New York City taxi during 2015

Olga Kholkovskaia

December 30, 2019

1 Motivation

NYC taxi data is widely used for evaluation of different routing and scheduling algorithms. Those datasets are are publicly available. Up to year 2016 they contained exact origin and destination coordinates of taxi trips, and could be directly used for simulation and evaluation. This project is closely connected to my bachelor thesis where we use one day of taxi operation suggest to test our solution. The problem is that the day was chosen at random without any data analysis because of the amount of given data. The aim of this project is analyze several metrics with respect to yellow taxi operation in 2015 in New York, how it changes during the year, if it has some patterns with respect to week days and time, and if it somehow correlates with weather. The project was partially inspired by [Sch15] and [Vel19].
Project code include three scripts for data preparation (mainly hive, some spark), data analysis (spark), and visualization (pandas, matplotlib). All scripts together with graphs are on project's github page:
`https://github.com/kholkolg/bdt_semestral_project` Results of spark queries are in
`https://drive.google.com/drive/folders/1stwV7UCoiQXmanqwFL6oO-IR5OzAb9gM?usp=sharing`

2 Data

Input data for first dataset is downloaded from New York City Taxi and Limousine Commission homepage [New19] where they provide access to trip record data gathered from different New York taxi service providers. For our

project we have chosen Yellow taxi data[1]. Each month of data is contained in separate .csv file where the row represents the trip with 19 columns filled with various information including trip origin and destination coordinates, trip's length, start and end time, fares, taxes and other payment details.
Second dataset [Ben17] contains around 5 years of hourly measurements of various weather parameters, such as temperature, humidity, air pressure from 30 US and Canadian cities (including New York). Each parameter is a separate .csv file having measurement timestamp as row and city name as column.

3 Data preparation

File data_preparation.ipnb

During the first stage we need to upload necessary files to hadoop cluster, filter out irrelevant data, and prepare databases for future work. Taxi dataset does not require any additional preprocessing here. Weather dataset contains data for more than 30 cities from which we are only interested in New York. Besides that, both datasets are processed in similar way: raw data is first loaded into temporary external table, then moved into internal table partitioned by month.

4 Daily statistics

Files data_analysis.ipnb and data_visualization.ipnb

After the data was prepared, we analysed following parameter: number of trips, total travelled distance, and number of passengers on a daily basis. Trip data may contain some corrupted like entries with end time earlier than start time, zero trip duration or trip length together with some missing values that are removed from dataset. The results are presented in Fig. 1, Fig. 2, and Fig. 3. Most of time values lie inside a bounded range with several 'down' peaks in trip and passenger counts that at the first glance seem to correlate with state holidays. Peak in travelled distance in the beginning of December looks suspicious, and may be resulted by outliers. Although the

[1]Yellow taxis accounted for most of taxi traffic during that year

service provider does not use ridesharing[2] passengers that made one common request may travel together, but the amount of trips with more than two passengers as can bee seen from Fig. 4 is very low with most of trips being "individual" with respect to passenger.

