Atharva Jadhav Portfolio

Professional Background

Currently in Final Year, pursuing Bachelor of Technology in Electronics and Telecommunication Engineering from Pimpri Chinchwad College of Engineering, Pune.

I have several Technical skills including Data Analytics using python, Machine Learning, Deep Learning, Strong understanding of Excel, DBMS and SQL. I am interested in working in Computer Vision and Natural Language Processing.

I have also worked on a few personal projects such as Movie Recommendation using TF-IDF in NLP, Drive Drowsiness Detection using Computer Vision and CNN, Online Multiplayer Game utilizing a server and client to play from anywhere across the world.

I am also adept in competitive programming and have solved a decent number of problems on codechef, leetcode and geeksforgeeks, 500+ combined.

I am a highly motivated and results-driven individual interested in machine learning and data analysis. I am adept at leveraging data-driven insights to solve complex problems and drive informed decision-making. With a passion for innovation and a commitment to continuous learning, I am seeking an opportunity to contribute to a dynamic team and make a significant impact in the field of machine learning.

Table of Contents

1. Professional Background 1
2. Table of Contents 2
3. Data Analytics Process 4
4. Instagram user analytics 5
    a. Description 5
    b. Approach 5
    c. Insights 6
    d. Results 12
5. Operation Analytics and Investigating Metric Spike 13
    a. Description 13
    b. Approach 13
    c. Insights 14
    d. Results 23
6. Hiring Process Analytics 24
    a. Description 24
    b. Approach 24
    c. Insights 25
    d. Results 32
7. IMDB Movie Analysis 33
    a. Description 33
    b. Approach 35
    c. Insights 36
    d. Results 41
8. Bank Loan Case Study 42
    a. Description 42
    b. Approach 42
    c. Insights 43
    d. Results 55
9. Impact of Car Features on Price and Profitability 56
    a. Description 56
    b. Approach 56
    c. Data Cleaning 57
    d. Insights 59
    e. Making DashBoard 71

f. Results 71

10. ABC Call Volume Trend Analysis 72

a. Description 72

b. Approach 72

c. Data Cleaning 73

d. Insights 75

e. Results 81

11. Appendix 82

Data Analytics Process

Plan: Suppose we wish to buy a laptop. First we decide, from where to buy the device, from the offline market or through e-commerce websites.

Prepare: Determine how much we are willing to spend on the laptop and set a budget.

Process: Then we research online and compare different laptop models based on their specifications, features, and prices.
Then we determine the specific features and specifications desired in a laptop, such as processor, RAM, storage capacity, screen size, and any additional requirements like a dedicated graphics card or touchscreen.
Also determine the type of laptop we want, such as a traditional laptop, 2-in-1 convertible, gaming laptop, or ultrabook, based on my usage and preferences.

Analyze: To analyze we read customer reviews and ratings for the shortlisted laptop models to gather insights about their performance, reliability, and user experience.
Verify if the laptop comes with a warranty and reliable customer support options in case any issues arise in the future.
We compare the prices of shortlisted laptops across different online stores to find the best deal.

Share: If we have any specific queries or concerns, we can reach out to the online store's customer support team to seek clarification.

Act: Add the chosen laptop to the online store's shopping cart and proceed to checkout.
Enter the required shipping information, such as the delivery address and contact details. Choose a preferred payment method, such as credit card, debit card, or online payment platforms, to complete the purchase. Choose a preferred payment method, such as credit card, debit card, or online payment platforms, to complete the purchase.

Instagram User Analytics

Project Description:

This project focuses on analyzing user behavior on Instagram to provide insights for marketing and investor assessments. By using SQL fundamentals, we will answer specific questions from the marketing and management teams.

In the marketing section, we will find the oldest users, identify users who haven't posted photos, determine the winner of a contest, suggest popular hashtags, and provide insights on the best day to launch AD campaigns.

In the investor metrics section, we will analyze user engagement by calculating the average number of posts per user and assess the presence of fake accounts by identifying users who have liked every single photo.

Through this project, we aim to provide valuable data-driven recommendations to support marketing campaigns, inform decision-making, and evaluate Instagram's performance and authenticity compared to other platforms.

Approach:

As an individual working on this project, I followed a structured approach to analyze user behavior on Instagram and find meaningful insights. I began by carefully examining the provided database and familiarizing myself with its structure. Then, I utilized SQL fundamentals to retrieve the necessary information for each task, employing appropriate queries and functions. I focused on data accuracy and quality throughout the project, ensuring reliable results. By leveraging my SQL skills and maintaining a systematic workflow, I successfully executed the project and created a comprehensive report that fulfilled the objectives of providing marketing insights and investor metrics.

Tech-Stack Used:

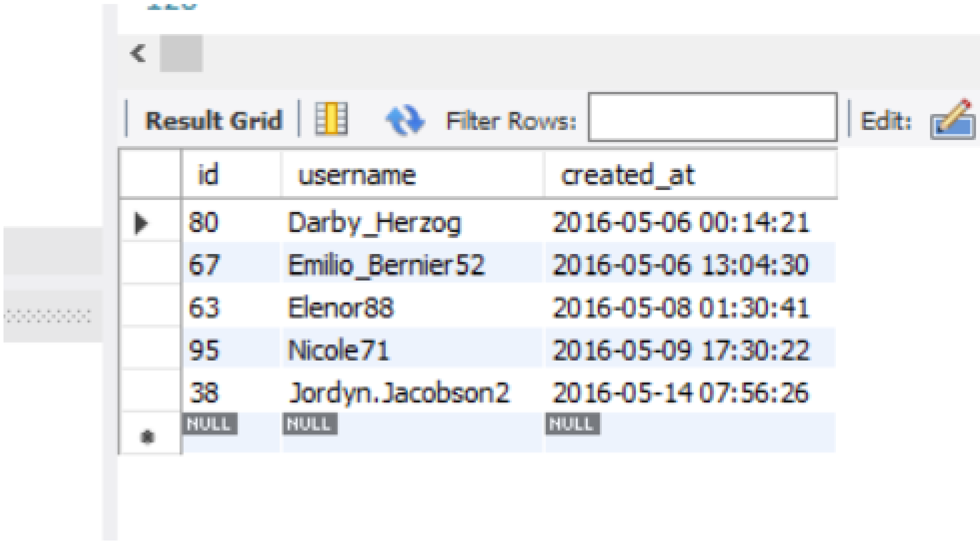
For this project, I utilized MySQL Workbench 8.0 as the primary software tool. MySQL Workbench is an integrated development environment (IDE) for MySQL databases, providing a graphical interface for designing, querying, and managing databases.

Insights:

1. Rewarding Most Loyal Users: Identifying the oldest users of Instagram helps recognize and reward long-term user loyalty.
For finding most loyal users following query was used:

```
SELECT id, username, created_at
FROM users
ORDER BY created_at
LIMIT 5;
```

Here we get the information of the oldest users from the "users" table, including their ID, username, and the date they joined, sorted in order of their joining date, we limit it to 5, to get top 5 users.
Following is the result of query:



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

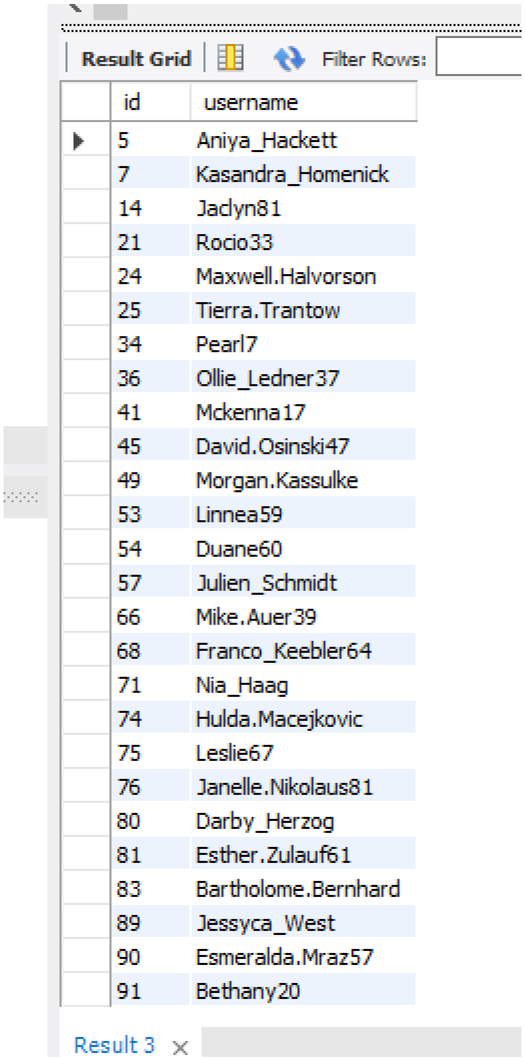
2. Remind Inactive Users to Start Posting: Users who have never posted a single photo on Instagram represent an opportunity for re-engagement through targeted promotional emails.

For finding inactive users following query was used:

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

Here we select the ID and username from the "users" table, where we find users who have not posted any photos based on the left join with the "photos" table, matching the user IDs. We filter out users who have a null value for the photo ID, indicating that they haven't posted any photos.

Following is the result of query:



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

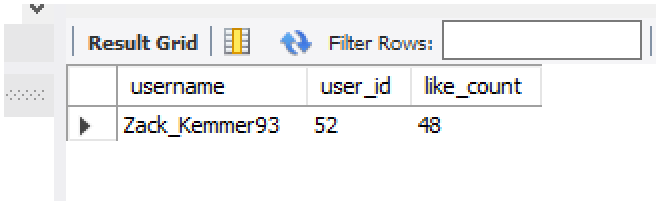
3. Declaring Contest Winner: The winner of a contest can be determined by the user with the most likes on a single photo, ensuring a fair and accurate declaration.

For finding user with most likes on a photo following query was used:

```
SELECT u.username, p.user_id, l.like_count
FROM (
   SELECT photo_id, COUNT(*) AS like_count
   FROM likes
   GROUP BY photo_id
   ORDER BY like_count DESC
   LIMIT 1
) l
JOIN photos p ON l.photo_id = p.id
JOIN users u ON p.user_id = u.id;
```

Here we are finding the username, user ID, and number of likes for the user with the most popular photo. We count the likes for each photo, find the photo with the highest count, and then match it with the user who owns the photo. By joining the photos, users and likes table, we retrieve the username and user ID for the user with the most likes on their photo.

Following is the result of query:

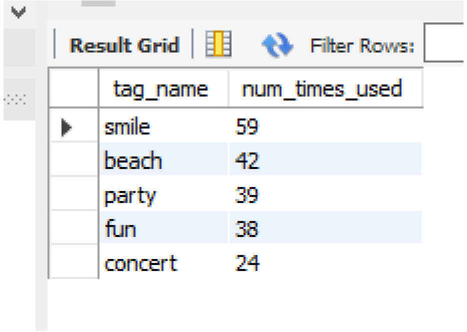


username	user_id	like_count
Zack_Kemmer93	52	48

4. Hashtag Researching: Identifying the top five most commonly used hashtags on Instagram allows for effective hashtag selection to reach a broader audience.
For finding top 5 hashtags used most commonly, following query was used:

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

To find the top 5 most commonly used hashtags, we join the "tags" table with the "photo_tags" table using their respective IDs. By grouping the records based on the tag name, we count the number of times each tag has been used. We then sort the results in descending order based on the count of usage, selecting only the top 5.
Following is the result of query:

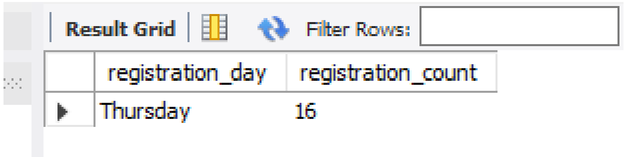


tag_name	num_times_used
smile	59
beach	42
party	39
fun	38
concert	24

5. Launch AD Campaign: Analyzing user registration patterns reveals the best day to launch advertisements, maximizing their potential impact and reach.
For finding day on which most users register, following query was used:

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

By selecting the day name from the "created_at" column of the "users" table, we group the registrations by day of the week. We count the number of registrations for each day and sort the results in descending order based on the registration count. We select only the top result, which represents the day of the week with the highest number of user registrations on Instagram.
Following is the result of query:

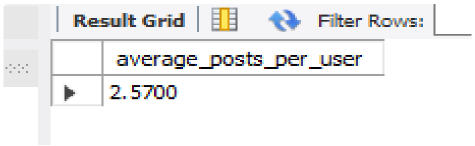


registration_day	registration_count
Thursday	16

6. User Engagement: Calculating the average number of posts per user provides insights into user activity levels and potential trends in engagement.
For finding average number of posts per user, following query was used:

```
SELECT COUNT(photos.id) / COUNT(DISTINCT users.id) AS average_posts_per_user
FROM users
LEFT JOIN photos ON users.id = photos.user_id;
```

By joining the "users" table with the "photos" table using their respective IDs, we count the total number of photos. We also count the distinct number of users. Then, we divide the total number of photos by the distinct number of users to calculate the average posts per user. This gives an average number of posts per user.
Following is the result of query:



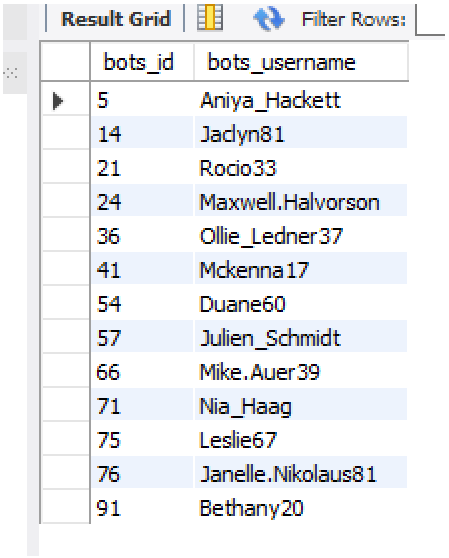
average_posts_per_user
2.5700

7. Bots & Fake Accounts: Identifying users who have liked every single photo helps detect the presence of bots and fake accounts, ensuring a more authentic user community.

```
SELECT u.id as bots_id, u.username AS bots_username
FROM (
   SELECT l.user_id
   FROM (
      SELECT p.id
      FROM photos p
   ) AS all_photos
   LEFT JOIN likes l ON all_photos.id = l.photo_id
   GROUP BY l.user_id
   HAVING COUNT(DISTINCT l.photo_id) = (SELECT COUNT(*) FROM photos)
) AS bots
JOIN users u ON bots.user_id = u.id;
```

To find bots, we find the users who have liked every single photo on Instagram. First, we select all the photo IDs from the "photos" table. Next, we join this list of photo IDs (stored as "all_photos") with the "likes" table on the photo ID to find the corresponding user IDs who have liked those photos. We group the results by the user ID and filter them using the "HAVING" clause, where we count the distinct photo IDs liked by each user and compare it to the total count of photos in the "photos" table. Then, we join this result (stored as "bots") with the "users" table using the matching user IDs.

Following is the result of query:



bots_id	bots_username
5	Aniya_Hackett
14	Jaclyn81
21	Rocio33
24	Maxwell.Halvorson
36	Ollie_Ledner37
41	Mckenna17
54	Duane60
57	Julien_Schmidt
66	Mike.Auer39
71	Nia_Haag
75	Leslie67
76	Janelle.Nikolaus81
91	Bethany20

Results:

While working on this project, I have gained a better understanding of user analytics and SQL fundamentals. By analyzing user data on Instagram, I was able to provide insights on various aspects such as rewarding loyal users, identifying inactive users, declaring contest winners, researching popular hashtags, determining the best day to launch ad campaigns, assessing user engagement, and detecting fake accounts.

This project has helped me enhance my SQL skills, particularly in querying and manipulating data to derive meaningful insights. It has also improved my ability to interpret data and provide actionable recommendations based on the analysis. Overall, this project has deepened my understanding of user behavior analysis and its application in making informed decisions for product development and marketing strategies.

Operation Analytics and Investigating Metric Spike

Project Description:

Case Study 1 (Job Data):

This project focuses on analyzing a job_data table to answer specific questions using SQL fundamentals. The table contains information about jobs, including job_id, actor_id, event, language, time_spent, org, and ds columns. The questions include calculating the number of jobs reviewed per hour per day, analyzing throughput, determining the percentage share of each language, and identifying duplicate rows.

Case Study 2 (Investigating Metric Spike):

This project involves analyzing three tables: users, events, and email_events. The goal is to calculate metrics related to user engagement, user growth, weekly retention, weekly engagement per device, and email engagement. SQL fundamentals will be used to extract insights from the dataset and provide valuable information for marketing campaigns and decision-making.

Approach:

As an individual working on this project, I followed a structured approach to analyze the job_data table. I began by understanding the table structure and column definitions. Using SQL queries and functions, I calculated the number of jobs reviewed per hour per day for November 2020, analyzed throughput by calculating events per second and 7-day rolling average, determined the percentage share of each language in the last 30 days, and identified duplicate rows. I prioritized data accuracy, optimized queries for efficiency, and maintained documentation of my workflow. The project aimed to provide valuable insights for marketing and investor assessments, achieved through the successful application of SQL fundamentals.

Tech-Stack Used:

For this project, I utilized MySQL Workbench 8.0 as the primary software tool. MySQL Workbench is an integrated development environment (IDE) for MySQL databases, providing a graphical interface for designing, querying, and managing databases.

Insights:

Case Study 1(Job Data):

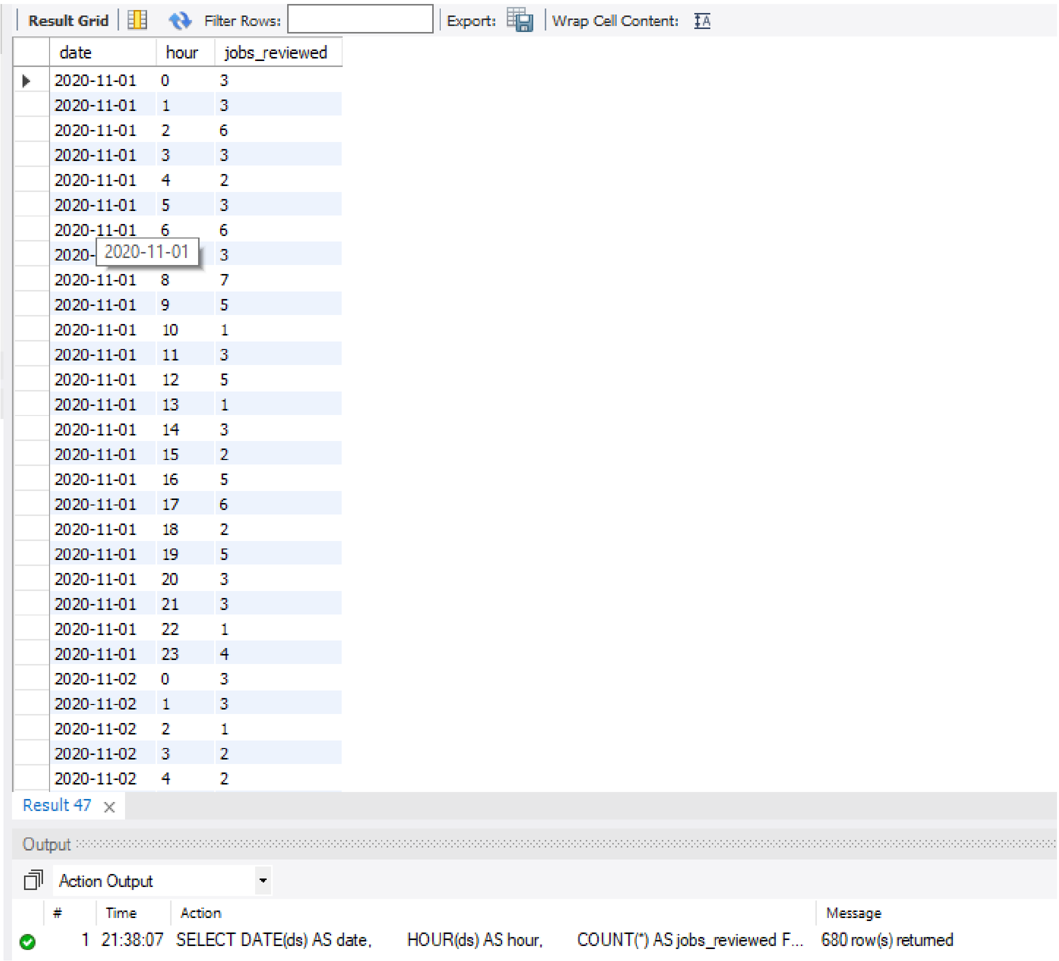
Number of jobs reviewed: The number of jobs reviewed per day per hour indicates the level of activity in job review processes.

To calculate the number of jobs reviewed per hour per day for November 2020, I utilized SQL queries to filter the data for the specified time period and grouped the results by hour and day.

Query:

```
SELECT DATE(ds) AS date,
       HOUR(ds) AS hour,
       COUNT(*) AS jobs_reviewed
FROM sheet1
WHERE ds BETWEEN '2020-11-01' AND '2020-11-30'
GROUP BY DATE(ds), HOUR(ds)
ORDER BY DATE(ds), HOUR(ds);
```

Output:



date	hour	jobs_reviewed
2020-11-01	0	3
2020-11-01	1	3
2020-11-01	2	6
2020-11-01	3	3
2020-11-01	4	2
2020-11-01	5	3
2020-11-01	6	6
2020-11-01	[illegible]	3
2020-11-01	8	7
2020-11-01	9	5
2020-11-01	10	1
2020-11-01	11	3
2020-11-01	12	5
2020-11-01	13	1
2020-11-01	14	3
2020-11-01	15	2
2020-11-01	16	5
2020-11-01	17	6
2020-11-01	18	2
2020-11-01	19	5
2020-11-01	20	3
2020-11-01	21	3
2020-11-01	22	1
2020-11-01	23	4
2020-11-02	0	3
2020-11-02	1	3
2020-11-02	2	1
2020-11-02	3	2
2020-11-02	4	2

```
#  Time      Action                                                                    Message
1  21:38:07  SELECT DATE(ds) AS date,    HOUR(ds) AS hour,    COUNT(*) AS jobs_reviewed F...   680 row(s) returned
```

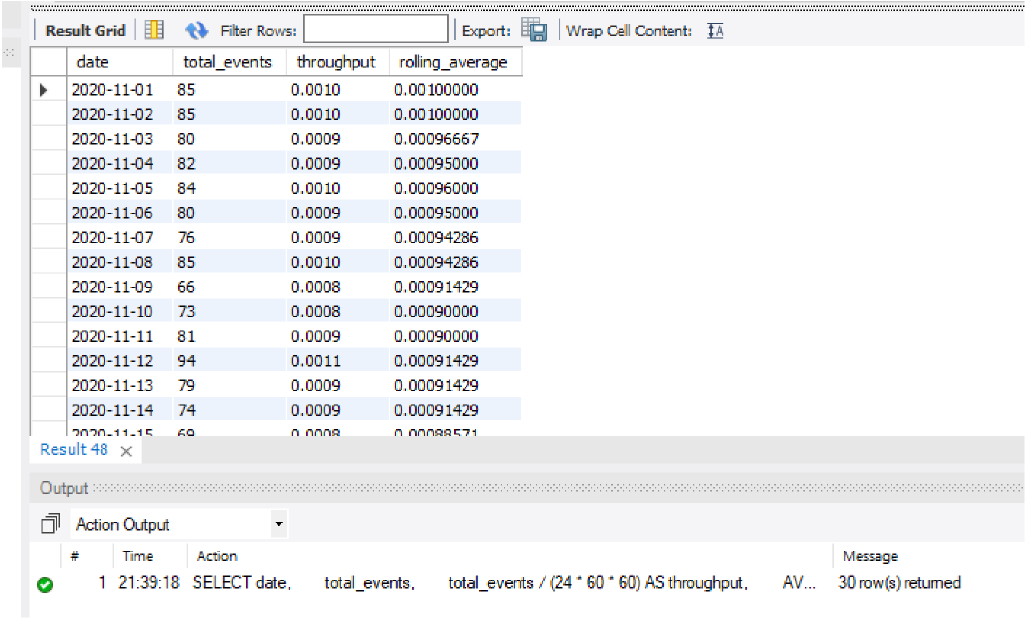
Throughput: Throughput refers to the number of events happening per second, representing the system's processing capacity.

To calculate the 7-day rolling average of throughput, I used SQL queries to aggregate the events per second and then calculated the average over a rolling window of 7 days. This helps identify any trends or variations in the throughput metric. I prefer the 7-day rolling average because it provides a smoother representation of the metric, reducing the impact of daily fluctuations and offering a more comprehensive view of the system's performance.

Query:

```
SELECT date,
       total_events,
       total_events / (24 * 60 * 60) AS throughput,
       AVG(total_events / (24 * 60 * 60)) OVER (ORDER BY date ROWS BETWEEN
6 PRECEDING AND CURRENT ROW) AS rolling_average
FROM (
    SELECT DATE(ds) AS date, COUNT(*) AS total_events
    FROM sheet1
    WHERE ds BETWEEN '2020-11-01' AND '2020-11-30'
    GROUP BY DATE(ds)
) AS subquery
GROUP BY date, total_events
ORDER BY date;
```

Output:



date	total_events	throughput	rolling_average
2020-11-01	85	0.0010	0.00100000
2020-11-02	85	0.0010	0.00100000
2020-11-03	80	0.0009	0.00096667
2020-11-04	82	0.0009	0.00095000
2020-11-05	84	0.0010	0.00096000
2020-11-06	80	0.0009	0.00095000
2020-11-07	76	0.0009	0.00094286
2020-11-08	85	0.0010	0.00094286
2020-11-09	66	0.0008	0.00091429
2020-11-10	73	0.0008	0.00090000
2020-11-11	81	0.0009	0.00090000
2020-11-12	94	0.0011	0.00091429
2020-11-13	79	0.0009	0.00091429
2020-11-14	74	0.0009	0.00091429
2020-11-15	69	0.0008	0.00088571

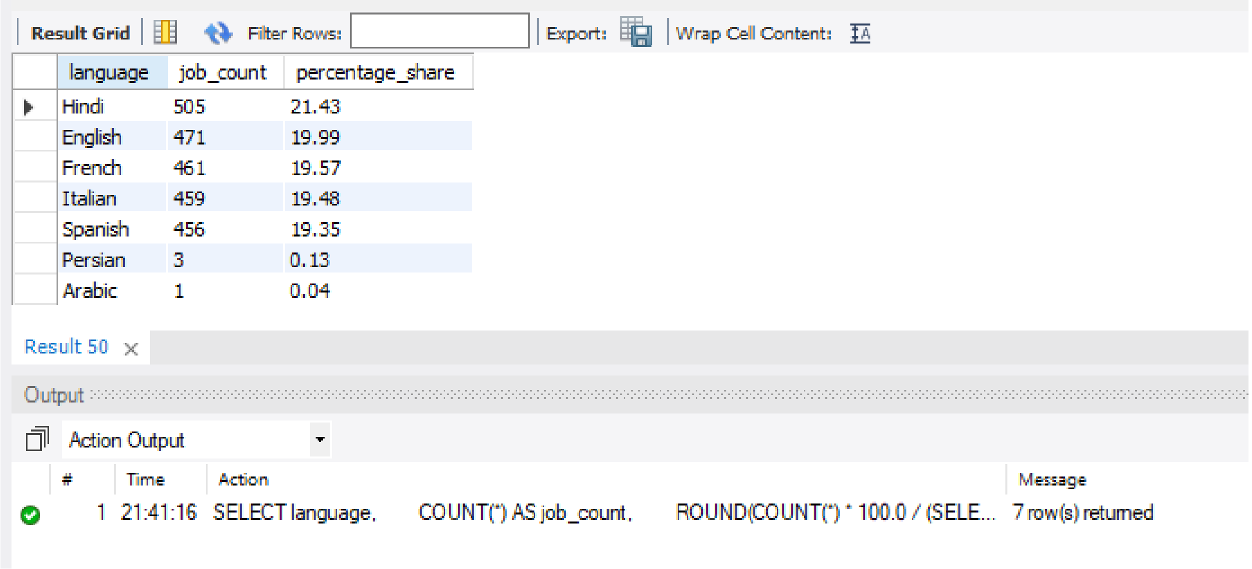
#	Time	Action	Message
1	21:39:18	SELECT date, total_events, total_events / (24 * 60 * 60) AS throughput, AV...	30 row(s) returned

Percentage share of each language: The percentage share of each language in different contents provides insights into language preferences and content distribution. To calculate the percentage share of each language in the last 30 days, I used SQL queries to filter the data for the specified time period and performed calculations to determine the language distribution. By dividing the count of each language by the total count of contents, I obtained the percentage share for each language.

Query:

```
SELECT language,
       COUNT(*) AS job_count,
       ROUND(COUNT(*) * 100.0 / (SELECT COUNT(*) FROM sheet1  WHERE ds
BETWEEN '2020-11-01' AND '2020-11-30'), 2) AS percentage_share
FROM sheet1
WHERE ds BETWEEN '2020-11-01' AND '2020-11-30'
GROUP BY language
ORDER BY percentage_share DESC;
```

Output:



language	job_count	percentage_share
Hindi	505	21.43
English	471	19.99
French	461	19.57
Italian	459	19.48
Spanish	456	19.35
Persian	3	0.13
Arabic	1	0.04

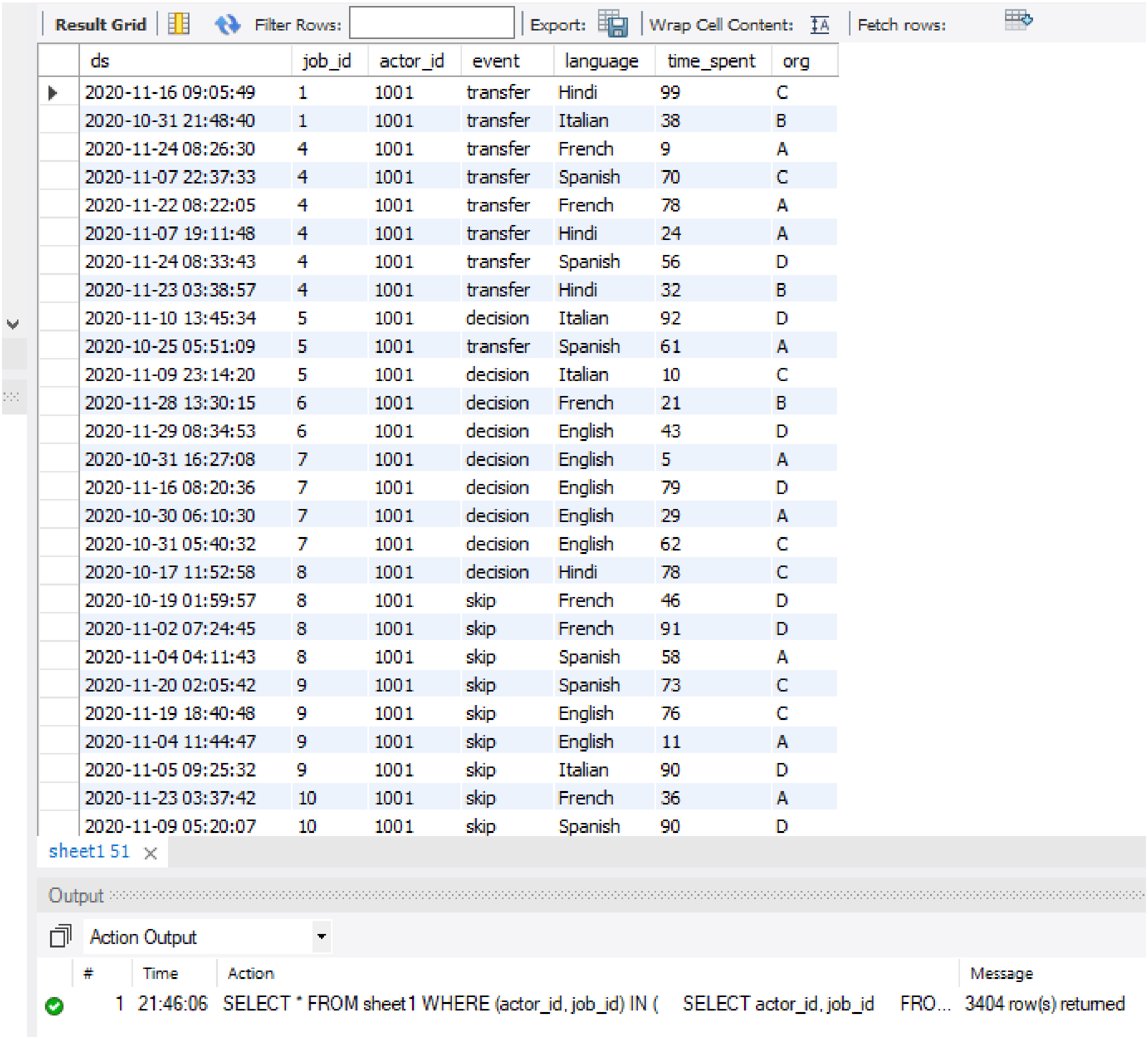
```
#  Time      Action                                                                  Message
1  21:41:16  SELECT language,   COUNT(*) AS job_count,   ROUND(COUNT(*) * 100.0 / (SELE...  7 row(s) returned
```

Duplicate rows: Duplicate rows refer to rows in the dataset that have the same values. If actor_id and job_id are the same for an entry, then that entry is considered to be duplicate. Comparing the values across columns, I can identify rows with identical values and retrieve them from the table.

Query:

```
SELECT *
FROM sheet1
WHERE (actor_id, job_id) IN (
    SELECT actor_id, job_id
    FROM sheet1
    GROUP BY actor_id, job_id
    HAVING COUNT(*) > 1
)
ORDER BY actor_id, job_id;
```

Output:



ds	job_id	actor_id	event	language	time_spent	org
2020-11-16 09:05:49	1	1001	transfer	Hindi	99	C
2020-10-31 21:48:40	1	1001	transfer	Italian	38	B
2020-11-24 08:26:30	4	1001	transfer	French	9	A
2020-11-07 22:37:33	4	1001	transfer	Spanish	70	C
2020-11-22 08:22:05	4	1001	transfer	French	78	A
2020-11-07 19:11:48	4	1001	transfer	Hindi	24	A
2020-11-24 08:33:43	4	1001	transfer	Spanish	56	D
2020-11-23 03:38:57	4	1001	transfer	Hindi	32	B
2020-11-10 13:45:34	5	1001	decision	Italian	92	D
2020-10-25 05:51:09	5	1001	transfer	Spanish	61	A
2020-11-09 23:14:20	5	1001	decision	Italian	10	C
2020-11-28 13:30:15	6	1001	decision	French	21	B
2020-11-29 08:34:53	6	1001	decision	English	43	D
2020-10-31 16:27:08	7	1001	decision	English	5	A
2020-11-16 08:20:36	7	1001	decision	English	79	D
2020-10-30 06:10:30	7	1001	decision	English	29	A
2020-10-31 05:40:32	7	1001	decision	English	62	C
2020-10-17 11:52:58	8	1001	decision	Hindi	78	C
2020-10-19 01:59:57	8	1001	skip	French	46	D
2020-11-02 07:24:45	8	1001	skip	French	91	D
2020-11-04 04:11:43	8	1001	skip	Spanish	58	A
2020-11-20 02:05:42	9	1001	skip	Spanish	73	C
2020-11-19 18:40:48	9	1001	skip	English	76	C
2020-11-04 11:44:47	9	1001	skip	English	11	A
2020-11-05 09:25:32	9	1001	skip	Italian	90	D
2020-11-23 03:37:42	10	1001	skip	French	36	A
2020-11-09 05:20:07	10	1001	skip	Spanish	90	D

#	Time	Action	Message
1	21:46:06	SELECT * FROM sheet1 WHERE (actor_id, job_id) IN (SELECT actor_id, job_id FRO...	3404 row(s) returned

Case Study 2 (Investigating Metric Spike):

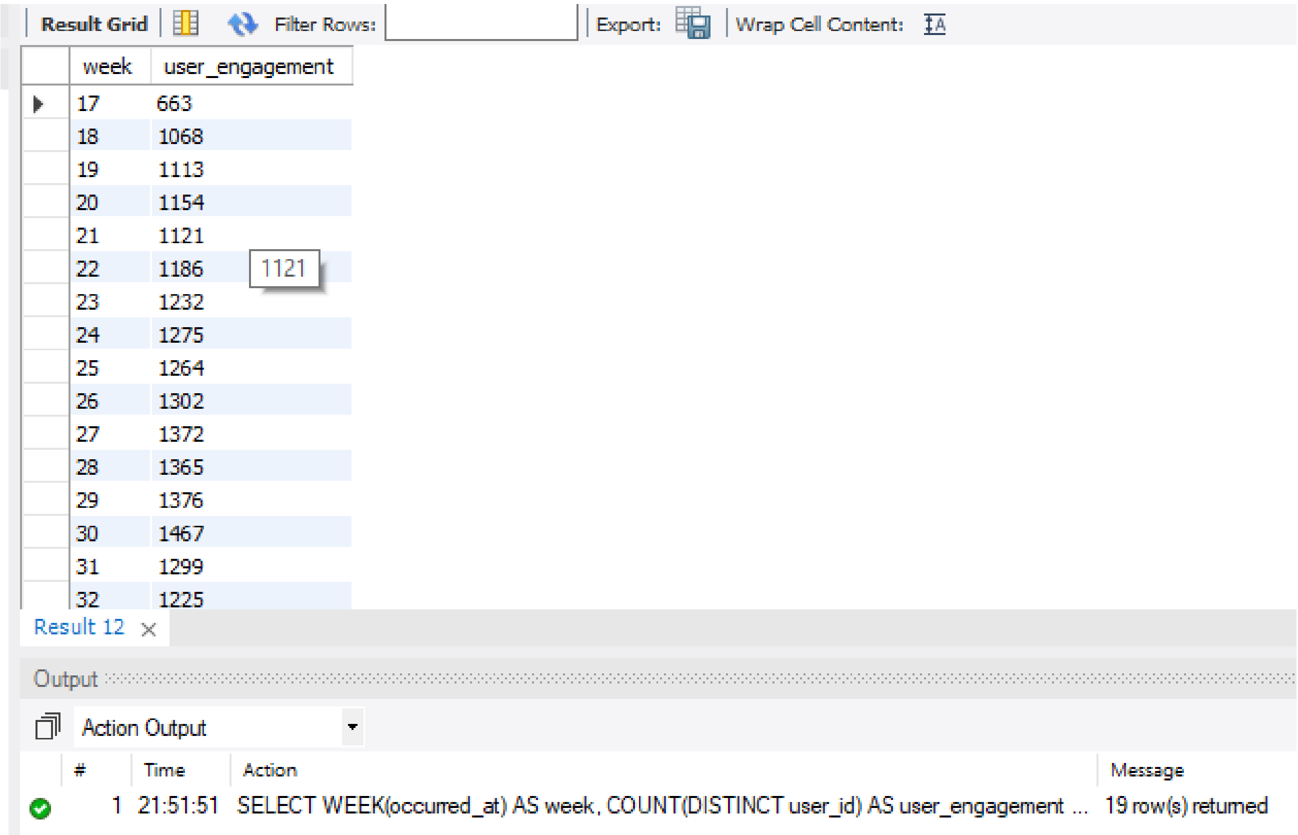
User Engagement: User engagement is a measure of how active users are and indicates their satisfaction with a product or service.

To calculate the weekly user engagement, I utilized SQL queries to analyze unique user engagement events within the specified week.

Query:

```
SELECT WEEK(occurred_at) AS week, COUNT(DISTINCT user_id) AS
user_engagement
FROM events
WHERE event_type = 'engagement'
GROUP BY week;
```

Output:



week	user_engagement
17	663
18	1068
19	1113
20	1154
21	1121
22	1186
23	1232
24	1275
25	1264
26	1302
27	1372
28	1365
29	1376
30	1467
31	1299
32	1225

```
#  Time      Action                                                                  Message
1  21:51:51  SELECT WEEK(occurred_at) AS week, COUNT(DISTINCT user_id) AS user_engagement ...  19 row(s) returned
```

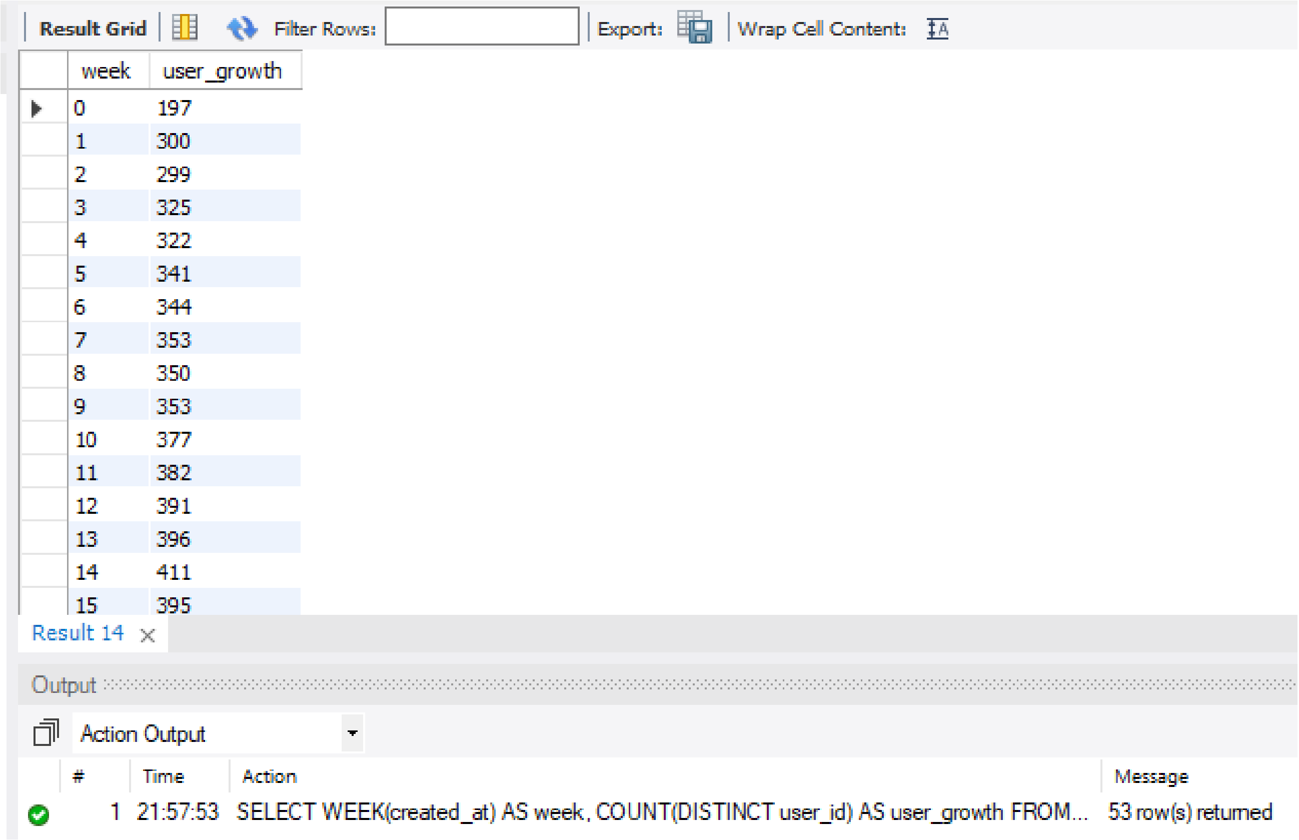
User Growth: User growth measures the increase in the number of users over a specific period, reflecting the product's adoption and popularity.

To calculate user growth for a product, I used SQL queries to track the number of new users added over time. By comparing the count of users added in different weeks, we can identify growth of product.

