Bike Share Demand Forecasting Methods

Chance Robinson and Kebur Fantahun

11/22/2021

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.

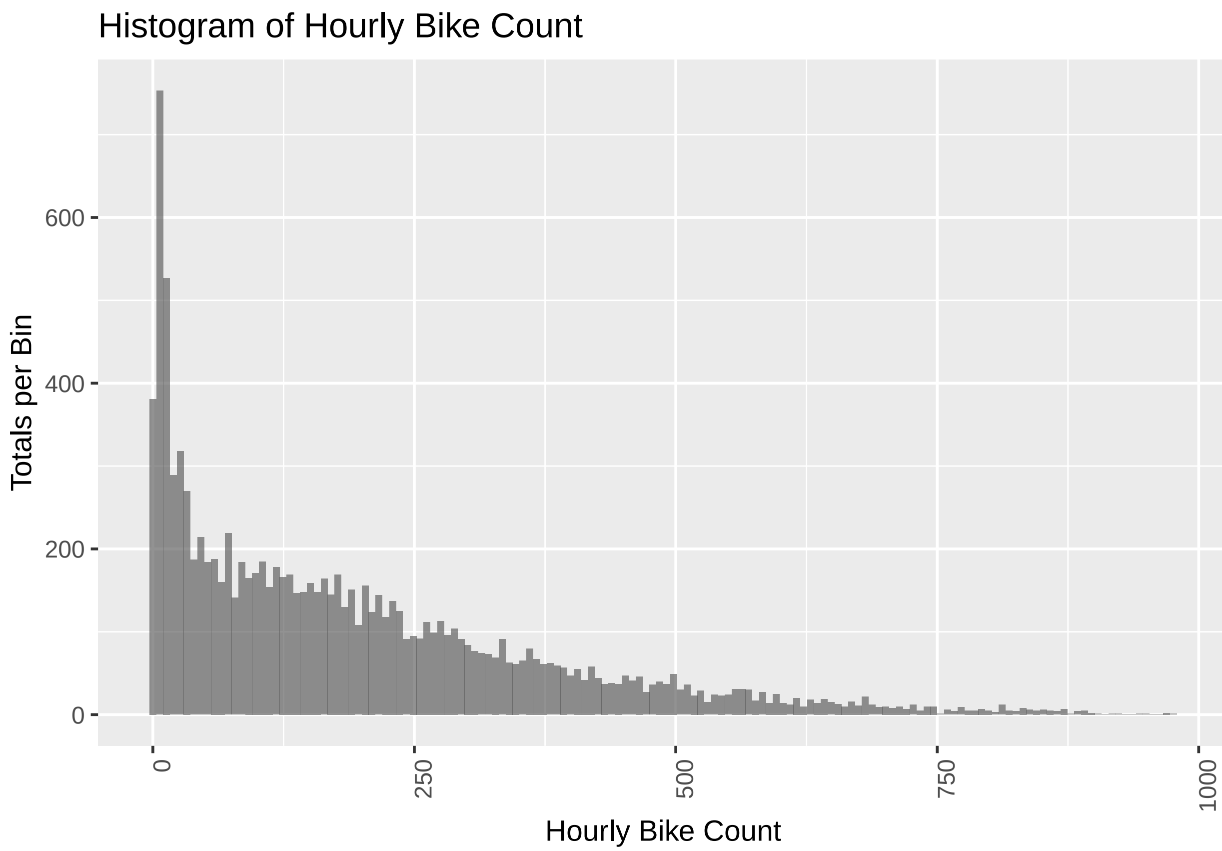
2 Data Analysis

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

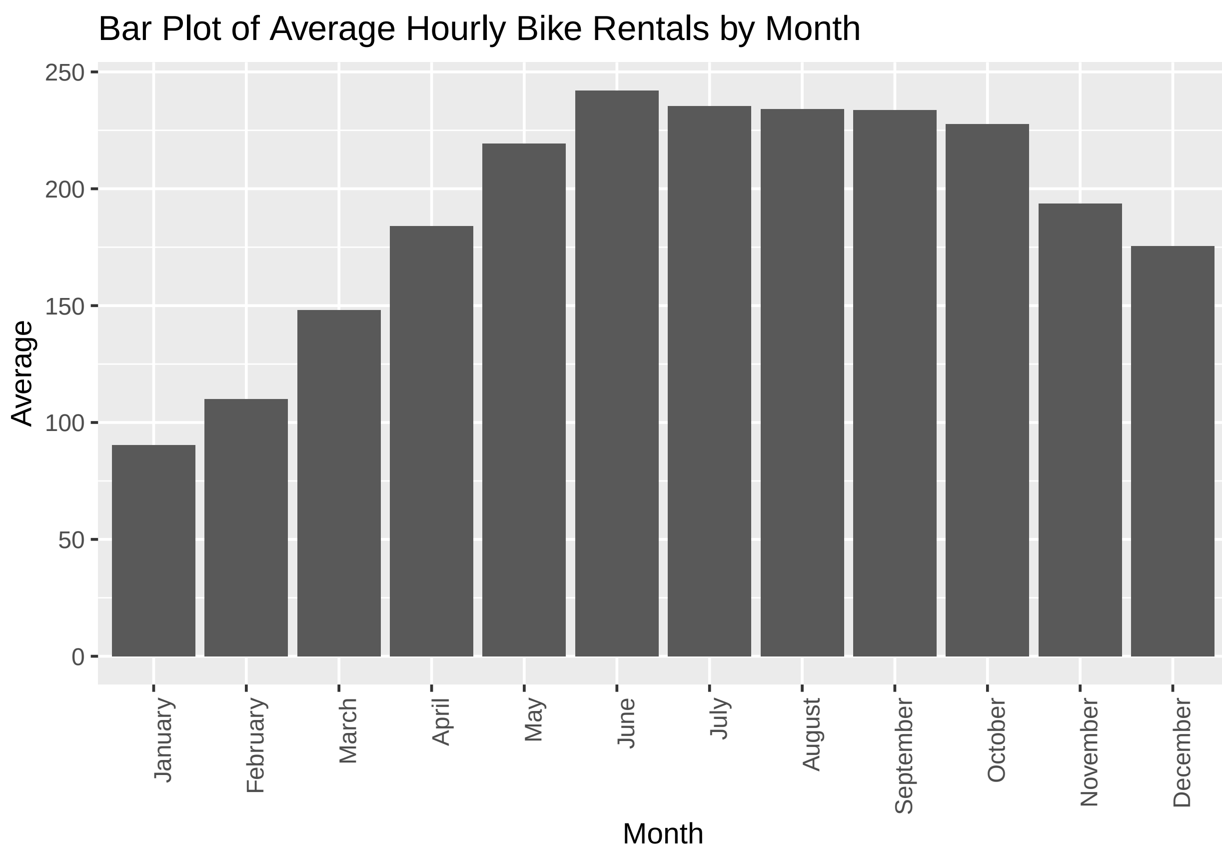
There are 12 columns provided in the `train.csv` data set with 10,886 observations. The `test.csv` data set has 6,493 records, or roughly 37% of the overall combined samples from the two files. This is due to the fact that the training data consist of the first 19 days of each month and the test the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

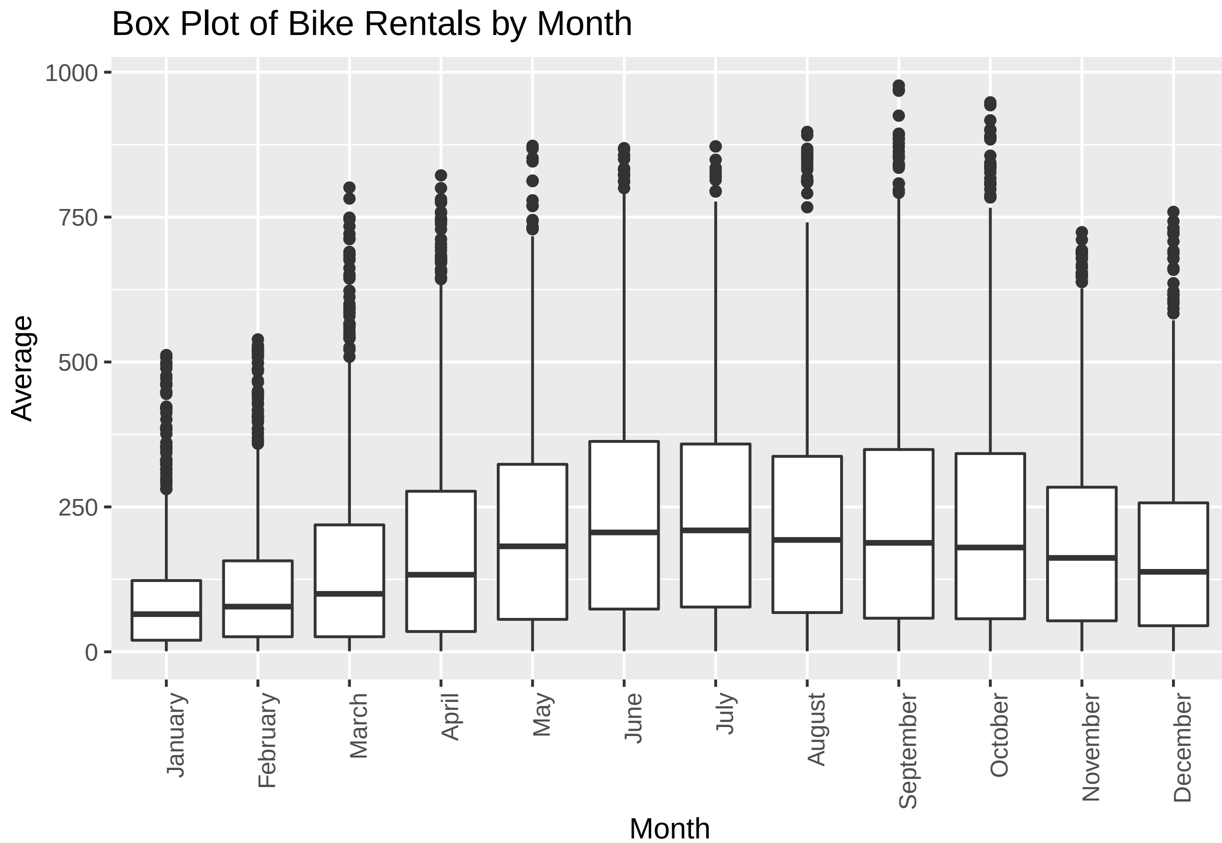
Var1	Freq
Min.	1.0000
1st Qu.	42.0000
Median	145.0000
Mean	191.5741
3rd Qu.	284.0000
Max.	977.0000



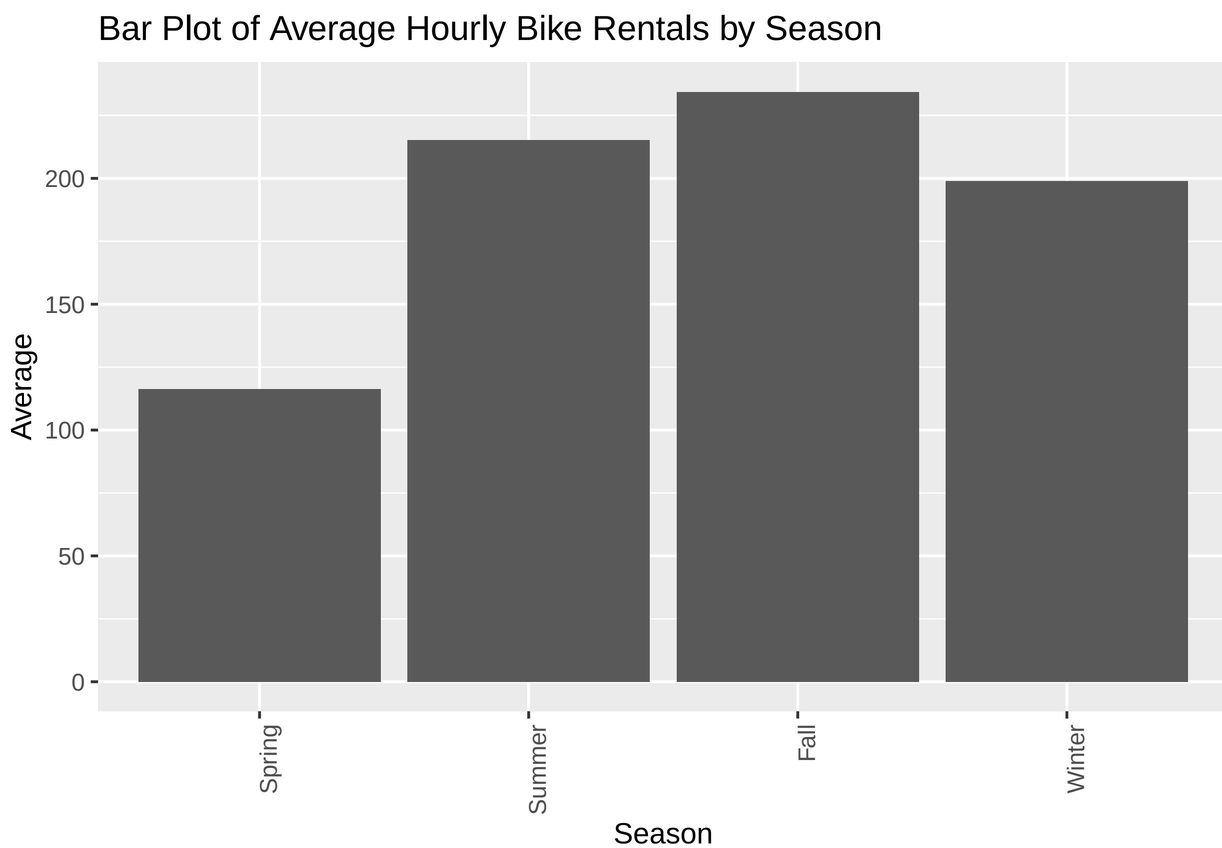
The response variable of `count` appears to be heavily right-skewed, with the median value at ~145 users. Additional summary statistics are show below.



