DATA ANALYSIS PORTFOLIO

ANURAG JOHN PHILLIPS

+91- 97218-33777

https://www.linkedin.com/in/anurag-john-phillips/

PROFESSIONAL BACKGROUND

I am an accomplished MBA graduate with over 5 years of outstanding performance in Sales and Business Development roles, both in Bengaluru, India, and Belleville, Canada. My extensive experience spans the Food and Beverage industry, as well as the Supply Chain sector, where I consistently ranked among the top performers within my organizations.

My tenure as a Business Development Manager has equipped me with a robust set of soft skills, including relationship building, negotiation, and strategic planning, all of which I believe are transferable assets in my transition to the Data Analytics field.

To facilitate this transition, I undertook a transformative internship with Trainity, where I excelled as one of the top performers. During this internship, I honed my technical skills in Excel, SQL, Statistics, Data Analysis, Tableau, Power BI, and Python. I successfully completed various data analysis projects, which are showcased in this portfolio.

I am enthusiastic about leveraging my unique blend of business acumen and analytical capabilities to excel as a Data Analyst. My comprehensive skill set and proven track record of achieving results make me well-prepared for this exciting new chapter in my career.

TABLE OF CONTENT

MODULE 1

DATA ANALYTICS PROCESS

DATA ANALYTICS PROCESS

DESCRIPTION:

We use Data Analytics in everyday life without even knowing it.
For eg : Going to a market to buy something .

- **Plan** We first decide which things I need before going to market. Is it a shirt, jeans , footwear etc.
- **Prepare** Next I need to check how much I am willing to spend and how to get that money.
- **Process** Then I need to check how much I want from the data. Like if I am going to buy footwear what do I want - slippers / shoes / sandals etc.
- **Analyze** You obviously won't buy things which are out of trend, Also you need to check does the jeans which you have and the color of t-shirt you want to buy, will it make a good combination.
- **Share** Now you communicate your idea to the shopkeeper to find the best suitable fit for you.
- **Act** Then you finally buy it!

YOUR TASK:

Your task is to give an example (s) of such a real-life situation where we use Data Analytics and link it with the data analytics process. You can prepare a PPT/PDF on a real-life scenario explaining it with the above process (Plan, Prepare, Process, Analyze, Share, Act) and submit it as part of this task.

PROJECT LINK

https://docs.google.com/presentation/d/1yz_krrMPR9XC3aGxlpn49aud7oWAVKD_/edit?usp=drive_link&ouid=109466755193972209405&rtpof=true&sd=true

DATA ANALYTICS PROCESS

OVERVIEW

My goal is to join a gym and achieve a fit physique with 6-pack abs.

Plan- Setting Fitness Goals

- I have a clear and specific fitness goal i.e. of obtaining 6-pack abs.
- This goal will provide a roadmap for effective fitness progress.
- It will help me measure my achievements, track my progress, and stay accountable.
- This plan will include specific strategies, such as workout routines and nutrition plans, to ensure that I stay on course.

Prepare - Assessing Current Fitness Level

- I will assess my current fitness level as part of my preparation phase.
- I will measure my body composition, including body fat percentage and muscle mass, to understand my starting point.
- I will conduct research to find local gyms that align with my fitness goals.
- By being prepared it will help me in monitoring my progress accurately and making necessary adjustments to my fitness plan.

Process - Analyzing Fitness Data

- I will be analyzing my body composition data and other data gathered to find out the most effective workout routines and dietary approaches for my fitness goal
- This will help me in maximizing my chances of achieving my fitness goal of getting 6 pack abs

Analyze - Tailoring the Fitness Routine

- I will be making a customized plan based on my metabolic rate, my exercise preference and my body composition
- By tailoring my fitness routine to suit my needs, I can achieve better results in a more targeted and efficient manner as it is customized according to me, I am more likely to stick to my fitness plan and will be able to maintain it for a longer time.

Share - Seeking Guidance and Support

- I will be taking the help of my personal trainer and will also be joining the fitness communities or pages on social media to share my goals, track progress and learn about new exercises
- This will help me in getting valuable feedback, recommendations and advice from experts in the fitness industry
- Interacting with like-minded people will help me stay motivated

Act - Implementing the Fitness Plan

- This is the most important step in my fitness journey, I can plan 100 things but if I don't act upon my plan, it won't become a reality
- I will be eating properly, tracking progress consistently and will be going to the gym regularly
- This small but consistent change in my plan will ensure that I reach my goal

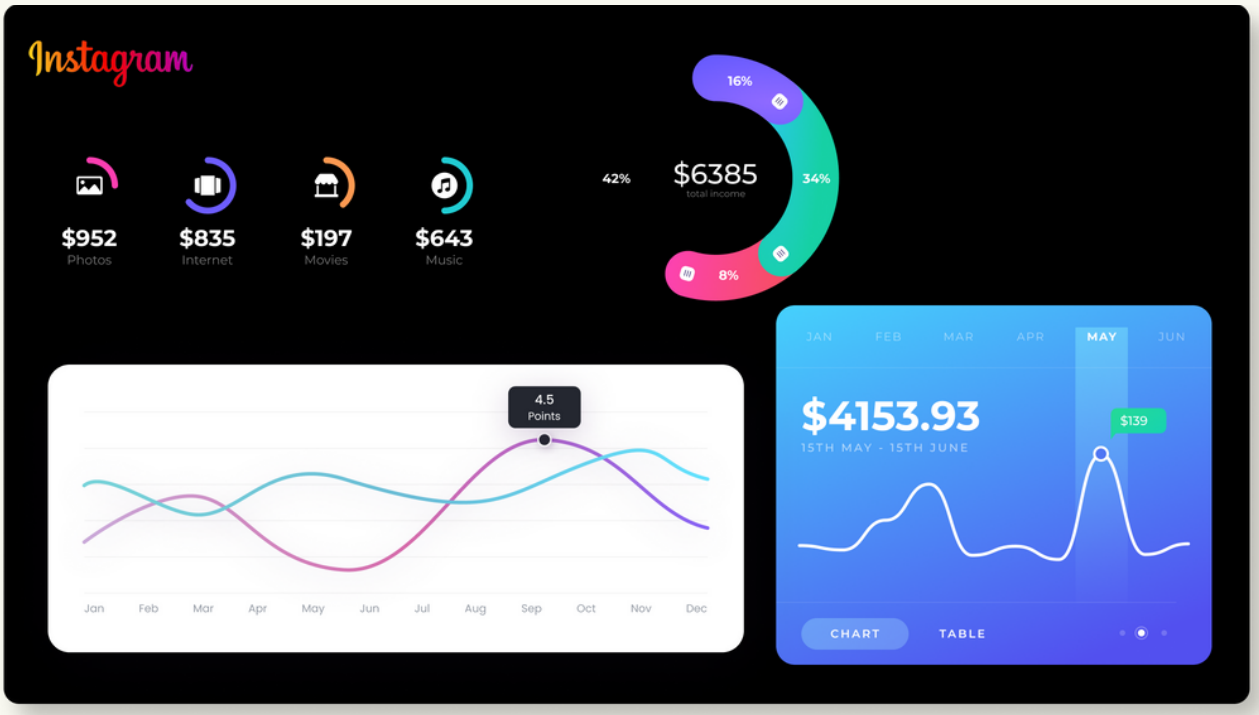
Conclusion - Data Analytics for Fitness Success

- Data analytics plays a crucial role in my fitness journeys, providing a systematic approach to achieving my goal of achieving a fit physique with 6-pack abs
- Through the Data Analytics process people like me can make informed decisions, track progress, and optimize their workouts for better results.

MODULE 2

INSTAGRAM USER ANALYTICS

SQL FUNDAMENTALS



INSTAGRAM USER ANALYTICS

PROJECT DESCRIPTION:

This project focuses on analysing user engagement and providing valuable insights for the Instagram platform.
Through this analysis, I aim to derive valuable information for the Marketing, Product, and Development teams.
I will track User interactions, Engagement Patterns of the customers with Instagram, and key metrics of likes, number of photos, what hashtags are being used etc. to gain insights that can guide decision-making for launching AD campaigns, developing new features, and contributing to overall business growth.

APPROACH:

My approach to solving this project was firstly to analyse the database. I made this table (picture attached below) to make myself aware of what all columns are there in each table and which are the common columns through which will be able to link the tables.

comments	Id	Comment_text	User_id	Photo_id	Created_at
follows	Follower_id	Followee_id	Created_at		
likes	User_id	Photo_id	Created_at		
photos	Id	Image_url	User_id	Created_dat	
photo_tags	Photo_id	Tag_id			
tags	Id	Tag_name	Created_at		
users	Id	Username	Created_at		

Using this approach helped me to derive meaningful conclusions regarding user behaviour, engagement, contest outcomes, hashtag popularity, and optimal ad campaign launch days.

TECH-STACK USED:

MySQL Workbench Version 8.0.33 for MacBook was used to analyse Ig_Clone database and extract insights.

INSTAGRAM USER ANALYTICS

INSIGHTS

Analysing the database helped in answering the questions asked by the Marketing Team

- **Rewarding Most Loyal Users:** Identified the five oldest users on Instagram, recognizing their loyalty and longevity on the platform. Here I checked when was each account created on and then sorted it by ascending to know the five oldest accounts.

```
SELECT
*
FROM
users
ORDER BY created_at ASC
LIMIT 5;
```

- **Reminding Inactive Users to Start Posting:** Identifying users who have never posted a single photo and can target them with promotional emails to encourage their active participation. Here I found which all ids are there in Users table but not there in Photos table. These are the users who have never posted any photos.

```
SELECT
username
FROM
users
LEFT JOIN
photos ON users.id = photos.user_id
WHERE
photos.user_id IS NULL;
```

- **Declaring Contest Winner:** By analysing , the contest winner was determined by identifying the user with the highest number of likes on a single photo. Here I joined the “likes” table with the “photos” table to know which photo got the most likes.

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

INSTAGRAM USER ANALYTICS

INSIGHTS

- **Hashtag Researching:** By examining the data, the top five most commonly used hashtags on the platform were identified, which provided insights for effective content reach. Here we analysed “photo_tags” table and counted the occurrence of each tag and then ranked the top 5 tags.

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

- **Launch AD Campaign:** To determine the best day to launch ads, I analysed the user registration data and identified the day of the week with the highest number of registrations. I used Google here to know search for the function which can tell me the name of the day based on the date provided.

```
SELECT
DAYNAME(created_at) AS registration_day,
COUNT(*) AS registration_count
FROM
users
GROUP BY registration_day
ORDER BY registration_count DESC
LIMIT 1;
```

- **User Engagement**- Analysing the database helped in answering the question if the users are still as active on Instagram as before or if are they making fewer posts. To know this I calculated the average number of posts per user by dividing the total number of photos by the total number of users

```
SELECT
COUNT(*) AS total_photos,
(SELECT
COUNT(*)
FROM
users) AS total_users,
COUNT(*) / (SELECT
COUNT(*)
FROM
users) AS avg_posts_per_user
FROM
photos;
```

INSTAGRAM USER ANALYTICS

- **Bots & Fake Accounts-** It is impossible for a human being to like every single photo on Instagram but it can be done through Internet bots. Hence it is very important to analyse the bots/fake accounts to make the platform as authentic and genuine as possible. Here we identified how many likes has each distinct id done, then compared it to the total, if both of them matches, then we conclude that it is a fake account or done by a bot

```
SELECT
users.id AS user_id,
users.username,
COUNT(likes.photo_id) AS liked_photo_count
FROM
users
INNER JOIN
likes ON users.id = likes.user_id
GROUP BY users.id
HAVING COUNT(DISTINCT likes.photo_id) = (SELECT
COUNT(*)
FROM
photos);
```

RESULT:

Through the project, I not only gained valuable insights into user engagement and behaviour on Instagram, but it also helped me in working with a database which is similar to a live database, it has helped me make the queries by relating different SQL tables with different SQL queries.

After I completed this project it helped me understand SQL more, how to think while working with SQL and how to join two different tables. How to use Google properly to get my answers when I am stuck.

I believe my knowledge of SQL has increased as I have completed this project.

PROJECT LINK

https://drive.google.com/file/d/1CuQyyzZGK2W4wea6cNIdOaAnW3WOWKUn/view?usp=drive_link

MODULE 3

OPERATION ANALYTICS & INVESTIGATING METRIC SPIKE

ADVANCED SQL

OPERATION ANALYTICS & INVESTIGATING METRIC SPIKE

PROJECT DESCRIPTION:

Case Study 1: In this project, we are examining the number of jobs reviewed per hour per day in November 2020, the 7-day rolling average of throughput, language distribution in the last 30 days, and identifying duplicate rows using SQL queries.

Case Study 2: In this project, we are analyzing data from three tables namely "users," "events," and "email_events" to measure User Engagement, Growth, and Email Service Performance. Weekly insights to help understand User activeness, Product Quality, and Email Engagement Metrics.

APPROACH:

Case 1 Approach: I used SQL queries to analyse the dataset. I created more rows in the dataset. I calculated the number of jobs reviewed per hour per day and the 7-day rolling average of throughput (used Google to understand more about the 7-day rolling average). Additionally, I used grouping and filtering techniques to determine the percentage share of each language in the last 30 days and identify any duplicate rows in the data.

Case 2 Approach: Here I had to create a database and then upload all the data which was in CSV to SQL, make three different tables and use SQL queries to extract, filter, and aggregate data from the provided tables. The final results were presented in a structured format to present the insights gained from the analysis. I have presented both the queries used and the output for better understanding.

TECH-STACK USED:

MySQL Workbench Version 8.0.33 for MacBook.

Microsoft Excel for Mac Version 16.74.

OPERATION ANALYTICS & INVESTIGATING METRIC SPIKE

CASE STUDY 1:

A: Calculate the number of jobs reviewed per hour per day for November 2020.

```
SELECT ds AS date,

SUM(jobs_done) AS jobs_done_everyday,
SUM(jobs_done) / 24 AS hours_spent_everyday

FROM (

SELECT ds, COUNT(job_id) AS jobs_done

FROM job_data

WHERE ds >= '2020-11-01' AND ds <= '2020-11-30'

GROUP BY ds

) subquery
GROUP BY date;
```

date	jobs_done_everyd...	hours_spent_everyd...
2020-11-30	17	0.7083
2020-11-29	11	0.4583
2020-11-28	10	0.4167
2020-11-27	7	0.2917
2020-11-26	7	0.2917
2020-11-25	7	0.2917

B: Calculate the 7-day rolling average of throughput.
For throughput, do you prefer daily metric or 7-day rolling and why?

```
SELECT
date_value,
AVG(throughput_per_second) OVER (ORDER BY date_value ROWS BETWEEN 6 PRECEDING AND CURRENT ROW) AS
rolling_avg_throughput,
throughput_per_second
FROM (
SELECT ds AS date_value, COUNT(*) / (24 * 60 * 60) AS throughput_per_second
FROM job_data
GROUP BY ds
) my_subquery
ORDER BY date_value;
```

date_value	rolling_avg_throughput	throughput_per_second
2020-11-25	0.00010000	0.0001
2020-11-26	0.00010000	0.0001
2020-11-27	0.00010000	0.0001
2020-11-28	0.00010000	0.0001
2020-11-29	0.00010000	0.0001
2020-11-30	0.00011667	0.0002

For throughput we prefer, 7-day rolling average because it is like a special way of looking at the throughput that gives us a steadier and smoother view of how things are going over time. It does this by considering the data from the last 7 days and finding their average. This way, it takes out the ups and downs that happen daily.

OPERATION ANALYTICS & INVESTIGATING METRIC SPIKE

CASE STUDY 1:

C: Calculate the percentage share of each language in the last 30 days?

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

language	percentage_sha...
English	11.86
Arabic	11.86
Persian	35.59
Hindi	16.95
French	11.86
Italian	11.86

By using this query, we will get the percentage share of each language in the **job_data** table.

D: Let's say you see some duplicate rows in the data. How will you display duplicates from the table?

```
SELECT actor_id, COUNT(*) AS duplicates
FROM job_data
GROUP BY actor_id
HAVING COUNT(*) > 1;
```

actor_id	duplicates
1001	11
1006	9
1003	10
1005	5
1002	7
1007	7
1004	10

When we execute this query, it will display the actor_ids that have duplicates in the job_data table, along with the count of occurrences for each actor_id.

