Data Analysis Portfolio

Prepared By: Prasiddhi S Tomar

Professional Background

I am highly motivated and skilled individual with a strong educational background in Mathematics(M.Sc). With a passion for data analysis and a firm grasp of mathematical principles, I possesses the necessary expertise to excel in the field of data analysis. Also, possesses experience as a data analyst trainee with Trainity, coupled with hands-on projects involving SQL and Excel, has provided me with valuable practical knowledge in the field. I am committed to leveraging the skills and experience to build a successful career as a data analyst.

I served as a data analyst trainee at Trainity, a leading data analysis firm. During this role, I actively participated in various data analysis projects, contributing to the development of data-driven insights. The responsibilities included data cleaning and preprocessing, constructing SQL queries to extract and manipulate data, performing statistical analysis, and creating informative visualizations to present findings.

I have a clear vision of building a long-term career as a data analyst. The passion for mathematics, coupled with the analytical skills and practical experience, drives my desire to continue exploring and mastering advanced data analysis techniques.

Table of contents

01 Professional Background

02 Table of Content

03 Data Analytics Process

04 Instagram User Analytics

05 Operation Analytics and Investigating Metric Spike

06 Hiring Process Analytics

07 IMDB Movie Analysis

08 Bank Loan Case Study

09 Impact of Car Features

10 ABC Call Volume Trend Analysis Project

11 Appendix

THE TASK:

Your task is to give the example(s) of such a real-life situation where we use Data Analytics and link it with the data analytics process. You can prepare a PPT/PDF on a real-life scenario explaining it with the above process (Plan, Prepare, Process, Analyze, Share, Act) and submit it as part of this task.

DESCRIPTION:

- Preparation for any government exam is a very self-directed process. Data analytics can be utilized to prepare for any competitive exam (plan, prepare, process, analyze, share, act).

- Exam preparation can be stressful, so using data analytics strategies can help you stay organized.

PLAN:

•Before beginning the preparation for any government exam, ensure that you first understand the criteria for that exam and that you are a right fit for the exam.

PREPARE:

- When you begin to prepare for your exam, list down the subjects and take priority what and how to complete the syllabus on time.

- You must also comprehend which subject requires more attention and how to study the weak subject.

PROCESS:

- Once the student has planned and begun preparing, they must see the data that is available to them or are there any other requirements for the data/study material to undergo for the exam.

- If book/notes does not consist of a few more topics which are in the syllabus, they must go and get another book, institute notes available, etc.

ANALYZE:

- The student will not purchase notes/books that do not contain the requisite qualifications to cover the syllabus.
- Now, the student must determine whether the notes they made for themselves are adequate.
- If not, it is necessary to specify the type of deletion or addition so that it can be revoked at the time of revision.

SHARE:

- After completing all of the preceding steps, the student can communicate or share their ideas, thoughts, and discuss them with their respective teachers/guidance to gain a better understanding of his/her approach and growth.

- The teacher will also provide feedback that can be used to improve the results.

ACT:

- The student can now administer test series and incorporate the results into the final exam. The student is now fully prepared to take the exam and act more confidently.

INSTAGRAM USER ANALYTICS

(SQL)

PROJECT 2

TASK:

User analysis is the process by which we track how users engage and interact with our digital product (software or mobile application) in an attempt to derive business insights for marketing, product & development teams.
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.
You are working with the product team of Instagram and the product manager has asked you to provide insights on the questions asked by the management team.

:

PROJECT DESCRIPTION

The idea is based on user analytics for Instagram. The research primarily focuses on MYSQL database queries to gain an understanding of user analysis via which we can monitor how people connect with and interact with our digital offering. When it comes to project management, it uses MySQL to query data and has tables with a variety of columns, including hastags, likes, photos, etc. The most important things I discovered concerned learning how a user behaves on the platform in order that these insights may support or update several other concepts in the project's database..

APPROACH,TECH-STACK AND INSIGHTS

1. APPROACH

Learning something new requires persistence and time. I began by studying the platform learning sections to master the fundamentals of MySQL, which gave me the foundation. With the aid of Google.

2. TECH STACK

With the aid of Google. I downloaded the most recent version of MySQL and began working on the project.

3. INSIGHTS

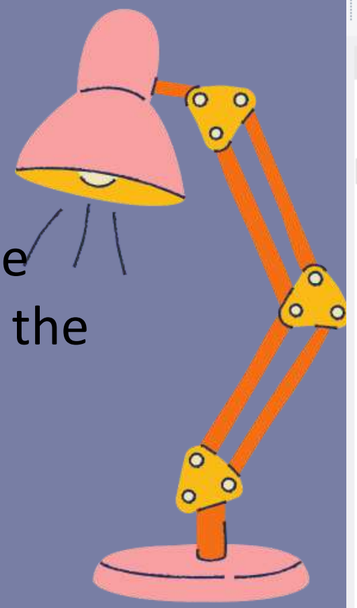
Working with MySQL is enjoyable because it's simple to comprehend and manage when constructing a correct query. Running a query and still receiving errors can be difficult at times, but if you know how to handle them, you can do so without difficulty. I learned a lot from all of these things, and I'll use them in future endeavors.

RESULTS:

The project taught me various aspects of MYSQL and how to solve problems when they arose.

The required detailed report answering the questions is provided below.

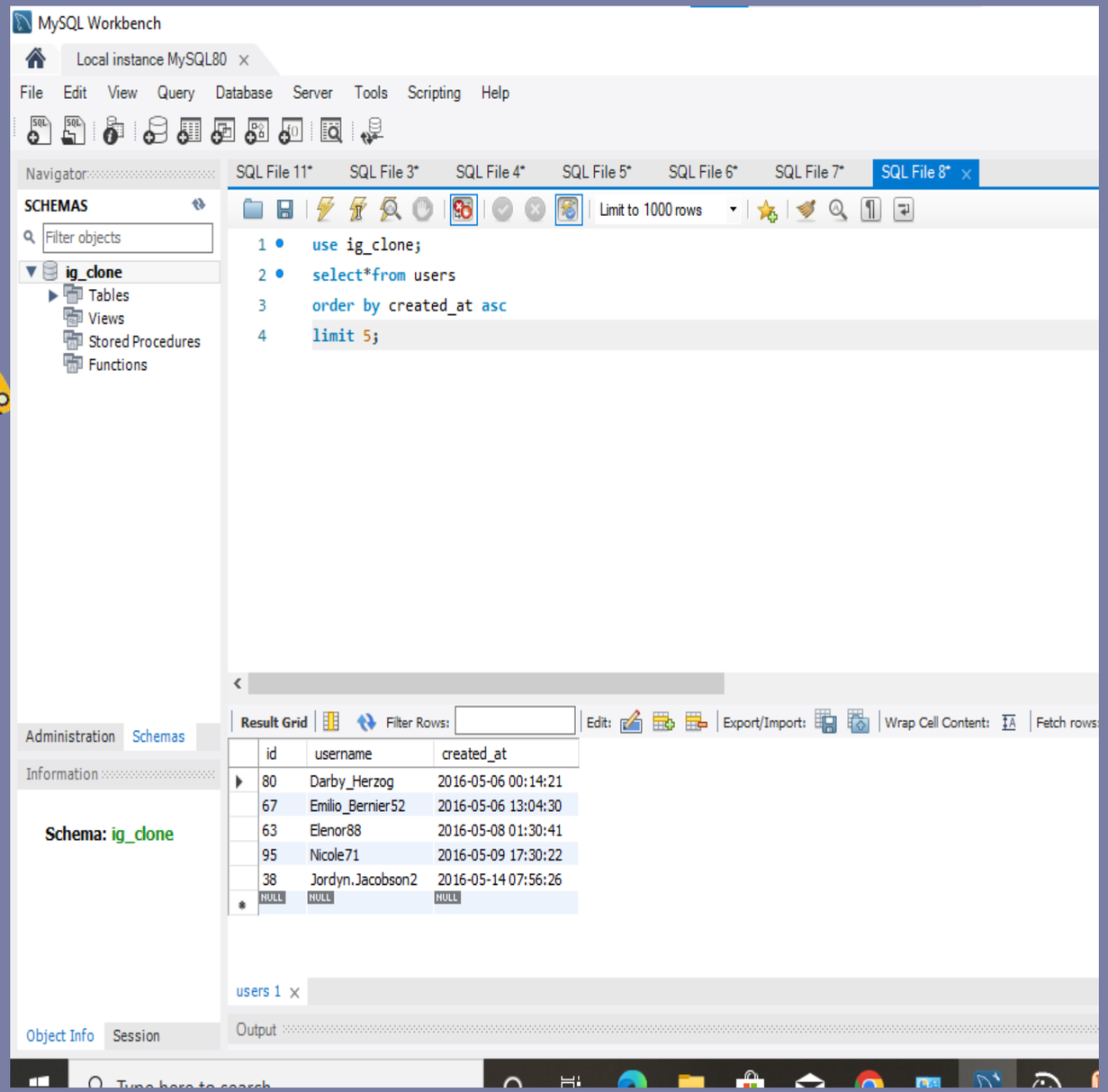
(A)MARKETING



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

Point 1

Rewarding Most Loyal Users: People who have been using the platform for the longest time.
Your Task: Find the 5 oldest users of the Instagram from the database provided

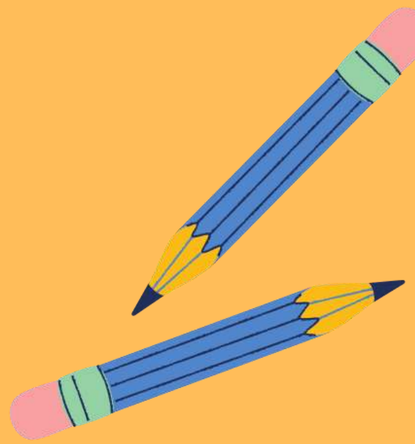


Point 2

Remind Inactive Users to Start Posting:

By sending them promotional emails to post their 1st photo.

<u>Your Task</u>: Find the users who have never posted a single photo on Instagram



MySQL Workbench

Local instance MySQL80

File Edit View Query Database Server Tools Scripting Help

SQL File 11* SQL File 3* SQL File 4* SQL File 5* SQL File 6* SQL File 7* SQL File 8* SQL File 9*

Navigator

SCHEMAS

Filter objects

ig_clone
- Tables
- Views
- Stored Procedures
- Functions

Limit to 1000 rows

```
1  select username
2  from users
3  where id not in (select user_id from photos);
```

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

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

Administration Schemas

Information

Schema: ig_clone

users 1

Object Info Session

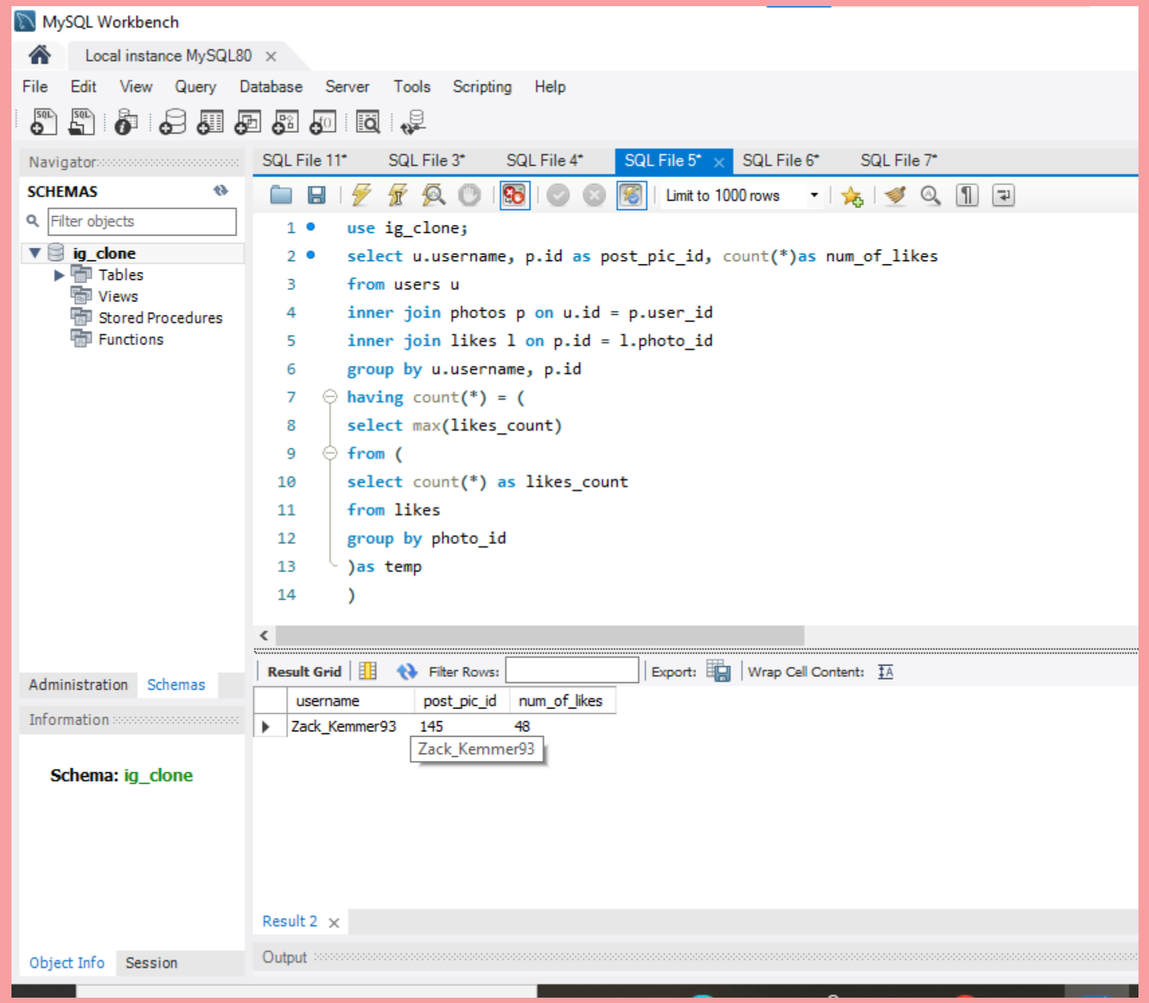
Output

Point 3

Declaring Contest Winner:

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

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

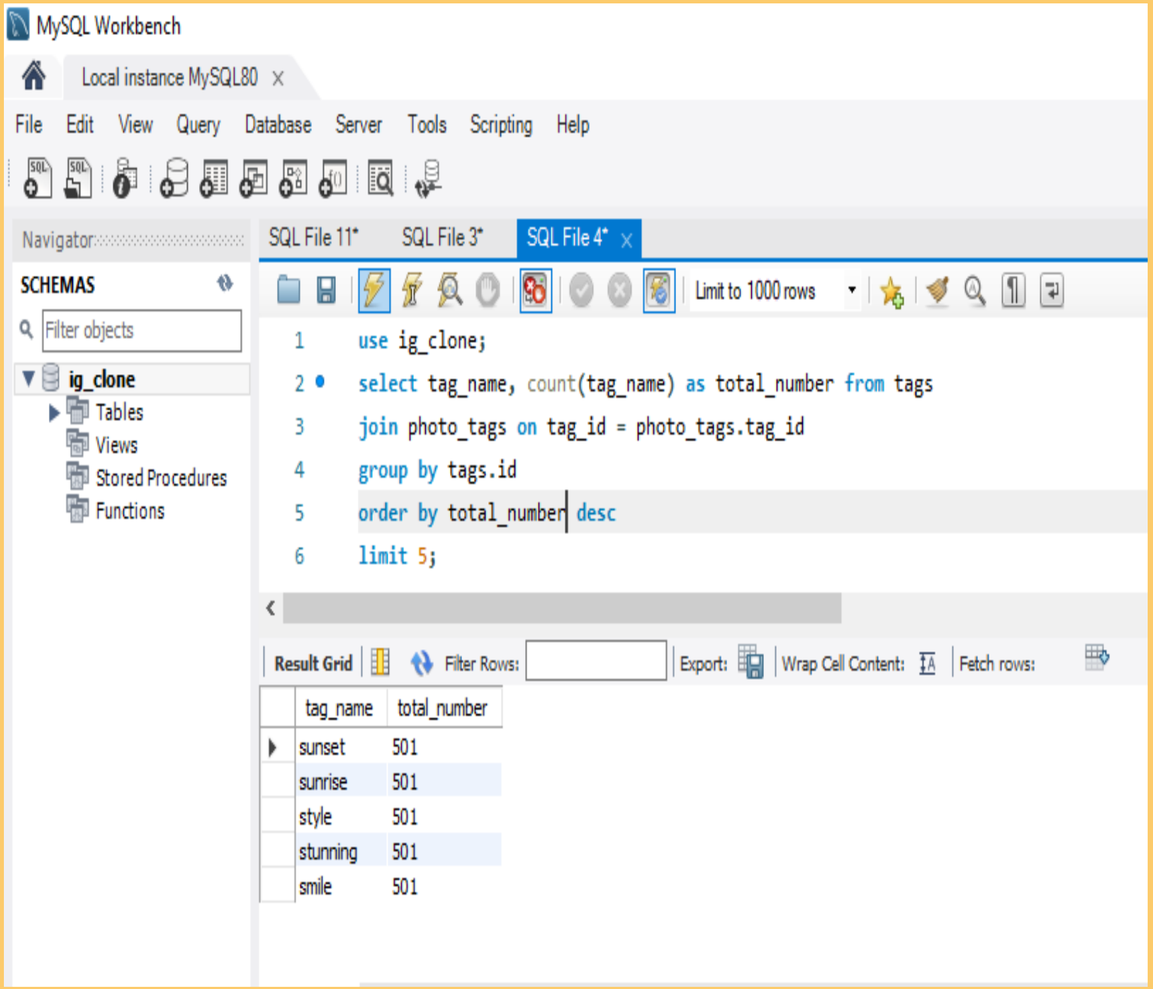


POINT 4:

Hashtag Researching:

A partner brand wants to know, which hashtags to use in the post to reach the most people on the platform.

<u>Your Task:</u> Identify and suggest the top 5 most commonly used hashtags on the platform

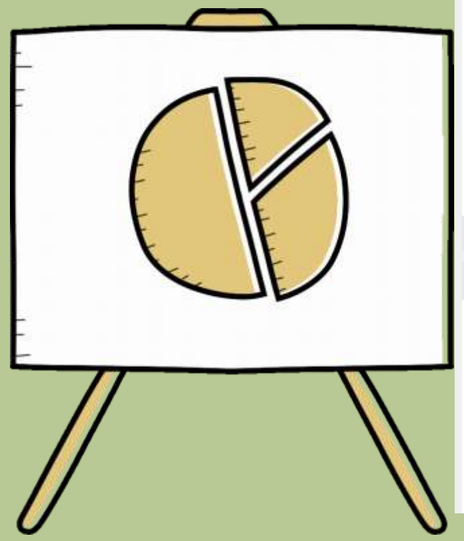


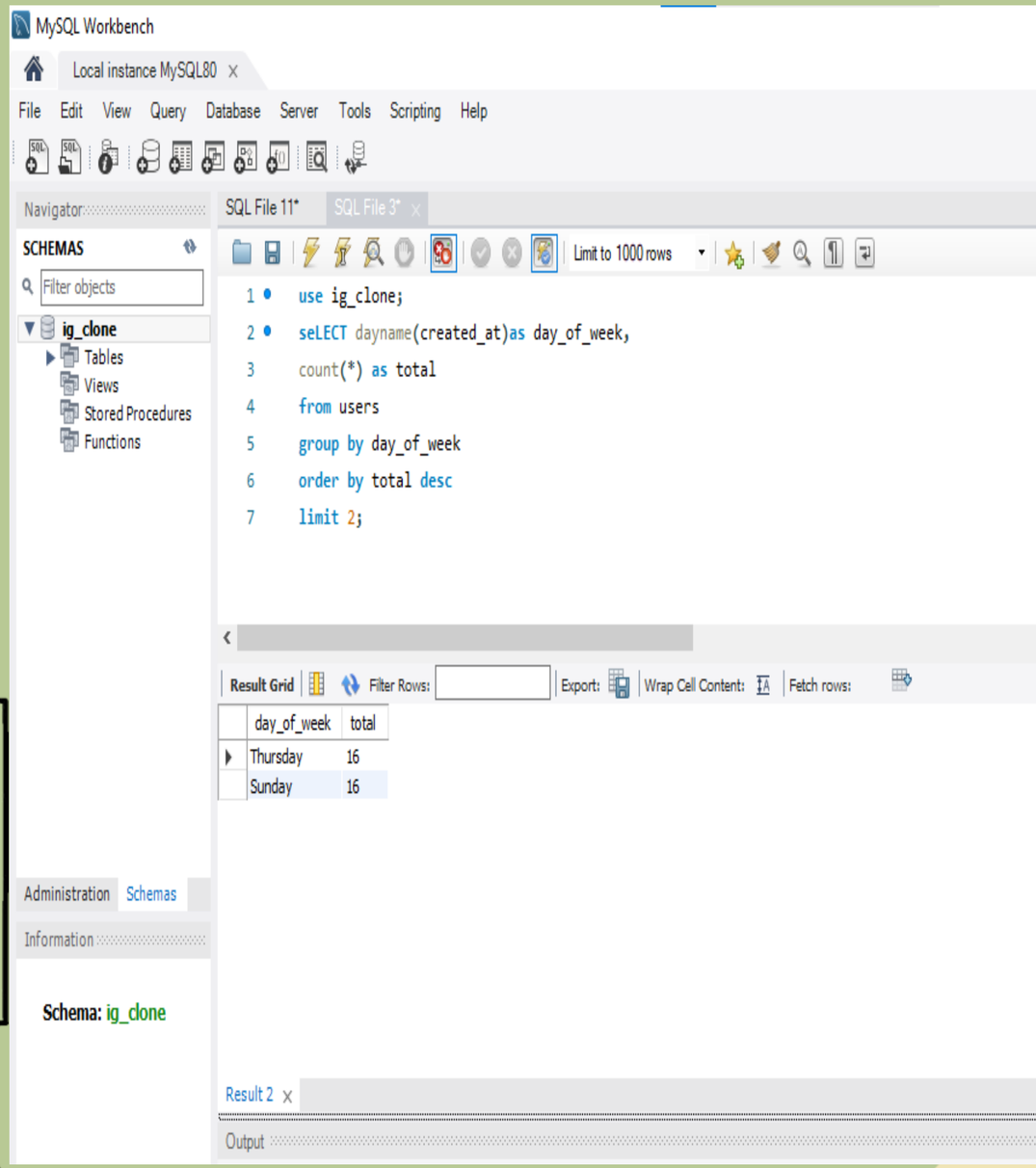
POINT 5:

Launch AD Campaign:

The team wants to know, which day would be the best day to launch ADs.