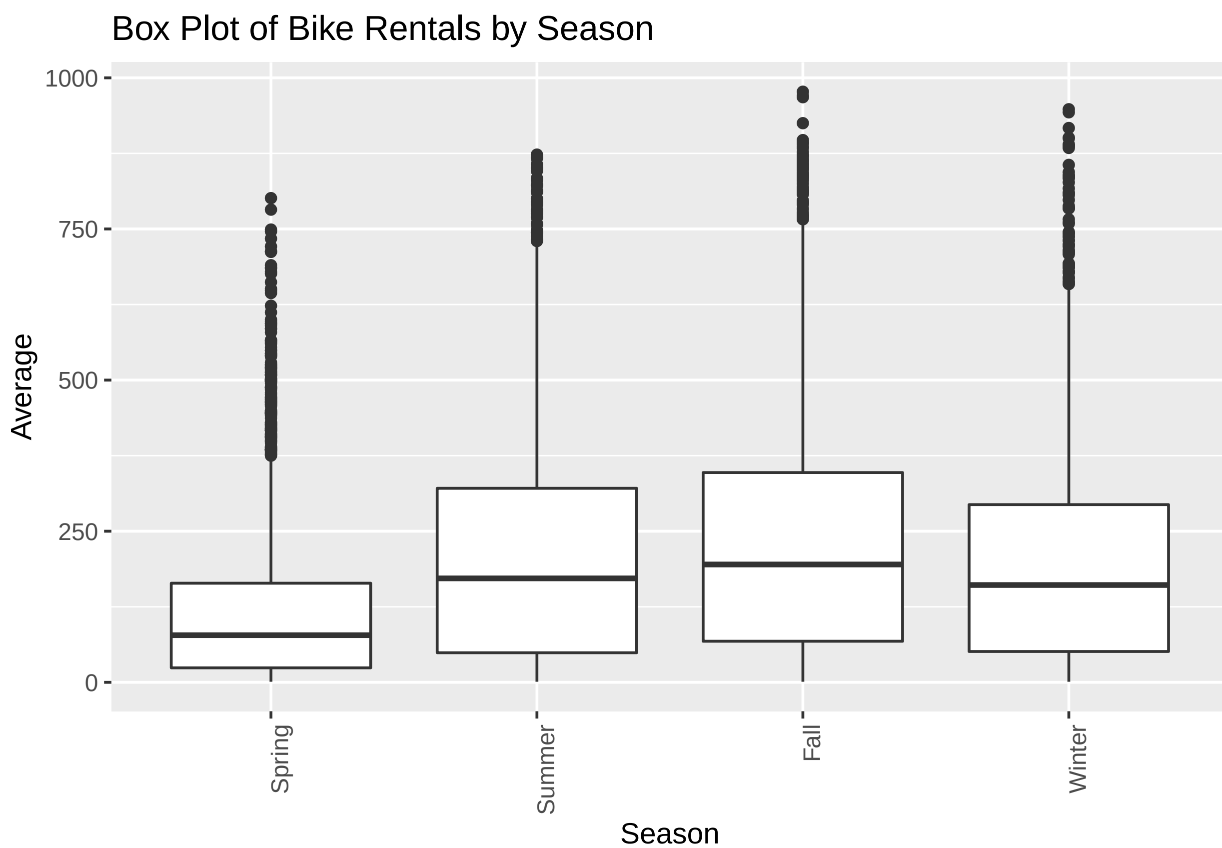
June appears to be the month with heaviest demand.

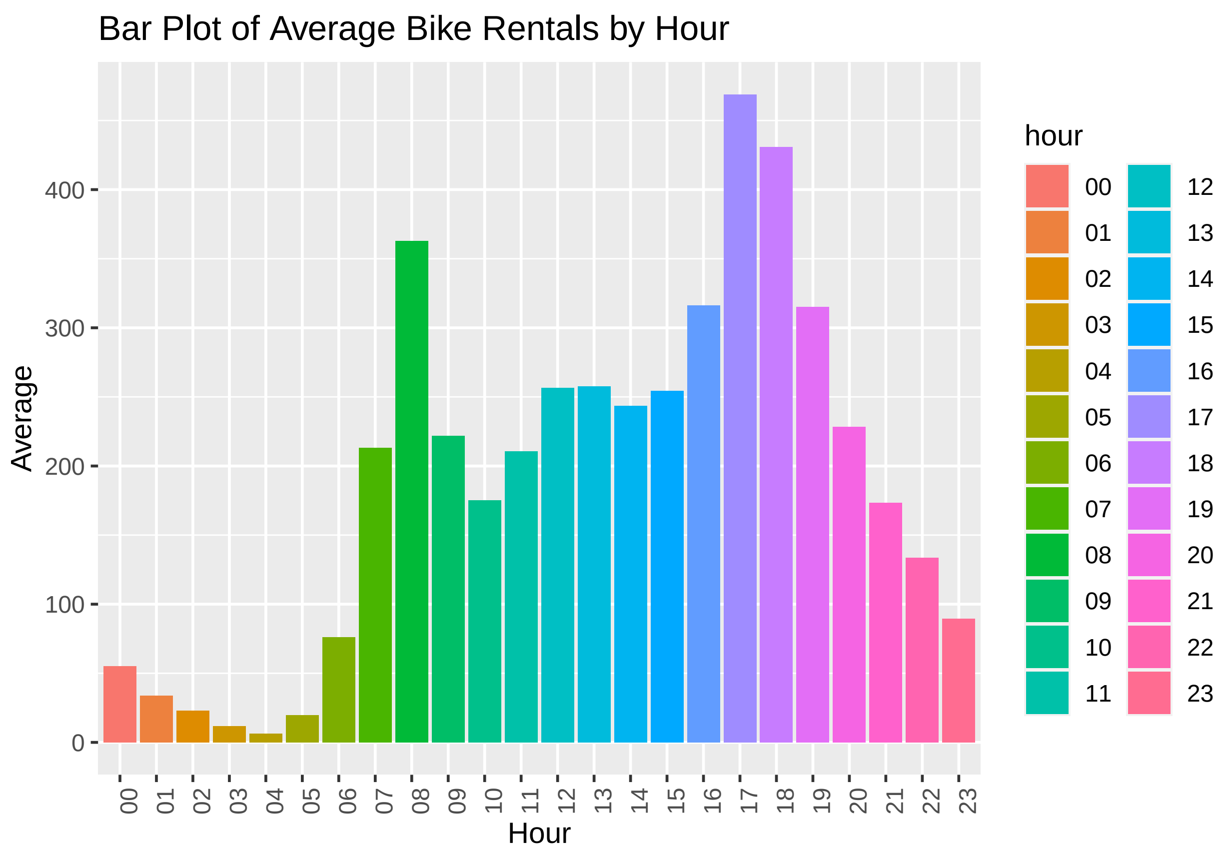


It also appears to show relatively few outliers compared to months like January.

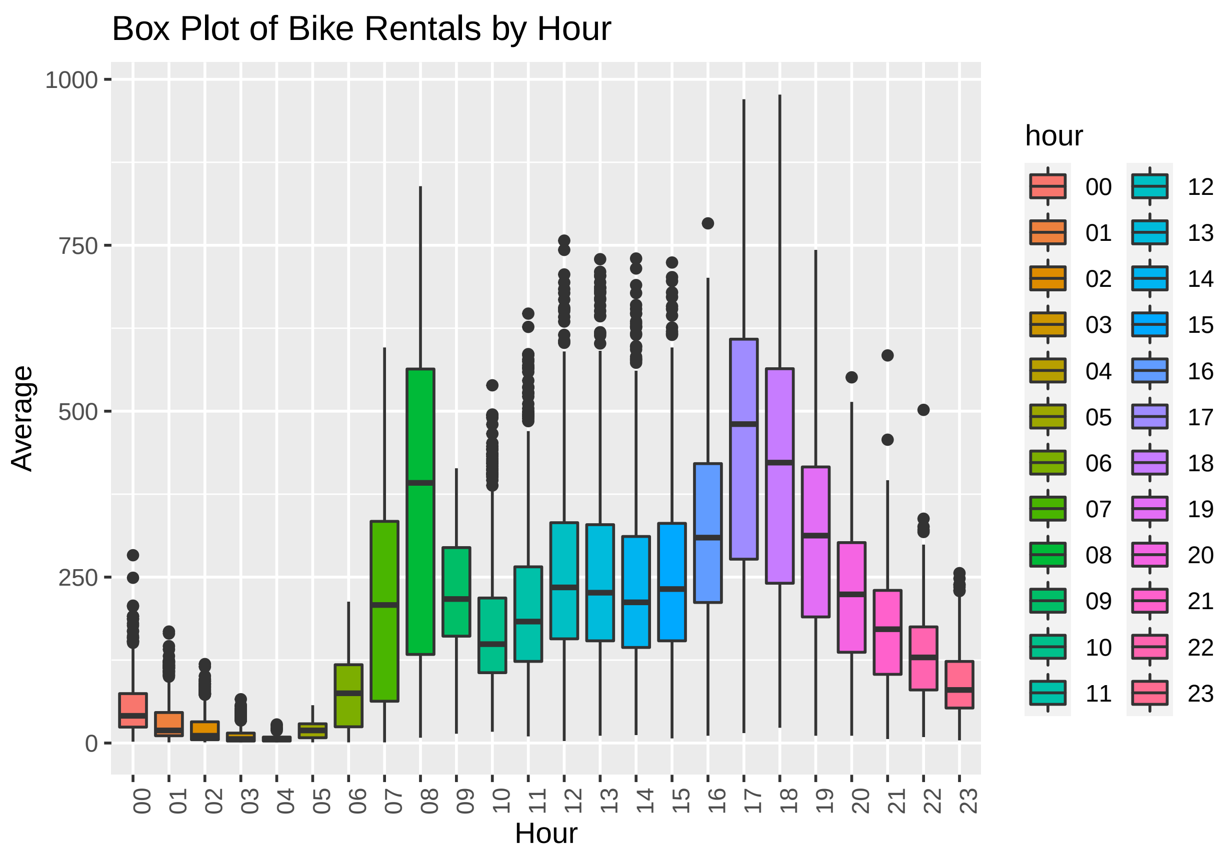


Fall appears to have more rentals on average than the other seasons.

