DATA ANALYSIS PORTFOLIO

PREPARED BY
HUMAIRA KHAN

PROFESSIONAL BACKGROUND

I am a third-year cybersecurity student pursuing bachelors as my degree from the University of Pune. I have secured 9.0 CGPA and am well-versed with Python, SQL, NoSQL, Statistics, Excel, Machine Learning and Data Analysis.

I have worked on several projects such as steganography using python, breast cancer prediction using SVM algorithm and movie recommendation system.

Coming to my data analysis projects, I have analyzed census data, analyzed NASA planetary budget exploration using SQL, uber trips analysis.

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 my theoretical knowledge in a practical way. And I believe by putting significant efforts, I will learn.

Table Of Contents

DATA ANALYTICS PROCESS
ORDERING FOOD FROM ONLINE APPS

PLAN: SO, FIRST WE WILL DECIDE FROM WHAT APP WE WILL ORDER FOOD FROM, IS IT ZOMATO OR SWIGGY? WHAT WE FEEL LIKE EATING, INDIAN OR CHINESE? SWEET OR SPICY?

PREPARE: NEXT I NEED TO HOW MUCH AM I WILLING TO SPEND AND HOW TO GET THAT MONEY.

PROCESS: THEN I NEED TO CHECK HOW MUCH I WANT FROM THE DATA. IF I WANT TO EAT INDIAN FOOD, WHAT EXACTLY AND HOW MUCH AMOUNT I WANT. 2 CHAPATIS W SABZI, WILL IT BE ENOUGH?

ANALYZE: BUT WAIT. I AM DIET CONSCIOUS. WILL MY ORDER SYNCHRONIZE WITH MY CALORIES INTAKE? WILL IT PROVIDE ME WITH THE PERFECT AMOUNT OF NUTRIENTS?

SHARE: NOW THAT YOU HAVE ANALYSED AND DECIDED ON THE PERFECT ORDER, YOU WOULD WANT TO SHARE YOUR IDEA BY CLICKING ON WHAT YOU WANT TO BUY.

ACT: AND THEN YOU FINALLY ORDER THE FOOD AND LATER EAT IT.

INSTAGRAM USER ANALYTICS USING SQL

PROJECT DESCRIPTION:

THIS PROJECT IS ABOUT ANALYSING HOW USERS ENGANGE WITH A VERY POPULAR SOCIAL APP NAMED INSTAGRAM. WHICH USER HAS THE MOST NUMBER OF LIKES AND FOLLOWERS? WHICH USER IS BARELY ACTIVE? HOW MUCH TIME DOES AN AVERAGE USER SPENDS DAILY ON INSTAGRAM? WHICH DAY ARE PEOPLE MOST ACTIVE? TO ANSWER ALL THESE CURIOUS QUESTIONS,WE WILL BE USING SQL AND ANALYSE EVERYTHING ABOUT IT BY THE INSTAGRAM DATA WE ALREADY HAVE.

APPROACH:

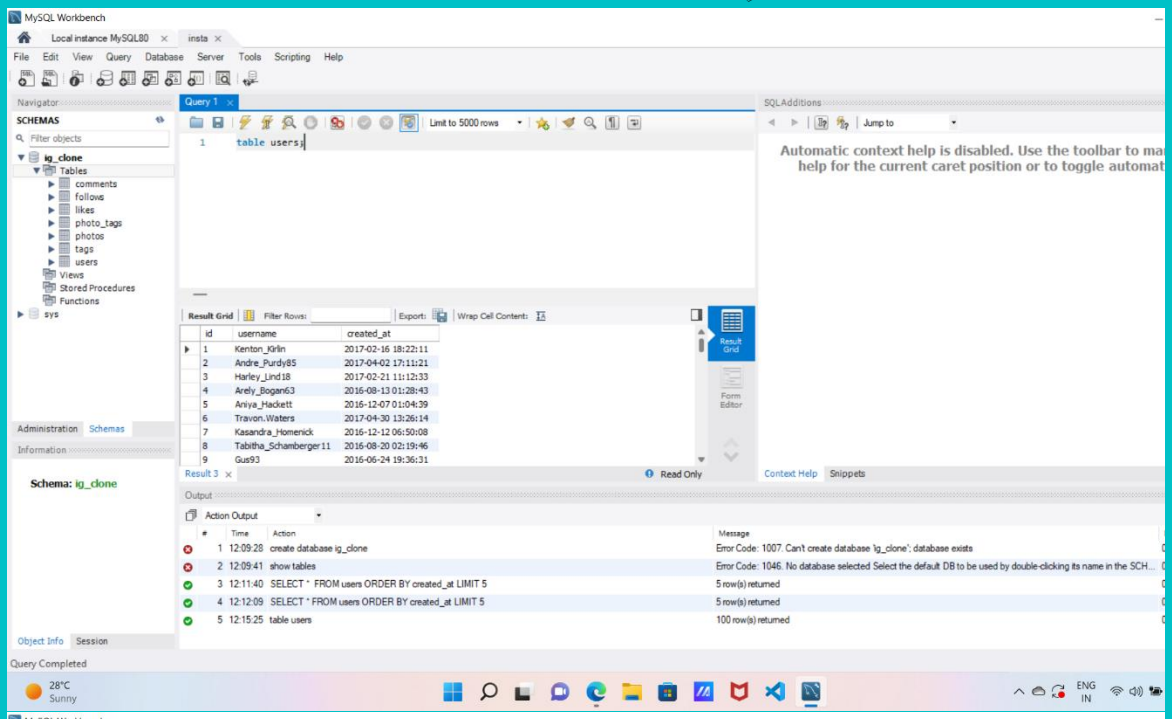
AFTER INSTALLATION AND EXECUTING THE DATABASE NAMED IG_CLONE, I FIRST USED THE SHOW TABLE COMMANDS AND THEN THE SHOW COLUMNS COMMAND TO UNDERSTAND ABOUT ALL THE TABLE,COLUMNS AND THE DATA INSIDE IT. THEN READING THE SQL QUERY QUESTIONS AND UNDERSTANDING WHAT ALL DATA WE EXACTLY NEED, I STARTED QUERYING ABOUT THE SAME.

INSIGHTS:

LETS START WITH THE QUERIES IN ORDER AND WHAT ALL WE LEARNED BY IT.

MARKETING QUERIES

THE 5 OLDEST USERS.

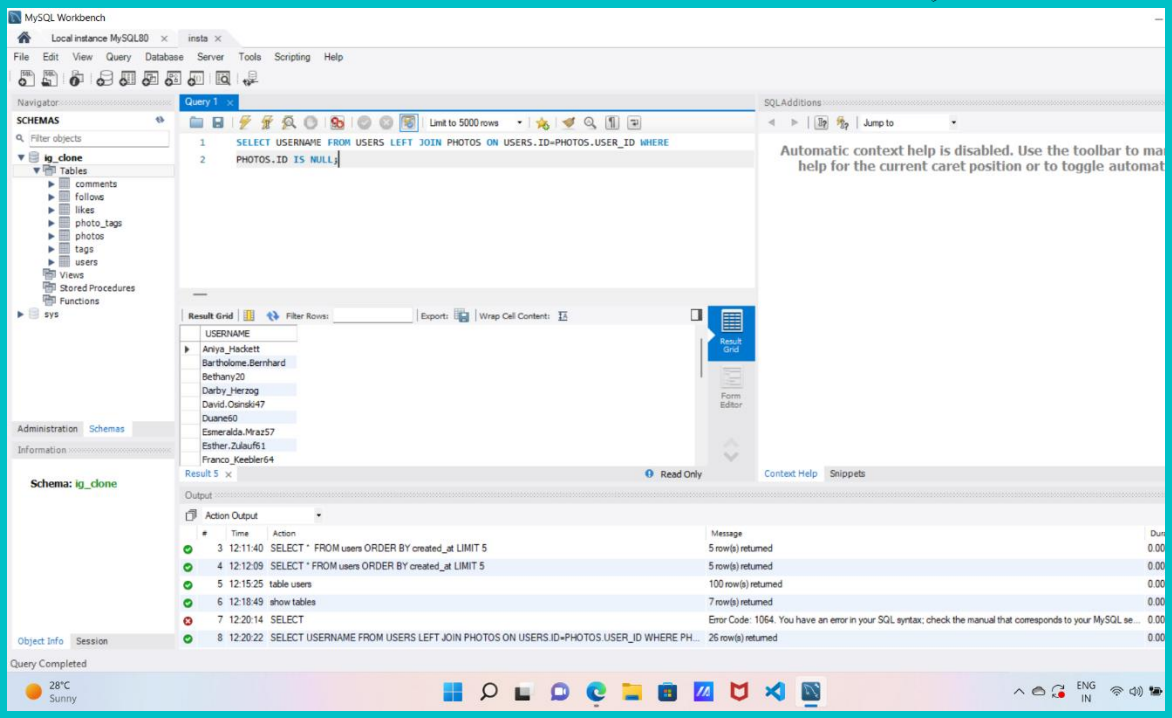


A very straight forward query which we executed by selecting all columns from table users and ordering it by the date the users created their account and because we need 5 oldest, we used the limit function.

These are loyal to instagram app and hence we can reward them for the same.

2)

SECOND QUERY WAS ABOUT KNOWING THE INACTIVE PEOPLE WHO HAVENT POSTED A SINGLE PICTURE.

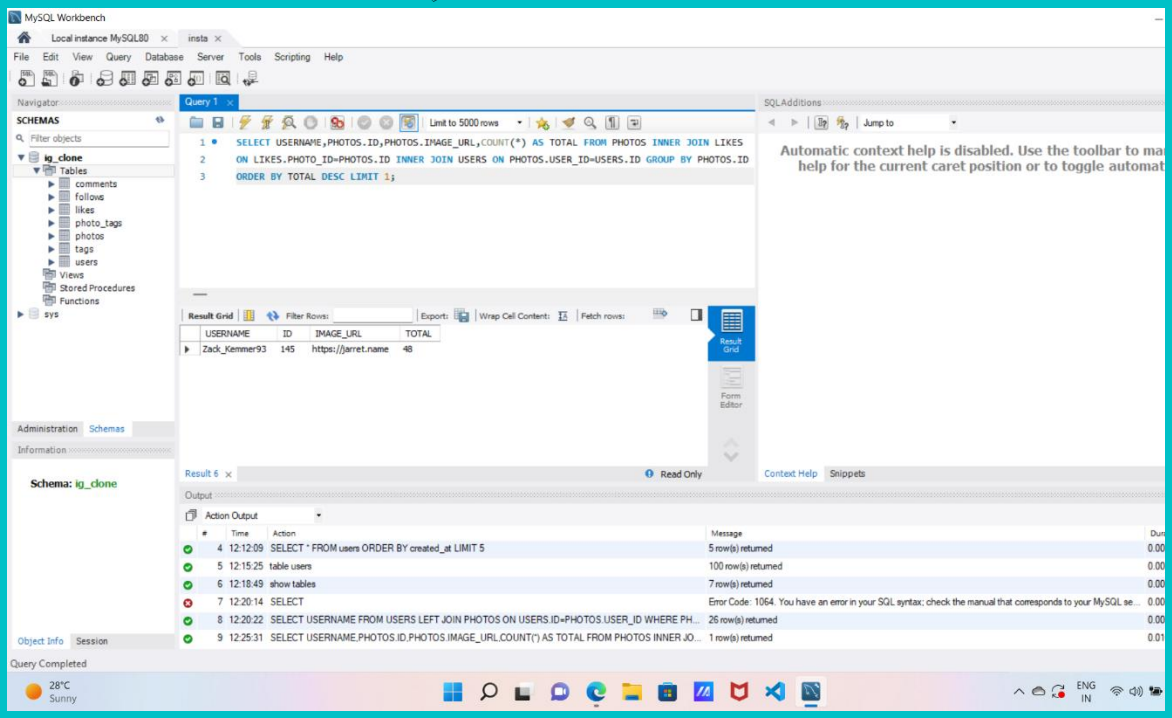


HERE WE NEEDED THE LEFT JOIN AS WE NEEDED THE USERS FROM THE USERS TABLE AND THE MATCHED RECORDS FROM PHOTOS TABLE. USER ID IS IN BOTH THE TABLES HENCE WE USED USER ID TO JOIN THESE TABLES. WHEREVER THE PHOTOS.ID IS NULL I.E NO PICTURE, IT RETURNS THE USERNAME.

WE LEARNED THERE ARE PLENTY OF PEOPLE WHO STILL HAVENT POSTED A SINGLE PICTURE, COULD BE BECAUSE OF THEIR BUSY LIFES, COULD BE BECAUSE THEY MIGHT HAVE SOME PROBLEM WITH THIS APP, SO WE CAN REMIND THEM TO POST.

3)

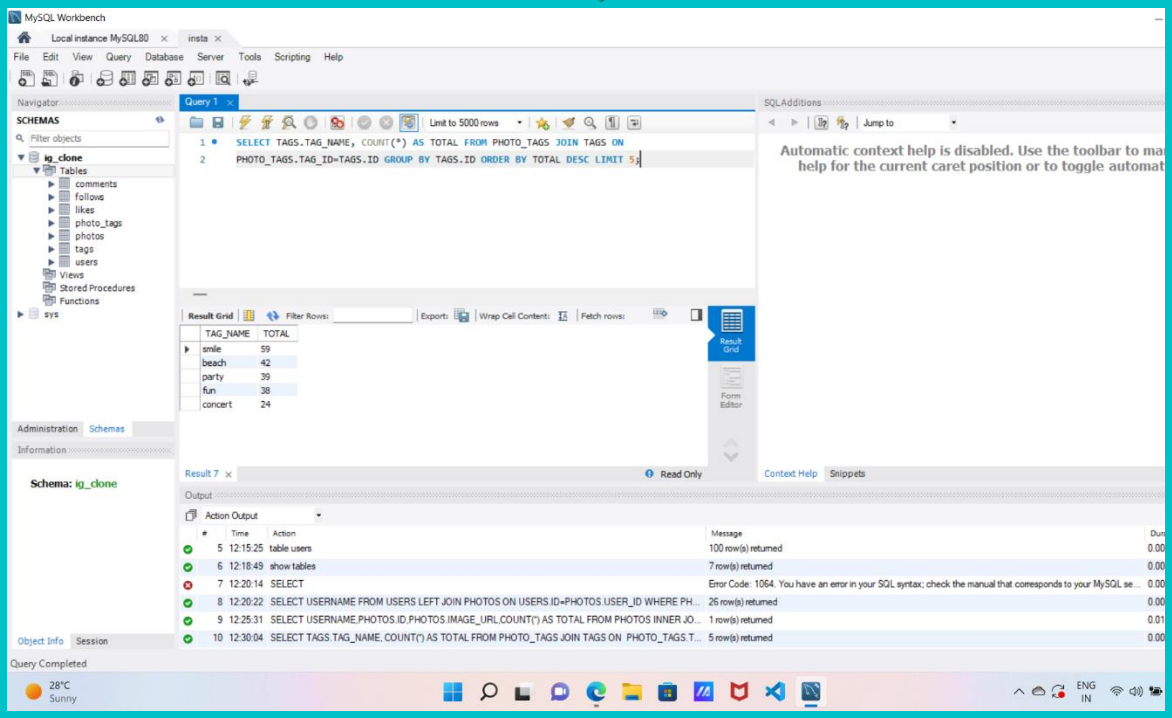
THE NEXT QUERY WAS TO FIND WHICH USER RECEIVED THE HIGHEST NUMBER OF LIKES.



WE USED THE COUNT FUNCTION TO COUNT THE LIKES AND USED THE RELATED COLUMN I.E PHOTO ID HERE TO JOIN TABLES, LIKES AND USERS TO PHOTOS. GROUP BY WAS USED TO GROUP ROWS THAT HAVE THE SAME VALUES AND ORDER BY THE TOTAL NUMBER OF LIKES.
HERE, ZACK RECEIVED THE HIGHEST NUMBER OF LIKES WITH A TOTAL OF 48.

4)

THE 4TH QUERY WAS TO POST THE FIRST FIVE POPULAR HASHTAGS SO THAT WE CAN REACH MOST PEOPLE.

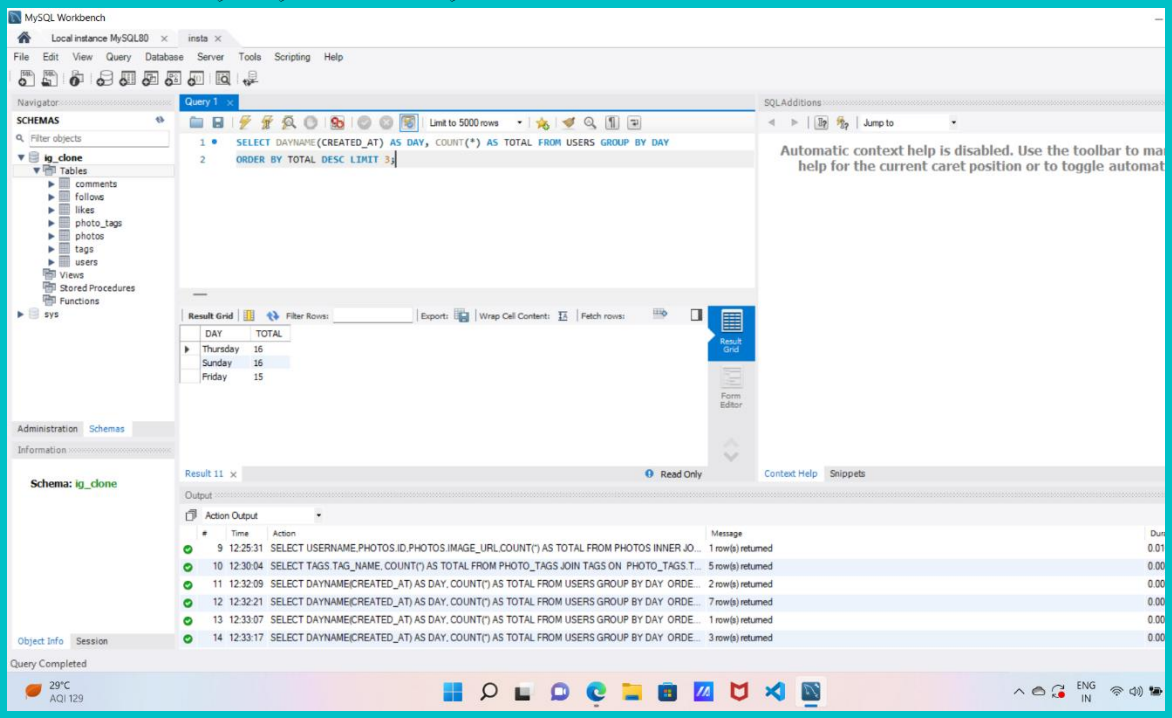


SIMILAR TO THE ABOVE QUERY, WE SELECTED THE COUNT OF TAG NAMES, USED TAG ID TO JOIN PHOTOS AND TAGS TABLE , GROUP BY TO GROUP ROWS WITH SAME VALUES I.E WITH SAME TAGS.ID AND ORDER BY TOTAL I.E COUNT OF TAGS NAME IN DESCENDING ORDER SO THE HIGHEST NUMBER COMES ABOVE.

ALL THE HASHTAGS SOUND PRETTY HAPPY, NO WONDER THEY ARE THE MOST USED ONES.

5)

5TH QUERY WAS ABOUT KNOWING WHICH DAY WOULD BE THE BEST TO LAUNCH AD CAMPAIGNS. WHAT DAY DP MOST USERS REGISTER ON THIS APP?

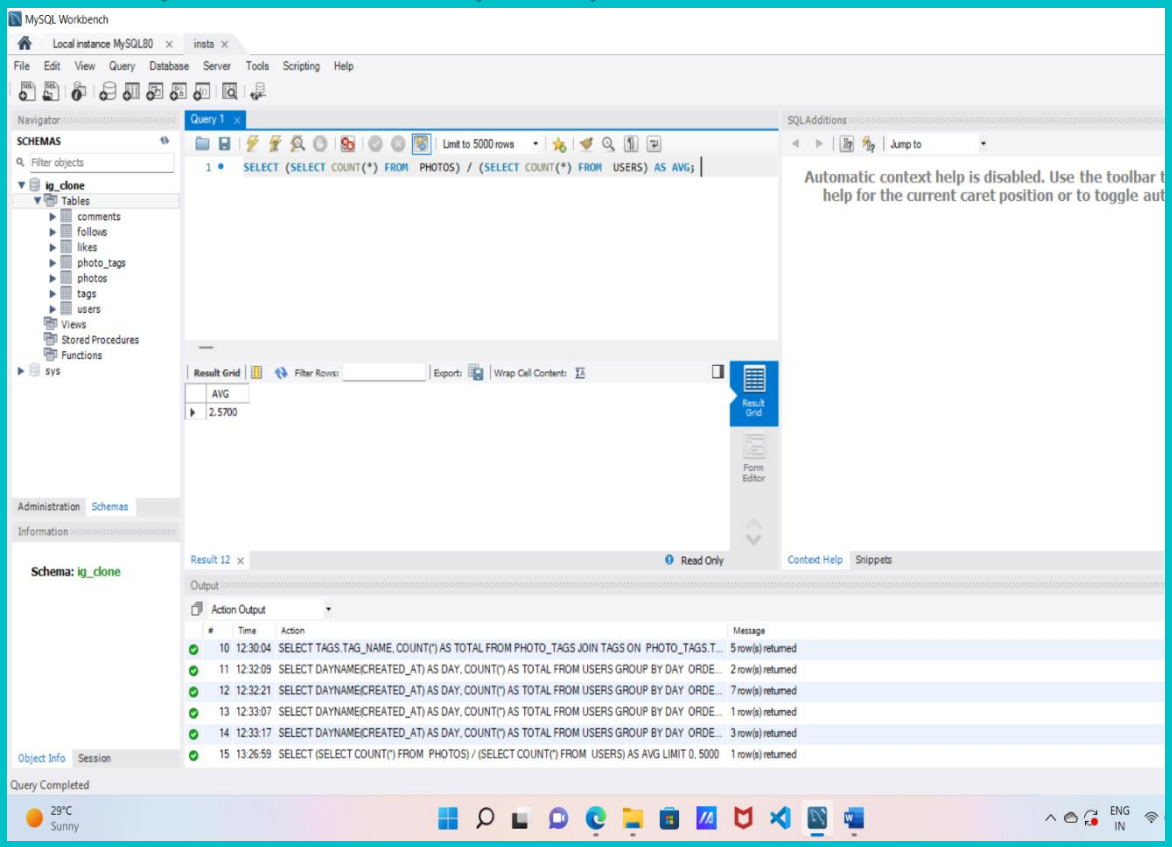


WE SELECT THE DAYNAME OF WHEN THE USER CREATED HIS ACCOUNT(USING CREATED_AT), COUNT FUNCTION TO COUNT THE DAYS, GROUP BY TO GROUP ROWS WITH SAME VALUES I.E GROUPING DAYS AND ORDER BY TOTAL IN DESCENDING. WE PRINTED TOP THREE ROWS AND GOT TO KNOW BOTH THURSDAY AND SUNDAY WE HAVE HIGHEST NUMBER OF REGISTRATIONS , 16.

INVESTOR METRICS

1)

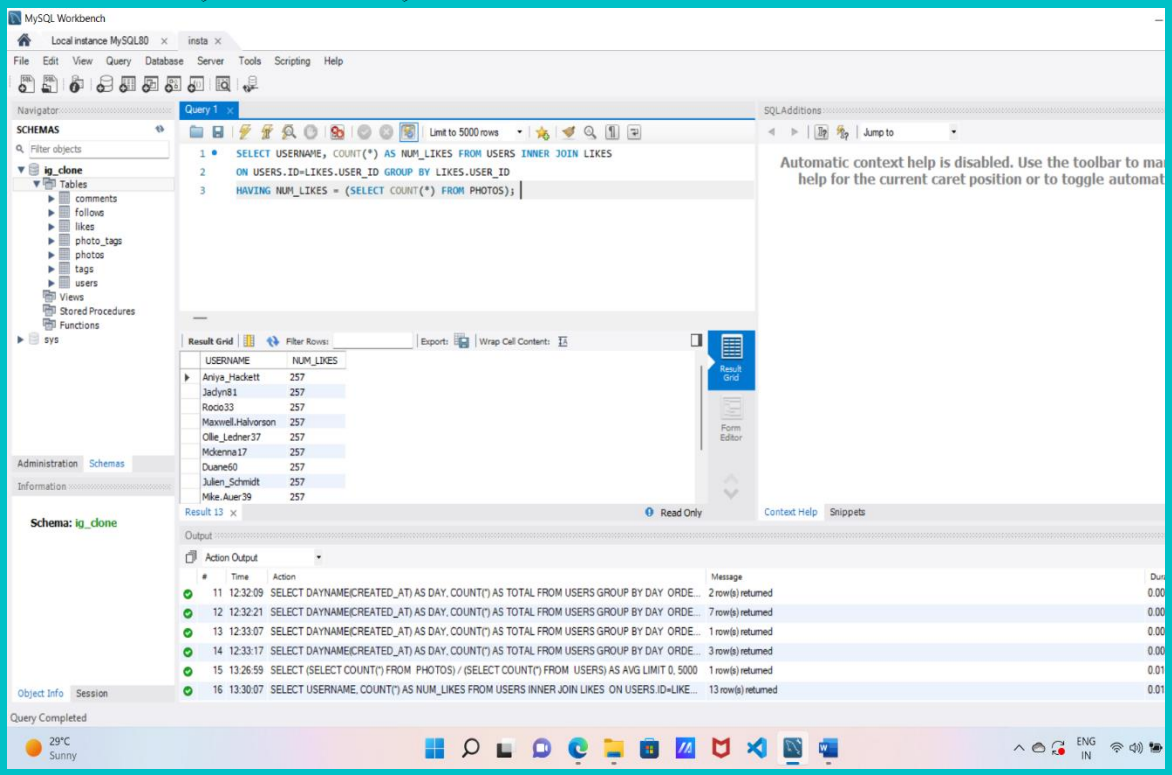
OUR FIRST QUERY HERE WAS TO FIND THE AVERAGE NUMBER OF PHOTOS PER USER



WE DO THIS BY USING THE FORMULA FOR AVERAGE I.E COUNT OF PHOTOS DIVIDED BY COUNT OF USERS.
THE AVERAGE COMES OUT TO BE 25700

2)

THE LAST QUERY WAS TO FIND THE BOT ACCOUNTS, WE KNOW THIS BY FINDING WHO ALL HAS LIKED ALL PHOTOS SINCE NORMAL USERS CANNOT DO THIS.



