Data583 EDA

Vimaljeet Singh

2023-03-09

What affects the number of bike rides the most? we could do a random forest?

Maybe want to do feature selection as answering one of the scientific questions we set out; which factors affect the most? and a

what are examples statistical analysis techniques we learnt in previous courses?

EDA

Workflow for current and next step in the project

To determine which model would fit this data, we need to consider several factors such as the nature of the problem, the available data, and the type of output that we are trying to predict.

First, we need to define the problem and the goal of the model. In this case, we are trying to **predict the number of bikes rented per hour (cnt)**, given the various features such as weather, time of day, and other factors. This is a regression problem, as we are trying to predict a discrete variable.

Second, we need to examine the available data and determine if there are any missing values, outliers, or other anomalies. If the data is incomplete, we may need to consider techniques such as imputation or data cleaning to address these issues.

Third, we need to select appropriate features for the model. Some features, such as season or weather, may have a strong correlation with the number of bikes rented, while others may not be as important. Feature selection can be done using techniques such as correlation analysis or principal component analysis.

Finally, we can select a model that is appropriate for the problem at hand. Some popular models for regression problems include linear regression, decision trees, random forests, and neural networks. We can use techniques such as cross-validation to evaluate different models and select the one that performs best on the data.

```r
myday = read.csv("day.csv", header = TRUE)
myday$dteday = as.Date(myday$dteday, format = "%Y-%m-%d") # converting 'dteday' column to date
```

The dteday column has been explicitly converted to date format, it was ‘char’ by default.

```r
head(myday, 5)
```

```
##   instant     dteday season yr mnth holiday weekday workingday weathersit
## 1       1 2011-01-01      1  0    1       0       6          0          2
## 2       2 2011-01-02      1  0    1       0       0          0          2
## 3       3 2011-01-03      1  0    1       0       1          1          1
## 4       4 2011-01-04      1  0    1       0       2          1          1
## 5       5 2011-01-05      1  0    1       0       3          1          1
##       temp    atemp      hum windspeed casual registered  cnt
## 1 0.344167 0.363625 0.805833  0.160446    331        654  985
## 2 0.363478 0.353739 0.696087  0.248539    131        670  801
## 3 0.196364 0.189405 0.437273  0.248309    120       1229 1349
## 4 0.200000 0.212122 0.590435  0.160296    108       1454 1562
## 5 0.226957 0.229270 0.436957  0.186900     82       1518 1600
```

```
min(myday$dteday)
```

```
## [1] "2011-01-01"
```

```
max(myday$dteday)
```

```
## [1] "2012-12-31"
```

This is the date range for the data set as seen above is from January 1, 2011 to December 31, 2012.

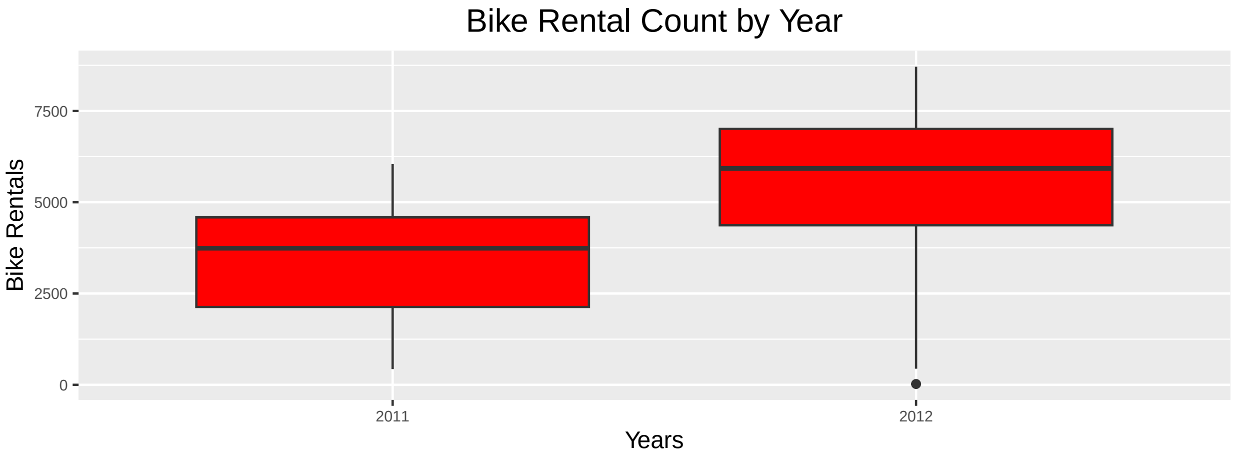
```
colSums(is.na(myday))
```

```
##    instant     dteday     season         yr       mnth    holiday    weekday
##          0          0          0          0          0          0          0
## workingday weathersit       temp      atemp        hum  windspeed     casual
##          0          0          0          0          0          0          0
## registered        cnt
##          0          0
```

