Draft 1 of our python group

Section 1. group information

Team member 1: Rongbo Hu

Member's email:hu763@purdue.edu

Member's Purdue username: Hu763

Team member 2: Jingyi Zhang

Member's email:zhan4129@purdue.edu

Member's Purdue username: Zhan4129

The project we choose : Path 1 Bike traffic

Section 2. Description of the dataset

We consider this dataset has 5 variables : Day, High temp , Low temp , Precipitation and Bridge

(1) *Day*

We find that in Monday, Saturday and Sunday, the average bike amount is larger than in other day,which means more people may go out at these three days. This make sense since at weekends many people do not need to work and want to go out and ride a bike to have fun. Also, some people live in different places during weekday (working/ study time) and weekends, so they need to ride the bike through the bridge in Monday and weekends to transfer themselves into these two places, , we will discuss detail in further draft.

(2) High Temp

The highest temperature in one day is 96.1˚F . It may affect people’s tendency to go out. From this chart we learn that when the highest temperature is higher, people tend to be more willing to go out and take a bike , we will discuss detail in further draft.

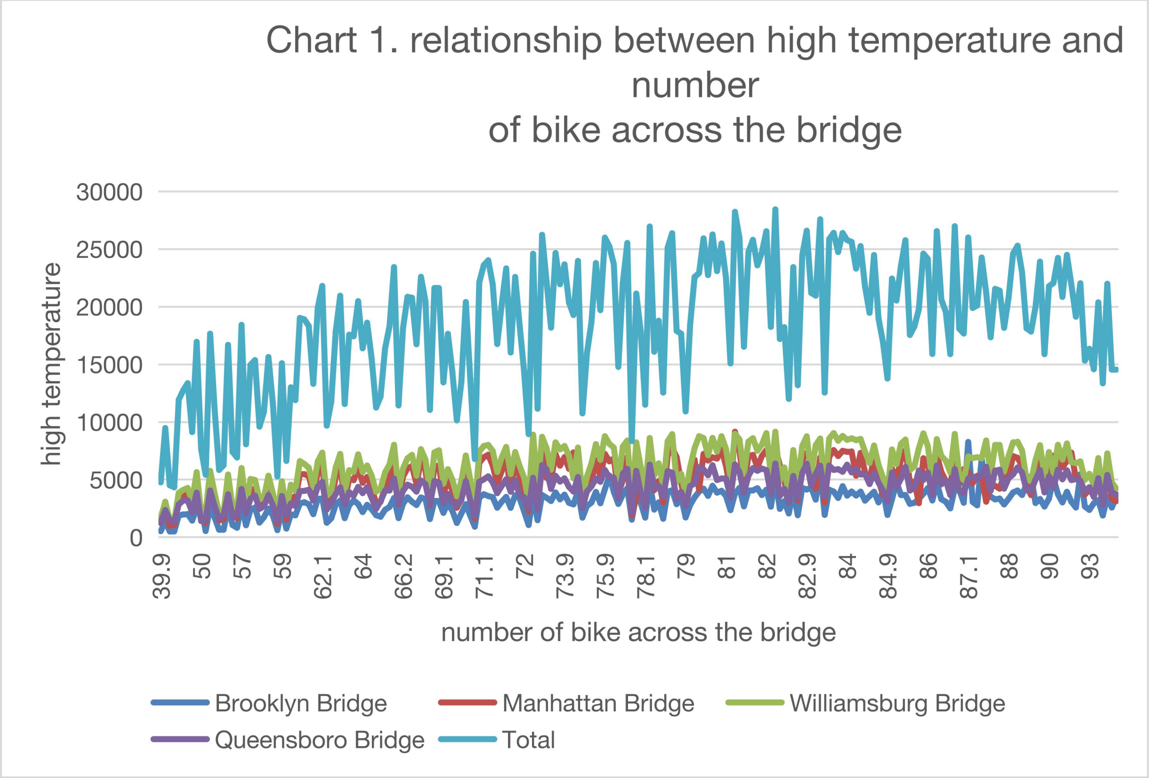


Chart 1 relationship between high temp and number of transport across the bridge

(3)Low Temp

The lowest temperature in one day in 26.1˚F. It may affect people’s tendency to go out. From this chart 2 we learn that when the lowest temperature is higher, people tend to be more willing to take a bike , we will discuss detail in further draft.

