Exploratory Data Analysis

New York Taxi Trips Dataset (2013)

Kovid Sharma

10th March 2021

Table of Contents

Data Engineering Exercise

You are part of an analytics team engaged to support NYC taxis to identify opportunities to improve their revenue or cut costs. This exercise is based on the public NYC taxi data containing trip details and trip fares for 2013. Given the size of the data, it's recommended to select a month or week of data for analysis.

The task is to build a data model and conduct analysis for the NYC taxis leadership team to give them interesting and useful insights as well as identify opportunities to drive their priorities.
In addition, the data model should lend itself to the following use cases.

- Enable the development of visualisation dashboards showing performance over time
- Enable the development of models to predict fare and tip amounts
- We are looking for a data model that supports insights generation and predictive modelling at scale
- We are looking for structured responses and data stories communicated clearly through slides, notebooks, data visualisation or dashboards

1. Variable Identification

Predictor (Input) - pickup_datetime, passenger_count, trip_time_in_secs, trip_distance,
Target (output) - fare_amount, surcharge, total_amount

Number of DISTINCT:

1. medallion – 13415
2. hack_license – 32062
3. vendor_id – 2 ("VTS", "CMT")
4. rate_code – 12 (0,1,2,3,4,5,6,7,28,79,210,221)
5. store_and_fwd_flag ISNULL – 6952551
6. store_and_fwd_flag IS NOT NULL – 7037625
7. payment_type – 5 ("CRD","CSH","DIS","NOC","UNK")
8. passenger_count – 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 129, 208

Trips lasting less than a minute:

trip_time_in_secs < 60; -- 287

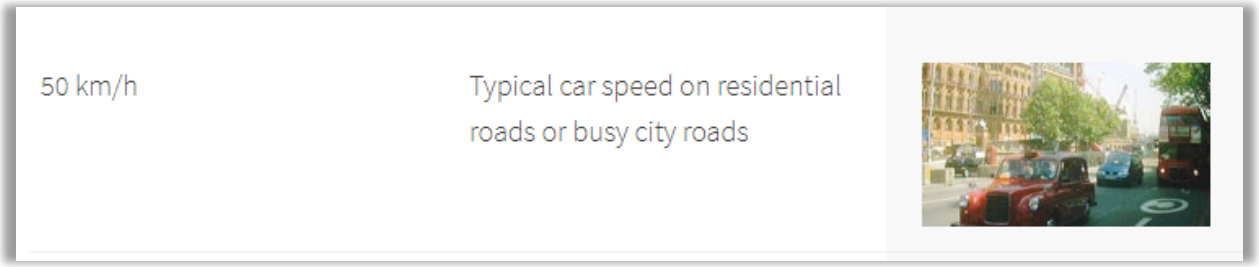
2. Deletion: List Wise Deletion

In list wise deletion, we delete observations where any of the variable is incorrect/missing.

Identifying errorenous values:

1. trip_distance, trip_time_in_secs

Assuming the average speed in NYC is 70km/h (19.44 meters/sec), any car going above this speed has some error. Deleting rows with speed exceeding 70km/h.



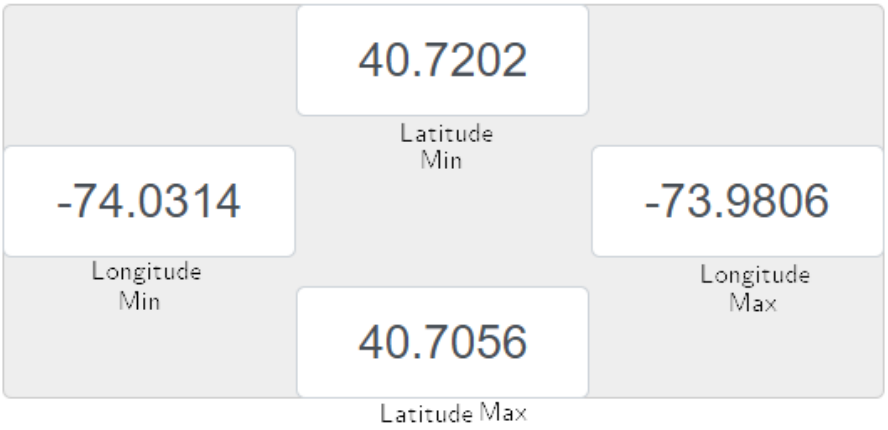

```
speed = ((trip_distance * 1.609 * 1000) / trip_time_in_secs) * 3.6 > 70
```

