pHExploration

Alexandra Lawrence

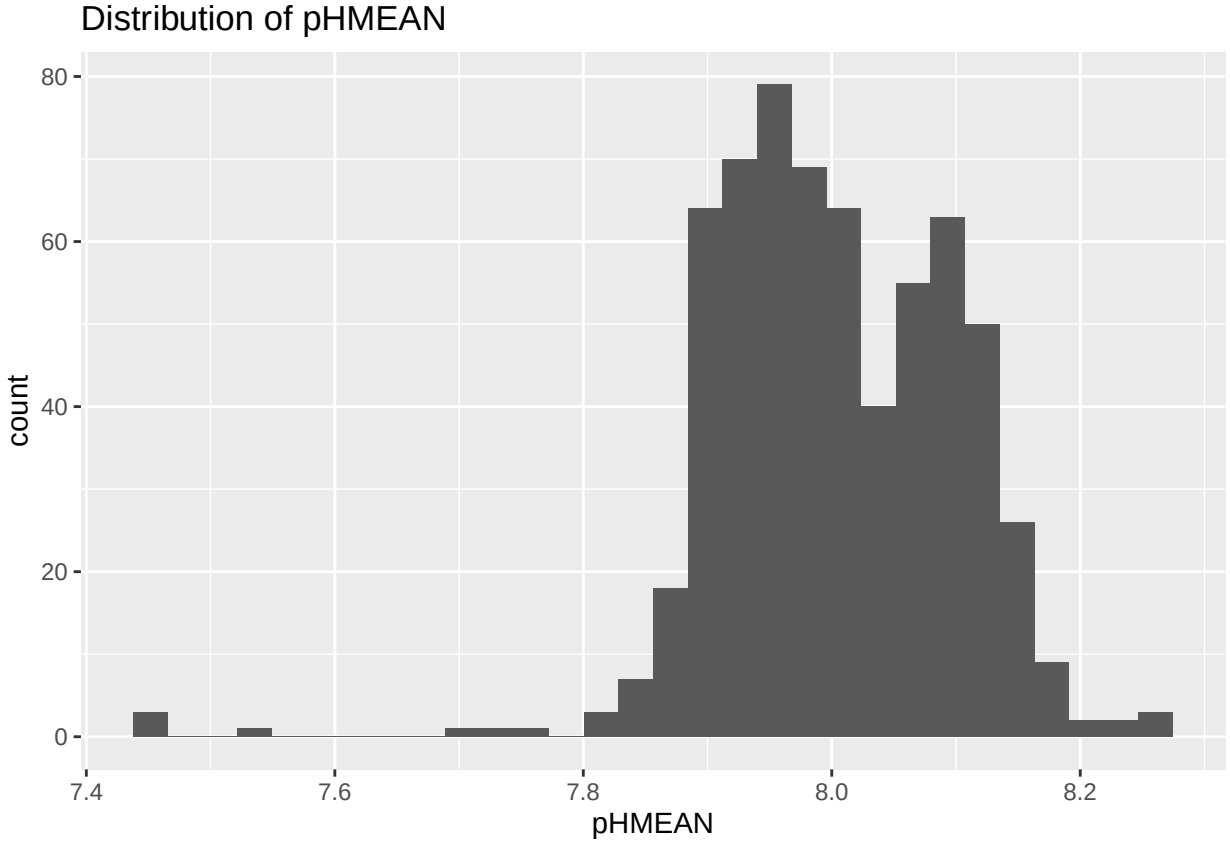
```
## Joining, by = "Date"
## Joining, by = "Date"
```

Exploring pH

In order to understand the variable that I want to explore, I want to look at the distribution of pH along with any patterns it may follow over time.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 48 rows containing non-finite values (stat_bin).
```

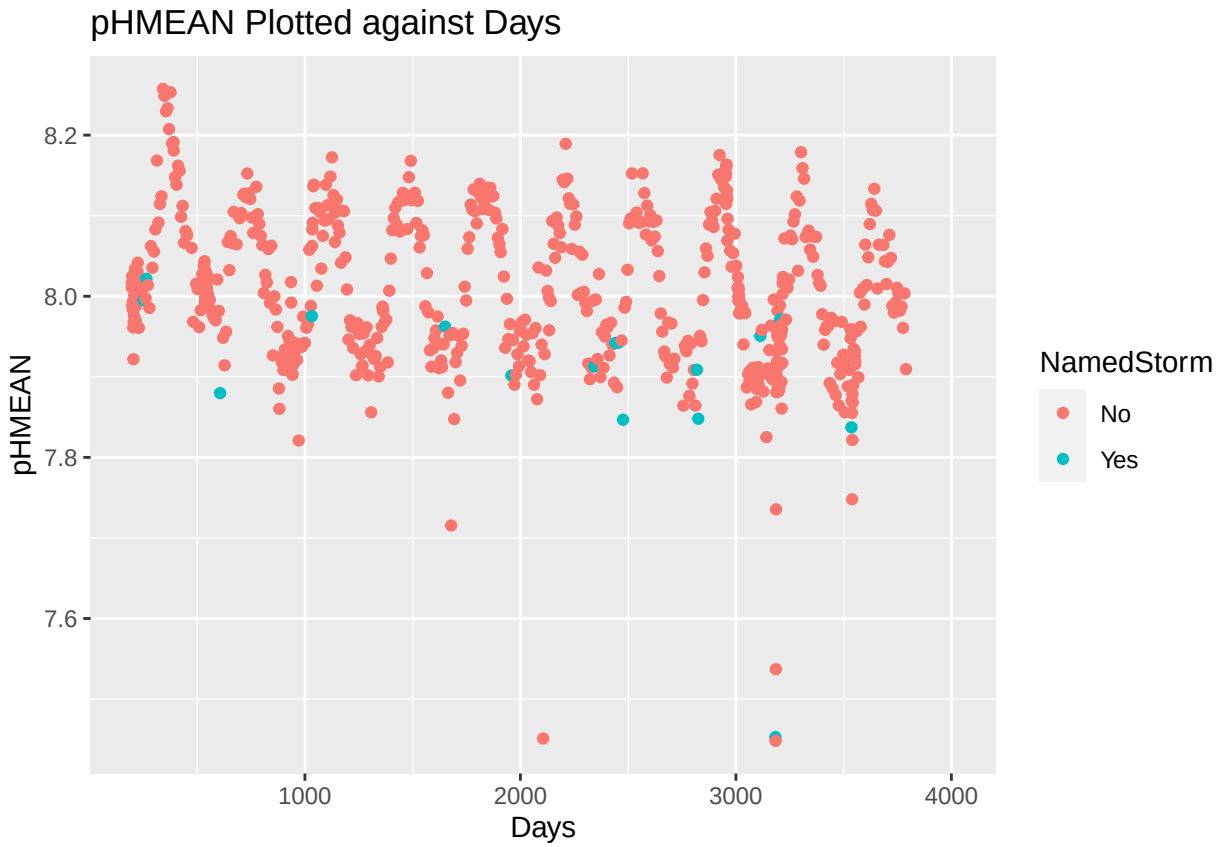


Min	Max	Mean	Median	Q1	Q3	Standard_Deviation	IQR
7.448	8.257	8.002	7.995	7.935	8.078	0.098	0.143

According to the histogram, pHMEAN appears to be bimodal and very slightly left-skewed. The average pH appears to be slightly basic, with a value of 8.002. The range of pH between the first and third quarters is

0.14. Additionally, on average, every value has about a 0.098 average distance from the mean.

```
## Warning: Removed 48 rows containing missing values (geom_point).
```



It is clear that the pH levels off of the coast vary throughout the year, with a higher pH in the colder months, and a lower pH value in the warmer months. Additionally, it looks as though the lowest value gets lower every year, but this will have to be examined in further detail. Values recorded during a named storm also appear to be on the smaller end, so correlations will be examined in greater detail.

```
##     pHMEAN       Date
## 1 7.448091 2018/09/20
```

The smallest pH value was recorded on 9/20/2018, which was about a week after Hurricane Florence hit North Carolina. Therefore, we will look into the influence of storms as well as how long these effects might last. However, is this because of the storm or just a coincidence because pH tends to lower in warmer seasons and storms happen more often in summer?

```
##
##  Welch Two Sample t-test
##
## data:  pHMEAN by NamedStorm
## t = 3.4617, df = 16.507, p-value = 0.003091
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.04156864 0.17207792
## sample estimates:
##  mean in group No mean in group Yes
##          8.004825          7.898002
```

```
## Warning: Removed 48 rows containing non-finite values (stat_boxplot).
```

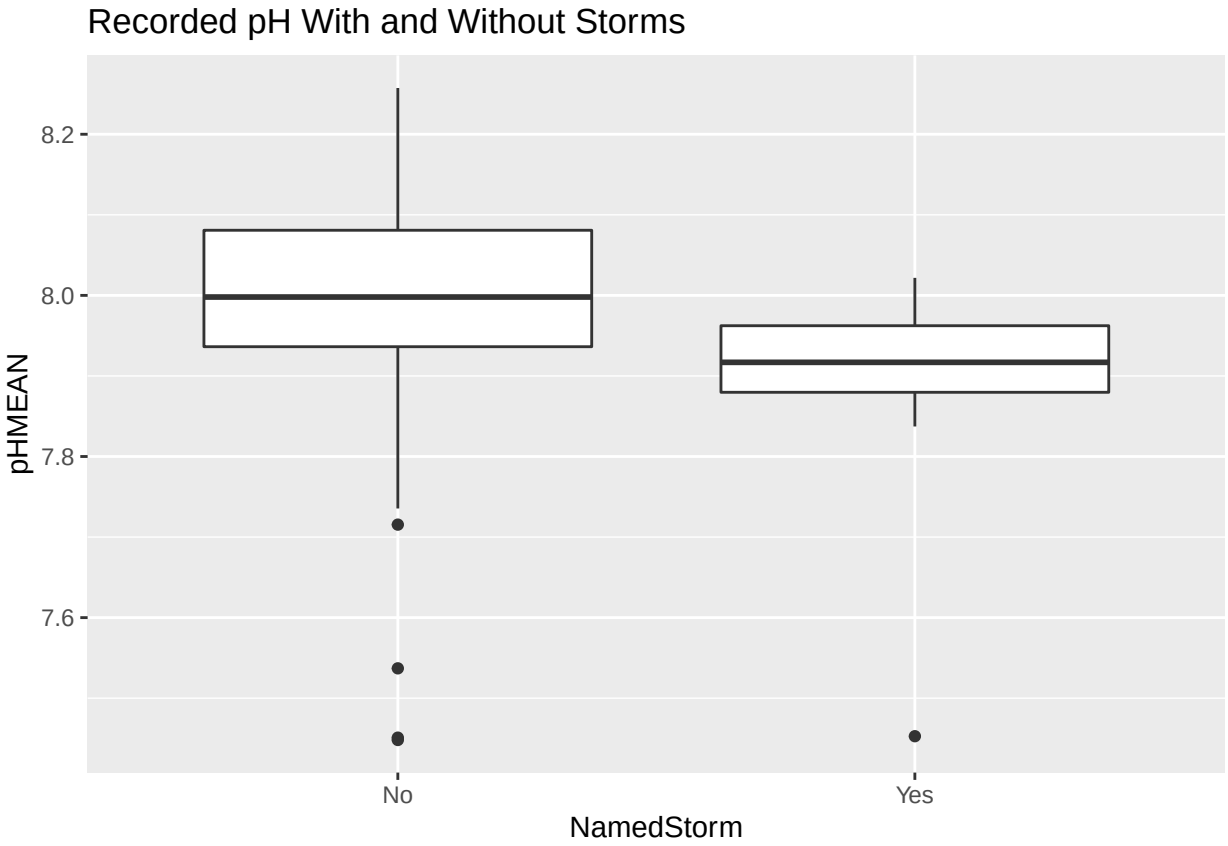
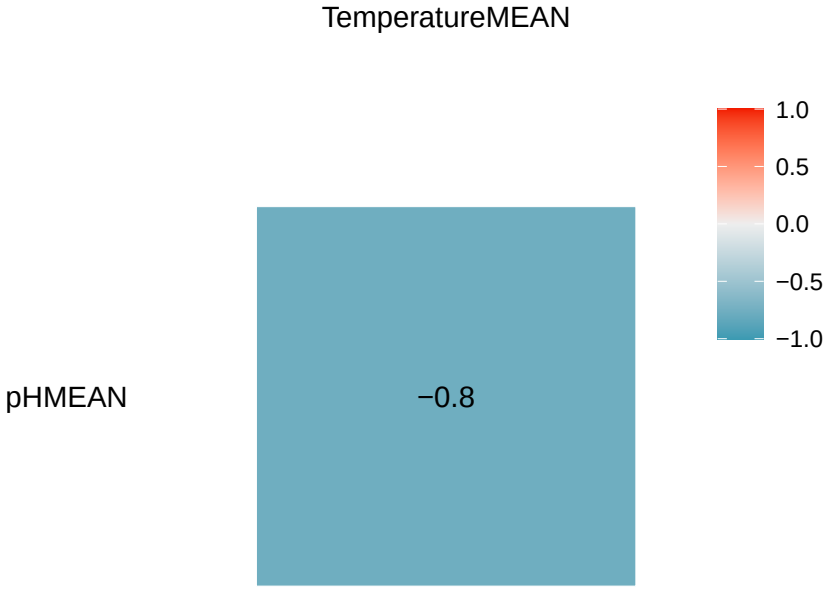


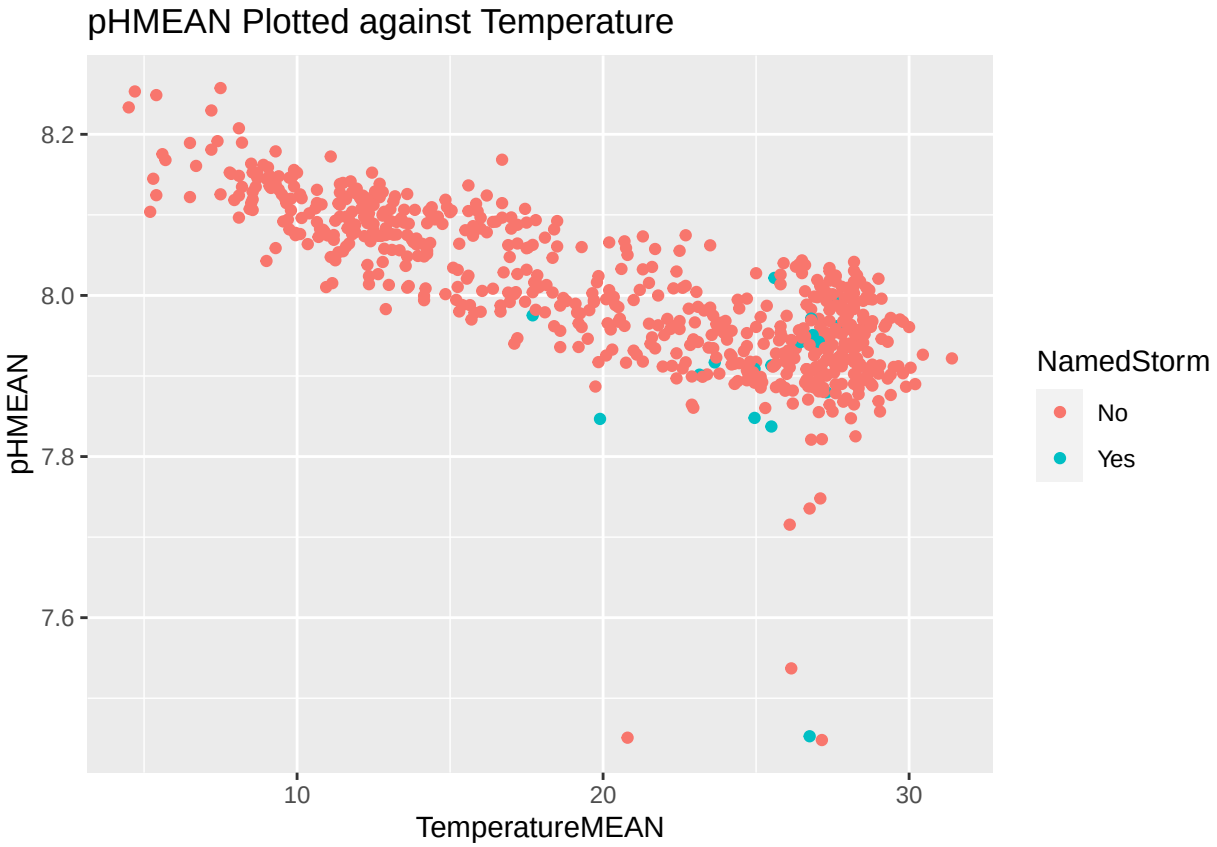
Table 2: Pairwise Wilcox Test

group1	group2	p.value
Yes	No	9e-05

According to the boxplot and pairwise wilcox test, with a p-value of 0.00009, there is in fact a difference in the mean pH value between values recorded during a named storm and those recorded during typical weather.



```
## Warning: Removed 48 rows containing missing values (geom_point).
```



There appears to be a somewhat negative linear relationship between temperature and pH, according to this plot. As the mean temperature increases, the mean pH decreases. This makes sense in the context of the data. Additionally, according to the correlation plot there is a relatively strong negative correlation between these two variables.

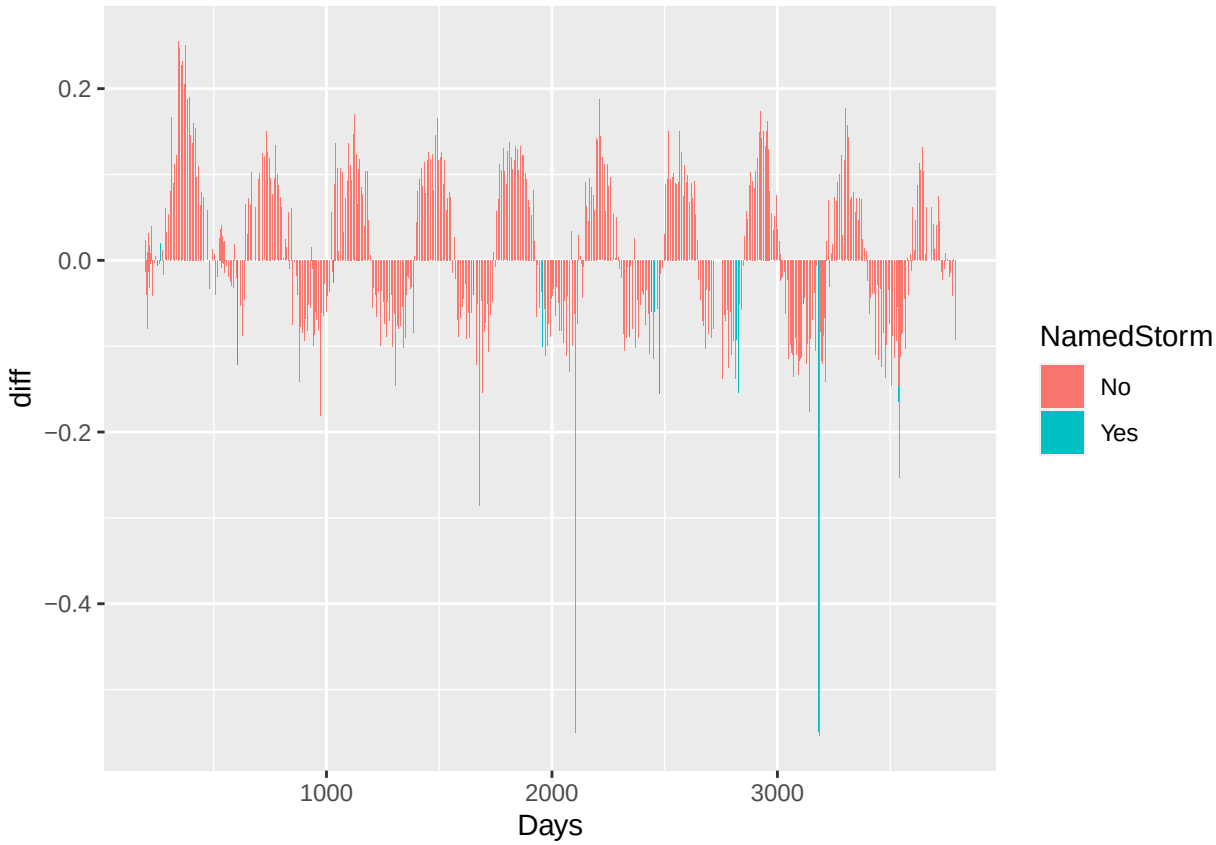
Table 3: Linear Model of pH and Temperature

term	estimate	std.error	statistic	p.value
(Intercept)	8.2140314	0.0076408	1075.01666	0
TemperatureMEAN	-0.0104838	0.0003566	-29.40315	0

For every one degree increase in temperature, the pH is estimated to decrease by 0.0105 on average. The p-value is close to zero, meaning that there is a relationship between pH and temperature.

