Assignment 4

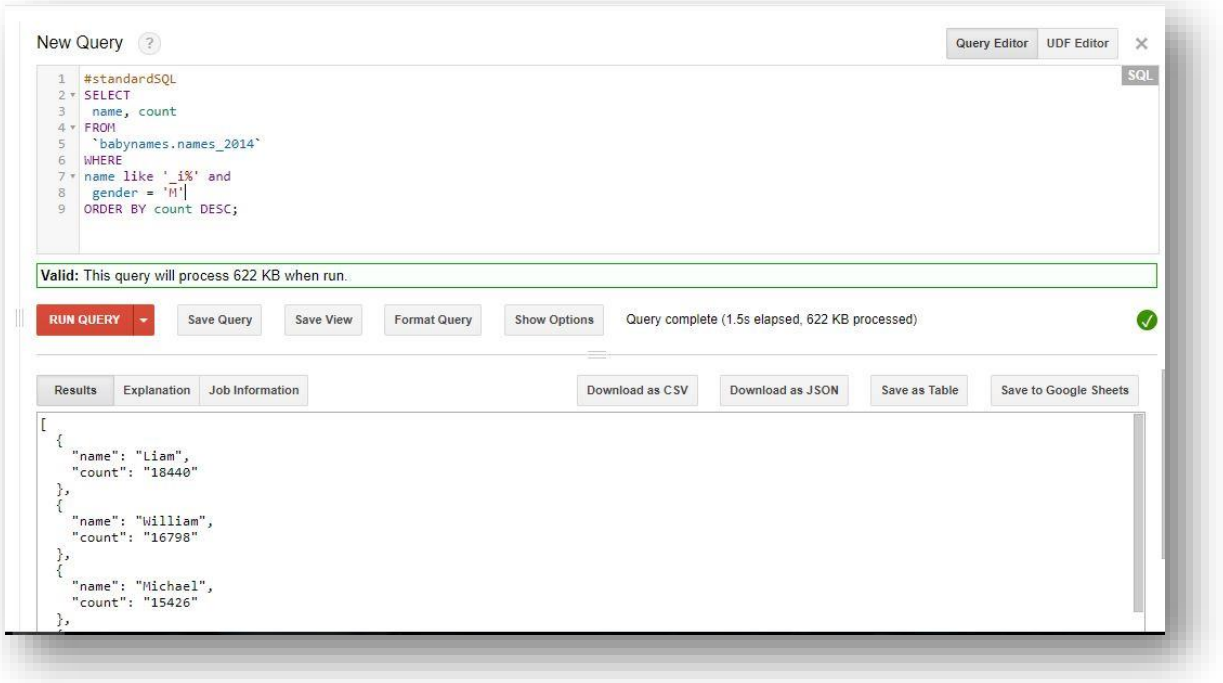
Name: Himanshu Purandare

USC ID: 7422196389

Part 1: Google BigQuery

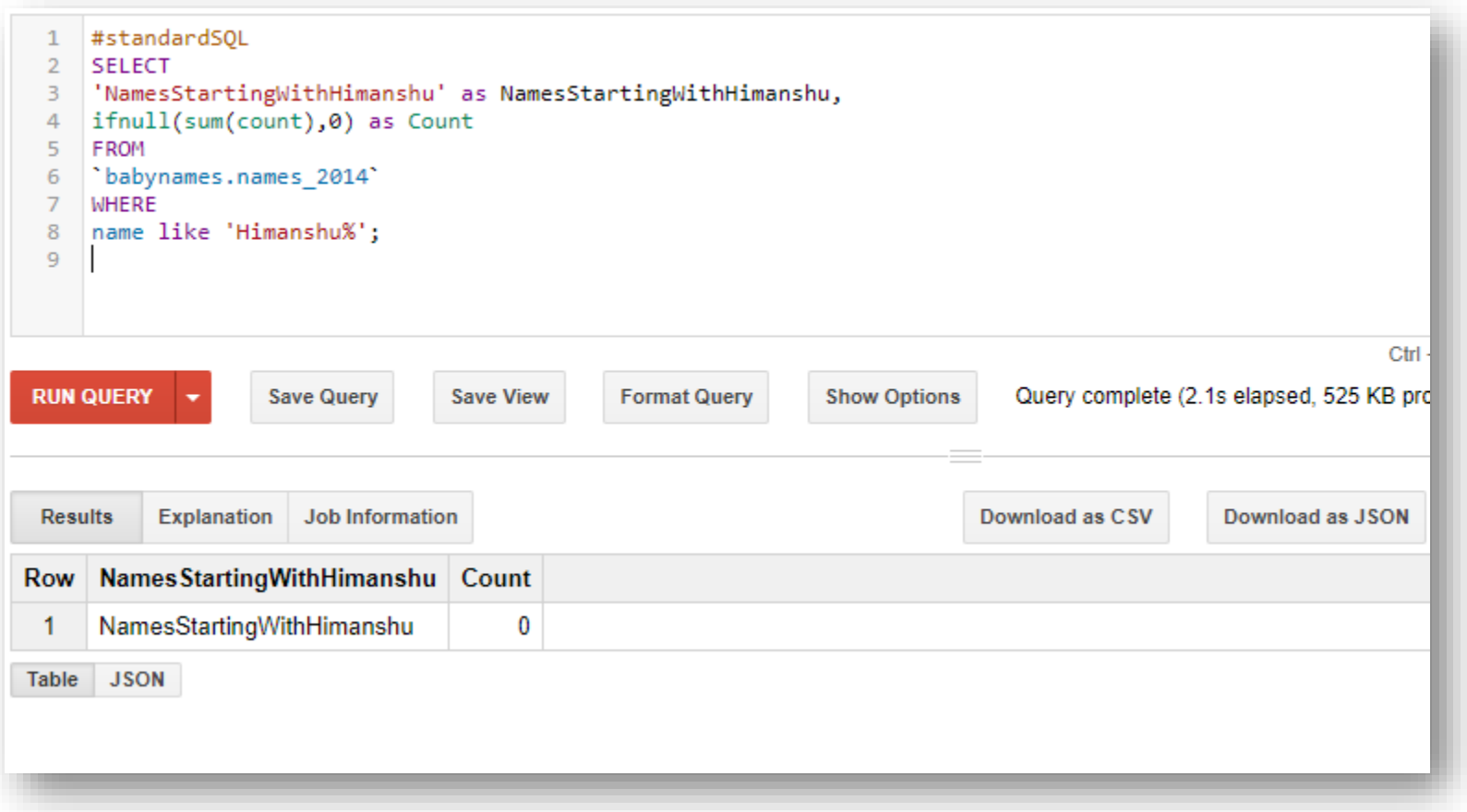
Query 1:

```
#standardSQL
SELECT
name, count
FROM `babynames.names_2014`
WHERE
name like '_i%' AND
gender = 'M'
ORDER BY count DESC;