Query:

```
SELECT WEEK(created_at) AS week, COUNT(DISTINCT user_id) AS user_growth
FROM user
GROUP BY week;
```

Output:



week	user_growth
0	197
1	300
2	299
3	325
4	322
5	341
6	344
7	353
8	350
9	353
10	377
11	382
12	391
13	396
14	411
15	395

```
#   Time      Action                                                              Message
1   21:57:53  SELECT WEEK(created_at) AS week, COUNT(DISTINCT user_id) AS user_growth FROM...   53 row(s) returned
```

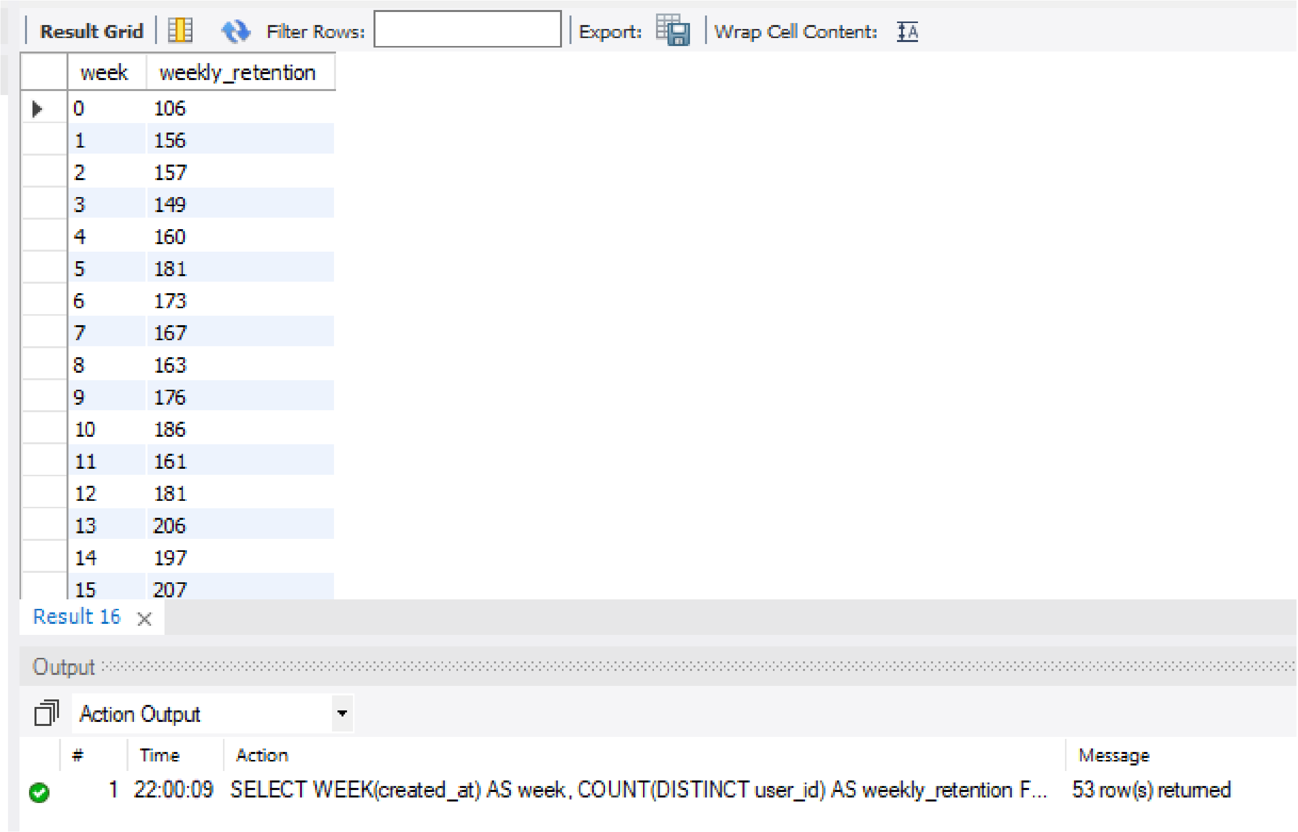
Weekly Retention: Weekly retention evaluates the percentage of users who continue to use a product or service after signing up, indicating its ability to retain users.

To calculate the weekly retention of users, I employed SQL queries to track the state of users and identify users who remained active in consecutive weeks. By comparing the count of retained users to the initial sign-up cohort, I determined the retention rate for each week.

Query:

```
SELECT WEEK(created_at) AS week, COUNT(DISTINCT user_id) AS
weekly_retention
FROM user
WHERE state = 'active'
GROUP BY week;
```

Output:



week	weekly_retention
0	106
1	156
2	157
3	149
4	160
5	181
6	173
7	167
8	163
9	176
10	186
11	161
12	181
13	206
14	197
15	207

```
#   Time      Action                                                                   Message
1   22:00:09  SELECT WEEK(created_at) AS week, COUNT(DISTINCT user_id) AS weekly_retention F...   53 row(s) returned
```

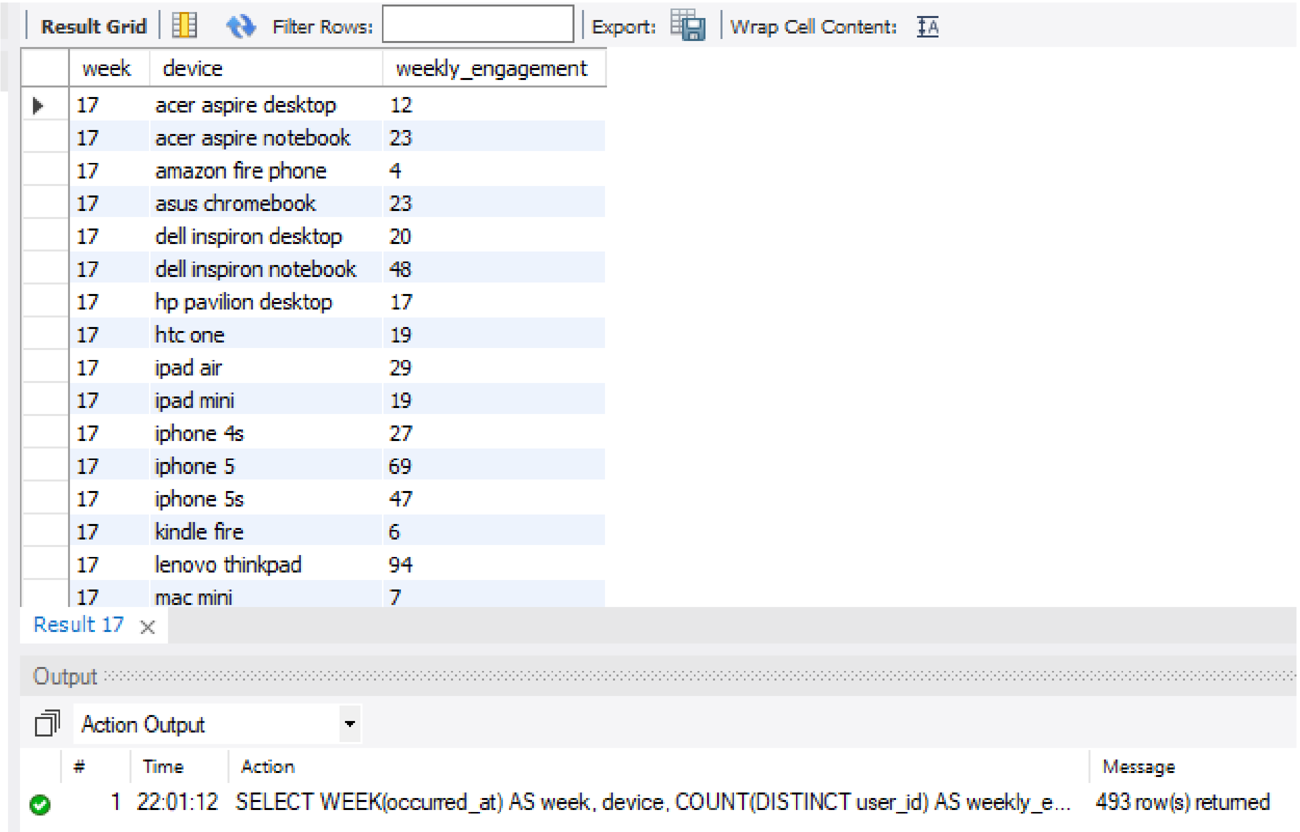
Weekly Engagement: Weekly engagement measures the level of user activity and satisfaction with a product or service on a weekly basis.

To calculate the weekly engagement per device, I utilized SQL queries to analyze user interactions and activities categorized by device type. By selecting distinct users and their devices and grouping by week and device, I assessed the level of engagement for each device category.

Query:

```
SELECT WEEK(occurred_at) AS week, device, COUNT(DISTINCT user_id) AS
weekly_engagement
FROM events
GROUP BY week, device;
```

Output:



week	device	weekly_engagement
17	acer aspire desktop	12
17	acer aspire notebook	23
17	amazon fire phone	4
17	asus chromebook	23
17	dell inspiron desktop	20
17	dell inspiron notebook	48
17	hp pavilion desktop	17
17	htc one	19
17	ipad air	29
17	ipad mini	19
17	iphone 4s	27
17	iphone 5	69
17	iphone 5s	47
17	kindle fire	6
17	lenovo thinkpad	94
17	mac mini	7

#	Time	Action	Message
1	22:01:12	SELECT WEEK(occurred_at) AS week, device, COUNT(DISTINCT user_id) AS weekly_e...	493 row(s) returned

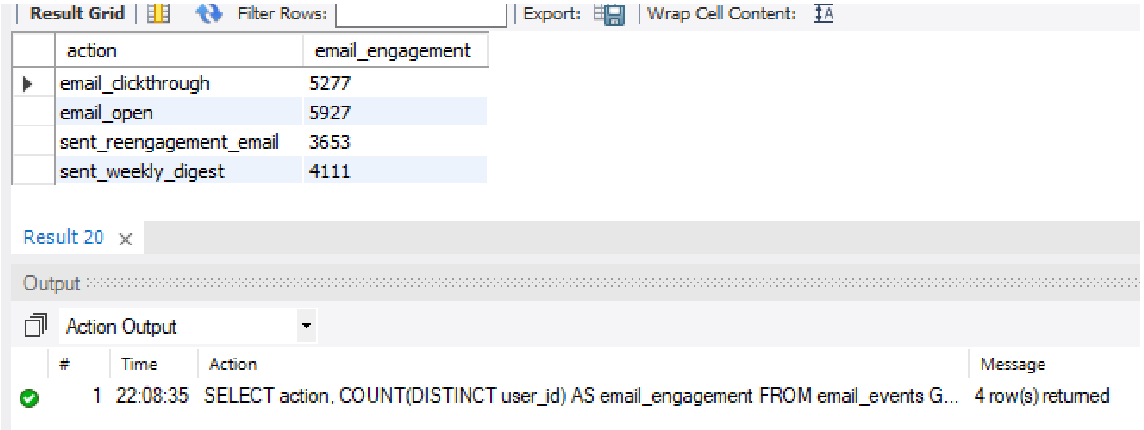
Email Engagement: Email engagement reflects user involvement and interaction with the email service.

To calculate email engagement metrics, I used SQL queries to analyze user email activities such as opens, clicks, and responses.

Query:

```
SELECT action, COUNT(DISTINCT user_id) AS email_engagement
FROM email_events
GROUP BY action;
```

Output:



action	email_engagement
email_clickthrough	5277
email_open	5927
sent_reengagement_email	3653
sent_weekly_digest	4111

#	Time	Action	Message
1	22:08:35	SELECT action, COUNT(DISTINCT user_id) AS email_engagement FROM email_events G...	4 row(s) returned

Results:

While working on this project, I have gained a better understanding of user analytics and SQL fundamentals. By analyzing Operation Analytics and Investigating Metric Spike, I was able to provide insights on various aspects such as number of jobs analyzed per hour per day, 7 day rolling average of throughput, percentage share of each language, user engagement, user growth, weekly retention, weekly engagement per device, and email engagement.

This project has helped me enhance my SQL skills, particularly in querying and manipulating data to derive meaningful insights. It has also improved my ability to interpret data and provide actionable recommendations based on the analysis. Overall, this project has deepened my understanding of user behavior analysis and its application in making informed decisions for product development and marketing strategies.

Hiring Process Analytics

Project Description:
This project is all about analyzing a company's data on people who applied for different positions in different departments. We'll be using statistics and Excel formulas to make sense of the information and draw important conclusions about the company.
We'll go through several steps to understand the data, check for missing values, group different categories together, spot any outliers, and summarize the data.

Here are the tasks we'll be working on:
1. Hiring: We'll figure out how many guys and girls got hired by the company.
2. Average Salary: We'll calculate the average salary offered by the company.
3. Class Intervals: We'll create groups based on salary ranges.
4. Charts and Plots: We'll make cool graphs like Pie Charts or Bar Graphs to show the percentage of people in different departments.
5. Charts: We'll use more graphs to show the different levels of job positions.

This project will give us important insights to help the company make decisions and improve their hiring process.

Approach:
In analyzing the dataset of a company's registrations for different posts in various departments, I followed a structured approach using Google Sheets. Here's how I tackled the project:
1. Hiring: I used the COUNTIF function in Google Sheets to determine the number of males and females hired by the company.
2. Average Salary: To calculate the average salary offered by the company, I utilized the AVERAGE function in Google Sheets.
3. Class Intervals: I utilized MIN and MAX function in Google Sheets to draw class intervals and FREQUENCY function to find the distribution.
4. Charts and Plots: Using the Insert Chart feature in Google Sheets, I created Pie Charts to show the proportion of people working in different departments.
5. Post Tiers: I utilized the AVERAGEIF and MAXIF functions in Google Sheets to categorize different job positions into their respective tiers based on specific conditions, which were then used to create charts or graphs.

By leveraging the functions in Google Sheets, as well as utilizing the charting features, I obtained valuable insights to support decision-making and enhance the company's hiring process.

Tech-Stack Used:

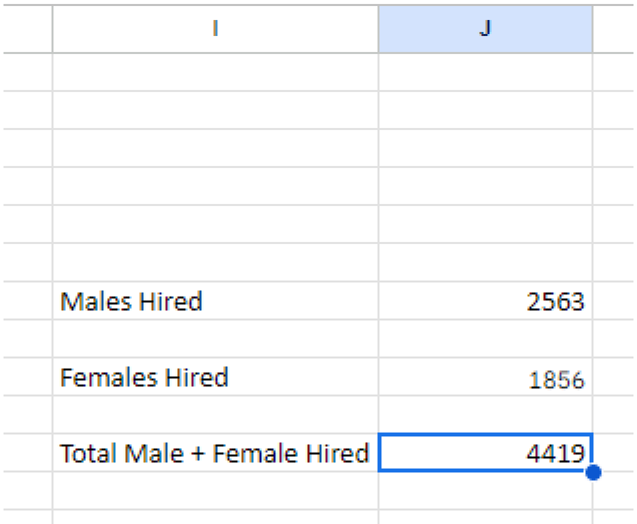
For this project, I utilized Google Sheets as the primary software tool. Google Sheets is a spreadsheet application included as part of the free, web-based Google Docs Editors suite offered by Google.

Insights:

1. ***Gender Distribution*** **:** It is important to analyze the number of males and females hired to gain insights into the gender diversity of the company. This information helps evaluate the company's efforts in promoting equality and inclusivity in its hiring process.

To find number of male employees and female employees following formulas were used:

```
=COUNTIFS(D2:D,"Male",C2:C,"Hired")
=COUNTIFS(D2:D,"Female",C2:C,"Hired")
=SUM(J7,J9)
```

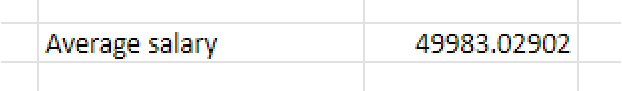


I	J
Males Hired	2563
Females Hired	1856
Total Male + Female Hired	4419

2. ***Salary Analysis*****:** Calculating the average salary offered by the company provides valuable information about the overall compensation provided to employees. This insight allows us to assess the company's competitiveness in terms of salary and understand the salary structure within the organization.

To calculate average salary of the employees we can use the following function:

```
=AVERAGE(G2:G)
```



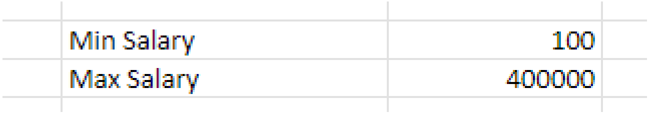
Average salary	49983.02902

As the G column contains the salary offered by the company, taking its average will give the average salary of the company.

***3. Salary Distribution Visualization*:** Drawing class intervals for salaries helps us group salary data into meaningful ranges. This visualization allows us to observe patterns and identify any concentration or gaps in salary levels across the company, giving us a better understanding of the salary distribution.

To find class intervals we have to find upper and lower limits. We can do that as following:

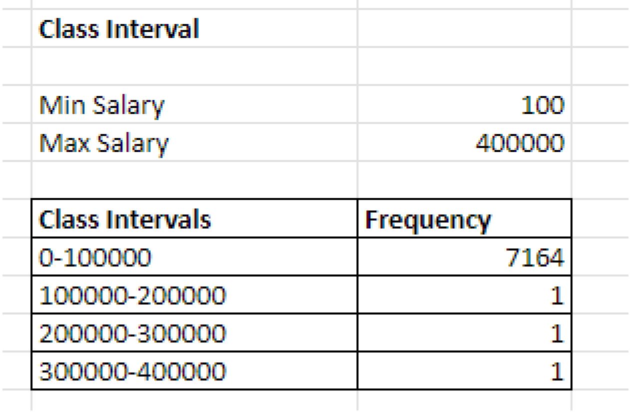
```
=MIN(G2:G)
=MAX(G2:G)
```



Min Salary	100
Max Salary	400000

We can now divide the classes into appropriate limits, and find number of elements in those classes by:

```
=FREQUENCY(G2:G,L8:L11)
```

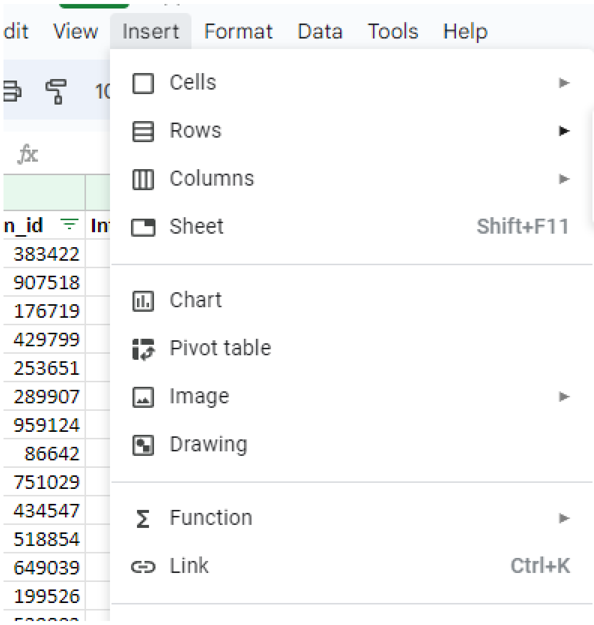


Class Interval	
Min Salary	100
Max Salary	400000

Class Intervals	Frequency
0-100000	7164
100000-200000	1
200000-300000	1
300000-400000	1

4. ***Charts and Plots***: Creating a Pie Chart, Bar Graph, or other graphical representations helps visualize the proportion of people working in different departments. This visual representation provides a clear overview of the departmental distribution within the company, enabling us to identify the size and composition of each department.

To Create a Pie Chart of different departments we can use the chart in the insert menu.



Then we select Data Range and select chart type:

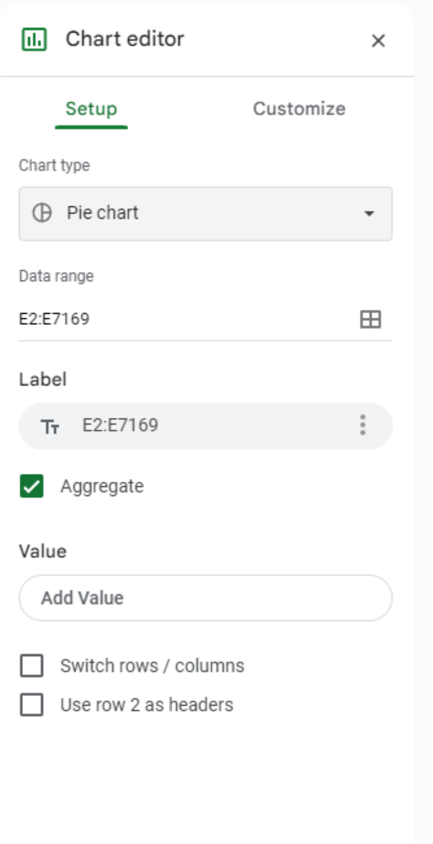


Chart type

Pie chart

Data range

E2:E7169

Label

E2:E7169

☑ Aggregate

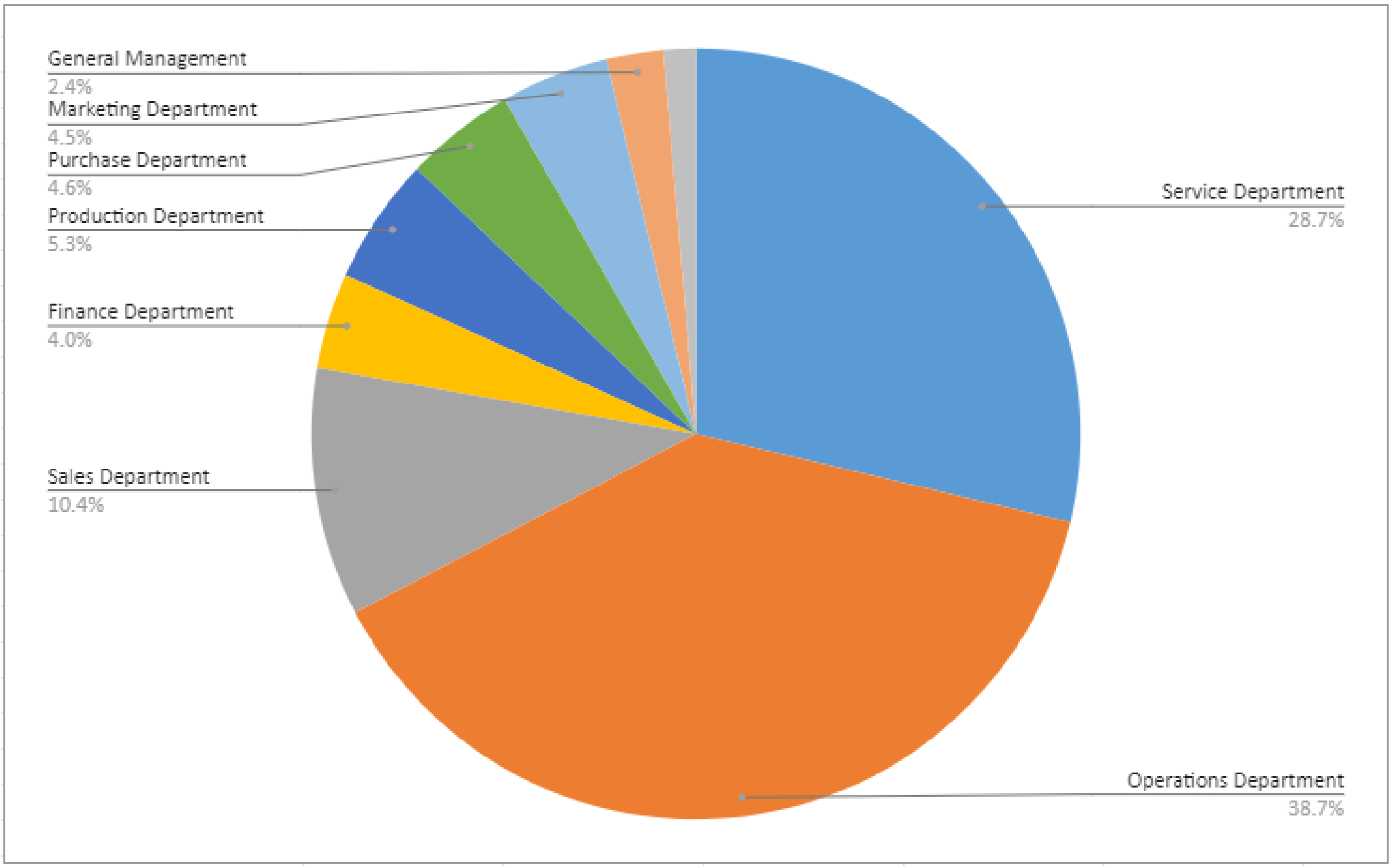
Value

Add Value

□ Switch rows / columns

□ Use row 2 as headers

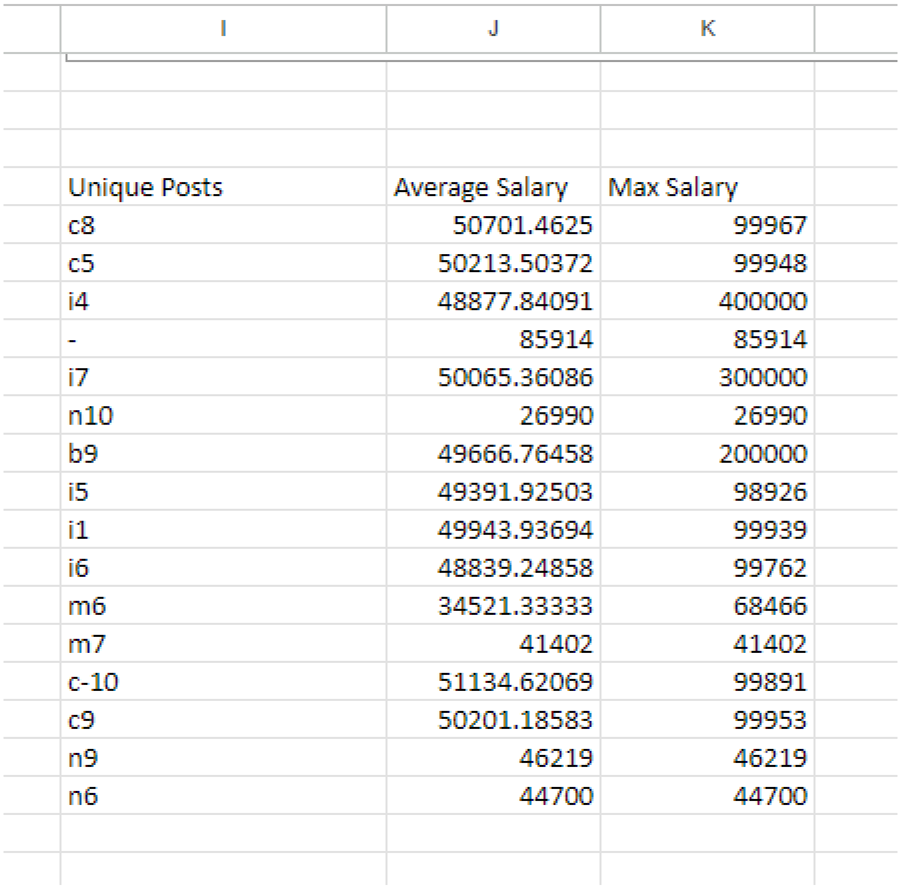
Department Pie Chart



5. *Visual Representation of Post Tiers*: Utilizing various charts and graphs, such as stacked bar graphs or grouped column charts, allows us to visually represent different post tiers within the company. This representation helps us compare job levels and understand the hierarchical structure of positions within the organization.

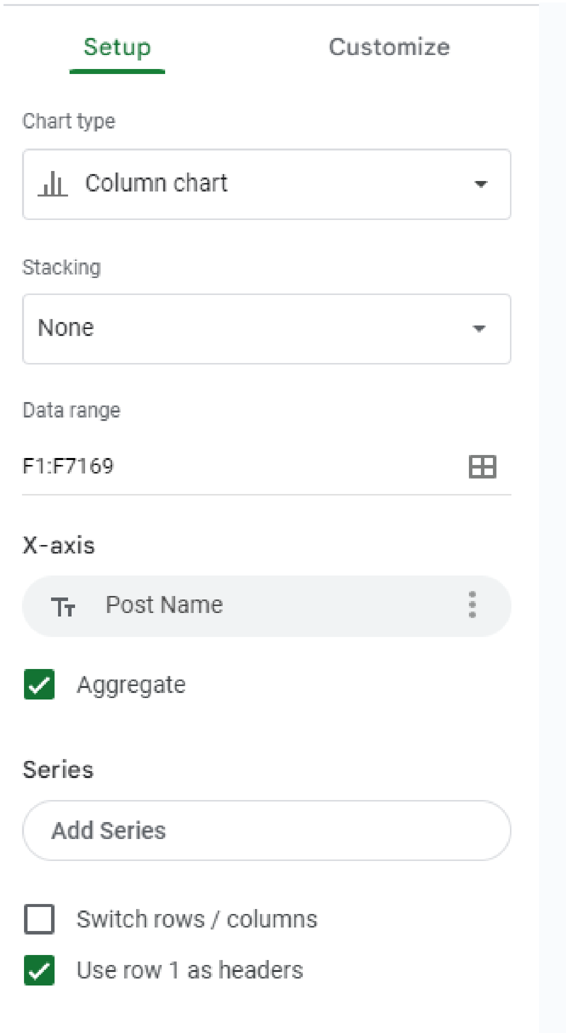
To plot charts and graphs of various post tiers, we have to find parameters for each post, such as average salary and maximum salary.
To find average salary and maximum salary we can use following formulae:

```
=AVERAGEIF(F2:F,I57,G2:G)
=MAXIFS(G2:G,F2:F,I57)
```



I	J	K
Unique Posts	Average Salary	Max Salary
c8	50701.4625	99967
c5	50213.50372	99948
i4	48877.84091	400000
-	85914	85914
i7	50065.36086	300000
n10	26990	26990
b9	49666.76458	200000
i5	49391.92503	98926
i1	49943.93694	99939
i6	48839.24858	99762
m6	34521.33333	68466
m7	41402	41402
c-10	51134.62069	99891
c9	50201.18583	99953
n9	46219	46219
n6	44700	44700

Now, to find how many jobs are there per post can be found using a column chart



Setup Customize

Chart type

Column chart

Stacking

None

Data range

F1:F7169

X-axis

Post Name

☑ Aggregate

Series

Add Series

☐ Switch rows / columns

☑ Use row 1 as headers

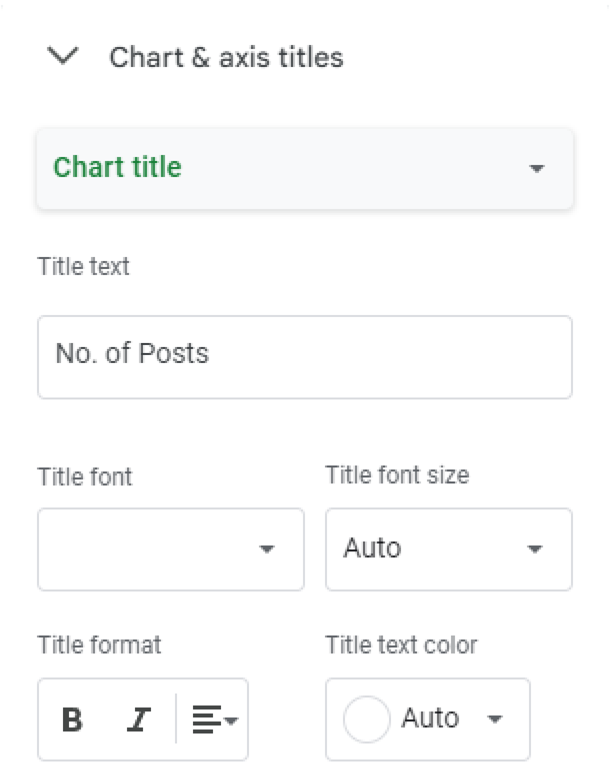


Chart & axis titles

Chart title

Title text

No. of Posts

Title font

Title font size

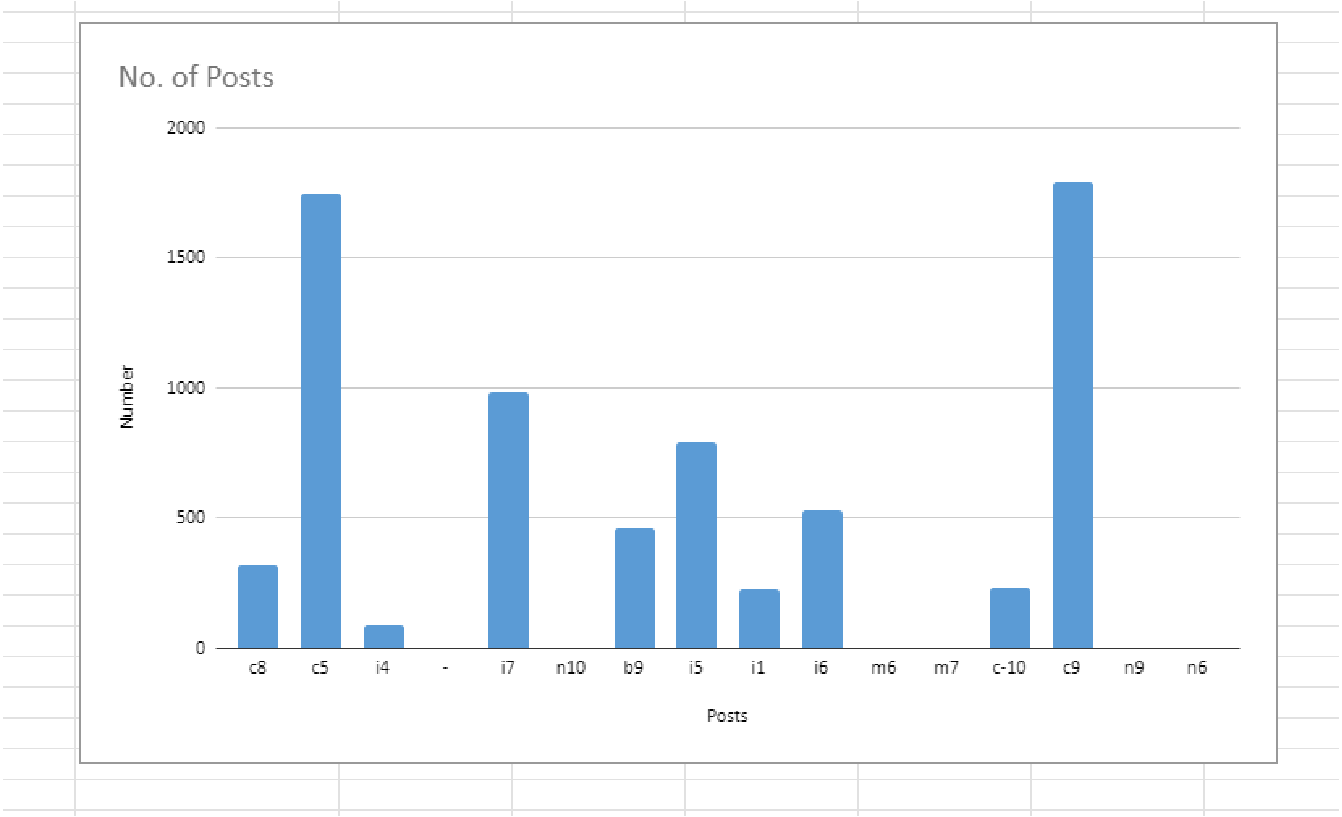
Auto

Title format

B *I*

Title text color

Auto



To find distribution of Average and Maximum salary we use data we derived earlier.

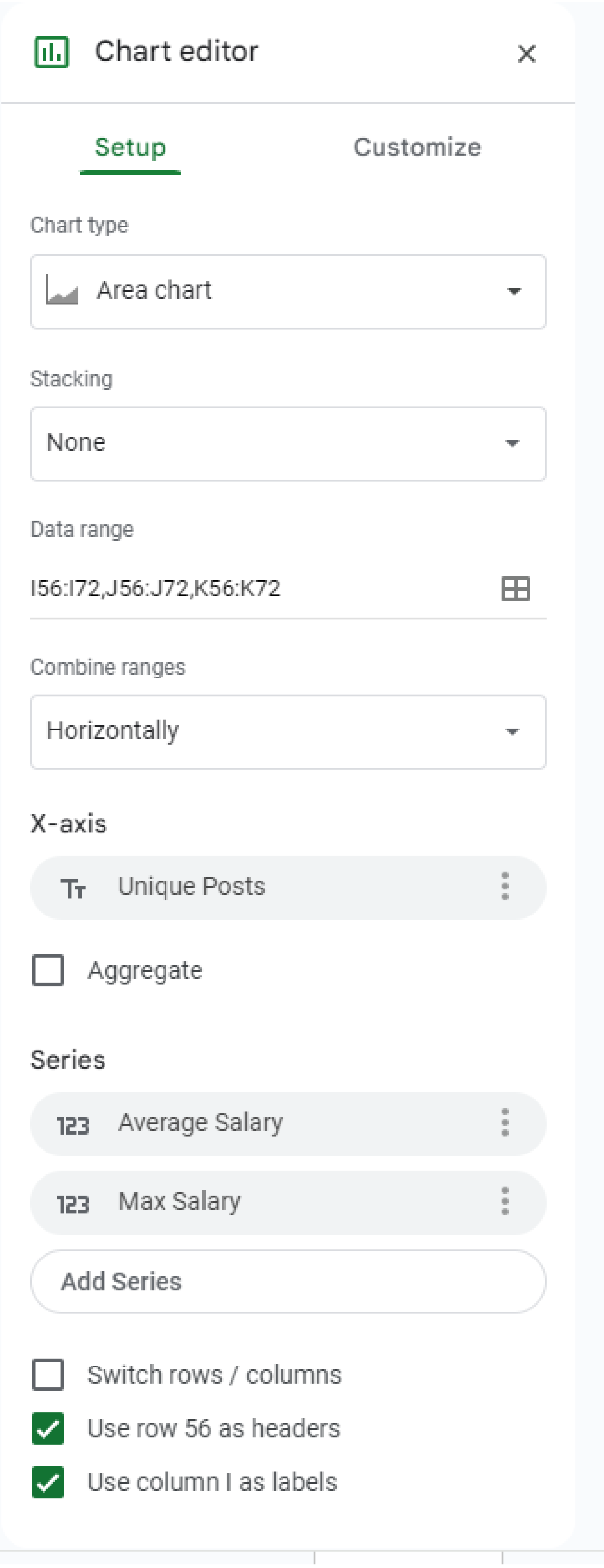


Chart editor

Setup | Customize

Chart type

Area chart

Stacking

None

Data range

I56:I72,J56:J72,K56:K72

Combine ranges

Horizontally

X-axis

Unique Posts

☐ Aggregate

Series

123 Average Salary

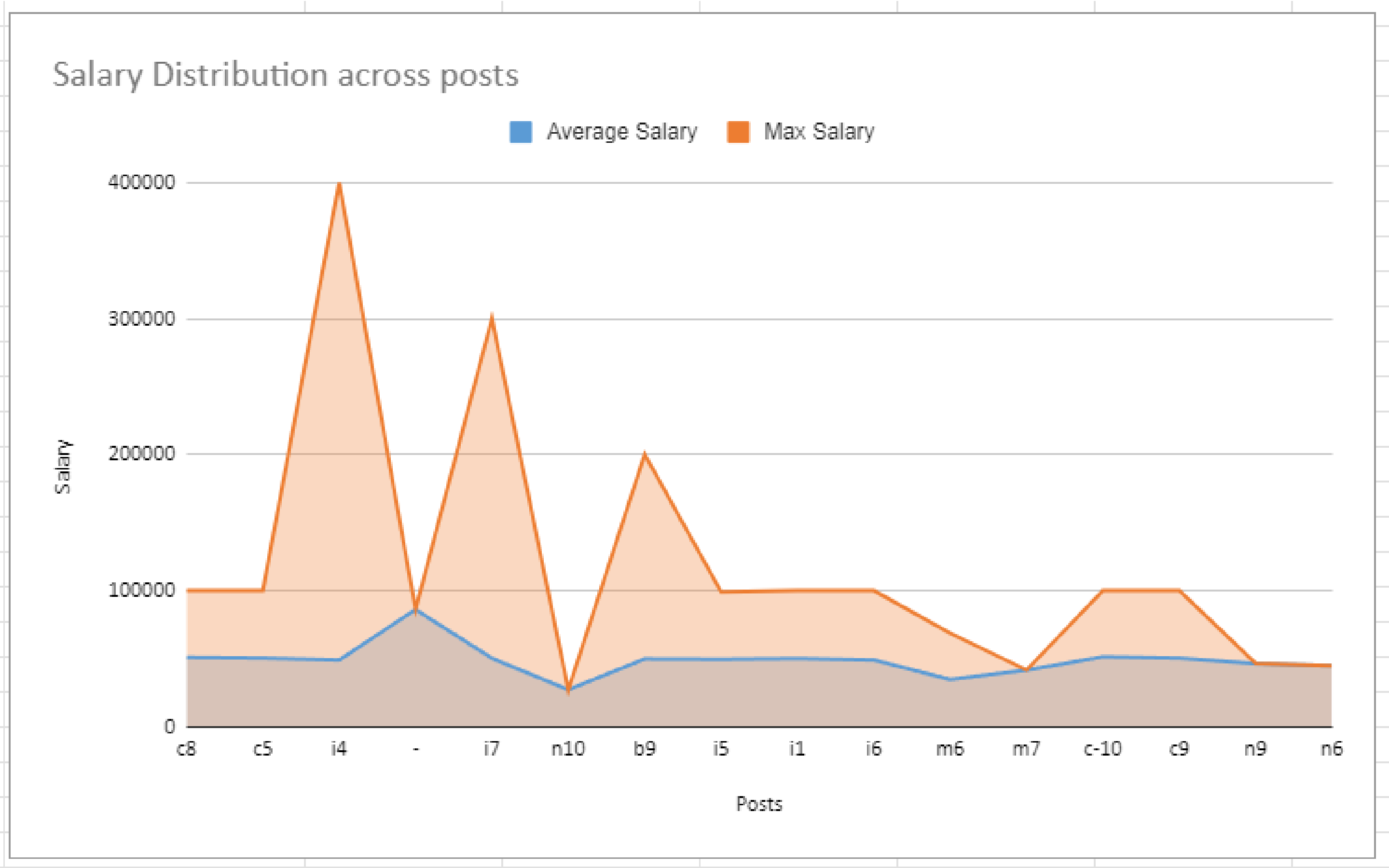
123 Max Salary

Add Series

☐ Switch rows / columns

☑ Use row 56 as headers

☑ Use column I as labels



This Chart gives Distribution of salary for posts, and makes it easier to understand where the difference between average salary and maximum salary is most.

Results:

While working on this project, I have gained a better understanding of hiring process analytics and Excel fundamentals. By analyzing hiring data, I was able to provide insights on various aspects such as gender gender distribution, average salary, salary distribution, Department size and human resource distribution, salary distribution across various posts.

This project has helped me enhance my Excel skills, particularly in functions and data visualization to derive meaningful insights. It has also improved my ability to interpret data and provide actionable recommendations based on the analysis.

IMDB Movie Analysis

Link for Google Sheets:

https://docs.google.com/spreadsheets/d/134qUS6AE1FmzjW2wKUemQOFuQD-CYimsf3JP27GIoPo/edit?usp=sharing

Project Description:

This project aims to analyze a dataset containing information about various movies from the IMDB database. The goal is to gain insights into different aspects of the movies, such as genre, duration, language, directors, and budgets, and their impact on the IMDB scores and financial success. By employing statistics and Excel formulas, we will extract meaningful conclusions to help understand the factors that contribute to a movie's popularity and success.

A. Movie Genre Analysis:
Task 1: Determine the most common genres of movies in the dataset.
Task 2: Calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores for each genre.

B. Movie Duration Analysis:
Task 1: Analyze the distribution of movie durations.
Task 2: Visualize the relationship between movie duration and IMDB score.

C. Language Analysis:
Task 1: Determine the most common languages used in movies.
Task 2: Analyze the impact of language on IMDB scores using descriptive statistics.

D. Director Analysis:
Task 1: Identify the top directors based on their average IMDB score.
Task 2: Analyze the contribution of top directors to the success of movies using percentile calculations.

E. Budget Analysis:
Task 1: Analyze the correlation between movie budgets and gross earnings.
Task 2: Identify the movies with the highest profit margin.

By completing the above tasks and analyzing the data using statistics and Excel formulas, we will gain valuable insights into the impact of movie genres, duration, language, directors, and budgets on IMDB scores and financial success. These findings will assist in making informed decisions to improve movie-making strategies and achieve greater popularity and profitability for future films.

Approach:

Data Engineering:

Genre column has multiple genres in the same column and can't be used directly to find distinct genre count and use each genre effectively. So we first split it to multiple columns by "split text to column" method and then list all distinct genres in a column for future use. We can use "=UNIQUE()" formula to find distinct elements in any column, and will do so in any future analysis.

A. Movie Genre Analysis:

We will use Excel's COUNTIF function to count the occurrences of each genre.

To Calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores for each genre. We will first need to manipulate the 'genres' column to separate multiple genres for a single movie. Then, we will use Excel's AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV functions to calculate the required statistics for each genre.

B. Movie Duration Analysis:

We will calculate descriptive statistics (mean, median, and standard deviation) for movie durations using Excel's functions.

We will create a scatter plot to visualize the relationship between movie duration and IMDB score. Additionally, we will add a trendline to assess the direction and strength of the relationship.

C. Language Analysis:

We will use Excel's COUNTIF function to count the number of movies for each language.

We will calculate the mean, median, and standard deviation of the IMDB scores for each language using Excel's functions.