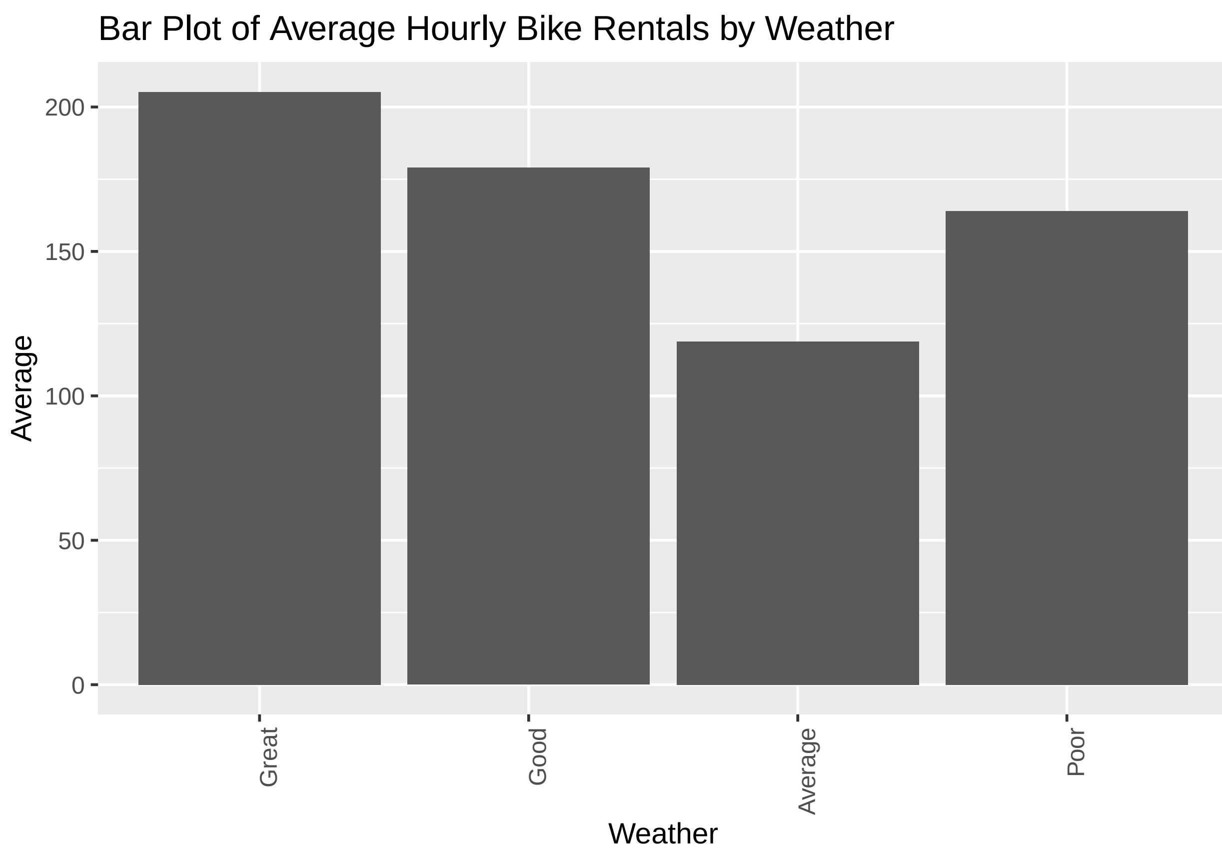

Box Plot of Bike Rentals by Season
1000
750
500
250
0
Average
Spring
Summer
Fall
Winter
Season



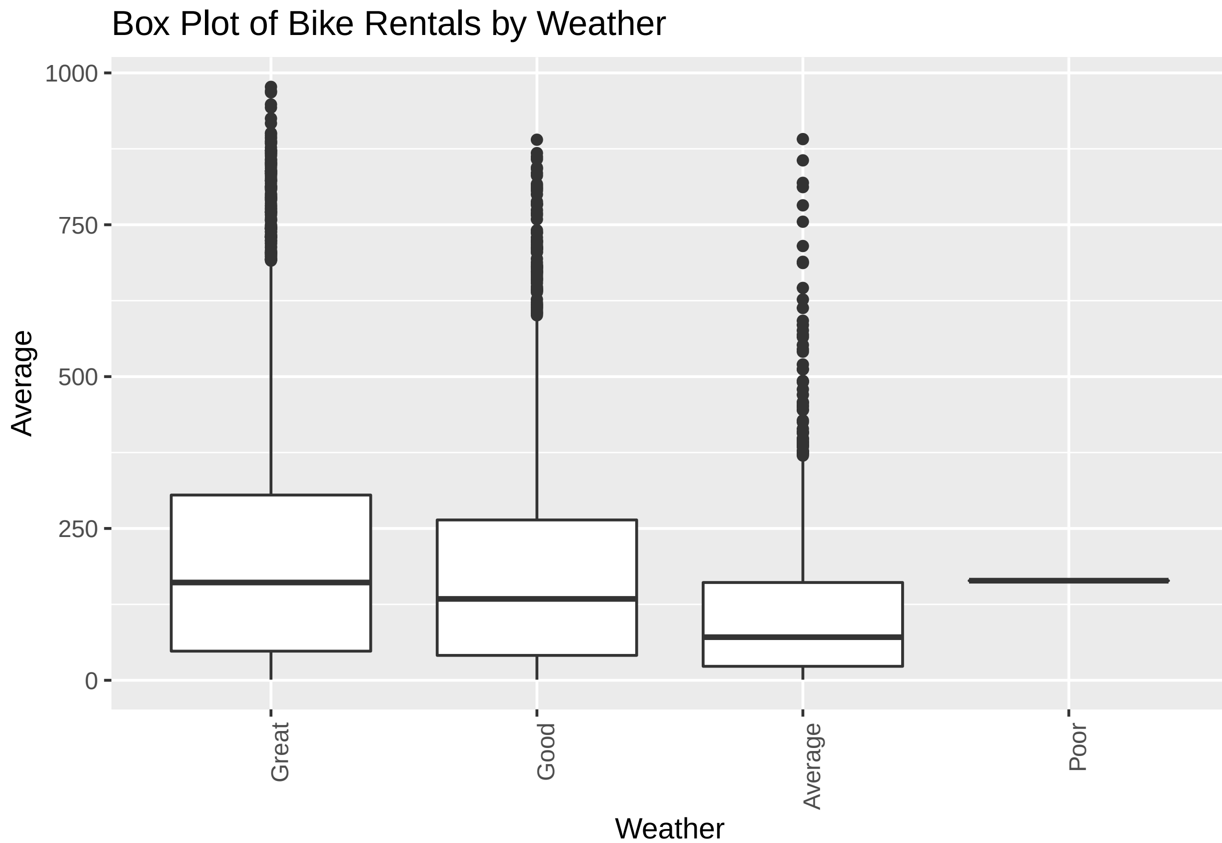
The 5pm hour clearly has the highest peak compared to the other hours of the day.



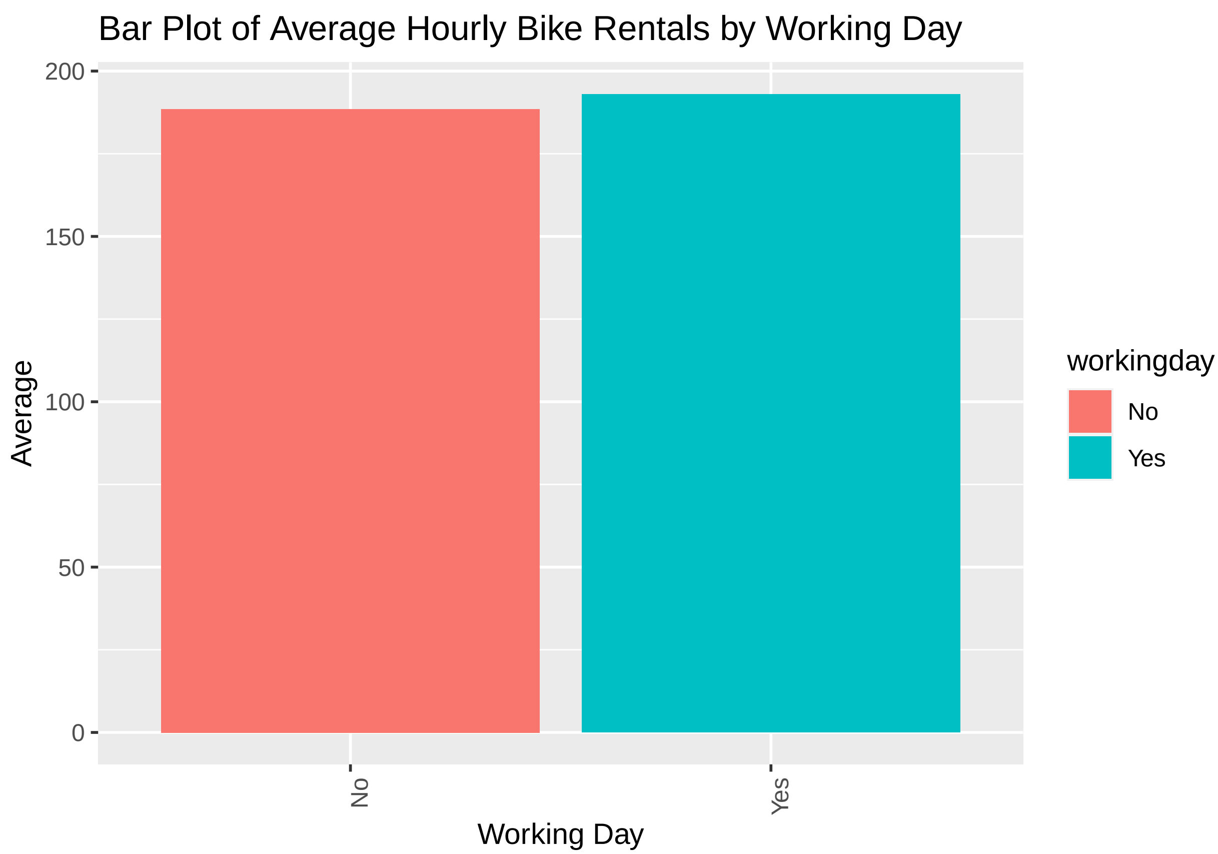
Box Plot of Bike Rentals by Hour
Average
1000
750
500
250
0
Hour
00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23
hour
00 01 02 03 04 05 06 07 08 09 10 11
12 13 14 15 16 17 18 19 20 21 22 23



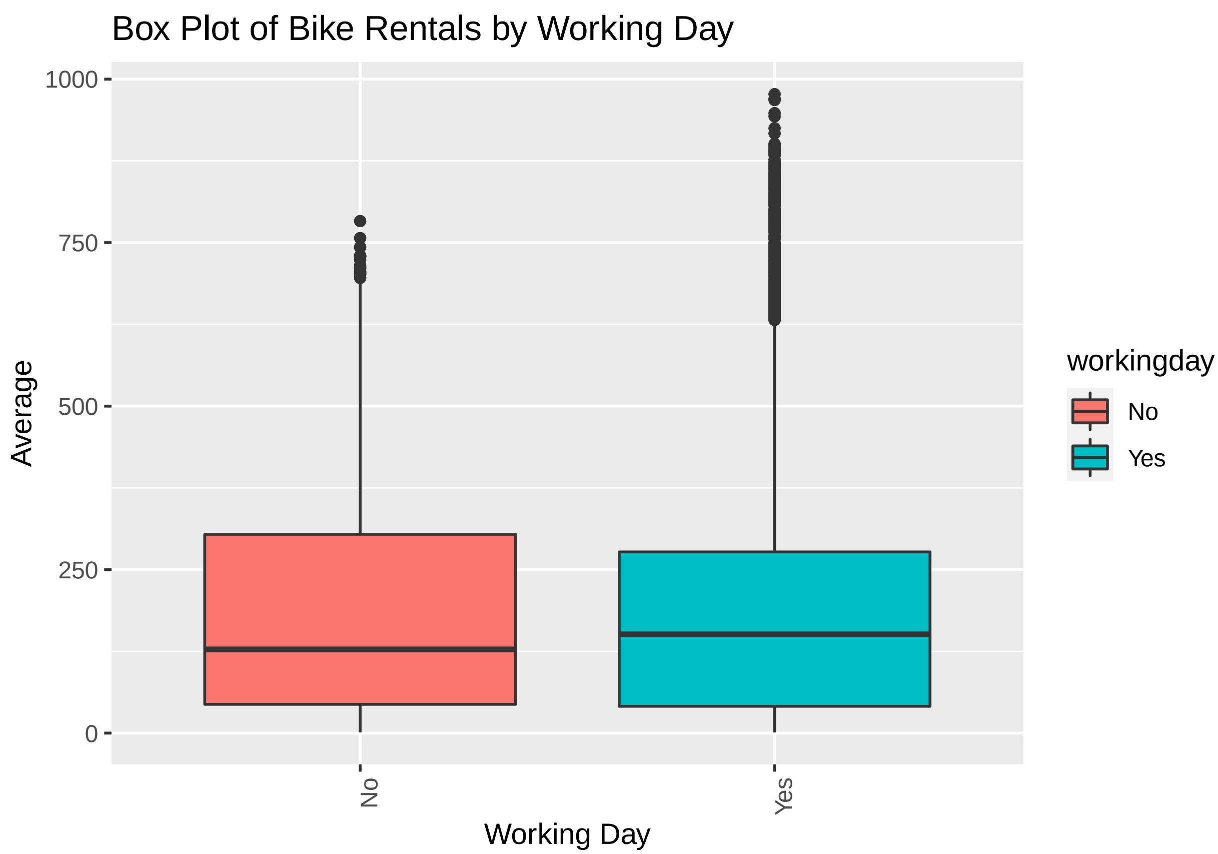
As expected, more riders are out when the weather is great, or better than average.

