Data Analytics Portfolio

Prepared By
Stan Pereira

Professional Background

My name is Stan Pereira and I am a professional leader with over 15 years of industry and career experience in IT Support.

I started my career in IT Support at the Taj Group of Hotels and have worked my way up to a managerial role in 8 years. I then shifted to KA Hospitality Pvt Ltd as Manager-IT and worked there for over 3 years.

After this, I have been an IT consultant for around 6 years.

A major part of the managerial role in IT Support involved analysis, visualization and problem solving of IT data. This introduced me to the World of Data Analytics.

Even though, my experience is mostly data collection, analyzing, interpreting, visualizing and problem solving using SQL and Excel, I have introduced and learnt to do the same using Tableau and PowerBI. I have also learnt Python to analyze huge data. Through the Trainity certification, I have done projects to better my skills.

I have an ITIL 4 Certification and a GNIIT Certification which, alongwith my MBA(IT) certification has allowed me to understand programming and processes better.

I may be a fresher in Data Analytics, but my real-world work experience and willingness to adapt & learn gives me an edge wrt hand holding on company running, and understanding a client's data and processes better.

Table of Contents

Project 1: Data Analytics Process

Project Description

We use Data Analytics in everyday life without even knowing it. The project is to give an example of such a real-life situation where we use Data Analytics and link it with the data analytics process. We need to prepare a PPT/PDF on a real-life scenario explaining it with the above process (Plan, Prepare, Process, Analyze, Share, Act).

Project Case Study

Data Analytics

"Data Analytics is the science of analyzing raw data to make conclusions about that information. It involves using various tools and methods to identify patterns, trends, and relationships within the data that can help organizations make informed decisions and improve their operations."

Simply put, Data Analytics is the process of drawing conclusions about information after examining the available raw data.

Data Analytics is used in everyday life, even though we do not realize it. A couple of scenarios would be Shopping Patterns, Cell Phone Usage Patterns, Social Media Statistics, Child's Marks Progress

Case Study

The case study that we are going to be presenting is an example of a real-life scenario where we use Data Analytics.

"Placing an order for Food via Delivery"

We will use the data analytics process to further explain the scenario. The various steps of this process are :

- Plan
- Prepare
- Process
- Analyze
- Share
- Act

Explaining the Scenario

Placing a Food Order to be delivered to you at a specified location, is something many people can relate to as it is a scenario that they have faced multiple times.

The decision made to place an order includes raw data, analysis, interpretation and final conclusion/decision.

The case study uses the various data points needed to place a food order using the delivery method to finally result in a satisfiable outcome.

Data Analytics Steps

Plan

The first step in the Data Analytics process is to 'Plan'. The planning process includes collection of various data points and data.

Some of the data points used in our scenario are:

- Mode of Ordering – Online/Phone
- Food Type(Indian/Chinese/Thai)
- Restaurant Selection based on Distance, Reviews, Price
- Menu Selection (Starters, Soups, Main Course, Desserts)
- Portion Size
- Likes/Dislikes

Prepare

The 'Prepare' step includes data that is needed to complete the process.

Some of the data points used in our scenario are:

- Finance Options Available – Cash on Delivery, Credit Card, UPI etc.
- Contact Details – Phone Number/Food App
- Restaurant Availability – Open / Closed
- Food Availability

Process

This step allows you to prepare the best way to go ahead using the data collected in the above two steps.

Analyze

The 'Analyze' step allows for us to methodically think and interpretate the data collected to explain the relationships. Some of the analysis interpretations in this scenario are:

- You will try to stay within the Budget allotted for the order
- The Food Quantity will not be less than the required amount
- Comparison of Reviews for Food you like in various restaurants
- Restaurant Reviews

Share

The 'Share' step enables you to share the information to attain insights on the data collected and analysis done. This can also help get further information/data necessary.

In this scenario, you can communicate with the restaurant/food app the order and get further options such as any offers based on amount, items ordered, payment options etc.

Act

The 'Act' step is the final step which is the act of placing the order, accepting the delivery, paying the order and finally enjoying the meal.

Conclusion

Data Analytics plays an important role in our day-to-day lives. It helps us make crucial and informed decisions in a variety of ways.

Data Analytics also helps us to gain valuable insights from massive data. The analysis of this data, helps us identify trends, predict outcomes and optimize processes

In conclusion, data analytics is a powerful utility that will help drive innovation and growth in all fields, whether it is customer support or production efficiency.

Effective use of the Data Analytics process allows for extracting better insights and conclusions.

Project 2: Instagram User Analytics

Project Description

As a data analyst working with the product team at Instagram, I was tasked with analysing user interactions and engagement with the Instagram app to provide valuable insights that can help the business grow.

The user analysis involves tracking how users engage with a digital product, and the insights derived from this analysis will be used by various teams within the business.

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, and the development team might use them to improve the overall user experience.

Approach

- The initial step is the creation of the database & tables, along with importing the data into the tables. This is achieved by running the DDL & DML SQL queries provided in MySQL through the MySQL Workbench.
- After the previous step, analysis on the data was made and insights were generated from the database by running SQL queries in MySQL Workbench.

Tech Stack Used

MySQL Workbench 8.0 CE - Version 8.0.34 build 3263449 CE (64 bits) Community

Management Queries, Analysis and Insight

Marketing Analysis

1. 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.

Task: Identify the five oldest users on Instagram

SQL Query:

```
SELECT
   username as Username,
   created_at as Creation_Date
FROM
   users
ORDER BY Creation_Date
LIMIT 5;
```

Username	Creation_Date
Darby_Herzog	2016-05-06 00:14:21
Emilio_Bernier52	2016-05-06 13:04:30
Elenor88	2016-05-08 01:30:41
Nicole71	2016-05-09 17:30:22
Jordyn.Jacobson2	2016-05-14 07:56:26

Insight: The marketing team should reward those who have been using the platform for the longest time, including the user 'Darby_Herzog' who has not posted anything. It may entice him to post something.

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

Task: Identify users who have never posted a single photo on Instagram.

SQL Query:

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

Inactive_Users
Aniya_Hackett
Bartholome.Bernhard
Bethany20
Darby_Herzog
David.Osinski47
Duane60
Esmeralda.Mraz57
Esther.Zulauf61
Franco_Keebler64
Hulda.Macejkovic
Jadyn81
Janelle.Nikolaus81
Jessyca_West
Julien_Schmidt
Kasandra_Homenick
Leslie67
Linnea59
Maxwell.Halvorson
Mckenna17
Mike.Auer39
Morgan.Kassulke
Nia_Haag
Ollie_Ledner37
Pearl7
Rocio33
Tierra.Trantow

Insight: The marketing team should encourage the 26 users to start posting by sending them promotional emails. Follow-ups can also be made, so as to remove users who are still inactive.

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

Task: Determine the winner of the contest and provide their details to the team.

SQL Query:

```
SELECT
   users.id AS ID, users.username AS Username,
   photos.image_url AS Photo,
   COUNT(likes.photo_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 likes.photo_id
ORDER BY Total_Likes DESC
LIMIT 1;
```

ID	Username	Photo	Total_Likes
52	Zack_Kemmer93	https://jarret.name	48

Insight: The user 'Zack_Kemmer93' has got the most likes (48) for one of his photos. All users should be informed of the contest winner and his photo. A small reward needs to be given to winner, which might entice other users to post more.

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

Task: Identify and suggest the top five most commonly used hashtags on the platform.

SQL Query:

```
WITH HTR as (
   SELECT
      tag_id, tag_name,
      COUNT(tag_id) as Total_Count,
      RANK() OVER (ORDER BY COUNT(tag_id) DESC) as Tag_Rank
   FROM
      photo_tags
         INNER JOIN
      tags ON photo_tags.tag_id=tags.id
   GROUP BY tag_id
)
SELECT *
FROM
   HTR
WHERE
   Tag_Rank <= 5;
```

tag_id	tag_name	Total_Count	Tag_Rank
21	smile	59	1
20	beach	42	2
17	party	39	3
13	fun	38	4
5	food	24	5
11	lol	24	5
18	concert	24	5

Insight: Even though the top five hashtags were asked for, the hashtag that was ranked 5th was tied with two others, hence a total of 7 hashtags are mentioned.

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

Task: Determine the day of the week when most users register on Instagram. Provide insights on when to schedule an ad campaign.
SQL Query:

```
WITH DayTable as (
   SELECT
      DAYNAME(created_at) as Day_Of_Creation,
      COUNT(DAYNAME(created_at)) as Total_Reg_Users,
      RANK() OVER (order by count(dayname(created_at)) DESC) as DayRank
   FROM
      users
   GROUP BY Day_Of_Creation
)
SELECT *
FROM
   DayTable
WHERE
   DayRank=1;
```

	Day_Of_Creation	Total_Reg_Users	DayRank
▶	Thursday	16	1
	Sunday	16	1

Insight: There are two days of the week that have the most registrations. As the goal is to increase registrations, the marketing team may want to launch ads on both the days of most registrations, as this indicates that there is a high demand and interest.

Investor Metrics

1. User Engagement: Investors want to know if users are still active and posting on Instagram or if they are making fewer posts.

Task: Calculate the average number of posts per user on Instagram. Also, provide the total number of photos on Instagram divided by the total number of users.
SQL Query:

```
SELECT
   (SELECT
      COUNT(DISTINCT id) / COUNT(DISTINCT user_id)
   FROM
      photos) AS Total_Posts_Per_Active_User,
   (SELECT
      COUNT(DISTINCT id)
   FROM
      photos) /
   (SELECT
      COUNT(DISTINCT id)
   FROM
      users) AS Ratio_Total_Photos_to_Total_Users;
```

Total_Posts_Per_Active_User	Ratio_Total_Photos_to_Total_Users
3.4730	2.5700

Insight: The Average number of posts is taken from Active Users, but the ratio is taken from All Users. The information shows the gap between the users that are registered and actual users making posts. This gap needs to decrease.

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

Task: Identify users (potential bots) who have liked every single photo on the site.
SQL Query:

```
SELECT
```

```
    user_id AS ID, username AS Username,
    COUNT(photo_id) AS TotalCount
FROM
    users
        JOIN
    likes ON users.id = likes.user_id
GROUP BY user_id
HAVING TotalCount = (SELECT
        COUNT(id)
    FROM
        photos)
ORDER BY username;
```

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

Insight: As genuine users do not normally like every single photo, the 13 users in the list are Bots or Fake Accounts. This list distorts the data and can provide incorrect analysis. The users need to be removed from the database.

Project Impact

Impact for the Team: The analysis from the data will help the product manager and the team make informed decisions about the future direction of the Instagram App.
Impact for Me: The project allowed me to learn the fundamentals of SQL and its working. It has also shown me how to use it to extract various insights for data analysis.

Project 3: Operation and Metric Analytics

Project Description

Operational Analytics is a crucial process that involves analysing a company's end-to-end operations. This analysis helps identify areas for improvement within the company. As a Data Analyst, I will 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. As a Data Analyst, I will need to answer these questions daily, making it crucial to understand how to investigate these metric spikes.

In this project, I will take on the role of a Lead Data Analyst at a company like Microsoft. I will be provided with various datasets and tables, and my task will be to derive insights from this data to answer questions posed by different departments within the company.

Case Study 1 (Operation Analytics)

- The first case involves an analysis of the job data to improve operational efficiency.
- Various metrics such as Throughput, Productivity, Percentage Share etc., are extracted from the provided data and recommendations for improvements are to be made

Case Study 2 (Metric Analytics)

- The second case study involves investigating data to draw better conclusions.
- Metrics to identify patterns and trends such as User Growth, User Engagement, Cohort Retention Analysis and Email Engagement Metrics are extracted and used to find the best ways to improve productivity.

Approach

Database Creation - The database and tables are created as per the given specifications.
Data Import and Cleaning - The data is imported into the database and ensured that the dataset is valid, accurate, and includes all the needed values.
Perform Analysis - The data is analysed using SQL to identify various metrics like throughput, retention analysis etc.
Data Visualisation - The final step is to use Excel to create insightful visualisations so as to better understand the data.
Notes:

- While creating the Job_Data Table, the datatype specified for the column 'ds' was varchar, even though the better datatype would have been date.
- As part of the data cleaning process, the csv file had the 'ds' column in the text format MM/DD/YYYY. But the required format was YYYY/MM/DD, and hence needed to be converted using SQL statements.
- To perform the analysis and ascertain insights, there was a need to research on functions/terms in SQL and Metrics like CAST function, LOAD DATA statement, ROWS BETWEEN, Throughput, Rolling Average, Cohort Analysis, Engagement metrics etc.

Tech Stack Used

- MySQL Workbench 8.0 CE - Version 8.0.34 build 3263449 CE (64 bits) Community
- Microsoft 365 Online Excel Version 16.0.17012.41002

Case Study 1

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

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

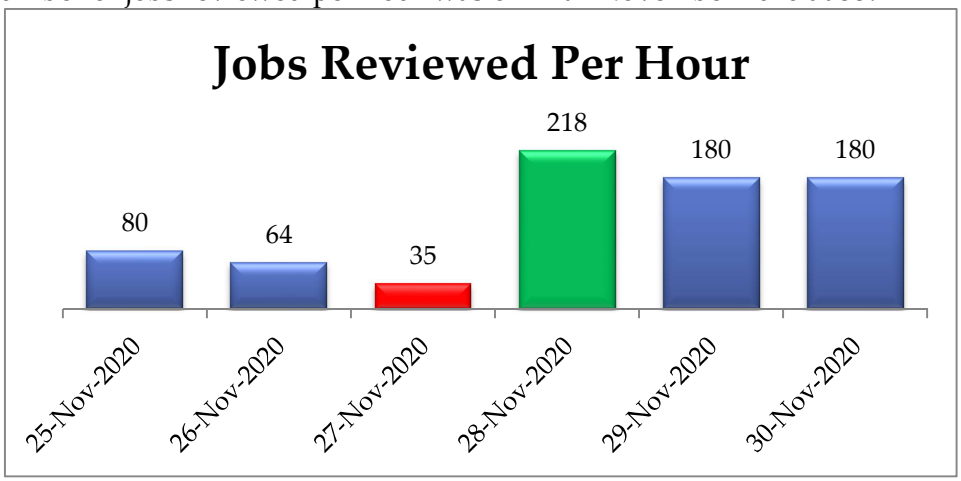
SQL Query:

```
SELECT
   CAST(ds AS DATE) AS Job_Date,
   COUNT(job_id) AS Job_Count,
   SUM(time_spent)/3600 AS Time_Spent_Hour,
   ROUND(COUNT(job_id)/(SUM(time_spent)/3600)) AS
Jobs_Reviewed_Hour_Day
FROM
   job_data
WHERE
   CAST(ds AS DATE) >= '2020/11/01' AND CAST(ds AS DATE) <= '2020/11/30'
GROUP BY Job_Date
ORDER BY Job_Date;
```

Job_Date	Job_Count	Time_Spent_Hour	Jobs_Reviewed_Hour_Day
2020-11-25	1	0.0125	80
2020-11-26	1	0.0156	64
2020-11-27	1	0.0289	35
2020-11-28	2	0.0092	218
2020-11-29	1	0.0056	180
2020-11-30	2	0.0111	180

Insights:

- 28th November 2020 had the highest number of jobs reviewed per hour at 218
- The lowest number of jobs reviewed per hour was on 27th November 2020 at 35.



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

Your Task: Write an SQL query to calculate the 7-day rolling average of throughput.

SQL Query:

```
SELECT
   CAST(ds AS DATE) AS Job_Date,
   ROUND(COUNT(job_id) / SUM(time_spent),4) AS Daily_Throughput,
   ROUND(AVG(COUNT(job_id) / SUM(time_spent)) OVER (ORDER BY CAST(ds AS DATE) ROWS
BETWEEN 6 PRECEDING AND CURRENT ROW),4) AS
7_Day_Throughput
FROM
   job_data
GROUP BY Job_Date
ORDER BY Job_Date;
```

Job_Date	Daily_Throughput	7_Day_Throughput
2020-11-25	0.0222	0.0222
2020-11-26	0.0179	0.0200
2020-11-27	0.0096	0.0166
2020-11-28	0.0606	0.0276
2020-11-29	0.0500	0.0321
2020-11-30	0.0500	0.0351

Insights:

What is Throughput

Throughput analysis is a method of measuring the efficiency of a business process by calculating the rate at which units move through the process from start to finish. It can help a business make decisions that minimize costs and maximize profits. Throughput analysis involves two components: inventory and flow time. Inventory is the number of units that are involved in the process at a given time, and flow time is the amount of time a unit spends in the process from start to finish. The formula for throughput rate is:

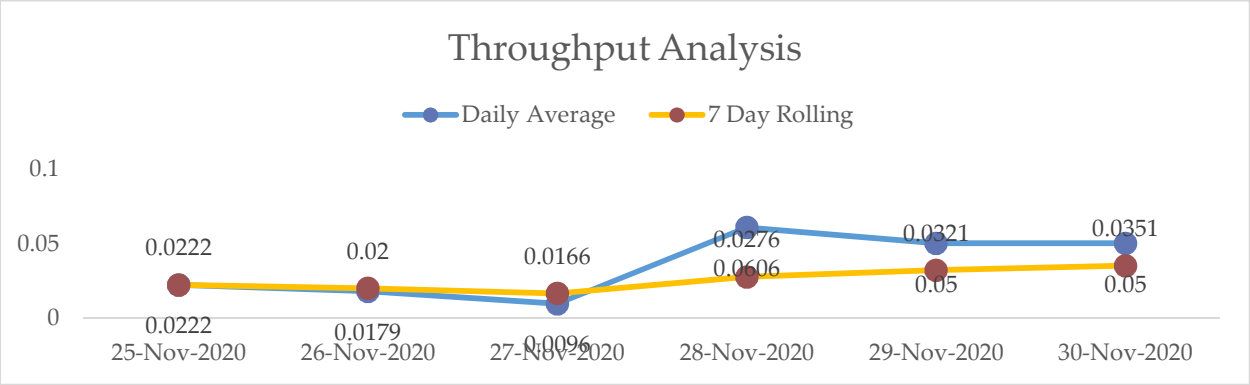
$$R=I/T$$

7-day rolling average v/s daily metric. Which is better in this situation?

A rolling average, sometimes referred to as a moving average, is a metric that calculates trends over short periods of time using a set of data. Specifically, it helps calculate trends when they might otherwise be difficult to detect.

An average/daily metric fluctuates wildly as each day can result in a different calculation. The rolling average creates a trend that can show a better picture over a long time.

In this situation, a 7-day rolling metric makes more sense, due to the fluctuation of the daily metric.



- The graph shows the daily average and the 7 day rolling average
- As observed via the graph, the rolling average is more consistent and less fluctuating. This will help to understand the trend better.

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

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

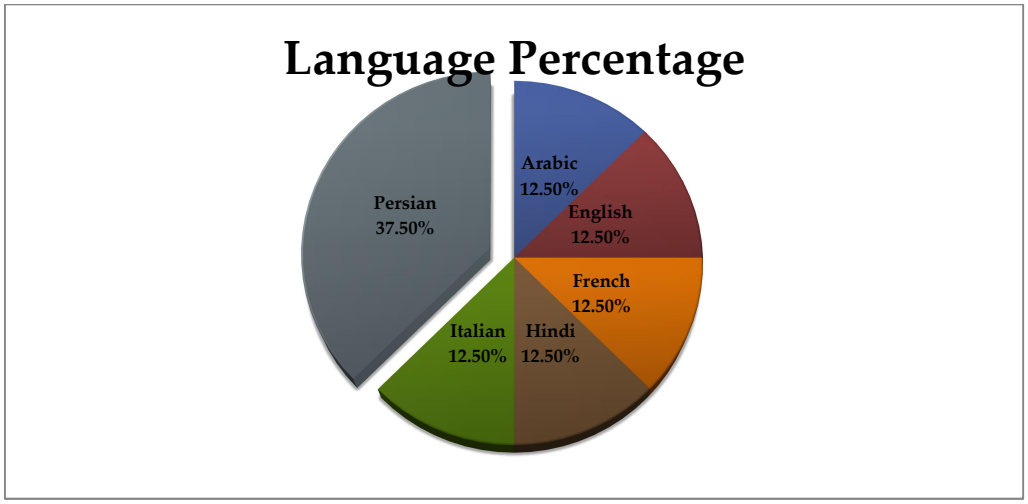
SQL Query:

```
SELECT
    DISTINCT language AS Language,
    ROUND(((COUNT(language) OVER (PARTITION BY language)) /
(COUNT(language) OVER ()) * 100),2) AS Language_Percentage
FROM job_data
WHERE
   CAST(ds AS DATE) >= '2020/11/01'
      AND CAST(ds AS DATE) <= '2020/11/30';
```

Language	Language_Percentage
Arabic	12.50
English	12.50
French	12.50
Hindi	12.50
Italian	12.50
Persian	37.50

Insights:

- The Pie Chart shows the language percentage of jobs reviewed.
- As per the chart, Persian Language jobs are reviewed the most with a share of 37.50%
- All the other languages have the same percentage share of 12.5%



04) Duplicate Rows Detection: Identify duplicate rows in the data.
Your Task: Write an SQL query to display duplicate rows from the job_data table.
SQL Query:

```
SELECT ds AS Job_Date, job_id AS Job_Id, actor_id AS Actor_Id, event AS Event,
    language AS Language, time_spent AS Time_Spent_Sec, org AS Organisation,
    IF (Job_Count>1,"Duplicate","Not Duplicate") AS Duplicate
FROM
    (SELECT *,
        COUNT(*) OVER
(PARTITION BY ds, job_id,
actor_id, event,
language,time_spent, org) AS
Job_Count
    FROM job_data) AS Tot_Count
ORDER BY Job_Date, Duplicate;
```

Job_Date	Job_Id	Actor_Id	Event	Language	Time_Spent_Sec	Organisation	Duplicate
2020/11/25	20	1003	transfer	Italian	45	C	Not Duplicate
2020/11/26	23	1004	skip	Persian	56	A	Not Duplicate
2020/11/27	11	1007	decision	French	104	D	Not Duplicate
2020/11/28	23	1005	transfer	Persian	22	D	Not Duplicate
2020/11/28	25	1002	decision	Hindi	11	B	Not Duplicate
2020/11/29	23	1003	decision	Persian	20	C	Not Duplicate
2020/11/30	21	1001	skip	English	15	A	Not Duplicate
2020/11/30	22	1006	transfer	Arabic	25	B	Not Duplicate

Insights:

- None of the fields are duplicates.

Note : All the columns in the table can be repeated and can have duplicates, so we need to assign a combination of two or more columns to allow for duplicate checks. For this project, I used a combination of all fields for the duplicate check (even though, this is not an ideal check). If a time field was present, it could be used as a primary key).

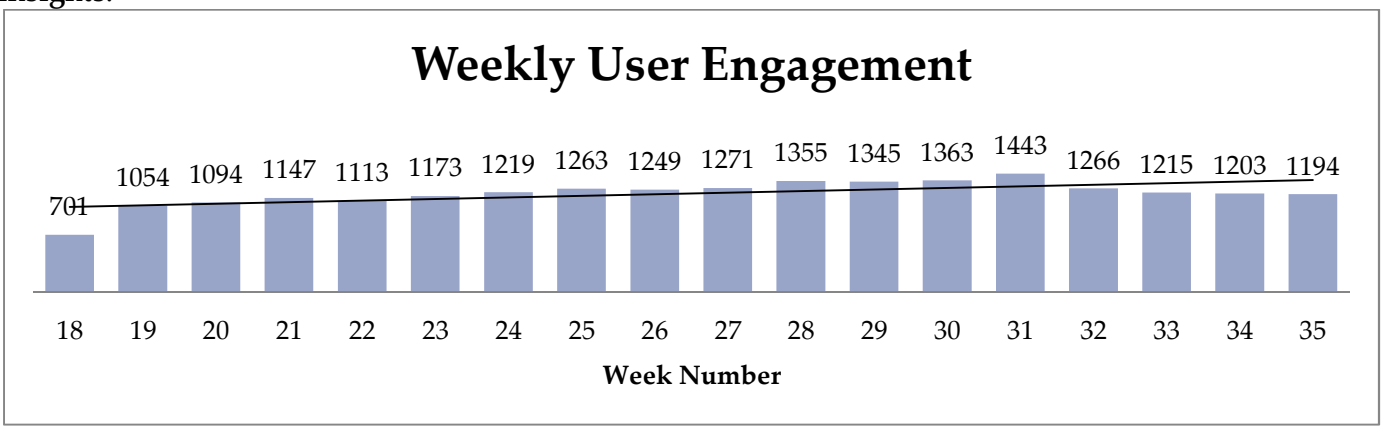
Case Study 2

01) Weekly User Engagement: Measure the activeness of users on a weekly basis.
Your Task: Write an SQL query to calculate the weekly user engagement.
SQL Query:

```
SELECT
   WEEK(occurred_at, 3) AS Week_Number,
   COUNT(DISTINCT user_id) AS User_Count
FROM
   events
GROUP BY Week_Number
ORDER BY Week_Number;
```

Week_Number	User_Count
18	701
19	1054
20	1094
21	1147
22	1113
23	1173
24	1219
25	1263
26	1249
27	1271
28	1355
29	1345
30	1363
31	1443
32	1266
33	1215
34	1203
35	1194

Insights:



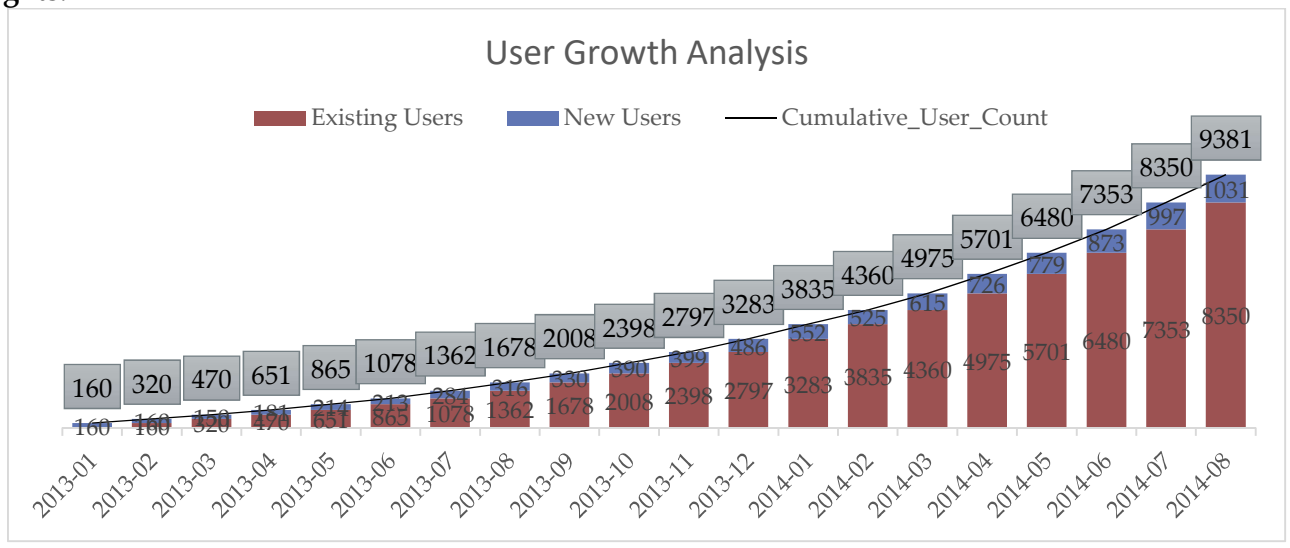
- The highest weekly user engagement was in Week 31 at 1443 users
- There is a slight upward trend in user engagement over the weeks
- The lowest weekly user engagement was in Week 19 at 1054 users. (Week 18 is not taken into account as it does not have all days)

02) User Growth Analysis: Analyse the growth of users over time for a product.
Your Task: Write an SQL query to calculate the user growth for the product.
SQL Query:

```
SELECT
        DATE_FORMAT(created_at,'%Y-%m') AS Creation_Month,
  COUNT(user_id) AS New_Users,
  SUM(COUNT(user_id)) OVER (ORDER BY
DATE_FORMAT(created_at,'%Y-%m') ROWS BETWEEN
UNBOUNDED PRECEDING AND 1 PRECEDING) AS
Existing_Users,
  SUM(COUNT(user_id)) OVER (ORDER BY
DATE_FORMAT(created_at,'%Y-%m') ROWS BETWEEN
UNBOUNDED PRECEDING AND CURRENT ROW) AS
Cumulative_User_Count
FROM users
GROUP BY Creation_Month;
```

Creation_Month	New_Users	Existing_Users	Cumulative_User_Count
2013-01	160	NULL	160
2013-02	160	160	320
2013-03	150	320	470
2013-04	181	470	651
2013-05	214	651	865
2013-06	213	865	1078
2013-07	284	1078	1362
2013-08	316	1362	1678
2013-09	330	1678	2008
2013-10	390	2008	2398
2013-11	399	2398	2797
2013-12	486	2797	3283
2014-01	552	3283	3835
2014-02	525	3835	4360
2014-03	615	4360	4975
2014-04	726	4975	5701
2014-05	779	5701	6480
2014-06	873	6480	7353
2014-07	997	7353	8350
2014-08	1031	8350	9381

Insights:



- We can notice a steady upward increase of weekly user registrations.
- The total number of users are 9381.

