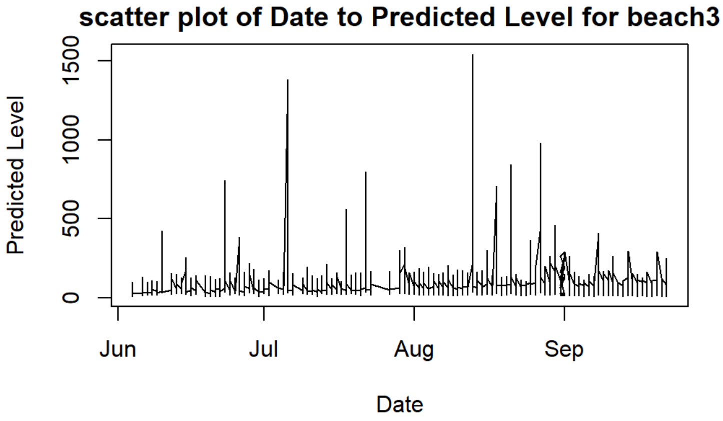
Data Science - “E.Coli levels” Project Paper

What dates created the highest and lowest records of E.Coli in the summer of 2016 and out of the beaches in the data set which have the lowest and highest over all E.Coli levels? This data set from data.cityofchicago.org involves tracking E. Coli levels in the summer of 2016 at various beaches around chicago. E.Coli are commonly found in warm-blooded organisms Most E.Coli strains are harmless, but some strains can cause serious food poisoning and will require professional medical attention. Despite E Coli being commonly found in warm blooded mammals it can also live in bodies of water that have not been filtered, like lakes. With hundreds of people who go to the beaches each summer it is easy to not be weary of this annoying bacteria which can ruin your summer plans. That's why I figured out at what time in the summer are e coli the largest and at which beach so you don't have to.

My first step was to divide the main data set up into different sub datasets via the beach name so that I could see each beach’s stats over the course of the summer. Then to better see the main trend I wanted to focus on (E.Coli levels throughout the summer) I made line plots for the main data set; the y & x axis were the data and the Predicted level of E.Coli. The second step I did was to make a line plot of all the subset beach data via their predicted level of e coli and the date that that information was recorded on; so I could see the trends combined throughout all the beaches. To further my inquiries I made a joint bar graph displaying the predicted level for all the beaches and whether or not a swim advisory was issued in order to convey the minimum value required to get a Swim Advisory issued. A swim advisory is an advisory given out when the e coli levels are above the threshold of 230-240. It issues an instruction saying to not place your head below or drink the water. To finalize my conclusions I identified the highest record of e coli out of all the beaches and located which beach it was at then compared it to the levels at the same time at the other beaches. As well as found the lowest levels of e coli, its designated beach, and looked at the other beaches at the same time to find the safest beach to go to.

Plotting each beach with Predicted Levels of E.Coli and the dates for said prediction led me to discover that overall each beach has higher predicted levels of e coli toward the end of the summer, except for 2 outliers. This Is backed by osterman who has the lowest record of e coli out of all the beaches which was recorded on the very first day of the records. As Well as the main line plot showing that the average level of E.Coli rose compared to the first half of summer.

