Bike Share Demand Forecasting Methods

Chance Robinson and Kebur Fantahun

11/28/2021

Contents

1 Introduction

Bike rentals where the customer can pick up and drop off a bike at their leisure at several locations has become popular. This dataset outlines attributes related to the travel of customers. Data gathered by the rental companies includes things like the date, temperature, count of users, humidity and more. The collection of attributes has the potential to assist researchers in developing an understanding of the mobility in a city.

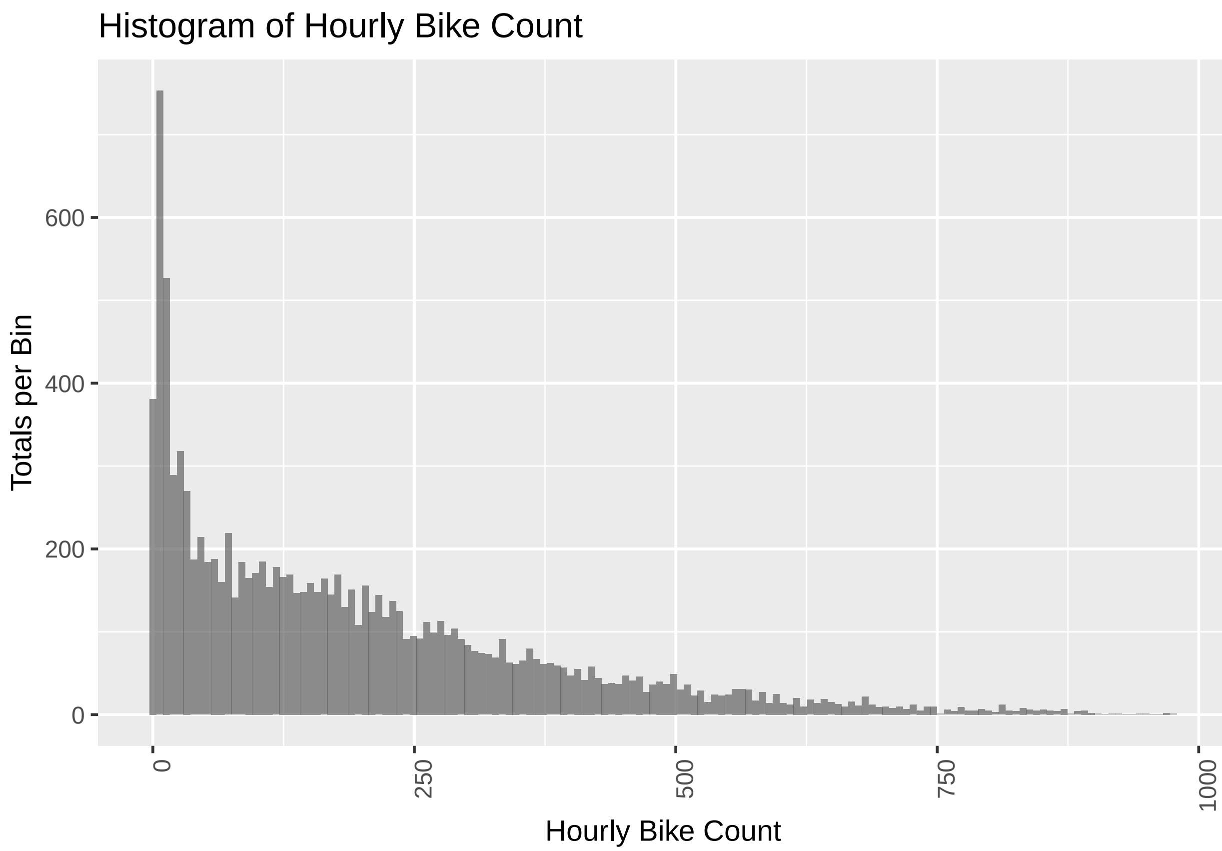
2 Data Analysis

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

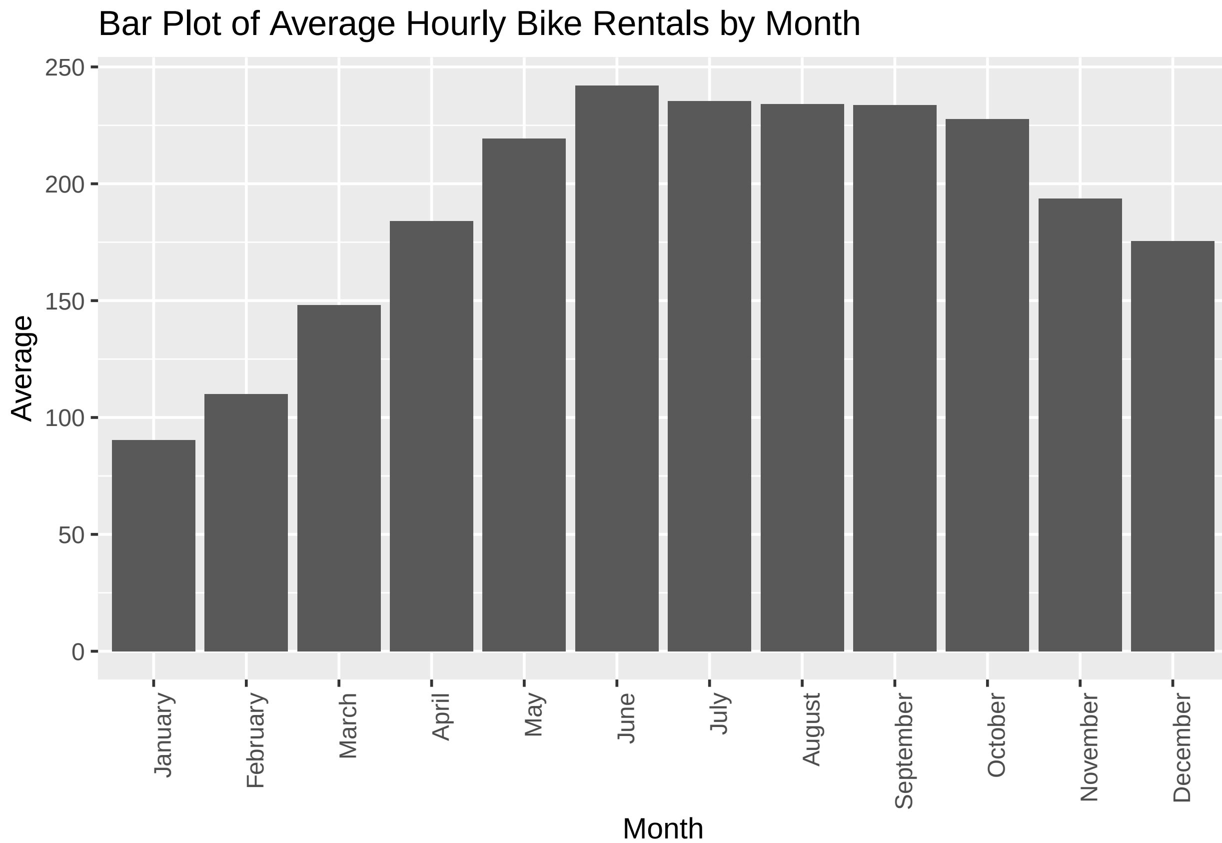
There are 12 columns provided in the `train.csv` data set with 10,886 observations. The `test.csv` data set has 6,493 records, or roughly 37% of the overall combined samples from the two files. This is due to the fact that the training data consist of the first 19 days of each month and the test the 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 (*primary response variable*)

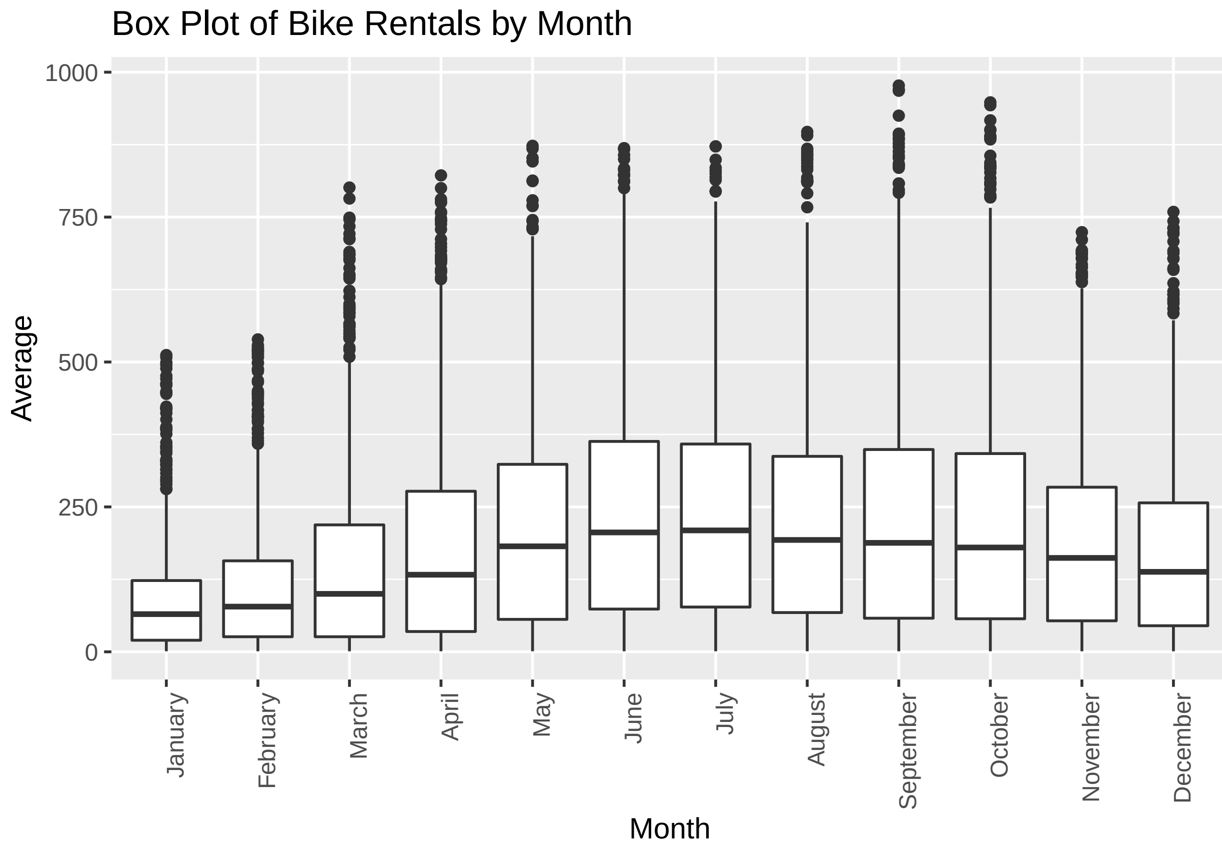
Var1	Freq
Min.	1.0000
1st Qu.	42.0000
Median	145.0000
Mean	191.5741
3rd Qu.	284.0000
Max.	977.0000



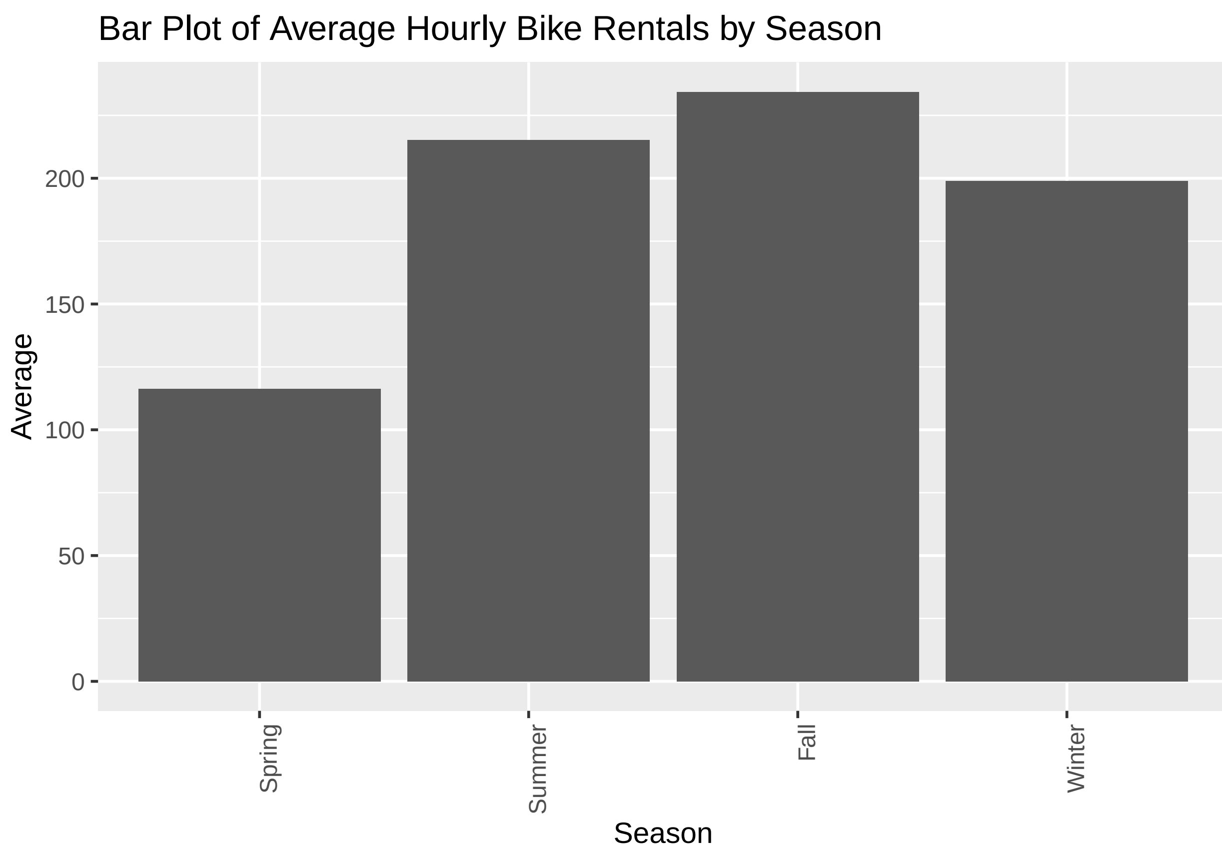
The response variable of `count` appears to be heavily right-skewed, with the median value at ~145 users. Additional summary statistics are show below.



