PROJECT REPORT - Operation Analytics and Investigating Metric Spike

Project Description

- I am working for a company like Microsoft as Data Analyst Lead and I am provided with different data-sets and tables. That data was collected by different teams within my company.
- I worked closely with the ops(operations) team, support team, marketing team, etc and helped them derive insights out of the data they collected.
- I did the entire analysis of the company's end-to-end operations to answer the following set of questions asked by different departments within the company.

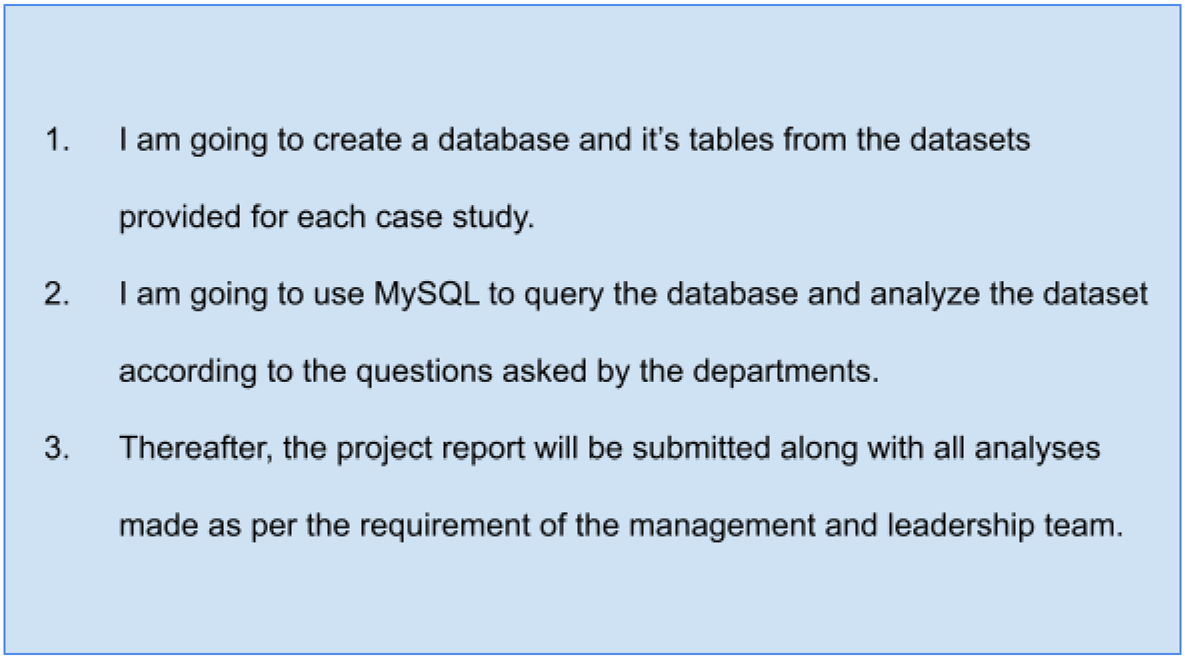
Case Study 1 (Job Data)

Questions asked by Departments	***KEY FINDINGS EXPECTED***
Calculate the number of jobs reviewed per hour per day for November 2020?	*Amount of jobs reviewed over time.*
● Calculate the number of events happening per second which is called as throughput. ● Also find a 7 day rolling average of throughput metric. ● For throughput, do you prefer daily metric or 7-day rolling and why?	*7 day rolling average of throughput*
Calculate the percentage share of each language in the last 30 days?	*Percentage share of each language for different contents*
Let's say you see some duplicate rows in the data. How will you display duplicates from the table?	*Duplicate rows that have the same value present in them.*

Case Study 2 (Investigating metric spike)

Questions asked by Departments	***KEY FINDINGS EXPECTED***
To measure the activeness of a user. Measuring if the user finds quality in a product/service.	*Weekly user engagement*

Calculate the user growth for a product?	*Amount of users growing over time*
Users get retained weekly after signing-up for a product. Calculate the weekly retention of users-sign up cohort?	*Weekly retention of users-sign up cohort*
To measure the activeness of a user. Measuring if the user finds quality in a product/service weekly.	*Weekly user engagement per device*
Find out how users are engaging with the email service.	*Email engagement metrics*



1. I am going to create a database and it's tables from the datasets provided for each case study.
2. I am going to use MySQL to query the database and analyze the dataset according to the questions asked by the departments.
3. Thereafter, the project report will be submitted along with all analyses made as per the requirement of the management and leadership team.

Approach

After downloading the dataset files containing tables (in comma separated value or csv format with .csv file extension) for each of the two case studies given, I started creating those tables on MySql Workbench 8.0.32 (running MySql Server 8.0.32) using commands for creating the ***ops_analytics*** database(the CREATE DATABASE command) followed by the USE command.

Then, I started importing the downloaded csv dataset files with the help of Table Data Import Wizard on MySql Workbench by mentioning the filepath of csv dataset files downloaded to my local machine and also started naming the tables during the import process.

I created one table job_data for case study 1 and three tables(users, events, email_events) for case study 2, a total of 4 tables using Table Data Import Wizard application of MySql Workbench.

Then, I started exploring the required tables of database ***ops_analytics*** by writing SQL queries for each of the problem statements given.

I read, analyzed the problem statement & tried to find the required table(s) of the ops_analytics database that I may need to work with and also the particular column(s)/attributes to select from the relevant tables.

Tech-stack Used

Software	Version	Purpose of using
MySQL Community Edition, Windows (x86, 32-bit) (**installer file - mysql-installer-community-8.0.32.0.msi**)	8.0.32.0	This installer provides all MySQL Softwares that are needed, including MySQL Server and Workbench)
MySQL Server	8.0.32	Server provides a Relational Database Management System which has querying and connectivity features. It was possible to query with SQL and connect to the MySQL server with the help of this software
MySQL Workbench	8.0.32	Provides an SQL editor to write queries to interact with the database for dataset analysis.
Google Docs	Web version of GDrive	Writing project report in detail
Google Sheets	Web version of GDrive	For creating visualizations from the csv file which are exported from MySql Workbench

