**Pickup Forecasting For NYC Taxi**

Problem Statement:

In response to the challenge of efficiently managing transportation demand in urban areas, particularly in New York City, where yellow taxi services are prevalent, there's a pressing need for accurate prediction models. Leveraging location coordinates (latitude and longitude) and time as inputs, the objective is to develop a robust forecasting model. This model should accurately predict the number of yellow taxi pickups within a specified query region and its surrounding areas.

Context:

In New York City, yellow taxis are a ubiquitous mode of transportation, primarily utilized in Manhattan but accessible across all five boroughs. Unlike other cities, New Yorkers frequently hail taxis from the street rather than relying on reservations. Moreover, external events like concerts and sports games can significantly influence taxi demand, necessitating a predictive mechanism to anticipate these fluctuations.

Approach:

The proposed solution involves employing time-series forecasting and regression techniques. By analyzing historical data and employing machine learning algorithms, patterns in passenger behavior can be identified and used to predict fluctuations in taxi demand. This predictive capability will empower taxi dispatchers with valuable insights, enabling them to optimize taxi allocation, determine the ideal number of dispatches, and adapt to changing passenger dynamics over time.

Significance:

A successful forecasting model will offer substantial benefits to taxi companies and city planners alike. Dispatchers will be equipped to make informed decisions that maximize profit margins by strategically placing taxis where they're most needed. Additionally, policymakers can leverage this data to address concerns such as traffic congestion and sustainability, fostering a more efficient and equitable transportation ecosystem for residents and visitors alike.

**Goals for this project:**

Understanding travel patterns:

* Finding traffic hotspots
* Travel patterns vary by time of day
* Impact of weekdays on demand

NYC Cab Data:

It's obtained from NYC Taxi & Limousine Commission Government Site. Data is collected from this site using Web Scraping Techniques.

Data Dictionary

*Yellow Taxi Trip Records*

*This data dictionary describes yellow taxi trip data.*

| VendorID | A code indicating the TPEP provider that provided the record.  1= Creative Mobile Technologies, LLC; 2= VeriFone Inc. |
| --- | --- |
| tpep\_pickup\_datetime | The date and time when the meter was engaged. |
| tpep\_dropoff\_datetime | The date and time when the meter was disengaged. |
| Passenger\_count | The number of passengers in the vehicle.  This is a driver-entered value. |
| Trip\_distance | The elapsed trip distance in miles reported by the taximeter. |
| PULocationID | TLC Taxi Zone in which the taximeter was engaged |
| DOLocationID | TLC Taxi Zone in which the taximeter was disengaged |
| RateCodeID | The final rate code in effect at the end of the trip.  1= Standard rate  2=JFK  3=Newark  4=Nassau or Westchester  5=Negotiated fare  6=Group ride |
| Store\_and\_fwd\_flag | This flag indicates whether the trip record was held in vehicle because the vehicle did not have a connection to the server.  Y= store and forward trip  N= not a store and forward trip |
| Payment\_type | A numeric code signifying how the passenger paid for the trip.  1= Credit card  2= Cash  3= No charge  4= Dispute  5= Unknown  6= Voided trip |
| Fare\_amount | The time-and-distance fare calculated by the meter. |
| Extra | Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges. |
| MTA\_tax | $0.50 MTA tax that is automatically triggered based on the metered rate in use. |
| Improvement\_surcharge | $0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015. |
| Tip\_amount | Tip amount This field is automatically populated for credit card tips. Cash tips are not included. |
| Tolls\_amount | Total amount of all tolls paid in trip. |
| Total\_amount | The total amount charged to passengers. Does not include cash tips. |
| Congestion\_Surcharge | Total amount collected in trip for NYS congestion surcharge. |
| Airport\_fee | $1.25 for pick up only at LaGuardia and John F. Kennedy Airports |

**Data Analysis/Cleaning.**

The data obtained is in its raw form and needs a lot of cleaning in order to proceed with applying any machine learning model. As a result, this section focuses on doing univariate analysis and removing outlier/illegitimate values which may be caused due to some errors.

