ADD JAR /opt/cloudera/parcels/CDH/lib/hive/lib/hive-hcatalog-core-1.1.0-cdh5.11.2.jar;

set hive.resultset.use.unique.column.names=false

drop table nyc\_taxi\_ankit1441\_2;

create external table if not exists nyc\_taxi\_ankit1441\_2(vendorID int ,

tpep\_pickup\_datetime string,

tpep\_dropoff\_datetime string,

Passenger\_count int,

Trip\_distance double,

RateCodeID int,

Store\_and\_fwd\_flag string ,

PULocationID int,

DOLocationID int,

Payment\_type int,

Fare\_amount double,

Extra double,

MTA\_tax double,

Tip\_amount double,

Tolls\_amount double,

Improvement\_surcharge double,

Total\_amount double ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

STORED AS TEXTFILE

location '/common\_folder/nyc\_taxi\_data/'

tblproperties ("skip.header.line.count"="1");

select \* from nyc\_taxi\_ankit1441\_2;

-- Total number of rows in the dataset

select count(\*) from nyc\_taxi\_ankit1441\_2;

-- Total 1174569 rows

-- Basic Data Quality Check.

-- 1.)

select count(\*) as NumberOfRecordsProvided, vendorID from nyc\_taxi\_ankit1441\_2 group by vendorID;

-- Result Obtained :

-- Vendor 2: 647183, Vendor 1: 527386

-- Basic data quality check 2:

-- Data does not fall in the month of November or December of 2017.

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime,1,7) not in ('2017-11','2017-12');

-- Number of results:14

-- Any null values for the column vendorID:

select count(\*) from nyc\_taxi\_ankit1441\_2 where vendorID=null;

-- Zero rows.

-- Taking a look at the passenger count

select Passenger\_count, count(\*) from nyc\_taxi\_ankit1441\_2

where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') group by Passenger\_count;

select count(\*),Passenger\_count from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime,1,7) in ('2017-11','2017-12') group by Passenger\_count;

select count(Passenger\_count),vendorID from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime,1,7) in ('2017-11','2017-12') and Passenger\_count=0 group by vendorID;

-- Vendor 1 has 6813 records where the passenger count is zero while for vendor 2 the result is 11 records.

-- Checking for trip distance

select count(Trip\_distance), vendorID from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime,1,7) in ('2017-11','2017-12') and Trip\_distance<='0' group by vendorID;

-- Vendor 2 has 3184 records where the trip distance is 0 or negative while for vendor 1 it is 4217

-- Checking for null values

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and tpep\_pickup\_datetime = null;

-- No NULL values

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and tpep\_dropoff\_datetime = null;

-- no null values for dropoff time

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Passenger\_count = null;

-- no null values

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and vendorID = null;

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Trip\_distance = null;

-- no null values

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and PULocationID = null;

-- no null values

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and DOLocationID = null;

-- no null values

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and RateCodeID = null;

-- no null values

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Store\_and\_fwd\_flag = null;

-- no null values

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Payment\_type = null;

-- no null values

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Fare\_amount = null;

-- no null values

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Extra = null;

-- no null values

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and MTA\_tax = null;

-- no null values

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Improvement\_surcharge = null;

-- no null values

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Tip\_amount = null;

-- no null values

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Tolls\_amount = null;

-- no null values

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Total\_amount = null;

-- no null values

-- Checking RateCode for the months of Nov and Dec of 2017

select RateCodeID,count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') group by RateCodeID;

-- we can see that there is RateCodeID as 99 and is present in 9 0f the records. Now let us analyse as to which vendor is making mistakes for RateCodeID

select count(vendorID), vendorID from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and RateCodeID='99' group by vendorID;

-- Out of the nine recors, 8 were entered by VendorID=1 and only 1 by VendorID=2

-- Checking for Store\_and\_fwd\_flag

select Store\_and\_fwd\_flag, count(vendorID) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Store\_and\_fwd\_flag not in ('Y','N') group by Store\_and\_fwd\_flag;

-- Zero records other than Y/N

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Payment\_type not in (1,2,3,4,5,6);