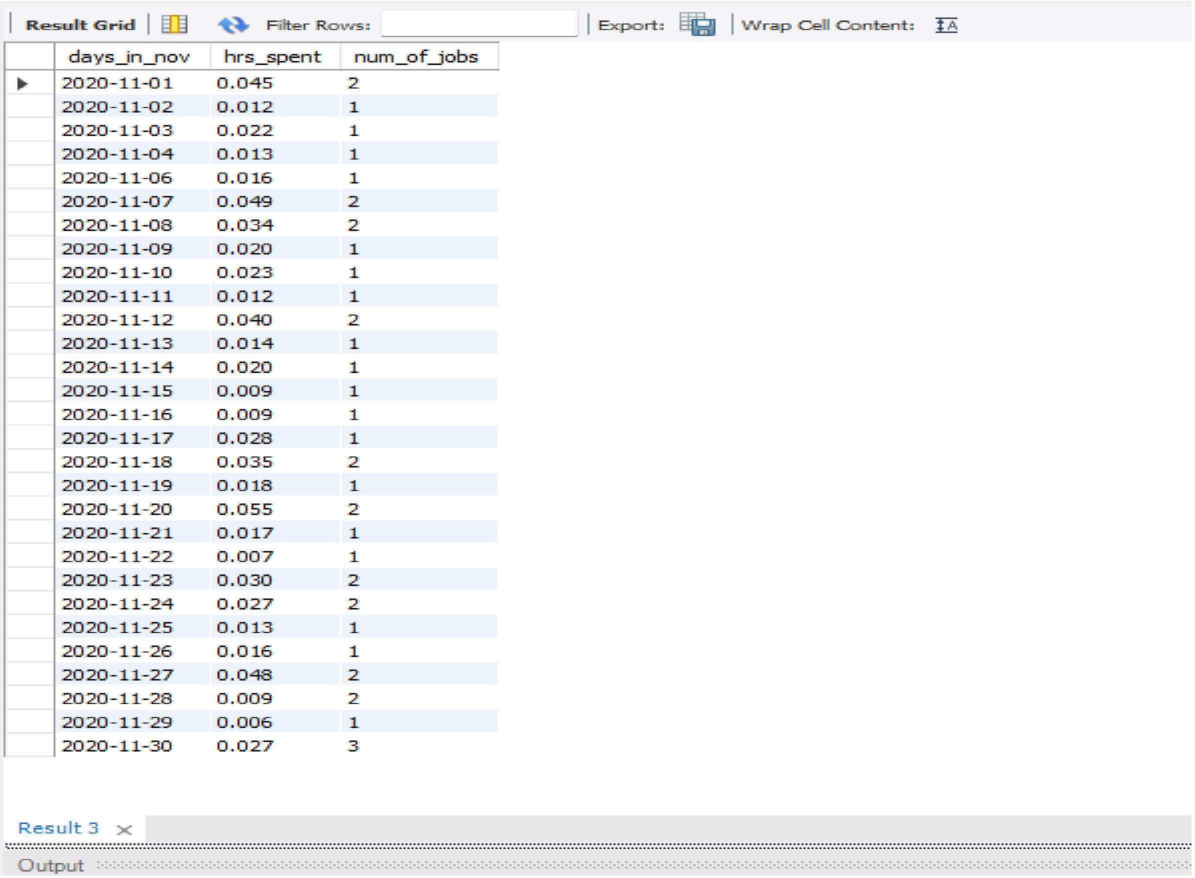
Insights

01. Case Study 1 (Job Data)

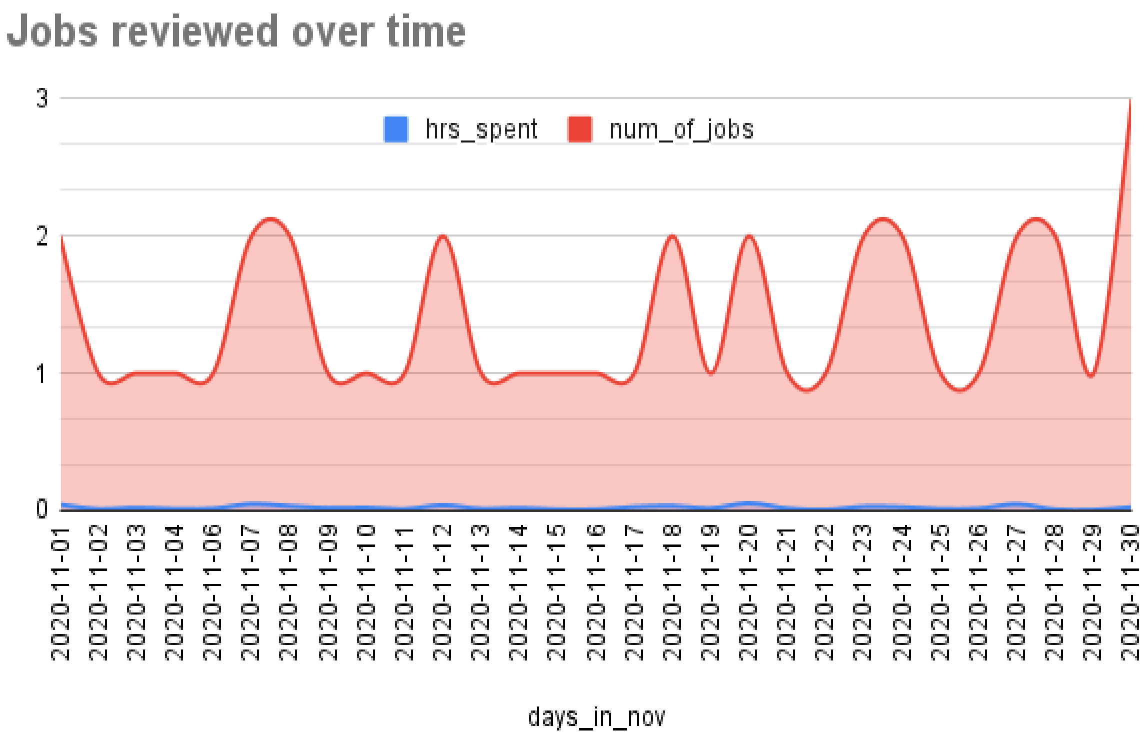
a. Amount of jobs reviewed over time

i. Calculated the number of jobs reviewed over time for the whole month of November 2020.

ii. I have found the reviewed jobs count with hours spent for each date of November after the analysis for this particular question asked.



days_in_nov	hrs_spent	num_of_jobs
2020-11-01	0.045	2
2020-11-02	0.012	1
2020-11-03	0.022	1
2020-11-04	0.013	1
2020-11-06	0.016	1
2020-11-07	0.049	2
2020-11-08	0.034	2
2020-11-09	0.020	1
2020-11-10	0.023	1
2020-11-11	0.012	1
2020-11-12	0.040	2
2020-11-13	0.014	1
2020-11-14	0.020	1
2020-11-15	0.009	1
2020-11-16	0.009	1
2020-11-17	0.028	1
2020-11-18	0.035	2
2020-11-19	0.018	1
2020-11-20	0.055	2
2020-11-21	0.017	1
2020-11-22	0.007	1
2020-11-23	0.030	2
2020-11-24	0.027	2
2020-11-25	0.013	1
2020-11-26	0.016	1
2020-11-27	0.048	2
2020-11-28	0.009	2
2020-11-29	0.006	1
2020-11-30	0.027	3

