Analysis of Cyclist Counts at East River Bridge Locations

Part 1: Data Source & Scientific Questions of Interest

My data is from New York City Department of Transportation (NYC DOT). The Traffic Information Management System (TIMS) collects the count data, to keep count of cyclists entering and leaving Queens, Manhattan and Brooklyn via the East River Bridges. The data includes the date, precipitation, highest temperature, lowest temperature, and the number of cyclists via Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge and Queensboro Bridge. It can be used to uncover what factors may have an impact on cyclist counts and how do they work.

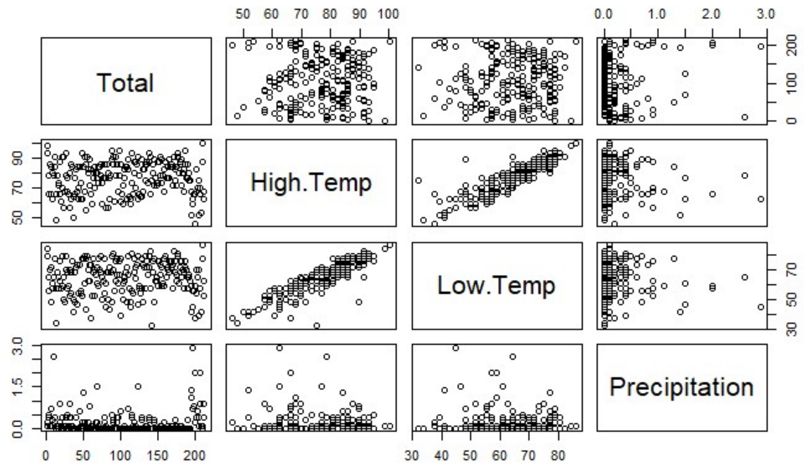
Part 2: Summary of Findings & Conclusions

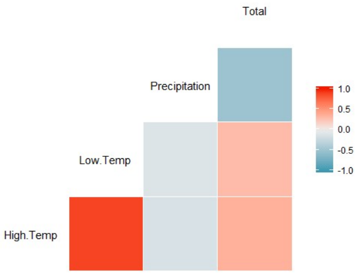
Through the analysis of correlation and Box Cox, I found that, by the large, the numerical variables are linearly correlated with cyclist counts. Cyclist counts have a positive correlation with temperatures, and a negative correlation with precipitations.

However, the dummy variables are more complicated, Through F-test, it's suggestive but inconclusive that there are some differences in weekdays or in weekends. On the one hand, in weekdays, we can't reject the hypothesis that means of cyclist counts in different weekdays are the same. On the other hand, by Tukey multiple comparisons of means, we found the difference of cyclists counts between Sunday and Saturday shouldn't be ignored. The cyclists count decrease as weekdays > Saturdays > Sundays. So, we would have 3 categories (weekdays, Saturday, Sunday), which means we would need 2 dummy variables.

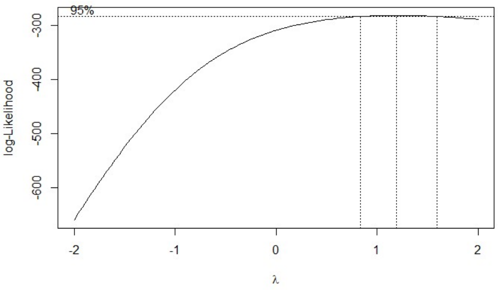
In this way, I constructed the model of cyclists counts via all East River Bridges, and also the models of cyclists count via each bridge mentioned above. The model can be used to predict the cyclists count of specific day with predicted precipitation and temperature from Weather forecast.

Part 3: Detailed Description

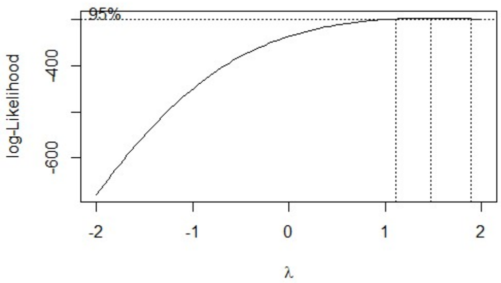


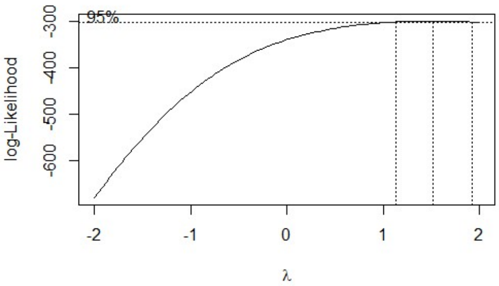


Correlation



boxcox(Total~Precipitation,data=bicycle)

