

BT2102 – Lab 3

Kaustubh Jagtap – A0168820B



national university of singapore

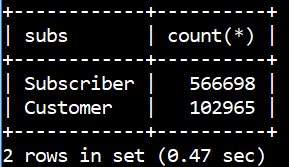
**Queries:**

4a) What were the number of trips per customer type?

Query:

SELECT subscription\_type AS subs, COUNT(\*)   
FROM trip WHERE duration <= 86400  
GROUP BY subs;

ANS:

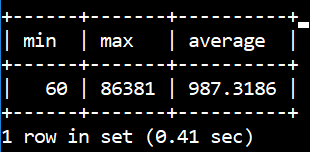


4b) What were the shortest, average and longest trip duration?

Query:

MIN(duration) AS min, MAX(duration) AS max, AVG(duration)   
AS average FROM trip  
WHERE duration <= 86400;

ANS:

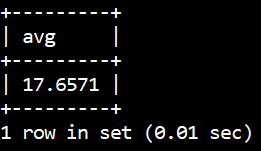


4c) On average, how many docks are at each station?

Query:

SELECT AVG(dock\_count) as avg from station;

Ans:

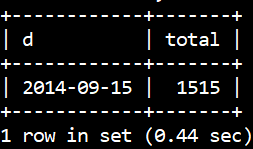


4d) Which was the busiest day of all, in terms of trips taken?

Query:

SELECT subq1.d, COUNT(subq1.d) AS total FROM  
(SELECT cast(start\_date as date) as d   
FROM trip WHERE duration<=86400)   
AS subq1  
GROUP BY subq1.d  
ORDER BY total DESC  
LIMIT 1;

ANS:

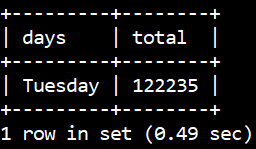


4e) Which was the busiest day of the week?

Query:

SELECT subq1.days, count(subq1.days) as total   
FROM (select DAYNAME(CAST(start\_date as date)) AS days   
FROM trip WHERE duration<=86400)   
AS subq1  
GROUP BY days  
ORDER BY total DESC  
LIMIT 1;

ANS:

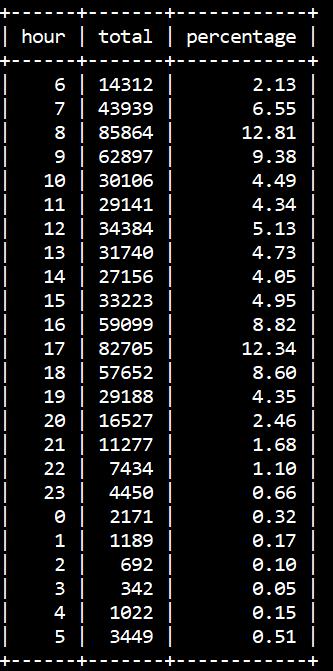


4f) For this question, I wanted to find some meaningful information that could be monetized. One potential source of additional revenue could be from peak surcharge during the busiest hours. Hence, the question I arrived at was: **What is the breakdown of number of rides by hour of the day?** Addressing this question also allows us to smoothen our operational capabilities, by planning for the optimal number of bikes to cater for peak demans.

Query:

SELECT subq1.hour, COUNTsubq1.hour) as total, TRUNCATE(COUNT(subq1.hour)/(SELECT COUNT(\*) from trip) \* 100, 2)  
AS percentage  
FROM (SELECT HOUR(start\_date) as hour from trip)   
AS subq1  
GROUP BY subq1.hour  
ORDER BY subq1.hour < 6, subq1.hour;

