Data Analysis Portfolio

Preented by:- Samarjith M N

Professional Background

Samarjith M N, Electronics and Communication Engineering graduate from the class of 2023, with a passion for leveraging technology to solve real-world challenges. I recently completed an internship at GTTC, Mysore, where I served as a PLC and IoT Intern.

My skill set includes proficiency in Python, R, Excel, Power BI, data analysis, statistics, and Tableau. I am well-versed in using these tools to extract meaningful insights from data, facilitating informed decision-making. The analytical skills acquired during my academic journey, coupled with practical exposure during my internship, have equipped me to navigate the complexities of the corporate world.

As a recent graduate, I bring a fresh perspective and a strong desire to tackle the challenges of the corporate environment. I consider myself highly adaptable and flexible, eager to learn and grow in a dynamic setting. While I possess a solid foundation in theoretical concepts, I am enthusiastic about channelling my efforts towards practical applications, confident that continuous learning and dedication will be key to my success. I am excited about the prospect of contributing to innovative projects and expanding my expertise in the field.

Table of Contents

Data Analytics Process: Buying Mobile

1. Plan

- Determine the objective Our objective is to purchase a mobile phone that meets our specific requirements.
- Define the target market Our target market is individuals aged between 20-30, seeking a stylish and powerful mobile phone for both personal and professional use.

2. Prepare

- Identify potential suppliers We identify potential suppliers of mobile phones in our target market by researching online, attending mobile phone exhibitions, and contacting relevant industry professionals.

- Conduct market research We gather relevant information about our target market, including demographics, preferences, and behaviour patterns. This information helps us to better understand our target audience and their needs.

3. Process

- Obtain supplier information We obtain comprehensive information about each supplier, including their product offerings, pricing, customer reviews, and after-sales services.
- Select the appropriate supplier After evaluating all the available suppliers based on our objective and target market, we select the supplier that best meets our needs.

4. Analyze

- Assess supplier products We analyze the features, specifications, and design of the mobile phones offered by the selected supplier.
- Evaluate the supplier's overall performance We assess the supplier's overall performance, including its reputation, reliability, and commitment to customer satisfaction.
- Make a final decision Based on our evaluation, we decide which mobile phone model to purchase from the selected supplier.

5. Share

- Provide feedback to the supplier We communicate our concerns, preferences, and expectations to the supplier, so they can better understand our requirements and deliver a higher quality product.
- Offer advice to other potential customers We provide insights and recommendations to individuals considering purchasing a mobile phone, helping them make an informed decision.

6. Act

- Complete the purchase We proceed with the purchase of the selected mobile phone model, adhere. adhering to the terms and conditions specified by the supplier.
- Share the purchasing experience After the purchase, we share our experiences with others, including the effectiveness of the supplier's services, the quality of the purchased mobile phone, and any recommendations we may have for future customers.

Instagram User Analytics

Project Description:

The project focuses on learaging SQL and MySQL Workbench to analyze user interactions and engagement with the Instagram app. Its primary goal is to extract meaningful insights that can guide decision-making for the product team. By delving into user data, the aim is to identify patterns, trends, and user behaviors that could influence future app development strategies.

Approach:

Data Gathering: Accessing and extracting relevant user engagement data from Instagram's databases.

Data Preparation: Cleaning and pre-processing the data to handle any inconsistencies, missing values, or outliers.

SQL Analysis: Utilizing SQL queries to perform various analyses, including user engagement metrics, popular features, user demographics, and more.

Insight Generation: Deriving actionable insights by interpreting the analyzed data and identifying key trends or user behaviors.

Tech-Stack Used:

MySQL Workbench (Version 8.0 CE): I chose MySQL Workbench as my primary tool for this project. It is a comprehensive Integrated Development Environment (IDE) for SQL, allowing developers to create, edit, test, and debug SQL statements

SQL: Leveraged for its querying power to efficiently extract and analyze the required data from Instagram's databases.

Insights:

User Engagement Patterns: Identified peak usage times, frequently used features, and patterns in user interactions within the app.

Demographic Insights: Uncovered demographic trends among engaged users, such as age groups or geographic locations.

Feature Analysis: Highlighted which app features drive higher engagement and user retention.

Data Analytics Tasks

A) 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. Your Task: Identify the five oldest users on Instagram from the provided database.

```
select * from users
order by created_at
limit 5;
```

Result Grid | Filter Rows: | Edit:

id	username	created_at
80	Darby_Herzog	2016-05-06 00:14:21
67	Emilio_Bernier52	2016-05-06 13:04:30
63	Elenor88	2016-05-08 01:30:41
95	Nicole71	2016-05-09 17:30:22
38	Jordyn.Jacobson2	2016-05-14 07:56:26
NULL	NULL	NULL

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

```
SELECT *
FROM users
where id not in (select distinct user_id from photos )
order by id;
```

id	username	created_at
5	Aniya_Hackett	2016-12-07 01:04:39
7	Kasandra_Homenick	2016-12-12 06:50:08
14	Jaclyn81	2017-02-06 23:29:16
21	Rocio33	2017-01-23 11:51:15
24	Maxwell.Halvorson	2017-04-18 02:32:44
25	Tierra.Trantow	2016-10-03 12:49:21
34	Pearl7	2016-07-08 21:42:01
36	Ollie_Ledner37	2016-08-04 15:42:20
41	Mckenna17	2016-07-17 17:25:45
45	David.Osinski47	2017-02-05 21:23:37
49	Morgan.Kassulke	2016-10-30 12:42:31
53	Linnea59	2017-02-07 07:49:34
54	Duane60	2016-12-21 04:43:38
57	Julien_Schmidt	2017-02-02 23:12:48
66	Mike.Auer39	2016-07-01 17:36:15

Result Grid | Filter Rows: | Edit: | Export/Import:

id	username	created_at
66	Mike.Auer39	2016-07-01 17:36:15
68	Franco_Keebler64	2016-11-13 20:09:27
71	Nia_Haag	2016-05-14 15:38:50
74	Hulda.Macejkovic	2017-01-25 17:17:28
75	Leslie	2016-09-21 05:14:01
76	Janelle.Nikolaus81	2016-07-21 09:26:09
80	Darby_Herzog	2016-05-06 00:14:21
81	Esther.Zulauf61	2017-01-14 17:02:34
83	Bartholome.Bernhard	2016-11-06 02:31:23
89	Jessyca_West	2016-09-14 23:47:05
90	Esmeralda.Mraz57	2017-03-03 11:52:27
91	Bethany20	2016-06-03 23:31:53
NULL	NULL	NULL

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

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

```
select distinct(user_id),max(photo_id)as max_likes,created_at from likes
group by user_id,created_at
order by max_likes desc
limit 1;
```

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

user_id	max_likes	created_at
3	257	2023-11-07 15:54:09

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

```
select tag_id ,count(*) as counts from photo_tags
group by tag_id
order by counts desc
limit 5;
```

Result Grid | Filter Rows: | Export:

tag_id	counts
21	59
20	42
17	39
13	38
18	24

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

```
select dayname(created_at) as days, count(*) as counts from users
group by days
order by counts desc
limit 1;
```

Result Grid | Filter Rows: | Export: | W

days	counts
Thursday	16

B) 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.
   Your 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.

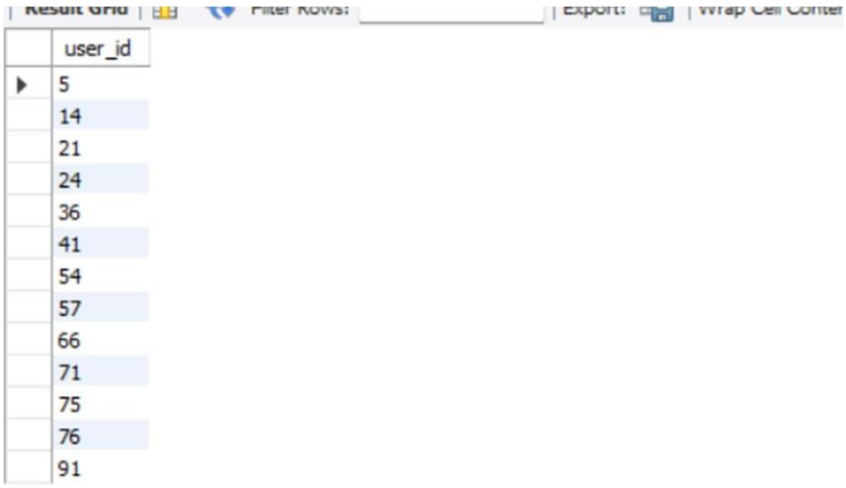
```
-- Calculate the total number of photos on Instagram divided by the total number of users
SELECT COUNT(*) AS total_photos, COUNT(DISTINCT user_id) AS total_users,
       COUNT(*) / COUNT(DISTINCT user_id) AS photos_per_user_average
FROM photos;
```

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

total_photos	total_users	photos_per_user_average
257	74	3.4730

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

```
SELECT user_id
FROM (
    SELECT l.user_id, COUNT(*) AS total_likes
    FROM likes l
    JOIN photos p ON l.photo_id = p.id
    GROUP BY l.user_id
) AS user_likes
WHERE total_likes = (SELECT COUNT(*) FROM photos);
```



user_id
5
14
21
24
36
41
54
57
66
71
75
76
91

Result:

Achievements: Generated detailed insights into user behaviors and preferences within the Instagram app.

Valuable Information: Provided actionable information for the product team to consider for future feature development, user experience enhancements, and targeted marketing strategies.

Impact of Analysis: Contributed to informed decision-making processes, potentially influencing the direction of Instagram's app development to better meet user needs.

Drive Link:

https://drive.google.com/file/d/1-Zo9HAy9x9FsMNGYfywtxzQVJNr9aTES/view?usp=sharing

Operation Analytics and Investigating Metric Spike

Project Overview:

The project focuses on operational analytics, involving the analysis of diverse company operations data to identify areas for improvement. As the Lead Data Analyst at a company similar to Microsoft, the aim was to utilize advanced SQL skills to investigate metric spikes, particularly sudden changes in key metrics impacting user engagement, sales, and other crucial business areas.

Approach:

To handle the analysis, a structured approach was adopted:

Data Collection & Understanding: Acquiring diverse datasets representing various operational aspects.

Data Cleaning & Preparation: Ensuring data quality by addressing inconsistencies, missing values, and transforming data for analysis.

SQL Analysis: Utilizing MySQL Workbench (version 8.0) extensively to perform complex SQL queries and manipulations.

Identification of Metric Spikes: Employing SQL queries to detect sudden changes in key metrics.

Collaboration & Reporting: Communicating insights and findings to relevant departments, fostering collaboration for actionable steps.

Tech-Stack Used:

MySQL Workbench (v8.0): Utilized as the primary tool for SQL analysis, enabling efficient querying, data manipulation, and visualization.

Insights:

Key insights and observations derived during the analysis:

- Identified Sudden User Engagement Decline: Detected a significant drop in daily user engagement metrics, coinciding with a software update.
- Sales Dip Corresponding to Marketing Strategy Change: Uncovered a decline in sales following alterations in marketing strategies.
- Operational Bottlenecks Impacting Customer Support: Discovered delays in handling customer support tickets due to a system integration issue.
- Insights and observations from the tasks:

- Used SQL to calculate jobs reviewed per hour for each day in November 2020.
- Created an SQL query for a 7-day rolling average of throughput.
- Preferred the 7-day rolling average as it smooths out fluctuations and provides a more stable representation of throughput trends.
- Calculated the percentage share of each language in the last 30 days using SQL.
- Developed an SQL query to identify and display duplicate rows from the job_data table.
- Crafted an SQL query to measure weekly user engagement based on user actions.
- Employed SQL to calculate user growth over time for the product.
- Constructed an SQL query to analyze weekly retention of users based on their sign-up cohort.
- Utilized SQL to calculate weekly user engagement per device.
- Created an SQL query to analyze how users are engaging with the email service, likely by examining various metrics related to email interactions such as opens, clicks, etc.

Case Study 1: Job Data Analysis

Creating database and table

```sql
create database casestudy_1;
CREATE TABLE job_data
(
    ds DATE,
    job_id INT NOT NULL,
    actor_id INT NOT NULL,
    event VARCHAR(15) NOT NULL,
    language VARCHAR(15) NOT NULL,
    time_spent INT NOT NULL,
    org CHAR(2)
);

INSERT INTO job_data (ds, job_id, actor_id, event, language, time_spent, org)
VALUES ('2020-11-30', 21, 1001, 'skip', 'English', 15, 'A'),
    ('2020-11-30', 22, 1006, 'transfer', 'Arabic', 25, 'B'),
    ('2020-11-29', 23, 1003, 'decision', 'Persian', 20, 'C'),
    ('2020-11-28', 23, 1005,'transfer', 'Persian', 22, 'D'),
    ('2020-11-28', 25, 1002, 'decision', 'Hindi', 11, 'B'),
    ('2020-11-27', 11, 1007, 'decision', 'French', 104, 'D'),
    ('2020-11-26', 23, 1004, 'skip', 'Persian', 56, 'A'),
    ('2020-11-25', 20, 1003, 'transfer', 'Italian', 45, 'C');

```

Jobs Reviewed Over Time:

Objective: 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.

```
SELECT DATE(ds) AS review_date,
       COUNT(*) AS jobs_reviewed_per_day,sum(time_spent)/3600 as jobs_reviewed_per_hr
FROM job_data
WHERE ds between '2020-11-01' and '20-11-30'
GROUP BY review_date
ORDER BY review_date;
```

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

review_date	jobs_reviewed_per_day	jobs_reviewed_per_hr
2020-11-25	1	0.0125
2020-11-26	1	0.0156
2020-11-27	1	0.0289
2020-11-28	2	0.0092
2020-11-29	1	0.0056
2020-11-30	2	0.0111

Throughput Analysis:

Objective: 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. Additionally, explain whether you prefer using the daily metric or the 7-day rolling average for throughput, and why.

```
29
30  SELECT ds,
31         time_spent,
32         AVG(time_spent)
33         OVER (ORDER BY ds ROWS BETWEEN 6 PRECEDING AND CURRENT ROW) AS rolling_7_day_average
34  FROM job_data;
35
```

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

ds	time_spent	rolling_7_day_average
2020-11-25	45	45.0000
2020-11-26	56	50.5000
2020-11-27	104	68.3333
2020-11-28	22	56.7500
2020-11-28	11	47.6000
2020-11-29	20	43.0000
2020-11-30	15	39.0000
2020-11-30	25	36.1429

Result Grid | Form Editor | Result 6 | Read Only

Reason:

Daily Metric: This provides the raw, day-to-day values of throughput without any smoothing or averaging. It might be useful when examining short-term trends or

fluctuations. Daily metrics are sensitive to sudden changes and can provide insights into specific days' performances.

7-Day Rolling Average: This smoothed metric helps in identifying overall trends and patterns by averaging the data over a 7-day period, reducing the impact of daily fluctuations or irregularities. It's particularly useful for understanding long-term trends, identifying broader patterns, and eliminating noise in the data.

Language Share Analysis:

Objective: 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.

```
36 •  select language,
37        round(100*count(*)/(select count(*) from job_data),2) as percentage_share
38    from job_data
39    group by language
40    order by language desc;
```

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

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

Duplicate Rows Detection:

Objective: Identify duplicate rows in the data.

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

```
42 •  select *
43    from
44    (select *,
45    row_number() over(partition by ds,actor_id,job_id) as row_num
46    from job_data) a
47    where row_num>1;
```

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

ds	job_id	actor_id	event	language	time_spent	org	row_num

Case Study 2: Investigating Metric Spike

Weekly User Engagement:

Objective: Measure the activeness of users on a weekly basis.

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

```sql
SELECT
    DATE_ADD(created_at, INTERVAL -WEEKDAY(created_at) DAY) AS week_start_date,
    COUNT(DISTINCT user_id) AS active_users_count
FROM
    users
GROUP BY
    week_start_date;
```

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

week_start_date	active_users_count
2012-12-31 02:52:00	1
2012-12-31 04:38:00	1
2012-12-31 08:07:00	1
2012-12-31 08:28:00	1
2012-12-31 09:29:00	1
2012-12-31 09:41:00	1
2012-12-31 09:54:00	1
2012-12-31 10:39:00	1
2012-12-31 10:56:00	1
2012-12-31 11:51:00	1
2012-12-31 12:16:00	1
2012-12-31 12:27:00	1

Result 1

User Growth Analysis:

Objective: Analyze the growth of users over time for a product.

