ECE 20875 Mini Project Report

Name: TJ Bielefeld
Purdue Username: tbielefe@purdue.edu
Path: #2

Dataset Description:

The data used in this project comes from a CVS file named, "nyc_bicycle_counts_2016", which represents the number of people who used a bike on a specific date in 2016. The data spans from April 1, 2016 to October 31, 2016, and includes a variety of metrics on each day including the date, day of week, high temperature, low temperature, precipitation, and total and individual bridge traffic levels (Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge, Queensboro Bridge).

Methods and Analyses:

Problem 1:

To determine the most effective sensor deployment strategy, linear regression models were trained using each possible combination of three bridges as predictors to estimate total daily bicycle traffic. There are 4 total configurations and for each configuration, the coefficient of determination (R^2) was computed to measure predictive performance. The combination with the highest R^2 value was selected as the optimal sensor configuration, since a higher R^2 indicates that the model explains a greater proportion of the variance in total traffic while minimizing the number of required sensors.

Problem 2:

A multiple-variable linear regression model was trained to estimate daily total bicycle traffic using weather variables in the data as predictors: high temperature, low temperature, and precipitation. Then, the model performance was evaluated using R^2, and regression coefficients were used to assess the influence of each variable on ridership patterns. Furthermore, a large negative coefficient would indicate the number of riders would decrease drastically for a single unit increase in a predictor. For example, it should be expected that an increase in 1 inch of precipitation would decrease the total number of riders that day. So, using linear regression, it would be expected that the coefficient connected to precipitation would be large and negative. Additionally, the correlation function was used to ensure accuracy and further draw conclusions about the predictors and their relationship with total bicycle traffic.

Problem 3:

Weekly traffic patterns were analyzed by grouping total bicycle counts by day of the week and computing the daily mean. Days were then ranked according to their average mean riders to help identify high and low traffic trends. To predict the weekday based on a single observed traffic total, a nearest-mean classification method was used in which the observed count was assigned to the weekday whose daily average was closest. Creating a bar chart of the average means is another way to visualize the data and see trends such as whether weekdays or weekends have higher traffic and then draw conclusions on NYC's commuter or recreational biking.

Results:

Problem 1:

3-Bridge Combinations with R^2 Value:

Combination	R^2
Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge	0.99704
Brooklyn Bridge, Manhattan Bridge, Queensboro Bridge	0.99417
Brooklyn Bridge, Williamsburg Bridge, Queensboro Bridge	0.97977
Manhattan Bridge, Williamsburg Bridge, Queensboro Bridge	0.98730

Four three-bridge linear regression models were evaluated. The model using Brooklyn Bridge, Manhattan Bridge, and Williamsburg Bridge achieved the highest performance with an R^2 score of 0.99704, indicating that this sensor combination explains more than 99.7% of the variance in total bicycle traffic. The other combinations exhibited lower predictive accuracy, with R^2 values of 0.99417 (Brooklyn, Manhattan, Queensboro), 0.9873 (Manhattan, Williamsburg, Queensboro), and 0.97977 (Brooklyn, Williamsburg, Queensboro). These results demonstrate that excluding Queensboro Bridge causes the smallest reduction in prediction accuracy relative to the full dataset.

The consistently high R^2 values across all sensor combinations suggest that bicycle traffic across the bridges is strongly correlated, however, there exists a marginally superior performance of the model excluding Queensboro Bridge. This indicates that it contributes the least unique information toward predicting total traffic.

Ultimately, based on regression performance comparisons, the recommended sensor configuration is to deploy sensors on Brooklyn Bridge, Manhattan Bridge, and Williamsburg Bridge. This configuration achieves an R^2 value of 0.99704, representing near-perfect predictive accuracy while using only three sensors.

Problem 2:

Weather Linear Regression Model with Coefficients:

R^2	High Temp	Low Temp	Precipitation
0.49946	390.91831	-162.32008	-7951.48638

Correlation Matrix of Linear Regression Parameters:

	High Temp	Low Temp	Precipitation	Total Traffic
High Temp	1.000000	0.917376	-0.052069	0.574179
Low Temp	0.917376	1.000000	0.040390	0.442149
Precipitation	-0.052069	0.040390	1.000000	-0.420711
Total Traffic	0.574179	0.442149	-0.420711	1.000000

The weather regression model achieved an R^2 value of 0.499, indicating that approximately 50% of daily variation in bicycle traffic can be explained by weather conditions alone. The coefficient for high

temperature was positive (+390.9), demonstrating increased bike traffic on warmer days. Precipitation produced a strong negative coefficient (−7,951.5), confirming that rainfall significantly reduces bicycle traffic. This was explained above and expected result held true. The coefficient for low temperature was negative, which likely results from multicollinearity with high temperature since the two variables are closely correlated.

These results indicate that weather variables are moderate predictors of daily bicycle traffic. While temperature and precipitation have meaningful influence on traffic, the moderate R^2 demonstrates that weather alone cannot fully determine daily traffic trends. Non-weather factors such as weekday schedules and events also contribute to total bicycle traffic.