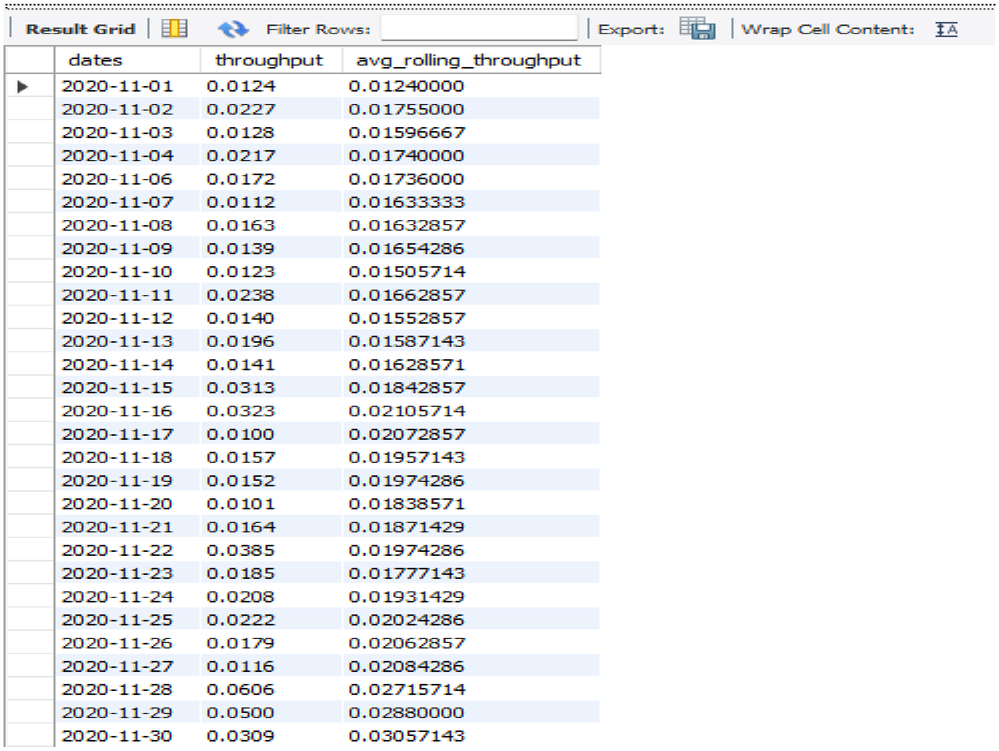


b. <u>7 day rolling average of throughput</u>

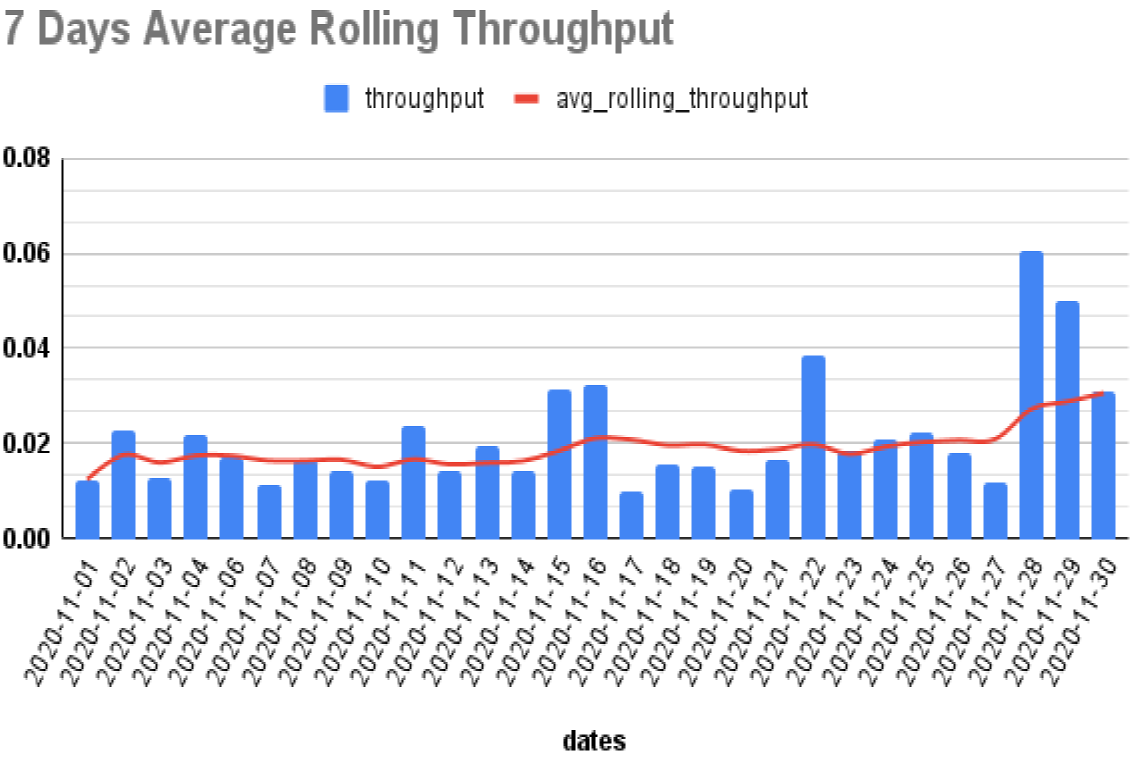
i. I found the metric value of throughput which is the number of events happening per second.

ii. Then I calculated the average throughput by utilizing the AVG() function.

iii. I utilized the MySql window function where I used the ROWS BETWEEN clause to limit the rows within the window created that has the average of throughput values. By writing ROWS BETWEEN 6 PRECEDING AND CURRENT ROW, I am getting the 7 day's average throughput value within the window created.



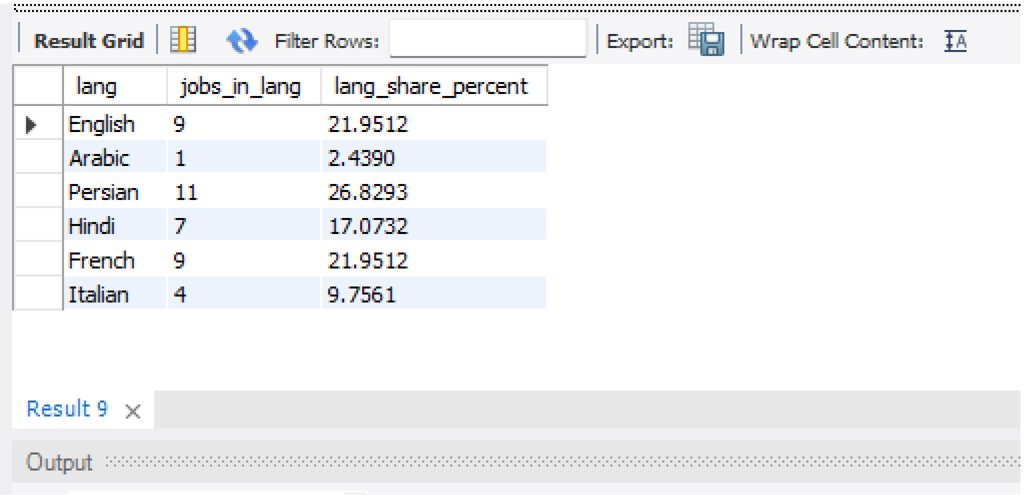
dates	throughput	avg_rolling_throughput
2020-11-01	0.0124	0.01240000
2020-11-02	0.0227	0.01755000
2020-11-03	0.0128	0.01596667
2020-11-04	0.0217	0.01740000
2020-11-06	0.0172	0.01736000
2020-11-07	0.0112	0.01633333
2020-11-08	0.0163	0.01632857
2020-11-09	0.0139	0.01654286
2020-11-10	0.0123	0.01505714
2020-11-11	0.0238	0.01662857
2020-11-12	0.0140	0.01552857
2020-11-13	0.0196	0.01587143
2020-11-14	0.0141	0.01628571
2020-11-15	0.0313	0.01842857
2020-11-16	0.0323	0.02105714
2020-11-17	0.0100	0.02072857
2020-11-18	0.0157	0.01957143
2020-11-19	0.0152	0.01974286
2020-11-20	0.0101	0.01838571
2020-11-21	0.0164	0.01871429
2020-11-22	0.0385	0.01974286
2020-11-23	0.0185	0.01777143
2020-11-24	0.0208	0.01931429
2020-11-25	0.0222	0.02024286
2020-11-26	0.0179	0.02062857
2020-11-27	0.0116	0.02084286
2020-11-28	0.0606	0.02715714
2020-11-29	0.0500	0.02880000
2020-11-30	0.0309	0.03057143



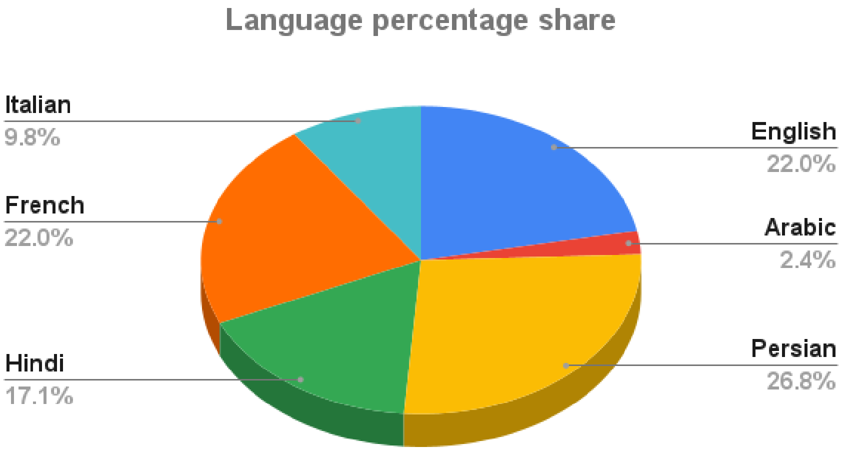
c. Percentage share of each language for different contents

i. I got the number of jobs in each language from the dataset table.

ii. After that, I calculated the share of percentage of each language present in the table.



