Elinor Velasquez
Capstone Two Problem Statement
Seoul, South Korea Bike Sharing Demand

The problem to be solved is the prediction of bike count required per day and at a given collection of locations for the stable supply of rental bikes. The client is Seoul Bike, which participates in a bike share program in Seoul. An accurate prediction of bike count is critical to the success of the Seoul bike share program. Based on our analysis the client may move bikes to different locations, by adding or subtracting bikes at certain locations depending on the predictions of bike locality and usage. It is important to make the rental bikes available and accessible to the public at the right time as it lessens the waiting time: Providing the city with a stable supply of rental bikes is a major concern for the program. The goal of the project is to predict the required number of rental bikes per day in a group of locations using machine learning and data mining.

The data set is concerned with rental bikes as a ride sharing program. There are 8760 instances and 14 attributes (features). The features are aspects of the weather, the number of bikes rented per hour and date information. The criteria for success is the following: The score to rate the prediction will be the R^2. M.A.E. and other types of error will be computed to rate the success of the model used to predict bike needs. The critical part is the prediction of bike count per day in a group of locations for a stable supply of rental bikes. The stakeholders are affiliated with the Seoul bike share program: the program operators and the bicyclists who use the program to ride bicycles in Seoul.

Data sources are the following: The data was acquired from the Machine Learning Repository, Center for Machine Learning and Intelligent Systems, U.C. Irvine. Histograms will be plotted to see if the data is normally distributed or skewed. A Pearson correlation heatmap will be generated to see which features are linked to bike usage. The first prediction will be usage based on the mean as the simplest regression map to get a baseline for the problem. A random forest regression will be used to get a better estimate. A support vector machine may be used to predict bike usage. A convolutional neural network may also be used to predict bike usage. Constraints of the problem include the weather and whether the day is a work day or a holiday. The day of the week can be a determinant in whether the bikes are heavily or lightly used.

For this capstone, the deliverables include: A GitHub repository containing the work completed for each step of the project, including: a slide deck, and a project report: The deliverables will be code, a project report, and a slide deck located in a GitHub repository.