

|  |
| --- |
| Data Analysis Project Portfolio. |
|  |
| February 13  TRAINITY  Authored by: Lohith Kumar A |

Professional Background:-

* I am a Bachelor Of Engineering(B.E) graduate from ECE branch. I have secured aggregate 7.5 CGPA .
* My skills includes SQL, Data Analysis, Python, Microsoft Excel, etc.
* I have worked on several companies virtual experience such as TATA,KPMG,ACCENTURE..
* I have also published a research paper titled- "Detection of spatial objects using remotely sensed images" in a journal .
* I have attended various workshops bases on Data analytics.
* As I am a fresher it would be great to experience the real challenges of the corporate world and understand how things work.
* Being a fresher, I think I am very flexible and adaptive to learn new things.
* I have theoretical knowledge. But I am waiting to use my theoretical knowledge in a practical way. And I believe by putting significant efforts I will learn.

Table Of Contents:-

Professional Background

Table of Contents

Projects

1. Data Analytics Process.
2. Instagram User Analytics.
3. Operation Analytics and Investigating Metric Spike.
4. Hiring Process Analytics.
5. IMDB Movie Analysis.
6. Bank Loan Case Study.
7. XYZ Ads Airing Report Analysis.
8. ABC Call Volume Trend Analysis.

Learnings.

Conclusion.

Project-1 Data Analytics Process.

Description:

We use Data Analytics in everyday life without even knowing it.  
For eg : Going to a market to buy something .

Plan:- We first decide which things I need before going to market . Is it a shirt , jeans , footwear etc.

Prepare:- Next I need to check how much I am willing to spend and how to get that money.

Process:- Then I need to check how much I want from the data. Like if I am going to buy footwear what do I want - slippers / shoes / sandals etc.

Analyze:- You obviously won't buy things which are out of trend, Also you need to check does the jeans which you have and the color of t-shirt you want to buy, will it make a good combination.

Share:- Now you communicate your idea to the shopkeeper to find the best suitable fit for you.

Act:- Then you finally buy it!

###### Your 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.

# DATA ANALYTICS PROCESS

# Going for a Trip.

Plan:- Will decide about the place to visit like devotional ( Tirupathi/ Kanchi/ Srisailam ) or Fantasy(Goa).

Prepare: -How much cost we can afford for the trip and how to save money for it.

Process:- Will decide about travelling mode like car/train/bus and food expenses.

Analyze:- Check which mode of transport will cost less and check for availability of darshanam tickets. Consider when college gives consecutive holidays.

.

Share:- Comminicate with friends about trip and come to conclusion.

Act:- Finally, book train tickets(with return tickets) for Tirupathi from Bengaluru.

##### Project 2:-Instagram User Analytics

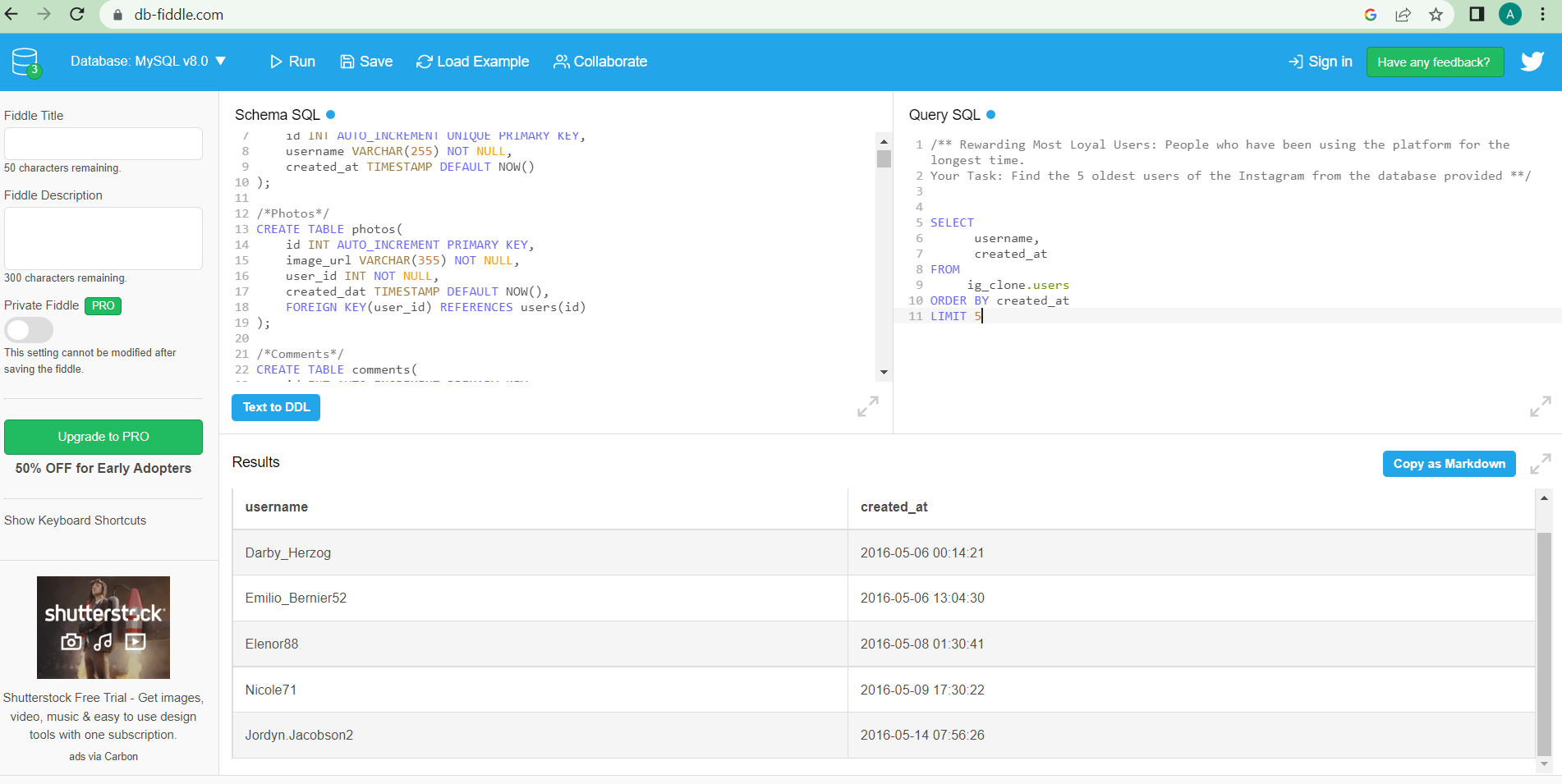
###### Description:

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.

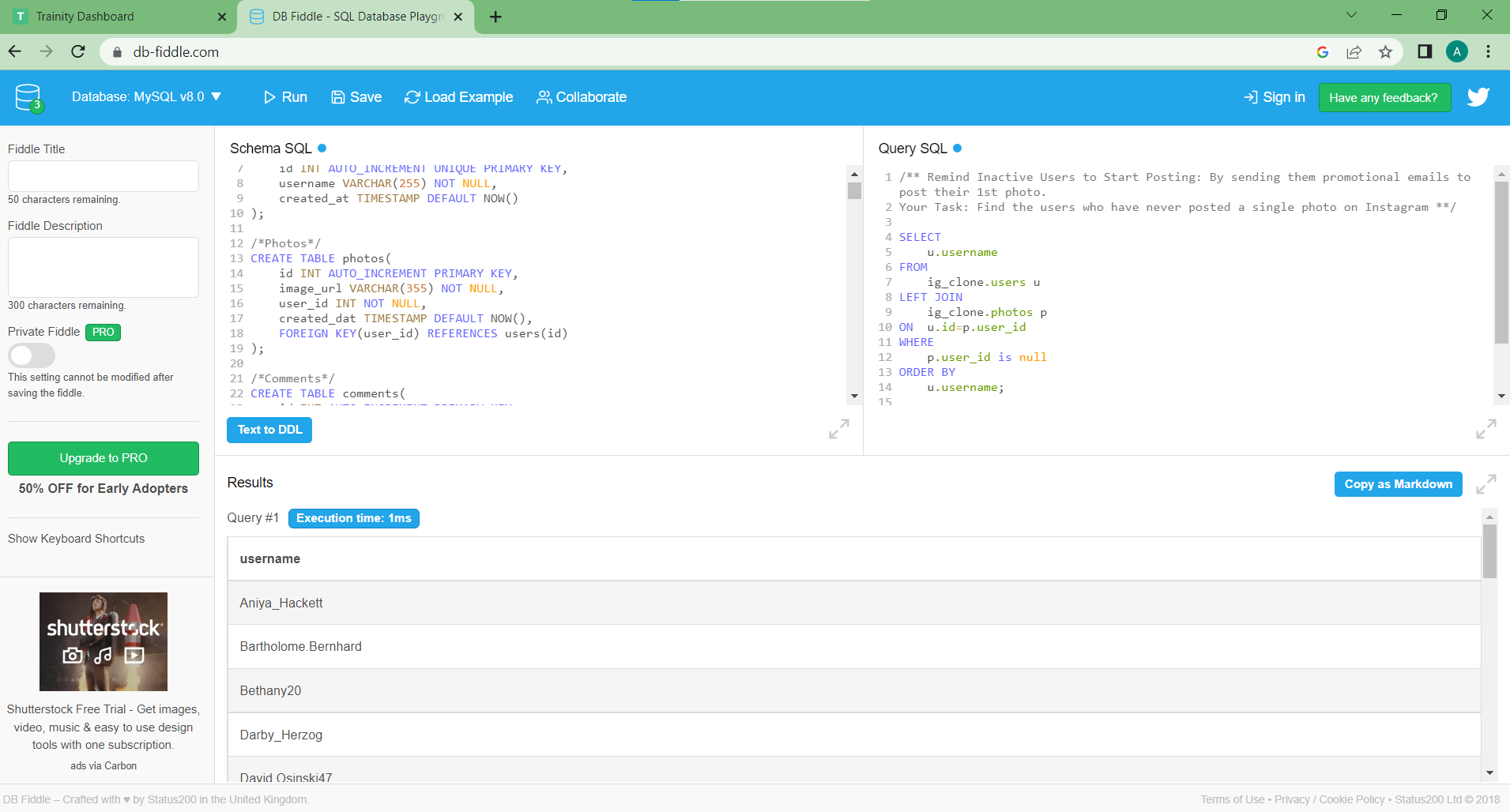
You are required to provide a detailed report answering the questions below :

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

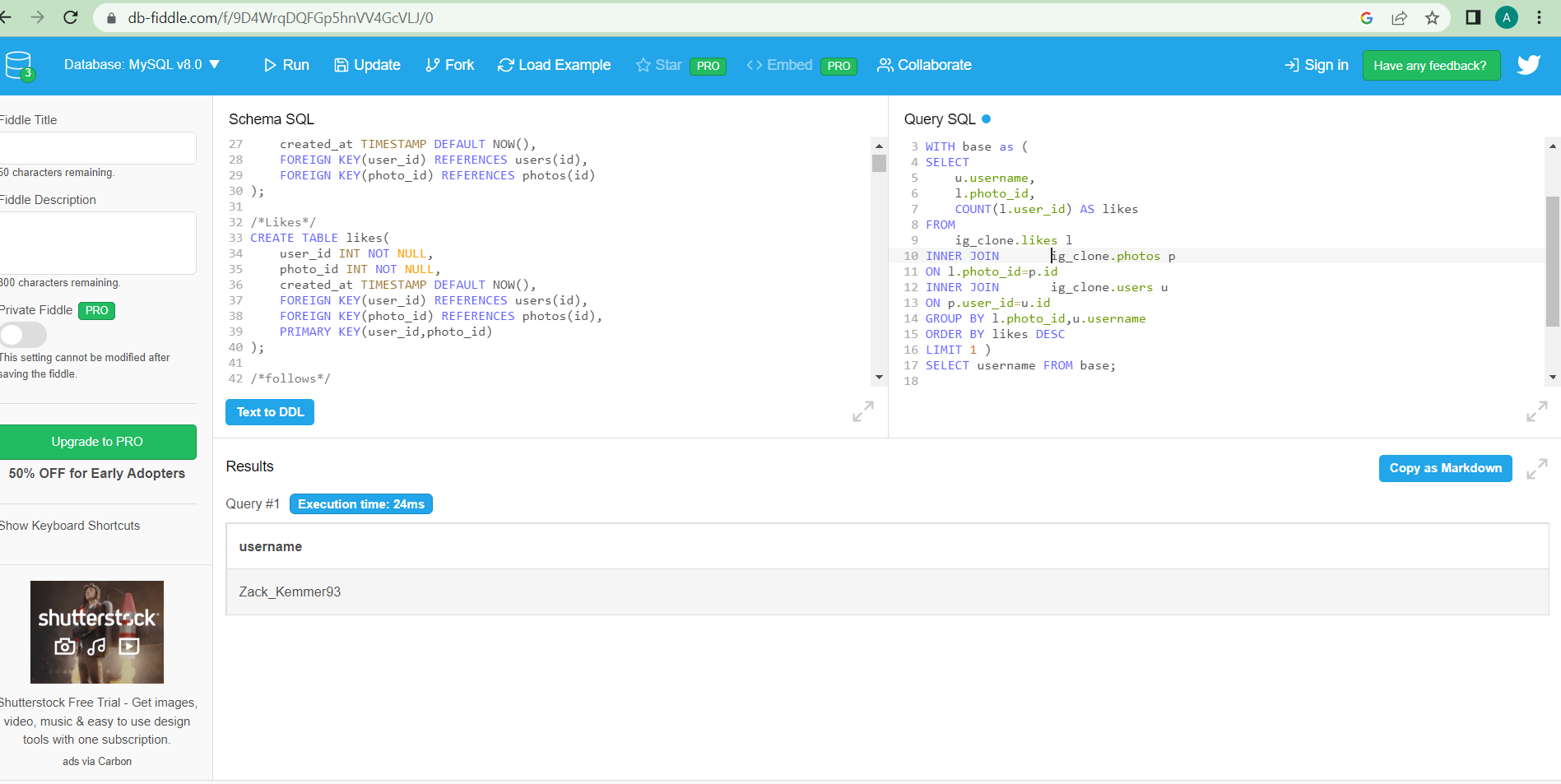
1. Rewarding Most Loyal Users: People who have been using the platform for the longest time.  
   Your Task: Find the 5 oldest users of the Instagram from the database provided.



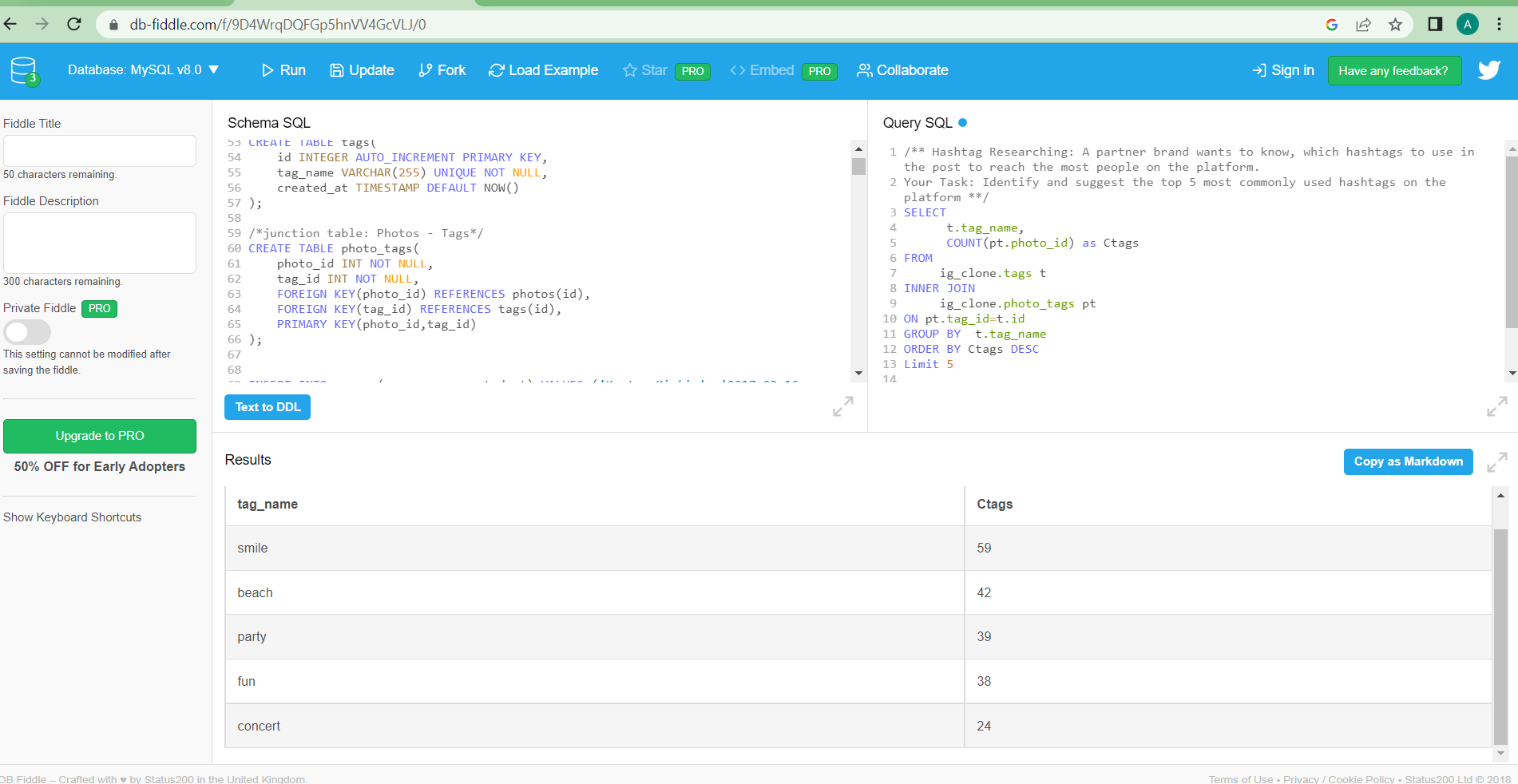
1. Remind Inactive Users to Start Posting: By sending them promotional emails to post their 1st photo.  
   Your Task: Find the users who have never posted a single photo on Instagram