<u>Your Task</u>: What day of the week do most users register on? Provide insights on when to schedule an ad campaign





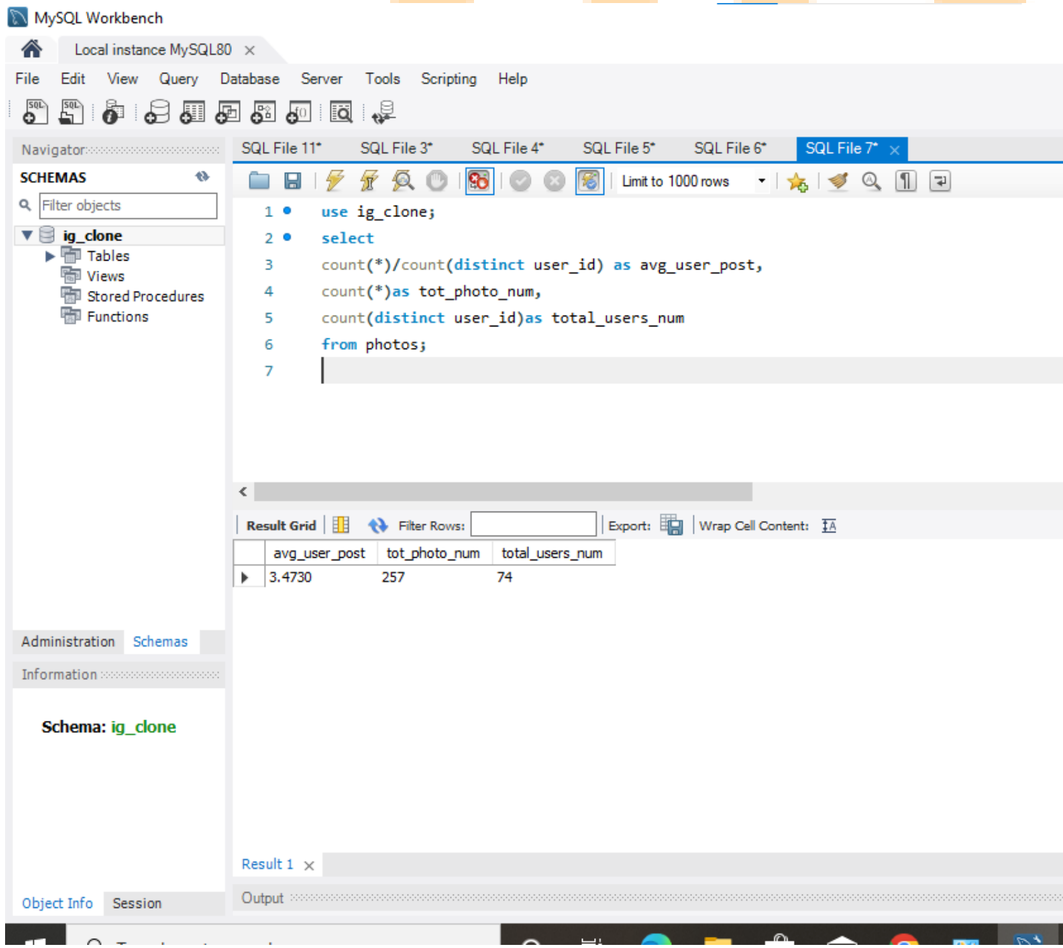
B) Investor Metrics

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

<u>POINT 1:</u>

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

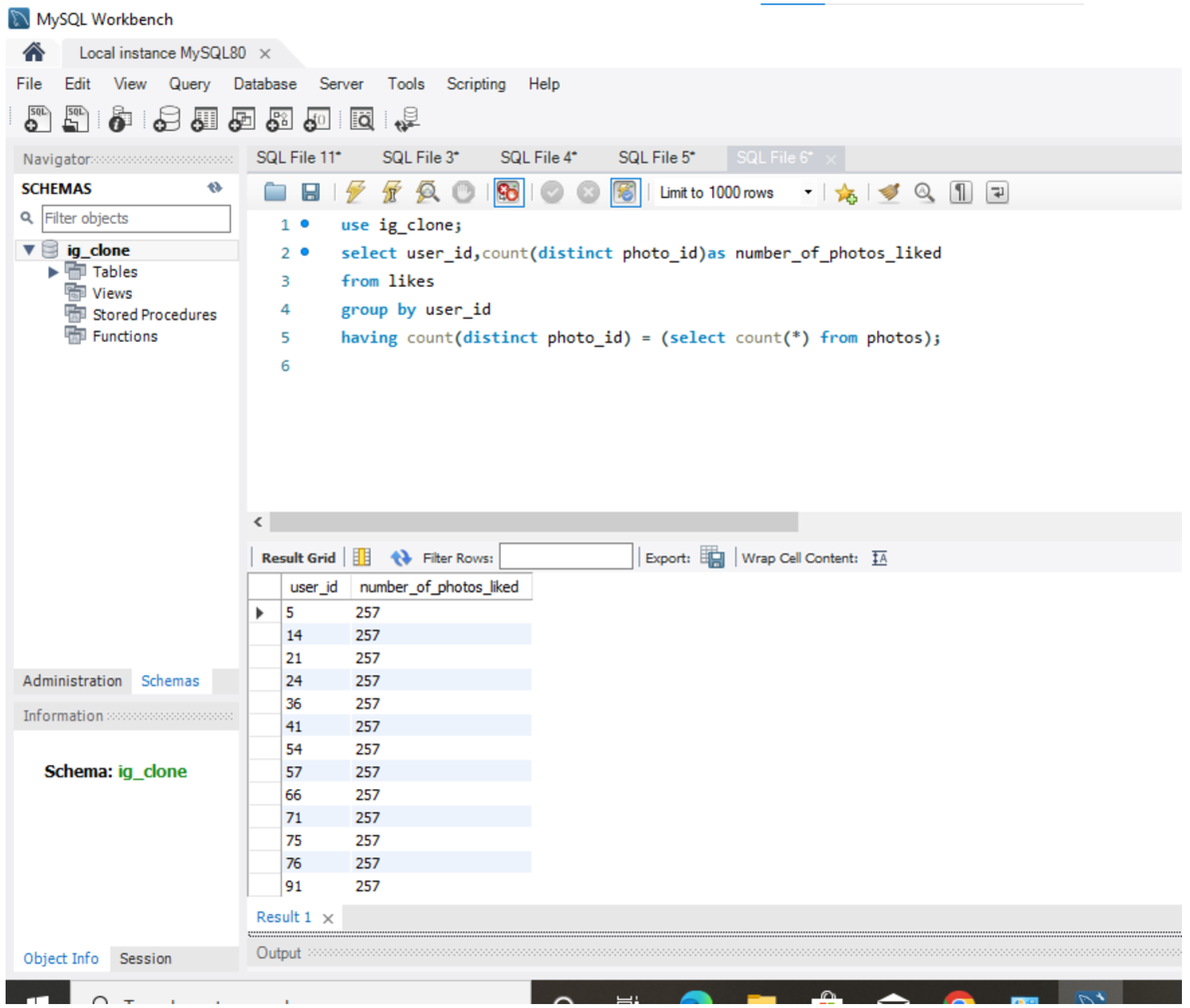
<u>Your Task:</u> Provide how many times does average user posts on Instagram. Also, provide the total number of photos on Instagram/total number of users



POINT 2

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

<u>Your Task:</u> Provide data on users (bots) who have liked every single photo on the site (since any normal user would not be able to do this).



Operation Analytics and Investigating Metric Spike

Advanced SQL

CONTENT

1. INTRODUCTION
2. CASE STUDY 1 (JOB DATA)
3. CASE STUDY 2(INVESTIGATIONG METRIC SPIKE)
4. SUMMARY

PROJECT 3

PROJECT DESCRIPTION:

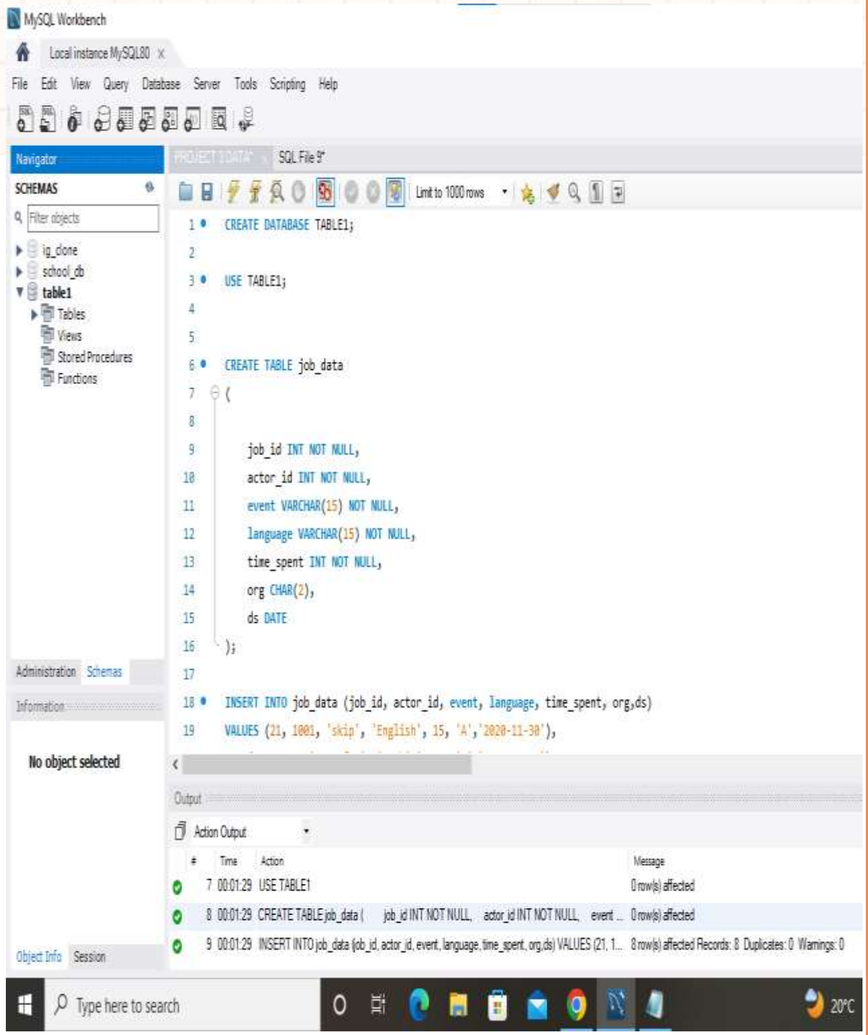
Operation Analytics is the analysis done for the complete end to end operations of a company. With the help of this, the company then finds the areas on which it must improve upon. You work closely with the ops team, support team, marketing team, etc and help them derive insights out of the data they 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 means 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.

You are working for a company like Microsoft designated as Data Analyst Lead and is provided with different data sets, tables from which you must derive certain insights out of it and answer the questions asked by different departments.

Create the Database and Tables: Created a database and then the tables using the structure and links provided.



Case Study 1 (Job Data)

Case Study 1 (Job Data)
Below is the structure of the table with the definition of each column that you must work on:
Table-1: job_data

job_id: unique identifier of jobs
actor_id: unique identifier of actor

event: decision/skip/transfer

language: language of the content
time_spent: time spent to review the job in second

org: organization of the actor
ds: date in the yyyy/mm/dd format. It is stored in the form of text and we use presto to run. no need for date function

CASE STUDY 1:

(A). Number of jobs reviewed: Amount of jobs reviewed over time.

The task: Calculate the number of jobs reviewed per hour per day for November 2020?

base Server Tools Scripting Help

PROJECT 3 DATA* | SQL File 9* | SQL File 9* | SQL File 10* | SQL File 6

Limit to 1000 rows

```
1  use table1;
2  SELECT ds, COUNT(distinct job_id)/(30*24) AS number_of_the_jobs
3    FROM job_data
4     WHERE ds >='2020-11-01' AND Ds>='2020-11-30'
5  GROUP BY ds;
```

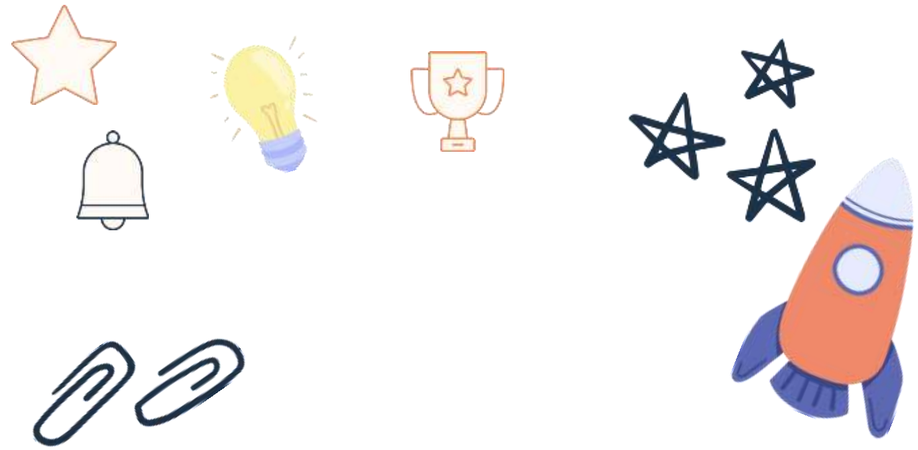
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

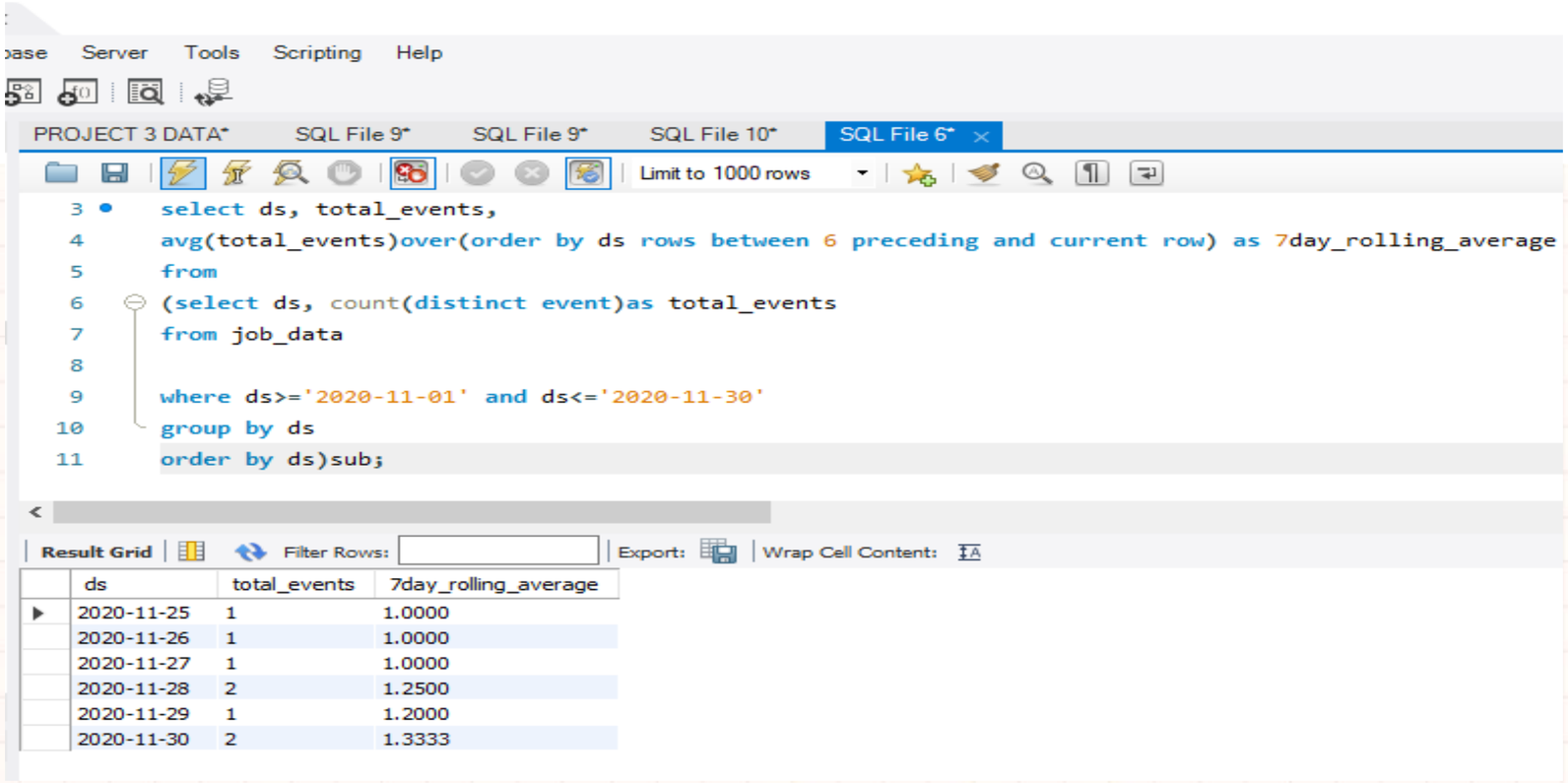
ds	number_of_the_jobs
2020-11-30	0.0028

CASE STUDY 1:

(B)Throughput: It is the no. of events happening per second.

The 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?



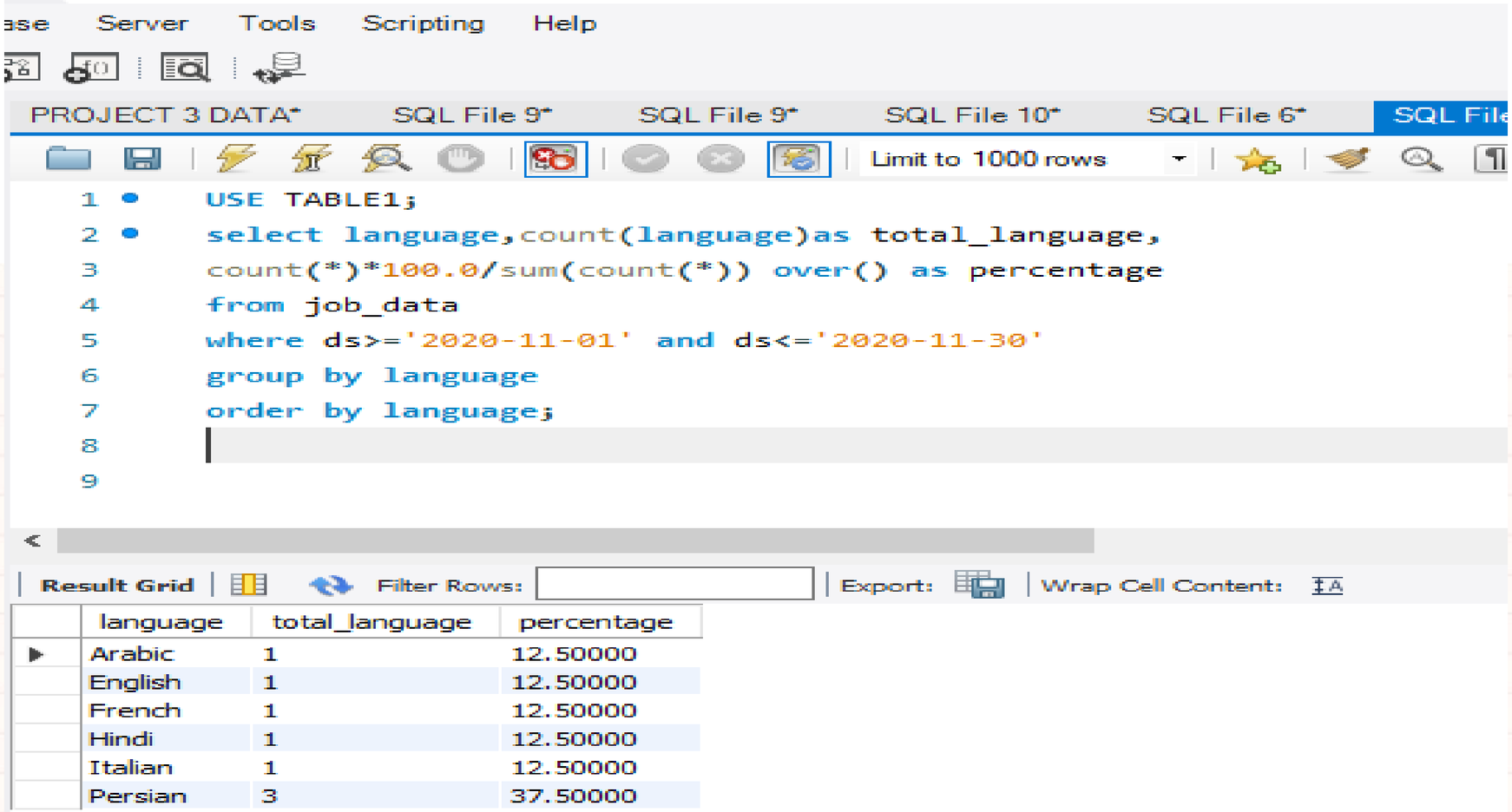


ds	total_events	7day_rolling_average
2020-11-25	1	1.0000
2020-11-26	1	1.0000
2020-11-27	1	1.0000
2020-11-28	2	1.2500
2020-11-29	1	1.2000
2020-11-30	2	1.3333

CASE STUDY 1:

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

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

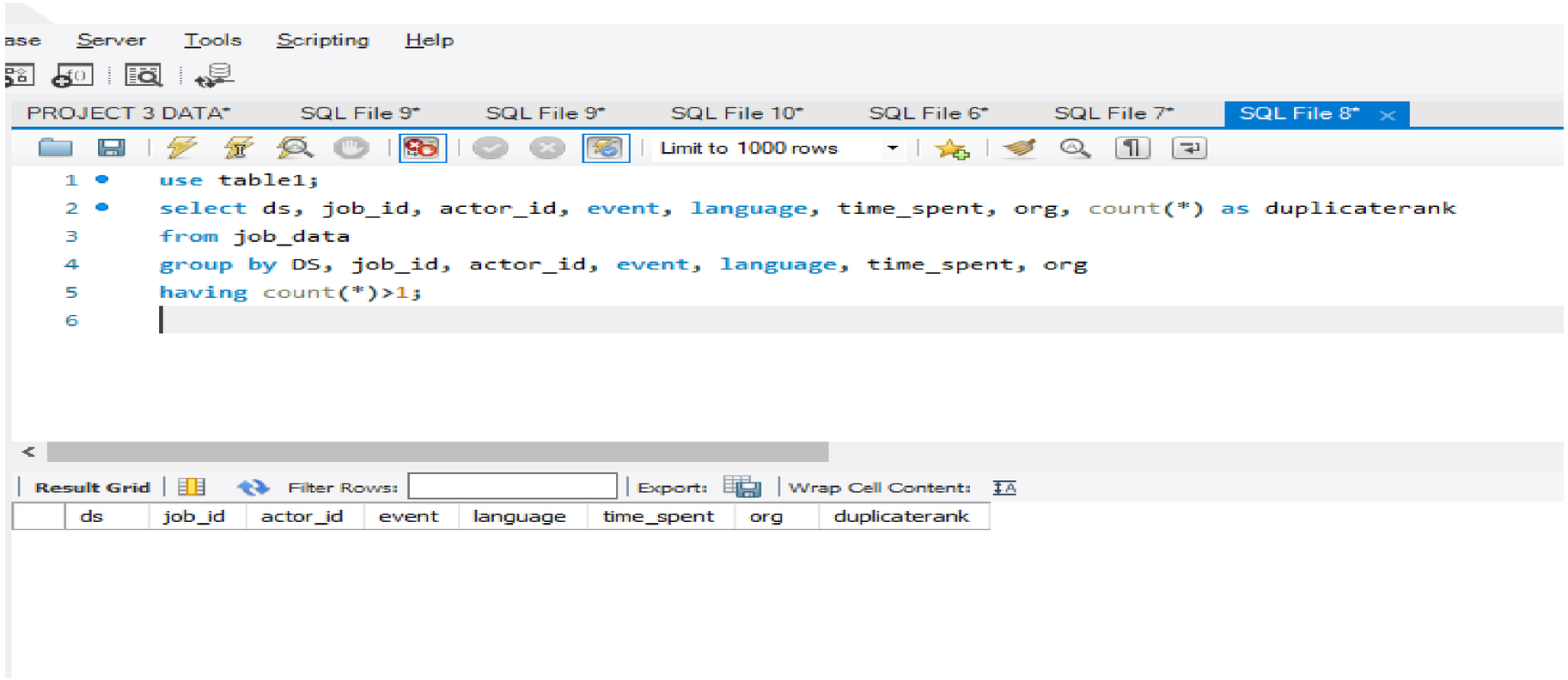


language	total_language	percentage
Arabic	1	12.50000
English	1	12.50000
French	1	12.50000
Hindi	1	12.50000
Italian	1	12.50000
Persian	3	37.50000

CASE STUDY 1:

(D).Duplicate rows: Rows that have the same value present in them.

The task: Let's say you see some duplicate rows in the data. How will you display duplicates from the table?



CASE STUDY 2:

The structure of the table with the definition of each column that you must work on is present in the project image

Table-1: users
This table includes one row per user, with descriptive information about that user's account.
Table-2: events
This table includes one row per event, where an event is an action that a user has taken. These events include login events, messaging events, search events, events logged as users progress through a signup funnel, events around received emails.
Table-3: email_events
This table contains events specific to the sending of emails. It is similar in structure to the events table above.

CASE STUDY 2:

(A)User Engagement: To measure the activeness of a user. Measuring if the user finds quality in a product/service.

Your task: Calculate the weekly user engagement?

```
1
2  SELECT EXTRACT( Week from occurred_at) as week_num, count(DISTINCT user_id) as num_users
3  FROM Trainity.Events
4  WHERE event_type = 'engagement'
5  GROUP BY week_num
6  order by week_num;
```

Query results

JOB INFORMATION | RESULTS | JS

Row	week_num	num_users
1	17	663
2	18	1068
3	19	1113
4	20	1154
5	21	1121
6	22	1186
7	23	1232
8	24	1275
9	25	1264
10	26	1302
11	27	1372
12	28	1365
13	29	1376
14	30	1467
15	31	1299

**THE OUTPUT IS NOT THE FULL OUTPUT AS THIS TABLE HAS 19 ROWS
INSIGHTS: The highest level of user engagement occurred during the 30th week, while the lowest level occurred during the 35th week.

CASE STUDY 2:
(B).User Growth: Amount of users growing over time for a product.

Your task: Calculate the user growth for product?

```sql
SELECT year, week_num, num_users, sum(num_users)
over(order by year,week_num) as cum_users
from (
SELECT EXTRACT(YEAR FROM created_at) as year, EXTRACT(Week from created_at) as week_num, count(DISTINCT user_id) as num_users
FROM Trainity.Users
WHERE state = 'active'
GROUP BY year, week_num
order by year, week_num)sub;
```

```sql
order by year, week_num)sub;
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS

Row	year	week_num	num_users	cum_users
1	2013	0	23	23
2	2013	1	30	53
3	2013	2	48	101
4	2013	3	36	137
5	2013	4	30	167
6	2013	5	48	215
7	2013	6	38	253
8	2013	7	42	295
9	2013	8	34	329
10	2013	9	43	372
11	2013	10	32	404
12	2013	11	31	435

**The highest number of users actively interacting with the product or service occurred during the 33rd week of 2014 the 35th week of 2014 had the fewest active users on the service

CASE STUDY 2:

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

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

```sql
with cte1 as (
  select distinct user_id, Extract(week from occurred_at) as signup_week
  from Trainity.Events
  where event_type ='signup_flow' and event_name ='complete_signup'
  and extract(week from occurred_at) = 17
)
, cte2 as (select distinct user_id, extract(week from occurred_at) as engagement_week
from Trainity.Events
where event_type ='engagement')
select count(user_id) as total_engaged_users,
sum(case when retention_week>0 then 1 else 0 end) as retained_users
from
(select a.user_id, a.signup_week,
b.engagement_week, b.engagement_week-a.signup_week as retention_week
from cte1 a left join cte2 b
on a.user_id = b.user_id
order by a.user_id)sub
```

Query results

JOB INFORMATION | RESULTS | JSON

Row	total_engaged_u	retained_users
1	278	206

CASE STUDEY 2:

(D).Weekly Engagement: To measure the activeness of a user. Measuring if the user finds quality in a product/service weekly.

Your task: Calculate the weekly engagement per device?

****The device that users use the most each week is the MacBook Pro.**

```
3  SELECT  EXTRACT(Month from occurred_at) as month_num,EXTRACT(Week from occurred_at) as week_num,device, count(user_id)
4  from Trainity.Events
5  group by month_num, week_num,device
6  order by month_num, week_num, device;
7
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH PREVIEW

Row	month_num	week_num	device	f0_
1	5	17	acer aspire desktop	80
2	5	17	acer aspire notebook	214
3	5	17	amazon fire phone	87
4	5	17	asus chromebook	265
5	5	17	dell inspiron desktop	193
6	5	17	dell inspiron notebook	517
7	5	17	hp pavilion desktop	143
8	5	17	htc one	201
9	5	17	ipad air	336
10	5	17	ipad mini	217
11	5	17	iphone 4s	231
12	5	17	iphone 5	746
13	5	17	iphone 5s	492
14	5	17	kindle fire	57
15	5	17	lenovo thinkpad	834
16	5	17	mac mini	64

```sql
SELECT action, EXTRACT(MONTH FROM occurred_at) AS month, count(action) as num_emails
FROM Trainity.Email_events
GROUP BY action, month
ORDER BY action, month
```

Query results

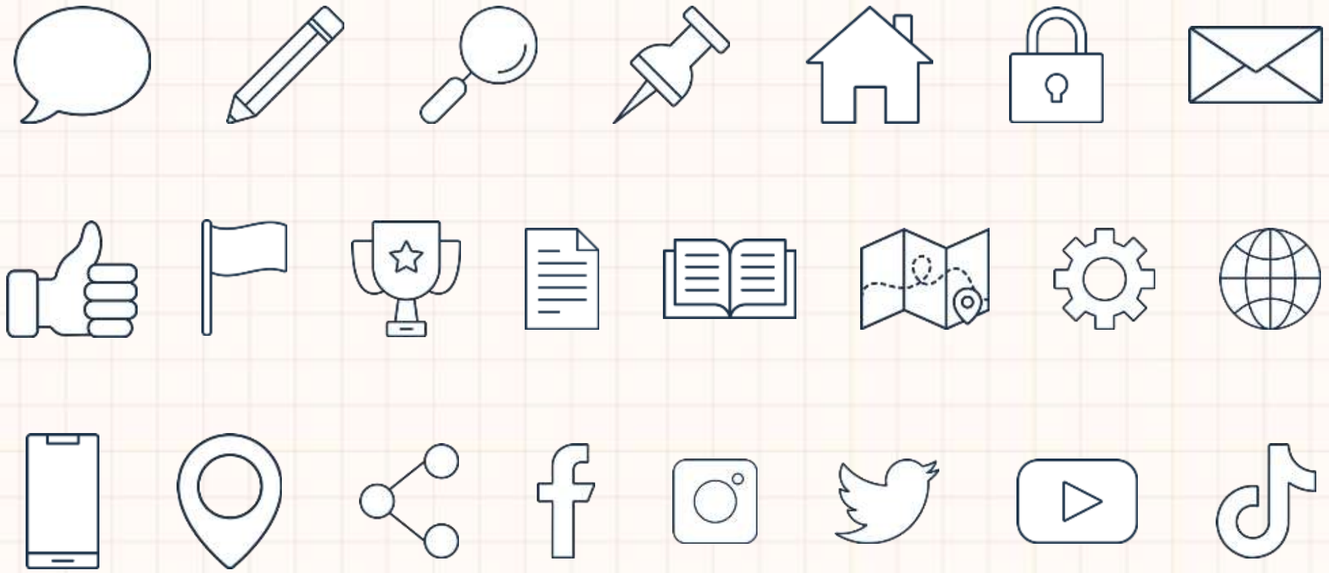
JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH

w	action	month	num_emails
1	email_clickthrough	5	2023
2	email_clickthrough	6	2274
3	email_clickthrough	7	2721
4	email_clickthrough	8	1992
5	email_open	5	4212
6	email_open	6	4658
7	email_open	7	5611
8	email_open	8	5978
9	sent_reengagement_email	5	758
10	sent_reengagement_email	6	889
11	sent_reengagement_email	7	933
12	sent_reengagement_email	8	1073
13	sent_weekly_digest	5	11730
14	sent_weekly_digest	6	13155
15	sent_weekly_digest	7	15902
16	sent_weekly_digest	8	16480

CASE STUDY 2:

(E).Email Engagement: Users engaging with the email service.

