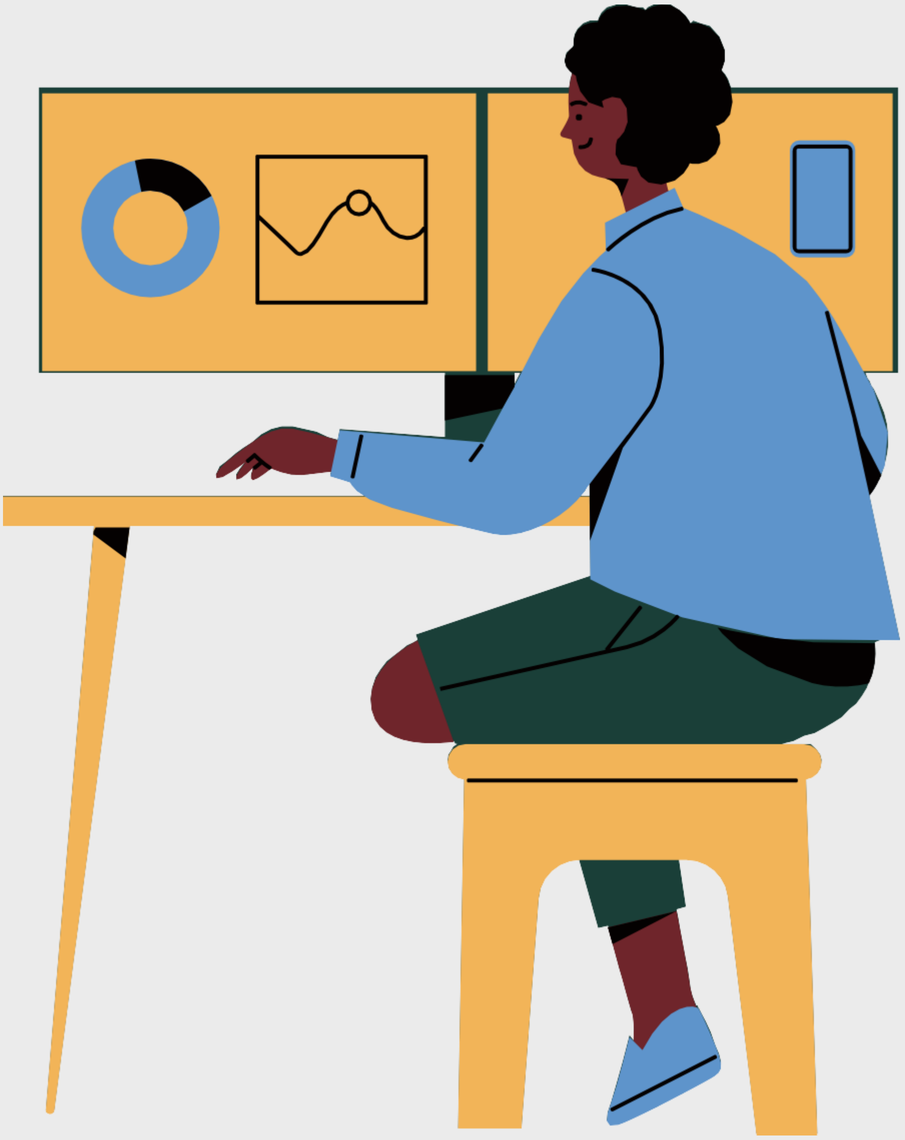
DATA ANALYSIS PORTFOLIO

BY: Tarang Gourshettiwar



Professional Background

I am a Civil Engineer, graduated in 2022 from Amravati University. I have secured a CGPA of 9.54 in my BE.

My technical skillsets include

- MySQL.
- MS: Excel.
- Python.
- Tableau.

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.

TABLE OF CONTENTS

INSTAGRAM USER ANALYTICS

Project Description:

The main aim of this project is to gain detailed insights for the Marketing Team and Investors. These insights are then used by teams across the business to launch a new marketing campaign, decide on features to build for an app, track the success of the app by measuring user engagement and improve the experience altogether while helping the business grow.

The insights required for marketing team include finding most loyal users, reminding inactive users to start posting, finding the most liked photo, to find top 5 hashtags used most commonly, to find what day of week most users register and when can the Ads be launched.

The insights required for investors are to find fake accounts and also check whether users are still as active as before.

Approach:

I've approached this problem statements one by one and wrote queries that can help me find the solution required for particular problem statement.

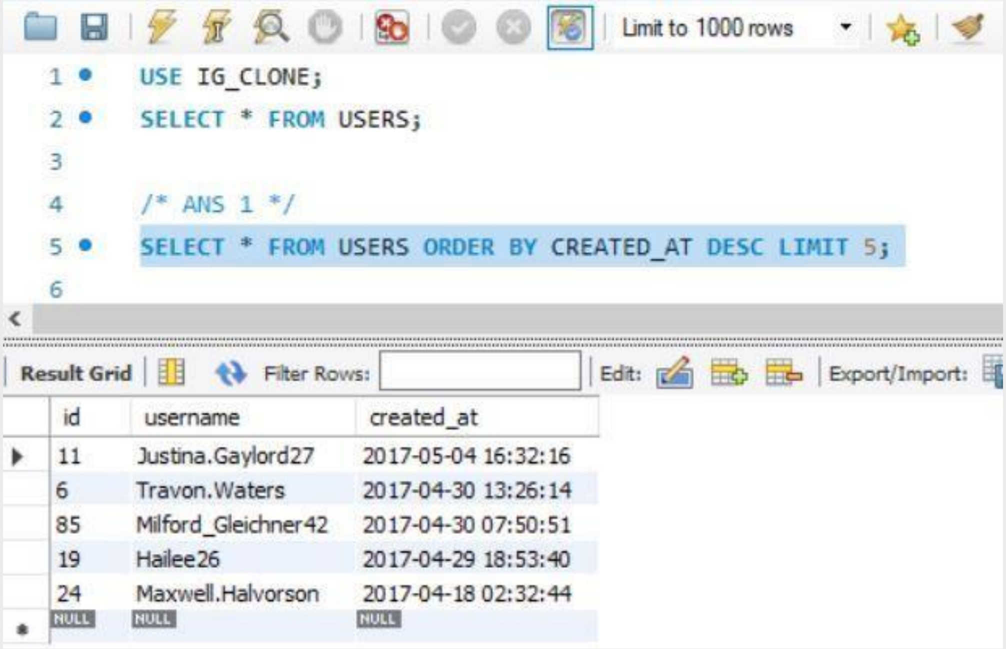
Tech-Stack Used:

In this project, MySQL version 8.0CE was used for accessing the datasets and writing queries.

Insights:

A] Marketing:

1. 5 oldest users of Instagram from the data are:

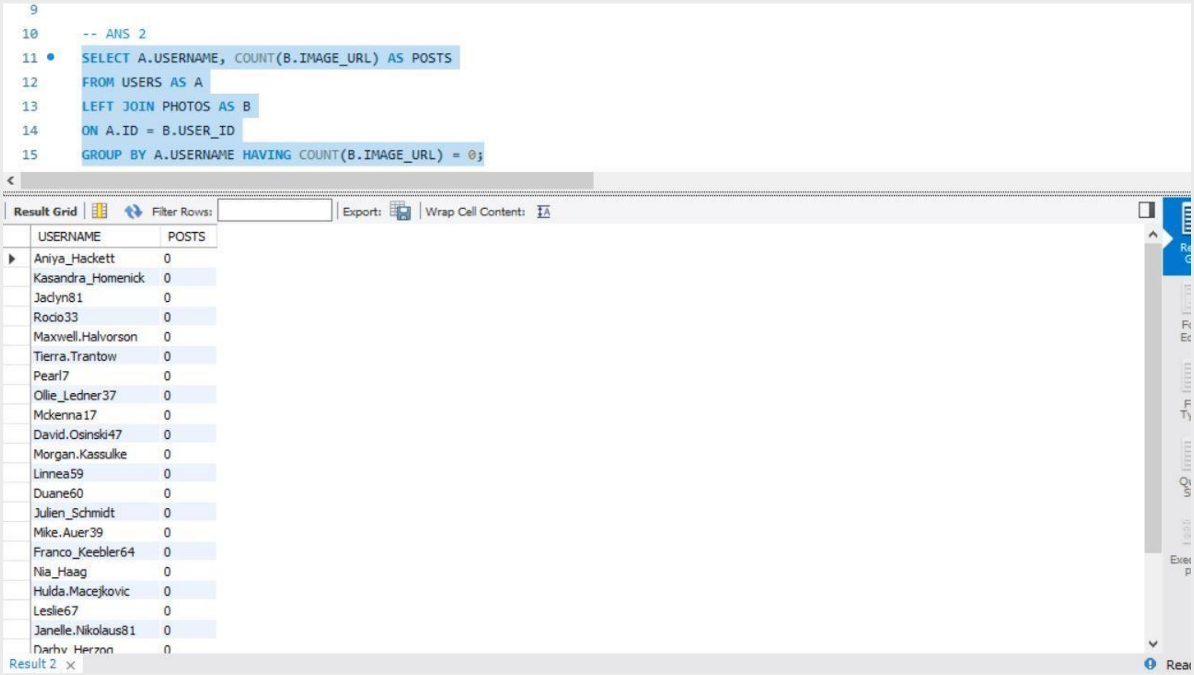


```sql
USE IG_CLONE;
SELECT * FROM USERS;

/* ANS 1 */
SELECT * FROM USERS ORDER BY CREATED_AT DESC LIMIT 5;
```

id	username	created_at
11	Justina.Gaylord27	2017-05-04 16:32:16
6	Travon.Waters	2017-04-30 13:26:14
85	Milford_Gleichner42	2017-04-30 07:50:51
19	Hailee26	2017-04-29 18:53:40
24	Maxwell.Halvorson	2017-04-18 02:32:44
NULL	NULL	NULL

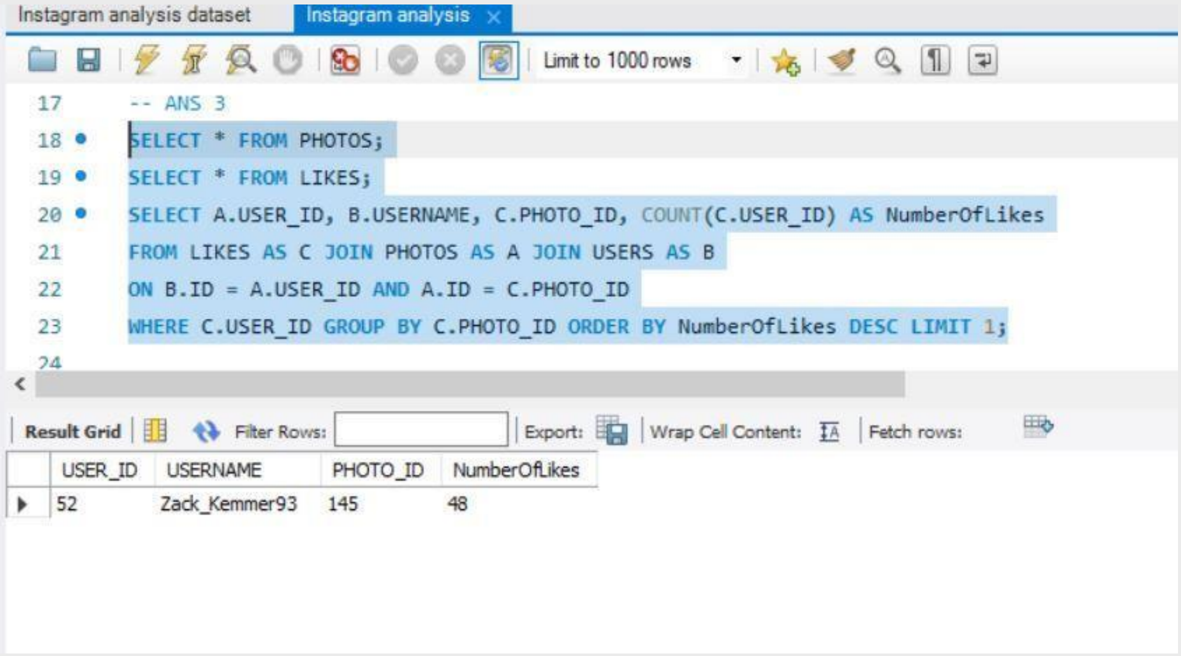
2. Users who have never posted a single photo:



```sql
-- ANS 2
SELECT A.USERNAME, COUNT(B.IMAGE_URL) AS POSTS
FROM USERS AS A
LEFT JOIN PHOTOS AS B
ON A.ID = B.USER_ID
GROUP BY A.USERNAME HAVING COUNT(B.IMAGE_URL) = 0;
```

USERNAME	POSTS
Aniya_Hackett	0
Kasandra_Homenick	0
Jaclyn81	0
Rocio33	0
Maxwell.Halvorson	0
Tierra.Trantow	0
Pearl7	0
Ollie_Ledner37	0
Mckenna17	0
David.Osinski47	0
Morgan.Kassulke	0
Linnea59	0
Duane60	0
Julien_Schmidt	0
Mike.Auer39	0
Franco_Keebler64	0
Nia_Haag	0
Hulda.Macejkovic	0
Leslie67	0
Janelle.Nikolaus81	0
Darby_Herzog	0

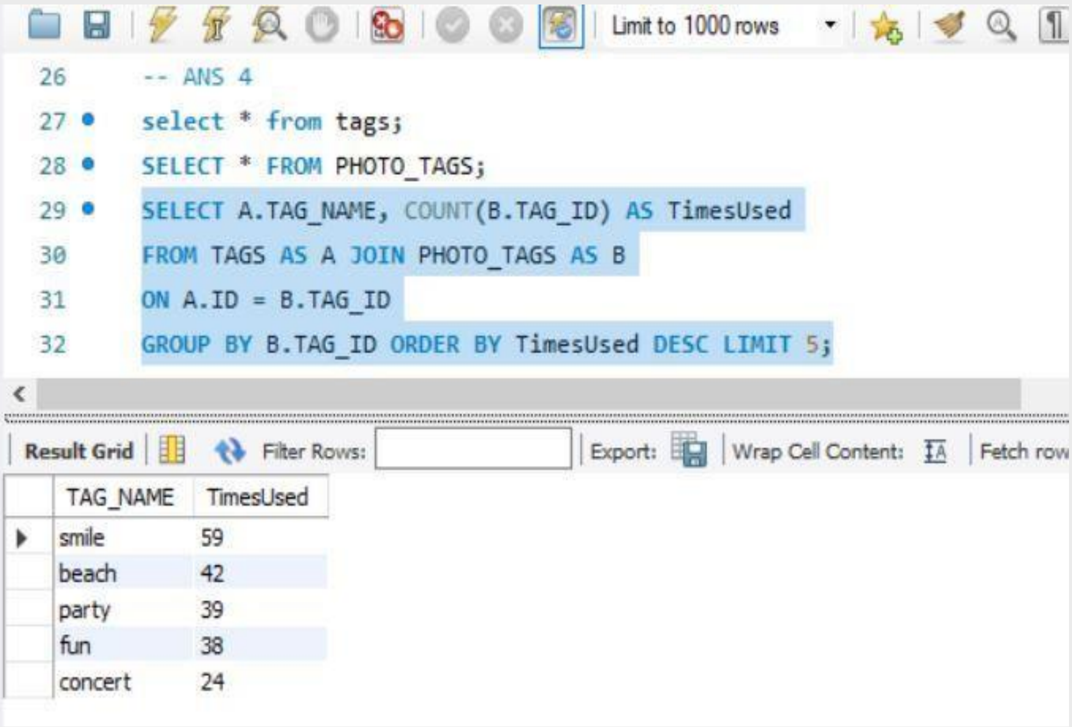
3. Winner of the contest with most likes on a post is:



```sql
-- ANS 3
SELECT * FROM PHOTOS;
SELECT * FROM LIKES;
SELECT A.USER_ID, B.USERNAME, C.PHOTO_ID, COUNT(C.USER_ID) AS NumberOfLikes
FROM LIKES AS C JOIN PHOTOS AS A JOIN USERS AS B
ON B.ID = A.USER_ID AND A.ID = C.PHOTO_ID
WHERE C.USER_ID GROUP BY C.PHOTO_ID ORDER BY NumberOfLikes DESC LIMIT 1;
```

USER_ID	USERNAME	PHOTO_ID	NumberOfLikes
52	Zack_Kemmer93	145	48

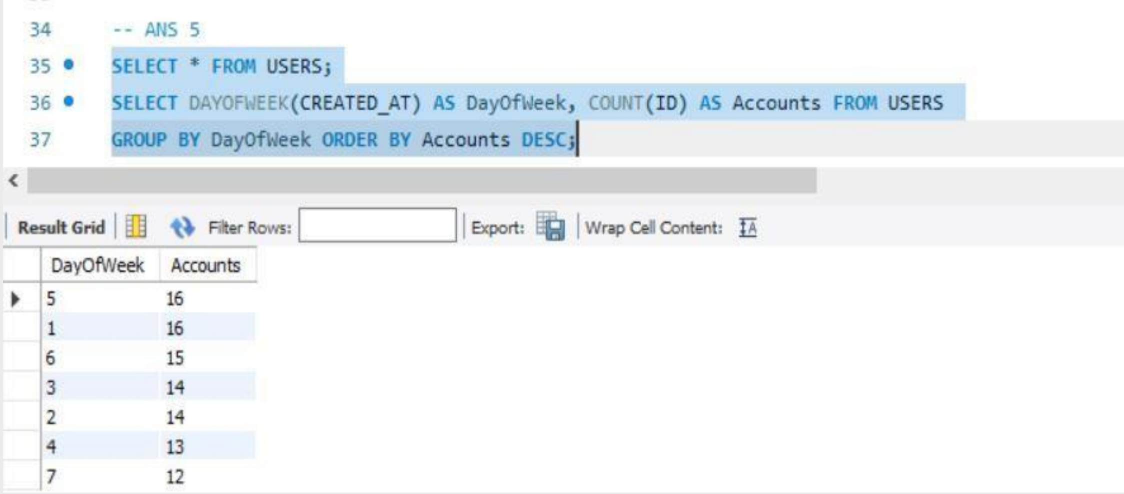
4. Following are top 5 most used hashtags:



```sql
-- ANS 4
select * from tags;
SELECT * FROM PHOTO_TAGS;
SELECT A.TAG_NAME, COUNT(B.TAG_ID) AS TimesUsed
FROM TAGS AS A JOIN PHOTO_TAGS AS B
ON A.ID = B.TAG_ID
GROUP BY B.TAG_ID ORDER BY TimesUsed DESC LIMIT 5;
```

TAG_NAME	TimesUsed
smile	59
beach	42
party	39
fun	38
concert	24

5. Following are the days of the week when most of the users register.

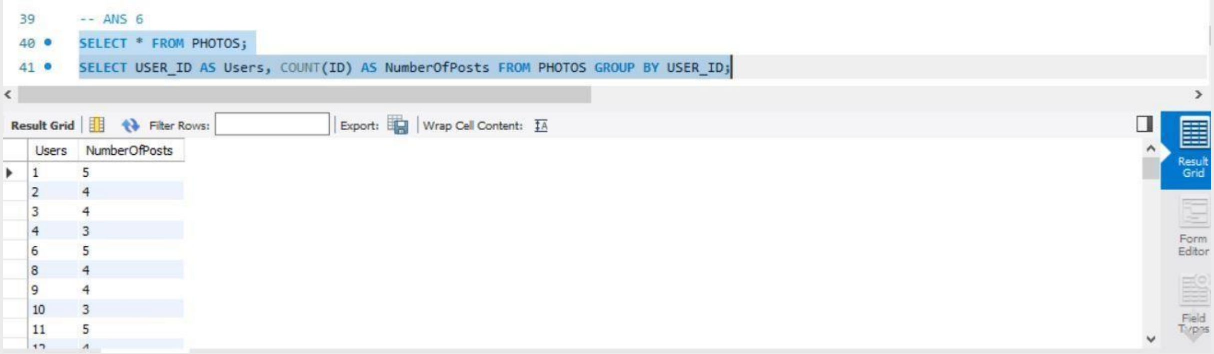


```
-- ANS 5
SELECT * FROM USERS;
SELECT DAYOFWEEK(CREATED_AT) AS DayOfWeek, COUNT(ID) AS Accounts FROM USERS
GROUP BY DayOfWeek ORDER BY Accounts DESC;
```

DayOfWeek	Accounts
5	16
1	16
6	15
3	14
2	14
4	13
7	12

B] Investor Metrics:

1. Number of posts per user



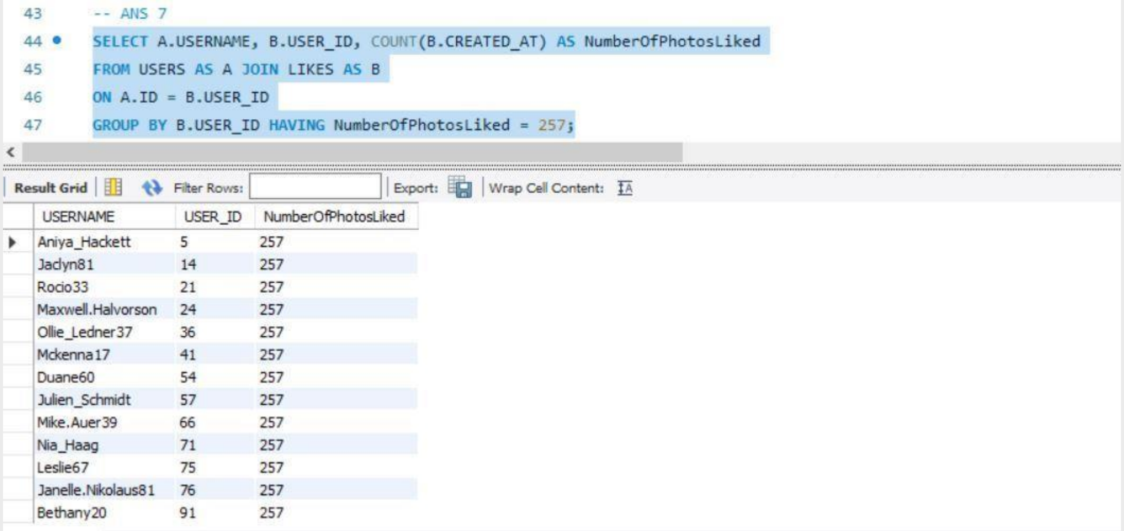
```
-- ANS 6
SELECT * FROM PHOTOS;
SELECT USER_ID AS Users, COUNT(ID) AS NumberOfPosts FROM PHOTOS GROUP BY USER_ID;
```

Users	NumberOfPosts
1	5
2	4
3	4
4	3
6	5
8	4
9	4
10	3
11	5

Total number of users = 100.

Total number of photos = 257

2. Following are the fake accounts that liked all the posts in Instagra



```
-- ANS 7
SELECT A.USERNAME, B.USER_ID, COUNT(B.CREATED_AT) AS NumberOfPhotosLiked
FROM USERS AS A JOIN LIKES AS B
ON A.ID = B.USER_ID
GROUP BY B.USER_ID HAVING NumberOfPhotosLiked = 257;
```

USERNAME	USER_ID	NumberOfPhotosLiked
Aniya_Hackett	5	257
Jaclyn81	14	257
Rocio33	21	257
Maxwell.Halvorson	24	257
Ollie_Ledner37	36	257
Mckenna17	41	257
Duane60	54	257
Julien_Schmidt	57	257
Mike.Auer39	66	257
Nia_Haag	71	257
Leslie67	75	257
Janelle.Nikolaus81	76	257
Bethany20	91	257

OPERATION ANALYTICS AND INVESTIGATING METRIC SPIKE

DESCRIPTION:

The project is based on Operation Analysis which helps to perform end to end operations for growth of the company and also gives insights for the reasons that are responsible for downfall of company’s fortune.

Investigating metric spike is also an important part of operation analytics as it helps you to understand and communicate more with other teams and solve their queries regarding business.

In this project, we gain insights on Number of jobs reviewed over time, finding 7 day rolling average, share of each language for different continents, finding duplicate rows, user engagement, amount of users growing over time, weekly engagement of the users per device, users engaging with email services.

Approach:

First, I imported the csv files in MySQL workbench. Then I’ve approached this problem statements one by one and wrote queries that can help me find the solution required for particular problem statement.

