Historical Hourly Weather Data Analysis

Pavithra Raghavan

4/20/2018

Weather data Analysis

Dataset Details

The dataset consists of six years (2012-2017) of hourly measurements of weather attributes for 36 cities. There is one dataset for each attribute of weather such as temperature, pressure, humidity, etc., and the details about the location of the 36 cities are also present.

Objective

The objective of this project is to analyze and find a pattern in the weather, as well as to analyze the taxi demand in New York.

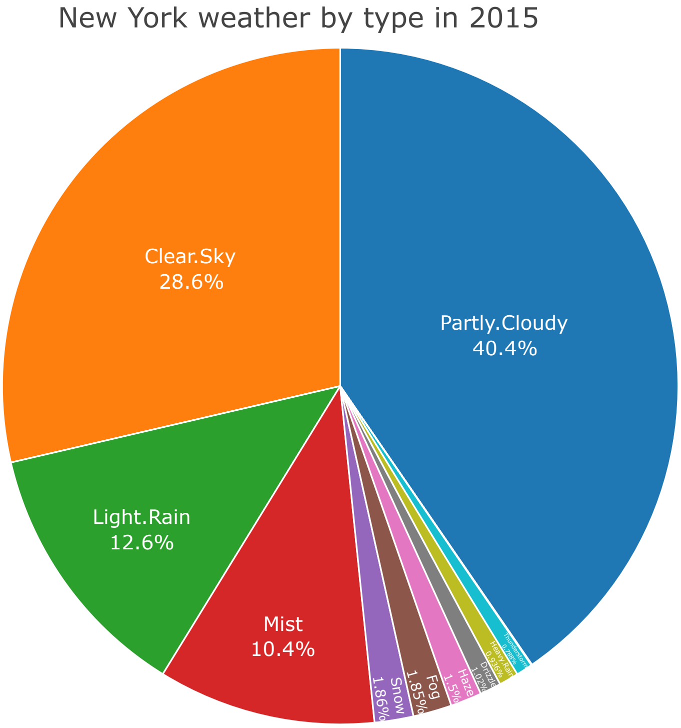
Preprocessing

The dataset consists of weather details of 36 cities. This project will focus on the weather in New York. A few alterations are done to the data for better analysis.

- Temperatures have been converted from Kelvin to °F.
- Weather descriptions in the dataset is very detailed, like "few clouds", "broken clouds", "scattered clouds". The description has been edited to include broader categories (the above 3 are edited to come under the category "patly.cloudy").
- A new column is added to portray the seasons.

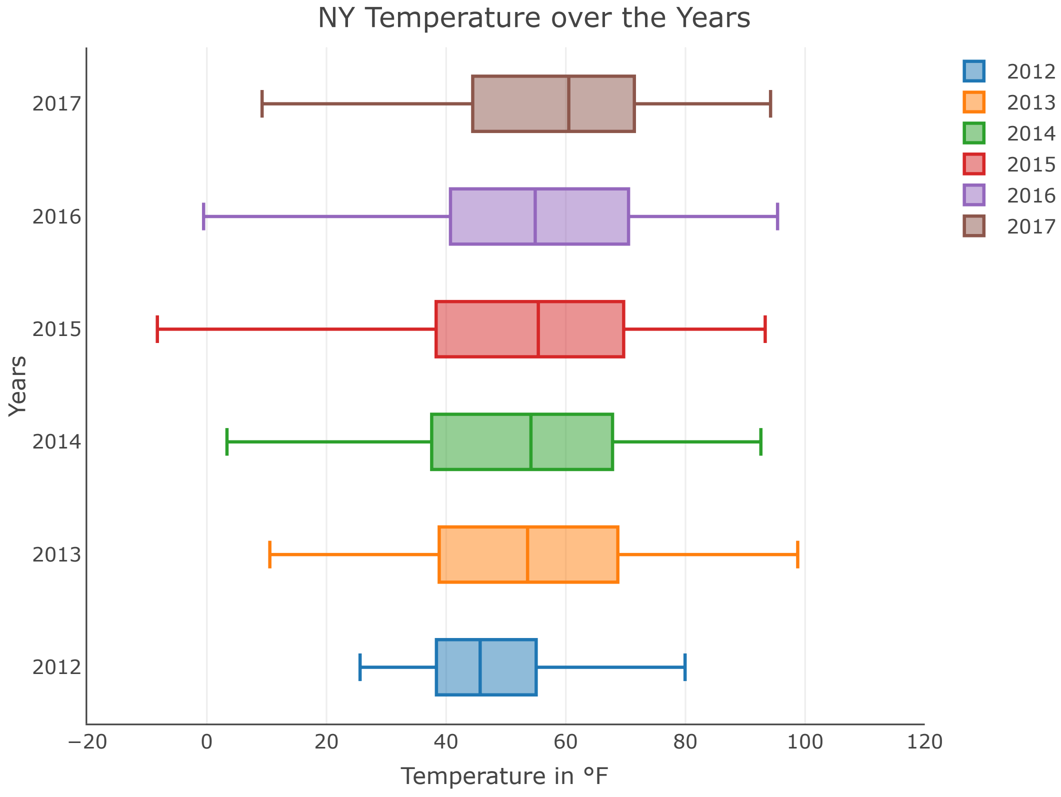
Weather Distribution in New York

Weather desscription dataset consists of the descriptions of weather in each day. The weather in NY ranges from a clear sky weather to drizzle to snow. Since this is a categorical data, frequency of each type of weather condition per year can be helpful to understand the general trend of weather in an area. The following bar plot shows the number of days each type of weather occured in NY in the year 2015. From the plot of 2015, we can see that most of the days in NY were clear sky weather or had a few clouds. There are only a few days of extreme weather like heavy rain or squalls.