D. Director Analysis:

We will calculate the average IMDB score for each director and use Excel's PERCENTILE function to identify the directors with the highest scores.

We will compare the scores of the top directors to the overall distribution of scores to assess their impact.

E. Budget Analysis:

We will calculate the correlation coefficient between movie budgets and gross earnings using Excel's CORREL function.

We will calculate the profit margin (gross earnings - budget) for each movie and use Excel's MAX function to identify the movies with the highest profit margin.

Tech-Stack Used:

For this project, I utilized Google Sheets as the primary software tool. Google Sheets is a spreadsheet application included as part of the free, web-based Google Docs Editors suite offered by Google.

Insights:

A. Movie Genre Analysis:

It is important to analyze the distribution of movies across different genres and understand relationships between genre and IMDB score, to predict what kind of movies audiences prefer. For this descriptive statistics can be used. For doing this following formulas were used.

To find distinct genres from column of genres, we first separated genres into multiple columns and then, made unique list of genres using following formula:

```
=UNIQUE(TOCOL(J:Q))
```

To count the number of movies per genre:

```
=COUNTIF(J:Q,AR3)
```

To find average IMDB Score per genre:

```
=ArrayFormula(AVERAGE(IF((J:J=AR3)+(K:K=AR3)+(L:L=AR3)+(M:M=AR3)+(N:N=AR3)+(O:O=AR3)+(P:P=AR3)+(Q:Q=AR3), AG:AG)))
```

To find median IMDB Score per genre:

```
=ArrayFormula(MEDIAN(IF((J:J=$AR3)+(K:K=$AR3)+(L:L=$AR3)+(M:M=$AR3)+(N:N=$AR3)+(O:O=$AR3)+(P:P=$AR3)+(Q:Q=$AR3), AG:AG)))
```

To find mode IMDB Score per genre:

```
=ArrayFormula(MODE(IF((J:J=$AR3)+(K:K=$AR3)+(L:L=$AR3)+(M:M=$AR3)+(N:N=$AR3)+(O:O=$AR3)+(P:P=$AR3)+(Q:Q=$AR3), AG:AG)))
```

To find minimum IMDB Score per genre:

```
=ArrayFormula(MIN(IF((J:J=$AR3)+(K:K=$AR3)+(L:L=$AR3)+(M:M=$AR3)+(N:N=$AR3)+(O:O=$AR3)+(P:P=$AR3)+(Q:Q=$AR3), AG:AG)))
```

To find maximum IMDB Score per genre:

```
=MAX(MAXIFS(AG:AG,J:J,AR3),MAXIFS(AG:AG,K:K,AR3),MAXIFS(AG:AG,L:L,AR3),MAXIFS(AG:AG,M:M,AR3),MAXIFS(AG:AG,N:N,AR3),MAXIFS(AG:AG,O:O,AR3),MAXIFS(AG:AG,P:P,AR3),MAXIFS(AG:AG,Q:Q,AR3))
```

To find variance in IMDB Score per genre:

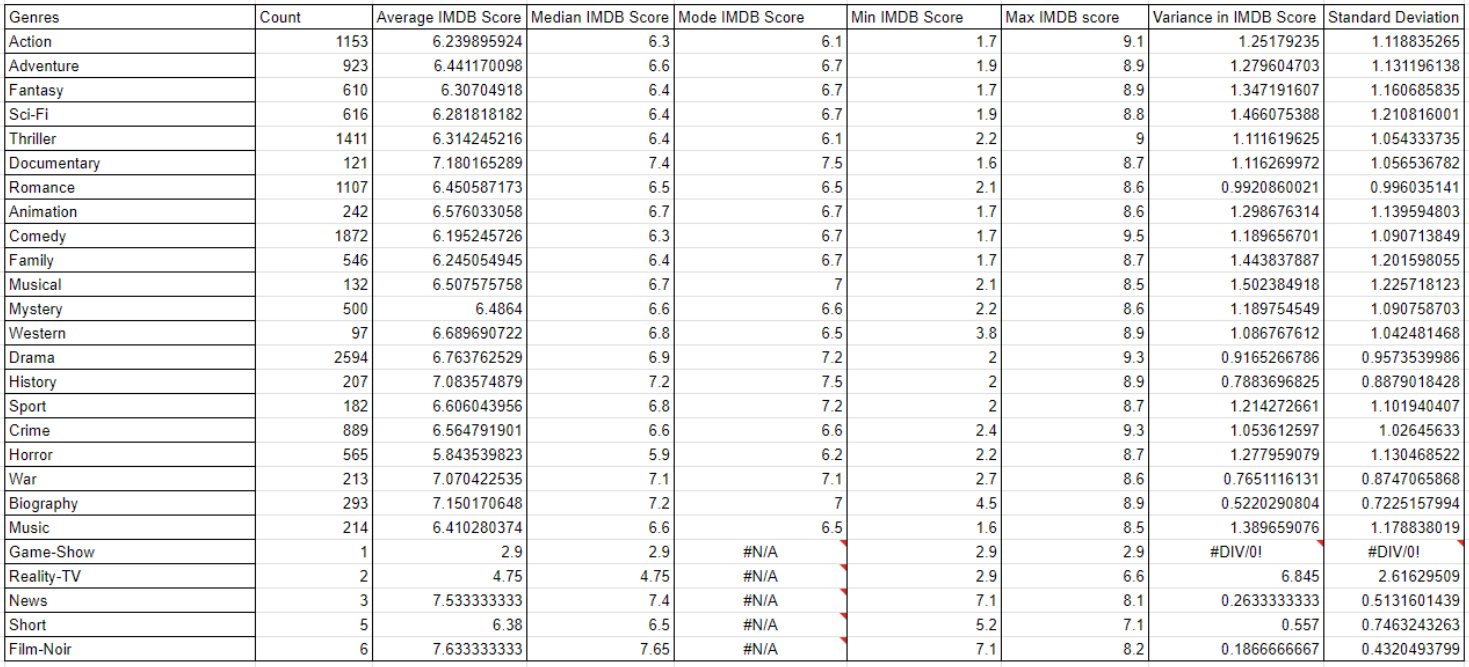
```
=ArrayFormula(VAR(IF((J:J=$AR3)+(K:K=$AR3)+(L:L=$AR3)+(M:M=$AR3)+(N:N=$AR3)+(O:O=$AR3)+(P:P=$AR3)+(Q:Q=$AR3), AG:AG)))
```

To find standard deviation in IMDB Score per genre:

```
=ArrayFormula(STDEV(IF((J:J=$AR3)+(K:K=$AR3)+(L:L=$AR3)+(M:M=$AR3)+(N:N=$AR3)+(O:O=$AR3)+(P:P=$AR3)+(Q:Q=$AR3), AG:AG)))
```

We can automatically generate complete table for each genre by autofilling formulas in google sheets

Output:



Genres	Count	Average IMDB Score	Median IMDB Score	Mode IMDB Score	Min IMDB Score	Max IMDB score	Variance in IMDB Score	Standard Deviation
Action	1153	6.239895924	6.3	6.1	1.7	9.1	1.25179235	1.118835265
Adventure	923	6.441170098	6.6	6.7	1.9	8.9	1.279604703	1.131196138
Fantasy	610	6.30704918	6.4	6.7	1.7	8.9	1.347191607	1.160685835
Sci-Fi	616	6.281818182	6.4	6.7	1.9	8.8	1.466075388	1.210816001
Thriller	1411	6.314245216	6.4	6.1	2.2	9	1.111619625	1.054333735
Documentary	121	7.180165289	7.4	7.5	1.6	8.7	1.116269972	1.056536782
Romance	1107	6.450587173	6.5	6.5	2.1	8.6	0.9920860021	0.996035141
Animation	242	6.576033058	6.7	6.7	1.7	8.6	1.298676314	1.139594803
Comedy	1872	6.195245726	6.3	6.7	1.7	9.5	1.189656701	1.090713849
Family	546	6.245054945	6.4	6.7	1.7	8.7	1.443837887	1.201598055
Musical	132	6.507575758	6.7	7	2.1	8.5	1.502384918	1.225718123
Mystery	500	6.4864	6.6	6.6	2.2	8.6	1.189754549	1.090758703
Western	97	6.689690722	6.8	6.5	3.8	8.9	1.086767612	1.042481468
Drama	2594	6.763762529	6.9	7.2	2	9.3	0.9165266786	0.9573539986
History	207	7.083574879	7.2	7.5	2	8.9	0.7883696825	0.8879018428
Sport	182	6.606043956	6.8	7.2	2	8.7	1.214272661	1.101940407
Crime	889	6.564791901	6.6	6.6	2.4	9.3	1.053612597	1.02645633
Horror	565	5.843539823	5.9	6.2	2.2	8.7	1.277959079	1.130468522
War	213	7.070422535	7.1	7.1	2.7	8.6	0.7651116131	0.8747065868
Biography	293	7.150170648	7.2	7	4.5	8.9	0.5220290804	0.7225157994
Music	214	6.410280374	6.6	6.5	1.6	8.5	1.389659076	1.178838019
Game-Show	1	2.9	2.9	#N/A	2.9	2.9	#DIV/0!	#DIV/0!
Reality-TV	2	4.75	4.75	#N/A	2.9	6.6	6.845	2.61629509
News	3	7.533333333	7.4	#N/A	7.1	8.1	0.2633333333	0.5131601439
Short	5	6.38	6.5	#N/A	5.2	7.1	0.557	0.7463243263
Film-Noir	6	7.633333333	7.65	#N/A	7.1	8.2	0.1866666667	0.4320493799

Note that Mode is N/A where each element appears only once and variance and standard deviance can’t be calculated for genre with single element.

B. Movie Duration Analysis:

To determine the ideal movie duration, that audience prefer is essential for a successful movie. So to find the relation between movie duration and IMDB Score we can use descriptive statistics to find average, median and standard deviation of movie duration. We can also create a scatterplot to better understand relationship and plot a trendline.

Formulae:

To find average duration of movies:

```
=AVERAGE(D:D)
```

To find median duration of movies:

```
=MEDIAN(D:D)
```

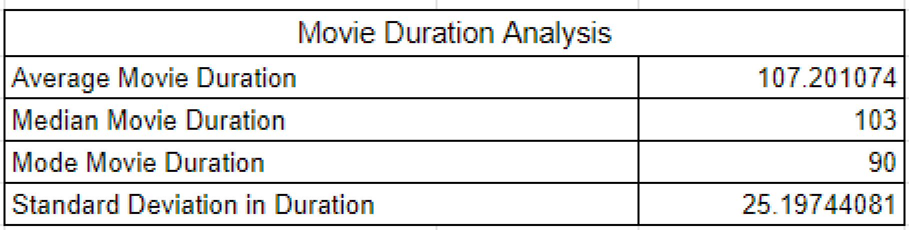
To find mode duration of movies:

```
=MODE(D:D)
```

To find standard deviation in duration of movies:

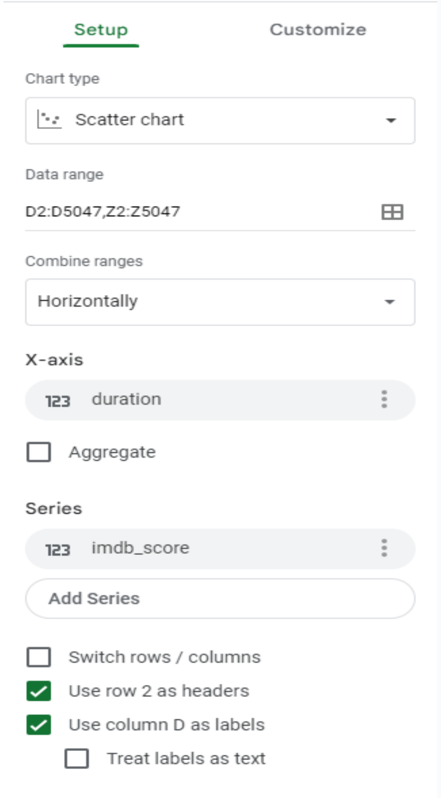
```
=STDEV(D:D)
```

Output:



Movie Duration Analysis	
Average Movie Duration	107.201074
Median Movie Duration	103
Mode Movie Duration	90
Standard Deviation in Duration	25.19744081

To Create a scatter plot of IMDB Score vs Movie duration we insert a chart and select scatterplot. We select data ranges and visualize the scatter plot. We add a trend line to better understand the relationship.



Setup | Customize

Chart type: Scatter chart

Data range: D2:D5047,Z2:Z5047

Combine ranges: Horizontally

X-axis: 123 duration

□ Aggregate

Series: 123 imdb_score

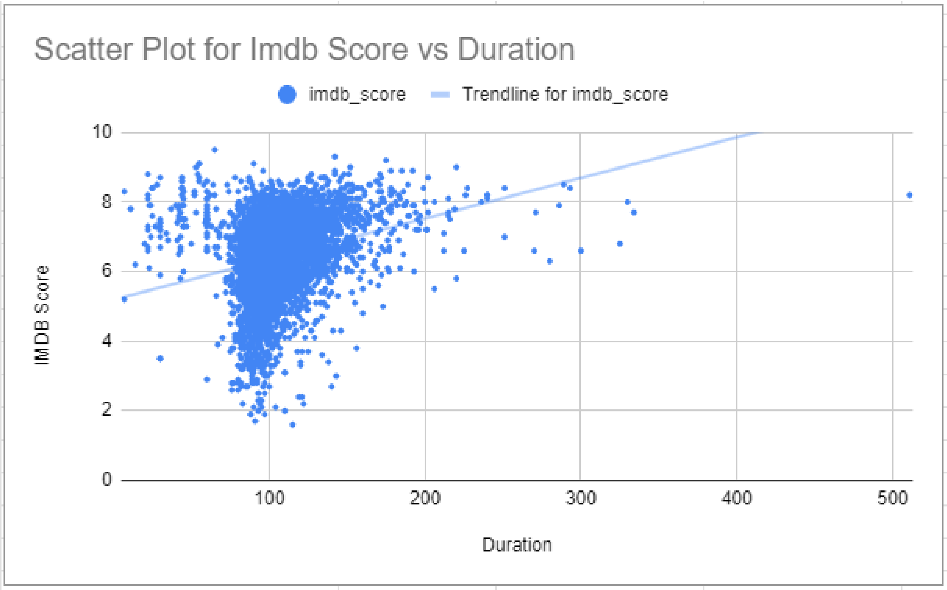
Add Series

□ Switch rows / columns

☑ Use row 2 as headers

☑ Use column D as labels

□ Treat labels as text



C. Language Analysis:

Determining which language the audience prefer to watch and where the majority of movies are successful is important for making a profit. That's why Descriptive statistics of Language and IMDB score are calculated.

Unique languages are found similarly unique genres are found, by using below formula:

```
=UNIQUE(AA3:AA)
```

To find count of movies in each language following formula is used:

```
=COUNTIF(AA:AA,AU68)
```

To find average IMDB Score per language:

```
=AVERAGEIF(AA:AA,AU68,AG:AG)
```

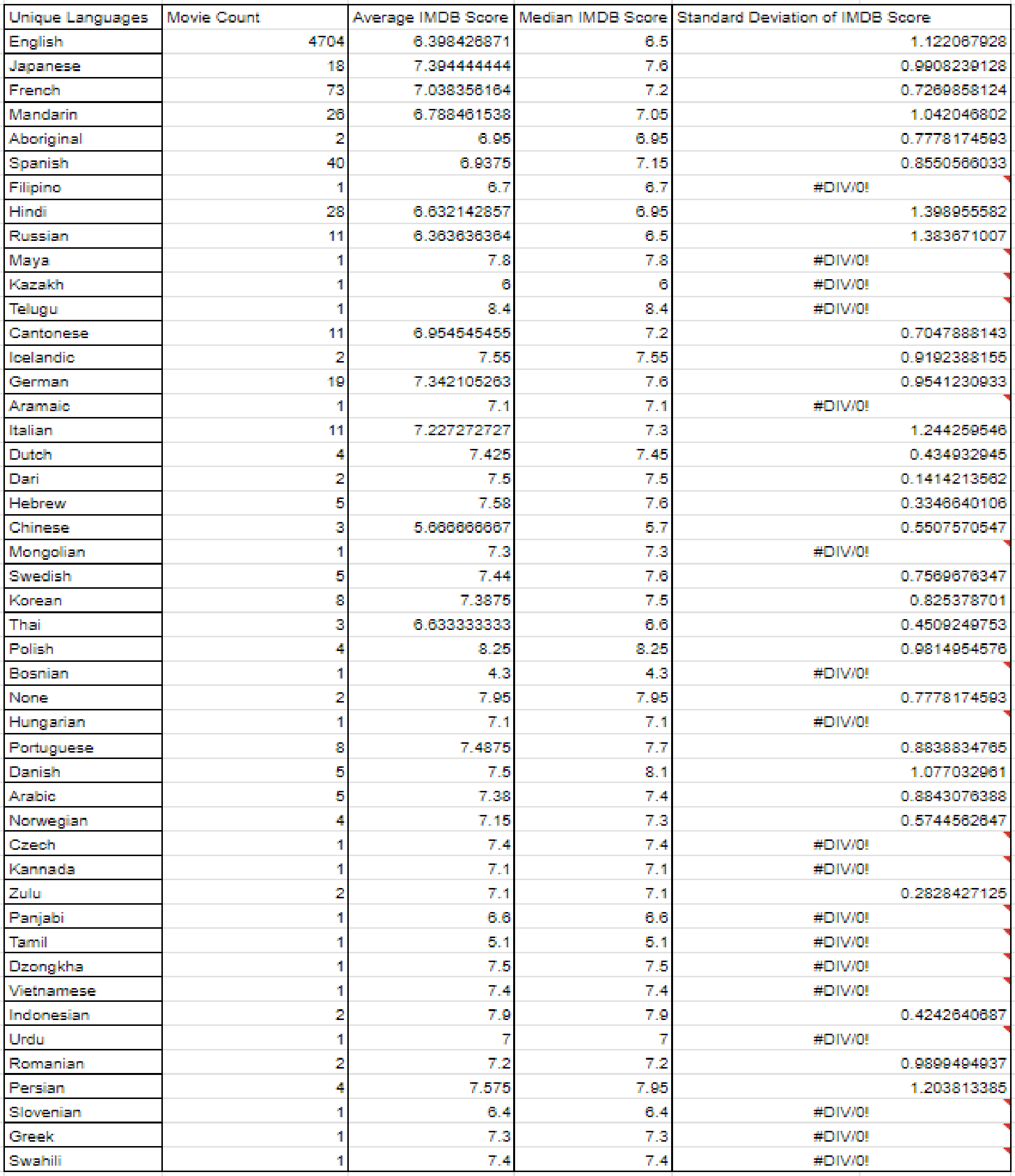
To find median IMDB Score per language:

```
=ArrayFormula(MEDIAN(if(AA:AA=$AU68,AG:AG)))
```

To find standard deviation of IMDB Score per language:

```
=ArrayFormula(STDEV(if(AA:AA=AU68,AG:AG)))
```

Output:



Unique Languages	Movie Count	Average IMDB Score	Median IMDB Score	Standard Deviation of IMDB Score
English	4704	6.398426871	6.5	1.122067928
Japanese	18	7.394444444	7.6	0.9908239128
French	73	7.038356164	7.2	0.7269858124
Mandarin	26	6.788461538	7.05	1.042046802
Aboriginal	2	6.95	6.95	0.7778174593
Spanish	40	6.9375	7.15	0.8550566033
Filipino	1	6.7	6.7	#DIV/0!
Hindi	28	6.632142857	6.95	1.398955582
Russian	11	6.363636364	6.5	1.383671007
Maya	1	7.8	7.8	#DIV/0!
Kazakh	1	6	6	#DIV/0!
Telugu	1	8.4	8.4	#DIV/0!
Cantonese	11	6.954545455	7.2	0.7047888143
Icelandic	2	7.55	7.55	0.9192388155
German	19	7.342105263	7.6	0.9541230933
Aramaic	1	7.1	7.1	#DIV/0!
Italian	11	7.227272727	7.3	1.244259546
Dutch	4	7.425	7.45	0.434932945
Dari	2	7.5	7.5	0.1414213562
Hebrew	5	7.58	7.6	0.3346640106
Chinese	3	5.666666667	5.7	0.5507570547
Mongolian	1	7.3	7.3	#DIV/0!
Swedish	5	7.44	7.6	0.7569676347
Korean	8	7.3875	7.5	0.825378701
Thai	3	6.633333333	6.6	0.4509249753
Polish	4	8.25	8.25	0.9814954576
Bosnian	1	4.3	4.3	#DIV/0!
None	2	7.95	7.95	0.7778174593
Hungarian	1	7.1	7.1	#DIV/0!
Portuguese	8	7.4875	7.7	0.8838834765
Danish	5	7.5	8.1	1.077032961
Arabic	5	7.38	7.4	0.8843076388
Norwegian	4	7.15	7.3	0.5744562647
Czech	1	7.4	7.4	#DIV/0!
Kannada	1	7.1	7.1	#DIV/0!
Zulu	2	7.1	7.1	0.2828427125
Panjabi	1	6.6	6.6	#DIV/0!
Tamil	1	5.1	5.1	#DIV/0!
Dzongkha	1	7.5	7.5	#DIV/0!
Vietnamese	1	7.4	7.4	#DIV/0!
Indonesian	2	7.9	7.9	0.4242640687
Urdu	1	7	7	#DIV/0!
Romanian	2	7.2	7.2	0.9899494937
Persian	4	7.575	7.95	1.203813385
Slovenian	1	6.4	6.4	#DIV/0!
Greek	1	7.3	7.3	#DIV/0!
Swahili	1	7.4	7.4	#DIV/0!

Note: Standard deviation of a single movie in a language can't be calculated.

D. Director Analysis:

Director of a movie plays a major role in the popularity of a movie, so finding popular directors with the most IMDB rating is detrimental in finding which movies are going to make it big in the market. To find top directors we have to first find average IMDB rating per director, and then we find top 1% directors by using "PERCENTILE" function in google sheets

Formulae:

To find all unique directors:

```
=QUERY(B3:B, "SELECT B WHERE B <> '' AND B IS NOT NULL", 0)
```

To find average IMDB score per directors:

```
=iferror(AVERAGEIF(B3:B,AR121,AG:AG),AG:AG)
```

To find value at 99%le of IMDB score:

```
=PERCENTILE(AS121:AS,99%)
```

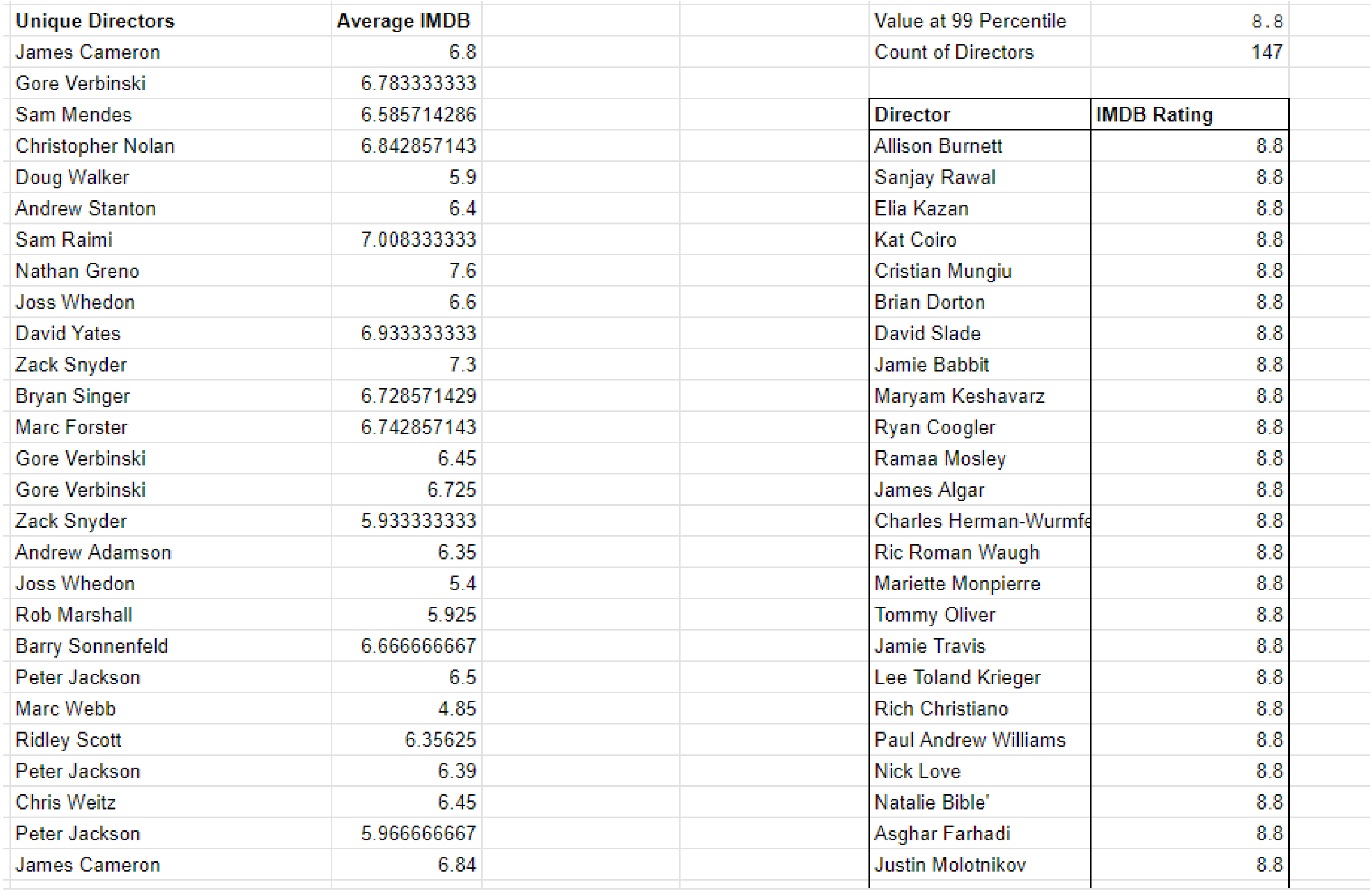
To count of top 1% directors:

```
=COUNTIF(AS121:AS, ">= "&PERCENTILE(AS121:AS, 99%))
```

To find list of top 1% directors:

```
=ARRAYFORMULA(FILTER(AR121:AS, AS121:AS >= AW120))
```

Output:



Unique Directors	Average IMDB
James Cameron	6.8
Gore Verbinski	6.783333333
Sam Mendes	6.585714286
Christopher Nolan	6.842857143
Doug Walker	5.9
Andrew Stanton	6.4
Sam Raimi	7.008333333
Nathan Greno	7.6
Joss Whedon	6.6
David Yates	6.933333333
Zack Snyder	7.3
Bryan Singer	6.728571429
Marc Forster	6.742857143
Gore Verbinski	6.45
Gore Verbinski	6.725
Zack Snyder	5.933333333
Andrew Adamson	6.35
Joss Whedon	5.4
Rob Marshall	5.925
Barry Sonnenfeld	6.666666667
Peter Jackson	6.5
Marc Webb	4.85
Ridley Scott	6.35625
Peter Jackson	6.39
Chris Weitz	6.45
Peter Jackson	5.966666667
James Cameron	6.84

Value at 99 Percentile	8.8
Count of Directors	147

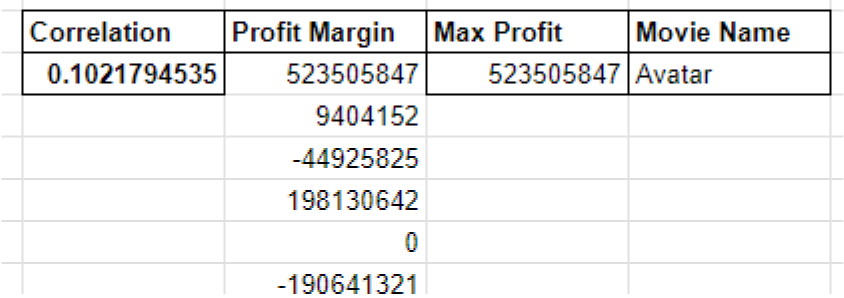
Director	IMDB Rating
Allison Burnett	8.8
Sanjay Rawal	8.8
Elia Kazan	8.8
Kat Coiro	8.8
Cristian Mungiu	8.8
Brian Dorton	8.8
David Slade	8.8
Jamie Babbit	8.8
Maryam Keshavarz	8.8
Ryan Coogler	8.8
Ramaa Mosley	8.8
James Algar	8.8
Charles Herman-Wurmfe	8.8
Ric Roman Waugh	8.8
Mariette Monpierre	8.8
Tommy Oliver	8.8
Jamie Travis	8.8
Lee Toland Krieger	8.8
Rich Christiano	8.8
Paul Andrew Williams	8.8
Nick Love	8.8
Natalie Bible'	8.8
Asghar Farhadi	8.8
Justin Molotnikov	8.8

These lists are very long and only some part of it is shown in output.

E. Budget Analysis:

The most important factor in determining the success of a movie is if it made a profit or not.
Profit can be defined as the subtraction of budget from gross income it made at box office.
We can also find correlation coefficient to find how likely it is to create a profit by setting an effective budget.
Also It would help to find the movie which made most profit to learn from it, to produce more successful movies.
Formulae:
To find correlation between budget and gross profit:
=CORREL(AD3:AD,I3:I)
To find profit by subtracting budget from gross:
=ArrayFormula((I3:I-AD3:AD))
To find most profit made by a movie:
=MAX(AL:AL)
To find movie with most profit:
=if(AM3=AL:AL,S:S)

Output:



Correlation	Profit Margin	Max Profit	Movie Name
0.1021794535	523505847	523505847	Avatar
	9404152		
	-44925825		
	198130642		
	0		
	-190641321		

Note: Output is very large; only the first few lines are shown for the profit margin column.

Results:

While working on this project, I have gained a better understanding of IMDB movie analysis and Advanced Excel methodologies. By analyzing movie data, I was able to provide insights on various aspects such as genre distribution and relation with IMDB score, relation between movie duration and score and visualization, language and IMDB Score ,Impacts of popularity of director on movie, and determining profit from budget and gross income.

This project has helped me enhance my Excel skills, particularly in functions and data visualization to derive meaningful insights. It has also improved my ability to interpret data and provide actionable recommendations based on the analysis.

Bank Loan Case Study

Link For Excel Sheet:

https://docs.google.com/spreadsheets/d/1mlateqxKvMcygMzQsjjxkTRCFzyoXbFu/edit?usp=sharing&ouid=107365393175079460343&rtpof=true&sd=true

Excel file contains different worksheets which have results of different tasks in them.

Project Description:

This project aims to analyze a dataset containing information about various bank loan applications. The goal is to gain insights about approval of bank loans, such as the relation between income and credit. The data provided has various missing or null Data, our task is to handle those missing values appropriately, by either deleting or imputing these data. There are various outliers in data, we have to find these outliers. We also have to check for data imbalance and perform various analyses on data, such as univariate and bivariate analysis. Finding correlation between various parameters would help us understand what factors affect most in bank loan application approval. Thus, by employing statistics and Excel formulas, we will extract meaningful conclusions to help understand the factors that contribute to a bank loan getting approved.

Approach:

As an individual working on this project, I followed a structured approach to analyze data about bank loan applications. I began by carefully examining the provided database and familiarizing myself with its structure and columns. I tried to find columns which had the most significance in the dataset. I handled missing values by eliminating columns which had most empty cells, and were not significant. And imputed data into cells that were necessary for analysis. Then, I utilized Excel fundamentals to retrieve the necessary information for each task, employing appropriate functions and statistical methods. I focused on data accuracy and quality throughout the project, ensuring reliable results. By leveraging my Excel skills and maintaining a systematic workflow, I successfully executed the project and created a comprehensive report that fulfilled the objectives of providing marketing insights and investor metrics.

Tech-Stack Used:

For this project, I utilized Microsoft Excel as the primary software tool.

Insights:

Task 1:

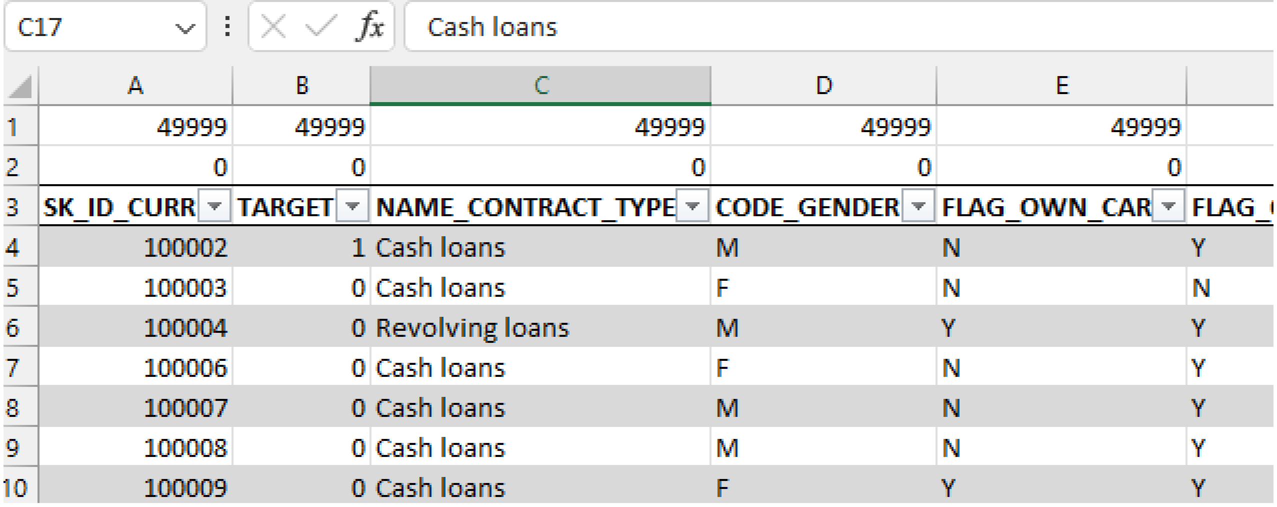
Identify Missing Data and Deal with it Appropriately (Data Cleaning):

To find data having missing values we utilized COUNTA function in Excel, which returns no. of cells which are not blank.

Formula:

=COUNTA(A4:A50002)

This gave us the number of rows in the TARGET column, which is the total number of rows which we have to consider for analysis.



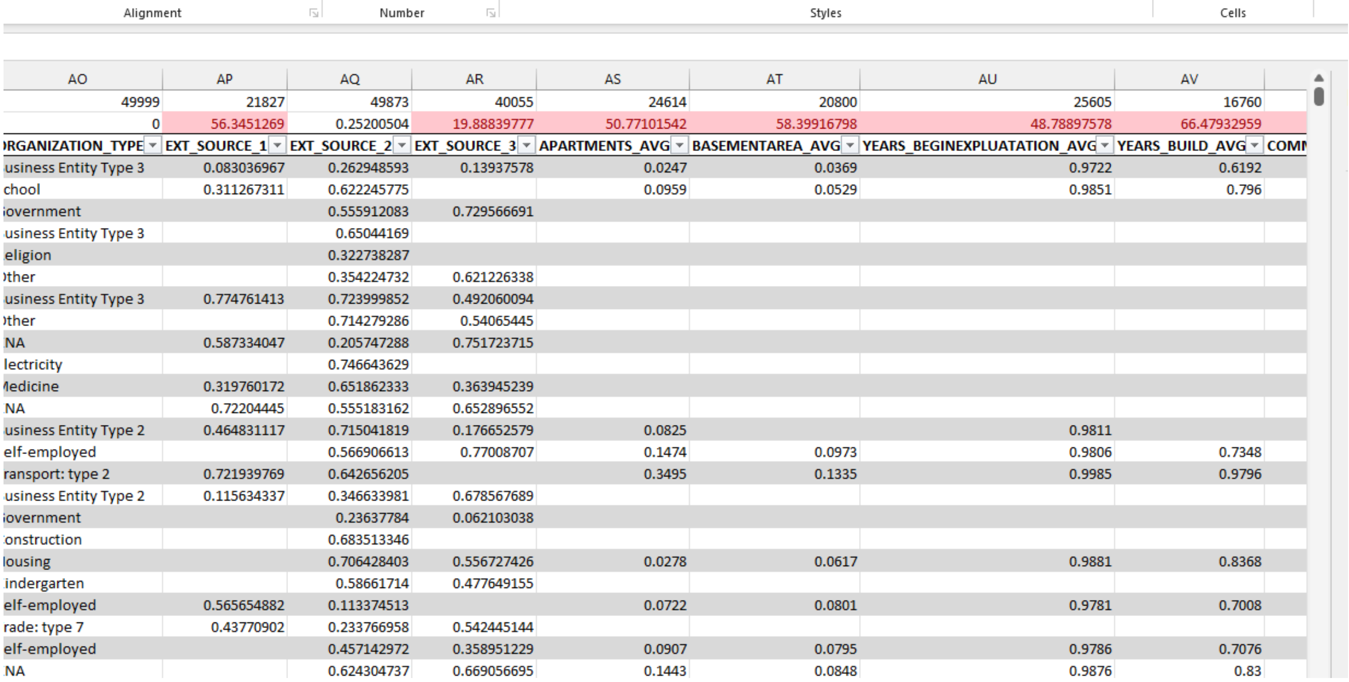
C17 | Cash loans

	A	B	C	D	E	
1	49999	49999	49999	49999	49999	
2	0	0	0	0	0	
3	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_
4	100002	1	Cash loans	M	N	Y
5	100003	0	Cash loans	F	N	N
6	100004	0	Revolving loans	M	Y	Y
7	100006	0	Cash loans	F	N	Y
8	100007	0	Cash loans	M	N	Y
9	100008	0	Cash loans	M	N	Y
10	100009	0	Cash loans	F	Y	Y

The columns which had missing data in them were found out by using the formula:

=(100-(V1/$A1)*100)

This formula gives us the percentage of missing values in the column.

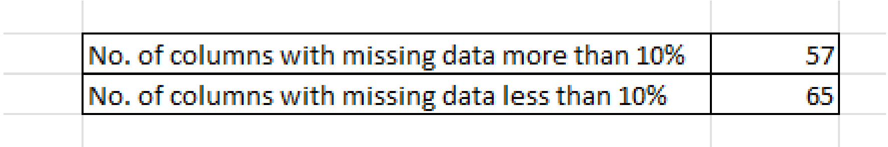


AO	AP	AQ	AR	AS	AT	AU	AV	
49999	21827	49873	40055	24614	20800	25605	16760	
0	56.3451269	0.25200504	19.88839777	50.77101542	58.39916798	48.78897578	66.47932959	
RGANIZATION_TYPE	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3	APARTMENTS_AVG	BASEMENTAREA_AVG	YEARS_BEGINEXPLUATATION_AVG	YEARS_BUILD_AVG	COMM
usiness Entity Type 3	0.083036967	0.262948593	0.13937578	0.0247	0.0369	0.9722	0.6192	
chool	0.311267311	0.622245775		0.0959	0.0529	0.9851	0.796	
iovernment		0.555912083	0.729566691					
usiness Entity Type 3		0.65044169						
eligion		0.322738287						
)ther		0.354224732	0.621226338					
usiness Entity Type 3	0.774761413	0.723999852	0.492060094					
)ther		0.714279286	0.54065445					
NA	0.587334047	0.205747288	0.751723715					
lectricity		0.746643629						
ledicine	0.319760172	0.651862333	0.363945239					
NA	0.72204445	0.555183162	0.652896552					
usiness Entity Type 2	0.464831117	0.715041819	0.176652579	0.0825		0.9811		
elf-employed		0.566906613	0.77008707	0.1474	0.0973	0.9806	0.7348	
ransport: type 2	0.721939769	0.642656205		0.3495	0.1335	0.9985	0.9796	
usiness Entity Type 2	0.115634337	0.346633981	0.678567689					
iovernment		0.23637784	0.062103038					
onstruction		0.683513346						
lousing		0.706428403	0.556727426	0.0278	0.0617	0.9881	0.8368	
indergarten		0.58661714	0.477649155					
elf-employed	0.565654882	0.113374513		0.0722	0.0801	0.9781	0.7008	
rade: type 7	0.43770902	0.233766958	0.542445144					
elf-employed		0.457142972	0.358951229	0.0907	0.0795	0.9786	0.7076	
NA		0.624304737	0.669056695	0.1443	0.0848	0.9876	0.83	

We highlighted columns with missing values by using conditional formatting.

