FA 541 – Final Project

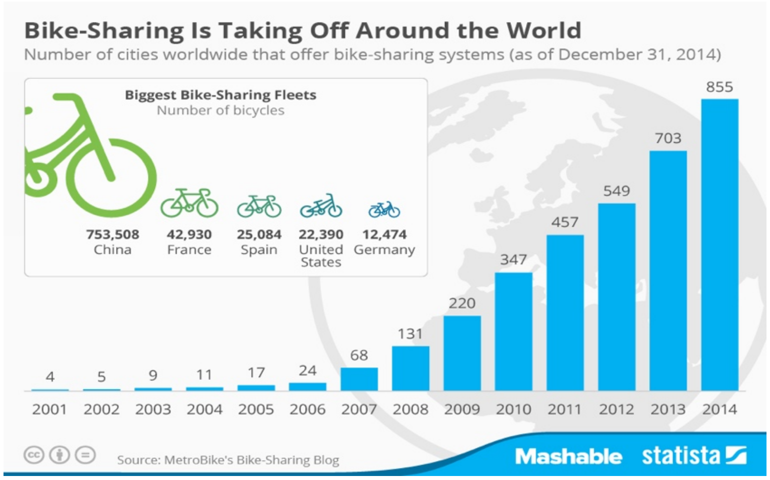
Prediction of bike rental demand based on various factors

Jahnavi Ravi

Table of Contents

Introduction

Bike sharing systems offer an automated way of renting bicycles through kiosks located throughout a city. Users can rent a bike from one location and return it to a different place as needed. There are currently over 500 bike-sharing programs around the world. The data generated by these systems, such as duration of travel, departure and arrival locations, and time elapsed, can be used by researchers to study mobility in a city. Bike sharing systems act as a sensor network, providing valuable insights into patterns of urban mobility.



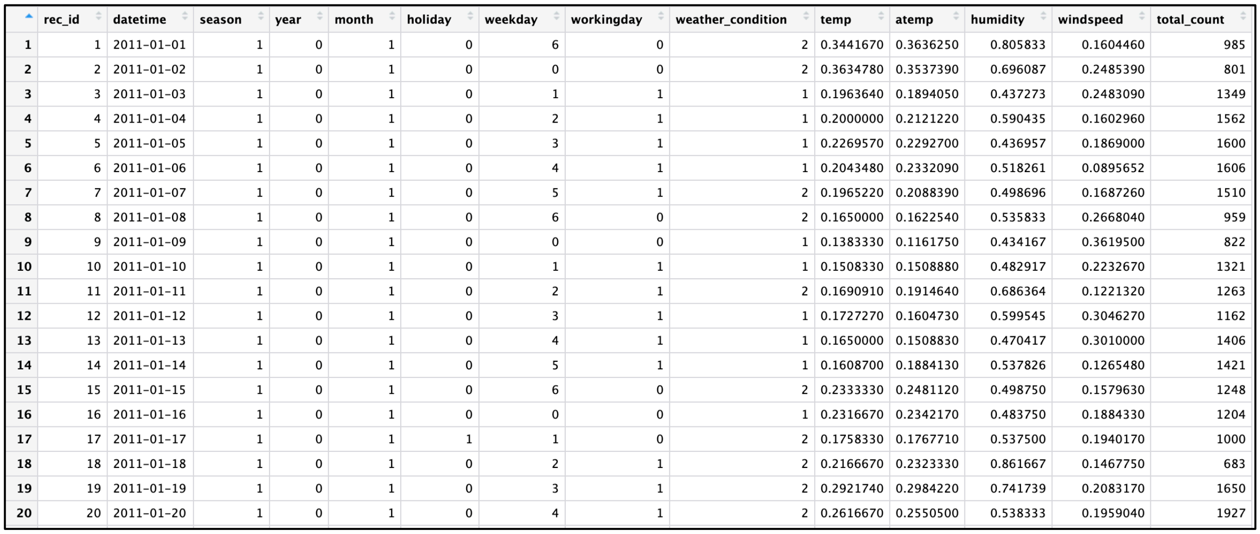
Predicting bike rental patterns can have significant benefits for various stakeholders involved in bike-sharing programs and rental companies. One of the key advantages is demand forecasting, where the ability to predict future bike rental demand allows for effective resource planning and management. By accurately estimating the demand, bike-sharing programs can adjust the number of available bikes, optimize maintenance schedules, and ensure they have enough staff to meet customer needs. This helps in avoiding situations where there is either a shortage or surplus of bikes, improving customer satisfaction and operational efficiency.