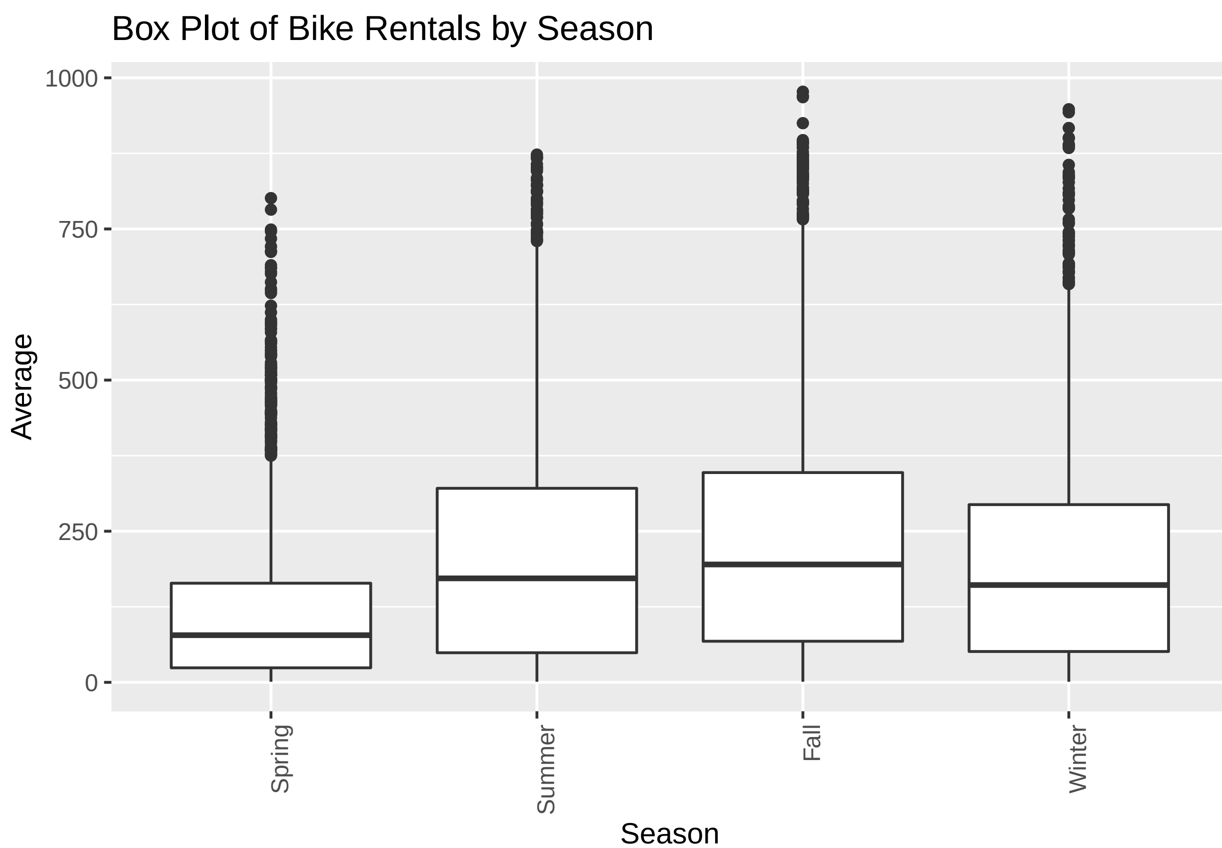
June appears to be the month with heaviest demand.

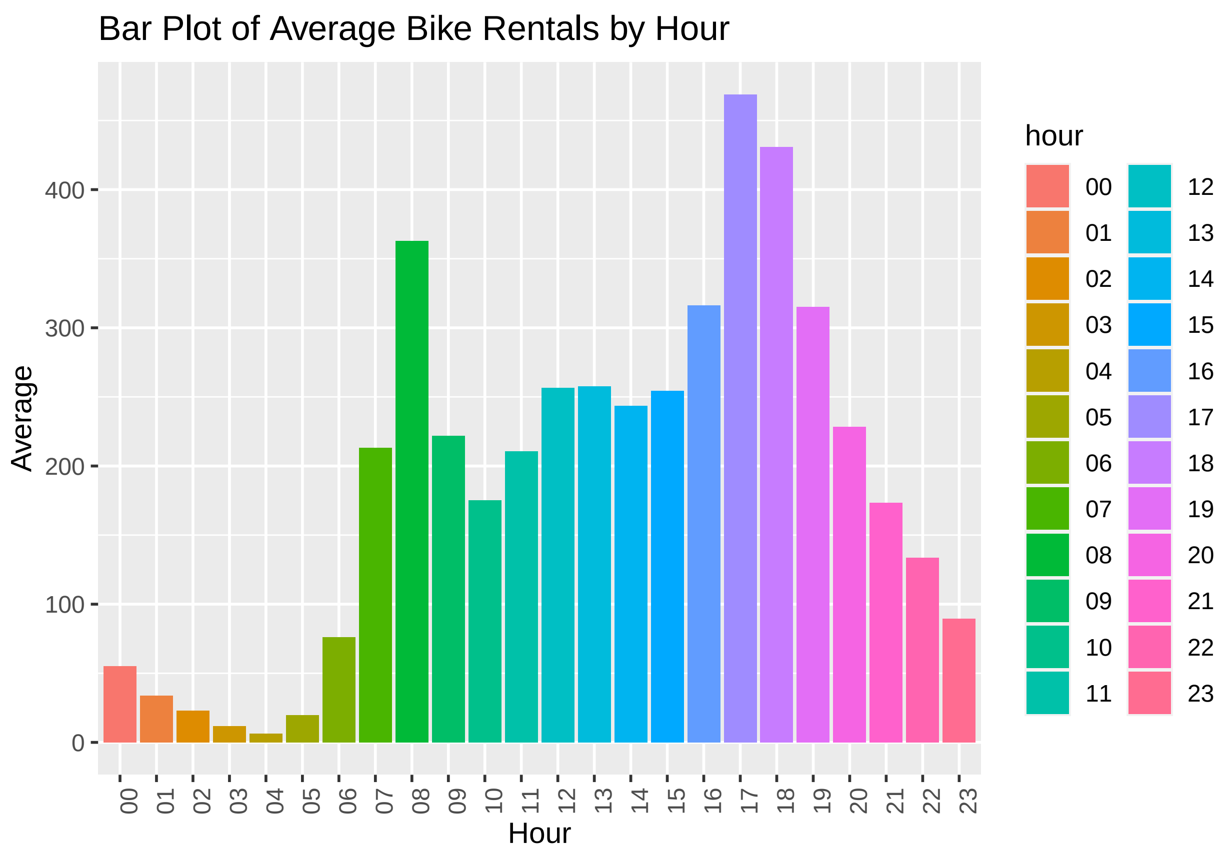


It also appears to show relatively few outliers compared to months like January.

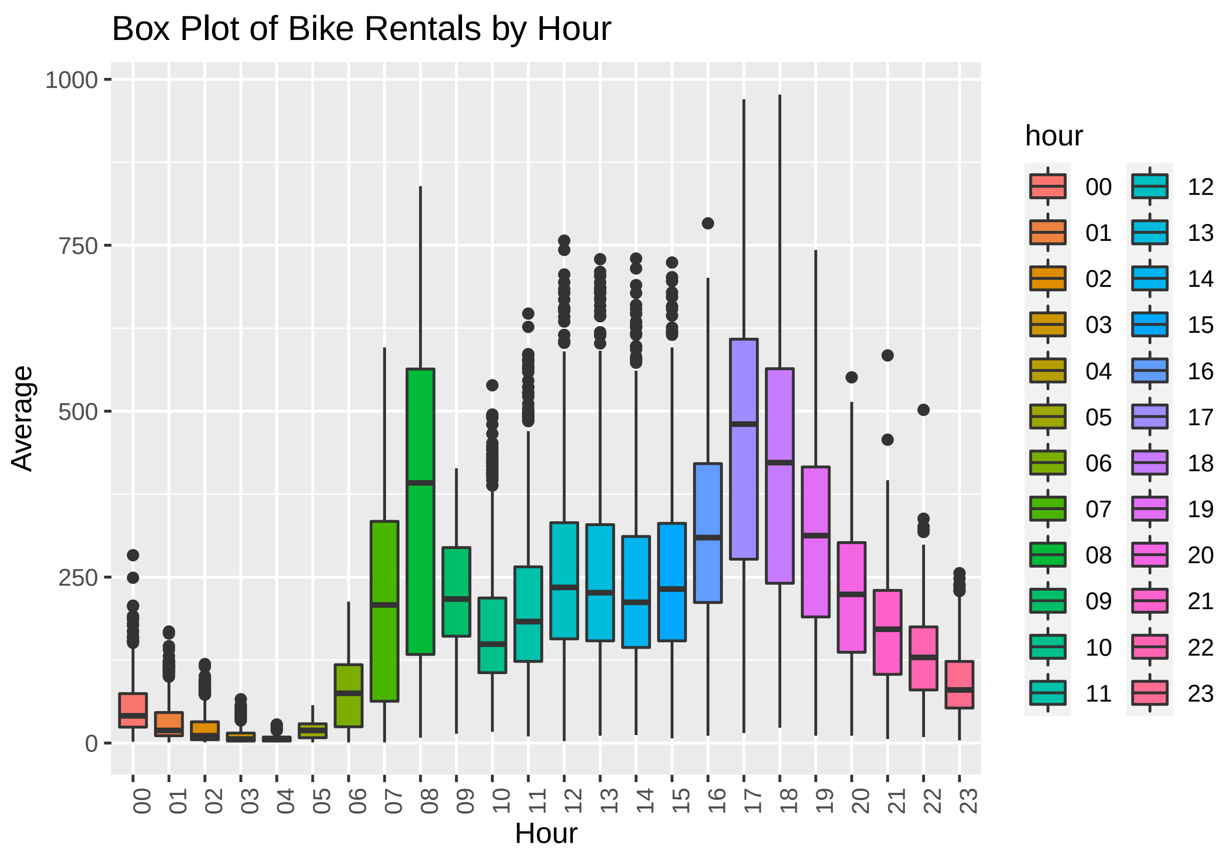


Fall looks to have more rentals on average than the other seasons.

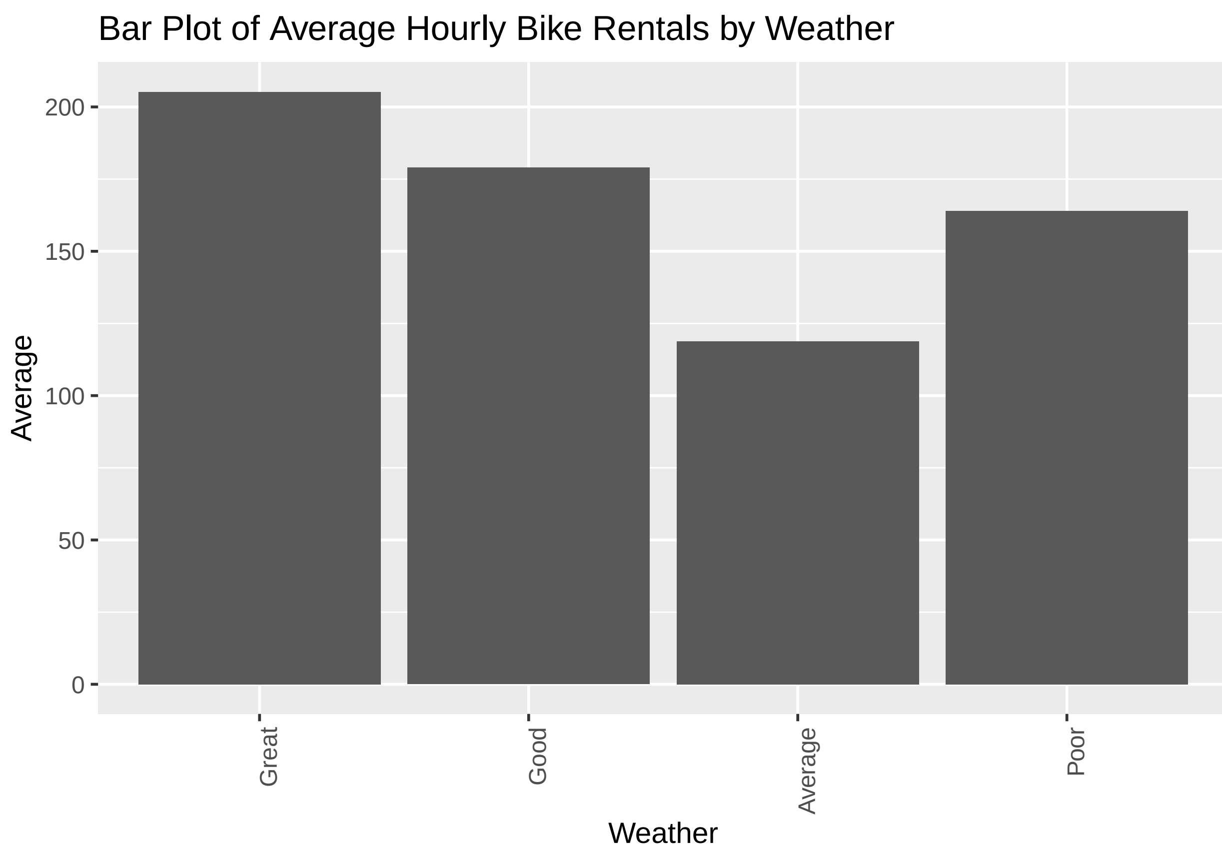

Box Plot of Bike Rentals by Season
Average
1000
750
500
250
0
Spring
Summer
Fall
Winter
Season