By combing high temp and low temp, we may use average temp or temp difference in further draft, but average temp and temp difference can be calculated from low and high temp so I do not list them as variables in draft0 now.

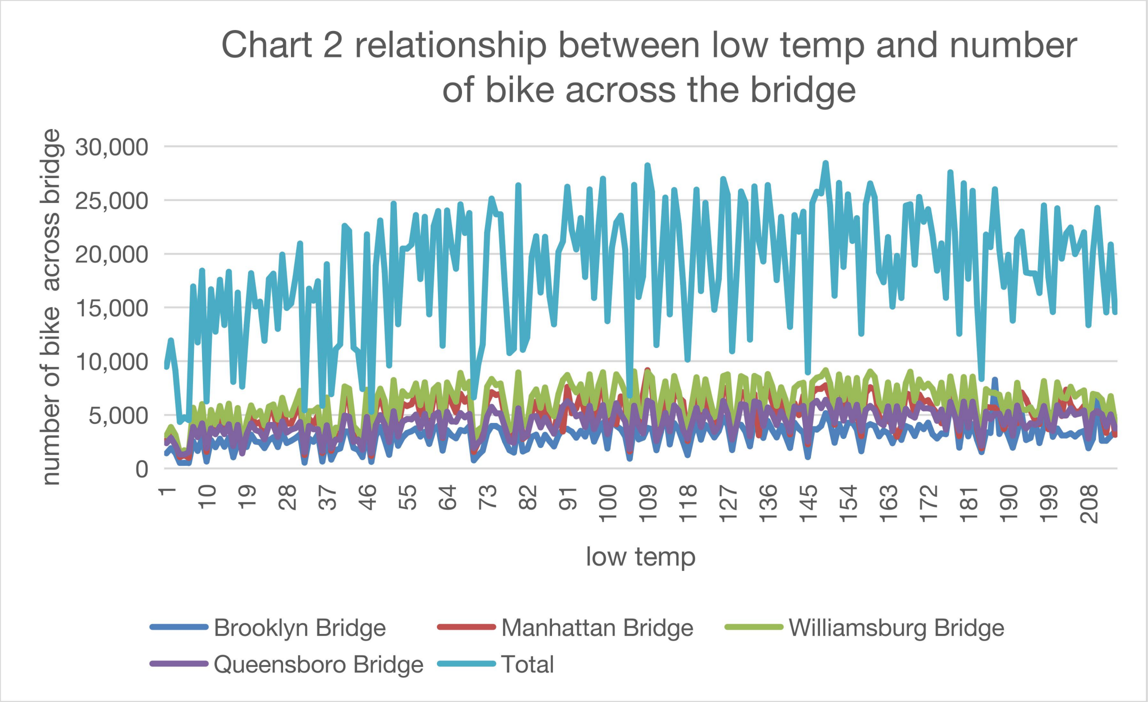


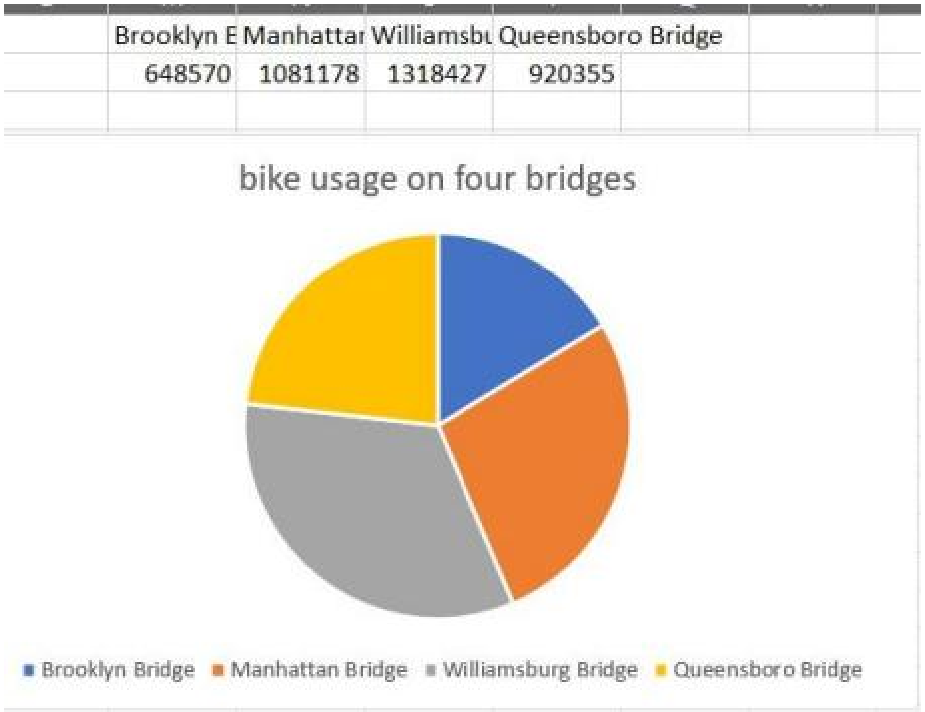
Chart 2 relationship between low temp and number of transport across the bridge

(4) Precipitation

Precipitation is raindrop height (in inch). It may affect people's tendency to go out. When I try to find relationship now, I find the range of precipitation can be related to different kinds of bike across amount, but in rainy days people should be less willing to face the rain and take a bike outside, so we will discuss the detail in further draft.

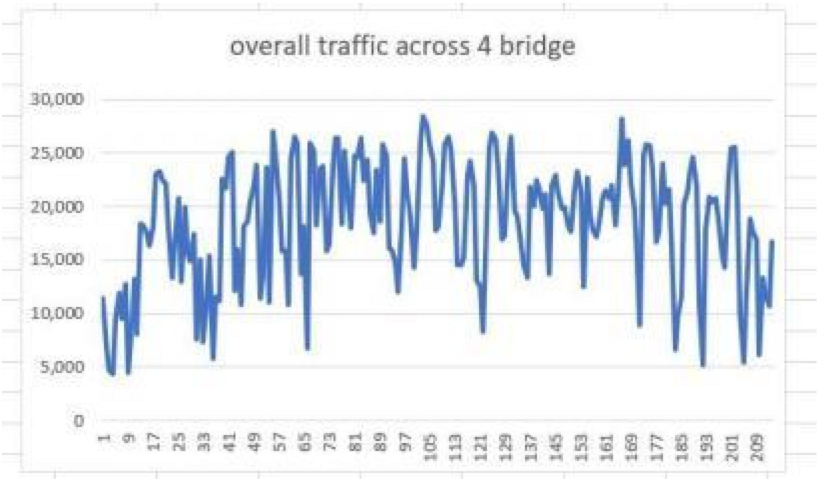
(5) Bridge

The bridge between Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge, and Queensboro Bridge visual that helps describe. Different bridge may need different calculating model.



Brooklyn E	Manhattar	Williamsbu	Queensboro Bridge
648570	1081178	1318427	920355

This picture compares the number of people who use bridges, we can find that the Williamsburg bridge has the most overall traffic across it.



The overall traffic across 4 bridges shows some regular pattern. We thought it was because of people's choice to ride a bike on weekdays or weekend

Section 3. method

Overall,we will take python and sklearn model as our tool to do analysis for each question. Here is the explanation for why we use python and sklearn model . Firstly, sklearn model allow array as input and returns list containing train-test split of input, which fits our design requirement (we take a list of temp, precipitation day and bridge as input and we hope output can be seperate into train and test).Next, skllearn model provides some useful functions. What we will use in this bike traffic question are “train_test_split ”(in which we can split arrays or matrices into random train and test subsets, we will use this is q1) and “linear regression” (in which we can get a linear relationship from datasets we want, we will use this in q2 and q3). What’s more, by using this sklearn model, we can train multiple models , these models help us solve the three questions we hope to solve.

