Grace Yeh ECE20875 – Distance education section

Chosen topic: Problem 1

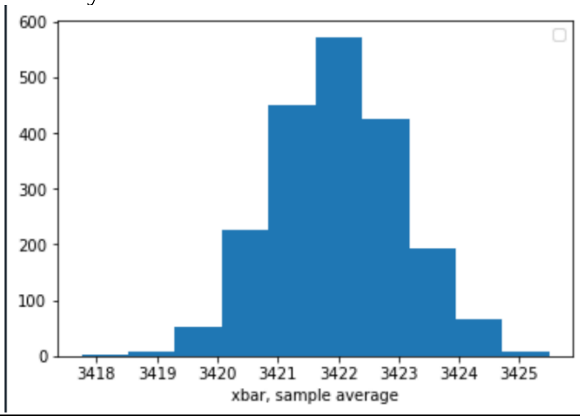
Description of data set:

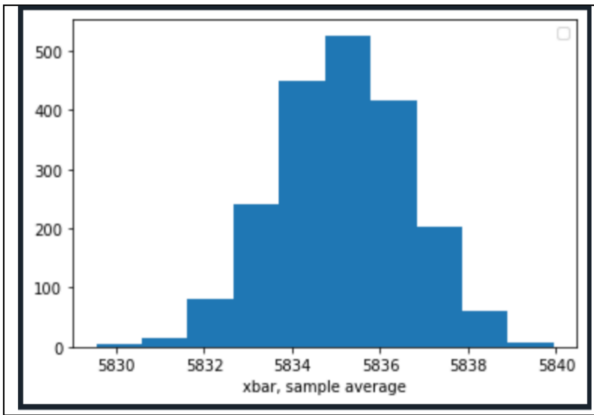
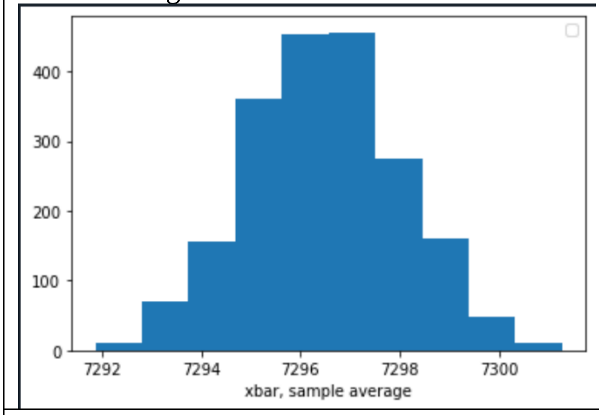
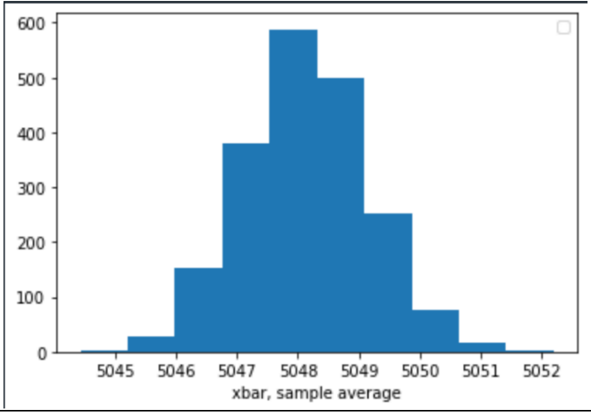
Data presents the number of bikes on 4 bridges daily across a span of 214 days. The following weather conditions are given: Maximum and minimum temperatures on each day, as well as the precipitation. Some of the data for precipitation is unrecorded, hence, I assumed unrecorded days to be 0.

Question 1 approach:

- For every bridge's bike flow data, extract the number of vehicles for these specific 2 days - Wednesday and Thursday
- Reason: A quick scan through the data showed that Wednesday and Thursdays had the most consistent traffic flow
- Conduct random sampling for each bridge
- The 3 bridges whose X bars could best be modelled by the normal distribution (lowest Mean Squared Error), are the most worthy of camera installation
- Reason: Bridges whose bike flow is best modelled by the normal distribution have the most consistent traffic flow across the days (less outliers), which would allow the machine learning algorithm in the sensors to determine a clear and undisputed mean bike flow for each day, rather than accounting for many outliers or positive/negatively skewed distributions

