1. **Background and introduction**

why I choose this topic

to predict the *duration of taxi rides in NYC* based on features like trip coordinates or pickup date and time. The [data](https://www.kaggle.com/c/nyc-taxi-trip-duration/data) comes in the shape of 1.5 million training observations (../input/train.csv) and 630k test observation (../input/test.csv). Each row contains one taxi trip.

1. **Data source and summary**

taxi trip data/weather data/

vendor\_id only takes the values 1 or 2, presumably to differentiate two taxi companies

pickup\_datetime and (in the training set) dropoff\_datetime are combinations of date and time that we will have to re-format into a more useful shape

passenger\_count takes a median of 1 and a maximum of 9 in both data sets

The pickup/dropoff\_longitute/latitute describes the geographical coordinates where the meter was activate/deactivated.

store\_and\_fwd\_flag is a flag that indicates whether the trip data was sent immediately to the vendor (“N”) or held in the memory of the taxi because there was no connection to the server (“Y”). Maybe there could be a correlation with certain geographical areas with bad reception?

trip\_duration: our target feature in the training data is measured in seconds.

data wrangling.

1. summary
2. sample making model

visualise the original data, engineer new features, and examine potential outliers