

Bike Share Demand Forecasting Methods

Chance Robinson and Kebur Fantahun

12/1/2021

Contents

1	Introduction	2
2	Data Analysis	3
3	Methods	21
3.1	ARMA Model	21
3.2	Vector Auto-Regressive (VAR) Model	31
3.3	Neural Network Model	39
3.4	Ensemble Model	41
4	Results	43
4.1	7 Day Forecast	43
4.2	1 Day Forecast	44
5	Appendix	45
5.1	Solution One	45
5.2	Solution Two	47

1 Introduction

Bike-sharing is a popular service that allows people to pick up a bike from one dock, use it and return it to the same or another docking station all for a small fee, sometimes for free. As opposed to transport like cars, bikes offer an inexpensive, eco-friendly option to urban people by promoting less pollution and dealing with less traffic. As such, bike-sharing programs have become very popular and in turn provide a wealth of data on the mobility in a city.

Data gathered by the rental companies includes things like temperature, count of users, humidity and more. The explicit collection of data has the potential to assist in understanding the highs and lows of bike demand which could influence; when to offer a sale, which areas require redistribution, and in general when the best time to offer rentals is. Future data to inquire about and include could be the demographics of the city, the style of current marketing and how often redistribution is needed.

2 Data Analysis

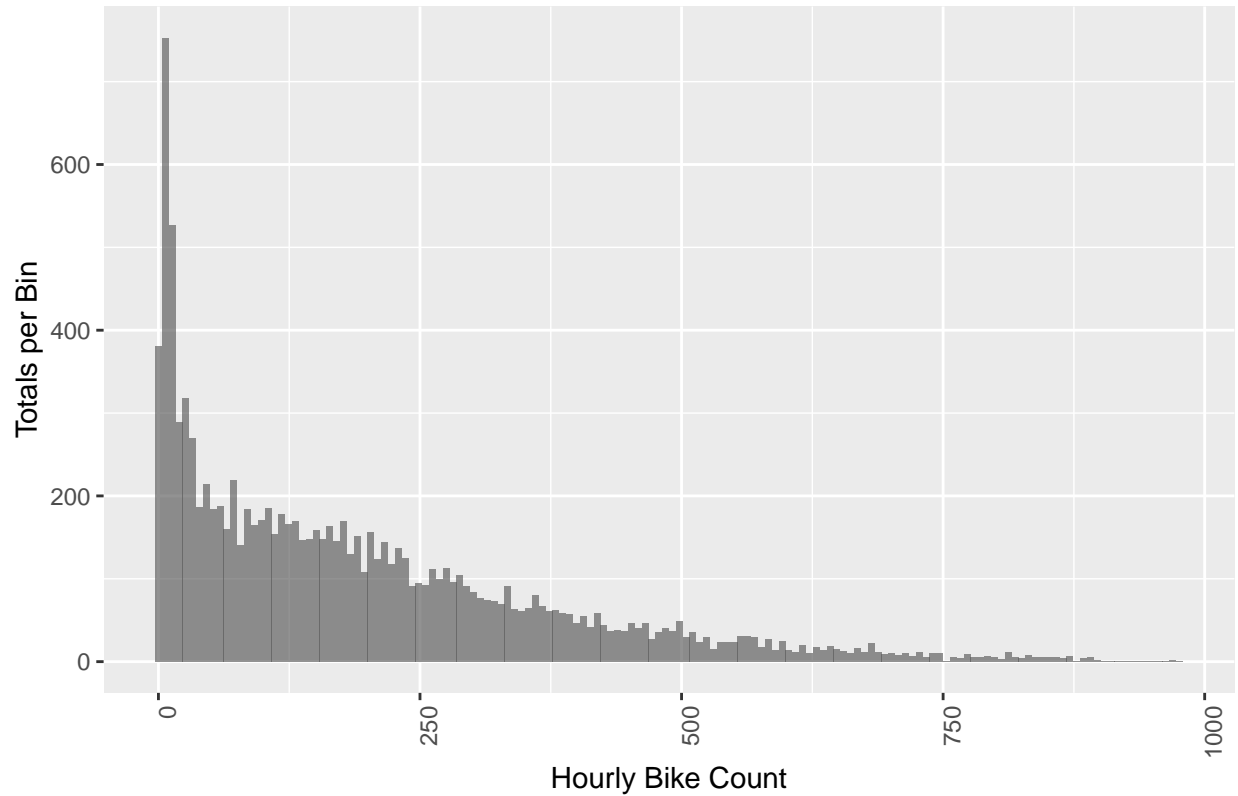
The data collected for this project consists of hourly bike share rentals from January 1st, 2011 through December 31st, 2012.

There are 12 columns provided in the `train.csv` data set with 10,886 observations. The `test.csv` data set has 6,493 records, or roughly 37% of the overall combined samples from the two files. This is due to the fact that the training data consist of the first 19 days of each month and the test the remaining 11~12 days. The test set does not include the response variables for casual, registered or total users.

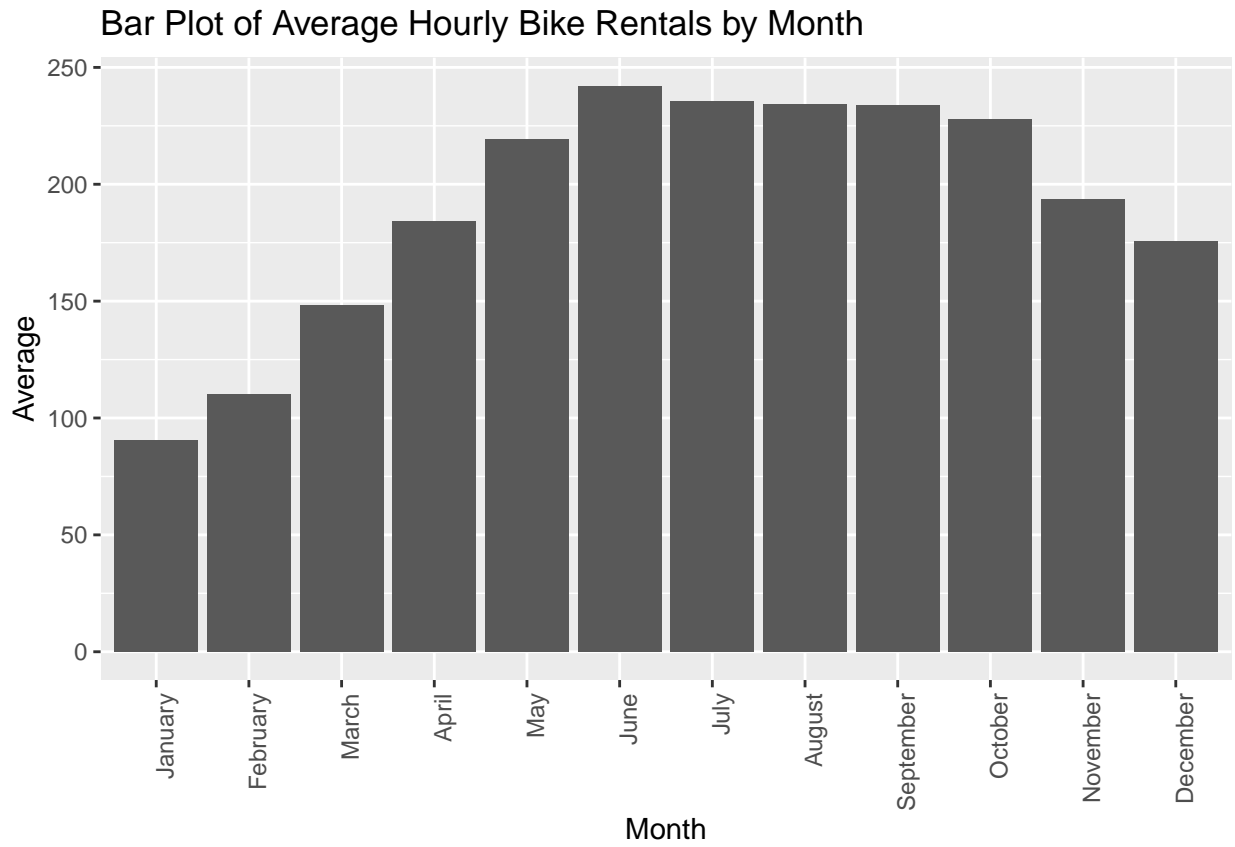
Column Name	Type	Description
1. datetime	Character	YYYY-MM-DD HH24 (example: 2011-01-01 04:00:00)
2. season	Integer	(1-4)
3. holiday	Integer	(0 or 1)
4. workingday	Integer	(0 or 1)
5. weather	Integer	(1-4)
6. temp	Float	temperature in Celsius
7. atemp	Float	“feels like” temperature in Celsius
8. humidity	Integer	relative humidity
9. windspeed	Float	wind speed
10. casual	Integer	count of casual users
11. registered	Integer	count of registered users
12. count	Integer	count of total users (<i>primary response variable</i>)

Var1	Freq
Min.	1.0000
1st Qu.	42.0000
Median	145.0000
Mean	191.5741
3rd Qu.	284.0000
Max.	977.0000

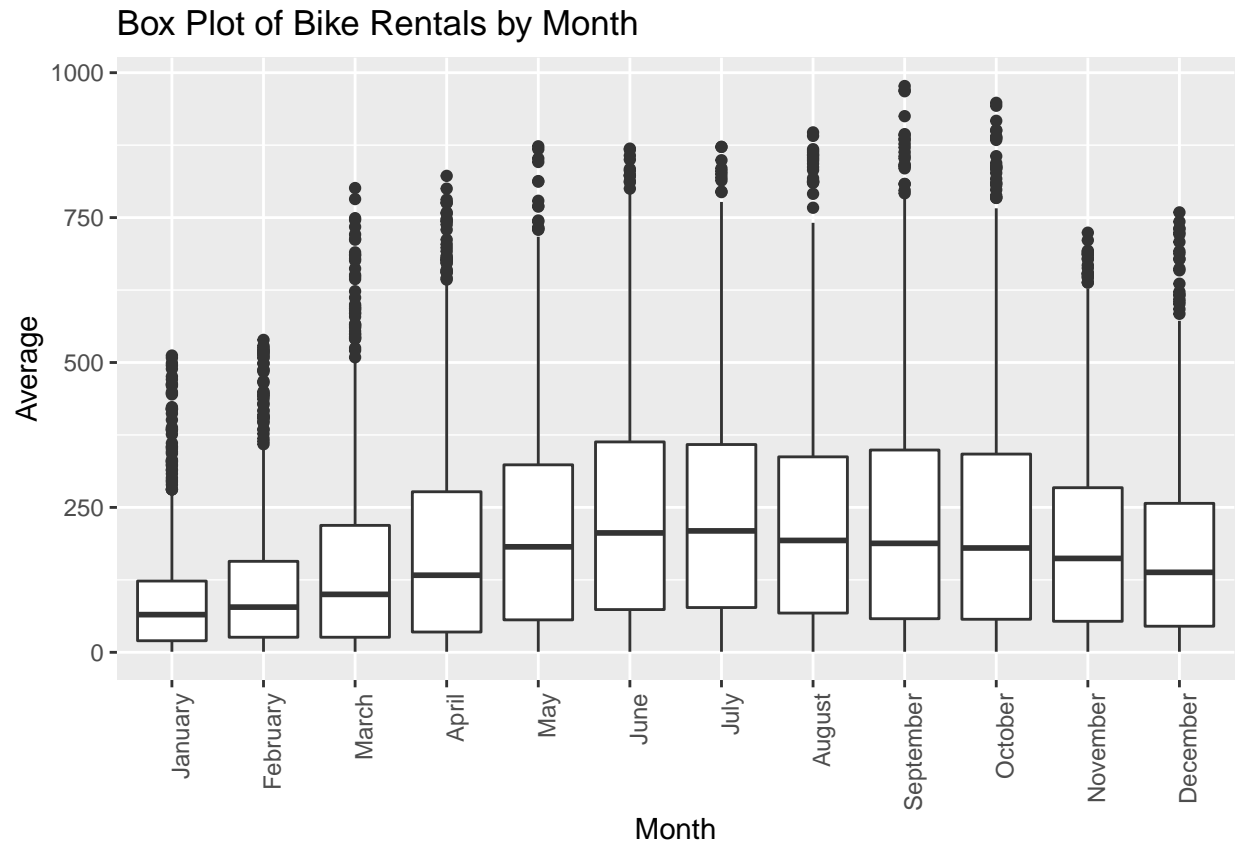
Histogram of Hourly Bike Count



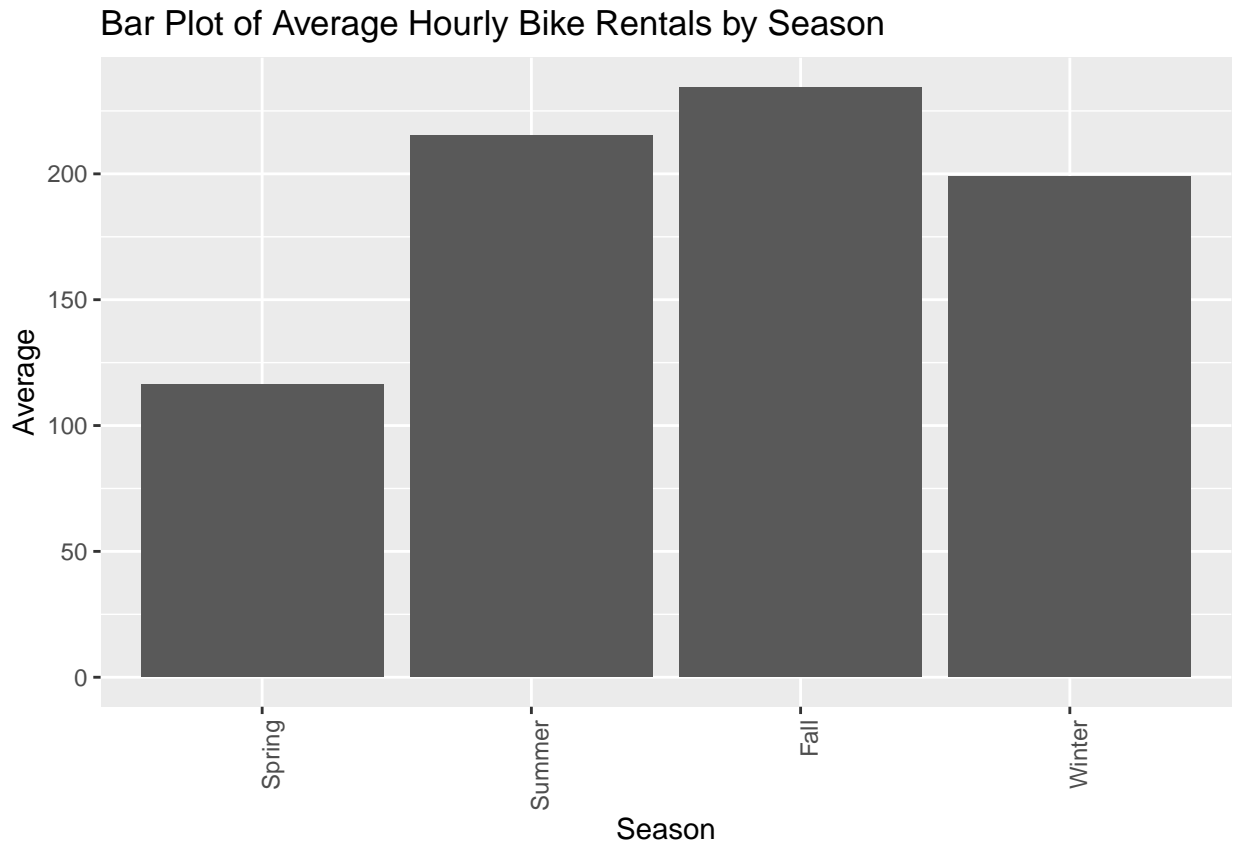
The response variable of `count` appears to be heavily right-skewed, with the median value at ~145 users. Additional summary statistics are show below.