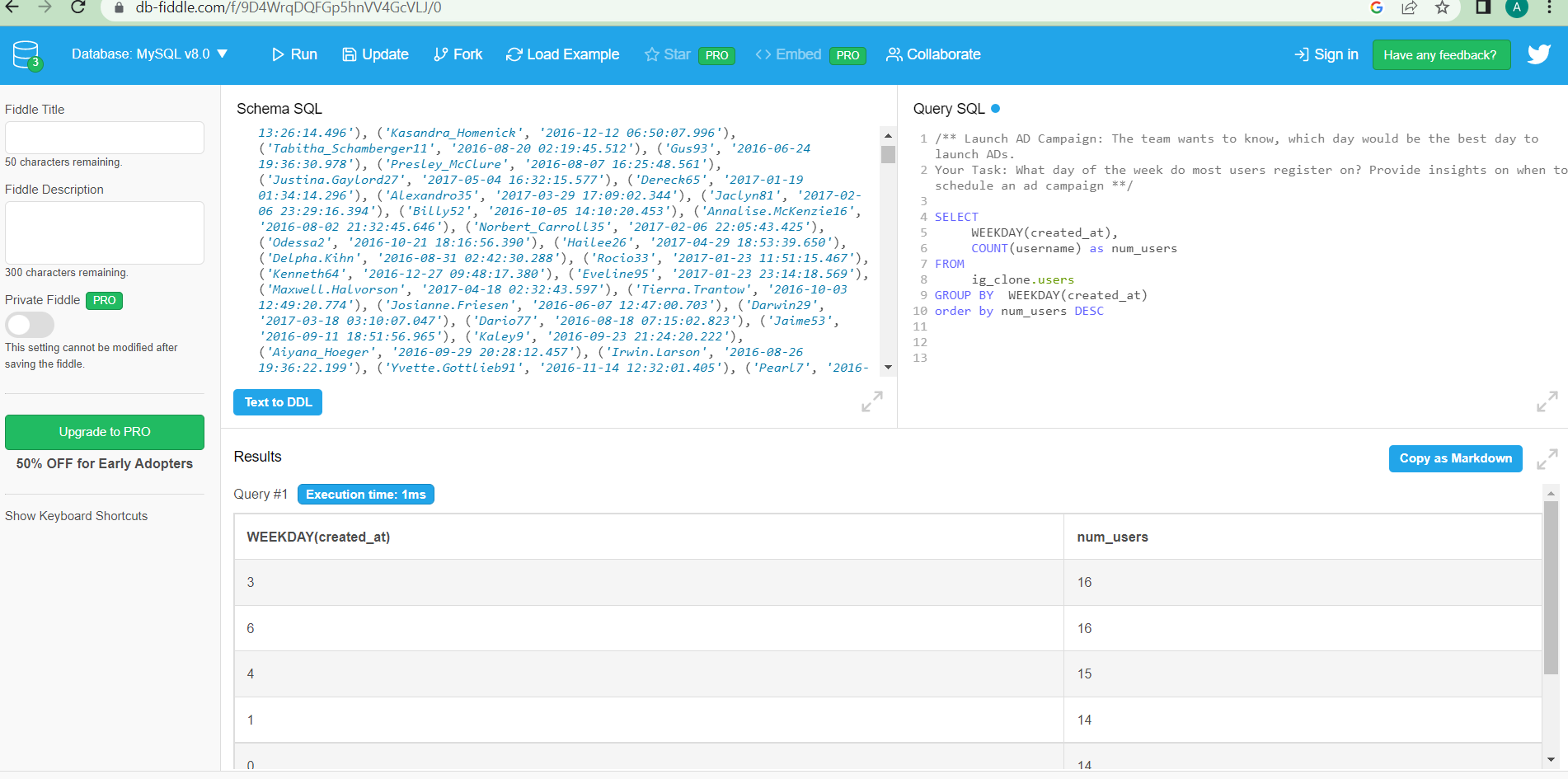
1. Declaring Contest Winner: The team started a contest and the user who gets the most likes on a single photo will win the contest now they wish to declare the winner.  
   Your Task: Identify the winner of the contest and provide their details to the team.



1. Hashtag Researching: A partner brand wants to know, which hashtags to use in the post to reach the most people on the platform.  
   Your Task: Identify and suggest the top 5 most commonly used hashtags on the platform.



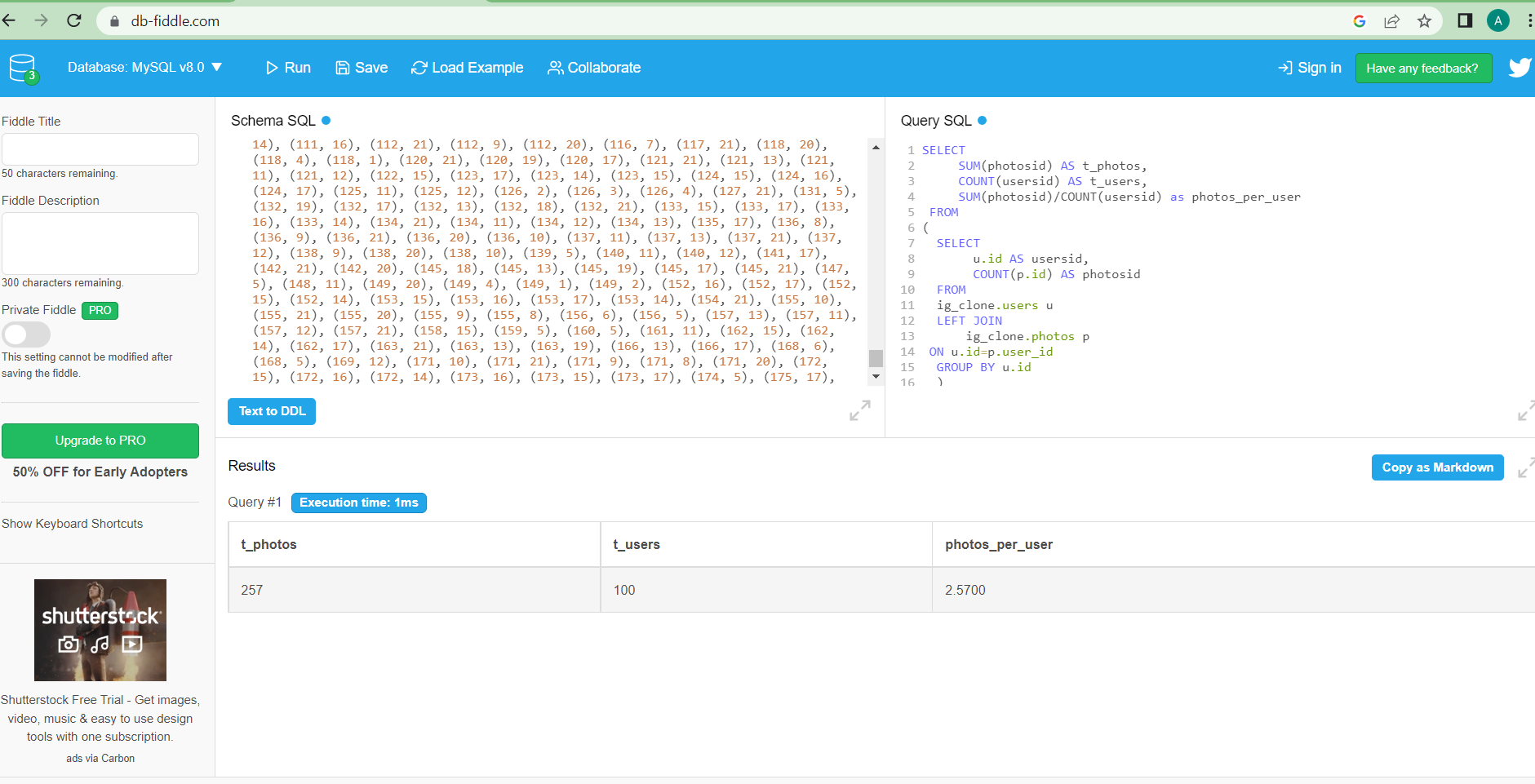
1. Launch AD Campaign: The team wants to know, which day would be the best day to launch ADs.  
   Your Task: What day of the week do most users register on? Provide insights on when to schedule an ad campaign



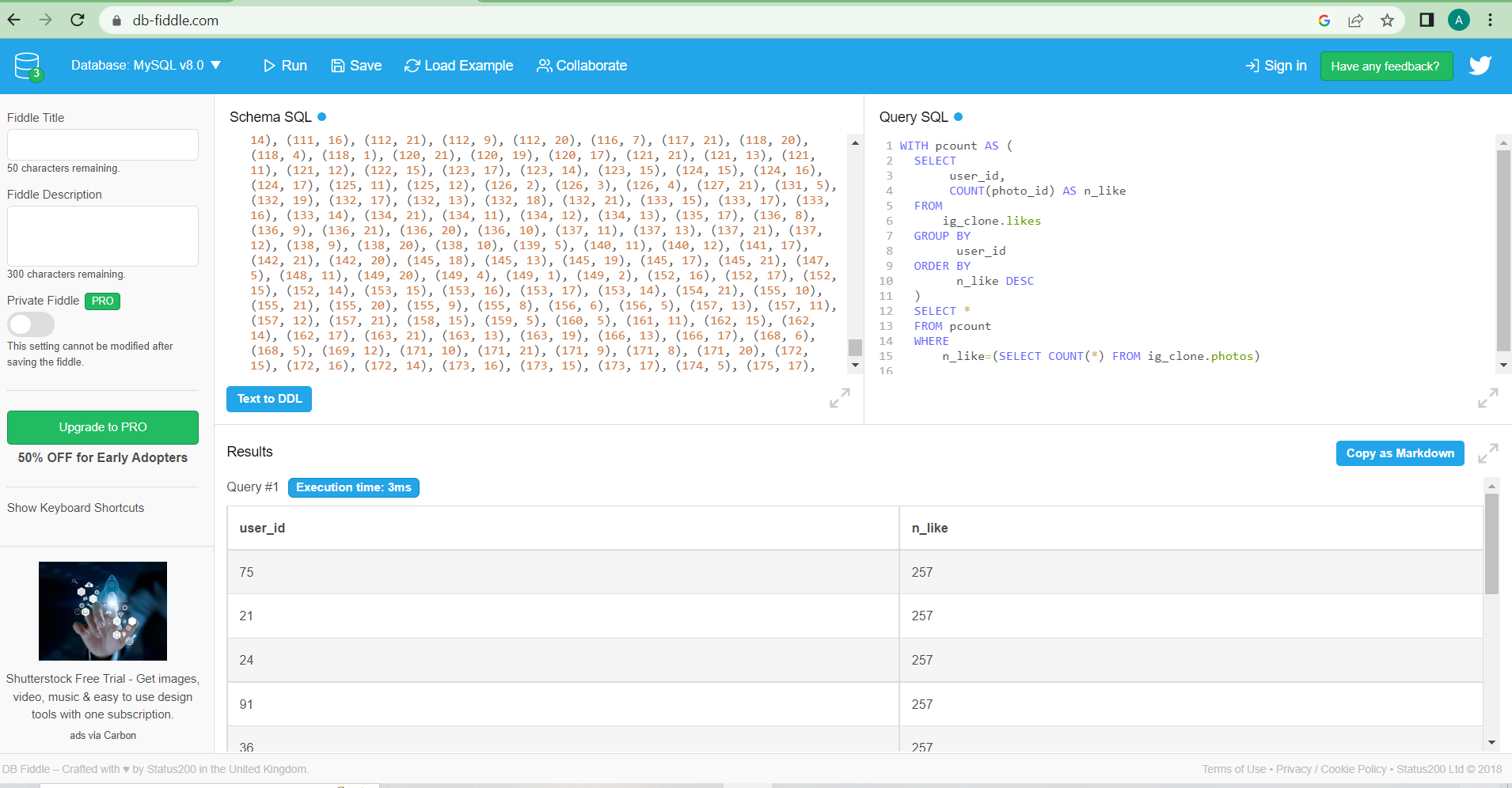
B) Investor Metrics: Our investors want to know if Instagram is performing well and is not becoming redundant like Facebook, they want to assess the app on the following grounds

1. User Engagement: Are users still as active and post on Instagram or they are making fewer posts  
   Your Task: Provide how many times does average user posts on Instagram. Also, provide the total number of photos on Instagram/total number of users





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



##### Project-3:- Operation Analytics and Investigating Metric Spike.

###### 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.

You are required to provide a detailed report for the below two operations mentioning the answers for the related questions:

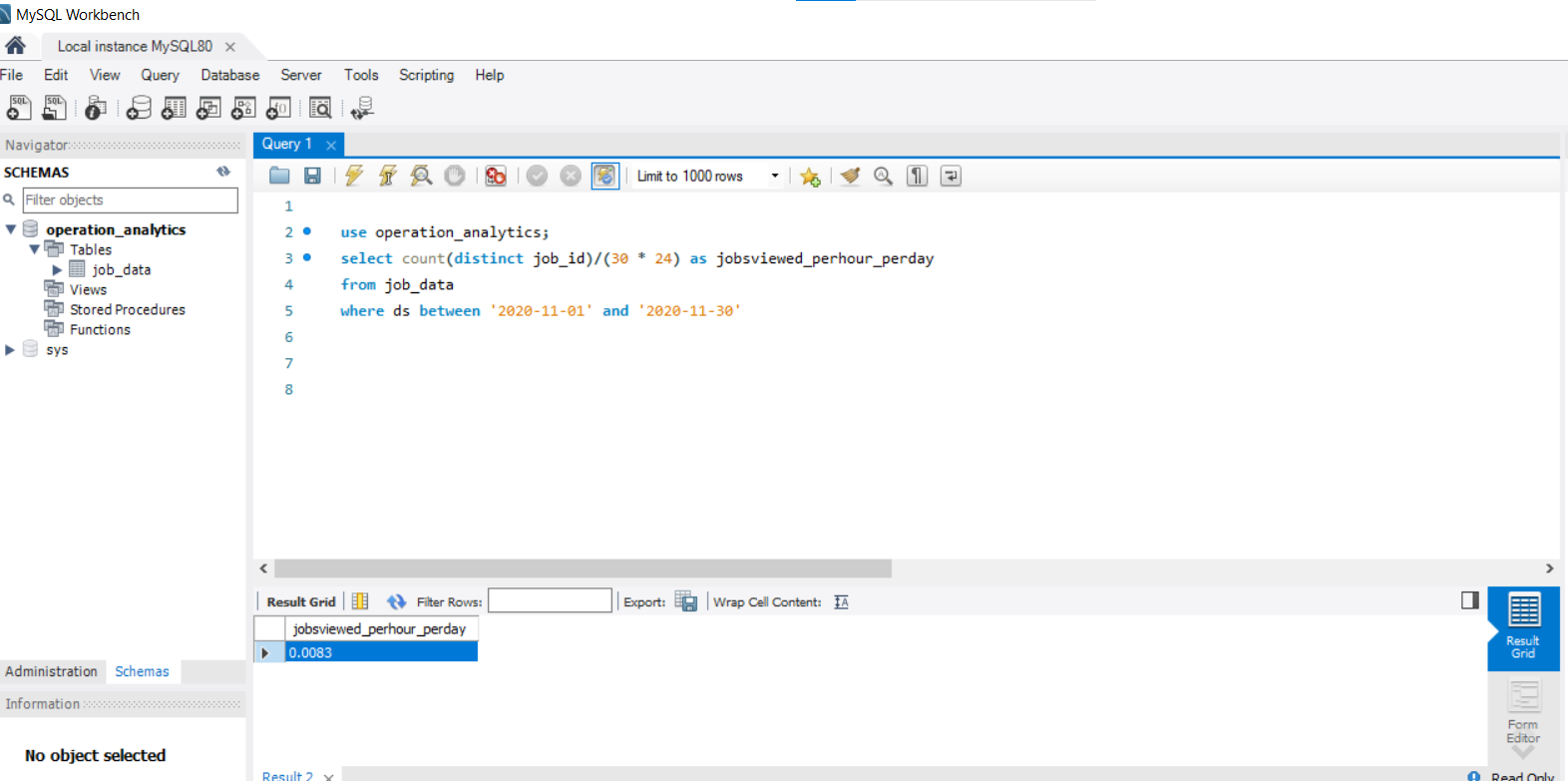
Case Study 1 (Job Data)

Below is the structure of the table with the definition of each column that you must work on:

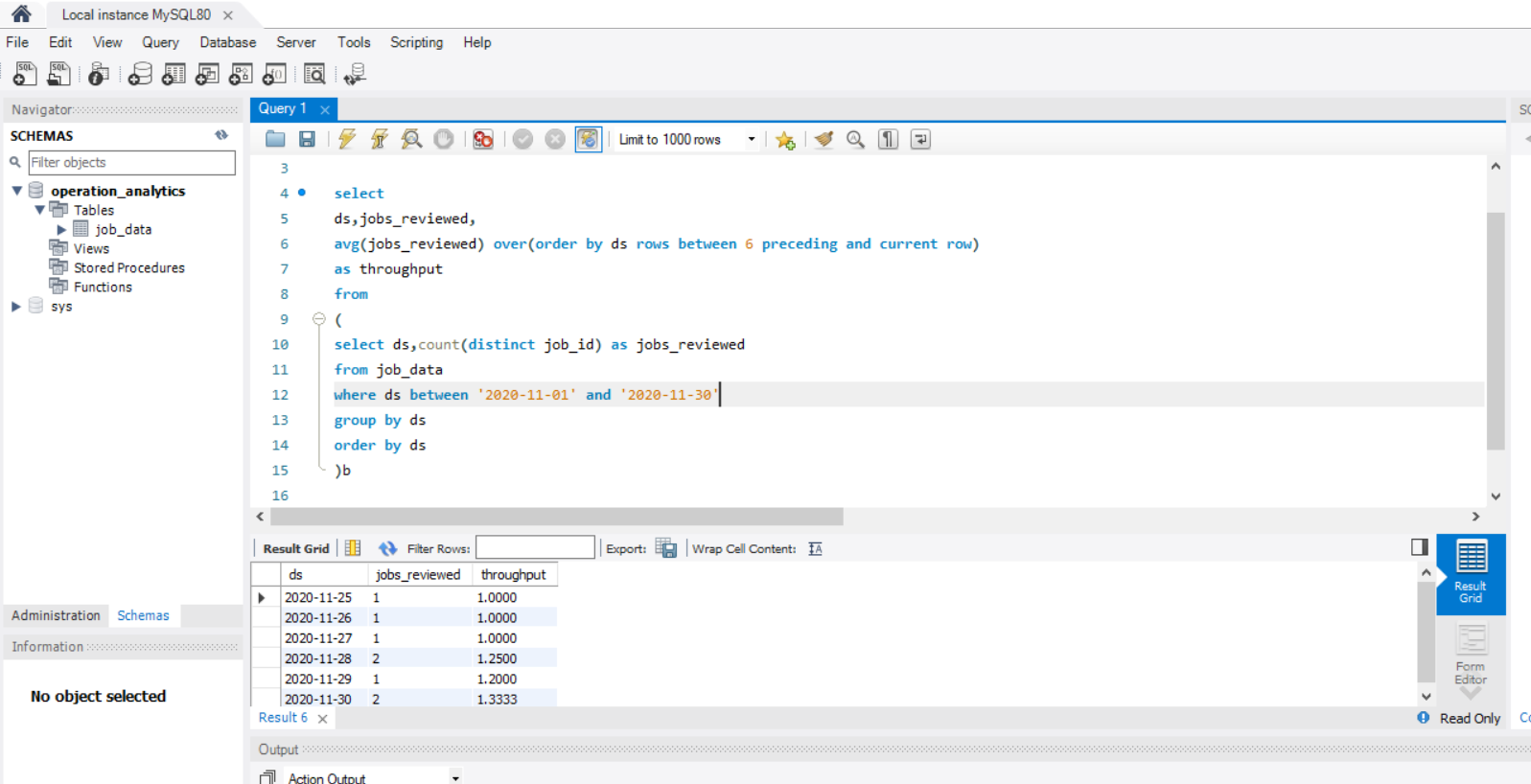
* Table-1: job\_data  
  + job\_id: unique identifier of jobs
  + actor\_id: unique identifier of actor
  + event: decision/skip/transfer
  + language: language of the content
  + time\_spent: time spent to review the job in seconds
  + org: organization of the actor
  + ds: date in the yyyy/mm/dd format. It is stored in the form of text and we use presto to run. no need for date function