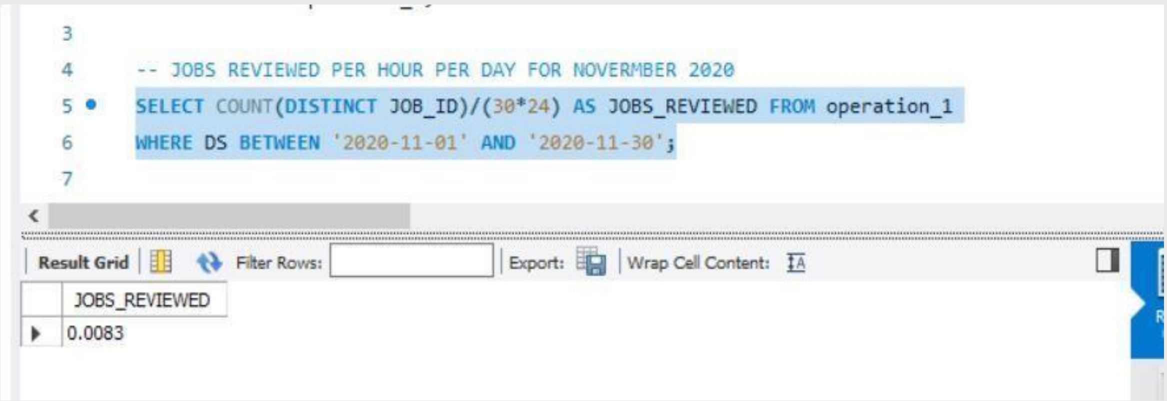
Tech-Stack Used:

In this project, MySQL version 8.0CE was used for accessing the csv files and writing the queries. Including this, Mode.com was used for few queries as the data was humungous and could not be loaded on MySQL.

INSIGHTS:

CASE-STUDY_1:

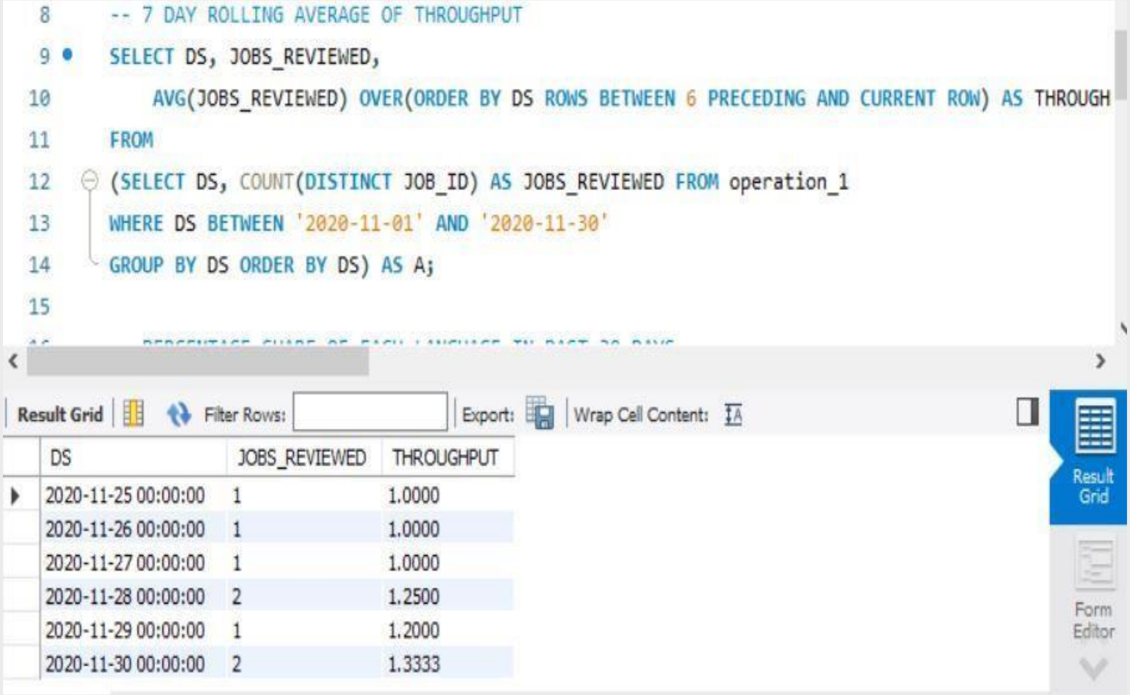
1. Calculate the number of jobs reviewed per hour per day for November 2020



```
-- JOBS REVIEWED PER HOUR PER DAY FOR NOVERMBER 2020
SELECT COUNT(DISTINCT JOB_ID)/(30*24) AS JOBS_REVIEWED FROM operation_1
WHERE DS BETWEEN '2020-11-01' AND '2020-11-30';
```

JOBS_REVIEWED
0.0083

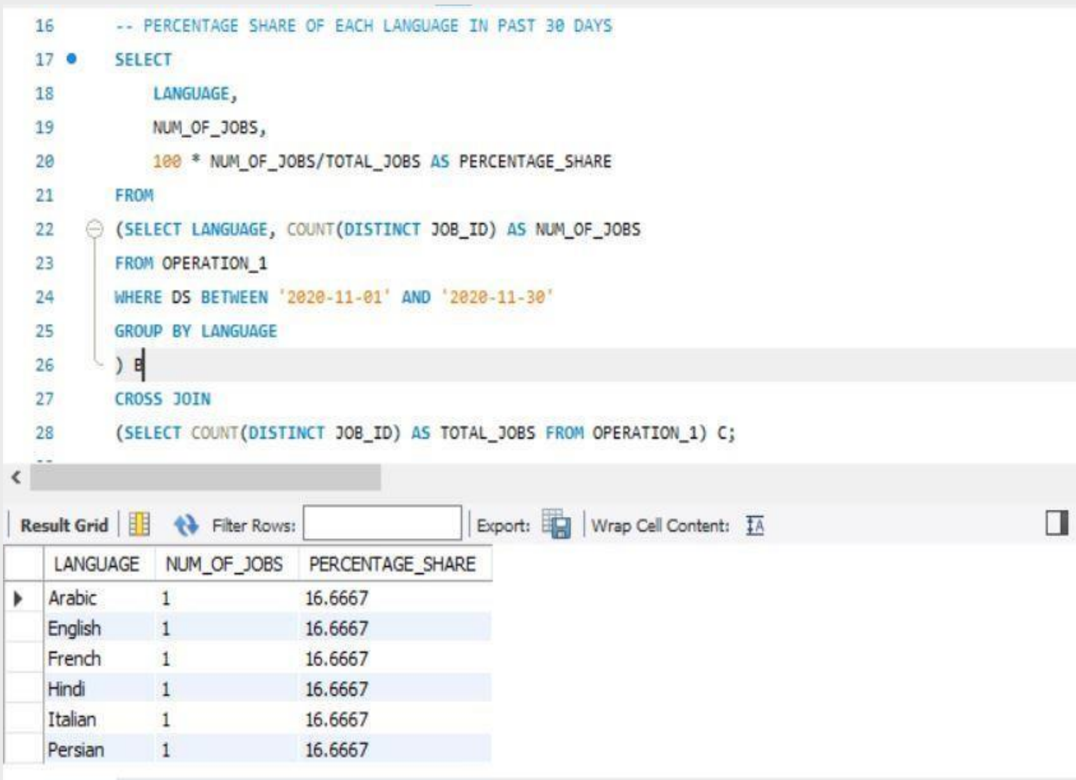
2. 7 day Rolling average of Throughput:



```
-- 7 DAY ROLLING AVERAGE OF THROUGHPUT
SELECT DS, JOBS_REVIEWED,
    AVG(JOBS_REVIEWED) OVER(ORDER BY DS ROWS BETWEEN 6 PRECEDING AND CURRENT ROW) AS THROUGH
FROM
(SELECT DS, COUNT(DISTINCT JOB_ID) AS JOBS_REVIEWED FROM operation_1
WHERE DS BETWEEN '2020-11-01' AND '2020-11-30'
GROUP BY DS ORDER BY DS) AS A;
```

DS	JOBS_REVIEWED	THROUGHPUT
2020-11-25 00:00:00	1	1.0000
2020-11-26 00:00:00	1	1.0000
2020-11-27 00:00:00	1	1.0000
2020-11-28 00:00:00	2	1.2500
2020-11-29 00:00:00	1	1.2000
2020-11-30 00:00:00	2	1.3333

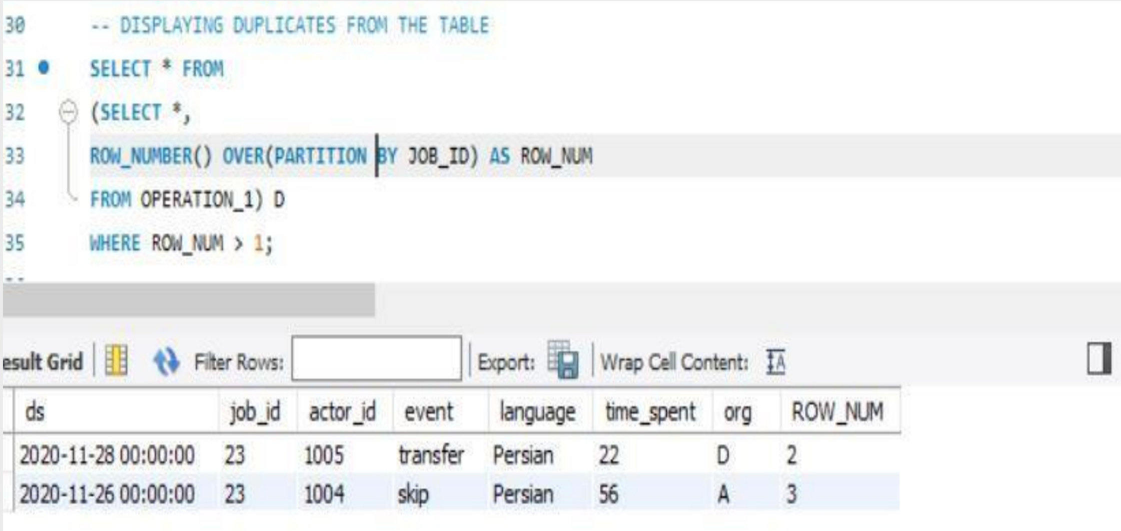
3. Percentage share of each language in past 30 days:



```
-- PERCENTAGE SHARE OF EACH LANGUAGE IN PAST 30 DAYS
SELECT
    LANGUAGE,
    NUM_OF_JOBS,
    100 * NUM_OF_JOBS/TOTAL_JOBS AS PERCENTAGE_SHARE
FROM
(SELECT LANGUAGE, COUNT(DISTINCT JOB_ID) AS NUM_OF_JOBS
FROM OPERATION_1
WHERE DS BETWEEN '2020-11-01' AND '2020-11-30'
GROUP BY LANGUAGE
) B
CROSS JOIN
(SELECT COUNT(DISTINCT JOB_ID) AS TOTAL_JOBS FROM OPERATION_1) C;
```

LANGUAGE	NUM_OF_JOBS	PERCENTAGE_SHARE
Arabic	1	16.6667
English	1	16.6667
French	1	16.6667
Hindi	1	16.6667
Italian	1	16.6667
Persian	1	16.6667

4. Displaying duplicates from the table:

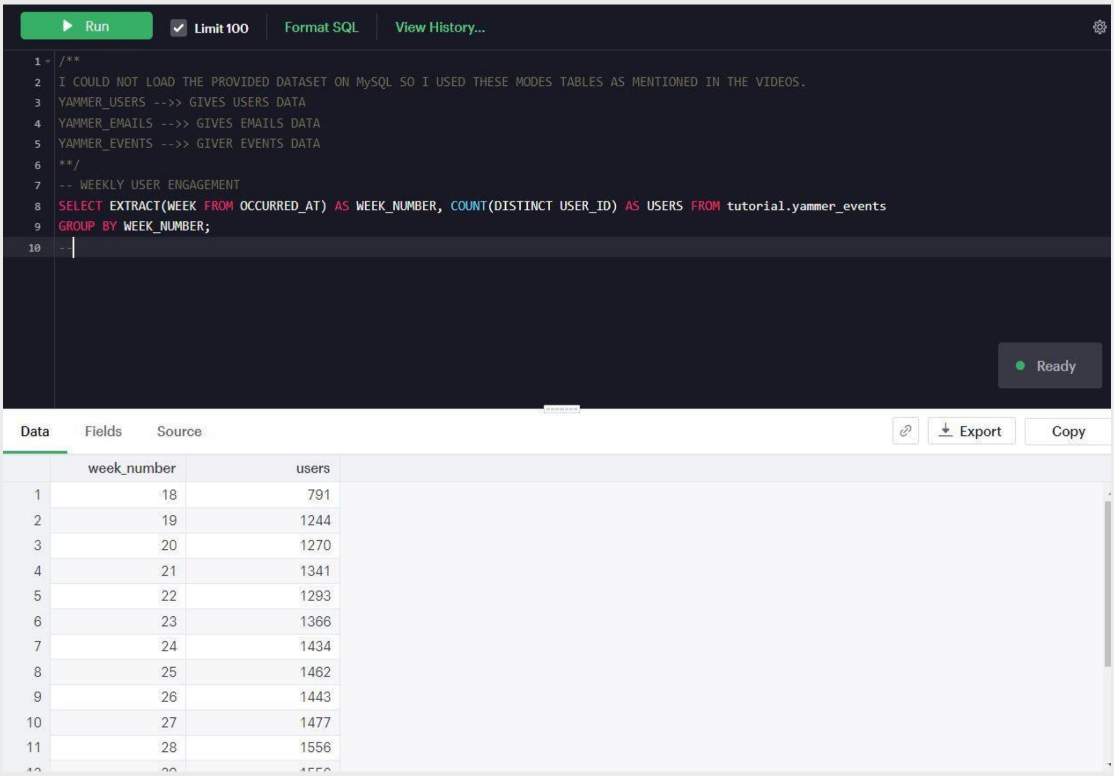


```
-- DISPLAYING DUPLICATES FROM THE TABLE
SELECT * FROM
(SELECT *,
ROW_NUMBER() OVER(PARTITION BY JOB_ID) AS ROW_NUM
FROM OPERATION_1) D
WHERE ROW_NUM > 1;
```

ds	job_id	actor_id	event	language	time_spent	org	ROW_NUM
2020-11-28 00:00:00	23	1005	transfer	Persian	22	D	2
2020-11-26 00:00:00	23	1004	skip	Persian	56	A	3

CASE-STUDY_2:

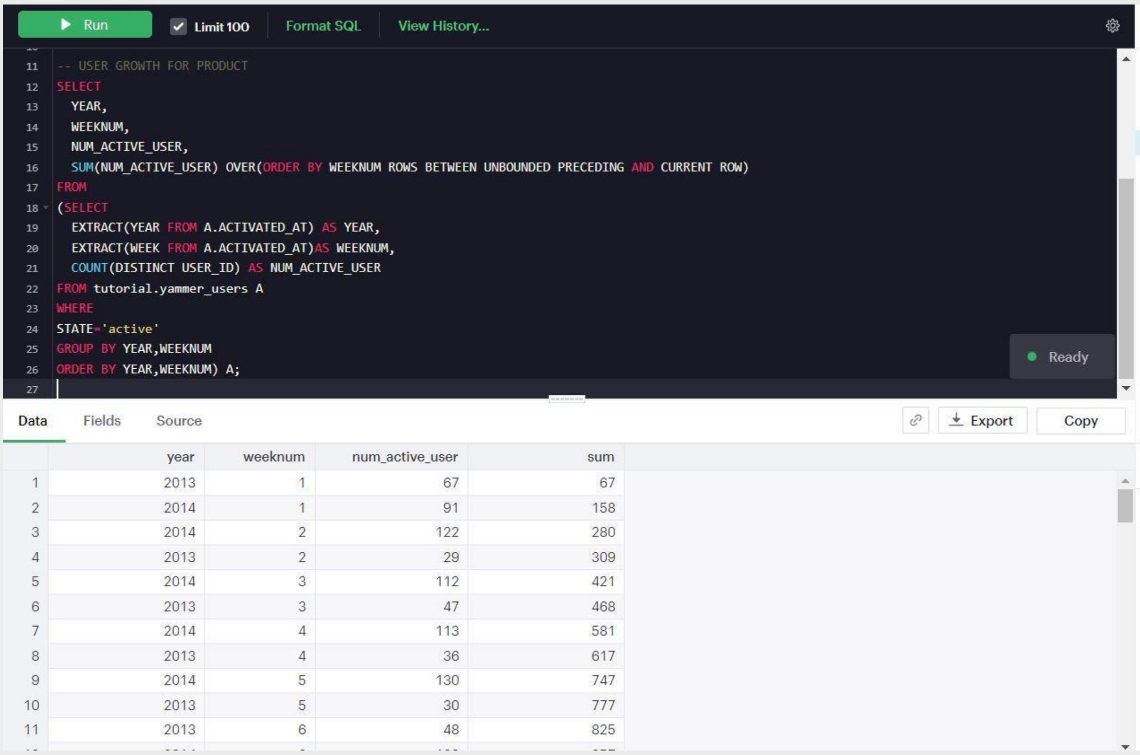
1. Measuring activeness of user:



```sql
/**
I COULD NOT LOAD THE PROVIDED DATASET ON MySQL SO I USED THESE MODES TABLES AS MENTIONED IN THE VIDEOS.
YAMMER_USERS -->> GIVES USERS DATA
YAMMER_EMAILS -->> GIVES EMAILS DATA
YAMMER_EVENTS -->> GIVER EVENTS DATA
**/
-- WEEKLY USER ENGAGEMENT
SELECT EXTRACT(WEEK FROM OCCURRED_AT) AS WEEK_NUMBER, COUNT(DISTINCT USER_ID) AS USERS FROM tutorial.yammer_events
GROUP BY WEEK_NUMBER;
--
```

	week_number	users
1	18	791
2	19	1244
3	20	1270
4	21	1341
5	22	1293
6	23	1366
7	24	1434
8	25	1462
9	26	1443
10	27	1477
11	28	1556

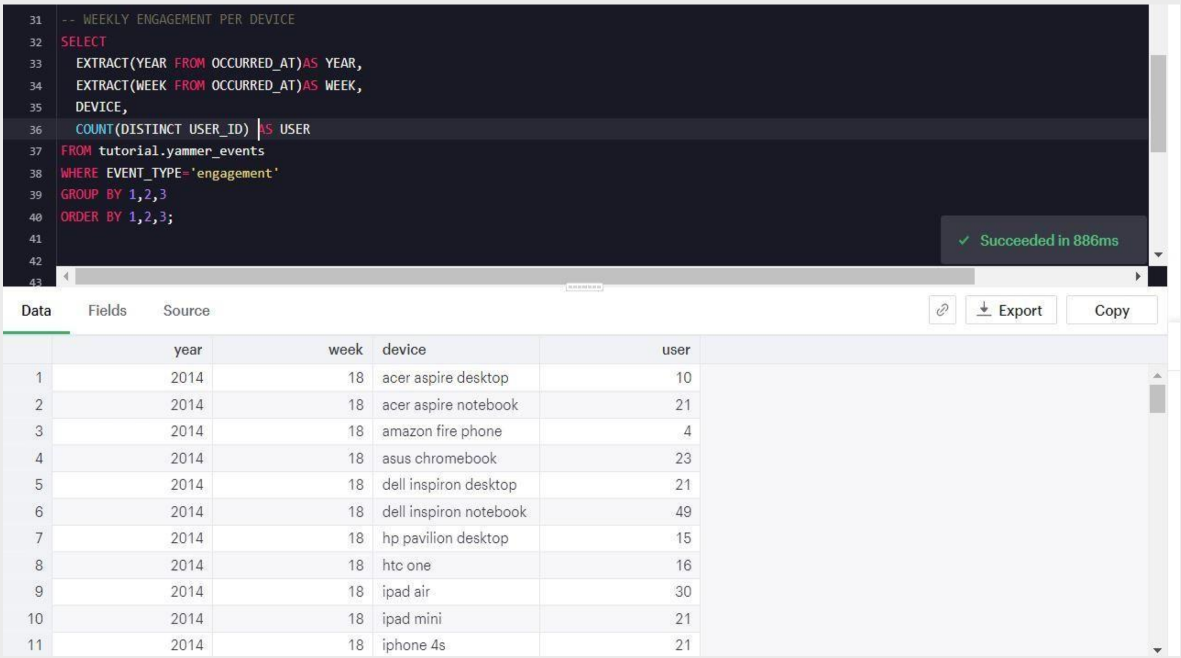
2. User growth for product:



```sql
-- USER GROWTH FOR PRODUCT
SELECT
  YEAR,
  WEEKNUM,
  NUM_ACTIVE_USER,
  SUM(NUM_ACTIVE_USER) OVER(ORDER BY WEEKNUM ROWS BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW)
FROM
(SELECT
  EXTRACT(YEAR FROM A.ACTIVATED_AT) AS YEAR,
  EXTRACT(WEEK FROM A.ACTIVATED_AT)AS WEEKNUM,
  COUNT(DISTINCT USER_ID) AS NUM_ACTIVE_USER
FROM tutorial.yammer_users A
WHERE
STATE='active'
GROUP BY YEAR,WEEKNUM
ORDER BY YEAR,WEEKNUM) A;
```

	year	weeknum	num_active_user	sum
1	2013	1	67	67
2	2014	1	91	158
3	2014	2	122	280
4	2013	2	29	309
5	2014	3	112	421
6	2013	3	47	468
7	2014	4	113	581
8	2013	4	36	617
9	2014	5	130	747
10	2013	5	30	777
11	2013	6	48	825

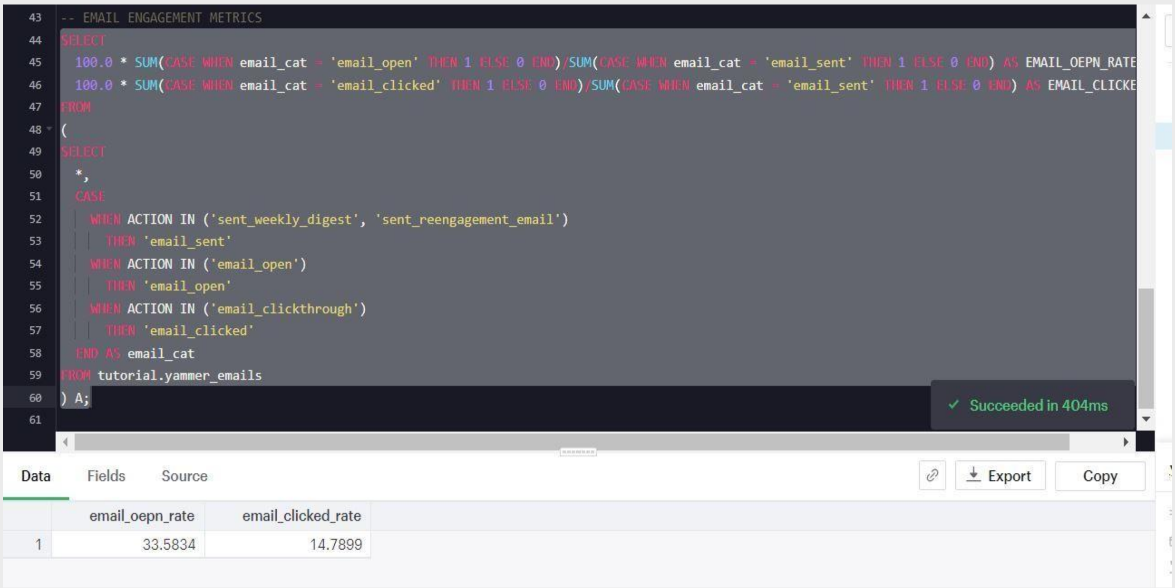
3. Weekly engagement per device:



```sql
-- WEEKLY ENGAGEMENT PER DEVICE
SELECT
  EXTRACT(YEAR FROM OCCURRED_AT)AS YEAR,
  EXTRACT(WEEK FROM OCCURRED_AT)AS WEEK,
  DEVICE,
  COUNT(DISTINCT USER_ID) AS USER
FROM tutorial.yammer_events
WHERE EVENT_TYPE='engagement'
GROUP BY 1,2,3
ORDER BY 1,2,3;
```

	year	week	device	user
1	2014	18	acer aspire desktop	10
2	2014	18	acer aspire notebook	21
3	2014	18	amazon fire phone	4
4	2014	18	asus chromebook	23
5	2014	18	dell inspiron desktop	21
6	2014	18	dell inspiron notebook	49
7	2014	18	hp pavilion desktop	15
8	2014	18	htc one	16
9	2014	18	ipad air	30
10	2014	18	ipad mini	21
11	2014	18	iphone 4s	21

4. Email engagement metrics:



```sql
-- EMAIL ENGAGEMENT METRICS
SELECT
  100.0 * SUM(CASE WHEN email_cat = 'email_open' THEN 1 ELSE 0 END)/SUM(CASE WHEN email_cat = 'email_sent' THEN 1 ELSE 0 END) AS EMAIL_OEPN_RATE
  100.0 * SUM(CASE WHEN email_cat = 'email_clicked' THEN 1 ELSE 0 END)/SUM(CASE WHEN email_cat = 'email_sent' THEN 1 ELSE 0 END) AS EMAIL_CLICKE
FROM
(
SELECT
  *,
  CASE
    WHEN ACTION IN ('sent_weekly_digest', 'sent_reengagement_email')
      THEN 'email_sent'
    WHEN ACTION IN ('email_open')
      THEN 'email_open'
    WHEN ACTION IN ('email_clickthrough')
      THEN 'email_clicked'
  END AS email_cat
FROM tutorial.yammer_emails
) A;
```

	email_oepn_rate	email_clicked_rate
1	33.5834	14.7899

HIRING PROCESS ANALYTICS

PROJECT DESCRIPTION:

In this project, a detailed analysis of the company's hiring process was done and insights were found out by working on the dataset provided.

