DATA ANALYSIS PORTFOLIO

CREATED BY :
ARIN KUMAR

TABLE OF CONTENTS

Professional Background

Currently in my 2nd year persuing BE-CSE in Chandigarh University, I have secured 8.05 CGPA(till 3rd sem) and have several skills in Data Analysis, Machine Learning, Python, R, SQL and Excel.

I have worked in several projects in machine learning and data analysis in Python, R, SQL and Excel.

As I am a fresher it would be great to experience the real challenges of the corporate world and understand how things work. Being a fresher, I think I am very flexible and adaptive to learn new things. I have theoretical knowledge. But I am waiting to use mytheoretical knowledge in a practical way. And I believe by putting significant efforts I will learn.

INSTAGRAM USER ANALYTICS

Project Description:

This project is about providing insights on user engagement and behavior on Instagram for the marketing and investor teams. The project will involve analyzing data from a provided database using SQL commands to answer specific questions related to user loyalty, inactive users, contest winners, hashtags, and ad campaign scheduling.

Approach:

I will begin by creating the database and running SQL commands to extract the necessary data. I will then analyze the data to answer the questions posed by the marketing and investor teams, and present my findings in a report.

Tech-Stack Used:

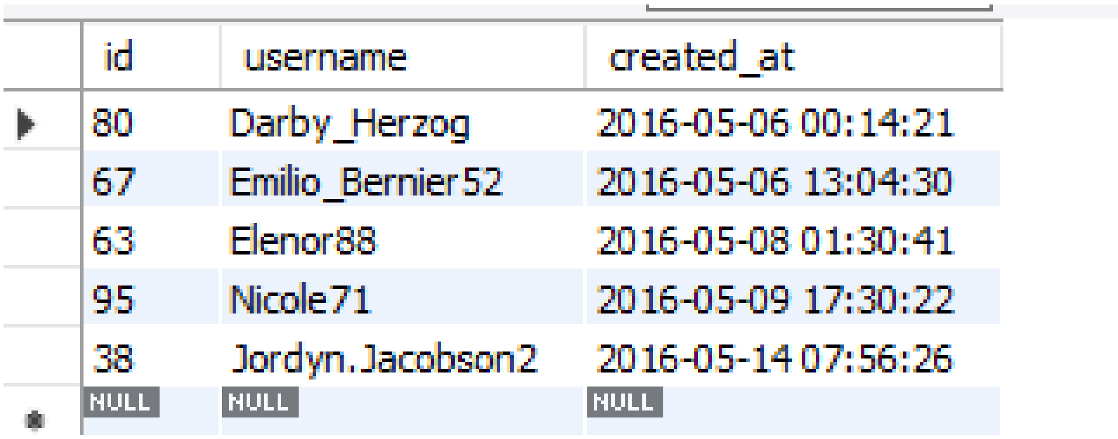
I will be using SQL and a relational database management system to perform the analysis and extract the data needed to answer the questions. The version of SQL and RDBMS used will depend on the specific database provided.

Insights:

Through this project, I will gain a deeper understanding of user behavior and engagement on Instagram, and be able to provide valuable insights to the marketing and investor teams. These insights can be used to improve the overall user experience and drive growth for the platform.

A) Marketing: The marketing team wants to launch some campaigns, and they need your help with the following

1. **Rewarding Most Loyal Users:** People who have been using the platform for the longest time.

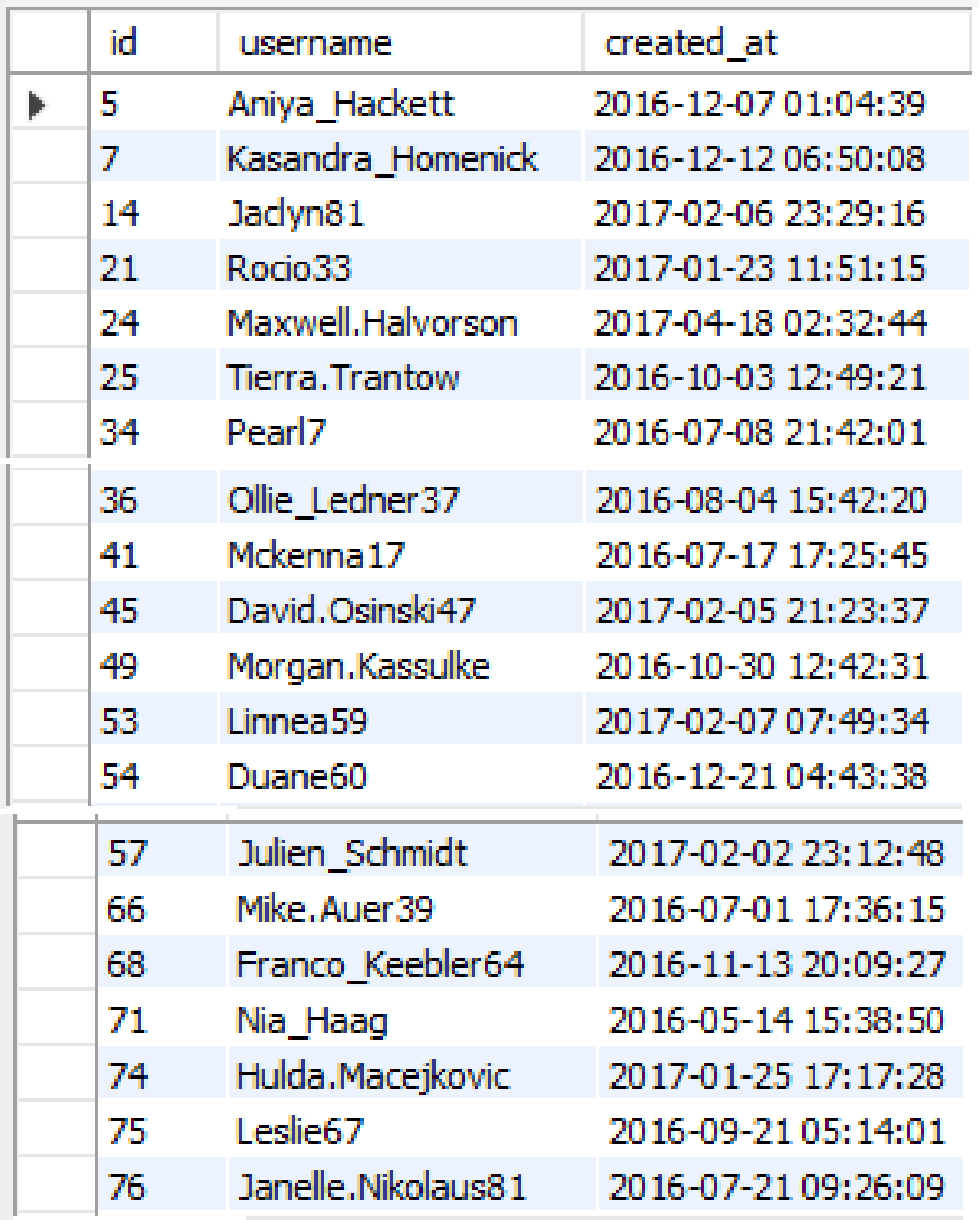


id	username	created_at
80	Darby_Herzog	2016-05-06 00:14:21
67	Emilio_Bernier52	2016-05-06 13:04:30
63	Elenor88	2016-05-08 01:30:41
95	Nicole71	2016-05-09 17:30:22
38	Jordyn.Jacobson2	2016-05-14 07:56:26
NULL	NULL	NULL

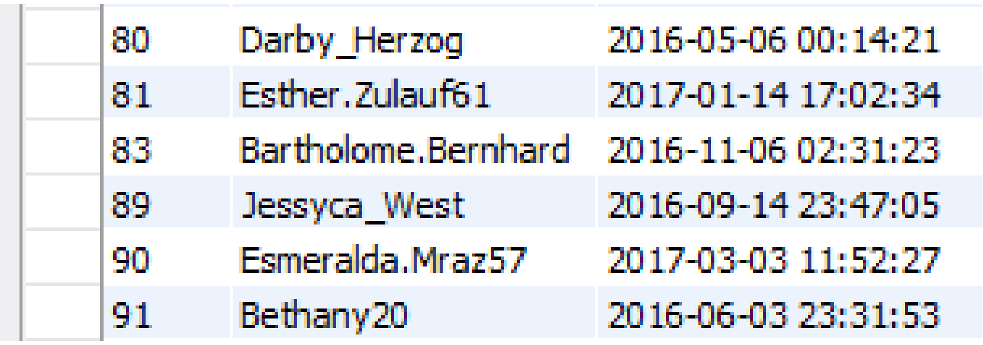
Darby_Herzog is the oldest user of Instagram who has created the account at 06-05-2016 at 12:14 am.

2. **Remind Inactive Users to Start Posting:** By sending them promotional emails to post their 1st photo.

26 users have created their accounts but has not posted yet on Instagram.

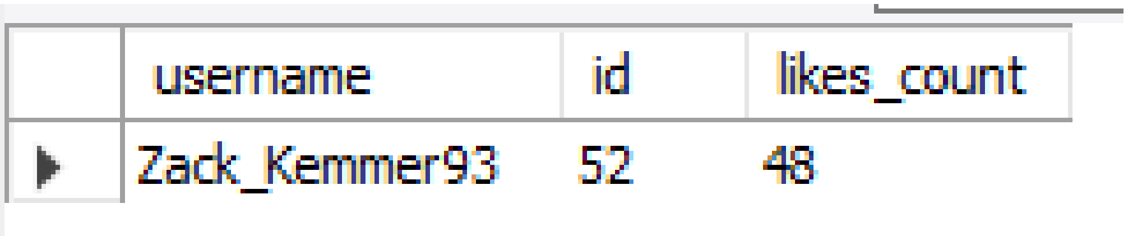


	id	username	created_at
▶	5	Aniya_Hackett	2016-12-07 01:04:39
	7	Kasandra_Homenick	2016-12-12 06:50:08
	14	Jaclyn81	2017-02-06 23:29:16
	21	Rocio33	2017-01-23 11:51:15
	24	Maxwell.Halvorson	2017-04-18 02:32:44
	25	Tierra.Trantow	2016-10-03 12:49:21
	34	Pearl7	2016-07-08 21:42:01
	36	Ollie_Ledner37	2016-08-04 15:42:20
	41	Mckenna17	2016-07-17 17:25:45
	45	David.Osinski47	2017-02-05 21:23:37
	49	Morgan.Kassulke	2016-10-30 12:42:31
	53	Linnea59	2017-02-07 07:49:34
	54	Duane60	2016-12-21 04:43:38
	57	Julien_Schmidt	2017-02-02 23:12:48
	66	Mike.Auer39	2016-07-01 17:36:15
	68	Franco_Keebler64	2016-11-13 20:09:27
	71	Nia_Haag	2016-05-14 15:38:50
	74	Hulda.Macejkovic	2017-01-25 17:17:28
	75	Leslie67	2016-09-21 05:14:01
	76	Janelle.Nikolaus81	2016-07-21 09:26:09



80	Darby_Herzog	2016-05-06 00:14:21
81	Esther.Zulauf61	2017-01-14 17:02:34
83	Bartholome.Bernhard	2016-11-06 02:31:23
89	Jessyca_West	2016-09-14 23:47:05
90	Esmeralda.Mraz57	2017-03-03 11:52:27
91	Bethany20	2016-06-03 23:31:53

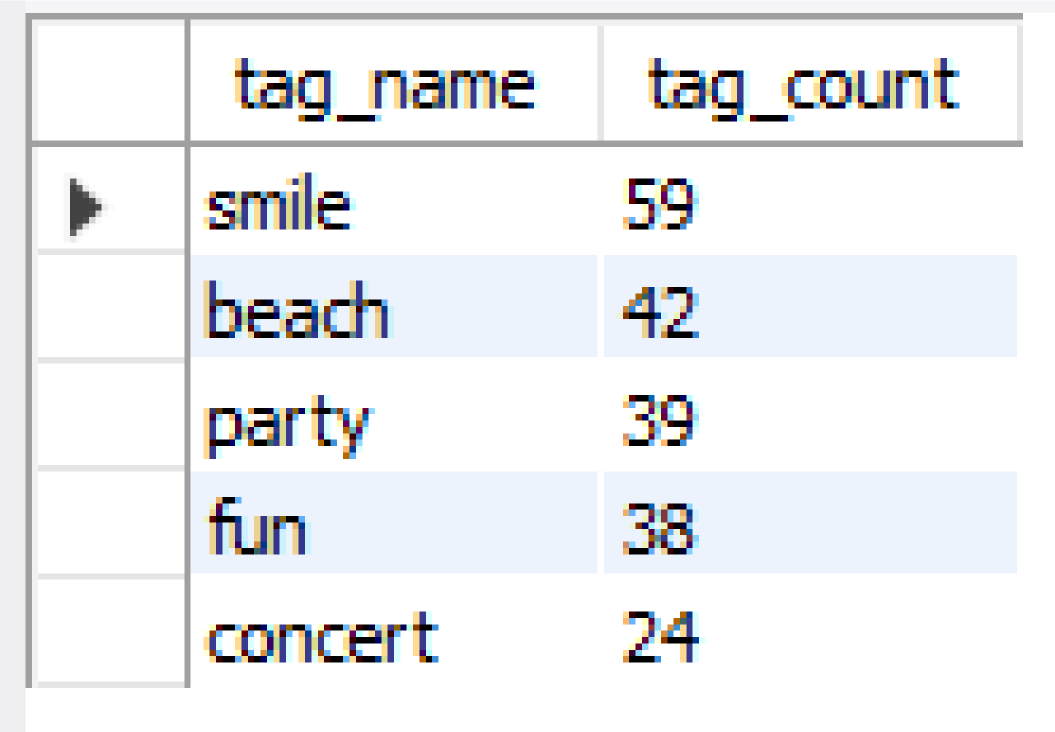
3. **Declaring Contest Winner:** The team started a contest and the user who gets the most likes on a single photo will win the contest now they wish to declare the winner.



username	id	likes_count
Zack_Kemmer93	52	48

The username Zack_Kemmer93 with id 52 is the contest winner with 48 likes.

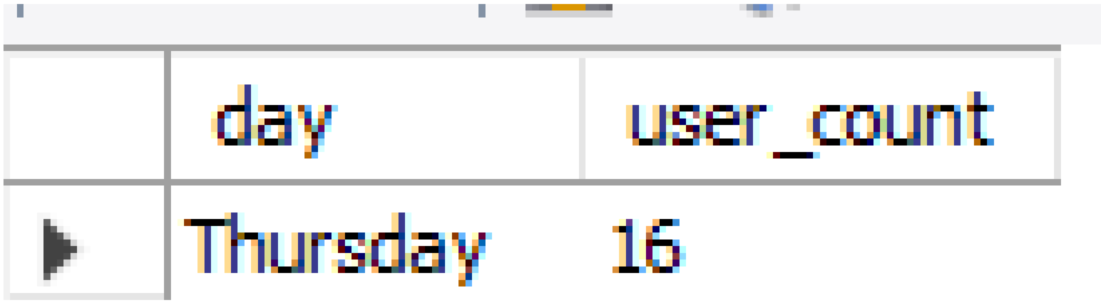
4. **Hashtag Researching:** A partner brand wants to know, which hashtags to use in the post to reach the most people on the platform.



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

The tag_name with smile has the maximum tag counts with 59 tags.

5. **Launch AD Campaign:** The team wants to know, which day would be the best day to launch ADs.

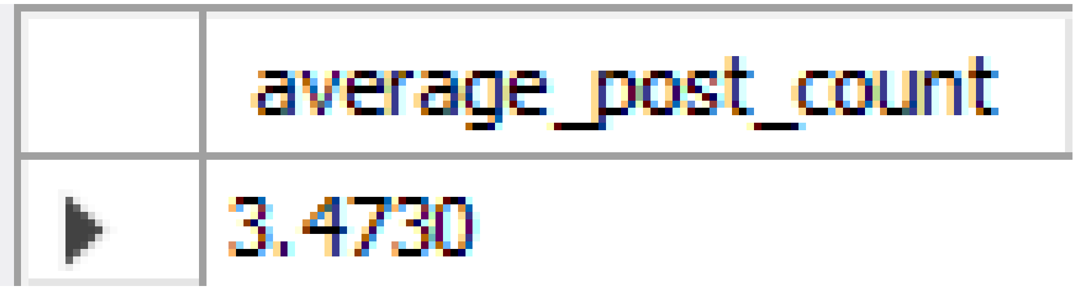


day	user_count
Thursday	16

Thursday is the best day to launch ADS as most users register on this day.

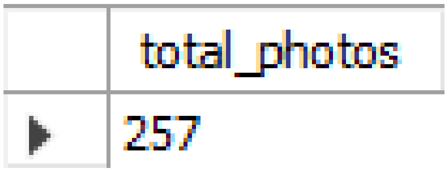
B) Investor Metrics: Our investors want to know if Instagram is performing well and is not becoming redundant like Facebook, they want to assess the app on the following grounds

1. **User Engagement:** Are users still as active and post on Instagram or they are making fewer posts

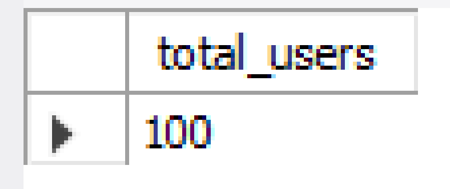


	average_post_count
▶	3.4730

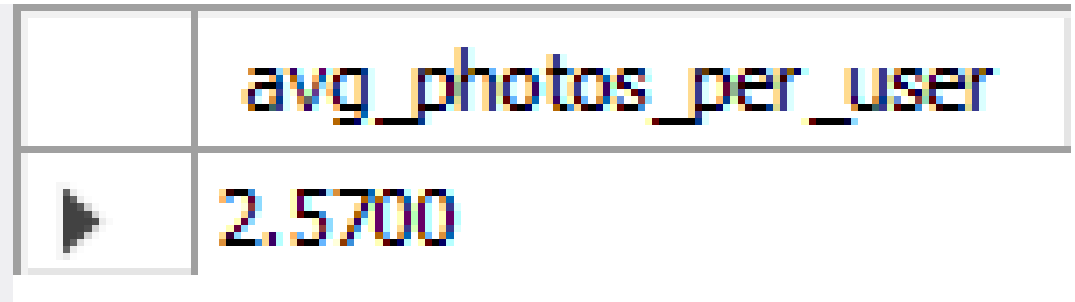
A user on an average posts 3.473 posts on Instagram.



	total_photos
▶	257



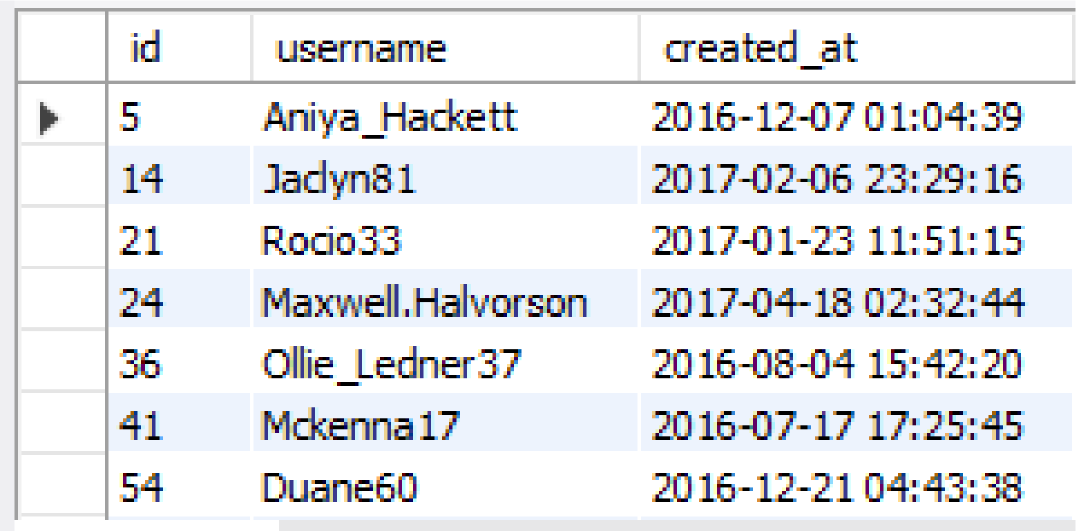
	total_users
▶	100

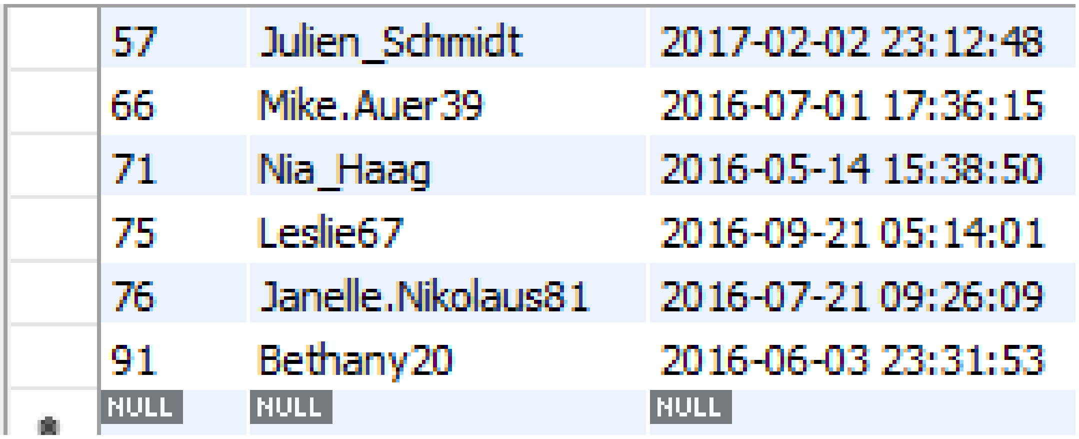


	avg_photos_per_user
▶	2.5700

The average photos per user is 2.57

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





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

There are a total of 13 fake and dummy accounts.

Result:

By completing this project, I will have provided detailed insights on user engagement and behavior on Instagram for the marketing and investor teams. These insights can be used to make informed decisions related to product development, marketing campaigns, and overall performance of the platform.

OPERATION ANALYTICS AND INVESTIGATING METRIC SPIKE

Project Description:

This project is about analyzing two datasets, job_data and users/events/email_events. The job_data dataset includes information about the jobs reviewed, including the unique identifier of the job, actor, event, language, time spent, organization, and date. The users/events/email_events dataset includes information about user activity, such as logins, messaging events, search events, and email events. The goal of this project is to answer various questions related to job review and user engagement using SQL.

Approach:

The approach taken in this project is to first create a database and tables based on the given structure and links. Then, use SQL to perform the analysis and answer the questions related to job review and user engagement.

Tech Stack Used:

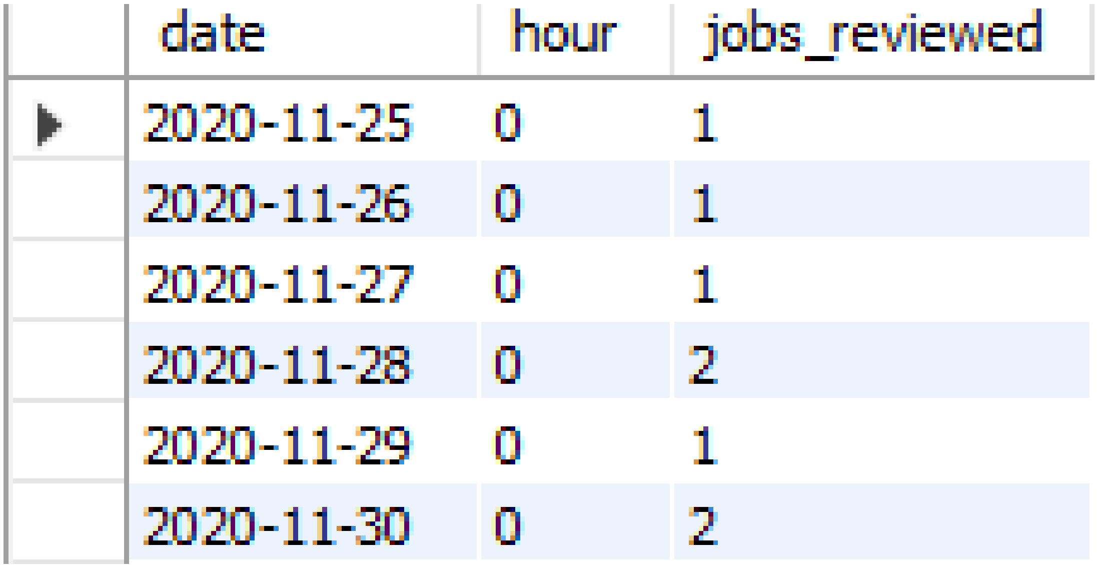
The tech stack used in this project is SQL. The specific version of SQL used will depend on the database management system used.