convert to km scale to m convert to km/h

2. To double back, any record with less than 60 secs and trip distance less than 0.5 would be deleted

3. GPS Co-ordinates Decimal Degrees

3.1 The latitude of New York City, NY, USA is 40.730610, and the longitude is -73.935242.

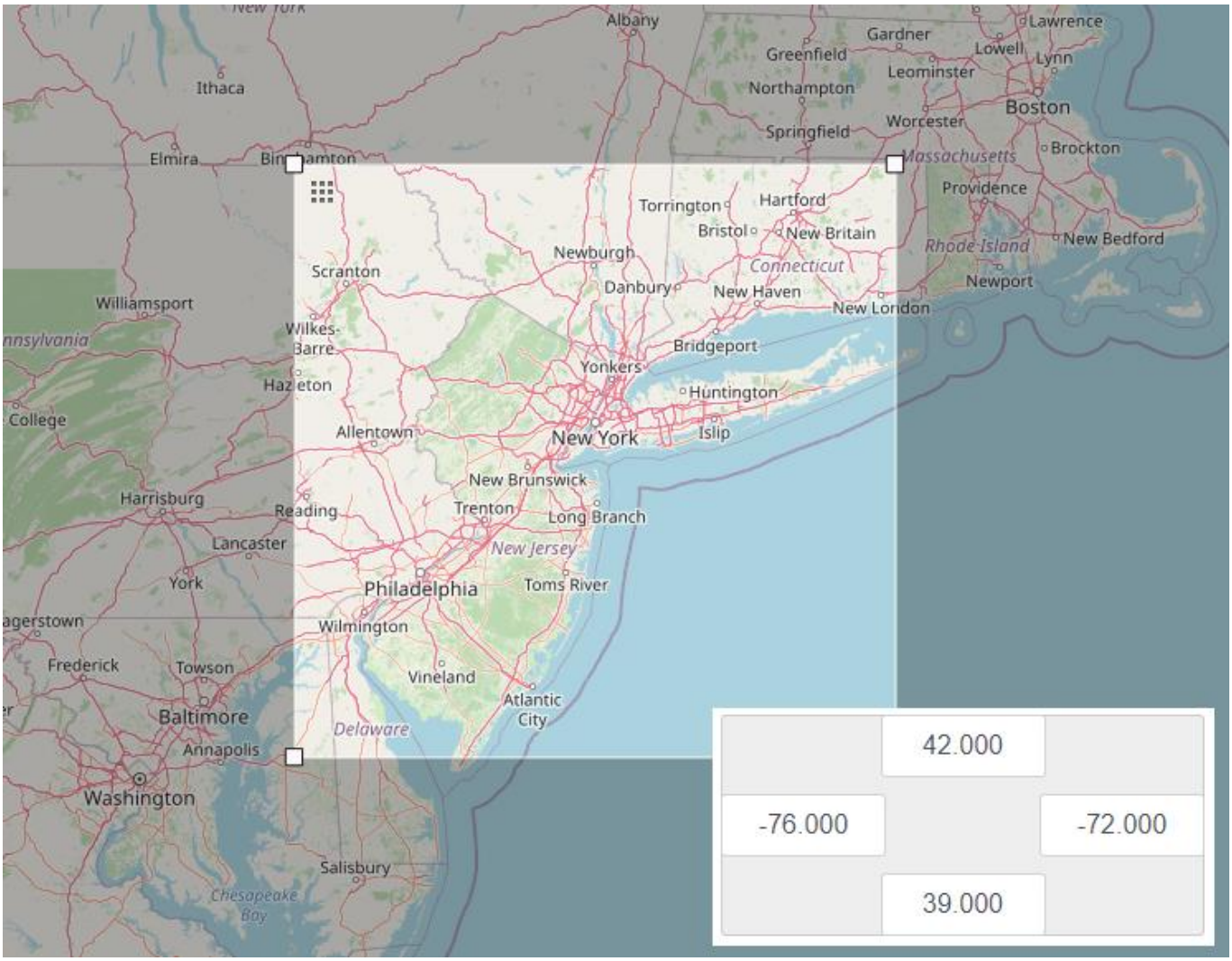


3.2 The range for latitude is -90 to 90 and -180 to 180 for longitude (Numbers are in decimal degrees format). The max and min ranges shown below are incorrect and need to be deleted.

MIN				MAX			
pickup		dropoff		pickup		dropoff	
longitude	latitude	longitude	latitude	longitude	latitude	longitude	latitude
-2001.505100	-3447.915000	-2491.213900	-3481.127700	176.543730	646.441160	2228.768300	3577.132100

3.3. Erroreneous data values:

For latitude I've chosen the range between 39 and 42; for longitude the range is set between -76 to -72. Any taxi going out of this range has an error and the entire row would be deleted.



Alter the co-ordinate Precision, Scale

NUMERIC(precision, scale)

The scale of a numeric is the count of decimal digits in the fractional part, to the right of the decimal point. So the number 23.5141 has a precision of 6 and a scale of 4. Integers can be considered to have a scale of zero.