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

```sql
SELECT
    DATE_ADD(created_at, INTERVAL -DAYOFMONTH(created_at) + 1 DAY) AS month_start_date,
    COUNT(DISTINCT user_id) AS total_users
FROM users
GROUP BY month_start_date
ORDER BY month_start_date;
```

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

month_start_date	total_users
2013-01-01 00:14:00	2
2013-01-01 00:17:00	1
2013-01-01 00:34:00	1
2013-01-01 01:04:00	1
2013-01-01 02:11:00	1
2013-01-01 02:52:00	1
2013-01-01 02:54:00	1
2013-01-01 02:58:00	1
2013-01-01 04:20:00	1
2013-01-01 04:38:00	1
2013-01-01 05:13:00	1
2013-01-01 05:44:00	1
2013-01-01 06:19:00	1

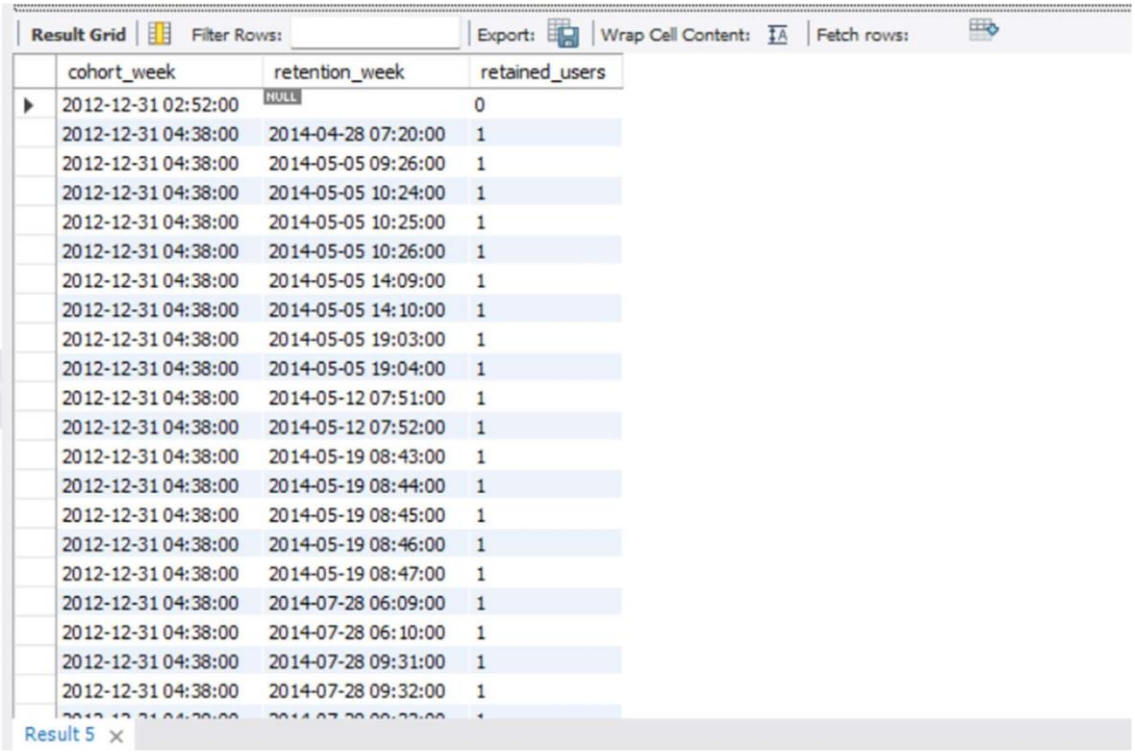
Result 3

Weekly Retention Analysis:

Objective: Analyze 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
WITH user_signups AS (
    SELECT
        user_id,
        DATE_ADD(created_at, INTERVAL -WEEKDAY(created_at) DAY) AS signup_week
    FROM users
),
user_activity AS (
    SELECT
        user_id,
        DATE_ADD(occurred_at, INTERVAL -WEEKDAY(occurred_at) DAY) AS activity_week
    FROM events
)
SELECT
    us.signup_week AS cohort_week,
    ua.activity_week AS retention_week,
    COUNT(DISTINCT ua.user_id) AS retained_users
FROM user_signups us
LEFT JOIN
    user_activity ua ON us.user_id = ua.user_id AND ua.activity_week >= us.signup_week
GROUP BY
    us.signup_week, ua.activity_week
ORDER BY
    us.signup_week, ua.activity_week;
```



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

cohort_week	retention_week	retained_users
2012-12-31 02:52:00	NULL	0
2012-12-31 04:38:00	2014-04-28 07:20:00	1
2012-12-31 04:38:00	2014-05-05 09:26:00	1
2012-12-31 04:38:00	2014-05-05 10:24:00	1
2012-12-31 04:38:00	2014-05-05 10:25:00	1
2012-12-31 04:38:00	2014-05-05 10:26:00	1
2012-12-31 04:38:00	2014-05-05 14:09:00	1
2012-12-31 04:38:00	2014-05-05 14:10:00	1
2012-12-31 04:38:00	2014-05-05 19:03:00	1
2012-12-31 04:38:00	2014-05-05 19:04:00	1
2012-12-31 04:38:00	2014-05-12 07:51:00	1
2012-12-31 04:38:00	2014-05-12 07:52:00	1
2012-12-31 04:38:00	2014-05-19 08:43:00	1
2012-12-31 04:38:00	2014-05-19 08:44:00	1
2012-12-31 04:38:00	2014-05-19 08:45:00	1
2012-12-31 04:38:00	2014-05-19 08:46:00	1
2012-12-31 04:38:00	2014-05-19 08:47:00	1
2012-12-31 04:38:00	2014-07-28 06:09:00	1
2012-12-31 04:38:00	2014-07-28 06:10:00	1
2012-12-31 04:38:00	2014-07-28 09:31:00	1
2012-12-31 04:38:00	2014-07-28 09:32:00	1

Result 5

Weekly Engagement Per Device:

Objective: 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
SELECT
    DATE_ADD(occurred_at, INTERVAL -WEEKDAY(occurred_at) DAY) AS week_start_date,
    device,
    COUNT(DISTINCT user_id) AS active_users_count
FROM events
GROUP BY week_start_date, device
ORDER BY week_start_date, device;
```

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

week_start_date	device	active_users_count
2014-04-28 00:14:00	macbook air	1
2014-04-28 00:15:00	macbook air	1
2014-04-28 00:16:00	macbook air	1
2014-04-28 00:17:00	iphone 5	1
2014-04-28 00:18:00	iphone 5	1
2014-04-28 00:19:00	iphone 5	1
2014-04-28 00:21:00	macbook air	1
2014-04-28 00:22:00	macbook air	1
2014-04-28 00:23:00	macbook air	1
2014-04-28 00:24:00	macbook air	1
2014-04-28 00:34:00	iphone 4s	1
2014-04-28 00:35:00	iphone 4s	1
2014-04-28 00:41:00	iphone 5s	1
2014-04-28 00:42:00	iphone 5s	1
2014-04-28 00:43:00	iphone 5s	1
2014-04-28 01:10:00	ipad air	1
2014-04-28 01:11:00	ipad air	1

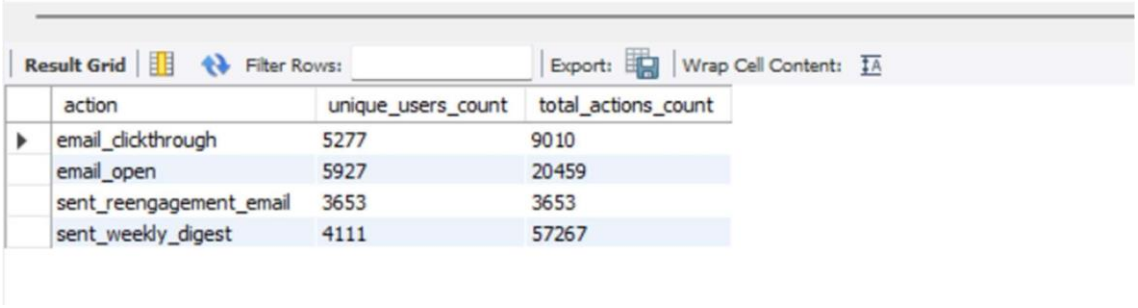
Result 7

Email Engagement Analysis:

Objective: Analyze how users are engaging with the email service.

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

```sql
SELECT
    action,
    COUNT(DISTINCT user_id) AS unique_users_count,
    COUNT(*) AS total_actions_count
FROM
    email_events
GROUP BY action
ORDER BY action;
```



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

action	unique_users_count	total_actions_count
email_clickthrough	5277	9010
email_open	5927	20459
sent_reengagement_email	3653	3653
sent_weekly_digest	4111	57267

Result

Contributions to Decision-Making: Provided actionable insights to various departments, aiding in strategic decision-making to rectify issues impacting user engagement, sales, and customer support.

Improved Operational Efficiency: The analysis facilitated a better understanding of operational bottlenecks, leading to targeted improvements and enhanced overall efficiency.

Drive Link

Case study 1:

Sql queries:
https://drive.google.com/file/d/1bAWlklY5OkW_fj99poDxnjBGhEmyIpsC/view?usp=sharin g

Report: https://drive.google.com/file/d/1ZDarTm9H-zM2DWUDDtQEDaaneXPf-2Ej/view?usp=drive_link

Case study 2:

Sql queries:
https://drive.google.com/file/d/1ruvkbhqtHwlmoKiUno_aWpqYRT5DUrqb/view?usp=drive _link

Report:
https://drive.google.com/file/d/1PN8WrG6JnwlkXmyHmFg1Teq34xB03vHY/view?usp=driv e_link

Hiring Process Analytics

Project Description

The project aimed to analyze the hiring process data of the company to derive meaningful insights that could optimize the hiring process. The objectives were to understand gender distribution among hires, determine the average salary, visualize salary distribution, departmental analysis, and analyze position tiers.

Approach

- Data Cleaning: Checked for missing values, handled them using appropriate strategies, and identified and addressed outliers.
- Calculations and Visualizations: Employed Excel functions for calculating averages, creating class intervals for salary distribution, and visualizing data using charts and graphs.
- Analysis: Conducted statistical analysis and visual representations to derive insights into hiring trends, salary distribution, departmental proportions, and position tiers.

Tech-Stack Used

Software: Microsoft Excel

Purpose: Excel was chosen for its robust data analysis features, statistical functions, and visualization tools, making it suitable for handling and analyzing the dataset.

Insights

- Gender Distribution: Identified that the company hired X males and Y females, showcasing a gender distribution within the hiring process.
- Average Salary: The average salary offered by the company was calculated to be.
- Salary Distribution: Created class intervals to understand the spread of salaries within the organization, indicating the ranges in which most salaries fell.
- Departmental Analysis: Visualized the proportion of employees in different departments, highlighting the workforce distribution across departments.
- Position Tier Analysis: Explored the distribution of position tiers within the company, providing insights into the hierarchical structure.

Data Analytics Tasks:

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

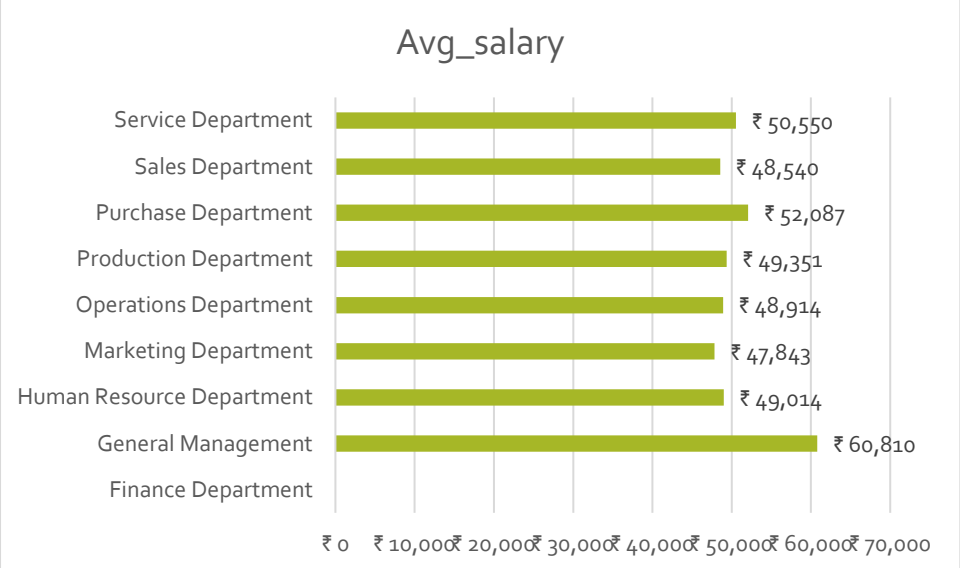
Your Task: Determine the gender distribution of hires. How many males and females have been hired by the company?

Male	2821
Female	1876

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.

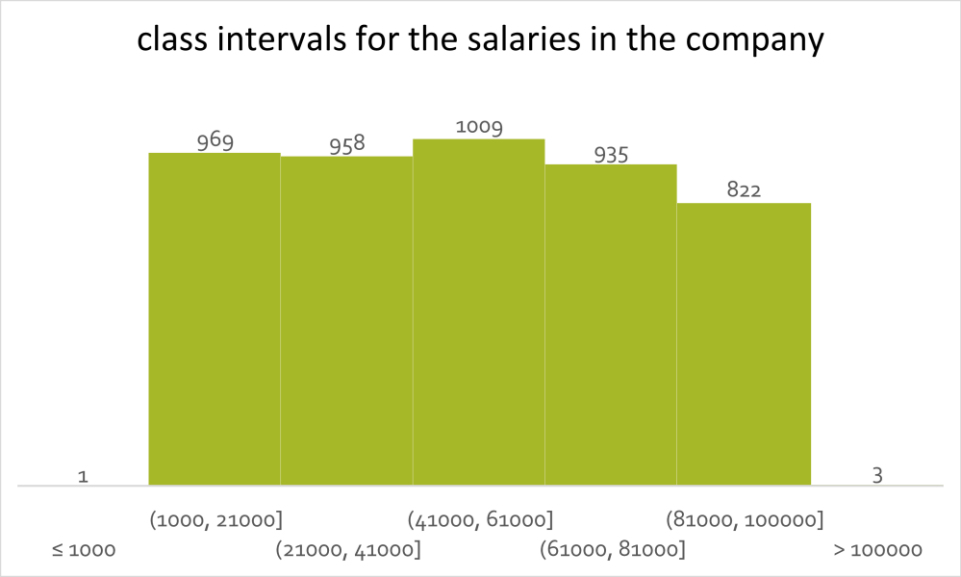
Your Task: What is the average salary offered by this company? Use Excel functions to calculate this.

- average salary offered by this company= **₹ 49,983**
- average salary offered by different Department in this company



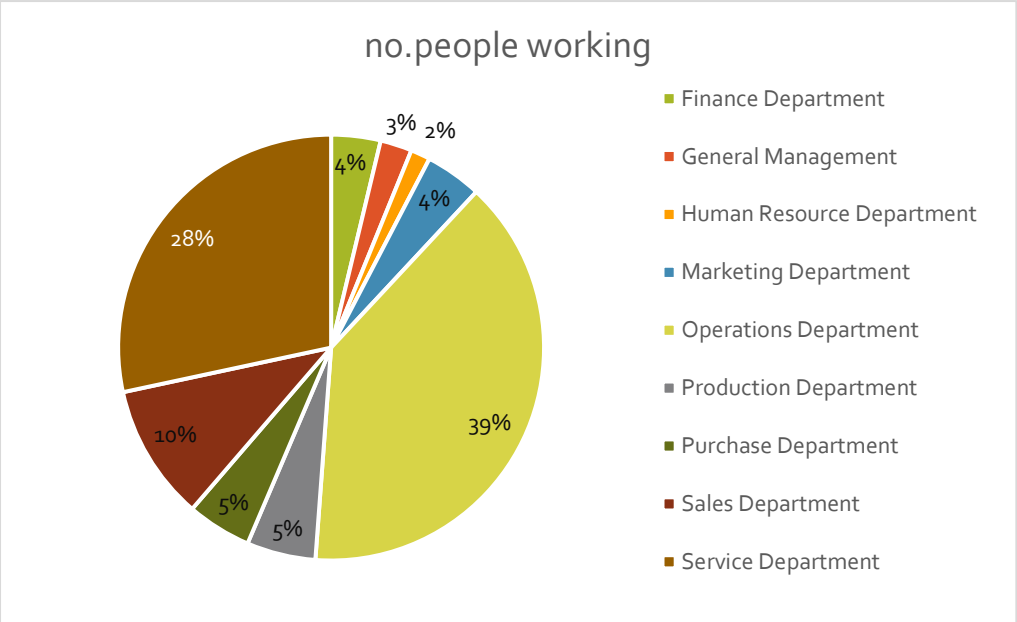
C. Salary Distribution: Class intervals represent ranges of values, in this case, salary ranges. The class interval is the difference between the upper and lower limits of a class.

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



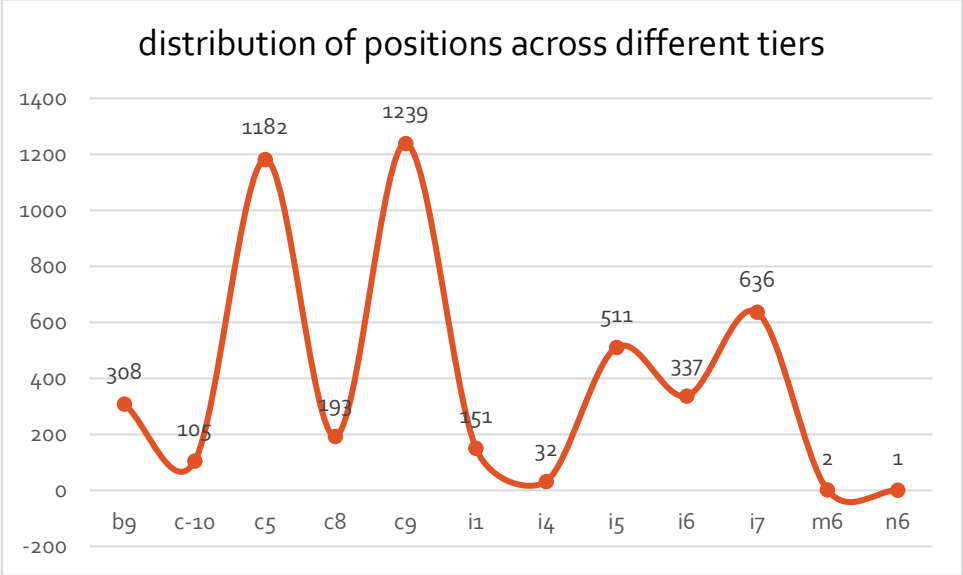
D. Departmental Analysis: Visualizing data through charts and plots is a crucial part of data analysis.

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

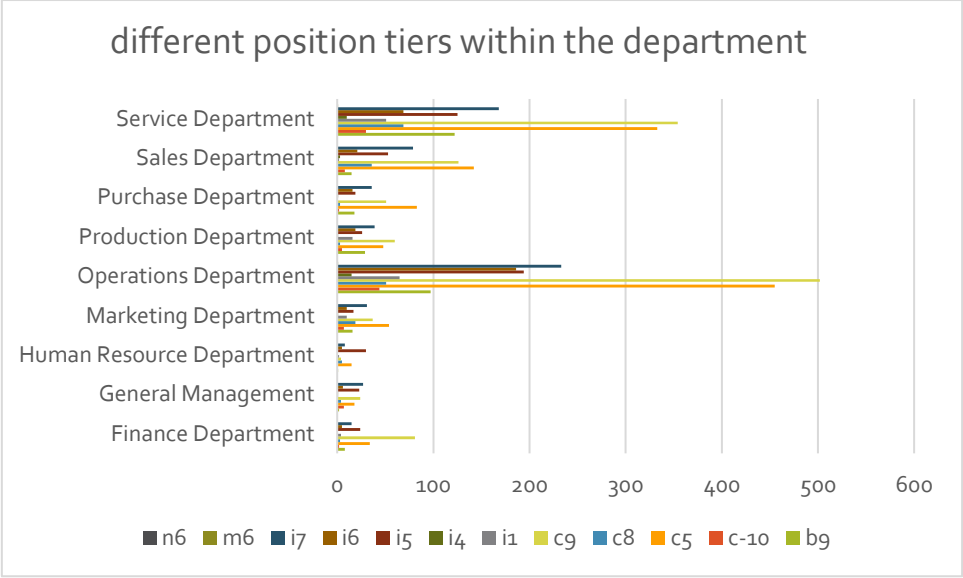


E. Position Tier Analysis: Different positions within a company often have different tiers or levels.

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.



the different position tiers within the department



Result

Through this project, a comprehensive analysis of the hiring process data was conducted, providing valuable insights into gender distribution, salary metrics,

departmental composition, and position tiers. These insights could potentially aid in refining the hiring strategy and decision-making processes of the company.