-- Zero records other than the allowed values

-- checking if fare amount is less than 0

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Fare\_amount<=0;

-- in 870 rows the fare amount is less than or equal to zero

select count(vendorID), vendorID from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Fare\_amount<=0 group by vendorID;

-- From vendor id 2 number of negative or zero fare amount records is 639 while it is 231 for vendor ID 1

-- Taking the same approach for Extra(only negative)

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Extra<0;

--286 records

select count(vendorID), vendorID from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Extra<0 group by vendorID;

-- 285 for vendor 2 and 1 for vendor 1

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and MTA\_tax not in (0.5,0);

--548 rows

select count(vendorID), vendorID from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and MTA\_tax not in(0.5,0) group by vendorID;

-- 547 for vendor 2 and 1 for vendor 1

-- checking for Improvement\_surcharge

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Improvement\_surcharge!=0.30;

--849 rows

select count(vendorID), vendorID from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Improvement\_surcharge!=0.3 group by vendorID;

-- 788 for vendor 2 and 61 for vendor 1.

-- checking for tip\_amount

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Tip\_amount<0;

-- 4 rows

select count(vendorID), vendorID from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Tip\_amount<0 group by vendorID;

-- All by vendor 2

-- checking for tolls\_amount

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and tolls\_amount<0;

--3 rows

select count(vendorID), vendorID from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and tolls\_amount<0 group by vendorID;

-- All by vendor 2

-- checking for total\_amount

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Total\_amount<=0;

--681 rows

select count(vendorID), vendorID from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and Total\_amount<=0 group by vendorID;

--42 by vendor 1 and 639 by vendor 2

-- checking timestamp of pickup and drop

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and (UNIX\_TIMESTAMP(tpep\_pickup\_datetime)=UNIX\_TIMESTAMP(tpep\_dropoff\_datetime));

-- there are 6481 records where pickup time is equal to dropoff time

select count(\*) from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and (UNIX\_TIMESTAMP(tpep\_pickup\_datetime)>UNIX\_TIMESTAMP(tpep\_dropoff\_datetime));

-- 73 records where pickup time is greater than dropoff time

select count(vendorID), vendorID from nyc\_taxi\_ankit1441\_2 where substr(tpep\_pickup\_datetime, 1, 7) in('2017-11','2017-12') and (UNIX\_TIMESTAMP(tpep\_pickup\_datetime)>=UNIX\_TIMESTAMP(tpep\_dropoff\_datetime)) group by vendorID;

-- 3062 such records given by vendor 2 and 3492 by vendor 1

-- On analysing different columns, it seems that Vendor 2 is doing a bad job in providing the correct data

-- Assumptions made during analysis

-- 1. Passenger\_count can never be equal to zero. No cap has been made on the upper limit of the passenger\_count as nothing has been said in the data dictionary

-- 2. Trip\_distance can never be 0 or less

-- 3. RateCodeID, Store\_and\_fwd\_flag, Payment\_type cannot fall out pof the given values in the dictionary

-- 4. Fare\_amount cannot be less than or equal to 0

-- 5. Extra cannot be negative

-- 6. MTA\_tax can be 0 or 0.5 only

-- 7. Improvent\_surcharge cannot be anything other than 0.3

-- 8. Tip\_amount, Tolls\_amount cannot be negative

-- 9. Total\_amount cannot be 0 or less

-- 10. Pickup time cannot be greater than or equal to dropoff time

SET hive.exec.max.dynamic.partitions=100000;