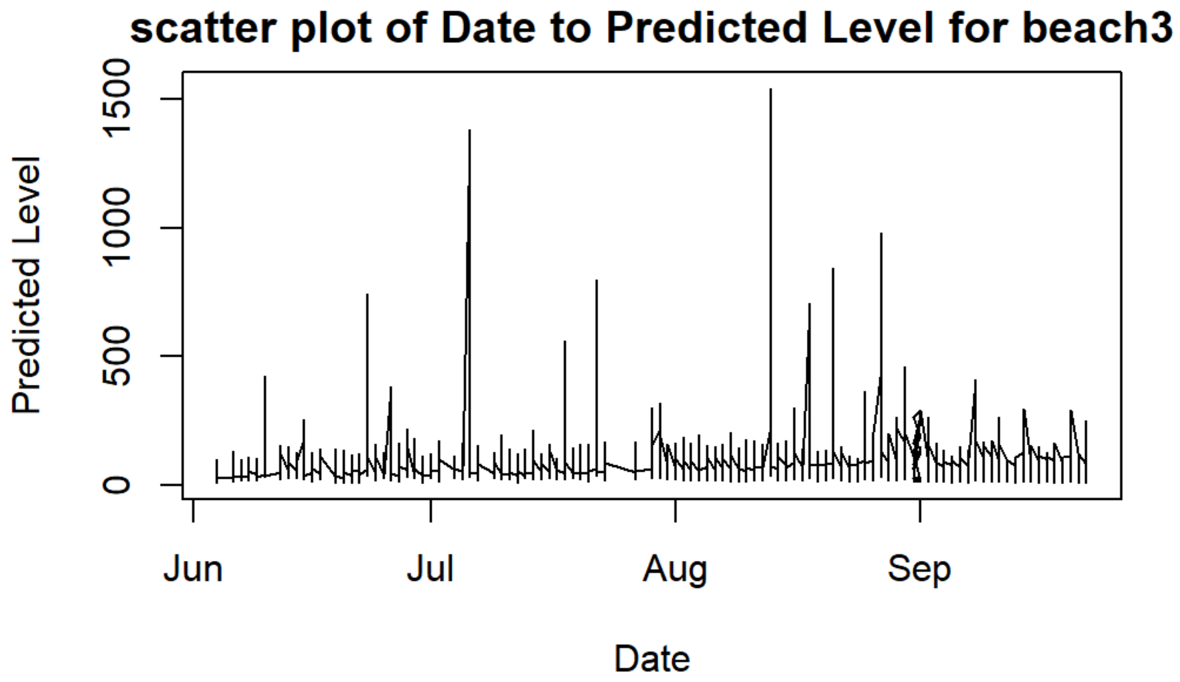


I also discovered that the threshold of 230-240 in order to get a swim advisory issued does not have any higher levels as in a requirement of getting out of the water if the E.Coli levels are above a certain amount above 230-240. For example the highest recorded level of predicted coeli at a beach was: 1542.5 and the probability that there was going to be this many e coli in the area was: 0.908. Despite these outrageously high levels of the summer-plan-ruining-bacteria the beaches still didn't shut down. How could these annoying ecoli be avoided when no one, let's be honest, bothers to check the e coli levels of a beach before going there other than very worried parents?

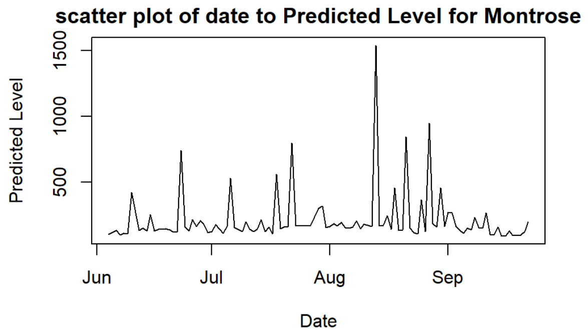
Eureka! A partial solution to this problem: at least out of the beaches that I have been given the data for (Montrose, Ohio, 63rdStreet, Osterman, Calumet, Foster, Leone, Rainbow, and OakStreet) I can tell you that Montrose has the highest levels of e coli in the beginning, the highest recorded level of e coli in the middle of the summer and osterman had the highest levels at the end of the summer, which is shocking because osterman had the lowest predicted levels out of all the beaches in the beginning of the summer. So in conclusion the beach that you should at all costs avoid this summer is montrose. The best and safest time of the summer to go to most beaches would be in the beginning of the swim season. If you prefer the end of summer then the safest beaches out of the ones in the “Beach” data set would be Ohio Street and Oak Street beaches with ohio street having the lowest recorded predicted level at the end of the summer and OakStreet beach having the lowest overall E.Coli levels out of all the beaches with never going above 45. This project was really fun to do and I learned useful information that I wouldn't have otherwise. I got to confirm some things that I thought to be true and I got to be surprised by new information. If I were to do another analysis like this I would try to do so for a data set that has no real world significance, for fun.

