: These are customers who had a late payment of more than X days on at least one of the first Y installments of the loan.

2) All other cases: These are cases where the payment was made on time.

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments.

ii) The Problem

Data Analytics Tasks

- Identify Missing Data and Handle Accordingly: As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

- Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

- Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

- Perform Univariate, Segmented Univariate and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

- Identify Top Correlations for Different Scenarios: Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

iii) Design

Steps taken to clean the data

- First downloaded the dataset and opened it using Microsoft Excel.
- Converted the raw data into a table and removed duplicate rows.
- Removed the missing values in categorical columns with its mode and numerical columns with its median with the help of histograms.
- Removed special characters for better understandability.

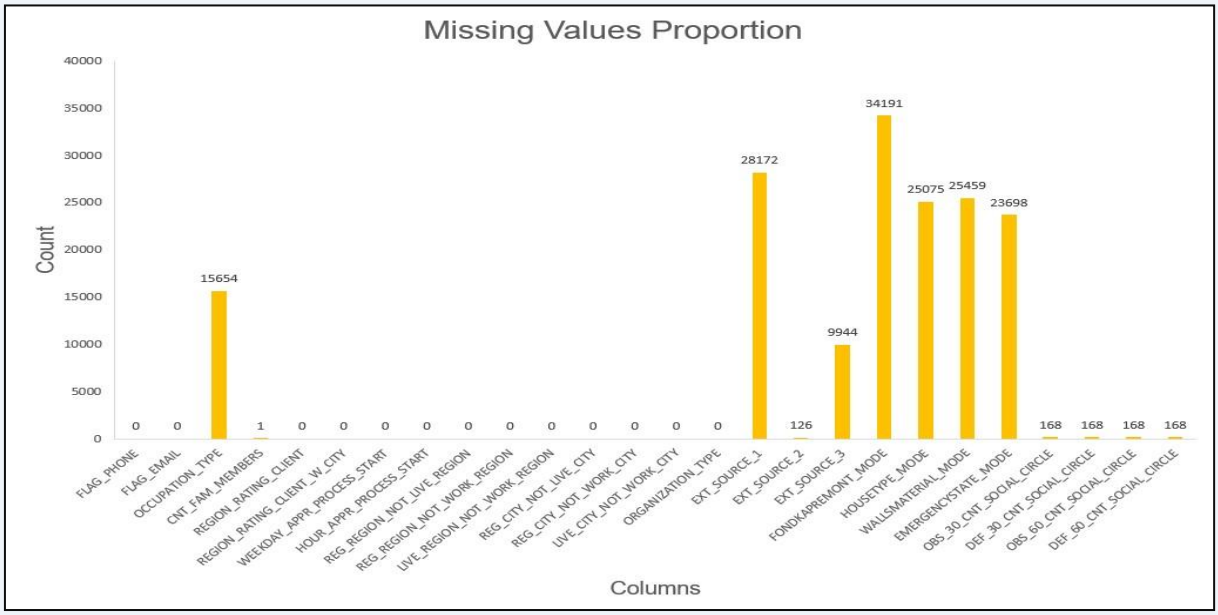
Software Used: Microsoft Excel Professional

Version: 2021

iv) Insights

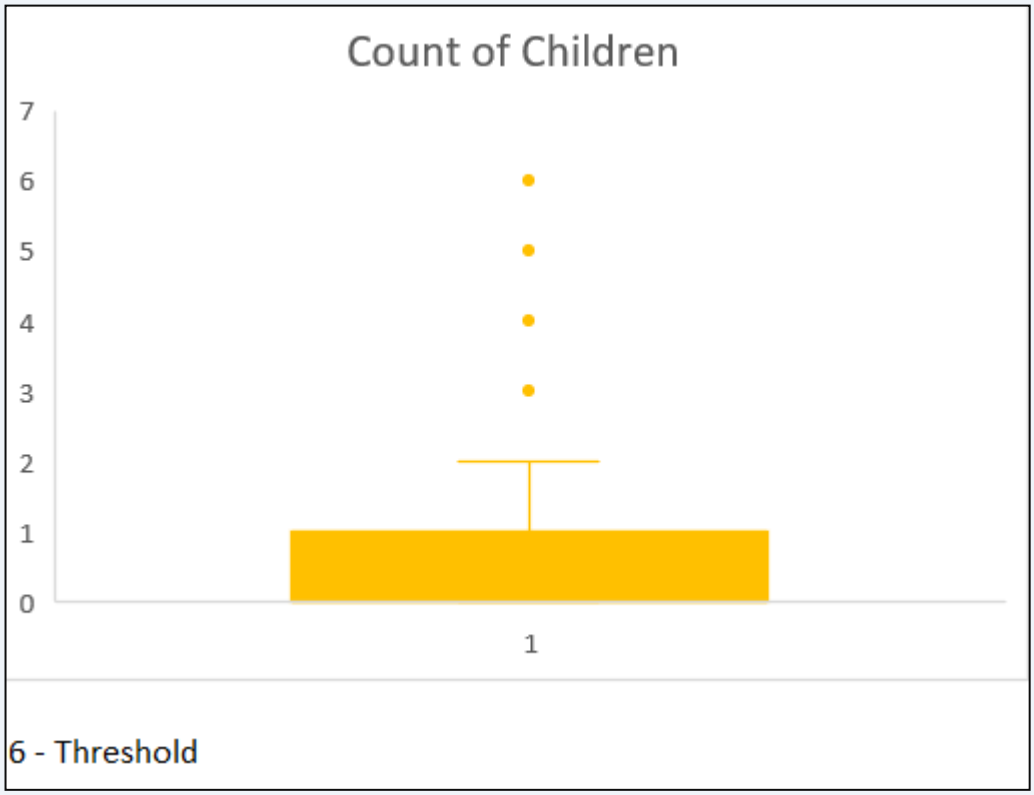
Data Analytics Tasks

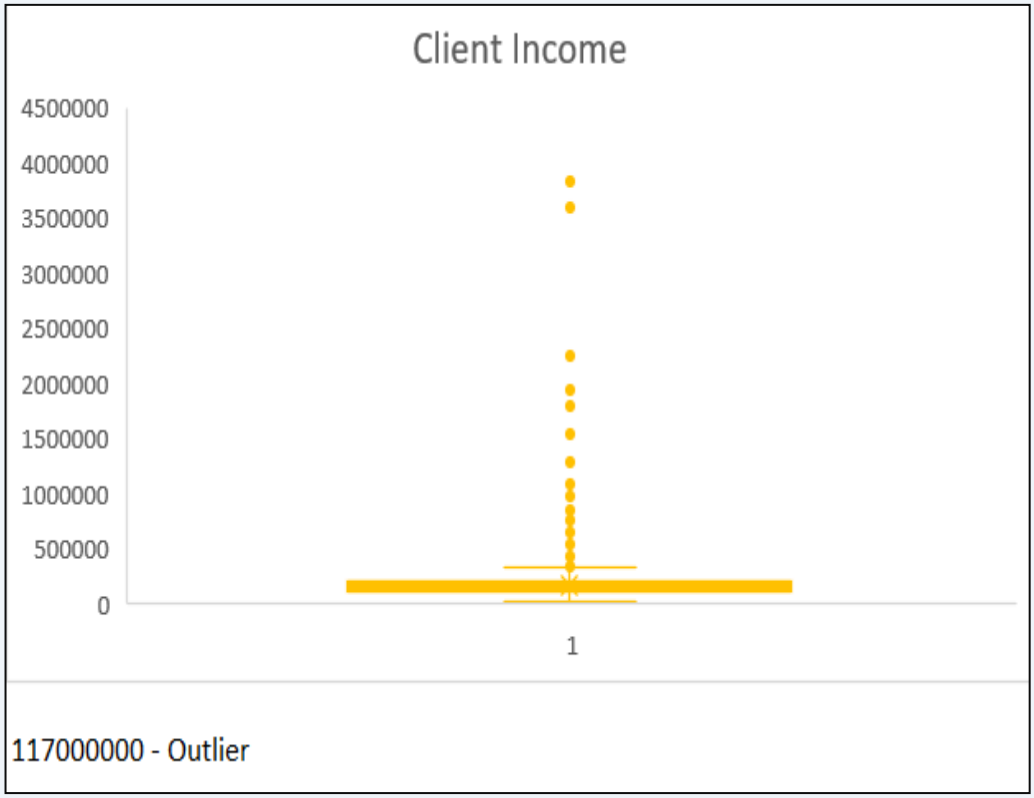
- Identify Missing Data and Handle Accordingly: I created column charts to show the proportion of missing values for each column. I then created histograms for the numerical columns for deciding the best method to fill the missing values i.e. mean, median or mode. I filled the categorical columns with their mode.



SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family
100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children
100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied
100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied
100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied
100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied
100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family
100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied
100021	0	Revolving loans	F	N	Y	1	81000	270000	13500	270000	Unaccompanied
100022	0	Revolving loans	F	N	Y	0	112500	157500	7875	157500	Other_A
100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied
100024	0	Revolving loans	M	Y	Y	0	135000	427500	21375	427500	Unaccompanied
100025	0	Cash loans	F	Y	Y	1	202500	1132573.5	37561.5	927000	Unaccompanied
100026	0	Cash loans	F	N	N	1	450000	497520	32521.5	450000	Unaccompanied
100027	0	Cash loans	F	N	Y	0	83250	239850	23850	225000	Unaccompanied
100029	0	Cash loans	M	Y	N	2	135000	247500	12703.5	247500	Unaccompanied
100030	0	Cash loans	F	N	Y	0	90000	225000	11074.5	225000	Unaccompanied

- <u>Identify Outliers in the Dataset:</u> I created box plots for identifying the outliers. I did not remove the outliers for some numerical columns as some outliers contained valuable information. Instead, I applied threshold values and business rules to identify the outliers.





- Analyze Data Imbalance: I calculated proportions for each unique value in the variable with the help of UNIQUE, COUNTIFS and SUM functions of Microsoft Excel. I also created bar charts to visualize the data imbalance of variables with respect to the target variable.

