Introduction

To prepare analysis for New York Taxis, I have been using public data set from BigQuery (bigquery-public-data.new_york).
To understand available data better, I did research on NY Taxis.

In New York City, taxicabs come in two varieties: yellow and green; they are widely recognizable symbols of the city.
Taxis painted yellow (medallion taxis) are able to pick up passengers anywhere in the five boroughs.
Those painted apple green (street hail livery vehicles, commonly known as "boro taxis"), which began to appear in August 2013, are allowed to pick up passengers in Upper Manhattan, the Bronx, Brooklyn, Queens (excluding LaGuardia Airport and John F. Kennedy International Airport), and Staten Island.
It also oversees over 40,000 other for-hire vehicles, including "black cars", commuter vans and ambulettes.
(Source: https://en.wikipedia.org/wiki/Taxicabs_of_New_York_City)

In each trip record dataset, one row represents a single trip made by a TLC-licensed vehicle.
Yellow taxis are traditionally hailed by signaling to a driver who is on duty and seeking a passenger (street hail), but now they may also be hailed using an e-hail app like Curb or Arro. Yellow taxis are the only vehicles permitted to respond to a street hail from a passenger in all five boroughs.
Green taxis, also known as boro taxis and street-hail liveries, were introduced in August of 2013 to improve taxi service and availability in the boroughs. Green taxis may respond to street hails, but only in the areas indicated in green on the map (i.e. above W 110 St/E 96th St in Manhattan and in the boroughs).
FHV data includes trip data from high-volume for-hire vehicle bases (bases for companies dispatching 10,000+ trip per day, meaning Uber, Lyft, Via, and Juno), community livery bases, luxury limousine bases, and black car bases.
(Source: https://www1.nyc.gov/assets/tlc/downloads/pdf/trip_record_user_guide.pdf)

I have decided to explore yellow taxis for 2016 year and have been using 'tlc_yellow_trips_2016' table.

Analysis consisted of the following steps:
1. Descriptive analysis of categorical data (vendors, passenger count, rate codes, payment types, store and forward trips)
2. Descriptive analysis of some numerical data (trip distances, trip time, total amount) and total numbers.
3. Exploratory analysis to answer questions:
    Q1: 'How does number of trips, average trip time, average trip cost and average trip distance changes over time of the day?'
    Q2: 'How does average trip price per minute and per mile changes over time of the day? When is it more profitable to have shorter rides?'
    Q3: 'Does this statistic differ for each weekday, is it different on weekends?'
    Q4: 'How does statistic changes throughout a year by month?'
    Q5: 'How does daily statistic changes throughout a year?'
    Q6: 'Does public holidays and other special days affect NY taxis?'

Please refer to additional files:
- Taxi Queries.ipynb
- 2016_Events.xlsx

Descriptive analysis of categorical data and total numbers for numerical data

Number of trips	Total trips distance, miles	Total profit, USD
131,165,043	625,469,034.50	2,148,287,056.26

Total number of passengers	Total trip time, minutes	Total tips, USD
217,355,302	1,485,148	236,029,188.66

Vendor Name	Record Count
VeriFone Inc.	70,074,613
Creative Mobile Technologies, LLC	61,084,518
Other	5,912
Grand total	**131,165,043**

Rate Code	Record Count
Standard rate	127,524,384
JFK	2,886,607
Negotiated fare	420,170
Newark	260,694
Nassau or Weatchester	61,004
Other	10,929
Group ride	1,255
Grand total	**131,165,043**