Use the dataset attached in the Dataset section below the project images then answer the questions that follows

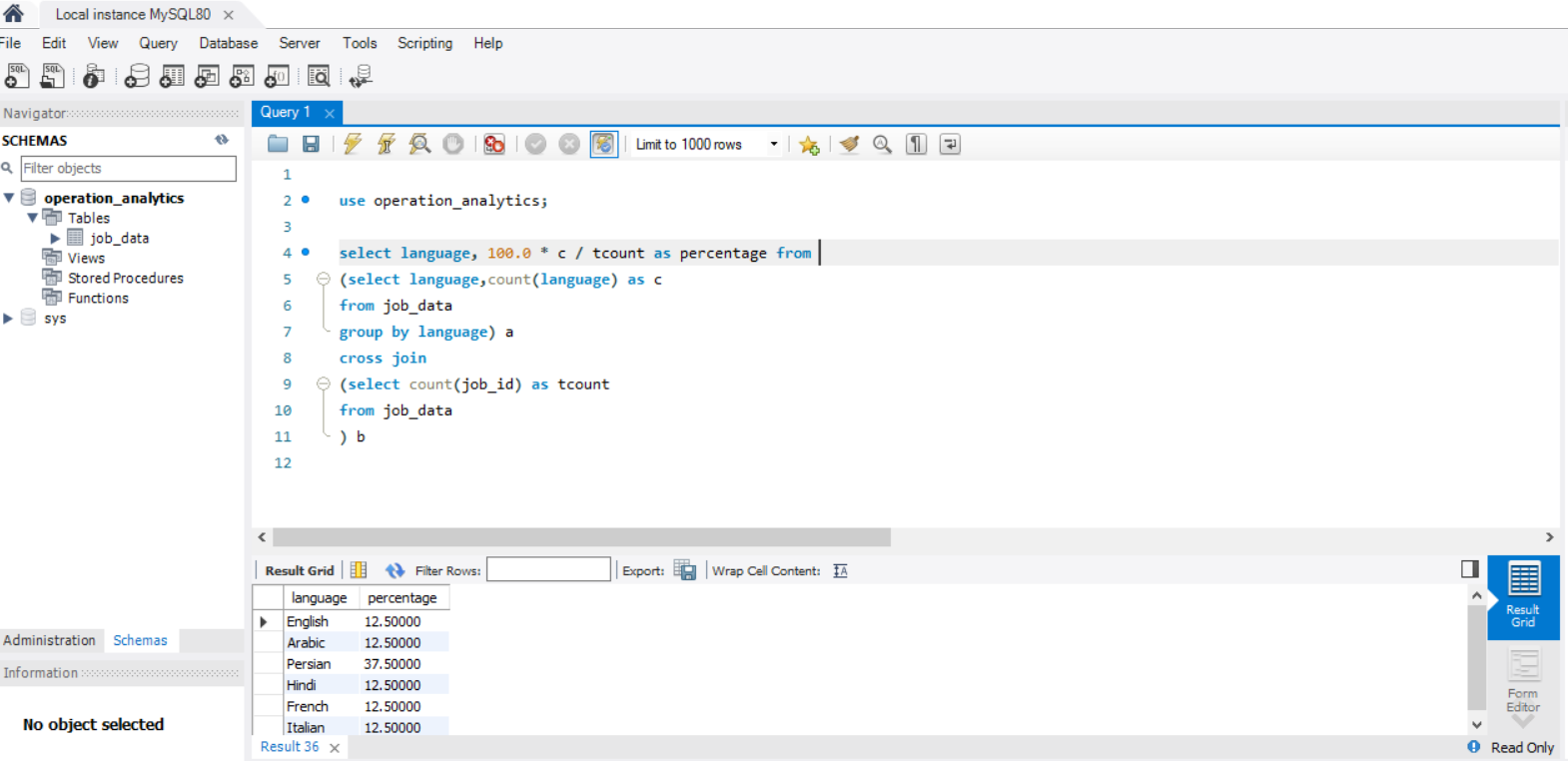
1. Number of jobs reviewed: Amount of jobs reviewed over time.  
   Your task: Calculate the number of jobs reviewed per hour per day for November 2020?



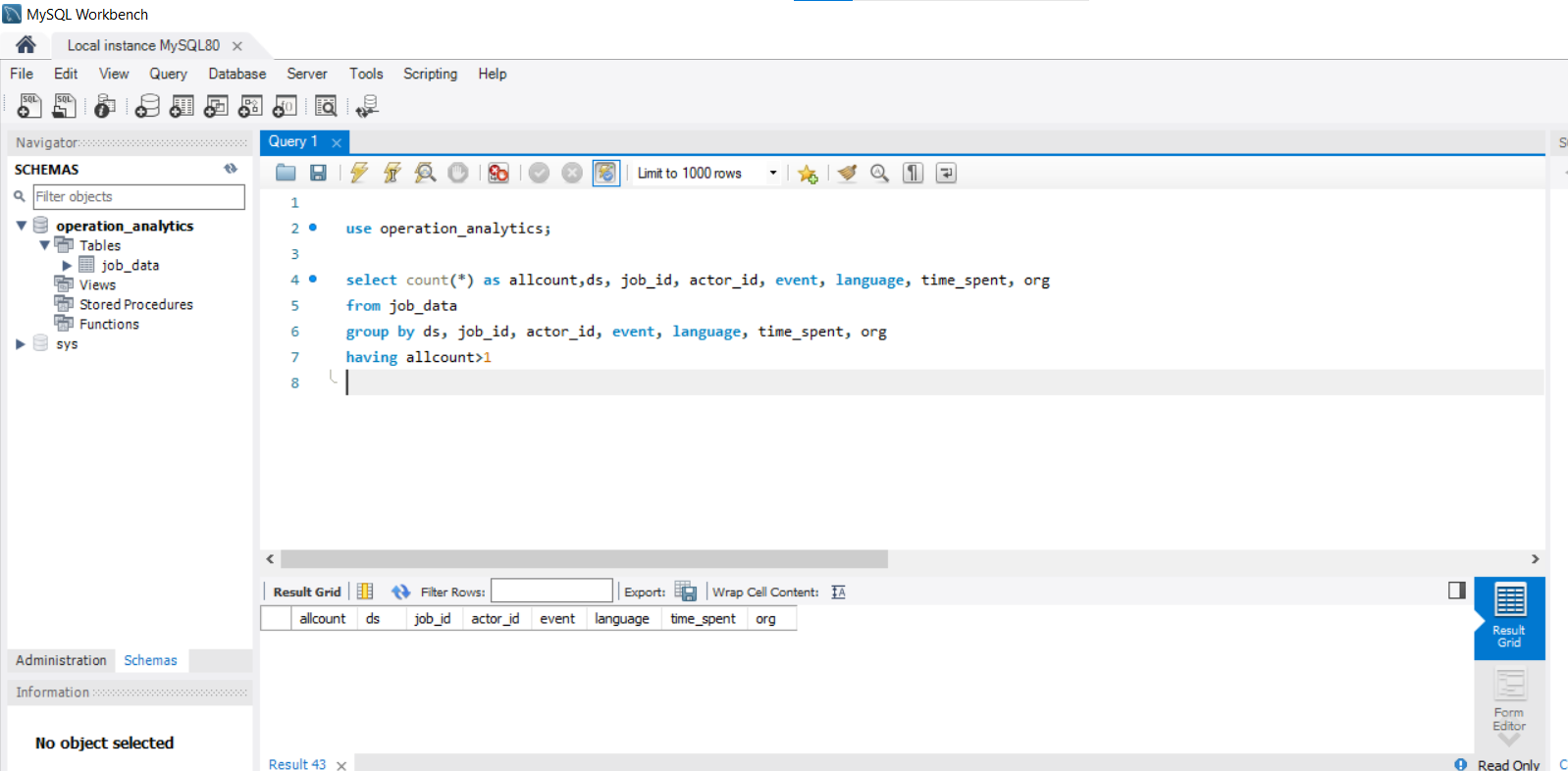
1. Throughput: It is the no. of events happening per second.  
   Your task: Let’s say the above metric is called throughput. Calculate 7 day rolling average of throughput? For throughput, do you prefer daily metric or 7-day rolling and why?



1. Percentage share of each language: Share of each language for different contents.  
   Your task: Calculate the percentage share of each language in the last 30 days?



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



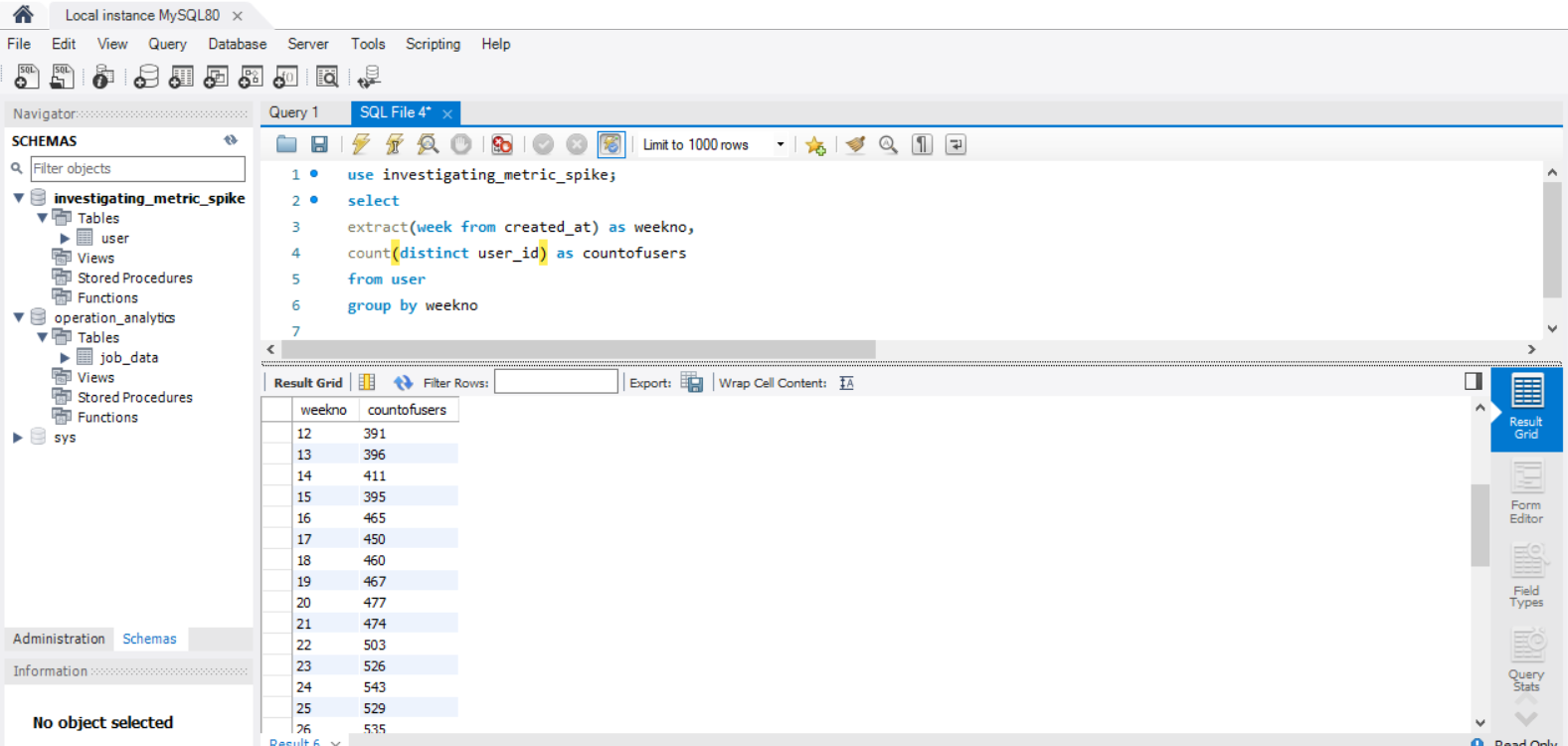
Case Study 2 (Investigating metric spike)

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

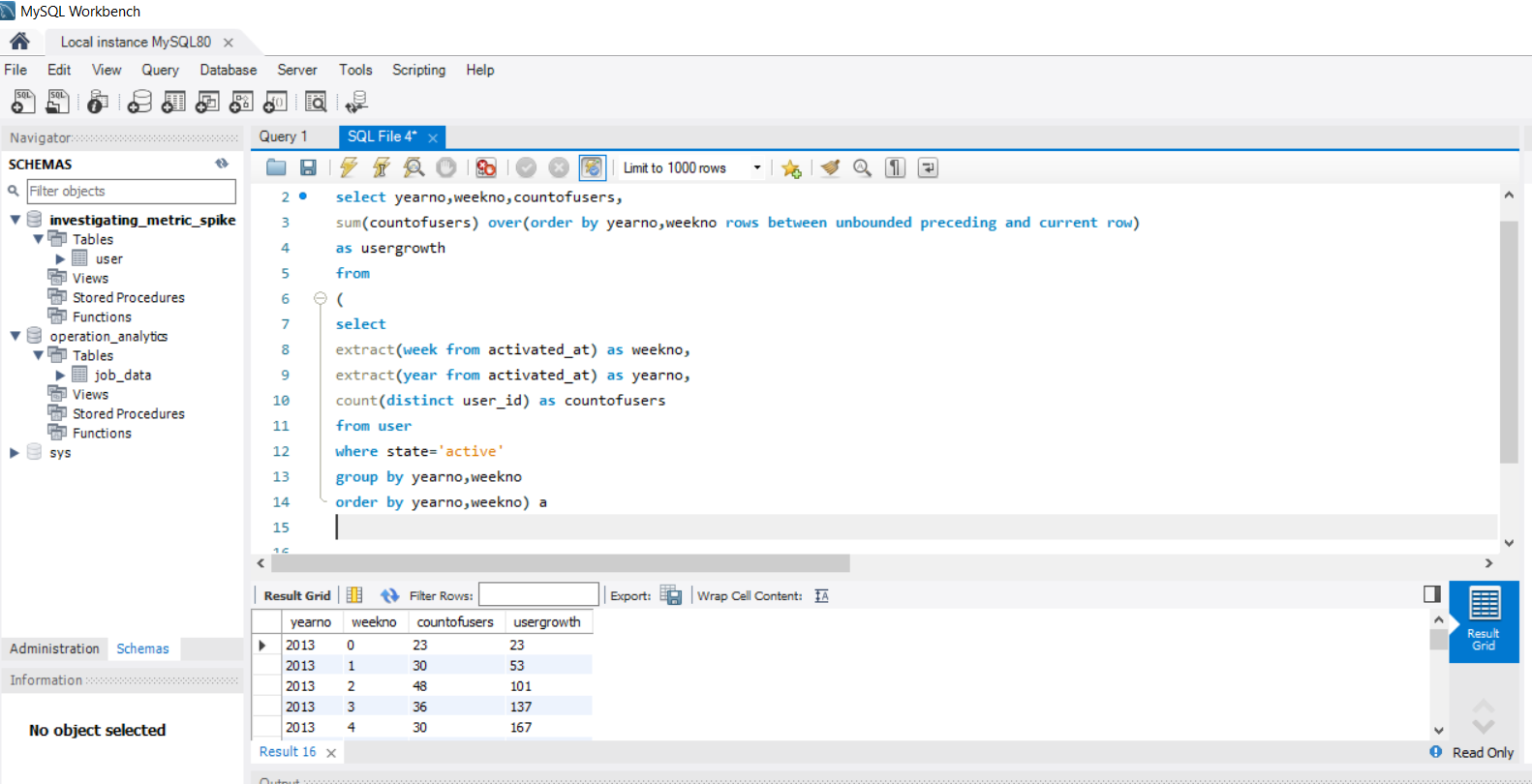
* Table-1: users  
  This table includes one row per user, with descriptive information about that user’s account.
* Table-2: events  
  This table includes one row per event, where an event is an action that a user has taken. These events include login events, messaging events, search events, events 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.

Use the dataset attached in the Dataset section below the project images then answer the questions that follows

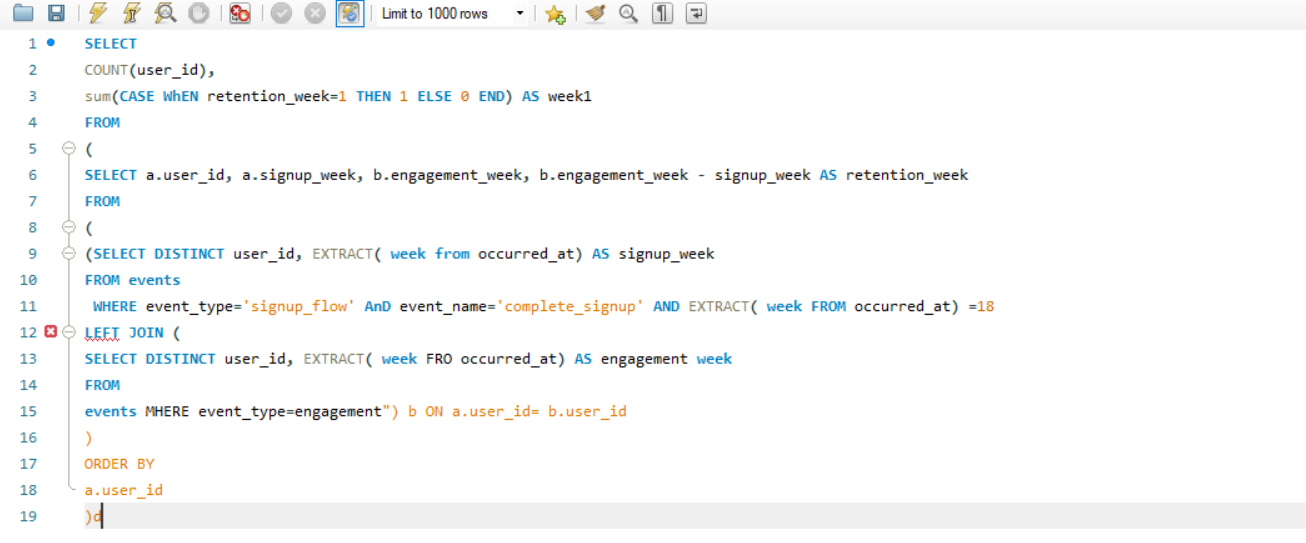
1. User Engagement: To measure the activeness of a user. Measuring if the user finds quality in a product/service.  
   Your task: Calculate the weekly user engagement?