Drive Link:

Excel Sheet link: https://docs.google.com/spreadsheets/d/1uFyCKUQZ-Wm2vcKI6NjjZyjHZbUfrgKW/edit?usp=sharing&ouid=118439998565682353976&rtpof=true&sd=true

IMDB Movie Analysis

Description:

Problem Statement: 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.

Project Description

The project aimed to investigate the factors influencing the success of movies on IMDb, focusing on determining what attributes correlate with higher IMDb ratings. By analyzing IMDb movie data, the objective was to provide insights to aid movie producers, directors, and investors in making informed decisions for future projects.

Approach

- Data Collection: Describe how the IMDb movie dataset was obtained and its characteristics.
- Data Cleaning: Detail the steps taken for data preprocessing, including handling missing values, duplicates, data type conversion, and feature engineering.
- Data Analysis Techniques: Explain the methods used to explore relationships between variables (correlation analysis, statistical modeling, etc.).

Tech-Stack Used:

Microsoft excel (2021)

Data Analytics Tasks:

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.

Hint: Use Excel's COUNTIF function to count the number of movies for each genre. You might need to manipulate the 'genres' column to separate multiple genres for a single movie. Use Excel's functions like AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV to calculate descriptive statistics. Compare the statistics to understand the impact of genre on movie ratings.

- Top 10 Common Genres of movies

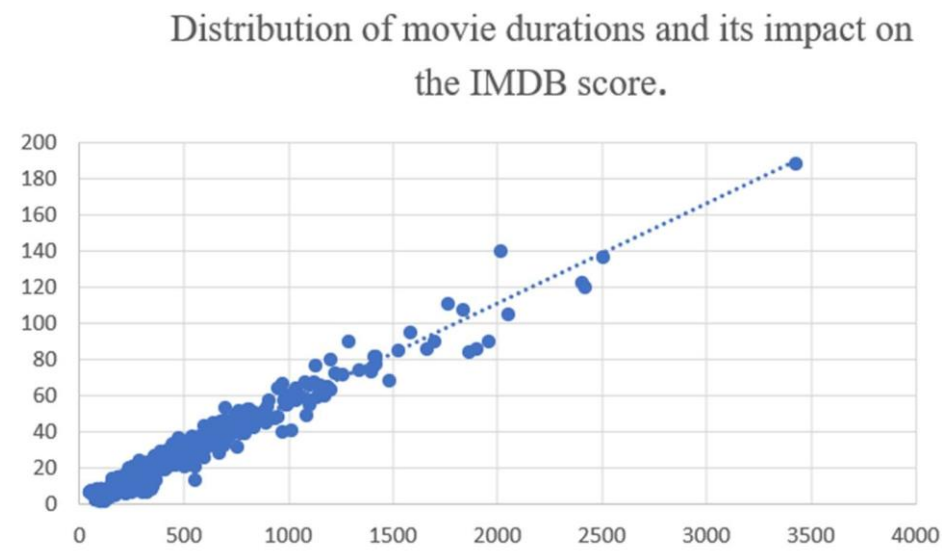
GENERE	NUMBER OF MOVIES
Comedy|Drama|Romanc	147
Comedy	144
Drama	144
Comedy|Drama	139
Comedy|Romance	132
Drama|Romance	117
Crime|Drama|Thriller	82
Action|Crime|Thriller	57
Action|Crime|Drama|Thr	50
Action|Adventure|Sci-Fi	48

GENERE	NUMBER OF MOVIE	Mean	Median	Range	Mode	Variance	Standard deviation
Comedy|Drama|Romance	147	6.486395	6.5	8	6.5	0.572416	0.754004227
Comedy	144	5.824306	6	8	6.2	1.597377	1.259477765
Drama	144	7.080556	7.2	8.8	7.3	0.690389	0.828006121
Comedy|Drama	139	6.560432	6.7	8.8	6.7	0.767771	0.87306775
Comedy|Romance	132	5.929545	6	8.4	6.1	0.703624	0.835639715
Drama|Romance	117	6.981197	7.1	8.1	6.7	0.54654	0.736117283
Crime|Drama|Thriller	82	6.859756	7	8.5	6.1	0.612311	0.777717071
Action|Crime|Thriller	57	6.35614	6.5	7.6	6.5	0.579292	0.754406364
Action|Crime|Drama|Thriller	50	6.498	6.5	9	6.1	0.514078	0.709785883
Action|Adventure|Sci-Fi	48	6.652083	6.8	8.4	6.6	1.541698	1.228649256
Comedy|Crime	47	6.065957	6.1	8.3	6.7	1.461859	1.196142131
Horror	47	5.808511	5.8	8	5.9	0.977317	0.978020099
Action|Adventure|Thriller	45	6.748889	6.8	8	6.8	0.595737	0.763216067
Drama|Thriller	43	6.665116	6.8	8.5	7	0.740897	0.850686104
Crime|Drama|Mystery|Thriller	42	6.938095	6.7	8.6	6.6	0.599489	0.764993738
Crime|Drama	41	7.47561	7.5	9.3	7.5	0.78789	0.876740217
Horror|Thriller	36	5.797222	5.9	7.9	5.9	1.217992	1.088190678
Action|Adventure|Sci-Fi|Thriller	34	6.367647	6.15	8.8	6.4	0.767103	0.862868195
Horror|Mystery|Thriller	33	5.860606	5.7	8.5	4.8	1.079962	1.023345492
Drama|Mystery|Thriller	30	6.756667	6.85	8.4	7.5	0.888057	0.926528767
Biography|Drama	29	7.22069	7.3	8.2	7.3	0.319557	0.55546148
Action|Comedy|Crime	27	5.937037	6.1	7.3	6.6	0.932422	0.947569268
Adventure|Animation|Comedy|Family|Fantasy	27	6.42963	6.8	8.3	7.3	1.814473	1.321843498
Horror|Mystery	26	5.807692	6.05	7.2	6.2	0.821538	0.888786154
Action|Adventure|Fantasy	25	6.356	6.4	8.3	5.8	0.9909	0.975327637
Biography|Drama|Sport	23	7.304348	7.3	8.3	7.6	0.336798	0.567587006
Action|Thriller	22	6.340909	6.4	8.5	6.5	1.077771	1.014288416
Drama|Sport	22	7.054545	6.9	8.2	6.8	0.400693	0.61844914
Action|Comedy|Crime|Thriller	21	6.161905	6.2	7.6	6.6	0.370476	0.59399871
Adventure|Animation|Comedy|Family	21	6.57619	6.6	8.3	6.7	0.707905	0.82109379
Biography|Drama|History	21	7.2	7.3	8.9	7.5	0.543	0.719126454
Action|Crime|Drama|Mystery|Thriller	19	6.336842	6.5	7.6	6	0.811345	0.876722681

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.

Hint: Calculate descriptive statistics such as mean, median, and standard deviation for movie durations. Use Excel's functions like AVERAGE, MEDIAN, and STDEV. Create a scatter plot to visualize the relationship between movie duration and IMDB score. Add a trendline to assess the direction and strength of the relationship.



director_Name	sum_duration	sum_ibdm_score	Mean	Median	stdv
Adam McKay	715	41.5	119.1667	285	14.85953
Adam Shankman	850	47.7	106.25	163	13.08386
Adrian Lyne	450	25.6	112.5	213	10.96586
Alan J. Pakula	252	12.6	126	79	15
Alan Parker	410	21.1	136.6667	317	6.236096
Alan Taylor	238	13.7	119	230	7
Albert Hughes	459	28	114.75	117	10.37726
Alejandro AmenÃ¡bar	367	22.9	122.3333	448	16.43844
Alejandro G. IÃ±Ã¡rritu	657	39.2	131.4	0	15.60256
Alex Kendrick	362	20.2	120.6667	589	7.408704
Alex Proyas	571	34.1	114.2	295	9.579144
Alexander Payne	584	37.1	116.8	729	8.352245
Alexandre Aja	397	24.9	99.25	192	10.84839
Alfonso CuarÃ³n	448	31.2	112	0	18.61451
Alfred Hitchcock	238	16.7	119	13000	11
Amy Heckerling	287	17.2	95.66667	143	2.624669
Anand Tucker	192	13.3	96	14	4
Andrew Adamson	483	28.6	120.75	80	29.26922
Andrew Bergman	212	9.6	106	31	11
Andrew Davis	462	26	115.5	99	9.233093
Andrew Dominik	257	13.7	128.5	181	31.5
Andrew Fleming	386	24.6	96.5	26	3.640055
Andrew Niccol	462	28	115.5	487	8.13941
Andrew Stanton	330	23.2	110	475	15.57776
Andrey Konchalovskiy	207	10.7	103.5	96	6.5
Andrzej Bartkowiak	526	26.3	105.2	43	7.44043
Andy Fickman	617	34.6	102.8333	99	5.273097
Andy Tennant	717	37.5	119.5	72	13.51234
Ang Lee	1035	58	129.375	0	11.01065

C. Language Analysis: Situation: 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.

Hint: Use Excel's COUNTIF function to count the number of movies for each language.

Calculate the mean, median, and standard deviation of the IMDB scores for each language. Compare the statistics to understand the impact of language on movie ratings.

Language	count_language	Lan_mean	lan_media	lan_stdv
English	3657	6.414876	6.5	1.067351
French	34	7.355882	7.3	0.511739
Spanish	23	7.082609	7.2	0.841661
Mandarin	15	7.08	7.4	0.745833
German	10	7.77	7.8	0.675352
Japanese	10	7.66	8	0.939361
Cantonese	7	7.342857	7.3	0.324509
Italian	7	7.185714	7	1.069618
Hindi	5	7.22	7.4	0.716659
Korean	5	7.7	7.7	0.509902
Portuguese	5	7.76	8	0.875443
Norwegian	4	7.15	7.3	0.497494
Danish	3	7.9	8.1	0.432049
Dutch	3	7.566667	7.8	0.329983
Persian	3	8.133333	8.4	0.449691
Thai	3	6.633333	6.6	0.368179
Aboriginal	2	6.95	6.95	0.55
Dari	2	7.5	7.5	0.1
Indonesian	2	7.9	7.9	0.3
Arabic	1	7.2	7.2	0
Aramaic	1	7.1	7.1	0
Bosnian	1	4.3	4.3	0
Czech	1	7.4	7.4	0

Language	count_language
English	3657
French	34
Spanish	23
Mandarin	15
German	10
Japanese	10
Cantonese	7
Italian	7
Hindi	5
Korean	5
Portuguese	5

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.

Hint: Calculate the average IMDB score for each director. Use Excel's PERCENTILE function to identify the directors with the highest scores. Compare the scores of these directors to the overall distribution of scores.

PERCENTIE
7.5

director_Name	Average of imdb_score
Akira Kurosawa	8.7
Charles Chaplin	8.6
Michael Curtiz	8.6
Tony Kaye	8.6
Damien Chazelle	8.5
Majid Majidi	8.5
Ron Fricke	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Asghar Farhadi	8.4
Richard Marquand	8.4

director_Name	Average of imdb_score	PERCENTIE
Akira Kurosawa	8.7	7.5
Charles Chaplin	8.6	
Michael Curtiz	8.6	
Tony Kaye	8.6	
Damien Chazelle	8.5	
Majid Majidi	8.5	
Ron Fricke	8.5	
Sergio Leone	8.433333333	
Christopher Nolan	8.425	
Asghar Farhadi	8.4	
Richard Marquand	8.4	
Alfred Hitchcock	8.35	
Billy Wilder	8.3	
Fritz Lang	8.3	
Lee Unkrich	8.3	
Lenny Abrahamson	8.3	
Pete Docter	8.233333333	
Hayao Miyazaki	8.225	
Quentin Tarantino	8.2	
Elia Kazan	8.2	
George Roy Hill	8.2	

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.

Hint: Calculate the correlation coefficient between movie budgets and gross earnings using Excel's CORREL function. Calculate the profit margin (gross earnings - budget) for each movie and identify the movies with the highest profit margin using Excel's MAX function.

- Movie with the highest profit margin is Avatar

AvatarÂ	237000000	760505847	523505847	0.099496423

movie_title	budget	gross	Profit_margin	Correlation
AvatarÂ	237000000	760505847	523505847	0.099496423
Jurassic WorldÂ	150000000	652177271	502177271	0.099496423
TitanicÂ	200000000	658672302	458672302	0.099496423
Star Wars: Episode IV - A New HopeÂ	11000000	460935665	449935665	0.099496423
E.T. the Extra-TerrestrialÂ	10500000	434949459	424449459	0.099496423
The AvengersÂ	220000000	623279547	403279547	0.099496423
The AvengersÂ	220000000	623279547	403279547	0.099496423
The Lion KingÂ	45000000	422783777	377783777	0.099496423
Star Wars: Episode I - The Phantom MenaceÂ	115000000	474544677	359544677	0.099496423
The Dark KnightÂ	185000000	533316061	348316061	0.099496423

movie_title	budget	gross	Profit_margin	Correlation
AvatarÂ	237000000	760505847	523505847	0.099496423
Jurassic WorldÂ	150000000	652177271	502177271	0.099496423
TitanicÂ	200000000	658672302	458672302	0.099496423
Star Wars: Episode IV - A New HopeÂ	11000000	460935665	449935665	0.099496423
E.T. the Extra-TerrestrialÂ	10500000	434949459	424449459	0.099496423
The AvengersÂ	220000000	623279547	403279547	0.099496423
The AvengersÂ	220000000	623279547	403279547	0.099496423
The Lion KingÂ	45000000	422783777	377783777	0.099496423
Star Wars: Episode I - The Phantom MenaceÂ	115000000	474544677	359544677	0.099496423
The Dark KnightÂ	185000000	533316061	348316061	0.099496423
The Hunger GamesÂ	78000000	407999255	329999255	0.099496423
DeadpoolÂ	58000000	363024263	305024263	0.099496423
The Hunger Games: Catching FireÂ	130000000	424645577	294645577	0.099496423
Jurassic ParkÂ	63000000	356784000	293784000	0.099496423
Despicable Me 2Â	76000000	368049635	292049635	0.099496423
American SniperÂ	58800000	350123553	291323553	0.099496423
Finding NemoÂ	94000000	380838870	286838870	0.099496423
Shrek 2Â	150000000	436471036	286471036	0.099496423
The Lord of the Rings: The Return of the KingÂ	94000000	377019252	283019252	0.099496423
Star Wars: Episode VI - Return of the JediÂ	32500000	309125409	276625409	0.099496423
Forrest GumpÂ	55000000	329691196	274691196	0.099496423
Star Wars: Episode V - The Empire Strikes BackÂ	18000000	290158751	272158751	0.099496423
Home AloneÂ	18000000	285761243	267761243	0.099496423
Star Wars: Episode III - Revenge of the SithÂ	113000000	380262555	267262555	0.099496423
Spider-ManÂ	139000000	403706375	264706375	0.099496423
MinionsÂ	74000000	336029560	262029560	0.099496423

Insights

- Summarize major observations, such as correlations between IMDb ratings and variables like genre, director, budget, actors, release year, etc.
- Highlight any significant trends or patterns discovered during analysis.
- Discuss insights obtained from the 'Five Whys' approach and their implications.

Result

- Describe the insights gained and their potential impact on decision-making for movie stakeholders.
- Explain how the project contributes to understanding the factors influencing movie success on IMDb.

Drive Link

Excel link: https://drive.google.com/file/d/1JsYhlcBSrf8f0G5taOYzf1eAoar3HHQ-/view?usp=sharing

Bank Loan Case Study

Project Description:

The project involves analyzing a loan application dataset to understand patterns influencing loan default, aiming to aid decision-making on loan approvals. The objective is to identify key factors behind loan default to make better decisions about loan approval, denial, or adjustment of terms.

Approach:

Utilizing Microsoft Excel 2022 for data analysis, I followed a structured approach to address the tasks outlined, leveraging Excel's functions and features for data cleaning, outlier detection, imbalance analysis, and various exploratory analyses.

Tech-Stack Used:

Software: Microsoft Excel 2021

Purpose: Excel was chosen for its robust functionalities in data manipulation, analysis, and visualization, making it suitable for this EDA task.

Insights:

1. Missing Data Handling:

- Identified missing values using functions like COUNT, ISBLANK, and IF.
- Employed appropriate imputation techniques (e.g., median imputation) for missing numerical values.
- Visualized missing data proportions using a bar/column chart.

2. Outlier Detection:

- Detected outliers using Excel's statistical functions (QUARTILE, IQR) and conditional formatting.
- Utilized box plots or scatter plots for visualizing and identifying outliers in numerical variables.

3. Data Imbalance Analysis:

- Determined class frequencies using COUNTIF and SUM functions to assess data imbalance.
- Visualized class distributions via pie or bar charts, highlighting any imbalance.

4. Univariate, Segmented Univariate, and Bivariate Analysis:

- Conducted univariate analysis to understand variable distributions.
- Utilized filters, sorting, and pivot tables for segmented and bivariate analysis.
- Visualized distributions and relationships using histograms, bar charts, box plots, and scatter plots.

5. Correlation Analysis:

- Segmented dataset based on scenarios to identify top correlations for each segment.
- Calculated correlation coefficients using Excel's CORREL function.
- Visualized correlations with matrices or heatmaps to showcase strong indicators of loan default.