03) Weekly Retention Analysis: Analyse the retention of users on a weekly basis after signing up for a product.

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

SQL Query:

```
SELECT
  first_week,
  SUM(CASE WHEN Week_Number = 0 THEN 1 ELSE 0 END) AS week_0,
  SUM(CASE WHEN Week_Number = 1 THEN 1 ELSE 0 END) AS week_1,
  SUM(CASE WHEN Week_Number = 2 THEN 1 ELSE 0 END) AS week_2,
  SUM(CASE WHEN Week_Number = 3 THEN 1 ELSE 0 END) AS week_3,
  SUM(CASE WHEN Week_Number = 4 THEN 1 ELSE 0 END) AS week_4,
  SUM(CASE WHEN Week_Number = 5 THEN 1 ELSE 0 END) AS week_5,
  SUM(CASE WHEN Week_Number = 6 THEN 1 ELSE 0 END) AS week_6,
  SUM(CASE WHEN Week_Number = 7 THEN 1 ELSE 0 END) AS week_7,
  SUM(CASE WHEN Week_Number = 8 THEN 1 ELSE 0 END) AS week_8,
  SUM(CASE WHEN Week_Number = 9 THEN 1 ELSE 0 END) AS week_9,
  SUM(CASE WHEN Week_Number = 10 THEN 1 ELSE 0 END) AS week_10,
  SUM(CASE WHEN Week_Number = 11 THEN 1 ELSE 0 END) AS week_11,
  SUM(CASE WHEN Week_Number = 12 THEN 1 ELSE 0 END) AS week_12,
  SUM(CASE WHEN Week_Number = 13 THEN 1 ELSE 0 END) AS week_13,
  SUM(CASE WHEN Week_Number = 14 THEN 1 ELSE 0 END) AS week_14,
  SUM(CASE WHEN Week_Number = 15 THEN 1 ELSE 0 END) AS week_15,
  SUM(CASE WHEN Week_Number = 16 THEN 1 ELSE 0 END) AS week_16,
  SUM(CASE WHEN Week_Number = 17 THEN 1 ELSE 0 END) AS week_17
FROM
        (SELECT b.user_id, Week_Occurred, First_Week, (Week_Occurred-First_Week) AS Week_Number
         FROM
                (SELECT user_id, week(occurred_at,3) AS Week_Occurred
                 FROM events
                 GROUP BY user_id, Week_Occurred
                 ORDER BY user_id, Week_Occurred) AS a
         RIGHT JOIN
                (SELECT user_id, min(week(occurred_at,3)) AS First_Week
                 FROM events
                 WHERE event_type='signup_flow'
                 GROUP BY user_id
                 ORDER BY user_id) AS b
         ON a.user_id = b.user_id) AS Week_Number_Select
GROUP BY First_Week
ORDER BY First_Week;
```

First_Week	week_0	week_1	week_2	week_3	week_4	week_5	week_6	week_7	week_8	week_9	week_10	week_11	week_12	week_13	week_14	week_15	week_16	week_17
18	81	64	27	19	15	19	13	10	7	9	9	9	8	9	7	5	3	3
19	160	104	67	45	34	22	18	24	11	17	12	11	14	10	9	9	4	0
20	186	144	77	64	40	26	20	19	23	18	15	15	13	11	8	10	0	0
21	177	121	79	50	40	28	21	32	23	23	23	18	18	10	9	0	0	0
22	186	117	78	53	34	24	33	30	18	18	13	14	15	9	0	0	0	0
23	197	133	83	60	49	42	29	25	26	17	17	12	7	0	0	0	0	0
24	198	146	85	56	44	41	34	28	22	21	14	10	0	0	0	0	0	0
25	222	135	89	57	41	31	28	25	15	17	12	0	0	0	0	0	0	0
26	210	151	100	62	44	30	24	19	15	15	0	0	0	0	0	0	0	0
27	199	130	82	60	43	34	33	26	14	0	0	0	0	0	0	0	0	0
28	223	152	95	83	52	39	26	23	0	0	0	0	0	0	0	0	0	0
29	215	144	91	52	33	19	20	0	0	0	0	0	0	0	0	0	0	0
30	228	156	82	59	40	31	0	0	0	0	0	0	0	0	0	0	0	0
31	234	154	94	64	47	0	0	0	0	0	0	0	0	0	0	0	0	0
32	189	126	69	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	250	163	82	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	259	173	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	266	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Insights:

User Retention Percentage (Cohort Analysis) - Weekwise

Start_Week	User_Count	Week_0	Week_1	Week_2	Week_3	Week_4	Week_5	Week_6	Week_7	Week_8	Week_9	Week_10	Week_11	Week_12	Week_13	Week_14	Week_15	Week_16	Week_17
18	81	100.00	79.00	33.00	23.00	19.00	23.00	16.00	12.00	9.00	11.00	11.00	11.00	10.00	11.00	9.00	6.00	4.00	4.00
19	160	100.00	65.00	42.00	28.00	21.00	14.00	11.00	15.00	7.00	11.00	8.00	7.00	9.00	6.00	6.00	6.00	3.00	
20	186	100.00	77.00	41.00	34.00	22.00	14.00	11.00	10.00	12.00	10.00	8.00	8.00	7.00	6.00	4.00	5.00		
21	177	100.00	68.00	45.00	28.00	23.00	16.00	12.00	18.00	13.00	13.00	13.00	10.00	10.00	6.00	5.00			
22	186	100.00	63.00	42.00	28.00	18.00	13.00	18.00	16.00	10.00	10.00	7.00	8.00	8.00	5.00				
23	197	100.00	68.00	42.00	30.00	25.00	21.00	15.00	13.00	13.00	9.00	9.00	6.00	4.00					
24	198	100.00	74.00	43.00	28.00	22.00	21.00	17.00	14.00	11.00	11.00	7.00	5.00						
25	222	100.00	61.00	40.00	26.00	18.00	14.00	13.00	11.00	7.00	8.00	5.00							
26	210	100.00	72.00	48.00	30.00	21.00	14.00	11.00	9.00	7.00	7.00								
27	199	100.00	65.00	41.00	30.00	22.00	17.00	17.00	13.00	7.00									
28	223	100.00	68.00	43.00	37.00	23.00	17.00	12.00	10.00										
29	215	100.00	67.00	42.00	24.00	15.00	9.00	9.00											
30	228	100.00	68.00	36.00	26.00	18.00	14.00												
31	234	100.00	66.00	40.00	27.00	20.00													
32	189	100.00	67.00	37.00	25.00														
33	250	100.00	65.00	33.00															
34	259	100.00	67.00																
35	266	100.00																	

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

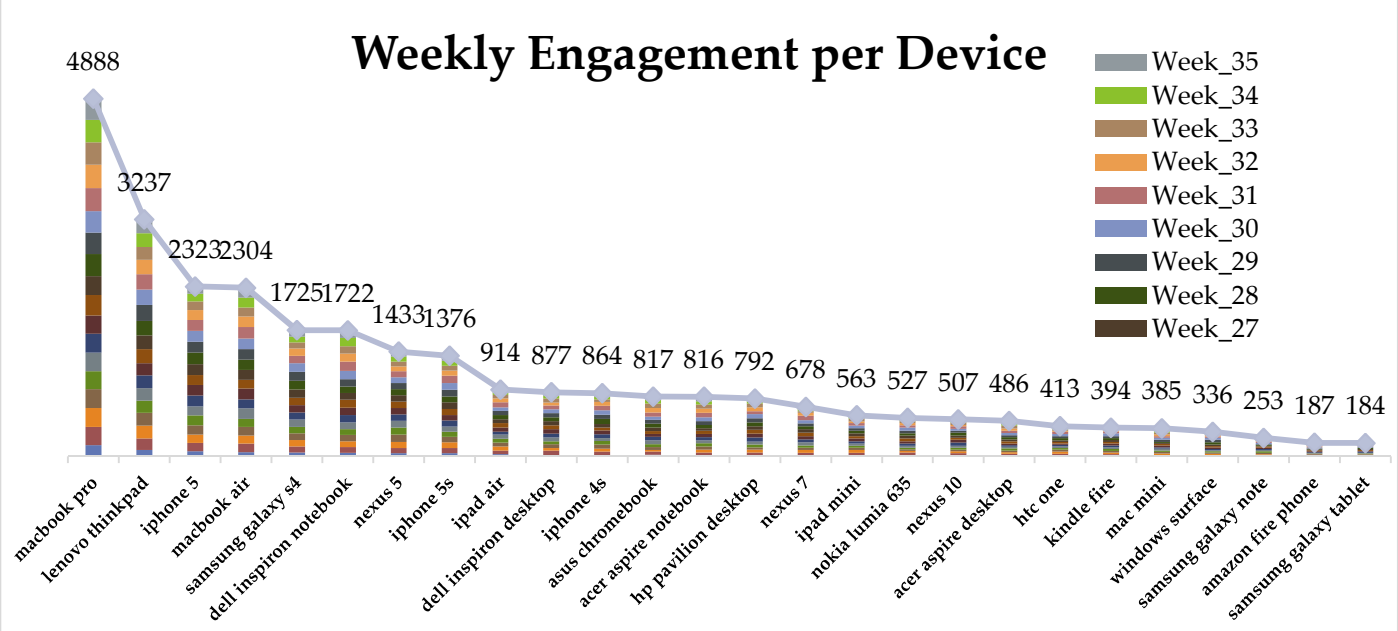
Your Task: Write an SQL query to calculate the weekly engagement per device.

SQL Query:

```
SELECT
        Device,
        SUM(CASE WHEN Week_Number = 18 THEN Tot_Count ELSE 0 END) AS Week_18,
        SUM(CASE WHEN Week_Number = 19 THEN Tot_Count ELSE 0 END) AS Week_19,
        SUM(CASE WHEN Week_Number = 20 THEN Tot_Count ELSE 0 END) AS Week_20,
        SUM(CASE WHEN Week_Number = 21 THEN Tot_Count ELSE 0 END) AS Week_21,
        SUM(CASE WHEN Week_Number = 22 THEN Tot_Count ELSE 0 END) AS Week_22,
        SUM(CASE WHEN Week_Number = 23 THEN Tot_Count ELSE 0 END) AS Week_23,
        SUM(CASE WHEN Week_Number = 24 THEN Tot_Count ELSE 0 END) AS Week_24,
        SUM(CASE WHEN Week_Number = 25 THEN Tot_Count ELSE 0 END) AS Week_25,
        SUM(CASE WHEN Week_Number = 26 THEN Tot_Count ELSE 0 END) AS Week_26,
        SUM(CASE WHEN Week_Number = 27 THEN Tot_Count ELSE 0 END) AS Week_27,
        SUM(CASE WHEN Week_Number = 28 THEN Tot_Count ELSE 0 END) AS Week_28,
        SUM(CASE WHEN Week_Number = 29 THEN Tot_Count ELSE 0 END) AS Week_29,
        SUM(CASE WHEN Week_Number = 30 THEN Tot_Count ELSE 0 END) AS Week_30,
        SUM(CASE WHEN Week_Number = 31 THEN Tot_Count ELSE 0 END) AS Week_31,
        SUM(CASE WHEN Week_Number = 32 THEN Tot_Count ELSE 0 END) AS Week_32,
        SUM(CASE WHEN Week_Number = 33 THEN Tot_Count ELSE 0 END) AS Week_33,
        SUM(CASE WHEN Week_Number = 34 THEN Tot_Count ELSE 0 END) AS Week_34,
        SUM(CASE WHEN Week_Number = 35 THEN Tot_Count ELSE 0 END) AS Week_35,
        SUM(Tot_Count) AS Total_Count
FROM
        (SELECT WEEK(occurred_at,3) AS Week_Number,device AS Device,COUNT(DISTINCT user_id) AS
Tot_Count
        FROM events WHERE event_type='engagement'
        GROUP BY user_id, Device, Week_Number
        ORDER BY user_id, device) AS Query
GROUP BY Device
ORDER BY Total_Count DESC;
```

Device	Week_18	Week_19	Week_20	Week_21	Week_22	Week_23	Week_24	Week_25	Week_26	Week_27	Week_28	Week_29	Week_30	Week_31	Week_32	Week_33	Week_34	Week_35	Total_Count
macbook pro	154	248	261	256	244	254	259	251	276	259	301	295	291	317	317	307	308	290	4888
lenovo thinkpad	90	155	176	177	164	170	176	164	196	188	195	220	209	208	196	177	190	186	3237
iphone 5	70	114	113	128	136	122	151	143	134	150	159	148	147	151	133	119	105	100	2323
macbook air	57	119	110	119	107	145	122	149	119	134	140	145	146	156	143	124	134	135	2304
samsung galaxy s4	56	80	90	92	84	103	95	102	100	114	119	120	117	104	99	85	76	89	1725
dell inspiron notebook	49	78	82	84	81	91	100	102	108	90	91	100	114	125	111	101	111	104	1722
nexus 5	43	73	84	99	94	95	87	85	90	86	83	83	81	81	65	66	71	67	1433
iphone 5s	45	70	77	75	71	71	80	78	80	92	79	93	92	100	71	65	69	68	1376
ipad air	30	52	53	54	51	57	42	58	57	55	57	55	50	70	53	45	39	36	914
dell inspiron desktop	21	58	36	52	41	53	53	57	53	59	52	55	52	53	42	57	35	48	877
iphone 4s	21	47	40	56	46	41	52	52	39	49	68	58	61	63	52	35	34	50	864
asus chromebook	23	42	26	39	38	51	48	41	40	47	52	51	47	56	59	61	48	48	817
acer aspire notebook	21	34	40	40	47	39	43	42	44	35	47	50	52	62	56	56	45	63	816
hp pavilion desktop	15	37	40	31	42	38	55	56	50	47	56	55	57	39	52	51	34	37	792
nexus 7	20	29	41	31	29	44	37	47	49	45	41	39	43	60	39	24	29	31	678
ipad mini	21	29	37	32	25	32	32	38	31	41	35	34	36	34	23	31	27	25	563
nokia lumia 635	19	34	22	21	25	25	31	32	37	41	31	34	42	33	30	26	27	17	527
nexus 10	16	30	25	23	24	28	43	38	30	29	38	28	25	35	19	31	22	23	507
acer aspire desktop	10	26	22	23	28	25	21	23	29	28	29	27	29	32	31	37	36	30	486
htc one	16	19	32	27	20	24	21	19	20	23	28	26	32	30	13	18	19	26	413
kindle fire	6	26	20	22	30	21	25	25	24	26	25	29	36	24	14	12	14	15	394
mac mini	8	12	19	25	18	24	17	29	22	11	15	28	28	23	24	21	32	29	385
windows surface	10	10	15	19	17	15	16	21	19	22	31	33	28	18	18	10	14	20	336
samsung galaxy note	7	15	11	18	20	19	12	19	14	10	14	10	17	15	14	12	13	13	253
amazon fire phone	4	9	12	10	4	5	16	11	12	13	10	6	12	12	14	12	14	11	187
samsumg galaxy tablet	8	11	6	9	6	11	14	11	12	12	15	9	13	9	7	6	12	13	184

Insights:



- The device with highest weekly engagement is Macbook Pro with an average engagement of 271.5.
- The device with lowest weekly engagement is Samsung Galaxy Tablet with an average engagement of 10.2.

05) Email Engagement Analysis: Analyse how users are engaging with the email service.

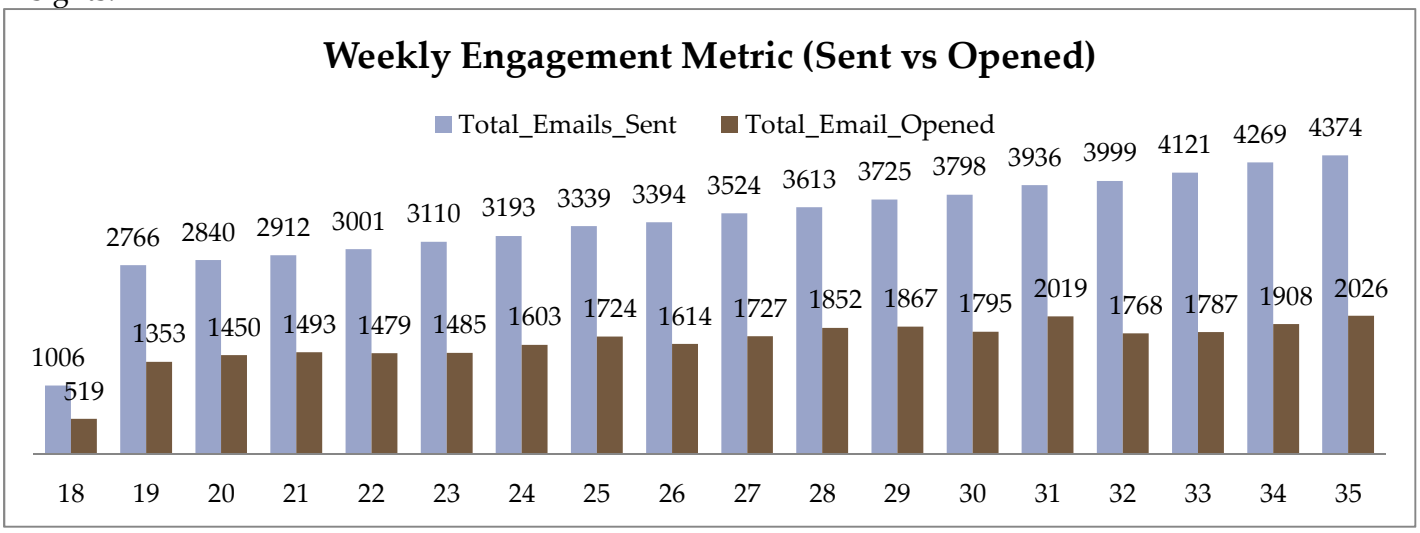
Your Task: Write an SQL query to calculate the email engagement metrics.

SQL Query:

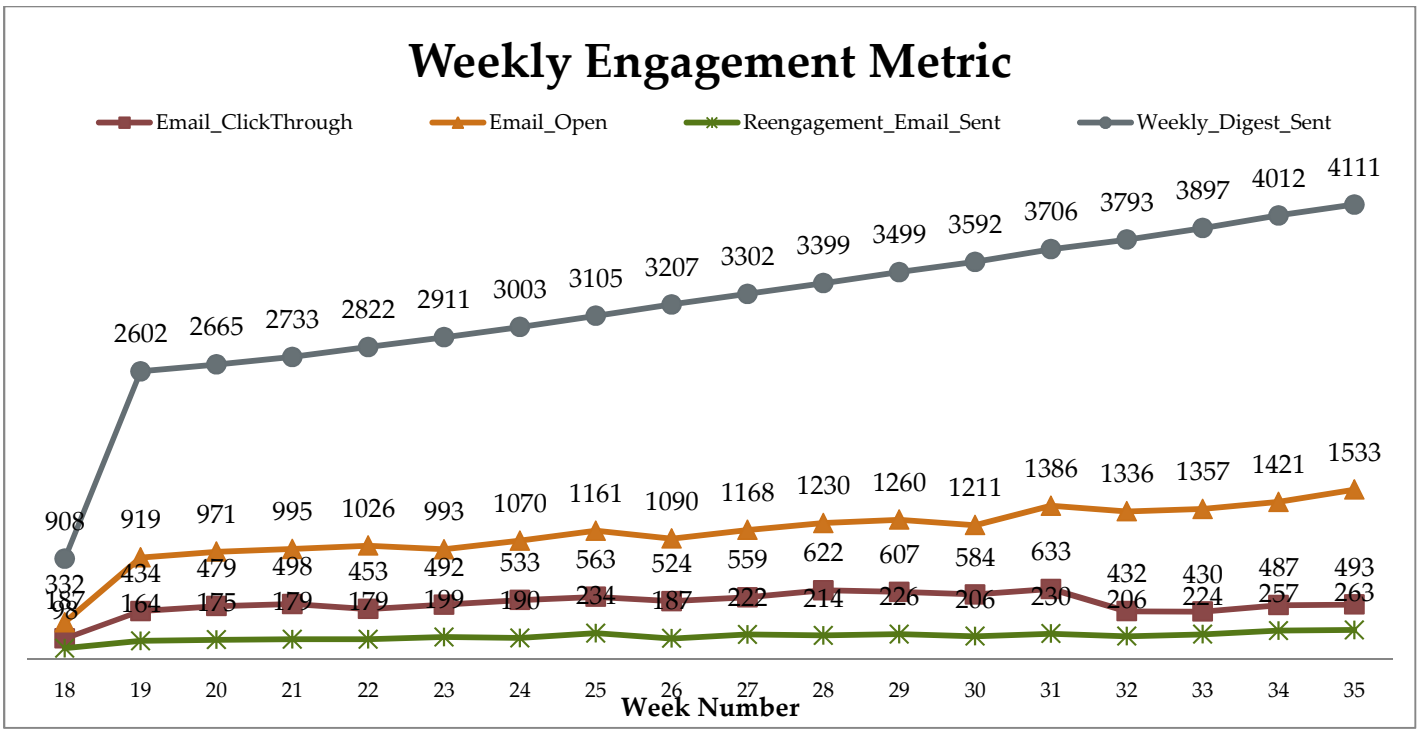
```
SELECT
        WEEK(occurred_at, 3) AS Week_Number,
        COUNT((CASE WHEN action = 'email_clickthrough' THEN user_id END)) AS Email_ClickThrough,
        COUNT((CASE WHEN action = 'email_open' THEN user_id END)) AS Email_Open,
        COUNT((CASE WHEN action = 'email_clickthrough' THEN user_id END)) + COUNT((CASE WHEN
action = 'email_open' THEN user_id END)) AS Total_Email_Opened,
        COUNT((CASE WHEN action = 'sent_reengagement_email' THEN user_id END)) AS
Reengagement_Email_Sent,
        COUNT((CASE WHEN action = 'sent_Weekly_digest' THEN user_id END)) AS Weekly_Digest_Sent,
        COUNT((CASE WHEN action = 'sent_reengagement_email' THEN user_id END)) + COUNT((CASE
WHEN action = 'sent_Weekly_digest' THEN user_id END)) AS Total_Emails_Sent,
        COUNT(DISTINCT user_id) AS Tot_User_Count
FROM
        email_events
GROUP BY Week_Number
ORDER BY Week_Number;
```

Week_Number	Email_ClickThrough	Email_Open	Total_Email_Opened	Reengagement_Email_Sent	Weekly_Digest_Sent	Total_Emails_Sent	Tot_User_Count
18	187	332	519	98	908	1006	1006
19	434	919	1353	164	2602	2766	2724
20	479	971	1450	175	2665	2840	2801
21	498	995	1493	179	2733	2912	2876
22	453	1026	1479	179	2822	3001	2945
23	492	993	1485	199	2911	3110	3047
24	533	1070	1603	190	3003	3193	3143
25	563	1161	1724	234	3105	3339	3272
26	524	1090	1614	187	3207	3394	3340
27	559	1168	1727	222	3302	3524	3461
28	622	1230	1852	214	3399	3613	3557
29	607	1260	1867	226	3499	3725	3675
30	584	1211	1795	206	3592	3798	3748
31	633	1386	2019	230	3706	3936	3883
32	432	1336	1768	206	3793	3999	3953
33	430	1357	1787	224	3897	4121	4061
34	487	1421	1908	257	4012	4269	4209
35	493	1533	2026	263	4111	4374	4309

Insights:



- Less than 50% of the emails sent are opened by users



- Most of the emails sent are Weekly Digests
- Less than 50% Users are opening emails, out of which less than 50% are clicking through.

Results

The project allowed me to advance my SQL skills through research and further learning. It has also allowed me to provide valuable insights based on the given data.

The project was a great learning experience as I was able to research new concepts in SQL, Excel and Data Visualisation. I also gained hands on experience in a real-world project, which allowed me to learn new business concepts related to various metrics.

Through the use of SQL queries, I was able to extract insightful analysis from operational data. The insight gained from this data will help improve the company's operations and understand sudden changes in key metrics.

Project 4: Hiring Process Analytics

Project Description

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 at a multinational company like Google, my task is to analyse the company's hiring process data and draw meaningful insights from it.
Using a dataset containing records of previous hires, I have to analyze this data and answer certain questions that can help the company improve its hiring process.

