Understanding the Weather and Seasonal effects on Bike Sharing Systems

Abhiraj Vinnakota

12/9/2019

Summary

A few sentences describing the inferential question(s), the method used and the most important results.

Understand the effects factors such as temperature, humidity as well uncovering the usage patterns on holiday or a particular day of the week etc, both on registered and casual users

GIT Link :

Introduction

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

It is of vital importance for us to understand the effects of seasonal and weather patterns on on the bike-sharing system to be able to progress towards event detection using this data, which this study hopes to serve. Some of the questions of interest include:

1> What is the extent of influence of weather and seasonal patterns on the the bike usage, if any?
2> Are there any differences of usage patterns between the registered and the casual users?

Data

The data was obtained from the UCI machine learning repository. It was used in the paper: **Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg**

This data was at a daily level for 2 consecutive years of 2011 and 2012. There are no missing values of any sort in the data

Distribution of the response variables:

The 2 response variables, registered and casual users' distribution was viewed. The distribution of the registered users seemed pretty normal. However, for the casual users, the square root transformation had to be applied to normalize the data.

The insights from EDA for the independent variables are briefly described below: (See Appendix 1.1 for the EDA plots not shown below)

Season : Seasons 2 and 3 (Spring and Summer) seem to have more users (both kinds) as compared to seasons 1 and 4.

Month : The distribution across months seemed to be a more zommed in view of the same trend that was observed in the 'season' variable.

Weekday: Casual users hardly seem to use the bikes during the work-week while the usage for registered users seem to be on the higher end during the work-days of the week.

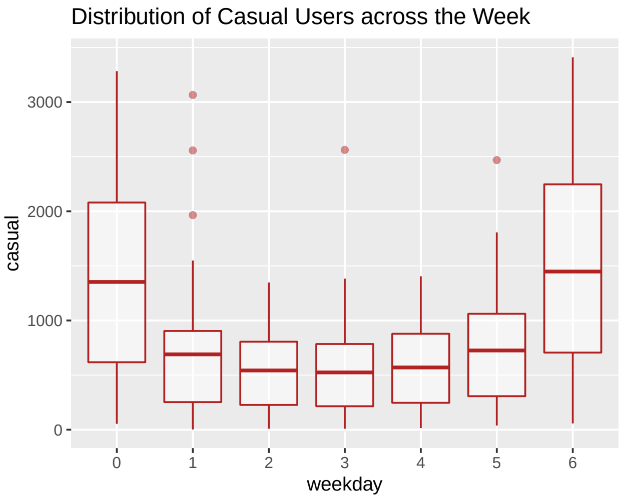
Weather variables (atemp, temp, humidity, windspeed) : The variables 'atemp' and 'temp' have a slightly positive correation with the users (both kinds) while 'humidity' and 'wind speed' seem to have a slightly negative correlation. This seems to be in accordance with the warmer months having more users as well as precipitation having a negative impact on bike usage.

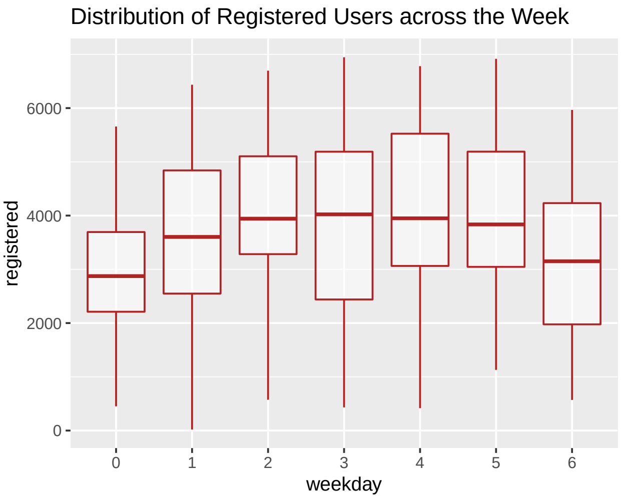
Working Day : This is a flag that indicates wheather that particular day is a working day or not (differentiatied weekdays from weekends). The registered users have most usage on working days while casual users use biked mostly on weekends. This is the same observation from the 'weekday' variable.

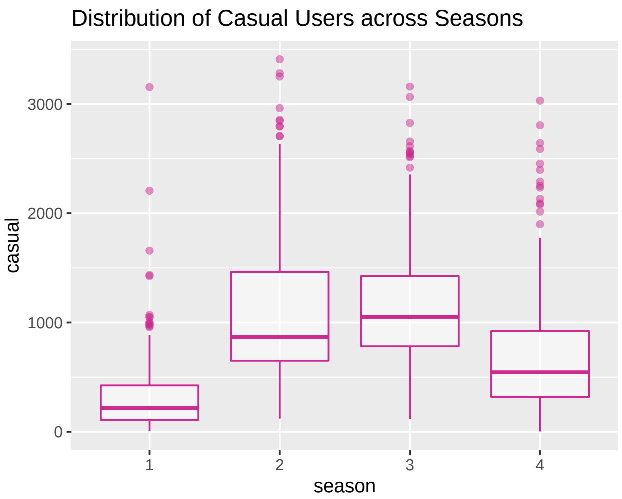
Holiday : This is a flag that indicated wheather that particular day is a holiday (special days) or not. Same inference from working day that was made could be made for this variable as well. Casual users are seen more on holidays while registered uers are seen more on non-holidays.