We think the ideal response should be both deterministic and stochastic . To realize the deterministic portion, we will try to do the linear regression and find a linear regression relationship. To realize the stochastic portion (both random and unpredictable), we will calculate r square and get residual plots.

Here is how we combine sklearn model into our analysis for each question. For question 1, we will calculate the R square for traffic of each bridge to decide which three bridges should we choose to install sensors. By importing “train_test_split”into sklearn model, we will set bike number of each bridge as input and train four distribution models which represent Rsquare for bike number of each bridge. By comparing these models, we can decide which bridge should we cancel for installing sensors. We will also get residual plots to evaluate the accuracy of model.

For question 2, we will use multiple variable linear regression method to find a linear regression relationship between weather forecast data (low/high temperature and precipitation) and the total number of bicyclists. With the help of sklearn model, we will set {high temp, low temp , precipitation} as x dataset and {bike number} as y dataset. Then using “lineargression” in sklearn model, we will get linear regression relationship between x and y dataset. Then we can put next day’s weather forecast into this linear regression and predict the total number of bicyclists that day.

For question 3, we will do linear regression of number of bikes from Monday to Sunday based on known statistics to predict what day(Monday to Sunday) is today based on the number of bicyclists on the bridges. Using skearn model, we will do a regression for each week and get the regression plots. By finding the similarity for these plots, we will get a linear regression between bike number and weekday to confirm what day it is .

Section 4. Result

Q1:You want to install sensors on the bridges to estimate overall traffic across all the bridges. But you only have enough budget to install sensors on three of the four *bridges. Which bridges should you install the sensors on to get the best prediction of overall traffic?*

A1: We try to calculate the R square to decide which three bridges we should install. Through python calculation we can get that if we choose "Brooklyn", "Manhattan", "Williamsburg" bridges as our case , the R square will be largest and near to 1, so we would choose these three bridges to install sensors and give up Queensboro bridge.

Here is the lamda for case we choose:

0.215443469

Here is the R square we get for the four cases:

Brooklyn, Manhattan, Williamsburg (Score without Queensboro): r-squared - 0.995728908804532 99.5728908804532 %

Brooklyn, Manhattan, Queensboro (Score without Williamsburg): r-squared - 0.9881826020010465 98.81826020010465 %

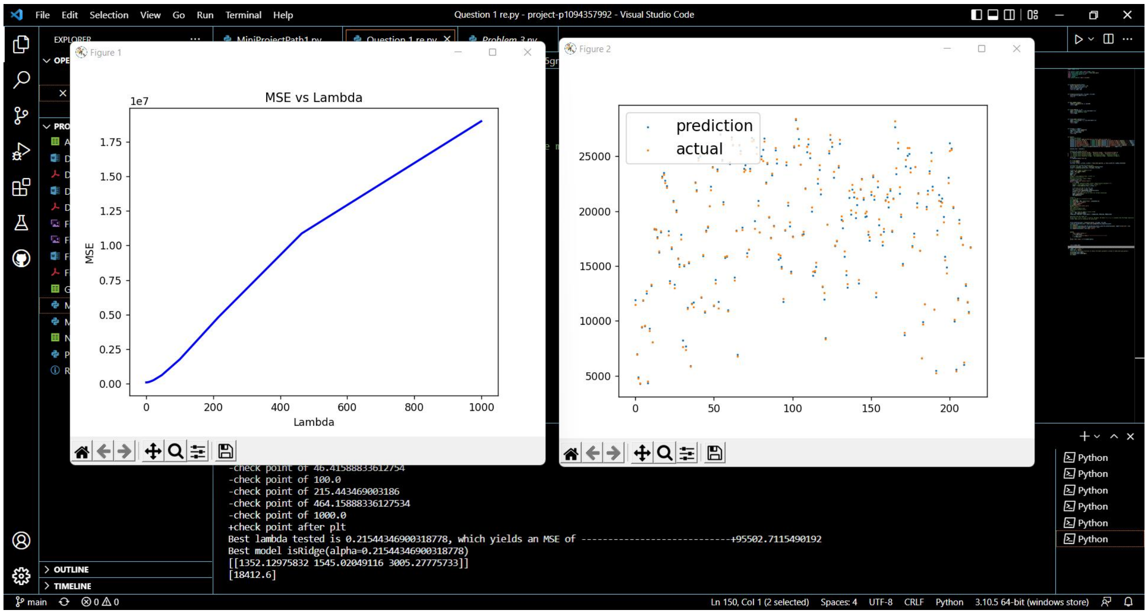
Brooklyn, Williamsburg, Queensboro (Score without Manhattan): r-squared - 0.9472389433167309 94.72389433167308 %