WE USE COUNT TO FUNCTION TO COUNT THE NUMBER OF LIKES FROM USERS WHICH IS EQUAL TO THE TOTAL NUMBER OF PHOTOS. USING USER_ID TO JOIN TABLES USERS AND LIKES, WE USE A SUB QUERY TO FIND OUT THE TOTAL NUMBER OF PHOTOS.

AS WE CAN SEE THERE ARE PLENTY OF BOT ACCOUNTS WHICH NEEDS TO BE ERADICATED.

OPERATION ANALYTICS AND INVESTIGATING METRIC SPIKE

PROJECT DESCRIPTION

OPERATION ANALYTICS IS THE ANALYSIS DONE FOR THE COMPLETE END TO END OPERATIONS OF A COMPANY. WITH THE HELP OF THIS, THE COMPANY THEN FINDS THE AREAS ON WHICH IT MUST IMPROVE UPON, WORKING CLOSELY WITH THE OPS TEAM, SUPPORT TEAM, MARKETING TEAM AND HELP THEM DERIVE INSIGHTS OUT OF THE DATA THEY COLLECT. BEING ONE OF THE MOST IMPORTANT PARTS OF A COMPANY, THIS KIND OF ANALYSIS IS FURTHER USED TO PREDICT THE OVERALL GROWTH OR DECLINE OF A COMPANY'S FORTUNE. IT ALSO USED FOR BETTER AUTOMATION, BETTER UNDERSTANDING BETWEEN CROSS-FUNCTIONAL TEAMS, AND MORE EFFECTIVE WORKFLOWS.
INVESTIGATING METRIC SPIKE IS ALSO AN IMPORTANT PART OF OPERATION ANALYTICS AS BEING A DATA ANALYST YOU MUST BE ABLE TO UNDERSTAND OR MAKE OTHER TEAMS UNDERSTAND QUESTIONS LIKE- WHY IS THERE A DIP IN DAILY ENGAGEMENT? WHY HAVE SALES TAKEN A DIP? ETC. QUESTIONS LIKE THESE MUST BE ANSWERED DAILY AND FOR THAT ITS VERY IMPORTANT TO INVESTIGATE METRIC SPIKE.

APPROACH

I GATHERED INFORMATION FROM THE DESCRIPTION. TOOK A LOOK AT THE DATASET TO UNDERSTAND ABOUT THE
TABLES AND THEN STARTED WORKING ON THIS PROJECT.
THEN I USED MYSQL WORKBENCH TO CREATE TABLES AND FINALLY EXECUTED MY QUERIES.

A FEW OF THEM HAD ERRORS BUT I REVISED MY QUERY AND EVENTUALLY GOT THEM RIGHT.

INSIGHTS

THE KNOWLEDGE THAT I GAINED FROM THIS PROJECT IS HOW TO CREATE THE DATABASE ,HOW TO USE THE DATABASE , HOW CAN WE PERFORM SQL OPERATION BASED ON SCENARIOS AND HOW WE CAN USE THIS DATA TO IMPROVE THE EXPERIENCE ALTOGETHER WHILE HELPING THE BUSINESS GROW.

JOB DATA

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?

DS	THROUGHPUT
30-11-2020	144
29-11-2020	180
28-11-2020	218.18
27-11-2020	34.62
25-11-2020	80

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?

DS	THROUGHPUT_FOR_7 DAYS
25-11-2020	0.02
27-11-2020	0.01
28-11-2020	0.02
29-11-2020	0.02
30-11-2020	0.03

C.Percentage share of each language: Share of each language for different contents.

Your task: Calculate the percentage share of each language in the last 30 days?

language	percentage_of_jobs
persian	200
italian	100
french	100
hindi	100
arabic	100
persian	100
italian	50
french	50
hindi	50
arabic	50

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?

NO DUPLICATE DATA FOUND IN THE TABLE

Ds	job_id	actor_id	Event	language	time_spent	org

INVESTIGATING METRIC SPIKE

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?

DATE_TRUNC	WEEKLY_ACTIVE_USERS
28-04-2014 00:00	701
05-05-2014 00:00	1054
12-05-2014 00:00	1094
19-05-2014 00:00	1147
26-05-2014 00:00	1113
02-06-2014 00:00	1173
09-06-2014 00:00	1219
16-06-2014 00:00	1262
23-06-2014 00:00	1249
30-06-2014 00:00	1271
07-07-2014 00:00	1355
14-07-2014 00:00	1345
21-07-2014 00:00	1363
28-07-2014 00:00	1442
04-08-2014 00:00	1266
11-08-2014 00:00	1215

18-08-2014 00:00	1203
25-08-2014 00:00	1194

B.USER GROWTH: AMOUNT OF USERS GROWING OVER TIME FOR A PRODUCT.

YOUR TASK: CALCULATE THE USER GROWTH FOR PRODUCT?

DAY	ALL_USERS	ACTIVATED_USERS
01-01-2013 00:00	13	7
02-01-2013 00:00	11	7
03-01-2013 00:00	14	6
04-01-2013 00:00	11	1
05-01-2013 00:00	3	2
06-01-2013 00:00	4	3
07-01-2013 00:00	13	4
08-01-2013 00:00	13	2
09-01-2013 00:00	11	6
10-01-2013 00:00	12	6
11-01-2013 00:00	11	6
12-01-2013 00:00	4	3
13-01-2013 00:00	3	2
14-01-2013 00:00	13	8
15-01-2013 00:00	15	11
16-01-2013 00:00	14	7
17-01-2013 00:00	13	9
18-01-2013 00:00	14	10
19-01-2013 00:00	4	1
20-01-2013 00:00	4	1
21-01-2013 00:00	13	7
22-01-2013 00:00	13	5
23-01-2013 00:00	14	7
24-01-2013 00:00	14	5
25-01-2013 00:00	16	8
26-01-2013 00:00	3	3
27-01-2013 00:00	4	1
28-01-2013 00:00	14	7

29-01-2013 00:00	14	3
30-01-2013 00:00	14	6

C.Weekly Retention: Users getting retained weekly after signing-up for a product.

Your Task: Calculate the weekly retention of users-sign up cohort?

Week	Average age during week	10+ weeks	9 weeks	8 weeks	7 weeks	6 weeks	5 weeks	4 weeks	3 weeks
28-04-2014 00:00	124.0072389	701	0	0	0	0	0	0	0
05-05-2014 00:00	124.3816908	1054	0	0	0	0	0	0	0
12-05-2014 00:00	131.9386442	1094	0	0	0	0	0	0	0
19-05-2014 00:00	132.3266284	1147	0	0	0	0	0	0	0
26-05-2014 00:00	132.3453634	1113	0	0	0	0	0	0	0
02-06-2014 00:00	131.8311091	1173	0	0	0	0	0	0	0
09-06-2014 00:00	131.0425824	1219	0	0	0	0	0	0	0
16-06-2014 00:00	136.4805654	1255	0	0	0	0	0	0	0
23-06-2014 00:00	136.2789056	1034	210	0	0	0	0	0	0
30-06-2014 00:00	136.4192975	917	151	199	0	0	0	0	0
07-07-2014 00:00	135.8887505	899	100	130	223	0	0	0	0
14-07-2014 00:00	143.4488157	832	62	82	152	215	0	0	0
21-07-2014 00:00	141.702782	791	44	60	95	144	228	0	0
28-07-2014 00:00	144.0786604	805	30	43	83	91	155	234	0

04-08-2014 00:00	140.732238	678	24	34	52	52	82	154	189	0	0
11-08-2014 00:00	125.994310 1	562	19	33	39	33	59	94	126	0	0
18-08-2014 00:00	128.021718 1	522	15	26	26	19	40	64	69	59	0
25-08-2014 00:00	128.269810 4	474	15	14	23	20	31	47	48	73	2

D.WEEKLY ENGAGEMENT: TO MEASURE THE ACTIVENESS OF A USER.

MEASURING IF THE USER FINDS QUALITY IN PRODUCT/SERVICE_WEEKLY.

YOUR TASK: CALCULATE THE WEEKLY ENGAGEMENT PER DEVICE?

WEEK	WEEKLY_ACTIVE_USERS	COMPUTER	PHONE	TABLET
28-04-2014 00:00	701	347	211	103
05-05-2014 00:00	1054	590	335	176
12-05-2014 00:00	1094	632	352	191
19-05-2014 00:00	1147	647	385	181
26-05-2014 00:00	1113	605	379	176
02-06-2014 00:00	1173	663	388	197
09-06-2014 00:00	1219	671	397	195
16-06-2014 00:00	1262	685	400	227
23-06-2014 00:00	1249	706	395	210
30-06-2014 00:00	1271	676	424	218
07-07-2014 00:00	1355	735	434	227
14-07-2014 00:00	1345	762	421	218
21-07-2014 00:00	1363	767	436	218
28-07-2014 00:00	1442	798	414	241
04-08-2014 00:00	1266	770	354	166
11-08-2014 00:00	1215	722	326	153
18-08-2014 00:00	1203	754	311	145
25-08-2014 00:00	1194	725	312	150

D.EMAIL ENGAGEMENT: USERS ENGAGING WITH THE EMAIL SERVICE.

YOUR TASK: CALCULATE THE EMAIL ENGAGEMENT METRCS?

WEEK	WEEKLY_EMAILS	REENGAGEMENT_EMAIL	EMAIL_OPENS	EMAIL_C
14-07-2014 00:00	3499	226	1260	607
28-07-2014 00:00	3706	230	1386	633
19-05-2014 00:00	2733	179	995	498
05-05-2014 00:00	2602	164	919	434
11-08-2014 00:00	3897	224	1357	430
12-05-2014 00:00	2665	175	971	479
25-08-2014 00:00	4111	263	1533	493
28-04-2014 00:00	908	98	332	187
26-05-2014 00:00	2822	179	1026	453
07-07-2014 00:00	3399	214	1230	622
09-06-2014 00:00	3003	190	1070	533
16-06-2014 00:00	3105	234	1161	563
30-06-2014 00:00	3302	222	1168	559
21-07-2014 00:00	3592	206	1211	584
02-06-2014 00:00	2911	199	993	492
18-08-2014 00:00	4012	257	1421	487
23-06-2014 00:00	3207	187	1090	524
04-08-2014 00:00	3793	206	1336	432

HIRING PROCESS ANALYTICS

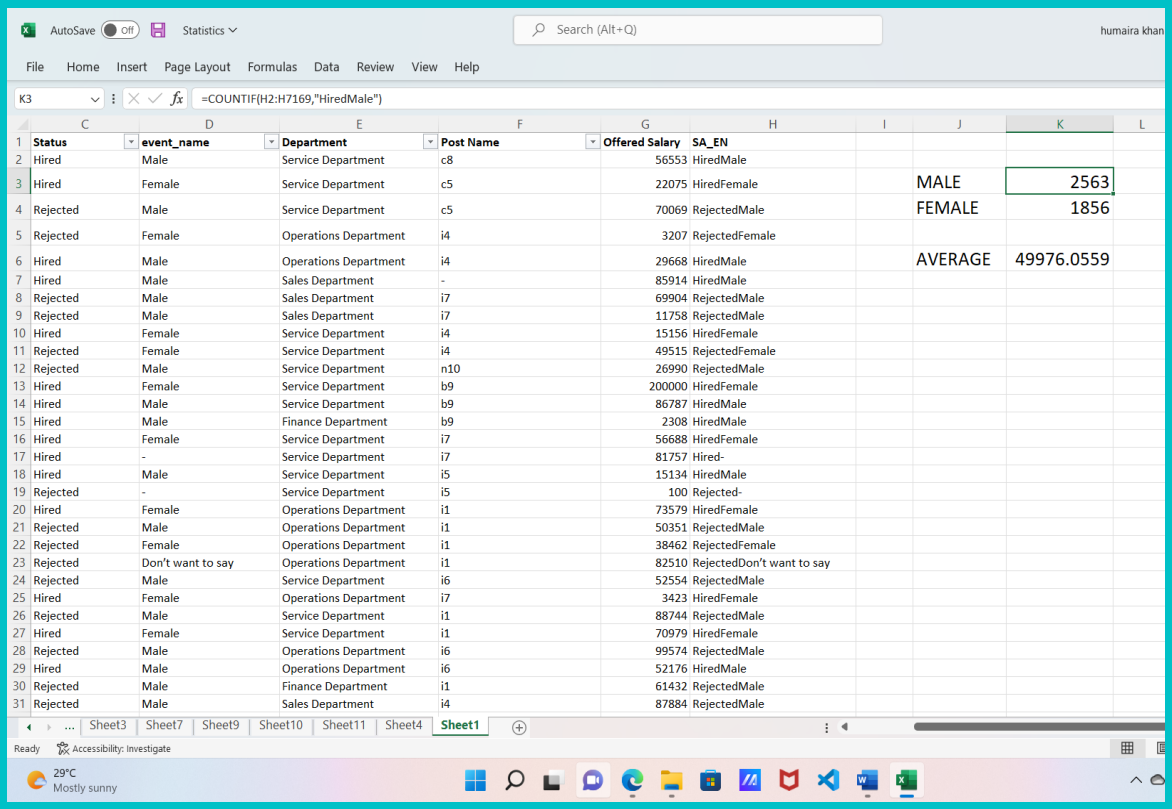
PROJECT DESRIPTION: THIS PROJECT IS TO KNOW ABOUT THE MAJOR UNDERLYING TRENDS ABOUT THE HIRING PROCESS. TRENDS SUCH AS- NUMBER OF REJECTIONS, NUMBER OF INTERVIEWS, TYPES OF JOBS, VACANCIES ETC. ARE IMPORTANT FOR A COMPANY TO ANALYZE BEFORE HIRING FRESHERS OR ANY OTHER INDIVIDUAL. BEING A DATA ANALYST, OUR JOB IS TO GO THROUGH THESE TRENDS AND DRAW INSIGHTS OUT OF IT FOR HIRING DEPARTMENT TO WORK UPON.

APPROACH AND INSIGHTS

I FIRST DOWNLOADED THE DATASET PROVIDED AND OPENED IT IN EXCEL. TRIED TO UNDERSTAND WHAT THAT DATASET CONTAINS, AND THEN STARTED WITH THE QUERIES USING ONLY EXCEL SHEET.

THE FIRST TASK WAS TO FIND THE NUMBER OF MALE AND FEMALE WORKERS HIRED. SO MY FIRST INSTINCT WAS TO JOIN STATUS COLUMN AND EVENT NAME COLUMN SO THAT THE NEW COLUMN CREATED HAS DETAILS OF BOTH WHO HAS BEEN HIRED ALONG WITH THE GENDER.

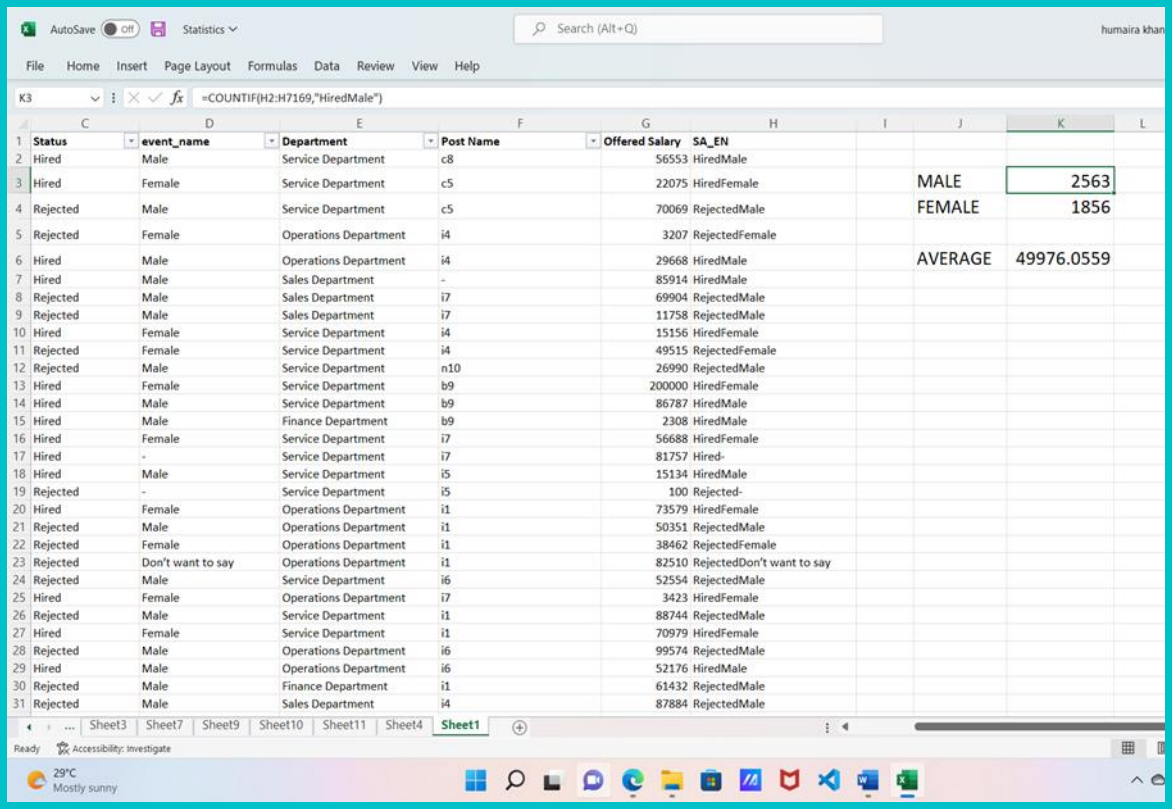
I USED THE COUNTIF FUNCTION,
=COUNTIF(G2:G7169,"HIREDMALE")
=COUNTIF(G2:G7169,"HIREDFEMALE")
BELOW IS THE ANSWER.



	C	D	E	F	G	H	I	J	K	L
1	Status	event_name	Department	Post Name	Offered Salary	SA_EN				
2	Hired	Male	Service Department	c8	56553	HiredMale				
3	Hired	Female	Service Department	c5	22075	HiredFemale		MALE	2563	
4	Rejected	Male	Service Department	c5	70069	RejectedMale		FEMALE	1856	
5	Rejected	Female	Operations Department	i4	3207	RejectedFemale				
6	Hired	Male	Operations Department	i4	29668	HiredMale		AVERAGE	49976.0559	
7	Hired	Male	Sales Department	-	85914	HiredMale				
8	Rejected	Male	Sales Department	i7	69904	RejectedMale				
9	Rejected	Male	Sales Department	i7	11758	RejectedMale				
10	Hired	Female	Service Department	i4	15156	HiredFemale				
11	Rejected	Female	Service Department	i4	49515	RejectedFemale				
12	Rejected	Male	Service Department	n10	26990	RejectedMale				
13	Hired	Female	Service Department	b9	200000	HiredFemale				
14	Hired	Male	Service Department	b9	86787	HiredMale				
15	Hired	Male	Finance Department	b9	2308	HiredMale				
16	Hired	Female	Service Department	i7	56688	HiredFemale				
17	Hired	-	Service Department	i7	81757	Hired-				
18	Hired	Male	Service Department	i5	15134	HiredMale				
19	Rejected	-	Service Department	i5	100	Rejected-				
20	Hired	Female	Operations Department	i1	73579	HiredFemale				
21	Rejected	Male	Operations Department	i1	50351	RejectedMale				
22	Rejected	Female	Operations Department	i1	38462	RejectedFemale				
23	Rejected	Don't want to say	Operations Department	i1	82510	RejectedDon't want to say				
24	Rejected	Male	Service Department	i6	52554	RejectedMale				
25	Hired	Female	Operations Department	i7	3423	HiredFemale				
26	Rejected	Male	Service Department	i1	88744	RejectedMale				
27	Hired	Female	Service Department	i1	70979	HiredFemale				
28	Rejected	Male	Operations Department	i6	99574	RejectedMale				
29	Hired	Male	Operations Department	i6	52176	HiredMale				
30	Rejected	Male	Finance Department	i1	61432	RejectedMale				
31	Rejected	Male	Sales Department	i4	87884	RejectedMale				

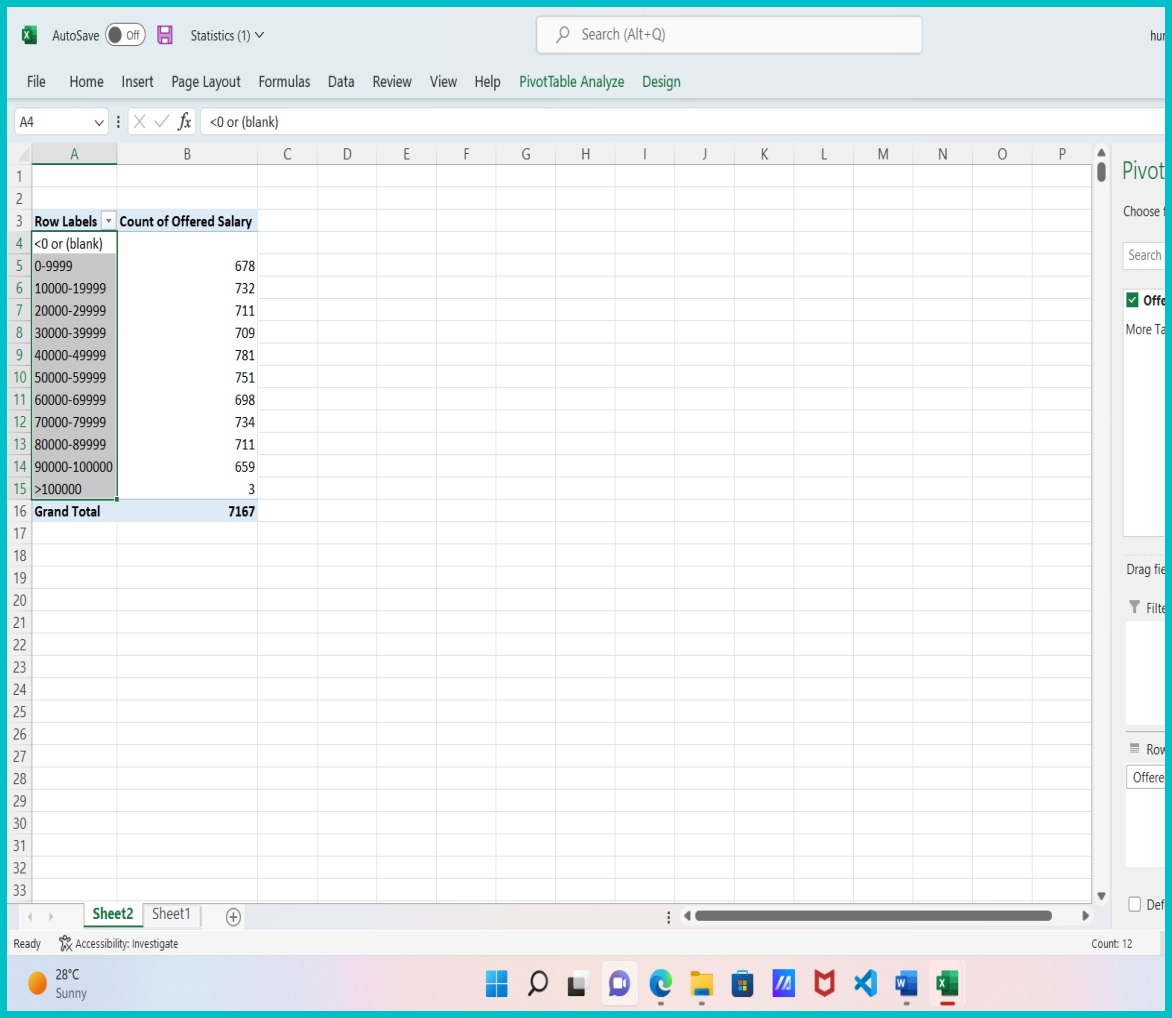
2563 MALES AND 1856 FEMALES WERE HIRED.