New York weather by type in 2015

Temperature distribution in New York

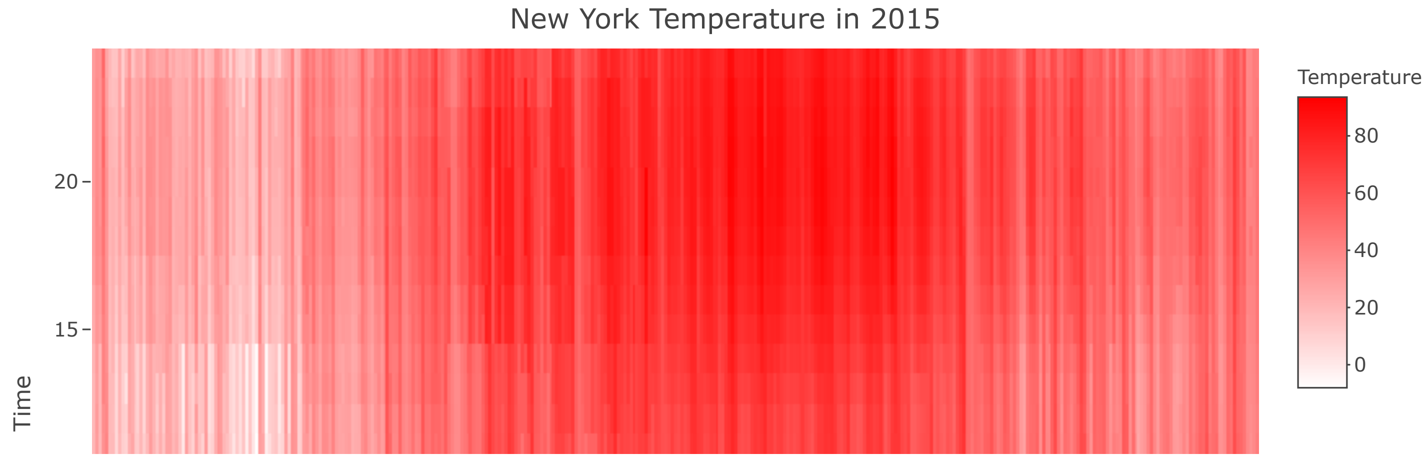
Temperature attribute in the dataset is numerical. For analysing the temperature in New York for various years, a box plot is used. With a box plot, the variation of temperature per year can be seen. The following is the box plot for NY temperatures in 6 years (2012-2017). It can be seen that the trend of median temperature is increasing.

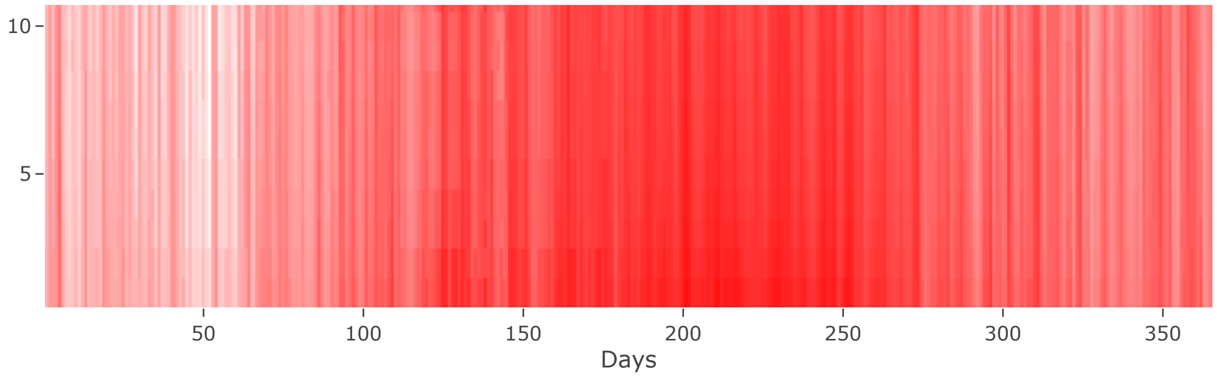


Following are the yearly mean temperatures in NY (in °F). Though no general trend for the mean temperature can be guaged, it has increased from 46.98°F in 2012 to 57.8°F in 2017.