```
## Warning: position_stack requires non-overlapping x intervals
```

```
## Warning: position_stack requires non-overlapping x intervals
```



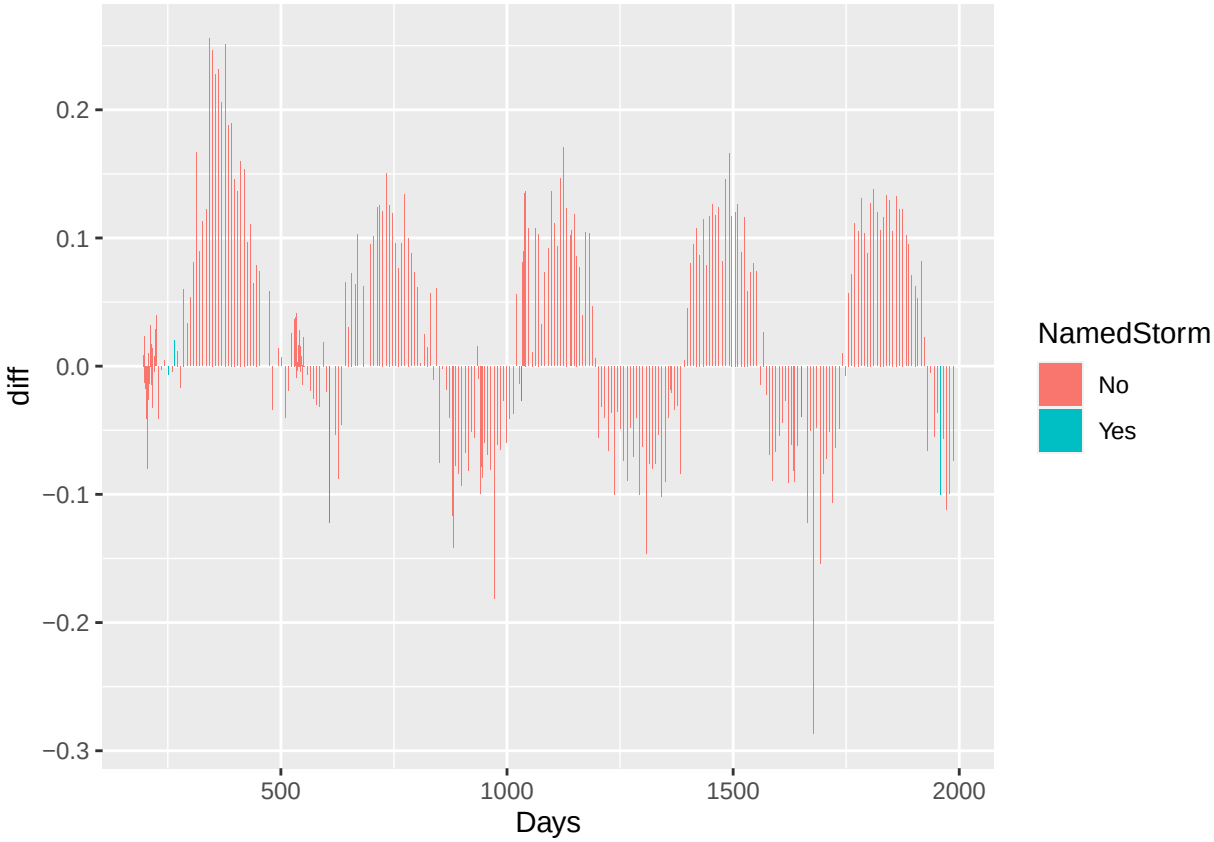

```
##          Date   pHMEAN      diff NamedStorm
## 1  2010/12/08 8.257426 0.2554793         No
## 2  2010/12/15 8.248694 0.2467472         No
## 3  2010/12/22 8.229639 0.2276926         No
## 4  2010/12/29 8.233435 0.2314883         No
## 5  2011/01/05 8.207517 0.2055706         No
## 6  2011/01/12 8.253215 0.2512684         No
## 7  2011/01/19 8.189668 0.1877210         No
## 8  2011/01/26 8.191532 0.1895856         No
## 9  2011/01/27 8.180976 0.1790294         No
## 10 2016/01/20 8.189257 0.1873104         No
## 11 2019/01/16 8.178843 0.1768957         No
```

```
##          Date   pHMEAN       diff NamedStorm
## 1  2012/08/29 7.820919 -0.1810281         No
## 2  2014/08/06 7.715420 -0.2865267         No
## 3  2014/08/20 7.847530 -0.1544166         No
## 4  2015/10/07 7.450993 -0.5509536         No
## 5  2016/10/12 7.846731 -0.1552157        Yes
## 6  2017/09/26 7.847970 -0.1539768        Yes
## 7  2018/08/08 7.825095 -0.1768513         No
## 8  2018/09/19 7.452792 -0.5491547        Yes
## 9  2018/09/20 7.448091 -0.5538559         No
## 10 2018/09/21 7.537025 -0.4649218         No
## 11 2018/09/22 7.735478 -0.2664689         No
## 12 2019/09/07 7.837277 -0.1646693        Yes
## 13 2019/09/10 7.748015 -0.2539317         No
```

```
## 14 2019/09/11 7.821565 -0.1803822         No
```

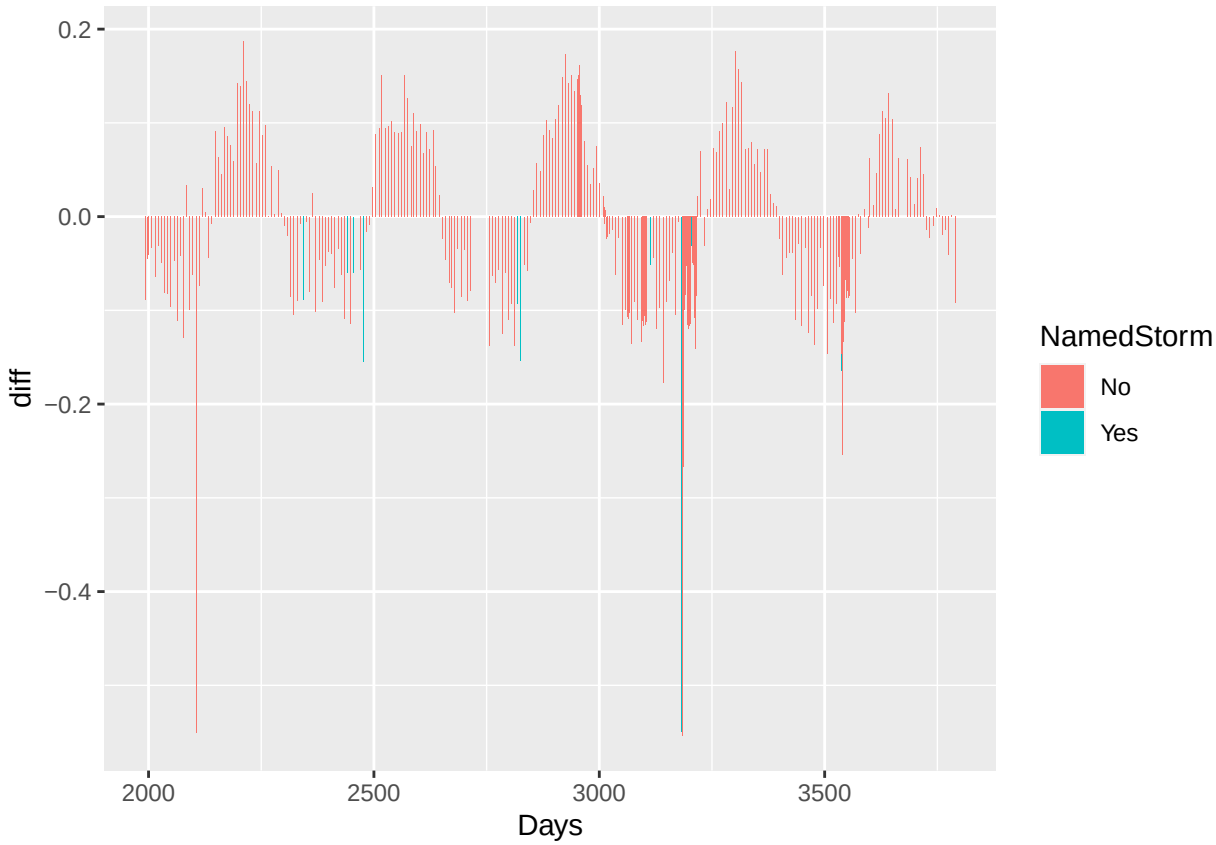
```
## Warning: position_stack requires non-overlapping x intervals
```

```
## Warning: position_stack requires non-overlapping x intervals
```



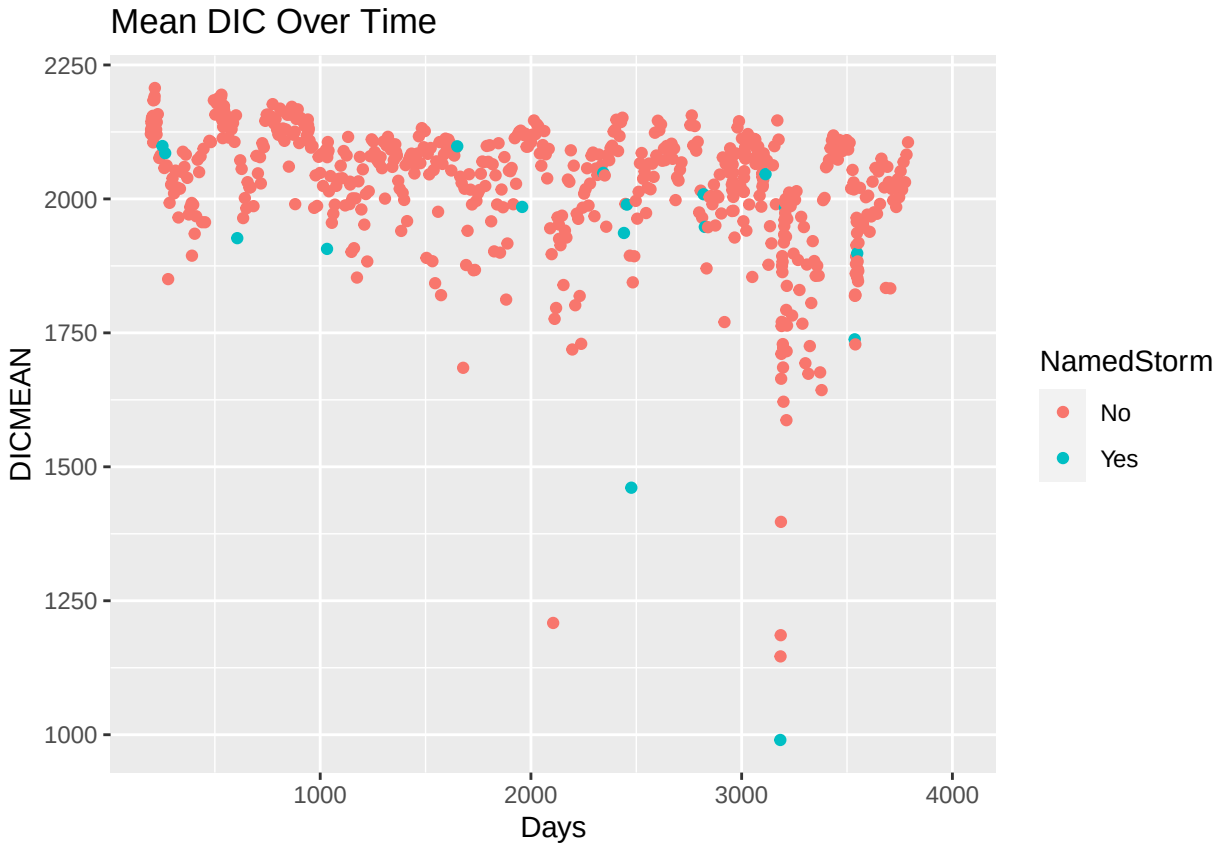

```
## Warning: position_stack requires non-overlapping x intervals
```

```
## Warning: position_stack requires non-overlapping x intervals
```



Dissolved Inorganic Carban

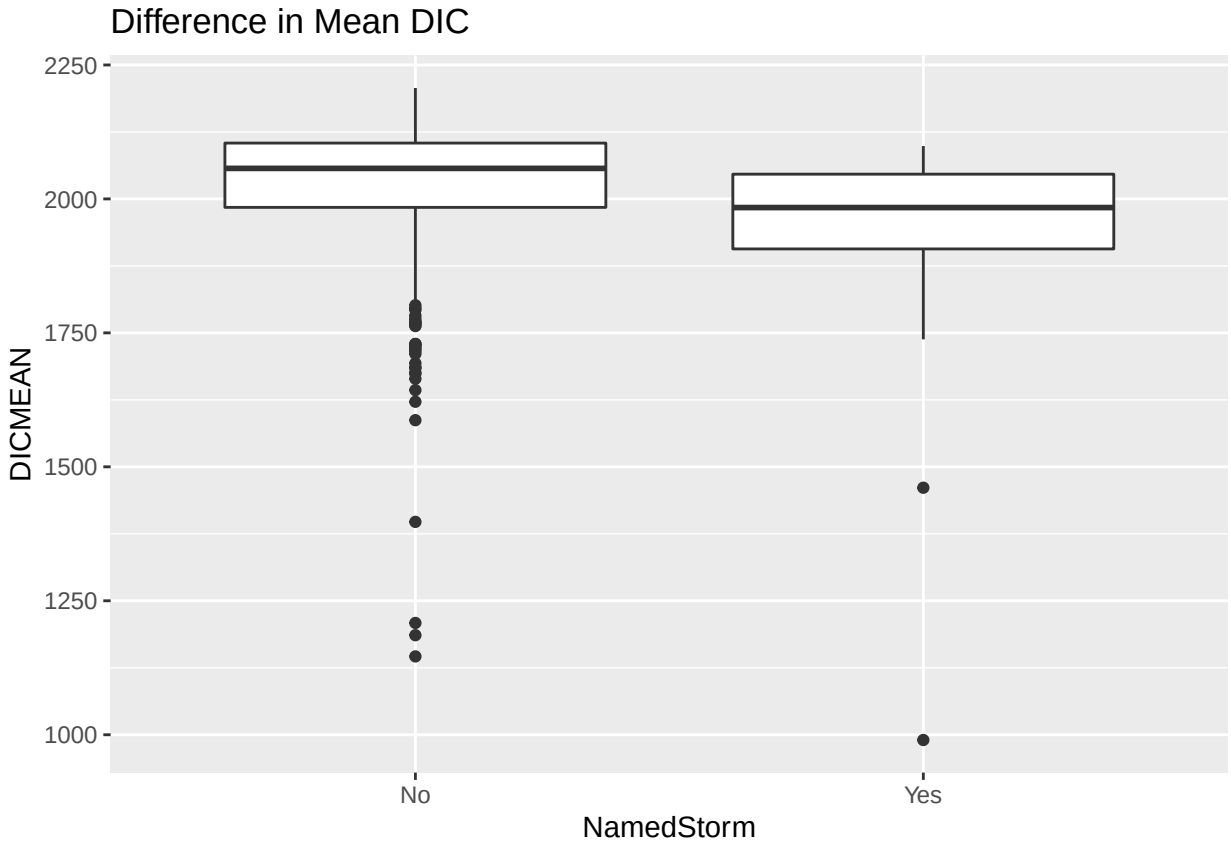
```
## Warning: Removed 46 rows containing missing values (geom_point).
```



Dissolved inorganic carbon appears to have a less prominent seasonal trend