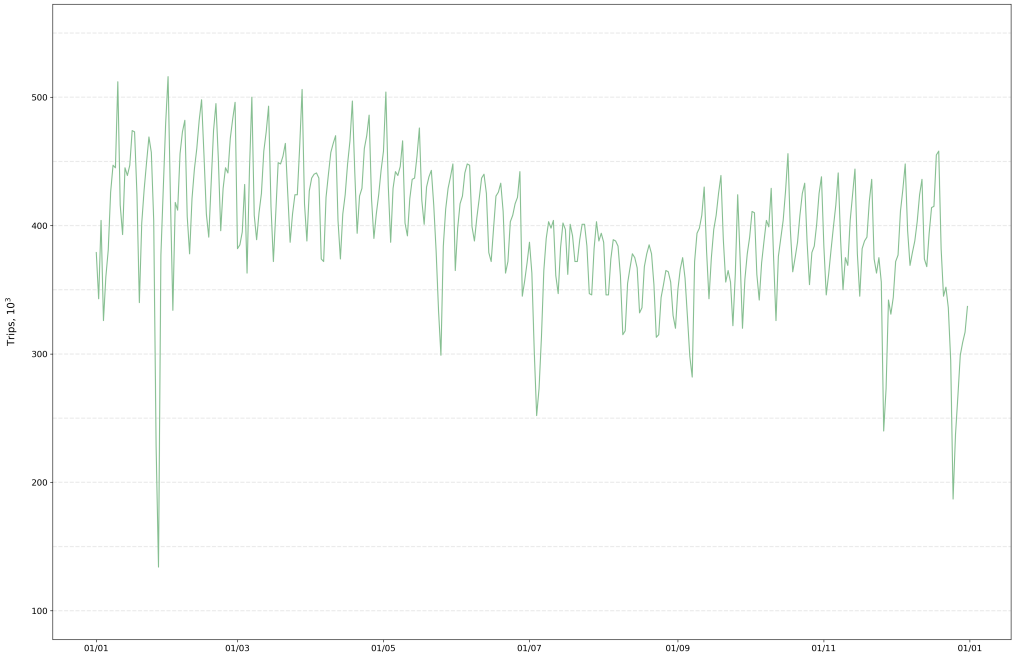

Figure 1: Yellow taxi daily values for total trip count. New York City, 2015

[2]Meaning that passengers that made separate transportation requests are not allowed to share one vehicle during whole or part of their trip

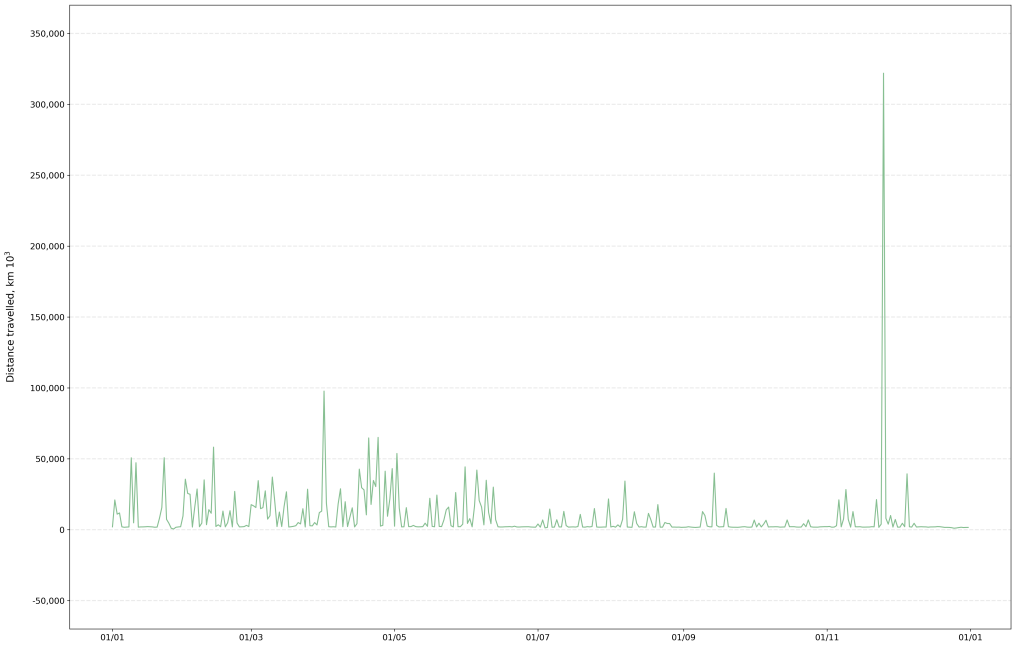

Figure 2: Yellow taxi daily values for total distance travelled. New York City, 2015

5 Daily distributions

Daily trip counts distribution is shown in Fig. 5. As for average speeds, trip lengths and durations histograms computed over the whole range would not be very informative having almost all data in the first bin. As we mainly interested in the central regions of the city, we decided to limit average speed to 40 km/h, trip length to 3 km, and trip duration to 30 minutes. The result is summarized in Fig. 6, Fig. 7, and Fig. 8.