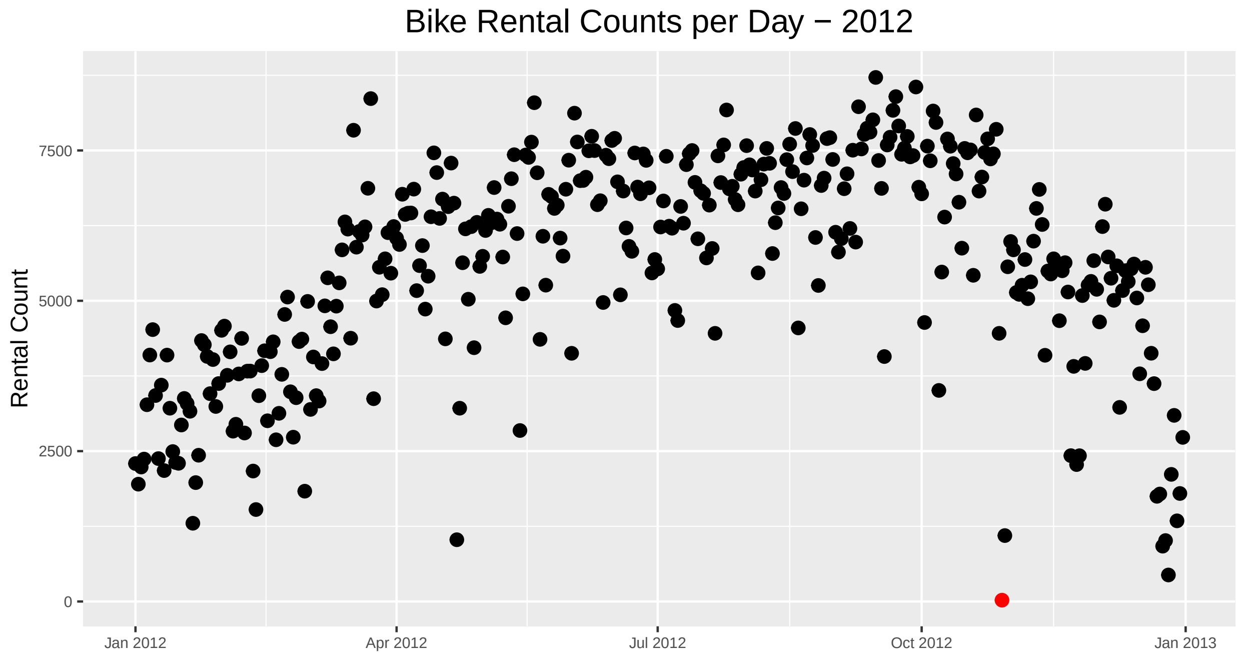
As shown in project proposal, the following varaibles were divided by the maximum value in the column, so we will get back original values and save it into new columns to help make charts and understand and visualize data better. - temp : Normalized temperature in Celsius. The values are divided to 41 (max) - atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max) - hum: Normalized humidity. The values are divided to 100 (max) - windspeed: Normalized wind speed. The values are divided to 67 (max)

To see if Bike Rental Count is different for both years

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.1     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```



- The number of bikes rented out has increased from 2011 in 2012.
- Also, we notice the outlier in 2012 which can be seen as a dot towards the bottom of the plot, lets explore that.
- The following chart marks it in red so that it is easy to see.