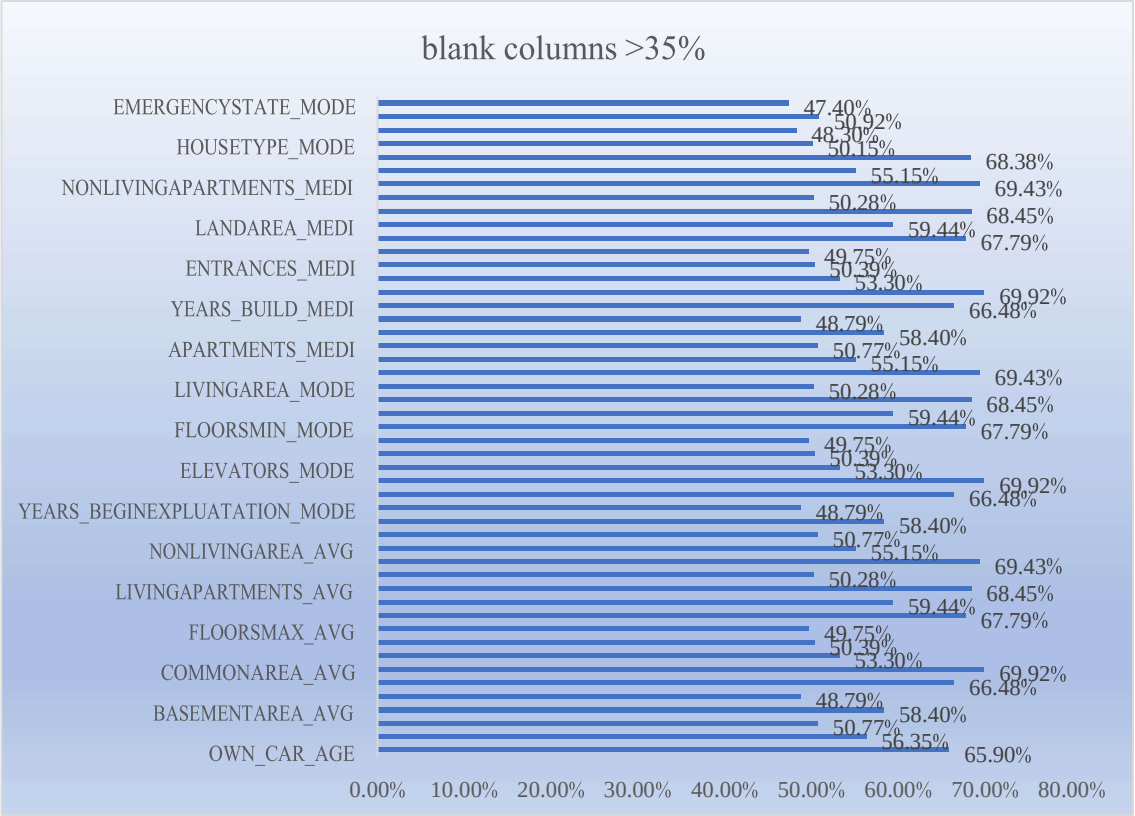
Data Analytics Tasks:

A. Identify Missing Data and Deal with it Appropriately:

1. Identify Missing Data:
    - In cell A2, enter the header "Missing Percentage."
    - In cell A3, use the formula `=(COUNTBLANK(B:B)/COUNT(B:B))*100` and drag it right for other columns.
    - use conditional formatting to highlight cells with a percentage greater than 35 of missing values.
2. Handling Missing Data:
    - If you choose to remove rows with missing values:
    - In a new column, use the `FILTER` or `IF` function to exclude rows with missing values.
    - If you choose to impute missing values:
    - For numerical columns, use functions like `AVERAGE` or `MEDIAN` to fill in missing values.
3. Visualization:

- Create a bar chart:
- Select the range B2:DS3.
- Go to the 'Insert' tab and choose 'Bar Chart'.

This visualization will provide a quick overview of the proportion of missing values for each variable, helping you make informed decisions on handling missing data.



B) Identify Outliers in the Dataset:

1. Detecting Outliers:

- Use the `QUARTILE` function to calculate the first (Q1) and third (Q3) quartiles for each numerical variable.
- Calculate the Interquartile Range (IQR) using the formula `IQR = Q3 - Q1`.
- Identify potential outliers as values beyond the range
  Lower Bound= [Q1 - 1.5 * IQR]
  Higher Bound= [Q3 + 1.5 * IQR]
- Use conditional formatting to highlight cells containing potential outliers.

QUARTILE 1	17348.63
QUARTILE 2(median)	26145

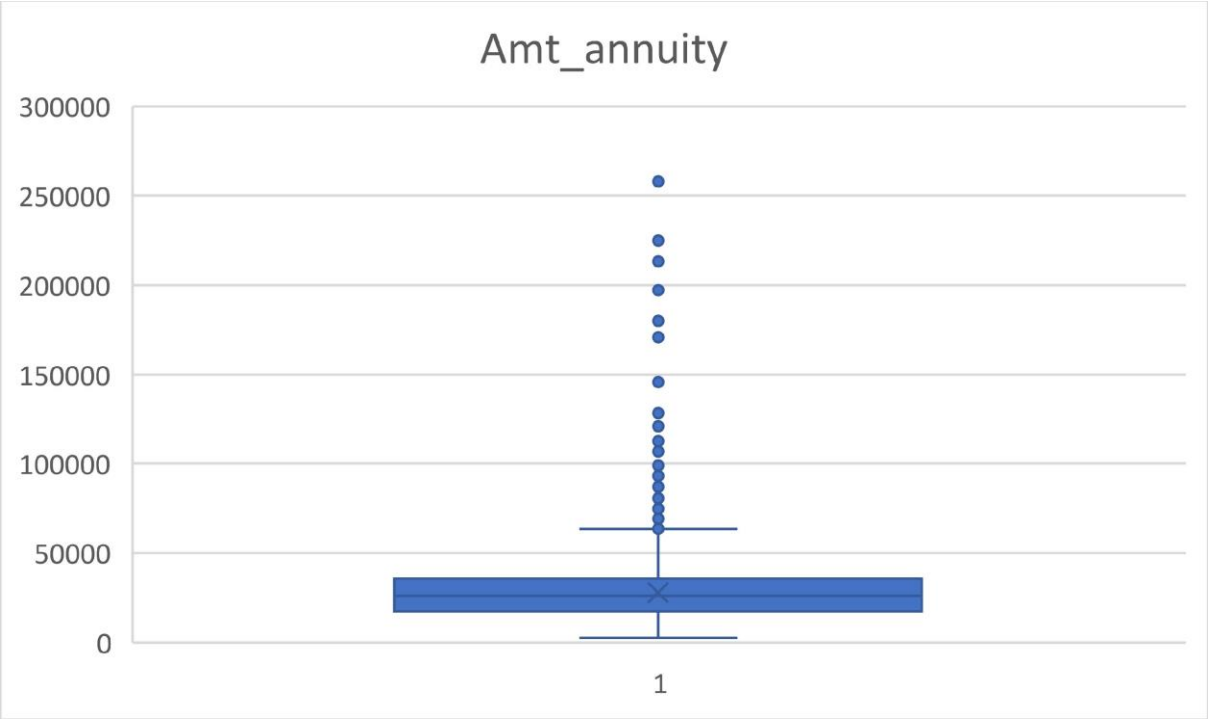
QUARTILE 3	35814.38
Interquartile Range (IQR):	18465.75
Lower Bound:	-10350
Upper Bound:	35814.38

2. Validating Outliers:

- Consider applying business rules or thresholds to determine if the identified outliers are valid or require further investigation.
- You may choose to exclude extreme values if they are not consistent with the expected data range.

3. Visualization:

- Create box plots or scatter plots to visualize the distribution of numerical variables and highlight the outliers.
- Box plots provide a visual representation of the quartiles and help identify outliers.
- Scatter plots allow you to visually inspect individual data points and their positions relative to the expected range.

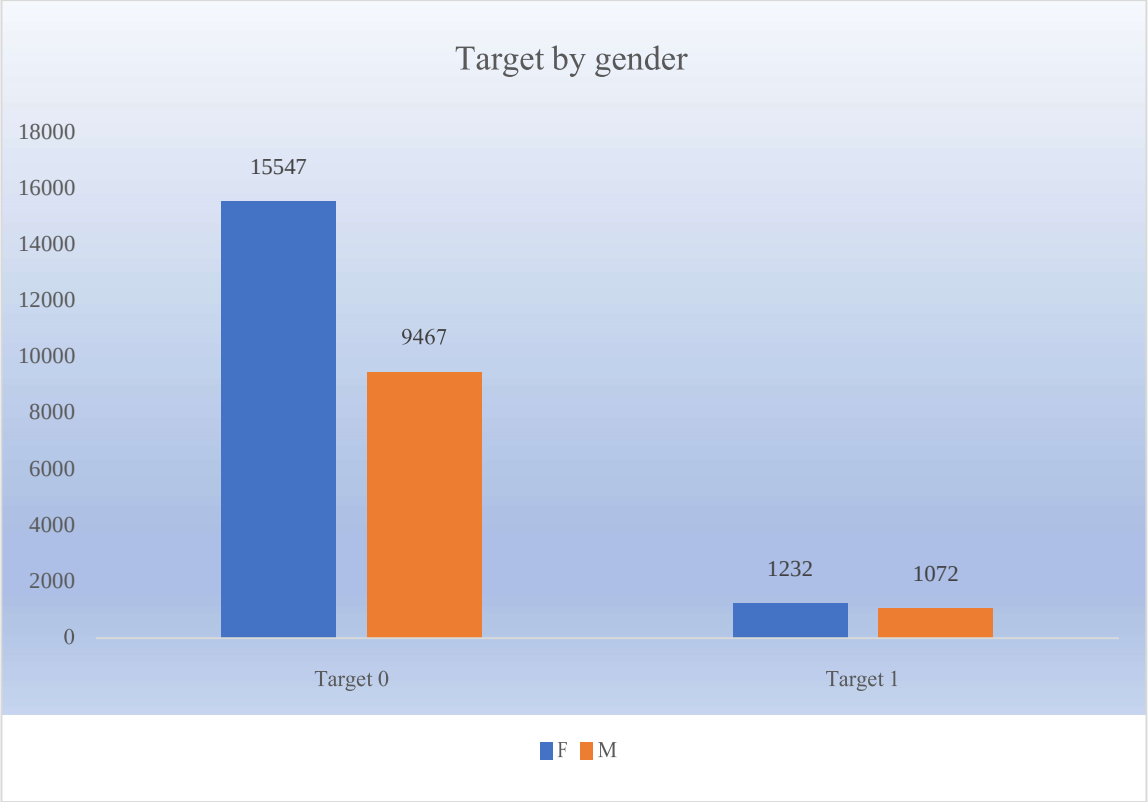


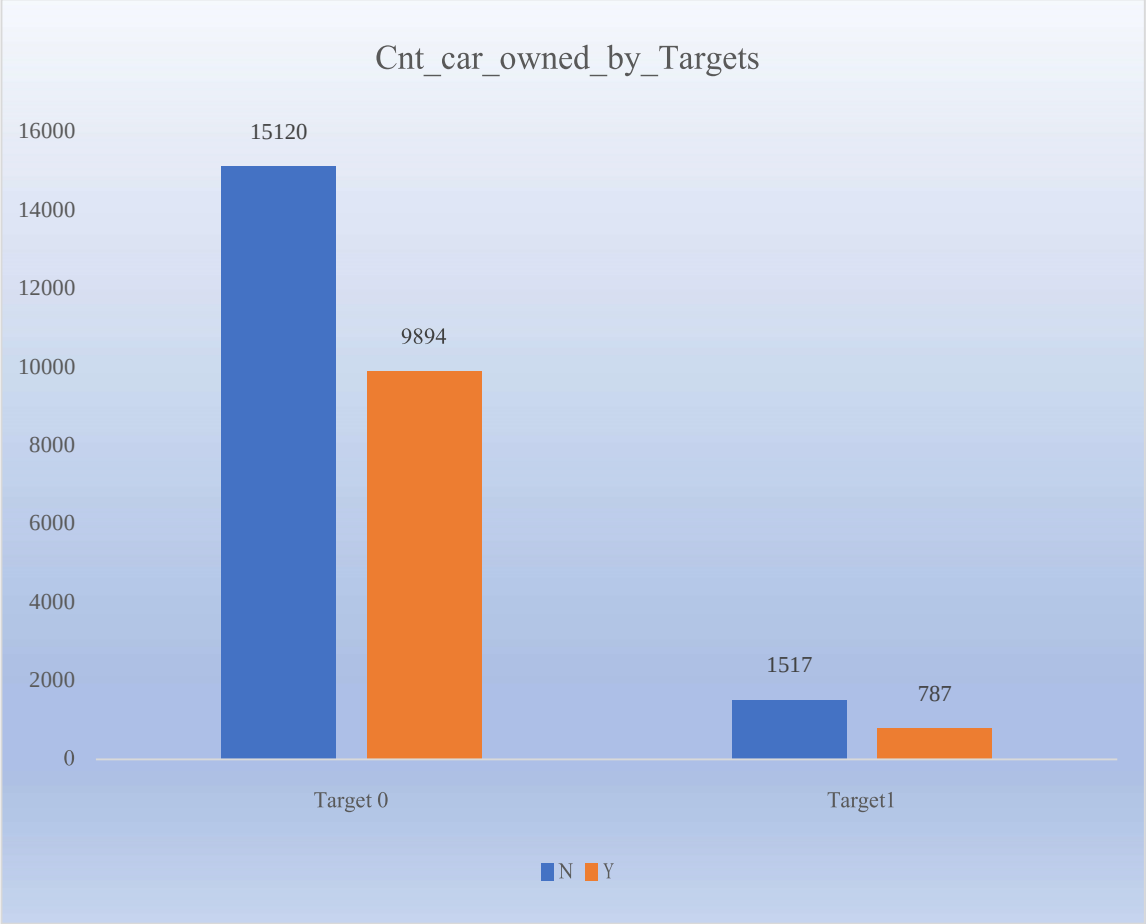
These visualizations will help you better understand the distribution of numerical variables and identify potential outliers for further investigation.

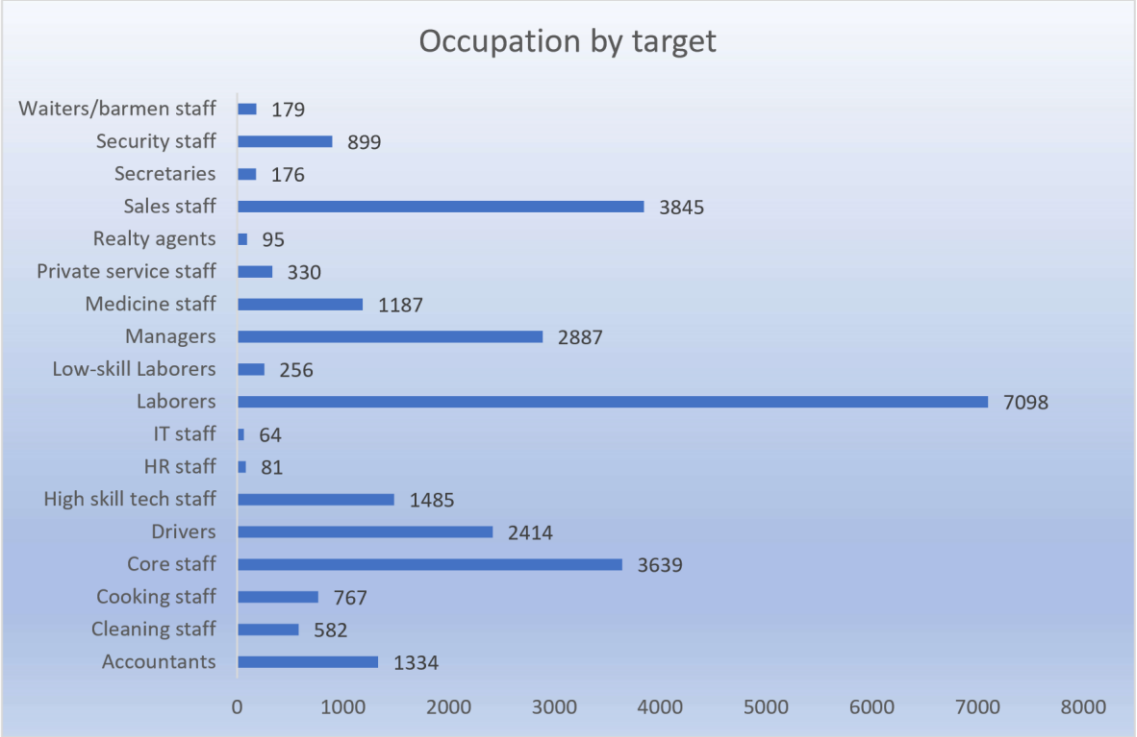
C. Analyze Data Imbalance:

1. Use the `COUNTIF` function to count the occurrences of each class in the target variable. Let's assume your target variable is in column A:
2. Calculate the proportions of each class by dividing the class frequency by the total number of samples:
3. Compare the class proportions to assess data imbalance. If the proportions are significantly different, there may be an imbalance issue.
4. Create a pie chart or bar chart to visually represent the distribution of the target variable.
   Highlight the class imbalance using different colors or annotations.
5. Select the class frequencies and proportions, including labels.
6. Go to the "Insert" tab and select either "Pie Chart" or "Doughnut Chart" from the Chart options.
7. Customize the chart to make it visually informative, such as adding data labels or a legend.

By following these steps, you can determine if there is data imbalance in your loan application dataset and create a visual representation of the class distribution. Adjust the formulas and chart options based on your specific dataset and requirements.