NAME_INCOME_TYPE	Target - 1	Target - 0	Proportion (Target - 1)	Proportion (Target - 0)
Working	2455	23534	61.08%	51.23%
State servant	198	3310	4.93%	14.77%
Commercial associate	863	10675	21.47%	55.90%
Pensioner	501	8411	12.47%	99.86%
Unemployed	2	4	0.05%	33.33%
Student	0	5	0.00%	62.50%
Businessman	0	2	0.00%	66.67%
Maternity leave	0	1	0.00%	100.00%

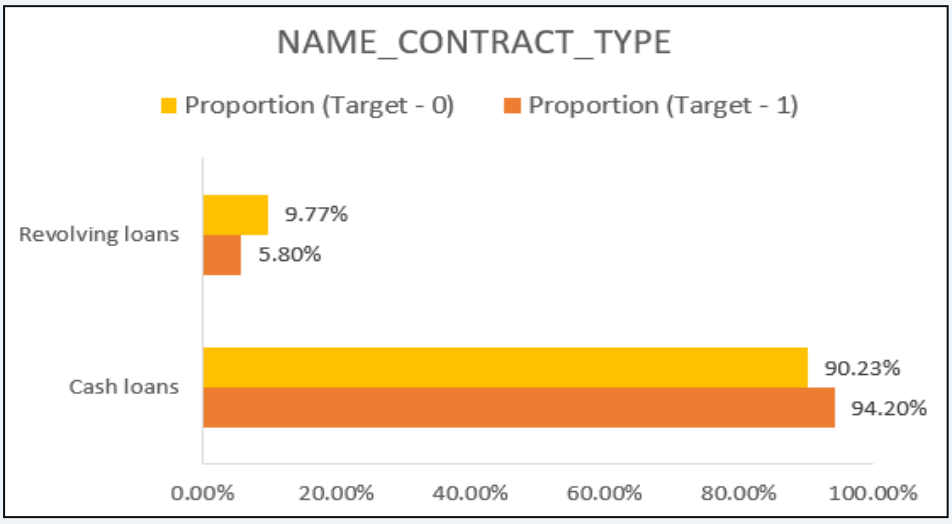
NAME_EDUCATION_TYPE	Target - 1	Target - 0	Proportion (Target - 1)	Proportion (Target - 0)
Secondary / secondary special	3204	32339	79.72%	70.39%
Higher education	605	11555	15.05%	25.15%
Incomplete higher	137	1481	3.41%	3.22%
Lower secondary	73	547	1.82%	1.19%
Academic degree	0	20	0.00%	0.04%

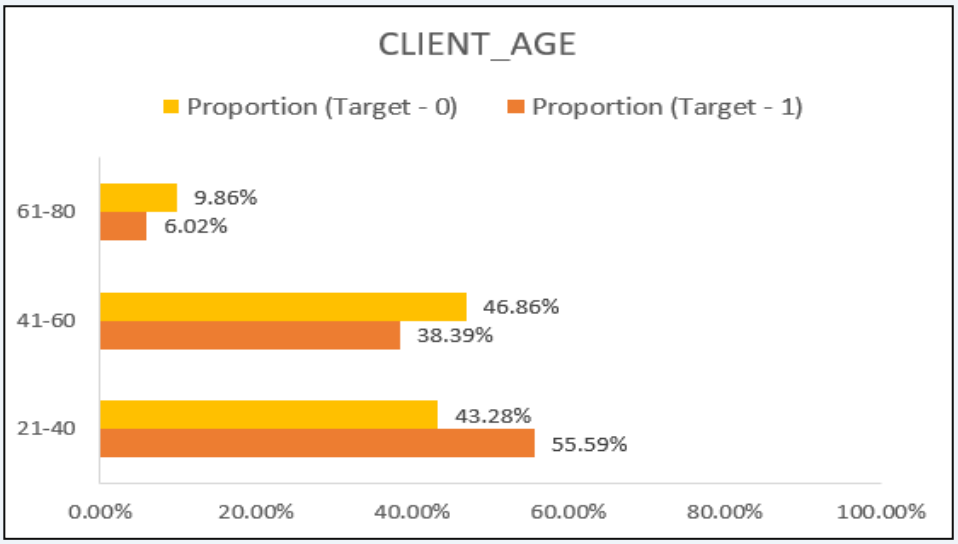
CLIENT_AGE	Target - 1	Target - 0	Proportion (Target - 1)	Proportion (Target - 0)
21-40	2234	19883	55.59%	43.28%
41-60	1543	21528	38.39%	46.86%
61-80	242	4531	6.02%	9.86%

DAYS_EMPLOYED	Target - 1	Target - 0	Proportion (Target - 1)	Proportion (Target - 0)
(-17531 - 82468)	3516	37529	87.48%	81.69%
(282469 - 382468)	503	8413	12.52%	18.31%

FLAG_MOBILE	Target - 1	Target - 0	Proportion (Target - 1)	Proportion (Target - 0)
1	4019	45941	100.00%	100.00%
0	0	1	0.00%	0.00%

FLAG_EMP_PHONE	Target - 1	Target - 0	Proportion (Target - 1)	Proportion (Target - 0)
1	3516	37527	87.48%	81.68%
0	503	8415	12.52%	18.32%

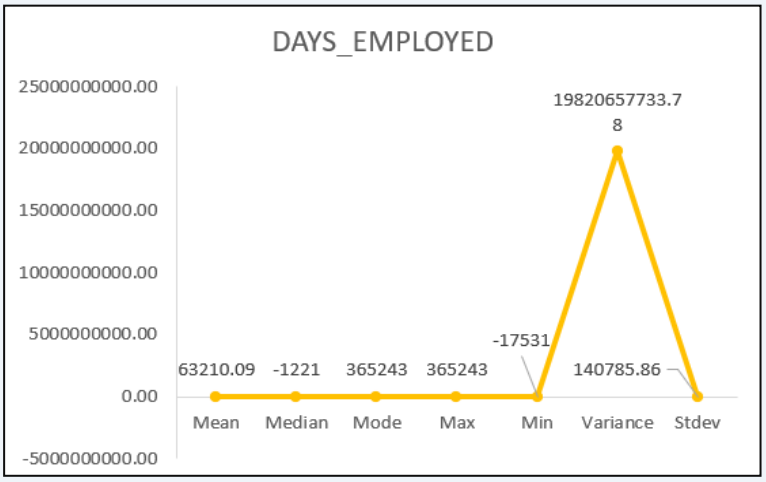


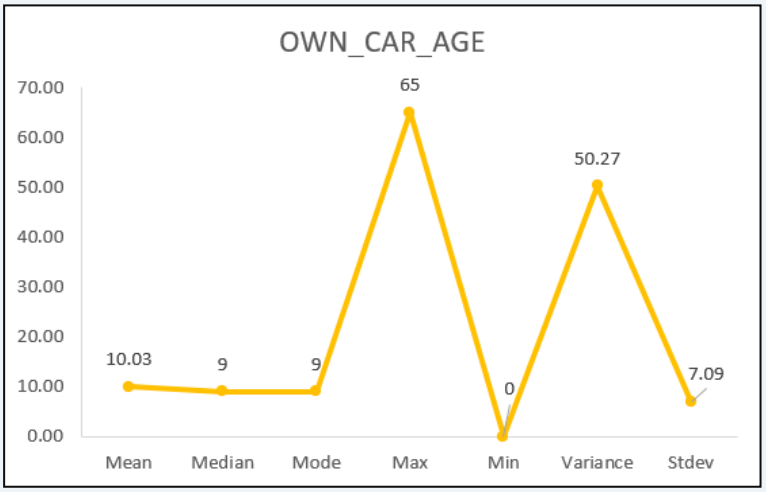


- <u>Perform Univariate, Segmented Univariate and Bivariate Analysis:</u> I performed univariate analysis for all the numerical columns and calculated their descriptive statistics like mean, median, mode, max, min, variance and standard deviation. I created line charts to visualize these analyses.

CNT_CHILDREN		AMT_INCOME_TOTAL		AMT_CREDIT		AMT_ANNUITY	
Mean	0.42	Mean	168422.52	Mean	599626.10	Mean	27106.25
Median	0	Median	144450	Median	514602	Median	24939
Mode	0	Mode	135000	Mode	450000	Mode	9000
Max	6	Max	3825000	Max	4050000	Max	258026
Min	0	Min	25650	Min	45000	Min	2052
Variance	0.52	Variance	9836785689.23	Variance	161943124118.08	Variance	212108545.64
Stdev	0.72	Stdev	99180.57	Stdev	402421.58	Stdev	14563.95

DAYS_EMPLOYED		DAYS_REGISTRATION		DAYS_ID_PUBLISH		OWN_CAR_AGE	
Mean	63210.09	Mean	-4978.04	Mean	-2996.94	Mean	10.03
Median	-1221	Median	-4491	Median	-3261	Median	9
Mode	365243	Mode	-3	Mode	-4360	Mode	9
Max	365243	Max	0	Max	0	Max	65
Min	-17531	Min	-22392	Min	-6232	Min	0
Variance	19820657733.78	Variance	12431626.59	Variance	2277560.85	Variance	50.27
Stdev	140785.86	Stdev	3525.85	Stdev	1509.16	Stdev	7.09





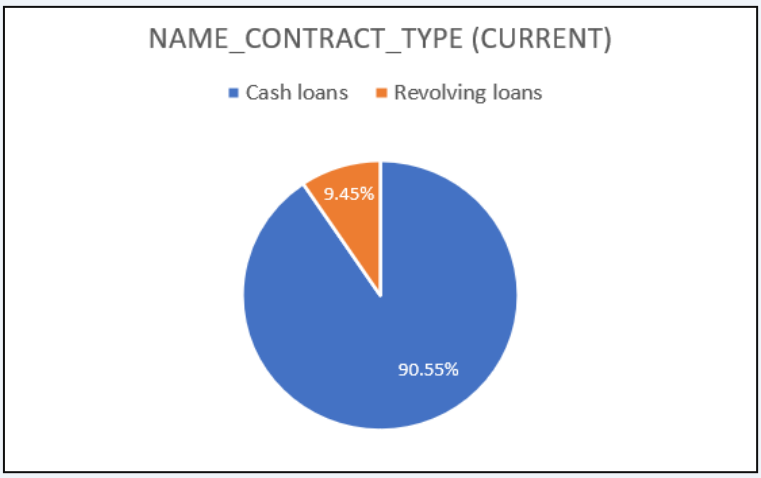
I then performed segmented univariate analysis for all the categorical columns and calculated the proportion of each unique value in the column. I also considered the previous application dataset for segmented univariate analysis. I then created pie charts to visualize these proportions.

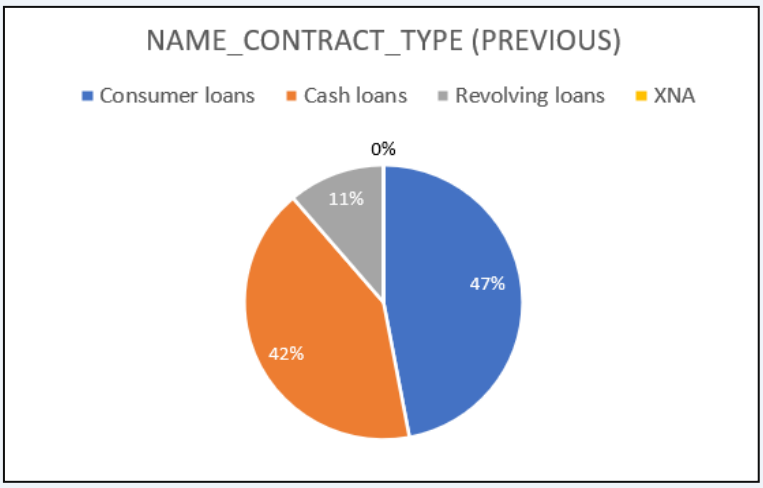
NAME_EDUCATION_TYPE	Count	Proportion
Secondary / secondary special	35543	71.14%
Higher education	12160	24.34%
Incomplete higher	1618	3.24%
Lower secondary	620	1.24%
Academic degree	20	0.04%

NAME_FAMILY_STATUS	Count	Proportion
Single / not married	7304	14.62%
Married	32066	64.18%
Civil marriage	4855	9.72%
Widow	2595	5.19%
Separated	3140	6.28%
Unknown	1	0.00%

NAME_HOUSING_TYPE	Count	Proportion
House / apartment	44335	88.74%
Rented apartment	768	1.54%
With parents	2397	4.80%
Municipal apartment	1843	3.69%
Office apartment	427	0.85%
Co-op apartment	191	0.38%

OCCUPATION_TYPE	Count	Proportion
Laborers	24588	49.21%
Core staff	4432	8.87%
Accountants	1620	3.24%
Managers	3487	6.98%
Drivers	3037	6.08%
Sales staff	5154	10.32%
Cleaning staff	738	1.48%
Cooking staff	963	1.93%
Private service staff	447	0.89%
Medicine staff	1402	2.81%
Security staff	1140	2.28%
High skill tech staff	1852	3.71%
Waiters/barmen staff	228	0.46%
Low-skill Laborers	357	0.71%
Realty agents	123	0.25%
Secretaries	212	0.42%
IT staff	80	0.16%
HR staff	101	0.20%





I then performed bivariate analysis to explore relationships between the variables and the target variable and created scatter charts to visualize these relationships.