Insights:

The insights gained from this project are related to the job review process and user engagement with a product. The results of the analysis provide information about the number of jobs reviewed per hour per day, the average throughput, the percentage share of each language, and the weekly user engagement, growth, retention, and email engagement.

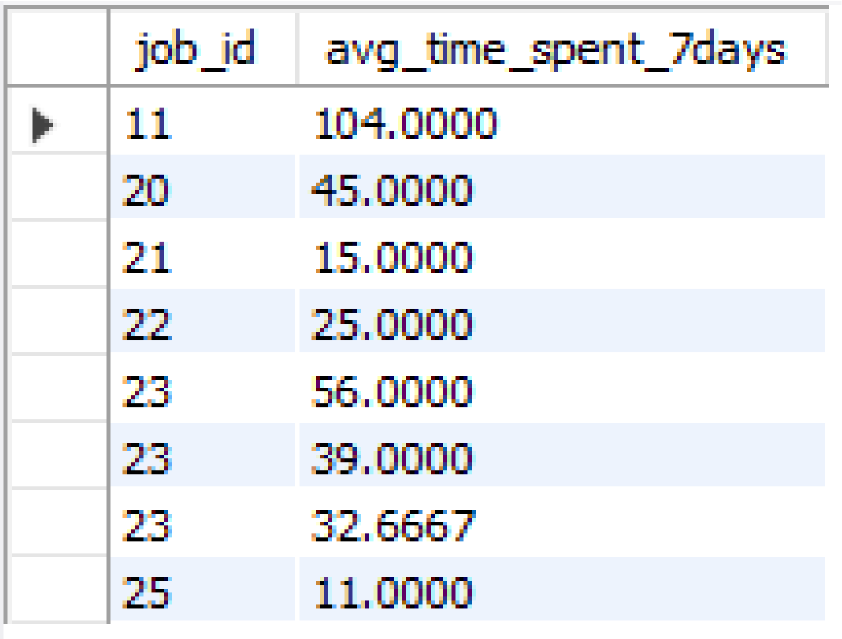
CASE STUDY – 1

A. **Number of jobs reviewed:** Amount of jobs reviewed over time.
   Your task: Calculate the number of jobs reviewed per hour per day for November 2020?



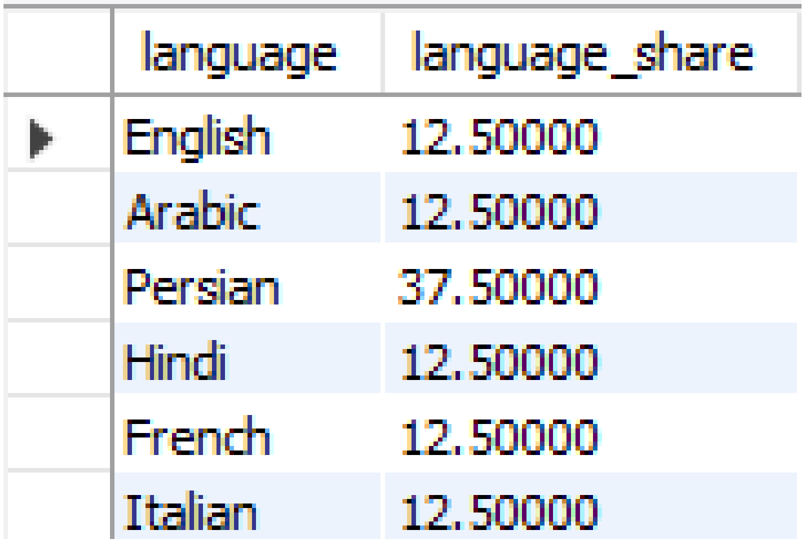
date	hour	jobs_reviewed
2020-11-25	0	1
2020-11-26	0	1
2020-11-27	0	1
2020-11-28	0	2
2020-11-29	0	1
2020-11-30	0	2

B. Throughput: It is the no. of events happening per second.
Your task: Let’s say the above metric is called throughput. Calculate 7 day rolling average of throughput? For throughput, do you prefer daily metric or 7-day rolling and why?



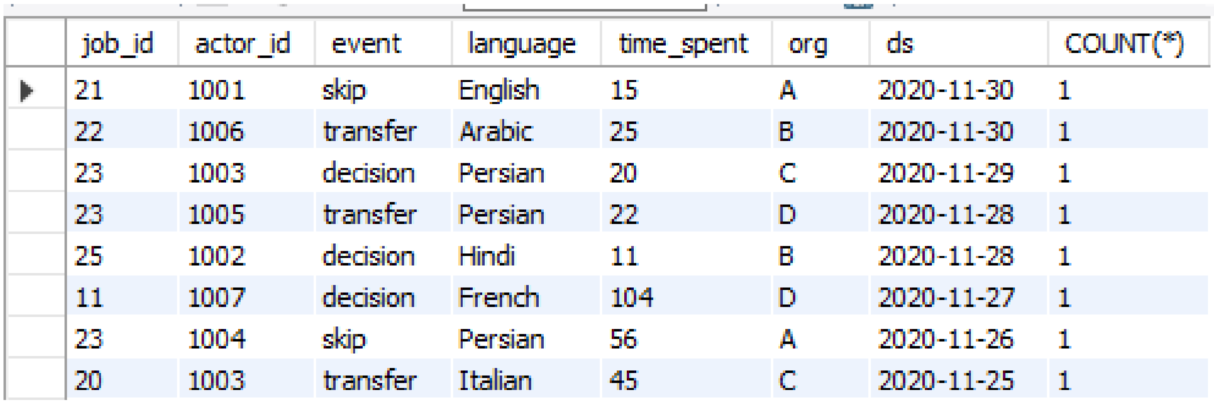
	job_id	avg_time_spent_7days
▶	11	104.0000
	20	45.0000
	21	15.0000
	22	25.0000
	23	56.0000
	23	39.0000
	23	32.6667
	25	11.0000

C. **Percentage share of each language:** Share of each language for differentcontents.
Your task: Calculate the percentage share of each language in the last 30 days?



language	language_share
English	12.50000
Arabic	12.50000
Persian	37.50000
Hindi	12.50000
French	12.50000
Italian	12.50000

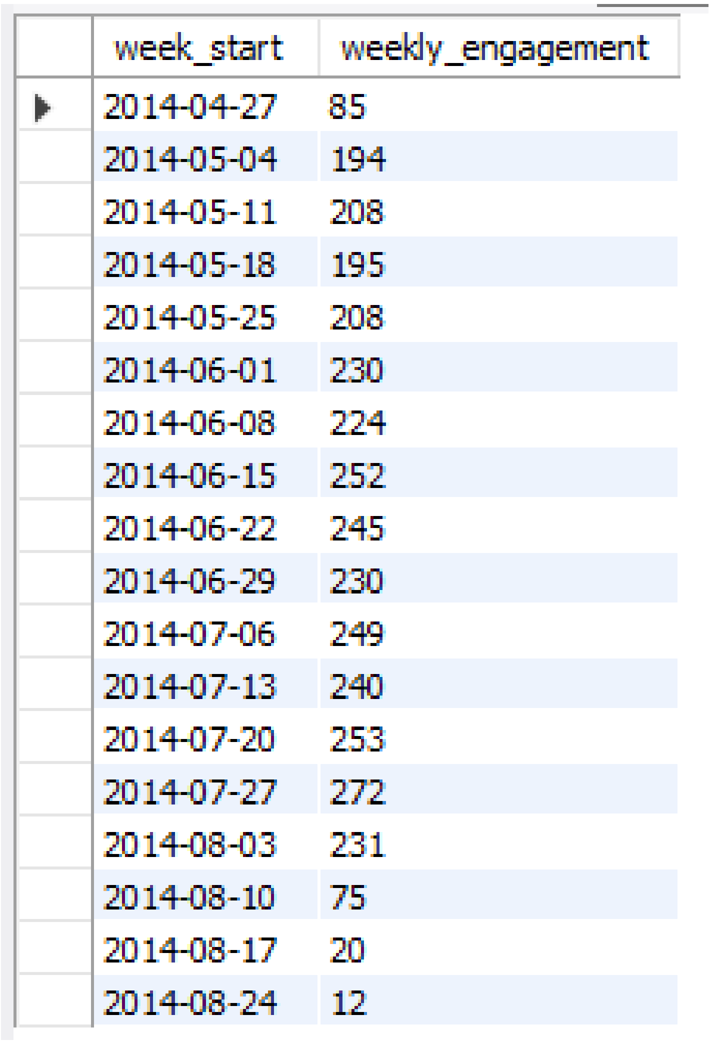
D. **Duplicate rows:** Rows that have the same value present in them.
Your task: Let's say you see some duplicate rows in the data. How will you display duplicates from the table?



job_id	actor_id	event	language	time_spent	org	ds	COUNT(*)
21	1001	skip	English	15	A	2020-11-30	1
22	1006	transfer	Arabic	25	B	2020-11-30	1
23	1003	decision	Persian	20	C	2020-11-29	1
23	1005	transfer	Persian	22	D	2020-11-28	1
25	1002	decision	Hindi	11	B	2020-11-28	1
11	1007	decision	French	104	D	2020-11-27	1
23	1004	skip	Persian	56	A	2020-11-26	1
20	1003	transfer	Italian	45	C	2020-11-25	1

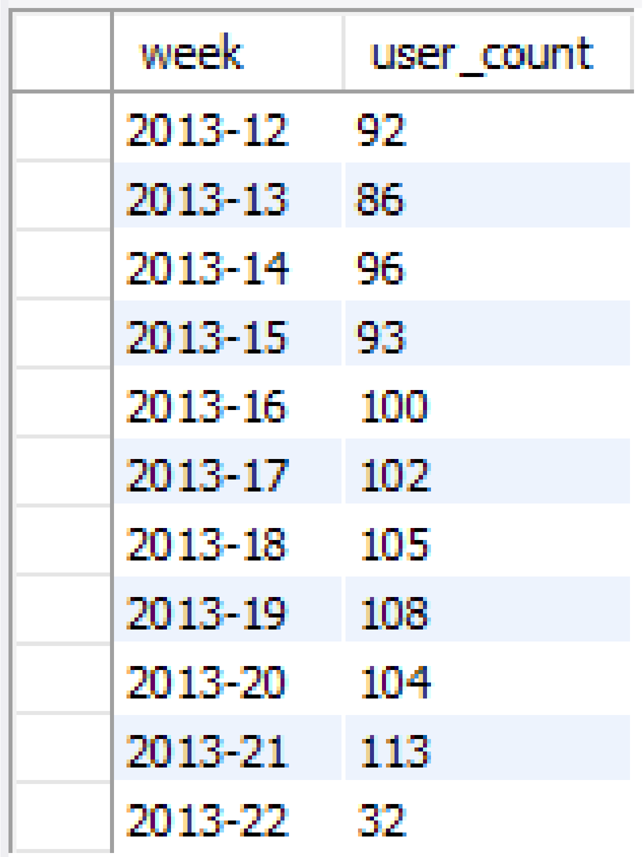
CASE STUDY 2

A. **User Engagement:** To measure the activeness of a user. Measuring if the user finds quality in a product/service.
   Your task: Calculate the weekly user engagement?



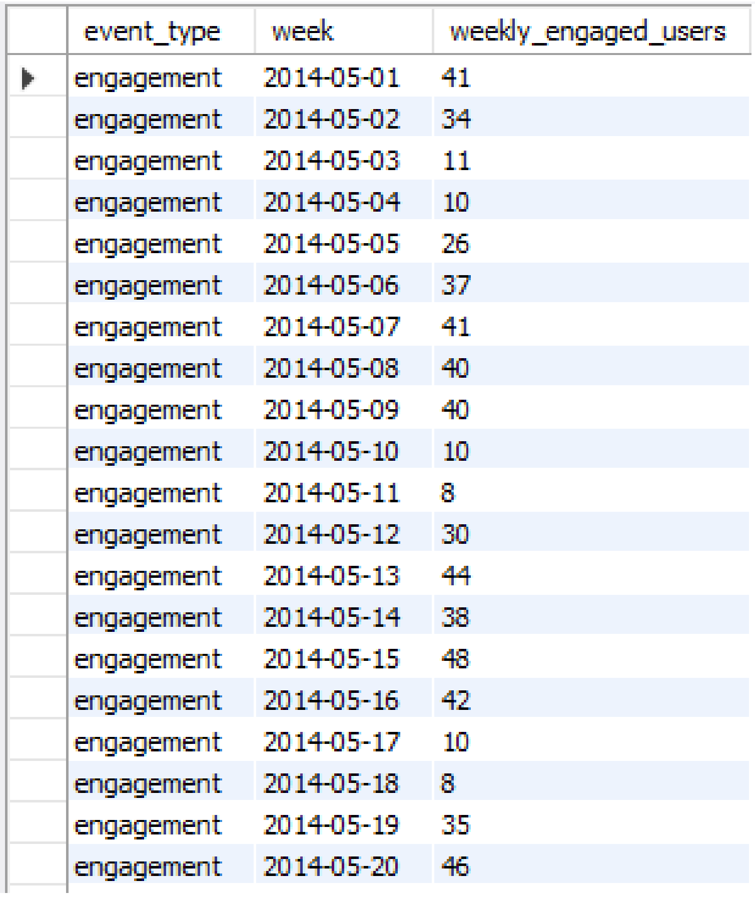
week_start	weekly_engagement
2014-04-27	85
2014-05-04	194
2014-05-11	208
2014-05-18	195
2014-05-25	208
2014-06-01	230
2014-06-08	224
2014-06-15	252
2014-06-22	245
2014-06-29	230
2014-07-06	249
2014-07-13	240
2014-07-20	253
2014-07-27	272
2014-08-03	231
2014-08-10	75
2014-08-17	20
2014-08-24	12

B. **User Growth:** Amount of users growing over time for a product.
Your task: Calculate the user growth for product?

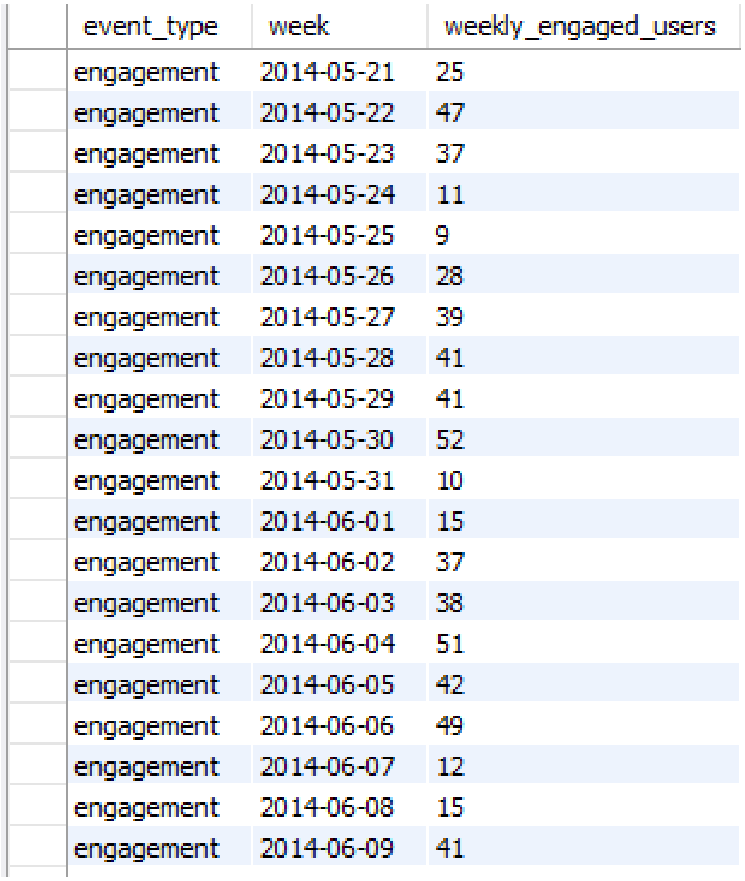


	week	user_count
	2013-12	92
	2013-13	86
	2013-14	96
	2013-15	93
	2013-16	100
	2013-17	102
	2013-18	105
	2013-19	108
	2013-20	104
	2013-21	113
	2013-22	32

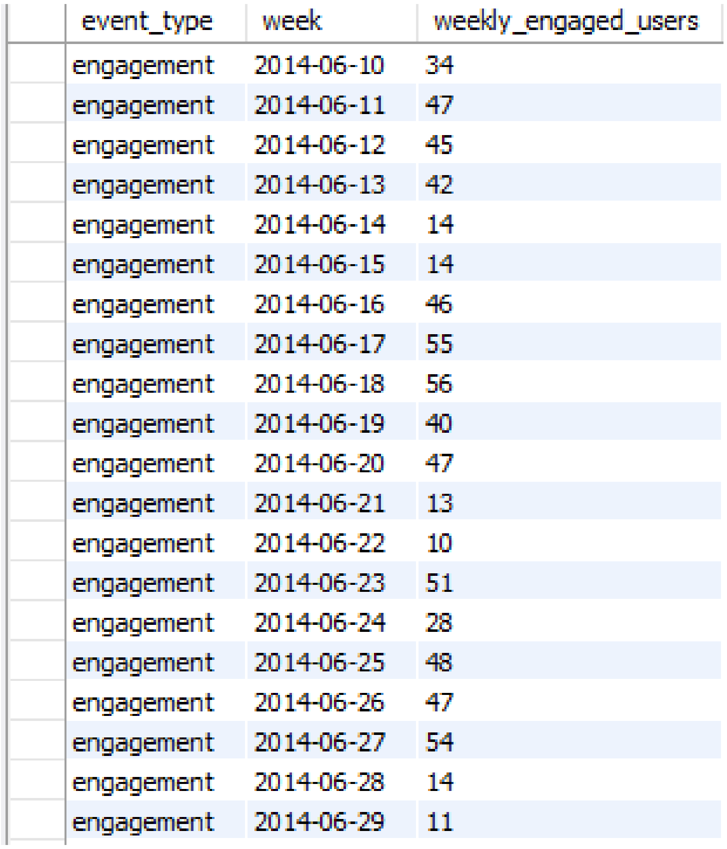
C. **Weekly Engagement:** To measure the activeness of a user. Measuring if the user finds quality in a product/service weekly.
Your task: Calculate the weekly engagement per device?



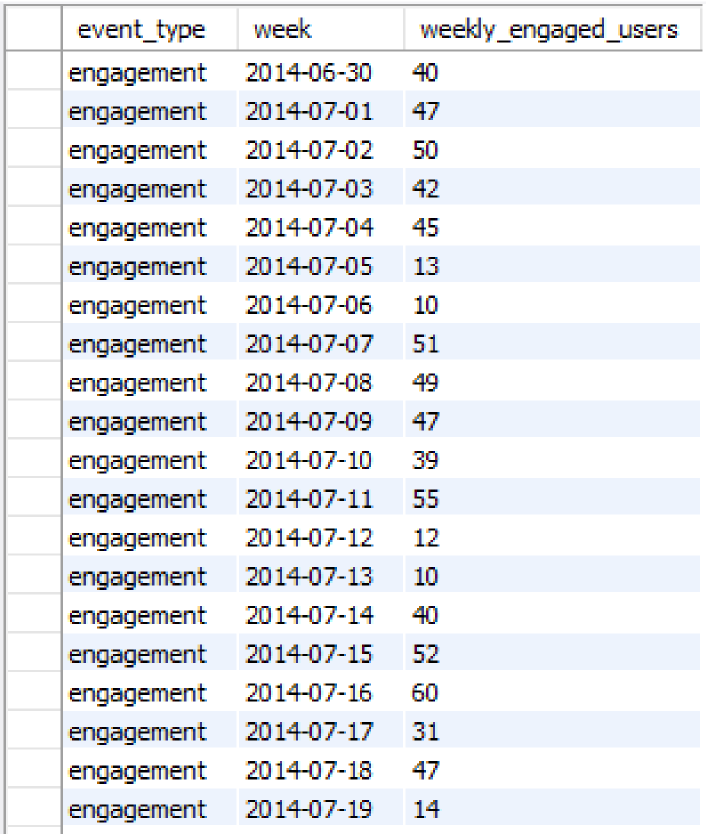
event_type	week	weekly_engaged_users
engagement	2014-05-01	41
engagement	2014-05-02	34
engagement	2014-05-03	11
engagement	2014-05-04	10
engagement	2014-05-05	26
engagement	2014-05-06	37
engagement	2014-05-07	41
engagement	2014-05-08	40
engagement	2014-05-09	40
engagement	2014-05-10	10
engagement	2014-05-11	8
engagement	2014-05-12	30
engagement	2014-05-13	44
engagement	2014-05-14	38
engagement	2014-05-15	48
engagement	2014-05-16	42
engagement	2014-05-17	10
engagement	2014-05-18	8
engagement	2014-05-19	35
engagement	2014-05-20	46



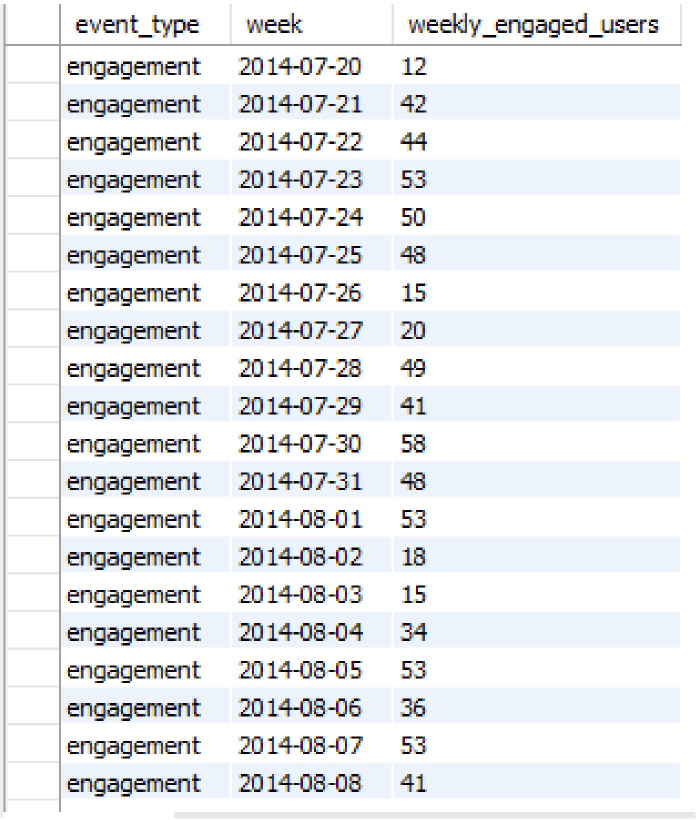
event_type	week	weekly_engaged_users
engagement	2014-05-21	25
engagement	2014-05-22	47
engagement	2014-05-23	37
engagement	2014-05-24	11
engagement	2014-05-25	9
engagement	2014-05-26	28
engagement	2014-05-27	39
engagement	2014-05-28	41
engagement	2014-05-29	41
engagement	2014-05-30	52
engagement	2014-05-31	10
engagement	2014-06-01	15
engagement	2014-06-02	37
engagement	2014-06-03	38
engagement	2014-06-04	51
engagement	2014-06-05	42
engagement	2014-06-06	49
engagement	2014-06-07	12
engagement	2014-06-08	15
engagement	2014-06-09	41



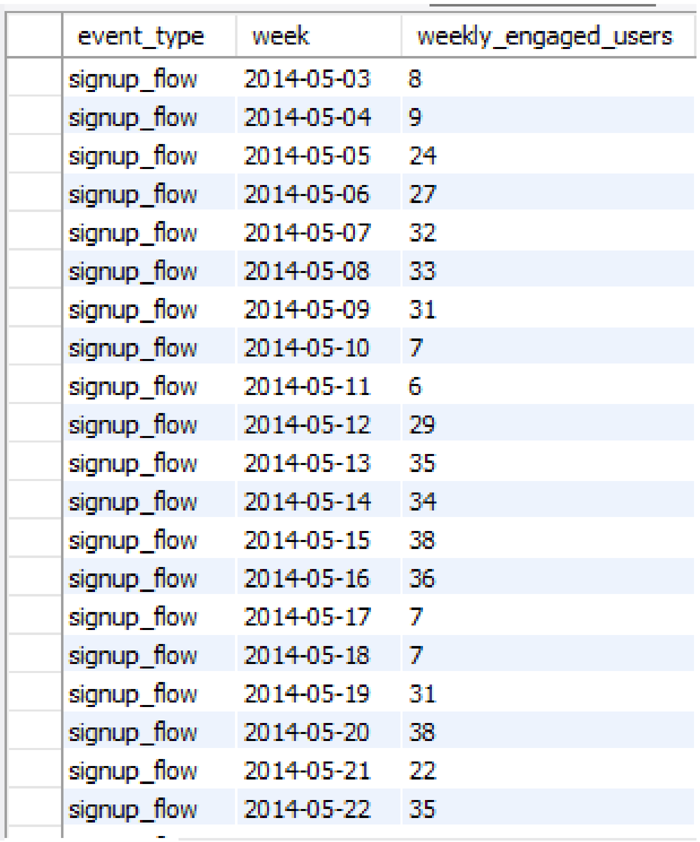
event_type	week	weekly_engaged_users
engagement	2014-06-10	34
engagement	2014-06-11	47
engagement	2014-06-12	45
engagement	2014-06-13	42
engagement	2014-06-14	14
engagement	2014-06-15	14
engagement	2014-06-16	46
engagement	2014-06-17	55
engagement	2014-06-18	56
engagement	2014-06-19	40
engagement	2014-06-20	47
engagement	2014-06-21	13
engagement	2014-06-22	10
engagement	2014-06-23	51
engagement	2014-06-24	28
engagement	2014-06-25	48
engagement	2014-06-26	47
engagement	2014-06-27	54
engagement	2014-06-28	14
engagement	2014-06-29	11