Target by gender
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
15547
9467
1232
1072
Target 0
Target 1
F
M

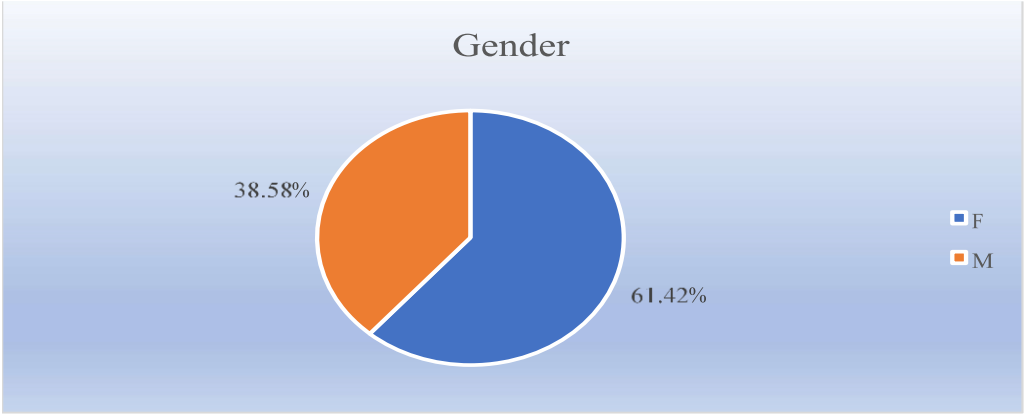

Cnt_car_owned_by_Targets
16000
14000
12000
10000
8000
6000
4000
2000
0
15120
9894
1517
787
Target 0
Target1
N
Y

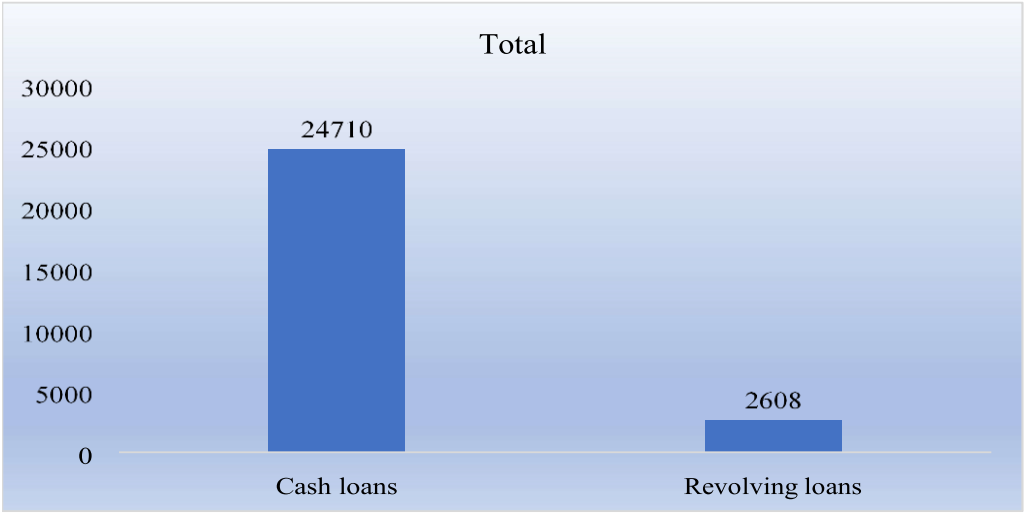


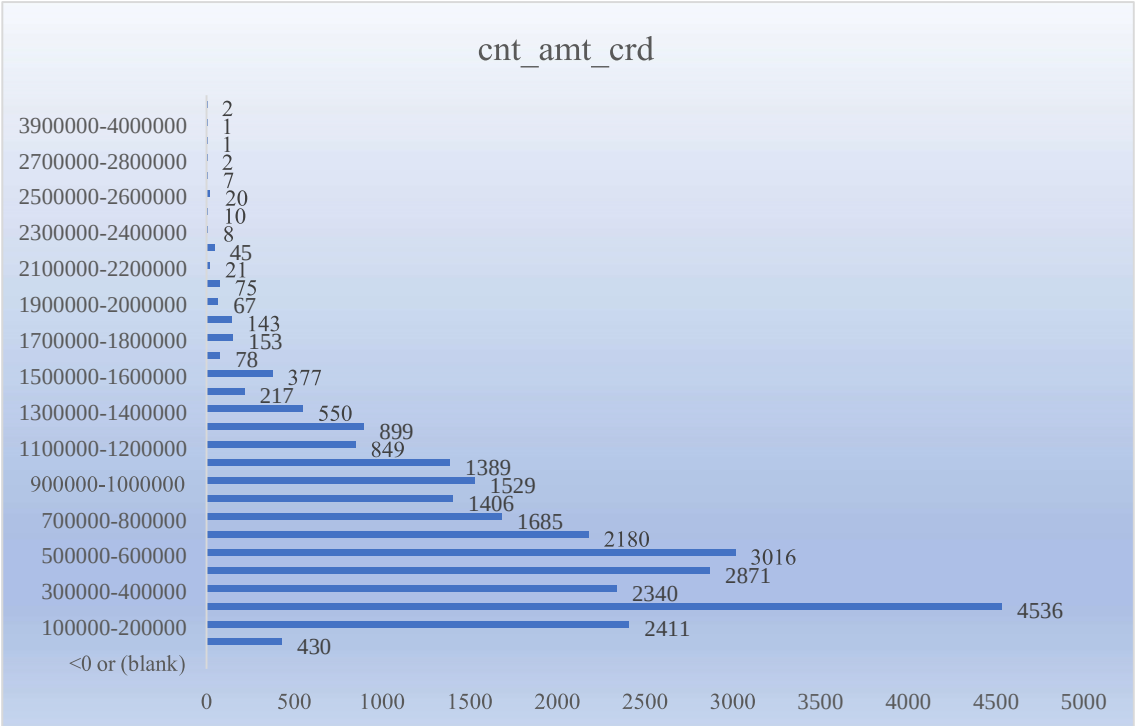
D.Perform Univariate, Segmented Univariate, and Bivariate Analysis:

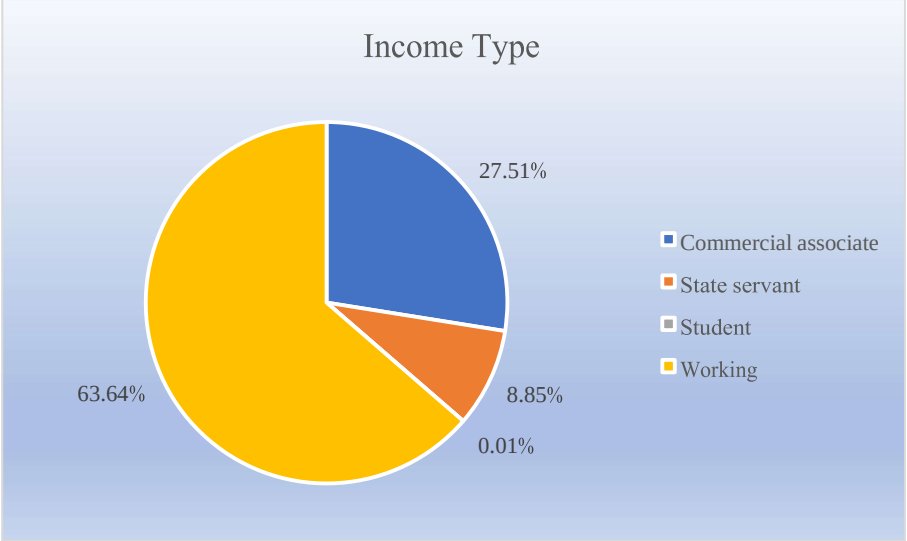
Univariate Analysis:

- List the variables you want to analyze (e.g., income, loan amount, credit score).
- Use functions like `COUNT`, `AVERAGE`, `MEDIAN`, `MIN`, `MAX`, and
  `STDEV` to calculate descriptive statistics for each variable.
- Visualize the distribution of each variable using histograms or box plots. You can use the "Insert" tab and select the appropriate chart type.
- Determine the criteria for segmenting the data (e.g., age groups, income brackets).
- Use Excel filters to segment the data based on your criteria.
- Repeat the univariate analysis steps for each segment separately.
- Create stacked bar charts or grouped bar charts to compare variable distributions across different segments. This can be done by selecting the relevant data and using the "Insert" tab to create the desired chart.





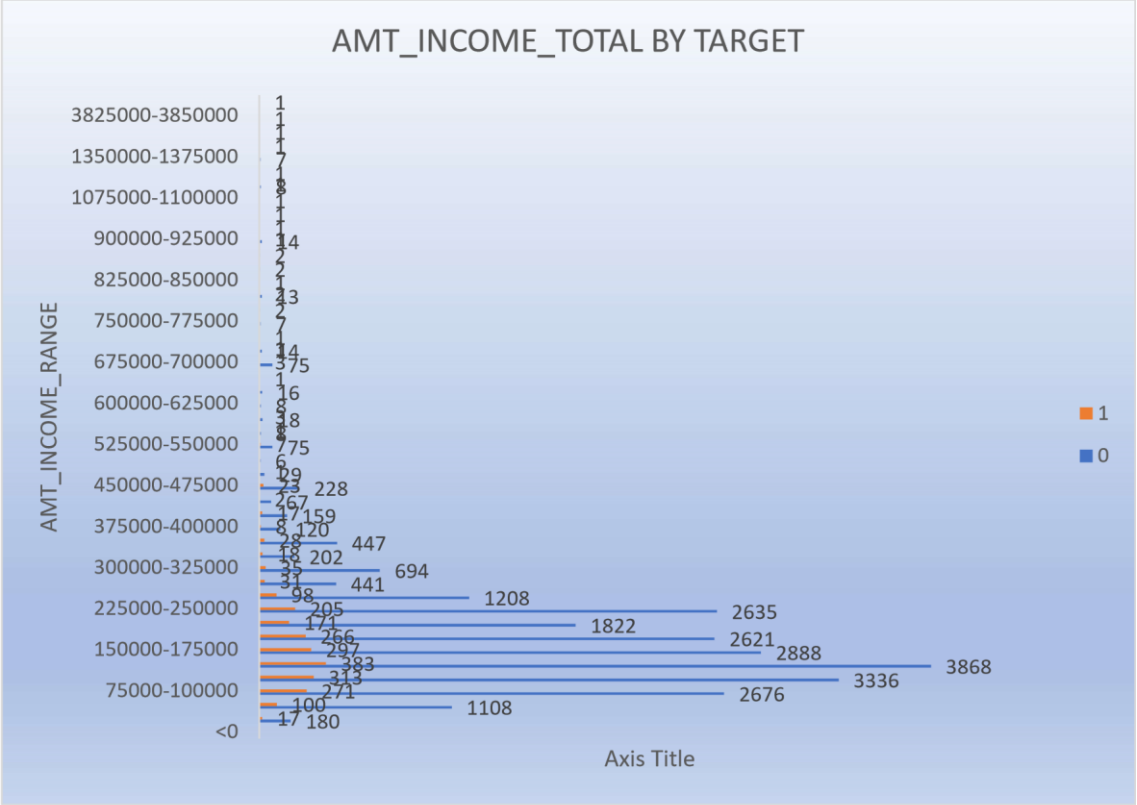


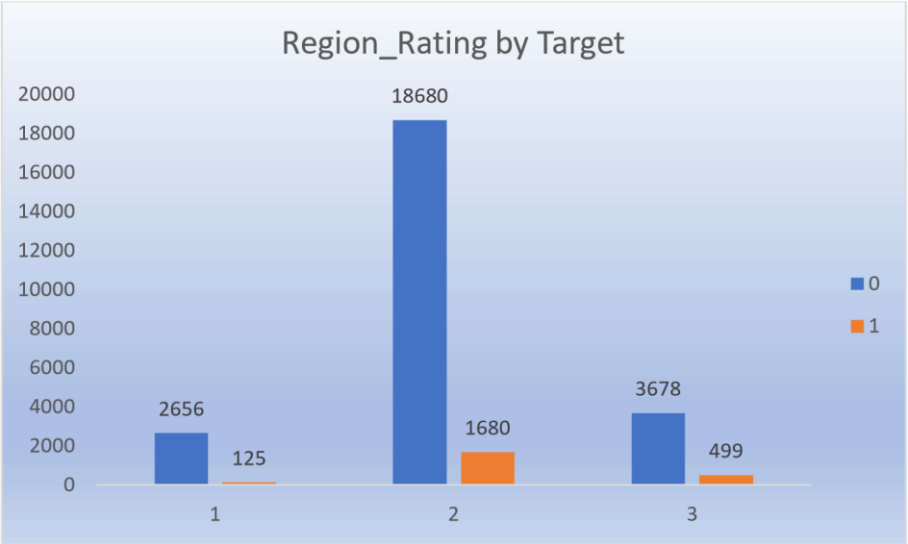


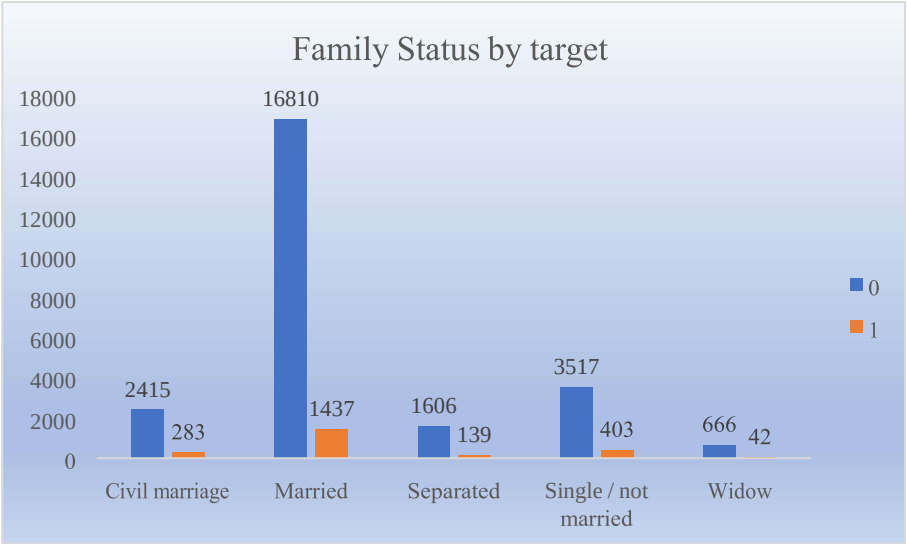
Bivariate Analysis:

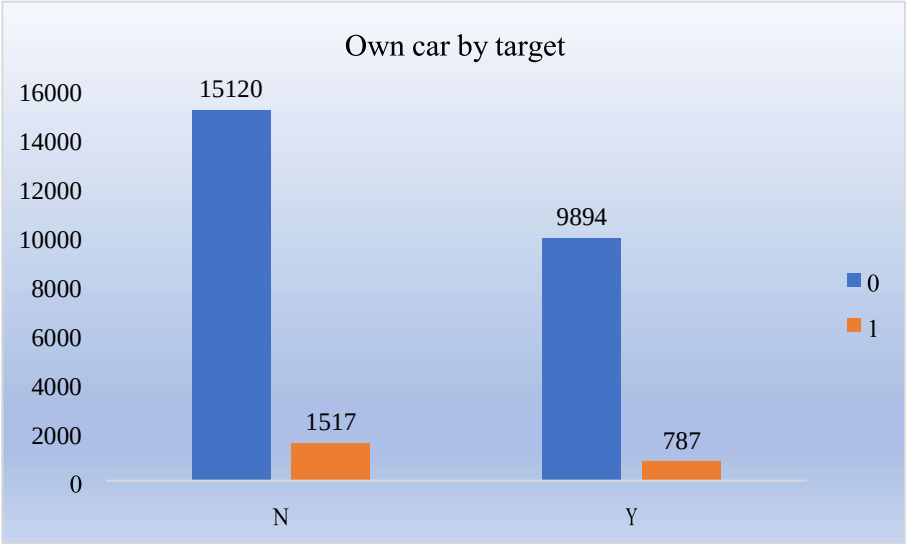
- Determine the relationships you want to explore (e.g., relationship between income and loan default).
- Use scatter plots to visualize the relationship between two continuous variables. Select the data and use the "Insert" tab to create a scatter plot.
- For categorical variables, use heatmaps to visualize the correlation. You can create a heatmap by using conditional formatting on a correlation matrix.
- Use the `COUNTIF` function or pivot tables to create cross-tabulations that show how two categorical variables relate to each other.

- Create graphical representations like stacked bar charts or grouped bar charts to compare the distribution of the target variable across different levels of another variable.
- Utilize Excel's pivot tables for dynamic and interactive analysis, especially in segmented and bivariate analyses.









By following these steps, you can perform a comprehensive analysis of consumer and loan attributes, gaining insights into the driving factors of loan default. Adjust the steps based on your specific dataset and analysis goals.

E. Identify Top Correlations for Different Scenarios:

1. Identify the scenarios for segmentation (e.g., clients with payment difficulties and all other cases).
2. Use Excel filters or other methods to segment the data based on the identified scenarios.
3. Identify the target variable for loan default (e.g., 1 for default, 0 for non-default).
4. Use the `CORREL` function to calculate correlation coefficients between each variable and the target variable.

5. Create correlation matrices or heatmaps to visualize the correlations between variables within each segment:Highlight the top correlated variables using different colors or shading for better emphasis.
6. Select the range of correlation coefficients for the variables within each segment.
7. Go to the "Insert" tab and choose "Heatmap" or use conditional formatting to create a visually appealing correlation heatmap.
8. Adjust the color scale to make strong correlations stand out.

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH
CNT_CHILDREN	1	0.00952242	0.012408771	0.016697551	0.00458965	-0.017700172	-0.1790284	-0.028866522	-0.140729592	0.10657788
AMT_INCOME_TOTAL	0.00952242	1	0.011070999	0.011559155	0.008663754	-0.012250841	-0.0055071	-0.010187961	0.016877994	0.013724345
AMT_CREDIT	0.012408771	0.011070999	1	0.736843501	0.981090827	0.048988455	0.18628417	0.07853943	0.037521632	0.039868297
AMT_ANNUITY	0.016697551	0.011559155	0.736843501	1	0.74027537	0.05037331	0.0820702	0.033755336	-0.014784462	0.059835163
AMT_GOODS_PRICE	0.00458965	0.008663754	0.981090827	0.74027537	1	0.057351504	0.17661649	0.085359746	0.033589001	0.044063135
REGION_POPULATION_RELATIVE	-0.017700172	-0.012250841	0.048988455	0.05037331	0.057351504	1	0.02517085	0.011165554	0.058260731	0.020898197
DAYS_BIRTH	-0.179028358	-0.00550713	0.186284168	0.082070203	0.176616493	0.025170845	1	0.305038457	0.240980889	0.118073716
DAYS_EMPLOYED	-0.028866522	-0.010187961	0.07853943	0.033755336	0.085359746	0.011165554	0.30503846	1	0.141461758	0.109653148
DAYS_REGISTRATION	-0.140729592	0.016877994	0.037521632	-0.014784462	0.033589001	0.058260731	0.24098089	0.141461758	1	0.01983706
DAYS_ID_PUBLISH	0.10657788	0.013724345	0.039868297	0.059835163	0.044063135	0.020898197	0.11807372	0.109653148	0.01983706	1

a. Target 1

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH_years	DAYS_EMPLOYED_year	DAYS_ID_PUBLISH_year	REGION_RATING_CLIENT
CNT_CHILDREN	1	-0.004911747	-0.01601349	-0.00448395	-0.019907888	-0.026197114	-0.2557575	-0.070314647	0.129553223	0.035395697
AMT_INCOME_TOTAL	-0.004911747	1	0.365279441	0.440204156	0.372806512	0.179846528	0.054125487	0.026995747	0.014928611	-0.20806829
AMT_CREDIT	-0.016013487	0.365279441	1	0.762470519	0.986611816	0.093268561	0.160451708	0.089589592	0.034060456	-0.107726011
AMT_ANNUITY	-0.00448395	0.440204156	0.762470519	1	0.766884412	0.110554518	0.101612037	0.054256785	0.024018313	-0.125911444
AMT_GOODS_PRICE	-0.019907888	0.372806512	0.986611816	0.766884412	1	0.096656435	0.15528024	0.090603251	0.034520555	-0.108551053
REGION_POPULATION_RELATIVE	-0.026197114	0.179846528	0.093268561	0.110554518	0.096656435	1	0.044620857	-0.010412732	0.000656732	-0.523154439
DAYS_BIRTH_years	-0.2557575	0.054125487	0.160451708	0.101612037	0.15528024	0.044620857	1	0.345553495	0.072472675	-0.045952464
DAYS_EMPLOYED_year	-0.070314647	0.026995747	0.089589592	0.054256785	0.090603251	-0.010412732	0.345553495	1	0.064585448	0.017966232
DAYS_ID_PUBLISH_year	0.129553223	0.014928611	0.034060456	0.024018313	0.034520555	0.000656732	0.072472675	0.064585448	1	-0.002768905
REGION_RATING_CLIENT	0.035395697	-0.20806829	-0.10772601	-0.125911444	-0.108551053	-0.523154439	-0.045952464	0.017966232	-0.002768905	1

b. Target 1

By following these steps, you can effectively identify and visualize the top correlations for different scenarios within your loan application dataset using Excel functions and features.
Adjust the steps based on your specific dataset and analysis goals.