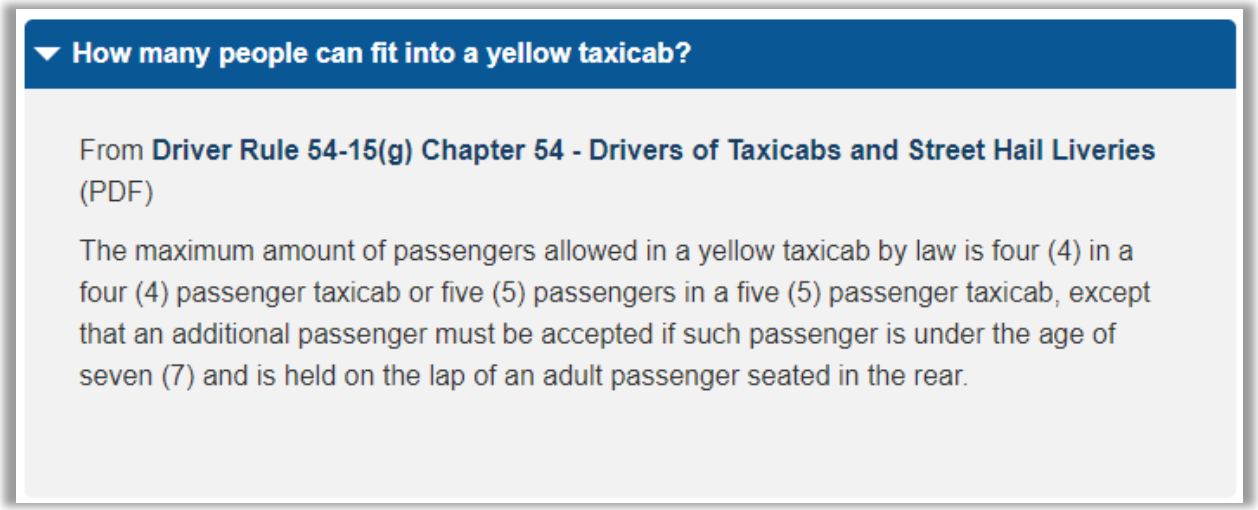
GPS floating point scale (Decimal degrees):

decimal places	decimal degrees	N/S or E/W at equator
3	0.001	111.32 m
4	0.0001	11.132 m
5	0.00001	1.1132 m
6	0.000001	111.32 mm

Switching from NUMERIC(10,6) to NUMERIC(7,4) saves roughly 112mb of disk space, without losing any valueable insight from the data.

4. passenger_count

Passenger count had some erroreneous values of 0, 129 and 208 passengers. Further reading at [4] states that yellow taxicabs are not allowed more than 6 passengers. With this reason, all records with passenger_count less than 0 and more than 6 will be deleted.



2. Make models

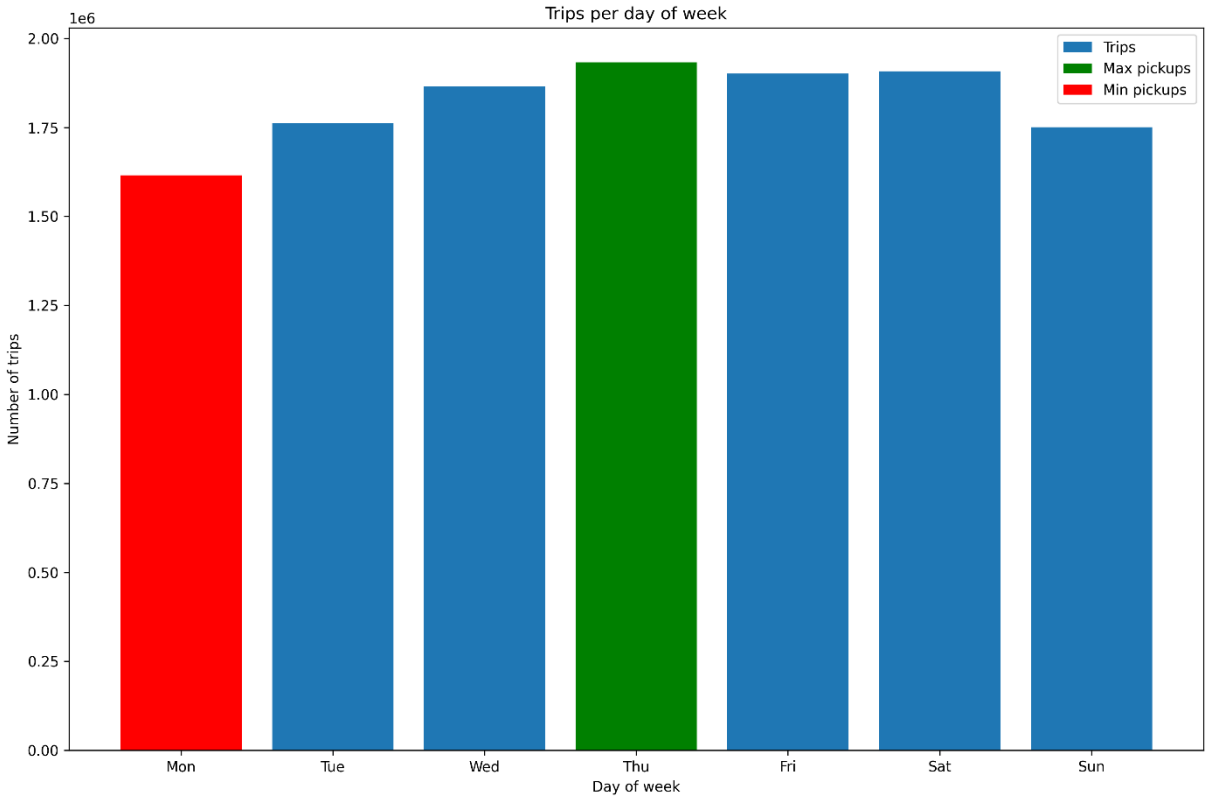
driver_analysis		
	Attribute	Datatype
1	id	INT (PK)
2	medallion	TEXT
3	hack_license	TEXT
4	vendor_id	TEXT