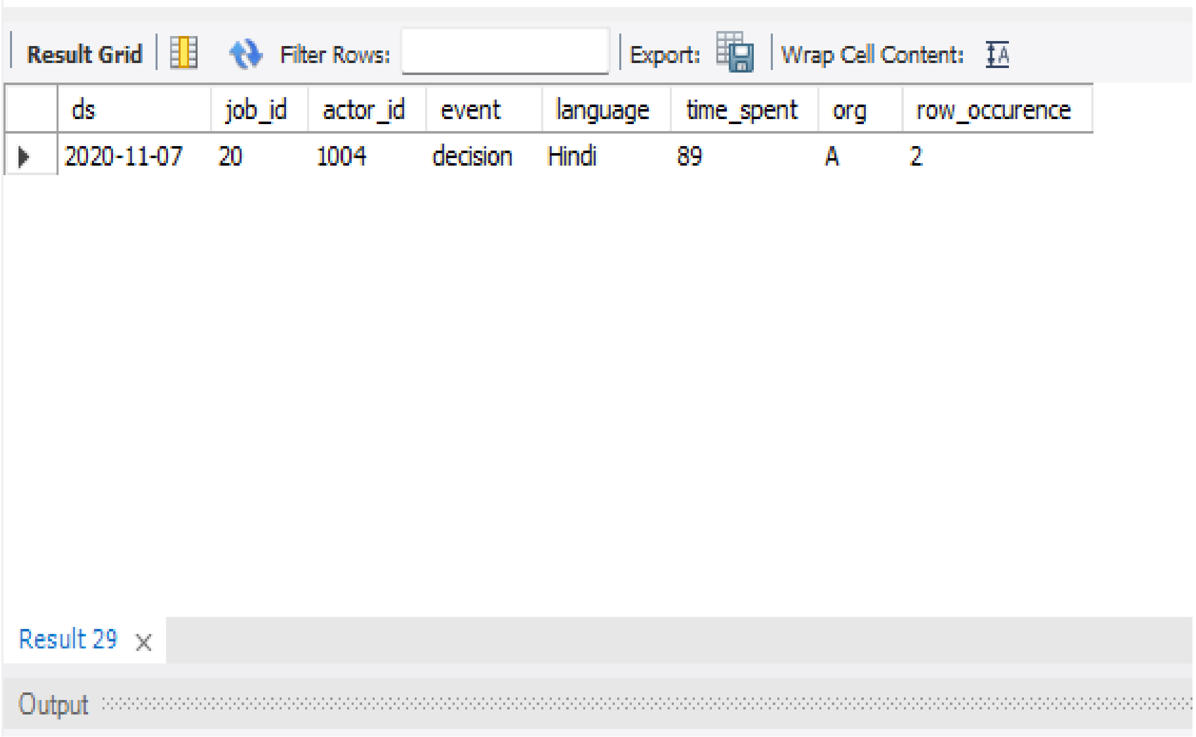
lang	jobs_in_lang	lang_share_percent
English	9	21.9512
Arabic	1	2.4390
Persian	11	26.8293
Hindi	7	17.0732
French	9	21.9512
Italian	4	9.7561



d. <u>Duplicate rows that have the same value present in them</u>

i. To display duplicates from the job_data table, I selected all field names or columns and used the COUNT() function to get the number of occurrences of the duplicate rows from the table. Then using GROUP BY clause with all column names along with a condition placed using HAVING clause where the query is checking for any row with duplicate occurrence .

ii. Screenshot of findings(from MySql Workbench) -

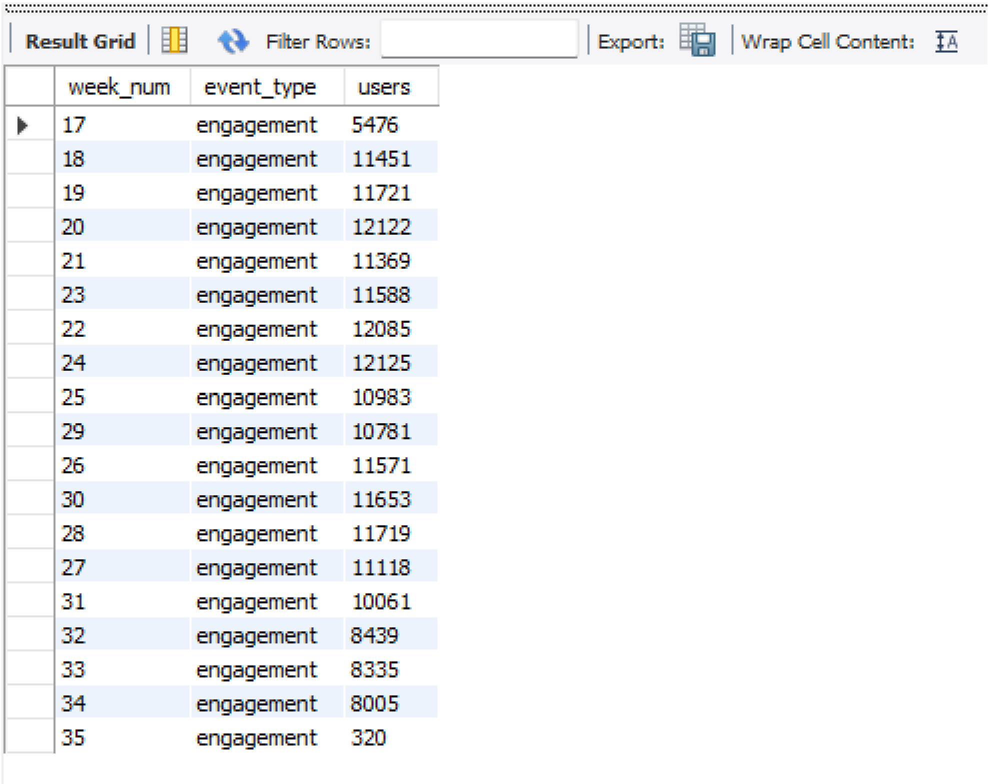


ds	job_id	actor_id	event	language	time_spent	org	row_occurence
2020-11-07	20	1004	decision	Hindi	89	A	2