CNT_CHILDREN		
	Target -1	Target - 0
Mean	0.48	0.41
Median	0	0
Mode	0	0
Max	5	6
Min	0	0
Variance	0.58	0.51
Stdev	0.76	0.72

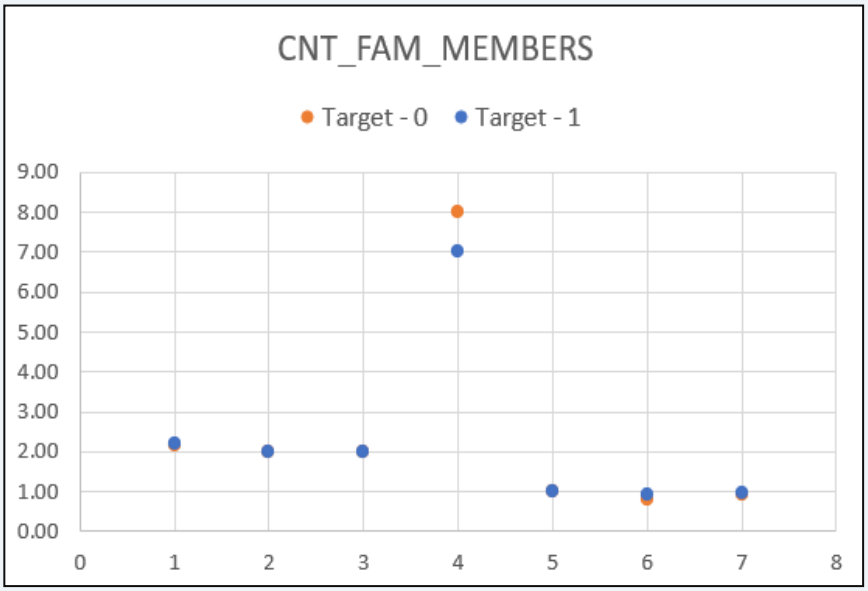
AMT_INCOME_TOTAL		
	Target - 1	Target - 0
Mean	161270.16	169048.21
Median	135000	148500
Mode	135000	135000
Max	1890000	3825000
Min	25650	25650
Variance	8081676990.82	9985455725.10
Stdev	89898.15	99927.25

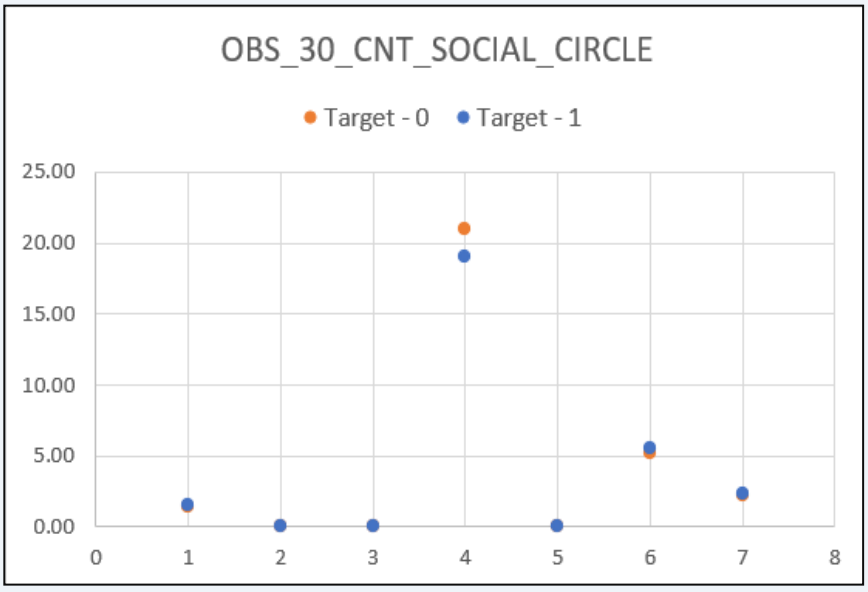
AMT_CREDIT		
	Target - 1	Target - 0
Mean	555549.31	603481.93
Median	497520	517923
Mode	450000	450000
Max	2961000	4050000
Min	45000	45000
Variance	115861767398.12	165789494929.43
Stdev	340384.73	407172.56

DAYS_EMPLOYED		
	Target - 1	Target - 0
Mean	44126.26	64879.54
Median	-1029	-1245
Mode	365243	365243
Max	365243	365243
Min	-14183	-17531
Variance	14754713368.64	20229179296.82
Stdev	121468.98	142229.32

DAYS_REGISTRATION		
	Target - 1	Target - 0
Mean	-4472.73	-5022.25
Median	-3969	-4543
Mode	-5011	-3
Max	0	0
Min	-17555	-22392
Variance	10723457.12	12556766.28
Stdev	3274.67	3543.55

DAYS_ID_PUBLISH		
	Target - 1	Target - 0
Mean	-2758.85	-3017.77
Median	-2838	-3297
Mode	-4033	-4360
Max	-2	0
Min	-6070	-6232
Variance	2321887.58	2268290.46
Stdev	1523.77	1506.08





- <u>Identify Top Correlations for Different Scenarios:</u> I segmented the dataset into 2 scenarios i.e. target – 1 (client with payment difficulties) and target – 0 (all other cases) with the help of FILTER function of MS Excel. I calculated the correlation for all the numerical columns of both the scenarios.

  I shaded the high correlations in yellow color and low correlations in green color with the help of conditional formatting which made it easier to indicate the top indicators of loan default for each scenario.

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	CLIENT_AGE
CNT_CHILDREN	1	-0.010	0.003	0.026	-0.005	-0.016	-0.259
AMT_INCOME_TOTAL	-0.010	1	0.307	0.377	0.307	0.100	-0.010
AMT_CREDIT	0.003	0.307	1	0.749	0.982	0.069	0.142
AMT_ANNUITY	0.026	0.377	0.749	1	0.749	0.074	0.008
AMT_GOODS_PRICE	-0.005	0.307	0.982	0.749	1	0.077	0.140
REGION_POPULATION_RELATIVE	-0.016	0.100	0.069	0.074	0.077	1	0.018
CLIENT_AGE	-0.259	-0.010	0.142	0.008	0.140	0.018	1
DAYS_EMPLOYED	-0.194	-0.117	0.016	-0.080	0.020	0.008	0.582
DAYS_REGISTRATION	0.155	0.030	-0.043	0.021	-0.044	-0.046	-0.289
DAYS_ID_PUBLISH	-0.047	0.000	-0.044	-0.021	-0.050	-0.007	-0.248
OWN_CAR_AGE	0.021	0.002	-0.004	0.019	-0.010	-0.023	-0.018
FLAG_EMP_PHONE	0.195	0.117	-0.017	0.079	-0.022	-0.008	-0.585
FLAG_WORK_PHONE	0.064	-0.068	-0.053	-0.036	-0.028	-0.026	-0.165
FLAG_CONT_MOBILE	0.010	-0.044	0.031	0.035	0.028	-0.003	0.008
FLAG_PHONE	-0.012	0.006	0.037	0.006	0.051	0.077	0.050
FLAG_EMAIL	0.014	0.085	-0.003	0.097	-0.002	0.052	-0.080
CNT_FAM_MEMBERS	0.888	0.001	0.059	0.075	0.053	-0.014	-0.204
REGION_RATING_CLIENT	0.060	-0.156	-0.045	-0.062	-0.052	-0.429	-0.047
REGION_RATING_CLIENT_W_CITY	0.059	-0.165	-0.053	-0.080	-0.057	-0.431	-0.040
HOUR_APPR_PROCESS_START	-0.014	0.070	0.045	0.044	0.057	0.157	-0.059
REG_REGION_NOT_LIVE_REGION	-0.014	0.057	0.007	0.032	0.008	-0.001	-0.042
REG_REGION_NOT_WORK_REGION	-0.004	0.108	0.024	0.066	0.026	0.021	-0.077
LIVE_REGION_NOT_WORK_REGION	0.001	0.108	0.035	0.074	0.036	0.060	-0.054
REG_CITY_NOT_LIVE_CITY	0.005	-0.006	-0.052	-0.018	-0.053	-0.034	-0.151
REG_CITY_NOT_WORK_CITY	0.048	-0.001	-0.040	0.001	-0.046	-0.043	-0.228
LIVE_CITY_NOT_WORK_CITY	0.056	0.007	-0.008	0.012	-0.015	-0.025	-0.145

v) Conclusion

I learned to handle large amounts of data with the help of Microsoft Excel.

Additionally, I learned how to handle outliers with the help of business rules and thresholds.

I gained good banking domain knowledge and understood the key factors responsible for the defaults in loans.

I also understood the importance of univariate, segmented univariate and bivariate analysis for analyzing the data.

Dataset Link: Bank Loan Case Study

NOTE- The above file is too large and it is recommended to download the dataset for viewing/editing.

Impact of Car Features

i) Description

For a given dataset, as a Data Analyst, the client has asked How can a car manufacturer optimize pricing and product development decisions to maximize profitability while meeting consumer demand?

This problem could be approached by analyzing the relationship between a car's features, market category, and pricing, and identifying which features and categories are most popular among consumers and most profitable for the manufacturer.

By using data analysis techniques such as regression analysis and market segmentation, the manufacturer could develop a pricing strategy that balances consumer demand with profitability, and identify which product features to focus on in future product development efforts.

ii) The Problem

Data Analytics Tasks

- Variation in popularity of car model across different market categories: How does the popularity of a car model vary across different market categories?

- Relationship between car engine and price: What is the relationship between a car's engine power and its price?

- Car features important for determining a car's price: Which car features are most important in determining a car's price?

- Variation in average car price across different manufacturer: How does the average price of a car vary across different manufacturers?

- Relationship between fuel efficiency and number of cylinders: What is the relationship between fuel efficiency and the number of cylinders in a car's engine?

iii) Design

Steps taken to clean the data

- First downloaded the dataset and opened it using Microsoft Excel.
- Converted the raw data into a table and removed duplicate rows.
- Removed the missing values in categorical columns with its mode and numerical columns with its median with the help of histograms.
- Checked for potential outliers and removed them.

