Bike Sharing

REHAM ALSHEHRI

2/5/2022

Contents

#install.packages("corrplot") #library(randomForest) #if (!require("libraryname")) install.packages("libraryname")

1 INTRODUCTION

Bike sharing services are a new version of conventional bike rentals in which the whole procedure from registration to renting and returning has been automated. Consumers may sign up to become subscribers of these platforms or simply use services on a temporary contract, which allows individuals to grab up the bike place at a single point and leave it off at many sites across cities. Such platform is transforming the way people utilize public transit, particularly in and around major metropolitan areas. Enterprises are in competition with other modes of transportation. Understanding the quantity of utilization from each station throughout the city is incredibly significant in demonstrating development trends of the service, particularly when compared to other, more conventional forms of transportation.

The goal of this case study is to investigate and develop a linear regression model in order to forecast bike sharing consumption using the weather information from the dataset. The work will undertake exploratory investigation, data analysis, data modelling, conclusion and limitation of the work on Hadi Fanaee Tork's Bike Sharing Demand data set, which was compiled utilizing data provided by Capital Bikeshare. Bike sharing services are a kind of bike hire in which the procedure of getting a subscription, borrowing a bike,

and returning it is computerized via the use of a system of kiosks stations located around a metropolis. The services let individuals to hire a bike from a single place and deposit it to another on an as-needed base. There are already over 500 bike-sharing schemes worldwide.

```
## Warning: package 'readxl' was built under R version 4.0.5
```

2 Data Analysis

section that explains the process and techniques used, including data cleaning, data exploration and data visualization. For this purpose various types of the methods and approaches are considered to perform the data analysis.

```
## tibble [730 x 16] (S3: tbl_df/tbl/data.frame)
##  $ instant   : num [1:730] 1 2 3 4 5 6 7 8 9 10 ...
##  $ dteday    : chr [1:730] "43101" "43132" "43160" "43191" ...
##  $ season    : num [1:730] 1 1 1 1 1 1 1 1 1 1 ...
##  $ yr        : num [1:730] 0 0 0 0 0 0 0 0 0 0 ...
##  $ mnth      : num [1:730] 1 1 1 1 1 1 1 1 1 1 ...
##  $ holiday   : num [1:730] 0 0 0 0 0 0 0 0 0 0 ...
##  $ weekday   : num [1:730] 6 0 1 2 3 4 5 6 0 1 ...
##  $ workingday: num [1:730] 0 0 1 1 1 1 1 0 0 1 ...
##  $ weathersit: num [1:730] 2 2 1 1 1 1 2 2 1 1 ...
##  $ temp      : num [1:730] 14.11 14.9 8.05 8.2 9.31 ...
##  $ atemp     : num [1:730] 18.18 17.69 9.47 10.61 11.46 ...
##  $ hum       : num [1:730] 80.6 69.6 43.7 59 43.7 ...
##  $ windspeed : num [1:730] 10.7 16.7 16.6 10.7 12.5 ...
##  $ casual    : num [1:730] 331 131 120 108 82 88 148 68 54 41 ...
##  $ registered: num [1:730] 654 670 1229 1454 1518 ...
##  $ cnt       : num [1:730] 985 801 1349 1562 1600 ...
```

2.1 Data Exploration

There are total '16' columns in the data set and their deatils are given below with their description.

Original column names: Description instant: record index dteday : date season : season (1:springer, 2:summer, 3:fall, 4:winter) yr : year (0: 2011, 1:2012) mnth : month (1 to 12) hr : hour (0 to 23) holiday : whether day is holiday or not (extracted from [Web Link]) weekday : day of the week workingday : if day is neither weekend nor holiday is 1, otherwise is 0. weathersit : 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 temp : Normalized temperature in Celsius. atemp: Normalized feeling temperature in Celsius. hum: Normalized humidity. The values are divided to 100 (max) windspeed: Normalized wind speed. The values are divided to 67 (max) casual: count of casual users registered: count of registered users cnt: count of total rental bikes including both casual and registered

2.2 Data Cleaning

Data cleaning is a very important process and some of the irrelevant variables from the dataset are removed for more accurate results.For this purpose, 6 variables (instant, dteday, yr, weathersit, atemp,casual,registered) and new dataset is created on the basis of remaining variables.

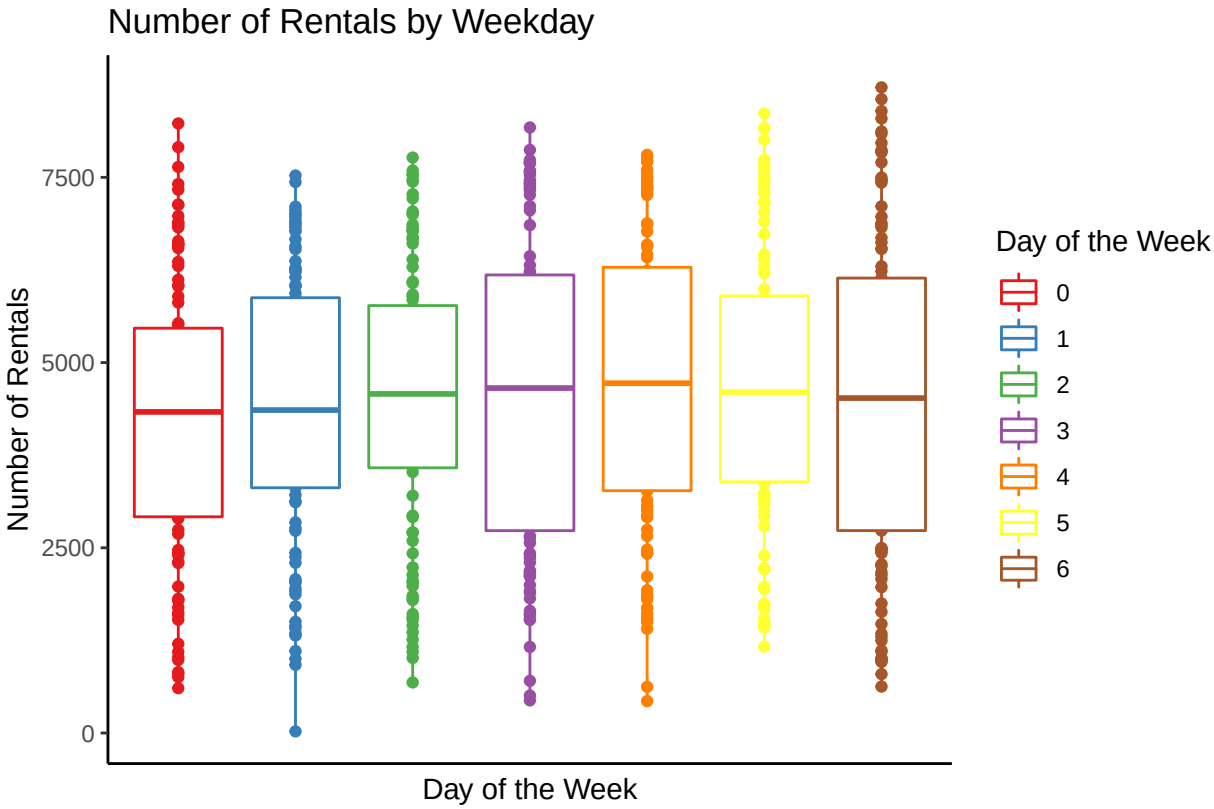
##Data Visualization

Data visulization is very helpful process to find the useful insights from it. To fulfil this need, boxplots and barcahrts are plotted for the various types of the varibales.

2.2.1 Weekdays

The rental numbers for each weekdays are being displayed here using the boxplot. We can see that there are various types of the outliers in the dataset. For each weekday, the rental numbers are shown here.