```
##   DICMEAN       Date
## 1   990.2 2018/09/19
```

The smallest dissolved inorganic carbon recording occurred during Hurricane Florence



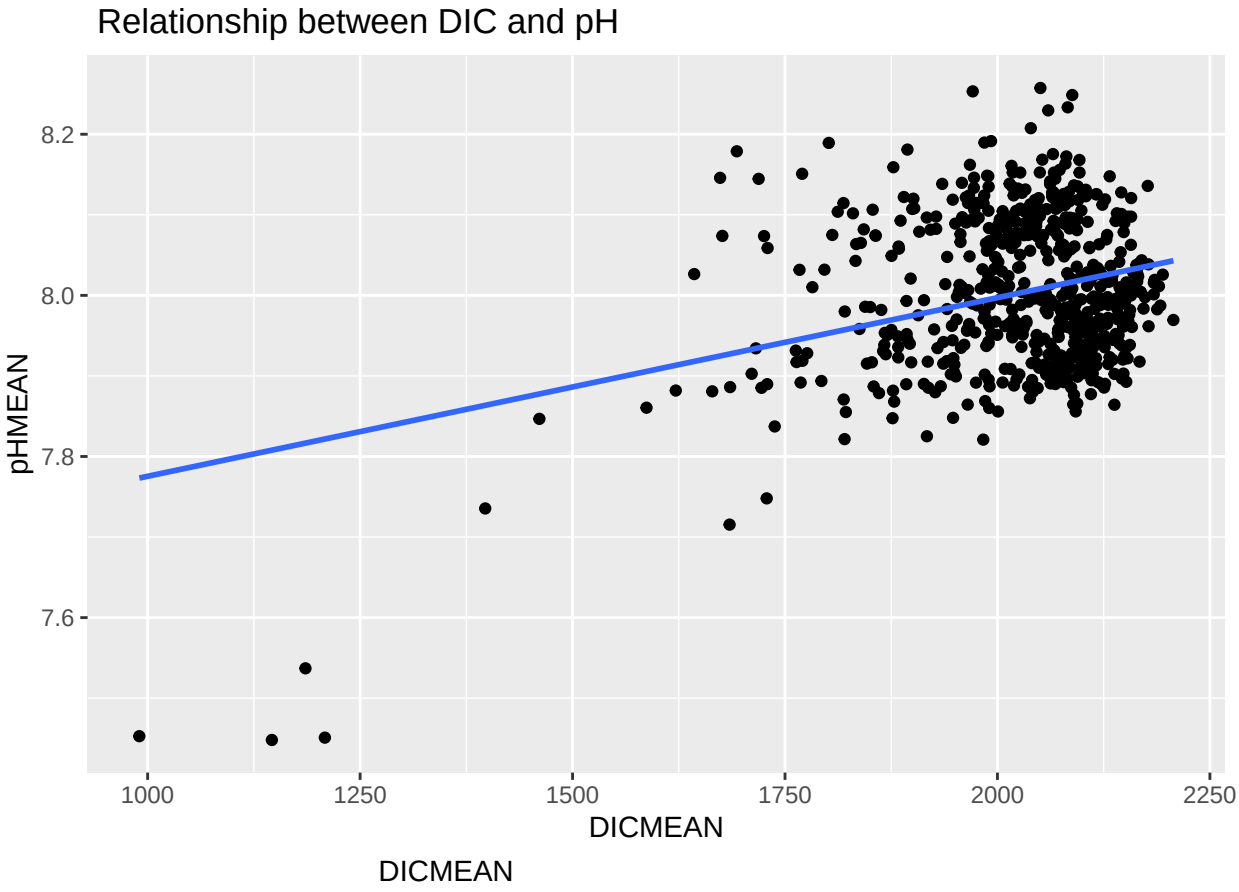

```
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic p.value parameter conf.low conf.high
##      <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>     <dbl>
## 1     134.     2025.     1891.      1.98  0.0646      16.2    -9.16      278.
## # ... with 2 more variables: method <chr>, alternative <chr>
```

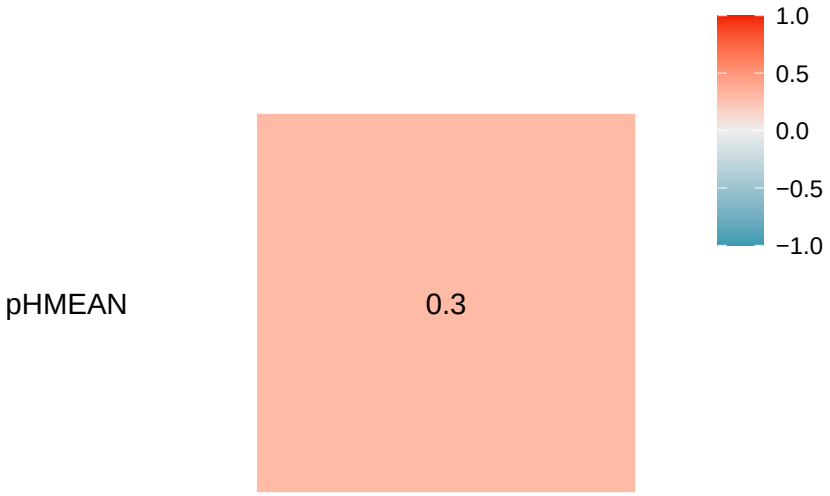
While the boxplots show a difference in means between storms, the p-value is not significant, indicating that there is not a difference in the mean DIC when there is a storm.

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 48 rows containing non-finite values (stat_smooth).
```

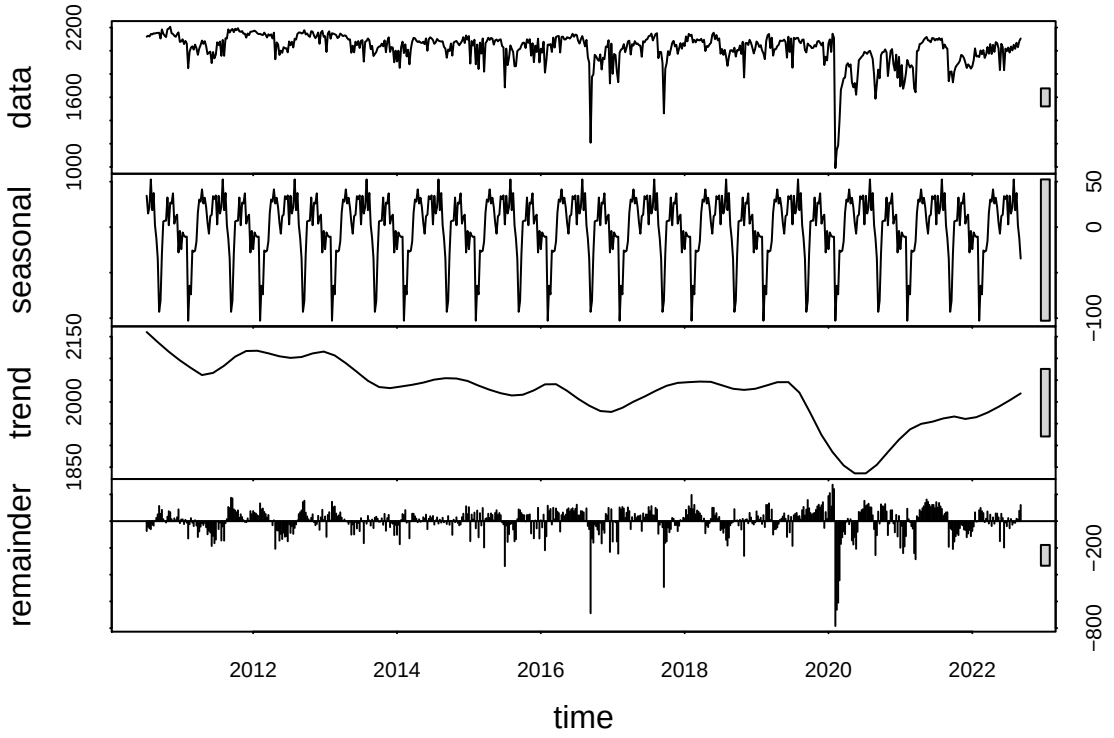
```
## Warning: Removed 48 rows containing missing values (geom_point).
```

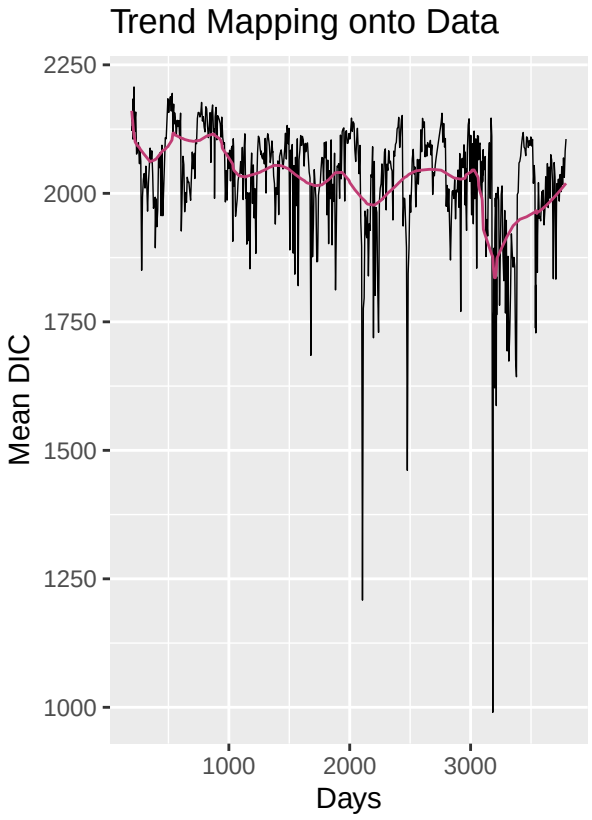




There appears to be a slight positive correlation between the two variables

DIC Time Series





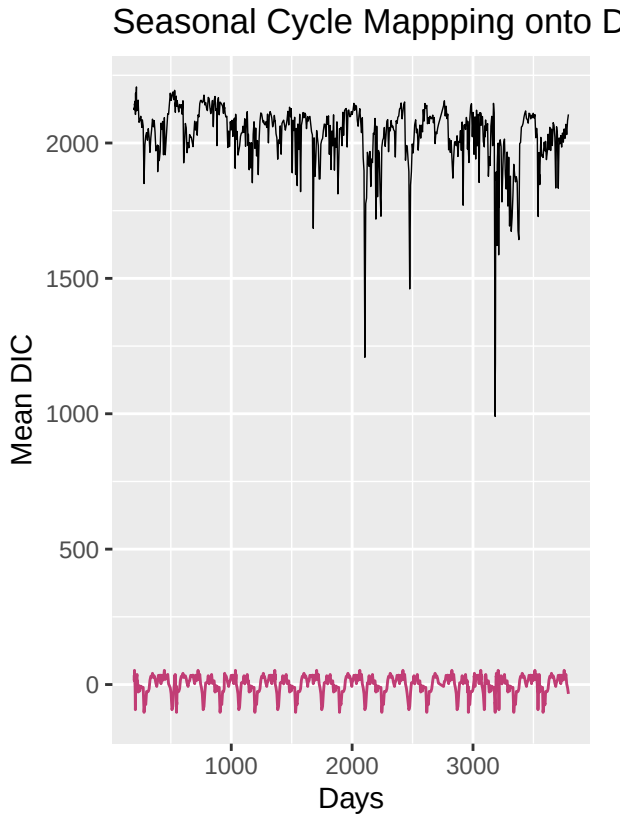
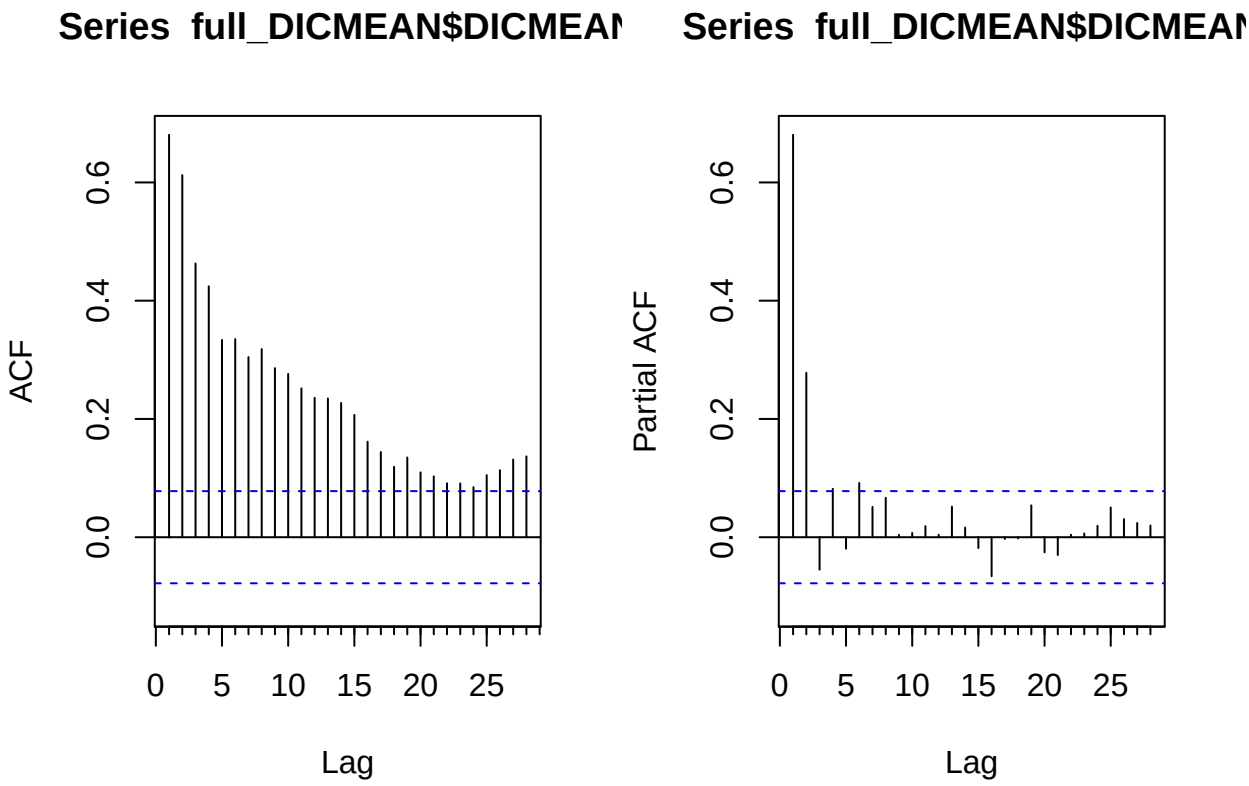
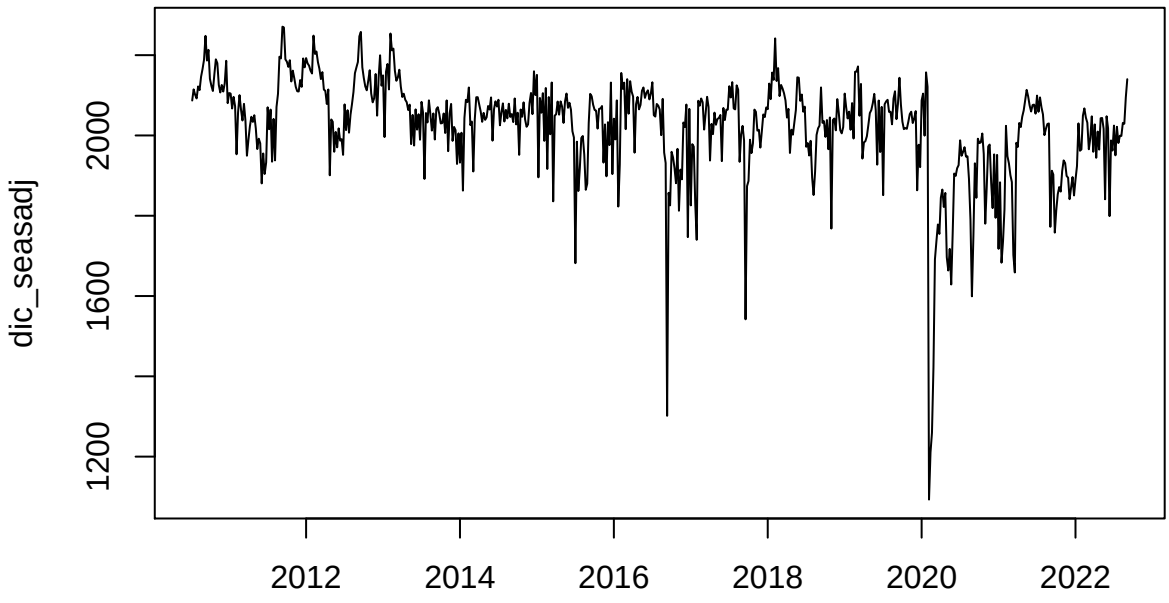


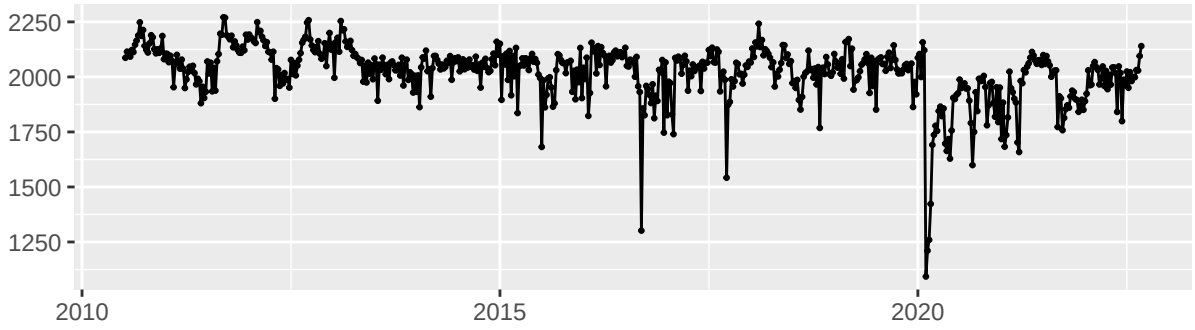
Table 4: Seasonal Mann Kendall test for pH

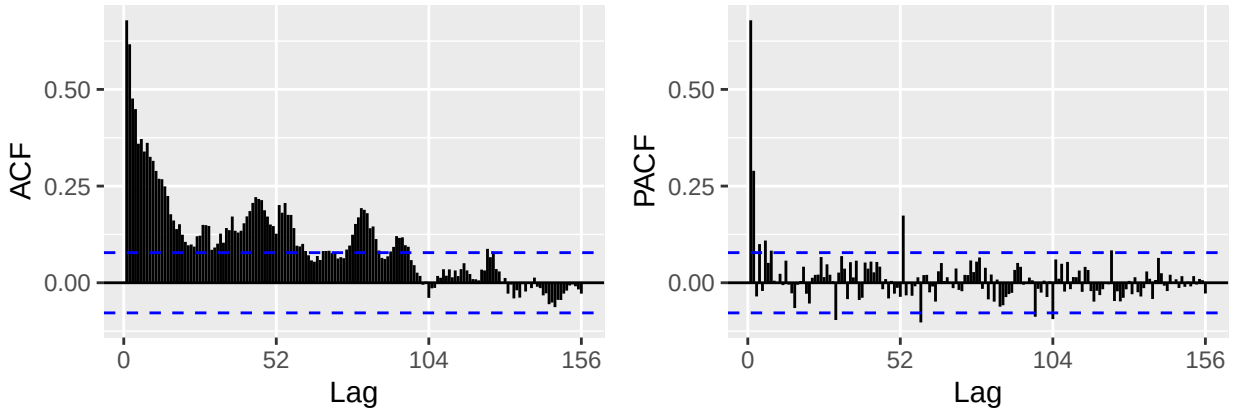
p.value	kendall_score	statistic
0	-1174	-0.3316384

```
##
##  Seasonal Mann-Kendall trend test (Hirsch-Slack test)
##
## data:  dic.ts
## z = -10.909, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##        S     varS
## -1174.00 11562.67
```