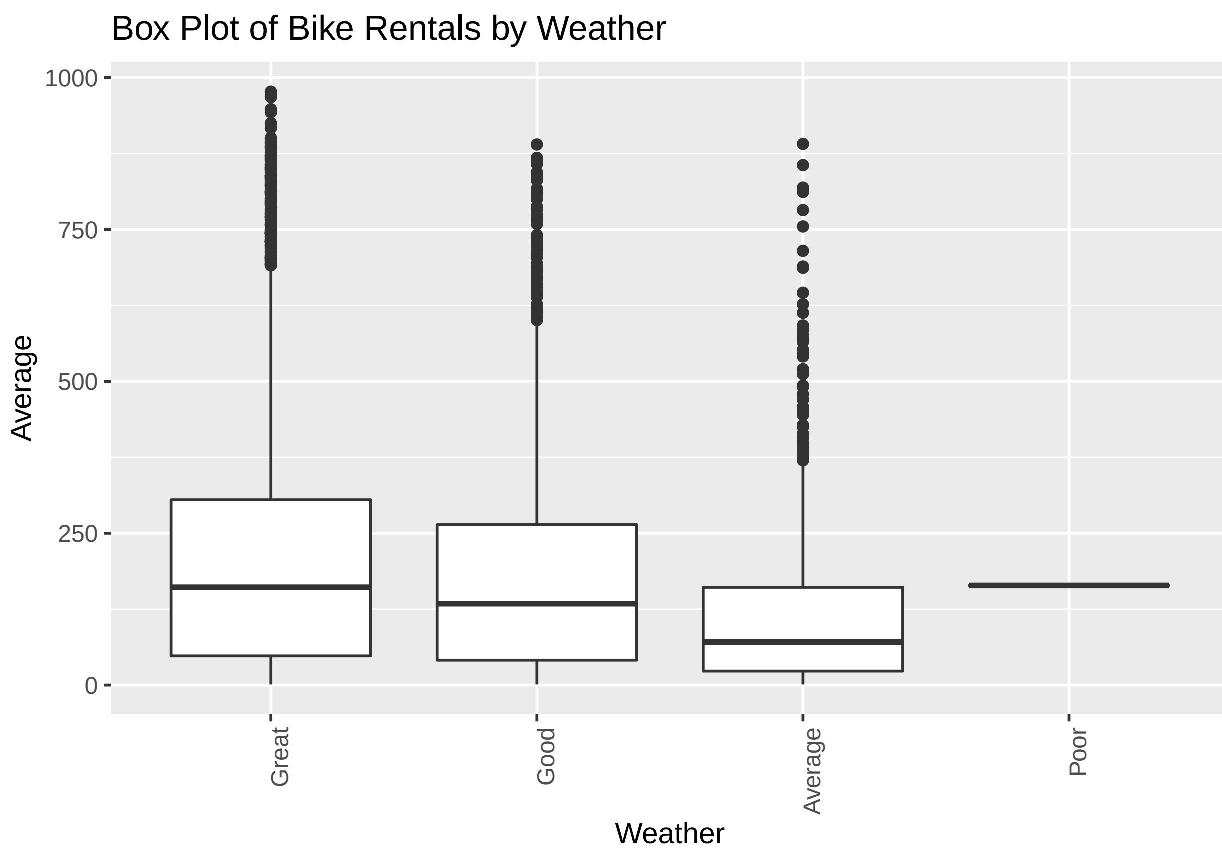
The 5pm hour clearly has the highest peak compared to the other hours of the day.



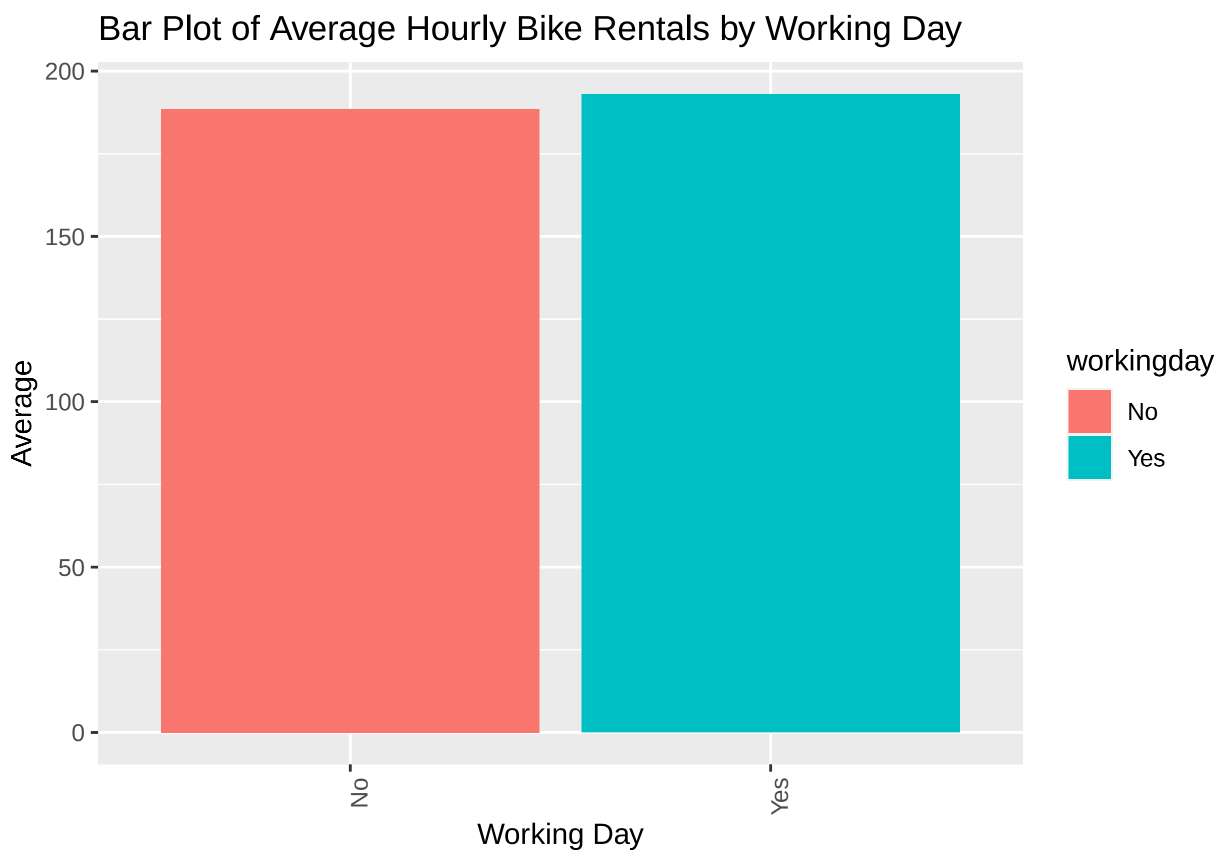
Box Plot of Bike Rentals by Hour
Average
1000
750
500
250
0
00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour
hour
00 01 02 03 04 05 06 07 08 09 10 11
12 13 14 15 16 17 18 19 20 21 22 23



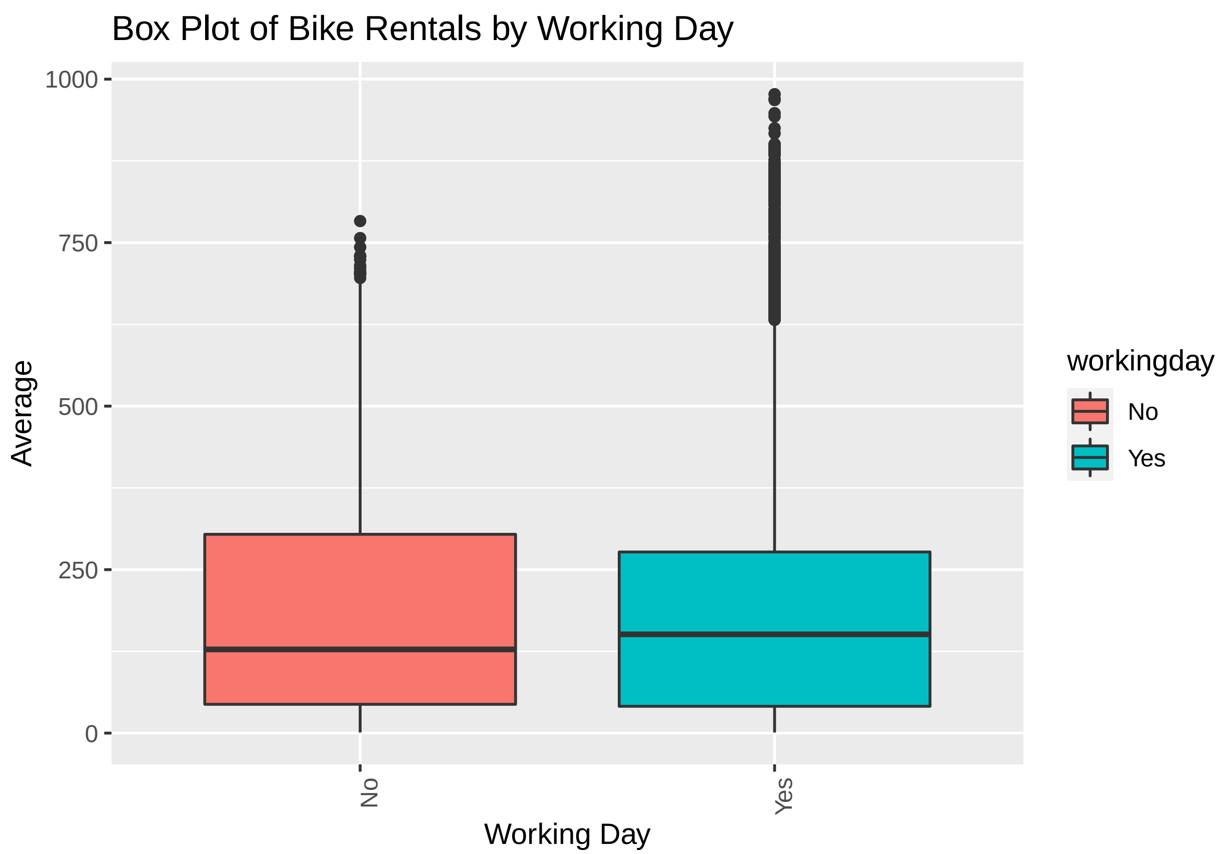
As expected, more riders are out when the weather is great, or better than average.



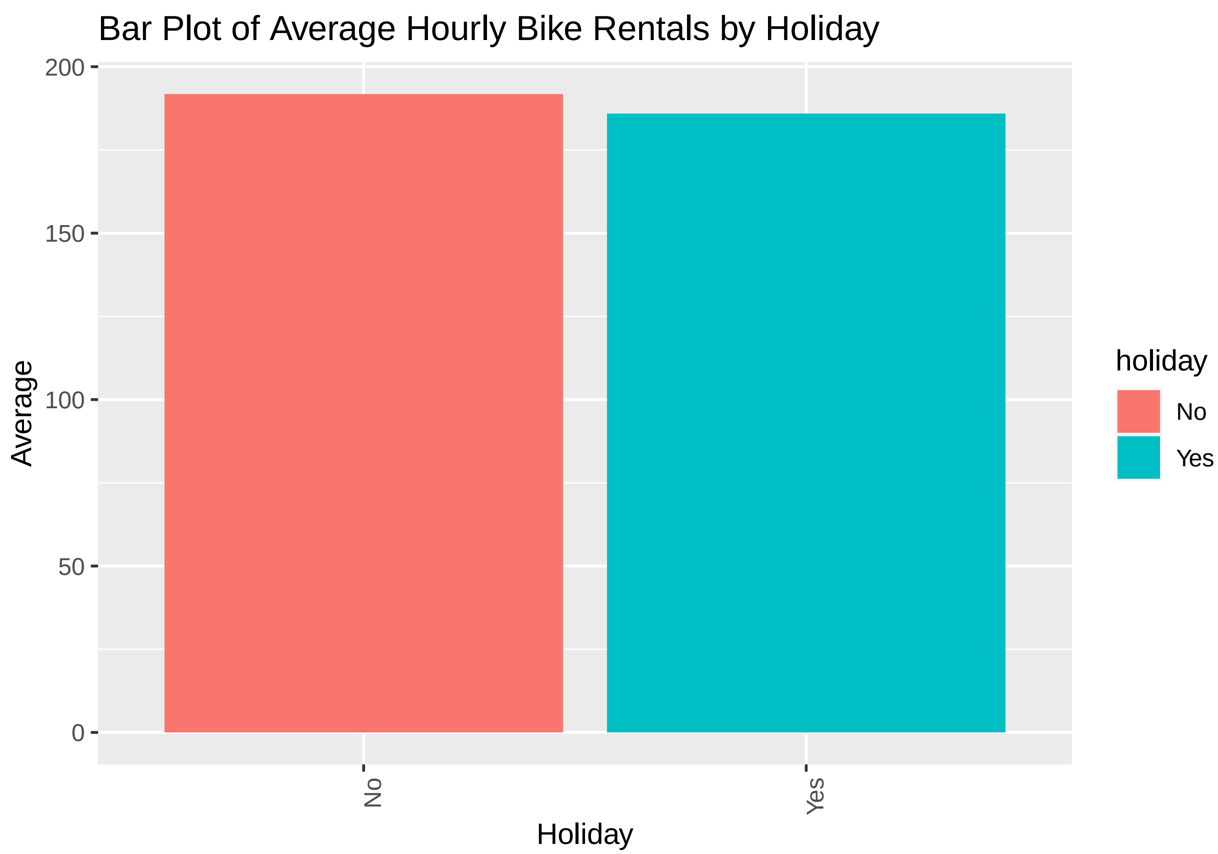
Note the small spread of riders when the weather was listed as poor, even though the averages were consistent with other categories.