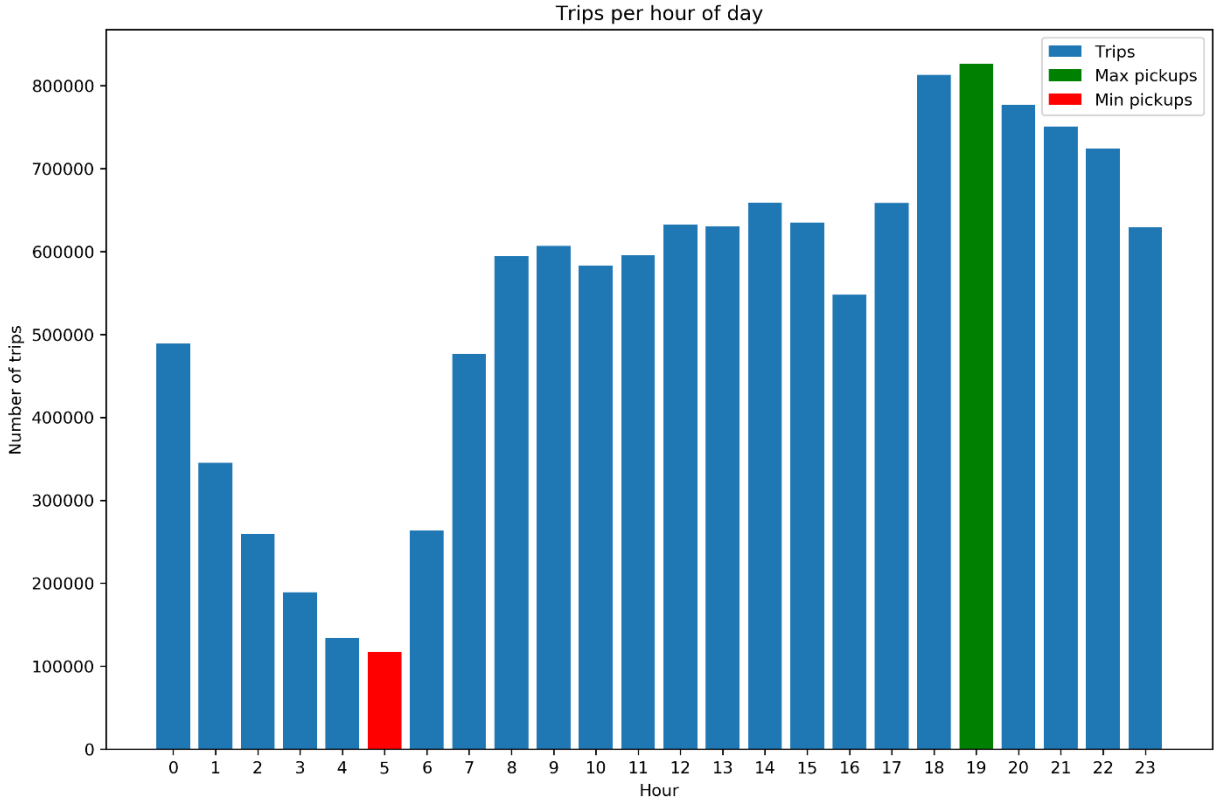
fare_analysis		
	Attribute	Datatype
1	id	INT (PK)
2	payment_type	TEXT
3	fare_amount	NUMERIC(5,2)
4	surcharge	NUMERIC(4,2)
5	mta_tax	NUMERIC(3,2)
6	tip_amount	NUMERIC(5,2)
7	tolls_amount	NUMERIC(4,2)
8	total_amount	NUMERIC(5,2)

data_analysis		
	Attribute	Datatype
1	id	INT (PK)
2	rate_code	INT
3	store_and_fwd_flag	BOOL
4	pickup_datetime	TIMESTAMP
5	dropoff_datetime	TIMESTAMP
6	passenger_count	INT
7	trip_time_in_secs	INT
8	trip_distance	NUMERIC(5,2)
9	pickup_longitude	NUMERIC(7,4)
10	pickup_latitude	NUMERIC(7,4)
11	dropoff_longitude	NUMERIC(7,4)
12	dropoff_latitude	NUMERIC(7,4)