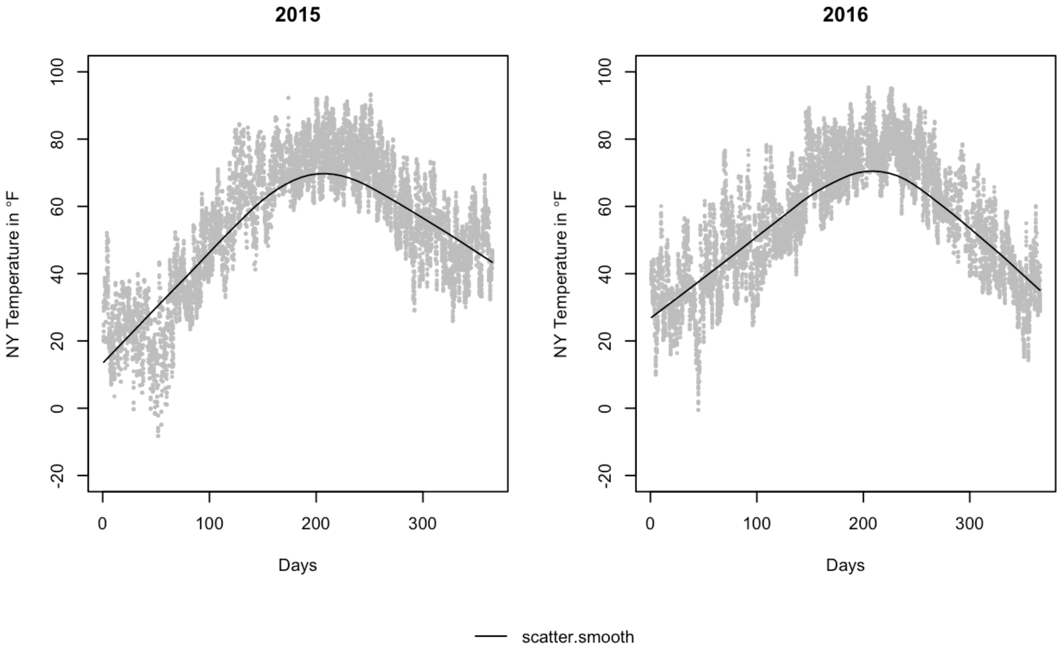
```
## Temperature NY 2012: mean =  46.98786
## Temperature NY 2013: mean =  53.94688
## Temperature NY 2014: mean =  52.46131
## Temperature NY 2015: mean =  53.22784
## Temperature NY 2016: mean =  55.22956
## Temperature NY 2017: mean =  57.82655
```

The dataset has hourly temperature measurements for 5 years. Temperature varies gradually over time, the following heatmap proves this point. There won't be much change from one hour to another, and from one day to another. But the change from one season to another would be great.





It would be helpful to see if there is a trend that temperature follows through the year, that is, we could plot the temperatures through-out a year, and find a pattern that best describes it. The following scatter plots show the temperatures in NY per day, for 2 years (2015, 2016). A scatter smooth trend line is shown which adds a smooth line to the scatter plot. It can be seen that the temperature follows a sinusoidal pattern.



Temerature Trend

Central Limit Theorem

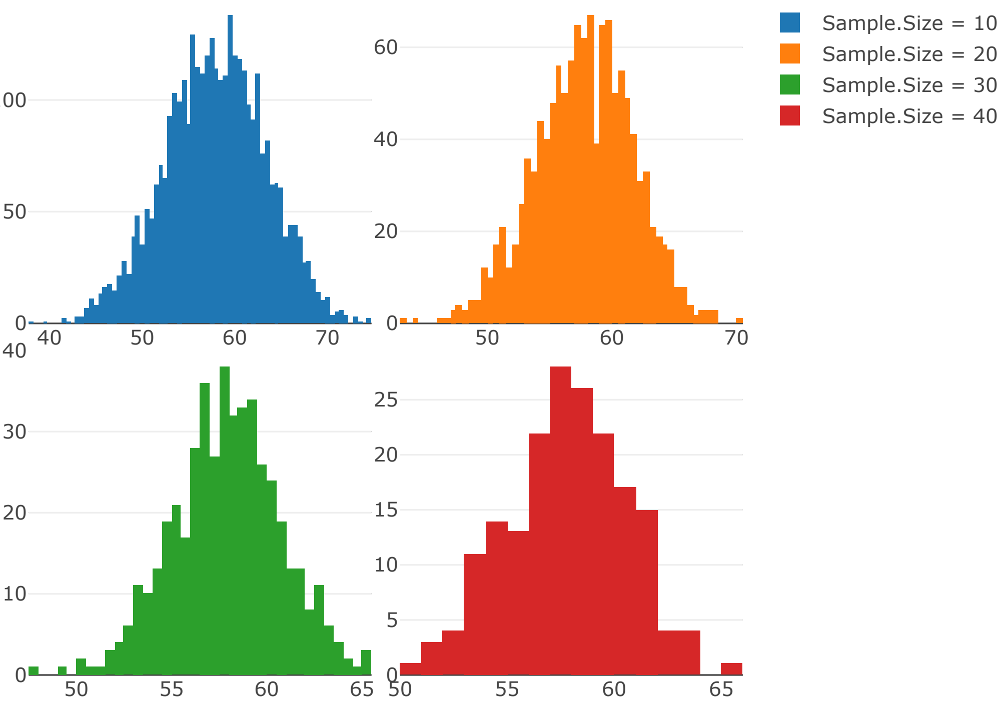
The central limit theorem states that the distribution of sample means has an approximate normal distribution, as long as all the samples have the same size and are independent. This is important because many statistical procedures require data to be approximately normal. When the data is not normal, Central Limit Theorem helps in creating a normally distributed attribute from the original data.