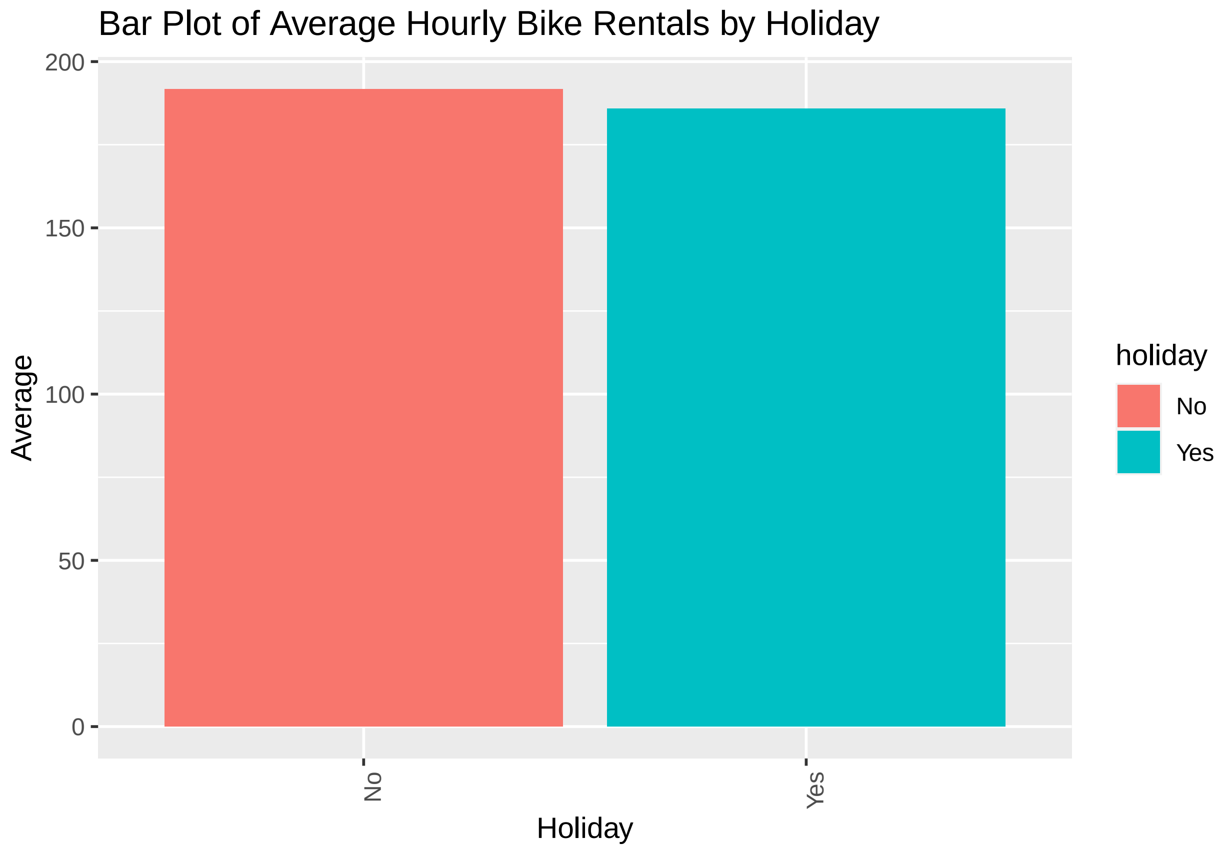


Note the small spread of riders when the weather was listed as poor, even though the averages were consistent with other categories.

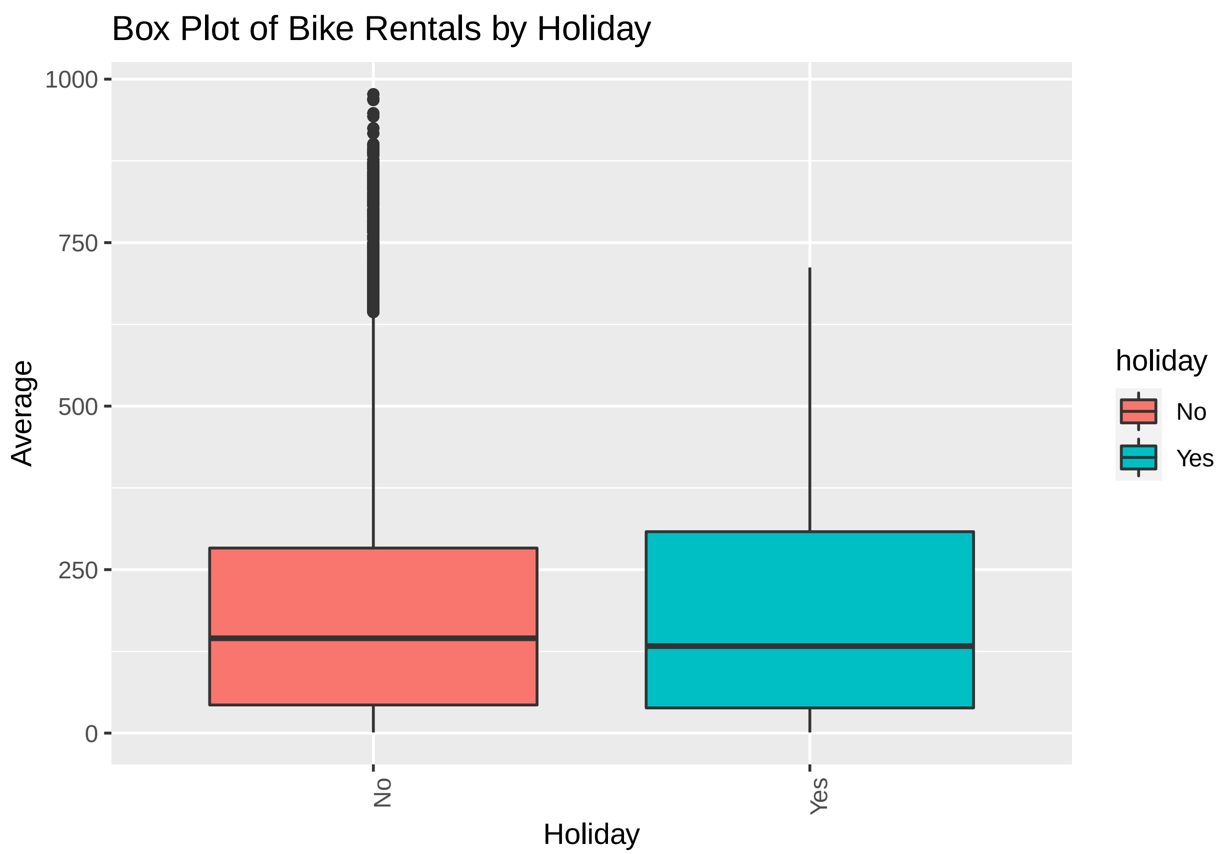


Surprisingly, whether or not the day was on a working day or not had little affect on the mean or median.

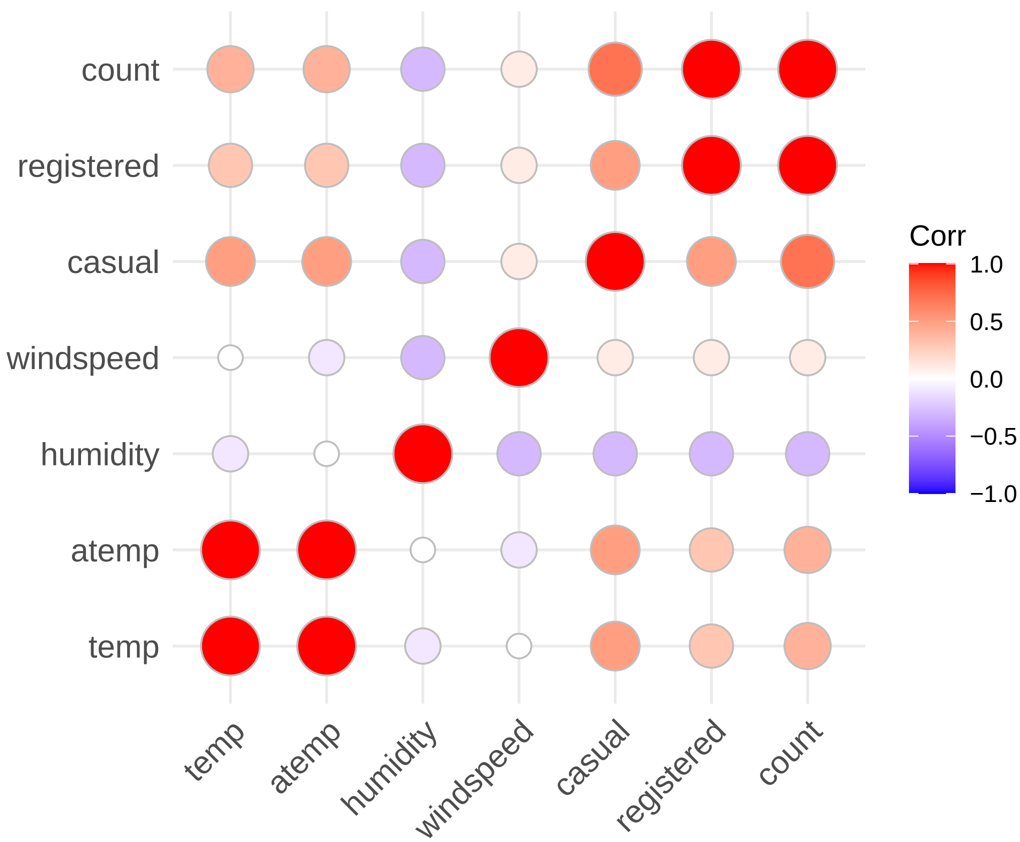

Box Plot of Bike Rentals by Working Day
1000
750
500
250
0
Average
No
Yes
Working Day
workingday
No
Yes