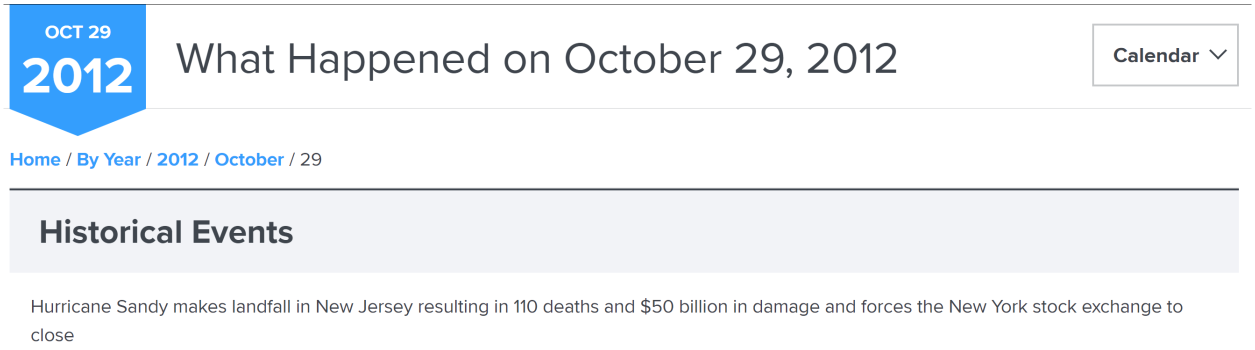


Let us explore this further:

```
myday_2012 <- subset(myday, yr == 1)
mydata = myday_2012[myday_2012$cnt == min(myday_2012$cnt), c("cnt","dteday")]
cbind(yearly_mean = mean(myday_2012$cnt), mydata)
```

```
##     yearly_mean cnt     dteday
## 668    5599.934  22 2012-10-29
```

We see that the number of bikes rented out on October 29, 2012 was underwhelmingly lower than the yearly average for 2012 making it an outlier. We investigated further to see why that is and found the following:



What Happened on October 29, 2012

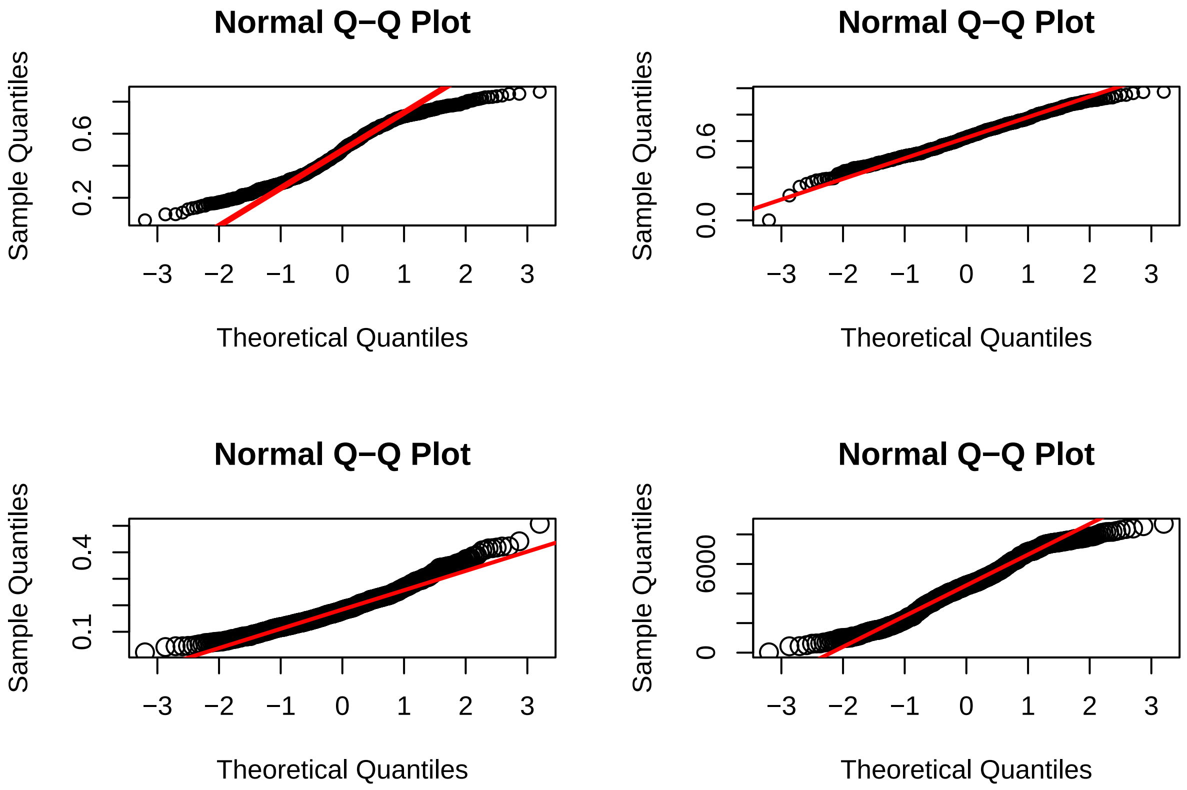
Home / By Year / 2012 / October / 29

Historical Events

Hurricane Sandy makes landfall in New Jersey resulting in 110 deaths and $50 billion in damage and forces the New York stock exchange to close

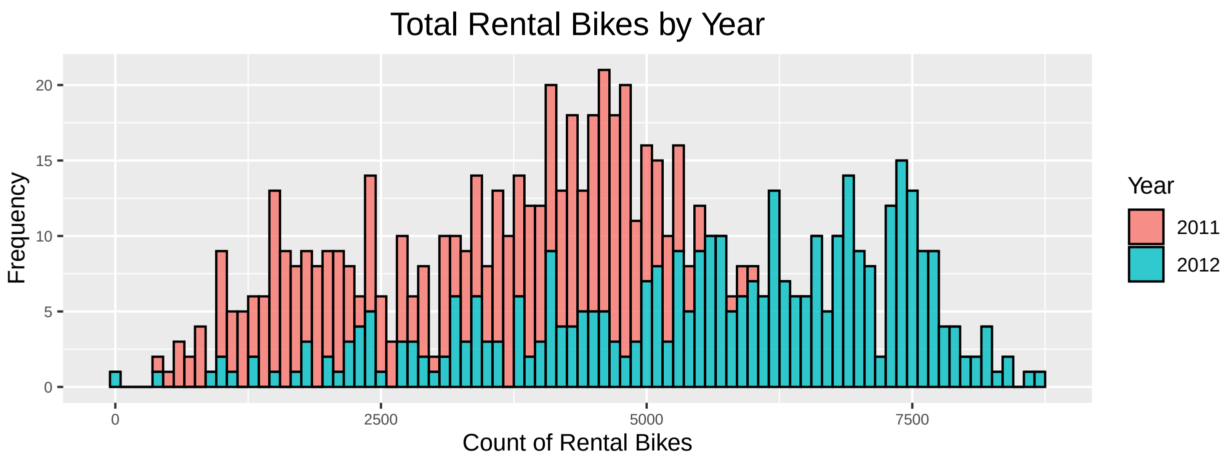
It was because of this natural calamity that the number of bike rental counts was super low. We have also successfully used this data set to find a natural calamity.