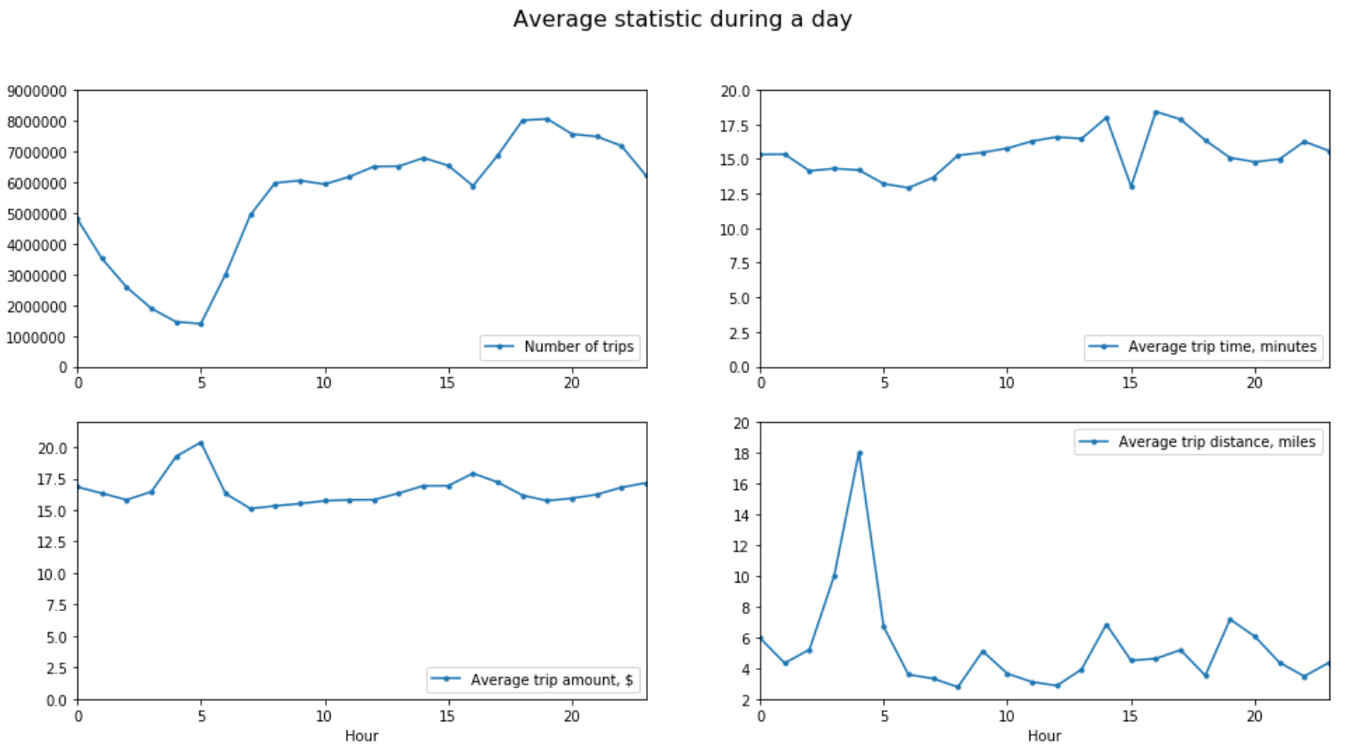
Store and Forward Trip	Record Count
No	130,417,250
Yes	747,793
Grand total	**131,165,043**

Passenger count	Record Count
1	92,987,719
2	19,038,307
5	6,773,026
3	5,456,807
6	4,234,423
4	2,660,369
0	13,454
7	361
8	316
9	261
Grand total	**131,165,043**

Payment	Record Count
Credit card	86,187,728
Cash	44,241,313
No charge	554,082
Dispute	181,908
Unknown	12
Grand total	**131,165,043**

Exploratory analysis

How does number of trips, average trips time, average trips cost and average trips distance changes over time of the day? Can we see high and low demand during a day?



Number of trips has pick around 18 - 19 hours, with period 17 - 23 hours being the busiest over the day with 8M+ trips. Nighttime demand gradually drops almost 5 times from 5M trips at 0 hours to 1M+ trips at 5 hours, while lesser people stay awake. As more people wake up and travel to work, morning demand climbs ~6 times over period of 5 to 8 hours to around 6M trips. Interesting, that number of trips has drop at 16 hours, will try to investigate it later.

Average trip time varies from 13 minutes to 18.5 minutes with picks at 14, 16 and 17 hours. Lowest trip time happens at 6 and 15 hours. Overall, variation of trip time is not too big, but dip at 15 hours is obvious and need further investigation.

