There are two competing theories on how cab drivers operate. The first theory is with regards to the intertemporal substitution hypothesis. Cab drivers will work longer on days in which they are earning a higher hourly wage. This is to maximize their time split between working and leisure. A cab driver working longer on days they are earning more money will allow them to not work (and spend their leisure time) on days that they are not earning as much. The second theory is with regards to behavioral economics. Cab drivers drive until they hit some predetermined quota. On a day that they are earning a higher hourly wage, they will reach their quota earlier and therefore work fewer hours.

## Dataset

The city of Chicago released a data set of what is believed to be the majority of taxi rides taken from 2013 to 2016. This includes approximately 105 million unique taxi rides. The data includes:

* Unique trip identifier
* Unique taxi identifier
* Trip start and end times – rounded to the nearest 15 minute interval for anonymity
* Trip duration in seconds
* Trip distance in miles
* Fare amounts broken down by base fare, tips, tolls, and extras
* Payment type
* Taxi company
* Approximate geolocation data – to protect anonymity

Limitation of the dataset include:

* Imperfect data collection
  + Geolocation
  + Trip duration and trip distance sometimes have zero values
  + Fare amount
    - Sometimes zero values
    - Sometimes have highly implausible values – such as in the thousands of dollars
* Taxis are identified by ID – there is no way to confidently identify if there are one or multiple people driving a certain taxi, or if the same people/persons are consistently driving the same taxi
* Cannot answer question such as what is the actual demand for taxi rides in Chicago

## Data Cleaning and Wrangling

Acquiring the Chicago taxi data set was straightforward – it is available to download directly from the city of Chicago’s website. Since it is a large download (40GB for the whole dataset), I added wget functionality to Git Bash and downloaded via that method instead of attempting to download the data inside a web browser.

Such a large dataset will not be able to fit into memory, so a PostgreSQL server was setup to house the data. The inspiration for this was taken from a previous analysis already done on the data (<http://toddwschneider.com/posts/chicago-taxi-data/>).

The raw data was first loaded into a temporary table. Unique Taxi IDs were generated and standardized for readability. Then the raw data was inserted into a working table with appropriate data types for each field.

This report is only focusing on 2016 data due to the size of the dataset. I created another table to house only 2016 taxi trips to speed up SQL queries. This amounted to 19 million taxi rides. Additionally, since I am focusing mainly on three variables – taxi ID, trip start time, and total trip fare, I created another table which housed only those variables to further speed up queries. From researching SQL query optimization, I found that SQL queries will load the whole row into memory regardless if only certain fields are being queried. This was done to reduce disk input/output.

As already stated, there are a number of missing or implausible fields in the dataset due to bad/incomplete data collection. Missing geolocation data, trip duration, and trip distance are not a factor in this analysis.

The main problem is with the total fare amount. As stated above, values can be zero or highly implausible values (such as in the thousands of dollars). To clean then, I performed these two steps:

* Trips with a total fare amount of zero were excluded
* Trips with a total fare amount greater than $200 were excluded
  + $200 was chosen as the exclusion point because that was approximately five standard deviations above the mean. There were a number of trips that had fare totals in the thousands. While possible, these trips are highly unlikely (and unexplainable) and do not represent fares that a cab driver could be expecting to earn on an average day.