We know that the temperature distribution follows a sinusoidal pattern. But Central Limit Theorem can be used to obtain normally distribute sample means from this dataset. The following are the mean and standard deviations for 10k random samples of sizes 10, 20, 30 and 40:

```
## Mean NY Temperature in 2017:  57.82655
```

```
## Sample Size =  10  Mean =  57.8844  SD =  5.439273
## Sample Size =  20  Mean =  57.78559  SD =  3.955918
## Sample Size =  30  Mean =  57.88624  SD =  2.832131
## Sample Size =  40  Mean =  57.801  SD =  2.742092
```

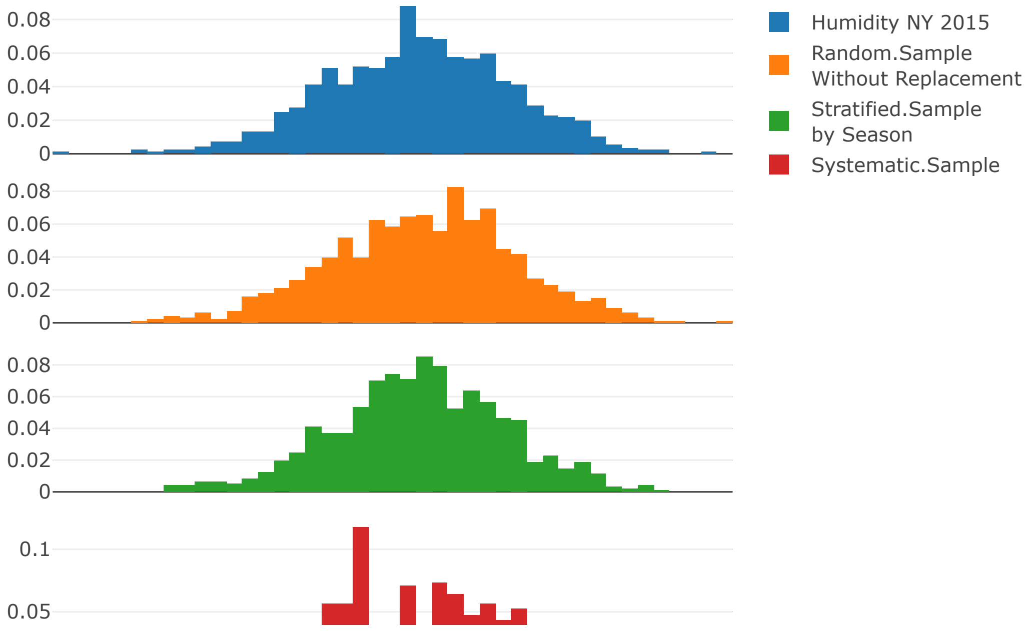
The following histograms show that the sample means follow normal distribution:

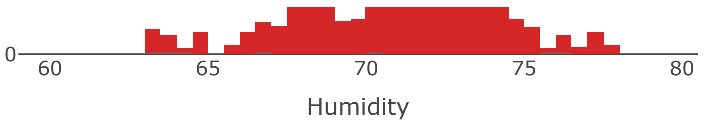


Sampling of Humidity attribute

It is not always feasible to analyse all the data from an entire population. So, a subset of data is taken and analysed instead, and then use statistics on that subset (sample) to draw conclusions about the entire population. There are many different sampling techniques like simple random sampling, stratified sampling and systematic sampling. In random sampling, every item in the population has equal chance of getting picked in a sample as every other item. Whereas stratified sampling can be used when there are categories in the data. In this, the same percentage of items are selected from each category in the population. In systematic sampling, each sample is selected such that its items are at a fixed interval. The starting item is selected at random. The fixed interval is calculated by dividing the total population by the sample size.

Humidity in air is another attribute that affects the weather in a place. Air has a maximum amount of humidity that it can hold. The dataset used for this project contains relative humididy, that is the amount of moisture in air with respect to the total amount of moisture that air can hold for a particular temperature. For this project, simple random samples without replacement, stratified sampling by Season and systematic sampling are used as the sampling methods on the humidity data.