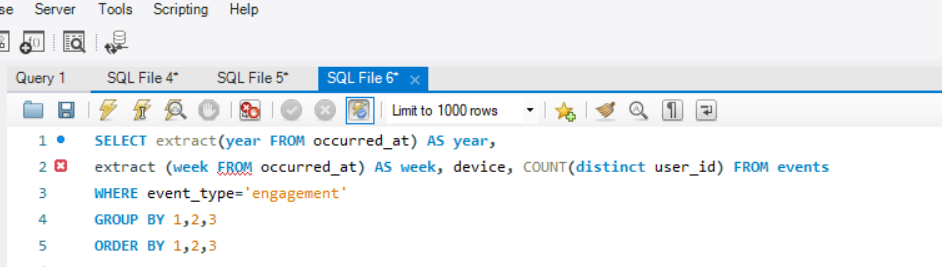
1. User Growth: Amount of users growing over time for a product.  
   Your task: Calculate the user growth for product?



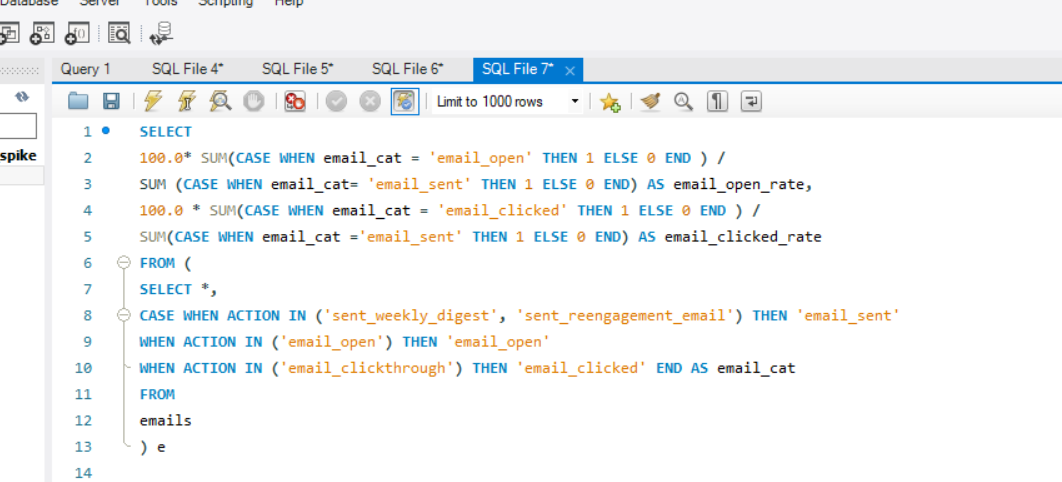
1. Weekly Retention: Users getting retained weekly after signing-up for a product.  
   Your task: Calculate the weekly retention of users-sign up cohort?



1. 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?



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



##### Project-4:-Hiring Process Analytics.

###### Description:

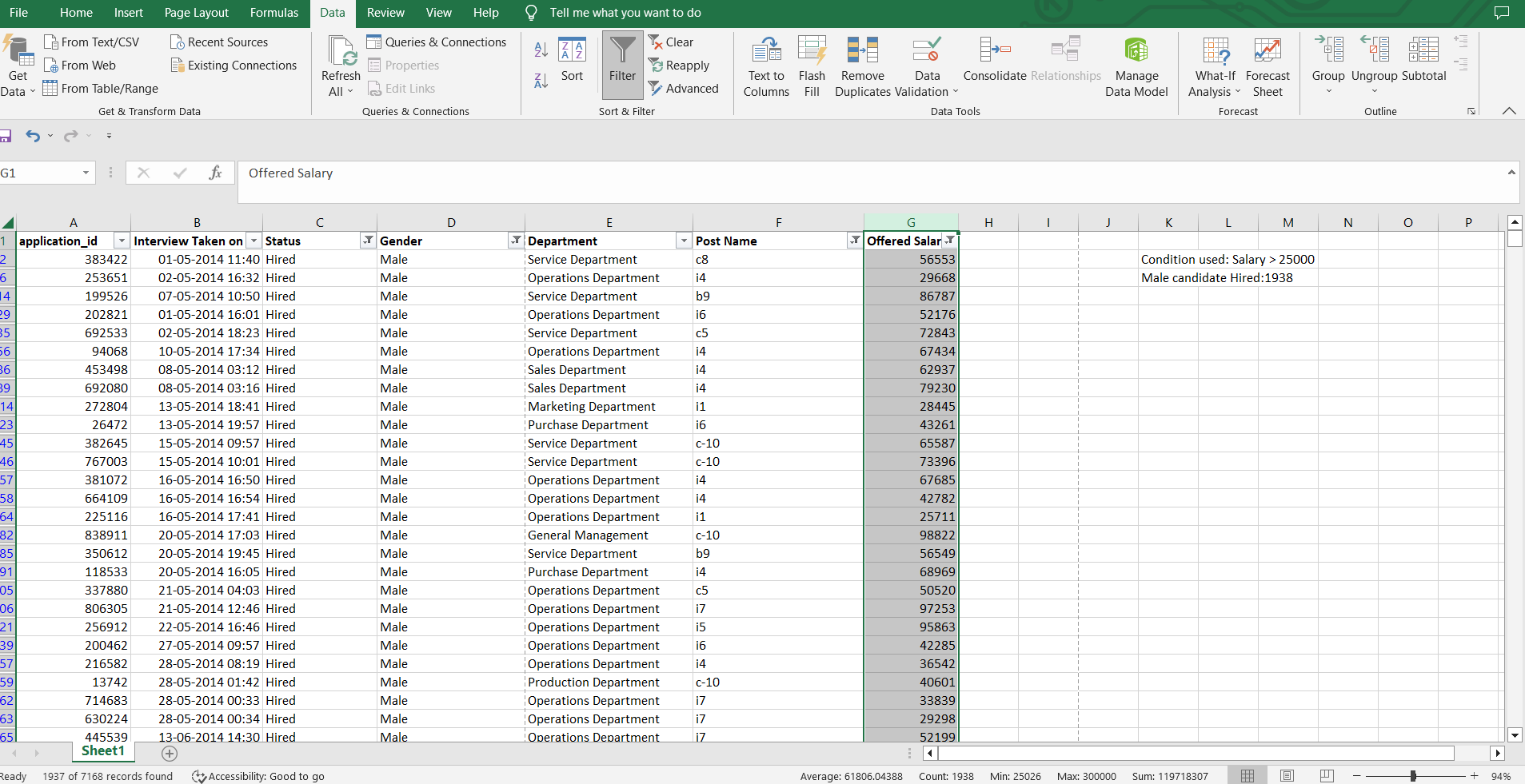
Hiring process is the fundamental and the most important function of a company. Here, the MNCs get to know about the major underlying trends about the hiring process. Trends such as- number of rejections, number of interviews, types of jobs, vacancies etc. are important for a company to analyse before 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.  
  
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.

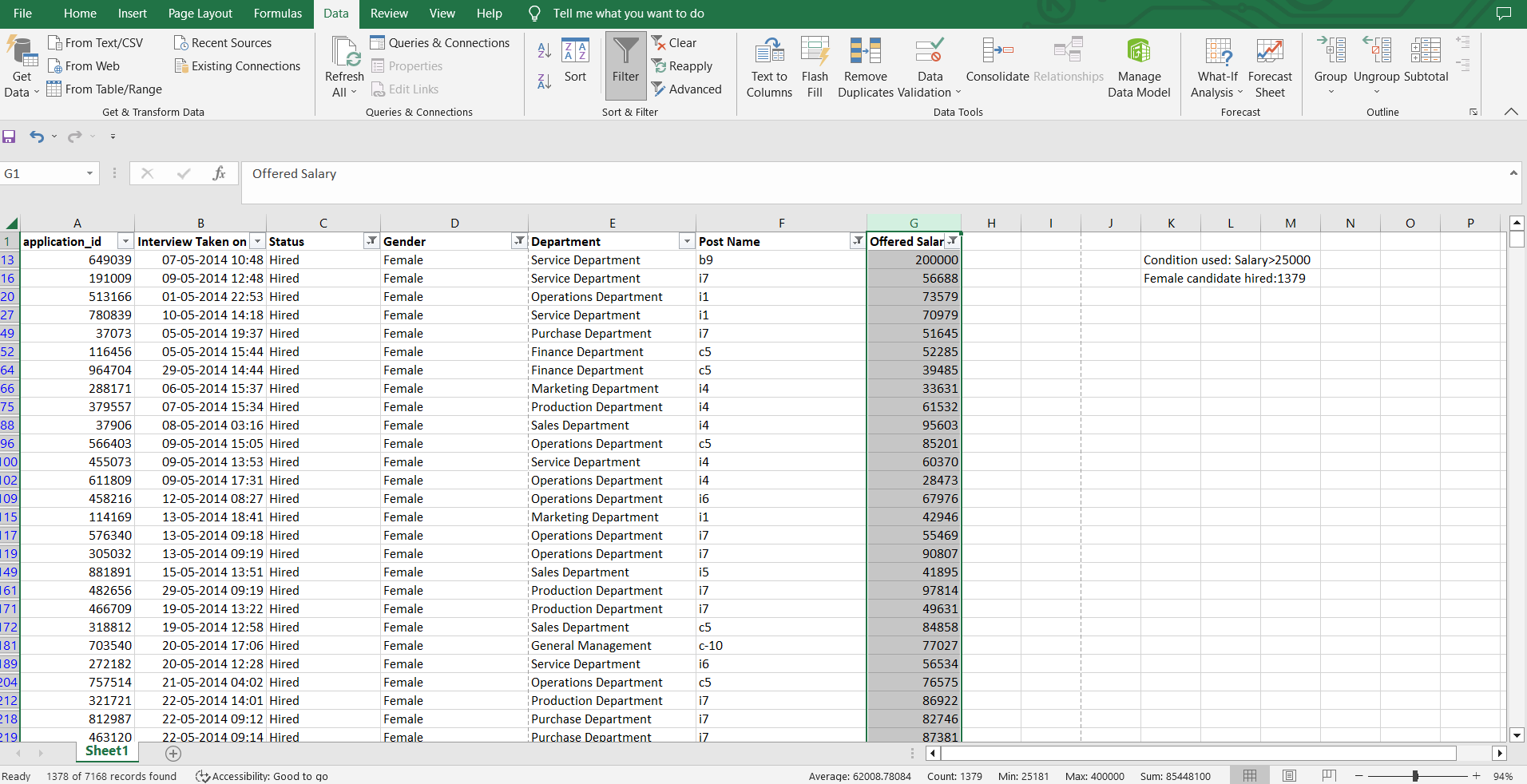
**You are required to provide a detailed report for the below data record mentioning the answers of the questions that follows:**  
  
You are given a dataset of a company where the details about people who registered for a particular post in a department of this company. You are required to use your knowledge in statistics and use different formulas in excel and draw necessary conclusions about the company.  
  
Use the below **Steps for EDA**

1. Understanding data columns and data
2. Checking for missing data
3. Clubbing columns with multiple categories
4. Checking for outliers
5. Removing outliers
6. Drawing Data Summary

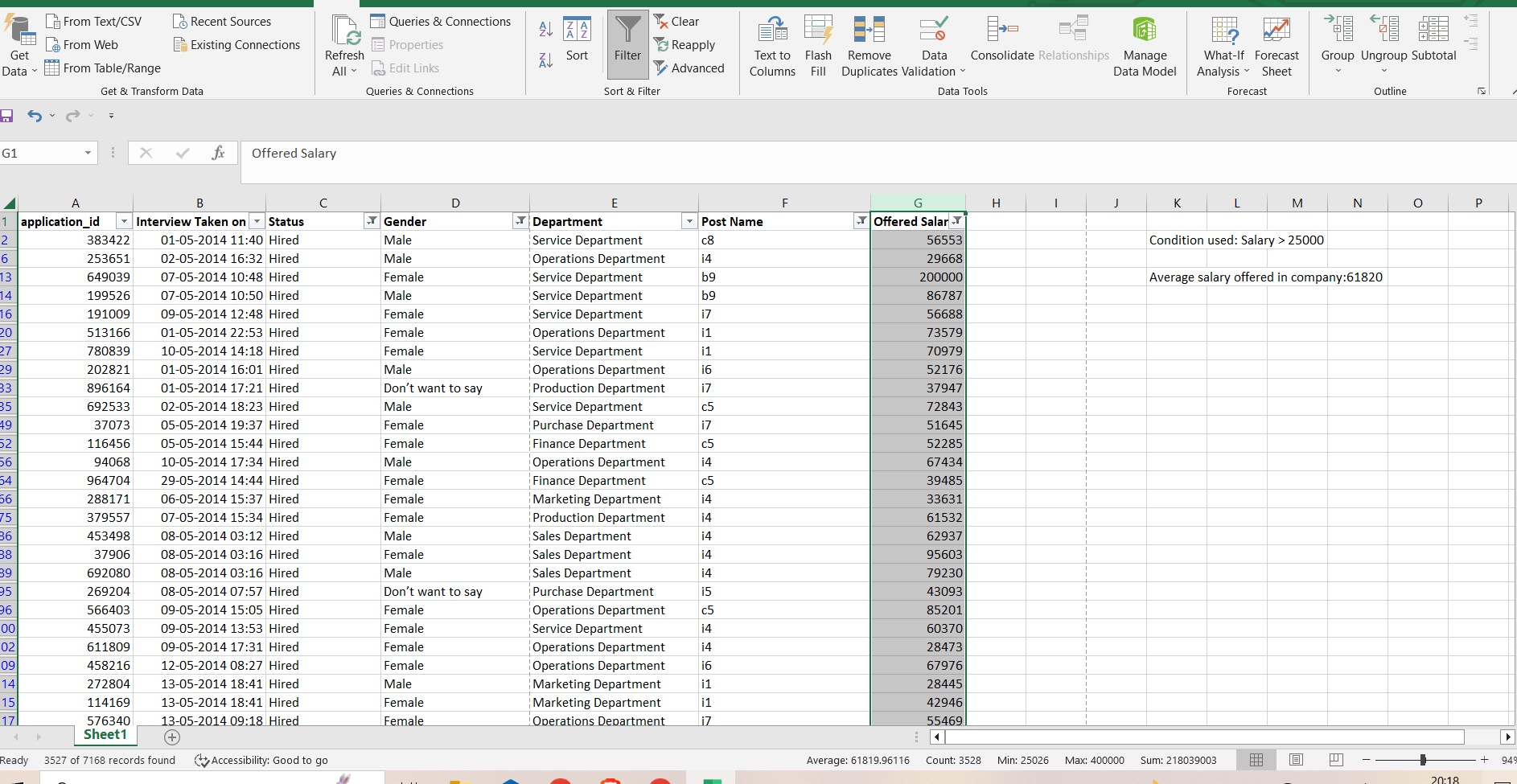
After downloading the dataset, use Excel or Google Sheets to answer the below questions:

1. Hiring: Process of intaking of people into an organization for different kinds of positions.  
   Your task: How many males and females are Hired ?

