YellowSpark

NYC 2013 Taxi Data Analysis

Maxime Lovino, Marco Rodrigues Lopes, David Wittwer

Steps taken

- Reading and cleaning data using Spark (Scala)
- Data analysis using Spark (Scala)
- Linear regression models for fares with Spark ML Lib (Scala)
- ML models for trip duration estimation with Spark ML Lib (Scala)
- Visualisations with Jupyter Notebooks using PySpark (Python)

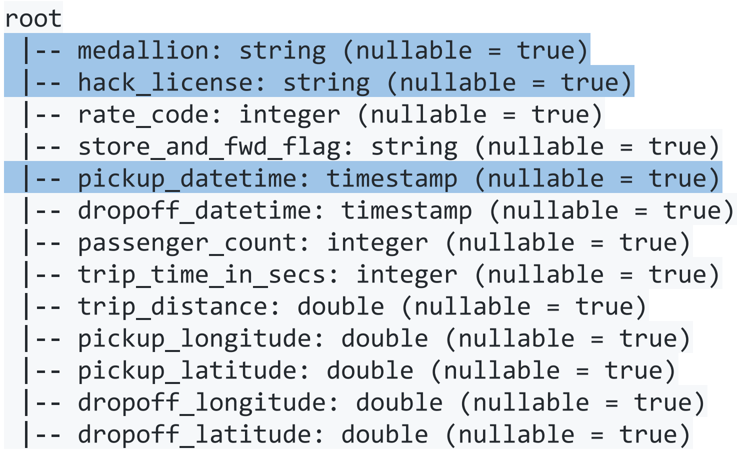
The Data

The Data

- One file per month for rides information
- One file per month for fares information
- CSV Format
- Download and then stored on S3
- 165'163'063 rides after cleaning
- 50 GB of data
- Reading takes 45 minutes on 2 m4.2xlarge
- Read, cleaned and saved as Parquet dataframe in S3 => 8.8 GB

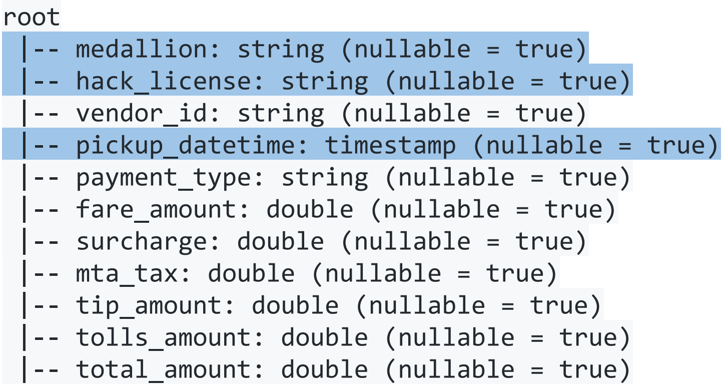
The Data

Rides



```
root
 |-- medallion: string (nullable = true)
 |-- hack_license: string (nullable = true)
 |-- rate_code: integer (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- passenger_count: integer (nullable = true)
 |-- trip_time_in_secs: integer (nullable = true)
 |-- trip_distance: double (nullable = true)
 |-- pickup_longitude: double (nullable = true)
 |-- pickup_latitude: double (nullable = true)
 |-- dropoff_longitude: double (nullable = true)
 |-- dropoff_latitude: double (nullable = true)
```

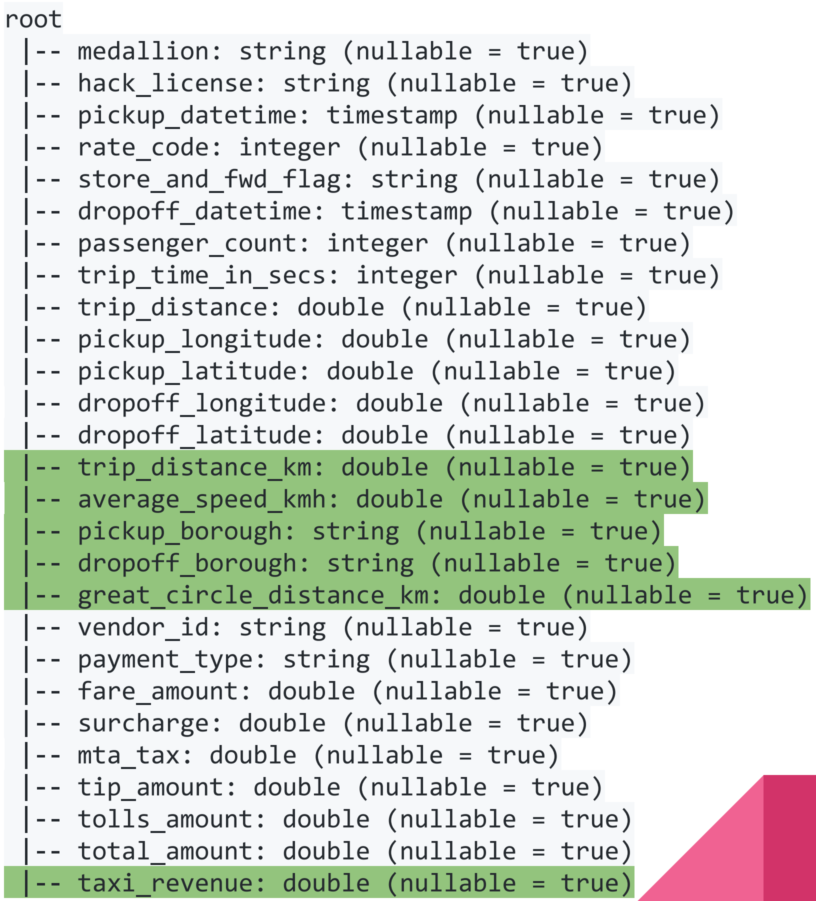
Fares



```
root
 |-- medallion: string (nullable = true)
 |-- hack_license: string (nullable = true)
 |-- vendor_id: string (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- payment_type: string (nullable = true)
 |-- fare_amount: double (nullable = true)
 |-- surcharge: double (nullable = true)
 |-- mta_tax: double (nullable = true)
 |-- tip_amount: double (nullable = true)
 |-- tolls_amount: double (nullable = true)
 |-- total_amount: double (nullable = true)
```

The Data

- Extracted features
  - Average speed
  - Great circle distance
  - Kilometer distance
  - Taxi Revenue
  - Borough mapping from coordinates