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

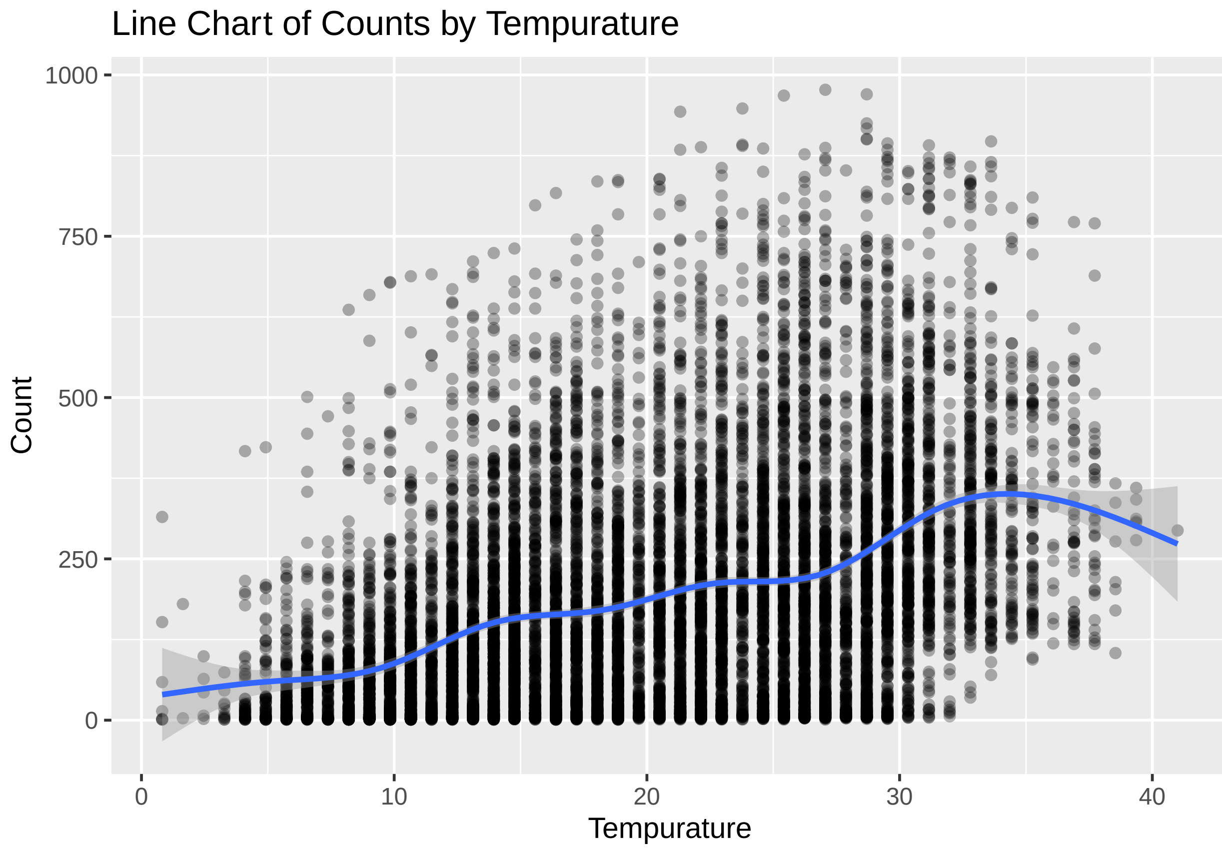

Box Plot of Bike Rentals by Holiday
1000
750
500
250
0
Average
No
Yes
Holiday
holiday
No
Yes

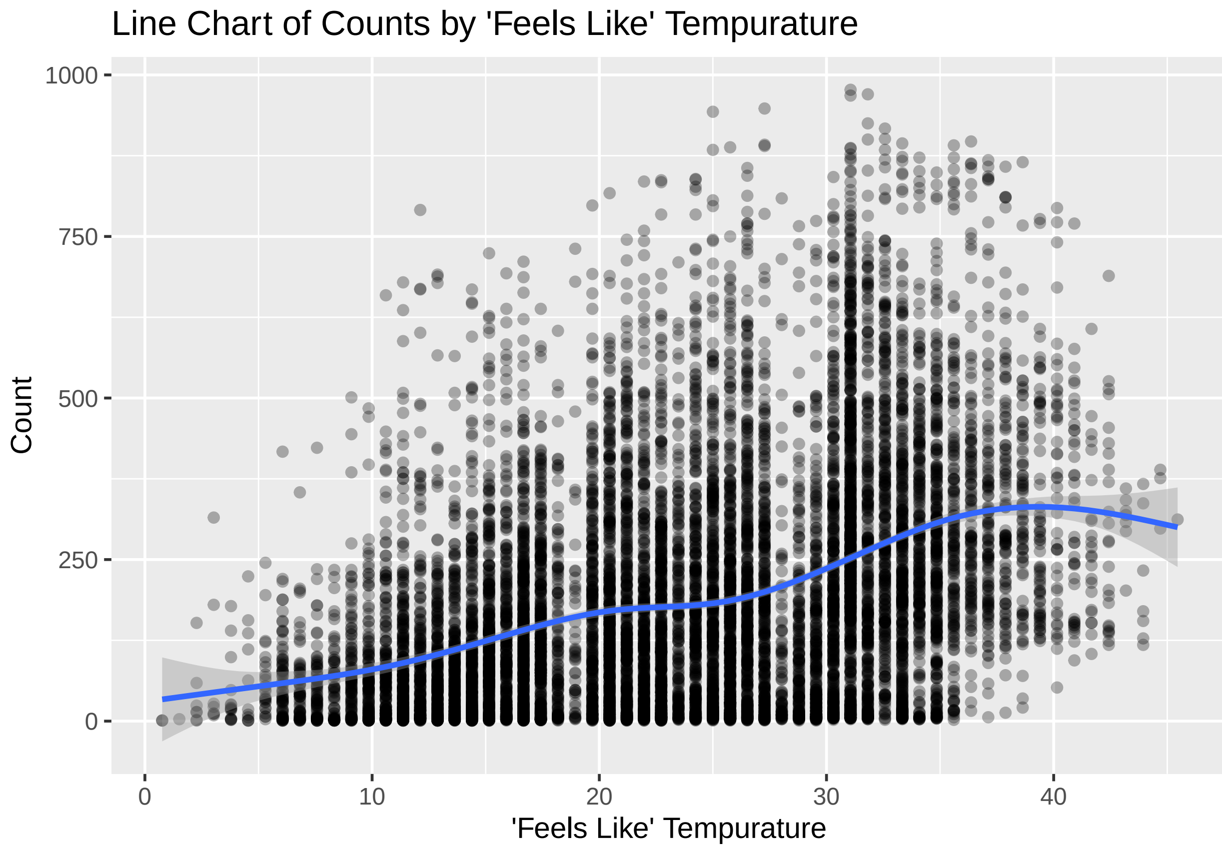
Feature 1	Feature 2	Correlation Coefficient
temp	atemp	0.9849481
registered	count	0.9709481
casual	count	0.6904136
casual	registered	0.4972497
temp	casual	0.4670971
atemp	casual	0.4620665
temp	count	0.3944536
atemp	count	0.3897844
temp	registered	0.3185713
atemp	registered	0.3146354
windspeed	count	0.1013695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

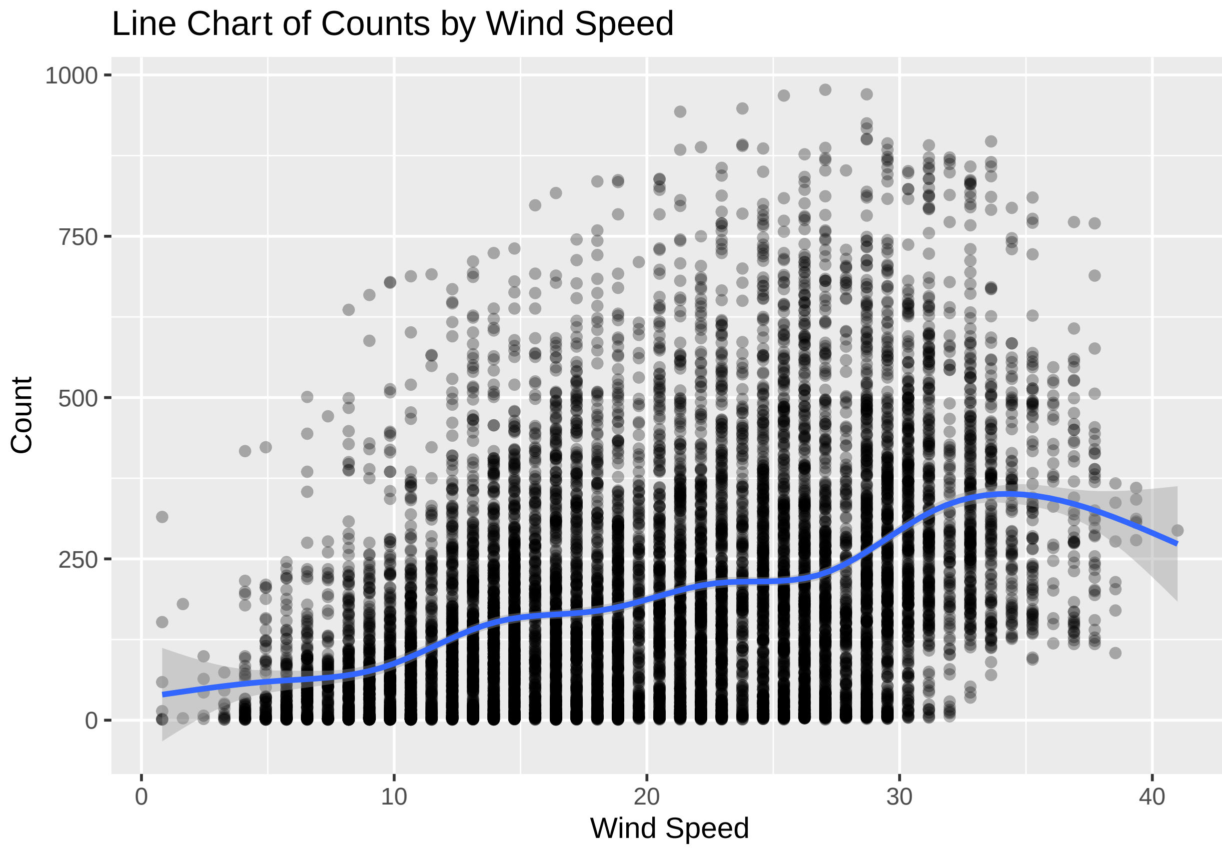


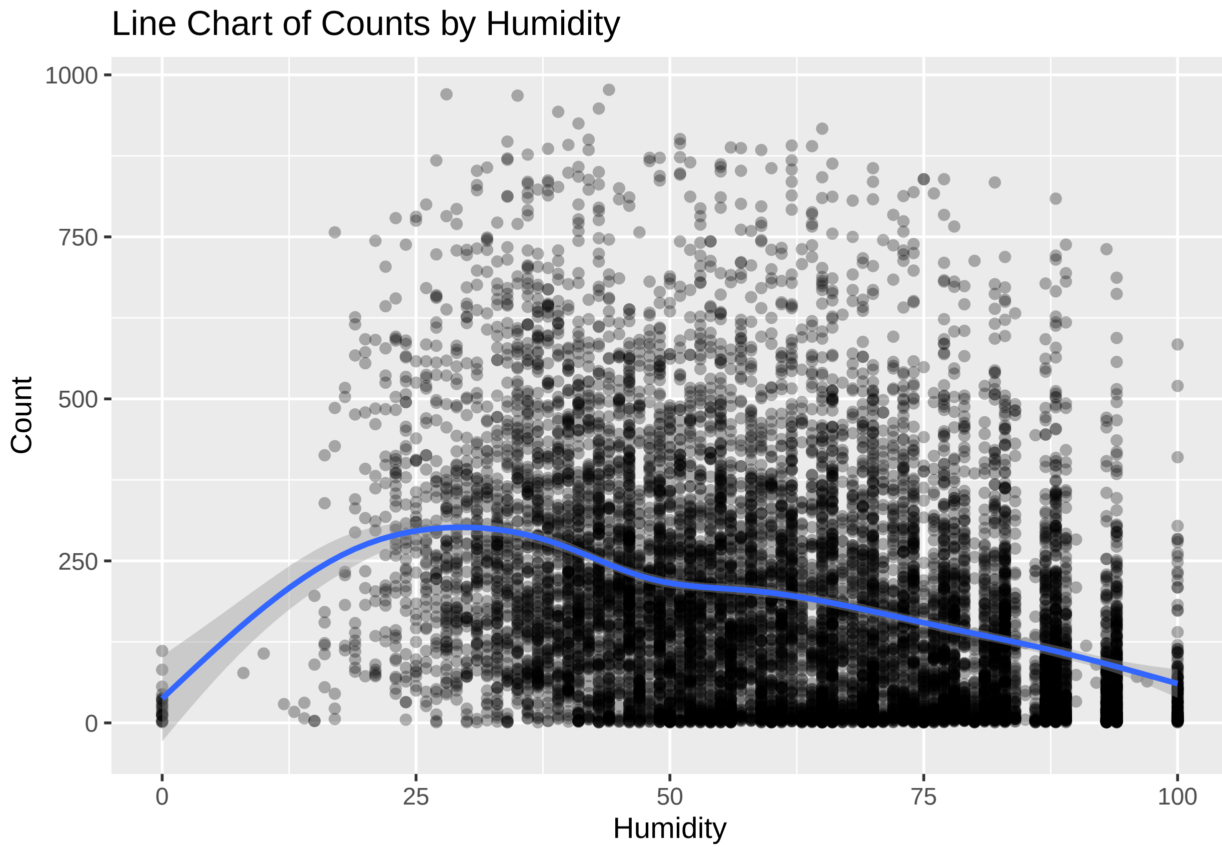
The plot below shows a strong correlation between casual, registered and total (count) users. We will use count as the primary response variable for our modeling, and discard the other response variables.