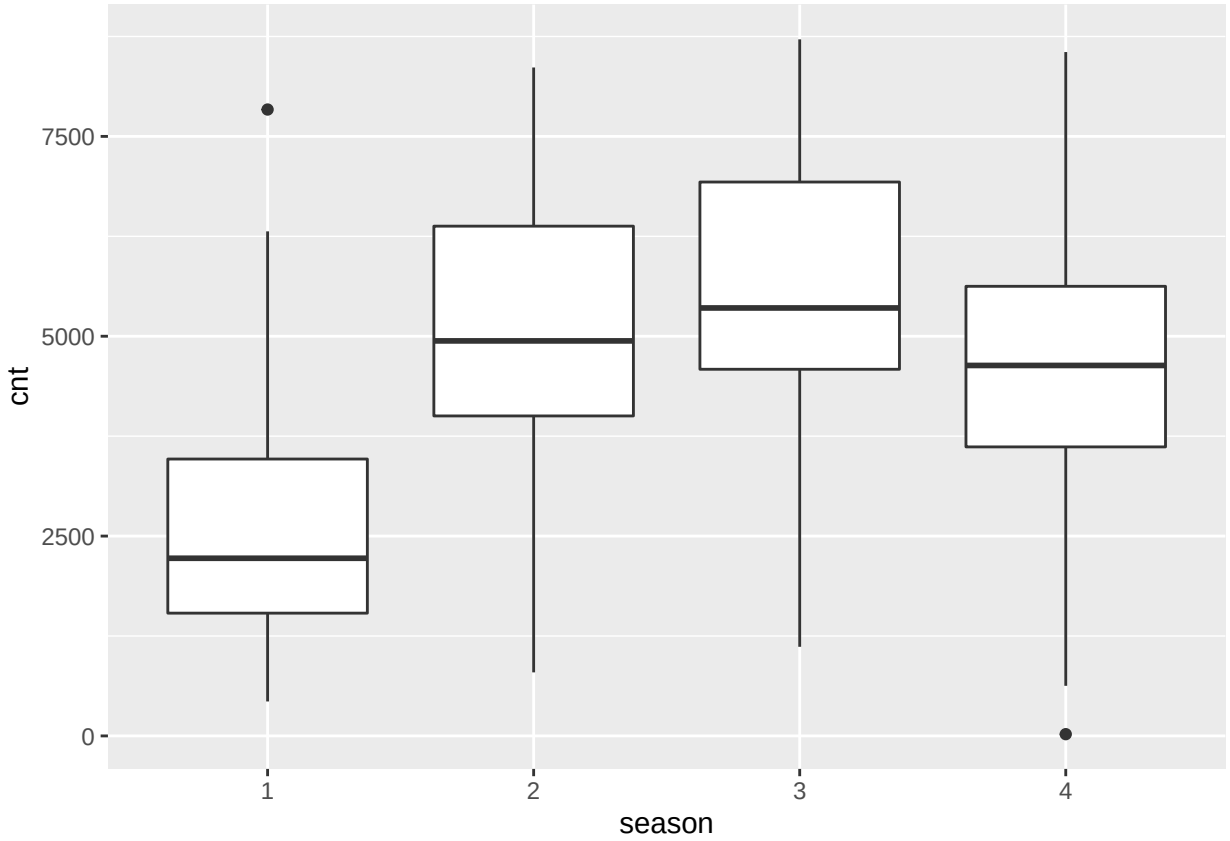
```
## Scale for 'colour' is already present. Adding another scale for 'colour',
## which will replace the existing scale.
```



ed on BoomBikes 2018 Kaggle data https://www.kaggle.com/yasserh/bike−sharing−dataset

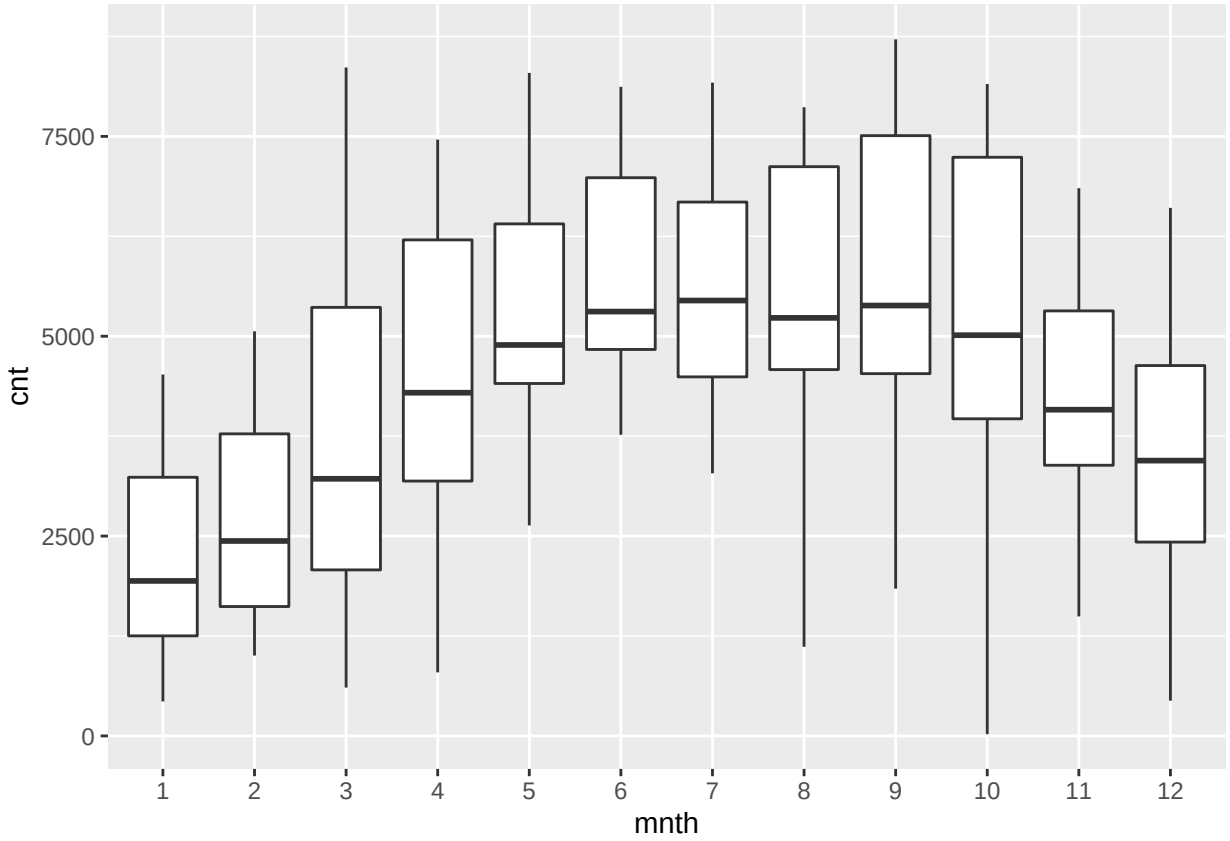
2.2.2 Season

There are total 4 seasons in the dataset and their total count is calculated here. Findings shows that season 3 has highest number of the count bikes as compared to others.