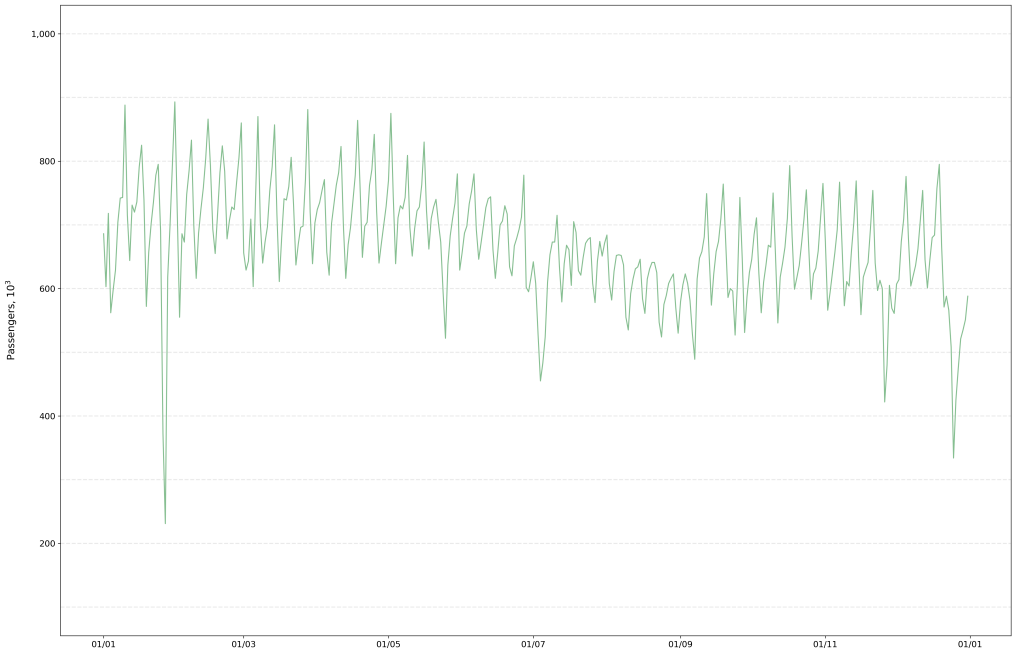

Figure 3: Yellow taxi daily values for number of transferred passengers. New York City, 2015

5.1 Weekday patterns

All weekdays demonstrate similar pattern. Amount of trips decreases during night time reaching its minimum between 3 A.M. and 5 A.M. (See Fig. 9, then starts to grow and stays stable during up to 4 P.M. where it once again increases and remains at that level until 10 P.M. Weekends have same patterned but it all starts and ends later, start of this shift can be noticed already from Wednesday. Speeds shown in Fig. 10 negatively correlate with number of taxi trips as a part of overall traffic. Average trip during night time is about twice as short as during day hours. Those result look similar to congestion levels computed in [Tom18].

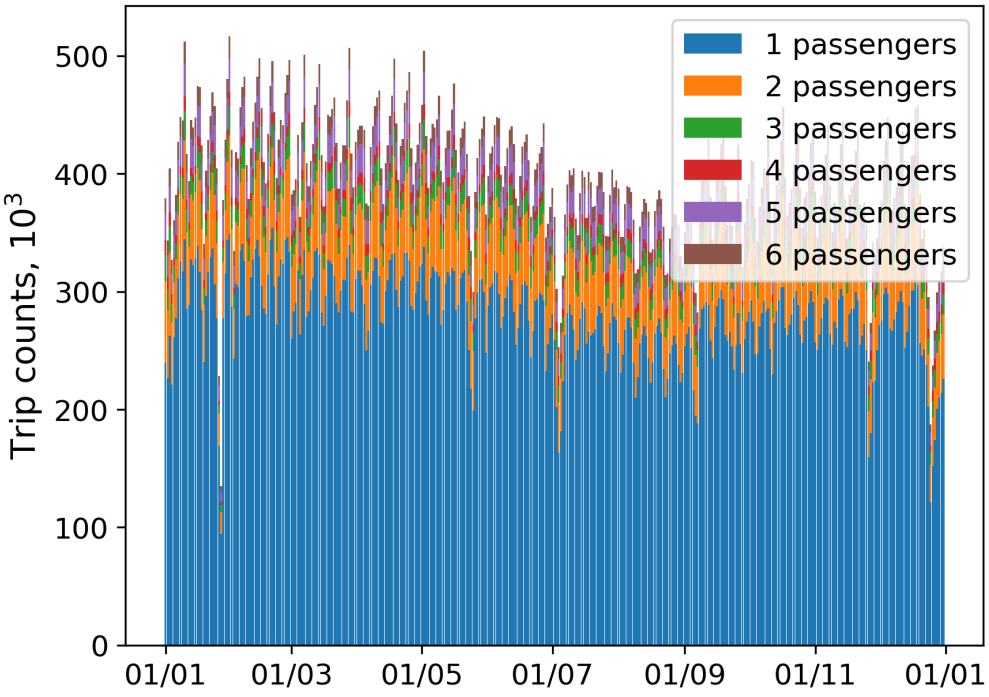

Figure 4: Trips distribution by number of passengers. New York City, 2015

5.2 Connection with weather

Brief answer: no. I would spare everybody's time and won't add more pictures that could be found in data_visualization.py. Number of trips does not correlates with weather measurements except may be the fact that when weather is generally bad people tend not to go outside and thus use less taxi.