Your task: Calculate the email engagement metrics?



APPROACH

List every table I'll require for the query. You need patience and determination to learn something new. In order to understand the advanced ideas of MySQL, I started by studying the platform learning portions. This gave me the groundwork I needed to tackle the case study.

INSIGHTS

Less than 0.01 jobs were rated each hour during the month of November, which had a relatively low amount of reviews.
Among the various languages present, Persian was the most widely spoken.
The device that customers use the most frequently on a weekly basis is the MACbook pro.
Retention rates decreased week after week with a maximum retention of users remaining for only one week.
August was the month in which customers received the most weekly digest emails.

RESULTS And TECH-STACK USED

One of the most important skills for anyone working in a data-driven profession is SQL. It makes it possible to effectively gather data, create metrics, and analyse them.

MY SQL WORKBENCH

Hiring Process Analytics

(STATISTICS)

PROJECT DESCRIPTION

Hiring process is the fundamental and the most important function of a company. Here, the MNCs get 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 analyse before hiring freshers or any other individual. Thus, making an opportunity for a Data Analyst job here too!

Being a Data Analyst, your job is to go through these trends and draw insights out of it for hiring department to work upon.

EXAMPLE: You are working for a MNC such as Google as a lead Data Analyst and the company has provided with the data records of their previous hirings and have asked you to answer certain questions making sense out of that data. You are required to provide a detailed report for the below data record mentioning the answers of the questions that follows:
You are given a dataset of a company where the details about people who registered for a particular post in a department of this company. You are required to use your knowledge in statistics and use different formulas in excel and draw necessary conclusions about the company.

01 APPROACH

I spent some time getting to know the data before starting the analysis. I then move on to my analysis. Once the analysis is complete, explain the results in an understandable and succinct manner. Utilizing visuals to assist you express your results, such as charts and graphs.

03 TECH-STACK USED

MS-EXCEL

02 INSIGHTS AND RESULTS

I received information regarding the number of men and women employed and the questions related while working on the project. Also, obtained the pictorial representations, such as pie charts and bar charts, that detail some of the requested data.

The question and the task to be completed are listed below with the acquired results.

Using the below Steps for EDA

Understanding data columns and data

Checking for missing data

Clubbing columns with multiple categories

Checking for outliers

Removing outliers

Drawing Data Summary

(A)Hiring: Process of intaking of people into an organization for different kinds of positions.

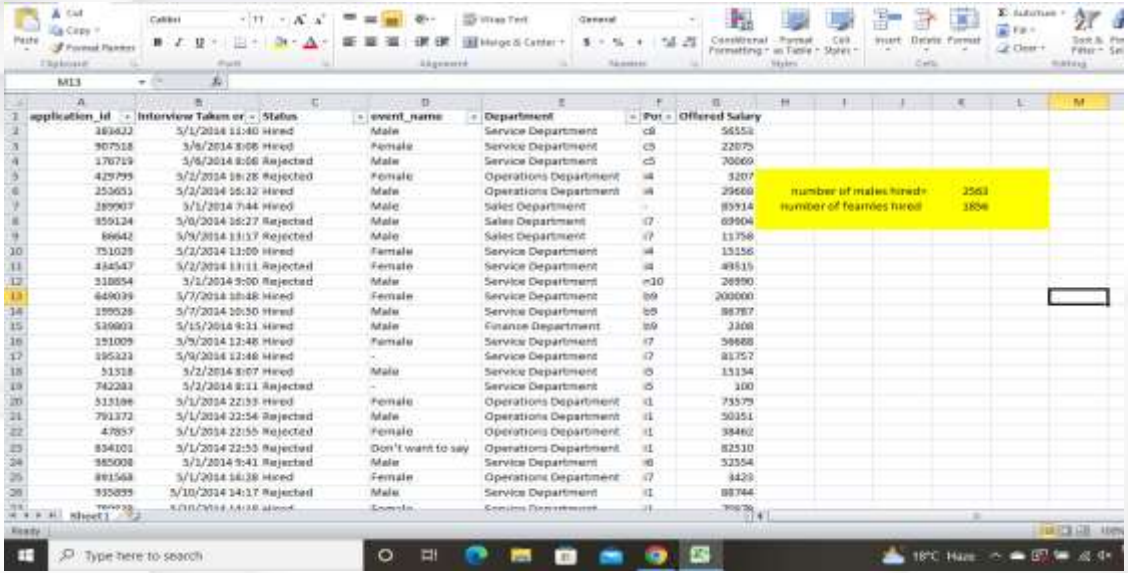
The task: How many males and females are Hired ?

Analysis: FUNCTIONS USED FOR THE OUTPUT:

=COUNTIFS(D2:C7169,"=MALE",C2:C7169,"=HIRED")

=COUNTIFS(D2:D7169,"=FEMALE",C2:C7169,"=HIRED")

OUTPUT:



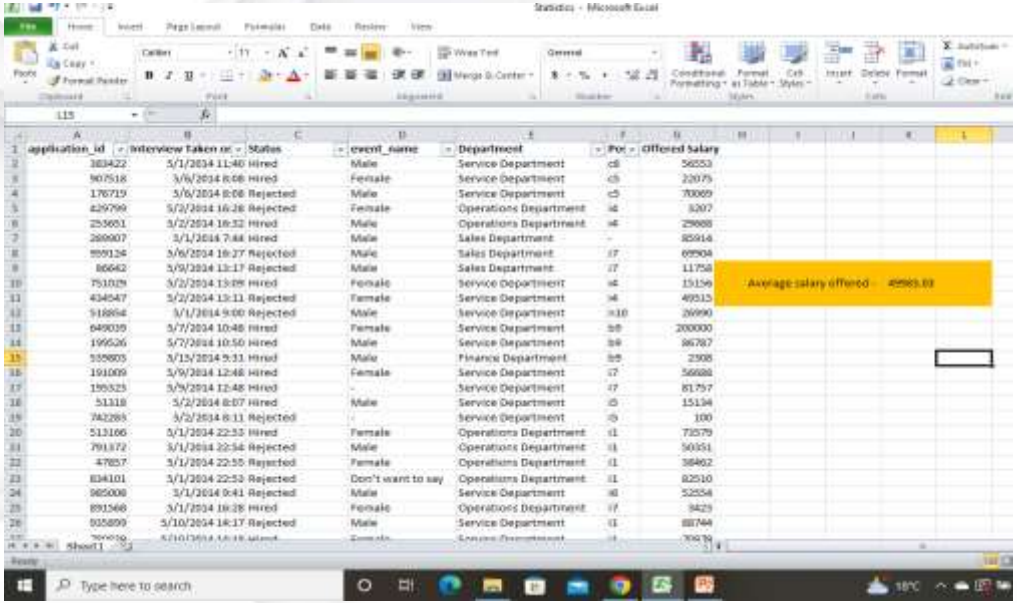
(B)Average Salary: Adding all the salaries for a select group of employees and then dividing the sum by the number of employees in the group.

The task: What is the average salary offered in this company ?

Analysis: FUNCTION USED FOR THE OUTPUT:

=AVERAGE(G2:G7169)

OUTPUT:

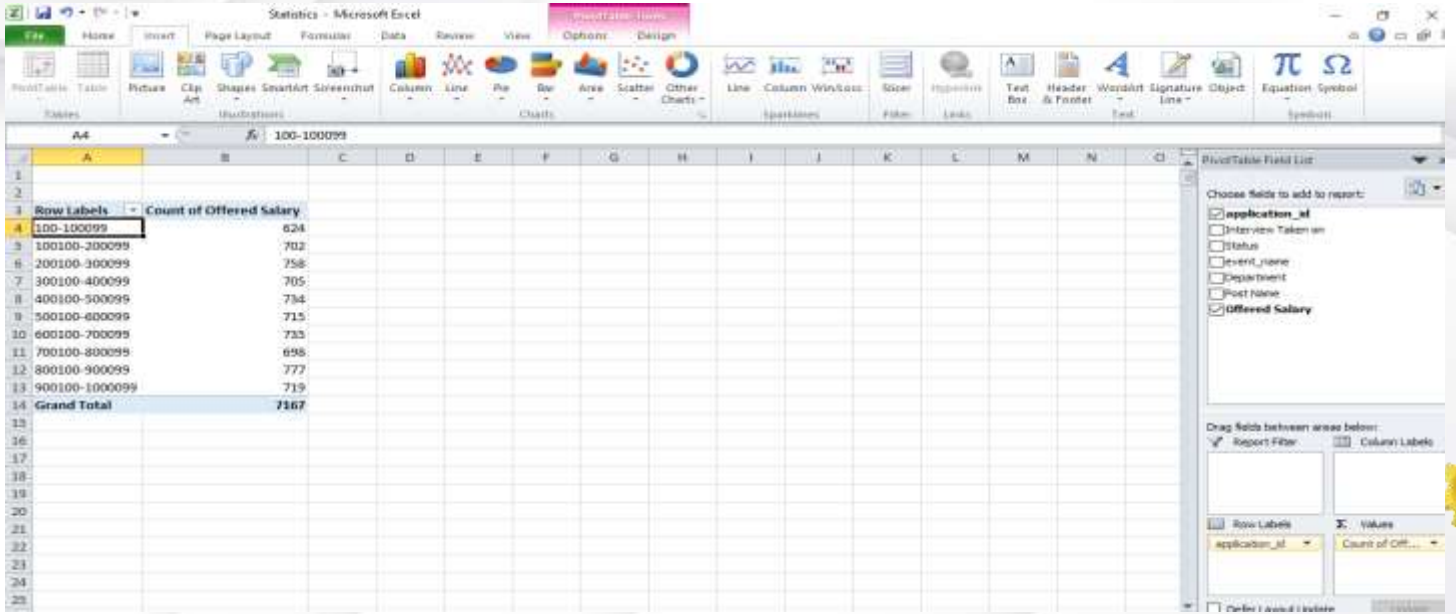


(C)Class Intervals: The class interval is the difference between the upper class limit and the lower class limit.

The task: Draw the class intervals for salary in the company ?

Analysis: Using Pivot table, selecting the fields: application_id and offered salary than grouping them.

OUTPUT:



Row Labels	Count of Offered Salary
100-100099	624
100100-200099	702
200100-300099	758
300100-400099	705
400100-500099	734
500100-600099	715
600100-700099	733
700100-800099	698
800100-900099	777
900100-1000099	719
Grand Total	7167

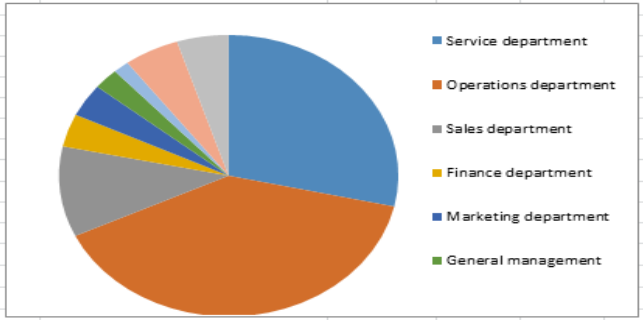
(D)Charts and Plots: This is one of the most important part of analysis to visualize the data.

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

Analysis: The functions used are:

=COUNTIFS($E$2:$E$7169,"=service department",$C$2:$C$7169,"=hired")
=COUNTIFS($E$2:$E$7169,"=operations department",$C$2:$C$7169,"=hired")
=COUNTIFS($E$2:$E$7169,"=sales department",$C$2:$C$7169,"=hired")
=COUNTIFS($E$2:$E$7169,"=finance department",$C$2:$C$7169,"=hired")
=COUNTIFS($E$2:$E$7169,"=finance department",$C$2:$C$7169,"=hired")
=COUNTIFS($E$2:$E$7169,"=General management",$C$2:$C$7169,"=hired")
=COUNTIFS($E$2:$E$7169,"=human resource department",$C$2:$C$7169,"=hired")
=COUNTIFS($E$2:$E$7169,"=production department",$C$2:$C$7169,"=hired")
=COUNTIFS($E$2:$E$7169,"=purchase department",$C$2:$C$7169,"=hired")

DEPARTMENT	TOTAL
Service department	1332
Operations department	1843
Sales department	485
Finance department	176
Marketing department	176
General management	113
Human resource department	70
Production department	246
Purchase department	230



(E)Charts: Use different charts and graphs to perform the task representing the data.

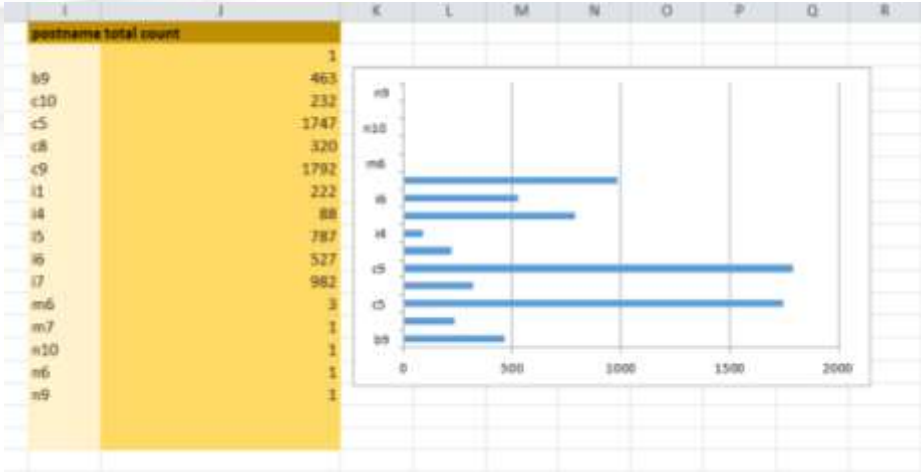
The task: Represent different post tiers using chart/graph?

Analysis:

The functions used are

=COUNTIF($F$2:$F$7169,"= ")
=COUNTIF($F$2:$F$7169,"=b9")
=COUNTIF($F$2:$F$7169,"=c10")
=COUNTIF($F$2:$F$7169,"=c5")
=COUNTIF($F$2:$F$7169,"=c8")
=COUNTIF($F$2:$F$7169,"=c9")
=COUNTIF($F$2:$F$7169,"=i1")
=COUNTIF($F$2:$F$7169,"=i4")
=COUNTIF($F$2:$F$7169,"=i5")
=COUNTIF($F$2:$F$7169,"=i6")
=COUNTIF($F$2:$F$7169,"=i7")
=COUNTIF($F$2:$F$7169,"=m6")
=COUNTIF($F$2:$F$7169,"=m7")
=COUNTIF($F$2:$F$7169,"=n10")
=COUNTIF($F$2:$F$7169,"=n6")
=COUNTIF($F$2:$F$7169,"=n9")

OUTPUT:



postname	total count
	1
b9	463
c10	232
c5	1747
c8	320
c9	1792
i1	222
i4	88
i5	787
i6	527
i7	982
m6	3
m7	1
n10	1
n6	1
n9	1

IMDB Movie Analysis

Final Project-1

Description:

For Final Project, the dataset is provided, having various columns of different IMDB Movies.

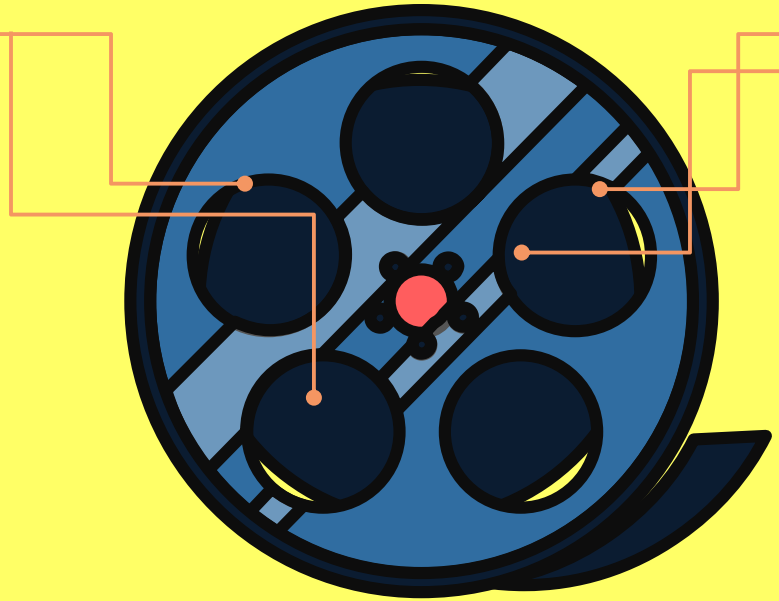
It is required to Frame the problem.

For this task, will need to define a problem wanted to shed some light on.

APPROACH

In order to produce charts and graphs in Excel that satisfy business criteria, we will first clean the data. Before starting the analytics process, I must clean the data by deleting any unnecessary columns or null values. The next step is to locate the films with the largest box office returns, the top 250 films on IMDB with more than 25000 user ratings, the best directors, the hottest genres, and the stars that are both the favorites of critics and viewers. After analysis is complete, visualization is used to aid in communicating findings.

TECH-STACK USED



TECH-STACK USED

DATA CLEANING

This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)
Your task: Clean the data

1.To ensure that any changes I made would not damage the original data, I first created a copy of the raw data on which I could do the analysis. Afterwards, delete any columns that won't be used in the analysis we'll be performing.

2. We now need to eliminate the rows from of the dataset that have any of their column values as blank or NULL after removing the unnecessary columns.

3. Finally, using the "Remove Duplicate Values/Cells" option found in the "Data" tab, we must remove the duplicate values from the dataset.

(A)DATA CLEANING

This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)
Your task: Clean the data

RAW DATA BEFORE CLEANING

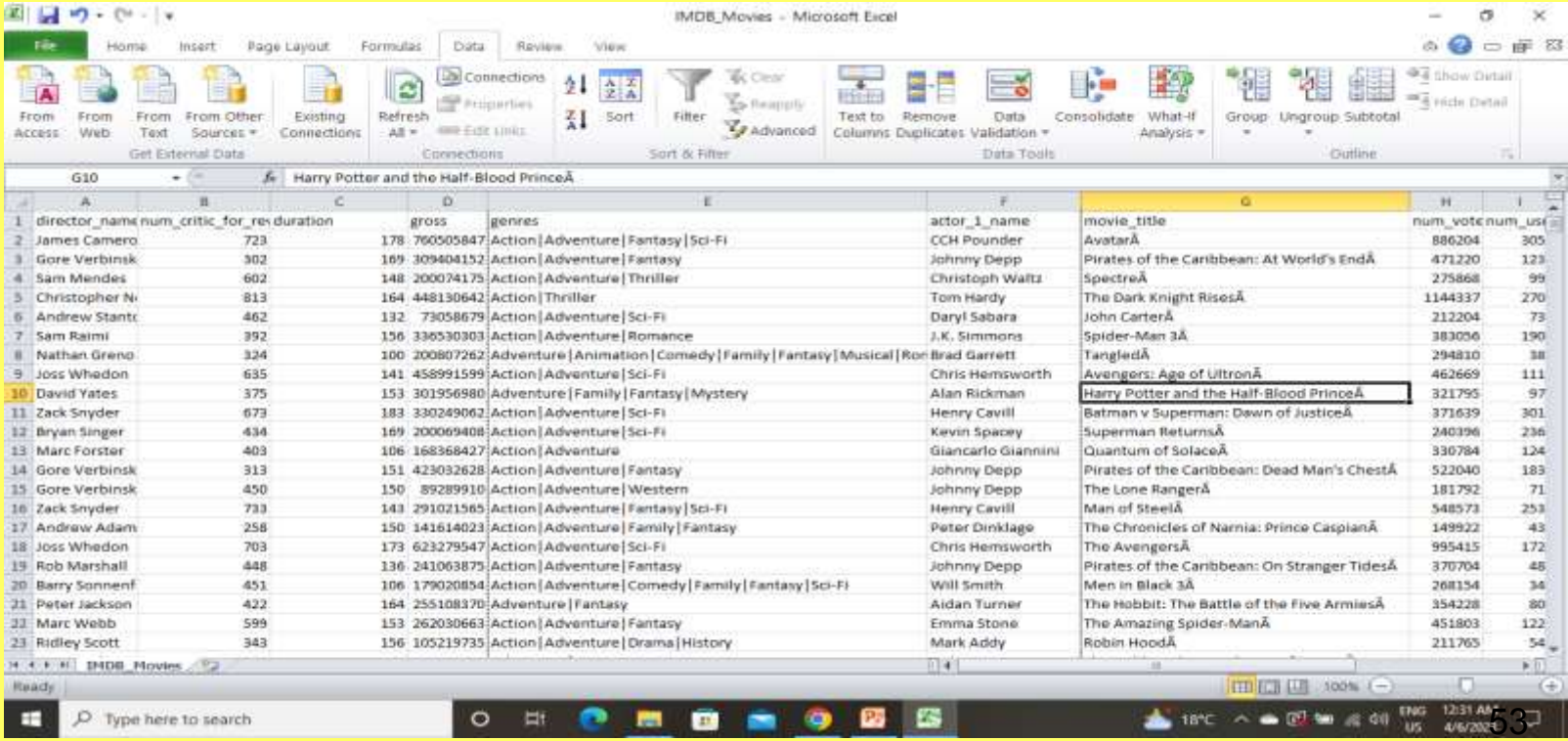
IMDB_Movies - Microsoft Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	color	director_r	num_criti	duration	director_f	actor_3_fa	actor_2_n	actor_1_fa	gross	genres	actor_1_n	movie_tit	num_vote	cast_total	actor_3_n	facenumb	plot_keyv	movie_im	num_user	language	countr
2	Color	James Car	723	178	0	855	Joel Davic	1000	7.61E+08	Action|Ac	CCH Poun	AvatarÂ	886204	4834	Wes Studi	0	avatar|fut	http://ww	3054	English	USA
3	Color	Gore Verb	302	169	563	1000	Orlando B	40000	3.09E+08	Action|Ac	Johnny De	Pirates of	471220	48350	Jack Dave	0	goddess|h	http://ww	1238	English	USA
4	Color	Sam Mend	602	148	0	161	Rory Kinn	11000	2E+08	Action|Ac	Christoph	SpectreÂ	275868	11700	Stephanie	1	bomb|es	http://ww	994	English	UK
5	Color	Christoph	813	164	22000	23000	Christian I	27000	4.48E+08	Action|Th	Tom Hard	The Dark I	1144337	106759	Joseph Gc	0	deception	http://ww	2701	English	USA
6		Doug Walker			131		Rob Walke	131		Documen	Doug Wall	Star Wars:	8	143		0		http://ww			
7	Color	Andrew S	462	132	475	530	Samantha	640	73058679	Action|Ac	Daryl Sabc	John Carte	212204	1873	Polly Wall	1	alien|am	http://ww	738	English	USA
8	Color	Sam Raim	392	156	0	4000	James Fra	24000	3.37E+08	Action|Ac	J.K. Simm	Spider-Ma	383056	46055	Kirsten Du	0	sandman|h	http://ww	1902	English	USA
9	Color	Nathan Gr	324	100	15	284	Donna Mu	799	2.01E+08	Adventur	Brad Garr	TangledÂ	294810	2036	M.C. Gain	1	17th centu	http://ww	387	English	USA
10	Color	Joss Whec	635	141	0	19000	Robert Do	26000	4.59E+08	Action|Ac	Chris Hem	Avengers:	462669	92000	Scarlett Jc	4	artificial i	http://ww	1117	English	USA
11	Color	David Yate	375	153	282	10000	Daniel Rac	25000	3.02E+08	Adventur	Alan Rickr	Harry Pott	321795	58753	Rupert Gri	3	blood|bo	http://ww	973	English	UK
12	Color	Zack Snyd	673	183	0	2000	Lauren Co	15000	3.3E+08	Action|Ac	Henry Cav	Batman v	371639	24450	Alan D. Pu	0	based on	http://ww	3018	English	USA
13	Color	Bryan Sing	434	169	0	903	Marlon Br	18000	2E+08	Action|Ac	Kevin Spa	Superman	240396	29991	Frank Lan	0	crystal|es	http://ww	2367	English	USA
14	Color	Marc Fors	403	106	395	393	Mathieu A	451	1.68E+08	Action|Ac	Giancarlo	Quantum	330784	2023	Rory Kinn	1	action her	http://ww	1243	English	UK
15	Color	Gore Verb	313	151	563	1000	Orlando B	40000	4.23E+08	Action|Ac	Johnny De	Pirates of	522040	48486	Jack Dave	2	box office	http://ww	1832	English	USA
16	Color	Gore Verb	450	150	563	1000	Ruth Wils	40000	89289910	Action|Ac	Johnny De	The Lone I	181792	45757	Tom Wilki	1	horse|out	http://ww	711	English	USA
17	Color	Zack Snyd	733	143	0	748	Christoph	15000	2.91E+08	Action|Ac	Henry Cav	Man of St	548573	20495	Harry Lens	0	based on	http://ww	2536	English	USA
18	Color	Andrew A	258	150	80	201	Pierfrance	22000	1.42E+08	Action|Ac	Peter Dinl	The Chron	149922	22697	DamiÃin A	4	brother be	http://ww	438	English	USA
19	Color	Joss Whec	703	173	0	19000	Robert Do	26000	6.23E+08	Action|Ac	Chris Hem	The Aveng	995415	87697	Scarlett Jc	3	alien inva	http://ww	1722	English	USA
20	Color	Rob Marsh	448	136	252	1000	Sam Claffi	40000	2.41E+08	Action|Ac	Johnny De	Pirates of	370704	54083	Stephen C	4	blackbear	http://ww	484	English	USA
21	Color	Barry Sonr	451	106	188	718	Michael S	10000	1.79E+08	Action|Ac	Will Smith	Men in Bl	268154	12572	Nicole Sch	1	alien|crin	http://ww	341	English	USA
22	Color	Peter Jack	422	164	0	773	Adam Bro	5000	2.55E+08	Adventur	Aidan Tur	The Hobbi	354228	9152	James Ne	0	army|elf|	http://ww	802	English	New Z
23	Color	Marc Web	599	153	464	963	Andrew G	15000	2.62E+08	Action|Ac	Emma Sto	The Amaz	451803	28489	Chris Zylk	0	lizard|out	http://ww	1225	English	USA