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

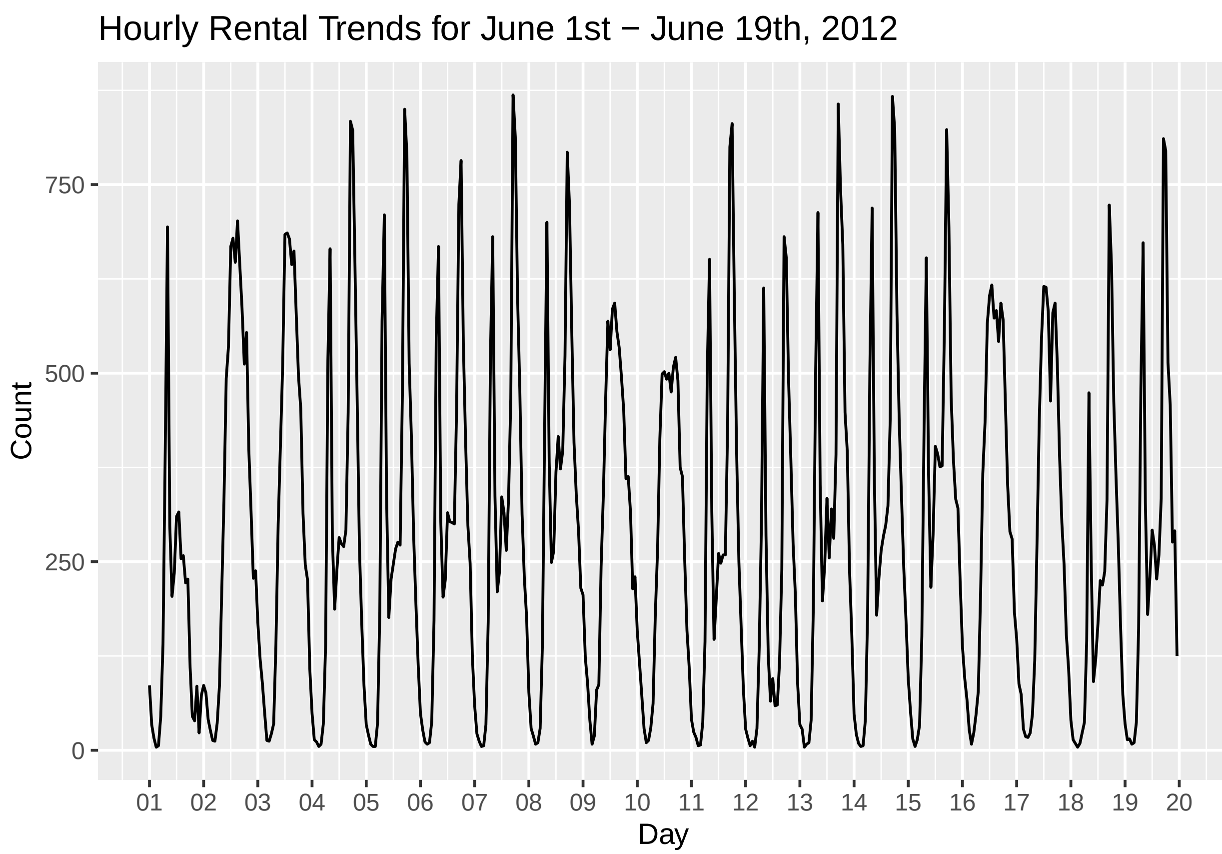

Line Chart of Counts by Tempurature
Count
1000
750
500
250
0
0
10
20
30
40
Tempurature


Line Chart of Counts by 'Feels Like' Tempurature
1000
750
500
250
0
Count
0
10
20
30
40
'Feels Like' Tempurature


Line Chart of Counts by Wind Speed
Count
1000
750
500
250
0
0
10
20
30
40
Wind Speed

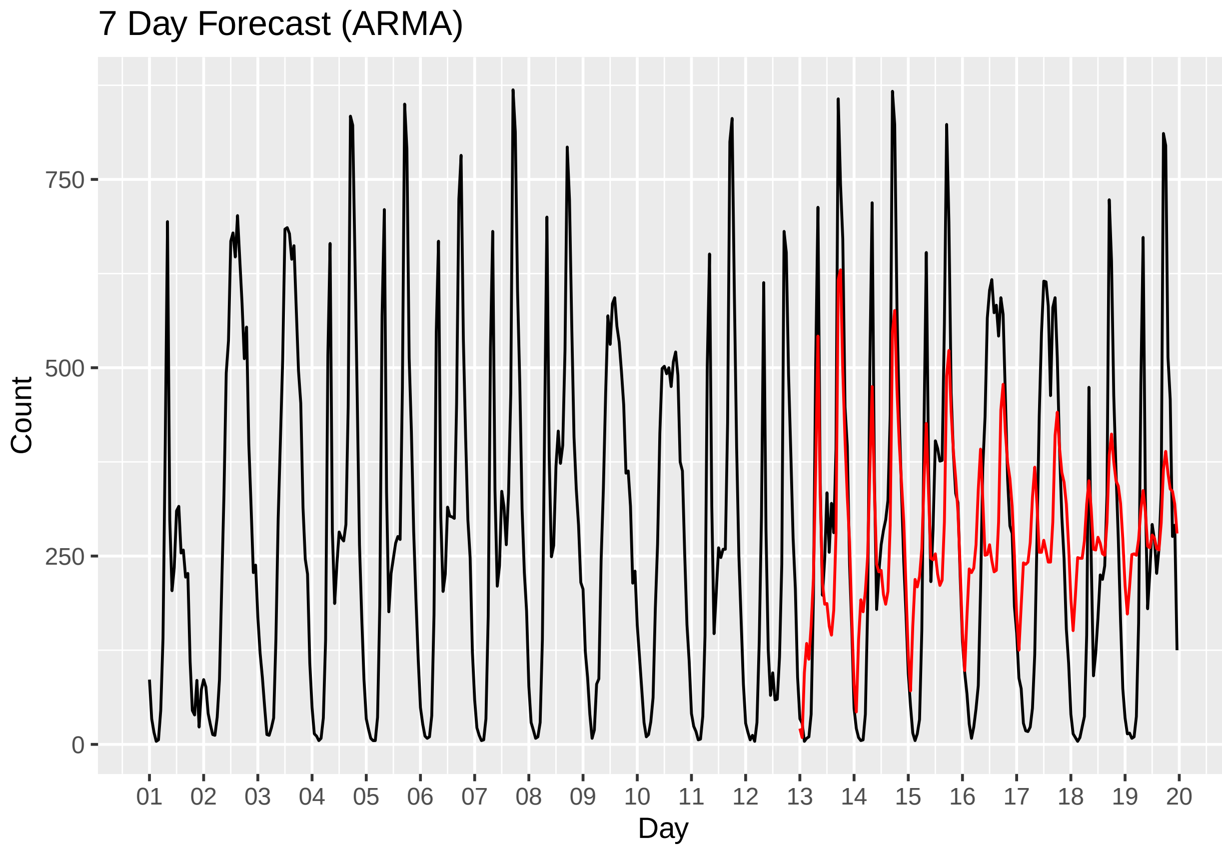


Line Chart of Counts by Humidity
1000
750
500
250
0
Count
0
25
50
75
100
Humidity

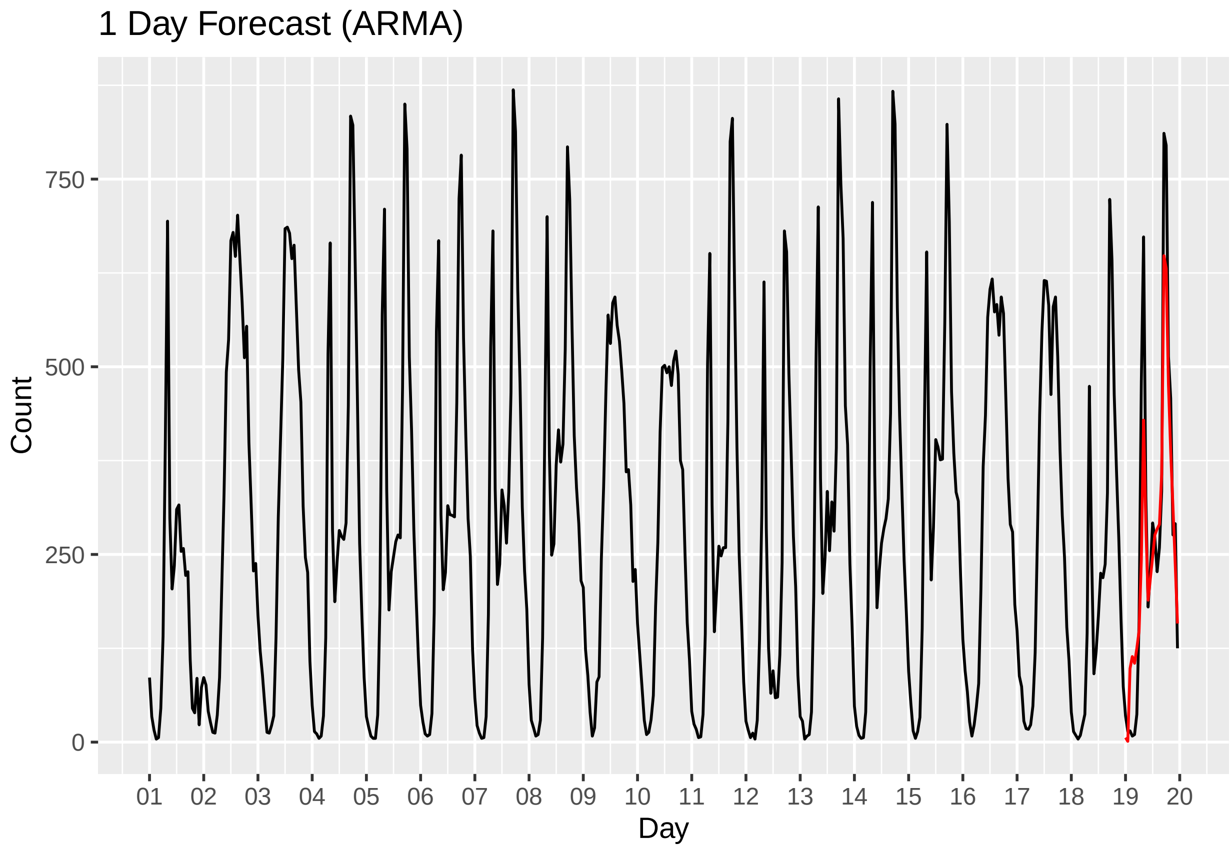

Hourly Rental Trends for June 1st − June 19th, 2012
Count
750
500
250
0
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
Day