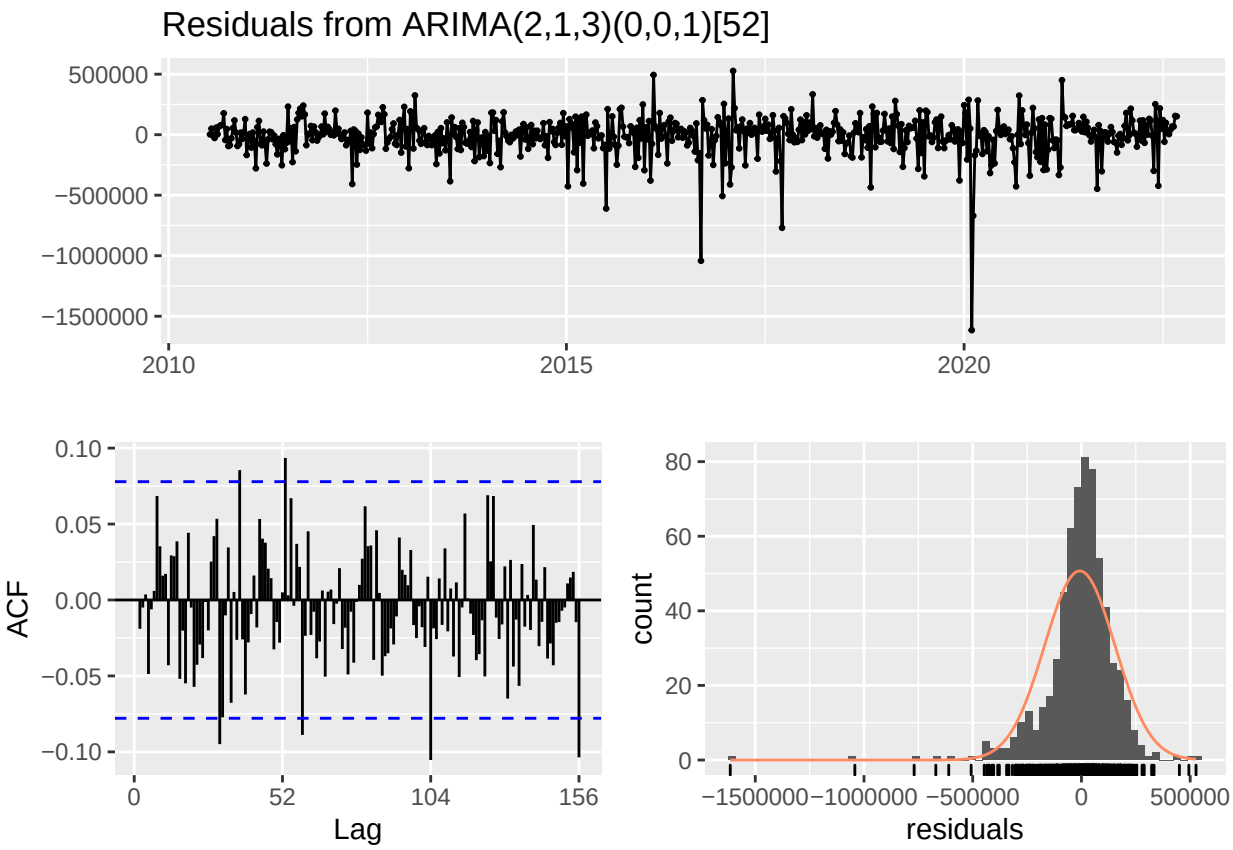





```
## Series: dic_seasadj
## ARIMA(2,1,3)(0,0,1)[52]
##
## Coefficients:
##          ar1     ar2      ma1      ma2     ma3     sma1
##      -0.0612  0.6456  -0.4368  -0.6611  0.1249  -0.1425
## s.e.  0.0773  0.0699   0.0899   0.1010  0.0666   0.0467
```

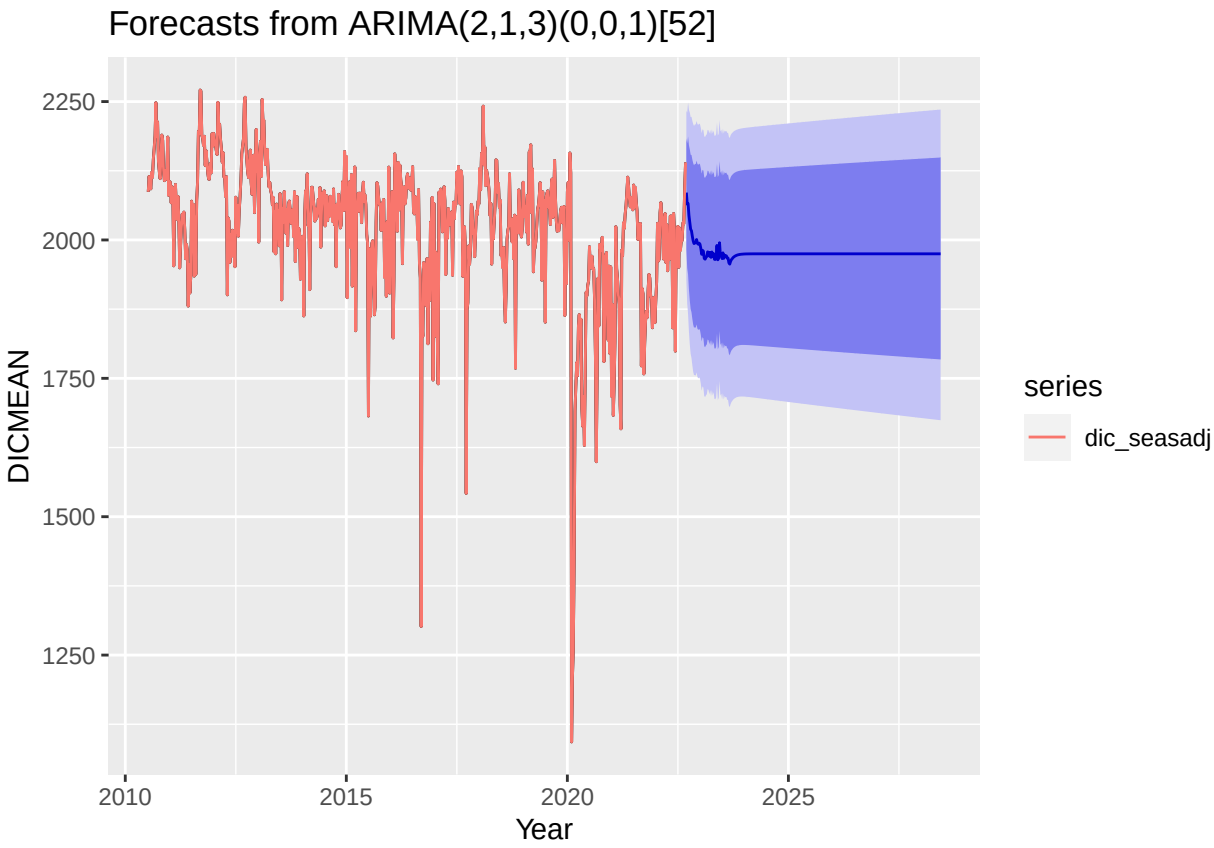
```
##
## sigma^2 estimated as 7796:  log likelihood=-3726.95
## AIC=7467.89   AICc=7468.07   BIC=7499.03
```

```
## Series: dic_seasadj
## ARIMA(2,1,3)(0,0,1)[52]
## Box Cox transformation: lambda= 1.999924
##
## Coefficients:
##          ar1     ar2      ma1      ma2     ma3     sma1
##       -0.0365  0.6653  -0.4766  -0.6420  0.1426  -0.1291
## s.e.   0.0716  0.0653   0.0852   0.0961  0.0651   0.0465
##
## sigma^2 estimated as 2.645e+10:  log likelihood=-8478.62
## AIC=16971.24   AICc=16971.42   BIC=17002.38
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,1,3)(0,0,1)[52]
## Q* = 7.195, df = 6, p-value = 0.3032
##
## Model df: 6.   Total lags used: 12
```

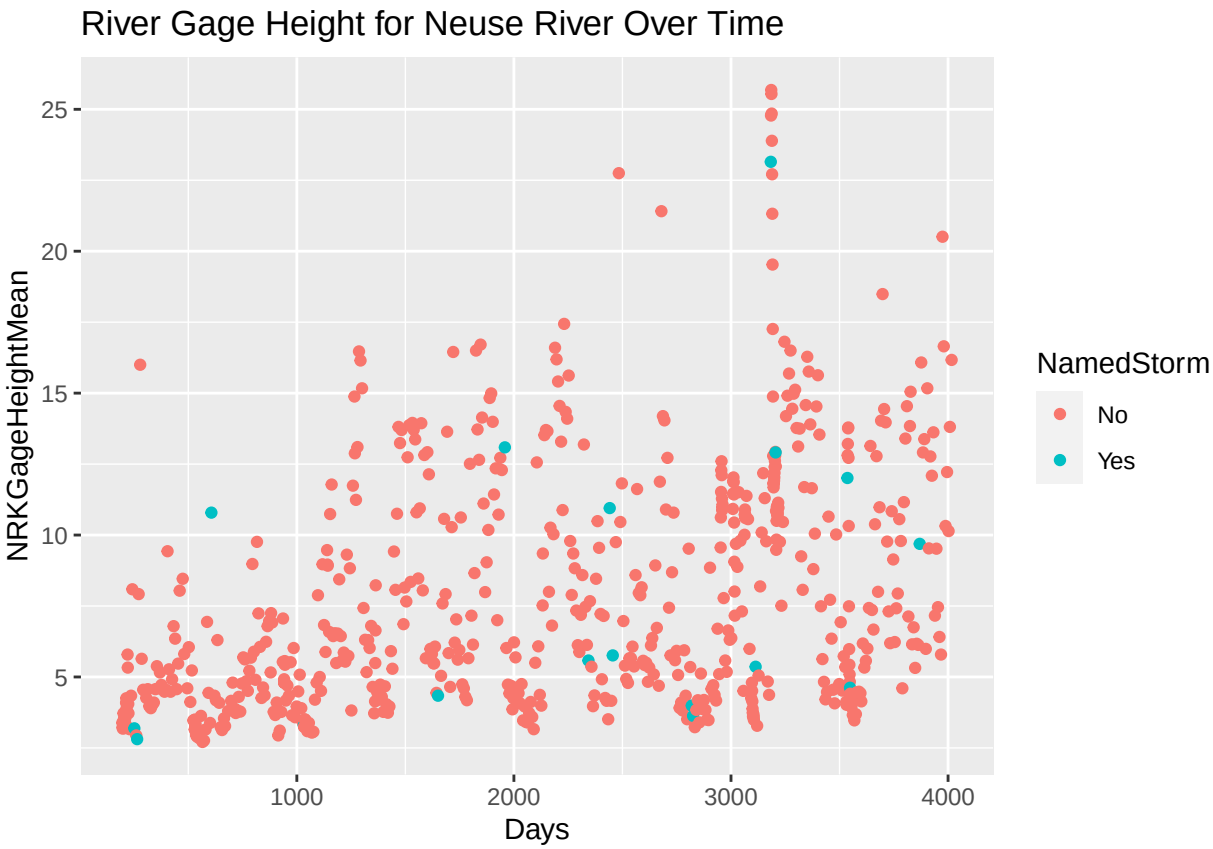
```
## [1] 16971.42
```



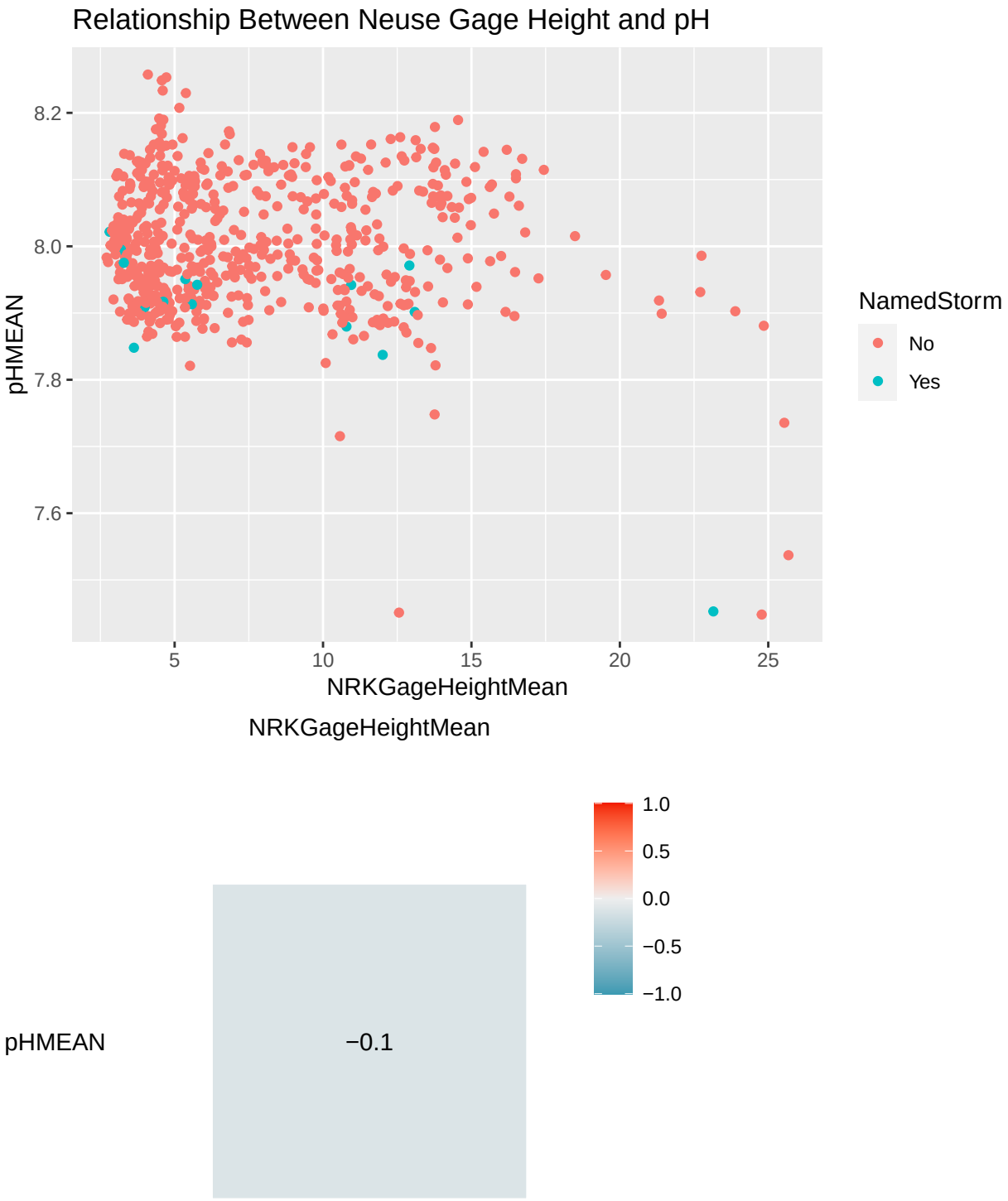
This model appears to forecast a slight decrease in dissolved inorganic carbon

River Gage

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

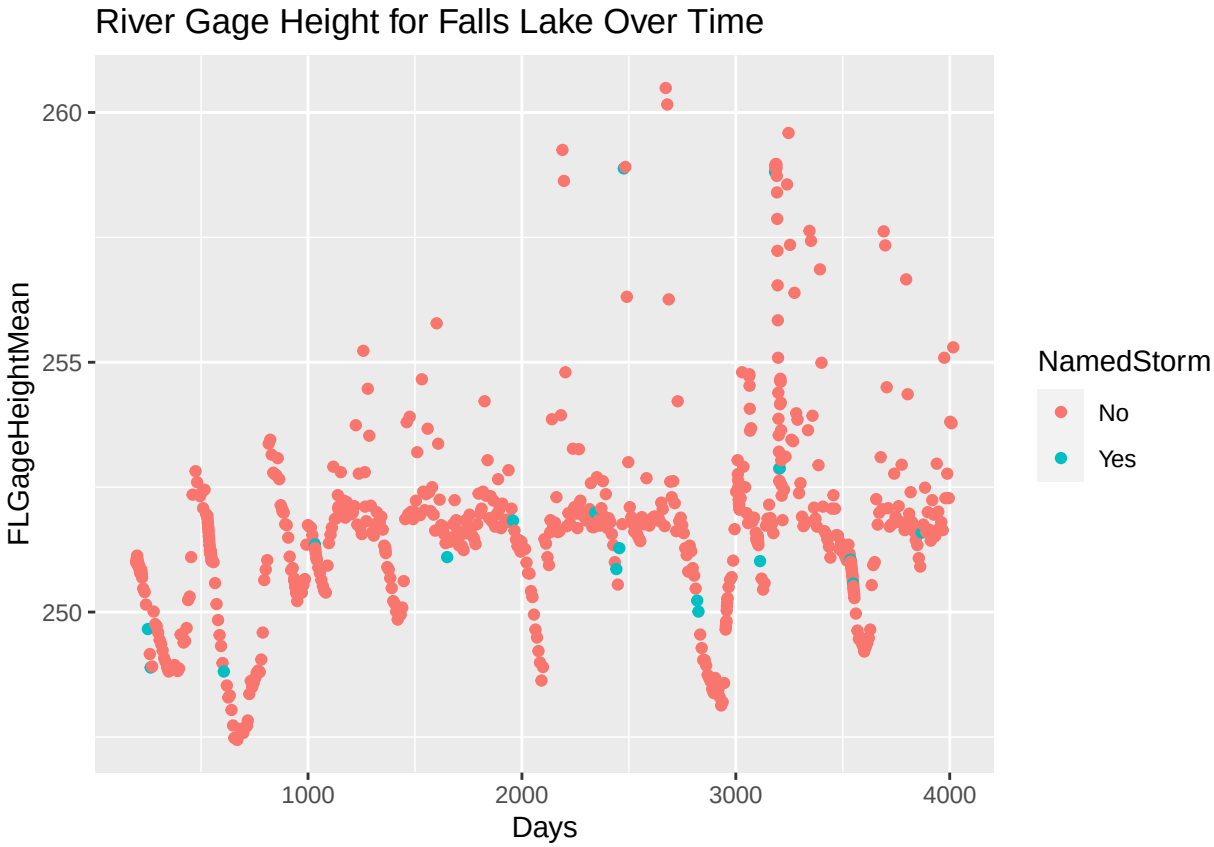



```
## Warning: Removed 52 rows containing missing values (geom_point).
```