(A)DATA CLEANING

This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)
Your task: Clean the data

RAW DATA AFTER CLEANING



MOVIES WITH HIGHEST PROFIT

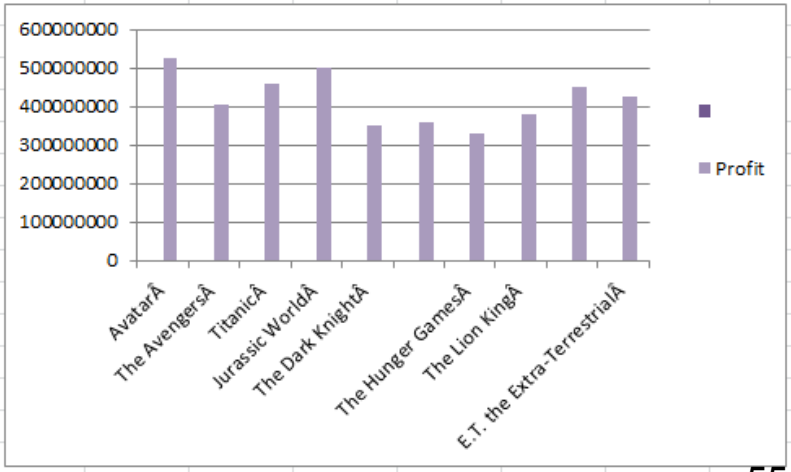
(B)Movies with highest profit

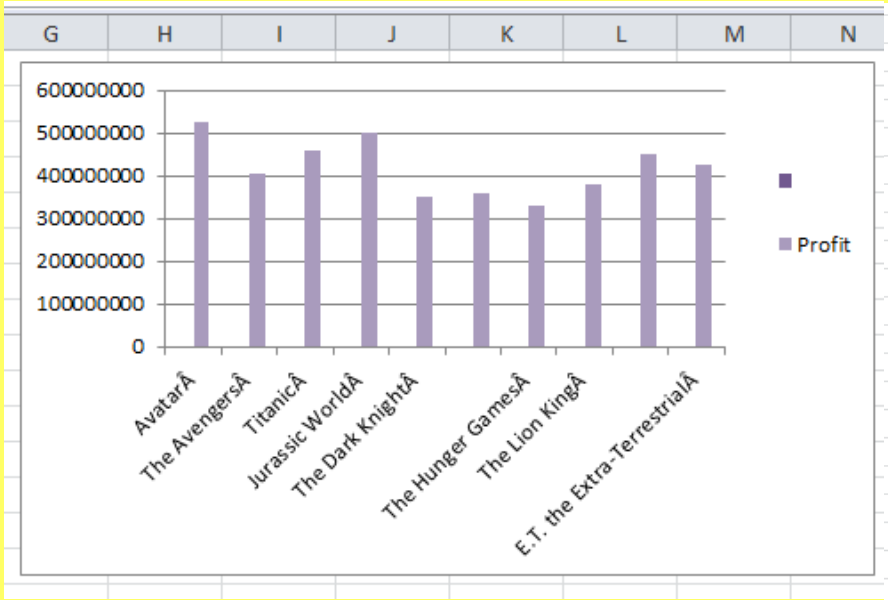
Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) VS budget (x- axis) and observe the outliers using the appropriate chart type.

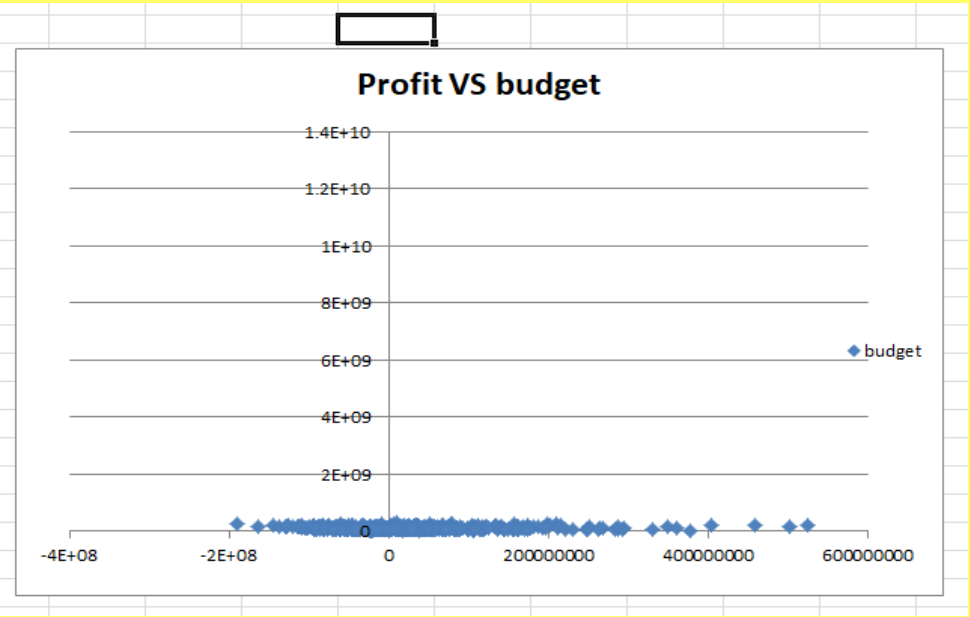
Your task: Find the movies with the highest profit?

- In order to determine which films made the Highest profit we must first calculate the profit by deducting the budget from the gross revenue.
- Then, plotting the graph Profit VS TOP 10 Profitable Movies
- AVATAR IS THE TOP PROFITABLE MOVIE
- Rest the output is given below.

	A	B	C	D	E
1	top 10 profitable movie_title	gross	budget		Profit
2	AvatarÂ	760505847	2.37E+08		523505847
18	The AvengersÂ	623279547	2.2E+08		403279547
27	TitanicÂ	658672302	2E+08		458672302
30	Jurassic WorldÂ	652177271	1.5E+08		502177271
56	The Dark KnightÂ	533316061	1.85E+08		348316061
132	Star Wars: Episode I - The Phantom MenaceÂ	474544677	1.15E+08		359544677
221	The Hunger GamesÂ	407999255	78000000		329999255
284	The Lion KingÂ	422783777	45000000		377783777
705	Star Wars: Episode IV - A New HopeÂ	460935665	11000000		449935665
749	E.T. the Extra-TerrestrialÂ	434949459	10500000		424449459
509					
510					
511					
512					
513					




G
H
I
J
K
L
M
N
600000000
500000000
400000000
300000000
200000000
100000000
0
Profit
AvatarÂ
The AvengersÂ
TitanicÂ
Jurassic WorldÂ
The Dark KnightÂ
The Hunger GamesÂ
The Lion KingÂ
E.T. the Extra-TerrestrialÂ


Profit VS budget
1.4E+10
1.2E+10
1E+10
8E+09
6E+09
4E+09
2E+09
0
budget
-4E+08
-2E+08
0
200000000
400000000
600000000


Top 250:Movies

(C)Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films. Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!

Your task: Find IMDB Top 250

	A	B	C	D	E	F
1	TOP_IMDB_250	num_voted_users	imdb_score		RANK	
2	The Shawshank RedemptionÂ	1689764	9.3		1	
3	The GodfatherÂ	1155770	9.2		2	
4	The Dark KnightÂ	1676169	9		3	
5	The Godfather: Part IIÂ	790926	9		4	
6	The Lord of the Rings: The Return of the KingÂ	1215718	8.9		5	
7	Schindler's ListÂ	865020	8.9		6	
8	Pulp FictionÂ	1324680	8.9		7	
9	The Good, the Bad and the UglyÂ	503509	8.9		8	
10	InceptionÂ	1468200	8.8		9	
11	The Lord of the Rings: The Fellowship of the RingÂ	1238746	8.8		10	
12	Fight ClubÂ	1347461	8.8		11	
13	Forrest GumpÂ	1251222	8.8		12	
14	Star Wars: Episode V - The Empire Strikes BackÂ	837759	8.8		13	
15	The Lord of the Rings: The Two TowersÂ	1100446	8.7		14	
16	The MatrixÂ	1217752	8.7		15	
17	GoodfellasÂ	728685	8.7		16	
18	Star Wars: Episode IV - A New HopeÂ	911097	8.7		17	
19	One Flew Over the Cuckoo's NestÂ	680041	8.7		18	
20	City of GodÂ	533200	8.7		19	

1.Using the sort and filter option, we will first remove any rows with num_voted_users > 25000 in order to discover the top 250 on IMDB.

2. Next, the dataset will be arranged in decreasing order using the imdb_score.

3. Only the top 250 rows will be chosen for further research.

4. Next, we'll use the RANK() function and the formula ==RANK(C2,$N$2:$N$251,0)+COUNTIFS(C$2:$N2,C2)-1

 According to the above statistics, "The Shawshank Redemption" got the highest IMDB ratings.

TOP FOREIGN LANGUAGE FILMS IN IMDB TOP 250 CATEGORY

Then we will filter out (unselect ‘English’) from the language column and we will get the desired output.
The Good, the Bad, and the Ugly has the greatest IMDB ratings among films in all other languages (aside from English), according to the above table. Its country of origin is Itlay.

G	H	I	J	K	L	M	N
movie_title	num_vote	num_user	languag	country	budget	title_year	imdb_score
The Good, the Bad and the UglyÂ	503509	780	Italian	Italy	1200000	1966	8.9
City of GodÂ	533200	749	Portugues	Brazil	3300000	2002	8.7
Seven SamuraiÂ	229012	596	Japanese	Japan	2000000	1954	8.7
Spirited AwayÂ	417971	902	Japanese	Japan	1.9E+07	2001	8.6
SamsaraÂ	22457	69	None	USA	4000000	2011	8.5


(D)BEST DIRECTORS

(D)TGroup the column using the director_name column.
Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.
Your task: Find the best directors

- Select Director_name,imdb_score column and create a pivot table.
- In values find average for Imdb_score and sort it in desending order and find rank using array formula =RANK(B2,$B$2:$B$13,0)+COUNTIF($B$2:B2,B2)-1 (Drag and drop)

and results are given:

2			
3	TOP 10 DIRECTORS	Average of imdb_score	RANK
4	Tony Kaye	8.6	1
5	Charles Chaplin	8.6	2
6	Ron Fricke	8.5	3
7	Majid Majidi	8.5	4
8	Damien Chazelle	8.5	5
9	Alfred Hitchcock	8.5	6
10	Sergio Leone	8.433333333	7
11	Christopher Nolan	8.425	8
12	Richard Marquand	8.4	9
13	S.S. Rajamouli	8.4	10

The highest mean IMDB Score, 8.6, was possessed by Charles Chaplin and Tony Kaye, according to the aforementioned table.

(E)POPULAR GENRES

(E) Perform this step using the knowledge gained while performing previous steps.
Your task: Find popular genres

Row Labels	Average of imdb_score	RANK
Crime|Drama|Fantasy|Mystery	8.5	1
Adventure|Animation|Drama|Family|Musical	8.5	2
Adventure|Drama|Thriller|War	8.4	3
Adventure|Animation|Fantasy	8.4	4
Action|Adventure|Drama|Fantasy|War	8.4	5
Documentary|Drama|Sport	8.3	6
Documentary|War	8.3	7
Biography|Drama|History|Music	8.3	8
Adventure|Animation|Comedy|Drama|Family|Fantasy	8.3	9
Adventure|Drama|War	8.25	10
Drama|Fantasy|War	8.2	11
Biography|Crime|Documentary|History	8.2	12
Drama|Mystery|War	8.2	13

- Select genere,imdb_score column and create pivot Table

- find mean for Imdb_score and sort it from largest to smallest

- Sort popular genres in ascending(A to Z)

- And find rank using array formula

=RANK(B3,$B$3:$B$14,0)+COUNTIF($B$3:B3,B3)-1

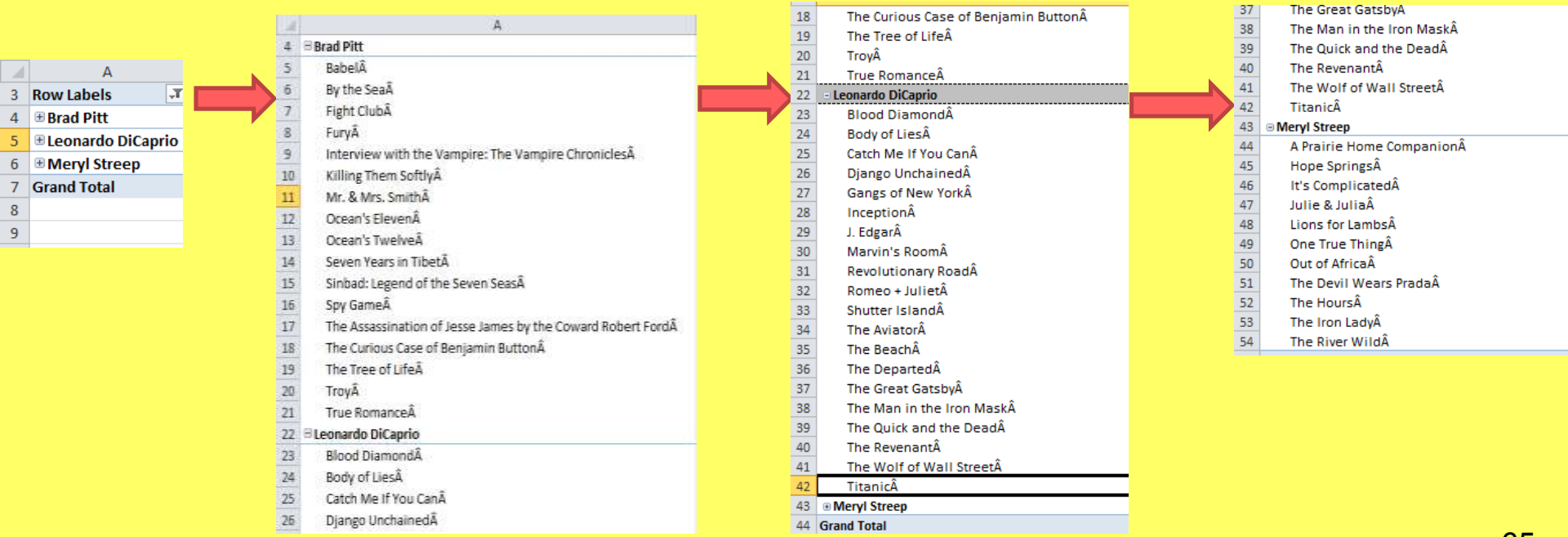
(F) Charts: Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names 'Meryl Streep',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

Append the rows of all these columns and store them in a new column named Combined. Group the combined column using the actor_1_name column.

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

Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.
Your task: Find the critic-favorite and audience-favorite actors

(F) 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



	A
3	Row Labels
4	⊞Brad Pitt
5	⊞Leonardo DiCaprio
6	⊞Meryl Streep
7	Grand Total
8	
9	

	A
4	⊟Brad Pitt
5	BabelÂ
6	By the SeaÃ
7	Fight ClubÃ
8	FuryÃ
9	Interview with the Vampire: The Vampire ChroniclesÃ
10	Killing Them SoftlyÃ
11	Mr. & Mrs. SmithÃ
12	Ocean's ElevenÃ
13	Ocean's TwelveÃ
14	Seven Years in TibetÃ
15	Sinbad: Legend of the Seven SeasÃ
16	Spy GameÃ
17	The Assassination of Jesse James by the Coward Robert FordÃ
18	The Curious Case of Benjamin ButtonÂ
19	The Tree of LifeÂ
20	TroyÂ
21	True RomanceÂ
22	⊟Leonardo DiCaprio
23	Blood DiamondÂ
24	Body of LiesÂ
25	Catch Me If You CanÂ
26	Django UnchainedÂ

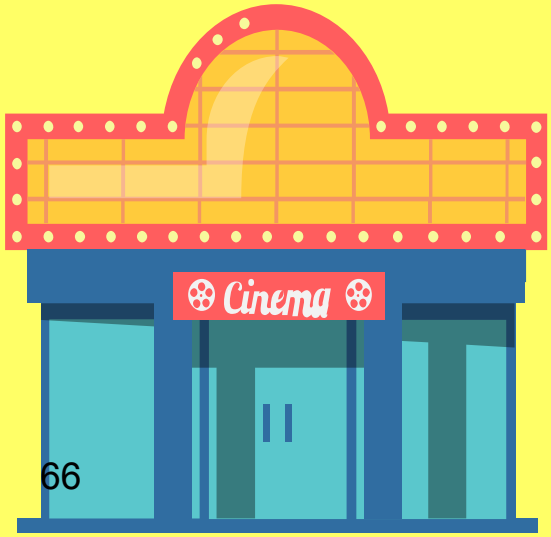
18	The Curious Case of Benjamin ButtonÂ
19	The Tree of LifeÂ
20	TroyÂ
21	True RomanceÂ
22	⊟Leonardo DiCaprio
23	Blood DiamondÂ
24	Body of LiesÂ
25	Catch Me If You CanÂ
26	Django UnchainedÂ
27	Gangs of New YorkÂ
28	InceptionÂ
29	J. EdgarÂ
30	Marvin's RoomÂ
31	Revolutionary RoadÂ
32	Romeo + JulietÂ
33	Shutter IslandÂ
34	The AviatorÂ
35	The BeachÂ
36	The DepartedÂ
37	The Great GatsbyÂ
38	The Man in the Iron MaskÂ
39	The Quick and the DeadÂ
40	The RevenantÂ
41	The Wolf of Wall StreetÂ
42	TitanicÂ
43	⊞Meryl Streep
44	Grand Total

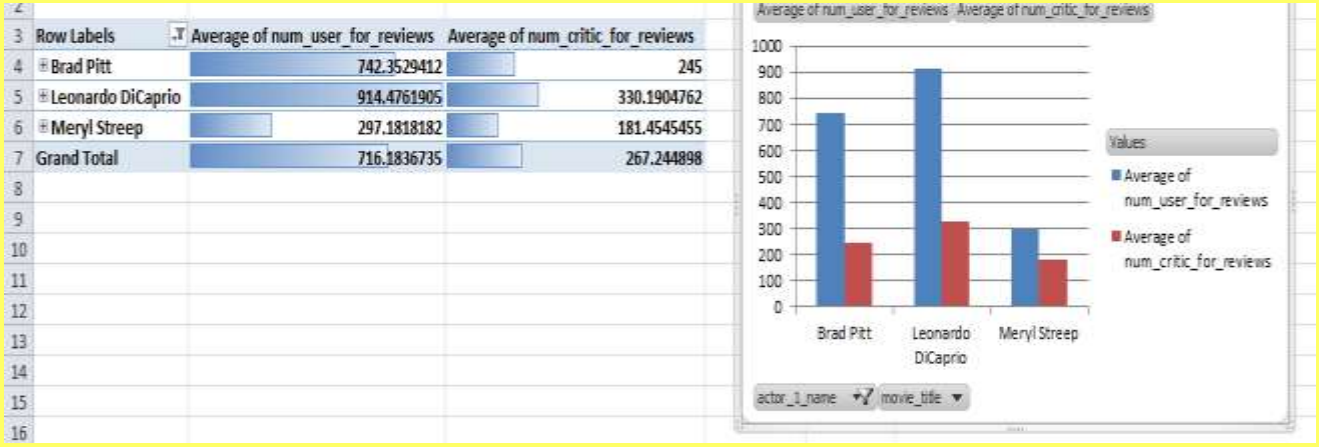
37	The Great GatsbyÂ
38	The Man in the Iron MaskÂ
39	The Quick and the DeadÂ
40	The RevenantÂ
41	The Wolf of Wall StreetÂ
42	TitanicÂ
43	⊟Meryl Streep
44	A Prairie Home CompanionÂ
45	Hope SpringsÂ
46	It's ComplicatedÂ
47	Julie & JuliaÂ
48	Lions for LambsÂ
49	One True ThingÂ
50	Out of AfricaÂ
51	The Devil Wears PradaÂ
52	The HoursÂ
53	The Iron LadyÂ
54	The River WildÂ

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

Leonardo DiCaprio is the one with the highest mean od number of users for review and also for number of critic review

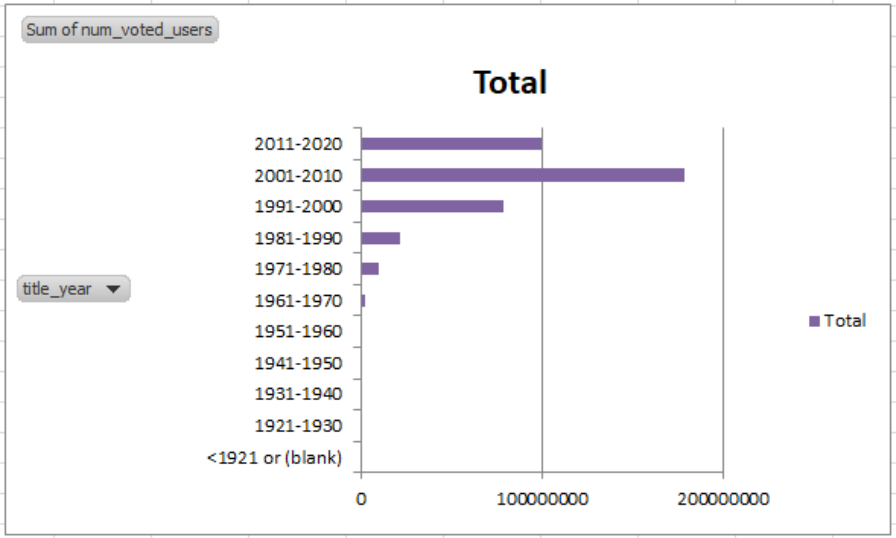
	A	B	C	D
3	Row Labels	Average of num_user_for_reviews	Average of num_critic_for_reviews	
4	Brad Pitt	742.3529412	245	
5	Leonardo DiCaprio	914.4761905	330.1904762	
6	Meryl Streep	297.1818182	181.4545455	
7	Grand Total	716.1836735	267.244898	
8				





Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade

Row Labels	Sum of num_voted_users
<1921 or (blank)	
1921-1930	116387
1931-1940	966520
1941-1950	72324
1951-1960	1097601
1961-1970	2607791
1971-1980	10245198
1981-1990	22100624
1991-2000	78706936
2001-2010	178631587
2011-2020	100252431
Grand Total	**394797399**



INSIGHTS AND RESULT:

Both during the pre-production and post-production phases of a movie, analysis is a key component. Additionally, it's not a given that the movie with the highest IMDB rating will also make the most money.

The quantity of tickets sold by theatres around the world is the real basis for profit calculation.

As a conclusion, I'd like to state that, prior to the production of a film, not only movie producers but also a variety of financiers, stakeholders, and cinema outlet owners perform IMDB Movie Analysis or any other similar analysis.

Bank Loan Case Study

Final Project-2

Description

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.

Business Understanding:

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. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company:

Approved: The company has approved loan application
Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
Refused: The company had rejected the loan (because the client does not meet their requirements etc.).
Unused Offer: Loan has been cancelled by the client but on different stages of the process.
In this case study, I will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Approach

There are two sets of substantial data in this case study: the current application and the prior application. Both contained numerous unnecessary columns that would be of little use for risk analytics, as well as numerous empty values. Before using pivot tables and charts to analyse this sizable amount of data for univariate and bivariate analysis, I first cleaned the data, found some outliers, and deleted them. Understanding the given data's columns, planning different combinations from the data at hand to produce results utilising pivot tables and graphs.