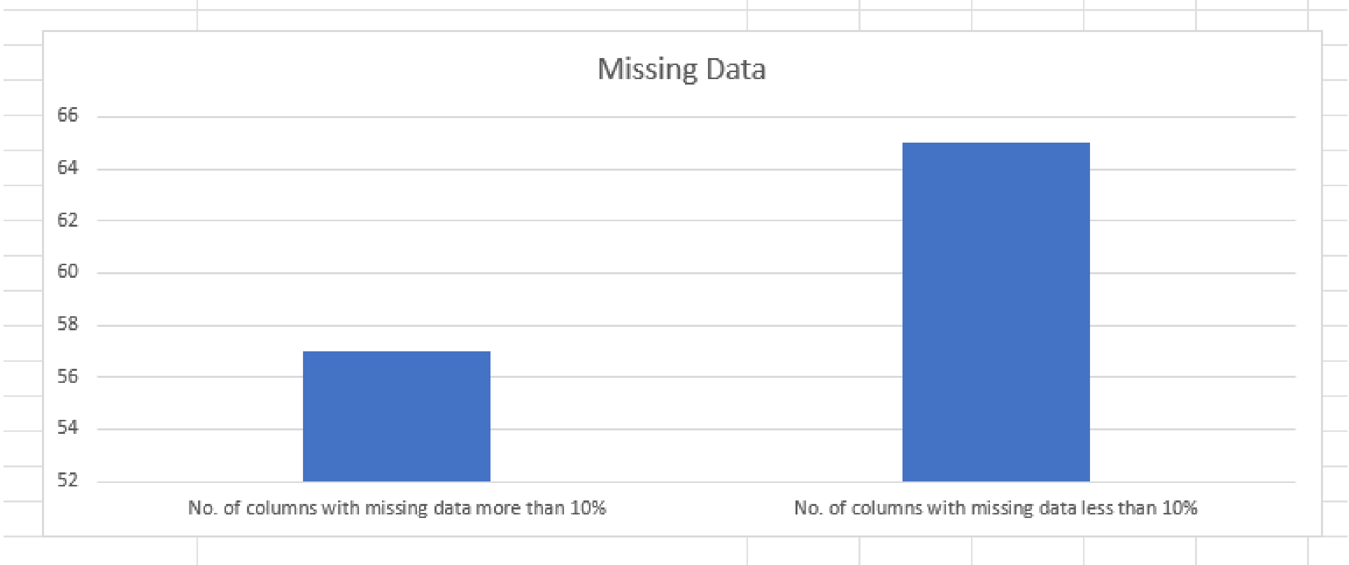
We found no. of columns with missing values greater than 10% by this formula:

=COUNTIF(A2:DR2,">10")
=COUNTIF(A2:DR2,"<10")

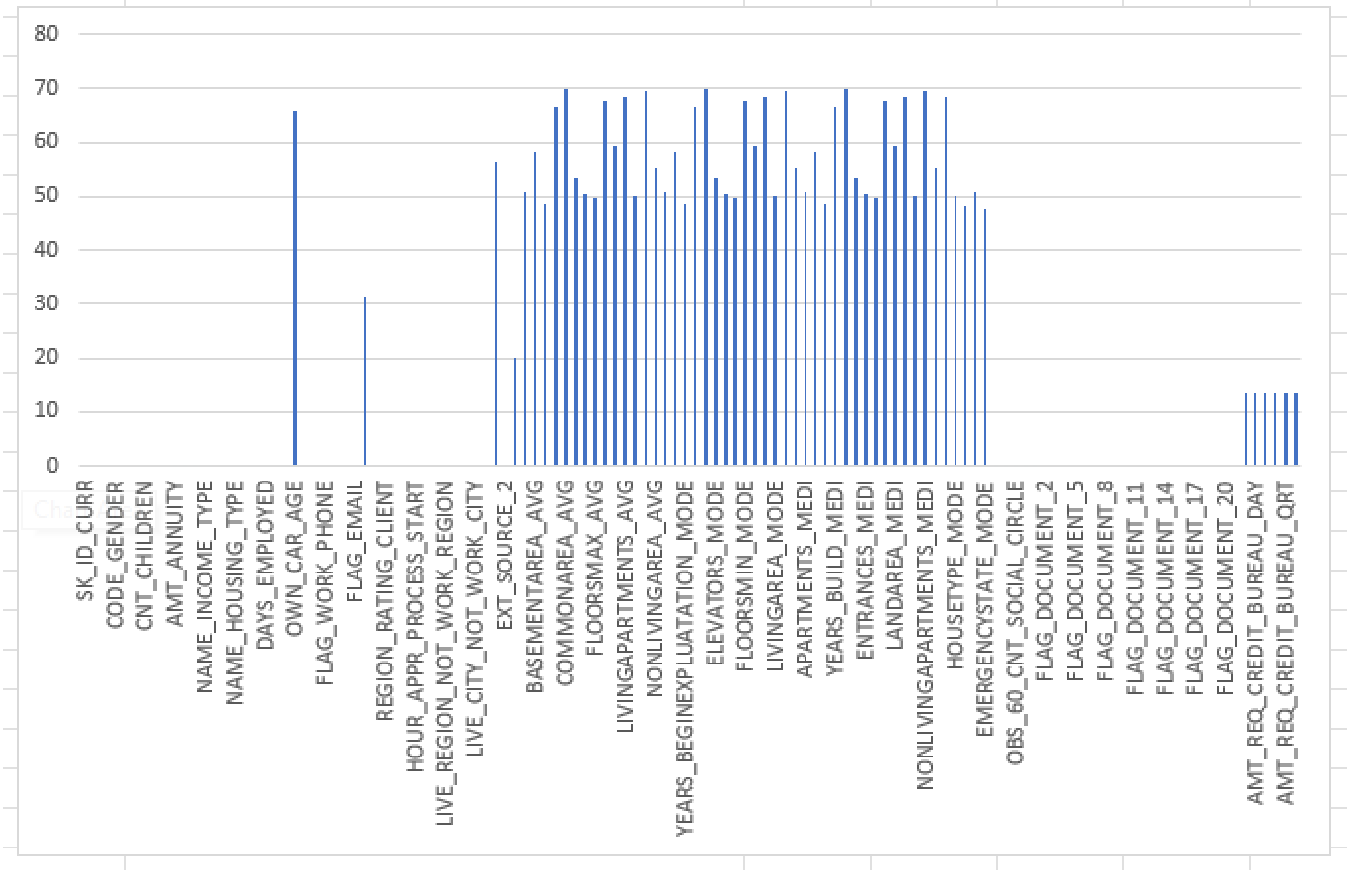


No. of columns with missing data more than 10%	57
No. of columns with missing data less than 10%	65

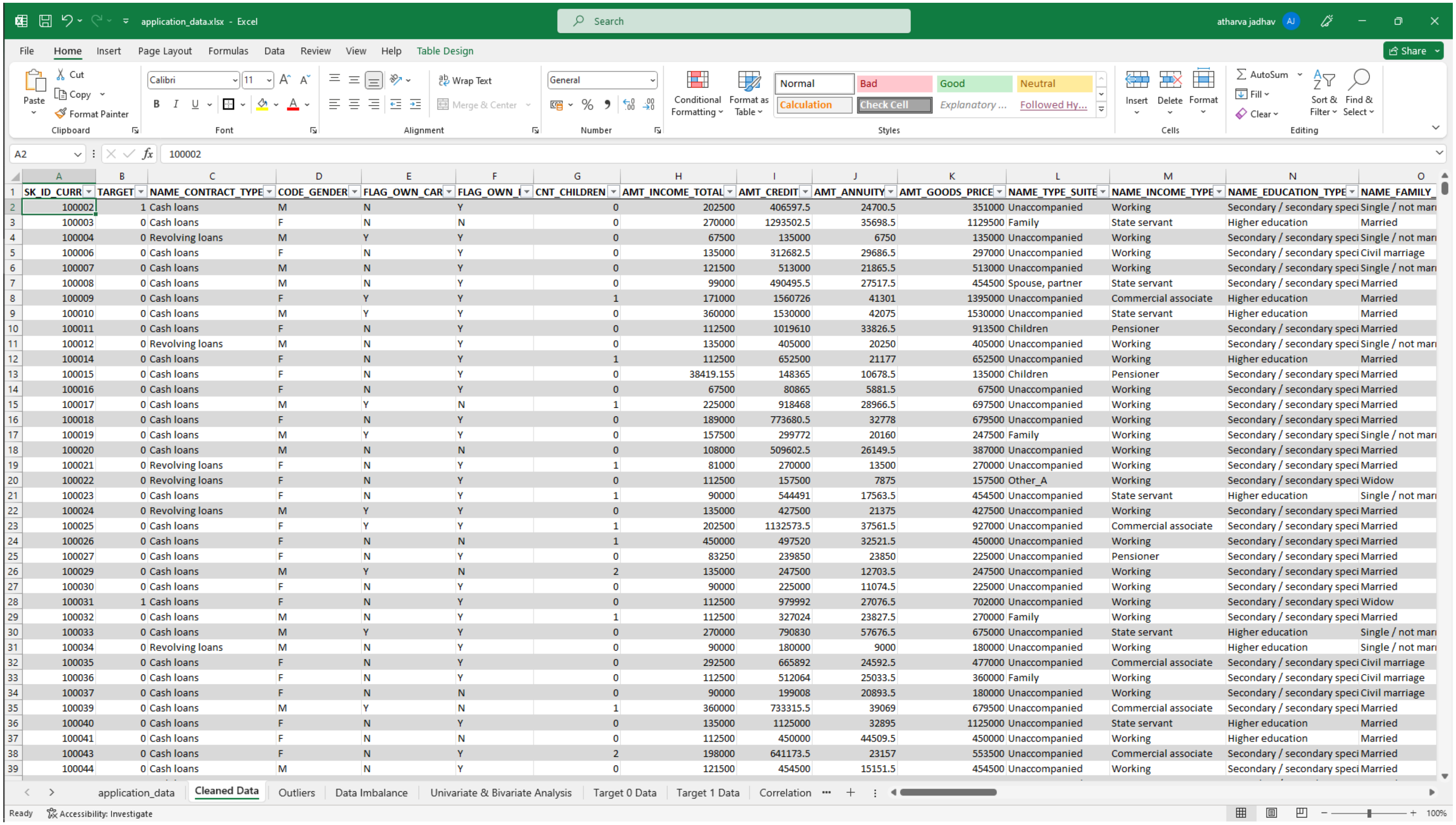
We plotted a bar graph to better understand the number of columns containing missing values



To Visualize columns and their respective missing values.



We saved all 65 columns with least missing values in a new worksheet called Cleaned Data.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_I	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_
2	100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied	Working	Secondary / secondary speci	Single / not marr
3	100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State servant	Higher education	Married
4	100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied	Working	Secondary / secondary speci	Single / not marr
5	100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied	Working	Secondary / secondary speci	Civil marriage
6	100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied	Working	Secondary / secondary speci	Single / not marr
7	100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner	State servant	Secondary / secondary speci	Married
8	100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied	Commercial associate	Higher education	Married
9	100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied	State servant	Higher education	Married
10	100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Secondary / secondary speci	Married
11	100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied	Working	Secondary / secondary speci	Single / not marr
12	100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied	Working	Higher education	Married
13	100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children	Pensioner	Secondary / secondary speci	Married
14	100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied	Working	Secondary / secondary speci	Married
15	100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied	Working	Secondary / secondary speci	Married
16	100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied	Working	Secondary / secondary speci	Married
17	100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family	Working	Secondary / secondary speci	Single / not marr
18	100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied	Working	Secondary / secondary speci	Married
19	100021	0	Revolving loans	F	N	Y	1	81000	270000	13500	270000	Unaccompanied	Working	Secondary / secondary speci	Married
20	100022	0	Revolving loans	F	N	Y	0	112500	157500	7875	157500	Other_A	Working	Secondary / secondary speci	Widow
21	100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied	State servant	Higher education	Single / not marr
22	100024	0	Revolving loans	M	Y	Y	0	135000	427500	21375	427500	Unaccompanied	Working	Secondary / secondary speci	Married
23	100025	0	Cash loans	F	Y	Y	1	202500	1132573.5	37561.5	927000	Unaccompanied	Commercial associate	Secondary / secondary speci	Married
24	100026	0	Cash loans	F	N	N	1	450000	497520	32521.5	450000	Unaccompanied	Working	Secondary / secondary speci	Married
25	100027	0	Cash loans	F	N	Y	0	83250	239850	23850	225000	Unaccompanied	Pensioner	Secondary / secondary speci	Married
26	100029	0	Cash loans	M	Y	N	2	135000	247500	12703.5	247500	Unaccompanied	Working	Secondary / secondary speci	Married
27	100030	0	Cash loans	F	N	Y	0	90000	225000	11074.5	225000	Unaccompanied	Working	Secondary / secondary speci	Married
28	100031	1	Cash loans	F	N	Y	0	112500	979992	27076.5	702000	Unaccompanied	Working	Secondary / secondary speci	Widow
29	100032	0	Cash loans	M	N	Y	1	112500	327024	23827.5	270000	Family	Working	Secondary / secondary speci	Married
30	100033	0	Cash loans	M	Y	Y	0	270000	790830	57676.5	675000	Unaccompanied	State servant	Higher education	Single / not marr
31	100034	0	Revolving loans	M	N	Y	0	90000	180000	9000	180000	Unaccompanied	Working	Higher education	Single / not marr
32	100035	0	Cash loans	F	N	Y	0	292500	665892	24592.5	477000	Unaccompanied	Commercial associate	Secondary / secondary speci	Civil marriage
33	100036	0	Cash loans	F	N	Y	0	112500	512064	25033.5	360000	Family	Working	Secondary / secondary speci	Civil marriage
34	100037	0	Cash loans	F	N	N	0	90000	199008	20893.5	180000	Unaccompanied	Working	Secondary / secondary speci	Civil marriage
35	100039	0	Cash loans	M	Y	N	1	360000	733315.5	39069	679500	Unaccompanied	Commercial associate	Secondary / secondary speci	Married
36	100040	0	Cash loans	F	N	Y	0	135000	1125000	32895	1125000	Unaccompanied	State servant	Higher education	Married
37	100041	0	Cash loans	F	N	N	0	112500	450000	44509.5	450000	Unaccompanied	Working	Higher education	Married
38	100043	0	Cash loans	F	N	Y	2	198000	641173.5	23157	553500	Unaccompanied	Commercial associate	Secondary / secondary speci	Married
39	100044	0	Cash loans	M	N	Y	0	121500	454500	15151.5	454500	Unaccompanied	Working	Secondary / secondary speci	Married

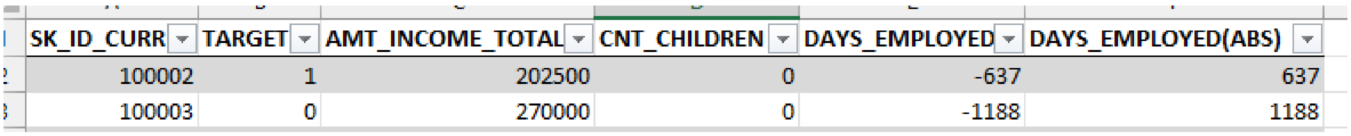
Task 2:

Identify Outliers in the Dataset:

To find Outliers in the Dataset we utilized functions like QUARTILE, IQR, and conditional formatting to identify potential outliers.

We first copied the columns of interest into a new worksheet for finding Outliers.

Columns Copied are:



SK_ID_CURR	TARGET	AMT_INCOME_TOTAL	CNT_CHILDREN	DAYS_EMPLOYED	DAYS_EMPLOYED(ABS)
100002	1	202500	0	-637	637
100003	0	270000	0	-1188	1188

We used the QUARTILE function to find quartile 1 and quartile 3, along with the IQR and upper limit and lower limit ranges.

Formulae:

=QUARTILE.EXC(Table5[AMT_INCOME_TOTAL],1)

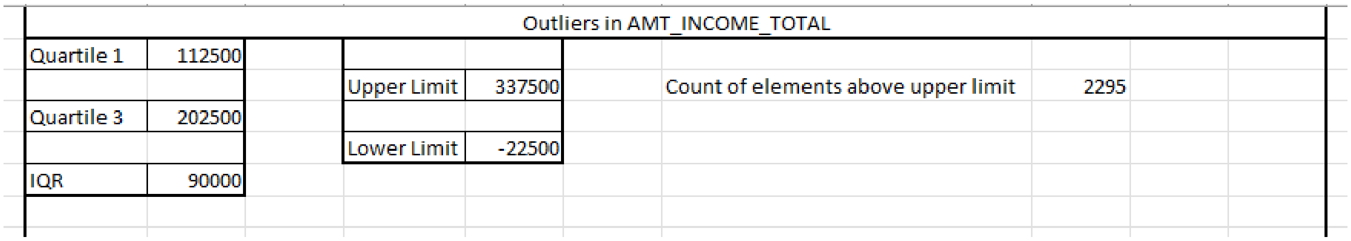
=QUARTILE.EXC(Table5[[#All],[AMT_INCOME_TOTAL]],3)

=I4-I2 (IQR)

=I4+1.5*I6 (Upper limit)

=I2-1.5*I6 (Lower limit)

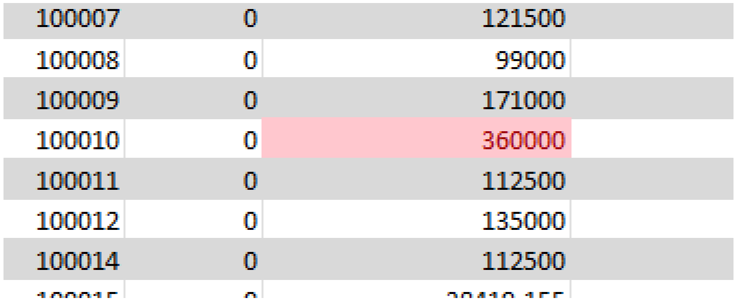
=COUNTIF(C2:C50000,">337500") (Count of elements outside limits)



Outliers in AMT_INCOME_TOTAL

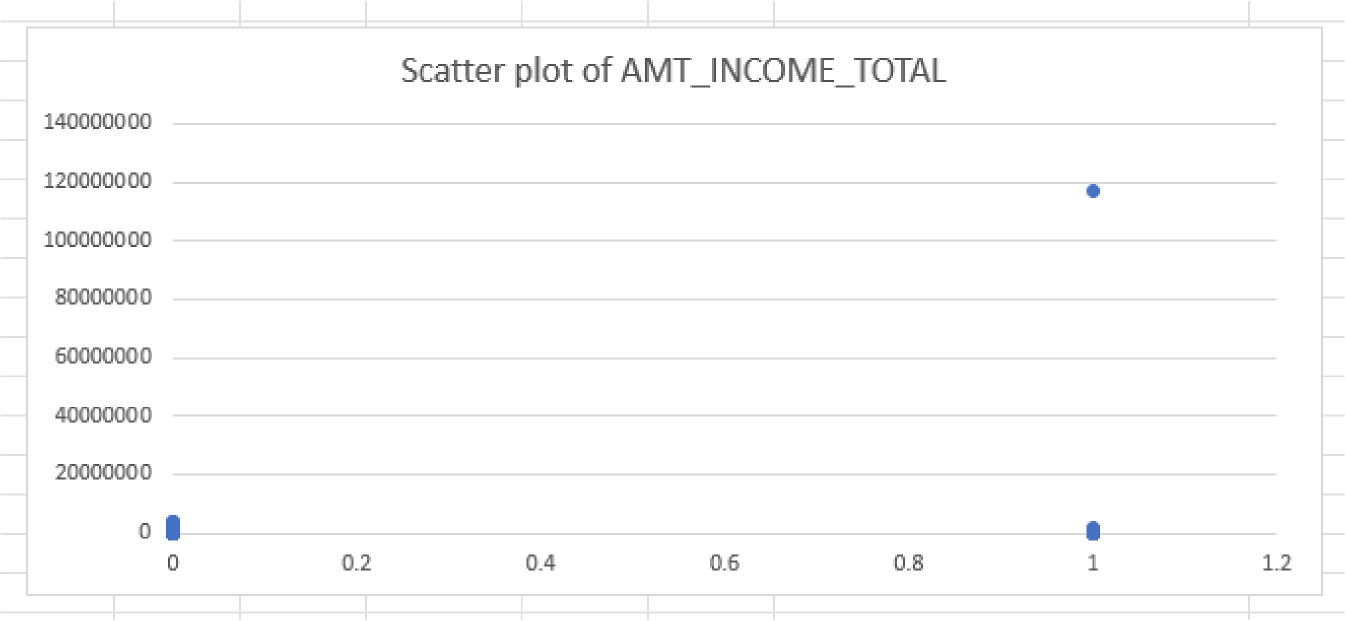
Quartile 1	112500				
		Upper Limit	337500	Count of elements above upper limit	2295
Quartile 3	202500				
		Lower Limit	-22500		
IQR	90000				

We used Conditional Formatting to highlight the cells which contain values outside the limits.



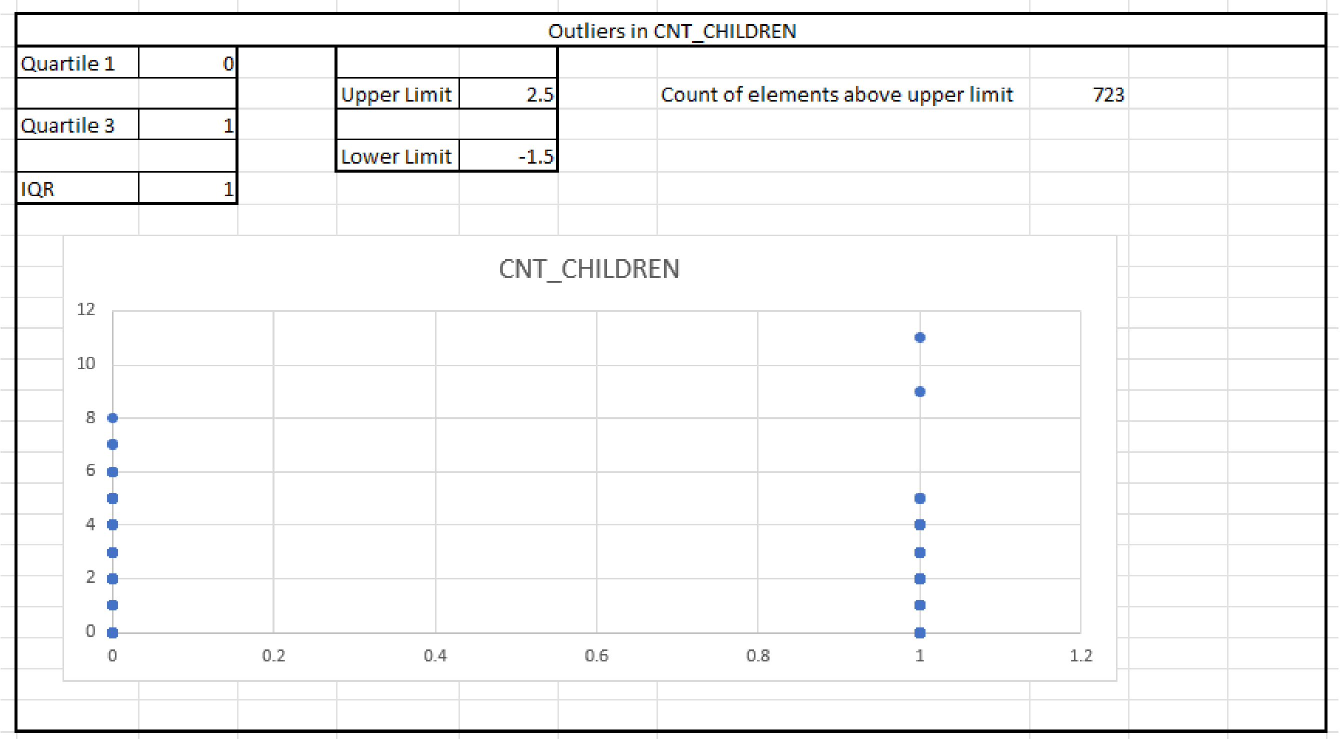
100007	0	121500
100008	0	99000
100009	0	171000
100010	0	360000
100011	0	112500
100012	0	135000
100014	0	112500

We also plotted A scatter plot to visualize the outliers



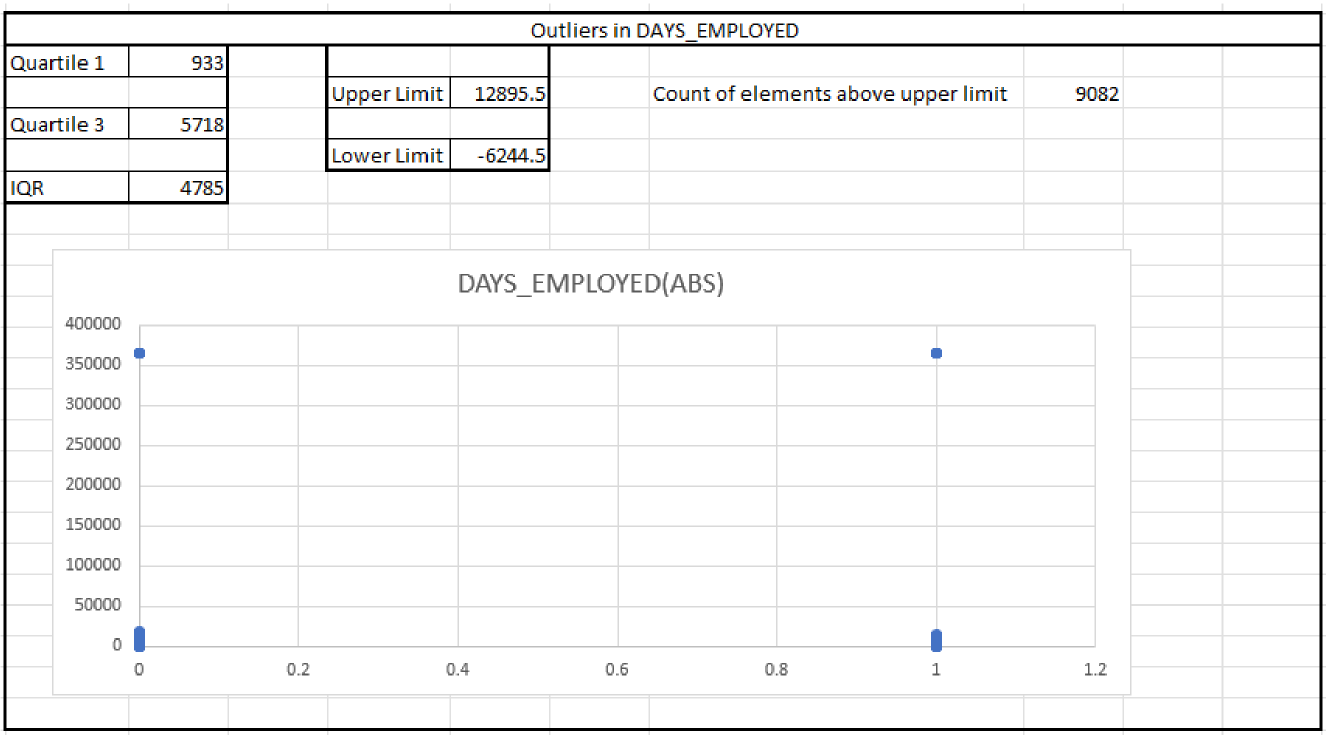
In the above plot the point which lies outside the general trend, and is very much out of the scope can be called an outlier.

Similar Steps were done for other columns and following results were obtained.



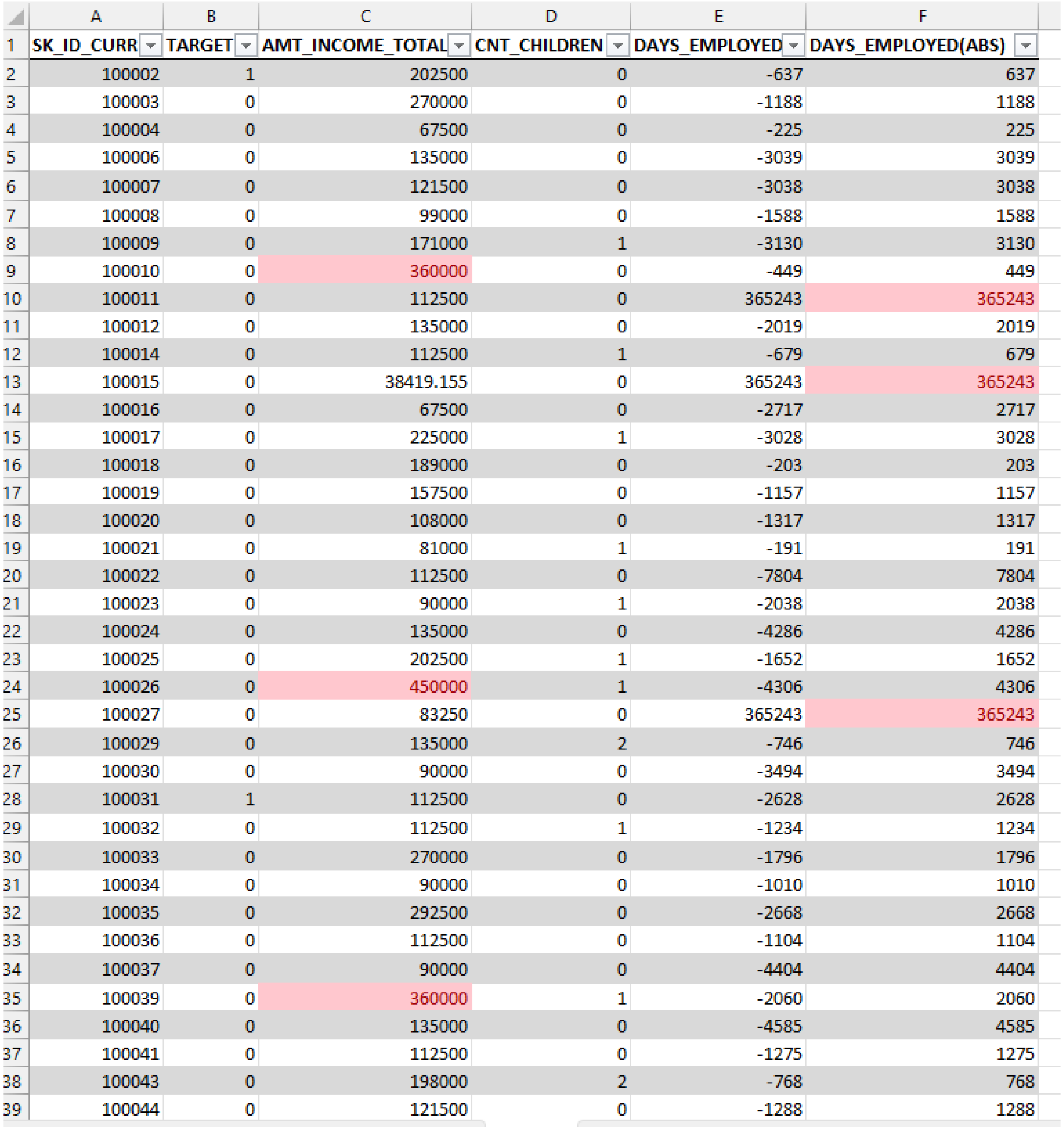
Outliers in CNT_CHILDREN

Quartile 1	0			
		Upper Limit	2.5	Count of elements above upper limit	723
Quartile 3	1			
		Lower Limit	-1.5	
IQR	1			



Outliers in DAYS_EMPLOYED

Quartile 1	933				
		Upper Limit	12895.5	Count of elements above upper limit	9082
Quartile 3	5718				
		Lower Limit	-6244.5		
IQR	4785				



	A	B	C	D	E	F
1	SK_ID_CURR	TARGET	AMT_INCOME_TOTAL	CNT_CHILDREN	DAYS_EMPLOYED	DAYS_EMPLOYED(ABS)
2	100002	1	202500	0	-637	637
3	100003	0	270000	0	-1188	1188
4	100004	0	67500	0	-225	225
5	100006	0	135000	0	-3039	3039
6	100007	0	121500	0	-3038	3038
7	100008	0	99000	0	-1588	1588
8	100009	0	171000	1	-3130	3130
9	100010	0	360000	0	-449	449
10	100011	0	112500	0	365243	365243
11	100012	0	135000	0	-2019	2019
12	100014	0	112500	1	-679	679
13	100015	0	38419.155	0	365243	365243
14	100016	0	67500	0	-2717	2717
15	100017	0	225000	1	-3028	3028
16	100018	0	189000	0	-203	203
17	100019	0	157500	0	-1157	1157
18	100020	0	108000	0	-1317	1317
19	100021	0	81000	1	-191	191
20	100022	0	112500	0	-7804	7804
21	100023	0	90000	1	-2038	2038
22	100024	0	135000	0	-4286	4286
23	100025	0	202500	1	-1652	1652
24	100026	0	450000	1	-4306	4306
25	100027	0	83250	0	365243	365243
26	100029	0	135000	2	-746	746
27	100030	0	90000	0	-3494	3494
28	100031	1	112500	0	-2628	2628
29	100032	0	112500	1	-1234	1234
30	100033	0	270000	0	-1796	1796
31	100034	0	90000	0	-1010	1010
32	100035	0	292500	0	-2668	2668
33	100036	0	112500	0	-1104	1104
34	100037	0	90000	0	-4404	4404
35	100039	0	360000	1	-2060	2060
36	100040	0	135000	0	-4585	4585
37	100041	0	112500	0	-1275	1275
38	100043	0	198000	2	-768	768
39	100044	0	121500	0	-1288	1288

In the above image all highlighted cells are outliers.

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.

To find Data Imbalance we find the number of each element in the TARGET column. For doing this we use the COUNTIF formula.

Formulae:

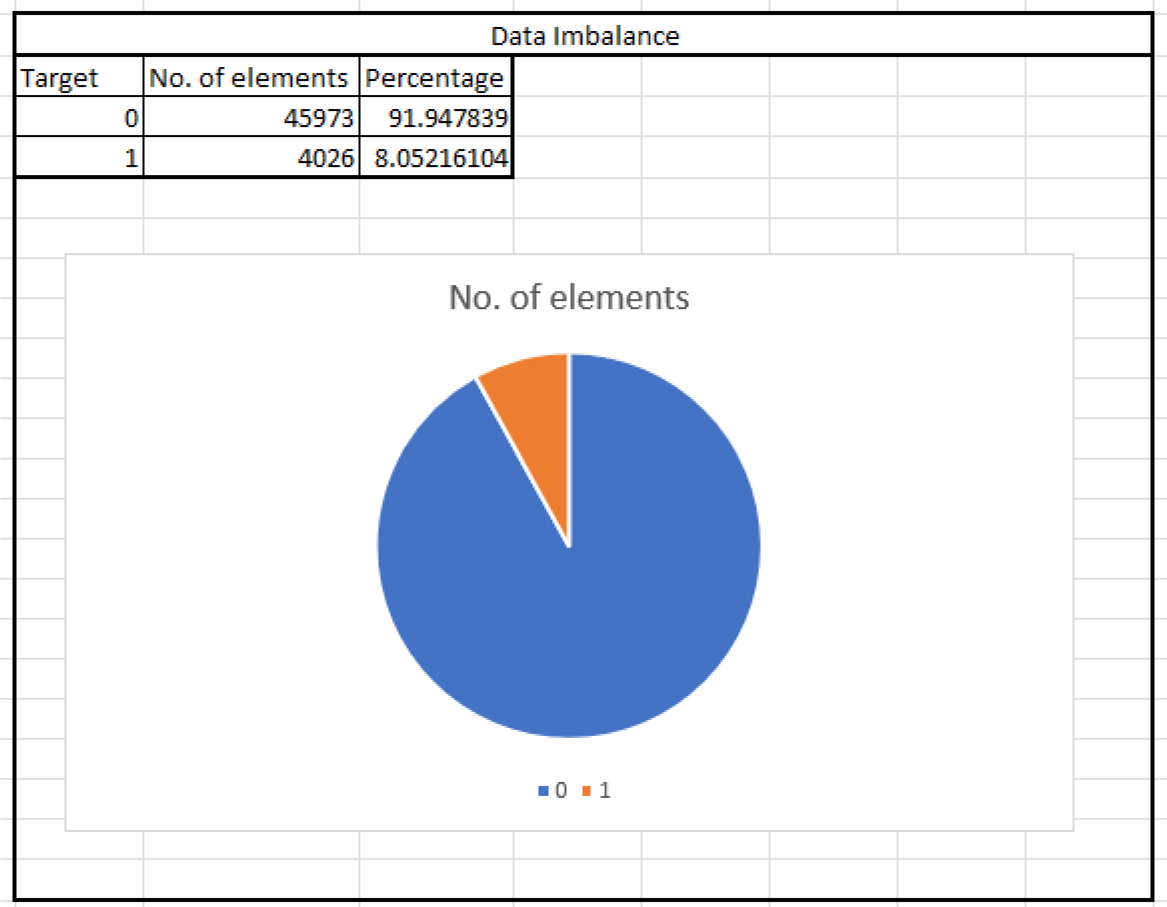
=COUNTIF(B:B,0)

=COUNTIF(B:B,1)

=F4/49999*100 (Percentage)

=F5/49999*100 (Percentage)

We also plot this Data in Pie Chart to visualize the Data Imbalance.



Data Imbalance

Target	No. of elements	Percentage
0	45973	91.947839
1	4026	8.05216104

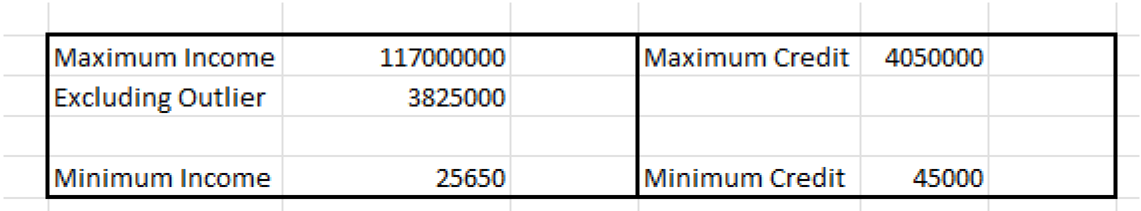
As we can see the no. of 0 in TARGET is very large compared to no. of 1. This will result in a very large data imbalance. Which might skew the results and give less accurate results.

Task 4:

Perform Univariate, Segmented Univariate, and Bivariate Analysis:

To perform Univariate/ Segmented Univariate analysis, we have to utilize functions such as COUNT, AVERAGE, or MEDIAN to find out the total number of applicants over a particular range or how much credit one shall receive according to their income, and other such relations.

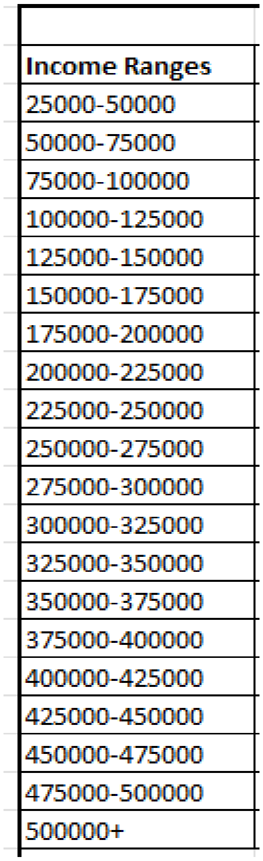
We start by selecting two columns, Credit and Income, we have selected this columns as they have higher correlation. We find maximum and minimum values of these columns excluding outliers.



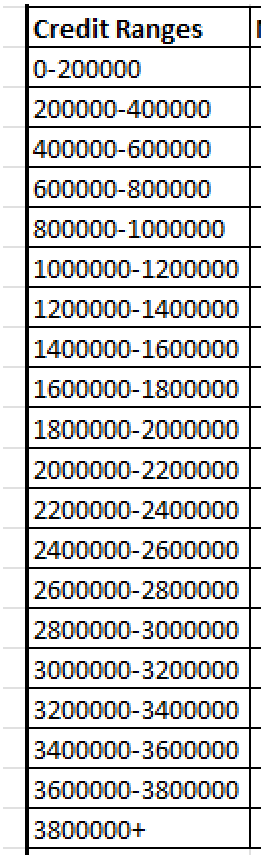
Maximum Income	117000000		Maximum Credit	4050000	
Excluding Outlier	3825000				
Minimum Income	25650		Minimum Credit	45000	

This Data helps us to define ranges to find how many applicants fall in each range.

We Define Ranges on particular Intervals



Income Ranges
25000-50000
50000-75000
75000-100000
100000-125000
125000-150000
150000-175000
175000-200000
200000-225000
225000-250000
250000-275000
275000-300000
300000-325000
325000-350000
350000-375000
375000-400000
400000-425000
425000-450000
450000-475000
475000-500000
500000+



Credit Ranges
0-200000
200000-400000
400000-600000
600000-800000
800000-1000000
1000000-1200000
1200000-1400000
1400000-1600000
1600000-1800000
1800000-2000000
2000000-2200000
2200000-2400000
2400000-2600000
2600000-2800000
2800000-3000000
3000000-3200000
3200000-3400000
3400000-3600000
3600000-3800000
3800000+

We find number of applicants over these ranges by utilizing functions such as:

=FREQUENCY(C:C,X5:X23)

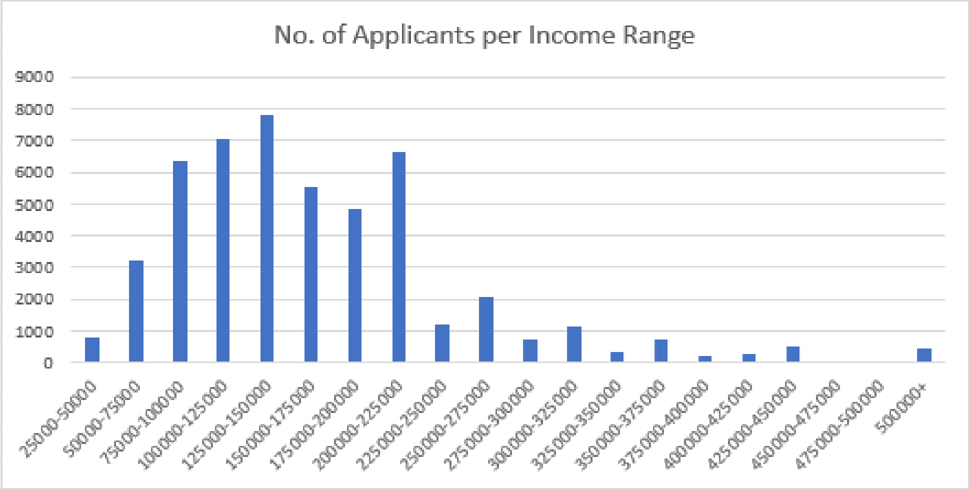
=FREQUENCY(D:D,Y5:Y23)

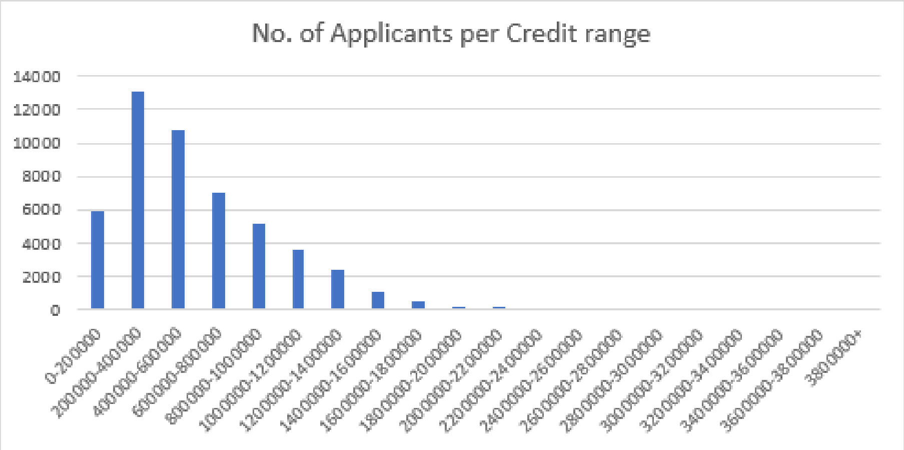
We also plot bar charts to visualize the frequency of applicants in each range.

Univariate Analysis

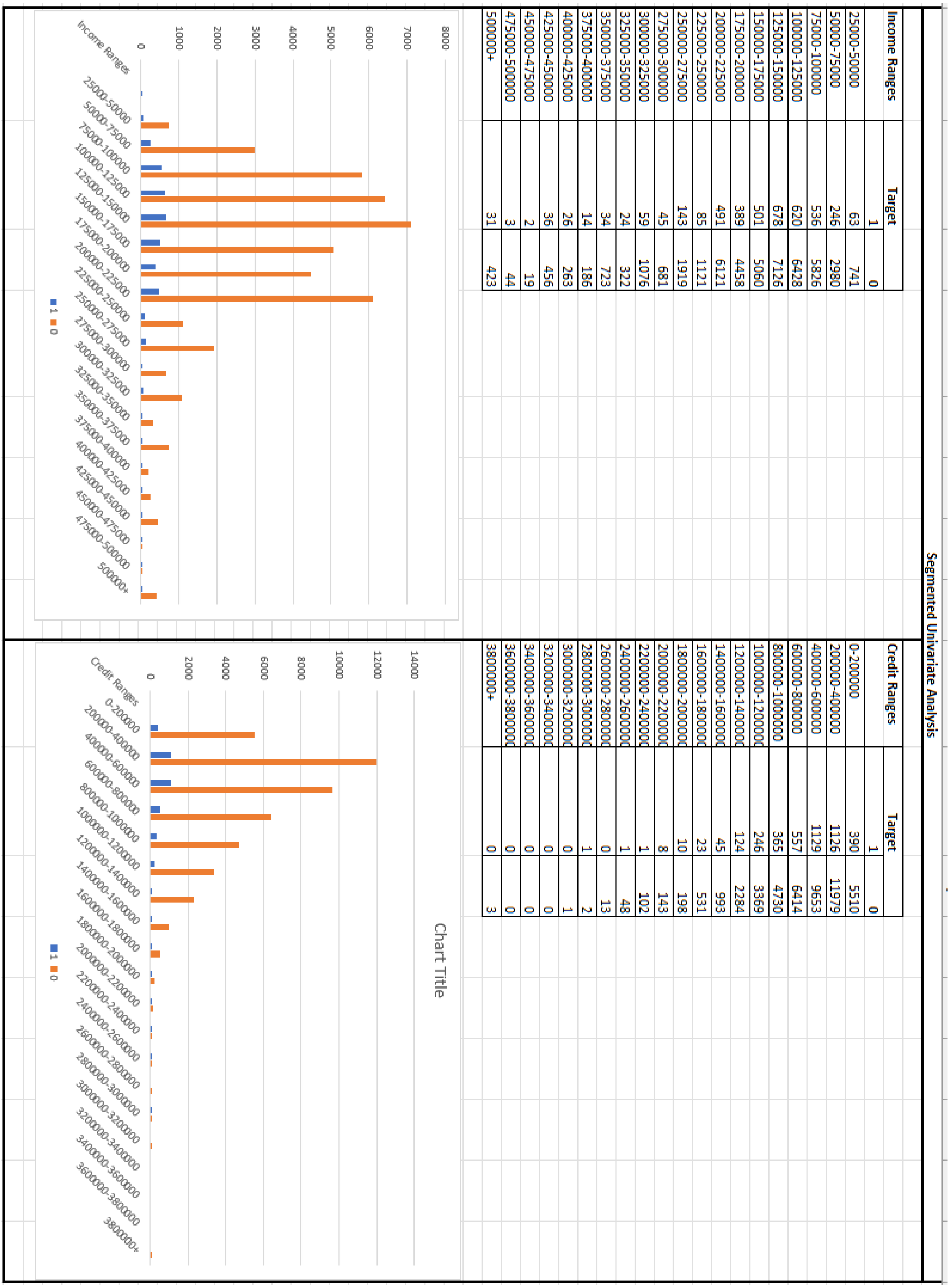
Income Ranges	No. of Applicants
25000-50000	804
50000-75000	3226
75000-100000	6362
100000-125000	7048
125000-150000	7804
150000-175000	5561
175000-200000	4847
200000-225000	6612
225000-250000	1206
250000-275000	2062
275000-300000	726
300000-325000	1135
325000-350000	346
350000-375000	757
375000-400000	200
400000-425000	289
425000-450000	492
450000-475000	21
475000-500000	47
500000+	454

Credit Ranges	No. of Applicants
0-200000	5900
200000-400000	13105
400000-600000	10782
600000-800000	6971
800000-1000000	5095
1000000-1200000	3615
1200000-1400000	2408
1400000-1600000	1038
1600000-1800000	554
1800000-2000000	208
2000000-2200000	151
2200000-2400000	103
2400000-2600000	49
2600000-2800000	13
2800000-3000000	3
3000000-3200000	1
3200000-3400000	0
3400000-3600000	0
3600000-3800000	0
3800000+	3


No. of Applicants per Income Range


No. of Applicants per Credit range

Similarly for segmented univariate analysis we split the Data into two classes according to TARGET.