02. Case Study 2 (Investigating metric spike)

a. Weekly user engagement

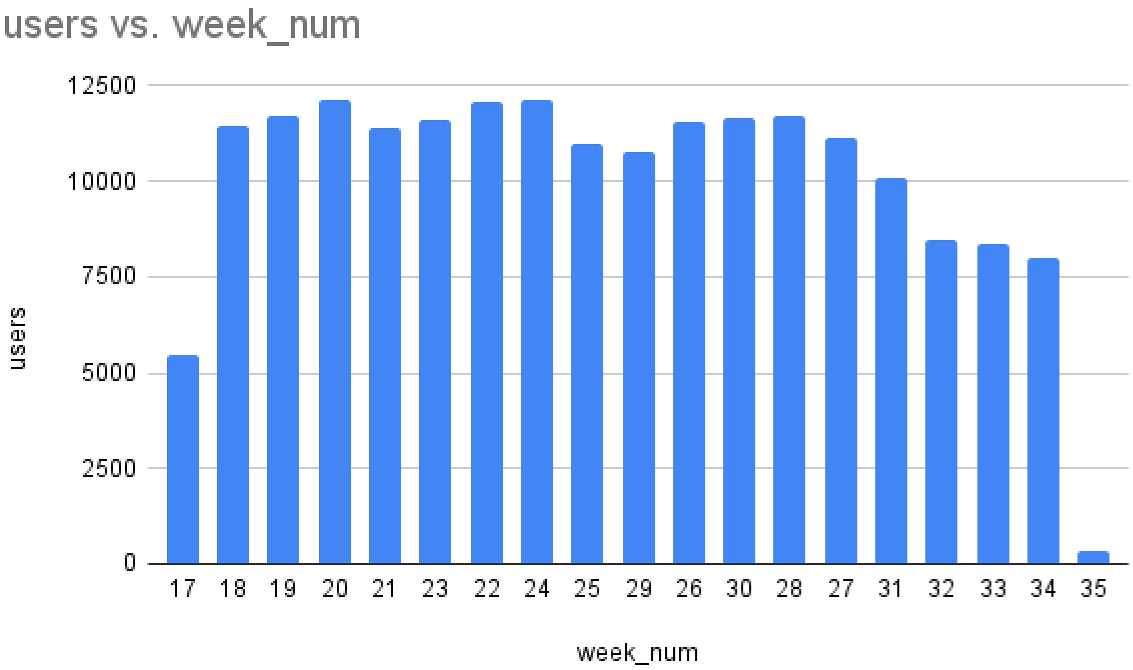
i. Found the weekly count of users who are engaged using the product.



week_num	event_type	users
17	engagement	5476
18	engagement	11451
19	engagement	11721
20	engagement	12122
21	engagement	11369
23	engagement	11588
22	engagement	12085
24	engagement	12125
25	engagement	10983
29	engagement	10781
26	engagement	11571
30	engagement	11653
28	engagement	11719
27	engagement	11118
31	engagement	10061
32	engagement	8439
33	engagement	8335
34	engagement	8005
35	engagement	320

ii. With this information, we can get a clear picture on how many users are engaging on a weekly basis to use the product. We can get the pattern to measure the activeness of a user, measuring if the user finds quality in a product/service.

iii. The number of users each week are increasing as well as decreasing. So the activity varies over the week.

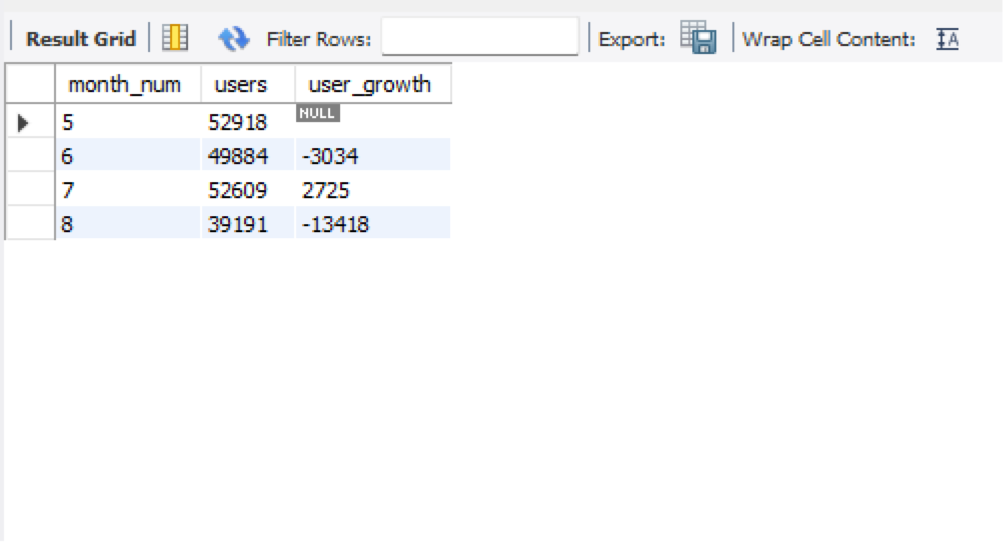


b. Amount of users growing over time

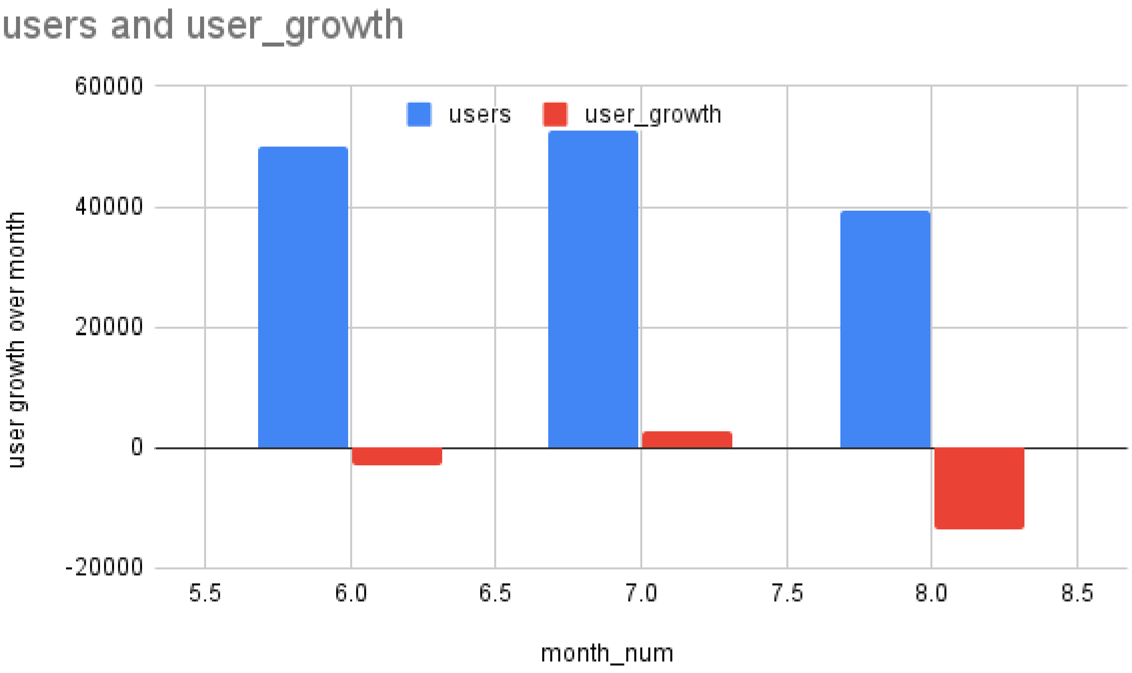
i. Found the total number of users using the product on May(5th), June(6th), July(7th) and August(8th) of the year 2014.

ii. From this , we can clearly see the monthly user growth. On June'14, growth declined by 3034 users. The next month's growth was positive. On July'14 product users increased by

2725 from the previous month. On August’14 there user growth was rapidly decreased by 13418.



month_num	users	user_growth
5	52918	NULL
6	49884	-3034
7	52609	2725
8	39191	-13418