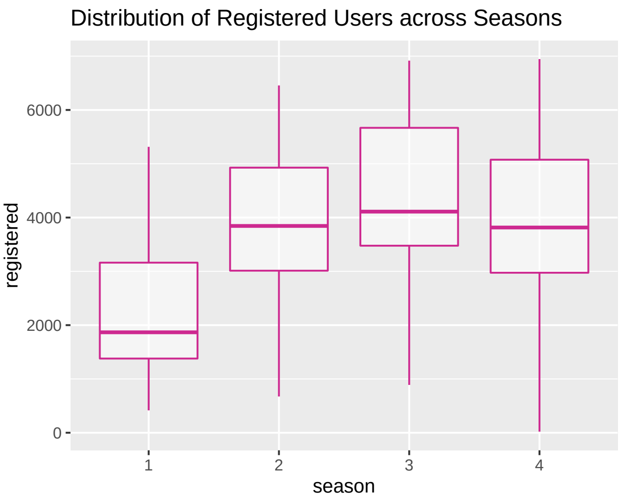
Weather Situation : This variable can take the values of 1,2,3 or 4 which indicate the following:
1: Clear, Few clouds, Partly cloudy, Partly cloudy
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

More worse the weather situation, lesser users (both kinds) were observed.


Distribution of Casual Users across the Week


Distribution of Registered Users across the Week





Model

Linear Regression Model Owing to explainability, a linear regression model was preferred. 2 seperate models were considered owing to the 2 response variables of interest. For both the models, a similar model building process was followed. No transformations were considered on the original variable for simplified interpretations (especially, since we are going to move to a time series model later on)

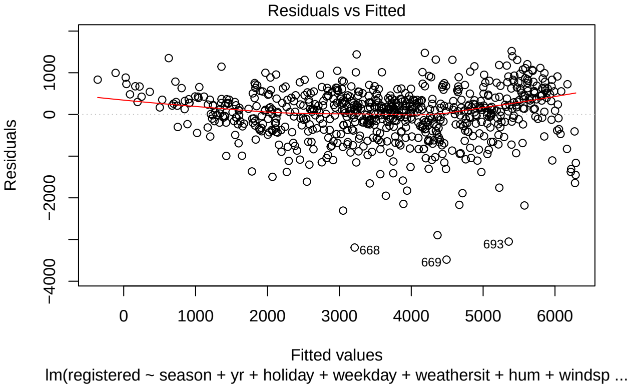
All variables were used to start with. Looking at the p-values, the variables, day('instant'), temp, workingday (wasn't converging) were removed in that order. Since, 'month' and 'season' were both explaining the same trend a call was taken to remove the 'month' variable (only some months were significant while all seasons were significant).

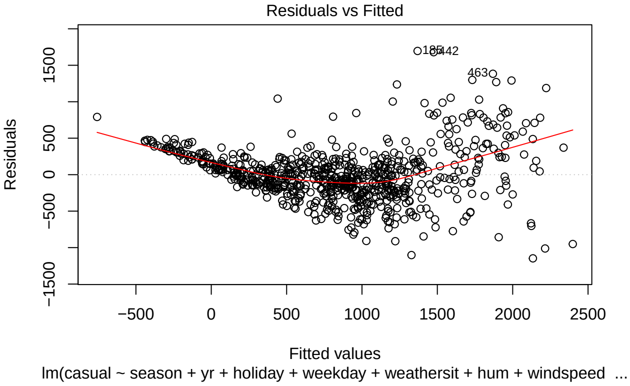
(See Appendix 1.2 for all the models)

Model Validation:

The VIFs were checked for both the models. They turned out to be fine. No issues were spotted (See Appendix 1.3 for VIF ouputs)

However, on viewing the residuals against the fitted values, there was a clear pattern that was observed within the data. We have used all the variables available to us in the model and still there is a trend that was clearly visible. This promted me to look at time series models as the data is ideal time series data.

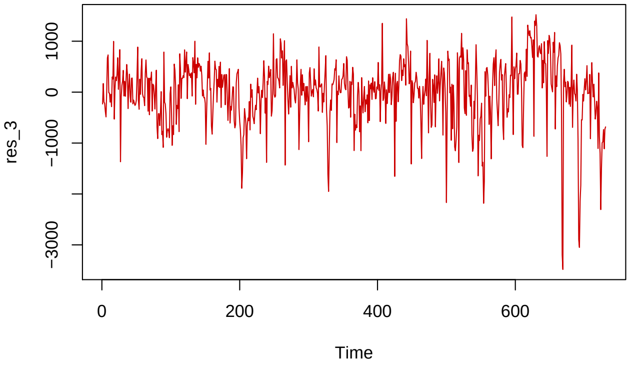


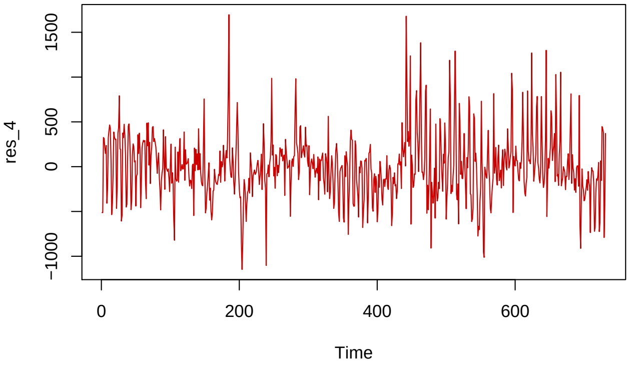


Time Series Model

Plotting the original data, we can clearly see the crests and troughs in the data, indicating the seasonal variation. The strong zig-zag nature of the curve is indicative of the weekly trend. The 2 crests are indicative of the 2 years of the data, with the middle trough seperating the 2 years. Clearly, the trends are non-stationary. (See Appendix 1.4)

Plotting the residuals of the models that were previously built.





These trends are appear stationary. The ADF and the KPSS tests were carried out on both the trends to be sure of stationarity.(Refer to Appendix 1.5)

1>Registered Users

Looking at the above plots, it is pretty clear that only the 1st lag seems to matter (epecially the PACF plot). the auto arima function gave an output of (1,0,0) as well. Hence, an AR1 model was used to describe this data.

Table 1: AR1 model for registered users

term	estimate	std.error
ar1	0.4980125	0.032042
intercept	-1.1269820	40.299995

[1] AIC: 11299.75

[2] Log Likelihood: -5646.87

2>Casual Users

The ACF plots suggest a strong weekly trend, the 7th lag seems to matter the most (same day in the previous week). The PACF plot suggests a similar trend.

