A Bridge Too Far Notebook

Importing Libraries and Data Set:

```
# A tibble: 214 × 10
   date  day     HighTempF LowTempF precipitation brooklynbridge manhattanbridge
   <chr> <chr>       <dbl>    <dbl> <chr>                  <dbl>           <dbl>
 1 4/1   Saturd…      46       37   0.00                     606            1446
 2 4/2   Sunday       62.1     41   0.00                    2021            3943
 3 4/3   Monday       63       50   0.03                    2470            4988
 4 4/4   Tuesday      51.1     46   1.18                     723            1913
 5 4/5   Wednes…      63       46   0.00                    2807            5276
 6 4/6   Thursd…      48.9     41   0.73                     461            1324
 7 4/7   Friday       48       43   T                       1222            2955
 8 4/8   Saturd…      55.9     39.9 0.00                    1674            3163
 9 4/9   Sunday       66       45   0.00                    2375            4377
10 4/10  Monday       73.9     55   0.00                    3324            6359
# ℹ 204 more rows
# ℹ 3 more variables: williamsburgbridge <dbl>, queensborobridge <dbl>,
#   total <dbl>
```

Converting **`date`** from <chr> to <date>.

Creating a new column called **`month`**

Checking for missing values:

Recoding the values and impute for **`precipitation`**:

Converting **`precipitation`** from <chr> to <dbl>

```
 [1] "date"              "day"               "HighTempF"
 [4] "LowTempF"          "precipitation"     "brooklynbridge"
 [7] "manhattanbridge"   "williamsburgbridge" "queensborobridge"
[10] "total"             "month"             "rain"
[13] "average_temp"
```

Removed columns **`HighTempF`**, **`LowTempF`**, **`Precipitation`**.

```
NY_cycling_data <- NY_cycling_data[, c("date", "day", "brooklynbridge", "manhattanbridge"
NY_cycling_data
```

```
# A tibble: 214 × 10
   date       day       brooklynbridge manhattanbridge williamsburgbridge
```

```
   <date>     <chr>              <dbl>           <dbl>              <dbl>
 1 2017-04-01 Saturday             606            1446               1915
 2 2017-04-02 Sunday              2021            3943               4207
 3 2017-04-03 Monday              2470            4988               5178
 4 2017-04-04 Tuesday              723            1913               2279
 5 2017-04-05 Wednesday           2807            5276               5711
 6 2017-04-06 Thursday             461            1324               1739
 7 2017-04-07 Friday              1222            2955               3399
 8 2017-04-08 Saturday            1674            3163               4082
 9 2017-04-09 Sunday              2375            4377               4886
10 2017-04-10 Monday              3324            6359               6881
# ℹ 204 more rows
# ℹ 5 more variables: queensborobridge <dbl>, total <dbl>, month <dbl>,
#   rain <dbl>, average_temp <dbl>
```

```
unique(NY_cycling_data$month)
```

```
[1]  4  5  6  7  8  9 10
```

Converted **months**

```
# Convert numeric months to labels
month_labels <- c("April", "May", "June", "July", "August", "September", "October")
NY_cycling_data$month <- month_labels[NY_cycling_data$month - min(NY_cycling_data$month)
NY_cycling_data
```

```
# A tibble: 214 × 10
   date       day       brooklynbridge manhattanbridge williamsburgbridge
   <date>     <chr>              <dbl>           <dbl>              <dbl>
 1 2017-04-01 Saturday             606            1446               1915
 2 2017-04-02 Sunday              2021            3943               4207
 3 2017-04-03 Monday              2470            4988               5178
 4 2017-04-04 Tuesday              723            1913               2279
 5 2017-04-05 Wednesday           2807            5276               5711
 6 2017-04-06 Thursday             461            1324               1739
 7 2017-04-07 Friday              1222            2955               3399
 8 2017-04-08 Saturday            1674            3163               4082
 9 2017-04-09 Sunday              2375            4377               4886
10 2017-04-10 Monday              3324            6359               6881
# ℹ 204 more rows
# ℹ 5 more variables: queensborobridge <dbl>, total <dbl>, month <chr>,
#   rain <dbl>, average_temp <dbl>
```

Title: NYC Cycling Data Codebook

Dataset Information: - Name: NY_cycling_data

Variable List:

```
        Variable_Name Variable_Type                         Description
1                date          Date                           Month/Day
2             weekend   Categorical                 Weekend/Not Weekend
3               month   Categorical                               Month
4 average_temperature     Numerical          Average daily temperature
5                rain     Numerical                Daily precipitation
6      brooklynbridge     Numerical    # of riders over Brooklyn Bridge
7     manhattanbridge     Numerical   # of riders over Manhattan Bridge
8  williamsburgbridge     Numerical # of riders over Williamsburg Bridge
9               total     Numerical Total riders over these four bridges
   Measurement_Unit Missing_Values
1        YYYY-MM-DD           None
2                0/1           None
3 April/May/June...           None
4        Fahrenheit           None
5            Inches           None
6             Count           None
7             Count           None
8             Count           None
9             Count           None
```

Variable_Name	Variable_Type	Description	Measurement_Unit	Missing_Values
date	Date	Month/Day	YYYY-MM-DD	None
weekend	Categorical	Weekend/Not Weekend	0/1	None
month	Categorical	Month	April/May/June...	None
average_temperature	Numerical	Average daily temperature	Fahrenheit	None
rain	Numerical	Daily precipitation	Inches	None
brooklynbridge	Numerical	# of riders over Brooklyn Bridge	Count	None
manhattanbridge	Numerical	# of riders over Manhattan Bridge	Count	None
williamsburgbridge	Numerical	# of riders over Williamsburg Bridge	Count	None
total	Numerical	Total riders over these four bridges	Count	None

Summary Statistic on Numerical Data:

```
<table class="Rtable1">
<thead>
<tr>
<th class='rowlabel firstrow lastrow'></th>
```

```
<th class='firstrow lastrow'><span class='stratlabel'>TRUE<br><span class='stratn'>
(N=214)</span></span></th>
</tr>
</thead>
<tbody>
<tr>
<td class='rowlabel firstrow'>Brooklyn Bridge</td>
<td class='firstrow'></td>
</tr>
<tr>
<td class='rowlabel'>Mean (SD)</td>
<td>2680 (855)</td>
</tr>
<tr>
<td class='rowlabel lastrow'>Median [Min, Max]</td>
<td class='lastrow'>2860 [151, 4960]</td>
</tr>
<tr>
<td class='rowlabel firstrow'>Manhattan Bridge</td>
<td class='firstrow'></td>
</tr>
<tr>
<td class='rowlabel'>Mean (SD)</td>
<td>5350 (1750)</td>
</tr>
<tr>
<td class='rowlabel lastrow'>Median [Min, Max]</td>
<td class='lastrow'>5610 [484, 8240]</td>
</tr>
<tr>
<td class='rowlabel firstrow'>Williamsburg Bridge</td>
<td class='firstrow'></td>
</tr>
<tr>
<td class='rowlabel'>Mean (SD)</td>
<td>6050 (1760)</td>
</tr>
<tr>
<td class='rowlabel lastrow'>Median [Min, Max]</td>
<td class='lastrow'>6290 [874, 8870]</td>
</tr>
<tr>
<td class='rowlabel firstrow'>Queensboro Bridge</td>
<td class='firstrow'></td>
</tr>
<tr>
```

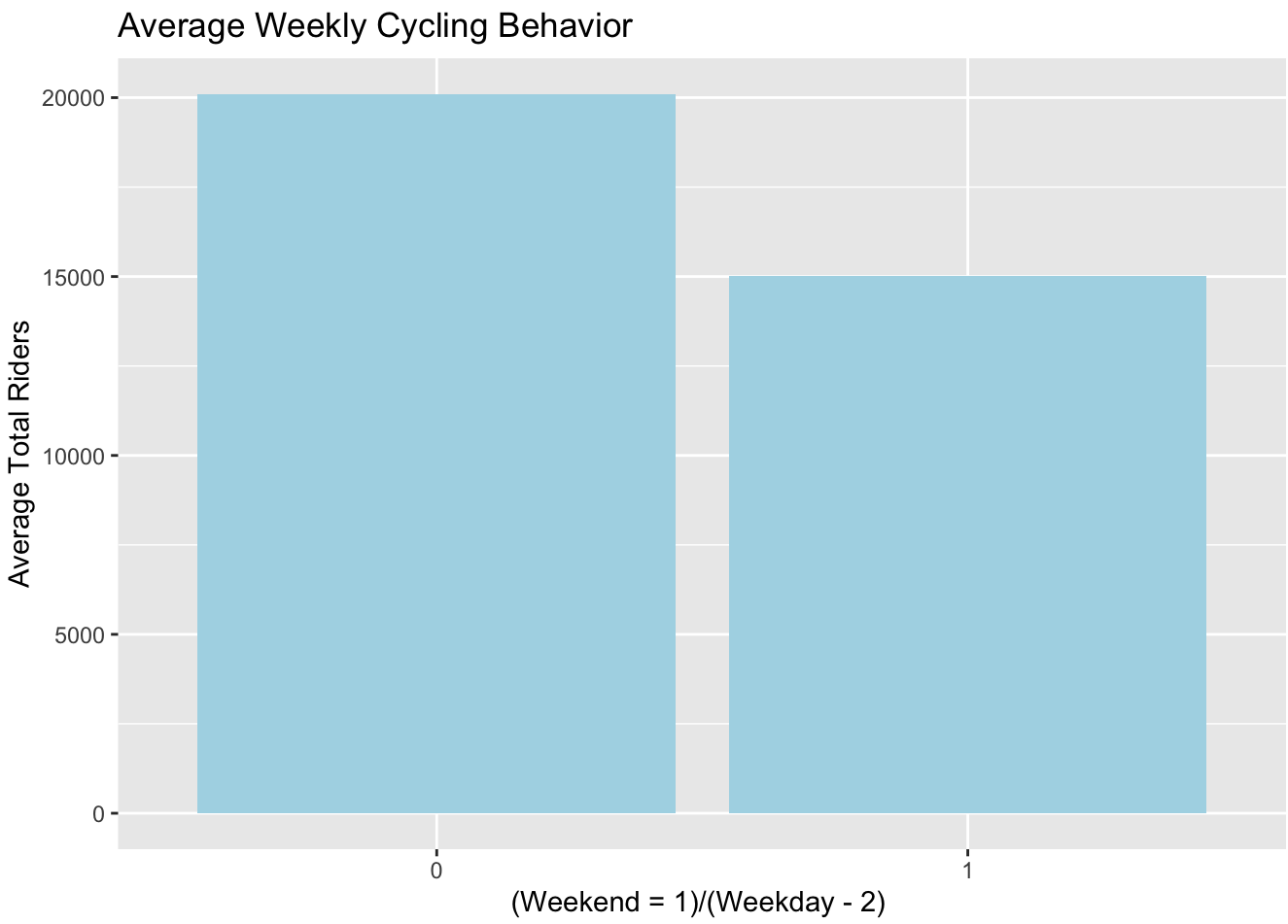
```
<td class='rowlabel'>Mean (SD)</td>
<td>4550 (1310)</td>
</tr>
<tr>
<td class='rowlabel lastrow'>Median [Min, Max]</td>
<td class='lastrow'>4680 [865, 6580]</td>
</tr>
<tr>
<td class='rowlabel firstrow'>Total</td>
<td class='firstrow'></td>
</tr>
<tr>
<td class='rowlabel'>Mean (SD)</td>
<td>18600 (5540)</td>
</tr>
<tr>
<td class='rowlabel lastrow'>Median [Min, Max]</td>
<td class='lastrow'>19400 [2370, 27000]</td>
</tr>
<tr>
<td class='rowlabel firstrow'>rain</td>
<td class='firstrow'></td>
</tr>
<tr>
<td class='rowlabel'>Mean (SD)</td>
<td>0.132 (0.394)</td>
</tr>
<tr>
<td class='rowlabel lastrow'>Median [Min, Max]</td>
<td class='lastrow'>0 [0, 3.03]</td>
</tr>
<tr>
<td class='rowlabel firstrow'>High Temp (°F)</td>
<td class='firstrow'></td>
</tr>
<tr>
<td class='rowlabel'>Mean (SD)</td>
<td>68.1 (9.57)</td>
</tr>
<tr>
<td class='rowlabel lastrow'>Median [Min, Max]</td>
<td class='lastrow'>70.0 [41.5, 86.0]</td>
</tr>
</tbody>
</table>
```

Adding Weekend

```
NY_cycling_data <- NY_cycling_data %>%
  mutate(weekend = ifelse(day %in% c("Saturday", "Sunday"), 1, 0))
```

Overall Months and Days

```
monthly_counts <- NY_cycling_data %>%
  group_by(month = factor(format(date, "%B"), levels = c("April", "May", "June", "July"))
  summarise(average_total_riders = mean(total))

daily_counts <- NY_cycling_data %>%
  group_by(weekend) %>%
  summarise(average_total_riders = mean(total))

# Daily pattern
ggplot(daily_counts, aes(x = factor(weekend), y = average_total_riders)) +
  geom_bar(stat = "identity", fill = "lightblue") +
  labs(title = "Average Weekly Cycling Behavior",
       x = "(Weekend = 1)/(Weekday - 2) ",
       y = "Average Total Riders")
```




```
NY_cycling_data %>%
  group_by(month) %>%
  summarise(n = n(), mean = mean(total), sd = sd(total)) %>%
  mutate(month = factor(month, levels = c("April", "May", "June", "July", "August", "Sept
  arrange(month)
```

```
# A tibble: 7 × 4
  month         n   mean    sd
  <fct>     <int>  <dbl> <dbl>
1 April        30 15029. 5518.
2 May          31 17196. 6387.
3 June         30 19894. 4999.
4 July         31 18805. 4763.
5 August       31 20367. 4684.
6 September    30 20435. 4979.
7 October      31 18652. 5640.
```

```
NY_cycling_data %>%
  group_by(weekend) %>%
  summarise(n = n(), mean = mean(total), sd = sd(total))
```

```
# A tibble: 2 × 4
  weekend     n   mean    sd
    <dbl> <int>  <dbl> <dbl>
1       0   152 20099. 5459.
2       1    62 15020. 3843.
```

```
daily_avg_counts <- NY_cycling_data %>%
  group_by(weekend) %>%
  summarise(avg_brooklyn = mean(brooklynbridge),
            var_brooklyn = sd(brooklynbridge, na.rm=T)^2,
            avg_manhattan = mean(manhattanbridge),
            var_manhattanbridge = sd(manhattanbridge, na.rm=T)^2,
            avg_williamsburgbridge = mean(williamsburgbridge),
            var_williamsburgbridge = sd(williamsburgbridge, na.rm=T)^2,
            avg_queensborobridge = mean(queensborobridge),
            var_queensborobridge = sd(queensborobridge, na.rm=T)^2)
daily_avg_counts
```

```
# A tibble: 2 × 9
  weekend avg_brooklyn var_brooklyn avg_manhattan var_manhattanbridge
    <dbl>        <dbl>        <dbl>         <dbl>               <dbl>
1       0        2783.      694637.         5859.            2889125.
2       1        2428.      740689.         4087.            1231612.
# ℹ 4 more variables: avg_williamsburgbridge <dbl>,
#   var_williamsburgbridge <dbl>, avg_queensborobridge <dbl>,
#   var_queensborobridge <dbl>
```

#1 Brooklyn Bridge

How are our independent variables related to the outcome?

```
NY_cycling_data %>%
  group_by(weekend) %>%
  summarise(n = n(), mean = mean(brooklynbridge), sd = sd(brooklynbridge)) %>%arrange(wee
```

```
# A tibble: 2 × 4
  weekend     n  mean    sd
    <dbl> <int> <dbl> <dbl>
1       0   152 2783.  833.
2       1    62 2428.  861.
```