June appears to be the month with heaviest demand.

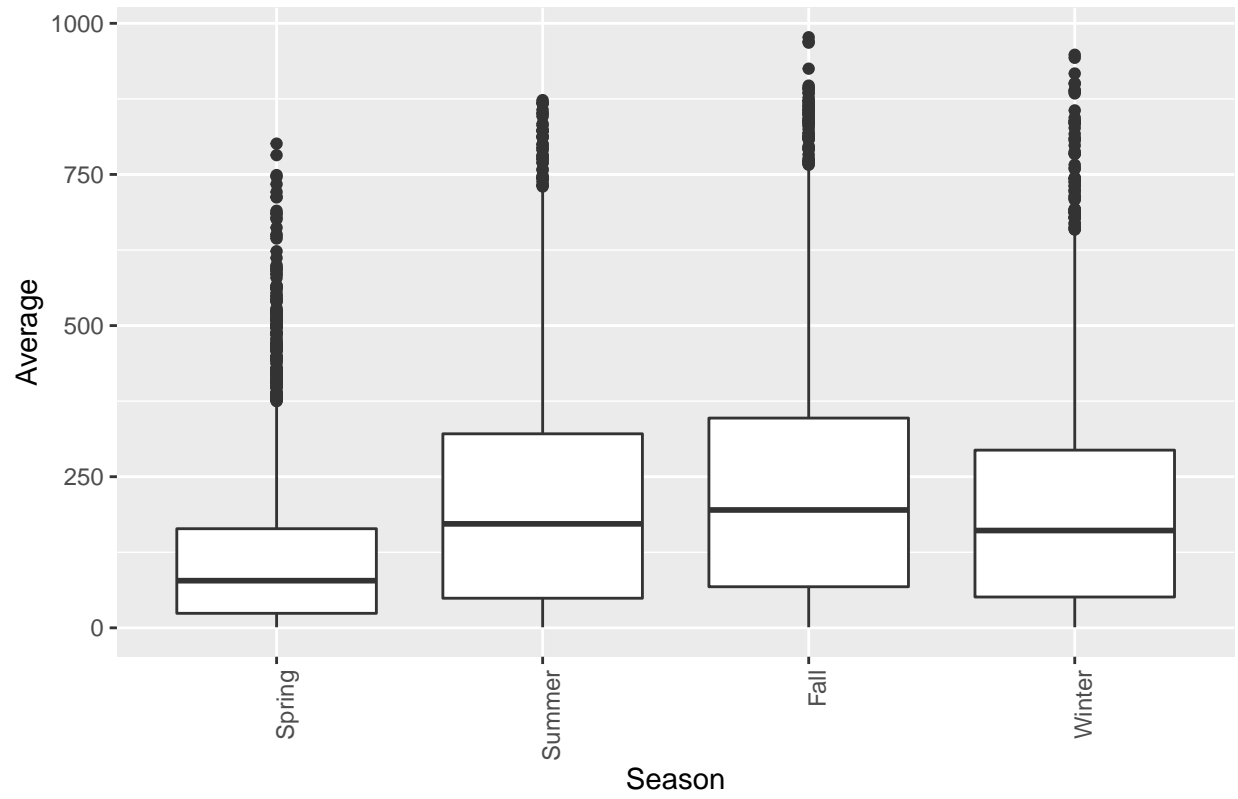


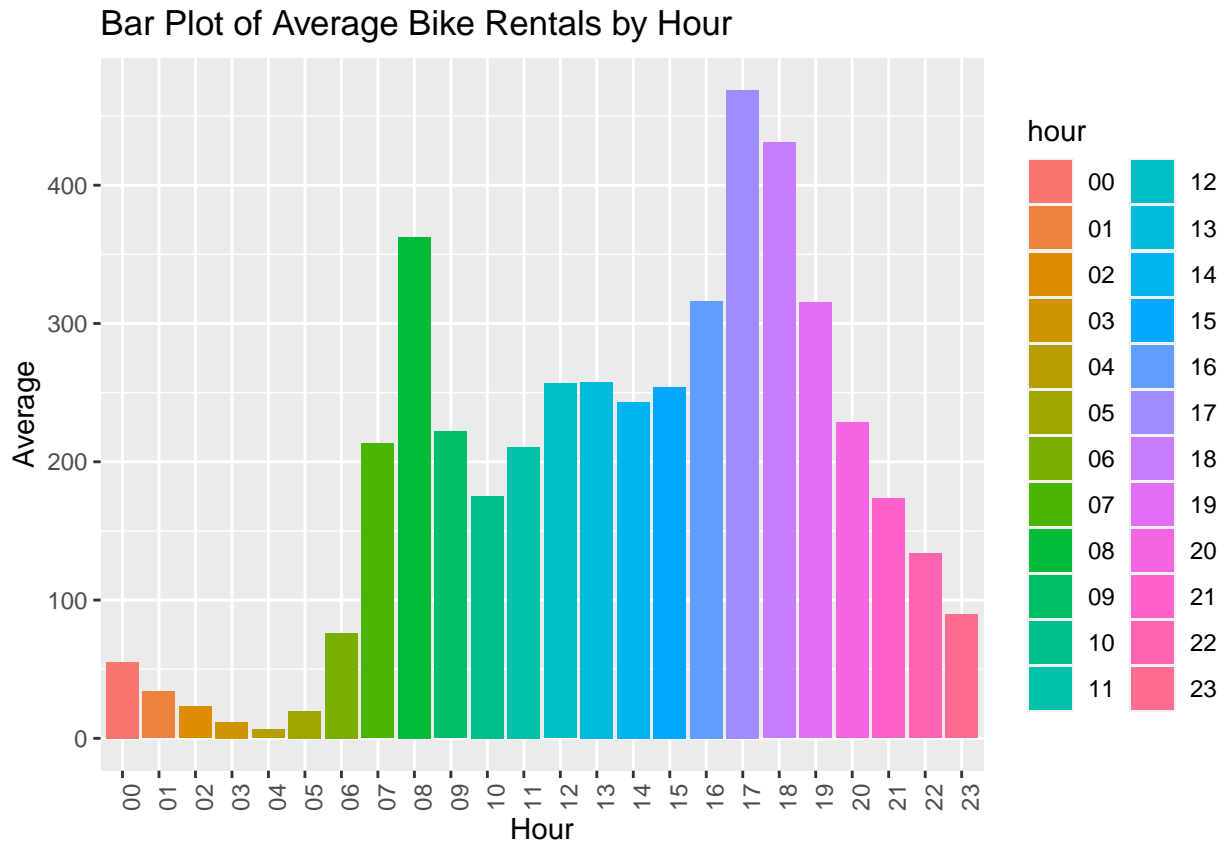
It also appears to show relatively few outliers compared to months like January.



Fall looks to have more rentals on average than the other seasons.

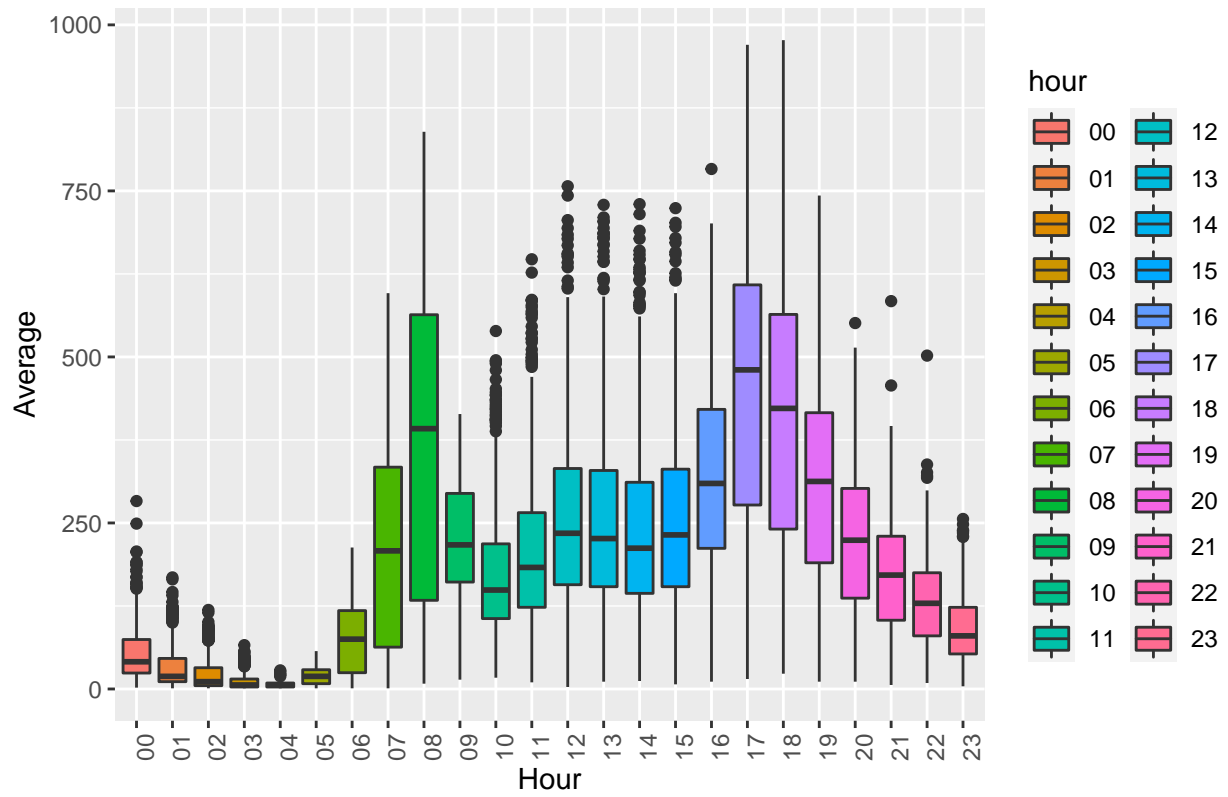
Box Plot of Bike Rentals by Season



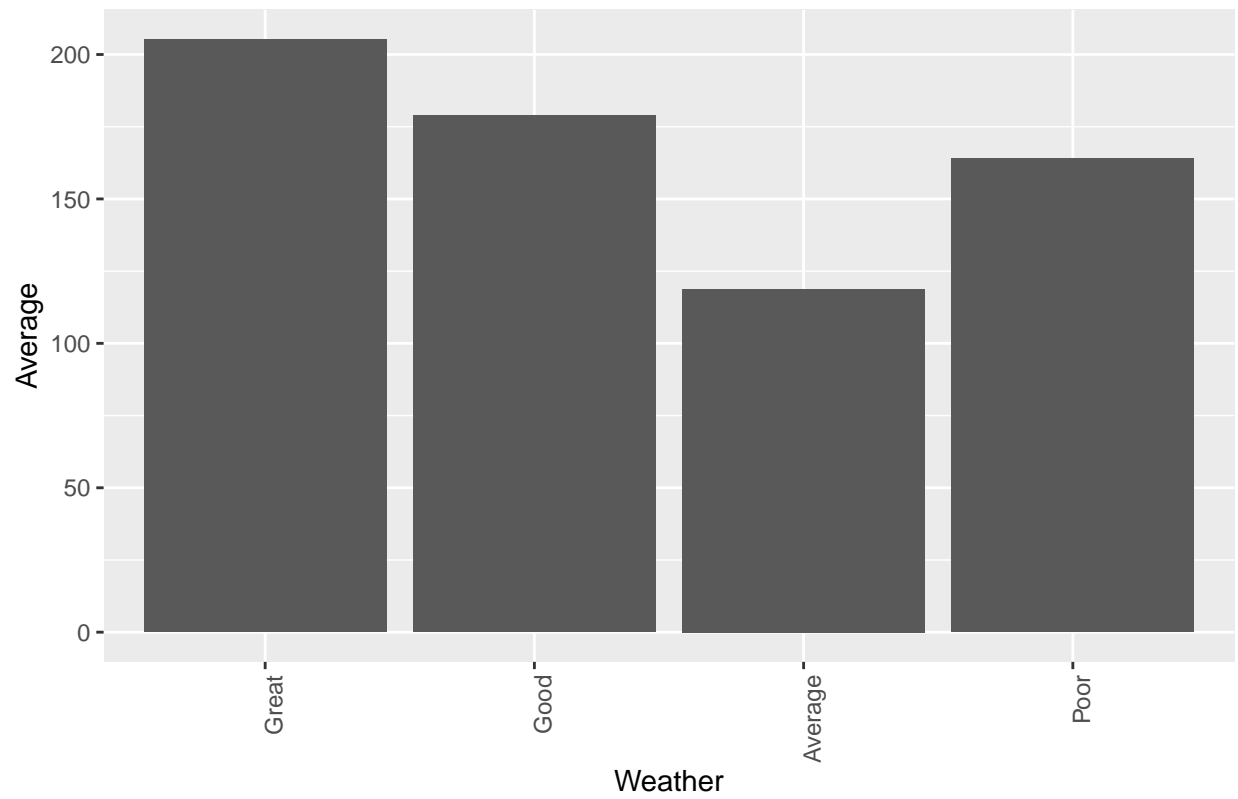


The 5pm hour clearly has the highest peak compared to the other hours of the day.