References

[Ben17] David Beniaguev. *Historical Hourly Weather Data*. 2017. URL: `https://www.kaggle.com/selfishgene/historical-hourly-weather-data` (visited on 10/30/2019).

[New19] City of New York. *TLC Trip Record Data*. 2019. URL: `https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page` (visited on 10/30/2019).

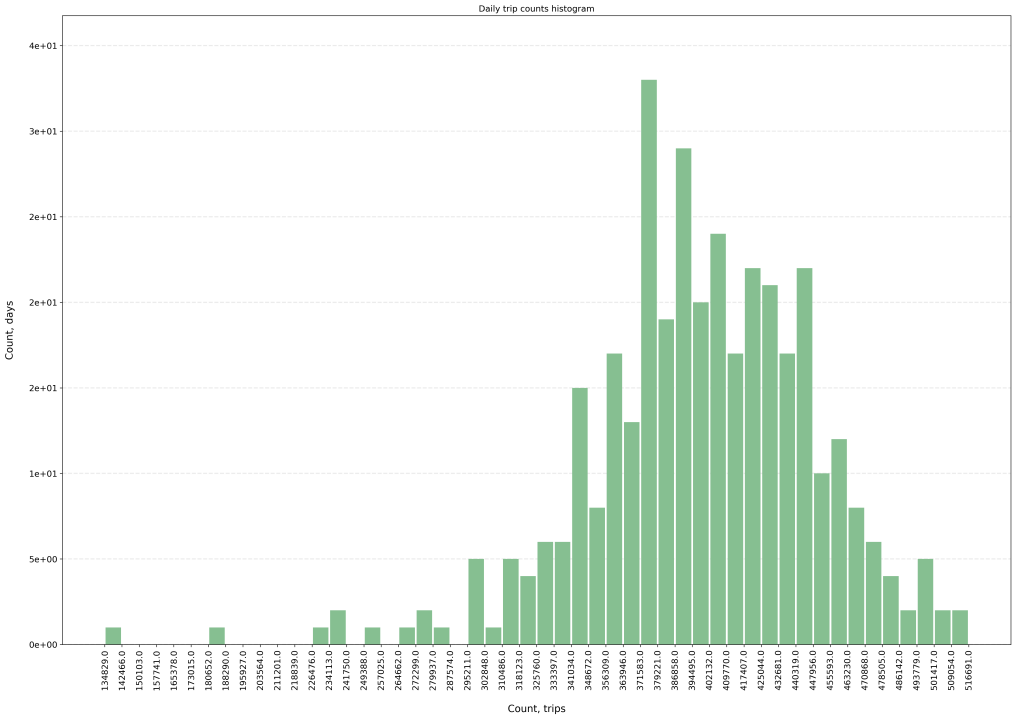

Figure 5: Daily trip counts distribution . New York City, 2015

[Sch15] Todd W. Schneider. *Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance*. 2015. URL: `https://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/` (visited on 10/30/2019).

[Tom18] TomTom. *New York in the Traffic Index*. 2018. URL: `https://www.tomtom.com/en_gb/traffic-index/new-york-traffic` (visited on 10/30/2019).

[Vel19] Jovan Veljanoski. *How to analyse 100 GB of data on your laptop with Python*. 2019. URL: `https://towardsdatascience.com/how-to-analyse-100s-of-gbs-of-data-on-your-laptop-with-python-f83363dda94` (visited on 10/30/2019).