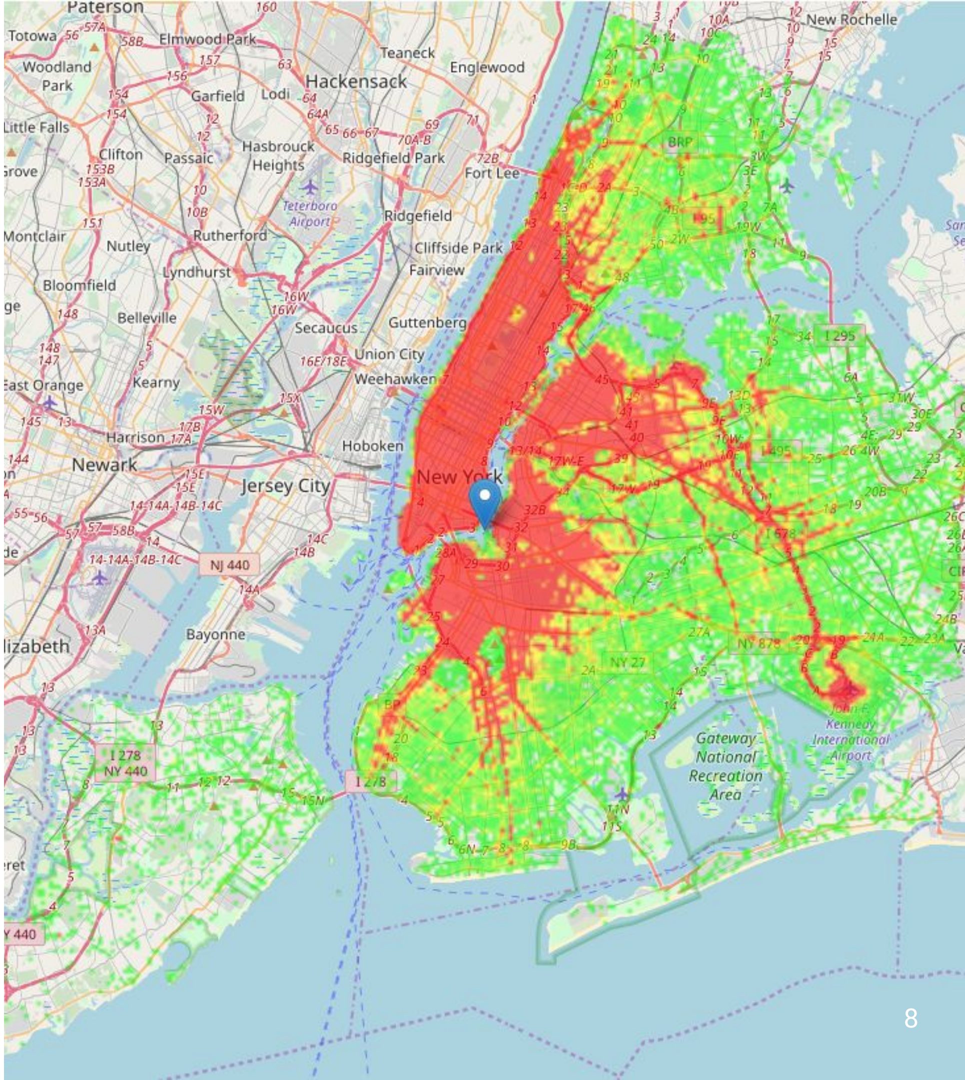


```
root
 |-- medallion: string (nullable = true)
 |-- hack_license: string (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- rate_code: integer (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- passenger_count: integer (nullable = true)
 |-- trip_time_in_secs: integer (nullable = true)
 |-- trip_distance: double (nullable = true)
 |-- pickup_longitude: double (nullable = true)
 |-- pickup_latitude: double (nullable = true)
 |-- dropoff_longitude: double (nullable = true)
 |-- dropoff_latitude: double (nullable = true)
 |-- trip_distance_km: double (nullable = true)
 |-- average_speed_kmh: double (nullable = true)
 |-- pickup_borough: string (nullable = true)
 |-- dropoff_borough: string (nullable = true)
 |-- great_circle_distance_km: double (nullable = true)
 |-- vendor_id: string (nullable = true)
 |-- payment_type: string (nullable = true)
 |-- fare_amount: double (nullable = true)
 |-- surcharge: double (nullable = true)
 |-- mta_tax: double (nullable = true)
 |-- tip_amount: double (nullable = true)
 |-- tolls_amount: double (nullable = true)
 |-- total_amount: double (nullable = true)
 |-- taxi_revenue: double (nullable = true)
```

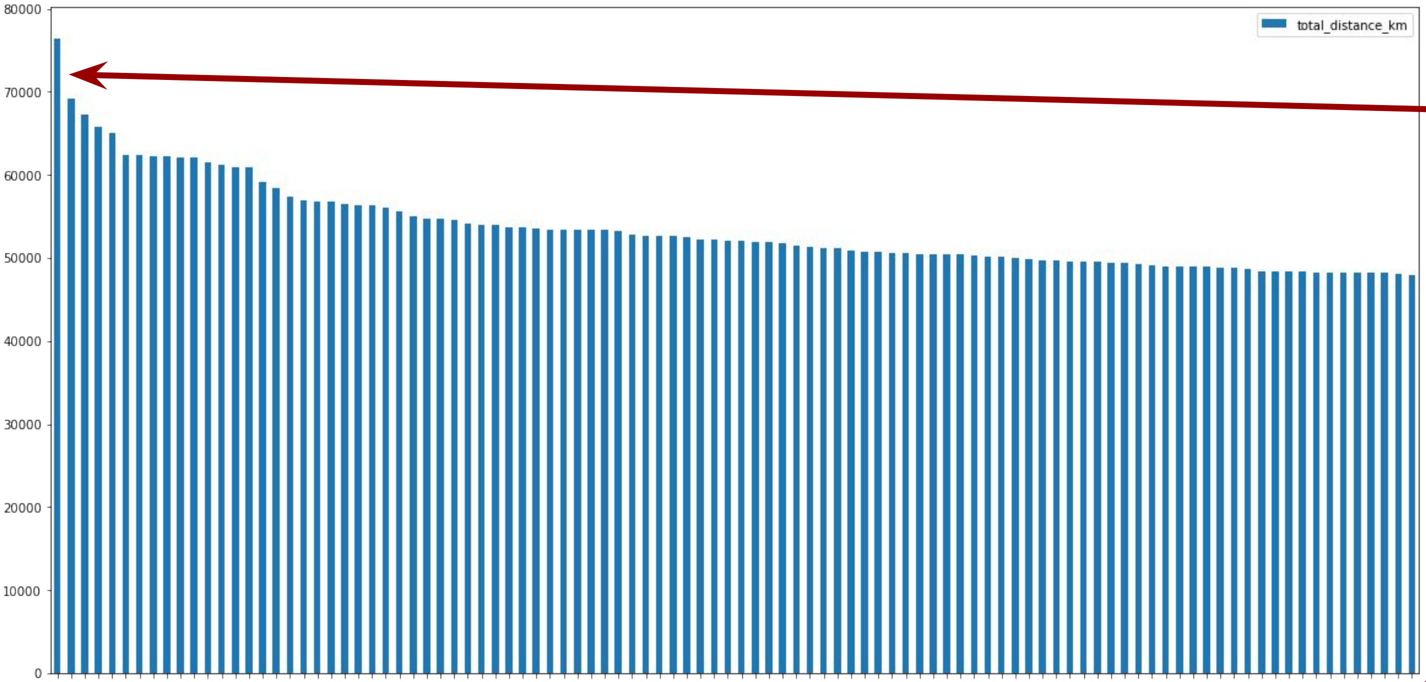
The Data - Cleaning

- All rides with average speed above 120 km/h
- Pickup or dropoff outside NYC boroughs
- Standard fare with distance smaller than great circle distance
- 0 passengers in the car
- Fare of 0 $ or less
- Trips longer than 24 hours (why does this exist?)
- Trips slower than 1 km/h on average

Data Analysis



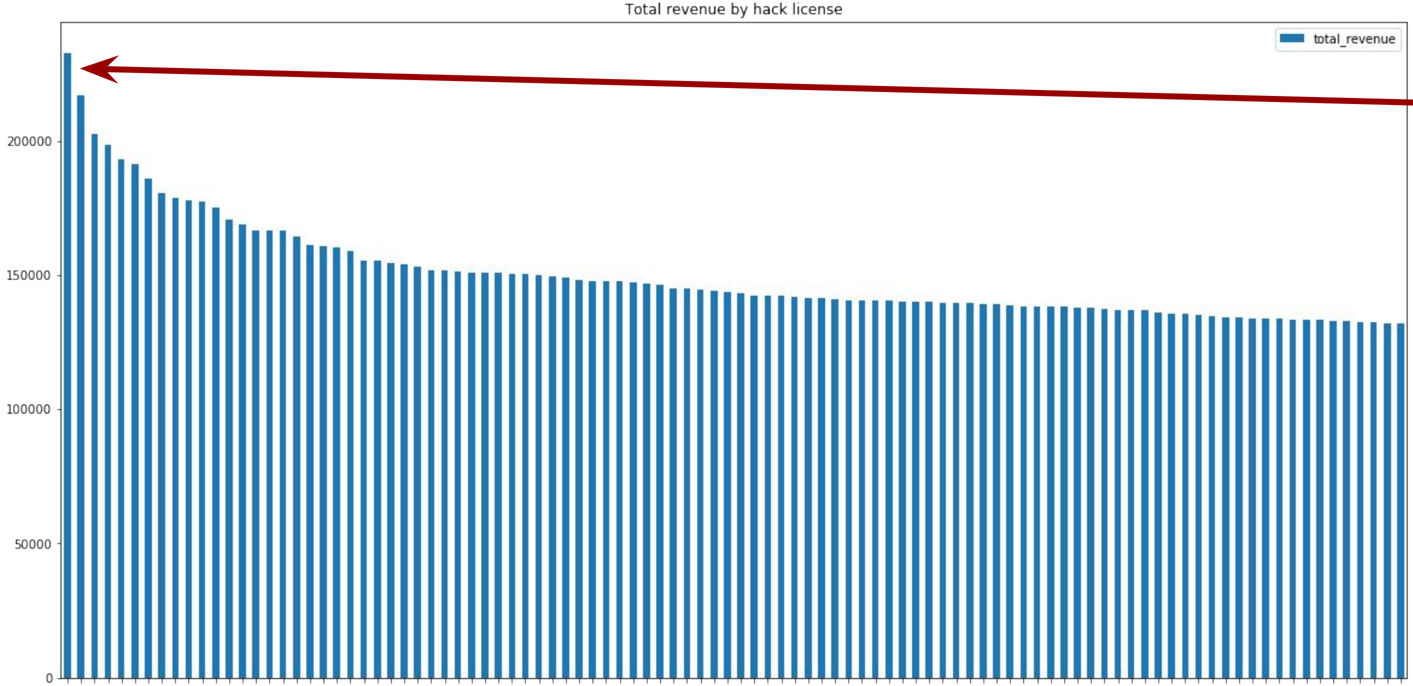
Analysis by drivers - Total distance - Top 100