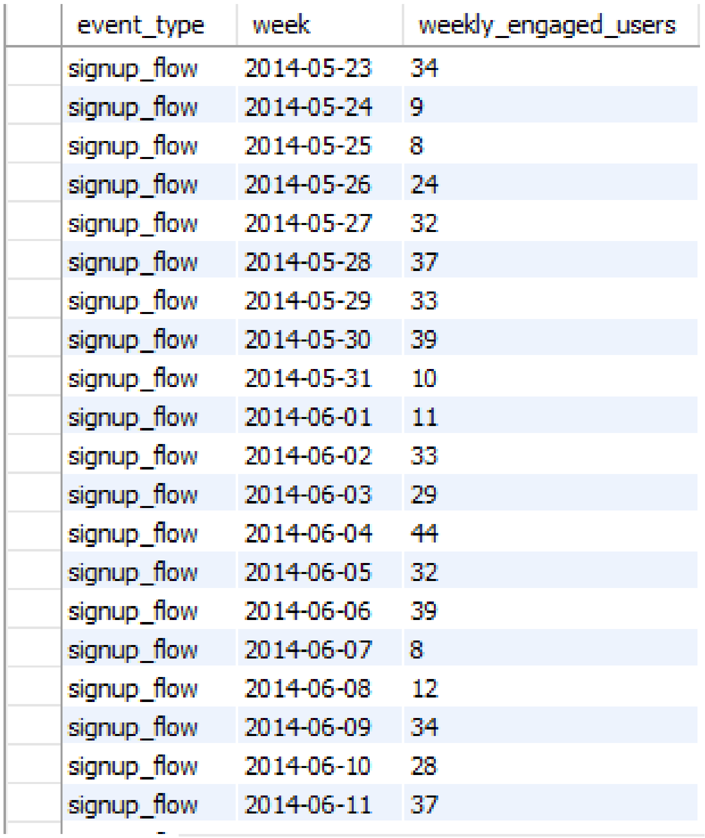
event_type	week	weekly_engaged_users
engagement	2014-06-30	40
engagement	2014-07-01	47
engagement	2014-07-02	50
engagement	2014-07-03	42
engagement	2014-07-04	45
engagement	2014-07-05	13
engagement	2014-07-06	10
engagement	2014-07-07	51
engagement	2014-07-08	49
engagement	2014-07-09	47
engagement	2014-07-10	39
engagement	2014-07-11	55
engagement	2014-07-12	12
engagement	2014-07-13	10
engagement	2014-07-14	40
engagement	2014-07-15	52
engagement	2014-07-16	60
engagement	2014-07-17	31
engagement	2014-07-18	47
engagement	2014-07-19	14



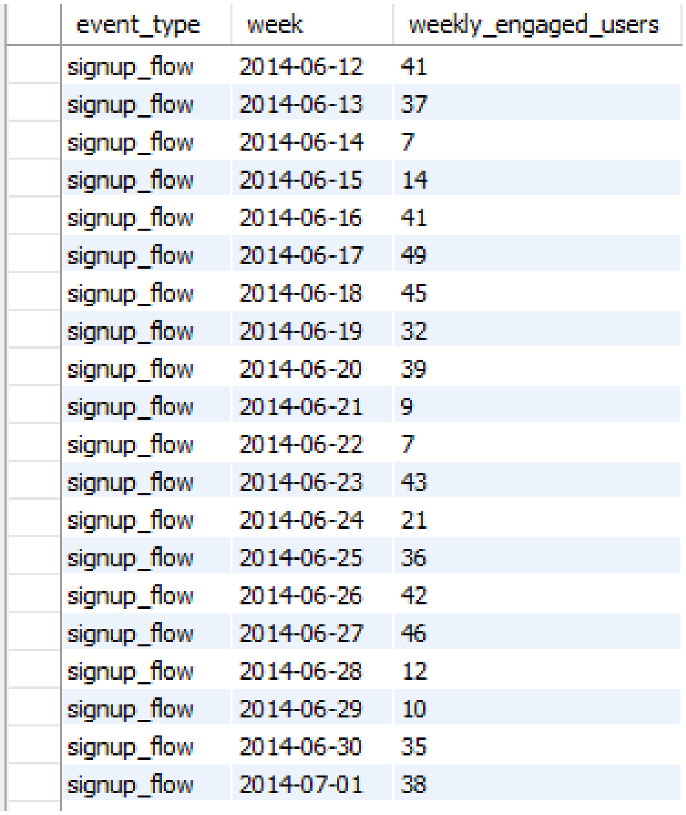
event_type	week	weekly_engaged_users
engagement	2014-07-20	12
engagement	2014-07-21	42
engagement	2014-07-22	44
engagement	2014-07-23	53
engagement	2014-07-24	50
engagement	2014-07-25	48
engagement	2014-07-26	15
engagement	2014-07-27	20
engagement	2014-07-28	49
engagement	2014-07-29	41
engagement	2014-07-30	58
engagement	2014-07-31	48
engagement	2014-08-01	53
engagement	2014-08-02	18
engagement	2014-08-03	15
engagement	2014-08-04	34
engagement	2014-08-05	53
engagement	2014-08-06	36
engagement	2014-08-07	53
engagement	2014-08-08	41



event_type	week	weekly_engaged_users
signup_flow	2014-05-03	8
signup_flow	2014-05-04	9
signup_flow	2014-05-05	24
signup_flow	2014-05-06	27
signup_flow	2014-05-07	32
signup_flow	2014-05-08	33
signup_flow	2014-05-09	31
signup_flow	2014-05-10	7
signup_flow	2014-05-11	6
signup_flow	2014-05-12	29
signup_flow	2014-05-13	35
signup_flow	2014-05-14	34
signup_flow	2014-05-15	38
signup_flow	2014-05-16	36
signup_flow	2014-05-17	7
signup_flow	2014-05-18	7
signup_flow	2014-05-19	31
signup_flow	2014-05-20	38
signup_flow	2014-05-21	22
signup_flow	2014-05-22	35

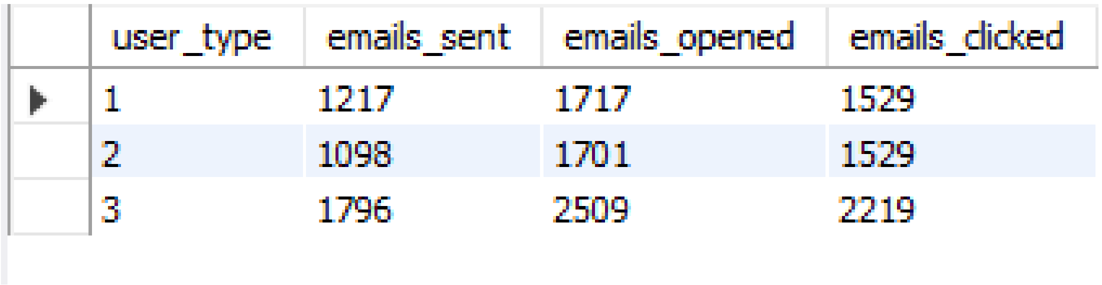


event_type	week	weekly_engaged_users
signup_flow	2014-05-23	34
signup_flow	2014-05-24	9
signup_flow	2014-05-25	8
signup_flow	2014-05-26	24
signup_flow	2014-05-27	32
signup_flow	2014-05-28	37
signup_flow	2014-05-29	33
signup_flow	2014-05-30	39
signup_flow	2014-05-31	10
signup_flow	2014-06-01	11
signup_flow	2014-06-02	33
signup_flow	2014-06-03	29
signup_flow	2014-06-04	44
signup_flow	2014-06-05	32
signup_flow	2014-06-06	39
signup_flow	2014-06-07	8
signup_flow	2014-06-08	12
signup_flow	2014-06-09	34
signup_flow	2014-06-10	28
signup_flow	2014-06-11	37



event_type	week	weekly_engaged_users
signup_flow	2014-06-12	41
signup_flow	2014-06-13	37
signup_flow	2014-06-14	7
signup_flow	2014-06-15	14
signup_flow	2014-06-16	41
signup_flow	2014-06-17	49
signup_flow	2014-06-18	45
signup_flow	2014-06-19	32
signup_flow	2014-06-20	39
signup_flow	2014-06-21	9
signup_flow	2014-06-22	7
signup_flow	2014-06-23	43
signup_flow	2014-06-24	21
signup_flow	2014-06-25	36
signup_flow	2014-06-26	42
signup_flow	2014-06-27	46
signup_flow	2014-06-28	12
signup_flow	2014-06-29	10
signup_flow	2014-06-30	35
signup_flow	2014-07-01	38

E. Email Engagement: Users engaging with the email service.
Your task: Calculate the email engagement metrics?



user_type	emails_sent	emails_opened	emails_clicked
1	1217	1717	1529
2	1098	1701	1529
3	1796	2509	2219

Result:

The result of this project is a report that can be presented to the leadership team. The report includes abrief description of the project, the approach taken, the tech stack used, the insights gained, and the results of the analysis. The results provide valuable information about the job review process and user engagement with a product, which can be used to make informed decisions and improvements.

HIRING PROCESS ANALYTICS

Project Description:

This project is about analyzing a dataset of a company which has the details of people who have registered for a particular post in a department of the company.The aim of this project is to use statistical knowledge and various formulas in Excel to draw necessary conclusions about the company. This report will include details about the number of males and females who are hired, the average salary offered, class intervals for salary, representation of people working in different departments, and representation of different post tiers using charts and graphs.

Approach:

The approach taken for this project was to first understand the data columns anddata present in the dataset. After that, the missing data was checked, and the columns with multiple categories were clubbed. Outliers were checked and removed. A data summary was drawn to get a clear understanding of the data.
Excel or Google Sheets was used to answer the questions and perform theanalysis.

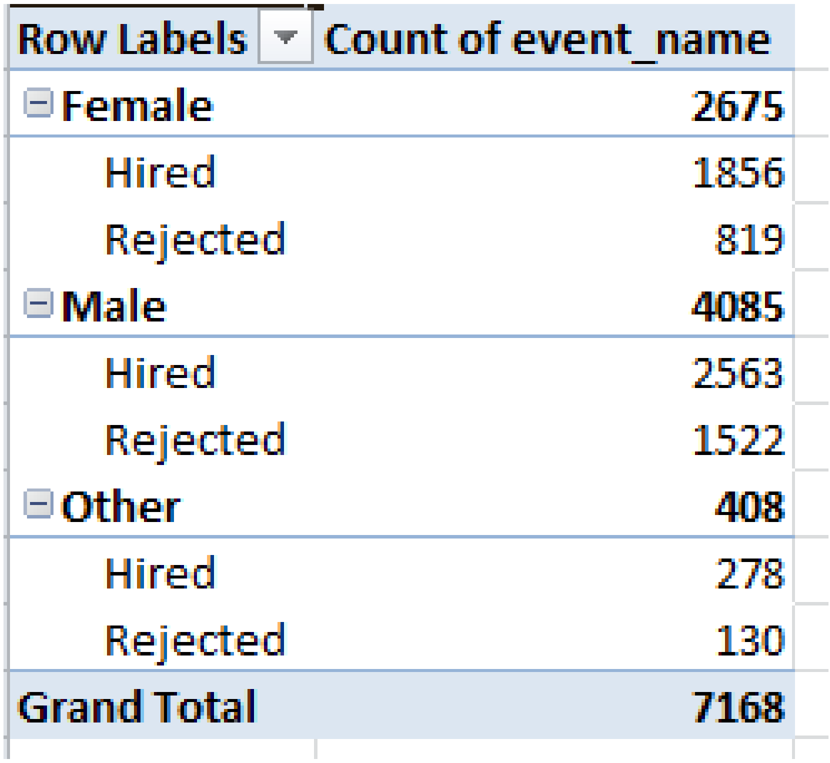
Tech-Stack Used:

The software used for this project was Microsoft Excel and the version used was Excel 365. The purpose of using Excel was to perform data analysis, create charts and graphs, and perform calculations using various formulas.

Insights:

The insights gained from this project were about the number of males and females who were hired in the company, the average salary offered, the class intervals for salary, the proportion of people working in different departments,and the representation of different post tiers using charts and graphs. The dataanalysis helped in understanding the company's hiring process and the salary offered to employees.

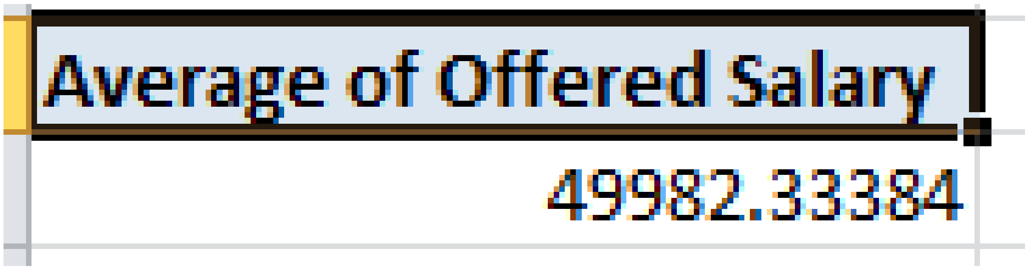
A. **Hiring:** Process of intaking of people into an organization fordifferent kinds of positions.
   Your task: How many males and females are Hired ?



Row Labels	Count of event_name
⊟ Female	**2675**
Hired	1856
Rejected	819
⊟ Male	**4085**
Hired	2563
Rejected	1522
⊟ Other	**408**
Hired	278
Rejected	130
Grand Total	**7168**

We can conclude that there are 1856 females and 2563 males are hired.

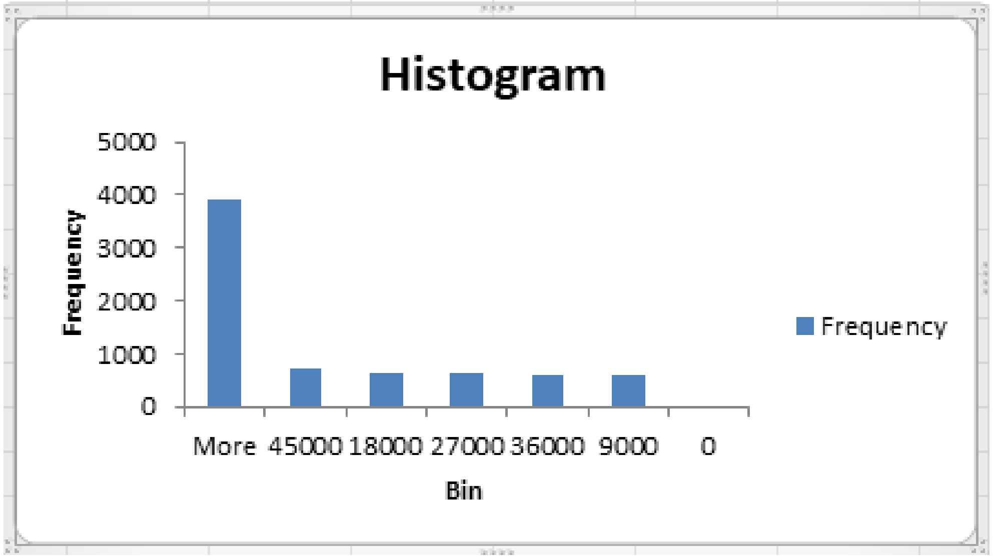
B. **Average Salary:** Adding all the salaries for a select group ofemployees and then dividing the sum by the number of employees in the group.
Your task: What is the average salary offered in this company ?



Average of Offered Salary

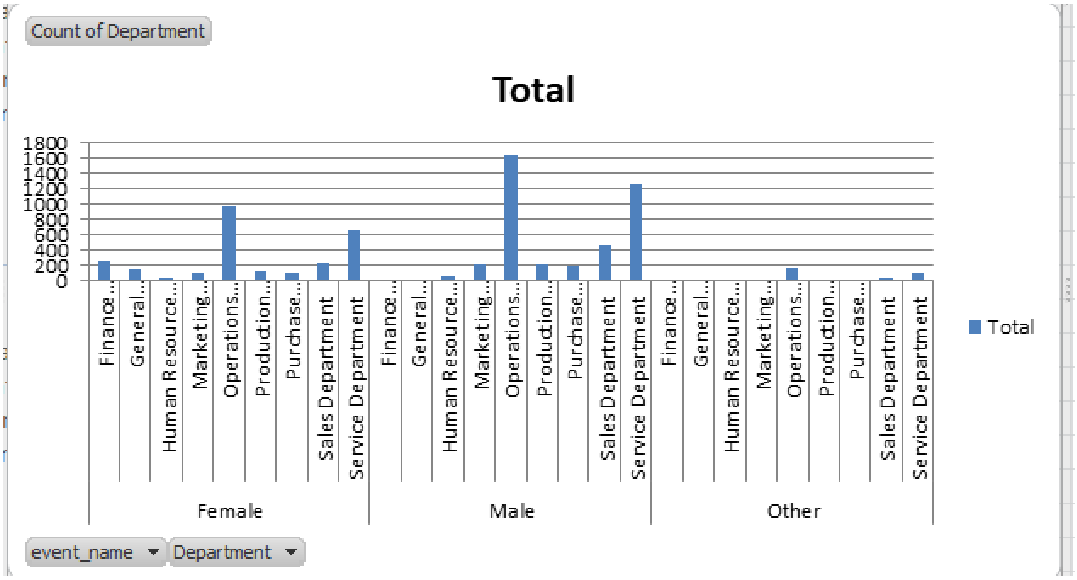
49982.33384

C. **Class Intervals:** The class interval is the difference between theupper class limit and the lower class limit.
Your task: Draw the class intervals for salary in the company ?

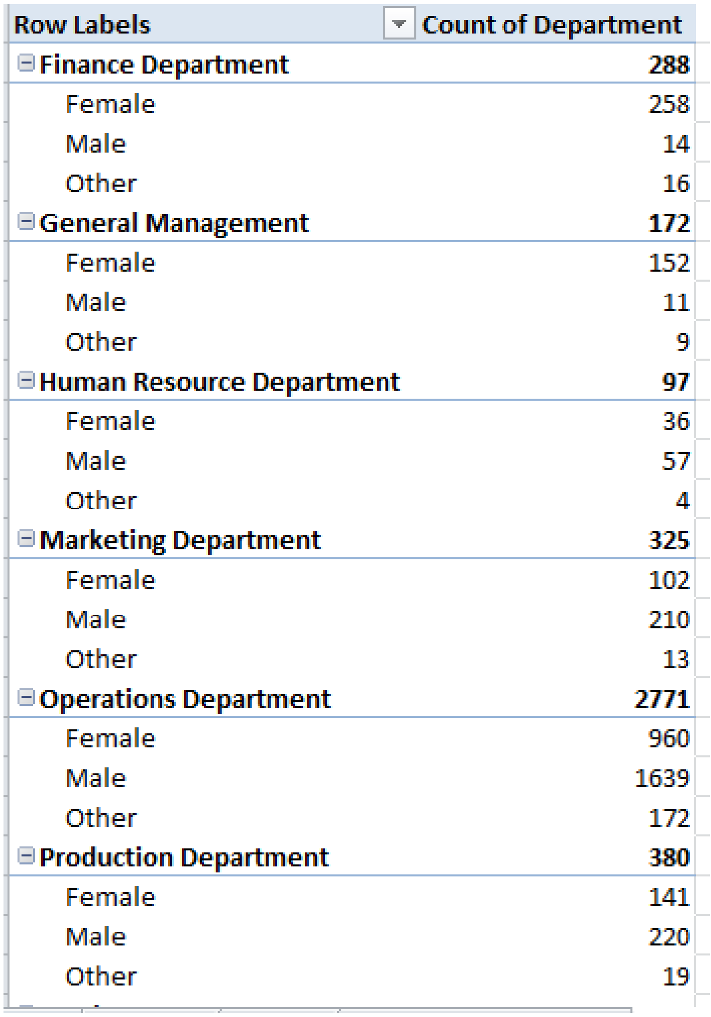


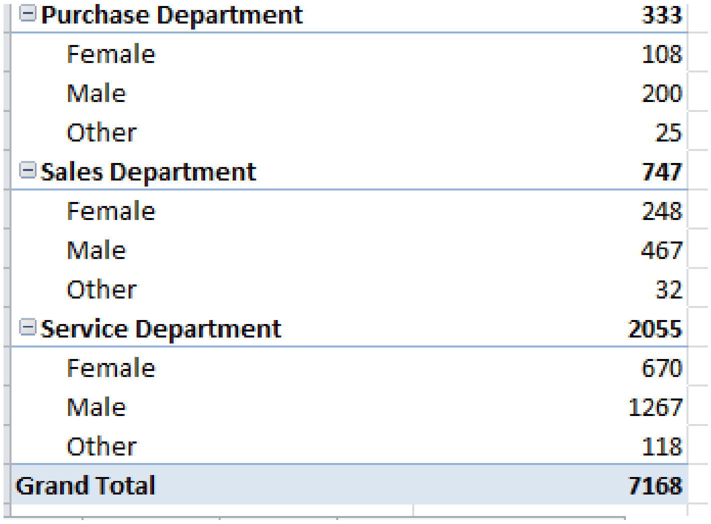
D. **Charts and Plots:** This is one of the most important part ofanalysis to visualize the data.

Your task: Draw Pie Chart / Bar Graph (or any other graph) to show proportion of people working different department ?



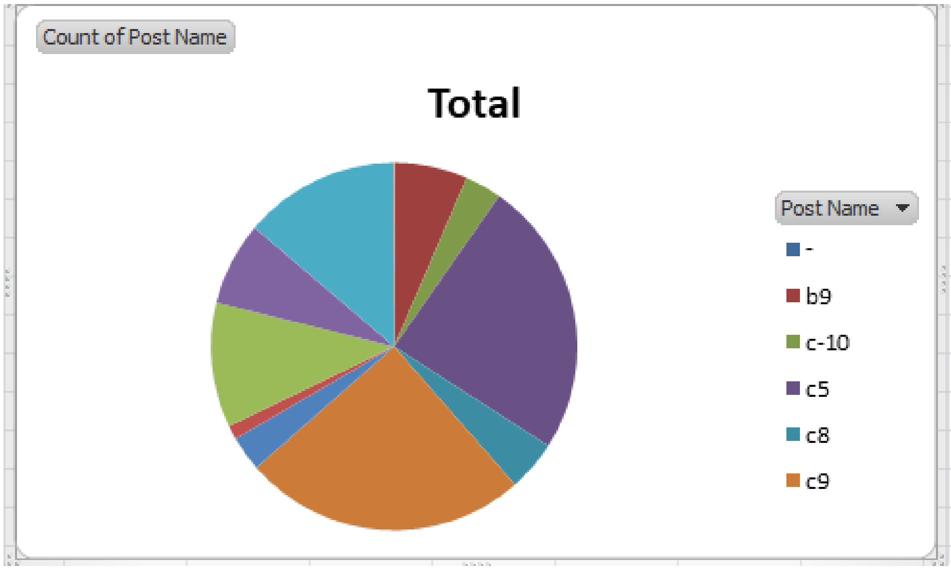
PIVOT TABLE OF THE CHART





Row Labels	Count of Department
⊟ Finance Department	**288**
Female	258
Male	14
Other	16
⊟ General Management	**172**
Female	152
Male	11
Other	9
⊟ Human Resource Department	**97**
Female	36
Male	57
Other	4
⊟ Marketing Department	**325**
Female	102
Male	210
Other	13
⊟ Operations Department	**2771**
Female	960
Male	1639
Other	172
⊟ Production Department	**380**
Female	141
Male	220
Other	19
⊟ Purchase Department	**333**
Female	108
Male	200
Other	25
⊟ Sales Department	**747**
Female	248
Male	467
Other	32
⊟ Service Department	**2055**
Female	670
Male	1267
Other	118
Grand Total	**7168**

E. **Charts:** Use different charts and graphs to perform the taskrepresenting the data.
Your task: Represent different post tiers using chart/graph?



Result:

The result of this project was a detailed report with answers to the questions mentioned in the prompt. The report includes insights gained from the data analysis and representation of the data using charts and graphs. The report provides a clear understanding of the company's hiring process and the salaryoffered to employees.