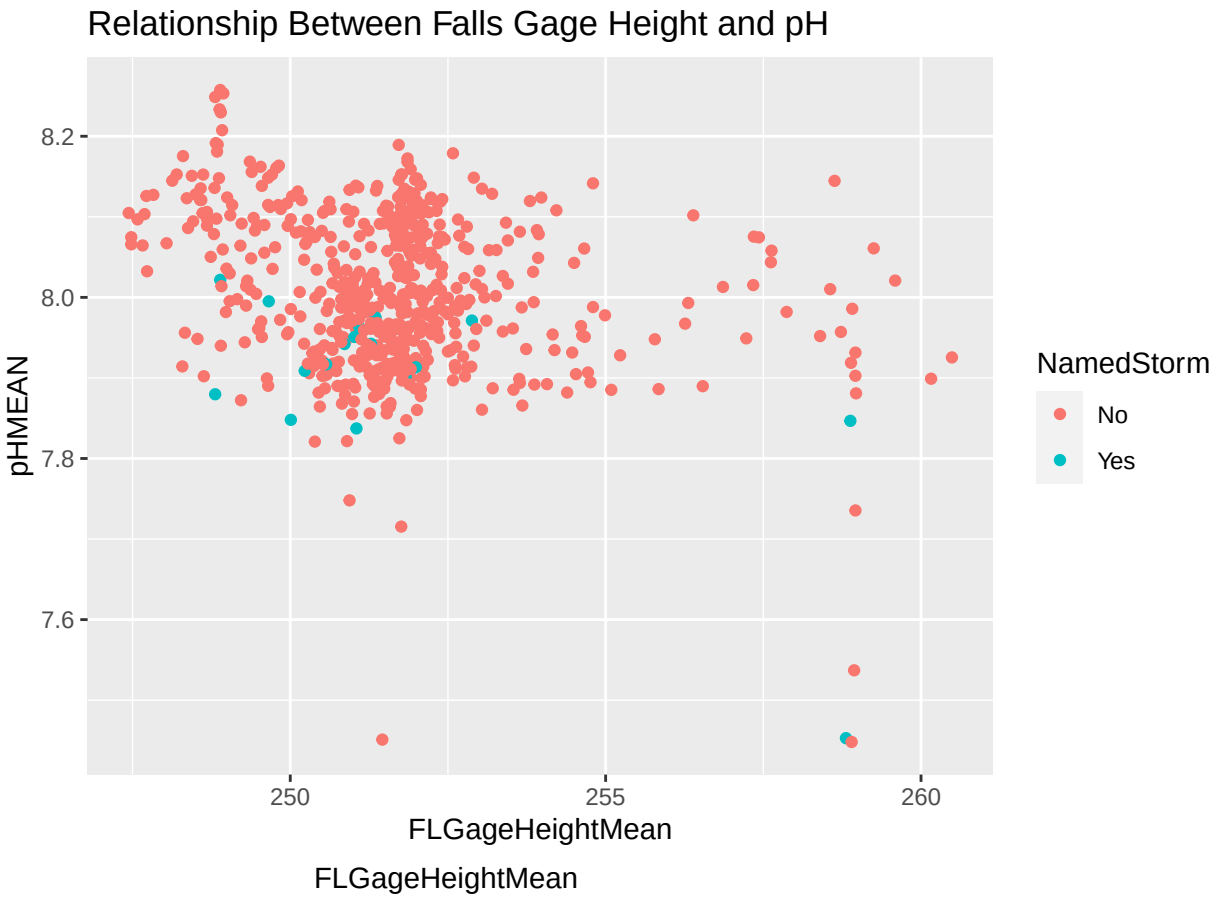


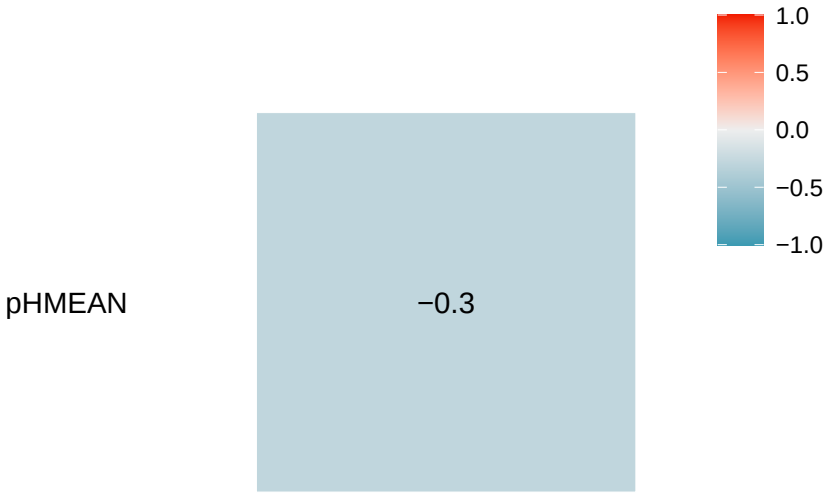
There is only a very weak negative correlation between the Neuse River gage height at Kinston, NC and the pH recorded off of Pivers Island.

```
## Warning: Removed 3 rows containing missing values (geom_point).
```



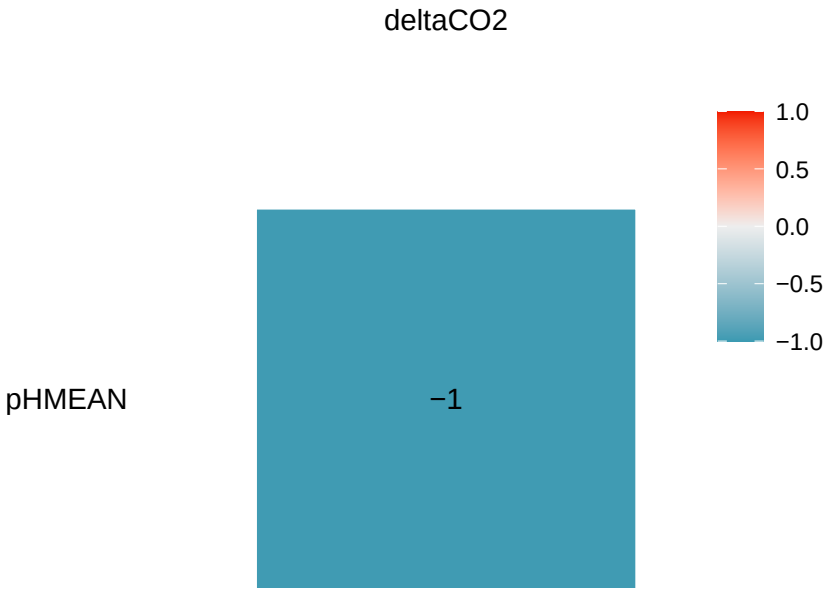

```
## Warning: Removed 51 rows containing missing values (geom_point).
```



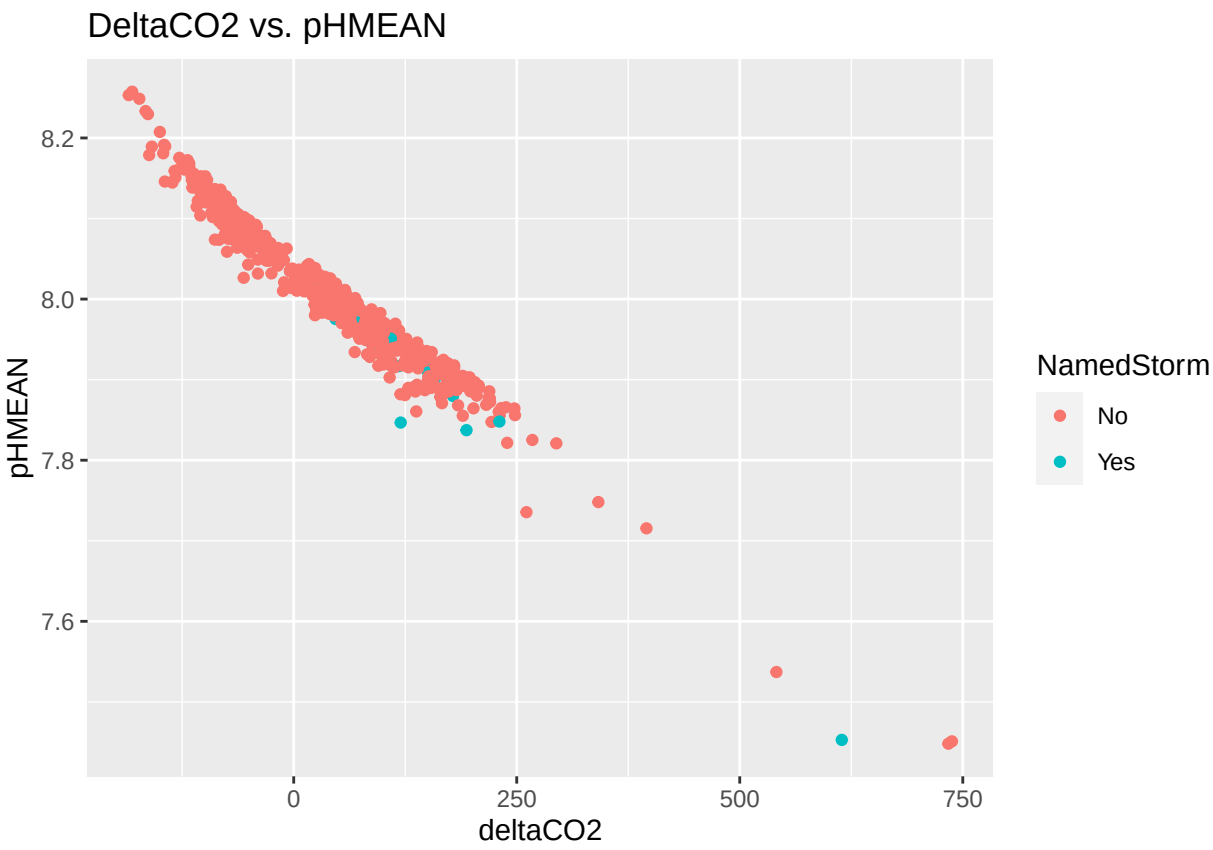


While slightly stronger between the correlation between the Neuse River and pH, there is only a weak negative correlation between the Falls Lake gage height and the pH recorded off of Pivers Island.

Additional Variables




```
## Warning: Removed 48 rows containing missing values (geom_point).
```



There appears to be a negative linear relationship between water-atmospheric carbon dioxide and pH. As the water-atmospheric CO2 increases, the pH decreases. Additionally, there appear to be more named storms recorded at higher values of water-atmospheric CO2 increases. According to the correlation plot, there is a strong negative correlation between these two variables.

Table 5: Linear Model of pH and DeltaCO2

term	estimate	std.error	statistic	p.value
(Intercept)	8.0413128	0.0007747	10379.7161	0
deltaCO2	-0.0008537	0.0000064	-133.6373	0

For every 100 µatm increase in water-atmospheric carbon dioxide, the pH is estimated to decrease by 0.0854 on average. The p-value is close to zero, meaning there is a relationship between the two variables.

```
##   deltaCO2       Date
## 1 737.8547 2015/10/07
```

The largest value for deltaCO2 occurs on a day when the second smallest pH is recorded.

```
## Warning: Removed 48 rows containing non-finite values (stat_boxplot).
```

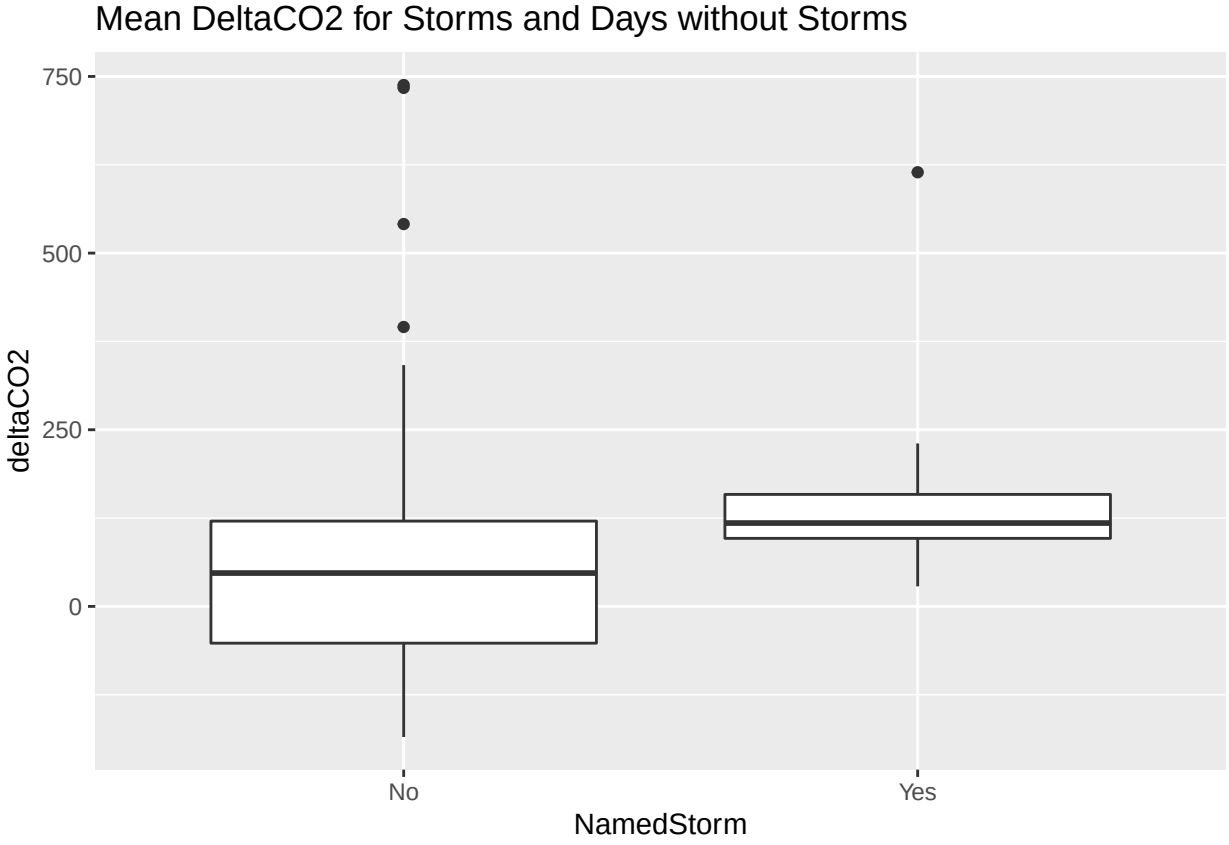
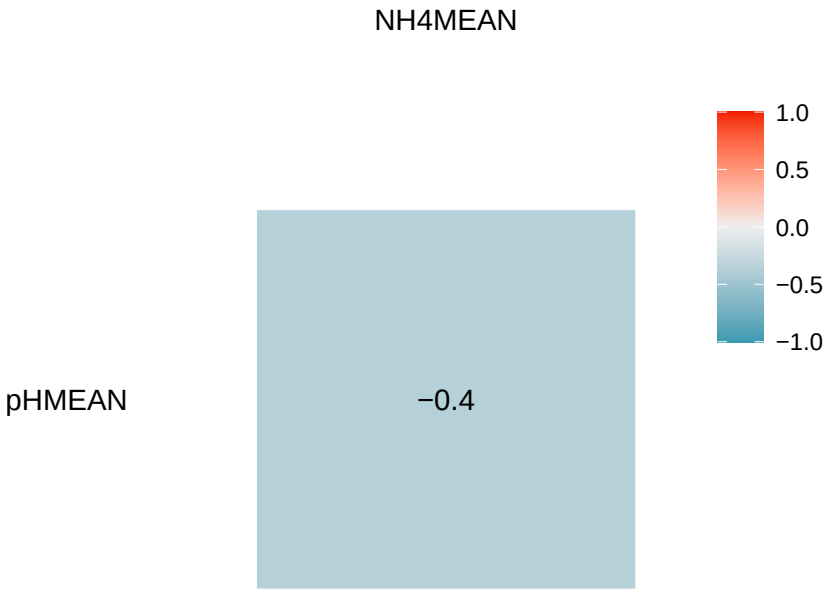


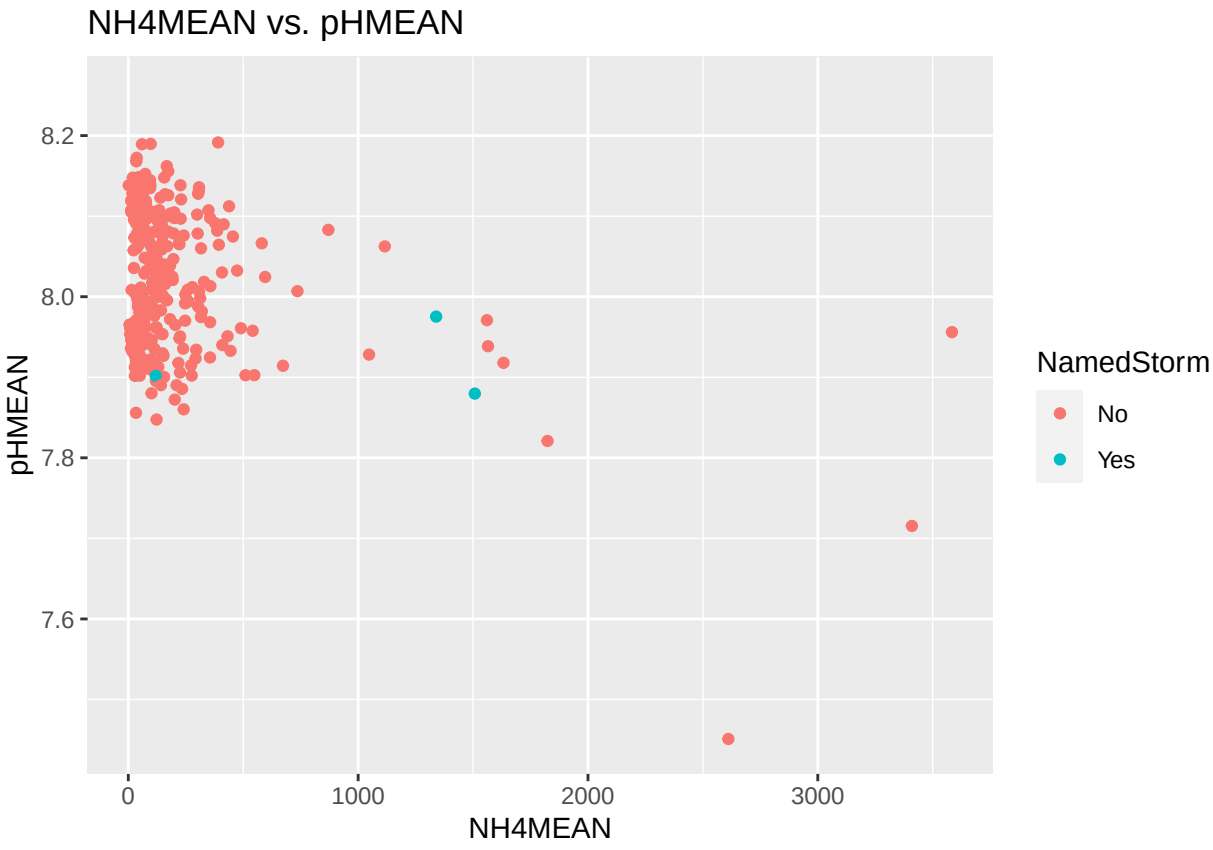
Table 6: Pairwise Wilcox Test

group1	group2	p.value
Yes	No	0.0005083

After examining the boxplot and administering a pairwise wilcox test, with a p-value of 0.0005, it is apparent that there is a difference in the mean water-atmospheric carbon dioxide for days with named stroms and days with typical weather.




```
## Warning: Removed 382 rows containing missing values (geom_point).
```



There does not appear to be a strong relationship between NH4 and pH. The correlation plot also displays a weak correlation between the two variables.