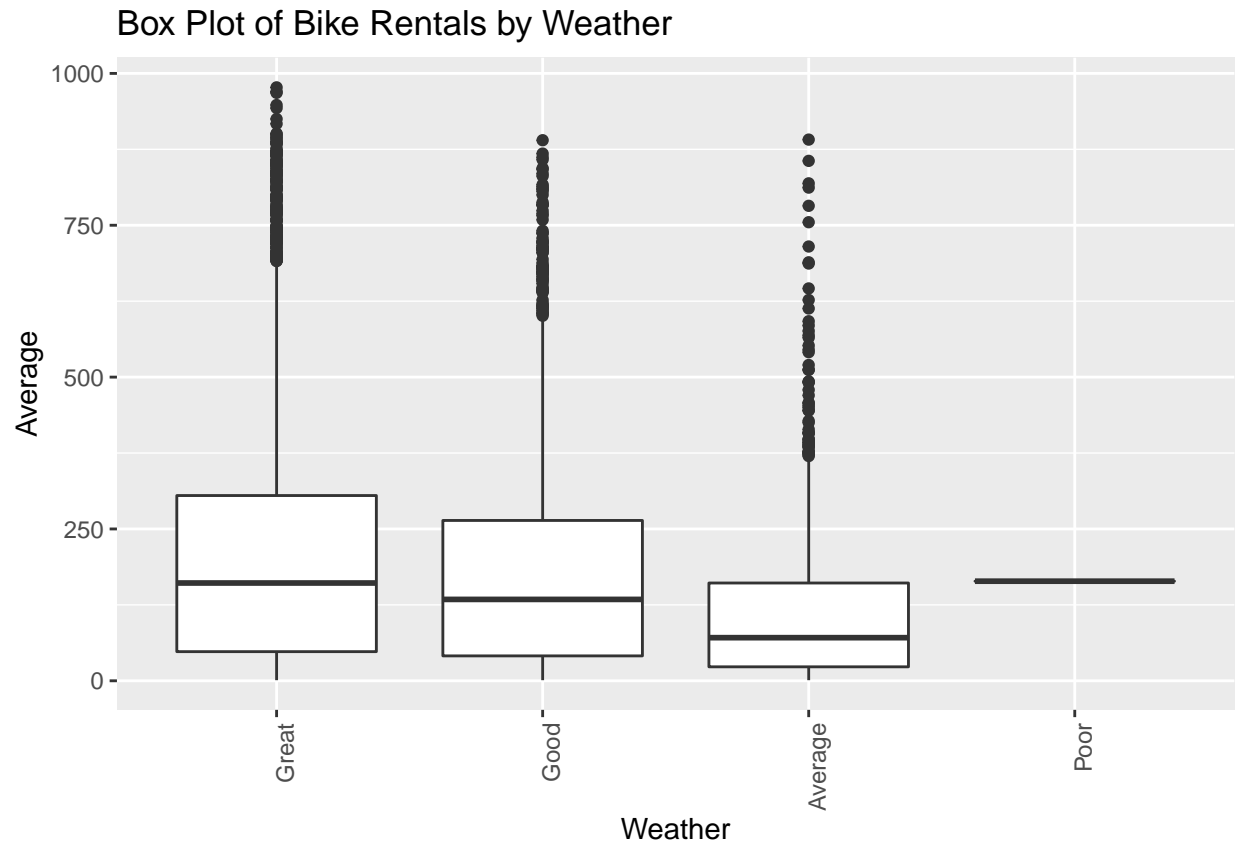
Box Plot of Bike Rentals by Hour



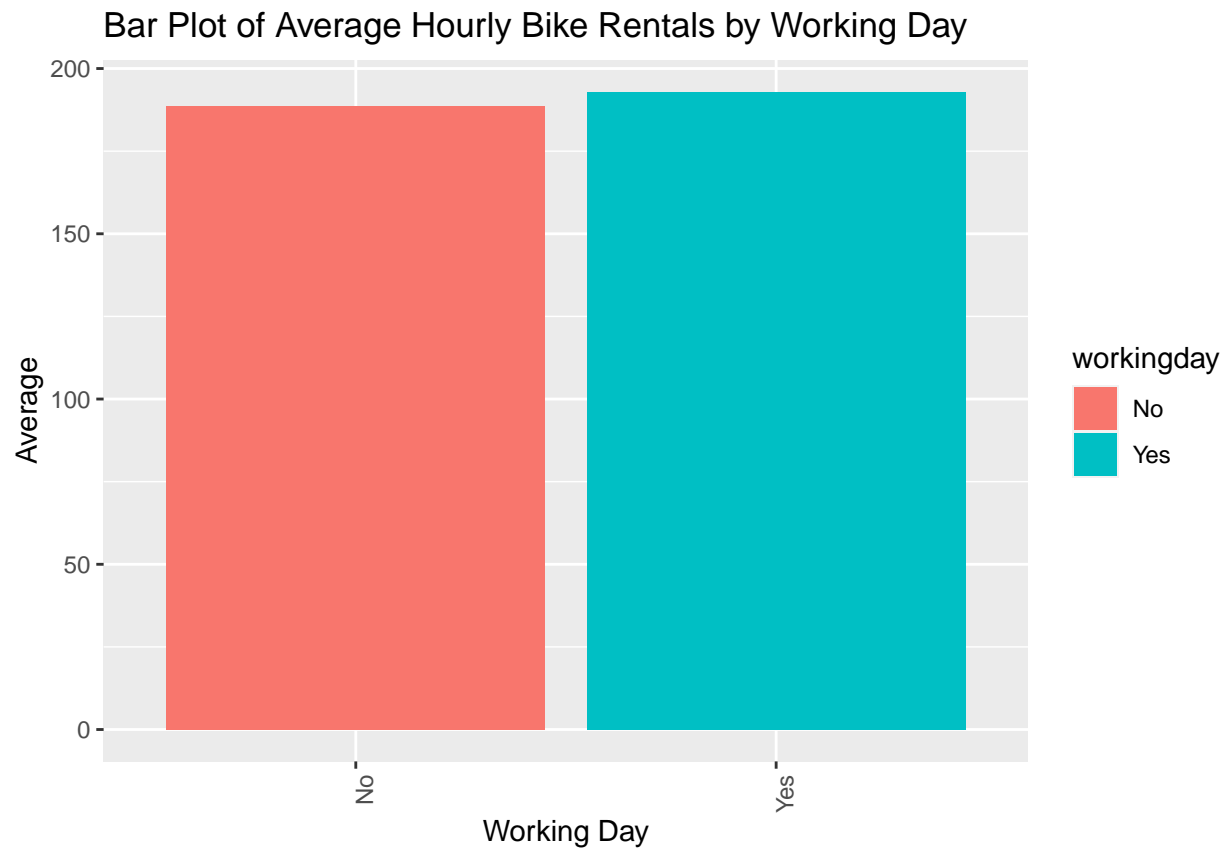
Bar Plot of Average Hourly Bike Rentals by Weather



As expected, more riders are out when the weather is great, or better than average.

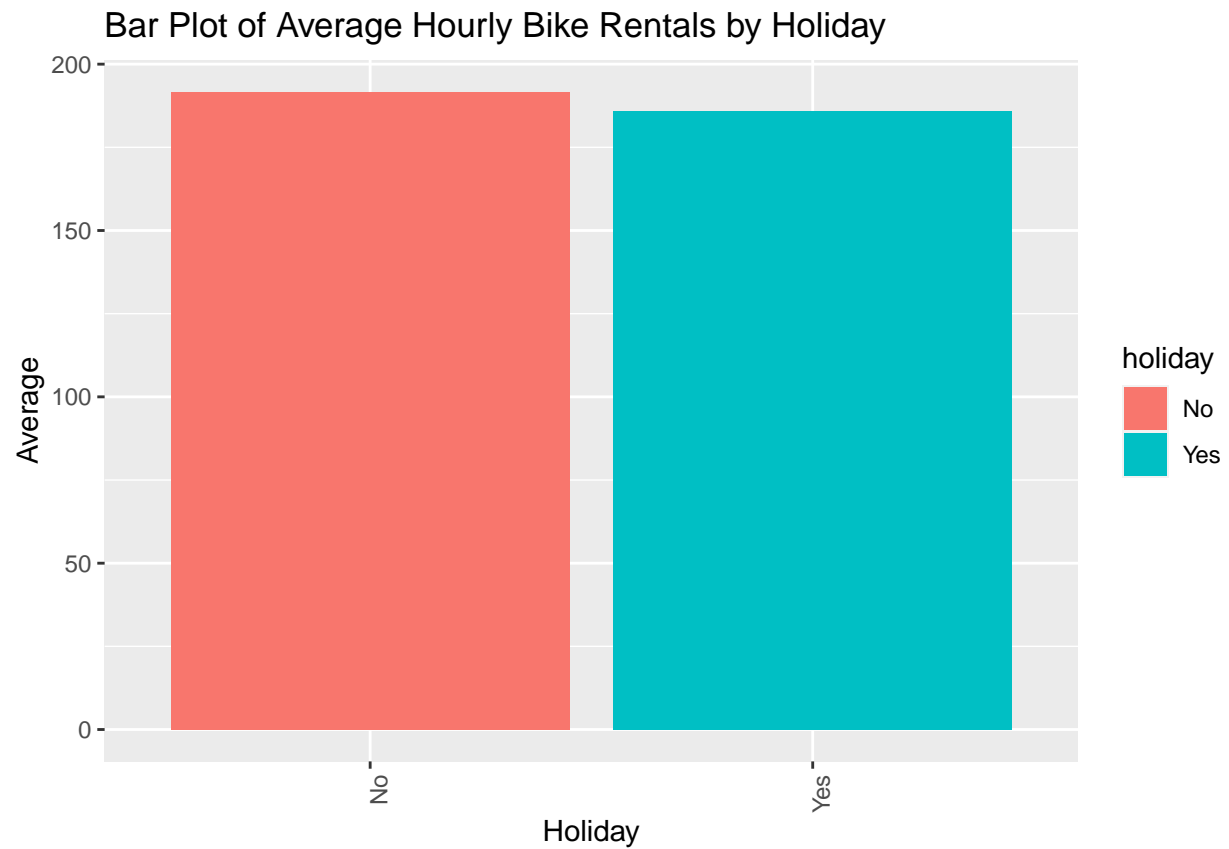


Note the small spread of riders when the weather was listed as poor, even though the averages were consistent with other categories.



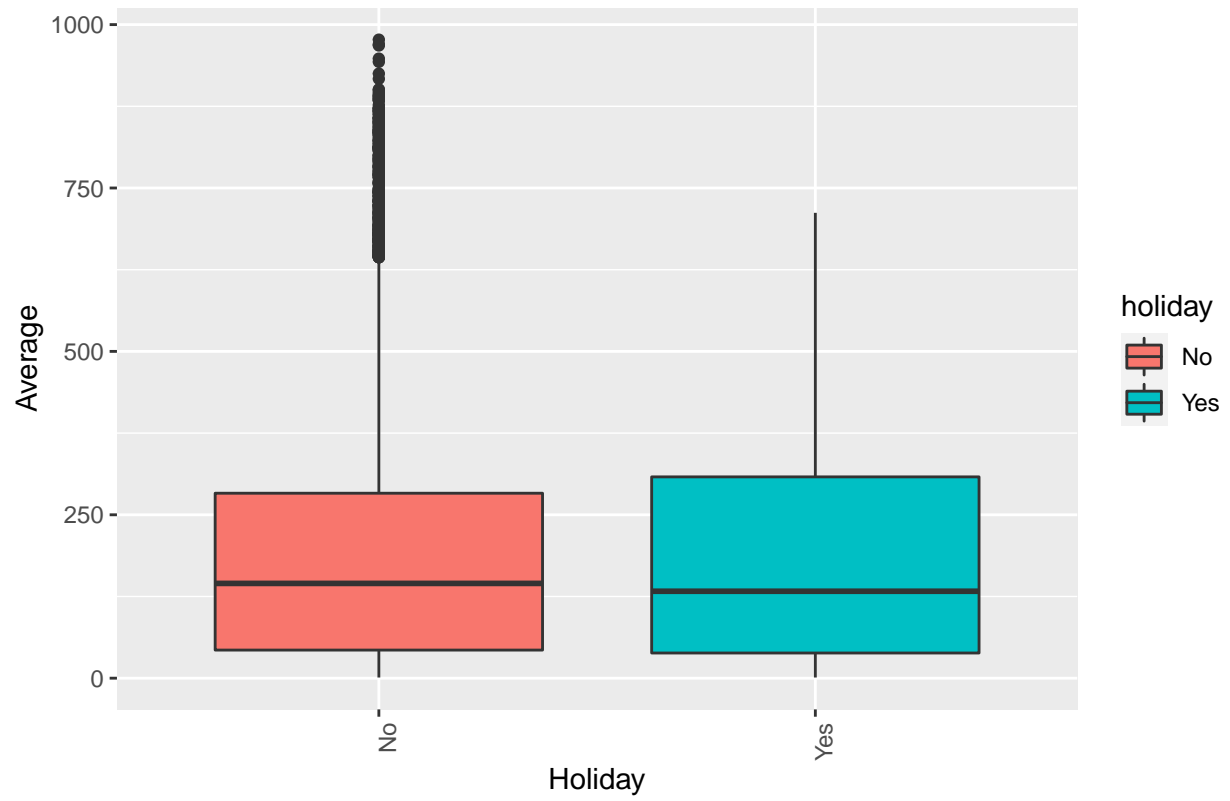
Surprisingly, whether or not the day was on a working day or not had little affect on the mean or median.



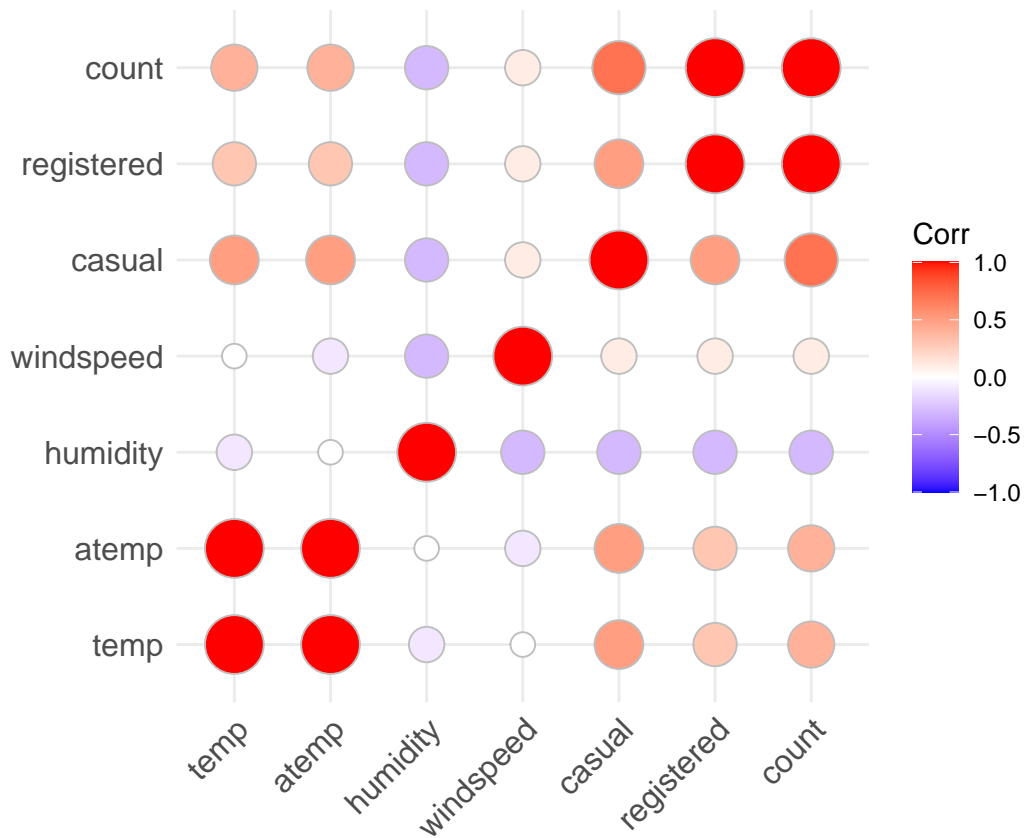


The same was true for days falling on a holiday, it appeared to have little affect on the counts.