SECOND QUERY WAS TO FIND THE AVERAGE SALARY. AS YOU CAN SEE IN THE PICTURE, AVERAGE COMES OUT TO BE 49976.0559 USING FORMULA =AVERAGE(G2:G7169)



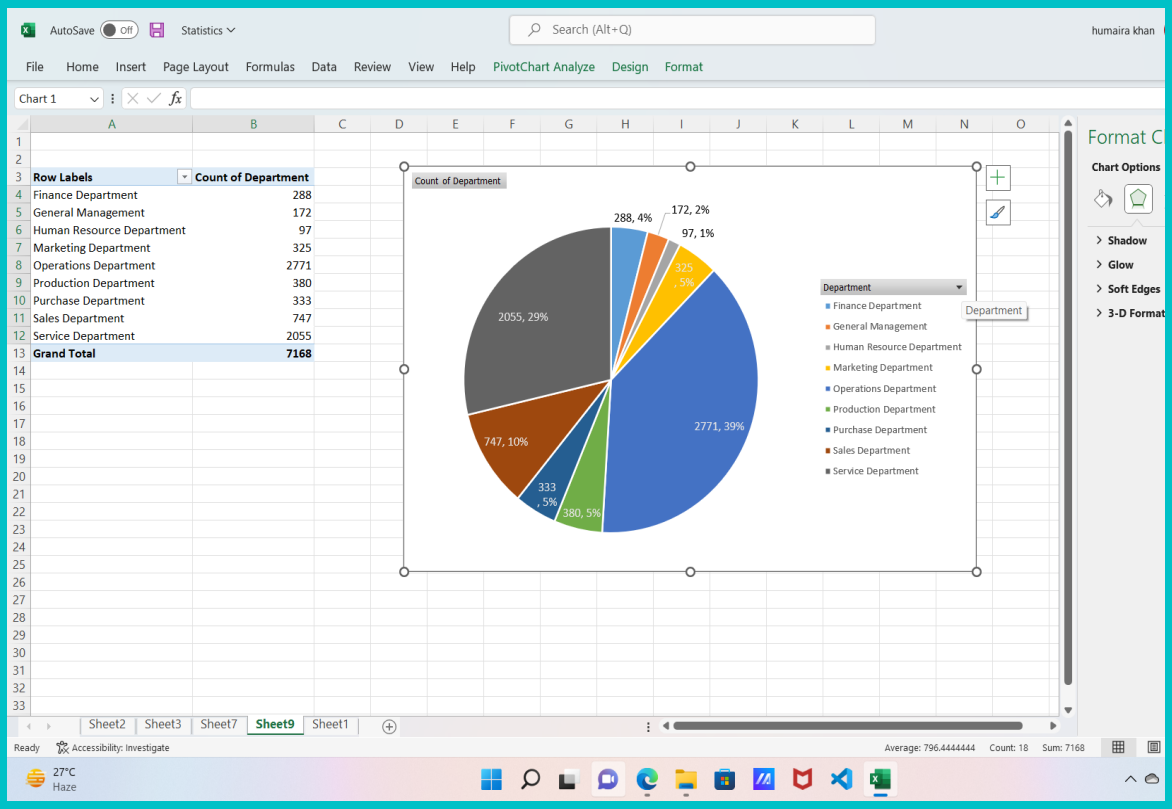
	C	D	E	F	G	H
1	Status	event_name	Department	Post Name	Offered Salary	SA_EN
2	Hired	Male	Service Department	c8	56553	HiredMale
3	Hired	Female	Service Department	c5	22075	HiredFemale
4	Rejected	Male	Service Department	c5	70069	RejectedMale
5	Rejected	Female	Operations Department	i4	3207	RejectedFemale
6	Hired	Male	Operations Department	i4	29668	HiredMale
7	Hired	Male	Sales Department	-	85914	HiredMale
8	Rejected	Male	Sales Department	i7	69904	RejectedMale
9	Rejected	Male	Sales Department	i7	11758	RejectedMale
10	Hired	Female	Service Department	i4	15156	HiredFemale
11	Rejected	Female	Service Department	i4	49515	RejectedFemale
12	Rejected	Male	Service Department	n10	26990	RejectedMale
13	Hired	Female	Service Department	b9	200000	HiredFemale
14	Hired	Male	Service Department	b9	86787	HiredMale
15	Hired	Male	Finance Department	b9	2308	HiredMale
16	Hired	Female	Service Department	i7	56688	HiredFemale
17	Hired	-	Service Department	i7	81757	Hired-
18	Hired	Male	Service Department	i5	15134	HiredMale
19	Rejected	-	Service Department	i5	100	Rejected-
20	Hired	Female	Operations Department	i1	73579	HiredFemale
21	Rejected	Male	Operations Department	i1	50351	RejectedMale
22	Rejected	Female	Operations Department	i1	38462	RejectedFemale
23	Rejected	Don't want to say	Operations Department	i1	82510	RejectedDon't want to say
24	Rejected	Male	Service Department	i6	52554	RejectedMale
25	Hired	Female	Operations Department	i7	3423	HiredFemale
26	Rejected	Male	Service Department	i1	88744	RejectedMale
27	Hired	Female	Service Department	i1	70979	HiredFemale
28	Rejected	Male	Operations Department	i6	99574	RejectedMale
29	Hired	Male	Operations Department	i6	52176	HiredMale
30	Rejected	Male	Finance Department	i1	61432	RejectedMale
31	Rejected	Male	Sales Department	i4	87884	RejectedMale

THIRD QUERY WAS TO DRAW CLASS INTERVALS FOR SALARY. I USED PIVOT TABLE AND THE GROUPING FUNCTION TO DO THE SAME.



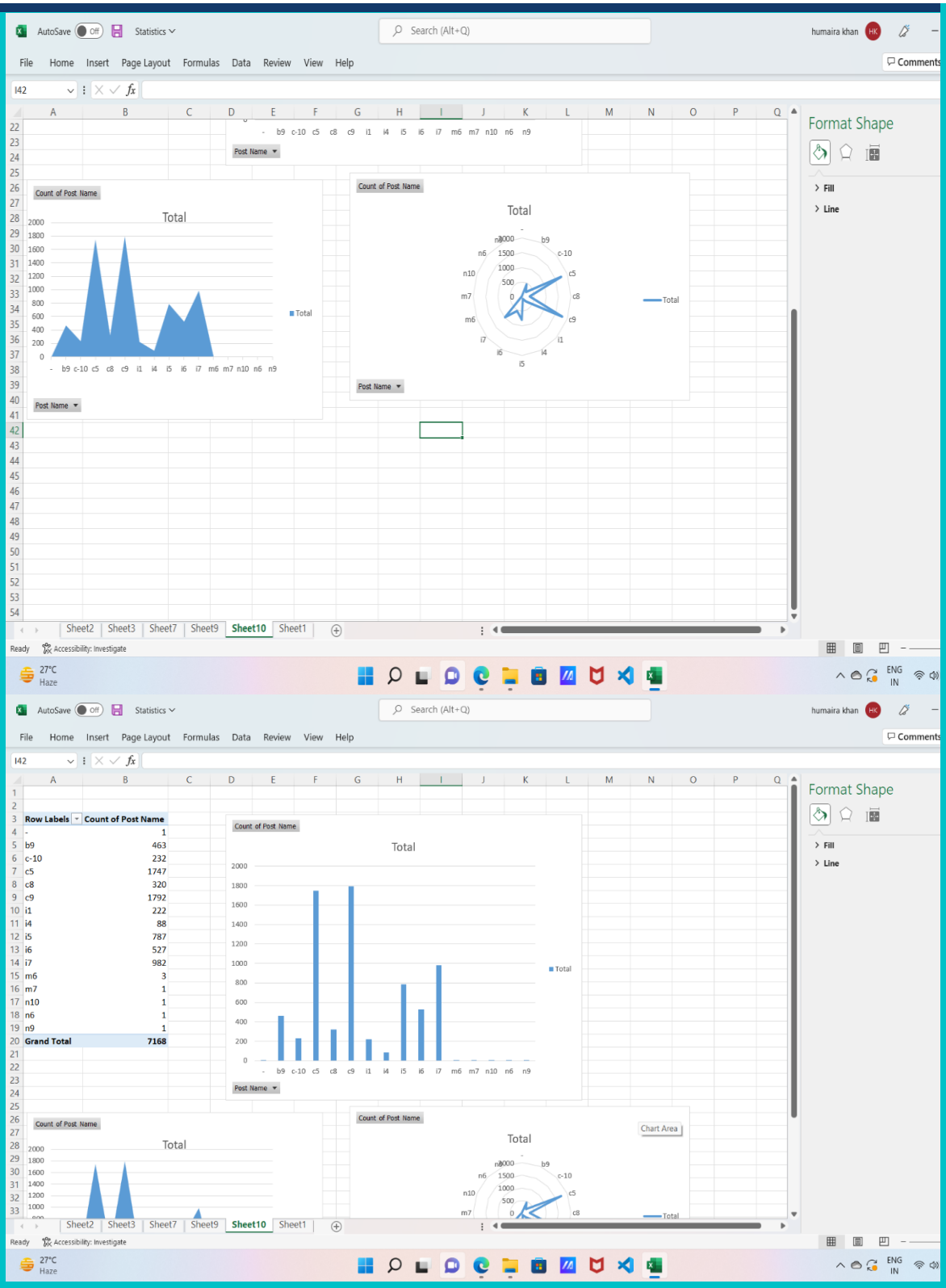
SET THE RANGE TO 10000 RANGING FROM 0 TO 100000.

4TH QUERY WAS TO DRAW PIE CHART TO SHOW PROPORTION OF PEOPLE WORKING DIFFERENT DEPARTMENT.



AGAIN I USED THE PIVOT TABLE ON THE DEPARTMENT COLUMN FIRST AND THEN CHOSE A PIE CHART. I ADDED LABELS TO THE CHART SO IT BECOMES EASIER TO UNDERSTAND ALONG WITH DATA LABELS.

LAST QUERY WAS TO USE DIFFERENT CHARTS.



I USED IT ON POST NAMES AND THEIR TOTAL COUNT AND CREATED 3 GRAPHS NAMELY BAR CHART, AREA CHART AND RADAR CHART.

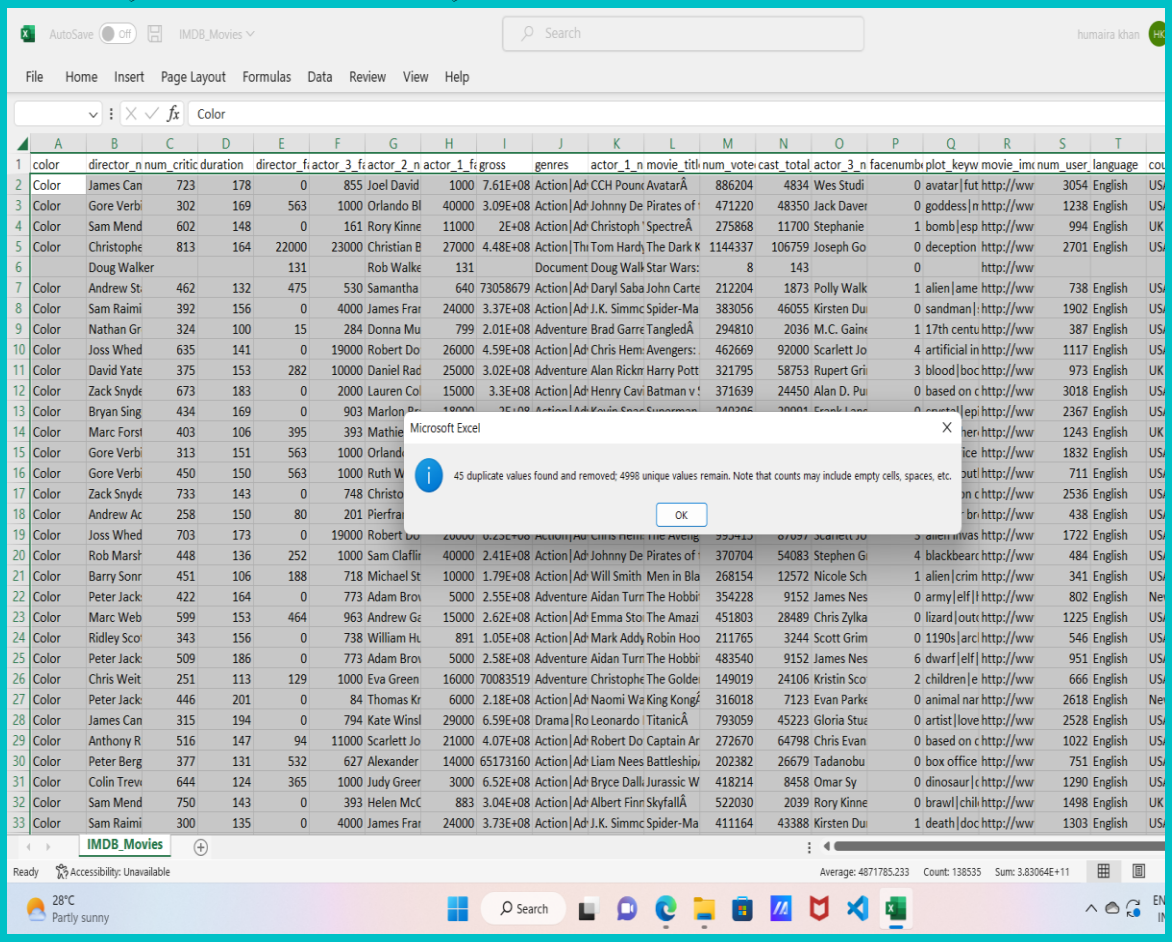
MOVIE IDMB ANALYSIS

DESCRIPTION:
THE PROJECT IS ABOUT FINDING OUT THE VARIOUS INSIGHTS IN IMDB_MOVIES DATASET . WE ANALYZE THIS DATA AND SOME FOLLOWING QUESTIONS:
1.CLEAN THE DATA
2.FIND THE MOVIES WITH THE HIGHEST PROFIT?
3.FIND IMDB TOP 250
4.FIND TOP 10 DIRECTORS
5.FIND POPULAR GENRES
6.FIND THE CRITIC-FAVORITE AND AUDIENCE-FAVORITE ACTORS.

QUERIES

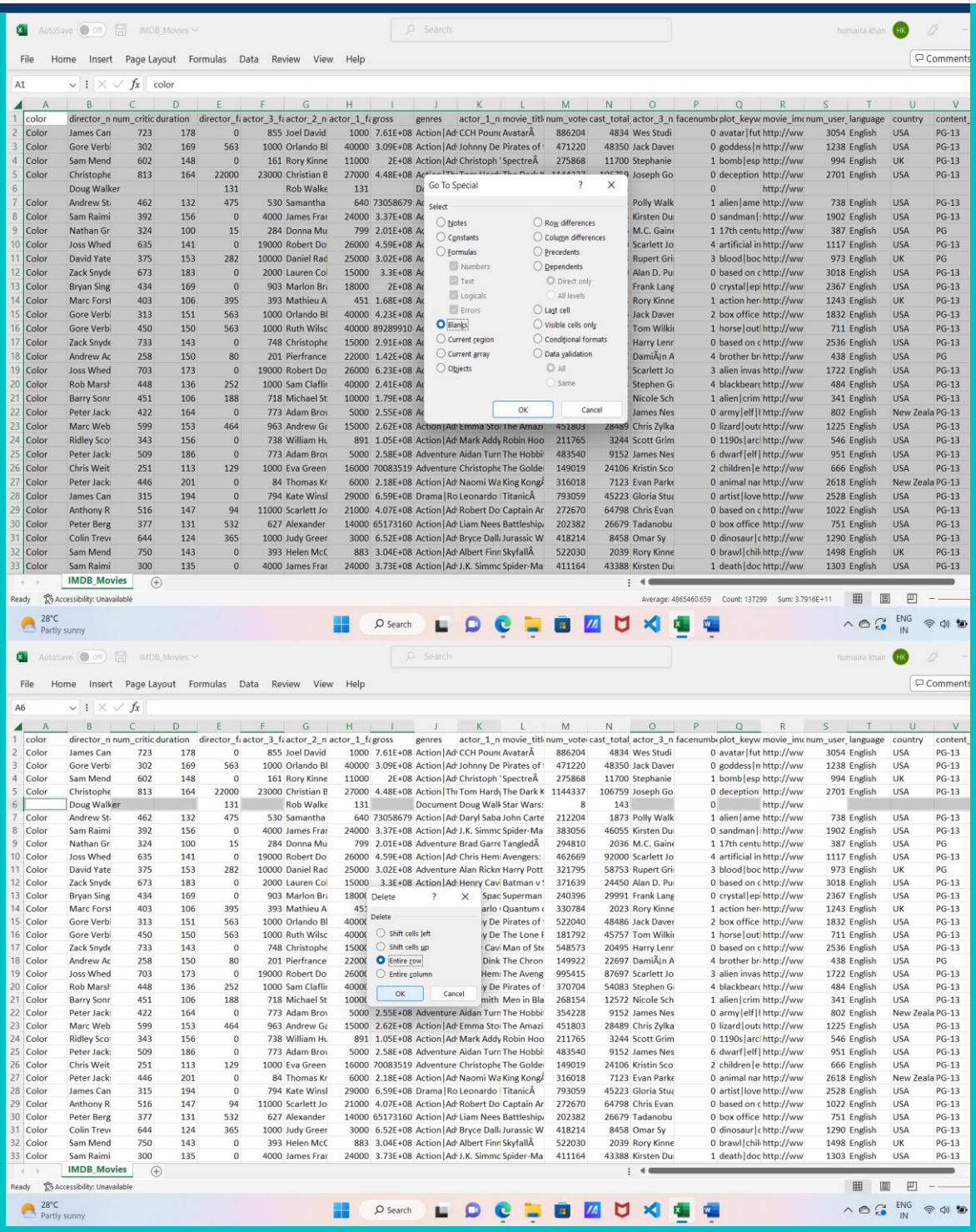
THE FIRST TASK WAS TO CLEAN THE DATASET.

I STARTED THIS WITH REMOVING DUPLICATE DATA.



45 DUPLICATE VALUES WERE FOUND AND REMOVED.

NEXT, I SELECTED ALL THE BLANK CELLS AND DELETED ALL THE ROWS.

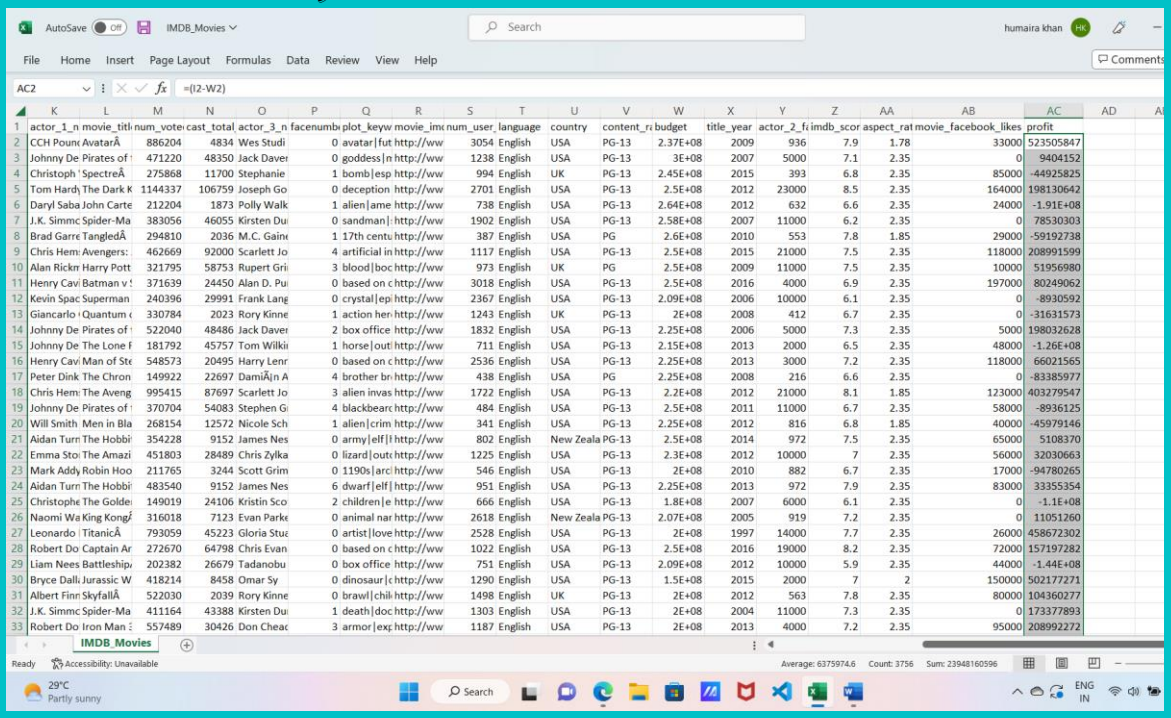


ALL THE BLANK CELLS ARE DELETED NOW.

WE CAN CHECK IF THE COLUMN HAS ANY BLANK ROWS BY SELECTING THE FIRST CELL OF THE COLUMN AND PRESSING CONTROL DOWN. IF THE CURSOR MOVES TO THE LAST CELL OF THE COLUMN, THERE ARE NO BLANK CELLS.

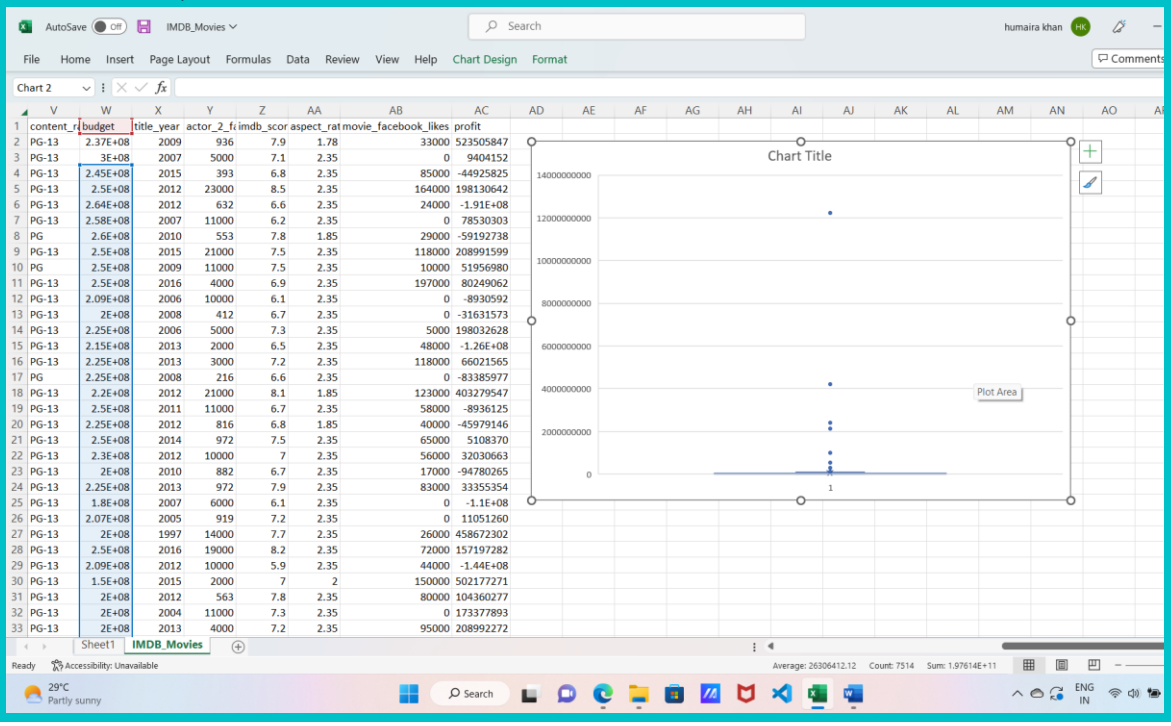
2) SECOND TASK

I FIRST CREATED A NEW COLUMN PROFIT WHICH IS THE DIFFERENCE BETWEEN GROSS AND BUDGET.
THEN I USED THE SORT FUNCTION ON PROFIT COLUMN(ASC TO DESCENDING)



