A Project on Exploratory Data Analysis

Narendran Santhanam

About the dataset

I chose the bikeshare dataset from the UCI machine learning repository found here: https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset (https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset). This dataset contains the hourly count of rental bikes between years 2011 and 2012 in Capital bikeshare system, with the corresponding weather and seasonal information.

Dataset Overview

Let us get a high level overview of the dataset and its attributes.

```
## No. of columns: 17
```

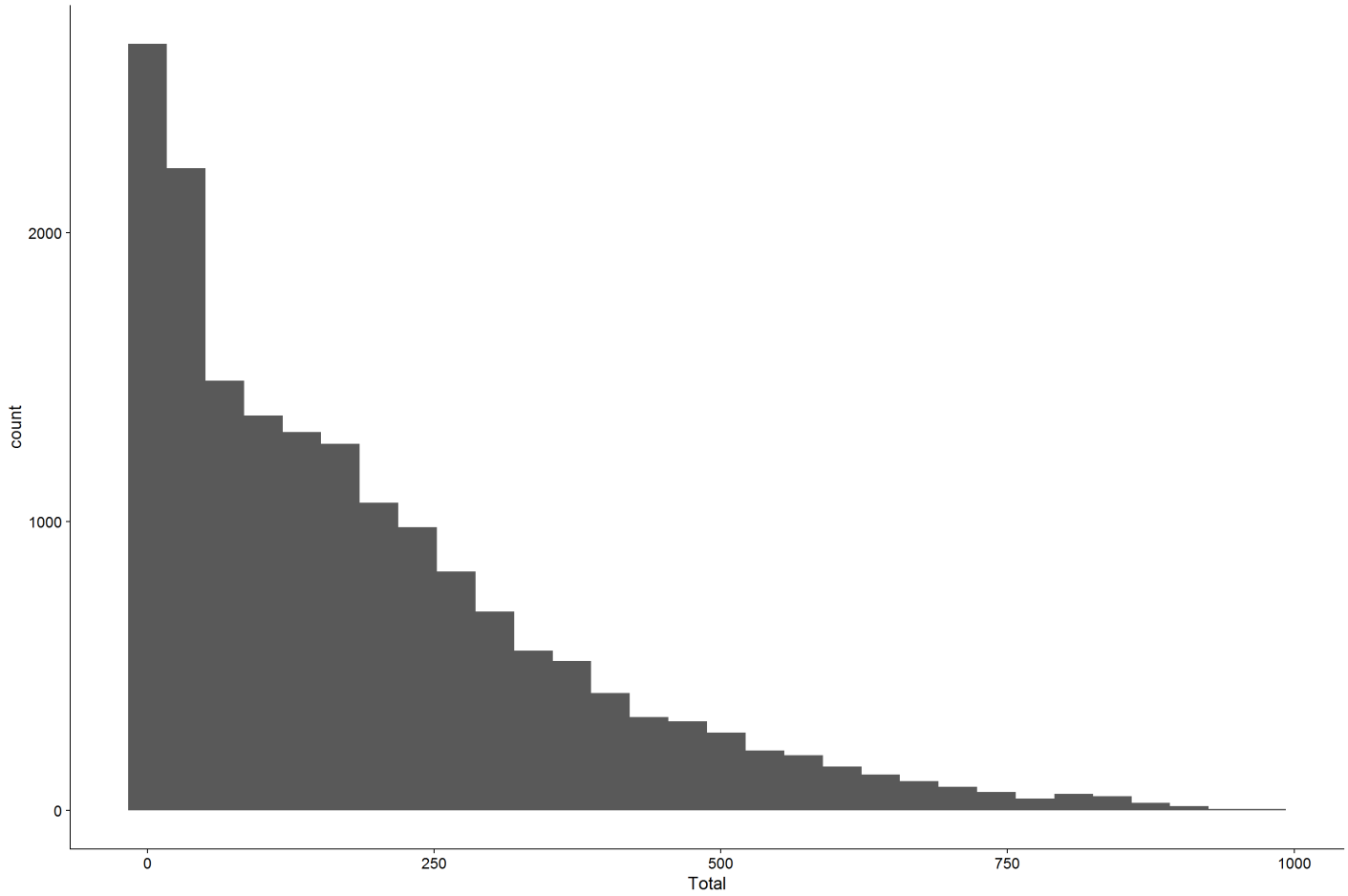
```
## No. of rows   : 17379
```

```
## 'data.frame':    17379 obs. of  16 variables:
##  $ Date      : Date, format: "2011-01-01" "2011-01-01" ...
##  $ Season    : Factor w/ 4 levels "Spring","Summer",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Year      : Ord.factor w/ 2 levels "2011"<"2012": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Month     : Ord.factor w/ 12 levels "1"<"2"<"3"<"4"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Hour      : Ord.factor w/ 24 levels "0"<"1"<"2"<"3"<..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Holiday   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Weekday   : Factor w/ 7 levels "Sunday","Monday",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ WorkingDay: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Weather   : Factor w/ 4 levels "Clear with a few clouds",..: 1 1 1 1 1 2 1 1 1 1 ...
##  $ Humidity  : num  0.81 0.8 0.8 0.75 0.75 0.75 0.8 0.86 0.75 0.76 ...
##  $ WindSpeed : num  0 0 0 0 0 ...
##  $ Casual    : int  3 8 5 3 0 0 2 1 1 8 ...
##  $ Registered: int  13 32 27 10 1 1 0 2 7 6 ...
##  $ Total     : int  16 40 32 13 1 1 2 3 8 14 ...
##  $ Fahrenheit: num  37.9 36.2 36.2 37.9 37.9 ...
##  $ TimeOfDay : Factor w/ 3 levels "Daytime","Evening",..: 3 3 3 3 3 3 3 3 1 1 ...
```

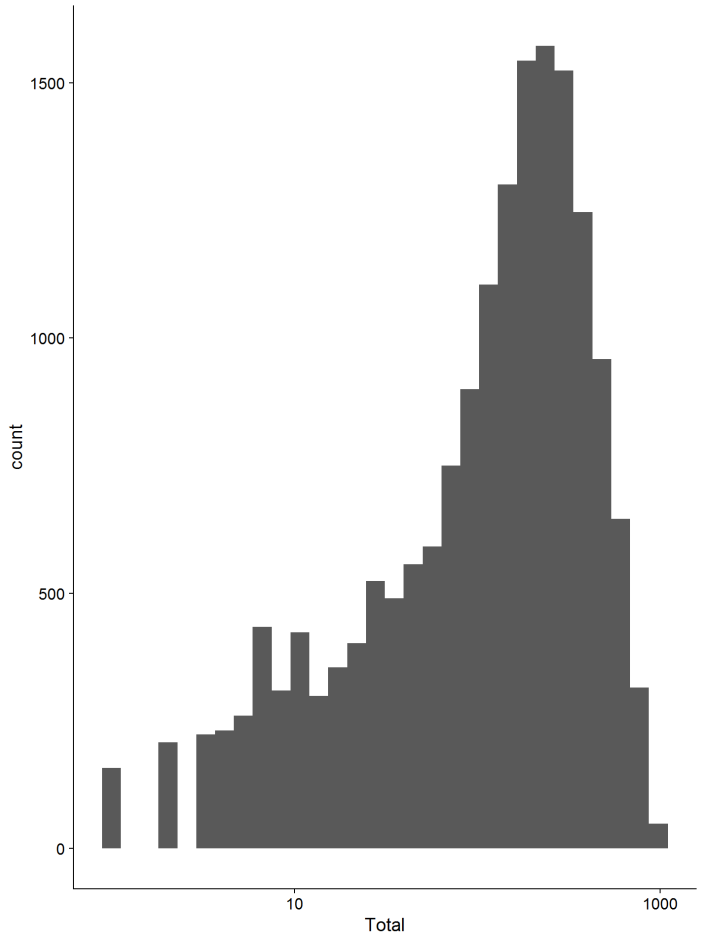
Our dataset contains 16 variables (instant is an ID column) and more than 17000 observations.

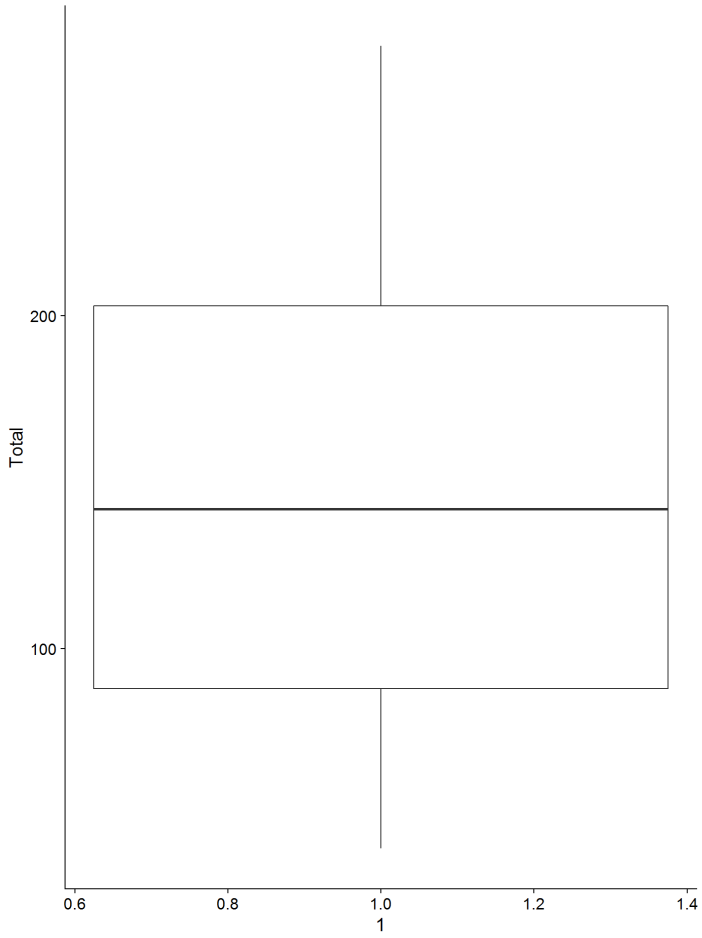
Univariate Plots

Our variables of interest are Total, Registered and Casual. Total is the total count of rental bikes for a certain hour of the day, which is the sum of Registered bikes (bikes rented by regular subscribers) and Casual rental bikes. Let us take a look at the distribution of these variables.



The histogram looks skewed to the right. Let’s apply a log transformation to see the distribution. I’ll also plot Total on a boxplot and remove outliers to analyze the range in which it lies.

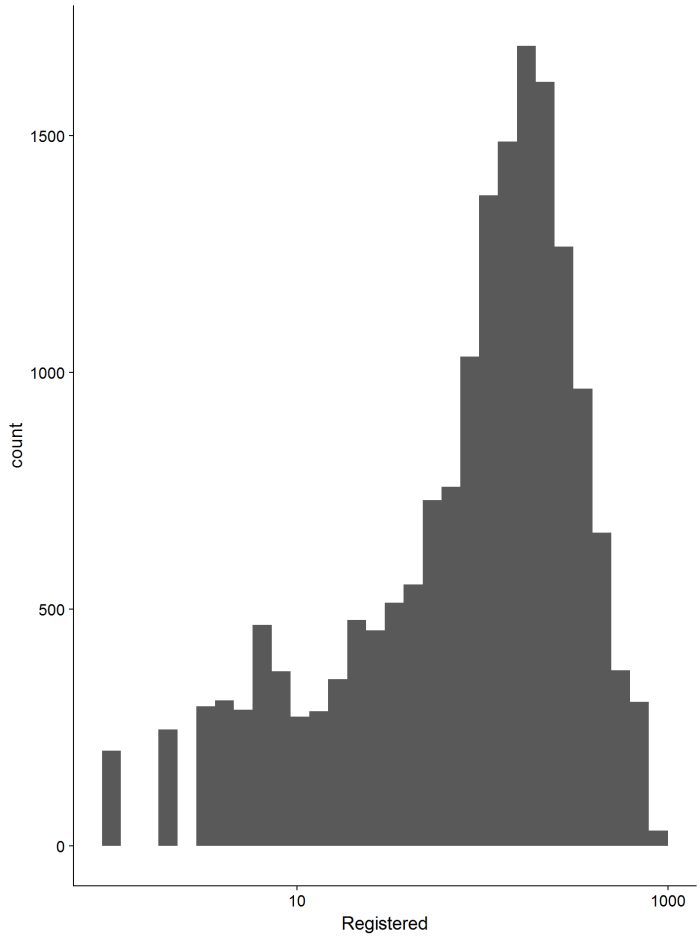


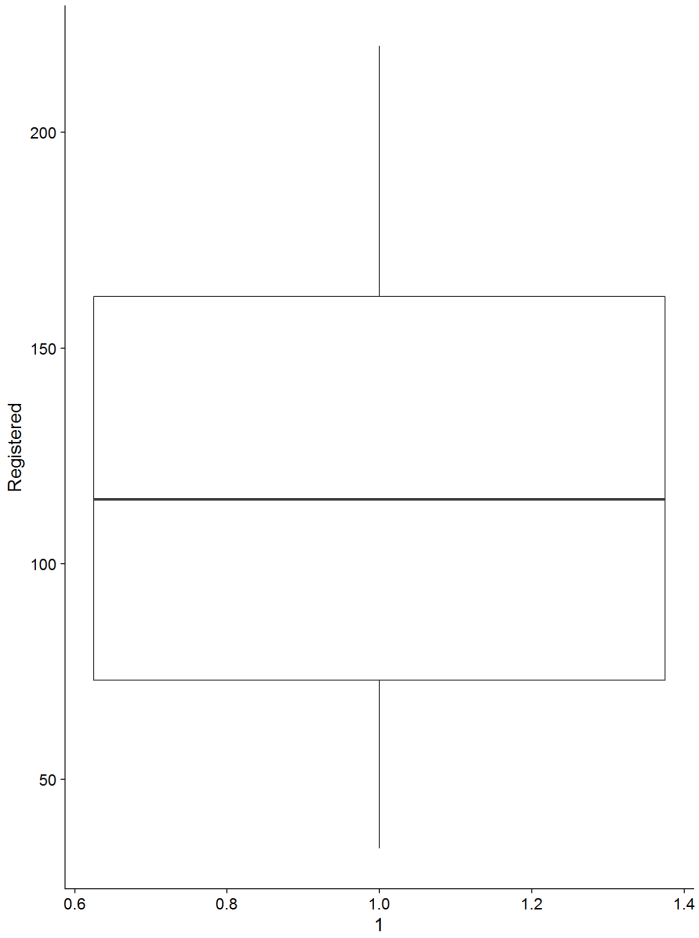



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    40.0   142.0   189.5   281.0   977.0
```

That looks better. 40-300 seems to be the most common range.

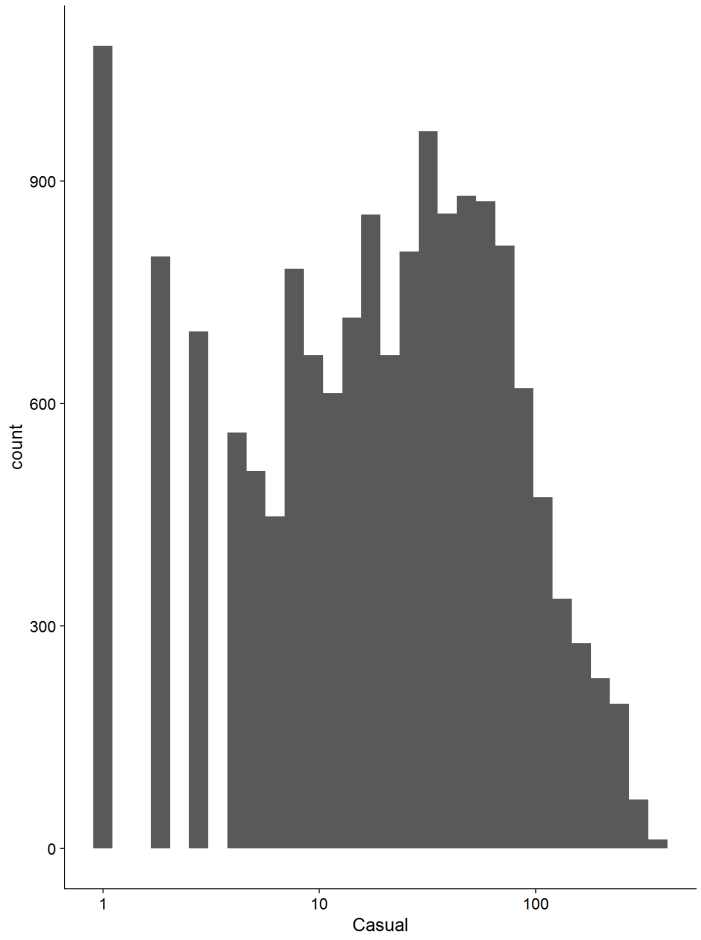
Let’s plot the Registered and Casual variables with the same transformation and boxplots with outliers excluded.

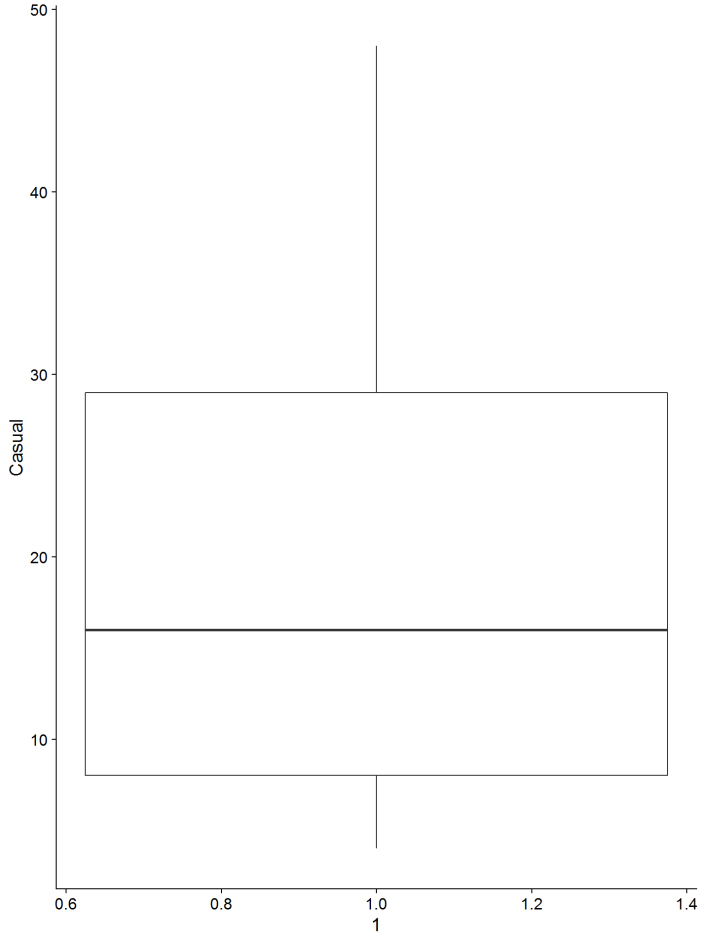





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    34.0   115.0   153.8   220.0   886.0
```

“Registered” seems to have an almost identical distribution as “Total”. Even the numbers seem pretty close to “Total”. Does this mean that the number of casual bike renters was pretty low, and that adding the casual numbers to registered numbers did not affect the distribution significantly?