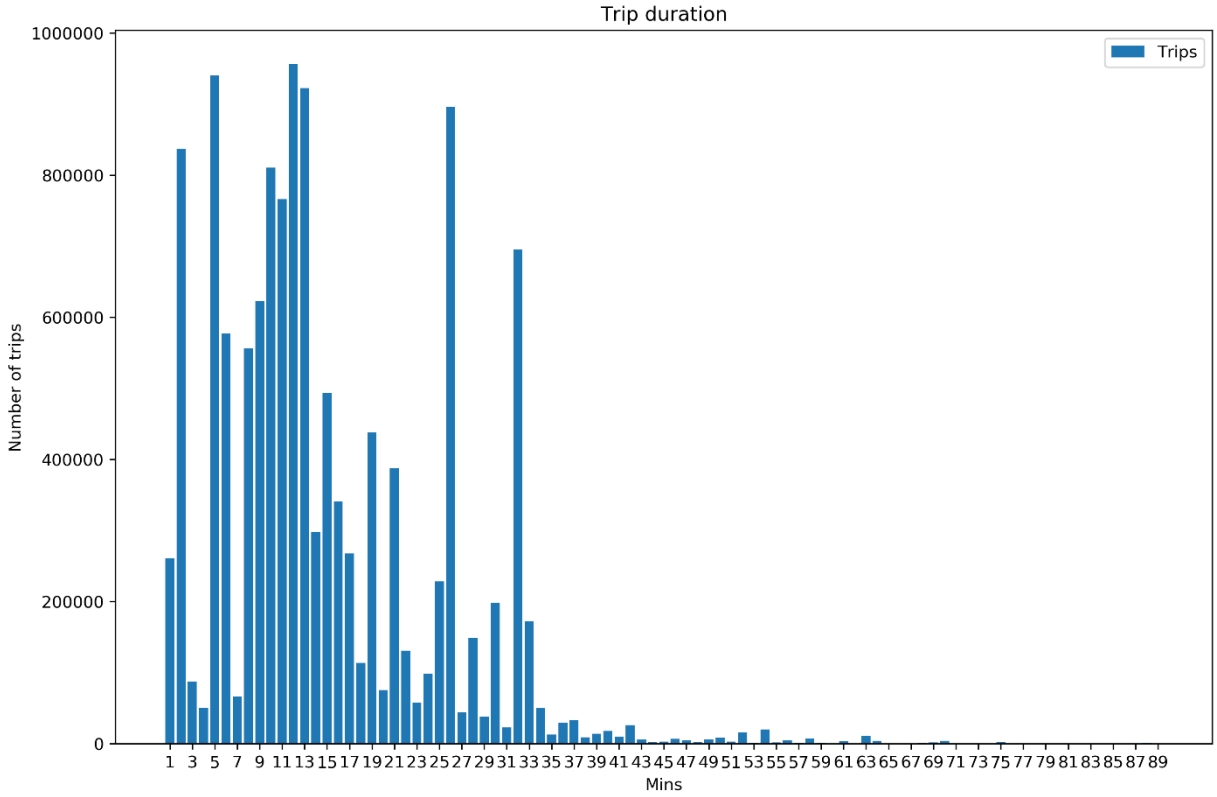
Analysis:

1. Split datetime into date and hour_of_day to get a better estimate of the weekend rush/slump and the surcharge, based on peak hours and night time.