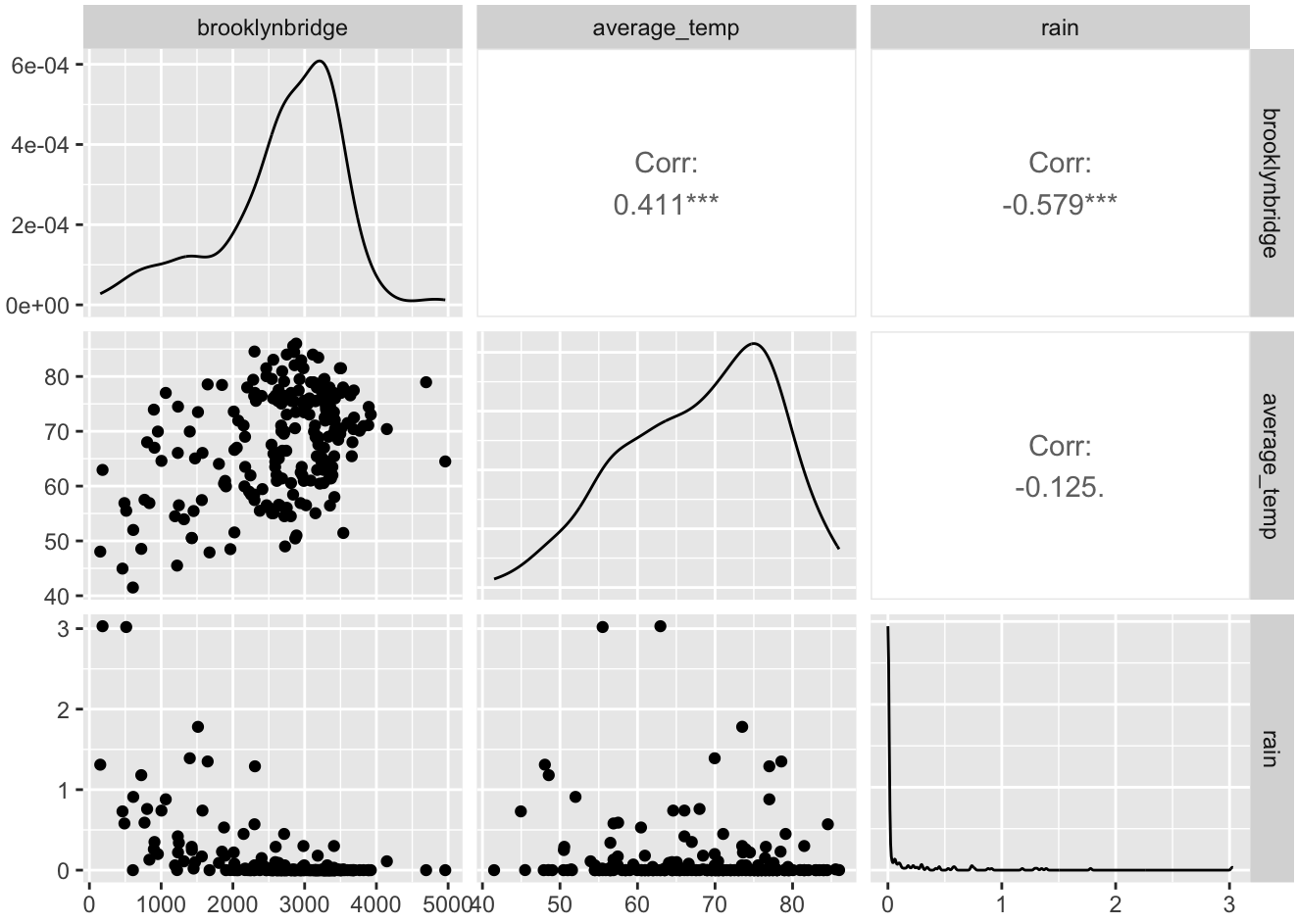
```
NY_cycling_data %>%
  group_by(month) %>%
  summarise(n = n(), mean = mean(brooklynbridge), sd = sd(brooklynbridge)) %>%
  mutate(month = factor(month, levels = c("April", "May", "June", "July", "August", "Sept
  arrange(month)
```

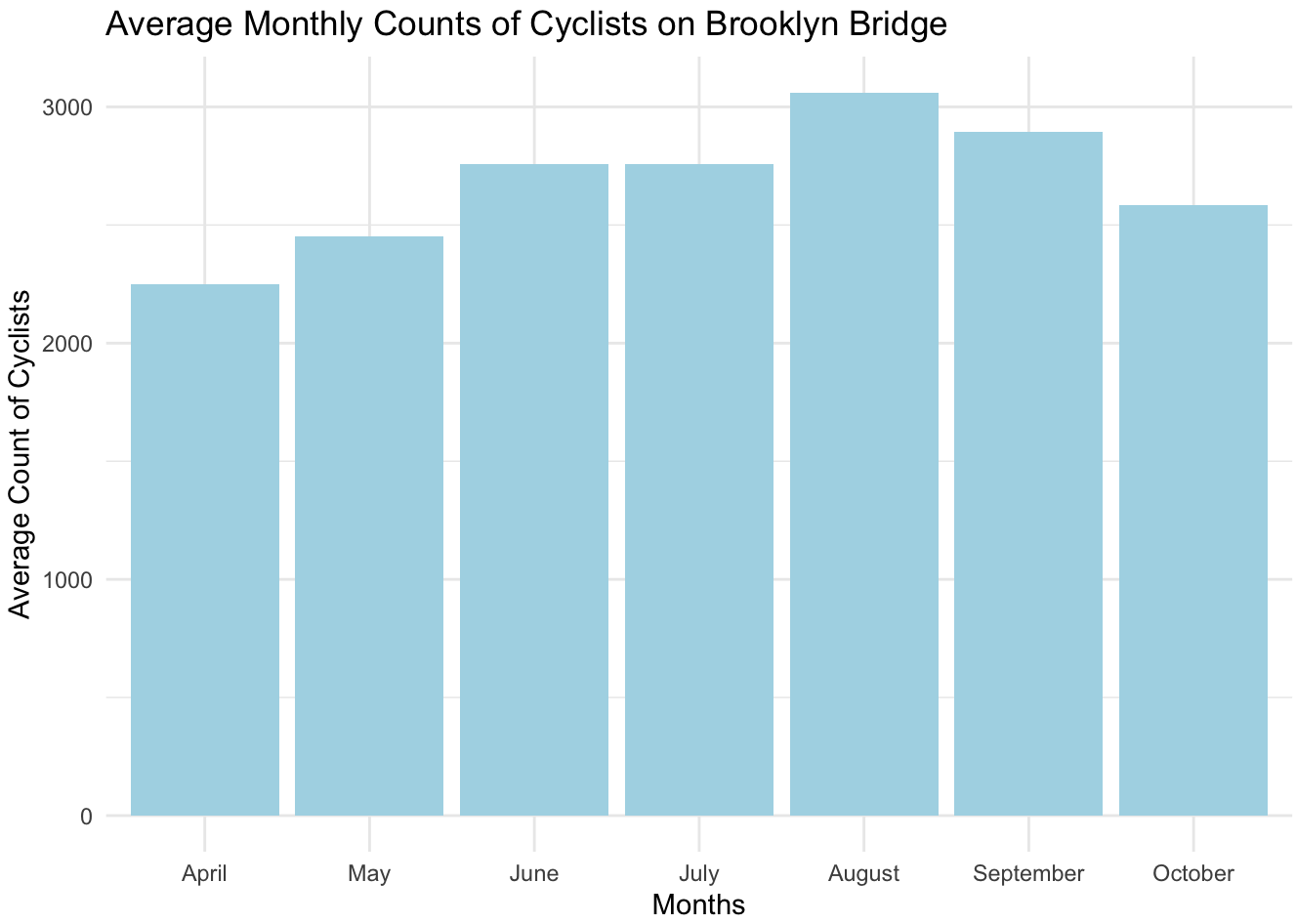
```
# A tibble: 7 × 4
  month         n  mean    sd
  <fct>     <int> <dbl> <dbl>
1 April        30 2250.  980.
2 May          31 2451.  972.
3 June         30 2757.  683.
4 July         31 2756.  659.
5 August       31 3060.  822.
6 September    30 2896.  754.
7 October      31 2585.  865.
```

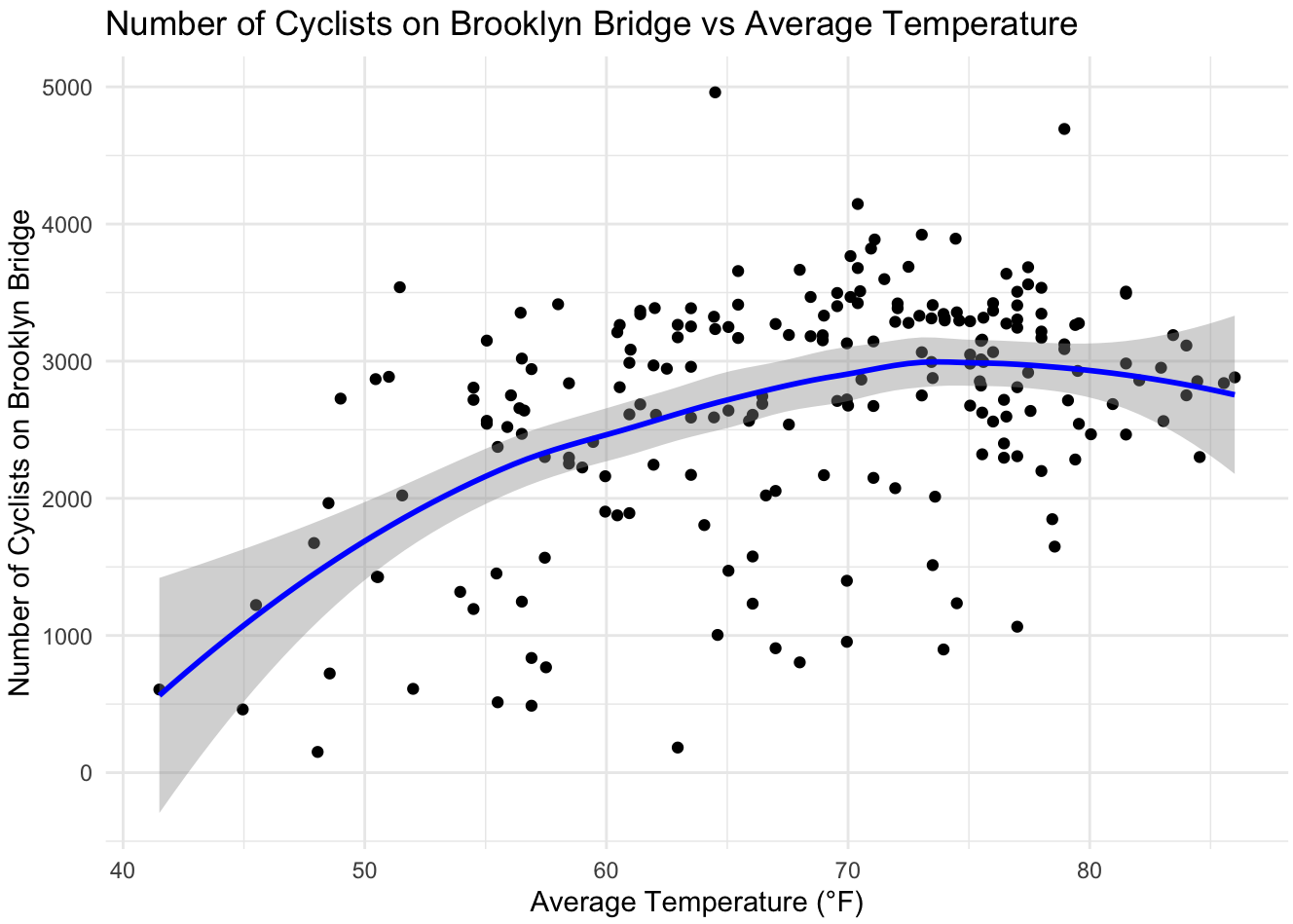
```
library(GGally)
```

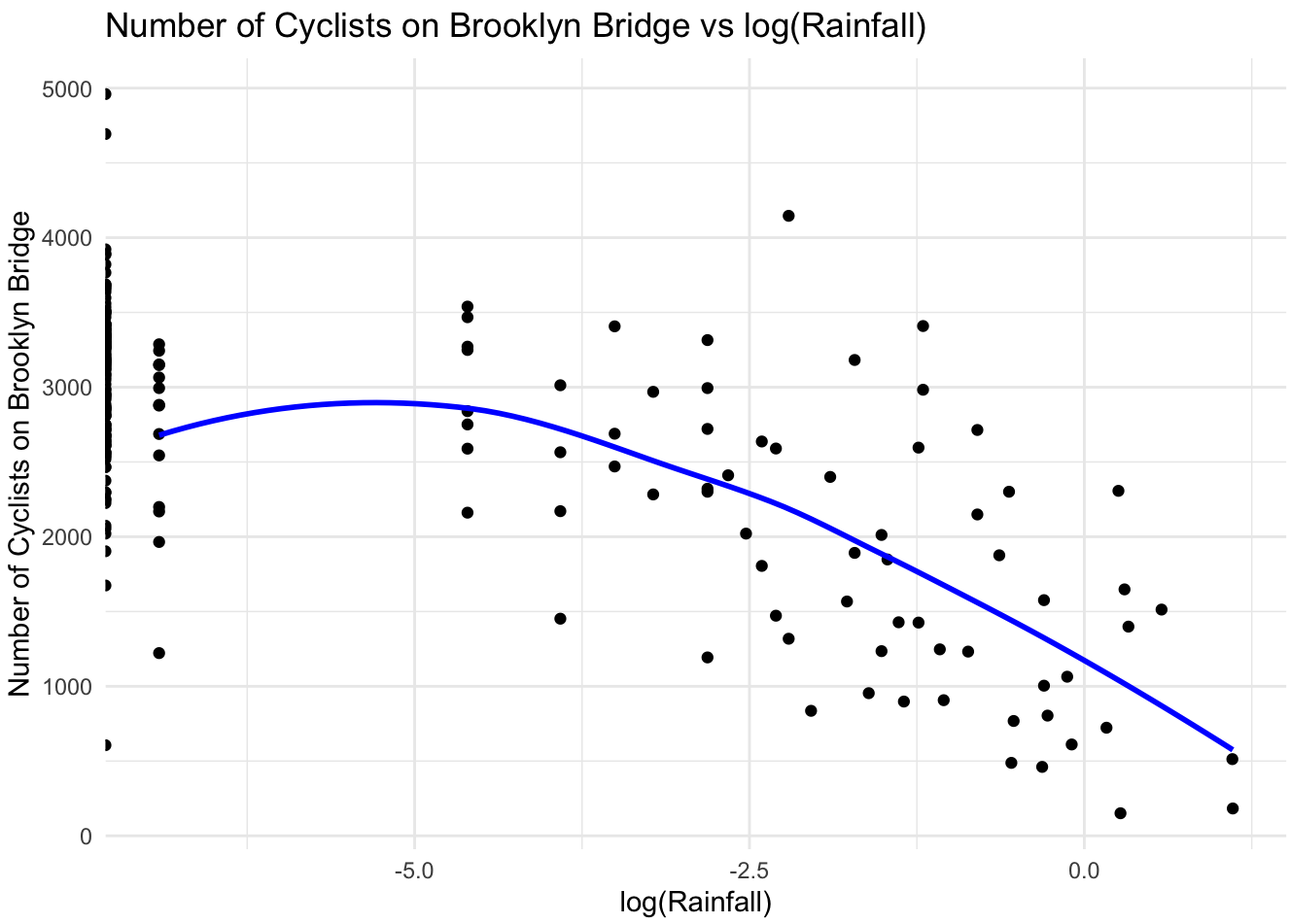
```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

```
bb_correlation_data <- NY_cycling_data[, c("brooklynbridge", "average_temp", "rain")]
ggpairs(bb_correlation_data)
```


brooklynbridge
average_temp
rain
Corr:
0.411***
Corr:
-0.579***
Corr:
-0.125.
6e-04
4e-04
2e-04
0e+00
80
70
60
50
40
3
2
1
0
0
1000
2000
3000
4000
5000
40
50
60
70
80
0
1
2
3


Average Monthly Counts of Cyclists on Brooklyn Bridge
Average Count of Cyclists
3000
2000
1000
0
April
May
June
July
August
September
October
Months


Number of Cyclists on Brooklyn Bridge vs Average Temperature
Number of Cyclists on Brooklyn Bridge
5000
4000
3000
2000
1000
0
40
50
60
70
80
Average Temperature (°F)



Modeling:

I will be centering since there's interaction:

```
#Centering on its mean
brooklynbridge_data <-NY_cycling_data %>% select(brooklynbridge,rain, average_temp, weeke
centered_brooklynbridge_data <- brooklynbridge_data %>%
  mutate(rain = rain - mean(rain), average_temp = average_temp - mean(average_temp))
centered_brooklynbridge_data
```

```
# A tibble: 214 × 5
   brooklynbridge   rain average_temp weekend month
            <dbl>  <dbl>        <dbl>   <dbl> <chr>
 1            606 -0.132        -26.6       1 April
 2           2021 -0.132        -16.6       1 April
 3           2470 -0.102        -11.6       0 April
 4            723  1.05         -19.6       0 April
 5           2807 -0.132        -13.6       0 April
```

```
 6            461  0.598       -23.2        0 April
 7           1222 -0.131       -22.6        0 April
 8           1674 -0.132       -20.2        1 April
 9           2375 -0.132       -12.6        1 April
10           3324 -0.132        -3.66       0 April
# ℹ 204 more rows
```

```
#Check the relationship between varaibles --

#Against the Null Model
bb_null_poisson_model <- glm(brooklynbridge ~ 1, family = poisson(link = "log"), data = c
summary(bb_null_poisson_model)
```

```
Call:
glm(formula = brooklynbridge ~ 1, family = poisson(link = "log"),
    data = centered_brooklynbridge_data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.89359    0.00132    5978   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 70021  on 213  degrees of freedom
Residual deviance: 70021  on 213  degrees of freedom
AIC: 72087

Number of Fisher Scoring iterations: 4
```

```
#average_temp + Rain
bb_poisson_model2 <- glm(brooklynbridge ~ average_temp + as.numeric(rain), family = poiss

#Average_ + Rain + month
bb_poisson_model3 <- glm(brooklynbridge ~average_temp + as.numeric(rain) + factor(month)
```

```
#Full Model (Sequence) Added average_temp + rain + month + days
bb_full_poisson_model <- glm(brooklynbridge ~ average_temp + as.numeric(rain) + factor(mo

anova(bb_null_poisson_model,bb_poisson_model2,bb_poisson_model3, bb_full_poisson_model, t
```