76’390 km

42528 unique licenses

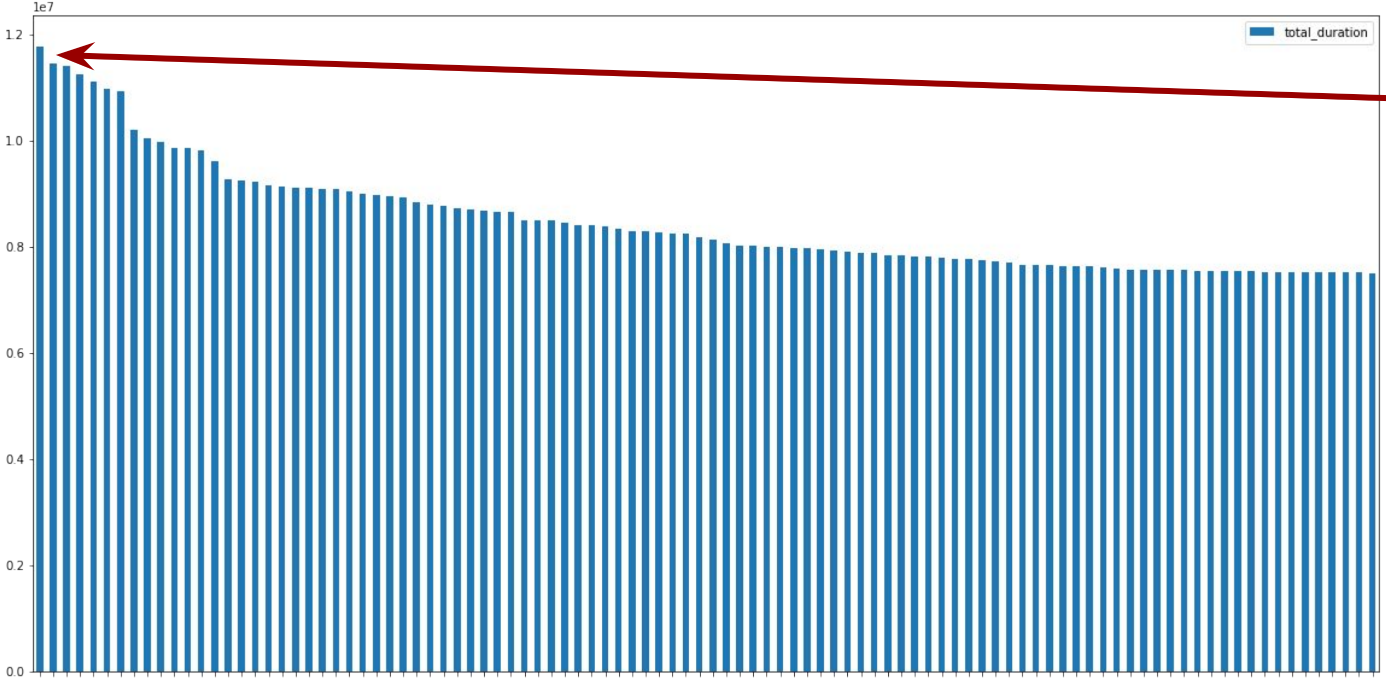
Analysis by drivers - Total revenue - Top 100



232’793 $

42528 unique licenses

Analysis by drivers - Total time on rides - Top 100

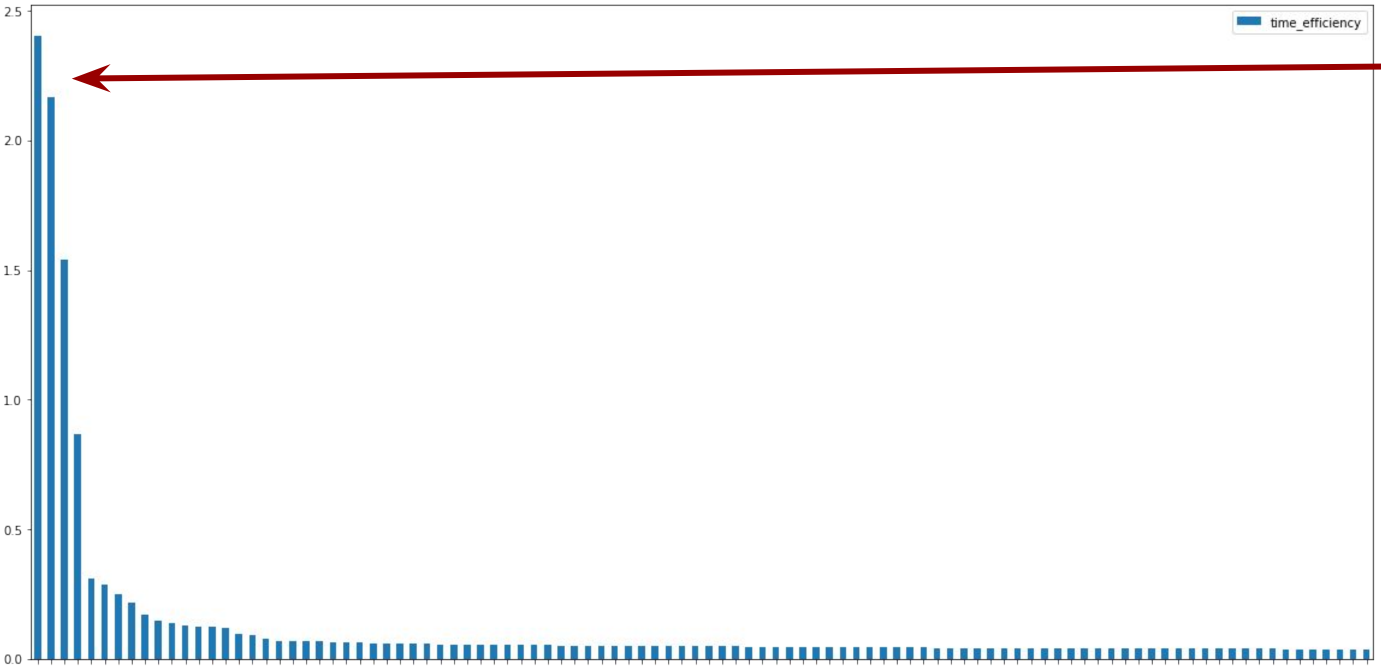


136 24-hours days

9 hours / day on rides

42528 unique licenses

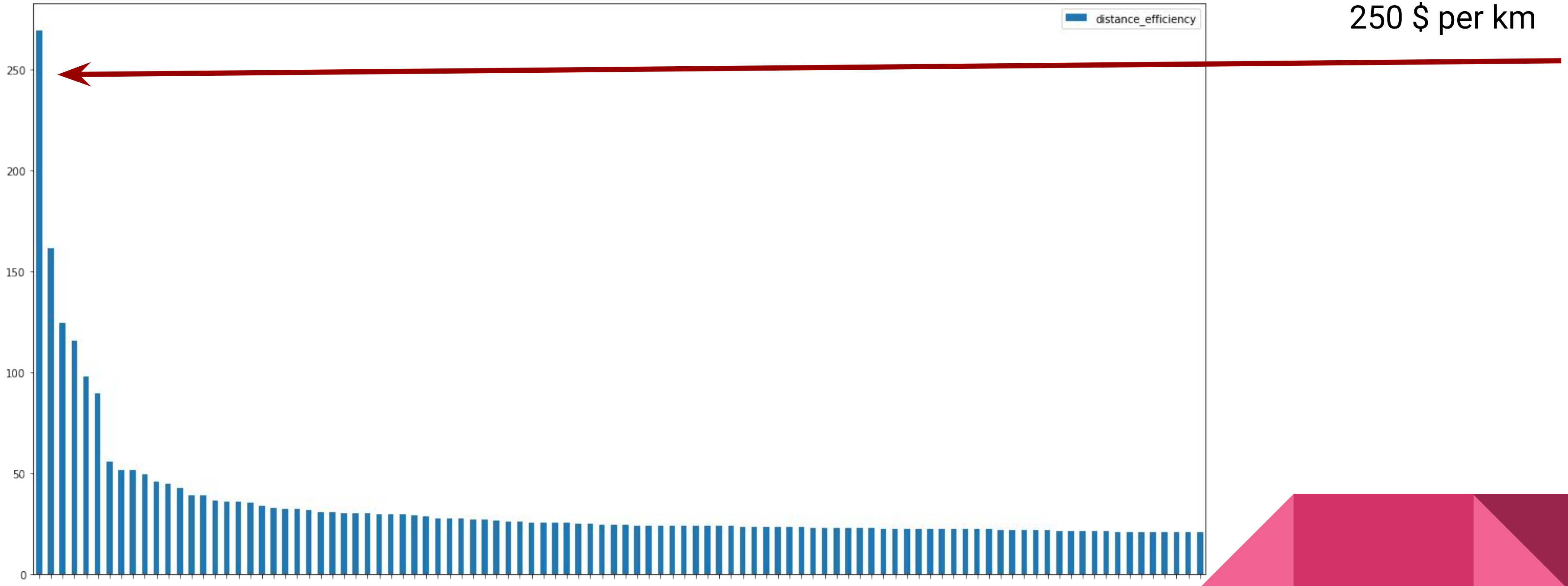
Analysis by drivers - Time efficiency - Top 100