Software Used: Microsoft Excel Professional

Version: 2021

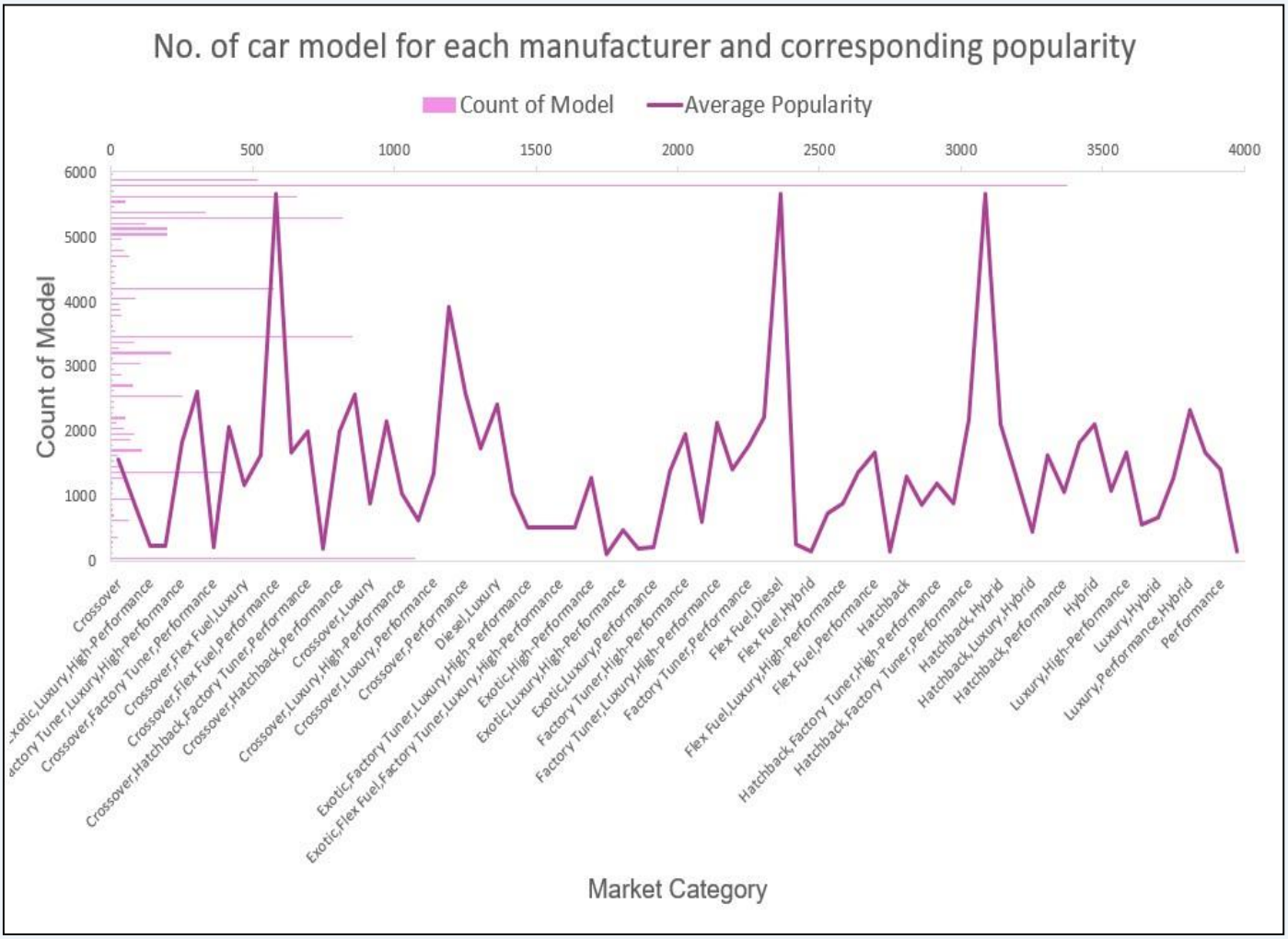
iv) Insights

Data Analytics Tasks

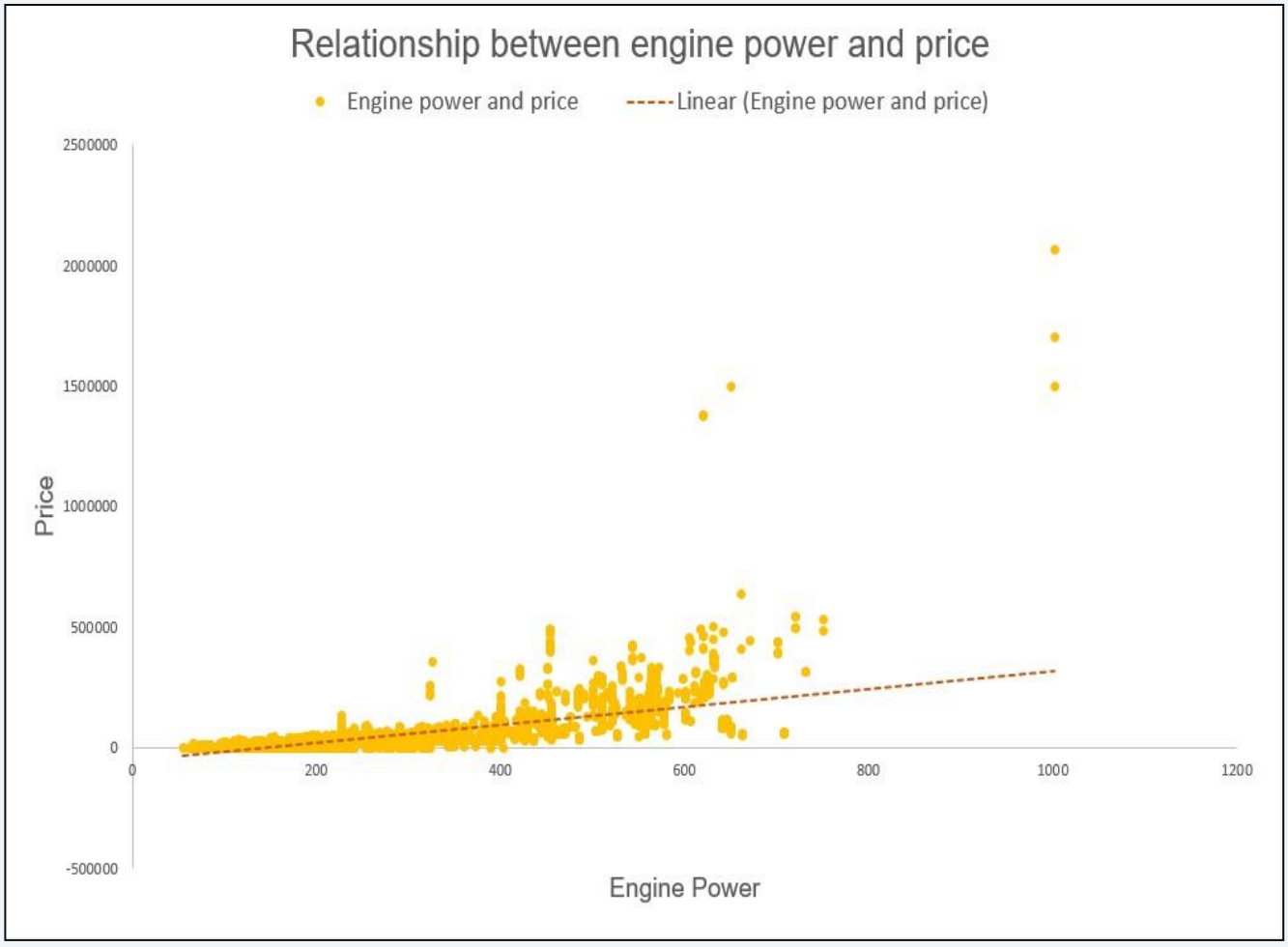
- Variation in popularity of car model across different market categories: I calculated the count of car model and their average popularity for each market category with the help of pivot table.

Market Category	Count of Model	Average Popularity
Crossover	1075	1556
Crossover,Diesel	7	873
Crossover,Exotic,Luxury,High-Performance	1	238
Crossover,Exotic,Luxury,Performance	1	238
Crossover,Factory Tuner,Luxury,High-Performance	26	1823
Crossover,Factory Tuner,Luxury,Performance	5	2607
Crossover,Factory Tuner,Performance	4	210
Crossover,Flex Fuel	64	2074
Crossover,Flex Fuel,Luxury	10	1173
Crossover,Flex Fuel,Luxury,Performance	6	1624
Crossover,Flex Fuel,Performance	6	5657
Crossover,Hatchback	72	1676
Crossover,Hatchback,Factory Tuner,Performance	6	2009
Crossover,Hatchback,Luxury	6	204
Crossover,Hatchback,Performance	6	2009
Crossover,Hybrid	42	2563
Crossover,Luxury	406	889
Crossover,Luxury,Diesel	34	2149
Crossover,Luxury,High-Performance	9	1037
Crossover,Luxury,Hybrid	24	631
Crossover,Luxury,Performance	112	1349
Crossover,Luxury,Performance,Hybrid	2	3916
Crossover,Performance	69	2586
Diesel	84	1731
Diesel,Luxury	47	2416
Exotic,Factory Tuner,High-Performance	21	1046
Exotic,Factory Tuner,Luxury,High-Performance	51	523

I then created a combo chart for visualizing the relationship between market category and popularity.



- <u>Relationship between car engine and price:</u> I created a scatter chart and added a trendline to find the relationship between car engine and price.
  We can see that there is a positive correlation between the car engine and price which indicates that as the number of car engine increases its price increases. This is the reason why the trendline is going in an upward direction.



- <u>Car features important for determining a car's price:</u> I calculated regression with the help of Data Analysis add-in for identifying the car features important for determining a car's price. I then created a bar chart to show the coefficient values for each variable to visualize their relative importance.

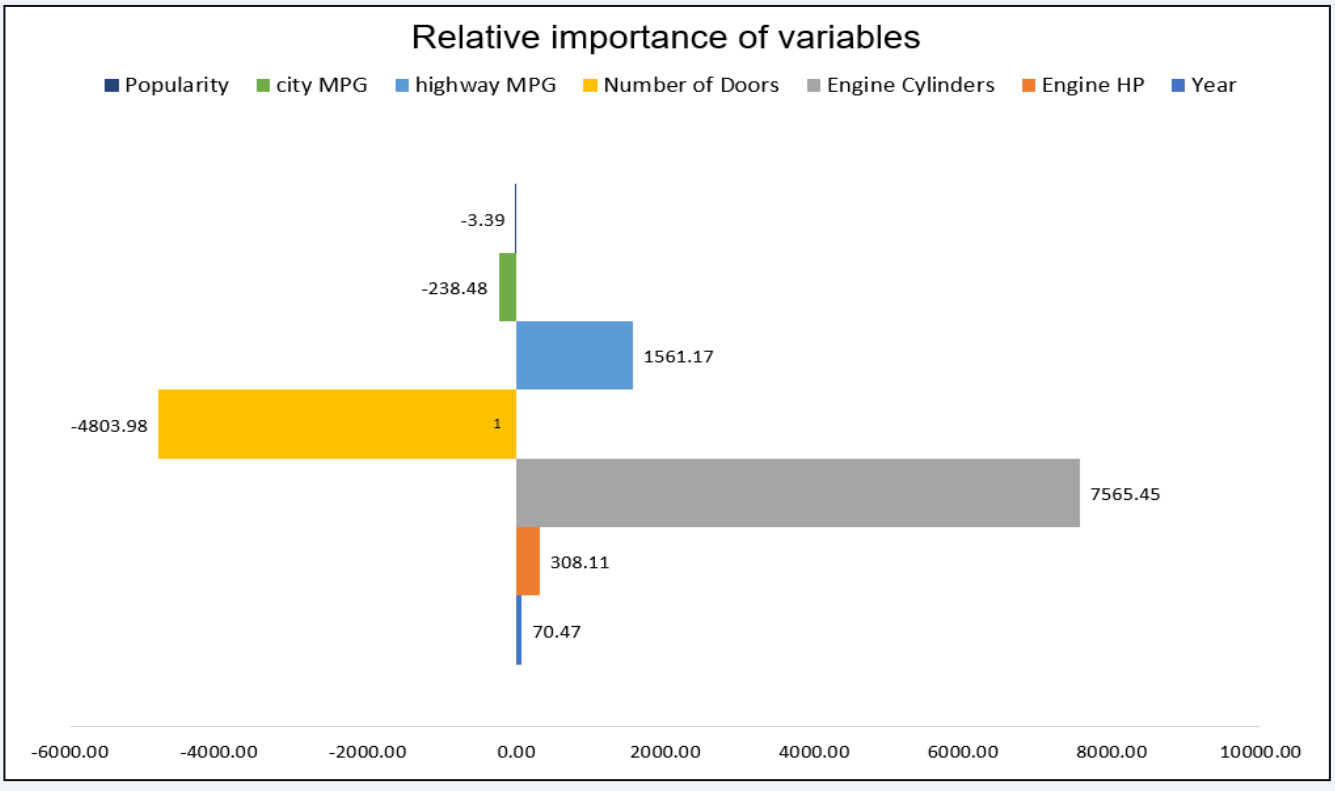
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.68125395
R Square	0.46410695
Adjusted R Square	0.46377168
Standard Error	45064.50536851
Observations	11197

