Final Project DS210 Report Ibrahim Quraishi

* The goal was to find clusters of days in the bike sharing dataset and use the centroids as representatives of common weather days and bike sharing usage. For example, we are looking to find a cluster of “mild days” around a certain number of average daily users or “bad weather days” with another smaller number of average daily users.
* Unfortunately, the kmeans clustering failed when run in rust, possibly due to a memory error? I couldn’t successfully get the code to compile and struggled heavily with formatting the csv correctly for the algorithm.
* The kmeans clustering would have divided the data into a set number of clusters “k”. The initial cluster centers are randomly chosen before the first iteration. Then points are assigned to clusters based on their minimum squared distance to the centers. New centers are chosen by averaging the points of each new cluster and the algorithm is repeated until a chosen endpoint, in this case it would be maximum iterations of 500.
* The goal of the project was to perform this kmeans clustering on the weather data, hopefully revealing groups of similar weather days. This combined with the count of ridesharing users could give a picture of how different types of weather affect people’s choice to bike for the day. For example, we could find that cold, rainy days have many less bikers than warm sunny days.
* This could be used by lawmakers to determine if a city’s biking infrastructure is flexible toward changing weather conditions. If the clusters don’t vary much in rider count, then the city’s infrastructure can probably adequately support bikers in detrimental conditions.
* The data was found at <https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset> by UCI’s machine learning repository and contains daily count of rental bikes between 2011 and 2012 in capital’s bikesharing system in Washington DC with associated weather data such as windspeed, temperature, humidity..etc.
* Although a kmeans clustering was not successful and a graph couldn’t be analyzed, during data exploration in python a multiple regression was performed to determine if weather was a good predictor of bikesharing ridership.
* Cities that have a strong correlation could be identified as needing increased biking infrastructure spending, while cities that don't have a strong correlation could be looked at as examples for building all-weather infrastructure.
* The variables considered in the regression are cnt as the dependent variable and weathersit, atemp, hum, and windspeed as independent variables. Weathersit is 1 for clear/light clouds, 2 for mist/light clouds, 3 for light precipitation and 4 for heavy precipitation. atemp is in celcius and is the normalized feeling temperature. hum is the normalized humidity. windpeed is the normalized winspeed, and cnt is the count of total rental bikes including casual and registered.
* A screenshot of a computer

  Description automatically generated with low confidence
* There is a strong correlation between each independent variable and count of bikesharing users. This can be seen in the very small p values of each coefficient. There is also a somewhat high adjusted R^2 of 0.461 which makes sense, as there are many other outside factors to a city's daily bikesharing users such as the economy, infrastructure, construction, fitness...etc. This means the model is somewhat good at predicting bikesharing users.