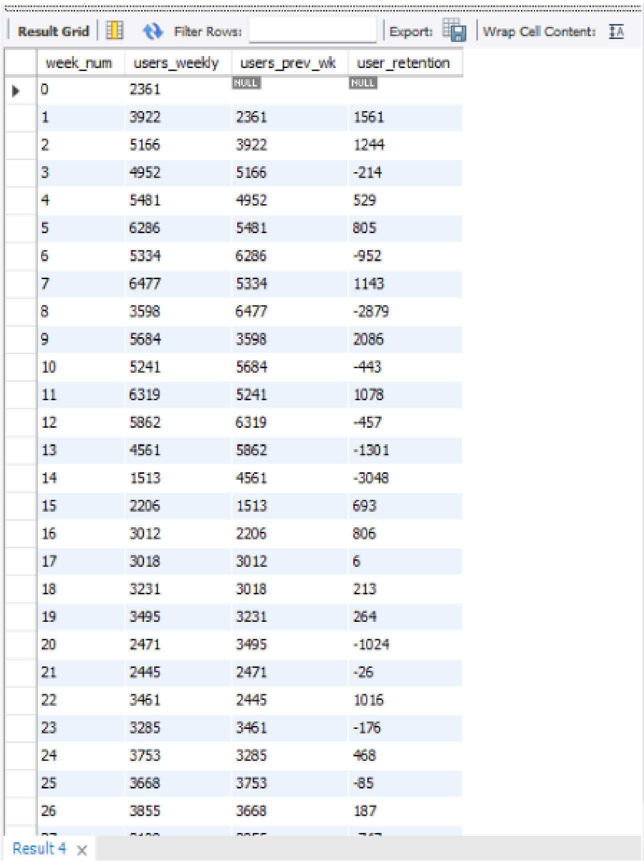


c. Weekly retention of users sign-up cohort

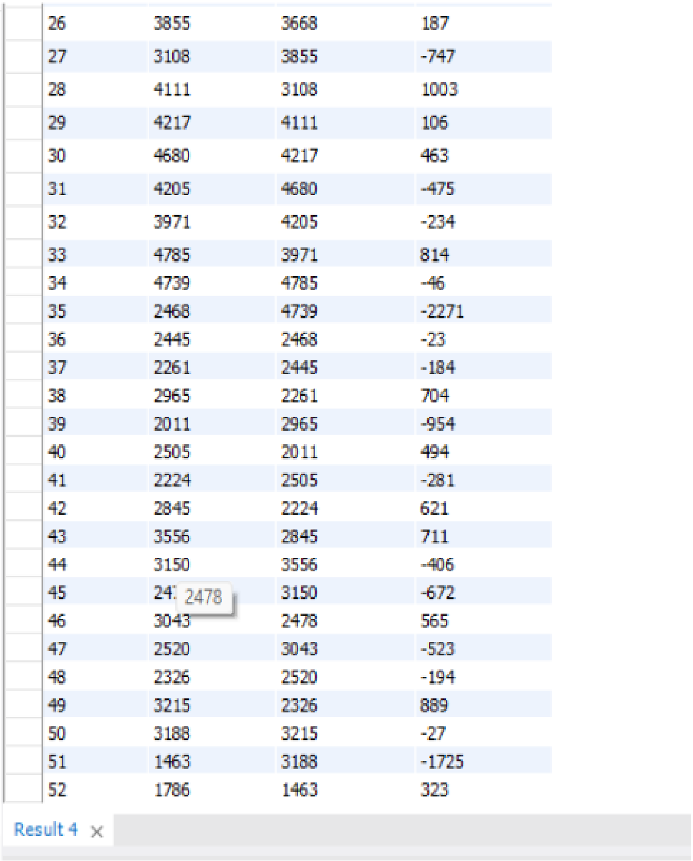
i. After analysis by writing sql queries on the workbench editor, I found the weekly count of users for a whole year of 2014.

ii. This data is telling about how users are getting retained weekly after signing-up for a product.

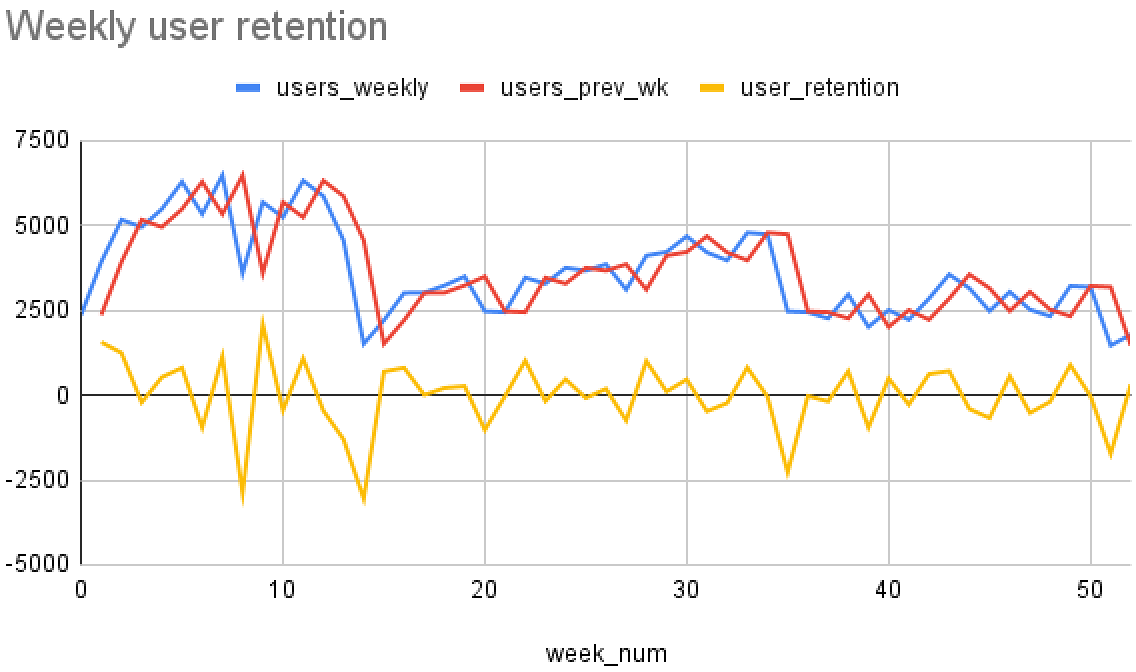
iii. We can see, many retentions over weeks (positive records of user_retention column) as well as declines also (negative records of user_retention column)



week_num	users_weekly	users_prev_wk	user_retention
0	2361	NULL	NULL
1	3922	2361	1561
2	5166	3922	1244
3	4952	5166	-214
4	5481	4952	529
5	6286	5481	805
6	5334	6286	-952
7	6477	5334	1143
8	3598	6477	-2879
9	5684	3598	2086
10	5241	5684	-443
11	6319	5241	1078
12	5862	6319	-457
13	4561	5862	-1301
14	1513	4561	-3048
15	2206	1513	693
16	3012	2206	806
17	3018	3012	6
18	3231	3018	213
19	3495	3231	264
20	2471	3495	-1024
21	2445	2471	-26
22	3461	2445	1016
23	3285	3461	-176
24	3753	3285	468
25	3668	3753	-85
26	3855	3668	187