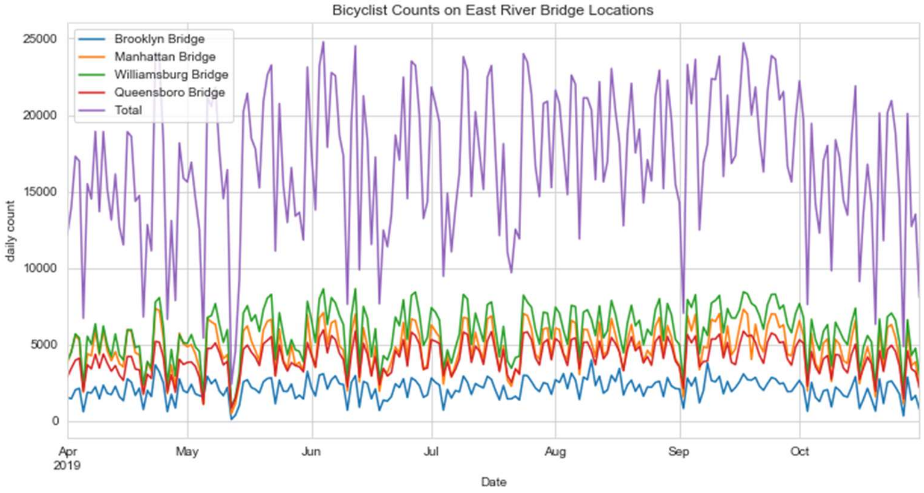




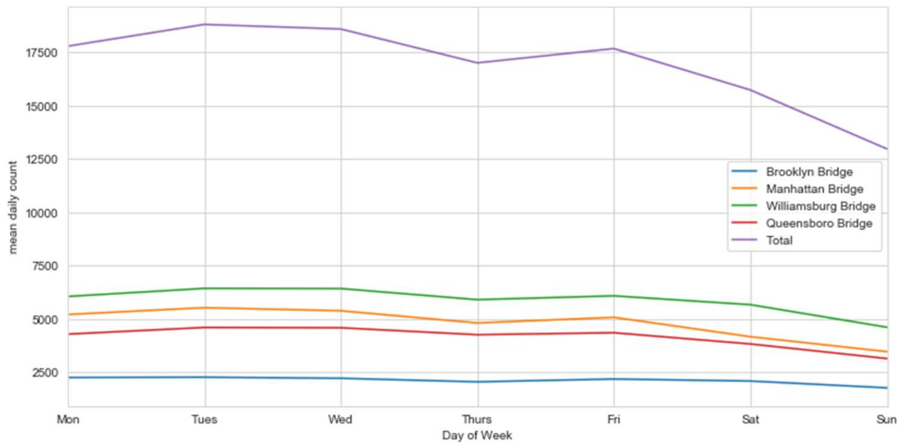
boxcox(Total~High.Temp,data=bicycle)

boxcox(Total~Low.Temp,data=bicycle)

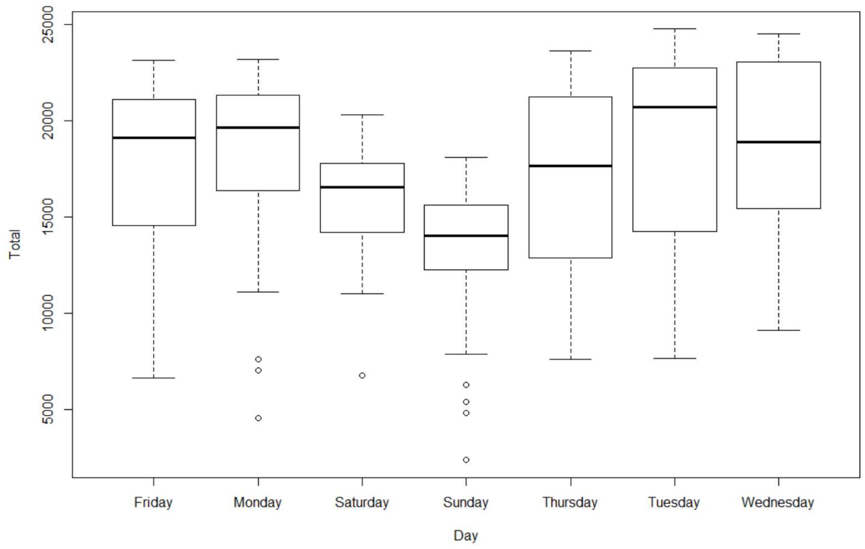
Through the analysis of correlation and Box Cox, I found that, by the large, the numerical variables are linearly correlated with cyclist counts. Cyclist counts have a positive correlation with temperatures, and a negative correlation with precipitations.