Box Plot of Bike Rentals by Holiday



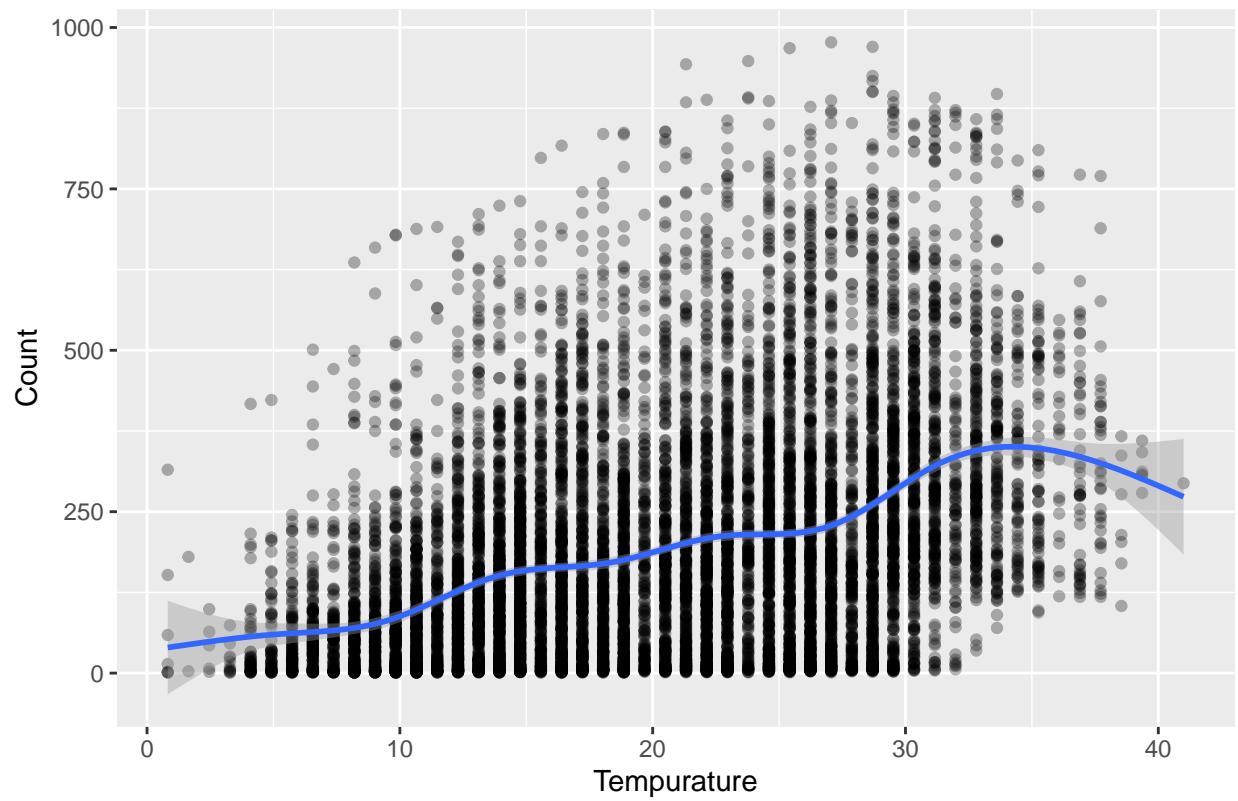
Feature 1	Feature 2	Correlation Coefficient
temp	atemp	0.9849481
registered	count	0.9709481
casual	count	0.6904136
casual	registered	0.4972497
temp	casual	0.4670971
atemp	casual	0.4620665
temp	count	0.3944536
atemp	count	0.3897844
temp	registered	0.3185713
atemp	registered	0.3146354
windspeed	count	0.1013695
humidity	registered	-0.2654579
humidity	count	-0.3173715
humidity	windspeed	-0.3186070
humidity	casual	-0.3481869



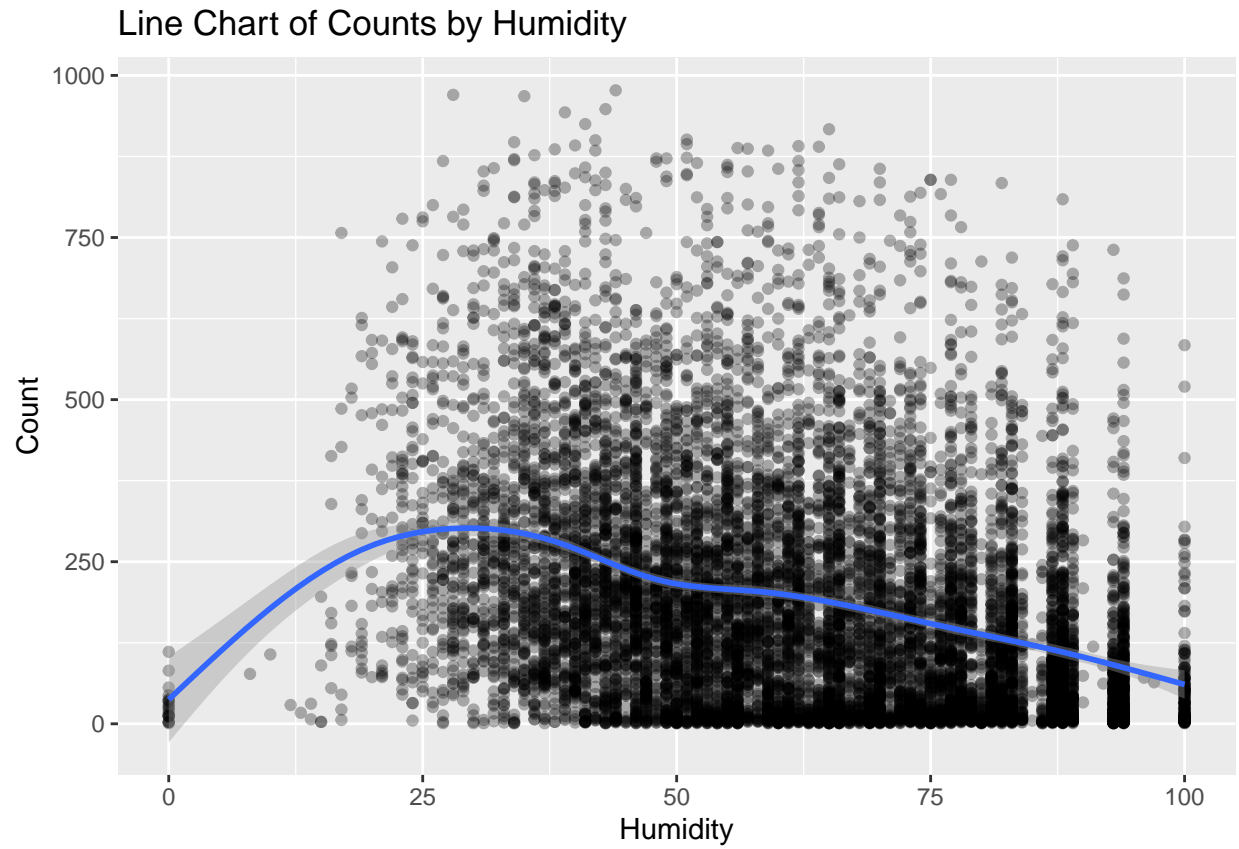
The plot above shows a strong correlation between casual, registered and total (count) users. We will use count as the primary response variable for our modeling, and discard the other secondary response variables.

Note that temp and atemp are also closely related to one another as would be expected.

Line Chart of Counts by Temperature

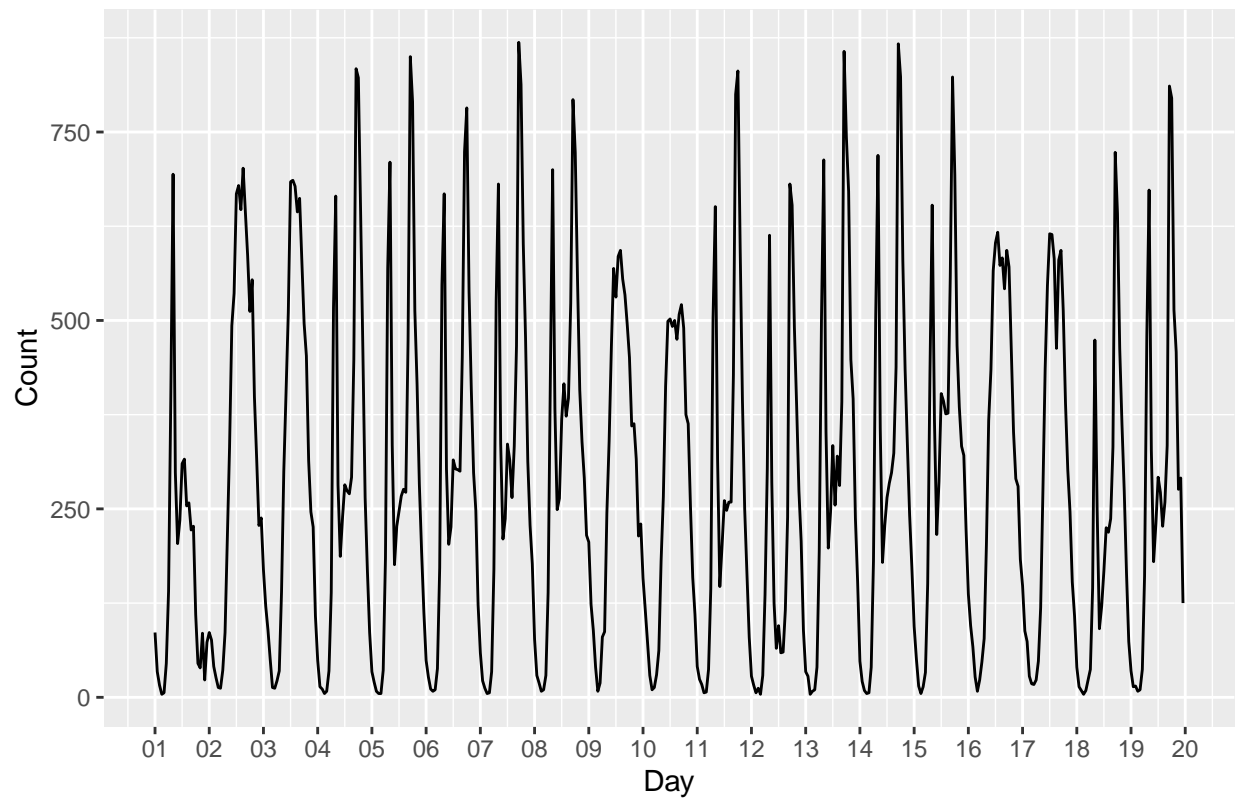


In general, recordings with warmer temperatures have more riders up to a threshold, which looks to start to decline around 34 degrees Celsius (~93 degrees Fahrenheit)



Humidity shows a negative trend, with recordings during high-humidity periods showing fewer riders.

Hourly Rental Trends for June 1st – June 19th, 2012

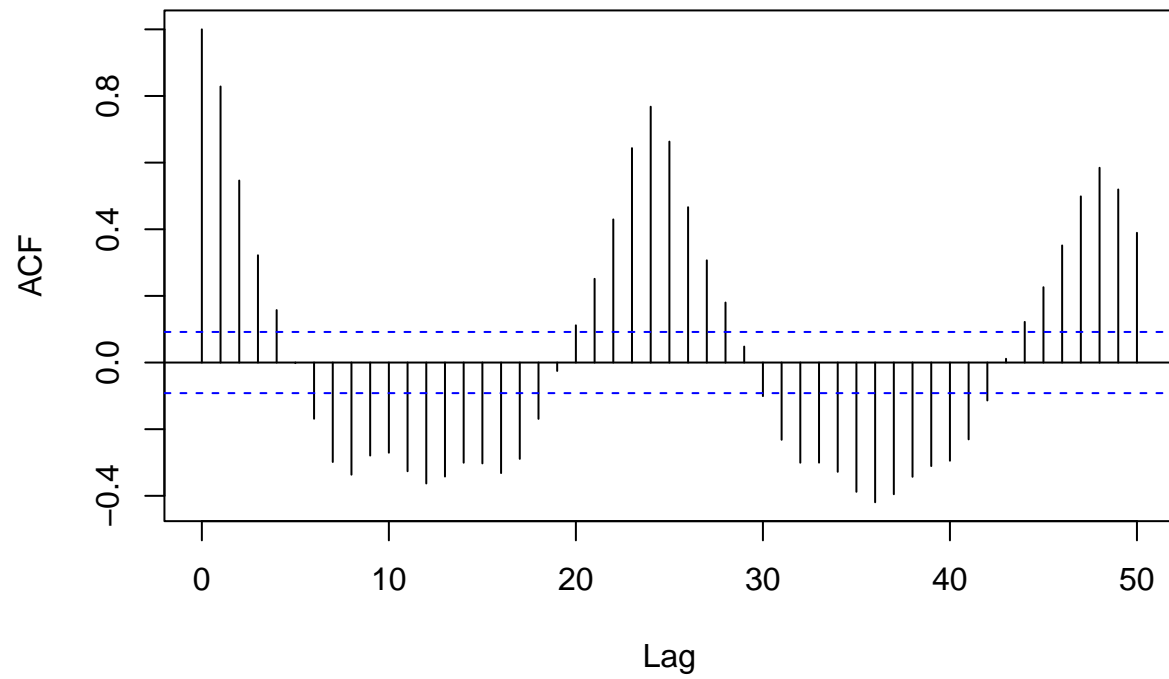


This plot shows the hourly bike rental counts for the first 19 days of June, 2019.