TECH-STACK USED

EXCEL

INSIGHTS

Worked with a large data set, and it took some time for me to understand the loan approval process for the banking sector, EDA. Learnt some concepts about risk analysis, univariate and bivariate analysis, correlation, and how to use it in Excel through the use of various charts, graphs, and scatter plots.

Overall strategy used in the analysis:

- By examining prior information, the study seeks to ascertain the credibility of a borrower or the risk involved in releasing a loan.
- We received 3 csv files, 2 of which contained loan application data and were named application_data.csv and previous_application, respectively, while the third file was named columns_description. Since my data has many missing values and undesired rows and columns, I first imported my data into Excel and checked for any missing values. Then I looked for outliers, and finally I looked for data imbalance. Additionally, the outcomes of segmented univariate, bivariate, and univariate analysis will be performed.
- As a final step, we will examine the relationship between the target variable and the Client's payment troubles.

A. IMPORTING THE DATASET PROVIDED:

1. PREVIOUS APPLICATION SHEET:

	A	B	C	D	E	F	G	H	I	J
1	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_T	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYME	AMT_GOODS_PRICE	WEEKDAY_APPR	HOUR_APPR
2	1000001	158271	Consumer loans	6404.31	58905	65124	0	58905	THURSDAY	8
3	1000002	101962	Consumer loans	6264	39145.5	35230.5	3915	39145.5	SUNDAY	8
4	1000005	176456	Consumer loans	14713.605	123486.075	120307.5	12349.575	123486.075	THURSDAY	13
5	1000008	152059	Consumer loans	26331.66	249255	224325	24930	249255	MONDAY	14
6	1000009	343078	Consumer loans	9302.85	42705	45243	0	42705	SATURDAY	11
7	1000011	198678	Cash loans	92435.04	855000	879831		855000	SUNDAY	18
8	1000013	215520	Consumer loans	13347.9	134955	133479	13500	134955	THURSDAY	18
9	1000016	157990	Consumer loans	6078.15	63720	56970	6750	63720	WEDNESDAY	16
10	1000017	310743	Consumer loans	7002.72	67864.23	75024	4.23	67864.23	SATURDAY	11
11	1000020	299072	Consumer loans	18393.165	151555.5	164961	0	151555.5	TUESDAY	10
12	1000024	448518	Consumer loans	2970.765	24705	24430.5	2475	24705	THURSDAY	15
13	1000026	227096	Consumer loans	23357.655	123300	129811.5	0	123300	WEDNESDAY	13
14	1000027	277601	Cash loans	8806.455	45000	46485		45000	FRIDAY	11
15	1000028	409793	Consumer loans	5613.12	60835.5	54751.5	6084	60835.5	SATURDAY	14
16	1000029	441872	Consumer loans	8659.305	46557	49014	0	46557	FRIDAY	15
17	1000031	131335	Revolving loans	2250	45000	45000		45000	FRIDAY	11
18	1000032	138624	Consumer loans	10561.23	103459.5	115155	0	103459.5	MONDAY	15
19	1000035	436351	Revolving loans	2250	45000	45000		45000	SATURDAY	9
20	1000039	388903	Consumer loans	3354.685	29416.5	26469	6750	29416.5	FRIDAY	13
21	1000040	151267	Cash loans	9686.34	45000	51898.5		45000	TUESDAY	13
22	1000043	423289	Consumer loans	5644.215	44995.5	49450.5	0	44995.5	SUNDAY	13
23	1000044	391316	Consumer loans	13243.59	59501.25	46485	14501.25	59501.25	THURSDAY	15

previous_application Sheet1

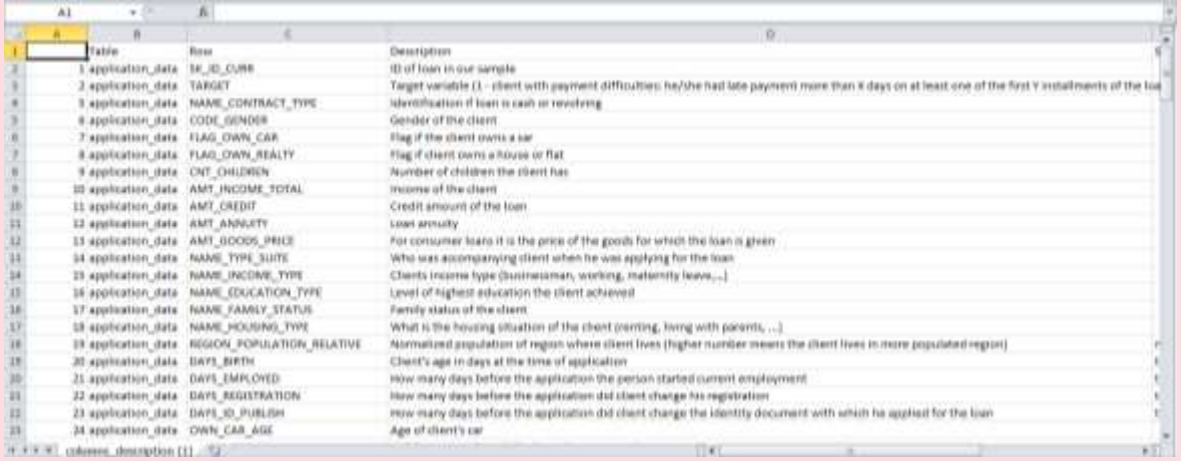
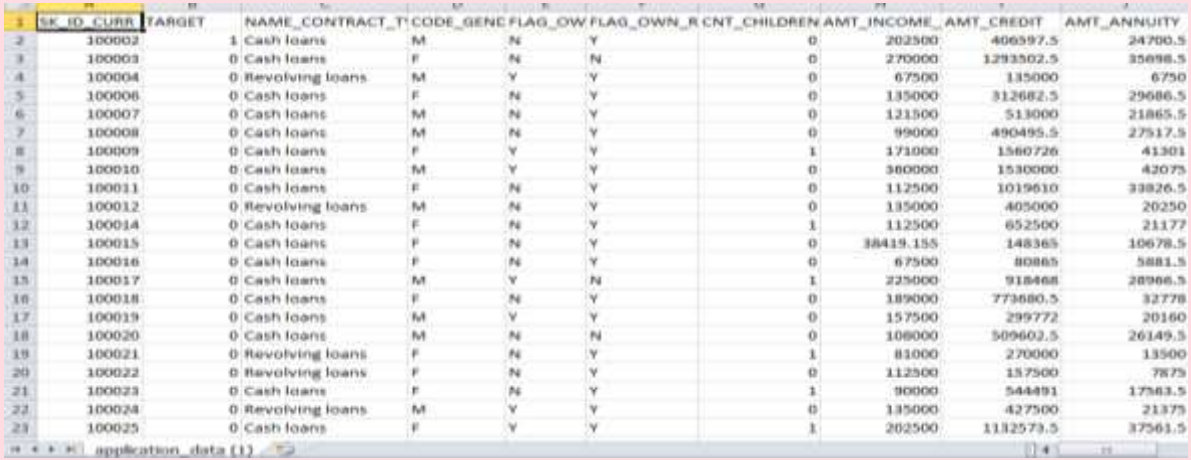


	Table	Row	Description
1	application_data	SK_ID_CURR	ID of loan in our sample
2	application_data	TARGET	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loa
3	application_data	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
4	application_data	CODE_GENDER	Gender of the client
5	application_data	FLAG_OWN_CAR	Flag if the client owns a car
6	application_data	FLAG_OWN_REALTY	Flag if client owns a house or flat
7	application_data	CNT_CHILDREN	Number of children the client has
8	application_data	AMT_INCOME_TOTAL	Income of the client
9	application_data	AMT_CREDIT	Credit amount of the loan
10	application_data	AMT_ANNUITY	Loan annuity
11	application_data	AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
12	application_data	NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan
13	application_data	NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave,...)
14	application_data	NAME_EDUCATION_TYPE	Level of highest education the client achieved
15	application_data	NAME_FAMILY_STATUS	Family status of the client
16	application_data	NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents, ...)
17	application_data	REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means the client lives in more populated region)
18	application_data	DAYS_BIRTH	Client's age in days at the time of application
19	application_data	DAYS_EMPLOYED	How many days before the application the person started current employment
20	application_data	DAYS_REGISTRATION	How many days before the application did client change his registration
21	application_data	DAYS_ID_PUBLISH	How many days before the application did client change the identity document with which he applied for the loan
22	application_data	OWN_CAR_AGE	Age of client's car

columns_description (1)

2. COLUMN_DESCRIPTION SHEET



SK_ID_CURR	TARGET	NAME_CONTRACT_T	CODE_GENE	FLAG_OW	FLAG_OWN_R	CNT_CHILDREN	AMT_INCOME_	AMT_CREDIT	AMT_ANNUITY
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5
100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5
100012	0	Revolving loans	M	N	Y	0	135000	405000	20250
100014	0	Cash loans	F	N	Y	1	112500	652500	21177
100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5
100017	0	Cash loans	M	Y	N	1	225000	918468	28986.5
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778
100019	0	Cash loans	M	Y	Y	0	157500	299772	20160
100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5
100021	0	Revolving loans	F	N	Y	1	81000	270000	13500
100022	0	Revolving loans	F	N	Y	0	112500	157500	7875
100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5
100024	0	Revolving loans	M	Y	Y	0	135000	427500	21375
100025	0	Cash loans	F	Y	Y	1	202500	1132573.5	37561.5

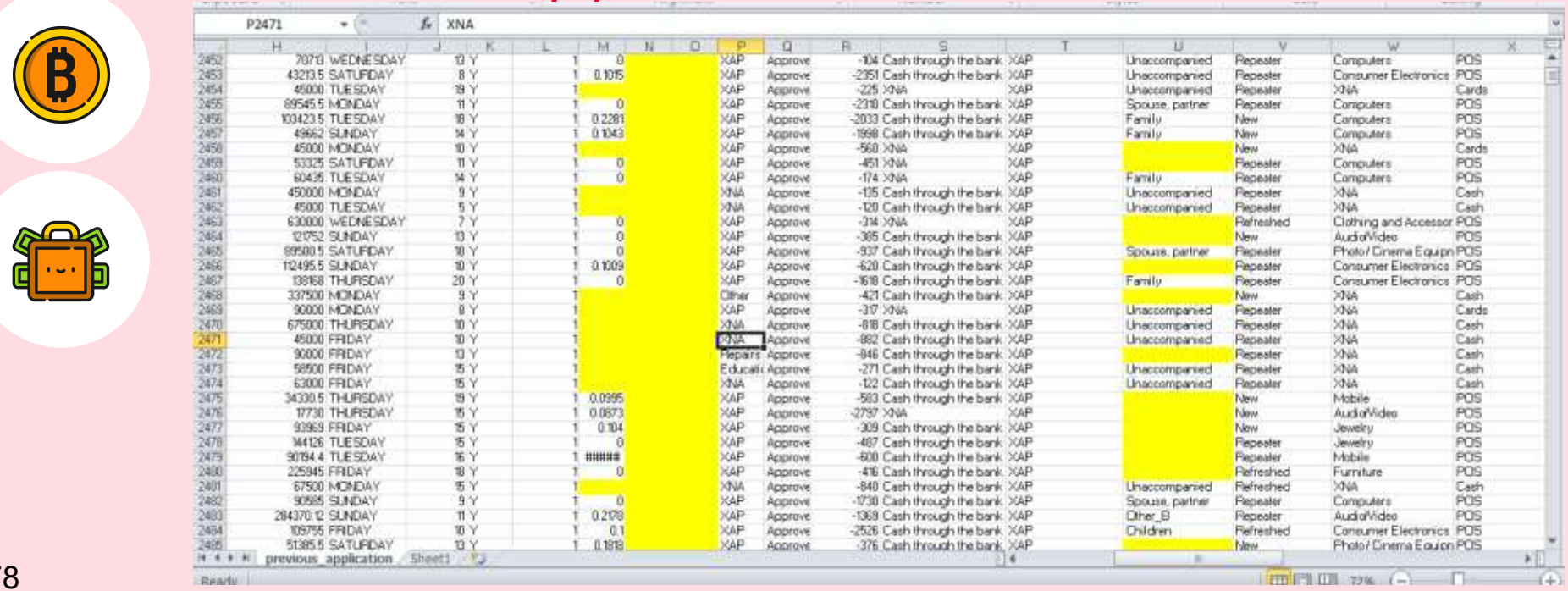
application_data (1)

3. APPLICATION_DATA SHEET

B. CLEANING THE DATA:

Analysis of the percentage of null values is required, and columns with more than 40% null data must be removed. Additionally, those columns with less than 40% of null data must be replaced with the mean, median, or the category variable with the highest frequency.

first, draw attention to the empty cells.



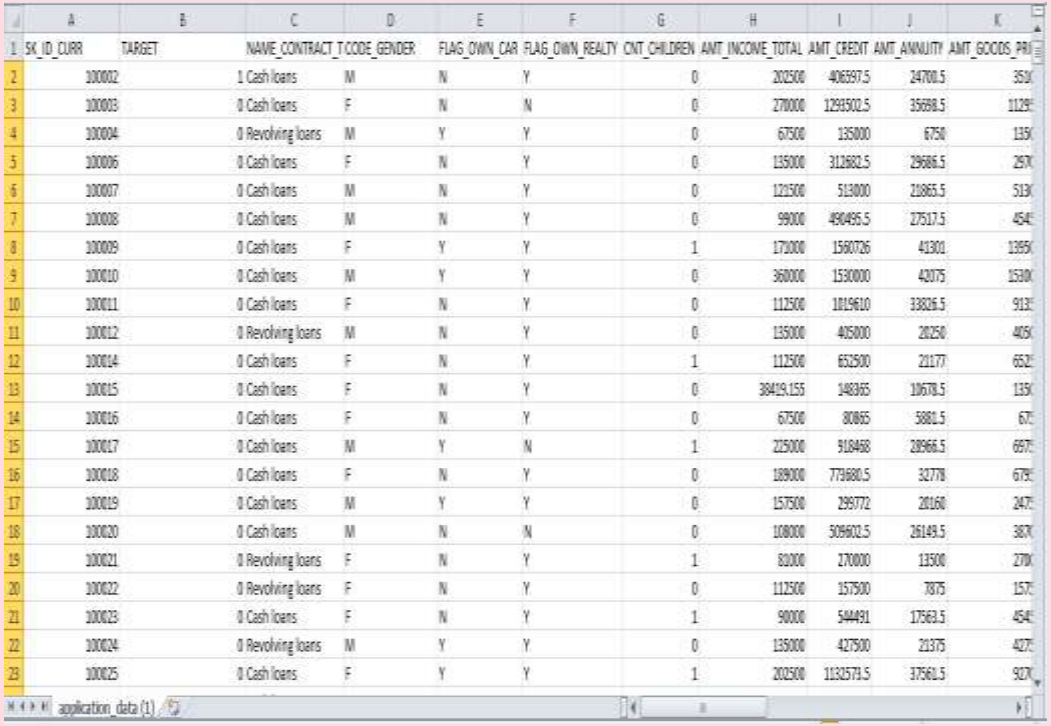
-The columns must be removed since they contain more than 50% NULL data.
-Drop the columns that are unnecessary for performing the data analysis after that.
-Remove duplicates if any.
-Imputing the missing values using mean median and mode

-Filling in blanks in the Application Dataset's Occupation_Type column with the categorical variable with the highest frequency 'Labourers' is the most frequent categorical variable.

-Replacing blanks in the Application Dataset's AMT_ANNUTIY column with the median value of AMT_ANNUITY since the column contains outliers
--> AMT_ANNUITY median: 24903
Blanks in the Application Dataset's AMT_GOODS_PRICE column should be replaced with the median of AMT_GOODS_PRICE because the column contains outliers.
--> AMT_GOODS_PRICE median: 450000.

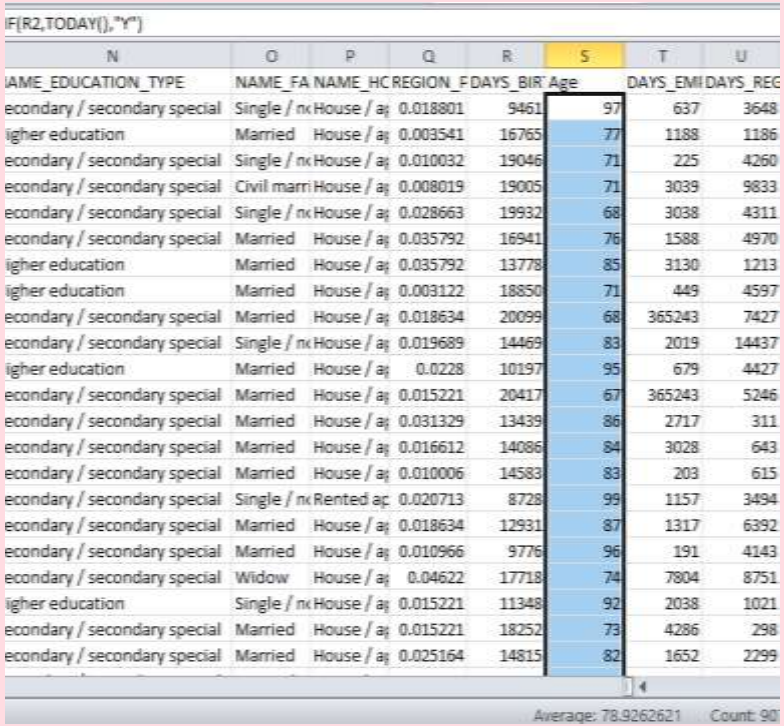
Filling in the Application Dataset's Name_Type_Suite column's blanks with the highest-occurring categorical variable
-- 'Unaccompanied' is the categorical variable with the highest frequency.

-Remove negative numeric values and convert to positive and find age using formula
-=DATEDIF(R2,TODAY(),"Y")



SK_ID_CURR	TARGET	NAME_CONTRACT_T	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRI
100002		1 Cash loans	M	N	Y	0	202500	406597.5	24700.5	351
100003		0 Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129
100004		0 Revolving loans	M	Y	Y	0	67500	135000	6750	135
100006		0 Cash loans	F	N	Y	0	135000	312682.5	29686.5	297
100007		0 Cash loans	M	N	Y	0	121500	513000	21865.5	513
100008		0 Cash loans	M	N	Y	0	99000	490495.5	27517.5	454
100009		0 Cash loans	F	Y	Y	1	171000	1560726	41301	1395
100010		0 Cash loans	M	Y	Y	0	360000	1530000	42075	1530
100011		0 Cash loans	F	N	Y	0	112500	1019610	33826.5	913
100012		0 Revolving loans	M	N	Y	0	135000	405000	20250	405
100014		0 Cash loans	F	N	Y	1	112500	652500	21177	652
100015		0 Cash loans	F	N	Y	0	38419.155	148365	10678.5	135
100016		0 Cash loans	F	N	Y	0	67500	80865	5881.5	67
100017		0 Cash loans	M	Y	N	1	225000	918468	28966.5	657
100018		0 Cash loans	F	N	Y	0	189000	773680.5	32778	675
100019		0 Cash loans	M	Y	Y	0	157500	299772	20160	247
100020		0 Cash loans	M	N	N	0	108000	509602.5	26149.5	387
100021		0 Revolving loans	F	N	Y	1	81000	270000	13500	270
100022		0 Revolving loans	F	N	Y	0	112500	157500	7875	157
100023		0 Cash loans	F	N	Y	1	90000	544491	17563.5	454
100024		0 Revolving loans	M	Y	Y	0	135000	427500	21375	427
100025		0 Cash loans	F	Y	Y	1	202500	1132573.5	37561.5	927

application_data (1)



=DATEDIF(R2,TODAY(),"Y")

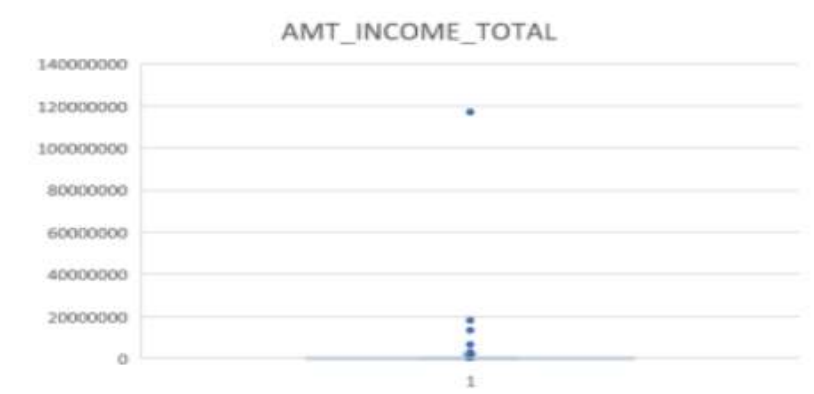
NAME_EDUCATION_TYPE	NAME_FA	NAME_HC	REGION_F	DAYS_BIR	Age	DAYS_EMI	DAYS_REG
econdary / secondary special	Single / no	House / ap	0.018801	9461	97	637	3648
igher education	Married	House / ap	0.003541	16765	77	1188	1186
econdary / secondary special	Single / no	House / ap	0.010032	19046	71	225	4260
econdary / secondary special	Civil marri	House / ap	0.008019	19005	71	3039	9833
econdary / secondary special	Single / no	House / ap	0.028663	19932	68	3038	4311
econdary / secondary special	Married	House / ap	0.035792	16941	76	1588	4970
igher education	Married	House / ap	0.035792	13778	85	3130	1213
igher education	Married	House / ap	0.003122	18850	71	449	4597
econdary / secondary special	Married	House / ap	0.018634	20099	68	365243	7427
econdary / secondary special	Single / no	House / ap	0.019689	14469	83	2019	14437
igher education	Married	House / ap	0.0228	10197	95	679	4427
econdary / secondary special	Married	House / ap	0.015221	20417	67	365243	5246
econdary / secondary special	Married	House / ap	0.031329	13439	86	2717	311
econdary / secondary special	Married	House / ap	0.016612	14086	84	3028	643
econdary / secondary special	Married	House / ap	0.010006	14583	83	203	615
econdary / secondary special	Single / no	Rented ap	0.020713	8728	99	1157	3494
econdary / secondary special	Married	House / ap	0.018634	12931	87	1317	6392
econdary / secondary special	Married	House / ap	0.010966	9776	96	191	4143
econdary / secondary special	Widow	House / ap	0.04622	17718	74	7804	8751
igher education	Single / no	House / ap	0.015221	11348	92	2038	1021
econdary / secondary special	Married	House / ap	0.015221	18252	73	4286	298
econdary / secondary special	Married	House / ap	0.025164	14815	82	1652	2299

Average: 78.9262621 Count: 90

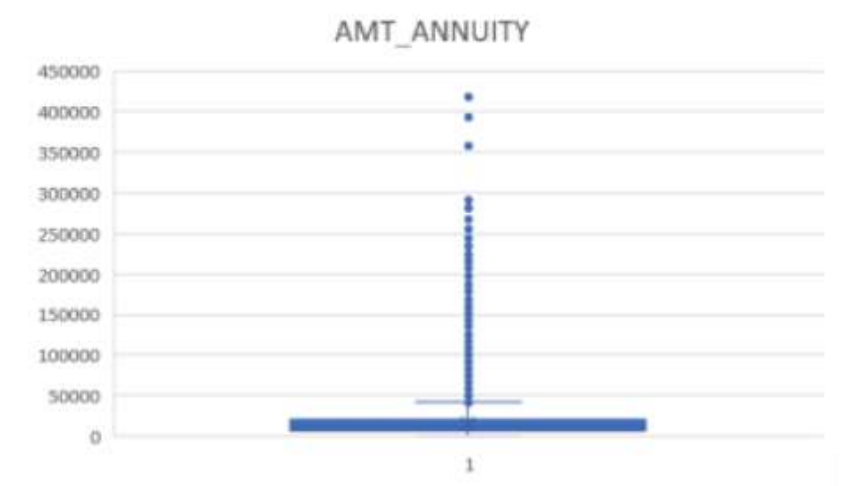
C. Identify if there are outliers in the dataset.

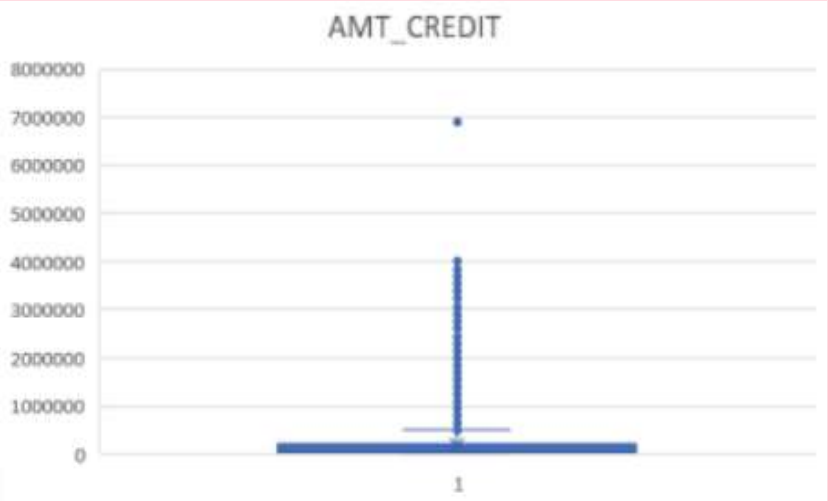
• Application_data.csv

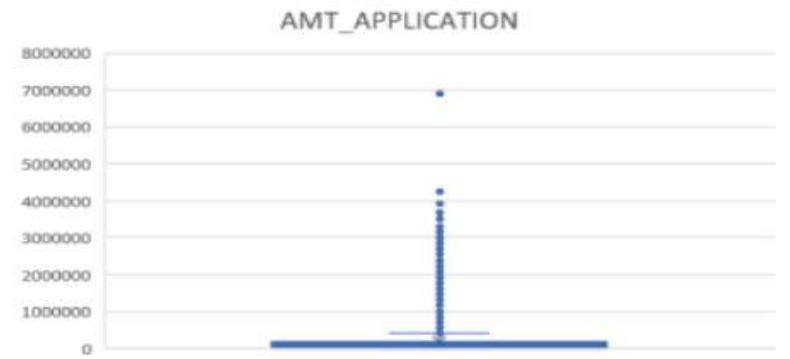
For outliers, we employ the IQR approach, where a number is considered an outlier if it is higher than the upper value or lower than the lower value.