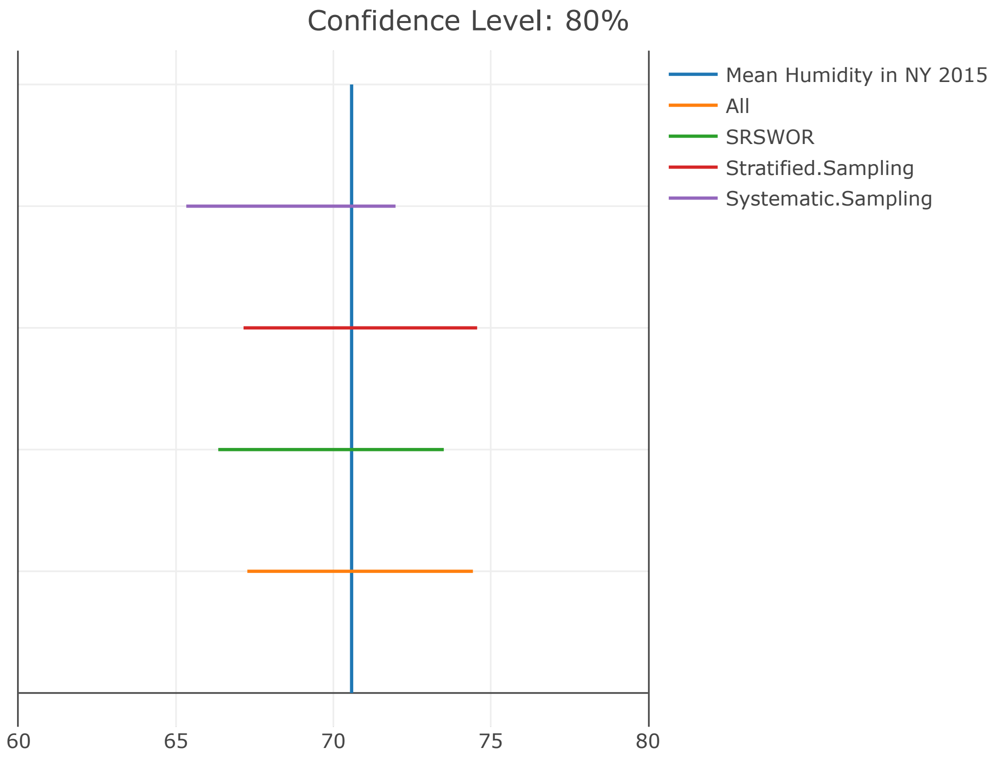


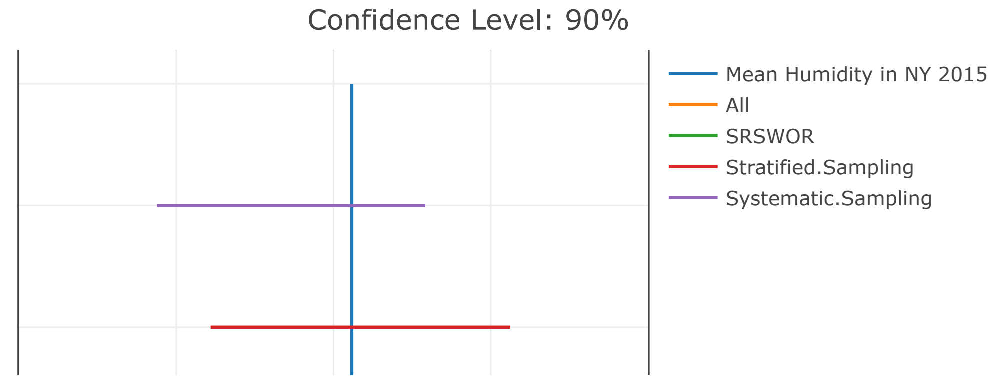


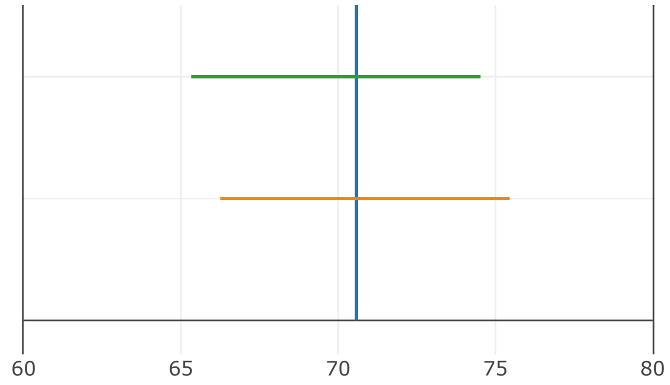
We have seen how the sample means of a dataset follows normal distribution. So, 95% of the sample means (including the mean of the entire population) lies within 2 standard deviations from the mean of mean samples. This means that, for a given sample mean, there is a 95% chance that the population mean lies within 2 standard deviations from it. We say that the confidence level is 95%. Following are the ranges (confidence interval CI) for each sampling method which have 80% and 90% chances of containing the mean of the entire population.

```
## Humidity NY 2015: mean =  70.85  and sd =  17.69188
## 80% Conf Level (alpha = 0.20), CI = 67.27 - 74.43
## 90% Conf Level (alpha = 0.10), CI = 66.25 - 75.45
## SRSWOR: mean =  69.925  and sd =  17.68584
## 80% Conf Level (alpha = 0.20), CI = 66.34 - 73.51
## 90% Conf Level (alpha = 0.10), CI = 65.33 - 74.52
## Stratified Sampling: mean =  70.85714  and sd =  18.7715
## 80% Conf Level (alpha = 0.20), CI = 67.15 - 74.57
## 90% Conf Level (alpha = 0.10), CI = 66.09 - 75.62
## Systematic Sampling: mean =  68.65  and sd =  16.41068
## 80% Conf Level (alpha = 0.20), CI = 65.32 - 71.98
## 90% Conf Level (alpha = 0.10), CI = 64.38 - 72.92
```

The follwing figures show the above ranges for each sample, with 80% and 90% confidence respectively. The vertical line represents the mean of the entire humidity data.