OPERATION ANALYTICS & INVESTIGATING METRIC SPIKE

CASE STUDY 2:

A:Calculate the weekly user engagement?

```
SELECT
week_num AS week_number,
COUNT(DISTINCT user_id) AS
weekly_user_engagement
FROM
(
SELECT
user_id,
DATE_FORMAT(occurred_at, '%U') AS
week_num
FROM
events
WHERE
event_type = 'engagement'
) weekly_events
GROUP BY
week_number;
```

week_number	weekly_user_engageme...
17	87
18	197
19	211
20	198
21	210
22	233
23	227
24	258
25	249
26	233
27	252
28	241
29	253
30	275
31	232
32	282
33	292
34	304
35	20

B: Calculate the user growth for product?

```
SELECT
month as 'year_month',
new_users,
ROUND((new_users - LAG(new_users) OVER
(ORDER BY month)) / LAG(new_users) OVER
(ORDER BY month) * 100, 2) AS
growth_percentage
FROM
(
SELECT
DATE_FORMAT(created_at, '%Y-%m') AS
month,
COUNT(*) AS new_users
FROM
users
GROUP BY
DATE_FORMAT(created_at, '%Y-%m')
) user_growth
ORDER BY
'year_month';
```

year_month	new_users	growth_percentage
2013-01	332	NULL
2013-02	328	-1.20
2013-03	383	16.77
2013-04	410	7.05
2013-05	486	18.54
2013-06	485	-0.21
2013-07	608	25.36
2013-08	636	4.61
2013-09	699	9.91
2013-10	826	18.17
2013-11	816	-1.21
2013-12	972	19.12
2014-01	1083	11.42
2014-02	1054	-2.68
2014-03	1231	16.79
2014-04	1419	15.27
2014-05	1597	12.54
2014-06	1728	8.20
2014-07	1983	14.76
2014-08	1990	0.35

C: Calculate the weekly retention of users-sign up cohort?

```
SELECT
week_period, cohort_size, cohort_retained,
cohort_retained / cohort_size AS percent_retained
FROM ( SELECT
week_period, COUNT(DISTINCT user_id) AS cohort_size,
COUNT(DISTINCT CASE WHEN is_retained = 1 THEN
user_id END) AS cohort_retained
FROM ( SELECT
cohort.week_period, cohort.user_id,
MAX(occurred_at = cohort.activated_at) AS is_retained
FROM ( SELECT
user_id, activated_at,
DATE_FORMAT(activated_at, '%Y-%U') AS
week_period
FROM users
WHERE state = 'active'
) cohort
INNER JOIN ( SELECT
user_id, DATE_FORMAT(occurred_at, '%Y-%U') AS
week_period, MIN(occurred_at) AS occurred_at
FROM events
GROUP BY user_id, week_period
) cohort_2 ON cohort.user_id = cohort_2.user_id AND
cohort.week_period = cohort_2.week_period
GROUP BY cohort.week_period, cohort.user_id
) cohort_3
GROUP BY week_period
) cohort_4;
```

week_period	cohort_size	cohort_retained	percent_retained
2014-17	75	72	0.9600
2014-18	163	163	1.0000
2014-19	185	185	1.0000
2014-20	176	176	1.0000
2014-21	183	183	1.0000
2014-22	196	196	1.0000
2014-23	196	196	1.0000
2014-24	229	229	1.0000
2014-25	207	207	1.0000
2014-26	201	201	1.0000
2014-27	222	222	1.0000
2014-28	215	215	1.0000
2014-29	221	221	1.0000
2014-30	238	238	1.0000
2014-31	193	193	1.0000
2014-32	245	245	1.0000
2014-33	261	261	1.0000
2014-34	259	259	1.0000
2014-35	18	18	1.0000

OPERATION ANALYTICS & INVESTIGATING METRIC SPIKE

CASE STUDY 2:

D: Calculate the weekly engagement per device?

```
SELECT
DATE_FORMAT(occurred_at, '%Y-%U')
AS week,
device,
COUNT(*) AS weekly_engagement
FROM
events
WHERE
event_type = 'engagement'
GROUP BY
DATE_FORMAT(occurred_at, '%Y-%U'),
device
ORDER BY
week, device;
```

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	week	acer aspire desktop	acer aspire notebook	amazon fire phone	asus chromebook	dell inspiron desktop	dell inspiron notebook	hp pavilion desktop	htc one	ipad air	ipad mini	iphone 4s	iphone 5	iphone 5s	kindle fire	lenovo thinkpad	mac mini	macbook air	macbook pro	nexus 10	nexus 5	nexus 7	nokia lumia 635	samsung galaxy tablet	samsung galaxy note	samsung galaxy s4	windows surface
2	17	9	21	11	19	17	20	17	4	9	25	26	67	23	NULL	84	3	32	93	NULL	24	30	24	16	8	45	NULL
3	18	21	18	36	88	24	84	47	16	66	64	29	58	63	22	242	5	174	334	8	43	39	18	NULL	31	88	8
4	19	NULL	48	19	25	6	45	48	38	71	23	50	127	74	24	156	40	166	250	33	149	37	16	NULL	18	112	19
5	20	6	31	11	38	81	73	9	11	104	26	20	117	143	19	165	5	135	295	11	109	35	22	NULL	38	120	12
6	21	20	14	NULL	94	40	67	38	33	55	61	59	215	73	20	243	3	99	243	45	124	11	5	11	24	69	19
7	22	NULL	62	8	77	59	112	34	17	85	31	55	88	99	56	215	12	204	256	77	106	82	78	11	NULL	79	26
8	23	50	68	23	149	72	67	122	3	48	34	58	188	95	34	174	14	188	223	23	90	32	24	21	NULL	65	14
9	24	45	41	10	73	67	74	59	25	65	43	94	219	100	23	278	21	206	328	27	98	52	6	10	6	113	45
10	25	12	56	4	64	64	119	33	27	96	34	54	183	63	6	247	15	117	283	30	70	59	43	29	NULL	113	37
11	26	24	13	17	59	43	98	75	9	64	19	70	220	143	31	218	24	150	338	21	48	87	73	6	25	80	41
12	27	32	71	20	53	51	70	36	28	31	22	130	200	113	37	181	40	116	378	43	87	59	9	43	15	111	36
13	28	30	61	5	40	61	107	65	44	76	22	53	189	69	39	196	20	162	281	23	78	30	20	NULL	12	200	37
14	29	16	13	16	68	47	96	27	29	70	75	46	147	80	29	220	9	192	249	34	61	33	45	3	12	115	20
15	30	60	85	17	62	45	108	54	7	132	37	109	188	100	5	292	44	94	327	41	118	34	49	15	18	144	15
16	31	25	51	NULL	42	51	108	22	5	100	28	87	268	53	7	161	4	158	330	30	75	26	47	29	15	110	25
17	32	24	49	45	76	22	144	40	20	75	43	40	182	147	22	188	9	99	495	43	51	26	13	14	18	91	6
18	33	32	79	26	34	59	131	33	24	42	88	62	234	106	27	147	50	117	405	39	68	37	22	22	50	148	29
19	34	26	79	38	129	67	67	57	63	86	20	67	107	157	14	277	38	141	366	55	100	86	14	8	2	88	60
20	35	NULL	7	NULL	NULL	4	35	NULL	6	NULL	8	24	NULL	NULL	NULL	22	NULL	12	21	NULL	4	NULL	5	NULL	NULL	NULL	8
21																											

E: Calculate the email engagement metrics

```
SELECT
week as Year_Week,
num_emails_sent as Total_Emails_Sent,
num_emails_opened as
Total_Emails_Opened,
num_emails_clicked as
Total_Emails_Clicked,
100.0 * num_emails_opened /
num_emails_sent AS Email_Opening_Rate,
100.0 * num_emails_clicked /
num_emails_sent AS Email_Clicking_Rate
FROM (
SELECT
DATE_FORMAT(occurred_at, '%Y-%U')
AS week,
COUNT(CASE WHEN action IN ('sent
weekly_digest', 'sent_reengagement_email')
THEN 1 END) AS num_emails_sent,
COUNT(CASE WHEN action =
'email_open' THEN 1 END) AS
num_emails_opened,
COUNT(CASE WHEN action =
'email_clickthrough' THEN 1 END) AS
num_emails_clicked
FROM
email_events
WHERE
action IN ('sent weekly_digest',
'sent_reengagement_email', 'email_open',
'email_clickthrough')
GROUP BY
DATE_FORMAT(occurred_at, '%Y-%U')
) Email_Eng;
```

Year_Week	Total_Emails_Sent	Total_Emails_Open...	Total_Emails_Clicked	Email_Opening_Rate	Email_Clicking_Rate
2014-22	192	987	488	514.06250	254.16667
2014-23	197	1075	538	545.68528	273.09645
2014-24	226	1155	554	511.06195	245.13274
2014-30	231	1383	630	598.70130	272.72727
2014-33	264	1432	490	542.42424	185.60606
2014-19	173	972	477	561.84971	275.72254
2014-20	191	1004	507	525.65445	265.44503
2014-25	196	1096	530	559.18367	270.40816
2014-26	219	1165	556	531.96347	253.88128
2014-32	200	1337	418	668.50000	209.00000
2014-27	213	1228	621	576.52582	291.54930
2014-34	261	1528	490	585.44061	187.73946
2014-31	222	1351	445	608.55856	200.45045
2014-29	213	1219	590	572.30047	276.99531
2014-18	157	912	430	580.89172	273.88535
2014-21	164	1014	443	618.29268	270.12195
2014-28	213	1250	599	586.85446	281.22066
2014-17	73	310	166	424.65753	227.39726
2014-35	48	41	38	85.41667	79.16667

OPERATION ANALYTICS & INVESTIGATING METRIC SPIKE

RESULT

This project helped me a lot in understanding more about SQL. My learnings started from creating the database, then how to load the data where I spent a lot of time, because I kept getting error. Then eventually changing the user_type in events table from int to text and altering all the columns which contained dates from text to datetime. This all was new to me, so it helped me in learning new things, which although took a lot of time but this effort will help me in future.

It was also my first time in analysing such a big dataset.

Throughout this project, I have learned the importance of operational analytics and how it can empower businesses with valuable insights.

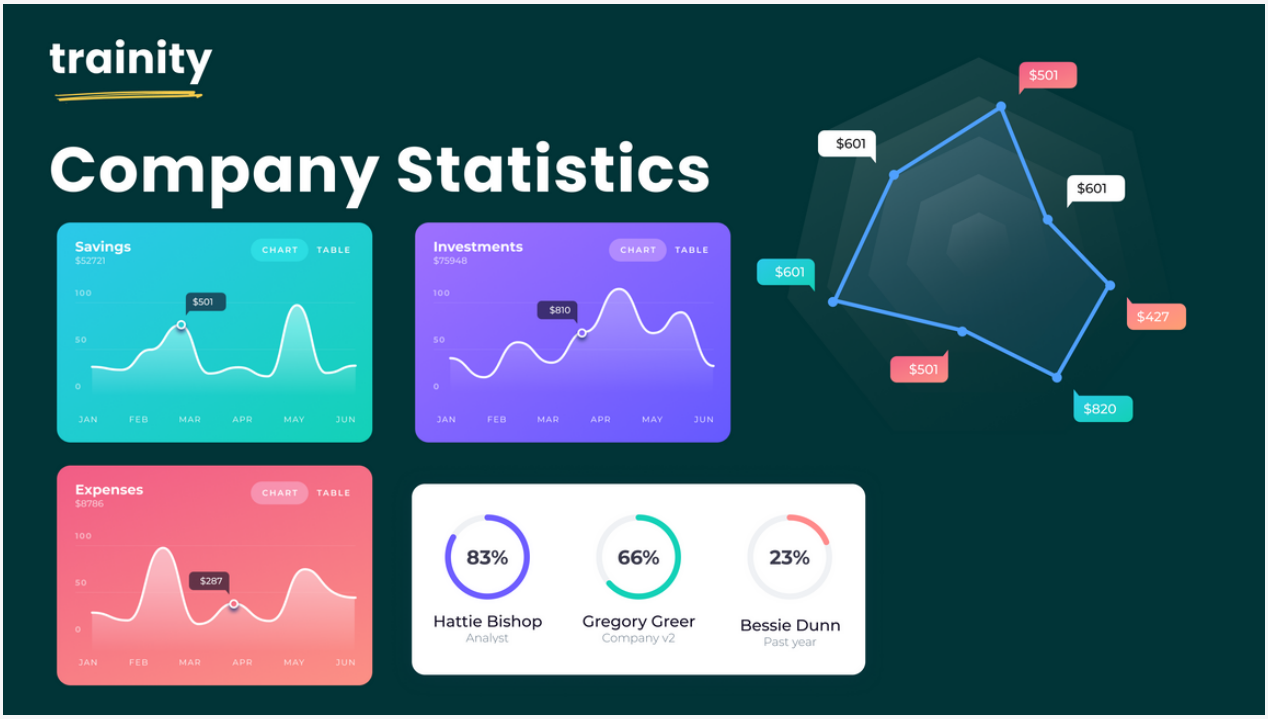
PROJECT LINK

https://drive.google.com/file/d/1dgzw4ppQnDiVJVrUwQNvBSmP42rc1DNi/view?usp=drive_link

MODULE 4

HIRING PROCESS ANALYTICS

STATISTICS



HIRING PROCESS ANALYTICS

PROJECT DESCRIPTION:

The project aims to analyse the hiring process data of a multinational company (MNC) to gain insights into their recruitment trends and patterns. The dataset provided information such as the interview status, gender, department, post name, and offered salary of the applicants. The project's goal wass to answer various questions like the count of males and females hired, the average salary offered, class intervals for salary, and visualizing data using pie charts and bar graphs.

APPROACH:

To execute the project, I first familiarized myself with the dataset and identified relevant columns for analysis. I used exploratory data analysis (EDA) and drew necessary conclusions about the company's hiring process. Then, I used Excel's formulas to calculate the required statistics, such as average salary and frequency of post tiers. I visualized the data using pie charts, histogram and bar charts.

TECH-STACK USED:

Microsoft Excel for Mac Version 16.74.

INSIGHTS:

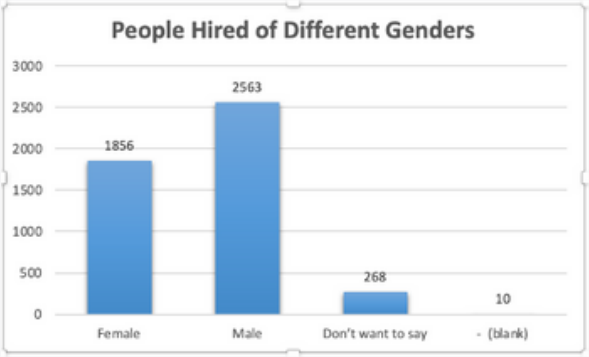
A. Hiring Analysis - Determine the gender distribution of hires. How many males and females have been hired by the company?

FEMALE	1856
MALE	2563
DON'T WANT TO SAY	268
- (BLANK)	10

How I found out?

1.Filter the data based on status column to "Hired"

2.Then filter the "event_name" coloum to get the numbers for Female, Male, Don't want to say and the people who left it blank



HIRING PROCESS ANALYTICS

B. Salary Analysis - What is the average salary offered by this company? Use Excel functions to calculate this.

Excel formula Used

=AVERAGE(G2:G7167)

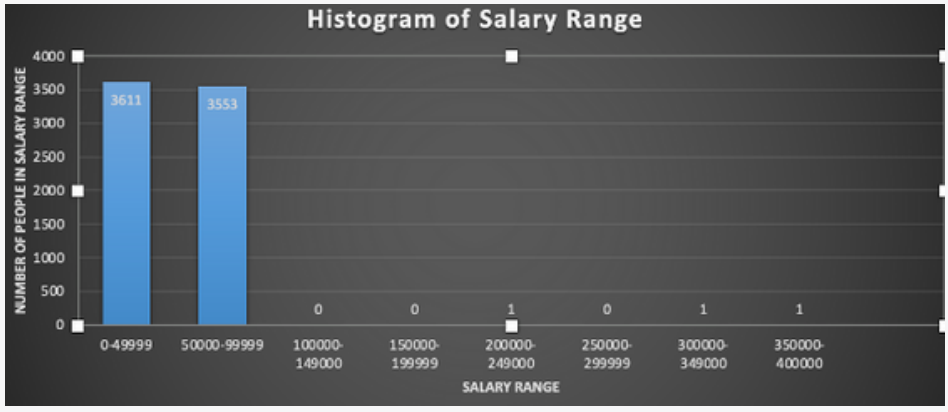
Service Department	i7	16756
Service Department	i7	30952
Service Department	c9	64150
Service Department	c9	40152
Service Department	c9	49282
Service Department	c5	57742
Service Department	c5	69932
Service Department	c5	14489
Operations Department	c5	54201
		49983.02902

The average salary offered by the company is **$49,983.03**

C. Salary Distribution - Create class intervals for the salaries in the company. This will help you understand the salary distribution.

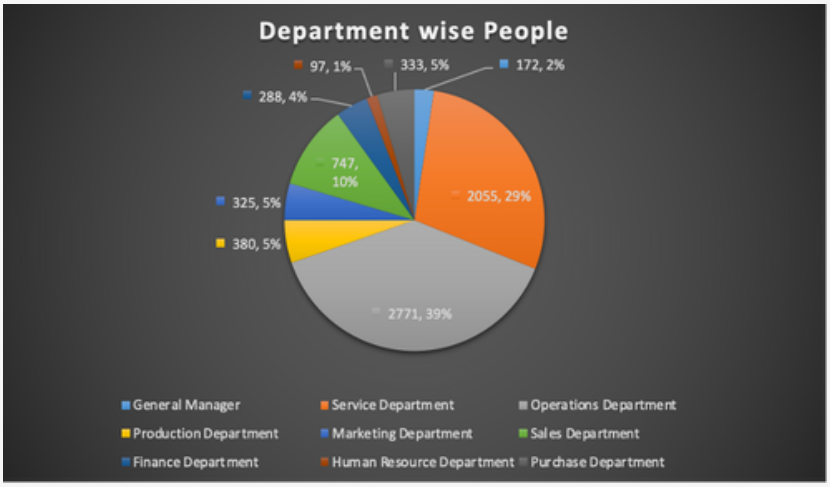
Excel formula Used

=FREQUENCY(Table1[Offered Salary],K21)
-FREQUENCY(Table1[Offered Salary],J21)



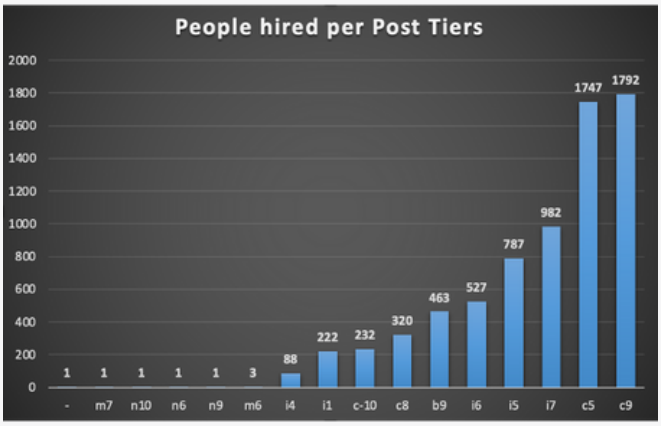
D. Departmental Analysis - Use a pie chart, bar graph, or any other suitable visualization to show the proportion of people working in different departments.