```



Query2:

```
#standardSQL
SELECT
' NamesStartingWithHimanshu ' as NamesStartingWithHimanshu,
ifnull(sum(count),0) as Count
FROM
`babynames.names_2014`
WHERE
name like 'Himanshu%';
```



Part 2: DataLab and Notebooks

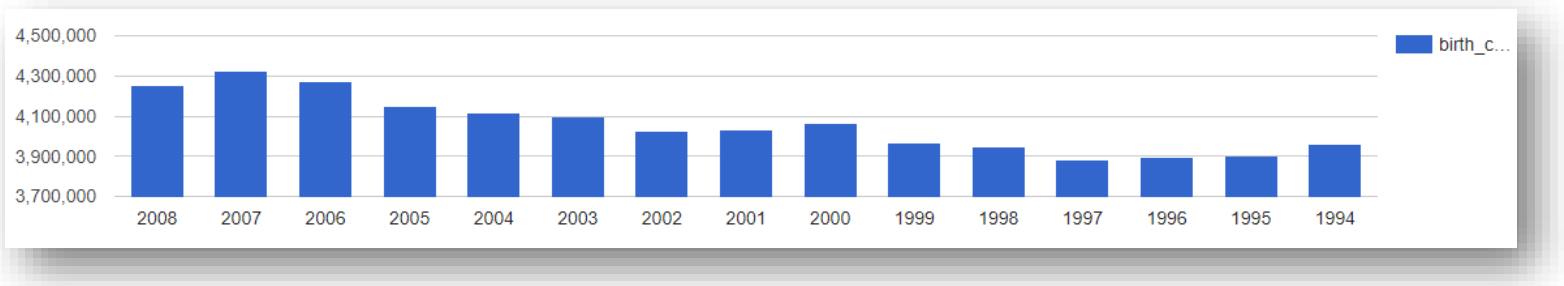
HW4_dataLab1:

Query 1:

```
%%bq query
#SQL query to return total births by year
SELECT CAST(source_year AS string) AS year, COUNT(is_male) AS birth_count
FROM `publicdata.samples.natality`
GROUP BY year
ORDER BY year DESC
LIMIT 15
```

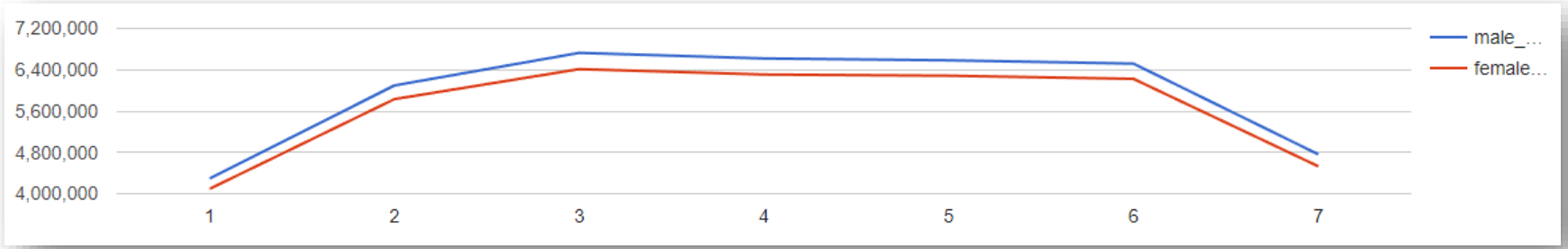
Query 2:

```
%%bq query --name total_births
#SQL query to return total births by year
SELECT CAST(source_year AS string) AS year, COUNT(is_male) AS birth_count
FROM `publicdata.samples.natality`
GROUP BY year
ORDER BY year DESC
LIMIT 15

%chart columns --data total_births --fields year,birth_count
```



Query 3:

```
%%bq query --name births_by_weekday
SELECT CAST(wday AS string) AS weekday, SUM(CASE WHEN is_male THEN 1 ELSE 0 END) AS male_births, SUM(CASE WHEN is_male THEN 0
ELSE 1 END) AS female_births
FROM `publicdata.samples.natality`
WHERE wday IS NOT NULL
GROUP BY weekday
ORDER BY weekday ASC

%chart line --data births_by_weekday --fields weekday,male_births,female_births
```



