Operation Analytics and Investigating Metric Spike

Project Description:

In this project I have conducted two case studies:

1. Understand the **Operation Analytics** of a company to identify areas for improvement based on data-driven insights by answering questions like-
   - Number of jobs reviewed
   - Throughput
   - Percentage share of languages preferred by users
   - Duplicate data

   Such analysis can help predict the company's growth or decline, leading to better automation, cross-functional collaboration, and efficient workflows.

2. **Investigate the metric spike** to understand daily fluctuations and addressing questions about engagement and sales such as-
   - Weekly engagement of users
   - User growth for a product overtime
   - Weekly retention of users
   - Weekly device-wise engagement of users
   - Email engagement of users

   By analysing these metrics and answering the associated questions, we will gain insights into various aspects of the data, user behaviours, and the performance of the product/service.

Approach:

I utilized the provided datasets to extract relevant information and answer the questions at hand.

To analyse the data and perform calculations, I used MySQL Workbench. This allowed me to efficiently manipulate the data, perform calculations such as counting, aggregating, and averaging, and derive insights from the datasets.

Additionally, I used Tableau for data visualization, enabling me to create clear and interactive visual representations of the analysed data.

Tech-Stack Used:

I opted to work on MySQL Workbench, Tableau and Microsoft Excel for this project due to several reasons.

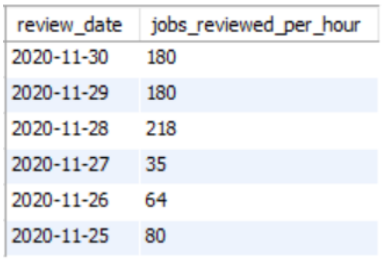
- Firstly, it is a cost-effective and open-source software that offers a user-friendly interface, making it well-suited for managing extensive datasets.
- Additionally, it provides various tools for optimizing performance, monitoring, and ensuring the security of the database.
- Furthermore, one of the significant advantages of MySQL Workbench is its compatibility with other software tools, including Tableau, which I employed for data visualization.
- Tableau is a robust tool that enables users to visually analyse, interpret, and present complex datasets effectively.
- Microsoft Excel is one of the best tools in the market that provide a lot of operations for data analysis like pivot tables and charts that makes the work a lot easier, quick and robust.

Together, MySQL Workbench, Excel and Tableau offer a seamless experience, allowing me to extract valuable insights and deliver meaningful, data-driven solutions for the Instagram product team.

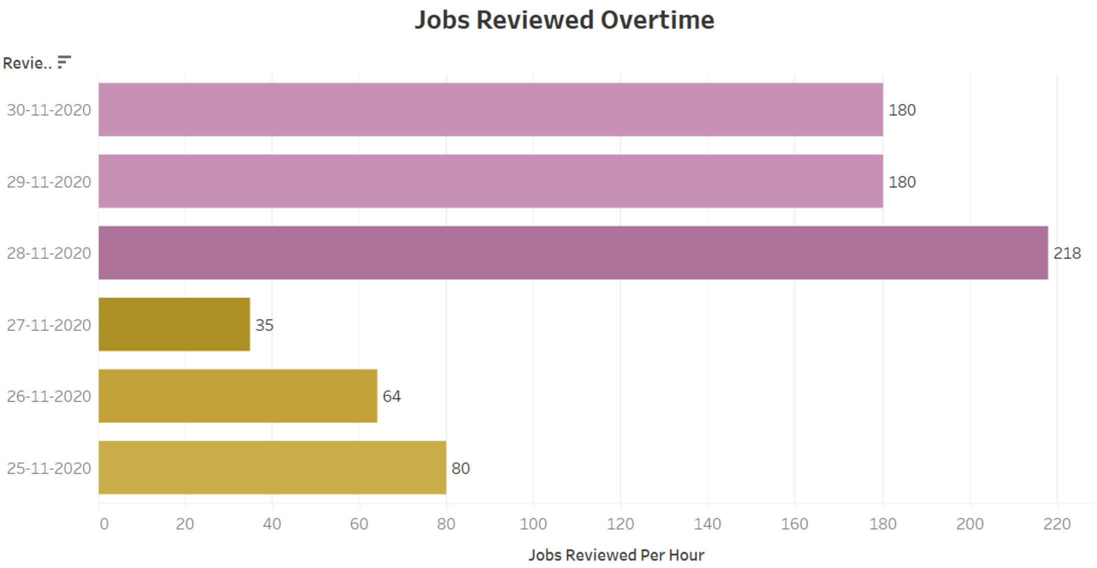
Insights:

Case Study 1(Operation Analytics)

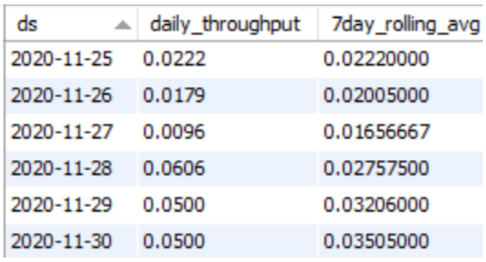
A. Number of Jobs reviewed - Calculating the number of jobs reviewed per hour per day for November 2020 showed the following result:



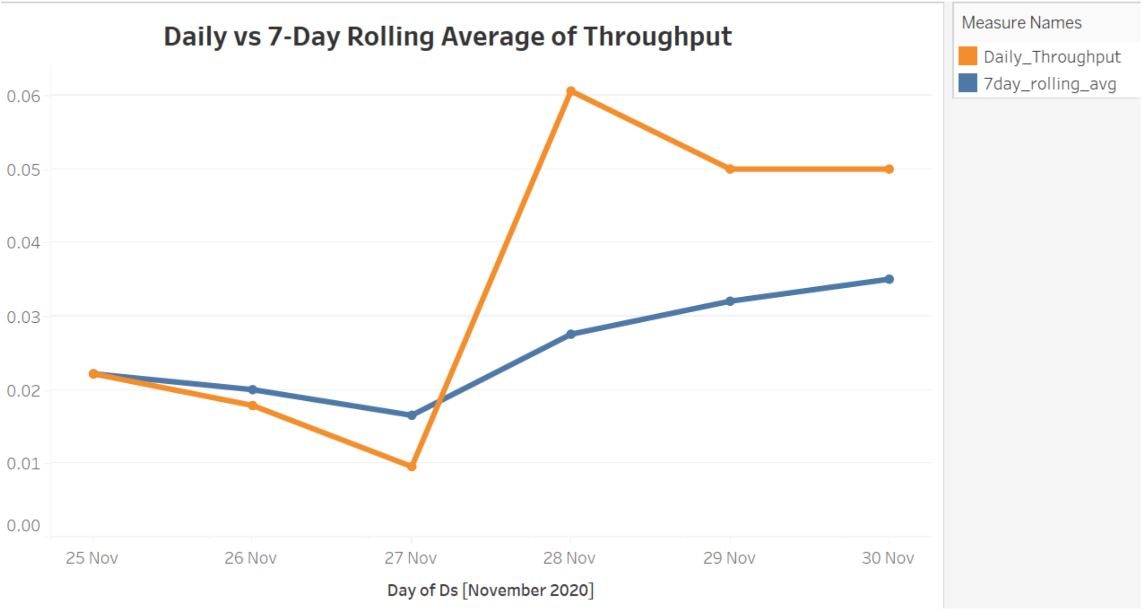
review_date	jobs_reviewed_per_hour
2020-11-30	180
2020-11-29	180
2020-11-28	218
2020-11-27	35
2020-11-26	64
2020-11-25	80



B. Throughput- It is the number of events happening per second. I calculate both daily metric and the 7-day rolling average of throughput and the result found is depicted below:

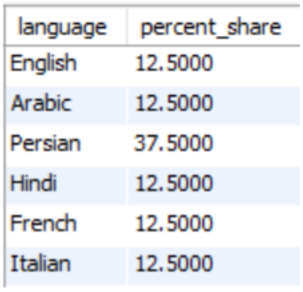


ds	daily_throughput	7day_rolling_avg
2020-11-25	0.0222	0.02220000
2020-11-26	0.0179	0.02005000
2020-11-27	0.0096	0.01656667
2020-11-28	0.0606	0.02757500
2020-11-29	0.0500	0.03206000
2020-11-30	0.0500	0.03505000