IMDB MOVIE ANALYSIS

Project Description

The project is about analyzing a dataset of movies in order to extract valuable insights and answer some business questions. The dataset includes informationabout movie titles, their cast, directors, ratings, budgets, and revenues. The analysis will be performed using Excel or Google Sheets.

Approach

The first step is to clean the dataset by removing null values and dropping unnecessary columns. Then, we will create a new column called "profit" to calculate the difference between gross and budget for each movie. We will use this column to find the movies with the highest profit and plot a scatter chart to observe the outliers.

Next, we will create a new column called "IMDb_Top_250" to store the top 250 movies with the highest IMDb rating and num_voted_users greater than 25,000.We will also add a new column called "Rank" to indicate the rank of each movie.We will extract all the non-English movies from this column and store them in a new column called "Top_Foreign_Lang_Film".

Then, we will group the dataset by director name and find the top 10 directors with the highest mean IMDb score. In case of a tie, we will sort them alphabetically. We will also find the popular genres by analyzing the frequency ofeach genre in the dataset.

After that, we will create three new columns to store the movies with the leadactors "Meryl Streep", "Leonardo DiCaprio", and "Brad Pitt". We will append these columns and group the resulting column by actor name. Then, we will findthe actors with the highest mean num_critic_for_reviews and num_users_for_review.

Finally, we will create a new column called "decade" to represent the decade towhich each movie belongs. We will observe the change in the number of voted users over decades using a bar chart and find the sum of users voted in each decade. We will store this information in a new data frame called "df_by_decade".

Tech-Stack Used

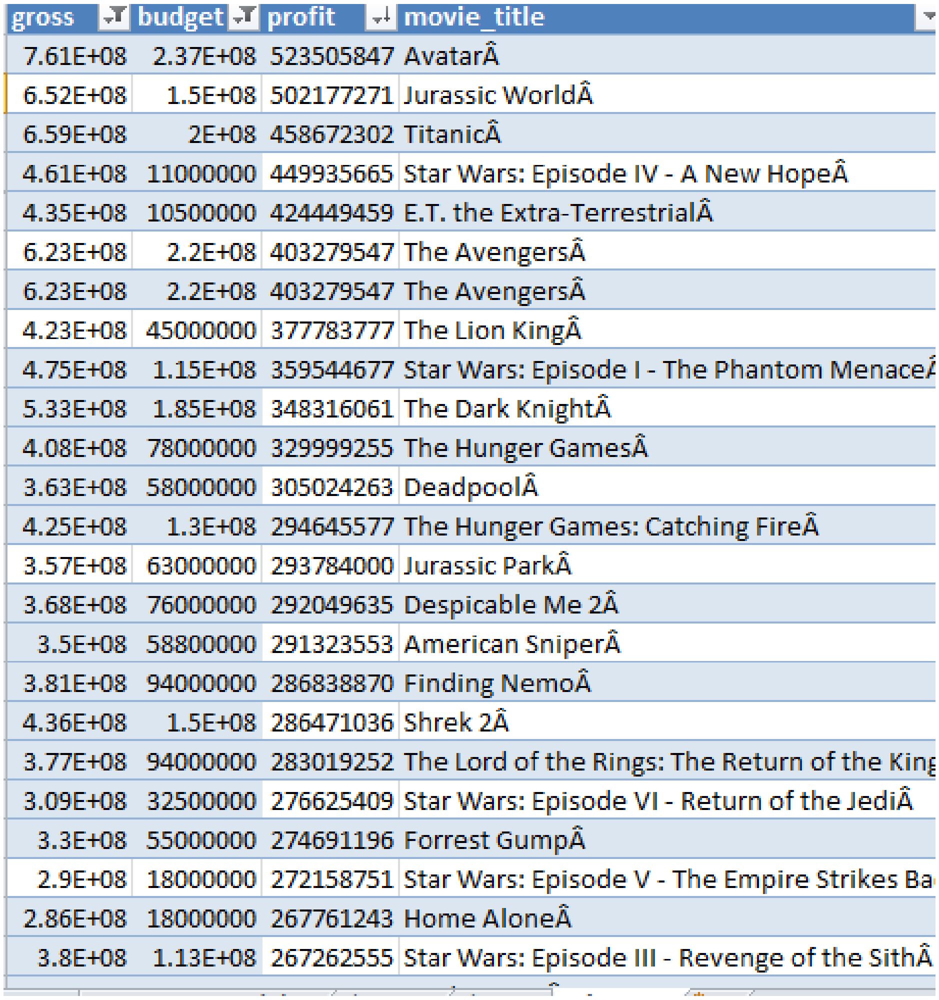
The analysis was performed using Excel, which is a spreadsheet software that allows us to perform data analysis, create charts and pivot tables, and use variousfunctions and formulas. It was chosen because it is a widely used tool that is accessible and easy to learn.

Insights

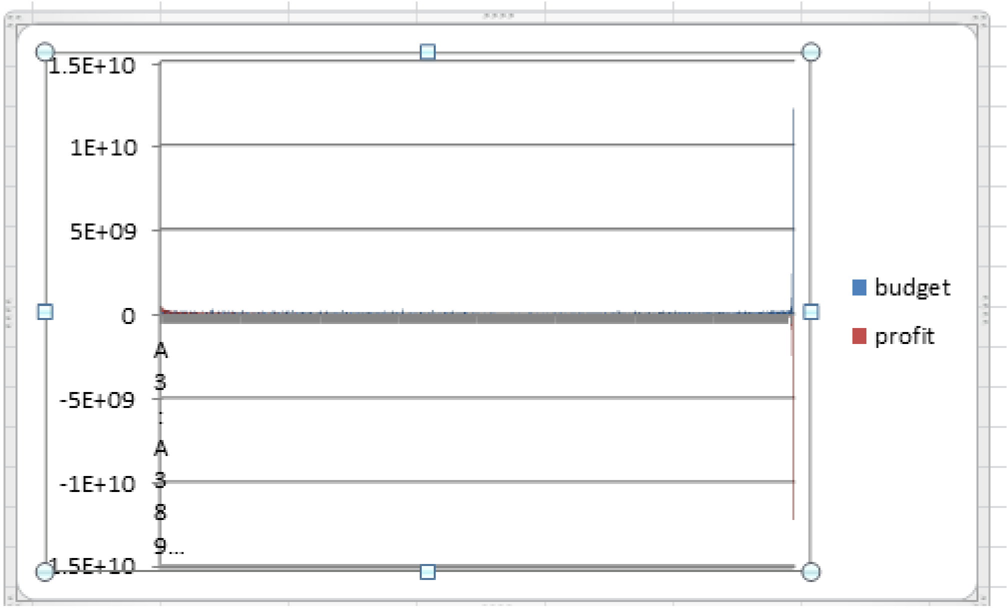
The analysis revealed several insights about the movie industry. We found that the highest profit movies have a budget of around 200 million dollars and a profitof over 1 billion dollars. We also observed some outliers with a very high profit and low budget.

A. **Movies with highest profit:** Create a new column called profit which contains the difference of the two columns: gross and budget. Sort thecolumn using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate charttype.

Your task: Find the movies with the highest profit?

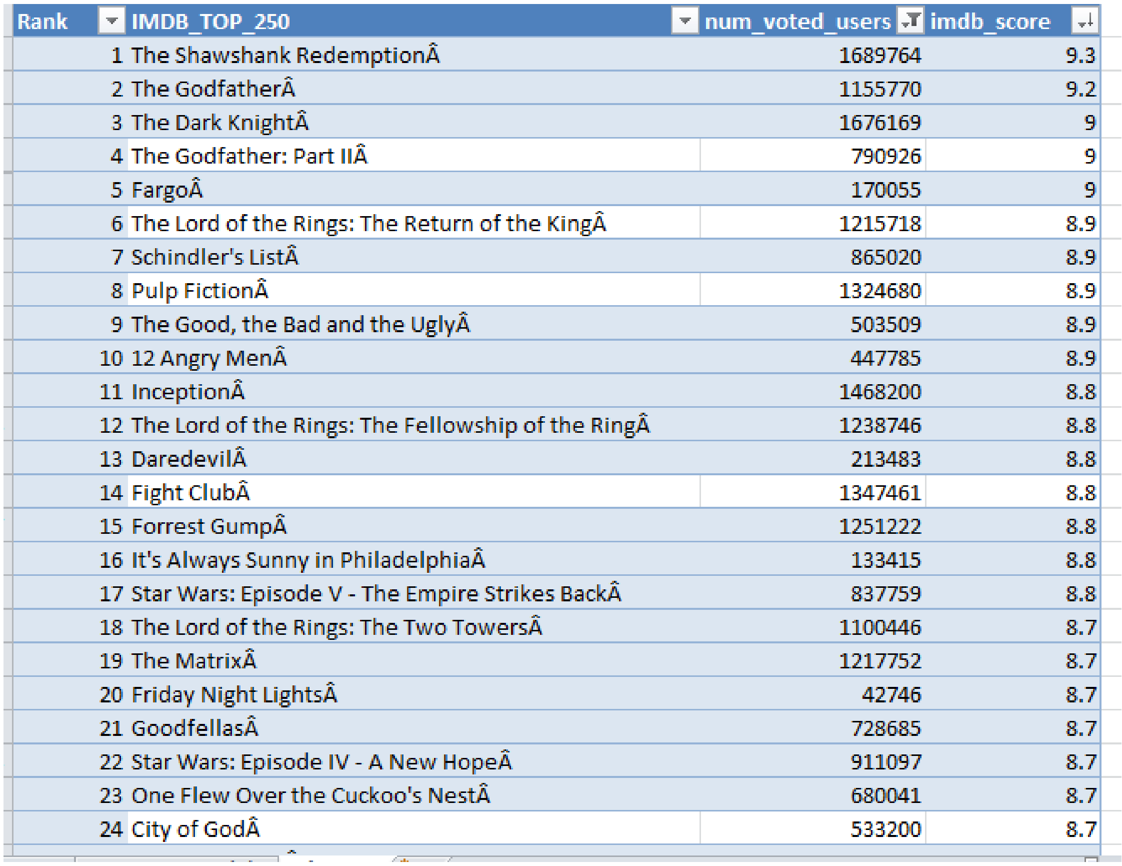


gross	budget	profit	movie_title
7.61E+08	2.37E+08	523505847	AvatarÂ
6.52E+08	1.5E+08	502177271	Jurassic WorldÂ
6.59E+08	2E+08	458672302	TitanicÂ
4.61E+08	11000000	449935665	Star Wars: Episode IV - A New HopeÂ
4.35E+08	10500000	424449459	E.T. the Extra-TerrestrialÂ
6.23E+08	2.2E+08	403279547	The AvengersÂ
6.23E+08	2.2E+08	403279547	The AvengersÂ
4.23E+08	45000000	377783777	The Lion KingÂ
4.75E+08	1.15E+08	359544677	Star Wars: Episode I - The Phantom Menace
5.33E+08	1.85E+08	348316061	The Dark KnightÂ
4.08E+08	78000000	329999255	The Hunger GamesÂ
3.63E+08	58000000	305024263	DeadpoolÂ
4.25E+08	1.3E+08	294645577	The Hunger Games: Catching FireÂ
3.57E+08	63000000	293784000	Jurassic ParkÂ
3.68E+08	76000000	292049635	Despicable Me 2Â
3.5E+08	58800000	291323553	American SniperÂ
3.81E+08	94000000	286838870	Finding NemoÂ
4.36E+08	1.5E+08	286471036	Shrek 2Â
3.77E+08	94000000	283019252	The Lord of the Rings: The Return of the King
3.09E+08	32500000	276625409	Star Wars: Episode VI - Return of the JediÂ
3.3E+08	55000000	274691196	Forrest GumpÂ
2.9E+08	18000000	272158751	Star Wars: Episode V - The Empire Strikes Ba
2.86E+08	18000000	267761243	Home AloneÂ
3.8E+08	1.13E+08	267262555	Star Wars: Episode III - Revenge of the SithÂ



Avatar is the most profitable movie with profit of 523505847.

B. **Top 250:** Create a new column IMDb_Top_250 and store the top 250movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the correspondingfilms.

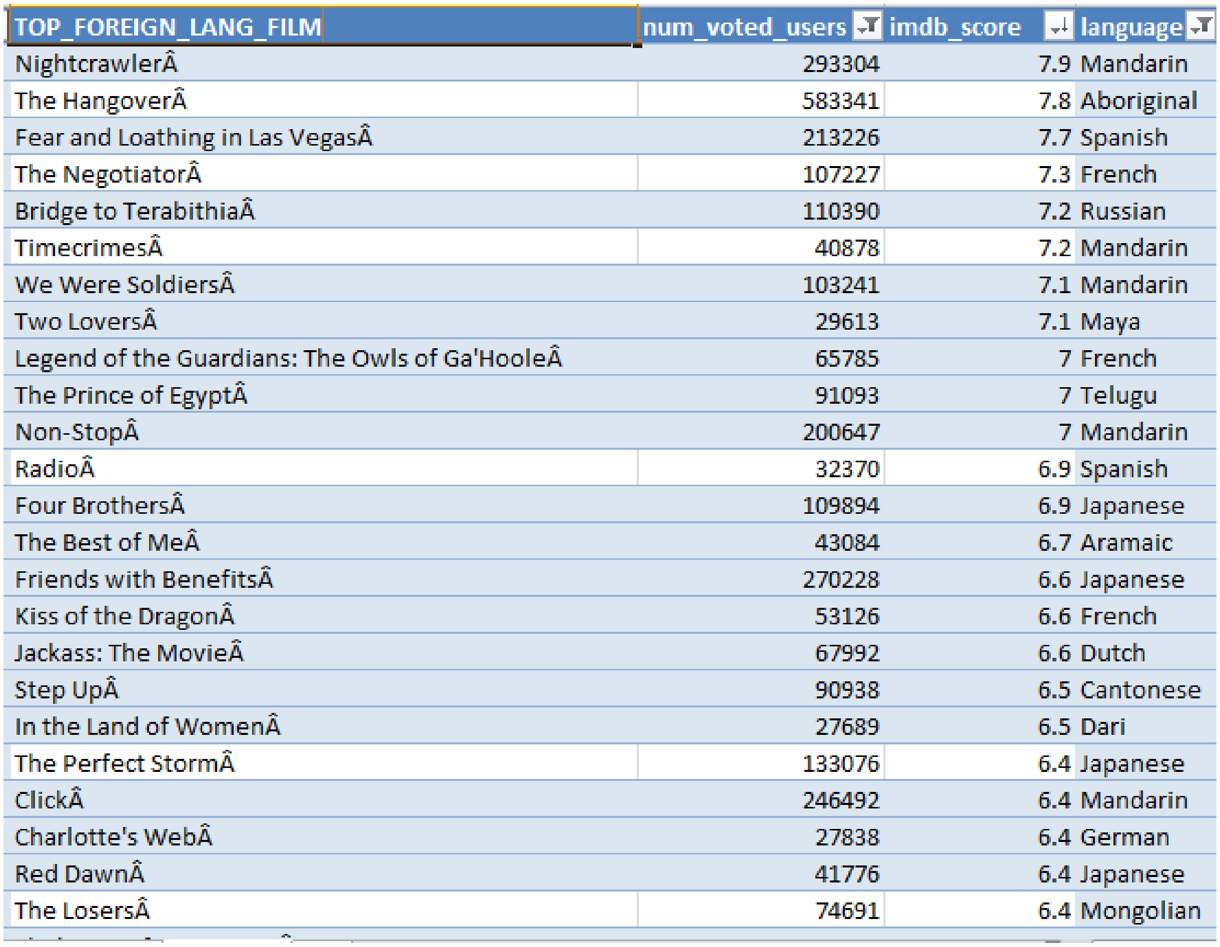


Rank	IMDB_TOP_250	num_voted_users	imdb_score
1	The Shawshank RedemptionÂ	1689764	9.3
2	The GodfatherÂ	1155770	9.2
3	The Dark KnightÂ	1676169	9
4	The Godfather: Part IIÂ	790926	9
5	FargoÂ	170055	9
6	The Lord of the Rings: The Return of the KingÂ	1215718	8.9
7	Schindler's ListÂ	865020	8.9
8	Pulp FictionÂ	1324680	8.9
9	The Good, the Bad and the UglyÂ	503509	8.9
10	12 Angry MenÂ	447785	8.9
11	InceptionÂ	1468200	8.8
12	The Lord of the Rings: The Fellowship of the RingÂ	1238746	8.8
13	DaredevilÂ	213483	8.8
14	Fight ClubÂ	1347461	8.8
15	Forrest GumpÂ	1251222	8.8
16	It's Always Sunny in PhiladelphiaÂ	133415	8.8
17	Star Wars: Episode V - The Empire Strikes BackÂ	837759	8.8
18	The Lord of the Rings: The Two TowersÂ	1100446	8.7
19	The MatrixÂ	1217752	8.7
20	Friday Night LightsÂ	42746	8.7
21	GoodfellasÂ	728685	8.7
22	Star Wars: Episode IV - A New HopeÂ	911097	8.7
23	One Flew Over the Cuckoo's NestÂ	680041	8.7
24	City of GodÂ	533200	8.7

The Shawshank Redemption is highest IMDB rated movie with 9.3.

Extract all the movies in the IMDb_Top_250 column which are not in the Englishlanguage and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!

Your task: Find IMDB Top 250



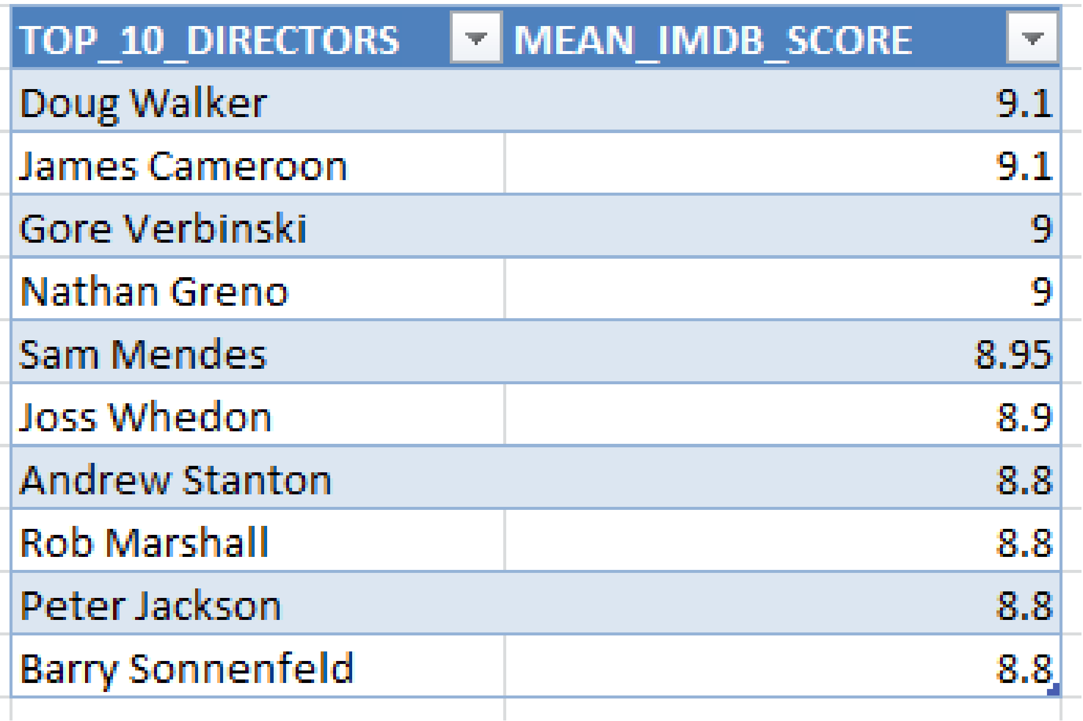
TOP_FOREIGN_LANG_FILM	num_voted_users	imdb_score	language
NightcrawlerÂ	293304	7.9	Mandarin
The HangoverÂ	583341	7.8	Aboriginal
Fear and Loathing in Las VegasÂ	213226	7.7	Spanish
The NegotiatorÂ	107227	7.3	French
Bridge to TerabithiaÂ	110390	7.2	Russian
TimecrimesÂ	40878	7.2	Mandarin
We Were SoldiersÂ	103241	7.1	Mandarin
Two LoversÂ	29613	7.1	Maya
Legend of the Guardians: The Owls of Ga'HooleÂ	65785	7	French
The Prince of EgyptÂ	91093	7	Telugu
Non-StopÂ	200647	7	Mandarin
RadioÂ	32370	6.9	Spanish
Four BrothersÂ	109894	6.9	Japanese
The Best of MeÂ	43084	6.7	Aramaic
Friends with BenefitsÂ	270228	6.6	Japanese
Kiss of the DragonÂ	53126	6.6	French
Jackass: The MovieÂ	67992	6.6	Dutch
Step UpÂ	90938	6.5	Cantonese
In the Land of WomenÂ	27689	6.5	Dari
The Perfect StormÂ	133076	6.4	Japanese
ClickÂ	246492	6.4	Mandarin
Charlotte's WebÂ	27838	6.4	German
Red DawnÂ	41776	6.4	Japanese
The LosersÂ	74691	6.4	Mongolian

Nightcrawler is the highest IMDB rated non English language film with imdb rating7.9.

C. **Best Directors:** TGroup the column using the director_name column.

Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie inIMDb score between two directors, sort them alphabetically.

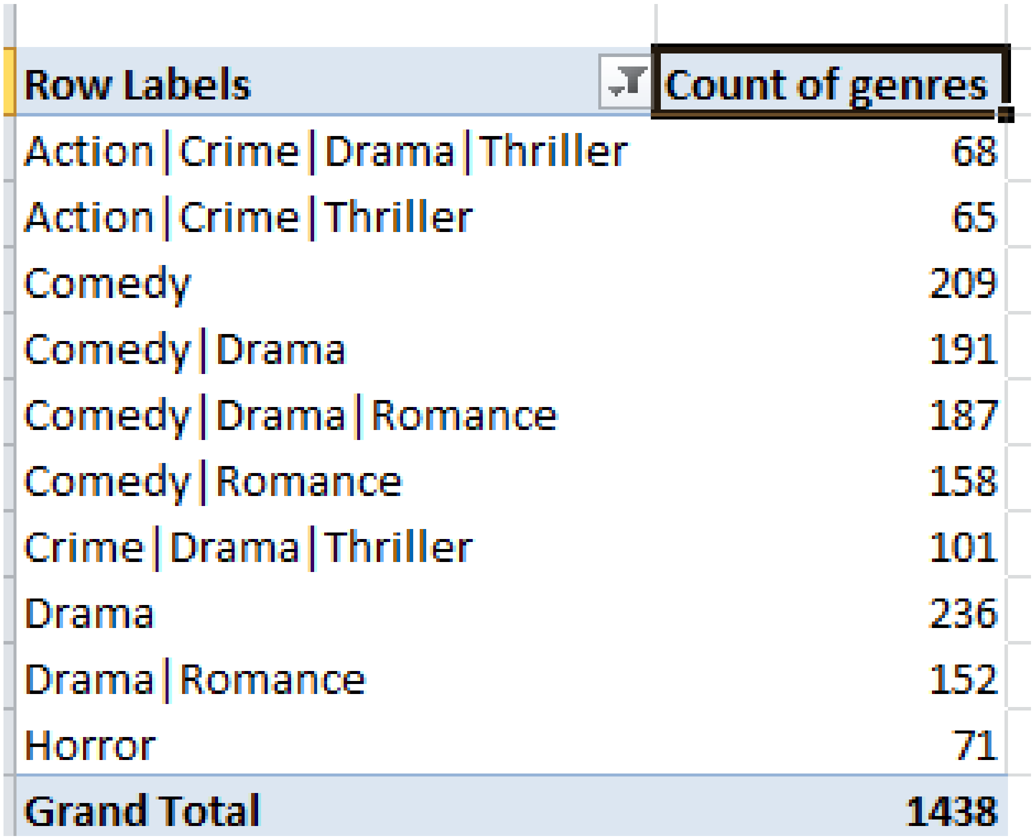
Your task: Find the best directors



TOP_10_DIRECTORS	MEAN_IMDB_SCORE
Doug Walker	9.1
James Cameroon	9.1
Gore Verbinski	9
Nathan Greno	9
Sam Mendes	8.95
Joss Whedon	8.9
Andrew Stanton	8.8
Rob Marshall	8.8
Peter Jackson	8.8
Barry Sonnenfeld	8.8

Doug Walker and James Cameroon has the joint highest mean imdb score of 9.1.

D. **Popular Genres:** Perform this step using the knowledge gained whileperforming previous steps.
   Your task: Find popular genres



Row Labels	Count of genres
Action|Crime|Drama|Thriller	68
Action|Crime|Thriller	65
Comedy	209
Comedy|Drama	191
Comedy|Drama|Romance	187
Comedy|Romance	158
Crime|Drama|Thriller	101
Drama	236
Drama|Romance	152
Horror	71
Grand Total	**1438**

Drama has the highest count of genres with 236 movies.

