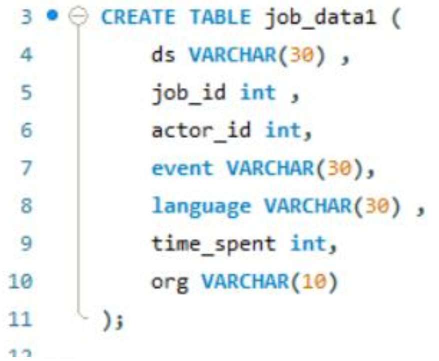
Case Study 1

Project Description

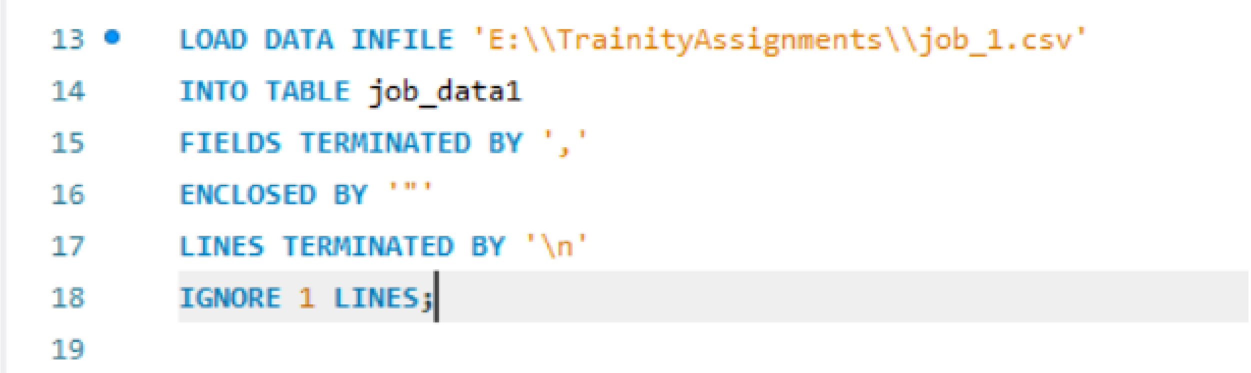
In this project we are trying to analyse the given dataset and draw key insights from it. I am given data for job_data table. Which is having column like job_id,actor_id,language,event,time spent,organization, and ds i.e date

Approach

The csv data sample were given. I have created the dataset then I have created the tables. Then I have loaded data from the csv files.



```sql
CREATE TABLE job_data1 (
    ds VARCHAR(30) ,
    job_id int ,
    actor_id int,
    event VARCHAR(30),
    language VARCHAR(30) ,
    time_spent int,
    org VARCHAR(10)
);
```



```sql
LOAD DATA INFILE 'E:\\TrainityAssignments\\job_1.csv'
INTO TABLE job_data1
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES;
```

I was having some issues loading the fields related to date fields so I had to take date fields as varchar

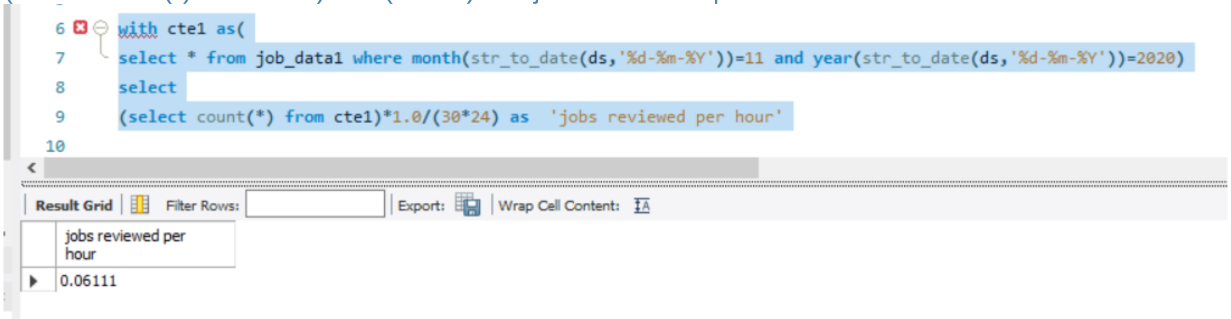
Tech-Stack Used

MySQL Server as the database, MySQL WORKBENCH as a query editor.
G-Drive to upload the assignment.