See if variables are normally distributed



We see that the variables, temp (top-left), humidity (top-right), windspeed (bottom-left) and cnt (bottom-right) are not normally distributed. Let us confirm the normality of the cnt variable using histograms and a test.

Plotting histogram to see distribution of count variable for both years



The graph shows us that the cnt variable is not normally distributed. A Shapiro test should corroborate this.

Shapiro Wilk test for 'cnt'

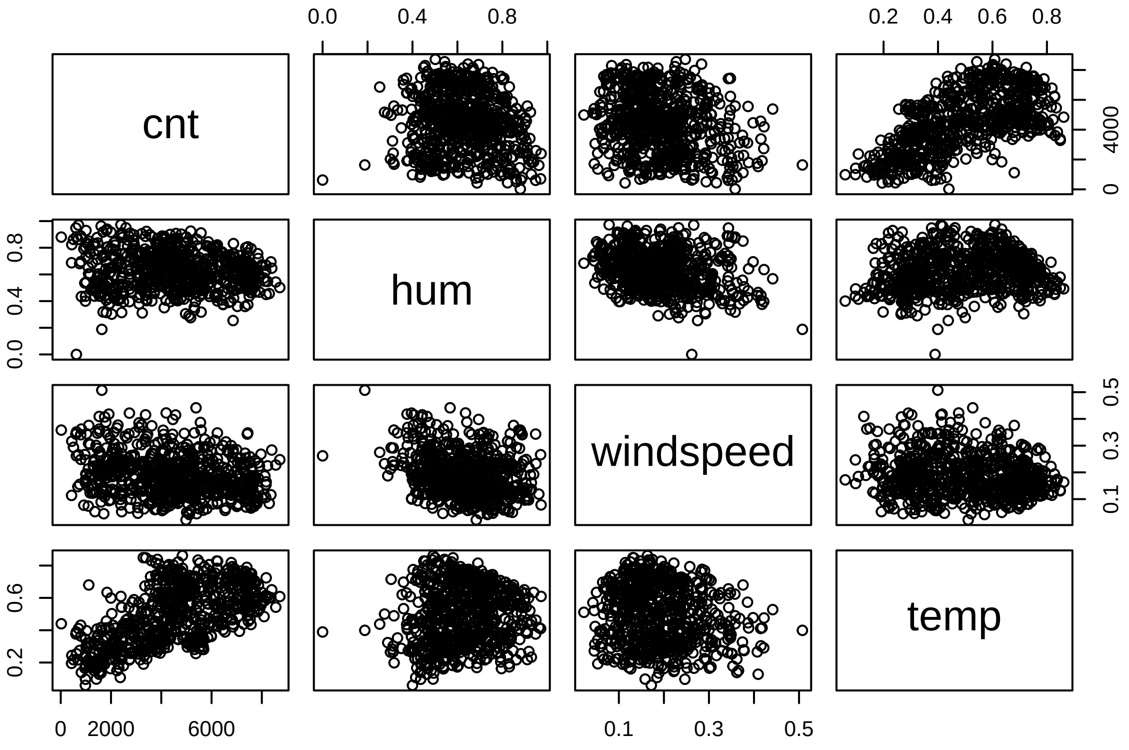
```
shapiro.test(myday$cnt)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  myday$cnt
## W = 0.98012, p-value = 2.081e-08
```

As the pvalue is really low ($\alpha = 0.05$), we do not have enough evidence for the null hypothesis and we conclude that cnt data is not normally distributed.