2.5 $ per second

42528 unique licenses

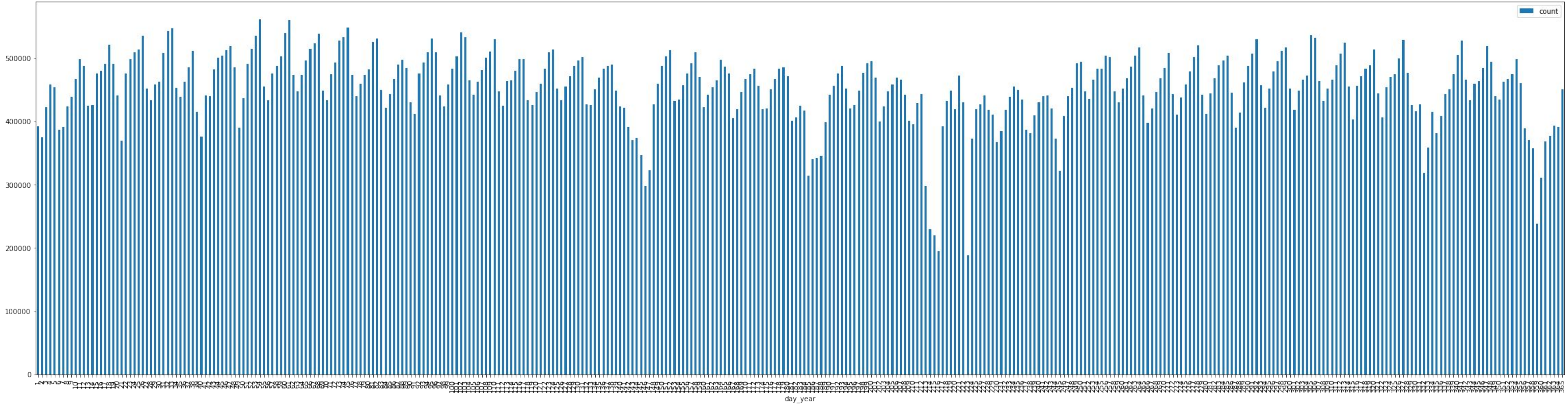
Analysis by drivers - Distance efficiency - Top 100



