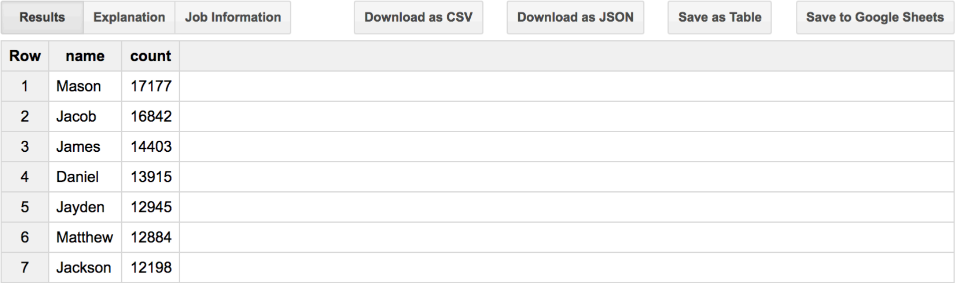
CSCI585 HW4 Report

NAME: Hao Wu
USCID:1699530173
EMAIL:hwu638@usc.edu

Part 1: Google BigQuery

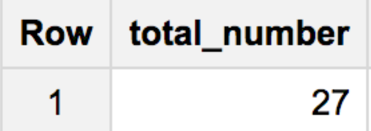
Query 1**: select name, count from babynames.names_2014 where gender = 'M' and name like '_a%' order by coun**



Row	name	count
1	Mason	17177
2	Jacob	16842
3	James	14403
4	Daniel	13915
5	Jayden	12945
6	Matthew	12884
7	Jackson	12198

Query 2**:**

select sum(count) as total_number from babynames.names_2014 where name like 'Hao%'



Row	total_number
1	27

Part 2: DataLab and Notebooks

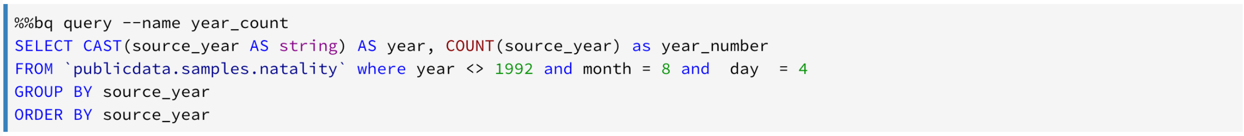
Query in the 2nd cell**:**

```
%%bq query
SELECT wday
FROM `publicdata.samples.natality` where year = 1992 and month = 8 and day  = 4
```

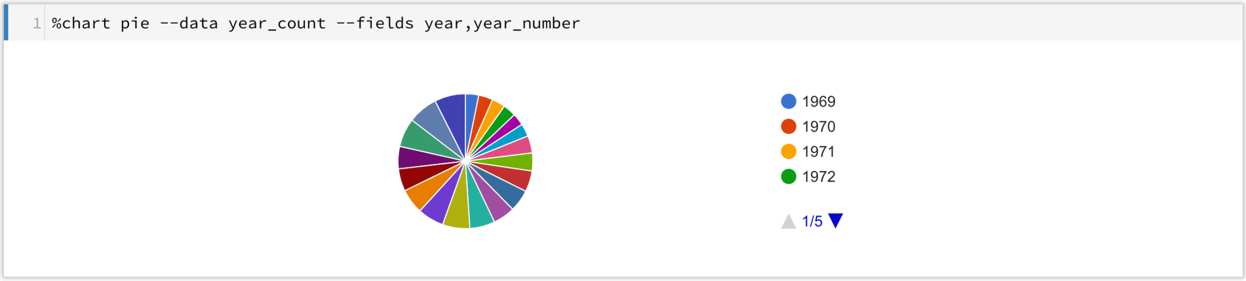
Query in the 3rd cell:

```
%%bq query --name year_count
SELECT CAST(source_year AS string) AS year, COUNT(source_year) as year_number
FROM `publicdata.samples.natality` where year <> 1992 and month = 8 and  day  = 4
GROUP BY source_year
ORDER BY source_year
```