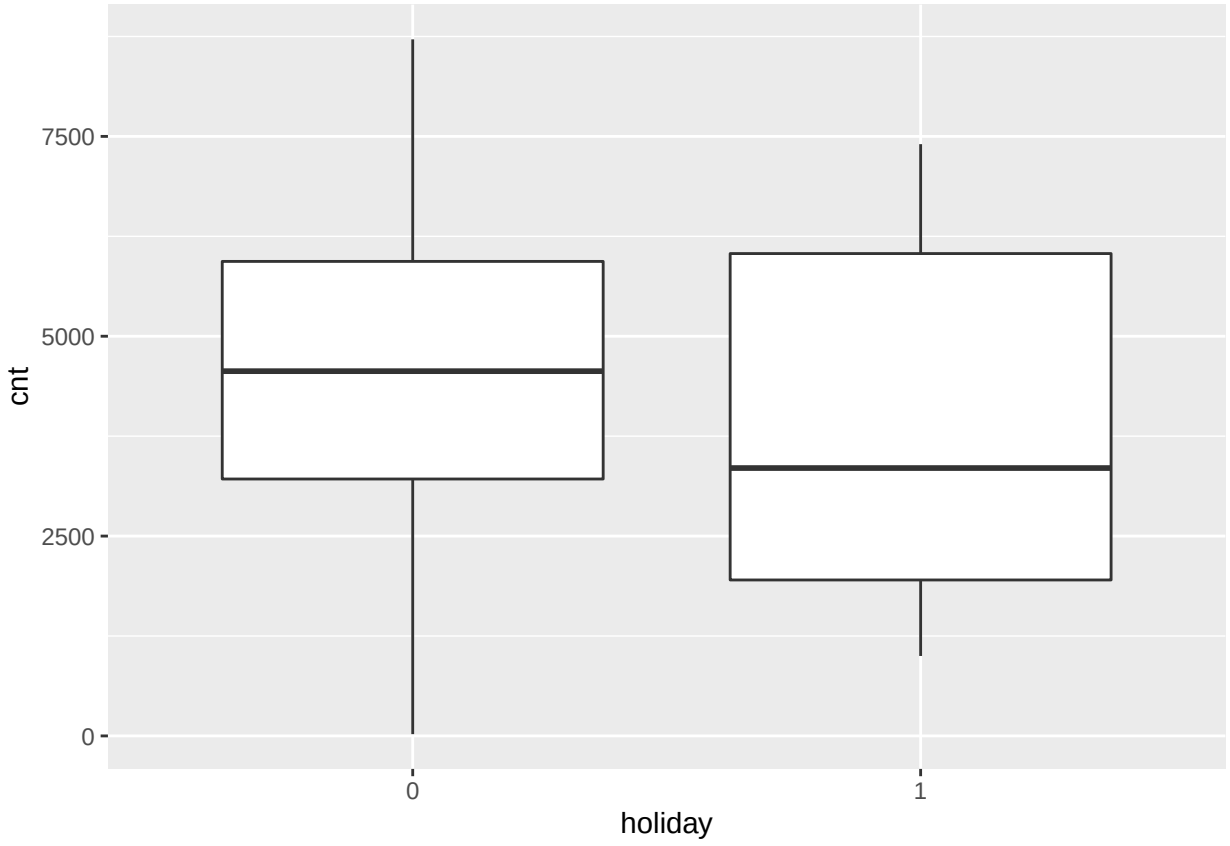
2.2.3 Month

The monthly count for the bikes is also computed and output showsh the counting for each month.



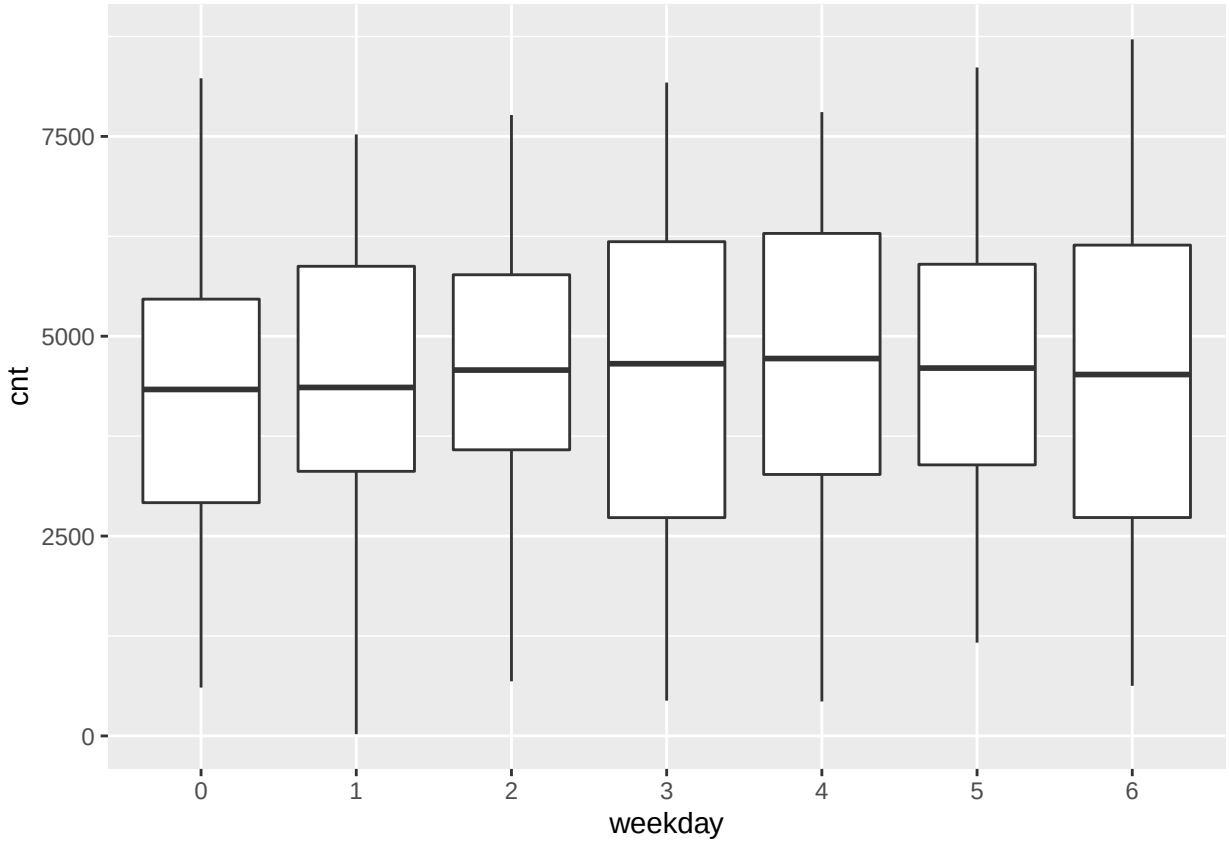
2.2.4 Holiday

The holiday counting is also calculated and shows that on holidays the count for each days.