HW4_Datalab2

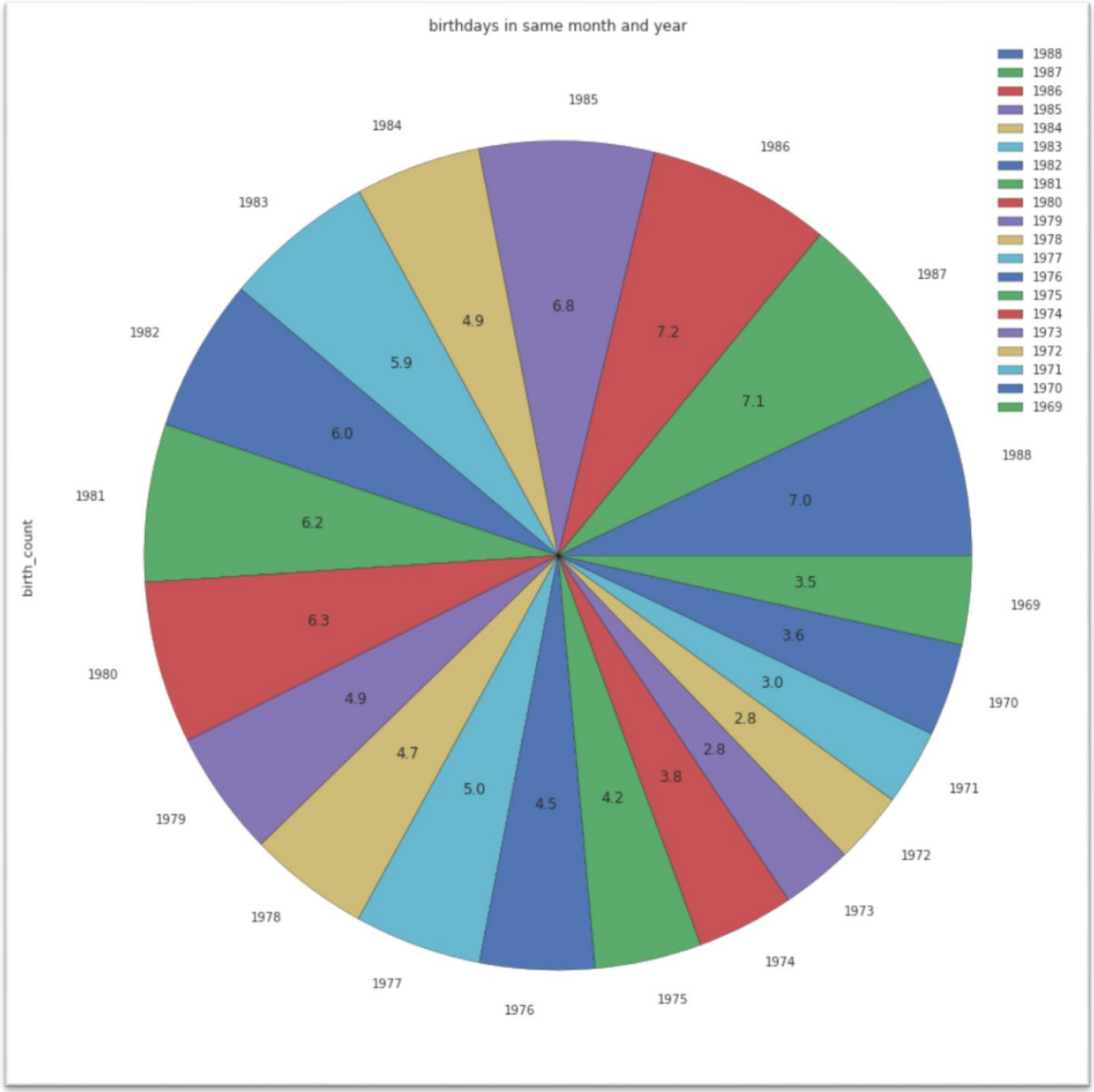
Cell 1:

Cell 2:

```
%%bq query
SELECT count(*) Number_of_People
FROM `publicdata.samples.natality`
WHERE source_year = 1994 and month=12;
```

Cell 3:

```
import google.datalab.bigquery as bq
import pandas as p

Birth_Count = bq.Query('SELECT CAST(source_year AS string) AS year, COUNT(is_male) AS birth_count FROM `publicdata.samples.natality`
WHERE month=12 AND day=02 GROUP BY year ORDER BY year DESC')
df = Birth_Count.execute(output_options=bq.QueryOutput.dataframe()).result()
list1=[]
list1=list(df.head(100).year)

df.plot.pie(labels=list1,x='year',y='birth_count', title = 'birthdays in same month and year',autopct='%.1f',figsize=(15, 15))
```


birthdays in same month and year
birth_count
1988
1987
1986
1985
1984
1983
1982
1981
1980
1979
1978
1977
1976
1975
1974
1973
1972
1971
1970
1969
7.0
7.1
7.2
6.8
4.9
5.9
6.0
6.2
6.3
4.9
4.7
5.0
4.5
4.2
3.8
2.8
2.8
3.0
3.6
3.5

