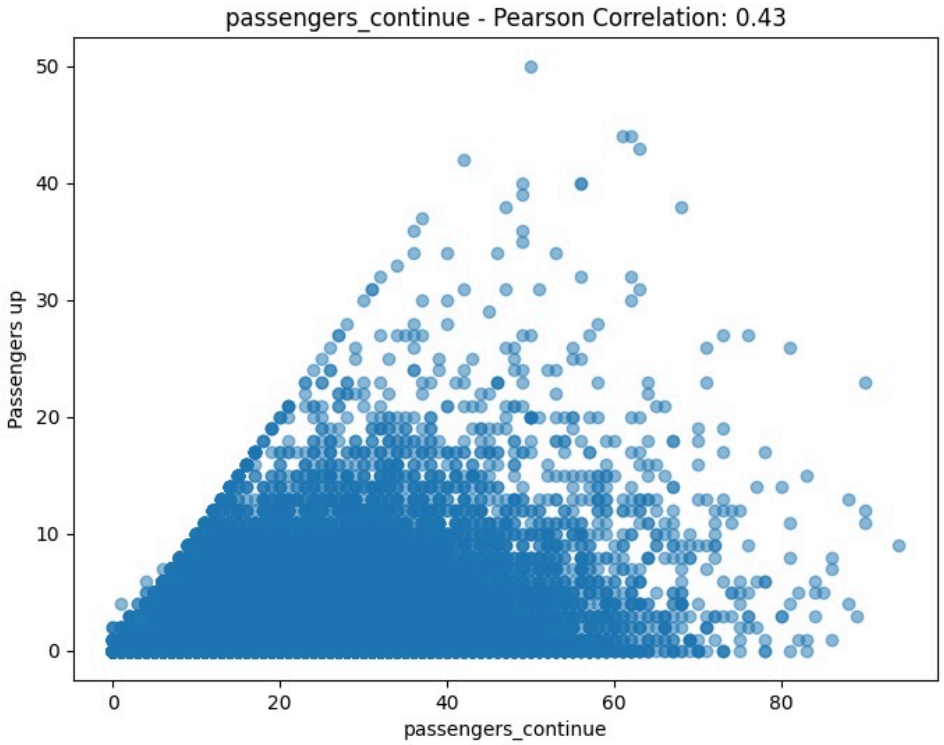
Part A:

Baseline model:

We preprocessed the model by inserting the corresponding arrival_time to the missing door_closing_time, then we made a new column that is $cell_i(door\_closing\_time) - cell_i(arrival\_time)$ in seconds, that is the feature time_in_station. We believe that there is a correlation between the amount of time the bus spent at the station and the amount of people that entered. In addition, after finding the Pearson correlation for each feature, we conclude that the feature with most impact on the results is 'passengers_continue':



We preprocessed this feature such that all the negative values will be zero, since we saw that the negative values are in the last station.

We used 80% of the data for training and the rest for tests and a linear regression model.

We got the following result:

Mean Squared Error: 3.47

Second baseline:

We used a RandomForestRegressor to fit the model with the arrival_time feature.
We used 80% of the data for training and the rest for tests.

We got the following result:

Random Forest MSE: 5.12.

The final model:

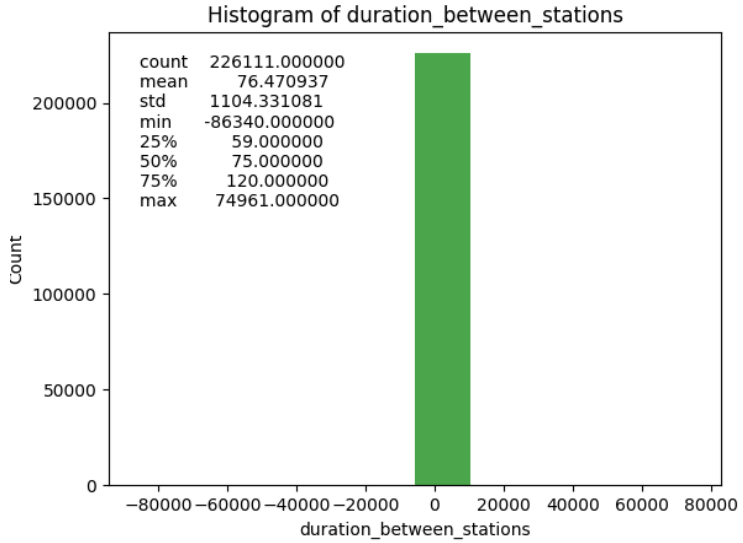
The final model was combined from the first baseline model and from another 3771 models. We recognized that the number of passengers boarding at each station is dependent on the passengers going up and decided to train a separate model for each station (total 3771 models) to capture this dependence accurately. We processed the data as outlined in our baseline approach and then grouped the data by station to train individual models tailored to each station.
However, we encountered a challenge when we had test data for a station that did not have a corresponding trained model. The baseline approach gives us a reasonable prediction for stations that have very little data. To address this, we opted to use the baseline model for predictions at these stations, as the low frequency of occurrences made training a dedicated model for them unnecessary.
Indeed, training a model specifically tailored to each station proved beneficial. We observed an improvement in our predictions for passenger boarding counts, as evidenced by the Mean Squared Error (MSE) at **2.84**, which is better than our previous model by 0.63.

Part B:

Baseline model:

We first try to predict the time spent on each station by training a Linear Regression model that has the features of distance between stations (that were calculated from the lon, lat of each station) and the passengers_up feature, that are related for the the time the bus spent on a station and the ride for the next station. However, with testing our results, we got the MSE of the prediction with 1217529.86, which is quite a bad approximation and we can't work with it.

To understand the data, we plotted the histogram on our label (that is, the duration between stations) and we got the following histogram:



We can see that the labels have a reasonable mea, but a significant std and a few outliers.