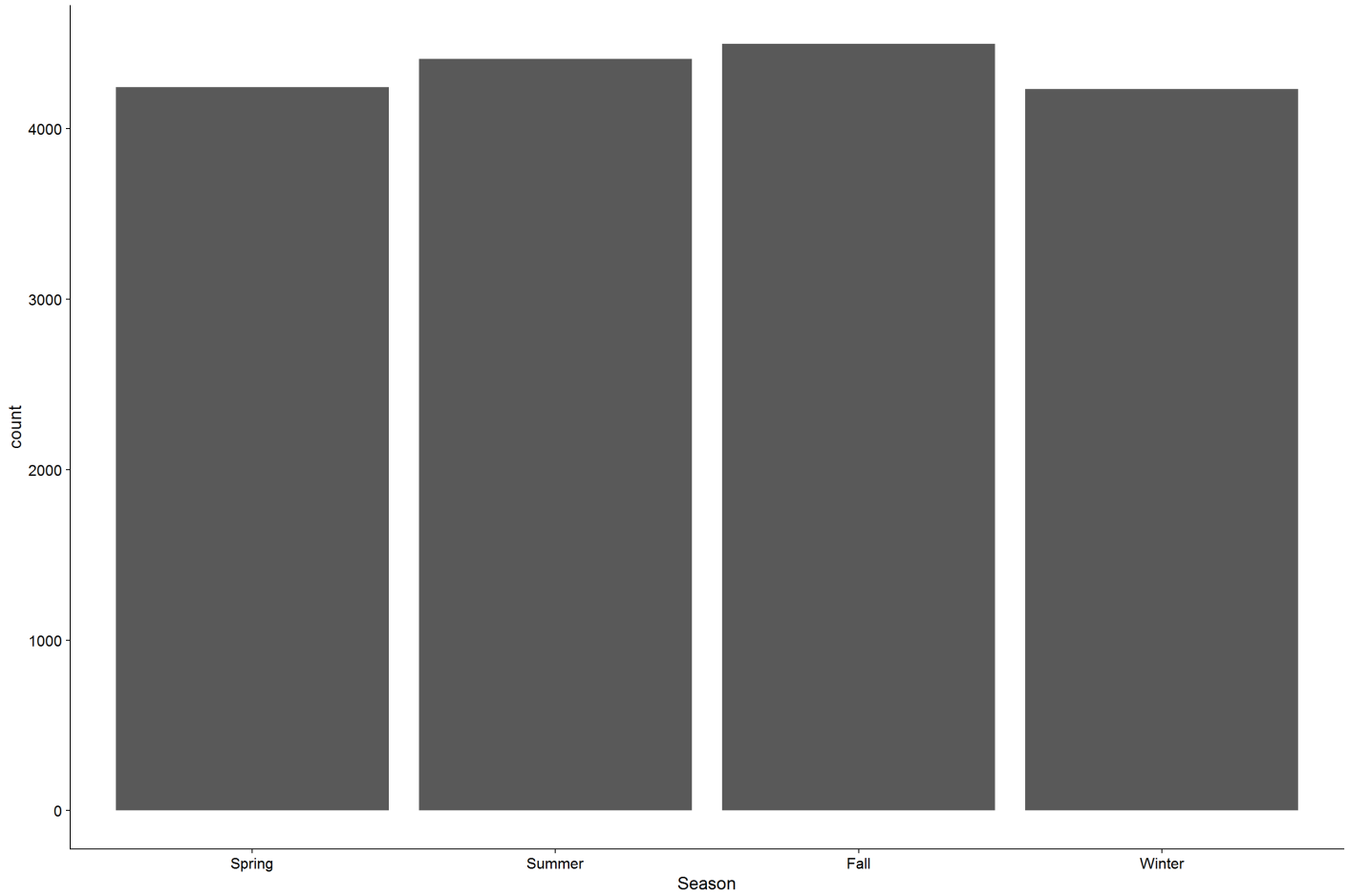



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    4.00   17.00   35.68   48.00  367.00
```

It looks like the number of casual bike renters is usually pretty low compared to registered bike renters. Most of the casual numbers seem to be below 100.

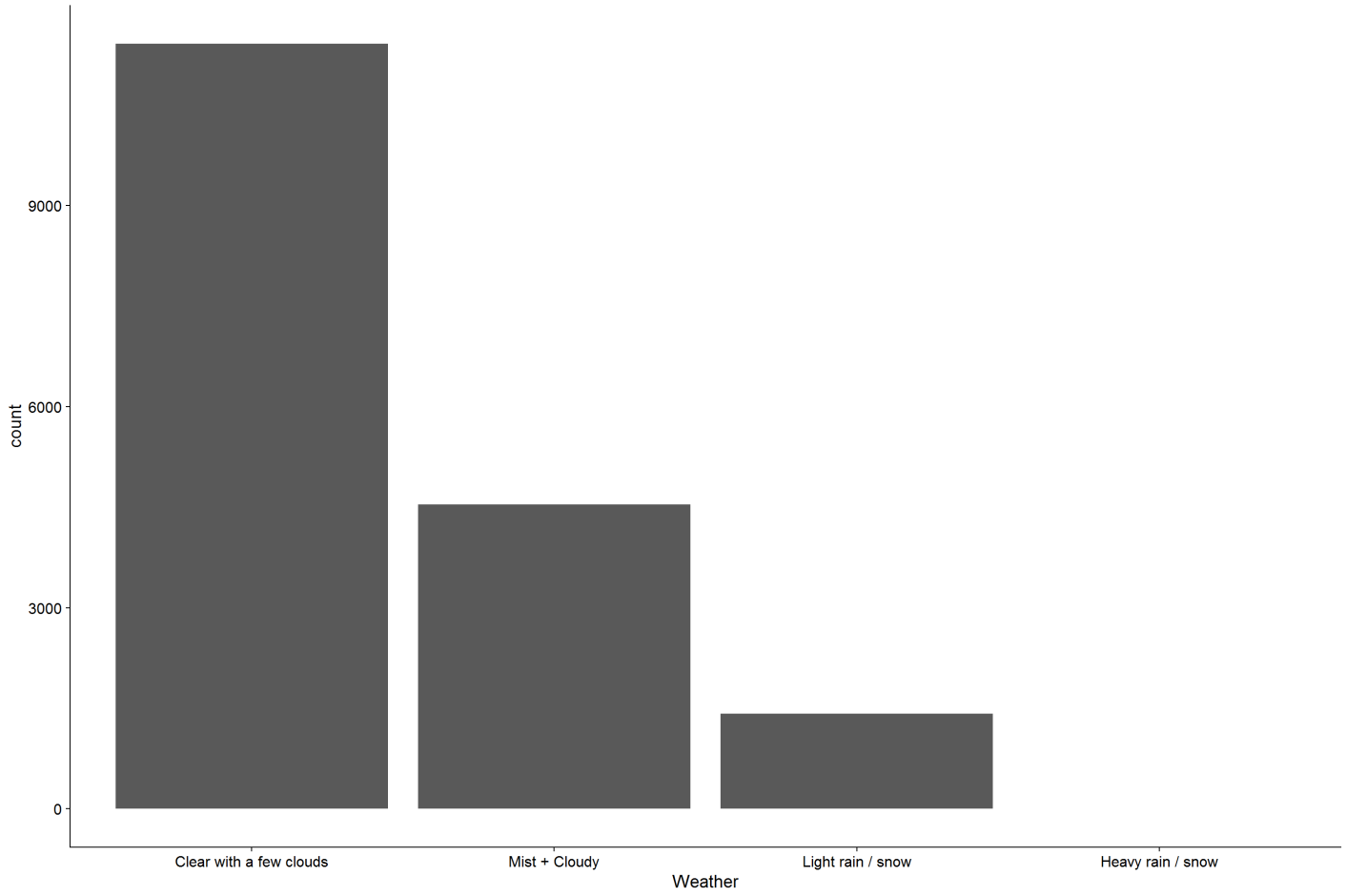
```
##      Total          Registered        Casual
##  Min.   :  1.0   Min.   :  0.0   Min.   :  0.00
##  1st Qu.: 40.0   1st Qu.: 34.0   1st Qu.:  4.00
##  Median :142.0   Median :115.0   Median : 17.00
##  Mean   :189.5   Mean   :153.8   Mean   : 35.68
##  3rd Qu.:281.0   3rd Qu.:220.0   3rd Qu.: 48.00
##  Max.   :977.0   Max.   :886.0   Max.   :367.00
```

Let’s take a look at some other variables in the dataset: Season, Weather, Temperature, etc.



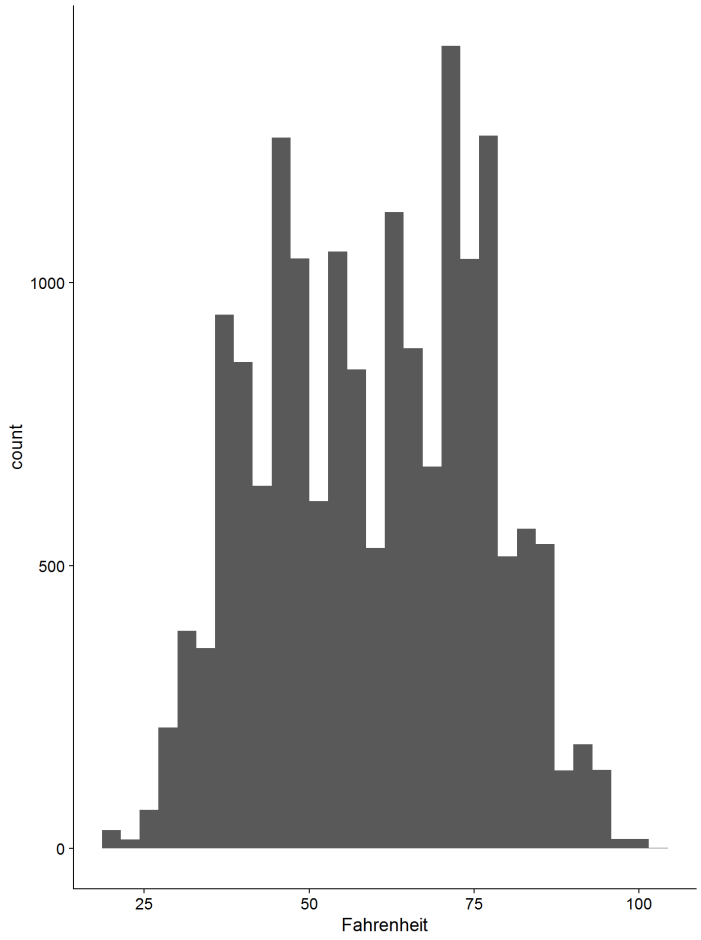

```
## Spring Summer   Fall Winter
##   4242   4409   4496   4232
```

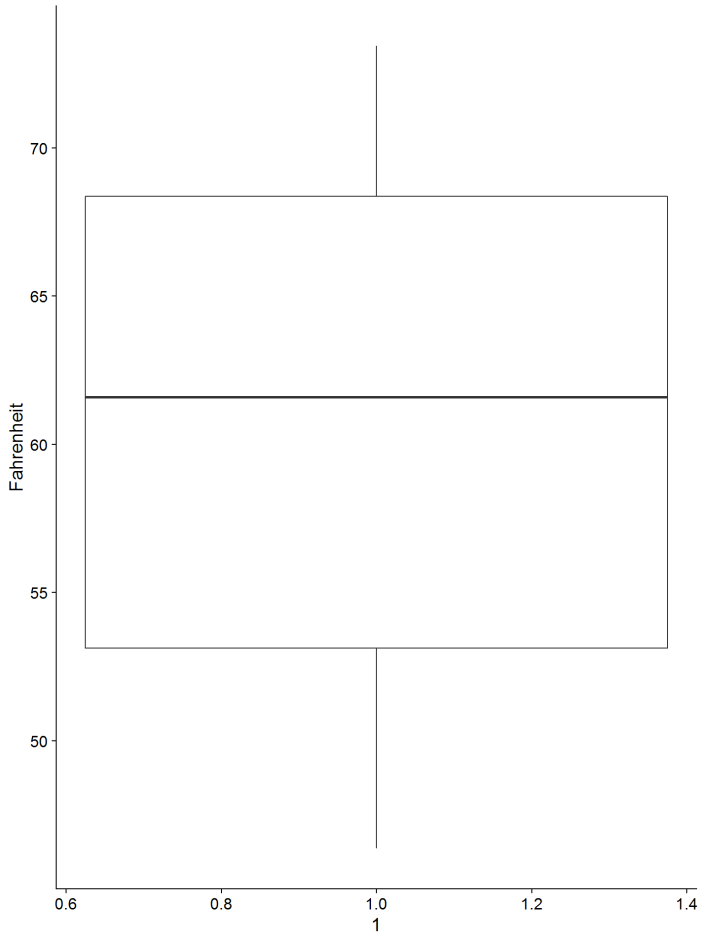
Not much difference here. Looks like all seasons share an almost equal number of days.




```
## Clear with a few clouds           Mist + Cloudy       Light rain / snow
##                   11413                    4544                    1419
##       Heavy rain / snow
##                       3
```

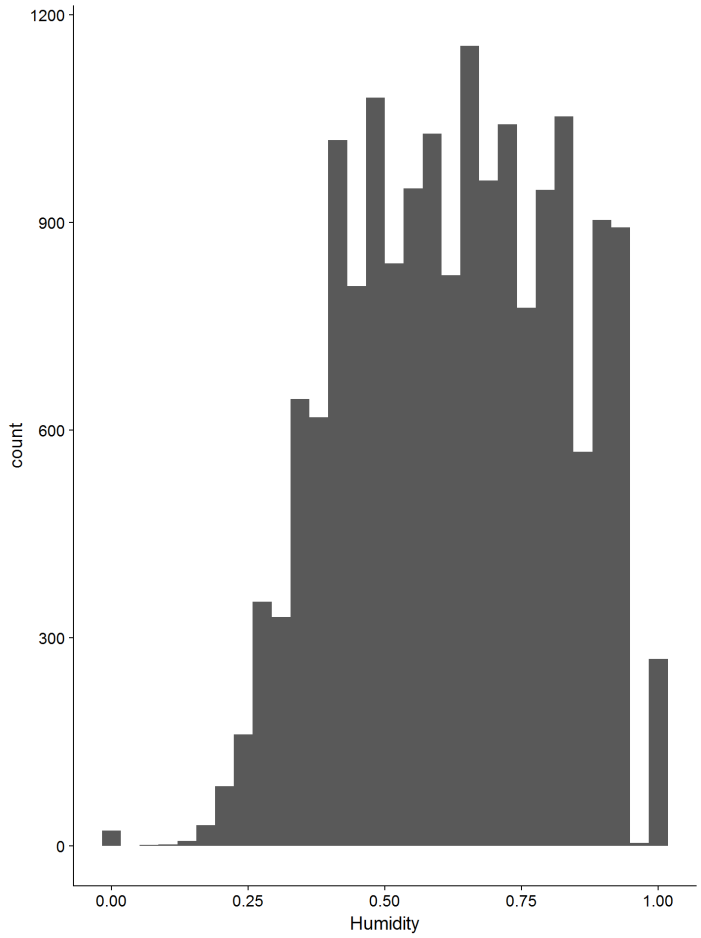
It seems like the weather was clear most of the time. Quite a few days had cloudy skies, with very few days that had light rain and snow. There were 3 days that had heavy snow.

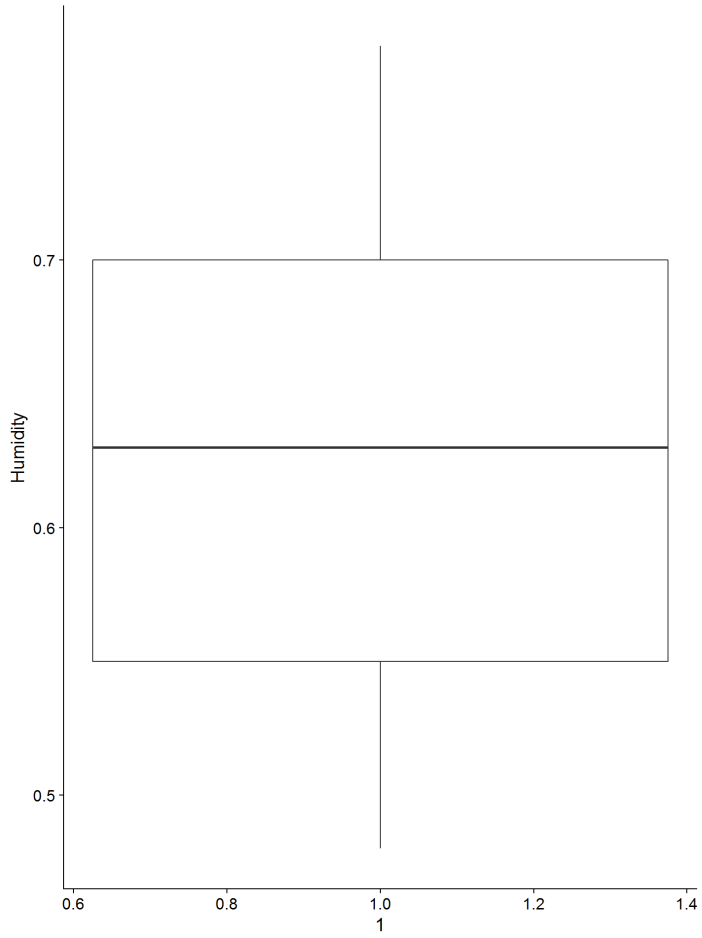





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.29   46.36   59.90   59.65   73.44  102.20
```

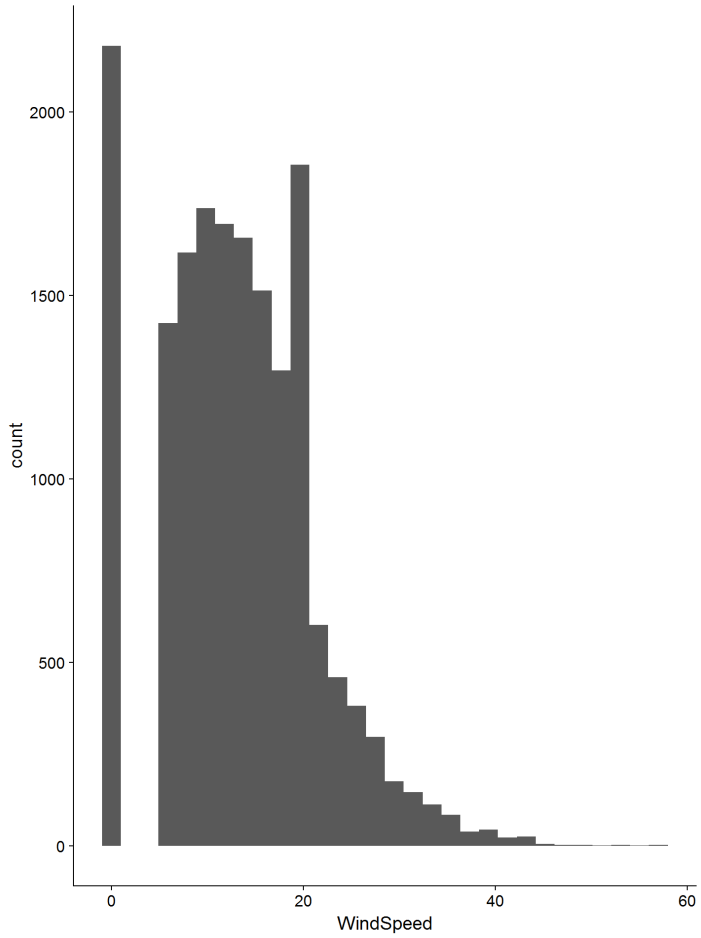
Most of the time the temperature was between 40F and 75F. There were a few days with abnormally low and very high temperatures too.

