Robert Hsu

Extra Credit Report

CMPSC 465

Professor Kamesh Madduri

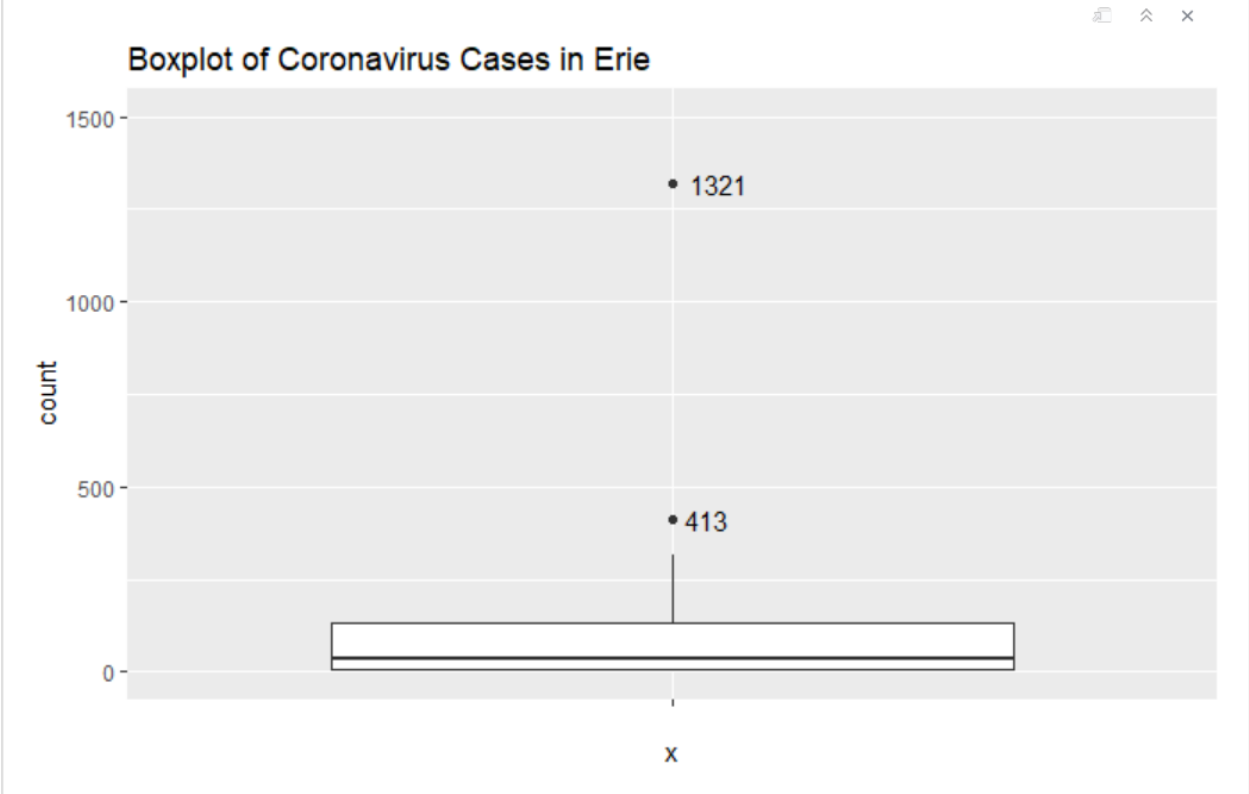
Erie, NY Findings Report

The data I used came from the following website:

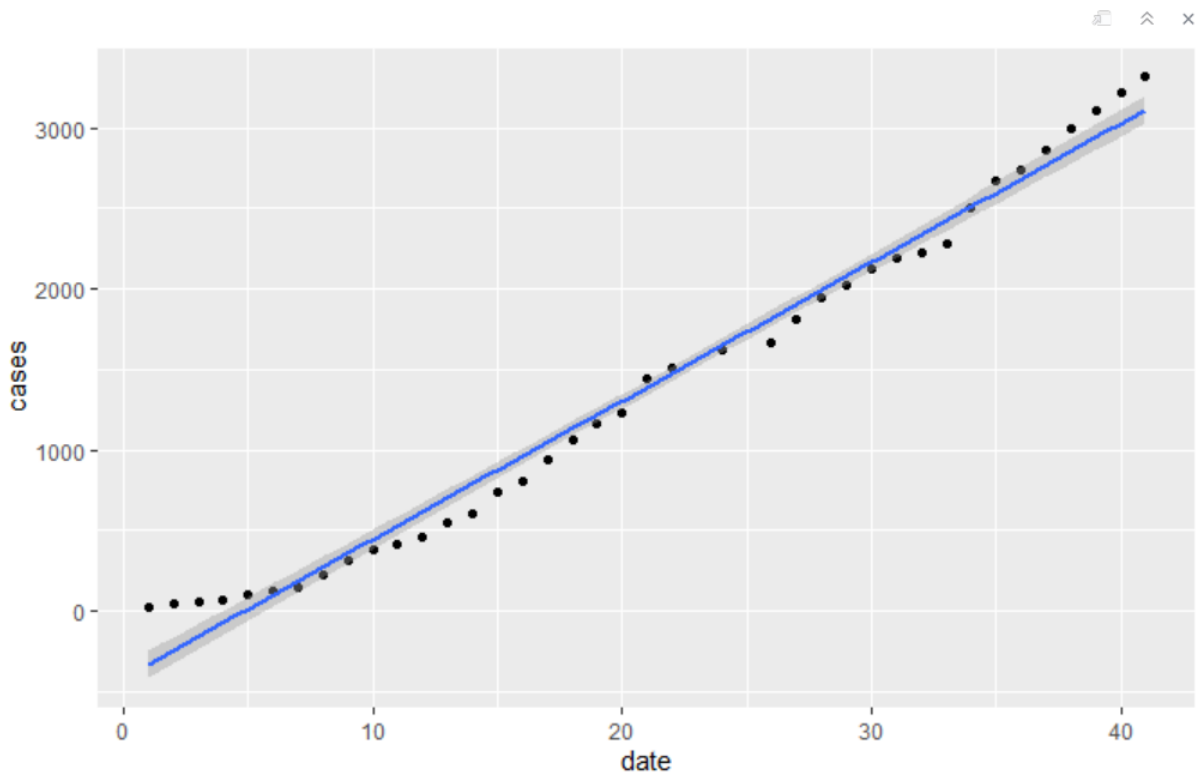
https://erieny.maps.arcgis.com/apps/opsdashboard/index.html#/dd7f1c0c352e4192ab162a1dfadc58e1

I used data as of 4/29 10 PM.

I did a boxplot test to determine if there are any municipalities that have significantly more cases than the others. As you can see datapoints 1321 and 413 are considered outliers. These datapoints correspond to Buffalo and Amherst/Williamville, respectively.



