The goal of this homework is to train a simple model for predicting the duration of a ride - similar to what we did in this module.

Q1. Downloading the data

We'll use the same NYC taxi dataset, but instead of "Green Taxi Trip Records", we'll use "For-Hire Vehicle Trip Records".

Download the data for January and February 2021.

Note that you need "For-Hire Vehicle Trip Records", not "High Volume For-Hire Vehicle Trip Records".

Read the data for January. How many records are there?

- 1054112
- 1154112
- 1254112
- 1354112

Q2. Computing duration

Now let's compute the `duration` variable. It should contain the duration of a ride in minutes.

What's the average trip duration in January?

- 15.16
- 19.16
- 24.16
- 29.16

Data preparation

Check the distribution of the duration variable. There are some outliers.

Let's remove them and keep only the records where the duration was between 1 and 60 minutes (inclusive).

How many records did you drop?

Q3. Missing values

The features we'll use for our model are the pickup and dropoff location IDs.

But they have a lot of missing values there. Let's replace them with "-1".

What's the fractions of missing values for the pickup location ID? I.e. fraction of "-1"s after you filled the NAs.

- 53%
- 63%
- 73%
- 83%

Q4. One-hot encoding

Let's apply one-hot encoding to the pickup and dropoff location IDs. We'll use only these two features for our model.

- Turn the dataframe into a list of dictionaries
- Fit a dictionary vectorizer
- Get a feature matrix from it

What's the dimensionality of this matrix? (The number of columns).

- 2
- 152
- 352
- 525
- 725

Q5. Training a model

Now let's use the feature matrix from the previous step to train a model.

- Train a plain linear regression model with default parameters
- Calculate the RMSE of the model on the training data

What's the RMSE on train?

- 5.52
- 10.52
- 15.52
- 20.52

Q6. Evaluating the model

Now let's apply this model to the validation dataset (Feb 2021).

What's the RMSE on validation?

- 6.01
- 11.01
- 16.01
- 21.01