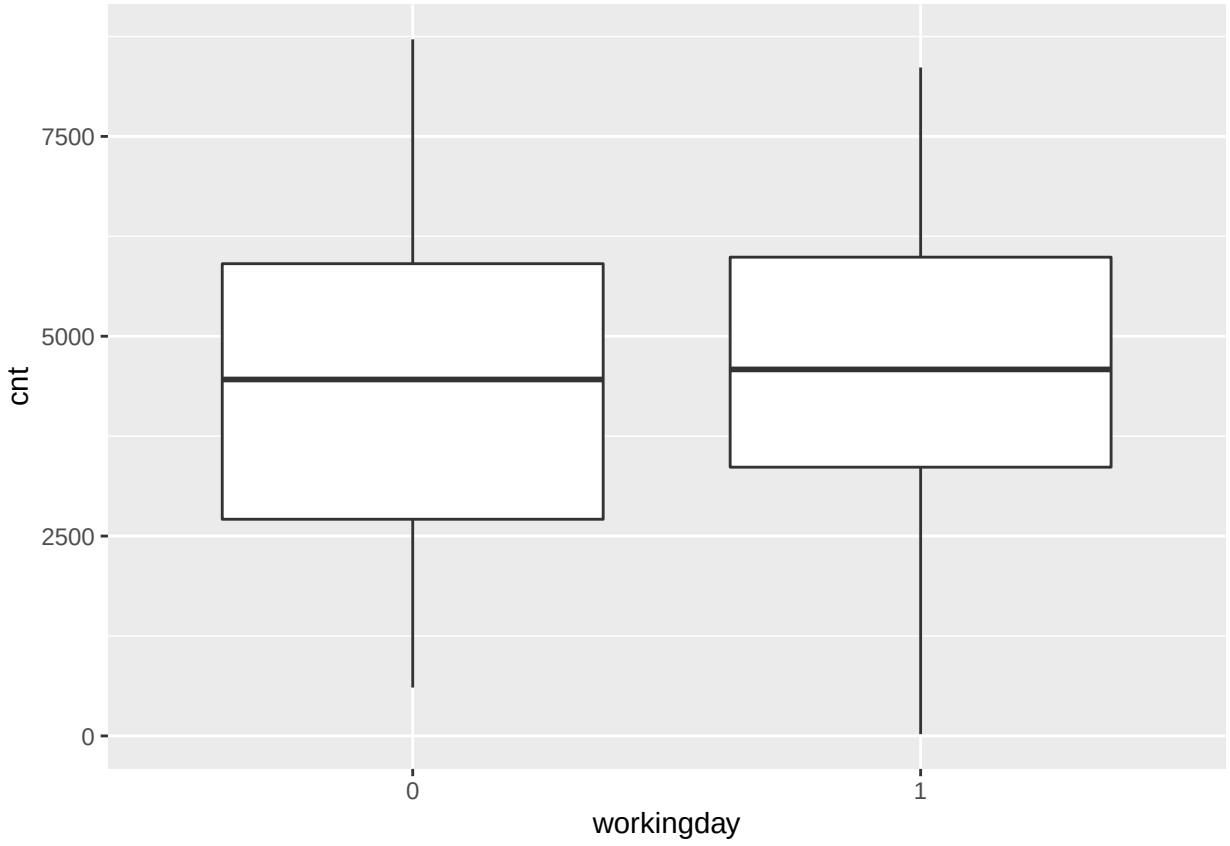
2.2.5 Week day

The counting of the bikes for each weekday is presented here using the boxplot. The output is displaying the number for each day.



2.2.6 Workingday

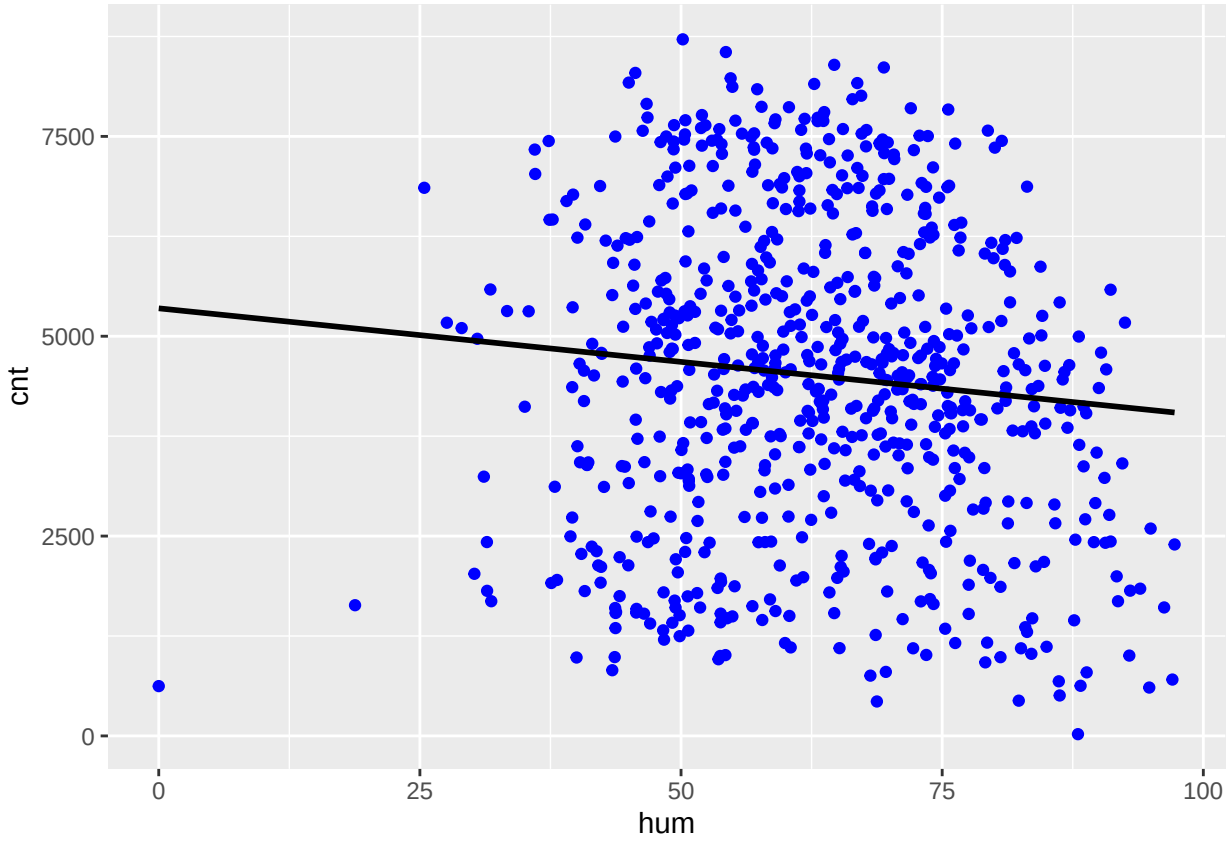
The count numbers on working is high as compared to the no working days.



2.2.7 Hum (Humidity)

The regression line between humidity and count is drawled using the ggplot. The output shows their relationship. They are uncorrelated.

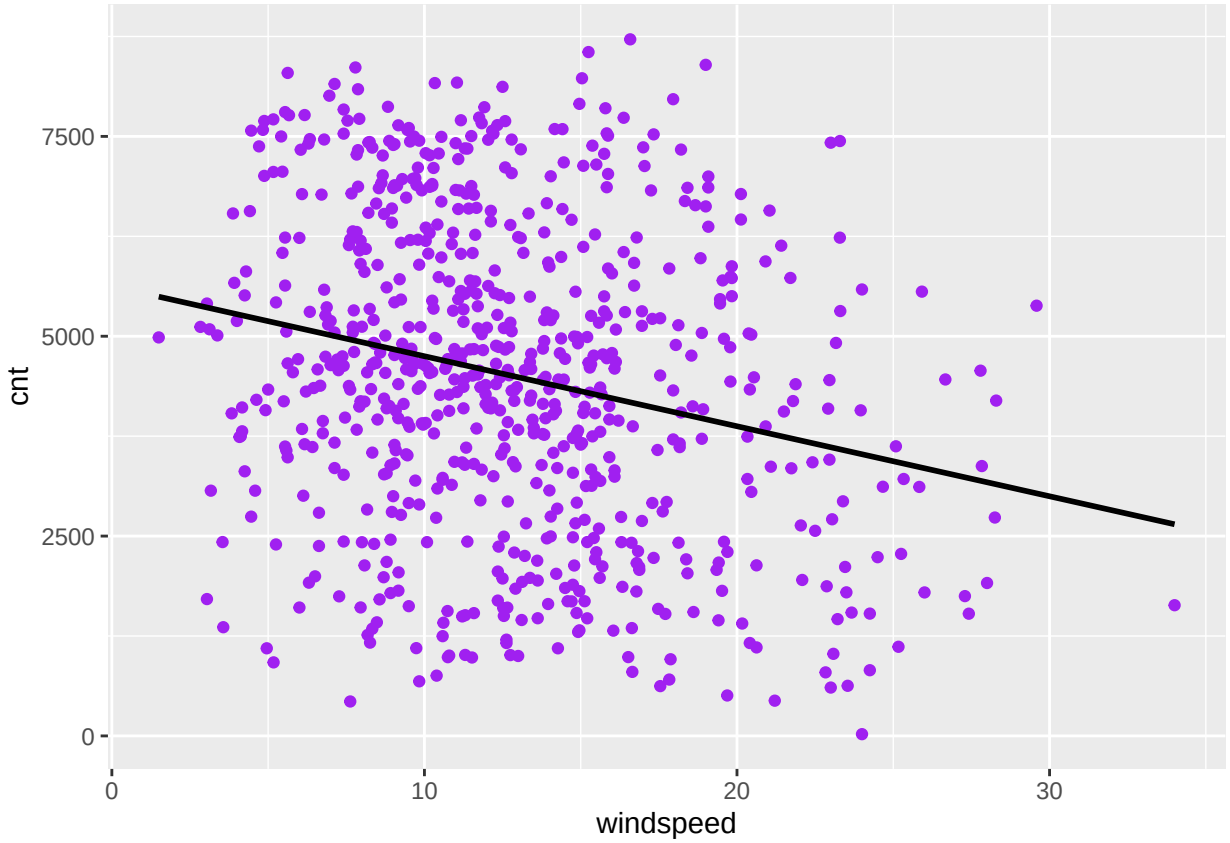
```
## `geom_smooth()` using formula 'y ~ x'
```



2.2.8 Wrind Speed

The realtionship between windspeed and count is determined using the regresssion line that can be scene form the the below output. These are not related to each other.