Using correlation analysis supported the regression findings. High temperature exhibited a moderate positive correlation with total traffic (r = 0.57), while low temperature also demonstrated a positive relationship (r = 0.44). Precipitation showed a moderate negative correlation with ridership (r = −0.42), confirming its inverse effect on bicycle use. High and low temperature were very strongly correlated with each other (r = 0.92), which helps to furthermore prove the parameters are closely related and result in multicollinearity seen in the regression model.

Problem 3:

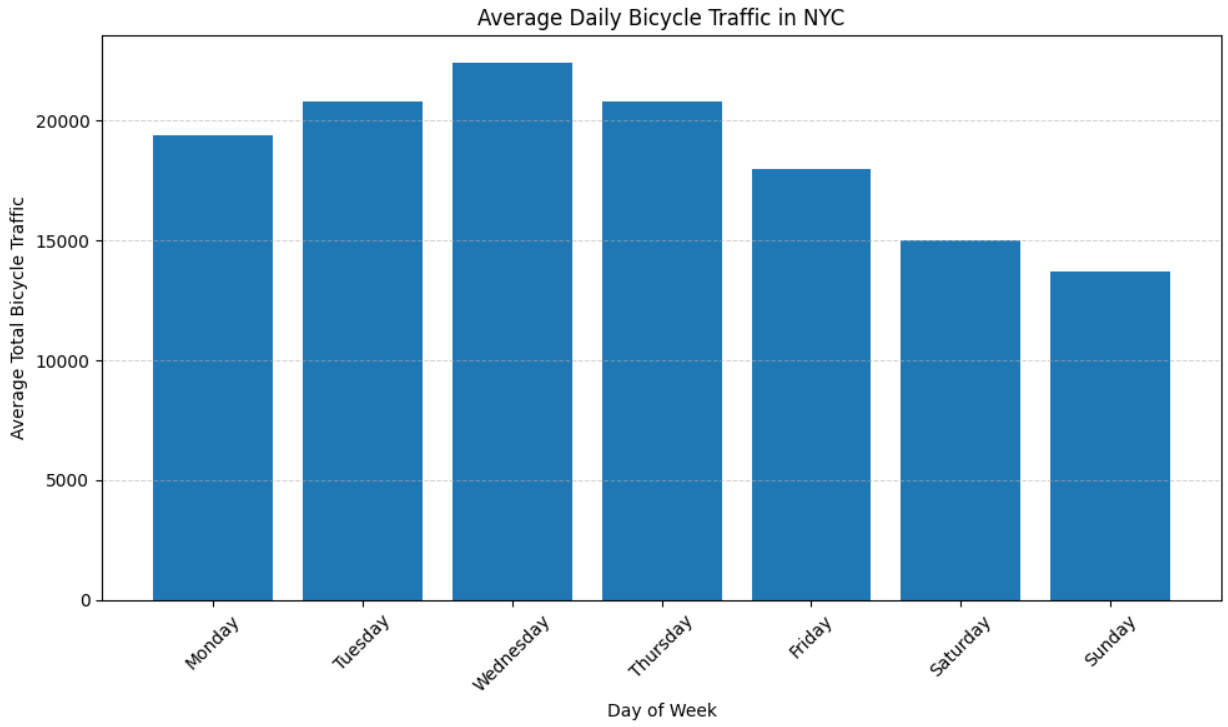

Figure 1: Average Daily Bicycle Traffic in NYC during 2026

Average Daily Bicycle Traffic by Day of Week:

Day of Week	**Average Bicycle Traffic**
Monday	19393.71
Tuesday	20782.27

Wednesday	22422.27
Thursday	20781.3
Friday	17984.58
Saturday	15000.65
Sunday	13716.39

Ranking of Highest Average Daily Bicycle Traffic by Day of Week

Rank	**Day of Week**	**Average Bicycle Traffic**
1	Wednesday	22422.27
2	Tuesday	20782.27
3	Thursday	20781.3
4	Monday	19393.71
5	Friday	17984.58
6	Saturday	15000.65
7	Sunday	13716.39

Weekend vs Weekday Average Traffic:

Part of Week	**Average Bicycle Traffic**
Weekend	14358.52
Weekday	20272.82

The highest average bicycle traffic occurred on Wednesday (22,422 riders), followed by Tuesday (20,782) and Thursday (20,781), which reflects peak mid-week commuter activity. The lowest traffic volumes occurred during the weekend, with Sunday averaging 13,716 riders and Saturday averaging 15,001 riders. Weekday ridership significantly exceeded weekend levels, with an average of 20,273 riders per weekday compared to 14,359 riders per weekend day, indicating that cycling activity is dominated by commuter behavior rather than recreational travel in New York City.

Using nearest-mean classification, daily traffic totals were mapped to their most likely weekday. As an example, an observed total of 18,000 riders was classified as Friday, whose historical average (17,985 riders) was the closest match. This method demonstrates that weekday traffic trends can be used to estimate the likely day of collection. However, overlap between daily ridership ranges limits the precision of single-day classification. This is apparent when looking at whether an observed total of 20,800 riders should be predicted to be a Tuesday (20,782) or Thursday (20,781), as these daily means are extremely close to each other and a nearest-mean classification would struggle to determine a correct day.