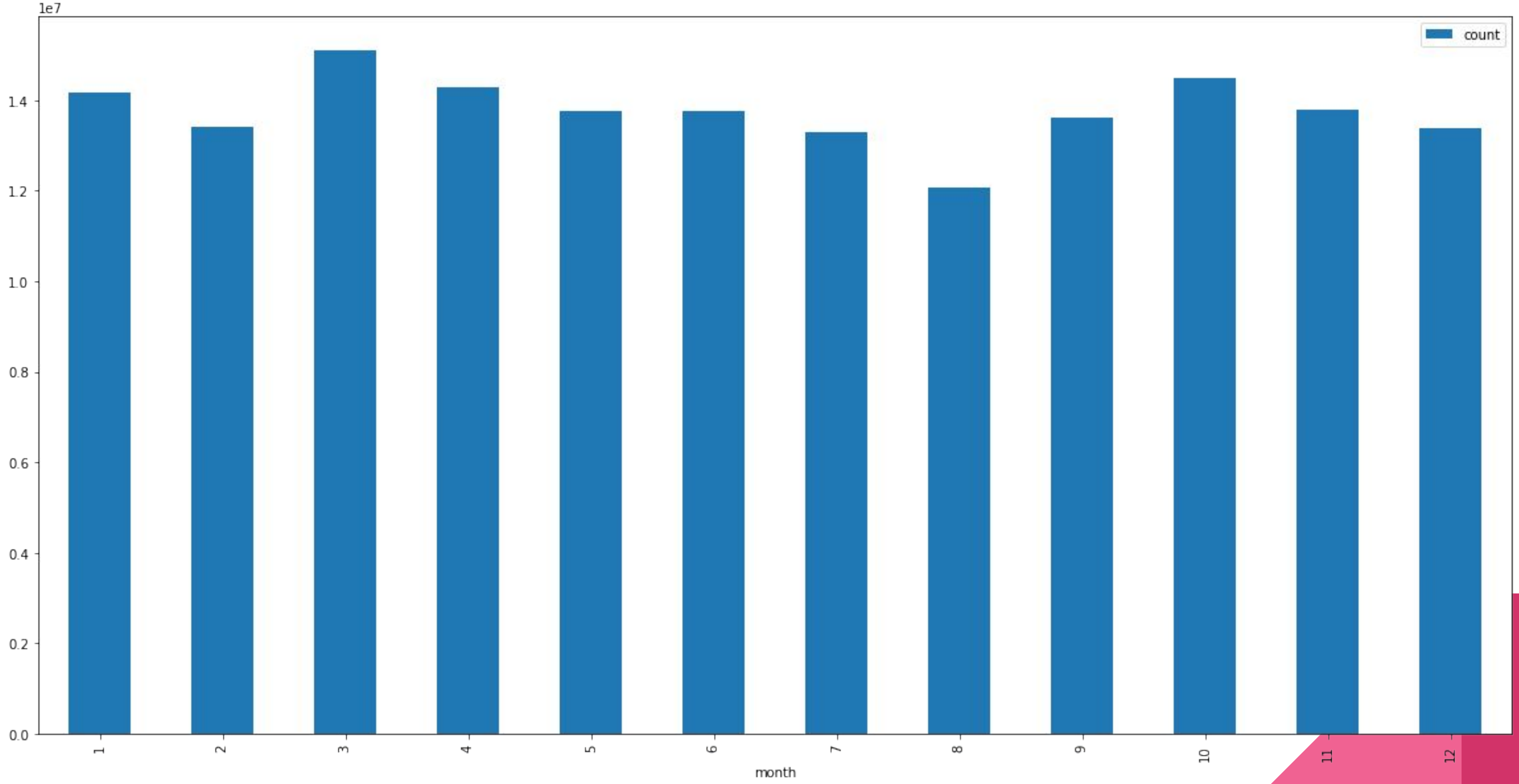
250 $ per km

42528 unique licenses

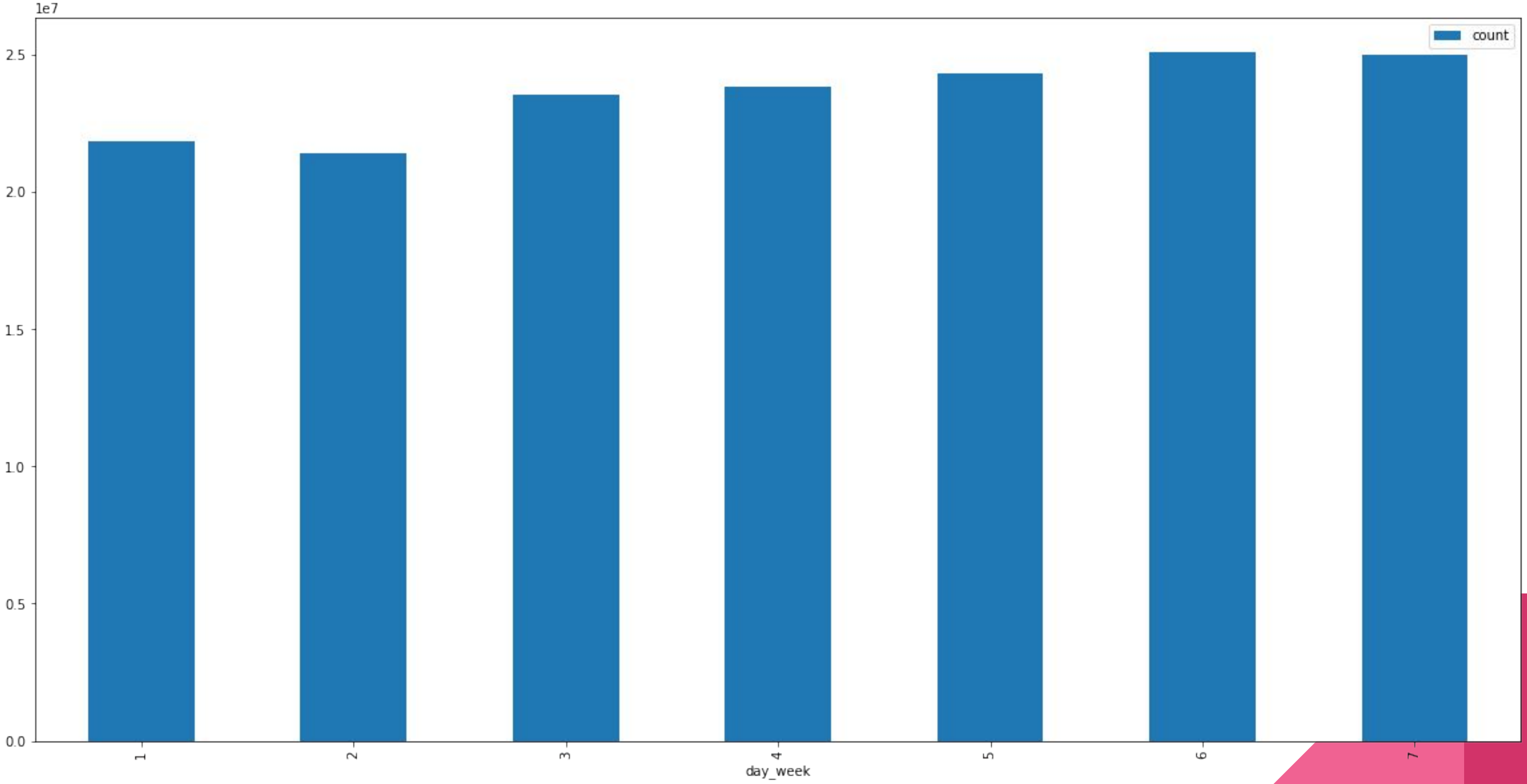
Analysis by dates - Year



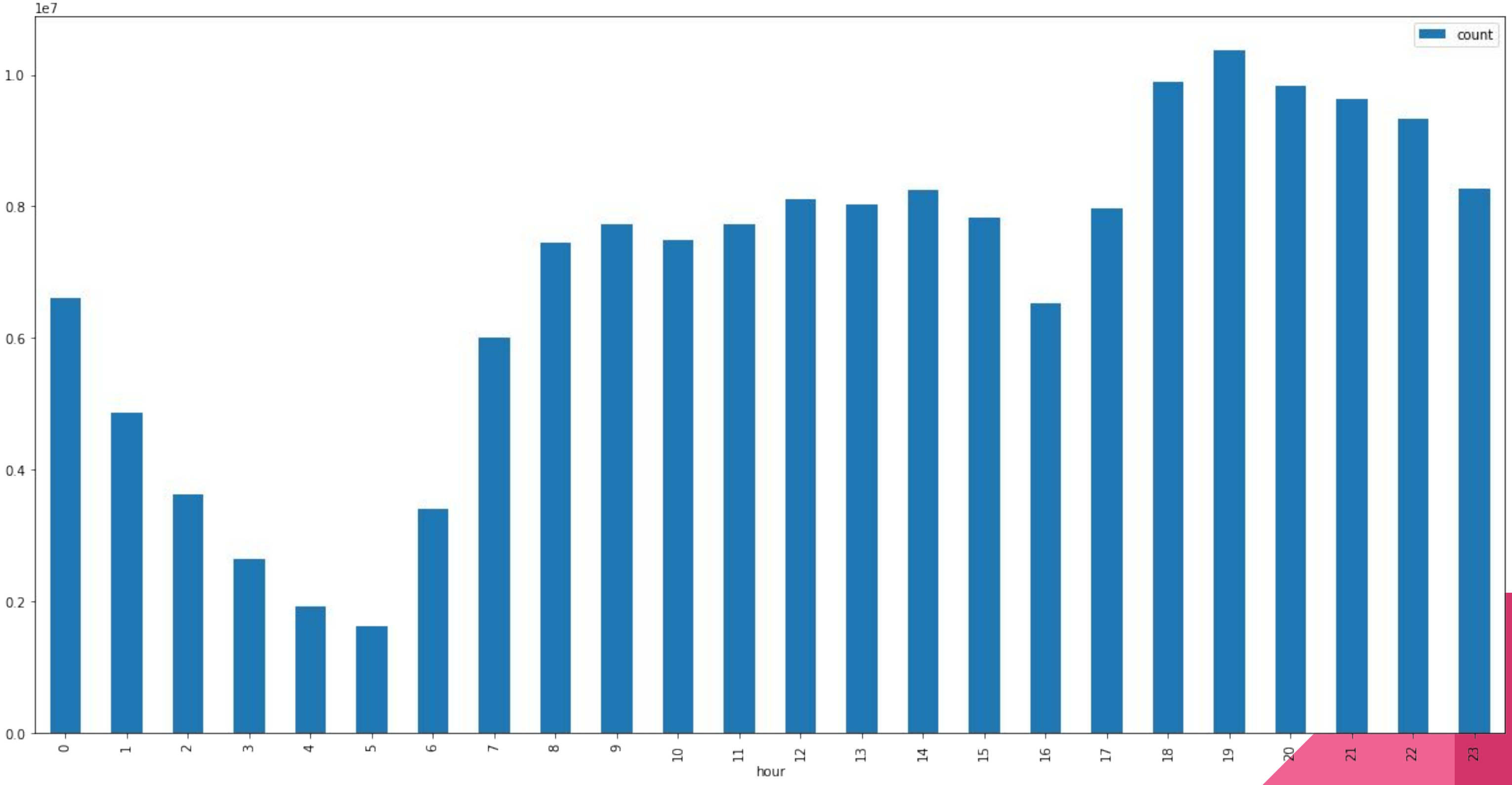
Analysis by dates - Month



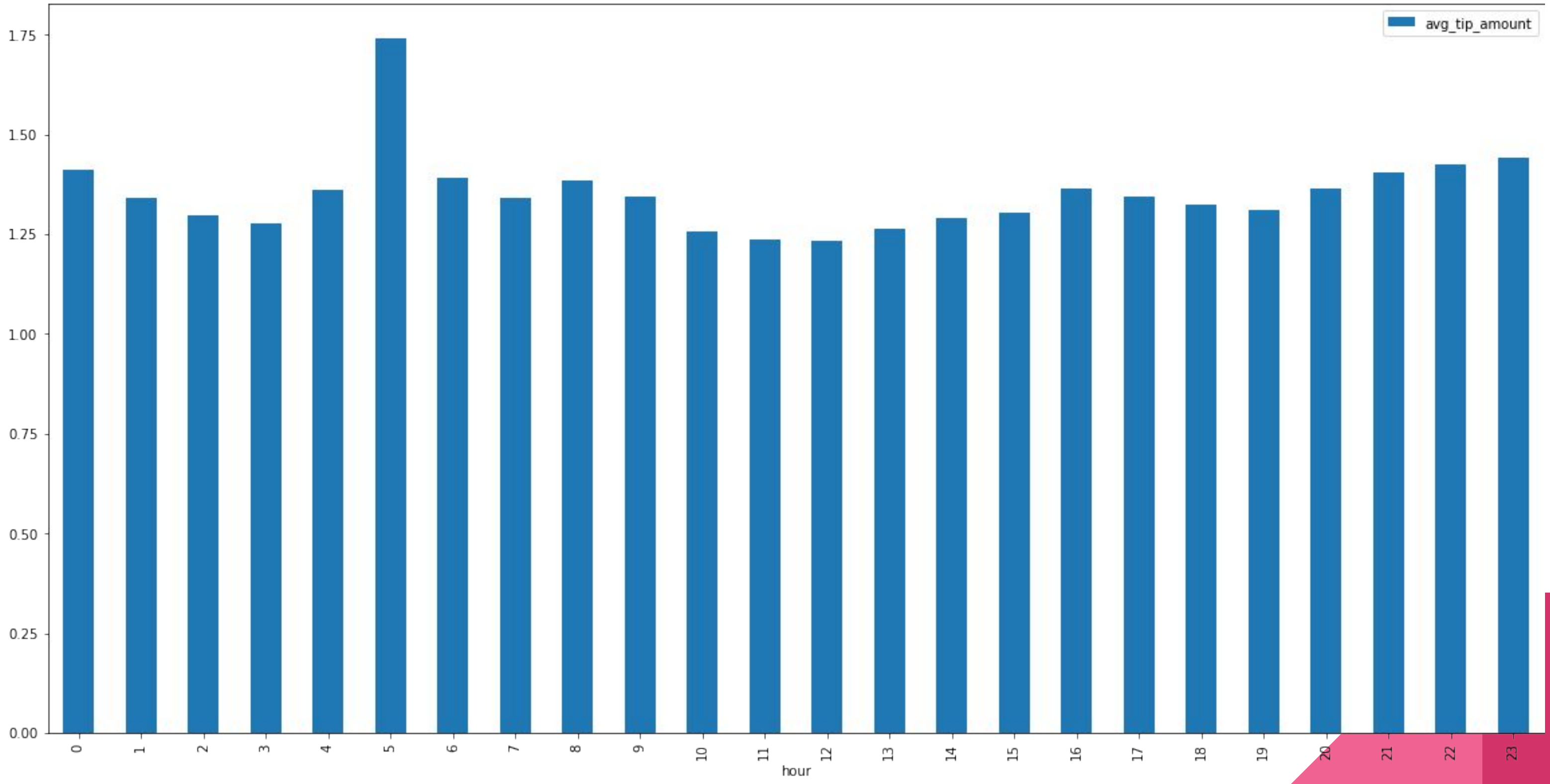
Analysis by dates - Day of week