Part 3: Big Public Data, Visualization and Interpretation

Query 1:

```
#standardSQL
SELECT
TIMESTAMP_TRUNC(pickup_datetime,DAY) as single_day,
SUM(passenger_count) as passengers_sum
FROM `nyc-tlc.yellow.trips`
WHERE pickup_datetime < "2015-01-01 00:00:00"
GROUP BY
 1
ORDER BY
 1;
```

```sql
#standardSQL
SELECT
TIMESTAMP_TRUNC(pickup_datetime,DAY) as single_day,
SUM(passenger_count) as passengers_sum
FROM `nyc-tlc.yellow.trips`
WHERE pickup_datetime < "2015-01-01 00:00:00"
GROUP BY
 1
ORDER BY
 1;
```

RUN QUERY | Save Query | Save View | Format Query | Show Options | Query

Results | Explanation | Job Information | Download

Row	single_day	passengers_sum
1	2009-01-01 00:00:00 UTC	602881
2	2009-01-02 00:00:00 UTC	696549
3	2009-01-03 00:00:00 UTC	811114
4	2009-01-04 00:00:00 UTC	667293
5	2009-01-05 00:00:00 UTC	609774

Table | JSON

First < Prev Rows 1 - 5 of 2191

HW4_nyc_taxi:

Query 1:

Erroneous Query:

```
import google.datalab.bigquery as bq
import pandas as p

bq.query('#standardSQL
SELECT
TIMESTAMP_TRUNC(pickup_datetime,DAY) as single_day,
SUM(passenger_count) as passengers_sum
FROM `nyc-tlc.yellow.trips`
where pickup_datetime < "2015-01-01 00:00:00"
GROUP BY
 1
ORDER BY
 1;
')
```

```
import google.datalab.bigquery as bq
import pandas as p
```

```
bq.query('#standardSQL
SELECT
TIMESTAMP_TRUNC(pickup_datetime,DAY) as single_day,
SUM(passenger_count) as passengers_sum
FROM `nyc-tlc.yellow.trips`
where pickup_datetime < "2015-01-01 00:00:00"
GROUP BY
 1
ORDER BY
 1;
')
```

```
      File "<ipython-input-4-d62027e16e18>", line 1
        bq.query('#standardSQL
                               ^
    SyntaxError: EOL while scanning string literal
```

Error Reason:

The error occured because I did not paste the bigQuery in a single line string. Instead I pasted it as a multiline string which is not accepted. Hence the error. The corrected version of the datalab query is as below:

Correct Version:

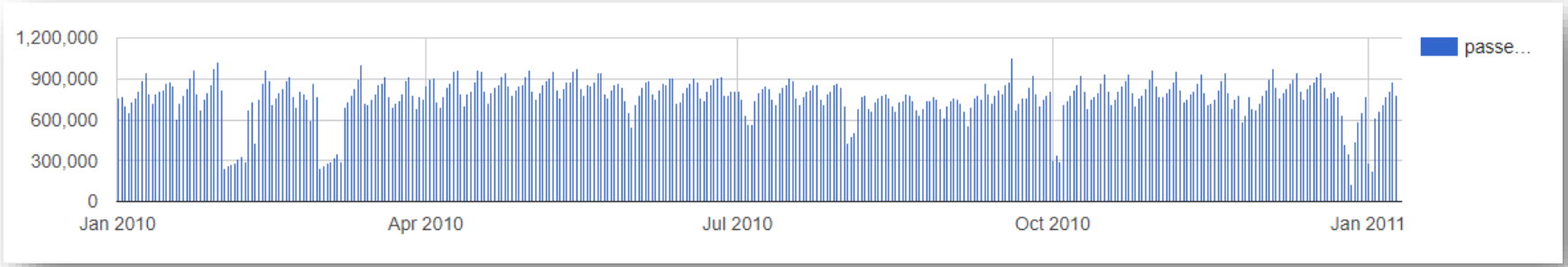
```
Passenger_sum = bq.Query('SELECT TIMESTAMP_TRUNC(pickup_datetime,DAY) as single_day, SUM(passenger_count) as passengers_sum
FROM `nyc-tlc.yellow.trips` where pickup_datetime <= "2015-01-01" GROUP BY 1 ORDER BY 1')
df = Passenger_sum.execute(output_options=bq.QueryOutput.dataframe()).result()
df.head(2191)
```