%chart pie --data year_count --fields year,year_number



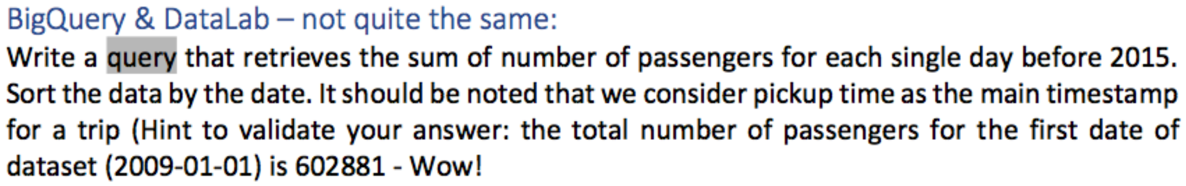
```
%%bq query --name year_count
SELECT CAST(source_year AS string) AS year, COUNT(source_year) as year_number
FROM `publicdata.samples.natality` where year <> 1992 and month = 8 and  day  = 4
GROUP BY source_year
ORDER BY source_year
```



```
%chart pie --data year_count --fields year,year_number
```

Part 3: Big Public Data, Visualization and Interpretation

Query for this Question:



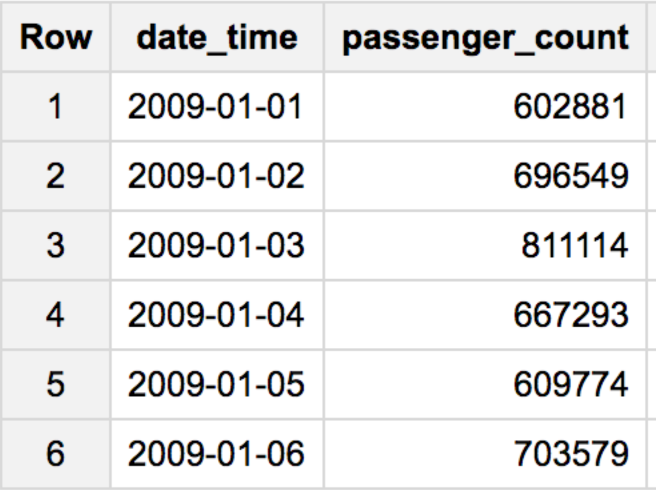
BigQuery & DataLab – not quite the same:

Write a query that retrieves the sum of number of passengers for each single day before 2015. Sort the data by the date. It should be noted that we consider pickup time as the main timestamp for a trip (Hint to validate your answer: the total number of passengers for the first date of dataset (2009-01-01) is 602881 - Wow!

```
%%bq query --name passenger_count_by_date
#standardSQL
SELECT date_time as day,passenger_count as number
FROM
(SELECT
DATE(pickup_datetime) as date_time
,
SUM(Passenger_count) as passenger_count
FROM
 `bigquery-public-data.new_york.tlc_yellow_trips_2010`
GROUP BY
 date_time
UNION ALL
SELECT
DATE(pickup_datetime) as date_time
,
SUM(Passenger_count) as passenger_count
FROM
```

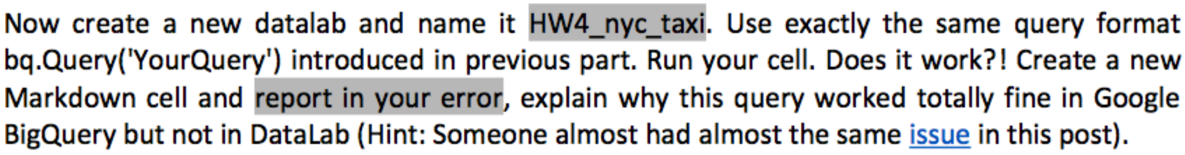
```
  `bigquery-public-data.new_york.tlc_yellow_trips_2009`
GROUP BY
  date_time
UNION ALL
SELECT
DATE(pickup_datetime) as date_time
,
SUM(Passenger_count) as passenger_count
FROM
  `bigquery-public-data.new_york.tlc_yellow_trips_2011`
GROUP BY
  date_time
UNION ALL
SELECT
DATE(pickup_datetime) as date_time
,
SUM(Passenger_count) as passenger_count
FROM
  `bigquery-public-data.new_york.tlc_yellow_trips_2012`
GROUP BY
  date_time
UNION ALL
SELECT
DATE(pickup_datetime) as date_time
,
SUM(Passenger_count) as Passenger_count
FROM
  `bigquery-public-data.new_york.tlc_yellow_trips_2013`
GROUP BY
  date_time
UNION ALL
SELECT
DATE(pickup_datetime) as date_time
,
SUM(Passenger_count) as Passenger_count
FROM
  `bigquery-public-data.new_york.tlc_yellow_trips_2014`
```

```
GROUP BY
  date_time
)
ORDER BY
  date_time
```