Segmented Univariate Analysis

Income Ranges	Target	
	1	0
25000-50000	63	741
50000-75000	246	2980
75000-100000	536	5826
100000-125000	620	6428
125000-150000	678	7126
150000-175000	501	5060
175000-200000	389	4458
200000-225000	491	6121
225000-250000	85	1121
250000-275000	143	1919
275000-300000	45	681
300000-325000	59	1076
325000-350000	24	322
350000-375000	34	723
375000-400000	14	186
400000-425000	26	263
425000-450000	36	456
450000-475000	2	19
475000-500000	3	44
500000+	31	423

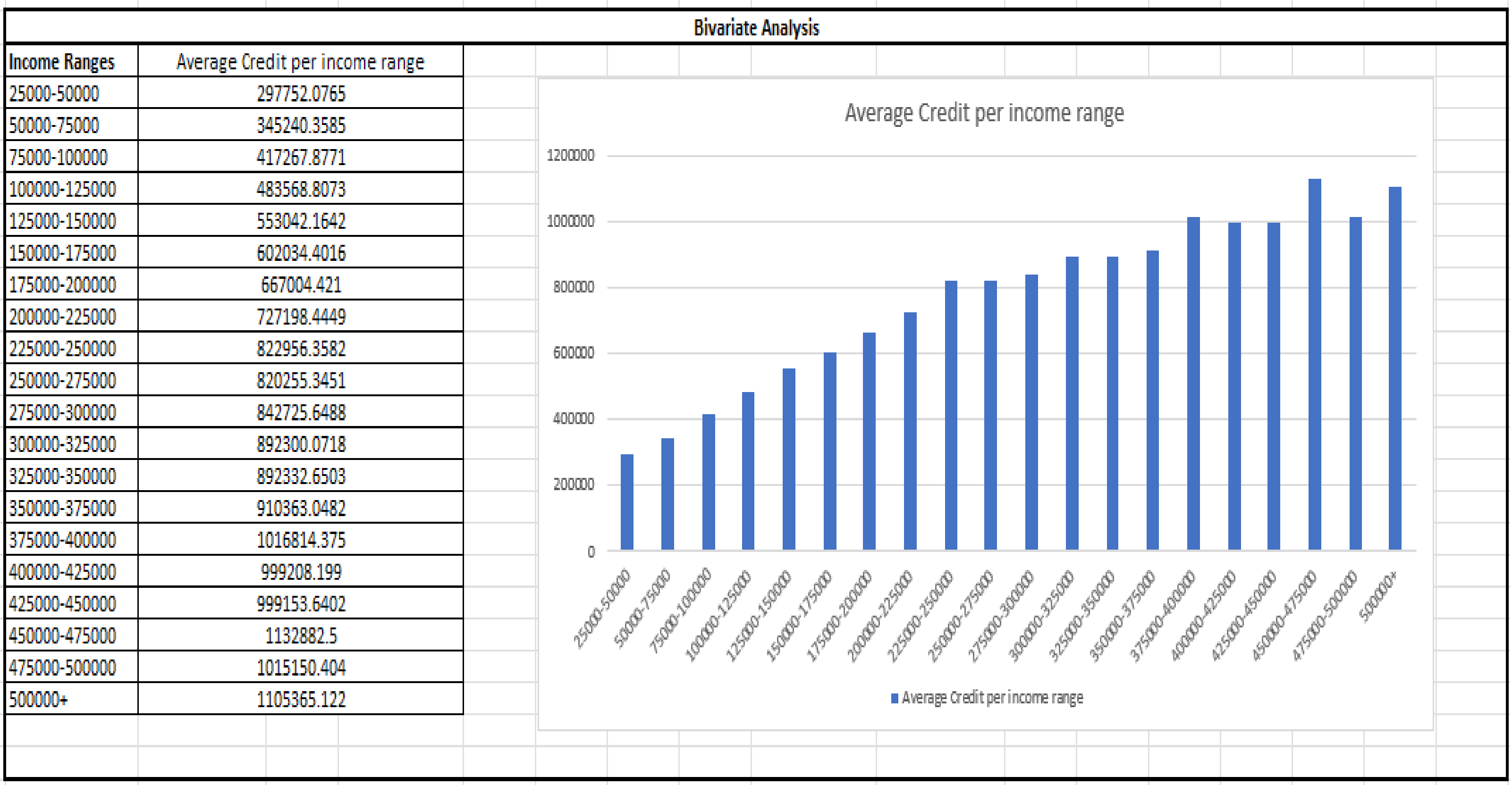
Credit Ranges	Target	
	1	0
0-200000	390	5510
200000-400000	1126	11979
400000-600000	1129	9653
600000-800000	557	6414
800000-1000000	365	4730
1000000-120000	246	3369
1200000-140000	124	2284
1400000-160000	45	993
1600000-180000	23	531
1800000-200000	10	198
2000000-220000	8	143
2200000-240000	1	102
2400000-260000	1	48
2600000-280000	0	13
2800000-300000	1	2
3000000-320000	0	1
3200000-340000	0	0
3400000-360000	0	0
3600000-380000	0	0
3800000+	0	3

To perform Bivariate Analysis, we need to find the average of credit per income range, for that we use the AVERAGEIF function.

=AVERAGEIFS($D$2:$D$50000,$C$2:$C$50000,">"&X4,$C$2:$C$50000,"<="&X5)

Above formula checks two conditions, if element is greater than lower limit and smaller than upper limit, and only then is considered for average.

We plotted a Bar Graph Similar to above analysis



Bivariate Analysis

Income Ranges	Average Credit per income range
25000-50000	297752.0765
50000-75000	345240.3585
75000-100000	417267.8771
100000-125000	483568.8073
125000-150000	553042.1642
150000-175000	602034.4016
175000-200000	667004.421
200000-225000	727198.4449
225000-250000	822956.3582
250000-275000	820255.3451
275000-300000	842725.6488
300000-325000	892300.0718
325000-350000	892332.6503
350000-375000	910363.0482
375000-400000	1016814.375
400000-425000	999208.199
425000-450000	999153.6402
450000-475000	1132882.5
475000-500000	1015150.404
500000+	1105365.122

Task 5:

Identify Top Correlations for Different Scenarios:

To find Correlation of different columns we utilized the CORREL function of Excel. We first separated the data into three tables, one having only 0 Target, one having 1 Target and both combined Target. We found correlation tables for all these by using the CORREL function, and made it better for visualization using conditional formatting.

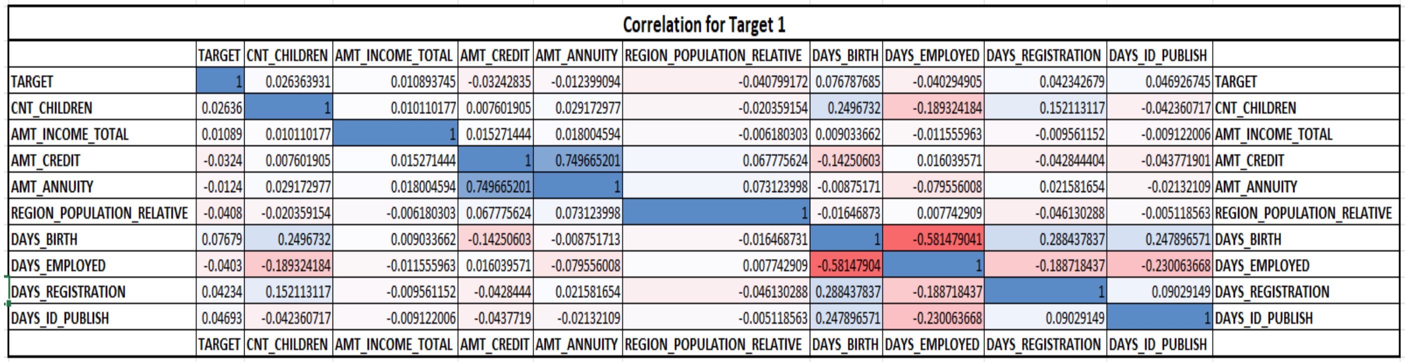
Formulae:

=CORREL($C:$C,B:B)

=CORREL('Target 0 Data'!C:C,'Target 0 Data'!$F:$F)

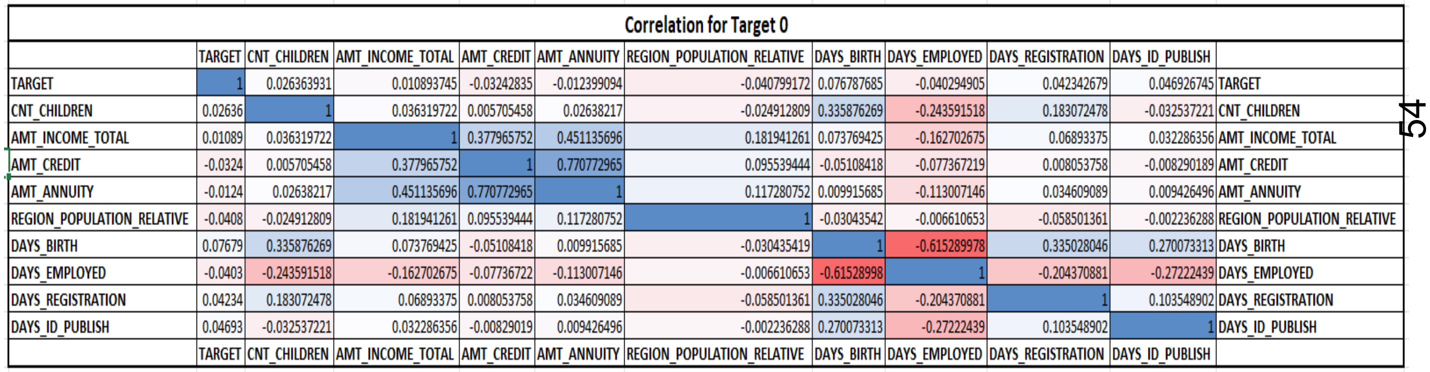
=CORREL('Target 1 Data'!B:B,'Target 1 Data'!$F:$F)

For better visualization heatmaps of correlation matrix were created.



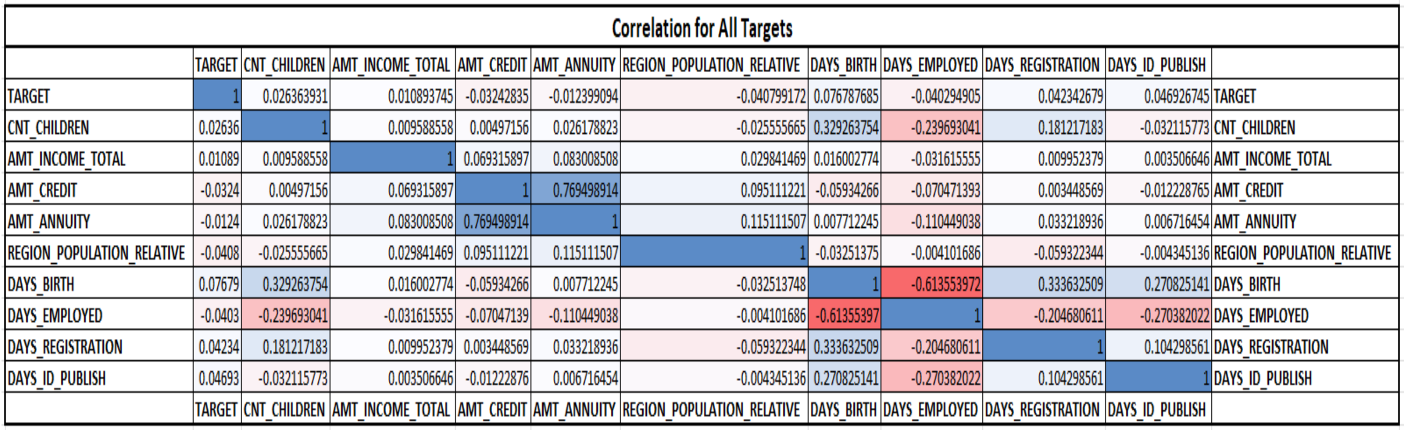
Correlation for Target 1

	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	
TARGET	1	0.026363931	0.010893745	-0.03242835	-0.012399094	-0.040799172	0.076787685	-0.040294905	0.042342679	0.046926745	TARGET
CNT_CHILDREN	0.02636	1	0.010110177	0.007601905	0.029172977	-0.020359154	0.2496732	-0.189324184	0.152113117	-0.042360717	CNT_CHILDREN
AMT_INCOME_TOTAL	0.01089	0.010110177	1	0.015271444	0.018004594	-0.006180303	0.009033662	-0.011555963	-0.009561152	-0.009122006	AMT_INCOME_TOTAL
AMT_CREDIT	-0.0324	0.007601905	0.015271444	1	0.749665201	0.067775624	-0.14250603	0.016039571	-0.042844404	-0.043771901	AMT_CREDIT
AMT_ANNUITY	-0.0124	0.029172977	0.018004594	0.749665201	1	0.073123998	-0.00875171	-0.079556008	0.021581654	-0.02132109	AMT_ANNUITY
REGION_POPULATION_RELATIVE	-0.0408	-0.020359154	-0.006180303	0.067775624	0.073123998	1	-0.01646873	0.007742909	-0.046130288	-0.005118563	REGION_POPULATION_RELATIVE
DAYS_BIRTH	0.07679	0.2496732	0.009033662	-0.14250603	-0.008751713	-0.016468731	1	-0.581479041	0.288437837	0.247896571	DAYS_BIRTH
DAYS_EMPLOYED	-0.0403	-0.189324184	-0.011555963	0.016039571	-0.079556008	0.007742909	-0.58147904	1	-0.188718437	-0.230063668	DAYS_EMPLOYED
DAYS_REGISTRATION	0.04234	0.152113117	-0.009561152	-0.0428444	0.021581654	-0.046130288	0.288437837	-0.188718437	1	0.09029149	DAYS_REGISTRATION
DAYS_ID_PUBLISH	0.04693	-0.042360717	-0.009122006	-0.0437719	-0.02132109	-0.005118563	0.247896571	-0.230063668	0.09029149	1	DAYS_ID_PUBLISH
	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	



Correlation for Target 0

	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	
TARGET	1	0.026363931	0.010893745	-0.03242835	-0.012399094	-0.040799172	0.076787685	-0.040294905	0.042342679	0.046926745	TARGET
CNT_CHILDREN	0.02636	1	0.036319722	0.005705458	0.02638217	-0.024912809	0.335876269	-0.243591518	0.183072478	-0.032537221	CNT_CHILDREN
AMT_INCOME_TOTAL	0.01089	0.036319722	1	0.377965752	0.451135696	0.181941261	0.073769425	-0.162702675	0.06893375	0.032286356	AMT_INCOME_TOTAL
AMT_CREDIT	-0.0324	0.005705458	0.377965752	1	0.770772965	0.095539444	-0.05108418	-0.077367219	0.008053758	-0.008290189	AMT_CREDIT
AMT_ANNUITY	-0.0124	0.02638217	0.451135696	0.770772965	1	0.117280752	0.009915685	-0.113007146	0.034609089	0.009426496	AMT_ANNUITY
REGION_POPULATION_RELATIVE	-0.0408	-0.024912809	0.181941261	0.095539444	0.117280752	1	-0.03043542	-0.006610653	-0.058501361	-0.002236288	REGION_POPULATION_RELATIVE
DAYS_BIRTH	0.07679	0.335876269	0.073769425	-0.05108418	0.009915685	-0.030435419	1	-0.615289978	0.335028046	0.270073313	DAYS_BIRTH
DAYS_EMPLOYED	-0.0403	-0.243591518	-0.162702675	-0.07736722	-0.113007146	-0.006610653	-0.61528998	1	-0.204370881	-0.27222439	DAYS_EMPLOYED
DAYS_REGISTRATION	0.04234	0.183072478	0.06893375	0.008053758	0.034609089	-0.058501361	0.335028046	-0.204370881	1	0.103548902	DAYS_REGISTRATION
DAYS_ID_PUBLISH	0.04693	-0.032537221	0.032286356	-0.00829019	0.009426496	-0.002236288	0.270073313	-0.27222439	0.103548902	1	DAYS_ID_PUBLISH
	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	



Correlation for All Targets

	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	
TARGET	1	0.026363931	0.010893745	-0.03242835	-0.012399094	-0.040799172	0.076787685	-0.040294905	0.042342679	0.046926745	TARGET
CNT_CHILDREN	0.02636	1	0.009588558	0.00497156	0.026178823	-0.025555665	0.329263754	-0.239693041	0.181217183	-0.032115773	CNT_CHILDREN
AMT_INCOME_TOTAL	0.01089	0.009588558	1	0.069315897	0.083008508	0.029841469	0.016002774	-0.031615555	0.009952379	0.003506646	AMT_INCOME_TOTAL
AMT_CREDIT	-0.0324	0.00497156	0.069315897	1	0.769498914	0.095111221	-0.05934266	-0.070471393	0.003448569	-0.012228765	AMT_CREDIT
AMT_ANNUITY	-0.0124	0.026178823	0.083008508	0.769498914	1	0.115111507	0.007712245	-0.110449038	0.033218936	0.006716454	AMT_ANNUITY
REGION_POPULATION_RELATIVE	-0.0408	-0.025555665	0.029841469	0.095111221	0.115111507	1	-0.03251375	-0.004101686	-0.059322344	-0.004345136	REGION_POPULATION_RELATIVE
DAYS_BIRTH	0.07679	0.329263754	0.016002774	-0.05934266	0.007712245	-0.032513748	1	-0.613553972	0.333632509	0.270825141	DAYS_BIRTH
DAYS_EMPLOYED	-0.0403	-0.239693041	-0.031615555	-0.07047139	-0.110449038	-0.004101686	-0.61355397	1	-0.204680611	-0.270382022	DAYS_EMPLOYED
DAYS_REGISTRATION	0.04234	0.181217183	0.009952379	0.003448569	0.033218936	-0.059322344	0.333632509	-0.204680611	1	0.104298561	DAYS_REGISTRATION
DAYS_ID_PUBLISH	0.04693	-0.032115773	0.003506646	-0.01222876	0.006716454	-0.004345136	0.270825141	-0.270382022	0.104298561	1	DAYS_ID_PUBLISH
	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	

Results:

While working on this project, I have gained a better understanding of Bank Loan Application Process and Analytics and Advanced Excel methodologies. By analyzing Application Data, I was able to provide insights on various aspects such as Cleaning the Data, Outliers in the Data, Data Imbalance, Univariate and Bivariate Analysis, and correlation between various parameters in bank loan application.
This project has helped me enhance my Excel skills, particularly in functions and data visualization to derive meaningful insights. It has also improved my ability to interpret data and provide actionable recommendations based on the analysis.

Impact of Car Features on Price and Profitability

Link for Excel sheet:

https://docs.google.com/spreadsheets/d/10QceKTy49wcBmaG8W8PhtQBoad8L1MI_/edit?usp=sharing&ouid=107365393175079460343&rtpof=true&sd=true

Project Description:

This project aims to analyze a dataset containing information about various Car Brands, Car models they make and their respective car features along with their prices. The goal is to gain insights about impact of car features on price and profitability, performing various analysis tasks and also build a dashboard to better visualize the insights. The data provided has various missing or null Data, our task is to handle those missing values appropriately, by either deleting or imputing these data. There are various outliers in data, we have to find these outliers. We utilize various excel features such as pivot tables and charts to better represent data. We find trends in car features and their popularities by implementing various methodologies and data analysis techniques such as regression. Thus, by employing statistics and Excel formulas, we will extract meaningful conclusions to help understand the factors that contribute to popularity and profitability of particular cars.

Approach:

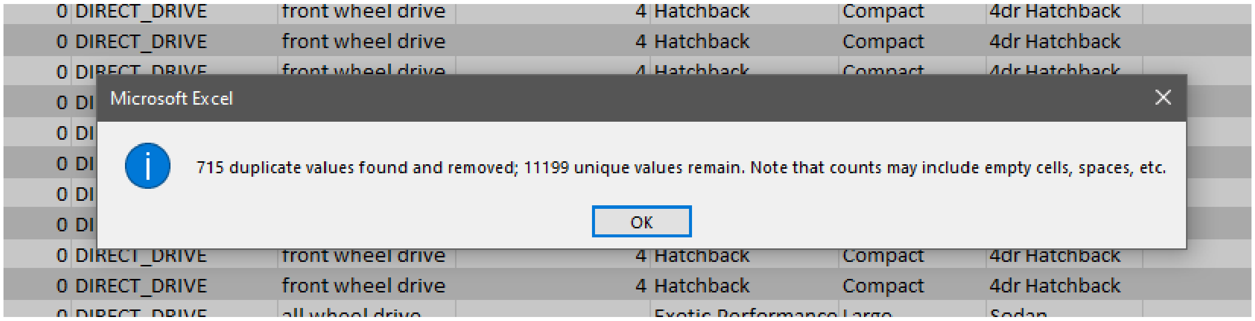
As an individual working on this project, I followed a structured approach to analyze data about Car Brands, models and features. I began by carefully examining the provided database and familiarizing myself with its structure and columns. I tried to find columns which had the most significance in the dataset. I handled missing values by eliminating columns which had most empty cells, and were not significant. And imputed data into cells that were necessary for analysis. Then, I utilized Excel fundamentals to retrieve the necessary information for each task, employing appropriate functions and statistical methods. I focused on data accuracy and quality throughout the project, ensuring reliable results. By leveraging my Excel skills and maintaining a systematic workflow, I successfully executed the project and created a comprehensive report that fulfilled the objectives of providing marketing insights and investor metrics.

Tech-Stack Used:

For this project, I utilized Microsoft Excel as the primary software tool.

Data Cleaning:

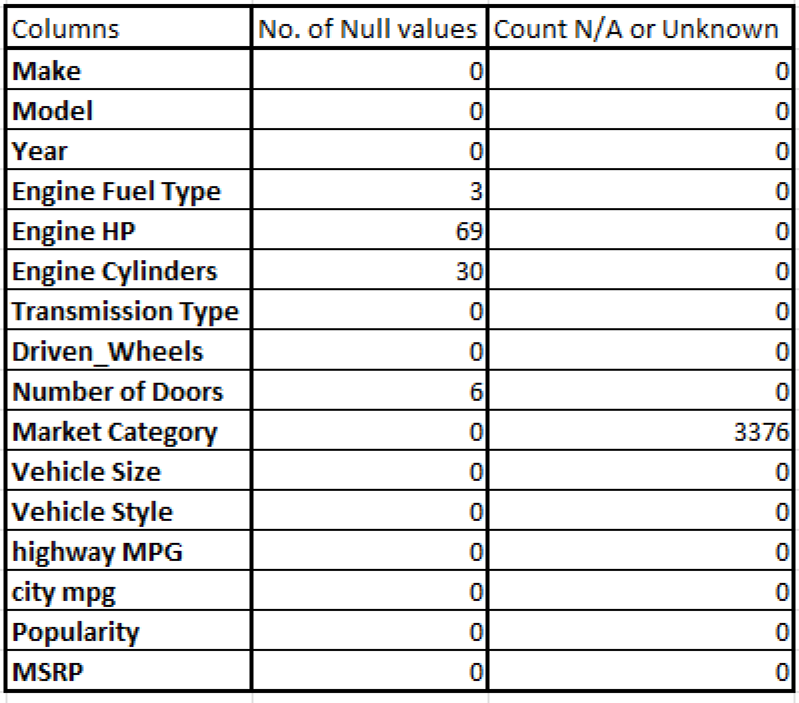
Given Data had various missing and duplicate values. For accurate analysis we need to handle this missing data, and eliminate the duplicate data as it is redundant and might skew the results. For Removing the duplicate data we used excel's Remove Duplicates feature in the Data Tools Tab. We had 715 duplicate rows, which were removed completely.



0	DIRECT_DRIVE	front wheel drive	4	Hatchback	Compact	4dr Hatchback
0	DIRECT_DRIVE	front wheel drive	4	Hatchback	Compact	4dr Hatchback
0	DIRECT_DRIVE	front wheel drive	4	Hatchback	Compact	4dr Hatchback
0	DIRECT_DRIVE	front wheel drive	4	Hatchback	Compact	4dr Hatchback

715 duplicate values found and removed; 11199 unique values remain. Note that counts may include empty cells, spaces, etc.

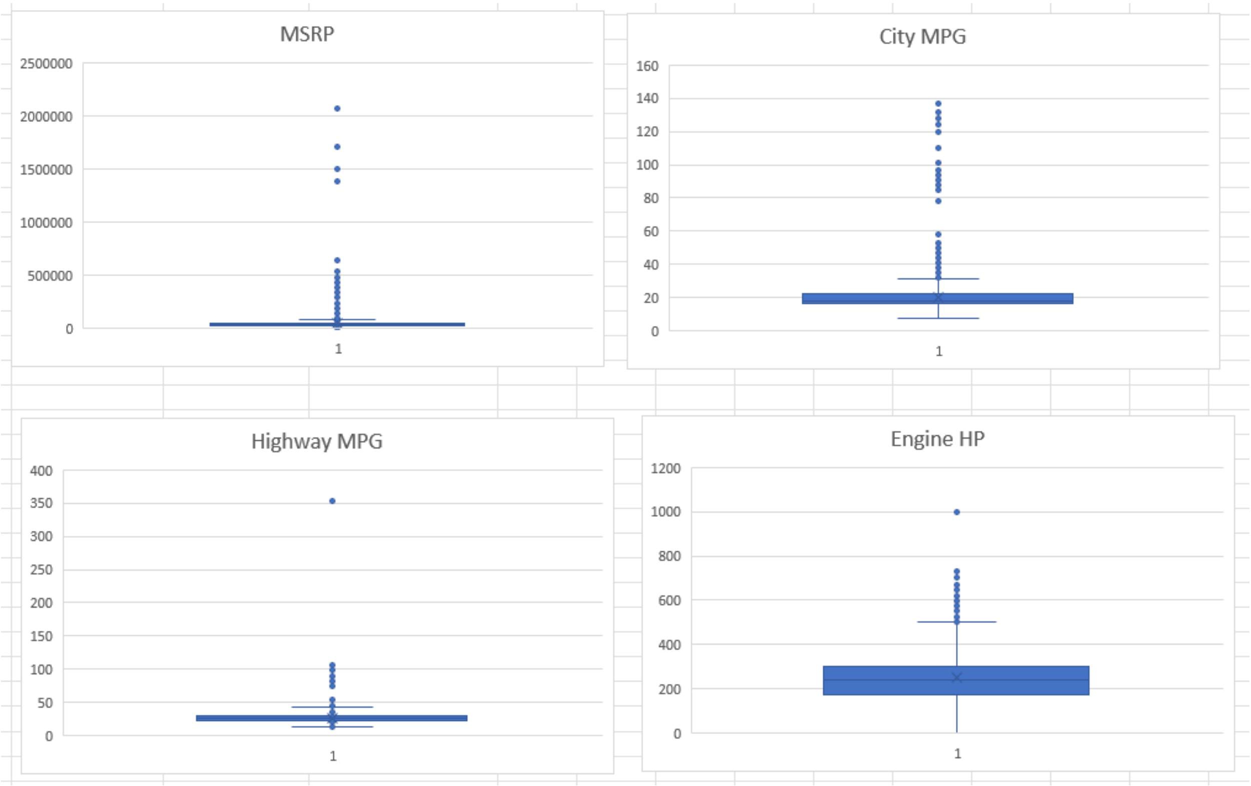
To find missing values we used the COUNTBLANK formula in excel.
=COUNTBLANK(A$2:A$11160)



Columns	No. of Null values	Count N/A or Unknown
Make	0	0
Model	0	0
Year	0	0
Engine Fuel Type	3	0
Engine HP	69	0
Engine Cylinders	30	0
Transmission Type	0	0
Driven_Wheels	0	0
Number of Doors	6	0
Market Category	0	3376
Vehicle Size	0	0
Vehicle Style	0	0
highway MPG	0	0
city mpg	0	0
Popularity	0	0
MSRP	0	0

We removed rows which had less no. of nulls and imputed values in columns such as Engine HP and Engine Cylinders according to the given data.

Data also had some outliers or false values, which needed to be handled. We plotted these outliers using BOX and Whisker chart type.
As seen in the The chart below the features have outliers, some of which are justified but, feature Highway MPG has value which is a bit out of range. So we check with the data of similar Cars and adjust it accordingly.


MSRP
2500000
2000000
1500000
1000000
500000
0
1
City MPG
160
140
120
100
80
60
40
20
0
1
Highway MPG
400
350
300
250
200
150
100
50
0
1
Engine HP
1200
1000
800
600
400
200
0
1

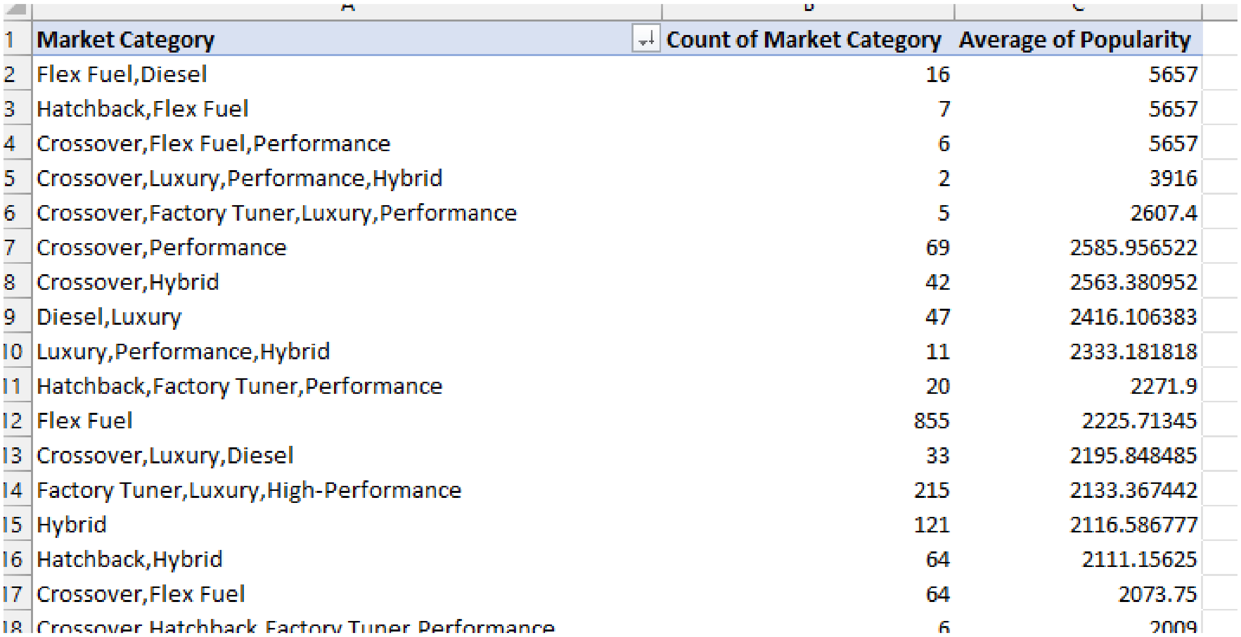
Insights:

Analysis:

Task 1:

Insight Required: How does the popularity of a car model vary across different market categories?

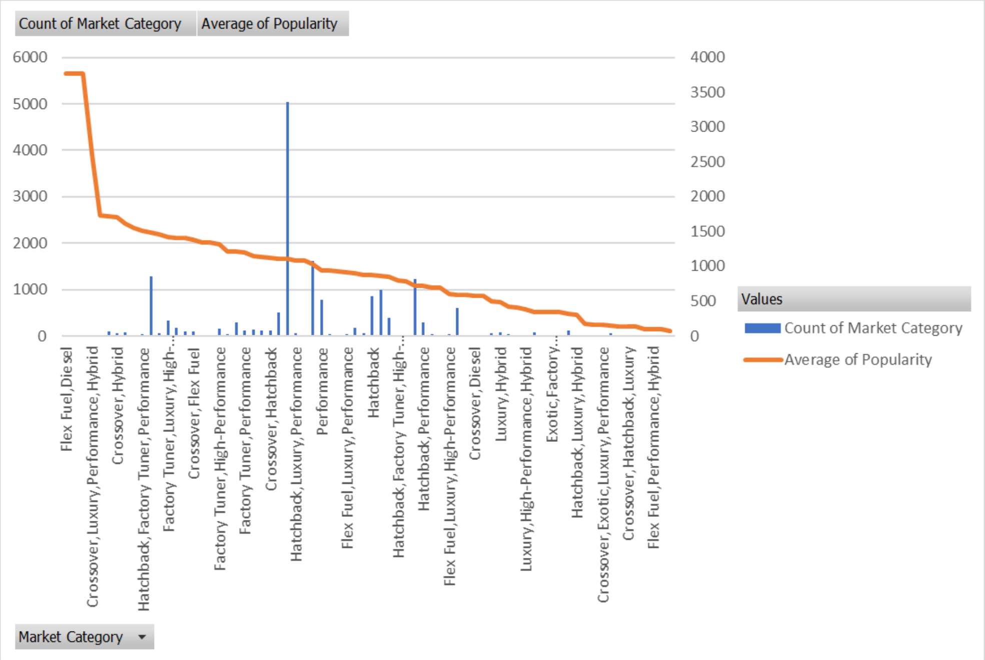
To perform this task we utilized a pivot table in excel that shows the number of car models in each market category and their corresponding popularity scores.



	Market Category	Count of Market Category	Average of Popularity
2	Flex Fuel,Diesel	16	5657
3	Hatchback,Flex Fuel	7	5657
4	Crossover,Flex Fuel,Performance	6	5657
5	Crossover,Luxury,Performance,Hybrid	2	3916
6	Crossover,Factory Tuner,Luxury,Performance	5	2607.4
7	Crossover,Performance	69	2585.956522
8	Crossover,Hybrid	42	2563.380952
9	Diesel,Luxury	47	2416.106383
10	Luxury,Performance,Hybrid	11	2333.181818
11	Hatchback,Factory Tuner,Performance	20	2271.9
12	Flex Fuel	855	2225.71345
13	Crossover,Luxury,Diesel	33	2195.848485
14	Factory Tuner,Luxury,High-Performance	215	2133.367442
15	Hybrid	121	2116.586777
16	Hatchback,Hybrid	64	2111.15625
17	Crossover,Flex Fuel	64	2073.75
18	Crossover,Hatchback,Factory Tuner,Performance	6	2009

This pivot table shows Market Category with its count and average popularity for each.

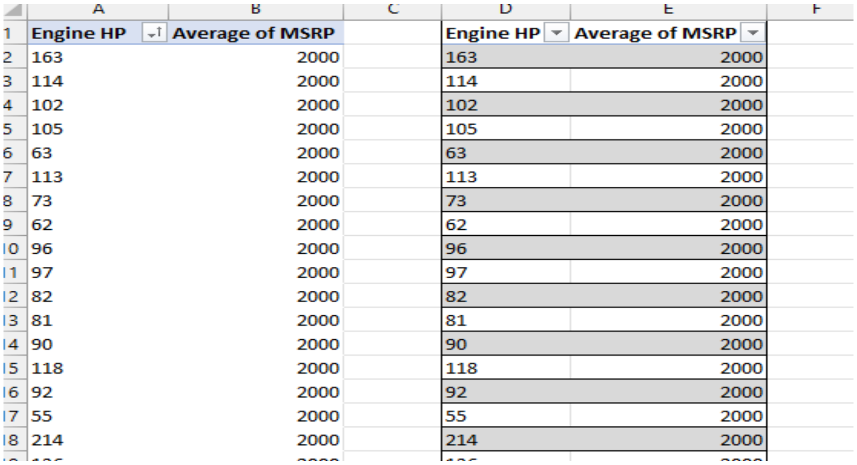
From the above pivot table we plot a combo chart of column-line charts. We select a secondary axis for count to better visualize the chart.



Task 2:

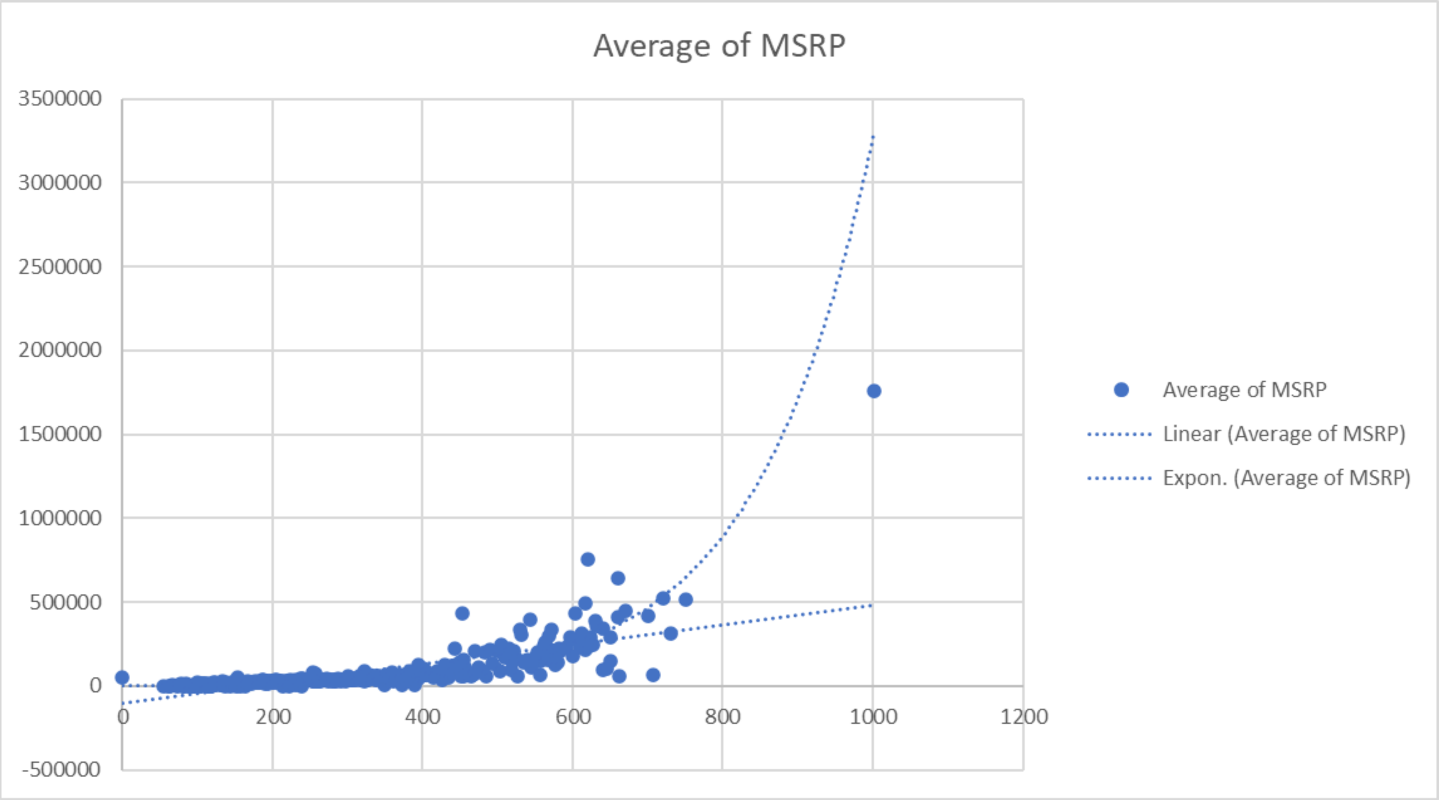
Insight Required: What is the relationship between a car's engine power and its price?

To find the relationship between a car's engine power that is Engine HP and its MSRP, we utilize power pivot to find average MSRP for each Engine HP. We then copy this data into a new table and then create a scatter plot of Engine HP vs average MSRP.



	A	B	C	D	E	F
1	Engine HP	Average of MSRP		Engine HP	Average of MSRP	
2	163	2000		163	2000	
3	114	2000		114	2000	
4	102	2000		102	2000	
5	105	2000		105	2000	
6	63	2000		63	2000	
7	113	2000		113	2000	
8	73	2000		73	2000	
9	62	2000		62	2000	
10	96	2000		96	2000	
11	97	2000		97	2000	
12	82	2000		82	2000	
13	81	2000		81	2000	
14	90	2000		90	2000	
15	118	2000		118	2000	
16	92	2000		92	2000	
17	55	2000		55	2000	
18	214	2000		214	2000	

We now Create a scatter plot for the above table.

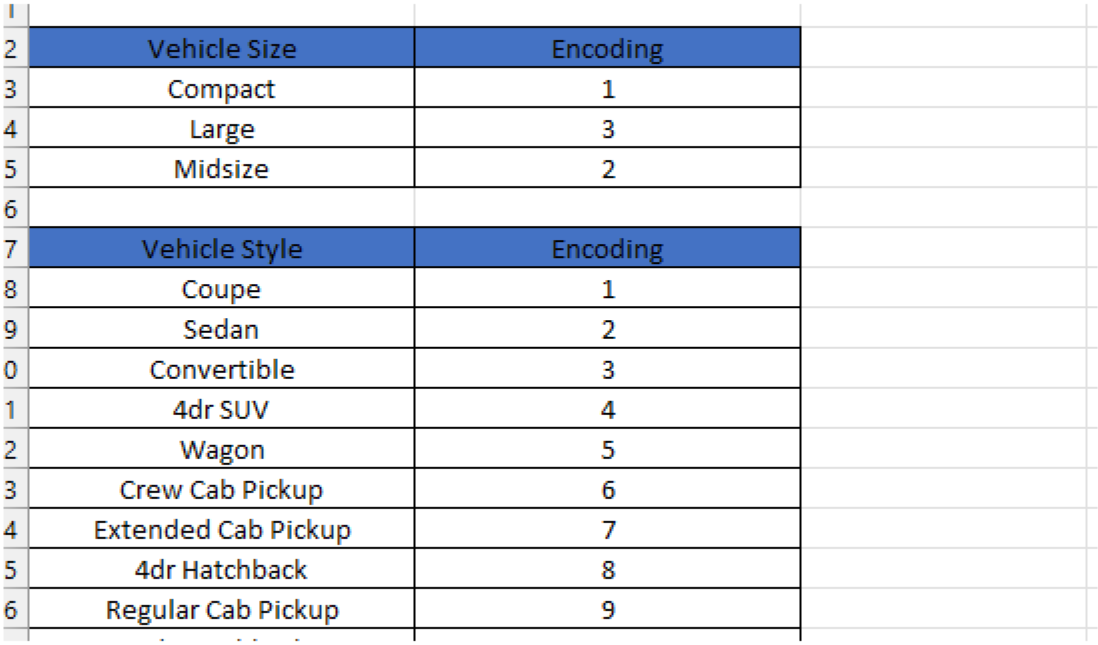


We have also added trendlines to understand how MSRP is changing according to the change inEngine HP. Trend seems to increase exponentially rather than linearly, but to predict more accurately we need to have more data available.

Task 3:

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

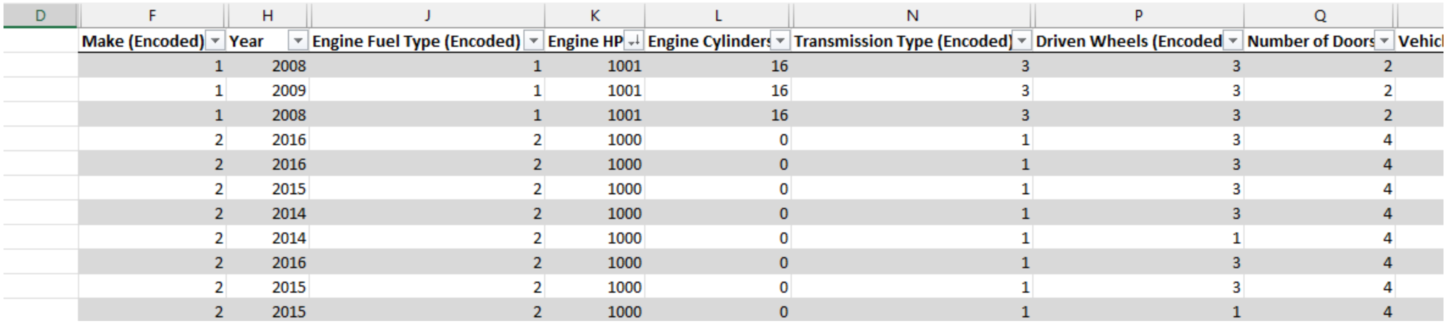
To perform this analysis, we need to consider every feature which is correlated with the price of a car. For this we need to perform regression analysis and then plot coefficients of each feature to check which have most impact on MSRP. But for regression analysis we need to have numerical data, so we first convert the data into numerical data by converting categorical data into encoded data.