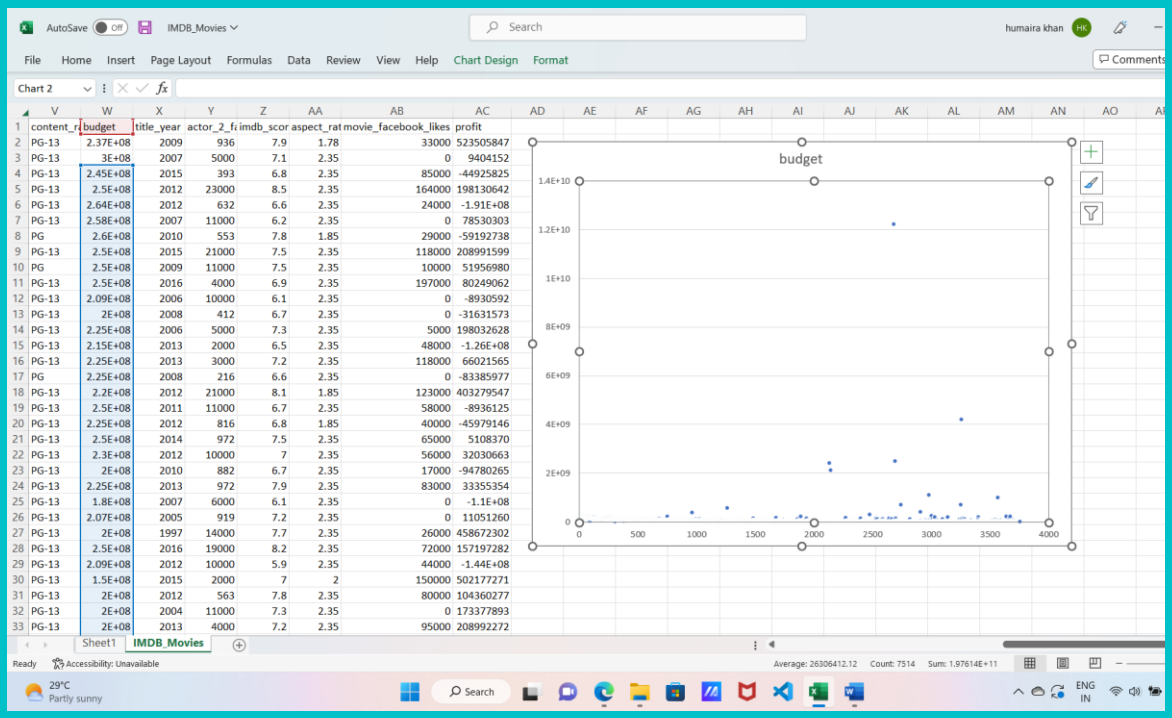
actor_1_n	movie_titl	num_vote	cast_total	actor_3_n	facenumb	plot_keyw	movie_im	num_user	language	country	content_ri	budget	title_year	actor_2_fa	imdb_scor	aspect_rat	movie_facebook_likes	profit
CCH Poun	AvatarÂ	886204	4834	Wes Studi	0	avatar|fut	http://ww	3054	English	USA	PG-13	2.37E+08	2009	936	7.9	1.78	33000	523505847
Johnny De	Pirates of	471220	48350	Jack Daver	0	goddess|n	http://ww	1238	English	USA	PG-13	3E+08	2007	5000	7.1	2.35	0	9404152
Christoph	SpectreÂ	275868	11700	Stephanie	1	bomb|esp	http://ww	994	English	UK	PG-13	2.45E+08	2015	393	6.8	2.35	85000	-44925825
Tom Hardy	The Dark K	1144337	106759	Joseph Go	0	deception	http://ww	2701	English	USA	PG-13	2.5E+08	2012	23000	8.5	2.35	164000	198130642
Daryl Saba	John Carte	212204	1873	Polly Walk	1	alien|ame	http://ww	738	English	USA	PG-13	2.64E+08	2012	632	6.6	2.35	24000	-1.91E+08
J.K. Simmc	Spider-Ma	383056	46055	Kirsten Du	0	sandman|	http://ww	1902	English	USA	PG-13	2.58E+08	2007	11000	6.2	2.35	0	78530303
Brad Garre	TangledÂ	294810	2036	M.C. Gain	1	17th centu	http://ww	387	English	USA	PG	2.6E+08	2010	553	7.8	1.85	29000	-59192738
Chris Hem	Avengers:	462669	92000	Scarlett Jo	4	artificial in	http://ww	1117	English	USA	PG-13	2.5E+08	2015	21000	7.5	2.35	118000	208991599
Alan Rickm	Harry Pott	321795	58753	Rupert Gri	3	blood|boc	http://ww	973	English	UK	PG	2.5E+08	2009	11000	7.5	2.35	10000	51956980
Henry Cavi	Batman v S	371639	24450	Alan D. Pu	0	based on c	http://ww	3018	English	USA	PG-13	2.5E+08	2016	4000	6.9	2.35	197000	80249062
Kevin Spac	Superman	240396	29991	Frank Lang	0	crystal|epi	http://ww	2367	English	USA	PG-13	2.09E+08	2006	10000	6.1	2.35	0	-8930592
Giancarlo	Quantum c	330784	2023	Rory Kinne	1	action her	http://ww	1243	English	UK	PG-13	2E+08	2008	412	6.7	2.35	0	-31631573
Johnny De	Pirates of	522040	48486	Jack Daver	2	box office	http://ww	1832	English	USA	PG-13	2.25E+08	2006	5000	7.3	2.35	5000	198032628
Johnny De	The Lone F	181792	45757	Tom Wilkii	1	horse|out	http://ww	711	English	USA	PG-13	2.15E+08	2013	2000	6.5	2.35	48000	-1.26E+08
Henry Cavi	Man of Ste	548573	20495	Harry Lenr	0	based on c	http://ww	2536	English	USA	PG-13	2.25E+08	2013	3000	7.2	2.35	118000	66021565
Peter Dink	The Chron	149922	22697	DamiÃ¡n A	4	brother br	http://ww	438	English	USA	PG	2.25E+08	2008	216	6.6	2.35	0	-83385977
Chris Hem	The Aveng	995415	87697	Scarlett Jo	3	alien invas	http://ww	1722	English	USA	PG-13	2.2E+08	2012	21000	8.1	1.85	123000	403279547
Johnny De	Pirates of	370704	54083	Stephen Gi	4	blackbear	http://ww	484	English	USA	PG-13	2.5E+08	2011	11000	6.7	2.35	58000	-8936125
Will Smith	Men in Bla	268154	12572	Nicole Sch	1	alien|crim	http://ww	341	English	USA	PG-13	2.25E+08	2012	816	6.8	1.85	40000	-45979146
Aidan Turn	The Hobbi	354228	9152	James Nes	0	army|elf|l	http://ww	802	English	New Zeala	PG-13	2.5E+08	2014	972	7.5	2.35	65000	5108370
Emma Sto	The Amazi	451803	28489	Chris Zylka	0	lizard|out	http://ww	1225	English	USA	PG-13	2.3E+08	2012	10000	7	2.35	56000	32030663
Mark Addy	Robin Hoo	211765	3244	Scott Grim	0	1190s|arc	http://ww	546	English	USA	PG-13	2E+08	2010	882	6.7	2.35	17000	-94780265
Aidan Turn	The Hobbi	483540	9152	James Nes	6	dwarf|elf|	http://ww	951	English	USA	PG-13	2.25E+08	2013	972	7.9	2.35	83000	33355354
Christophe	The Golde	149019	24106	Kristin Sco	2	children|e	http://ww	666	English	USA	PG-13	1.8E+08	2007	6000	6.1	2.35	0	-1.1E+08
Naomi Wa	King KongÂ	316018	7123	Evan Parke	0	animal nar	http://ww	2618	English	New Zeala	PG-13	2.07E+08	2005	919	7.2	2.35	0	11051260
Leonardo	TitanicÂ	793059	45223	Gloria Stua	0	artist|love	http://ww	2528	English	USA	PG-13	2E+08	1997	14000	7.7	2.35	26000	458672302
Robert Do	Captain Ar	272670	64798	Chris Evan	0	based on c	http://ww	1022	English	USA	PG-13	2.5E+08	2016	19000	8.2	2.35	72000	157197282
Liam Nees	Battleship/	202382	26679	Tadanobu	0	box office	http://ww	751	English	USA	PG-13	2.09E+08	2012	10000	5.9	2.35	44000	-1.44E+08
Bryce Dall	Jurassic W	418214	8458	Omar Sy	0	dinosaur|c	http://ww	1290	English	USA	PG-13	1.5E+08	2015	2000	7	2	150000	502177271
Albert Finn	SkyfallÂ	522030	2039	Rory Kinne	0	brawl|chil	http://ww	1498	English	UK	PG-13	2E+08	2012	563	7.8	2.35	80000	104360277
J.K. Simmc	Spider-Ma	411164	43388	Kirsten Du	1	death|doc	http://ww	1303	English	USA	PG-13	2E+08	2004	11000	7.3	2.35	0	173377893
Robert Do	Iron Man 3	557489	30426	Don Cheac	3	armor|exp	http://ww	1187	English	USA	PG-13	2E+08	2013	4000	7.2	2.35	95000	208992272

NOW TO PLOT BUDGET AND PROFIT DATA, AND TO FIND OUTLIERS, THE BEST CHART WOULD BE BOXPLOT AND SCATTERPLOT.



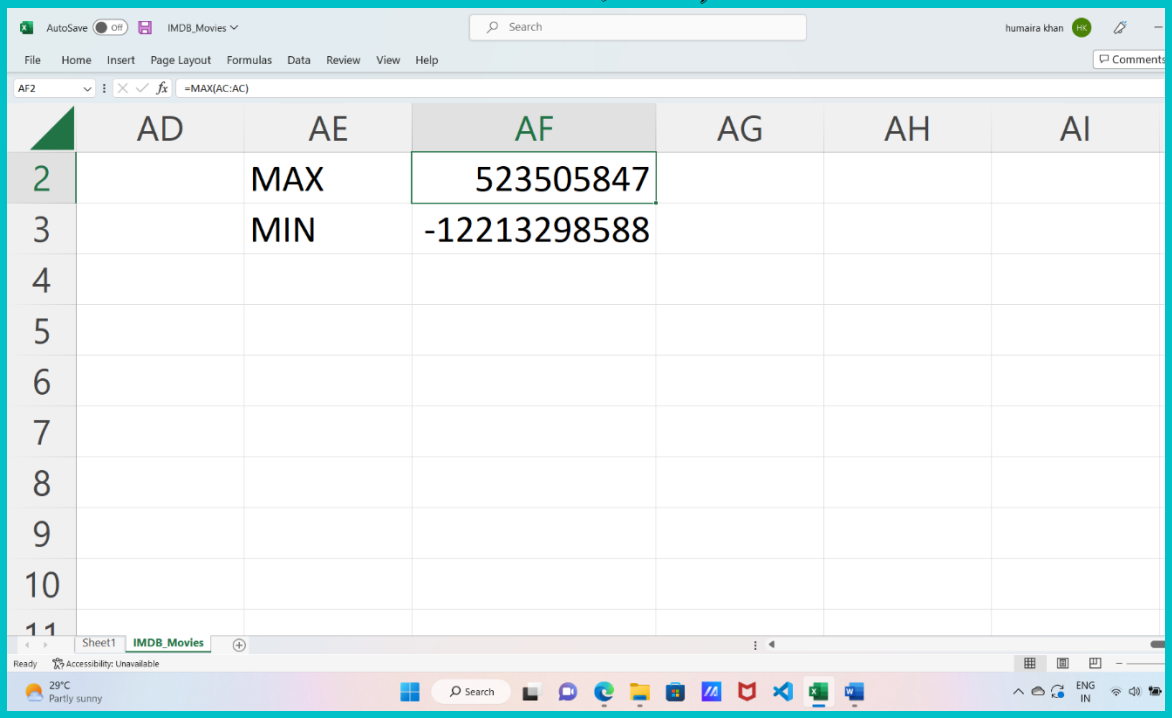
content_ri	budget	title_year	actor_2_fa	imdb_scor	aspect_rat	movie_facebook_likes	profit
PG-13	2.37E+08	2009	936	7.9	1.78	33000	523505847
PG-13	3E+08	2007	5000	7.1	2.35	0	9404152
PG-13	2.45E+08	2015	393	6.8	2.35	85000	-44925825
PG-13	2.5E+08	2012	23000	8.5	2.35	164000	198130642
PG-13	2.64E+08	2012	632	6.6	2.35	24000	-1.91E+08
PG-13	2.58E+08	2007	11000	6.2	2.35	0	78530303
PG	2.6E+08	2010	553	7.8	1.85	29000	-59192738
PG-13	2.5E+08	2015	21000	7.5	2.35	118000	208991599
PG	2.5E+08	2009	11000	7.5	2.35	10000	51956980
PG-13	2.5E+08	2016	4000	6.9	2.35	197000	80249062
PG-13	2.09E+08	2006	10000	6.1	2.35	0	-8930592
PG-13	2E+08	2008	412	6.7	2.35	0	-31631573
PG-13	2.25E+08	2006	5000	7.3	2.35	5000	198032628
PG-13	2.15E+08	2013	2000	6.5	2.35	48000	-1.26E+08
PG-13	2.25E+08	2013	3000	7.2	2.35	118000	66021565
PG	2.25E+08	2008	216	6.6	2.35	0	-83385977
PG-13	2.2E+08	2012	21000	8.1	1.85	123000	403279547
PG-13	2.5E+08	2011	11000	6.7	2.35	58000	-8936125
PG-13	2.25E+08	2012	816	6.8	1.85	40000	-45979146
PG-13	2.5E+08	2014	972	7.5	2.35	65000	5108370
PG-13	2.3E+08	2012	10000	7	2.35	56000	32030663
PG-13	2E+08	2010	882	6.7	2.35	17000	-94780265
PG-13	2.25E+08	2013	972	7.9	2.35	83000	33355354
PG-13	1.8E+08	2007	6000	6.1	2.35	0	-1.1E+08
PG-13	2.07E+08	2005	919	7.2	2.35	0	11051260
PG-13	2E+08	1997	14000	7.7	2.35	26000	458672302
PG-13	2.5E+08	2016	19000	8.2	2.35	72000	157197282
PG-13	2.09E+08	2012	10000	5.9	2.35	44000	-1.44E+08
PG-13	1.5E+08	2015	2000	7	2	150000	502177271
PG-13	2E+08	2012	563	7.8	2.35	80000	104360277
PG-13	2E+08	2004	11000	7.3	2.35	0	173377893
PG-13	2E+08	2013	4000	7.2	2.35	95000	208992272

WE CAN CLEARLY SEE THE OUTLIER HERE. BELOW IS SCATTER PLOT.



	V	W	X	Y	Z	AA	AB	AC
1	content_r	budget	title_year	actor_2_f	imdb_scor	aspect_rat	movie_facebook_likes	profit
2	PG-13	2.37E+08	2009	936	7.9	1.78	33000	523505847
3	PG-13	3E+08	2007	5000	7.1	2.35	0	9404152
4	PG-13	2.45E+08	2015	393	6.8	2.35	85000	-44925825
5	PG-13	2.5E+08	2012	23000	8.5	2.35	164000	198130642
6	PG-13	2.64E+08	2012	632	6.6	2.35	24000	-1.91E+08
7	PG-13	2.58E+08	2007	11000	6.2	2.35	0	78530303
8	PG	2.6E+08	2010	553	7.8	1.85	29000	-59192738
9	PG-13	2.5E+08	2015	21000	7.5	2.35	118000	208991599
10	PG	2.5E+08	2009	11000	7.5	2.35	10000	51956980
11	PG-13	2.5E+08	2016	4000	6.9	2.35	197000	80249062
12	PG-13	2.09E+08	2006	10000	6.1	2.35	0	-8930592
13	PG-13	2E+08	2008	412	6.7	2.35	0	-31631573
14	PG-13	2.25E+08	2006	5000	7.3	2.35	5000	198032628
15	PG-13	2.15E+08	2013	2000	6.5	2.35	48000	-1.26E+08
16	PG-13	2.25E+08	2013	3000	7.2	2.35	118000	66021565
17	PG	2.25E+08	2008	216	6.6	2.35	0	-83385977
18	PG-13	2.2E+08	2012	21000	8.1	1.85	123000	403279547
19	PG-13	2.5E+08	2011	11000	6.7	2.35	58000	-8936125
20	PG-13	2.25E+08	2012	816	6.8	1.85	40000	-45979146
21	PG-13	2.5E+08	2014	972	7.5	2.35	65000	5108370
22	PG-13	2.3E+08	2012	10000	7	2.35	56000	32030663
23	PG-13	2E+08	2010	882	6.7	2.35	17000	-94780265
24	PG-13	2.25E+08	2013	972	7.9	2.35	83000	33355354
25	PG-13	1.8E+08	2007	6000	6.1	2.35	0	-1.1E+08
26	PG-13	2.07E+08	2005	919	7.2	2.35	0	11051260
27	PG-13	2E+08	1997	14000	7.7	2.35	26000	458672302
28	PG-13	2.5E+08	2016	19000	8.2	2.35	72000	157197282
29	PG-13	2.09E+08	2012	10000	5.9	2.35	44000	-1.44E+08
30	PG-13	1.5E+08	2015	2000	7	2	150000	502177271
31	PG-13	2E+08	2012	563	7.8	2.35	80000	104360277
32	PG-13	2E+08	2004	11000	7.3	2.35	0	173377893
33	PG-13	2E+08	2013	4000	7.2	2.35	95000	208992272

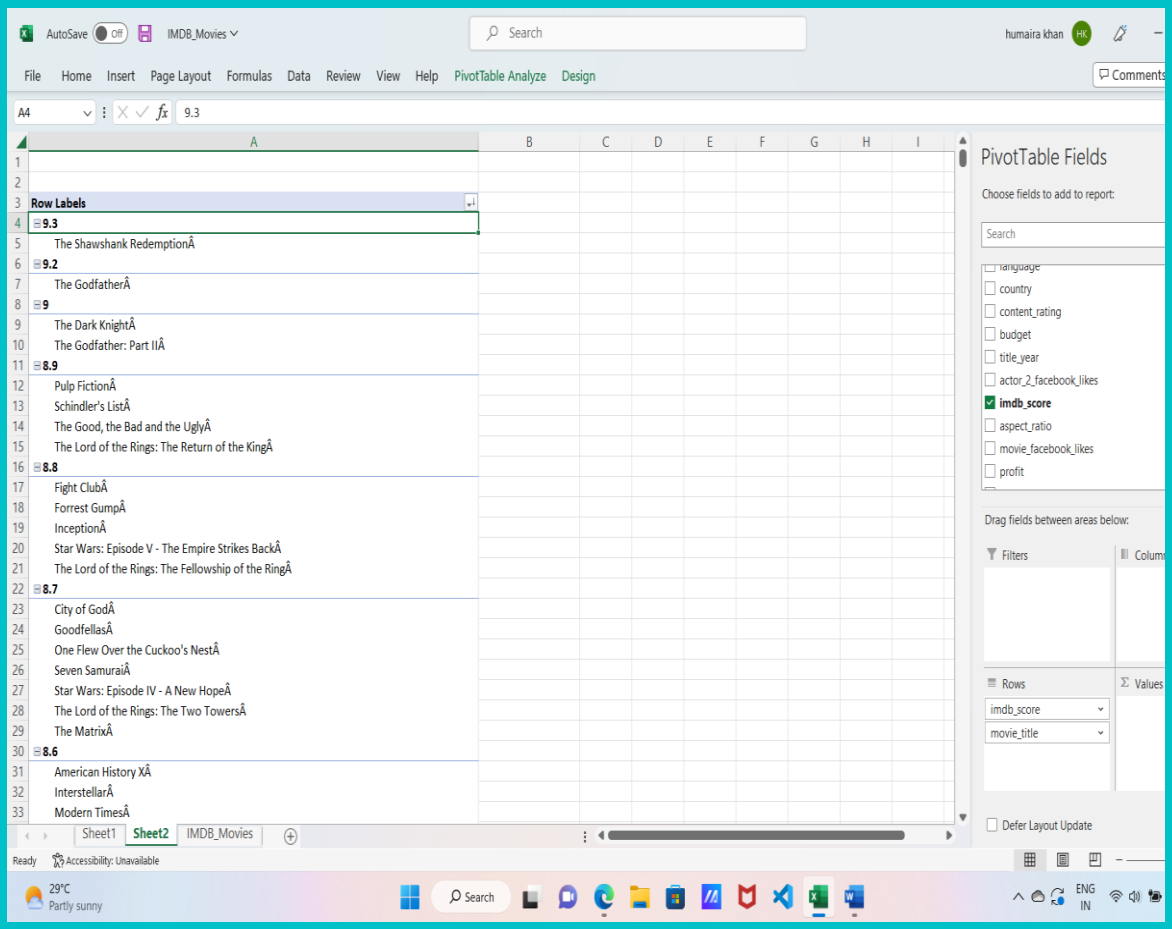
NOW TO FIND THE MAXIMUM PROFIT,

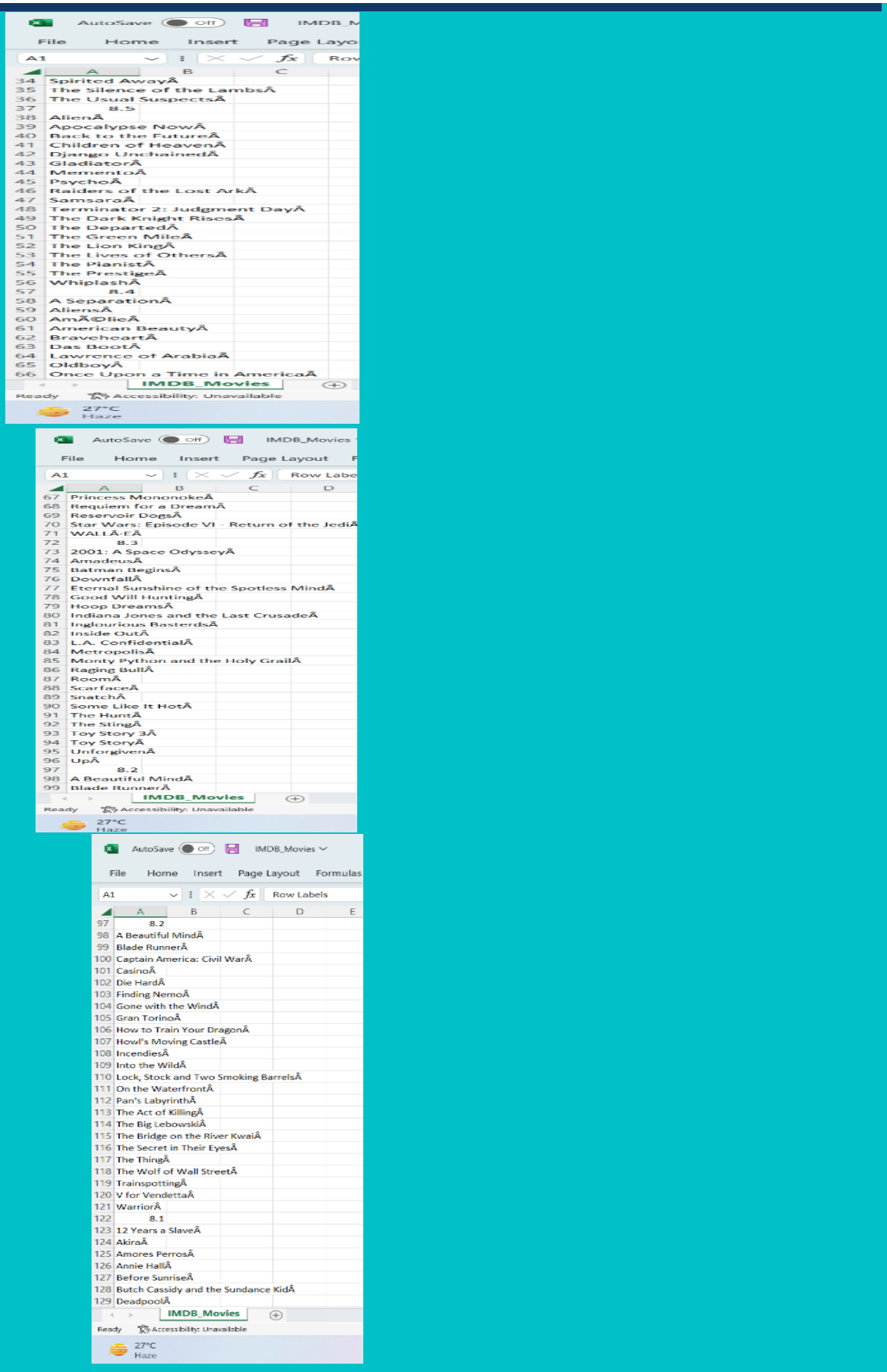


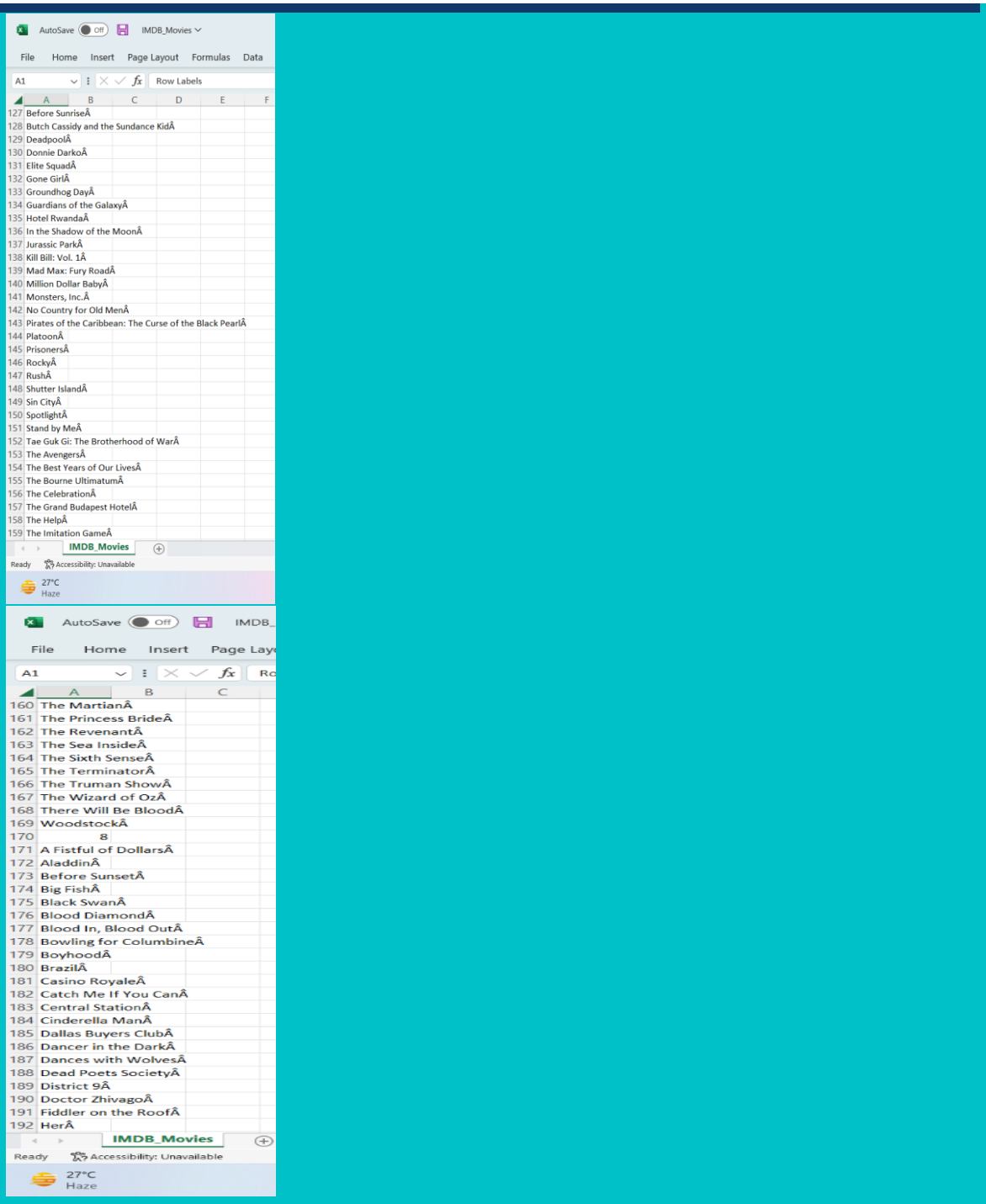
WE USED THE MAX FUNCTION AND WE GOT TO KNOW THAT AVATARA IS THE MOVIE WITH HIGHEST PROFIT.
TO FIND THE TOP 10 WE USED THE SORT FUNCTION.