```
Analysis of Deviance Table

Model 1: brooklynbridge ~ 1
Model 2: brooklynbridge ~ average_temp + as.numeric(rain)
Model 3: brooklynbridge ~ average_temp + as.numeric(rain) + factor(month)
Model 4: brooklynbridge ~ average_temp + as.numeric(rain) + factor(month) +
    factor(weekend)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       213      70021
2       211      33806  2    36216 < 2.2e-16 ***
3       205      32863  6      943 < 2.2e-16 ***
4       204      31677  1     1185 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pois_pearson_gof(bb_full_poisson_model) #>>> 1.445647e-193 BAD FIT
```

```
$pval
[1] 0

$df
[1] 204
```

```
#pois_dev_gof(bb_full_poisson_model)
#anova(bb_full_poisson_model, test= "LRT")
```

MFP:

```
#MFP
library(mfp)
```

```
Loading required package: survival
```

```
mfp::mfp(brooklynbridge ~ fp(average_temp),
         family = poisson,
         data = centered_brooklynbridge_data)
```

```
Call:
mfp::mfp(formula = brooklynbridge ~ fp(average_temp), data =
centered_brooklynbridge_data,
    family = poisson)
```

```
Deviance table:
         Resid. Dev
Null model   70021.32
Linear model     60013.28
Final model  54703.51

Fractional polynomials:
             df.initial select alpha df.final power1 power2
average_temp          4      1  0.05        4     -2     -1


Transformations of covariates:
                                                                      formula
average_temp I(((average_temp+26.7)/10)^-2)+I(((average_temp+26.7)/10)^-1)

Coefficients:
     Intercept  average_temp.1  average_temp.2
      8.156126        0.004966       -0.595707

Degrees of Freedom: 213 Total (i.e. Null);  211 Residual
Null Deviance:      70020
Residual Deviance: 54700    AIC: 56770
```

Interaction and Confouding:

```r
#confounding

# Fit the full Poisson regression model
bb_full_poisson_model <- glm(brooklynbridge ~  factor(month) + factor(weekend)+ average_t

# Includce interaction term between average_temp and rain
bb_interaction_model <- glm(brooklynbridge ~  factor(month) + factor(weekend) + average_t

# Compare the models using ANOVA
anova_result <- anova(bb_full_poisson_model, bb_interaction_model, test = "LRT")
print(anova_result)
```

```
Analysis of Deviance Table

Model 1: brooklynbridge ~ factor(month) + factor(weekend) + average_temp +
    as.numeric(rain)
Model 2: brooklynbridge ~ factor(month) + factor(weekend) + average_temp +
    as.numeric(rain) + average_temp * as.numeric(rain)
```

```
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       204      31677
2       203      28887  1     2790 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overdispersion Test and AIC/BIC:

```
bb_interaction_model <- glm(brooklynbridge ~ average_temp * as.numeric(rain) + factor(mon

AER::dispersiontest(bb_interaction_model)
```

```
    Overdispersion test

data:  bb_interaction_model
z = 5.318, p-value = 5.245e-08
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  142.8918
```

```
bb_aic_value_org <- AIC(bb_poisson_model3)
bb_bic_value_org <- BIC(bb_poisson_model3)
bb_aic_value <- AIC(bb_interaction_model)
bb_bic_value <- BIC(bb_interaction_model)

bb_aic_value_org
```

```
[1] 34944.93
```

```
bb_bic_value_org
```

```
[1] 34975.23
```

```
bb_aic_value
```

```
[1] 30973.54
```

```
bb_bic_value
```

```
[1] 31010.57
```

How is our outcome variable distributed?

```
centered_brooklynbridge_data %>% select(brooklynbridge) %>% skimr::skim()
```

Data summary	
Name	Piped data
Number of rows	214
Number of columns	1
_____	
Column type frequency:	
numeric	1
_____	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
brooklynbridge	0	1	2680.04	854.71	1151	2298	2857	3285	4960	▁▃▇▇▂

```
bb_negbin_model <- glm.nb(brooklynbridge ~ factor(month) + factor(weekend) +  average_tem

summary(bb_negbin_model)
```

```
Call:
glm.nb(formula = brooklynbridge ~ factor(month) + factor(weekend) +
    average_temp + as.numeric(rain) + average_temp * as.numeric(rain),
    data = centered_brooklynbridge_data, init.theta = 12.60541684,
    link = log)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                  7.846776   0.062089 126.380  < 2e-16 ***
factor(month)August          0.090407   0.086915   1.040  0.29826
factor(month)July           -0.023677   0.091860  -0.258  0.79660
```

```
factor(month)June              0.041951   0.085509   0.491  0.62371
factor(month)May               0.076232   0.073747   1.034  0.30128
factor(month)October           0.048764   0.075006   0.650  0.51560
factor(month)September         0.073456   0.082240   0.893  0.37176
factor(weekend)1              -0.120589   0.042815  -2.817  0.00485 **
average_temp                   0.014798   0.002873   5.151 2.59e-07 ***
as.numeric(rain)              -0.630129   0.057495 -10.960  < 2e-16 ***
average_temp:as.numeric(rain)  0.016346   0.005478   2.984  0.00285 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(12.6054) family taken to be 1)

    Null deviance: 460.89  on 213  degrees of freedom
Residual deviance: 218.15  on 203  degrees of freedom
AIC: 3438.9

Number of Fisher Scoring iterations: 1

              Theta:  12.61
          Std. Err.:  1.22

 2 x log-likelihood:  -3414.874
```

```
#GOF

anova(bb_negbin_model, test = "LRT")
```

```
Warning in anova.negbin(bb_negbin_model, test = "LRT"): tests made without
re-estimating 'theta'

Analysis of Deviance Table

Model: Negative Binomial(12.6054), link: log

Response: brooklynbridge

Terms added sequentially (first to last)

                             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                           213     460.89
factor(month)                 6   24.454       207     436.43 0.0004308 ***
factor(weekend)               1   10.078       206     426.36 0.0015007 **
average_temp                  1   35.195       205     391.16 2.983e-09 ***
as.numeric(rain)              1  162.807       204     228.35 < 2.2e-16 ***
average_temp:as.numeric(rain) 1   10.205       203     218.15 0.0014003 **
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pois_dev_gof(bb_negbin_model)
```

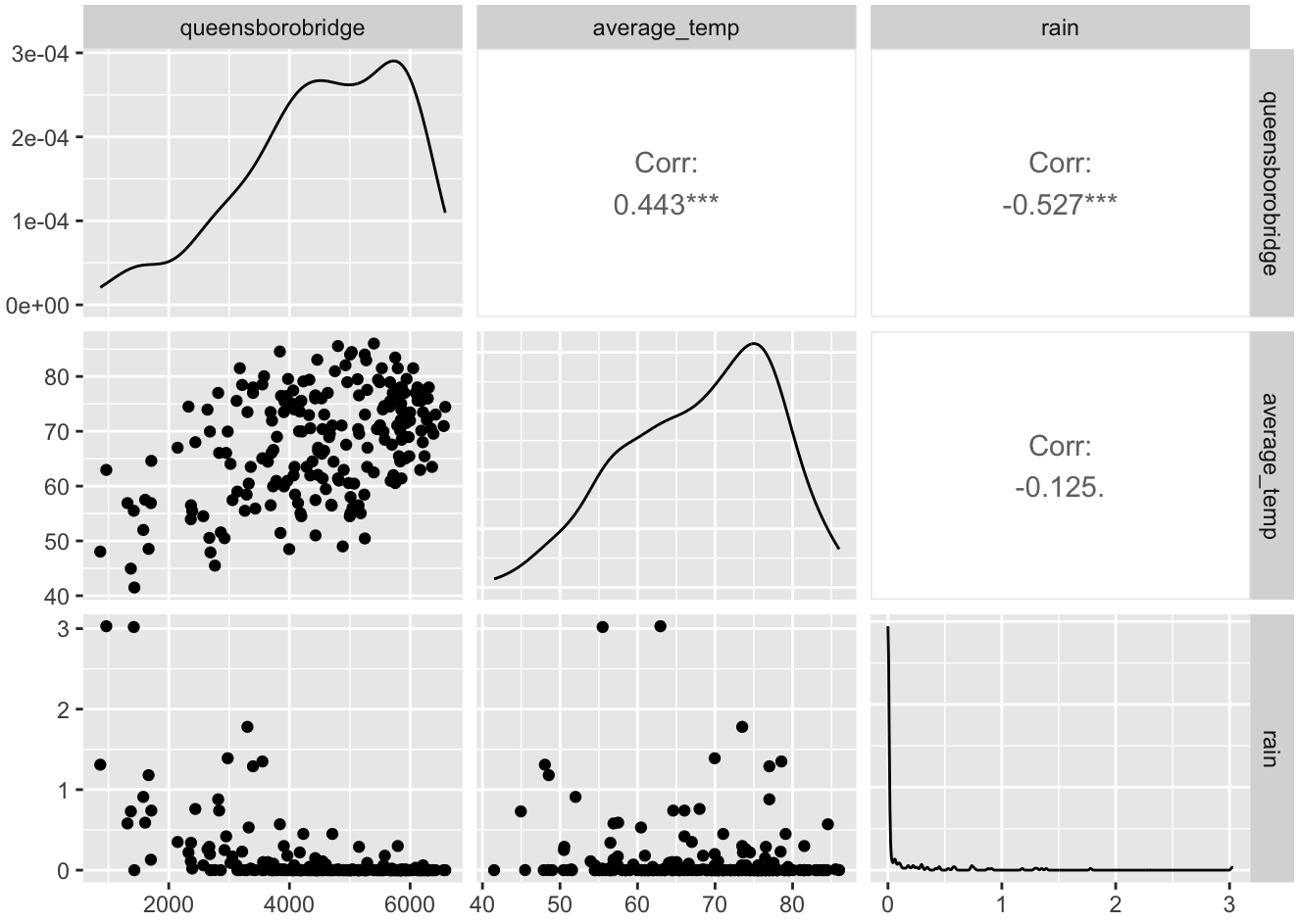
```
$pval
[1] 0.2216194

$df
[1] 203
```

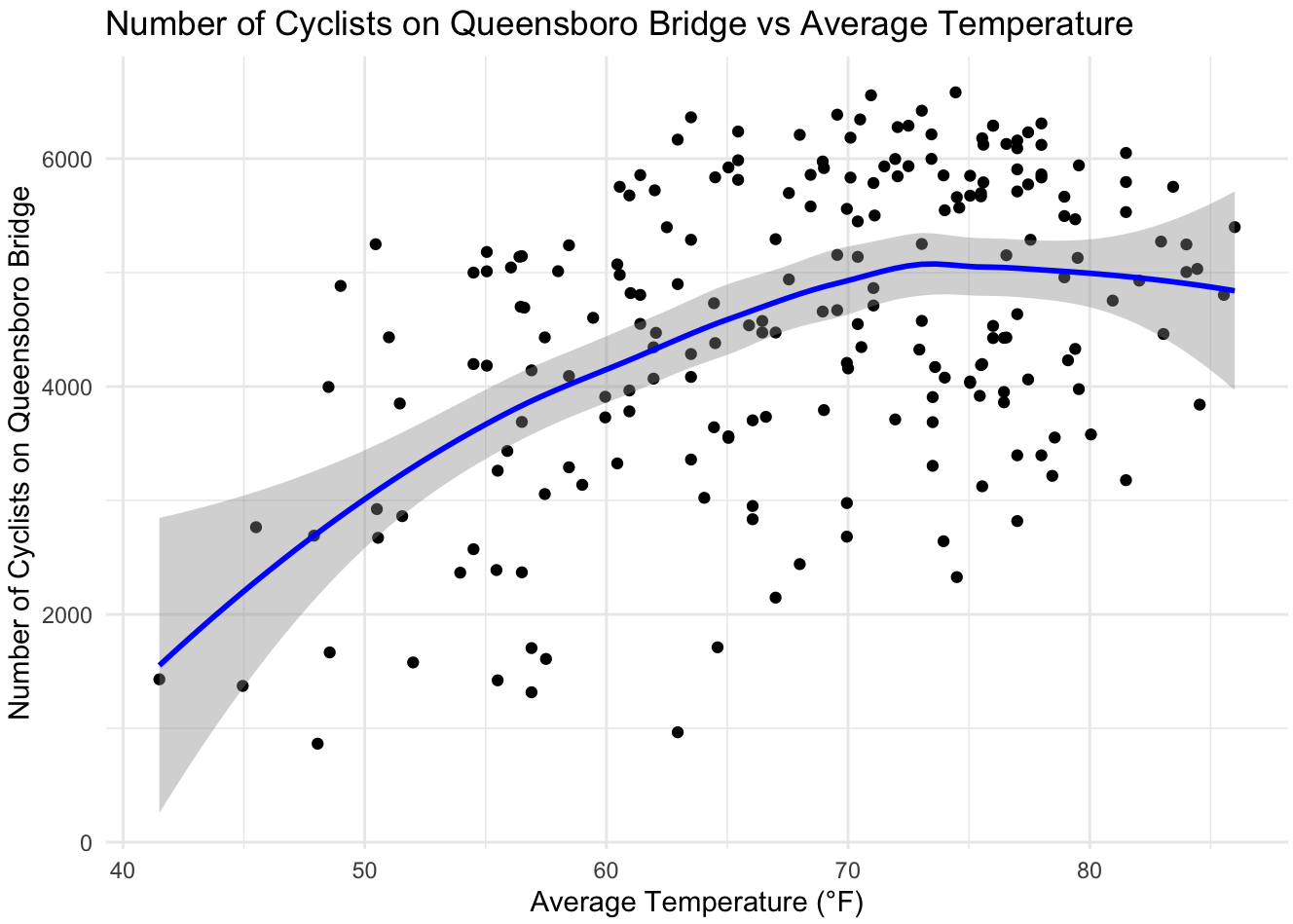
#2 Queensboro Bridge

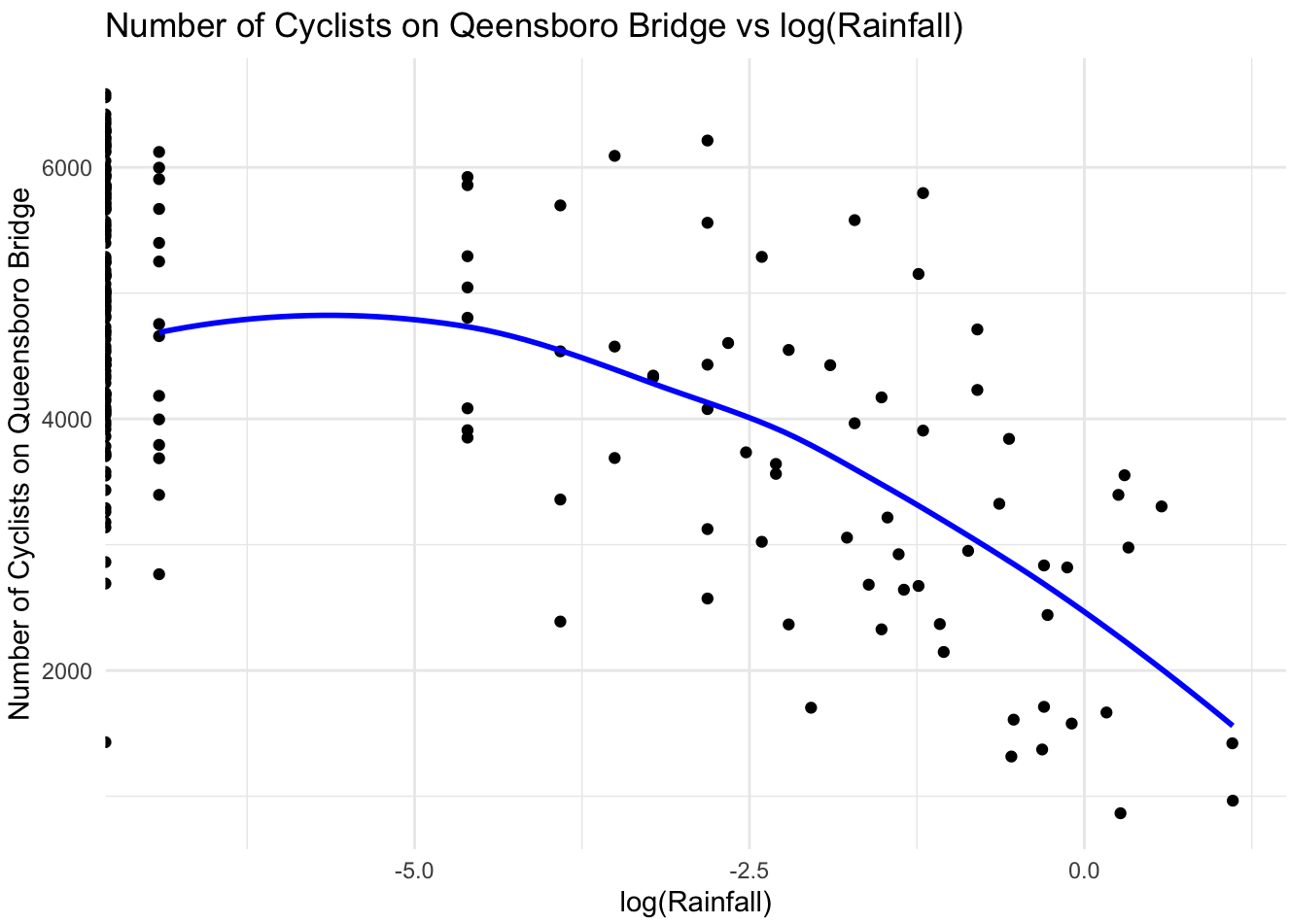
```
NY_cycling_data
```

```
# A tibble: 214 × 11
   date       day       brooklynbridge manhattanbridge williamsburgbridge
   <date>     <chr>              <dbl>           <dbl>               <dbl>
 1 2017-04-01 Saturday             606            1446                1915
 2 2017-04-02 Sunday              2021            3943                4207
 3 2017-04-03 Monday              2470            4988                5178
 4 2017-04-04 Tuesday              723            1913                2279
 5 2017-04-05 Wednesday           2807            5276                5711
 6 2017-04-06 Thursday             461            1324                1739
 7 2017-04-07 Friday              1222            2955                3399
 8 2017-04-08 Saturday            1674            3163                4082
 9 2017-04-09 Sunday              2375            4377                4886
10 2017-04-10 Monday              3324            6359                6881
# ℹ 204 more rows
# ℹ 6 more variables: queensborobridge <dbl>, total <dbl>, month <chr>,
#   rain <dbl>, average_temp <dbl>, weekend <dbl>
```

```
qb_correlation_data <- NY_cycling_data[, c("queensborobridge", "average_temp", "rain")]
ggpairs(qb_correlation_data)
```


queensborobridge
average_temp
rain
Corr:
0.443***
Corr:
-0.527***
Corr:
-0.125.
3e-04
2e-04
1e-04
0e+00
80
70
60
50
40
3
2
1
0
2000
4000
6000
40
50
60
70
80
0
1
2
3