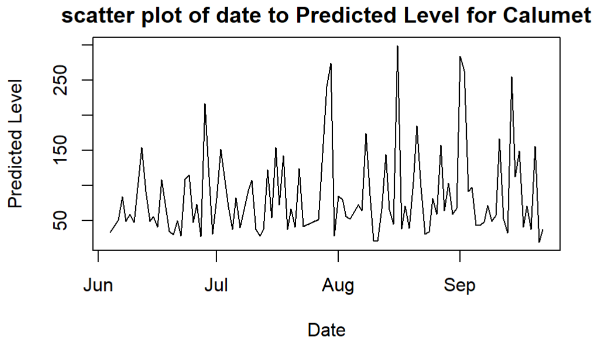
Further down are the Graphs and table.

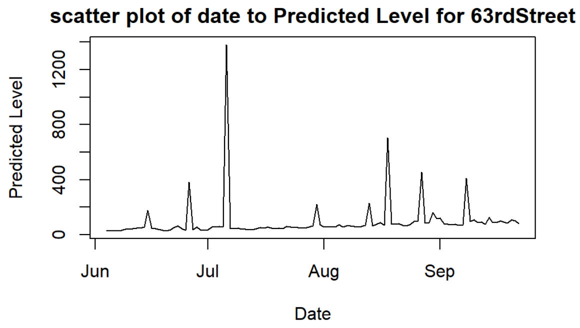
First step, dividing the beaches into their own subsets and graphing the main data set:

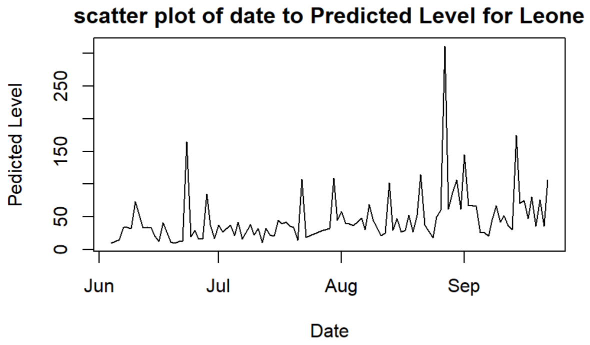

scatter plot of Date to Predicted Level for beach3

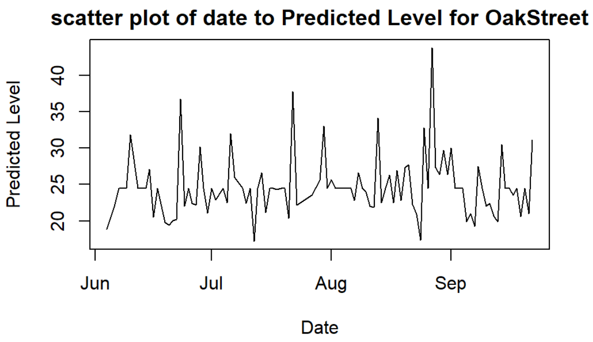
Second step, making the subsets into graphs to better visualize the necessary data:

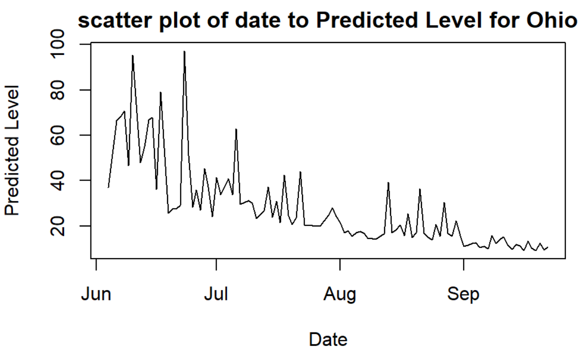

scatter plot of date to Predicted Level for Montrose

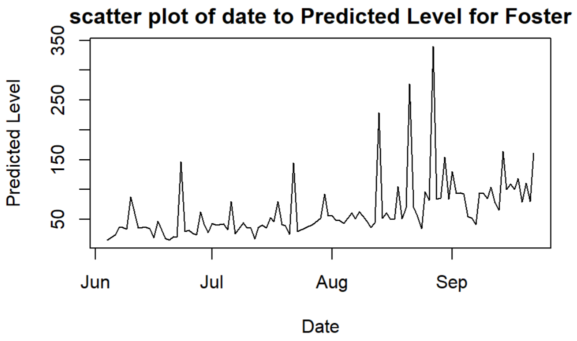

scatter plot of date to Predicted Level for Calumet

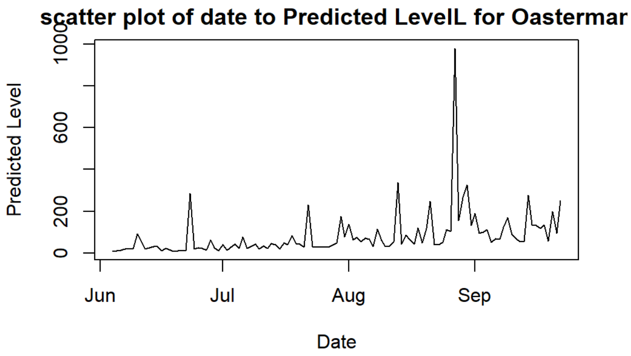

scatter plot of date to Predicted Level for 63rdStreet

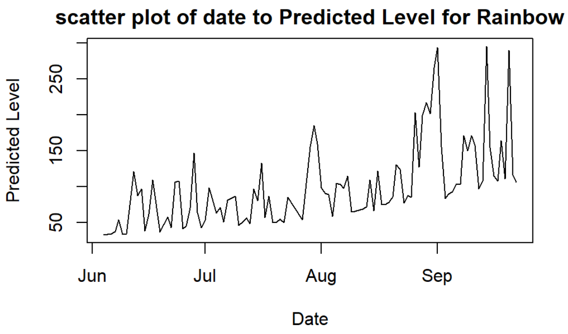

scatter plot of date to Predicted Level for Leone


scatter plot of date to Predicted Level for OakStreet
Predicted Level
20
25
30
35
40
Jun
Jul
Aug
Sep
Date

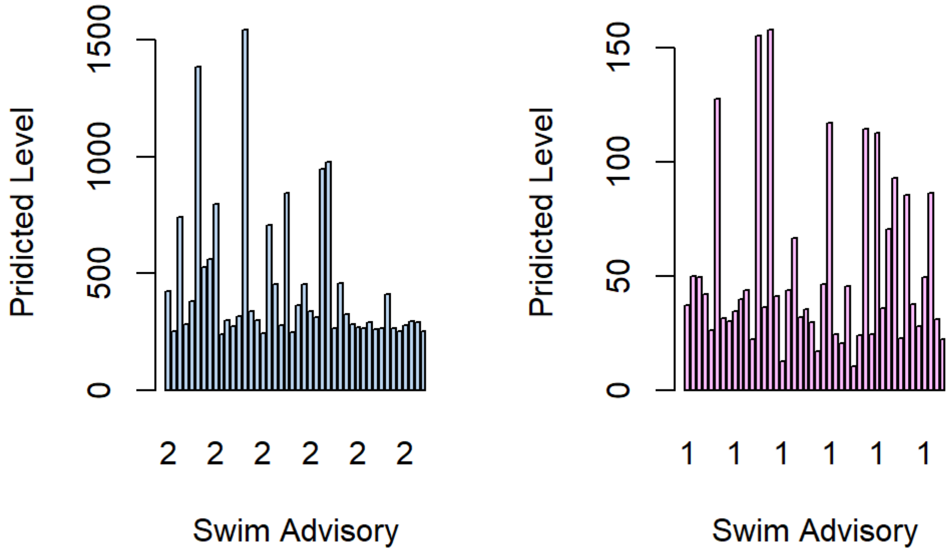

scatter plot of date to Predicted Level for Ohio
Predicted Level
20
40
60
80
100
Jun
Jul
Aug
Sep
Date