By doing so we get to know various trends in hiring processes such as number of rejections, number of interviews, types of jobs, vacancies, etc. can also be found out.

APPROACH:

The given dataset was first cleaned so that unwanted data was removed and also checked that all the blank spaces were deleted. Basically, the first steps were related to data cleaning.

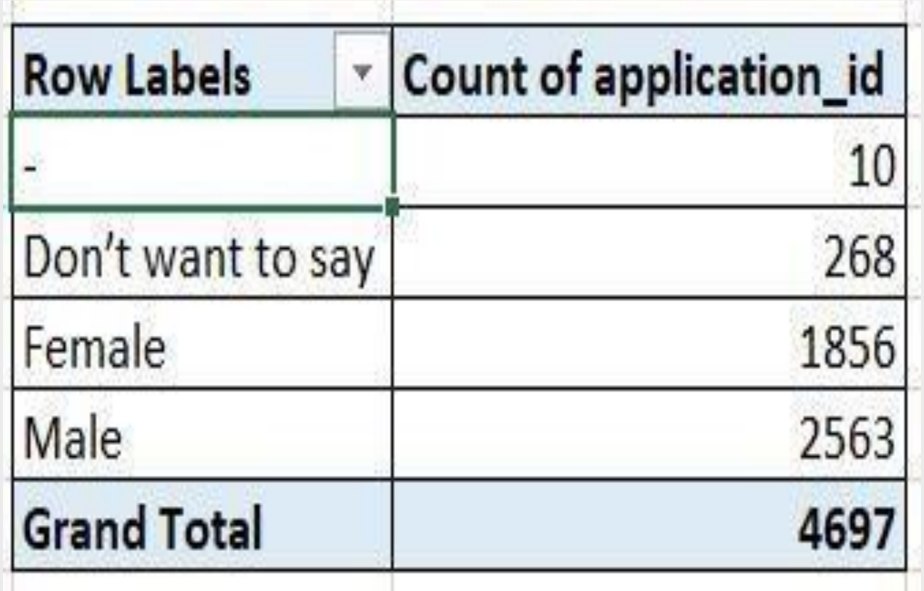
After that, by using excel, all the queries regarding this dataset were solved and the solution to those queries were found out by using various statistical formulas.

TECH-STACK USED:

In this project, MS-Excel 2020 was used for doing all the analysis and finding out solutions.

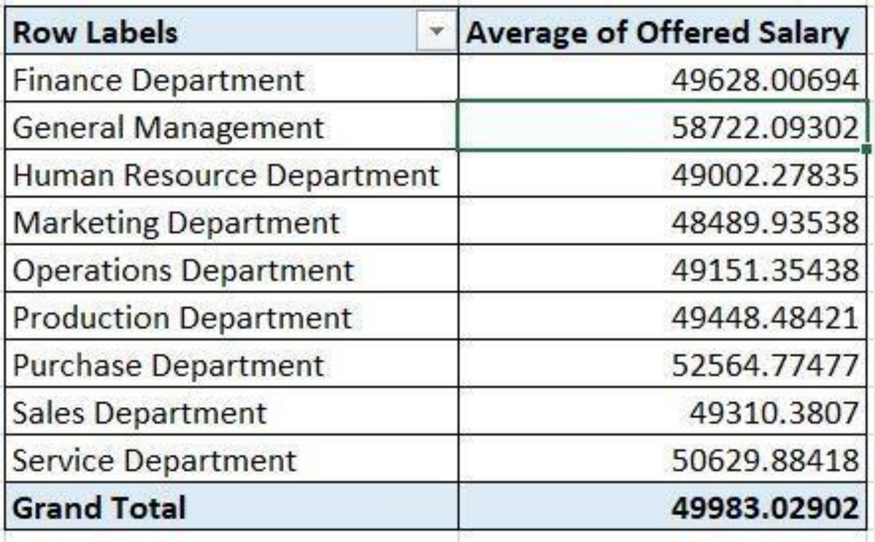
INSIGHTS:

1. How many males and females were hired?



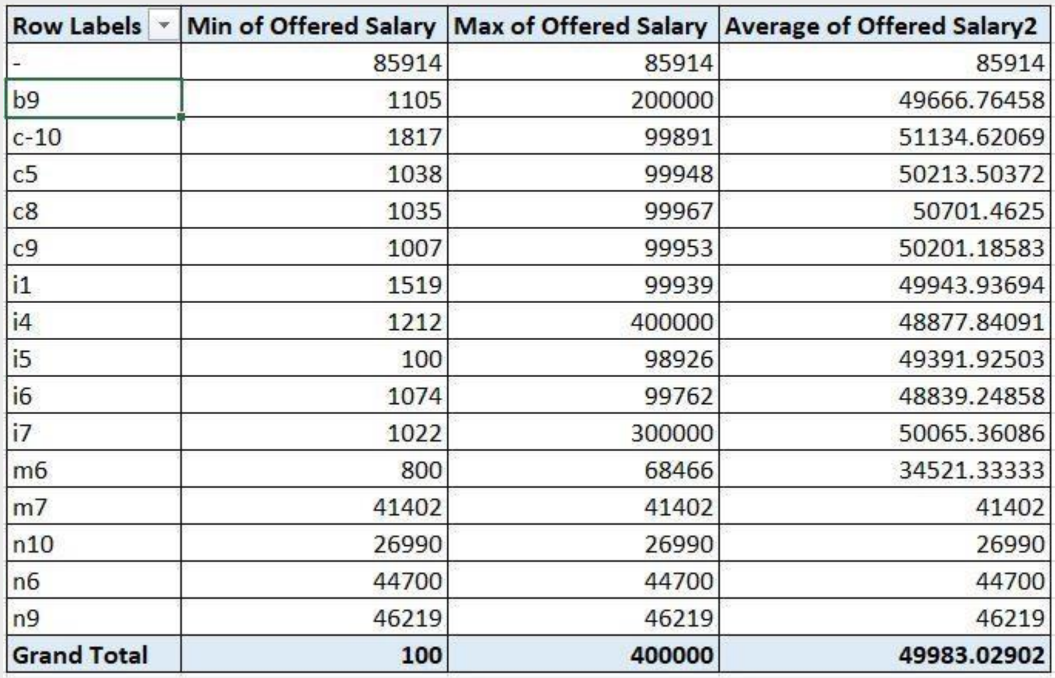
Row Labels	Count of application_id
-	10
Don’t want to say	268
Female	1856
Male	2563
Grand Total	**4697**

2. Average salary offered in the company Department wise?



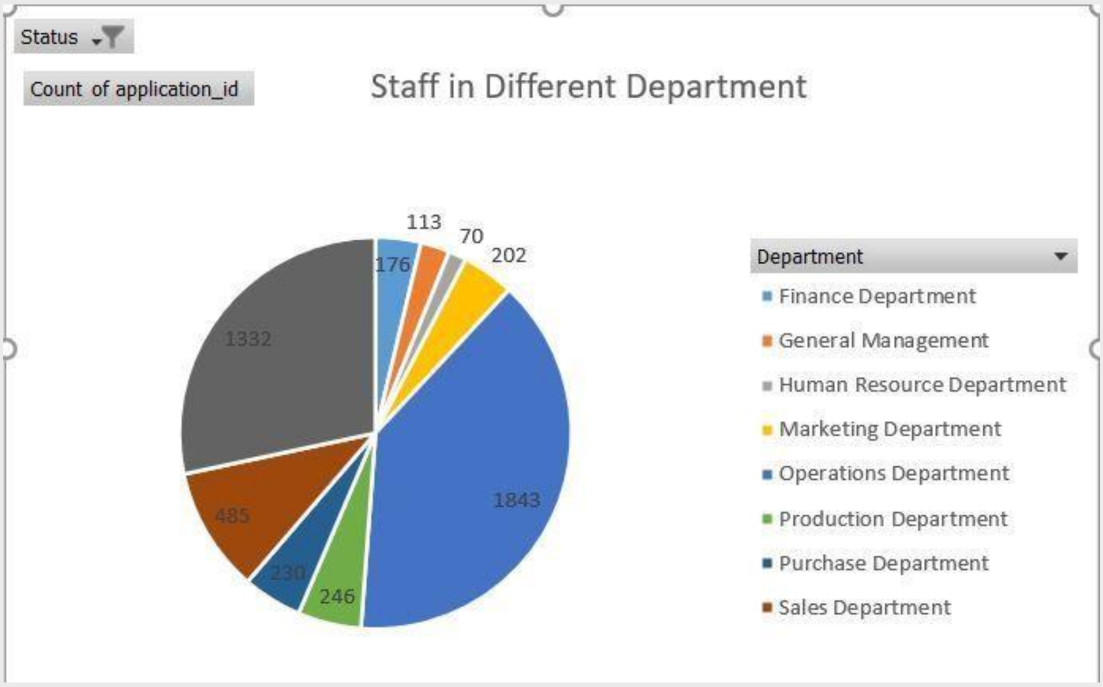
Row Labels	Average of Offered Salary
Finance Department	49628.00694
General Management	58722.09302
Human Resource Department	49002.27835
Marketing Department	48489.93538
Operations Department	49151.35438
Production Department	49448.48421
Purchase Department	52564.77477
Sales Department	49310.3807
Service Department	50629.88418
Grand Total	**49983.02902**

3. Class intervals for salaries in the company?

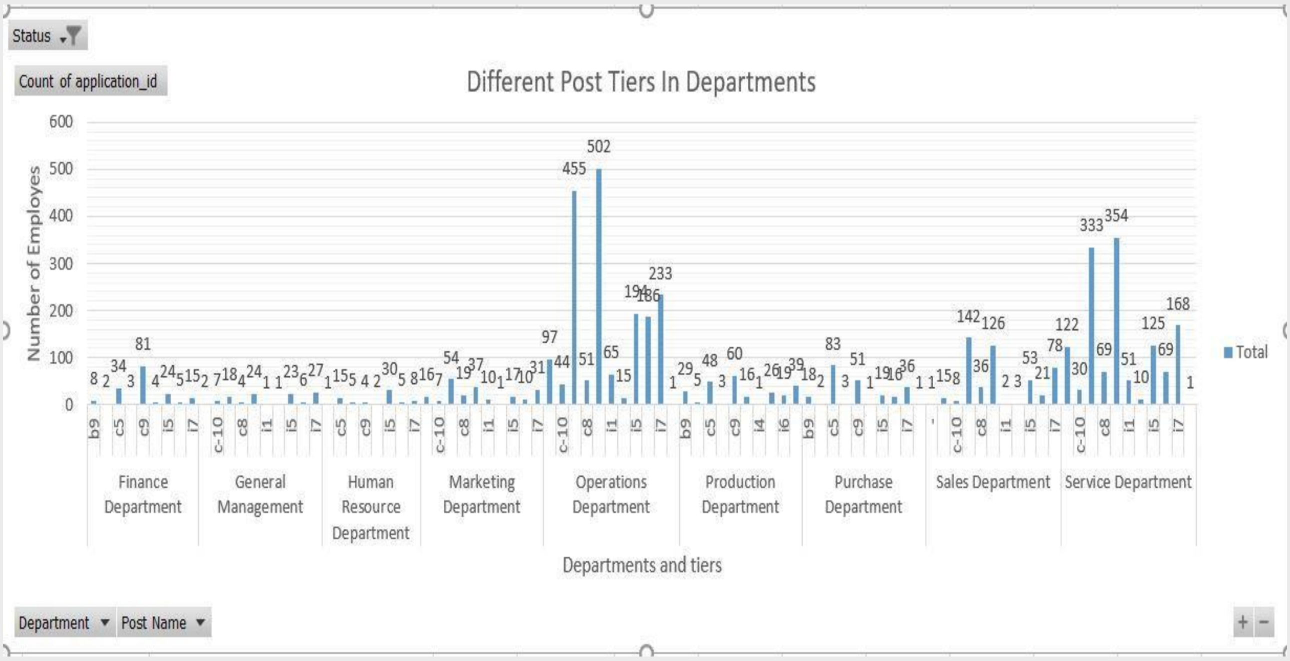


Row Labels	Min of Offered Salary	Max of Offered Salary	Average of Offered Salary2
-	85914	85914	85914
b9	1105	200000	49666.76458
c-10	1817	99891	51134.62069
c5	1038	99948	50213.50372
c8	1035	99967	50701.4625
c9	1007	99953	50201.18583
i1	1519	99939	49943.93694
i4	1212	400000	48877.84091
i5	100	98926	49391.92503
i6	1074	99762	48839.24858
i7	1022	300000	50065.36086
m6	800	68466	34521.33333
m7	41402	41402	41402
n10	26990	26990	26990
n6	44700	44700	44700
n9	46219	46219	46219
Grand Total	**100**	**400000**	**49983.02902**

4. Pie-Chart showing proportions of people working in different departments?



5. Different posts in Departments?



IMDB MOVIE ANALYSIS

PROJECT DESCRIPTION:

In this project, a detail analysis of movies according to the dataset of imdb was done and insights were found out by working on the dataset using Excel.

By doing so, we gained various information regarding profits, highest grossing, best genre, etc.

APPROACH:

The given dataset was first cleaned so that unwanted data was removed and also checked that all the blank spaces were deleted. Basically, the first steps were related to data cleaning.

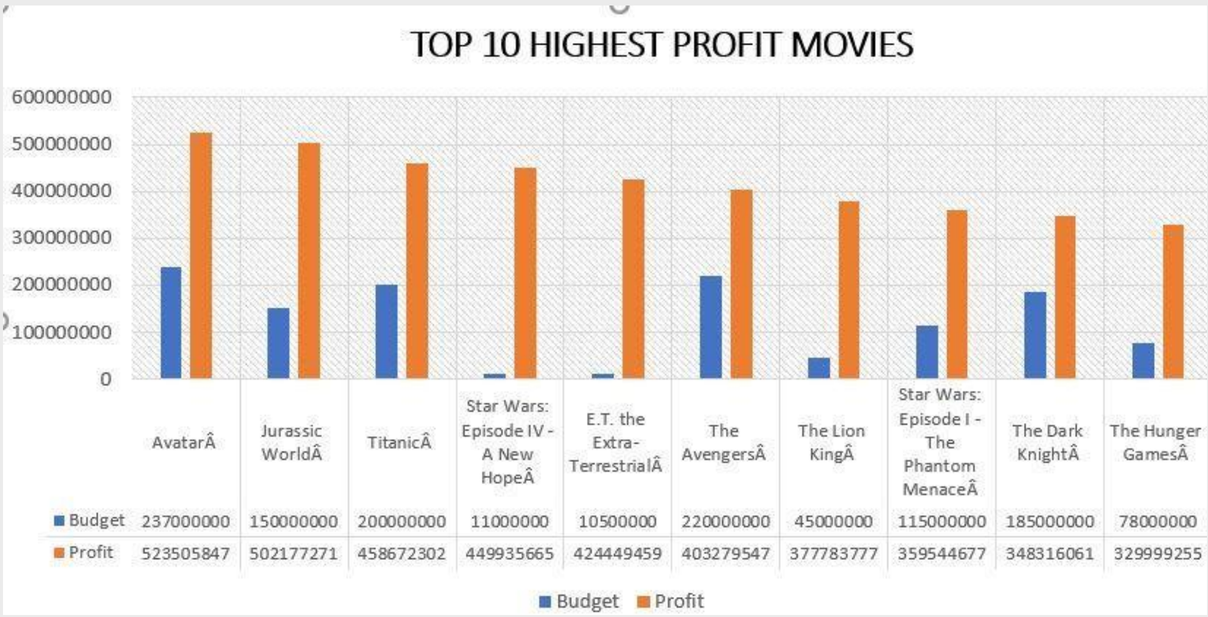
After that, by using excel, pivot tables, all the queries regarding this dataset were solved and the solution to those queries were found out by using various statistical and mathematical formulas.

TECH-STACK USED:

In this project, MS-Excel 2020 was used for doing all the analysis and finding out solutions.

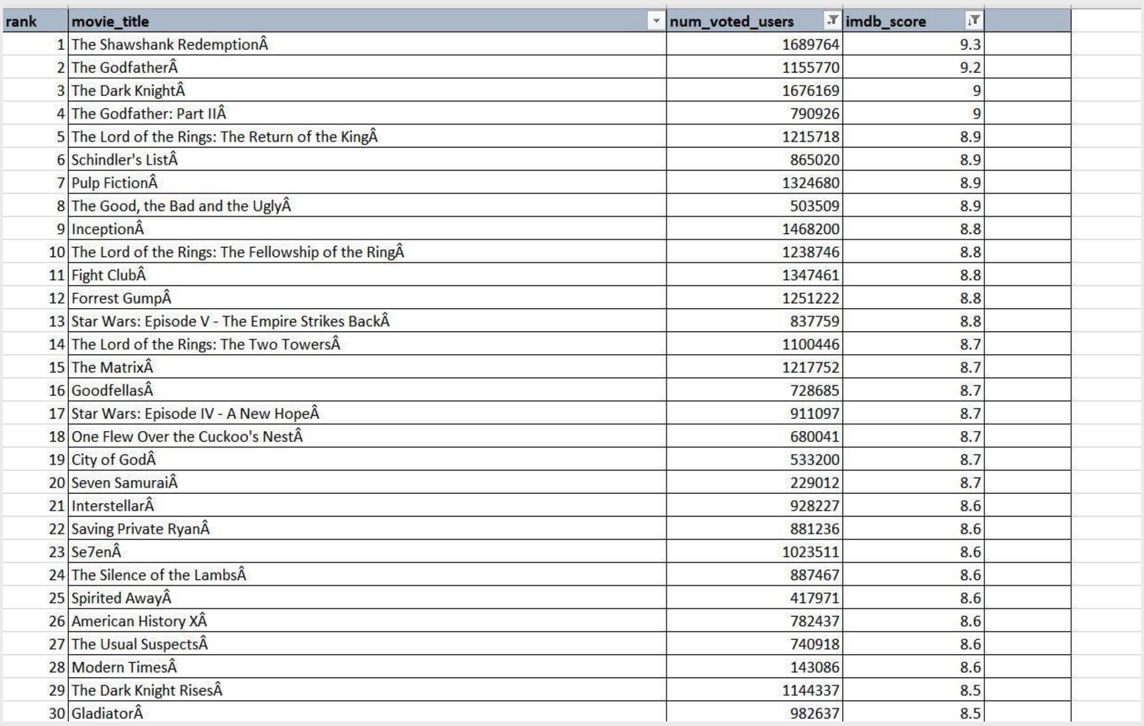
INSIGHTS:

1. Movies with highest profit:

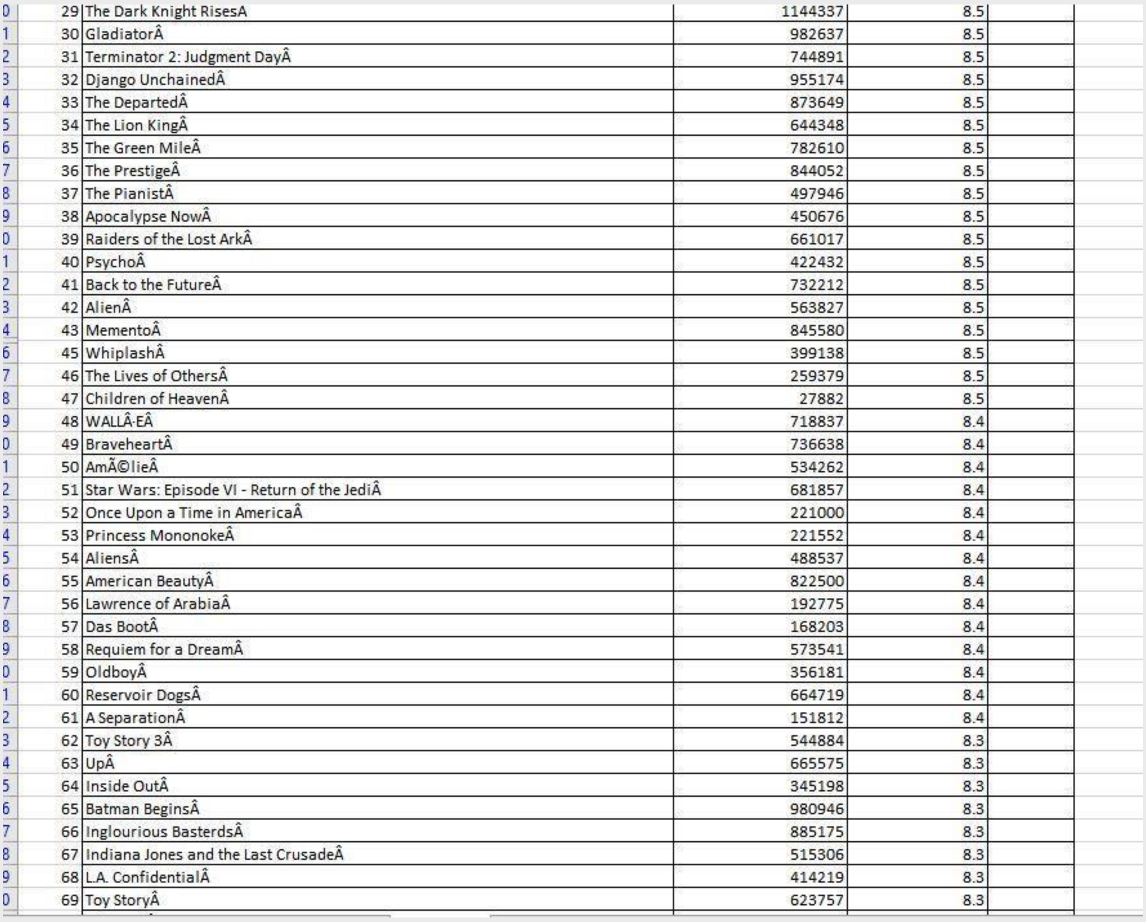


	AvatarÂ	Jurassic WorldÂ	TitanicÂ	Star Wars: Episode IV - A New HopeÂ	E.T. the Extra-TerrestrialÂ	The AvengersÂ	The Lion KingÂ	Star Wars: Episode I - The Phantom MenaceÂ	The Dark KnightÂ	The Hunger GamesÂ
■ Budget	237000000	150000000	200000000	11000000	10500000	220000000	45000000	115000000	185000000	78000000
■ Profit	523505847	502177271	458672302	449935665	424449459	403279547	377783777	359544677	348316061	329999255

