# Personal report

## Log:

1. Data explotion: plot some figures to find if there are regular patterns.
2. Data clean: transform classified features and decrease redundancy. Convert invalid values if they exist.
3. Choose model to train. Then use decision tree.
4. Summarize the completion process and reflect on it.

## Contribution：

1. Download dataset and hotcodeing the classified value, then saved the converted data into a new “csv” file to teammate.
2. Explore the dataset and plot some figures. Compare my results with teammates’ result.
3. Write code to use decision tree model to predict the number of London shared bikes.
4. Write most part of introduction document and make most part of ppt.

## Difficulties:

1. Consider how to explore the dataset, means to kown what we want to now and what figures can we plot to show more information clearly.
2. How to use classified model, decision tree, to predict constant value —— the number of London shared bikes. How to transform the other constant features, like wind speed.
3. To judge if there are invalid values, if there are, how to transform them.
4. The model perfects badly, What’s the problem? And how to improve the accuracy?

## Solution:

1. Considered relationship of each feature with the number of shared bikes. So drew figures of each feature and number.
2. Divide the number of shared bikes into 4 classes firstly. The dividing principle is like that, [min, Q1], [Q1, mean], [mean, Q3), [Q3, max]. Futher, tried the 2 classes by principle that,

[min, mean], [mean, max]. The first try showes accuracy of 0.4XX, the second try showes accuracy of 0.7XX.

1. By confirm the statistic analysis of dataset and see source data directly, confirming there are not invalid values. But there redundant feature. So I select the one has large standard deviation, less outliers as the better one.
2. The reason may as follows. The number of bicycles varies in a large range. Take the minimum value and the maximum value as the boundary. If there is a situation in the test set that is smaller than the minimum value of the training set and the maximum value is larger, it cannot be judged. Similarly, the training set takes the average value as the dividing boundary, which is easily affected by the extreme value, and it is easy to produce a large deviation from the test set. To try other kinds of division ways and analyse which one is better.

## Impovement:

1. Use other kinds of ways to classify the numder of shared bikes.
2. Use DecisionTreeRegressor to train model, and compared with DecisionTreeClassifier.
3. Use tensorflow to predict the result.