Number of Cyclists on Queensboro Bridge vs Average Temperature
Number of Cyclists on Queensboro Bridge
0
2000
4000
6000
40
50
60
70
80
Average Temperature (°F)



Month:

```
NY_cycling_data %>%
  group_by(month) %>%
  summarise(n = n(), mean = mean(queensborobridge), sd = sd(queensborobridge)) %>%
  mutate(month = factor(month, levels = c("April", "May", "June", "July", "August", "Sept
  arrange(month)
```

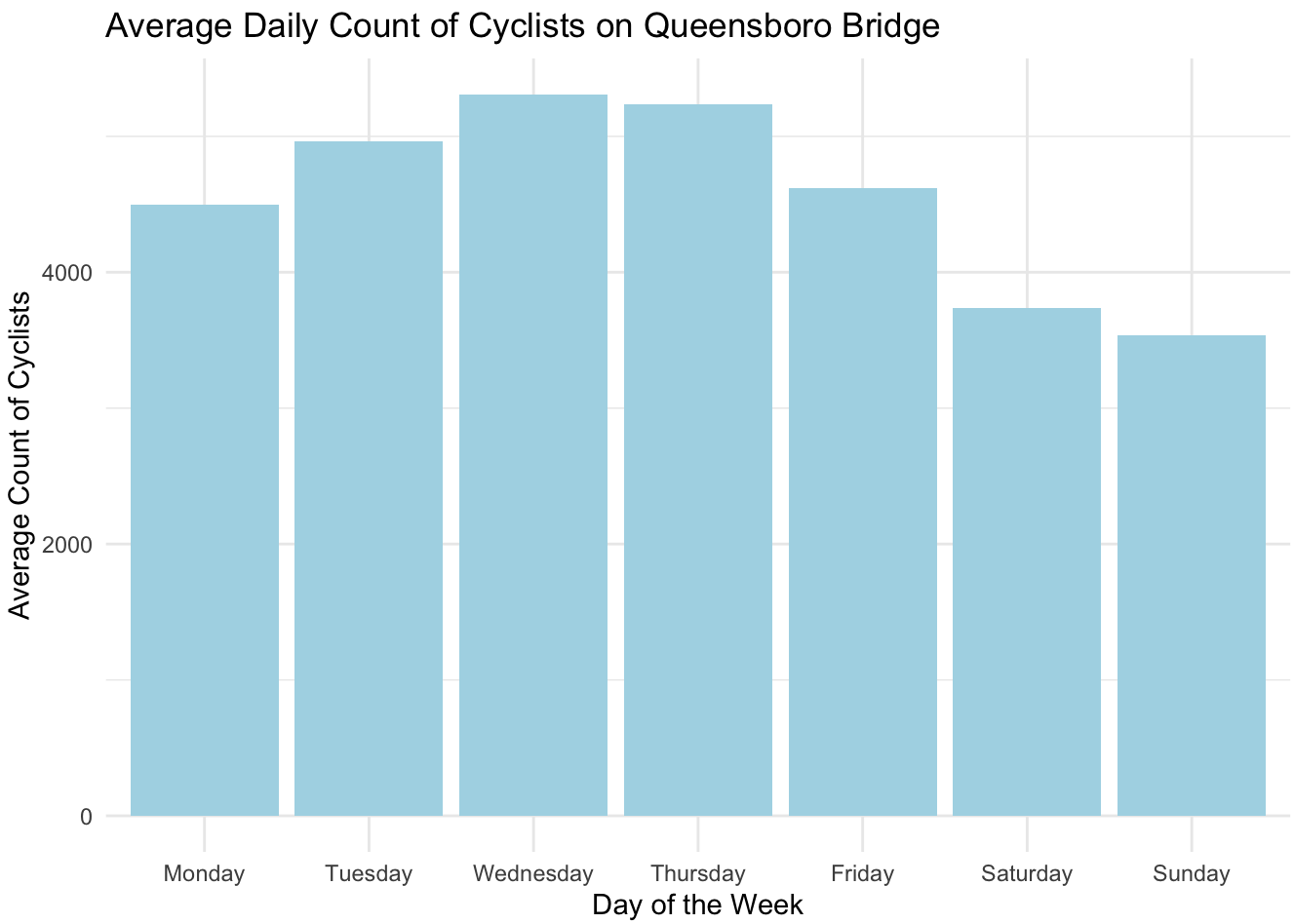
```
# A tibble: 7 × 4
  month         n  mean    sd
  <fct>     <int> <dbl> <dbl>
1 April        30 3483. 1147.
2 May          31 4075. 1438.
3 June         30 4737. 1139.
4 July         31 4551. 1073.
5 August       31 5169. 1075.
6 September    30 5054. 1158.
7 October      31 4773. 1349.
```

Days:

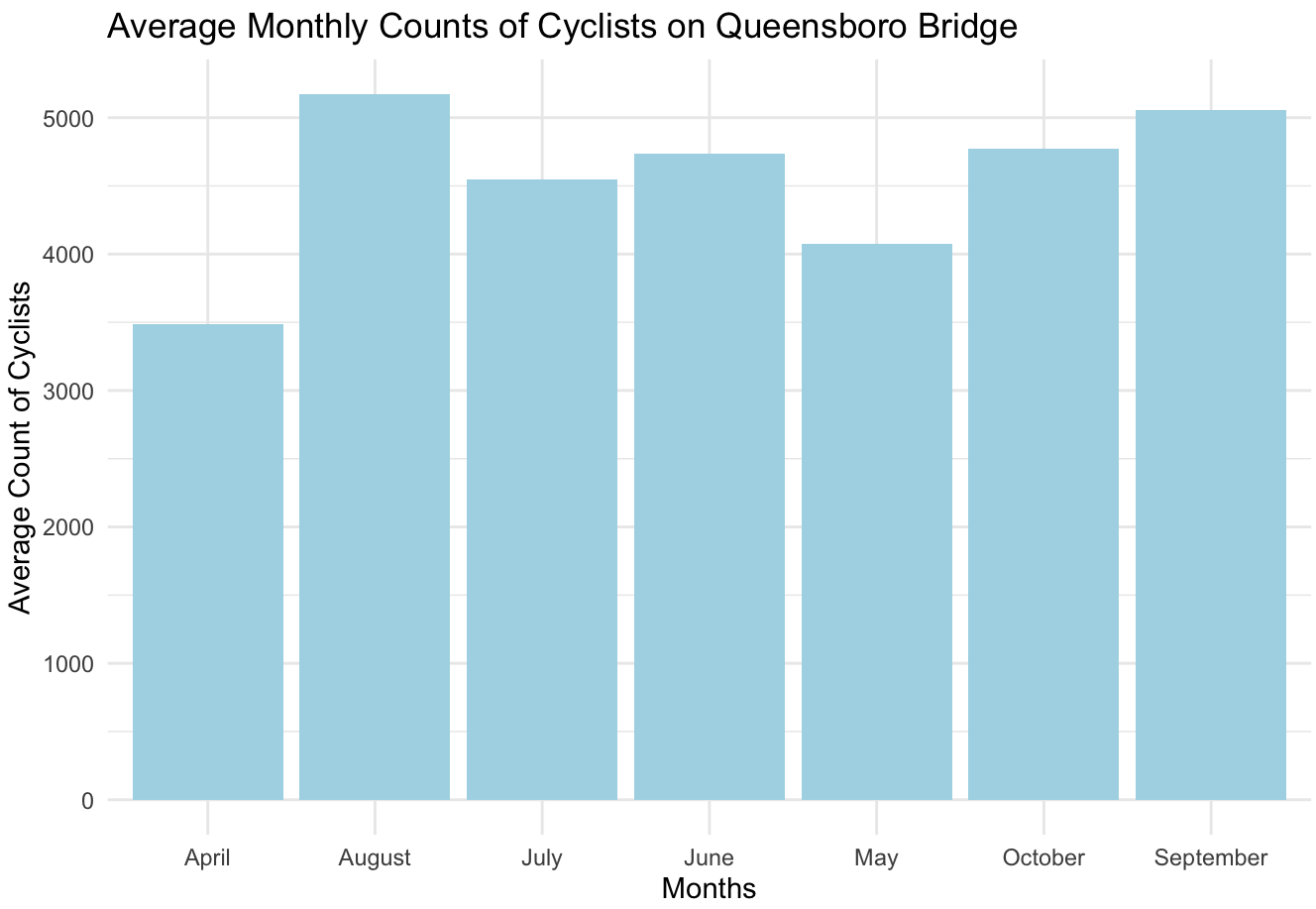
```
NY_cycling_data %>%
  group_by(weekend) %>%
  summarise(n = n(), mean = mean(queensborobridge), sd = sd(queensborobridge))
```

```
# A tibble: 2 × 4
  weekend     n  mean    sd
    <dbl> <int> <dbl> <dbl>
1       0   152 4923. 1271.
2       1    62 3638.  877.
```

```
library(ggplot2)
library(dplyr)


NY_cycling_data %>%
  mutate(day = factor(day, levels = c("Monday", "Tuesday", "Wednesday", "Thursday", "Frid
  group_by(day) %>%
  summarise(total_count = mean(queensborobridge)) %>%
  ggplot(aes(x = day, y = total_count)) +
  geom_bar(stat = "identity", fill = "lightblue") +
  labs(title = "Average Daily Count of Cyclists on Queensboro Bridge",
       x = "Day of the Week",
       y = "Average Count of Cyclists") +
  theme_minimal()
```




```
total_count_brooklyn <- NY_cycling_data %>%
  group_by(month) %>%
  summarise(total_count = mean(queensborobridge))

ggplot(total_count_brooklyn, aes(x = month, y = total_count)) +
  geom_bar(stat = "identity", fill = "lightblue") +
  labs(title = "Average Monthly Counts of Cyclists on Queensboro Bridge",
       x = "Months
       ",
       y = "Average Count of Cyclists") +
  theme_minimal()
```



Modeling:

```
queensborobridge_data <-NY_cycling_data %>% select(queensborobridge,rain, average_temp, w

centered_queensborobridge_data <- queensborobridge_data %>%
  mutate(rain = rain - mean(rain), average_temp = average_temp - mean(average_temp))
centered_queensborobridge_data
```

```
# A tibble: 214 × 5
   queensborobridge   rain average_temp weekend month
              <dbl>  <dbl>        <dbl>   <dbl> <chr>
 1             1430 -0.132       -26.6        1 April
 2             2862 -0.132       -16.6        1 April
 3             3689 -0.102       -11.6        0 April
 4             1666  1.05        -19.6        0 April
 5             4197 -0.132       -13.6        0 April
 6             1372  0.598       -23.2        0 April
 7             2765 -0.131       -22.6        0 April
```

```
 8             2691 -0.132       -20.2        1 April
 9             3261 -0.132       -12.6        1 April
10             4731 -0.132        -3.66       0 April
# ℹ 204 more rows
```

```
centered_queensborobridge_data <- queensborobridge_data %>%
  mutate(rain = rain - mean(rain), average_temp = average_temp - mean(average_temp))
centered_queensborobridge_data
```

```
# A tibble: 214 × 5
   queensborobridge   rain average_temp weekend month
              <dbl>  <dbl>        <dbl>   <dbl> <chr>
 1             1430 -0.132       -26.6        1 April
 2             2862 -0.132       -16.6        1 April
 3             3689 -0.102       -11.6        0 April
 4             1666  1.05        -19.6        0 April
 5             4197 -0.132       -13.6        0 April
 6             1372  0.598       -23.2        0 April
 7             2765 -0.131       -22.6        0 April
 8             2691 -0.132       -20.2        1 April
 9             3261 -0.132       -12.6        1 April
10             4731 -0.132        -3.66       0 April
# ℹ 204 more rows
```

```
#Check the relationship between varaibles --
#Against the Null Model
qb_null_poisson_model <- glm(queensborobridge ~ 1, family = poisson(link = "log"), data =
summary(qb_null_poisson_model )
```

```

Call:
glm(formula = queensborobridge ~ 1, family = poisson(link = "log"),
    data = centered_queensborobridge_data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 8.422990   0.001013    8312   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 89871  on 213  degrees of freedom
Residual deviance: 89871  on 213  degrees of freedom
AIC: 92057
```

```
Number of Fisher Scoring iterations: 4
```

```
#average_temp + Rain
qb_poisson_model2 <- glm(queensborobridge ~ average_temp + as.numeric(rain), family = poi

#Average_ + Rain + month
qb_poisson_model3 <- glm(queensborobridge ~average_temp + as.numeric(rain) + factor(month
```

```
#Full Model (Sequence) Added average_temp + rain + month + days
qb_full_poisson_model <- glm(queensborobridge ~ average_temp + as.numeric(rain) + factor(

anova(qb_null_poisson_model,qb_poisson_model2,qb_poisson_model3, qb_full_poisson_model, t
```

```
Analysis of Deviance Table

Model 1: queensborobridge ~ 1
Model 2: queensborobridge ~ average_temp + as.numeric(rain)
Model 3: queensborobridge ~ average_temp + as.numeric(rain) + factor(month)
Model 4: queensborobridge ~ average_temp + as.numeric(rain) + factor(month) +
    factor(weekend)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       213      89871
2       211      47806  2    42065 < 2.2e-16 ***
3       205      42077  6     5729 < 2.2e-16 ***
4       204      28862  1    13215 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#pois_pearson_gof(poisson_model) #>>> 1.445647e-193 BAD FIT
#pois_dev_gof(poisson_model)
#anova(poisson_model, test= "LRT")
```

MFP:

```
#library(mfp)
#mfp::mfp(williamsburgbridge ~ fp(average_temp),
 #        family = poisson,
  #       data = centered_queensborobridge_data)
```

```
centered_queensborobridge_data %>% select(queensborobridge) %>% skimr::skim()
```

Data summary	
Name	Piped data
Number of rows	214
Number of columns	1
_____	
Column type frequency:	
numeric	1
_____	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
queensborobridge	0	1	4550.49	1306.9	865	3746	4681	5692	6582	▁▂▆▇▇

Confounding and Interaction:

```
#confounding

# Fit the full Poisson regression model
qb_full_poisson_model <- glm(queensborobridge ~  factor(month) + factor(weekend)+ average

# Include interaction term between average_temp and rain
qb_interaction_model <- glm(queensborobridge ~  factor(month) + factor(weekend) + average

# Compare the models using ANOVA
qb_anova_result <- anova(qb_full_poisson_model, qb_interaction_model, test = "LRT")
print(qb_anova_result)
```

```
Analysis of Deviance Table

Model 1: queensborobridge ~ factor(month) + factor(weekend) + average_temp +
    as.numeric(rain)
Model 2: queensborobridge ~ factor(month) + factor(weekend) + average_temp +
    as.numeric(rain) + average_temp * as.numeric(rain)
```

```
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       204      28862
2       203      26986  1   1875.5 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

OverDispersion:

```
qb_interaction_model <- glm(queensborobridge ~  factor(month) + factor(weekend) + average

AER::dispersiontest(qb_full_poisson_model)
```

```
    Overdispersion test

data:  qb_full_poisson_model
z = 8.0441, p-value = 4.343e-16
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  127.1055
```

