**Course Two**

# Get Started with Python

**Data Project Questions & Considerations**

**PACE: Planning Stage**

* How can you best prepare to understand and organize the provided information?

Figure out the missing column names if any. Like a trip ID. Label, it for future use.

Figure out which are the variables required to achieve the desired result and remove the other variables to keep the dataframe concise.

Figure out if there are any missing values in the dataset and if it would affect the desired result. If any, get back to TLC and ask if there is a way that data could be provided.

Figure out if there is any column that we need to modify or create new variables from the data which might be useful

* What follow-along and self-review codebooks will help you perform this work?

Data Structures in Python and pandas notebook

* What are some additional activities a resourceful learner would perform before starting to code?

Think about the questions he would like to explore during the inspection.

Make sure he looks at the data without any bias. Do not make assumptions.

**PACE: Analyzing Stage**

* Will the available information be sufficient to achieve the goal based on your intuition and the analysis of the variables?

Yes. The available information does seem to have relevant data for analysis upon first inspection. Although a further analysis while digging deeper would give us a better understanding if other variables might be useful.

**PACE: Constructing Stage**

* How would you build summary dataframe statistics and assess the min and max range of the data?

We can use **describe()** functionality to assess the basic dataframe statistics. Info()

* Do any data variables averages look unusual? Can you describe the interval data?

No. Not upon initial inspection. Digging a bit deeper would give a better understanding of this. Groundwork has been laid for further EDA.

**PACE: Execute Stage**

* Given your current knowledge of the data, what would you initially recommend to your manager to investigate further prior to performing exploratory data analysis?

Questions like:

How many trips were taken on an hourly basis? ***For eg: 12:00 - 1:00 >>>> 150 trips***

How many trips were taken on an hourly basis grouped by location?

What was the average trip\_distance on hourly basis grouped by the location?

These are a few questions which might help us understand the data better and give us an analysis as well as a direction for building a machine learning model.

* What data initially presents as containing anomalies?

So, there are a few rows in which the **trip\_distance** is less and still the **total\_fare** is high.

* What additional types of data could strengthen this dataset?

No recommendation here. Might get some while exploring the dataset further.