The auto arima function gave an output of (2,0,3). Hence, going ahead with an AR2 MA3 model for the casual users.

Table 2: AR2MA3 model for casual users

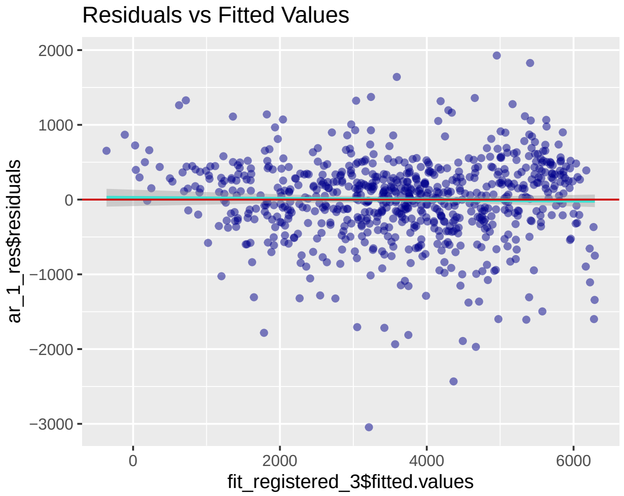
term	estimate	std.error
ar1	1.2051866	0.0140391
ar2	-0.9728781	0.0146501
ma1	-0.8170685	0.0367762
ma2	0.5689013	0.0419867
ma3	0.2774949	0.0363146
intercept	1.4801319	15.6425440

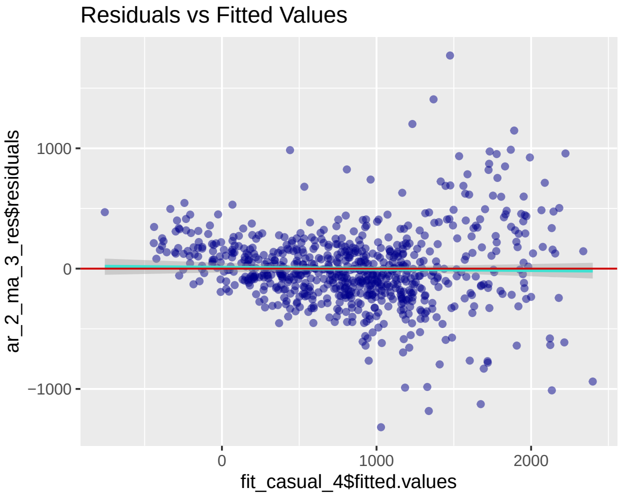
[1] AIC: 10502.26

[2] Log Likelihood: -5244.13

Model Validation

Looking at the plots of the residuals of the time series models against the fitted values of the linear models reveals more or less no pattern (far better than the earlier residual plots). This is indicative of the linear models doing a fine job, except that there was a certain autoregressiveness in the response variables which the linear models couldn't model which were taken care by the timeseries models. Hence, out interpretations of the linear models are fine.





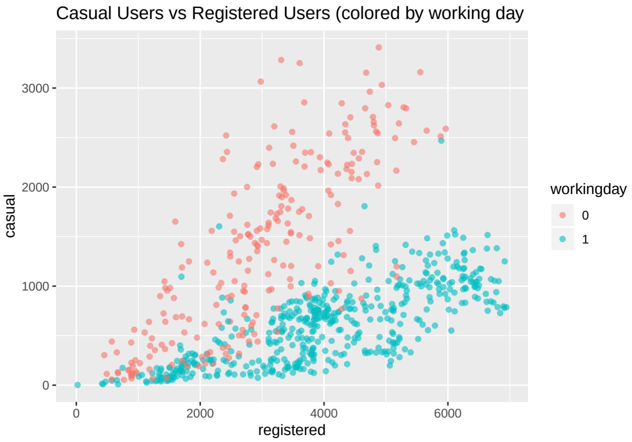
Results

Some of the inferences that we could get from the linear models are as follows:

1> The R squared for the registered users is at 0.8322 while that of casual users is at 0.713 which indicates that the registered user pattern is more dependent on weather and seasonal patterns as compared to casual users.
2> Every season is significantly different from the other, something that was indicative from the EDA.
3> The user base of these bikes has significantly increased in the 2nd year. The coefficient 1735 and 288 for the 'yr1' indicate the average increase in users in the 2nd year, as compared to the 1st.
4> From the coefficients, -1154 and 523 of 'holiday1' indicate the average dip and average increase in registered and casual users during a holiday as compared to a non-holiday.
5> The coefficients of weekdays 1 to 5 for registered users are around 1000, way higher than that of weekday 0 or 6, indicating the average usage during weekdays are much more as compared to weekends. Casual users observe an opposite trend.
6> The coefficients of 'weathersit' indicate a dip in the usage of both registered and casual users as the weather situation worsens.
7> The 'humidity', 'windspeed' and 'atemp' are all significant and normalized and hence need to be carefully interpreted. The coefficient of the 'atemp' variable for example indicates the increase that is expected beteen the day with the least temperature (normalized to 0) and the day with the maximum termperature (normalized to 1).

Limitations

The models used fail to capture the covariance between both the response variables. From the below graph, the association between both the variables is pretty clear. The working day variable does capture this relation is some aspect but fails to find a place in the linear models used due to dependence on other predictor variables.

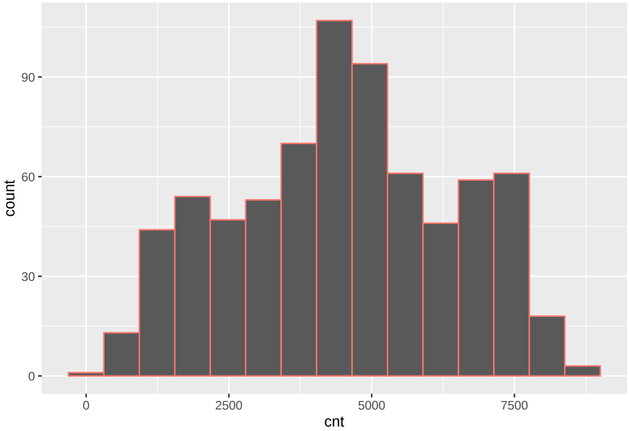


It will be interesting to use VAR (Vector Auto Regressive) models and see how to results change.