Scatter plot of reponse variable with 'hum', 'windspeed' and 'temp'

```
pairs(myday[, c("cnt", "hum", "windspeed", "temp")])
```



Correlations

```
cor.hum <- cor.test(x = myday$cnt, y = myday$hum)
cor.temp <- cor.test(x = myday$cnt, y = myday$temp)
cor.ws <- cor.test(x = myday$cnt, y = myday$windspeed)
cbind(corr_hum=cor.hum[4], corr_temp=cor.temp[4], corr_ws=cor.ws[4])
```

```
##           corr_hum   corr_temp corr_ws
## estimate -0.1006586 0.627494  -0.234545
```

- Correlation of cnt with humidity -0.1006586
- Correlation of cnt with temp 0.627494
- Correlation of cnt with windspeed -0.234545

The only somewhat correlation that cnt shows with is temp, it seems to have almost no correlation with humidity and windspeed. Find the chart below that shows correlation between raw temperatures and type of users.

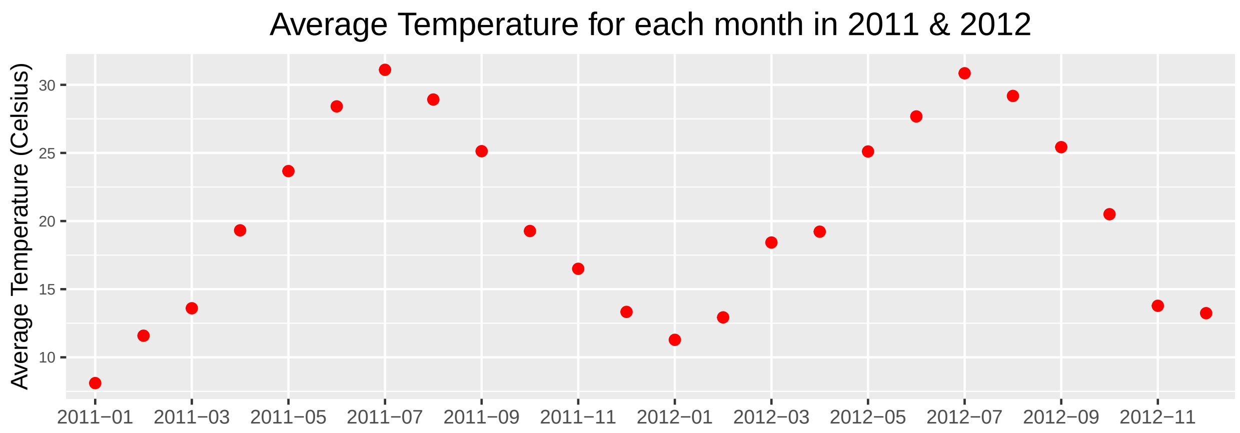
Check if 'weathersit' and 'holiday' variables are important using ANOVA