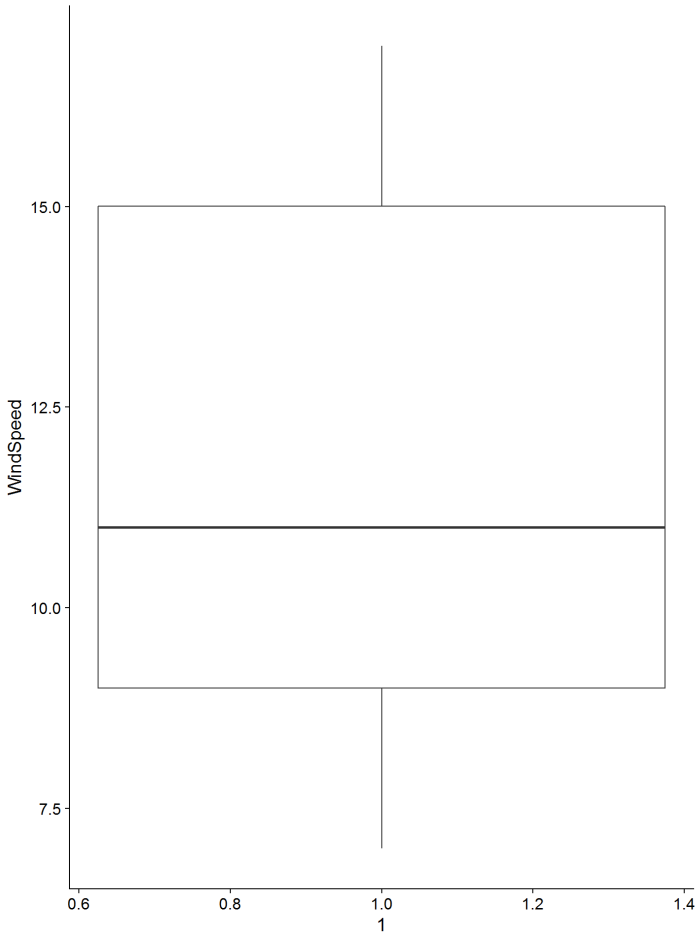





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.4800  0.6300  0.6272  0.7800  1.0000
```

The region is seen to have high humidity levels, with most days having higher than 50% humidity.

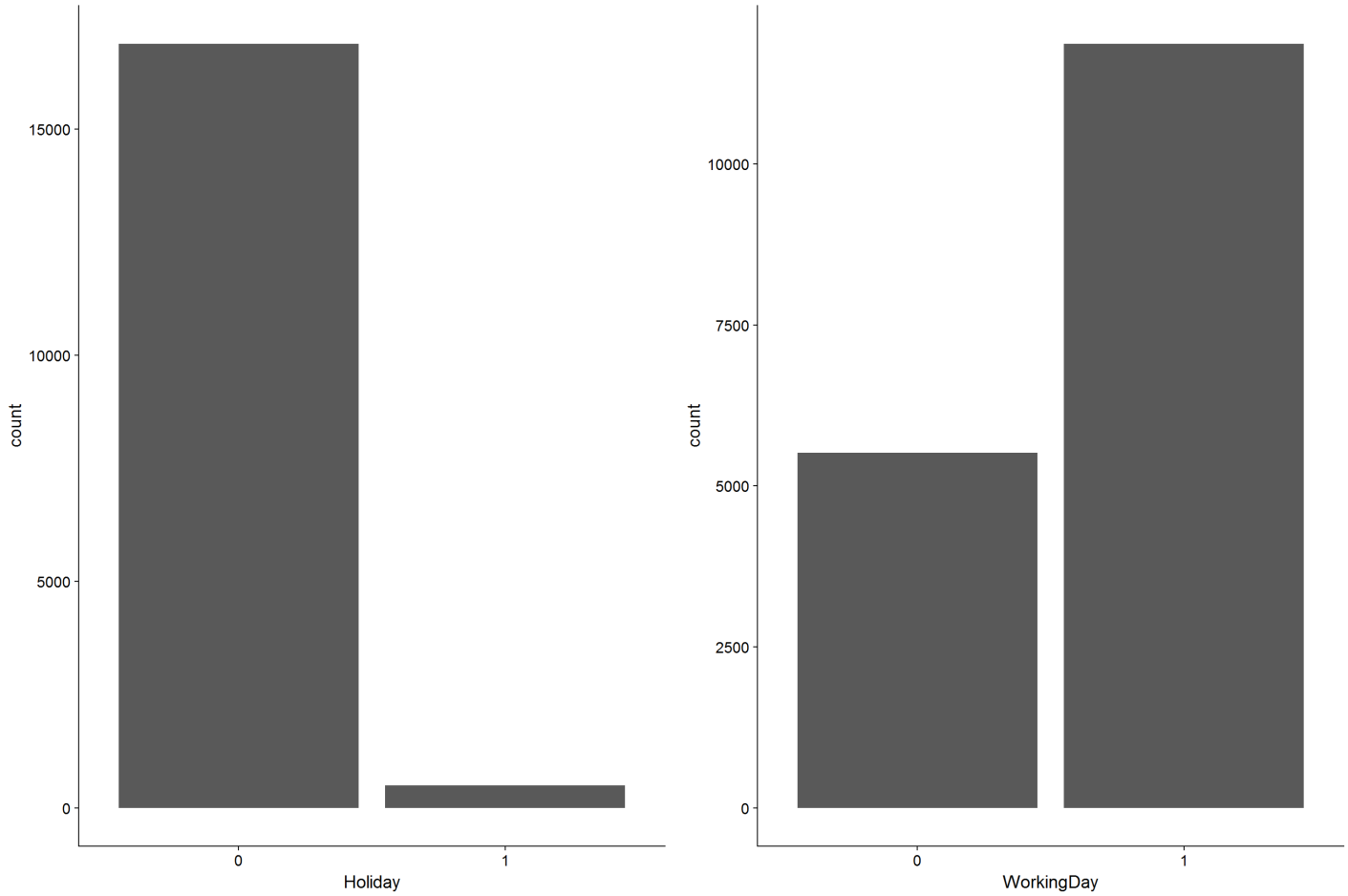




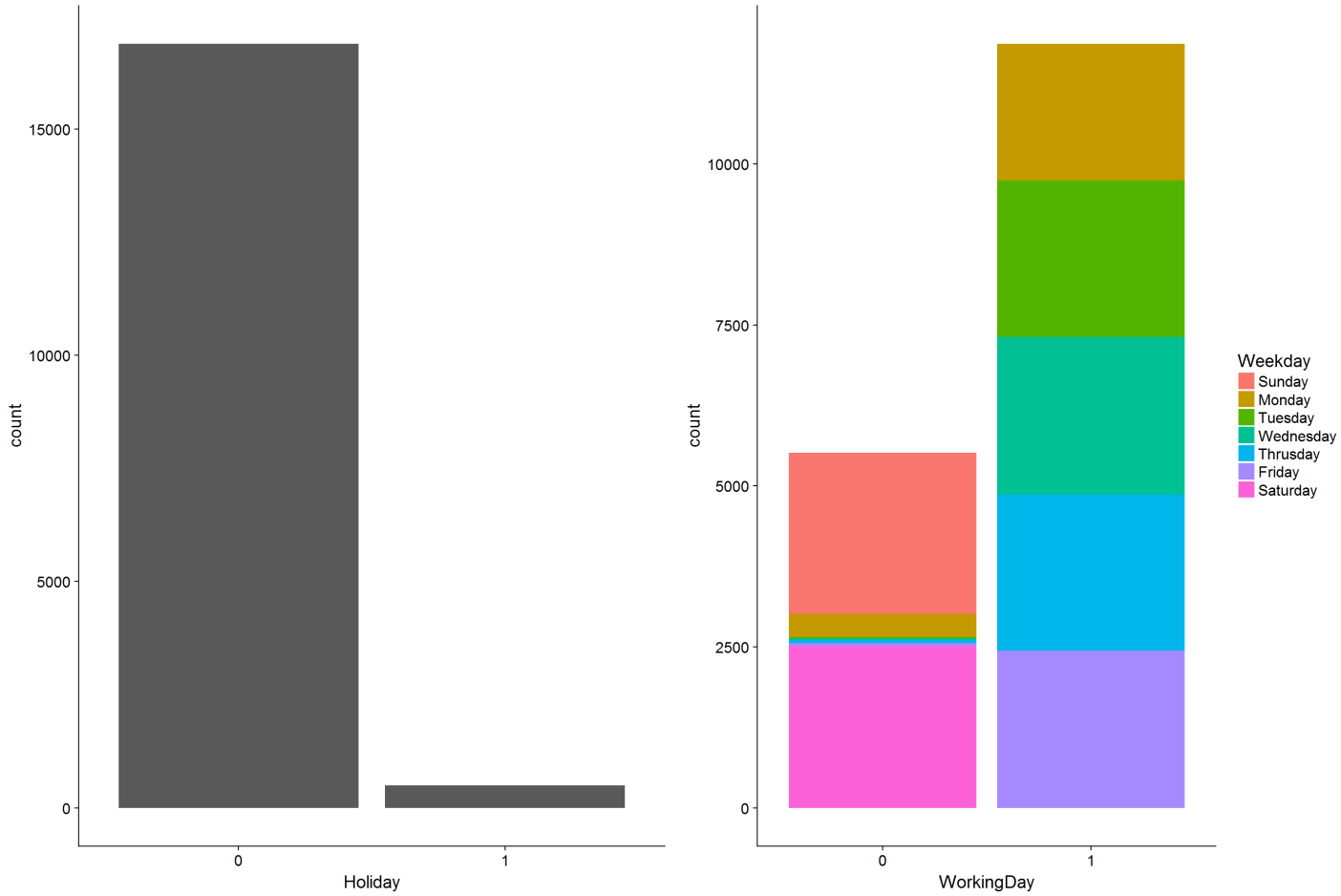

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.002  13.000  12.740  17.000  57.000
```

Nothing abnormal here - wind speed is generally low in this region, mostly less than 20 mph.

Finally, let’s look at the distribution of Holiday and WorkingDay.



There are very few holidays in the dataset, but a lot of non-working days - this could be because non-working days includes weekends whereas holidays only includes special occasions such as Independence Day, Labor Day, Memorial Day, etc. Let’s confirm this.



My assumption was correct: most of the non-working days are in the weekends.

Univariate Analysis

What is the structure of your dataset?

There are 17379 records in the dataset that correspond to each hour of each day in the years 2011 and 2012. The dataset gives information on the number of bikes rented by registered and casual renters during this time period. There are 16 features in the dataset with a combination of date, numeric, integer and factor (ordered and unordered) characteristics.

- Date : Date
- Factors (ordered) : Year, Month, Hour, Weekday
- Factors (unordered) : Holiday, WorkingDay, Weather
- Numeric : Temperature (Celsius and Fahrenheit), Humidity, WindSpeed
- Integer : Total, Registered, Casual

Other observations:

- Season seems to be more or less uniformly distributed, with all the four seasons sharing an almost equal number of days.
- Weather: Most of the days seem to be clear, with a few clouds. This is followed by misty, cloudy days and light snow days.
- The temperature graph shows that most of the days had temperatures between 40F and 75F. There were a few days outside this range too - in the peak winter and summer seasons.

- Humidity in the region is mostly above 50%, owing to the proximity to the coast.
- Windspeed is generally low, mostly below 20 mph.

What is/are the main feature(s) of interest in your dataset?

Our features of interest are Total, Registered and Casual. Total is the total number of bikes rented, which is a sum of Registered and Casual. I'd like to see if features such as temperature, humidity and others have an effect on registered or casual number of bike rentals. I suspect that weekdays will have a high number of registered bike rentals and holidays will have a high number of casual rentals. I will try to confirm the same in my analysis.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

I think Weekday, Workingday and Weather will be more closely correlated to the no. of bike rentals than others.

Did you create any new variables from existing variables in the dataset?

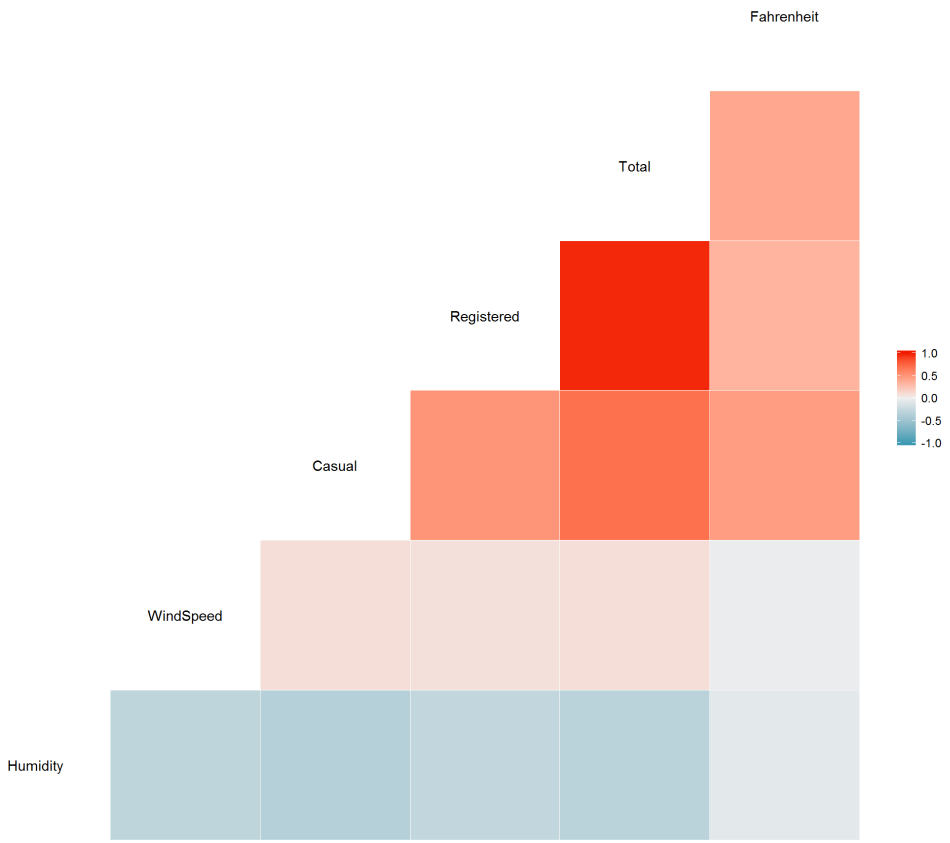
The temperature was given in a normalized format, so I converted it to Celsius and Fahrenheit units. The wind speed was also normalized so I have converted it to absolute mph units. Other than these transformations, I've converted integers to factors and ordered factors wherever applicable.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

I performed a log transform on the no. of bike rentals (for all three features), since the histograms based on the raw data looked right-skewed. The transformed data for Total peaks at around 200, and the one for Registered peaks at around 150. The transformed data for Casual has a few breaks in between 0 and 50, and peaks at around 40.

Bivariate Plots

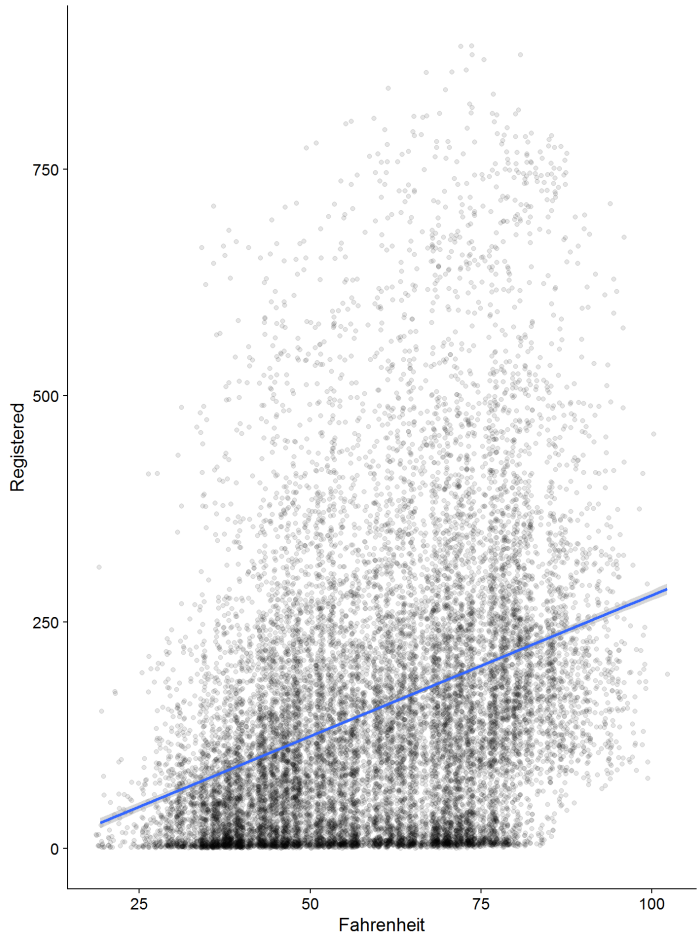
Let’s start this section by investigating how Registered and Casual are correlated with the numeric features - temperature, humidity and windspeed.

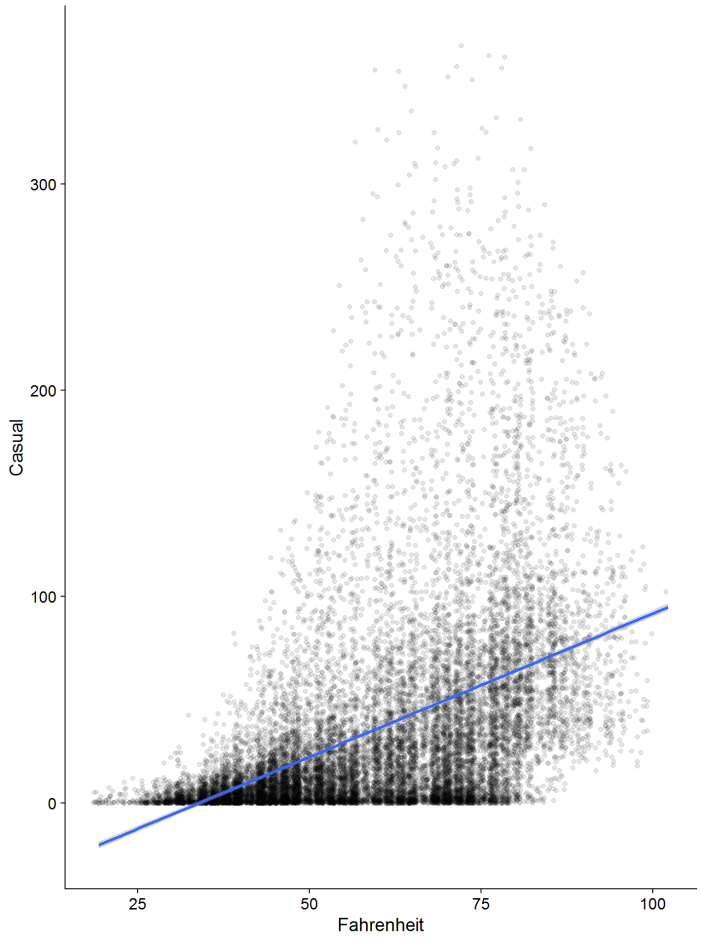


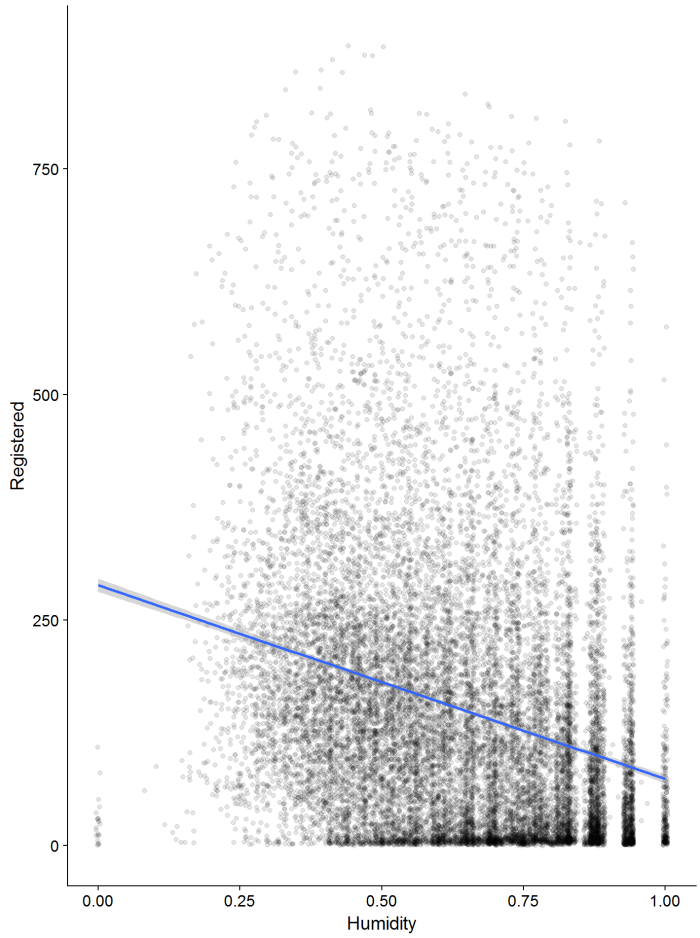
We can see that bike rentals (Registered and Casual) are positively correlated with temperature (Fahrenheit) and negatively with humidity. This makes sense: rentals would be low in winter when the temperature is low, and also on days of rain and snow when the humidity is high. There is a very weak correlation between the bike rentals and windspeed. We saw in univariate analysis that the windspeed is generally low in the region and thus won’t affect our dependent variables too much.

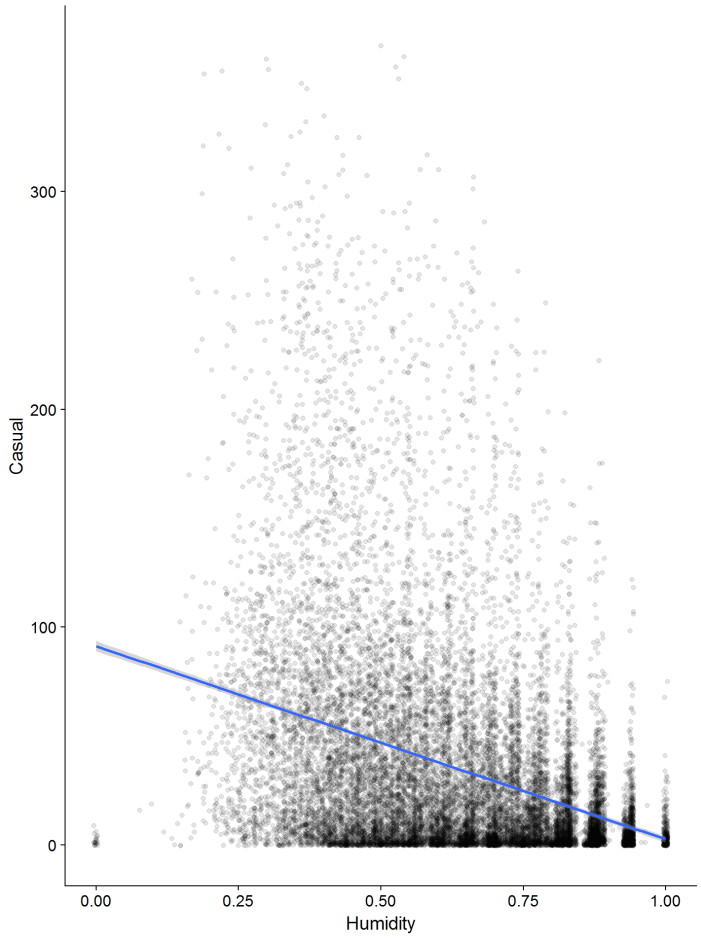
The strongest correlations are between Total with Registered and Casual, but this is only because Total is the sum of Registered and Casual and is not a valid relation.