Results

Bridge	Mean Squared Error
Brooklyn 	1.12
Manhattan	2.26

	
Williamsburg 	2.35
Queensboro 	1.09

Question 1 Conclusion

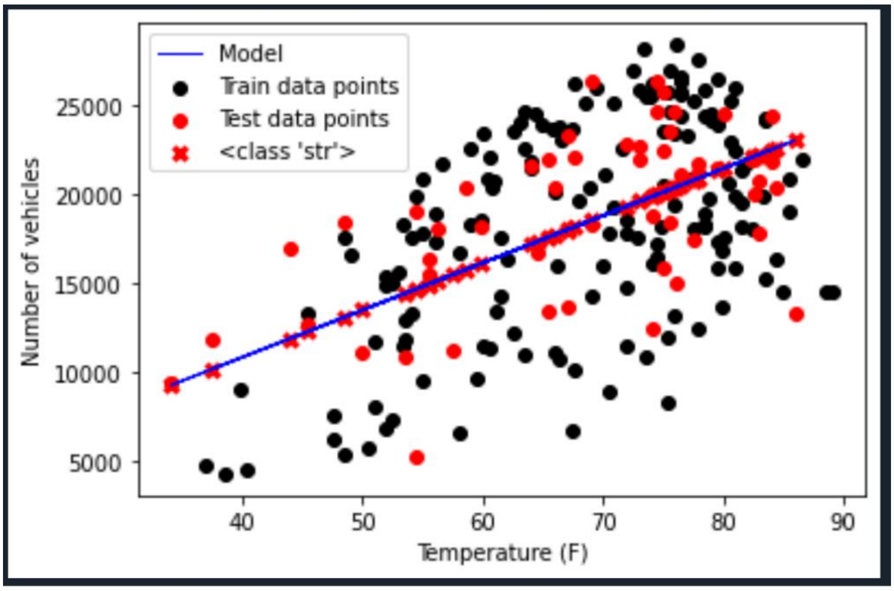
For a 10,000 sample size with 2000 repeats, the bridges which yielded the lowest Mean Squared Error are Brooklyn, Manhattan and Queensboro at 1.12, 2.26 and 1.09 respectively. Cameras should be installed at these 3 bridges.

Problem 2 approach:

- Analysis method: Conduct Linear Regression using the test-train machine learning analysis from sklearn library
- Reason: To predict the number of bikers based on the temperature, there has to be some kind of relationship between the independent and explanatory variable which allows an equation to be formed for prediction of future data points. I assume there is a linear relation between the variables
- X variable: Average daily temperature
- Y variable: Extracted data for total bikes on each day of the week

Results

Graph of Total vehicles against Temperature (F)



Equation of line	Coefficients [[265.02095194]] Intercept [279.55001838] Y = 265.02X + 279.55
MSE	15540403.994629279
Coefficient of determination	0.30509664515805823

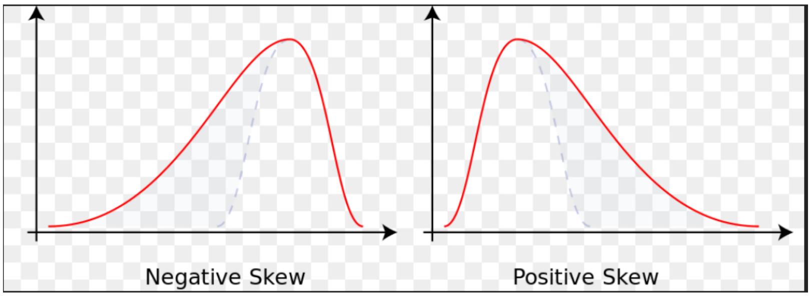
Conclusion:

Since the coefficient of determination is small at approximately 0.305, there is no linear relation between the temperature of the day and the total number of bikes on the road.