Table 7: Linear Model of pH and NH4

term	estimate	std.error	statistic	p.value
(Intercept)	8.0343996	0.0055237	1454.545014	0
NH4MEAN	-0.0000796	0.0000123	-6.454661	0

For every one nM increase in NH4, the pH decreases by -0.00008. While the small p-value suggests a relationship between the two variables, this may be a result of the influence from the outliers.

```
## Warning: Removed 381 rows containing non-finite values (stat_boxplot).
```

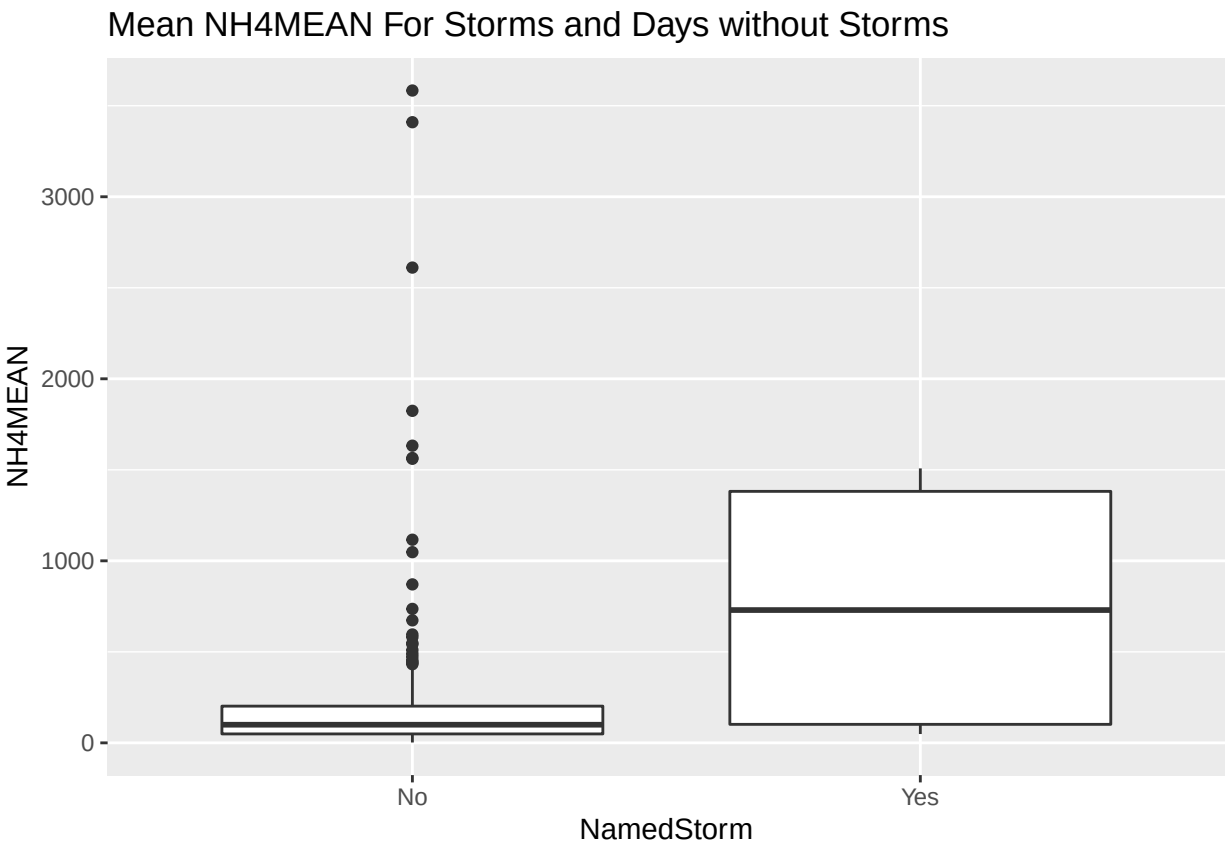


Table 8: Pairwise Wilcox Test

group1	group2	p.value
Yes	No	0.1976984

According to the boxplot there looks like there is a difference in the mean NH4 for days with a named storm, however, the pairwise wilcox test fails to reject the hypothesis that there is no difference in means.

Multiple Linear Regression

In this linear model, I will include the temperature, named storm indicator, water-atmospheric carbon dioxide, and NH4 because each variable appears to have a relationship with pH. I will also include an interaction term between the storm indicator and each of the numeric variables because there is a difference in means for values recorded during a storm.

Table 9: Multiple Linear Regression Model for pH

term	estimate	std.error	statistic	p.value
(Intercept)	8.0414305	0.0034884	2305.1744161	0.0000000
NamedStormYes	-0.0482955	0.0466487	-1.0353009	0.3013939
TemperatureMEAN	0.0002437	0.0001839	1.3253833	0.1860908
deltaCO2	-0.0008351	0.0000128	-65.3884154	0.0000000
NH4MEAN	-0.0000023	0.0000021	-1.0490303	0.2950407
NamedStormYes:TemperatureMEAN	0.0017248	0.0022417	0.7694073	0.4422800
NamedStormYes:deltaCO2	-0.0000229	0.0001669	-0.1370037	0.8911233
NamedStormYes:NH4MEAN	-0.0000070	0.0000104	-0.6755170	0.4998875

Several variables have insignificant p-values in this model, so I will perform backwards selection in order to ensure that I get the most accurate results. This test will eliminate any unnecessary predictors, which is important as any model can get a high r-squared value with a lot of variables, even if many are not considered significant.

Table 10: Reduced Model through backward Selection

term	estimate	std.error	statistic	p.value
(Intercept)	8.0397817	0.0032301	2488.996526	0.0000000
NamedStormYes	-0.0160253	0.0064739	-2.475359	0.0138757
TemperatureMEAN	0.0003153	0.0001755	1.796980	0.0733685
deltaCO2	-0.0008414	0.0000115	-72.866298	0.0000000

The reduced model only includes the named storm indicator, temperature, and water-atmospheric carbon dioxide. While temperature does not have a significant p-value, I will leave it in the model for the moment because of the results from exploratory data analysis. However, this may need to be investigated further.

Table 11: Model Selection Criteria for Full Model

Adjusted R Squared	AIC	BIC
0.9797	-1735.585	-1702.341

Table 12: Model Selection Criteria for Reduced Model

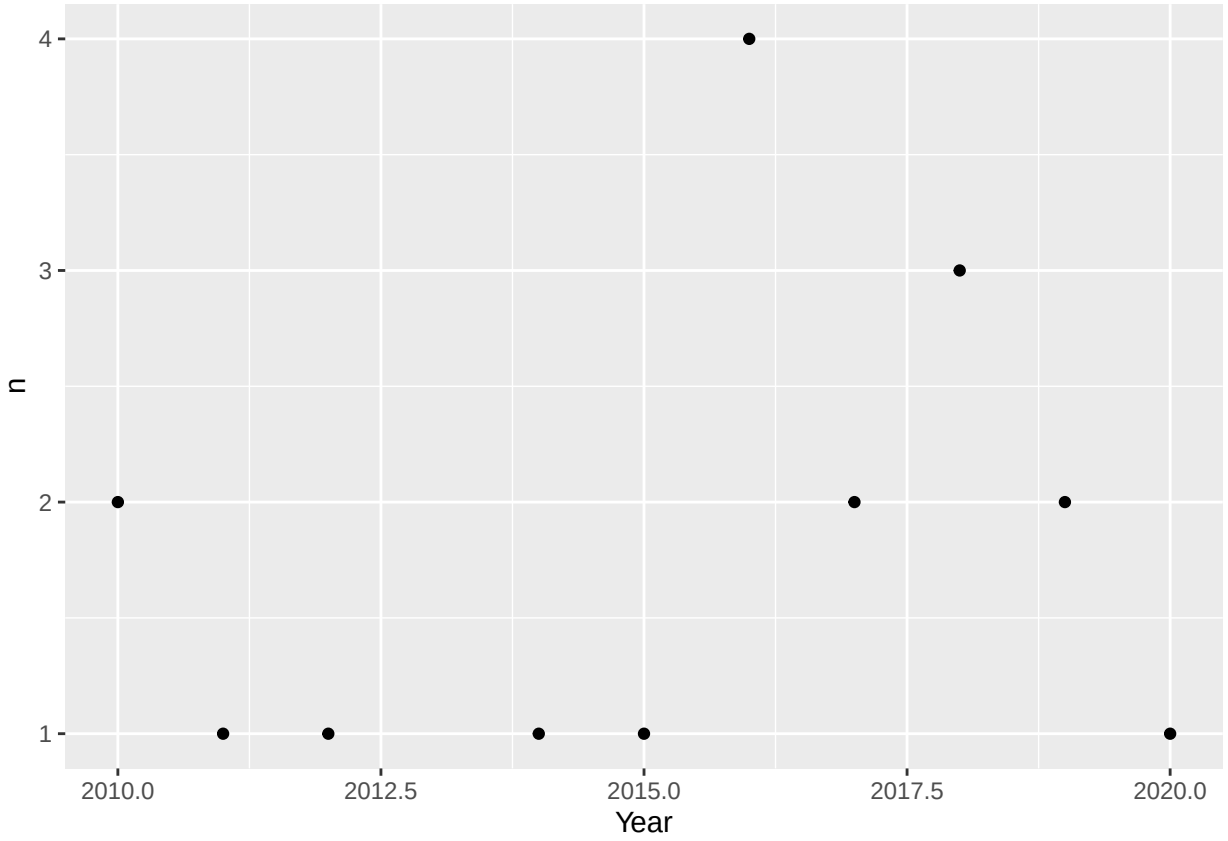
Adjusted R Squared	AIC	BIC
0.9798	-1739.935	-1721.466

The adjusted r-squared, AIC, and BIC values are similar for the full and reduced models. However, the reduced model has a slightly higher adjusted r-squared and lower AIC and BIC values, so I will choose this as the most accurate.

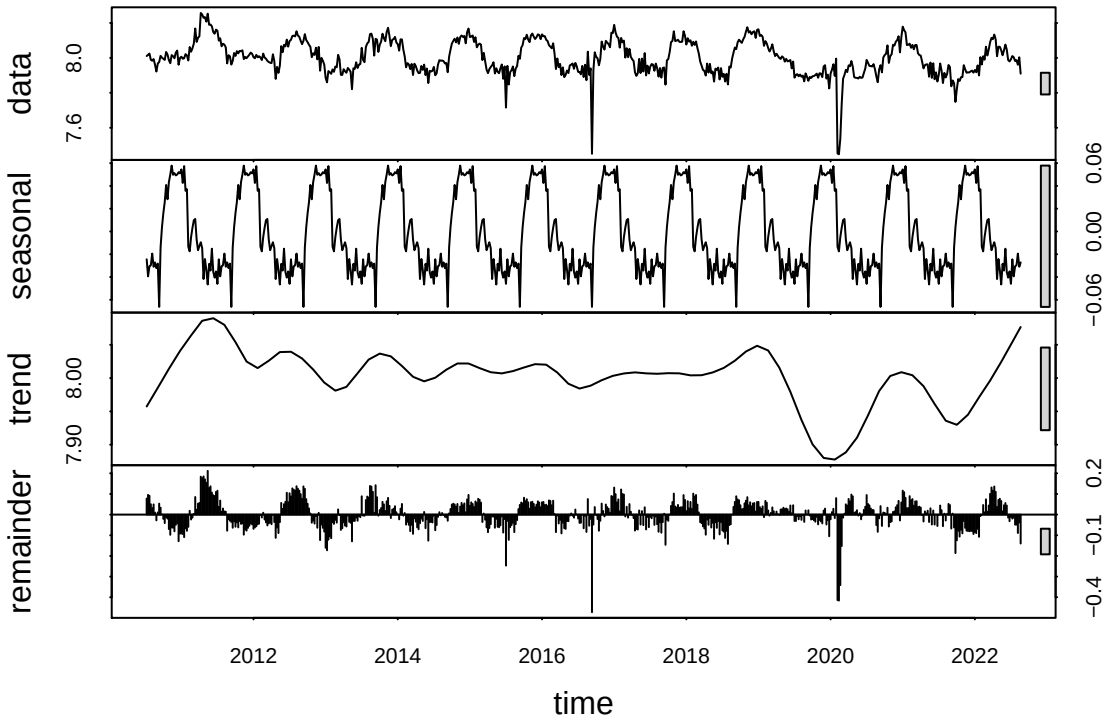
Checking Storm Frequency

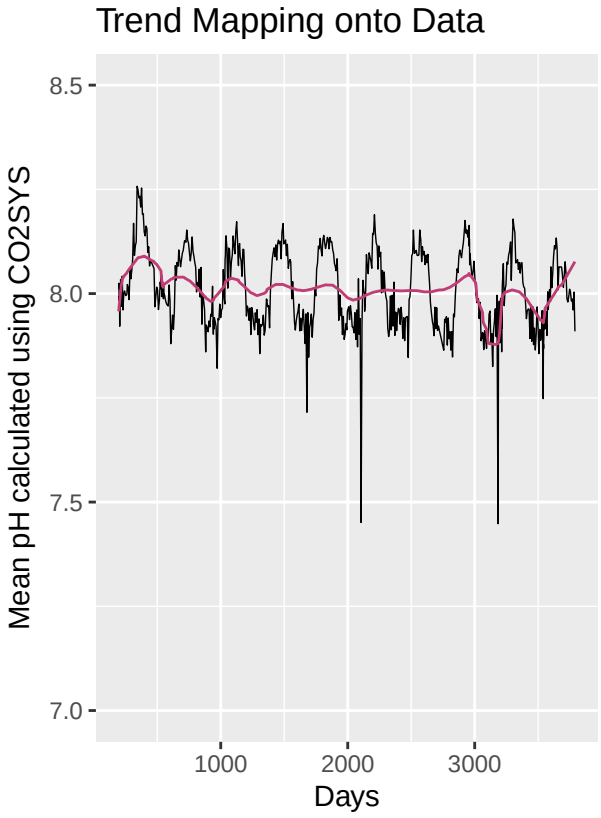
```
## # A tibble: 21 x 3
## # Groups:   Year [11]
##     Year NamedStorm     n
```

```
##    <dbl> <fct>      <int>
##  1  2010 No            41
##  2  2010 Yes            2
##  3  2011 No            65
##  4  2011 Yes            1
##  5  2012 No            68
##  6  2012 Yes            1
##  7  2013 No            58
##  8  2014 No            52
##  9  2014 Yes            1
## 10  2015 No            53
## # ... with 11 more rows
```



pH Time Series





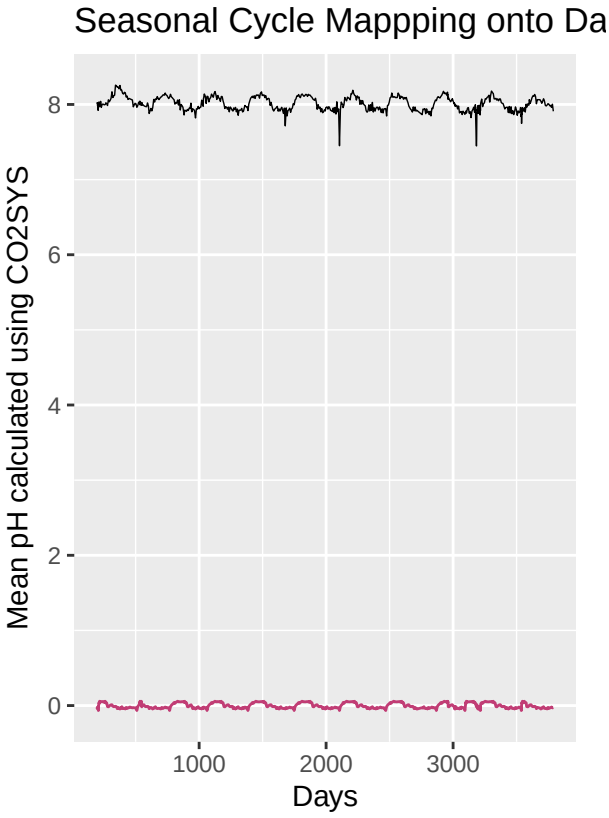
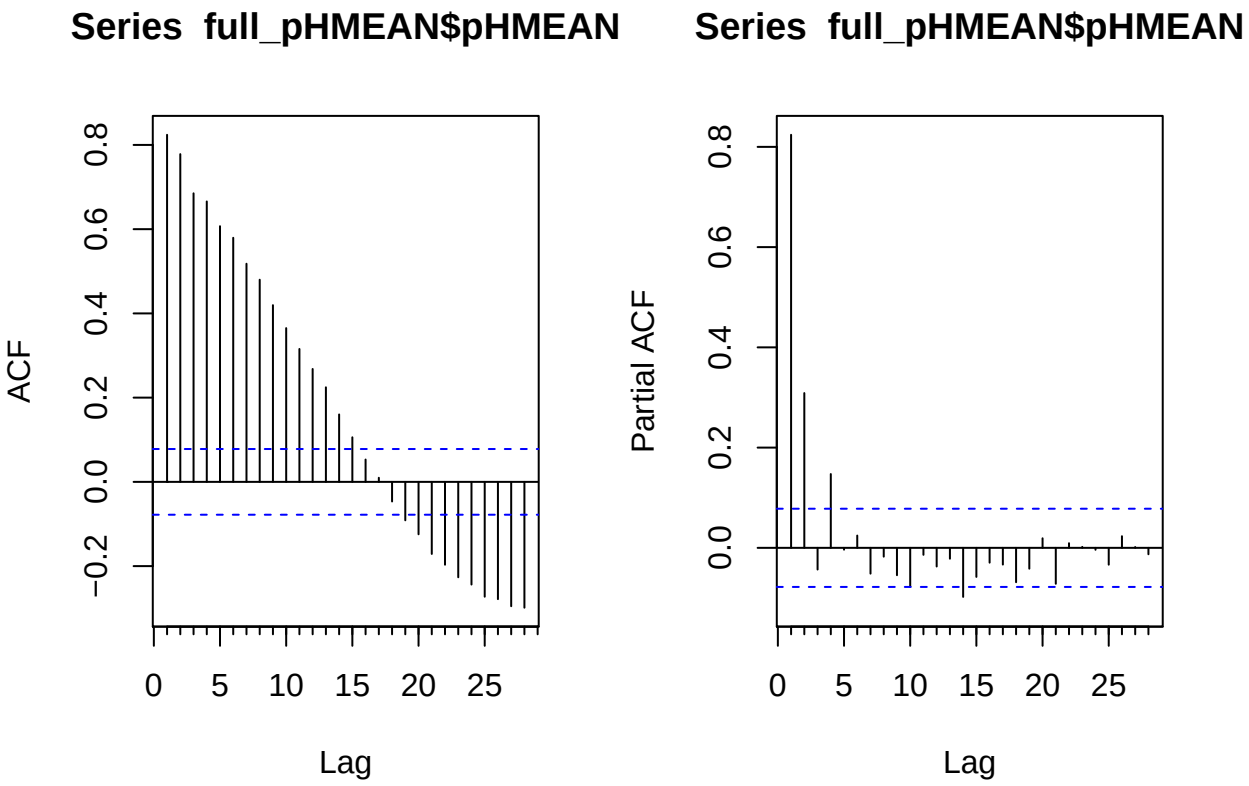