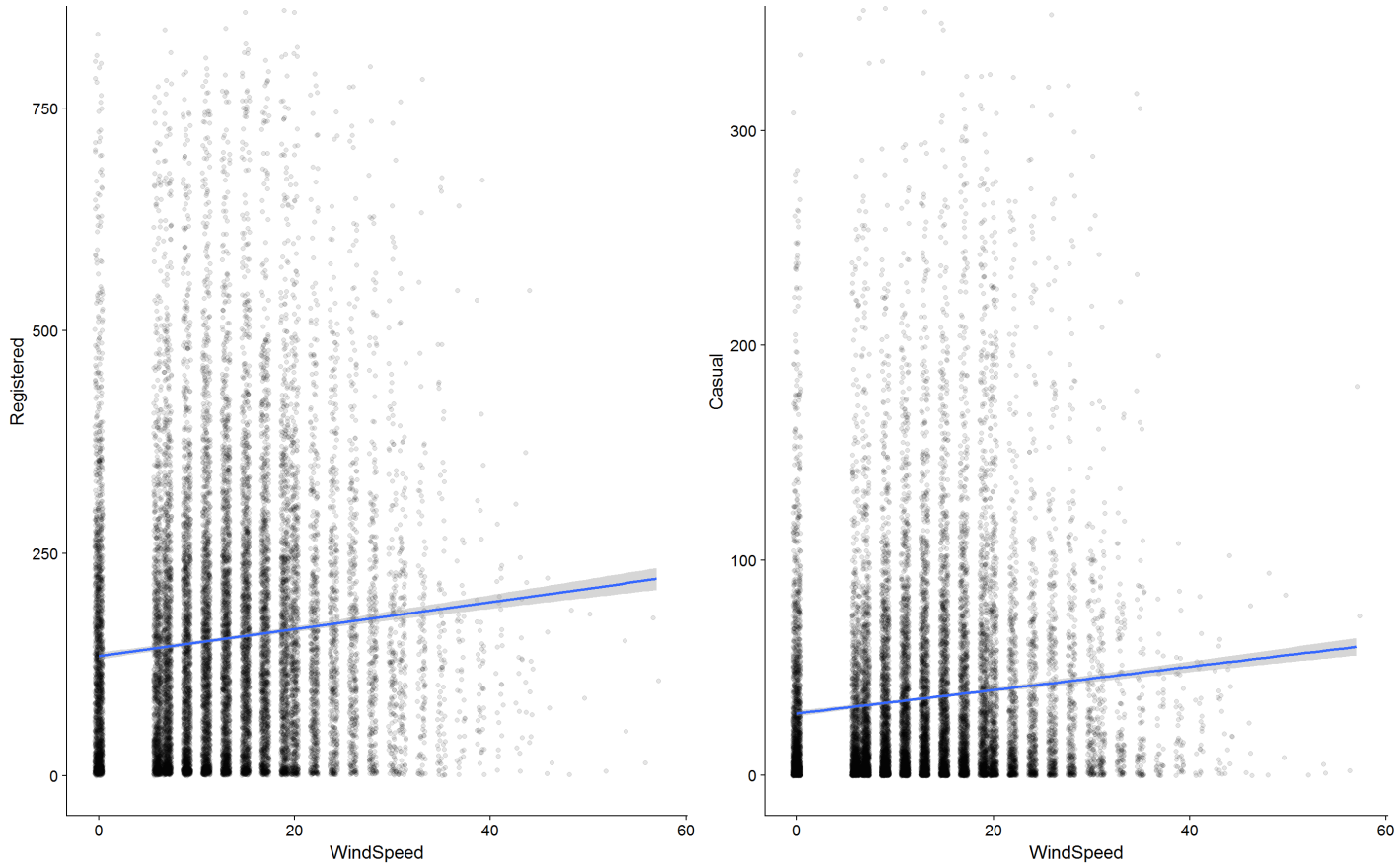
How do temperture, humidity and windspeed affect the bike rental numbers?


Registered
750
500
250
0
25
50
75
100
Fahrenheit


Casual
300
200
100
0
25
50
75
100
Fahrenheit


Registered
750
500
250
0
0.00
0.25
0.50
0.75
1.00
Humidity

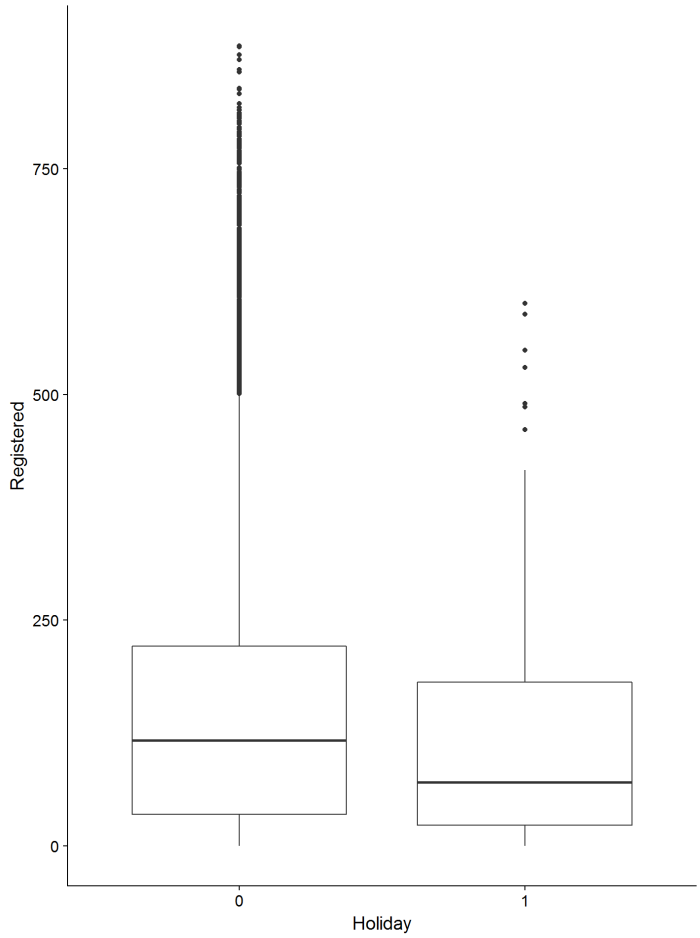

Casual
300
200
100
0
0.00
0.25
0.50
0.75
1.00
Humidity

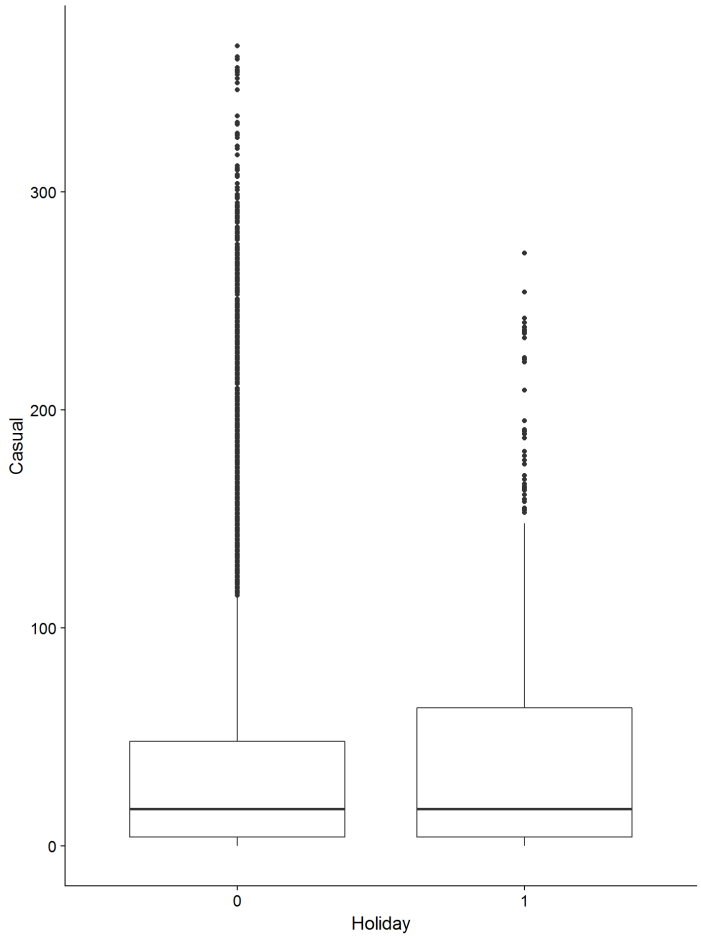


These plots confirm the correlation effects that we studied earlier.

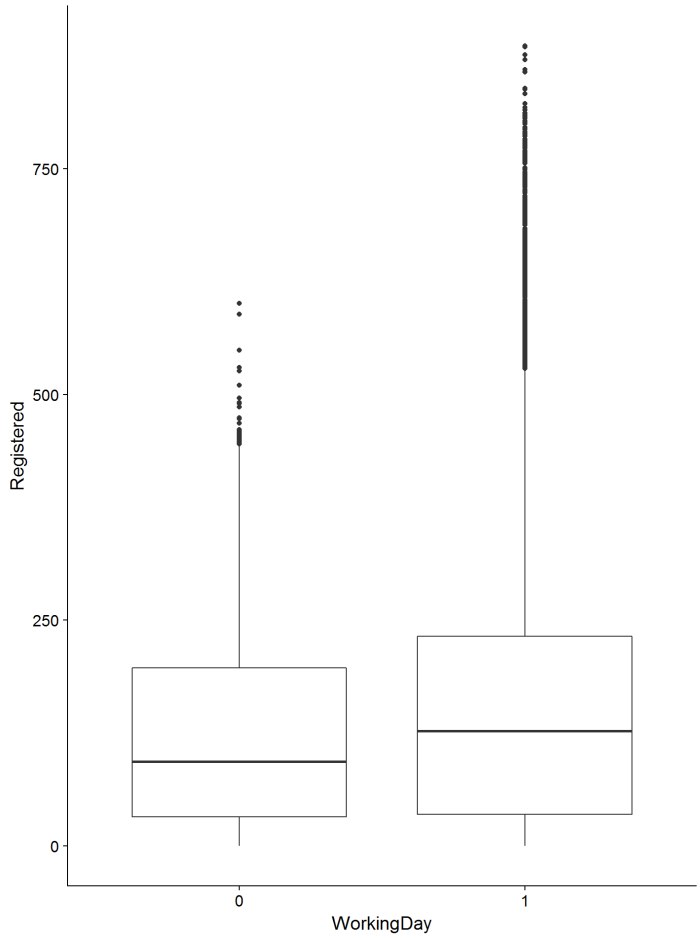
We can now investigate how other variables affect the no. of bike rentals (Registered and Casual).

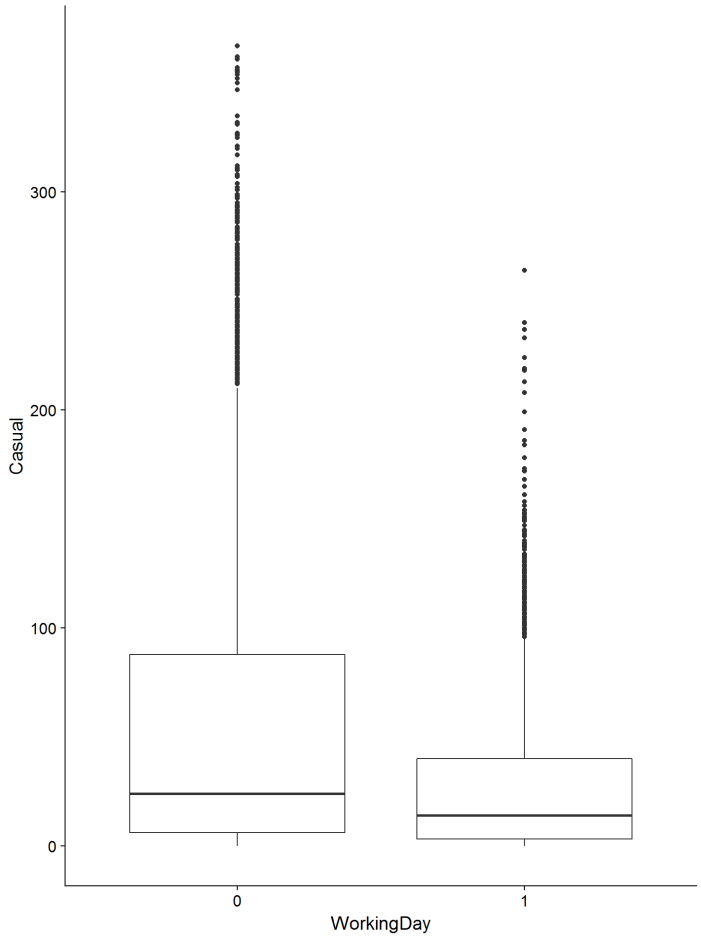
Let’s start with ‘Holiday’ and ‘WorkingDay’.






```
## # A tibble: 2 x 3
##   Holiday Registered   Casual
##    <fctr>      <dbl>    <dbl>
## 1       0   155.0202 35.40838
## 2       1   112.1520 44.71800
```

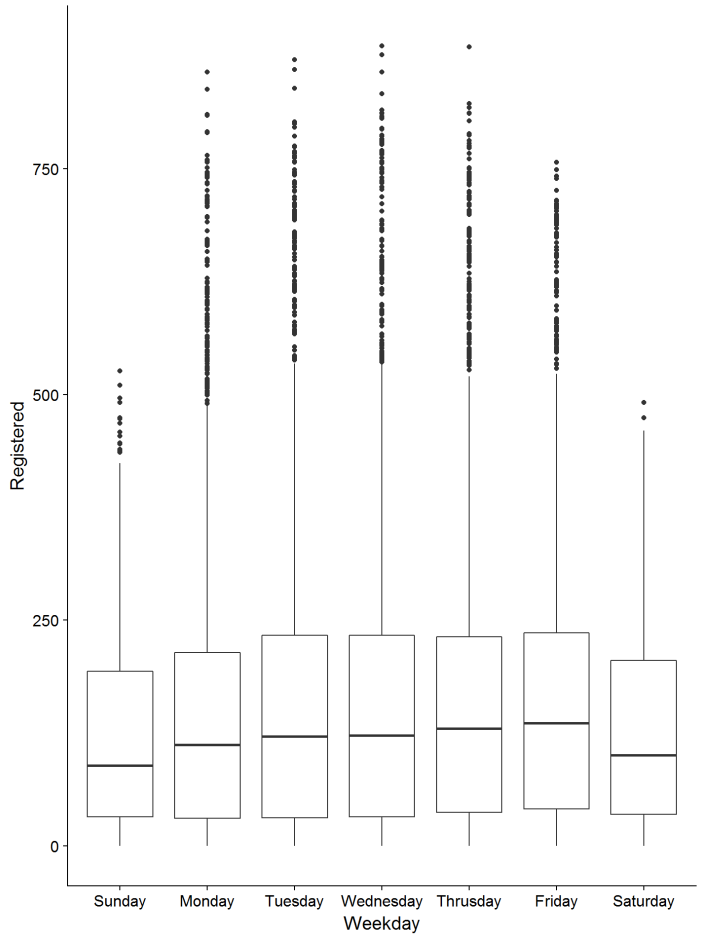


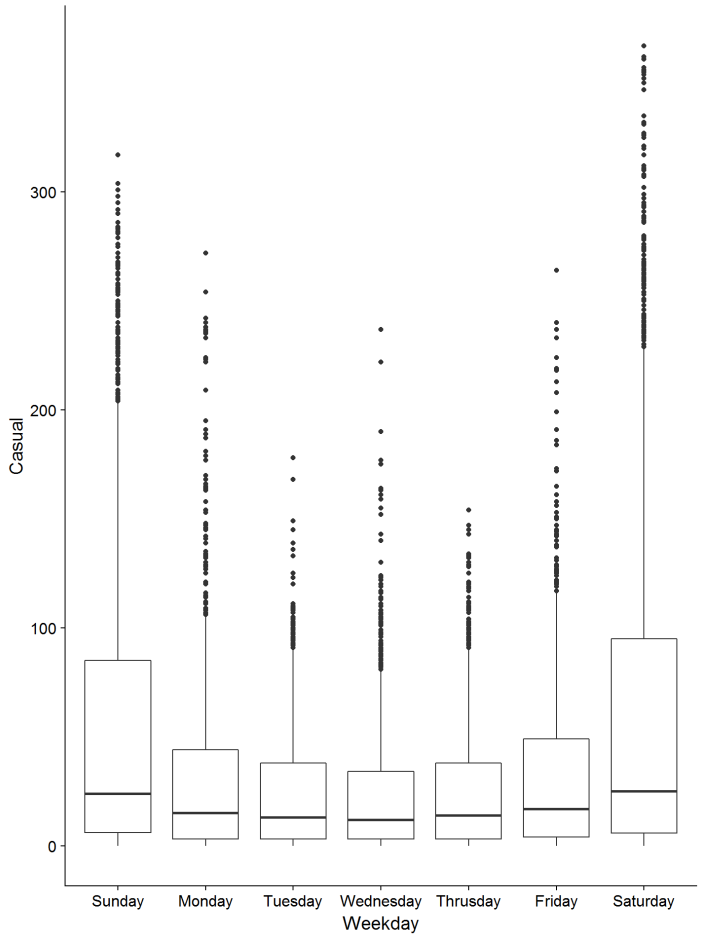


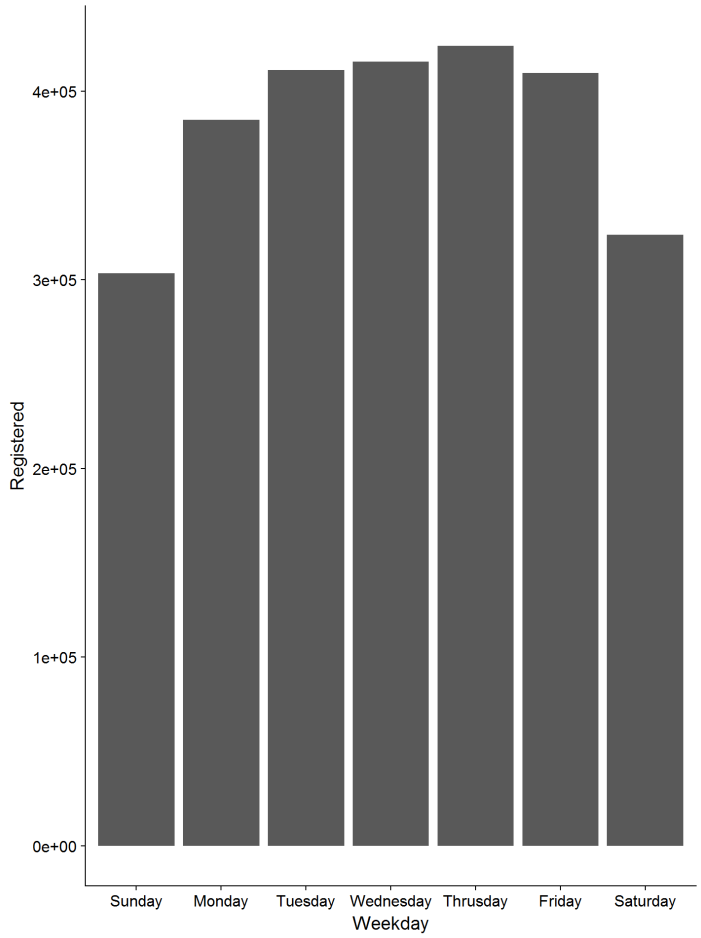

```
## # A tibble: 2 x 3
##   WorkingDay Registered   Casual
##       <fctr>      <dbl>    <dbl>
## 1          0   123.9639 57.44142
## 2          1   167.6464 25.56131
```

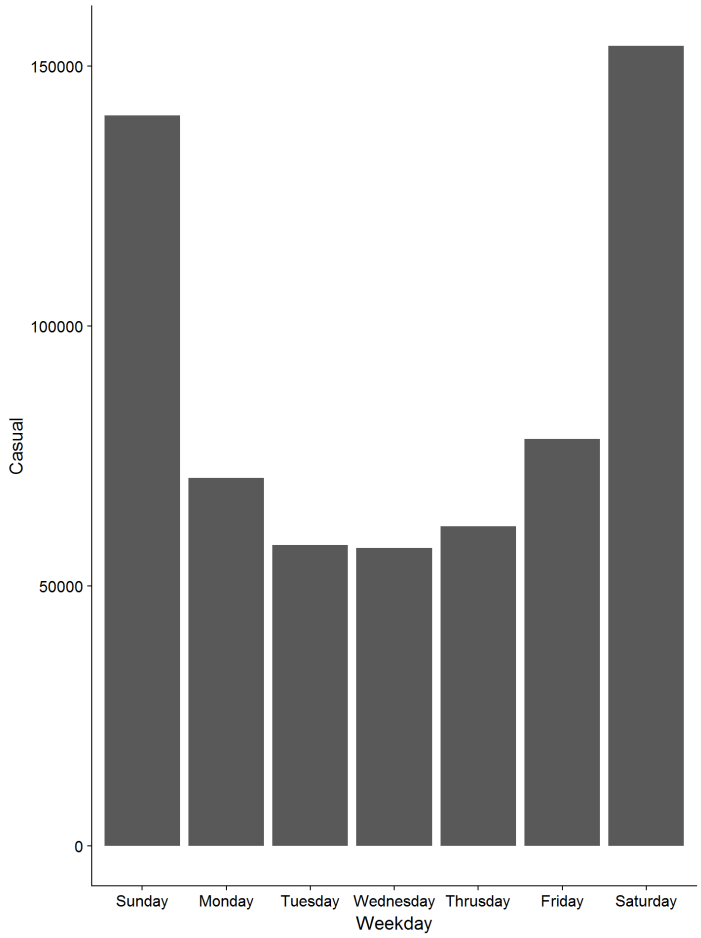
We can see that the no. of registered bikers on holidays and non-working days is low, whereas for casual bikers, it’s on the higher side.