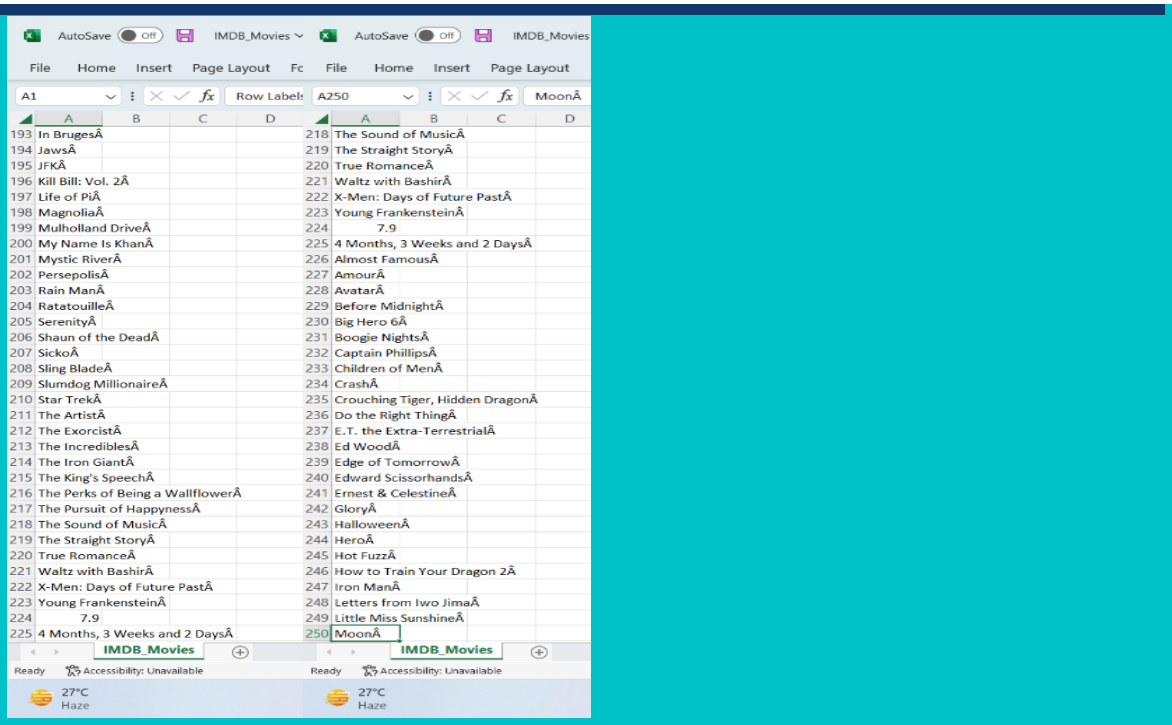
TOP 10 HIGHEST MOVIES	PROFIT
Avatar	523505847
Jurassic World	502177271
Titanic	458672302
Stars War:Episode IV-A New Hope	449935665
E.T. the Extra-Terrestrial	424449459
The Avengers	403279547
The Lion King	377783777
Stars War:Episode I- The Phantom Menace	359544677
The Dark Knight	348316061
The Hunger Games	329999255

THE THIRD QUERY WAS TO CALCULATE TOP 250 MOVIES WITH THE HIGHEST IDMB RATING.
FOR THIS WE USED THE PIVOT TABLE FUNCTION, SORTED THE VALUES IN DESCENDING ORDER AND IT SHOWS THE TOP RATED MOVIES.



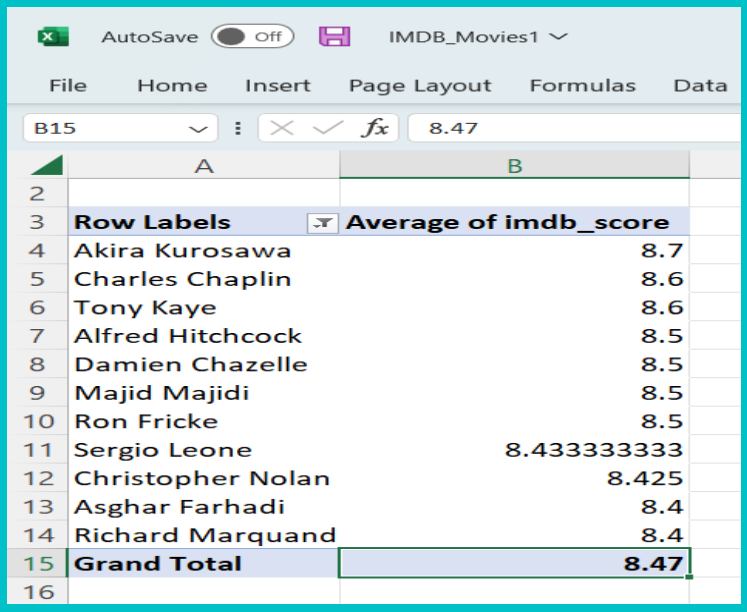

AutoSave Off
IMDB_M
File Home Insert Page Layo
A1
Rov
A
B
C
34 Spirited AwayÂ
35 The Silence of the LambsÂ
36 The Usual SuspectsÂ
37 8.5
38 AlienÂ
39 Apocalypse NowÂ
40 Back to the FutureÂ
41 Children of HeavenÂ
42 Django UnchainedÂ
43 GladiatorÂ
44 MementoÂ
45 PsychoÂ
46 Raiders of the Lost ArkÂ
47 SamsaraÂ
48 Terminator 2: Judgment DayÂ
49 The Dark Knight RisesÂ
50 The DepartedÂ
51 The Green MileÂ
52 The Lion KingÂ
53 The Lives of OthersÂ
54 The PianistÂ
55 The PrestigeÂ
56 WhiplashÂ
57 8.4
58 A SeparationÂ
59 AliensÂ
60 AmÃ©lieÂ
61 American BeautyÂ
62 BraveheartÂ
63 Das BootÂ
64 Lawrence of ArabiaÂ
65 OldboyÂ
66 Once Upon a Time in AmericaÂ
IMDB_Movies
Ready
Accessibility: Unavailable
27°C
Haze
AutoSave Off
IMDB_Movies
File Home Insert Page Layout
A1
Row Labe
A
B
C
D
67 Princess MononokeÂ
68 Requiem for a DreamÂ
69 Reservoir DogsÂ
70 Star Wars: Episode VI - Return of the JediÂ
71 WALLÂ·EÂ
72 8.3
73 2001: A Space OdysseyÂ
74 AmadeusÂ
75 Batman BeginsÂ
76 DownfallÂ
77 Eternal Sunshine of the Spotless MindÂ
78 Good Will HuntingÂ
79 Hoop DreamsÂ
80 Indiana Jones and the Last CrusadeÂ
81 Inglourious BasterdsÂ
82 Inside OutÂ
83 L.A. ConfidentialÂ
84 MetropolisÂ
85 Monty Python and the Holy GrailÂ
86 Raging BullÂ
87 RoomÂ
88 ScarfaceÂ
89 SnatchÂ
90 Some Like It HotÂ
91 The HuntÂ
92 The StingÂ
93 Toy Story 3Â
94 Toy StoryÂ
95 UnforgivenÂ
96 UpÂ
97 8.2
98 A Beautiful MindÂ
99 Blade RunnerÂ
IMDB_Movies
Ready
Accessibility: Unavailable
27°C
Haze
AutoSave Off
IMDB_Movies
File Home Insert Page Layout Formulas
A1
Row Labels
A
B
C
D
E
97 8.2
98 A Beautiful MindÂ
99 Blade RunnerÂ
100 Captain America: Civil WarÂ
101 CasinoÂ
102 Die HardÂ
103 Finding NemoÂ
104 Gone with the WindÂ
105 Gran TorinoÂ
106 How to Train Your DragonÂ
107 Howl's Moving CastleÂ
108 IncendiesÂ
109 Into the WildÂ
110 Lock, Stock and Two Smoking BarrelsÂ
111 On the WaterfrontÂ
112 Pan's LabyrinthÂ
113 The Act of KillingÂ
114 The Big LebowskiÂ
115 The Bridge on the River KwaiÂ
116 The Secret in Their EyesÂ
117 The ThingÂ
118 The Wolf of Wall StreetÂ
119 TrainspottingÂ
120 V for VendettaÂ
121 WarriorÂ
122 8.1
123 12 Years a SlaveÂ
124 AkiraÂ
125 Amores PerrosÂ
126 Annie HallÂ
127 Before SunriseÂ
128 Butch Cassidy and the Sundance KidÂ
129 DeadpoolÂ
IMDB_Movies
Ready
Accessibility: Unavailable
27°C
Haze


AutoSave Off
IMDB_Movies
File Home Insert Page Layout Formulas Data
A1
Row Labels
127 Before SunriseÂ
128 Butch Cassidy and the Sundance KidÂ
129 DeadpoolÂ
130 Donnie DarkoÂ
131 Elite SquadÂ
132 Gone GirlÂ
133 Groundhog DayÂ
134 Guardians of the GalaxyÂ
135 Hotel RwandaÂ
136 In the Shadow of the MoonÂ
137 Jurassic ParkÂ
138 Kill Bill: Vol. 1Â
139 Mad Max: Fury RoadÂ
140 Million Dollar BabyÂ
141 Monsters, Inc.Â
142 No Country for Old MenÂ
143 Pirates of the Caribbean: The Curse of the Black PearlÂ
144 PlatoonÂ
145 PrisonersÂ
146 RockyÂ
147 RushÂ
148 Shutter IslandÂ
149 Sin CityÂ
150 SpotlightÂ
151 Stand by MeÂ
152 Tae Guk Gi: The Brotherhood of WarÂ
153 The AvengersÂ
154 The Best Years of Our LivesÂ
155 The Bourne UltimatumÂ
156 The CelebrationÂ
157 The Grand Budapest HotelÂ
158 The HelpÂ
159 The Imitation GameÂ
IMDB_Movies
Ready Accessibility: Unavailable
27°C Haze
AutoSave Off
IMDB_
File Home Insert Page Lay
A1
160 The MartianÂ
161 The Princess BrideÂ
162 The RevenantÂ
163 The Sea InsideÂ
164 The Sixth SenseÂ
165 The TerminatorÂ
166 The Truman ShowÂ
167 The Wizard of OzÂ
168 There Will Be BloodÂ
169 WoodstockÂ
170 8
171 A Fistful of DollarsÂ
172 AladdinÂ
173 Before SunsetÂ
174 Big FishÂ
175 Black SwanÂ
176 Blood DiamondÂ
177 Blood In, Blood OutÂ
178 Bowling for ColumbineÂ
179 BoyhoodÂ
180 BrazilÂ
181 Casino RoyaleÂ
182 Catch Me If You CanÂ
183 Central StationÂ
184 Cinderella ManÂ
185 Dallas Buyers ClubÂ
186 Dancer in the DarkÂ
187 Dances with WolvesÂ
188 Dead Poets SocietyÂ
189 District 9Â
190 Doctor ZhivagoÂ
191 Fiddler on the RoofÂ
192 HerÂ
IMDB_Movies
Ready Accessibility: Unavailable
27°C Haze


AutoSave Off
IMDB_Movies
File Home Insert Page Layout Fc
A1
Row Labels
193 In BrugesÂ
194 JawsÂ
195 JFKÂ
196 Kill Bill: Vol. 2Â
197 Life of PiÂ
198 MagnoliaÂ
199 Mulholland DriveÂ
200 My Name Is KhanÂ
201 Mystic RiverÂ
202 PersepolisÂ
203 Rain ManÂ
204 RatatouilleÂ
205 SerenityÂ
206 Shaun of the DeadÂ
207 SickoÂ
208 Sling BladeÂ
209 Slumdog MillionaireÂ
210 Star TrekÂ
211 The ArtistÂ
212 The ExorcistÂ
213 The IncrediblesÂ
214 The Iron GiantÂ
215 The King's SpeechÂ
216 The Perks of Being a WallflowerÂ
217 The Pursuit of HappynessÂ
218 The Sound of MusicÂ
219 The Straight StoryÂ
220 True RomanceÂ
221 Waltz with BashirÂ
222 X-Men: Days of Future PastÂ
223 Young FrankensteinÂ
224 7.9
225 4 Months, 3 Weeks and 2 DaysÂ
IMDB_Movies
Ready
Accessibility: Unavailable
27°C
Haze
AutoSave Off
IMDB_Movies
File Home Insert Page Layout
A250
MoonÂ
218 The Sound of MusicÂ
219 The Straight StoryÂ
220 True RomanceÂ
221 Waltz with BashirÂ
222 X-Men: Days of Future PastÂ
223 Young FrankensteinÂ
224 7.9
225 4 Months, 3 Weeks and 2 DaysÂ
226 Almost FamousÂ
227 AmourÂ
228 AvatarÂ
229 Before MidnightÂ
230 Big Hero 6Â
231 Boogie NightsÂ
232 Captain PhillipsÂ
233 Children of MenÂ
234 CrashÂ
235 Crouching Tiger, Hidden DragonÂ
236 Do the Right ThingÂ
237 E.T. the Extra-TerrestrialÂ
238 Ed WoodÂ
239 Edge of TomorrowÂ
240 Edward ScissorhandsÂ
241 Ernest & CelestineÂ
242 GloryÂ
243 HalloweenÂ
244 HeroÂ
245 Hot FuzzÂ
246 How to Train Your Dragon 2Â
247 Iron ManÂ
248 Letters from Iwo JimaÂ
249 Little Miss SunshineÂ
250 MoonÂ
IMDB_Movies
Ready
Accessibility: Unavailable
27°C
Haze

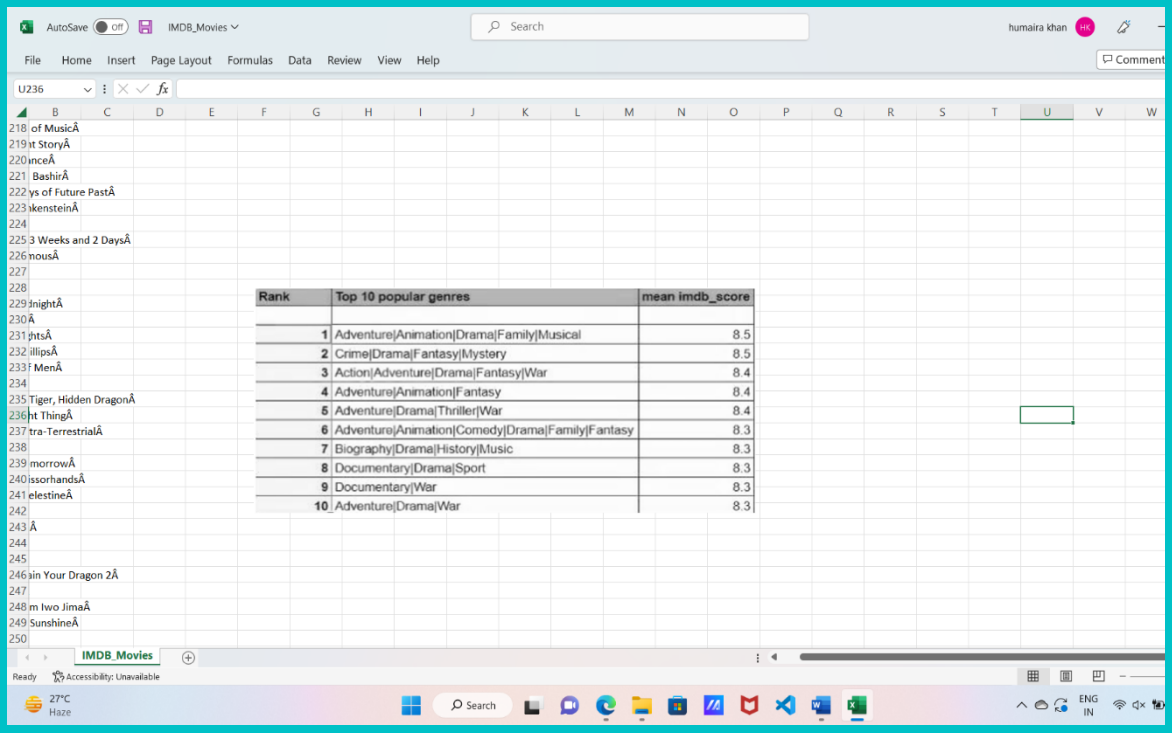
4) FORTH QUERY

NOW THE NEXT ONE WAS TO CALCULATE TOP 10 DIRECTORS WITH A MEAN OF HIGHEST RATING MOVIES. AGAIN WE USED THE PIVOT TABLE.



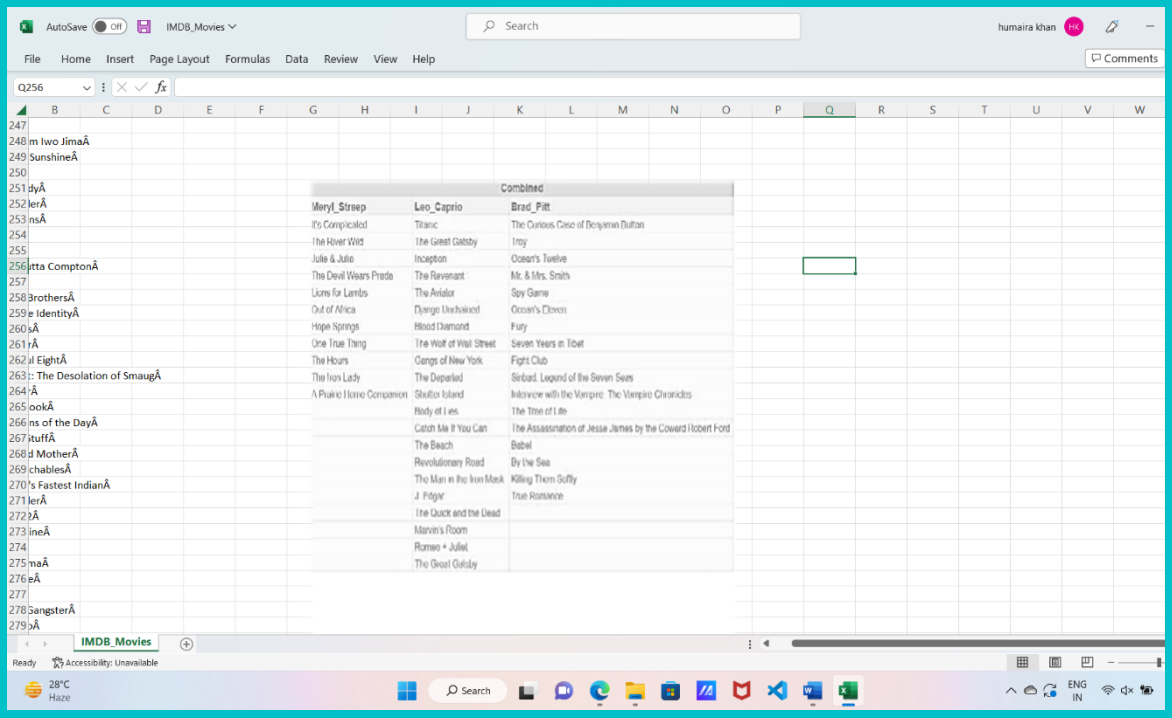
	A	B
2		
3	**Row Labels**	**Average of imdb_score**
4	Akira Kurosawa	8.7
5	Charles Chaplin	8.6
6	Tony Kaye	8.6
7	Alfred Hitchcock	8.5
8	Damien Chazelle	8.5
9	Majid Majidi	8.5
10	Ron Fricke	8.5
11	Sergio Leone	8.433333333
12	Christopher Nolan	8.425
13	Asghar Farhadi	8.4
14	Richard Marquand	8.4
15	**Grand Total**	**8.47**
16		

5)THE FIFTH QUERY WAS TO FIND THE POPULAR GENRES. WE DID THIS BY SEEING THE IDMB SCOREES.TOP 10 GENRES ARE:

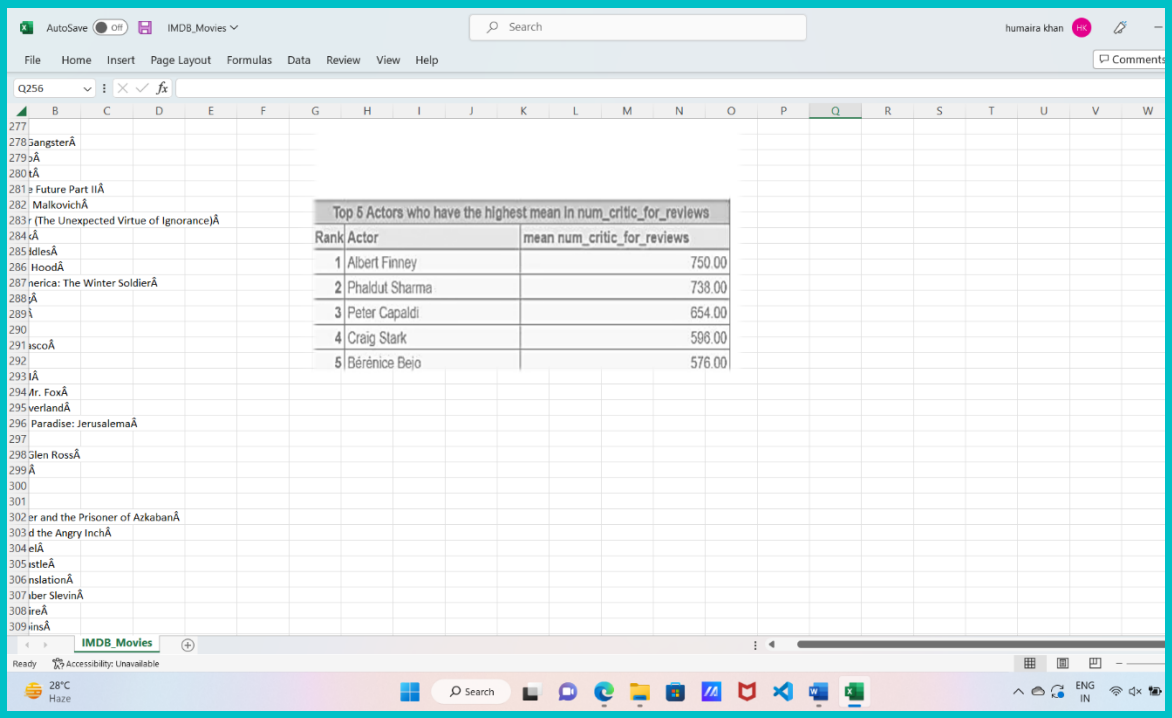


Rank	Top 10 popular genres	mean imdb_score
1	Adventure|Animation|Drama|Family|Musical	8.5
2	Crime|Drama|Fantasy|Mystery	8.5
3	Action|Adventure|Drama|Fantasy|War	8.4
4	Adventure|Animation|Fantasy	8.4
5	Adventure|Drama|Thriller|War	8.4
6	Adventure|Animation|Comedy|Drama|Family|Fantasy	8.3
7	Biography|Drama|History|Music	8.3
8	Documentary|Drama|Sport	8.3
9	Documentary|War	8.3
10	Adventure|Drama|War	8.3

6)THE LAST QUERY WAS REALY CHALLENGING. 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 THE NAMES 'MERYL STREEP', 'LEONARDO DICAPRIO', AND 'BRAD PITT' FOR THE SAID EXTRACTION.

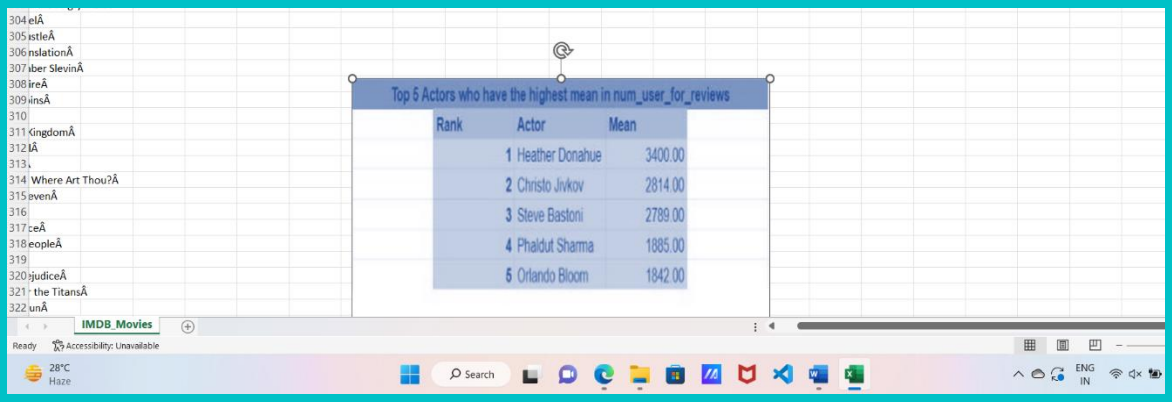


FIND THE MEAN OF THE NUM_CRITIC_FOR_REVIEWS AND NUM_USERS_FOR_REVIEW AND IDENTIFY THE ACTORS WHICH HAVE THE HIGHEST MEAN.



Top 5 Actors who have the highest mean in num_critic_for_reviews

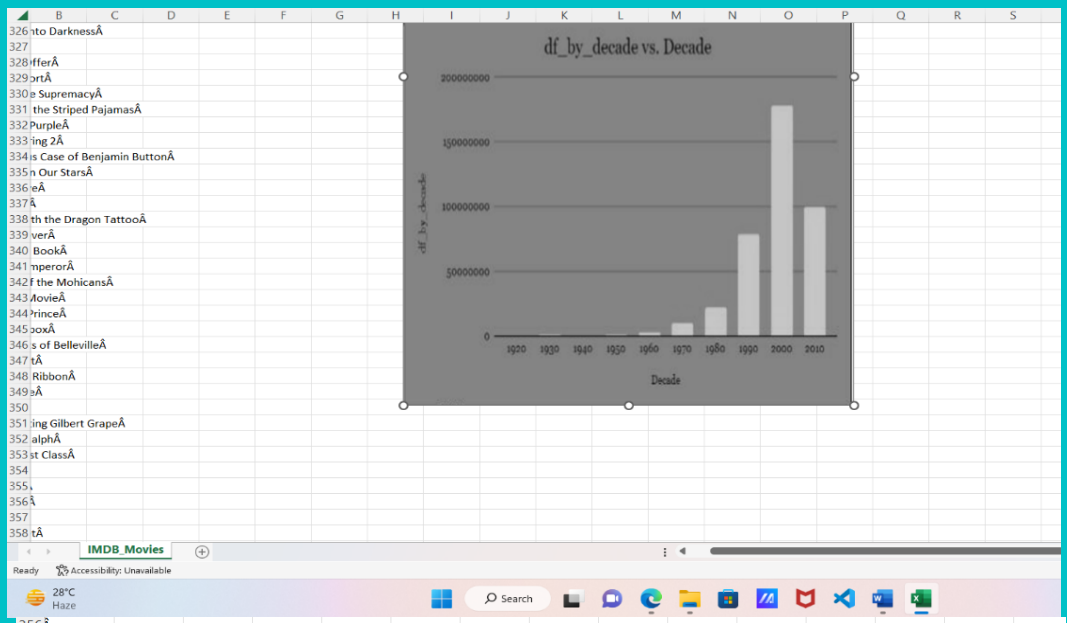
Rank	Actor	mean num_critic_for_reviews
1	Albert Finney	750.00
2	Phaldut Sharma	738.00
3	Peter Capaldi	654.00
4	Craig Stark	596.00
5	Bérénice Bejo	576.00



Top 5 Actors who have the highest mean in num_user_for_reviews

Rank	Actor	Mean
1	Heather Donahue	3400.00
2	Christo Jivkov	2814.00
3	Steve Bastoni	2789.00
4	Phaldut Sharma	1885.00
5	Orlando Bloom	1842.00

CREATE A COLUMN CALLED DECADE WHICH REPRESENTS THE DECADE TO WHICH EVERY MOVIE BELONGSTO. FOR EXAMPLE, THE TITLE_YEAR YEAR 1923, 1925 SHOULD BE STORED AS 1920S. SORT THE COLUMNBASED ON THE COLUMN DECADE, GROUP IT BY DECADE AND FIND THE SUM OF USERS VOTED IN EACHDECADE. STORE THIS IN A NEW DATA FRAME CALLED DF_BY_DECADE.