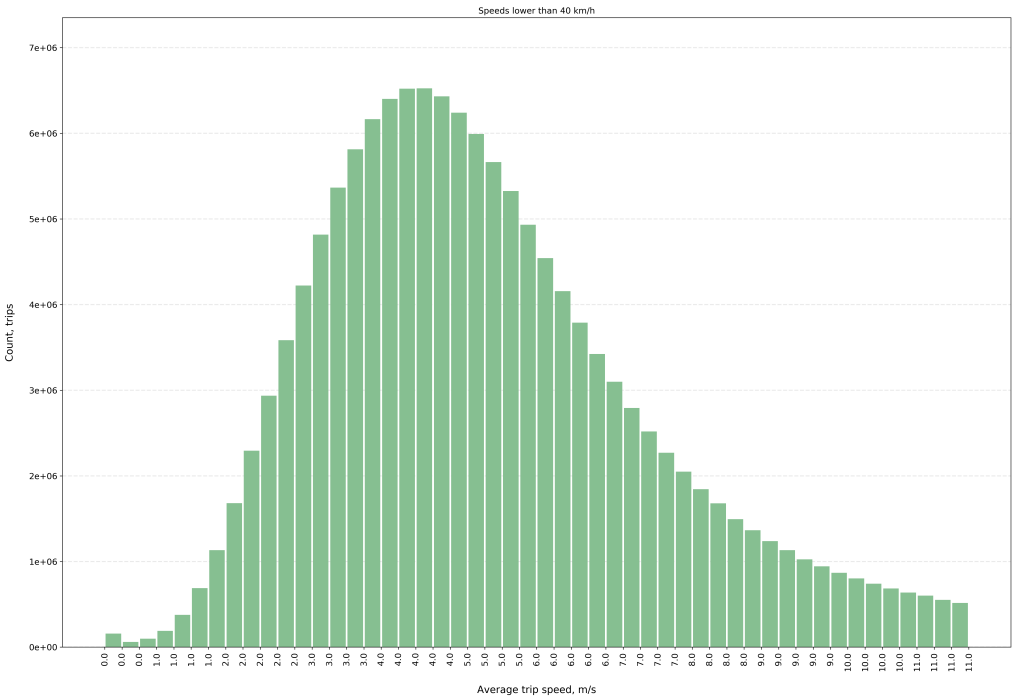

Figure 6: Distribution of average trip speeds lower than 40 km/s. New York City, 2015