```
qb_aic_value_org <- AIC(qb_poisson_model3)
qb_bic_value_org <- BIC(qb_poisson_model3)
qb_aic_value <- AIC(qb_interaction_model)
qb_bic_value <- BIC(qb_interaction_model)

qb_aic_value_org
```

```
[1] 44279.06
```

```
qb_bic_value_org
```

```
[1] 44309.35
```

```
qb_aic_value
```

```
[1] 29192.34
```

```
qb_bic_value
```

```
[1] 29229.37
```

```
qb_negbin_model <- glm.nb(queensborobridge ~ factor(month) + factor(weekend) +  average_t

summary(qb_negbin_model)
```

```
Call:
glm.nb(formula = queensborobridge ~ factor(month) + factor(weekend) +
    average_temp + as.numeric(rain) + average_temp * as.numeric(rain),
    data = centered_queensborobridge_data, init.theta = 26.71052056,
    link = log)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   8.340899   0.042686 195.401  < 2e-16 ***
factor(month)August           0.182421   0.059749   3.053 0.002265 **
factor(month)July             0.052082   0.063149   0.825 0.409517
factor(month)June             0.133452   0.058782   2.270 0.023191 *
factor(month)May              0.117206   0.050696   2.312 0.020783 *
factor(month)October          0.226185   0.051554   4.387 1.15e-05 ***
factor(month)September        0.195243   0.056535   3.453 0.000553 ***
factor(weekend)1             -0.272855   0.029434  -9.270  < 2e-16 ***
average_temp                  0.012305   0.001975   6.232 4.62e-10 ***
as.numeric(rain)             -0.418544   0.039411 -10.620  < 2e-16 ***
average_temp:as.numeric(rain)  0.011667   0.003756   3.106 0.001894 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(26.7105) family taken to be 1)

    Null deviance: 633.42  on 213  degrees of freedom
Residual deviance: 215.91  on 203  degrees of freedom
AIC: 3514.1

Number of Fisher Scoring iterations: 1

              Theta:  26.71
          Std. Err.:  2.59

 2 x log-likelihood:  -3490.13
```

```
#GOF
anova(qb_negbin_model, test = "LRT")
```

```
Warning in anova.negbin(qb_negbin_model, test = "LRT"): tests made without
re-estimating 'theta'
```

```
Analysis of Deviance Table

Model: Negative Binomial(26.7105), link: log

Response: queensborobridge

Terms added sequentially (first to last)

                             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                           213     633.42
factor(month)                 6   88.118       207     545.31 < 2.2e-16 ***
factor(weekend)               1   95.466       206     449.84 < 2.2e-16 ***
average_temp                  1   53.646       205     396.19 2.401e-13 ***
as.numeric(rain)              1  169.809       204     226.39 < 2.2e-16 ***
average_temp:as.numeric(rain) 1   10.471       203     215.91  0.001213 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#3 Manhattan Bridge

How are our independent variables related to the outcome?

```
NY_cycling_data %>%
  group_by(weekend) %>%
  summarise(n = n(), mean = mean(manhattanbridge), sd = sd(manhattanbridge)) %>%arrange(w
```

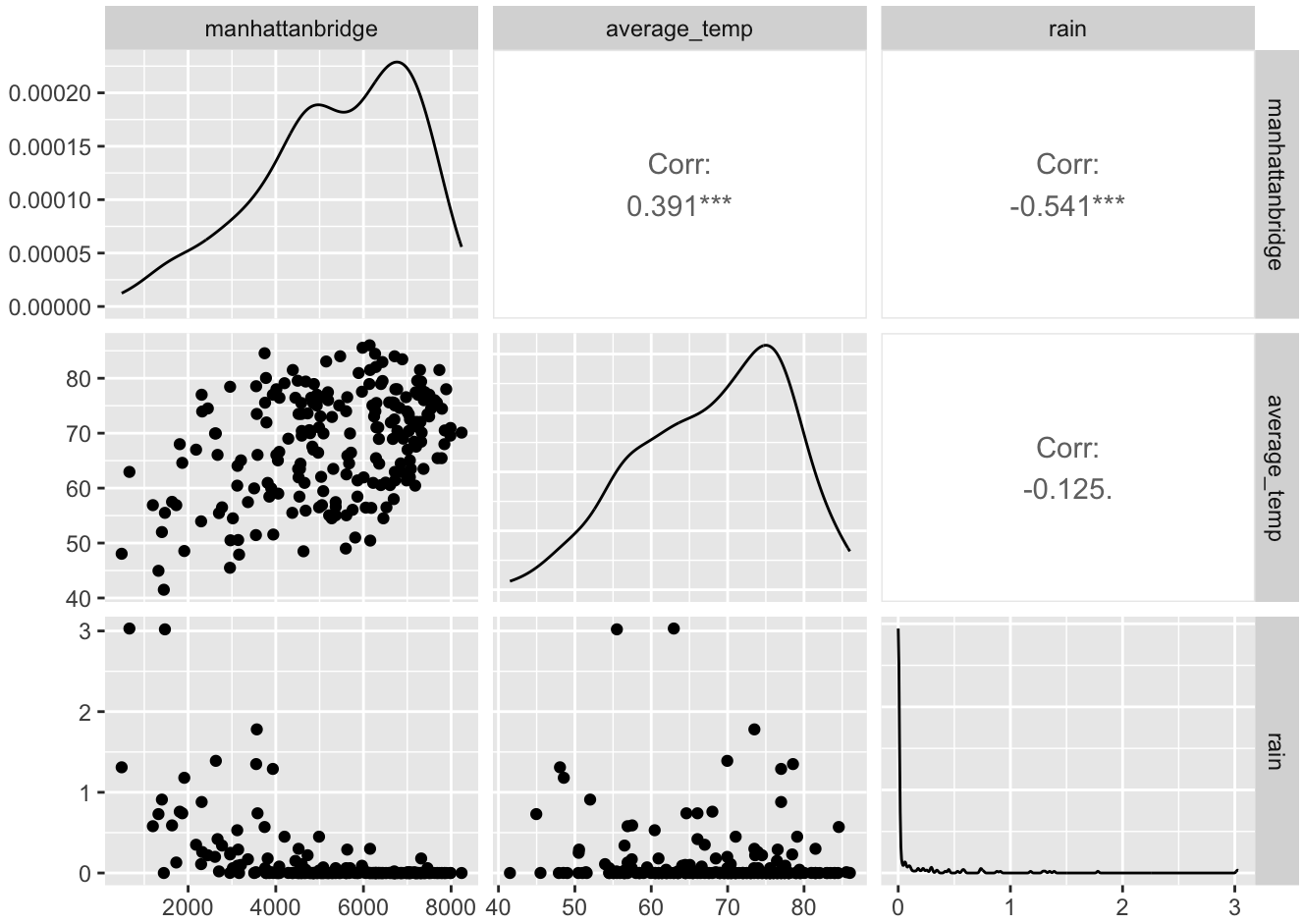
```
# A tibble: 2 × 4
  weekend     n  mean    sd
    <dbl> <int> <dbl> <dbl>
1       0   152 5859. 1700.
2       1    62 4087. 1110.
```

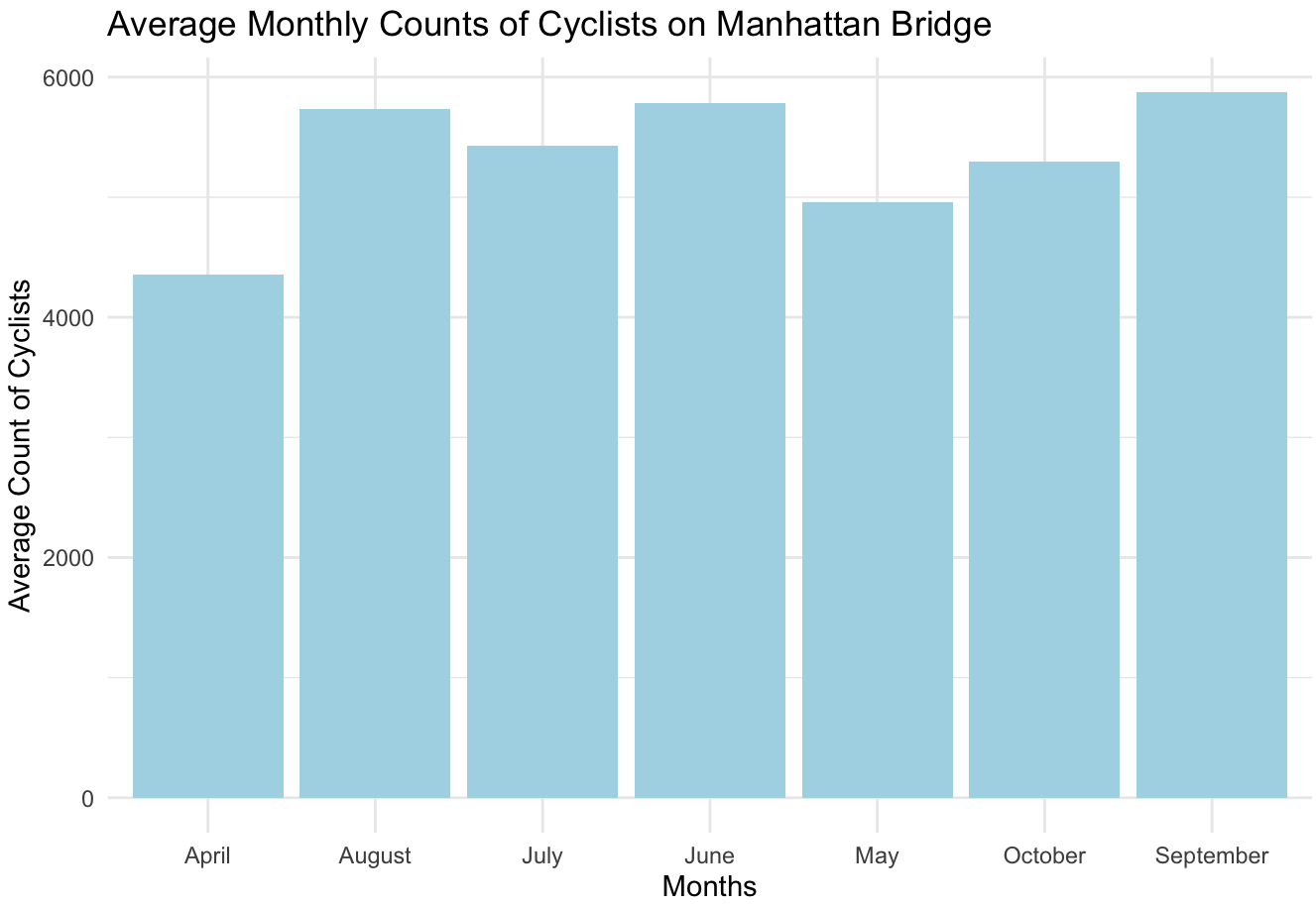
```
library(dplyr)
NY_cycling_data %>%
  group_by(month) %>%
  summarise(n = n(), mean = mean(manhattanbridge), sd = sd(manhattanbridge)) %>%
  mutate(month = factor(month, levels = c("April", "May", "June", "July", "August", "Sept
  arrange(month)
```

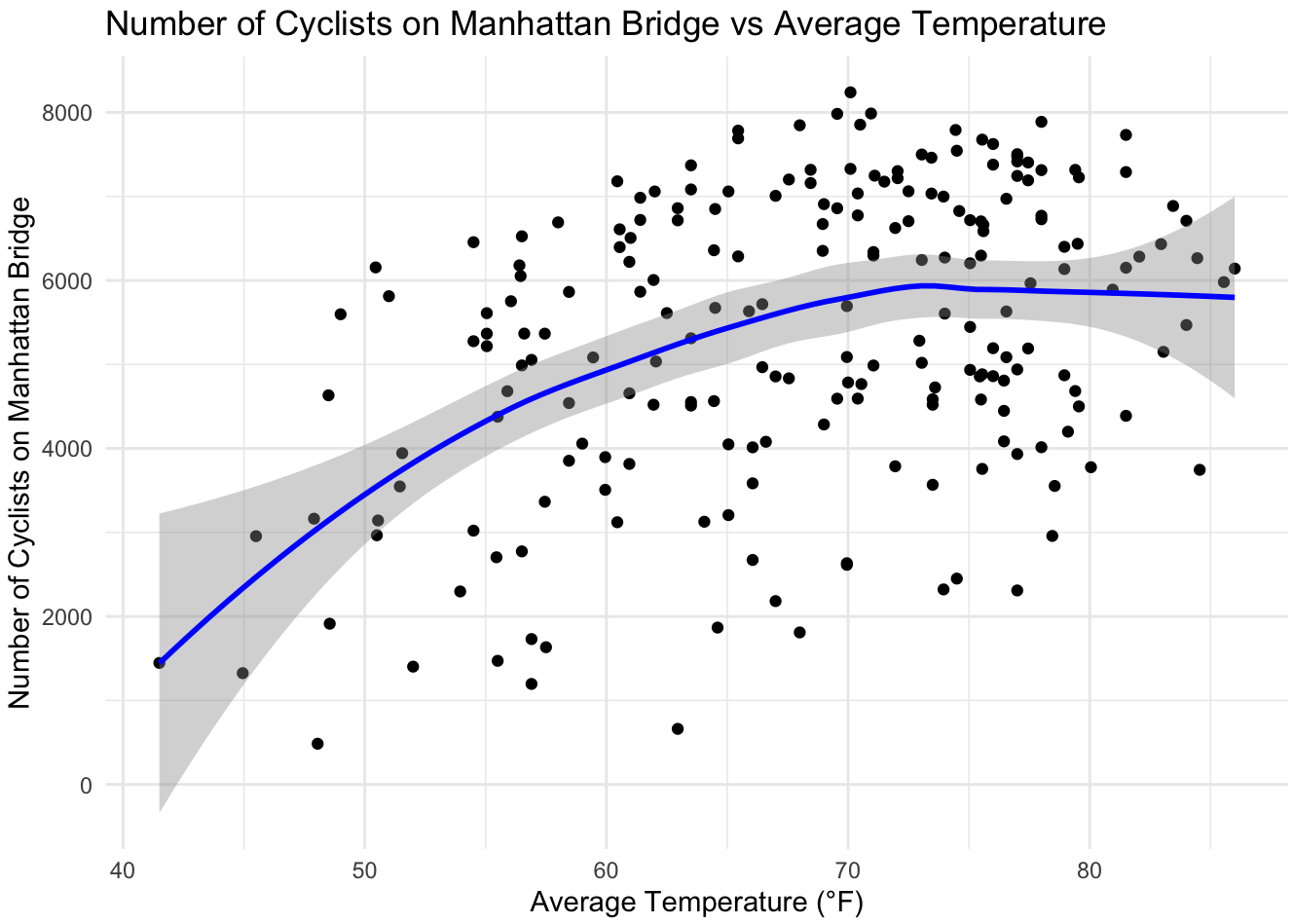
```
# A tibble: 7 × 4
  month         n  mean    sd
  <fct>     <int> <dbl> <dbl>
1 April        30 4354. 1693.
2 May          31 4959. 2064.
```

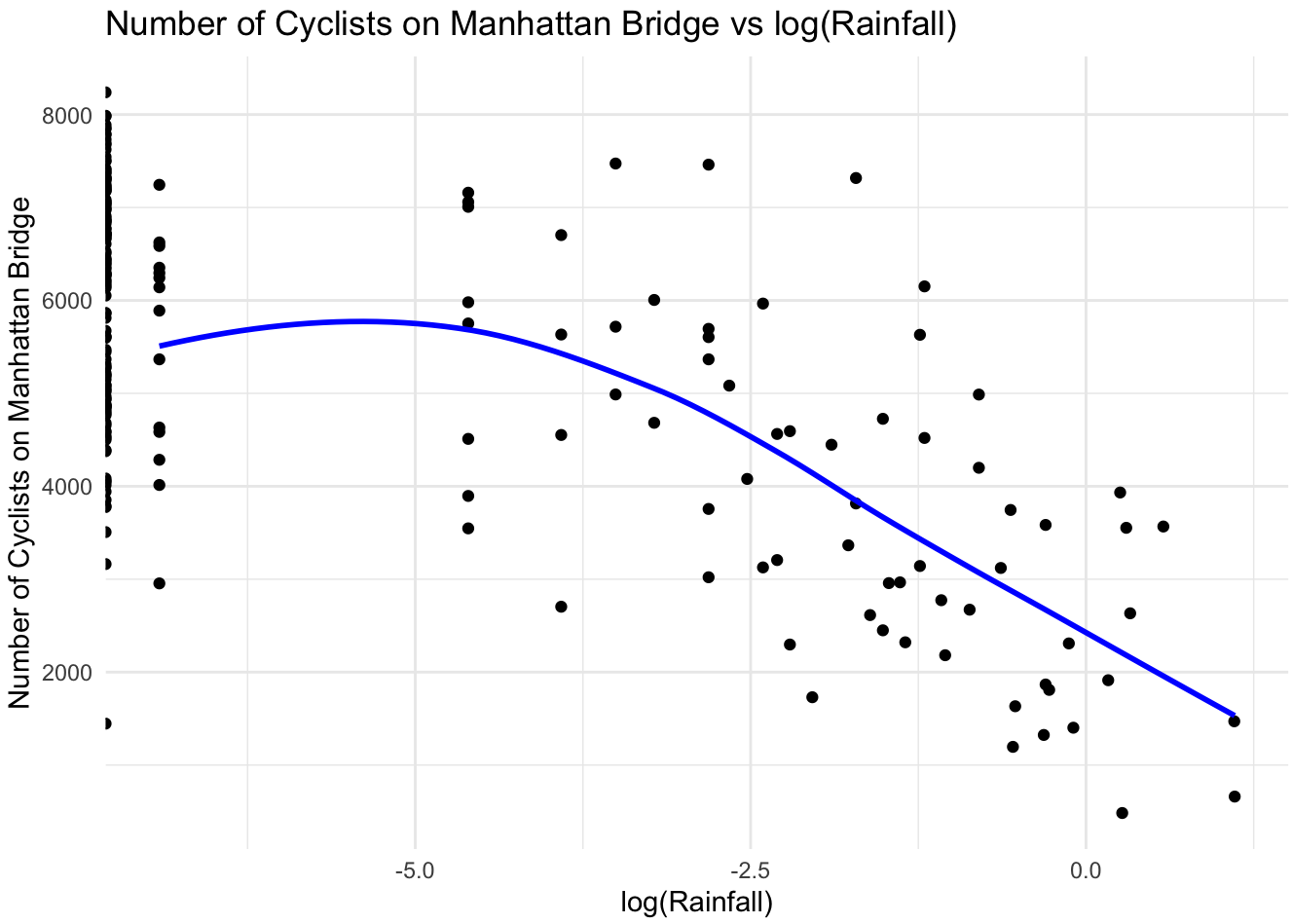
```
3 June         30 5780. 1664.
4 July         31 5425. 1566.
5 August       31 5730. 1515.
6 September    30 5871. 1614.
7 October      31 5298. 1710.
```

```
library(GGally)
mb_correlation_data <- NY_cycling_data[, c("manhattanbridge", "average_temp", "rain")]
ggpairs(mb_correlation_data)
```