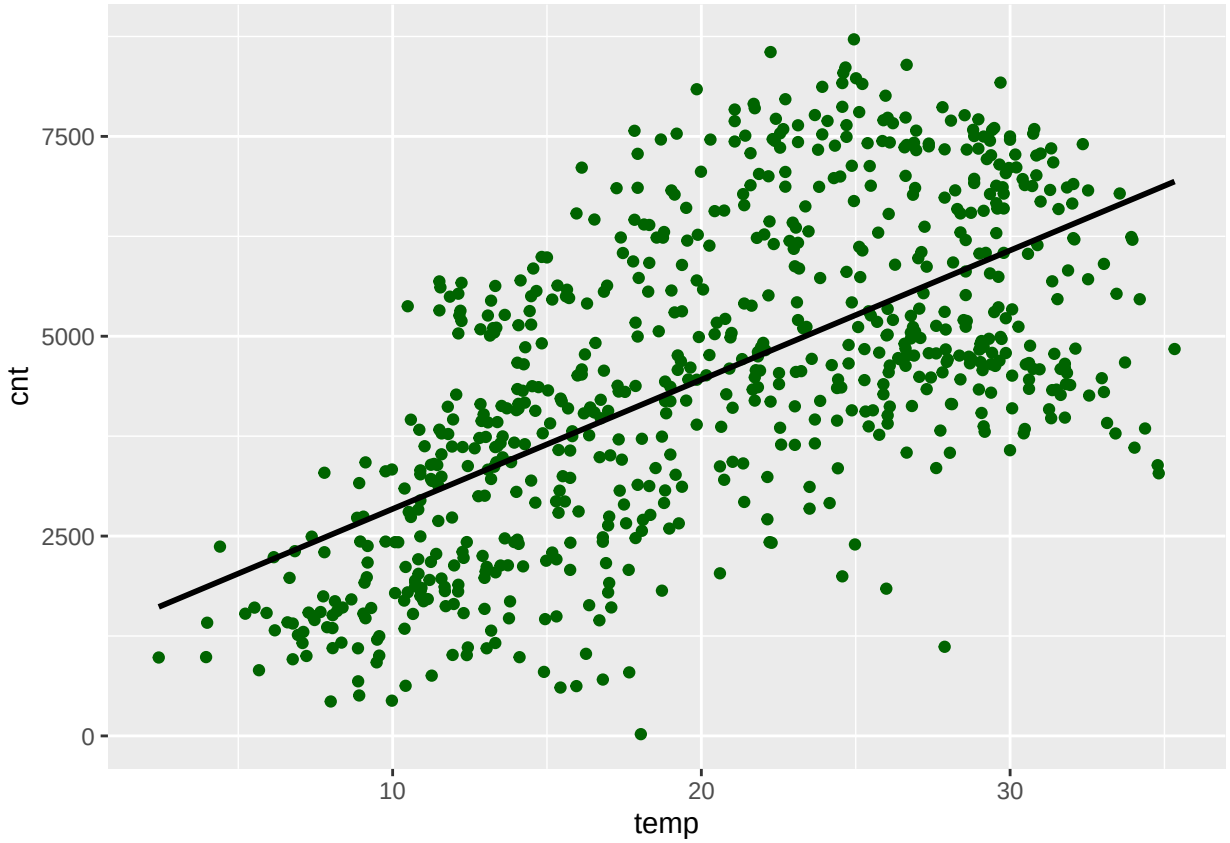
```
## `geom_smooth()` using formula 'y ~ x'
```



2.2.9 Temp (Temperature)

The regression line for thr temperature and cout is presented below and it is showing that the is linear and increasing.

```
## `geom_smooth()` using formula 'y ~ x'
```



3 Data splits for traning and tesing

The bike share dataset is divide for training and testing with the 75/25 ratio for training and testing respectively.

```
## <Analysis/Assess/Total>
## <546/184/730>
```

```
## # A tibble: 546 x 9
##    season mnth  holiday weekday workingday  temp   hum windspeed   cnt
##    <fct>  <fct> <fct>   <fct>   <fct>      <dbl> <dbl>     <dbl> <dbl>
##  1 1      1     0       6       0          14.1   80.6     10.7    985
##  2 1      1     0       0       0          14.9   69.6     16.7    801
##  3 1      1     0       1       1           8.05  43.7     16.6   1349
##  4 1      1     0       5       1           8.06  49.9     11.3   1510
##  5 1      1     0       6       0           6.76  53.6     17.9    959
##  6 1      1     0       0       0           5.67  43.4     24.3    822
##  7 1      1     0       2       1           6.93  68.6      8.18  1263
##  8 1      1     0       3       1           7.08  60.0     20.4   1162
##  9 1      1     0       6       0           9.57  49.9     10.6   1248
## 10 1      1     0       0       0           9.50  48.4     12.6   1204
## # ... with 536 more rows
```

```
## # A tibble: 184 x 9
```

```
##    season mnth  holiday weekday workingday  temp   hum windspeed   cnt
##    <fct>  <fct> <fct>   <fct>   <fct>      <dbl> <dbl>     <dbl> <dbl>
##  1 1      1     0       2       1           8.2   59.0     10.7   1562
##  2 1      1     0       3       1           9.31  43.7     12.5   1600
##  3 1      1     0       4       1           8.38  51.8      6.00  1606
##  4 1      1     0       1       1           6.18  48.3     15.0   1321
##  5 1      1     0       4       1           6.76  47.0     20.2   1406
##  6 1      1     0       5       1           6.60  53.8      8.48  1421
##  7 1      1     0       2       1           8.88  86.2      9.83   683
##  8 1      1     0       4       1          10.7   53.8     13.1   1927
##  9 1      1     0       1       1           3.99  49.2     10.6   1416
## 10 1      1     0       1       1           7.41  60.4     12.5   1501
## # ... with 174 more rows
```

3.1 Feature Engineering

It is a process from which we pick only the influential variables from the dataset to build a better predictive modeling. Due to this issue, this step is performed here. The target varibale in the dataset is 'cnt' and all other varibales are features in the dataset.