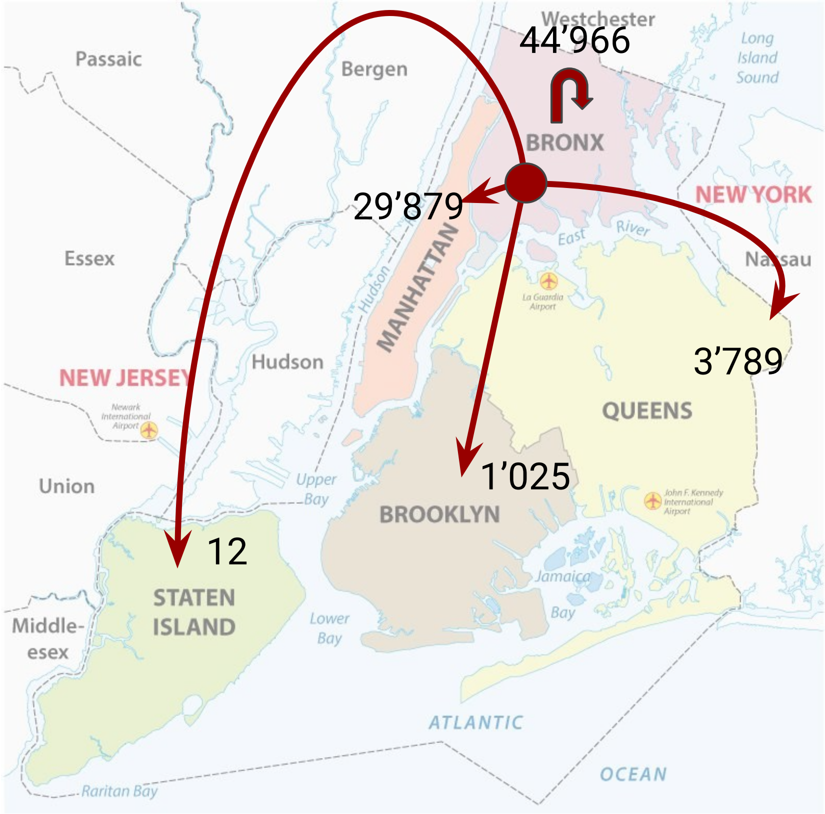
Analysis by hour of the day - Number of rides



Analysis by hour of the day - Average tips

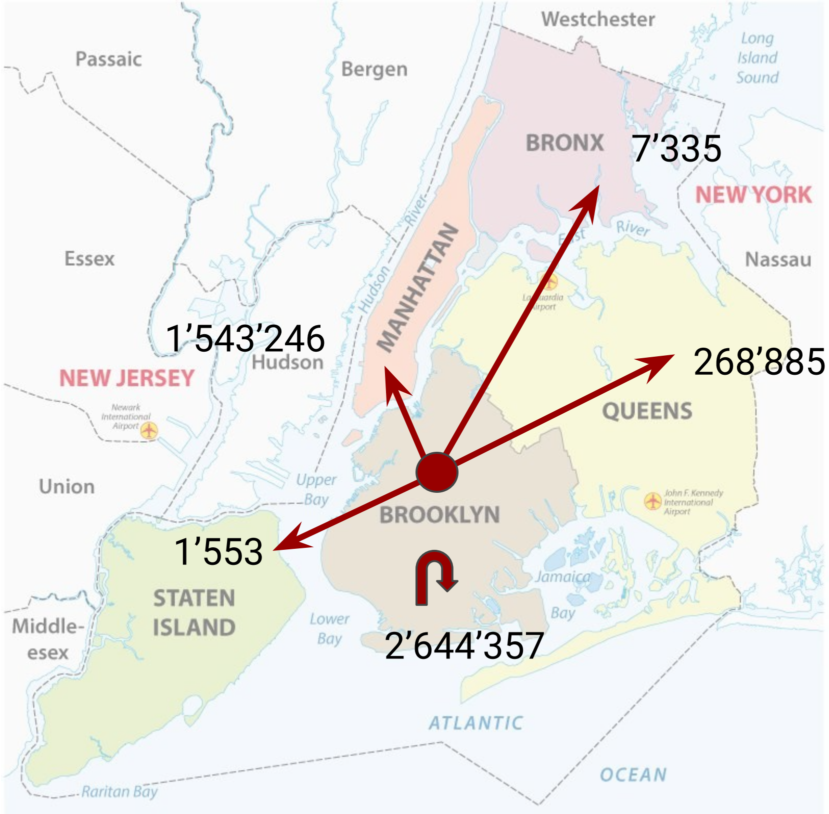


Analysis by boroughs - Bronx



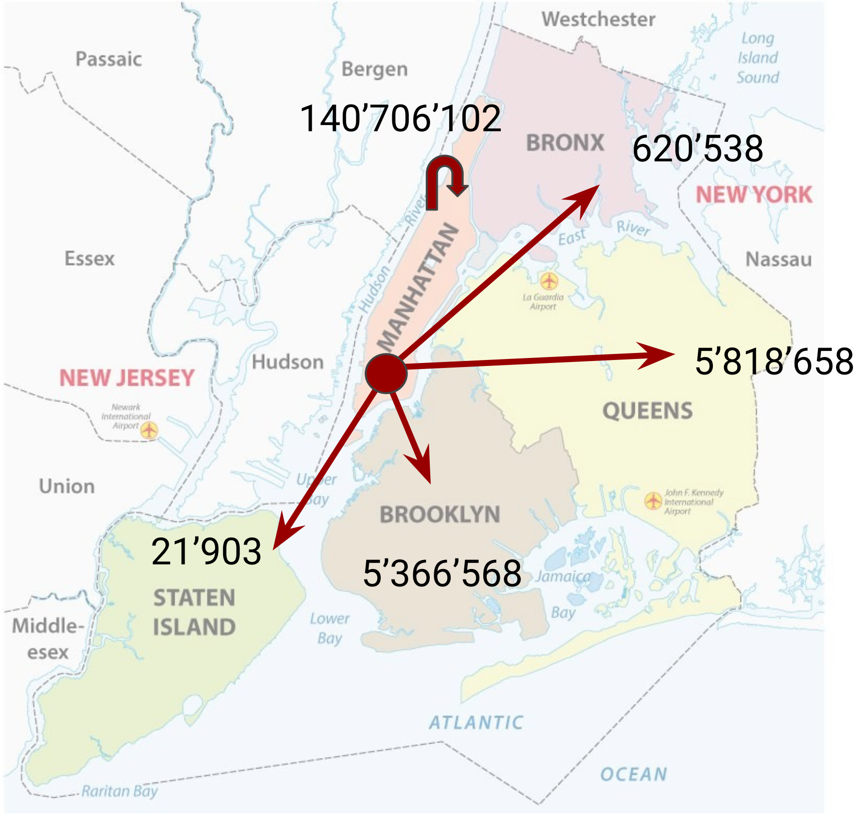
- 79’671 rides from the borough
- 0.04 % of all rides