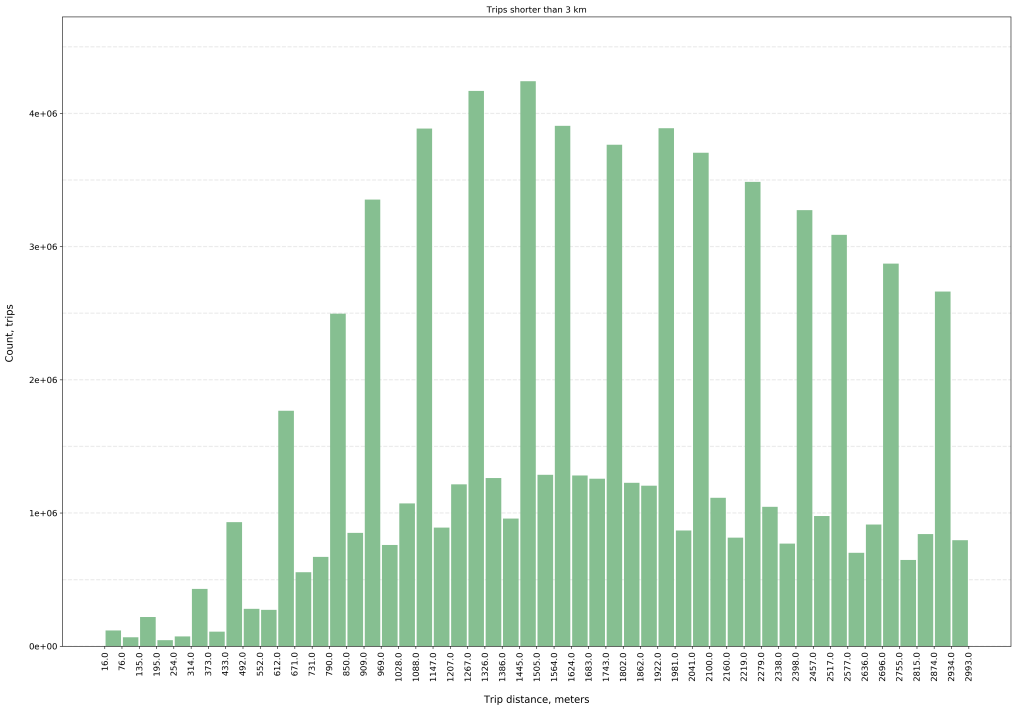

Figure 7: Distribution of trip lengths below 3 km. New York City, 2015

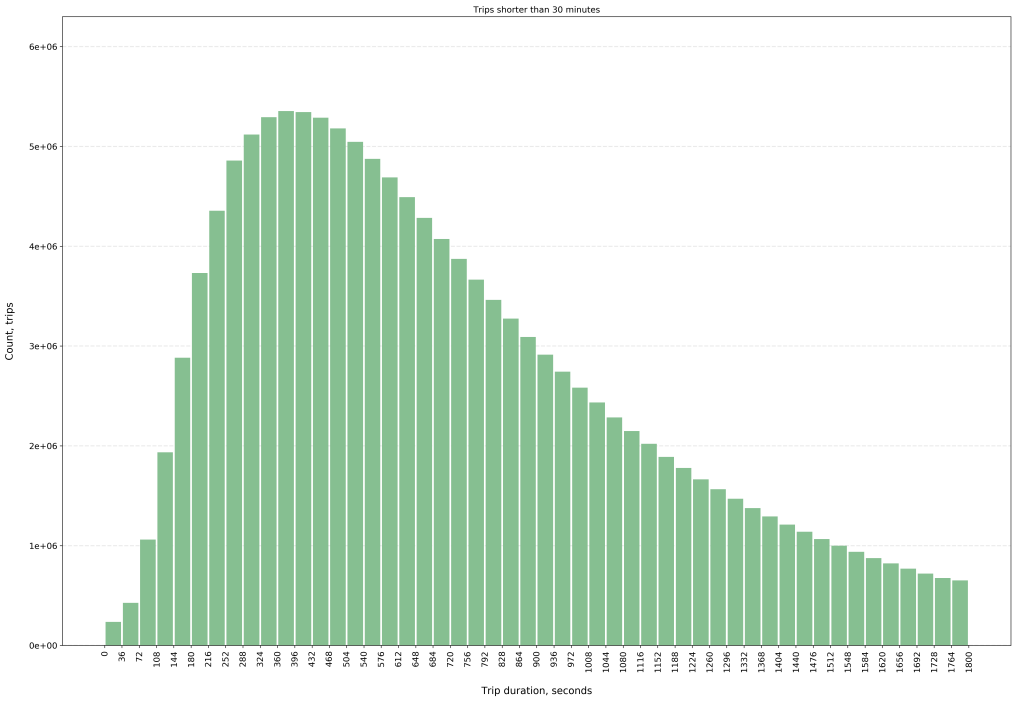

Figure 8: Distribution of trip durations below 30 minutes. New York City, 2015