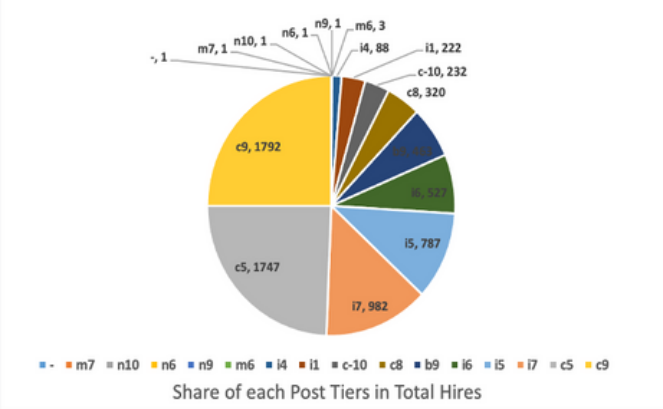
Department	No. of Employees Working
General Manager	172
Service Department	2055
Operations Department	2771
Production Department	380
Marketing Department	325
Sales Department	747
Finance Department	288
Human Resource Department	97
Purchase Department	333



HIRING PROCESS ANALYTICS

E. Position Tier Analysis - 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.





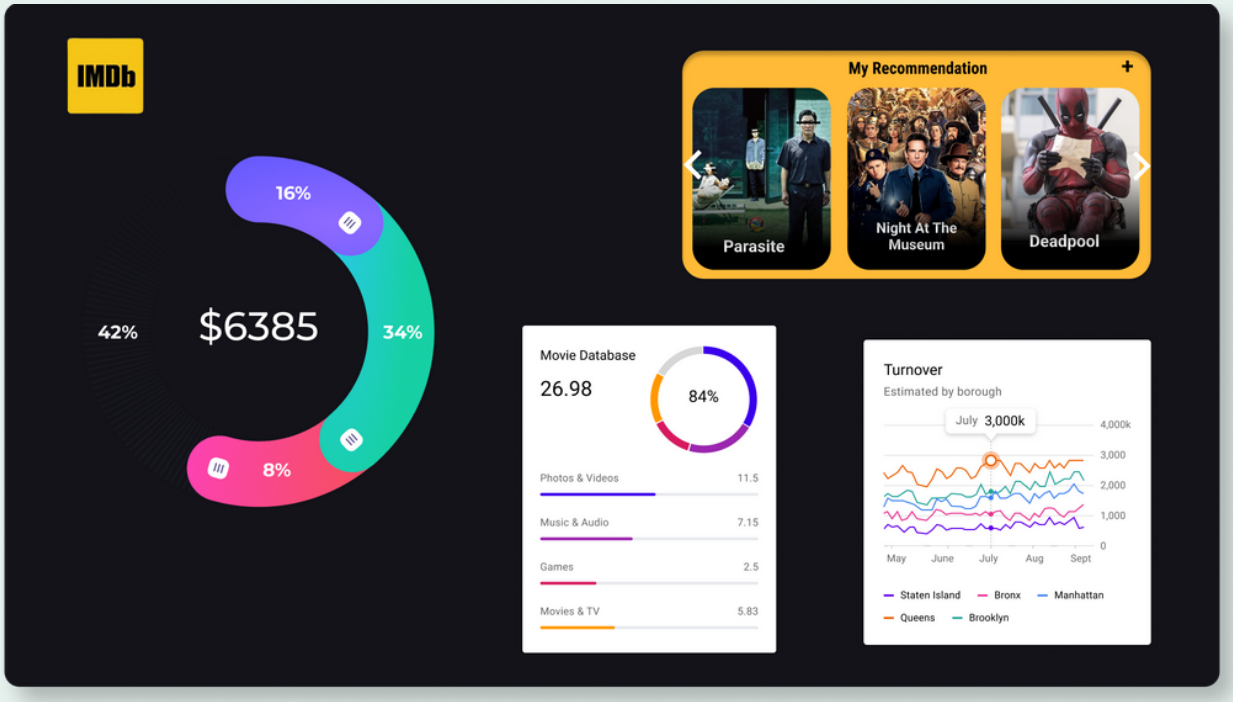
Posts	-	m7	n10	n6	n9	m6	i4	i1	c-10	c8	b9	i6	i5	i7	c5	c9
Frequency	1	1	1	1	1	3	88	222	232	320	463	527	787	982	1747	1792

RESULT

By completing this project, I learned how to perform data analysis on real data. It helped me in answering the questions which are asked by the company in analysing the Hiring Process. I got meaningful insights from the hiring process data, which will assist the company's HR department in making informed decisions.
These data analysis results can help in uncovering valuable information that can drive business decisions and improve organizational efficiency.

MODULE 5

IMDB MOVIE ANALYSIS



IMDB MOVIE ANALYSIS

PROJECT DESCRIPTION:

The objective of this project is to analyse the IMDB movie dataset and derive insights from it. This dataset has several columns, and we need to manipulate the data using Excel formulas.

Various tasks are required to be completed such as determining the most common genres, analysing the impact of genres on IMDB scores, examining the distribution of movie durations, identifying top directors based on their average IMDB scores, exploring the relationship between movie budgets and gross earnings, and calculate profit margins for each movie.

APPROACH:

To execute the project, I first familiarized myself with the dataset and identified relevant columns for analysis.

I then cleaned the data and removed unnecessary columns and rows which had null values, or which were not relevant for our tasks.

Before Cleaning

Columns – 28
Rows - 5044

After Cleaning

Columns – 9
Rows - 3818

TECH-STACK USED:

Microsoft Excel for Mac Version 16.74.

INSIGHTS:

A. Movie Genre Analysis - 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.

Genres	Count	Mean	Median	Mode	Range	Variance	Standard Deviation
Drama	1922	6.790426639	6.9	6.7	7.2	0.79110039	0.889438244
Comedy	1498	6.187583445	6.3	6.7	6.9	1.07431466	1.036491517
Thriller	1115	6.378295964	6.4	6.5	6.3	0.93313713	0.965990233
Action	957	6.293103448	6.3	6.1	6.9	1.06401097	1.031509072
Romance	874	6.432837529	6.5	6.5	6.4	0.9328609	0.965847243
Adventure	783	6.458109834	6.6	6.7	6.6	1.23305115	1.110428365
Crime	711	6.545428973	6.6	6.6	6.9	0.9585671	0.979064403
Fantasy	512	6.293164063	6.4	6.7	6.7	1.27664594	1.129887578
Sci-Fi	497	6.322736419	6.4	6.7	6.9	1.33655863	1.156096287
Family	448	6.213616071	6.3	6.7	6.7	1.35249875	1.162969798
Horror	391	5.926086957	6	5.9	6.3	0.99418952	0.997090528
Mystery	382	6.478534031	6.5	6.6	5.5	1.01029916	1.005136388
Biography	241	7.151037344	7.2	7	4.4	0.48184267	0.694148881
Animation	198	6.702525253	6.8	6.7	5.8	0.98146567	0.990689493
Music	158	6.463562753	6.7	6.2	6.9	1.41460189	1.189370375
War	155	7.070967742	7.1	7.1	4.3	0.65545036	0.809598886
Sport	151	6.603311258	6.8	7.2	6.4	1.08858896	1.043354668
History	148	7.160135135	7.2	7.7	3.4	0.44282175	0.665448533
Musical	103	6.559223301	6.7	7.1	6.4	1.30185037	1.140986578
Documentary	58	7.017241379	7.3	6.6	6.9	1.5923291	1.261875231
Western	57	6.812280702	6.8	6.8	4.2	0.88573935	0.941137263
Western	57	6.812280702	6.8	6.8	4.2	0.88573935	0.941137263
Short	2	6.8	6.8	#N/A	0.6	0.18	0.424264069
Film-Noir	1	7.7	7.7	#N/A	0	#DIV/0!	#DIV/0!

Here I have divided the Genre Column into different columns based on " | " .

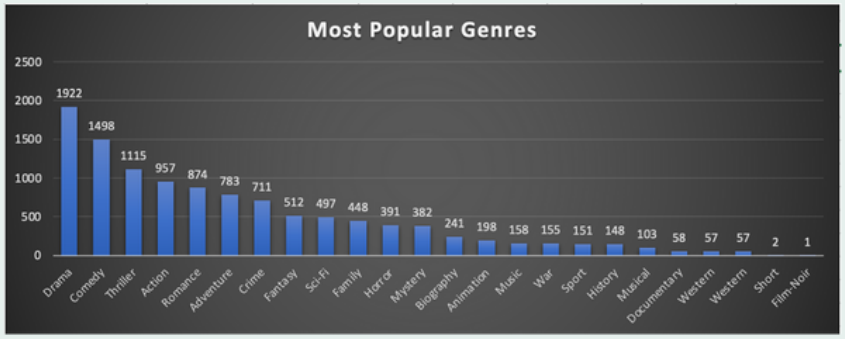
I used the function Text to Columns, which is under Data Tab

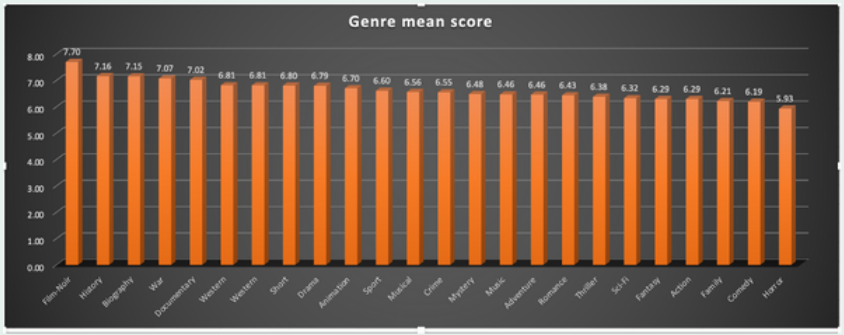
Then I extracted the Unique Genres and then performed the following descriptive analysis as shown in the table on the left. Following Excel Formulas were used :

IMDB MOVIE ANALYSIS

FINDINGS

Genres	Count	Mean
Drama	1922	6.790426639
Comedy	1498	6.187583445
Thriller	1115	6.378295964
Action	957	6.293103448
Romance	874	6.432837529
Adventure	783	6.458109834
Crime	711	6.545428973
Fantasy	512	6.293164063
Sci-Fi	497	6.322736419
Family	448	6.213616071
Horror	391	5.926086957
Mystery	382	6.478534031
Biography	241	7.151037344
Animation	198	6.702525253
Music	158	6.463562753
War	155	7.070967742
Sport	151	6.603311258
History	148	7.160135135
Musical	103	6.559223301
Documentary	58	7.017241379
Western	57	6.812280702
Western	57	6.812280702
Short	2	6.8
Film-Noir	1	7.7





•**Drama, Comedy, and Thriller are the most common genres**, with a high number of movies in each genre.

•**History, Biography, War and Documentary, have higher Mean IMDB scores**. This indicates that movies in these genres generally receive better ratings from viewers.

•**Horror and Comedy have lower Mean IMDB scores**, which indicates that movies in these genres might receive relatively lower ratings.

•**Adventure and Fantasy genres have a wider range of IMDB scores**, meaning that some movies in these genres have very high ratings, while others have lower ratings.

•**Drama and Romance**, **have similar mean, median, and mode scores**, showing a relatively balanced distribution of ratings.

IMDB MOVIE ANALYSIS

B. Movie Duration Analysis - Analyse the distribution of movie durations and identify the relationship between movie duration and IMDB score

duration	imdb_score	movie_title
142	9.3	The Shawshank Redemption¬†
175	9.2	The Godfather¬†
152	9	The Dark Knight¬†
220	9	The Godfather: Part II¬†
192	8.9	The Lord of the Rings: The Return of the King¬†
185	8.9	Schindler's List¬†
178	8.9	Pulp Fiction¬†
142	8.9	The Good, the Bad and the Ugly¬†
148	8.8	Inception¬†
171	8.8	The Lord of the Rings: The Fellowship of the Ring¬†
151	8.8	Fight Club¬†
142	8.8	Forrest Gump¬†
127	8.8	Star Wars: Episode V - The Empire Strikes Back¬†
172	8.7	The Lord of the Rings: The Two Towers¬†
136	8.7	The Matrix¬†
146	8.7	Goodfellas¬†
125	8.7	Star Wars: Episode IV - A New Hope¬†
133	8.7	One Flew Over the Cuckoo's Nest¬†
135	8.7	City of God¬†
202	8.7	Seven Samurai¬†
169	8.6	Interstellar¬†
169	8.6	Saving Private Ryan¬†
127	8.6	Se7en¬†
138	8.6	The Silence of the Lambs¬†
125	8.6	Spirited Away¬†
101	8.6	American History X¬†
106	8.6	The Usual Suspects¬†
87	8.6	Modern Times¬†
164	8.5	The Dark Knight Rises¬†
171	8.5	Gladiator¬†
153	8.5	Terminator 2: Judgment Day¬†
165	8.5	Django Unchained¬†
151	8.5	The Departed¬†

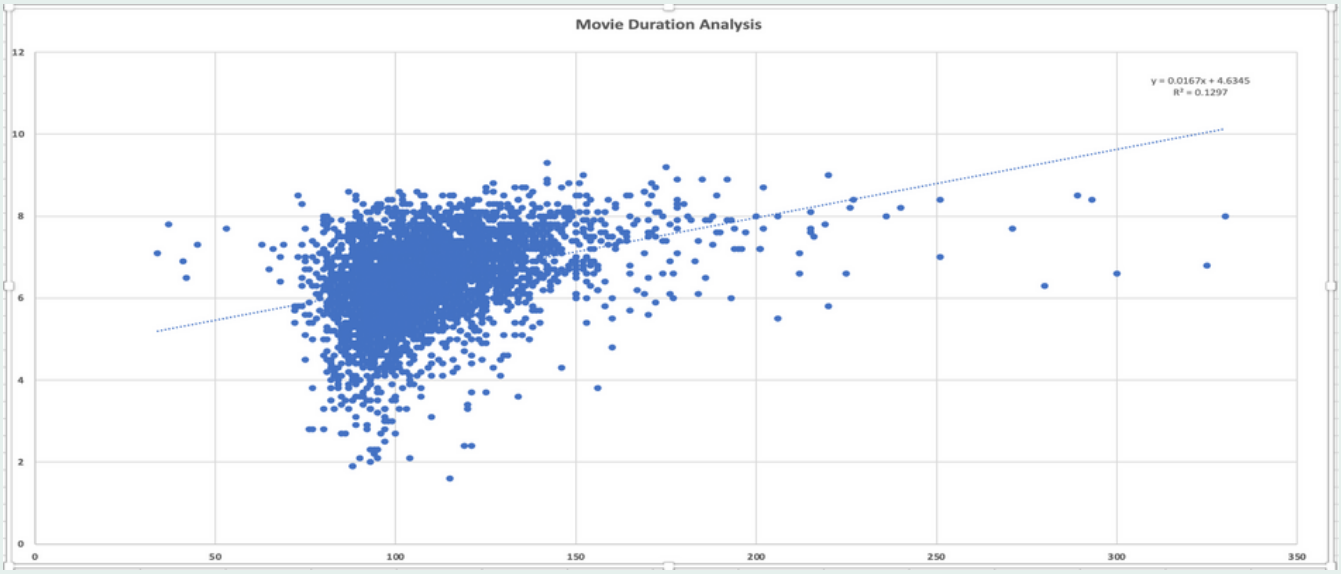
For Movie Duration Analysis Duration and IMDB_Score Columns from the dataset are chosen for analysis

	Mean	Median	Standard Deviation
Duration	110.01	106	22.79
IMDB Score	6.47	6.6	1.05

From the above statistics, we can observe the following:

- The Mean and Median of Duration are relatively close, suggesting that the distribution of movie durations is roughly symmetric.
- The Standard Deviation for movie durations is larger than the standard deviation for IMDb scores, indicating that movie durations have more variability compared to IMDb scores.

IMDB MOVIE ANALYSIS



To determine a relationship between movie duration and IMDb score, I plotted movie duration (x) and IMDb score (y) on a scatter plot and added the trendline to observe the direction and strength of the relationship visually.

The equation (y = 0.0167x + 4.6345) shows a positive linear relationship between movie duration and IMDb score. However, the coefficient of determination- R^2 value (0.1297) indicates that only 12.97% of the variation in IMDb scores can be explained by the variation in movie durations.

This suggests that there is a **weak positive relationship** between the duration of the movie and its IMDB Score, what it means it that the duration of the movie is not a big factor in the increase or decrease of their IMDB Scores, other factors also influence IMDb scores.

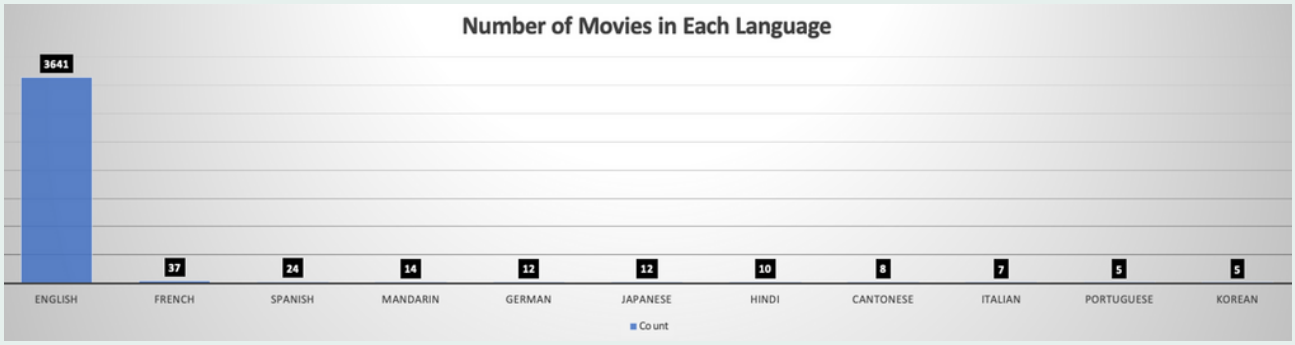
C. Language Analysis - Determine the most common languages used in movies and analyse their impact on the IMDB score using descriptive statistics.