Result:

The analysis has provided significant insights into the loan application dataset:

- Identification of missing data and appropriate handling methods.

- Detection and understanding of outliers and their potential impact on analysis.
- Assessment of data imbalance and its potential implications for classification models.
- Uncovering relationships between variables and loan default, highlighting key factors influencing default.

Drive Link:

https://docs.google.com/spreadsheets/d/18ogiujClVHCsqdIQ56u18lST_4FVsYPy/edit?usp=sharing& ouid=118439998565682353976&rtpof=true&sd=true

Analyzing the Impact of Car Features on Price and Profitability

Project Description:

Overview of the project and its purpose:

The project aims to analyze the impact of car features on pricing and profitability in the automotive industry. By leveraging a dataset containing information on over 11,000 car models, the objective is to provide insights that will help car manufacturers optimize pricing and product development decisions to maximize profitability while meeting consumer demand.

The primary business problem is how a car manufacturer can make informed decisions regarding pricing and product development to enhance profitability while aligning with consumer demand. This involves understanding the relationship between car features, market categories, and pricing.

The dataset, titled "Car Features and MSRP," was obtained from Kaggle and includes information on various car models, their specifications, and pricing. It consists of 16 variables, such as make, model, year, engine details, transmission type, market category, and manufacturer's suggested retail price (MSRP).

Description of data cleaning and pre-processing:

Before diving into the analysis, thorough data cleaning was performed to ensure accuracy and reliability of the results. Steps included handling missing values, removing duplicates, and addressing outliers. The dataset, last updated in 2017, was considered for historical trends, and any assumptions made during the analysis are documented.

Approach:

I thoroughly reviewed the Excel data provided by the Trainity Impact of Car Features project, focusing on the columns related to car features in the dataset. I gained a comprehensive understanding of each column and its respective constraints to facilitate the analysis. The project involved addressing a set of specific questions, and I utilized Microsoft Excel 2021 to solve these queries, leveraging Excel formulas for data manipulation.

The data cleaning process was crucial for ensuring the accuracy and reliability of the analysis. Key steps in data cleaning included:

- Removing null values to enhance data integrity.

- Eliminating columns deemed unnecessary for the analysis to streamline the dataset.

- Identifying and removing duplicate rows to prevent redundancy.

Before the data cleaning process, the dataset had 11,915 columns under the "Car data" category. After completing the cleaning steps, the dataset was refined to 11,098 columns, ensuring a more focused and relevant set of data for analysis. This process helped optimize the dataset for meaningful insights.

Tech-Stack Used:

The primary tool employed for the analysis was Microsoft Excel 2021. The use of Excel allowed for the application of various formulas and functionalities to derive insights and findings from the dataset.

Insights:

1. Market Category and Popularity: Luxury and Performance market categories tend to have higher popularity scores.

2. Engine Power and Price: There is a positive correlation between a car's engine power and its price.

3. Factors Influencing Price: Regression analysis identified significant variables affecting car prices, with a bar chart visually representing their relative importance.

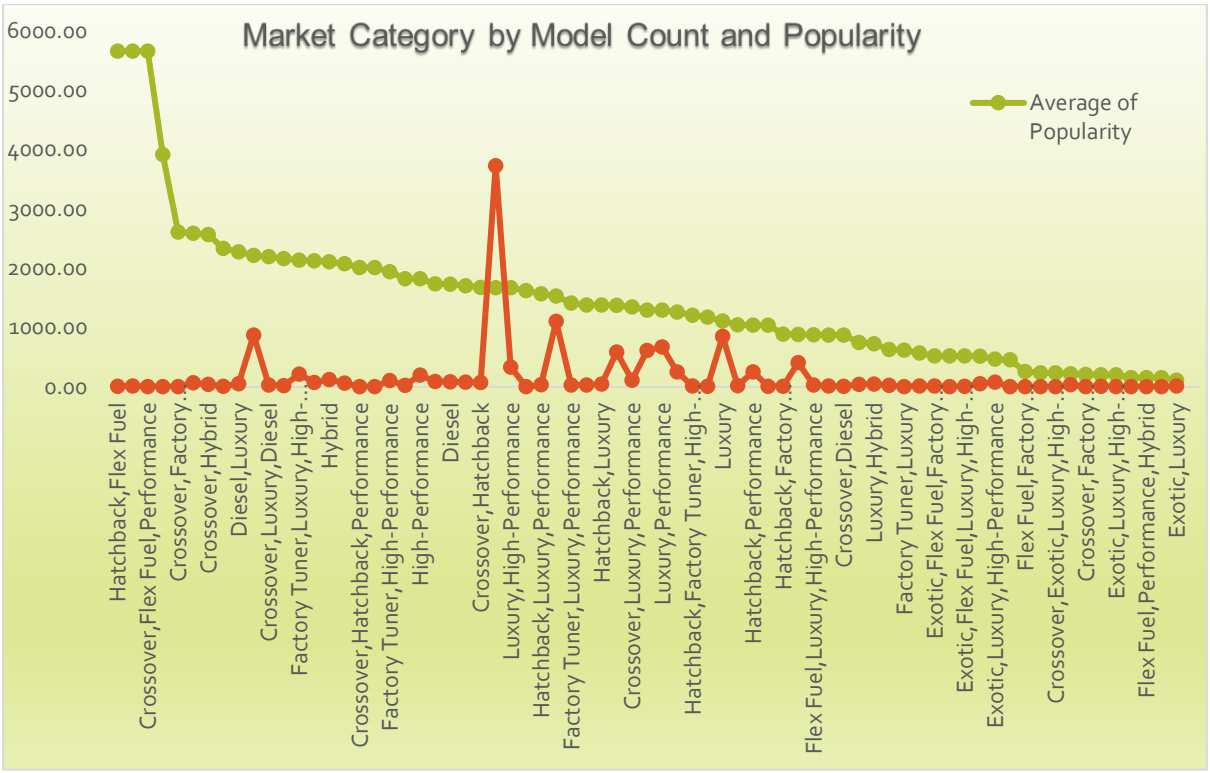
4. Manufacturer and Average Price: Certain manufacturers command higher average prices, providing insights into brand positioning.

5. Fuel Efficiency and Cylinder Count: Higher cylinder counts negatively impact highway MPG, and a scatter plot with a trendline visually represents this relationship.

Analytics Tasks:

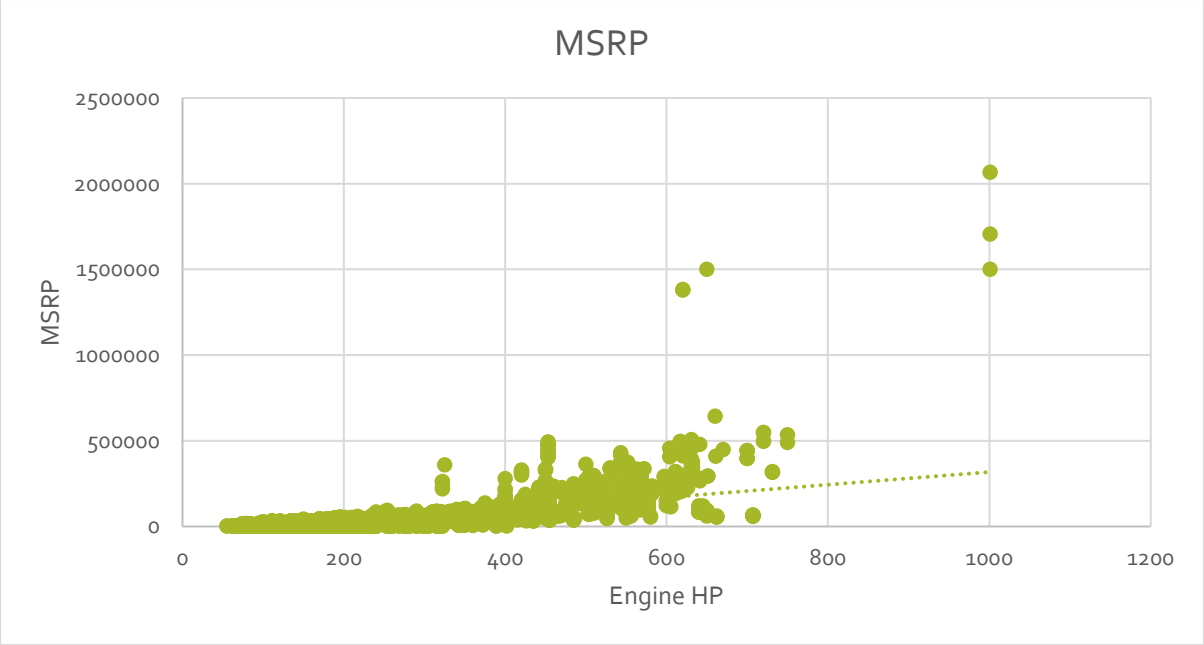
Task A: How does the popularity of a car model vary across different market categories?

Converted the columns Car Model, Popularity, and Market Categories into a pivot chart and used a line chart to illustrate the popularity and count of models across various Market Categories.



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

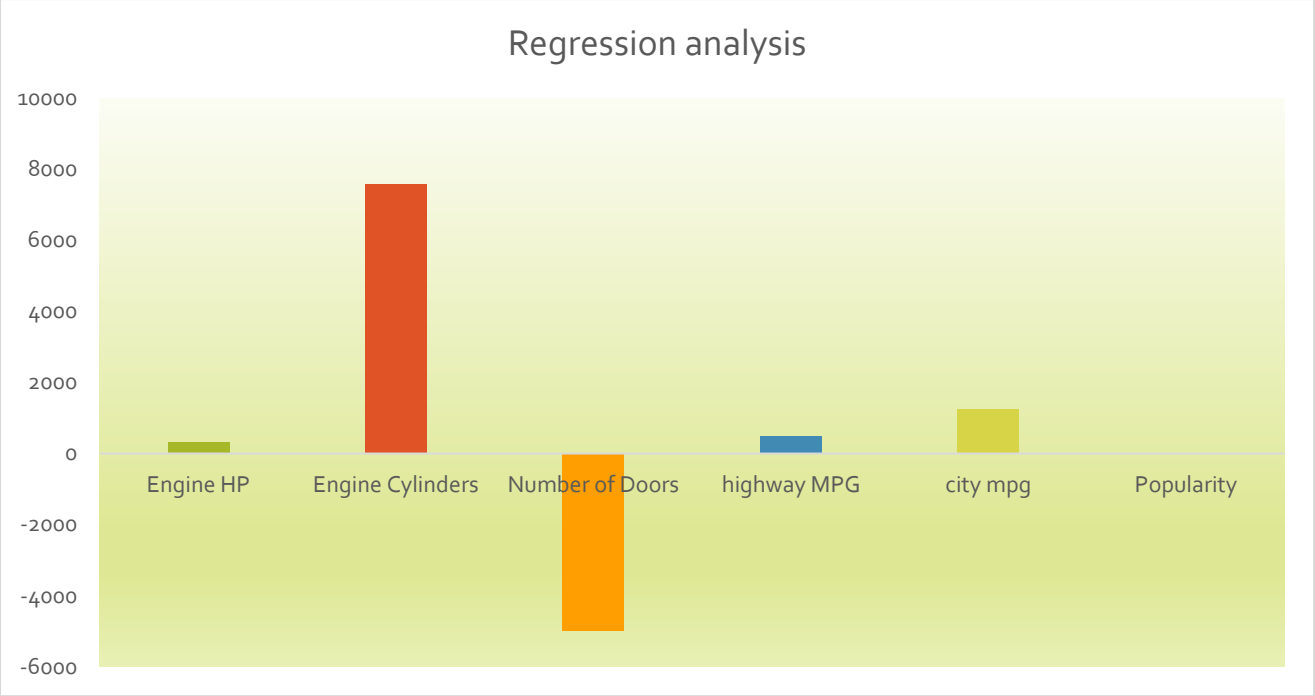
For this analysis, I considered the columns Engine HP and MSRP for comparison. I utilized a Scatter Plot chart to examine the relationship between them, and I also incorporated a Trendline to identify the trend. Upon reviewing the scatter plot, it is evident that with an increase in Engine HP, the price of the car also increases. The Trendline reinforces this observation.



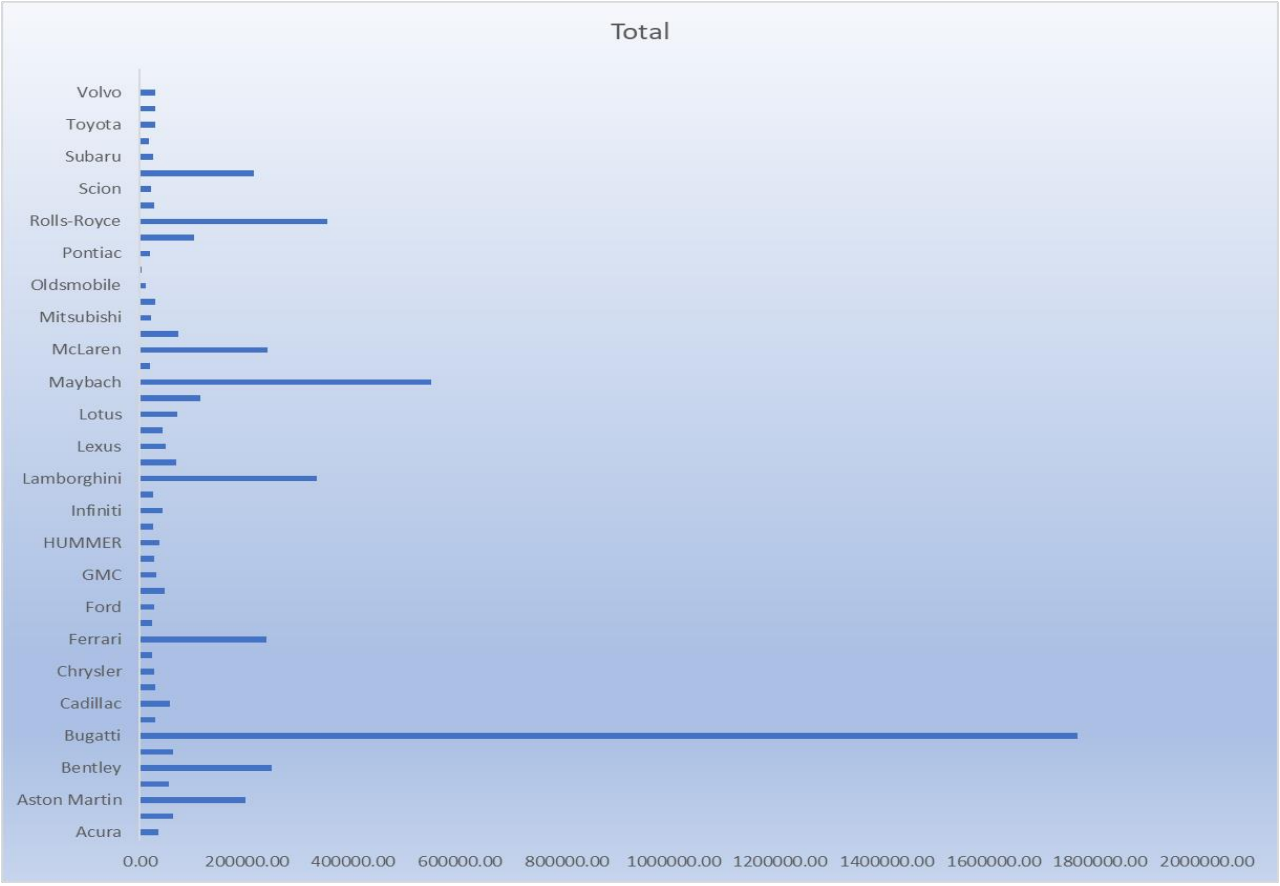
Task C: Which car features are most important in determining a car's price?

Utilizing regression analysis, I identified variables with the strongest relationship to a car's price. Subsequently, I created a bar chart illustrating the coefficient values for each variable, providing a visual representation of their relative importance. The analysis revealed that "Engine Cylinders" have a more pronounced correlation with Car Price, whereas the column "Number of Doors" showed the least correlation.