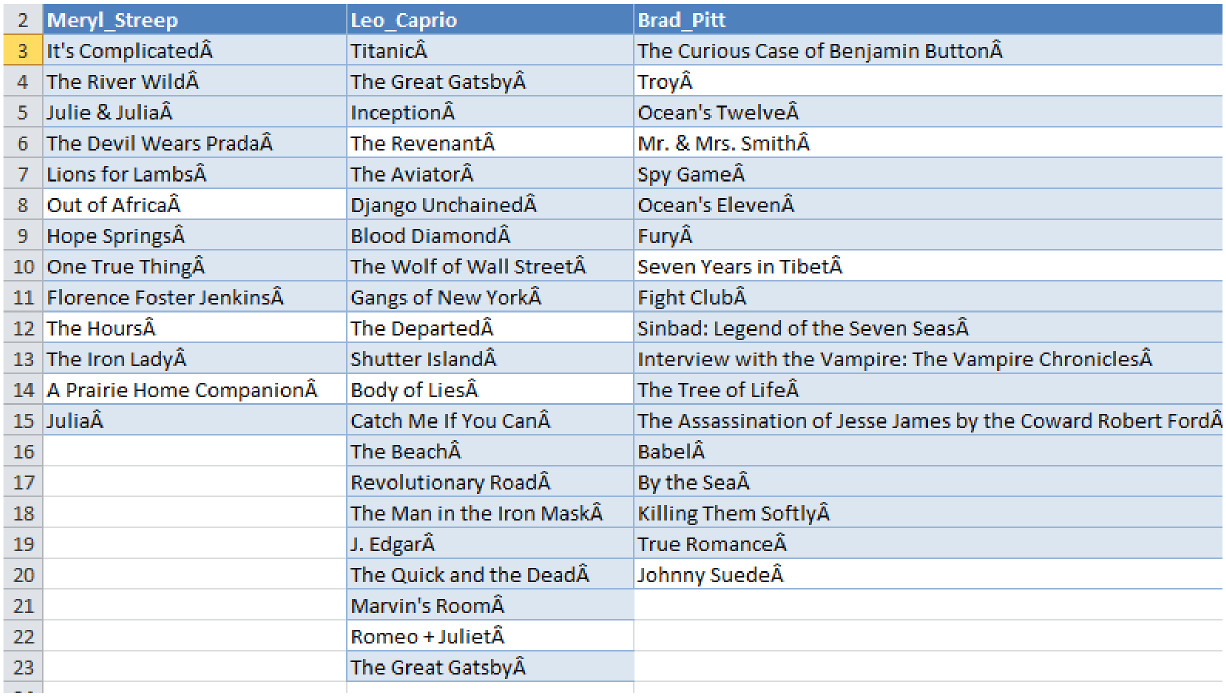
E. **Charts:** Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep','Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use thenames 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.

Append the rows of all these columns and store them in a new columnnamed Combined.

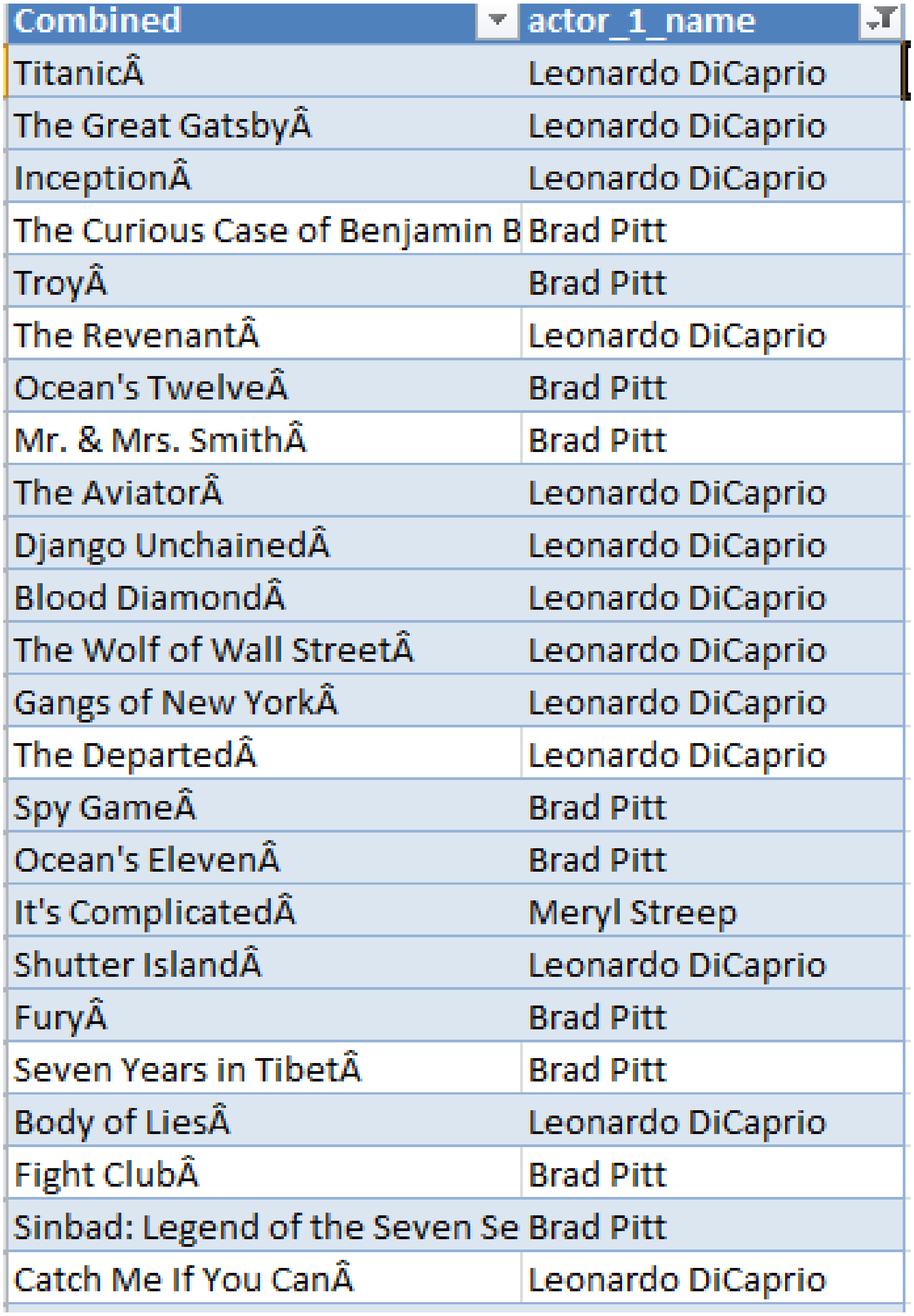
Group the combined column using the actor_1_name column.

Find the mean of the num_critic_for_reviews and num_users_for_review and identify theactors which have the highest mean.

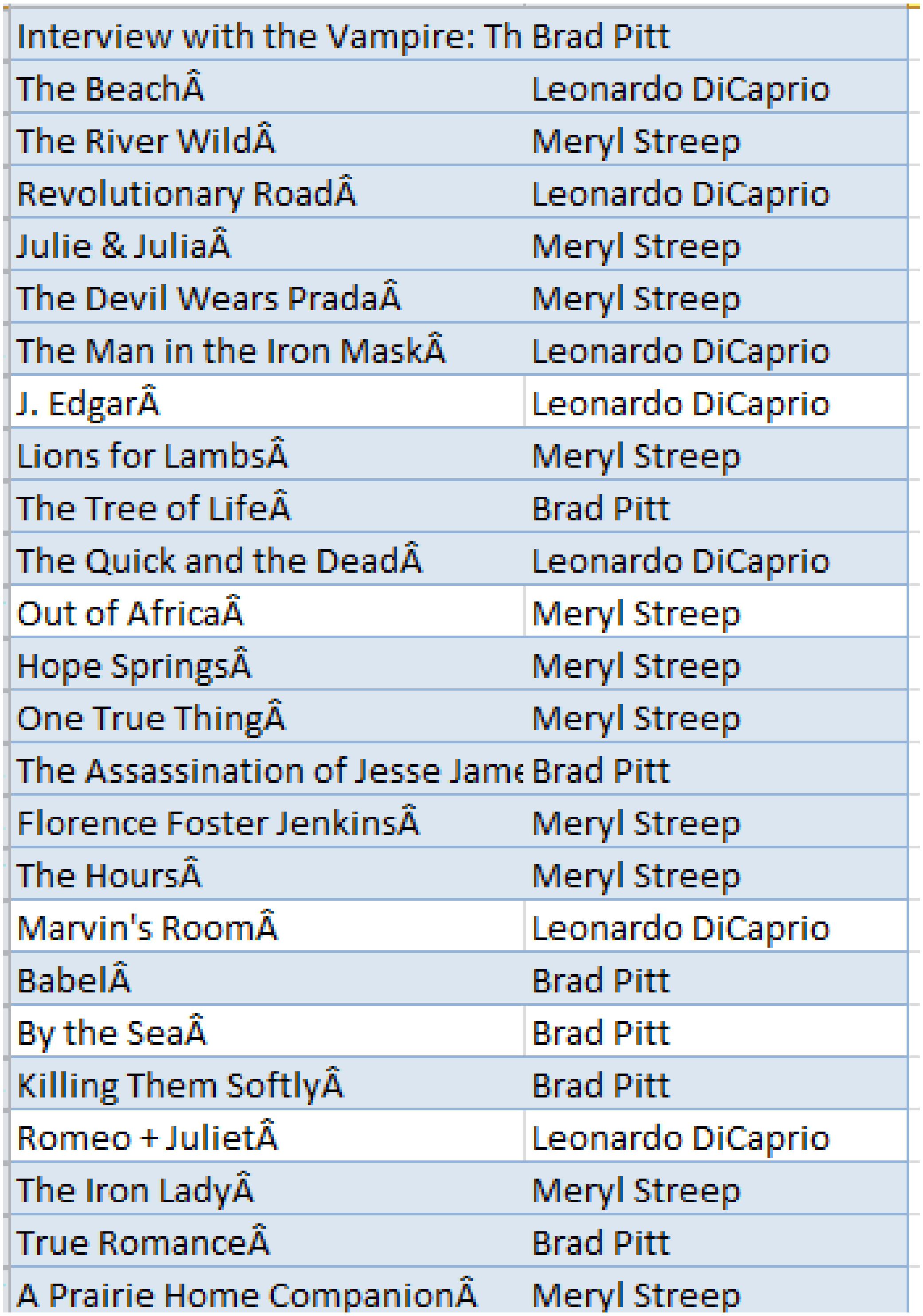
Your task: Find the critic-favorite and audience-favorite actors



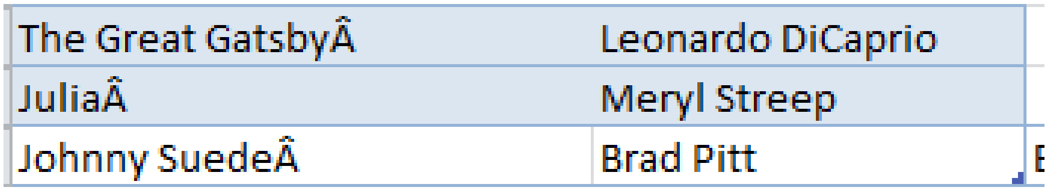
2	Meryl_Streep	Leo_Caprio	Brad_Pitt
3	It's ComplicatedÂ	TitanicÂ	The Curious Case of Benjamin ButtonÂ
4	The River WildÂ	The Great GatsbyÂ	TroyÂ
5	Julie & JuliaÂ	InceptionÂ	Ocean's TwelveÂ
6	The Devil Wears PradaÂ	The RevenantÂ	Mr. & Mrs. SmithÂ
7	Lions for LambsÂ	The AviatorÂ	Spy GameÂ
8	Out of AfricaÂ	Django UnchainedÂ	Ocean's ElevenÂ
9	Hope SpringsÂ	Blood DiamondÂ	FuryÂ
10	One True ThingÂ	The Wolf of Wall StreetÂ	Seven Years in TibetÂ
11	Florence Foster JenkinsÂ	Gangs of New YorkÂ	Fight ClubÂ
12	The HoursÂ	The DepartedÂ	Sinbad: Legend of the Seven SeasÂ
13	The Iron LadyÂ	Shutter IslandÂ	Interview with the Vampire: The Vampire ChroniclesÂ
14	A Prairie Home CompanionÂ	Body of LiesÂ	The Tree of LifeÂ
15	JuliaÂ	Catch Me If You CanÂ	The Assassination of Jesse James by the Coward Robert FordÂ
16		The BeachÂ	BabelÂ
17		Revolutionary RoadÂ	By the SeaÂ
18		The Man in the Iron MaskÂ	Killing Them SoftlyÂ
19		J. EdgarÂ	True RomanceÂ
20		The Quick and the DeadÂ	Johnny SuedeÂ
21		Marvin's RoomÂ	
22		Romeo + JulietÂ	
23		The Great GatsbyÂ	



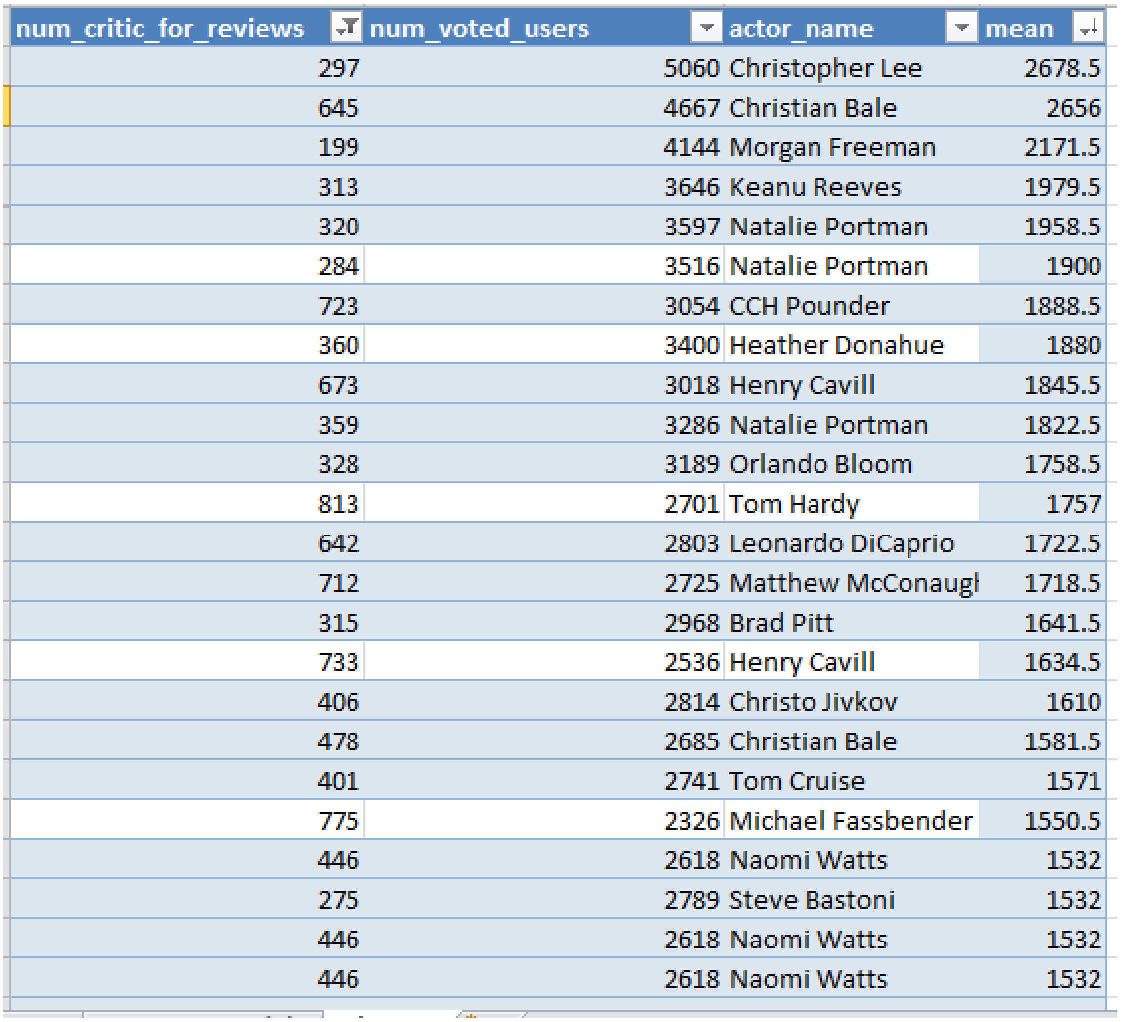
Combined	actor_1_name
TitanicÂ	Leonardo DiCaprio
The Great GatsbyÂ	Leonardo DiCaprio
InceptionÂ	Leonardo DiCaprio
The Curious Case of Benjamin B	Brad Pitt
TroyÂ	Brad Pitt
The RevenantÂ	Leonardo DiCaprio
Ocean's TwelveÂ	Brad Pitt
Mr. & Mrs. SmithÂ	Brad Pitt
The AviatorÂ	Leonardo DiCaprio
Django UnchainedÂ	Leonardo DiCaprio
Blood DiamondÂ	Leonardo DiCaprio
The Wolf of Wall StreetÂ	Leonardo DiCaprio
Gangs of New YorkÂ	Leonardo DiCaprio
The DepartedÂ	Leonardo DiCaprio
Spy GameÂ	Brad Pitt
Ocean's ElevenÂ	Brad Pitt
It's ComplicatedÂ	Meryl Streep
Shutter IslandÂ	Leonardo DiCaprio
FuryÂ	Brad Pitt
Seven Years in TibetÂ	Brad Pitt
Body of LiesÂ	Leonardo DiCaprio
Fight ClubÂ	Brad Pitt
Sinbad: Legend of the Seven Se	Brad Pitt
Catch Me If You CanÂ	Leonardo DiCaprio



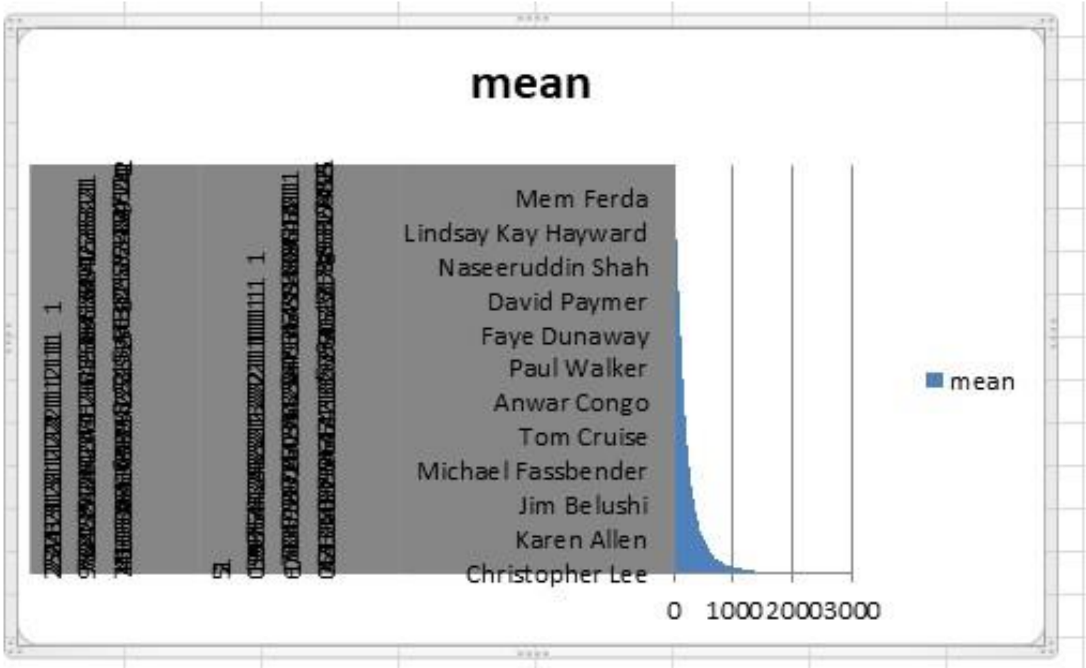
Interview with the Vampire: Th	Brad Pitt
The BeachÂ	Leonardo DiCaprio
The River WildÂ	Meryl Streep
Revolutionary RoadÂ	Leonardo DiCaprio
Julie & JuliaÂ	Meryl Streep
The Devil Wears PradaÂ	Meryl Streep
The Man in the Iron MaskÂ	Leonardo DiCaprio
J. EdgarÂ	Leonardo DiCaprio
Lions for LambsÂ	Meryl Streep
The Tree of LifeÂ	Brad Pitt
The Quick and the DeadÂ	Leonardo DiCaprio
Out of AfricaÂ	Meryl Streep
Hope SpringsÂ	Meryl Streep
One True ThingÂ	Meryl Streep
The Assassination of Jesse Jame	Brad Pitt
Florence Foster JenkinsÂ	Meryl Streep
The HoursÂ	Meryl Streep
Marvin's RoomÂ	Leonardo DiCaprio
BabelÂ	Brad Pitt
By the SeaÂ	Brad Pitt
Killing Them SoftlyÂ	Brad Pitt
Romeo + JulietÂ	Leonardo DiCaprio
The Iron LadyÂ	Meryl Streep
True RomanceÂ	Brad Pitt
A Prairie Home CompanionÂ	Meryl Streep



The Great GatsbyÂ	Leonardo DiCaprio	
JuliaÂ	Meryl Streep	
Johnny SuedeÂ	Brad Pitt	E



num_critic_for_reviews	num_voted_users	actor_name	mean
297	5060	Christopher Lee	2678.5
645	4667	Christian Bale	2656
199	4144	Morgan Freeman	2171.5
313	3646	Keanu Reeves	1979.5
320	3597	Natalie Portman	1958.5
284	3516	Natalie Portman	1900
723	3054	CCH Pounder	1888.5
360	3400	Heather Donahue	1880
673	3018	Henry Cavill	1845.5
359	3286	Natalie Portman	1822.5
328	3189	Orlando Bloom	1758.5
813	2701	Tom Hardy	1757
642	2803	Leonardo DiCaprio	1722.5
712	2725	Matthew McConaugl	1718.5
315	2968	Brad Pitt	1641.5
733	2536	Henry Cavill	1634.5
406	2814	Christo Jivkov	1610
478	2685	Christian Bale	1581.5
401	2741	Tom Cruise	1571
775	2326	Michael Fassbender	1550.5
446	2618	Naomi Watts	1532
275	2789	Steve Bastoni	1532
446	2618	Naomi Watts	1532
446	2618	Naomi Watts	1532



Result

The analysis of the movie dataset provided valuable insights into the movie industry, including the factors that contribute to the success of a movie, the mostpopular genres and actors, and the changes in audience preferences over time.

The findings of this analysis can be used by movie studios and producers to makeinformed decisions about the production and marketing of movies, which can ultimately lead to greater success and profitability.

BANK LOAN CASE STUDY

Project Description:

In this project, we will be analyzing two datasets - application_data.csv and previous_application.csv to identify if a client has payment difficulties and if thereare any factors affecting this. We will be performing exploratory data analysis to understand the data and draw insights from it. We will also be identifying missingvalues, outliers, and data imbalances in the data and taking appropriate steps to handle them.

Approach:

The approach for this analysis will involve the following steps:

1. Data Understanding: Understanding the data, its structure, and variables.
2. Data Cleaning: Identifying missing values and outliers and replacing themwith appropriate methods.
3. Data Exploration: Exploring the data through univariate, segmentedunivariate, and bivariate analysis.
4. Correlation Analysis: Identifying the top 10 correlations for clients withpayment difficulties and all other cases.
5. Visualization and Insights: Presenting the most important results throughvisualizations and summarizing the insights.