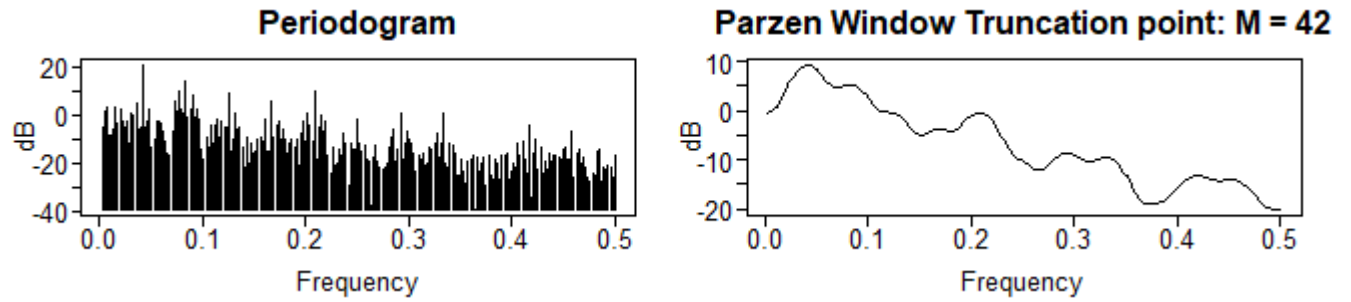
3 Methods

3.1 ARMA Model

Auto-Correlation Plot of Count

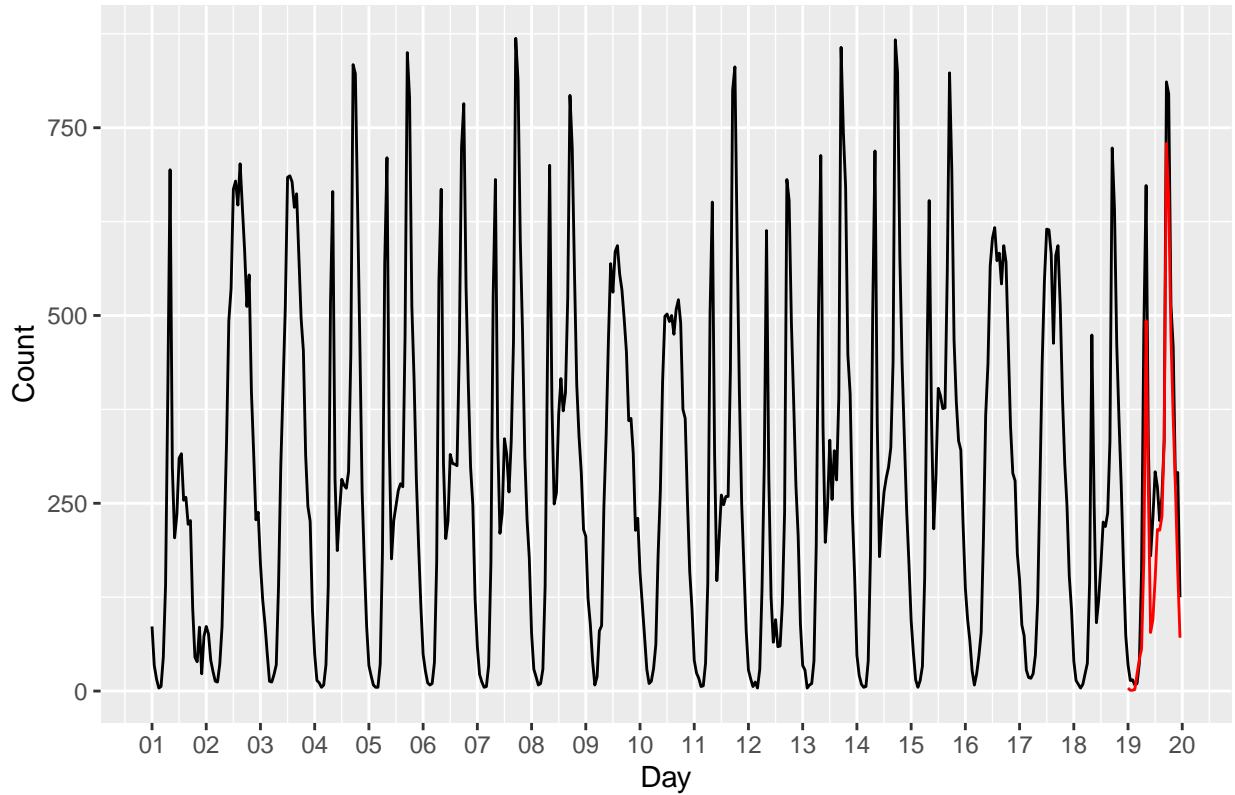


Strong sinusoidal trend with a period of 24, which would likely reflect the hourly cycles from how the data was recorded.



A peak can be seen in the spectral density at around 0.04. Which would equate to a period of 24, which is what would be expected again from the frequency of our data set. You can also visually confirm this from the number of cycles in the realization and divide that by the total number of observations. ($19 / 456 \sim 0.0417$) From there the period can be calculated with $1 / \text{frequency}$. (or ~ 24)

1 Day Forecast (ARMA w/ S=24)



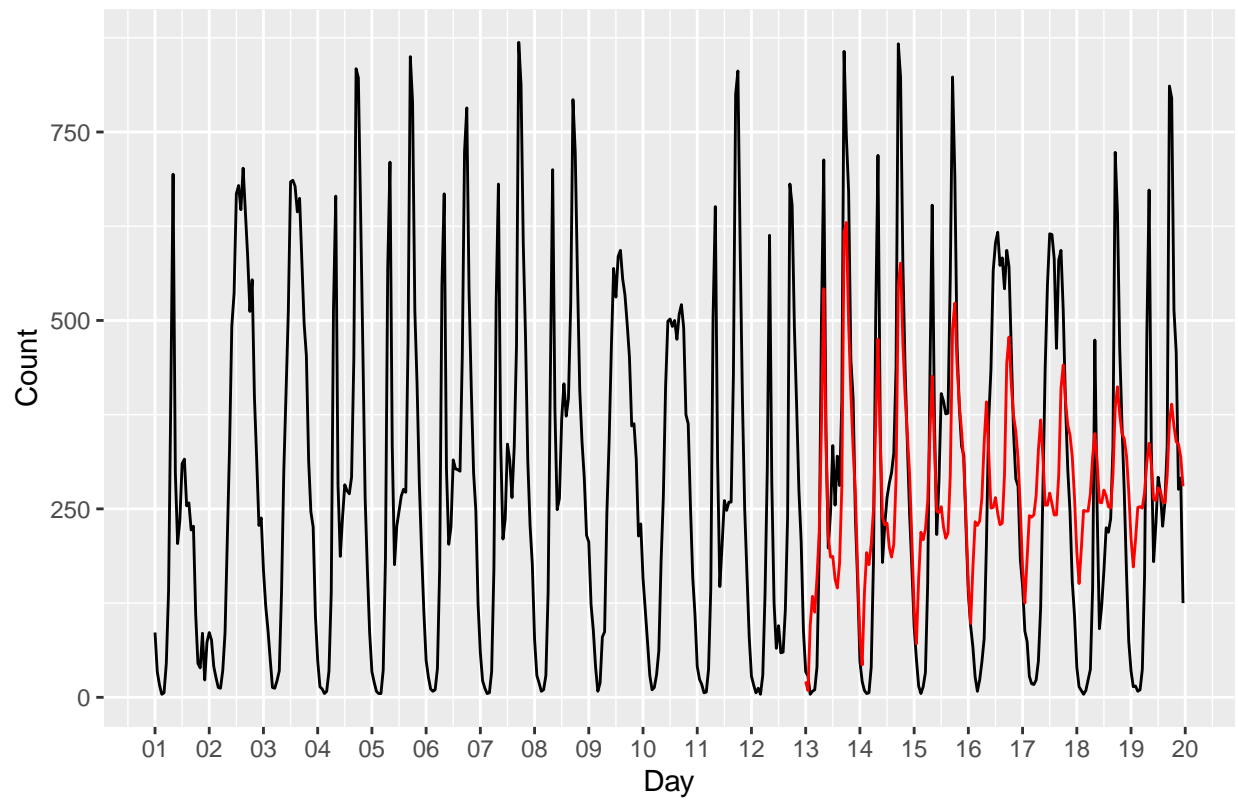
An attempt to make the time series more stationary was made with a 24th order difference. AIC recommended an ARMA(10,0) model on the residuals. We then fit the original data with the recommended phi's with a seasonal component to make for an ARIMA(10,0,0); $s = 24$ model. As shown later in our write-up, the ARIMA model was not quite as effective as those without differencing or a seasonal adjustment. It will therefore be removed from further analysis and comparisons to other models.

- **ASE** 1.1158596×10^4
- **RMSE** 105.634

3.1.1 7 Day Forecast

forecast.2012.june.f	forecast.2012.june.ll	forecast.2012.june.ul
20.619552	-125.197859	166.4370
8.911618	-207.711423	225.5347
94.694990	-146.374034	335.7640
134.202311	-112.847671	381.2523
113.430174	-135.316297	362.1766
157.646842	-91.725498	407.0192
219.581047	-29.805245	468.9673
364.417908	114.959994	613.8758
541.616862	288.552028	794.6817
335.145181	80.782281	589.5081
213.852335	-40.533652	468.2383
186.045602	-68.508632	440.5998
187.474938	-67.649885	442.5998
156.752819	-99.253081	412.7587
145.344549	-111.380193	402.0693
178.127261	-78.890784	435.1453
279.721850	22.014834	537.4289
616.726511	358.326264	875.1268
630.378566	371.644524	889.1126
496.166968	237.150483	755.1835
404.480535	145.168894	663.7922
329.354454	69.892950	588.8160
269.212504	9.661879	528.7631
168.783976	-90.807099	428.3751
80.451287	-195.977744	356.8803
43.294727	-255.760098	342.3496
138.644931	-170.556056	447.8459
192.498563	-119.901585	504.8987
175.823818	-137.678401	489.3260
202.958462	-110.849294	516.7662

7 Day Forecast (ARMA)

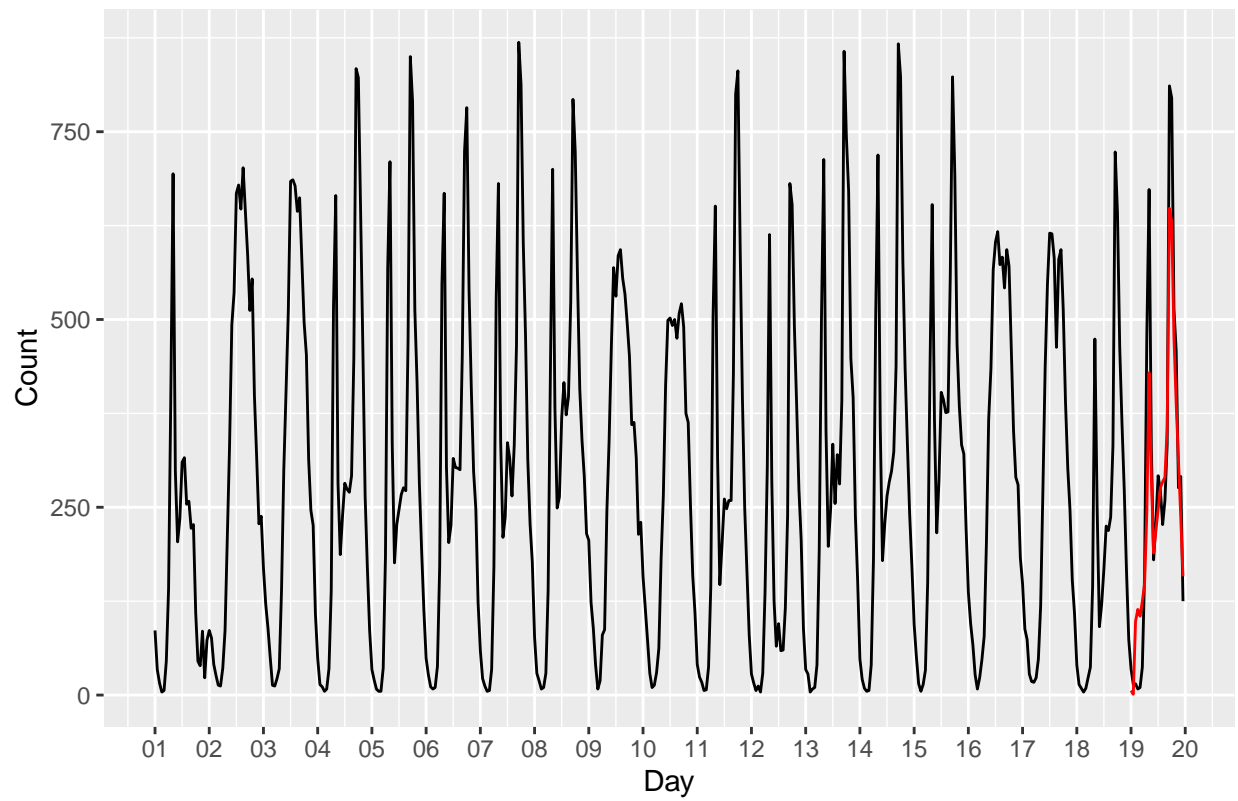


- **ASE** 2.9004653×10^4
- **RMSE** 170.308

3.1.2 1 Day Forecast

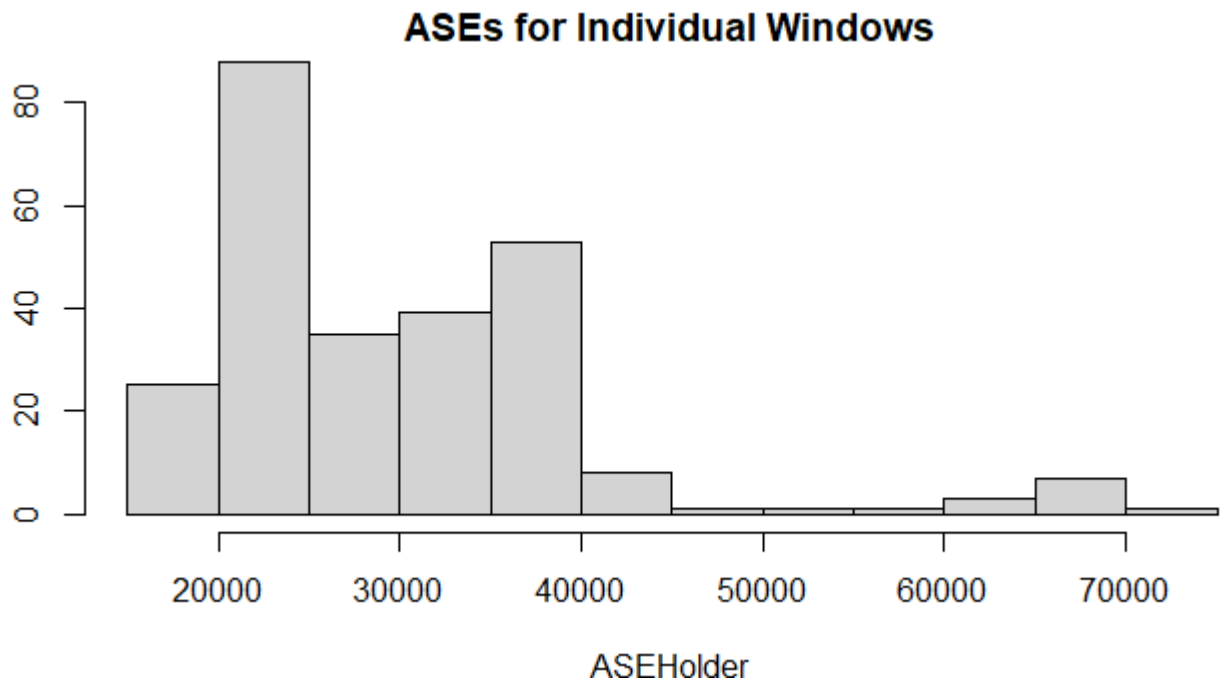
forecast.2012.june.f	forecast.2012.june.ll	forecast.2012.june.ul
5.971962	-139.845450	151.7894
-9.007633	-225.630674	207.6154
97.988929	-143.080095	339.0580
113.603486	-133.446495	360.6535
104.702879	-144.043592	353.4493
124.582433	-124.789908	373.9548
147.799162	-101.587129	397.1855
243.440802	-6.017112	492.8987
429.097035	176.032202	682.1619
289.925441	35.562540	544.2883
188.579045	-65.806942	442.9650
218.132166	-36.422068	472.6864
247.243190	-7.881633	502.3680
276.429765	20.423865	532.4357
284.112290	27.387547	540.8370
289.677293	32.659248	546.6953
356.850571	99.143556	614.5576
647.701032	389.300784	906.1013
630.764523	372.030482	889.4986
476.373727	217.357242	735.3902
386.779872	127.468232	646.0915
312.687851	53.226347	572.1494
237.363586	-22.187039	496.9142
158.248563	-101.342512	417.8396