• **previous_application.csv**

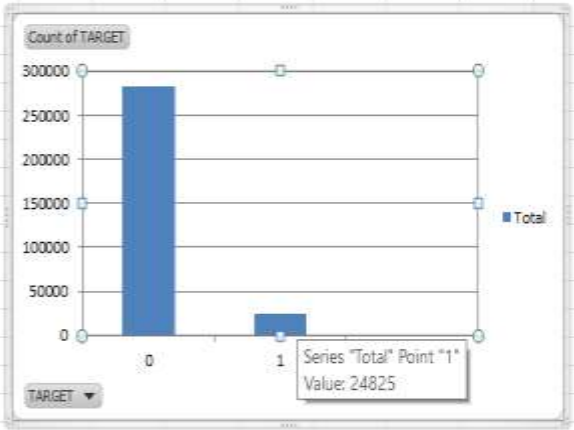






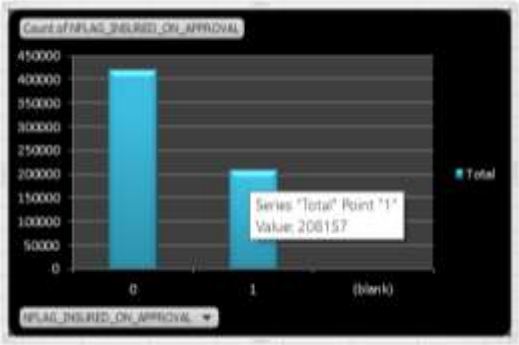
- **D. Data Imbalance :**

Row Labels	Count of TARGET
0	282686
1	24825
(blank)	
Grand Total	307511



This chart shows that there is 282686 of 0 value and 24825 of 1 value which shows data is imbalance.

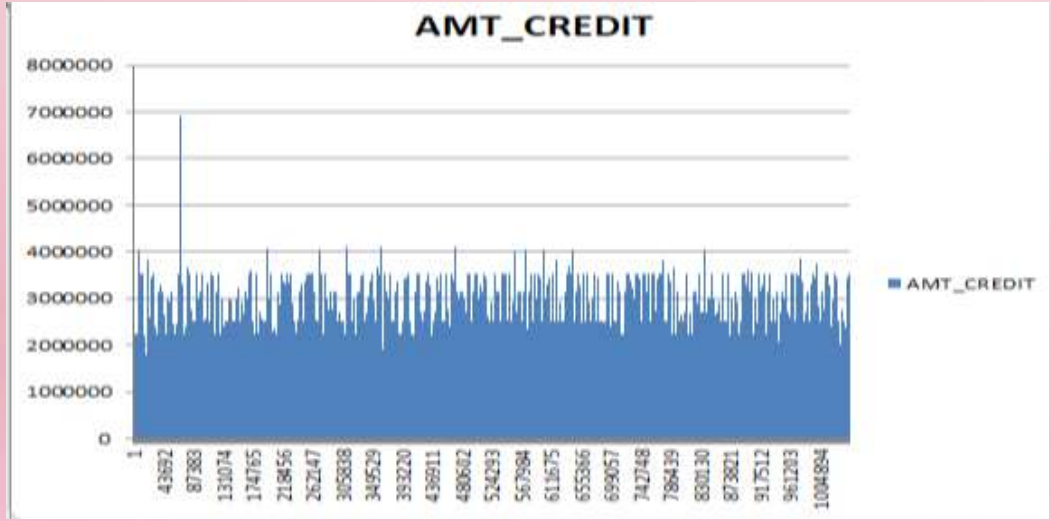
Row Labels	Count of NFLAG_INSURED_ON_APPROVAL
0	419710
1	208157
(blank)	
Grand Total	627867



The ratio of NFLAG_INSURED_ON_APPROVAL is 665527:33162.

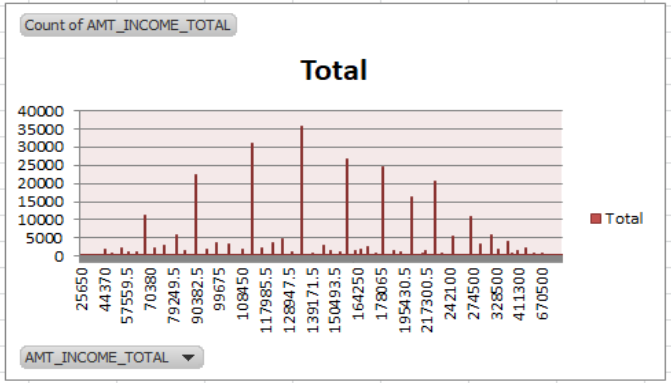
E. Explain the results of univariate ,segmented variate and bivariate analysis

UNIVARIATE ANALYSIS:



Univariate Analysis:

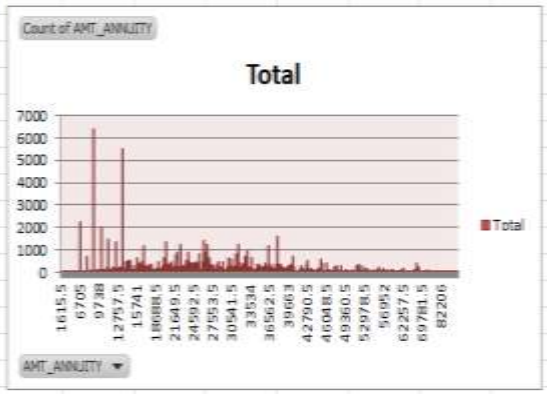
Row Labels	Count of AMT_INCOME_TOTAL
25650	2
26100	3
26460	1
26550	2
27000	66
27585	1
27900	7
27967.5	1
28161	1
28350	11
28404	1
28575	3
28593	1
28723.5	1



A particular chart displays the income total, which varies.

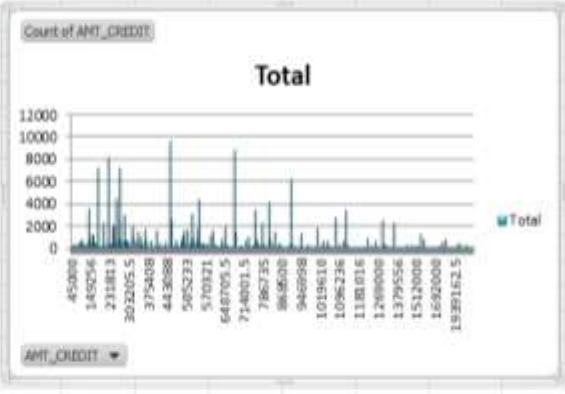
Univariate Analysis:

Row Labels	Count of AMT_ANNUITY
1615.5	1
1980	2
1993.5	1
2052	1
2164.5	2
2173.5	1
2187	2
2205	1
2295	2
2299.5	1
2317.5	2
2412	1
2425.5	1
2511	1
2542.5	3

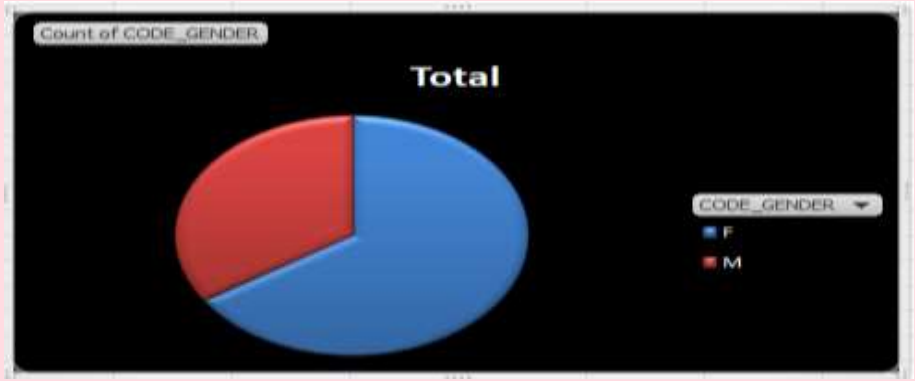


This graph displays an annuity count, which is not frequent.

Row Labels	Count of AMT_CREDIT
45000	230
47970	218
48519	1
49455	19
49500	40
49752	53
50940	393
52128	115
52767	16
53910	56
54000	55
54400.5	5
54504	3
54729	4
54738	3



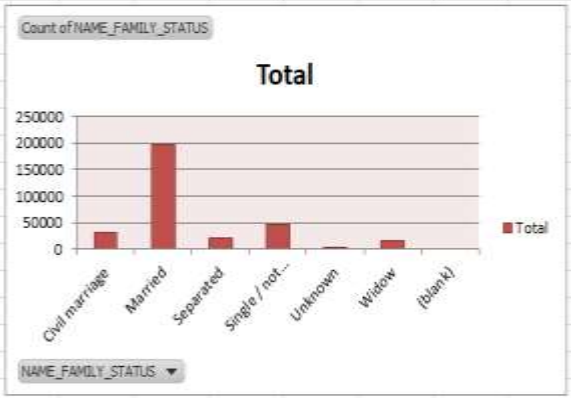
This graph displays the credit count, which is likewise not very common.



➔The pie chart shows there is more female than men .

Univariate Analysis:

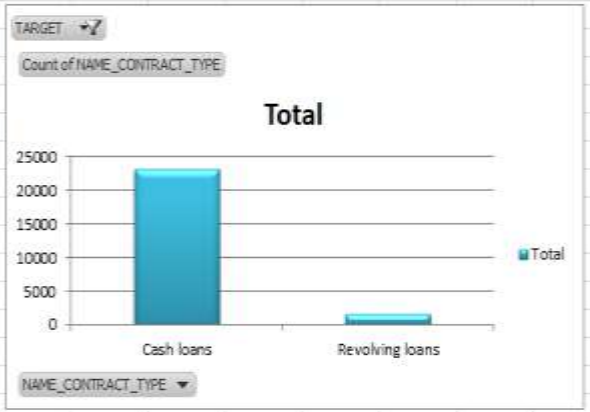
Row Labels	Count of NAME_FAMILY_STATUS
Civil marriage	29775
Married	196432
Separated	19770
Single / not married	45444
Unknown	2
Widow	16088
(blank)	
Grand Total	**307511**



➔The graph demonstrates that married couples borrow more money than any other group.

Segmented Univariate analysis:

TARGET	1
Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	23221
Revolving loans	1604
Grand Total	24825

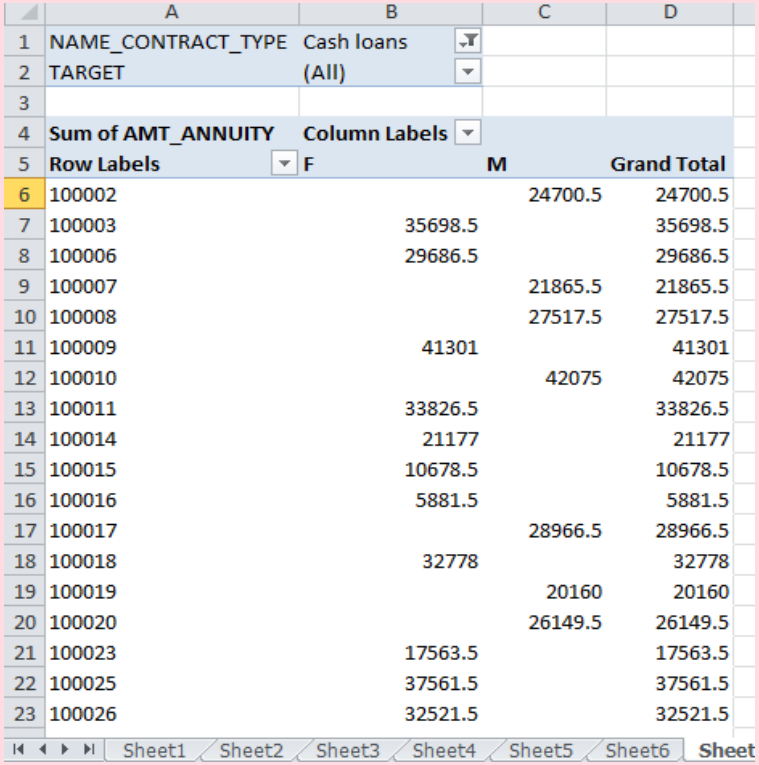


The chart shows when the target value is 1 the cash loan Is 23221 and revolving loans is 1604.

TARGET	0
Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	255011
Revolving loans	27675
Grand Total	282686

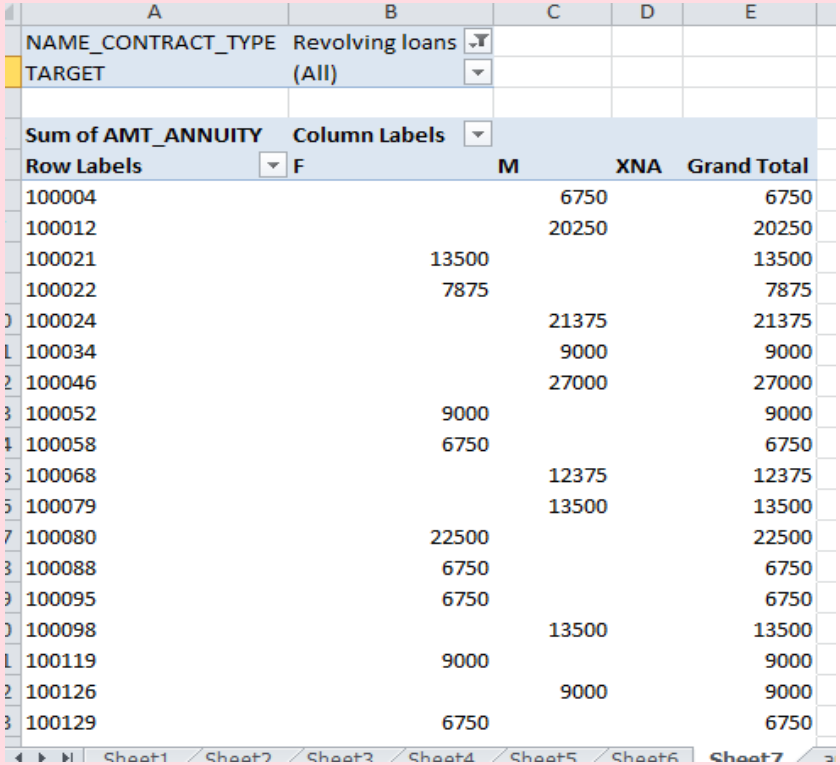
The chart shows when the target value is 0 the cash loan Is 255011 and revolving loans is 27675.

Bivariate Analysis



NAME_CONTRACT_TYPE	Cash loans		
TARGET	(All)		
Sum of AMT_ANNUITY	Column Labels		
Row Labels	F	M	Grand Total
100002		24700.5	24700.5
100003	35698.5		35698.5
100006	29686.5		29686.5
100007		21865.5	21865.5
100008		27517.5	27517.5
100009	41301		41301
100010		42075	42075
100011	33826.5		33826.5
100014	21177		21177
100015	10678.5		10678.5
100016	5881.5		5881.5
100017		28966.5	28966.5
100018	32778		32778
100019		20160	20160
100020		26149.5	26149.5
100023	17563.5		17563.5
100025	37561.5		37561.5
100026	32521.5		32521.5

Sheet1 Sheet2 Sheet3 Sheet4 Sheet5 Sheet6 Sheet



NAME_CONTRACT_TYPE	Revolving loans		
TARGET	(All)		
Sum of AMT_ANNUITY	Column Labels		
Row Labels	F	M	XNA	Grand Total
100004		6750		6750
100012		20250		20250
100021	13500			13500
100022	7875			7875
100024		21375		21375
100034		9000		9000
100046		27000		27000
100052	9000			9000
100058	6750			6750
100068		12375		12375
100079		13500		13500
100080	22500			22500
100088	6750			6750
100095	6750			6750
100098		13500		13500
100119	9000			9000
100126		9000		9000
100129	6750			6750

Sheet1 Sheet2 Sheet3 Sheet4 Sheet5 Sheet6 Sheet7

The data shows that how many male and female has taken Cash loans

The data shows that how many male and female has taken Revolving loans

Bivariate Analysis

TARGET	0			
Count of CODE_GENDER	Column Labels			
Row Labels	F	M	XNA	Grand Total
Businessman	3	7		10
Commercial associate	41551	24705	1	66257
Maternity leave	2	1		3
Pensioner	43018	9362		52380
State servant	15009	5445		20454
Student	7	11		18
Unemployed	11	3		14
Working	88677	54870	3	143550
Grand Total	**188278**	**94404**	**4**	**282686**

The information reveals how many men and women were in the income type when the target figure was 0.

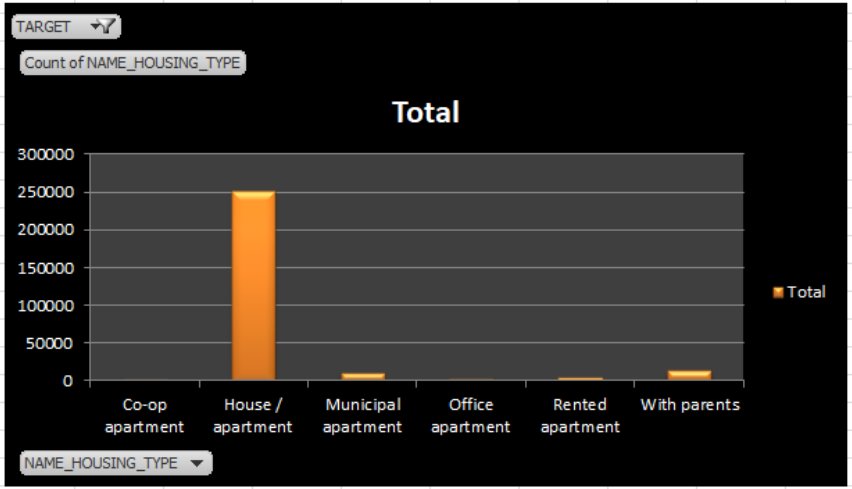
TARGET	1		
Count of CODE_GENDER	Column Labels		
Row Labels	F	M	Grand Total
Commercial associate	2968	2392	5360
Maternity leave	2		2
Pensioner	2243	739	2982
State servant	847	402	1249
Unemployed	6	2	8
Working	8104	7120	15224
Grand Total	**14170**	**10655**	**24825**

The information reveals how many men and women were in the income type when the target figure was 1.

F. The top 10 correlation for the Client with payment difficulties and all other cases (Target variable).

TARGET	0

Row Labels	Count of NAME_HOUSING_TYPE
Co-op apartment	1033
House / apartment	251596
Municipal apartment	10228
Office apartment	2445
Rented apartment	4280
With parents	13104
Grand Total	**282686**



Housing type encounter difficulties

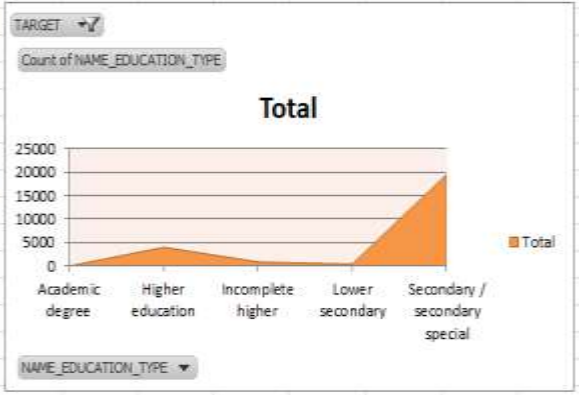
The graph reveals that the 251596 individuals who own a home or apartment have experienced the most payment issues.

TARGET	1
Row Labels	Count of NAME_FAMILY_STATUS
Civil marriage	2961
Married	14850
Separated	1620
Single / not married	4457
Widow	937
Grand Total	24825

Family status facing problem

➔The graph demonstrates that the 14850 married individuals have experienced payment issues.

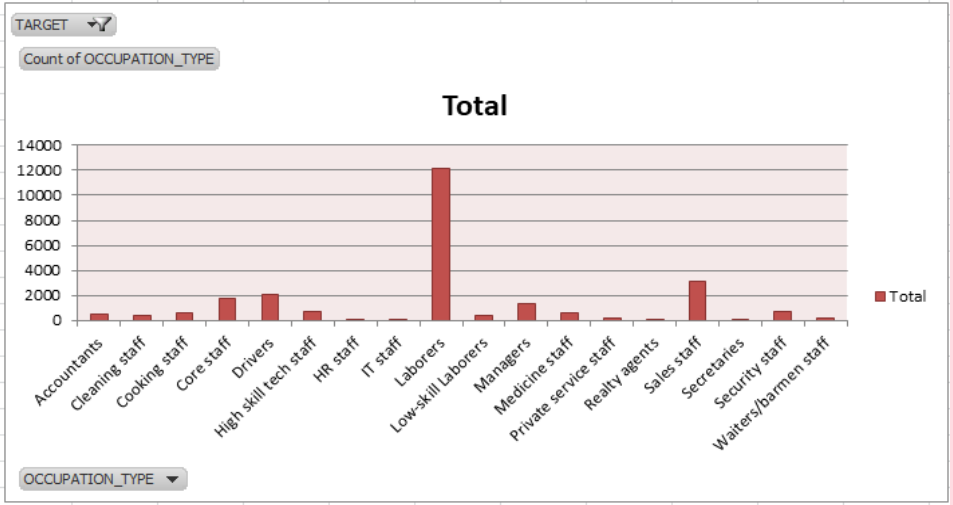
TARGET	1
Row Labels	Count of NAME_EDUCATION_TYPE
Academic degree	3
Higher education	4009
Incomplete higher	872
Lower secondary	417
Secondary / secondary special	19524
Grand Total	24825



Education type facing problem

➔The graph demonstrates that the 19524 students who enrolled in secondary special education experienced financial challenges.

1	TARGET	1
2		
3	**Row Labels**	**Count of OCCUPATION_TYPE**
4	Accountants	474
5	Cleaning staff	447
6	Cooking staff	621
7	Core staff	1738
8	Drivers	2107
9	High skill tech staff	701
10	HR staff	36
11	IT staff	34
12	Laborers	12116
13	Low-skill Laborers	359
14	Managers	1328
15	Medicine staff	572
16	Private service staff	175
17	Realty agents	59
18	Sales staff	3092
19	Secretaries	92
20	Security staff	722
21	Waiters/barmen staff	152
22	**Grand Total**	**24825**



The graphic displays the employment types that have experienced payment issues. As we can see, the 12116 Labourers class has seen the most payment issues, followed by the sales personnel.

INSIGHTS(after completing the analysis):

I now have a better comprehension of the bank loan research portion of this project. It also improves my EDA abilities. The analysis's findings thus indicate that both persons with secondary special education and those who work as labourers will have difficulty repaying their loans. Additionally, the person living in a house or flat will have trouble repaying the loan. Additionally, those who are married will also have trouble repaying their loans.

Impact of Car Features

Analyzing the Impact of Car Features on Price and Profitability

TABLE OF CONTENT

1. PROJECT DESCIPTION
2. APPROACH
3. TECH-STACK USED
4. INSIGHTS
5. RESULTS

PROJECT DESCRIPTION

The automotive industry has been rapidly evolving over the past few decades, with a growing focus on fuel efficiency, environmental sustainability, and technological innovation. With increasing competition among manufacturers and a changing consumer landscape, it has become more important than ever to understand the factors that drive consumer demand for cars.

In recent years, there has been a growing trend towards electric and hybrid vehicles and increased interest in alternative fuel sources such as hydrogen and natural gas. At the same time, traditional gasoline-powered cars remain dominant in the market, with varying fuel types and grades available to consumers.

I assisted the firm boost its competitiveness and profitability by identifying the product attributes on which to concentrate future development efforts using a variety of DATA ANALYSIS methods.

APPROACH

Firstly, understanding of the dataset.
Then eliminating duplicates and blank cells.
Next creating a pivot table, creating some new tables and charts to visualise insights, and using a regression analysis technique to determine the features influencing consumer desires. Visualisation is used to help people comprehend all the aspects that affect the sales and developing dashboard.

TECH-STACK USED

- MICROSOFT EXCEL

DATASET DESCRIPTION

Before cleaning the dataset :

Number of rows = 11915
Number of columns = 16

After cleaning the dataset:

Number of columns after removing the DUPLICATE values = 11199
(715 rows removed)
Removed blank columns as well.