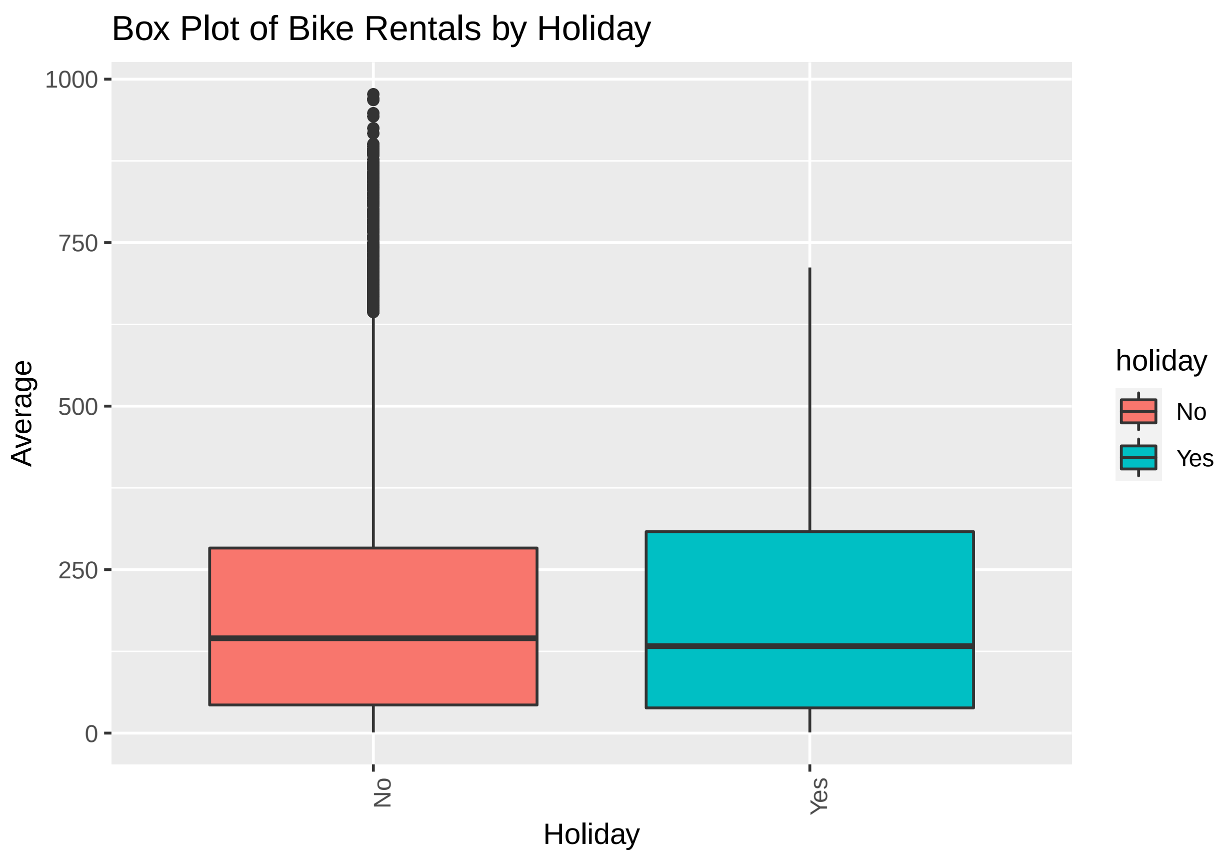
Surprisingly, whether or not the day was on a working day or not had little affect on the mean or median.



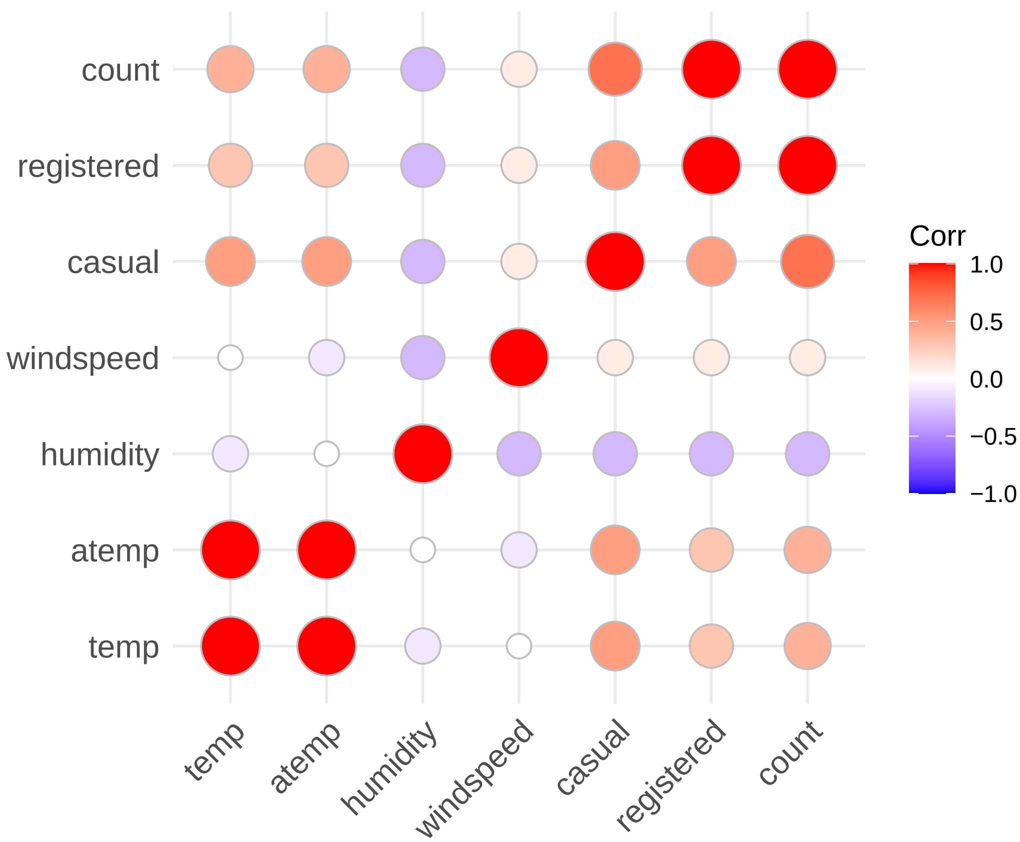
Box Plot of Bike Rentals by Working Day
1000
750
500
250
0
Average
No
Yes
Working Day
workingday
No
Yes



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

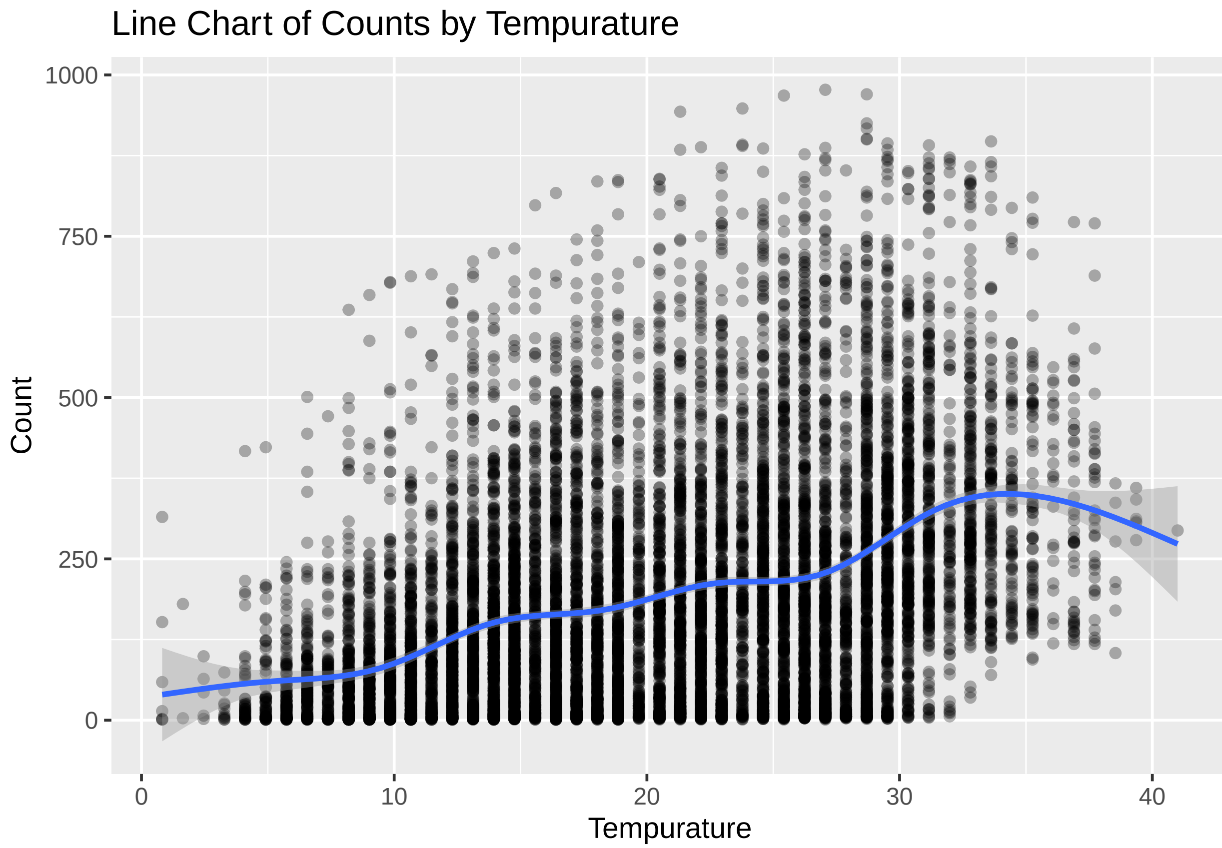

Box Plot of Bike Rentals by Holiday
1000
750
500
250
0
Average
No
Yes
Holiday
holiday
No
Yes

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

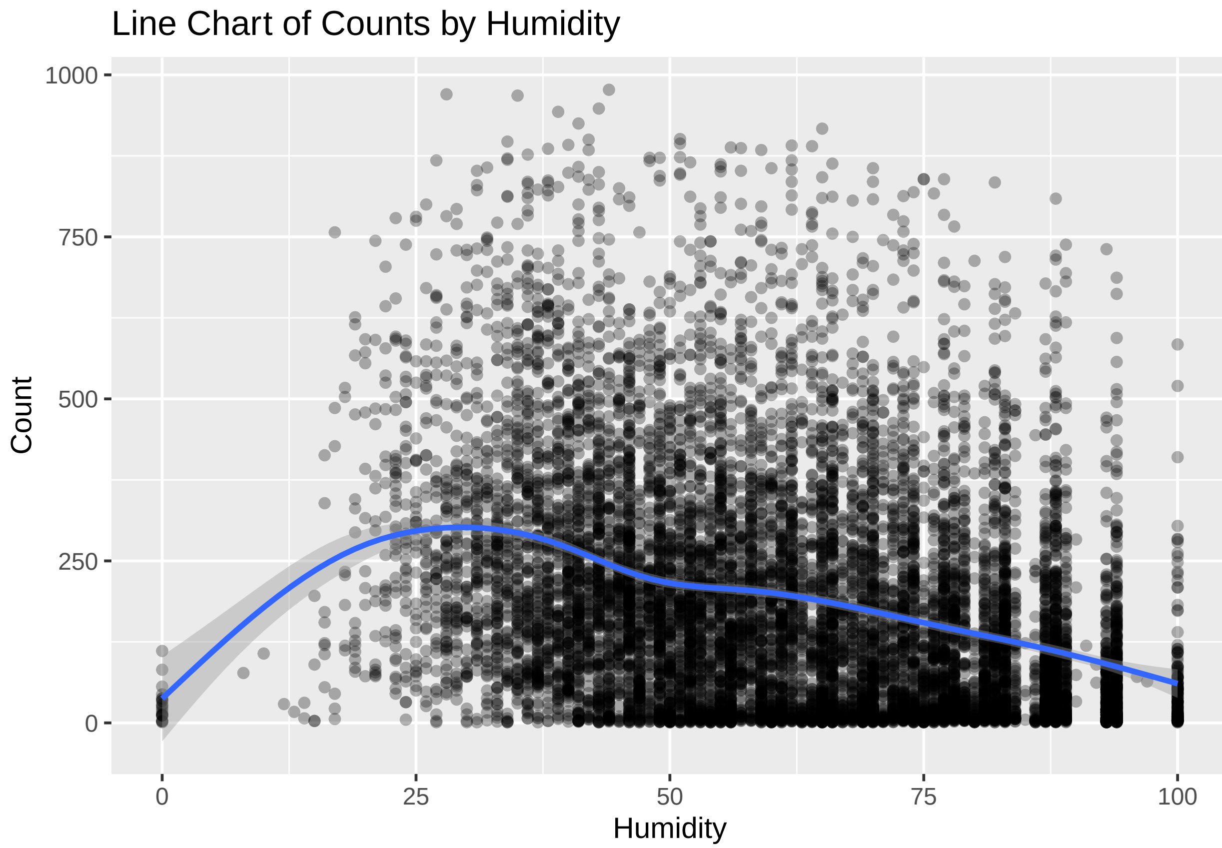


The plot above shows a strong correlation between casual, registered and total (count) users. We will use count as the primary response variable for our modeling, and discard the other secondary response variables.