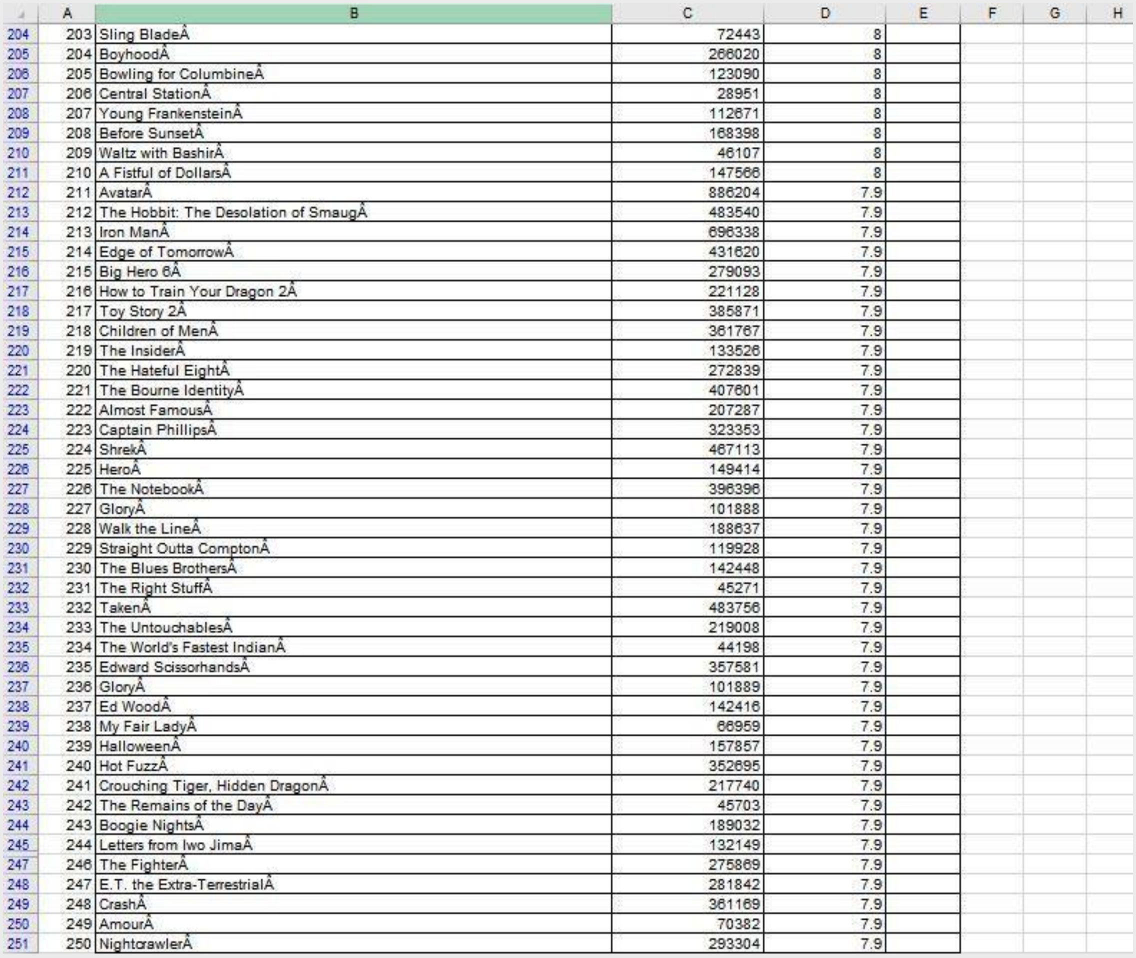
2. IMDB Top 250 movies



rank	movie_title	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	The Lord of the Rings: The Return of the KingÂ	1215718	8.9
6	Schindler's ListÂ	865020	8.9
7	Pulp FictionÂ	1324680	8.9
8	The Good, the Bad and the UglyÂ	503509	8.9
9	InceptionÂ	1468200	8.8
10	The Lord of the Rings: The Fellowship of the RingÂ	1238746	8.8
11	Fight ClubÂ	1347461	8.8
12	Forrest GumpÂ	1251222	8.8
13	Star Wars: Episode V - The Empire Strikes BackÂ	837759	8.8
14	The Lord of the Rings: The Two TowersÂ	1100446	8.7
15	The MatrixÂ	1217752	8.7
16	GoodfellasÂ	728685	8.7
17	Star Wars: Episode IV - A New HopeÂ	911097	8.7
18	One Flew Over the Cuckoo's NestÂ	680041	8.7
19	City of GodÂ	533200	8.7
20	Seven SamuraiÂ	229012	8.7
21	InterstellarÂ	928227	8.6
22	Saving Private RyanÂ	881236	8.6
23	Se7enÂ	1023511	8.6
24	The Silence of the LambsÂ	887467	8.6
25	Spirited AwayÂ	417971	8.6
26	American History XÂ	782437	8.6
27	The Usual SuspectsÂ	740918	8.6
28	Modern TimesÂ	143086	8.6
29	The Dark Knight RisesÂ	1144337	8.5
30	GladiatorÂ	982637	8.5



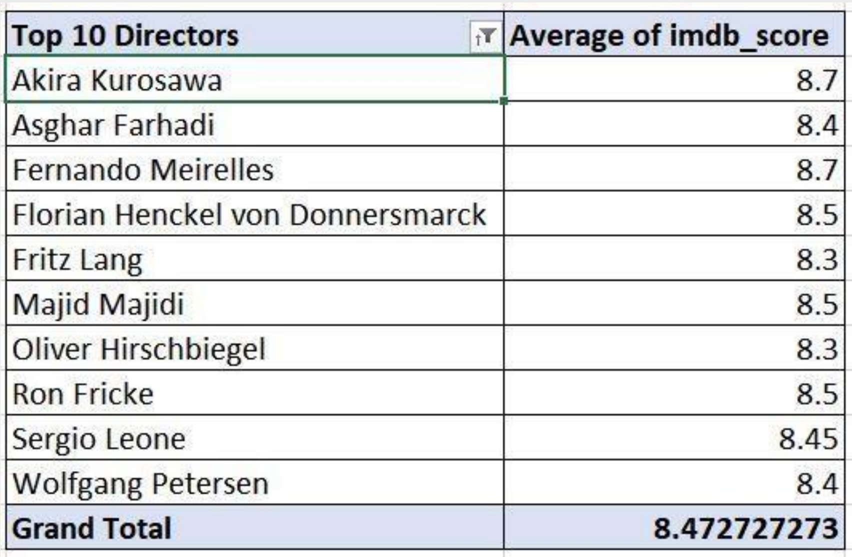
0	29	The Dark Knight RisesA	1144337	8.5
1	30	GladiatorÂ	982637	8.5
2	31	Terminator 2: Judgment DayÂ	744891	8.5
3	32	Django UnchainedÂ	955174	8.5
4	33	The DepartedÂ	873649	8.5
5	34	The Lion KingÂ	644348	8.5
6	35	The Green MileÂ	782610	8.5
7	36	The PrestigeÂ	844052	8.5
8	37	The PianistÂ	497946	8.5
9	38	Apocalypse NowÂ	450676	8.5
0	39	Raiders of the Lost ArkÂ	661017	8.5
1	40	PsychoÂ	422432	8.5
2	41	Back to the FutureÂ	732212	8.5
3	42	AlienÂ	563827	8.5
4	43	MementoÂ	845580	8.5
6	45	WhiplashÂ	399138	8.5
7	46	The Lives of OthersÂ	259379	8.5
8	47	Children of HeavenÂ	27882	8.5
9	48	WALLÂ·EÂ	718837	8.4
0	49	BraveheartÂ	736638	8.4
1	50	AmÃ©lieÂ	534262	8.4
2	51	Star Wars: Episode VI - Return of the JediÂ	681857	8.4
3	52	Once Upon a Time in AmericaÂ	221000	8.4
4	53	Princess MononokeÂ	221552	8.4
5	54	AliensÂ	488537	8.4
6	55	American BeautyÂ	822500	8.4
7	56	Lawrence of ArabiaÂ	192775	8.4
8	57	Das BootÂ	168203	8.4
9	58	Requiem for a DreamÂ	573541	8.4
0	59	OldboyÂ	356181	8.4
1	60	Reservoir DogsÂ	664719	8.4
2	61	A SeparationÂ	151812	8.4
3	62	Toy Story 3Â	544884	8.3
4	63	UpÂ	665575	8.3
5	64	Inside OutÂ	345198	8.3
6	65	Batman BeginsÂ	980946	8.3
7	66	Inglourious BasterdsÂ	885175	8.3
8	67	Indiana Jones and the Last CrusadeÂ	515306	8.3
9	68	L.A. ConfidentialÂ	414219	8.3
0	69	Toy StoryÂ	623757	8.3



	A	B	C	D	E	F	G	H
204	203	Sling BladeÂ	72443	8				
205	204	BoyhoodÂ	266020	8				
206	205	Bowling for ColumbineÂ	123090	8				
207	206	Central StationÂ	28951	8				
208	207	Young FrankensteinÂ	112671	8				
209	208	Before SunsetÂ	168398	8				
210	209	Waltz with BashirÂ	46107	8				
211	210	A Fistful of DollarsÂ	147566	8				
212	211	AvatarÂ	886204	7.9				
213	212	The Hobbit: The Desolation of SmaugÂ	483540	7.9				
214	213	Iron ManÂ	696338	7.9				
215	214	Edge of TomorrowÂ	431620	7.9				
216	215	Big Hero 6Â	279093	7.9				
217	216	How to Train Your Dragon 2Â	221128	7.9				
218	217	Toy Story 2Â	385871	7.9				
219	218	Children of MenÂ	361767	7.9				
220	219	The InsiderÂ	133526	7.9				
221	220	The Hateful EightÂ	272839	7.9				
222	221	The Bourne IdentityÂ	407601	7.9				
223	222	Almost FamousÂ	207287	7.9				
224	223	Captain PhillipsÂ	323353	7.9				
225	224	ShrekÂ	467113	7.9				
226	225	HeroÂ	149414	7.9				
227	226	The NotebookÂ	396396	7.9				
228	227	GloryÂ	101888	7.9				
229	228	Walk the LineÂ	188637	7.9				
230	229	Straight Outta ComptonÂ	119928	7.9				
231	230	The Blues BrothersÂ	142448	7.9				
232	231	The Right StuffÂ	45271	7.9				
233	232	TakenÂ	483756	7.9				
234	233	The UntouchablesÂ	219008	7.9				
235	234	The World's Fastest IndianÂ	44198	7.9				
236	235	Edward ScissorhandsÂ	357581	7.9				
237	236	GloryÂ	101889	7.9				
238	237	Ed WoodÂ	142416	7.9				
239	238	My Fair LadyÂ	66959	7.9				
240	239	HalloweenÂ	157857	7.9				
241	240	Hot FuzzÂ	352695	7.9				
242	241	Crouching Tiger, Hidden DragonÂ	217740	7.9				
243	242	The Remains of the DayÂ	45703	7.9				
244	243	Boogie NightsÂ	189032	7.9				
245	244	Letters from Iwo JimaÂ	132149	7.9				
247	246	The FighterÂ	275869	7.9				
248	247	E.T. the Extra-TerrestrialÂ	281842	7.9				
249	248	CrashÂ	361169	7.9				
250	249	AmourÂ	70382	7.9				
251	250	NightcrawlerÂ	293304	7.9				

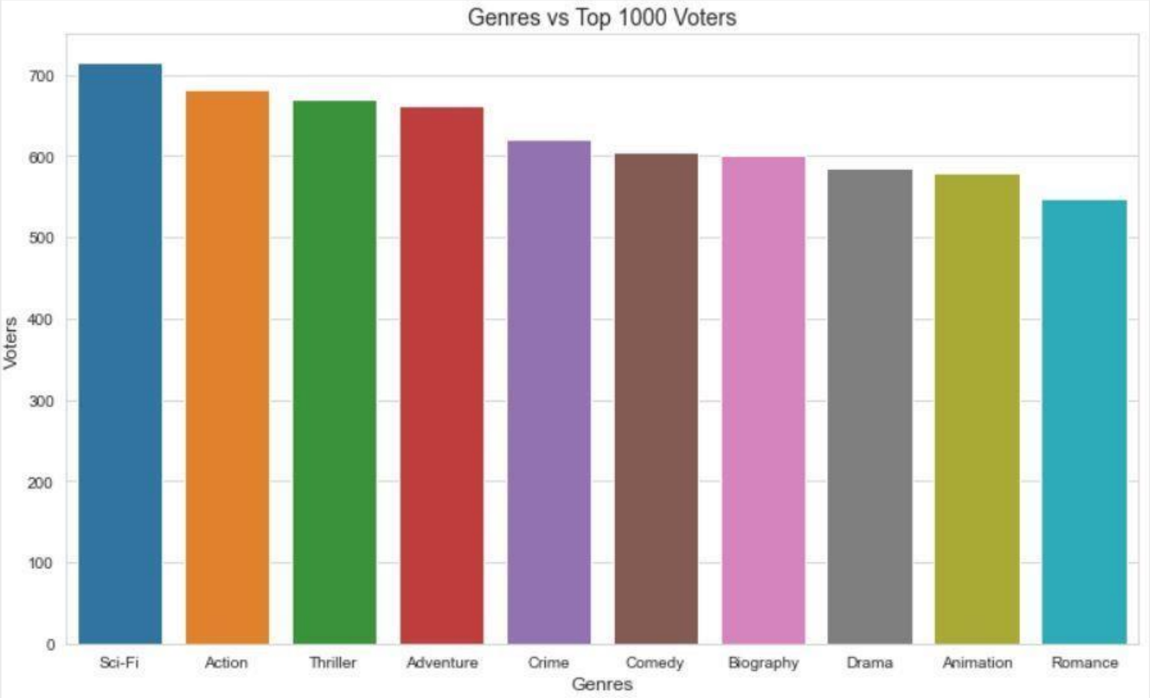
Note: total 250 movies were taken but not all the info is shared. Please refer above ones. (I can provide excel sheet for this too if needed)

3. Best Directors:



Top 10 Directors	Average of imdb_score
Akira Kurosawa	8.7
Asghar Farhadi	8.4
Fernando Meirelles	8.7
Florian Henckel von Donnersmarck	8.5
Fritz Lang	8.3
Majid Majidi	8.5
Oliver Hirschbiegel	8.3
Ron Fricke	8.5
Sergio Leone	8.45
Wolfgang Petersen	8.4
Grand Total	**8.472727273**

4. Popular Genres:



BANK LOAN CASE STUDY

PROJECT DESCRIPTION:

This project is done to gain hands on experience in handling huge datasets using EDA (Exploratory Data Analysis). This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

APPROACH:

On the given dataset, EDA was performed accordingly. After that, all the queries regarding this dataset were solved and the solution to those queries were found out by using various methods.

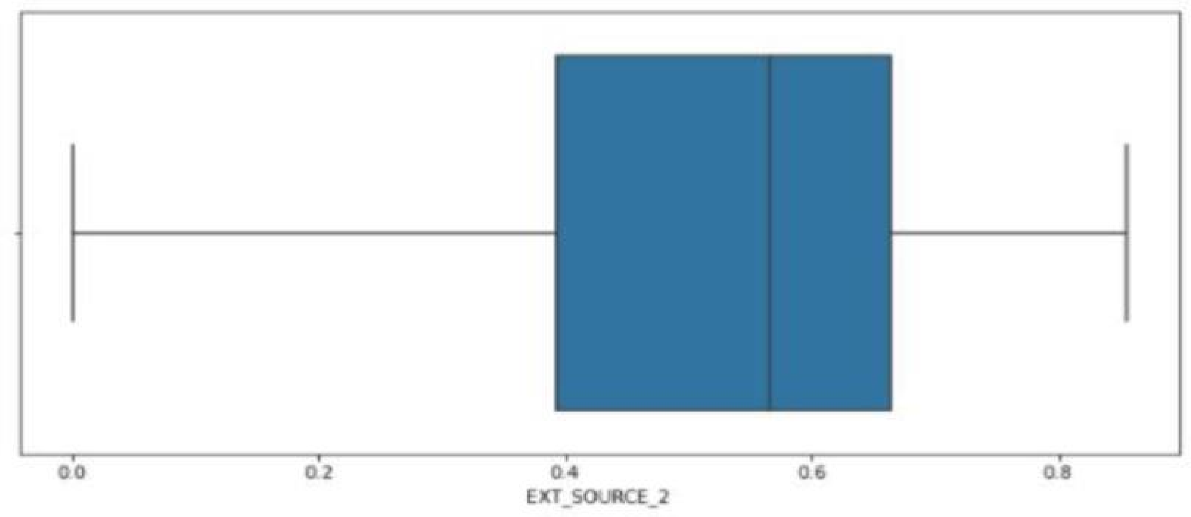
TECH-STACK USED:

In this project, MS-Excel 2020 was used for doing all the analysis and finding out solutions.

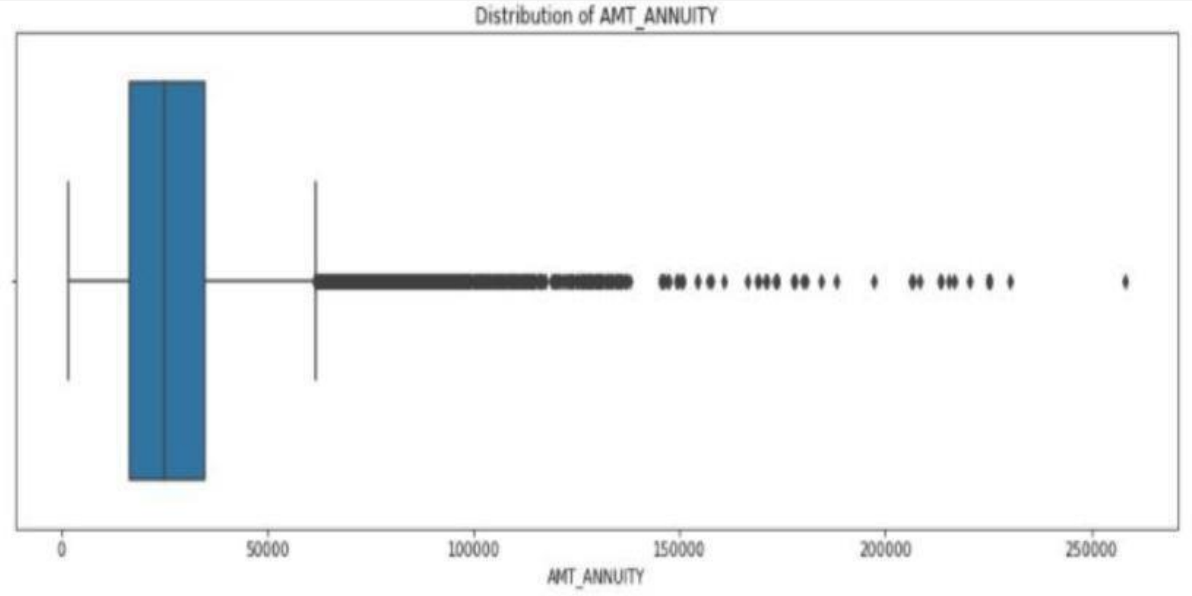
INSIGHTS:

1. We first found out the missing data in the provided dataset and worked on it accordingly to gain required results. We found out whether the missing data had any impact on our dataset, if not, then it can be removed.

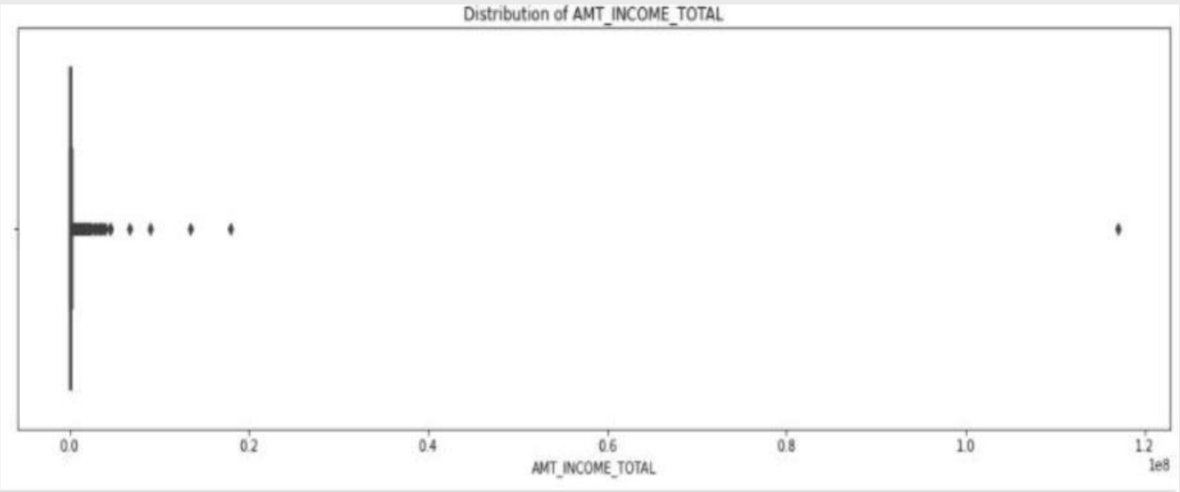
2. Outliers:



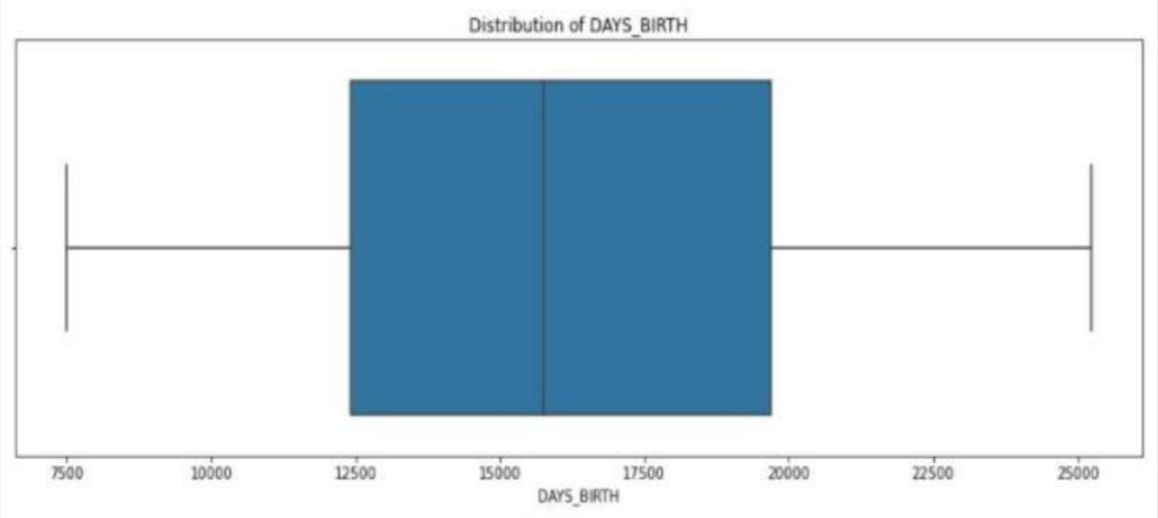
There are no outliers in EXT_SOURCE_2 as seen above.



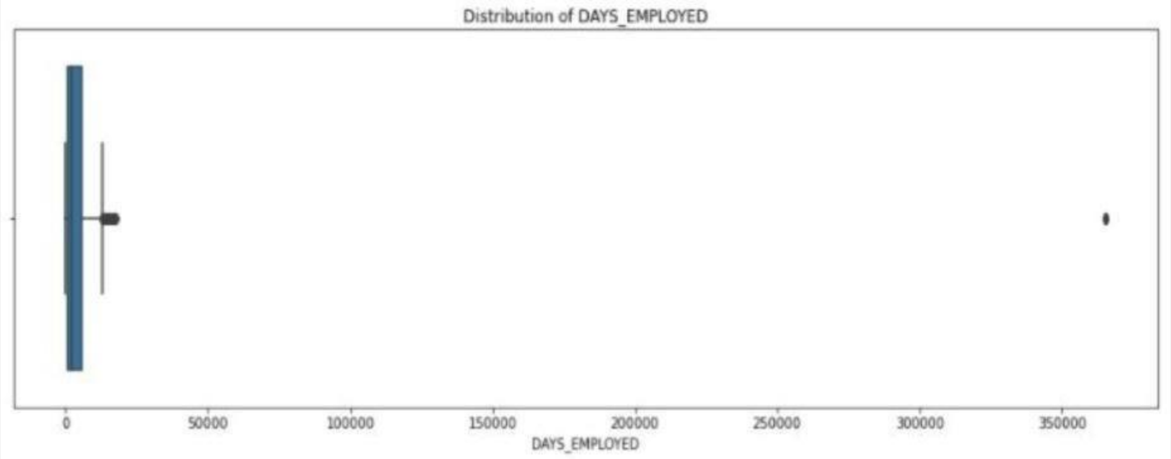
As seen above, there is a value above 250000 in AMT_ANNUITY.



As seen above, there is a outlier in AMT_INCOME_TOTAL.