Manhattan, Williamsburg, Queensboro (Score without Brooklyn): r-squared - 0.9821507190448353 98.21507190448354 %

Here is the regression relationship we get between bike through bridges and bike number:

Total Traffic = 1352.1297582(Brooklyn Traffic) + 1545.02049*(Manhattan Traffic) +3005.2777*(Williamsburg Traffic) +18412.6*

Here is the figure show distribution plot of bike number for each bridge.



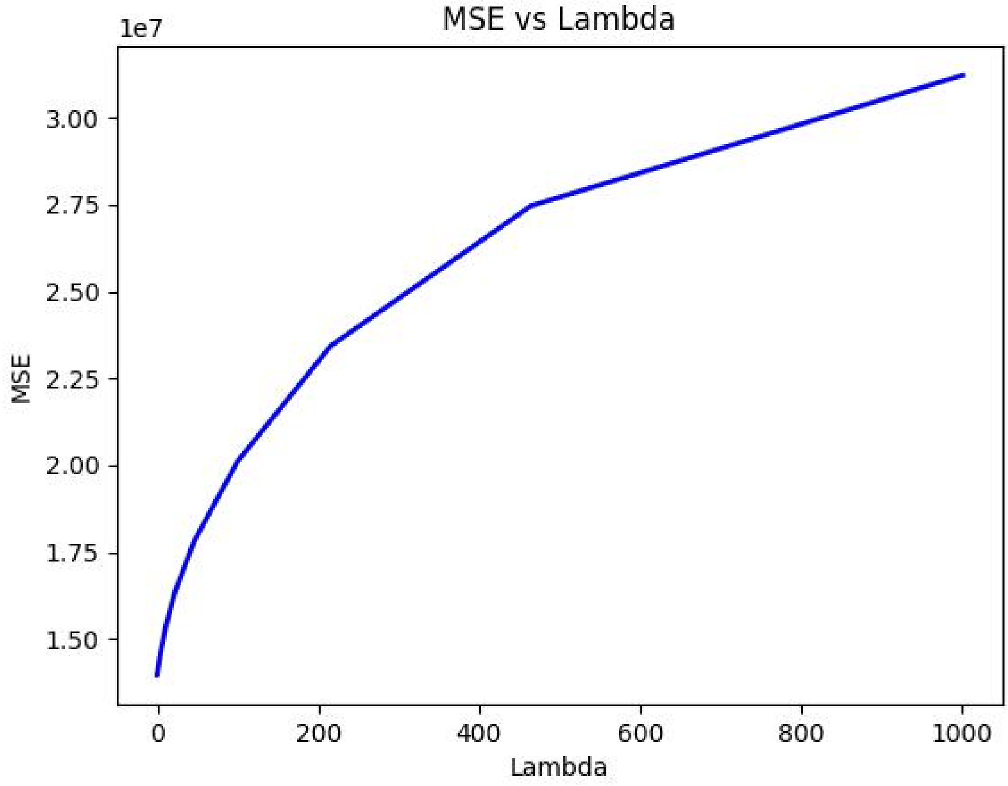
```
-check point of 46.41588833612754
-check point of 100.0
-check point of 215.443469003186
-check point of 464.15888336127534
-check point of 1000.0
+check point after plt
Best lambda tested is 0.21544346900318778, which yields an MSE of ---------------------------+95502.7115490192
Best model isRidge(alpha=0.21544346900318778)
[[1352.12975832 1545.02049116 3005.27775733]]
[18412.6]
```