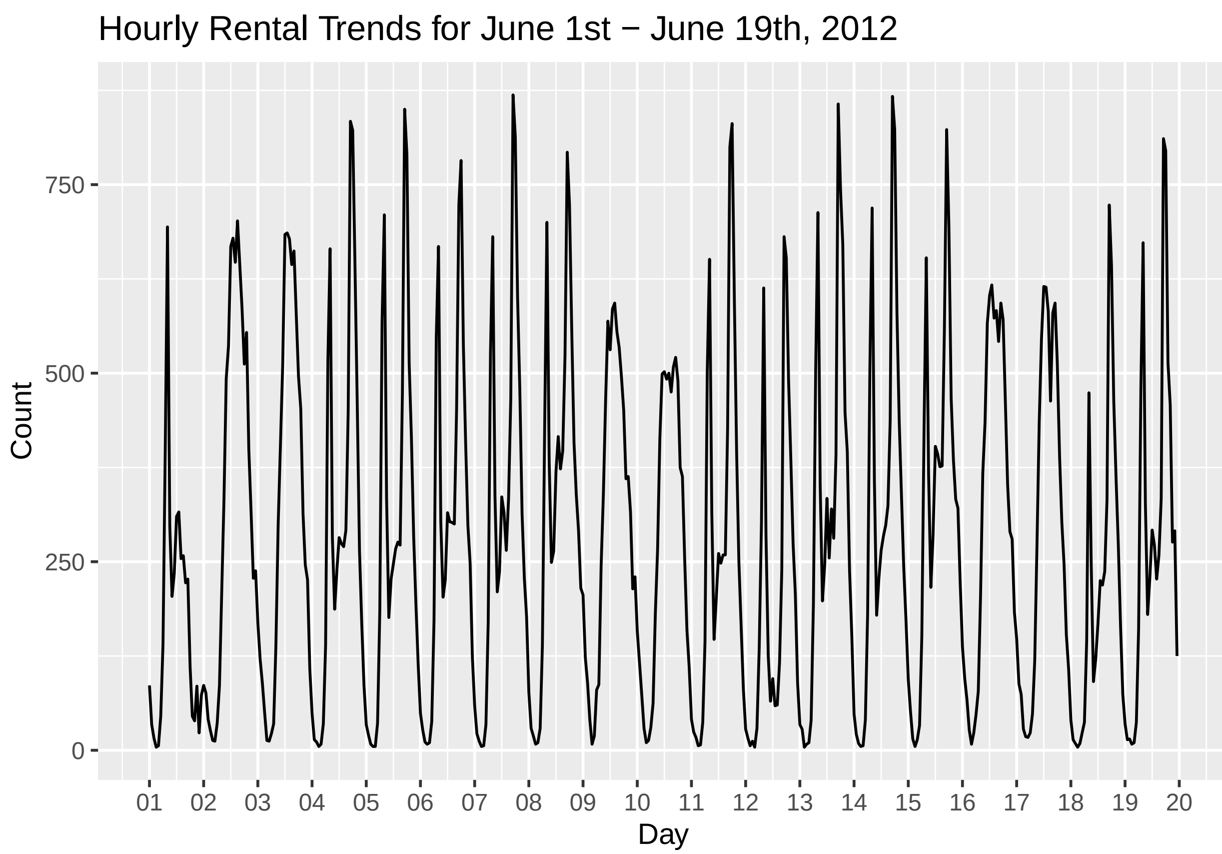
Note that temp and atemp are also closely related to one another as would be expected.



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)



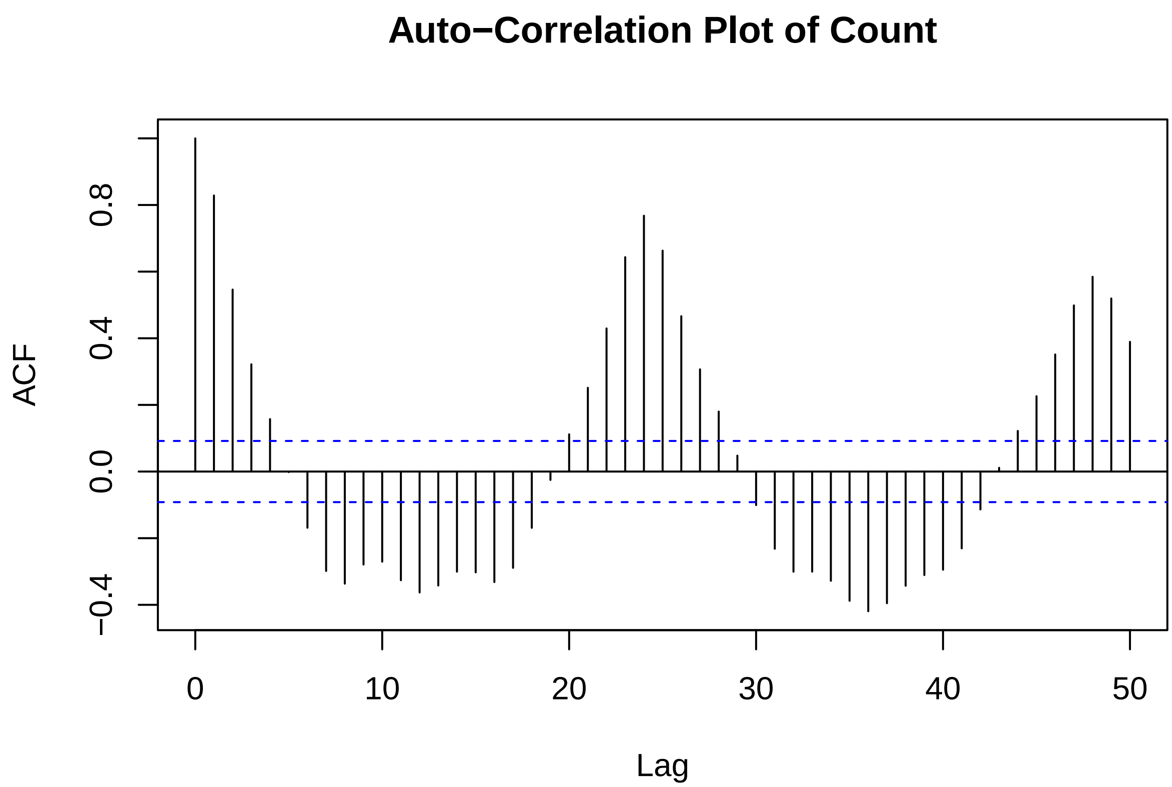
Humidity shows a negative trend, with recordings during high-humidity periods showing fewer riders.



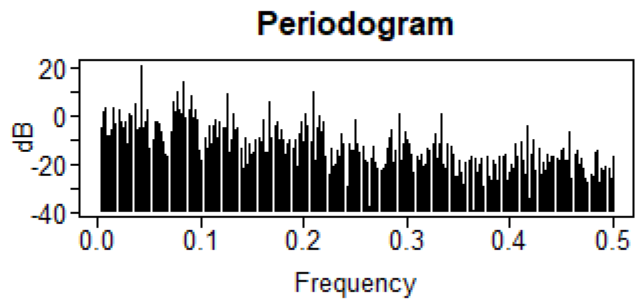
This plot shows the hourly bike rental counts for the first 19 days of June, 2019.

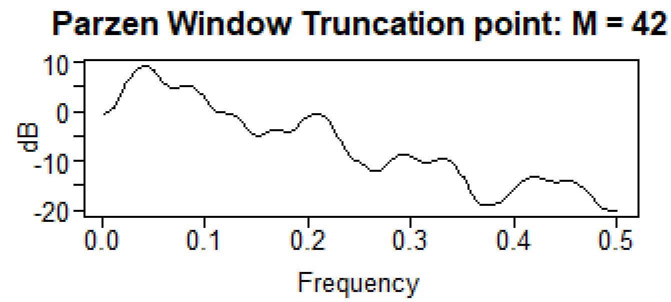
3 Methods

3.1 ARMA Model

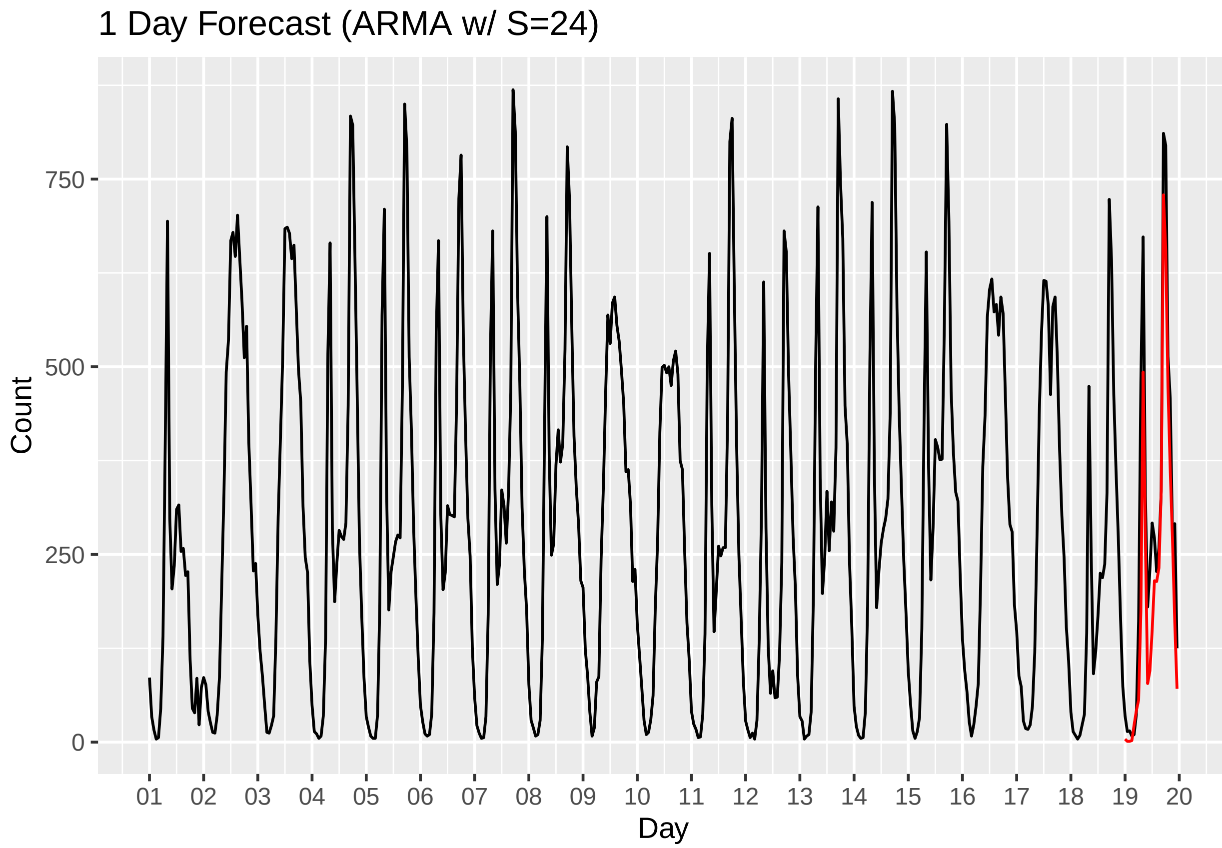


Strong sinusoidal trend with a period of 24, which would likely reflect the hourly cycles from how the data was recorded.