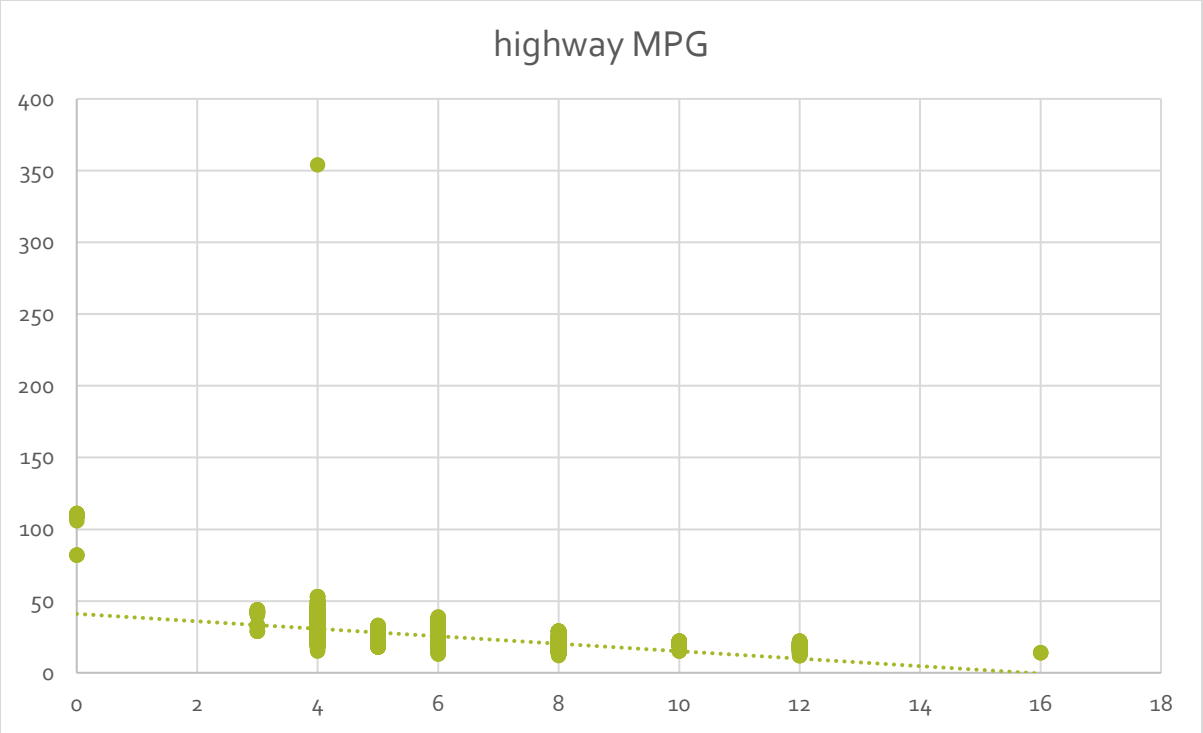
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.*
Intercept	-97167.87274	3898.078612	-24.92711985	1.7136E-133	-104808.8004	-89526.94512	-104808.8004	-89526.94
Engine HP	320.4942139	6.3774589	50.25421863	0	307.9932598	332.995168	307.9932598	332.995
Engine Cylinders	7578.79133	461.2602827	16.43061762	5.88653E-60	6674.639109	8482.94355	6674.639109	8482.94
Number of Doors	-4980.209981	496.4047724	-10.03255863	1.38198E-23	-5953.251655	-4007.168308	-5953.251655	-4007.168
highway MPG	503.5834871	109.2773107	4.608307836	4.10488E-06	289.3805157	717.7864585	289.3805157	717.7864
city mpg	1253.468123	125.6629389	9.974843293	2.46287E-23	1007.146405	1499.789841	1007.146405	1499.789
Popularity	-3.553387511	0.297352947	-11.95006655	1.02989E-32	-4.136252193	-2.97052283	-4.136252193	-2.97052



Task D: How does the average price of a car vary across different manufacturers? Considering the columns MSRP and Car Model, I created a bar chart to visualize the relationship between manufacturers and average prices. The analysis indicated that Bugatti has the highest average MSRP among other brands, while Plymouth has the least.

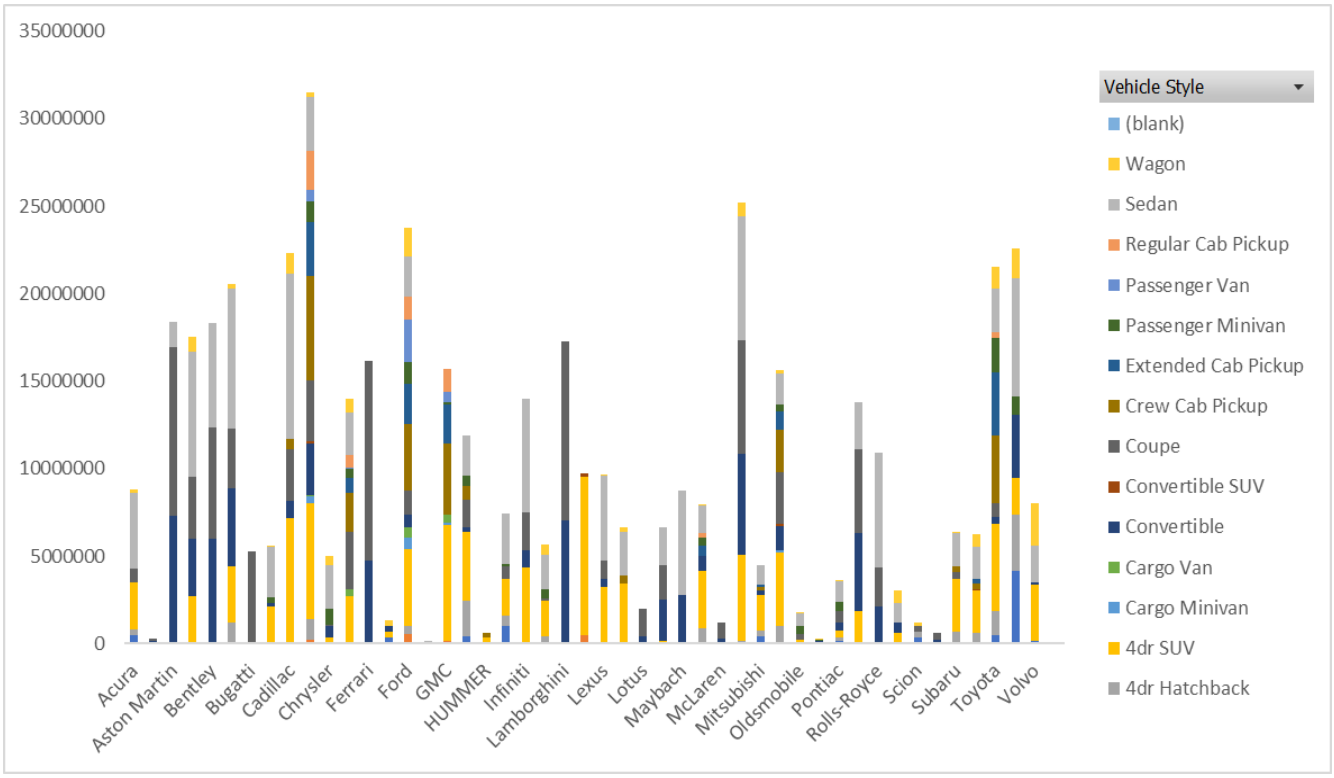


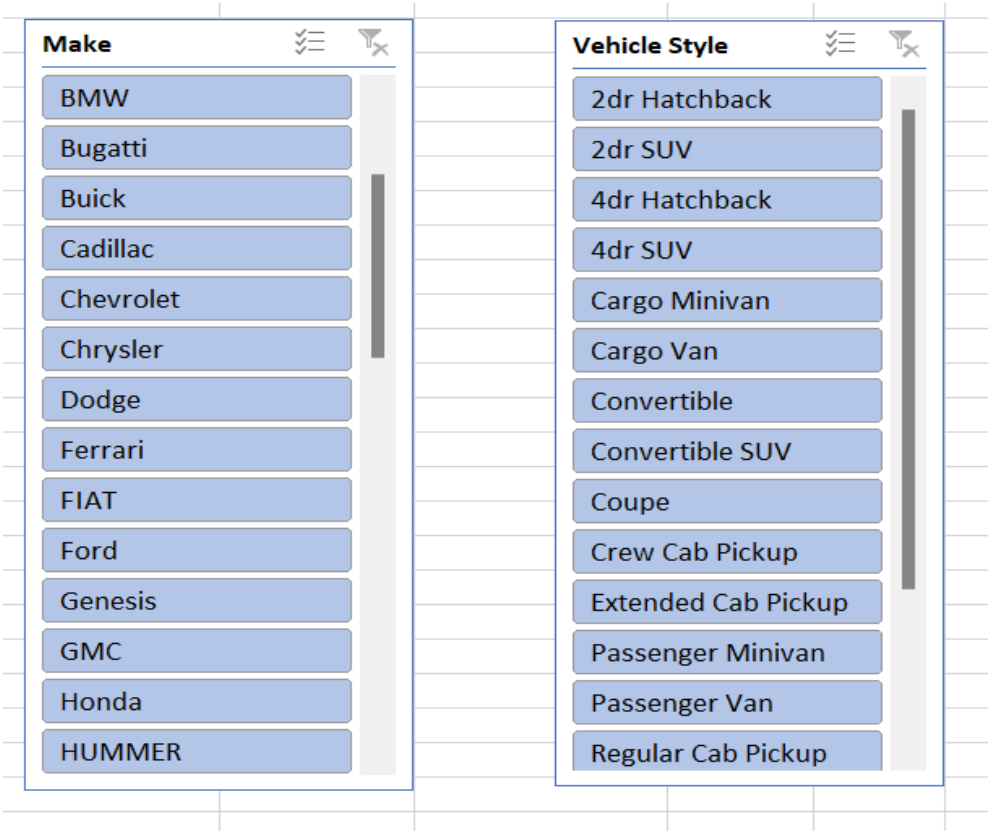
Task E: What is the relationship between fuel efficiency and the number of cylinders in a car's engine?
I constructed a scatter plot with the number of cylinders on the x-axis and highway MPG on the y-axis. A trendline was added to visually estimate the slope of the relationship. The analysis revealed a downward trend in the Trendline with an increase in Engine Cylinders. The correlation coefficient between the number of cylinders and highway MPG was calculated, resulting in a negative correlation of -0.6147.



Building the Dashboard:

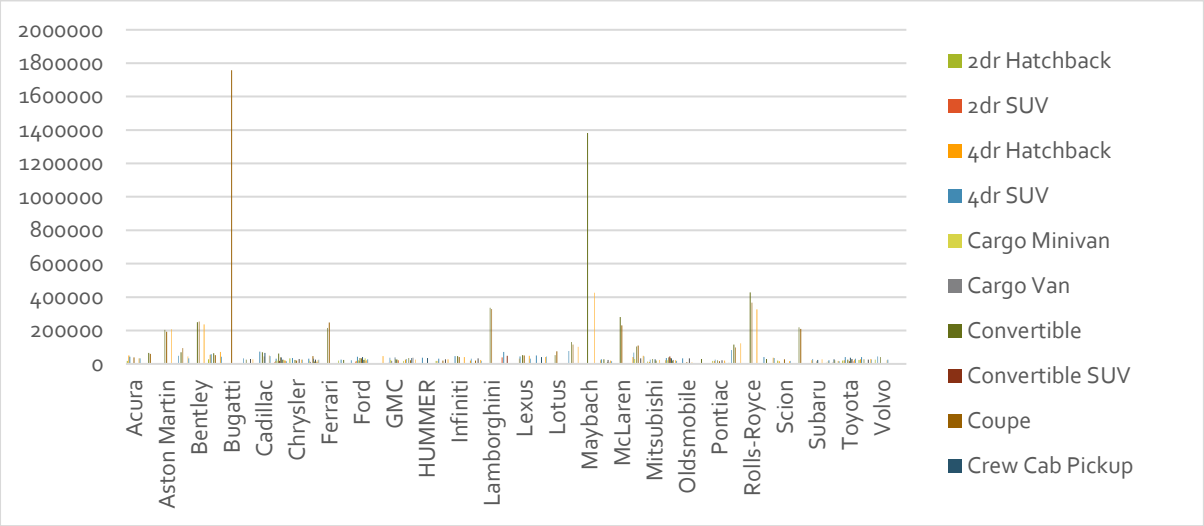
Task 1: How does the distribution of car prices vary by brand and body style?
I employed a Stacked Column chart to illustrate the distribution of car prices by brand and body style. Filters and slicers were utilized for interactivity, and the total MSRP for each brand and body style was calculated using Pivot Tables.

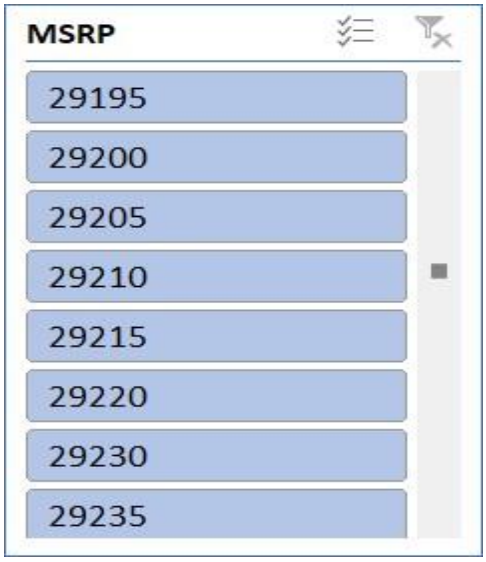




Task 2: Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?

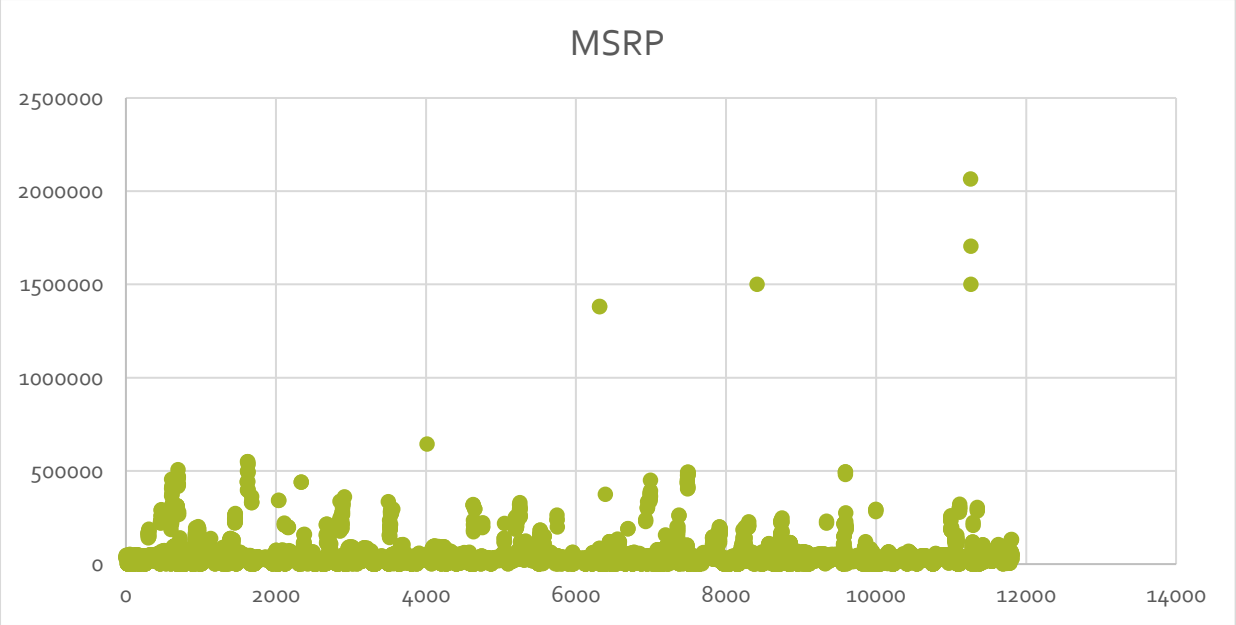
A Clustered Column chart was used to compare average MSRPs across different car brands and body styles. The analysis revealed that Bugatti has the highest average MSRP, while Plymouth has the lowest.

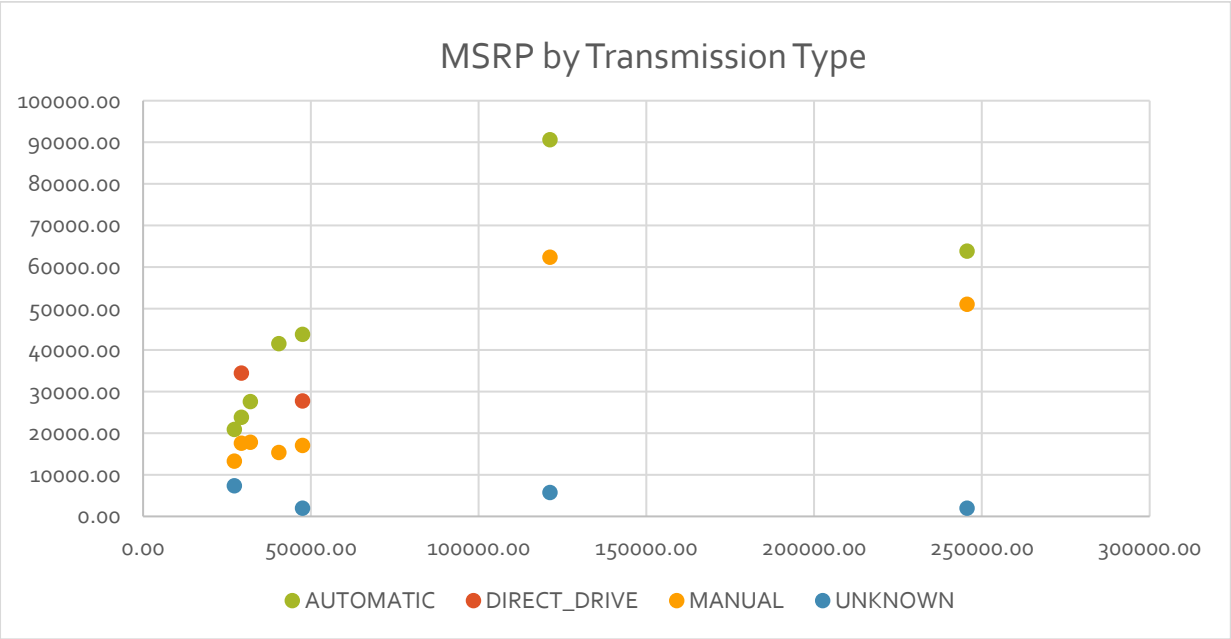




Task 3: How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?