**Removing Missing Values**

Missing values contributes to around 3% of the data and mostly the values are missing from important columns like passenger\_count. As a result, these data are dropped.

Missing Values Percentage:

VendorID 0.000000

tpep\_pickup\_datetime 0.000000

tpep\_dropoff\_datetime 0.000000

passenger\_count 3.208546

trip\_distance 0.000000

RatecodeID 3.208546

store\_and\_fwd\_flag 3.208546

PULocationID 0.000000

DOLocationID 0.000000

payment\_type 0.000000

fare\_amount 0.000000

extra 0.000000

mta\_tax 0.000000

tip\_amount 0.000000

tolls\_amount 0.000000

improvement\_surcharge 0.000000

total\_amount 0.000000

congestion\_surcharge 3.208546

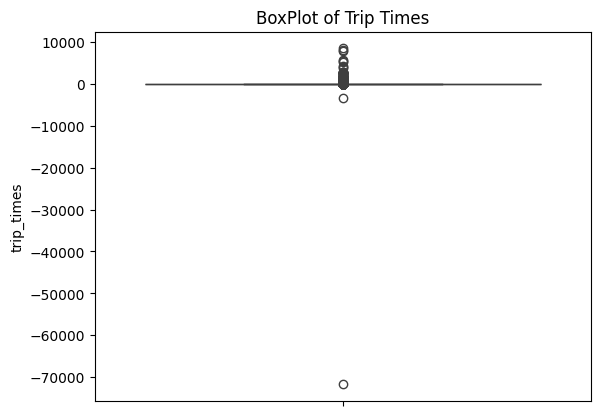
airport\_fee 3.208546

dtype: float64

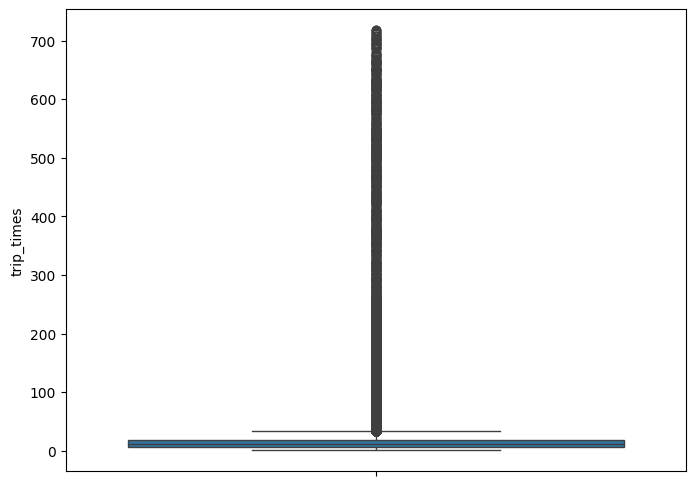
### **Trip Duration**

According to the NYC Taxi & Limousine Commission, a trip may not last longer than 12 hours in any 24-hour period. In order to comply with this rule, trip duration is calculated and any outliers are removed from the dataset.

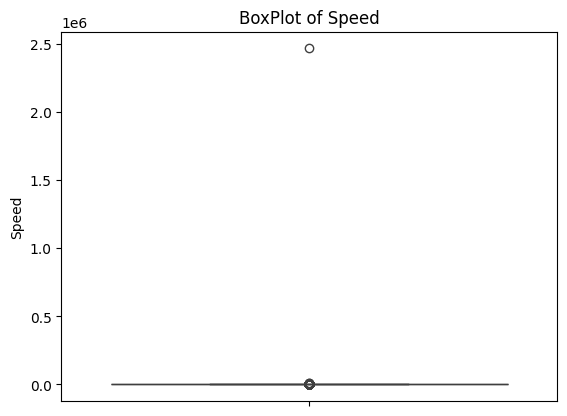
The timestamps are converted to Unix so as to get duration(trip-time) & speed also pickup times in Unix are used while binning in our data. We have time in the format “YYYY-MM-DD HH:MM:SS” .