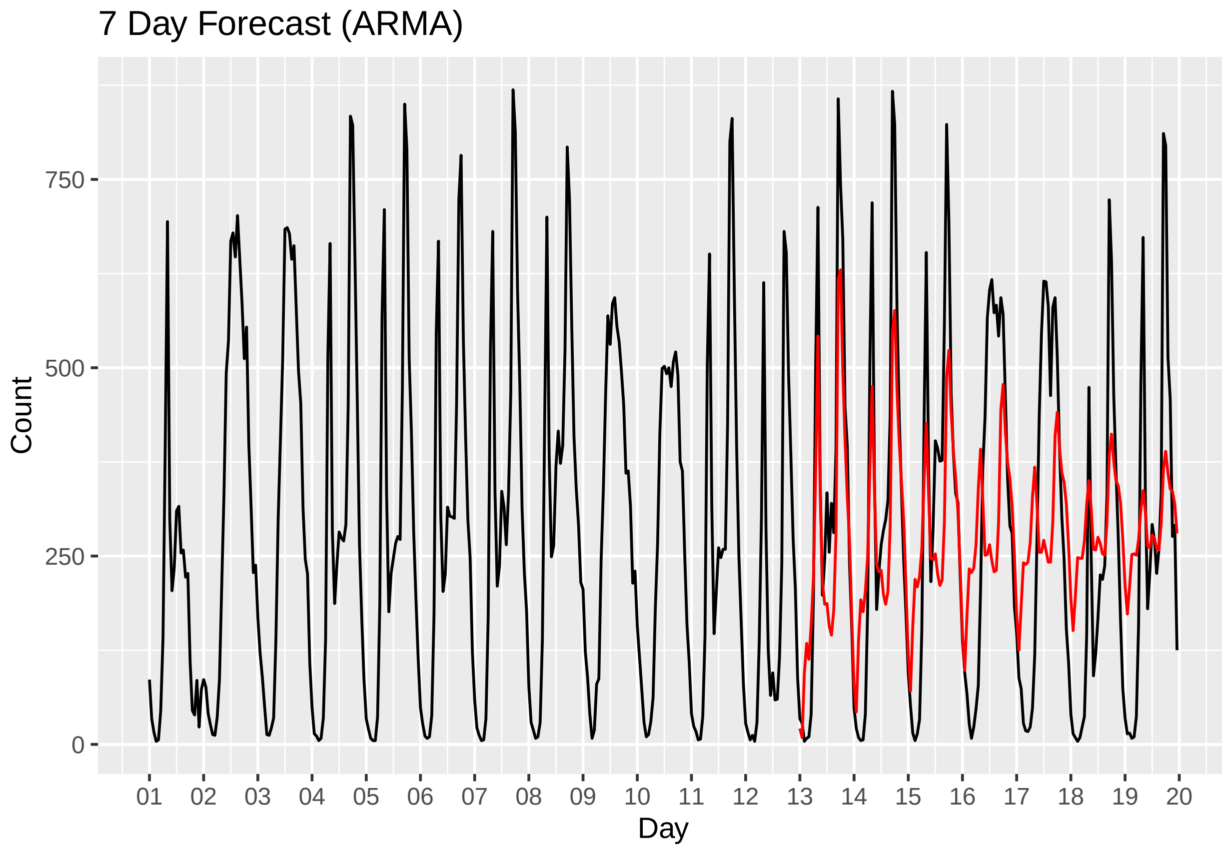


A peak can be seen in the spectral density at around 0.04. Which would equate to a period of 24, which is what would be expected again from the frequency of our data set. You can also visually confirm this from the number of cycles in the realization and divide that by the total number of observations.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

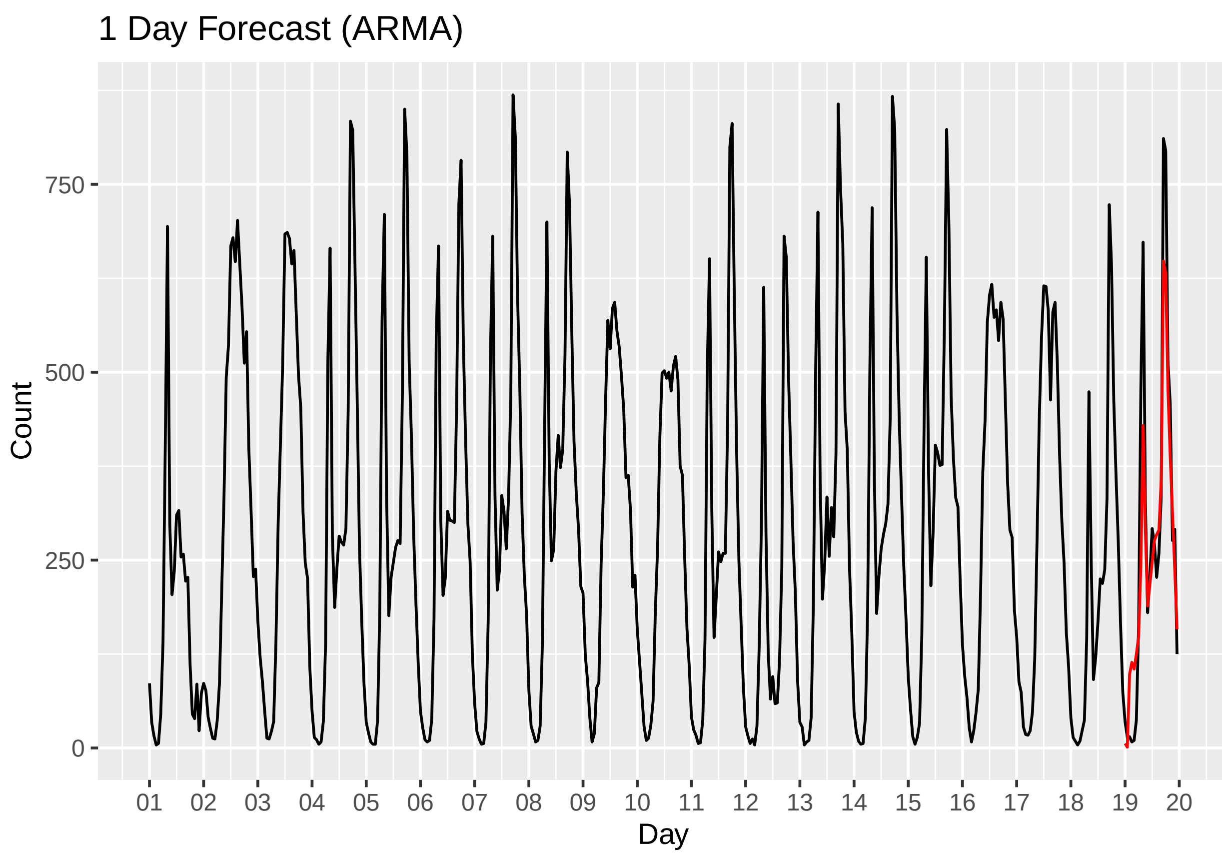


An attempt to make the time series more stationary was made with a 24th order difference. AIC recommended an ARMA(10,0) model on the residuals. We then fit the original data with the recommended phi's with a seasonal component to make for an ARIMA(10,0,0); s = 24 model. As shown later in our write-up, 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