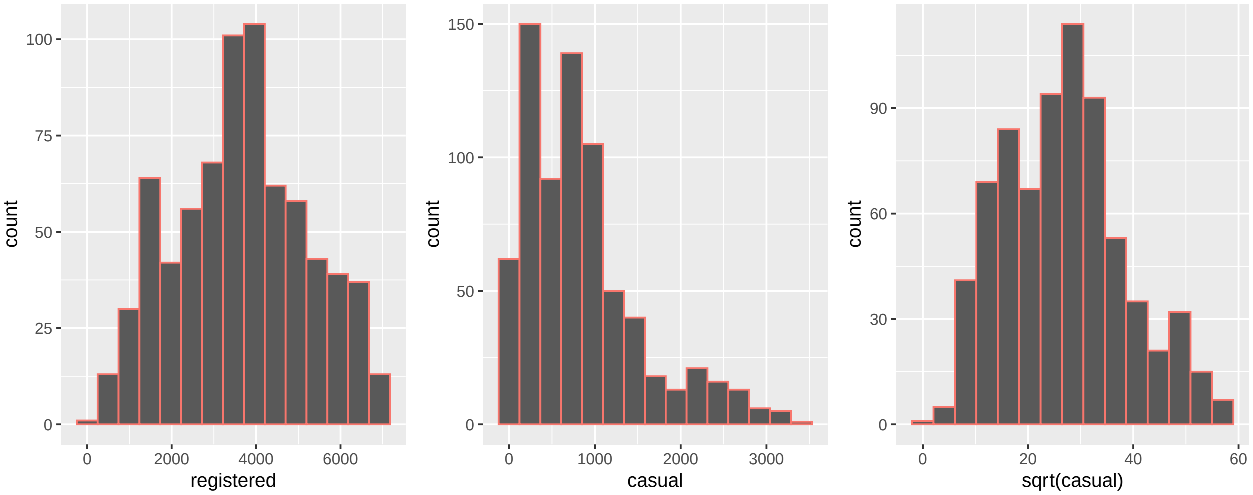
Appendix:

1.1 EDA

Distribution of the summation of both the response variables of interest:

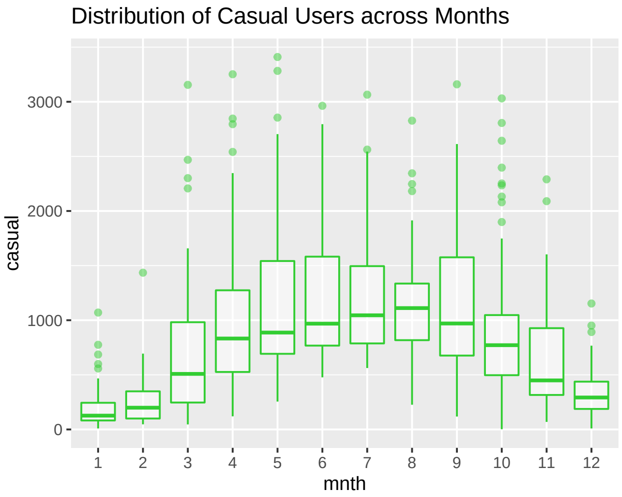


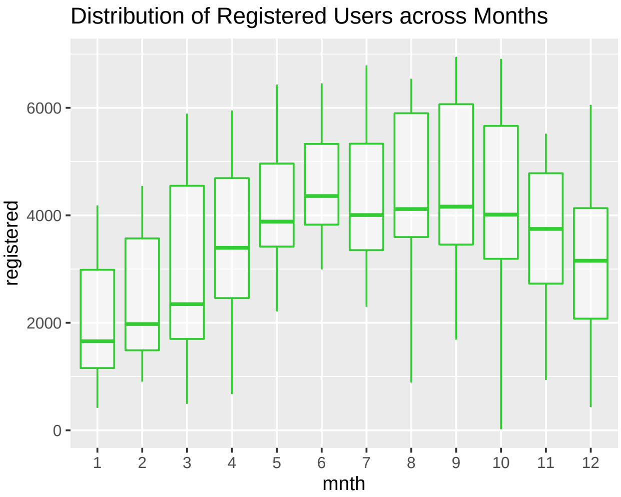
Distribution of the 2 response variables



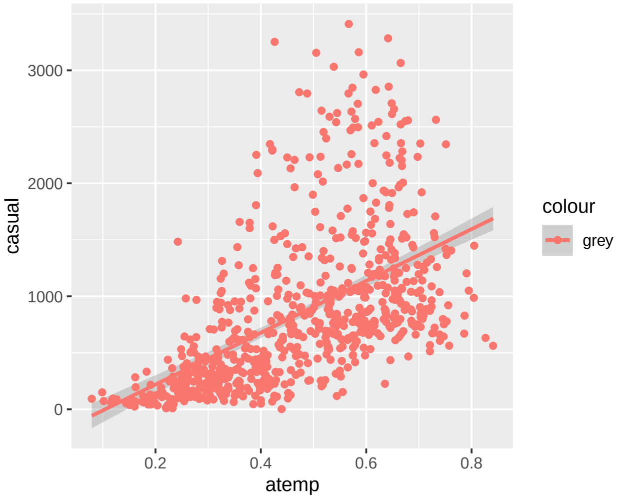
Independent Variables:

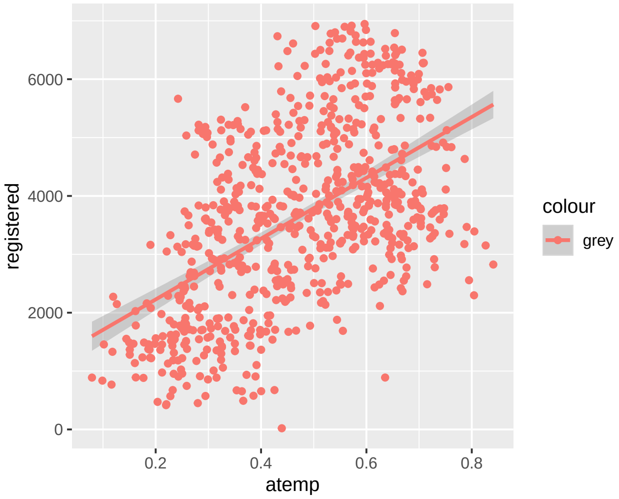
Month


Distribution of Casual Users across Months

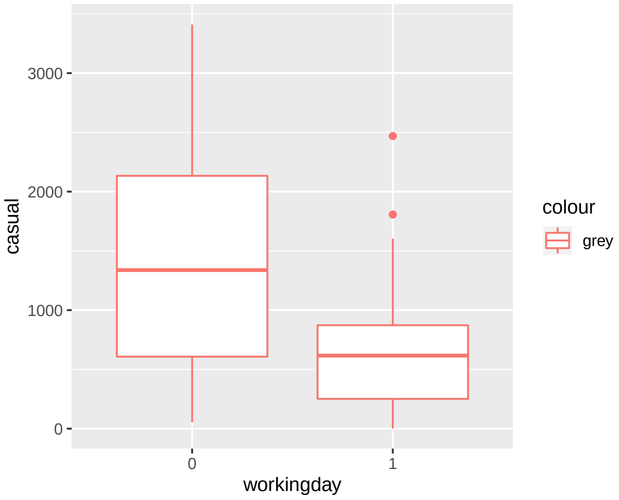

Distribution of Registered Users across Months

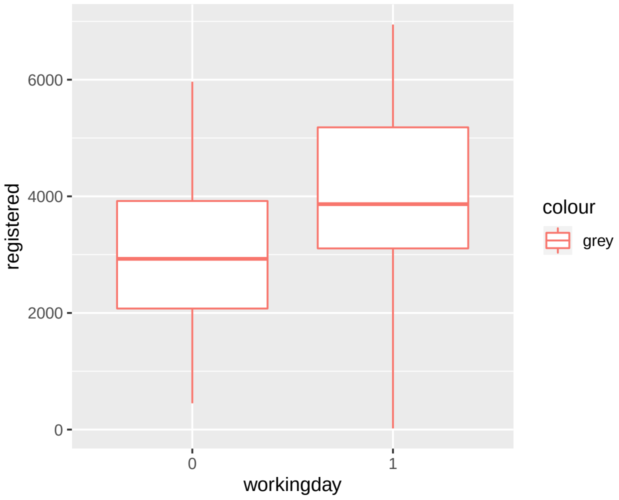
Feeling Temperature



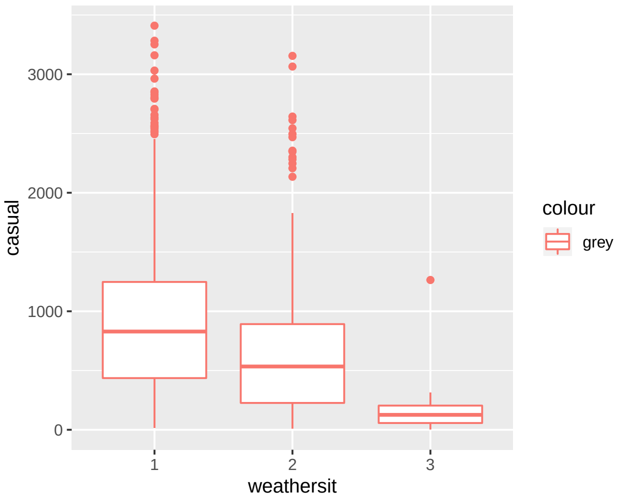


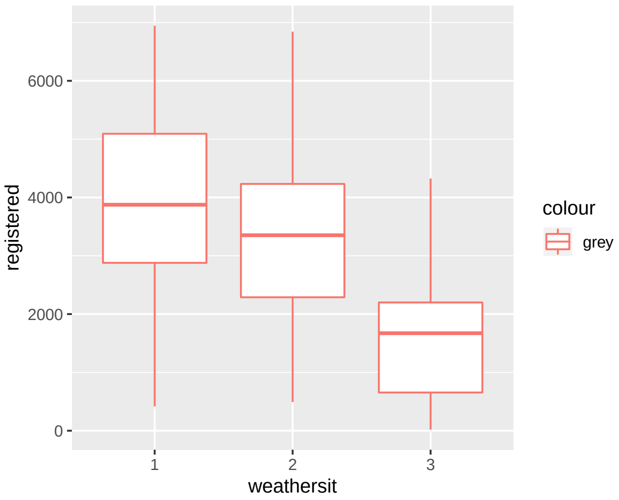
Working Day



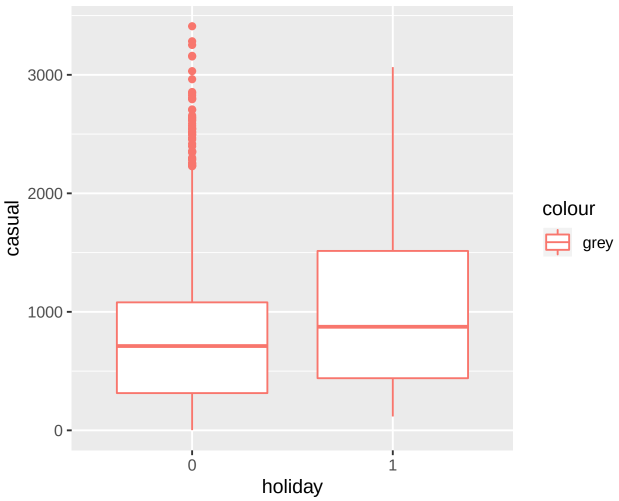


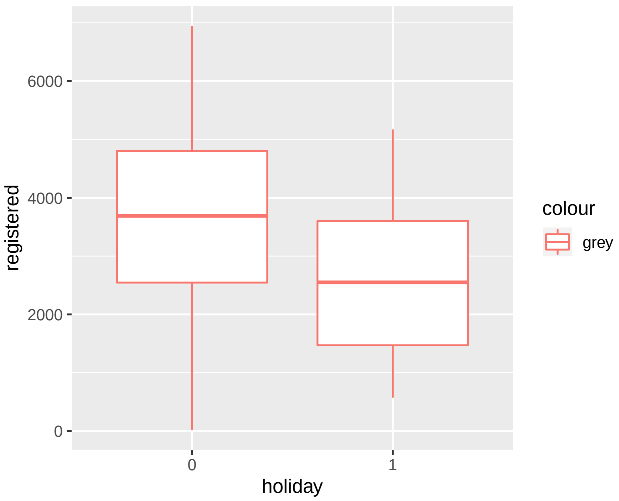
Weather Situation



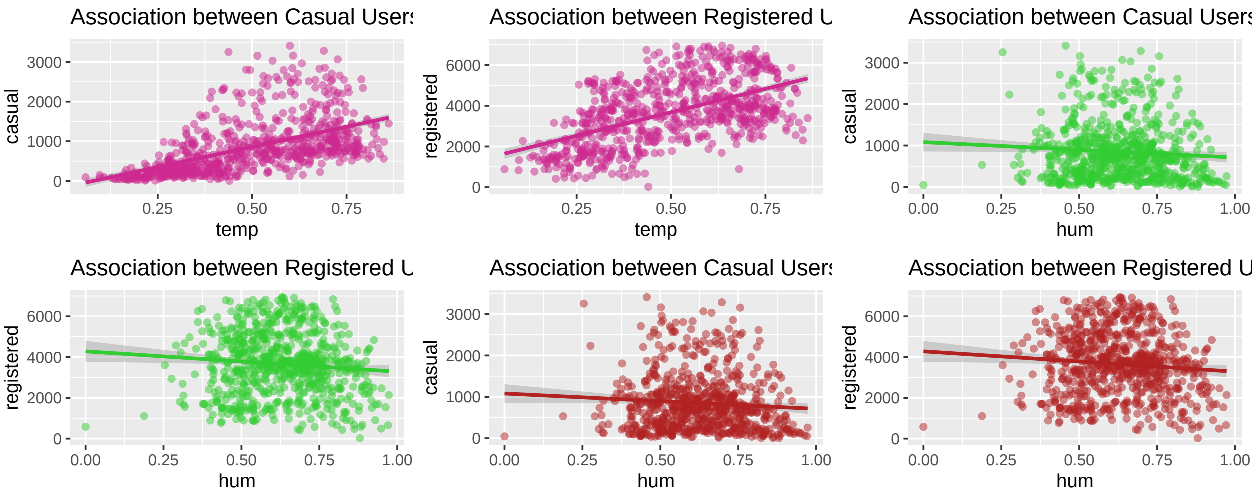


Holiday





Weather Variables (Temperature, Humidity & Wind Intensity



1.2 Linear Regression

Final Linear Model for Registered Users

Table 3: Linear Model for Registered Users

term	estimate	std.error	statistic	p.value
(Intercept)	558.2841	185.67934	3.006711	0.0027336
season2	809.0749	87.72508	9.222845	0.0000000
season3	806.3306	112.91144	7.141266	0.0000000
season4	1356.6674	75.30657	18.015260	0.0000000
yr1	1735.5498	47.99221	36.163153	0.0000000
holiday1	-1154.0150	147.86639	-7.804444	0.0000000
weekday1	948.2287	90.87405	10.434538	0.0000000
weekday2	1097.0539	88.79625	12.354733	0.0000000
weekday3	1178.5904	88.97880	13.245743	0.0000000
weekday4	1182.1364	88.95551	13.289074	0.0000000
weekday5	1063.0051	88.94372	11.951435	0.0000000
weekday6	294.9399	88.48515	3.333214	0.0009028
weathersit2	-347.1812	63.43347	-5.473154	0.0000001