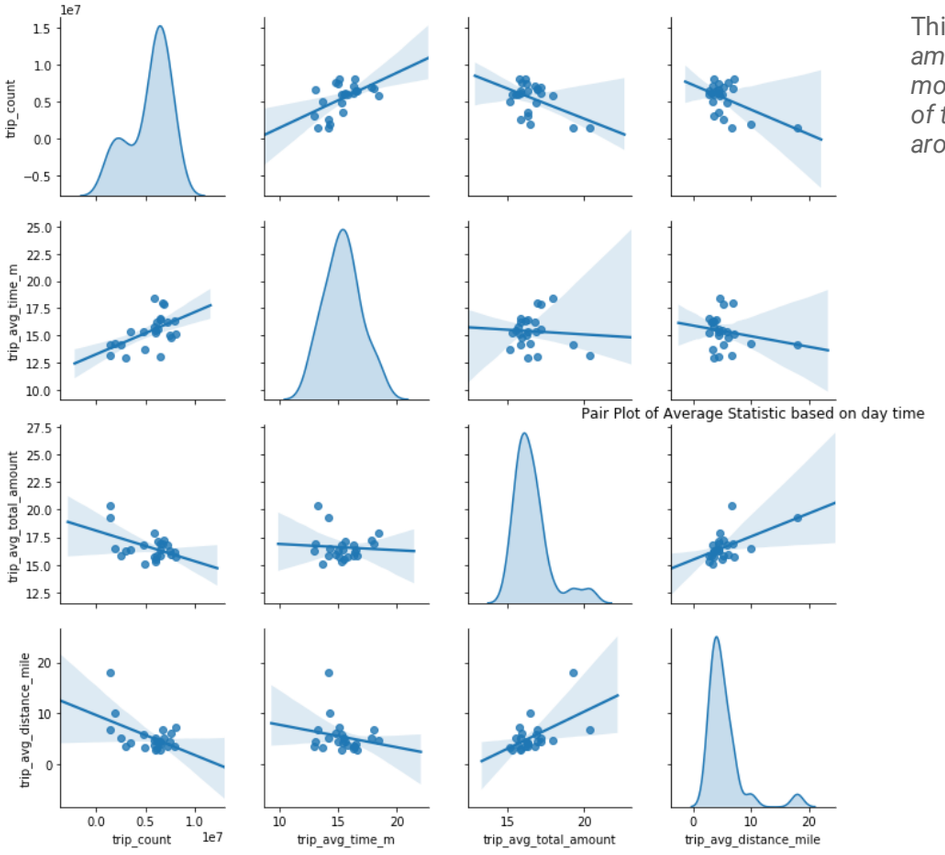
Average trip amount varies from ~USD15 to ~USD21 with peak at 5 hours (~USD21) and lesser peaks at 16 hours (~USD18) and 23 hours (~USD17.5). Low average trip amount happens at 2 hours (~USD16), 7 hours (~USD15) and 19 hours (~USD15.5). Pattern of average trip amount chart looks as mirror image to number of trips chart pattern. We will check correlation later.

Average trip distance varies from ~2.5 miles to ~18 miles with peak at 4 hours (~18 miles) and local peaks at 0 hours (~6 miles), 9 hours (~5 miles), 14 hours (~7 miles), 19 hours (~7.5 miles). Shortest trips are at 8 hours (~2.5 miles), 12 hours (~2.5 miles), 18 hours (~3 miles), 22 hours (~3 miles). There is an interesting pattern of 5 points drop starting from 4 hours in the morning: trip distance is highest at first point and decreases over next 4 points. It repeats throughout a day but not applicable during nighttime.

Summary.
People of New York use yellow taxis mostly from 8 to 23 hours, with period 17 to 23 hours being the busiest over the day. Lowest taxi usage is during nighttime from 1 to 6 hours.
Average trip time varies from 13 to 18.5 minutes throughout a day with unusual dip at 15 hours (~12.5 minute).
Average trip amount varies from ~USD15 to ~USD21 and possibly has negative correlation with number of trips.
Average trip distance varies from ~2.5 miles to ~18 miles with most of the trips happens within 2.5 - 7.5 miles (for comparison, Manhattan Island is 13.4 miles long and 2.3 miles wide). There is unusual peak at 4 hours (~18 miles).

