PROJECT MILESTONE 5

PACE: Plan Stage

- What am I trying to solve or accomplish?

I'm trying to develop a multiple linear regression model that helps estimate taxi fares before the ride, based on data that the taxi company has gathered over the course of a year.

- What are the initial observations exploring the data?

The dataset has 22.699 rows and 18 columns, of which 7 columns are of int64, 8 columns of float64, and 3 columns of object dtypes.
There are no duplicates and no missing values in the dataset.

- What resources do I find myself using as I complete this stage?

- taxi trip data

- Jypiter notebook

PACE: Analyze Stage

- What are some purposes of EDA before constructing a multiple linear regression model?

1. Outliers and extreme data values can significantly impact linear regression equations. After visualizing data, make a plan for addressing outliers by dropping rows, substituting extreme data with average data, and/or removing data values greater than 3 standard deviations.

2. EDA activities also include identifying missing data to help the analyst make decisions on their exclusion or inclusion by substituting values with data set means, medians, and other similar methods.

3. It's important to check for things like multicollinearity between predictor variables, as well to understand their distributions, as this will help me decide what statistical inferences can be made from the model and which ones cannot.

4. Additionally, it can be useful to engineer new features by multiplying variables together or taking the difference from one variable to another. For example, in this dataset I can create a `duration` variable by subtracting `tpep_dropoff` from `tpep_pickup time`.

- Do I have any ethical considerations at this stage?

Not so far.

PACE: Construct Stage

- Do I notice anything odd?

No, I didn’t notice anything odd.

- Can I improve the model? Is there anything I would change about the model?

Model can be improved by excluding trips with RatecodeID of 2 because they lead to the international, and all have same fares of $52, no matter the starting point.

I don’t want to predict on those trips with fixed fares, because they don’t bring any predictive power to the model.

- What resources do I find myself using as I complete this stage?

Jypiter notebook

PACE: Execute Stage

- What key insights emerged from the model(s)?

The feature with the greatest effect on fare amount was ride duration, which was not unexpected. The model revealed that for every 2.82 miles traveled, the fare increased by a mean of $5.86. Or, reduced: for every 1 mile traveled, the fare increased by a mean of $2.08.

- What business recommendations do I propose based on the models built?

In order to increase fare amounts, the regression model insights recommend to stimulate longer trips.

- To interpret model results, why is it important to interpret the beta coefficients?

They indicate how much the response variable changes for every unit change in the predictor variable, holding all other predictors constant.

- What potential recommendations would I make?

Recommend to stimulate longer trips and to avoid short trips.

- What business/organizational recommendations would I propose based on the models built?

The taxi company can use these findings to create an app that allows users to see the estimated fare before their ride begins.

- Given what I already know about the data and the models I were using, what other questions could I address for the stakeholders?

Request additional data from under-represented itineraries.

- Do I have any ethical considerations at this stage?

No ethical considerations for this stage.