Table 13: Seasonal Mann Kendall test for pH

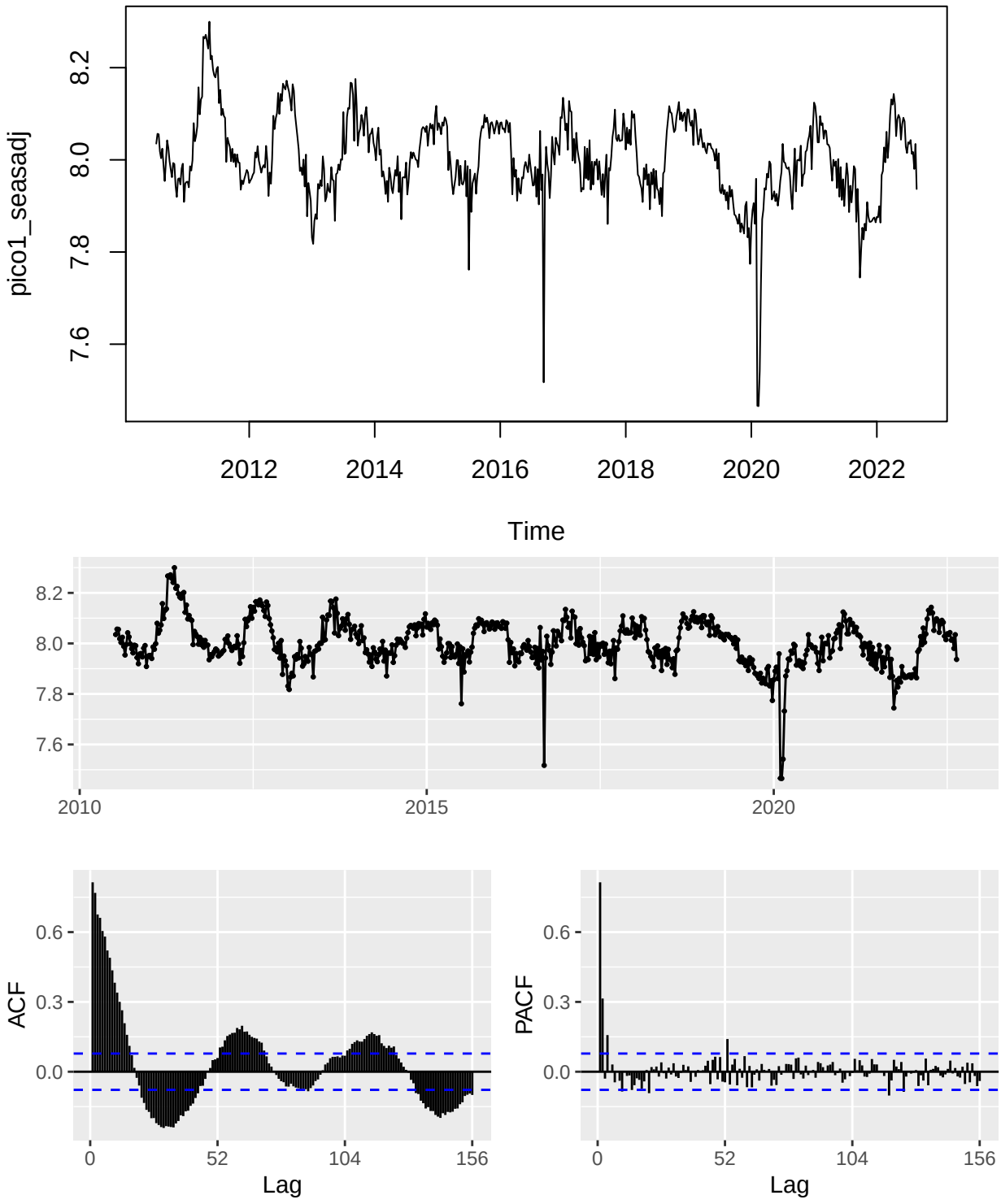
p.value	kendall_score	statistic
0.0021753	-328	-0.0932878

```
##
##  Seasonal Mann-Kendall trend test (Hirsch-Slack test)
##
## data:  pico1.ts
## z = -3.0559, p-value = 0.002244
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##        S     varS
##  -328.00 11450.67
```

p-value is less than 0.05, so we can reject the null hypothesis – There is a possible trend in the data




```
## Series: pico1.ts
## ARIMA(1,1,3)
##
## Coefficients:
##           ar1     ma1      ma2      ma3
##       -0.7071  0.3294  -0.2159  -0.1328
## s.e.   0.1300  0.1343   0.0685   0.0410
##
## sigma^2 estimated as 0.002793:  log likelihood=960.33
## AIC=-1910.66   AICc=-1910.57   BIC=-1888.43
```



Test for Stationarity:

```
## Warning in adf.test(pico1_seasadj): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  pico1_seasadj
```

```
## Dickey-Fuller = -4.7127, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary
```

```
## Warning in kpss.test(pico1_seasadj): p-value smaller than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  pico1_seasadj
## KPSS Level = 0.85542, Truncation lag parameter = 6, p-value = 0.01
```

Conflicting conclusions for the two tests, so I will first-difference the time series to see if that will create stationarity

```
## Warning in adf.test(pico1_seasadj_diff): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  pico1_seasadj_diff
## Dickey-Fuller = -8.8457, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary
```

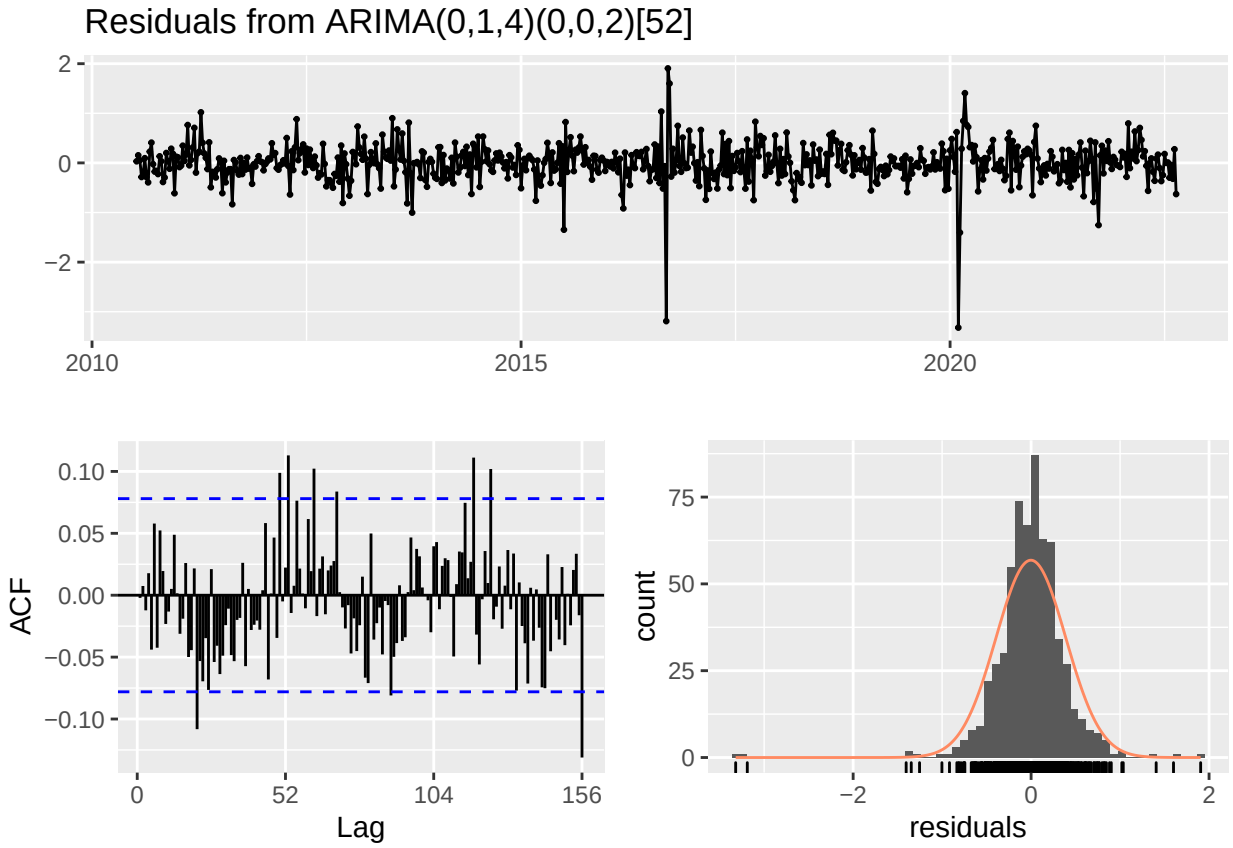
```
## Warning in kpss.test(pico1_seasadj_diff): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  pico1_seasadj_diff
## KPSS Level = 0.011185, Truncation lag parameter = 6, p-value = 0.1
```

Because the p-value for the KPSS is larger than an alpha value of 0.05, the null hypothesis cannot be rejected. However, we can reject the null hypothesis for the ADF test. Therefore, it can now be concluded that the time series is stationary.

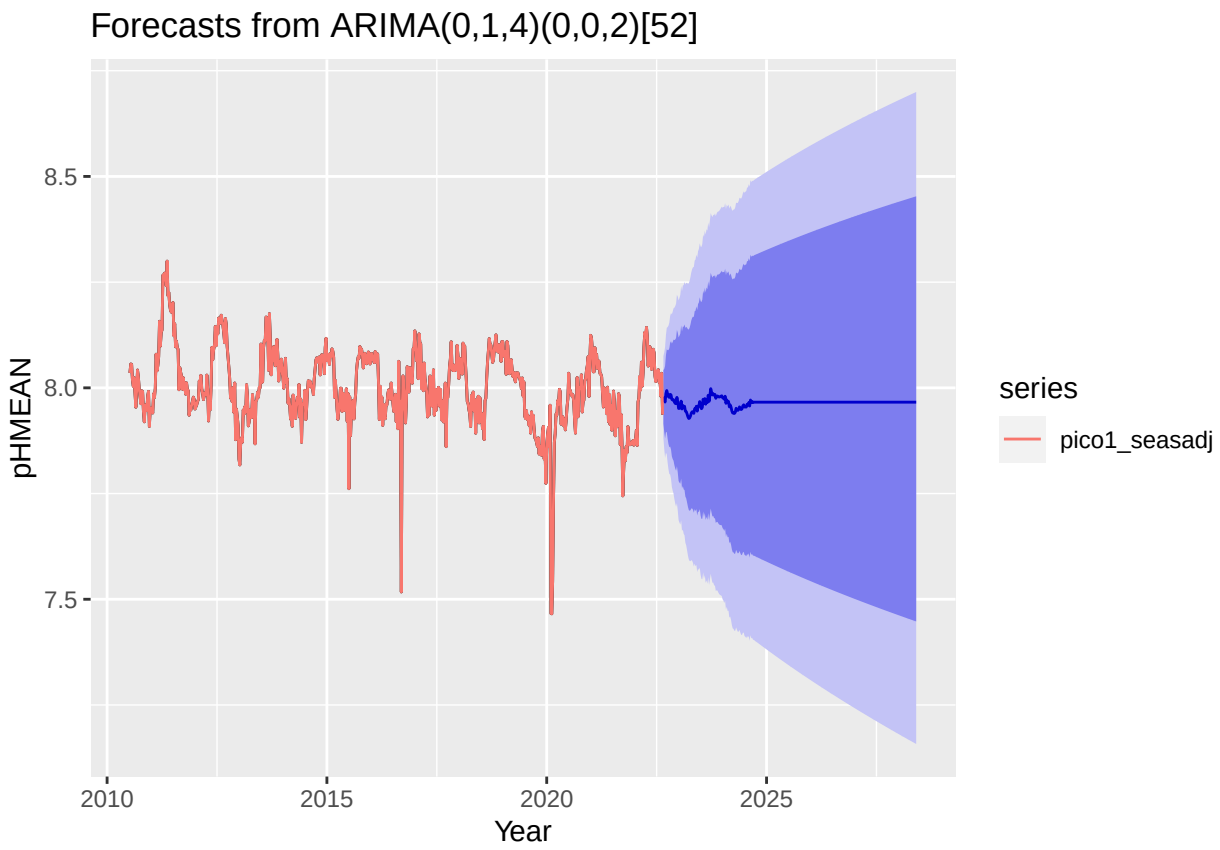
```
## Series: pico1_seasadj
## ARIMA(0,1,4)(0,0,2)[52]
##
## Coefficients:
##           ma1     ma2      ma3     ma4     sma1     sma2
##       -0.3853  0.0486  -0.1800  0.0887  -0.1576  -0.1442
## s.e.   0.0403  0.0426   0.0397  0.0377   0.0499   0.0579
##
## sigma^2 estimated as 0.002499:  log likelihood=994.37
## AIC=-1974.74   AICc=-1974.56   BIC=-1943.62
```

```
## Series: pico1_seasadj
## ARIMA(0,1,4)(0,0,2)[52]
## Box Cox transformation: lambda= 1.999924
##
## Coefficients:
##           ma1     ma2      ma3     ma4     sma1     sma2
##       -0.3857  0.0518  -0.1778  0.0930  -0.1588  -0.1468
## s.e.   0.0403  0.0426   0.0398  0.0377   0.0503   0.0585
##
## sigma^2 estimated as 0.1555:  log likelihood=-306.76
## AIC=627.53   AICc=627.71   BIC=658.65
```