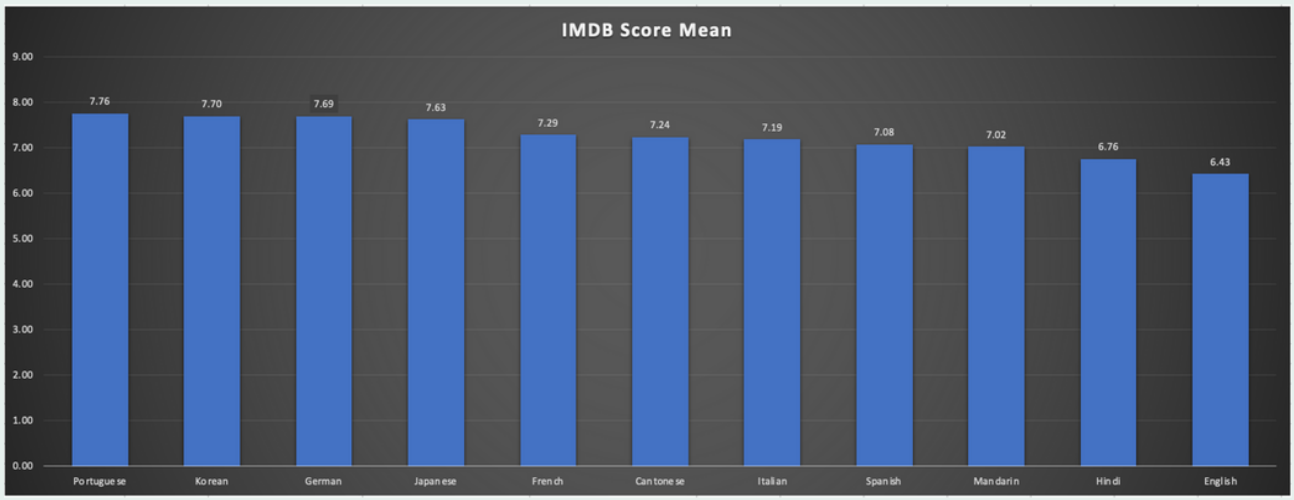
Language	Count	IMDB Score Mean	IMDB Score Median	IMDB Score Std Dev
English	3641	6.43	6.5	1.048477755
French	37	7.29	7.2	0.561328861
Spanish	24	7.08	7.15	0.841829874
Mandarin	14	7.02	7.25	0.765786244
German	12	7.69	7.75	0.669407246
Japanese	12	7.63	7.8	0.899621132
Hindi	10	6.76	7.05	1.111755369
Cantonese	8	7.24	7.3	0.440575922
Italian	7	7.19	7	1.155318962
Portuguese	5	7.76	8	0.978774744
Korean	5	7.70	7.7	0.570087713
Norwegian	4	7.15	7.3	0.574456265
Persian	3	8.13	8.4	0.550757055
Danish	3	7.90	8.1	0.529150262
Dutch	3	7.57	7.8	0.404145188
Thai	3	6.63	6.6	0.450924975
Indonesian	2	7.90	7.9	0.424264069
Hebrew	2	7.65	7.65	0.494974747
Dari	2	7.50	7.5	0.141421356
Aboriginal	2	6.95	6.95	0.777817459
Telugu	1	8.40	8.4	#DIV/0!
Romanian	1	7.90	7.9	#DIV/0!
Maya	1	7.80	7.8	#DIV/0!
Swedish	1	7.60	7.6	#DIV/0!
Dzongkha	1	7.50	7.5	#DIV/0!
Czech	1	7.40	7.4	#DIV/0!
Vietnamese	1	7.40	7.4	#DIV/0!
Mongolian	1	7.30	7.3	#DIV/0!
Zulu	1	7.30	7.3	#DIV/0!
Arabic	1	7.20	7.2	#DIV/0!
Aramaic	1	7.10	7.1	#DIV/0!
Hungarian	1	7.10	7.1	#DIV/0!
Icelandic	1	6.90	6.9	#DIV/0!
Filipino	1	6.70	6.7	#DIV/0!
Russian	1	6.50	6.5	#DIV/0!
Kazakh	1	6.00	6	#DIV/0!
Bosnian	1	4.30	4.3	#DIV/0!

- Languages with a larger number of movies have more reliable statistics, as the data is more representative of the overall movie population that is why for our analysis, **we have taken only those movies whose language count is equal to or more than 5 movies per language.**
- ·**Mean and Median of IMDB Score -** helps us understand the overall reception of movies in each language. Languages with a high number of movies and relatively high IMDb Score Means and Medians indicates that these languages have consistently produced well-received movies.
- ·**Standard Deviation of IMDb Score -** higher standard deviations suggests that there are many good and bad movies in these languages

IMDB MOVIE ANALYSIS

FINDINGS





- English has the most movies in the dataset, with a mean IMDB score of 6.43 and a median IMDB score of 6.5
- French has the second-highest count with a mean IMDB score of 7.29 and a median IMDB score of 7.2
- Spanish has the third highest count, with a mean IMDB score of 7.08 and a median IMDB score of 7.15
- Portuguese, Korean and German language movies have the highest Mean IMDB Score, but the number of movies in each language is less compared to English

IMDB MOVIE ANALYSIS

D. Director Analysis - Identify the top directors based on their average IMDB score and analyse their contribution to the success of movies using percentile calculations.

Director Names	Average IMDB SCORE	Number of Movies	Percentile
Christopher Nolan	8.4	8	95 Percentile
Quentin Tarantino	8.2	8	95 Percentile
James Cameron	7.9	7	95 Percentile
Alejandro G. IV±V"rritu	7.8	5	95 Percentile
David Fincher	7.8	10	95 Percentile
Martin Scorsese	7.7	16	90 Percentile
Peter Jackson	7.7	12	90 Percentile
Francis Ford Coppola	7.7	9	90 Percentile
Wes Anderson	7.6	7	90 Percentile
Paul Greengrass	7.6	7	90 Percentile
Brad Bird	7.6	5	90 Percentile
Steven Spielberg	7.5	25	90 Percentile
Paul Thomas Anderson	7.5	6	90 Percentile
Sam Mendes	7.5	8	90 Percentile
Darren Aronofsky	7.5	6	80 Percentile
Danny Boyle	7.4	8	80 Percentile
Alexander Payne	7.4	5	80 Percentile
George Lucas	7.4	5	80 Percentile
John Lasseter	7.4	5	80 Percentile
Mike Leigh	7.4	5	80 Percentile
Jean-Pierre Jeunet	7.3	5	80 Percentile
Terry Gilliam	7.3	7	80 Percentile
Richard Linklater	7.3	11	80 Percentile
Edward Zwick	7.3	8	80 Percentile
Robert Zemeckis	7.3	13	80 Percentile
Bryan Singer	7.3	8	80 Percentile
Ang Lee	7.3	8	80 Percentile
Marc Forster	7.2	7	80 Percentile
Clint Eastwood	7.2	19	80 Percentile
James Wan	7.2	7	80 Percentile
Jason Reitman	7.2	6	80 Percentile
Zack Snyder	7.2	8	80 Percentile
David O. Russell	7.2	7	80 Percentile

EXCEL FORMULAS USED

To get a Unique Directors Name-

=UNIQUE(Table4[director_name], FALSE,FALSE)

To get the Mean IMDB Score of Each Director-

=AVERAGEIF($A$2:$A$3818, F2, $B$2:$B$3818)

To calculate Percentile-

=PERCENTILE($B$2:$B$3818, 0.95)

=PERCENTILE($B$2:$B$3818, 0.95

	G	H
	Total Directors	1724
	95 Percentile	7.7
	90 Percentile	7.5
	80 Percentile	7.1
	70 Percentile	6.9
	60 Percentile	6.7
	50 Percentile	6.5
	40 Percentile	6.2
	30 Percentile	5.95
	20 Percentile	5.6
	10 Percentile	5.1
	5 Percentile	4.4
	1 Percentile	3.223
	0.5 Percentile	2.8

To get the number of Movies for each Director-

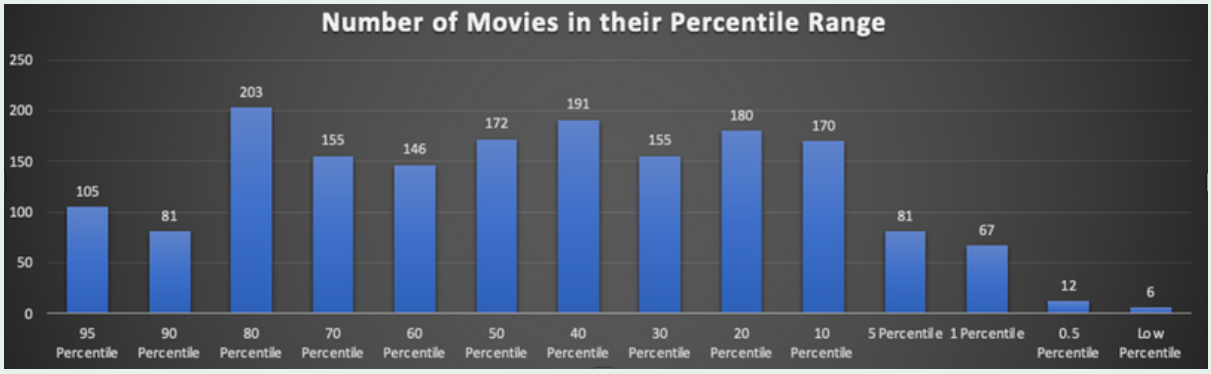
=COUNTIF(Table4[director_name],F2)

To automatically populate Percentile column according to their Average IMDB Score-

=IF(B2>=$H$2,$G$2,IF(B2>=$H$3, $G$3,IF(B2>=$H$4, $G$4,IF(B2>=$H$5, $G$5,IF(B2>=$H$6, $G$6,IF(B2>=$H$7, $G$7,IF(B2>=$H$8, $G$8,IF(B2>=$H$9, $G$9,IF(B2>=$H$10, $G$10,IF(B2>=$H$11, $G$11,IF(B2>=$H$12, $G$12,IF(B2>=$H$13, $G$13,IF(B2>=$H$14, $G$14,"Low Percentile")))))))))))))

For our analysis I have taken those directors who have number of movies equal to or greater than 5

Bin Limit	Bin Label	Bin Counts
7.7	95 Percentile	105
7.5	90 Percentile	81
7.1	80 Percentile	203
6.9	70 Percentile	155
6.7	60 Percentile	146
6.5	50 Percentile	172
6.2	40 Percentile	191
5.95	30 Percentile	155
5.6	20 Percentile	180
5.1	10 Percentile	170
4.4	5 Percentile	81
3.223	1 Percentile	67
2.8	0.5 Percentile	12
	Low Percentile	6



We see the **TOP 25 Directors** in our list and the number of movies in their Percentile Range.

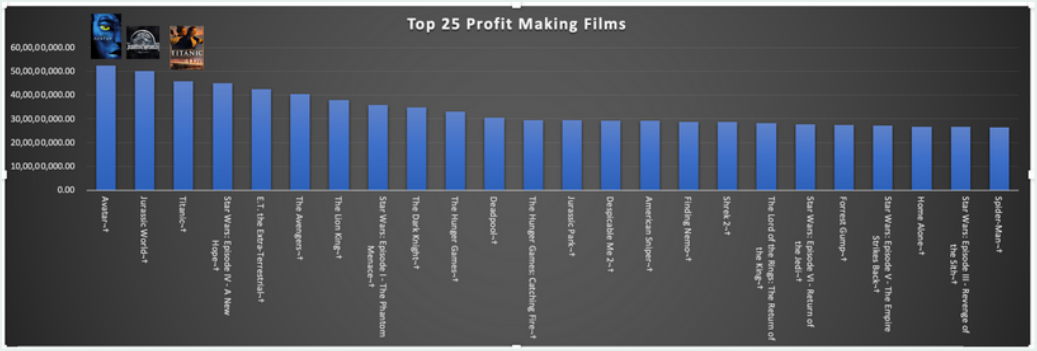
IMDB MOVIE ANALYSIS

E. Budget Analysis - Analyse the correlation between movie budgets and gross earnings and identify the movies with the highest profit margin.

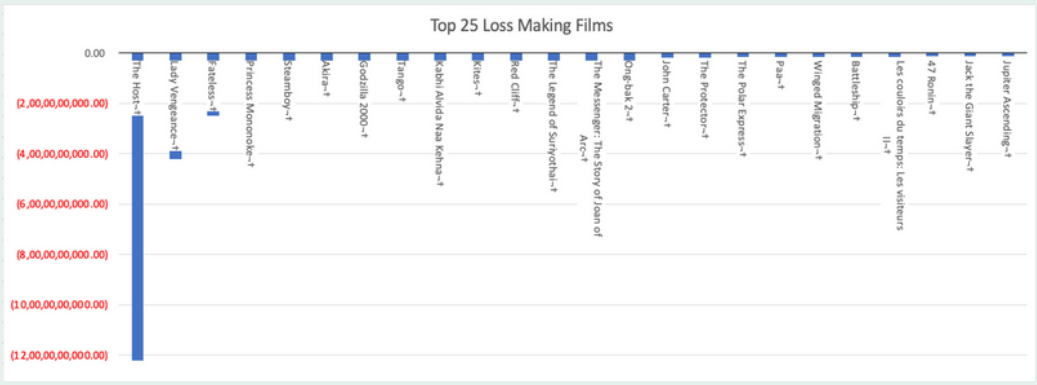
Correlation Coefficient	0.1000617

A correlation coefficient of 0.1000617 indicates a weak positive correlation between Movie Budgets and Gross Earnings.

This means that there is a slight tendency for movies with higher budgets to have higher gross earnings, however there are many movies with high budgets that did not have high gross earnings, and vice versa.



Top 25 Profit Making Movies in our Dataset



Bottom 25 Loss Making Movies in our Dataset

RESULT

The insights that I got are:

- **Movie Genre Analysis**- Drama, Comedy and Thriller are the most common genres
- **Movie Duration Analysis**- Shows that there is a weak positive relationship between the duration of the movie and its IMDB Score and other factors also influence IMDb scores
- **Language Analysis**- English, French and Spanish are the most common languages in our IMDB dataset
- **Director Analysis**- Christopher Nolan, Quentin Tarantino and James Cameron are the top 3 directors who are above the 95 percentile range and have the highest Mean IMDB Score
- **Budget Analysis**- Avatar, Jurassic World and Titanic are the top 3 movies with the highest Profit Margins

This analysis has contributed to a better understanding of the factors influencing movie ratings and financial success, and it has provided valuable information for further exploration in the realm of movie analysis.

IMDB MOVIE ANALYSIS

WORKING EXCEL FILES

I have used these two Excel files to do my analysis, reason for dividing the Excel files into two files was, that my system was hanging a lot when I was trying to do the analysis in just one file.

The links of these files are:

- https://docs.google.com/spreadsheets/d/1AIxcFU-PB3QepWV_ZjZWSRxYgen0f6bv/edit?usp=sharing&ouid=109466755193972209405&rtpof=true&sd=true
- https://docs.google.com/spreadsheets/d/15pi7R1bcRgHN0OUxFtUBvBsc4SVm454v/edit?usp=sharing&ouid=109466755193972209405&rtpof=true&sd=true
- https://docs.google.com/spreadsheets/d/1VHInwRA8T787v9x_VETPmYuBhoGbM86Jj-bYPnQ_-Dc/edit?usp=sharing_

The video file link is :

- https://screenpal.com/watch/c0ivqfVkuWS

MODULE 6

BANK LOAN ANALYSIS

BANK LOAN ANALYSIS

PROJECT DESCRIPTION:

In this project, we're diving into the world of ”loan defaults” using Exploratory Data Analysis (EDA).

Our main goal is to give loans to good applicants and not turn them away.

A finance company has two big risks:

- They might say "no" to someone who deserves the loan, and that means losing business.
- They could give a loan to someone who might not be able to pay it back, and that's a financial risk.

The data we're working with has two different situations:

- People who are having a hard time making their payments. This means they're paying late more often, especially in the beginning.
- People who are making their payments on time.

The objective of this project is to use EDA to comprehend how customer characteristics and loan-related factors impact the probability of loan default.

APPROACH:

To execute the project, I first familiarized myself with the dataset and identified relevant columns for analysis. I then cleaned the data by removing duplicates and unnecessary rows and columns. I then Imputed the missing cells by using Median and Mode Imputation methods. I then analysed and visualized the dataset and ultimately gathered Insights.

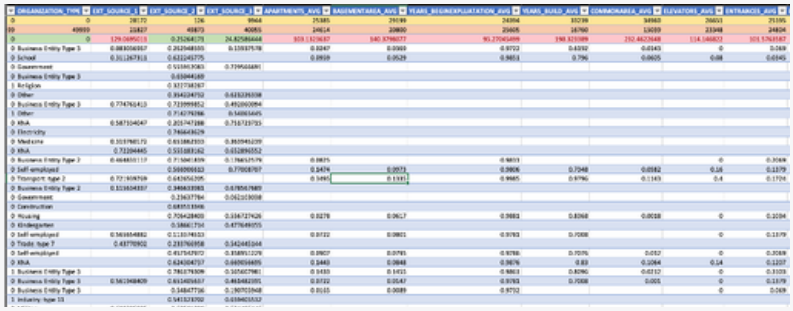
TECH-STACK USED:

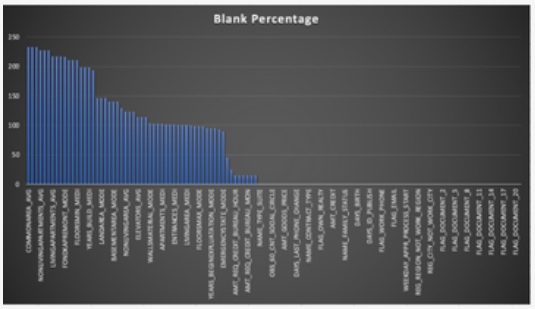
Microsoft Excel for Mac Version 16.74.