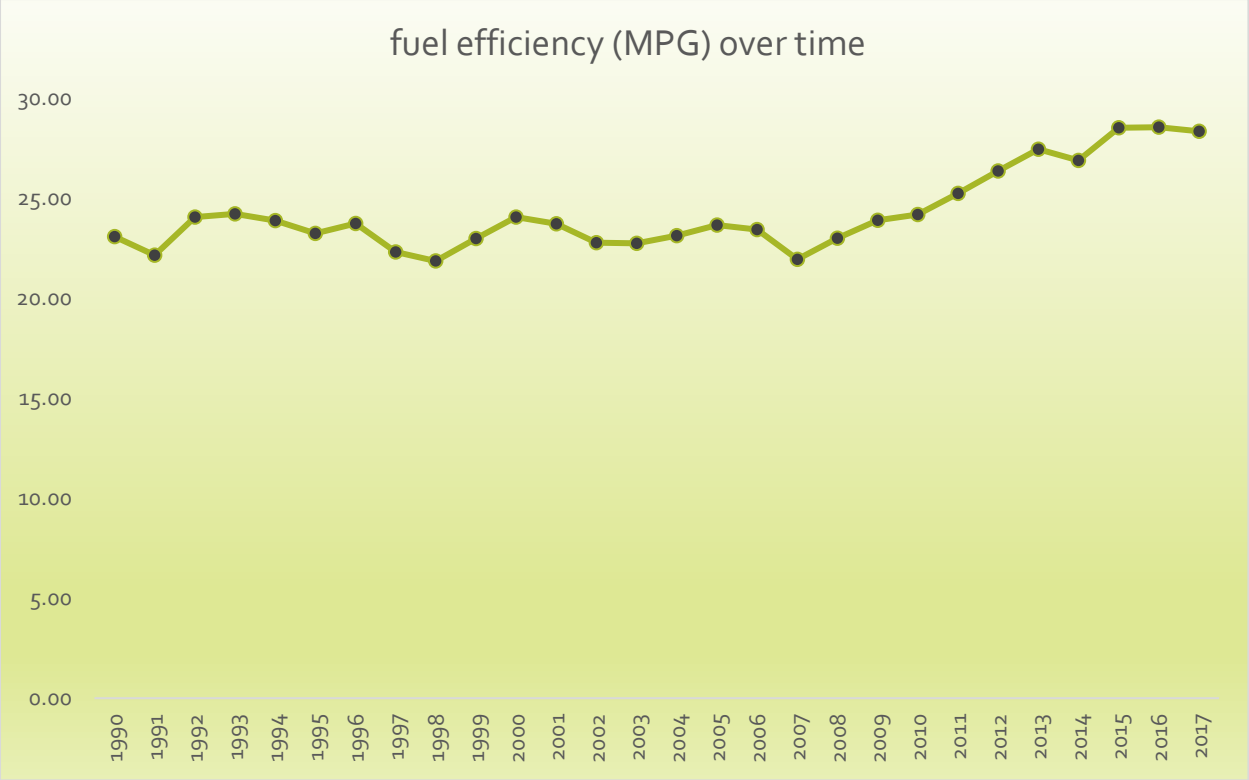
I utilized a Scatter Plot chart to visualize the relationship between MSRP and transmission type, with different symbols for each body style. The analysis indicated higher prices for Automated Manual transmission types and lower prices for Manual transmission types.



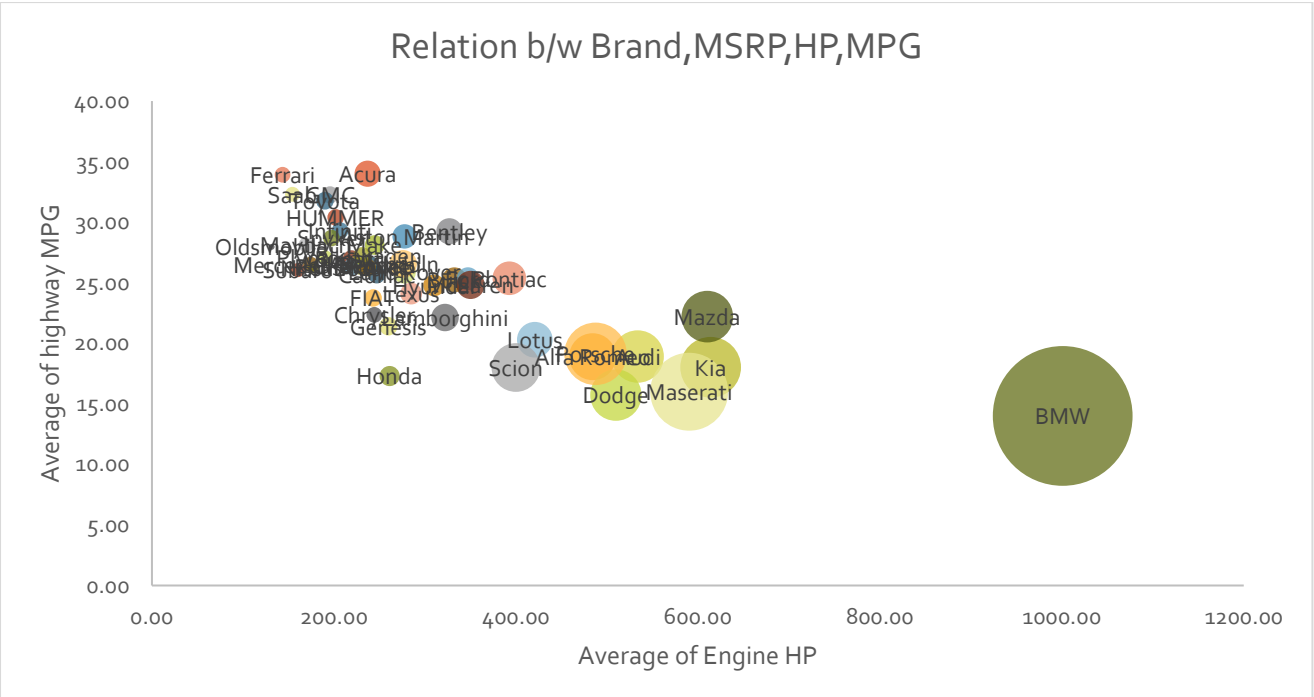


Task 4: How does the fuel efficiency of cars vary across different body styles and model years?

A Line chart was used to display the trend of fuel efficiency (MPG) over time for each body style. The average MPG for each combination of body style and model year was calculated using Pivot Tables. Despite a dip in the year 2007, the analysis showed a gradual increase in fuel efficiency year by year.



Task 5: How does the car's horsepower, MPG, and price vary across different Brands? A Bubble chart was employed to visualize the relationship between horsepower, MPG, and price across different car brands. Different colors were assigned to each brand, and the bubbles were labeled with the car model name. The average horsepower, MPG, and MSRP for each car brand were calculated using Pivot Tables, revealing a positive correlation between Engine HP and MSRP.



Result:

This project provided me with a comprehensive understanding of utilizing Excel Pivot Tables for data analysis and chart creation. I learned how to convert raw data into meaningful insights, clean and manipulate data, and present findings visually. The achievements include:

1. Identification of the most popular Car Model categories.

2. Observation of a positive correlation between Engine HP and Car Price.

3. Recognition of Engine Cylinders' significance in determining Car Price.

4. Identification of Bugatti having the highest average MSRP and Plymouth having the lowest.

5. Exploration of the impact of transmission types on MSRP.

6. Analysis of the upward trend in Fuel Efficiency over the years.

7. Recognition of the correlation between Engine HP, MRSP, and Highway MPG.

This project has contributed significantly to data analysis skills, and I trust that the findings will be valuable for further decision-making processes.

Drive-link for the Excel sheet:

https://docs.google.com/spreadsheets/d/1a4V6WYUOFdGSCpWanWAy4uBryYJL6ltF/edit?usp=sharing&ouid=114833682369349459947&rtpof=true&sd=true

ABC Call Volume Trend Analysis

Description:

A Customer Experience (CX) team plays a crucial role in a company as 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.

In the current era, several AI-powered tools are being used to enhance customer experience. These include Interactive Voice Response (IVR), Robotic Process Automation (RPA), Predictive Analytics, and Intelligent Routing.

One of the key roles in a CX team is that of the customer service representative, also known as a call center agent. These agents handle various types of support, including email, inbound, outbound, and social media support.

Inbound customer support, 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.

In this project, we will delve into the world of Customer Experience (CX) analytics, specifically focusing on the inbound calling team of a company. We are provided with a dataset that spans 23 days and includes various details such as the agent's name and ID, the queue time (how long a customer had to wait before connecting with an agent), the time of the call, the duration of the call, and the call status (whether it was abandoned, answered, or transferred). We will use our analytical skills to understand the trends in the call volume of the CX team and derive valuable insights from it.

Data Pre-Processing:

Handling Null Values:

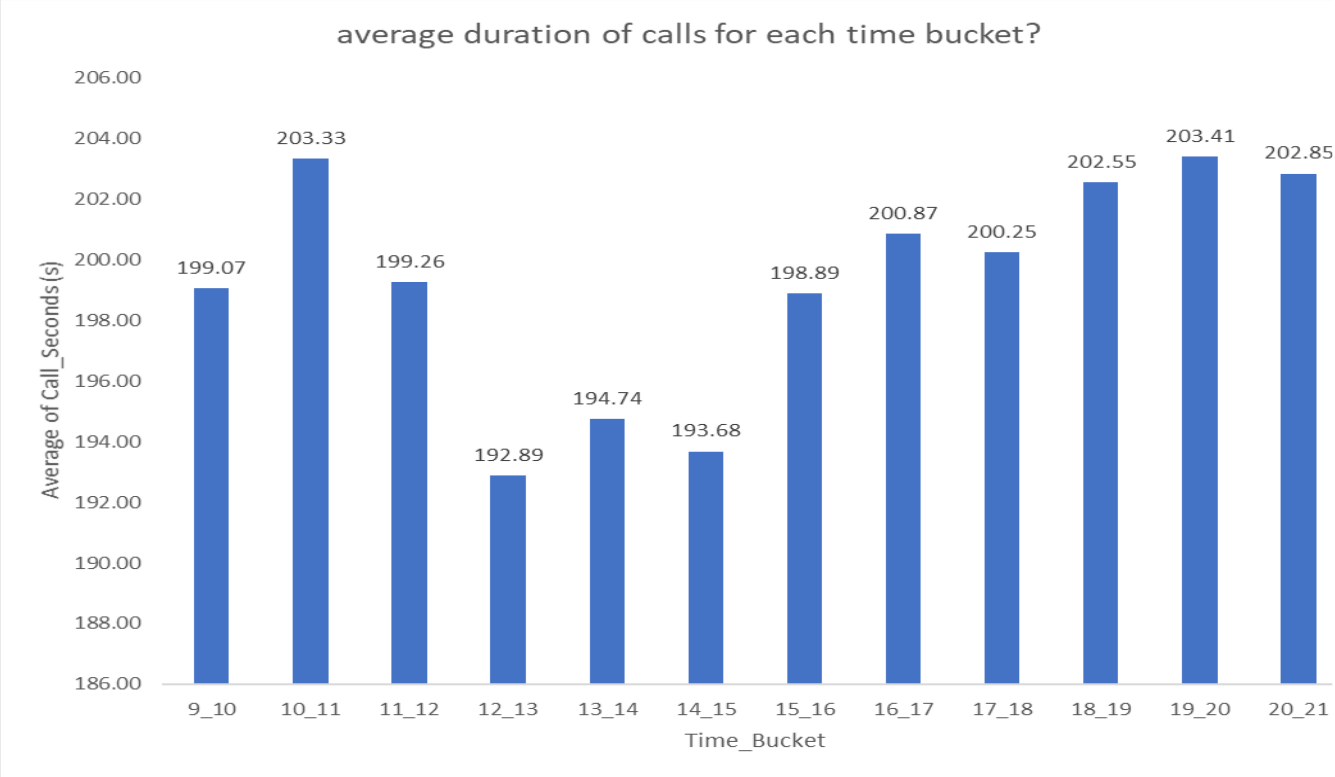
We found that all the rows where Agent_Name and Agent_ID were Null denote abandoned calls. Some calls where Wrapped_By was Null were answered or transferred calls. So we replaced them with the value 'Agent.' The rest of the Null values were replaced by the value 'Not Available.'

Data Analytics Tasks:

Task A: 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?

Result:

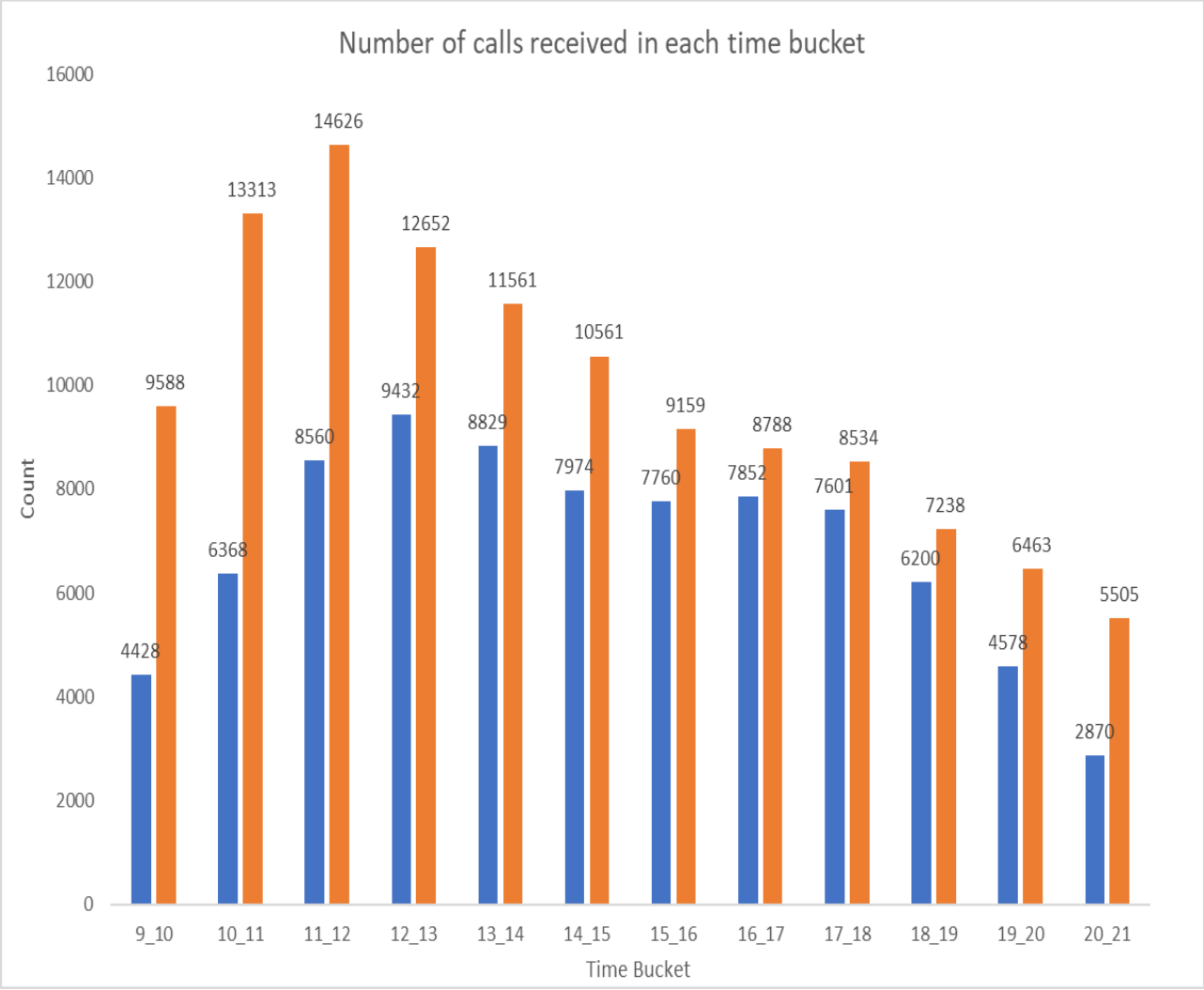


The overall Average Call Duration is 198.62seconds. We observe that the Average Call Duration first peaks in the morning hours before dropping to a below-average value during lunch hours and then increases again to an above-average value.

Task B: 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?

Result:



We observe that the number of received calls first increases with time before dropping down. We can also observe that the number of abandoned calls is very high in the morning hours, and as the day progresses, the number of abandoned calls reduces.

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

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

Result:

time_bucket	total_calls	Abandoned Calls	Answered Calls	Desired Abandoned Calls	Desired Answered Calls	Additional Answered Calls	Additional Agents Needed	Current_agent	required_age
9_10	9588	2876.4	6711.6	958.8	8629.2	1917.6	160	9588	97
10_11	13313	3993.9	9319.1	1331.3	11981.7	2662.6	222	13313	135
11_12	14626	4387.8	10238.2	1462.6	13163.4	2925.2	244	14626	148
12_13	12652	3795.6	8856.4	1265.2	11386.8	2530.4	211	12652	128
13_14	11561	3468.3	8092.7	1156.1	10404.9	2312.2	193	11561	117
14_15	10561	3168.3	7392.7	1056.1	9504.9	2112.2	176	10561	107
15_16	9159	2747.7	6411.3	915.9	8243.1	1831.8	153	9159	93
16_17	8788	2636.4	6151.6	878.8	7909.2	1757.6	146	8788	89
17_18	8534	2560.2	5973.8	853.4	7680.6	1706.8	142	8534	86
18_19	7238	2171.4	5066.6	723.8	6514.2	1447.6	121	7238	73
19_20	6463	1938.9	4524.1	646.3	5816.7	1292.6	108	6463	65
20_21	5505	1651.5	3853.5	550.5	4954.5	1101	92	5505	55

We can observe that to maintain a maximum of a 10% abandon rate, we need to increase the availability of agents in the morning hours by a large margin, as in these hours, the number of incoming calls is quite high, and the number of agents available currently is quite low. During afternoon hours and late evening hours, we need to increase the availability of agents by a slight margin to maintain a maximum of 10% abandon rate.

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

Result:

	09_10	10_11	11_12	12_13	13_14	14_15	15_16	16_17	17_18	18_19	19_20	20_21
SUNDAY	234	460	387	293	316	256	245	255	248	249	231	242
MONDAY	738	1000	1004	902	811	681	458	393	368	313	272	198
TUESDAY	253	293	355	345	309	315	324	294	313	233	233	182
WEDNESDAY	255	310	376	332	291	284	259	245	239	212	187	151
THURSDAY	234	296	358	294	288	276	261	257	228	186	168	132
FRIDAY	187	270	351	266	240	218	193	219	214	172	152	135
SATURDAY	230	328	420	379	316	318	295	289	287	244	192	184

From the above heatmap, we observe that for the day of the week, Monday requires the most number of agents in individual time buckets as well as for the overall day, as it is the start of the week. For the rest of the days, the agent requirement remains more or less the same, with Saturday’s and Sunday’s requirements on the lower side as they are weekends.

Conclusion:

This project helped me understand the importance of Data Analytics in Customer Experience Analysis as it provides valuable insights that help in making Data-Driven Decisions. In this project, I was able to gain insights like call abandon rates, distribution of call duration, number of calls, agents, and how to create a manpower plan to decrease abandoned calls. I also gained experience in Data Preprocessing, like Data Cleaning,

handling Outliers, Feature Engineering, etc. in this project. I can now communicate the insights to relevant stakeholders as per the requirements.

Drive Link:

https://docs.google.com/spreadsheets/d/1HP72iqVC7BfFokknVGyn6Sn0sD4rZxTE/edit?usp=drive_link&ouid=114833682369349459947&rtpof=true&sd=true

Appendix

GitHub Repository:

https://github.com/samarjithMnagesha?tab=repositories