NYC_Taxi_Data_Analysis

DATA LINK :

November : https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2017-11.csv

December : https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2017-12.csv

Creating Table November:

```
create external table if not exists yellow_trip_data_nov_s3 (VendorID tinyint,tpep_pickup_datetime string,tpep_dropoff_datetime string,passenger_count tinyint,trip_distance float,RatecodeID tinyint,store_and_fwd_flag char(1),PULocationID smallint,DOLocationID smallint,payment_type float,fare_amount float,extra float,mta_tax float,tip_amount float,tolls_amount float,improvement_surcharge float,total_amount float)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

tblproperties ("skip.header.line.count"="2");
```

Creating Table December :

```
create external table if not exists yellow_trip_data_dec_s3 (VendorID tinyint,tpep_pickup_datetime string,tpep_dropoff_datetime string,passenger_count tinyint,trip_distance float,RatecodeID tinyint,store_and_fwd_flag char(1),PULocationID smallint,DOLocationID smallint,payment_type float,fare_amount float,extra float,mta_tax float,tip_amount float,tolls_amount float,improvement_surcharge float,total_amount float)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

tblproperties ("skip.header.line.count"="2");
```

Combining Tables:

```
create table yellow_trip_data_s3 as(select * from yellow_trip_data_nov_s3

UNION ALL

select  * from yellow_trip_data_dec_s3);
```

Displaying Tables:

```
select * from yellow_trip_data_s3 limit 100;
```

Basic Data Quality Checks::

Summarises the number of records of each provider.

```
select vendorid,count(*) as counts from yellow_trip_data_s3 group by vendorid;
```

checking data and cleaning .

```
select distinct month(tpep_pickup_datetime) as month from yellow_trip_data_s3;

select distinct year(tpep_pickup_datetime) as year from yellow_trip_data_s3;

select distinct year(tpep_pickup_datetime) as year from yellow_trip_data_s3 where
month(tpep_pickup_datetime) in (11,12);

select distinct year(tpep_dropoff_datetime) as year from yellow_trip_data_s3 where
month(tpep_pickup_datetime) in (11,12) and year(tpep_pickup_datetime)=2017;

select distinct month(tpep_dropoff_datetime) as month from yellow_trip_data_s3 where
month(tpep_pickup_datetime) in (11,12) and year(tpep_pickup_datetime)=2017 and
year(tpep_dropoff_datetime) in (2017,2018);

create table yellow_trip_data_final_s3 as (select * from yellow_trip_data_s3 where
month(tpep_pickup_datetime) in (11,12) and year(tpep_pickup_datetime)=2017 and
year(tpep_dropoff_datetime) in (2017,2018));
```

The average fare for November and December.

```
select month(tpep_pickup_datetime) as month, round(avg(fare_amount),2) as average_fare from
yellow_trip_data_final_s3 where fare_amount>0 group by month(tpep_pickup_datetime);
```

Number of passengers per trip

```
select passenger_count, count(*) as counts from yellow_trip_data_final_s3 where passenger_count!=0

group by passenger_count order by passenger_count;
```

The most preferred mode of payment

```
select payment_type, count(*) as counts from yellow_trip_data_final_s3 group by payment_type

order by payment_type;
```

The average tip paid

```
select round(avg(tip_amount),2) as average, round(percentile_approx(tip_amount, 0.25),2) as
25th_percentile, round(percentile_approx(tip_amount ,0.50),2) as 50th_percentile,
round(percentile_approx(tip_amount ,0.75),2) as 75th_percentile from yellow_trip_data_final_s3

where tip_amount>=0;
```

Explore the 'Extra' (charge) variable

```
SELECT round(SUM( IF( extra == 0.0, 0 , 1 ) )/ COUNT(*) * 100,2) as
total_trips_percentage_with_extra_charge FROM yellow_trip_data_final_s3;
```

Which month has a greater average 'speed'

```
select month(tpep_pickup_datetime) as month,
round(avg(trip_distance/((unix_timestamp(tpep_dropoff_datetime) -
unix_timestamp(tpep_pickup_datetime))/3600)),2) as average_speed from yellow_trip_data_final_s3
where trip_distance!=0  group by month(tpep_pickup_datetime);
```

Tip paid classeswise;

```
create table tip_amount_bucket_counts as (select tip_amount,

        case

        when (tip_amount >= 0  and tip_amount < 5)  then '[0-5)'

        when (tip_amount >= 5  and tip_amount < 10) then '[5-10)'

        when (tip_amount >= 10 and tip_amount < 15) then '[10-15)'

        when (tip_amount >= 15 and tip_amount < 20) then '[15-20)'

        when (tip_amount >= 20) THEN '>=20'

        end as tip_amount_bucket

          from yellow_trip_data_final_s3

          where tip_amount>=0);
```

multiple travellers pay more compared to solo travellers?

```
        select passenger_count,round(avg(tip_amount),2) as average_tip_amount

        from yellow_trip_data_final_s3

        where passenger_count!=0

        group by passenger_count

        order by passenger_count;
```

Overall speed average

```
select round(avg(trip_distance/((unix_timestamp(tpep_dropoff_datetime) -
unix_timestamp(tpep_pickup_datetime))/3600)),2) as average_speed from yellow_trip_data_final_s3
where trip_distance!=0;
```