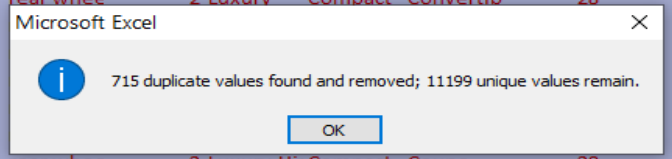
Adding a new column:

Fuel efficiency = AVERAGE OF MSRP)

DATA CLEANING

Before diving into the analysis of the given dataset, it is important to perform thorough data cleaning to ensure accurate and reliable results.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Make	Model	Year	Engine	Engine	Engine	Transm	Driven	Numbe	Market	Vehicle	Vehicle	highwa	city mp	Popula
2	BMW	1 Series M	2011	premium	335	6	MANUAL	rear whee	2	Factory Tu	Compact	Coupe	26	19	39
3	BMW	1 Series	2011	premium	300	6	MANUAL	rear whee	2	Luxury,Pe	Compact	Convertib	28	19	39
4	BMW	1 Series	2011	premium	300	6	MANUAL	rear whee	2	Luxury,Hi	Compact	Coupe	28	20	39
5	BMW	1 Series	2011	premium	230	6	MANUAL	rear whee	2	Luxury,Pe	Compact	Coupe	28	18	39
6	BMW	1 Series	2011	premium	230	6	MANUAL	rear whee	2	Luxury	Compact	Convertib	28	18	39
7	BMW	1 Series	2012	premium	230	6	MANUAL							18	39
8	BMW	1 Series	2012	premium	300	6	MANUAL							17	39
9	BMW	1 Series	2012	premium	300	6	MANUAL							20	39
10	BMW	1 Series	2012	premium	230	6	MANUAL							18	39
11	BMW	1 Series	2013	premium	230	6	MANUAL							18	39
12	BMW	1 Series	2013	premium	300	6	MANUAL	rear whee	2	Luxury,Hi	Compact	Coupe	28	20	39
13	BMW	1 Series	2013	premium	230	6	MANUAL	rear whee	2	Luxury,Pe	Compact	Coupe	28	19	39
14	BMW	1 Series	2013	premium	300	6	MANUAL	rear whee	2	Luxury,Pe	Compact	Convertib	28	19	39
15	BMW	1 Series	2013	premium	230	6	MANUAL	rear whee	2	Luxury	Compact	Convertib	28	19	39
16	BMW	1 Series	2013	premium	320	6	MANUAL	rear whee	2	Luxury,Hi	Compact	Convertib	25	18	39
17	BMW	1 Series	2013	premium	320	6	MANUAL	rear whee	2	Luxury,Hi	Compact	Coupe	28	20	39
18	Audi	100	1992	regular un	172	6	MANUAL	front whe	4	Luxury	Midsize	Sedan	24	17	31
19	Audi	100	1992	regular un	172	6	AUTOMAT	all wheel	4	Luxury	Midsize	Wagon	20	16	31
20	Audi	100	1992	regular un	172	6	MANUAL	all wheel	4	Luxury	Midsize	Sedan	21	16	31
21	Audi	100	1993	regular un	172	6	MANUAL	front whe	4	Luxury	Midsize	Sedan	24	17	31
22	Audi	100	1993	regular un	172	6	AUTOMAT	all wheel	4	Luxury	Midsize	Wagon	20	16	31
23	Audi	100	1993	regular un	172	6	MANUAL	all wheel	4	Luxury	Midsize	Sedan	21	16	31



THIS TASK HAS BEEN PERFORMED IN TWO PARTS

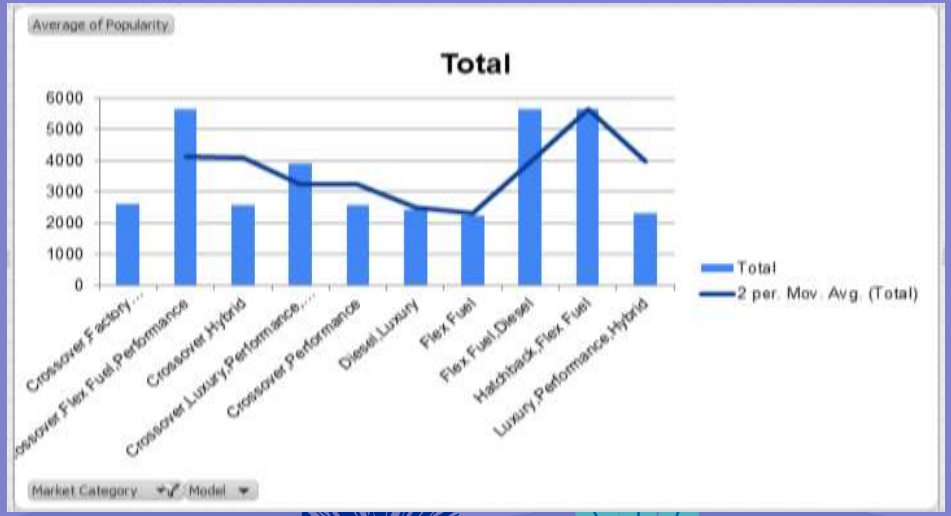
1. ANALYSIS
2. DASHBOARDS

1. ANALYSIS:

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

- Task 1.A: Create a pivot table that shows the number of car models in each market category and their corresponding popularity scores.
- Task 1.B: Create a combo chart that visualizes the relationship between market category and popularity.

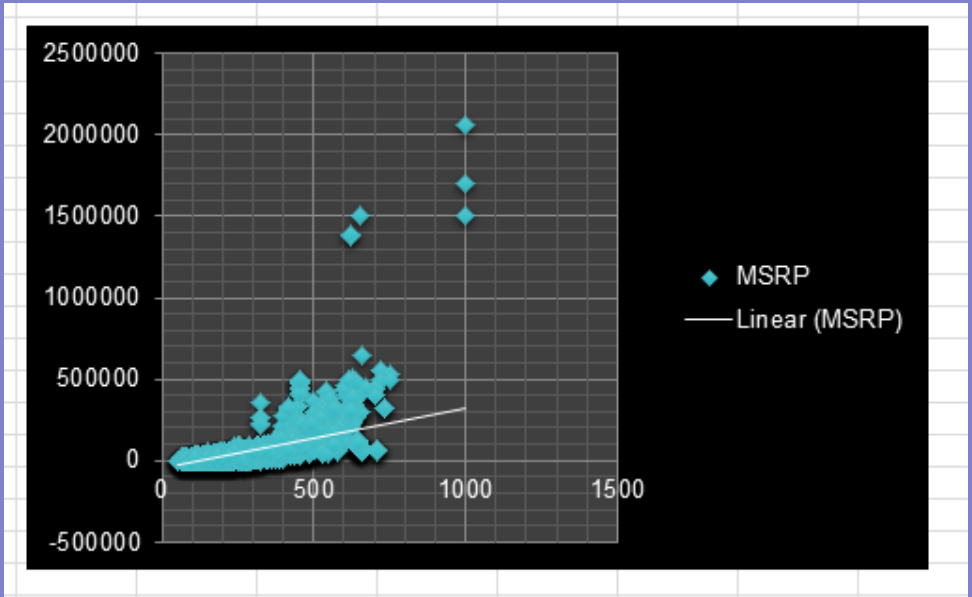
Row Labels	Average of Popularity
Crossover,Factory Tuner,Luxury,Performance	2607.4
Crossover,Flex Fuel,Performance	5657
Crossover,Hybrid	2563.380952
Crossover,Luxury,Performance,Hybrid	3916
Crossover,Performance	2585.956522
Diesel,Luxury	2416.106383
Flex Fuel	2225.71345
Flex Fuel,Diesel	5657
Hatchback,Flex Fuel	5657
Luxury,Performance,Hybrid	2333.181818
Grand Total	2370.964151



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

- Task 2: Create a scatter chart that plots engine power on the x-axis and price on the y-axis. Add a trendline to the chart to visualize the relationship between these Variable.

	A	B
1	Engine HP	MSRP
2	335	46135
3	300	40650
4	300	36350
5	230	29450
6	230	34500
7	230	31200
8	300	44100
9	300	39300
10	230	36900
11	230	37200
12	300	39600
13	230	31500
14	300	44400
15	230	37200
16	320	48250
17	320	43550
18	172	2000
19	172	2000
20	172	2000
21	172	2000
22	172	2000
23	172	2000
24	172	2000
25	172	2000
26	172	2000



➔The relationship between a car's engine power and its price is determined.

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

- Task 3: Use regression analysis to identify the variables that have the strongest relationship with a car's price. Then create a bar chart that shows the coefficient values for each variable to visualize their relative importance.

Number of doors	engine cylinder	Engine HP	highway MPG	city mpg
2	6	335	26	19
2	6	300	28	19
2	6	300	28	20
2	6	230	28	18
2	6	230	28	18
2	6	230	28	18
2	6	300	26	17
2	6	300	28	20
2	6	230	28	18
2	6	230	27	18
2	6	300	28	20
2	6	230	28	19
2	6	300	28	19
2	6	230	28	19
2	6	320	25	18
2	6	320	28	20
4	6	172	24	17
4	6	172	20	16
4	6	172	21	16
4	6	172	24	17
4	6	172	20	16
4	6	172	21	16
4	6	172	21	16
4	6	172	22	16
4	6	172	22	17
4	6	172	22	16

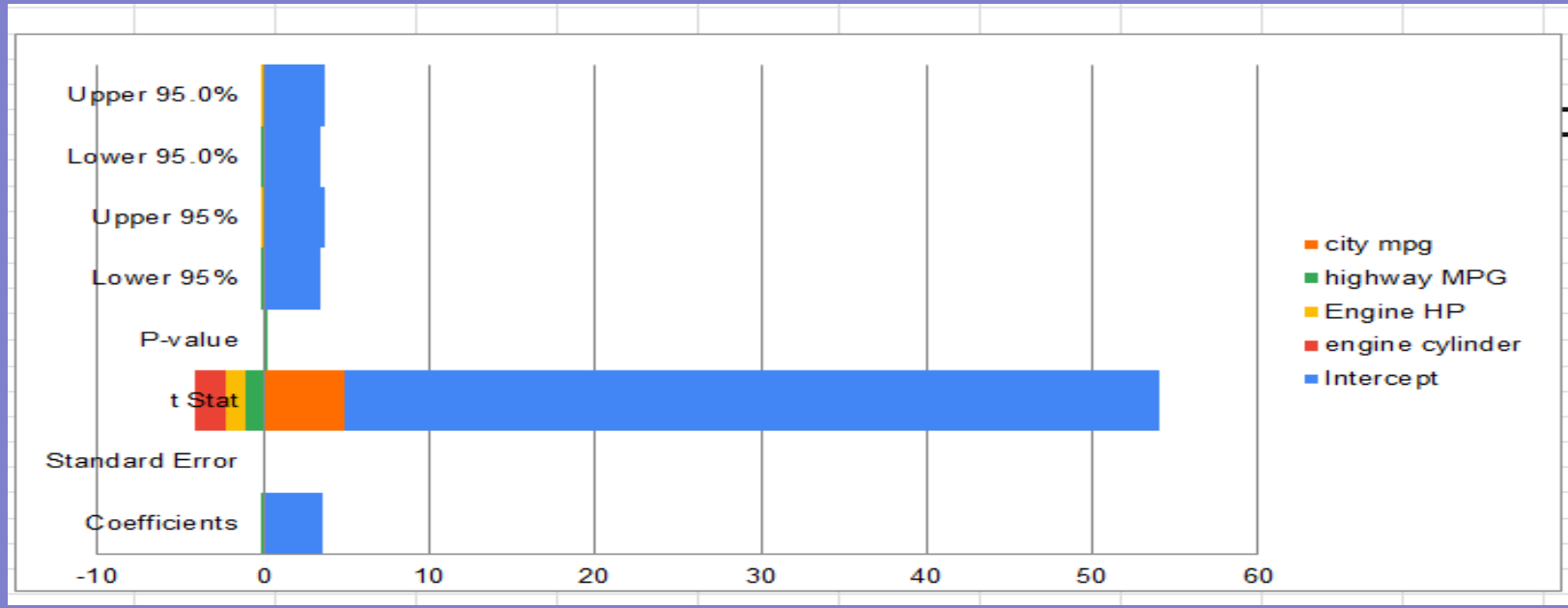
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.16002
R Square	0.025606
Adjusted F	0.025255
Standard E	0.863487
Observatio	11100

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	4	217.3958	54.34895	72.89187	4.57E-61
Residual	11095	8272.55	0.745611		
Total	11099	8489.946			

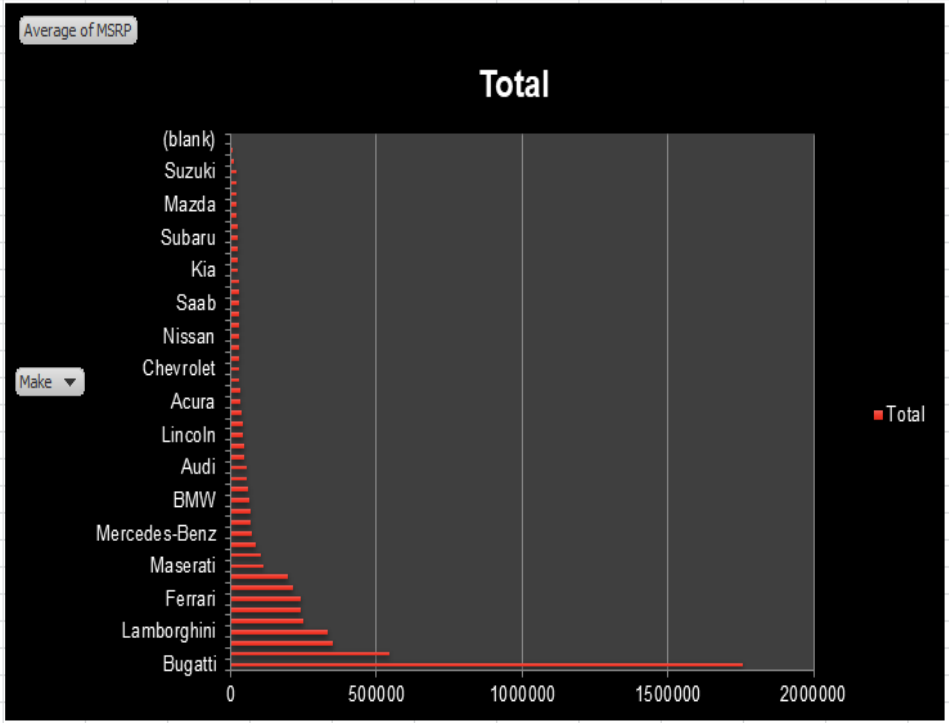
	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	3.564343	0.065936	54.05787	0	3.435097	3.693588	3.435097	3.693588
engine cyli	-0.0371	0.008825	-4.20419	2.64E-05	-0.0544	-0.0198	-0.0544	-0.0198
Engine HP	-0.00028	0.000122	-2.30201	0.021353	-0.00052	-4.2E-05	-0.00052	-4.2E-05
highway M	-0.00229	0.002092	-1.0954	0.273366	-0.00639	0.001809	-0.00639	0.001809
city mpg	0.011854	0.002402	4.934239	8.16E-07	0.007145	0.016563	0.007145	0.016563


Upper 95.0%
Lower 95.0%
Upper 95%
Lower 95%
P-value
t Stat
Standard Error
Coefficients
-10
0
10
20
30
40
50
60
city mpg
highway MPG
Engine HP
engine cylinder
Intercept

Insight Required: How does the average price of a car vary across different manufacturers?

- Task 4.A: Create a pivot table that shows the average price of cars for each manufacturer.
- Task 4.B: Create a bar chart or a horizontal stacked bar chart that visualizes the relationship between manufacturer and average price

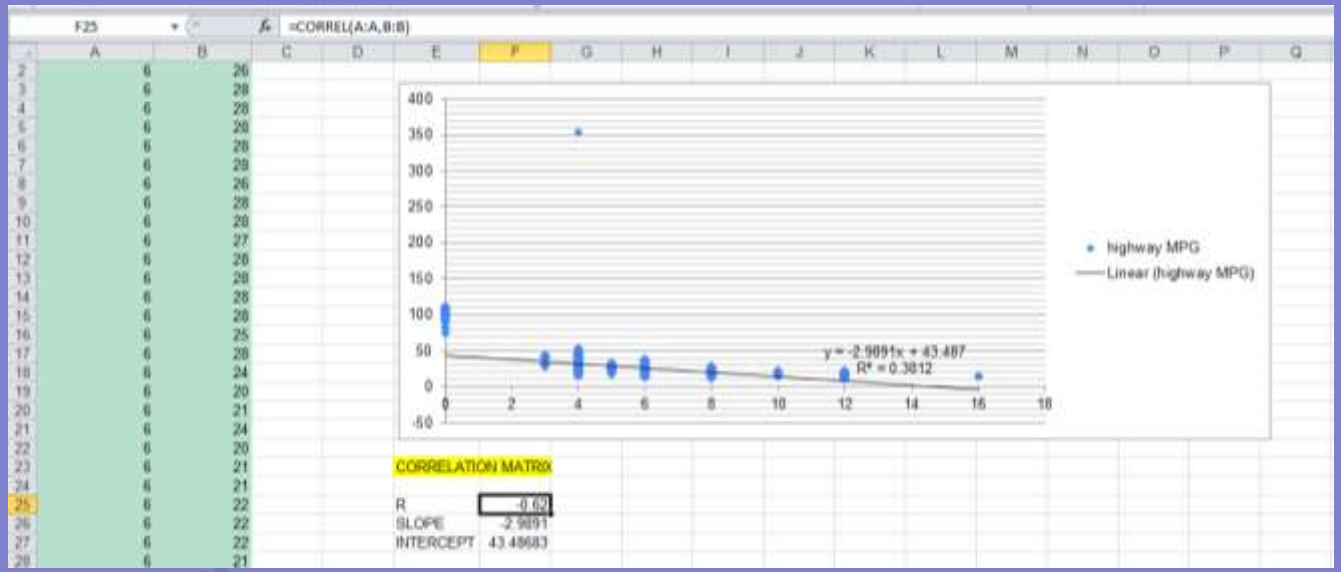
Row Labels	Average of MSRP
Bugatti	1757223.667
Maybach	546221.875
Rolls-Royce	351130.6452
Lamborghini	331567.3077
Bentley	247169.3243
McLaren	239805
Ferrari	238218.8406
Spyker	214990
Aston Martin	198123.4615
Maserati	113684.4909
Porsche	101622.3971
Tesla	85255.55556
Mercedes-Benz	72069.52786
Lotus	68377.14286
Land Rover	68067.08633
BMW	62162.55864
Alfa Romeo	61600
Cadillac	56368.26515
Audi	54574.1215
Lexus	47549.06931
Genesis	46616.66667
Lincoln	43860.825
Infiniti	42640.27134
HUMMER	36464.41176



➔Bugatti has the highest average MSRP.

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

- Task 5.A: Create a scatter plot with the number of cylinders on the x-axis and highway MPG on the y-axis. Then create a trendline on the scatter plot to visually estimate the slope of the relationship and assess its significance.
- Task 5.B: Calculate the correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.



➔Strong Irreversible relation was found between Fuel Efficiency and number of Cylinders.

2. DASHBOARDS:

Task 2.1: How does the distribution of car prices vary by brand and body style?

● Hints: Stacked column chart to show the distribution of car prices by brand and body style. Use filters and slicers to make the chart interactive. Calculate the total MSRP for each brand and body style using SUMIF or Pivot Tables.

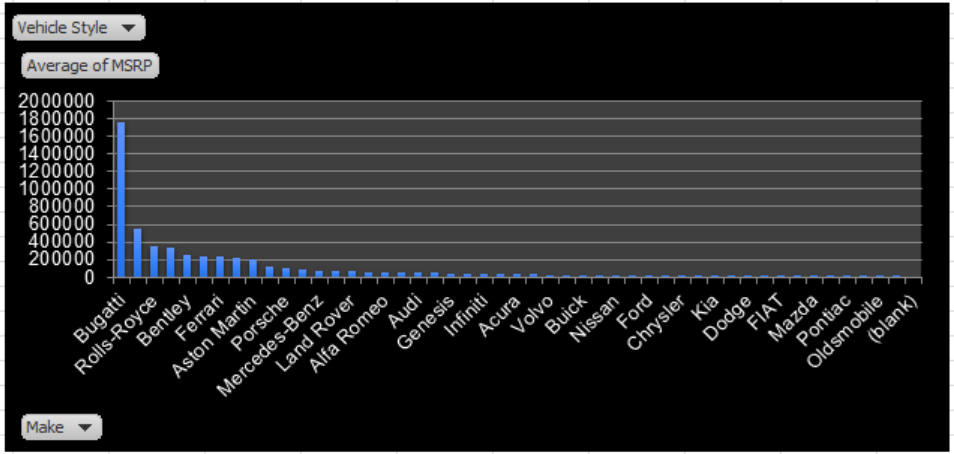
Vehicle Style	(All)
Row Labels	Sum of MSRP
Acura	8631522
Alfa Romeo	308000
Aston Martin	18029235
Audi	17518293
Bentley	18290530
BMW	20140669
Bugatti	5271671
Buick	5516496
Cadillac	22321833
Chevrolet	31487928
Chrysler	4997194
Dodge	13149377
Ferrari	16437100
FIAT	1405555
Ford	23521829
Genesis	139850
GMC	15638049
Honda	11541579
HUMMER	619895
Hyundai	6455902
Infiniti	13986009
Kia	5842650
Lamborghini	17241500
Land Rover	9461325

We can plainly see how the price distribution of cars varies depending on the brand and body style.

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

- Hints: Clustered column chart to compare the average MSRPs across different car brands and body styles. Calculate the average MSRP for each brand and body style using AVERAGEIF or Pivot Tables.

1	Vehicle Style	(All)
2		
3	**Row Labels**	**Average of MSRP**
4	Bugatti	1757223.667
5	Maybach	546221.875
6	Rolls-Royce	351130.6452
7	Lamborghini	331567.3077
8	Bentley	247169.3243
9	McLaren	239805
10	Ferrari	238218.8406
11	Spyker	214990
12	Aston Martin	198123.4615
13	Maserati	113684.4909
14	Porsche	101622.3971
15	Tesla	85255.55556
16	Mercedes-Benz	72069.52786
17	Lotus	68377.14286
18	Land Rover	68067.08633
19	BMW	62162.55864
20	Alfa Romeo	61600
21	Cadillac	56368.26515
22	Audi	54574.1215
23	Lexus	47549.06931
24	Genesis	46616.66667
25	Lincoln	43860.825
26	Infiniti	42640.27134

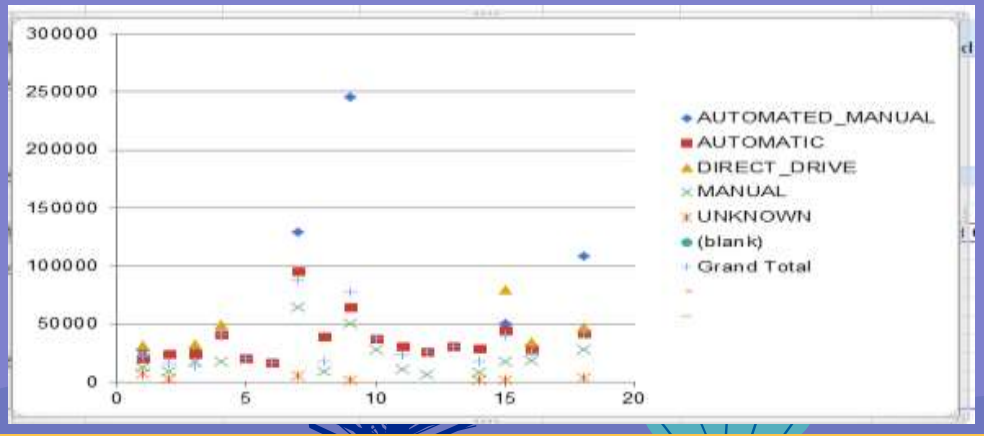


While Plymouth has the lowest average MSRP, Bugatti has the highest.

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

- Hints: Scatter plot chart to visualize the relationship between MSRP and transmission type, with different symbols for each body style. Calculate the average MSRP for each combination of transmission type and body style using AVERAGEIFS or Pivot Tables.

Column Labels												
4dr Hatchback	2dr Hatchback	2dr SUV	4dr SUV	Cargo Minivan	Cargo Van	Convertible	Convertible SUV	Coupe	Crew Cab Pickup	Extended Cab Pickup	Passenger Minivan	Passenger V
29347.04545	27470.41667		40451.15385			129082.2339		245977.4252				
23888.73529	20784.09901	24153.60606	41638.26534	20315.59322	17019.29762	95153.3131	38925.5	64523.41955	37718.95307	30711.45251	26589.50919	30578.06
32799.72973	31800		49800									
17500.36364	12840.65556	9173.018519	17422.08791			64794.34437	9594.8	50901.4973	28233.10811	11553.29707	6510	
	7361.5	2371				5783.5		2000				
22416.46757	16177.74029	14306.54945	40730.27362	20315.59322	17019.29762	88216.79217	17975	77595.28766	37183.11145	23041.77219	26176.56298	30578.06

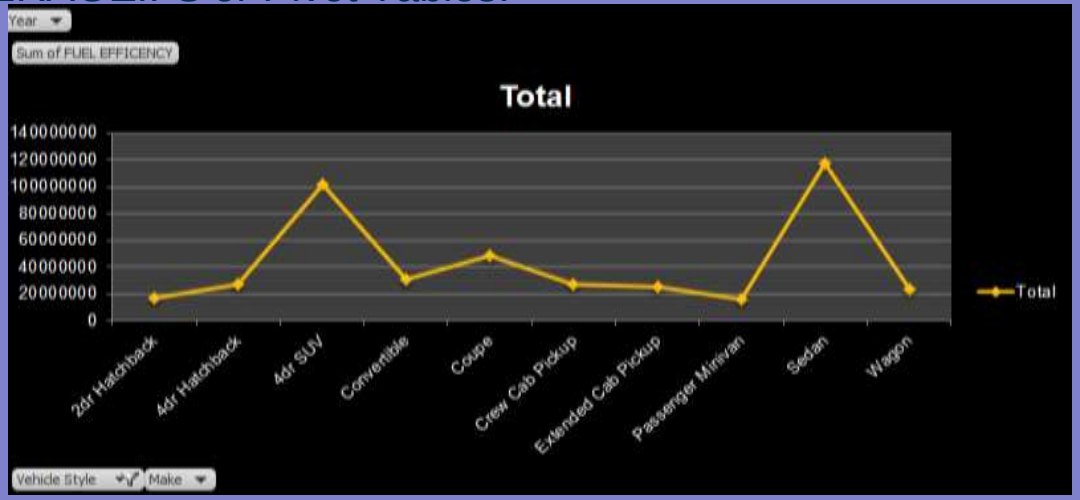


➔Prices for various body types vary depending on the transmission type.

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

- Hints: Line chart to show the trend of fuel efficiency (MPG) over time for each body style. Calculate the average MPG for each combination of body style and model year using AVERAGEIFS or Pivot Tables.