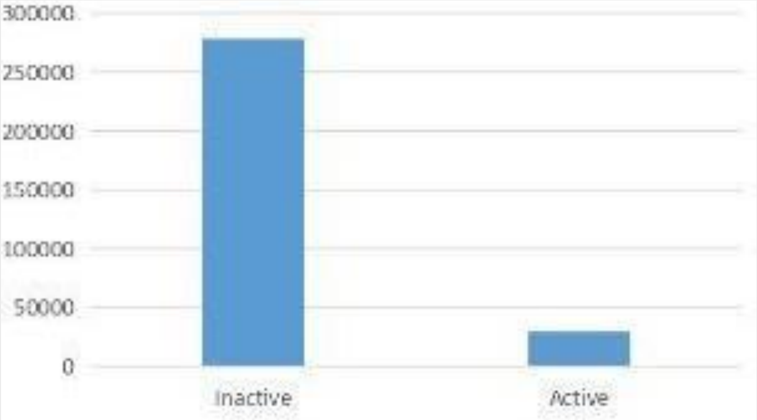
There are no outliers in DAYS_BIRTH.



In DAYS_EMPLOYED, an outlier can be seen on extreme right i.e, after 350000

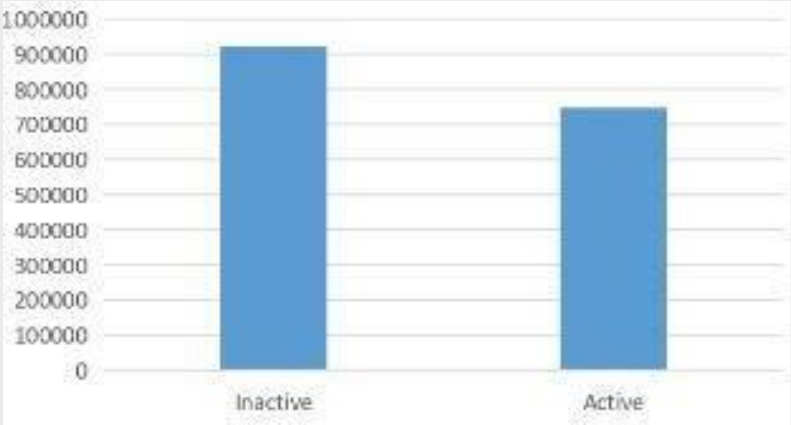
3. Imbalance in Data:

- Application Dataset:



The percentage of data imbalance is 10.5% and number of active variables are 278232 and inactive are 29279.

- Previous_Application Dataset:



The percentage of data imbalance is 82% and number of active variables are 922661 and inactive are 747553.

4. Univariate and Segmented univariate:

Univariate analysis is the simplest kind of data analysis in the field of statistics. This could be either descriptive or inferential in nature as is the case in any data analysis in statistics. The key thing about the univariate analysis to remember is that there is only one data involved here, since there are more variables involved in this dataset. So we will conduct bivariate analysis on the following dataset.

5. Bivariate analysis:

   Bivariate analysis is stated to be an analysis of any concurrent relation between two variables or attributes. This study explores the relationship of two variables as well as the depth of this relationship to figure out if there are any discrepancies between two variables and any causes of this difference. Correlation analysis has been conducted on the basis of given dataset.

XYZ ADS AIRING REPORT ANALYSIS

Project Description:

In this dataset, analysis is done on a dataset having information on different TV Airing Brands, their product, their category. This dataset includes the network through which Ads are airing, types of networks like Cable/ Broadcast and the show name also on which Ads got aired. You can also see the data of Dayparts, Time zone and the time & date at which Ads got aired. IT also includes other data like Pod Position (the lesser the valuable), duration for which Ads aired on screen, Equivalent sales &, total amount spent on the Ads aired.

Approach:

The first approach in this project has been data cleaning so that the data is perfectly operable upon. After that various analysis has been done to find the different solutions to questions according to needs.

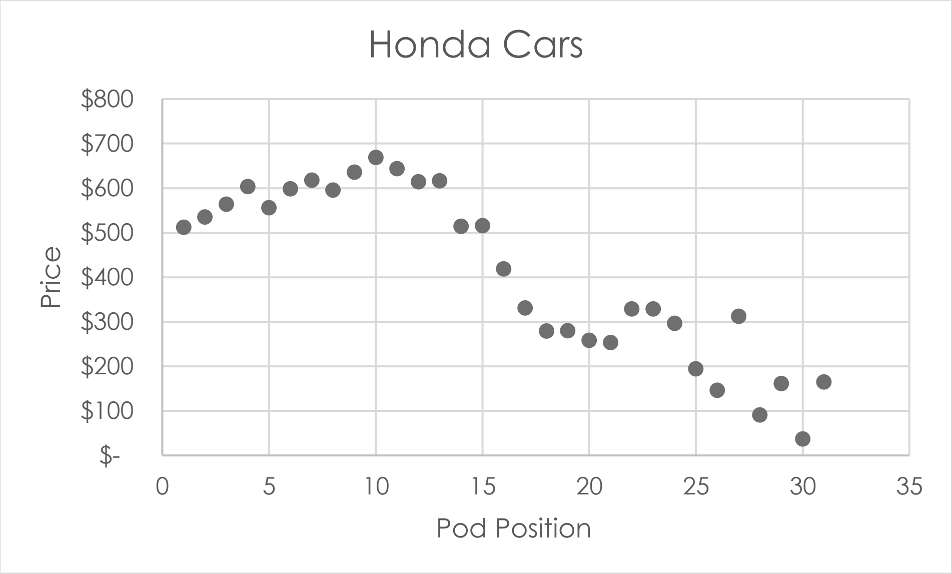
Tech-Stack Used:

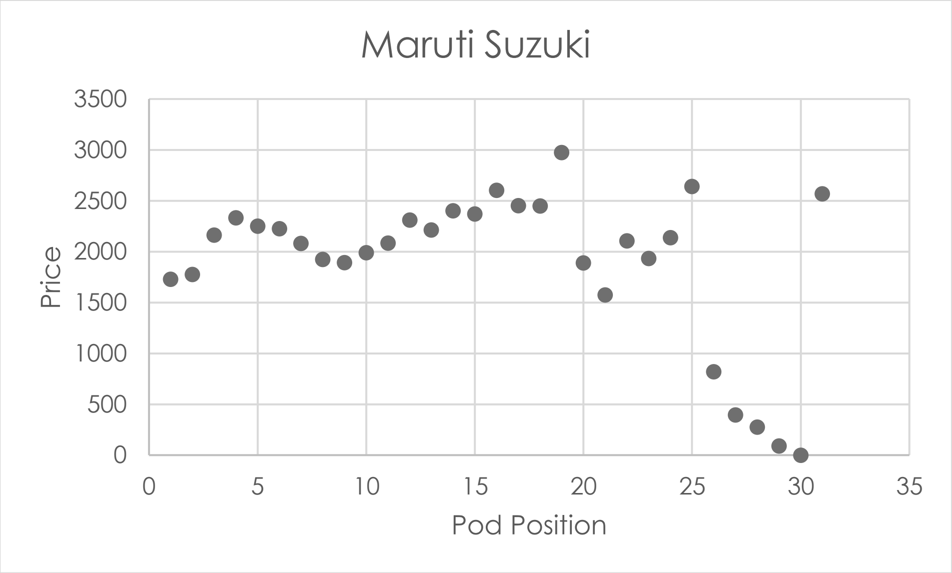
In this project, Microsoft - Excel was used to carry out all the analysis and cleaning part in the project.

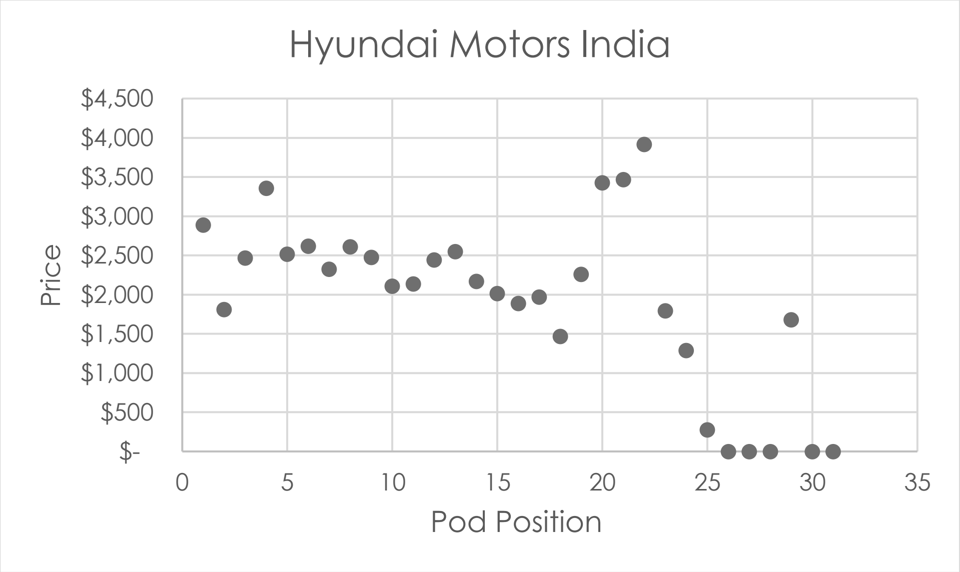
Insights:

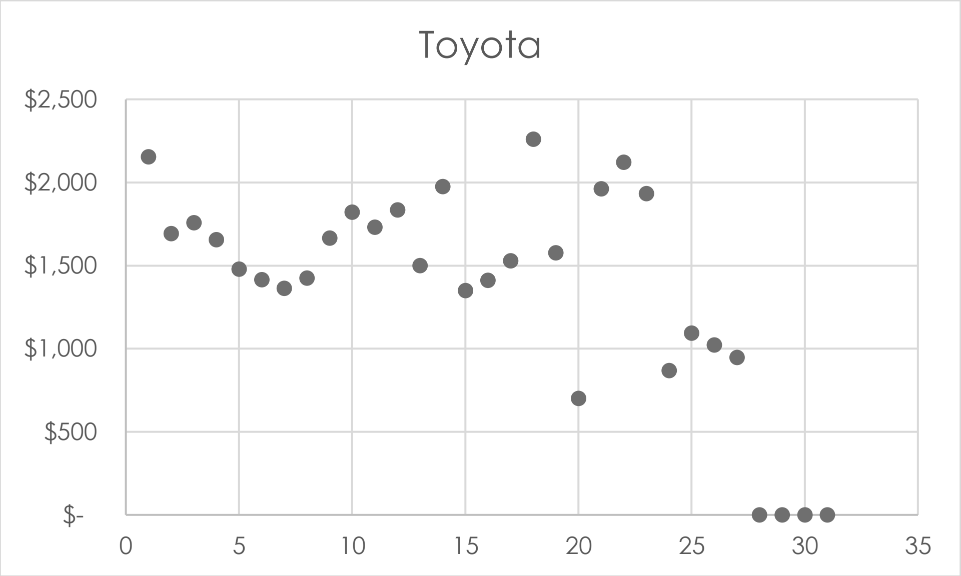
1. Pod positions:

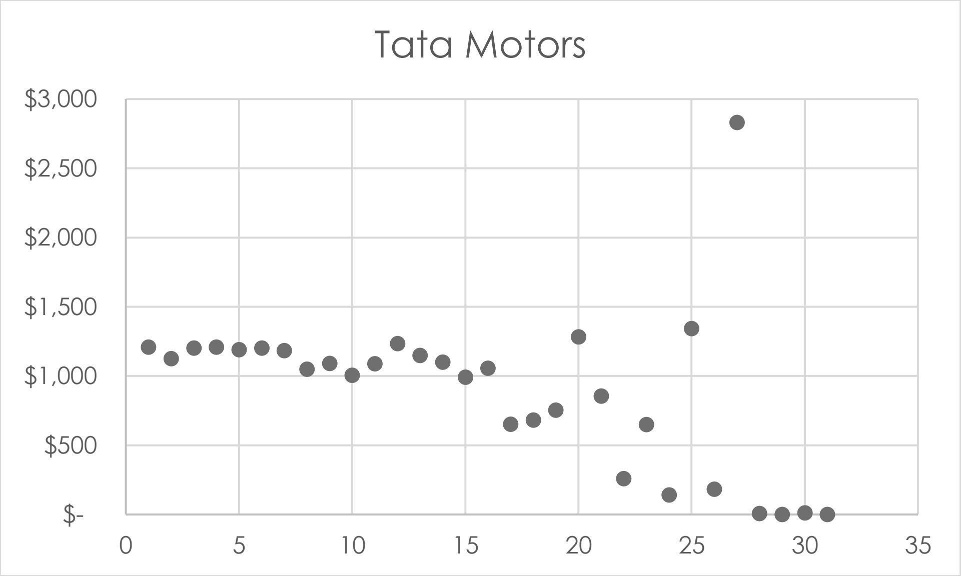
It is the position or sequence of ads in which it is shown in a particular break. And as per the dataset, For each car manufactures, as pod position increases, price increase earlier and then it starts declining. Below are few examples of it.

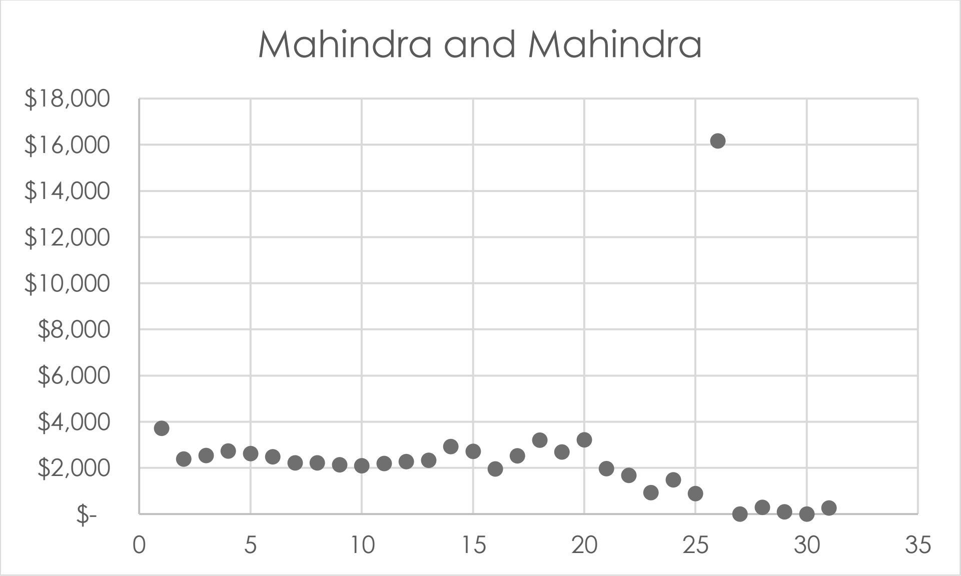

Honda Cars
Price
$800
$700
$600
$500
$400
$300
$200
$100
$-
0
5
10
15
20
25
30
35
Pod Position


Maruti Suzuki
Price
3500
3000
2500
2000
1500
1000
500
0
0
5
10
15
20
25
30
35
Pod Position


Hyundai Motors India
Price
$4,500
$4,000
$3,500
$3,000
$2,500
$2,000
$1,500
$1,000
$500
$-
0
5
10
15
20
25
30
35
Pod Position


Toyota
$2,500
$2,000
$1,500
$1,000
$500
$-
0
5
10
15
20
25
30
35


Tata Motors
$3,000
$2,500
$2,000
$1,500
$1,000
$500
$-
0
5
10
15
20
25
30
35

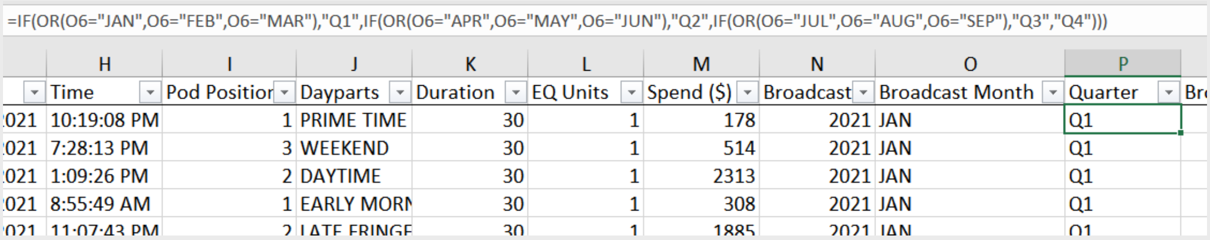

Mahindra and Mahindra
$18,000
$16,000
$14,000
$12,000
$10,000
$8,000
$6,000
$4,000
$2,000
$-
0
5
10
15
20
25
30
35

2. The share of various brands in TV airings:

In the main dataset, a column has been made beside broadcast month and that column's name is Quarters. This column all 4 quarters present in dataset.

- Q1 contains January, February, March.
- Q2 contains April, May, June.
- Q3 contains July, August, September.
- Q4 contains October, November, December.

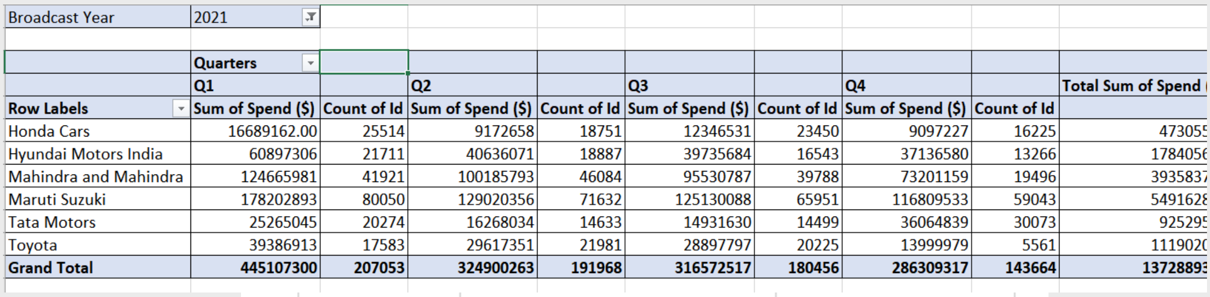
The formula to make these columns in excel is:
=IF(OR(O6="JAN",O6="FEB",O6="MAR"),"Q1",IF(OR(O6="APR",O6="MAY",O6="JUN"),"Q2",IF(OR(O6="JUL",O6="AUG",O6="SEP"),"Q3","Q4")))



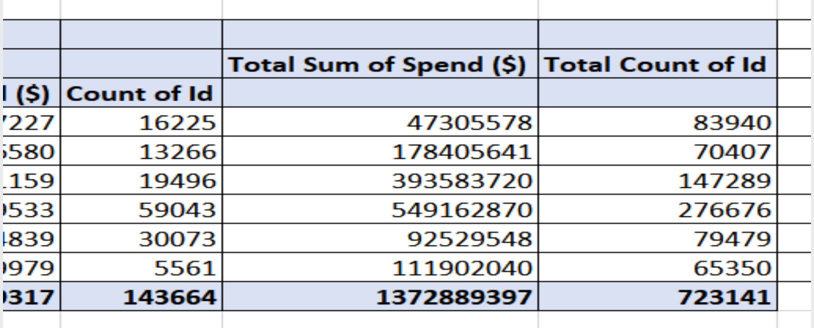
=IF(OR(O6="JAN",O6="FEB",O6="MAR"),"Q1",IF(OR(O6="APR",O6="MAY",O6="JUN"),"Q2",IF(OR(O6="JUL",O6="AUG",O6="SEP"),"Q3","Q4")))

	H	I	J	K	L	M	N	O	P	
	Time	Pod Positior	Dayparts	Duration	EQ Units	Spend ($)	Broadcast	Broadcast Month	Quarter	Bro
:021	10:19:08 PM	1	PRIME TIME	30	1	178	2021	JAN	Q1	
:021	7:28:13 PM	3	WEEKEND	30	1	514	2021	JAN	Q1	
:021	1:09:26 PM	2	DAYTIME	30	1	2313	2021	JAN	Q1	
:021	8:55:49 AM	1	EARLY MORN	30	1	308	2021	JAN	Q1	
:021	11:07:43 PM	2	LATE FRINGE	30	1	1885	2021	JAN	Q1	

After the above step, with the help of pivot tables, we performed the Analysis



Broadcast Year	2021								
	Quarters								
	Q1		Q2		Q3		Q4		Total Sum of Spend (
Row Labels	Sum of Spend ($)	Count of Id	Sum of Spend ($)	Count of Id	Sum of Spend ($)	Count of Id	Sum of Spend ($)	Count of Id	
Honda Cars	16689162.00	25514	9172658	18751	12346531	23450	9097227	16225	473055
Hyundai Motors India	60897306	21711	40636071	18887	39735684	16543	37136580	13266	1784056
Mahindra and Mahindra	124665981	41921	100185793	46084	95530787	39788	73201159	19496	3935837
Maruti Suzuki	178202893	80050	129020356	71632	125130088	65951	116809533	59043	5491628
Tata Motors	25265045	20274	16268034	14633	14931630	14499	36064839	30073	925295
Toyota	39386913	17583	29617351	21981	28897797	20225	13999979	5561	1119020
Grand Total	**445107300**	**207053**	**324900263**	**191968**	**316572517**	**180456**	**286309317**	**143664**	**13728893**



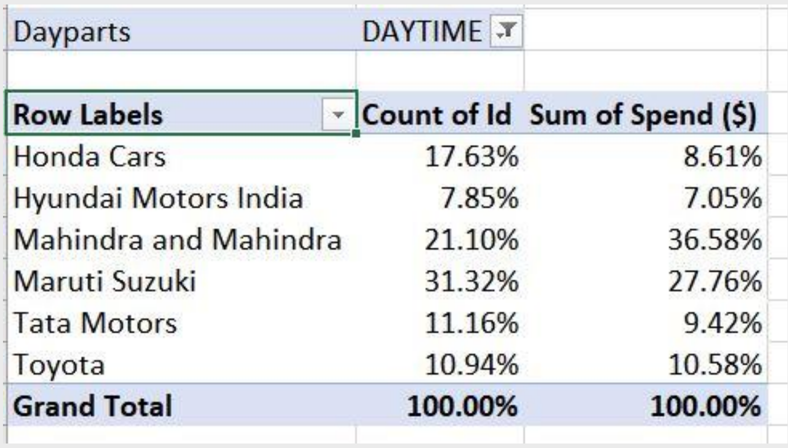
		Total Sum of Spend ($)	Total Count of Id
l ($)	Count of Id		
7227	16225	47305578	83940
6580	13266	178405641	70407
.159	19496	393583720	147289
)533	59043	549162870	276676
4839	30073	92529548	79479
)979	5561	111902040	65350
)317	**143664**	**1372889397**	**723141**

As seen above, for 2021 in Q1 the sum of spend for maximum companies is much greater than in Q4.