Approach

To analyse the data we will be using Exploratory Data Analysis (EDA) process. Exploratory Data Analysis is a process of examining or understanding the data and extracting insights or main characteristics of the data.
The goal of this project is to use my knowledge of Statistics and Excel to draw meaningful conclusions about the company's hiring process. These insights could potentially help the company improve its hiring process and make better hiring decisions in the future.
As part of the data cleaning, I made the below changes:

- Changed the column header from 'event_name' to 'Gender'
- Changed the value of "-" in the column Gender to "Don't want to say"
- Removed the word 'Department' from all Department Names
- Removed the employees that are Hired but the post is not mentioned (1 record)
- Changed the case of Post Name records to UPPERCASE and also C-10 to C10
- Removed records that do not have an 'Offered Salary' value (1 record) and converted 'Offered Salary' to 'Currency' data type.

Total Number of Records after data cleaning is 7166

- **Handling Missing Data -** A check will be conducted to see if there are any missing values in the dataset. If there are, a decision will be made to the best strategy on handling them.
- **Clubbing Columns -** To simplify the analysis, any columns with multiple categories will be combined.
- **Outlier Detection -** Outliers may skew the analysis and therefore there needs to be a check for outliers.
- **Removing Outliers -** Once the outliers are found, a decision needs to be made as to the best strategy to handle them. Some of the ways could be to remove them, replace them, or leaving them alone, depending on the situation.
- **Data Summary -** After cleaning and preparing the data, a summarization of the findings needs to be done. This will involve calculating averages, medians, or other statistical measures. It can also involve creating visualizations to better understand the data.

Tech Stack Used

Microsoft Office 2010 Professional Plus Version 14.0.7268.5000

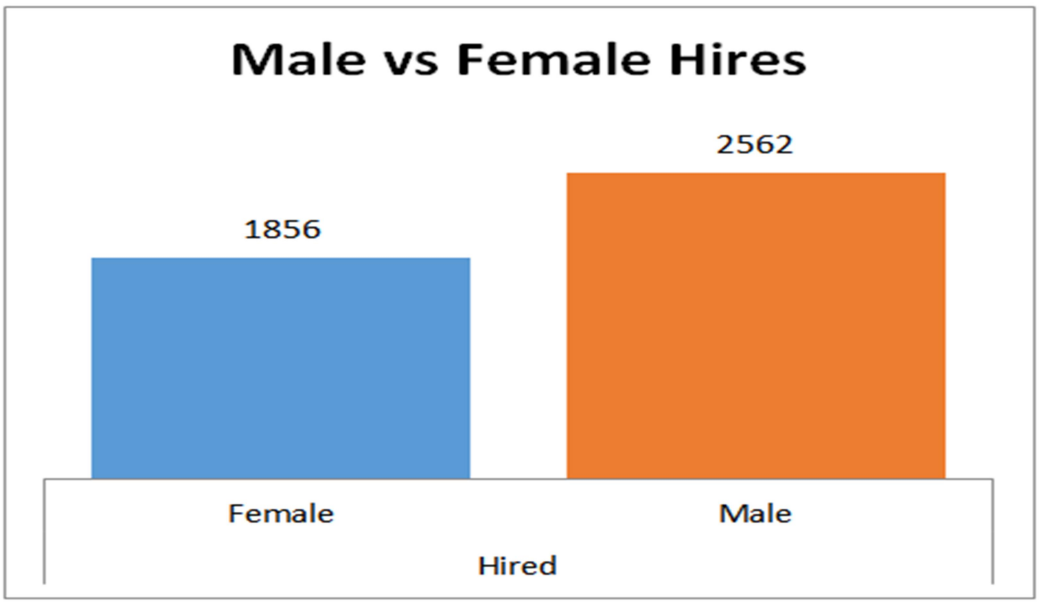
Charts & Insights

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

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

Using a Pivot Table		
Status	**Gender**	**Count of application_id**
Hired	Female	1856
	Male	2562
Hired Total		**4418**
Grand Total		**4418**

Using Formulas	
	Count
Total Hired	**4696**
Female	**1856**
Male	**2562**
Total Male & Female	**4418**



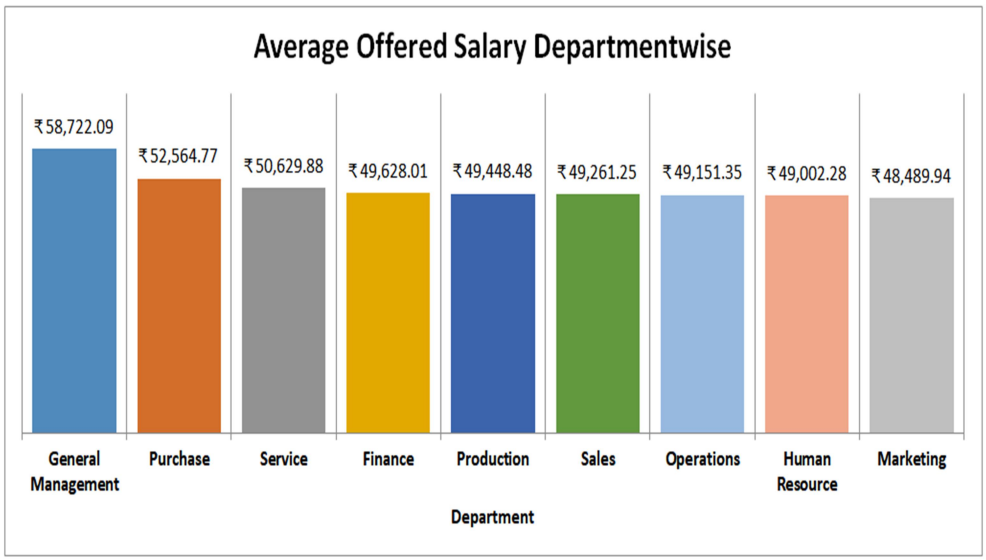
Q. B) Salary Analysis: The average salary is calculated by adding up the salaries of a group of employees and then dividing the total by the number of employees.

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

Using Step-by-Step Formulas	
Total Salary Offered to Applicants	**₹ 35,81,42,455.00**
Total Count of Applicants	**7166**
Average Salary	**₹ 49,978.01**

Using Average Formula	
Average Salary of Employees	**₹ 49,978.01**

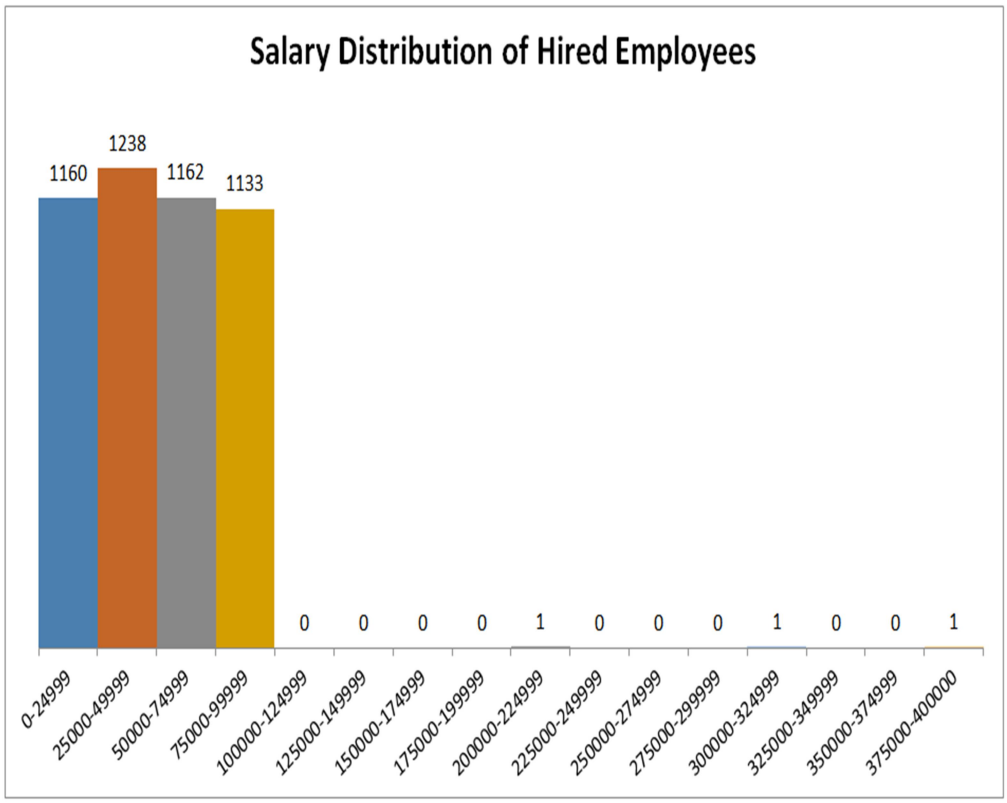
Using Pivot Table	
Average of Offered Salary	
Department	**Total**
General Management	₹ 58,722.09
Purchase	₹ 52,564.77
Service	₹ 50,629.88
Finance	₹ 49,628.01
Production	₹ 49,448.48
Sales	₹ 49,261.25
Operations	₹ 49,151.35
Human Resource	₹ 49,002.28
Marketing	₹ 48,489.94
Grand Total	**₹ 49,978.01**



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

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

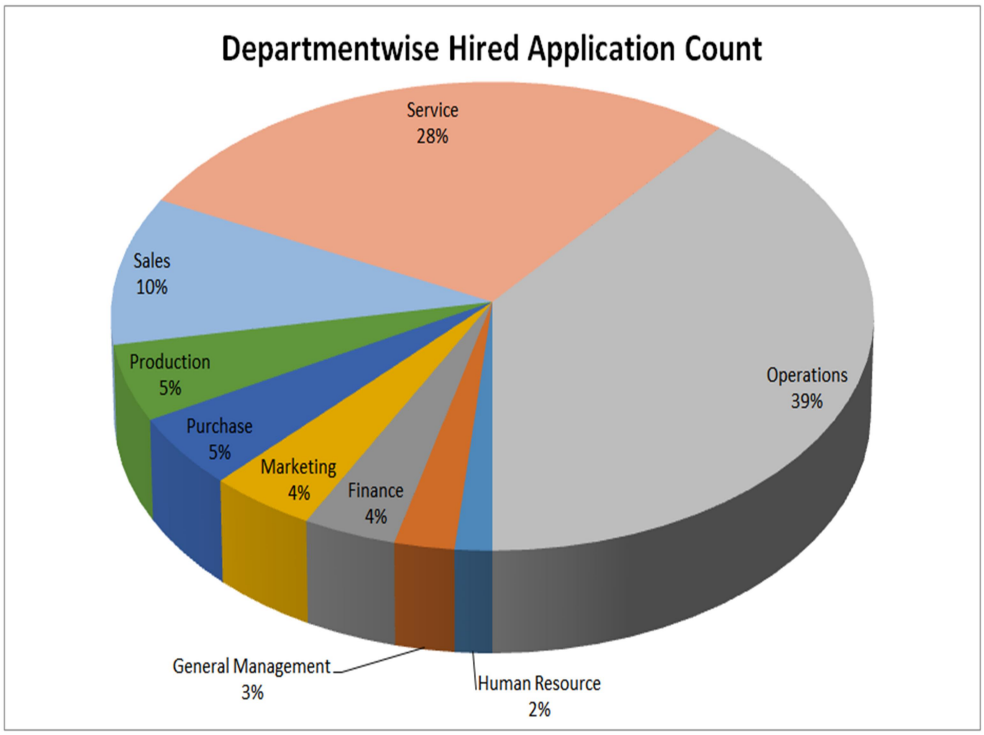
Class Interval Used	25000
Using Pivot Table	
Count of application_id	**Column Labels**
Row Labels	**Hired**
0-24999	1160
25000-49999	1238
50000-74999	1162
75000-99999	1133
100000-124999	0
125000-149999	0
150000-174999	0
175000-199999	0
200000-224999	1
225000-249999	0
250000-274999	0
275000-299999	0
300000-324999	1
325000-349999	0
350000-374999	0
375000-400000	1
Grand Total	**4696**



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

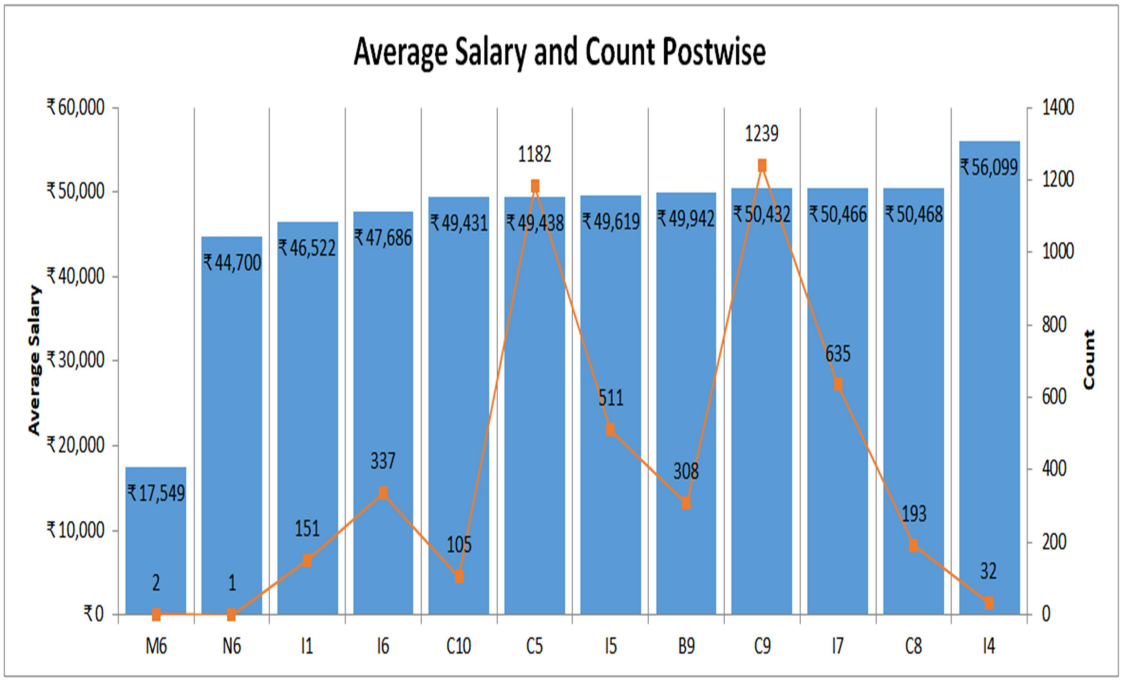
Your Task: Use a pie chart, bar graph, or any other suitable visualization to show the proportion of people working in different departments.

Status	Hired
Row Labels	**Count of application_id**
Human Resource	70
General Management	113
Finance	176
Marketing	202
Purchase	230
Production	246
Sales	484
Service	1332
Operations	1843
Grand Total	**4696**



Q. E) Position Tier Analysis: Different positions within a company often have different tiers or levels.
Your Task: Use a chart or graph to represent the different position tiers within the company. This will help you understand the distribution of positions across different tiers.

Status	Hired	
Row Labels	**Average of Offered Salary**	**Count of application_id**
M6	₹17,549	2
N6	₹44,700	1
I1	₹46,522	151
I6	₹47,686	337
C10	₹49,431	105
C5	₹49,438	1182
I5	₹49,619	511
B9	₹49,942	308
C9	₹50,432	1239
I7	₹50,466	635
C8	₹50,468	193
I4	₹56,099	32
Grand Total	**₹49,745**	**4696**



Insights

- For the data period, 2562 males and 1856 females have been hired for various positions in the company.
- The average salary offered to applicants in this company is ₹ 49978.01.
- The highest average salary offered is for the "General Management Department" and the lowest average salary is for the "Marketing Department".
- The maximum hired employees are offered salary in the range of ₹ 25,000 to ₹ 49999.
- The "Operations Department" has the maximum employees, whereas the "Human resource department" has the least number of employees.
- The maximum number of employees have the Post Name C9
- The Post Name with the highest salary average is I4 and the lowest salary average is M6.

Results

The Hiring Process Analytics are important for a company as it helps to make better decisions when it comes to hiring process, which can potentially improve the company's hiring process.
The Hiring Process Analytics are checked on a monthly, quarterly or yearly basis as per the company's requirement.
The project has helped me understand the Exploratory Data Analysis process better. It has also helped me improve my knowledge of Excel and Statistics and its working, by allowing me to utilize basic and advanced concepts to attain the insights.

Link to Statistics Excel File

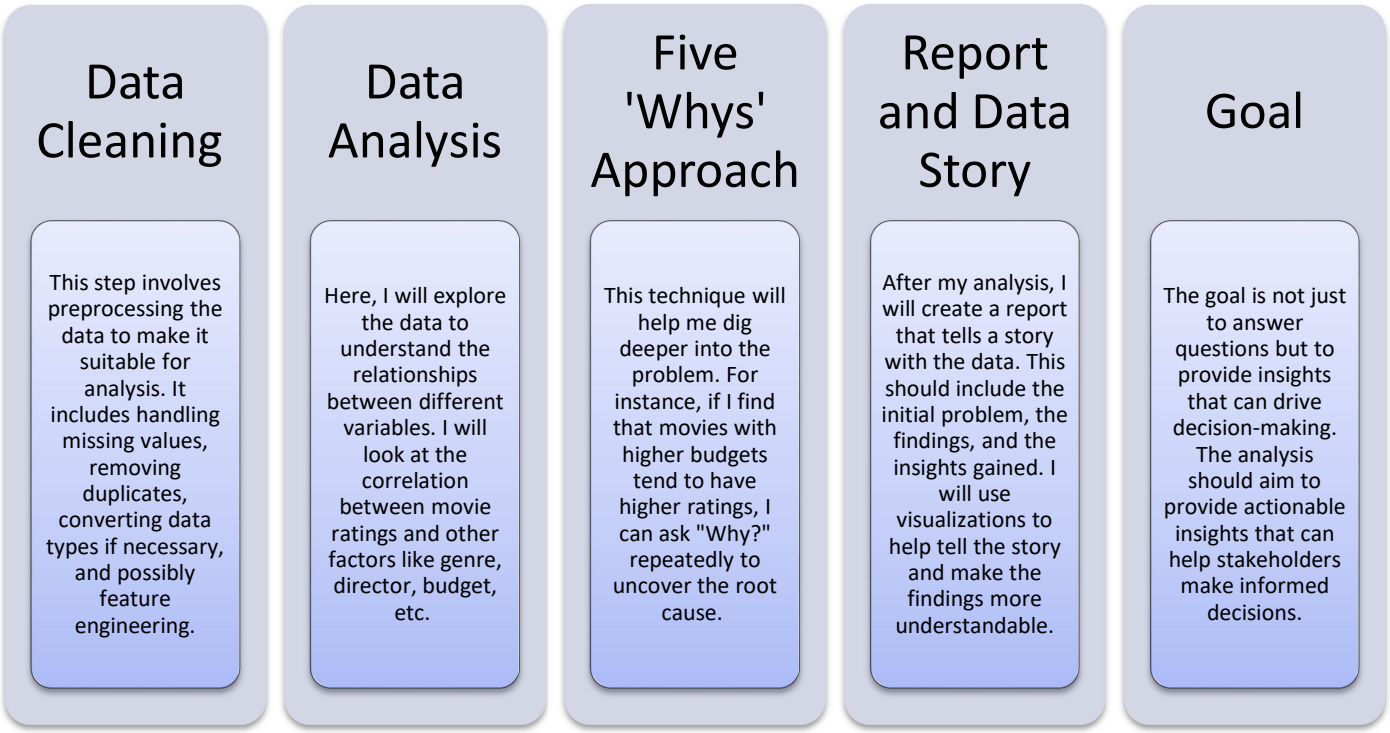
Project 5 - IMDb Movie Analysis

Project Description

IMDb (Internet Movie Database) is an online database of information related to films, television series, podcasts, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews.

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 projects.

Approach



Data Cleaning

As part of the data cleaning, I made the below changes:

- Determined the main columns required : IMDb Score, Movie Title, Genre, Movie Duration, Language, Director Name, Budget & Gross Earnings
- Created a new column "Profit Margin" = ("Gross Earnings"-"Budget")
- Removed rows that had blanks in the required columns
- Removed duplicates in the required parameters – The Movie Title "The Host" was duplicate, but it was two different movies having the same name.

Total Number of Records before "Data Cleaning" was 5043
Total Number of Records after "Data Cleaning" became 3786

Tech Stack Used

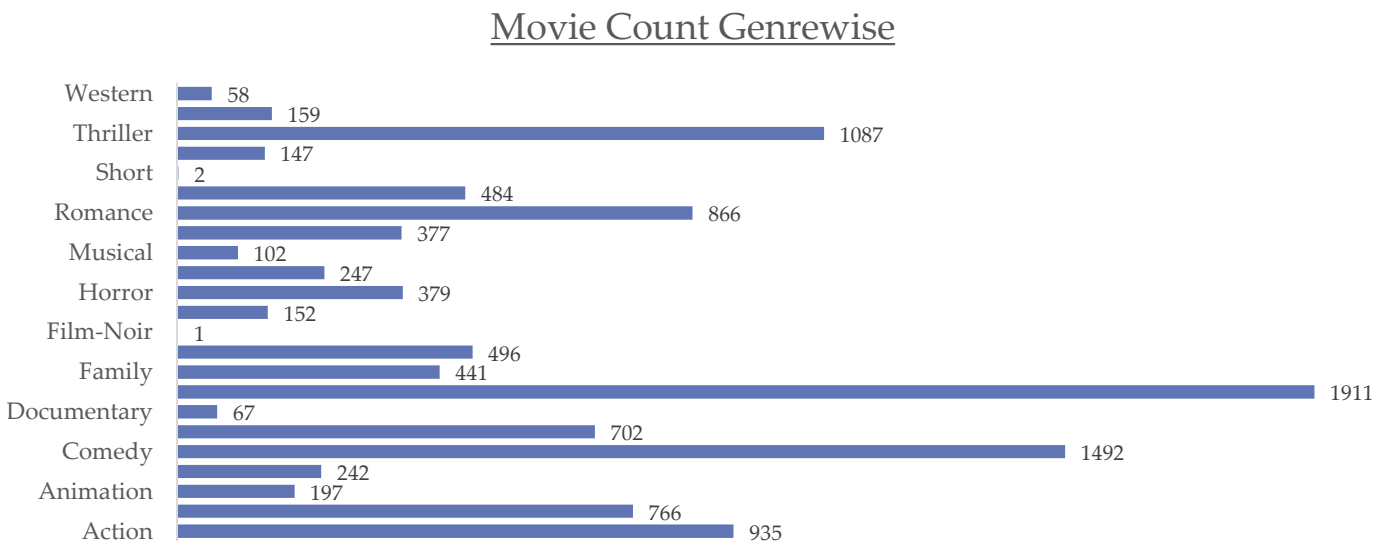
Microsoft 365 Online Excel Free Version 16.0.17012.41002

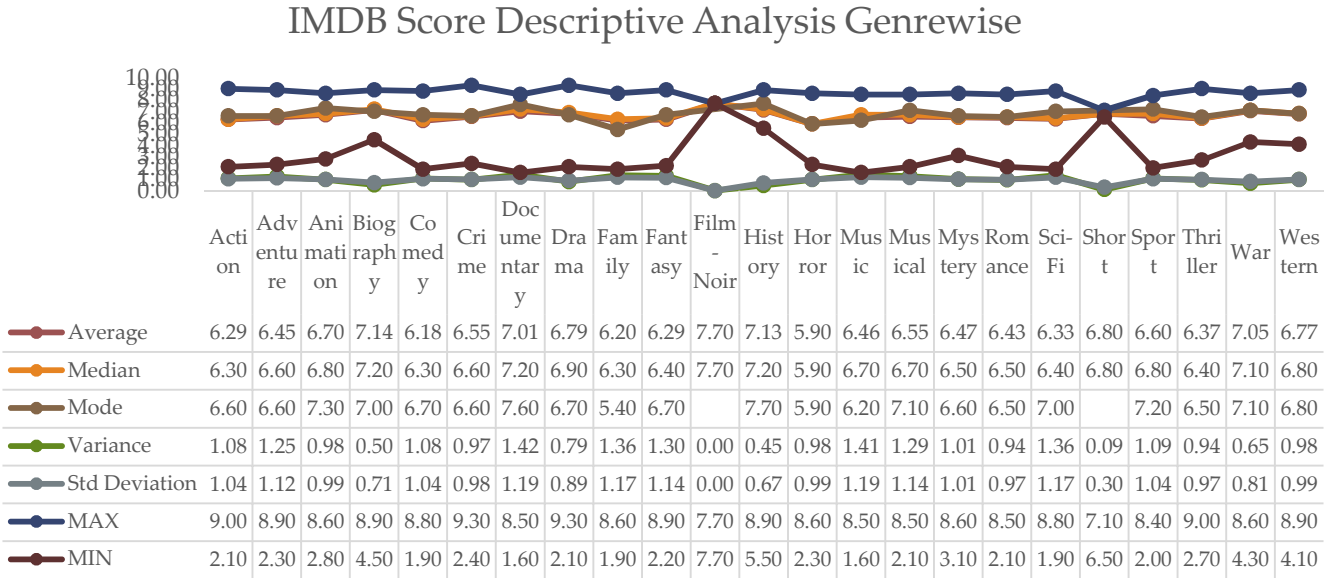
Charts & Insights

Q. A) Movie Genre Analysis: Analyze the distribution of movie genres and their impact on the IMDB score.
Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