Vehicle Size	Encoding
Compact	1
Large	3
Midsize	2

Vehicle Style	Encoding
Coupe	1
Sedan	2
Convertible	3
4dr SUV	4
Wagon	5
Crew Cab Pickup	6
Extended Cab Pickup	7
4dr Hatchback	8
Regular Cab Pickup	9

We use this type of conversion to encode data into numerical values.



Make (Encoded)	Year	Engine Fuel Type (Encoded)	Engine HP	Engine Cylinders	Transmission Type (Encoded)	Driven Wheels (Encoded	Number of Doors	Vehicl
1	2008	1	1001	16	3	3	2	
1	2009	1	1001	16	3	3	2	
1	2008	1	1001	16	3	3	2	
2	2016	2	1000	0	1	3	4	
2	2016	2	1000	0	1	3	4	
2	2015	2	1000	0	1	3	4	
2	2014	2	1000	0	1	3	4	
2	2014	2	1000	0	1	1	4	
2	2016	2	1000	0	1	3	4	
2	2015	2	1000	0	1	3	4	
2	2015	2	1000	0	1	1	4	

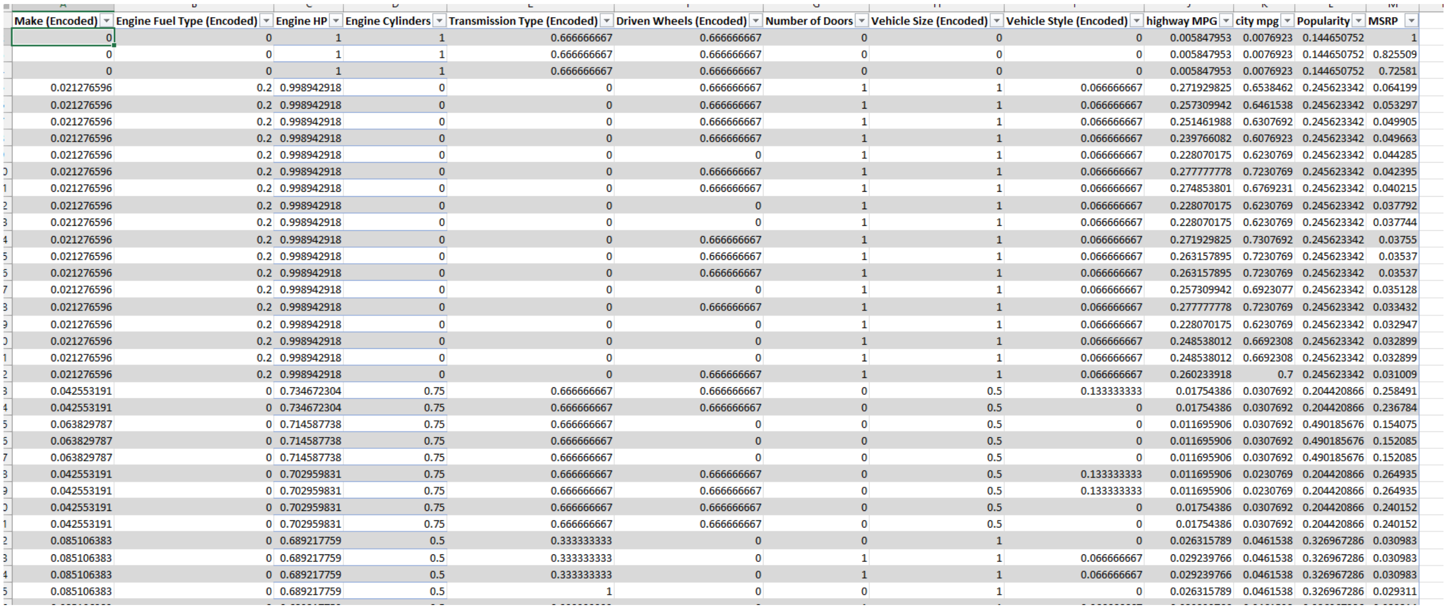
We get this type of data. But we have to normalize it first. As the parameters have very large differences in their ranges.

To normalize we find maximum and minimum values in each column and then, normalize them using the following formula.

=(Analysis_Task3!$F2-Analysis_Task3!F$11199)/Analysis_Task3!F$11201

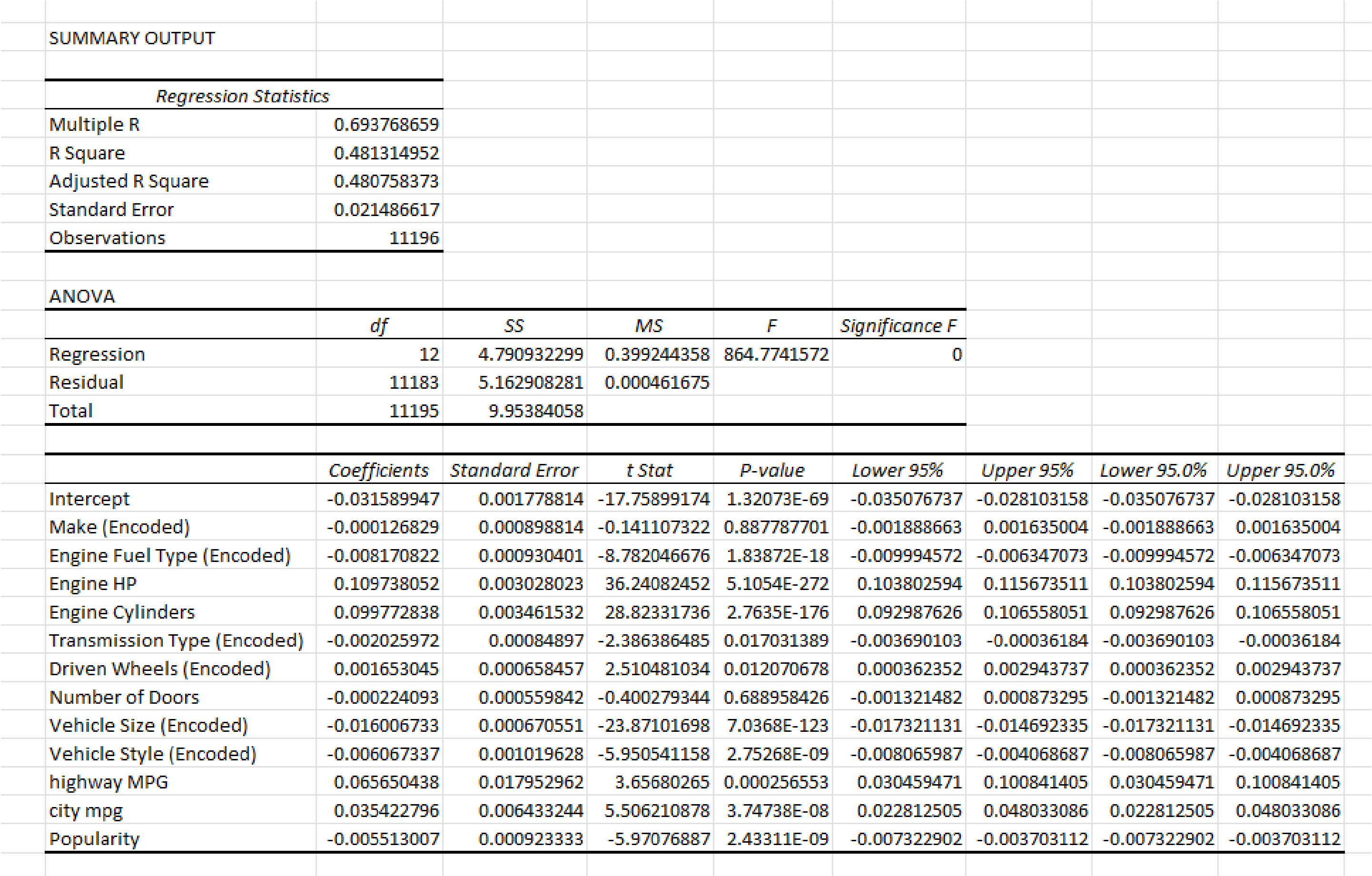
Here we subtract minimum values from each value and then divide with the difference between maximum and minimum values, to get normalized values between 0 and 1.

Thus all of the values will get converted into range from 0 to 1.



Make (Encoded)	Engine Fuel Type (Encoded)	Engine HP	Engine Cylinders	Transmission Type (Encoded)	Driven Wheels (Encoded)	Number of Doors	Vehicle Size (Encoded)	Vehicle Style (Encoded)	highway MPG	city mpg	Popularity	MSRP
0	0	1	1	0.666666667	0.666666667	0	0	0	0.005847953	0.0076923	0.144650752	1
0	0	1	1	0.666666667	0.666666667	0	0	0	0.005847953	0.0076923	0.144650752	0.825509
0	0	1	1	0.666666667	0.666666667	0	0	0	0.005847953	0.0076923	0.144650752	0.72581
0.021276596	0.2	0.998942918	0	0	0.666666667	1	1	0.066666667	0.271929825	0.6538462	0.245623342	0.064199
0.021276596	0.2	0.998942918	0	0	0.666666667	1	1	0.066666667	0.257309942	0.6461538	0.245623342	0.053297
0.021276596	0.2	0.998942918	0	0	0.666666667	1	1	0.066666667	0.251461988	0.6307692	0.245623342	0.049905
0.021276596	0.2	0.998942918	0	0	0.666666667	1	1	0.066666667	0.239766082	0.6076923	0.245623342	0.049663
0.021276596	0.2	0.998942918	0	0	0	1	1	0.066666667	0.228070175	0.6230769	0.245623342	0.044285
0.021276596	0.2	0.998942918	0	0	0.666666667	1	1	0.066666667	0.277777778	0.7230769	0.245623342	0.042395
0.021276596	0.2	0.998942918	0	0	0.666666667	1	1	0.066666667	0.274853801	0.6769231	0.245623342	0.040215
0.021276596	0.2	0.998942918	0	0	0	1	1	0.066666667	0.228070175	0.6230769	0.245623342	0.037792
0.021276596	0.2	0.998942918	0	0	0	1	1	0.066666667	0.228070175	0.6230769	0.245623342	0.037744
0.021276596	0.2	0.998942918	0	0	0.666666667	1	1	0.066666667	0.271929825	0.7307692	0.245623342	0.03755
0.021276596	0.2	0.998942918	0	0	0.666666667	1	1	0.066666667	0.263157895	0.7230769	0.245623342	0.03537
0.021276596	0.2	0.998942918	0	0	0.666666667	1	1	0.066666667	0.263157895	0.7230769	0.245623342	0.03537
0.021276596	0.2	0.998942918	0	0	0	1	1	0.066666667	0.257309942	0.6923077	0.245623342	0.035128
0.021276596	0.2	0.998942918	0	0	0.666666667	1	1	0.066666667	0.277777778	0.7230769	0.245623342	0.033432
0.021276596	0.2	0.998942918	0	0	0	1	1	0.066666667	0.228070175	0.6230769	0.245623342	0.032947
0.021276596	0.2	0.998942918	0	0	0	1	1	0.066666667	0.248538012	0.6692308	0.245623342	0.032899
0.021276596	0.2	0.998942918	0	0	0	1	1	0.066666667	0.248538012	0.6692308	0.245623342	0.032899
0.021276596	0.2	0.998942918	0	0	0.666666667	1	1	0.066666667	0.260233918	0.7	0.245623342	0.031009
0.042553191	0	0.734672304	0.75	0.666666667	0.666666667	0	0.5	0.133333333	0.01754386	0.0307692	0.204420866	0.258491
0.042553191	0	0.734672304	0.75	0.666666667	0.666666667	0	0.5	0	0.01754386	0.0307692	0.204420866	0.236784
0.063829787	0	0.714587738	0.75	0.666666667	0	0	0.5	0	0.011695906	0.0307692	0.490185676	0.154075
0.063829787	0	0.714587738	0.75	0.666666667	0	0	0.5	0	0.011695906	0.0307692	0.490185676	0.152085
0.063829787	0	0.714587738	0.75	0.666666667	0	0	0.5	0	0.011695906	0.0307692	0.490185676	0.152085
0.042553191	0	0.702959831	0.75	0.666666667	0.666666667	0	0.5	0.133333333	0.011695906	0.0230769	0.204420866	0.264935
0.042553191	0	0.702959831	0.75	0.666666667	0.666666667	0	0.5	0.133333333	0.011695906	0.0230769	0.204420866	0.264935
0.042553191	0	0.702959831	0.75	0.666666667	0.666666667	0	0.5	0	0.01754386	0.0307692	0.204420866	0.240152
0.042553191	0	0.702959831	0.75	0.666666667	0.666666667	0	0.5	0	0.01754386	0.0307692	0.204420866	0.240152
0.085106383	0	0.689217759	0.5	0.333333333	0	0	1	0	0.026315789	0.0461538	0.326967286	0.030983
0.085106383	0	0.689217759	0.5	0.333333333	0	1	1	0.066666667	0.029239766	0.0461538	0.326967286	0.030983
0.085106383	0	0.689217759	0.5	0.333333333	0	1	1	0.066666667	0.029239766	0.0461538	0.326967286	0.030983
0.085106383	0	0.689217759	0.5	1	0	0	1	0	0.026315789	0.0461538	0.326967286	0.029311

We now use this data to perform regression analysis by using the Data Analysis feature in the Data menu. We get following output:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.693768659
R Square	0.481314952
Adjusted R Square	0.480758373
Standard Error	0.021486617
Observations	11196

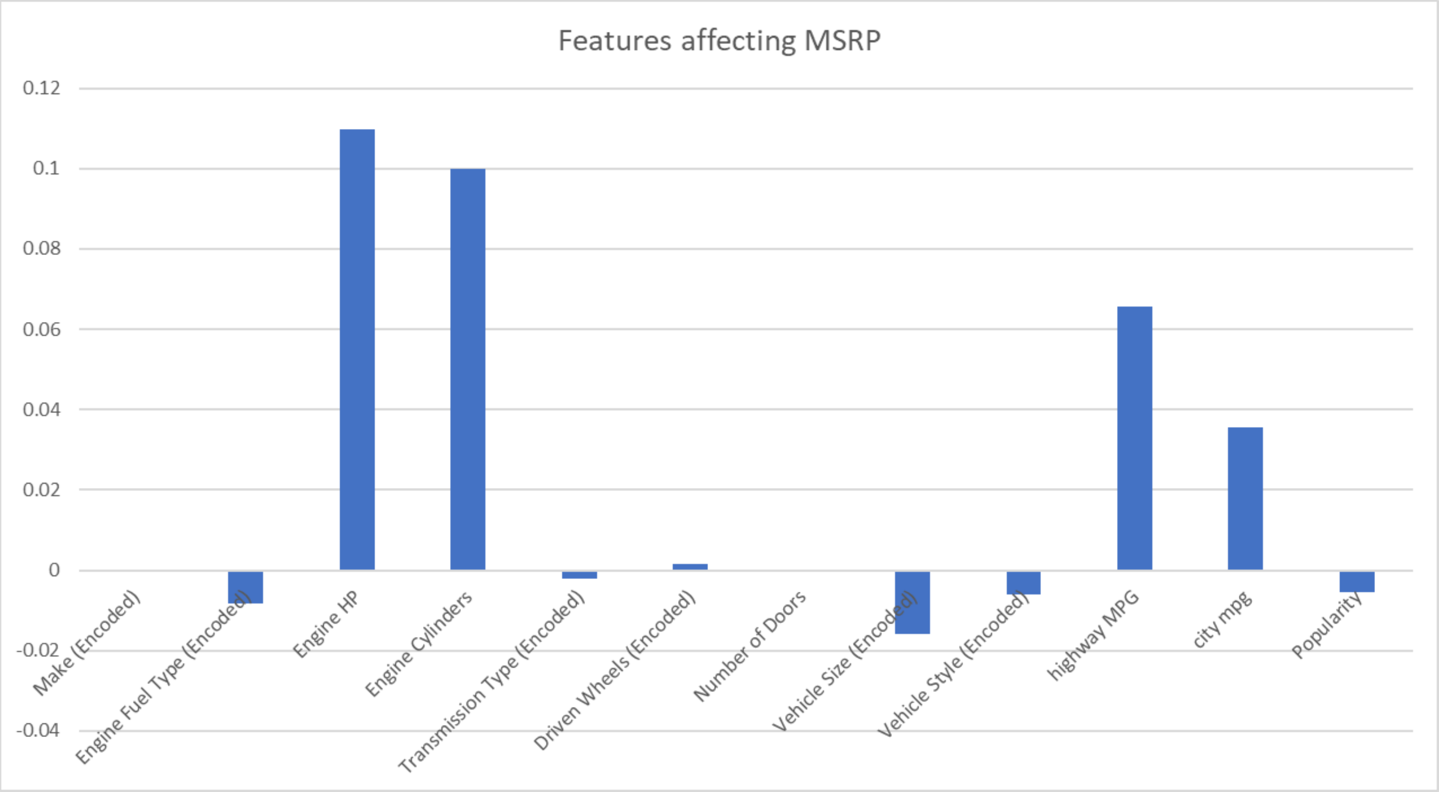
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	12	4.790932299	0.399244358	864.7741572	0
Residual	11183	5.162908281	0.000461675		
Total	11195	9.95384058			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.031589947	0.001778814	-17.75899174	1.32073E-69	-0.035076737	-0.028103158	-0.035076737	-0.028103158
Make (Encoded)	-0.000126829	0.000898814	-0.141107322	0.887787701	-0.001888663	0.001635004	-0.001888663	0.001635004
Engine Fuel Type (Encoded)	-0.008170822	0.000930401	-8.782046676	1.83872E-18	-0.009994572	-0.006347073	-0.009994572	-0.006347073
Engine HP	0.109738052	0.003028023	36.24082452	5.1054E-272	0.103802594	0.115673511	0.103802594	0.115673511
Engine Cylinders	0.099772838	0.003461532	28.82331736	2.7635E-176	0.092987626	0.106558051	0.092987626	0.106558051
Transmission Type (Encoded)	-0.002025972	0.00084897	-2.386386485	0.017031389	-0.003690103	-0.00036184	-0.003690103	-0.00036184
Driven Wheels (Encoded)	0.001653045	0.000658457	2.510481034	0.012070678	0.000362352	0.002943737	0.000362352	0.002943737
Number of Doors	-0.000224093	0.000559842	-0.400279344	0.688958426	-0.001321482	0.000873295	-0.001321482	0.000873295
Vehicle Size (Encoded)	-0.016006733	0.000670551	-23.87101698	7.0368E-123	-0.017321131	-0.014692335	-0.017321131	-0.014692335
Vehicle Style (Encoded)	-0.006067337	0.001019628	-5.950541158	2.75268E-09	-0.008065987	-0.004068687	-0.008065987	-0.004068687
highway MPG	0.065650438	0.017952962	3.65680265	0.000256553	0.030459471	0.100841405	0.030459471	0.100841405
city mpg	0.035422796	0.006433244	5.506210878	3.74738E-08	0.022812505	0.048033086	0.022812505	0.048033086
Popularity	-0.005513007	0.000923333	-5.97076887	2.43311E-09	-0.007322902	-0.003703112	-0.007322902	-0.003703112

By using this we can plot a bar graph to see which features are affecting the MSRP most.

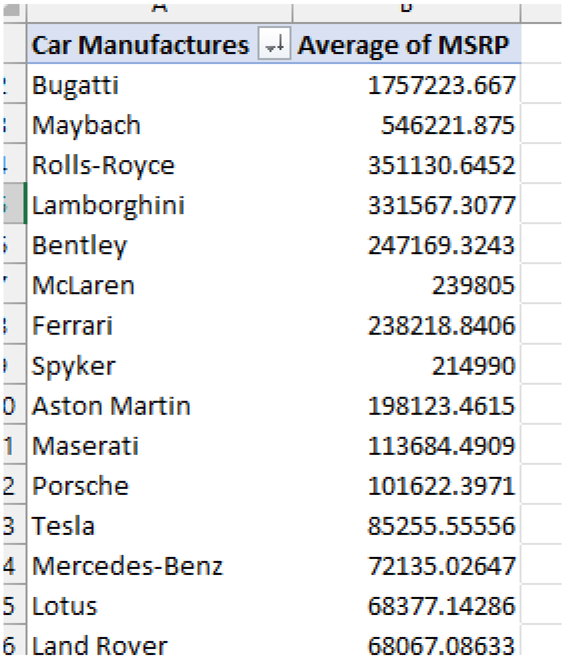
We select coefficients of each features and create a bar graph as below:



As we can see from above graph that Engine HP, Cylinders and MPG are some of the deciding factors for the MSRP of a car.

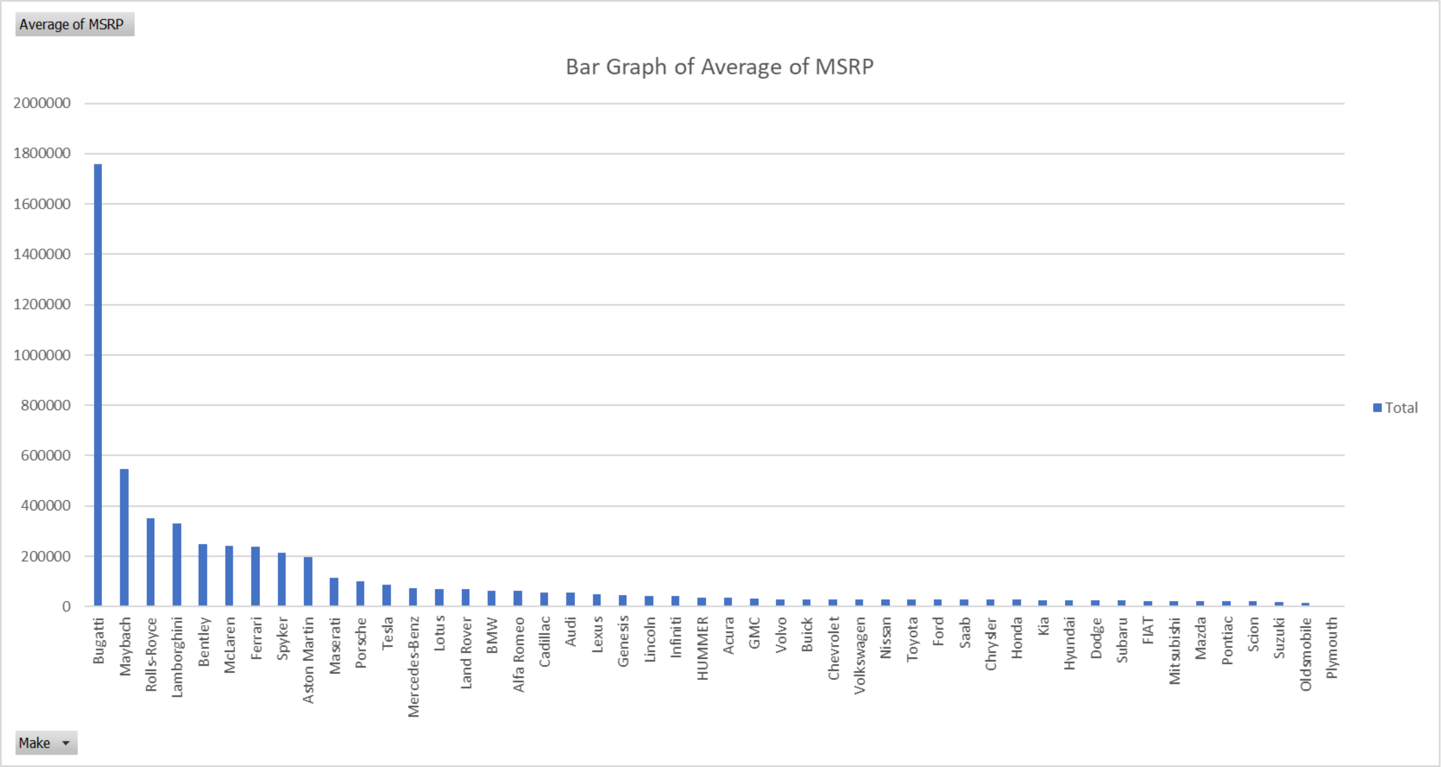
Task 4:

Insight Required: How does the average price of a car vary across different manufacturers?
For doing this task we utilized pivot tables and found the average price of a car for each car manufacturer.



Car Manufactures	Average of MSRP
Bugatti	1757223.667
Maybach	546221.875
Rolls-Royce	351130.6452
Lamborghini	331567.3077
Bentley	247169.3243
McLaren	239805
Ferrari	238218.8406
Spyker	214990
Aston Martin	198123.4615
Maserati	113684.4909
Porsche	101622.3971
Tesla	85255.55556
Mercedes-Benz	72135.02647
Lotus	68377.14286
Land Rover	68067.08633

By using this we can plot a column chart to visualize this data.



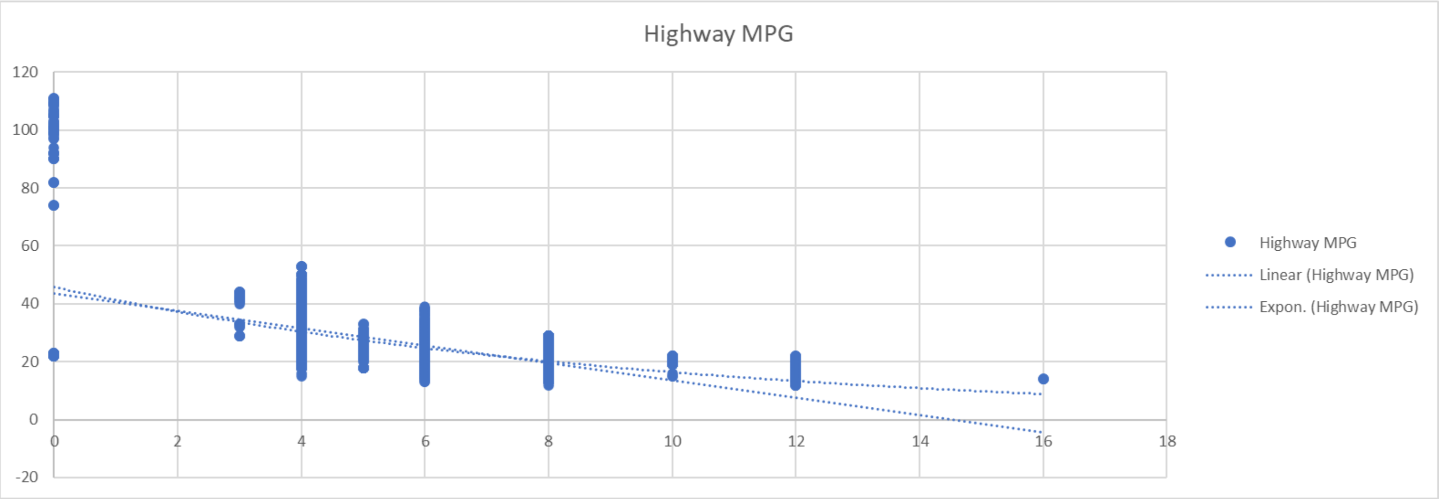
We can see here that brands such as Bugatti have very high average MSRP as they are into high end cars and don't have any cars in lower price segments.

Task 5:

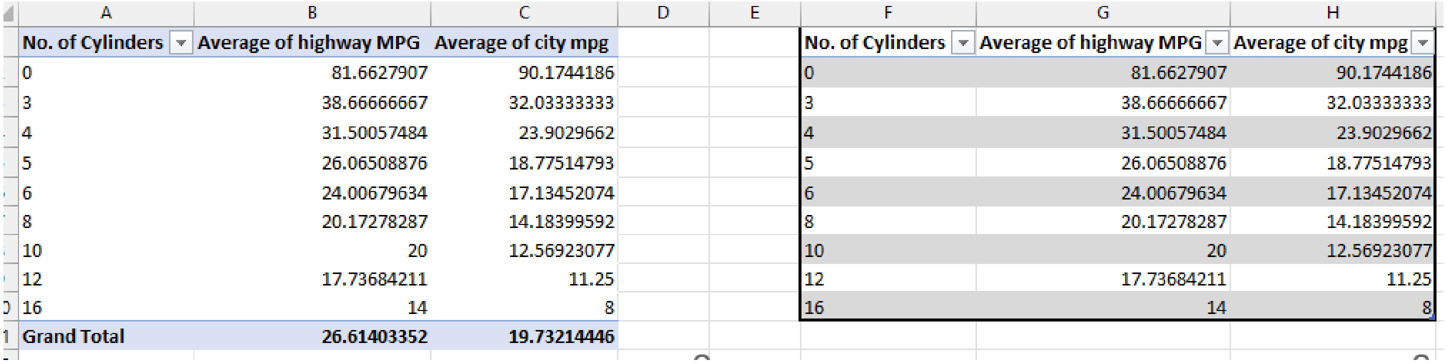
Insight Required: What is the relationship between fuel efficiency and the number of cylinders in a car's engine?

To find relationships between fuel efficiency and number of cylinders in a car’s engine we have to create a scatter plot of number of cylinders vs its MPG and see if there exists any trend by plotting a trendline.

We select two columns Engine Cylinders and Highway MPG and create a scatter plot.

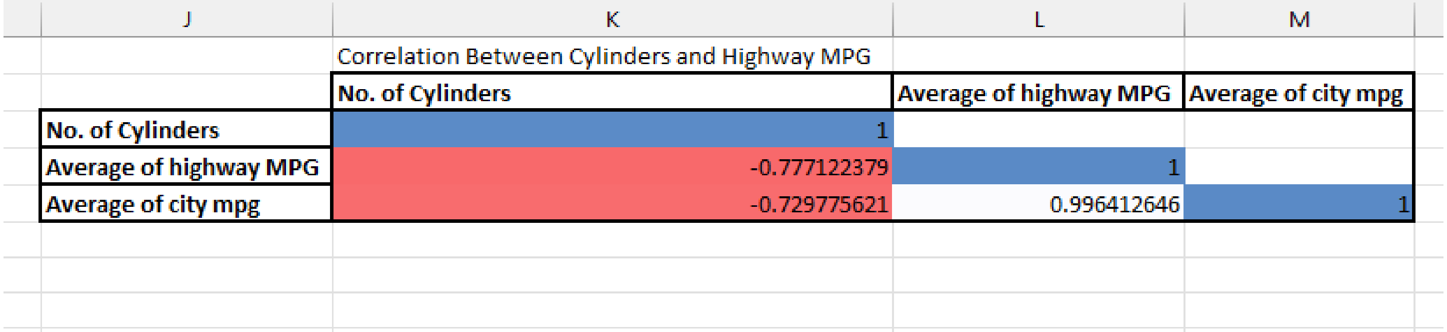


We also create a correlation matrix to check if there exist any correlation between them. We create a pivot table of no. of Cylinders and Highway MPG and City MPG. We create a correlation matrix by using the CORREL function in excel and conditional formatting.



A	B	C
No. of Cylinders	Average of highway MPG	Average of city mpg
0	81.6627907	90.1744186
3	38.66666667	32.03333333
4	31.50057484	23.9029662
5	26.06508876	18.77514793
6	24.00679634	17.13452074
8	20.17278287	14.18399592
10	20	12.56923077
12	17.73684211	11.25
16	14	8
Grand Total	26.61403352	19.73214446

F	G	H
No. of Cylinders	Average of highway MPG	Average of city mpg
0	81.6627907	90.1744186
3	38.66666667	32.03333333
4	31.50057484	23.9029662
5	26.06508876	18.77514793
6	24.00679634	17.13452074
8	20.17278287	14.18399592
10	20	12.56923077
12	17.73684211	11.25
16	14	8



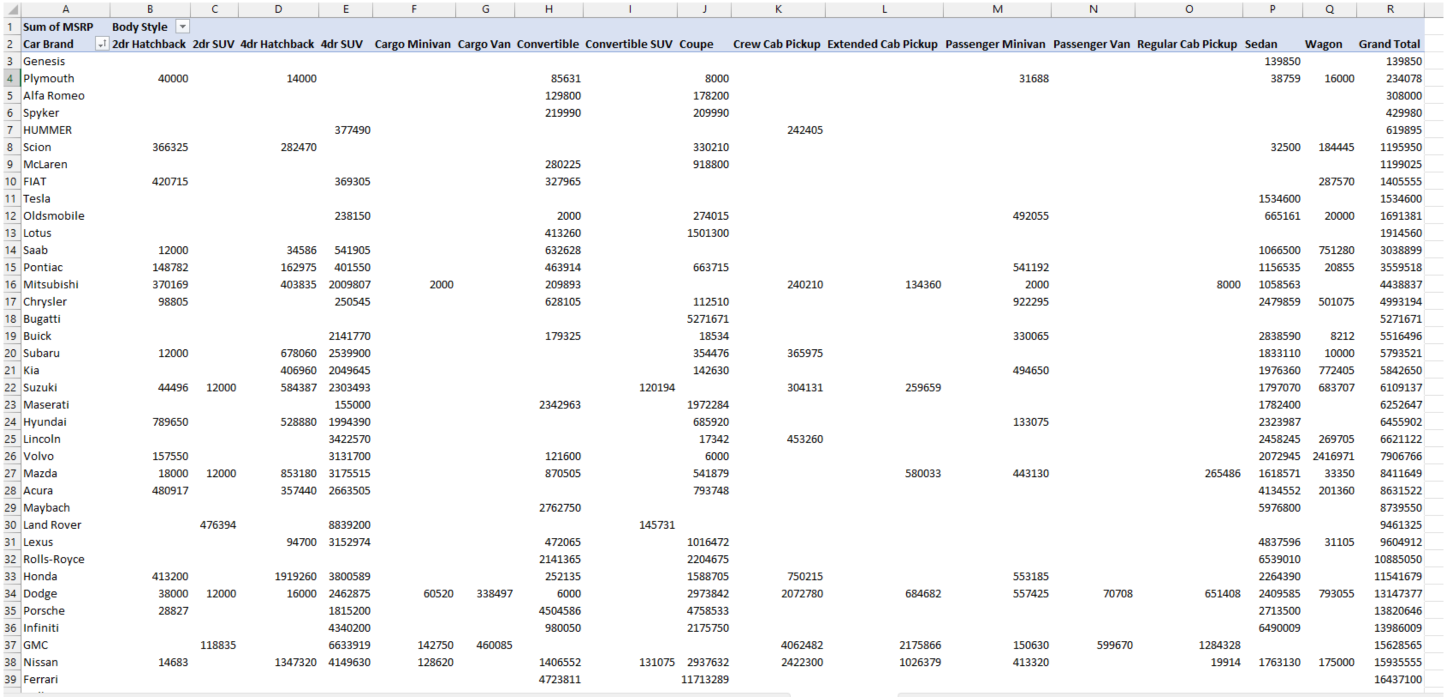
J	K	L	M
	Correlation Between Cylinders and Highway MPG		
	No. of Cylinders	Average of highway MPG	Average of city mpg
No. of Cylinders	1		
Average of highway MPG	-0.777122379	1	
Average of city mpg	-0.729775621	0.996412646	1

As we can see there is less correlation between MPG and no. cylinders. But there is high correlation between highway MPG and city MPG.

Dashboard:

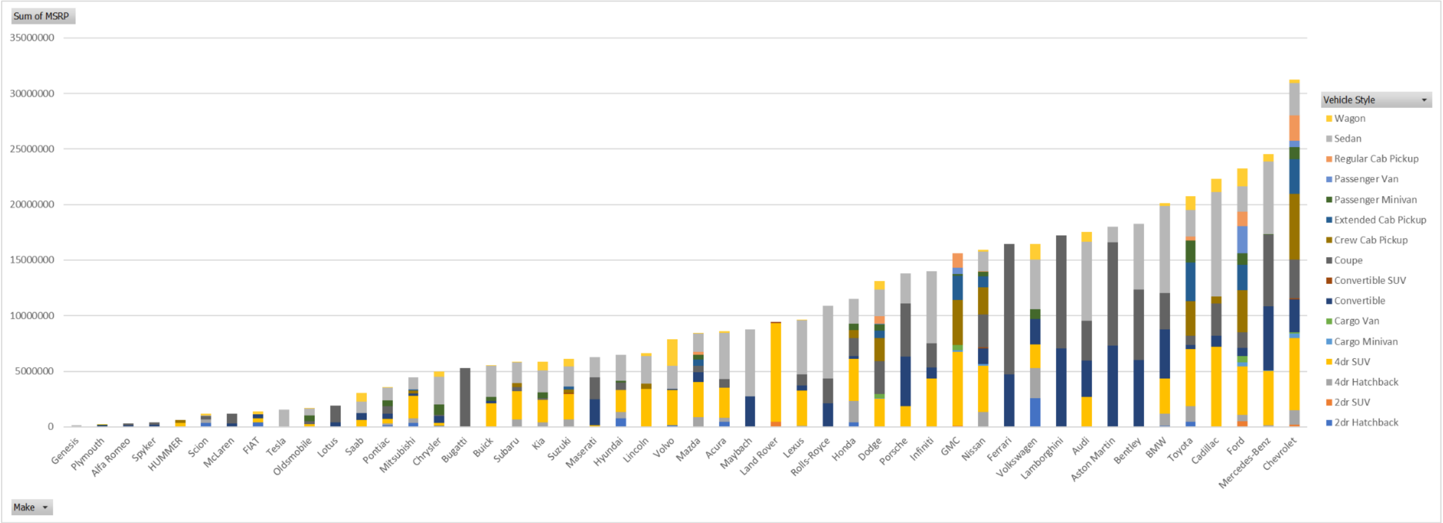
Task 1: How does the distribution of car prices vary by brand and body style?

We created a stacked column chart of car price for each brand and each body style in that column. To create this chart, we first need to create a pivot table consisting of Sum of MSRP for each category. These categories being, brand of car in row and body style in columns and we get an interactive table consisting of total MSRP for each body style that each car brand makes. Table looks like below:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Sum of MSRP	Body Style																
2	Car Brand	2dr Hatchback	2dr SUV	4dr Hatchback	4dr SUV	Cargo Minivan	Cargo Van	Convertible	Convertible SUV	Coupe	Crew Cab Pickup	Extended Cab Pickup	Passenger Minivan	Passenger Van	Regular Cab Pickup	Sedan	Wagon	Grand Total
3	Genesis															139850		139850
4	Plymouth	40000		14000				85631		8000			31688			38759	16000	234078
5	Alfa Romeo							129800		178200								308000
6	Spyker							219990		209990								429980
7	HUMMER				377490						242405							619895
8	Scion	366325		282470						330210						32500	184445	1195950
9	McLaren							280225		918800								1199025
10	FIAT	420715			369305			327965									287570	1405555
11	Tesla															1534600		1534600
12	Oldsmobile				238150			2000		274015			492055			665161	20000	1691381
13	Lotus							413260		1501300								1914560
14	Saab	12000		34586	541905			632628								1066500	751280	3038899
15	Pontiac	148782		162975	401550			463914		663715			541192			1156535	20855	3559518
16	Mitsubishi	370169		403835	2009807	2000		209893			240210	134360	2000		8000	1058563		4438837
17	Chrysler	98805			250545			628105		112510			922295			2479859	501075	4993194
18	Bugatti									5271671								5271671
19	Buick				2141770			179325		18534			330065			2838590	8212	5516496
20	Subaru	12000		678060	2539900					354476	365975					1833110	10000	5793521
21	Kia			406960	2049645					142630			494650			1976360	772405	5842650
22	Suzuki	44496	12000	584387	2303493				120194		304131	259659				1797070	683707	6109137
23	Maserati				155000			2342963		1972284						1782400		6252647
24	Hyundai	789650		528880	1994390					685920			133075			2323987		6455902
25	Lincoln				3422570					17342	453260					2458245	269705	6621122
26	Volvo	157550			3131700			121600		6000						2072945	2416971	7906766
27	Mazda	18000	12000	853180	3175515			870505		541879		580033	443130		265486	1618571	33350	8411649
28	Acura	480917		357440	2663505					793748						4134552	201360	8631522
29	Maybach							2762750								5976800		8739550
30	Land Rover		476394		8839200				145731									9461325
31	Lexus			94700	3152974			472065		1016472						4837596	31105	9604912
32	Rolls-Royce							2141365		2204675						6539010		10885050
33	Honda	413200		1919260	3800589			252135		1588705	750215		553185			2264390		11541679
34	Dodge	38000	12000	16000	2462875	60520	338497	6000		2973842	2072780	684682	557425	70708	651408	2409585	793055	13147377
35	Porsche	28827			1815200			4504586		4758533						2713500		13820646
36	Infiniti				4340200			980050		2175750						6490009		13986009
37	GMC		118835		6633919	142750	460085				4062482	2175866	150630	599670	1284328			15628565
38	Nissan	14683		1347320	4149630	128620		1406552	131075	2937632	2422300	1026379	413320		19914	1763130	175000	15935555
39	Ferrari							4723811		11713289								16437100

We can plot a stacked column chart from this table where each column in the chart for a particular brand would have the total sum of MSRP of all body styles and sections would have different colors to identify each body style. Chart looks like below:

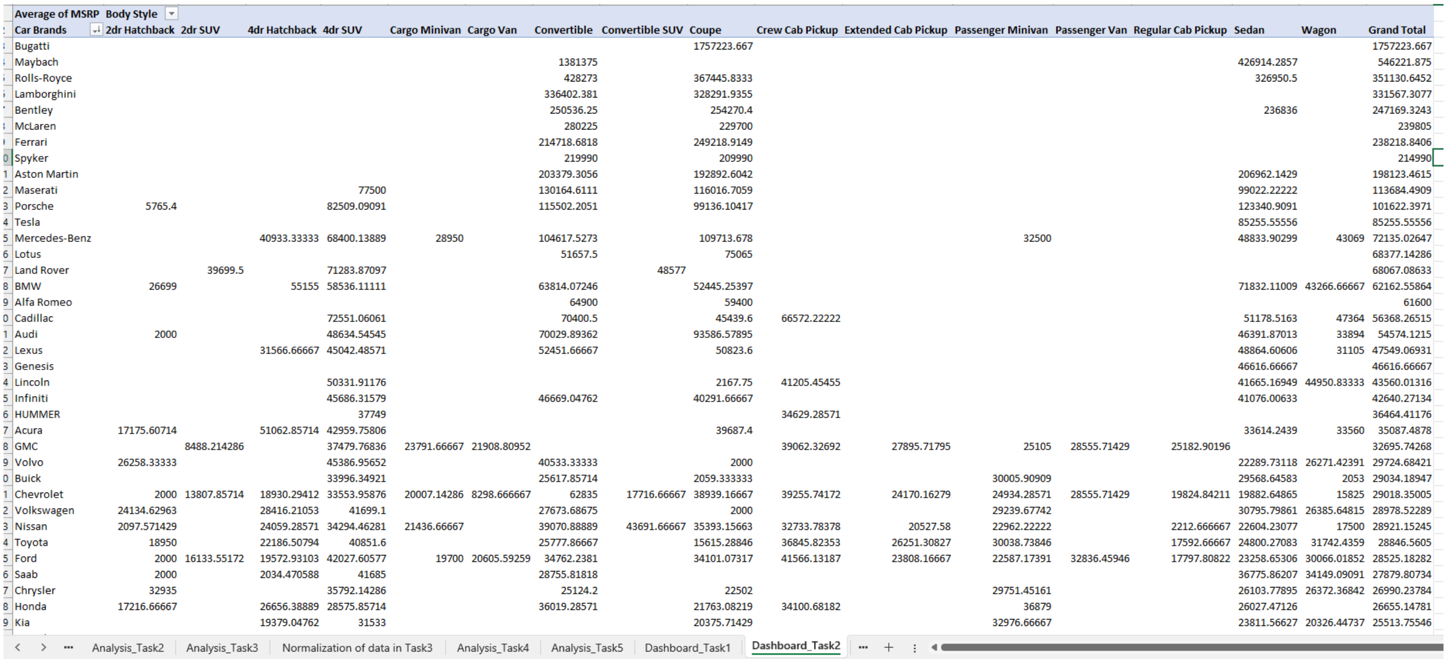


We can also change chart features by changing the filter in the pivot table.