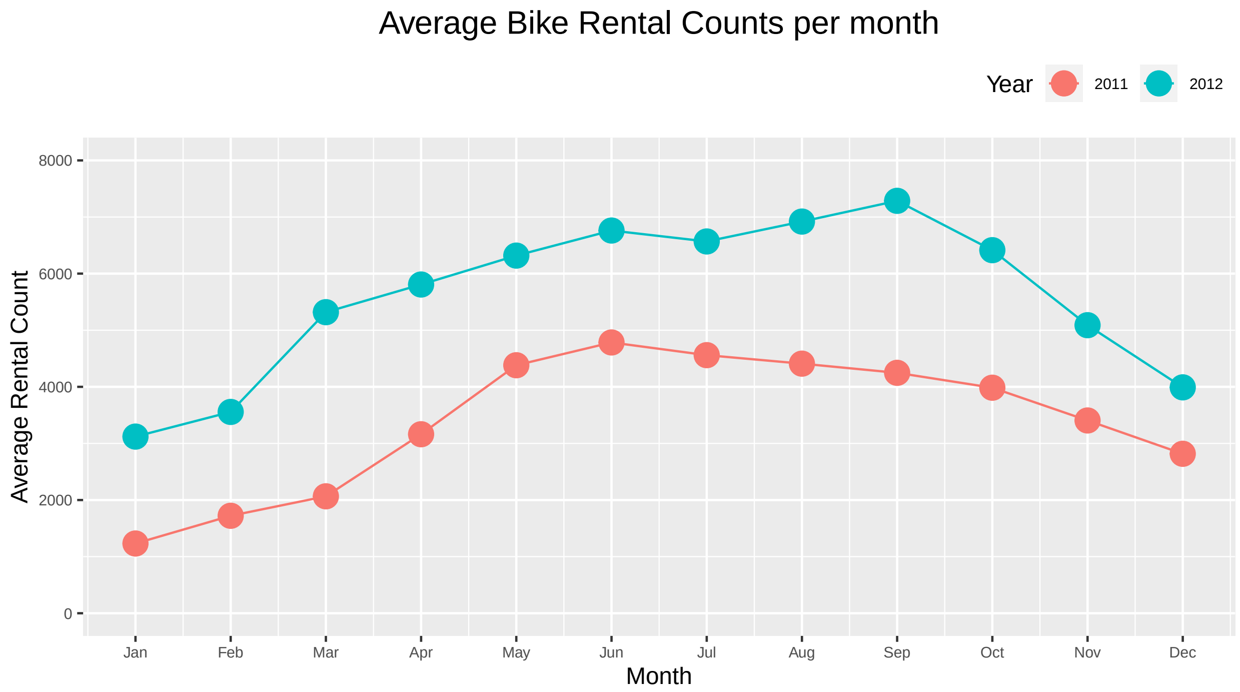
```
lm <- lm(cnt ~ weathersit + holiday, data = myday)
anova (lm)
```

```
## Analysis of Variance Table
##
## Response: cnt
##             Df     Sum Sq   Mean Sq F value  Pr(>F)
## weathersit   1  242288753 242288753 71.1154 < 2e-16 ***
## holiday      1   16964654  16964654  4.9794 0.02595 *
## Residuals  728 2480281985   3406981
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

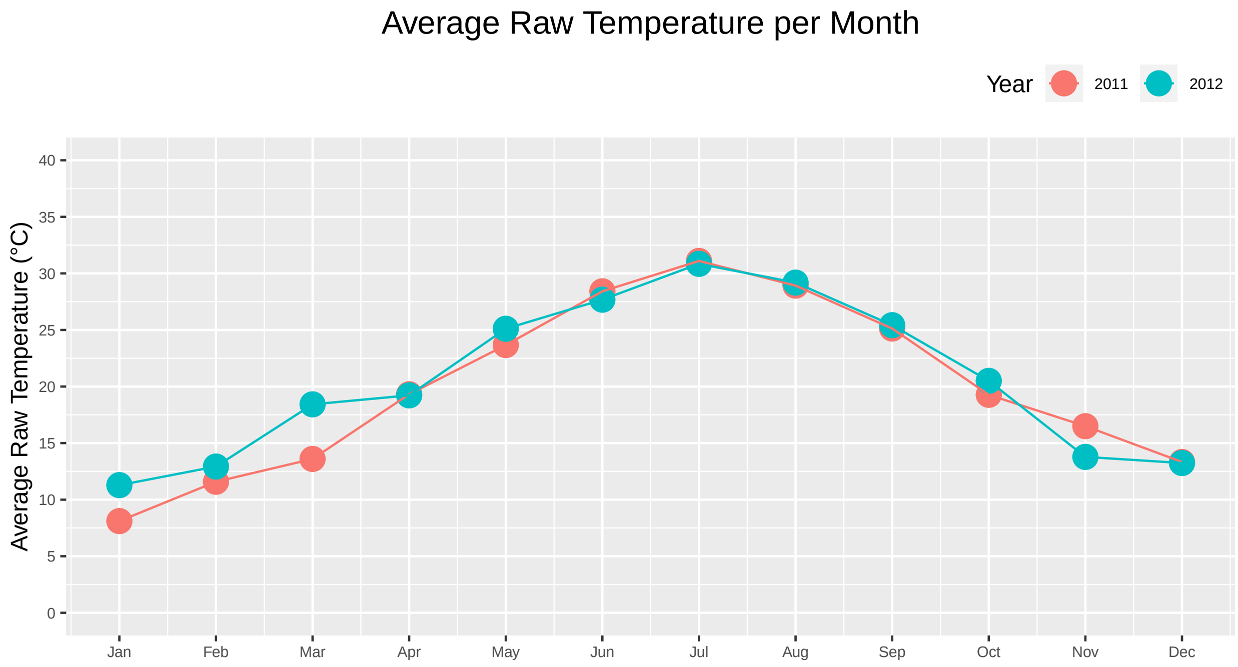
F values greater than 1, and pvalues less than 0.05 show that both these variables are significant in predicting count of bike rentals.