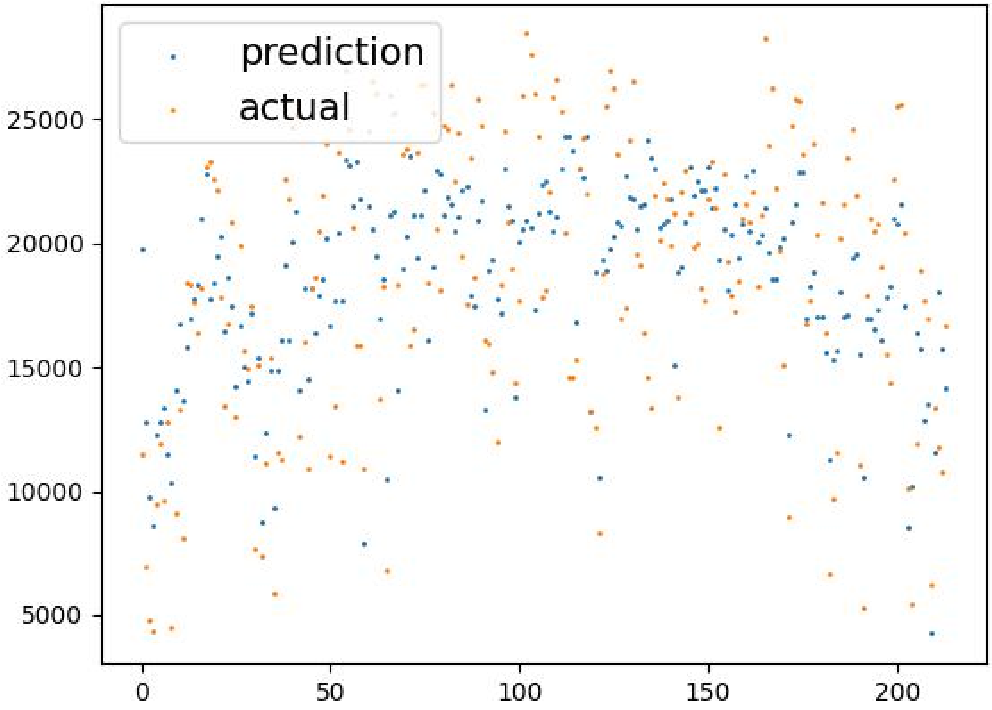
Q2:The city administration is cracking down on helmet laws, and wants to deploy police officers on days with high traffic to hand out citations. Can they use the next day's weather forecast(low/high temperature and precipitation) to predict the total number of bicyclists that day?

A2: We try to find a linear regression model , using low/high temperature and precipitation as x factors (x1 x2 x3) and bike number as y factor .Through python coding , we find the linear regression relationship:

Total Traffic = 406.17916154797587(High Temp) + -170.7826243577951*(Low Temp) + -8034.912567405353*(Precipitation) + -422.8260224649057*

The following plot shows this linear regression relationship. With this linear regression relationship, it is convenient for us to take next day’s weather forecast into linear regression relationship directly to get the total number of bike that day.





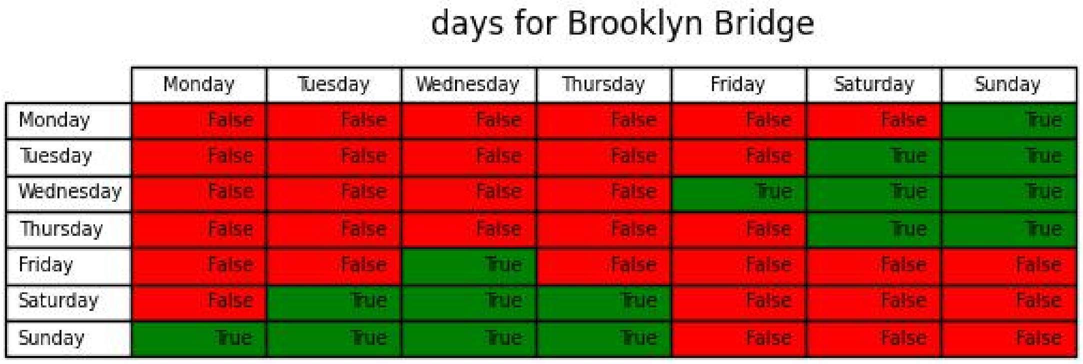
Q3:Can you use this data to predict what day(Monday to Sunday) is today based on the number of bicyclists on the bridges?