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

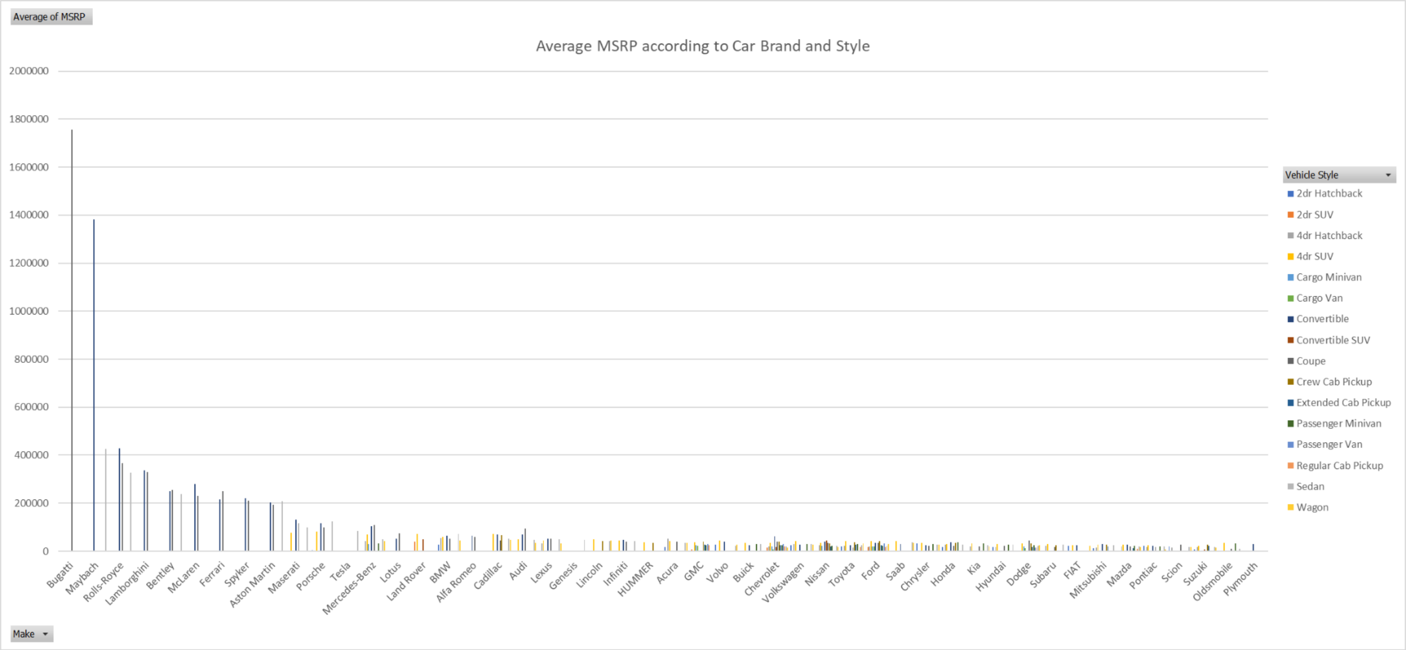
We created a clustered column chart of average car price for each brand and each body style in that column. To create this chart, we first need to create a pivot table consisting of the average of MSRP for each category. These categories being, brand of car in row and body style in columns and we get an interactive table consisting of average MSRP for each body style that each car brand makes.

Pivot table looks like below:



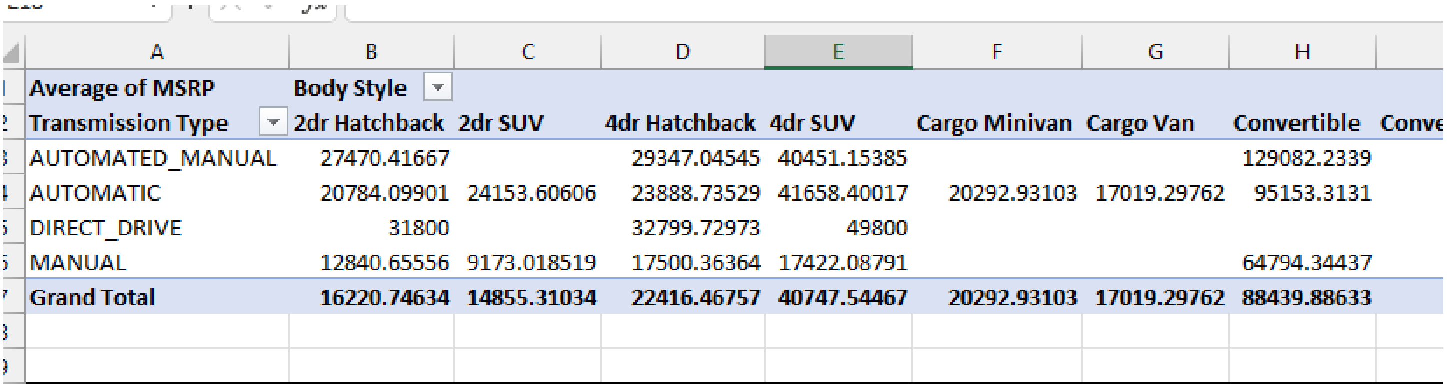
Average of MSRP	Body Style																
Car Brands	2dr Hatchback	2dr SUV	4dr Hatchback	4dr SUV	Cargo Minivan	Cargo Van	Convertible	Convertible SUV	Coupe	Crew Cab Pickup	Extended Cab Pickup	Passenger Minivan	Passenger Van	Regular Cab Pickup	Sedan	Wagon	Grand Total
Bugatti									1757223.667								1757223.667
Maybach							1381375								426914.2857		546221.875
Rolls-Royce							428273		367445.8333						326950.5		351130.6452
Lamborghini							336402.381		328291.9355								331567.3077
Bentley							250536.25		254270.4						236836		247169.3243
McLaren							280225		229700								239805
Ferrari							214718.6818		249218.9149								238218.8406
Spyker							219990		209990								214990
Aston Martin							203379.3056		192892.6042						206962.1429		198123.4615
Maserati				77500			130164.6111		116016.7059						99022.22222		113684.4909
Porsche	5765.4			82509.09091			115502.2051		99136.10417						123340.9091		101622.3971
Tesla															85255.55556		85255.55556
Mercedes-Benz			40933.33333	68400.13889	28950		104617.5273		109713.678			32500			48833.90299	43069	72135.02647
Lotus							51657.5		75065								68377.14286
Land Rover		39699.5		71283.87097				48577									68067.08633
BMW	26699		55155	58536.11111			63814.07246		52445.25397						71832.11009	43266.66667	62162.55864
Alfa Romeo							64900		59400								61600
Cadillac				72551.06061			70400.5		45439.6	66572.22222					51178.5163	47364	56368.26515
Audi	2000			48634.54545			70029.89362		93586.57895						46391.87013	33894	54574.1215
Lexus			31566.66667	45042.48571			52451.66667		50823.6						48864.60606	31105	47549.06931
Genesis															46616.66667		46616.66667
Lincoln				50331.91176					2167.75	41205.45455					41665.16949	44950.83333	43560.01316
Infiniti				45686.31579			46669.04762		40291.66667						41076.00633		42640.27134
HUMMER				37749						34629.28571							36464.41176
Acura	17175.60714		51062.85714	42959.75806					39687.4						33614.2439	33560	35087.4878
GMC		8488.214286		37479.76836	23791.66667	21908.80952				39062.32692	27895.71795	25105	28555.71429	25182.90196			32695.74268
Volvo	26258.33333			45386.95652			40533.33333		2000						22289.73118	26271.42391	29724.68421
Buick				33996.34921			25617.85714		2059.333333			30005.90909			29568.64583	2053	29034.18947
Chevrolet	2000	13807.85714	18930.29412	33553.95876	20007.14286	8298.666667	62835	17716.66667	38939.16667	39255.74172	24170.16279	24934.28571	28555.71429	19824.84211	19882.64865	15825	29018.35005
Volkswagen	24134.62963		28416.21053	41699.1			27673.68675		2000			29239.67742			30795.79861	26385.64815	28978.52289
Nissan	2097.571429		24059.28571	34294.46281	21436.66667		39070.88889	43691.66667	35393.15663	32733.78378	20527.58	22962.22222		2212.666667	22604.23077	17500	28921.15245
Toyota	18950		22186.50794	40851.6			25777.86667		15615.28846	36845.82353	26251.30827	30038.73846		17592.66667	24800.27083	31742.4359	28846.5605
Ford	2000	16133.55172	19572.93103	42027.60577	19700	20605.59259	34762.2381		34101.07317	41566.13187	23808.16667	22587.17391	32836.45946	17797.80822	23258.65306	30066.01852	28525.18282
Saab	2000		2034.470588	41685			28755.81818								36775.86207	34149.09091	27879.80734
Chrysler	32935			35792.14286			25124.2		22502			29751.45161			26103.77895	26372.36842	26990.23784
Honda	17216.66667		26656.38889	28575.85714			36019.28571		21763.08219	34100.68182		36879			26027.47126		26655.14781
Kia			19379.04762	31533					20375.71429			32976.66667			23811.56627	20326.44737	25513.75546

We can plot a clustered column chart from this table where each cluster in the chart for a particular brand would have the column of average MSRP of all body styles and sections would have different colors to identify each body style. Chart looks like below:



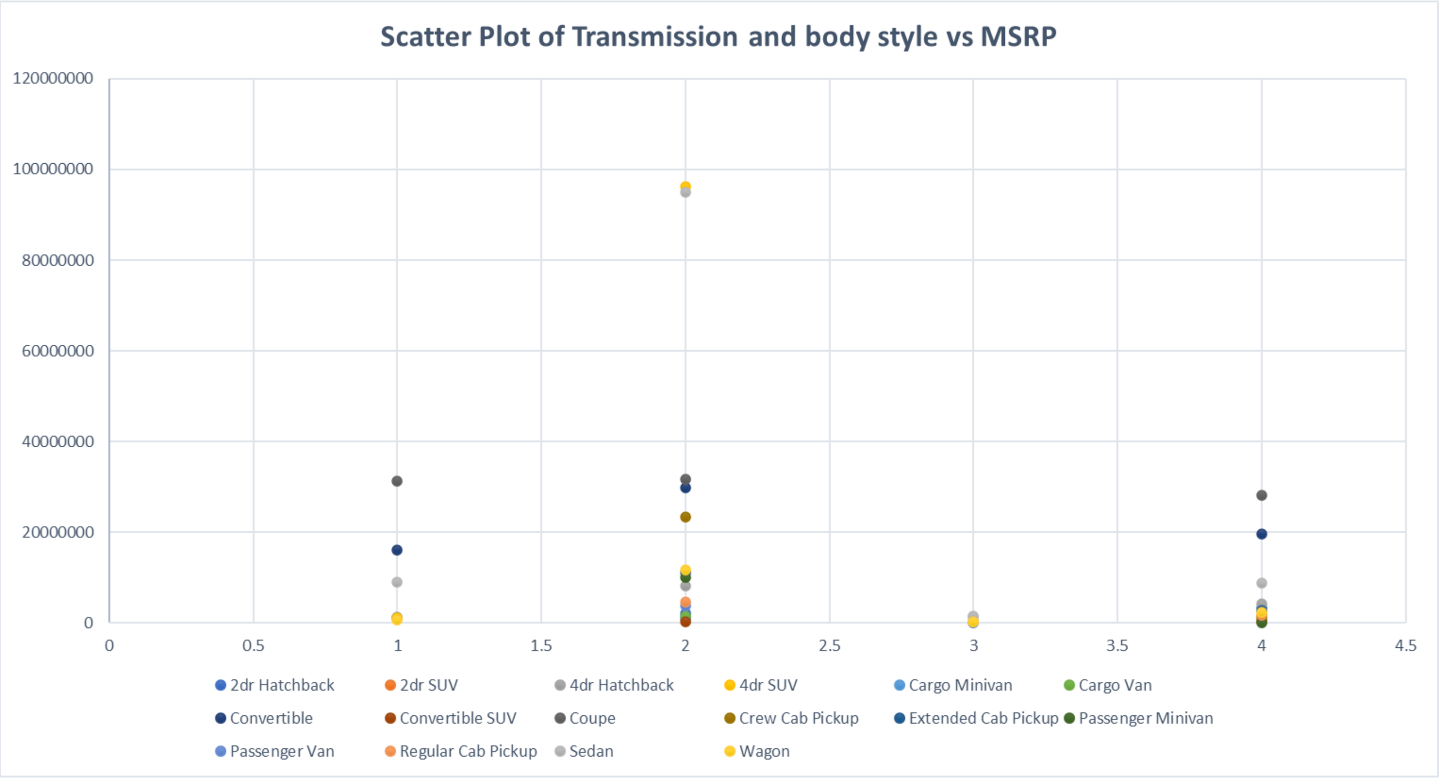
Task 3: How do the different features such as transmission type affect the MSRP, and how does this vary by body style?

To find the effect of transmission type on MSRP we have to create a pivot table and add body style as column and transmission type as row. We consider the average of MSRP to better visualize the data. We get the following pivot table.

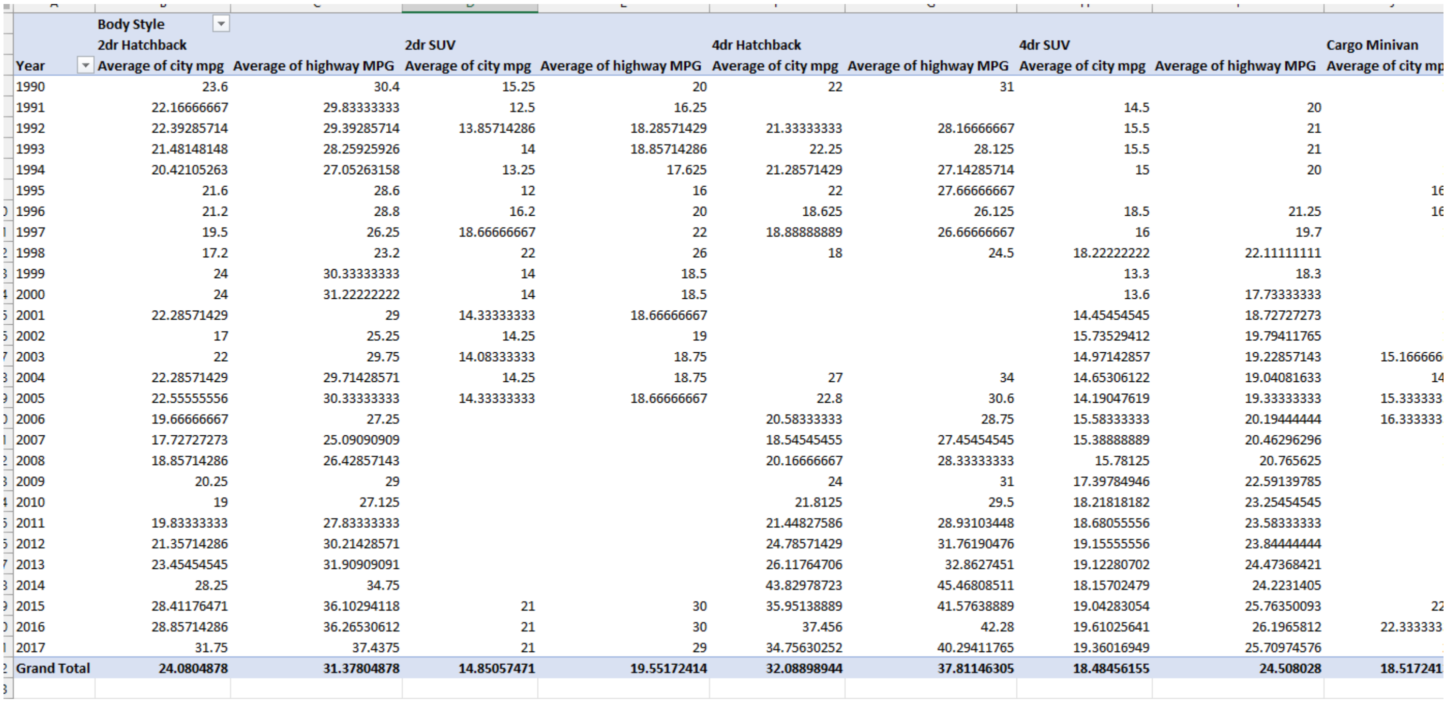


Average of MSRP	Body Style							
Transmission Type	2dr Hatchback	2dr SUV	4dr Hatchback	4dr SUV	Cargo Minivan	Cargo Van	Convertible	Conve
AUTOMATED_MANUAL	27470.41667		29347.04545	40451.15385			129082.2339	
AUTOMATIC	20784.09901	24153.60606	23888.73529	41658.40017	20292.93103	17019.29762	95153.3131	
DIRECT_DRIVE	31800		32799.72973	49800				
MANUAL	12840.65556	9173.018519	17500.36364	17422.08791			64794.34437	
Grand Total	**16220.74634**	**14855.31034**	**22416.46757**	**40747.54467**	**20292.93103**	**17019.29762**	**88439.88633**	

We copy down the contents of the pivot table to create a scatter plot. Scatter plot look like below:

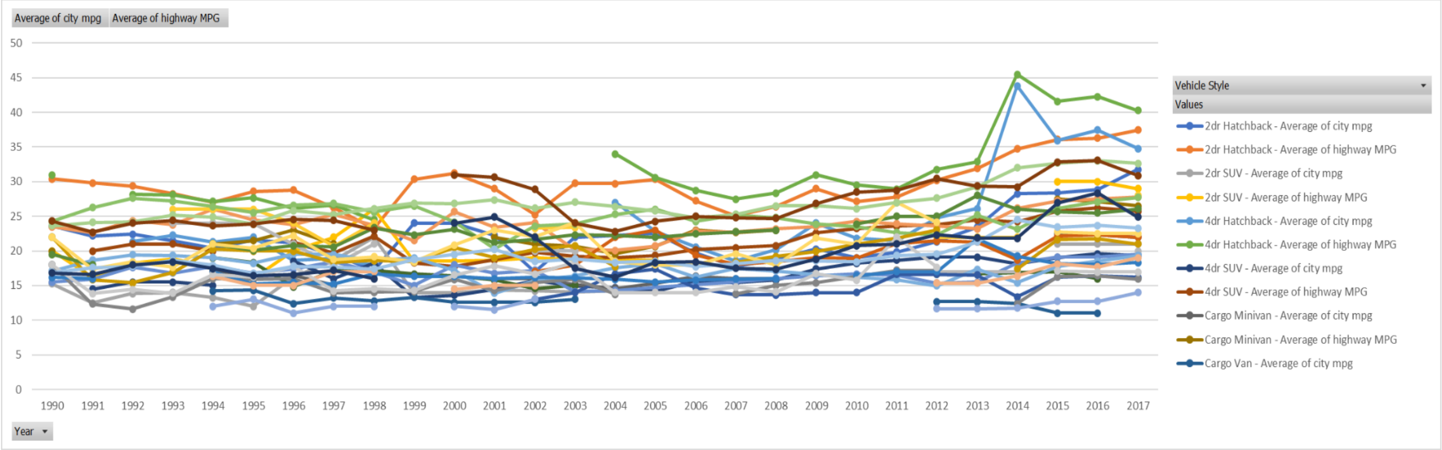


Task 4: How does the fuel efficiency of cars vary across different body styles and model years? To find how fuel efficiency of a car varies across different body styles across different years we create a pivot table consisting of average highway MPG and City MPG across body styles as columns and years as rows. We get following table:

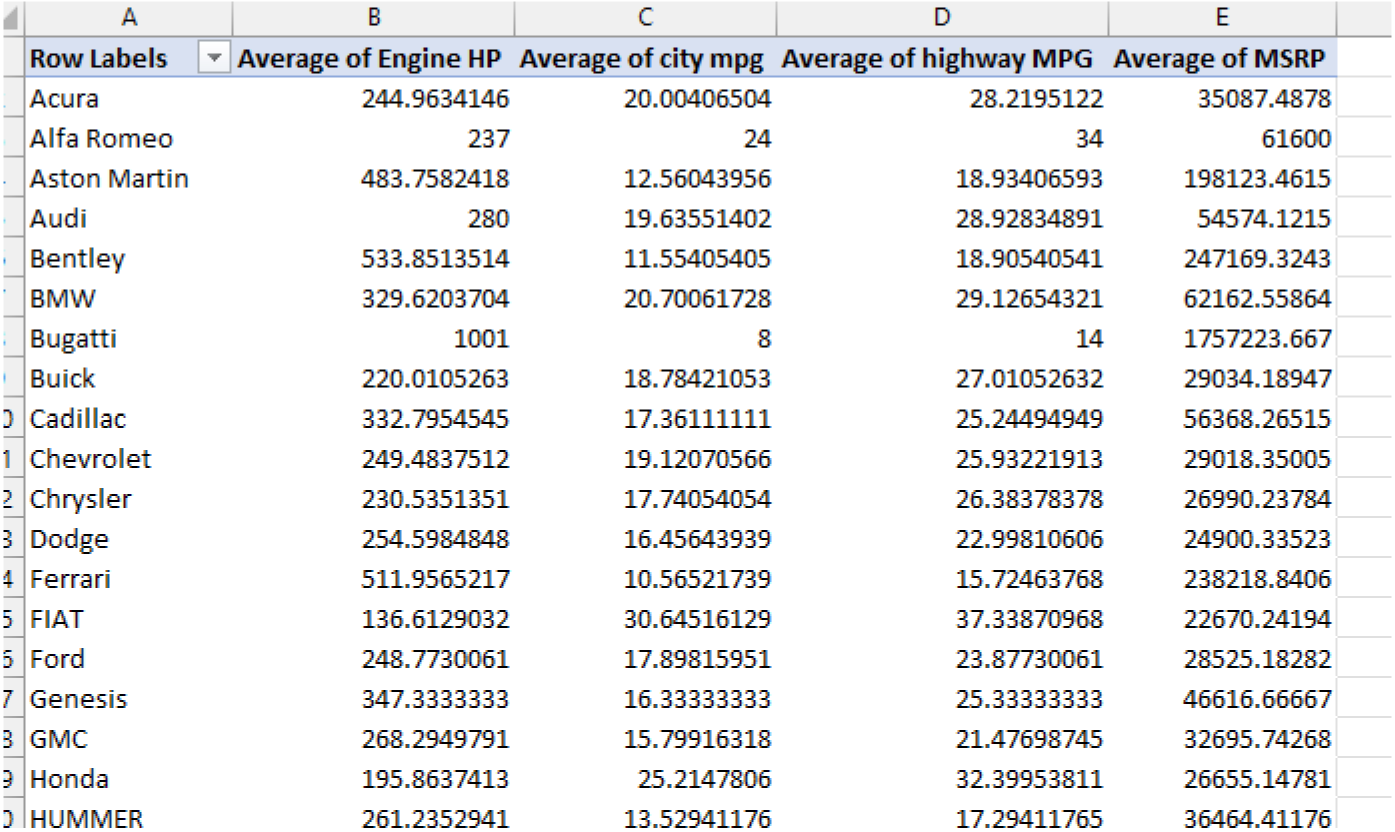


	Body Style								
	2dr Hatchback		2dr SUV		4dr Hatchback		4dr SUV		Cargo Minivan
Year	Average of city mpg	Average of highway MPG	Average of city mpg	Average of highway MPG	Average of city mpg	Average of highway MPG	Average of city mpg	Average of highway MPG	Average of city mp
1990	23.6	30.4	15.25	20	22	31			
1991	22.16666667	29.83333333	12.5	16.25			14.5	20	
1992	22.39285714	29.39285714	13.85714286	18.28571429	21.33333333	28.16666667	15.5	21	
1993	21.48148148	28.25925926	14	18.85714286	22.25	28.125	15.5	21	
1994	20.42105263	27.05263158	13.25	17.625	21.28571429	27.14285714	15	20	
1995	21.6	28.6	12	16	22	27.66666667			16
1996	21.2	28.8	16.2	20	18.625	26.125	18.5	21.25	16
1997	19.5	26.25	18.66666667	22	18.88888889	26.66666667	16	19.7	
1998	17.2	23.2	22	26	18	24.5	18.22222222	22.11111111	
1999	24	30.33333333	14	18.5			13.3	18.3	
2000	24	31.22222222	14	18.5			13.6	17.73333333	
2001	22.28571429	29	14.33333333	18.66666667			14.45454545	18.72727273	
2002	17	25.25	14.25	19			15.73529412	19.79411765	
2003	22	29.75	14.08333333	18.75			14.97142857	19.22857143	15.166666
2004	22.28571429	29.71428571	14.25	18.75	27	34	14.65306122	19.04081633	14
2005	22.55555556	30.33333333	14.33333333	18.66666667	22.8	30.6	14.19047619	19.33333333	15.333333
2006	19.66666667	27.25			20.58333333	28.75	15.58333333	20.19444444	16.333333
2007	17.72727273	25.09090909			18.54545455	27.45454545	15.38888889	20.46296296	
2008	18.85714286	26.42857143			20.16666667	28.33333333	15.78125	20.765625	
2009	20.25	29			24	31	17.39784946	22.59139785	
2010	19	27.125			21.8125	29.5	18.21818182	23.25454545	
2011	19.83333333	27.83333333			21.44827586	28.93103448	18.68055556	23.58333333	
2012	21.35714286	30.21428571			24.78571429	31.76190476	19.15555556	23.84444444	
2013	23.45454545	31.90909091			26.11764706	32.8627451	19.12280702	24.47368421	
2014	28.25	34.75			43.82978723	45.46808511	18.15702479	24.2231405	
2015	28.41176471	36.10294118	21	30	35.95138889	41.57638889	19.04283054	25.76350093	22
2016	28.85714286	36.26530612	21	30	37.456	42.28	19.61025641	26.1965812	22.333333
2017	31.75	37.4375	21	29	34.75630252	40.29411765	19.36016949	25.70974576	
Grand Total	**24.0804878**	**31.37804878**	**14.85057471**	**19.55172414**	**32.08898944**	**37.81146305**	**18.48456155**	**24.508028**	**18.517241**

We use this table to create a line plot with markers to visualize the data as a timeline and across different body types.

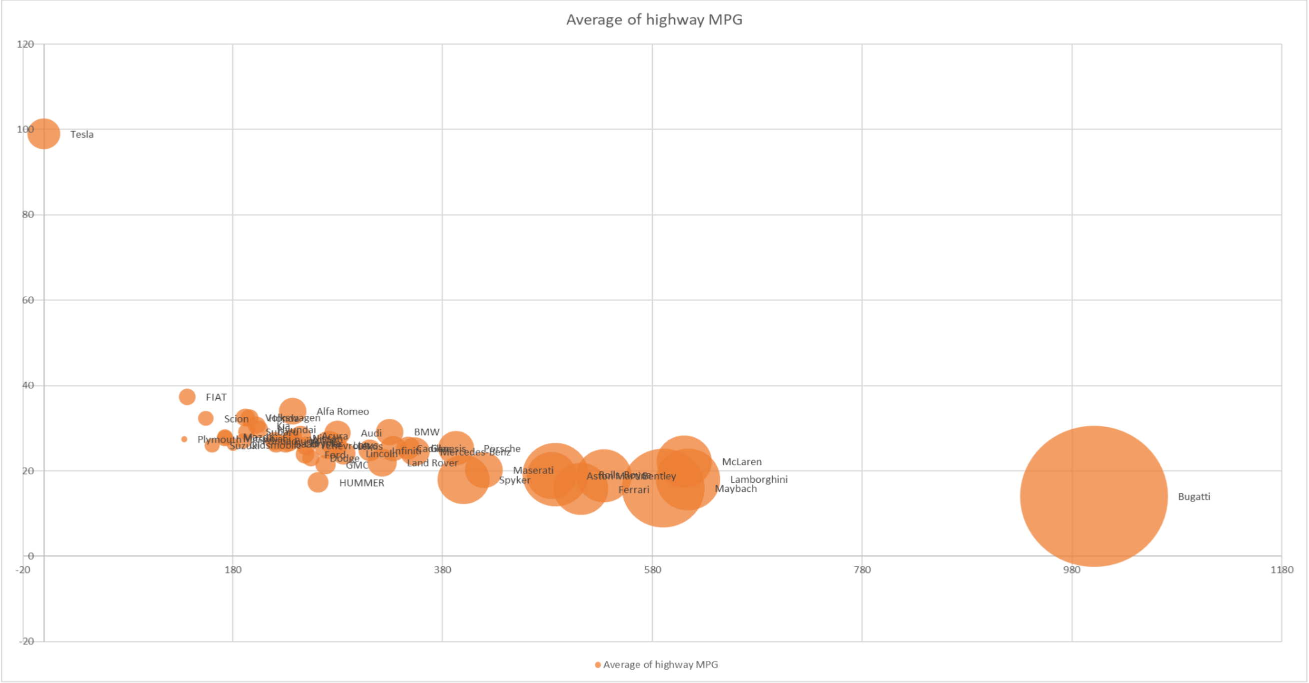


Task 5: How does the car's horsepower, MPG, and price vary across different Brands?
To find relationships between a car's horsepower and MPG and price across different brands we can create a bubble plot to better visualize, for this we find average MPG , price and horsepower across each brand and create a pivot table. This pivot table looks like below:



A	B	C	D	E
Row Labels	Average of Engine HP	Average of city mpg	Average of highway MPG	Average of MSRP
Acura	244.9634146	20.00406504	28.2195122	35087.4878
Alfa Romeo	237	24	34	61600
Aston Martin	483.7582418	12.56043956	18.93406593	198123.4615
Audi	280	19.63551402	28.92834891	54574.1215
Bentley	533.8513514	11.55405405	18.90540541	247169.3243
BMW	329.6203704	20.70061728	29.12654321	62162.55864
Bugatti	1001	8	14	1757223.667
Buick	220.0105263	18.78421053	27.01052632	29034.18947
Cadillac	332.7954545	17.36111111	25.24494949	56368.26515
Chevrolet	249.4837512	19.12070566	25.93221913	29018.35005
Chrysler	230.5351351	17.74054054	26.38378378	26990.23784
Dodge	254.5984848	16.45643939	22.99810606	24900.33523
Ferrari	511.9565217	10.56521739	15.72463768	238218.8406
FIAT	136.6129032	30.64516129	37.33870968	22670.24194
Ford	248.7730061	17.89815951	23.87730061	28525.18282
Genesis	347.3333333	16.33333333	25.33333333	46616.66667
GMC	268.2949791	15.79916318	21.47698745	32695.74268
Honda	195.8637413	25.2147806	32.39953811	26655.14781
HUMMER	261.2352941	13.52941176	17.29411765	36464.41176

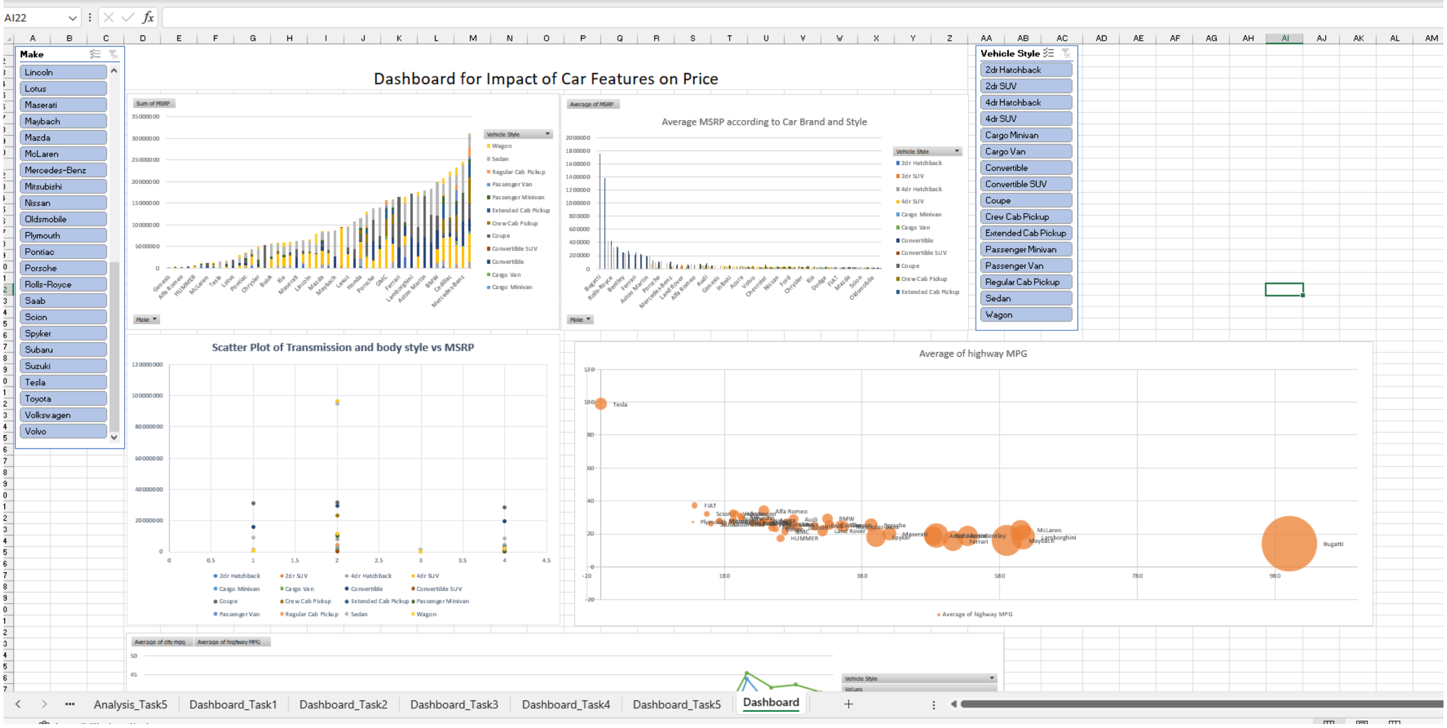
We copy this contents to create a table to create a bubble plot.
On the X axis there would be average horsepower and on the Y axis average MPG. Each bubble would represent each car brand and would be labeled to better identify the brands



Making Dashboard:

We have created each chart to visualize different parameters and relationships and trends, we can now create a dashboard by combining all these charts into one single worksheet to have a better understanding of the data. We copy all these charts into one worksheet and add slicers to change parameters which are shown the data point for. We add two slicers in the worksheet, make, and vehicle style, which represent car brand and body style. We then make connections with these slicers with all the charts to make the dashboard functional. Now we can easily find different trends and relationships between price and parameters.

Dashboard looks like below:



Results:

While working on this project, I have gained a better understanding of Impact of Car Features on Price and Profitability as well as popularity of the Car. I have improved my understanding of Advanced Excel methodologies. By analyzing Car features Data, I was able to provide insights on various aspects such as Features most affecting MSRP, Outliers in the Data, relation between Engine HP and MSRP, Regression Analysis, average MSRP across different brands and relation between no. of cylinders and fuel efficiency. I was also able to create different visualizations to improve data understanding and create a dashboard for ease of understanding between various parameters in Car Features.

This project has helped me enhance my Excel skills, particularly in data visualization and creating pivot tables and charts to derive meaningful insights. It has also improved my ability to interpret data and provide actionable recommendations based on the analysis.

ABC Call Volume Trend Analysis

Excel Sheet:

https://docs.google.com/spreadsheets/d/1lqsTTNInPiGetvTCOAvujjXyQyGqSLMX/edit?usp=sharing&ouid=107365393175079460343&rtpof=true&sd=true

Project Description:

This project aims to analyze a dataset containing information about call trends in a company's customer care. The goal is to gain insights about trends in calls received, such as average call duration, call volume analysis, distribution of calls across various time buckets, and employee distribution according to call volume.. The data provided has various missing or null Data, our task is to handle those missing values appropriately, by either deleting or imputing these data. We utilize various excel features such as pivot tables and charts to better represent data. We find trends in call volume and employee distribution by implementing various methodologies and formulae and visualization techniques in Excel. Thus, by employing statistics and Excel formulas, we will extract meaningful conclusions to help understand how to better utilize manpower in handling the call volumes across various time buckets.

Approach:

As an individual working on this project, I followed a structured approach to analyze data about Call volume and Employees. I began by carefully examining the provided database and familiarizing myself with its structure and columns. I tried to find columns which had the most significance in the dataset. I handled missing values by eliminating columns which had most empty cells, and were not significant. And imputed data into cells that were necessary for analysis. Then, I utilized Excel fundamentals to retrieve the necessary information for each task, employing appropriate functions and statistical methods. I focused on data accuracy and quality throughout the project, ensuring reliable results. By leveraging my Excel skills and maintaining a systematic workflow, I successfully executed the project and created a comprehensive report that fulfilled the objectives of providing marketing insights and investor metrics.

Tech-Stack Used:

For this project, I utilized Microsoft Excel as the primary software tool.

Data Cleaning:

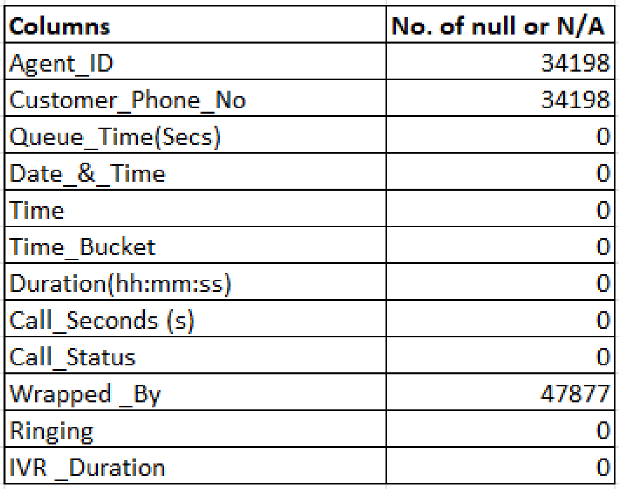
Given Data had various missing values, for better analysis of the data, we had to handle these missing values.

We found missing values by using following formulae:

=COUNTIF(B:B,"#N/A")

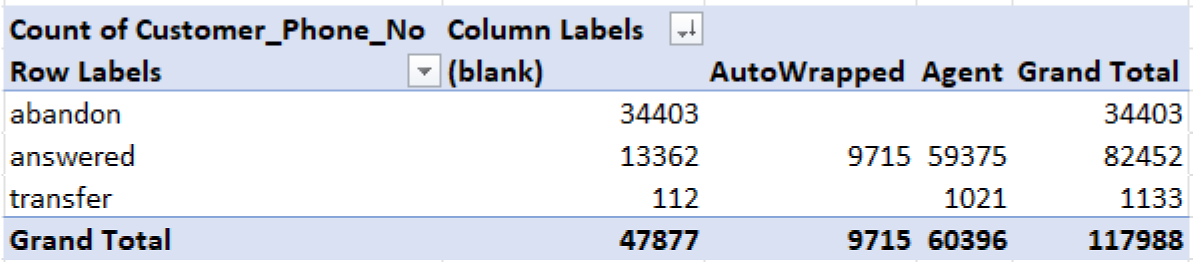
=COUNTBLANK(Table1[Wrapped _By])

Output:



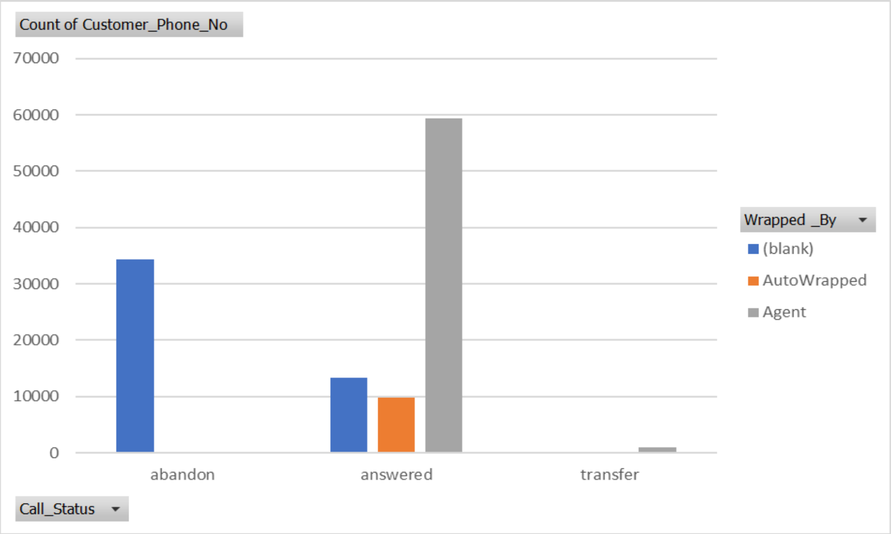
Columns	No. of null or N/A
Agent_ID	34198
Customer_Phone_No	34198
Queue_Time(Secs)	0
Date_&_Time	0
Time	0
Time_Bucket	0
Duration(hh:mm:ss)	0
Call_Seconds (s)	0
Call_Status	0
Wrapped _By	47877
Ringing	0
IVR _Duration	0

As there is a lot of missing data in Wrapped_By column we begin by handling these values. We create a pivot table to understand the data.



Count of Customer_Phone_No	Column Labels			
Row Labels	(blank)	AutoWrapped	Agent	Grand Total
abandon	34403			34403
answered	13362	9715	59375	82452
transfer	112		1021	1133
Grand Total	**47877**	**9715**	**60396**	**117988**

As we can see here most of the blank values are in abandon calls. To better understand the data we visualize it using Bar Chart

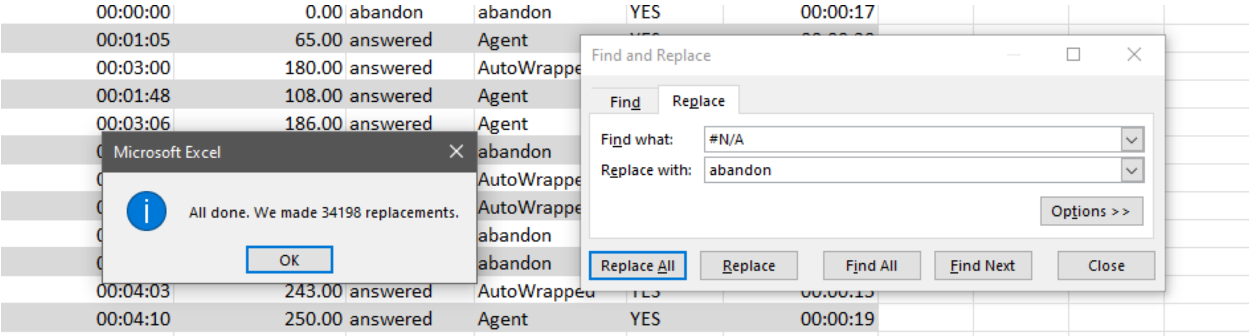


As most of the calls answered and transferred are by agent we check with the call_status column and impute appropriate values in the Wrapped_By column. We impute value Agent, if call is answered or transferred, else abandon.

Formula:

=IF(I7="abandon","abandon","Agent")

Agent ID column also had missing values where call were abandoned, so we replaced #NA with abandon.



00:00:00	0.00 abandon	abandon	YES	00:00:17
00:01:05	65.00 answered	Agent		
00:03:00	180.00 answered	AutoWrappe		
00:01:48	108.00 answered	Agent		
00:03:06	186.00 answered	Agent		
(		abandon		
(		AutoWrappe		
(		AutoWrappe		
(		abandon		
(		abandon		
00:04:03	243.00 answered	AutoWrapped		
00:04:10	250.00 answered	Agent	YES	00:00:19

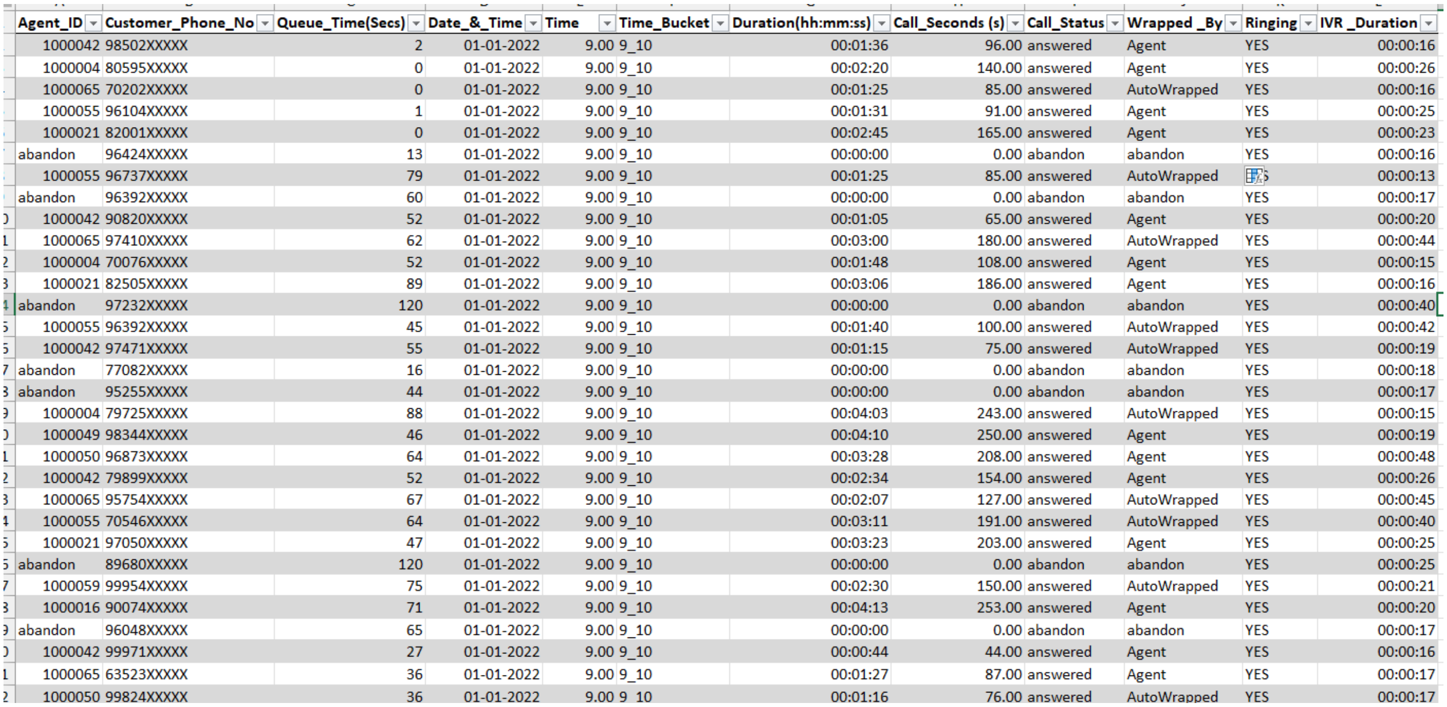
Microsoft Excel

All done. We made 34198 replacements.