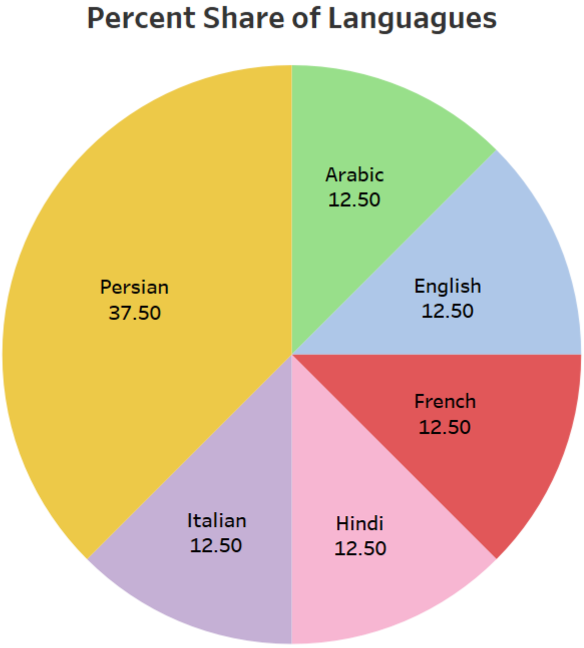


- Daily throughput provides the metric’s throughput value for each individual day and allows detailed analysis of short-term fluctuations and patterns. However, a rolling average provides the throughput average that gets updated with each passing day that means a more generalised and smoothed representation of the metric over time (7 days in this case). Thus, it helps identify long-term trends and patterns to highlight the overall direction of the metric, over a specific timeframe. So, I would prefer 7-day rolling average of the throughput over the daily throughput.

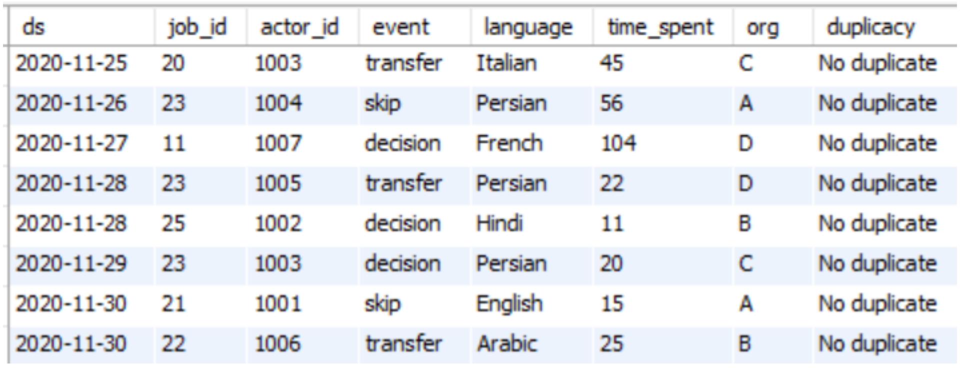
C. **Percentage share of each language-** Share of each language for different contents was found to be as follows:



language	percent_share
English	12.5000
Arabic	12.5000
Persian	37.5000
Hindi	12.5000
French	12.5000
Italian	12.5000