1 Day Forecast (ARMA)



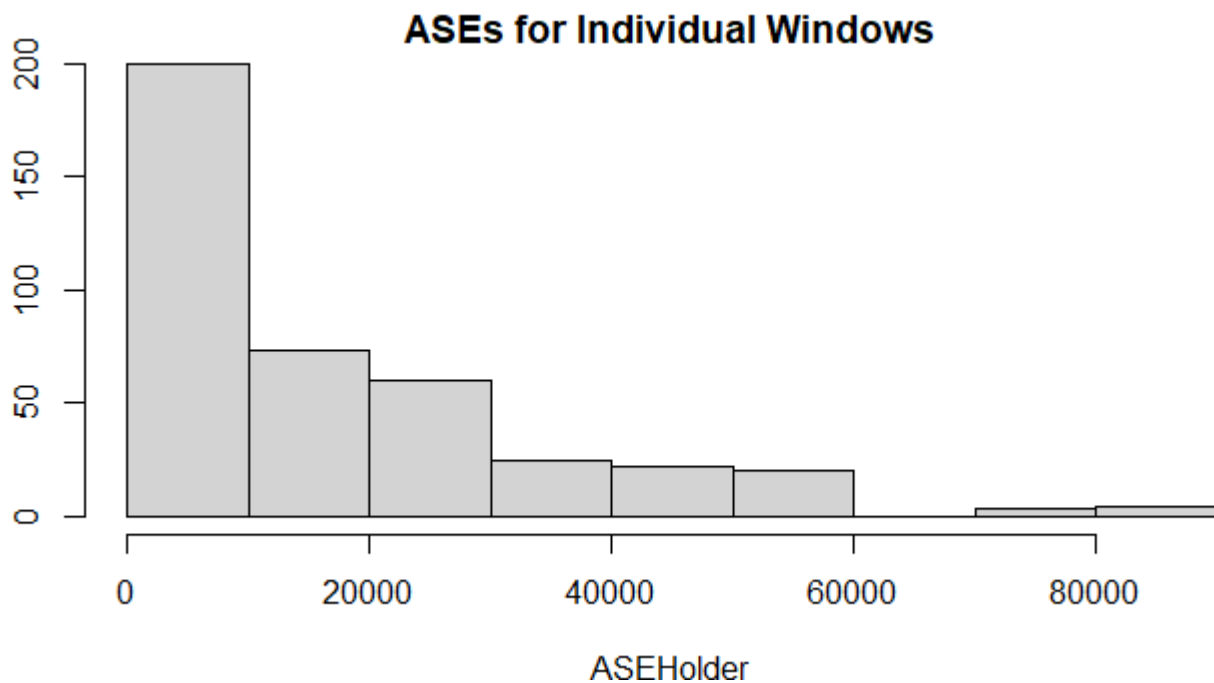
- **ASE** 9397.58
- **RMSE** 96.941

3.1.3 Long-Term Rolling Window ASE (7-Days)



- 30,053 (mean)
- 27,033 (median)

3.1.4 Short-Term Rolling Window ASE (1-Day)

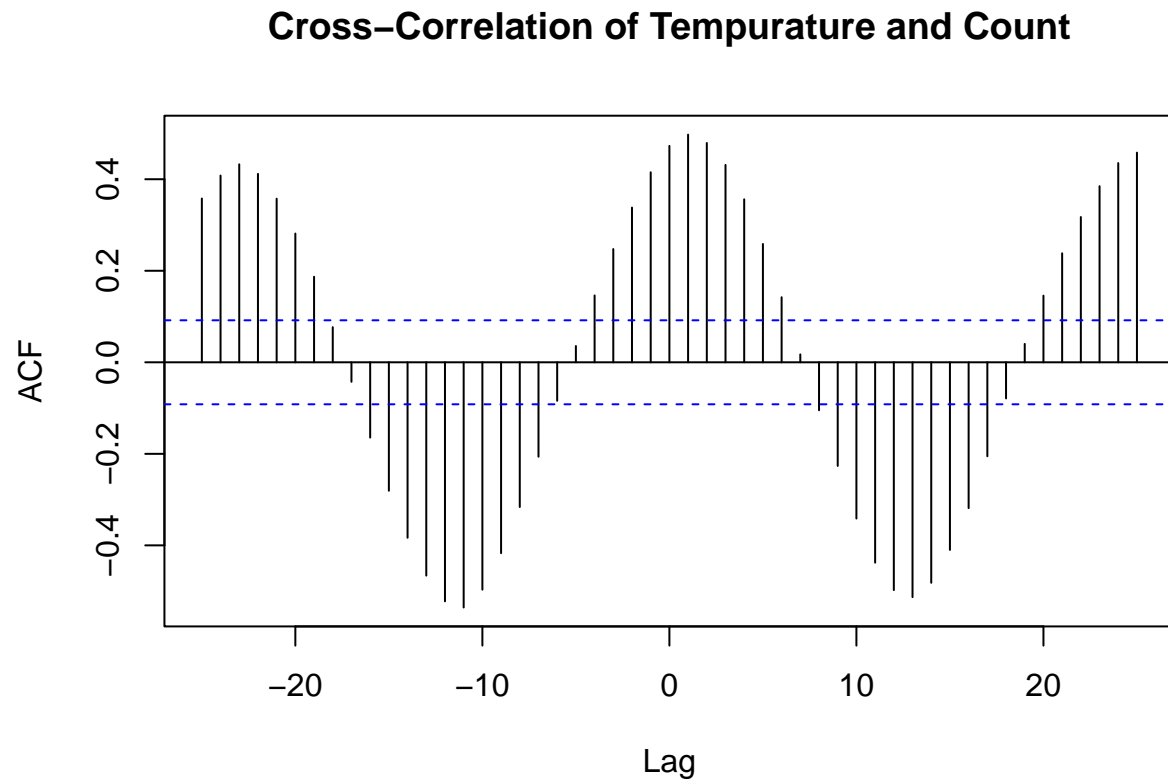


- 17,002 (mean)
- 10,648 (median)

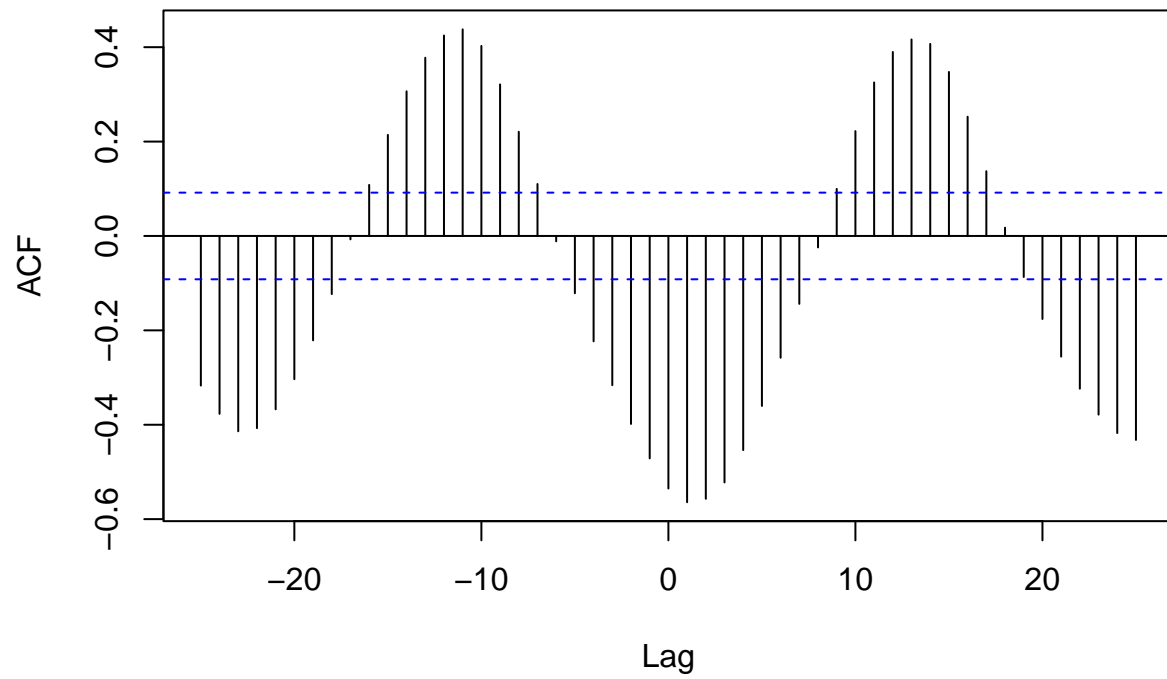
Note that the mean is much higher than the median, which would indicate a more skewed distribution as evidenced by the chart above.