BANK LOAN ANALYSIS

INSIGHTS:

A. Identify Missing Data and Deal with it appropriately- Identify the missing data in the dataset and decide on an appropriate method to deal with it.





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	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
100031	1	Cash loans	F	N	Y	0	112500	979992	27076.5	702000	Unaccompanied
100032	0	Cash loans	M	N	Y	1	112500	327024	23827.5	270000	Family
100033	0	Cash loans	M	Y	Y	0	270000	790830	57676.5	675000	Unaccompanied
100034	0	Revolving loans	M	N	Y	0	90000	180000	9000	180000	Unaccompanied
100035	0	Cash loans	F	N	Y	0	292500	665892	24592.5	477000	Unaccompanied
100036	0	Cash loans	F	N	Y	0	112500	512064	25033.5	360000	Family
100037	0	Cash loans	F	N	N	0	90000	199008	20893.5	180000	Unaccompanied

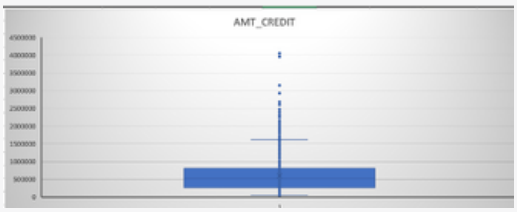
Steps performed to clean our working Dataset-

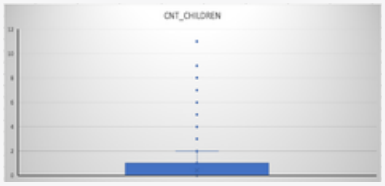
- **Before Cleaning**- Our Dataset has 122 Columns, some columns have 50,003 rows while others have 307478 Rows
- For my analysis, to make the data uniform across all the columns I took the first 50,003 rows and deleted the other rows, made it into a table and then removed Duplicates which in under the **Table** Tab
- I then calculated the Percentage of Missing Values in each Column and whichever columns had a Blank Percentage of more than 30% I deleted those columns from the Table
- For all the columns which had a Blank Percentage of less than 30% but had missing cells, I imputed values in those cells for Numerical Columns, I used **Median Imputation,** for Text Columns, I used **Mode Imputation**
- **After Cleaning and Imputation**- Our Dataset has 73 Columns and 50,000 Rows
- This completes the first task, Here I have identified the missing data in the dataset and used Median and Mode Imputation to deal with it.

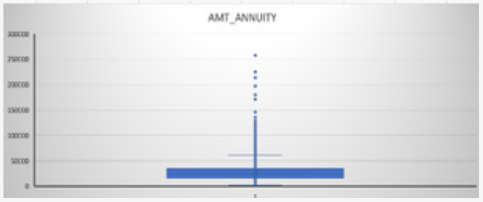
BANK LOAN ANALYSIS

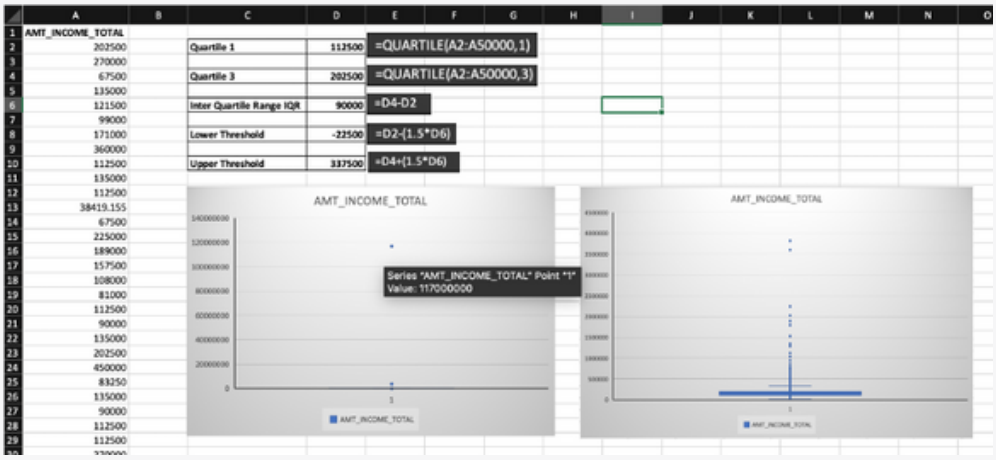
INSIGHTS:

B. Identify Outliers in the Dataset- Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.









Steps performed to identify the outliers-

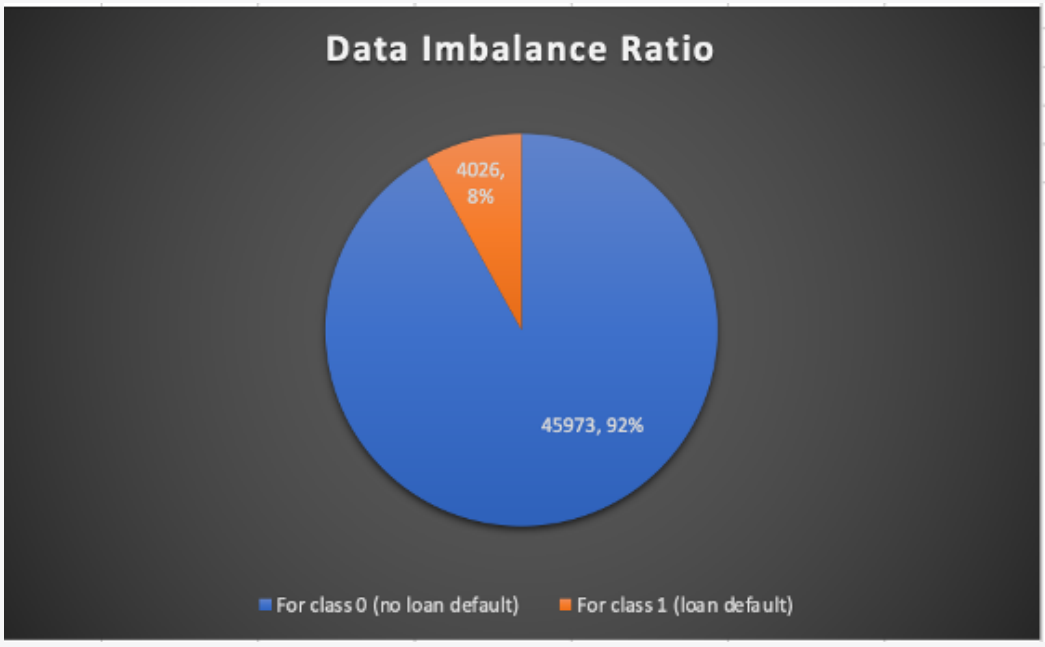
- I calculated the **Q1, Q3, IQR, Lower Threshold and Upper Threshold** using Excel functions as shown in the picture above. Data points that fall below the Lower Threshold or above the Upper Threshold are considered potential outliers and may require further investigation to determine if they are valid data points or data entry errors.
- I also plotted a Box and Whisker Chart to know the outliers, here any data points below the Lower Whisker or above the Upper Whisker are considered as Outliers, which can be easily seen in the different charts of different columns.
- We observe here that most of the columns have outliers in them except **Days_Birth** and **Days_Id_Publish**

BANK LOAN ANALYSIS

INSIGHTS:

C. Analyse Data Imbalance- Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

For class 0 (no loan default)	45973
For class 1 (loan default)	4026
Data Imbalance Ratio	8.75%



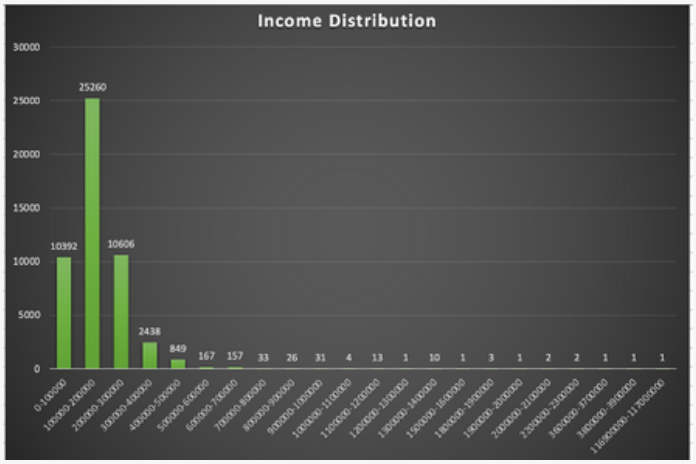
Steps performed to analyze the data imbalance-

- With the help of the pivot table, I found the values of Class 0 and Class 1.
- To know the Data Imbalance Ratio, I divided Class 1 (Loan Default Cases) by Class 0 (No Loan Default Cases) and got 8.75 % as the Data Imbalance Ratio
- The data imbalance ratio of approximately 8.75% indicates that the 'loan default class' is about 8.75% of the 'no loan default class', highlighting the data imbalance in the target variable.

BANK LOAN ANALYSIS

INSIGHTS:

D. Perform Univariate, Segmented Univariate, and Bivariate Analysis- Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features



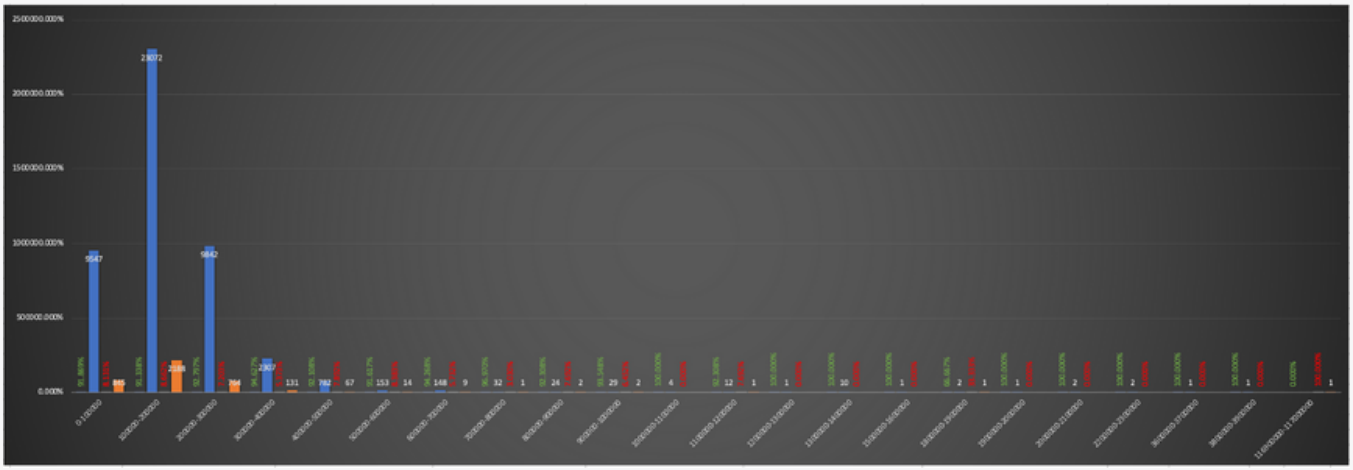
Univariate Analysis

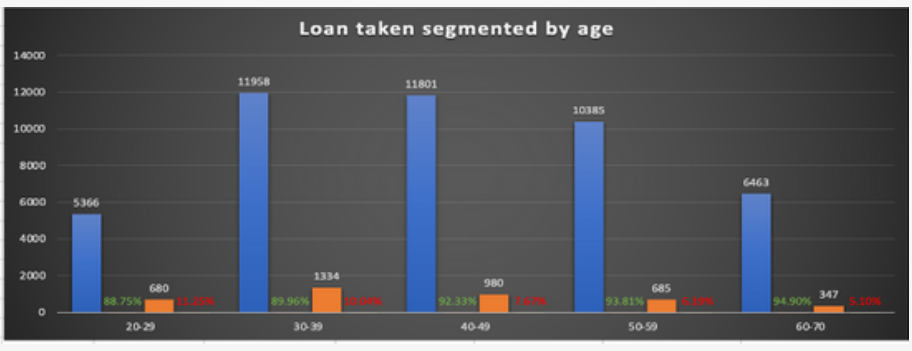
Here we observe that most of the people who come to take the loans are earning in between $0 to $300,000 , whereas the highest number of people who take the loan earn between $100,000 to $200,000

Segmented Univariate Analysis

Amt_Income_Total - Target

Here we observe that generally as Income is Increasing the loan default percentage is decreasing.

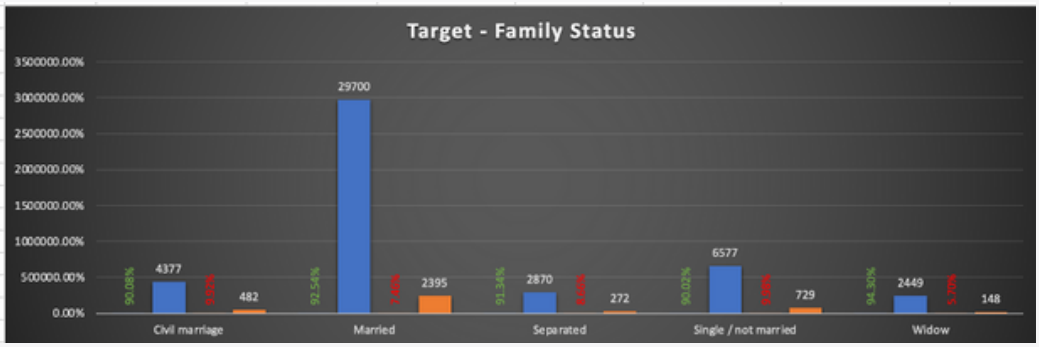




Age - Target

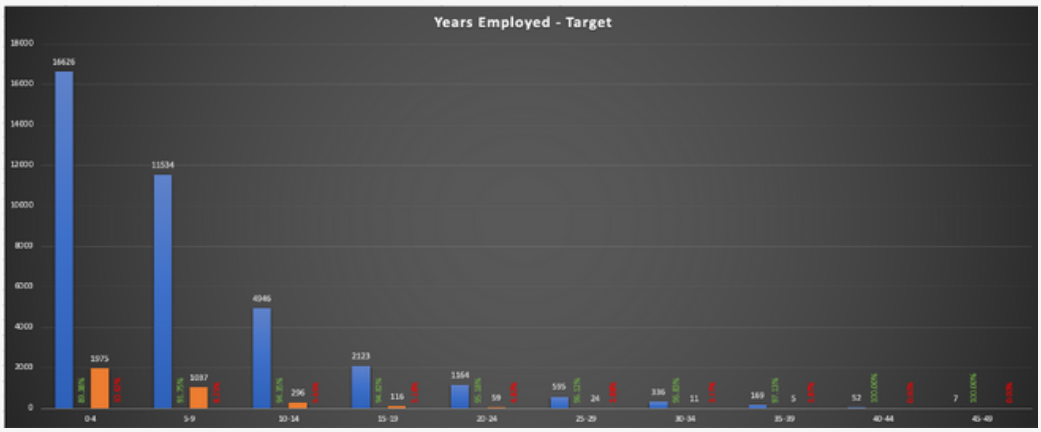
Here we observe that generally as Age is Increasing the loan default percentage is decreasing.

BANK LOAN ANALYSIS



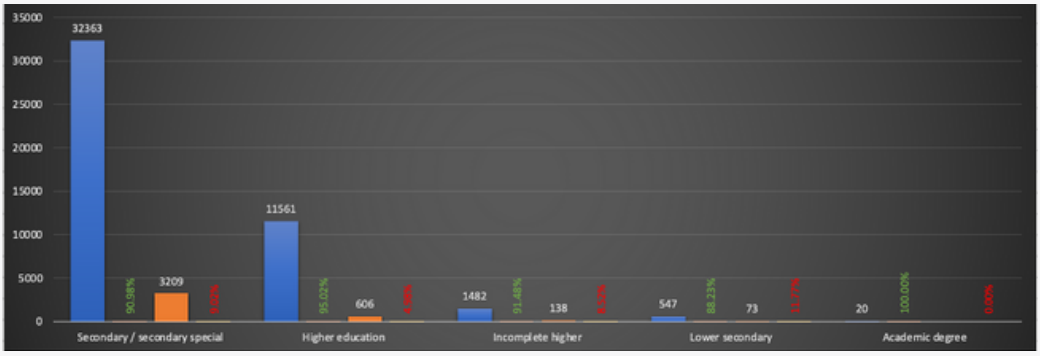
Family Status - Target

Here we observe that People who are Married or Widow have lower loan default percentage.



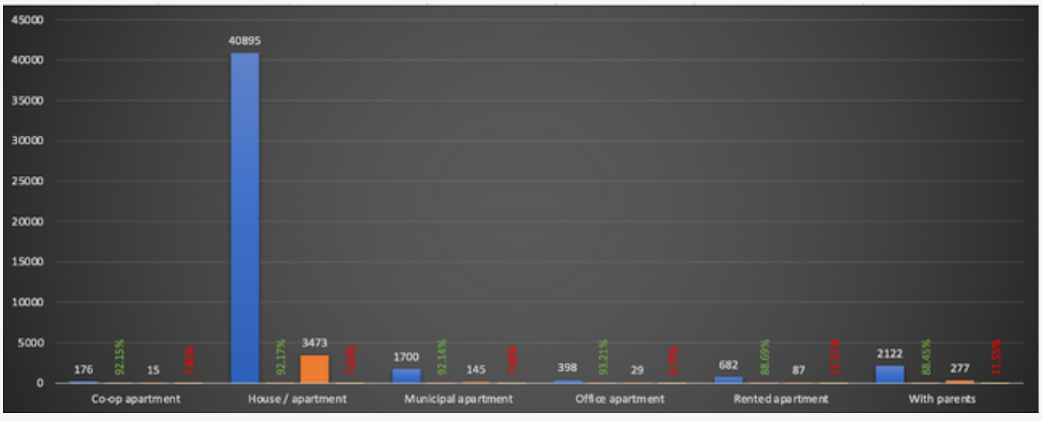
Years Employed - Target

Here we observe that People who are employed for more years have lower loan default percentage.