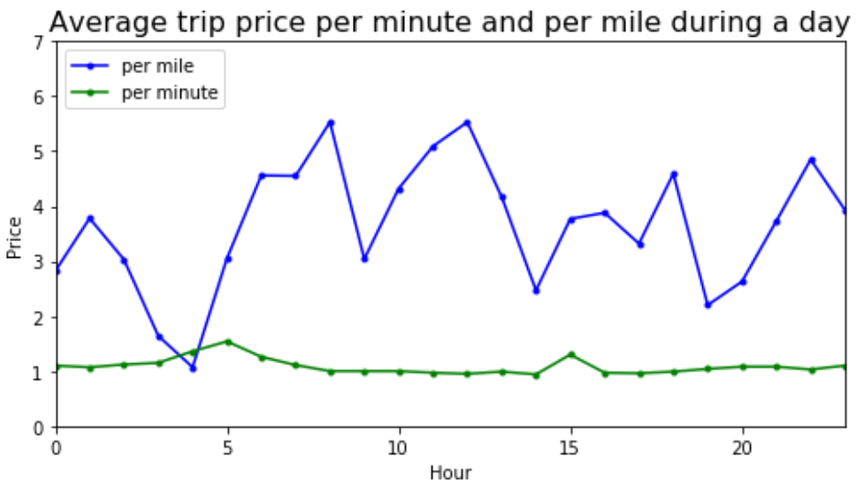
Exploratory analysis

To check possible correlations between our variables, let's look at pairwise relationship charts.



This shows, that *average trip amount does have negative moderate correlation with number of trips with data points gathered around one area.*

How does average trip price per minute and per mile changes over time of the day?
When is it more profitable for taxi drivers to have shorter rides (in minutes and in miles)?



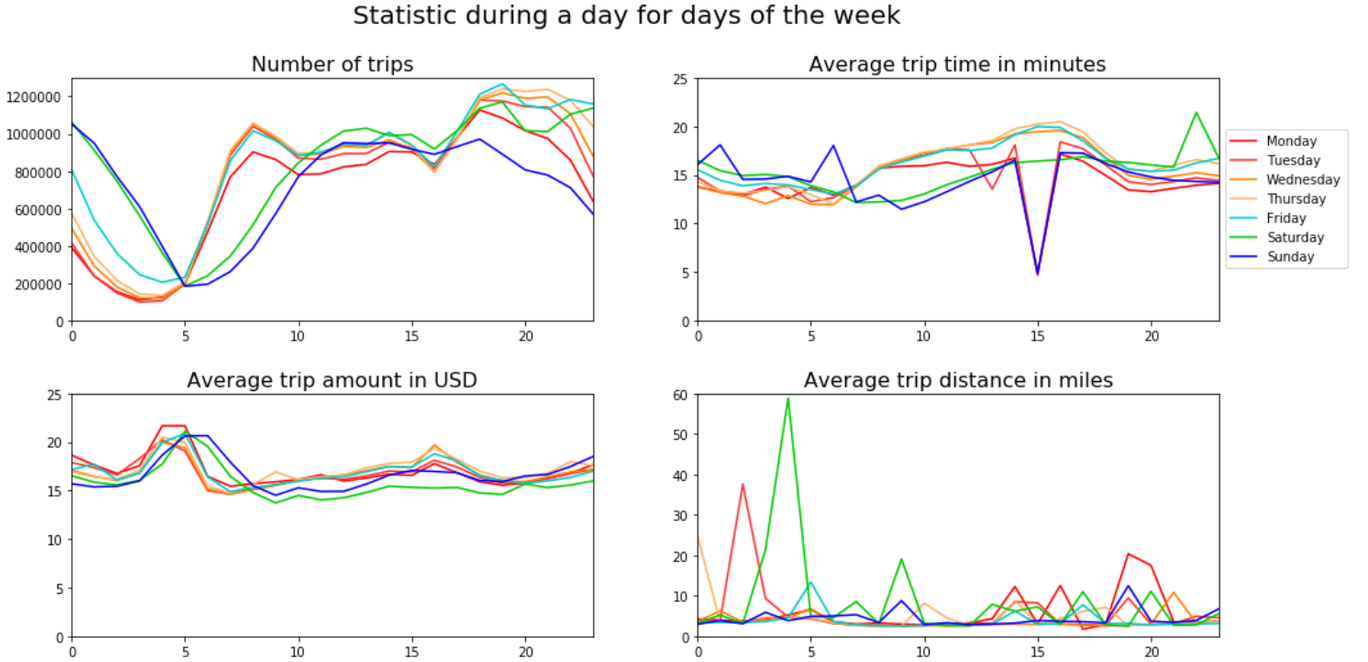
To answer, we will find average trip price per minute and per mile during a day.

It is pretty obvious that average trip price per minute is very stable at ~USD1, while average price per mile varies from ~USD1 to ~USD5.5.

It means that total amount of the trip doesn't depend much on traveling time and traffic jams, it depends more on the trip distance.
Higher average price per mile means that taxi drivers, who pick up passengers for shorter rides at 8, 12, 18, 22 hours will get higher income.

Exploratory analysis

Now let's split statistic by weekday and **check if weekdays and weekends follow the same pattern**.