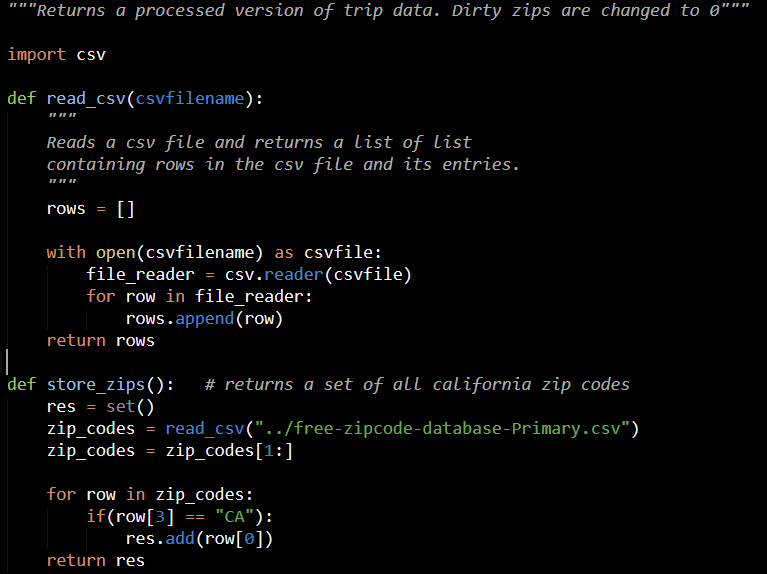
Ans:

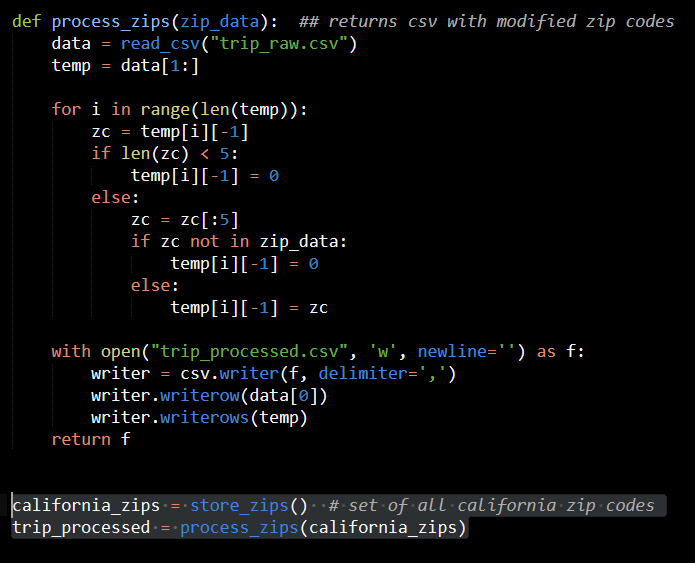


**Data Processing – trip**

I found a list of all California zip codes online, on the following website: <http://federalgovernmentzipcodes.us/>

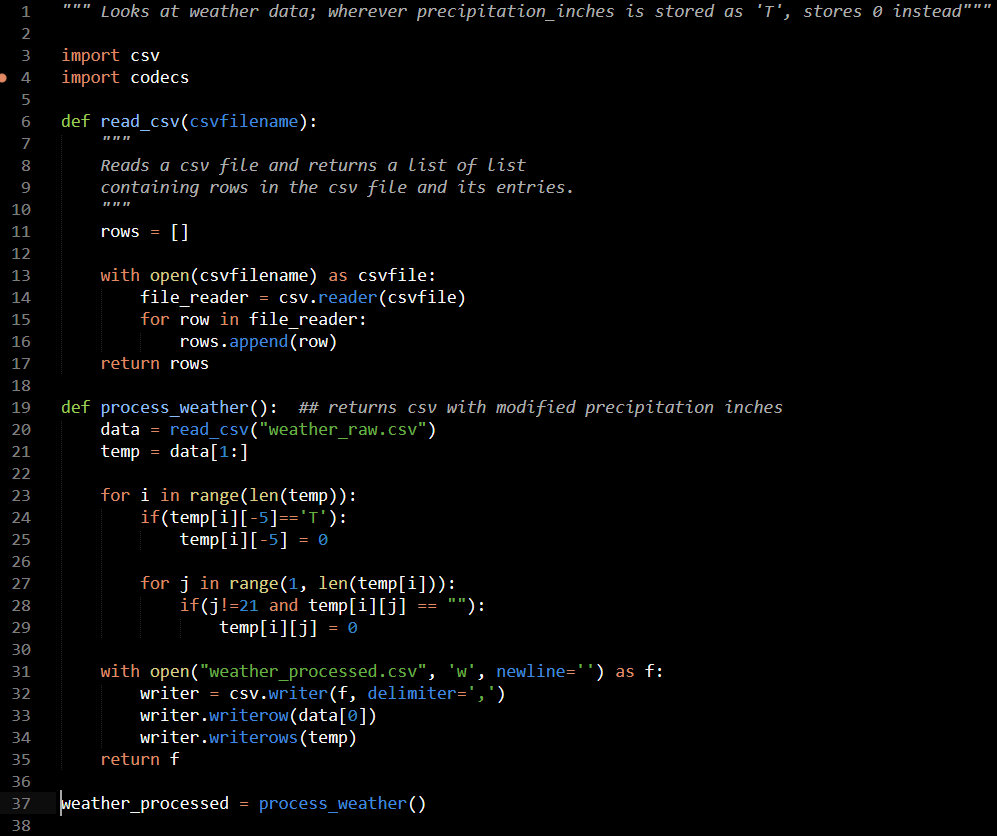
To filter out the wrong zip codes, I matched all the entries with this list and altered the trips data. This was output into a new csv file, from which the data was loaded.