SET hive.exec.max.dynamic.partitions.pernode=100000;

drop table nyc\_taxi\_partition\_orc\_ankit1441\_2;

create external table if not exists nyc\_taxi\_partition\_orc\_ankit1441\_2 (vendorID int ,

tpep\_pickup\_datetime string,

tpep\_dropoff\_datetime string,

Passenger\_count int,

Trip\_distance double,

RateCodeID int,

Store\_and\_fwd\_flag string ,

PULocationID int,

DOLocationID int,

Payment\_type int,

Fare\_amount double,

Extra double,

MTA\_tax double,

Tip\_amount double,

Tolls\_amount double,

Improvement\_surcharge double,

Total\_amount double )partitioned by (yr int, mnth int)

stored as orc location '/user/hive/warehouse/ankit1441\_2\_orc'

tblproperties ("orc.compress"="SNAPPY");

insert overwrite table nyc\_taxi\_partition\_orc\_ankit1441\_2 partition(yr, mnth)

select vendorID, tpep\_pickup\_datetime, tpep\_dropoff\_datetime, Passenger\_count, Trip\_distance,

RateCodeId, Store\_and\_fwd\_flag, PULocationID, DOLocationID, Payment\_type, Fare\_amount, Extra, MTA\_tax,

Tip\_amount, Tolls\_amount, Improvement\_surcharge, Total\_amount,

year(tpep\_pickup\_datetime) as yr, month(tpep\_pickup\_datetime) as mnth

from nyc\_taxi\_ankit1441\_2

where Passenger\_count not in(0) and Trip\_distance > 0

and (UNIX\_TIMESTAMP(tpep\_pickup\_datetime)<UNIX\_TIMESTAMP(tpep\_dropoff\_datetime))

and RateCodeID != 99 and Fare\_amount > 0 and Extra >= 0

and MTA\_tax in (0,0.5) and Tip\_amount >= 0

and Tolls\_amount >= 0 and Improvement\_surcharge = 0.30

and Total\_amount > 0 and year(tpep\_pickup\_datetime) = 2017 and month(tpep\_pickup\_datetime) in (11,12);

select count(\*) from nyc\_taxi\_partition\_orc\_ankit1441\_2;

-- total 1157886 rows as compared to 1174569 rows originally

select \* from nyc\_taxi\_partition\_orc\_ankit1441\_2;

-- There won't be any null values because we have already checked it earlier in the original dataset

-- Analysis Check 1

--Compare the overall average fare per trip for November and December.

select avg(Fare\_amount) as Average\_amount, mnth as Month from nyc\_taxi\_partition\_orc\_ankit1441\_2 group by mnth;

--Rounding it off

select round(avg(Fare\_amount),2) as Average\_amount, mnth as Month from nyc\_taxi\_partition\_orc\_ankit1441\_2 group by mnth;

-- 13.05 for month of November and 12.85 for month of December

-- Check 2: Explore the â€˜number of passengers per tripâ€™ - how many trips are made by each level of â€˜Passenger\_countâ€™? Do most people travel solo or with other people?

select count(\*) as `Number Of Trips`, Passenger\_count from nyc\_taxi\_partition\_orc\_ankit1441\_2 group by Passenger\_count order by `Number Of Trips` desc;

-- 8,19798 trips are there where the people have travelled solo and this forms the most of the total number of trips

-- 3. Which is the most preferred mode of payment?

select count(\*) as `Number of payments`, Payment\_type from nyc\_taxi\_partition\_orc\_ankit1441\_2 group by Payment\_type order by `Number of Payments` desc;

-- Payment type 1 is the most preffered:782119 transactions

--4 What is the average tip paid per trip?

--Compare the average tip with the 25th, 50th and 75th percentiles and comment whether the â€˜average tipâ€™ is a representative statistic (of the central tendency) of â€˜tip amount paidâ€™.

select round(avg(Tip\_amount),2) as `Average Tip Paid` from nyc\_taxi\_partition\_orc\_ankit1441\_2;

select round(avg(Tip\_amount),2) as `Average Tip Paid`,

percentile\_approx(Tip\_amount,0.25) as `25th Percentil`,

percentile\_approx(Tip\_amount,0.50) as `50th Percentil`,

percentile\_approx(Tip\_amount,0.75) as `75th Percentil`

from nyc\_taxi\_partition\_orc\_ankit1441\_2;

-- Average : 1.85, 25th Percentile 0, 50th Percentile : 1.36, 75th percentile : 2.45

--5 Explore the â€˜Extraâ€™ (charge) variable - what fraction of total trips have an extra charge is levied?

