**CSCI585 HW4 Report**

**NAME: Hao Wu**

**USCID:1699530173**

**EMAIL:hwu638@usc.edu**

**Part 1: Google BigQuery**

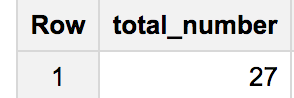
Query 1**: select name, count from babynames.names\_2014**

**where gender = 'M' and name like '\_a%' order by coun**

Query 2**:**

**select sum(count) as total\_number from babynames.names\_2014**

**where name like 'Hao%'**

****

**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**

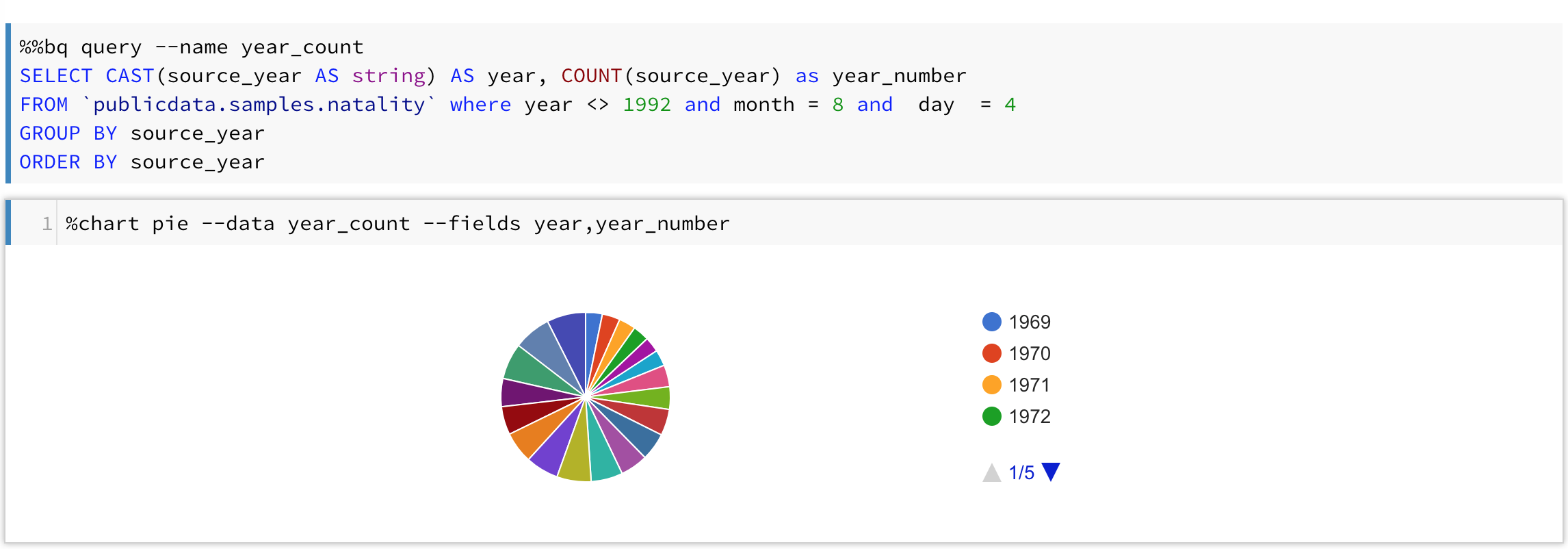
**SELECT CAST(source\_year AS string) AS year, COUNT(source\_year) as year\_number**

**FROM `publicdata.samples.natality` where year <> 1992 and month = 8 and day = 4**

**GROUP BY source\_year**

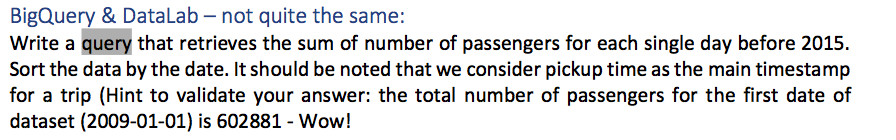
**ORDER BY source\_year**

**%chart pie --data year\_count --fields year,year\_number**

****

**Part 3: Big Public Data, Visualization and Interpretation**

**Query for this Question:**

****

#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\_time, sum(Passenger\_count) as passenger\_count FROM