Insights

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?

```
with cte1 as(
select * from job_data1 where month(str_to_date(ds,'%d-%m-%Y'))=11 and year(str_to_date(ds,'%d-%m-%Y'))=2020)
select
(select count(*) from cte1)*1.0/(30*24) as  'jobs reviewed per hour'
```

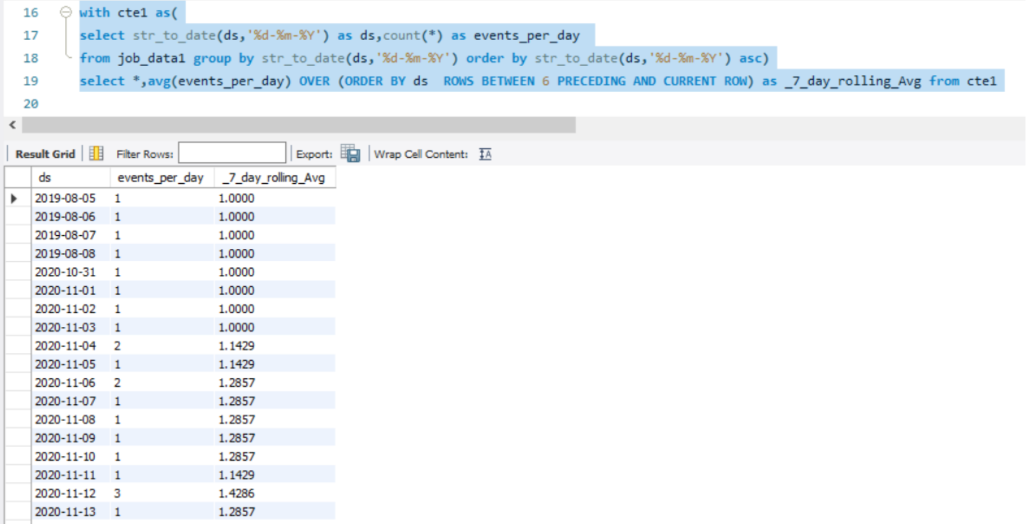


```
with cte1 as(
select * from job_data1 where month(str_to_date(ds,'%d-%m-%Y'))=11 and year(str_to_date(ds,'%d-%m-%Y'))=2020)
select
(select count(*) from cte1)*1.0/(30*24) as  'jobs reviewed per hour'
```

jobs reviewed per hour
0.06111

2. 7 day rolling average.
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?==> **According to me 7day rolling average is preferred as it gives more realistic results**

```
with cte1 as(
select str_to_date(ds,'%d-%m-%Y') as ds,count(*) as events_per_day from job_data1 group by str_to_date(ds,'%d-%m-%Y') order by str_to_date(ds,'%d-%m-%Y') asc)
select *,avg(events_per_day) OVER (ORDER BY ds  ROWS BETWEEN 6 PRECEDING AND CURRENT ROW) as _7_day_rolling_Avg from cte1
```