Very obvious now, that weekends have separate pattern as well as Monday being different from the rest of the weekdays.

On weekends people do not use as much taxis from 5 hours to 10 hours in the morning, but use more taxis during nighttime from 0 to 5 hours compare to weekdays - probably they use yellow taxi to come back home after Friday night and Saturday night. Friday and Saturday 17 to 23 hours also have higher number of trips (because more people are going out at this time) with little dip around 20-21 hours when many people probably spend time over dinner or movie or elsewhere before heading to the next destination. Sunday evening has lowest number of trips, one of the possible reasons is that more people spend this time at home getting ready for a week.

Trip amount pattern is still the same, but weekends have peak at 5 hours compare to peak at 4 hours for weekdays.

Number of trips and average trip amount have dip and peak respectively at 16 hours over weekdays, which means there are lesser taxis available. This is true, because 16 to 17 hours is the time for cabs to change shifts.

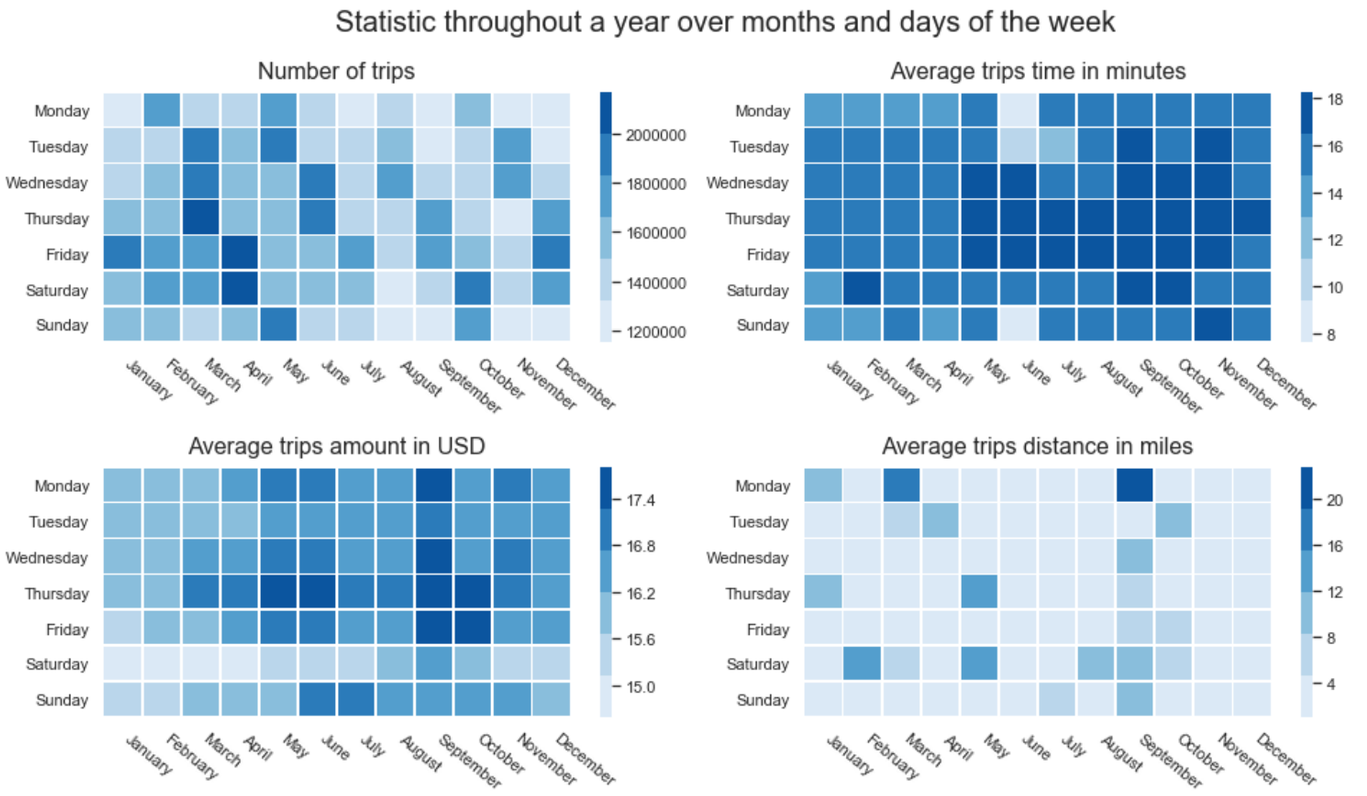
Average trip time in minutes for Sundays and Mondays has big dip at 15 hours to 5 minutes. Other charts do not show any anomalies around 15 hours, have to investigate reasons in future (maybe church events? poor data?). Average trip time on Saturdays at 22 hours has peak (~22 minutes) as probably more people coming back home by taxi rather than by public transport. Generally during weekends people have shorter trips.