In addition to demand forecasting, predicting bike rental patterns also aids in resource allocation. By understanding the expected rental patterns, bike-sharing programs can allocate their resources, such as bikes and docking stations, more efficiently. For example, if the prediction indicates high demand during specific times of the day or in certain areas, they can increase the number of available bikes in those locations or redistribute bikes from low-demand areas to high-demand areas. This ensures that bikes are available when and where they are needed the most, optimizing the utilization of resources and enhancing the overall user experience. This report provides an analysis of a Bike Sharing System dataset using R programming.

Exploratory Data Analysis

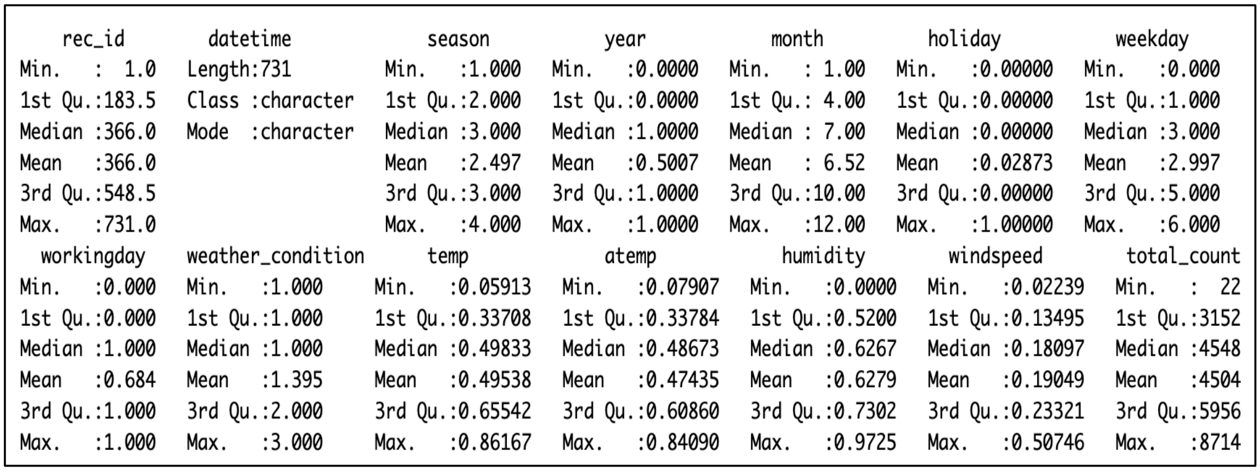
The Data Set

The data set contains daily bike rental data spanning two years from 2011 to 2012, with a total of 731 data points and 16 columns or indicators. The data set was analyzed to understand the various factors affecting bike rental patterns like weather, climate, temperature and humidity and to identify the patterns in the data.



	rec_id	datetime	season	year	month	holiday	weekday	workingday	weather_condition	temp	atemp	humidity	windspeed	total_count
1	1	2011-01-01	1	0	1	0	6	0	2	0.3441670	0.3636250	0.805833	0.1604460	985
2	2	2011-01-02	1	0	1	0	0	0	2	0.3634780	0.3537390	0.696087	0.2485390	801
3	3	2011-01-03	1	0	1	0	1	1	1	0.1963640	0.1894050	0.437273	0.2483090	1349
4	4	2011-01-04	1	0	1	0	2	1	1	0.2000000	0.2121220	0.590435	0.1602960	1562
5	5	2011-01-05	1	0	1	0	3	1	1	0.2269570	0.2292700	0.436957	0.1869000	1600
6	6	2011-01-06	1	0	1	0	4	1	1	0.2043480	0.2332090	0.518261	0.0895652	1606
7	7	2011-01-07	1	0	1	0	5	1	2	0.1965220	0.2088390	0.498696	0.1687260	1510
8	8	2011-01-08	1	0	1	0	6	0	2	0.1650000	0.1622540	0.535833	0.2668040	959
9	9	2011-01-09	1	0	1	0	0	0	1	0.1383330	0.1161750	0.434167	0.3619500	822
10	10	2011-01-10	1	0	1	0	1	1	1	0.1508330	0.1508880	0.482917	0.2232670	1321
11	11	2011-01-11	1	0	1	0	2	1	2	0.1690910	0.1914640	0.686364	0.1221320	1263
12	12	2011-01-12	1	0	1	0	3	1	1	0.1727270	0.1604730	0.599545	0.3046270	1162
13	13	2011-01-13	1	0	1	0	4	1	1	0.1650000	0.1508830	0.470417	0.3010000	1406
14	14	2011-01-14	1	0	1	0	5	1	1	0.1608700	0.1884130	0.537826	0.1265480	1421
15	15	2011-01-15	1	0	1	0	6	0	2	0.2333330	0.2481120	0.498750	0.1579630	1248
16	16	2011-01-16	1	0	1	0	0	0	1	0.2316670	0.2342170	0.483750	0.1884330	1204
17	17	2011-01-17	1	0	1	1	1	0	2	0.1758330	0.1767710	0.537500	0.1940170	1000
18	18	2011-01-18	1	0	1	0	2	1	2	0.2166670	0.2323330	0.861667	0.1467750	683
19	19	2011-01-19	1	0	1	0	3	1	2	0.2921740	0.2984220	0.741739	0.2083170	1650
20	20	2011-01-20	1	0	1	0	4	1	2	0.2616670	0.2550500	0.538333	0.1959040	1927