select count(\*) from nyc\_taxi\_partition\_orc\_ankit1441\_2 where Extra>0;

-- 537343 trips where extra charge has been levied

select sub.extra\_charge\_group, sub.count/sum(sub.count) over () from (

select case

when extra = 0 then 'Extra Charge not Applied'

else 'Extra charge Applied' end as extra\_charge\_group,

count(1) as count

from nyc\_taxi\_partition\_orc\_ankit1441\_2

group by

case

when extra = 0 then 'Extra Charge not Applied'

else 'Extra charge Applied' end) sub;

-- Extra charge applied : 46.3%

-- Extra charge not applied : 0.54

-- Anlysis Part II

--1. What is the correlation between the number of passengers on any given trip, and the tip paid per trip? Do multiple travellers tip more compared to solo travellers?

select corr(Passenger\_count, Tip\_amount) as Correl from nyc\_taxi\_partition\_orc\_ankit1441\_2;

-- Correlation b/w Passenger\_count and Tip\_amount is -0.00426 which means that the correlation is not strong and the tip generally does not depend on the number of passengers

--2. Segregate the data into five segments of â€˜tip paidâ€™: [0-5), [5-10), [10-15) , [15-20) and >=20. Calculate the percentage share of each bucket (i.e. the fraction of trips falling in each bucket).

select (sum(if(tip\_amount >=0 and tip\_amount < 5, 1,0))/count(\*))\*100 as `[0-5)`,

(sum(if(tip\_amount >=5 and tip\_amount < 10, 1,0))/count(\*))\*100 as `[5-10)`,

(sum(if(tip\_amount >=10 and tip\_amount < 15, 1,0))/count(\*))\*100 as `[10-15)`,

(sum(if(tip\_amount >=15 and tip\_amount < 20, 1,0))/count(\*))\*100 as `[15-20)`,

(sum(if(tip\_amount >=20, 1,0))/count(\*))\*100 as `>=20`

from nyc\_taxi\_partition\_orc\_ankit1441\_2

-- 0-5: 92.19

-- 5-10 : 5.66

-- 10-15 : 1.83

-- 15-20: 0.22

-- >=20 0.09

--3. Calculating the speed in meter per sec using tpep\_dropoff\_datetime, tpep\_pickup\_datetime & trip\_distance column and comparing average speed for Novermber and Decemeber

select mnth, round(avg(Trip\_distance/(unix\_timestamp(tpep\_dropoff\_datetime)-unix\_timestamp(tpep\_pickup\_datetime)))\*1609.344,2) as Speed from nyc\_taxi\_partition\_orc\_ankit1441\_2 group by mnth;

-- 4.92 for November and 4.96 for December

--4 Analyse the average speed of the most happening days of the year, i.e. 31st December (New yearâ€™s eve) and 25th December (Christmas) and compare it with the overall average.

--Christmas Average Speed

select round(avg(Trip\_distance/(unix\_timestamp(tpep\_dropoff\_datetime)-unix\_timestamp(tpep\_pickup\_datetime)))\*1609.344,2) as Speed\_Christmas from nyc\_taxi\_partition\_orc\_ankit1441\_2 where day(tpep\_pickup\_datetime)=25 and month(tpep\_pickup\_datetime) = 12;

-- New Year avg speed--6.82

select round(avg(Trip\_distance/(unix\_timestamp(tpep\_dropoff\_datetime)-unix\_timestamp(tpep\_pickup\_datetime)))\*1609.344,2) as Speed\_New\_Year from nyc\_taxi\_partition\_orc\_ankit1441\_2 where day(tpep\_pickup\_datetime)=31 and month(tpep\_pickup\_datetime) = 12;

--5.92

--Overall avg

select round(avg(Trip\_distance/(unix\_timestamp(tpep\_dropoff\_datetime)-unix\_timestamp(tpep\_pickup\_datetime)))\*1609.344,2) as Speed\_New\_Year from nyc\_taxi\_partition\_orc\_ankit1441\_2;

-- 4.94 which is less as compared to New Year and Christmas.