3 Methods

3.1 ARMA Model



7 Day Forecast (ARMA)
Count
750
500
250
0
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
Day

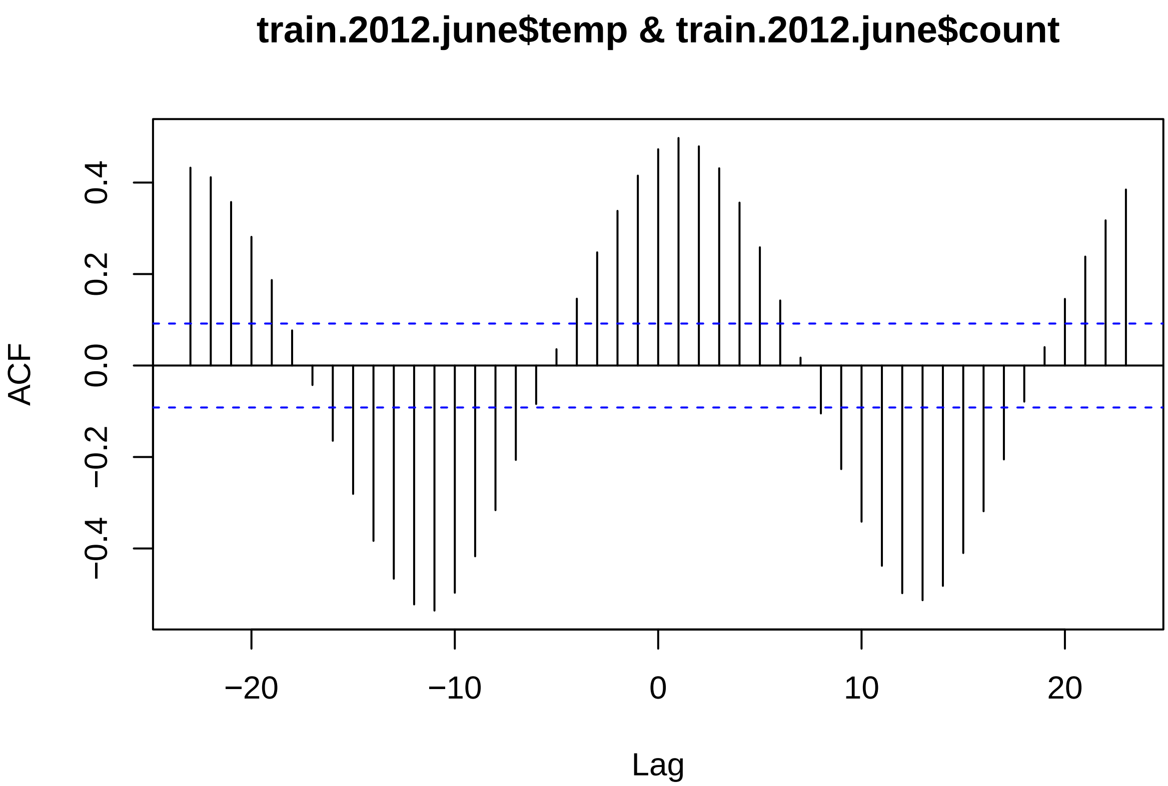


1 Day Forecast (ARMA)
Count
0
250
500
750
Day
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20

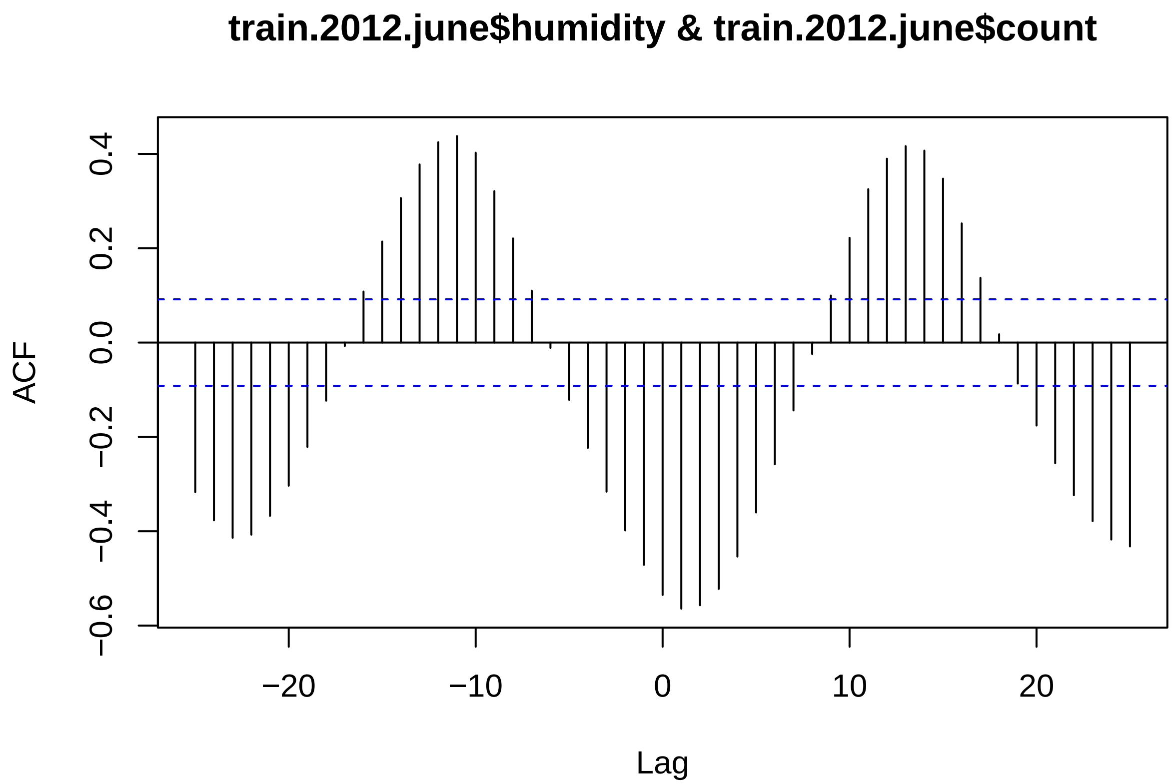
3.2 Vector Auto-Regressive (VAR) Model

3.2.1 Cross-Correlation

Temperature

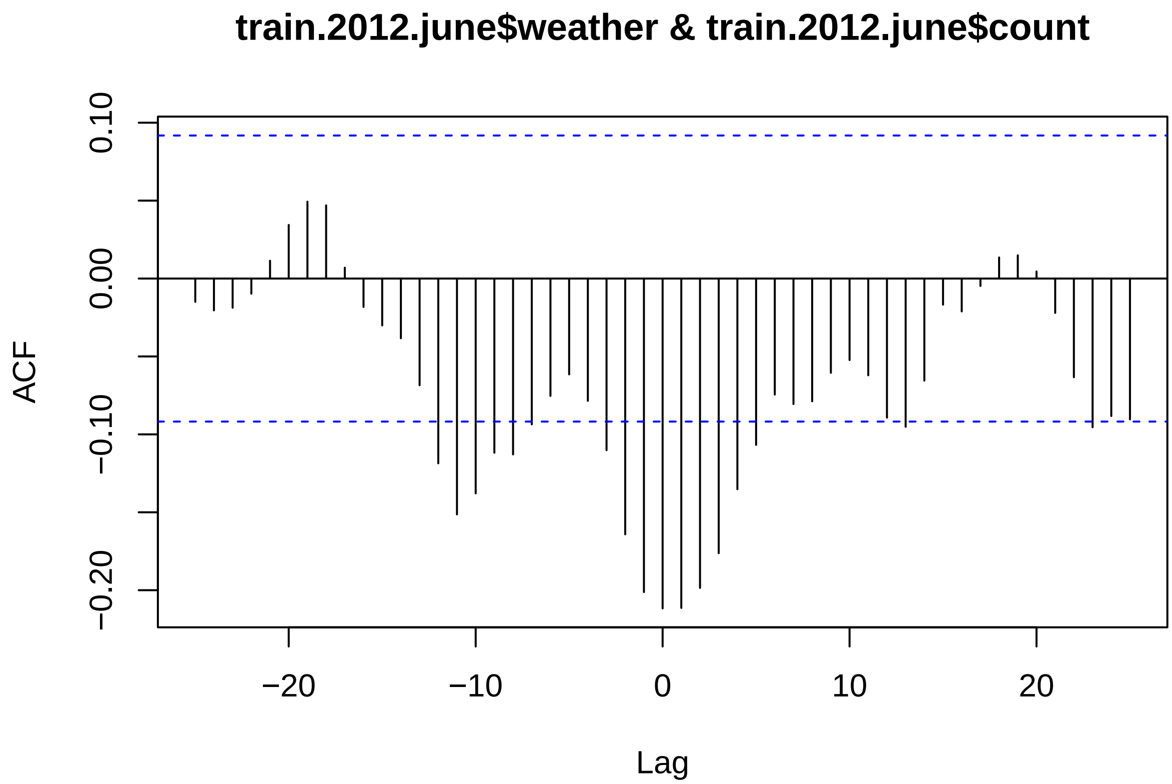


Humidity



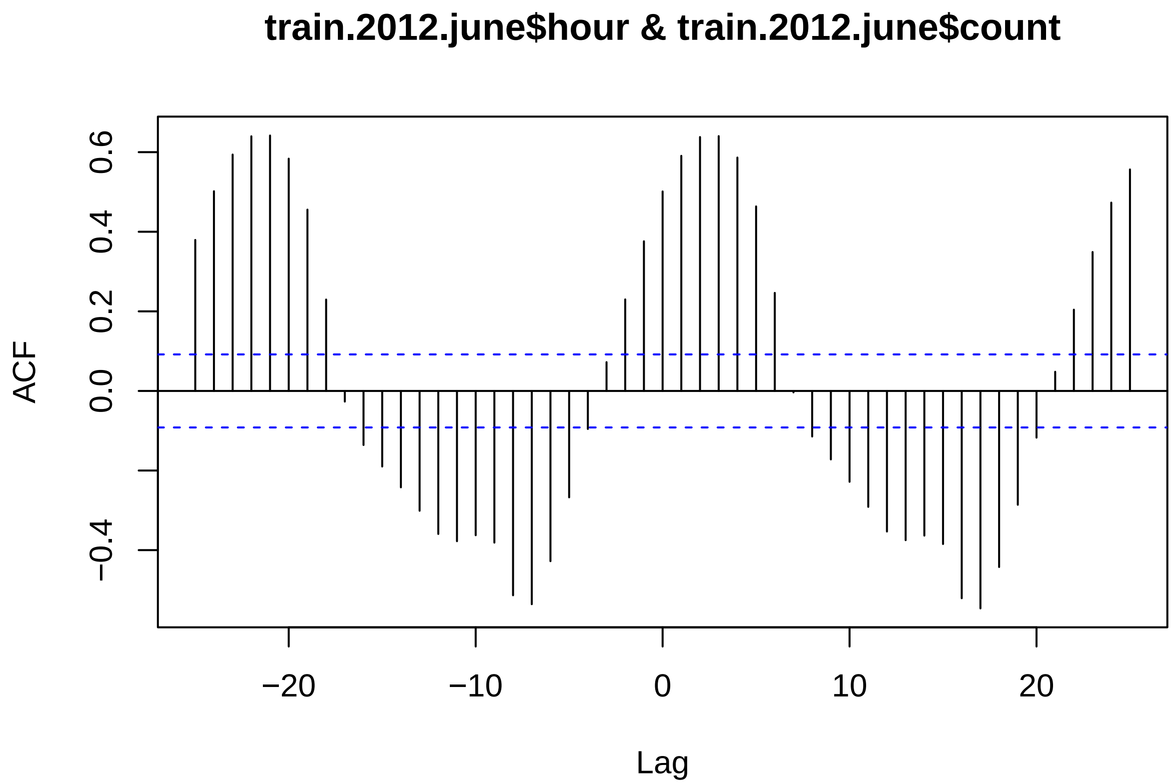
train.2012.june$humidity & train.2012.june$count
ACF
0.4
0.2
0.0
−0.2
−0.4
−0.6
−20
−10
0
10
20
Lag

Weather



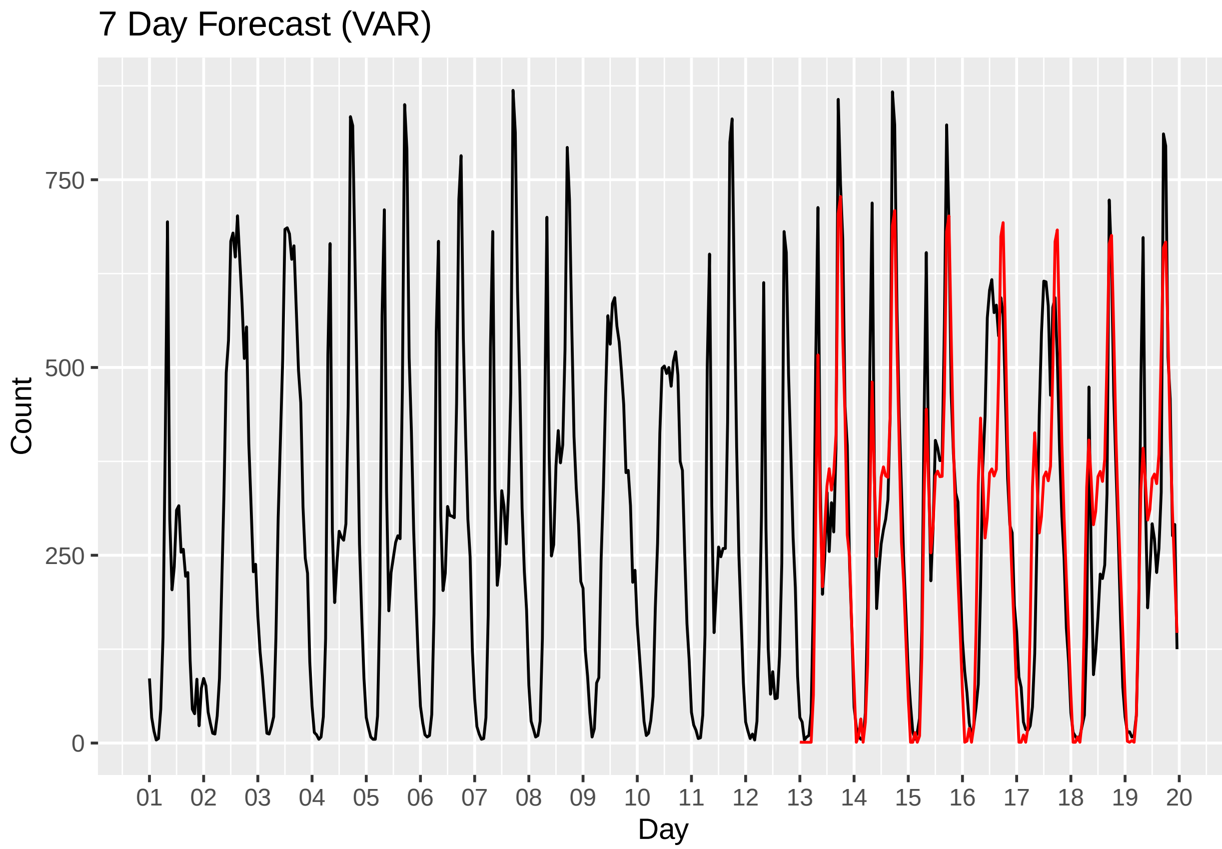
train.2012.june$weather & train.2012.june$count
ACF
0.10
0.00
−0.10
−0.20
−20
−10
0
10
20
Lag