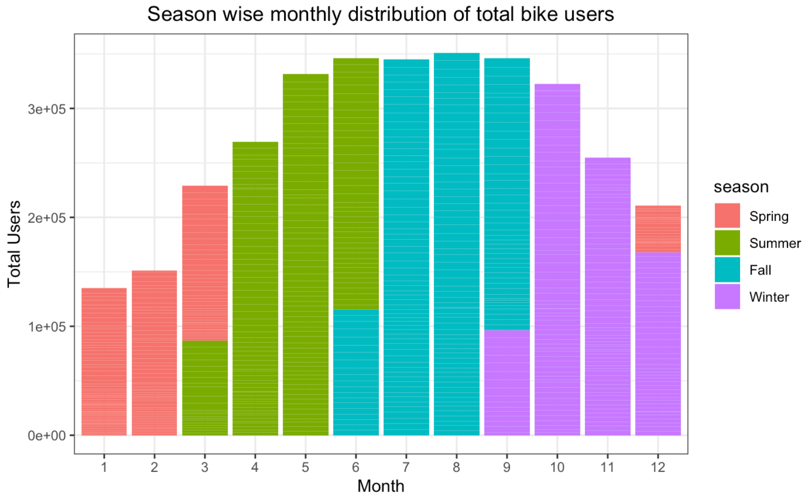
Summary Statistics



```
     rec_id         datetime              season           year            month          holiday           weekday     
 Min.   :  1.0   Length:731         Min.   :1.000   Min.   :0.0000   Min.   : 1.00   Min.   :0.00000   Min.   :0.000  
 1st Qu.:183.5   Class :character   1st Qu.:2.000   1st Qu.:0.0000   1st Qu.: 4.00   1st Qu.:0.00000   1st Qu.:1.000  
 Median :366.0   Mode  :character   Median :3.000   Median :1.0000   Median : 7.00   Median :0.00000   Median :3.000  
 Mean   :366.0                      Mean   :2.497   Mean   :0.5007   Mean   : 6.52   Mean   :0.02873   Mean   :2.997  
 3rd Qu.:548.5                      3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:10.00   3rd Qu.:0.00000   3rd Qu.:5.000  
 Max.   :731.0                      Max.   :4.000   Max.   :1.0000   Max.   :12.00   Max.   :1.00000   Max.   :6.000  
   workingday    weather_condition      temp             atemp            humidity        windspeed        total_count  
 Min.   :0.000   Min.   :1.000     Min.   :0.05913   Min.   :0.07907   Min.   :0.0000   Min.   :0.02239   Min.   :  22  
 1st Qu.:0.000   1st Qu.:1.000     1st Qu.:0.33708   1st Qu.:0.33784   1st Qu.:0.5200   1st Qu.:0.13495   1st Qu.:3152  
 Median :1.000   Median :1.000     Median :0.49833   Median :0.48673   Median :0.6267   Median :0.18097   Median :4548  
 Mean   :0.684   Mean   :1.395     Mean   :0.49538   Mean   :0.47435   Mean   :0.6279   Mean   :0.19049   Mean   :4504  
 3rd Qu.:1.000   3rd Qu.:2.000     3rd Qu.:0.65542   3rd Qu.:0.60860   3rd Qu.:0.7302   3rd Qu.:0.23321   3rd Qu.:5956  
 Max.   :1.000   Max.   :3.000     Max.   :0.86167   Max.   :0.84090   Max.   :0.9725   Max.   :0.50746   Max.   :8714  
```