3. Competitive analysis of all brands:

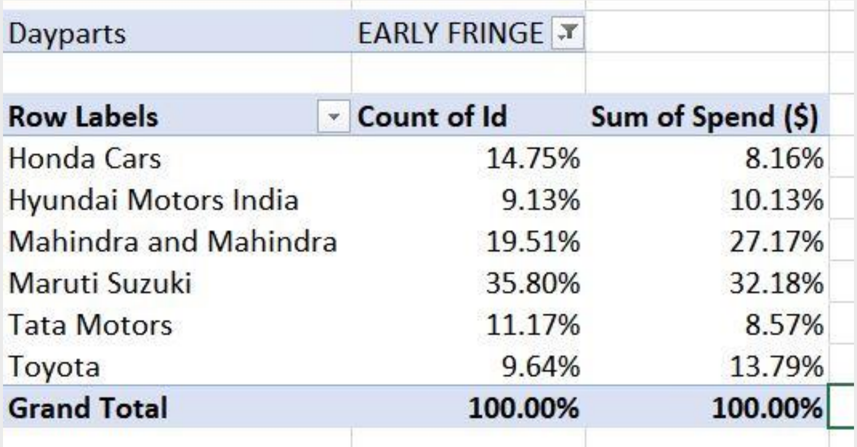
Percentage of ads and amount spend for all companies at different time

At Day time:



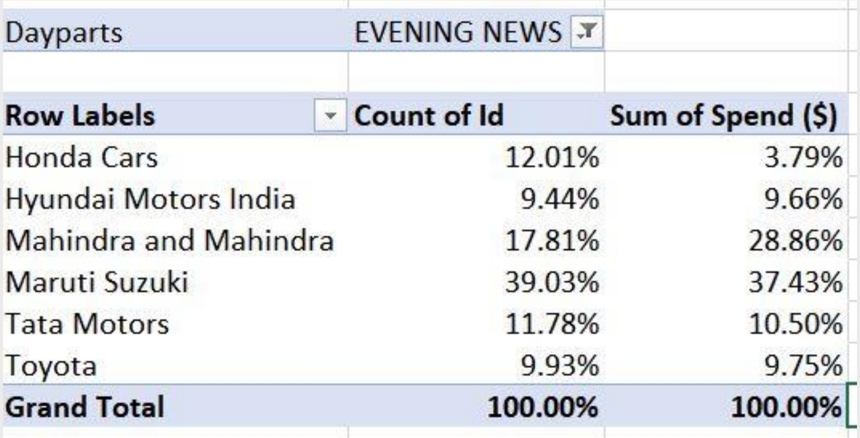
Dayparts	DAYTIME	
Row Labels	**Count of Id**	**Sum of Spend ($)**
Honda Cars	17.63%	8.61%
Hyundai Motors India	7.85%	7.05%
Mahindra and Mahindra	21.10%	36.58%
Maruti Suzuki	31.32%	27.76%
Tata Motors	11.16%	9.42%
Toyota	10.94%	10.58%
Grand Total	**100.00%**	**100.00%**

At Early Fringe:



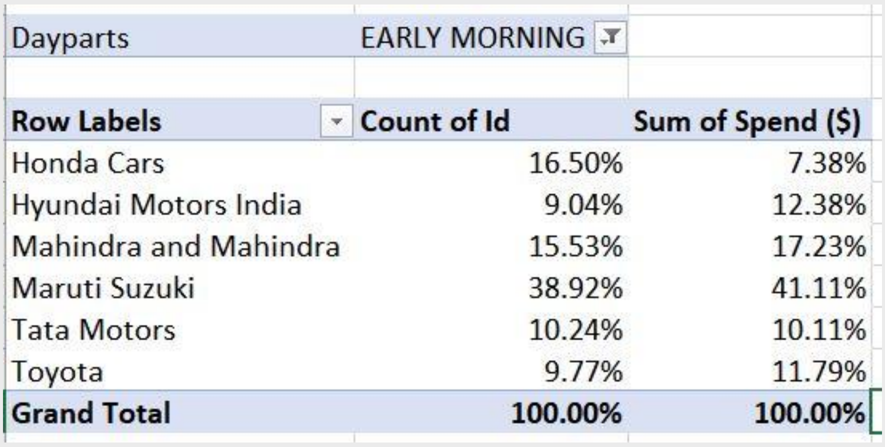
Dayparts	EARLY FRINGE	
Row Labels	**Count of Id**	**Sum of Spend ($)**
Honda Cars	14.75%	8.16%
Hyundai Motors India	9.13%	10.13%
Mahindra and Mahindra	19.51%	27.17%
Maruti Suzuki	35.80%	32.18%
Tata Motors	11.17%	8.57%
Toyota	9.64%	13.79%
Grand Total	**100.00%**	**100.00%**

At Evening News:



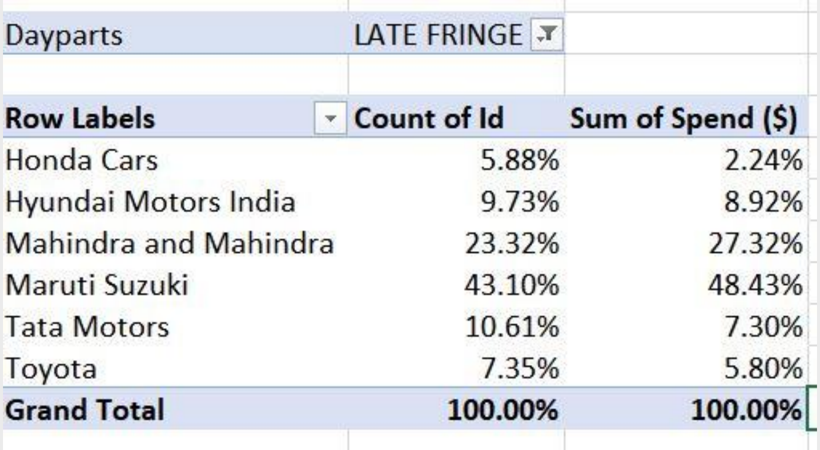
Dayparts	EVENING NEWS	
Row Labels	**Count of Id**	**Sum of Spend ($)**
Honda Cars	12.01%	3.79%
Hyundai Motors India	9.44%	9.66%
Mahindra and Mahindra	17.81%	28.86%
Maruti Suzuki	39.03%	37.43%
Tata Motors	11.78%	10.50%
Toyota	9.93%	9.75%
Grand Total	**100.00%**	**100.00%**

At Early Morning:



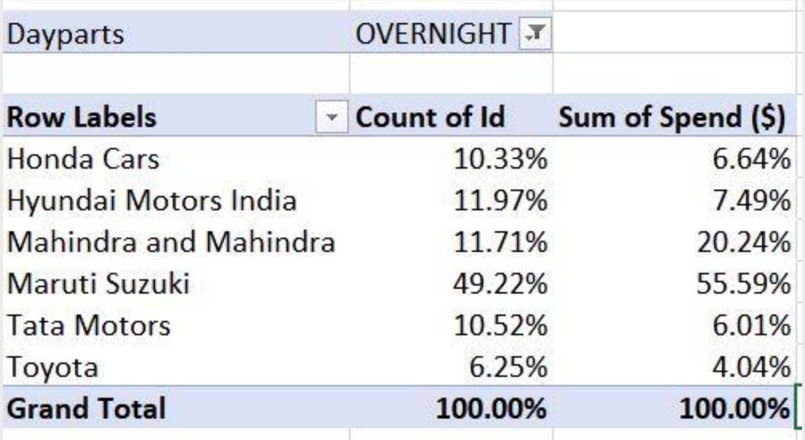
Dayparts	EARLY MORNING	
Row Labels	**Count of Id**	**Sum of Spend ($)**
Honda Cars	16.50%	7.38%
Hyundai Motors India	9.04%	12.38%
Mahindra and Mahindra	15.53%	17.23%
Maruti Suzuki	38.92%	41.11%
Tata Motors	10.24%	10.11%
Toyota	9.77%	11.79%
Grand Total	**100.00%**	**100.00%**

At Late Fringe:



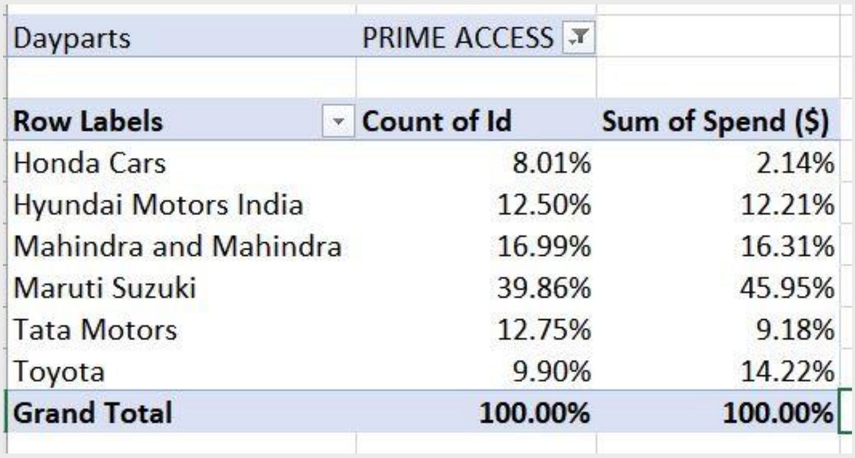
Dayparts	LATE FRINGE	
Row Labels	**Count of Id**	**Sum of Spend ($)**
Honda Cars	5.88%	2.24%
Hyundai Motors India	9.73%	8.92%
Mahindra and Mahindra	23.32%	27.32%
Maruti Suzuki	43.10%	48.43%
Tata Motors	10.61%	7.30%
Toyota	7.35%	5.80%
Grand Total	**100.00%**	**100.00%**

At Overnight:



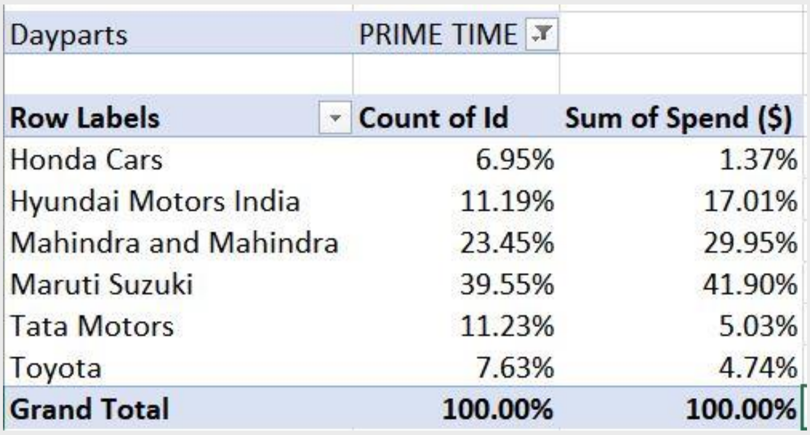
Dayparts	OVERNIGHT	
Row Labels	**Count of Id**	**Sum of Spend ($)**
Honda Cars	10.33%	6.64%
Hyundai Motors India	11.97%	7.49%
Mahindra and Mahindra	11.71%	20.24%
Maruti Suzuki	49.22%	55.59%
Tata Motors	10.52%	6.01%
Toyota	6.25%	4.04%
Grand Total	**100.00%**	**100.00%**

At Prime Access:



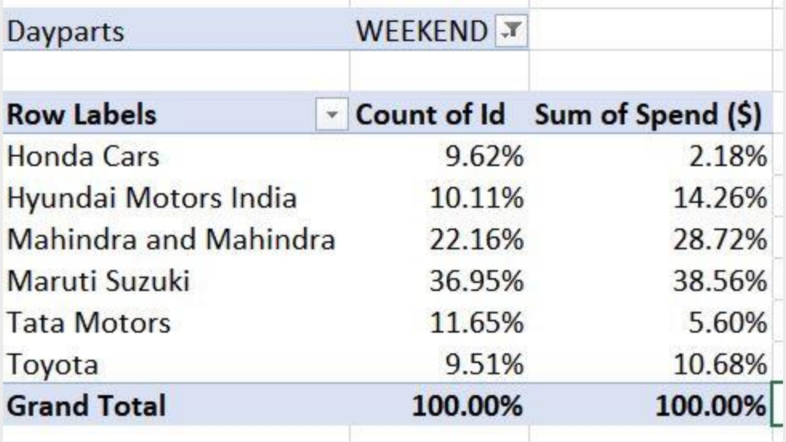
Dayparts	PRIME ACCESS	
Row Labels	**Count of Id**	**Sum of Spend ($)**
Honda Cars	8.01%	2.14%
Hyundai Motors India	12.50%	12.21%
Mahindra and Mahindra	16.99%	16.31%
Maruti Suzuki	39.86%	45.95%
Tata Motors	12.75%	9.18%
Toyota	9.90%	14.22%
Grand Total	**100.00%**	**100.00%**

At Prime Time:



Dayparts	PRIME TIME	
Row Labels	**Count of Id**	**Sum of Spend ($)**
Honda Cars	6.95%	1.37%
Hyundai Motors India	11.19%	17.01%
Mahindra and Mahindra	23.45%	29.95%
Maruti Suzuki	39.55%	41.90%
Tata Motors	11.23%	5.03%
Toyota	7.63%	4.74%
Grand Total	**100.00%**	**100.00%**

At Weekends:



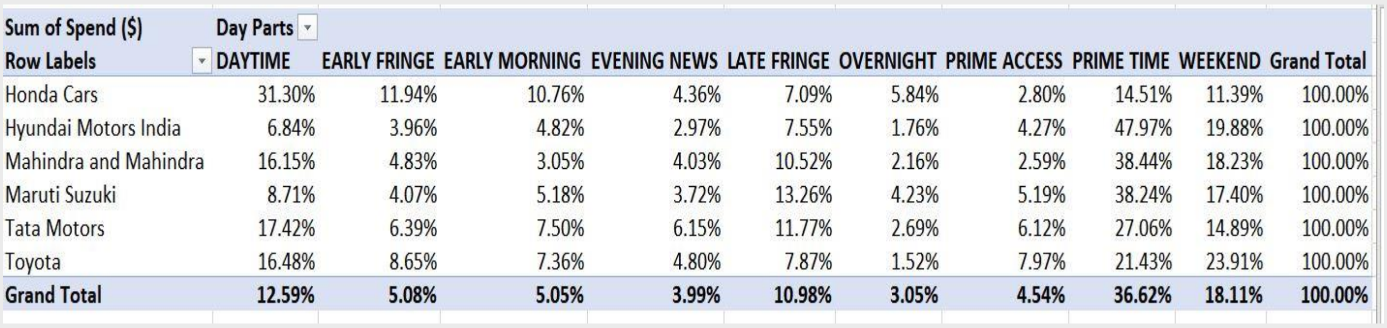
Dayparts	WEEKEND	
Row Labels	**Count of Id**	**Sum of Spend ($)**
Honda Cars	9.62%	2.18%
Hyundai Motors India	10.11%	14.26%
Mahindra and Mahindra	22.16%	28.72%
Maruti Suzuki	36.95%	38.56%
Tata Motors	11.65%	5.60%
Toyota	9.51%	10.68%
Grand Total	**100.00%**	**100.00%**

As we can see above from all the tables,

- Honda cars provide more ads in the daytime and spend the most amount for it.
- Whereas Hyundai Motors India has more ads on the prime access time and spend more on prime time.

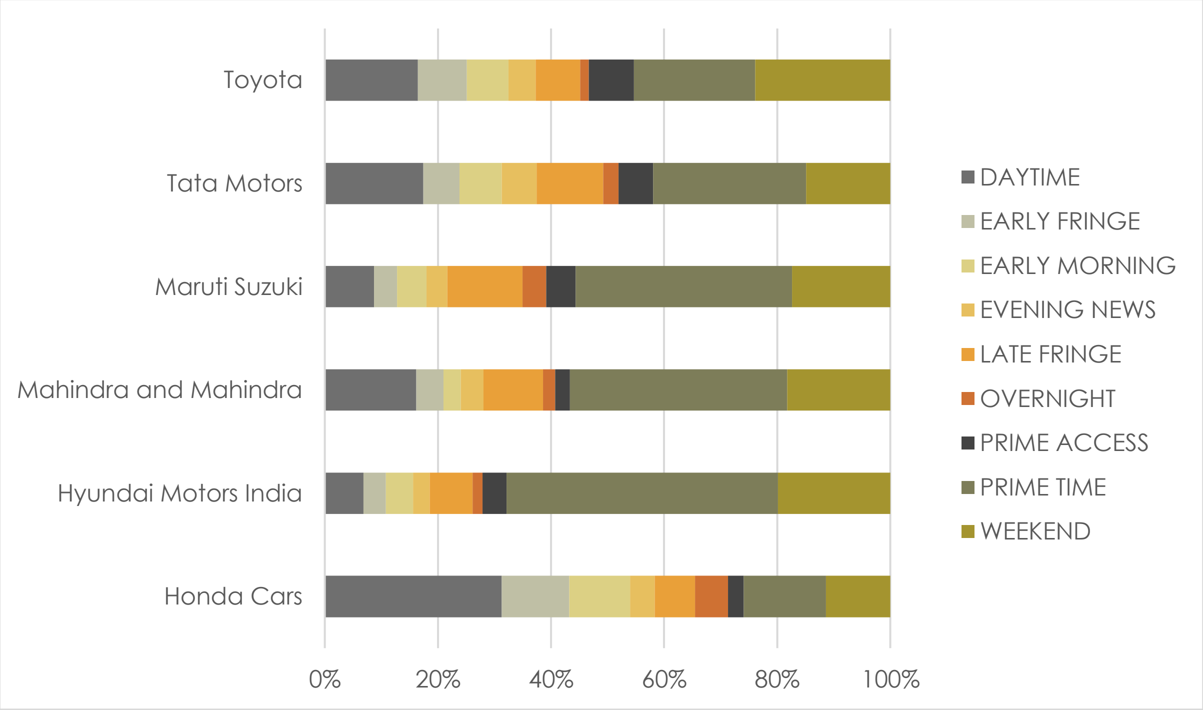
- Mahindra & Mahindra have their more ads on the prime time but spend the most on daytime ads.
- Maruti Suzuki airs their ads most on overnight and spend more on it also.
- Tata Motors have their many ads on the prime access time and they spend the most for it.
- Whereas the Toyota cars have their ads more on the daytime and spend the most on prime access.

By combining the dayparts and network types and converting those rows to 100%, we get the table like,



Sum of Spend ($)	Day Parts									
Row Labels	DAYTIME	EARLY FRINGE	EARLY MORNING	EVENING NEWS	LATE FRINGE	OVERNIGHT	PRIME ACCESS	PRIME TIME	WEEKEND	Grand Total
Honda Cars	31.30%	11.94%	10.76%	4.36%	7.09%	5.84%	2.80%	14.51%	11.39%	100.00%
Hyundai Motors India	6.84%	3.96%	4.82%	2.97%	7.55%	1.76%	4.27%	47.97%	19.88%	100.00%
Mahindra and Mahindra	16.15%	4.83%	3.05%	4.03%	10.52%	2.16%	2.59%	38.44%	18.23%	100.00%
Maruti Suzuki	8.71%	4.07%	5.18%	3.72%	13.26%	4.23%	5.19%	38.24%	17.40%	100.00%
Tata Motors	17.42%	6.39%	7.50%	6.15%	11.77%	2.69%	6.12%	27.06%	14.89%	100.00%
Toyota	16.48%	8.65%	7.36%	4.80%	7.87%	1.52%	7.97%	21.43%	23.91%	100.00%
Grand Total	**12.59%**	**5.08%**	**5.05%**	**3.99%**	**10.98%**	**3.05%**	**4.54%**	**36.62%**	**18.11%**	**100.00%**

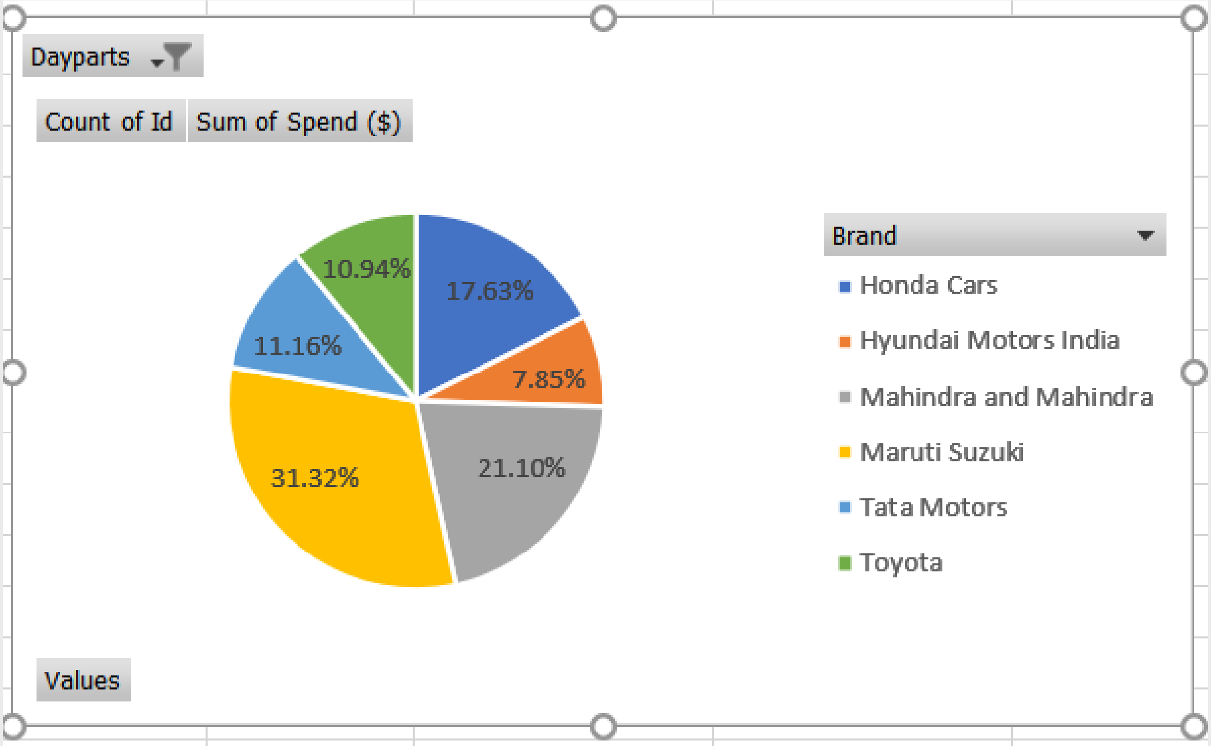
To see above table in a graphical way,



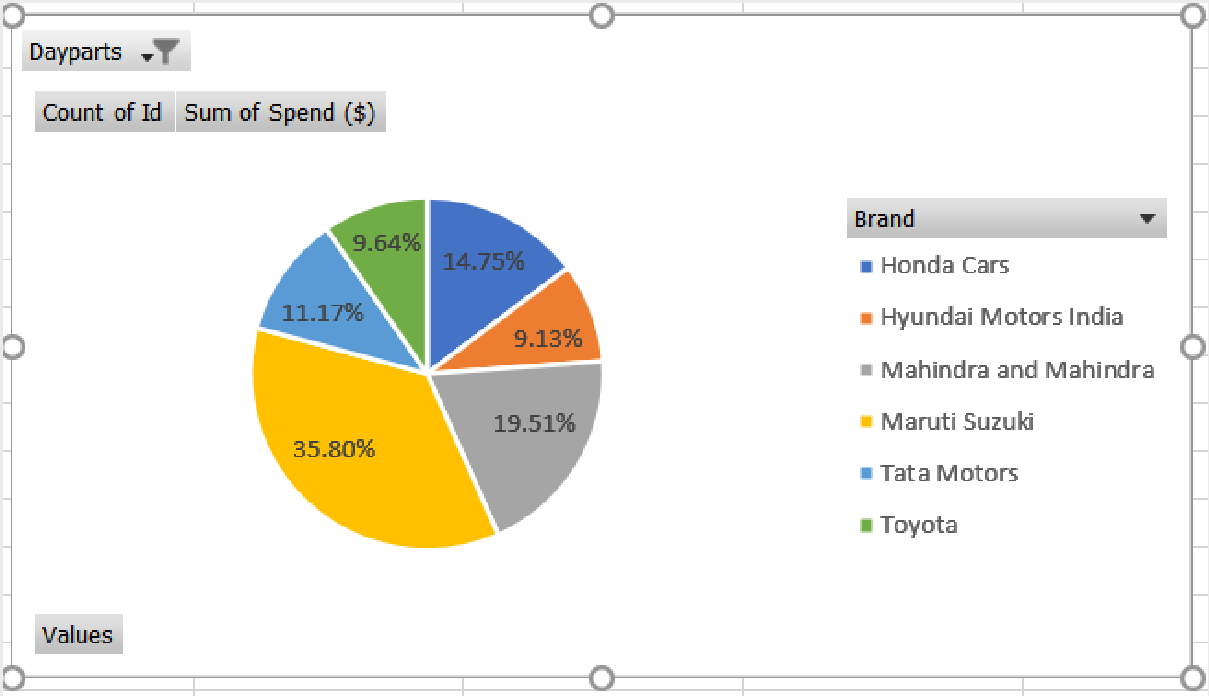
In this, we can clearly see that Honda cars concentrate more on spending during the daytime and all remaining car brands except Toyota focus more on the primetime where Toyota concentrates more on the weekend.

We can also see pie charts for above ads table for a particular day part.