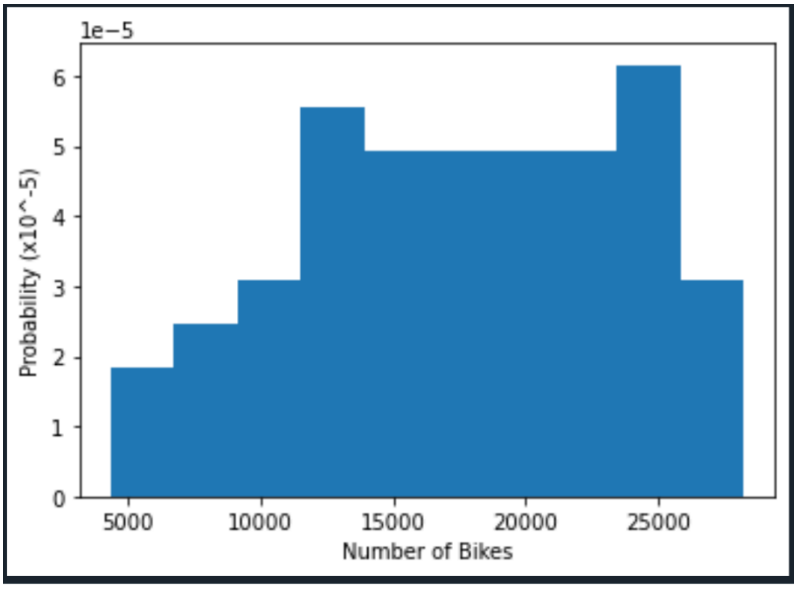
However, I do acknowledge that there could be a non-linear relation between temperature and the total vehicles on the road such as a quadratic/cubic/exponential relationship

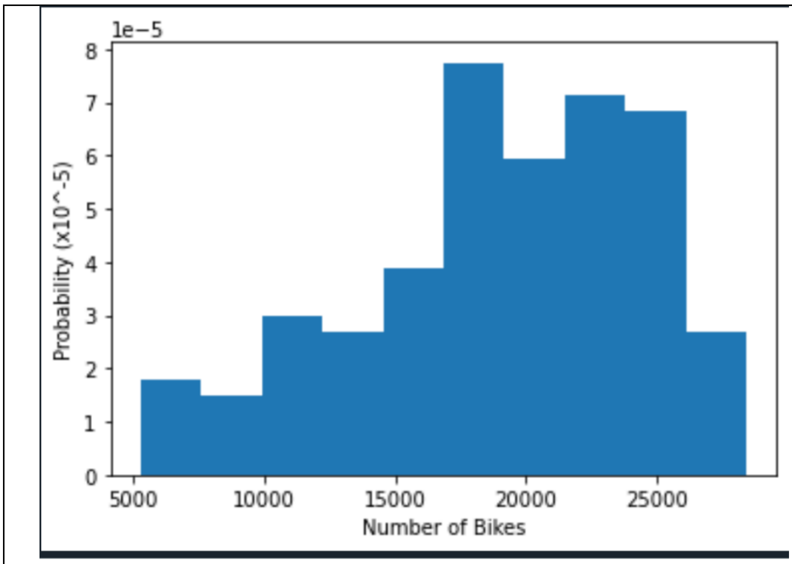
Problem 3 Approach:

- Separate the array containing total daily amount of traffic into 2 lists – no precipitation and precipitation
- Hypothesize that traffic would be less (<15000) on days with precipitation, meaning a positively skewed histogram
- Hypothesize that traffic would be higher (>15000) on days with precipitation, meaning a negatively skewed histogram
- Plot 2 probability histograms based of the 2 lists



<u>Results</u>

Histogram	Analysis
Histogram showing total traffic on days **with** precipitation  	Hypothesis that probability histogram would be positively skewed is false. On days with precipitation, the probability of >15000 bikes is still high at about 5 * 10^-5 percent
Histogram showing total traffic on days **without** precipitation	Hypothesis that probability histogram would be negatively skewed is True as seen in the hisrogram tending towards the right. The probabilities of bikes >15000 averages about 7 * 10 ^(-5). On days without precipitation, there is a higher

	probability of more than 15000 bikes on the road as opposed to the lower spectrum

<u>Conclusion:</u>

There is a lower probability (6e^-5) of >15000 bikes on the road on days with precipitation, while there is a higher probability (7e^-5) of >15000 bikes on the road on days without precipitation.

We cannot conclude that rainy days equate to less bikes since the histogram is not positively skewed so there is still a higher probability of >15000 bikes on the road than probability of <15000 bikes on a rainy day.

<u>References:</u>

NeilDANeilDA 78311 gold badge66 silver badges1717 bronze badges, & Rayryengrayryeng 94.5k1919 gold badges160160 silver badges172172 bronze badges. (1964, March 01). Estimating skewness of histogram in MATLAB. Retrieved December 04, 2020, from https://stackoverflow.com/questions/28056124/estimating-skewness-of-histogram-in-matlab