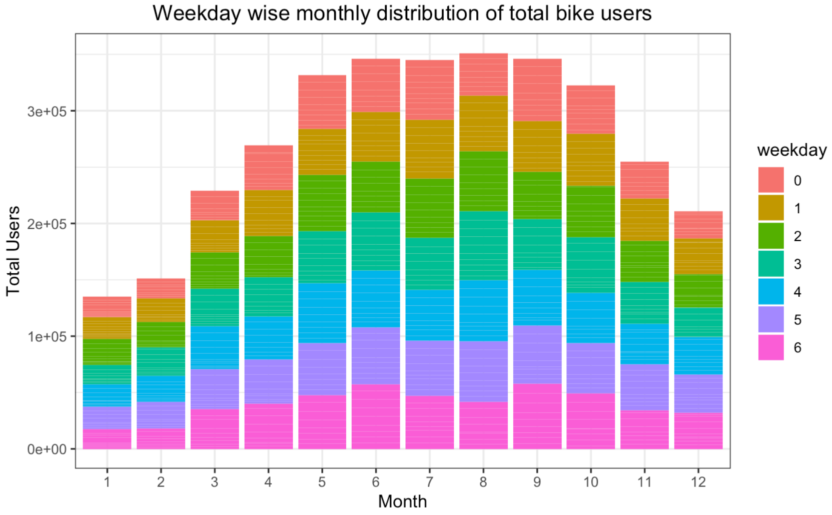
Plot for season-wise monthly distribution of counts

The data was visualized to understand the patterns in the data. The ggplot2 library was used to plot various graphs. The first plot showed the season-wise monthly distribution of counts. It was observed that the demand for bike rentals was high in summer and fall. These two seasons fall under the months of May, June, July, August, September and October as we can see from the plot below.



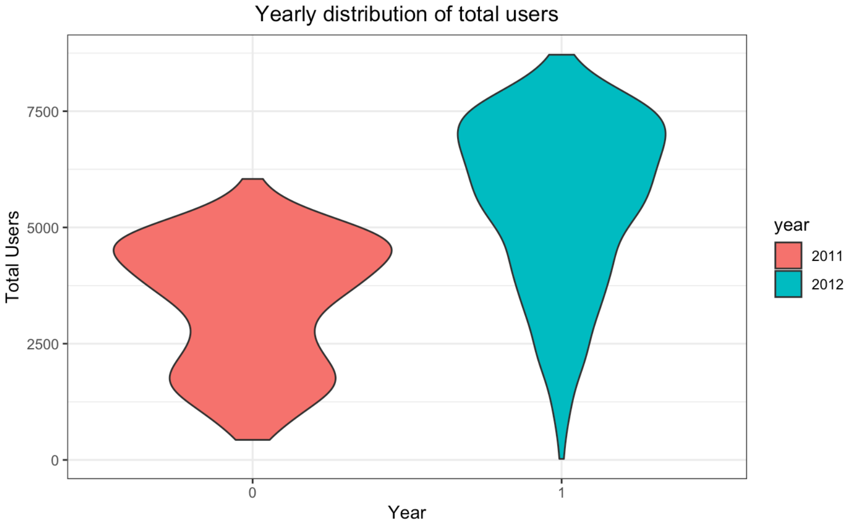
Plot for weekday-wise monthly distribution of counts

The second plot showed the weekday-wise monthly distribution of counts. It was observed that the demand for bike rentals was high on weekdays as compared to weekends during almost all 12 months.