Query 2: the total count of passengers for 2009

```
%%bq query --name passenger_1
SELECT
TIMESTAMP_TRUNC(pickup_datetime,DAY) as single_day,
SUM(passenger_count) as passengers_sum
FROM `nyc-tlc.yellow.trips` where pickup_datetime >= '2009-01-01' and pickup_datetime <= '2010-01-10' GROUP BY 1 ORDER BY 1;

%chart line --data passenger_1 --fields single_day,passengers_sum
```

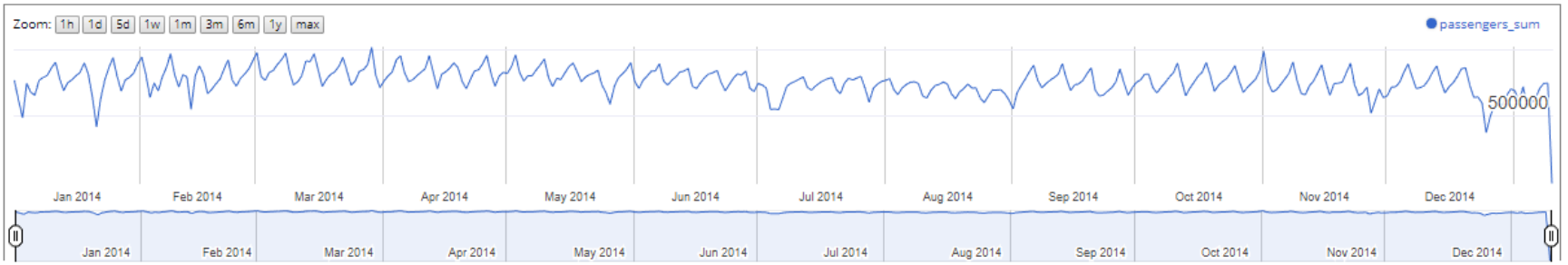
Query 3: the total count of passengers for 2010

```
%%bq query --name passenger_2
SELECT
TIMESTAMP_TRUNC(pickup_datetime,DAY) as single_day,
SUM(passenger_count) as passengers_sum
FROM `nyc-tlc.yellow.trips` where pickup_datetime >= '2010-01-01' and pickup_datetime <= '2011-01-10' GROUP BY 1 ORDER BY 1;

%chart columns --data passenger_2 --fields single_day,passengers_sum
```



Query 4: the total count of passengers for 2014

```
%%bq query --name passenger_3
SELECT
TIMESTAMP_TRUNC(pickup_datetime,DAY) as single_day,
SUM(passenger_count) as passengers_sum
FROM `nyc-tlc.yellow.trips` where pickup_datetime >= '2014-01-01' and pickup_datetime <= '2015-01-10' GROUP BY 1 ORDER BY 1;

%chart annotation --data passenger_3 --fields single_day,passengers_sum
```



Q: General semi-periodical pattern hypothesis:

A: There exists a semi periodic pattern in the graph of number of passengers travelled per day in yellow taxi in New York City. Here we can observe following things:

- The number of days between any two consecutive local maxima or any two consecutive local minima is usually around 7 days for all three years.
- Usually the local maxima occurred usually on weekends.
- The local minima occurred usually on Mondays.
- We can write a query to find out all maxima or all minima though the graph. We can then show that the difference between any two consecutive local minima or two consecutive local maxima is always around 7.