Tech-Stack Used:

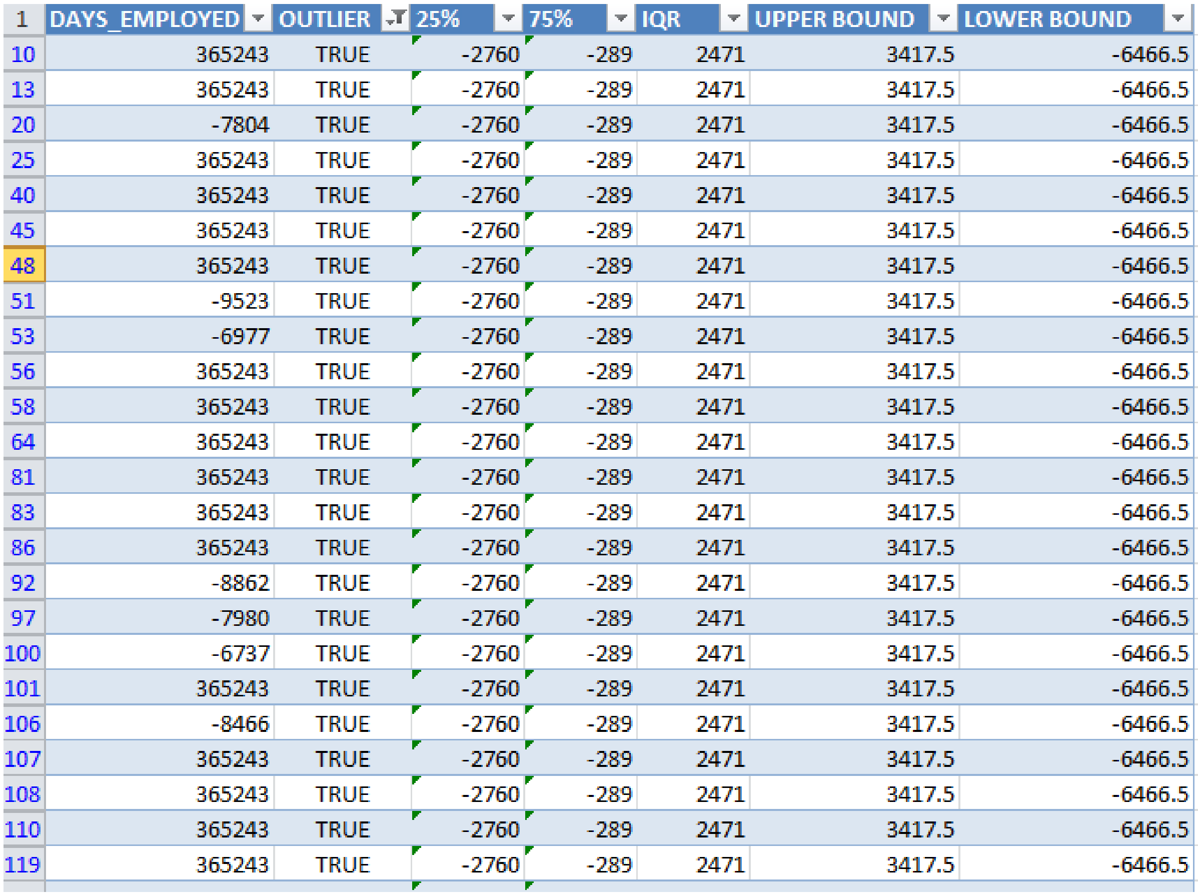
For this project, we will be using Excel to perform exploratory data analysis anddraw insights from the data.

Insights:

During our exploratory data analysis, we observed that the application_data.csv file contains information about 307,511 clients and 122 variables, while the previous_application.csv file contains information about 1,670,214 previous loanapplications and 37 variables. We observed missing values in several columns in both datasets and used various methods to handle them. We also observed outliers in some columns, which we did not remove as they seemed valid and could provide valuable insights. We identified data imbalance in the target variable, with only 8.07% of clients having payment difficulties.

In our bivariate analysis, we identified several variables that were strongly correlated with payment difficulties, including the number of days before the application when the client changed his registration, the number of days beforethe application when the client's ID document was changed, and the number ofdays before the application when the client registered his phone number.

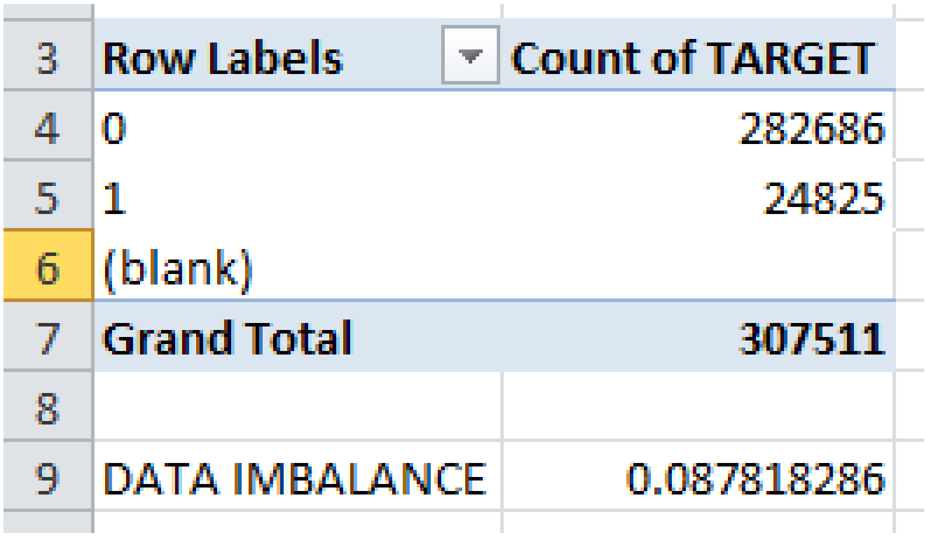
A. Identify if there are **outliers** in the dataset. Also, mention why do you think itis an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.



1	DAYS_EMPLOYED	OUTLIER	25%	75%	IQR	UPPER BOUND	LOWER BOUND
10	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
13	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
20	-7804	TRUE	-2760	-289	2471	3417.5	-6466.5
25	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
40	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
45	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
48	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
51	-9523	TRUE	-2760	-289	2471	3417.5	-6466.5
53	-6977	TRUE	-2760	-289	2471	3417.5	-6466.5
56	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
58	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
64	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
81	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
83	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
86	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
92	-8862	TRUE	-2760	-289	2471	3417.5	-6466.5
97	-7980	TRUE	-2760	-289	2471	3417.5	-6466.5
100	-6737	TRUE	-2760	-289	2471	3417.5	-6466.5
101	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
106	-8466	TRUE	-2760	-289	2471	3417.5	-6466.5
107	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
108	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
110	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
119	365243	TRUE	-2760	-289	2471	3417.5	-6466.5

These are the outliers present in the DAYS_EMPLOYED column because thisvalues has exceeded the upper bound or the values was less than the lowerbound.

B. Identify if there is data imbalance in the data. Find the ratio of dataimbalance.

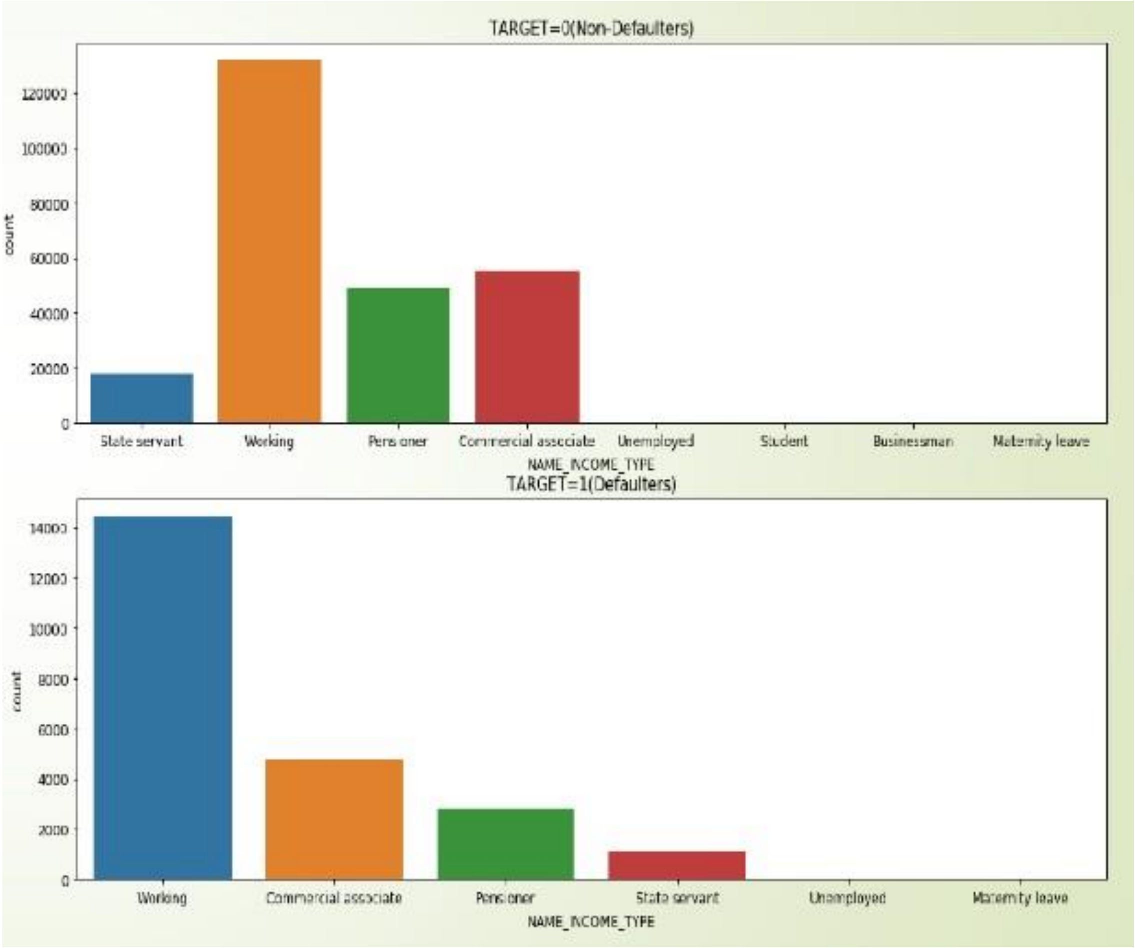


3	Row Labels	Count of TARGET
4	0	282686
5	1	24825
6	(blank)	
7	**Grand Total**	**307511**
8		
9	DATA IMBALANCE	0.087818286

C. Explain the results of univariate, bivariate analysis, etc. in business terms.

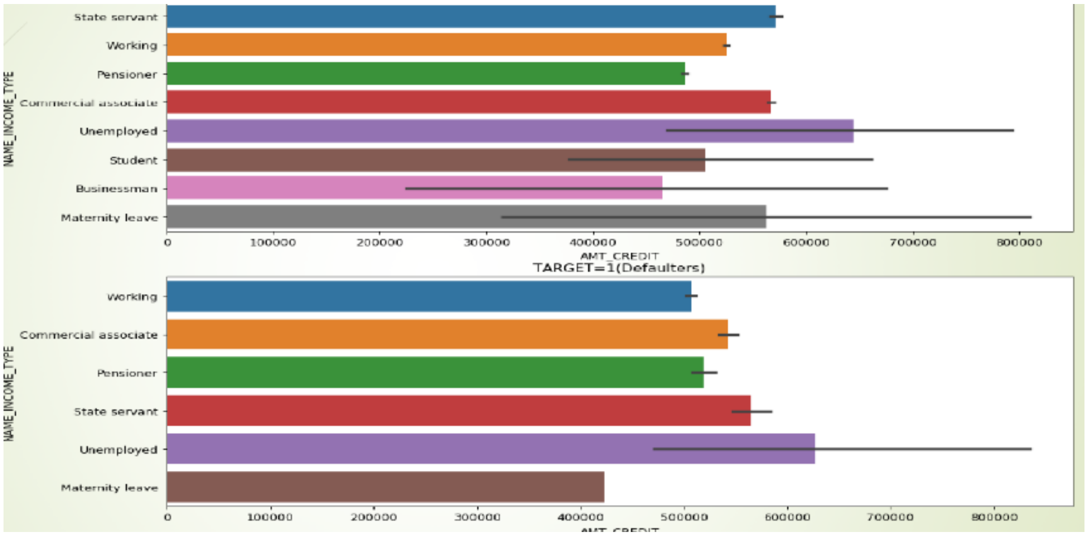
UNIVARIATE ANALYSIS

Target variable for defaulters and non defaulters

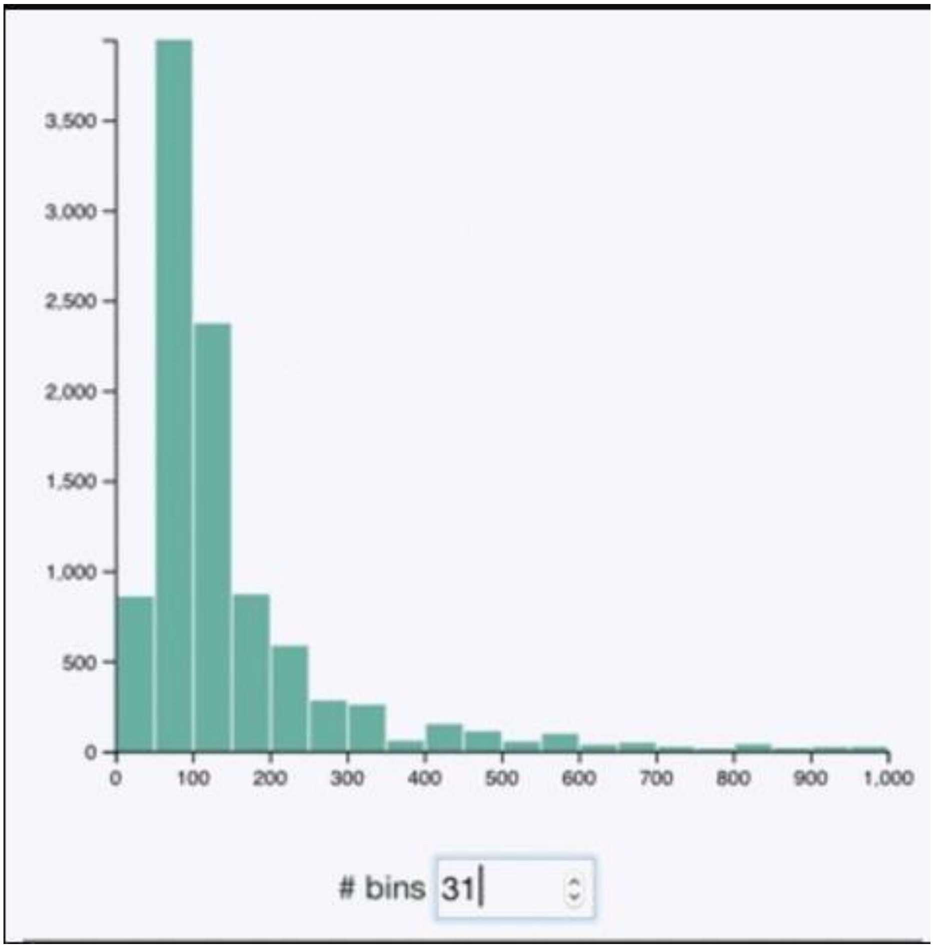


BIVARIATE ANALYSIS

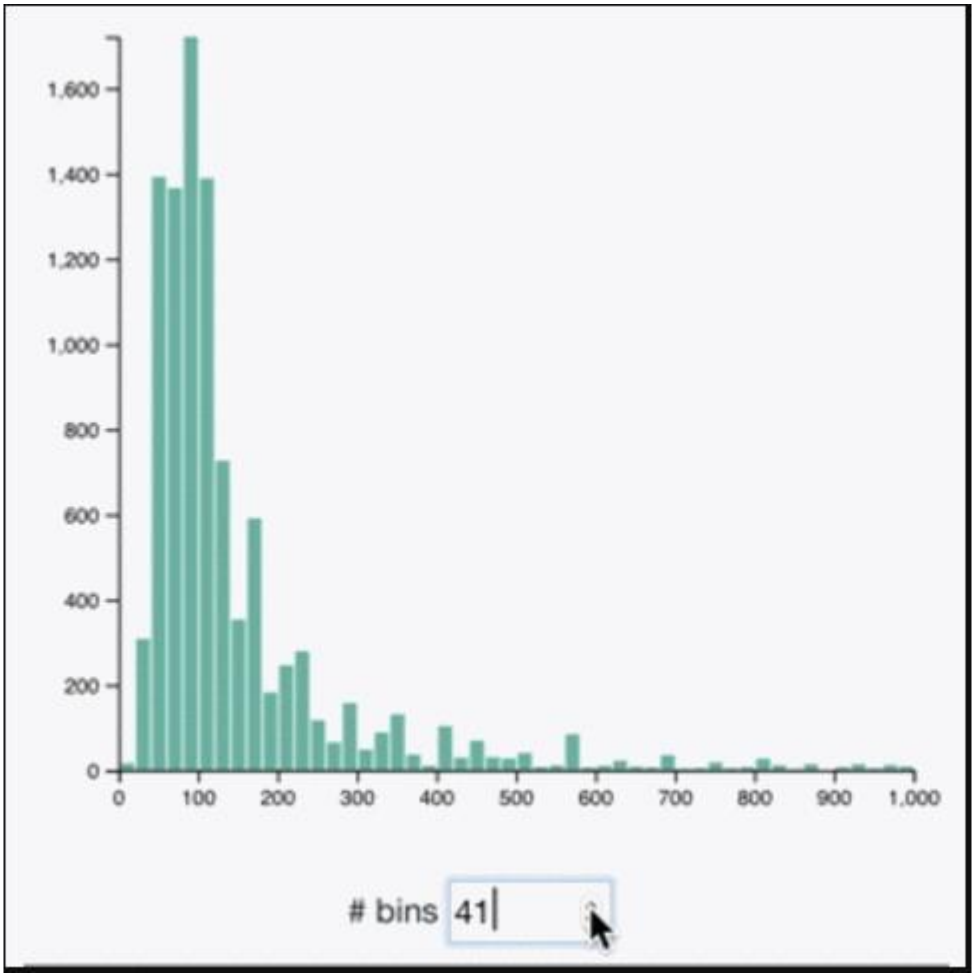
Name_income_type vs amt_credit



D. Find the top 10 **correlation** for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there.Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5,Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it isa categorical variable and not a continuous variable which is increasing or decreasing.



bins 31



Result:

Through this project, we were able to gain valuable insights into the factors affecting payment difficulties for clients. We identified several variables that werestrongly correlated with payment difficulties and could be used to predict if a client is likely to have payment difficulties. This information can be used by the company to identify high-risk clients and take appropriate steps to mitigate risk.

XYZ ADS AIRING REPORT ANALYSIS

Project Description:

This project aims to analyze the TV Ad Airings of some brands from the Automobile category to provide insights that can be used by the company to improve their advertisement strategy. Thedataset includes different variables such as the network through which Ads are airing, the typesof network like Cable/ Broadcast, the show name also on which Ads got aired, Dayparts, Time zone, the time & date at which Ads got aired, Pod Position, duration for which Ads aired on screen, Equivalent sales, total amount spent on the Ads aired, and other data.

Approach:

To start the analysis, we will first clean and preprocess the data by removing duplicates, checking for missing values, and correcting any data entry errors. We will then perform exploratory data analysis to understand the distribution of data, identify trends, and discover any outliers. After that, we will answer the questions mentioned in the case study objectives byapplying statistical analysis, creating graphs and tables, and interpreting the results.

Tech-Stack Used:

We will be using Microsoft Excel to perform our analysis. Microsoft Excel is a powerful tool thatallows us to manipulate and analyze data, create graphs and tables, and perform statistical analysis. We will be using various Excel functions, pivot tables, and charts to analyze the data and derive insights.

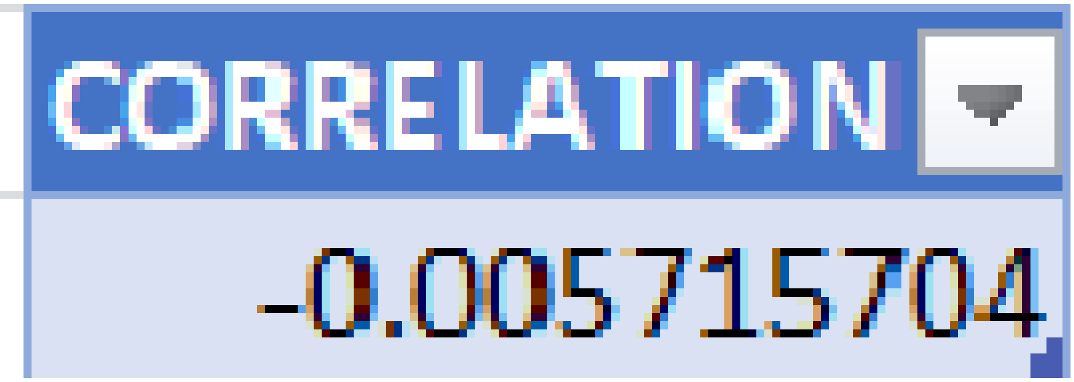
Insights:

After analyzing the data, we can derive several insights that can help the company to improve its advertising strategy. For example, we can analyze the Pod Position variable to see if it affectsthe amount spent on Ads for a specific period of time by a company. We can also analyze the share of various brands in TV airings and how it changed from Q1 to Q4 in 2021. Additionally, we can conduct a competitive analysis for the brands and define an advertisement strategy of different brands and how it differs across the brands. Finally, we can suggest a media plan to the CMO of Mahindra and Mahindra for their digital ad campaign in Q1 of 2022.

a. What is Pod Position? Does the Pod position number affect the amount spent on Ads for aspecific period of time by a company? (Explain in Details with examples from the dataset provided)

Pod position refers to the placement of an advertisement within a set of other ads. It indicatesthe position of the ad within the commercial break or pod. The ad airing positions can be pre- roll (before the show starts), mid-roll (in the middle of the show), or post-roll (after the show ends).

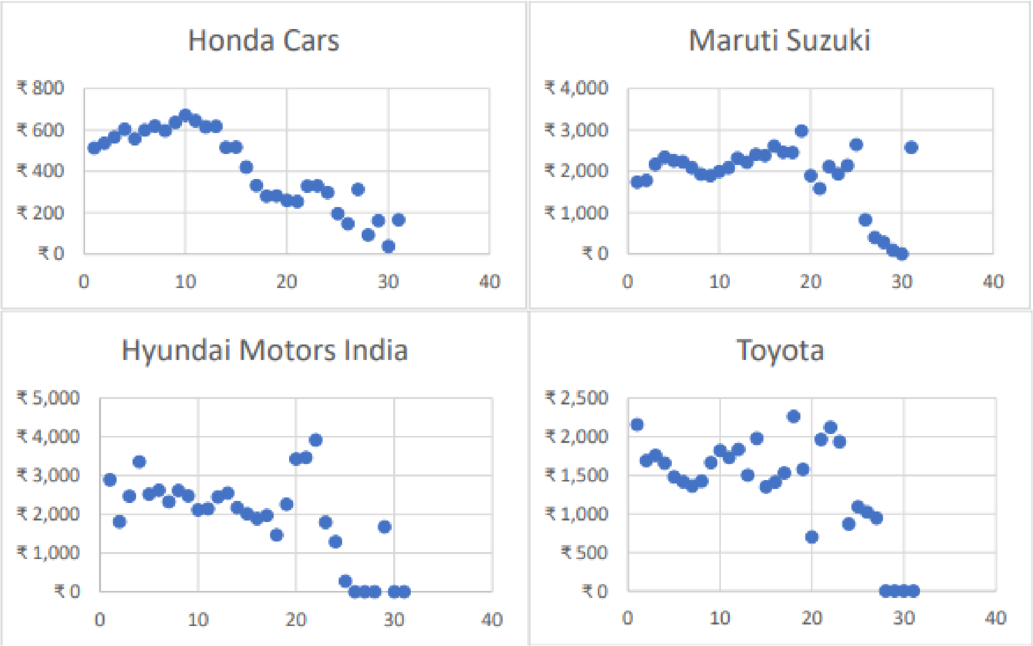
Yes, the pod position number does affect the amount spent on ads for a specific period by a company. Companies usually pay more for prime positions such as the first ad in a commercialbreak, as these positions are more likely to grab the viewer's attention.



CORRELATION
-0.005715704

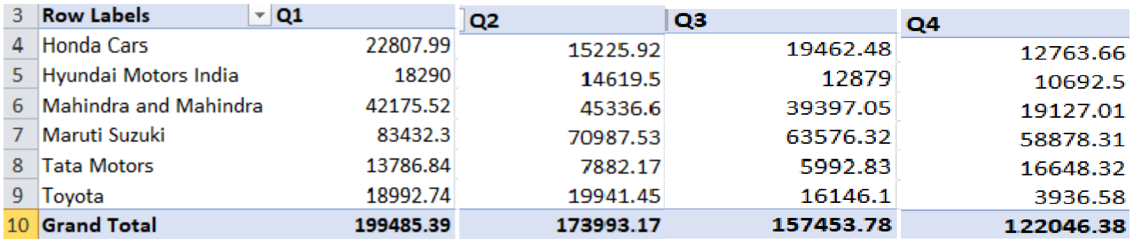
The correlation between pod position and amount spent is -0.0057, which indicates a very weaknegative correlation. This means that there is a slight tendency for companies to spend less on ads as the pod position increases, but the relationship is not strong.

In the given dataset, we can observe the variation in spending across different pod positions by analyzing the amount spent by different brands. For example, Honda spent the most on the first position ads, whereas Toyota spent the most on the second position ads. Mahindra and Maruti Suzuki, on the other hand, spent more on the third position ads. This indicates that different brands have different strategies when it comes to pod position and the amount they are willing to spend.



b. What is the share of various brands in TV airings and how has it changed from Q1 to Q4 in 2021?