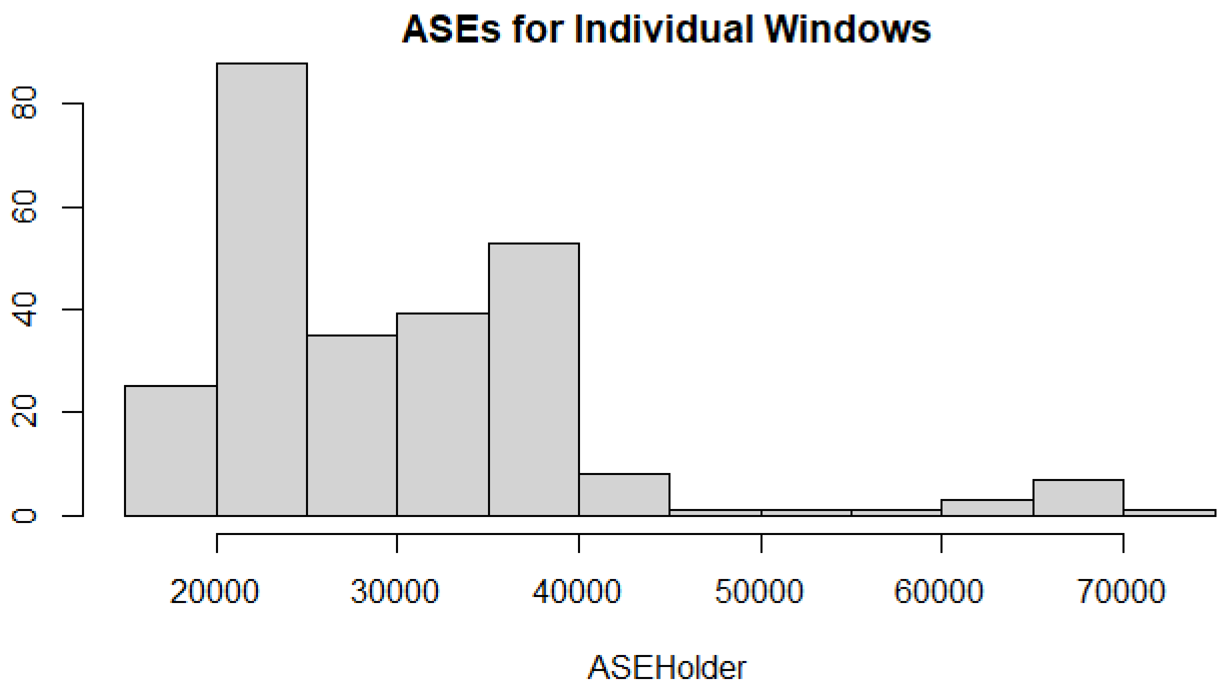


- **ASE** 2.9004653×10^4
- **RMSE** 170.308



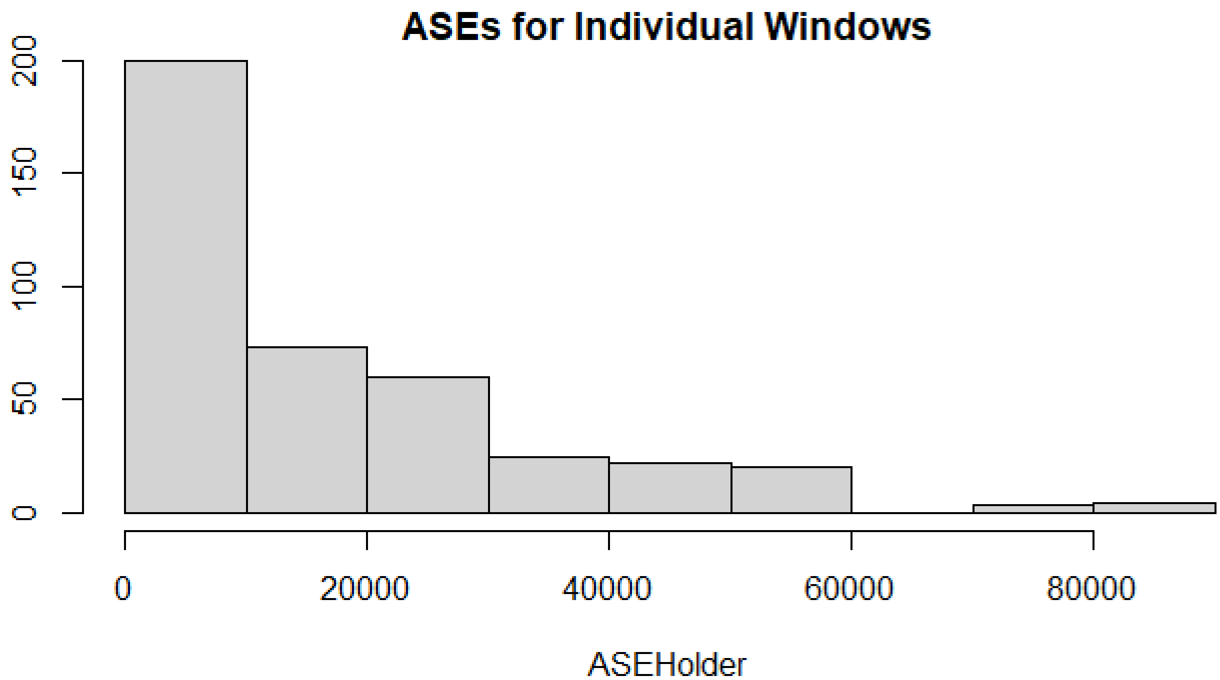
- **ASE** 9397.58
- **RMSE** 96.941

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



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

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

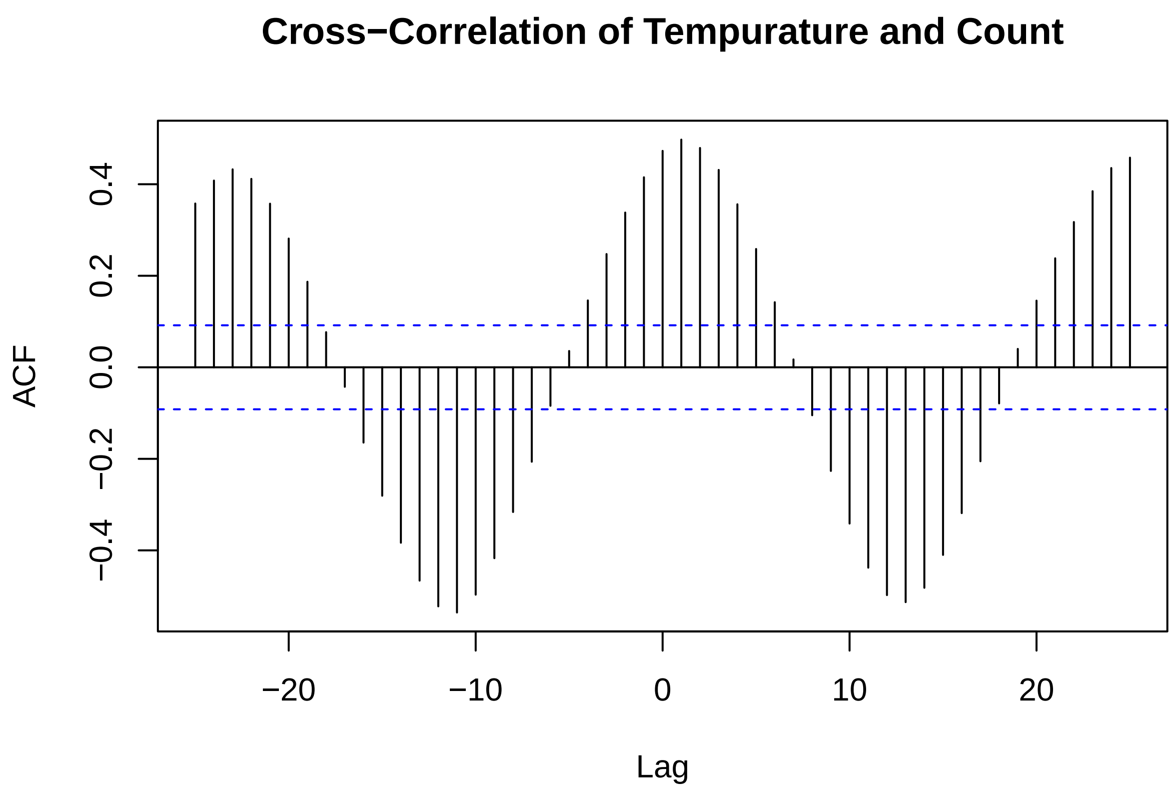


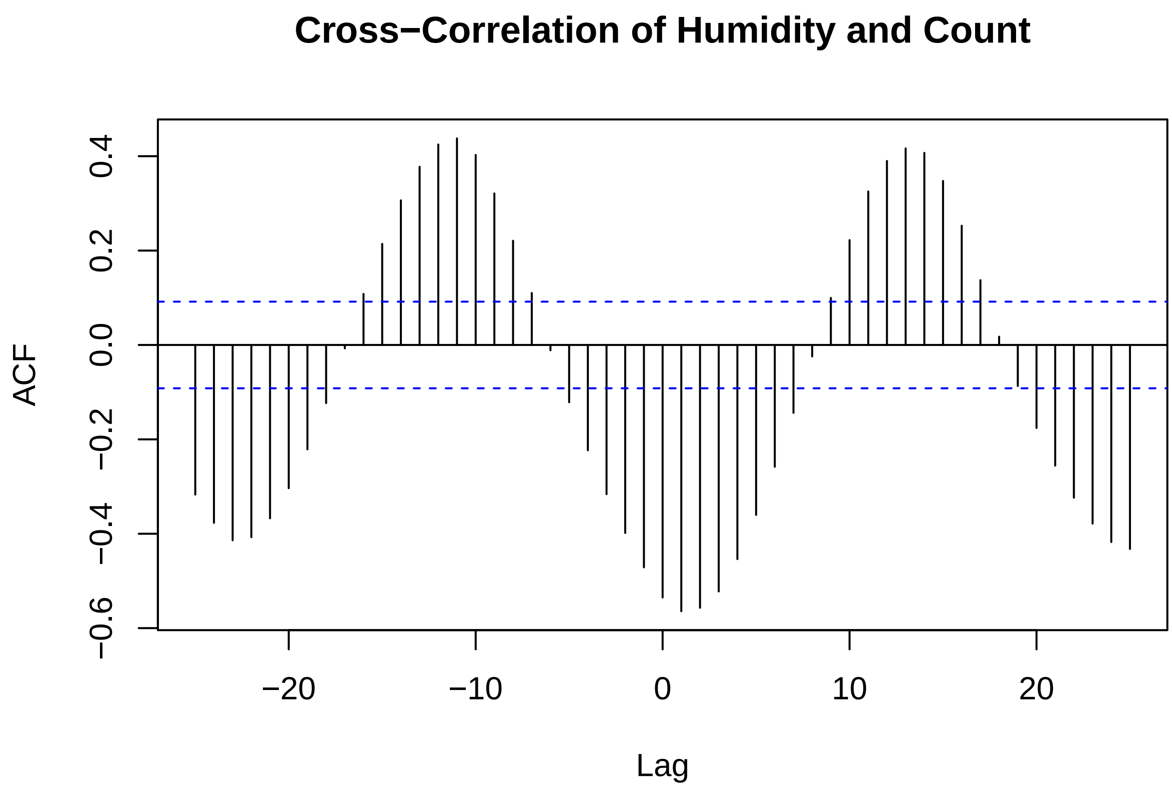
- 17,002 (mean)
- 10,648 (median)

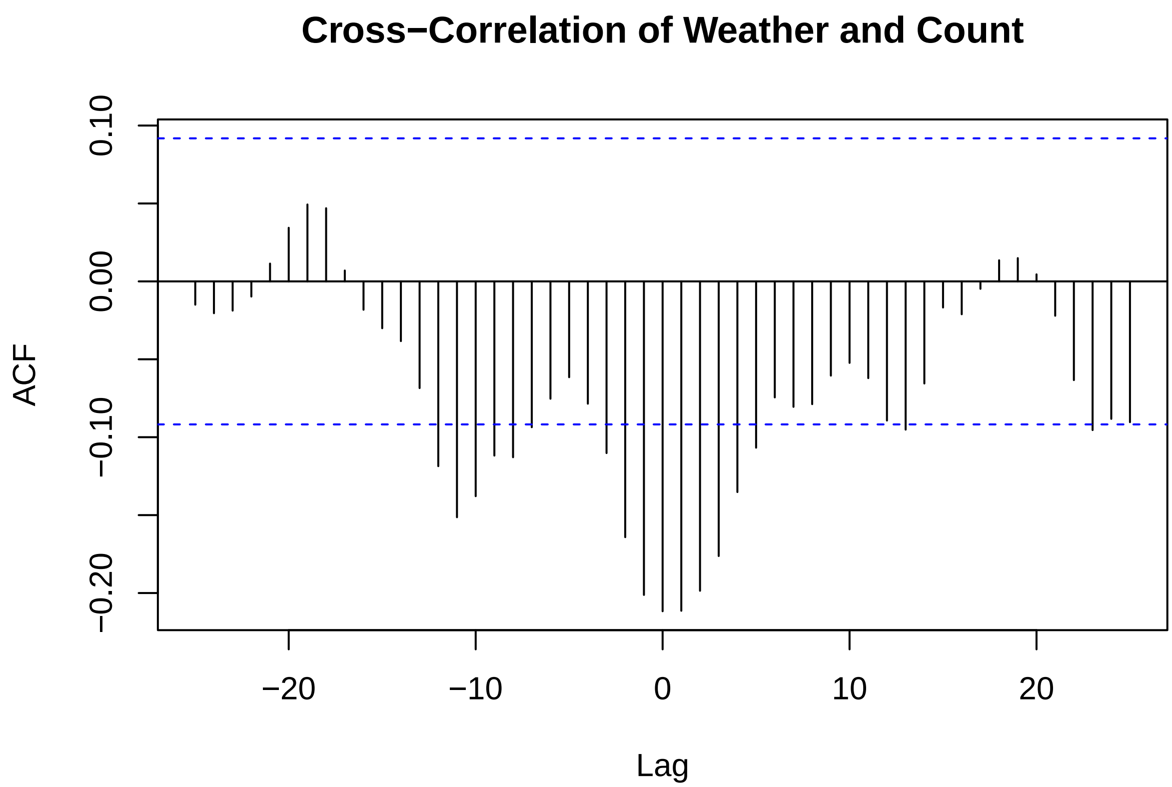
Note that the mean is much higher than the median, which would indicate a more skewed distribution as evidences by the chart above.

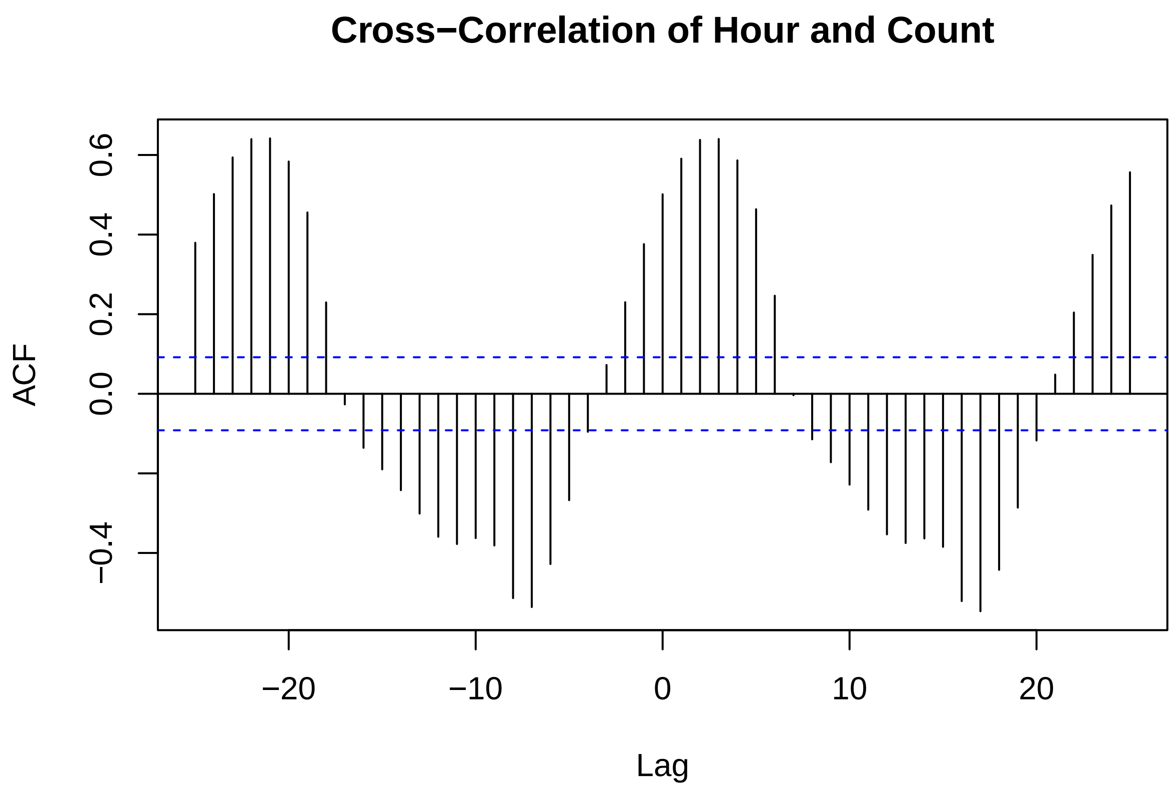
3.2 Vector Auto-Regressive (VAR) Model

3.2.1 Cross-Correlation

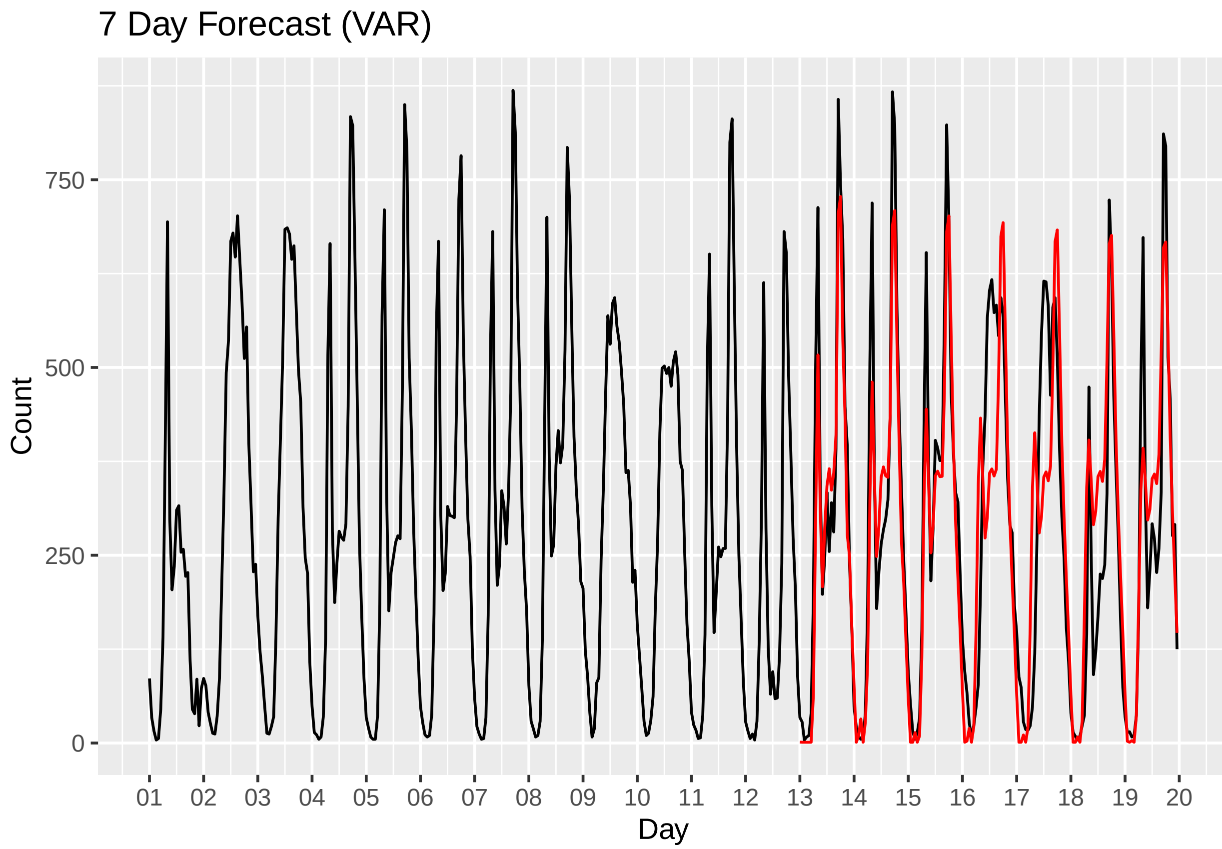



Cross−Correlation of Humidity and Count
ACF
0.4
0.2
0.0
−0.2
−0.4
−0.6
−20
−10
0
10
20
Lag


Cross−Correlation of Weather and Count
ACF
0.10
0.00
−0.10
−0.20
−20
−10
0
10
20
Lag