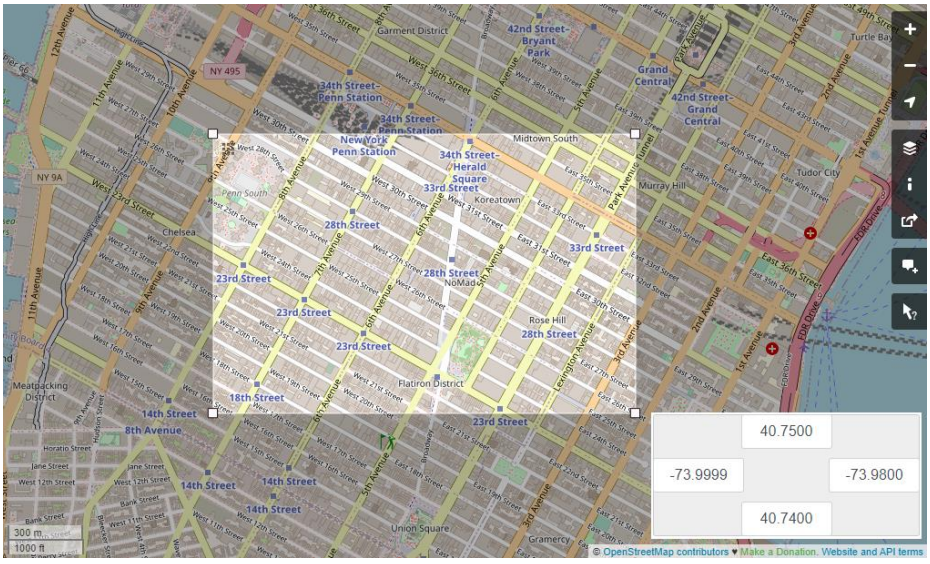




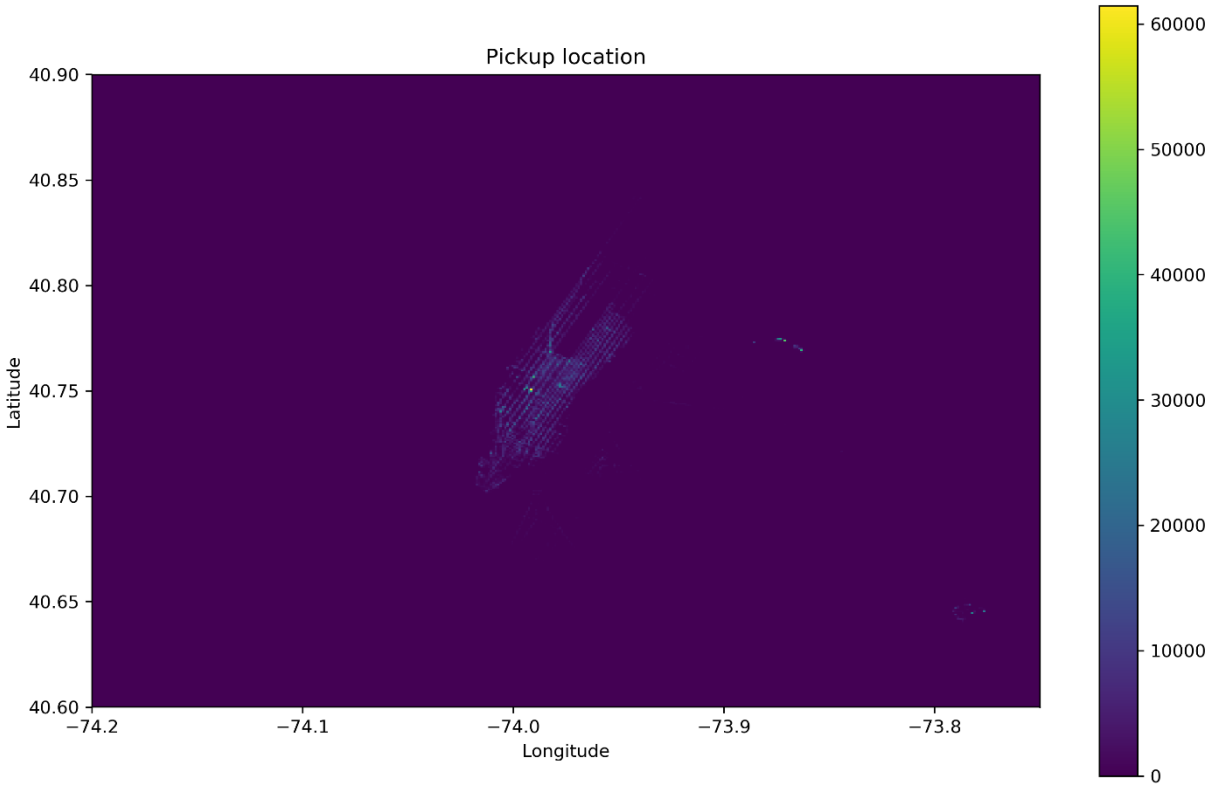
2. Convert trip_time_in_secs to trip_time_in_mins to get a better understanding of the time spent while travelling vs. the cost/distance