Average Monthly Counts of Cyclists on Manhattan Bridge
Average Count of Cyclists
6000
4000
2000
0
April
August
July
June
May
October
September
Months


Number of Cyclists on Manhattan Bridge vs Average Temperature
Number of Cyclists on Manhattan Bridge
8000
6000
4000
2000
0
40
50
60
70
80
Average Temperature (°F)



Modeling:

I will be centering since there's interaction:

```
#Centering on its mean

manhattanbridge_data <-NY_cycling_data %>% select(manhattanbridge,rain, average_temp, wee

centered_manhattanbridge_data <- manhattanbridge_data %>%
  mutate(rain = rain - mean(rain), average_temp = average_temp - mean(average_temp))
centered_manhattanbridge_data
```

```
# A tibble: 214 × 5
   manhattanbridge   rain average_temp weekend month
             <dbl>  <dbl>        <dbl>   <dbl> <chr>
 1            1446 -0.132       -26.6        1 April
 2            3943 -0.132       -16.6        1 April
 3            4988 -0.102       -11.6        0 April
```

```
 4            1913  1.05        -19.6        0 April
 5            5276 -0.132       -13.6        0 April
 6            1324  0.598       -23.2        0 April
 7            2955 -0.131       -22.6        0 April
 8            3163 -0.132       -20.2        1 April
 9            4377 -0.132       -12.6        1 April
10            6359 -0.132        -3.66       0 April
# ℹ 204 more rows
```

```
#Check the relationship between varaibles --
#Against the Null Model
mb_null_poisson_model <- glm(manhattanbridge ~ 1, family = poisson(link = "log"), data =
summary(mb_null_poisson_model)
```

```
Call:
glm(formula = manhattanbridge ~ 1, family = poisson(link = "log"),
    data = centered_manhattanbridge_data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 8.584008   0.000935    9181   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 139331  on 213  degrees of freedom
Residual deviance: 139331  on 213  degrees of freedom
AIC: 141546

Number of Fisher Scoring iterations: 4
```

```
#average_temp + Rain
mb_poisson_model2 <- glm(manhattanbridge ~ average_temp + as.numeric(rain), family = pois

#Average_ + Rain + month
mb_poisson_model3 <- glm(manhattanbridge ~average_temp + as.numeric(rain) + factor(month)
```

```
#Full Model (Sequence) Added average_temp + rain + month + days
mb_full_poisson_model <- glm(manhattanbridge ~ average_temp + as.numeric(rain) + factor(m

mb_interaction_model <- glm(manhattanbridge ~  factor(month) + factor(weekend) + average_
```

```
anova(mb_null_poisson_model,mb_poisson_model2,mb_poisson_model3, mb_full_poisson_model, t
```

```
Analysis of Deviance Table

Model 1: manhattanbridge ~ 1
Model 2: manhattanbridge ~ average_temp + as.numeric(rain)
Model 3: manhattanbridge ~ average_temp + as.numeric(rain) + factor(month)
Model 4: manhattanbridge ~ average_temp + as.numeric(rain) + factor(month) +
    factor(weekend)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       213     139331
2       211      73891  2    65440 < 2.2e-16 ***
3       205      71226  6     2665 < 2.2e-16 ***
4       204      49048  1    22178 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pois_pearson_gof(mb_full_poisson_model) #>>> 1.445647e-193 BAD FIT
```

```
$pval
[1] 0

$df
[1] 204
```

```
#pois_dev_gof(bb_full_poisson_model)
#anova(bb_full_poisson_model, test= "LRT")
```

MFP:

```
#MFP

mfp::mfp(manhattanbridge ~ fp(average_temp),
         family = poisson,
         data = centered_manhattanbridge_data)
```

```
Call:
mfp::mfp(formula = manhattanbridge ~ fp(average_temp), data =
centered_manhattanbridge_data,
    family = poisson)
```

```
Deviance table:
         Resid. Dev
Null model   139330.7
Linear model     120378.5
Final model  112140.3

Fractional polynomials:
             df.initial select alpha df.final power1 power2
average_temp          4      1  0.05        4      1      1


Transformations of covariates:

formula
average_temp
I(((average_temp+26.7)/10)^1)+I(((average_temp+26.7)/10)^1*log(((average_temp+26.7)/10)))

Coefficients:
     Intercept  average_temp.1  average_temp.2
        7.3486          0.8667         -0.3879

Degrees of Freedom: 213 Total (i.e. Null);  211 Residual
Null Deviance:      139300
Residual Deviance: 112100   AIC: 114400
```

Interaction and Confouding:

```r
# Fit the full Poisson regression model
mb_full_poisson_model <- glm(manhattanbridge ~  factor(month) + factor(weekend)+ average_

# Include interaction term between average_temp and rain
mb_interaction_model <- glm(manhattanbridge ~  factor(month) + factor(weekend) + average_

# Compare the models using ANOVA
anova_result <- anova(mb_full_poisson_model, mb_interaction_model, test = "LRT")
print(anova_result)
```

```
Analysis of Deviance Table

Model 1: manhattanbridge ~ factor(month) + factor(weekend) + average_temp +
    as.numeric(rain)
Model 2: manhattanbridge ~ factor(month) + factor(weekend) + average_temp +
    as.numeric(rain) + average_temp * as.numeric(rain)
```

```
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       204      49048
2       203      46481  1   2566.8 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overdispersion Test and AIC/BIC:

```
mb_interaction_model <- glm(manhattanbridge ~ average_temp * as.numeric(rain) + factor(mo

AER::dispersiontest(mb_interaction_model)
```

```

    Overdispersion test

data:  mb_interaction_model
z = 5.8533, p-value = 2.41e-09
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  218.9714
```

```
# Obtain AIC and BIC
bb_aic_value_org <- AIC(mb_poisson_model3)
bb_bic_value_org <- BIC(mb_poisson_model3)
bb_aic_value <- AIC(mb_interaction_model)
bb_bic_value <- BIC(mb_interaction_model)

bb_aic_value_org
```

```
[1] 73457.29
```

```
bb_bic_value_org
```

```
[1] 73487.58
```

```
bb_aic_value
```

```
[1] 48716.9
```

```
bb_bic_value
```

```
[1] 48753.93
```

How is our outcome variable distributed?

```
centered_manhattanbridge_data %>% select(manhattanbridge) %>% skimr::skim()
```

Data summary	
Name	Piped data
Number of rows	214
Number of columns	1
_____	
Column type frequency:	
numeric	1
_____	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
manhattanbridge	0	1	5345.49	1746.55	484	4308	5607.5	6759.5	8239	▁▂▆▇▃

```
mb_negbin_model <- glm.nb(manhattanbridge ~ factor(month) + factor(weekend) +  average_te

summary(mb_negbin_model)
```

```
Call:
glm.nb(formula = manhattanbridge ~ factor(month) + factor(weekend) +
    average_temp + as.numeric(rain) + average_temp * as.numeric(rain),
    data = centered_manhattanbridge_data, init.theta = 16.45457534,
    link = log)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                  8.586228   0.054290 158.156  < 2e-16 ***
factor(month)August          0.032358   0.076011   0.426   0.6703
factor(month)July           -0.019426   0.080330  -0.242   0.8089
```

```
factor(month)June              0.090065   0.074774   1.204   0.2284
factor(month)May               0.075805   0.064473   1.176   0.2397
factor(month)October           0.094629   0.065580   1.443   0.1490
factor(month)September         0.093956   0.071918   1.306   0.1914
factor(weekend)1              -0.333854   0.037442  -8.917  < 2e-16 ***
average_temp                   0.014117   0.002512   5.620 1.91e-08 ***
as.numeric(rain)              -0.547218   0.050117 -10.919  < 2e-16 ***
average_temp:as.numeric(rain)  0.010468   0.004773   2.193   0.0283 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(16.4546) family taken to be 1)

    Null deviance: 546.42  on 213  degrees of freedom
Residual deviance: 216.68  on 203  degrees of freedom
AIC: 3678.2

Number of Fisher Scoring iterations: 1

              Theta:  16.45
          Std. Err.:  1.58

 2 x log-likelihood:  -3654.23
```

```
anova(mb_negbin_model, test = "LRT")
```

```
Warning in anova.negbin(mb_negbin_model, test = "LRT"): tests made without
re-estimating 'theta'

Analysis of Deviance Table

Model: Negative Binomial(16.4546), link: log

Response: manhattanbridge

Terms added sequentially (first to last)

                              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                            213     546.42
factor(month)                  6   32.384       207     514.04 1.377e-05 ***
factor(weekend)                1   84.754       206     429.28 < 2.2e-16 ***
average_temp                   1   43.113       205     386.17 5.167e-11 ***
as.numeric(rain)               1  164.130       204     222.04 < 2.2e-16 ***
average_temp:as.numeric(rain)  1    5.358       203     216.68   0.02062 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#4 Williamsburg Bridge

How are our independent variables related to the outcome?

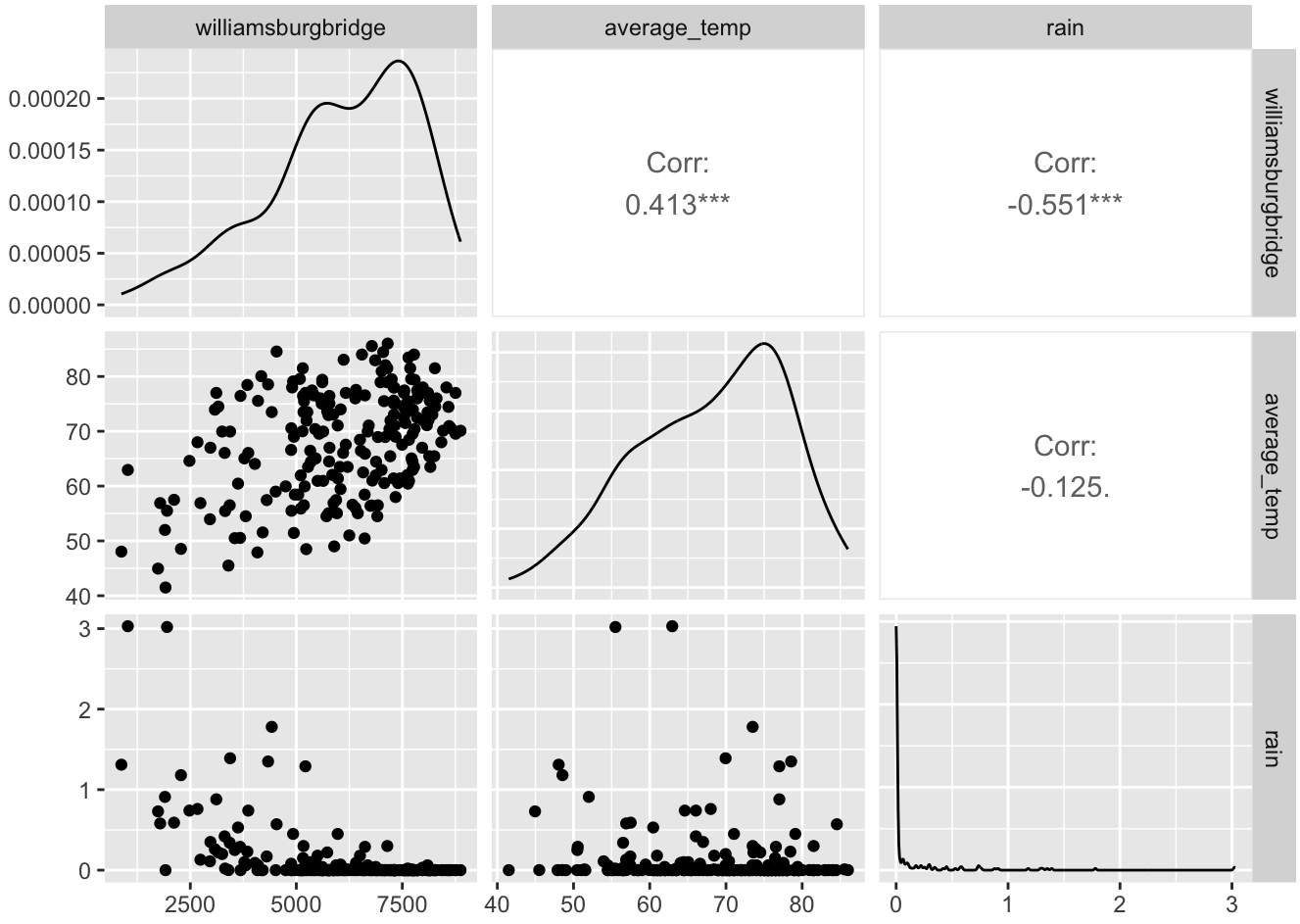
```
NY_cycling_data %>%
  group_by(weekend) %>%
  summarise(n = n(), mean = mean(williamsburgbridge), sd = sd(williamsburgbridge)) %>%arr
```

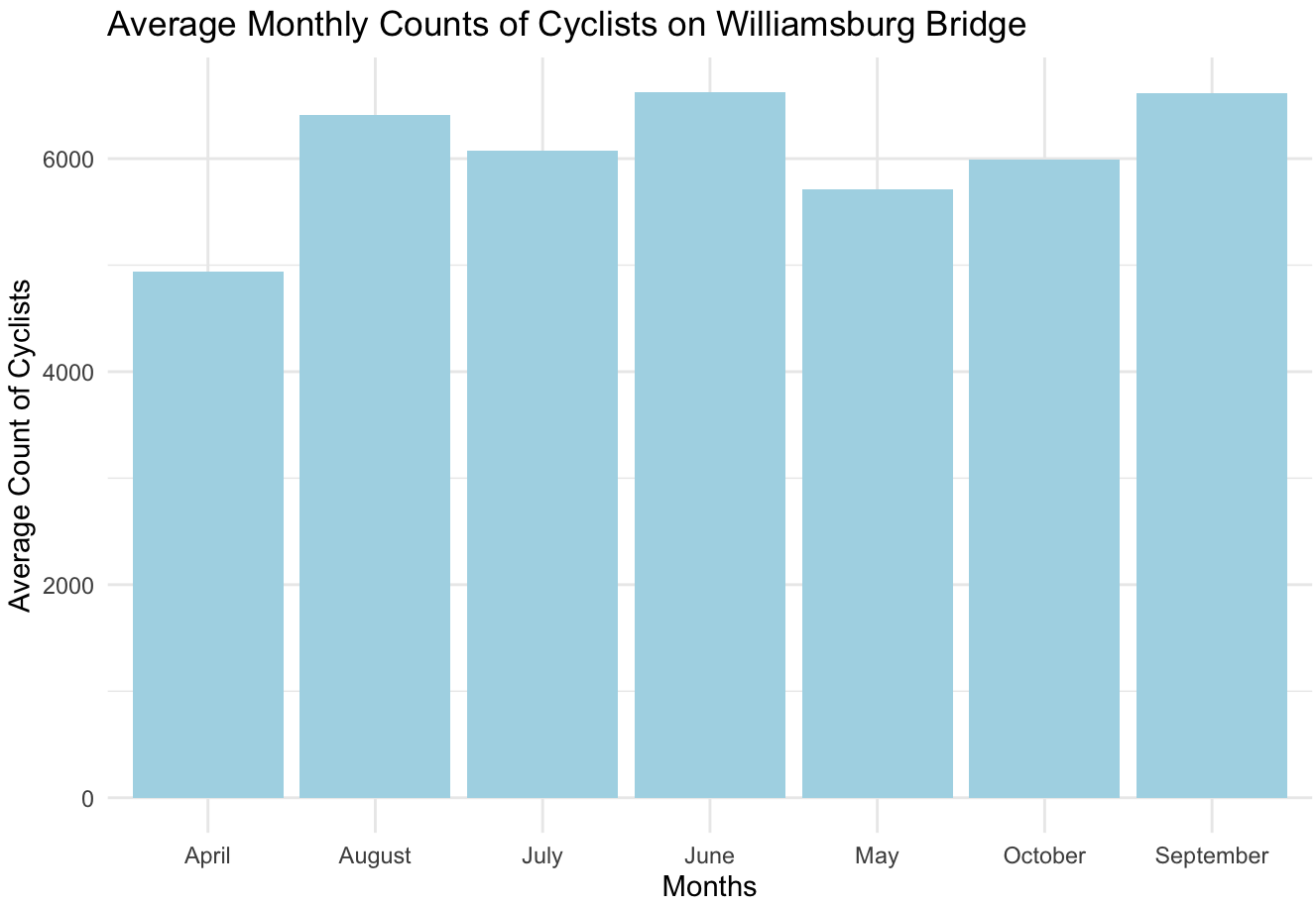
```
# A tibble: 2 × 4
  weekend     n  mean    sd
    <dbl> <int> <dbl> <dbl>
1       0   152 6535. 1725.
2       1    62 4868. 1180.
```