Decade	df_by_decade
1920	12467
1930	958599
1940	72324
1950	1097601
1960	2607791
1970	10196874
1980	22100624
1990	78706936
2000	178518291
2010	100145237

EXPLORATORY DATA ANALYSIS FOR BANK LOAN

PROJECT DESCRIPTION - THE LOAN PROVIDING COMPANIES FIND IT HARD TO GIVE LOANS TO THE PEOPLE DUE TO THEIR INSUFFICIENT OR NON-EXISTENT CREDIT HISTORY. BECAUSE OF THAT, SOME CONSUMERS USE IT AS THEIR ADVANTAGE BY BECOMING A DEFAULTER. WE HAVE TO USE EDA TO ANALYZE THE PATTERNS PRESENT IN THE DATA. THIS WILL ENSURE THAT THE APPLICANTS CAPABLE OF REPAYING THE LOAN ARE NOT REJECTED.
APPROACH - THIS CASE STUDY AIMS TO IDENTIFY PATTERNS WHICH INDICATE IF AN APPLICANT HAS DIFFICULTY IN PAYING HIS/HER 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.
I HAVE MAINLY FOCUSED ON ANALYZING PREVIOUS_APPLICATION.CSV I.E. DATA ABOUT PREVIOUS APPLICATION OF AN APPLICANT.
FOR THE EXPLORATORY DATA ANALYSIS, MENTIONED

STEPS HAVE BEEN FOLLOWED. -> IMPORT MODULES,

READ THE DATASET

SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMEN
2030495	271877	Consumer loans	1730.430	17145.0	17145.0	0.
2802425	108129	Cash loans	25188.615	607500.0	679671.0	Nal
2523466	122040	Cash loans	15060.735	112500.0	136444.5	Nal
2819243	176158	Cash loans	47041.335	450000.0	470790.0	Nal
1784265	202054	Cash loans	31924.395	337500.0	404055.0	Nal

```
Data columns (total 37 columns):
 #   Column                       Non-Null Count    Dtype
---  ------                       --------------    -----
 0   SK_ID_PREV                   1670214 non-null  int64
 1   SK_ID_CURR                   1670214 non-null  int64
 2   NAME_CONTRACT_TYPE           1670214 non-null  object
 3   AMT_ANNUITY                  1297979 non-null  float64
 4   AMT_APPLICATION              1670214 non-null  float64
 5   AMT_CREDIT                   1670213 non-null  float64
 6   AMT_DOWN_PAYMENT             774370 non-null   float64
 7   AMT_GOODS_PRICE              1284699 non-null  float64
 8   WEEKDAY_APPR_PROCESS_START   1670214 non-null  object
 9   HOUR_APPR_PROCESS_START      1670214 non-null  int64
 10  FLAG_LAST_APPL_PER_CONTRACT  1670214 non-null  object
 11  NFLAG_LAST_APPL_IN_DAY       1670214 non-null  int64
 12  RATE_DOWN_PAYMENT            774370 non-null   float64
 13  RATE_INTEREST_PRIMARY        5951 non-null     float64
 14  RATE_INTEREST_PRIVILEGED     5951 non-null     float64
 15  NAME_CASH_LOAN_PURPOSE       1670214 non-null  object
 16  NAME_CONTRACT_STATUS         1670214 non-null  object
 17  DAYS_DECISION                1670214 non-null  int64
 18  NAME_PAYMENT_TYPE            1670214 non-null  object
 19  CODE_REJECT_REASON           1670214 non-null  object
 20  NAME_TYPE_SUITE              849809 non-null   object
 21  NAME_CLIENT_TYPE             1670214 non-null  object
 22  NAME_GOODS_CATEGORY          1670214 non-null  object
 23  NAME_PORTFOLIO               1670214 non-null  object
 24  NAME_PRODUCT_TYPE            1670214 non-null  object
 25  CHANNEL_TYPE                 1670214 non-null  object
 26  SELLERPLACE_AREA             1670214 non-null  int64
 27  NAME_SELLER_INDUSTRY         1670214 non-null  object
 28  CNT_PAYMENT                  1297984 non-null  float64
 29  NAME_YIELD_GROUP             1670214 non-null  object
 30  PRODUCT_COMBINATION          1669868 non-null  object
 31  DAYS_FIRST_DRAWING           997149 non-null   float64
 32  DAYS_FIRST_DUE               997149 non-null   float64
 33  DAYS_LAST_DUE_1ST_VERSION    997149 non-null   float64
 34  DAYS_LAST_DUE                997149 non-null   float64
 35  DAYS_TERMINATION             997149 non-null   float64
 36  NFLAG_INSURED_ON_APPROVAL    997149 non-null   float64
```

PREV_AP_DF CONTAINS 37 FEATURES AND 1670214 ROWS(OUT OF WHICH 15 FEATURES ARE FLOAT64, 6 FEATURES ARE INTEGER, 16 FEATURES ARE OBJECT DATATYPE)

FOLLOWING ARE THE COMMON FEATURES AMONG APPLICATION DATA AND PREVIOUS APPLICATION DATA -

['SK_ID_CURR', 'NAME_CONTRACT_TYPE', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'NAME_TYPE_SUITE', 'WEEKDAY_APPR_PROCESS_START', 'HOUR_APPR_PROCESS_START']

SK_ID_CURR IS AN UNIQUE IDENTIFIER, WHICH WILL USE TO MERGE THE RELEVANT COLUMNS OF 2 DATAFRAMES (APPLICATION DATA AND PREVIOUS APPLICATION DATA).

DATA CLEANING

MISSING VALUE HANDLING, TYPE CASTING, FIXING ROWS AND COLUMNS – REMOVING UNNECESSARY ROWS/COLUMNS (THROUGH MISSING VALUE HANDLING AND CORRELATION), HANDLING OUTLIERS.
FIRST I HAVE CALCULATED THE MISSING VALUE PERCENTAGES FOR EACH FEATURE IN PREVIOUS APPLICATION DATA.

	category	percentage
5	RATE_INTEREST_PRIMARY	99.643698
6	RATE_INTEREST_PRIVILEGED	99.643698
2	AMT_DOWN_PAYMENT	53.636480
4	RATE_DOWN_PAYMENT	53.636480
7	NAME_TYPE_SUITE	49.119754
10	DAYS_FIRST_DRAWING	40.298129
11	DAYS_FIRST_DUE	40.298129
12	DAYS_LAST_DUE_1ST_VERSION	40.298129
13	DAYS_LAST_DUE	40.298129
14	DAYS_TERMINATION	40.298129
15	NFLAG_INSURED_ON_APPROVAL	40.298129
3	AMT_GOODS_PRICE	23.081773
0	AMT_ANNUITY	22.286665
8	CNT_PAYMENT	22.286366
9	PRODUCT_COMBINATION	0.020716
1	AMT_CREDIT	0.000060

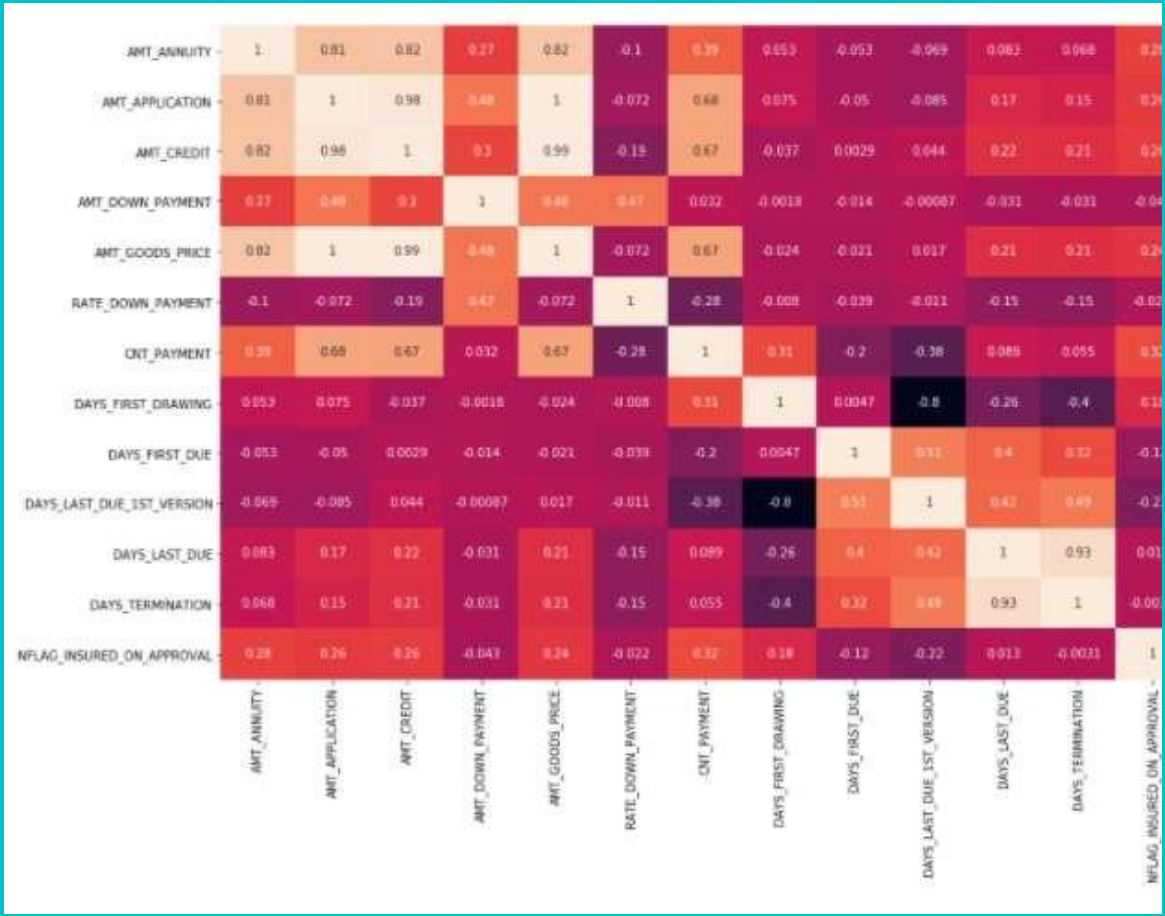
THERE ARE 16 FEATURES IN PREV_APP_DF THAT HAVE MISSING VALUES.

PERMANENTLY DROPPING THE FEATURES (RATE_INTEREST_PRIMARY AND RATE_INTEREST_PRIVILEGED) AS 99% DATA IS MISSING.
DROPPING ROWS CONTAINING MISSING VALUES FOR THE FEATURES (AMT_CREDIT AND PRODUCT_COMBINATION) FOR VERY LOW % OF MISSING DATA. DROPPING ENTRIES WOULD NOT CAUSE IMPACT THE ANALYSIS AS PERCENTAGE OF MISSING VALUE IS VERY LOW (~2%).

->

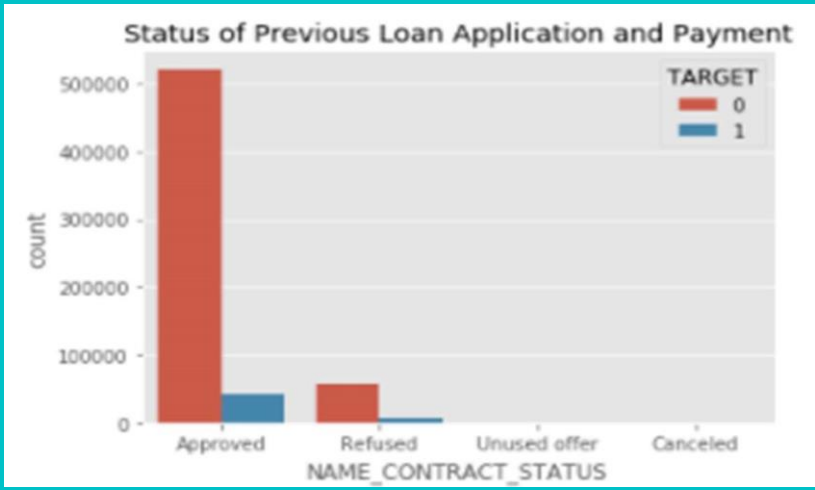
UNIVARIATE ANALYSIS, BIVARIATE AND MULTIVARIATE ANALYSIS

FIRST, I EXTRACTED THE NUMERICAL VARIABLES IN THE DATASET AND CHECKED OUT THE CORRELATION COEFFICIENTS WITH THE HELP OF A HEATMAP.



'DAYS_LAST_DUE' AND 'DAYS_TERMINATION' ARE HIGHLY CORRELATED
'DAYS_FIRST_DRAWING' AND 'DAYS_LAST_DUE_1ST_VERSION' HAVE HIGH NEGATIVE CORRELATION
'AMT_ANNUITY','AMT_APPLICATION','AMT_CREDIT','AMT_GOODS_PRICE' ARE HIGHLY CORRELATED
THE FEATURES CAN BE REMOVED BEFORE MODELLING THIS DATA, AS THEY WOULD CAUSE COLLINEARITY
'DAYS_TERMINATION','DAYS_LAST_DUE_1ST_VERSION','AMT_APPLICATION','AMT_CREDIT','AMT_GOODS_ PRICE' FOR EDA PURPOSE WE ARE NOT REMOVING THEM.

CHECKING DATA IMBALANCE IN PREVIOUS APPLICATION DATA



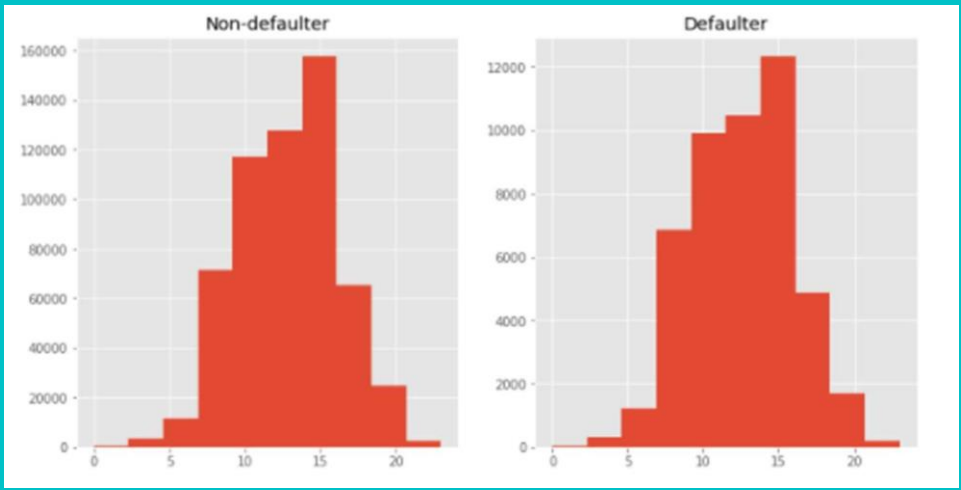
THE APPLICANTS WHOSE PREVIOUS LOANS WERE APPROVED ARE MORE LIKELY TO PAY CURRENT LOAN IN TIME, THAN THE APPLICANTS WHOSE PREVIOUS LOANS WERE REJECTED.

7% OF THE PREVIOUSLY APPROVED LOAN APPLICANTS THAT DEFAULTED IN CURRENT LOAN

90 % OF THE PREVIOUSLY REFUSED LOAN APPLICANTS THAT WERE ABLE TO PAY CURRENT LOAN

THIS DATA IS HIGHLY IMBALANCED AS NUMBER OF DEFAULTERS IS VERY LESS IN TOTAL POPULATION.

ANALYSIS OF NUMERIC FEATURES OF PREVIOUS APPLICATION DATA



NUMBER OF

DEFAULTERS ARE LESS FOR LARGER AMOUNT OF ANNUITY OF PREVIOUS APPLICATION.

FOR HIGHER DOWN PAYMENT, DEFAULTER CASES ARE LESS.

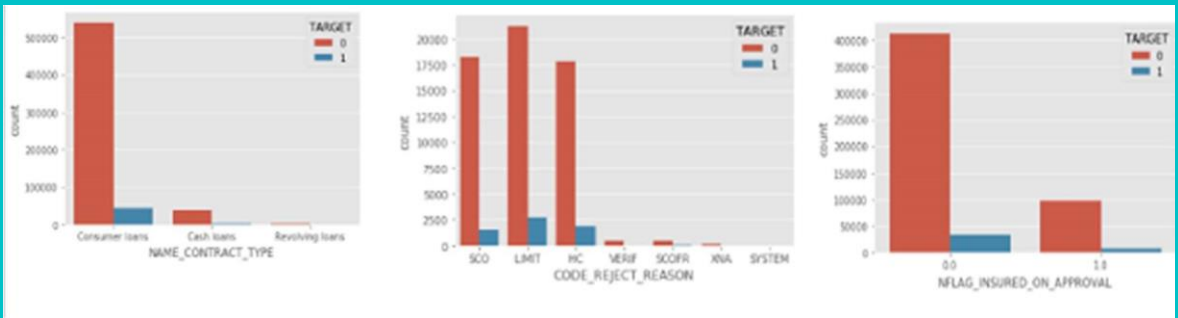
MOST OF THE LOANS ARE APPLIED AROUND 15:00 HOURS. THIS FEATURE IS DOES NOT HAVE VISIBLE IMPACT ON TARGET VARIABLE.

ANALYSIS OF CATEGORICAL FEATURES OF

PREVIOUS APPLICATION DATA HIGHEST

NUMBER OF LOANS ARE APPLIED FOR

CONSUMER LOANS.



AS SEEN IN THE ABOVE PLOT, 'SCO', 'LIMIT' AND 'HC' ARE THE MOST COMMON REASON OF REJECTION.
MOST OF THE PEOPLE DID NOT REQUEST INSURANCE DURING PREVIOUS LOAN APPLICATION.
MOST OF THE APPLICANTS ARE REPEATER.
'CASH THROUGH THE BANK' IS THE MOST

FREQUENTLY USED PAYMENT METHOD. THEN FOR A

GIVEN CATEGORICAL FEATURE, I OBTAINED

PERCENTAGE OF DEFAULTERS.

	Value	Percentage of Defaulter
23	Insurance	10.526316
0	Vehicles	10.257410
14	Jewelry	9.124951
17	Auto Accessories	9.029763
3	Mobile	8.615336
15	Office Appliances	8.307692
8	Computers	8.074335
20	Weapon	8.064516
21	Direct Sales	8.024691
5	Audio/Video	7.698706
7	Photo / Cinema Equipment	7.455000
18	Sport and Leisure	7.354150
2	Consumer Electronics	7.066548
4	Construction Materials	6.978320
9	XNA	6.885879
24	Additional Service	6.730769
6	Gardening	6.723063
11	Homewares	6.706444
19	Medicine	6.196747
25	Education	5.882353
1	Furniture	5.860781
10	Clothing and Accessories	5.807427
13	Other	5.765921
12	Medical Supplies	5.564190
16	Tourism	4.444444
22	Fitness	4.268293
26	Animals	0.000000

Highest percentage of default cases are for t

	Value	Percentage of Defaulter
1	walk-in	9.165550
0	XNA	7.665995
2	x-sell	6.036420

highest.

	Value	Percentage of Defaulter
1	walk-in	9.165550
0	XNA	7.665995
2	x-sell	6.036420

From all the walk-in applicants 9% defaulte
loan.

	Value	Percentage of Defaulter
4	AP+ (Cash loan)	15.000000
1	Country-wide	7.908171
2	Regional / Local	7.551291
0	Stone	7.294692
3	Credit and cash offices	6.124197
5	Contact center	4.545455
6	Car dealer	0.000000

WHO PREVIOUSLY APPLIED FOR INSURANCE AND VEHICLES.

FOR CARDS DEFAULTER RATE 15% LOAN APPLICANTS DEFAULTED FOR

	Value	Percentage of Defaulter
0	Auto technology	10.522088
9	Jewelry	9.019221
3	Connectivity	8.780637
2	Consumer electronics	7.451983
7	Industry	7.211664
4	Construction	6.597424
5	XNA	6.226598
1	Furniture	5.924492
6	Clothing	5.857399
8	Tourism	4.778157
10	MLM partners	4.654655

AP + CASH LOAN

IN SELLER INDUSTRY "AUTO TECHNOLOGY" HAS HIGHEST RATE OF DEFAULTER
MLM PARTNERS HAS LOWEST NUMBER OF DEFAULTERS

	Value	Percentage of Defaulter
4	XNA	17.119695
2	high	8.340935
1	middle	7.558098
0	low_normal	6.844973
3	low_action	6.608936

DEFAULTER PERCENTAGE IS HIGHEST WHERE NAME_YIELD_GROUP IS NOT KNOWN.
HIGHEST PERCENTAGE OF DEFAULT CASES IS FOR CARD STREET.

	Value	Percentage of Defaulter
13	Card Street	17.195005
4	POS mobile with interest	8.761056
0	POS other with interest	7.953141
3	POS mobile without interest	7.888514
2	POS household with interest	7.752151
11	POS others without interest	7.256127
15	Card X-Sell	6.666667
5	POS household without interest	6.649376
9	Cash Street: middle	6.475391
10	Cash Street: high	6.417625
8	Cash X-Sell: high	6.410114
1	POS industry with interest	6.350635
12	Cash X-Sell: middle	6.017039
7	Cash Street: low	5.976676
6	POS industry without interest	4.711940
14	Cash X-Sell: low	3.986711

SUMMARY OF PREVIOUS APPLICATION DATA

THERE ARE FEATURE COLUMNS IN THE DATASET THAT ARE HIGHLY CORRELATED TO EACH OTHER. WHICH MEANS BOTH WILL HAVE SIMILAR IMPACT ON THE TARGET VALUE. THOSE FEATURES CAN BE REMOVED BEFORE FEEDING THIS DATA TO A MODEL TO AVOID COLLINEARITY.
FEATURE COLUMNS WITH 50% OR MORE MISSING DATA CAN BE DROPPED.
THIS DATASET IS HIGHLY IMBALANCED
THE APPLICANTS WHOSE PREVIOUS LOANS WERE APPROVED ARE MORE LIKELY TO PAY CURRENT LOAN IN TIME, THAN THE APPLICANTS WHOSE PREVIOUS LOANS WERE REJECTED. NAME_CONTRACT_STATUS IS AN IMPORTANT FEATURE.
7% OF THE PREVIOUSLY APPROVED LOAN APPLICANTS

THAT DEFAULTED IN CURRENT LOAN 90 % OF THE

PREVIOUSLY REFUSED LOAN APPLICANTS THAT WERE

ABLE TO PAY CURRENT LOAN.

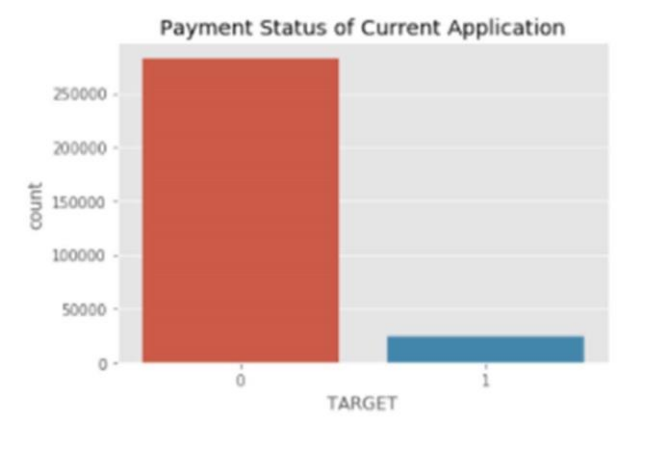
'SCO', 'LIMIT' AND 'HC' ARE THE MOST COMMON REASON OF REJECTION.
MOST OF THE PEOPLE DID NOT REQUEST INSURANCE DURING PREVIOUS LOAN APPLICATION.
FOR "CARDS" DEFAULTER PERCENTAGE IS HIGHEST (17%). 15% LOAN APPLICANT DEFAULTED FOR AP+ (CASH LOAN).

CURRENT APPLICATION DATA ANALYSIS:

FOR ANALYZING CURRENT APPLICATION DATA, I HAVE TAKEN A DIFFERENT APPROACH THAN THAT OF PREVIOUS APPLICATION DATA. BY TAKING A CLOSE LOOK AT THE FEATURES, I COULD IDENTIFY THE FEATURES OF DIFFERENT ASPECTS OF THE LOAN APPLICANT.
THAT'S WHY I HAVE DIVIDED THE FEATURES INTO SMALL SEGMENTS AND ANALYZED THEM SEGMENT-WISE USING A SMALLER DATA-FRAME CONTAINING ONLY RELEVANT CATEGORIES.
DATA CLEANING, MISSING DATA HANDLING, TYPE CASTING ARE DONE SEGMENT-WISE.
PLOTS AND PERCENTAGE WISE DEFAULTER CALCULATION ARE DONE SEGMENT-WISE AS WELL.

CHECKING DATA IMBALANCE:

THIS DATA IS HIGHLY IMBALANCED AS NUMBER OF DEFAULTERS IS VERY LESS IN TOTAL POPULATION. DATA IMBALANCE RATIO DEFAULTER : NON-DEFAULTER = 8 : 92 = 2 : 23.



HOUSING INFORMATION OF APPLICANT

ALL OF THE FEATURES HAVE VERY HIGH (47-70%) MISSING DATA PERCENTAGE. HENCE ALL THESE FEATURES CAN BE DROPPED.

