**AMOD 5240H – Assignment 06a**

Nicholas Hopewell - 0496633

For my analysis project, I am initially interested in a data set released by the NYC Taxi and Limousine Commission which details information about taxi trips in New York City. Features included in the data set describe taxi pick-up times, drop-off times, number of passengers along for each ride, geographical coordinates for pick-up locations, geographical coordinates for drop-off locations, and trip duration. Most of these feature data types are continuous numeric as well as discrete numeric (passenger count). This data set, which includes 1458644 trip records in total, was made available by Big Query and was later put on Kaggle. The NYC taxi data set is feasible because it has been preprocessed enough to where the features are realistically manageable to a novice data analyst. Furthermore, it contains features which are related to the target variable “trip duration” but are not highly inter-related making it suitable for a predictive analysis.

This data set was initially interesting to me because I believe I could create new features from existing features such as distance traveled and relative speed. A couple things worth exploring would be whether newly creator features such as relative speed of cab driver are better predictors of trip duration than features such as time of day. Perhaps some drivers take less-congested roads, thus traveling further, but do so at higher speeds to reduce trip duration. This idea could maybe be explored through looking at drivers who traveled similar locations at similar times of days, and seeing whether some drivers did their routes significantly faster than other drivers (labeled as ‘fast’ drivers). Another interesting topic involves exploring the passenger count feature as it related to travel duration and determining the optimal number of passengers to carry to reduce trip duration.

Importantly though, I am considering openly available Canadian government data sets, and so my focus might entirely change if I find something more interesting that is also realistic.