Assessment report

Alexandra Saizan

# Assumptions

I made the following assumptions while conducting this analysis:

1. The demographic data provided is collected at the zone level
2. The word “recommend” in question 5 implies that the question is asking us to build a recommendation system; however, predictions from a supervised learning algorithm are sufficient to act as a recommendation for a passenger’s tip percentage based on historical data
3. Choropleth maps are sufficient for understanding high-level traffic trends

# Methods

## Data transformation

I used pandas to perform the data manipulation tasks, including loading the data, generating summary statistics, and merging datasets.

## Data visualization

For most of the visualizations, I used matplotlib. I used geopandas to visualize traffic trends because it is equipped to deal with shapefiles (the geographic data provided by TLC) and had the lowest learning curve.

## Machine learning

To generate a tip recommendation for a particular trip, I used a gradient boosting regression algorithm implemented in scikit-learn. I relied on the gradient boosting algorithm because I assumed an ensemble method would provide the best performance, and without hyperparameter tuning gradient boosting performed better than random forests. For this exercise, I wanted to stay within the scikit-learn framework to increase my development speed. I have used XGBoost and LightGBM before on a client project and would have tested out that algorithm if I had additional time.

# Conclusions

Most of the results reflect my expectations about taxi rides in NYC. Most pickups and dropoffs tend to occur during rush hour. Airport fares tend to be higher on average, and most airport trips also occur during the rush hour timeframe. Airport trips declined significantly after the pandemic begun, which reflects the decline in travel overall. Most dropoffs occur in Manhattan, which did not surprise me even as a non-NYC resident. The baseline machine learning performance (without hyperparameter optimization) was poor. I found that in general, the average predicted tip percentage is lower for zones with Black and Hispanic population rates that are higher than the median rates. This result suggests that the model may present fairness issues.

# Limitations

I am not familiar with recommendation systems, as most of my work in public sector consulting has focused on supervised learning or natural language processing problems. I used a supervised learning algorithm for this project because it was most expedient for me to implement and because it would be easiest for me to assess ethical challenges in the results.

I would have preferred to run AutoML to determine the best-performing regression algorithm in an automated fashion. However, setting up an AutoML package on my computer would have increased my development time, and because this project is just a notional example, I did not think it was worth it to spend time developing more robust models.

Additionally, I noticed that many of the taxi zones did not have recorded demographic information. Consequently, I dropped rows with missing values because assessing demographic information was central to question 6. Had those records contained complete demographic information, including them would have likely improved my model’s performance.

# References

I used the following blog post as a template for mapping with geopandas: <https://www.relataly.com/visualize-covid-19-data-on-a-geographic-heat-maps/291/>. I used additional TLC data, including the [taxi zone lookup table](https://s3.amazonaws.com/nyc-tlc/misc/taxi+_zone_lookup.csv) and the [taxi zone shapefiles](https://s3.amazonaws.com/nyc-tlc/misc/taxi_zones.zip), to supplement my analysis.