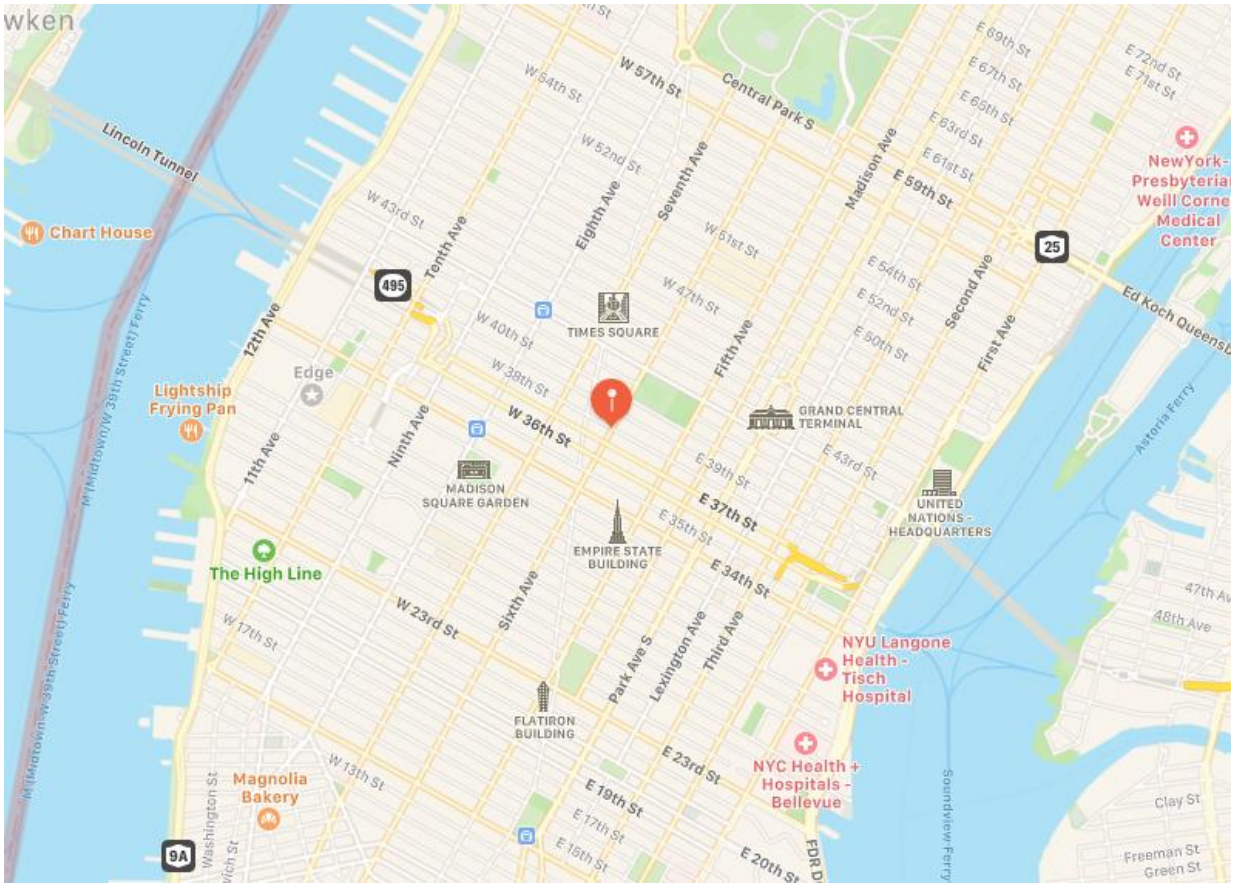


3. Get an idea of the location where most pickups and drop-off are done. If possible, plot a heatmap for the co-ordinates on the NYC map.



Max pickup region


Pickup location
Latitude
Longitude
40.90
40.85
40.80
40.75
40.70
40.65
40.60
−74.2
−74.1
−74.0
−73.9
−73.8
60000
50000
40000
30000
20000
10000
0


Lincoln Tunnel
Chart House
Lightship Frying Pan
Edge
The High Line
MADISON SQUARE GARDEN
TIMES SQUARE
EMPIRE STATE BUILDING
FLATIRON BUILDING
GRAND CENTRAL TERMINAL
UNITED NATIONS HEADQUARTERS
NYU Langone Health - Tisch Hospital
NYC Health + Hospitals - Bellevue
NewYork-Presbyterian Weill Cornell Medical Center
Magnolia Bakery
W 57th St
Central Park S
W 54th St
W 52nd St
W 43rd St
W 40th St
W 38th St
W 36th St
W 23rd St
W 17th St
W 13th St
E 59th St
E 39th St
E 37th St
E 35th St
E 34th St
E 23rd St
E 19th St
E 17th St
E 15th St
E 20th St
Tenth Ave
Eighth Ave
Seventh Ave
Madison Ave
Fifth Ave
Ninth Ave
Eleventh Ave
Twelfth Ave
Sixth Ave
Park Ave S
Lexington Ave
Third Ave
Second Ave
First Ave
Ed Koch Queensboro
Astoria Ferry
Soundview Ferry
FDR D
Clay St
Freeman St
Green St
47th Ave
48th Ave
495
25
9A

Performance:

Loading Entire Dataset:

Time to read from SQL: Patience ran out

Time to read from CSV: 0.88 mins

Time to read from pickle: 0.07 mins

Loading time for Driver/Data/Fare_Analysis Models:

Time to read from SQL: 1.18 mins

Time to read from CSV: 0.11 mins

Time to make pickle: 0.03 mins

Time to read from pickle: 0.01 mins

Total time: 1.33 mins

Exploring Parquet file format (entire dataset):

Time to read from paraquet: 0.24 mins

Time to read from CSV: 0.84 mins

Time to read from pickle: 0.06 mins

Challenges

1. The size of the dataset was too large for PostgreSQL to import in one go

```
\copy trip_data_raw FROM 'D:/trip_data/trip_data_2.csv' DELIMITER ',' CSV HEADER;


ERROR:  could not stat file "D:/trip_data/trip_data_2.csv": value too large

SQL state: XX000
```

Remediation:

Split csv file into chunks of 3 or 4 files to make it more manageable.