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          3
##
## Training data contained 546 data points and no missing data.
##
## Operations:
##
## Correlation filter removed no terms [trained]
```

```
## # A tibble: 4 x 4
##   variable  type    role      source
##   <chr>     <chr>   <chr>     <chr>
## 1 temp      numeric predictor original
## 2 hum       numeric predictor original
## 3 windspeed numeric predictor original
## 4 cnt       numeric outcome   original
```

```
## # A tibble: 184 x 4
##     temp   hum windspeed   cnt
##    <dbl> <dbl>     <dbl> <dbl>
##  1  8.2   59.0     10.7   1562
##  2  9.31  43.7     12.5   1600
##  3  8.38  51.8      6.00  1606
##  4  6.18  48.3     15.0   1321
##  5  6.76  47.0     20.2   1406
##  6  6.60  53.8      8.48  1421
##  7  8.88  86.2      9.83   683
```

```
##  8 10.7   53.8     13.1   1927
##  9  3.99  49.2     10.6   1416
## 10  7.41  60.4     12.5   1501
## # ... with 174 more rows
```

```
## # A tibble: 546 x 4
##     temp   hum windspeed   cnt
##    <dbl> <dbl>     <dbl> <dbl>
##  1 14.1   80.6     10.7    985
##  2 14.9   69.6     16.7    801
##  3  8.05  43.7     16.6   1349
##  4  8.06  49.9     11.3   1510
##  5  6.76  53.6     17.9    959
##  6  5.67  43.4     24.3    822
##  7  6.93  68.6      8.18  1263
##  8  7.08  60.0     20.4   1162
##  9  9.57  49.9     10.6   1248
## 10  9.50  48.4     12.6   1204
## # ... with 536 more rows
```

4 Results

In this section, linear regression model is applied on the dataset using the influential variables from the dataset. The performance of the model is determined on the testing dataset by using the training dataset.There are total three models are builds using the temp, hum and windspread. We assess the performance of the environmental features only on the bike counting and finds the best model for the problem. The p[performance is assessed using the 3 types of the regression models that are given below:

4.1 Model Training

Here we trained the above three regression models on the dataset and their results for training are shown below as an evidence.

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.398         0.397 1508.      359. 6.73e-62     1 -4770. 9545. 9558.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.415         0.412 1488.      192. 7.60e-64     2 -4762. 9532. 9549.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.460         0.457 1430.      154. 3.49e-72     3 -4740. 9490. 9511.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

The same models are implemented here as well using the tidymodels and findings are shown below:

```
##
## Call:
## lm(formula = cnt ~ temp, data = bi_train)
##
## Coefficients:
## (Intercept)         temp
##      1178.2        163.2
```