ANOVA

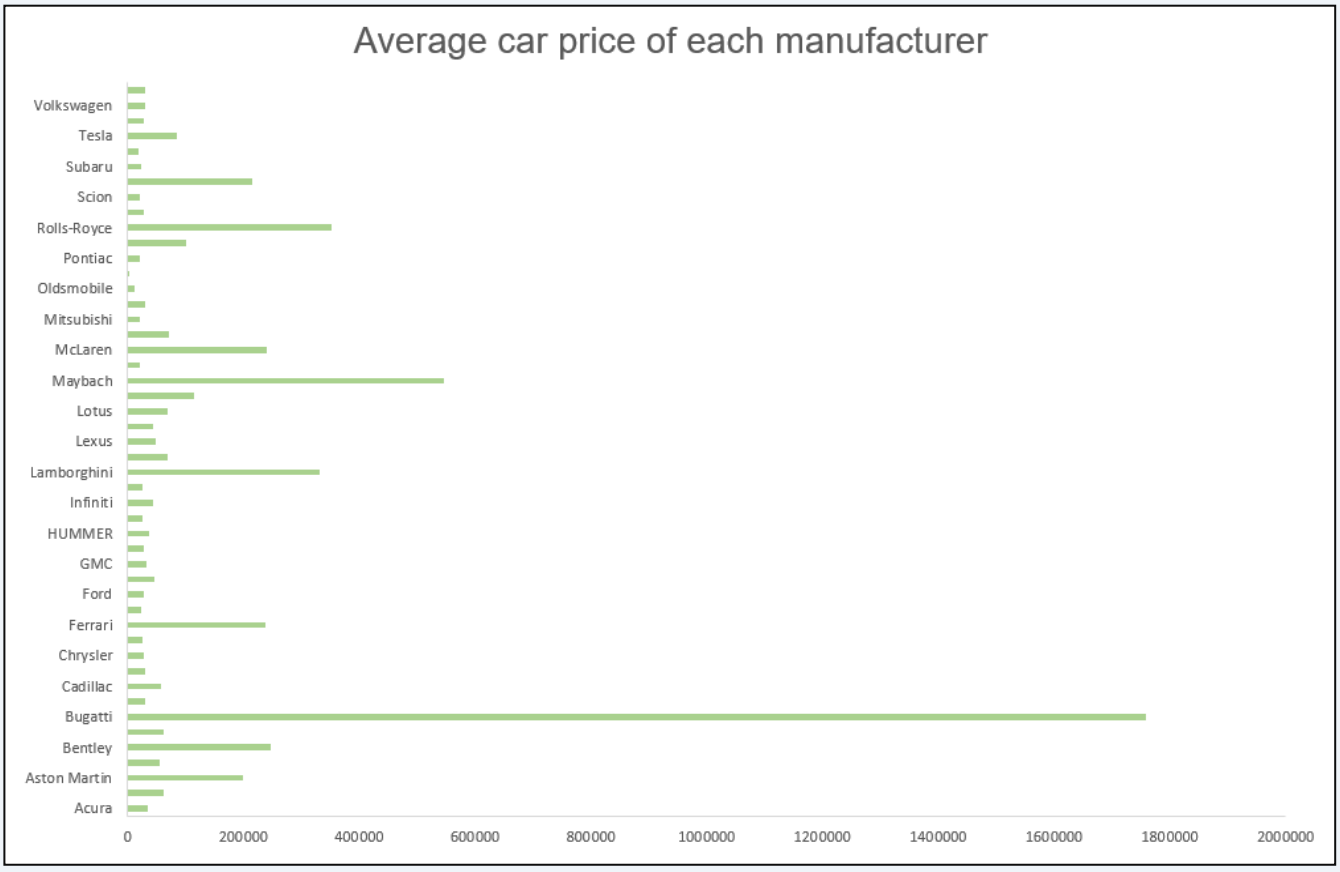
	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	1.96789E+13	2.81127E+12	1384.309415	0.00
Residual	11189	2.27227E+13	2030809644		
Total	11196	4.24016E+13			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-235571.67	159366.56	-1.48	0.14	-547958.17	76814.84	-547958.17	76814.84
Year	70.47	80.04	0.88	0.38	-86.42	227.36	-86.42	227.36
Engine HP	308.11	7.66	40.23	0.00	293.10	323.12	293.10	323.12
Engine Cylinders	7565.45	485.10	15.60	0.00	6614.56	8516.33	6614.56	8516.33
Number of Doors	-4803.98	524.92	-9.15	0.00	-5832.92	-3775.05	-5832.92	-3775.05
highway MPG	1561.17	172.04	9.07	0.00	1223.95	1898.40	1223.95	1898.40
city MPG	-238.48	143.93	-1.66	0.10	-520.61	43.66	-520.61	43.66
Popularity	-3.39	0.30	-11.37	0.00	-3.98	-2.81	-3.98	-2.81

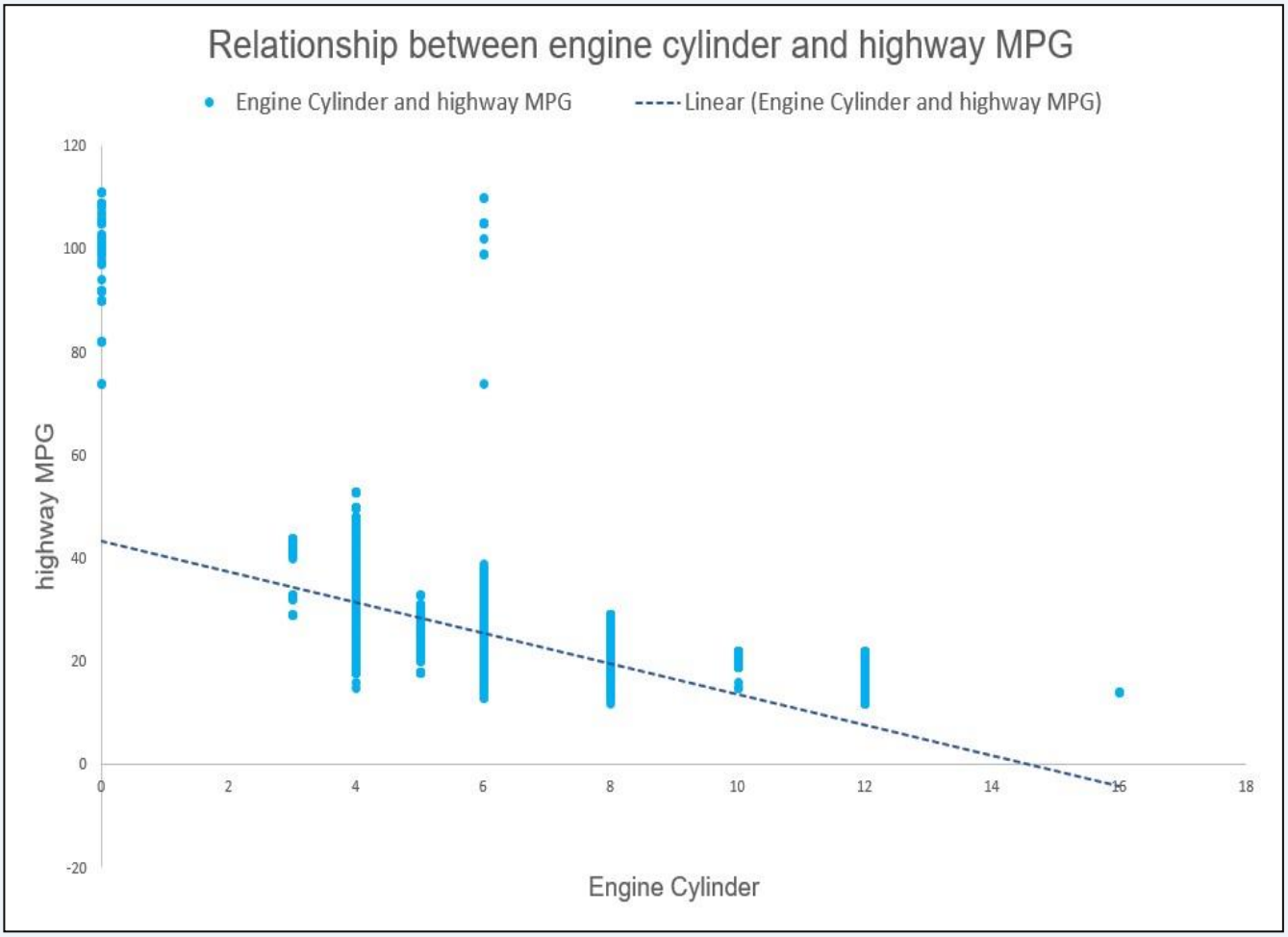


- Variation in average car price across different manufacturer: I calculated the average price of car for each manufacturer with the help of pivot table. I also created a stacked bar chart to visualize the relationship between manufacturer and average car price.