3.2 Vector Auto-Regressive (VAR) Model

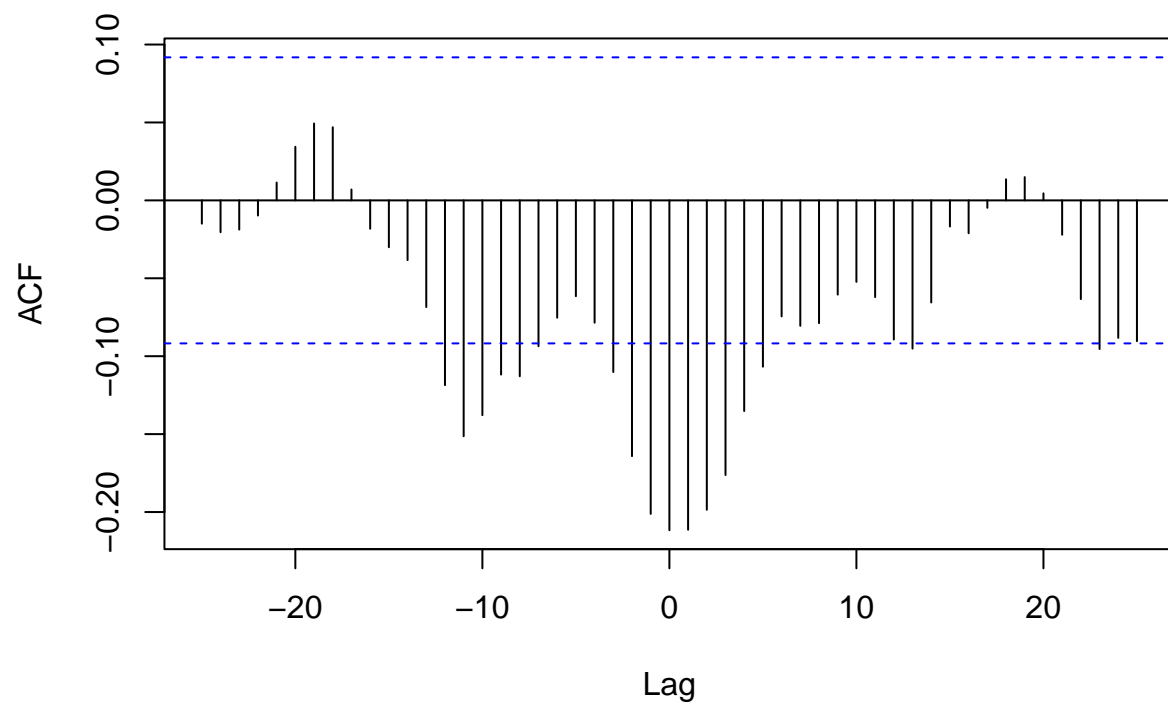
3.2.1 Cross-Correlation



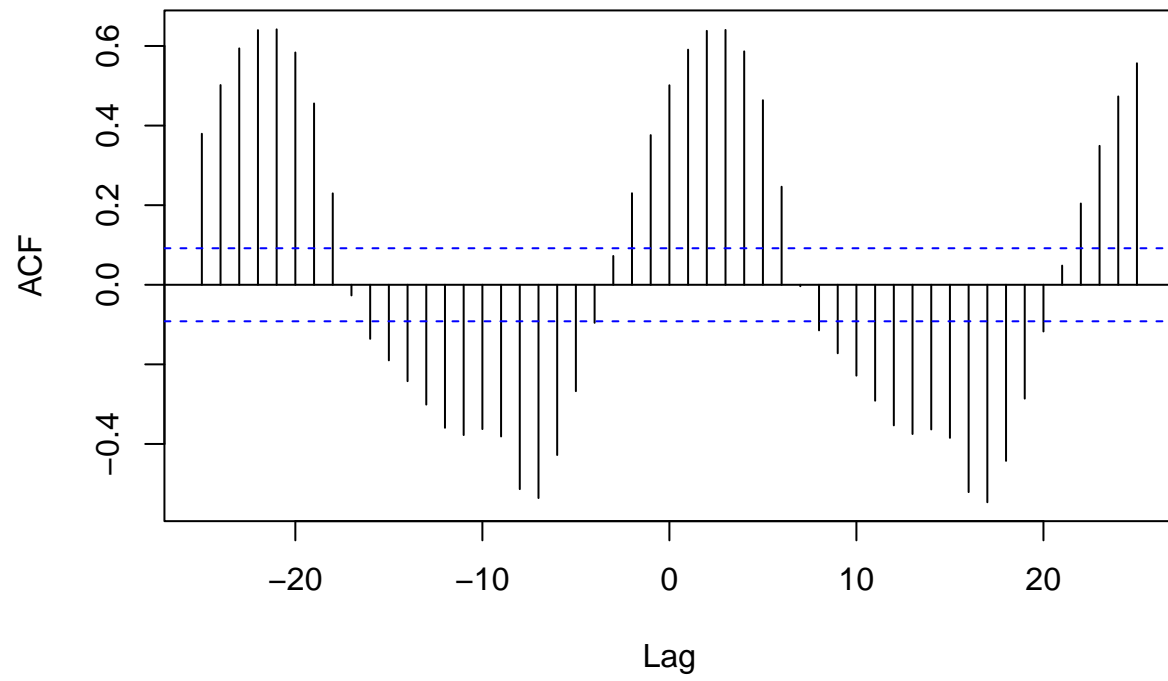
Cross-Correlation of Humidity and Count



Cross-Correlation of Weather and Count



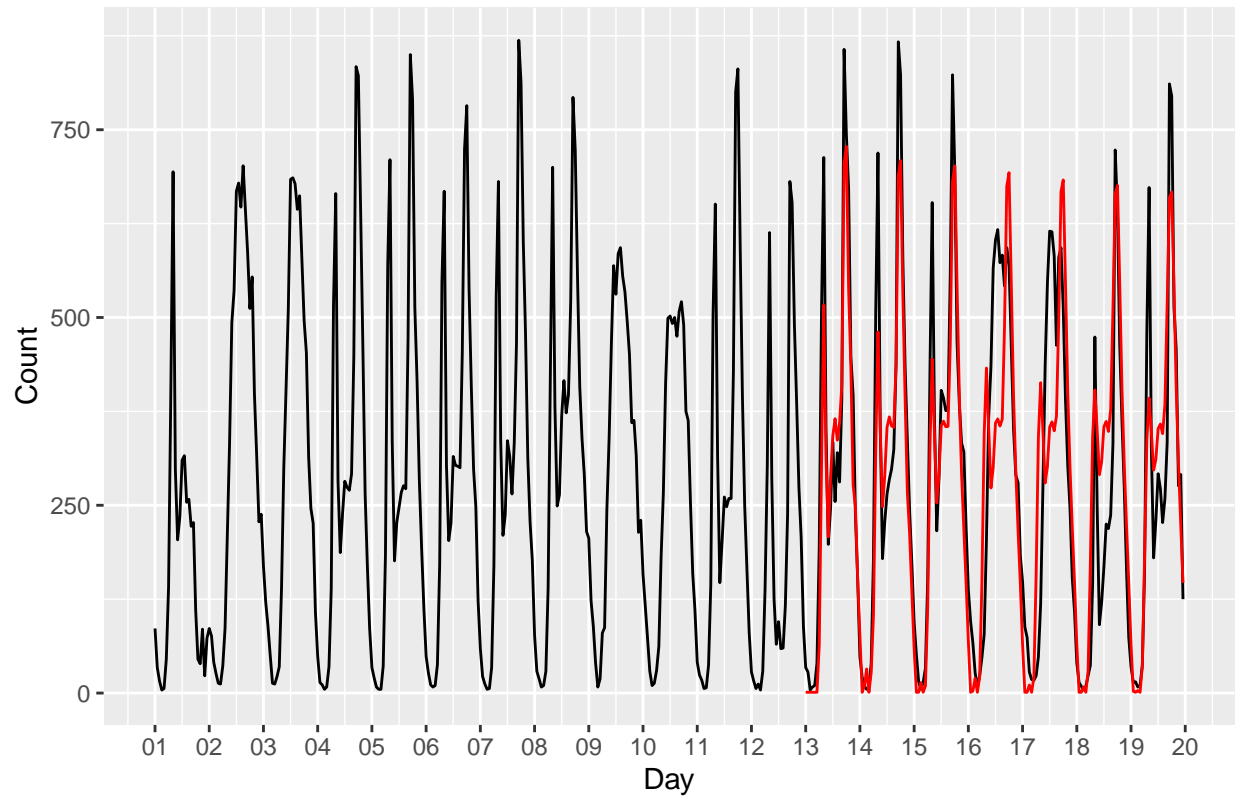
Cross-Correlation of Hour and Count



3.2.2 7 Day Forecast

fcst	lower	upper	CI
1.219709	-124.9655652	127.4050	126.1853
-33.905645	-204.1527835	136.3415	170.2471
-18.175382	-201.4361555	165.0854	183.2608
-9.434422	-196.6017650	177.7329	187.1673
-69.385457	-259.1037398	120.3328	189.7183
-9.284170	-201.9433057	183.3750	192.6591
63.956771	-132.2647452	260.1783	196.2215
300.352389	99.9509440	500.7538	200.4014
516.691458	307.2245262	726.1584	209.4669
314.638721	98.9642331	530.3132	215.6745
208.048713	-9.8363483	425.9338	217.8851
286.276141	63.8731667	508.6791	222.4030
342.136558	115.0404935	569.2326	227.0961
365.258262	134.7699355	595.7466	230.4883
336.655681	103.6149522	569.6964	233.0407
359.708093	124.6503420	594.7658	235.0578
410.119112	174.6359986	645.6022	235.4831
704.744074	467.7523316	941.7358	236.9917
727.698776	487.5583724	967.8392	240.1404
537.861510	293.5048840	782.2181	244.3566
434.419379	188.3045444	680.5342	246.1148
277.052371	29.4420989	524.6626	247.6103
248.003785	-0.0648516	496.0724	248.0686
154.417984	-94.5595826	403.3956	248.9776
76.672832	-174.8682067	328.2139	251.5410
-3.250914	-258.4093408	251.9075	255.1584
9.709669	-246.9340038	266.3533	256.6437
32.075866	-225.1360866	289.2878	257.2120
-24.870054	-282.2067737	232.4667	257.3367
27.732718	-229.7031103	285.1685	257.4358

7 Day Forecast (VAR)

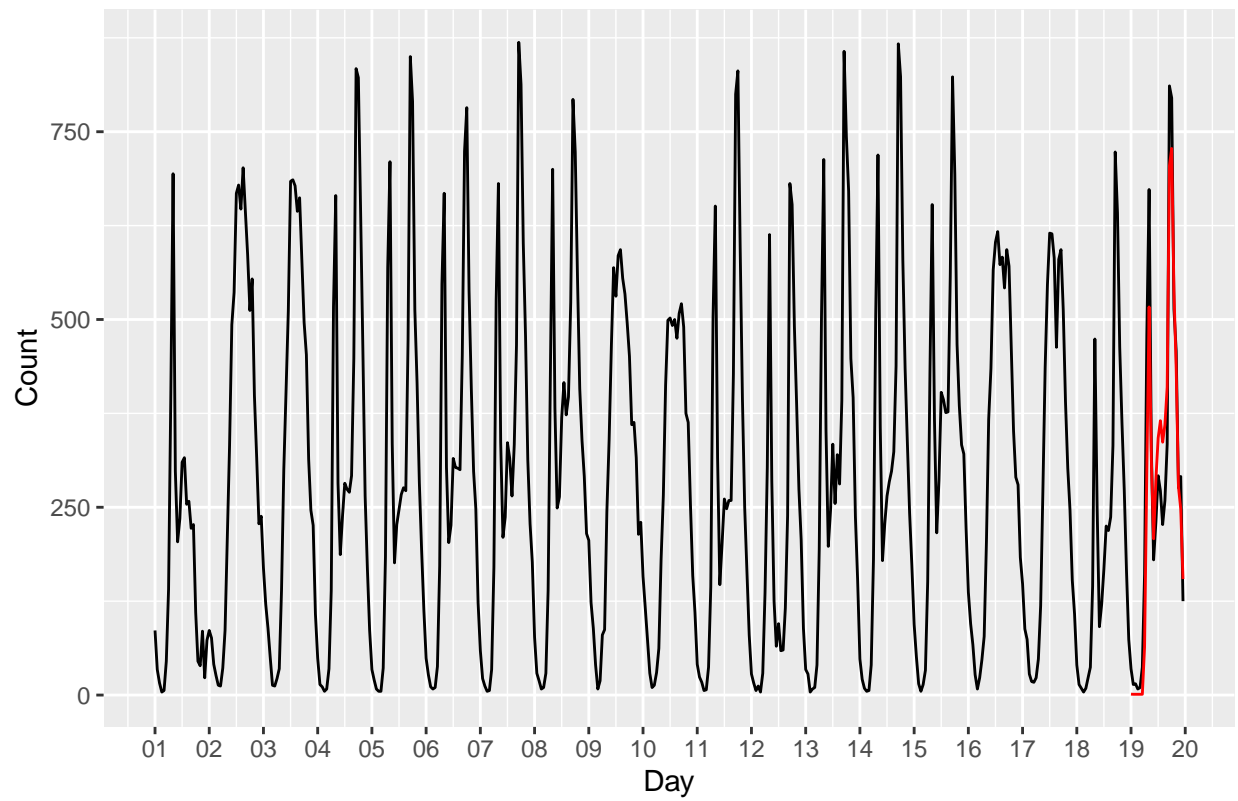


- **ASE** 1.25045×10^4
- **RMSE** 111.824

3.2.3 1 Day Forecast

fcst	lower	upper	CI
1.219709	-124.9655652	127.4050	126.1853
-33.905645	-204.1527835	136.3415	170.2471
-18.175382	-201.4361555	165.0854	183.2608
-9.434422	-196.6017650	177.7329	187.1673
-69.385457	-259.1037398	120.3328	189.7183
-9.284170	-201.9433057	183.3750	192.6591
63.956771	-132.2647452	260.1783	196.2215
300.352389	99.9509440	500.7538	200.4014
516.691458	307.2245262	726.1584	209.4669
314.638721	98.9642331	530.3132	215.6745
208.048713	-9.8363483	425.9338	217.8851
286.276141	63.8731667	508.6791	222.4030
342.136558	115.0404935	569.2326	227.0961
365.258262	134.7699355	595.7466	230.4883
336.655681	103.6149522	569.6964	233.0407
359.708093	124.6503420	594.7658	235.0578
410.119112	174.6359986	645.6022	235.4831
704.744074	467.7523316	941.7358	236.9917
727.698776	487.5583724	967.8392	240.1404
537.861510	293.5048840	782.2181	244.3566
434.419379	188.3045444	680.5342	246.1148
277.052371	29.4420989	524.6626	247.6103
248.003785	-0.0648516	496.0724	248.0686
154.417984	-94.5595826	403.3956	248.9776

1 Day Forecast (VAR)

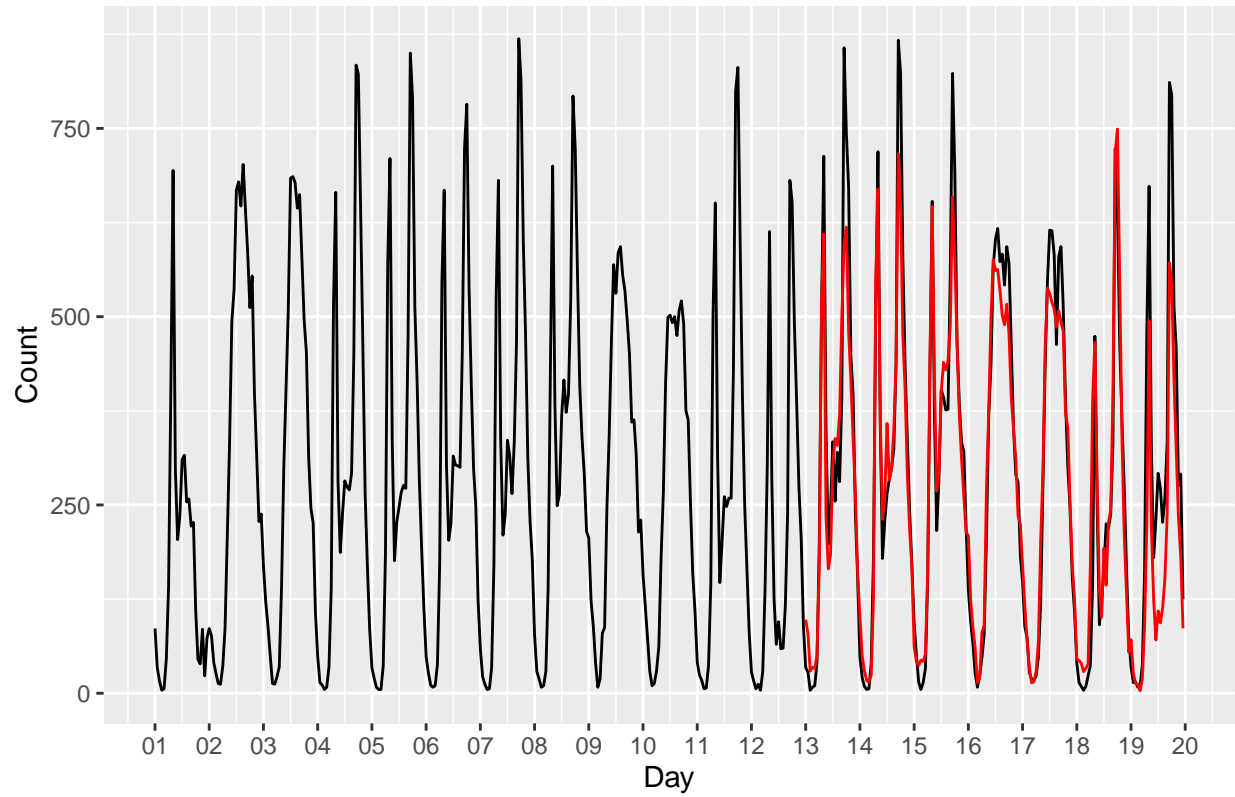


- **ASE** 5500.798
- **RMSE** 74.167

3.3 Neural Network Model

3.3.1 7 Day Forecast

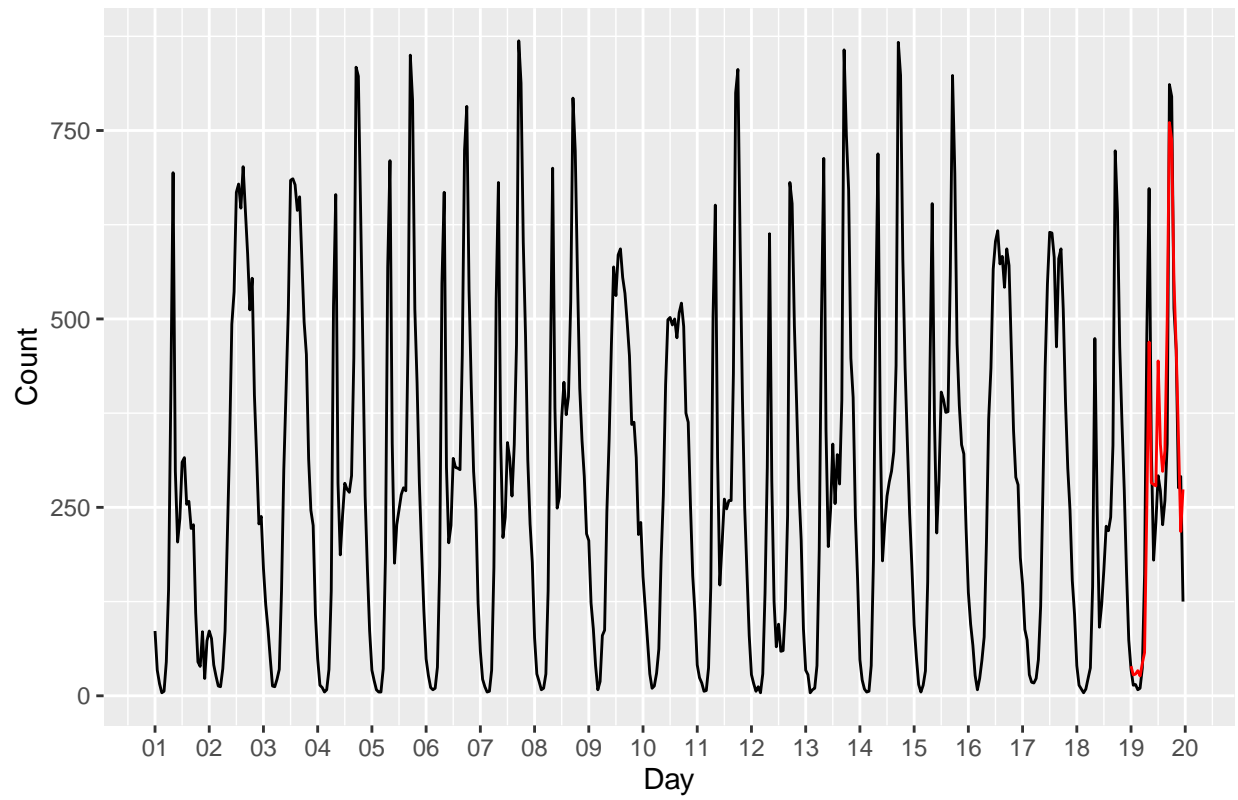
7 Day Forecast (MLP)