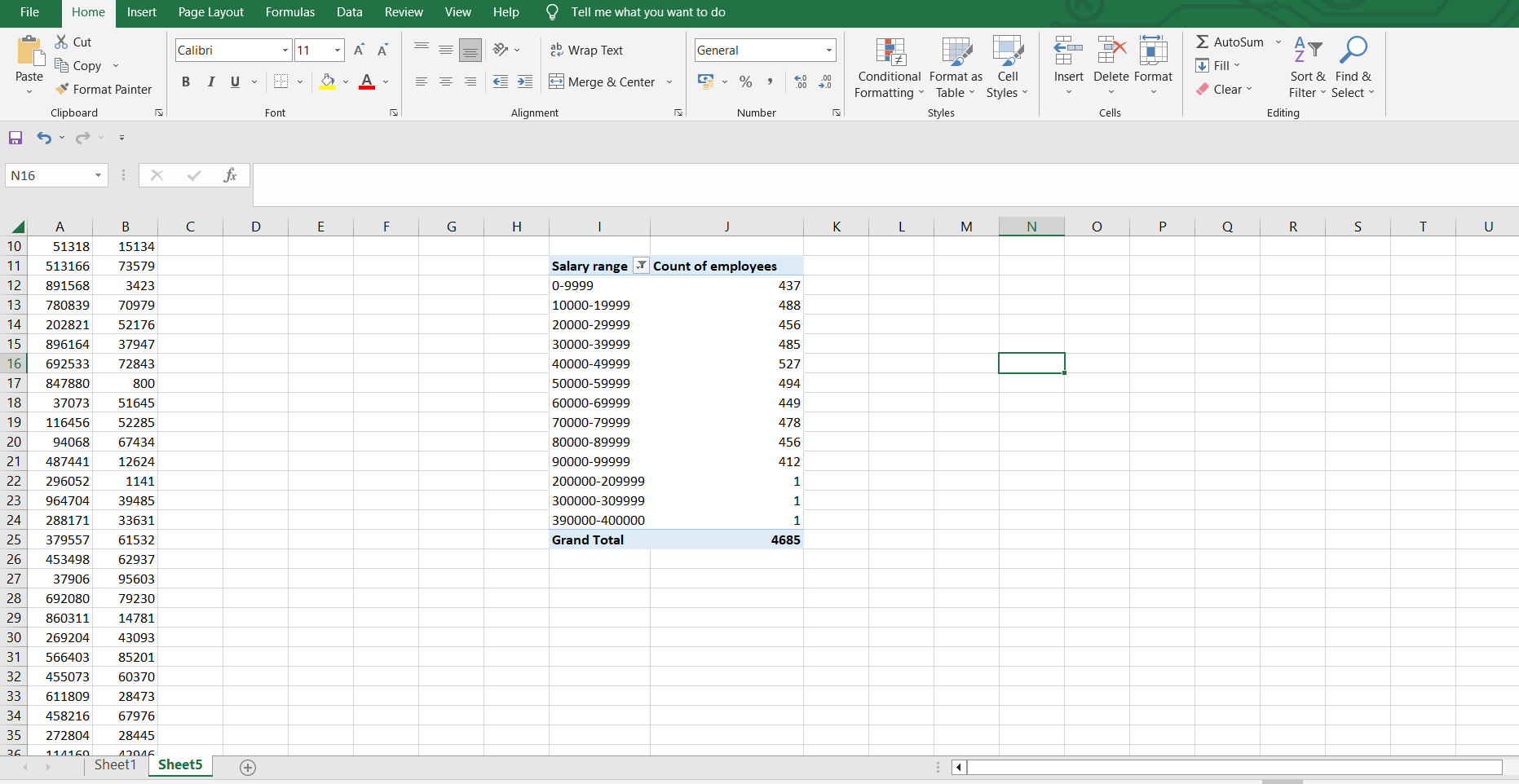




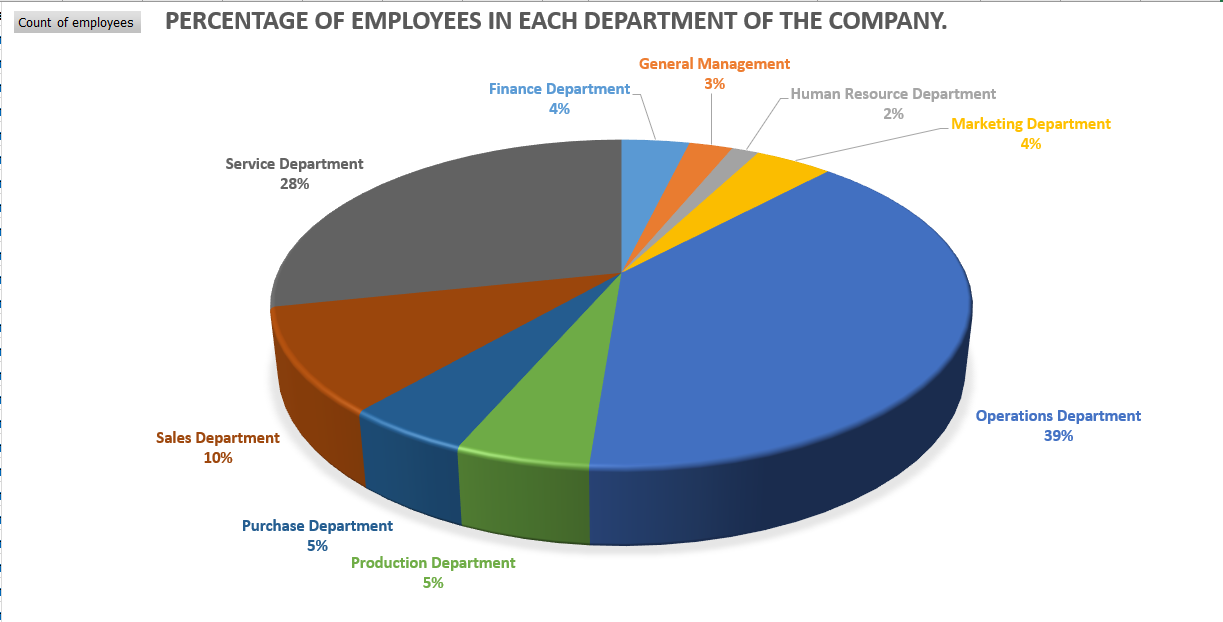
1. Average Salary: Adding all the salaries for a select group of employees and then dividing the sum by the number of employees in the group.  
   Your task: What is the average salary offered in this company ?

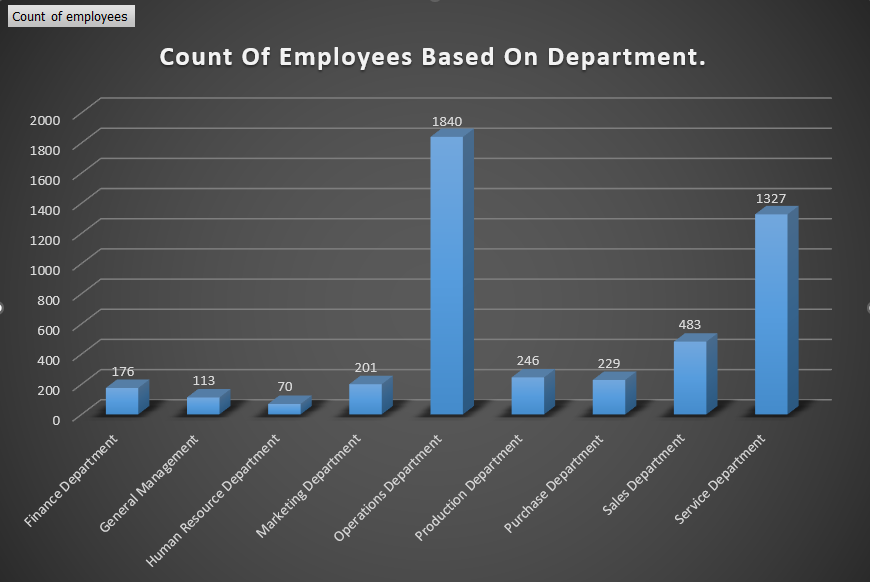


1. Class Intervals: The class interval is the difference between the upper class limit and the lower class limit.  
   Your task: Draw the class intervals for salary in the company ?

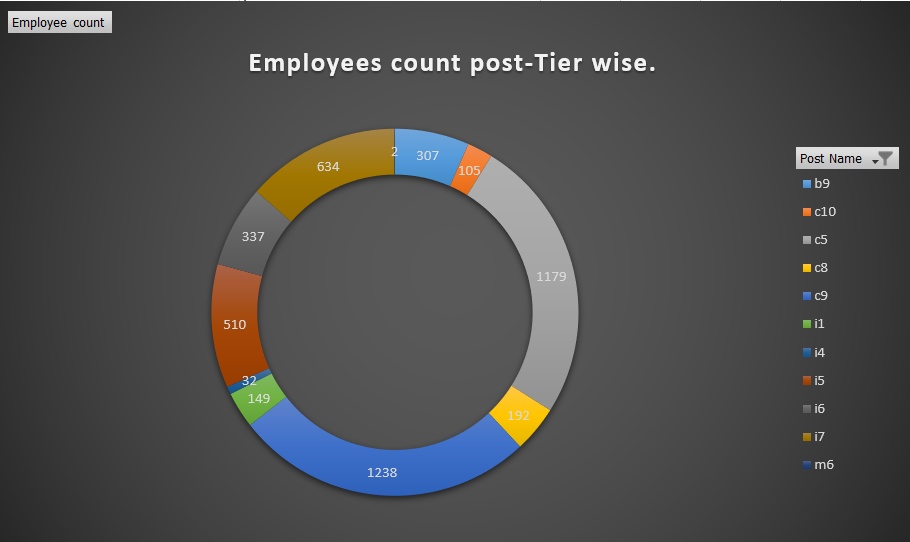


1. Charts and Plots: This is one of the most important part of analysis to visualize the data.  
   Your task: Draw Pie Chart / Bar Graph ( or any other graph ) to show proportion of people working different department ?





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



##### Project-5:-IMDB Movie Analysis.

###### Description:

For your Final Project, we are providing you with dataset having various columns of different IMDB Movies. You are required to Frame the problem. For this task, you will need to define a problem you want to shed some light on.  
  
We can do this by asking 'What?' This is where you frame the problem i.e. What is the problem?  
  
Use these questions to guide your thinking:

* What do you see happening?
* What is your hypothesis for the cause of the problem? (this will be broadly based on intuition initially)
* What is the impact of the problem on stakeholders?
* What is the impact of the problem not being solved?

Answering these questions will help you define a problem you are trying to solve and will allow you to find the right data to solve it.

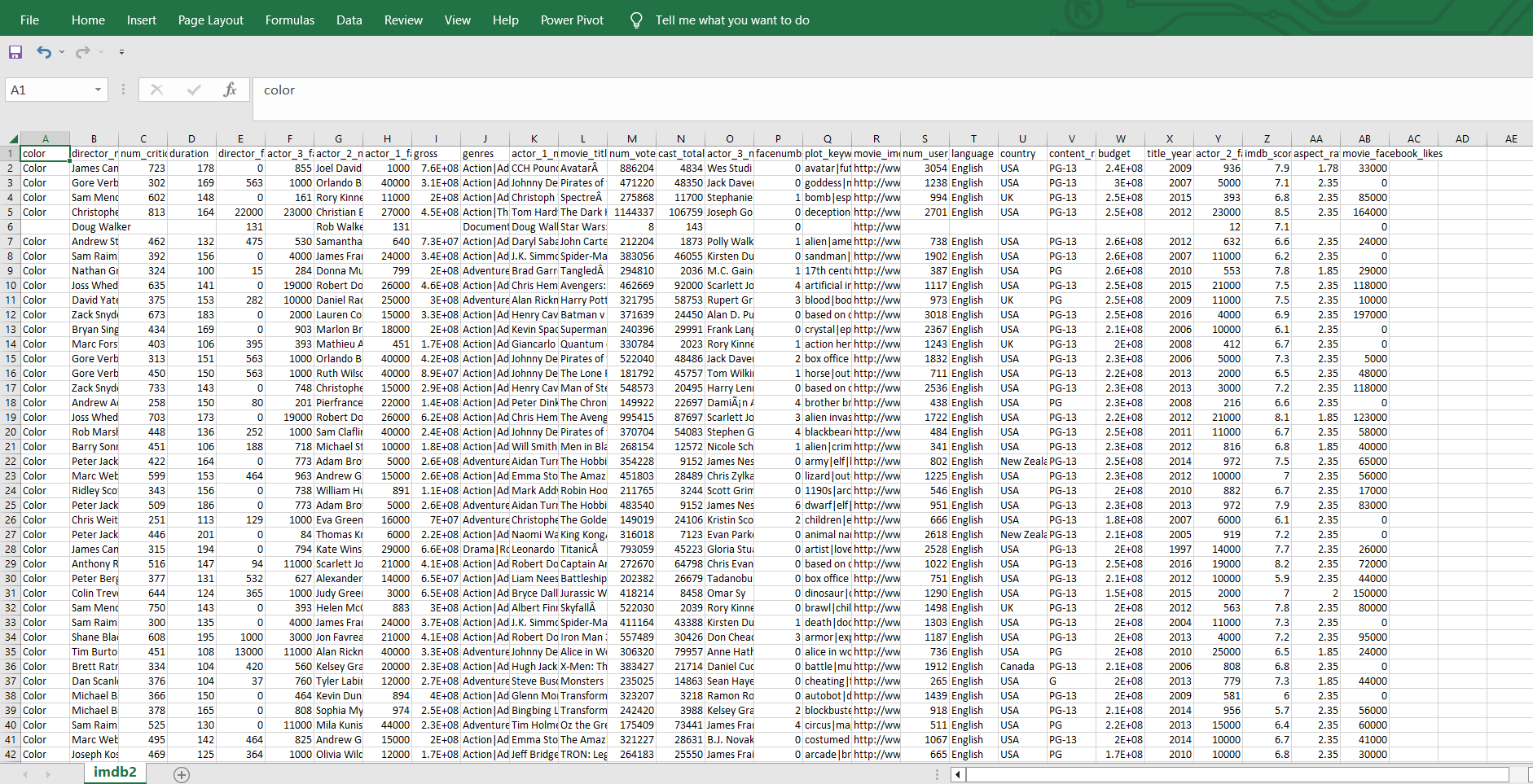
Once you have defined a problem, clean the data as necessary, and use your Data Analysis skills to explore the data set and derive insights.  
  
Make sure to use 5 Whys Analysis in your analysis and use this to create a report which conveys a data story.  
  
Once you have framed the problem and gathered initial insights from the data, you can ask the following questions as you dig deeper into your analysis.

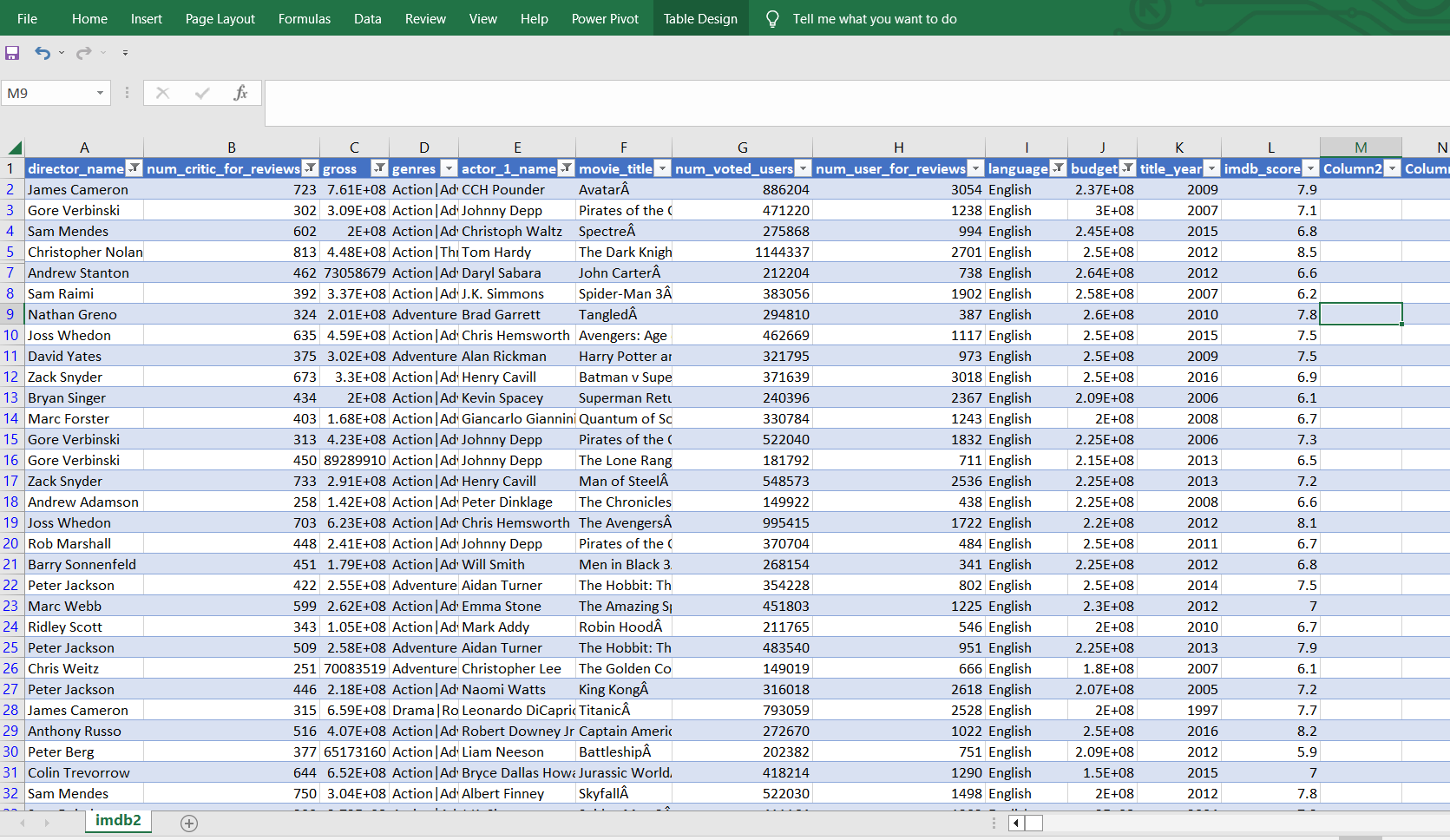
* What do you see happening?
* What are the specific symptoms of the problem?
* What is your hypothesis for the cause of the problem?

Five 'Whys' approach  
  
Once you have the problem better defined, you can use 5 Whys technique to determine its root cause by repeatedly asking the question “Why”.  
  
It's also called the Root Cause Analysis, developed by Sakichi Toyoda, founder of Toyota Industries. Here’s an example of how this technique could be used to figure out the cause of the following problem: A business went over budget on a recent project.  
  
Q: “Why did we go over budget on our project?”  
  
A: It took much longer than we expected to complete.  
  
Q: “Why did it take longer than expected to complete?”  
  
A: We had to redesign several elements of the product.  
  
Q: “Why did we have to redesign elements of the product?”  
  
A: Features of the product were confusing to use.  
  
Q: “Why were the features of the product confusing to use?”  
  
A: We made incorrect assumptions about what users wanted.  
  
Q: “Why did we make incorrect assumptions about what users wanted?”  
  
A: Our user experience research team didn’t ask effective questions.  
  
As you see above, what looked like a budgeting problem turned out to be a problem with the user experience team not working effectively.  
  
While asking Why is easy, what we're interested in is the answer. Each time you answer why the next time gets more difficult as you must think deeper behind the reasons for this. As you ask why, you may find that you have multiple answers for the same question.