When we observe the cyclists count over time, we can see some periodical change. After some transformation:



More specifically:



The mean of cyclist counts seems to be different among different day of week. To figure out the truth, I did some inference:

Null Hypothesis 1: μ1 = μ2 = μ3 = μ4 = μ5 = μ6 = μ7

```
Analysis of Variance Table
Model 1: Total ~ 1
Model 2: Total ~ 0 + Day
   Res.Df        RSS Df Sum of Sq     F    Pr(>F)
1     213 4846732741
2     207 4099246022  6 747486720 6.291 4.305e-06
```

Apparently, bicyclist counts on East River Bridge locations are different among different day of a week. (p-value = 4.3e-06)

Null Hypothesis 2: μ1 = μ2 = μ3 = μ4 = μ5 , μ6 = μ7

```
Analysis of Variance Table
Model 1: Total ~ 0 + Weekend
Model 2: Total ~ 0 + Day
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    212 4279899964
2    207 4099246022  5 180653943 1.8245 0.1094
```

It's suggestive but inconclusive that the difference of bicyclist counts is caused by weekend. (p-value = 0.11)

Null Hypothesis 3: μ1 = μ2 = μ3 = μ4 = μ5

```
Analysis of Variance Table
Model 1: Total ~ Sunday + Saturday
Model 2: Total ~ 0 + Day
  Res.Df        RSS Df Sum of Sq     F Pr(>F)
1    211 4165464780
2    207 4099246022  4  66218759 0.836 0.5037
```

We can't reject Null Hypothesis 3, there is no difference between weekdays. (p-value = 0.50)

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = Total ~ weekfactor, data = bicycle)
$weekfactor
                      diff        lwr        upr     p adj
sunday-saturday  -2762.067 -5469.9221  -54.21125 0.0444270
weekday-saturday  2242.217   149.2665 4335.16814 0.0325047
weekday-sunday    5004.284  2911.3332 7097.23481 0.0000002
```

The difference of cyclists counts between Sunday and Saturday shouldn't be ignored. (p-value = 0.04)

The cyclists count decrease as weekdays > Saturdays > Sundays. So, we would have 3 categories (weekdays, Saturday, Sunday), which means we would need 2 dummy variables.

Because of the collinearity of High.Temp and Low.Temp, I use Stepwise AIC backward regression to construct model.

```
Start:  AIC=3498.51
Total ~ High.Temp + Low.Temp + Precipitation + Sunday + Saturday
               Df Sum of Sq        RSS    AIC
- Low.Temp      1  12743078 2559446298 3497.6
<none>                      2546703220 3498.5
- High.Temp     1 171931088 2718634308 3510.5
- Saturday        1 174007590 2720710810 3510.7
- Sunday          1 574212354 3120915574 3540.0
- Precipitation   1 920698684 3467401904 3562.5
Step:  AIC=3497.58
Total ~ High.Temp + Precipitation + Sunday + Saturday
                 Df Sum of Sq          RSS      AIC
<none>                          2559446298 3497.6
- Saturday        1 169462120 2728908418 3509.3
- High.Temp       1 461752324 3021198621 3531.1
- Sunday          1 567184449 3126630747 3538.4
- Precipitation   1 926907396 3486353694 3561.7

lm(formula = Total ~ High.Temp + Precipitation + Sunday + Saturday,
    data = bicycle)
Residuals:
     Min        1Q    Median        3Q       Max
-11009.1   -2022.5     780.4    2267.9    7963.9
Coefficients:
              Estimate   Std. Error  t value Pr(>|t|)
(Intercept)    8684.92    1715.76     5.062 9.09e-07 ***
High.Temp       134.31      21.87     6.141 4.08e-09 ***
Precipitation -5123.59     588.92    -8.700 9.93e-16 ***
Sunday        -4777.09     701.94    -6.806 1.04e-10 ***
Saturday      -2602.79     699.69    -3.720 0.000256 ***
Residual standard error: 3499 on 209 degrees of freedom
Multiple R-squared:  0.4719,    Adjusted R-squared:  0.4618
F-statistic: 46.69 on 4 and 209 DF,  p-value: < 2.2e-16
```

By Stepwise AIC backward regression, we get the model for cyclist count at East River Bridges:

$$Total = 8684.92 + 134.31 * \beta1 - 5123.59 * \beta2 - 2602.79 * \beta3 - 4777.09 * \beta4$$

β1: Highest Temperature
β2: Precipitation
β3: Saturday (1 for Saturday, 0 for others)
β4: Sunday (1 for Sunday, 0 for others)

Similarly, construct model for 4 bridges each.

Brooklyn Bridge:

```
lm(formula = Brooklyn.Bridge ~ High.Temp + Precipitation + Sunday +
    Saturday, data = bicycle)
Coefficients:
```

```
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      891.844      278.197     3.206 0.001558 **
High.Temp            18.613          3.546     5.248 3.76e-07 ***
Precipitation -655.637         95.489    -6.866 7.39e-11 ***
Sunday             -398.439       113.814    -3.501 0.000567 ***
Saturday          -155.376       113.448    -1.370 0.172289
Residual standard error: 567.4 on 209 degrees of freedom
Multiple R-squared:    0.3254,    Adjusted R-squared:    0.3125
F-statistic: 25.21 on 4 and 209 DF,    p-value: < 2.2e-16
```

Manhattan Bridge:

```
lm(formula = Manhattan.Bridge ~ High.Temp + Precipitation + Sunday +
      Saturday, data = bicycle)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)      3243.594      542.243     5.982 9.45e-09 ***
High.Temp             29.549          6.912     4.275 2.91e-05 ***
Precipitation -1600.281      186.120    -8.598 1.92e-15 ***
Sunday             -1641.682      221.839    -7.400 3.26e-12 ***
Saturday          -1122.365      221.126    -5.076 8.52e-07 ***
Residual standard error: 1106 on 209 degrees of freedom
Multiple R-squared:    0.4605,    Adjusted R-squared:    0.4502
F-statistic:    44.6 on 4 and 209 DF,    p-value: < 2.2e-16
```

Williamsburg Bridge:

```
lm(formula = Williamsburg.Bridge ~ High.Temp + Precipitation +
      Sunday + Saturday, data = bicycle)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)      2724.986      599.477     4.546 9.27e-06 ***
High.Temp             49.572          7.642     6.487 6.23e-10 ***
Precipitation -1778.629      205.765    -8.644 1.43e-15 ***
Sunday             -1495.640      245.254    -6.098 5.11e-09 ***
Saturday            -641.913      244.466    -2.626    0.00928 **
Residual standard error: 1223 on 209 degrees of freedom
Multiple R-squared:    0.4598,     Adjusted R-squared:    0.4494
F-statistic: 44.47 on 4 and 209 DF,    p-value: < 2.2e-16
```

Queensboro Bridge:

```
lm(formula = Queensboro.Bridge ~ High.Temp + Precipitation +
      Sunday + Saturday, data = bicycle)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)      1824.496      350.492     5.206 4.61e-07 ***
```

```
High.Temp            36.574         4.468      8.186 2.64e-14 ***
Precipitation -1089.040        120.303    -9.052    < 2e-16 ***
Sunday              -1241.329        143.391    -8.657 1.31e-15 ***
Saturday             -683.140        142.930    -4.780 3.31e-06 ***
Residual standard error: 714.9 on 209 degrees of freedom
Multiple R-squared:    0.555,       Adjusted R-squared:    0.5465
F-statistic: 65.18 on 4 and 209 DF,    p-value: < 2.2e-16
```

Model of cyclist count via each bridge:

$$B = 891.84 + 18.61 * \beta 1 - 655.64 * \beta 2 - 398.44 * \beta 3 - 155.38 * \beta 4$$

$$M = 3243{,}59 + 29.59 * \beta 1 - 1600.28 * \beta 2 - 1641.68 * \beta 3 - 1122.37 * \beta 4$$

$$W = 2724.99 + 49.57 * \beta 1 - 1788.63 * \beta 2 - 1495.64 * \beta 3 - 641.91 * \beta 4$$

$$Q = 1824.50 + 36.57 * \beta 1 - 1089.04 * \beta 2 - 1241.33 * \beta 3 - 683.14 * \beta 4$$

β1: Highest Temperature
β2: Precipitation
β3: Saturday (1 for Saturday, 0 for others)
β4: Sunday (1 for Sunday, 0 for others)
B: number of cyclists via Brooklyn Bridge
M: number of cyclists via Manhattan Bridge
W: number of cyclists via Williamsburg Bridge
Q: number of cyclists via Queensboro Bridge

Weichen Li
wl2726