Movie Count and IMDB Score Descriptive Analysis Genrewise

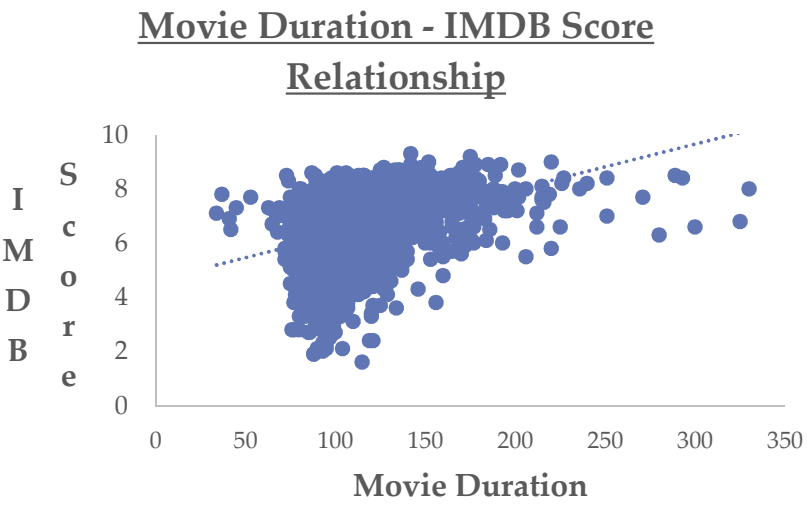
Unique Genres	Movie Genre Count	IMDB Score Average	IMDB Score Median	IMDB Score Mode	IMDB Score Variance	IMDB Score Standard Deviation	IMDB Score MAX	IMDB Score MIN	Lower Quartile	Upper Quartile
Action	935	6.29	6.30	6.60	1.08	1.04	9.00	2.10	5.70	6.90
Adventure	766	6.45	6.60	6.60	1.25	1.12	8.90	2.30	5.80	7.20
Animation	197	6.70	6.80	7.30	0.98	0.99	8.60	2.80	6.10	7.30
Biography	242	7.14	7.20	7.00	0.50	0.71	8.90	4.50	6.80	7.60
Comedy	1492	6.18	6.30	6.70	1.08	1.04	8.80	1.90	5.60	6.90
Crime	702	6.55	6.60	6.60	0.97	0.98	9.30	2.40	6.00	7.20
Documentary	67	7.01	7.20	7.60	1.42	1.19	8.50	1.60	6.75	7.70
Drama	1911	6.79	6.90	6.70	0.79	0.89	9.30	2.10	6.30	7.40
Family	441	6.20	6.30	5.40	1.36	1.17	8.60	1.90	5.40	7.00
Fantasy	496	6.29	6.40	6.70	1.30	1.14	8.90	2.20	5.60	7.00
Film-Noir	1	7.70	7.70	#N/A	0.00	0.00	7.70	7.70	7.70	7.70
History	152	7.13	7.20	7.70	0.45	0.67	8.90	5.50	6.70	7.60
Horror	379	5.90	5.90	5.90	0.98	0.99	8.60	2.30	5.20	6.60
Music	247	6.46	6.70	6.20	1.41	1.19	8.50	1.60	5.85	7.30
Musical	102	6.55	6.70	7.10	1.29	1.14	8.50	2.10	5.90	7.40
Mystery	377	6.47	6.50	6.60	1.01	1.01	8.60	3.10	5.90	7.20
Romance	866	6.43	6.50	6.50	0.94	0.97	8.50	2.10	5.90	7.10
Sci-Fi	484	6.33	6.40	7.00	1.36	1.17	8.80	1.90	5.70	7.10
Short	2	6.80	6.80	#N/A	0.09	0.30	7.10	6.50	6.65	6.95
Sport	147	6.60	6.80	7.20	1.09	1.04	8.40	2.00	6.15	7.20
Thriller	1087	6.37	6.40	6.50	0.94	0.97	9.00	2.70	5.80	7.00
War	159	7.05	7.10	7.10	0.65	0.81	8.60	4.30	6.50	7.60
Western	58	6.77	6.80	6.80	0.98	0.99	8.90	4.10	6.13	7.50





Q. B) Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.

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



Movie Duration Descriptive Analysis

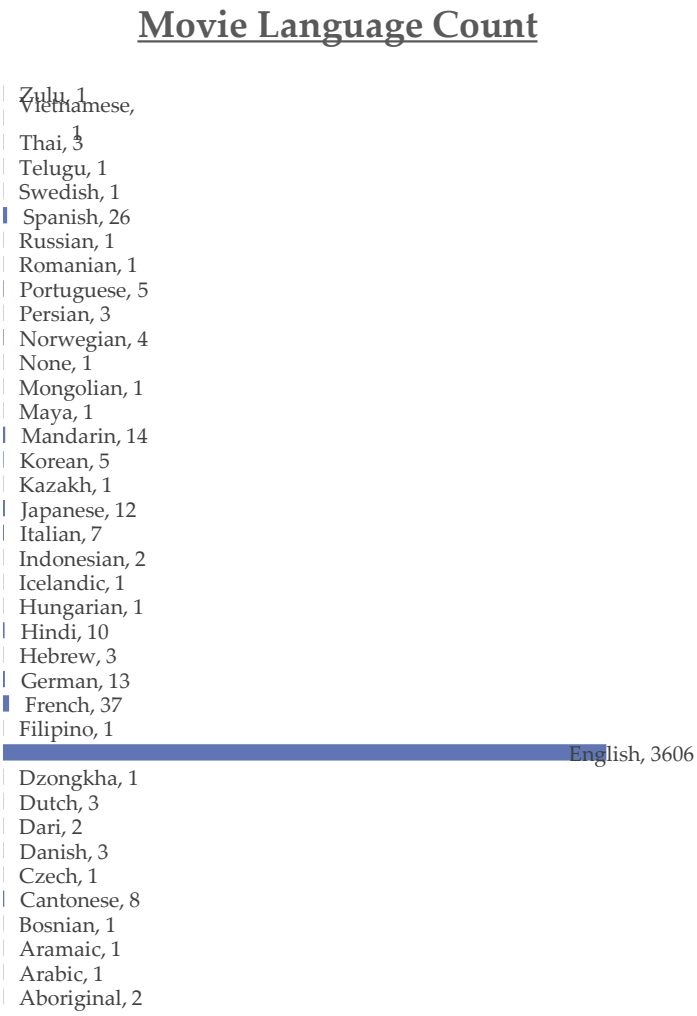
Movie Duration Group	Movie Count	Average Duration	Median Duration	Stdev Duration
Short (40 mins or less)	2	35.5	35.5	1.5
Short-Medium (41-110 mins)	2307	96.74	98	8.61
Medium (111-180 mins)	1425	127.19	123	14.29
Long (Above 180 mins)	52	216.4	202	37.31
Total	3786	109.81	105	22.76

Q.C) Language Analysis: Examine the distribution of movies based on their language.

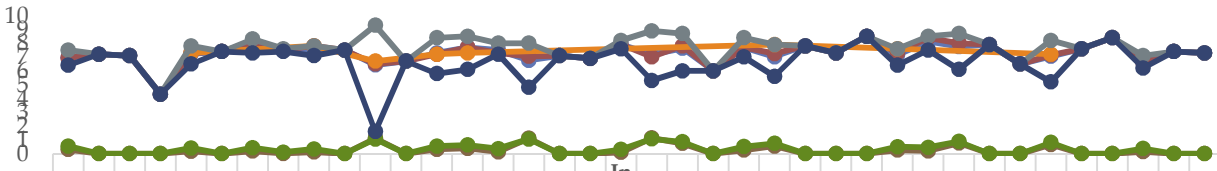
Task: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

Movie Count and IMDB Score Descriptive Analysis Languagewise

Unique Languages	Movie Language Count	IMDB Score Average	IMDB Score Median	IMDB Score Mode	IMDB Score Variance	IMDB Score Standard Deviation	IMDB Score MAX	IMDB Score MIN
Aboriginal	2	6.95	6.95	#N/A	0.3	0.55	7.5	6.4
Arabic	1	7.2	7.2	#N/A	0	0	7.2	7.2
Aramaic	1	7.1	7.1	#N/A	0	0	7.1	7.1
Bosnian	1	4.3	4.3	#N/A	0	0	4.3	4.3
Cantonese	8	7.24	7.3	7.3	0.17	0.41	7.8	6.5
Czech	1	7.4	7.4	#N/A	0	0	7.4	7.4
Danish	3	7.9	8.1	#N/A	0.19	0.43	8.3	7.3
Dari	2	7.5	7.5	#N/A	0.01	0.1	7.6	7.4
Dutch	3	7.57	7.8	7.8	0.11	0.33	7.8	7.1
Dzongkha	1	7.5	7.5	#N/A	0	0	7.5	7.5
English	3606	6.42	6.5	6.7	1.11	1.05	9.3	1.6
Filipino	1	6.7	6.7	#N/A	0	0	6.7	6.7
French	37	7.29	7.2	7.2	0.31	0.55	8.4	5.8
German	13	7.69	7.7	7.3	0.38	0.62	8.5	6.1
Hebrew	3	7.5	7.3	#N/A	0.13	0.36	8	7.2
Hindi	10	6.76	7.05	#N/A	1.11	1.05	8	4.8
Hungarian	1	7.1	7.1	#N/A	0	0	7.1	7.1
Icelandic	1	6.9	6.9	#N/A	0	0	6.9	6.9
Indonesian	2	7.9	7.9	#N/A	0.09	0.3	8.2	7.6
Italian	7	7.19	7	#N/A	1.14	1.07	8.9	5.3
Japanese	12	7.63	7.8	#N/A	0.74	0.86	8.7	6
Kazakh	1	6	6	#N/A	0	0	6	6
Korean	5	7.7	7.7	#N/A	0.26	0.51	8.4	7
Mandarin	14	7.02	7.25	7.9	0.54	0.74	7.9	5.6
Maya	1	7.8	7.8	#N/A	0	0	7.8	7.8
Mongolian	1	7.3	7.3	#N/A	0	0	7.3	7.3
None	1	8.5	8.5	#N/A	0	0	8.5	8.5
Norwegian	4	7.15	7.3	7.6	0.25	0.5	7.6	6.4
Persian	3	8.13	8.4	#N/A	0.2	0.45	8.5	7.5
Portuguese	5	7.76	8	#N/A	0.77	0.88	8.7	6.1
Romanian	1	7.9	7.9	#N/A	0	0	7.9	7.9
Russian	1	6.5	6.5	#N/A	0	0	6.5	6.5
Spanish	26	7.05	7.15	7.2	0.66	0.81	8.2	5.2
Swedish	1	7.6	7.6	#N/A	0	0	7.6	7.6
Telugu	1	8.4	8.4	#N/A	0	0	8.4	8.4
Thai	3	6.63	6.6	#N/A	0.14	0.37	7.1	6.2
Vietnamese	1	7.4	7.4	#N/A	0	0	7.4	7.4
Zulu	1	7.3	7.3	#N/A	0	0	7.3	7.3



IMDB Score Descriptive Analysis Languagewise



	Aboriginal	Arabic	Aramaic	Bosnian	Cantonese	Czech	Danish	Dari	Dutch	Dzongkha	English	Filipino	French	German	Hebrew	Hindi	Hungarian	Icelandic	Indonesian	Italian	Japanese	Kazakh	Korean	Mandarin	Maya	Mongolian	None	Norwegian	Persian	Portuguese	Romanian	Russian	Spanish	Swedish	Tulu	Thai	Vietnamese	Zulu
Average	7	7.2	7.1	4.3	7.2	7.4	7.9	7.5	7.6	7.5	6.4	6.7	7.3	7.7	7.5	6.8	7.1	6.9	7.9	7.2	7.6	6	7.7	7	7.8	7.3	8.5	7.2	8.1	7.8	7.9	6.5	7.1	7.6	8.4	6.6	7.4	7.3
Median	7	7.2	7.1	4.3	7.3	7.4	8.1	7.5	7.8	7.5	6.5	6.7	7.2	7.7	7.3	7.1	7.1	6.9	7.9	7	7.8	6	7.7	7.3	7.8	7.3	8.5	7.3	8.4	8	7.9	6.5	7.2	7.6	8.4	6.6	7.4	7.3
Mode					7.3				7.8		6.7		7.2	7.3										7.9				7.6					7.2						
Variance	0.3	0	0	0	0.2	0	0.2	0	0.1	0	1.1	0	0.3	0.4	0.1	1.1	0	0	0.1	1.1	0.7	0	0.3	0.5	0	0	0	0.3	0.2	0.8	0	0	0.7	0	0	0.1	0	0
Std Dev	0.6	0	0	0	0.4	0	0.4	0.1	0.3	0	1.1	0	0.6	0.6	0.4	1.1	0	0	0.3	1.1	0.9	0	0.5	0.7	0	0	0	0.5	0.5	0.9	0	0	0.8	0	0	0.4	0	0
MAX	7.5	7.2	7.1	4.3	7.8	7.4	8.3	7.6	7.8	7.5	9.3	6.7	8.4	8.5	8	8	7.1	6.9	8.2	8.9	8.7	6	8.4	7.9	7.8	7.3	8.5	7.6	8.5	8.7	7.9	6.5	8.2	7.6	8.4	7.1	7.4	7.3
MIN	6.4	7.2	7.1	4.3	6.5	7.4	7.3	7.4	7.1	7.5	1.6	6.7	5.8	6.1	7.2	4.8	7.1	6.9	7.6	5.3	6	6	7	5.6	7.8	7.3	8.5	6.4	7.5	6.1	7.9	6.5	5.2	7.6	8.4	6.2	7.4	7.3

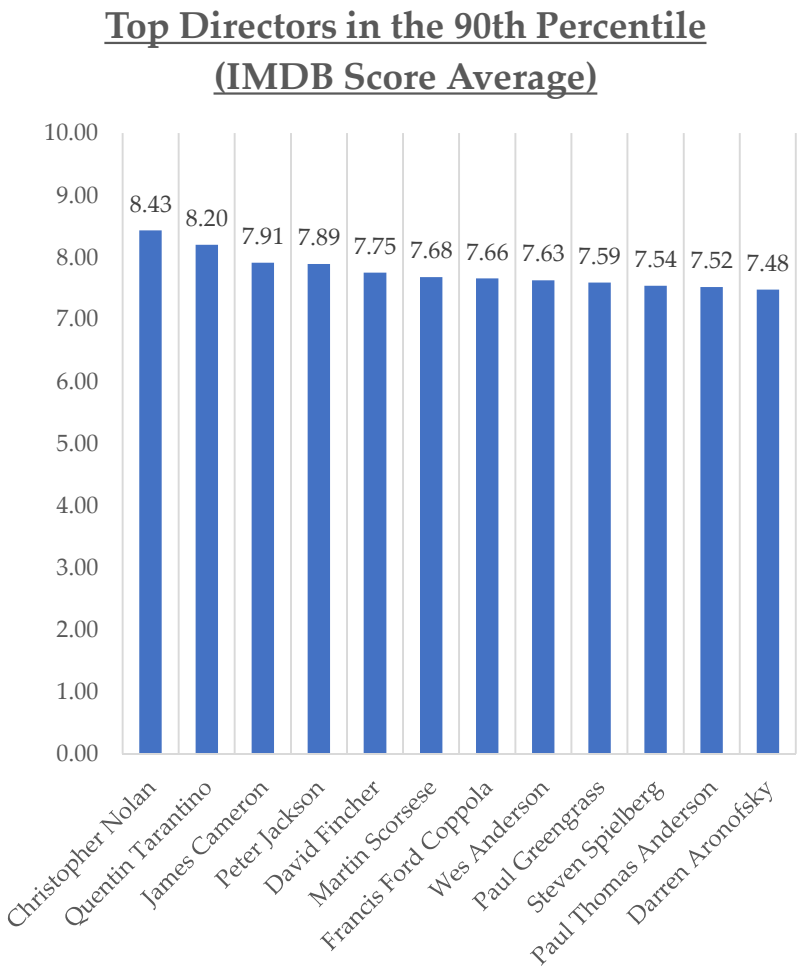
Q.D) Director Analysis: Influence of directors on movie ratings.

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

	Top Directors	All Directors
90th Percentile	7.48	7.5

Movie Count and IMDB Score Descriptive Analysis Directorwise
(Directors who have directed more than 5 movies)

Unique Directors	Movie Director Count	IMDB Score Average	IMDB Score Median	IMDB Score Mode	IMDB Score Variance	IMDB Score Standard Deviation	IMDB Score MAX	IMDB Score MIN	IMDB Percent Rank
Christopher Nolan	8	8.43	8.50	8.50	0.25	0.50	9.00	7.20	100.00%
Quentin Tarantino	8	8.20	8.20	#N/A	0.16	0.40	8.90	7.50	99.10%
James Cameron	7	7.91	7.90	#N/A	0.18	0.43	8.50	7.20	98.21%
Peter Jackson	9	7.89	7.90	7.90	0.53	0.73	8.90	6.70	97.32%
David Fincher	10	7.75	7.80	7.80	0.47	0.68	8.80	6.40	96.42%
Martin Scorsese	16	7.68	7.50	7.50	0.32	0.56	8.70	6.80	95.53%
Francis Ford Coppola	9	7.66	7.50	#N/A	0.96	0.98	9.20	6.30	94.64%
Wes Anderson	7	7.63	7.70	7.80	0.10	0.31	8.10	7.10	93.75%
Paul Greengrass	7	7.59	7.70	#N/A	0.16	0.40	8.10	6.90	92.85%
Steven Spielberg	25	7.54	7.60	7.60	0.45	0.67	8.90	6.20	91.96%
Paul Thomas Anderson	6	7.52	7.60	#N/A	0.27	0.52	8.10	6.70	91.07%
Darren Aronofsky	6	7.48	7.70	#N/A	0.69	0.83	8.40	5.80	90.17%
Sam Mendes	7	7.46	7.30	7.10	0.25	0.50	8.40	6.80	89.28%
Danny Boyle	8	7.44	7.45	7.60	0.24	0.49	8.20	6.60	88.39%
Richard Linklater	11	7.33	7.10	7.10	0.40	0.64	8.10	6.00	86.60%
Terry Gilliam	7	7.33	7.60	#N/A	0.58	0.76	8.30	5.90	86.60%
Robert Zemeckis	13	7.31	7.40	7.40	0.55	0.74	8.80	6.30	85.71%
Bryan Singer	8	7.29	7.35	#N/A	0.59	0.77	8.60	6.10	84.82%
Ang Lee	8	7.25	7.60	7.70	0.54	0.74	8.00	5.70	83.92%
Edward Zwick	7	7.24	7.50	#N/A	0.41	0.64	8.00	6.30	83.03%
Marc Forster	7	7.23	7.10	7.60	0.16	0.40	7.80	6.70	82.14%
Clint Eastwood	19	7.21	7.30	7.30	0.48	0.70	8.30	5.90	81.25%
James Wan	7	7.20	7.20	6.80	0.20	0.44	7.80	6.60	80.35%
Kenneth Branagh	6	7.18	7.20	7.00	0.29	0.54	7.80	6.20	79.46%
David O. Russell	7	7.17	7.10	#N/A	0.23	0.48	7.90	6.60	78.57%
Zack Snyder	7	7.14	7.20	7.70	0.27	0.52	7.70	6.10	77.67%
Doug Liman	7	7.13	7.30	7.90	0.41	0.64	7.90	6.10	75.00%
Gus Van Sant	7	7.13	7.10	#N/A	0.41	0.64	8.30	6.20	75.00%
Ridley Scott	16	7.13	7.05	8.50	0.85	0.92	8.50	5.30	75.00%
Ethan Coen	7	7.11	7.00	7.00	0.40	0.63	8.10	6.20	73.21%
Guy Ritchie	7	7.11	7.50	7.30	2.20	1.48	8.30	3.60	73.21%
Stephen Frears	8	7.09	7.35	7.60	0.44	0.66	7.70	5.80	72.32%
James Mangold	8	7.08	7.10	7.30	0.32	0.56	7.90	6.30	70.53%
Michael Mann	6	7.08	7.30	#N/A	0.73	0.85	7.90	5.40	70.53%
Tim Burton	14	7.05	7.00	7.00	0.47	0.68	8.00	5.70	69.64%
Jon Favreau	6	7.02	6.95	#N/A	0.44	0.66	7.90	6.10	67.85%

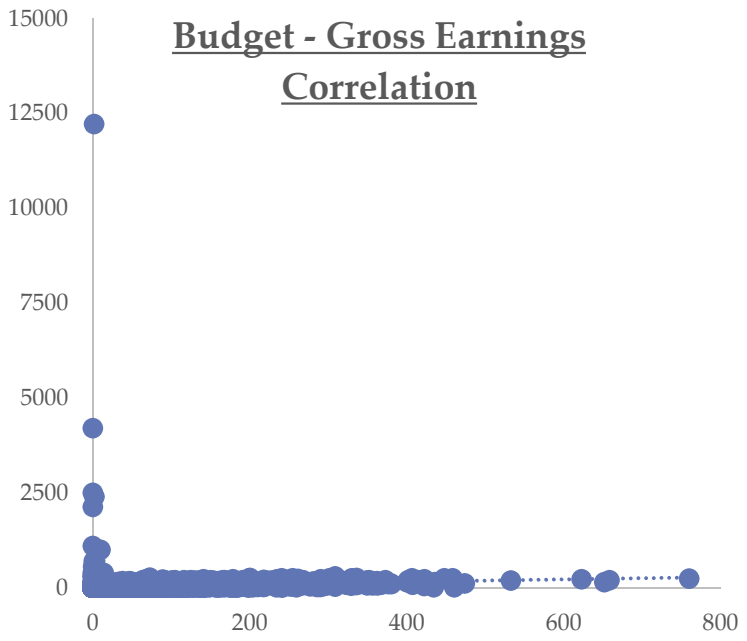


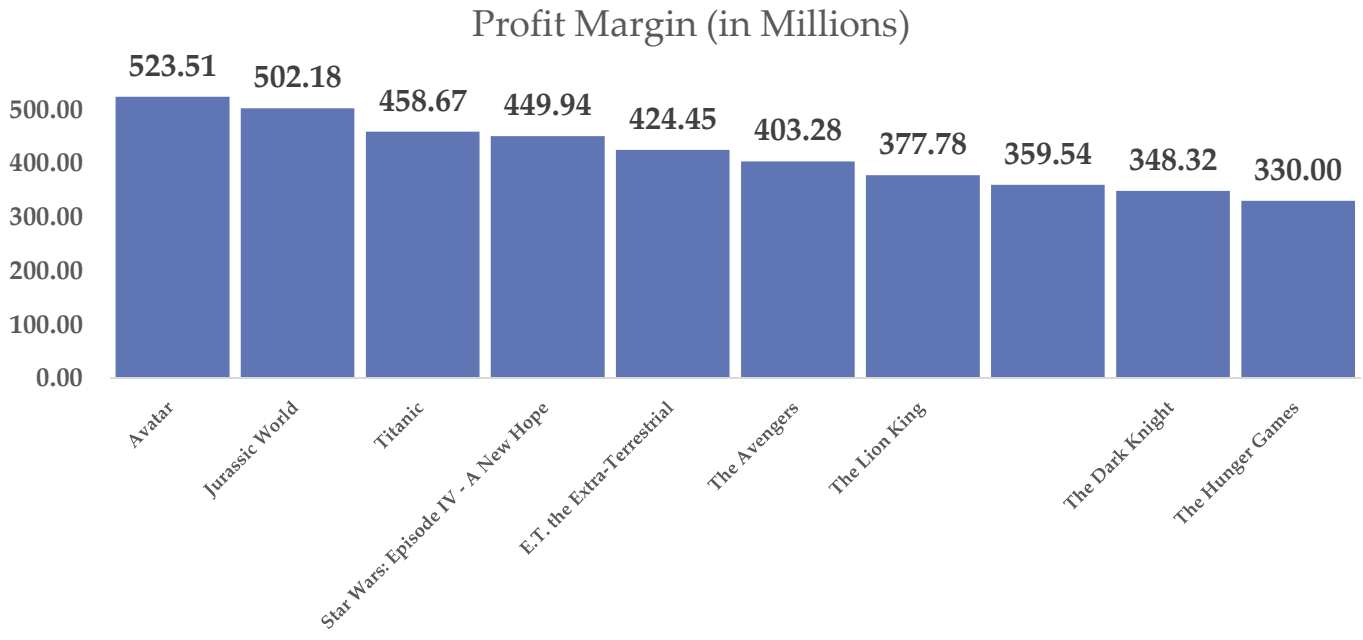
Q.E) Budget Analysis: Explore the relationship between movie budgets and their financial success.
Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

Correlation between budget and gross earnings	0.096568921
Movie with the highest profit margin	Avatar

Top 10 movies by Profit Margin

Movie Title	Profit Margin (in Millions)
Avatar	523.51
Jurassic World	502.18
Titanic	458.67
Star Wars: Episode IV - A New Hope	449.94
E.T. the Extra-Terrestrial	424.45
The Avengers	403.28
The Lion King	377.78
Star Wars: Episode I - The Phantom Menace	359.54
The Dark Knight	348.32
The Hunger Games	330.00





Insights

- Drama is the most common genre, followed by Comedy and Thriller.
- Biography has the best IMDB Average (Film-Noir is not taken into account as it has only 1 movie), followed by History and War.
- The mean, median and mode of all genres are mostly similar.
- Most movies made are in the 41-110 mins duration category.
- The average duration for a movie is 110 mins.
- The relationship between movie duration and imdb scores has a slight upward trend.
- The most popular language is English.
- There are 1751 unique directors, but only 113 have directed over 5 movies and are used to rate the best.
- Steven Spielberg has directed the most number of movies at 25.
- Christopher Nolan is the most highly rated director at 8.43.
- 12 directors are in the 90th percentile based on IMDB Score Average which above 7.48.
- Correlation between budget and gross earnings is 0.096 which means that there is little or no relation between Budget and Gross Earnings.
- The movie with the highest profit margin is 'Avatar'.

Results

The 'IMDb Movie Analysis' is important for movie producers, directors, and investors who want to understand what makes a movie successful. This allows them to make informed decisions in their future projects.

The project has helped me understand the relationship between various fields, like 'Genre & IMDb Score' and 'Budget & Earnings'. It has also helped me improve my knowledge of Data Analysis using Excel and Statistics.

This project allowed me to ask the question 'Why?', making me dive deeper to finally find the root cause.

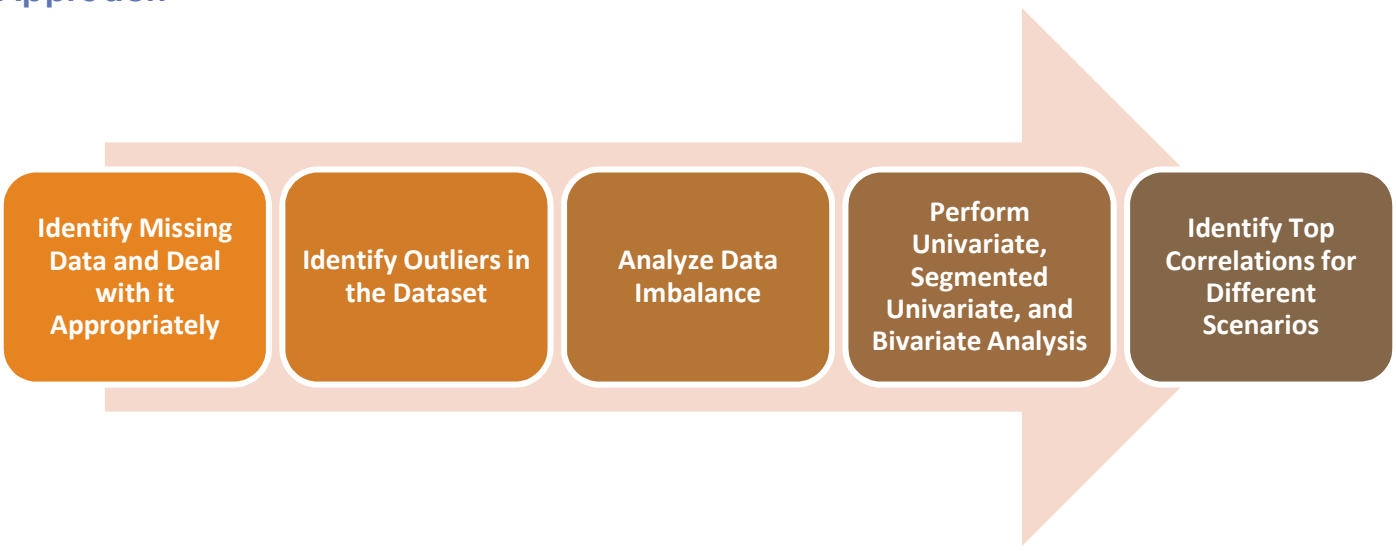
Link to IMDB Movie Analysis Excel File

Project 6 - Bank Loan Case Study

Project Description

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments.
This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants.
The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.
As a Data Analyst, my task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

Approach



Tech Stack Used

Microsoft Excel 2010 Version 14.0.7628.5000

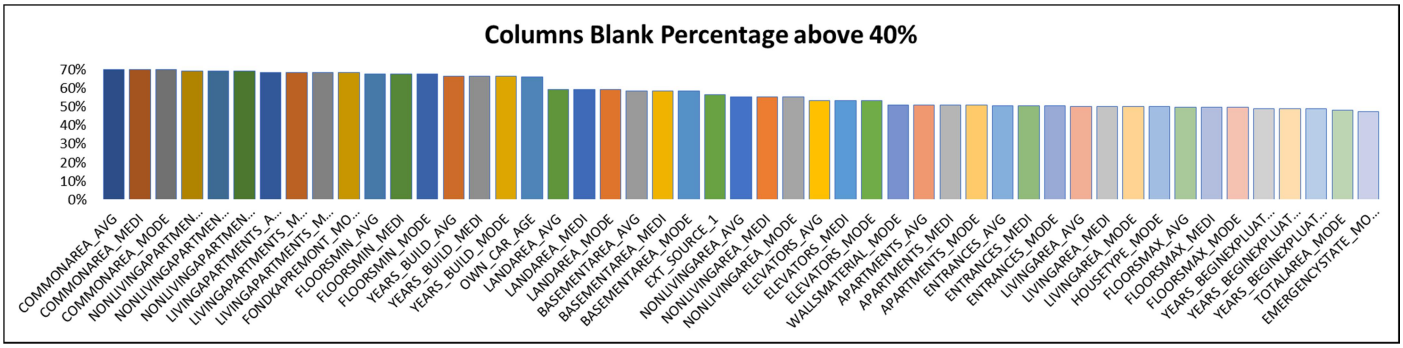
Insights

Task 1: Identify Missing Data and Deal with it Appropriately
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.