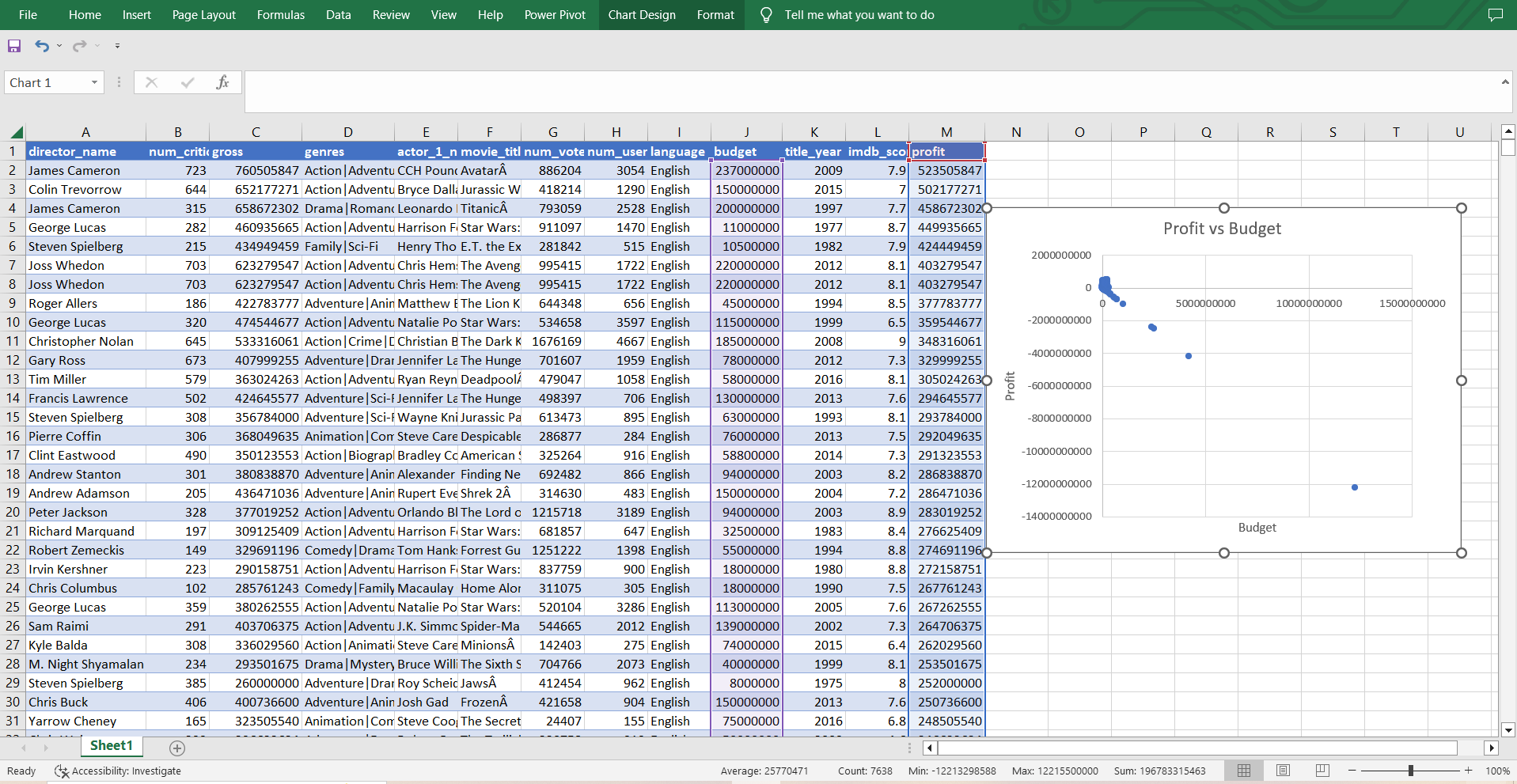
You are required to provide a detailed report for the below data record mentioning the answers of the questions that follows:

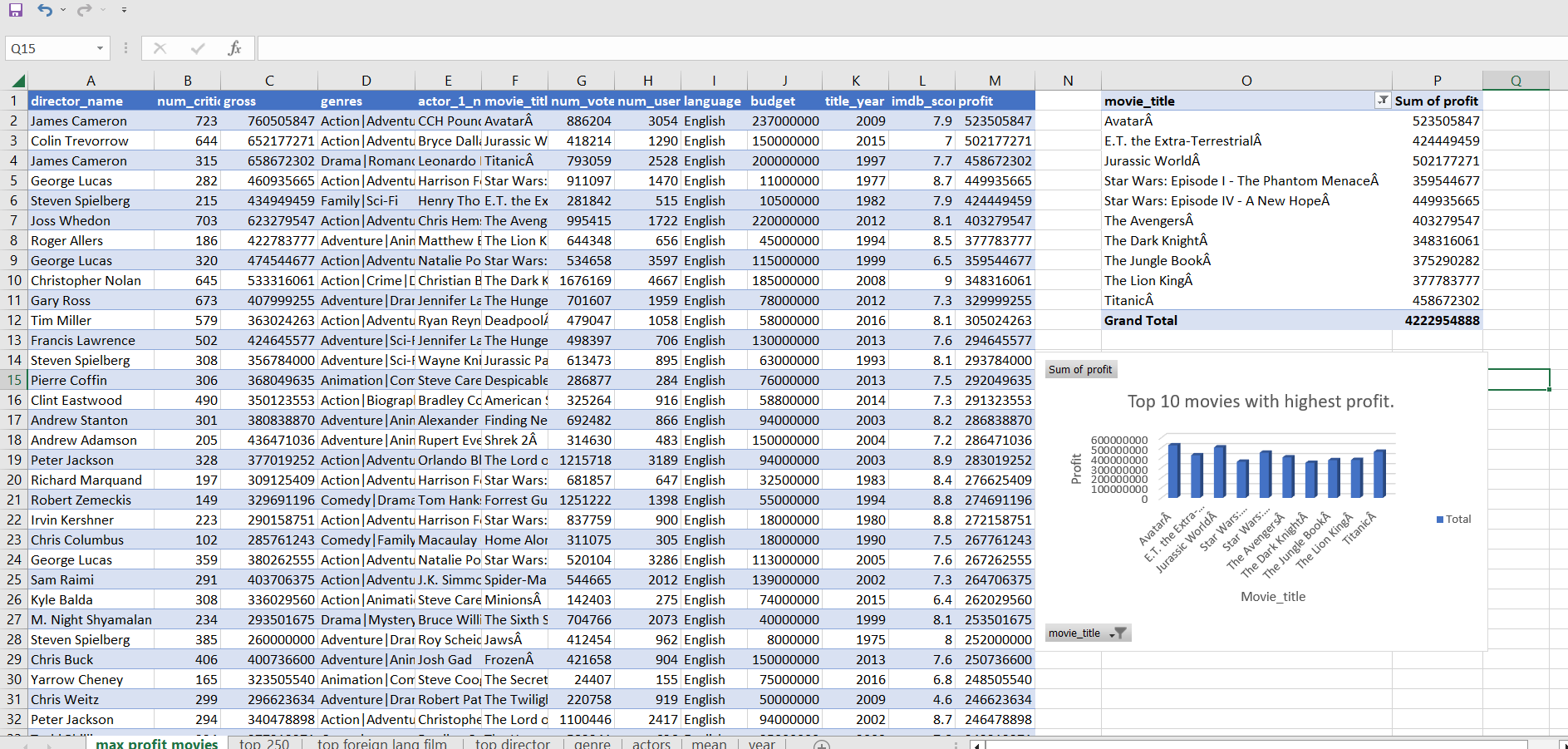
1. Cleaning the data:: This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)  
   Your task: Clean the data

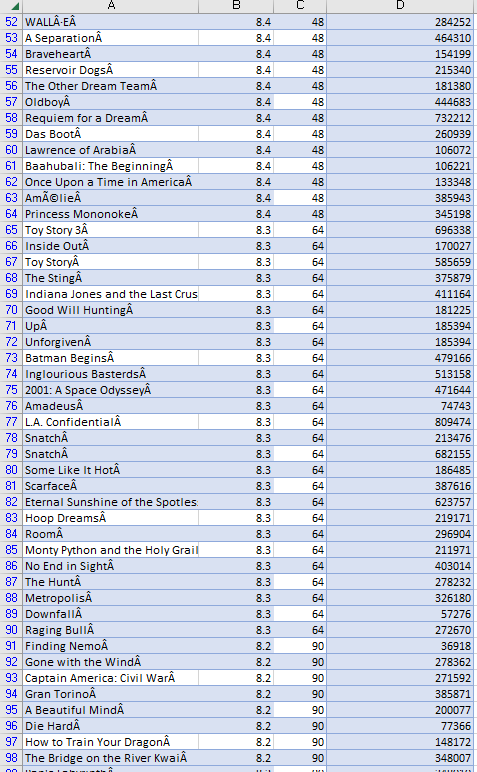
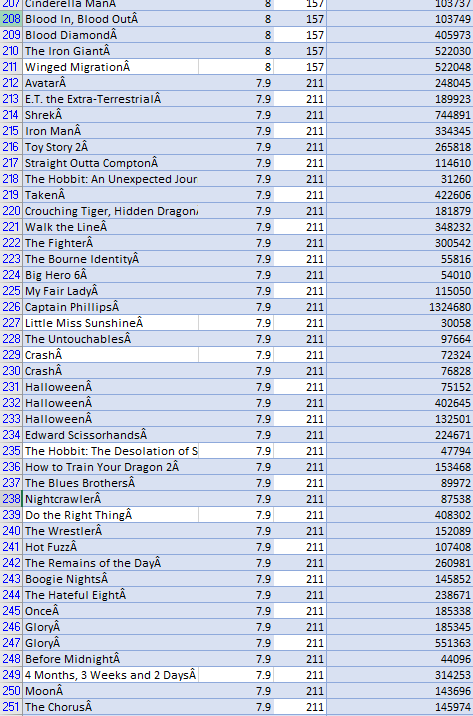
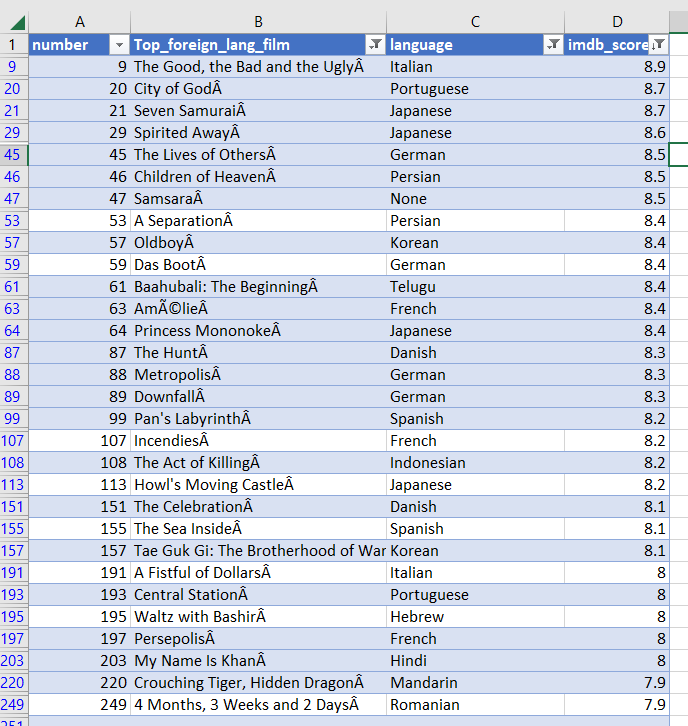


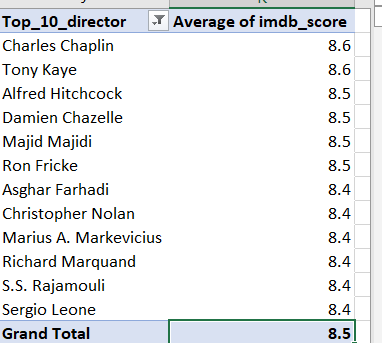


1. Movies with highest profit: Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.  
   Your task: Find the movies with the highest profit?

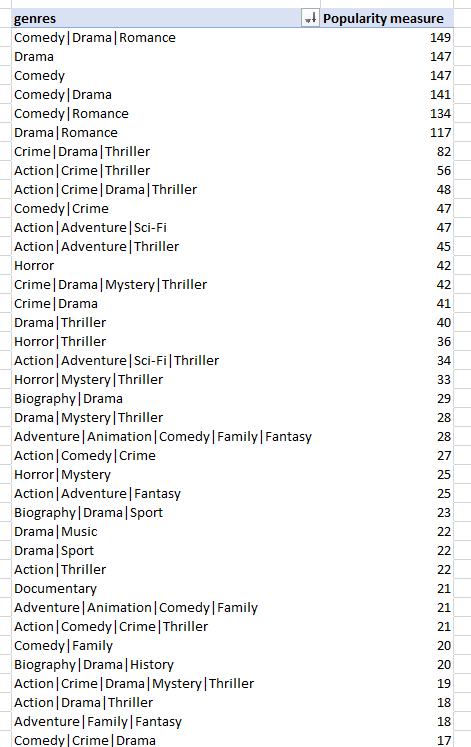
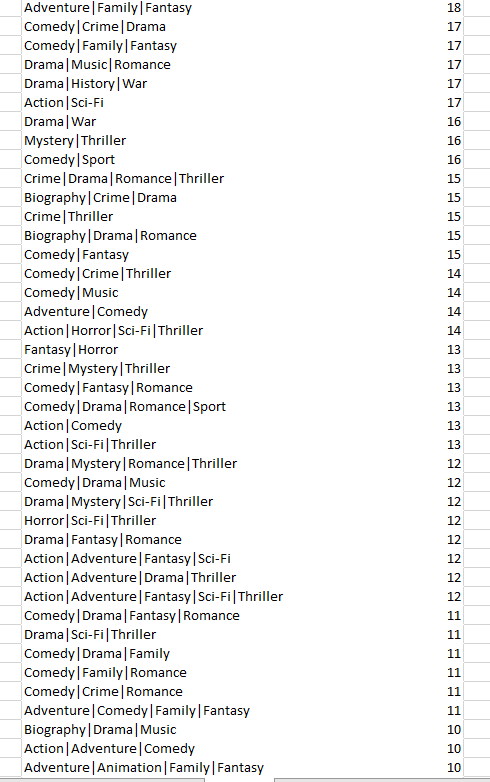




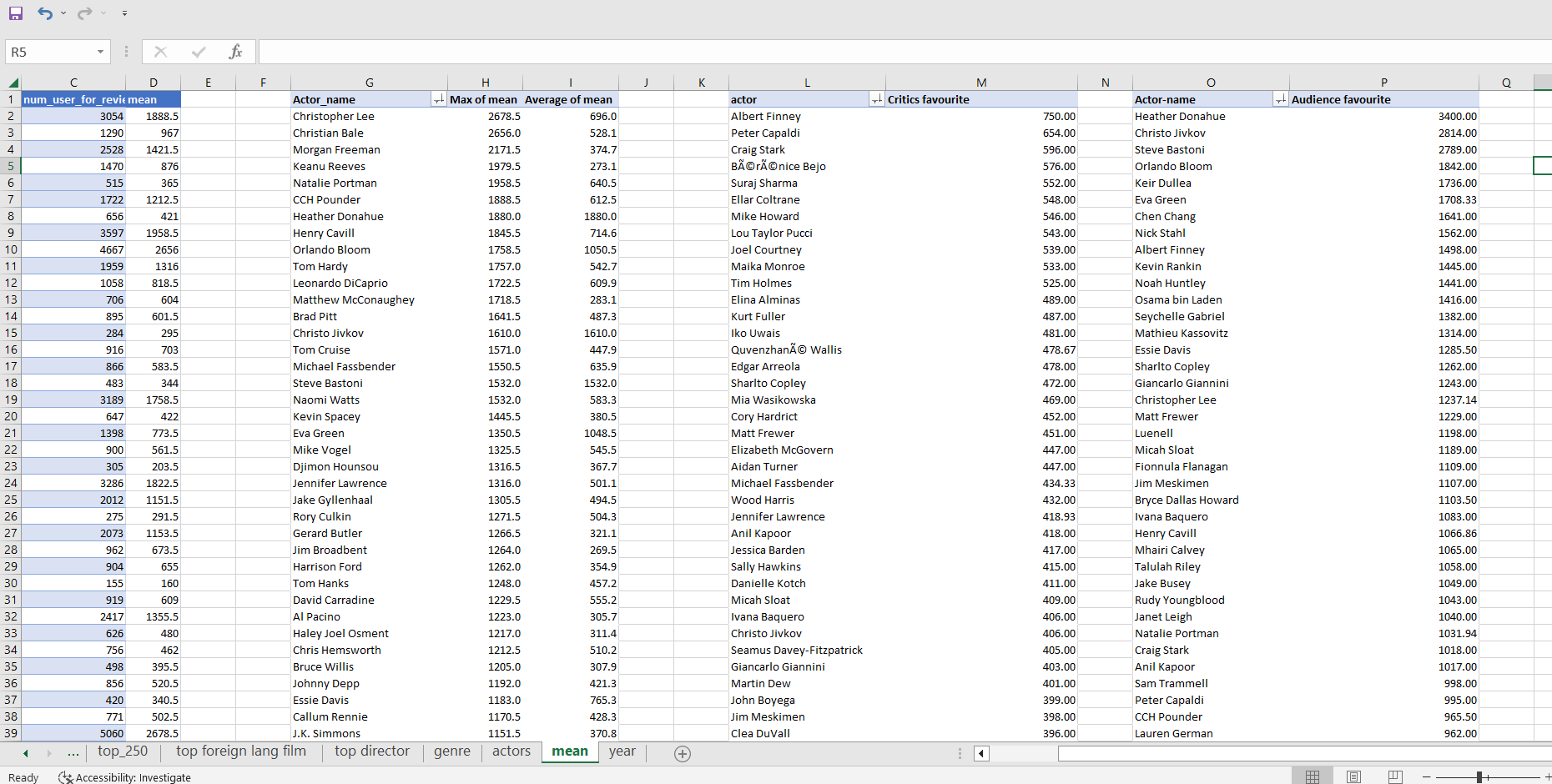
1. Top 250: Create a new column IMDb\_Top\_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb\_score). Also make sure that for all of these movies, the num\_voted\_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating 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.   
2. Best Directors: TGroup the column using the director\_name column.  
     