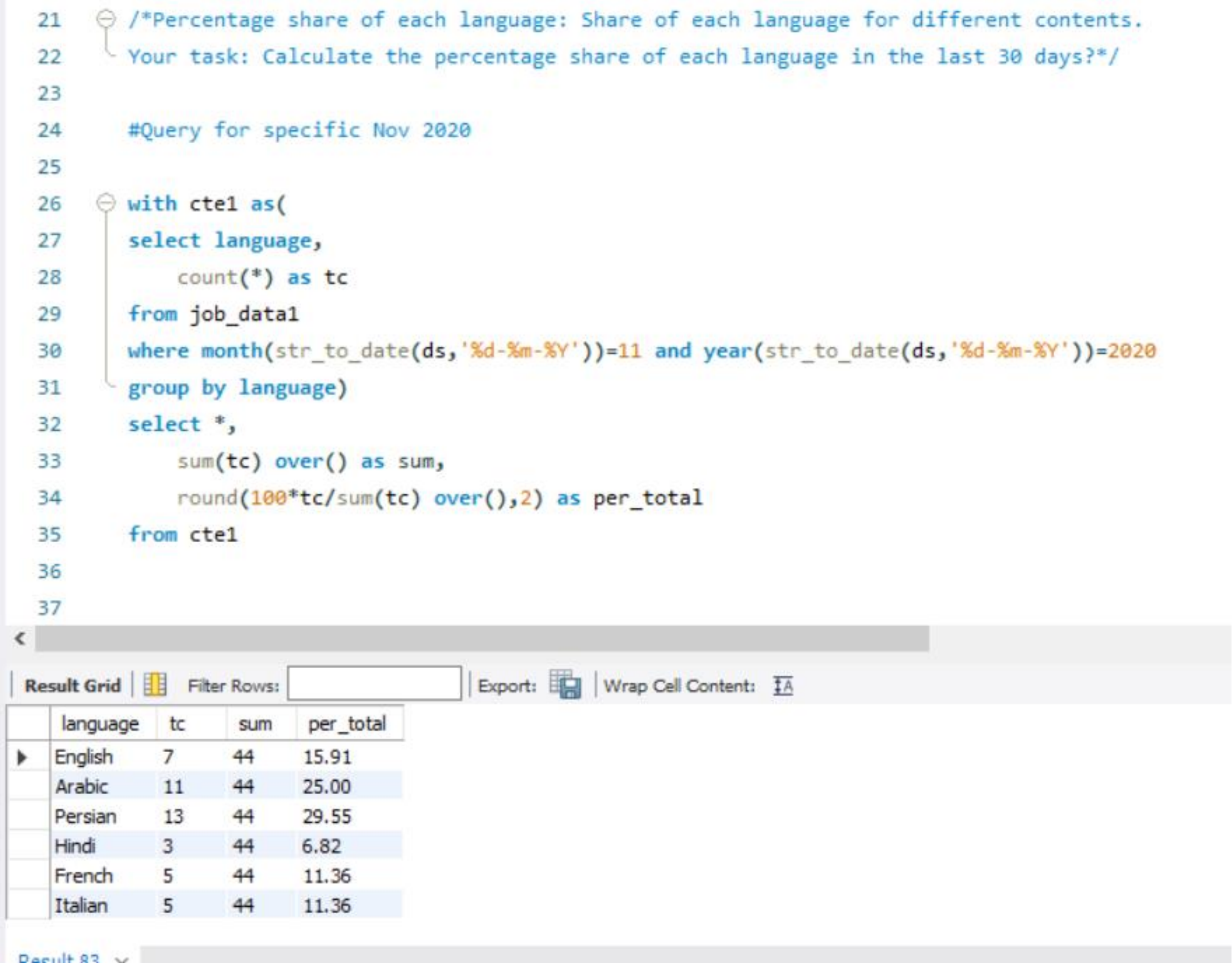
```
with cte1 as(
select str_to_date(ds,'%d-%m-%Y') as ds,count(*) as events_per_day
from job_data1 group by str_to_date(ds,'%d-%m-%Y') order by str_to_date(ds,'%d-%m-%Y') asc)
select *,avg(events_per_day) OVER (ORDER BY ds  ROWS BETWEEN 6 PRECEDING AND CURRENT ROW) as _7_day_rolling_Avg from cte1
```

ds	events_per_day	_7_day_rolling_Avg
2019-08-05	1	1.0000
2019-08-06	1	1.0000
2019-08-07	1	1.0000
2019-08-08	1	1.0000
2020-10-31	1	1.0000
2020-11-01	1	1.0000
2020-11-02	1	1.0000
2020-11-03	1	1.0000
2020-11-04	2	1.1429
2020-11-05	1	1.1429
2020-11-06	2	1.2857
2020-11-07	1	1.2857
2020-11-08	1	1.2857
2020-11-09	1	1.2857
2020-11-10	1	1.2857
2020-11-11	1	1.1429
2020-11-12	3	1.4286
2020-11-13	1	1.2857