- Understand the Data
- Find and Handle Missing Values (Rows and Columns)
    - Remove Irrelevant Columns
    - Remove Duplicates
    - Impute Missing Values with Mean/Median/Mode

Understand the Data

Application Data (before Cleaning)	
Total Rows	49999
Total Columns	122
Columns with Null %age >40%	49



Find and Handle Missing Values (Rows and Columns)

NO DUPLICATES FOUND – SK_ID_CURR

Columns Dropped as not required for Analysis	
EXT_SOURCE2	EXT_SOURCE3

Columns Dropped due to Blank Percentage over 40%	
COMMONAREA_AVG	NONLIVINGAREA_MEDI
COMMONAREA_MEDI	NONLIVINGAREA_MODE
COMMONAREA_MODE	ELEVATORS_AVG
NONLIVINGAPARTMENTS_AVG	ELEVATORS_MEDI
NONLIVINGAPARTMENTS_MEDI	ELEVATORS_MODE
NONLIVINGAPARTMENTS_MODE	WALLSMATERIAL_MODE
LIVINGAPARTMENTS_AVG	APARTMENTS_AVG
LIVINGAPARTMENTS_MEDI	APARTMENTS_MEDI
LIVINGAPARTMENTS_MODE	APARTMENTS_MODE
FONDKAPREMONT_MODE	ENTRANCES_AVG
FLOORSMIN_AVG	ENTRANCES_MEDI
FLOORSMIN_MEDI	ENTRANCES_MODE
FLOORSMIN_MODE	LIVINGAREA_AVG
YEARS_BUILD_AVG	LIVINGAREA_MEDI
YEARS_BUILD_MEDI	LIVINGAREA_MODE
YEARS_BUILD_MODE	HOUSETYPE_MODE
OWN_CAR_AGE	FLOORSMAX_AVG
LANDAREA_AVG	FLOORSMAX_MEDI
LANDAREA_MEDI	FLOORSMAX_MODE
LANDAREA_MODE	YEARS_BEGINEXPLUATATION_AVG
BASEMENTAREA_AVG	YEARS_BEGINEXPLUATATION_MEDI
BASEMENTAREA_MEDI	YEARS_BEGINEXPLUATATION_MODE
BASEMENTAREA_MODE	TOTALAREA_MODE
EXT_SOURCE_1	EMERGENCYSTATE_MODE
NONLIVINGAREA_AVG	

Columns Having Blank Percentage Data Imputed based on Mean/Median/Mode	
OCCUPATION_TYPE	Mode Used. New Type assigned as Blank had the most records - 'Unknown'
AMT_REQ_CREDIT_BUREAU_DAY	Median Used
AMT_REQ_CREDIT_BUREAU_HOUR	Median Used
AMT_REQ_CREDIT_BUREAU_MON	Median Used
AMT_REQ_CREDIT_BUREAU_QRT	Median Used
AMT_REQ_CREDIT_BUREAU_WEEK	Median Used
AMT_REQ_CREDIT_BUREAU_YEAR	Median Used
NAME_TYPE_SUITE	Mode Used. Changed the Blanks to 'Unaccompanied'
DEF_30_CNT_SOCIAL_CIRCLE	Median Used
DEF_60_CNT_SOCIAL_CIRCLE	Median Used
OBS_30_CNT_SOCIAL_CIRCLE	Median Used
OBS_60_CNT_SOCIAL_CIRCLE	Median Used
AMT_GOODS_PRICE	Median Used

Columns with Minimal Blank Percentage Rows with Blank Values dropped	
AMT_ANNUITY	01 Row Dropped
CNT_FAM_MEMBERS	01 Row Dropped
DAYS_LAST_PHONE_CHANGE	01 Row Dropped

Task 2: 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.

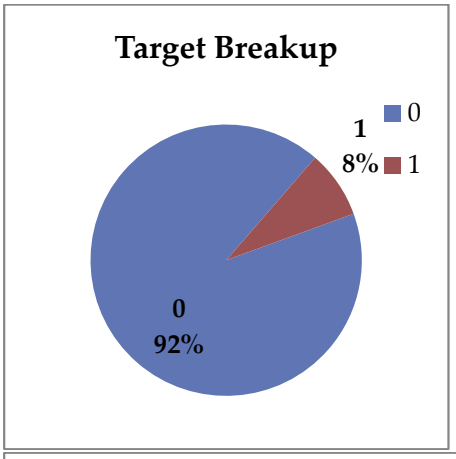
Outliers Analysis

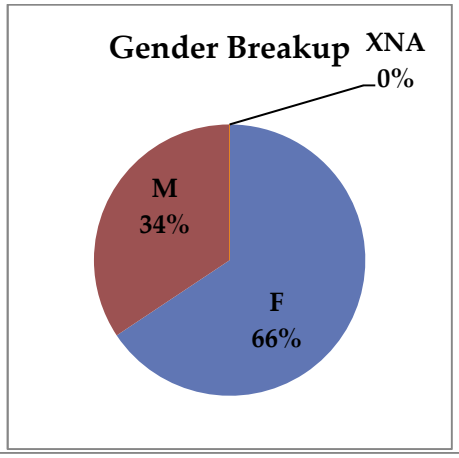
- There are few high outliers in AMT_ANNUITY which does not give a very healthy analysis and can skew the mean.
- There are a few outliers in AMT_GOODS_PRICE & AMT_CREDIT where the amount is more than normal.
- There is an outliers in AMT_INCOME_TOTAL which is extraordinarily high, which is unusual.
- The YEARS_WORKED data has quite a lot of data that has the value - 1000 years. This is impossible. On further investigating the data, it was found that the value was assigned for 'Pensioners'

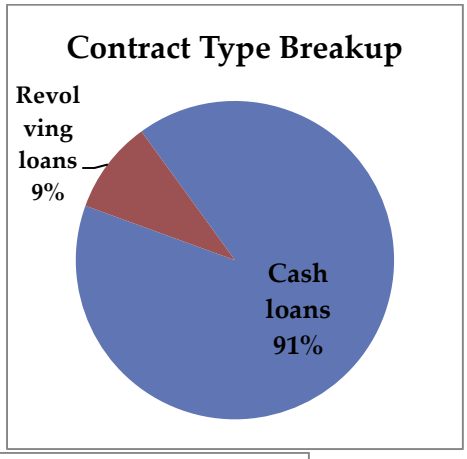
- There are a few outliers in DAYS_LAST_PHONE_CHANGE which suggests that many people are using the same phone for a long time, even over 8 years
- Most of the people in the data are over 20 years and below 75 years, which shows that the data is not skewed from the point of age.
- Lastly, there are few outliers in CNT_CHILDREN, where a few people have more than 4 children, which is impractical and can put financial pressure.

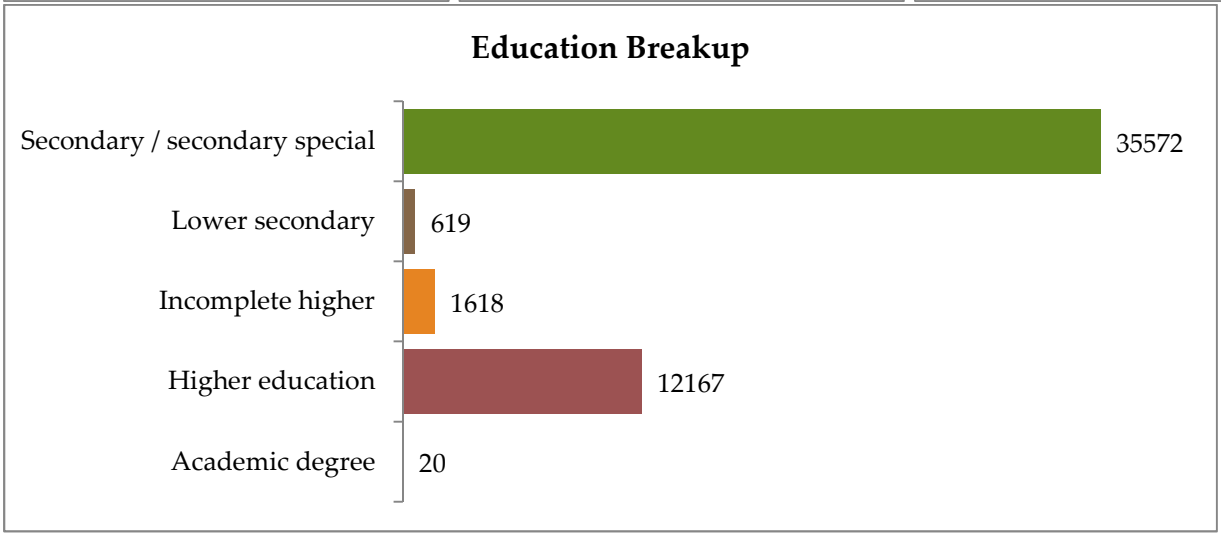
Task 3: Analyze Data Imbalance

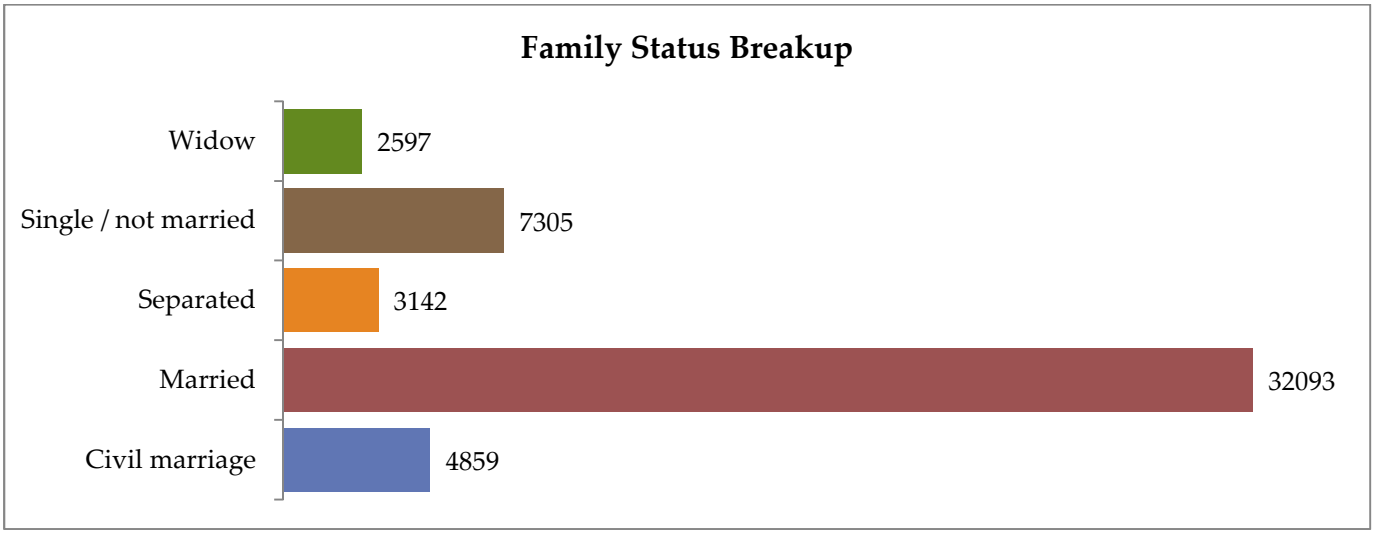
Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

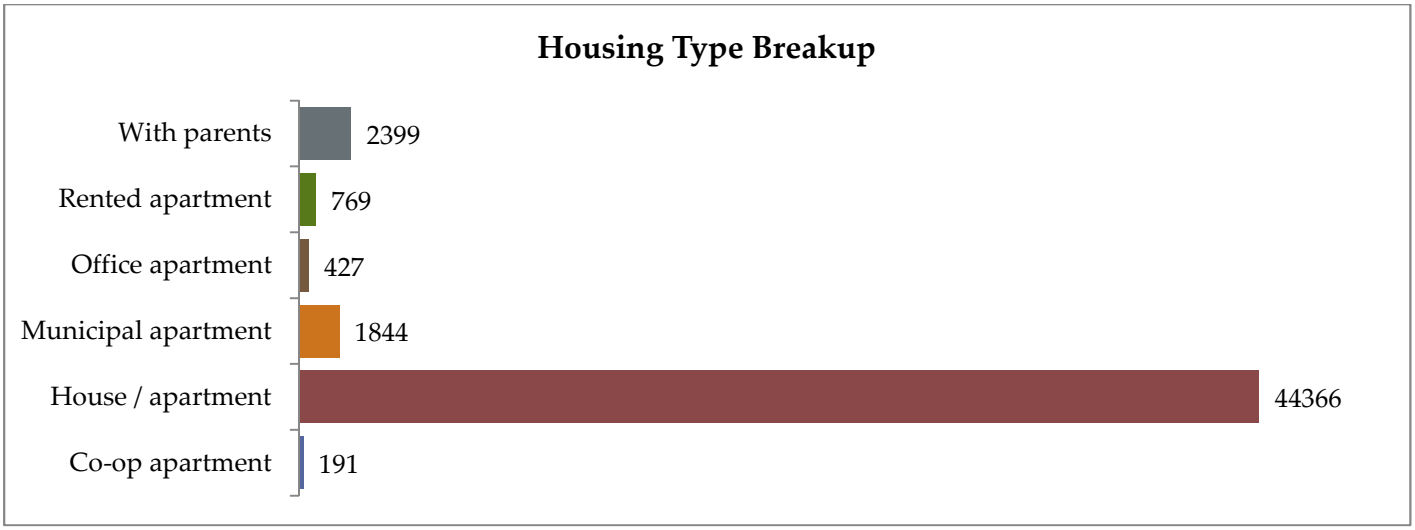












All the charts shown above have Data Imbalance

- The Pie Charts have RePayers, Females and Cash Loans are majority in the data.
- The Bar Charts have people who have studied till Secondary, Married and People staying in Apartments taking up most of the data

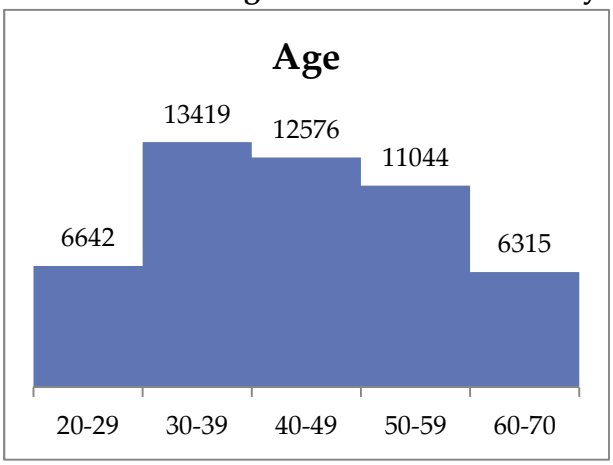
Task 4: Perform Univariate, Segmented Univariate, and Bivariate Analysis

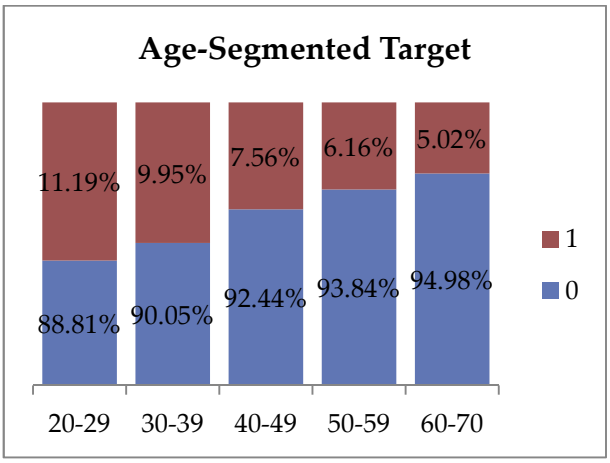
To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

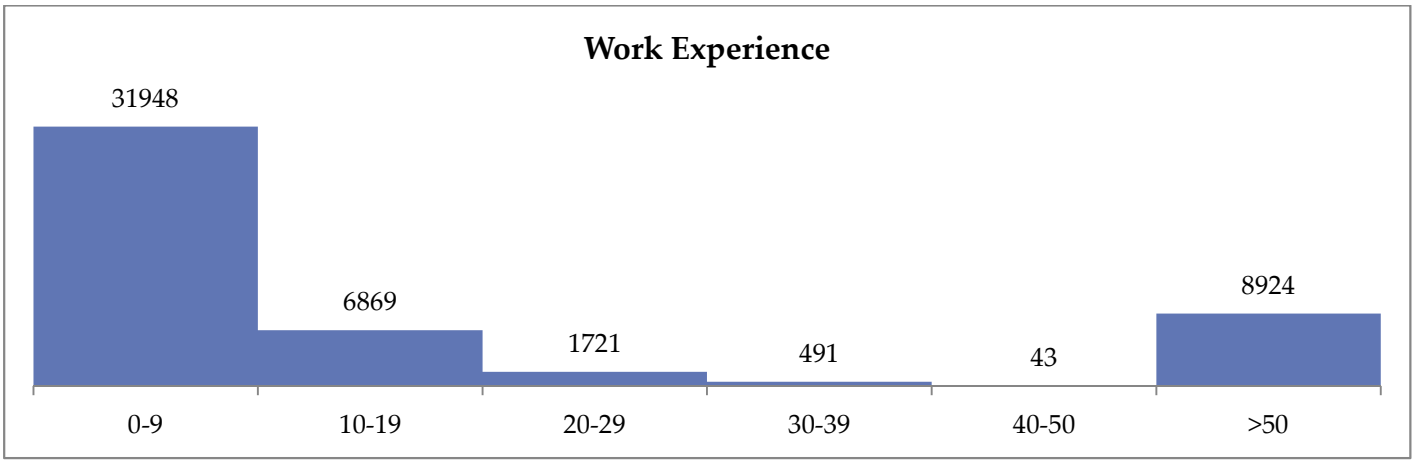
To attain an analysis of the data, we will be performing the below:

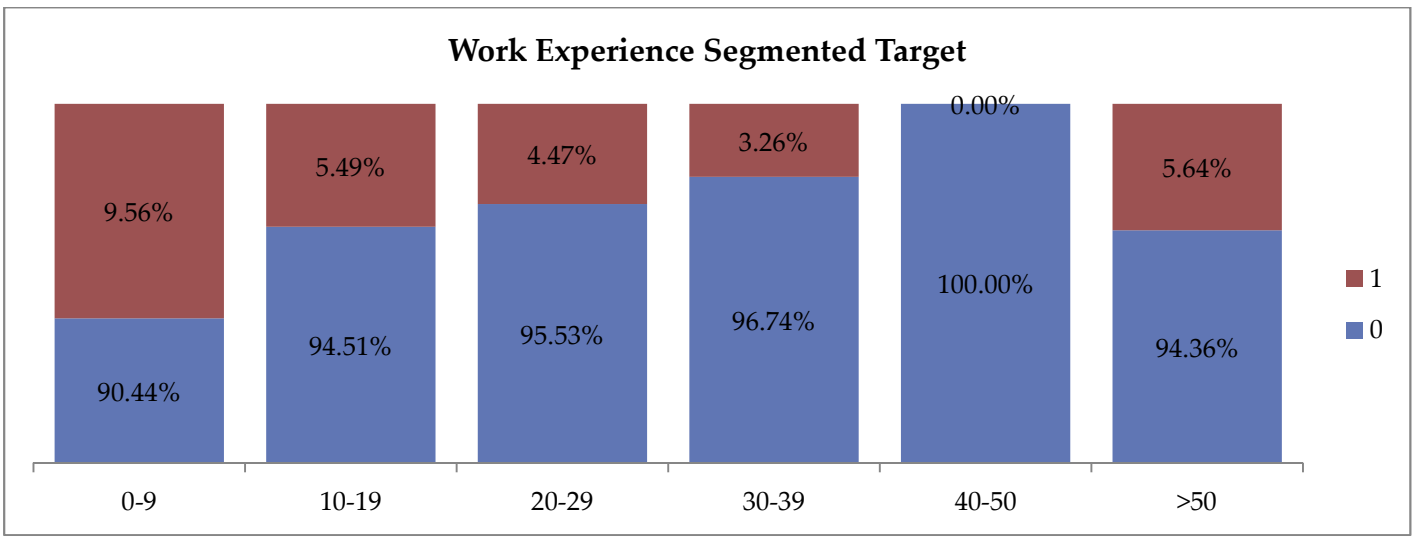
- Univariate Analysis to understand the distribution of individual variables
- Segmented Univariate Analysis to compare variable distributions for different scenarios
- Bivariate Analysis to explore relationships between variables and the target variable using Excel functions and features.

Univariate and Segmented Univariate Analysis

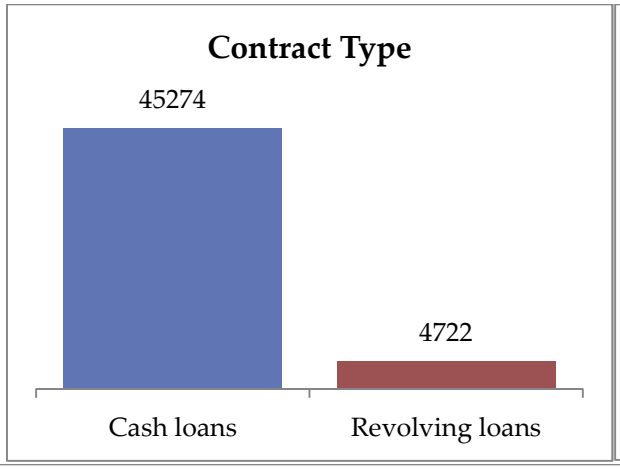


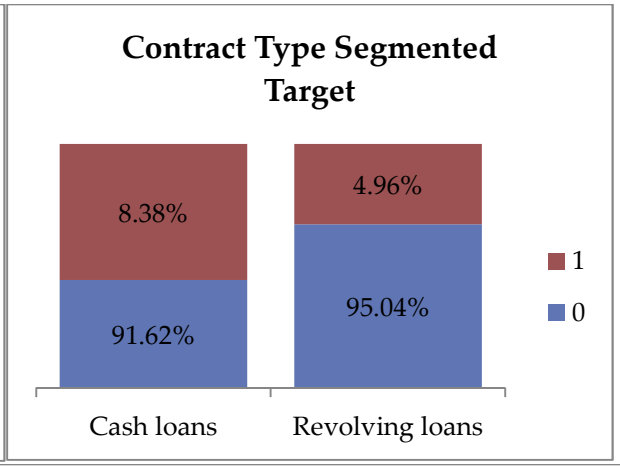


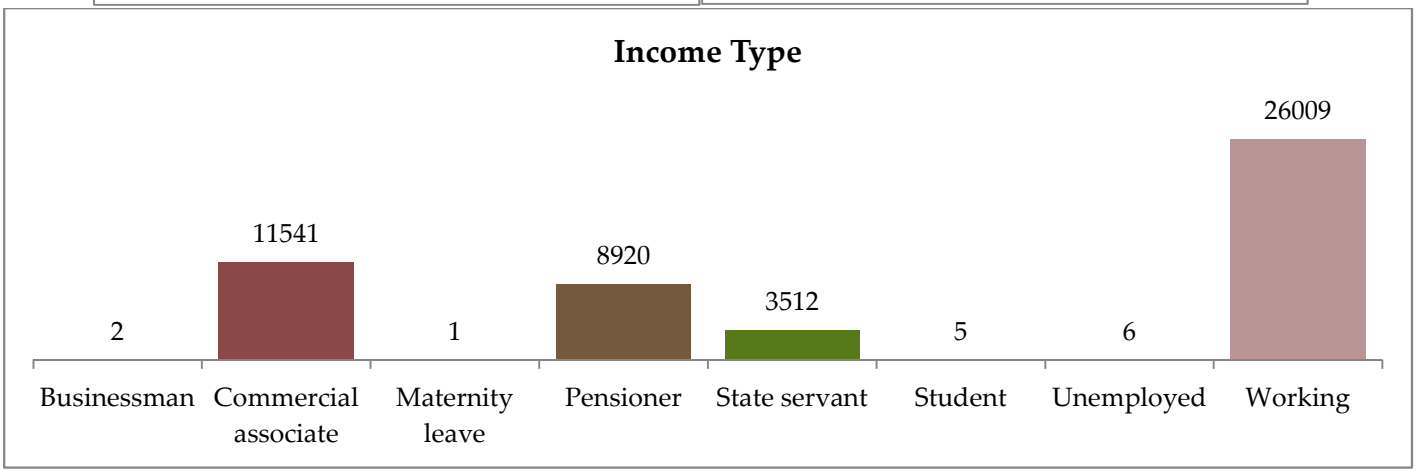


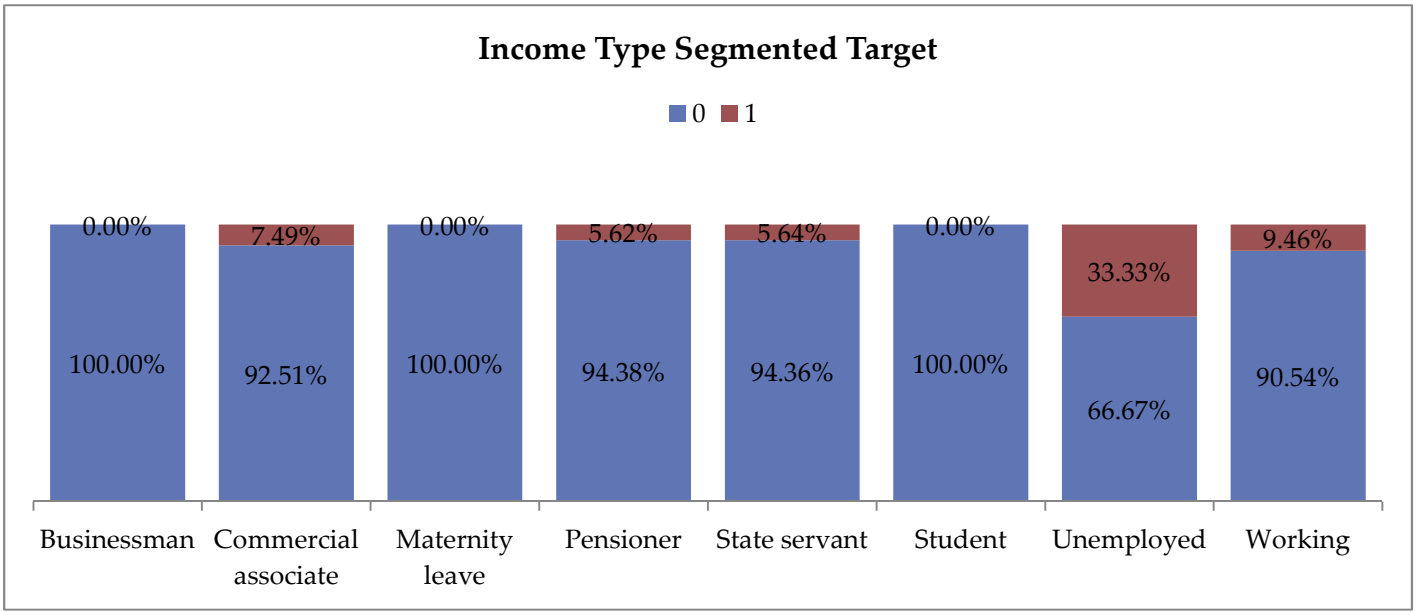


- Middle Age group have been given the most loans
- The percentage of Defaulters decrease as age increases
- Most people taking loans have less than 10 years work experience
- Defaulters decrease as work experience increase