Let’s look at ‘Weekday’.


Registered
750
500
250
0
Sunday
Monday
Tuesday
Wednesday
Thrusday
Friday
Saturday
Weekday


Casual
300
200
100
0
Sunday
Monday
Tuesday
Wednesday
Thrusday
Friday
Saturday
Weekday

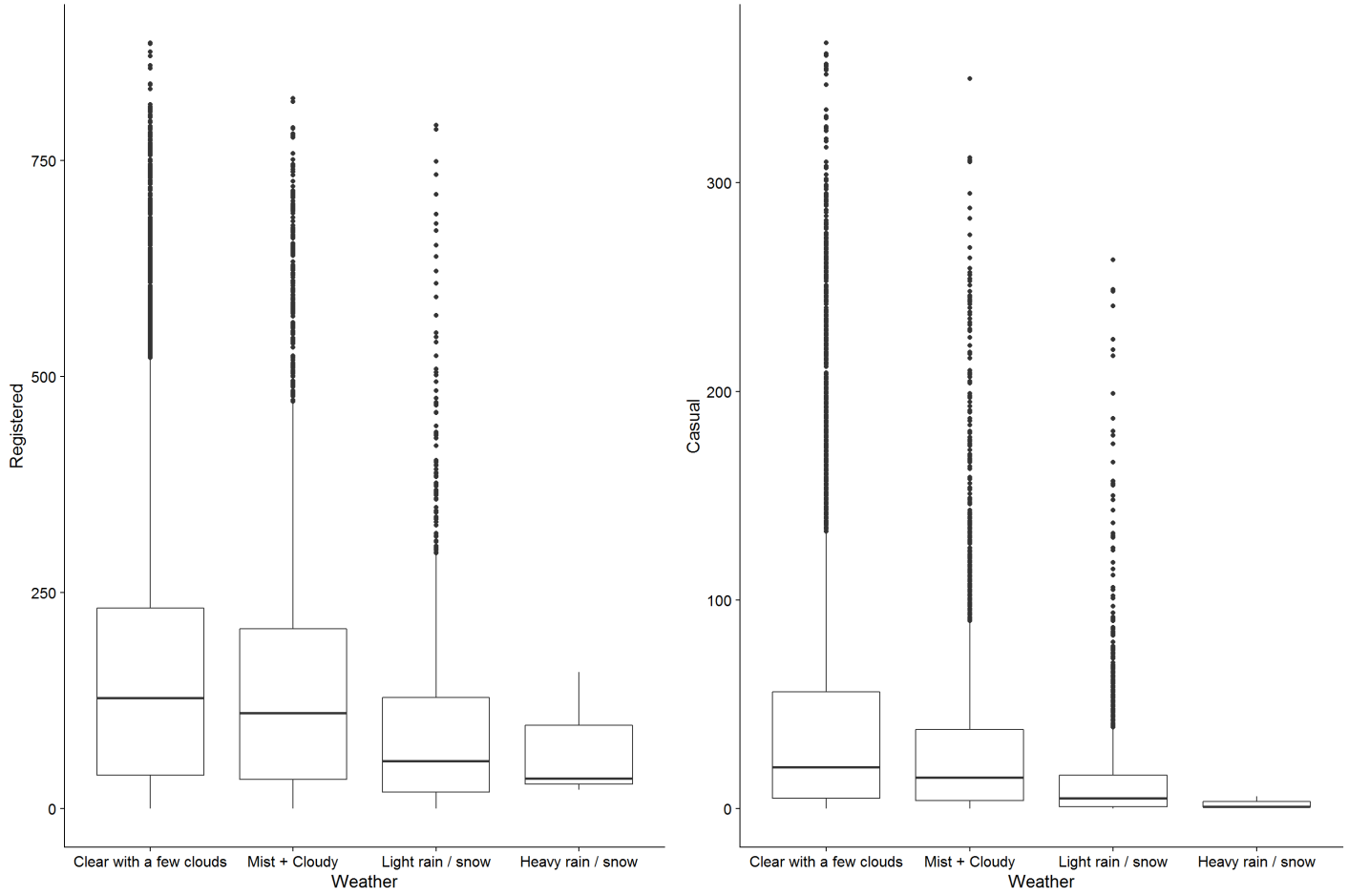

Registered
4e+05
3e+05
2e+05
1e+05
0e+00
Sunday
Monday
Tuesday
Wednesday
Thrusday
Friday
Saturday
Weekday

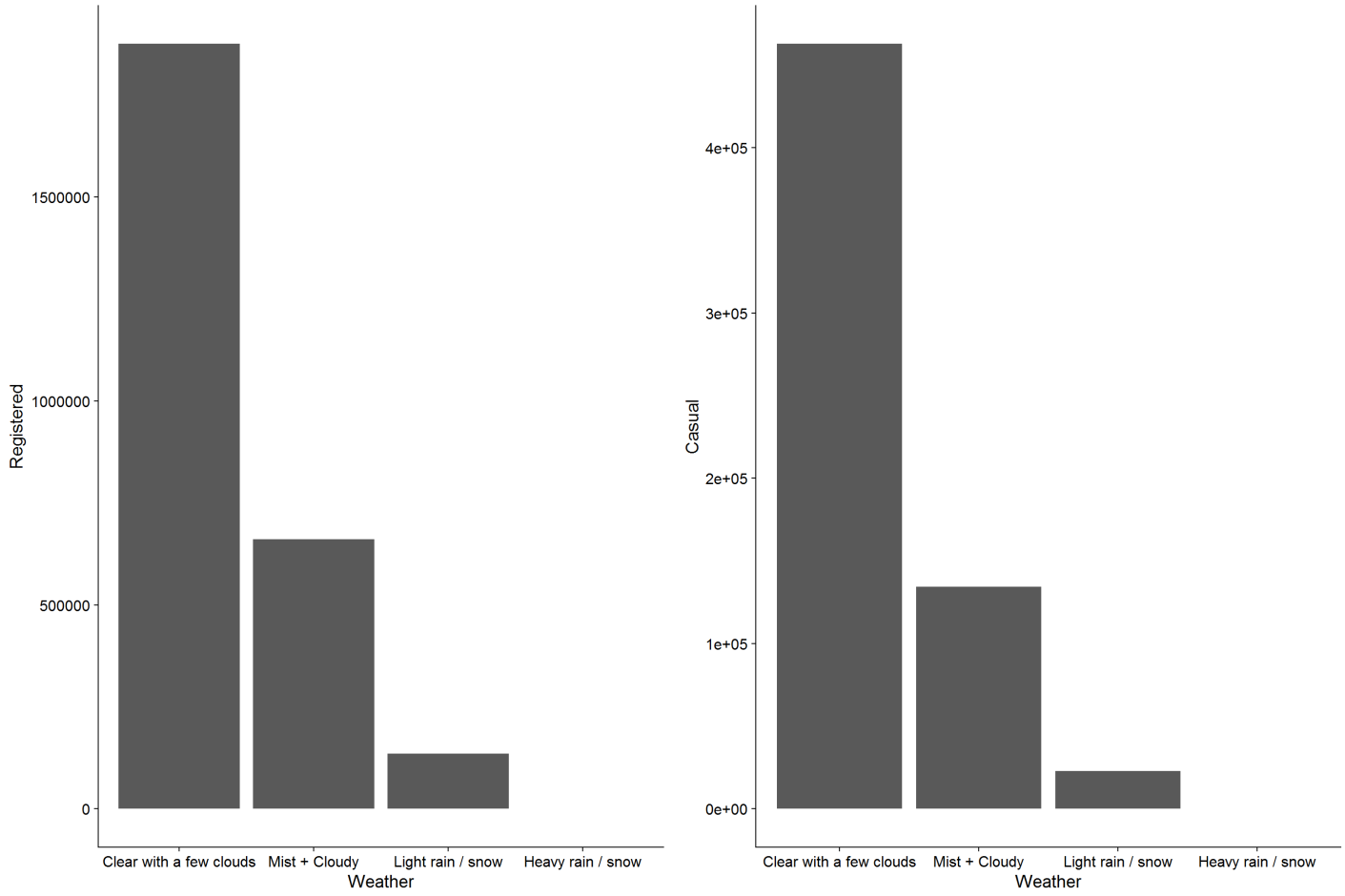

Casual
150000
100000
50000
0
Sunday
Monday
Tuesday
Wednesday
Thrusday
Friday
Saturday
Weekday

```
## # A tibble: 7 x 3
##     Weekday Registered   Casual
##      <fctr>      <dbl>    <dbl>
## 1    Sunday   121.3054 56.16347
## 2    Monday   155.1912 28.55345
## 3   Tuesday   167.6584 23.58051
## 4 Wednesday   167.9713 23.15919
## 5  Thrusday   171.5641 24.87252
## 6    Friday   164.6771 31.45879
## 7  Saturday   128.9630 61.24682
```

This presents a very clear trend! The no. of registered bikers is clearly higher during the weekdays, and the no. of casual bikers is clearly higher during the weekends. I think this is because most of the registered bikers use it for commuting to work, whereas most of the casual bikers may be tourists / visitors that use it during the weekends.

Let's see if weather has an impact on the no. of rentals. I think the no. of bike rentals, both registered and casual, will be higher on clear days and low on bad weather days.

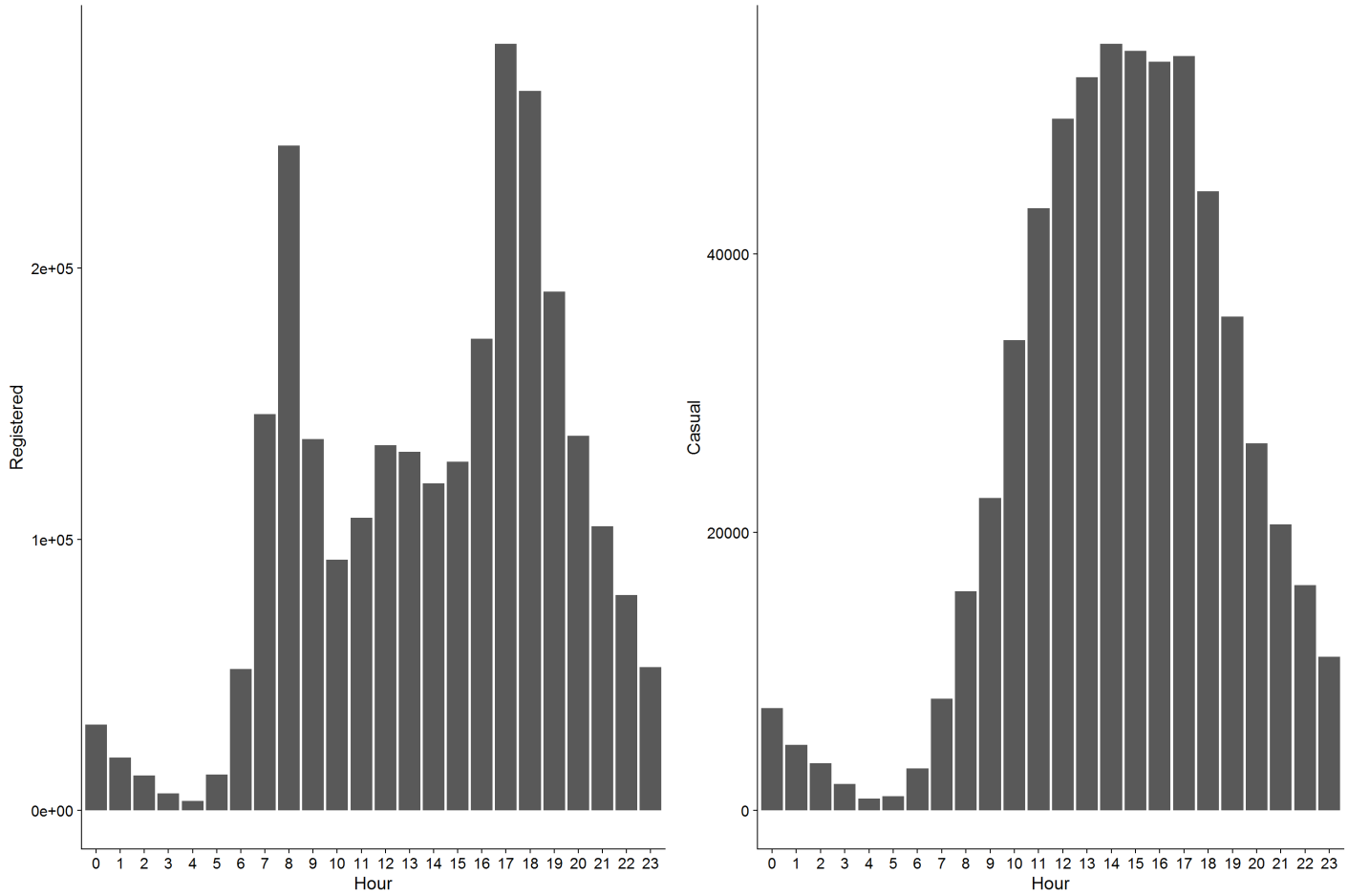





```
## # A tibble: 4 x 3
##                   Weather Registered    Casual
##                    <fctr>      <dbl>     <dbl>
## 1 Clear with a few clouds  164.32384 40.545431
## 2           Mist + Cloudy  145.57020 29.595290
## 3       Light rain / snow   95.52361 16.055673
## 4       Heavy rain / snow   71.66667  2.666667
```

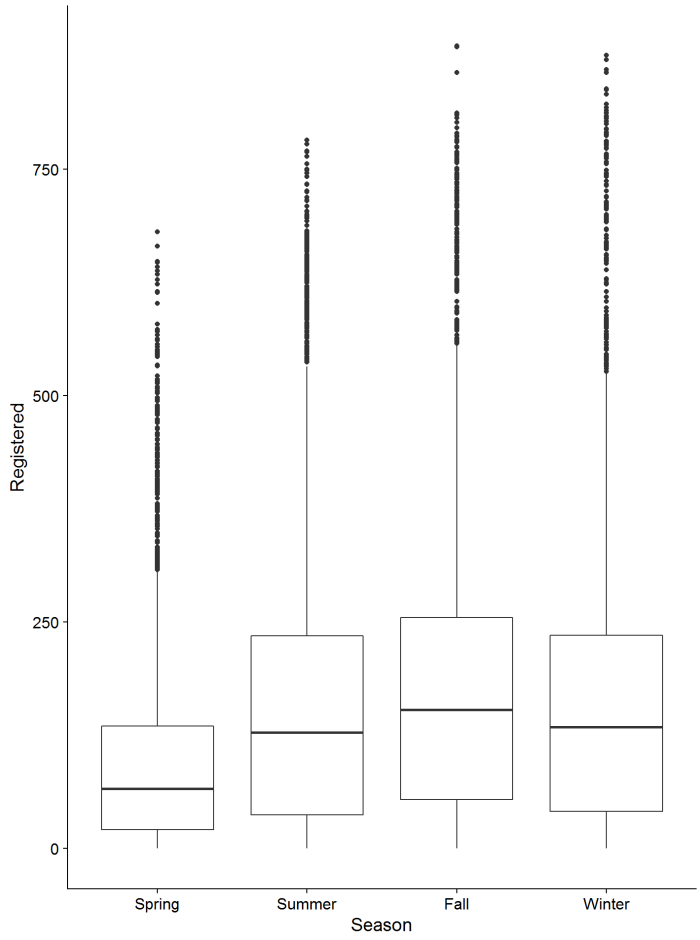
This confirms my theory. Clear days have a high no. of rentals, and the numbers are lower as the weather situation gets worse.

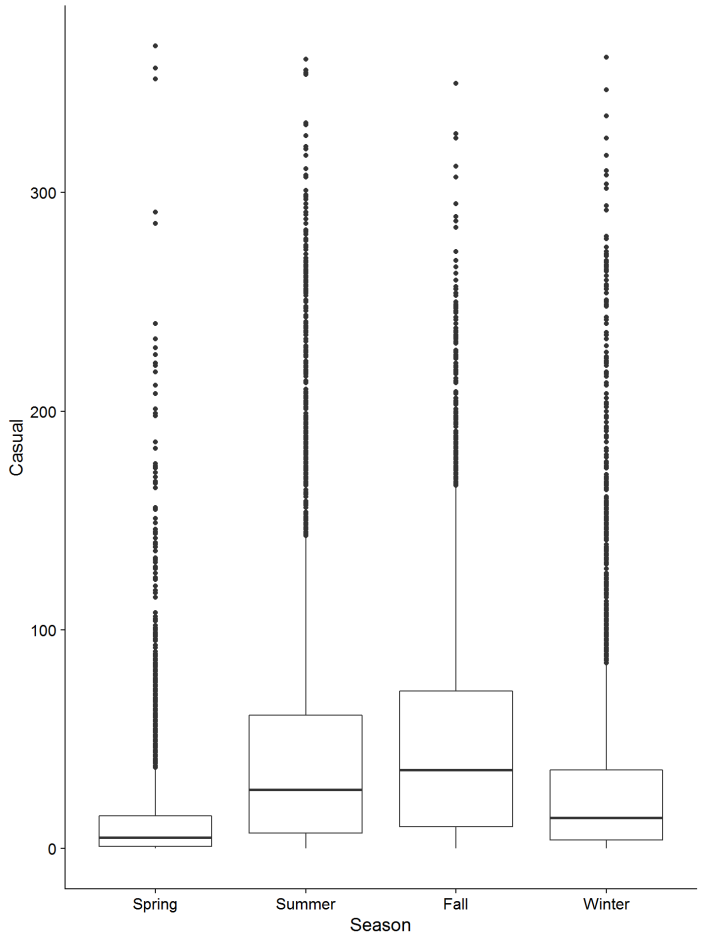
Since the data is available for each hour of the day, I think the time of the day may also have an impact. Let's investigate.

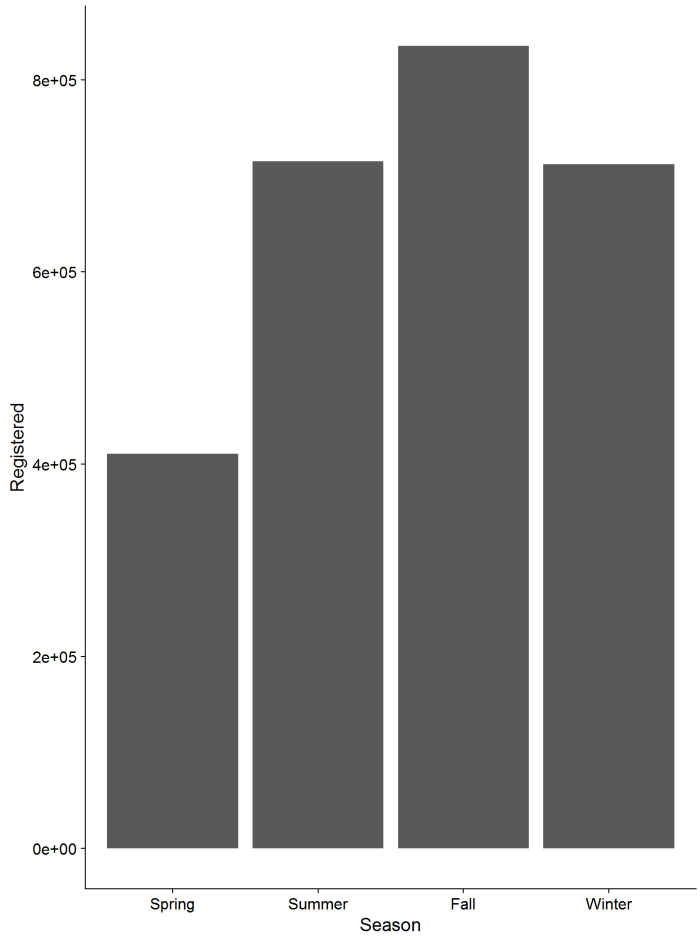


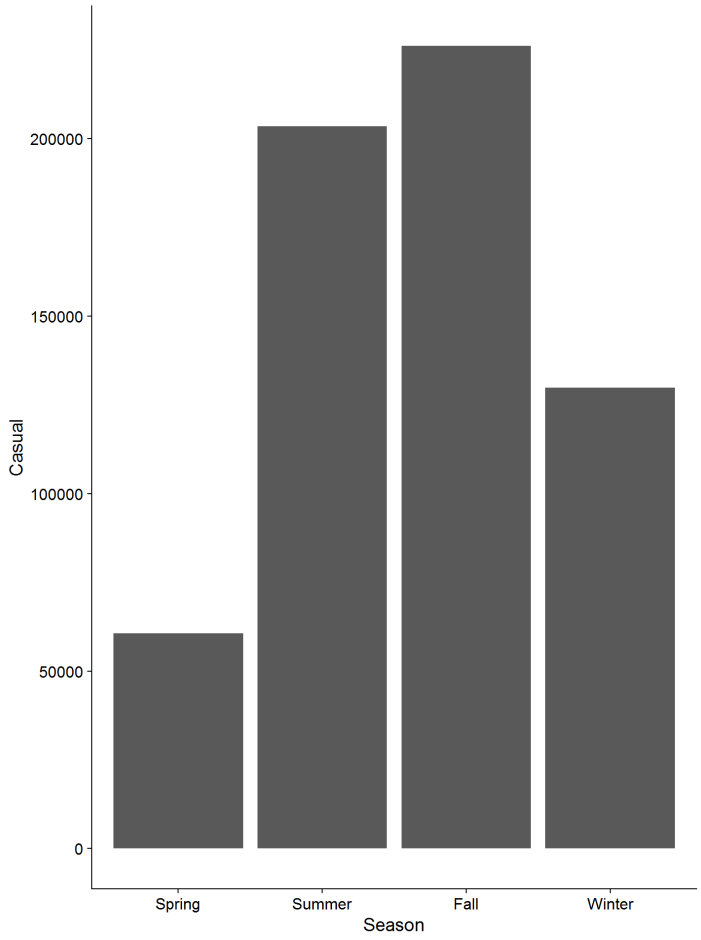
This clearly shows that the majority of the registered subscribers use the service to commute to work. The peaks at 8 AM and 5 PM show the concentration of usage during the times when subscribers go to and from work. The Casual plot, on the other hand, peaks during the day time, around 2-5 PM, and gradually goes down after that.