Year	(All)
Row Labels	Sum of FUEL EFFICENCY
2dr Hatchback	17097732.26
4dr Hatchback	27171188.4
4dr SUV	101170447.9
Convertible	30988167.26
Coupe	48607510.16
Crew Cab Pickup	27139037.53
Extended Cab Pickup	25200640.35
Passenger Minivan	16289600.93
Sedan	116923225.8
Wagon	23592318.49
Acura	293665.0528
Audi	1038580.441
BMW	252890.6119
Buick	157933.3705
Cadillac	1019050.365
Chevrolet	766037.8867
Chrysler	783472.9122
Dodge	1314140.592
FIAT	547951.8514
Ford	2221802.26
Kia	1498288.046
Lexus	40478.63671
Lincoln	248581.6532
Mazda	81314.62642

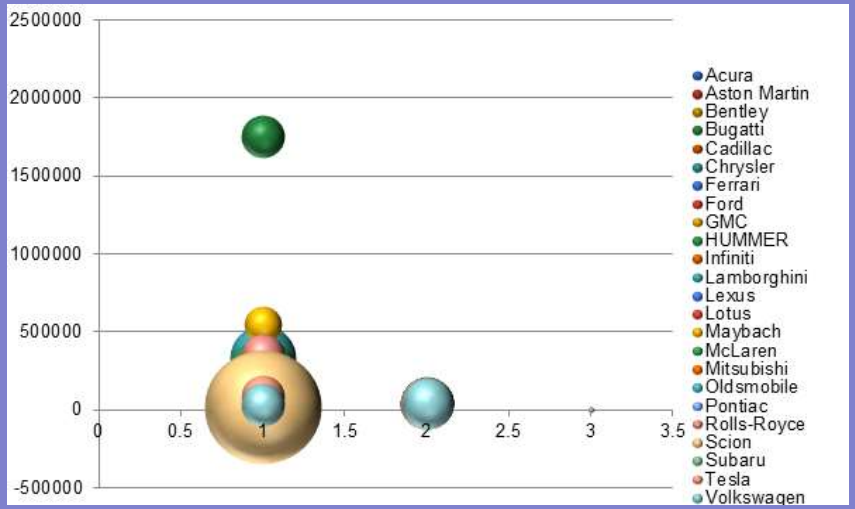


➔For different body types and brands, fuel efficiency varies.The highest MPG average is for sedans.

Task 5: How does the car's horsepower, MPG, and price vary across different Brands?

- Hints: Bubble chart to visualize the relationship between horsepower, MPG, and price across different car brands. Assign different colors to each brand and label the bubbles with the car model name. Calculate the average horsepower, MPG, and MSRP for each car brand using AVERAGEIFS or Pivot Tables.

Row Labels	Average of MSRP	Average of FUEL EFFICENCY	Average of Engine HP
Acura	35087.4878	41601.06049	244.9634146
Alfa Romeo	61600	42092.58667	237
Aston Martin	198123.4615	45581.1523	483.7582418
Audi	54574.1215	41042.43151	280
Bentley	247169.3243	41243.85662	533.8513514
BMW	62162.55864	43052.81028	329.6203704
Bugatti	1757223.667	47711.20256	1001
Buick	29034.18947	41252.93854	220.0105263
Cadillac	56368.26515	41149.04057	332.7954545
Chevrolet	29074.72576	41437.60235	249.4837512
Chrysler	26722.96257	41522.81746	229.1390374
Dodge	24857.04537	40653.63557	254.3534972
Ferrari	238218.8406	41638.07044	511.9565217
FIAT	22670.24194	42125.07798	143.559322
Ford	28511.30788	41194.52431	249.6921182
Genesis	46616.66667	41326.47039	347.3333333
GMC	32444.08506	41934.19857	267.6452282
Honda	26655.14781	40770.42791	196.7726218
HUMMER	36464.41176	39966.99058	261.2352941
Hyundai	24926.26255	41981.23506	205.2046332
Infiniti	42640.27134	40873.39176	310.6768293
Kia	25513.75546	39984.51274	207.5580357
Lamborghini	331567.3077	40867.54993	614.0769231
Land Rover	68067.08633	40391.16501	322.5179856
Lexus	47549.06931	40244.39156	277.4158416
Lincoln	43860.825	41169.85164	286.125



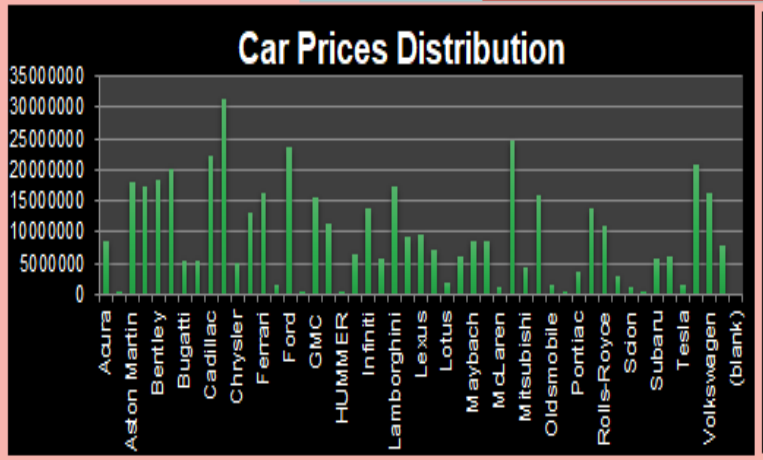
➔Various brands have various horsepower, efficiency, and price ranges.

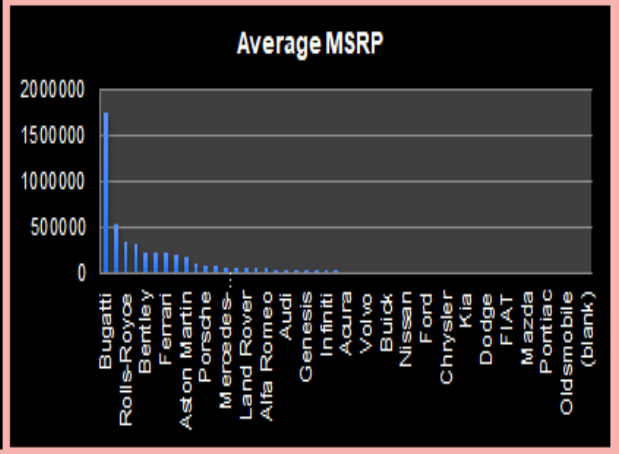
INSIGHTS:

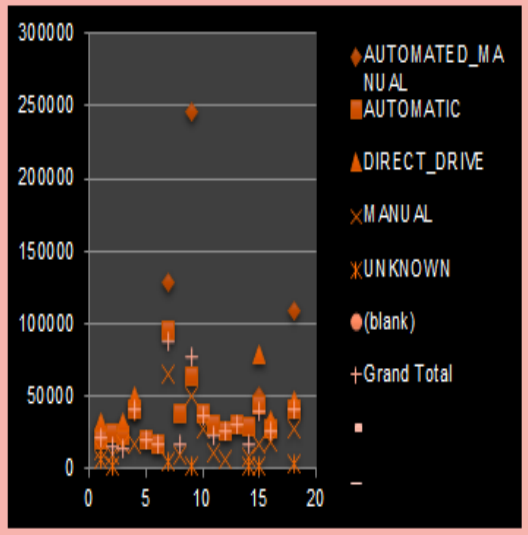
- The relationship between a car's engine power and its price is determined.
- Bugatti has the highest average MSRP
- Strong Irreversible relation was found between Fuel Efficiency and number of Cylinders.
- the price distribution of cars varies depending on the brand and body style
- While Plymouth has the lowest average MSRP, Bugatti has the highest.
- Prices for various body types vary depending on the transmission type.
- For different body types and brands, fuel efficiency varies. The highest MPG average is for sedans.
- Various brands have various horsepower, efficiency, and price ranges

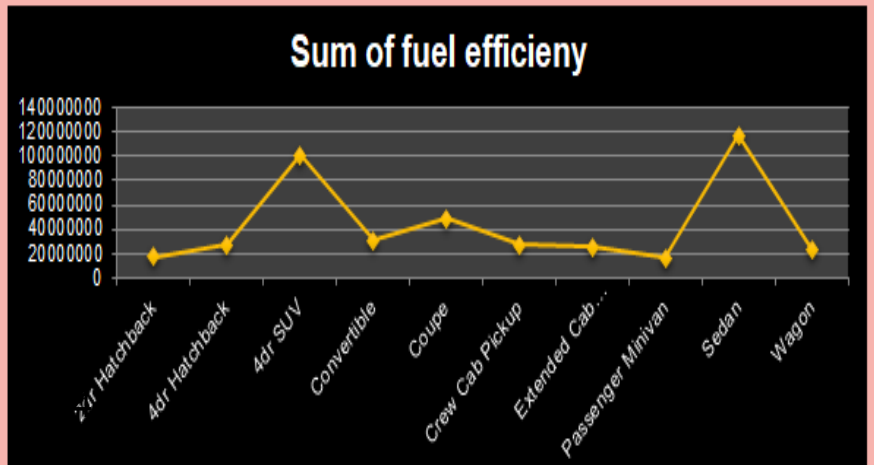
RESULTS:

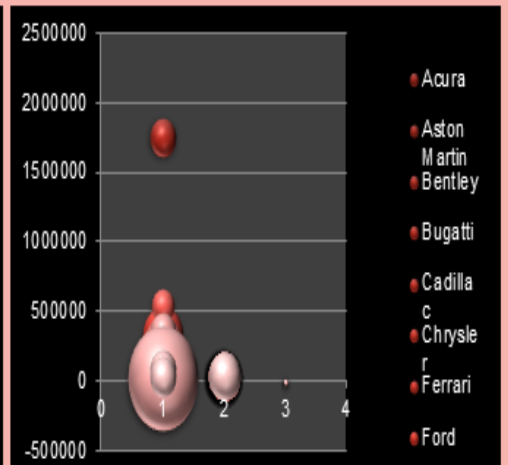
CAR FEATURE ANALYSIS DASHBOARD

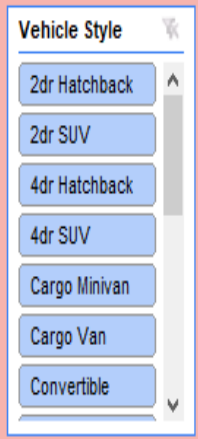


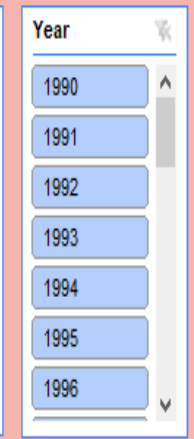












ABC Call Volume Trend Analysis

Final Project-4

PROJECT DESCRIPTION

For the final project we are provided with a dataset of a Customer Experience (CX) Inbound calling team for 23 days. 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).

A customer experience (CX) team consists of professionals who analyze customer feedback and data, and share insights with the rest of the organization. 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.

Business Understanding:

Advertising is a way of marketing your business in order to increase sales or make your audience aware of your products or services. Until a customer deals with you directly and actually buys your products or services, your advertising may help to form their first impressions of your business. Target audience for businesses could be local, regional, national or international or a mixture. So they use different ways for advertisement. Some of the types of advertisement are: Internet/online directories, Trade and technical press, Radio, Cinema, Outdoor advertising, National papers, magazines and TV. Advertising business is very competitive as a lot of players bid a lot of money in a single segment of business to target the same audience. Here comes the analytical skills of the company to target those audiences from those types of media platforms where they convert them to their customers at a low cost.

APPROACH

01 Check for consistency in the data.

02 In Excel, creating tables if required and add more columns as necessary.

03 Make charts and pivot tables for better understanding.

04 Lastly, drawing insights from the data.

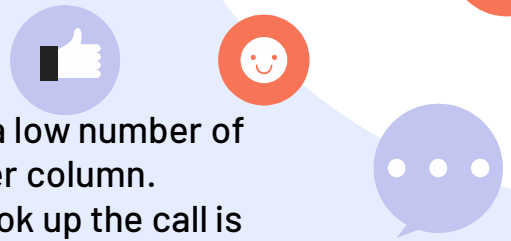
TECH-STACK USED

- MS-EXCEL
- MS-POWERPOINT
- LOOM

24/7

ABOUT THE DATASET

There are 13 columns and 117989 rows.
Name of the agent, called status, and wrapped
make up the category column.
The agent id is contained in the Agent Id column, which also has a low number of null values.The customer's phone number is listed in the customer column.
The number of seconds the customer waited before the agent took up the call is the queue time.For time, Date_&_Time, Time, and Time_Bucket are utilised.
The length of the call is contained in Call_seconds and Duration.

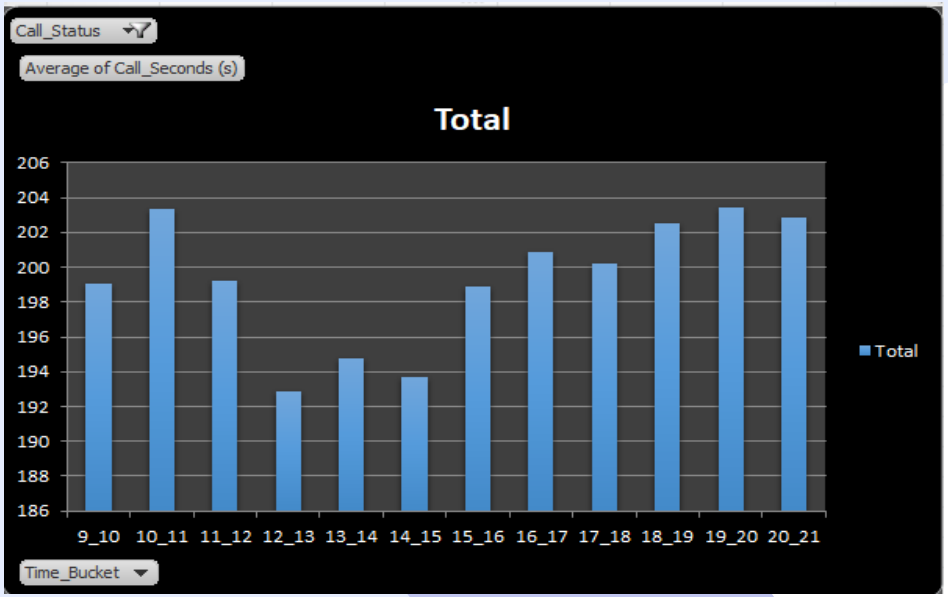


CLEANING THE DATASET

Since ringing only has one variable and IVR duration is irrelevant for our research, the columns Ringing and IVR duration are not used.

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

1	Call_Status	answered
2		
3	**Row Labels**	**Average of Call_Seconds (s)**
4	9_10	199.0691057
5	10_11	203.3310302
6	11_12	199.2550234
7	12_13	192.8887829
8	13_14	194.7401744
9	14_15	193.6770755
10	15_16	198.8889175
11	16_17	200.8681864
12	17_18	200.2487831
13	18_19	202.5509677
14	19_20	203.4060725
15	20_21	202.845993
16	**Grand Total**	**198.6227745**

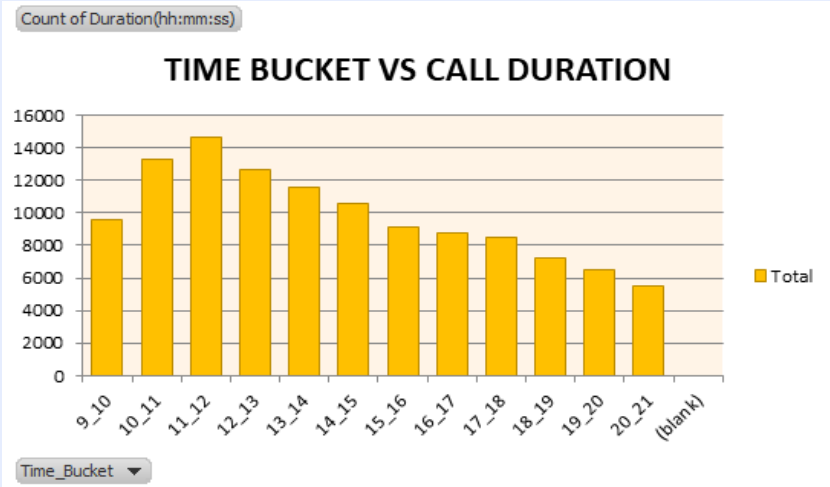


According to the graph above, the average call time is 198.6 seconds, while the longest call length was buck 19_20 and second largest is 10_11.

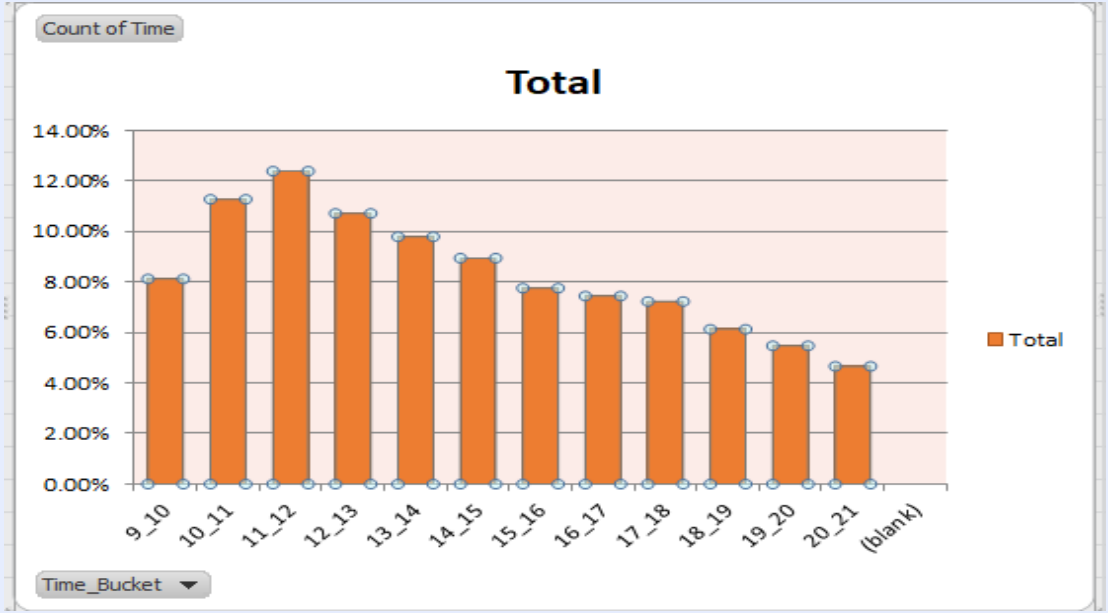
B.

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

Row Labels	Count of Duration(hh:mm:ss)
9_10	9588
10_11	13313
11_12	14626
12_13	12652
13_14	11561
14_15	10561
15_16	9159
16_17	8788
17_18	8534
18_19	7238
19_20	6463
20_21	5505
(blank)	
Grand Total	**117988**



3	Row Labels	Count of Time
4	9_10	8.13%
5	10_11	11.28%
6	11_12	12.40%
7	12_13	10.72%
8	13_14	9.80%
9	14_15	8.95%
10	15_16	7.76%
11	16_17	7.45%
12	17_18	7.23%
13	18_19	6.13%
14	19_20	5.48%
15	20_21	4.67%
16	(blank)	0.00%
17	**Grand Total**	**100.00%**



From above both graph we can see maximum no of calls and percentage of time is on time bucket 11_12

Assumption

An agent works 6 days a week and takes an average of 4 days off per month as unplanned leave.
The total working hours for an agent is 9 hours, with 1.5 hours being dedicated to lunch and snacks in the office.
On average, the agent is occupied for 60% of their 7.5 actual working hours in call with customers or users.
A month is comprised of 30 days.

1	Agent working hour	9			
2	Agent on-floor work hour	7.5			
3	Days of agent on floor	5			6 working days
4	Total time spent on call	4.5			
5					
6					
7					Out of 28 days, 24 days
8					Out of 28 days,20 days unplanned leave
9					

C.

As you can see current abandon rate is approximately 30%. Propose a manpower plan required during each time bucket [between 9am to 9pm] to reduce the abandon rate to 10%. (i.e. You have to calculate minimum number of agents required in each time bucket so that at least 90 calls should be answered out of 100.)

Count of Duration(hh:mm:ss)	Column Labels				
Row Labels	abandon	answered	transfer	(blank)	Grand Total
(blank)					
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

The data was analyzed by placing Date & Time in the Rows, Call Status in the Columns, and using the count of Call Duration in the Values section.

The average of abandon, answered, and transfer calls was calculated using an average excel formula.

	A	B
1	Time taken on an average to answer a call	198.6 sec
2		
3	Time requirement to answer 90% of the calls	254.7
4		
5	Total working person required per day	57
6		
7		
8	Daily Call volume (9am - 9pm)	5130
9	If we provide support in night (9pm - 9am)	1539
10		
11	Additional hours required	76.41
12		
13	Additional HC	17
14		
15	Total HC	74

	A	B	C
1	Time Bucket	Count of Time	Req.Agents
2	9_10	8.1%	5
3	10_11	11.3%	6
4	11_12	12.4%	7
5	12_13	10.7%	6
6	13_14	9.8%	6
7	14_15	9.0%	5
8	15_16	7.8%	4
9	16_17	7.4%	4
10	17_18	7.2%	4
11	18_19	6.1%	3
12	19_20	5.5%	3
13	20_21	4.7%	3
14	Grand Total	100.0%	57

- On average, it takes 198.6 seconds to answer a call.254.7 hours are needed to handle 90% of the calls.
- There must be 57 workers overall per day.5130 calls are made every day from 9 am to 9 night.
- If we offer support from 9 p.m. to 9 a.m., we can assist 1539 additional clients.
- 76.41 further hours are needed.
- Extra HC is necessary 17.HC in all, 74.

D. **Let's say customers also call this ABC insurance company in night but didn't get answer as there are no age to answer, this creates a bad customer experience for this Insurance company. Suppose every 100 calls th customer made during 9 Am to 9 Pm, customer also made 30 calls in night between interval [9 Pm to 9 Am] distribution of those 30 calls are as follows:**

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

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

Night time slot	Calls per slot	76.41135	Agents needed	Time distribution
21_22	3	7.641135	13	10%
22-23	3	7.641135	13	10%
23_24	2	5.09409	8	7%
00_01	2	5.09409	8	7%
01_02	1	2.547045	4	3%
02_03	1	2.547045	4	3%
03_04	1	2.547045	4	3%
04_05	1	2.547045	4	3%
05_06	3	7.641135	13	10%
06_07	4	10.18818	17	13%
07_08	4	10.18818	17	13%
08_09	5	12.735225	21	17%
Total	30	76.41135	126	100%

- First I calculated the Time Distribution by dividing each calls distribution by total calls i.e. 30.
- The number of agents required for each time bucket is calculated by 17 * Time Distribution.

No. of person required (total extra hours req./ working time for one agent)	17

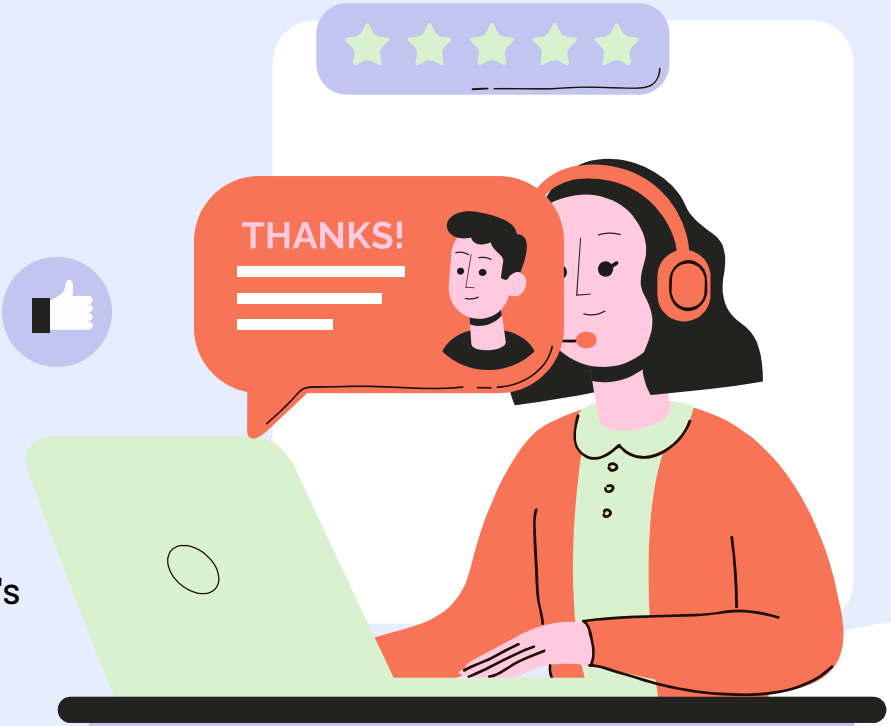
INSIGHTS

By taking into account the number of calls received at various times of the day, the business may optimise their customer care operations.

I concluded from my analysis of the data that call volume is lowest in the late afternoon.

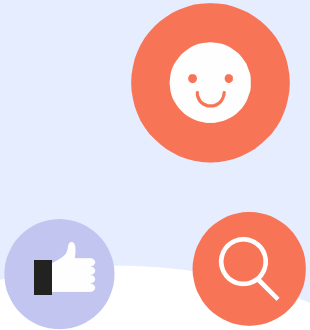
The corporation could decrease the number of agents available at such time in order to adapt personnel accordingly.

Calls can also be addressed promptly by employing 17 customer service representatives for the night shift and switching some of the present day employees to the night shift.
The business might divide its employees into three teams with various work schedules to assure 24-hour coverage.It's crucial to keep in mind that the results might have been impacted by outliers in the data, and that deleting those outliers might produce different results.



RESULT

I learned more about the function of an analyst in enhancing a company's customer service division through this project.
I learned how businesses use a variety of strategies to ensure that their customers are completely satisfied.



CONCLUSION

I successfully completed the projects that involved designing and implementing an SQL database for a retail company. They demonstrated the proficiency in constructing complex queries to extract specific information, generating comprehensive reports, and optimizing data retrieval processes. As part of the hands-on experience, undertook many Excel-based projects to analyze sales data for a multinational corporation. Also, utilized advanced Excel functions, such as pivot tables, data validation, and conditional formatting, to perform data cleaning, identify trends, and create insightful visualizations.

APPENDIX

- Excel working file link of ABC Volume Trend Analysis Project
https://docs.google.com/spreadsheets/d/1Of_1OdK9qV-rEvOoSvu1ojLvHyl1G0Dp/edit?usp=share_link&ouid=112479189862540444718&rtpof=true&sd=true

- Excel working file link of Impact of Car Features:
https://docs.google.com/spreadsheets/d/1IVeEzg7nxwzVDSeQSHPilQLyBgqW8ZQU/edit?usp=share_link&ouid=112479189862540444718&rtpof=true&sd=true