(SELECT date\_time, Passenger\_count

FROM

( 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\_time, Passenger\_count

FROM

( SELECT

DATE(pickup\_datetime) as date\_time

,

SUM(Passenger\_count) as Passenger\_count

FROM

`bigquery-public-data.new\_york.tlc\_green\_trips\_2013`

GROUP BY

date\_time))

GROUP BY

date\_time

UNION ALL

SELECT date\_time, sum(Passenger\_count) as passenger\_count FROM

(SELECT date\_time, Passenger\_count

FROM

( 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)

UNION ALL

SELECT date\_time, Passenger\_count

FROM

( SELECT

DATE(pickup\_datetime) as date\_time

,

SUM(Passenger\_count) as Passenger\_count

FROM

`bigquery-public-data.new\_york.tlc\_green\_trips\_2014`

GROUP BY

date\_time))

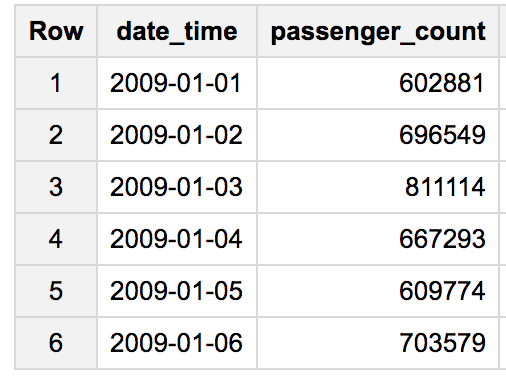
GROUP BY

date\_time

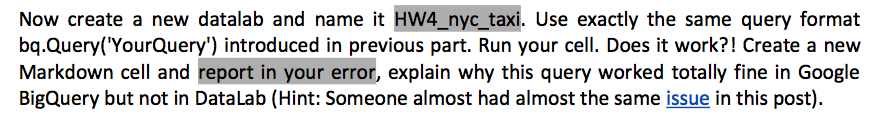
)

ORDER BY

date\_time

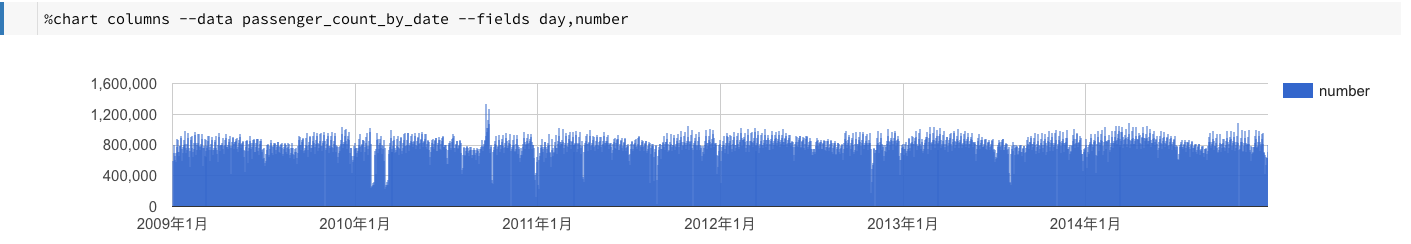


For the question:



**I didn’t encounter any issues.**

**panda dataframe:**

****

**Visualization:**

**2009**

****

**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

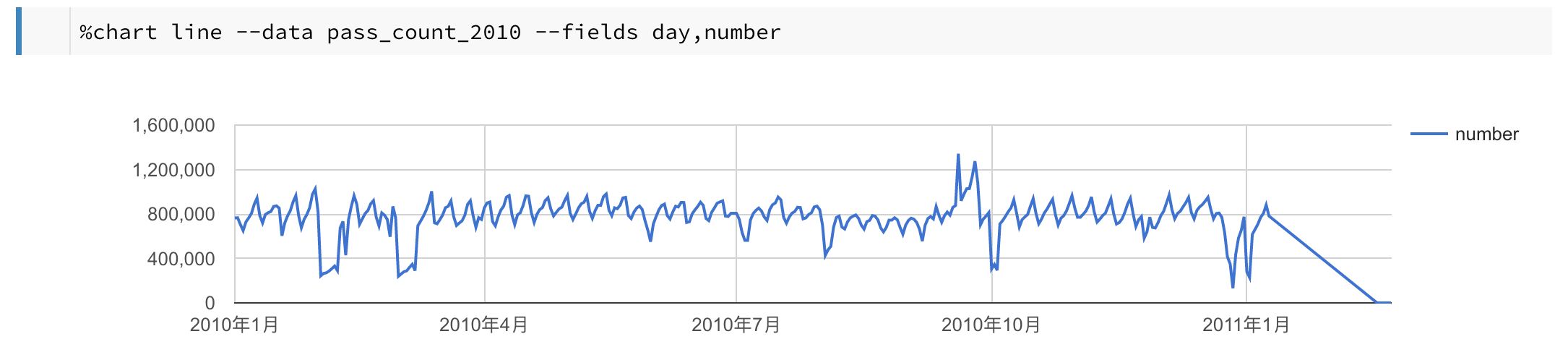
DATE)

Order by

Date

%chart line --data pass\_count\_2009 --fields day,number

**2010**

****

**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

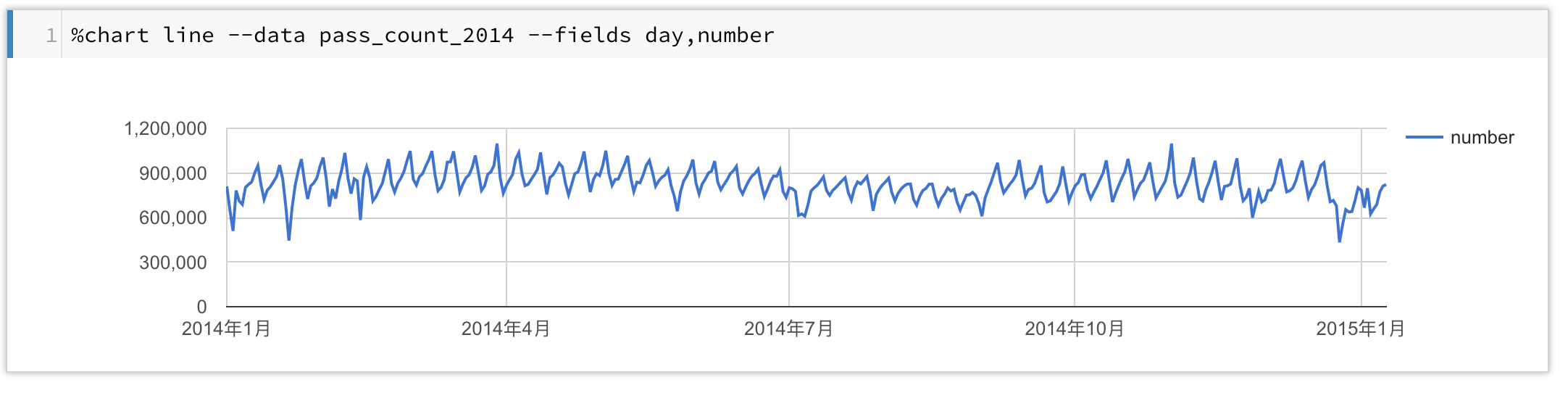
DATE)

Order by

Date

%chart line --data pass\_count\_2010 --fields day,number

**2014**

****

**Query:**

%%bq query --name pass\_count\_2014

#standardSQL

select DATE as day ,total\_passenger as number

from

(

select DATE , sum(total\_passenger) AS total\_passenger 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\_green\_trips\_2014`

Group by

DATE

)

GROUP BY

DATE

UNION ALL

select DATE , sum(total\_passenger) AS total\_passenger from(

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

UNION ALL

select DATE(pickup\_datetime) as DATE, sum(passenger\_count) as total\_passenger

from `bigquery-public-data.new\_york.tlc\_green\_trips\_2015`

where DATE(pickup\_datetime) < '2015-1-10'

Group by

DATE

)