After applying the outliers we can see the consistency

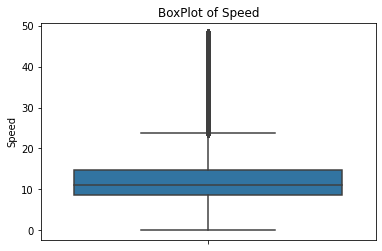


### **Speed**



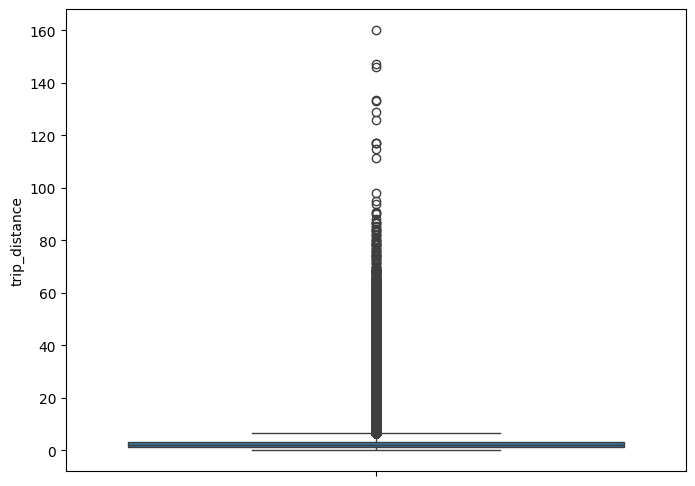
From the above observations, it is evident that there are outliers in the data. Both the box plot and percentile calculation verifies the same.

*Upon calculation, the avg speed in New York City is 12.19 miles/Hr, so a cab driver can travel 2 miles per 10 min on avg.*

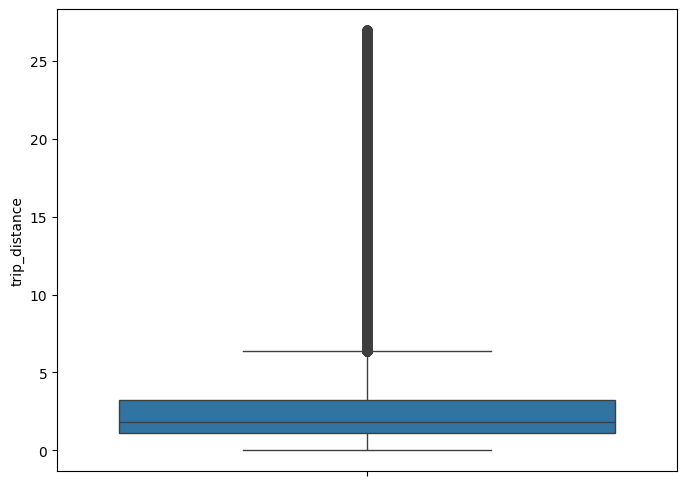


**Trip Distance**

We need not keep the data for the trips where the distance is suspiciously high as it might pollute the training data, hence we remove them.



After removing the rows with distance above 99th percentile value we get a consistent value plot of distance



### **Pickup and Drop Location outliers**

* It is inferred from the source <https://www.flickr.com/places/info/2459115> that New York is bounded by the location cordinates (lat,long) - (40.5774, -74.15) & (40.9176,-73.7004)
* Hence, any cordinates not within this cordinate range, are not considered by us, as we are only concerned with pickups which originate within New York.

So upon getting the latitude and longitude of each location based on the locationID col we got to know that significant amount of rows were outside our inference boundaries, hence were removed.

Outliers coordinates lying outside NY boundaries: 18550

# 2. Data Preparation