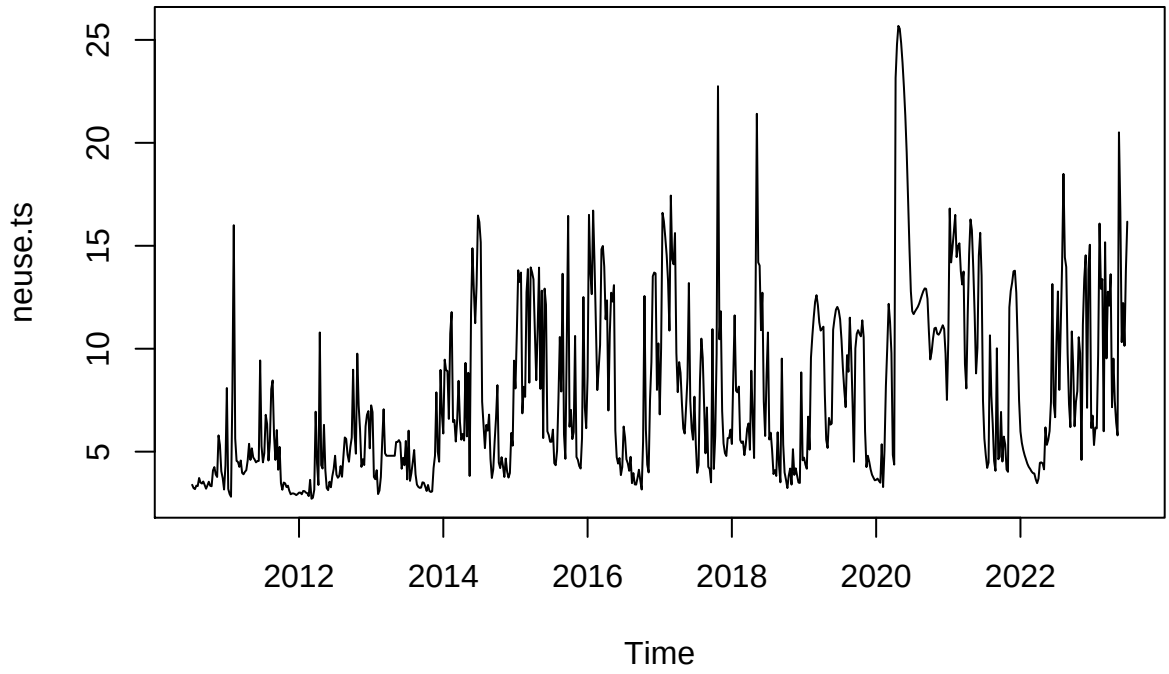

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,4)(0,0,2)[52]
## Q* = 7.3346, df = 6, p-value = 0.291
##
## Model df: 6.   Total lags used: 12
```

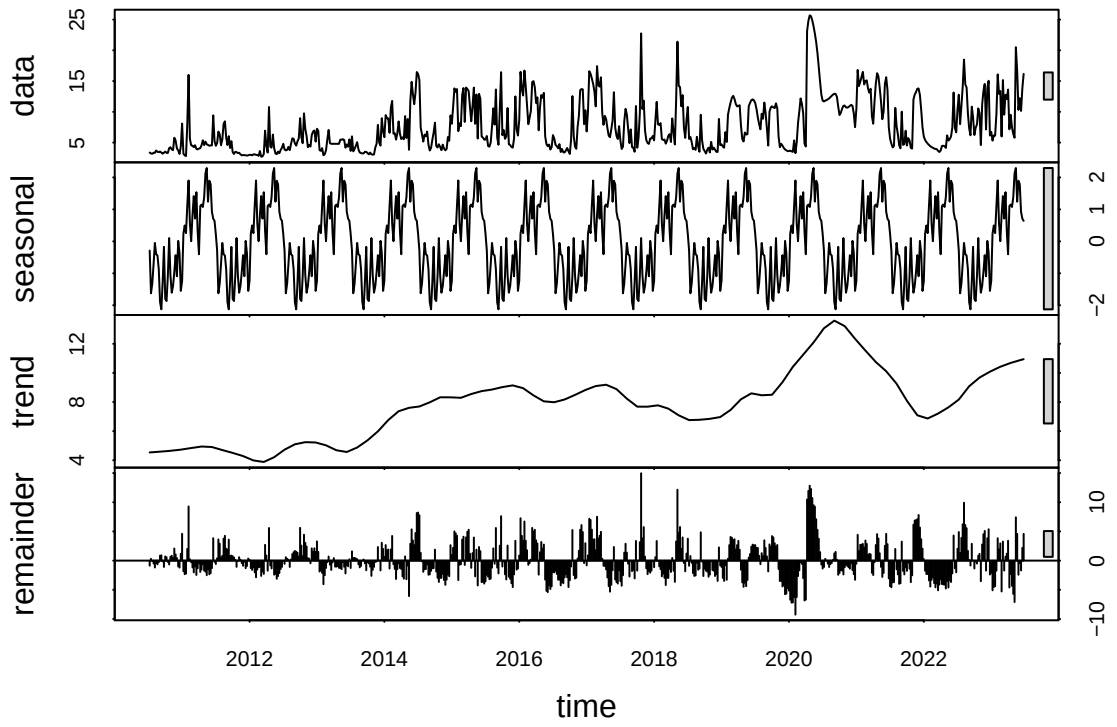
```
## [1] 627.7082
```



River Data Time Series

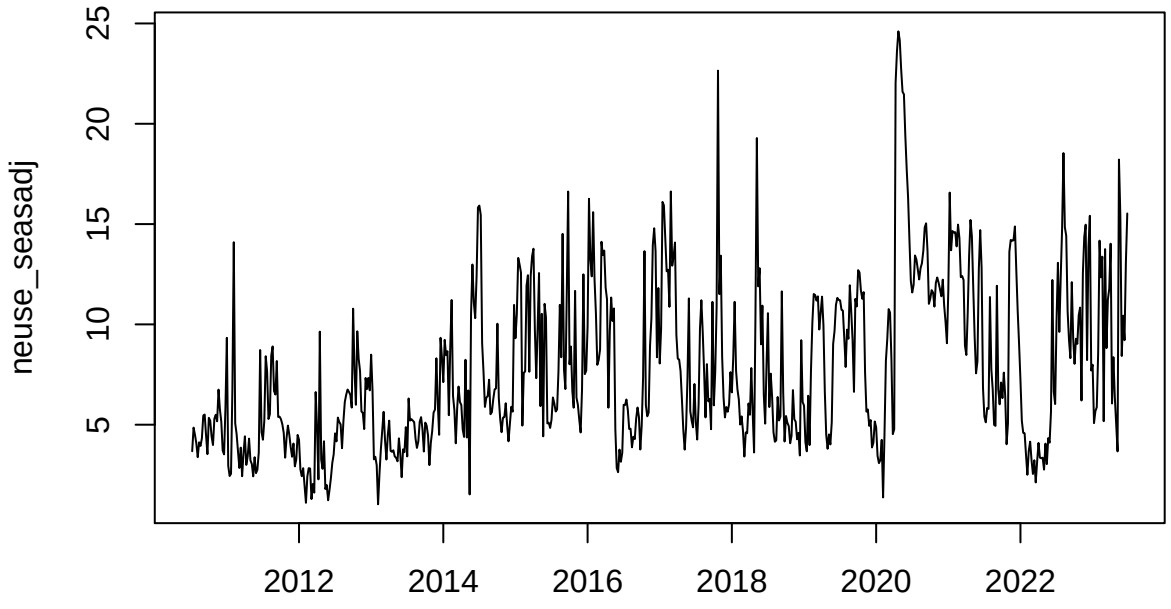
Neuse River



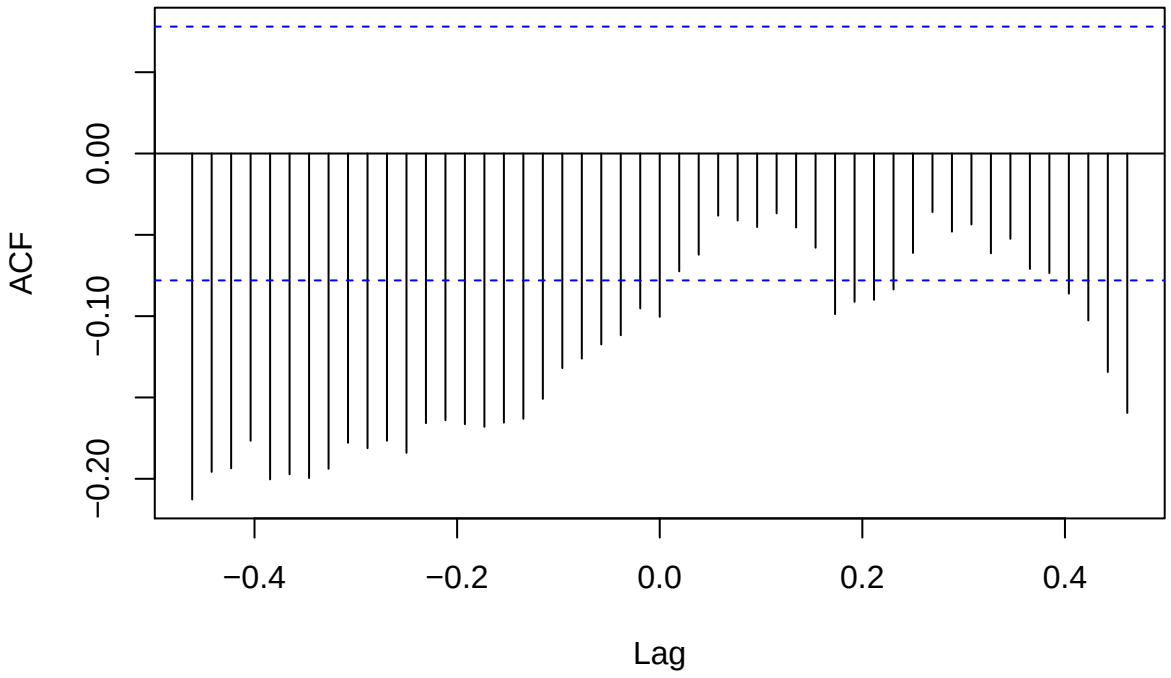



```
## Warning in adf.test(neuse.ts): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  neuse.ts
## Dickey-Fuller = -6.1267, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary

## Warning in kpss.test(neuse.ts): p-value smaller than printed p-value

##
##  KPSS Test for Level Stationarity
##
## data:  neuse.ts
## KPSS Level = 2.6583, Truncation lag parameter = 6, p-value = 0.01
```

Neuse River gage height is stationary. However, it appears as though there may be seasonality, so I will remove that before looking at the correlation.



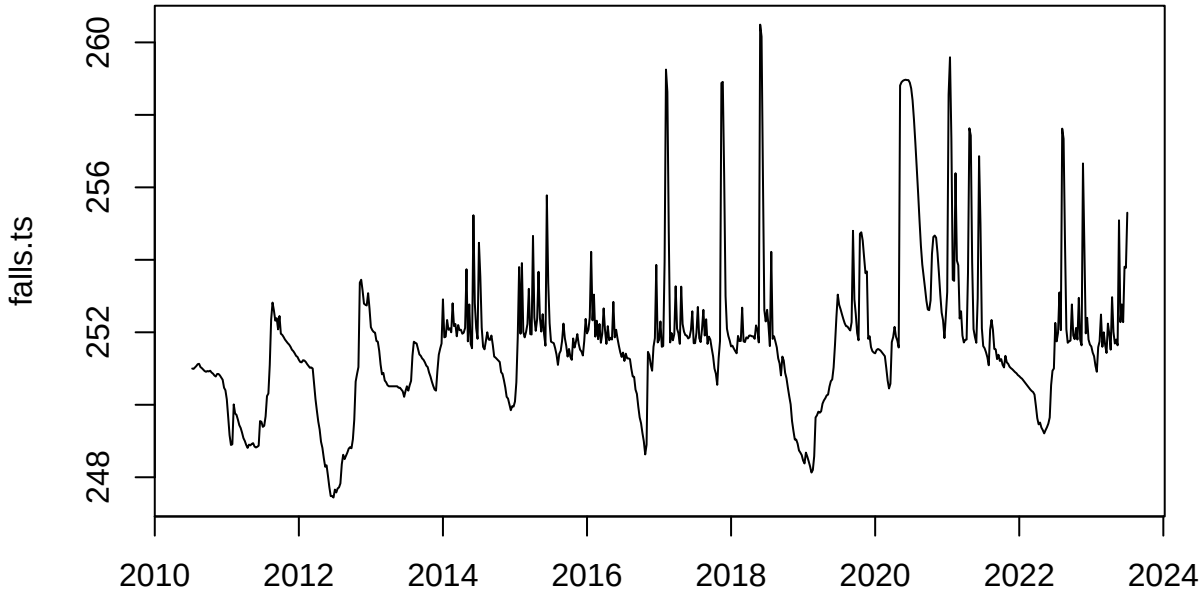
neuse_seasadj & pico1_seasadj

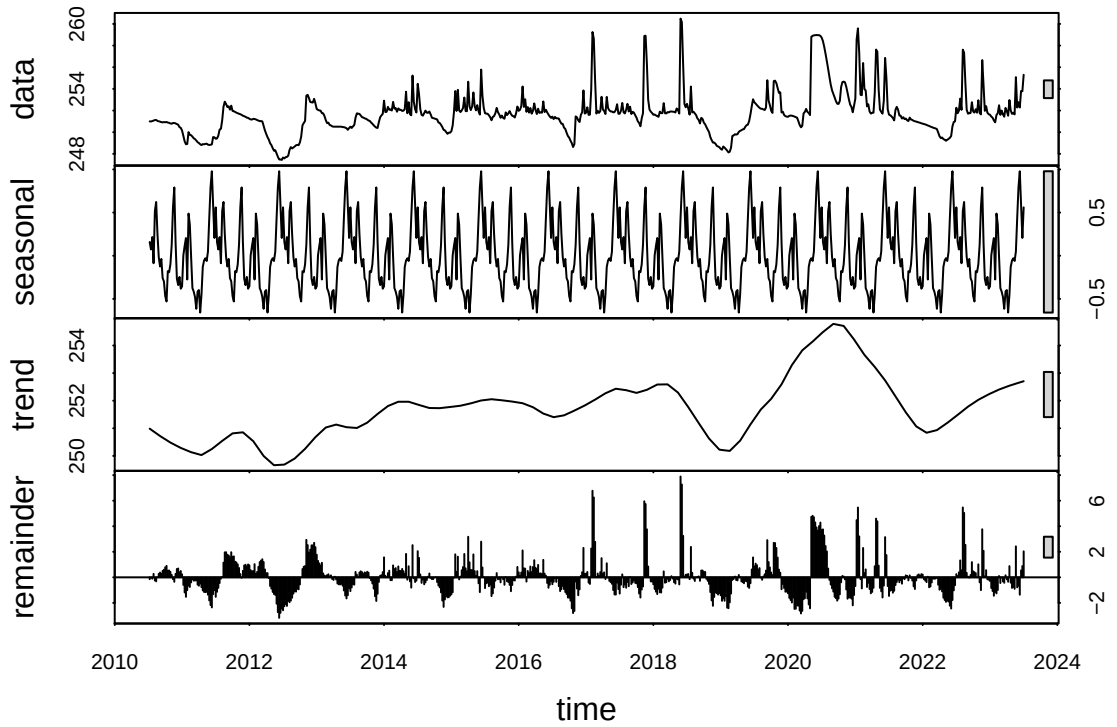



```
##
## Autocorrelations of series 'X', by lag
##
## -0.4615 -0.4423 -0.4231 -0.4038 -0.3846 -0.3654 -0.3462 -0.3269 -0.3077 -0.2885
##  -0.213  -0.196  -0.194  -0.177  -0.200  -0.197  -0.200  -0.194  -0.178  -0.181
## -0.2692 -0.2500 -0.2308 -0.2115 -0.1923 -0.1731 -0.1538 -0.1346 -0.1154 -0.0962
##  -0.177  -0.184  -0.166  -0.164  -0.166  -0.168  -0.165  -0.163  -0.151  -0.132
## -0.0769 -0.0577 -0.0385 -0.0192  0.0000  0.0192  0.0385  0.0577  0.0769  0.0962
```

```
##  -0.126  -0.117  -0.112  -0.095  -0.100  -0.072  -0.062  -0.038  -0.041  -0.045
##  0.1154  0.1346  0.1538  0.1731  0.1923  0.2115  0.2308  0.2500  0.2692  0.2885
##  -0.037  -0.045  -0.058  -0.099  -0.091  -0.090  -0.084  -0.061  -0.036  -0.048
##  0.3077  0.3269  0.3462  0.3654  0.3846  0.4038  0.4231  0.4423  0.4615
##  -0.044  -0.061  -0.052  -0.071  -0.074  -0.086  -0.103  -0.134  -0.160
```

Falls Lake



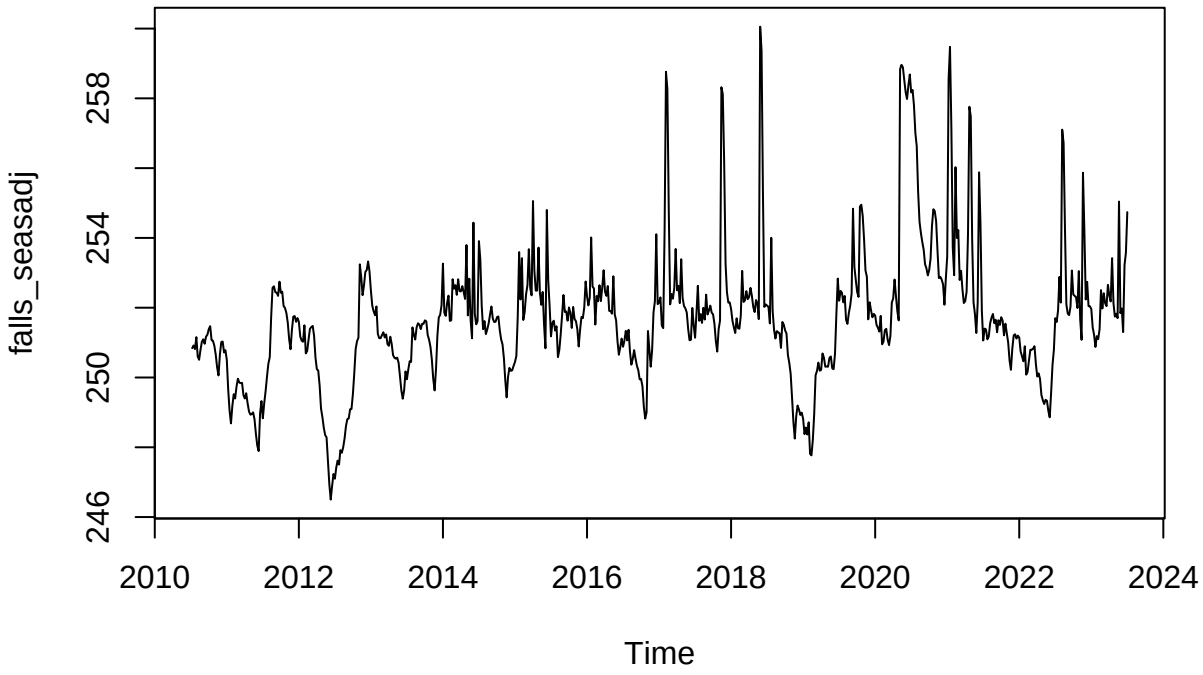


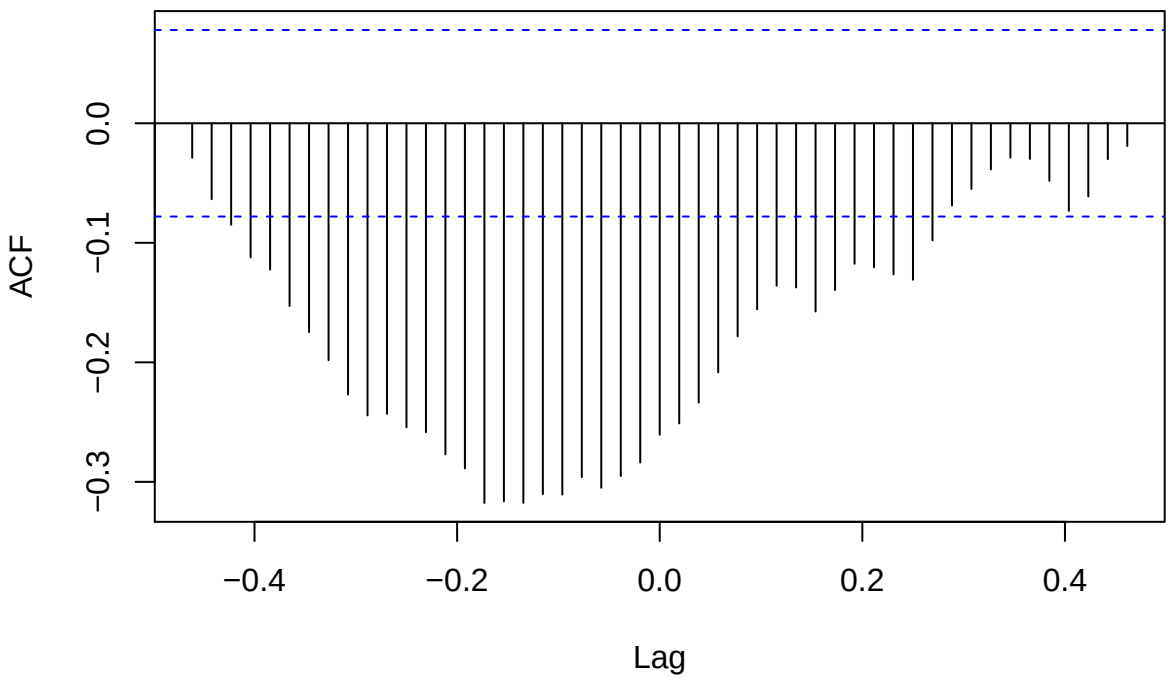

```
##
##  Phillips-Perron Unit Root Test
```

```
##
## data:  falls.ts
## Dickey-Fuller = -7.1968, Truncation lag parameter = 6, p-value = 0.01
```

```
## Warning in kpss.test(falls.ts): p-value smaller than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  falls.ts
## KPSS Level = 1.6263, Truncation lag parameter = 6, p-value = 0.01
```






```
##
## Autocorrelations of series 'X', by lag
##
## -0.4615 -0.4423 -0.4231 -0.4038 -0.3846 -0.3654 -0.3462 -0.3269 -0.3077 -0.2885
##  -0.029  -0.064  -0.085  -0.112  -0.122  -0.153  -0.175  -0.198  -0.227  -0.244
## -0.2692 -0.2500 -0.2308 -0.2115 -0.1923 -0.1731 -0.1538 -0.1346 -0.1154 -0.0962
##  -0.243  -0.254  -0.258  -0.277  -0.289  -0.318  -0.316  -0.318  -0.310  -0.311
## -0.0769 -0.0577 -0.0385 -0.0192  0.0000  0.0192  0.0385  0.0577  0.0769  0.0962
##  -0.296  -0.305  -0.295  -0.284  -0.261  -0.251  -0.234  -0.208  -0.178  -0.156
##  0.1154  0.1346  0.1538  0.1731  0.1923  0.2115  0.2308  0.2500  0.2692  0.2885
##  -0.136  -0.137  -0.157  -0.139  -0.118  -0.121  -0.126  -0.131  -0.098  -0.069
##  0.3077  0.3269  0.3462  0.3654  0.3846  0.4038  0.4231  0.4423  0.4615
##  -0.055  -0.039  -0.029  -0.030  -0.048  -0.073  -0.061  -0.030  -0.019
```