D. **Duplicate rows:** Checking for duplicate entries of rows in the given dataset showed no duplicates.

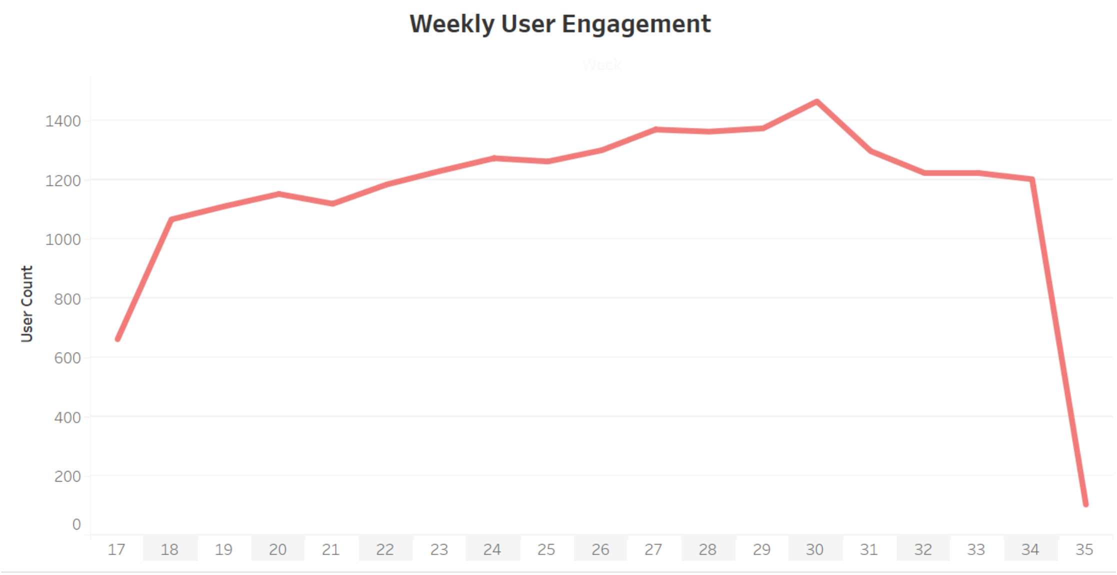


ds	job_id	actor_id	event	language	time_spent	org	duplicacy
2020-11-25	20	1003	transfer	Italian	45	C	No duplicate
2020-11-26	23	1004	skip	Persian	56	A	No duplicate
2020-11-27	11	1007	decision	French	104	D	No duplicate
2020-11-28	23	1005	transfer	Persian	22	D	No duplicate
2020-11-28	25	1002	decision	Hindi	11	B	No duplicate
2020-11-29	23	1003	decision	Persian	20	C	No duplicate
2020-11-30	21	1001	skip	English	15	A	No duplicate
2020-11-30	22	1006	transfer	Arabic	25	B	No duplicate

Case Study 2 (Investing Metric Spike)

A. User Engagement- The task was to find the effectiveness of a user by measuring the weekly user engagement and the result was found to be as follows:

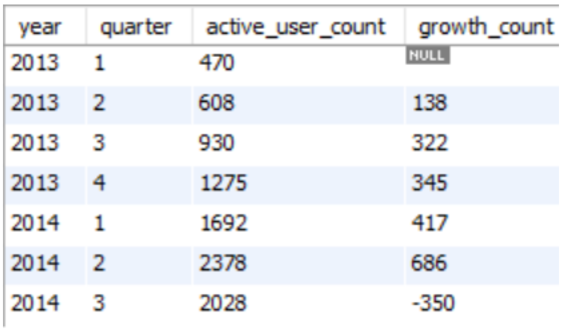
week	user_count
17	663
18	1068
19	1113
20	1154
21	1121
22	1186
23	1232
24	1275
25	1264
26	1302
27	1372
28	1365
29	1376
30	1467
31	1299
32	1225
33	1225
34	1204
35	104



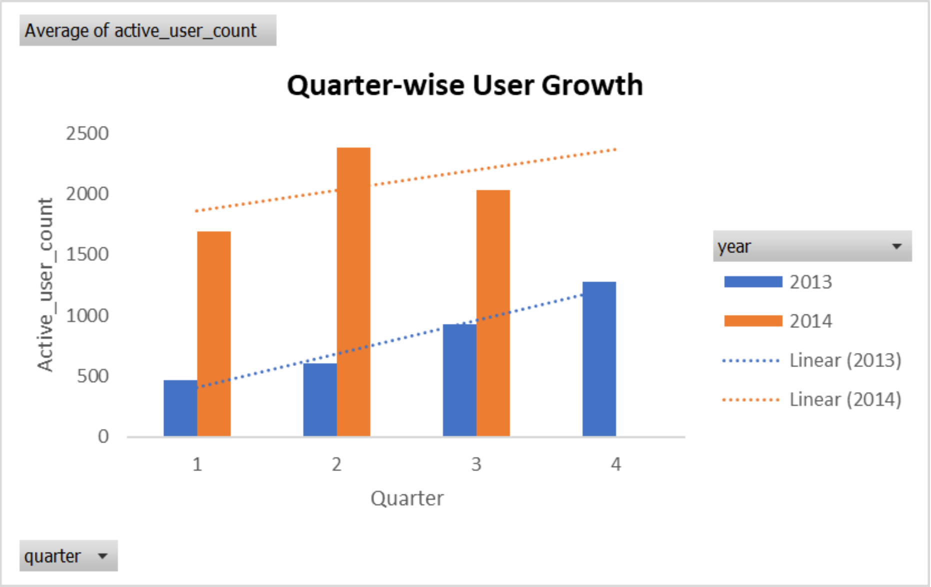
The user engagement showed an overall increase till week 30 but then started decreasing in the next couple of weeks.