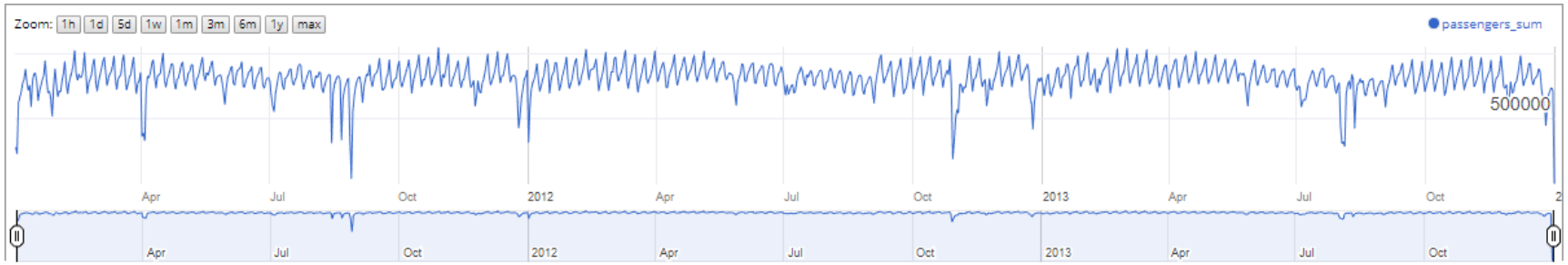
Q: Unusual patterns (anomaly) being repeated in all three figures:

A: We see two anomalies in these three graphs:

- Due to Martin Luther King Day, which is a holiday in USA, the number of passengers travelling in the period around 15th-20th January every year are very less.
- Due to the Christmas holidays, the number of passengers travelling in the last week of December is prominently less.

Query 5: Visualize the complete data for year 2011, 2012, and 2013.

```
%%bq query --name passenger_4
SELECT
TIMESTAMP_TRUNC(pickup_datetime,DAY) as single_day,
SUM(passenger_count) as passengers_sum
FROM `nyc-tlc.yellow.trips` where pickup_datetime >= '2011-01-01' and pickup_datetime <= '2013-12-31' GROUP BY 1 ORDER BY 1;

%chart annotation --data passenger_4 --fields single_day,passengers_sum
```



Q: Find the minimum point (you can query this or just find it manually):

A: Here we can see 3 prominent drops in the number of passengers travelled over the period of 3 years:

1. **28th August, 2011:** Hurricane Irene hit New York which decreased the number of passengers travelling in New York.
2. **29th October, 2012:** Hurricane Sandy hits the east coast of the United States, killing 148 directly and 138 indirectly, while leaving nearly $70 billion in damages and causing major power outages.
3. **03rd August, 2013:** The reason that the number of passengers travelled in yellow cab were less for few days:
    a. In August 2013 a new class of taxi was allowed on the road: referred to in regulation as “street-hail livery” (SHL) taxis, or alternatively as “boro” or green taxis, these taxis and their drivers are regulated as yellow taxis, with the exception that passenger pickups (whether pre-arranged or street-hails) throughout the five boroughs with the exclusion of the two city airports and Manhattan south of East 96th Street and West 110th Street, which was referred to as the “hail-exclusion zone.”
    b. However, the supply of boro taxis was rolled-out gradually: only 114 boro taxis carried out at least one trip during the program’s first month of August 2013, and hence passengers travelling in yellow cab increased again.

Bonus Part:

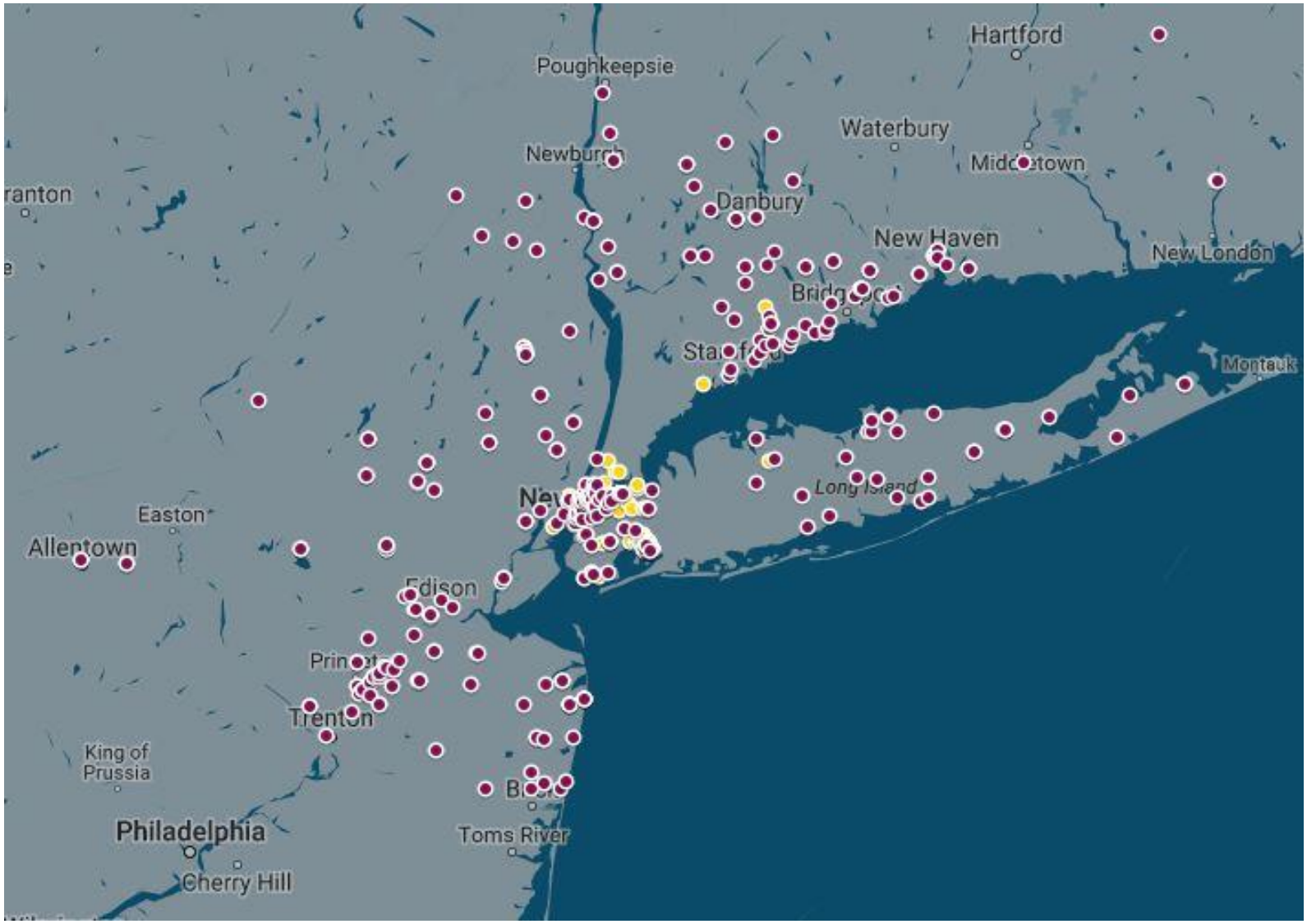
Query to find pickup longitude-latitude:

```
%%bq query
SELECT
pickup_longitude,
pickup_latitude
FROM `nyc-tlc.yellow.trips`
where pickup_datetime >= '2013-01-01' and
dropoff_datetime <= '2013-12-31' and
total_amount>=300 and
total_amount<=400 and
cast(pickup_datetime as time) >= '18:00:00' and
pickup_longitude <> 0.0 and
pickup_latitude <> 0.0 and
dropoff_longitude <> 0.0 and
dropoff_latitude <> 0.0
ORDER BY 1;
```

Query to find drop-off longitude-latitude:

```
%%bq query
SELECT
dropoff_longitude,
dropoff_latitude
FROM `nyc-tlc.yellow.trips`
where pickup_datetime >= '2013-01-01' and
dropoff_datetime <= '2013-12-31' and
total_amount>=300 and
total_amount<=400 and
cast(pickup_datetime as time) >= '18:00:00' and
pickup_longitude <> 0.0 and
pickup_latitude <> 0.0 and
dropoff_longitude <> 0.0 and
dropoff_latitude <> 0.0
ORDER BY 1;
```

Google Maps screenshot:



- The output of the above queries were pickup and drop-off locations (latitude and longitude) which was exported in csv files.
- I added numbers to the points in the CSV files.
- Then I uploaded it to myMaps in google maps and got this output. Pickup points are yellow in color and drop-off points are red in color.