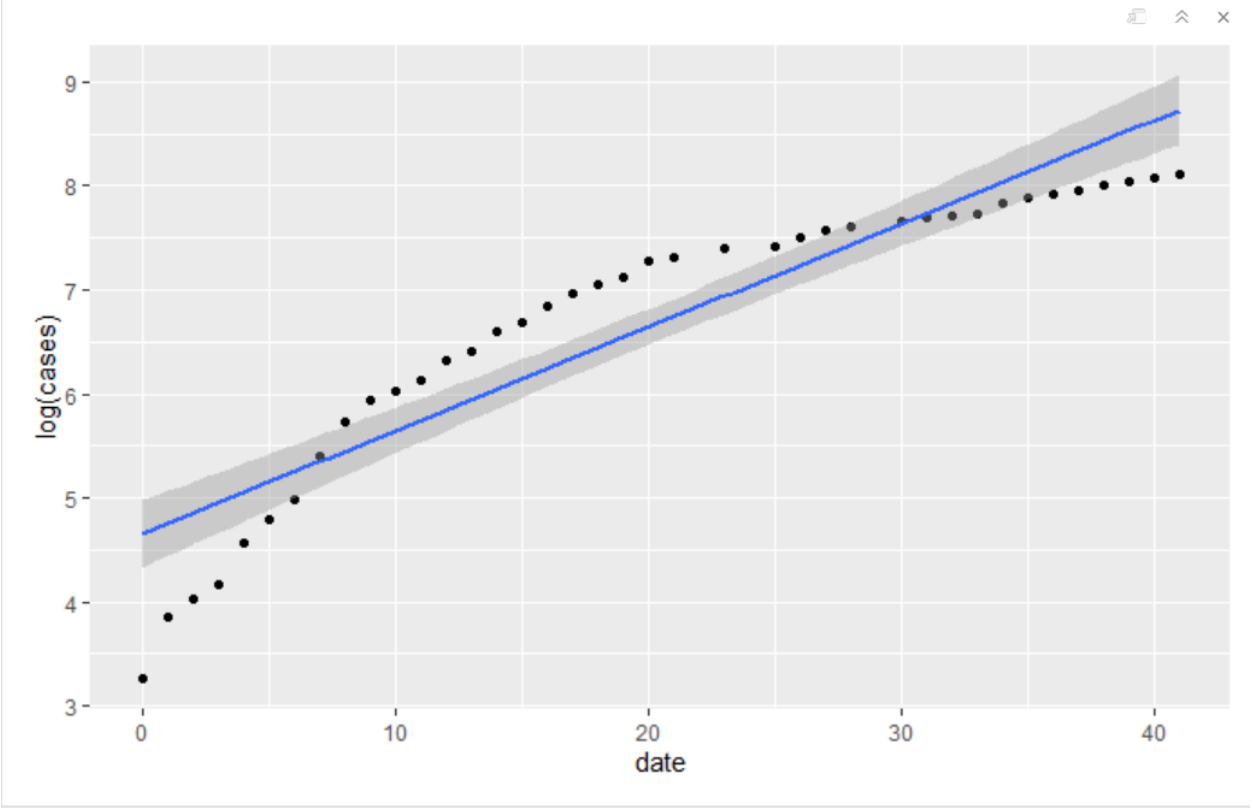
Pulling from the same website I decided to do an analysis on the daily increase of the cases in Erie.




```
Call:
lm(formula = cases ~ date, data = daily_increase)

Residuals:
    Min      1Q  Median      3Q     Max
-184.53  -98.99  -41.34   79.63  357.22

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -417.278     43.363  -9.623 1.29e-11 ***
date          86.058      1.799  47.845  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 136 on 37 degrees of freedom
Multiple R-squared:  0.9841,    Adjusted R-squared:  0.9837
F-statistic:  2289 on 1 and 37 DF,  p-value: < 2.2e-16
```

The data started from 3/20 which I have labeled as day 0 and ends on 4/30 which is day # 41. Three datapoints were missing, which were on 4/11, 4/13, and 4/18. The R^2 for the simple linear regression is 0.9841, which is really good fit. I used a simple linear regression model and determined that # of cases = -417.278 + 86.058 * number of days past.

To be sure I have the best model I also conducted an exponential linear regression model which is the following:



```
Call:
lm(formula = log(cases) ~ date, data = daily_increase)

Residuals:
     Min       1Q   Median       3Q      Max
-1.35149 -0.34646  0.06688  0.43191  0.60281

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.506482   0.161671   27.87   <2e-16 ***
date        0.103100   0.006706   15.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5071 on 37 degrees of freedom
Multiple R-squared:  0.8646,    Adjusted R-squared:  0.861
F-statistic: 236.4 on 1 and 37 DF,  p-value: < 2.2e-16
```

The equation for this model is log(cases) = 4.506482 + 0.103100 * days past

As you can see the R^2 is a lot lower than the previous model.

We can conclude that using the simple linear regression model to predict the data is the best approach.

The following is the data:

date	cases
0	26
1	47
2	56
3	64
4	96
5	121
6	146
7	221
8	310
9	380

10	414
11	463
12	553
13	603
14	734
15	802
16	945
17	1059
18	1163
19	1235
20	1440
21	1506
23	1624
25	1661
26	1809
27	1951
28	2023
30	2127
31	2192
32	2221
33	2284
34	2505
35	2671
36	2738
37	2857
38	2996
39	3109
40	3224
41	3315

Municipality	count
Buffalo	1321
Amherst	413
Sloan	318
Kenmore	198
Lancaster	170
Hamburg	158
Village	140
West_Seneca	132
Aurora	75
Clarence	60
Alden	48
Grand_Island	47
City_of_Tonawanda	37
Lackawanna	35
Elma	34
Concord	28
Evans	24
Newstead	22
Eden	9
Boston	8
Collins	8
Holland	7
North_Collins	6
Marilla	5
Brant	4
Colden	4
Wales	3
Sardinia	1