A3:We are looking forward to finding a significant difference between days, so we can use The Two-Sample t-Test to find if there is any significant difference between the two days. We make a hypothesis test with a significant level is 0.05.

H_0: The two variables do not have a significant difference. This means the μ_1 = μ_2

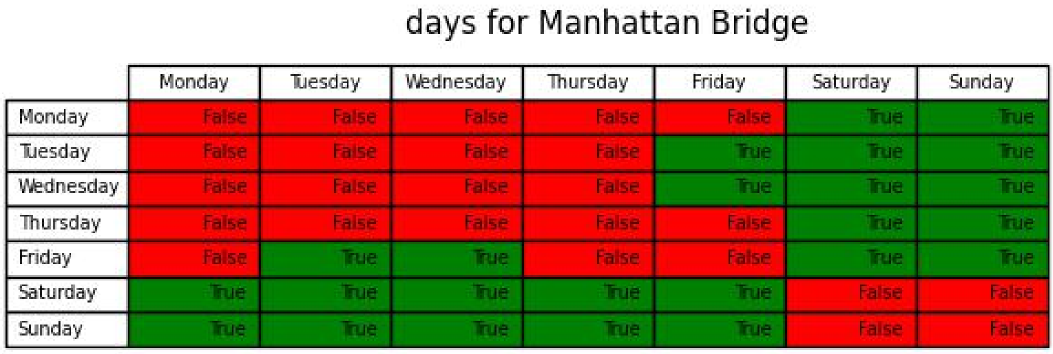
H_a: The two variables have a significant difference. This means the μ_1 ≠ μ_2

In the following table, each cell shows one hypothesis test. The green box means we have statistical evidence to reject H_0 and accept H_a – the two variables have a significant difference. The red box means we cannot reject H_0, the two variables do not have a significant difference.



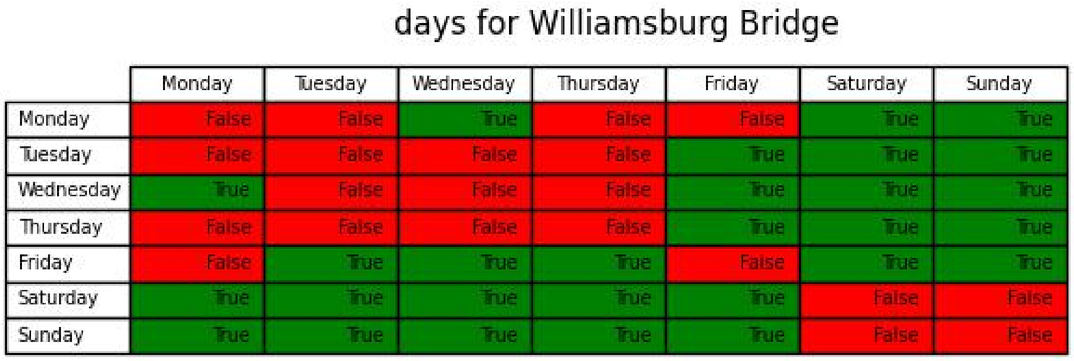
days for Brooklyn Bridge

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Monday	False	False	False	False	False	False	True
Tuesday	False	False	False	False	False	True	True
Wednesday	False	False	False	False	True	True	True
Thursday	False	False	False	False	False	True	True
Friday	False	False	True	False	False	False	False
Saturday	False	True	True	True	False	False	False
Sunday	True	True	True	True	False	False	False