Analysis by boroughs - Brooklyn



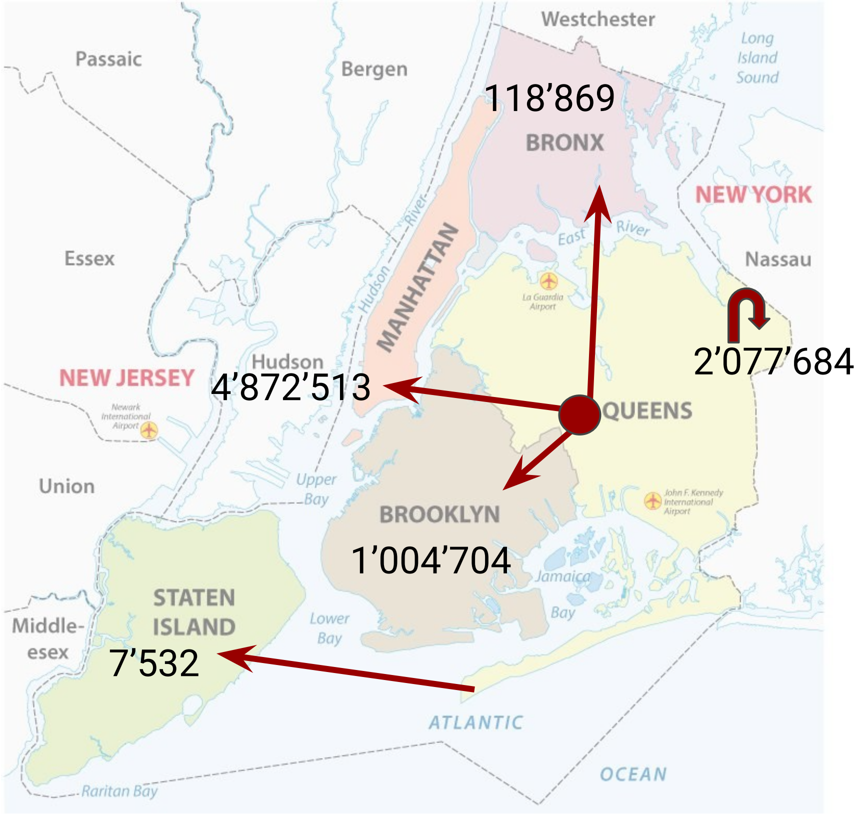
- 4’465’376 rides from the borough
- 2.7 % of all rides

Analysis by boroughs - Manhattan



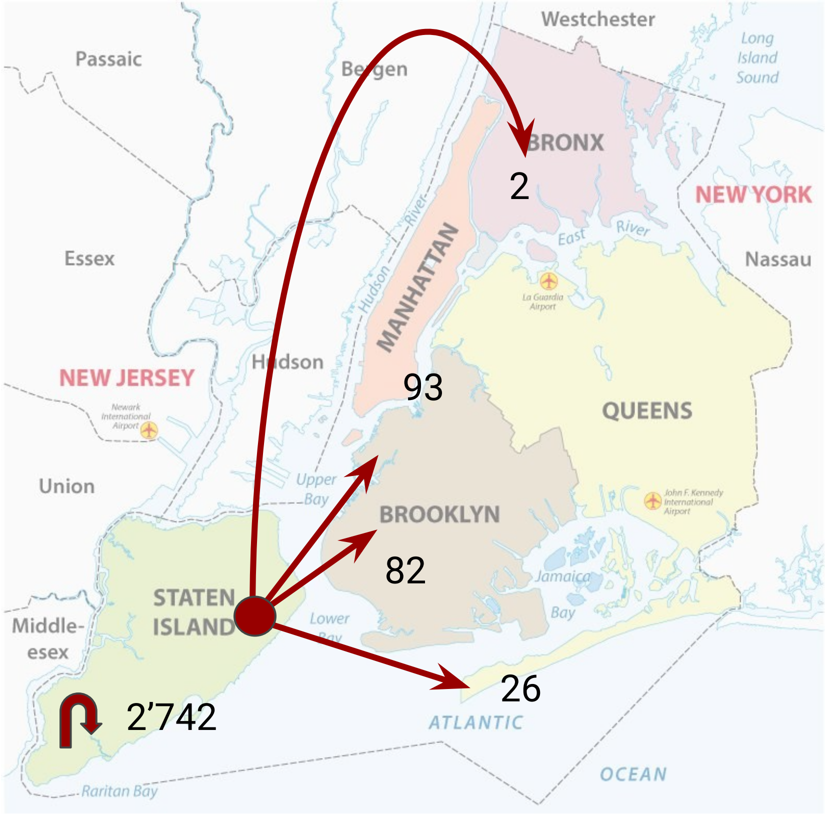
- 152’533’769 rides from the borough
- 92.3 % of all rides

Analysis by boroughs - Queens



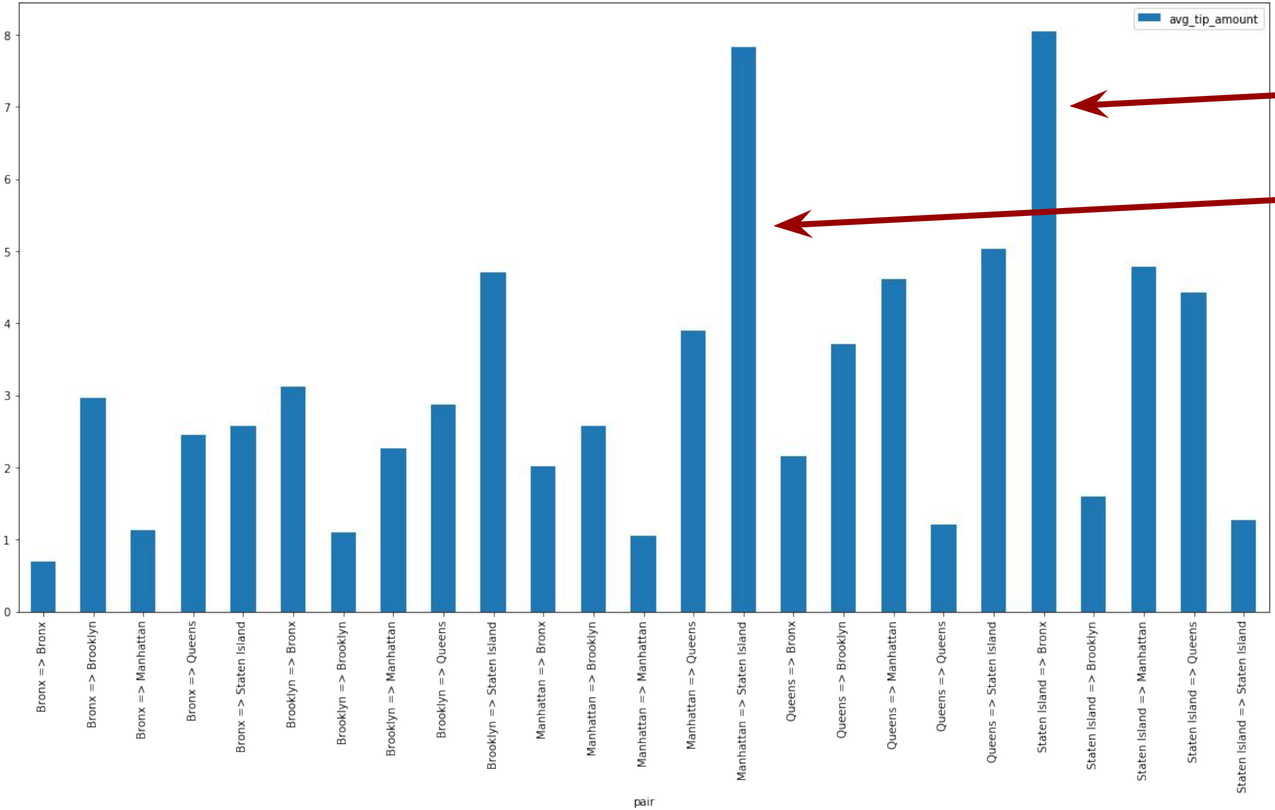
- 8'081'302 rides from the borough
- 4.89 % of all rides
- Queens contains 2 airports

Analysis by boroughs - Staten Island



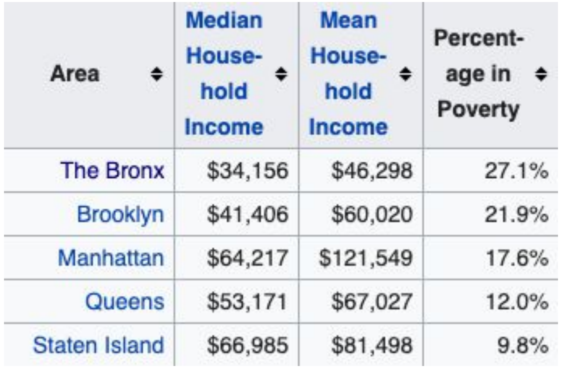
- 2945 rides from the borough
- 0.002 % of all rides

Analysis by boroughs - Who tips the most?