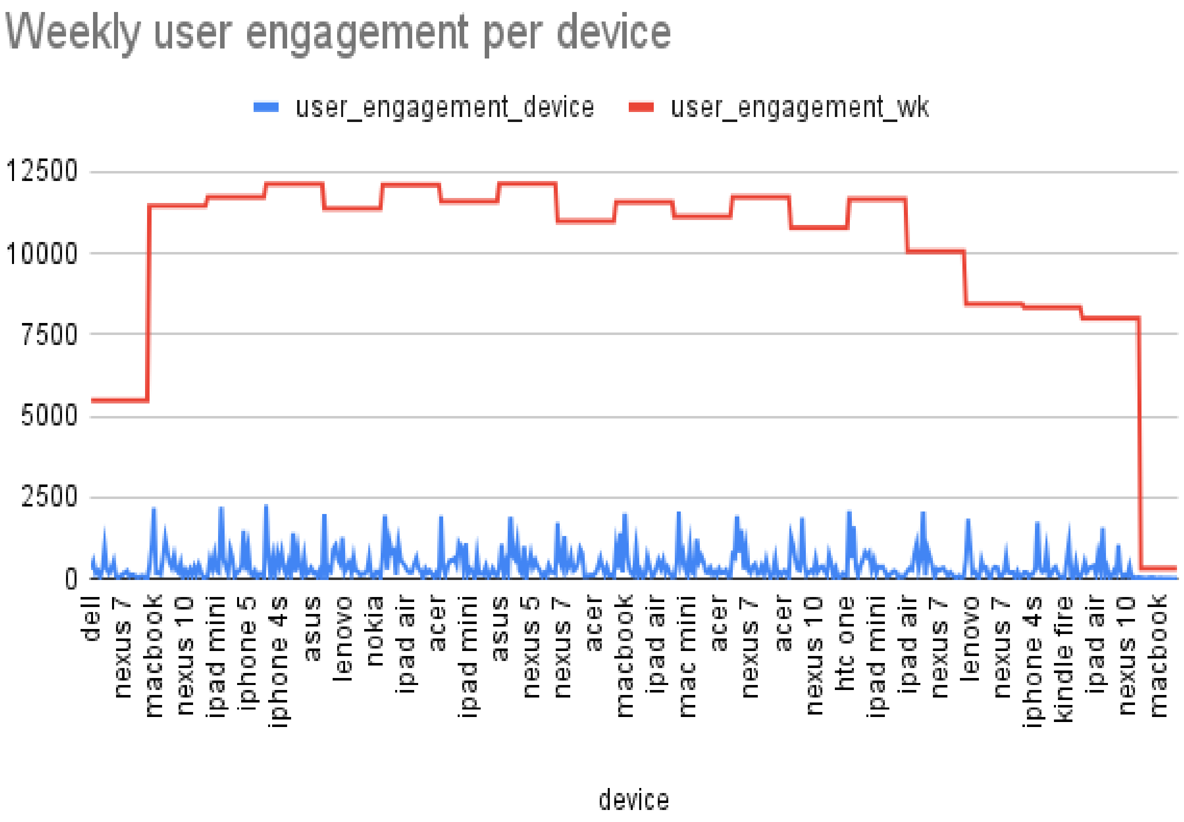
26	3855	3668	187
27	3108	3855	-747
28	4111	3108	1003
29	4217	4111	106
30	4680	4217	463
31	4205	4680	-475
32	3971	4205	-234
33	4785	3971	814
34	4739	4785	-46
35	2468	4739	-2271
36	2445	2468	-23
37	2261	2445	-184
38	2965	2261	704
39	2011	2965	-954
40	2505	2011	494
41	2224	2505	-281
42	2845	2224	621
43	3556	2845	711
44	3150	3556	-406
45	2478	3150	-672
46	3043	2478	565
47	2520	3043	-523
48	2326	2520	-194
49	3215	2326	889
50	3188	3215	-27
51	1463	3188	-1725
52	1786	1463	323

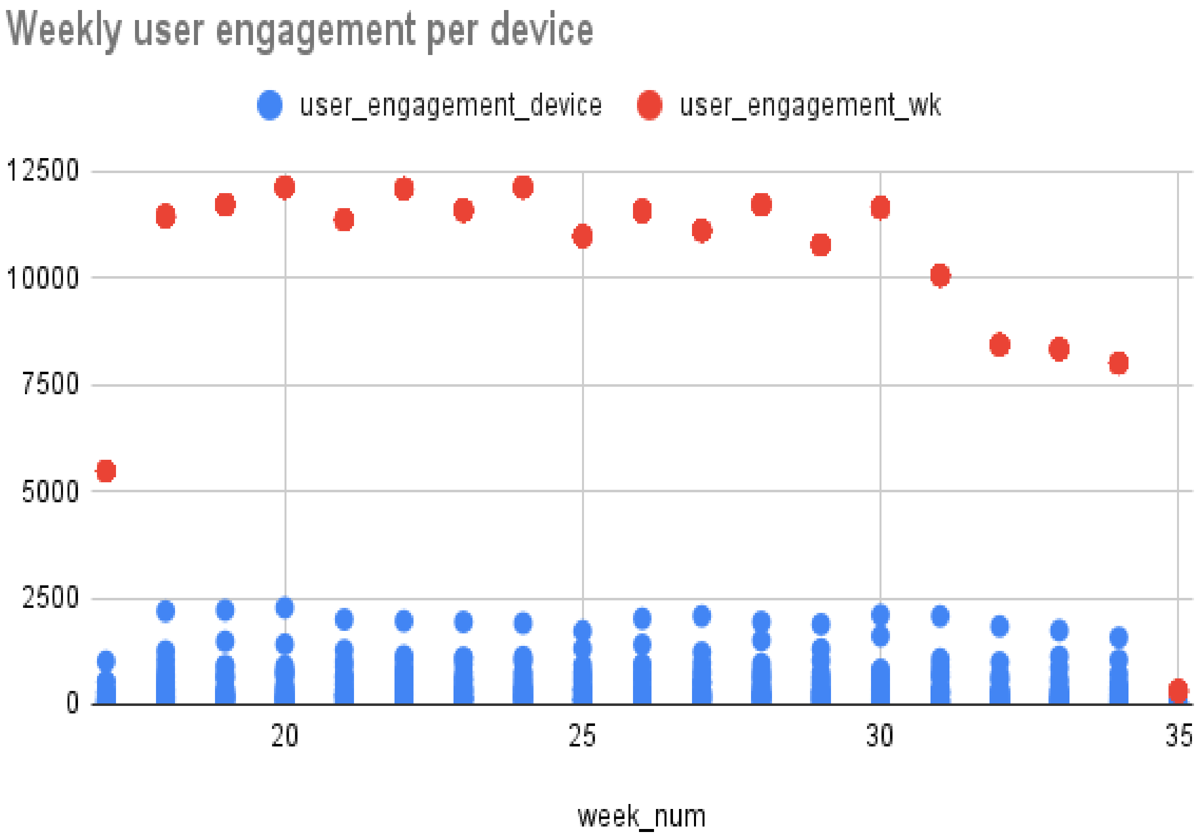


d. Weekly user engagement per device

i. I found user's engagement with the product from all devices for each week from 17th week to 35th week of the year. All users are doing some sort of engaging activity like loading the homepage, liking other user's messages, login to the product etc. Users are using various devices and that information is gathered from different IP addresses of user's devices. Devices are divided into mainly three categories upon which users are also categorized into 3 different user_type.

ii. To measure the activeness of a user and measure if the user finds quality in a product/service weekly, this study is very useful because I found that users are staying active with the product only if they find quality in product / service. The weekly engaged user count(user_engagement_device column) and per device weekly engaged user count(user_engagement_wk column) can give an idea about how much users are getting engaged with the product on a weekly basis and from which devices they are experiencing the product. This analysis can help to improve the user experience(UX) of the product.