Manufacturer	Average price of car
Acura	34999
Alfa Romeo	61600
Aston Martin	198123
Audi	54583
Bentley	247169
BMW	62163
Bugatti	1757224
Buick	29034
Cadillac	56368
Chevrolet	29075
Chrysler	26723
Dodge	24857
Ferrari	238219
FIAT	22670
Ford	28511
Genesis	46617
GMC	32444
Honda	26655
HUMMER	36464
Hyundai	24926
Infiniti	42640
Kia	25514
Lamborghini	331567
Land Rover	68067
Lexus	47549
Lincoln	43861
Lotus	68377



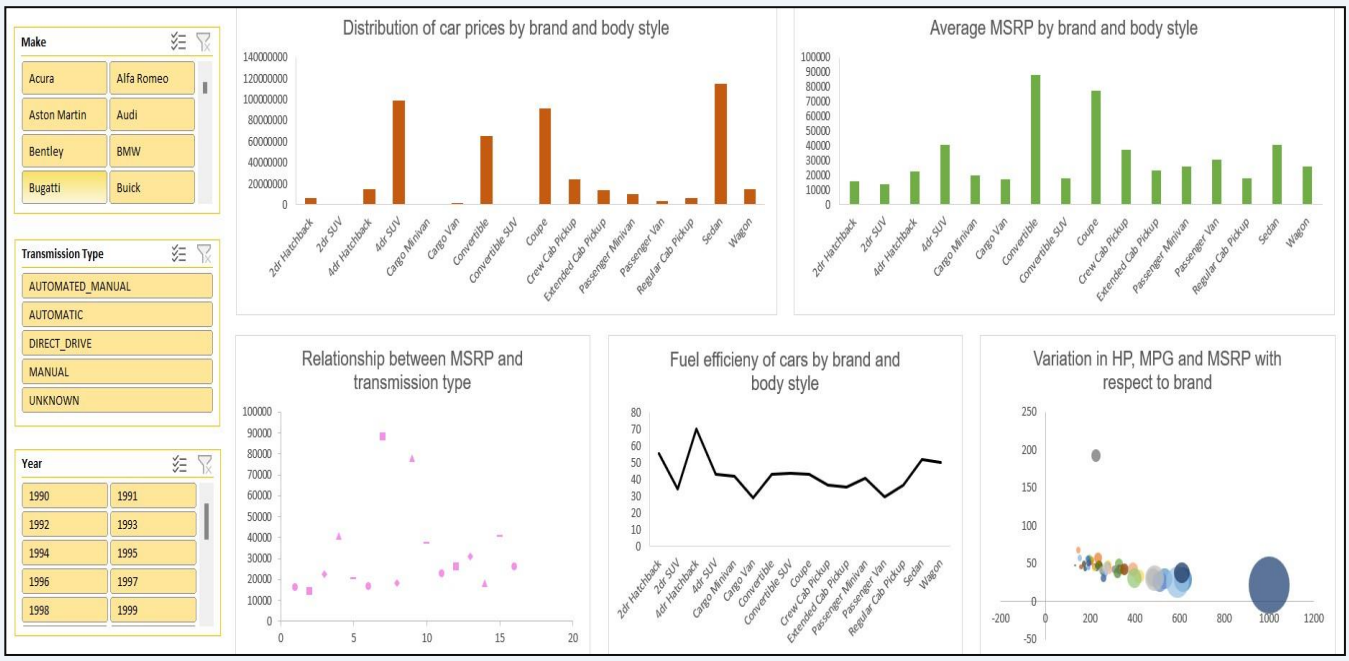
- <u>Relationship between fuel efficiency and number of cylinders:</u> I created a scatter chart and added a trendline to find the relationship between fuel efficiency and number of cylinders.
  I then calculated the correlation between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.
  We can clearly see that there is a negative correlation between engine cylinder and highway MPG which indicates that as the number of engine cylinder increases the highway MPG decreases. This is the reason why the trendline is going in a downward direction.



v) Conclusion

I created an interactive dashboard with the help of pivot tables and slicers and was able to answer these questions asked by the client:

- How does the distribution of car prices vary by brand and body style?
- Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?
- How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?
- How does the fuel efficiency of cars vary across different body styles and model years?
- How does the car's horsepower, MPG, and price vary across different brands?



Dashboard Link: Impact of Car Features

Call Volume Trend Analysis

i) Description

In this project, you'll be diving into the world of Customer Experience (CX) analytics, specifically focusing on the inbound calling team of a company.

You'll be provided with a dataset that 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).

You'll be using your analytical skills to understand the trends in the call volume of the CX team and derive valuable insights from it.

ii) The Problem

Data Analytics Tasks

- Average Call Duration: Determine the average duration of all incoming calls received by agents. This should be calculated for each time bucket.

- Call Volume Analysis: Visualize the total number of calls received. This should be represented as a graph or chart showing the number of calls against time. Time should be represented in buckets (e.g., 1-2, 2-3, etc.).

- Manpower Planning: The current rate of abandoned calls is approximately 30%. Propose a plan for manpower allocation during each time bucket (from 9 am to 9 pm) to reduce the abandon rate to 10%. In other words, you need to calculate the minimum number of agents required in each time bucket to ensure that at least 90 out of 100 calls are answered.

- Nightshift Manpower Planning: Customers also call ABC Insurance Company at night but don't get an answer because there are no agents available. Assume that for every 100 calls that customers make between 9 am and 9 pm, they also make 30 calls at night between 9 pm and 9am. Propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%.

iii) Design

Steps taken to clean the data

- First downloaded and read the dataset with pd.read_excel() command using Jupyter Notebook.
- Checked for duplicate rows and removed them.
- Removed the missing values in categorical columns with its mode and numerical columns with its median with the help of histograms.
- Checked for potential outliers and removed them with the help of box-plots.

I made an assumption during this project which was:

- 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.

Software Used: Jupyter Notebook

Version: 3.11.7

iv) Insights

Data Analytics Tasks

- Average Call Duration: I started by importing the datetime package. I then grouped the time_bucket column and calculated call_seconds for each time_bucket.
  I also calculated the average time spent by an agent and with the help of datetime.timedelta() function, I converted the seconds into time.

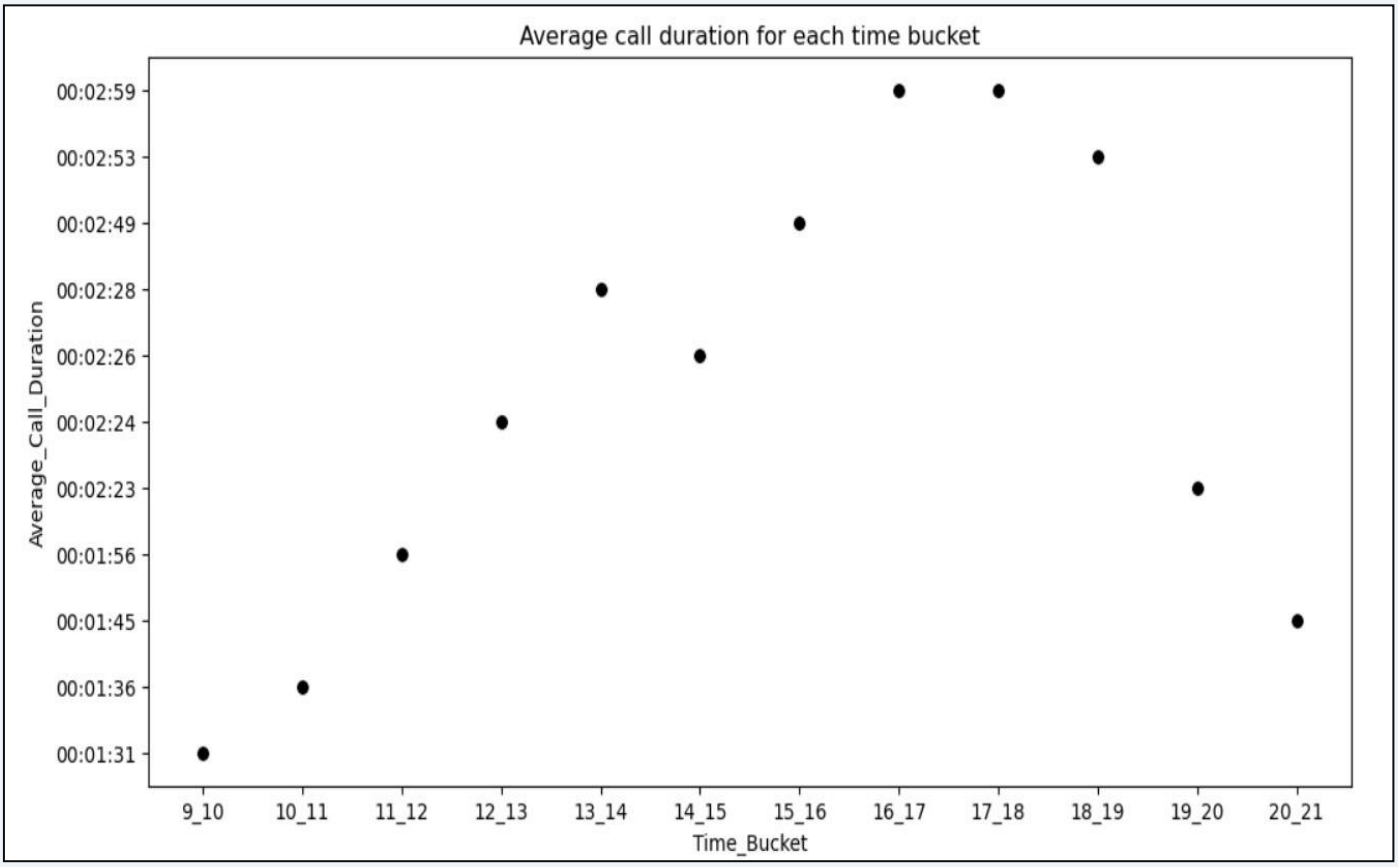
```
import datetime
time_bucket = new_df.groupby("Time_Bucket")["Call_Seconds (s)"].mean()  # calculates mean of call_seconds for each time_bucket
time_bucket = time_bucket.sort_values(ascending=True).reset_index()  # sorts the mean of call_seconds in ascending order
average_time_by_agent = int(time_bucket['Call_Seconds (s)'].mean())  # returns the average time spent by an agent
time_bucket['Call_Seconds (s)'] = time_bucket['Call_Seconds (s)'].apply(lambda x: datetime.timedelta(seconds=x))  # converts the seconds into time
time_bucket
```

```
time_bucket['Average_Call_Duration'] = time_bucket['Call_Seconds (s)'].apply(lambda x: str(x).split(' ')[-1])  # returns only the time
time_bucket['Average_Call_Duration'] = time_bucket['Average_Call_Duration'].apply(lambda x: str(x).split('.')[0]) # removes milliseconds
time_bucket = time_bucket.drop('Call_Seconds (s)',axis=1)  # we drop seconds since we have time duration now
time_bucket
```

	Time_Bucket	Average_Call_Duration
0	9_10	00:01:31
1	10_11	00:01:36
2	20_21	00:01:45
3	11_12	00:01:56
4	19_20	00:02:23
5	12_13	00:02:24
6	14_15	00:02:26
7	13_14	00:02:28
8	15_16	00:02:49
9	18_19	00:02:53
10	17_18	00:02:59
11	16_17	00:02:59

I created a scatterplot to visualize the average call_duration for each time_bucket.

```
import warnings
warnings.filterwarnings("ignore",category=FutureWarning)  # hides the FutureWarning
sns.scatterplot(data = time_bucket, x ='Time_Bucket', y ='Average_Call_Duration',color='black',s=50)
plt.gca().invert_yaxis()  # inverts the y axis
plt.title("Average call duration for each time bucket")
plt.show()
```