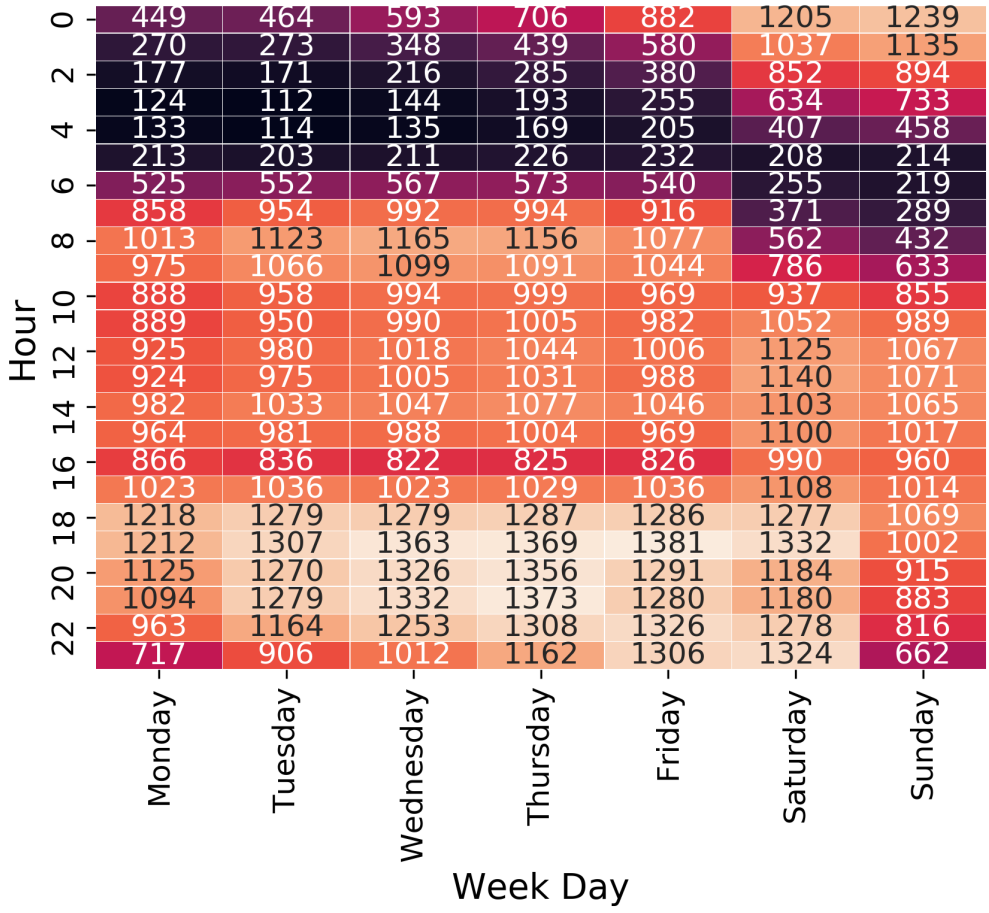

Figure 9: Yellow taxi trip counts by week days and time. New York City, 2015

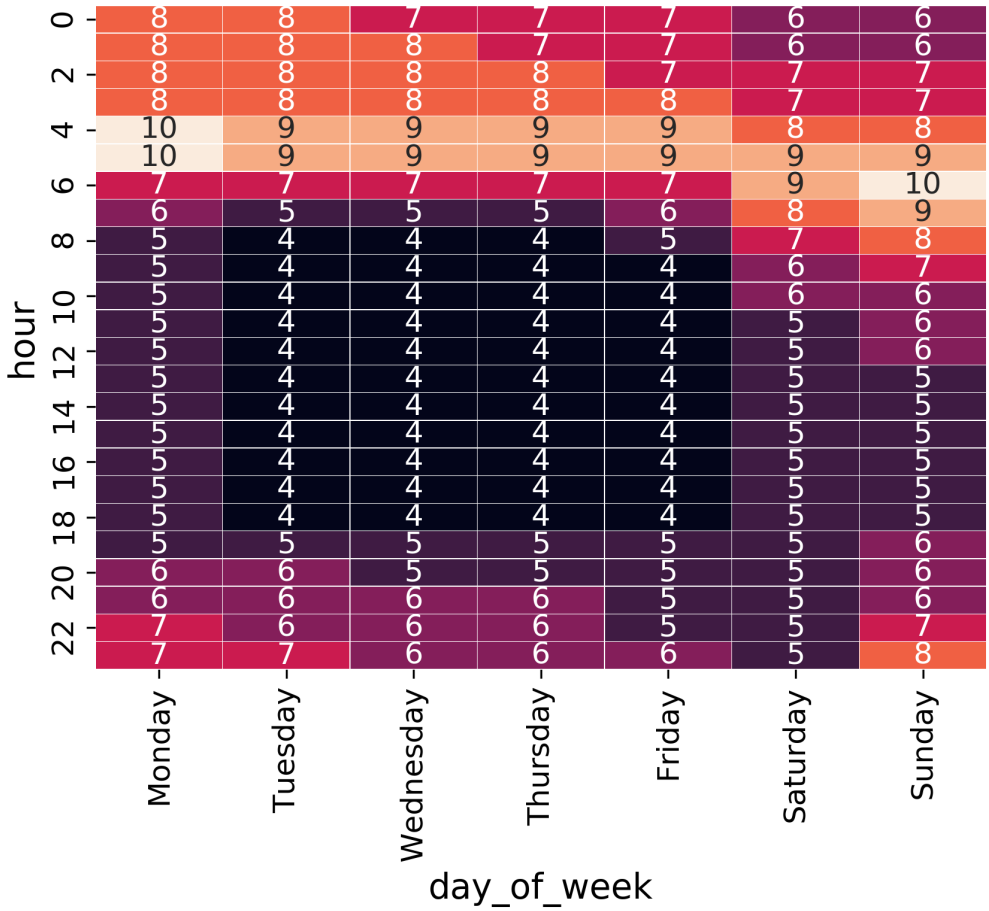

Figure 10: Yellow taxi computed average trip speeds by week days and time. New York City, 2015