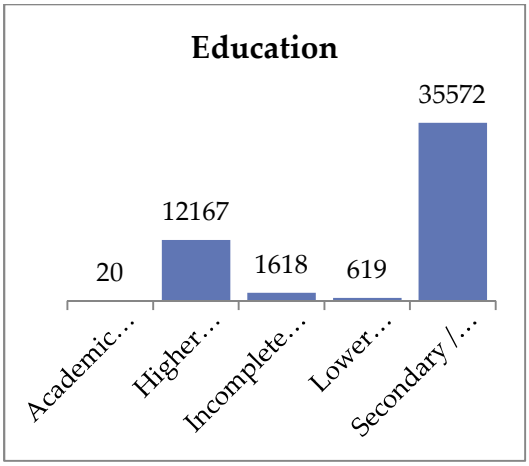


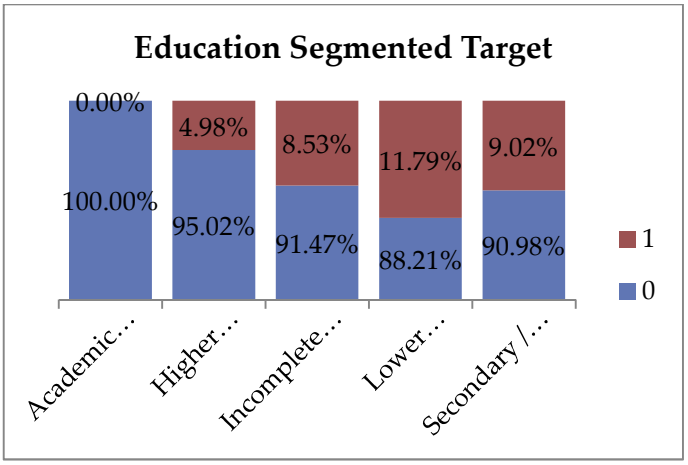


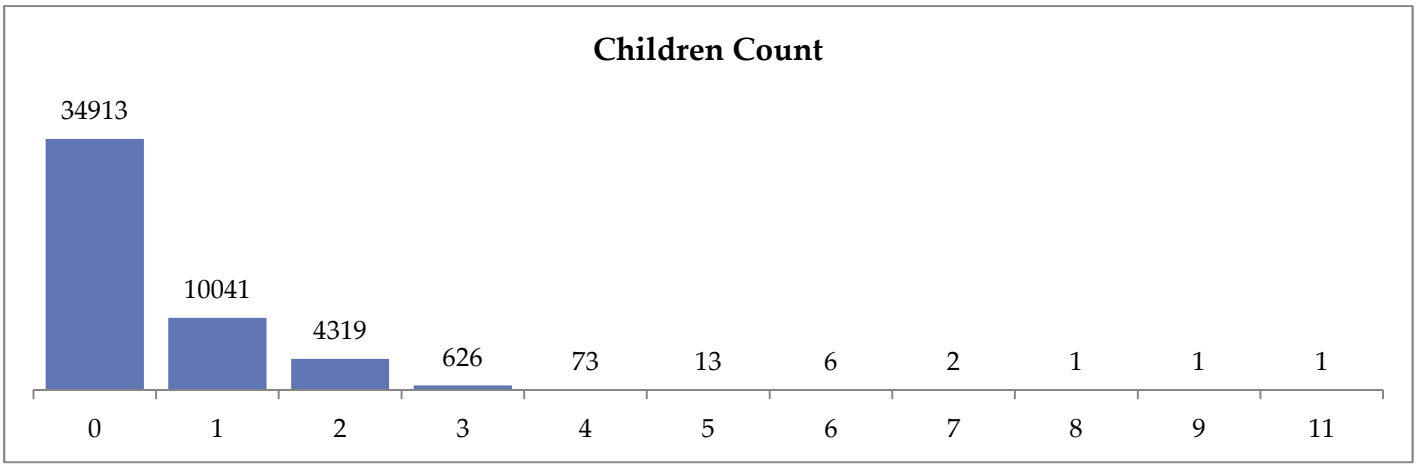


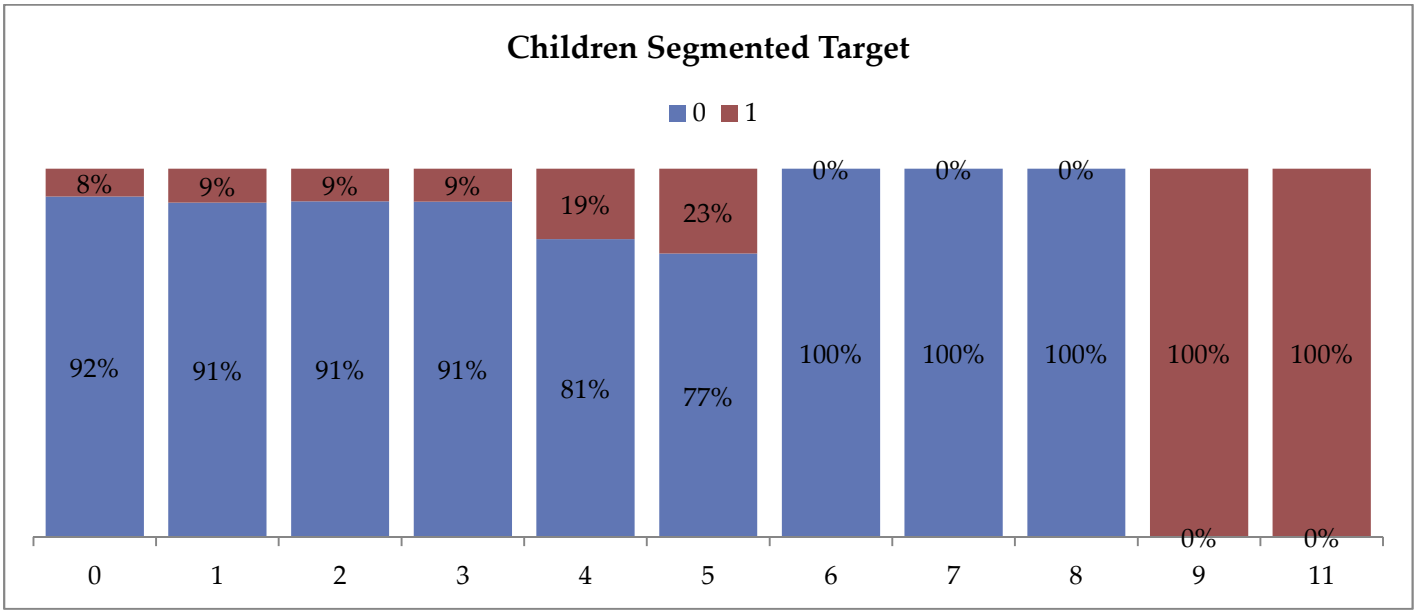


- Most people take Cash Loans
- Cash Loans has the highest Defaulter Percentage
- Majority of the Applied Loans are by Working Professionals
- Highest Defaulters are Unemployed
- Students and Businessmen are less likely to default

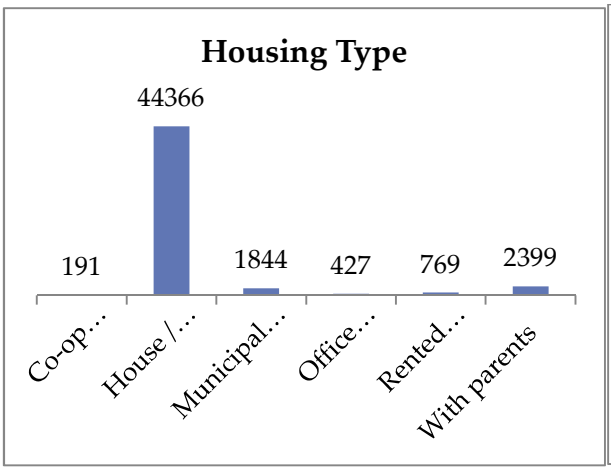


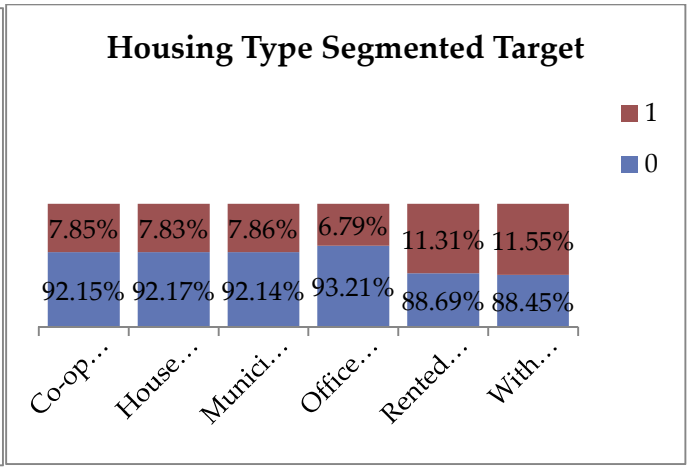


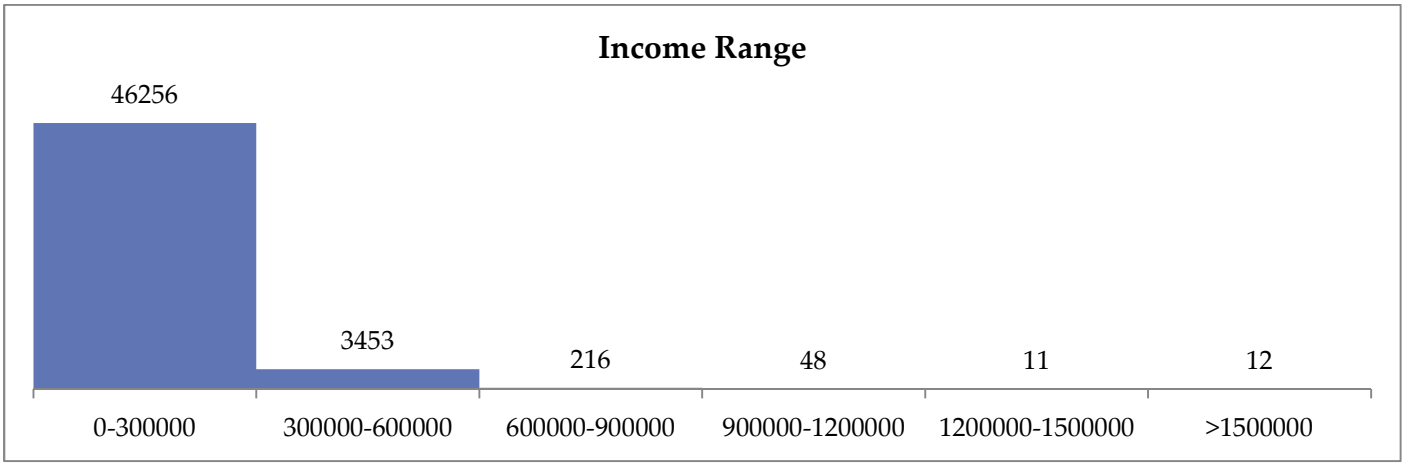


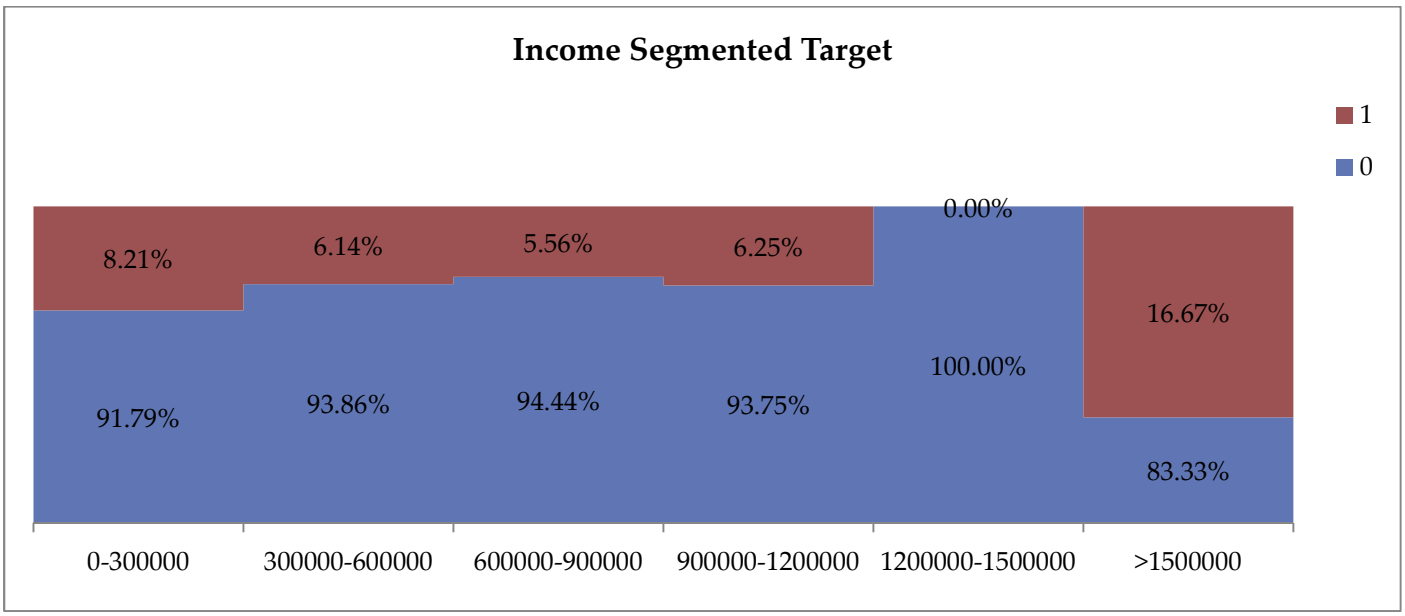


- Most people applying for loans have only had Secondary Education
- Lower Education has the highest defaulters
- Singles or People with no children are the highest loan appliers
- Outliers with over 8 kids are defaulters



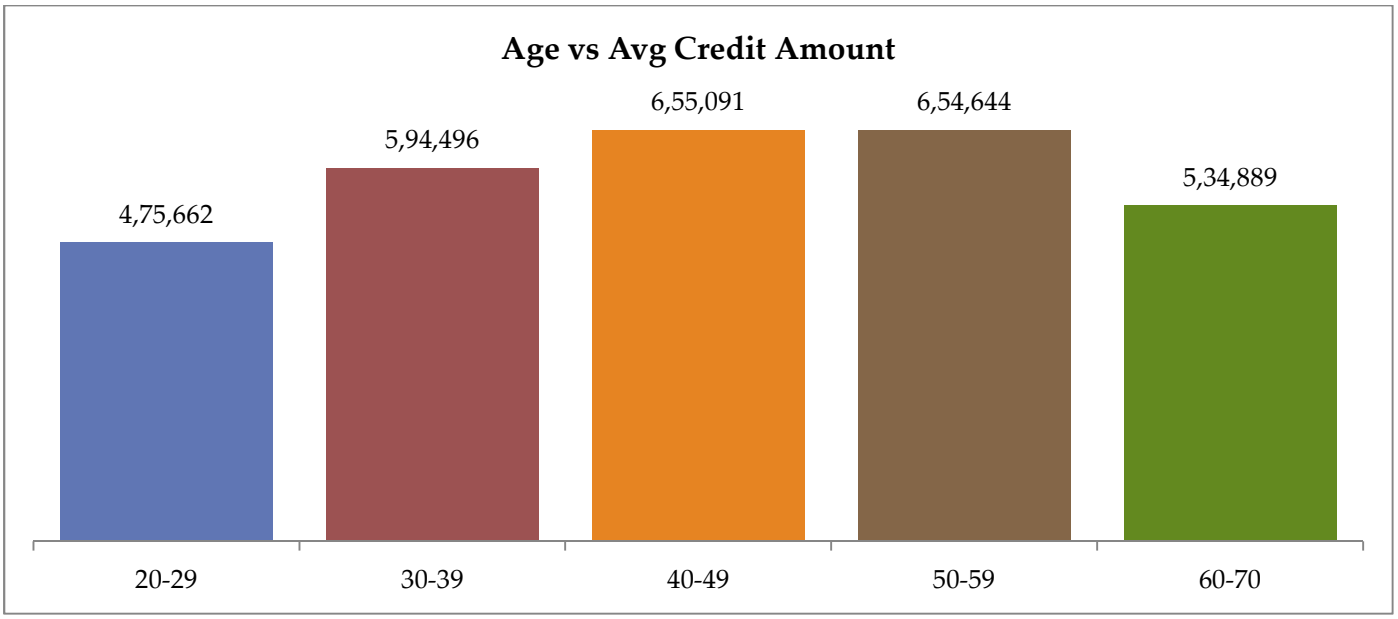




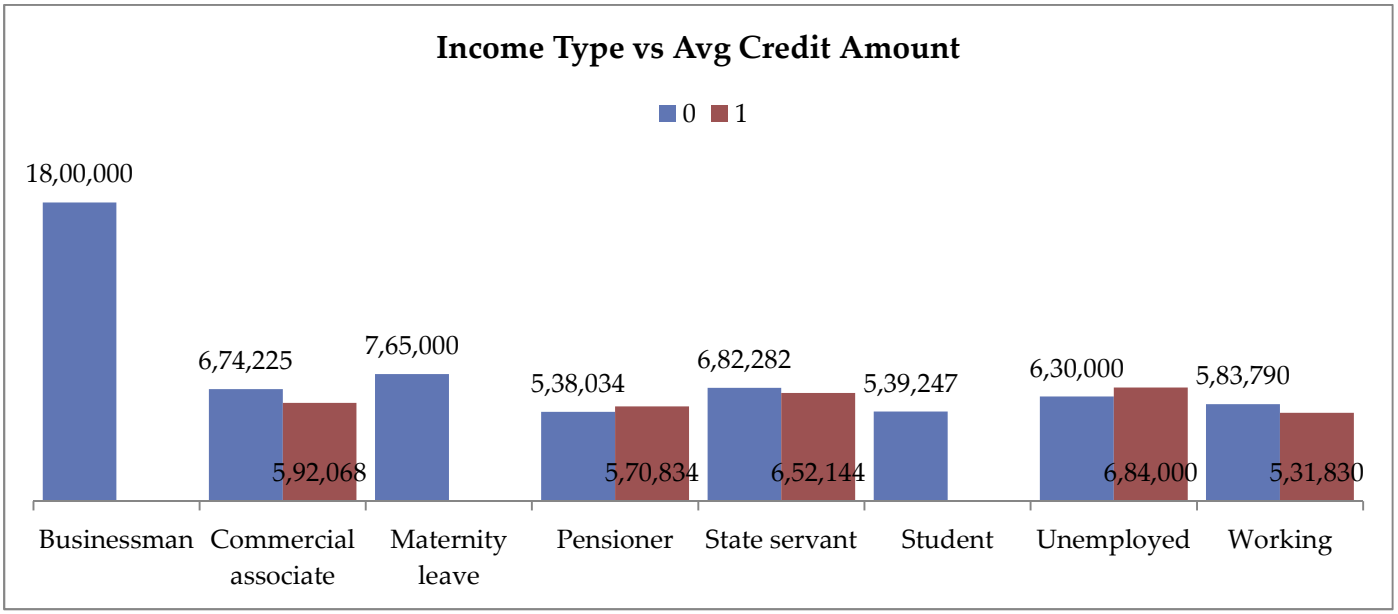


- Most people applying for loans stay in Apartments
- Most Defaulter Percentage stay with their parents or on rent
- Majority of the people applying for loans are in the low income group
- The High Income Groups are more likely to pay their loans, but the Very High Income Groups have the highest defaulter percentage

Bivariate Analysis



More credit is given to people as they grow older, but once they are reaching their retirement, the credit decreases. This is due to the fact that with age you default less



Businessmen get the highest credit amount whereas Students get the lowest

Task 5: Identify Top Correlations for Different Scenarios

Correlation Matrix for Defaulters

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	AGE	DAYS_EMPLOYED	YEARS_WORKED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	DAYS_LAST_PHONE_CHANGE	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1																							
AMT_INCOME_TOTAL	0.01011018	1																						
AMT_CREDIT	0.00760191	0.01527144	1																					
AMT_ANNUITY	0.02917298	0.01800459	0.7496652	1																				
AMT_GOODS_PRICE	-0.00107967	0.0132695	0.98226796	0.74950403	1																			
REGION_POPULATION_RELATIVE	-0.02035915	-0.0061803	0.06777562	0.073124	0.07663549	1																		
DAYS_BIRTH	0.2496732	0.00903366	-0.14250603	-0.00875171	-0.1410059	-0.01646873	1																	
AGE	-0.24961576	-0.00844428	0.14238416	0.00886209	0.14086284	0.01653014	-0.99968736	1																
DAYS_EMPLOYED	-0.18932418	-0.01155596	0.01603957	-0.07955601	0.02023535	0.00774291	-0.58147904	0.58106649	1															
YEARS_WORKED	-0.18978832	-0.01173541	0.01877681	-0.07811983	0.02318228	0.00770574	-0.58825779	0.58785843	0.99960013	1														
DAYS_REGISTRATION	0.15211312	-0.00956115	-0.0428444	0.02158165	-0.04332022	-0.04613029	0.28843784	-0.28793306	-0.18871844	-0.19244489	1													
DAYS_ID_PUBLISH	-0.04236072	-0.00912201	-0.0437719	-0.02132109	-0.04972323	-0.00511856	0.24789657	-0.24792541	-0.23006367	-0.23266468	0.09029149	1												
CNT_FAM_MEMBERS	0.89252187	0.01312168	0.06124869	0.07583846	0.05513581	-0.01725715	0.1991414	-0.19900248	-0.18356011	-0.18337761	0.15178655	-0.04403782	1											
OBS_30_CNT_SOCIAL_CIRCLE	0.01793193	-0.01128092	0.03346617	0.01381902	0.03272397	-0.00887544	-0.01115023	0.01122317	0.00352185	0.00470808	-0.0057933	-0.02731374	0.03999054	1										
DEF_30_CNT_SOCIAL_CIRCLE	-0.01361871	-0.00797944	-0.02494668	-0.03454537	-0.01909661	0.02780592	-0.02083879	0.02099375	0.02985635	0.0297283	0.00099818	-0.02842652	-0.00645388	0.36507385	1									
OBS_60_CNT_SOCIAL_CIRCLE	0.01514587	-0.01121117	0.03443931	0.01409863	0.03387918	-0.007065	-0.01257029	0.01262327	0.00420868	0.00541281	-0.0059266	-0.02621248	0.0375377	0.99806585	0.36805994	1								
DEF_60_CNT_SOCIAL_CIRCLE	-0.0185057	-0.00672696	-0.02900724	-0.04047103	-0.02059292	0.02714232	-0.02575665	0.0258594	0.0238941	0.02378984	-0.00641263	-0.02789635	-0.00887718	0.29795102	0.89051161	0.30142085	1							
DAYS_LAST_PHONE_CHANGE	0.01133933	0.01245711	-0.12453934	-0.10047094	-0.12883245	-0.06710568	0.12460949	-0.12387487	-0.01573254	-0.01936405	0.07860465	0.13808778	-0.00573115	-0.0219161	0.00415783	-0.0230033	0.015271	1						
AMT_REQ_CREDIT_BUREAU_HOUR	-0.0002876	-0.00110418	0.01780636	0.03739749	0.01526195	0.00935622	0.02489871	-0.0249157	-0.00304646	-0.00355628	0.00638373	0.01407583	0.00486301	-0.01408506	0.00272831	-0.01358072	-0.01317789	0.0124725						
AMT_REQ_CREDIT_BUREAU_DAY	-0.03060525	-0.00144685	-0.0085184	-0.01868834	-0.00631921	-0.00383354	-0.02267042	0.02261487	0.04947762	0.04939249	-0.00147508	-0.00643299	-0.03314229	-0.01702919	0.01223625	-0.01744512	-0.01032405	-0.00679158	0.3511998	1				
AMT_REQ_CREDIT_BUREAU_WEEK	-0.03060405	-0.00221861	0.00012537	0.03472145	0.00011449	0.01206424	-0.00966098	0.00993477	0.0203878	0.020222	-0.01817817	0.01953762	-0.0280826	0.0058366	-0.01162117	0.00555602	-0.00393597	-0.00103167	0.01931666	0.06198856	1			
AMT_REQ_CREDIT_BUREAU_MON	0.008161	-0.00086402	0.0834082	0.07129522	0.0789087	0.0753956	-0.0072774	0.00748929	-0.03306561	-0.03208909	-0.001526	-0.03791731	0.01616533	0.01607779	0.00808821	0.01698449	0.01303479	-0.05674853	-0.00102354	-0.01629214	-0.00095331	1		
AMT_REQ_CREDIT_BUREAU_QRT	-0.0115206	-0.00374923	-0.01936131	-0.00163066	-0.02036764	0.01531017	-0.00878324	0.0087175	0.01787588	0.01760253	-0.00629042	-0.03267147	0.00246556	0.03483581	0.0201013	0.03640041	0.02534777	0.00042674	0.03109885	0.02553544	0.01169736	0.0199464	1	
AMT_REQ_CREDIT_BUREAU_YEAR	-0.03080113	-0.00510098	-0.01645997	0.00156927	-0.02347544	0.02402393	-0.09012732	0.08962933	0.01769246	0.01811595	-0.02509419	-0.08164306	-0.00506845	0.05051753	0.02101665	0.05070851	0.02062616	-0.10067865	0.00615969	0.02760701	0.03091697	0.0387895	0.10363174	1