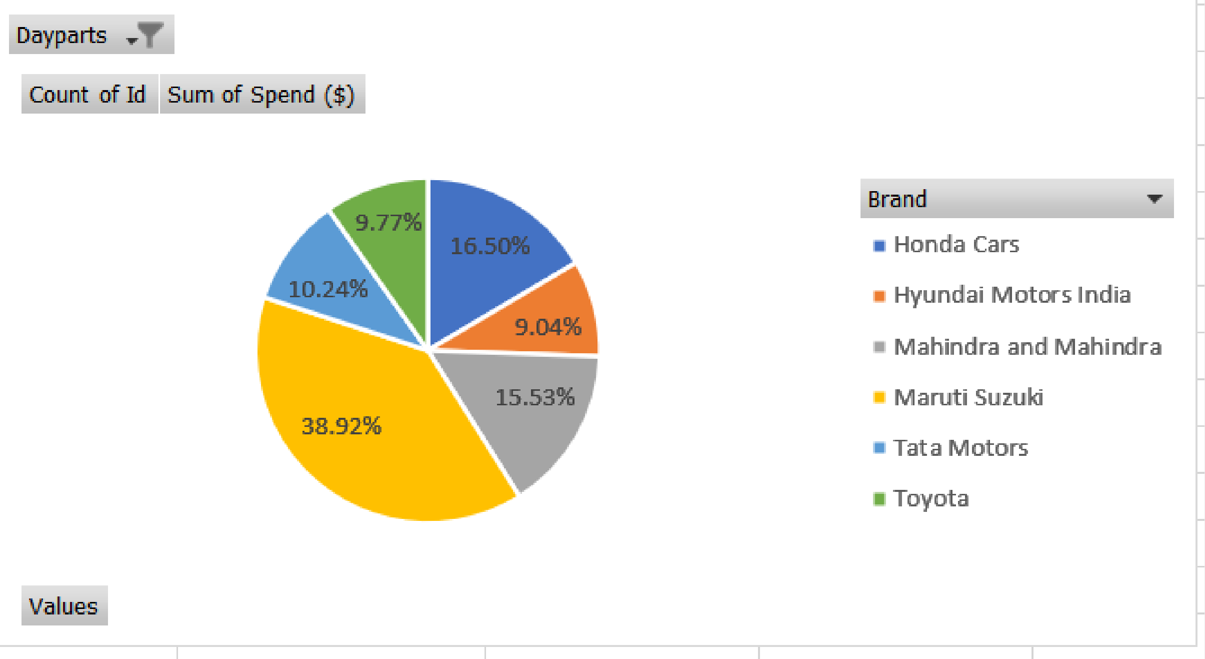
For Day Time:



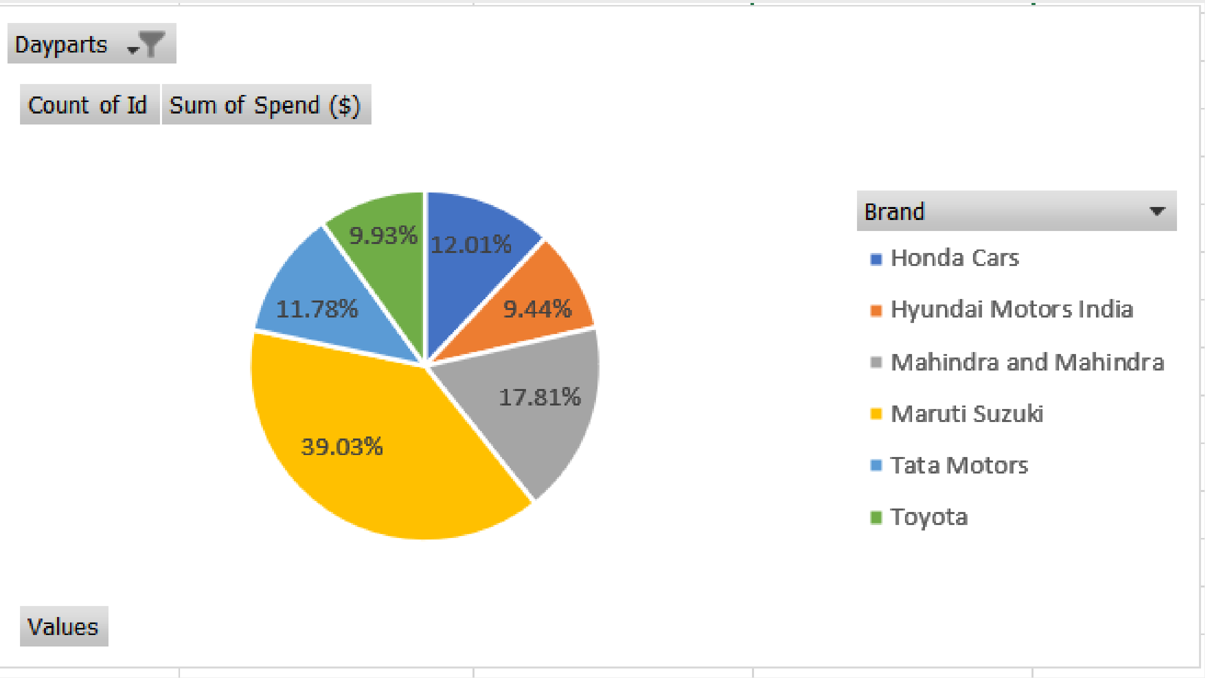
For Early Fringe:



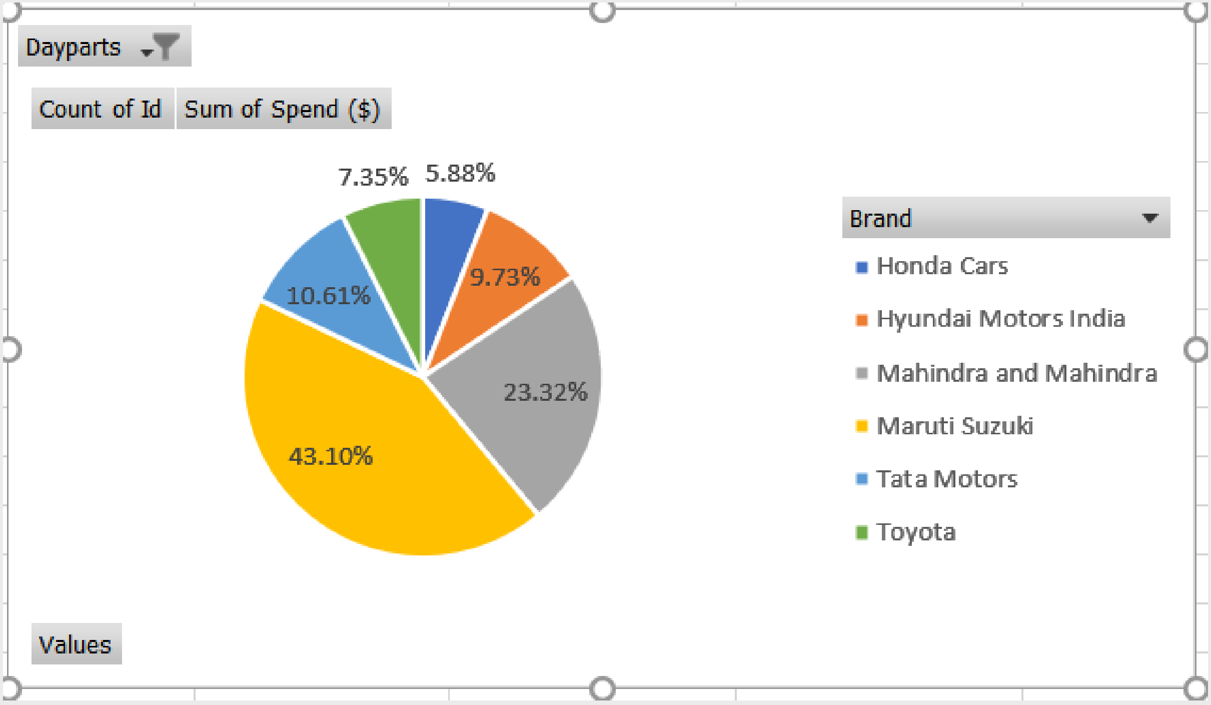
For Early Mornings:



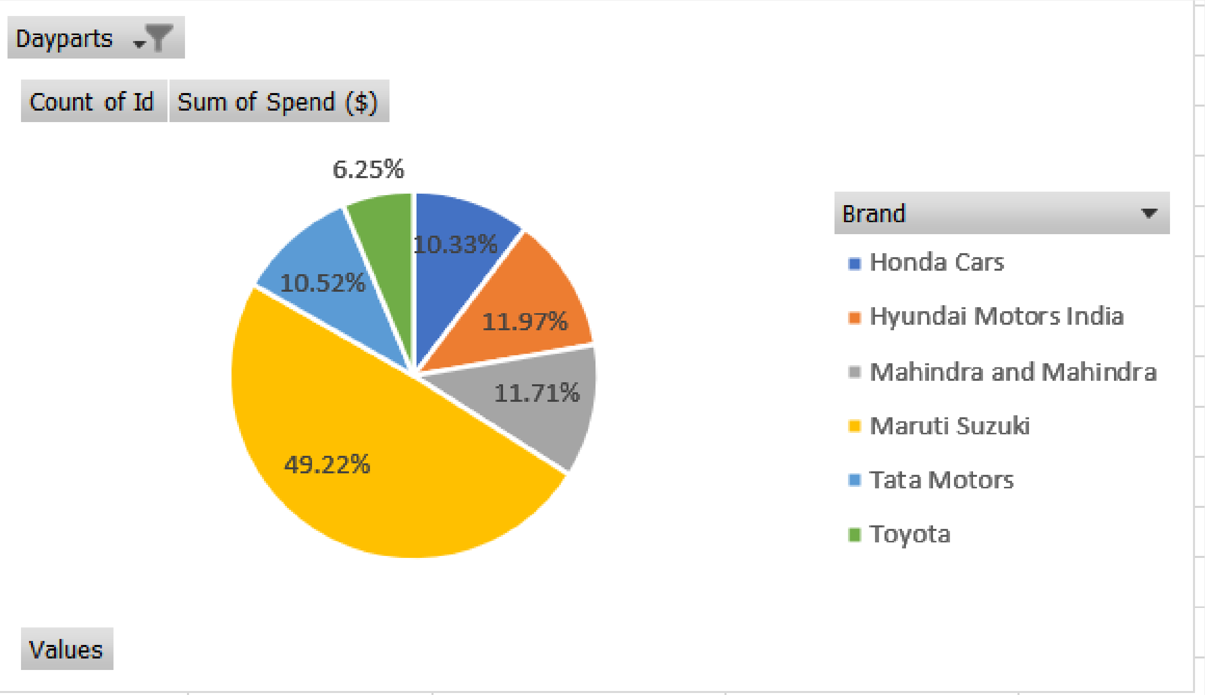
For Evening News:



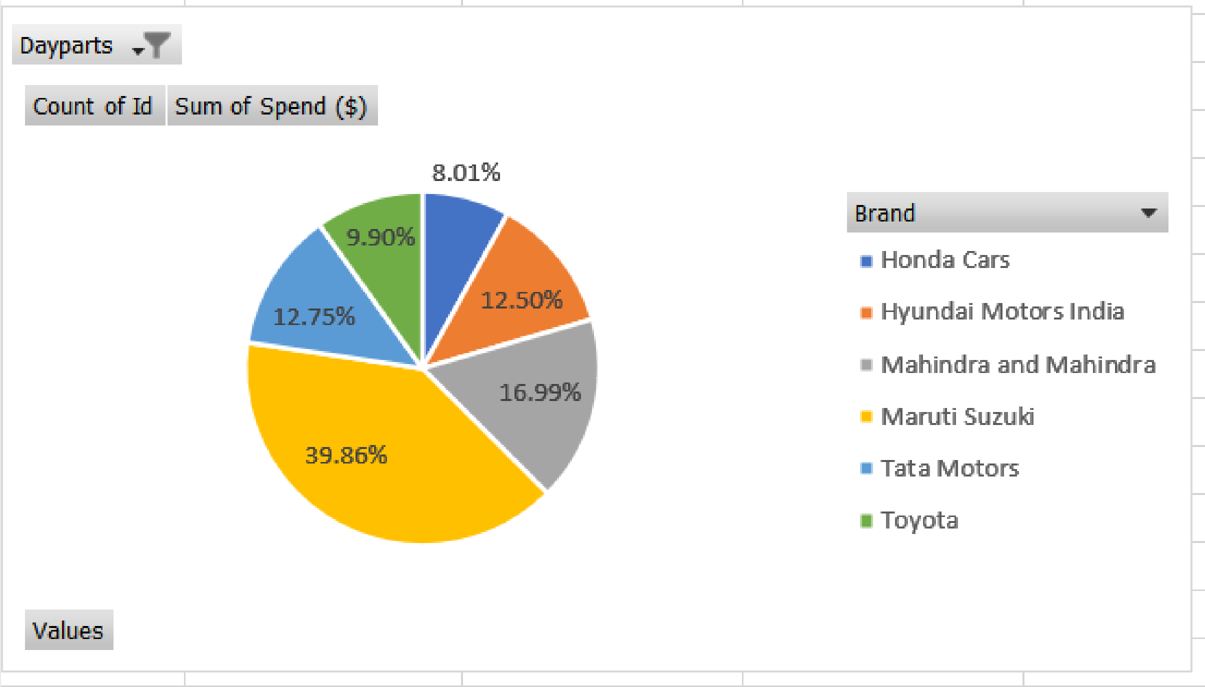
For Late Fringe:



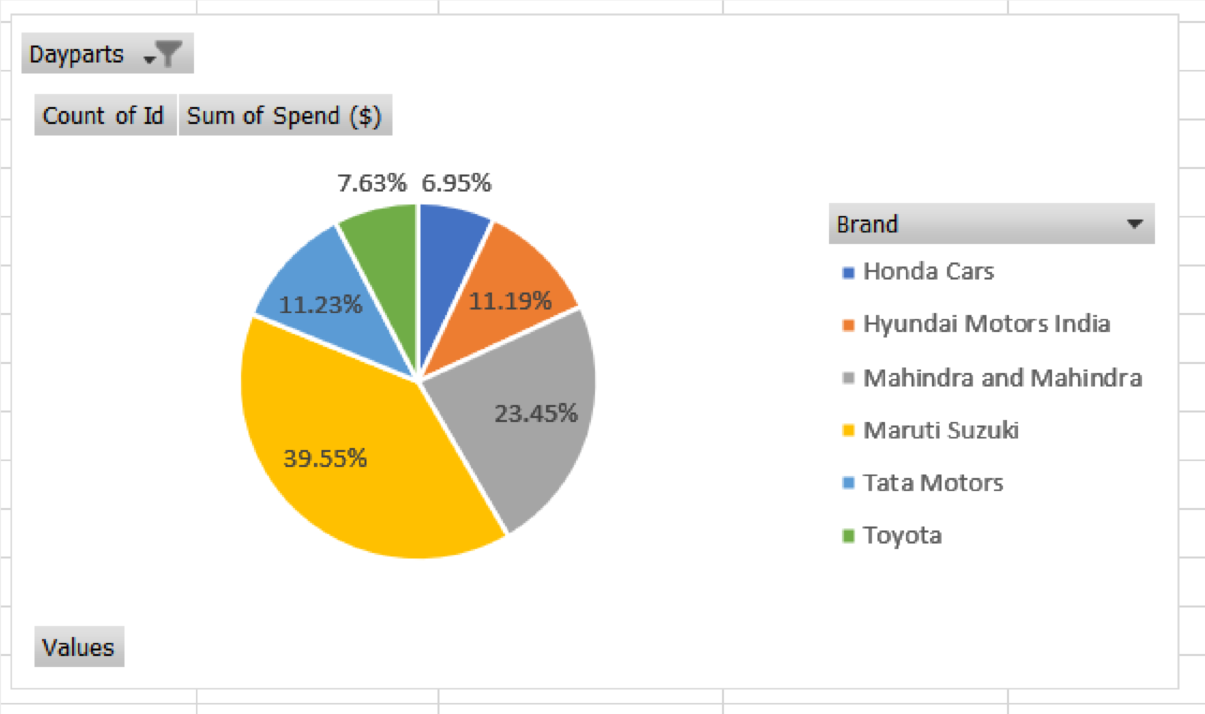
For Overnight:



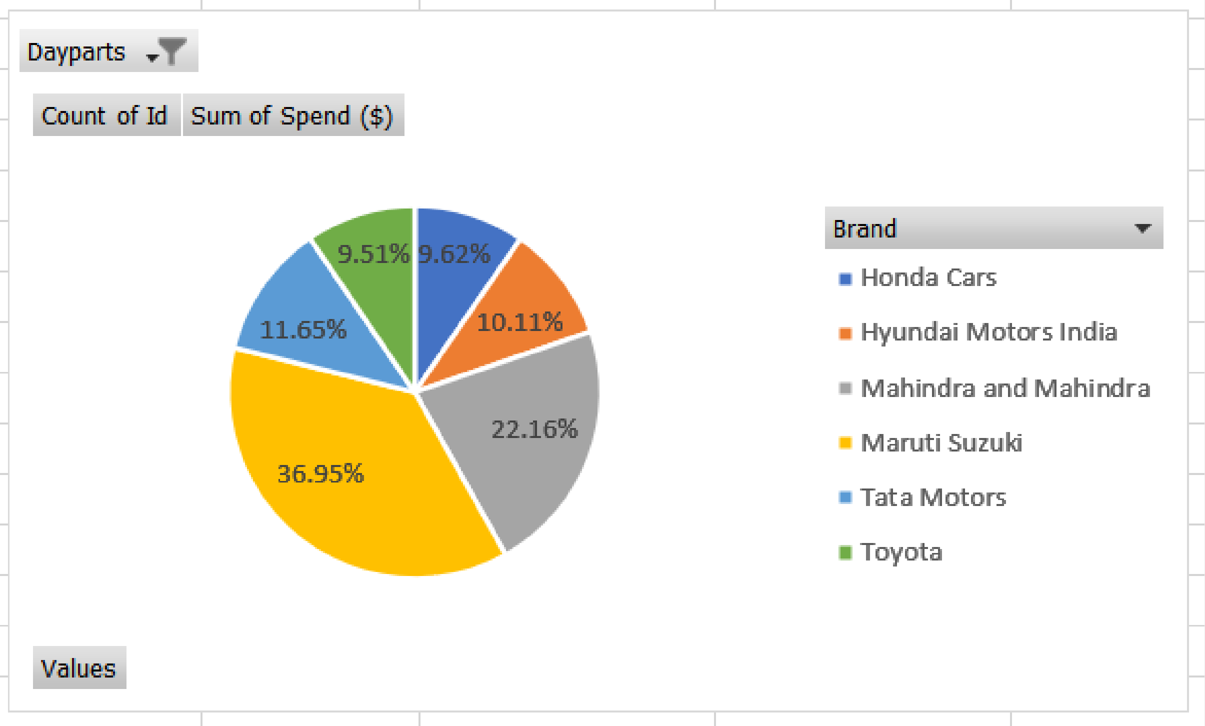
For Prime Access:



For Prime Time:

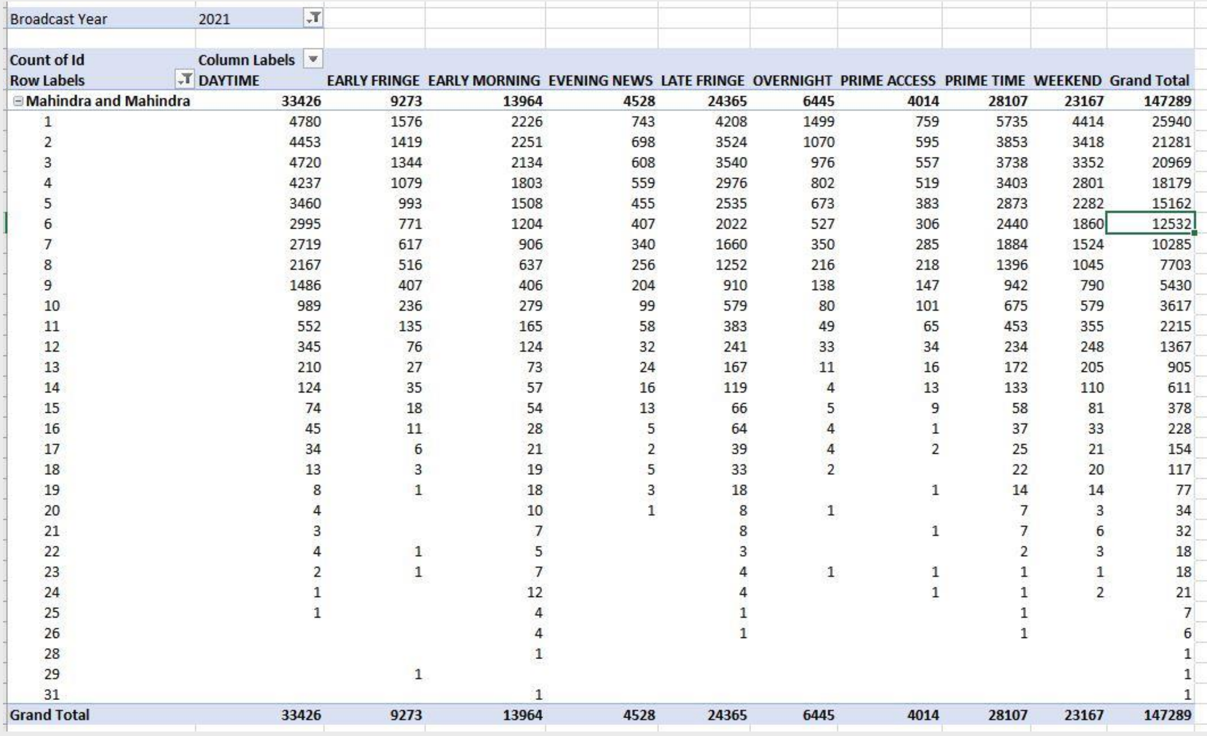


For Weekends:



4. Mahindra and Mahindra Ads:

We have to help Mahindra and Mahindra to conduct a digital ad campaign in January, February and March of 2022. So, we have to provide them with a media plan that must attract so many viewers. According to the data of 2021 a pivot table using the data of the manufacturer Mahindra & Mahindra is created as below,



Broadcast Year	2021									
Count of Id	Column Labels									
Row Labels	DAYTIME	EARLY FRINGE	EARLY MORNING	EVENING NEWS	LATE FRINGE	OVERNIGHT	PRIME ACCESS	PRIME TIME	WEEKEND	Grand Total
⊟ Mahindra and Mahindra	33426	9273	13964	4528	24365	6445	4014	28107	23167	147289
1	4780	1576	2226	743	4208	1499	759	5735	4414	25940
2	4453	1419	2251	698	3524	1070	595	3853	3418	21281
3	4720	1344	2134	608	3540	976	557	3738	3352	20969
4	4237	1079	1803	559	2976	802	519	3403	2801	18179
5	3460	993	1508	455	2535	673	383	2873	2282	15162
6	2995	771	1204	407	2022	527	306	2440	1860	12532
7	2719	617	906	340	1660	350	285	1884	1524	10285
8	2167	516	637	256	1252	216	218	1396	1045	7703
9	1486	407	406	204	910	138	147	942	790	5430
10	989	236	279	99	579	80	101	675	579	3617
11	552	135	165	58	383	49	65	453	355	2215
12	345	76	124	32	241	33	34	234	248	1367
13	210	27	73	24	167	11	16	172	205	905
14	124	35	57	16	119	4	13	133	110	611
15	74	18	54	13	66	5	9	58	81	378
16	45	11	28	5	64	4	1	37	33	228
17	34	6	21	2	39	4	2	25	21	154
18	13	3	19	5	33	2		22	20	117
19	8	1	18	3	18		1	14	14	77
20	4		10	1	8	1		7	3	34
21	3		7		8		1	7	6	32
22	4	1	5		3			2	3	18
23	2	1	7		4	1	1	1	1	18
24	1		12		4		1	1	2	21
25	1		4		1			1		7
26			4		1			1		6
28			1							1
29		1								1
31			1							1
Grand Total	33426	9273	13964	4528	24365	6445	4014	28107	23167	147289

The above data is for year 2021 showing all the data of Mahindra and Mahindra. It clearly can be seen that they have higher ads in pod position 1 and that too in daytime. So, we can say that maximum viewers are from daytime as seen above. So, this is the best plan they can follow in the future to increase their sales.

ABC Call Volume Trend Analysis

Project Description:

In this dataset, analysis is done on a dataset having information on calls for a ABC Company. Data includes Agent_Name, Agent_ID, Queue_Time [duration for which customer have to wait before they get connected to an agent], Time [time at which call was made by customer in a day], Time_Bucket [for easiness we have also provided you with the time bucket], Duration [duration for which a customer and executives are on call, Call_Seconds [for simplicity we have also converted those time into seconds], call status (Abandon, answered, transferred).

Nowadays we are getting a bunch of phone calls from Banks, Insurance companies, or from other organizations to invest money or for offering loans. So, if we need any customer service support to ensure this we can ask for their help by calling them. There are a lot of analytic ways to do the analysis of these call trends. Here, we are provided with a dataset of a Customer Experience (CX) Inbound calling team for 23 days.