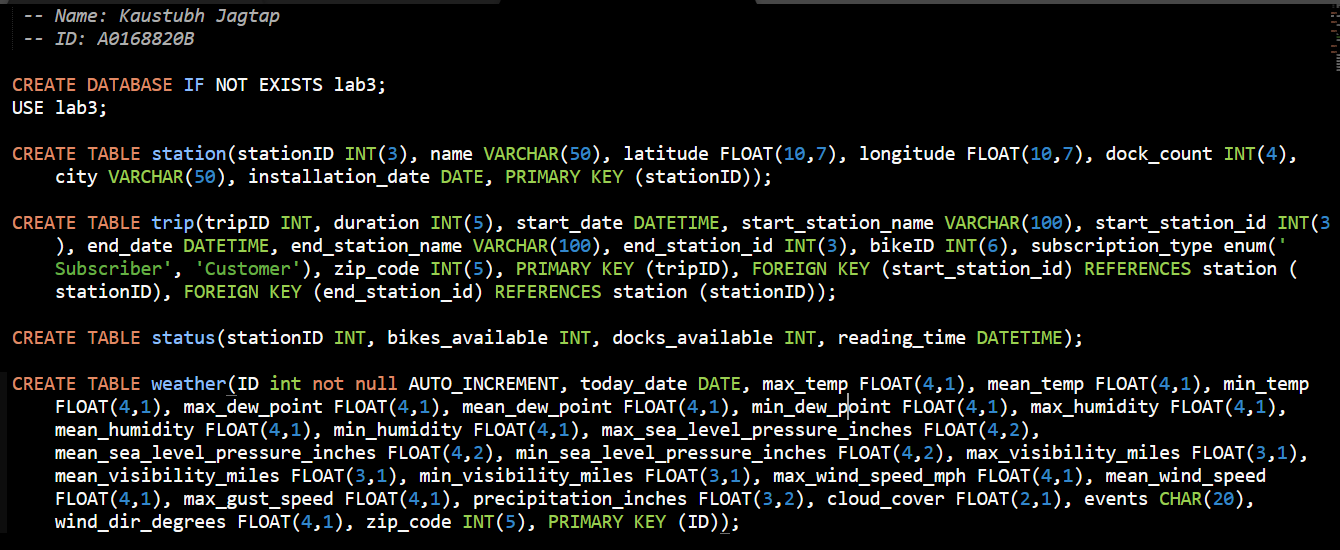


**Data processing – weather**

For the weather data, it was necessary to convert trace amount data (represented by ‘T’) to a numeric form; hence I converted those entries into 0.00. Missing columns were also replaced with 0.00.



**Schema**



**Load Data**