Row	date_time	passenger_count
1	2009-01-01	602881
2	2009-01-02	696549
3	2009-01-03	811114
4	2009-01-04	667293
5	2009-01-05	609774
6	2009-01-06	703579

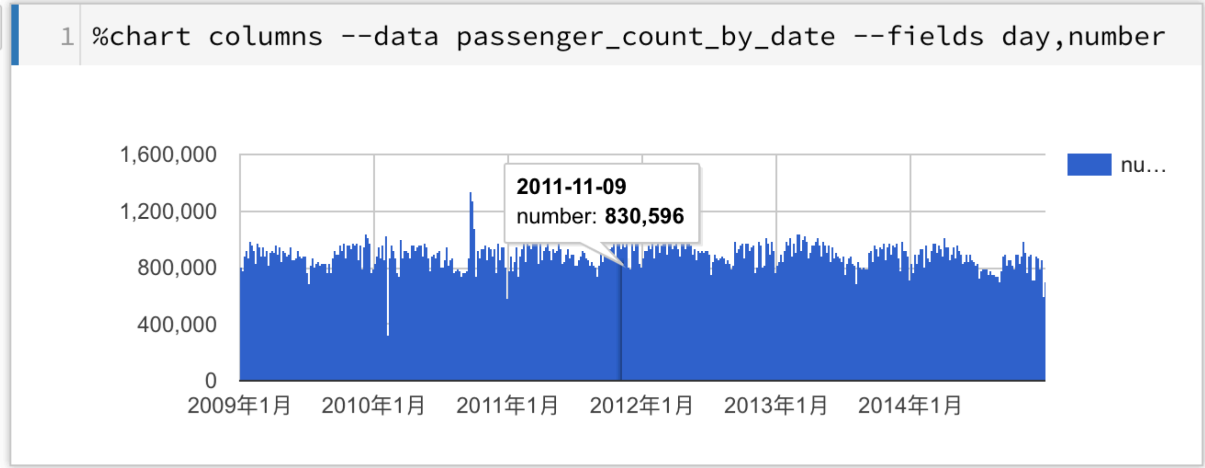
For the question:



Now create a new datalab and name it HW4_nyc_taxi. Use exactly the same query format bq.Query('YourQuery') introduced in previous part. Run your cell. Does it work?! Create a new Markdown cell and report in your error, explain why this query worked totally fine in Google BigQuery but not in DataLab (Hint: Someone almost had almost the same issue in this post).

I didn't encounter any issues.

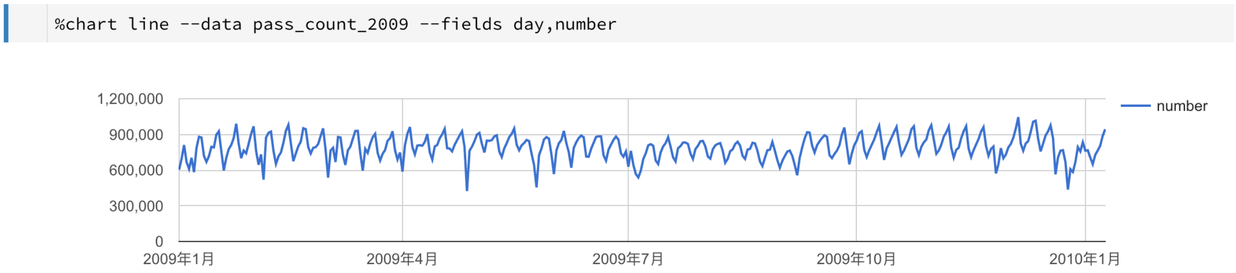
panda dataframe:



```
%chart columns --data passenger_count_by_date --fields day,number
```

Visualization:

2009



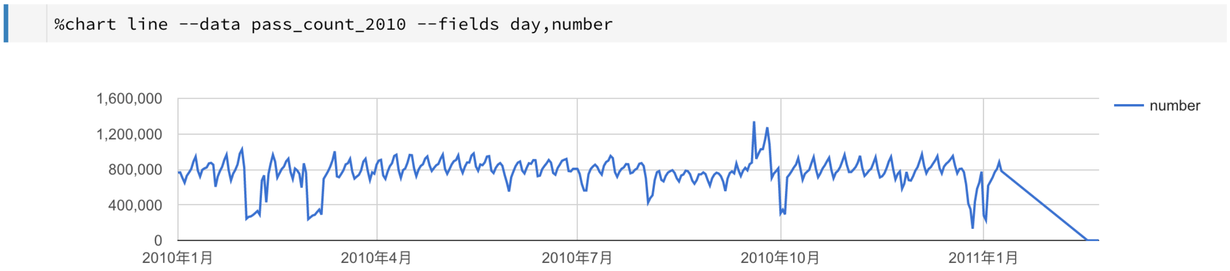
```
%chart line --data pass_count_2009 --fields day,number
```

Query:

```
%%bq query --name pass_count_2009
#standardSQL
select DATE as day ,total_passenger as number From(
select DATE(pickup_datetime) as DATE, sum(passenger_count) as
total_passenger
from `bigquery-public-data.new_york.tlc_yellow_trips_2009`
Group by
 DATE
UNION ALL
select DATE(pickup_datetime) as DATE, sum(passenger_count) as
total_passenger
from   `bigquery-public-data.new_york.tlc_yellow_trips_2010`
where DATE(pickup_datetime) < '2010-1-10'
Group by
 DATE)
Order by
Date
```

```
%chart line --data pass_count_2009 --fields day,number
```

2010



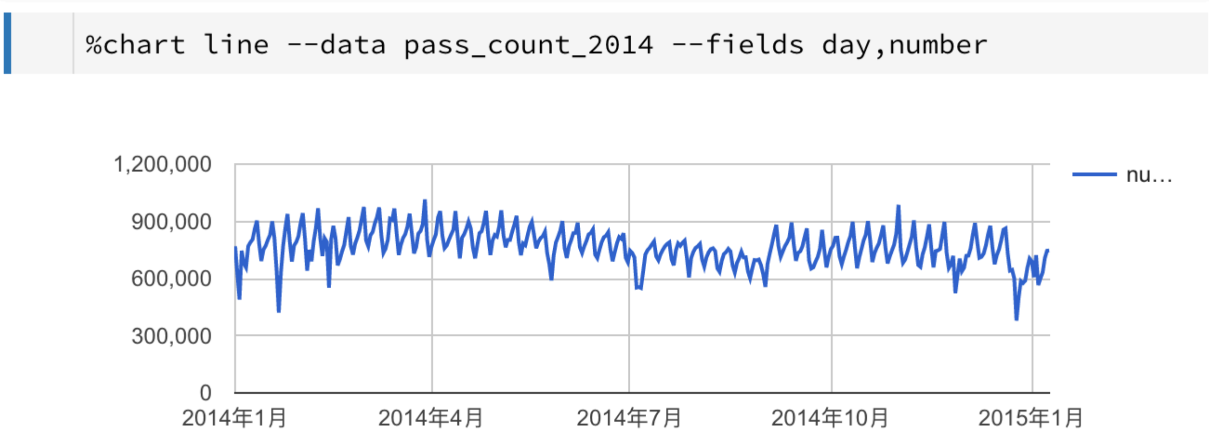
```
%chart line --data pass_count_2010 --fields day,number
```

Query:

```
%%bq query --name pass_count_2010
#standardSQL
select DATE as day ,total_passenger as number From(
select DATE(pickup_datetime) as DATE, sum(passenger_count) as
total_passenger
from `bigquery-public-data.new_york.tlc_yellow_trips_2010`
Group by
 DATE
UNION ALL
select DATE(pickup_datetime) as DATE, sum(passenger_count) as
total_passenger
from   `bigquery-public-data.new_york.tlc_yellow_trips_2011`
where DATE(pickup_datetime) < '2011-1-10'
Group by
 DATE)
Order by
Date
```

%chart line --data pass_count_2010 --fields day,number

2014



```
%chart line --data pass_count_2014 --fields day,number
```

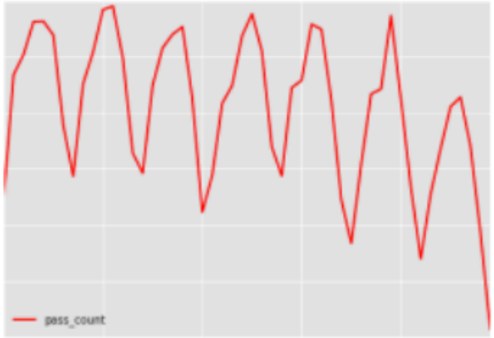
Query:

```
%%bq query --name pass_count_2014
#standardSQL
select DATE as day ,total_passenger as number
from
(

   select DATE(pickup_datetime) as DATE, sum(passenger_count) as 
total_passenger
          from   `bigquery-public-data.new_york.tlc_yellow_trips_2014`
          Group by
          DATE
UNION ALL
```