weathersit3	-1655.3910	162.27262	-10.201296	0.0000000
hum	-796.0408	231.14959	-3.443834	0.0006070
windspeed	-1614.5700	333.33853	-4.843635	0.0000016
atemp	3523.1209	260.76792	13.510561	0.0000000

[1] Adjusted R-squared : 0.8322

[2] Multiple R-Squared : 0.8358

Final Linear Model for Casual Users

1.4 Model Validation

```
# no problems with VIF
vif(fit_registered_3)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
```

Table 4: Linear Model for Casual Users

term	estimate	std.error	statistic	p.value
(Intercept)	584.1675	106.85418	5.466959	0.0000001
season2	365.5437	50.48376	7.240818	0.0000000
season3	161.5608	64.97794	2.486395	0.0131320
season4	187.2511	43.33719	4.320794	0.0000178
yr1	288.3673	27.61841	10.441123	0.0000000
holiday1	523.1523	85.09371	6.147956	0.0000000
weekday1	-744.6678	52.29592	-14.239500	0.0000000
weekday2	-798.6127	51.10020	-15.628369	0.0000000
weekday3	-800.4196	51.20525	-15.631593	0.0000000
weekday4	-788.8406	51.19185	-15.409497	0.0000000
weekday5	-607.5410	51.18506	-11.869498	0.0000000
weekday6	151.9342	50.92117	2.983715	0.0029449
weathersit2	-101.1363	36.50450	-2.770515	0.0057423
weathersit3	-292.4595	93.38416	-3.131789	0.0018084
hum	-450.7846	133.02127	-3.388816	0.0007405
windspeed	-882.8615	191.82865	-4.602345	0.0000049
atemp	1961.6873	150.06593	13.072169	0.0000000