Correlation Matrix for Non-Defaulters

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	AGE	DAYS_EMPLOYED	YEARS_WORKED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	DAYS_LAST_PHONE_CHANGE	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1																							
AMT_INCOME_TOTAL	0.0363546	1																						
AMT_CREDIT	0.005693	0.3779851	1																					
AMT_ANNUITY	0.026387	0.4511483	0.7707771	1																				
AMT_GOODS_PRICE	0.0015021	0.3846215	0.9870017	0.7758435	1																			
REGION_POPULATION_RELATIVE	-0.024923	0.1819764	0.0955331	0.1172846	0.0989572	1																		
DAYS_BIRTH	0.3359474	0.0737412	-0.051051	0.009911	-0.048736	-0.030412	1																	
AGE	-0.335761	-0.073614	0.0512097	-0.009707	0.0489161	0.0303618	-0.999707	1																
DAYS_EMPLOYED	-0.243613	-0.162693	-0.077379	-0.113004	-0.07512	-0.006618	-0.615291	0.6150742	1															
YEARS_WORKED	-0.245549	-0.161668	-0.074743	-0.111287	-0.07246	-0.006781	-0.623467	0.6232501	0.999533	1														
DAYS_REGISTRATION	0.1830823	0.0689103	0.0080533	0.0346046	0.0112642	-0.058497	0.335036	-0.334776	-0.204367	-0.208838	1													
DAYS_ID_PUBLISH	-0.032536	0.0323623	-0.008266	0.0094387	-0.00937	-0.002223	0.2700702	-0.270217	-0.272233	-0.274523	0.1035609	1												
CNT_FAM_MEMBERS	0.8792434	0.0416157	0.0648659	0.0778914	0.0628827	-0.022999	0.2844454	-0.284258	-0.233756	-0.234792	0.171489	-0.025058	1											
OBS_30_CNT_SOCIAL_CIRCLE	0.0161794	-0.033097	0.0008616	-0.010001	0.0004953	-0.019072	0.0123028	-0.012352	0.0056494	0.0055721	0.0109714	-0.011823	0.0242932	1										
DEF_30_CNT_SOCIAL_CIRCLE	-0.002833	-0.031999	-0.013516	-0.019745	-0.015219	0.0089004	0.0007102	-0.000764	0.0170239	0.0166534	0.003453	0.0023139	-0.002824	0.3061583	1									
OBS_60_CNT_SOCIAL_CIRCLE	0.0163343	-0.033069	0.0011701	-0.009685	0.0007178	-0.018015	0.0123099	-0.012352	0.0055107	0.0054417	0.0112891	-0.012125	0.0245776	0.9983575	0.3085654	1								
DEF_60_CNT_SOCIAL_CIRCLE	-0.003341	-0.032523	-0.018573	-0.023009	-0.019743	0.0032491	0.0022298	-0.002301	0.016508	0.0161195	0.0062859	0.0026435	-0.004596	0.2291725	0.850995	0.2312825	1							
DAYS_LAST_PHONE_CHANGE	-0.004803	-0.049511	-0.071191	-0.064449	-0.074233	-0.044133	0.0725007	-0.072429	0.0329686	0.0291783	0.0477788	0.0850658	-0.025007	-0.014342	0.0025038	-0.015119	0.0022878	1						
AMT_REQ_CREDIT_BUREAU_HOUR	0.0026143	0.0081277	3.49E-05	0.0101412	0.000809	-0.003133	0.0014926	-0.001468	-0.004296	-0.004401	-0.003688	0.0028246	0.0036848	0.0023638	-0.004399	0.0025849	-0.0032	-0.00128	1					
AMT_REQ_CREDIT_BUREAU_DAY	0.0011966	0.0094819	0.0134851	0.0091571	0.0136393	-0.00034	0.0019892	-0.001957	0.0016163	0.0015184	-0.003383	0.0035151	0.0006473	0.0009729	0.0036862	0.0008657	0.0027768	-0.000453	0.2307631	1				
AMT_REQ_CREDIT_BUREAU_WEEK	0.0043217	0.009497	0.0053717	0.0189105	0.0058066	0.0026421	-0.002388	0.0022825	-0.006486	-0.006243	0.0006618	-0.004665	0.0061138	-0.004288	-0.005038	-0.004878	-0.005732	-0.005992	0.012125	0.2491225	1			
AMT_REQ_CREDIT_BUREAU_MON	-0.011618	0.074876	0.0639709	0.0379869	0.0657025	0.070733	-0.002431	0.0023181	-0.032963	-0.032245	-0.010721	-0.013232	-0.00451	0.0081697	0.0076822	0.0081263	0.0039669	-0.047332	0.0095461	-0.000655	-0.010603	1		
AMT_REQ_CREDIT_BUREAU_QRT	-0.004724	0.0157967	0.0267999	0.010067	0.027519	-0.009716	-0.021539	0.0216167	0.0145838	0.0146865	0.0031317	-0.024582	-0.004241	0.0088515	0.0053543	0.0086808	0.0083102	-0.012883	0.003519	-0.007869	-0.014597	0.011892	1	
AMT_REQ_CREDIT_BUREAU_YEAR	-0.035752	0.0313498	-0.031578	-0.004172	-0.034428	0.0046453	-0.070236	0.0702236	0.0441713	0.0443584	-0.022957	-0.044693	-0.022929	0.0341607	0.014498	0.0345735	0.0151975	-0.11761	0.0040933	-0.000859	0.024733	0.0193038	0.1217536	1

Results

This project involved an extensive use of Excel and Statistics. It was an enormous challenge trying to work in Excel and manipulate such massive data.

Nonetheless, the project was very insightful and informative. It allowed me to learn to work with such datasets.

Throughout the project, I got a deeper understanding on EDA, right from analysing the data columns, to handling missing data and creating new analysis.

I was also able to learn about new add-ins in Excel such as Data Analyzer.

[Link to Excel File]

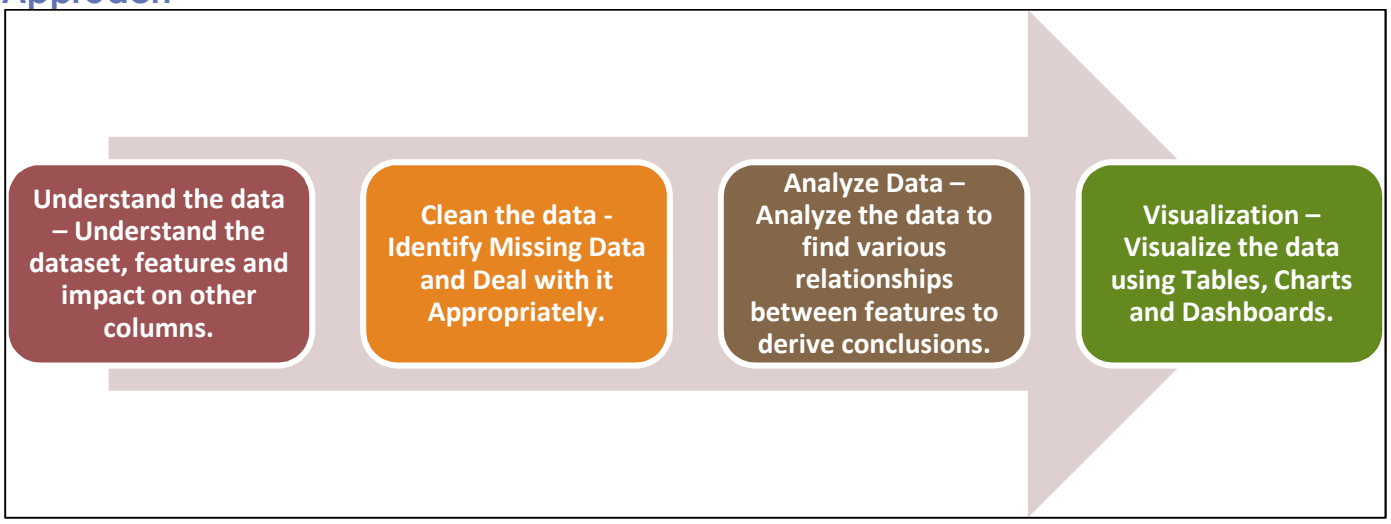
Project 7 - Analyzing the Impact of Car Features on Price and Profitability

Project Description

The automotive industry has been rapidly evolving over the past few decades, with a growing focus on fuel efficiency, environmental sustainability, and technological innovation. With increasing competition among manufacturers and a changing consumer landscape, it has become more important than ever to understand the factors that drive consumer demand for cars.

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. This could help the manufacturer improve its competitiveness in the market and increase its profitability over time.

Approach



Tech Stack Used

Microsoft Excel 2010 Version 14.0.7628.5000

Understanding the Data

Table Information (Original)	
Total Rows	11914
Total Columns	16
Total Blanks	108
Duplicates	715

Table Information (After Cleaning)	
Total Rows	11199
Total Columns	16
Total Blanks	0
Duplicates	0

Data Cleaning

Distribution of Blanks	
Columns	**Count of Blanks**
Make	0
Model	0
Year	0
Engine Fuel Type	3
Engine HP	69
Engine Cylinders	30
Transmission Type	0
Driven_Wheels	0
Number of Doors	6
Market Category	0
Vehicle Size	0
Vehicle Style	0
highway MPG	0
city mpg	0
Popularity	0
MSRP	0

Engine Fuel Type		
Engine Fuel Type	**Count**	**Change To**
diesel	150	Diesel
electric	66	Electric
flex-fuel (premium unleaded recommended/E85)	26	Flex-Fuel Premium
flex-fuel (premium unleaded required/E85)	53	Flex-Fuel Premium
flex-fuel (unleaded/E85)	887	Flex-Fuel Regular
flex-fuel (unleaded/natural gas)	6	Flex-Fuel Regular
natural gas	2	Natural Gas
premium unleaded (recommended)	1392	Premium Unleaded
premium unleaded (required)	1956	Premium Unleaded
regular unleaded	6658	Regular Unleaded
(blank)	3	Regular Unleaded. As the records have the same model and make, and the car used regular unleaded fuel type

Blank Number of Doors				
Make	**Model**	**Count of Make**	**New Value**	**Remarks**
Ferrari	FF	1	2	Same as the model of another record with simiar values
Tesla	Model S	5	4	Same as the model of another record with simiar values

Engine Cylinders					
Make	Model	Engine Fuel Type	Count of Make	New Value	Remarks
Chevrolet	Bolt EV	electric	2	0	Electric Vehicles have no cylinders
Mazda	RX-7	regular unleaded	3	4	Wankel 2 engine has no cylinders but is equivalent to a 4 cylinder engine
Mazda	RX-8	premium unleaded (required)	17	4	Wankel 2 engine has no cylinders but is equivalent to a 4 cylinder engine
Mitsubishi	i-MiEV	electric	3	0	Electric Vehicles have no cylinders
Toyota	RAV4 EV	electric	1	0	Electric Vehicles have no cylinders
Volkswagen	e-Golf	electric	4	0	Electric Vehicles have no cylinders

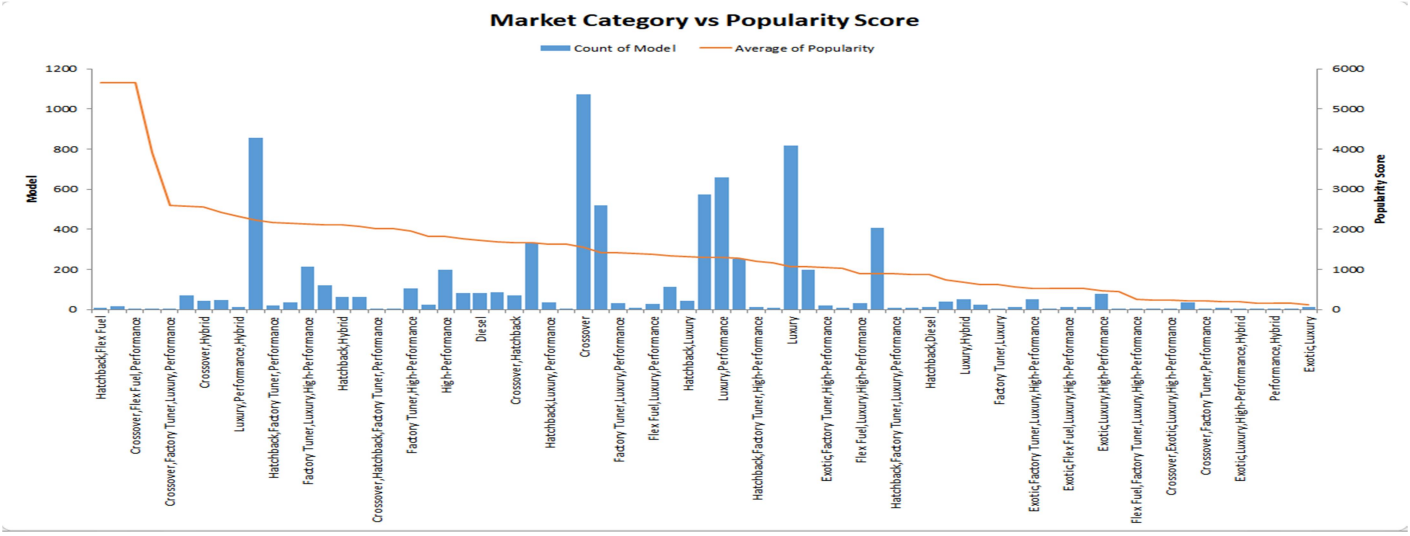
Updating the Blank Engine HP (Based on Certain Factors) Size of the engine, Number of cylinders, Type of fuel injection system							
Make	Model	Year	Engine Fuel Type	Engine Cylinders	Count of Make	New Engine HP	Remarks
Chevrolet	Impala	2015 – 2017	flex-fuel (unleaded/natural gas)	6	2	230	Online
FIAT	500e	2015 - 2017	electric	0	1	111	Online
Ford	Escape	2017	regular unleaded	4	4	168	Online
Ford	Focus	2015 - 2017	electric	0	1	143	Online
Ford	Freestar	2005	regular unleaded	6	6	201	Online
Honda	Fit EV	2013 - 2014	electric	0	1	123	Online
Kia	Soul EV	2015 - 2016	electric	0	2	109	Online
Lincoln	Continental	2017	premium unleaded (recommended)	6	4	400	Online
Lincoln	MKZ	2017	regular unleaded	4	4	245	Online
Mercedes-Benz	M-Class	2015	diesel	4	1	200	Online
Mitsubishi	i-MiEV	2014	electric	(blank)	1	66	Online
Nissan	Leaf	2014 - 2016	electric	0	3	107	Online
Tesla	Model S	2014 – 2016	electric	0	4	360	302,362 or 416
Toyota	RAV4 EV	2013 - 2014	electric	0	1	154	
Chevrolet	Impala	2015 – 2017	flex-fuel (unleaded/natural gas)	6	2	230	Online
FIAT	500e	2015 - 2017	electric	0	1	111	Online
Ford	Escape	2017	regular unleaded	4	4	168	Online
Ford	Focus	2015 - 2017	electric	0	1	143	Online
Ford	Freestar	2005	regular unleaded	6	6	201	Online
Honda	Fit EV	2013 - 2014	electric	0	1	123	Online
Kia	Soul EV	2015 - 2016	electric	0	2	109	Online
Lincoln	Continental	2017	premium unleaded (recommended)	6	4	400	Online
Lincoln	MKZ	2017	regular unleaded	4	4	245	Online
Mercedes-Benz	M-Class	2015	diesel	4	1	200	Online
Mitsubishi	i-MiEV	2014	electric	(blank)	1	66	Online
Nissan	Leaf	2014 - 2016	electric	0	3	107	Online
Tesla	Model S	2014 – 2016	electric	0	4	360	302,362 or 416
Toyota	RAV4 EV	2013 - 2014	electric	0	1	154	

Rename Columns	
Old Name	**New Name**
Driven Wheels	Drive Wheels
highway MPG	Highway MPG
city mpg	City MPG

- The #N/A in Market Category Column was left as is.
- All the Duplicate values were removed
- One Outlier 'Audi A6', had an highway MPG of 354, which was corrected to 34

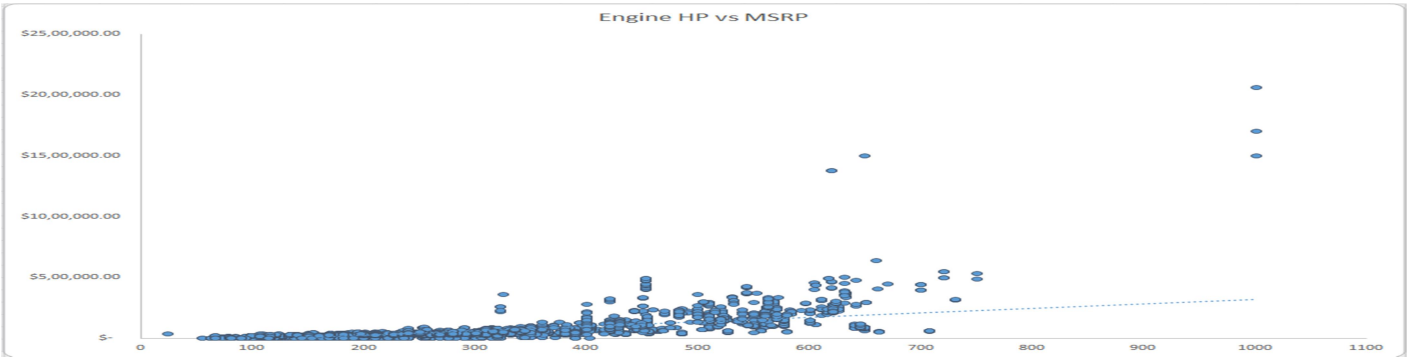
Data Analysis

How does the popularity of a car model vary across different market categories?
Task 1.A. Create a pivot table that shows the number of car models in each market category and their corresponding popularity scores.
Task 1.B. Create a combo chart that visualizes the relationship between market category and popularity.



Insight: 'Crossover' has the most number of models, but the 'Hatchback, Flex Fuel' is the most popular.

What is the relationship between a car's engine power and its price?
Task 2: Create a scatter chart that plots engine power on the x-axis and price on the y-axis. Add a trendline to the chart to visualize the relationship between these variables.



Insight: 'Engine HP' and 'MSRP' have a positive linear relation.
Note the Dodge Challenger which is high on HP, but low on MSRP. This could be due to other features such as low city and highway MPG, or fuel type.

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

Task 3: Use regression analysis to identify the variables that have the strongest relationship with a car's price. Then create a bar chart that shows the coefficient values for each variable to visualize their relative importance.

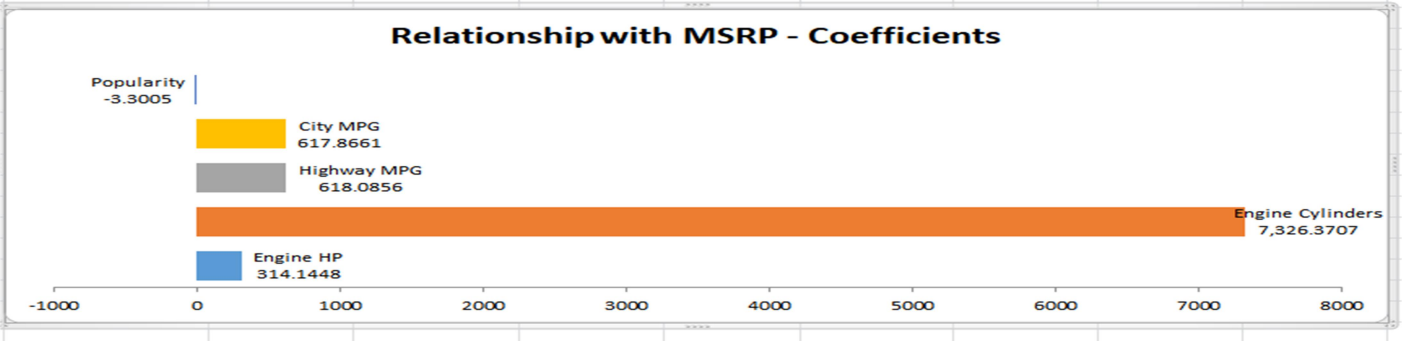
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.677769875
R Square	0.459372003
Adjusted R Square	0.4591305
Standard Error	45255.23123
Observations	11199

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	1.94783E+13	3.89565E+12	1902.140051	0
Residual	11193	2.29237E+13	2048035954		
Total	11198	4.24019E+13			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-102560.8823	3115.699988	-32.91744478	6.5455E-227	-108668.2025	-96453.56213	-108668.2025	-96453.56213
Engine HP	314.1448474	6.28046852	50.01933318	0	301.8340241	326.4556708	301.8340241	326.4556708
Engine Cylinders	7326.370725	454.4774819	16.12042624	8.19089E-58	6435.514895	8217.226555	6435.514895	8217.226555
Highway MPG	618.0856489	107.1773191	5.766944482	8.2858E-09	407.9992457	828.1720521	407.9992457	828.1720521
City MPG	617.8660941	101.4618006	6.089642511	1.16876E-09	418.9831128	816.7490754	418.9831128	816.7490754
Popularity	-3.300531052	0.296297403	-11.13925069	1.14296E-28	-3.881326095	-2.719736009	-3.881326095	-2.719736009

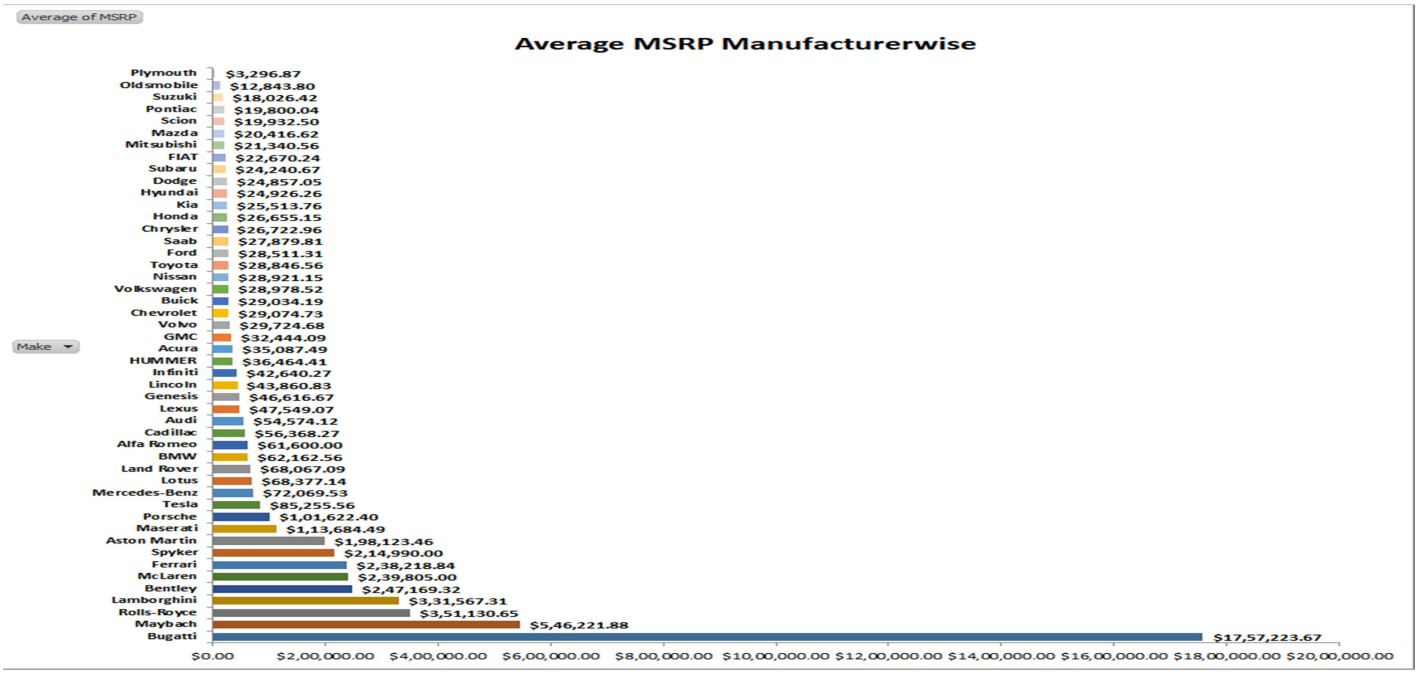


Insight: 'Engine Cylinder' and 'MSRP' have the highest positive coefficient Whereas, 'Popularity' has a negative coefficient.