Average distance in miles has peak at 4 hours on Saturday (~60miles), possible reason is that people travel far to other counties for stay over weekend. And on Monday around 2 hours they are coming back. But far trips usually also long trips, and average trip time doesn't show long trips at these times. On Monday evening 19 to 20 hours is another peak (~20 miles).

Summary.
There is obviously different patterns for weekends, Mondays and other weekdays. During weekdays when cabs change shift at 16-17 hours there are lesser yellow taxis available on the street, but demand is still there. People use more taxis on Saturday and Sunday night to come home after outing. (in future: check if pickup locations are more centralized and drop off locations are more remote for weekends nighttime)
People have shorter trips over weekends during day time and longer trips on Saturday's evenings. (in future: investigate average time dip at 15 hours)
From average distance split by weekday is not possible to draw any valuable insights - need to check data and find if there is relevance to some events.

Exploratory analysis

After we explore statistics during weekdays and during hours of day, let's look at the bigger picture.
Our question is "**How does statistic changes throughout a year by month?**"



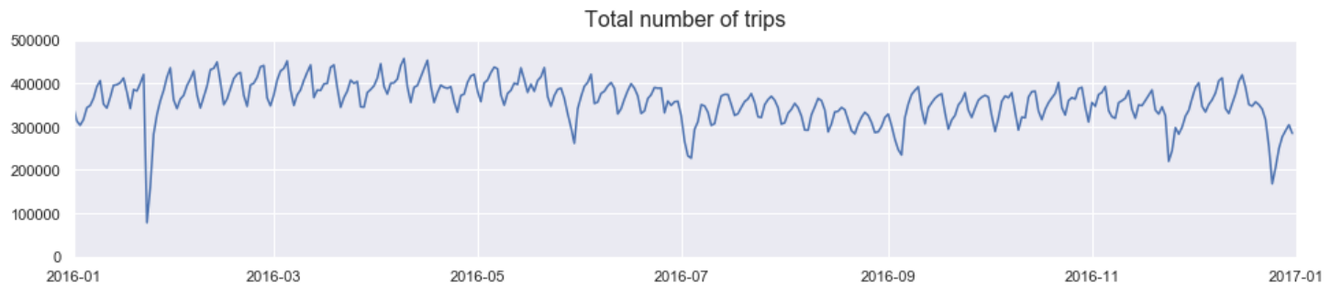
Generally there are more taxi trips during late winter (February) and spring (March, April). In March people took more taxis during weekdays, while in April there are more taxi trips during weekends and Fridays. There is obvious lesser demand during second half of the year. Probably it is relevant to New York's weather, have to investigate in future.

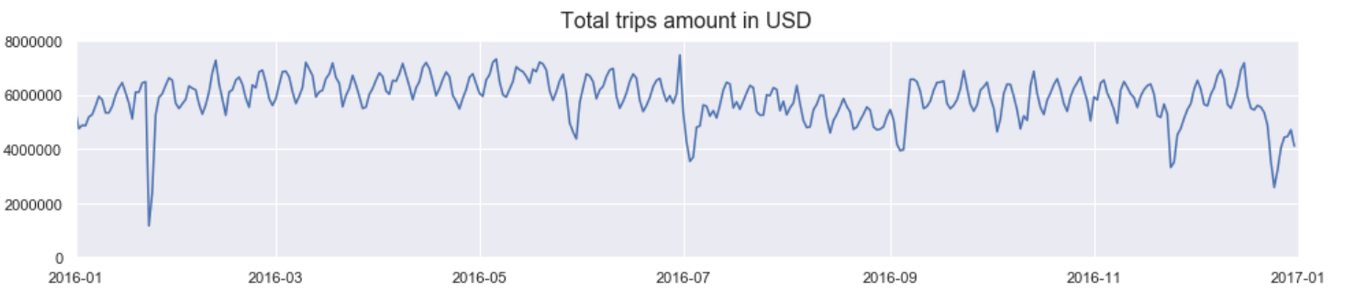
Even though second half of the year has lesser number of taxi trips, they tend to be longer in minutes and cost more. There are obvious outliers for average trip time in June (Mondays, Tuesdays and Sundays), have to investigate in future.
On Saturdays for all months trips generally costs lesser, have to investigate in future.