GROUP BY

DATE

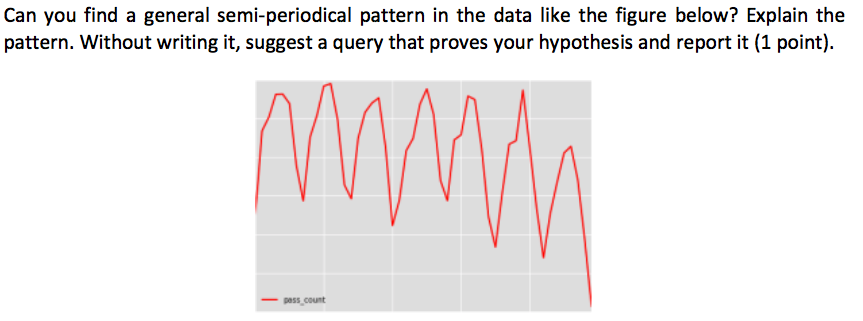
)

Order by

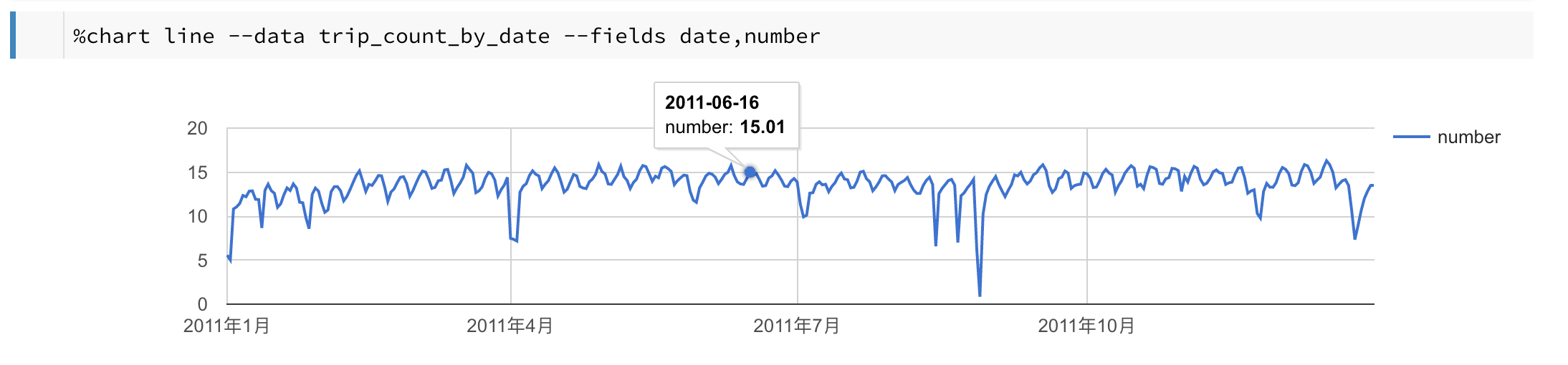
DATE

%chart line --data pass\_count\_2014 --fields day,number

**For question:**



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



**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

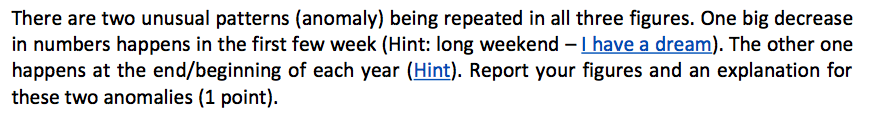
date

ORDER BY

Date

%chart line --data trip\_count\_by\_date --fields date,number

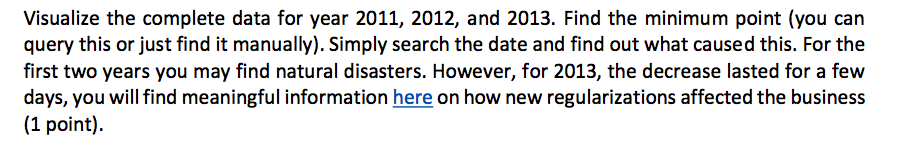
For question:



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:



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 date\_time, amount,p\_lo,p\_la,d\_la,d\_lo

FROM

( 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

)

UNION ALL

SELECT date\_time, amount ,p\_lo,p\_la,d\_la,d\_lo

FROM

(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\_green\_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.