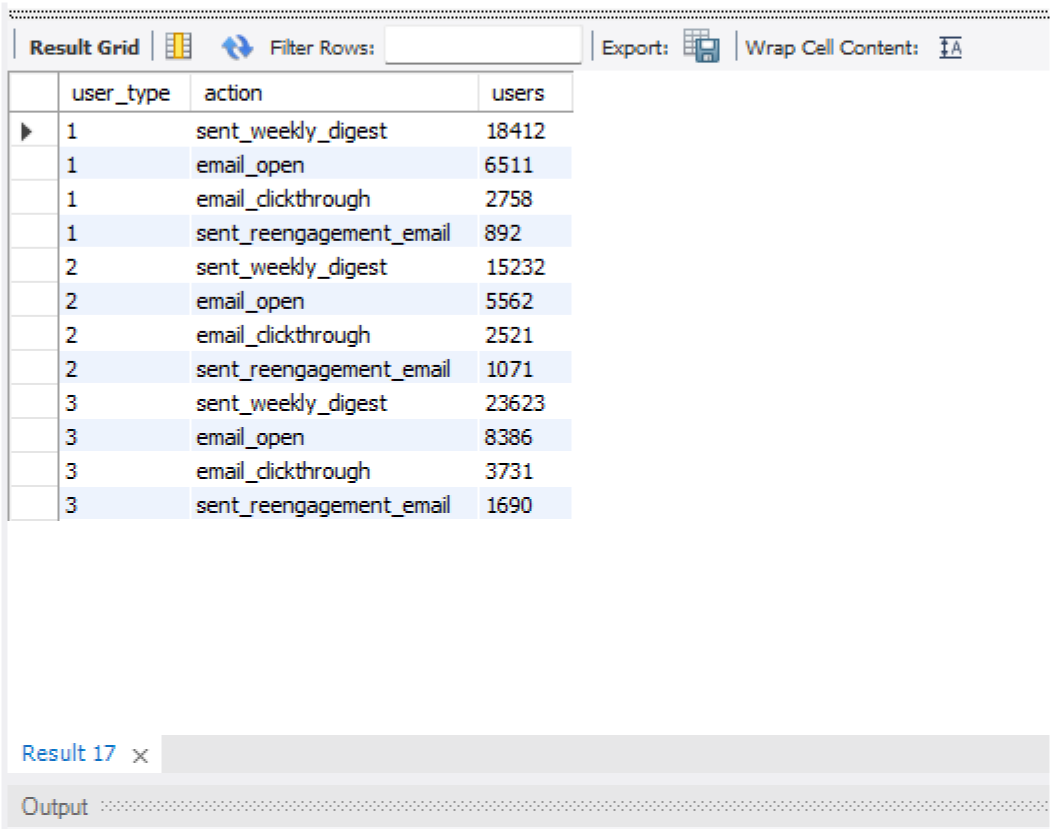




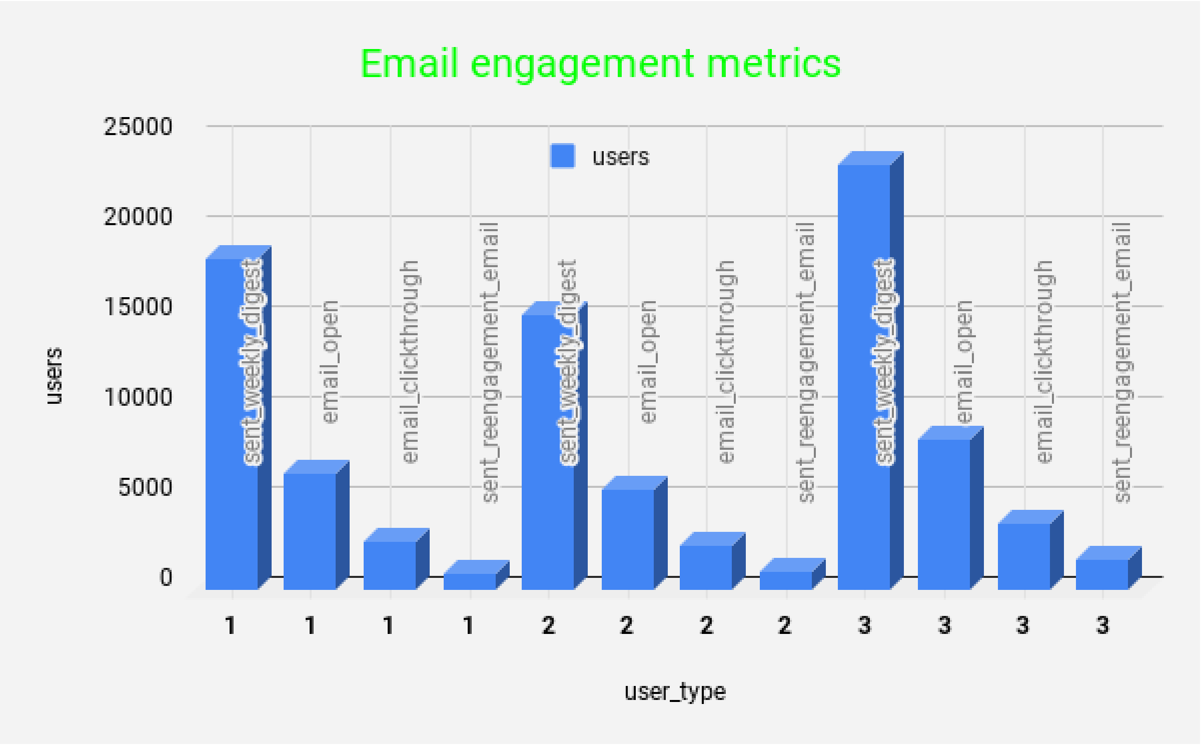
e. Email engagement metrics

i. I have found the number of users associated with each type of action related to events specific to the sending of emails and the user_type(s).

ii. This metric will help to understand how many users from the same user_type(users using similar kind of devices) are engaging with each different type of actions/events specific to the sending of emails.



user_type	action	users
1	sent_weekly_digest	18412
1	email_open	6511
1	email_clickthrough	2758
1	sent_reengagement_email	892
2	sent_weekly_digest	15232
2	email_open	5562
2	email_clickthrough	2521
2	sent_reengagement_email	1071
3	sent_weekly_digest	23623
3	email_open	8386
3	email_clickthrough	3731
3	sent_reengagement_email	1690



Result

★ While doing this project on Operation Analytics and Investigating Metric Spike,

- ○ I learned more usage of basic SQL commands and functions while solving problems practically. The structure of SQL queries, and SQL fundamental topics like logical operators, aggregate functions, sorting functions, and different types of joins(inner join, left join, right join, full join, etc). Apart from that, the learning on advanced SQL and doing this project on basic and advanced level SQL topics like window functions, over clause of window definition, three types of window functions, date and time functions, using nested query etc. helped me to write and execute SQL queries on MySQL Workbench for doing analysis according to the needs.

- ○ I learned how to use MySQL Workbench on my local machine and its features and functionalities.

- ○ I also understood the process of importing data from a csv file into a MySql database table with the help of the “table data import wizard” tool of MySql Workbench 8.0.32.

- ○ I learned how to build and execute the queries.

- ○ I have learned how to do some user analysis using SQL and MySQL

- Being part of the Microsoft Data Analytics Team and working closely with other departments within the organization, I tried to answer questions and figure out what can be derived from those findings. It helped me to understand the end to end operation analytics process.

Drive Links

GDrive Link 01

SQL Query - MySQL Workbench file

GDrive Link 02

Project_Report_pdf - Operation Analytics and Investigating Metric Spike(Advanced SQL)

GDrive Link 03

Job_data.csv - More rows are added into this database table to increase data points

---