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

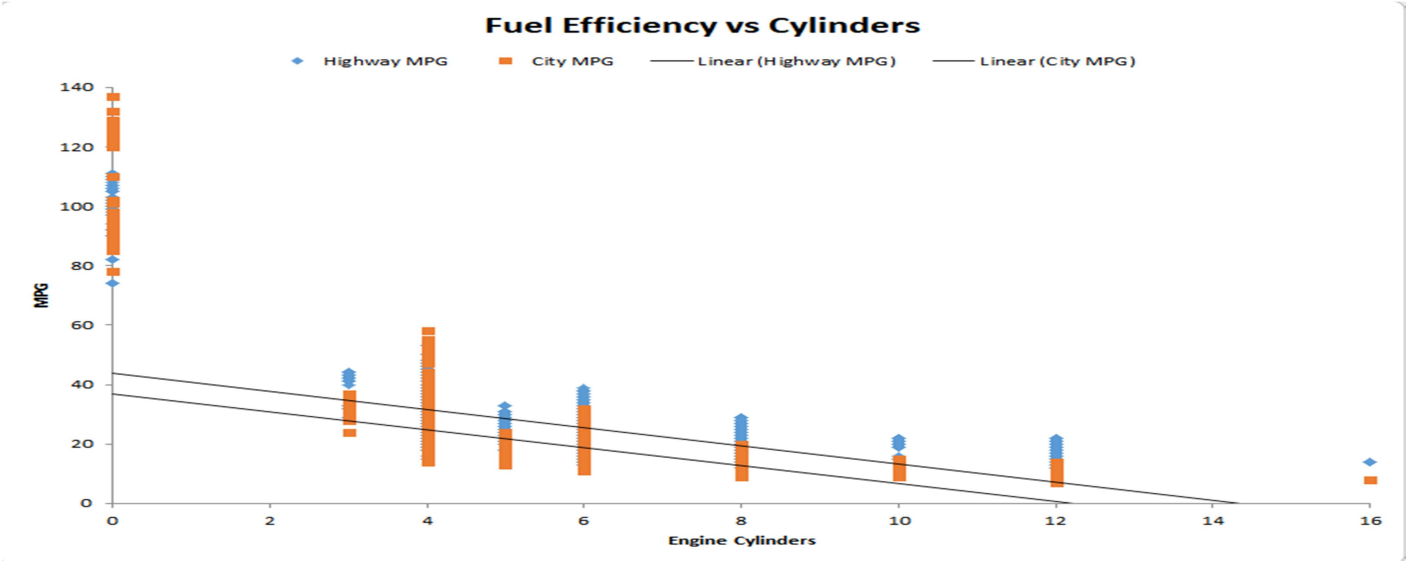
Task 4.A: Create a pivot table that shows the average price of cars for each manufacturer.

Task 4.B: Create a bar chart or a horizontal stacked bar chart that visualizes the relationship between manufacturer and average price.



Insight: Luxury Brands like Bugatti, Maybach and Rolls-Royce were the most expensive.

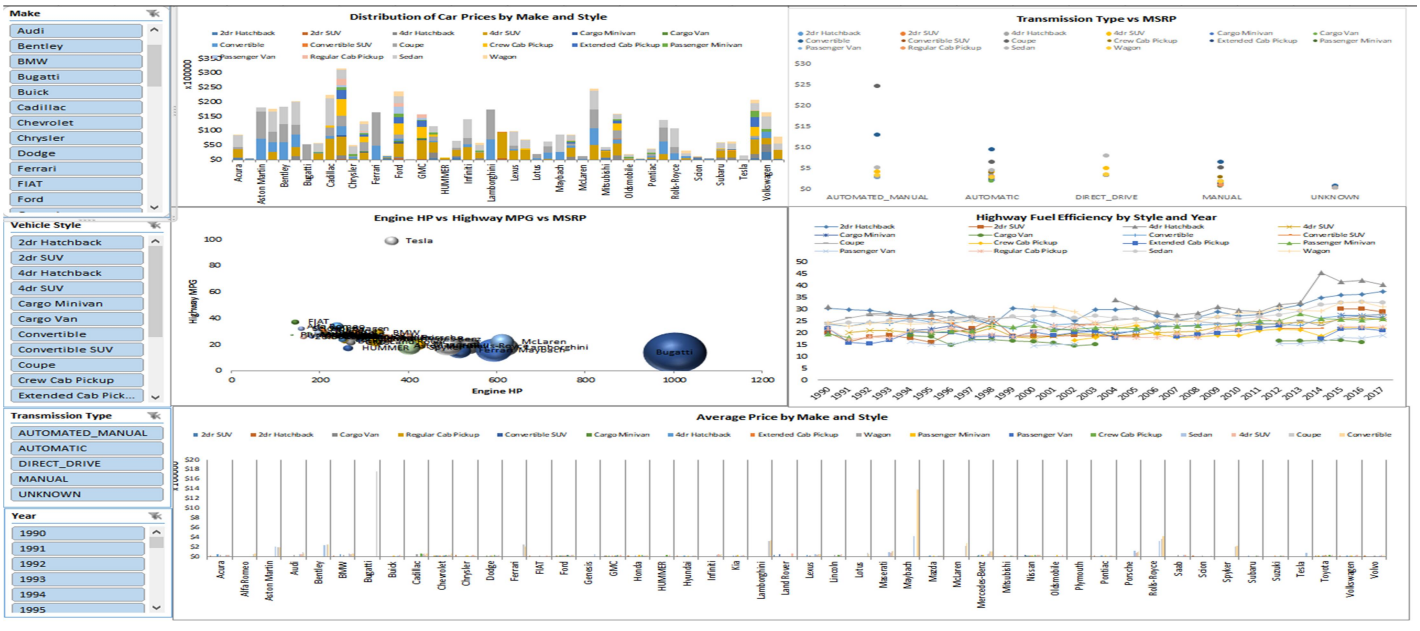
What is the relationship between fuel efficiency and the number of cylinders in a car's engine?
Task 5.A: Create a scatter plot with the number of cylinders on the x-axis and highway MPG on the y-axis. Then create a trendline on the scatter plot to visually estimate the slope of the relationship and assess its significance.
Task 5.B: Calculate the correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.



Insight: For both Highway and City MPG, as the cylinders increase, the fuel efficiency decreases.

Dashboard

Task 1: How does the distribution of car prices vary by brand and body style?
Task 2: Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?
Task 3: How do the different features such as transmission type affect the MSRP, and how does this vary by body style?
Task 4: How does the fuel efficiency of cars vary across different body styles and model years?
Task 5: How does the car's horsepower, MPG, and price vary across different Brands?



Summary

This particular project involved an extensive usage of Excel and Statistics.
Most of the analysis was done by using Pivot Tables and Charts, for which a lot of research was done.
Creating a dashboard was a new and interesting challenge, which was difficult, and could have been easily done using PowerBI or Tableau.
In Statistics, achieving the Regression Analysis for many columns together was difficult, but after researching it, was able to understand and complete the task.
The major challenge faced was in understanding the data. Updating the NULL/Missing values proved a challenge as most of the data required internet research. There were some data irregularities, which were dealt with.
Overall, this was a very challenging but creative and informative project.

Link to Excel File

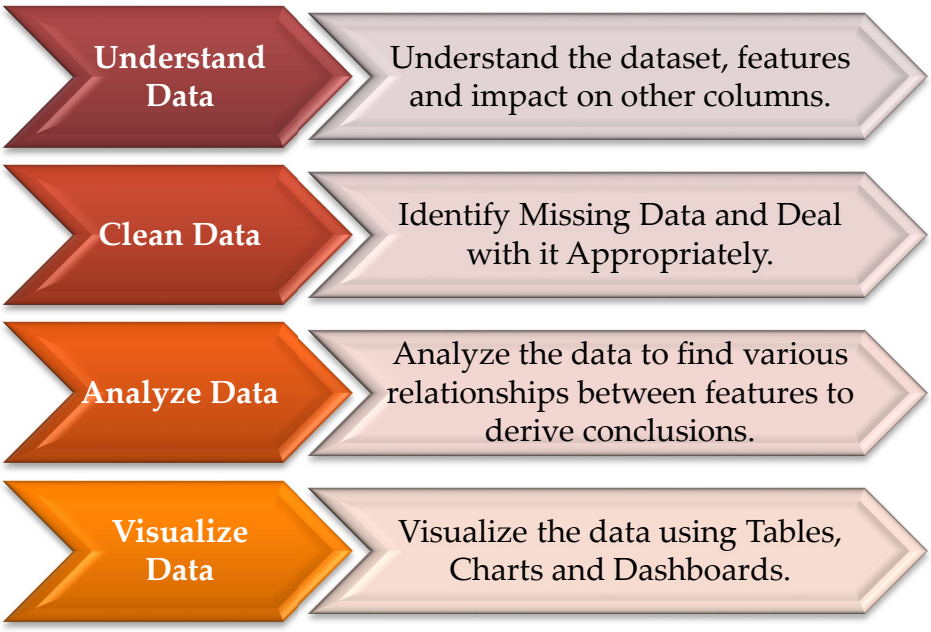
Car Data.xlsx

Project 8 - ABC Call Volume Trend Analysis

Project Description

A Customer Experience (CX) team plays a crucial role in a company. They analyze customer feedback and data, derive insights from it, and share these insights with the rest of the organization. This team is responsible for a wide range of tasks, including managing customer experience programs, handling internal communications, mapping customer journeys, and managing customer data, among others.
One of the key roles in a CX team is that of the customer service representative or call center agent. These agents handle various types of support, including email, inbound, outbound, and social media support.
The dataset provided spans 23 days and includes various details such as the agent's name and ID, the queue time, the time and duration of the call and the call status.
Inbound customer support, which is the focus of this project, involves handling incoming calls from existing or prospective customers. The goal is to attract, engage, and delight customers, turning them into loyal advocates for the business.

Approach



Tech Stack Used

Microsoft Excel 2010 Version 14.0.7628.5000

Understand Data

Table Information (Original)	
Total Rows	117988
Total Columns	13
Total Blanks	47877
Duplicates	0

Columns	Blank Count
Agent_Name	0
Agent_ID	0
Customer_Phone_No	0
Queue_Time(Secs)	0
Date_&_Time	0
Time	0
Time_Bucket	0
Duration(hh:mm:ss)	0
Call_Seconds (s)	0
Call_Status	0
Wrapped _By	47877
Ringing	0
IVR _Duration	0

Clean Data

Table Information (Clean)	
Total Rows	117988
Total Columns	13
Total Blanks	0
Duplicates	0

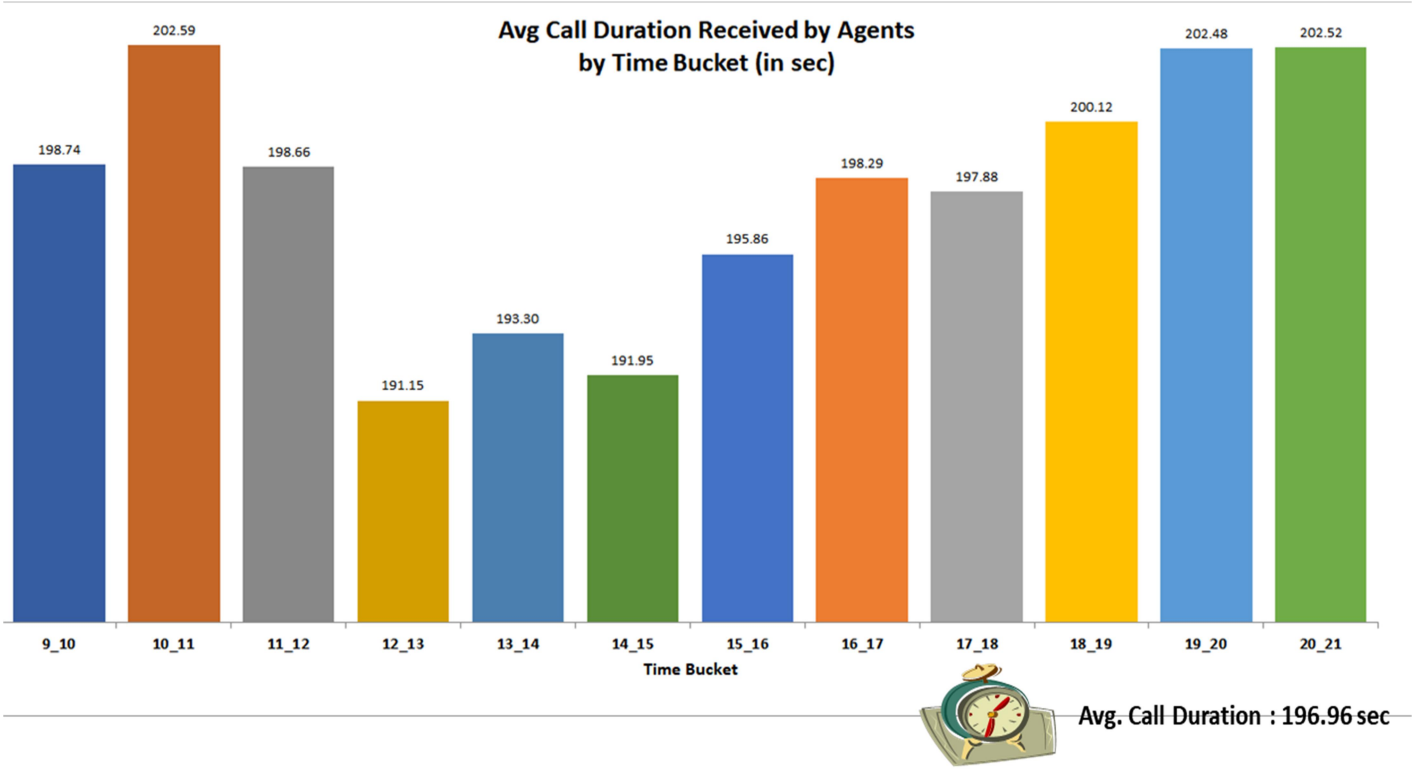
Blank/Null Value Update			
Wrapped By	**Call Status**	**Count**	**New Value**
Blank	Abandon	34403	Not Available
Blank	Answered	13362	Agent
Blank	Transfer	112	Agent

Value Changes		
Column Name	**Old Value**	**New Value**
Customer Phone No.	XXXXX	XXXXXXXXXX
Customer Phone No.	CzentXXXXX	No Change as this could be a Special Number made using the alphabets associated with the numbers on a dialpad (CZENT = 29368)

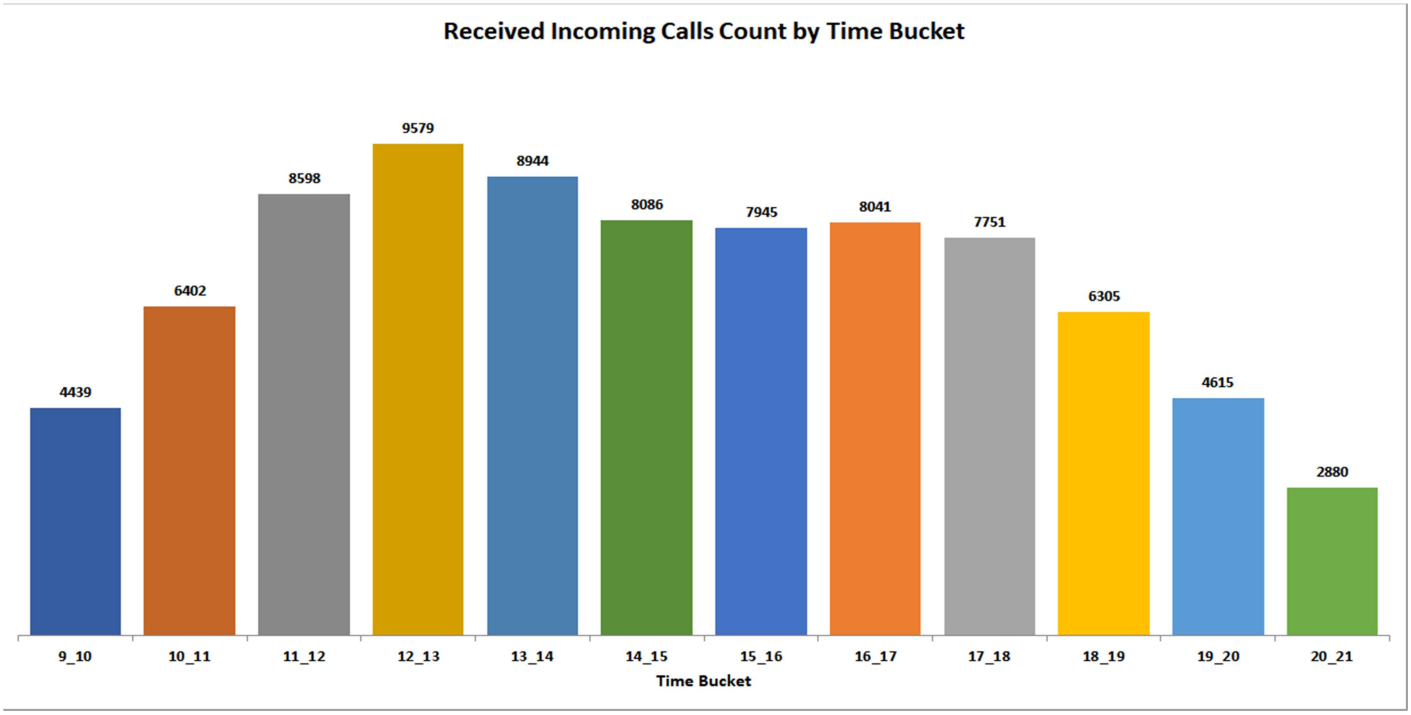
Column Name	Format Change
Agent_Name	Converted to Text
Agent_ID	Converted to Text
Customer_Phone_No	Converted to Text
Queue_Time(Secs)	Converted to Number without Decimal
Date_&_Time	Converted to "dd-mmm-yyyy hh:mm:ss" Format
Time	Converted to Number without Decimal
Time_Bucket	Converted to Text
Duration(hh:mm:ss)	No Format Change
Call_Seconds (s)	Converted to Number without Decimal
Call_Status	Converted to Text
Wrapped _By	Converted to Text
Ringing	Converted to Text
IVR _Duration	No Format Change

Analyze & Visualize Data

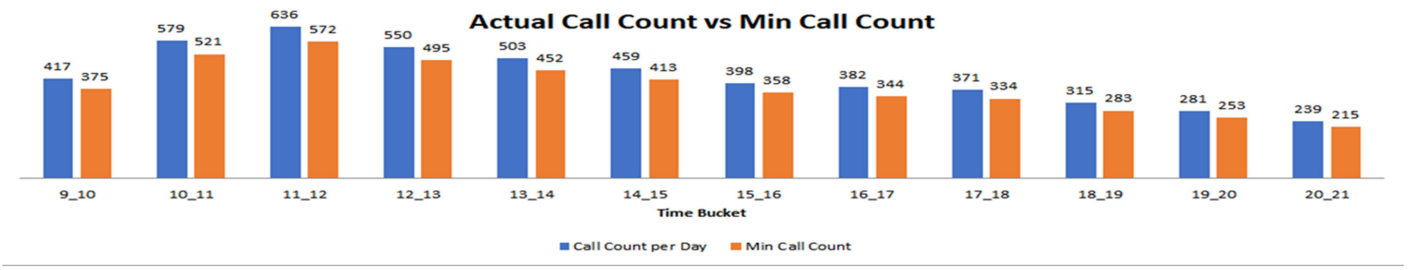
Average Call Duration: Determine the average duration of all incoming calls received by agents. This should be calculated for each time bucket.
Your Task: What is the average duration of calls for each time bucket?

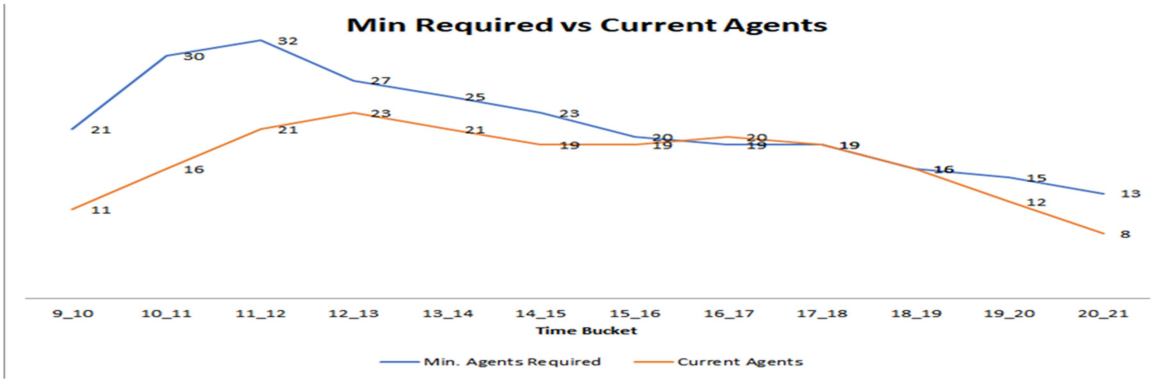


Call Volume Analysis: Visualize the total number of calls received. This should be represented as a graph or chart showing the number of calls against time. Time should be represented in buckets (e.g., 1-2, 2-3, etc.).
Your Task: Can you create a chart or graph that shows the number of calls received in each time bucket?



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%.
Your Task: What is the minimum number of agents required in each time bucket to reduce the abandon rate to 10%?



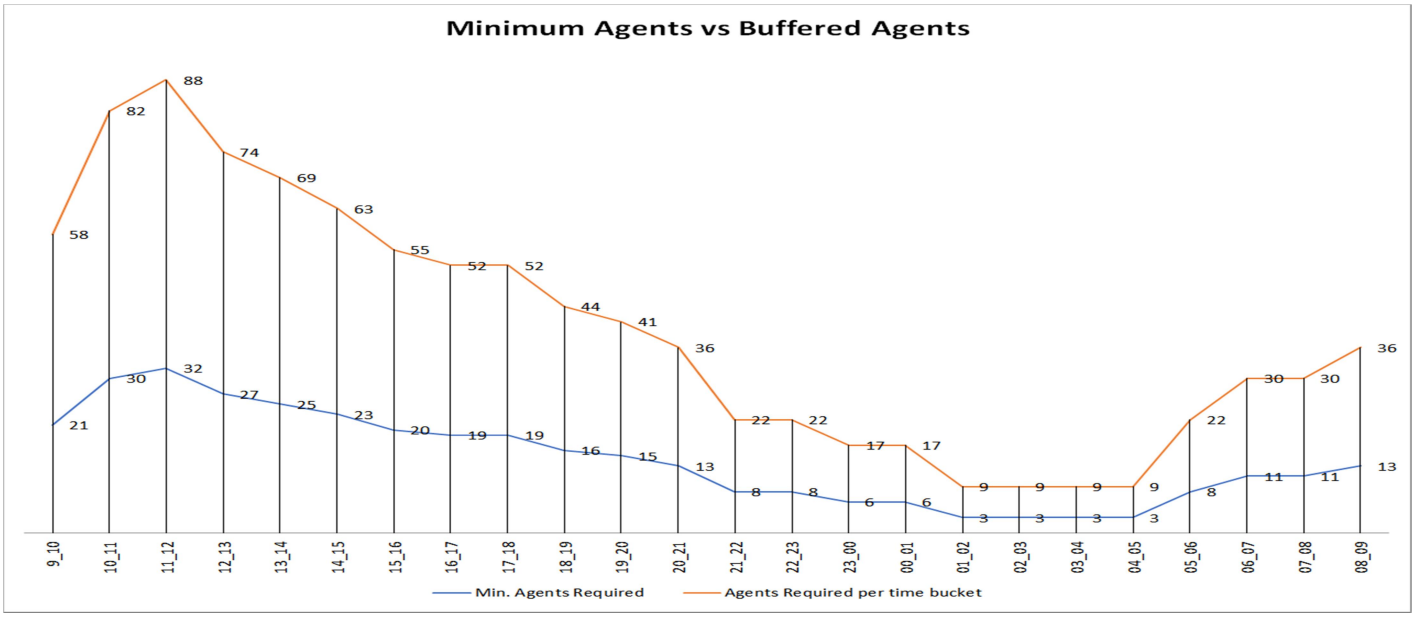


Night Shift Manpower Planning: Customers also call ABC Insurance Company at night but don't get an answer because there are no agents available. This creates a poor customer experience. Assume that for every 100 calls that customers make between 9 am and 9 pm, they also make 30 calls at night between 9 pm and 9 am. The distribution of these 30 calls is as follows:

Your Task: Propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%.

Assumptions and Calculations	
Total Days	30 days
Working Days in week	6 days
Unplanned Leaves per month	4 days
Working Hours in day	9
Lunch and Breaks (in hrs)	1.5
Actual Working Hours (AWH)	7.5
Occupancy %age of AWH	60%
Occupancy (in hrs)	4.50
Non-Occupancy (in hrs)	4.50
Non-Occupancy %age	50%
Waiting Probability	100%
Abandon Rate	10%
Shrinkage (4 days Unplanned & 4 Days Weekly Off)	26.67%
Average Call Duration for 2100 to 0900 hrs is an assumption based on actual data	

Time Bucket	Average Call Duration	Call Distribution	Call Count per Day	Abandon Rate	Min Call Count	Max Calls that can be received	Min. Agents Required	Non-Occupancy %age (60% of 7.5 hrs + 1.5 hrs Breaks)	Agents required after taking into account of occupancy	Shrinkage (8 days in a month - 4 unplanned and 4 w/off)	Agents Required per time bucket
9_10	198.74		417	10%	375	18	21	50%	42	26.67%	58
10_11	202.59		579	10%	521	18	30	50%	60	26.67%	82
11_12	198.66		636	10%	572	18	32	50%	64	26.67%	88
12_13	191.15		550	10%	495	19	27	50%	54	26.67%	74
13_14	193.30		503	10%	452	19	25	50%	50	26.67%	69
14_15	191.95		459	10%	413	19	23	50%	46	26.67%	63
15_16	195.86		398	10%	358	18	20	50%	40	26.67%	55
16_17	198.29		382	10%	344	18	19	50%	38	26.67%	52
17_18	197.88		371	10%	334	18	19	50%	38	26.67%	52
18_19	200.12		315	10%	283	18	16	50%	32	26.67%	44
19_20	202.48		281	10%	253	18	15	50%	30	26.67%	41
20_21	202.52		239	10%	215	18	13	50%	26	26.67%	36
21_22	196.96	3	154	10%	139	18	8	50%	16	26.67%	22
22_23	196.96	3	154	10%	139	18	8	50%	16	26.67%	22
23_00	196.96	2	103	10%	92	18	6	50%	12	26.67%	17
00_01	196.96	2	103	10%	92	18	6	50%	12	26.67%	17
01_02	196.96	1	51	10%	46	18	3	50%	6	26.67%	9
02_03	196.96	1	51	10%	46	18	3	50%	6	26.67%	9
03_04	196.96	1	51	10%	46	18	3	50%	6	26.67%	9
04_05	196.96	1	51	10%	46	18	3	50%	6	26.67%	9
05_06	196.96	3	154	10%	139	18	8	50%	16	26.67%	22
06_07	196.96	4	205	10%	185	18	11	50%	22	26.67%	30
07_08	196.96	4	205	10%	185	18	11	50%	22	26.67%	30
08_09	196.96	5	256	10%	231	18	13	50%	26	26.67%	36



Summary

This project involved research on Time Series and Manpower Planning at Call Centres.
Most of the analysis was done by using Pivot Tables and Formulas.
In Statistics, a lot of research was done on how a Call Centre Manpower Planning is done. New concepts like Erlang Formula was researched and understood. This helped understand better how to solve the problem.
Overall, though the project started out easy and familiar, the major challenge was the Manpower planning which required a lot of research and learning.

Link to Excel File

Call Volume Trend Analysis (Excel File)

Appendix

Powerpoint Presentations on Projects

Project 1 – Data Analytics

Project 2 – Instagram User Analytics

Project 3 – Operation & Metric Analytics

Project 4 – Hiring Process Analytics

Project 5 – IMDB Movie Analysis

Project 6 – Bank Loan Case Study

Project 7 – Analysing the Impact of Car Features on Price & Profitability

Project 8 – ABC Call Volume Trend Analysis