2. Pandas dataframe takes too much time to read the query using either psycopg2 or pyodbc

Remediation:

Pickle the tables for faster read times

Security

The medallion and hack_licence fields, though hashed, are not as secure as they might seem. Using [5], I was able to unhash some of the medallion values and all the hack_licence values that I tried. The prices for medallions peaked around 2013 at over $1,000,000 [6]. It's a good thing that the company has made sure such valueable information is not stored in plaintext, it can be made better with salting or choosing other techniques such as scrypt, bcrypt and PBKDF2.

medallion md5 – unhashed:

Hash	Type	Result
972C54ECDF892AB63BF75975779C425E	md5	1V62
A4E590BD8D71D53C7887649EFD8B5532	md5	8A54
F0581E1A2A56A5917506660C8FFA7878	Unknown	Not found.
B5FEE89ADD8F36CAE0896474B4CB5522	md5	3D14
A6C0D15006C7592E686798E025293BDD	Unknown	Not found.
9205B3DC797AB9B1AA6929F36D6EAED8	Unknown	Not found.
86F6DC3CEEE95AA60C65F68E07D5F85B	md5	5C77
451E378C8757925E2EE248E5142EDF99	Unknown	Not found.
C3FB6119AC1744CBFEA564666D9DD0CB	Unknown	Not found.
419FD46C50CFA0CFD57170AC0D580310	Unknown	Not found.
7F344E842AAC8F7FBC4AEA1C53CD209E	md5	4F57

Color Codes: Green: Exact match, Yellow: Partial match, Red: Not found.

hack_license md5 – unhashed:

Hash	Type	Result
C62170F2BCE4A72BF700E631FFA55790	md5	5457029
0610C5A089CF755E38567E01609DDC23	md5	5200905
FA90BA5AEAB5CE6F8656AA395268F4CA	md5	5438462
14034BB590CF08BA9B41D0C4D6852B40	md5	5388144
ED1DEEDFD1B4A55EED5D97BCD3363D25	md5	410134
CBB3E0D99C02B3998EE7A009857A375D	md5	5146204
ED0014C5C4E7EA18A5F92F5C2359E1CD	md5	485842
9D89720EC460E5C8F717A2E6A1EBEA6D	md5	5457227
26E669A9D8004756172F7B93E77C2366	md5	5446296
500F2B4324F29CC24F0D5B3C59199166	md5	5259914
A999EF8A46251492437FDA23C8F7DF3C	md5	5429725
69C7DA2905202CD02A44FDBE5D291DCA	md5	5426846
E80940B3F309D6526A958DD6E6CDF5FB	md5	399664
C778C85E82DFC902B2E63F700DABC1D5	md5	5409629
89C520125C85CF698A2771AABC065639	md5	5432365
2CD8683E2C15D4F6A5FB8D120FBAF9DD	md5	480337
A0199F8B131BACEDA37FC8E492D86796	md5	422722
92097029E14ED297701B9E9912E89137	md5	5296480
9D77E425BA981D6187D92AA91E369357	md5	5127995
F5951A3CB100069A1D5BEEF1427D0D76	md5	5299347

Color Codes: Green: Exact match, Yellow: Partial match, Red: Not found.

Hash	Type	Result
9E17E3E8F44897945B08127A0425A3A5	md5	477908
20585F0D2291E94A1401A5FCF94D1371	md5	388066
3B2E865206BC55C5E0DB0B325D75090F	md5	501701
32E7F03F79210B7D24C80699156D8D2C	md5	445747
F184B14153E2F18854A4636130CD3939	md5	500636
7BED4A1CB3DAFE307854FDBCB69F5D57	md5	5360755
0A08EC79EF0E7F3C14DC4AD92D27F97F	md5	468773
FEC567B6210A503FA424245ED2EED431	md5	5340600
E70869C61FA355A82C1E70711F3F054F	md5	483603
D03BCD997D8CEE75125577A9EDCB01D5	md5	429448
14BA17C349C90A92FB962D35A8F6E55B	md5	5390482
A978E00491FEF62927C2C598477D3E2D	md5	5067862
F7BFAEA11982396C5EEDA5680EE78C69	md5	408081
15E2E688CA84FFA4C7A567B959D26EC5	md5	5177671
9F3E40140E6C2F607F7CA1BAF5E60AAF	md5	5450963
4BCF96CCB6C35AA09430A1557468AC82	md5	5368797
8057FAEC813F45E992BAD750C1A586CC	md5	5170742
4DD96B922FB2326C2977C331805FD965	md5	271661
EA905FA18437A11183901209FB2438CF	md5	5326801
10B3C8B53A1A89F7E80FAE8DDC8295EA	md5	5404295

Color Codes: Green: Exact match, Yellow: Partial match, Red: Not found.

References:

[1] https://thinkmetric.uk/basics/speed/

[2] https://www.postgresql.org/docs/13/datatype-numeric.html#DATATYPE-NUMERIC-DECIMAL

[3] https://en.wikipedia.org/wiki/Boroughs_of_New_York_City

[4] https://www1.nyc.gov/assets/tlc/downloads/pdf/rule_book_current_chapter_54.pdf

[5] https://crackstation.net/

[6] https://en.wikipedia.org/wiki/Taxi_medallion