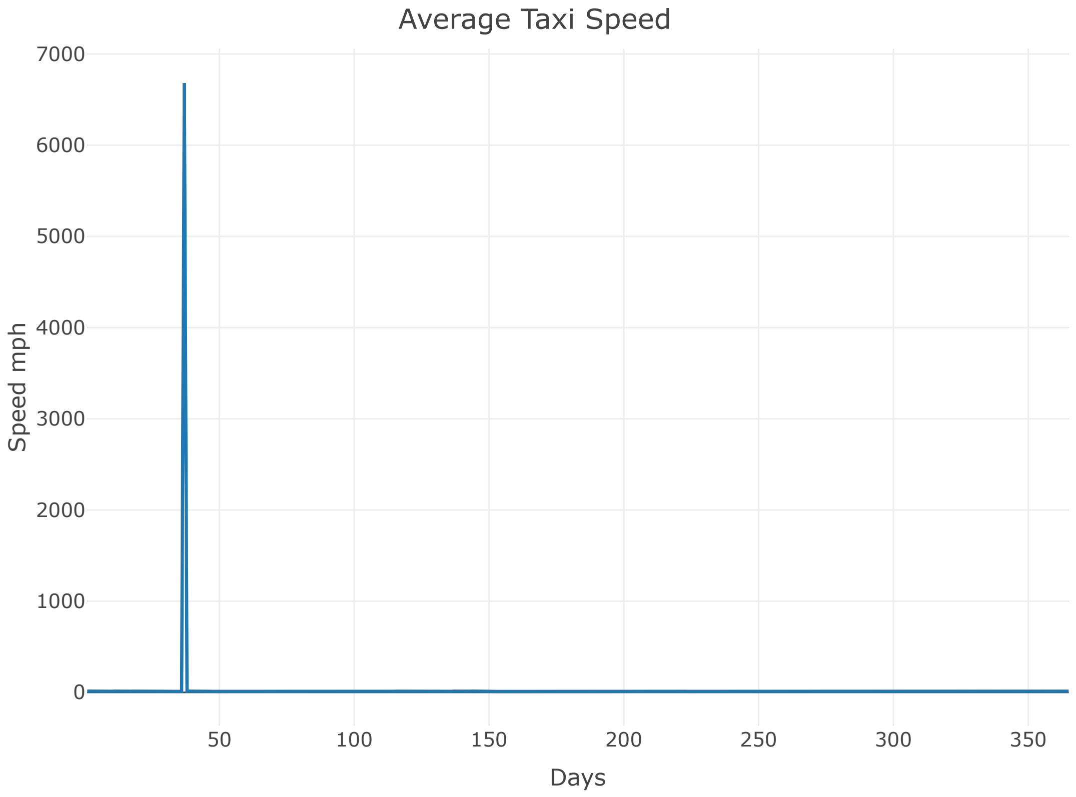
Effect of weather on Taxi Demand

Analysis of how weather affects taxi demand.

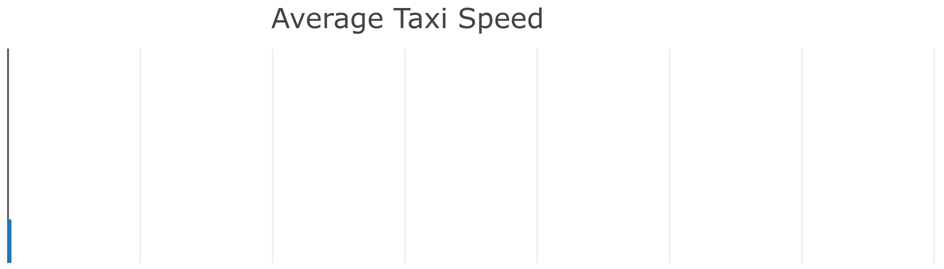
A dataset in https://www.kaggle.com/dhimananubhav/2015-nyc-taxi-trips-subset-12-million-rows/data (https://www.kaggle.com/dhimananubhav/2015-nyc-taxi-trips-subset-12-million-rows/data) will be used for this purpose.

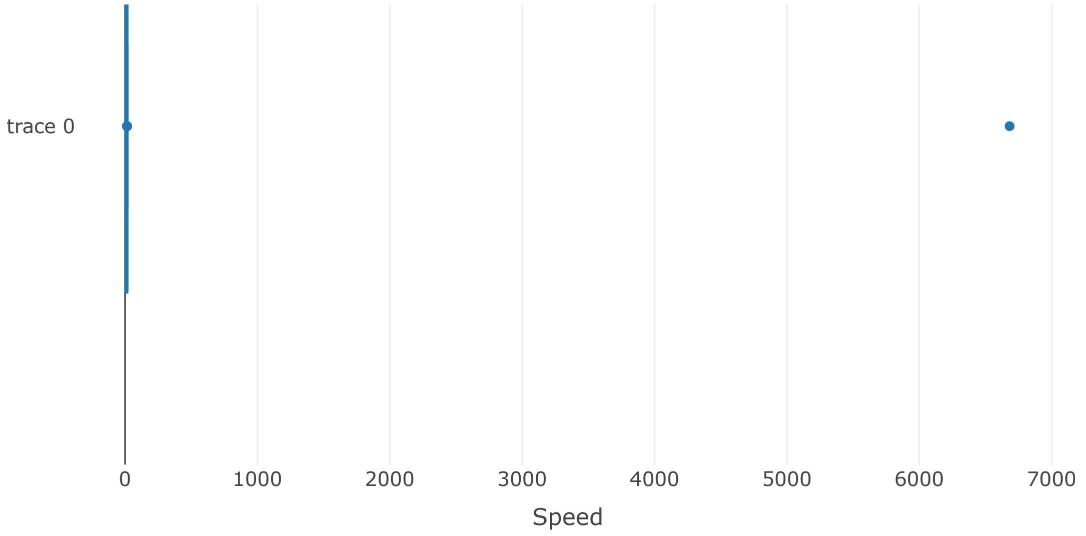
Data processing

The taxi data is not perfect. There are outliers which make the data skewed.