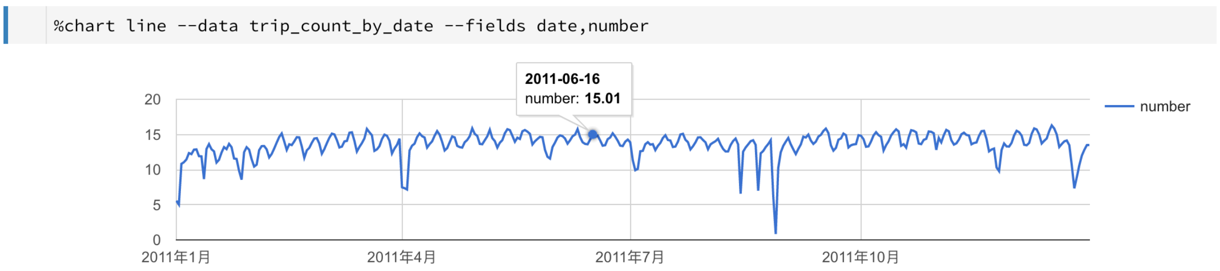
```
        select DATE(pickup_datetime) as DATE,sum( passenger_count) as
total_passenger
        from   `bigquery-public-data.new_york.tlc_yellow_trips_2015`
        where DATE(pickup_datetime) < '2015-1-10'
        Group by
        DATE

)
Order by
 DATE

%chart line --data pass_count_2014 --fields day,number
```

For question:

Can you find a general semi-periodical pattern in the data like the figure below? Explain the pattern. Without writing it, suggest a query that proves your hypothesis and report it (1 point).



Explanation: The data has fluctuated widely by some reasons (like holiday, ceremony).



```
%chart line --data trip_count_by_date --fields date,number
```

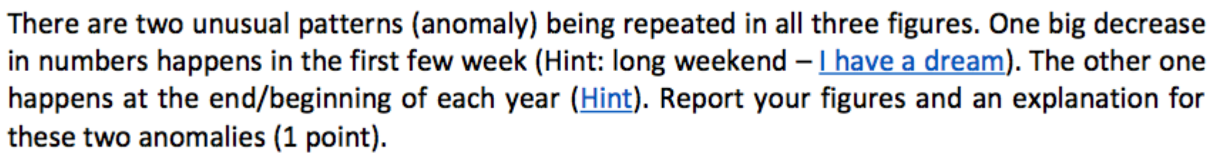
Query:

%%bq query --name trip_count_by_date

```
#standardSQL
SELECT
DATE(pickup_datetime) as date
,
SUM(Trip_distance)/100000 as number
FROM
  `bigquery-public-data.new_york.tlc_yellow_trips_2011`
GROUP BY
  date
ORDER BY
  Date
```

```
%chart line --data trip_count_by_date --fields date,number
```

For question:

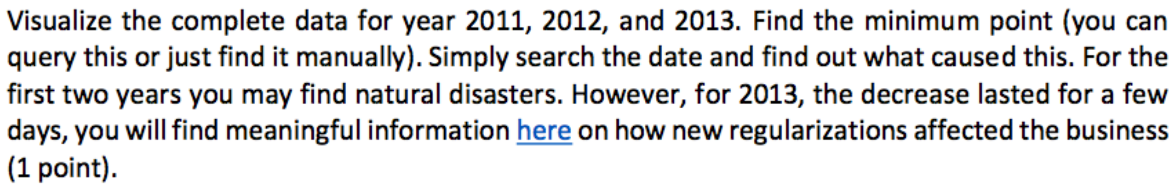


There are two unusual patterns (anomaly) being repeated in all three figures. One big decrease in numbers happens in the first few week (Hint: long weekend – I have a dream). The other one happens at the end/beginning of each year (Hint). Report your figures and an explanation for these two anomalies (1 point).