Staten Island => Bronx

Manhattan => Staten Island



Area ⬍	Median House-hold Income ⬍	Mean House-hold Income ⬍	Percent-age in Poverty ⬍
The Bronx	$34,156	$46,298	27.1%
Brooklyn	$41,406	$60,020	21.9%
Manhattan	$64,217	$121,549	17.6%
Queens	$53,171	$67,027	12.0%
Staten Island	$66,985	$81,498	9.8%

Machine Learning

Machine Learning - Estimating fares

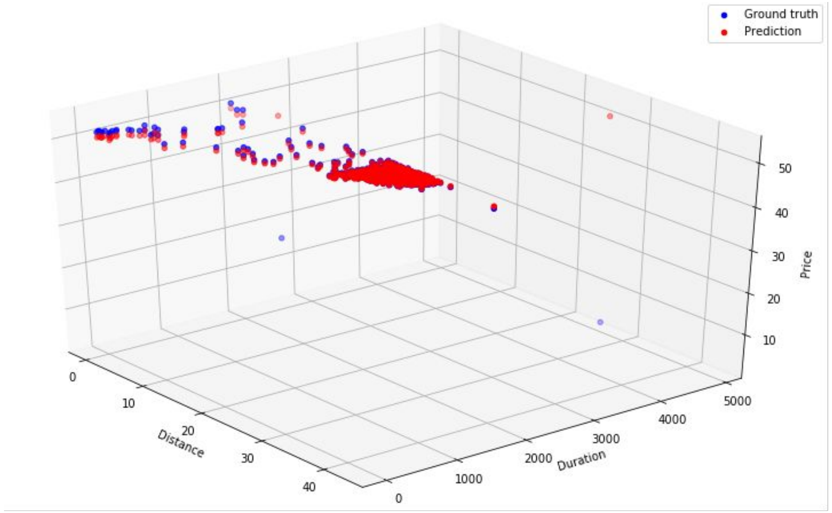
- Trained one model for each rate code from 1 to 4
- Train - Test Split 70% / 30%
- Linear Regression model
- Features are distance and duration
- Aim to estimate fare amount

Machine Learning - Estimating fares

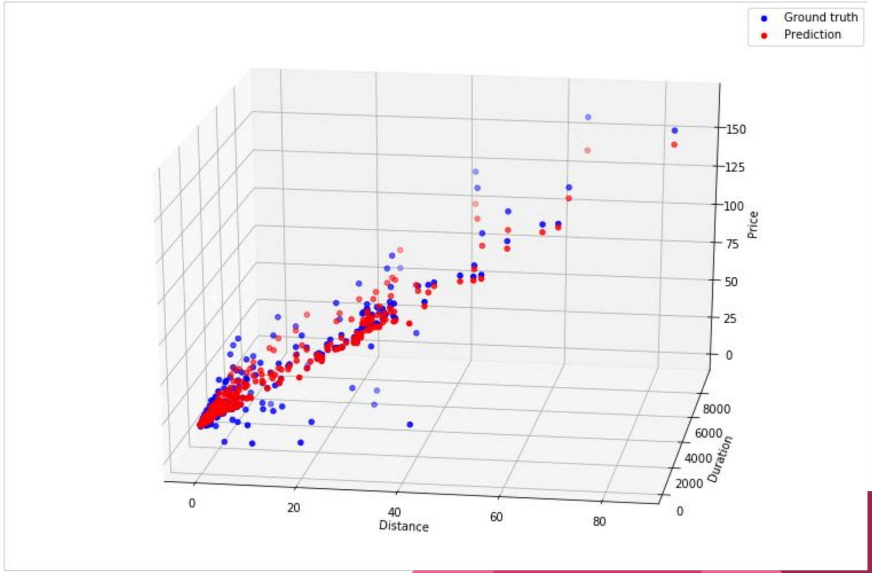
Rate Code	MSE (Mean Squared Error)	Formula
1	1.804926	0.006 * seconds + 1.201 * km + 2.095
2	4.070359	0.000 * seconds + 0.039 * km + 50.818
3	56.277111	0.003 * seconds + 1.371 * km + 20.203
4	16.970909	0.004 * seconds + 1.642 * km + 2.047

Machine Learning - Estimating fares

Rate code 2



Rate code 3



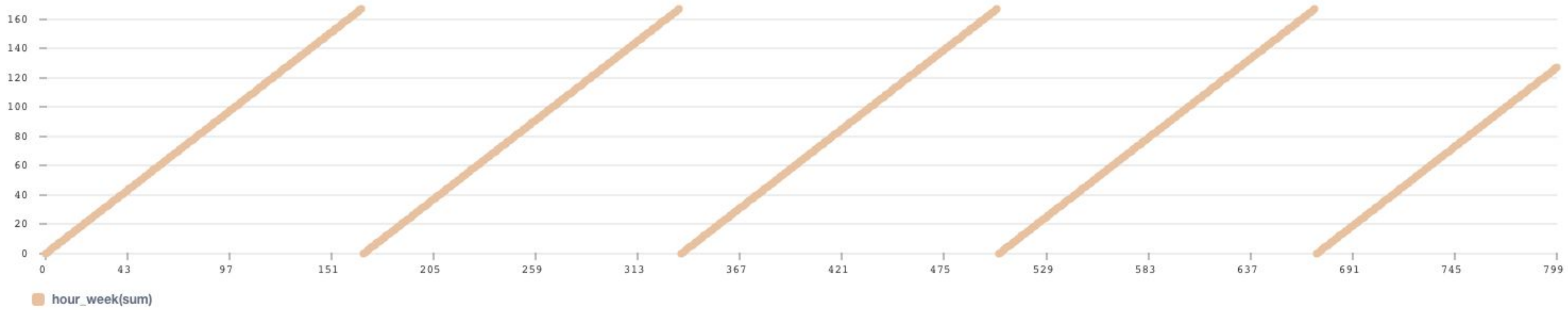
Machine Learning - Estimating trip duration

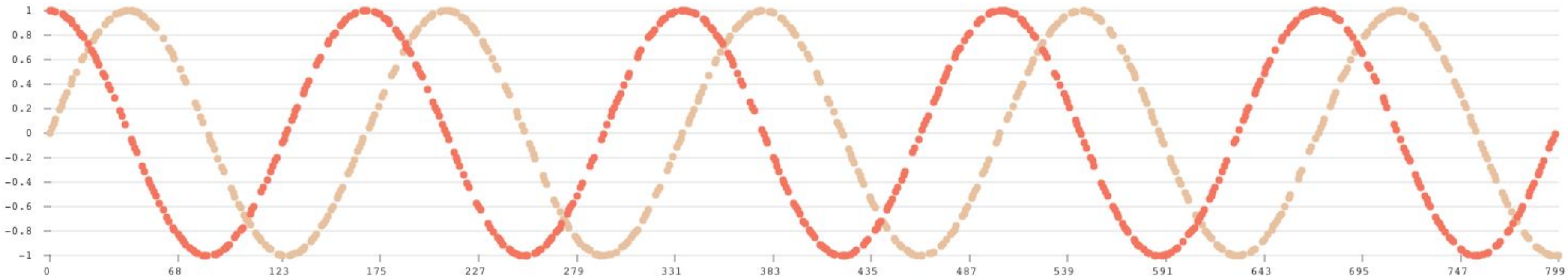
The “Billion Dollar Problem”

- Trained the model only for Manhattan
- 5 hours to train
- Extracted new features (more on next slide)
    - Hour of the week for pickup (sin and cos)
- Features used
    - Hour of the week for pickup (sin and cos)
    - Pickup coordinates
    - Dropoff coordinates
    - Distance
- Gradient Boost Regressor to estimate trip duration
    - Better model in our testings
- Accurate to +/- 4 minutes

Machine Learning - Estimating trip duration

Sinus and cosinus for pickup hour week





Conclusion and improvements

- We could work a full semester on this if we wanted
- Really interesting project and data
- Multi-year data would be interesting to train models
- We should spend way more time on data cleaning
- Cluster rides by behaviors
- We could actually detect anomalies