The boxplot below is skewed because of an outlier at 6681:





We can use the stats of boxplot for finding the outliers:

```
boxplot.stats(demand.per.day$speed)$out
```

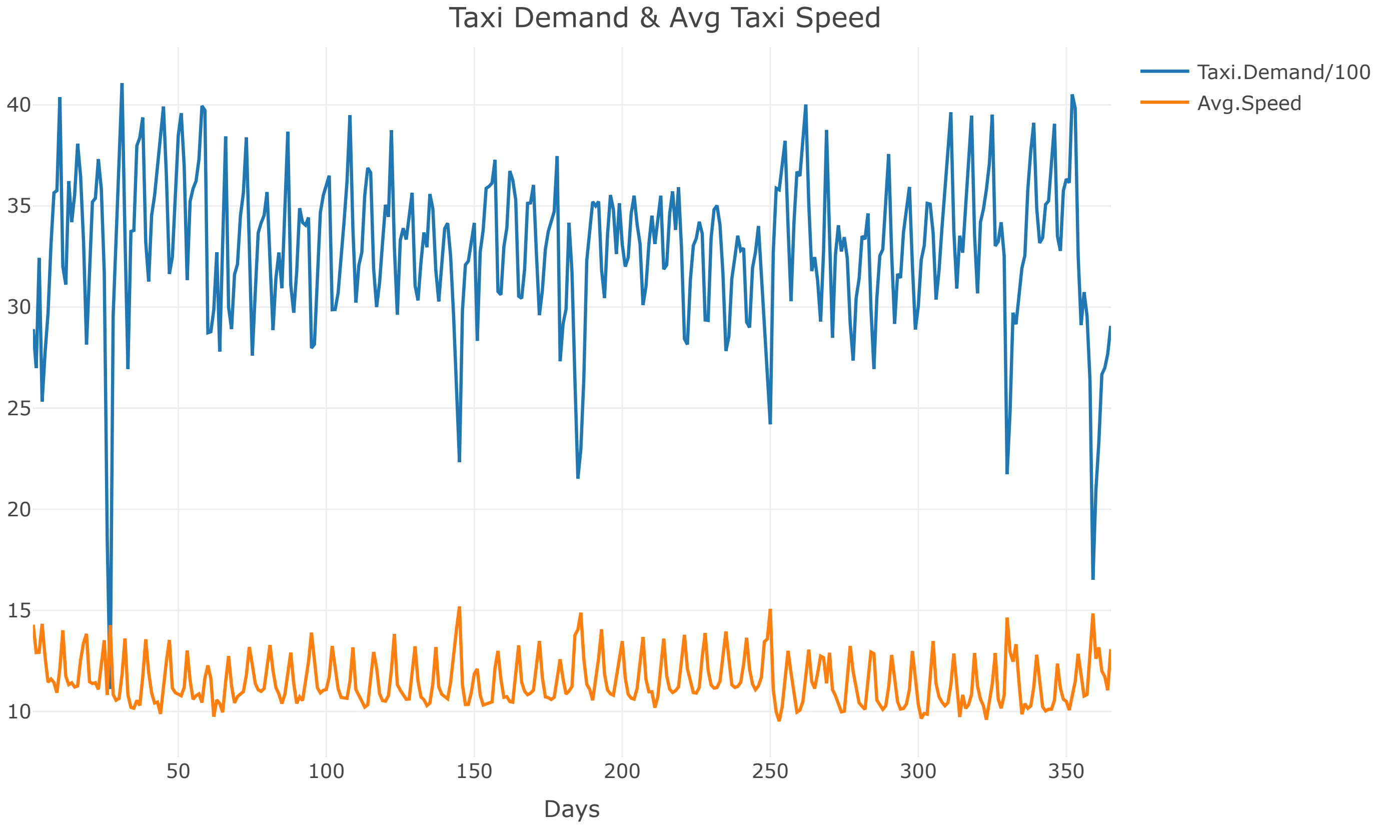
```
## [1] 6681.94487   15.19449   14.88484   15.07294   14.64162   14.84780
```

```
boxplot.stats(demand.per.day$speed, coef=2)$out
```

```
## [1] 6681.945
```

This speed of 6681mph is definitely not possible, must be a mistake or a typo. So we will analyze the data without that row which contains this very high speed.