On Mondays in March and September there are longer trips, as well as for the rest of the days in September, have to investigate in future.

Exploratory analysis

How does daily total statistic changes throughout a year?
Are there particular patterns for total statistic over dates?





There are obvious patterns in total statistic over dates, mostly related to weeks.
What are those bigger dips and higher peaks?

To answer this question we look for details about public holidays (Fedreal Holidays) and other special days (observances, religious holidays, etc.) from https://www.nycbynatives.com/events/events_nyc.php

Exploratory analysis

How does daily total statistic changes throughout a year?
Are there particular patterns for total statistic over dates?

It is very clear now that public holidays affect New York taxis and there are lesser trips on public holidays compare to other days during that month. Most of other special days do not cause any interruption to taxi business. Apart from special days, there are few outlying points.

Biggest dip of total number of trips happen on 23 January 2016.
Quick google search explains it as "a crippling and historic blizzard".

Peak of total trips time happen on 13 February 2016. Quick google search doesn't show major events for this date.
Dips of total trips time have values below 0, which means we should check data for these dates.
Date of 23 January (lowest from positive values) already explained by blizzard.

Majority of peaks in total trips distances happen on special day (Hindu and Christian Holidays and Observances) - most probably people are travelling to the religious places of worship.
26 September 2016: The First 2016 Presidential Debate was held at Hofstra University, located in Hempstead, NY, which is ~32 miles away from Manhattan.
28 May 2016: 2016 New York Red Bulls season game, held in Harrison, New Jersey, which is ~14 miles away from Manhattan.

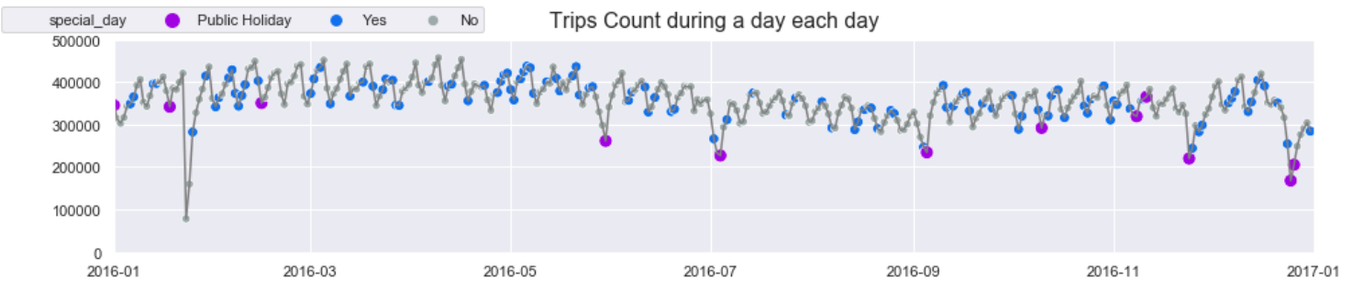
Biggest dip in total trips distance happen on Chinese New Year 8 February 2016 and has negative value. Have to check data for this date. The next dip is on 23 January 2016, when blizzard happen.
Lowest total trips amount is on 23 January and 24 January 2016, when blizzard happened.

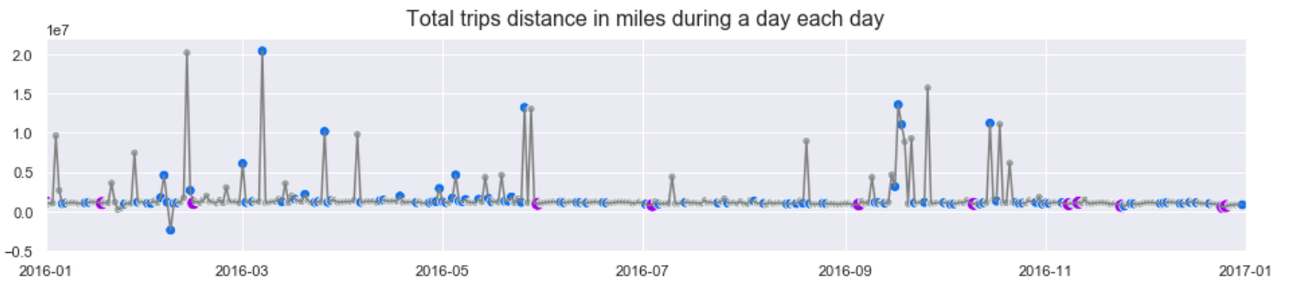
As a summary, *it is very obvious, that public holidays affect taxi business, but we also have seen, that bad weather and major events affect it too.*