Similar to weather, let's try plotting the numbers by season. Are the rentals high during summer?


Registered
750
500
250
0
Spring
Summer
Fall
Winter
Season


Casual
300
200
100
0
Spring
Summer
Fall
Winter
Season

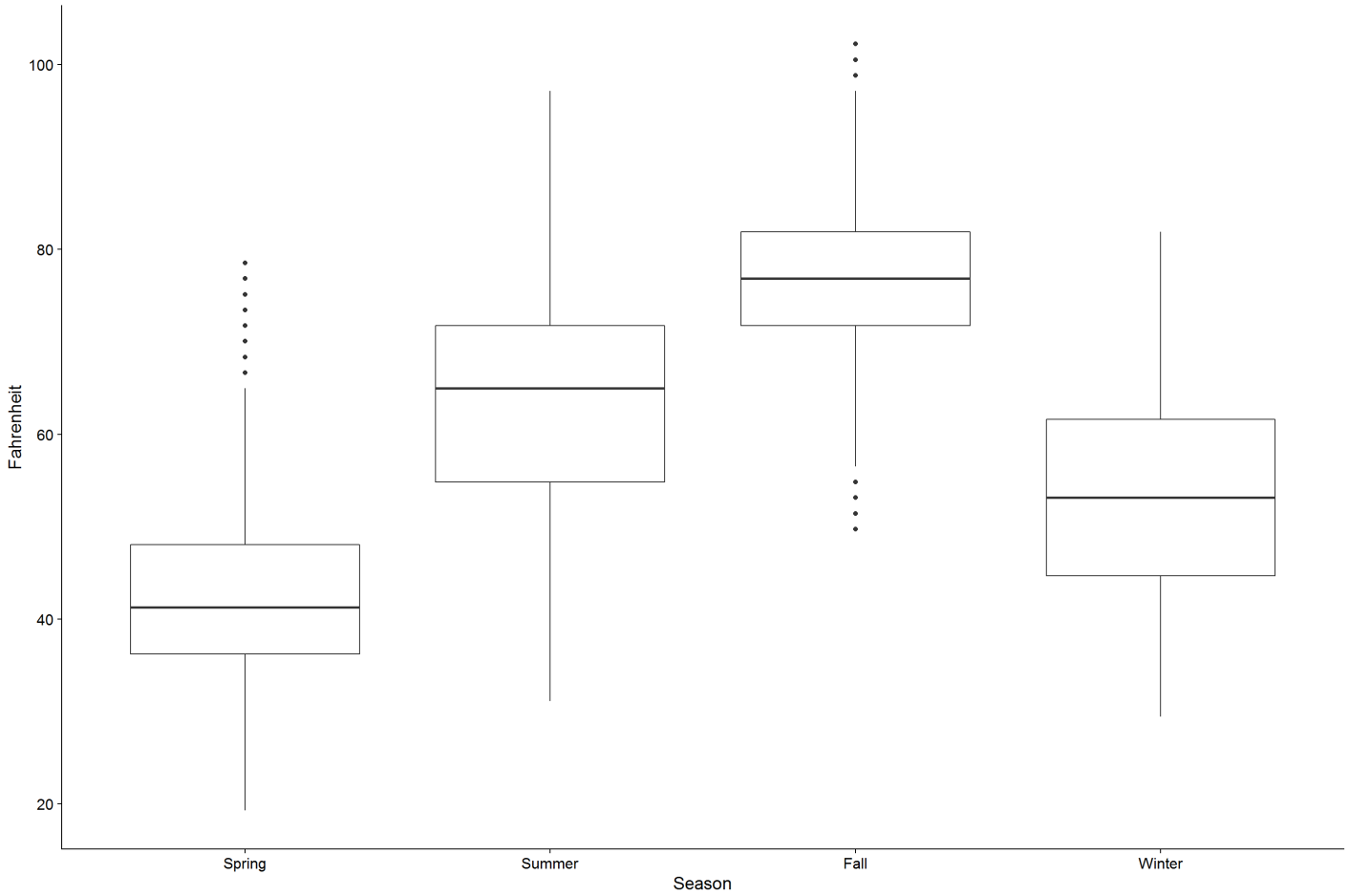

Registered
8e+05
6e+05
4e+05
2e+05
0e+00
Spring
Summer
Fall
Winter
Season


Casual
200000
150000
100000
50000
0
Spring
Summer
Fall
Winter
Season

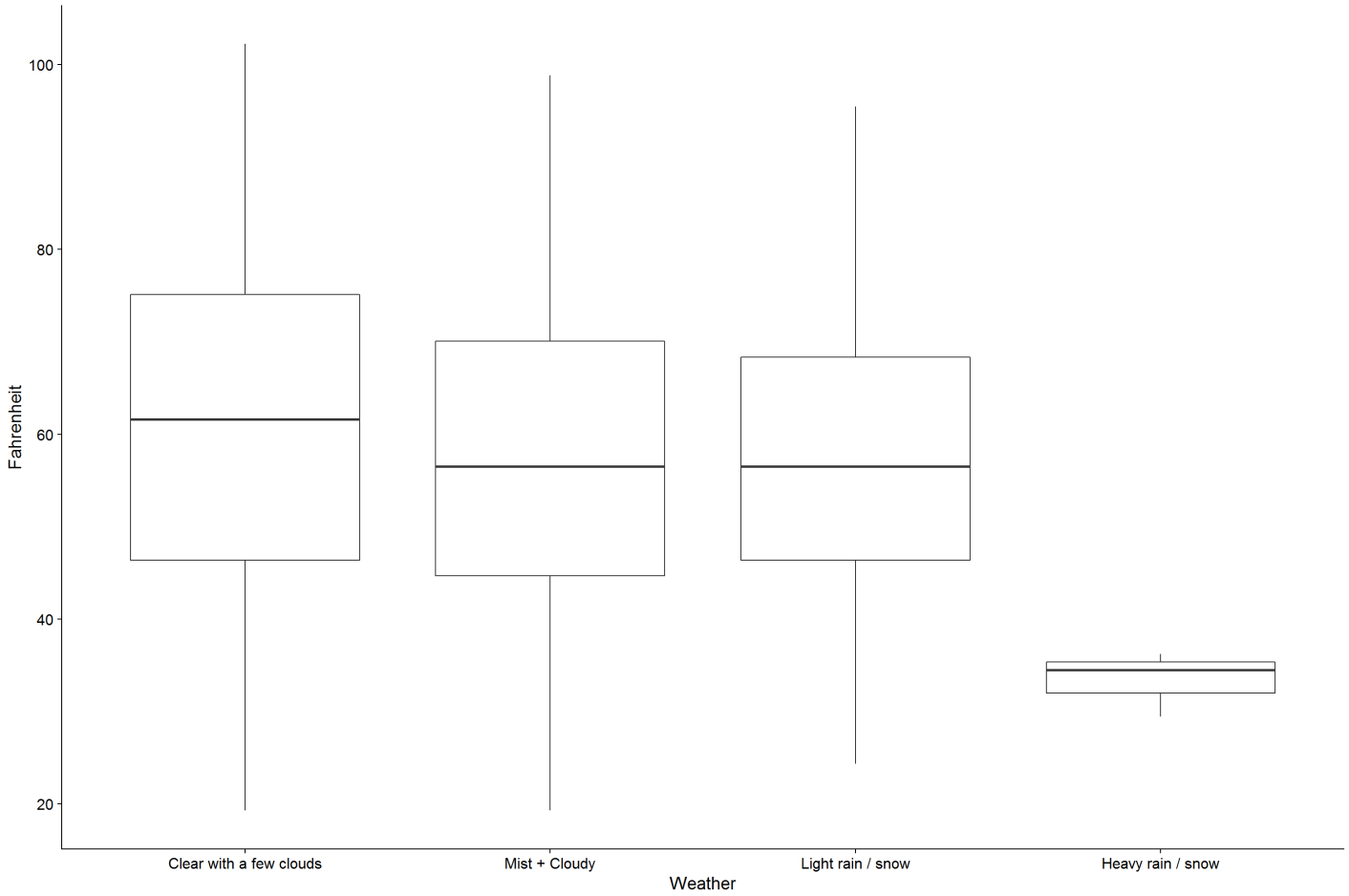
```
## # A tibble: 4 x 3
##   Season Registered   Casual
##   <fctr>      <dbl>    <dbl>
## 1 Spring   96.82367 14.29090
## 2 Summer  162.18349 46.16058
## 3   Fall  185.72909 50.28714
## 4 Winter  168.20203 30.66682
```

The graphs show that the numbers are highest during fall, not summer. This could be because fall is characterized by lower temperatures, fall foliage colors and clear skies. Surprisingly though, the numbers during spring are lower than in winter - I was expecting otherwise.

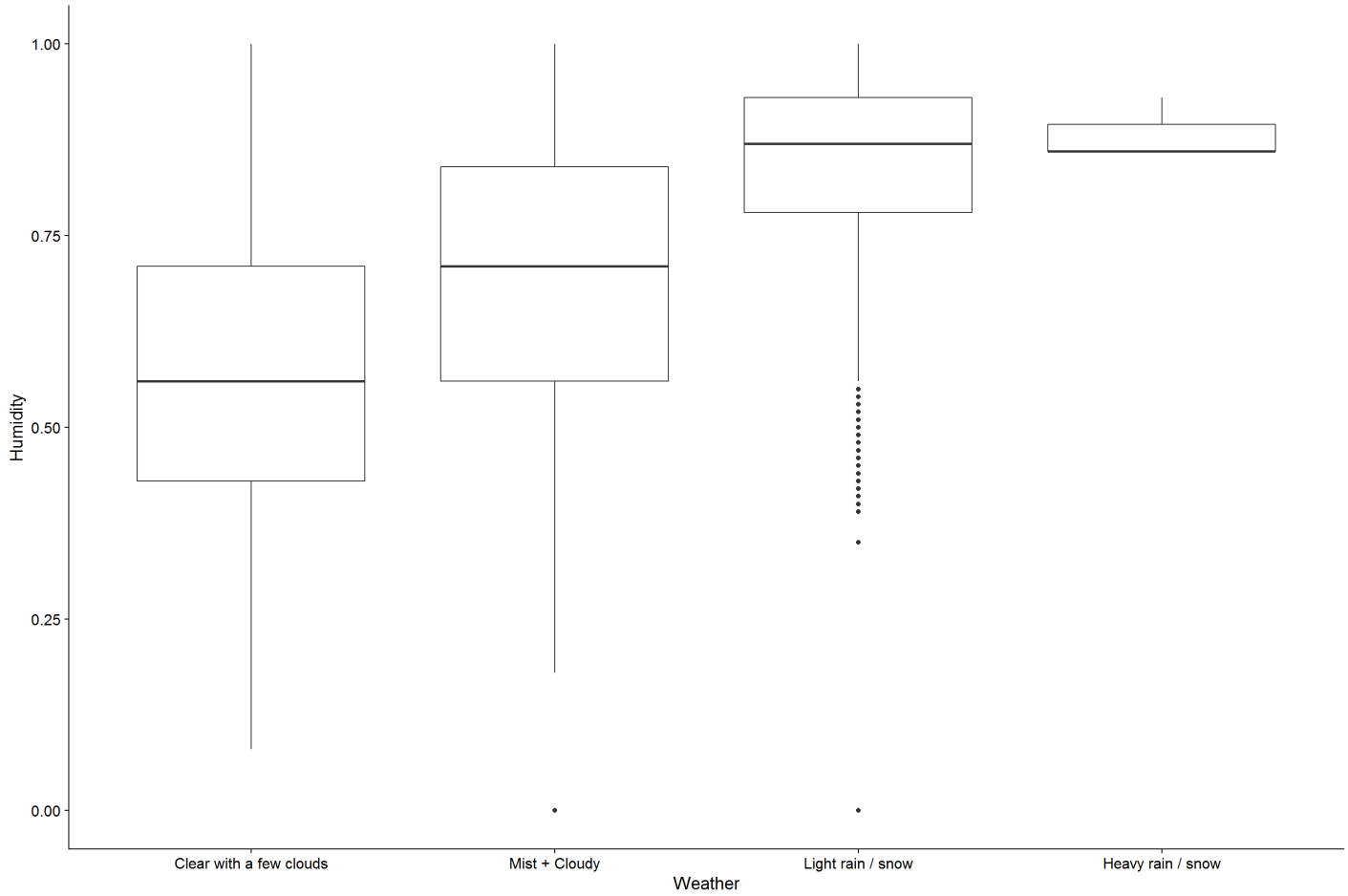
Let's see if there's any interplay between the independent variables in the dataset.



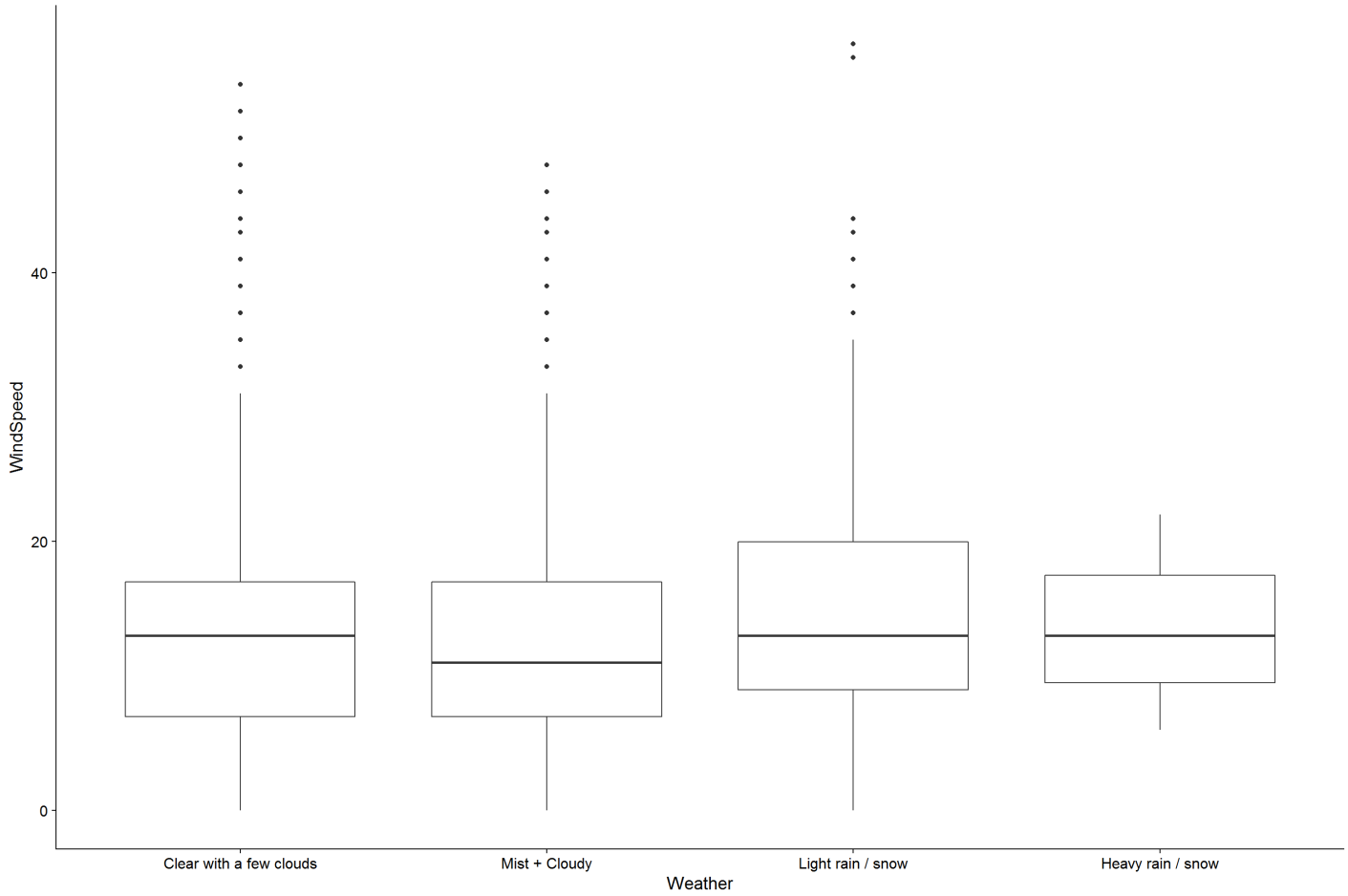
This is interesting. Even though temperature did not seem to have a direct impact on the rental numbers, the graph is strikingly similar to the one plotted between Season and the rental numbers.



Surprisingly, there isn't too much variability in the temperature with respect to weather, except during the days with heavy rain / snow. I was expecting a decreasing trend similar to the Weather vs. no. of rentals plot.



There is a clear relation between weather and humidity, with rainy and snowy days having high humidity. I wonder if temperature and humidity together affect the bike rentals. I will explore this in the multivariate plots.



The wind speed does not vary too much with respect to the weather conditions.

Bivariate analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

Analyzing the main features of interest (Registered and Casual) with the other variables revealed interesting insights. The most interesting was how the numbers varied based on whether a day was a working day or not. It was clear that the majority of registered bike rentals are used for commuting to work, whereas the majority of casual bike rentals are used by tourists / visitors.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

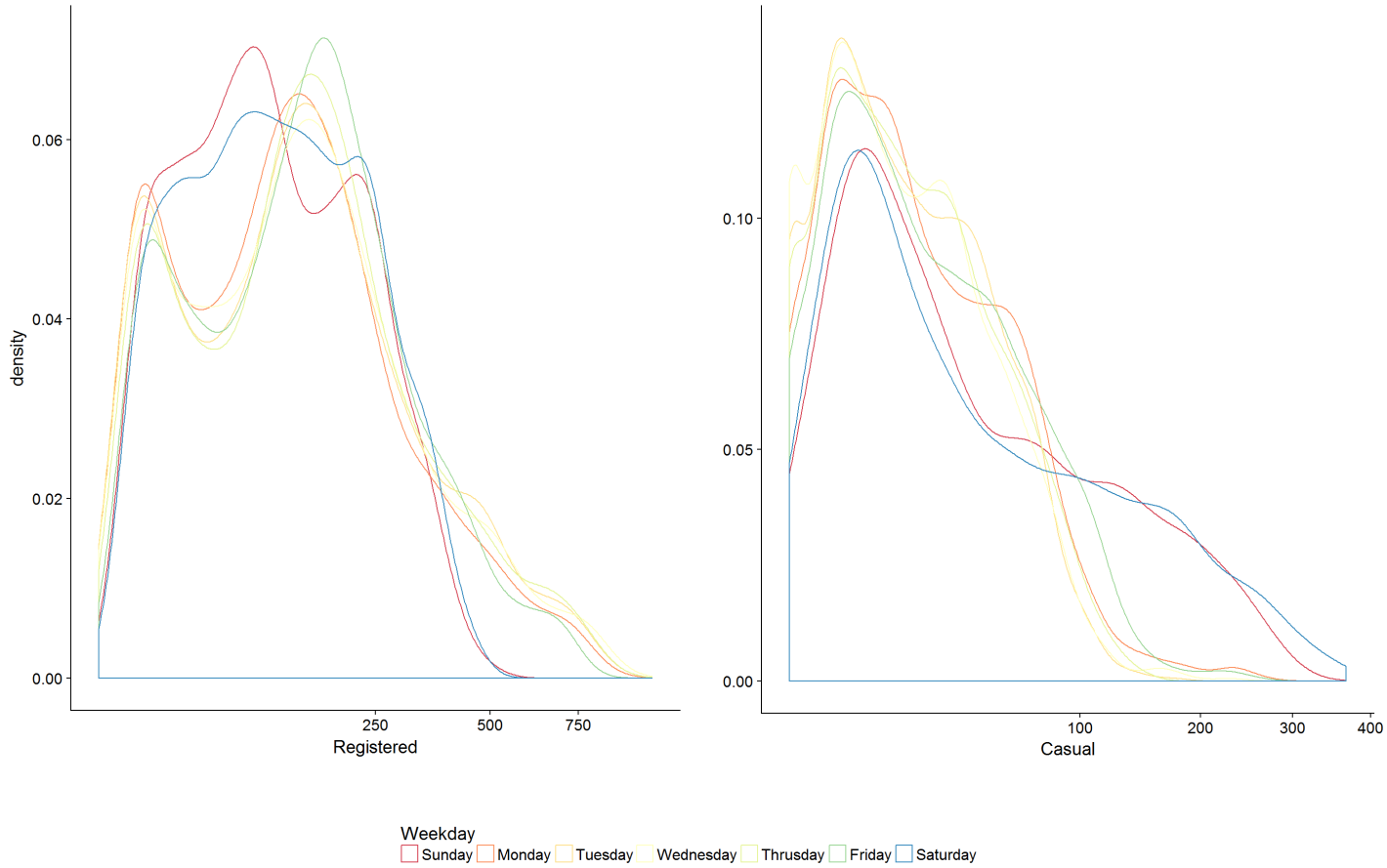
Yes. When I analyzed the impact of temperature and humidity on the no. of rentals, there wasn't a clear relationship from the plot. But, when I plotted temperature and season together, I could see that the relationship was almost identical to the one between season and the no. of rentals. This leads me to believe that temperature may indirectly impact the no. of rentals. Similarly, humidity also had a relationship with season, whereas it didn't seem to have any with the no. of rentals.