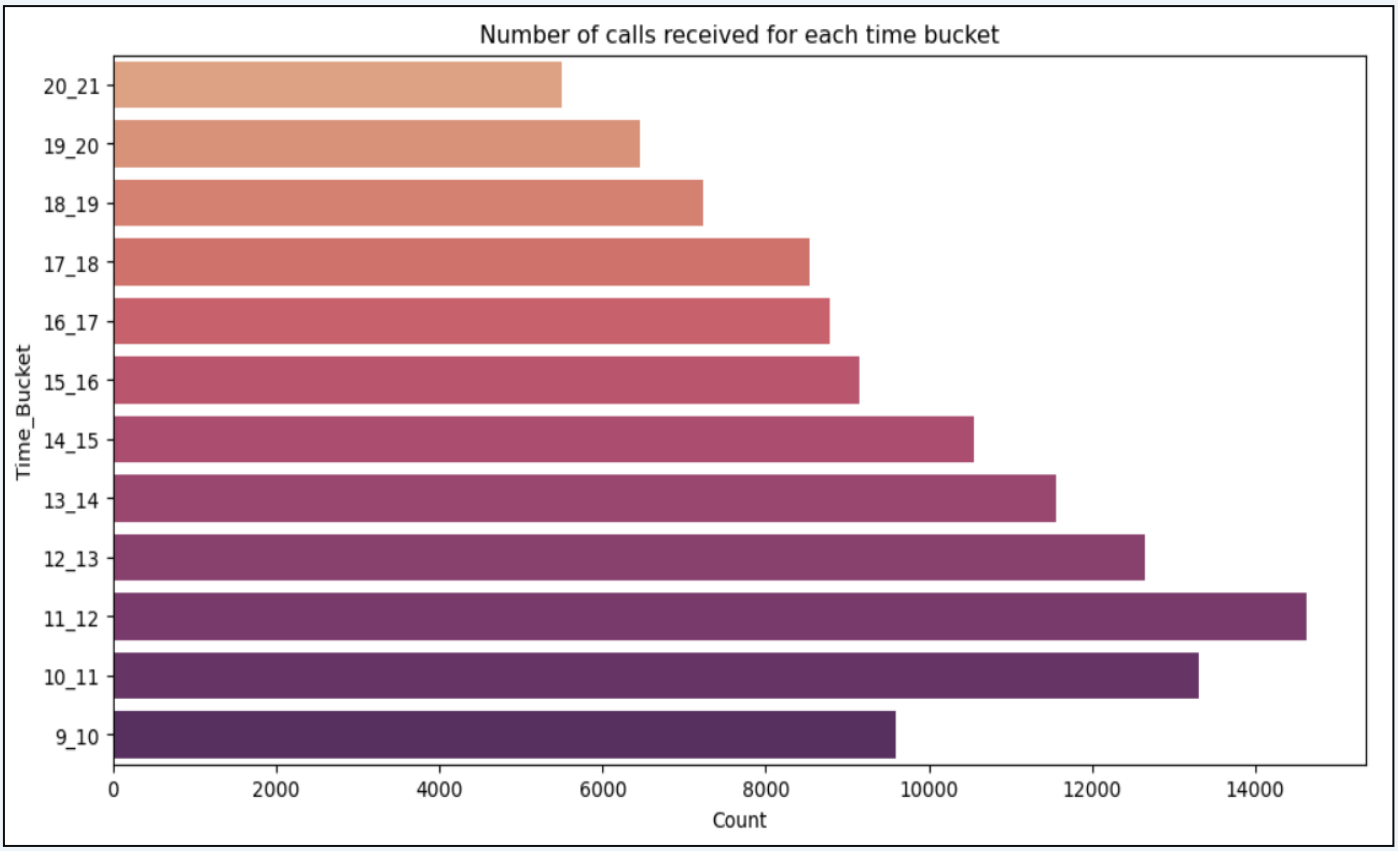
- <u>Call Volume Analysis:</u> I grouped the time_bucket column with the help of groupby() function and calculated the count of calls for each time_count with the help of size() function.

```
total_call = new_df.groupby('Time_Bucket').size().reset_index(name='Count')  # returns count of each time_bucket
total_call
```

	Time_Bucket	Count
0	10_11	13310
1	11_12	14625
2	12_13	12651
3	13_14	11555
4	14_15	10559
5	15_16	9157
6	16_17	8781
7	17_18	8533
8	18_19	7236
9	19_20	6460
10	20_21	5505
11	9_10	9587

I created a bar plot to visualize the number of calls received for each time_bucket and kept the orientation as ‘horizontal’ for this barplot as the default orientation is vertical.

```
order_1 = ["20_21","19_20","18_19","17_18","16_17","15_16","14_15","13_14","12_13","11_12","10_11","9_10"]
sns.barplot(data=total_call,x='Count',y='Time_Bucket',palette='flare',order = order_1, orient='horizontal')
plt.title("Number of calls received for each time bucket")
plt.show()
```



- <u>Manpower Planning:</u> I started by calculating the percentage of abandoned calls which was approximately 29% and then calculated the value to make it to 10%.
  I then calculated the reduction ratio to multiply it with count of abandoned calls (10%) and also calculated the manpower needed for each time_bucket.

```
total_abandoned_calls = len(new_df[new_df['Call_Status']=="abandon"]) # calculates the count of abandoned calls
total_calls = len(new_df['Call_Status'])  # returns the total number of calls
percentage_of_abandoned_calls = int(total_abandoned_calls / total_calls * 100)  # returns the percentage of abandoned calls
ten_percent_of_total_calls = int(total_calls / 10)  # calculates 10% of abandoned calls
reduction_ratio = round(ten_percent_of_total_calls / total_abandoned_calls,2) # calculates the reduction_ratio to get count of abandoned_calls (10%)
total_abandoned_calls, total_calls, percentage_of_abandoned_calls, ten_percent_of_total_calls, reduction_ratio
```

```
(34403, 117959, 29, 11795, 0.34)
```

```
abandoned_df = new_df[new_df['Call_Status']=="abandon"]  # creates a dataframe where call status is abandon
reduced_abandoned_rate = abandoned_df.groupby("Time_Bucket").size().reset_index(name="Count of Abandoned Calls")
# calculates count of abandoned calls for each time_bucket
reduced_abandoned_rate
```

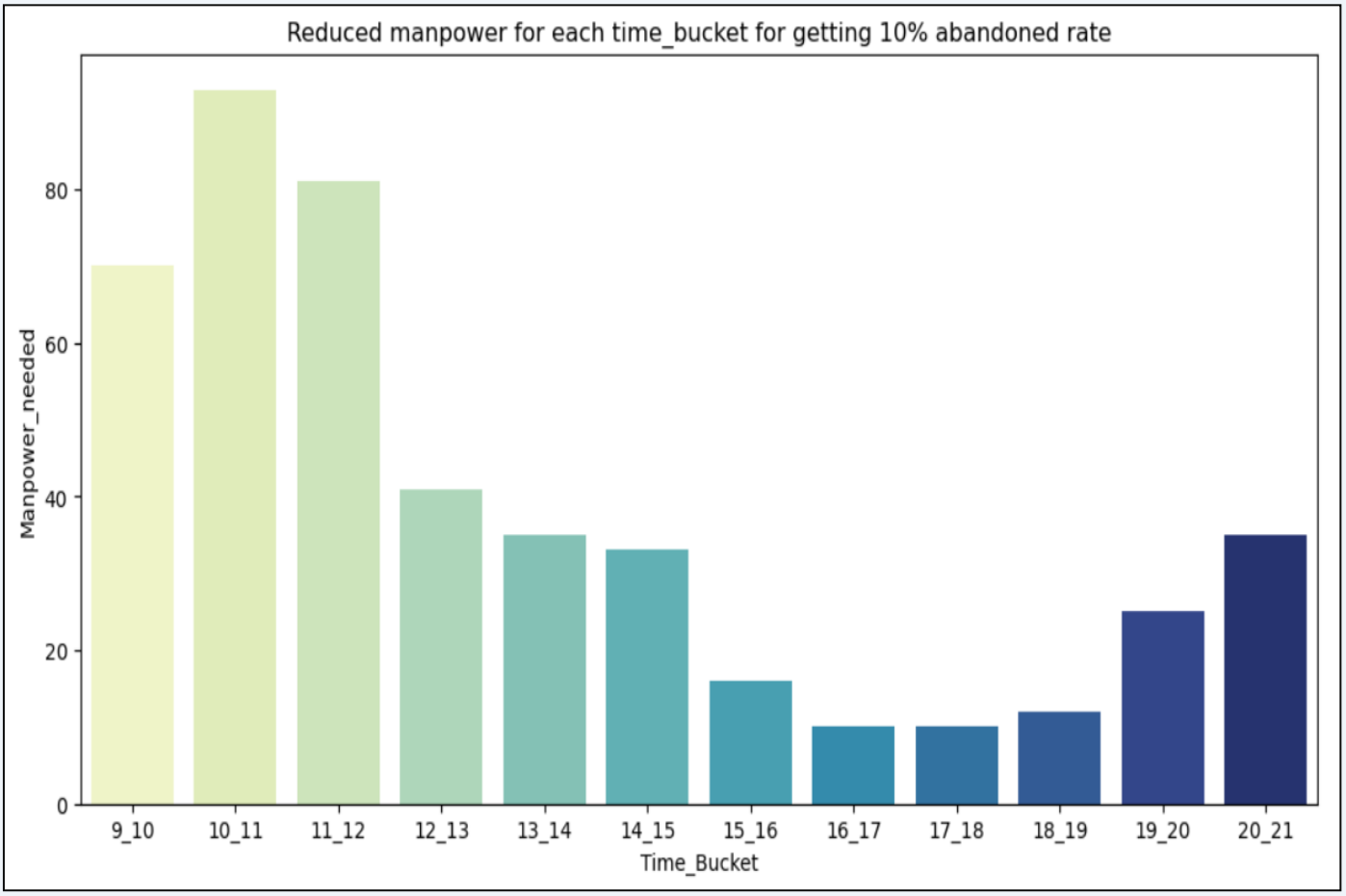
```
reduced_abandoned_rate['Count of Abandoned Calls (10%)'] = (reduced_abandoned_rate['Count of Abandoned Calls'] * reduction_ratio).astype(int)
# calculates the count of abandoned calls (10%) for each time_bucket
reduced_abandoned_rate
```

```
reduced_abandoned_rate['Manpower_needed'] = (reduced_abandoned_rate['Count of Abandoned Calls (10%)'] / average_calls_per_hour).astype(int)
# calculates the manpower required for each time_bucket to make rate of abandoned calls to 10%
reduced_abandoned_rate
```

	Time_Bucket	Count of Abandoned Calls	Count of Abandoned Calls (10%)	Manpower_needed
0	10_11	6911	2349	93
1	11_12	6028	2049	81
2	12_13	3073	1044	41
3	13_14	2617	889	35
4	14_15	2475	841	33
5	15_16	1214	412	16
6	16_17	747	253	10
7	17_18	783	266	10
8	18_19	933	317	12
9	19_20	1848	628	25
10	20_21	2625	892	35
11	9_10	5149	1750	70

I created a barplot to visualize the minimum manpower needed to reduce abandon rate to 10%.

```
order = ["9_10","10_11","11_12","12_13","13_14","14_15","15_16","16_17","17_18","18_19","19_20","20_21"]
sns.barplot(data = reduced_abandoned_rate,x='Time_Bucket',y='Manpower_needed',order=order,palette='YlGnBu')
plt.title("Reduced manpower for each time_bucket for getting 10% abandoned rate")
plt.show()
```