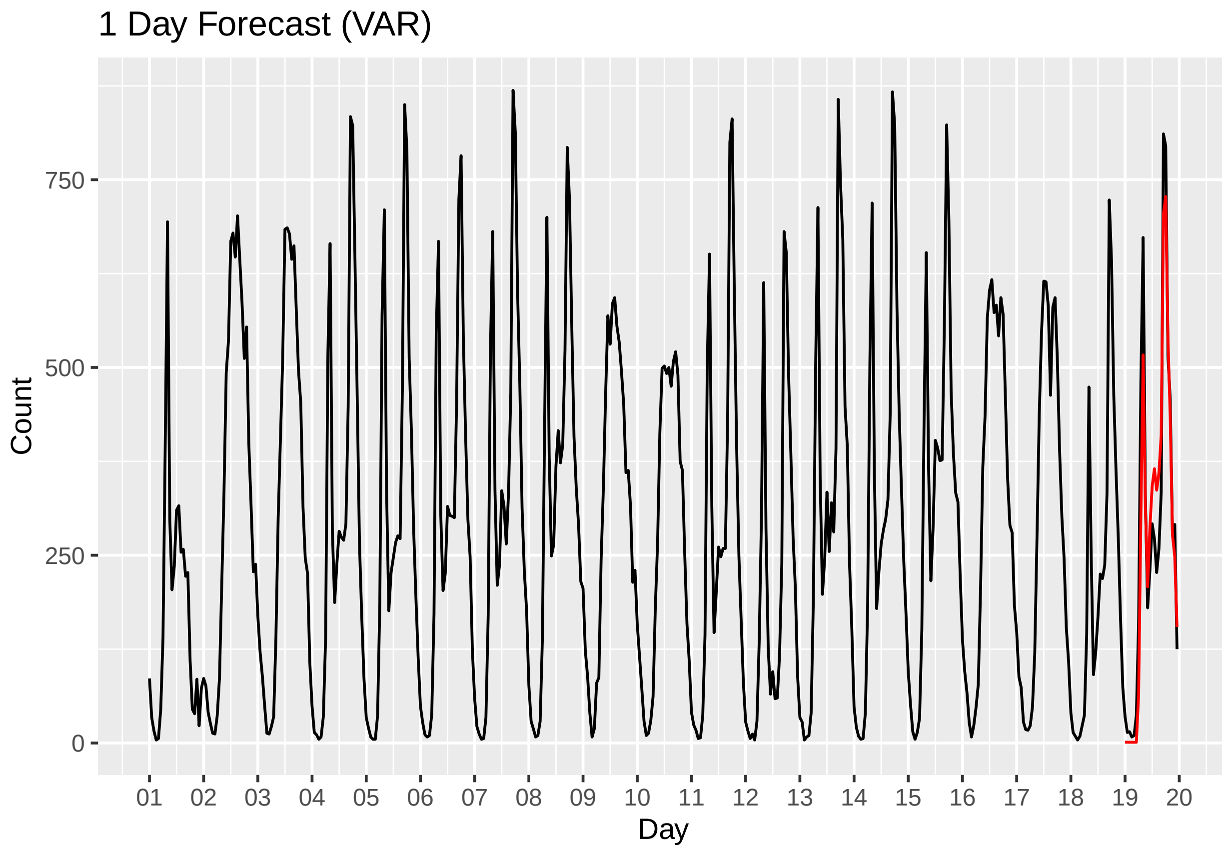

Cross−Correlation of Hour and Count
ACF
0.6
0.4
0.2
0.0
−0.4
−20
−10
0
10
20
Lag

3.2.2 7 Day Forecast



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

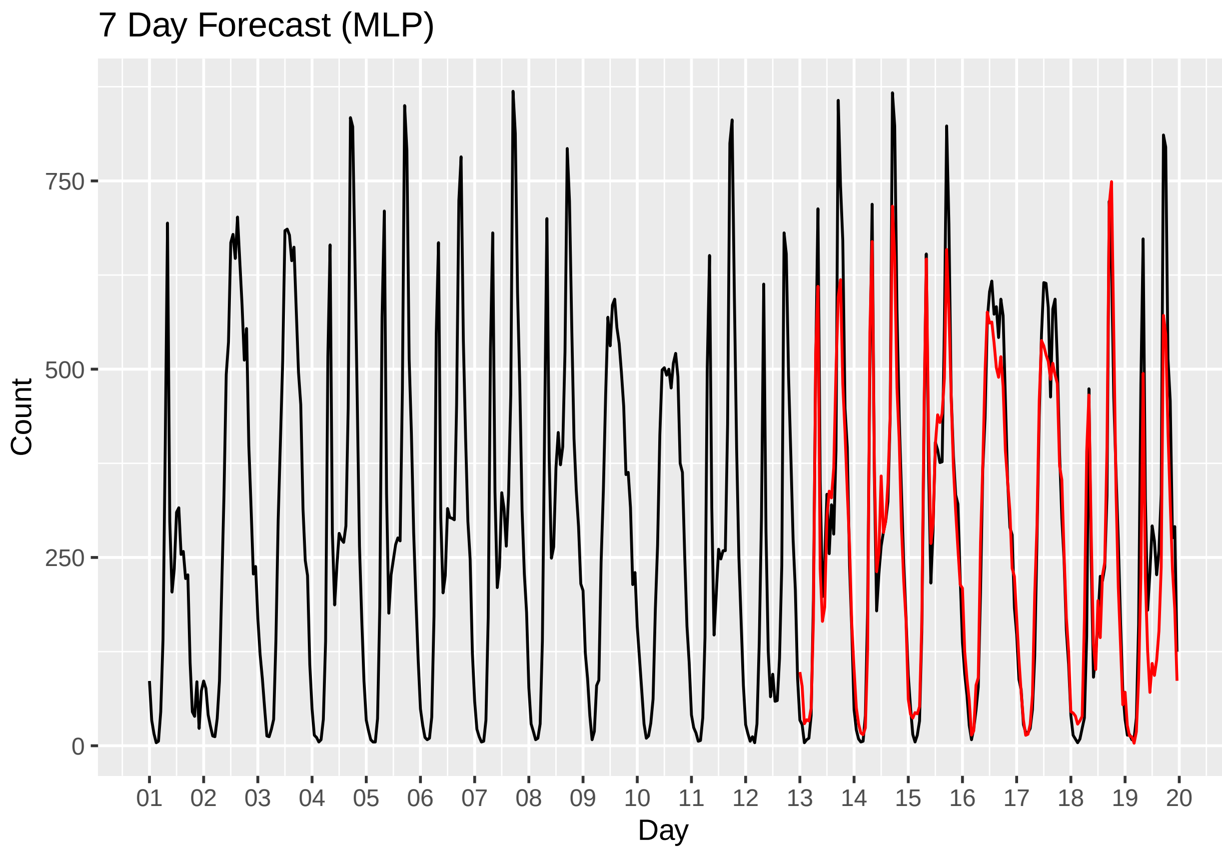
3.2.3 1 Day Forecast



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

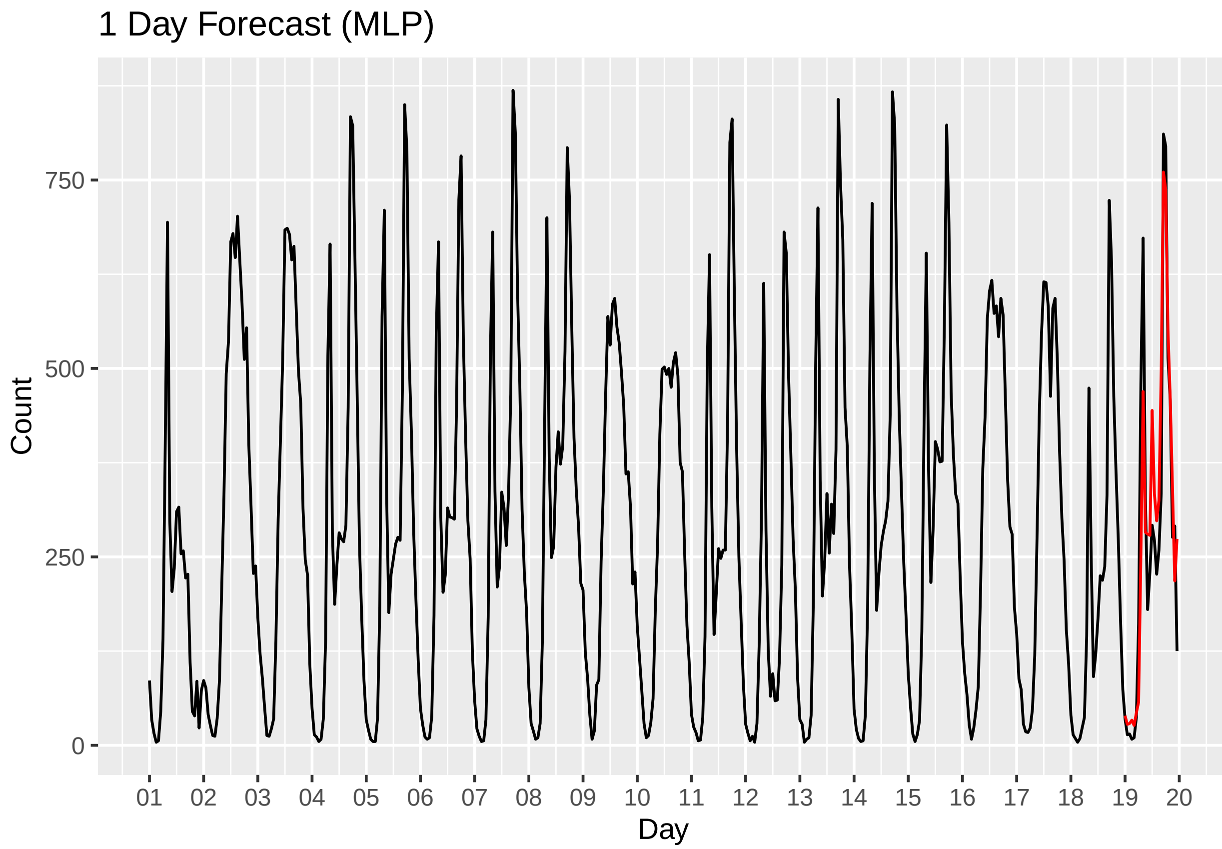
3.3 Neural Network Model

3.3.1 7 Day Forecast



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

3.3.2 1 Day Forecast



- **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

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

5 Appendix

5.1 Solution One

Log Transform the response variable

5.2 Solution Two

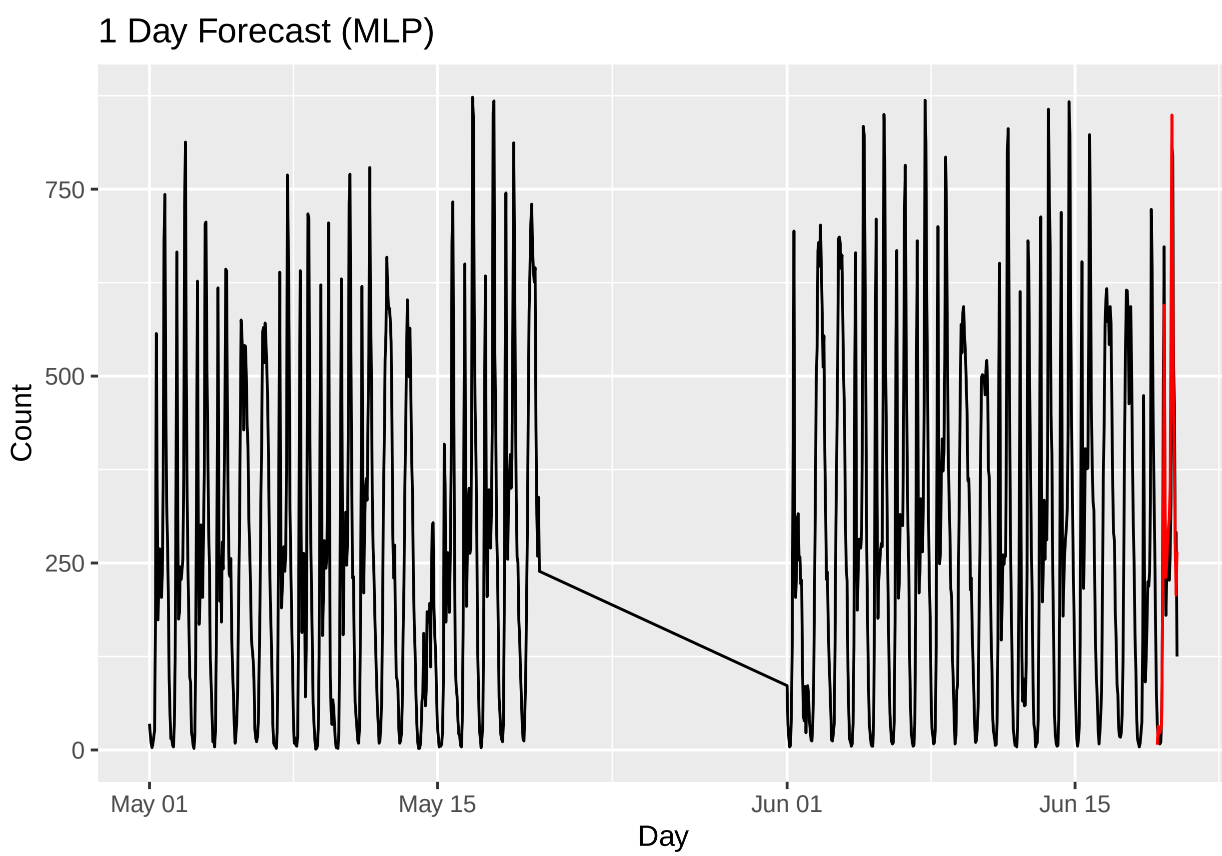
Can additional data from previous month(s) aid in lowering the ASE and RMSE? We will focus on using an MLP neural network with additional input features to test our hypothesis when forecasting the short-term counts for the month of June, 2012.

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.



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.