Education_Type - Target

Here we observe that as education level increases, loan default percentage decreases. People who have higher education tend to default less and less on their loan.



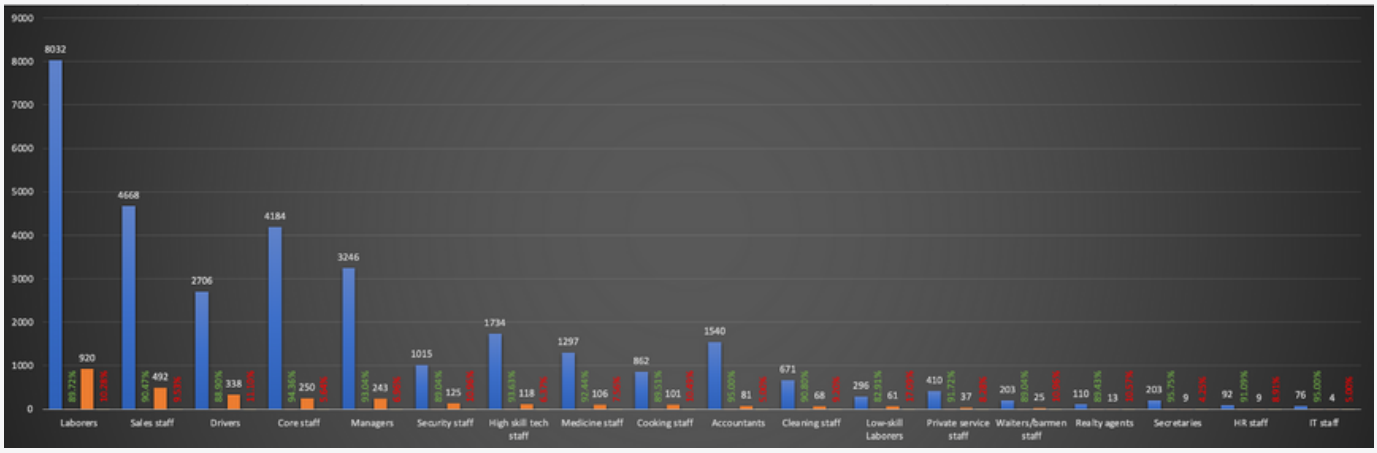
Housing_Type - Target

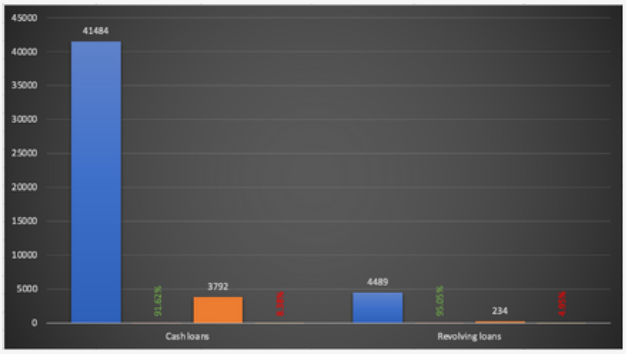
Here we observe that People living in a Rented Apartment or with Parents have higher loan default percentage as compared to people living in apartments or houses.

BANK LOAN ANALYSIS

Occupation_Type - Target

Here we observe that People working as Low Skill Labourers, Drivers, Labourers, Cleaning Staff, Waiters/Barmen Staff etc tend to have higher loan default percentage as compared to people working as IT Staff, Secretaries, Accountants, Core Staff and Managers.

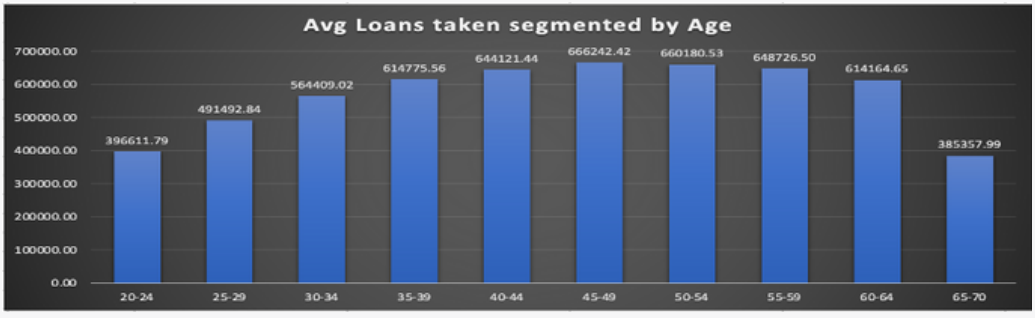




Contract_Type - Target

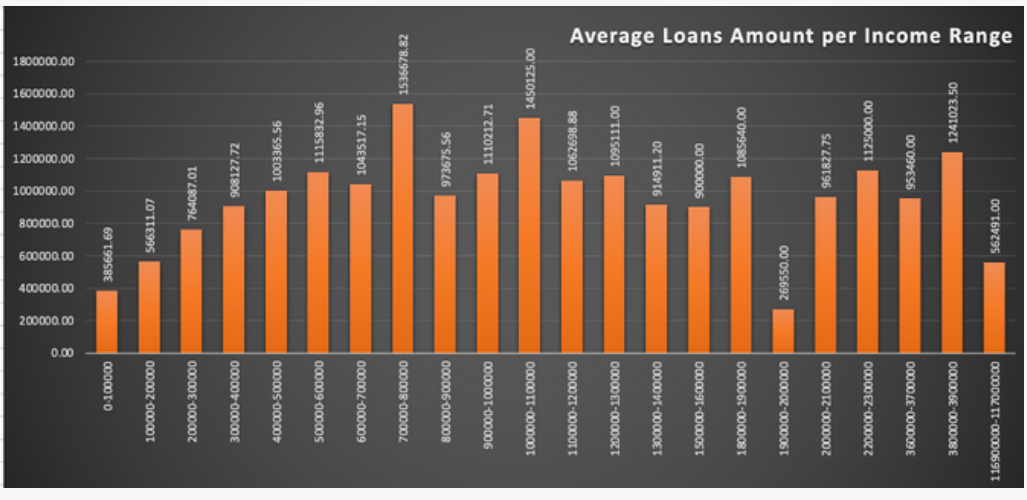
Here we observe that although most people took cash loans but people who took revolving loans have lower loan default percentage.

Bivariate Analysis



Amt_Credit - Age

Here we observe that People in the age between 35 to 65 tend to take higher amount loan



Amt_Credit - Income

Here we observe that Highest average Loan amount taken is by People whose Income is in 700K-800K.

Here we cant say that as their income increases the amount of loan also increases

BANK LOAN ANALYSIS

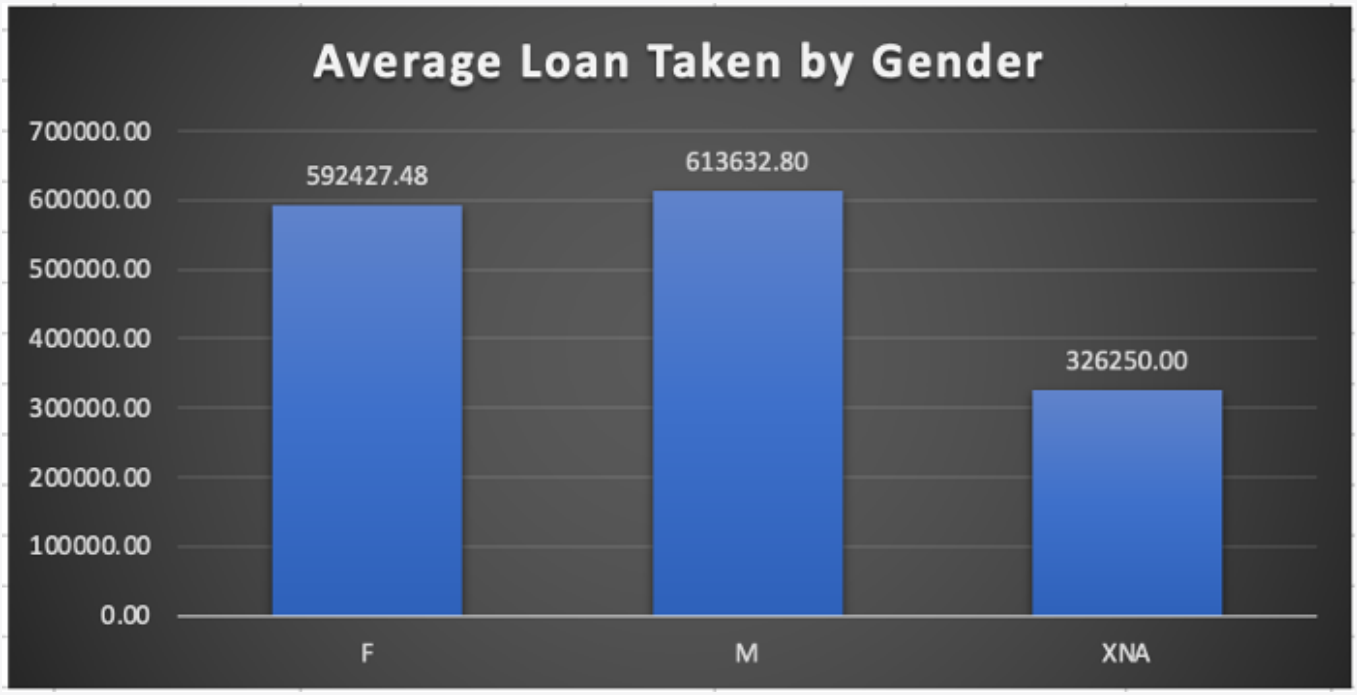
Segmented and Bivariate Analysis

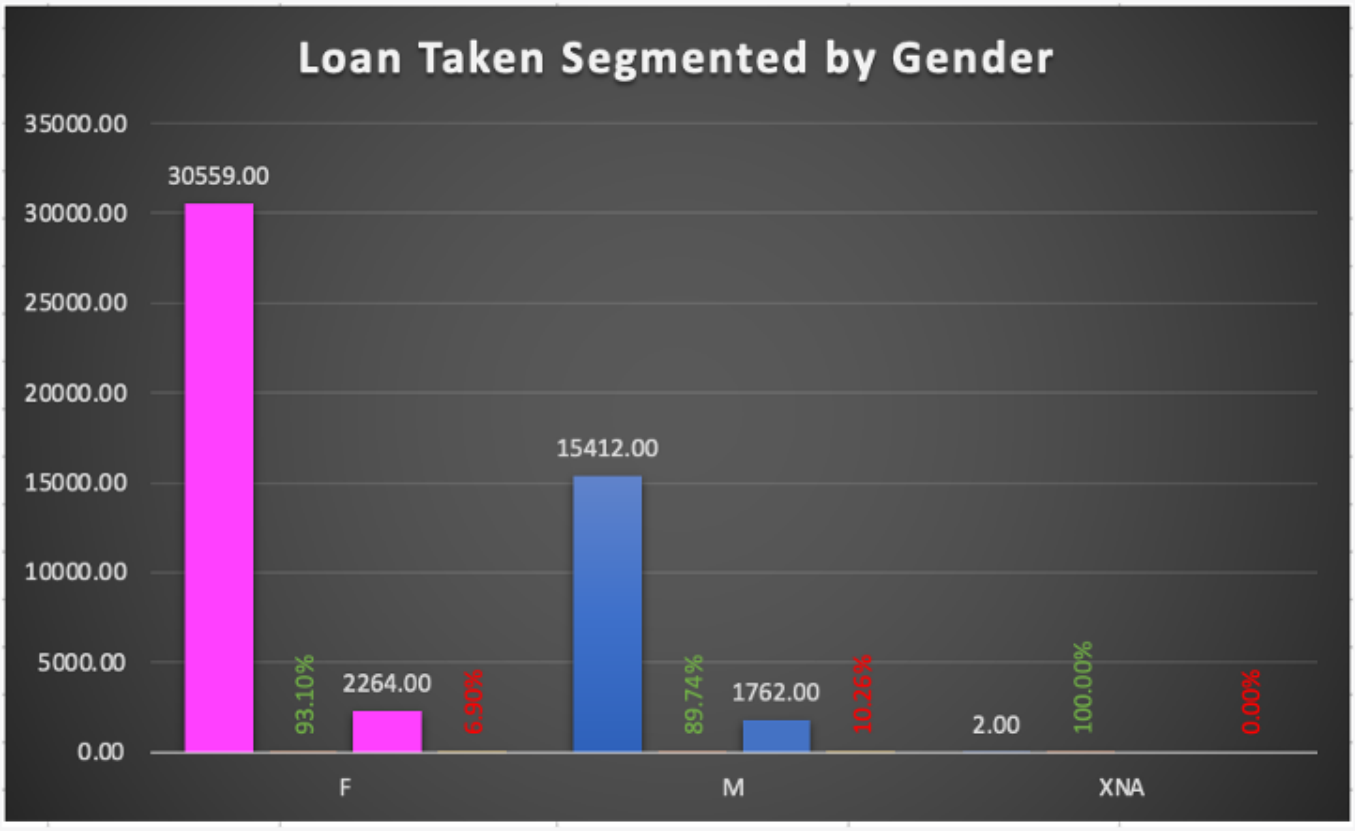
Gender - Amt_Credit

Here we observe that More Females took loans compared to Males and their Loan Default Percentage is Less.

Although fewer Males take loans their average Loan Amount is more.

This data shows that **Females should be preferred over males** when giving the loan.





BANK LOAN ANALYSIS

INSIGHTS:

E. Identify Top Correlations for Different Scenarios- Segment the dataset based on different scenarios and identify the top correlations for each segmented data using Excel functions.

Correlation
Target – 0 (Loan Payees)

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	EXT_SOURCE_2	EXT_SOURCE_23	DAYS_LAST_PHONE_CHANGE	CNT_CHILDREN
AMT_INCOME_TOTAL	1												
AMT_CREDIT	0.377965752	1											
AMT_ANNUITY	0.451135103	0.770772908	1										
AMT_GOODS_PRICE	0.384675888	0.987235417	0.77613285	1									
REGION_POPULATION_RELATIVE	0.181941261	0.095539444	0.117279106	0.09896058	1								
DAYS_BIRTH	0.073769425	-0.051084182	0.009910977	-0.04869335	-0.030435419	1							
DAYS_EMPLOYED	-0.162702675	-0.077367219	-0.113005115	-0.075172	-0.006610653	-0.61528998	1						
DAYS_REGISTRATION	0.06893375	0.008053758	0.034609001	0.01129847	-0.058501361	0.33502805	-0.204370881	1					
DAYS_ID_PUBLISH	0.032286356	-0.008290189	0.009427021	-0.00930065	-0.002236288	0.27007331	-0.27222439	0.103548902	1				
EXT_SOURCE_2	0.15617334	0.136258463	0.130022952	0.14329839	0.201089847	-0.0803327	-0.034096314	-0.053917105	-0.040900368	1			
EXT_SOURCE_23	-0.073654638	0.028831122	0.018547427	0.03105218	-0.013832665	-0.17888625	0.098594722	-0.099901082	-0.111944396	0.068846786	1		
DAYS_LAST_PHONE_CHANGE	-0.049497956	-0.071203379	-0.064450897	-0.07428594	-0.044127291	0.07253958	0.032951867	0.047780168	0.085063175	-0.184718008	-0.059902129	1	
CNT_CHILDREN	0.036319722	0.005705458	0.026384162	0.00155316	-0.024912809	0.33587627	-0.243591518	0.183072478	-0.032537221	-0.013466269	-0.039280661	-0.004822698	1

Top 10 Correlations
Loan Payees

AMT_GOODS_PRICE	AMT_CREDIT	0.987235417
AMT_GOODS_PRICE	AMT_ANNUITY	0.77613285
AMT_ANNUITY	AMT_CREDIT	0.770772908
AMT_ANNUITY	AMT_INCOME_TOTAL	0.451135103
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.384675888
AMT_CREDIT	AMT_INCOME_TOTAL	0.377965752
CNT_CHILDREN	DAYS_BIRTH	0.335876269
DAYS_REGISTRATION	DAYS_BIRTH	0.335028046
DAYS_ID_PUBLISH	DAYS_BIRTH	0.270073313
EXT_SOURCE_2	REGION_POPULATION_RELATIVE	0.201089847

Bottom 10 Correlations
Loan Payees

DAYS_EMPLOYED	DAYS_BIRTH	-0.615289978
DAYS_ID_PUBLISH	DAYS_EMPLOYED	-0.27222439
CNT_CHILDREN	DAYS_EMPLOYED	-0.243591518
DAYS_REGISTRATION	DAYS_EMPLOYED	-0.204370881
DAYS_LAST_PHONE_CHANGE	EXT_SOURCE_2	-0.184718008
EXT_SOURCE_23	DAYS_BIRTH	-0.17888625
DAYS_EMPLOYED	AMT_INCOME_TOTAL	-0.162702675
DAYS_EMPLOYED	AMT_ANNUITY	-0.113005115
EXT_SOURCE_23	DAYS_ID_PUBLISH	-0.111944396
EXT_SOURCE_24	DAYS_REGISTRATION	-0.099901082

Steps performed to calculate correlation-

- To find the correlation, the first thing I did was I choose all the numeric columns
- Then used Excels built in function called Correlation which is under the Data Tab to create the correlation matrix
- Then used conditional formatting to highlight the cells
- From this correlation matrix it was very easy to find the top 10 and bottom 10 Correlations

BANK LOAN ANALYSIS

Correlation
Target – 1 (Loan Defaulters)