- Nightshift Manpower Planning: I started by calculating the total number of answered calls while keeping the abandoned rate at max 10%. I then calculated the average number of calls per day and then got the average number of calls per night with the help of it which was (30%) of average number of call per day.

```
total_answered_calls = total_calls - ten_percent_of_total_calls  # calculates the total answered calls while keeping max 10% abandoned rate
average_call_per_day = int(total_answered_calls/number_of_working_days)  # calculates the average number of call for a day
average_call_per_night = int(0.30 * average_call_per_day) # 30% of average_call_per_day
total_answered_calls, average_call_per_day, average_call_per_night
```

```
(106164, 3538, 1061)
```

I then created a dataframe for the distribution of 30 calls coming in night for every 100 calls coming in between 9am – 9pm (i.e. 12 hours slot) with the help of pd.Dataframe() function of pandas library.

```
distribution = pd.DataFrame({  # creates a new dataframe
    '21_22' : [3],
    '22_23' : [3],
    '23_24' : [2],
    '12_1' : [2],
    '1_2' : [1],
    '2_3' : [1],
    '3_4' : [1],
    '4_5' : [1],
    '5_6' : [3],
    '6_7' : [4],
    '7_8' : [4],
    '8_9' : [5]
})
distribution_of_given_calls = pd.DataFrame({
    'Time_Bucket' : distribution.columns,
    'Distribution_of_calls' : distribution.iloc[0].values
})
distribution_of_given_calls
```

I then calculated the distribution of calls for night by multiplying the average_call_per_night with percentage distribution of calls for each time_bucket and converted the result into integer with the help of astype(int) function of pandas.

```
distribution_of_given_calls['Percentage Distribution'] = distribution_of_given_calls['Distribution_of_calls'] / number_of_calls_in_night
# percentage distribution of calls
distribution_of_given_calls
```

I then calculated the manpower_needed by dividing the distribution of calls for night with average_calls_per_hour for each time_bucket and converted the result into integer with the help of astype(int) function of pandas.

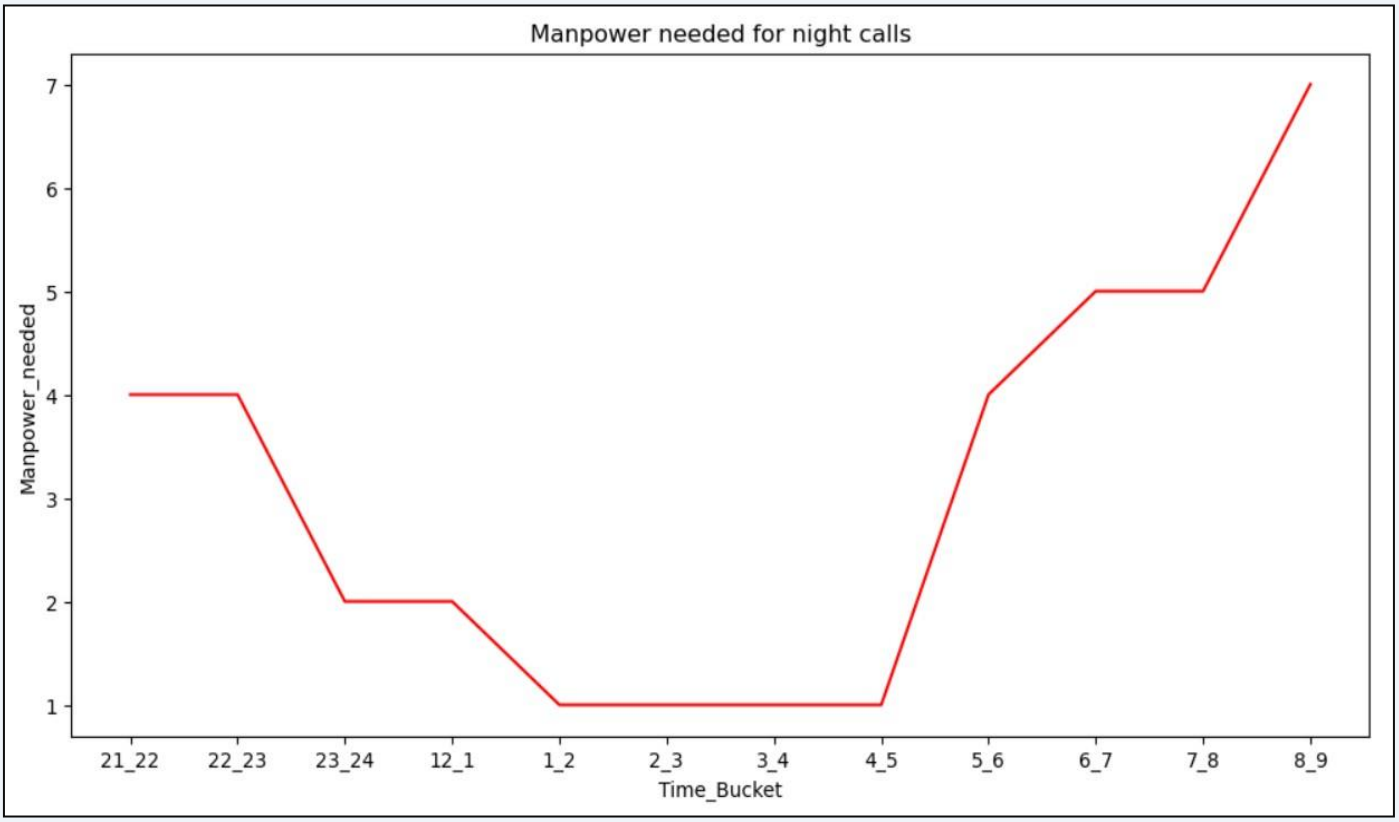
```
distribution_of_given_calls['Distribution_of_calls_for_night'] = (average_call_per_night * distribution_of_given_calls['Percentage Distribution']).astype
# distribution of calls for night for each time_bucket
distribution_of_given_calls
```

```
distribution_of_given_calls['Manpower_needed'] = (distribution_of_given_calls['Distribution_of_calls_for_night'] / average_calls_per_hour).astype(int)
# calculates the manpower required for each time_bucket in night
distribution_of_given_calls
```

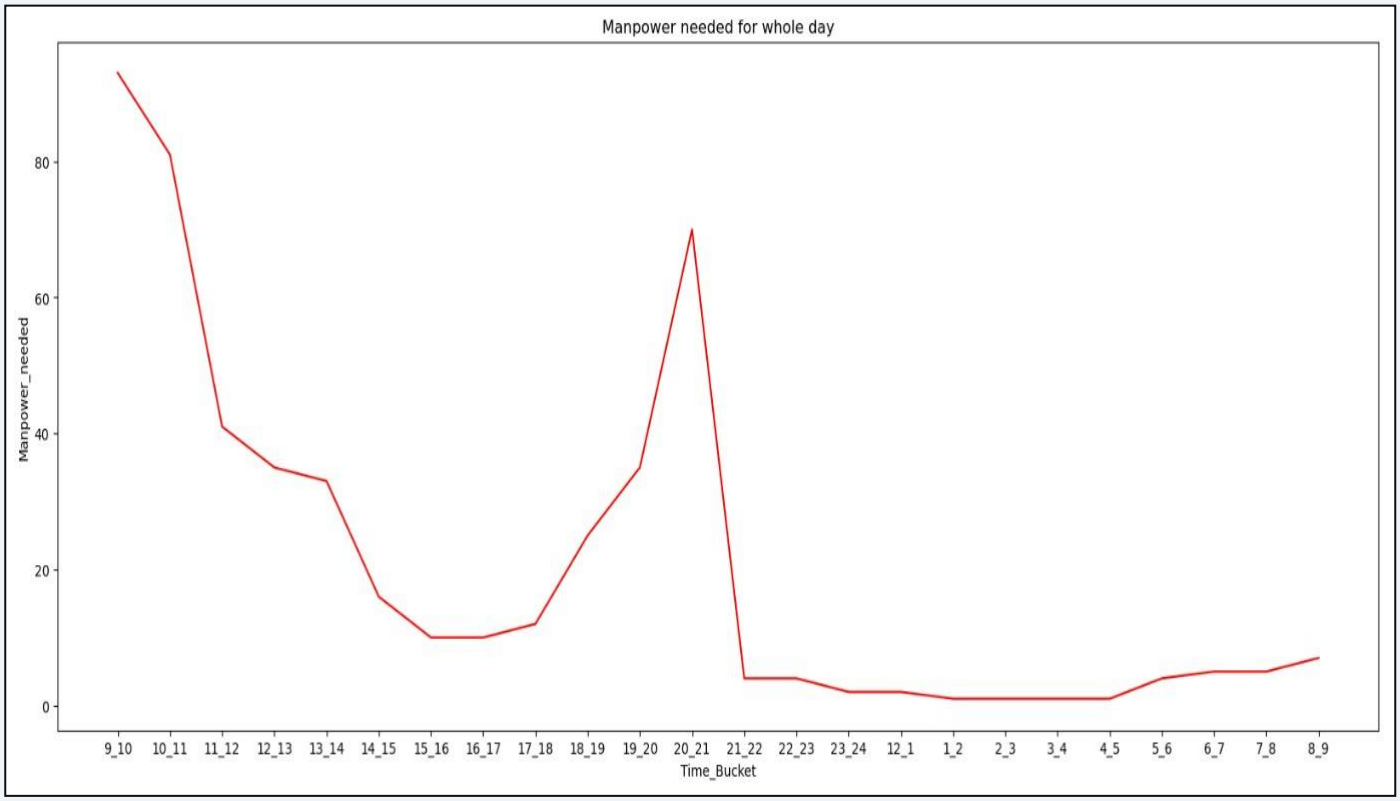
	Time_Bucket	Distribution_of_calls	Percentage Distribution	Distribution_of_calls_for_night	Manpower_needed
0	21_22	3	0.100000	106	4
1	22_23	3	0.100000	106	4
2	23_24	2	0.066667	70	2
3	12_1	2	0.066667	70	2
4	1_2	1	0.033333	35	1
5	2_3	1	0.033333	35	1
6	3_4	1	0.033333	35	1
7	4_5	1	0.033333	35	1
8	5_6	3	0.100000	106	4
9	6_7	4	0.133333	141	5
10	7_8	4	0.133333	141	5
11	8_9	5	0.166667	176	7

I then created a lineplot for visualizing the manpower needed for night calls.

```
sns.lineplot(data = distribution_of_given_calls, x='Time_Bucket', y='Manpower_needed',color='red')
plt.title("Manpower needed for night calls")
plt.show()
```



I finally created a lineplot for visualizing the manpower needed for the entire day.



v) Conclusion

By working on this project, I understood the role of a data analyst with respect to advertising domain.

I learned how to perform data analysis with the help of Jupyter Notebook (Python).

I understood the importance of pandas, matplotlib and seaborn libraries with the help of this project.

Additionally, I got to know about some functions in python like timedelta(), reset_index(), Categorical(), etc.

Moreover, I recognized the importance of Customer Experience Analytics in enhancing data-driven decision-making processes and improving customer satisfaction and retention. Looking forward, I plan to leverage these skills and experiences in future projects to drive impactful data-driven decisions in the advertising industry.

Notebook link: Call Volume Trend Analysis