The big decrease is happened at Martin Luther King Day. In that day, most people have spare time to go out. So, there should be a lot of people using taxi.

The other one is happened during Christmas Day. Same reason for it, people will take taxi more often than usual on holiday. Another reason is may that lots of people go to New York for travel.

For question:



Visualize the complete data for year 2011, 2012, and 2013. Find the minimum point (you can query this or just find it manually). Simply search the date and find out what caused this. For the first two years you may find natural disasters. However, for 2013, the decrease lasted for a few days, you will find meaningful information here on how new regularizations affected the business (1 point).

For 2011, the minimum point is at the day August 28. This is caused by hurricane Irene.

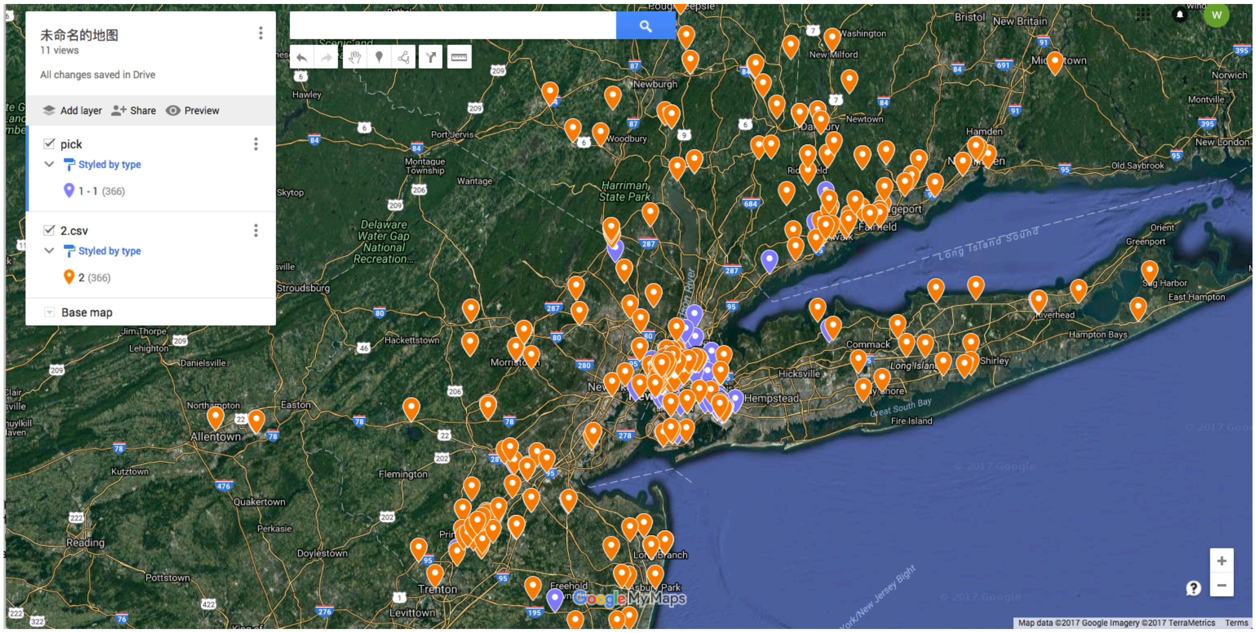
For 2012, the minimum point is at the day Nov. 29. This is caused by hurricane sandy.

For 2013, the minimum point is at the day August 04. The reason for this is because **"New York City licensed a new type of taxi in Aug. 2013: "boro" taxis are restricted from picking up passengers in Manhattan south of a boundary along East 96th Street and West 110th Street."**

Bonus Part:

```
#standardSQL
SELECT
pickup_datetime as date_time,
Total_amount as amount,
Pickup_longitude as p_lo,
Pickup_latitude as p_la,
Dropoff_longitude as d_lo,
Dropoff_latitude as d_la
FROM
  `bigquery-public-data.new_york.tlc_yellow_trips_2013`
where Total_amount between 300 and 400 and  extract(hour from
pickup_datetime) > 18
```

I use the above query to get the pickup longitude(p_lo), pickup latitude(p_la), dropoff_longitude(d_lo) and dropoff_latitude(d_la). Then I save the result as csv format and import the csv file into google map. The snapshot is in the below.



未命名的地图
11 views

All changes saved in Drive

☑ pick
Styled by type
1 - 1 (366)

☑ 2.csv
Styled by type
2 (366)

Base map