7dayRollingAvg.csv

3. /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?
#Query for specific Nov 2020

```
with cte1 as(
select language,
            count(*) as tc
from job_data1
where month(str_to_date(ds,'%d-%m-%Y'))=11 and year(str_to_date(ds,'%d-%m-%Y'))=2020
group by language)
select *,
            sum(tc) over() as sum,
            round(100*tc/sum(tc) over(),2) as per_total
from cte1
```



```
/*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?*/

#Query for specific Nov 2020

with cte1 as(
select language,
    count(*) as tc
from job_data1
where month(str_to_date(ds,'%d-%m-%Y'))=11 and year(str_to_date(ds,'%d-%m-%Y'))=2020
group by language)
select *,
    sum(tc) over() as sum,
    round(100*tc/sum(tc) over(),2) as per_total
from cte1
```

language	tc	sum	per_total
English	7	44	15.91
Arabic	11	44	25.00
Persian	13	44	29.55
Hindi	3	44	6.82
French	5	44	11.36
Italian	5	44	11.36

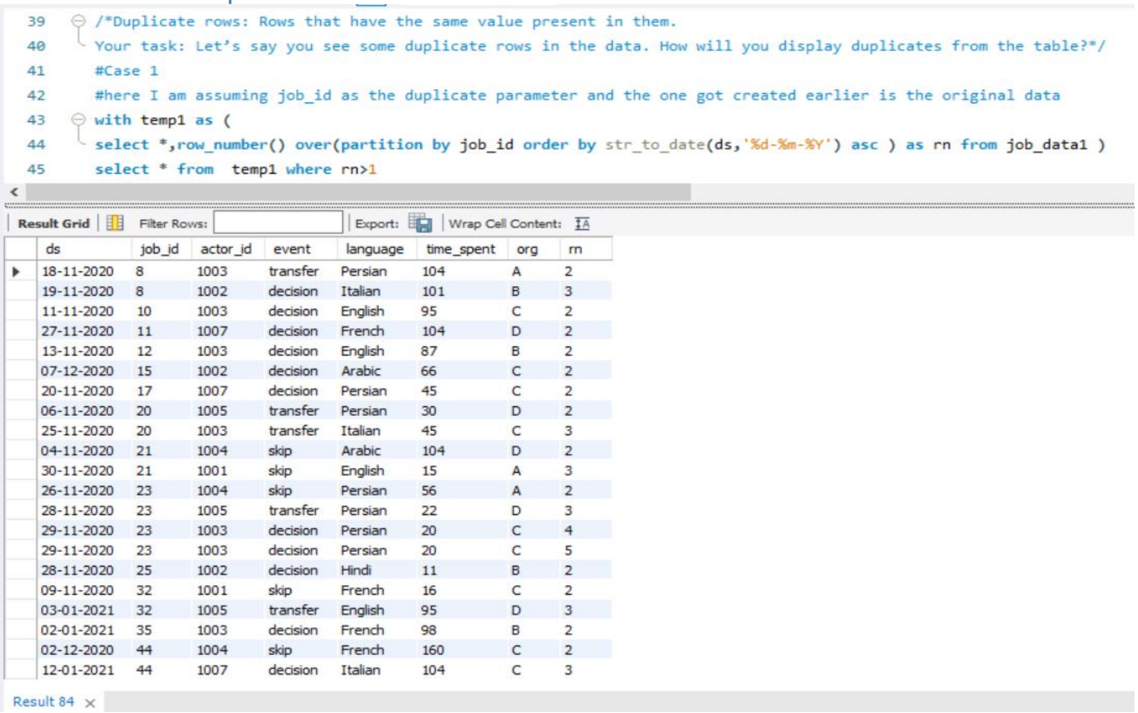
4. Duplicate rows: Rows that have the same value present in them.