[1] Adjusted R-squared : 0.713

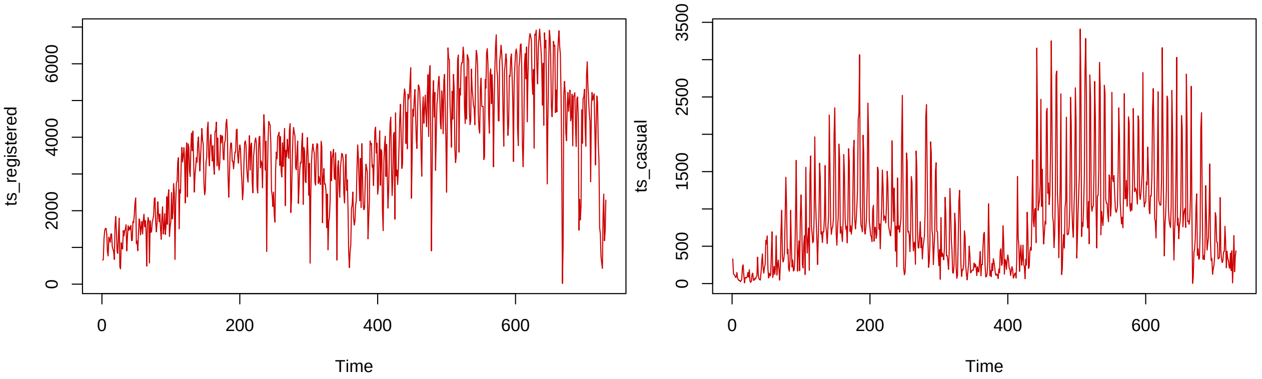
[2] Multiple R-Squared : 0.7193

```
## season     3.303088  3         1.220356
## yr         1.030220  1         1.014997
## holiday    1.091518  1         1.044757
## weekday    1.131766  6         1.010368
## weathersit 1.833911  2         1.163709
## hum        1.936593  1         1.391615
## windspeed  1.192354  1         1.091950
## atemp      3.226497  1         1.796245
```

```
# no problems with VIF
vif(fit_casual_4)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## season     3.303088  3         1.220356
## yr         1.030220  1         1.014997
## holiday    1.091518  1         1.044757
## weekday    1.131766  6         1.010368
## weathersit 1.833911  2         1.163709
## hum        1.936593  1         1.391615
## windspeed  1.192354  1         1.091950
## atemp      3.226497  1         1.796245
```

2.1 Actual Data Time Series Plots



2.2 Stationarity tests on the residuals of the linear models

Registered Users Linear Model residuals

```
#Tests for stationarity
adf_test <- adf.test(res_3, alternative = 'stationary')
```

```
## Warning in adf.test(res_3, alternative = "stationary"): p-value smaller
## than printed p-value
```

```
print(adf_test)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  res_3
## Dickey-Fuller = -5.5608, Lag order = 9, p-value = 0.01
## alternative hypothesis: stationary
```

```
#note that the alternative hypothesis here is "stationary"
#so that low p-values support stationarity
# p = 0.01 , hence supports stationarity

kpss_test <- kpss.test(res_3)
```

```
## Warning in kpss.test(res_3): p-value greater than printed p-value
```

```
print(kpss_test)
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  res_3
## KPSS Level = 0.090064, Truncation lag parameter = 6, p-value = 0.1
```

```
#by the way, KPSS stands for Kwiatkowski-Philips-Schmidt-Shin
#here, the null hypothesis is actually "stationary"
#so that high p-values support stationarity
# p = 0.1 , hence supports stationarity
```

Casual Users Linear Model residuals

```r
#Tests for stationarity
adf_test <- adf.test(res_4, alternative = 'stationary')
```

```
## Warning in adf.test(res_4, alternative = "stationary"): p-value smaller
## than printed p-value
```

```r
print(adf_test)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  res_4
## Dickey-Fuller = -5.8356, Lag order = 9, p-value = 0.01
## alternative hypothesis: stationary
```

```r
#note that the alternative hypothesis here is "stationary"
#so that low p-values support stationarity
# p = 0.01 , hence supports stationarity

kpss_test <- kpss.test(res_4)
```

```
## Warning in kpss.test(res_4): p-value greater than printed p-value
```

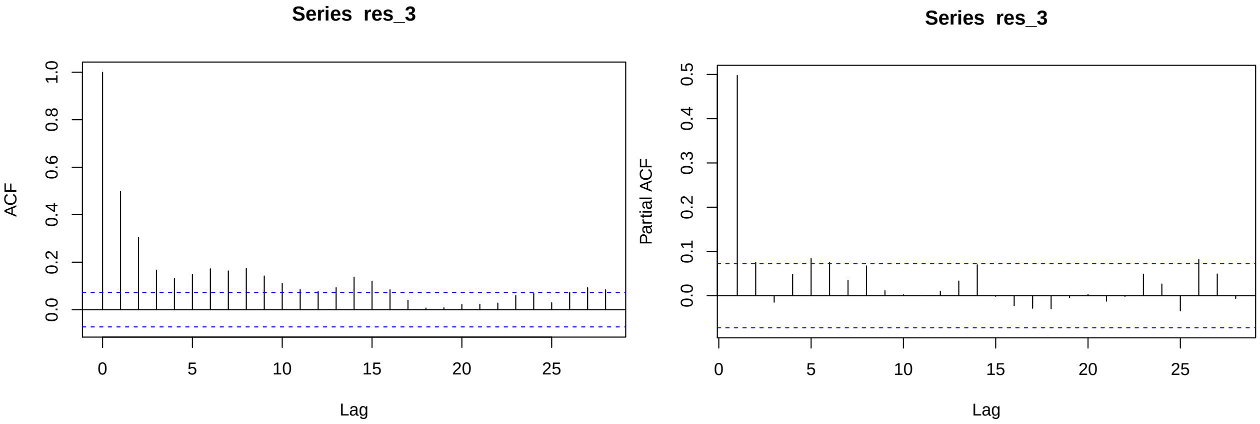
```r
print(kpss_test)
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  res_4
## KPSS Level = 0.11878, Truncation lag parameter = 6, p-value = 0.1
```

