Final Project: New York City Bike Traffic **Hojoung Jang**
Yul Hee Cho

This project aims to analyze bike traffic data across a number of bridges in New York city. The dataset is collected by the New York City Department of Transportation for a stretch of 6 months (from April to October 2016) and contains daily record of the number of bicycles crossing into or out of Manhattan via one of the East River bridges which are Brooklyn, Manhattan, Williamsburg and Queensboro bridges. In addition to the traffic data, this dataset also provides minimum, maximum temperature and precipitation data.

This data analysis project aims to answer three main questions:

1. If we want to install sensors on the bridges to estimate overall traffic across all the bridges but only have budget to install only on three out of four bridges, which bridges should we choose to best represent the overall bike traffic?
2. The city administration is cracking down on helmet laws, and wants to deploy police officers on days with high bike traffic to out citations. Can they use the nest day's weather forecast to predict the number of bicyclists that day?
3. Can we use the traffic data to predict whether it is raining based on the number of bicyclists on the bridges?

1. Problem 1 Analysis:

For Problem 1, we chose to create linear regression model using traffic data of three bridges. Since we have 4 bridges, we make unique combinations of three bridges and we actually end up with 4 unique combinations:

a. Brooklyn, Manhattan, Williamsburg
b. Brooklyn, Manhattan, Queensboro
c. Brooklyn, Williamsburg, Queensboro
d. Manhattan, Williamsburg, Queensboro

By training the datasets that are grouped of 3 bridges, we could then build a linear regression model to find the model that best fits the total bike traffic data. After getting trained dataset, we finally predict the overall traffic.

To find the best fitting answer, we compared each model's r-squared (r^2) value which is the representation of how the represents the data we are trying to

model. R-squared ranges from 0 to 1 with 0 meaning the model does not represent the dataset at all and 1 meaning the model perfectly represent the dataset.

Problem 1 Result:

- R-squared Value Table

	Brooklyn, Manhattan, Williamsburg	Brooklyn, Manhattan, Queensboro	Brooklyn, Williamsburg, Queensboro	Manhattan, Williamsburg, Queensboro
r-squared	0.996	0.988	0.947	0.982

Model Equation:

Total Traffic = 1.1386000788715267 * (Brooklyn Traffic) +
0.9471171505682368 * (Manhattan Traffic) +
1.6086469611158551 * (Williamsburg Traffic) +
382.7456681782314

By using this data, it could be determined that the prediction score was the highest when we install the traffic sensors to the bridges not including the Queensboro bridge. Therefore, it could be said that if the budget is limited to only 3 bridges amount of sensors, the maximum efficiency could be reached when installing the sensors to Brooklyn bridge, the Manhattan bridge, and the Williamsburg bridge.

2. Problem 2 Analysis:

In problem 2, we are trying to figure out the relationship between the weather data and traffic data. We can model a linear regression model. To do this, we set the three weather data, minimum temperature, maximum temperature and precipitation as features (independent variable) of our model and set the total traffic data as our output value (dependent variable). Once we get our model trained and built, we can evaluate our model by feeding test or unseen data and determine if this model is good enough to predict the number of bicyclists on that day.

Problem 2 Result:

Unfortunately, the model did not yield a reasonable score, so we played around with the configuration values a little bit to get our best model score. In the end, we chose to change the test set size to 19 to yield the best score of **0.433** (scale of 0 to 1).

Model Equation:

$$\text{Total Traffic} = 406.17916154797587 * (\text{High Temp}) - 170.7826243577951 * (\text{Low Temp}) - 8034.912567405353 * (\text{Precipitation}) - 422.8260224649057$$

We really wanted to create a model that could represent total traffic using weather data as our features, but as we have analyzed, our best model gave a score of **0.433.** This is not the best model we can present to solve a real world problem, and we concluded that the weather data is not perfect to predict people biking in New York bridges. Thus, the police officers should not fully rely on the weather and determine the number of bicyclists to enforce the helmet law.

3. Problem 3 Analysis:

In problem 3, we want to determine if we can predict if it is raining or not given that we have the number of bicyclists. This sounds very similar to problem 2 but, in fact, it was a totally different problem. Here, we noticed that our output is either a 'yes' or 'no' which are discrete values (1 or 0). We noticed that this is a classification problem and chose to use logistic regression to fit a model but later applied Naive Bayes classification method because it performed better than the logistic regression model.

Problem 3 Result:

On a scale from 0 to 1, the Gaussian Naive Bayes model gave a score of **0.837**.

- Table. Actual and Prediction Values on Test Set

	Rain (1=raining, 0=not raining)																																										
Actual	1	0	0	0	1	1	1	0	0	1	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0
Predicted	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0

- Table. Confusion Matrix

	Predicted		
Actual		True	False
	True	29	2
	False	5	7

As we can see, we are able to correctly predict 36 out of 43 test data.

We are able to predict about **83.7%** of the guesses we make on the raining condition using the bike traffic. Therefore, we concluded that this is a fair model and it is fair to say **'yes'** when we are asked to predict if it is raining based on the number of bicyclists.