# PROJECT MILESTONE 3

## PACE: Plan Stage

- What are the data columns and variables and which ones are most relevant to my deliverable?

There are 18 data columns in the dataset, but the most relevant to my deliverable are 'trip_distance' and 'total_amount' variables.

- What units are the variables in?

Both variables are in decimal units.

- What are my initial presumptions about the data that can inform my EDA, knowing I will need to confirm or deny with my future findings?

I think that these two variables show most for the taxi cab rides in the city. Of course, with my future findings I will confirm or deny this presumption.

- Is there any missing or incomplete data?

There are no null values, but there are many outliers in the dataset which should be excluded from future analysis.

- Are all pieces of this dataset in the same format?

From the total of 18 columns, 15 columns are numerical, and 3 columns are non-numerical.

- Which EDA practices will be required to begin this project?

Data loading and inspection, data cleaning and preprocessing, summary statistics, visualization, data quality assessment, feature selection.

## PACE: Analyze Stage

- What steps need to be taken to perform EDA in the most effective way to achieve the project goal?

1. Data loading and inspection
2. Visualization
3. Correlation analysis
4. Feature engineering
5. Outlier detection
6. Data transformation

- Do I need to add more data using the EDA practice of joining? What type of structuring needs to be done to this dataset, such as filtering, sorting, etc.?

Beside 'trip_distance' and 'total_amount', it may be useful to draw 'trip_duration' as new variable from the dataset based on pick-up and drop-off times. For that purpose, it is necessary to convert those two columns into date/time format.

- What initial assumptions do I have about the types of visualizations that might best be suited for the intended audience?

A box plot will be helpful to determine outliers and where the bulk of the data points reside in terms of 'trip_distance', 'duration' (new feature), and 'total_amount'.

A scatter plot will be helpful to visualize the trends and patters and outliers of critical variables, such as 'trip_distance' and 'total_amount'.

A bar chart will help determine average number of trips per month, weekday, weekend, etc.

## PACE: Construct Stage

- What data visualizations, machine learning algorithms, or other data outputs will need to be built in order to complete the project goals?

Box plots, scatter plots and bar charts will need to be built.

- What processes need to be performed in order to build the necessary data visualizations?

Some of the plots was drawn directly from the dataset, but for more complex plots it was needed first to extract the necessary data in new series and dataframes, and after that to plot from these sources.

- Which variables are most applicable for the visualizations in this data project?

'total_amount', 'tip_amount' and 'trip_distance'.

- Going back to the Plan stage, how do I plan to deal with the missing data (if any)?

There is no missing data in the original dataset.

## PACE: Execute Stage

- What key insights emerged from my EDA and visualizations(s)?

I have learned that the majority of trips were journeys of less than two miles, with the most of trips less than 5 miles, but there are outliers all the way out to 35 miles. There are no missing values.

The most costs falling in the $5-15 range.

Nearly all the tips falling in 0-3$ range.

Tip amounts for tips>10$ shows the similar distribution for both vendors. Most of the tips are between 10 and 13$ for both vendors.

Most of the time there was only one passenger in the taxi cab (71%). Two passengers were in the taxi cab rides 14,5% of the trip time. The rest were in significantly smaller percentage. There were still nearly 700 rides with as many as six passengers. Also, there are 33 rides with an occupancy count of zero, which doesn't make sense. These would likely be dropped unless a reasonable explanation can be found for them.

Mean tip amount varies very little by passenger count.

The number of monthly rides doesn't vary a lot, but it is clear that the number of rides is smaller in the summer months (July to September), and also in February. The peak of rides is in March.

The bar plot of mean trip distances by drop-off location presents a characteristic curve related to the cumulative density function of a normal distribution. In other words, it indicates that the drop-off points are relatively evenly distributed over the terrain.

The histogram of rides by drop-off location reveals that out of the 200+ drop-off locations, a disproportionate number of locations receive the majority of the traffic, while all the rest get relatively few trips. It's likely that these high-traffic locations are near popular tourist attractions like the Empire State Building or Times Square, airports, and train and bus terminals. However, it would be helpful to know the location that each ID corresponds with. Unfortunately, this is not in the data.

- What business and/or organizational recommendations do I propose based on the visualization(s) built?

Because a disproportionate number of locations receive the majority of the traffic, while all the rest get relatively few trips, I will propose to receive the location that each ID corresponds with in order to investigate the reasons of this disproportion.

- Given what I know about the data and the visualizations I was using, what other questions could I research for the team?

1. What to do with the outliers, taking into account that they are almost half of the total data in the dataset?

2. What to do with 33 rides with an occupancy count of zero, which doesn't make sense. These would likely be dropped unless a reasonable explanation can be found for them.

- How might I share these visualizations with different audiences?

I will prepare them to suit to every different audience.