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	EXT_SOURCE_2	EXT_SOURCE_23	DAYS_LAST_PHONE_CHANGE	CNT_CHILDREN
AMT_INCOME_TOTAL	1												
AMT_CREDIT	0.015271444	1											
AMT_ANNUITY	0.018004594	0.749665201	1										
AMT_GOODS_PRICE	0.013266279	0.982432318	0.749705184	1									
REGION_POPULATION_RELATIVE	-0.006180303	0.067775624	0.073123998	0.076596242	1								
DAYS_BIRTH	0.009033662	-0.142506035	-0.008751713	-0.140996151	-0.016468731	1							
DAYS_EMPLOYED	-0.011555963	0.016039571	-0.079556008	0.020213912	0.007742909	-0.58147904	1						
DAYS_REGISTRATION	-0.009561152	-0.042844404	0.021581654	-0.043371319	-0.046130288	0.28843784	-0.188718437	1					
DAYS_ID_PUBLISH	-0.009122006	-0.043771901	-0.02132109	-0.049784603	-0.005118563	0.24789657	-0.230063668	0.09029149	1				
EXT_SOURCE_2	-0.016228175	0.119184882	0.113693247	0.133319809	0.159220129	-0.11141366	-0.018800184	-0.067763159	-0.03748574	1			
EXT_SOURCE_23	-0.026469291	0.045724291	0.017777631	0.04751535	-0.021648936	-0.13943364	0.083171459	-0.052718825	-0.073389069	0.047324503	1		
DAYS_LAST_PHONE_CHANGE	0.012457111	-0.124539343	-0.100470941	-0.128807068	-0.067105681	0.12460949	-0.015732544	0.078604652	0.138087781	-0.204067856	-0.041135827	1	
CNT_CHILDREN	0.010110177	0.007601905	0.029172977	-0.001116682	-0.020359154	0.2496732	-0.189324184	0.152113117	-0.042360717	-0.01541713	-0.015321808	0.011339334	1

Top 10 Correlations
Loan Defaulters

AMT_GOODS_PRICE	AMT_CREDIT	0.982432318
AMT_GOODS_PRICE	AMT_ANNUITY	0.749705184
AMT_ANNUITY	AMT_CREDIT	0.749665201
DAYS_REGISTRATION	DAYS_BIRTH	0.288437837
CNT_CHILDREN	DAYS_BIRTH	0.2496732
DAYS_ID_PUBLISH	DAYS_BIRTH	0.247896571
EXT_SOURCE_2	*REGION_POPULATION_RELATIVE*	0.159220129
CNT_CHILDREN	DAYS_REGISTRATION	0.152113117
DAYS_LAST_PHONE_CHANGE	DAYS_ID_PUBLISH	0.138087781
EXT_SOURCE_2	*AMT_GOODS_PRICE*	0.133319809

Bottom 10 Correlations
Loan Defaulters

DAYS_EMPLOYED	DAYS_BIRTH	-0.581479041
DAYS_ID_PUBLISH	DAYS_EMPLOYED	-0.230063668
DAYS_LAST_PHONE_CHANGE	EXT_SOURCE_2	-0.204067856
CNT_CHILDREN	DAYS_EMPLOYED	-0.189324184
DAYS_REGISTRATION	DAYS_EMPLOYED	-0.188718437
DAYS_BIRTH	AMT_CREDIT	-0.142506035
DAYS_BIRTH	AMT_GOODS_PRICE	-0.140996151
EXT_SOURCE_23	*DAYS_BIRTH*	-0.139433636
DAYS_LAST_PHONE_CHANGE	AMT_GOODS_PRICE	-0.128807068
DAYS_LAST_PHONE_CHANGE	AMT_CREDIT	-0.124539343

Steps performed to calculate correlation

- To find the correlation, the first thing I did was I choose all the numeric columns
- Then used Excels built in function called Correlation which is under the Data Tab to create the correlation matrix
- Then used conditional formatting to highlight the cells
- From this correlation matrix it was very easy to find the top 10 and bottom 10 Correlations

BANK LOAN ANALYSIS

KEY INSIGHTS FROM OUR ANALYSIS:

- Outliers and Data Distribution: Most columns display outliers, except for **"Days_Birth"** and **"Days_ID_Publish."**
- **Data Imbalance:** The dataset has a substantial data imbalance ratio of 8.75%, emphasizing the need to address class imbalance.
- **Loan Amount and Income:** People with incomes between 100k to 200k tend to apply for loans more frequently. As income increases, loan default percentages decrease.
- **Age and Loan Default:** Loan default likelihood decreases as age increases, implying older applicants have better repayment behaviour.
- **Marital Status:** Married or widowed applicants show lower loan default percentages, suggesting a connection between marital status and loan repayment.
- **Employment Duration:** Longer job tenures are associated with lower loan default rates, indicating stable employment positively influences repayments.
- **Education Level:** Higher education correlates with reduced default rates, implying educated individuals prioritize loan repayment.
- **Housing Situation:** Rented apartments or living with parents are linked to higher default rates, possibly reflecting financial stability differences.
- **Occupation Impact:** Occupations like IT Staff, Secretaries, Accountants, Core Staff, and Managers show lower default percentages than low skilled labourers, Drivers, and Cleaning Staff.
- **Loan Type:** Revolving loans have lower defaults, though cash loans are more common.
- **Gender Dynamics:** Despite more female applicants, their loan default rate is lower than males. Males, however, ask for higher average loan amounts.

RESULTS

- The main challenge was managing the large dataset. I learned how to handle large datasets.
- The dataset had a lot of missing data and outliers, which I learned how to manage effectively.
- I learned how to use data analysis add-ins in Excel.
- This project helped me improve my Excel skills and gain a better understanding of how to navigate complex datasets.

PROJECT LINK

https://drive.google.com/drive/folders/1r6bM200g8DAag94qym5_Dhdz8aVdh0_?usp=drive_link

VIDEO LINK

https://tella.video/trainity-bank-loan-case-study-hn6h

MODULE 7

ANALYSING THE IMPACT OF CAR FEATURES ON PRICE AND PROFITABILITY

ANALYSING THE IMPACT OF CAR FEATURES ON PRICE AND PROFITABILITY

PROJECT DESCRIPTION:

The car industry is ever-evolving, with a focus on things like saving fuel and using new technology. Companies that make cars are competing a lot, and people's choices about cars are also changing. Some cars use electricity, and others use gasoline.

The question for us as Data Analyst is: How can car companies decide the price of cars and make them in a way that people like and they also make money? This means looking at things like what makes a car special, what kind of people want it, and how much it costs.

By analysing data, to look at patterns and groups, car companies can figure out what to do. This helps them make good choices about prices and what to make in the future, so they can do well in the market and make more money over time.

APPROACH:

To execute the project, I first familiarized myself with the dataset and identified relevant columns for analysis. I then cleaned the data by removing duplicates and unnecessary rows and columns. I then Imputed the missing cells by using data from the Table or from Online. I then analysed and visualized the dataset, created a Dashboard on Excel and ultimately gathered Insights.

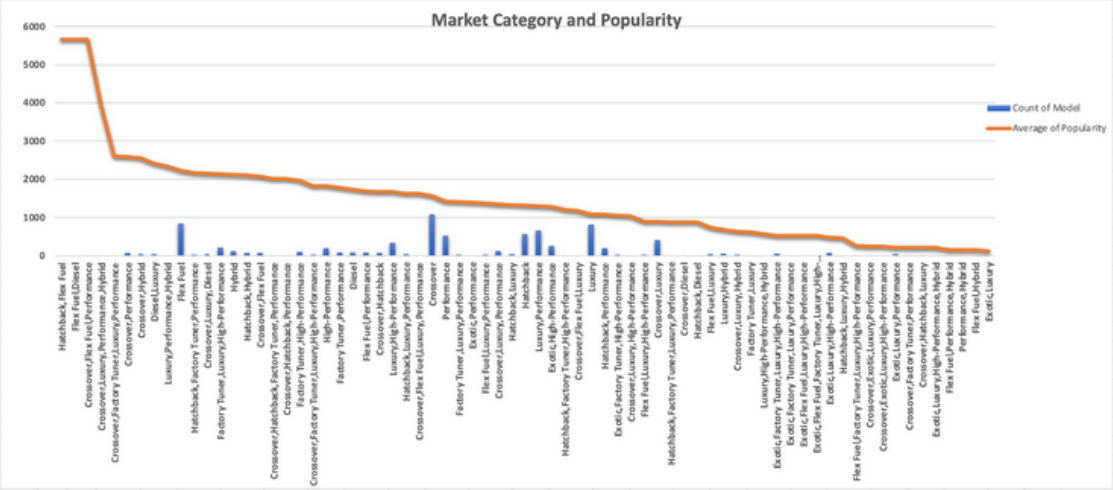
TECH-STACK USED:

Microsoft Excel for Mac Version 16.74.

ANALYSING THE IMPACT OF CAR FEATURES ON PRICE AND PROFITABILITY

INSIGHTS:

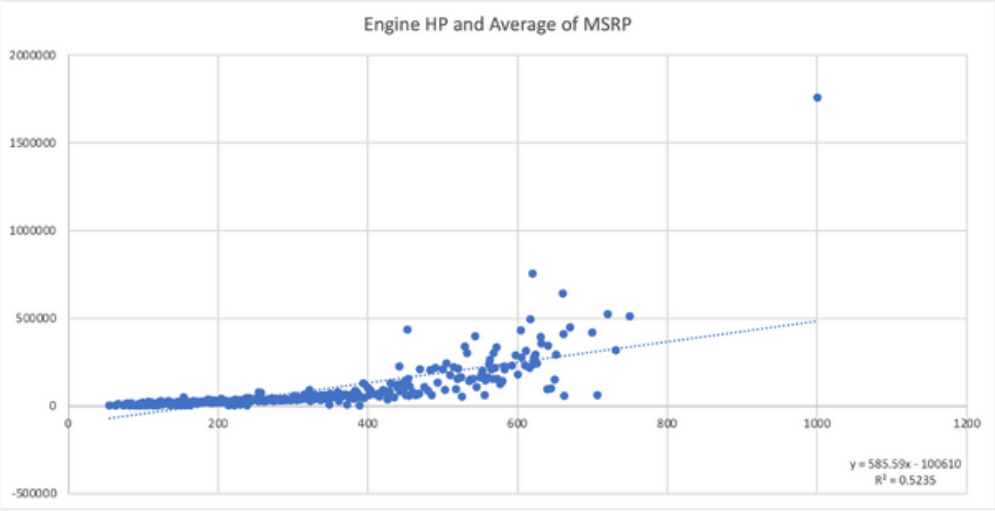
A. How does the popularity of a car model vary across different market categories?



We observe that -

- Hatchback, Flex Fuel and Diesel have the highest Popularity, whereas Exotic has the least Popularity
- Crossover, Flex Fuel and Luxury are the most sold cars

B. What is the relationship between a car's engine power and its price?

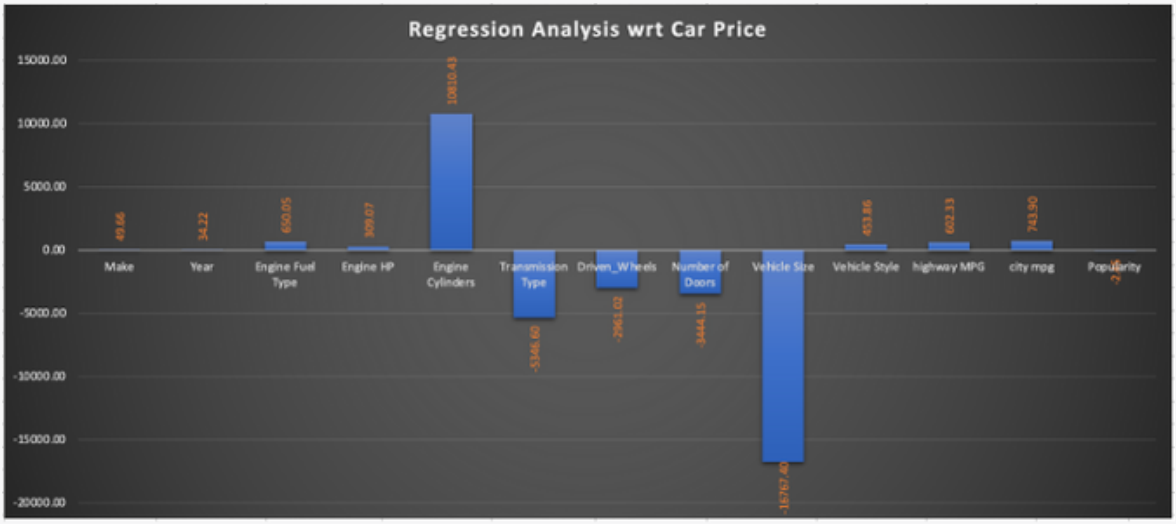


We observe that -

- The positive coefficient for Engine HP (585.59) suggests that, on average, as the Engine HP of a car increases, the Price tends to increase as well. This aligns with the common expectation that more powerful engines are often associated with higher-priced vehicles.
- The R^2 value indicates that 52.35% of the variability in Price can be explained by Engine HP according to this linear model. This suggests that while Engine HP is a significant factor in determining Price, there are other factors that contribute to the variability in Price.

ANALYSING THE IMPACT OF CAR FEATURES ON PRICE AND PROFITABILITY

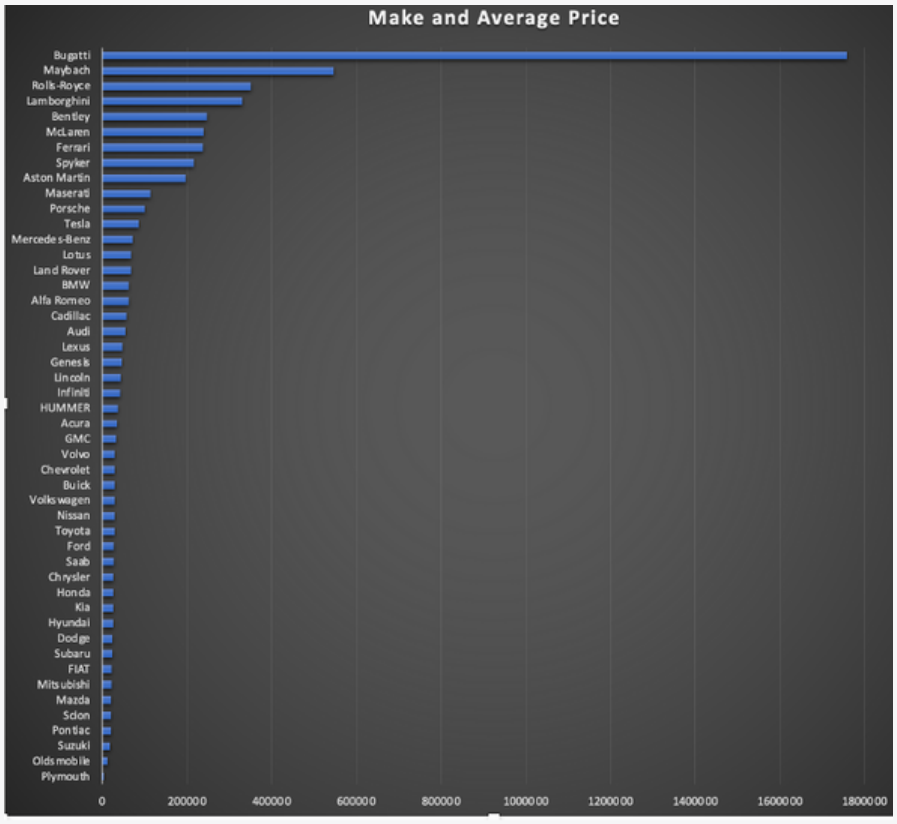
C. Which car features are most important in determining a car's price?



Steps performed to do Regression Analysis

- Here I first did Regression Analysis on the above columns shown in the charts.
- Regression Analysis is inside Data Analysis under the Data Tab
- On the Y axis, I choose Price. On the X axis- I choose all the related columns.
- I observed that **Engine Cylinders**, **City MPG**, **Highway MPG** and **Engine Fuel Type** have the **highest positive coefficients w.r.t Car Price**. This means that these variables are most important in determining a car's price.

D. How does the average price of a car vary across different manufacturers?

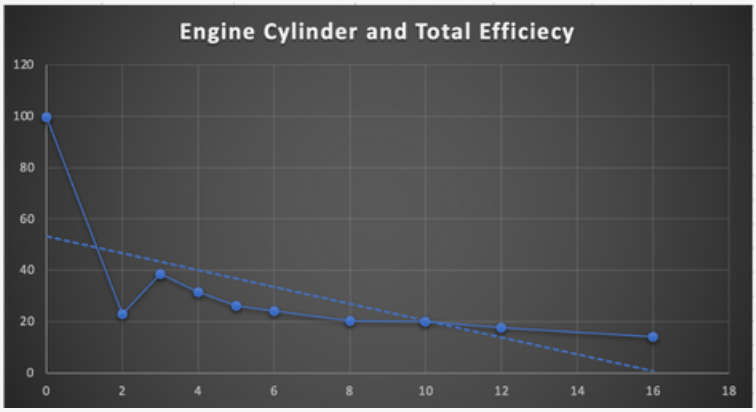


We observe that -

- There's a wide range of average prices across different makes, with some makes having relatively lower average prices (e.g., Plymouth, Suzuki) and others having significantly higher average prices (e.g., Bugatti, Maybach).
- This variability in average prices across different car brands can reflect factors such as brand reputation, target market, vehicle type, performance, features, and luxury status.

ANALYSING THE IMPACT OF CAR FEATURES ON PRICE AND PROFITABILITY

E. What is the relationship between fuel efficiency and the number of cylinders in a car's engine?



	highway MPG	*Engine Cylinders*
highway MPG	1	
Engine Cylinders	-0.616638877	1

A correlation coefficient of -0.616638877 indicates that as the number of engine cylinders increases, fuel efficiency tends to decrease.

In practical terms, this means that vehicles with more engine cylinders (larger engine sizes) are likely to have lower fuel efficiency. As larger engines with more cylinders often require more fuel to operate, resulting in lower miles per gallon (MPG) or worse fuel efficiency.

DASHBOARD TASKS

- 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 features 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?