Hour

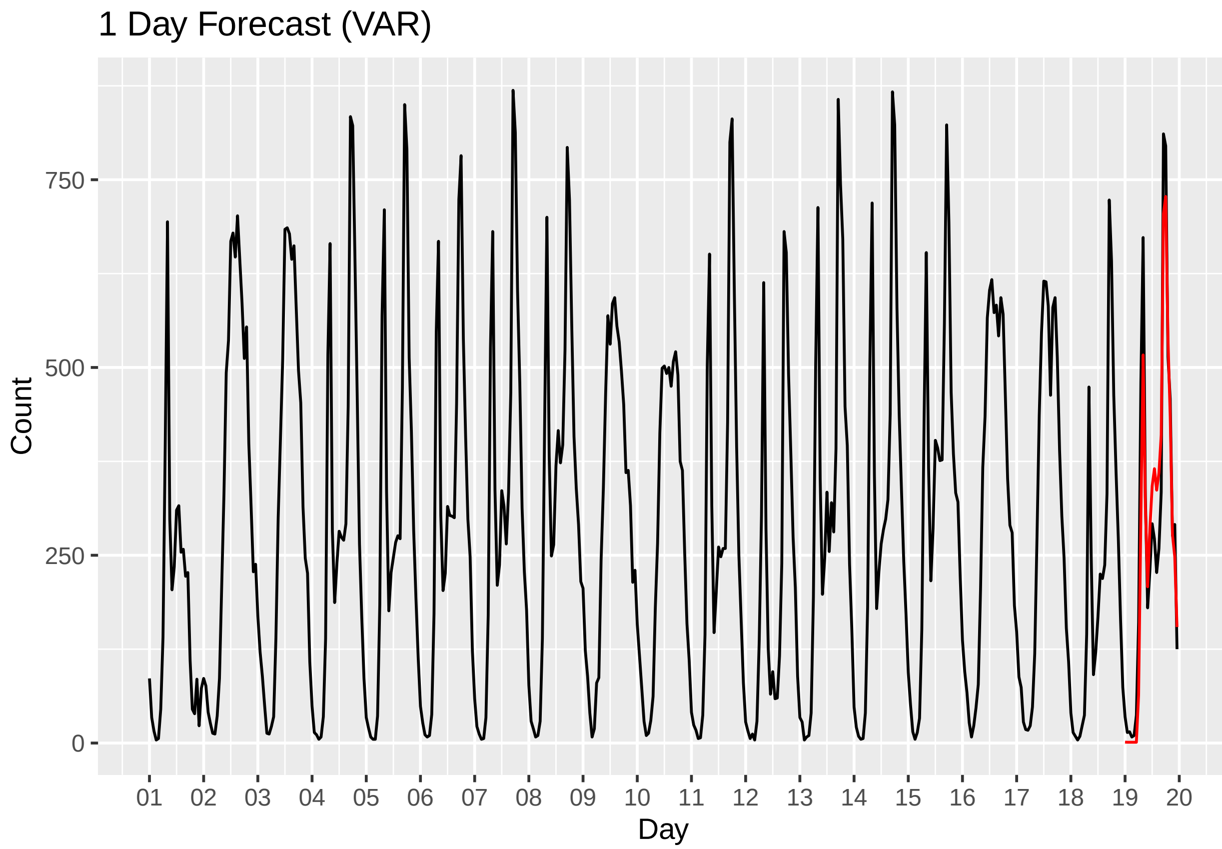


train.2012.june$hour & train.2012.june$count
ACF
0.6
0.4
0.2
0.0
−0.4
−20
−10
0
10
20
Lag

3.2.2 7 Day Forecast

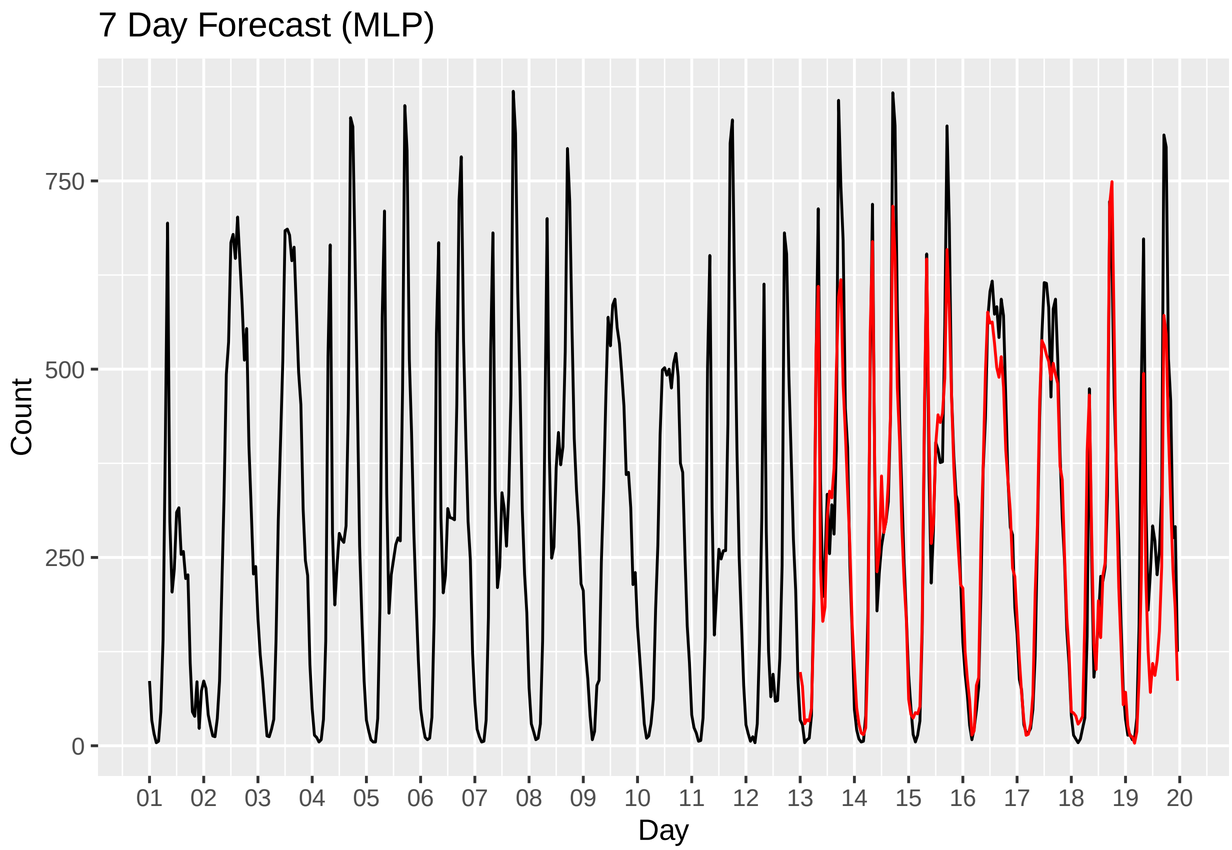


3.2.3 1 Day Forecast

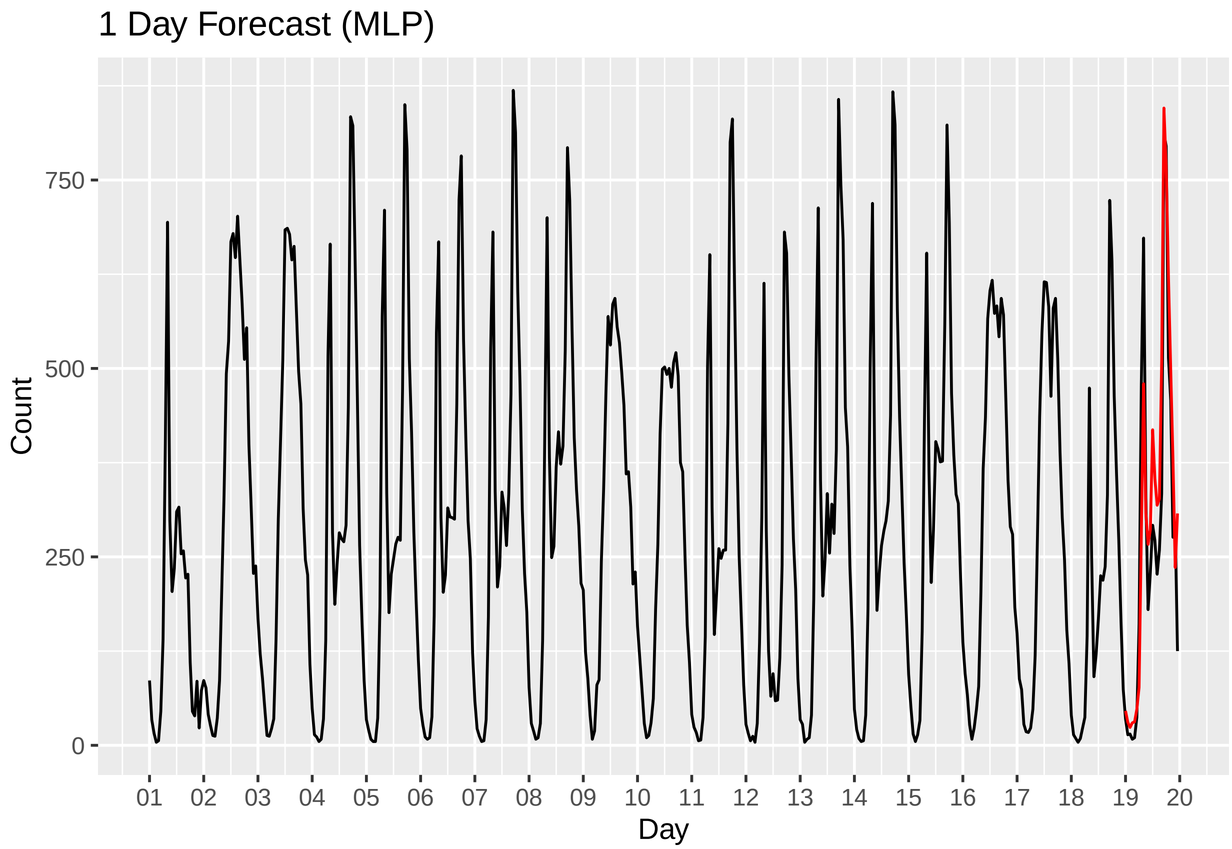


3.3 Neural Network Model

3.3.1 7 Day Forecast



3.3.2 1 Day Forecast



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

3.4.2 1 Day Forecast

	Actual	Predicted (ARMA)	Predicted (VAR)	Predicted (MLP)	Predicted (Ensemble)
433	35	6	1	46	23
434	14	1	1	31	16
435	15	98	1	24	12
436	8	114	1	30	15
437	10	105	1	31	16
438	37	125	1	48	24

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	9417.125
Ensemble	6446.292

4.2.2 RMSE

Model	RMSE
ARMA	96.94307
VAR	74.15496
MLP	97.04187
Ensemble	80.28880

5 Appendix

5.1 Solution One

Log Transform the response variable

5.2 Solution Two

Use all monthly data leading up to June 2012