**We can see that user engagement dropped significantly by week 35 but its important to note that we had the data for only 1 day for week 35, so it should be exlcluded from the overall conclusion.

B. User Growth- The task was to find out the number of users growing over time. The user growth for each quarter per year was found to be as follows:

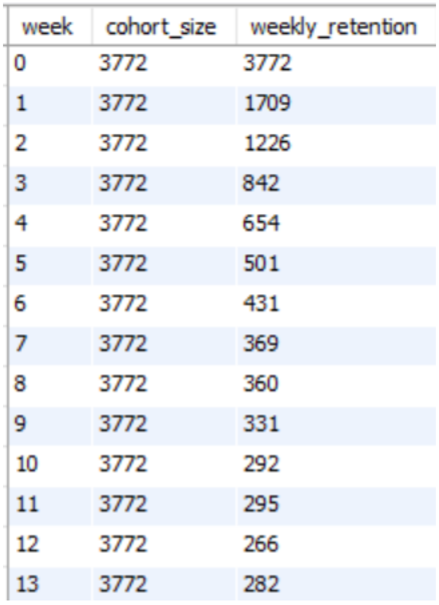


year	quarter	active_user_count	growth_count
2013	1	470	NULL
2013	2	608	138
2013	3	930	322
2013	4	1275	345
2014	1	1692	417
2014	2	2378	686
2014	3	2028	-350

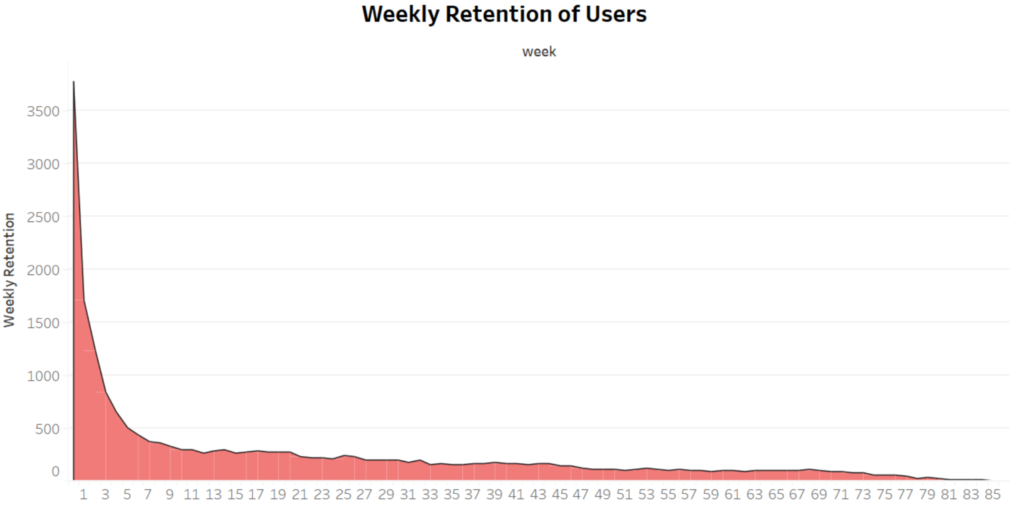


As represented by the graph, there is an overall increase in user growth overtime.

C. **Weekly Retention-** The result of calculation of weekly retention of users after signing up for a product is depicted below:



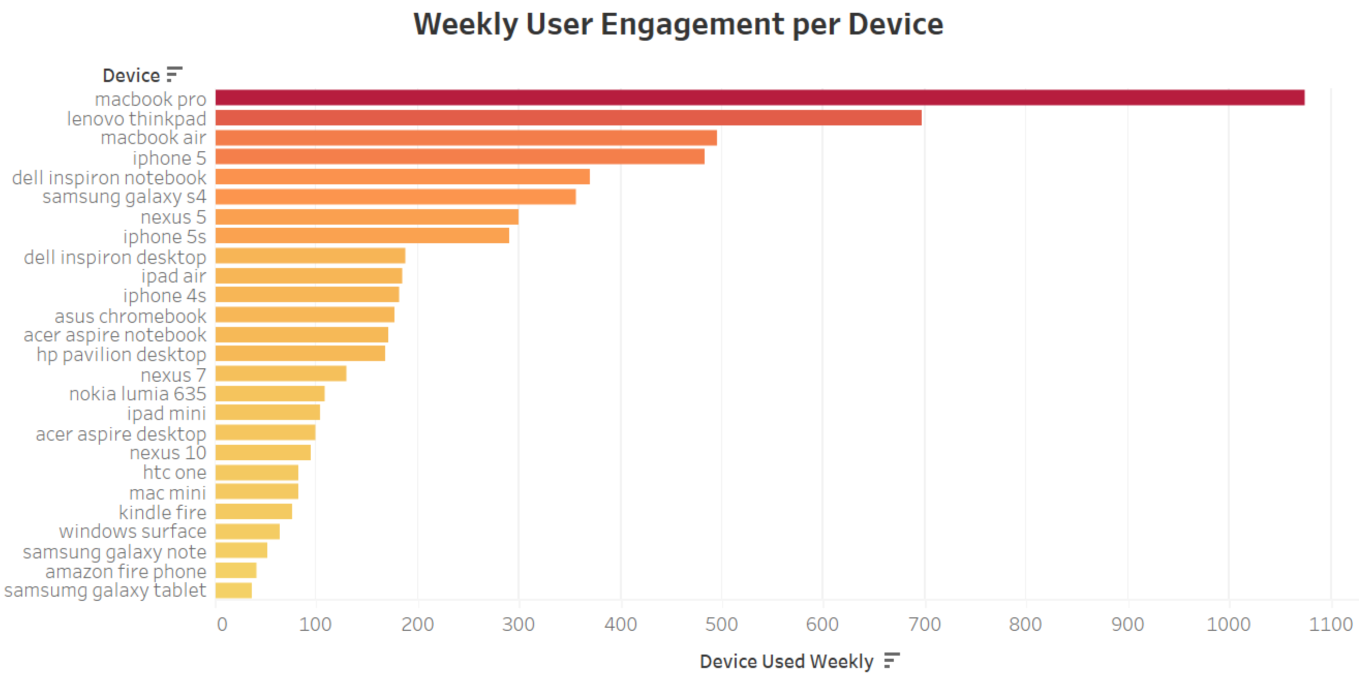
week	cohort_size	weekly_retention
0	3772	3772
1	3772	1709
2	3772	1226
3	3772	842
4	3772	654
5	3772	501
6	3772	431
7	3772	369
8	3772	360
9	3772	331
10	3772	292
11	3772	295
12	3772	266
13	3772	282



There is a significant drop in retention of users sign-up cohort in the initial 2 weeks which keeps decreasing overtime such that by end of 86 weeks only 2 users are retained.

D. **Weekly Engagement-** To measure the activeness of users, weekly user engagement per device was measured and the results are as follows:

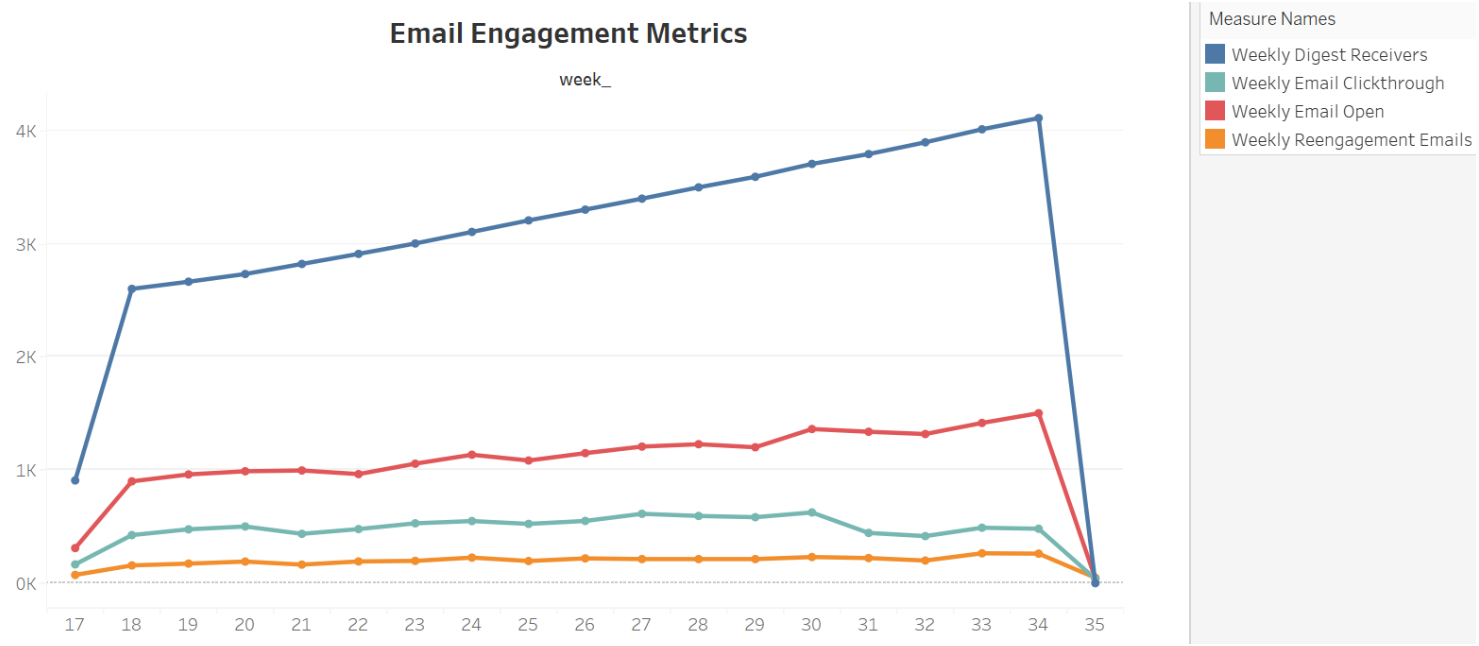
device	users_weekly	device_used_weekly
acer aspire desktop	26	98.8421
acer aspire notebook	43.1579	170.5263
amazon fire phone	10.5556	41.3333
asus chromebook	43.5263	176.6842
dell inspiron desktop	46.6316	188.2105
dell inspiron notebook	91.1053	370.4211
hp pavilion desktop	42.1053	167.5263
htc one	21.8421	83.0526
ipad air	51.4444	185.1667
ipad mini	30	104.2105
iphone 4s	46.6316	181.7368
iphone 5	123.1579	483.6316
iphone 5s	73.3158	290.3684
kindle fire	21.1579	76.5789
lenovo thinkpad	172.9474	697.7368
mac mini	20.4737	82.1053
macbook air	123.1579	494.6842
macbook pro	260.1579	1074.4737
nexus 10	27.0526	95.5263
nexus 5	76.3684	298.8947
nexus 7	36.3684	129.7895
nokia lumia 635	28.1579	108.7895
samsumg galaxy tablet	10.2778	36.3333
samsung galaxy note	13.4737	52.7368
samsung galaxy s4	91.5789	356.2105
windows surface	18.2105	64.5789



The graph shows the overall usage of devices by users on a weekly basis where MacBook Pro seems to be the highest used device.

E. Email Engagement- The metrics for weekly user engagement with email services provided by the company is as follows:

week	weekly_digest_receivers	weekly_email_open	weekly_email_clickthrough	weekly_reengagement_emails
17	909	311	167	74
18	2602	901	426	158
19	2665	962	477	174
20	2733	990	502	192
21	2822	997	437	165
22	2911	965	479	193
23	3003	1057	530	198
24	3105	1136	550	227
25	3207	1085	525	197
26	3302	1150	551	220
27	3399	1208	614	214
28	3499	1229	595	214
29	3592	1202	584	214
30	3706	1363	626	232
31	3793	1339	445	223
32	3897	1319	417	201
33	4012	1417	491	265
34	4111	1503	482	262
35	1	42	39	49



An overall increase is observed in weekly digest receivers and weekly email opening numbers and there is decrease in the weekly email clickthrough specially after week 30 whereas weekly reengagement emails show a somewhat steady graph.
The graph also signifies that although there is overall increasing engagement of users with the emails however the response of the users compared to the average number of emails sent by the company weekly, is significantly low.

Results:

- Working on this project was a little challenging in terms of understanding the requirements of the tasks at hand and making the best assumptions to analyse the given dataset.
- Advanced SQL methods like window and rolling average functions were new and interesting to learn, as well as learning new concepts like cohort analysis and retention.
- Understood and experienced using complicated nested queries in SQL
- Got practice in using tableau for data visualization and interpretation and using pivot tables and charts in excel for data analysis.