scatter plot of date to Predicted Level for Foster
Predicted Level
50
150
250
350
Jun
Jul
Aug
Sep
Date


scatter plot of date to Predicted LevelL for Oastermar
Predicted Level
0
200
600
1000
Jun
Jul
Aug
Sep
Date


scatter plot of date to Predicted Level for Rainbow
Predicted Level
50
150
250
Jun
Jul
Aug
Sep
Date

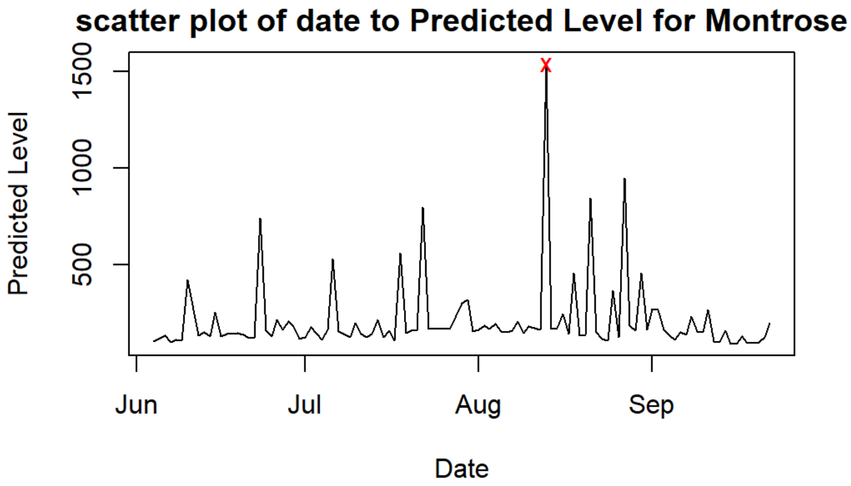
Third step, joint bar graph displaying the predicted level and whether or not a swim advisory was issued:



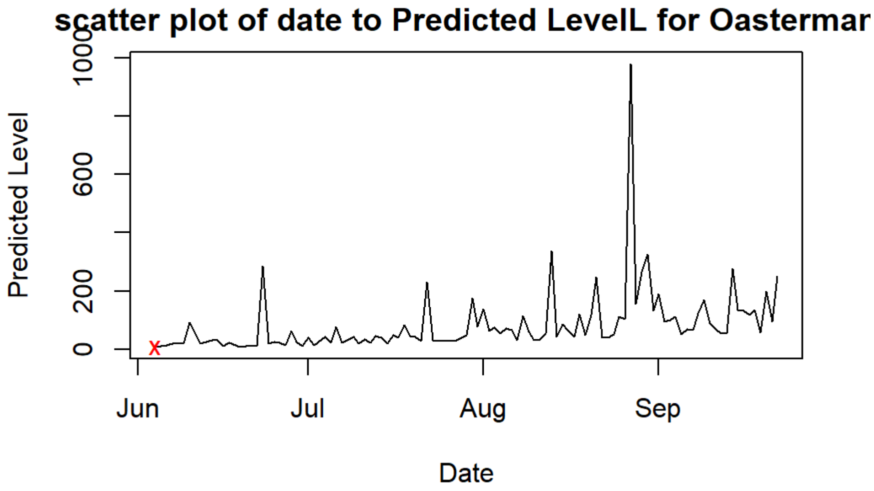
A summary of the Predicted Levels from main sub beach dat set:

Predicted Level	
Minimum	7.6
Median	51.40
Mean	83.51
Maximum	1542.5

Final step:
The beach with the highest level of E.Coli:



Beach with the lowest levels of E.Coli:


scatter plot of date to Predicted LevelL for Oastermar
Predicted Level
Jun
Jul
Aug
Sep
Date
x