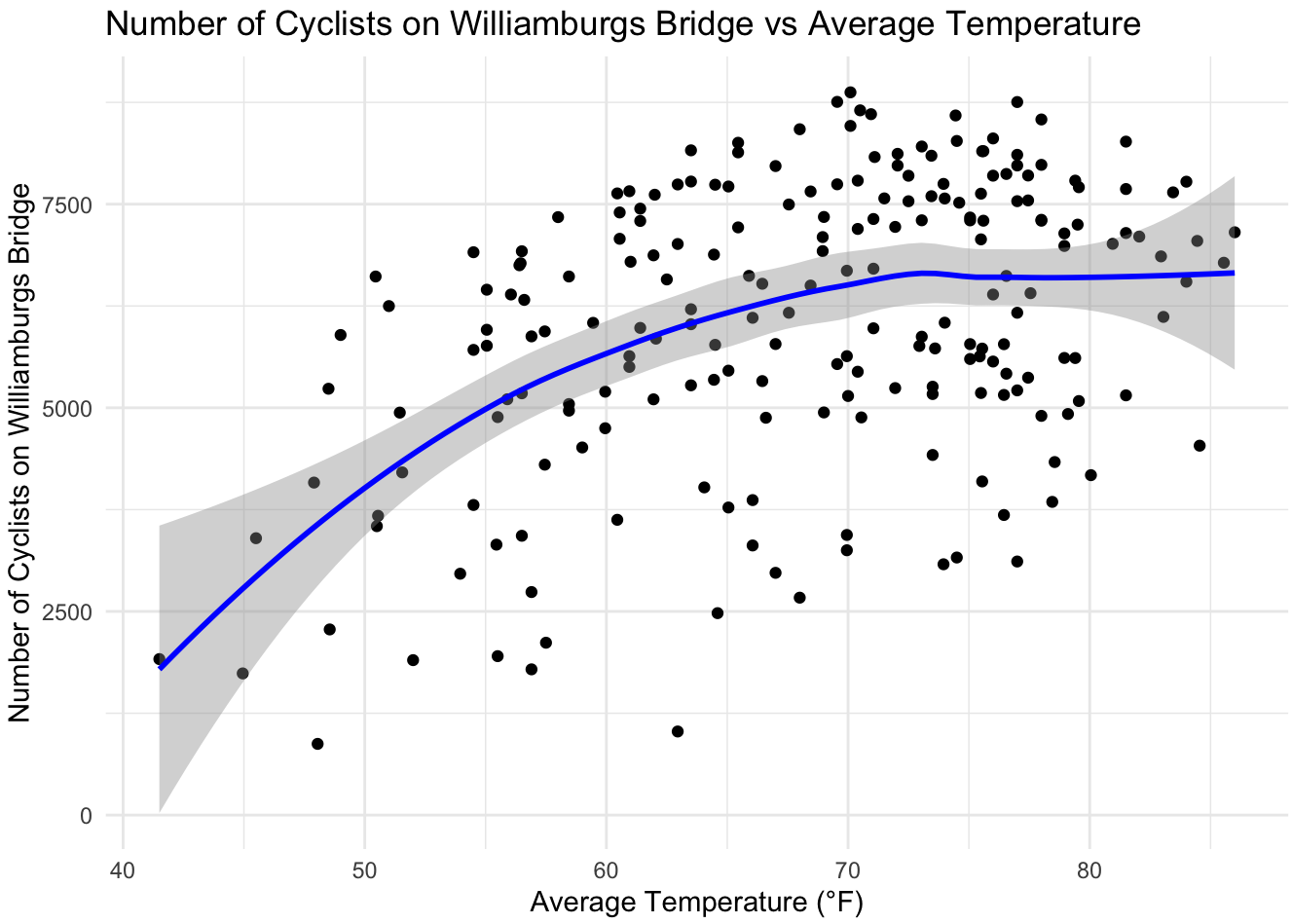
```
library(dplyr)
NY_cycling_data %>%
  group_by(month) %>%
  summarise(n = n(), mean = mean(williamsburgbridge), sd = sd(williamsburgbridge)) %>%
  mutate(month = factor(month, levels = c("April", "May", "June", "July", "August", "Sept
  arrange(month)
```

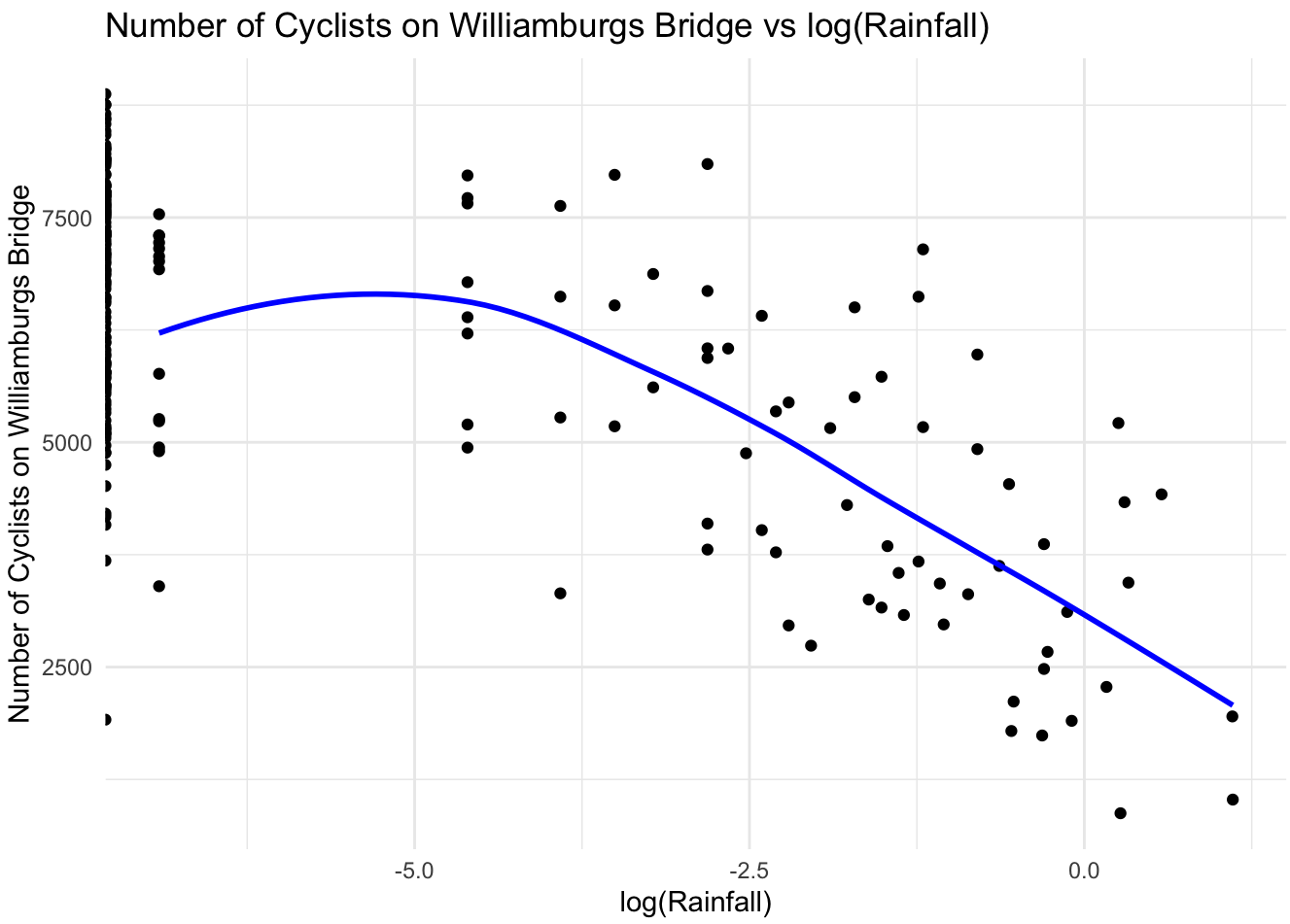
```
# A tibble: 7 × 4
  month         n  mean    sd
  <fct>     <int> <dbl> <dbl>
1 April        30 4942. 1734.
2 May          31 5710. 2007.
3 June         30 6620  1599.
4 July         31 6074. 1576.
5 August       31 6407. 1490.
6 September    30 6614. 1642.
7 October      31 5995. 1746.
```

```
library(GGally)
wb_correlation_data <- NY_cycling_data[, c("williamsburgbridge", "average_temp", "rain")]
ggpairs(wb_correlation_data)
```


williamsburgbridge
average_temp
rain
Corr:
0.413***
Corr:
-0.551***
Corr:
-0.125.
0.00020
0.00015
0.00010
0.00005
0.00000
80
70
60
50
40
3
2
1
0
2500
5000
7500
40
50
60
70
80
0
1
2
3


Average Monthly Counts of Cyclists on Williamsburg Bridge
Average Count of Cyclists
6000
4000
2000
0
April
August
July
June
May
October
September
Months


Number of Cyclists on Williamburgs Bridge vs Average Temperature
Number of Cyclists on Williamburgs Bridge
0
2500
5000
7500
40
50
60
70
80
Average Temperature (°F)



Modeling:

I will be centering since there's interaction:

```
#Centering on its mean
williamsburgbridge_data <-NY_cycling_data %>% select(williamsburgbridge,rain, average_tem

centered_williamsburgbridge_data <- williamsburgbridge_data %>%
  mutate(rain = rain - mean(rain), average_temp = average_temp - mean(average_temp))
```

```
#Check the relationship between varaibles --
#Against the Null Model
wb_null_poisson_model <- glm(williamsburgbridge ~ 1, family = poisson(link = "log"), data
summary(wb_null_poisson_model)
```

```
Call:
```

```
glm(formula = williamsburgbridge ~ 1, family = poisson(link = "log"),
    data = centered_williamsburgbridge_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 8.7080877  0.0008787    9910   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 123257  on 213  degrees of freedom
Residual deviance: 123257  on 213  degrees of freedom
AIC: 125503

Number of Fisher Scoring iterations: 4
```

```
#average_temp + Rain
wb_poisson_model2 <- glm(williamsburgbridge ~ average_temp + as.numeric(rain), family = p

#Average_ + Rain + month
wb_poisson_model3 <- glm(williamsburgbridge ~average_temp + as.numeric(rain) + factor(mon
```

```
#Full Model (Sequence) Added average_temp + rain + month + days
wb_full_poisson_model <- glm(williamsburgbridge ~ average_temp + as.numeric(rain) + facto

wb_interaction_model <- glm(williamsburgbridge ~  factor(month) + factor(weekend) + avera


anova(wb_null_poisson_model,wb_poisson_model2,wb_poisson_model3, wb_full_poisson_model, t
```

```
Analysis of Deviance Table

Model 1: williamsburgbridge ~ 1
Model 2: williamsburgbridge ~ average_temp + as.numeric(rain)
Model 3: williamsburgbridge ~ average_temp + as.numeric(rain) + factor(month)
Model 4: williamsburgbridge ~ average_temp + as.numeric(rain) + factor(month) +
    factor(weekend)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       213     123257
2       211      64038  2    59220 < 2.2e-16 ***
3       205      60339  6     3699 < 2.2e-16 ***
4       204      43805  1    16534 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pois_pearson_gof(wb_full_poisson_model)
```

```
$pval
[1] 0

$df
[1] 204
```

```
#pois_dev_gof(bb_full_poisson_model)
#anova(bb_full_poisson_model, test= "LRT")
```

MFP:

```
#MFP

mfp::mfp(williamsburgbridge ~ fp(average_temp),
         family = poisson,
         data = centered_williamsburgbridge_data)
```

```
Call:
mfp::mfp(formula = williamsburgbridge ~ fp(average_temp), data =
centered_williamsburgbridge_data,
    family = poisson)


Deviance table:
         Resid. Dev
Null model   123257.2
Linear model     104421.4
Final model  96228.34

Fractional polynomials:
             df.initial select alpha df.final power1 power2
average_temp          4      1  0.05        4      1      1


Transformations of covariates:

formula
average_temp
I(((average_temp+26.7)/10)^1)+I(((average_temp+26.7)/10)^1*log(((average_temp+26.7)/10)))

Coefficients:
```

```
     Intercept  average_temp.1  average_temp.2
        7.5640          0.8031         -0.3594

Degrees of Freedom: 213 Total (i.e. Null);  211 Residual
Null Deviance:      123300
Residual Deviance: 96230    AIC: 98480
```

Interaction and Confouding:

```
# Fit the full Poisson regression model
wb_full_poisson_model <- glm(williamsburgbridge ~  factor(month) + factor(weekend)+ avera

# Include interaction term between average_temp and rain
wb_interaction_model <- glm(williamsburgbridge ~  factor(month) + factor(weekend) + avera

# Compare the models using ANOVA
wb_anova_result <- anova(wb_full_poisson_model, wb_interaction_model, test = "LRT")
print(wb_anova_result)
```

```
Analysis of Deviance Table

Model 1: williamsburgbridge ~ factor(month) + factor(weekend) + average_temp +
    as.numeric(rain)
Model 2: williamsburgbridge ~ factor(month) + factor(weekend) + average_temp +
    as.numeric(rain) + average_temp * as.numeric(rain)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       204      43805
2       203      41193  1   2611.8 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overdispersion Test and AIC/BIC:

```
wb_interaction_model <- glm(williamsburgbridge ~ average_temp * as.numeric(rain) + factor

AER::dispersiontest(wb_interaction_model)
```

```
    Overdispersion test

data:  wb_interaction_model
z = 6.3159, p-value = 1.343e-10
```

```
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  189.7187
```

```
# Obtain AIC and BIC
bb_aic_value_org <- AIC(wb_poisson_model3)
bb_bic_value_org <- BIC(wb_poisson_model3)
bb_aic_value <- AIC(wb_interaction_model)
bb_bic_value <- BIC(wb_interaction_model)

bb_aic_value_org
```

```
[1] 62600.79
```

```
bb_bic_value_org
```

```
[1] 62631.09
```

```
bb_aic_value
```

```
[1] 43459.49
```

```
bb_bic_value
```

```
[1] 43496.51
```

How is our outcome variable distributed?

```
centered_manhattanbridge_data %>% select(manhattanbridge) %>% skimr::skim()
```

Data summary

Name	Piped data
Number of rows	214
Number of columns	1
Column type frequency:	
numeric	1

Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
manhattanbridge	0	1	5345.49	1746.55	484	4308	5607.5	6759.5	8239	▂▂▆▇▇

```
wb_negbin_model <- glm.nb(williamsburgbridge ~ factor(month) + factor(weekend) +  average

summary(wb_negbin_model)
```

```
Call:
glm.nb(formula = williamsburgbridge ~ factor(month) + factor(weekend) +
    average_temp + as.numeric(rain) + average_temp * as.numeric(rain),
    data = centered_williamsburgbridge_data, init.theta = 22.61322593,
    link = log)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                  8.693054   0.046327 187.646  < 2e-16 ***
factor(month)August          0.033580   0.064861   0.518  0.60466
factor(month)July           -0.022322   0.068547  -0.326  0.74469
factor(month)June            0.111304   0.063804   1.744  0.08108 .
factor(month)May             0.106213   0.055015   1.931  0.05353 .
factor(month)October         0.099706   0.055960   1.782  0.07479 .
factor(month)September       0.101551   0.061368   1.655  0.09797 .
factor(weekend)1            -0.267363   0.031948  -8.369  < 2e-16 ***
average_temp                 0.013237   0.002144   6.175 6.62e-10 ***
as.numeric(rain)            -0.463885   0.042748 -10.852  < 2e-16 ***
average_temp:as.numeric(rain)  0.011032   0.004072   2.709  0.00675 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(22.6132) family taken to be 1)

    Null deviance: 566.08  on 213  degrees of freedom
Residual deviance: 216.02  on 203  degrees of freedom
AIC: 3670.4

Number of Fisher Scoring iterations: 1
```

```
            Theta:  22.61
        Std. Err.:  2.18

 2 x log-likelihood:  -3646.364
```

```
#GOF
anova(wb_negbin_model, test = "LRT")
```

```
Warning in anova.negbin(wb_negbin_model, test = "LRT"): tests made without
re-estimating 'theta'

Analysis of Deviance Table

Model: Negative Binomial(22.6132), link: log

Response: williamsburgbridge

Terms added sequentially (first to last)

                            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                          213     566.08
factor(month)                6   41.745       207     524.34 2.064e-07 ***
factor(weekend)              1   76.868       206     447.47 < 2.2e-16 ***
average_temp                 1   51.747       205     395.72 6.312e-13 ***
as.numeric(rain)             1  171.615       204     224.11 < 2.2e-16 ***
average_temp:as.numeric(rain)  1    8.086       203     216.02  0.004461 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pois_dev_gof(bb_negbin_model)
```