To determine the share of various brands in TV airings, we can analyze the number of ad airingsby each brand in the given dataset. The table below shows the number of ad airings by each brand in each quarter of 2021.



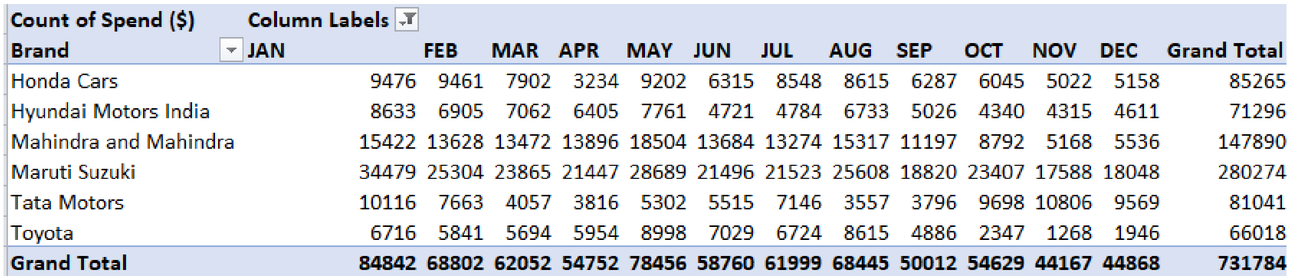
3	Row Labels	Q1	Q2	Q3	Q4
4	Honda Cars	22807.99	15225.92	19462.48	12763.66
5	Hyundai Motors India	18290	14619.5	12879	10692.5
6	Mahindra and Mahindra	42175.52	45336.6	39397.05	19127.01
7	Maruti Suzuki	83432.3	70987.53	63576.32	58878.31
8	Tata Motors	13786.84	7882.17	5992.83	16648.32
9	Toyota	18992.74	19941.45	16146.1	3936.58
10	**Grand Total**	**199485.39**	**173993.17**	**157453.78**	**122046.38**

We can see that Maruti Suzuki had the highest number of ad airings in each quarter of 2021, followed by Mahindra and Mahindra, Honda Cars, Toyota, Hyundai Motors India and Tata Motors . However, we can also observe a decline in the number of ad airings by each brand asit moves from one quarter to the next quarter.

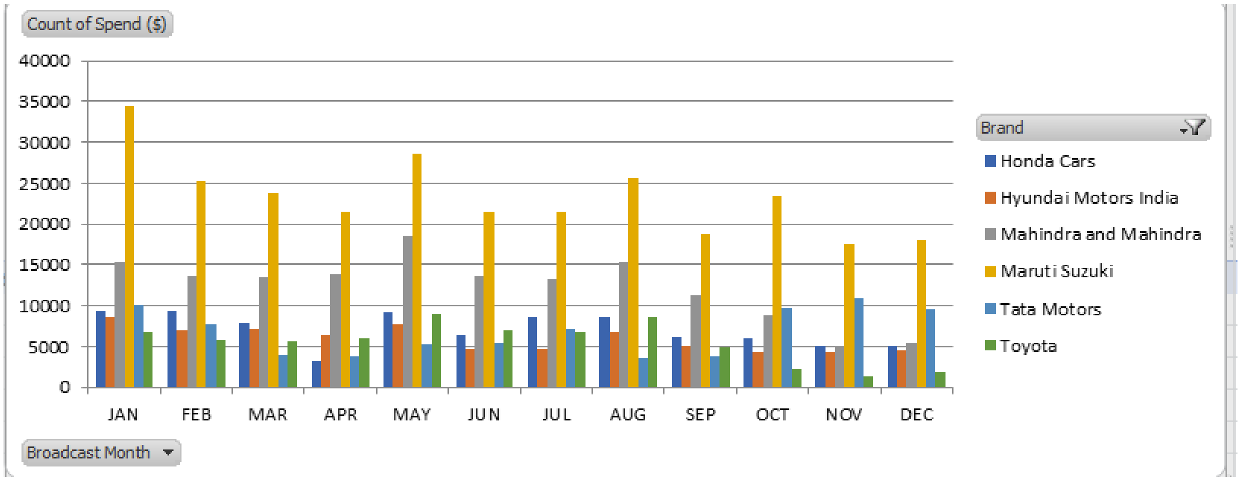
c. Conduct a competitive analysis for the brands and define advertisement strategy of different brands and how it differs across the brands.

To conduct a competitive analysis and define an advertisement strategy for different brands,we can analyze the following metrics:

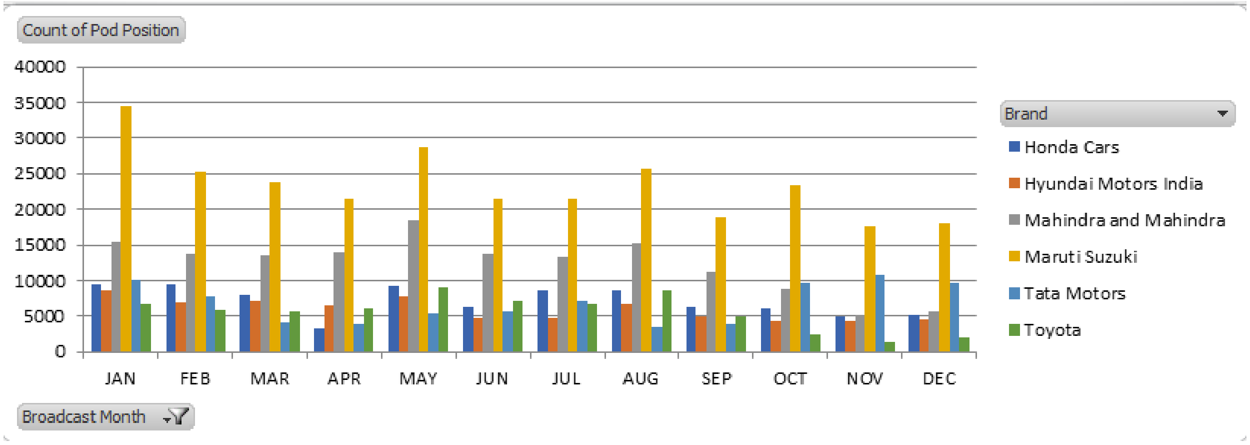
- Pod Position: As discussed earlier, pod position plays a crucial role in the effectiveness of an ad. Different brands have different strategies when it comes to pod position, andanalyzing this can help to identify the best practices and optimize ad spend.
- Ad Frequency: Ad frequency refers to the number of times a particular ad is aired duringa specific period. Analyzing ad frequency can help identify the most effective ad and optimize ad spend.
- Creative Quality: The quality of the ad creative can also impact its effectiveness. Analyzing the creative quality of different ads can help identify the best practices andoptimize ad spend



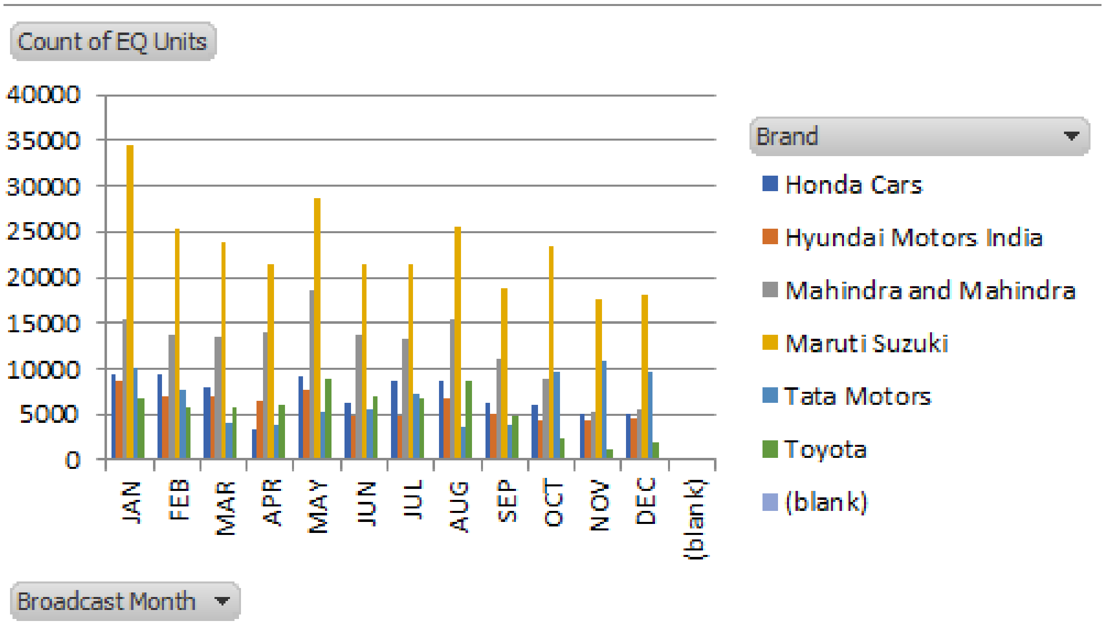
Count of Spend ($)	Column Labels												
Brand	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	Grand Total
Honda Cars	9476	9461	7902	3234	9202	6315	8548	8615	6287	6045	5022	5158	85265
Hyundai Motors India	8633	6905	7062	6405	7761	4721	4784	6733	5026	4340	4315	4611	71296
Mahindra and Mahindra	15422	13628	13472	13896	18504	13684	13274	15317	11197	8792	5168	5536	147890
Maruti Suzuki	34479	25304	23865	21447	28689	21496	21523	25608	18820	23407	17588	18048	280274
Tata Motors	10116	7663	4057	3816	5302	5515	7146	3557	3796	9698	10806	9569	81041
Toyota	6716	5841	5694	5954	8998	7029	6724	8615	4886	2347	1268	1946	66018
Grand Total	**84842**	**68802**	**62052**	**54752**	**78456**	**58760**	**61999**	**68445**	**50012**	**54629**	**44167**	**44868**	**731784**



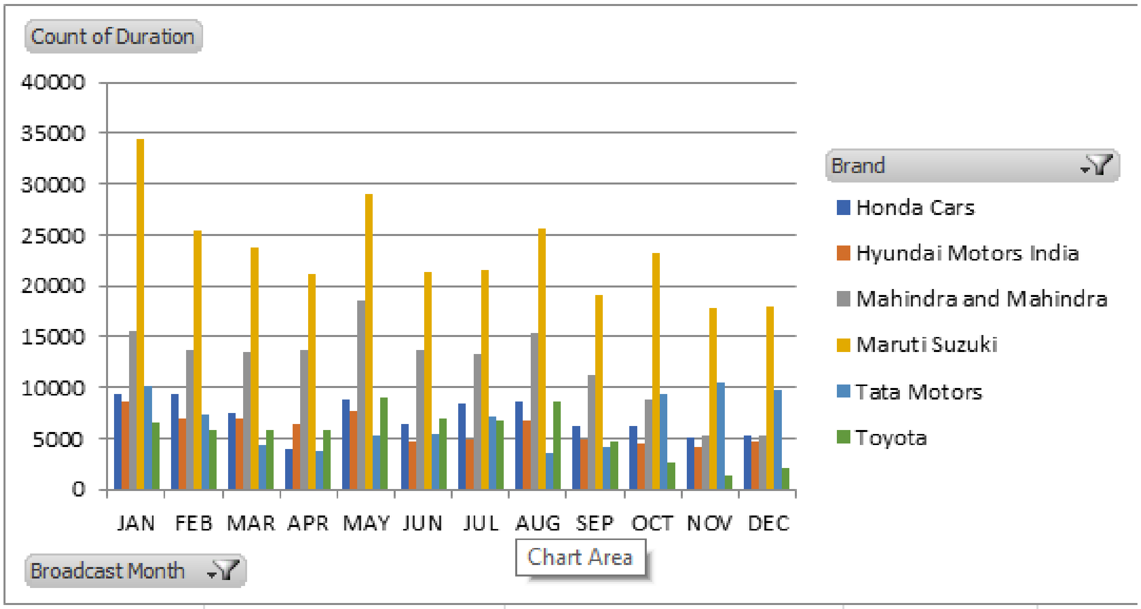
Here Maruti Suzuki spends the highest over the other brands and its highest was recorded inJanuary, whie Toyota and Honda Cars spends the least.



Mauti Suzuki has the highest count of Pod position and its highest was recorded in January,while Toyota and Honda Cars has the least count of POD position.



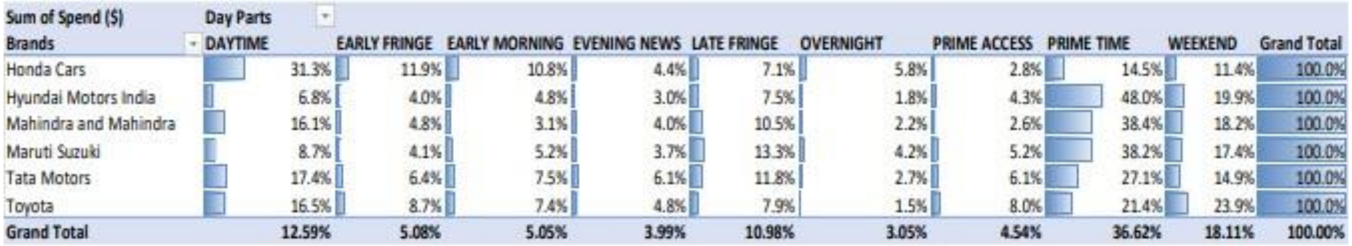
Maruti Suzuki has the highest count of EQ units and its highest was recorded in January.



Maruti Suzuki has the highest duration and its ad frequency.

From the insights above, we can easily find out that Maruti Suzuki has the highest hold over allother brands.

Analysis of every brand and why it is different from each other’s brand



Sum of Spend ($)	Day Parts									
Brands	DAYTIME	EARLY FRINGE	EARLY MORNING	EVENING NEWS	LATE FRINGE	OVERNIGHT	PRIME ACCESS	PRIME TIME	WEEKEND	Grand Total
Honda Cars	31.3%	11.9%	10.8%	4.4%	7.1%	5.8%	2.8%	14.5%	11.4%	100.0%
Hyundai Motors India	6.8%	4.0%	4.8%	3.0%	7.5%	1.8%	4.3%	48.0%	19.9%	100.0%
Mahindra and Mahindra	16.1%	4.8%	3.1%	4.0%	10.5%	2.2%	2.6%	38.4%	18.2%	100.0%
Maruti Suzuki	8.7%	4.1%	5.2%	3.7%	13.3%	4.2%	5.2%	38.2%	17.4%	100.0%
Tata Motors	17.4%	6.4%	7.5%	6.1%	11.8%	2.7%	6.1%	27.1%	14.9%	100.0%
Toyota	16.5%	8.7%	7.4%	4.8%	7.9%	1.5%	8.0%	21.4%	23.9%	100.0%
Grand Total	12.59%	5.08%	5.05%	3.99%	10.98%	3.05%	4.54%	36.62%	18.11%	100.00%

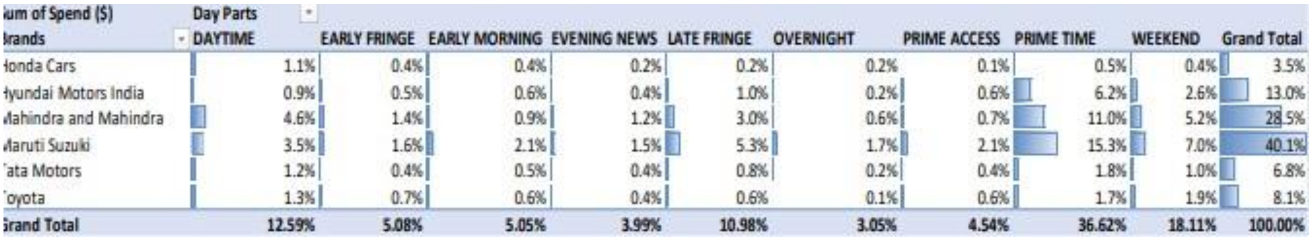
➢ Honda Cars spends the most in the daytime, early Fringe and early morning advertisement.

➢ Maruti Suzuki spends the most in late fringe and overnight advertisement.

➢ Toyota spends the most in prime access advertisement.

➢ Every car brands spends the most in prime time advertisement except the Honda Cars.

➢ Toyota spends the most in weekend advertisement.



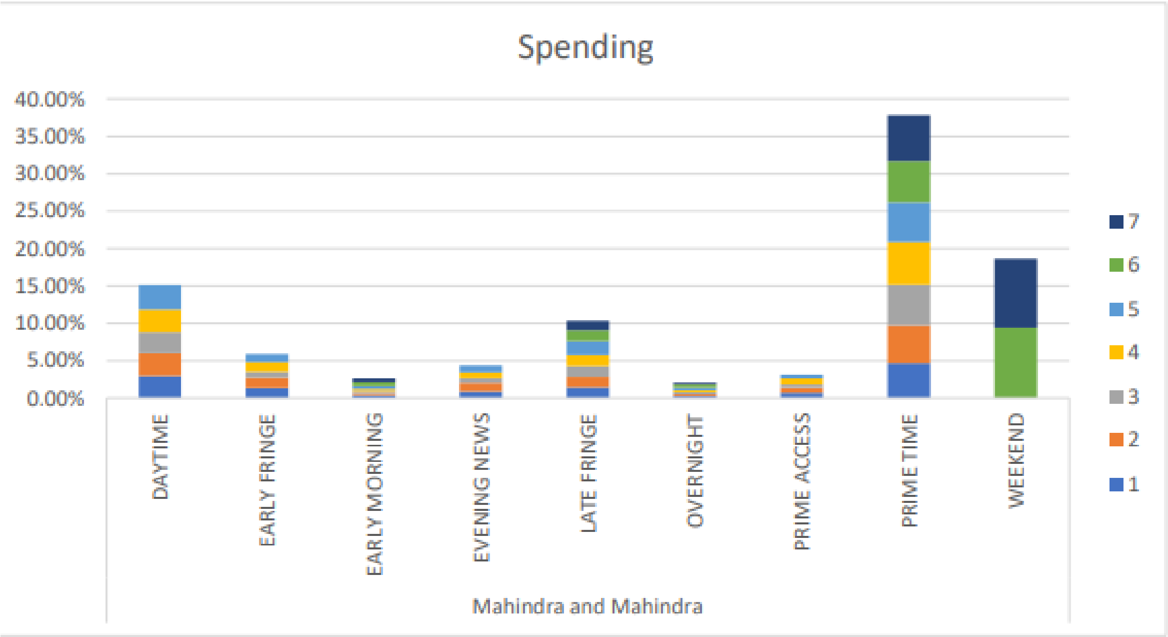
um of Spend ($)	Day Parts									
rands	DAYTIME	EARLY FRINGE	EARLY MORNING	EVENING NEWS	LATE FRINGE	OVERNIGHT	PRIME ACCESS	PRIME TIME	WEEKEND	Grand Total
onda Cars	1.1%	0.4%	0.4%	0.2%	0.2%	0.2%	0.1%	0.5%	0.4%	3.5%
yundai Motors India	0.9%	0.5%	0.6%	0.4%	1.0%	0.2%	0.6%	6.2%	2.6%	13.0%
ahindra and Mahindra	4.6%	1.4%	0.9%	1.2%	3.0%	0.6%	0.7%	11.0%	5.2%	28.5%
aruti Suzuki	3.5%	1.6%	2.1%	1.5%	5.3%	1.7%	2.1%	15.3%	7.0%	40.1%
ata Motors	1.2%	0.4%	0.5%	0.4%	0.8%	0.2%	0.4%	1.8%	1.0%	6.8%
oyota	1.3%	0.7%	0.6%	0.4%	0.6%	0.1%	0.6%	1.7%	1.9%	8.1%
rand Total	12.59%	5.08%	5.05%	3.99%	10.98%	3.05%	4.54%	36.62%	18.11%	100.00%

➢ The brands spend the least in the overnight advertisement and spends the most in the primetime advertisement

➢ Honda Cars spends the least for the advertisement. But they have the least products usedfor branding.

➢ Maruti Suzuki spends the most for the advertisement. But they have the most products usedfor branding.

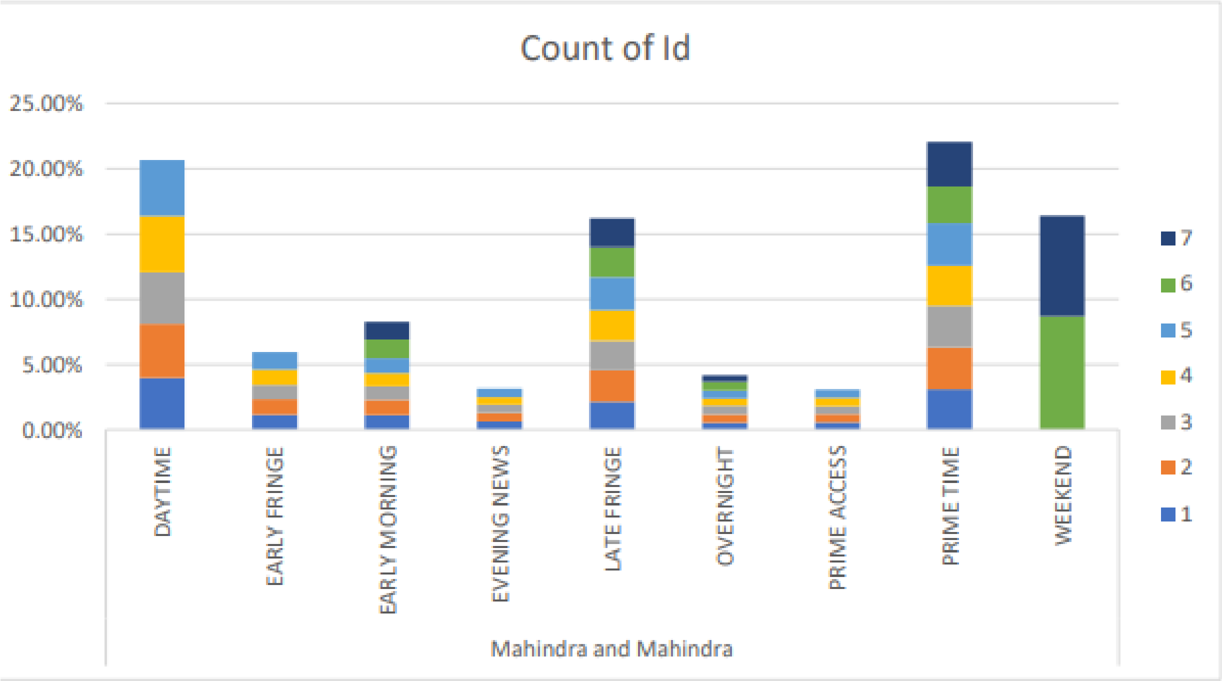
d. Mahindra and Mahindra wants to run a digital ad campaign to complement its existing TV ads in Q1 of 2022. Based on the data from 2021, suggest a media plan to the CMO of Mahindra and Mahindra. Which audience should they target? *Assume XYZ Ads has the ad viewership data and TV viewershipfor the people in India.



➢ Mahindra and Mahindra spends the most in the 6th &7th on Weekend advertisement in Q1.

➢ The company spends almost consistently in the whole week in Prime Time but with a slightincrease each time.

➢ Mahindra and Mahindra spends the least in the Overnight and Early Morning advertisementin Q1.

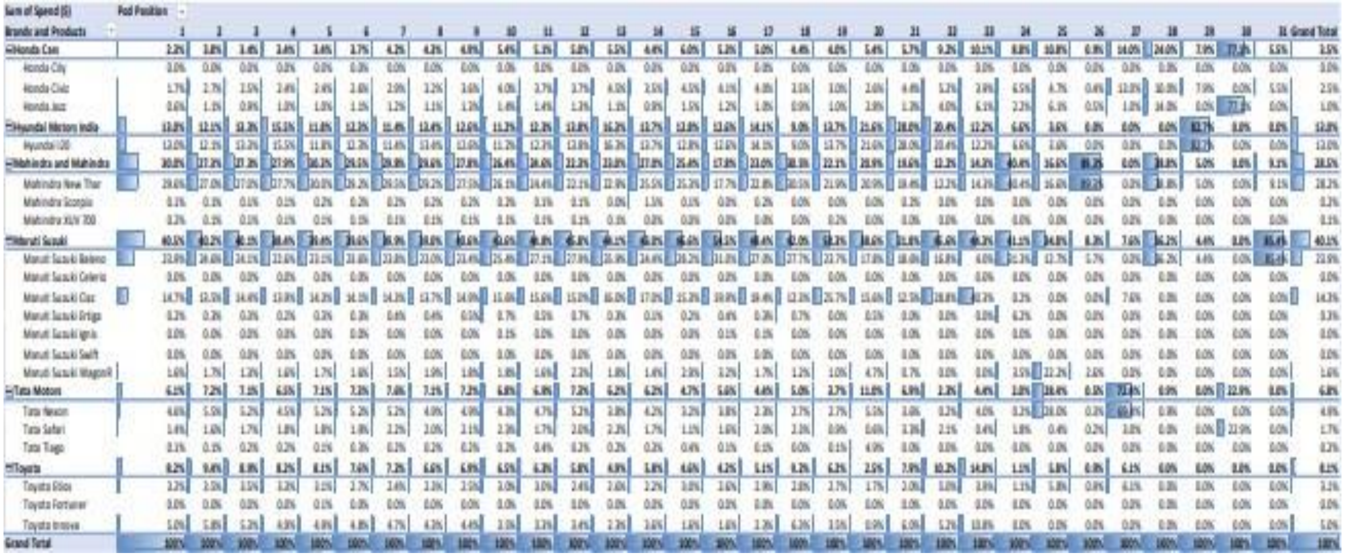


➢ The company spends around 38% of their money in Prime Time advertisement but the Adsshown is around 22%.