   Find out the top 10 directors for whom the mean of imdb\_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.  
   Your task: Find the best directors

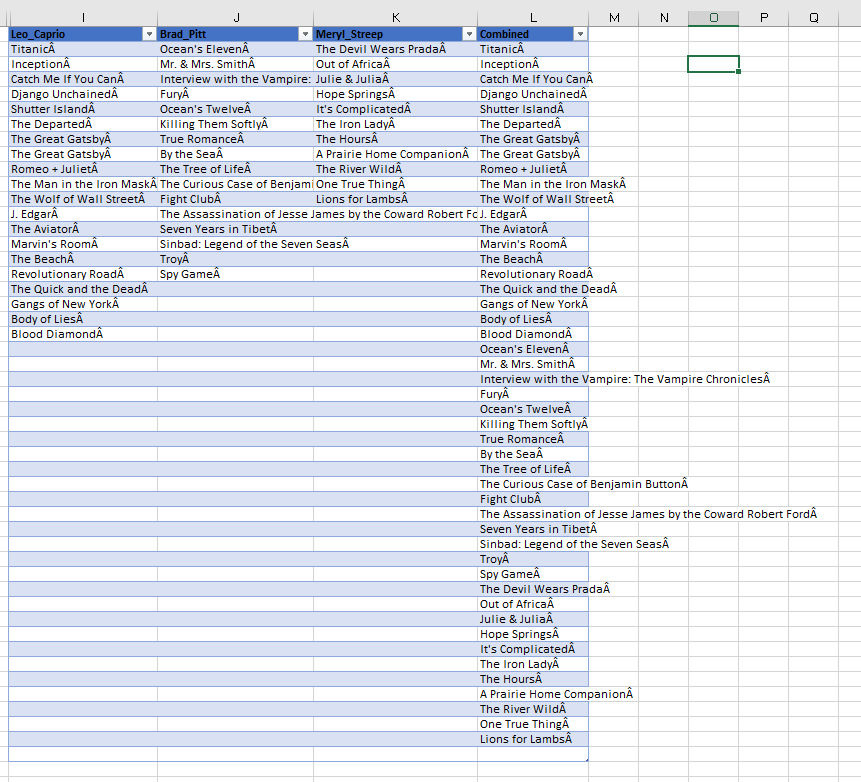


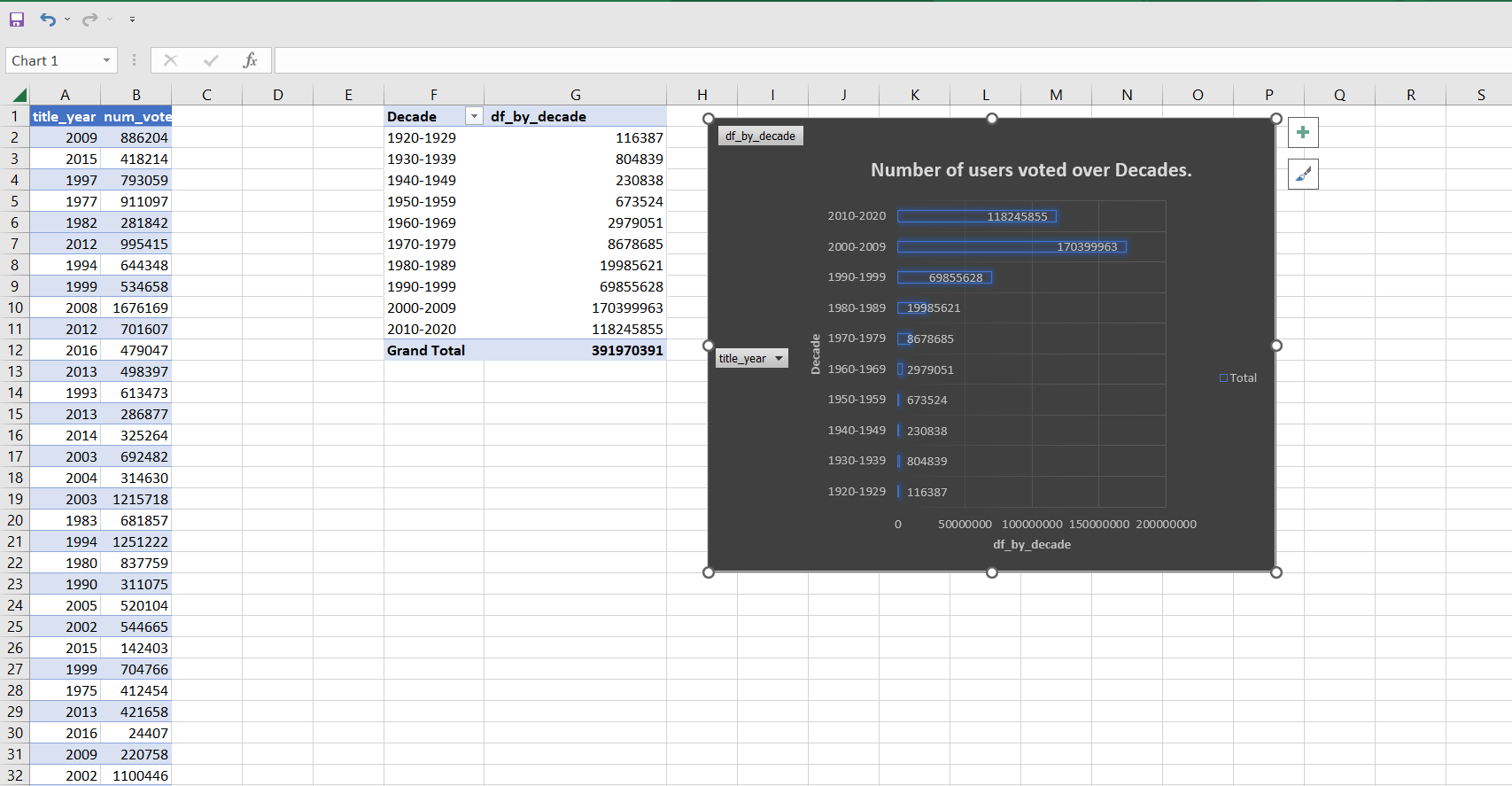
1. Popular Genres: Perform this step using the knowledge gained while performing previous steps.  
   Your task: Find popular genres

1. 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.







##### Project-6:-Bank Loan Case Study.

###### 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 the company receives a loan application, the company has to decide for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision:

* If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
* If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

* The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample
* All other cases: All other cases when the payment is paid on time.

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

1. **Approved:**The company has approved loan application
2. **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.
3. **Refused:**The company had rejected the loan (because the client does not meet their requirements etc.).
4. **Unused Offer:**Loan has been cancelled by the client but on different stages of the process.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

###### Business Objectives:

It aims to **identify patterns** which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. **Identification of such applicants using EDA** is the aim of this case study.  
  
In other words, the company wants to understand the **driving factors** (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.  
  
To develop your understanding of the domain, you are advised to independently research a little about **risk analytics** – understanding the types of variables and their significance should be enough).

###### Data Understanding:

Download the Dataset using the link given under dataset section on the right.

1. **`application\_data.csv`**contains all the information of the client at the time of application.  
   The data is about wheather a client has payment difficulties.
2. **`previous\_application.csv`**contains information about the client’s previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
3. **`columns\_descrption.csv`**is data dictionary which describes the meaning of the variables.

You are required to provide a detailed report for the below data record mentioning the answer to the questions that follows:

* Present the overall approach of the analysis. Mention the problem statement and the analysis approach briefly
* Indentify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)  
  *Hint: Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach.*
* Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.
* Identify if there is data imbalance in the data. Find the ratio of data imbalance.  
  *Hint: Since there are a lot of columns, you can run your analysis in loops for the appropriate columns and find the insights.*
* Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.
* Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.
* Include visualizations and summarize the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases.

Sol:- Overall Approach of Analysis with Problem Statement: Two csv files for different applicants for accessing loans are given. We will first import it into excel. The goal is to analyze the historical data to understand a borrower’s creditworthiness or to assess the risk involved in the granting of a loan. The result of the analysis help banks and financial institutions evaluate their risk and those of their customers. The company will utilize this knowledge for its portfolio and risk assessment.

# CSV files will be checked for any unnecessary data and unwanted columns/rows, and will be cleaned/removed if necessary. Will check for outliers, if any, to find if there is skewness in the given columns which would affect the final visualization and insight. Imbalance in data will be checked. Different types of analysis will be done to understand the relationships between different variable to find the Driving Factors. Different visualizations will be observed to understand the relationships.   EDA – Data Cleaning (Identifying rows/columns with missing data, Identifying Unnecessary columns and cleaning if required).  Applicant\_data.csv(Before Cleaning) Columns 122, Rows 307512   After Cleaning:-Columns 42 + 4 Columns, Rows 306216 There are date columns with negative values, those needs to be standardized.There are columns having more than 40% null data. These need to be removed. There are more than 50 unwanted columns or columns not desirable for our analysis so I will remove them from the dataset. There are columns with null values less than 40%. They can be treated in 2 ways. I can delete those columns but then I might lose some important information required for my analysis. I can retain it but then I will have to do treatment. If I impute them, I will introduce bias. The decision to delete or retain basically depends on the Understanding of the problem statement, the usefulness of the variable, total size of available data. Here it seems that those columns can be removed. So, I have removed them. There are still some columns will very little missing values which will be treated if necessary or left as it is.

# AMT\_ANNUITY has a smaller number of null values (12). It can be imputed with mean. If it has an outlier which is very large then the null values can be computed with Median.EXIT\_SOURCE\_2 has 656 null values which is also quite small as compared to total number of rows. Can be imputed with 0.OCCUPATION\_TYPE has 96005 null values. Can be imputed by the category which is the most popular (Mode), i.e., ‘Labourers’.

# 

# Previous\_Application.csv(Before Cleaning) Columns 37 ,Rows 1048576 There are columns with more than 40% null values and few unnecessary columns that need to be removed. After cleaning:-Columns-15, Rows- 1

# Columns with null values less than 40% are present. They can be treated in 2 ways. I can delete those columns but then I might lose some important information required for my analysis. I can retain it but then I will have to do treatment. If I impute them, I will introduce bias. The decision to delete or retain basically depends on the Understanding of the problem statement, the usefulness of the variable, total size of available data. Here it seems that those columns can be removed So, I have removed them. There are still some columns will very little missing values which will be treated if necessary or left as it is.

# 

# EDA – Identifying Outliers.  Applicatiton\_data.csv

# 

# Outliers are those values which are Less Than LOWER BOUND and Also Greater Than UPPER BOUND. Box and Whisker chart are shown in next page.

# In the above data of the next four Numeric columns, it can be seen that they contain outliers as well Except DAYS\_BIRTH and DAYS\_PUBLISH column. It can also be confirmed by the chart below.

# 

# 

# DATA IMPUTING BEFORE ANALYSIS AMT\_ANNUITY – Imputing with Mean. OCCUPATION\_TYPE – Imputing with ‘Labourers’

# Explain the results of univariate, segmented variate and bivariate analysis. Univariate Numerical Analysis.

# 

# 

# Univariate Categorical Analysis:-

# Segmented Univariate Analysis:-

# 

# 

# Find the top correlations for Client with payment difficulties and all other case.

# 

# Insights:-After performing the analysis, we can rectify whether a client will repay the loan or not. Also, the people who are likely to face problem in loan repayment are labourers. Also, people with Secondary /secondary special education might face problem in loan repayment. Moreover, those who are living in house/apartment are facing difficulty in loan repayment (may be because of extra home loan, EMIs and so on). •NAME\_EDUCATION\_TYPE: Academic degree has less defaults.

# NAME\_EDUCATION\_TYPE: People with Lower Secondary & Secondary education • NAME\_INCOME\_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot. •REGION\_RATING\_CLIENT: People who live in Rating 3 has highest defaults. • OCCUPATION\_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as their default rate is huge.

# • NAME\_INCOME\_TYPE: Student and Businessmen have no defaults. • REGION\_RATING\_CLIENT: RATING 1 is safer. • ORGANIZATION\_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%. • DAYS\_BIRTH: People above age of 50 have low probability of defaulting • DAYS\_EMPLOYED: Clients with 40+ year experience having less than 1% default rate.

# • AMT\_INCOME\_TOTAL: Applicant with Income more than 700,000 are less likely to default.•NAME\_CASH\_LOAN\_PURPOSE: Loans bought for Hobby, buying garage are being repaid mostly. • CNT\_CHILDREN: People with zero to two children tend to repay the loans.• CODE\_GENDER: Men are at relatively higher default rate • NAME\_FAMILY\_STATUS: People who have civil marriage or who are single default a lot. •

##### Project-7:-XYZ Ads Airing Report Analysis.

###### Link for dataset analyzed: https://docs.google.com/spreadsheets/d/1Y9-wkXwpgPs7PLKZvSTSpsSG81EWgebn/edit?usp=sharing&ouid=110945409935767923714&rtpof=true&sd=true

###### Description:

For your Final Project, we are providing you with a dataset having different TV Airing Brands, their product, their category. Dataset includes the network through which Ads are airing,  types of network like Cable/ Broadcast and the show name also on which Ads got aired. You can also see the data of Dayparts, Time zone and the time & date at which Ads got aired. IT also includes other data like Pod Position (the lesser the valuable), duration for which Ads aired on screen, Equivalent sales &, total amount spent on the Ads aired.

###### 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.

Case Study Objectives:

Attached is the dataset of TV Ad Airings of some brands from the Automobile category. Use this data to answer the following:

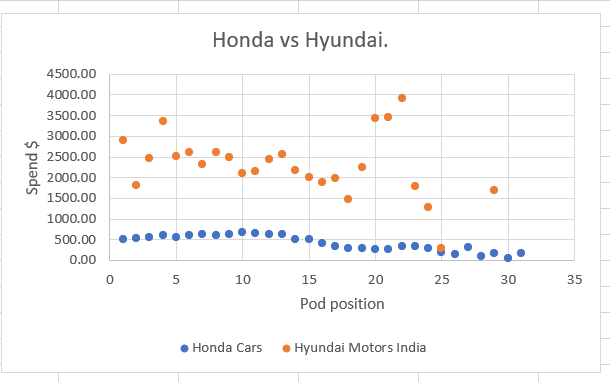
1. What is Pod Position? Does the Pod position number affect the amount spent on Ads for a specific period of time by a company? (Explain in Details with examples from the dataset provided)

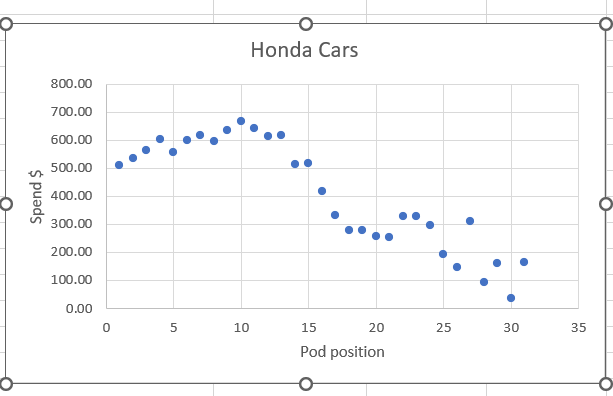
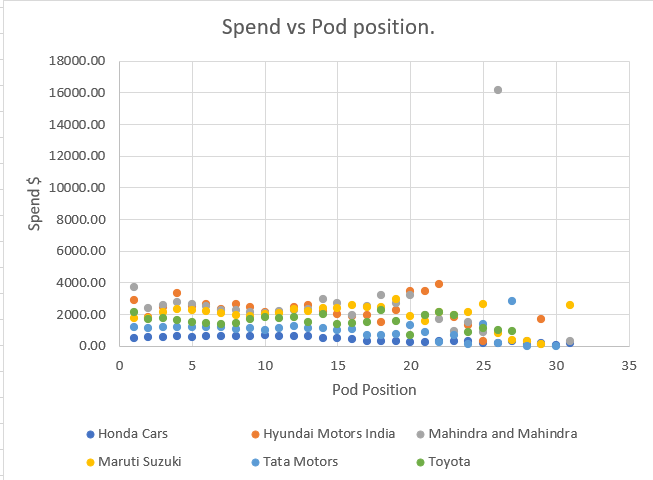
* Solution: Ad Pod is a term used in connection with TV advertising to specify multiple ads sequenced together and played back-to-back within a single ad break. ‘Pod Position’ is basically the Sequence in which the ads are played on TV, Like which ad plays in the beginning followed by which one and so on.

Observation:-

1.General condition is that as the Pod Position Increases, Price increases up to a certain position, then dips drastically as the position number increases further.

2. When looked for Honda’s data, it is observed that amount spent increases from Pod Position 1 to 10 but then it starts to decrease drastically as the Position further Increases. 3. When looked for Hyundai Data Price fluctuates for a certain period ,then drastically increases and then drastically decreases to the end.



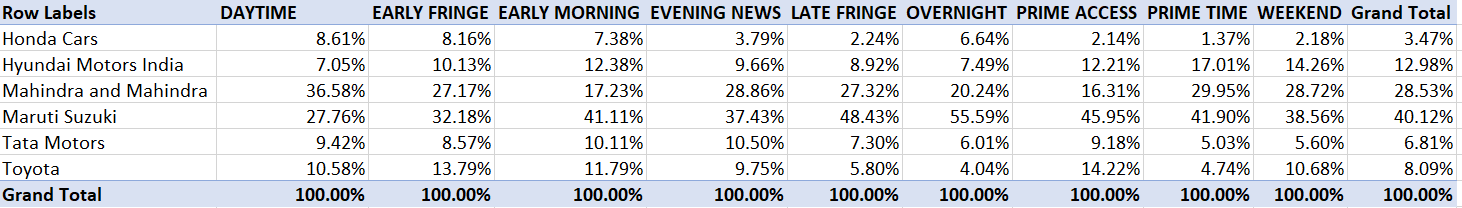
 

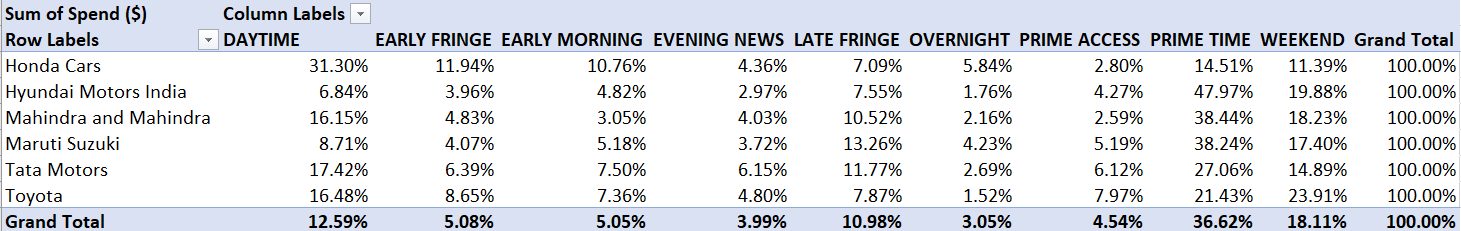
1. What is the share of various brands in TV airings and how has it changed from Q1 to Q4 in 2021?

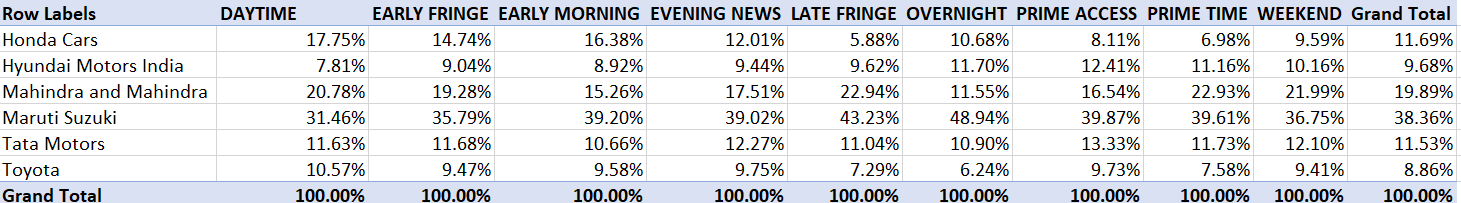
Solution:-Shares can be calculated in many terms. Here the share of Spends and Count of Ads has been taken into consideration. 1. Honda has overall decreased it percentage spend as the quarter progressed but the share in number of ads run doesn’t have a definite pattern. 2. Mahindra and Mahinda has Increased is spend percentage till Q3 but decreased in Q4, but the share in the number of ads run has overall decreased from Q1 to Q4. 3. Rest of the comparison can be seen in the table above. 4. Maruti Suzuki has the Maximum Share of spends in Q1 of total. The same goes for total share in ads count in Q1. 5. M&M have 40+ % spends share with 38+ ads count share, Suzuki has 27+% spends share but only around 20% ads count. This means the ads were costlier for Maruti Suzuki as compared to Mahindra and Mahinda. The trend is same for all Quarters.