MOST OF THE APPLICANTS LIVE IN HOUSE/APARTMENT

APPLICANTS LIVING WITH THEIR PARENTS OR IN RENTED APARTMENT HAVE HIGHER RATE OF DEFAULT.

	category	percentage
46	EMERGENCYSTATE_MODE	47.398304
44	TOTALAREA_MODE	48.268517
2	YEARS_BEGINEXPLUATATION_AVG	48.781019
30	YEARS_BEGINEXPLUATATION_MEDI	48.781019
16	YEARS_BEGINEXPLUATATION_MODE	48.781019
35	FLOORSMAX_MEDI	49.760822
7	FLOORSMAX_AVG	49.760822
21	FLOORSMAX_MODE	49.760822
43	HOUSETYPE_MODE	50.176091
39	LIVINGAREA_MEDI	50.193326
11	LIVINGAREA_AVG	50.193326
25	LIVINGAREA_MODE	50.193326
6	ENTRANCES_AVG	50.348768
34	ENTRANCES_MEDI	50.348768
20	ENTRANCES_MODE	50.348768
28	APARTMENTS_MEDI	50.749729
0	APARTMENTS_AVG	50.749729
14	APARTMENTS_MODE	50.749729
45	WALLSMATERIAL_MODE	50.840783
19	ELEVATORS_MODE	53.295980
5	ELEVATORS_AVG	53.295980
33	ELEVATORS_MEDI	53.295980
13	NONLIVINGAREA_AVG	55.179164
41	NONLIVINGAREA_MEDI	55.179164
27	NONLIVINGAREA_MODE	55.179164
1	BASEMENTAREA_AVG	58.515956
29	BASEMENTAREA_MEDI	58.515956
15	BASEMENTAREA_MODE	58.515956
23	LANDAREA_MODE	59.376738
37	LANDAREA_MEDI	59.376738
9	LANDAREA_AVG	59.376738
3	YEARS_BUILD_AVG	66.497784

	Value	Percentage of Defaulter
1	Rented apartment	12.313051
2	With parents	11.698113
3	Municipal apartment	8.539748
5	Co-op apartment	7.932264
0	House / apartment	7.795711
4	Office apartment	6.572411

ASSET DETAILS

MOST OF THE APPLICANT'S OWN REALTY
MOST OF THE APPLICANTS DO NOT OWN CARS
PEOPLE NOT OWNING REALITY AND CAR AND HAVE A SLIGHTLY HIGHER DEFAULT RATE THAN THE PEOPLE WHO OWN REALITY AND CAR.
DEFAULTER OR NOT, MOST APPLICANTS HAVE CAR AGE BETWEEN 0-25 YEARS.
SINCE FOR BOTH TARGET VALUE, TREND IS SIMILAR, THIS FEATURE CAN DROP.

FAMILY RELATED INFO

DEFAULT RATE IS HIGHEST FOR CIVIL MARRIAGE AND SINGLE APPLICANTS
MOST OF THE APPLICANTS ARE MARRIED (AND/OR) NO CHILDREN (AND/OR) 2 FAMILY MEMBERS.
APPLICANTS WITH RELATIVELY GREATER NUMBER OF CHILDREN (AND/OR) FAMILY MEMBERS HAVE HIGHER DEFAULT PERCENTAGE.
FOR SOME OF THE CASES WHERE COUNT CHILDREN/FAMILY MEMBERS IS HIGH, AND THE DEFAULT RATE IS VERY HIGH OR VERY LOW. THESE CASES CANNOT BE TAKEN AS A CONCLUSION AS NUMBER OF APPLICANTS HAVING A LARGE FAMILY IS VERY LOW.

EDUCATION AND OCCUPATION INFO

MOST OF THE APPLICANTS ARE WORKING.
APPLICANTS ON MATERNITY LEAVE AND UNEMPLOYED HAS HIGHEST PERCENTAGE OF DEFAULTER
BUSINESSMAN HAVE LOWEST (0) PERCENTAGE OF DEFAULTER HOWEVER APPLICANTS OF INCOME TYPE('UNEMPLOYED', 'STUDENT', 'BUSINESSMAN', 'MATERNITY LEAVE') ARE VERY FEW IN THE DATASET TO CONTRIBUTE IN THE ANALYSIS.
APPLICANTS HAVING "LOWER SECONDARY" EDUCATION HAVE HIGHEST PERCENTAGE OF DEFAULTER.
LOW SKILLED LABORERS HAVE VERY HIGH RATE OF DEFAULTERS IN COMPARISON TO OTHER OCCUPATIONS.

FEMALE APPLICANTS ARE MORE THAN MALE APPLICANTS

DEFAULTER PERCENTAGE IS

HIGHER FOR MALE APPLICANTS

PEOPLE OF AGE 30 HAVE HIGHER

DEFAULT RATE

	Value	Percentage of Defaulter
7	Maternity leave	40.000000
4	Unemployed	36.363636
0	Working	9.588472
2	Commercial associate	7.484466
1	State servant	5.754965
3	Pensioner	5.386366
5	Student	0.000000
6	Businessman	0.000000

DEFAULT CASES ARE LESS FOR APPLICANTS MORE THAN 40 YEARS OLD.

MOST OF THE APPLICANTS ARE
UNACCOMPANIED WHILE APPLYING
FOR LOAN

NUMBER CASH LOANS IS QUITE HIGHER THAN REVOLVING LOANS

ALL WEEKDAYS HAVE SIMILAR NUMBER OF APPLICANTS THAN WEEKEND (SATURDAY AND SUNDAY)

BOXPLOT IS SHOWING THE OUTLIERS FOR INCOME AND ANNUITY, THERE ARE FEW ENTRIES HAVING VERY LARGE ANNUITY AND INCOME THAN OTHERS.

TOP 10 CORRELATION OF DEFAULTERS

AMT_REQ_CREDIT_BUREAU_YEAR	AMT_REQ_CREDIT_BUREAU_YEAR	1.000000
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269
BASEMENTAREA_AVG	BASEMENTAREA_MEDI	0.998250
COMMONAREA_AVG	COMMONAREA_MEDI	0.998107
YEARS_BUILD_MEDI	YEARS_BUILD_AVG	0.998100
NONLIVINGAPARTMENTS_AVG	NONLIVINGAPARTMENTS_MEDI	0.998075
FLOORSMIN_AVG	FLOORSMIN_MEDI	0.997825
LIVINGAPARTMENTS_AVG	LIVINGAPARTMENTS_MEDI	0.997668
FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997187
NONLIVINGAPARTMENTS_MEDI	NONLIVINGAPARTMENTS_MODE	0.997032
ENTRANCES_MEDI	ENTRANCES_AVG	0.996700

TOP 10 CORRELATION OF NON DEFAULTERS

AMT_REQ_CREDIT_BUREAU_YEAR	AMT_REQ_CREDIT_BUREAU_YEAR	1.000000
YEARS_BUILD_AVG	YEARS_BUILD_MEDI	0.998522
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508
FLOORSMIN_MEDI	FLOORSMIN_AVG	0.997202
FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997018
ENTRANCES_MEDI	ENTRANCES_AVG	0.996899
ELEVATORS_AVG	ELEVATORS_MEDI	0.996161
COMMONAREA_MEDI	COMMONAREA_AVG	0.995857
LIVINGAREA_AVG	LIVINGAREA_MEDI	0.995568
APARTMENTS_AVG	APARTMENTS_MEDI	0.995163
BASEMENTAREA_MEDI	BASEMENTAREA_AVG	0.994081

RESULT - THIS DATA IS HIGHLY IMBALANCED AS NUMBER OF DEFAULTERS IS VERY LESS IN TOTAL POPULATION.
MOST OF THE APPLICANTS LIVE IN HOUSE/APARTMENT II. APPLICANTS LIVING WITH THEIR PARENTS OR IN RENTED APARTMENT HAVE HIGHER RATE OF DEFAULT.

SOCIAL CIRCLE INFO: THE FEATURES SHOW SIMILAR TREND FOR DEFAULTERS AND NON-DEFAULTERS, CAN BE DROPPED.
ASSET INFO - MOST OF THE APPLICANT'S OWN REALTY II. MOST OF THE APPLICANTS DO NOT OWN CARS III. PEOPLE NOT OWNING REALITY AND CAR AND HAVE A SLIGHTLY HIGHER DEFAULT RATE THAN THE PEOPLE WHO OWN REALITY AND CAR.

XYZ ADS AIRING REPORT ANALYSIS

DESCRIPTION:

IN SIMPLER TERMS, ADVERTISING ANALYTICS CAN BE REFERRED TO AS THE USE OF ANALYTICAL DATA AND TOOLS THAT HELP BUSINESSES AND MARKETERS EFFICIENTLY MONITOR THEIR OMNICHANNEL MARKETING EFFORTS.AS THESE DATA SETS OFFER ACTIONABLE INSIGHTS, MARKETERS CAN USE THEM TO REASSURE THAT THE CAMPAIGNS THEY RUN ARE TARGETED TO THE RIGHT AUDIENCE AND USE THE RIGHT MEDIUM TO DO SO.

AS PREVIOUSLY MENTIONED, ORGANIZATIONS NEVER RESORT TO A SINGLE MARKETING CHANNEL AS THEIR TARGET AUDIENCES WILL BE SPREAD ACROSS MULTIPLE CHANNELS. TO BE MORE PRECISE, ORGANIZATIONS DON'T HAVE A CHOICE OTHER THAN TO RUN CROSS-CHANNEL CAMPAIGNS AS MULTI-CHANNEL CUSTOMERS TEND TO SPEND TWO TO FOUR TIMES MORE THAN SINGLE-CHANNEL CUSTOMERS.

RESULTS

A.WHAT IS POD POSITION? DOES THE POD POSITION NUMBER AFFECT THE AMOUNT SPENT ON ADS FOR A SPECIFIC PERIOD OF TIME BY A COMPANY? (EXPLAIN IN DETAILS WITH EXAMPLES FROM THE DATASET PROVIDED)

AN AD POD IS A GROUP OF ADS THAT ARE SEQUENCED TOGETHER TO BE PLAYED BACK-TO-BACK WITHIN A SINGLE AD BREAK/PLACEMENT, SIMILAR TO AD BREAKS IN TRADITIONAL LINEAR TV. AD PODS GIVE PUBLISHERS THE OPPORTUNITY TO MAXIMISE REVENUE FROM EACH AD BREAK AND GIVE ADVERTISERS MORE CONTROL OVER AD POSITIONING.

THEY ALLOW PUBLISHERS TO RETURN MULTIPLE ADS FROM A SINGLE AD REQUEST, AND THEN THOSE ADS ARE PLAYED IN SEQUENCE.
FOUR REASONS WHY ADVERTISERS AND PUBLISHERS USE AD PODS

1) THEY OFFER MORE CONTROL AD PODS HELP ADVERTISERS AVOID RUNNING ADS ALONGSIDE DIRECT COMPETITORS, ENSURING THAT THEIR OFFERING STANDS OUT TO VIEWERS AND THAT THEIR MESSAGE DOESN'T GET SATURATED

2) AD PODS OFFER A BETTER WAY TO MONETIZE LONG-FORM CONTENT PUBLISHERS WITH LONGER-FORM CONTENT CAN LEVERAGE THE CONTROLS OFFERED BY AD PODDING TO SET UP MORE ADVANCED MONETIZATION STRATEGIES FOR THEIR STREAMING CONTENT.

3)– AD PODS ALLOW PUBLISHERS TO MEET BUYERS' NEEDS WITHOUT AN AD POD IN PLACE, ADVERTISERS HAVE HISTORICALLY BEEN SPINNING A WHEEL OF CHANCE WHEN THEY BOUGHT INTO STREAMING APPS ON CTV. THEY HAVE HAD VERY LITTLE CONTROL OVER FREQUENCY OR THE POSITION OF THEIR AD WITHIN THE AD BREAK.

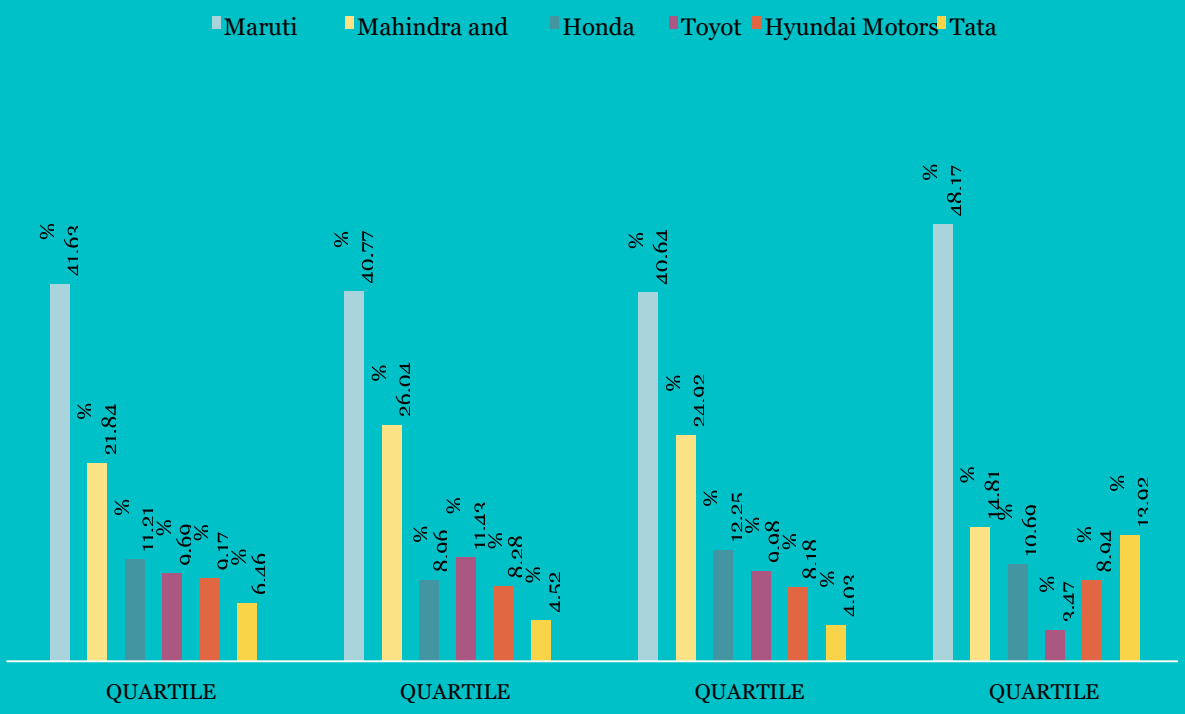
4)–AD PODS IMPROVE THE USER EXPERIENCE AD PODS ENABLE A BETTER USER EXPERIENCE BECAUSE THE ADS ARE TAILORED TO VIEWERS' PREFERENCES.

Pod Positions	Honda Cars	Hyundai Motors India	Mahindra and Mahindra	Maruti Suzuki	Tata Motors	Toyota
1	26%	28%	26%	37%	30%	26%
2	23%	24%	21%	21%	22%	22%
3	20%	20%	21%	18%	20%	21%
4	17%	16%	18%	14%	16%	18%
5	14%	12%	15%	11%	13%	14%

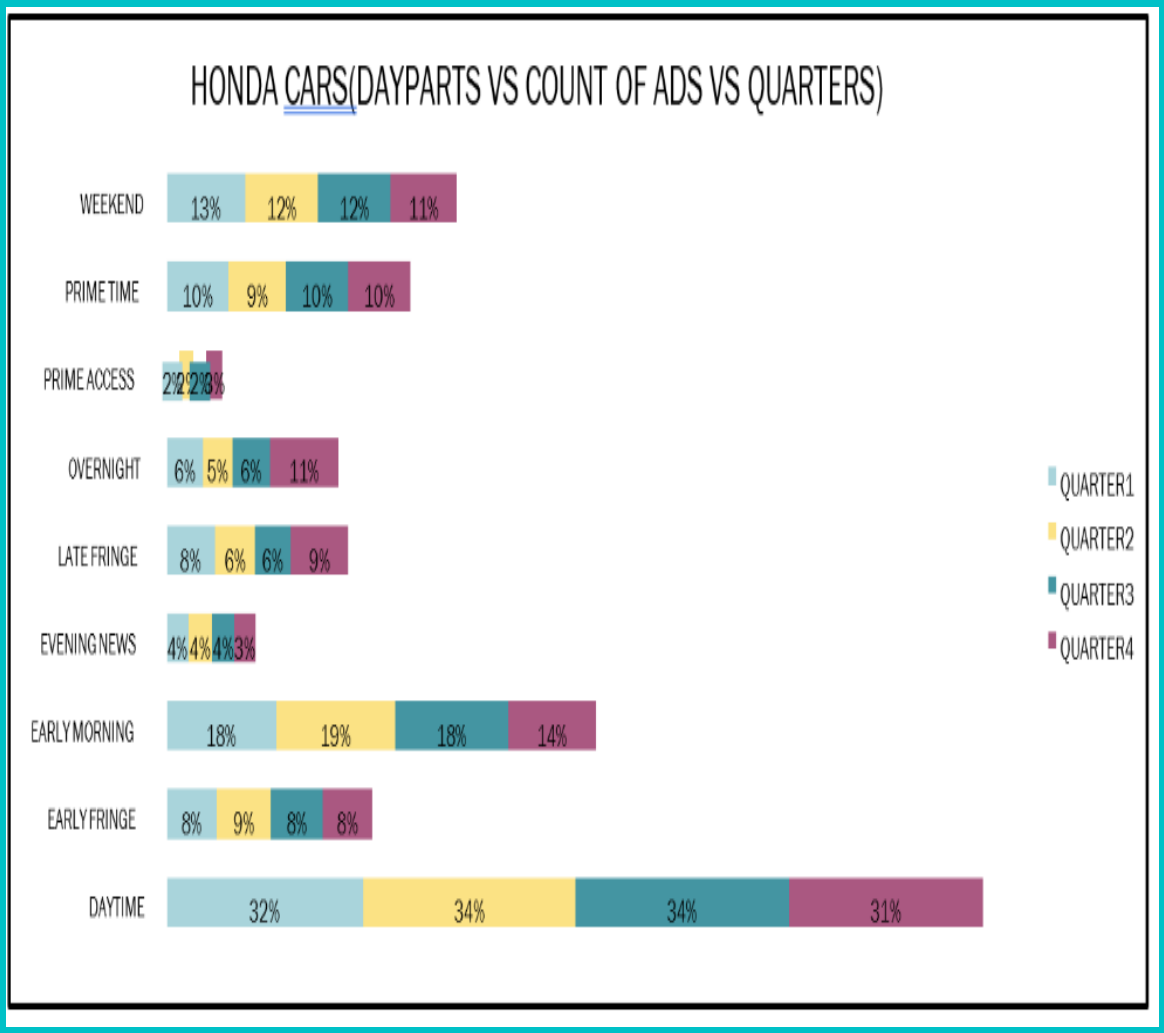
- THE ABOVE ANALYSIS SHOWS THAT ALL BRANDS HAVE THE HIGHEST CONTRIBUTION OF EQ UNITS IN 1 ST POD_POSITION AMONG THE TOP 5 POD_POSITIONS.

DISTRIBUTION OF ADS AS PER SUM OF EQ UNITS

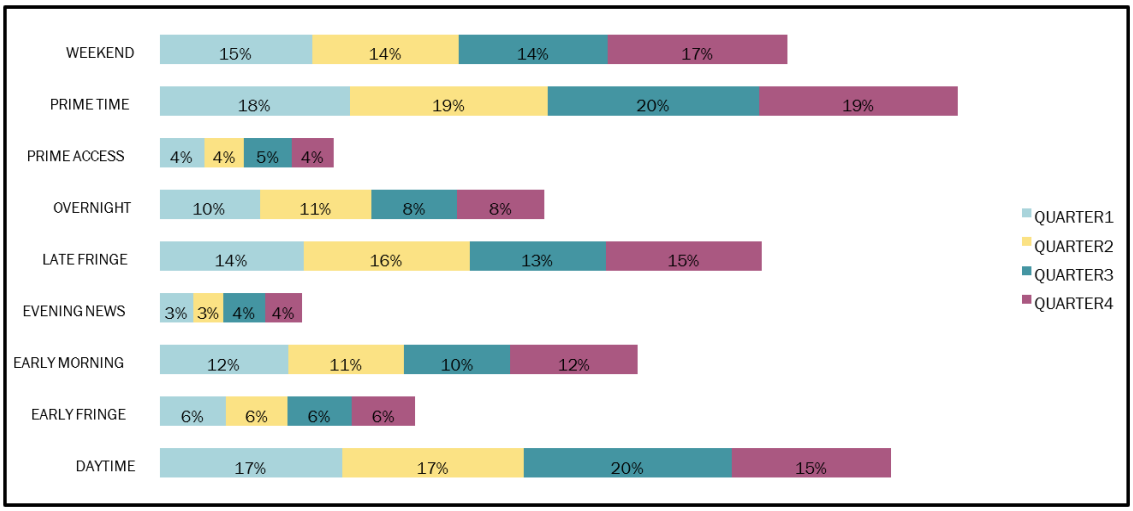
DISTRIBUTION OF ADS AS PER THEIR DURATION



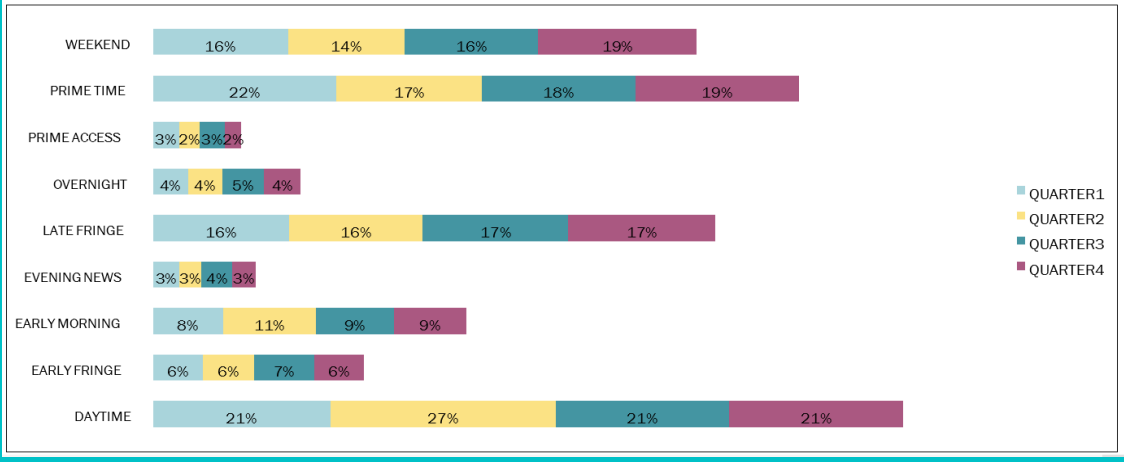
CONDUCT A COMPETITIVE ANALYSIS FOR THE BRANDS AND DEFINE ADVERTISEMENT STRATEGY OF DIFFERENT BRANDS AND HOW IT DIFFERS ACROSS THE BRANDS.