As Dashboard is interactive in nature. It is not possible to show it here so it can be accessed by clicking on the below link

https://docs.google.com/spreadsheets/d/1WIngqM0tFJ1T9tWo-H61uyJ_FZAV4fSD/edit?usp=sharing&ouid=109466755193972209405&rtpof=true&sd=true

ANALYSING THE IMPACT OF CAR FEATURES ON PRICE AND PROFITABILITY

INSIGHTS

- We observed that although Hatchback, Flex Fuel and Diesel have the highest Popularity, but the most sold cars were Crossover, Flex Fuel and Luxury.
- W also observed that Engine HP is a significant factor in determining Price. Higher the Engine HP, Higher the Price of the Car
- Engine Cylinders, City MPG, Highway MPG and Engine Fuel Type are most important in determining a car's price.
- We can divide our list into Luxury, Premium, Economic and Everyday vehicles. This variability in average prices depends on brand reputation, target market, vehicle type, performance, features, and luxury status.
- We also observed that vehicles with more engine cylinders (larger engine sizes) are likely to have lower fuel efficiency. As larger engines with more cylinders often require more fuel to operate, resulting in lower miles per gallon (MPG) or worse fuel efficiency.

MY LEARNINGS

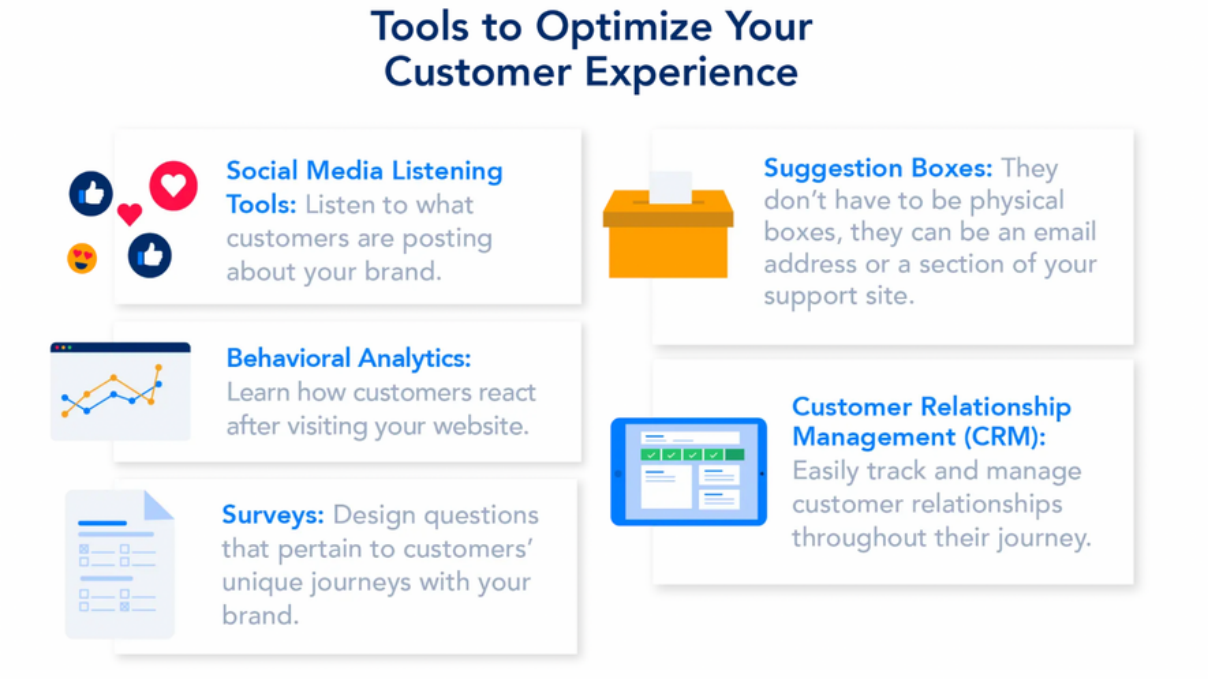
- This dataset was very interesting to me, as I personally love cars and to analyse this dataset, was something I loved doing.
- All my analysis was done in Excel and Pivot Table was used in all the task analysis, from making Charts to Dashboard everywhere.
- The dataset did not have a lot of missing data so instead of removing the missing value rows, I decided to go ahead by imputing values on it.
- I learned how to make a Dashboard in Excel and to do Regression Analysis by transforming Ordinal Variable Columns into Numerical Columns.
- This project helped me improve my Excel skills and gaining a better understanding of how to navigate complex datasets.

PROJECT LINK

https://drive.google.com/drive/folders/1VaLyJnATf4JZcgyvurSXTGmvehlatmdI?usp=drive_link

MODULE 8

ABC CALL VOLUME TREND ANALYSIS



ABC CALL VOLUME TREND ANALYSIS

PROJECT DESCRIPTION:

Our project focuses on enhancing the customer experience (CX) within an inbound call centre operated by company ABC. By analyzing data from the inbound calling team over a 23-day period, we aimed to gain insights into call volume patterns, duration analysis, and abandonment percentage.

These insights informed strategic decisions to optimize call centre operations and enhance customer interactions.

By delving into the inbound calling team's customer experience data, which includes details such as time buckets, call durations, and call statuses, we aimed to uncover patterns that contribute to improved customer satisfaction and operational efficiency.

Our overarching goal was to offer actionable recommendations on Manpower Planning both during the day and night time to reduce the abandonment percentage to 10% while keeping certain assumptions about the agents in mind.

Through a process of continuous analysis and insights sharing, our project endeavours to make a positive impact on both customer experiences and operational effectiveness.

APPROACH:

To execute the project, I first familiarized myself with the dataset and identified relevant columns for analysis. I then cleaned the data by removing columns which had a blank percentage of more than 30%. I then analysed and visualized the dataset, and ultimately gathered Insights.

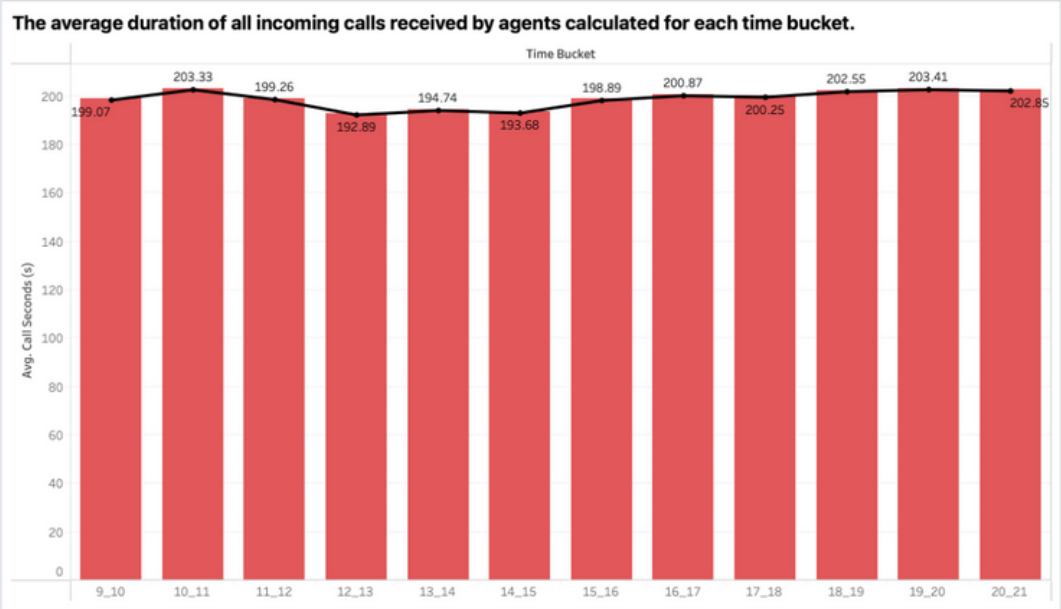
TECH-STACK USED:

Microsoft Excel for Mac Version 16.74.

ABC CALL VOLUME TREND ANALYSIS

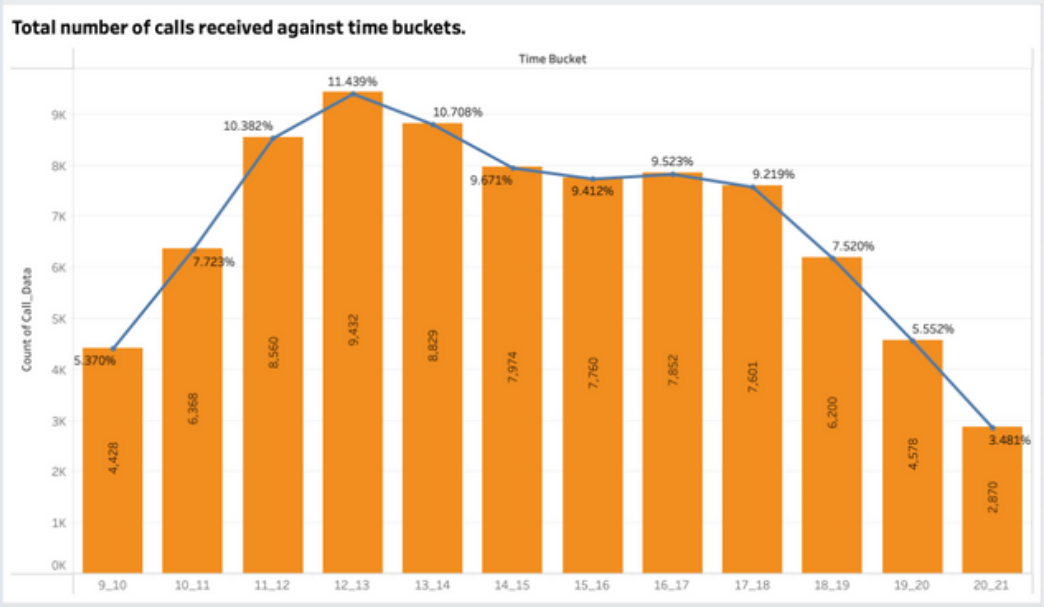
INSIGHTS:

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



We observe that the average duration across the different Time Buckets is relatively same with an average call duration of 198.62 Seconds

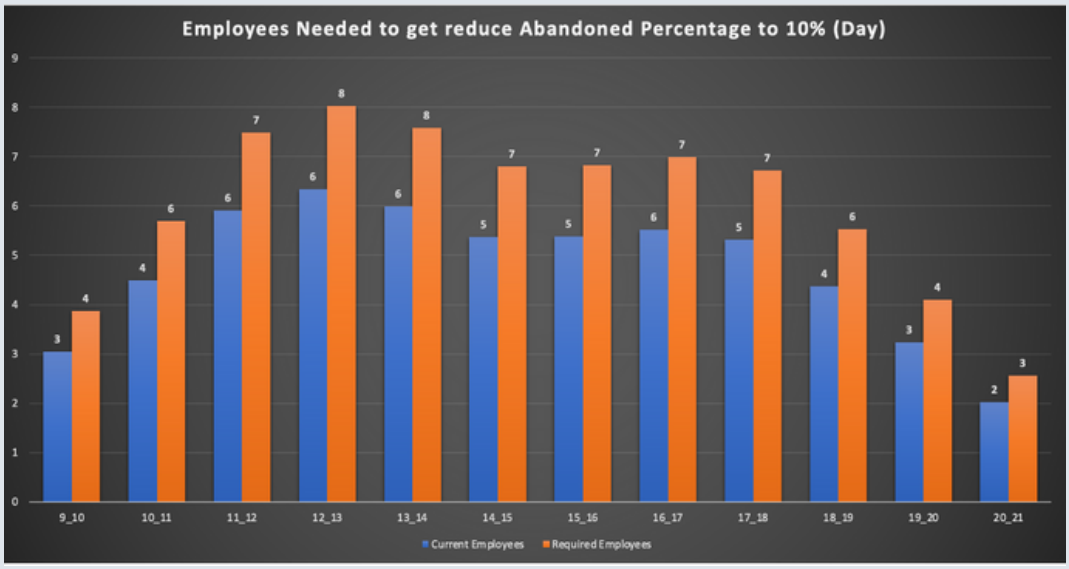
B. Call Volume Analysis- Visualize the total number of calls received against time buckets.



We observe that highest percent of the calls are received between 11 AM to 6 PM.

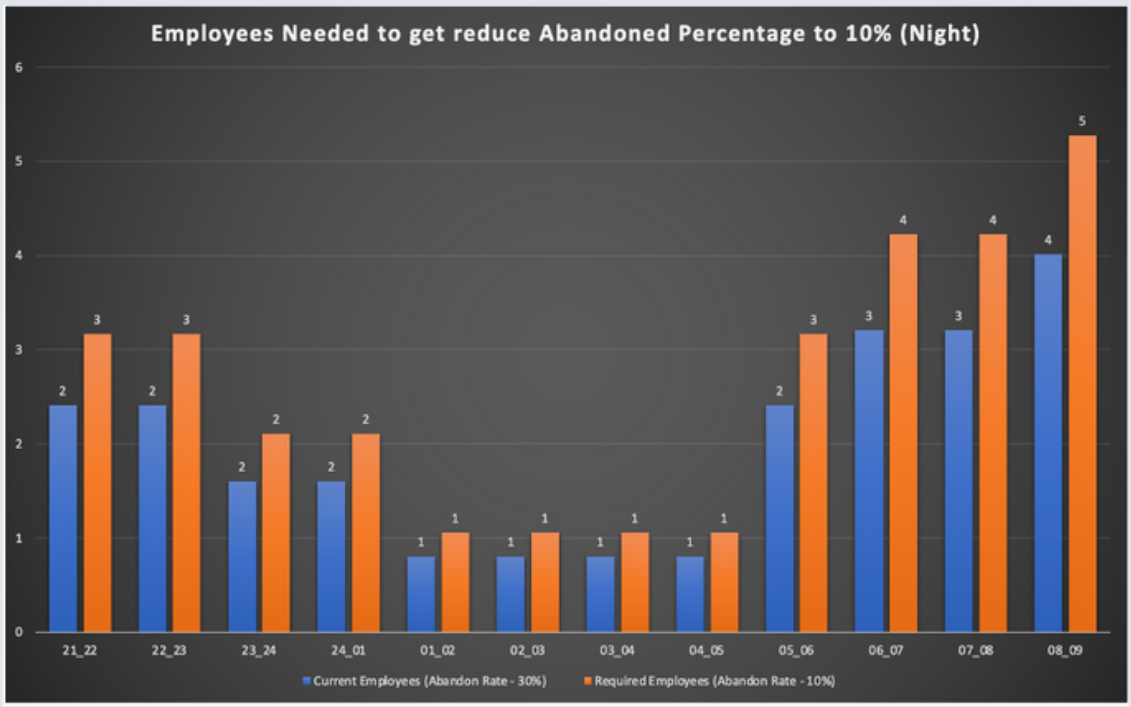
ABC CALL VOLUME TREND ANALYSIS

C. Manpower Planning- Propose a plan for manpower allocation to reduce the abandon rate from 30% to 10% during each time buckets



Through our calculations we found out the Present Employees in the day are 57, but to get abandon rate to 10% we need number of Employees to increase to 72.

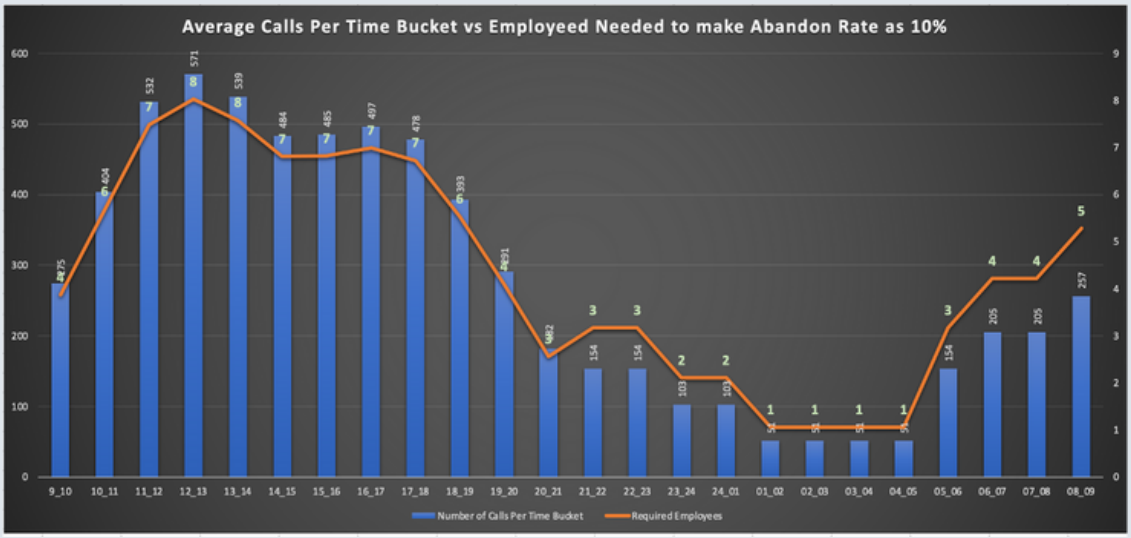
D. Night Shift Manpower Planning- Propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%.



Right now there are no Employees who are working in the night shift. So if we want the abandon rate to maintain at 10% during night also then we will have to employ 32 more employees.

ABC CALL VOLUME TREND ANALYSIS

Calls Received vs Manpower Planning of the whole day (24 Hours)



INSIGHTS

- This dataset was very interesting as by analysing we are required to solve a real-world problem which is faced by almost every company
- I took a lot of time to get the Abandon rate to 10% as I wanted to include all the assumptions in my calculation
- This project helped me in learning more about Manpower Planning
- This project helped me improve my Excel skills and gain a better understanding of how to navigate complex datasets.

MY LEARNINGS

- This dataset was very interesting as by analysing we are required to solve a real-world problem which is faced by almost every company
- I took a lot of time to get the Abandon rate to 10% as I wanted to include all the assumptions in my calculation
- This project helped me in learning more about Manpower Planning
- This project helped me improve my Excel skills and gain a better understanding of how to navigate complex datasets.

PROJECT LINK

https://drive.google.com/drive/folders/1Azz1Xcli9B-h87CtDI7xw1IDkfUEvavc?usp=drive_link

THANK YOU !