Find what: #N/A

Replace with: abandon

Thus, we got our cleaned Data.

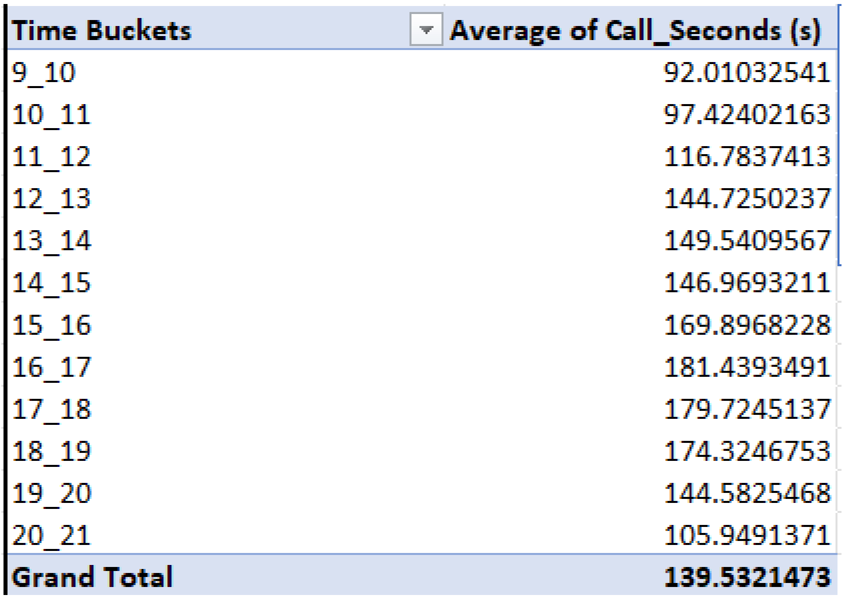


Agent_ID	Customer_Phone_No	Queue_Time(Secs)	Date_&_Time	Time	Time_Bucket	Duration(hh:mm:ss)	Call_Seconds (s)	Call_Status	Wrapped _By	Ringing	IVR _Duration
1000042	98502XXXXX	2	01-01-2022	9.00	9_10	00:01:36	96.00	answered	Agent	YES	00:00:16
1000004	80595XXXXX	0	01-01-2022	9.00	9_10	00:02:20	140.00	answered	Agent	YES	00:00:26
1000065	70202XXXXX	0	01-01-2022	9.00	9_10	00:01:25	85.00	answered	AutoWrapped	YES	00:00:16
1000055	96104XXXXX	1	01-01-2022	9.00	9_10	00:01:31	91.00	answered	Agent	YES	00:00:25
1000021	82001XXXXX	0	01-01-2022	9.00	9_10	00:02:45	165.00	answered	Agent	YES	00:00:23
abandon	96424XXXXX	13	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	abandon	YES	00:00:16
1000055	96737XXXXX	79	01-01-2022	9.00	9_10	00:01:25	85.00	answered	AutoWrapped	S	00:00:13
abandon	96392XXXXX	60	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	abandon	YES	00:00:17
1000042	90820XXXXX	52	01-01-2022	9.00	9_10	00:01:05	65.00	answered	Agent	YES	00:00:20
1000065	97410XXXXX	62	01-01-2022	9.00	9_10	00:03:00	180.00	answered	AutoWrapped	YES	00:00:44
1000004	70076XXXXX	52	01-01-2022	9.00	9_10	00:01:48	108.00	answered	Agent	YES	00:00:15
1000021	82505XXXXX	89	01-01-2022	9.00	9_10	00:03:06	186.00	answered	Agent	YES	00:00:16
abandon	97232XXXXX	120	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	abandon	YES	00:00:40
1000055	96392XXXXX	45	01-01-2022	9.00	9_10	00:01:40	100.00	answered	AutoWrapped	YES	00:00:42
1000042	97471XXXXX	55	01-01-2022	9.00	9_10	00:01:15	75.00	answered	AutoWrapped	YES	00:00:19
abandon	77082XXXXX	16	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	abandon	YES	00:00:18
abandon	95255XXXXX	44	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	abandon	YES	00:00:17
1000004	79725XXXXX	88	01-01-2022	9.00	9_10	00:04:03	243.00	answered	AutoWrapped	YES	00:00:15
1000049	98344XXXXX	46	01-01-2022	9.00	9_10	00:04:10	250.00	answered	Agent	YES	00:00:19
1000050	96873XXXXX	64	01-01-2022	9.00	9_10	00:03:28	208.00	answered	Agent	YES	00:00:48
1000042	79899XXXXX	52	01-01-2022	9.00	9_10	00:02:34	154.00	answered	Agent	YES	00:00:26
1000065	95754XXXXX	67	01-01-2022	9.00	9_10	00:02:07	127.00	answered	AutoWrapped	YES	00:00:45
1000055	70546XXXXX	64	01-01-2022	9.00	9_10	00:03:11	191.00	answered	AutoWrapped	YES	00:00:40
1000021	97050XXXXX	47	01-01-2022	9.00	9_10	00:03:23	203.00	answered	Agent	YES	00:00:25
abandon	89680XXXXX	120	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	abandon	YES	00:00:25
1000059	99954XXXXX	75	01-01-2022	9.00	9_10	00:02:30	150.00	answered	AutoWrapped	YES	00:00:21
1000016	90074XXXXX	71	01-01-2022	9.00	9_10	00:04:13	253.00	answered	Agent	YES	00:00:20
abandon	96048XXXXX	65	01-01-2022	9.00	9_10	00:00:00	0.00	abandon	abandon	YES	00:00:17
1000042	99971XXXXX	27	01-01-2022	9.00	9_10	00:00:44	44.00	answered	Agent	YES	00:00:16
1000065	63523XXXXX	36	01-01-2022	9.00	9_10	00:01:27	87.00	answered	Agent	YES	00:00:17
1000050	99824XXXXX	36	01-01-2022	9.00	9_10	00:01:16	76.00	answered	AutoWrapped	YES	00:00:17

Insights:

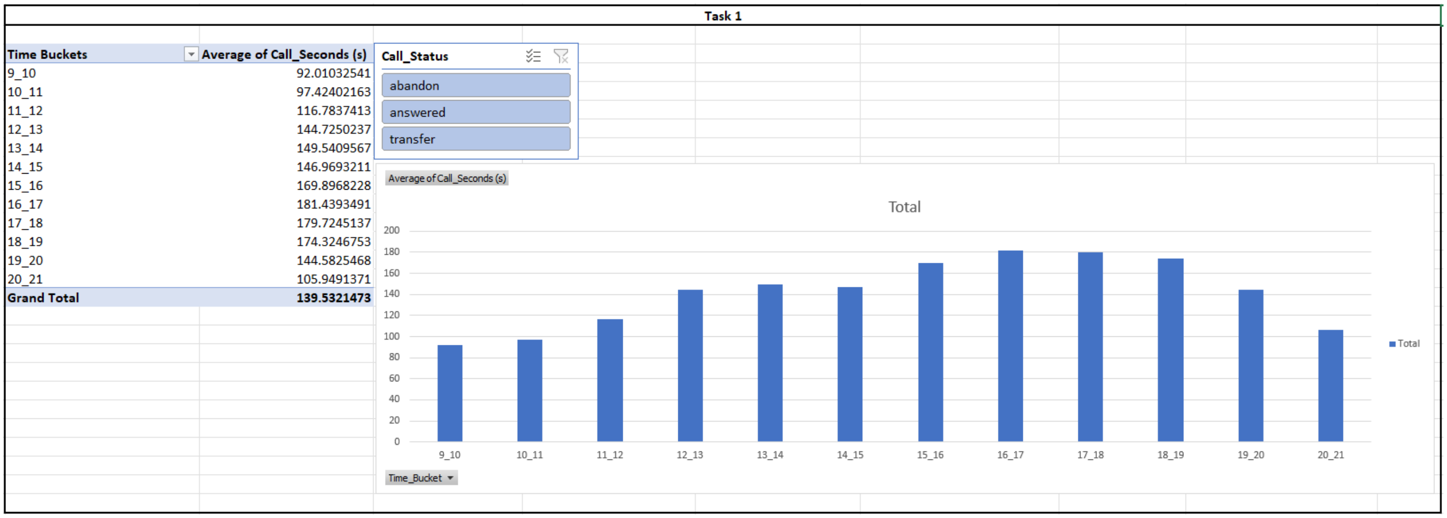
Question 1: What is the average duration of calls for each time bucket?

To find the average duration of calls for each time bucket, we can create a pivot table, we select time bucket as row and the average of call duration in seconds as the field value. We get the following Table.



Time Buckets	Average of Call_Seconds (s)
9_10	92.01032541
10_11	97.42402163
11_12	116.7837413
12_13	144.7250237
13_14	149.5409567
14_15	146.9693211
15_16	169.8968228
16_17	181.4393491
17_18	179.7245137
18_19	174.3246753
19_20	144.5825468
20_21	105.9491371
Grand Total	**139.5321473**

We can visualize this on a column chart and also create a slicer to visualize the data according to various filters.



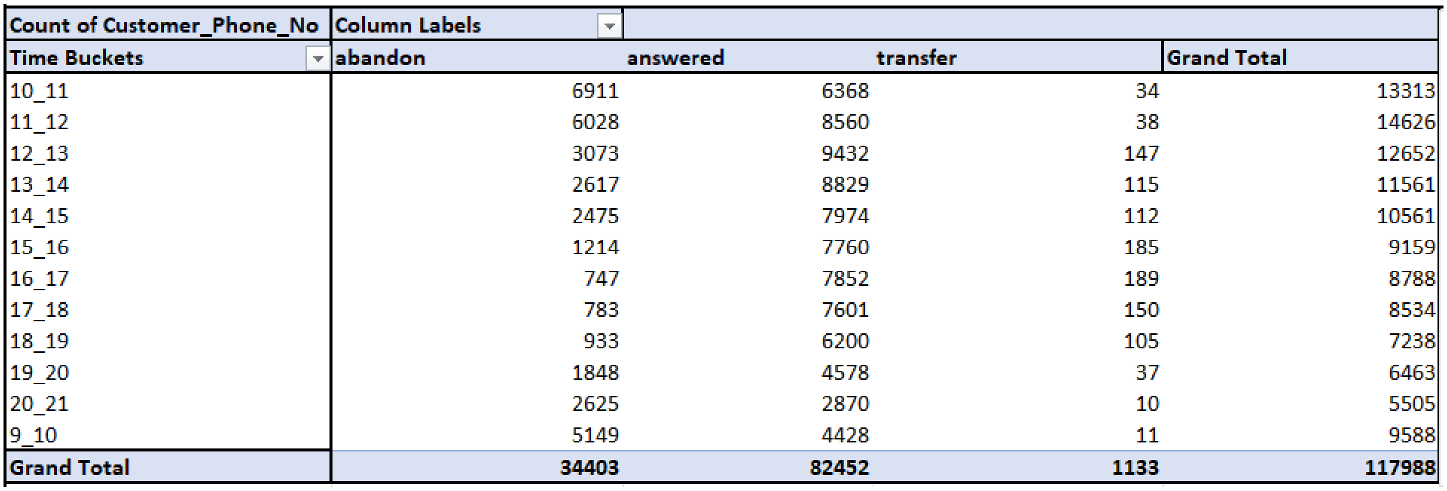
Task 1

Time Buckets	Average of Call_Seconds (s)
9_10	92.01032541
10_11	97.42402163
11_12	116.7837413
12_13	144.7250237
13_14	149.5409567
14_15	146.9693211
15_16	169.8968228
16_17	181.4393491
17_18	179.7245137
18_19	174.3246753
19_20	144.5825468
20_21	105.9491371
Grand Total	**139.5321473**

Question 2: Can you create a chart or graph that shows the number of calls received in each time bucket?

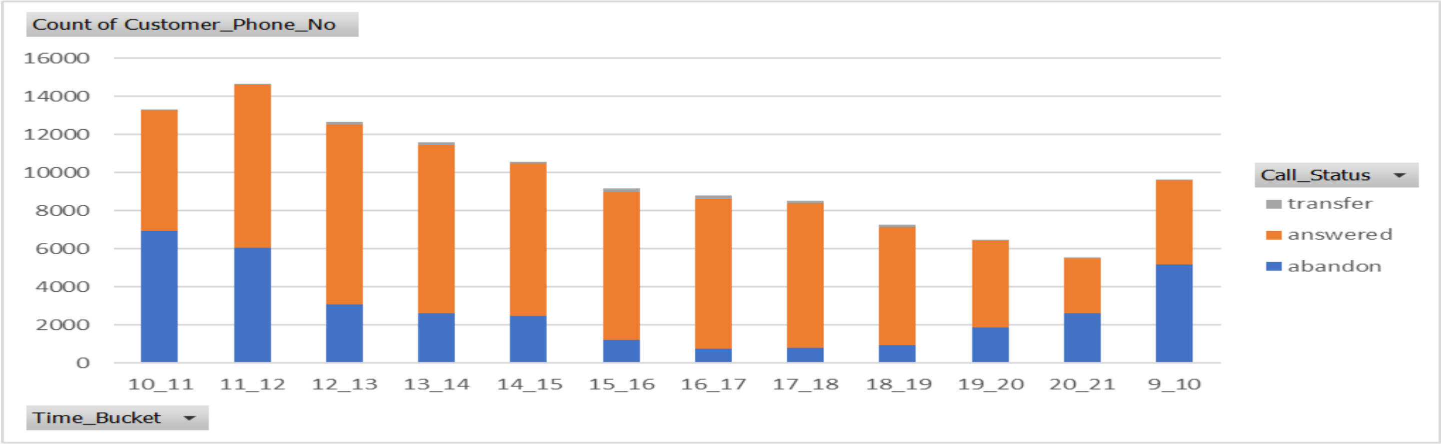
To Create a chart that shows the number of calls received in each time bucket, we first have to create a table which has a count of calls in each time bucket. To do this, we can create a pivot table. Pivot table has time buckets as rows and counts of customer phone numbers as fields, we also select call status as column to better understand trends.

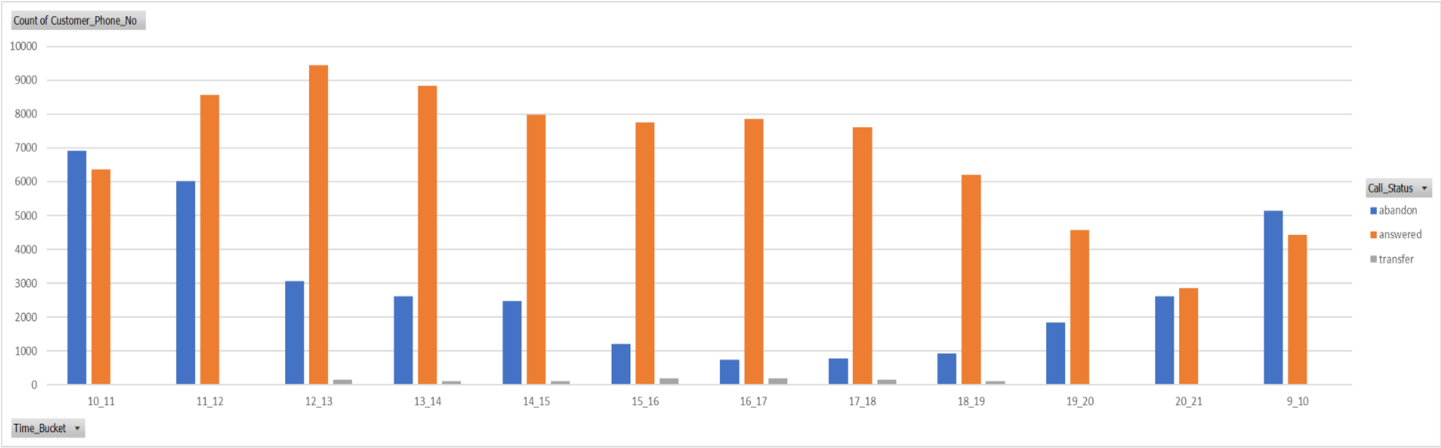
We get following pivot table:



Count of Customer_Phone_No	Column Labels			
Time Buckets	abandon	answered	transfer	Grand Total
10_11	6911	6368	34	13313
11_12	6028	8560	38	14626
12_13	3073	9432	147	12652
13_14	2617	8829	115	11561
14_15	2475	7974	112	10561
15_16	1214	7760	185	9159
16_17	747	7852	189	8788
17_18	783	7601	150	8534
18_19	933	6200	105	7238
19_20	1848	4578	37	6463
20_21	2625	2870	10	5505
9_10	5149	4428	11	9588
Grand Total	**34403**	**82452**	**1133**	**117988**

From this table we can create Column Charts that show the number of calls received in each time bucket. We create two charts one stacked column chart, and clustered column chart:



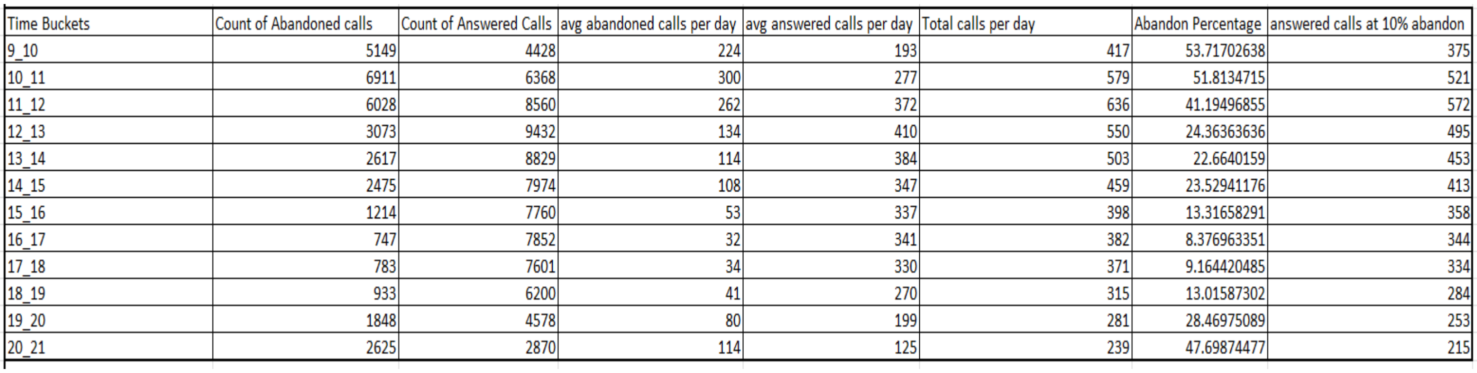


Question 3: What is the minimum number of agents required in each time bucket to reduce the abandon rate to 10%?

To find the minimum number of agents required to reduce abandonment rate to 10% first we have to find the current abandonment rate, total number of calls per day and total number of calls to be answered to reduce the rate to 10%.

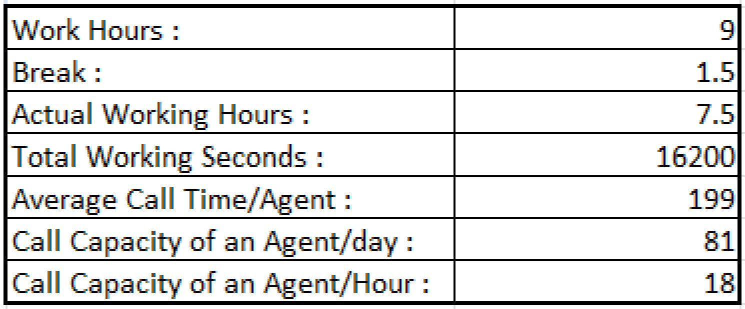
We create a table which consists of total abandoned call per time bucket, total answered call per bucket and find average calls abandoned and answered per day. From that we find the total number of calls received per day. Thus, we find an abandoned percentage. We find number of calls answered at abandon rate 10% by multiplying total number of call received by 0.9

Table below shows all values discussed above.



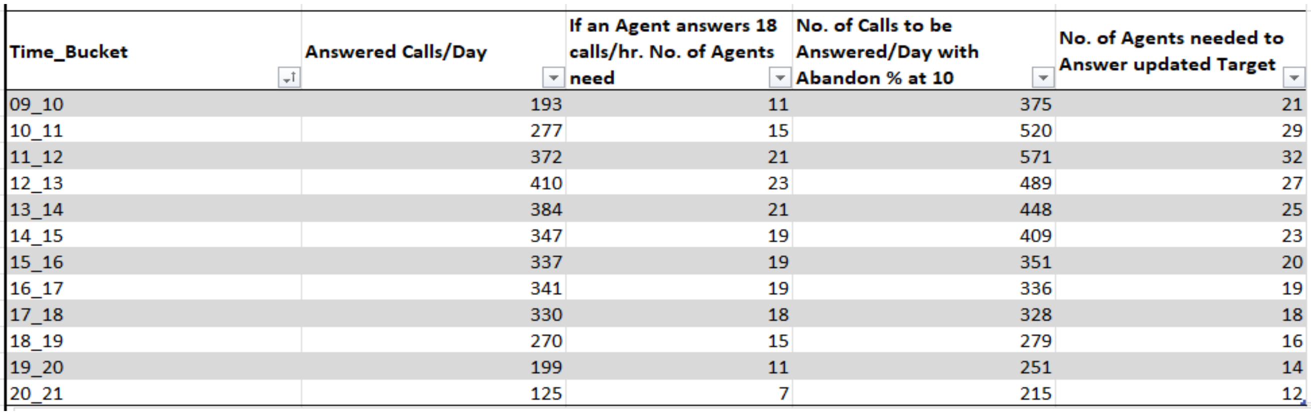
Time Buckets	Count of Abandoned calls	Count of Answered Calls	avg abandoned calls per day	avg answered calls per day	Total calls per day	Abandon Percentage	answered calls at 10% abandon
9_10	5149	4428	224	193	417	53.71702638	375
10_11	6911	6368	300	277	579	51.8134715	521
11_12	6028	8560	262	372	636	41.19496855	572
12_13	3073	9432	134	410	550	24.36363636	495
13_14	2617	8829	114	384	503	22.6640159	453
14_15	2475	7974	108	347	459	23.52941176	413
15_16	1214	7760	53	337	398	13.31658291	358
16_17	747	7852	32	341	382	8.376963351	344
17_18	783	7601	34	330	371	9.164420485	334
18_19	933	6200	41	270	315	13.01587302	284
19_20	1848	4578	80	199	281	28.46975089	253
20_21	2625	2870	114	125	239	47.69874477	215

As per the given data, an agent works for 9 hrs per day, of which 1.5 hrs is break. So effectively the agent works 7.5 hrs per day. Which means total working seconds are 16200 seconds. On average a call lasts 199 seconds. Thus, an agent can answer 81 calls per day effectively. And can answer 18 calls per hour.



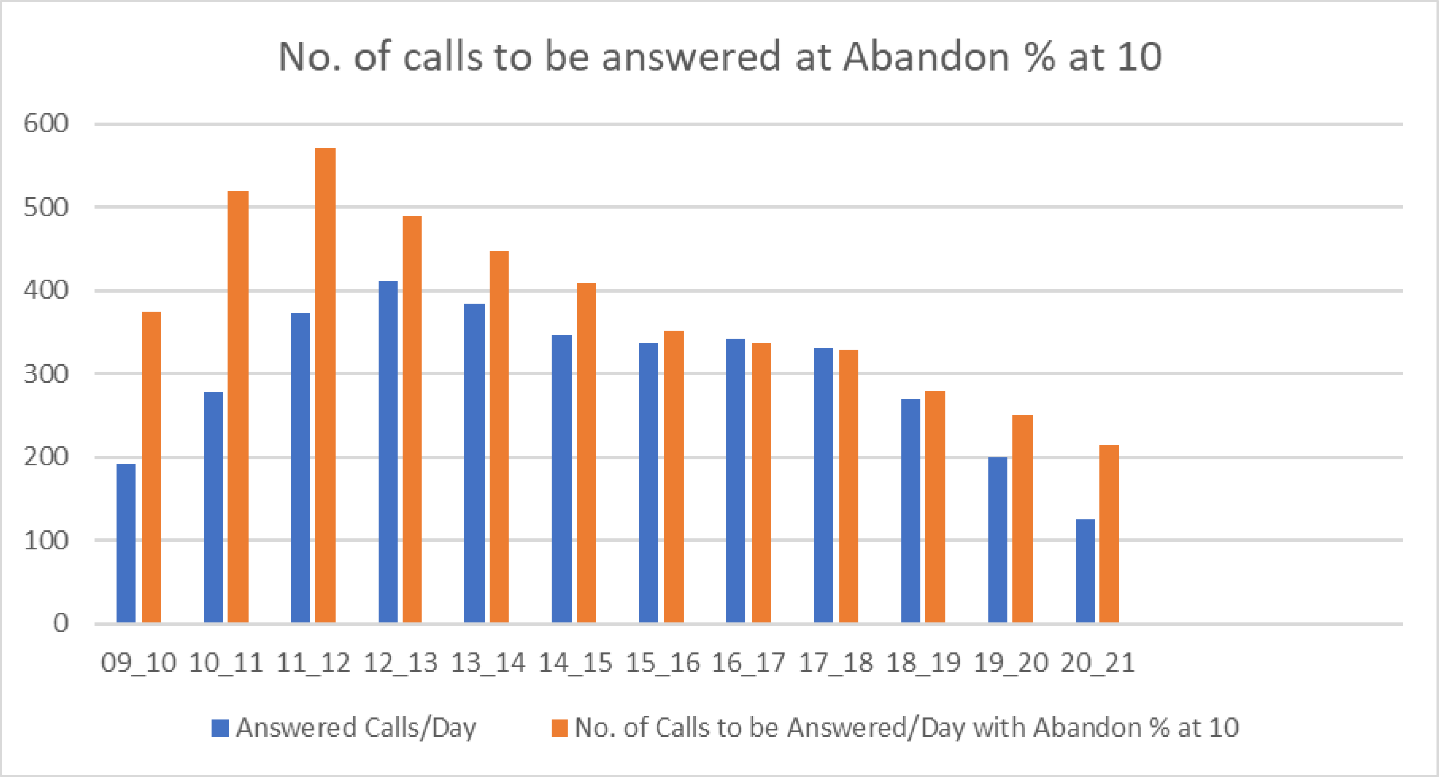
Work Hours :	9
Break :	1.5
Actual Working Hours :	7.5
Total Working Seconds :	16200
Average Call Time/Agent :	199
Call Capacity of an Agent/day :	81
Call Capacity of an Agent/Hour :	18

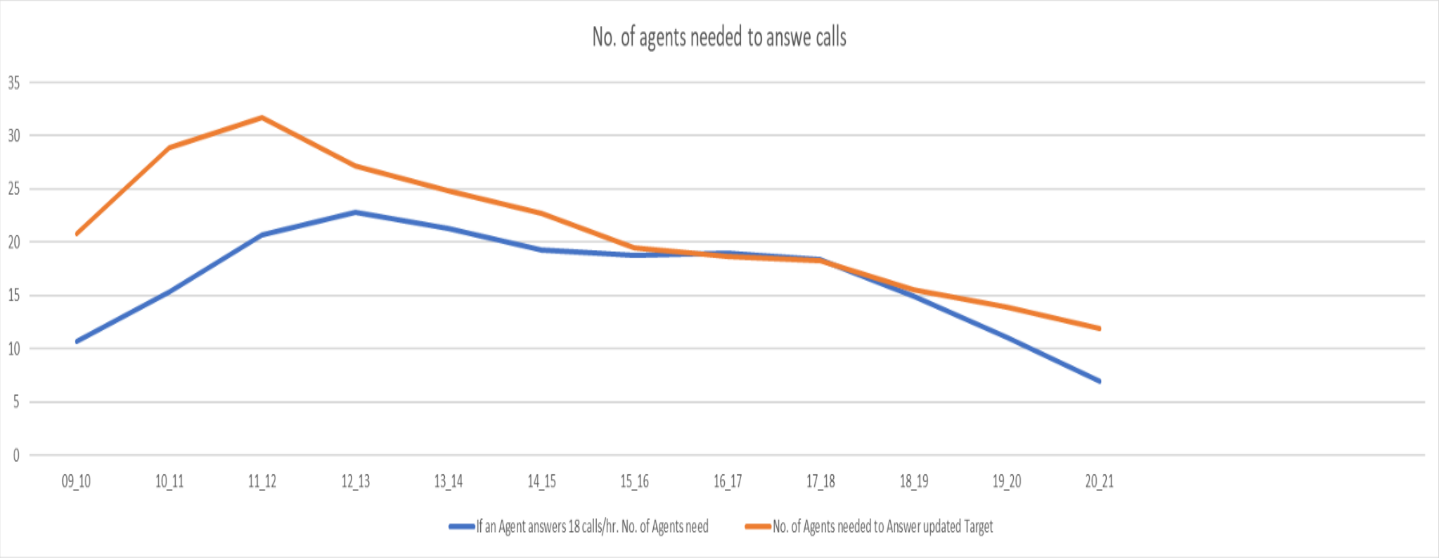
We find no. of agents by dividing no. of calls answered by 18 for each time bucket. Similarly we can find agents required to reduce abandonment rate.



Time_Bucket	Answered Calls/Day	If an Agent answers 18 calls/hr. No. of Agents need	No. of Calls to be Answered/Day with Abandon % at 10	No. of Agents needed to Answer updated Target
09_10	193	11	375	21
10_11	277	15	520	29
11_12	372	21	571	32
12_13	410	23	489	27
13_14	384	21	448	25
14_15	347	19	409	23
15_16	337	19	351	20
16_17	341	19	336	19
17_18	330	18	328	18
18_19	270	15	279	16
19_20	199	11	251	14
20_21	125	7	215	12

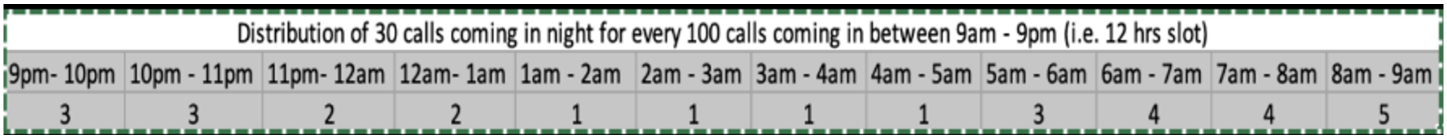
To better understand the trend we plot clustered column chart to find differences in number of agents, as well as line chart.





Question 4: Propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%.

To create a manpower plan for each time bucket throughout the day, we have to find manpower required at night. As per given information, for per 100 calls that customers make during each time bucket during day, they make 30 calls at night. And distribution of these 30 calls is as below.



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

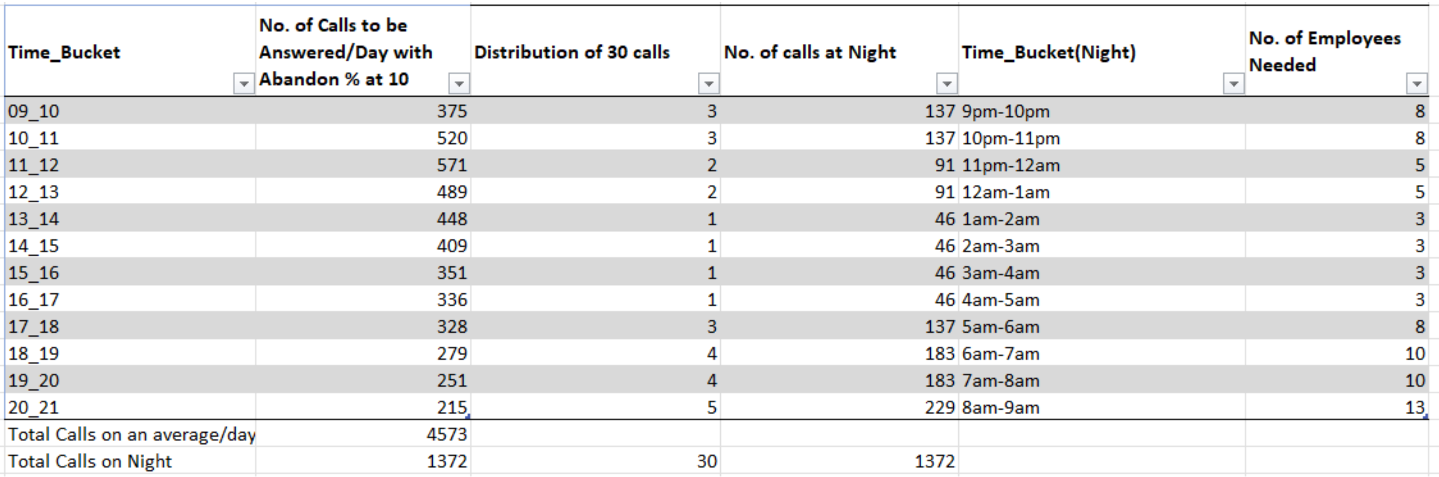
We find total number of calls made during night, by multiplying total number of calls made an average during day by 0.3

We can find distribution of calls for each time bucket during night by using formula below:

=ROUND(P$148*[@[Distribution of 30 calls]]/Q$148,0)

Here, we are multiplying total calls made at night by distribution of calls divided by 30.

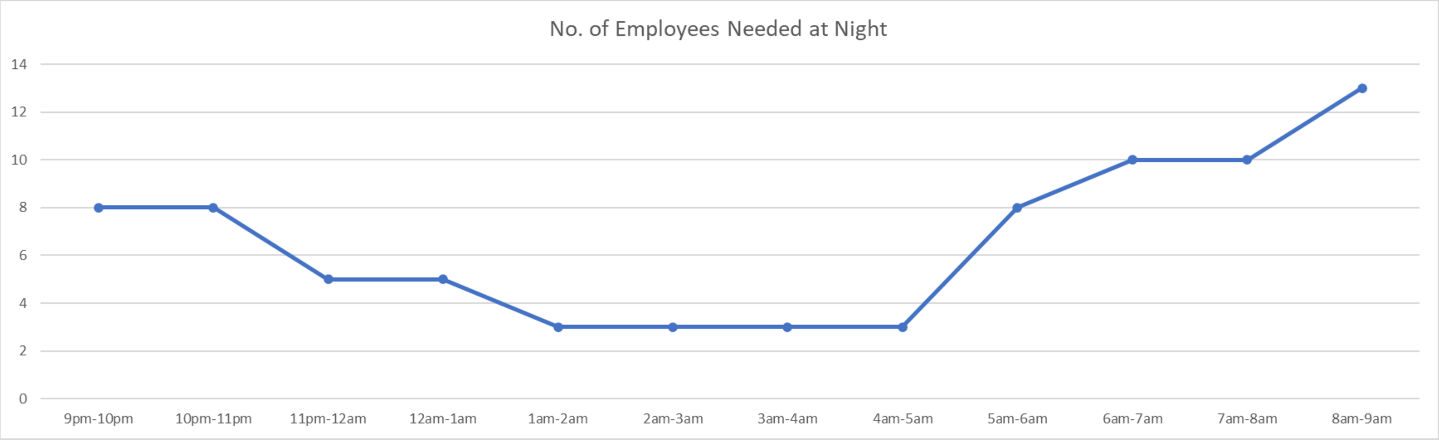
We get following results:



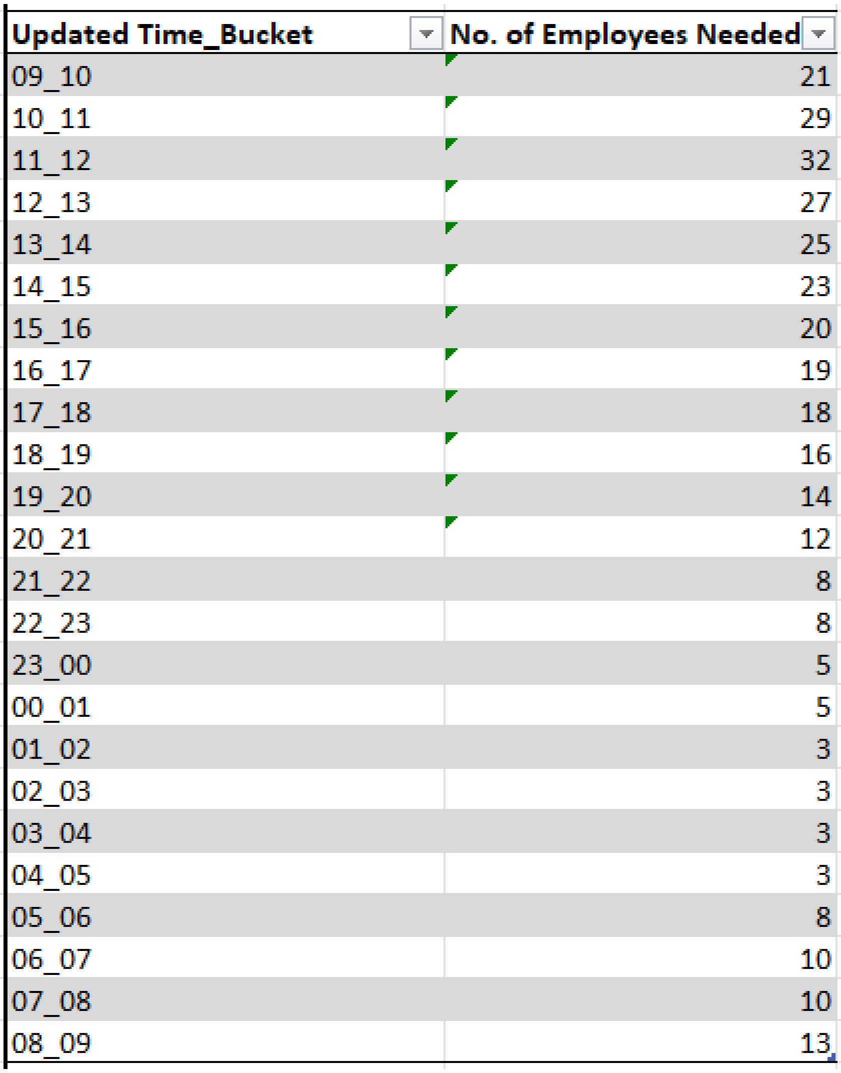
Time_Bucket	No. of Calls to be Answered/Day with Abandon % at 10	Distribution of 30 calls	No. of calls at Night	Time_Bucket(Night)	No. of Employees Needed
09_10	375	3	137	9pm-10pm	8
10_11	520	3	137	10pm-11pm	8
11_12	571	2	91	11pm-12am	5
12_13	489	2	91	12am-1am	5
13_14	448	1	46	1am-2am	3
14_15	409	1	46	2am-3am	3
15_16	351	1	46	3am-4am	3
16_17	336	1	46	4am-5am	3
17_18	328	3	137	5am-6am	8
18_19	279	4	183	6am-7am	10
19_20	251	4	183	7am-8am	10
20_21	215	5	229	8am-9am	13
Total Calls on an average/day	4573				
Total Calls on Night	1372	30	1372		

To find the number of employees, we divided the total number of calls by 18(number of calls an employee can answer per hour).

We plot this using a line chart to better visualize the data.

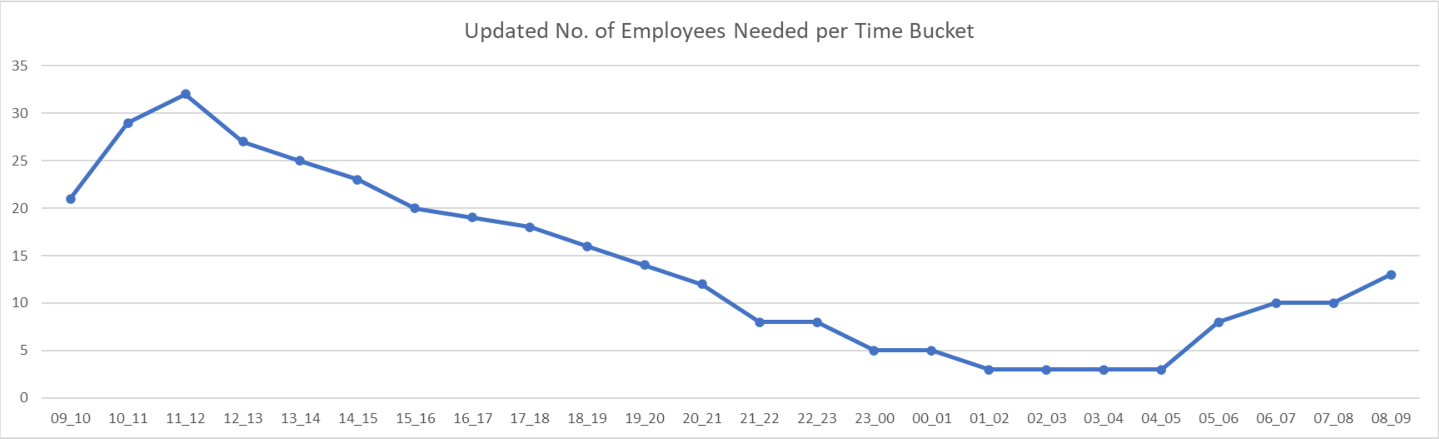


We also combine this data with no. of employees required for abandoned rate at 10% received from previous question and create a combined manpower table.



Updated Time_Bucket	No. of Employees Needed
09_10	21
10_11	29
11_12	32
12_13	27
13_14	25
14_15	23
15_16	20
16_17	19
17_18	18
18_19	16
19_20	14
20_21	12
21_22	8
22_23	8
23_00	5
00_01	5
01_02	3
02_03	3
03_04	3
04_05	3
05_06	8
06_07	10
07_08	10
08_09	13

We create a line Chart of combined Data to visualize manpower distribution for a complete day.



Results:

While working on this project, I have gained a better understanding of Call Volume Trends. I have improved my understanding of Advanced Excel methodologies. By analyzing Customer Experience Call Data, I was able to provide insights on various aspects such as average calls made throughout the day for each time bucket, total number of calls per time bucket, how to reduce abandon rate by increasing manpower, and how manpower distribution would look like for night shift. I was also able to create different visualizations to improve data understanding.

This project has helped me enhance my Excel skills, particularly in data visualization and creating pivot tables and charts to derive meaningful insights. It has also improved my ability to interpret data and provide actionable recommendations based on the analysis.

Appendix

LINK TO PROJECT REPORTS

1. DATA ANALYTICS PROCESS
https://drive.google.com/file/d/1YHjZQ2n5kCDr5dqxEtenZF0_YsOgLBeG/view?usp=drive_link

2. INSTAGRAM USER ANALYTICS
https://drive.google.com/file/d/1A-LkZFnR9MNt4lEMzT_lY7xtBLHpCIj-/view?usp=drive_link

3. OPERATION ANALYTICS AND INVESTIGATING METRIC SPIKE
https://drive.google.com/file/d/18xLe2WK0T3aZ_19d24rycsopw9t7nwIH/view?usp=drive_link

4. HIRING PROCESS ANALYTICS
https://drive.google.com/file/d/1kbWxPJmIbiU-AtWxFPetBeBPbtzQjTq-/view?usp=drive_link

5. IMDB MOVIE ANALYSIS
https://drive.google.com/file/d/1uqiD-G8iaM_eEpFZuxBCMyYPfUj1lEtI/view?usp=drive_link

6. BANK LOAN CASE STUDY
https://drive.google.com/file/d/13Ou0ZeU6McaOub29tOcV4BrpZNErvKQs/view?usp=drive_link

7. ANALYZING THE IMPACT OF CAR FEATURES ON PRICE AND PROFITABILITY
https://drive.google.com/file/d/1dVQLHpXHKUxFGdx0UFZlwDeGrgtjex38/view?usp=drive_link

8. ABC CALL VOLUME TREND ANALYSIS
https://drive.google.com/file/d/1PJZGza8caECbIOGXIulkKkla9bJ13NyB/view?usp=drive_link