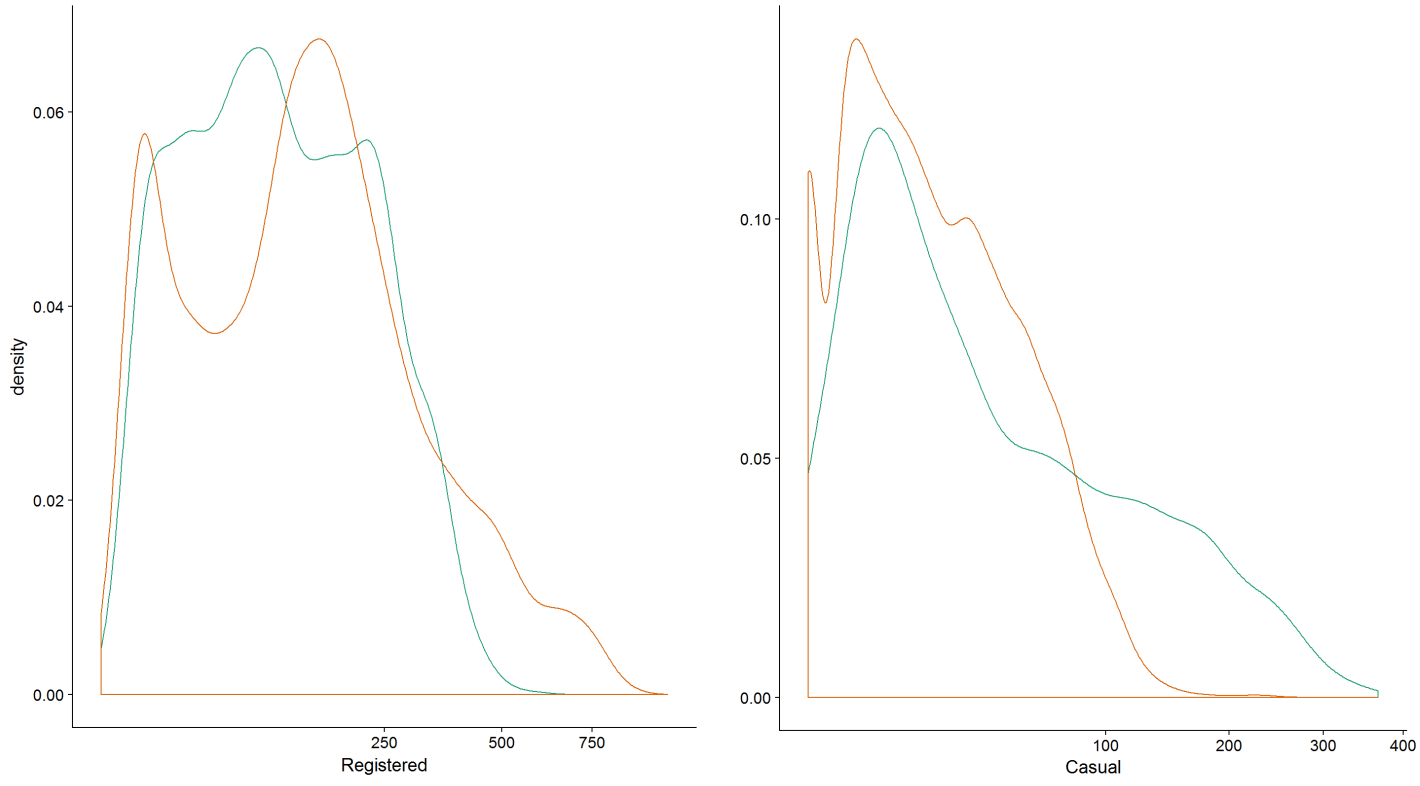
What was the strongest relationship you found?

The no. of rentals clearly depends on whether the day was a working day or not, and also the time of the day in case of the registered rentals. Also, the weather conditions and the season of the year definitely have an impact on the no. of rentals.

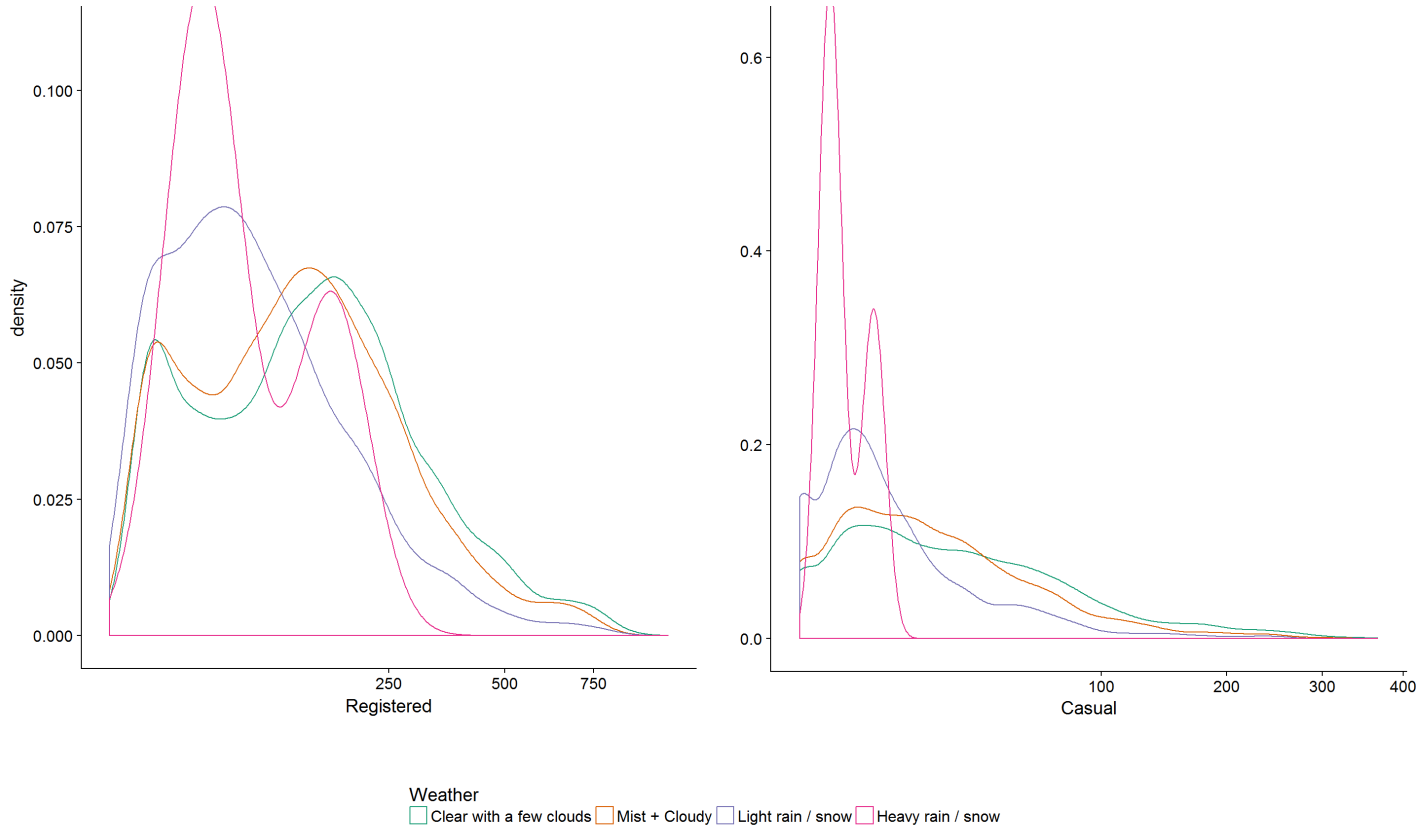
Multivariate plots

Let us look at some density plots, comparing Registered and Casual across different parameters.


density
0.06
0.04
0.02
0.00
250
500
750
Registered
0.10
0.05
0.00
100
200
300
400
Casual
Weekday
Sunday
Monday
Tuesday
Wednesday
Thrusday
Friday
Saturday

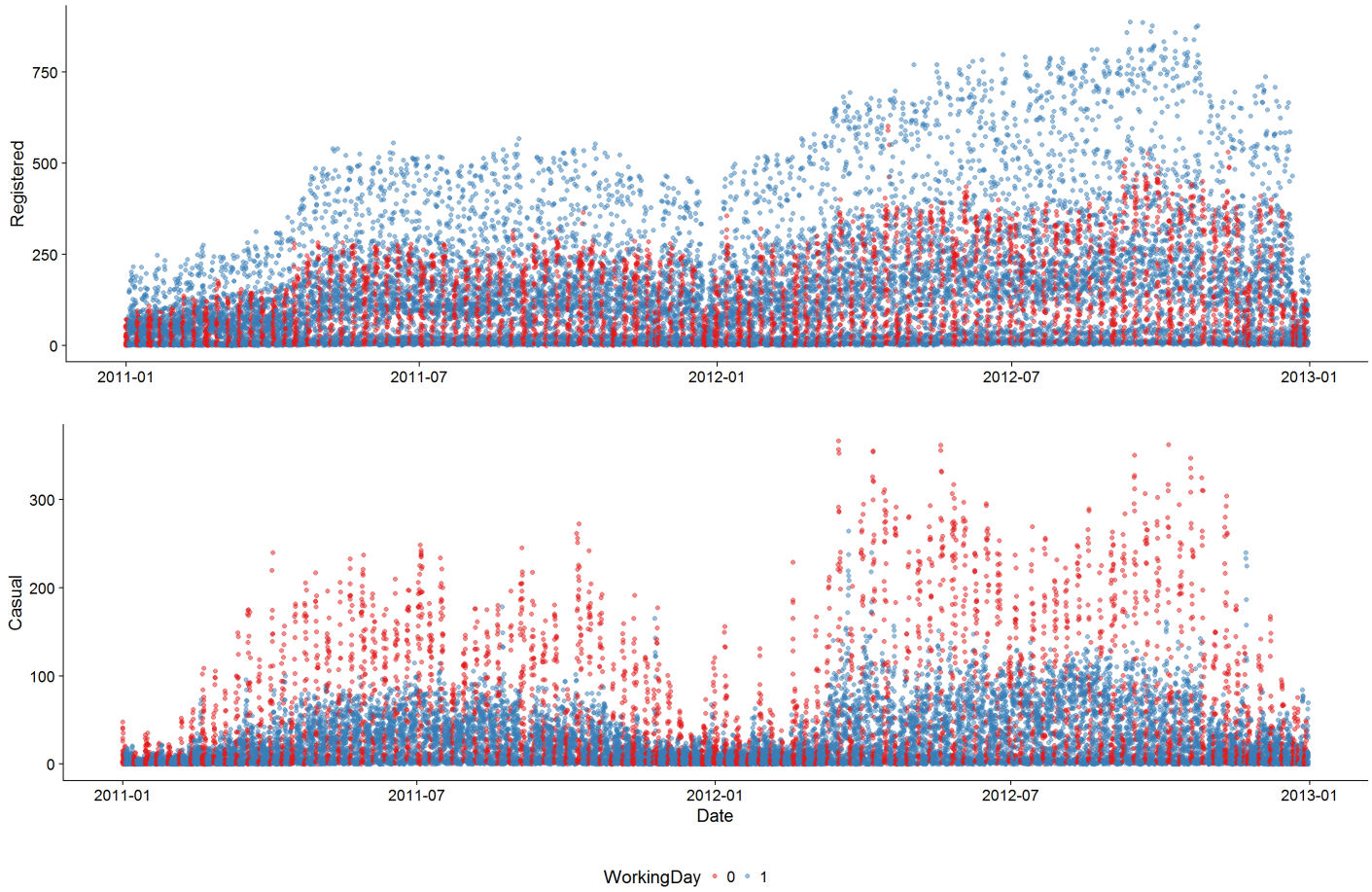

density
0.06
0.04
0.02
0.00
250
500
750
Registered
0.10
0.05
0.00
100
200
300
400
Casual
WorkingDay
0
1


0.125



These density plots again show that while weather has the same effect on Registered and Casual bike rentals, Weekday and WorkingDay have different effects.

Let us plot the bike rental numbers against date to see how the trend is. Let’s also color the plot by WorkingDay.

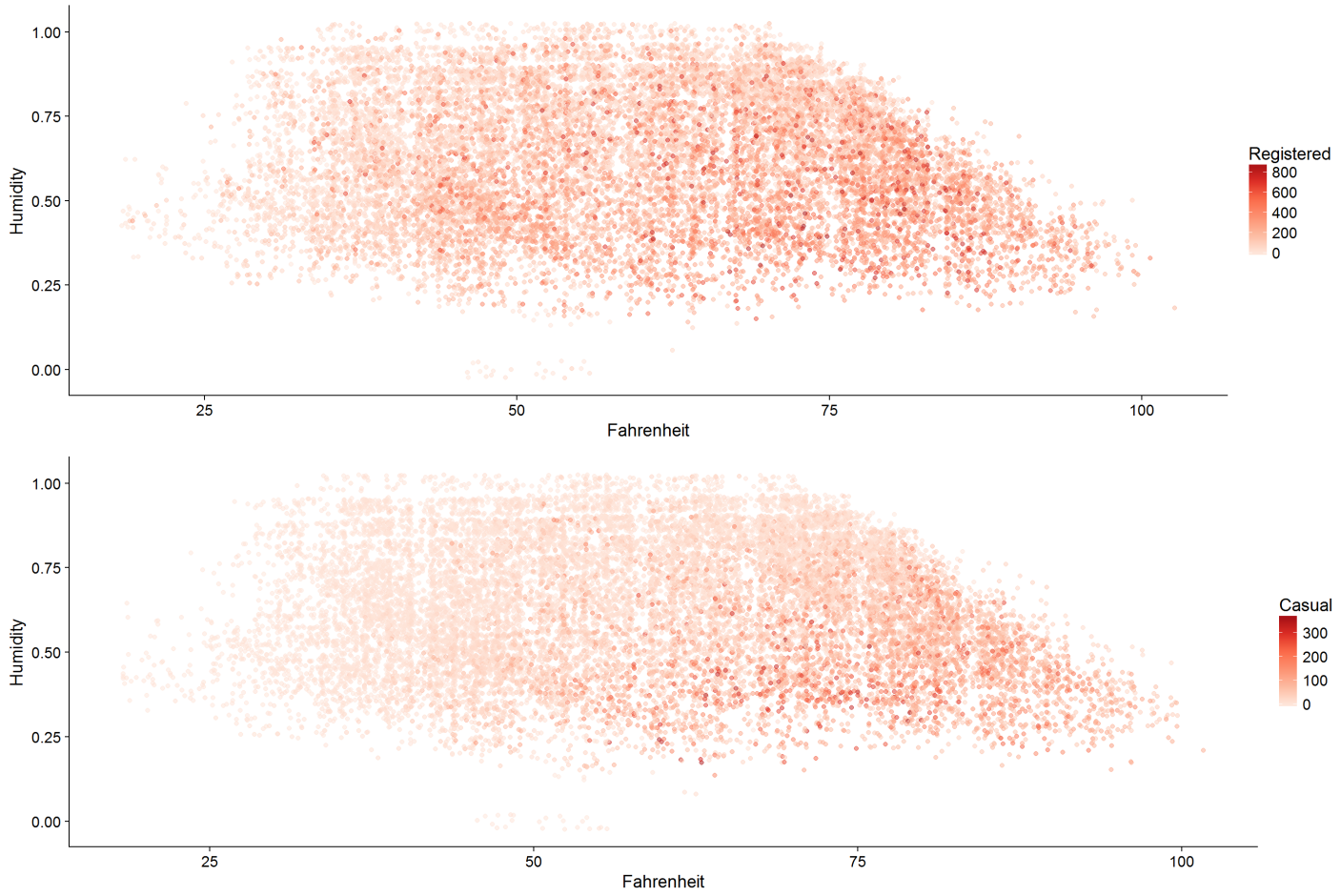


This represents the usage of registered and casual bike rentals over the entire time period given in the dataset. Although it is reinforcing the evidence that working day causes the difference between registered and casual, it does so in a visually pleasing plot.

It is also interesting to note the no. of rentals has increased in the second year, as can be seen by the higher peaks in 2012 than 2011 for both registered and casual rentals.

Earlier, we saw that clear days have high number of rentals, but humidity and temperature didn't seem to have a direct effect on the bike rental numbers. But, we did notice that humidity and temperature have a significant relation with the weather. Let us plot both temperature and humidity against weather, registered and casual to investigate this further.

This plot is dominated by clear days. Even on clear days, what is the ideal combination of temperature and humidity that leads to higher bike rentals?



These plots show a further distinction between the Registered and Casual rentals.

In the Registered plot, the dark spots indicating high rental numbers occur over a wide area, although it is just a bit concentrated around 60-80F at varying levels of humidity. In the Casual plot, though, we can see that the dark spots are concentrated in a very small area around 65-75F and around 40% humidity.

I think this is because if you are a regular subscriber, you are probably commuting to work (as we saw before) and you've gotta use it even though the conditions may not always be ideal. But if you're just a casual bike renter, you have more freedom to choose to do it only when the conditions are ideal.