```
$pval
[1] 0.2216194

$df
[1] 203
```

```
pois_dev_gof(wb_negbin_model)
```

```
$pval
[1] 0.2527467

$df
[1] 203
```

```
pois_dev_gof(qb_negbin_model)
```

```
$pval
[1] 0.2543887

$df
[1] 203
```

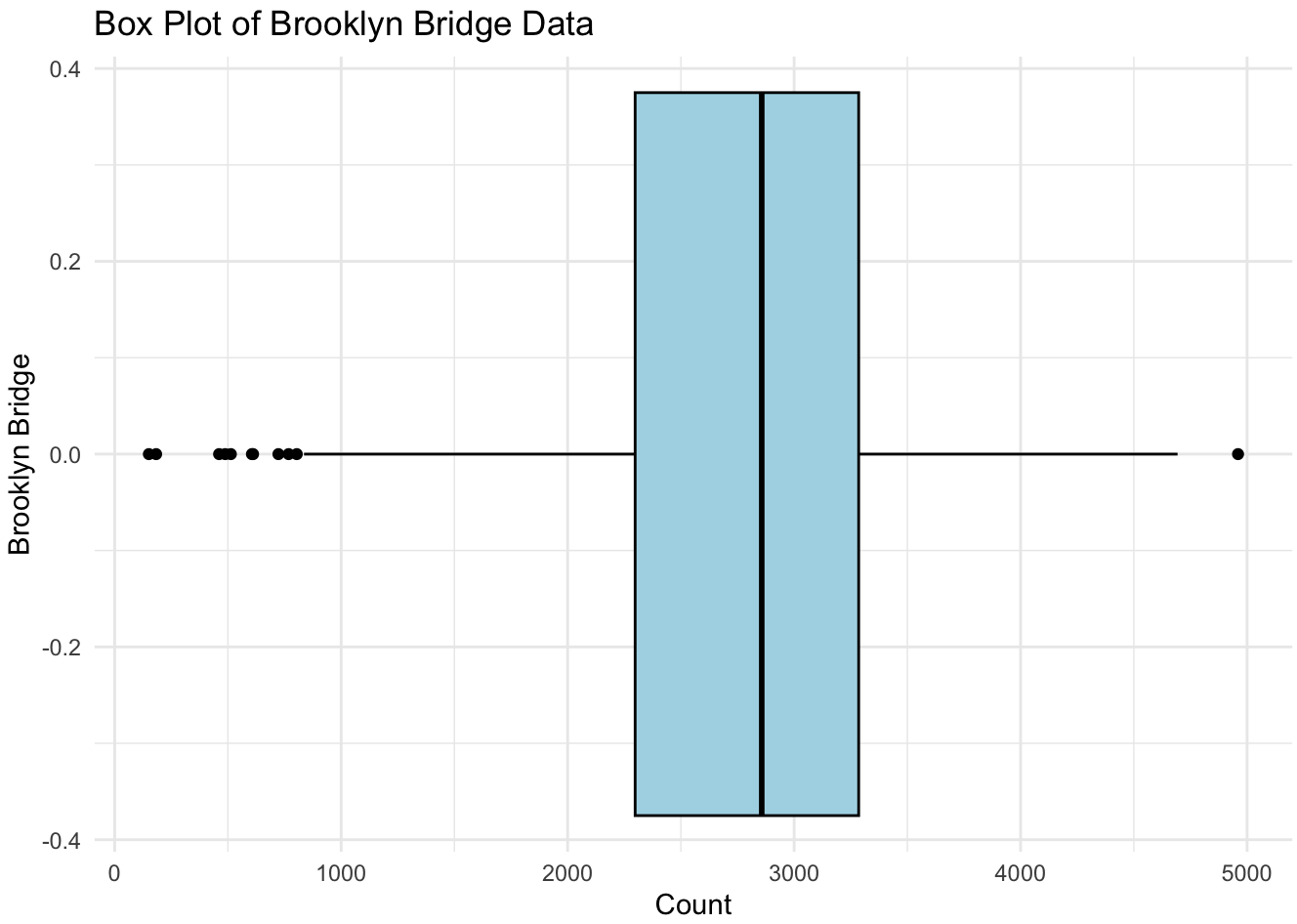
```
pois_dev_gof(mb_negbin_model)
```

```
$pval
[1] 0.2428304

$df
[1] 203
```

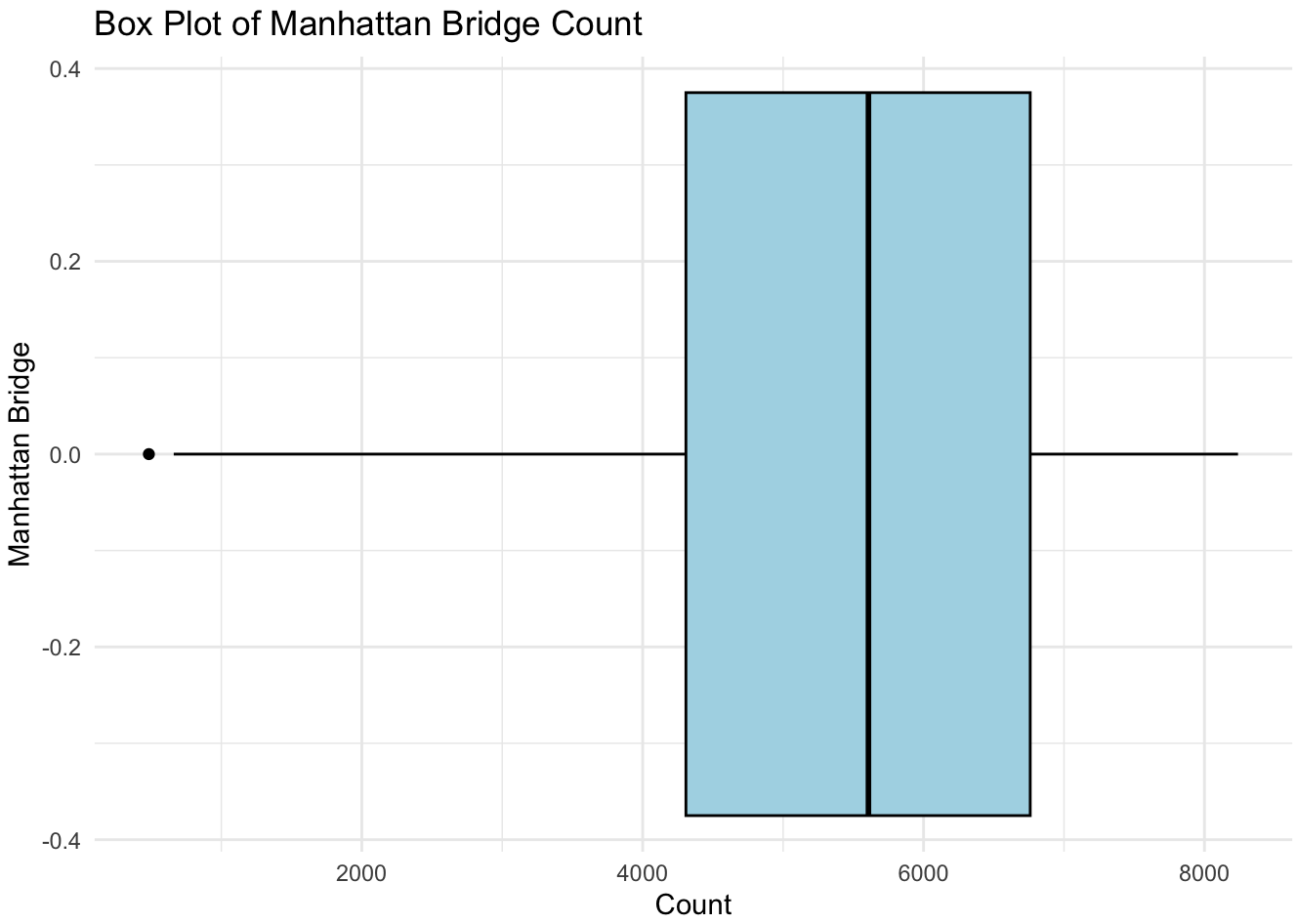
```
centered_brooklynbridge_data
```

```
# A tibble: 214 × 5
   brooklynbridge   rain average_temp weekend month
            <dbl>  <dbl>        <dbl>   <dbl> <chr>
 1            606 -0.132       -26.6        1 April
 2           2021 -0.132       -16.6        1 April
 3           2470 -0.102       -11.6        0 April
 4            723  1.05        -19.6        0 April
 5           2807 -0.132       -13.6        0 April
 6            461  0.598       -23.2        0 April
 7           1222 -0.131       -22.6        0 April
 8           1674 -0.132       -20.2        1 April
 9           2375 -0.132       -12.6        1 April
10           3324 -0.132        -3.66       0 April
# ℹ 204 more rows
```

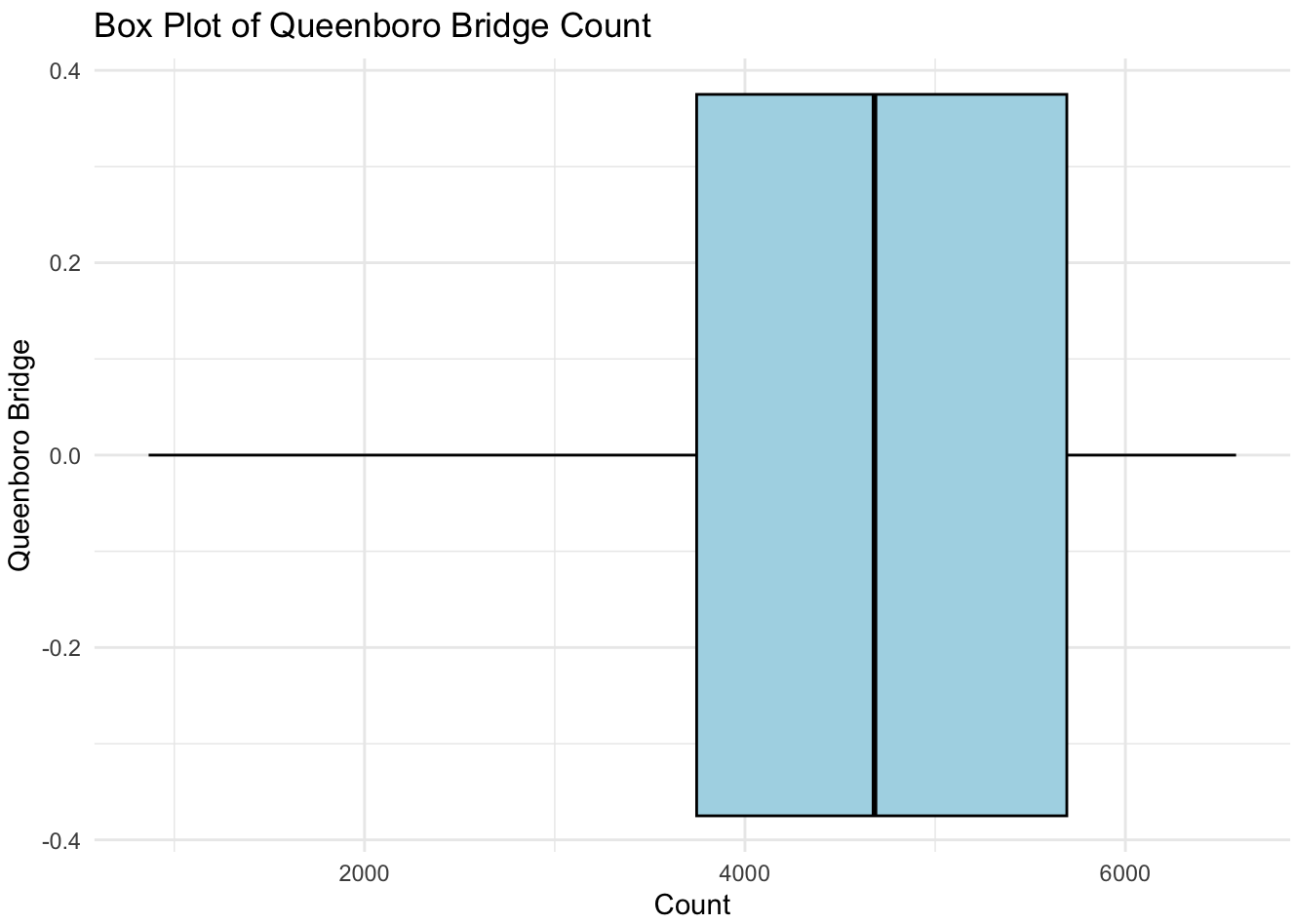
```
ggplot(NY_cycling_data, aes(x = brooklynbridge)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Box Plot of Brooklyn Bridge Data",
       y = "Brooklyn Bridge",
       x = "Count") +
  theme_minimal()
```



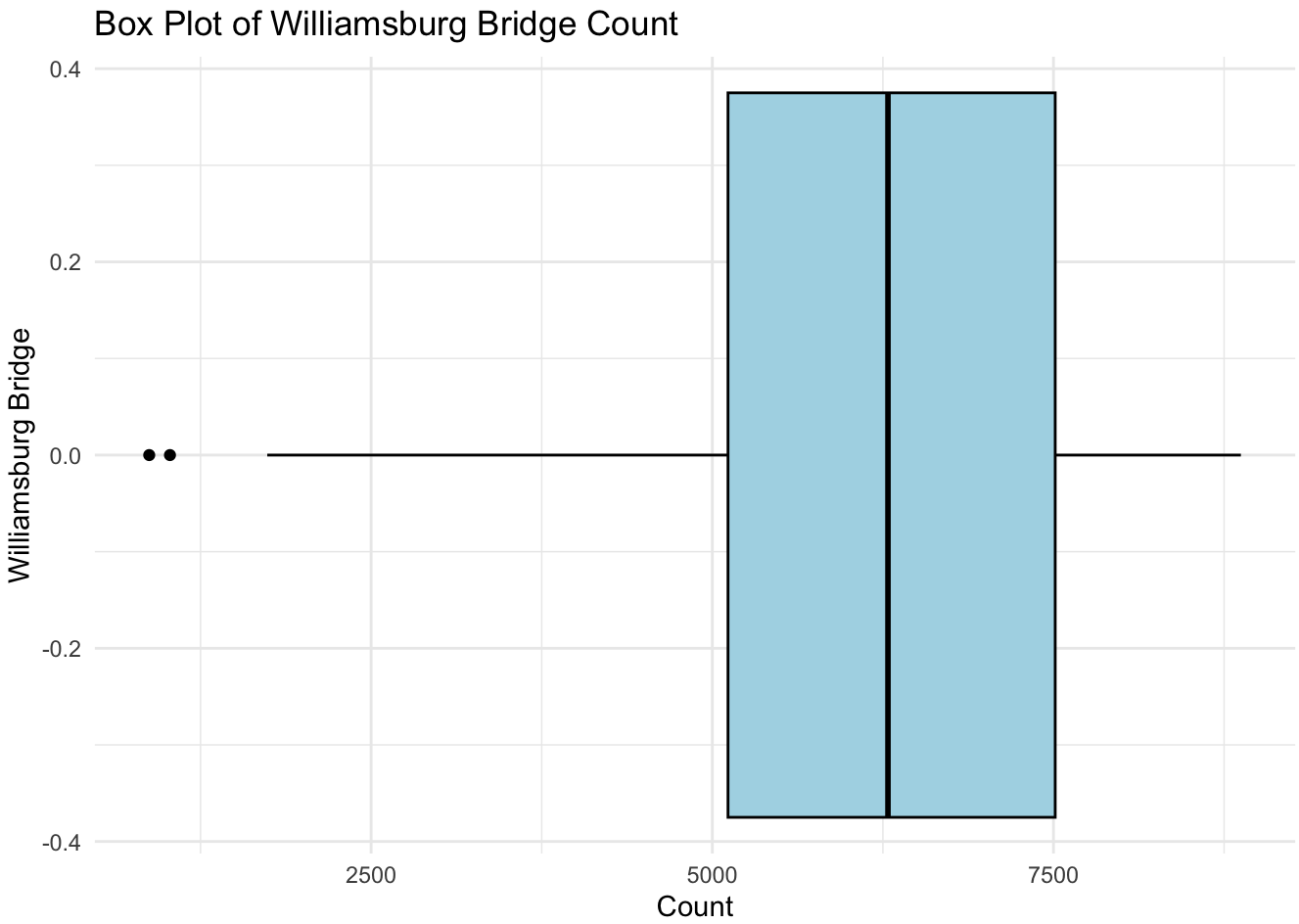

```
ggplot(NY_cycling_data, aes(x = manhattanbridge)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Box Plot of Manhattan Bridge Count",
       y = "Manhattan Bridge",
       x = "Count") +
  theme_minimal()
```




```
ggplot(NY_cycling_data, aes(x = queensborobridge)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Box Plot of Queenboro Bridge Count",
       y = "Queenboro Bridge",
       x = "Count") +
  theme_minimal()
```




```
ggplot(NY_cycling_data, aes(x = williamsburgbridge)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Box Plot of Williamsburg Bridge Count ",
       y = "Williamsburg Bridge",
       x = "Count") +
  theme_minimal()
```


Box Plot of Williamsburg Bridge Count
Williamsburg Bridge
0.4
0.2
0.0
-0.2
-0.4
2500
5000
7500
Count