Some exploratory charts

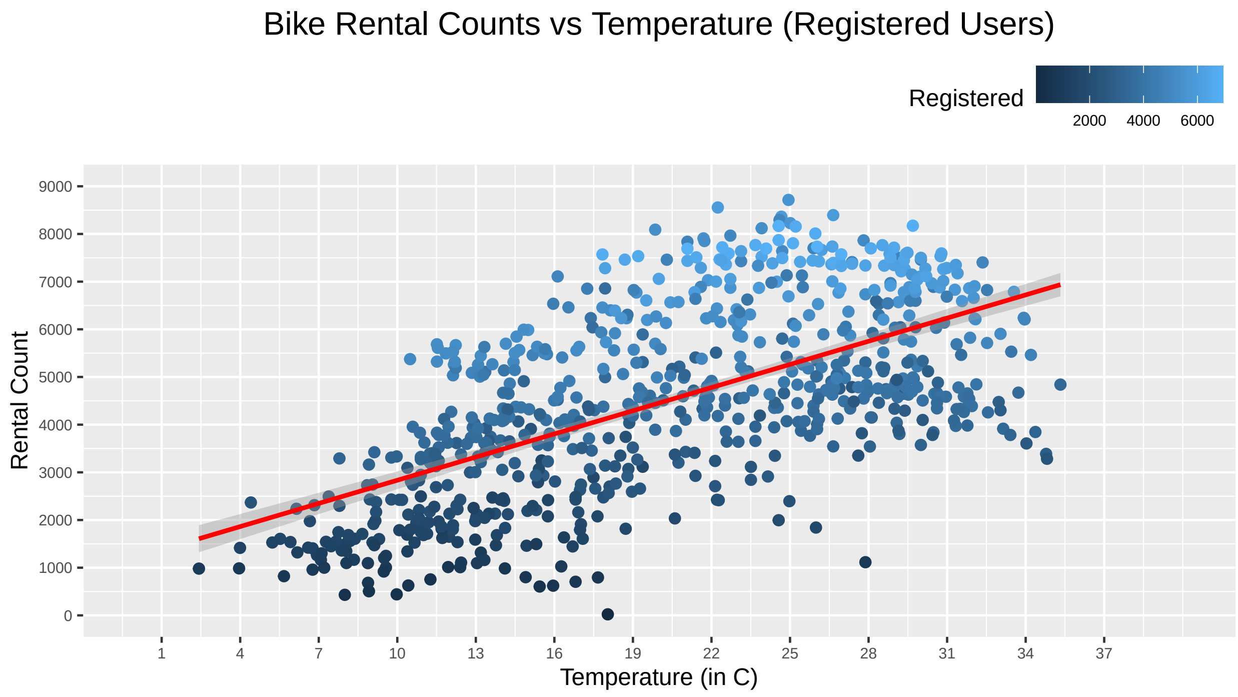




We note that the bike rentals went were higher in the year 2012 as compared to 2011, did temperature have a role to play in this? We can make a similar plot for average raw temperatures to see if the temperatures varied in these two years causing an increase in the bike rental count.



```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```


Bike Rental Counts vs Temperature (Registered Users)
Registered
Rental Count
Temperature (in C)