```
##
## Call:
## lm(formula = cnt ~ temp, data = bi_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4612.6 -1097.9  -125.3  1013.6  3747.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1178.249    187.406   6.287 6.64e-10 ***
## temp         163.175      8.608  18.956  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1508 on 544 degrees of freedom
## Multiple R-squared:  0.3978, Adjusted R-squared:  0.3967
## F-statistic: 359.3 on 1 and 544 DF,  p-value: < 2.2e-16
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1178.    187.        6.29 6.64e-10
## 2 temp            163.      8.61     19.0  6.73e-62
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   1917.     263.        7.28 1.19e-12
## 2 temp           157.       8.62     18.3  1.75e-58
## 3 windspeed      -48.9     12.4      -3.94 9.20e- 5
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   3979.     396.       10.0  6.73e-22
## 2 temp           162.       8.31     19.5  2.36e-64
## 3 windspeed      -70.4     12.4      -5.70 1.96e- 8
## 4 hum            -29.9      4.41     -6.77 3.33e-11
```

4.1.1 Model Testing

Here, the linear regression models are trained on the 75% of the dataset and now their performance is assessed using the testing dataset. The testing scores for 3 linear regression models are shown below.

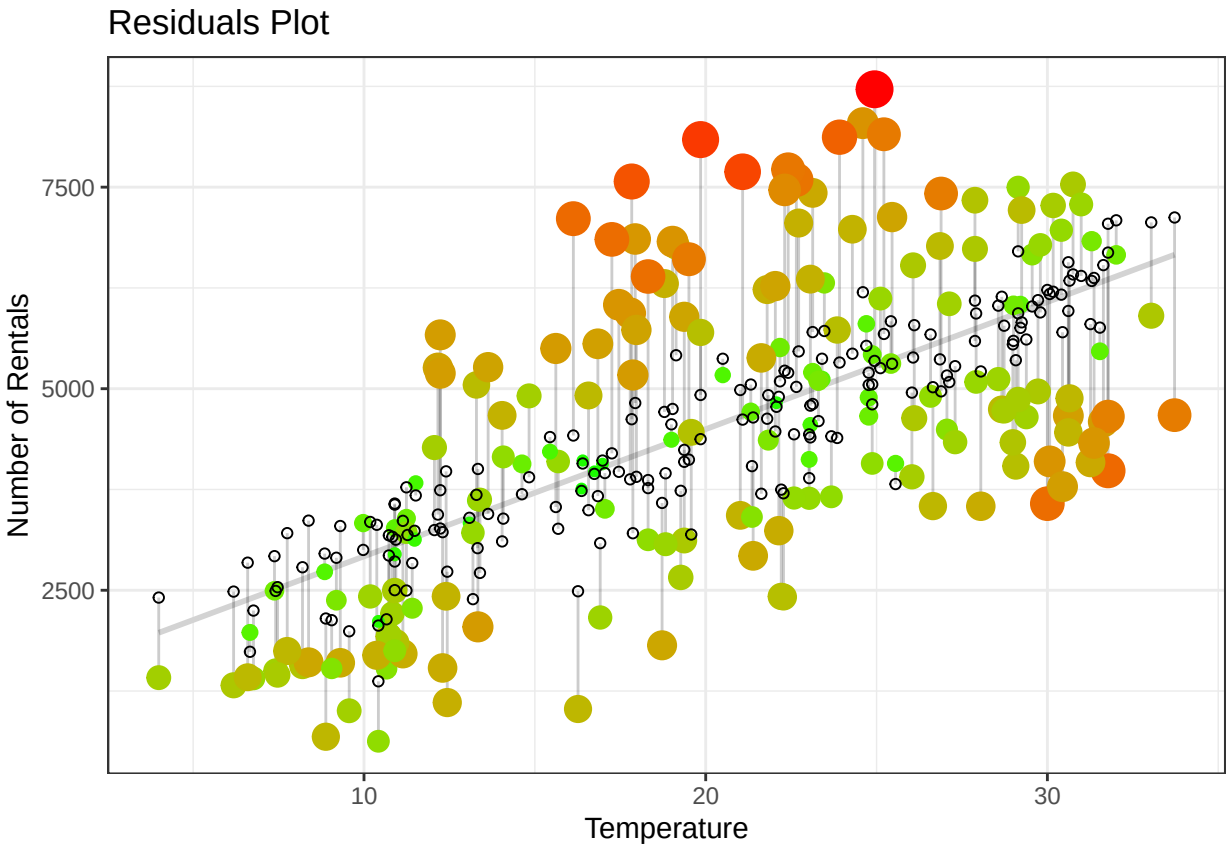
```
## # A tibble: 6 x 6
##    temp   hum windspeed   cnt .pred residuals
##   <dbl> <dbl>     <dbl> <dbl> <dbl>     <dbl>
## 1  8.2   59.0      10.7   1562 2682.    -1120.
## 2  9.31  43.7      12.5   1600 2769.    -1169.
## 3  8.38  51.8       6.00  1606 2942.    -1336.
## 4  6.18  48.3      15.0   1321 2158.     -837.
## 5  6.76  47.0      20.2   1406 1995.     -589.
## 6  6.60  53.8       8.48  1421 2540.    -1119.
```

4.1.2 Model Evaluation using Residuals metric

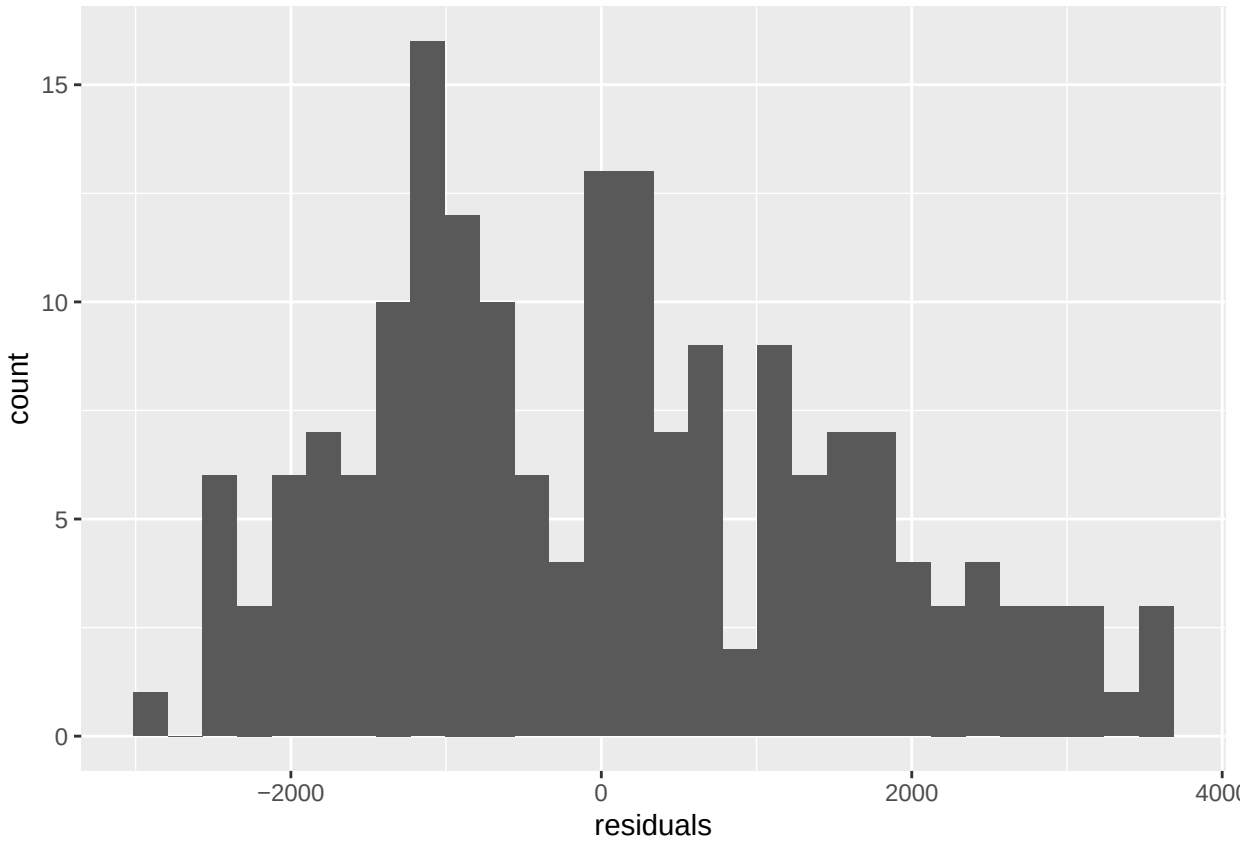
The residual is indeed a metric of just how far vertically apart a point would be from the linear interpolation. It indicates variance that the model does not explain. It is the difference between a forecaster and actual values. These plots are being utilized to examine the residuals for underlying patterns that may indicate that the linear regression model has an issue.

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

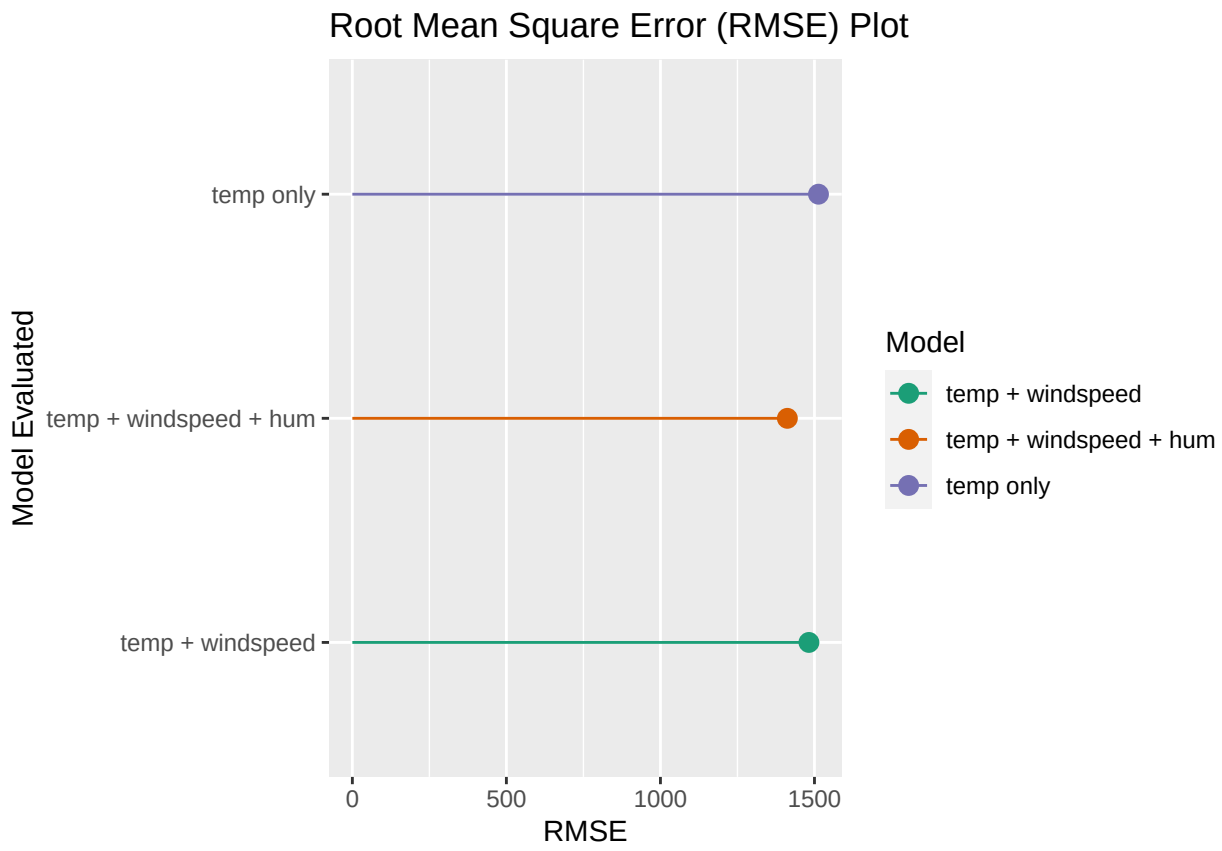


4.1.3 Model Evaluation using RMSE

The Root Mean Square Error (RMSE) is a measure of the residuals' standard deviation (prediction errors). The term "residuals" refers to the distance between the data points on the regression line and the term "RMSE" refers to the expansion of such residuals. On the other hands, it indicates the degree to which the data is packed it around best fit line. The RMSE error is often used to validate experimental findings in meteorology, prediction, and linear regression. The lower value of the RMSE on the test dataset shows that model is fitting very well on the dataset.

The model 3 which is consist of the temp+windspread+hum has the lowest RMSE score as compared to other 2 models on the test. The comparison of the models is showing below:

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       1513.
```



4.1.4 Model Evaluation using R-squared

R-squared is the amount of variation (percentage) in the dependent variable that the independent variable can explain. As a result, as a general rule, evaluate the strength of a link in terms of its R-squared value.

(a) R-squared value 0.3 is regarded to be a None or Very Weak impact size.
(b) To-squared value of 0.5 r 0.7 is typically regarded as a Moderate effect size.
(c) R-squared value greater than 0.7 is regarded to have a large impact size.