We see that the no. of rentals is impacted by time of the day, week and year, and also by environmental conditions such as temperature, humidity and so on. Let's try to build a model around these variables.

```
##
## Calls:
## Model1: lm(formula = Registered ~ Season + Fahrenheit + Humidity + Weekday +
##     WorkingDay + Weather + Hour + Month + Year + WindSpeed, data = bikehours)
## Model2: lm(formula = Casual ~ Season + Fahrenheit + Humidity + Weekday +
##     WorkingDay + Weather + Hour + Month + Year + WindSpeed, data = bikehours)
## Model3: lm(formula = sqrt(Registered) ~ Season + Fahrenheit + Humidity +
##     Weekday + WorkingDay + Weather + Hour + Month + Year + WindSpeed,
##     data = bikehours)
## Model4: lm(formula = sqrt(Casual) ~ Season + Fahrenheit + Humidity +
##     Weekday + WorkingDay + Weather + Hour + Month + Year + WindSpeed,
##     data = bikehours)
##
## ================================================================================================
==============
##                                                      Model1         Model2       Model3
    Model4
## ------------------------------------------------------------------------------------------------
--------------
##   (Intercept)                                        32.522***     12.254***     5.773***
   2.330***
##                                                      (6.787)       (2.512)      (0.228)
  (0.138)
##   Season: Summer/Spring                              28.740***      9.756***     1.272***
   0.811***
##                                                      (4.079)       (1.510)      (0.137)
  (0.083)
##   Season: Fall/Spring                                30.084***      1.914        1.499***
   0.367***
##                                                      (4.830)       (1.788)      (0.162)
  (0.098)
##   Season: Winter/Spring                              65.616***      2.641        2.898***
   0.447***
##                                                      (4.101)       (1.518)      (0.138)
  (0.083)
##   Fahrenheit                                          1.682***      1.076***     0.069***
   0.083***
##                                                      (0.093)       (0.035)      (0.003)
  (0.002)
##   Humidity                                          -52.798***    -28.535***    -1.806***
  -1.836***
##                                                      (4.656)       (1.723)      (0.156)
  (0.094)
##   Weekday: Monday/Sunday                             -7.465       -10.269***    -0.565***
  -0.622***
##                                                      (4.266)       (1.579)      (0.143)
  (0.086)
##   Weekday: Tuesday/Sunday                            -2.935       -13.443***    -0.450**
  -0.902***
##                                                      (4.735)       (1.752)      (0.159)
  (0.096)
##   Weekday: Wednesday/Sunday                          -0.467       -13.403***    -0.337*
  -0.953***
```

```
##                                                          (4.731)       (1.751)      (0.159)
  (0.096)
##   Weekday: Thrusday/Sunday                               -0.853       -13.289***    -0.269
  -0.866***
##                                                          (4.699)       (1.739)      (0.158)
  (0.095)
##   Weekday: Friday/Sunday                                 -4.885        -5.453**     -0.200
  -0.257**
##                                                          (4.694)       (1.737)      (0.158)
  (0.095)
##   Weekday: Saturday/Sunday                               10.314***      5.621***     0.475***
   0.229***
##                                                          (2.417)       (0.895)      (0.081)
  (0.049)
##   WorkingDay: 1/0                                        48.081***    -20.892***     1.735***
  -1.326***
##                                                          (4.096)       (1.516)      (0.137)
  (0.083)
##   Weather: Mist + Cloudy/Clear with a few clouds         -7.032***     -3.658***    -0.229***
  -0.229***
##                                                          (1.612)       (0.597)      (0.054)
  (0.033)
##   Weather: Light rain / snow/Clear with a few clouds    -54.643***    -11.284***    -2.411***
  -1.271***
##                                                          (2.714)       (1.005)      (0.091)
  (0.055)
##   Weather: Heavy rain / snow/Clear with a few clouds    -60.118        -3.951       -1.676
  -0.608
##                                                         (49.473)      (18.311)      (1.660)
  (1.002)
##   Hour.L                                                250.651***     55.700***    13.565***
   6.210***
##                                                          (3.564)       (1.319)      (0.120)
  (0.072)
##   Hour.Q                                               -233.705***    -68.148***   -11.054***
  -5.782***
##                                                          (3.411)       (1.263)      (0.114)
  (0.069)
##   Hour.C                                               -132.075***    -53.091***    -5.718***
  -4.484***
##                                                          (3.556)       (1.316)      (0.119)
  (0.072)
##   Hour^4                                                -53.875***     22.573***     0.969***
   2.839***
##                                                          (3.222)       (1.192)      (0.108)
  (0.065)
##   Hour^5                                               -128.088***     16.488***    -6.682***
  -0.134*
##                                                          (3.232)       (1.196)      (0.108)
  (0.065)
##   Hour^6                                                135.606***     -3.288**      4.183***
  -0.566***
##                                                          (3.178)       (1.176)      (0.107)
  (0.064)
```

```
##   Hour^7                                  174.714***      0.387        6.275***
   0.995***
##                                            (3.178)       (1.176)      (0.107)
  (0.064)
##   Hour^8                                  -87.079***      0.901       -3.889***
  -0.305***
##                                            (3.179)       (1.177)      (0.107)
  (0.064)
##   Hour^9                                  -73.806***     -4.421***    -1.641***
  -0.580***
##                                            (3.178)       (1.176)      (0.107)
  (0.064)
##   Hour^10                                  13.401***     -4.193***     1.066***
   0.097
##                                            (3.178)       (1.176)      (0.107)
  (0.064)
##   Hour^11                                 -73.047***     -0.077       -2.974***
  -0.196**
##                                            (3.178)       (1.176)      (0.107)
  (0.064)
##   Hour^12                                  15.640***      4.076***     0.760***
   0.216***
##                                            (3.179)       (1.177)      (0.107)
  (0.064)
##   Hour^13                                 136.218***      1.791        3.952***
   0.256***
##                                            (3.180)       (1.177)      (0.107)
  (0.064)
##   Hour^14                                  -9.157**      -1.087       -0.651***
  -0.128*
##                                            (3.182)       (1.178)      (0.107)
  (0.064)
##   Hour^15                                 -91.549***     -1.436       -2.492***
  -0.120
##                                            (3.184)       (1.178)      (0.107)
  (0.065)
##   Hour^16                                  17.008***     -0.106        0.476***
   0.049
##                                            (3.184)       (1.178)      (0.107)
  (0.065)
##   Hour^17                                   4.004        -0.631        0.526***
  -0.062
##                                            (3.182)       (1.178)      (0.107)
  (0.064)
##   Hour^18                                 -42.501***      0.395       -1.123***
   0.014
##                                            (3.178)       (1.176)      (0.107)
  (0.064)
##   Hour^19                                  38.189***      3.104**      0.746***
   0.201**
##                                            (3.173)       (1.174)      (0.106)
  (0.064)
##   Hour^20                                  33.923***      1.853        0.934***
   0.139*
```

```
##                                                  (3.170)       (1.173)      (0.106)
  (0.064)
##   Hour^21                                      -18.762***      0.237       -0.432***
  -0.013
##                                                  (3.168)       (1.173)      (0.106)
  (0.064)
##   Hour^22                                       32.126***      0.864        0.605***
   0.081
##                                                  (3.167)       (1.172)      (0.106)
  (0.064)
##   Hour^23                                       15.075***      1.087        0.257*
   0.081
##                                                  (3.168)       (1.173)      (0.106)
  (0.064)
##   Month: .L                                     -8.753         3.806*      -0.660***
   0.361***
##                                                  (5.233)       (1.937)      (0.176)
  (0.106)
##   Month: .Q                                    -12.974        -0.088       -0.352
  -0.492***
##                                                  (6.776)       (2.508)      (0.227)
  (0.137)
##   Month: .C                                     -4.131        -1.060        0.085
   0.090
##                                                  (3.714)       (1.374)      (0.125)
  (0.075)
##   Month: ^4                                     -4.815       -12.189***    -0.184
  -0.860***
##                                                  (3.259)       (1.206)      (0.109)
  (0.066)
##   Month: ^5                                      6.616        -5.019***     0.182
  -0.369***
##                                                  (3.443)       (1.274)      (0.116)
  (0.070)
##   Month: ^6                                     15.627***      9.520***     0.416***
   0.636***
##                                                  (2.512)       (0.930)      (0.084)
  (0.051)
##   Month: ^7                                     15.822***      1.876        0.654***
   0.073
##                                                  (2.642)       (0.978)      (0.089)
  (0.054)
##   Month: ^8                                     -5.530*        1.172       -0.250**
   0.156**
##                                                  (2.496)       (0.924)      (0.084)
  (0.051)
##   Month: ^9                                    -15.811***     -2.328**     -0.615***
  -0.187***
##                                                  (2.315)       (0.857)      (0.078)
  (0.047)
##   Month: ^10                                     2.943         2.362**      0.164*
   0.254***
##                                                  (2.300)       (0.851)      (0.077)
  (0.047)
```

```
##   Month: ^11                                                 2.225        -1.885*       0.068
  -0.129**
##                                                           (2.313)       (0.856)      (0.078)
  (0.047)
##   Year: .L                                               51.981***      8.362***     1.898***
   0.570***
##                                                           (0.928)       (0.344)      (0.031)
  (0.019)
##   WindSpeed                                              -0.276**      -0.266***    -0.012***
  -0.020***
##                                                           (0.086)       (0.032)      (0.003)
  (0.002)
## ------------------------------------------------------------------------------------------
--------------
##   R-squared                                                  0.682         0.590       0.776
       0.770
##   adj. R-squared                                             0.681         0.588       0.775
       0.769
##   sigma                                                     85.466        31.633       2.868
       1.731
##   F                                                        728.934       488.062    1177.336
    1137.416
##   p                                                          0.000         0.000       0.000
       0.000
##   Log-likelihood                                       -101937.588    -84664.294  -42944.142
  -34174.510
##   Deviance                                          126564577.315  17338291.982  142516.590
   51947.657
##   AIC                                                  203981.177    169434.589   85994.285
   68455.019
##   BIC                                                  204392.617    169846.029   86405.725
   68866.459
##   N                                                     17379         17379       17379
   17379
## ==========================================================================================
==============
```

I had used a square root transformation in the density plots that we saw earlier. When this transformation is applied, we are able to explain around 77% of the variance in the bike rental numbers with the variables available in the dataset. Without the transformation though, this number is much lesser.

Multivariate analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

There were two relationships that stood out in this analysis that we had not seen before. * The bike rentals had actually increased in 2012 compared to 2011. This was clear from the date plot of casual and registered rentals. * In the bivariate analysis, we saw that temperature and humidity did not have much impact on the no. of bike

rentals directly. Here, when plotted together with the rental numbers, we were able to see the ideal conditions that resulted in high rental numbers.

Were there any interesting or surprising interactions between features?

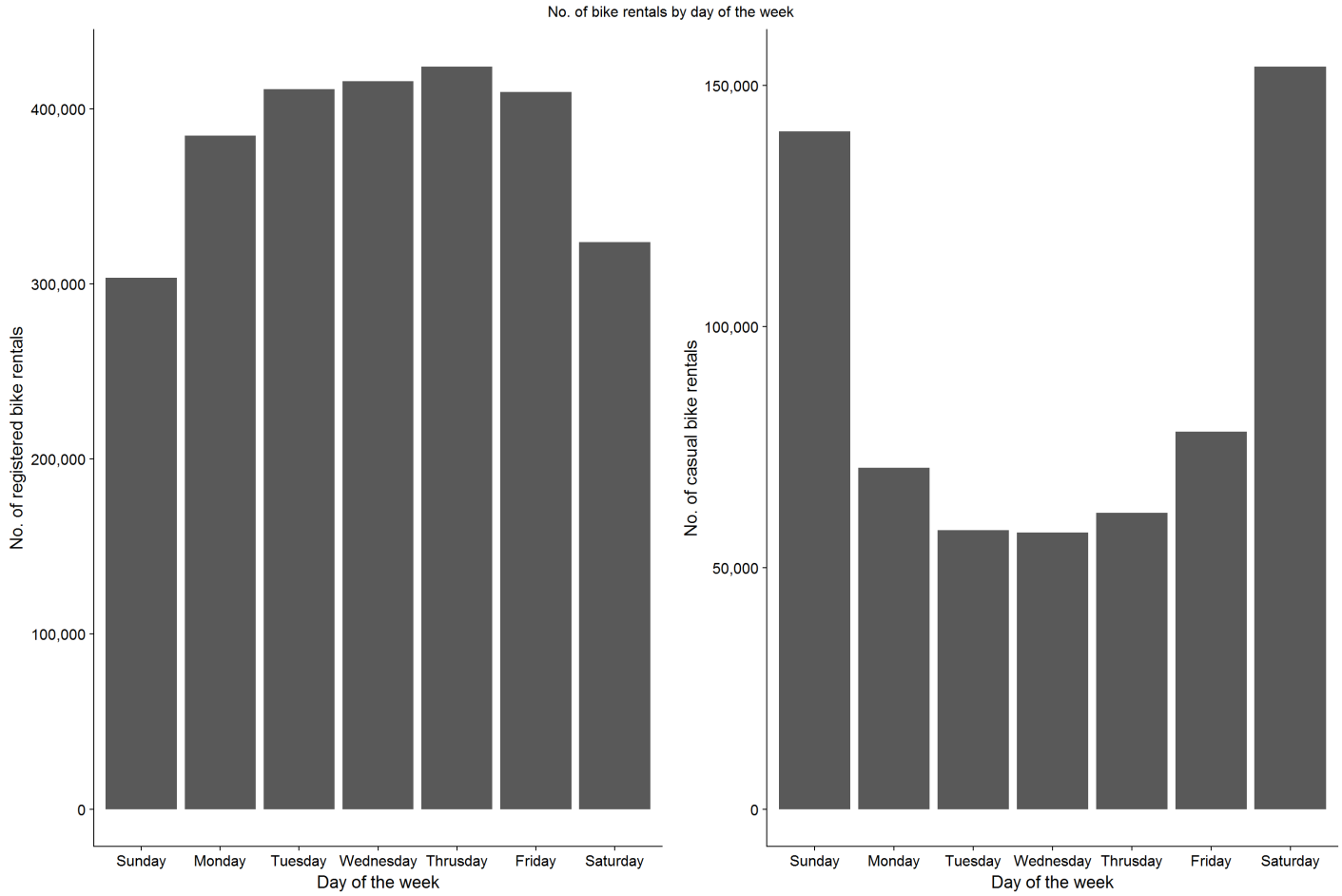
The combined effect of temperature and humidity was really surprising and unexpected.

Did you create any models with your dataset? Discuss the strengths and limitations of your model.

Yes. There were time-related and environmental variables that seemed to impact the no. of bike rentals. So I created a linear model to assess the predictive strength of the variables. Since I used a square root transformation for drawing the density plots, I checked if this transformation will give me a better model, and it did. The variables were able to explain 77% of the variance in the bike rental numbers.

Final Plots and Summary

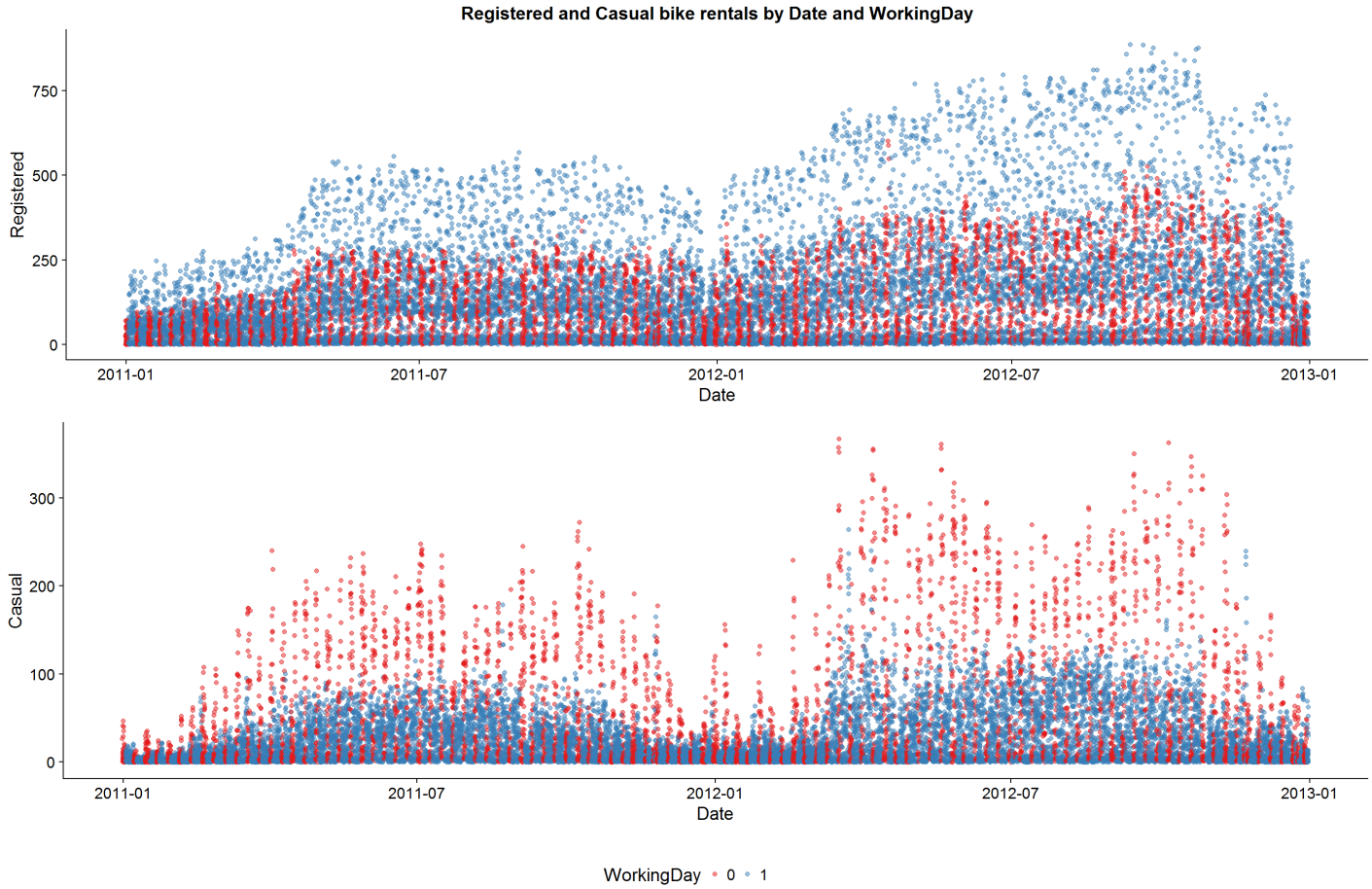
Plot One



Description One

The distribution of registered vs. casual bike rentals over the day of the week shows that registered bike rentals are probably used by people who commute to work, and thus higher over the weekdays. Casual bike rentals are used by people who are tourists or visitors and hence higher over the weekends.

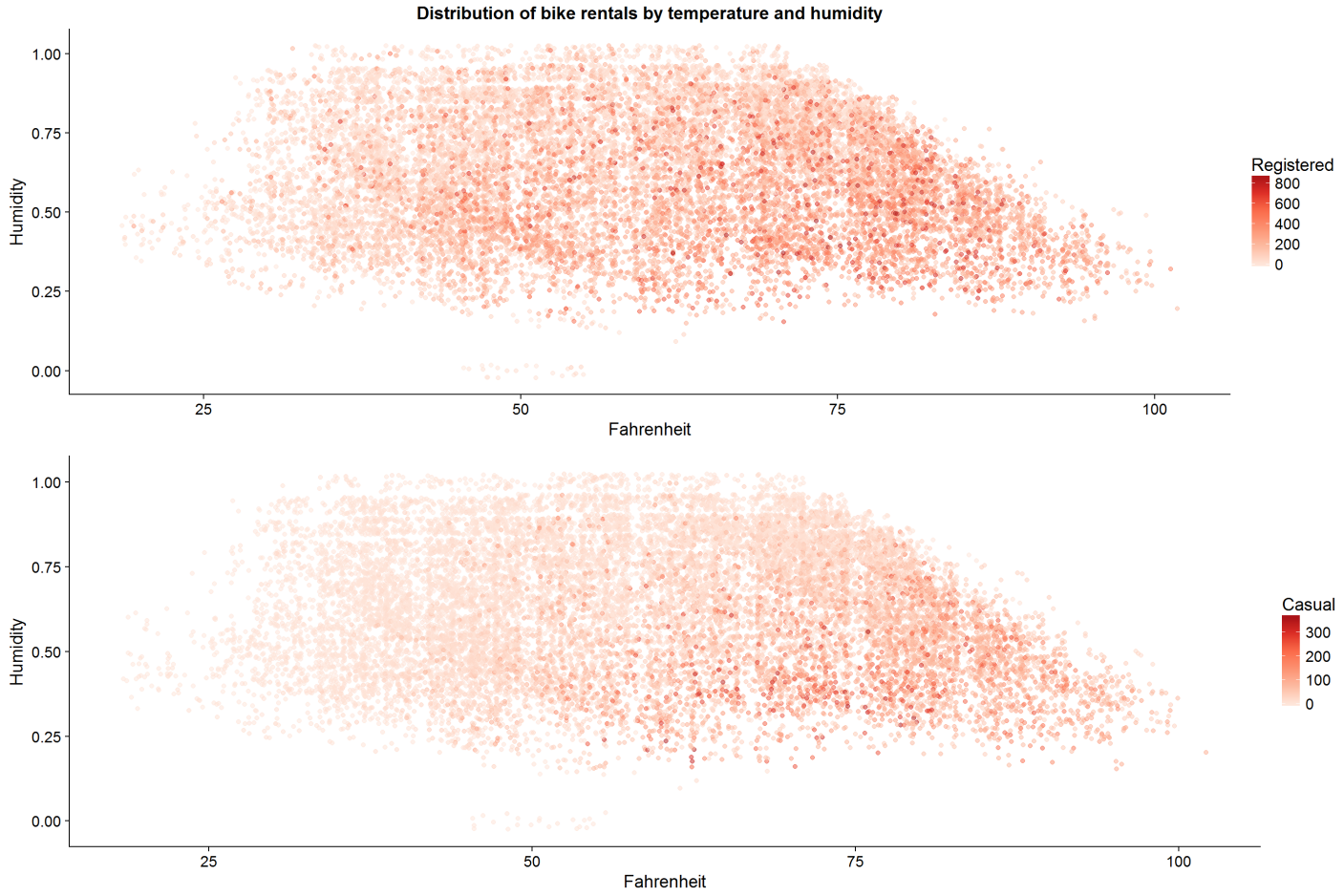
Plot Two



Description Two

This comparison plot shows two things - the increase in bike rentals in 2012 from 2011, and difference in usage of registered and casual bike rentals.

Plot Three



Description Three

This comparison plot shows a critical difference between the behavior registered and casual bike renters. While the registered bike rentals are not concentrated in a specific zone in the scatter plot, the casual rentals are. This shows that casual renters have a greater freedom to rent bikes only when it is ideal to do so - this may also explain why the casual rental numbers are quite low when compared to the registered ones.

Reflection

The bike sharing dataset contains more than 17000 observations of bike rental numbers in the registered and casual categories, supported by a host of variables. These variables can be classified into two types: temporal and environmental. Temporal, or time-based variables include weekday, workingday, hour, month, year and season. Environmental variables include weather, humidity, temperature and windspeed.

After understanding the distrbution of the individual variables in the dataset, I started exploring the relationships between the different variables and how they affect the no. of bike rentals. I found that there was a clear difference in how the day of the week affected the registered and casual bike rentals. This was further confirmed by the impact of Holiday and WorkingDay variables on the bike rentals. Apart from the day of the week, the hour of the day also had an impact on the no. of bike rentals.

While all these time-related variables had a clear relation with the bike rentals, the relation with environmental variables wasn’t easy to decipher. The weather did have some impact but it looked very generic and easy to predict - clear days obviously had higher bike rentals. I couldn’t easily figure out how temperature and humidity affected the dependent variables, until I plotted them together.

With all these variables, I was able to create a linear model, which was able to explain around 77% of the variance of both the registered and casual bike rentals.

I think there may be other variables not included in the dataset that could increase the predictive power of the model. For example, traffic conditions, road conditions and days of important public events could be important factors. Also, if we had data for a longer time period, that could also increase the accuracy.