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

Datetime Range: From 2013-01-01 00:00:00 to 2013-01-31 23:59:59

Total Rows: 14776615

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

Field Names:

Index(['medallion', 'hack\_license', 'vendor\_id', 'rate\_code',

'store\_and\_fwd\_flag', 'pickup\_datetime', 'dropoff\_datetime',

'passenger\_count', 'trip\_time\_in\_secs', 'trip\_distance',

'pickup\_longitude', 'pickup\_latitude', 'dropoff\_longitude',

'dropoff\_latitude', 'hour'],

dtype='object')

Field Descriptions:

medallion : A unique identifier for the taxi

hack\_license : A unique identifier for the taxi driver

pickup\_datetime : Date and time of pickup

dropoff\_datetime : Date and time of dropoff

Give some sample data for each field.

Sample Data:

medallion hack\_license \

0 89D227B655E5C82AECF13C3F540D4CF4 BA96DE419E711691B9445D6A6307C170

1 0BD7C8F5BA12B88E0B67BED28BEA73D8 9FD8F69F0804BDB5549F40E9DA1BE472

2 0BD7C8F5BA12B88E0B67BED28BEA73D8 9FD8F69F0804BDB5549F40E9DA1BE472

3 DFD2202EE08F7A8DC9A57B02ACB81FE2 51EE87E3205C985EF8431D850C786310

4 DFD2202EE08F7A8DC9A57B02ACB81FE2 51EE87E3205C985EF8431D850C786310

vendor\_id rate\_code store\_and\_fwd\_flag pickup\_datetime \

0 CMT 1 N 2013-01-01 15:11:48

1 CMT 1 N 2013-01-06 00:18:35

2 CMT 1 N 2013-01-05 18:49:41

3 CMT 1 N 2013-01-07 23:54:15

4 CMT 1 N 2013-01-07 23:25:03

dropoff\_datetime passenger\_count trip\_time\_in\_secs trip\_distance \

0 2013-01-01 15:18:10 4 382 1.0

1 2013-01-06 00:22:54 1 259 1.5

2 2013-01-05 18:54:23 1 282 1.1

3 2013-01-07 23:58:20 2 244 0.7

4 2013-01-07 23:34:24 1 560 2.1

pickup\_longitude pickup\_latitude dropoff\_longitude dropoff\_latitude \

0 -73.978165 40.757977 -73.989838 40.751171

1 -74.006683 40.731781 -73.994499 40.750660

...

1 0

2 18

3 23

4 23

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

MySQL Data Types:

medallion object

hack\_license object

vendor\_id object

rate\_code int64

store\_and\_fwd\_flag object

pickup\_datetime datetime64[ns]

dropoff\_datetime object

passenger\_count int64

trip\_time\_in\_secs int64

trip\_distance float64

pickup\_longitude float64

pickup\_latitude float64

dropoff\_longitude float64

dropoff\_latitude float64

hour int32

dtype: object

Maximum String Lengths:

medallion 32

hack\_license 32

vendor\_id 3

rate\_code 3

store\_and\_fwd\_flag 3

pickup\_datetime 19

...

dropoff\_longitude 14

dropoff\_latitude 14

hour 2

dtype: int64

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

Geographic Range:

Min Longitude: -2771.2854

Max Longitude: 112.40418

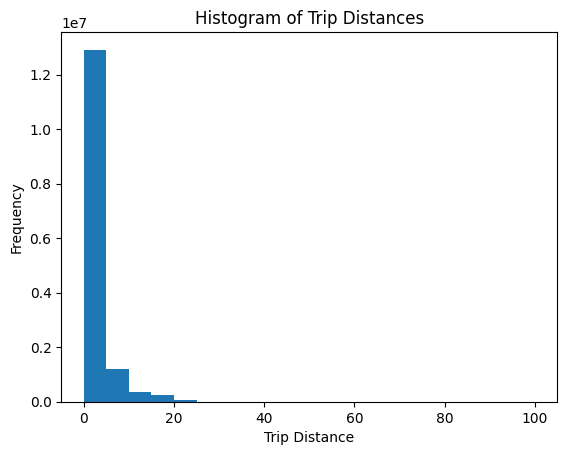
Min Latitude: -3547.9207

Max Latitude: 3310.3645

What is the average overall computed trip distance?

Average Trip Distance (Haversine): 12.178383325209287 miles

What are the distinct values for each field?



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

Minimum and Maximum Values for Numeric Fields (excluding lat and lon):

rate\_code passenger\_count trip\_time\_in\_secs trip\_distance \

min 0 0 0 0.0

max 210 255 10800 100.0

pickup\_longitude pickup\_latitude dropoff\_longitude dropoff\_latitude \

min -2771.28540 -3547.9207 -2350.9556 -3547.9207

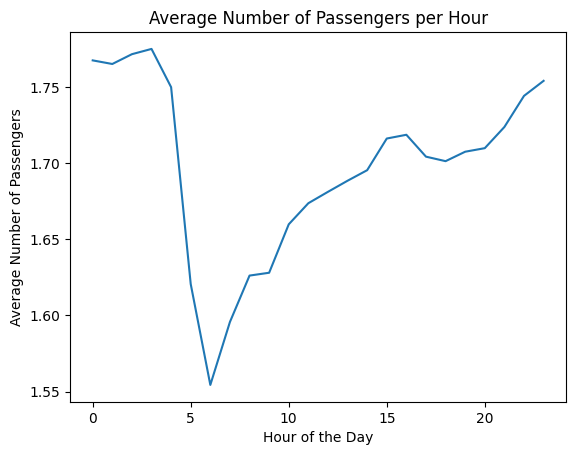
max 112.40418 3310.3645 2228.7375 3477.1055

hour

min 0

max 23

Create a chart which shows the average number of passengers each hour of the day.



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