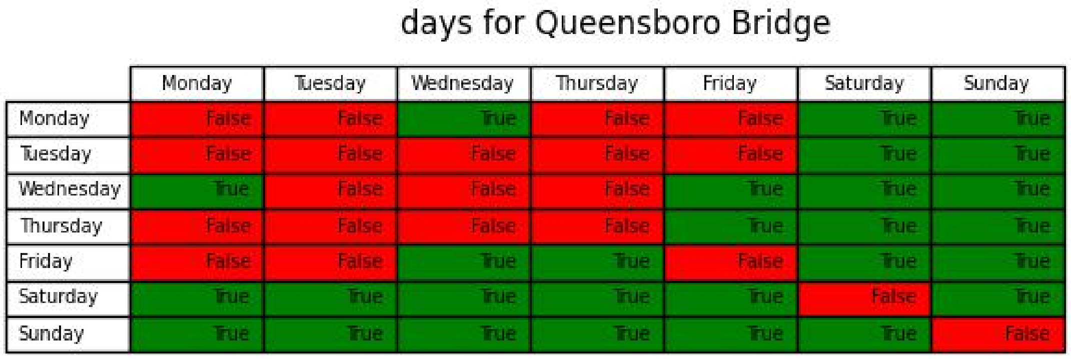


days for Manhattan Bridge

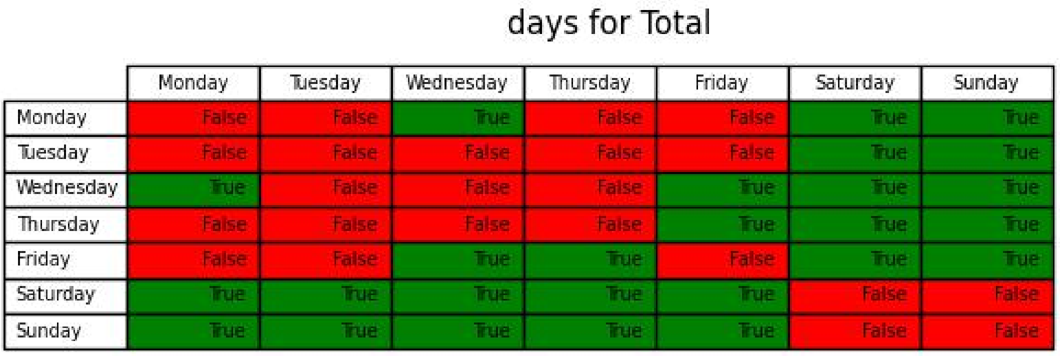
	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Monday	False	False	False	False	False	True	True
Tuesday	False	False	False	False	True	True	True
Wednesday	False	False	False	False	True	True	True
Thursday	False	False	False	False	False	True	True
Friday	False	True	True	False	False	True	True
Saturday	True	True	True	True	True	False	False
Sunday	True	True	True	True	True	False	False


days for Williamsburg Bridge

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Monday	False	False	True	False	False	True	True
Tuesday	False	False	False	False	True	True	True
Wednesday	True	False	False	False	True	True	True
Thursday	False	False	False	False	True	True	True
Friday	False	True	True	True	False	True	True
Saturday	True	True	True	True	True	False	False
Sunday	True	True	True	True	True	False	False


days for Queensboro Bridge

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Monday	False	False	True	False	False	True	True
Tuesday	False	False	False	False	False	True	True
Wednesday	True	False	False	False	True	True	True
Thursday	False	False	False	False	True	True	True
Friday	False	False	True	True	False	True	True
Saturday	True	True	True	True	True	False	True
Sunday	True	True	True	True	True	True	False


days for Total

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Monday	False	False	True	False	False	True	True
Tuesday	False	False	False	False	False	True	True
Wednesday	True	False	False	False	True	True	True
Thursday	False	False	False	False	True	True	True
Friday	False	False	True	True	False	True	True
Saturday	True	True	True	True	True	False	False
Sunday	True	True	True	True	True	False	False

For those 4bridges, we can see the pattern that there is no significant difference between days in the workday or the days on the weekend (the 4*4 red box in the lower right and nearly 5*5 red box in the top left). But weekends and workdays always show a significant difference.

In conclusion, we cannot use this data to predict what day (Monday to Sunday) is today based on the number of bicyclists on the bridges, but we can predict the weekend and workday.