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

Case 1

here I am assuming job_id as the duplicate parameter and the one got created earlier is the original data

```
with temp1 as (
select *,
          row_number() over(partition by job_id order by str_to_date(ds,'%d-%m-%Y') asc ) as rn
from job_data1 )
select * from  temp1 where rn>1
```



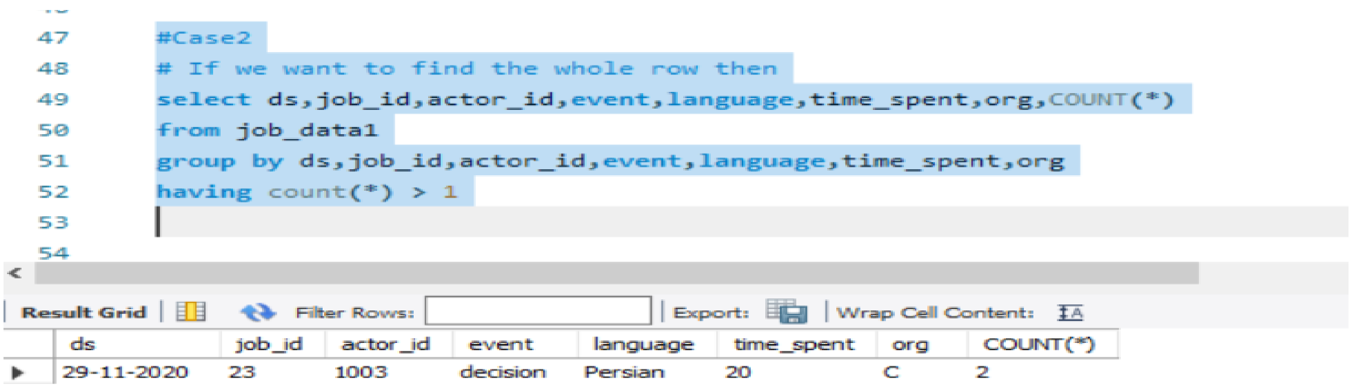
```
/*Duplicate rows: Rows that have the same value present in them.
Your task: Let’s say you see some duplicate rows in the data. How will you display duplicates from the table?*/
#Case 1
#here I am assuming job_id as the duplicate parameter and the one got created earlier is the original data
with temp1 as (
select *,row_number() over(partition by job_id order by str_to_date(ds,'%d-%m-%Y') asc ) as rn from job_data1 )
select * from  temp1 where rn>1
```

ds	job_id	actor_id	event	language	time_spent	org	rn
18-11-2020	8	1003	transfer	Persian	104	A	2
19-11-2020	8	1002	decision	Italian	101	B	3
11-11-2020	10	1003	decision	English	95	C	2
27-11-2020	11	1007	decision	French	104	D	2
13-11-2020	12	1003	decision	English	87	B	2
07-12-2020	15	1002	decision	Arabic	66	C	2
20-11-2020	17	1007	decision	Persian	45	C	2
06-11-2020	20	1005	transfer	Persian	30	D	2
25-11-2020	20	1003	transfer	Italian	45	C	3
04-11-2020	21	1004	skip	Arabic	104	D	2
30-11-2020	21	1001	skip	English	15	A	3
26-11-2020	23	1004	skip	Persian	56	A	2
28-11-2020	23	1005	transfer	Persian	22	D	3
29-11-2020	23	1003	decision	Persian	20	C	4
29-11-2020	23	1003	decision	Persian	20	C	5
28-11-2020	25	1002	decision	Hindi	11	B	2
09-11-2020	32	1001	skip	French	16	C	2
03-01-2021	32	1005	transfer	English	95	D	3
02-01-2021	35	1003	decision	French	98	B	2
02-12-2020	44	1004	skip	French	160	C	2
12-01-2021	44	1007	decision	Italian	104	C	3

Case 2

If we want to find the whole row then

```
select ds,job_id,actor_id,event,language,time_spent,org,COUNT(*)
from job_data1
group by ds,job_id,actor_id,event,language,time_spent,org
having count(*) > 1
```



```
#Case2
# If we want to find the whole row then
select ds,job_id,actor_id,event,language,time_spent,org,COUNT(*)
from job_data1
group by ds,job_id,actor_id,event,language,time_spent,org
having count(*) > 1
```

ds	job_id	actor_id	event	language	time_spent	org	COUNT(*)
29-11-2020	23	1003	decision	Persian	20	C	2

Result:

1. I was able to create meaningful insights from the dataset which can be crucial for business. I was able to sharpen my SQL skills from here. Successfully completing this project boosted my confidence.
2. Created the database from scratch, created dataset
3. Wrote many complex queries. Faced many issues like I was not able to load the date fields so had to change the data type and load data then had to convert the string to date filed in each time