## Methodology

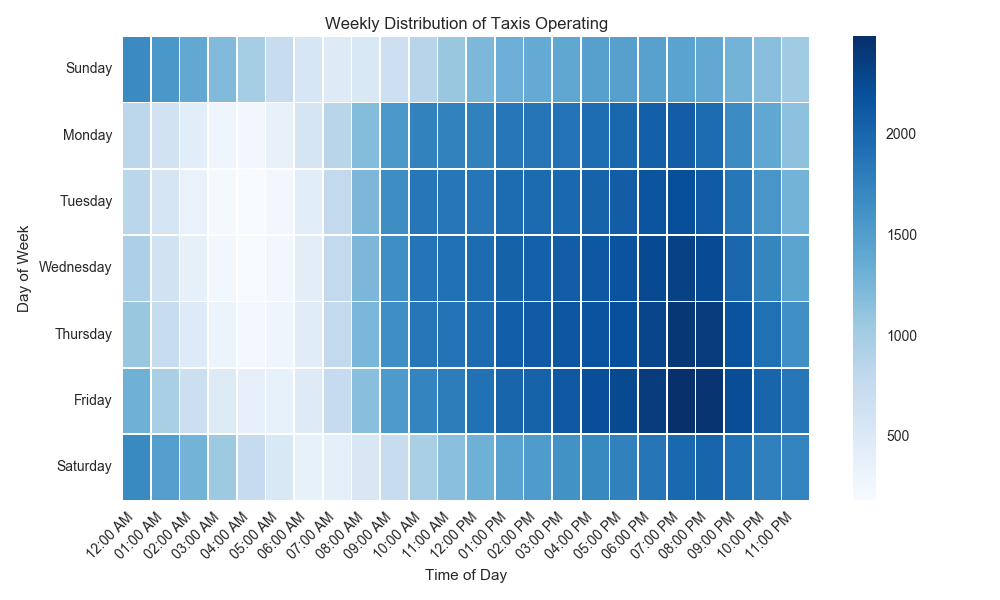
## Exploratory Analysis

I created heatmaps to explore the distribution of when taxis are operating as well as when taxis in demand.

Below is a heatmap of when taxis are most in demand – that is the distribution of trips taken. This heatmap confirms some obvious assumptions. Most trips are being taken weekday mornings between 8-9 AM (assumingly people going to work), weekday nights between 5-7 PM (assumingly people coming home from work), as well as Friday night into Saturday morning and Saturday night into Sunday morning between 8 PM and 3 AM (the weekend nightlife crowd).

## 

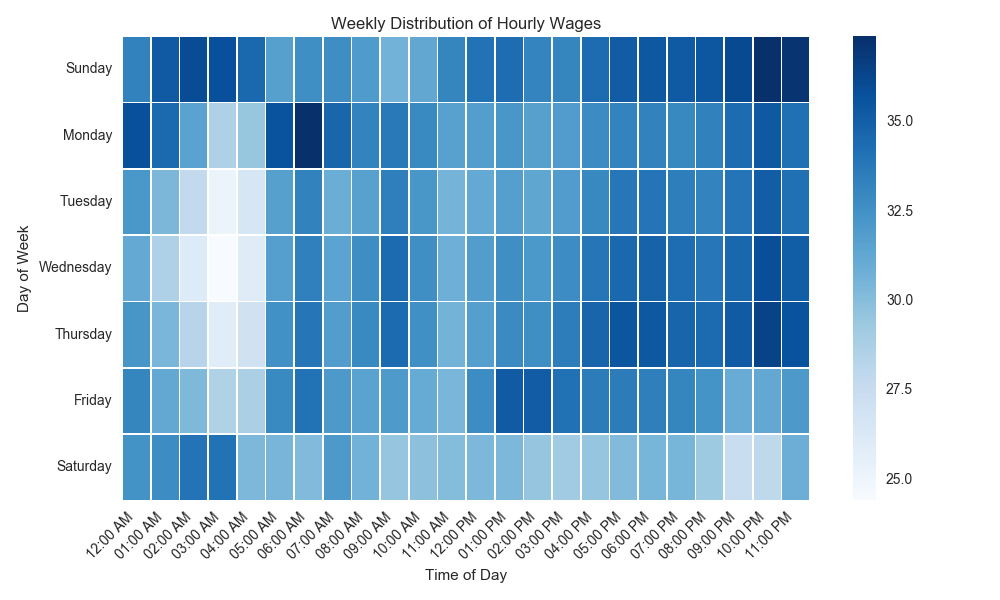
Below is a heatmap of when taxis are on the road. This heatmap closely resembles the above when taxis are in demand. Taxis are generally on the road to meet demand.