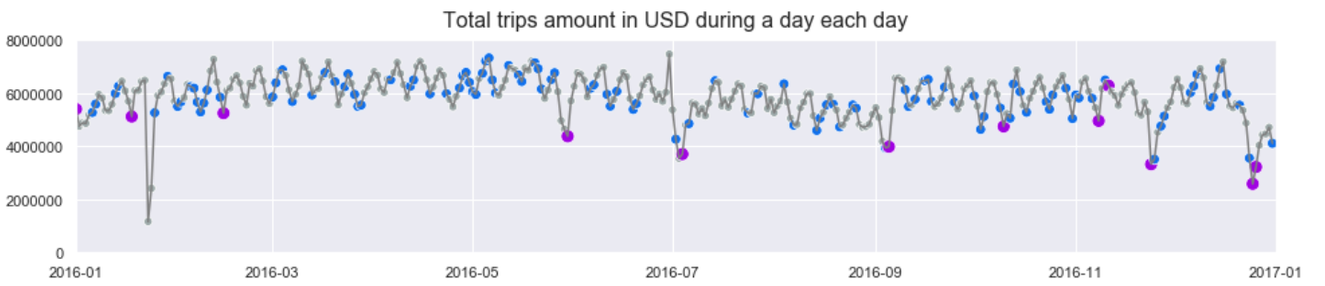
*Please see bigger size chart on the next page

Exploratory analysis

How does daily total statistic changes throughout a year?
Are there particular patterns for total statistic over dates?







Summary

How does number of trips, average trips time, average trips cost and average trips distance changes over time of the day? Can we see high and low demand during a day?

People of New York use yellow taxis mostly from 8 to 23 hours, with period 17 to 23 hours being the busiest over the day. Lowest taxi usage is during nighttime from 1 to 6 hours.
Average trip time varies from 13 to 18.5 minutes throughout a day with unusual dip at 15 hours (~12.5 minute).
Average trip amount varies from ~USD15 to ~USD21 and possibly has negative correlation with number of trips.
Average trip distance varies from ~2.5 miles to ~18 miles with most of the trips happens within 2.5 - 7.5 miles (for comparison, Manhattan Island is 13.4 miles long and 2.3 miles wide). There is unusual peak at 4 hours (~18 miles).

How does average trip price per minute and per mile changes over time of the day? When is it more profitable for taxi drivers to have shorter rides (in minutes and in miles)?
Total amount of the trip doesn't depend much on traveling time and traffic jams, it depends more on the trip distance.
Higher average price per mile means that taxi drivers who pick up passengers for shorter rides at 8, 12, 18, 22 hours will get higher income.

Does this statistic differ for each weekday, is it different on weekends?
There are obviously different patterns for weekends, Mondays and other weekdays.
During weekdays when cabs change shift at 16-17 hours there are lesser yellow taxis available on the street, but demand is still there. People use more taxis on Saturday and Sunday night to come home after outing.
People have shorter trips over weekends during day time and longer trips on Saturday's evenings.
From average distance split by weekday is not possible to draw any valuable insights - need to check data and find if there is relevance to some events.

How does statistic changes throughout a year by month?
Generally there are more taxi trips during late winter (February) and spring (March, April). In March people took more taxis during weekdays, while in April there are more taxi trips during weekends and Fridays. There is obvious lesser demand during second half of the year. Probably it is relevant to New York's weather, have to investigate in future.
Even though second half of the year has lesser number of taxi trips, they tend to be longer in minutes and cost more.
There are obvious outliers for average trip time in June (Mondays, Tuesdays and Sundays), have to investigate in future.
On Saturdays for all months trips generally costs lesser, have to investigate in future.
On Mondays in March and September there are longer trips, as well as for the rest of the days in September, have to investigate in future.

How does daily statistic changes throughout a year? Does public holidays and other special days affect NY taxis?
It is very obvious, that public holidays affect taxi business, but we also have seen, that bad weather and major events affect it too.