SQL Query:

CASE STUDY 1:

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

```
SELECT
    DATE(ds) AS review_date,
    round(COUNT(job_id) / (SUM(time_spent) / 3600)) AS
jobs_reviewed_per_hour
FROM
    job_data
WHERE
    MONTH(ds) = 11 AND YEAR(ds) = 2020
GROUP BY review_date;
```

B. Calculate 7 day rolling average of throughput?

```
with throughput_data as(
          select
               ds,
               count(event)/sum(time_spent)as throughput
            from job_data
            group by ds
            order by ds
            )
     select
          ds,
        avg(throughput) over(order by ds rows between 6 preceding
and current row) as 7day_rolling_avg
     from throughput_data;
```

C. Calculate the percentage share of each language in the last 30 days?

```
select language, count(job_id) as num_of_jobs,
count(job_id)*100/sum(count(*)) over() as percent_share
from job_data
where ds between '2020-11-01' and '2020-11-30'
group by language;
```

D. How will you display duplicates from the table?

```
select dupli_data.ds, dupli_data.job_id, dupli_data.actor_id,
dupli_data.event, dupli_data.language, dupli_data.time_spent,
dupli_data.org,
case when dupli_data.duplicates = 1 then 'No duplicate' else
'Duplicate' end as 'duplicacy'
from (select *,
row_number() over(partition by ds, job_id, actor_id, event,
language, time_spent, org) as duplicates
from job_data) dupli_data;
```

CASE STUDY 2:

A. Calculate the weekly user engagement?

```
SELECT week(occurred_at) as week,
COUNT(distinct(user_id)) as user_count
FROM events
WHERE event_type = 'engagement'
GROUP BY week;
```

B. Calculate the user growth for product?

```
select growth_data.year, growth_data.quarter,
growth_data.active_user_count,
active_user_count - lag(active_user_count,1) over(order by year,
quarter) as growth_count
from (
select year(created_at) as year,
quarter(created_at) as quarter,
count(distinct(user_id)) as active_user_count
from users
where state = 'active' and activated_at is not null
group by year, quarter) as growth_data;
```

C. Calculate the weekly retention of users-sign up cohort?

```
SELECT
  week,
  FIRST_VALUE(weekly_retention) OVER (ORDER BY week) AS
cohort_size,
  weekly_retention
FROM
  (SELECT
    TIMESTAMPDIFF(WEEK, u.activated_at, e.occurred_at) AS week,
    COUNT(DISTINCT u.user_id) AS weekly_retention
  FROM
    (SELECT user_id, activated_at
     FROM users
     WHERE state = 'active') u
  INNER JOIN
    (SELECT user_id, occurred_at
     FROM events
     WHERE event_type = 'engagement') e
  ON u.user_id = e.user_id
  GROUP BY 1) cte;
```

D. Calculate the weekly engagement per device?

```
SELECT
    device,
    AVG(users) AS users_weekly,
    AVG(device_used) AS device_used_weekly
```

```
FROM
    (SELECT
        WEEK(occurred_at) AS week,
            device,
            COUNT(DISTINCT (user_id)) AS users,
            COUNT(device) AS device_used
    FROM
        events
    WHERE
        event_name = 'login'
    GROUP BY week , device
    ORDER BY week) d
GROUP BY device;
```

E. Calculate the email engagement metrics?

```
SELECT
    WEEK(occurred_at) AS week,
    COUNT(DISTINCT (CASE
            WHEN action = 'sent_weekly_digest' THEN user_id
            ELSE 0
        END)) AS weekly_digest_receivers,
    COUNT(DISTINCT (CASE
            WHEN action = 'email_open' THEN user_id
            ELSE 0
        END)) AS weekly_email_open,
    COUNT(DISTINCT (CASE
            WHEN action = 'email_clickthrough' THEN user_id
            ELSE 0
        END)) AS weekly_email_clickthrough,
    COUNT(DISTINCT (CASE
            WHEN action = 'sent_reengagement_email' THEN user_id
            ELSE 0
        END)) AS weekly_reengagement_emails
FROM
    email_events
GROUP BY week
ORDER BY week;
```