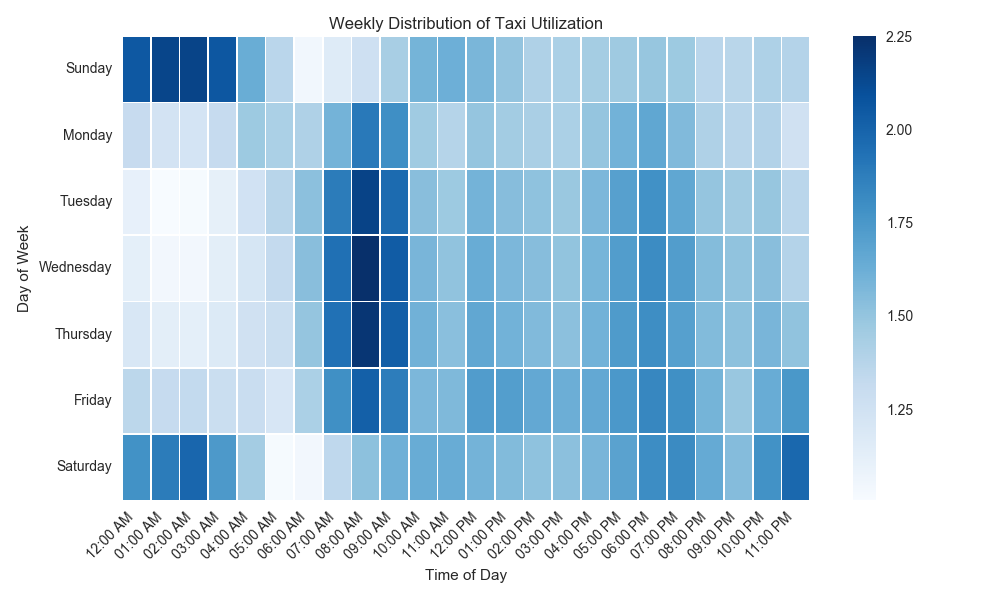
Below is a heatmap of hourly wages. This heatmap does not follow any discernable pattern. Some hotspots include weekday nights around 10 PM, Sunday morning between 1-4 AM, Sunday night from 5 PM to 1 AM, and Monday morning between 5-7 AM.

Some inferences could be drawn such as:

* Sunday late afternoon into the evening, early Monday morning, and Thursday evening could be seeing higher hourly wages due to increased trips to/from the airports for people who travel weekly for work
* Early Saturday and Sunday morning, 1-3 AM, could be seeing higher hourly wages due to people who traveled longer distances (such as from the suburbs) into the city for its nightlife going home

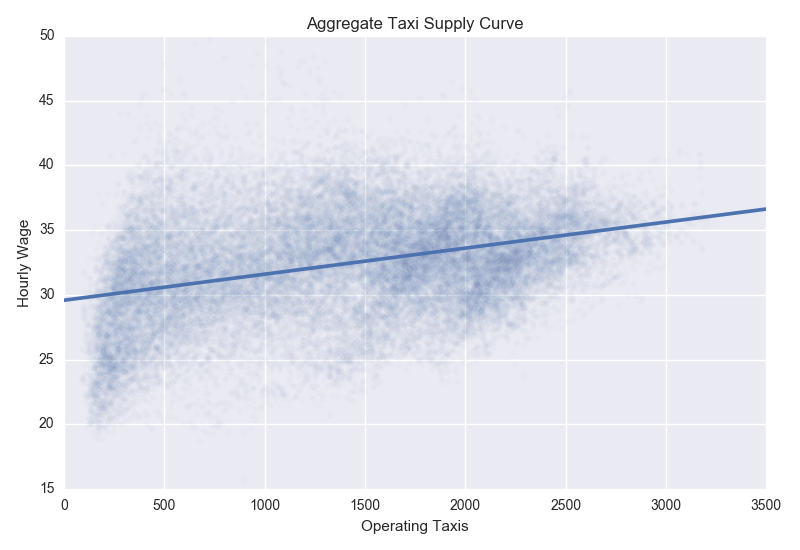


Below is a heatmap of taxi utilization. Utilization is a metric that tracks approximately how many trips each taxi is taking per hour – another way of looking at how much work each taxi is getting. It is calculated by counting the number of trips taken each hour and dividing by the number of active taxis on the road. Taxi utilization is highest weekday mornings between 7-9 AM and Friday night into Saturday morning and Saturday night into Sunday morning between 11 PM and 3 AM.



## Results

A simple linear regression found that on average, for every $1 increase in the hourly rate of taxis, 43 additional taxis are working.



## Limitations

Data was collected on a per taxi cab level, not a per driver level. There is no confirmable data on the actual drivers. Therefore, it is unknown is a cab is being used by one or multiple drivers – though inferences could be made if the data shows that certain cabs are making trips with no breaks. It is also unknown if the same cab is consistently used by the same driver or drivers on a daily basis.