20875 Mini Project

Introduction:

This analysis conducted on path one with the cyclist and weather data through four bridges in New York City was completed by Andrew Hockman (GitHub: spamdrew128; Purdue: ahockma), and Riley Arnholt (GitHub: r-arn49; Purdue: rarnholt).

The Data Set

The team was given a .csv containing data with three distinguishable categories. The first major category contains the dates and the day of the week for each subsequent data point. The second major category consists of a small portion of the weather data on those days, including high temperatures, low temperatures, and the precipitation levels. The third major category of data lists the four different bridges to be analyzed and the amount of traffic per bridge. This category also includes the total number of bicyclists per day across the four bridges. Each subcategory contains two hundred fourteen points.

The Analyses

The first involved four bridges to install sensors on to give the best estimate of the total traffic across all bridges, but due to budget constraints only three out of the four bridges can have one installed. One assumption the team made before the analysis took place was not weighing different times of the year. However, this assumption was carried over for all the bridges. The team tackled this question by first finding the average amount of travelers per weekday per bridge. For example, all the travelers on Monday on the Brooklyn bridge became one average taken for that bridge. With each of the averages calculated, the Mean Squared Error or MSE was determined between the averages for each day for each bridge and the total number. Then, using a process similar to cross validation, where one data set or "bridge" was removed, the data was finished being analyzed and a result was obtained.

The second question involved determining when the city should place their officers to enforce helmet laws across the bridges. The city wanted to determine whether or not the weather forecast could be used to determine when to place the officers. Some assumptions were made to simplify the analysis. First, the date of the data's collection was ignored, such that seasons and holidays were not taken into consideration. Secondly, the day of the week was also not taken into account, such as whether Saturday has higher traffic than Wednesday. This analysis was done solely by looking at the trend between the weather characteristics provided and the amount of total riders across the bridges. Linear regression was used to create a function that accepts three inputs of weather conditions and return a predicted amount of riders. The MSE of the function and the R^2 value were calculated to determine a method of deciding goodness of fit.

The final question to be answered was regarding if the amount of people traveling on the bridges on a given day could be used to predict whether or not it is raining on that day. One assumption that was made was to ignore the date and the day of the week. The team chose the boundary between rain and not rain to be 0.1 inches as meteorologists have determined this to be the difference between no rain and light rain. This analysis focused on the relationship between the precipitation and the amount of travelers on the bridges. Because shows a binary relationship between if it is raining or not raining, logistic regression was used for an effective way to see the trend between the state of rain and travelers.

The Results

After completing the analysis described for the first question, the Brooklyn Bridge was determined to be the bridge that needed to be left out of receiving sensors. When the cross validation finished, the one that did not include Brooklyn Bridge produced the smallest MSE between the averages of the bridges for the day and the total average for each day.

Bridge Left Out	Brooklyn	Manhattan	Williamsburg	Queensboro
MSE Value	9343441.875506	26288372.95779	39376705.13928	19063584.667665

Table 1: MSE values determined between the other bridges not left out and the total average

Once the model was created to show the relationship between the weather and the amount of travelers using linear regression, an R^2 value was determined. However, with the model taking in high temperature, low temperature, and the precipitation of the day, the model produced an R^2 value of 0.4868, indicating a very mild goodness of fit. Using the developed model, it cannot be recommended that the weather be used as a prediction device to show where to place the officers. Note: due to the data being unnormalized, the coefficients are not weighted equally. However, the sign of each does appear to make logical sense when thinking about how weather affects outdoor activities.

MSE	16195757.449488156
R^2	0.4868487345487541

Equation: Travelers = 390.28(High Temp) – 165.91(Low Temp) – 7844.66(Precip.) + 456.34

Table 2: Information regarding the developed model

Due to the binary nature of the final question, logistic regression was used to determine a relationship between the amount of travelers and rain. With a cutoff value chosen to determine what is rain and what is not enough to be considered rain, the model was created. With an accuracy value of 0.8263, the model made is fairly capable of predicting whether there is rain or no rain based on the number.