➢ The company spends around 15% of their money in Day Time advertisement but the Adsshown is around 21%.

➢ The most Ads shown in the 6th Day in Q1.

Additional Insights



Sum of Spend (₹)	Pod Position																														
Brands and Products	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31 Grand Total
⊟Honda Cars	2.2%	3.8%	3.4%	3.4%	3.4%	3.7%	4.2%	4.2%	4.9%	5.4%	5.1%	5.0%	5.5%	4.4%	6.0%	5.2%	5.0%	4.4%	4.0%	5.4%	5.7%	9.2%	10.1%	8.8%	10.8%	6.9%	14.0%	24.0%	7.9%	77.3%	5.5% 3.5%
Honda City	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0% 0.0%
Honda Civic	1.7%	2.7%	2.5%	2.4%	2.4%	2.6%	2.9%	3.2%	3.6%	4.0%	3.7%	3.7%	4.0%	3.5%	4.5%	4.1%	4.0%	3.5%	3.0%	2.6%	4.4%	5.2%	3.9%	6.5%	4.7%	0.4%	13.0%	10.0%	7.9%	0.0%	5.5% 2.5%
Honda Jazz	0.6%	1.1%	0.9%	1.0%	1.0%	1.1%	1.2%	1.1%	1.2%	1.4%	1.4%	1.3%	1.1%	0.9%	1.5%	1.2%	1.0%	0.9%	1.0%	2.8%	1.3%	4.0%	6.1%	2.2%	6.1%	0.5%	1.0%	14.0%	0.0%	77.3%	0.0% 1.0%
⊟Hyundai Motors India	13.0%	12.1%	13.2%	15.5%	11.8%	12.3%	11.4%	13.4%	12.6%	11.2%	12.3%	13.8%	16.2%	13.7%	12.8%	12.6%	14.1%	9.0%	13.7%	21.6%	28.0%	20.4%	12.2%	6.6%	3.6%	0.0%	0.0%	0.0%	82.1%	0.0%	0.0% 13.0%
Hyundai i20	13.0%	12.1%	13.2%	15.5%	11.8%	12.3%	11.4%	13.4%	12.6%	11.2%	12.3%	13.8%	16.3%	13.7%	12.8%	12.6%	14.1%	9.0%	13.7%	21.6%	28.0%	20.4%	12.2%	6.6%	3.6%	0.0%	0.0%	0.0%	82.1%	0.0%	0.0% 13.0%
⊟Mahindra and Mahindra	30.0%	27.2%	27.3%	27.9%	30.3%	29.5%	29.8%	29.6%	27.8%	26.4%	24.6%	22.3%	23.0%	27.0%	25.4%	17.8%	23.0%	30.5%	22.1%	20.9%	19.6%	12.3%	14.3%	40.4%	16.6%	89.3%	0.0%	38.8%	5.0%	0.0%	9.1% 28.5%
Mahindra New Thar	29.6%	27.0%	27.0%	27.7%	30.0%	29.2%	29.5%	29.2%	27.5%	26.1%	24.4%	22.1%	22.9%	25.5%	25.3%	17.7%	22.8%	30.5%	21.9%	20.9%	18.4%	12.2%	14.3%	40.4%	16.6%	89.3%	0.0%	38.8%	5.0%	0.0%	9.1% 28.2%
Mahindra Scorpio	0.1%	0.1%	0.1%	0.1%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.1%	0.1%	0.0%	1.5%	0.1%	0.0%	0.2%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0% 0.2%
Mahindra XUV 700	0.2%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0% 0.1%
⊟Maruti Suzuki	40.5%	40.2%	40.1%	38.4%	38.4%	38.6%	38.9%	38.8%	40.6%	40.6%	40.8%	45.8%	40.1%	40.2%	46.6%	54.5%	48.4%	42.0%	52.2%	38.6%	33.8%	45.6%	46.3%	41.1%	34.8%	1.3%	7.6%	36.2%	4.4%	0.0%	81.4% 40.1%
Maruti Suzuki Baleno	22.8%	24.6%	24.1%	22.6%	23.1%	21.8%	22.8%	21.0%	21.4%	25.4%	27.1%	27.9%	25.9%	24.4%	28.2%	31.0%	27.8%	27.7%	23.7%	17.8%	18.0%	16.8%	6.6%	25.3%	12.7%	1.7%	0.0%	36.2%	4.4%	0.0%	81.4% 22.8%
Maruti Suzuki Celerio	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0% 0.0%
Maruti Suzuki Ciaz	14.7%	12.5%	14.4%	13.8%	14.3%	14.1%	14.2%	13.7%	14.0%	11.6%	15.6%	15.0%	16.0%	17.0%	15.2%	19.8%	19.4%	12.3%	25.7%	15.6%	12.5%	28.8%	32.3%	3.2%	0.0%	0.0%	7.6%	0.0%	0.0%	0.0%	0.0% 14.3%
Maruti Suzuki Ertiga	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.4%	0.4%	0.5%	0.7%	0.5%	0.7%	0.3%	0.1%	0.2%	0.4%	0.3%	0.7%	0.0%	0.1%	0.0%	0.0%	0.0%	6.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0% 0.3%
Maruti Suzuki Ignis	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0% 0.0%
Maruti Suzuki Swift	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0% 0.0%
Maruti Suzuki WagonR	1.6%	1.7%	1.2%	1.6%	1.7%	1.6%	1.5%	1.9%	1.8%	1.8%	1.6%	2.2%	1.8%	1.4%	2.0%	3.2%	1.7%	1.2%	1.0%	4.7%	0.7%	0.0%	0.0%	3.5%	22.2%	2.6%	0.0%	0.0%	0.0%	0.0%	0.0% 1.6%
⊟Tata Motors	6.1%	7.2%	7.1%	6.5%	7.1%	7.2%	7.6%	7.1%	7.2%	6.8%	6.9%	7.2%	6.2%	6.2%	4.7%	5.6%	4.4%	5.0%	3.7%	11.8%	6.9%	2.3%	4.4%	3.2%	28.4%	0.1%	72.4%	0.0%	0.0%	22.9%	0.0% 6.8%
Tata Nexon	4.6%	5.5%	5.2%	4.5%	5.2%	5.2%	5.2%	4.9%	4.9%	4.3%	4.7%	5.2%	3.8%	4.2%	3.2%	3.8%	2.3%	2.7%	2.7%	5.5%	3.6%	0.2%	4.0%	0.2%	28.0%	0.3%	69.8%	0.0%	0.0%	0.0%	0.0% 4.8%
Tata Safari	1.4%	1.6%	1.7%	1.8%	1.8%	1.8%	2.2%	2.0%	2.1%	2.3%	1.7%	2.0%	2.2%	1.7%	1.1%	1.6%	2.0%	2.1%	0.9%	0.6%	3.3%	2.1%	0.4%	1.8%	0.4%	0.2%	3.0%	0.0%	0.0%	22.9%	0.0% 1.7%
Tata Tiago	0.1%	0.1%	0.2%	0.2%	0.1%	0.3%	0.2%	0.2%	0.3%	0.2%	0.4%	0.3%	0.2%	0.2%	0.4%	0.1%	0.1%	0.0%	0.1%	4.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0% 0.2%
⊟Toyota	8.2%	9.4%	8.8%	8.2%	8.1%	7.6%	7.2%	6.6%	6.9%	6.5%	6.3%	5.8%	4.9%	5.8%	6.6%	4.2%	5.1%	8.2%	6.2%	2.5%	7.9%	10.2%	14.8%	1.1%	5.8%	0.0%	6.1%	0.0%	0.0%	0.0%	0.0% 8.1%
Toyota Etios	3.2%	3.5%	3.5%	3.2%	3.1%	2.7%	2.4%	2.3%	2.5%	3.0%	3.0%	2.4%	2.6%	2.2%	3.0%	2.6%	2.9%	2.0%	2.7%	1.7%	2.0%	1.0%	1.0%	1.1%	5.8%	0.0%	6.1%	0.0%	0.0%	0.0%	0.0% 3.1%
Toyota Fortuner	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0% 0.0%
Toyota Innova	5.0%	5.8%	5.3%	4.9%	4.9%	4.8%	4.7%	4.3%	4.4%	3.5%	3.3%	3.4%	2.3%	3.6%	1.6%	1.6%	2.3%	6.3%	3.5%	0.9%	6.0%	5.2%	13.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0% 5.0%
Grand Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100% 100%

Result:

Through this project, we were able to analyze the TV Ad Airings of some brands from the Automobile category and provide insights that can be used by the company to improve their advertisement strategy. We were able to answer the questions mentioned in the case study objectives and provide a media plan to the CMO of Mahindra and Mahindra for their digital adcampaign in Q1 of 2022.

ABC CALL VOLUME TREND ANALYSIS

Project Description:

The project is about analyzing the inbound calls received by a customer experience team of ABC insurance company. The dataset includes the details of the agents, queue time, time of call, duration of the call, and call status. The objectives of the project are to calculate the average call time duration for all incoming calls received by agents in each time bucket, show the total volume/number of calls coming in via charts/graphs, propose a manpower plan required duringeach time bucket to reduce the abandon rate to 10%, and propose a manpower plan required during each time bucket in a day to attend to the calls received in the night.

Approach:

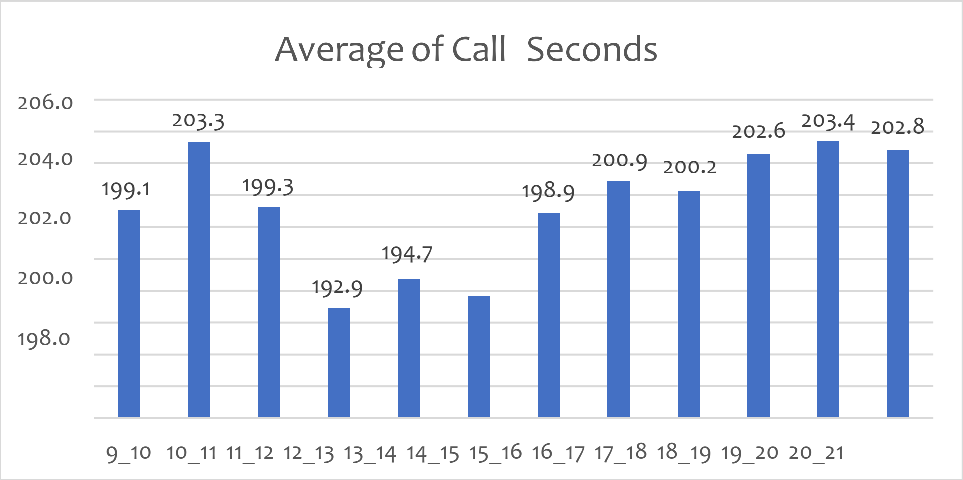
The approach involves downloading the dataset and analyzing it using Excel. The average call time duration for all incoming calls received by agents in each time bucket is calculated using pivot tables. The total volume/number of calls coming in is shown via charts/graphs. The manpower plan required during each time bucket to reduce the abandon rate to 10% is proposed using Erlang C formula. The manpower plan required during each time bucket in a dayto attend to the calls received in the night is also proposed using Erlang C formula.

Tech-Stack Used:

The project is executed using Excel.

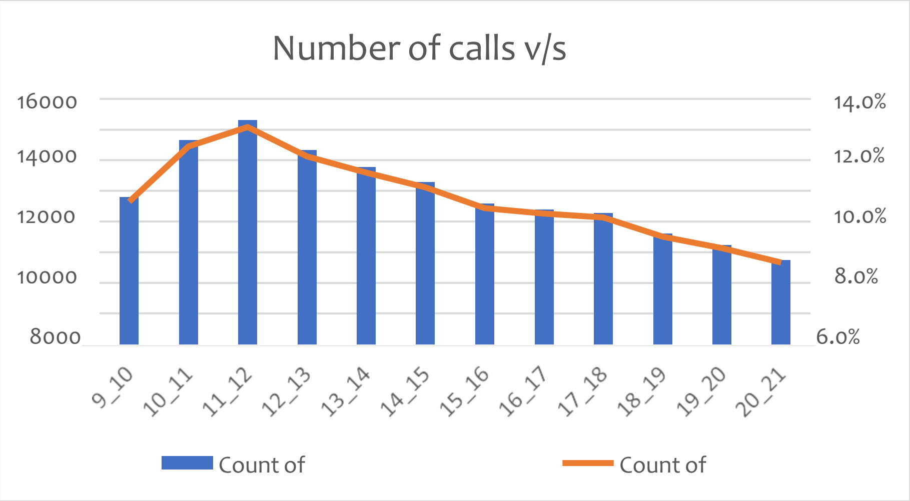
Insights:

a. Calculate the average call time duration for all incoming calls received by agents (in each Time_Bucket).



- Pivot Table is used to answer this question.
- Time_Bucket is measured in the Rows and average of Call_Seconds ismeasured in theValues section. And we put Call_Status in the Filters section.
- The total average of call time duration which are answered by the agents is 198.6seconds.
- The average call time duration for all incoming calls received by agents isthe highest inbetween 10 am to 11 am and from 7 pm to 8 pm
- The average call time duration for all incoming calls received by agentsis the least inbetween 12 noon to 1 pm.

b. Show the total volume/ number of calls coming in via charts/ graphs [Number of calls v/s Time]. You can select time in a bucket form (i.e. 1-2, 2-3, …..)



- We plotted Time_Bucket in the rows and took Count of Customer_Phone_No and Count ofTime in the Values section.
- We measured Count of Time as the percentage of Column Total.
- The customers call the most in between 11 am to 12 noon.
- The customers call the least in between 8 pm to 9 pm.

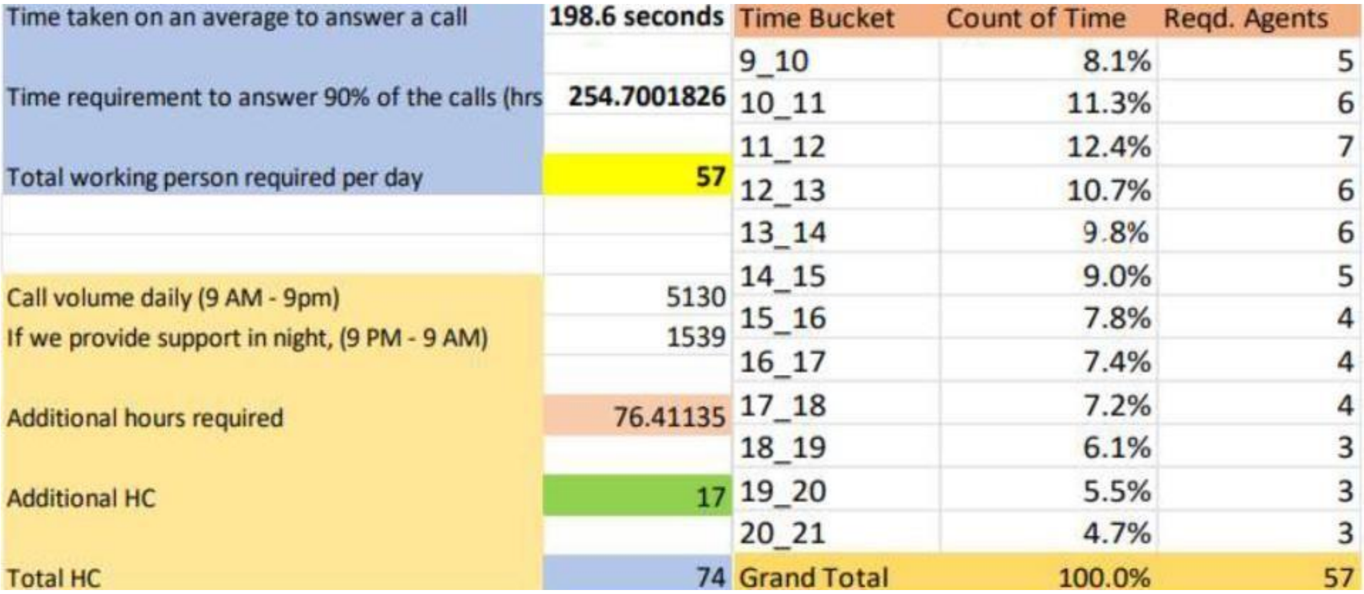
- **Assumption:** An agent work for 6 days a week; On an average total unplannedleaves per agent is 4 days a month; An agent total working hrs is 9 Hrs out of which 1.5 Hrs goes intolunch and snacks in the office. On average an agent occupied for 60% of his total actual working Hrs (i.e. 60% of 7.5 Hrs) on call with customers/ users. Total days in a month is 30days.

Agents working hour	**9**
Agents on-floor work hour	**7.5**
Working Days	**6**
Out of 28 days, an agent works	**24**
Unplanned leave days	**4**
Work days per month	**20**
Days an agent work in a week	**5**
Actual working hours	**60%**
Total time spent on call	**4.5**

- ***Note:*** *For easy calculation, we assumed there are 28 days in a month.*

c. As we can see current abandon rate is approximately 30%.

Propose a manpower plan required during each time bucket [between 9am to 9pm] to reduce the abandon rate to 10%. (i.e. We have to calculate minimum number of agents required in each time bucketso that at least 90 calls should be answered outof 100.)



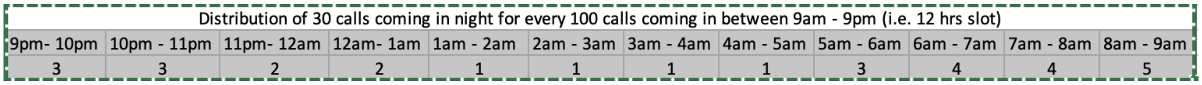
Time taken on an average to answer a call	198.6 seconds
Time requirement to answer 90% of the calls (hrs	254.7001826
Total working person required per day	57
Call volume daily (9 AM - 9pm)	5130
If we provide support in night, (9 PM - 9 AM)	1539
Additional hours required	76.41135
Additional HC	17
Total HC	74

Time Bucket	Count of Time	Reqd. Agents
9_10	8.1%	5
10_11	11.3%	6
11_12	12.4%	7
12_13	10.7%	6
13_14	9.8%	6
14_15	9.0%	5
15_16	7.8%	4
16_17	7.4%	4
17_18	7.2%	4
18_19	6.1%	3
19_20	5.5%	3
20_21	4.7%	3
Grand Total	100.0%	57

- First, we created pivot table. Date & Time is dragged down to Rows, Call Status to Columns,while taking count Call Duration in the Values section.
- Then, we calculated the average of abandon, answered and transfer byusing the averageexcel formula.
- 29% of the calls are abandoned, 1% is transferred, while 70% of the calls are answered in theday time.
- Total agents required to answer the 90% of the calls per day is 57.
- The minimum number of agents required for each time bucket is calculated by57 * count oftime (calculated in the 2nd question).

d. Let’s say customers also call this ABC insurance company in night

but didn't get answer asthere are no agents to answer, this creates a bad customer experience for this Insurance company. Suppose every 100 calls that customer made during 9 Am to 9 Pm,customer also made 30 calls in night between interval [9 Pm to 9 Am] and distribution of those 30 calls are as follows:

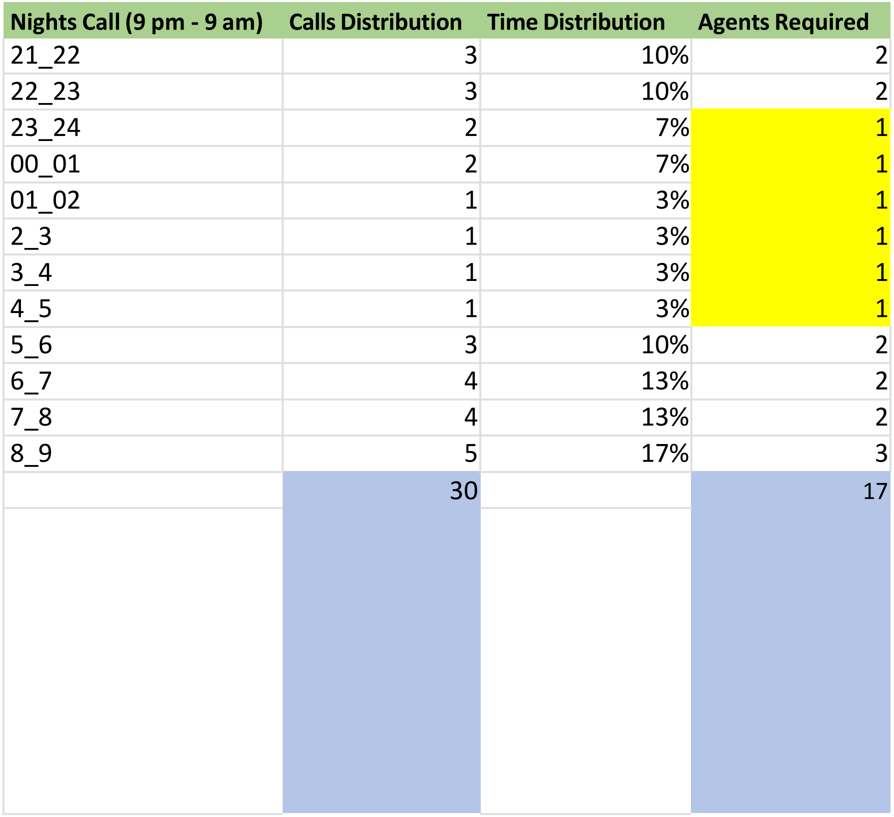


Distribution of 30 calls coming in night for every 100 calls coming in between 9am - 9pm (i.e. 12 hrs slot)											
9pm- 10pm	10pm - 11pm	11pm- 12am	12am- 1am	1am - 2am	2am - 3am	3am - 4am	4am - 5am	5am - 6am	6am - 7am	7am - 8am	8am - 9am
3	3	2	2	1	1	1	1	3	4	4	5

Now propose a manpower plan required during each time bucket in a day. MaximumAbandon rate assumption would be same 10%.

- We first calculated the Time Distribution by dividing each calls distribution bytotal calls i.e.30.
- The number of agents required for each time bucket is calculated by 17 * TimeDistribution.

- **Note:** 17 is calculated above by dividing the additional hours required toanswer the nightcalls by 4.5 (actual working hours of agents).



Nights Call (9 pm - 9 am)	Calls Distribution	Time Distribution	Agents Required
21_22	3	10%	2
22_23	3	10%	2
23_24	2	7%	1
00_01	2	7%	1
01_02	1	3%	1
2_3	1	3%	1
3_4	1	3%	1
4_5	1	3%	1
5_6	3	10%	2
6_7	4	13%	2
7_8	4	13%	2
8_9	5	17%	3
	30		17

Results of Insights

- The customers call the least in the evening. So, the company can reducethe number ofagents at that time for answering the calls.
- The company can hire 17 customer support agents for the night shift work.
- The company can shift some of the day workers for the night shift.
- The employees who are working 9 am to 9 pm. The manager can change someof the workers shift from 5 am to 2 pm and some workers from 2 pm to 11 pmto get the most callsanswered.
- The company can make the employers divide into 3 parts too, so that theagents are alwaysavailable 24/7.
- We found there were few outliers in the data. And if we have removed thatoutliers, thenthe answers would have been different.

Result:

- I learned how an analyst can make an impact in customer service department.
- I learned how a company deals with the customers to give them the most satisfaction.
- I got to know about the IVR Duration, which is an AI tool, who answer thecalls to get toknow the customer exact question and then transfer it to the right agent to get the customer’s queries get answered.
- This project was easy to get the answers as the data provided by the team have already calculated the time bucket and converted the calls duration intoseconds, so we do not hadto spend time on it to calculate.
- I learned about the behavioural analytics.

LEARNINGS

From these projects given by trainity, I have developed various skills in SQL and Excel. I have also learnt how to present these projects in a good fashion. I have learnt about visualization how it makes the presenter present its points in an easy fashion. I have also enjoyed myself creating these projects. With these projects I have also gained an experience how to work in a data analyst company going forward.

Lastly I want to say that I am ready and greatly excited to work as an intern in any Data Analysis Company to gain industrial exposure.