Approach:

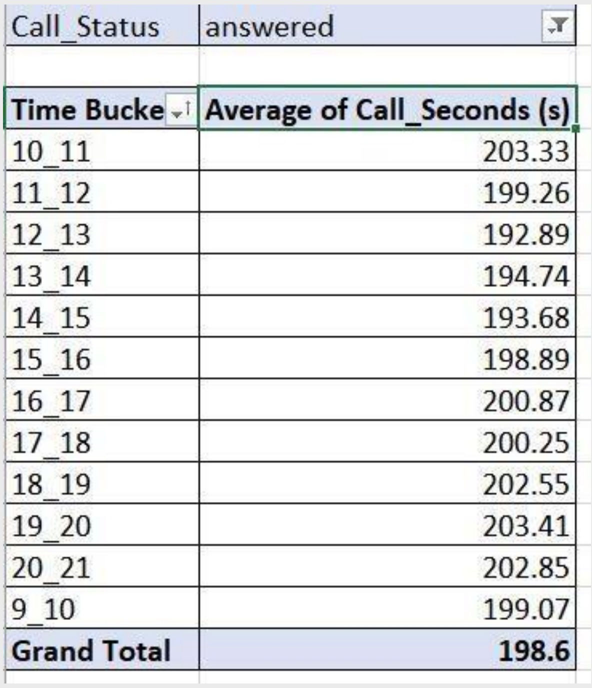
The first approach in this project has been data cleaning so that the data is perfectly operable upon. After that various analysis has been done to find the different solutions to questions according to needs.

Tech-Stack Used:

In this project, Microsoft - Excel was used to carry out all the analysis and cleaning part in the project. MS- Word was also used for making the final report and pdf.

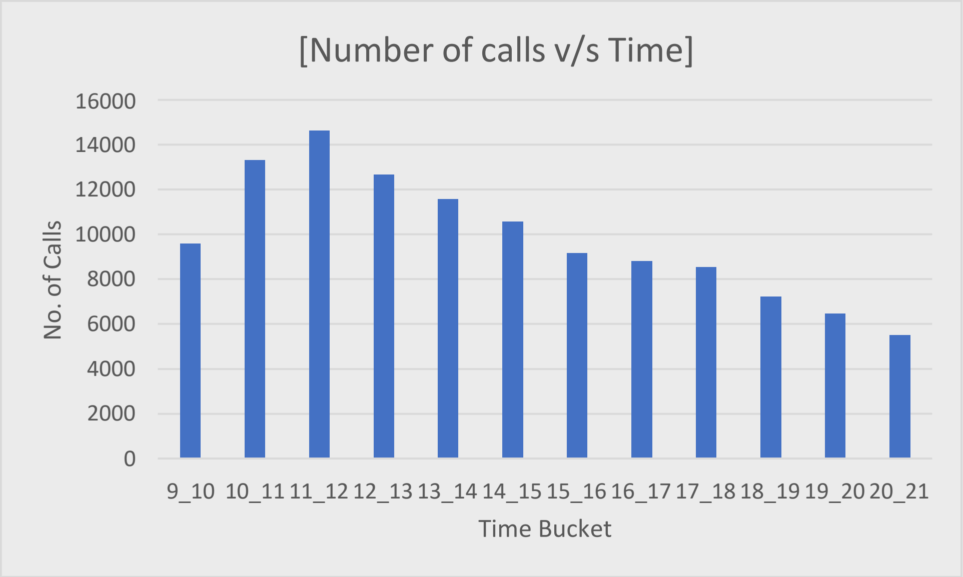
Insights:

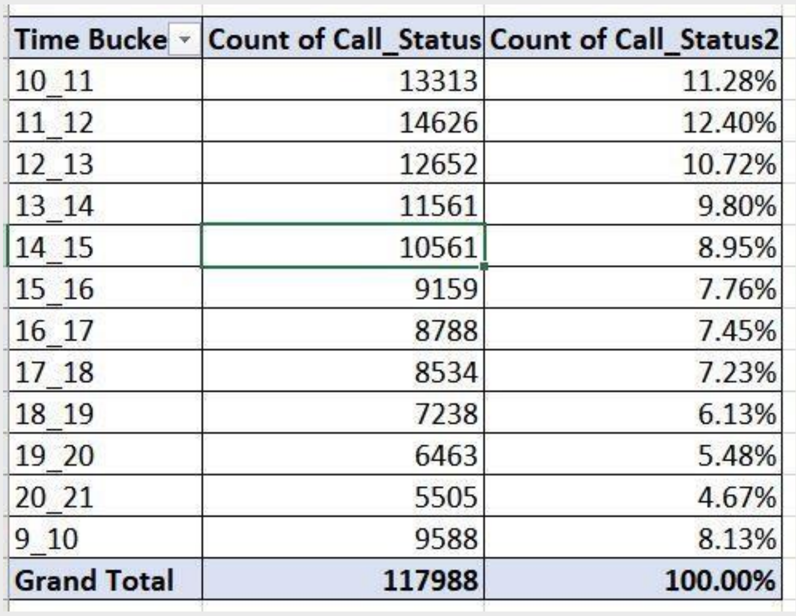
1. Find the average call time duration for all incoming calls received by agents in each time bucket.



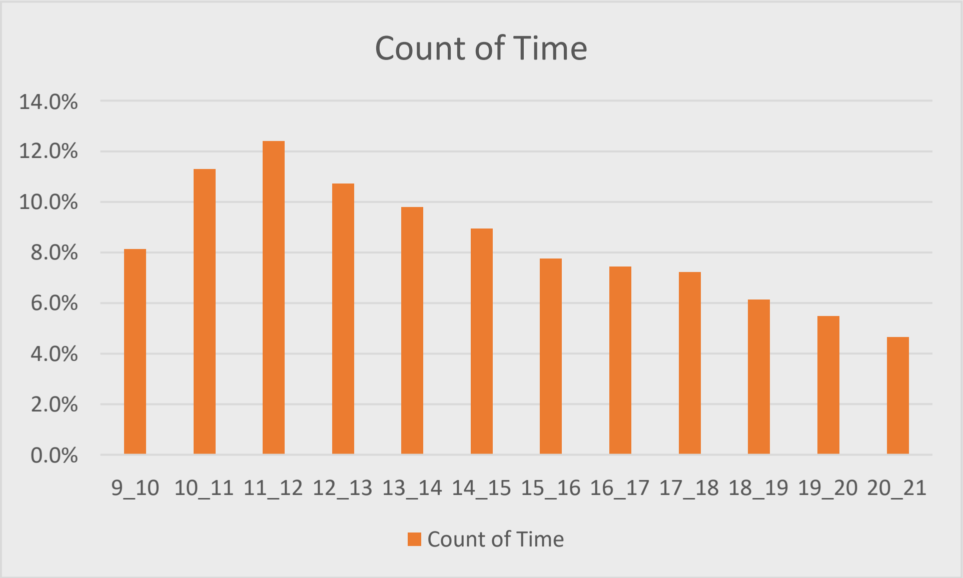
Call_Status	answered
Time Bucke	**Average of Call_Seconds (s)**
10_11	203.33
11_12	199.26
12_13	192.89
13_14	194.74
14_15	193.68
15_16	198.89
16_17	200.87
17_18	200.25
18_19	202.55
19_20	203.41
20_21	202.85
9_10	199.07
Grand Total	**198.6**

2. Show the total volume / number of calls coming in via charts or graphs:



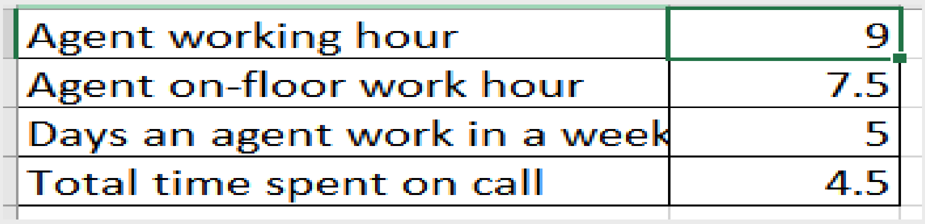


Time Bucke	Count of Call_Status	Count of Call_Status2
10_11	13313	11.28%
11_12	14626	12.40%
12_13	12652	10.72%
13_14	11561	9.80%
14_15	10561	8.95%
15_16	9159	7.76%
16_17	8788	7.45%
17_18	8534	7.23%
18_19	7238	6.13%
19_20	6463	5.48%
20_21	5505	4.67%
9_10	9588	8.13%
Grand Total	**117988**	**100.00%**

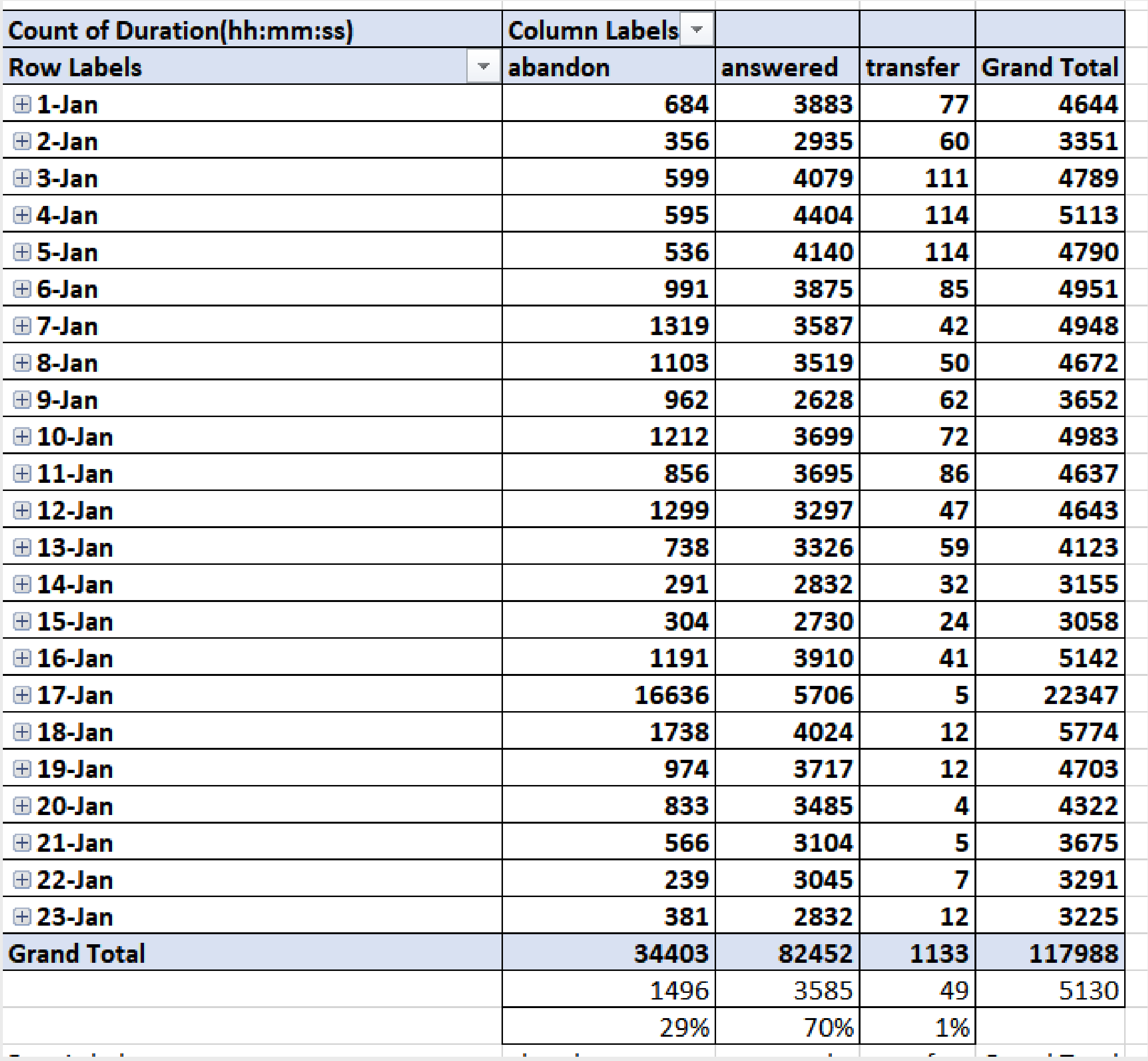


3. As you can see the current abandon rate is approximately 30%,propose a manpower plan required during each time bucket (between 9am - 9pm) to reduce the abandon rate to 10%.

Here, some details have been found out for further answering the question:

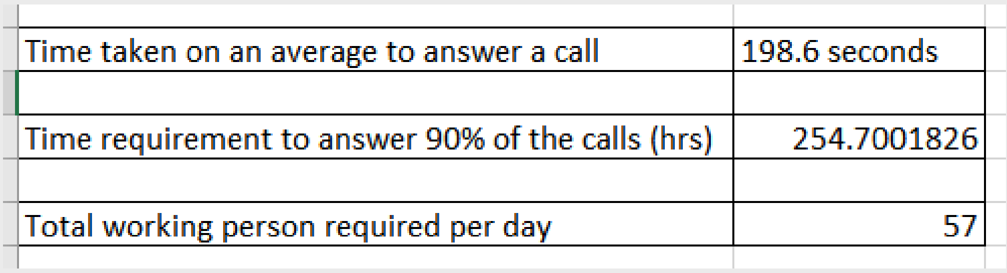


Agent working hour	9
Agent on-floor work hour	7.5
Days an agent work in a week	5
Total time spent on call	4.5



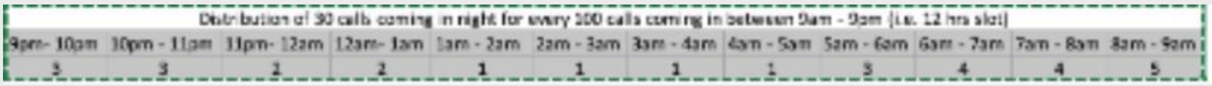
Count of Duration(hh:mm:ss)	Column Labels			
Row Labels	abandon	answered	transfer	Grand Total
⊞ 1-Jan	684	3883	77	4644
⊞ 2-Jan	356	2935	60	3351
⊞ 3-Jan	599	4079	111	4789
⊞ 4-Jan	595	4404	114	5113
⊞ 5-Jan	536	4140	114	4790
⊞ 6-Jan	991	3875	85	4951
⊞ 7-Jan	1319	3587	42	4948
⊞ 8-Jan	1103	3519	50	4672
⊞ 9-Jan	962	2628	62	3652
⊞ 10-Jan	1212	3699	72	4983
⊞ 11-Jan	856	3695	86	4637
⊞ 12-Jan	1299	3297	47	4643
⊞ 13-Jan	738	3326	59	4123
⊞ 14-Jan	291	2832	32	3155
⊞ 15-Jan	304	2730	24	3058
⊞ 16-Jan	1191	3910	41	5142
⊞ 17-Jan	16636	5706	5	22347
⊞ 18-Jan	1738	4024	12	5774
⊞ 19-Jan	974	3717	12	4703
⊞ 20-Jan	833	3485	4	4322
⊞ 21-Jan	566	3104	5	3675
⊞ 22-Jan	239	3045	7	3291
⊞ 23-Jan	381	2832	12	3225
Grand Total	34403	82452	1133	117988
	1496	3585	49	5130
	29%	70%	1%	

Therefore,



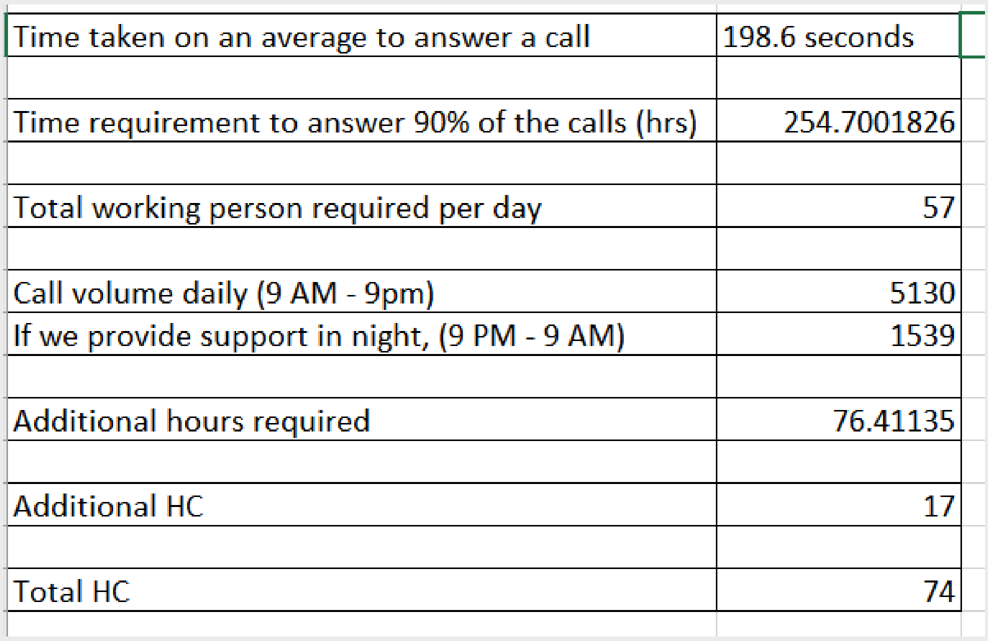
Time taken on an average to answer a call	198.6 seconds
Time requirement to answer 90% of the calls (hrs)	254.7001826
Total working person required per day	57

4. Let’s say customers also call this ABC insurance company at night but don't get an answer as there are no agents to answer. This creates a bad customer experience for this insurance company. So, suppose every 100 calls that customer made during 9 am to 9 pm, customer also made 30 calls in night between 9 pm to 9 am and the distribution is given as,

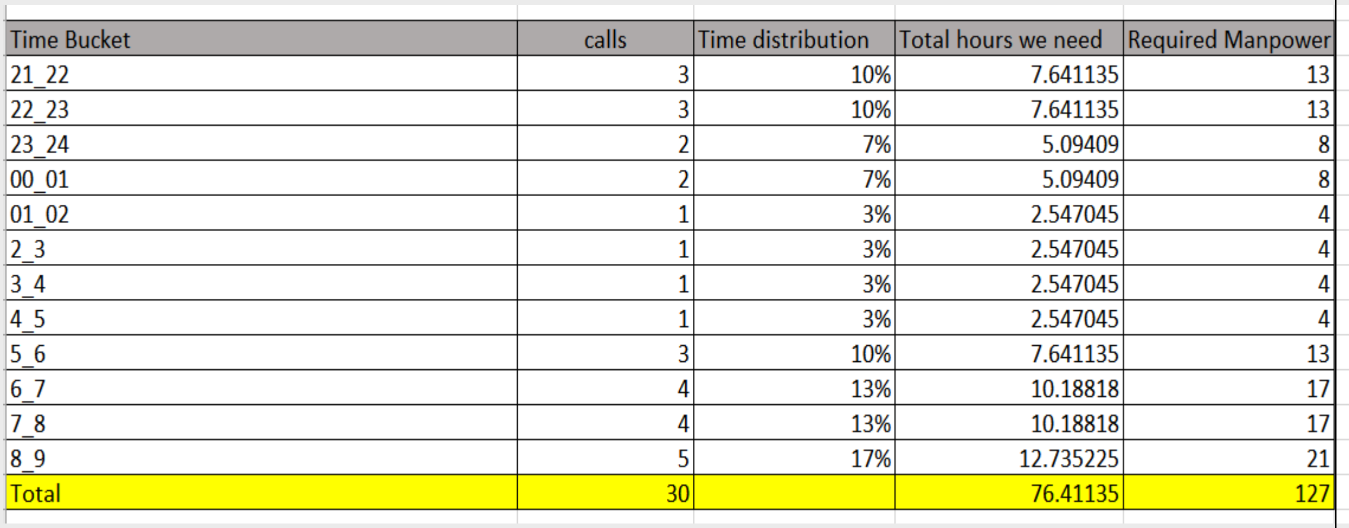


Distribution of 30 calls coming in night for every 100 calls coming in between 9am - 9pm (i.e. 12 hrs slot)											
9pm- 10pm	10pm - 11pm	11pm- 12am	12am- 1am	1am - 2am	2am - 3am	3am - 4am	4am - 5am	5am - 6am	6am - 7am	7am - 8am	8am - 9am
3	3	2	2	1	1	1	1	3	4	4	5

Now, prepare a manpower plan required during each time bucket in a day. Maximum abandon rate assumption would be the same 10%.



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	calls	Time distribution	Total hours we need	Required Manpower
21_22	3	10%	7.641135	13
22_23	3	10%	7.641135	13
23_24	2	7%	5.09409	8
00_01	2	7%	5.09409	8
01_02	1	3%	2.547045	4
2_3	1	3%	2.547045	4
3_4	1	3%	2.547045	4
4_5	1	3%	2.547045	4
5_6	3	10%	7.641135	13
6_7	4	13%	10.18818	17
7_8	4	13%	10.18818	17
8_9	5	17%	12.735225	21
Total	30		76.41135	127

Global Store Analysis

Project Description:

A large dataset containing information on a global store has been analyzed using Power BI and various queries regarding performance of the store from the year 2012-15 has been found out.

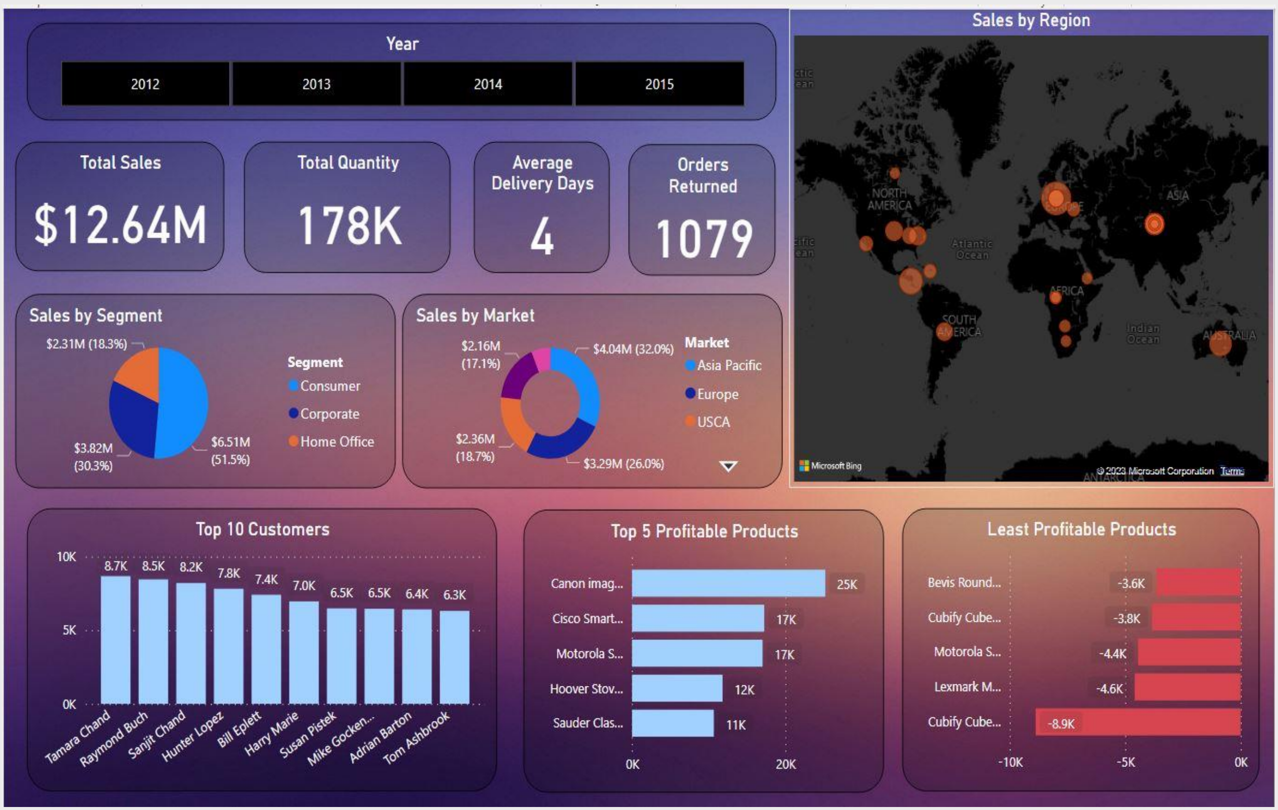
Approach:

The dataset was first cleaned using Excel and unwanted data such as null, unwanted columns were deleted. After that the dataset was loaded in Power BI and then further analysis has been carried out. While carrying out the analysis, new columns were also made using power pivot and DAX functions as per the requirements.

Tech Stack Used:

The tech stacks which were used for completing this project were Microsoft Excel, Microsoft Power BI, Microsoft Word (for making a report).

Insights:



Results:

By performing all the above stated projects, my hands on experience on many tools like MS: Excel, MySQL, Python, Power BI were increased and also, I got to know about many functionalities of all the tech stacks and many more useful information about how a data analyst does his work and what analysis does he carry out and how he helps the companies to gain profit from the data.