**Analysis of NYC Trip Data**

In this project, we will analyze a dataset that contains information about taxi rides in NYC. The data set is quite large so getting a basic idea of what the data contains is important. Each student should use one of the CSV files. Answer the following questions:

1. What datetime range does your data cover? How many rows are there total?

There are **13823840** rows

1. What are the field names? Give descriptions for each field.

* Medallion: This is a text field which stores an identification to an object such as a medal.
* hack\_license: This is similar to the medallion and also stores an identification to the license.
* vendor\_id: This is an identification for the vendor.
* rate\_code: This is a field which would most likely stores the link to a record which shows how the trip should be charged.
* store\_and\_fwd\_flag: A field which can be used to check for conditions.
* pickup\_datetime: This the time when the taxi rider was picked up.
* dropoff\_datetime: The time when the taxi rider was dropped off at their destination.
* passenger\_count. This is an integer that shows the number of passengers in the taxi.
* trip\_time\_in\_secs: The time the entire trip took from the pickup location to the destination.
* trip\_distance: The distance from the pickup location to the destination.
* pickup\_longitude: This is a decimal number that shows the longitude where the rider was picked up at.
* pickup\_latitude: A decimal number that shows the latitude where the rider was picked up at.
* dropoff\_longitude: A decimal number showing the longitude location where the location was used at.
* dropoff\_latitude: This is a decimal number that shows the latitude where the rider was dropped off at.

The longitude and latitudes are used together to show the exact location of a user on a map.

1. Give some sample data for each field.

|  |  |
| --- | --- |
| **Field** | **Data Sample** |
| Medallion | 9406D2C34715E1DA10AD4D4DDADF4DA5 |
| hack\_license | 0602DFD837433635FE860BDE2F14BC3A |
| vendor\_id | VTS |
| rate\_code | 1 |
| store\_and\_fwd\_flag | N |
| pickup\_datetime | 2013-07-01 01:47:00 |
| dropoff\_datetime | 2013-07-01 01:52:00 |
| passenger\_count | 1 |
| trip\_time\_in\_secs | 300 |
| trip\_distance | 1.17 |
| pickup\_longitude | -74.013229 |
| pickup\_latitude | 40.714718 |
| dropoff\_longitude | -74.003906 |
| dropoff\_latitude | 40.742031 |

4. What MySQL data types/len would you need to store each of the fields?

|  |  |  |
| --- | --- | --- |
| **Field** | **Data Type** | **Length** |
| Medallion | Varchar | 100 |
| hack\_license | VarChar | 100 |
| vendor\_id | Varchar | 3 |
| rate\_code | int | 1 |
| store\_and\_fwd\_flag | Varchar | 1 |
| pickup\_datetime | Datetime | Not applicable |
| dropoff\_datetime | Datetime | Not applicable |
| passenger\_count | int | 2 |
| trip\_time\_in\_secs | int | 10 |
| trip\_distance | Decimal | 3, 2; The significant digits is 3 and the number of decimal places is 2 |
| pickup\_longitude | Decimal | 2,6: The number of significant digits is 2 and the decimal places is 6 |
| pickup\_latitude | Decimal | 2, 6: The number of significant digits is 2 and the decimal places is 6 |
| dropoff\_longitude | Decimal | 2, 6: The number of significant digits is 2 and the decimal places is 6 |
| dropoff\_latitude | Decimal | 2, 6: The number of significant digits is 2 and the decimal places is 6 |

5. What is the geographic range of your data (min/max - X/Y)?

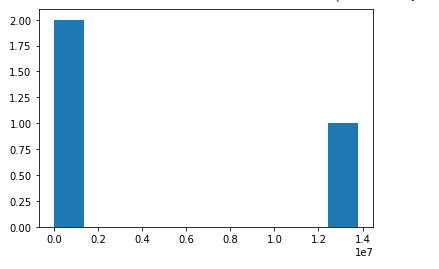
a. Plot this (approximately on a map)



6. What is the average computed trip distance? (You should use Haversine Distance)

The average trip distance is **12.743926394710524 KM**

a. Draw a histogram of the trip distances binned anyway you see fit.



7. What are the distinct values for each field? (If applicable)

|  |  |  |
| --- | --- | --- |
| Medallion | 13434 |  |
| Hack license' | 32788 |  |
| Vendor ID | 2 | 'VTS', CMT |
| rate\_code | 11 | 1, 5, 3, 2, 4, 8, 0, 6, 210, 7, 9 |
| store\_and\_fwd\_flag | 8 | ‘’, N, Y |
| pickup\_datetime | 2293629 |  |
| passenger\_count' | 8 | 1, 5, 4, 6, 2, 3, 0, 8 |
| trip\_time\_in\_secs | 7054 |  |
| dropoff\_datetime | 2296818 |  |
| trip\_distance | 4252 |  |
| pickup\_longitude | 37885 |  |
| pickup\_latitude | 62924 |  |
| dropoff\_longitude | 55140 |  |
| dropoff\_latitude | 88407 |  |

8. For other numeric types besides lat and lon, what are the min and max values?

**Max Values**

rate\_code 210

passenger\_count 8

trip\_time\_in\_secs 10800

trip\_distance 100.0

**Min Values**

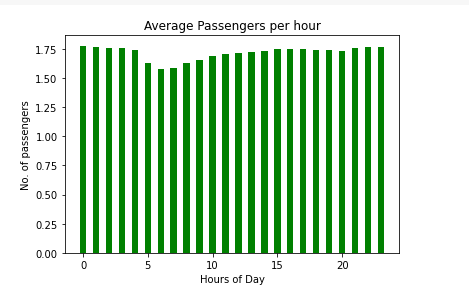
rate\_code 0

passenger\_count 0

trip\_time\_in\_secs 0

trip\_distance 0.0

9. Create a chart which shows the average number of passengers each hour of the day. (X axis should have 24 hours)



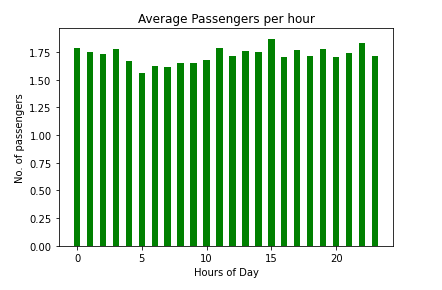
10. Create a new CSV file which has only one out of every thousand rows.

The code below illustrates the creation of the new CSV by picking every 1000th row.

*df3 = df.iloc[::1000, :]*

*df3.to\_csv('/content/drive/MyDrive/nyc\_trip\_reduced\_data.csv')*

11. Repeat step 9 with the reduced dataset and compare the two charts.



The two chart are very similar in that they show that the number of trips increase as at around 1900 hrs when most people are heading from work. Then there is a drop at 5 am when most people are asleep followed by s stead rise in the number of trips as the day progresses.