- **ASE** 5858.1
- **RMSE** 76.538

3.3.2 1 Day Forecast

1 Day Forecast (MLP)



- **ASE** 9280.075
- **RMSE** 96.333

3.4 Ensemble Model

3.4.1 7 Day Forecast

	Actual	Predicted (ARMA)	Predicted (VAR)	Predicted (MLP)	Predicted (Ensemble)
289	34	21	1	98	50
290	28	9	1	79	40
291	4	95	1	29	15
292	8	134	1	35	18
293	10	113	1	33	17
294	40	158	1	50	26
295	194	220	64	167	116
296	505	364	300	519	410
297	713	542	517	610	563
298	352	335	315	235	275
299	198	214	208	165	187
300	246	186	286	184	235
301	334	187	342	304	323
302	255	157	365	338	352
303	320	145	337	329	333
304	281	178	360	369	364
305	392	280	410	498	454
306	857	617	705	596	650
307	744	630	728	619	673
308	671	496	538	480	509
309	448	404	434	416	425
310	396	329	277	337	307
311	238	269	248	265	256
312	153	169	154	161	158
313	48	80	77	102	89
314	21	43	1	50	26
315	9	139	10	29	19
316	5	192	32	17	25
317	6	176	1	15	8
318	40	203	28	24	26

3.4.2 1 Day Forecast

	Actual	Predicted (ARMA)	Predicted (VAR)	Predicted (MLP)	Predicted (Ensemble)
433	35	6	1	39	20
434	14	1	1	28	14
435	15	98	1	29	15
436	8	114	1	33	17
437	10	105	1	26	14
438	37	125	1	43	22
439	161	148	64	57	61
440	480	243	300	251	275
441	673	429	517	469	493
442	328	290	315	283	299
443	180	189	208	280	244
444	230	218	286	279	282
445	292	247	342	444	393
446	272	276	365	334	349
447	227	284	337	298	317
448	259	290	360	325	342
449	334	357	410	497	454
450	811	648	705	761	733
451	795	631	728	738	733
452	514	476	538	551	544
453	458	387	434	461	448
454	276	313	277	342	310
455	291	237	248	218	233
456	125	158	154	274	214

4 Results

4.1 7 Day Forecast

4.1.1 ASE

Model	ASE
ARMA	29009.292
VAR	12496.690
MLP	5856.042
Ensemble	6067.696

4.1.2 RMSE

Model	RMSE
ARMA	170.32114
VAR	111.78860
MLP	76.52478
Ensemble	77.89542

Here, we include the ensemble model error to compare. Our ensemble is built out of a rounded average between VAR and MLP. It turns out that for the 7-day, “long-term”, forecast the MLP performs best. Followed by the ensemble, var and then finally arma. Var likely performs better because it considers other data that happened at the same time compared to a strict ARMA(25,1). Similarly, the MLP model may perform well because of its ability to use multiple layers to understand how all the time series work together, with lag in mind.

4.2 1 Day Forecast

4.2.1 ASE

Model	ASE
ARMA	9397.958
VAR	5498.958
MLP	9272.958
Ensemble	6729.208

4.2.2 RMSE

Model	RMSE
ARMA	96.94307
VAR	74.15496
MLP	96.29620
Ensemble	82.03175

VAR was the preferred method for short term forecasts. But note that the first six observations all have 1s as these were imputed. VAR had a tendency to predict negative values early on, which wouldn't make any sense given that our hourly counts should never be zero or less. And it turns out that there were not any observations with fewer than one.

5 Appendix

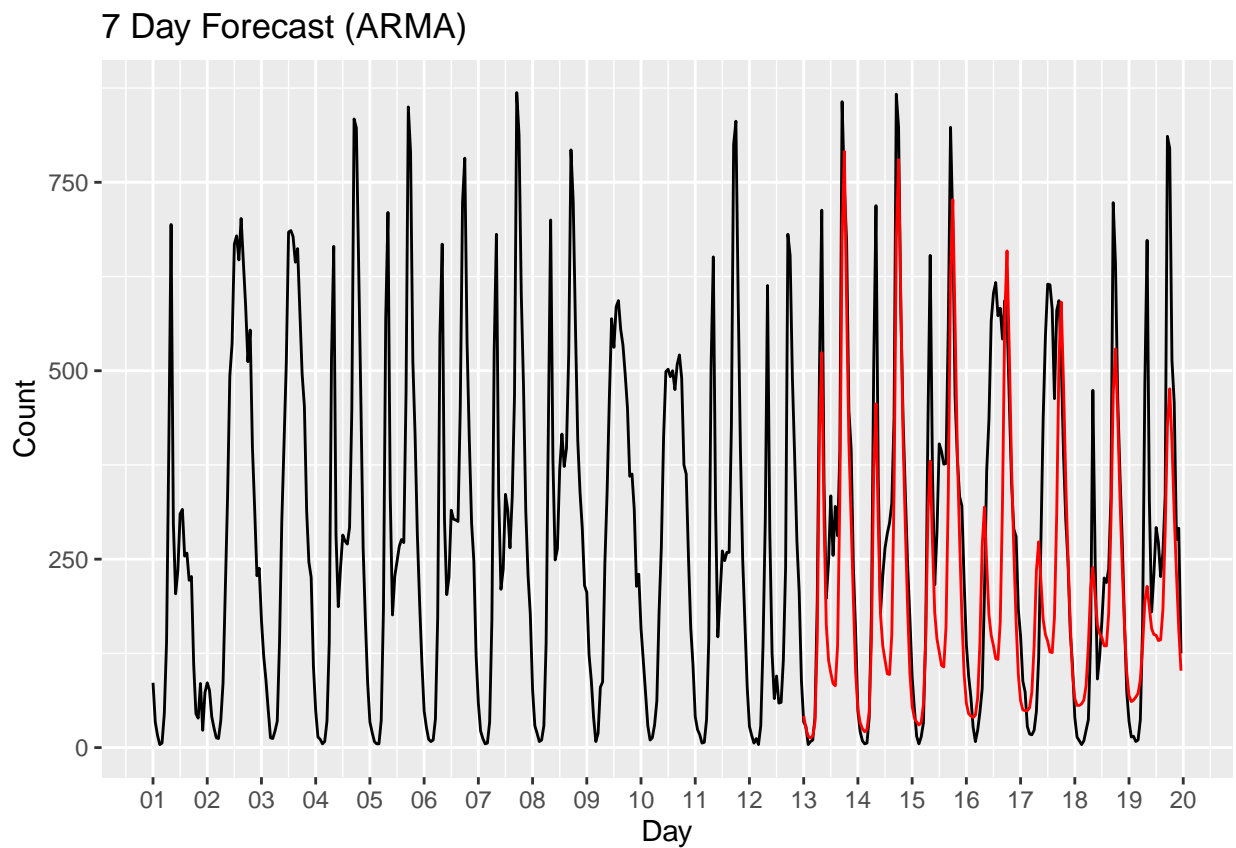
5.1 Solution One

The primary response variable, count, for the Bike Sales Demand data looks skewed. What kind of transformation is ideal for this time series realization?

Find and compare the ASEs for both the ARMA and the VAR models as they apply to the bike data. Which model performs better on the long-term forecast (7 days) and the short-term forecast (1 day)? (Hint: This is hourly data)

Apply a log transform to the primary response variable, count.

5.1.1 7 Day Forecast

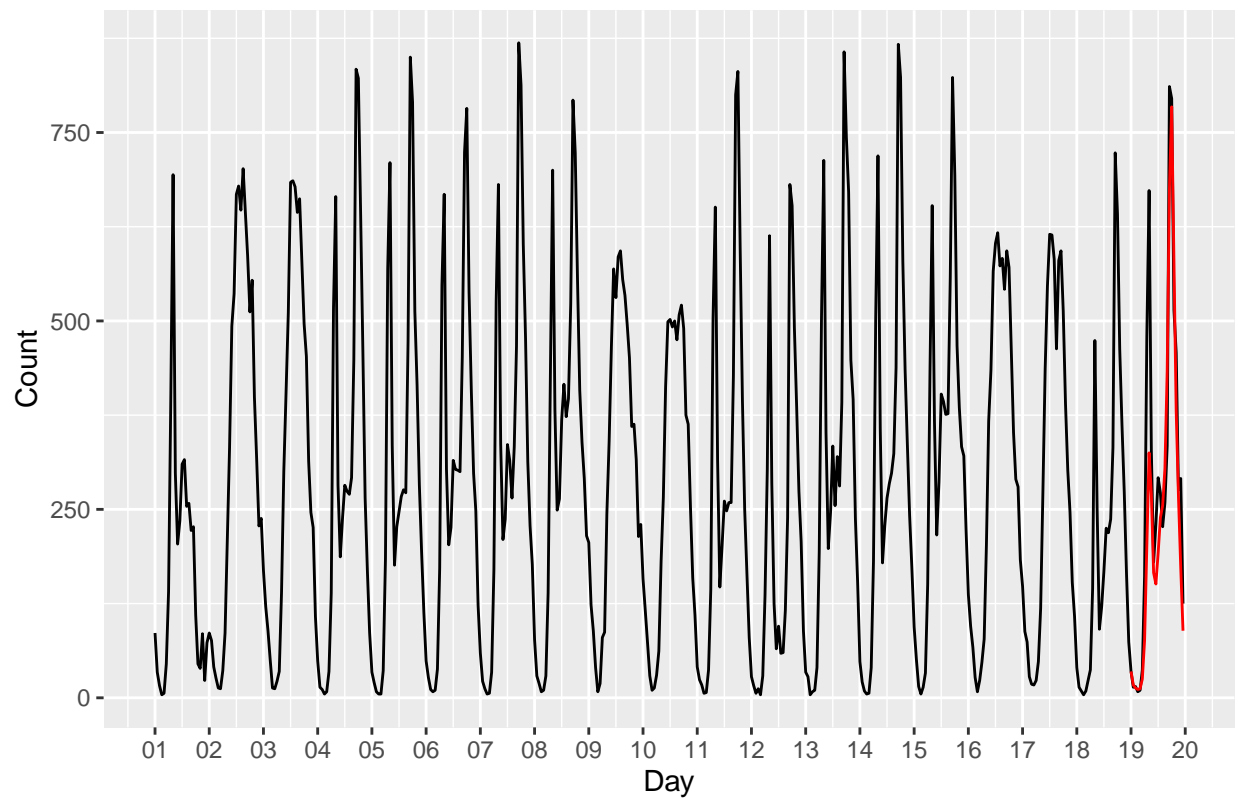


Note that the p and q estimates were made from the logged count, and AIC picked and ARMA(25,2) as opposed to an ARMA(25,1) as had been previously used.

- **ASE** 2.6306311×10^4
- **RMSE** 162.192

5.1.2 1 Day Forecast

1 Day Forecast (ARMA)



- **ASE** 1.1827925×10^4
- **RMSE** 108.756

5.2 Solution Two

Can additional data from previous month(s) aid in lowering the ASE and RMSE? We will focus on using an MLP neural network with additional input features to test our hypothesis when forecasting the short-term counts for the month of June, 2012.

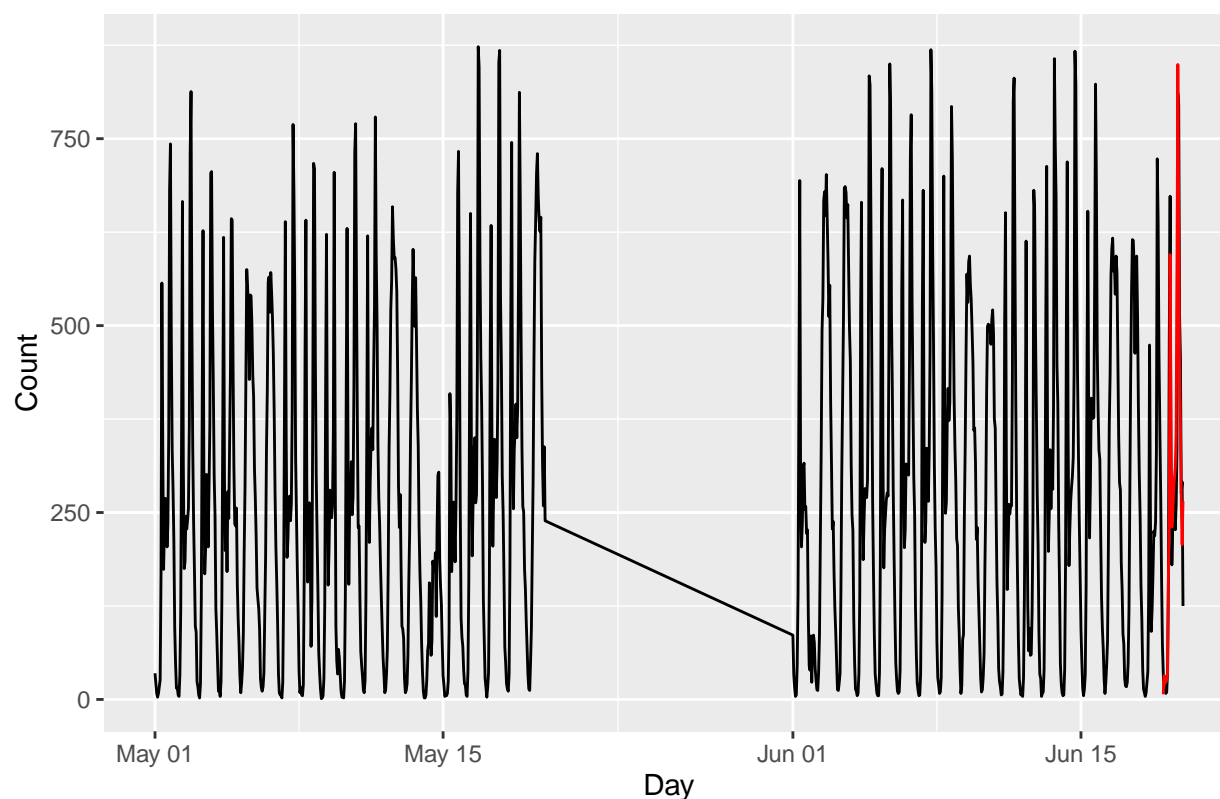
Including training data from May seemed to indicate that you could, albeit marginally as shown below, improve over the previous ARMA, VAR and MLP metrics that only used training data from June 2012.

However, attempting to use observations from more than 2 months out was detrimental to the model's performance.

Due to our data not being continuous in the sense that we only have bike rental counts for the first 19 days of each month, we cannot make use of all of the available training data easily.

Although not a very likely example to be encountered in a real world setting, it does offer some evidence that having more available data to train your model with can offer benefits. But having a continuous time-series and more observations would have been preferred.

1 Day Forecast (MLP)



Our MLP model with both May and June training data resulted in an ASE score of 8685.425.

And the RMSE stood at 93.196. Again, these are both improvements over previous short-term model forecasts.