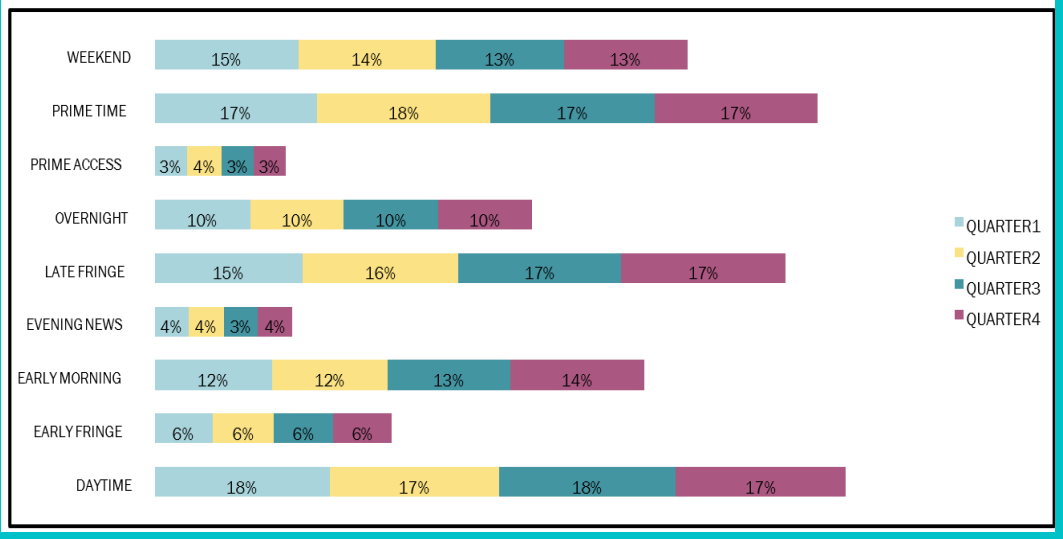
HYUNDAI MOTORS INDIA(DAY PARTS VS COUNT OF ADS VS QUARTERS)



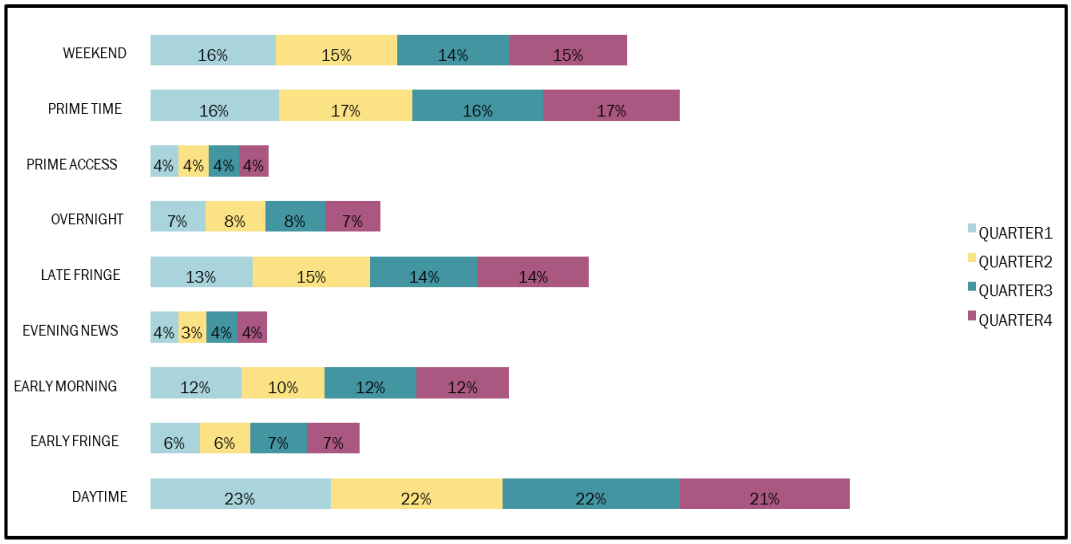
MAHINDRA AND MAHINDRA(DAY PARTS VS COUNT OF ADS VS QUARTERS)



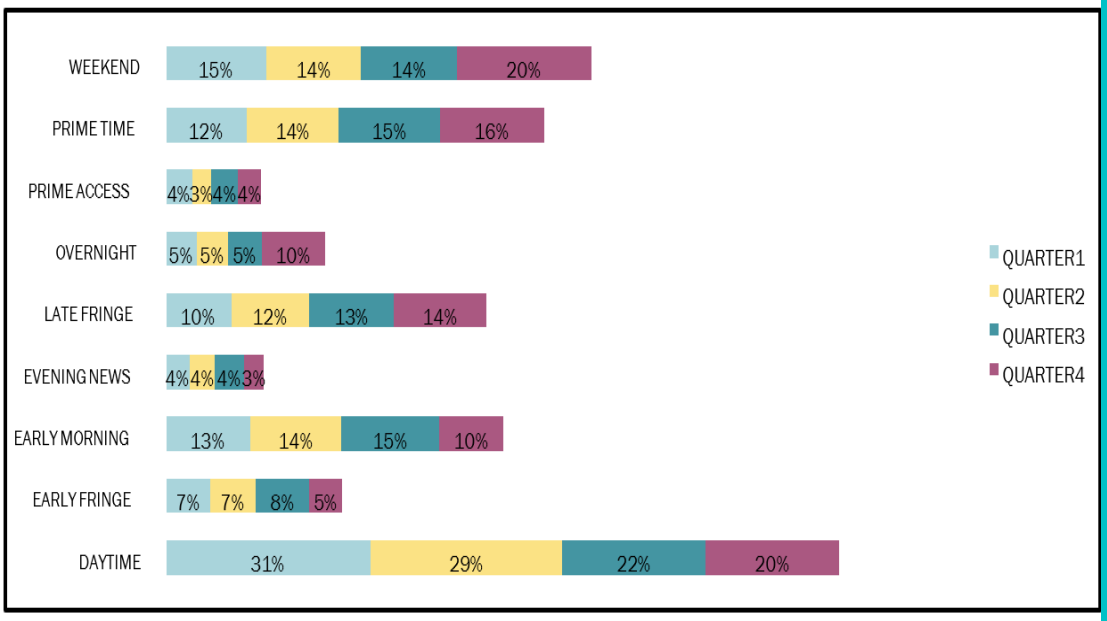
MARUTI SUZUKI (DAY PARTS VS COUNT OF ADS VS QUARTERS)



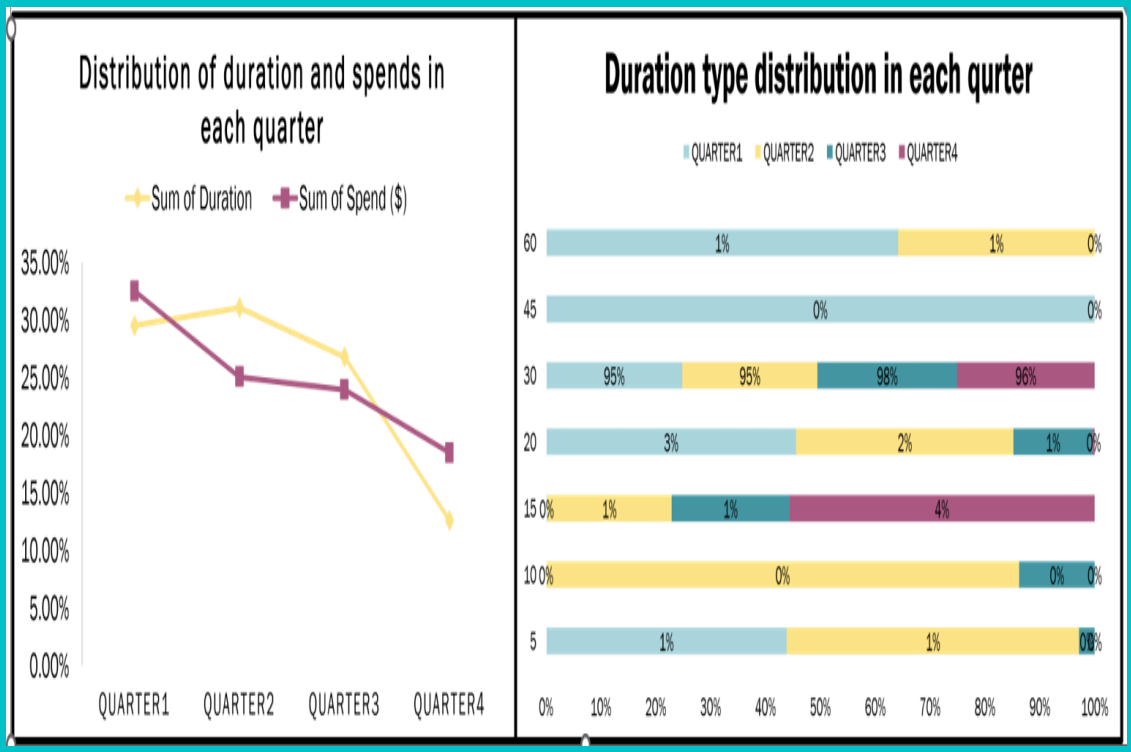
TATA MOTORS (DAYPARTS VS COUNT OF ADS VS QUARTERS)

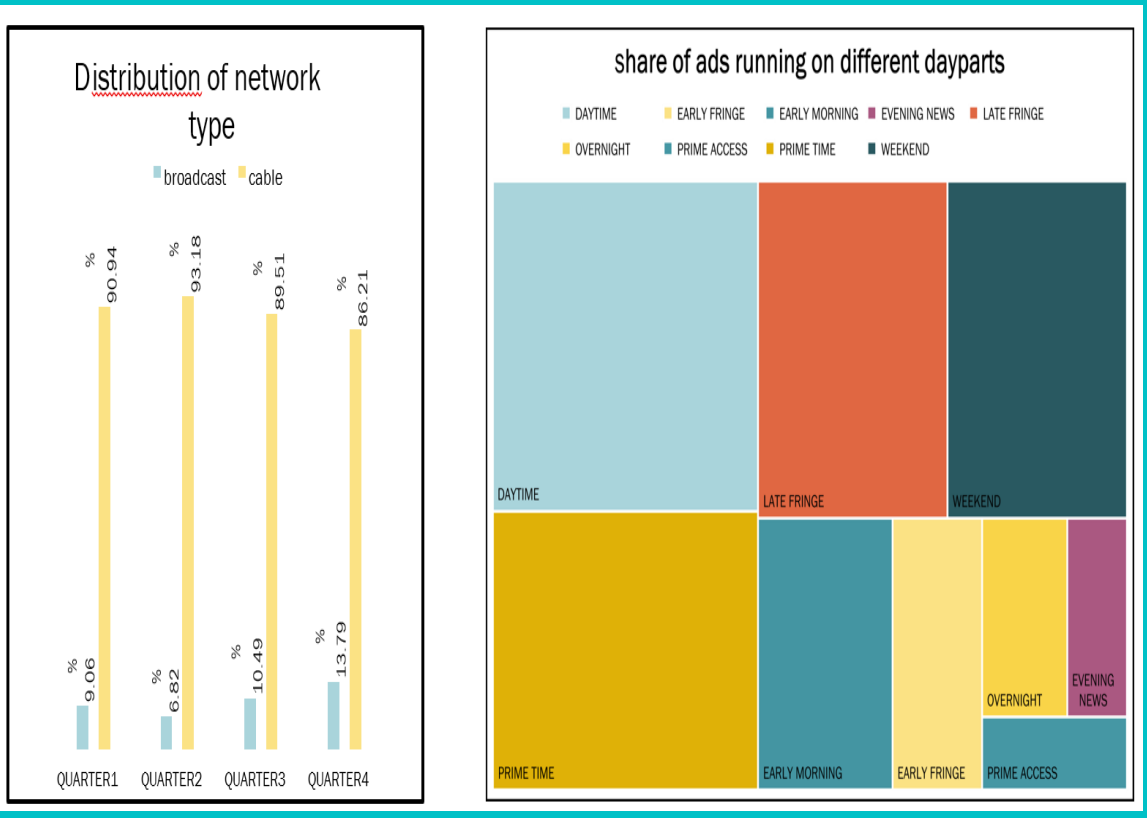


TOYOTA (DAYPARTS VS COUNT OF ADS VS QUARTERS)

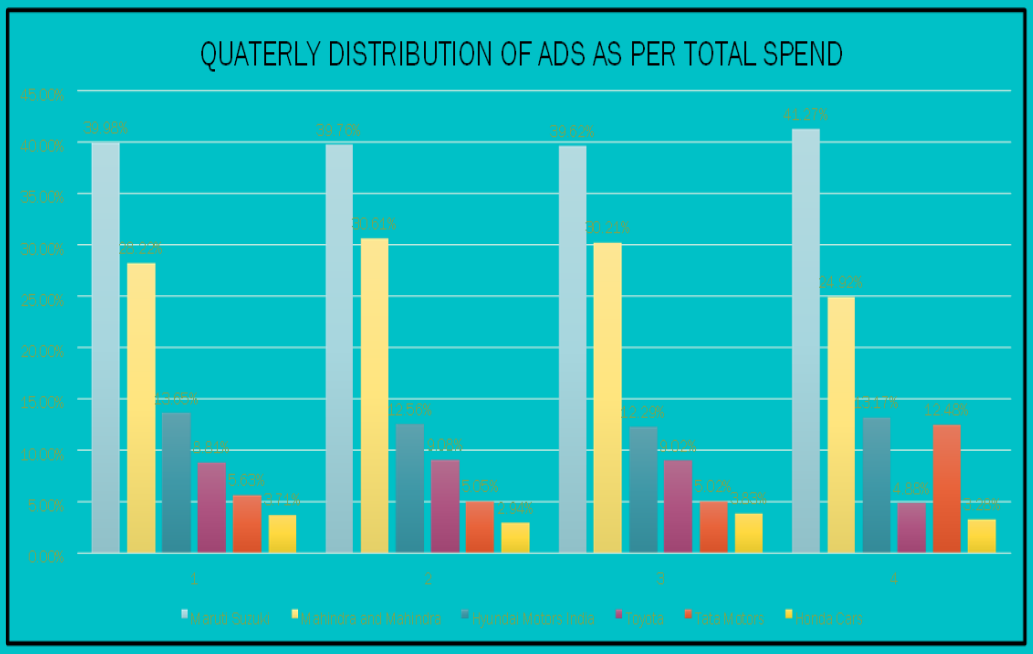


MAHINDRA AND MAHINDRA WANTS TO RUN A DIGITAL AD CAMPAIGN TO COMPLEMENT ITS EXISTING TV ADS IN Q1 OF 2022. BASED ON THE DATA FROM 2021, SUGGEST A MEDIA PLAN TO THE CMO OF MAHINDRA AND MAHINDRA. WHICH AUDIENCE SHOULD THEY TARGET? *ASSUME XYZ ADS HAS THE AD VIEWERSHIP DATA AND TV VIEWERSHIP FOR THE PEOPLE IN INDIA.




Distribution of network type
broadcast
cable
% 9.06
% 90.94
% 6.82
% 93.18
% 10.49
% 89.51
% 13.79
% 86.21
QUARTER1
QUARTER2
QUARTER3
QUARTER4
share of ads running on different dayparts
DAYTIME
EARLY FRINGE
EARLY MORNING
EVENING NEWS
LATE FRINGE
OVERNIGHT
PRIME ACCESS
PRIME TIME
WEEKEND
DAYTIME
LATE FRINGE
WEEKEND
PRIME TIME
EARLY MORNING
EARLY FRINGE
OVERNIGHT
EVENING NEWS
PRIME ACCESS

WHAT IS THE SHARE OF VARIOUS BRANDS IN TV AIRINGS AND HOW HAS IT CHANGED FROM Q1 TO Q4 IN 2021?



2022 Q1 MEDIA PLAN FOR

MARKETING STRATEGIES

- MAHINDRA AND MAHINDRA SHOULD TARGET AUDIENCE FROM NORTHEAST AND SOUTHERN PARTS OF INDIA AS A TARGET CUSTOMER IN Q1 2022 TO ENHANCE THEIR BUSINESS.

- SINCE THE MAIN PRODUCT OF MAHINDRA AND MAHINDRA IS PERSONAL AUTOMOBILES AND INSURANCE SO THEY NEED TO INCREASE THEIR MEDIA SPENT AND ADS DURATION ON SPORTS CHANNEL, NEWS CHANNEL AND CORPORATES ENTERTAINMENT NETWORK AS COMPARED TO FAMILY ENTERTAINMENT AND CHILD ENTERTAINMENT IN ORDER TO MAXIMIZE THEIR REACH ON TARGET AUDIENCE.

Distribution of amount spent by different brands on a monthly basis

Months	Honda Cars	Hyundai Motors India	Mahindra and Mahindra	Maruti Suzuki	Tata Motors	Toyota
January	10.35%	11.66%	11.34%	12.47%	9.79%	13.75%
February	13.85%	12.44%	9.30%	9.62%	10.90%	10.80%
March	10.95%	10.45%	11.89%	10.67%	6.52%	11.26%
April	5.04%	7.49%	8.03%	7.55%	5.30%	8.44%
May	6.85%	7.44%	8.15%	7.29%	6.55%	8.69%
June	7.89%	7.65%	8.86%	8.29%	5.46%	9.07%
July	10.15%	7.09%	8.25%	8.17%	6.56%	8.64%
August	8.06%	7.30%	7.91%	7.26%	2.58%	9.79%
September	6.81%	7.02%	7.80%	6.91%	7.56%	6.80%
October	6.20%	7.27%	6.71%	8.00%	11.56%	4.10%
November	5.84%	6.67%	5.45%	6.35%	13.27%	3.83%
December	8.00%	7.51%	6.31%	7.41%	13.95%	4.82%

ABC CALL VOLUME TREND ANALYSIS

PROJECT DESCRIPTION:

ABC IS A CALL CENTRE WHICH HAS A CUSTOMER EXPERIENCE TEAM FOR THE VOICE PROCESS. TYPICALLY, THESE TEAMS FULFIL VARIOUS ROLES AND RESPONSIBILITIES SUCH AS: CUSTOMER EXPERIENCE PROGRAMS (CX PROGRAMS), DIGITAL CUSTOMER EXPERIENCE,
DESIGN AND PROCESSES, INTERNAL COMMUNICATIONS, VOICE OF THE CUSTOMER (VOC),
USER EXPERIENCES, CUSTOMER EXPERIENCE MANAGEMENT, JOURNEY MAPPING, NURTURING CUSTOMER INTERACTIONS, CUSTOMER SUCCESS, CUSTOMER SUPPORT, HANDLING CUSTOMER DATA, LEARNING ABOUT THE CUSTOMER JOURNEY. I HAVE BEEN PROVIDED WITH THE DATA OF ABC CALL CENTRE FOR THE LAST 23 DAYS AND I SHOULD ANALYZE THE DATA AND HELP THE COMPANY ANSWER SOME OF THE BUSINESS QUESTIONS.

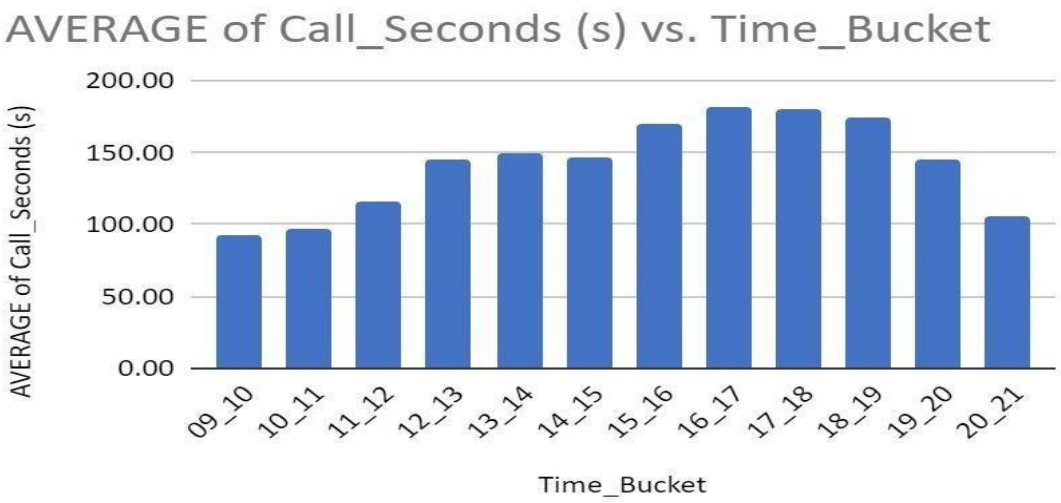
ABOUT THE DATASET :

- THERE ARE 117989 ROWS AND 13 COLUMNS.
- AGENT NAME IS A CATEGORICAL COLUMN CONTAINING NAMES OF THE AGENTS ATTENDING THE PHONE CALL.
- AGENT ID COLUMN CONTAINS THE AGENT ID AND THEY ALSO CONTAIN A LO OF NULL VALUES.
    - CUSTOMER COLUMN CONTAINS THE CONTACT NUMBER OF THE CUSTOMER.
    - QUEUE TIME CONTAINS THE SECONDS CUSTOMER HAS WAITED BEFORE THE AGENT HAS PICKED THE CALL.
    - **DATE_&_TIME, TIME & TIME_BUCKET** ARE USED FOR TIME INTELLIGENCE.
    - DURATION AND CALL_SECONDS CONTAINS THE DURATION OF THE CALL.
    - CALL STATUS AND WRAPPED ARE ALSO CATEGORICAL COLUMNS.

CLEANING THE DATASET :

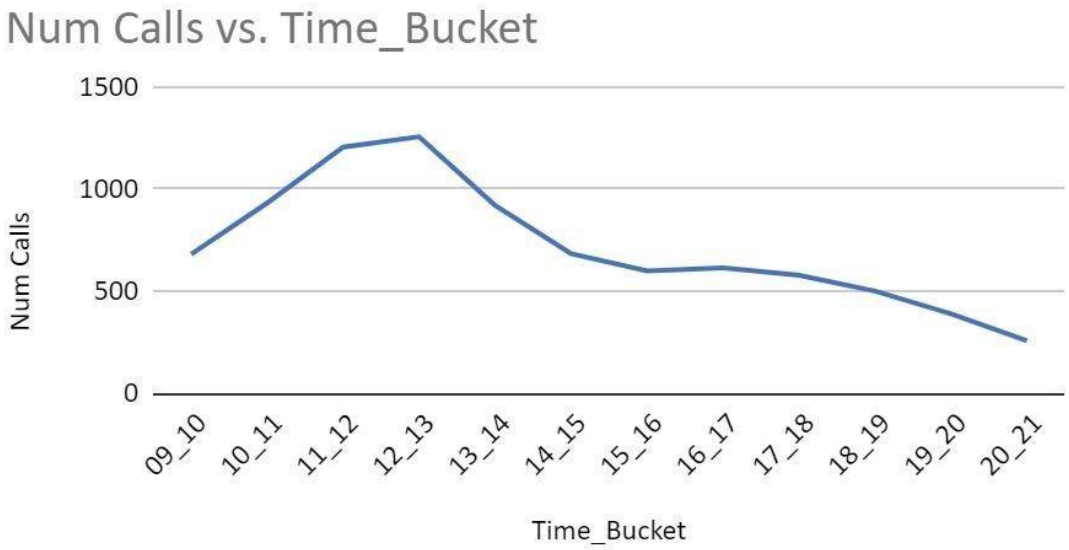
- THE AGENT NAME AND AGENT ID HAS NULL VALUES.
- THE COLUMNS RINGING AND IVR DURATION ARE NOT USED SINCE RINGING HAS ONLY ONE VARIABLE AND IVR DURATION IS NOT OF ANY USE TO OUR ANALYSIS.

AVERAGE CALL TIME DURATION FOR EACH TIME BUCKET



IT CAN BE SEEN THAT THE AVERAGE CALL DURATION INCREASES FROM 9 TO 5 AND THEN DECREASES.

NO. OF CALLS FOR EACH TIME BUCKET



IT CAN BE SEEN THAT THE NUMBER OF CALLS IN A DAY INCREASES FROM 9 TO 12 AND THEN IT STARTS DECREASING.

DECREASE THE DROP RATE

EXECUTIVE NAME	DROPPED	ANSWERED	TRANSFERRED
#N/A	34198		
GRAND TOTAL	**34403**	**82452**	**1133**

IT CAN BE SEEN THAT OUT OF 34403 CALLS DROPPED, 34198 CALLS HAVE NOT BEEN ANSWERED BY ANYONE. THIS CAN ALSO MEAN THAT THE AGENTS WERE BUSY IN ANY OTHER CALLS AND SO, THEY WERE UNABLE TO TEND TO THESE CALLS. THIS MEANS THAT WE NEED TO INCREASE THE MAN POWER TO REDUCE THE DROP RATE.

TIME_BUCKET	NUMBER OF AGENTS	ABANDON RATE	CALLS EACH DAY	AVG CALL DURATION	AVG QUEUE TIME
09_10	42	53.70%	416.87	92.01	82.86
10_11	51	51.91%	578.83	97.42	83.25
11_12	59	41.21%	635.91	116.78	72.32
12_13	60	24.29%	550.09	144.73	41.66
13_14	58	22.64%	502.65	149.54	41.80
14_15	60	23.44%	459.17	146.97	43.60
15_16	58	13.25%	398.22	169.90	29.88
16_17	58	8.50%	382.09	181.44	23.54
17_18	58	9.18%	371.04	179.72	23.75
18_19	59	12.89%	314.70	174.32	34.09
19_20	52	28.59%	281.00	144.58	58.69
20_21	27	47.68%	239.35	105.95	75.28
GRAND TOTAL	**66**	29.16%	5,129.91	139.53	**52.17**

It can be inferred that the abandon rates are high on particular times of a day and lowest on a particular time in a day. Even though the calls are low during the 9_10 bucket, the drop rate is very high. We have the average call duration in each time bucket. Using that, we can calculate approximately how much time the call centre agents are spending talking to the customers in total in a particular time bucket. By using that data, we can arrive at the approximate number of agents required to have a drop rate lesser than 10 %. It can also be seen that the time buckets in which the avg call duration is high and avg queue time is low have lesser drop rates.
Note: A tolerance level needs to be added in every calculation to account for the errors.

TIME_BUCKET	9 SHIFT	10 SHIFT	12 SHIFT	AGENTS REQ	AGENTS REQ CALCULATED
09_10	50	0	0	50	44.88
10_11	50	20	0	70	64.04
11_12	50	20	0	70	73.96
12_13	25	10	40	75	62.65
13_14	0	20	40	60	58.47
14_15	50	0	20	70	53.57
15_16	25	20	40	85	48.84
16_17	50	20	0	70	47.60
17_18	50	10	20	80	46.01
18_19	0	20	30	50	39.48
19_20	0	0	40	40	34.78
20_21	0	0	30	30	26.58

Agents required is calculated by using the formula
Agents req calculated = 1.1 * calls each day * 2 * (avg call duration + avg queue time) Where 1.1 and 2 are tolerance of no of calls and waiting time respectively.

So the total man power required = 9 AM shift + 10 AM shift + 12 PM shift = 50 + 20 + 40 = 110
New manpower to be added = req man power - available employees = 110 - 66 = 44

IT IS ADVISED TO ADD 44 NEW EMPLOYEES TO REDUCE THE ABANDON RATE FROM 30 % TO 10 %

4. Organize Night Calls

Men required can be calculated from the same formula given in the previous slide. For the night calls too, men required can be found from the same formula. Calls/ day can be found from the data for the whole 24 hours. A shift plan has been drafted in which the required number of agents in that specific period of time has been used to carefully decided based on the expected number of calls along with some tolerance.

8 AM SHIFT

Time Bucket	Men Required	Total Men	8 AM Shift	5 PM Shift	2 AM Shift
08_09	42.93	80	60	0	20
09_10	44.88	90	70	0	20
10_11	64.04	70	70	0	0
11_12	73.96	70	70	0	0
12_13	62.65	60	60	0	0
13_14	58.47	60	60	0	0
14_15	53.57	60	60	0	0
15_16	48.84	50	50	0	0
16_17	47.60	50	50	0	0

5 PM SHIFT

Time Bucket	Men Required	Total Men	8 AM Shift	5 PM Shift	2 AM Shift
17_18	46.01	50	0	50	0
18_19	39.48	50	0	50	0

19_20	34.78	50	0	50	0
20_21	26.58	50	0	50	0
21_22	25.76	25	0	25	0
22_23	25.76	25	0	25	0
22-00	17.17	50	0	50	0
00_01	17.17	50	0	50	0
01_02	8.59	70	0	50	20

2 AM SHIFT

TIME BUCKET	MEN REQUIRED	TOTAL MEN	8 AM SHIFT	5 PM SHIFT	2 AM SHIFT
02_03	8.59	70	0	50	20
03_04	8.59	20	0	0	20
04_05	8.59	10	0	0	10
05_06	8.59	10	0	0	10
06_07	25.76	20	0	0	20
07_08	34.34	20	0	0	20
08_09	34.34	20	0	0	20
02_03	42.93	80	60	0	20
09_10	44.88	90	70	0	20

SO, THE TOTAL NUMBER OF MAN POWER NEEDED IS 8 AM SHIFT + 5 PM SHIFT + 2 AM SHIFT = 70 + 50 + 20 = 140.
AFTER HIRING 140 - 66 = 74 AGENTS, WE CAN EXPECT THE CALL ABANDON RATES TO LESSER THAN 10 %

Conclusion

In conclusion, I would like to tell that after doing a thorough analysis we were able to derive the insights from the data and was able to plot various graphs using that data.

The data which once looked useless gave some very useful insights.