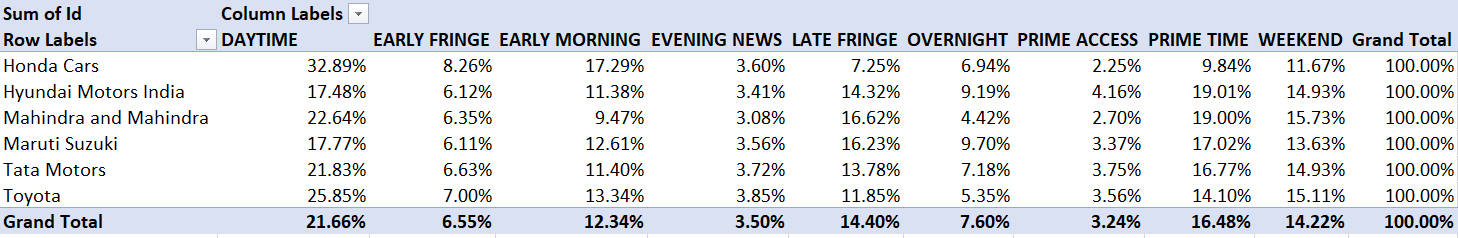


1. Conduct a competitive analysis for the brands and define advertisement strategy of different brands and how it differs across the brands.

Percentage of Sum spent - 



Percentage of Ad count- 

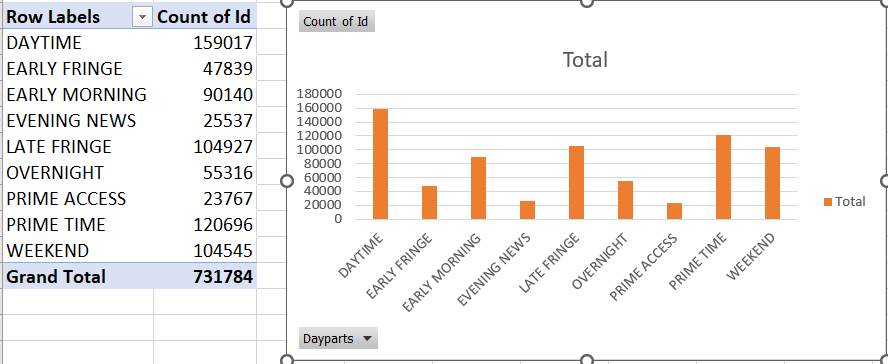


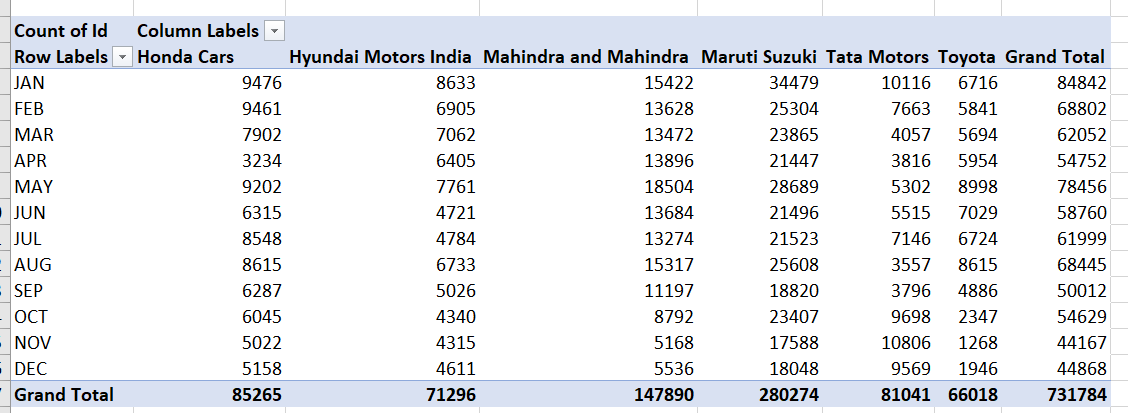
Insights:-

* Toyota doesn’t have a specific preference when it comes to selecting a specific time to show its ads. The ad percentage for Toyota is almost Evenly distribute across the Dayparts for ad count share but the money spent is different during the same period.
* Honda Spends the maximum(32.88%) share of money for ads during the DAYTIME Dayparts and minimum share during PRIME ACCESS(2.23%) Dayparts while of all the spends done during DAYTIME, Honda accounts for only 8.61% of the Total. 2. Maruti Suzuki bears 55.59% of total spends done during OVERNIGHT Dayparts and it also has the maximum share for Number of Ads in the OVERNIGHT Dayparts(49.22%). This shows Maruti’s strategy for showing their ads. 3. Mahindra has 36.5% share of the total amount spend during DAYTIME but gets only 21.1% share of Ads count from DAYTIME.
* Suzuki has the maximum percentage of spend across all Dayparts. Suzuki on the other hand has 31.32% share in ad count during DAYTIME with only 27.76% Spends share during DAYTIME, clearly stating that the cost of ads for Mahinda is more than that of Suzuki. This also means Mahinda is more preferred by People than Maruti Suzuki.

1. Mahindra and Mahindra wants to run a digital ad campaign to complement its existing TV ads in Q1 of 2022. Based on the data from 2021, suggest a media plan to the CMO of Mahindra and Mahindra. Which audience should they target? \*Assume XYZ Ads has the ad viewership data and TV viewership for the people in India.   
   P.S. Brownie points for any additional actionable insights you can draw from the dataset.

Mahindra & Mahindra should Target the Audience of DAYTIME and PRIMETIME viewers in the First Quarter of 2022. As observed from the Comparative Analysis Table Mahinda followed the same trend in 2021.

 Maximum number of Ads are played in the Month of January and Suzuki tops the list with 34478 ads. Mahindra’s got less than half of it. In the 1 st Quarter of 2022 Mahindra needs to increase the number of Ads it plays to get better results.

 Final Insights:-  
1. How is Ads Airing Analysis done.  
 2. Concept of Airing time, its significance.   
3. Pod Position and it significance.   
4. Excel visualization to understand Ads Airing Analysis.  
5. Evaluate formula.

Conclusion:-

Learnt how companies plan their advertisements on TV and how data analytics can be used to provide insights for better working efficiency.

##### Project-8:-ABC Call Volume Trend Analysis.

###### Link for dataset analyzed:-

###### https://docs.google.com/spreadsheets/d/1\_ma--e6is0Nchp4Kc0iDZq96vkWBj3in/edit?usp=sharing&ouid=110945409935767923714&rtpof=true&sd=true

###### Description:

For you final project we are providing you 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.  
  
Let’s look at some of the most impactful AI-empowered customer experience tools you can use today:  
  
Interactive Voice Response (IVR), Robotic Process Automation (RPA), Predictive Analytics, Intelligent Routing  
  
In a Customer Experience team there is a huge employment opportunities for Customer service representatives A.k.a. call centre agents, customer service agents. Some of the roles for them include: Email support, Inbound support, Outbound support, social media support.  
  
Inbound customer support is defined as the call centre which is responsible for handling inbound calls of customers. Inbound calls are the incoming voice calls of the existing customers or prospective customers for your business which are attended by customer care representatives. Inbound customer service is the methodology of attracting, engaging, and delighting your customers to turn them into your business' loyal advocates. By solving your customers' problems and helping them achieve success using your product or service, you can delight your customers and turn them into a growth engine for your business.

###### 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.

Case Study Objectives:

Attached is the dataset of Inbound calls of a ABC company from the insurance category. Use this data to answer the following:

# Approach:- The dataset consists of various columns such as queue time, Agent\_Id, Agent\_Name, IVR duration ,etc which are not useful for our analysis.

# In this project we are not cleaning the data i.e, we did not delete any of the redundant columns.

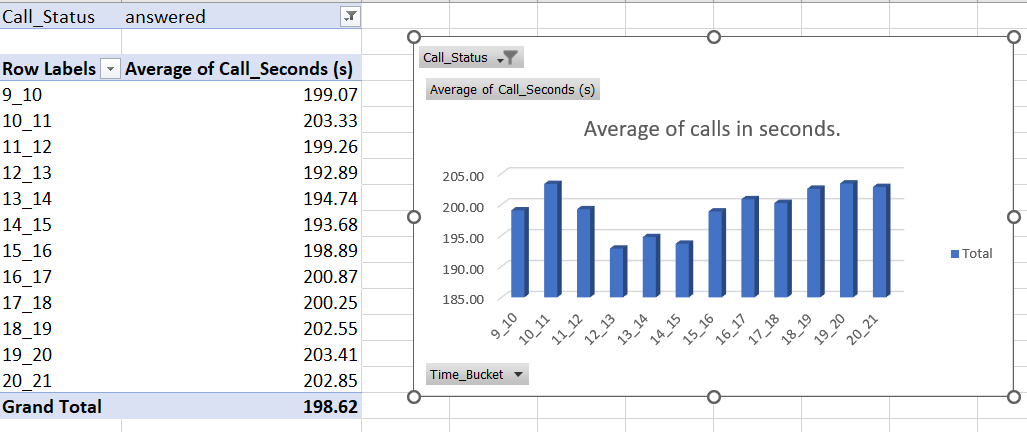
# Analysis done on few important columns such customer\_phone\_number, time\_bucket,etc.

# We initiated visualization using charts such as column chart for the insights we got through our analysis.

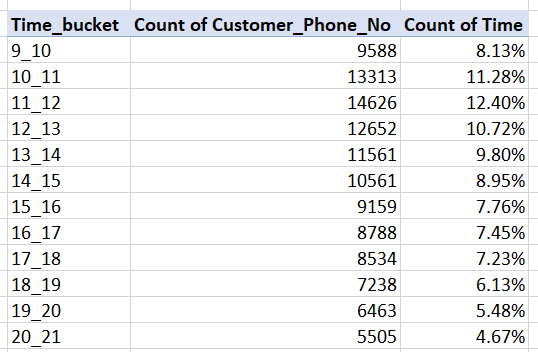
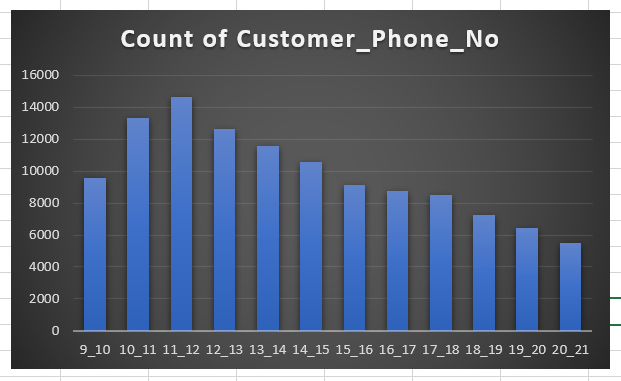
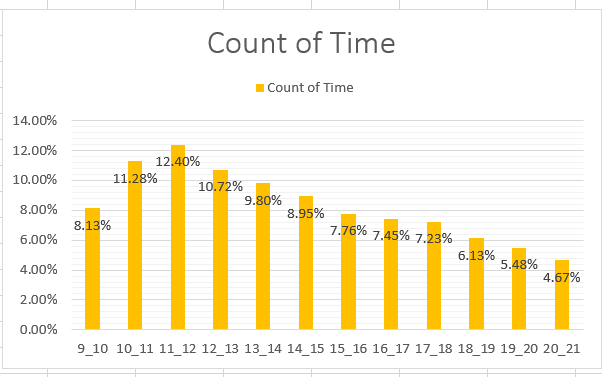
Assumption: An agent work for 6 days a week; On an average total unplanned leaves per agent is 4 days a month; An agent total working hrs is 9 Hrs out of which 1.5 Hrs goes into lunch and snacks in the office. On average an agent occupied for 60% of his total actual working Hrs (i.e 60% of 7.5 Hrs) on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

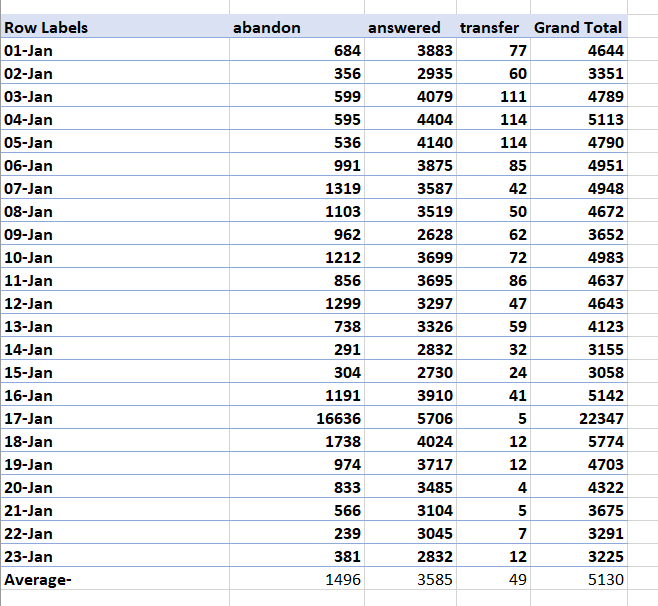
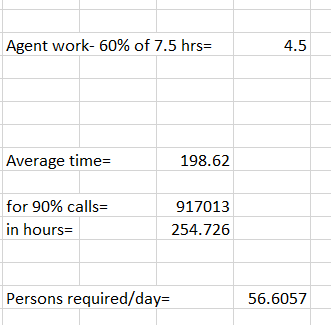
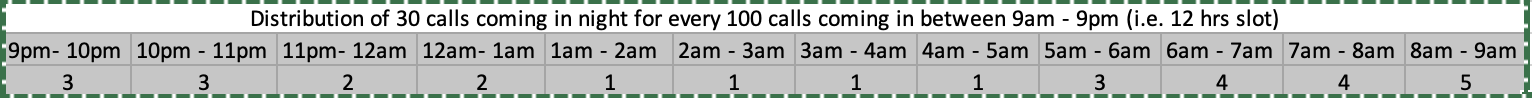
* A manpower plan required to reduce the abandon rate to 10%. The effective working days of an employee = 20 X 7/28 = 5 and The effective working hours (60% occupation time of agent) = 7.5 X 0.6 = 4.5hours.
* total working hours = 9
* on floor working hour = 7.5
* days worked in 1 week = 5
* total time spent = 4.5
* Time requirement to answer 90% calls = (Total average calls X 198.62X 0.9)/3600 = 254.726
* Total working person required per day = Requirement to answer 90% calls/ Total time spent Total working person required per day to take 90% of calls 57

1. Calculate the average call time duration for all incoming calls received by agents (in each Time\_Bucket).

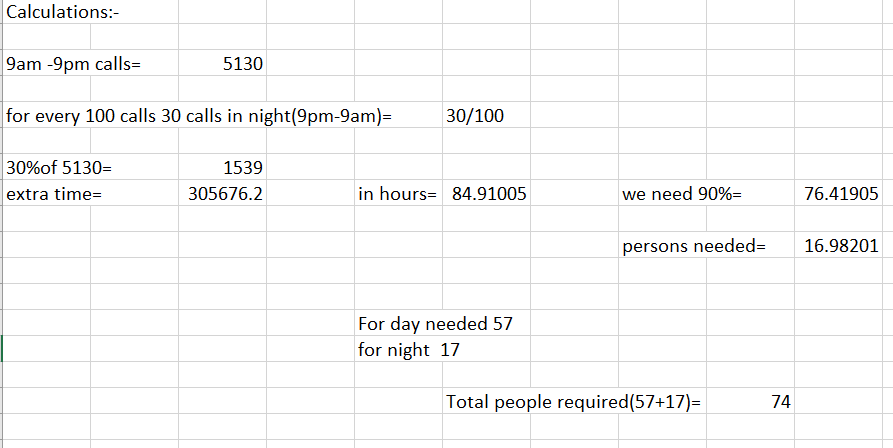
The average call duration for each time bucket 

1. 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, …..)

1. 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.)  
2. Let’s say customers also call this ABC insurance company in night but didn’t get answer as there are no agents to answer, this creates a bad customer experience for this Insurance company. Suppose every 100 calls that customer made during 9 Am to 9 Pm, customer also made 30 calls in night between interval [9 Pm to 9 Am] and distribution of those 30 calls are as follows:  
     
     
     
   Now propose a manpower plan required during each time bucket in a day. Maximum Abandon rate assumption would be same 10%.

No. of person required (total extra hours req./ working time for one agent) 17   
The actual number of agents required to attend 90% of the extra calls incoming at night is: The total number of agents required to attend 90% of all the incoming calls in a day is 57+17=74.



Final Insights :-

* ABC call volume trend analysis.
* Learnt how to use pivot tables.
* Excel visualization.

Conclusion:-

* Learnt how Customer care support teams process and data analytics can be used to improve working efficiency of company ,how the workload can be maintained for the employees.

Learnings from all projects above includes:-

* How Data analytics can be used in real world scenario.
* How social media platform like Instagram functions.
* How to use business analytics.
* How to use quartiles and IQR to find outliers and
* Calculate retention ratios and their enagement in weekly/ daily/yearly basis. Cohort analysis.
* Excel visualization.
* Create class interval in Excel.
* Seggregating data.
* Knowledge on how banks function to approve loans for people.
* How Ads airing affects the sales/development of a company product.
* Pod position and its significance.
* Concept of Call volume analysis,Call status,IVR duration,etc.
* Statistics concepts like mean,mode,median,etc.
* Probability utilization.

# Conclusion:-

# Got the Real-world industry experience on how companies process the data to produce insights and predict outcomes.

# Learnt various tools such as Microsoft Excel, MySQL,etc.

# Knowledge of Exploratory Data Analysis was used to solve real world examples.

# Theoritical and Practical Knowledge was blended in way to provide better opportunity to learn new concepts.

# Data Analytics domain , Critical analysis, Case studies, Business studies were the key learnings from the projects.

# THANK YOU.