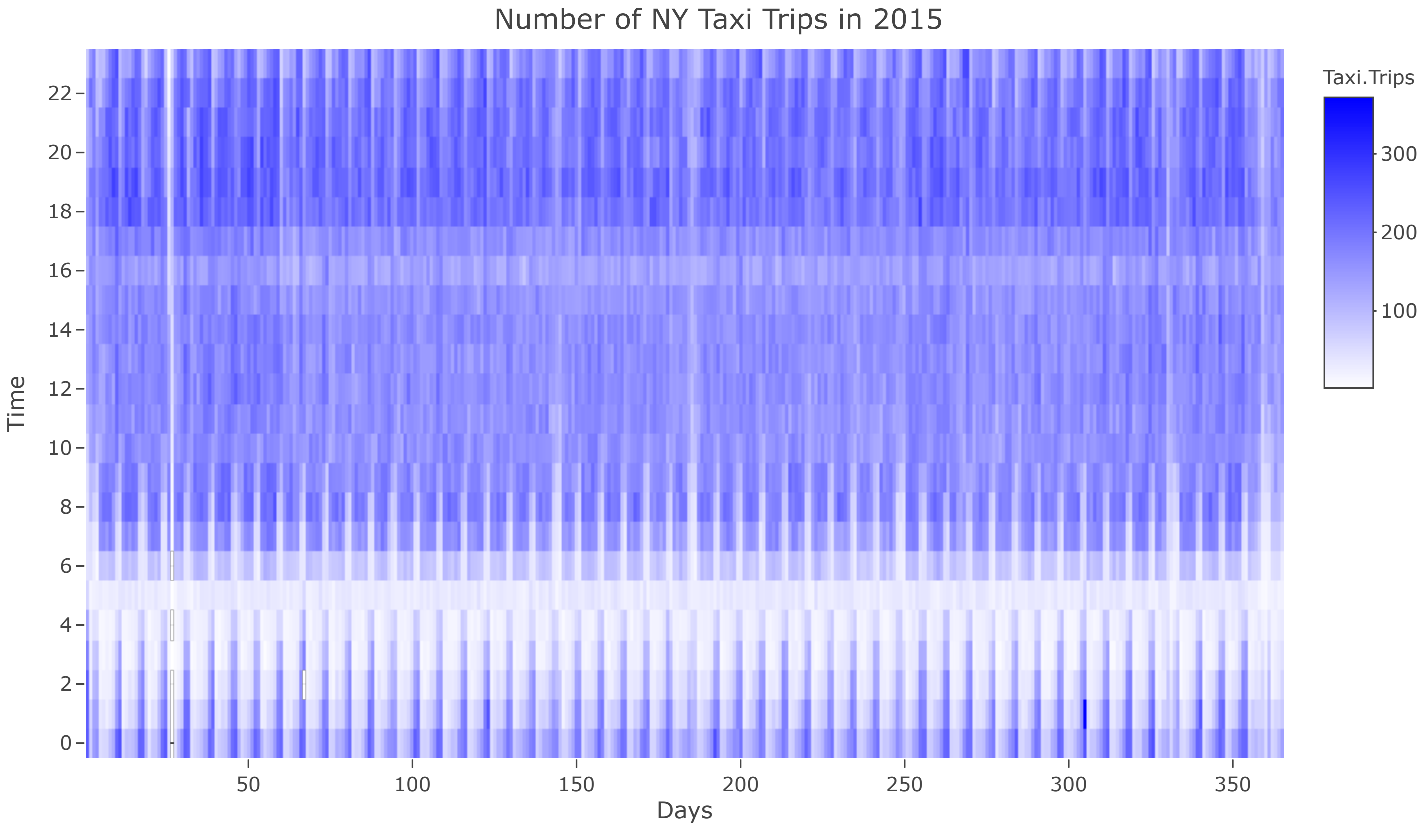
The following graph shows the average speed of taxi (which shows the speed of movement of traffic) and the number of taxi trips per day. We can see that both are inversely proportional. This makes sense, because when there are more taxi trips in a day, traffic is more.



Taxi Demand

The below heatmap shows the trend of number of taxi trips per hour. We can notice the following:

- Number of trips is less from midnight till 6AM when compared to day time. This trend is not seen on weekends where the number of taxi trips is high during the nights.
- There are discontinuities in the taxi trip trend at certain days. For example, number of trips drops gradually from *227* on 141st day of 2015 to *126* on 145th day. The 145th day is a holiday (Memorial Day, 25th May), and the number of trips reduced for the long weekend.
- Every other discontinuity is because of holidays, except for one: number of trips drops from *131* at 8PM on 25th day of 2015 to *38* on 26th.



26th January 2015 has reduced taxi trips

This is not a weekend or a holiday, but the number of taxi trips has gone down drastically from the previous day, and gets back up the next day. Let us see if weather is a reason behind this reduction.

pick_date	pick_timeslot	New.York	Taxi.number	Speed	days	isHol
2015-01-25	20	**few clouds**	**131**	13.64242	25	n
2015-01-25	21	**few clouds**	**126**	14.27737	25	n
2015-01-25	22	**few clouds**	**95**	16.27914	25	n
2015-01-25	23	**sky is clear**	**85**	16.32644	25	n
2015-01-26	20	**snow**	**38**	13.32915	26	n
2015-01-26	21	**overcast clouds**	**27**	14.09202	26	n
2015-01-26	22	**overcast clouds**	**14**	16.04324	26	n
2015-01-26	23	**snow**	**2**	13.08995	26	n
2015-01-27	20	**snow**	**96**	13.83844	27	n

2015-01-27	21	**broken clouds**	**90**	14.76170	27	n
2015-01-27	22	**broken clouds**	**96**	14.91296	27	n
2015-01-27	23	**overcast clouds**	**66**	17.50196	27	n

Since it snowed (quite heavily) from 26th until morning of 27th January, the number of taxi trips has reduced during this time. It later picked up during the day of 27th after the blizzard wore away.

Conclusion

Temperature follows a sinusoidal pattern, and the pattern (and median temperature) is slowly moving up through the years. By analysing the 2015 New York taxi trip data, we find that weather doesn't affect the taxi demand as much holidays do. Taxi trips reduce a little during extreme weather conditions, which happens very rarely as we have seen in the pie chart.

Refernces

1. Using R for Introductory Statistics - John Verzani
2. CS544 class notes