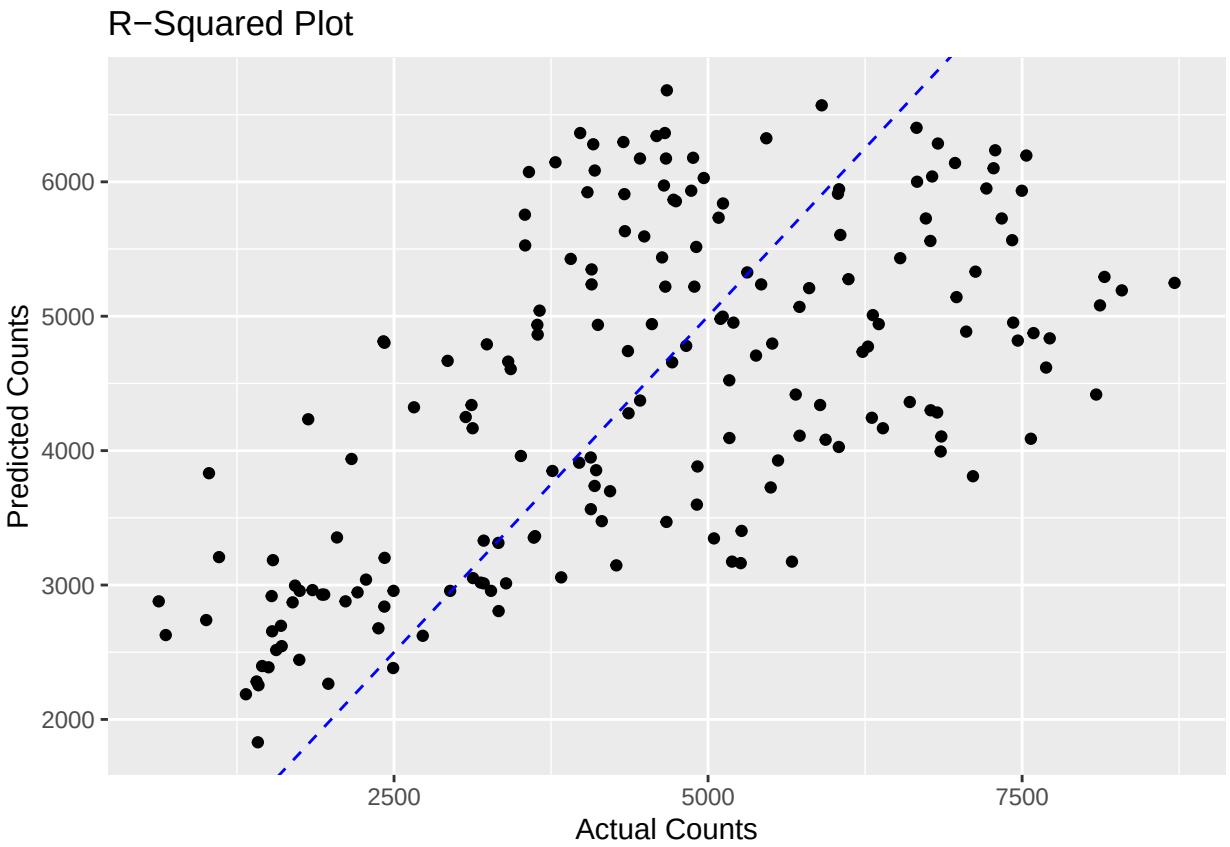
To illustrate the projected vs. real counts (in this example), it is a good practise to plot the predicted vs. actual counts. The performance of the model

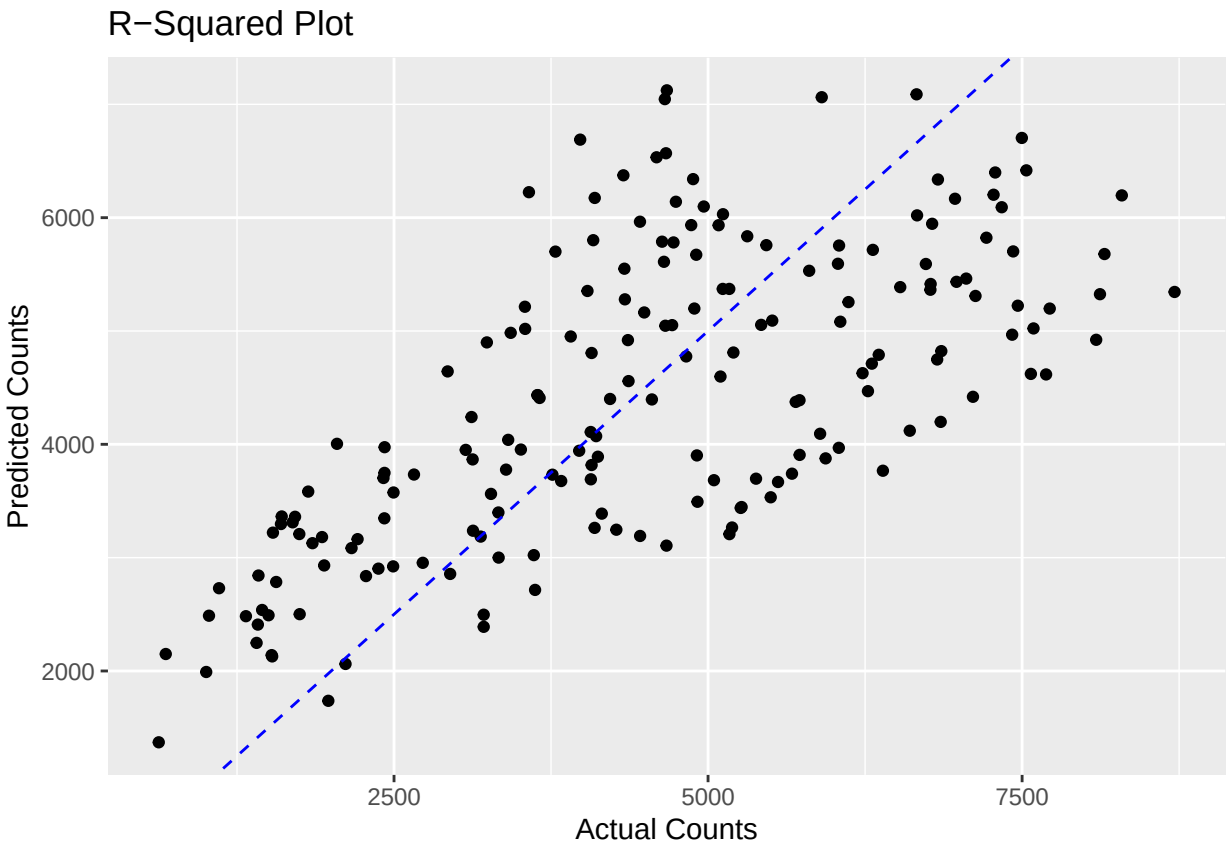
– patterns that are not linear – areas in which the model performs badly

The more variation the regression model accounts for, the closest the data points lie to the linear regression line. In theory, if a model can account for 100% of the variance, then estimated coefficients will always match the observed data, and so all measured values will be on the fitted on line of regression.

The R squared plots and estimates value for the model are shown below and findings shows that the estimates values are very low as compared to the others.

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rsq     standard       0.380
```


R−Squared Plot
Predicted Counts
6000
5000
4000
3000
2000
2500
5000
7500
Actual Counts



5 Conlusion

In this case study the bike share dataset in analyzed using the various steps to find the impact of the weather effects variables from the dataset to predict the 'cnt' variables. Various types of the steps are considered here like data cleaning, data exploration and data visualization to find the hidden insights from the dataset. The various types of the box plots are plotted using the target variables. The data visualization is very helpful to find the hidden facts from the dataset and determined the bike rental counting for each day, month and year. The findings shows that the feature engineering is also performed on the dataset to find the hidden insights from it. Three linear regression models are builds to predict 'cnt' variables using the 75/25 ratio for training and testing the models on ht dataset. The evaluation of the models is assessed using the residual error, RMSE and R-square. The plots are also plotted. The findings shows that the the tmp+windspread+hum variables achieved the lowest RMSE score as compared to the other two. So, these three variables are very helpful to build a better predictive model using the weather types variables as compared to other two regression models.

5.1 Limitations

In this analysis, we only utilized the weather related variables to predict the rental count for the bikes from very huge amount of the variables. Furthermore, the linear regression modeling is considered here to build the predictive model that is also one of the limited approach.

5.2 Future Work

For future work various types of the other features can also accommodate to forecast the rental bike count for this dataset with more diverse way for better results, These variables can be location, customer - occupation, income and area. Moreover, the various types of the other machine learning models can be used here like Random Forest,Gradient boosted tree, Neural network and Support vector machines.