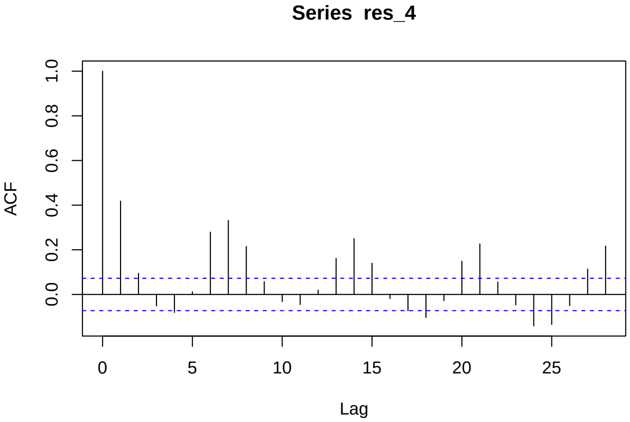
```r
#by the way, KPSS stands for Kwiatkowski-Philips-Schmidt-Shin
#here, the null hypothesis is actually "stationary"
#so that high p-values support stationarity
# p = 0.1 , hence supports stationarity
```

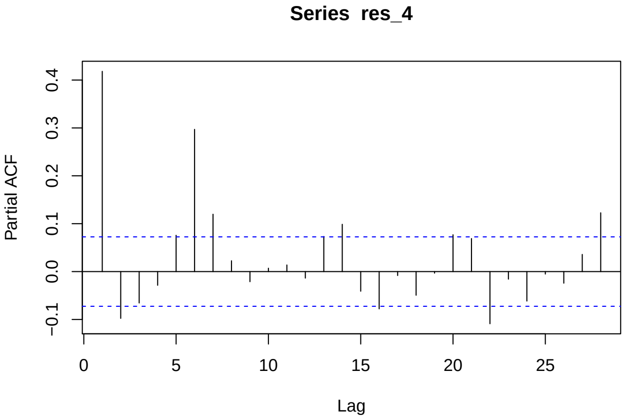
2.3 ACF and PACF plots for residuals of the linear models

Registered Users Linear Model residuals



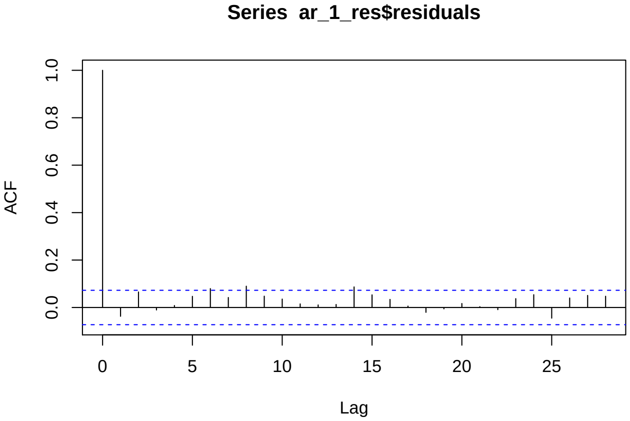
Casual Users Linear Model residuals

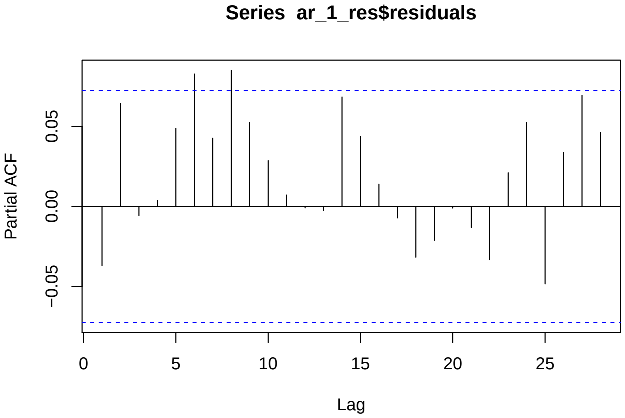




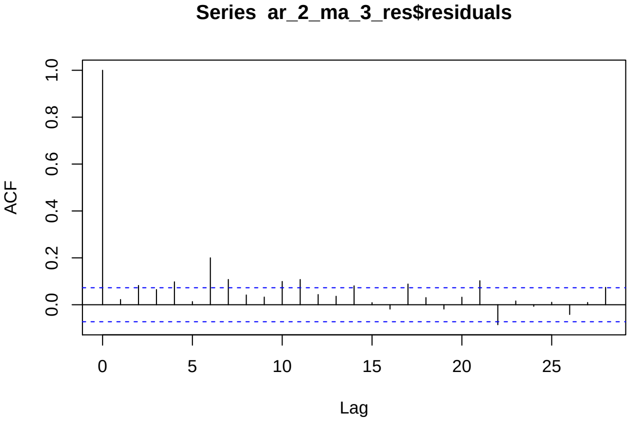
2.4 ACF and PACF plots for the residuals of the Time Series Models

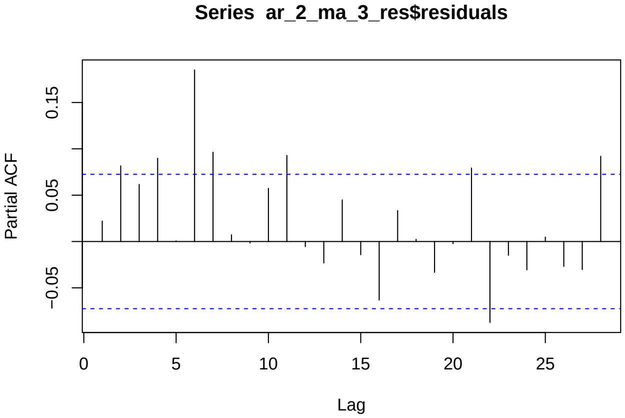
Registered Users AR1 model residuals





Casual Users AR2 MA3 Model residuals


Series ar_2_ma_3_res$residuals
ACF
Lag


Series ar_2_ma_3_res$residuals
Partial ACF
Lag