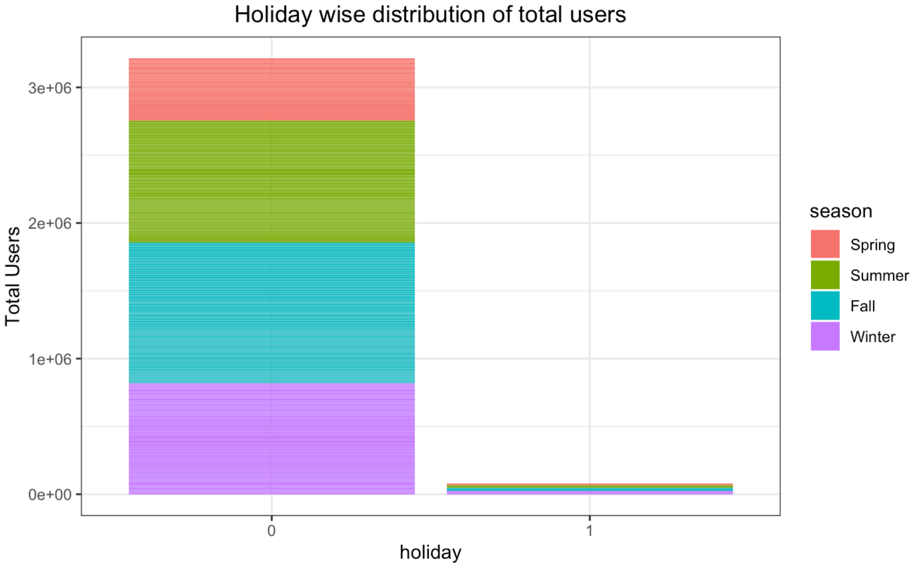
Violin plot for yearly distribution of counts

The third plot shows the yearly distribution of counts using a violin plot. It was observed that the demand for bike rentals was higher in 2012 as compared to 2011. This could be because people were becoming more aware of the bike rental system as time passed.



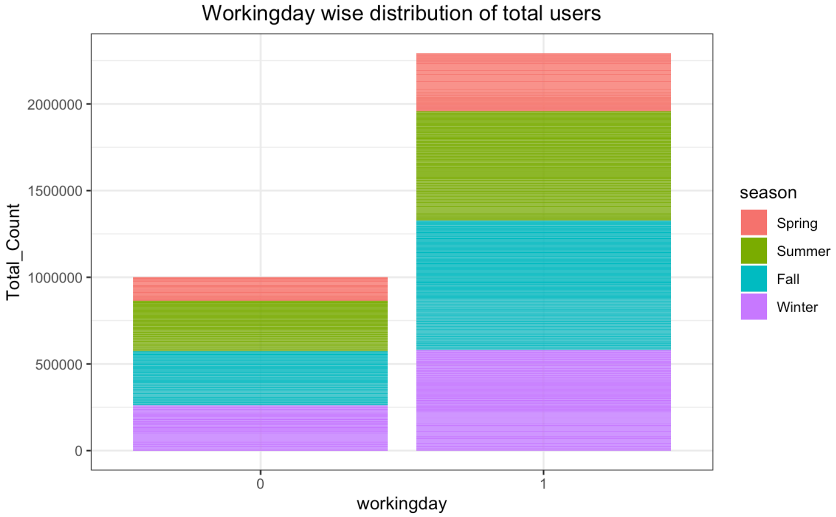
Holiday-wise distribution of total users

The fourth plot shows the holiday-wise distribution of total users. It was observed that the demand for bike rentals was high during holidays in summer and fall, followed by winter when there is no holiday, which is an interesting observation.



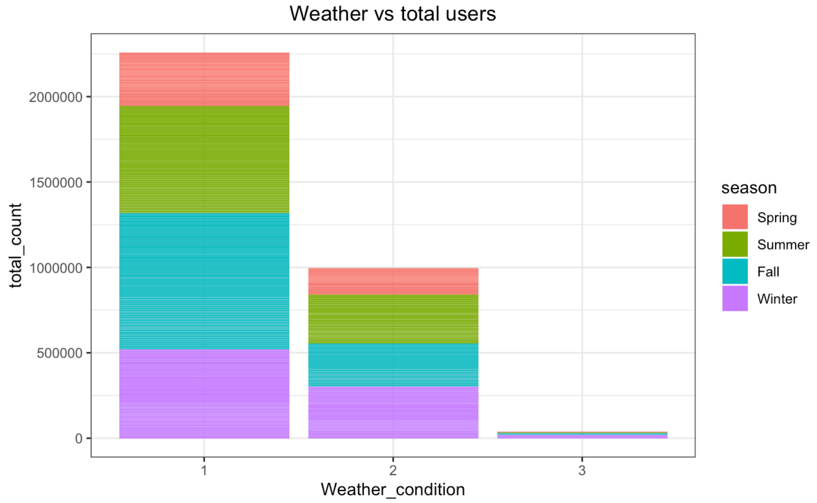
Working-day-wise distribution of counts

The fifth plot showed the working-day-wise distribution of total counts. It was observed that the demand for bike rentals was higher on working days as compared to non-working days. This could mean that people are using the bikes more for travelling to and from work or for other purposes rather than using them for recreational purposes.



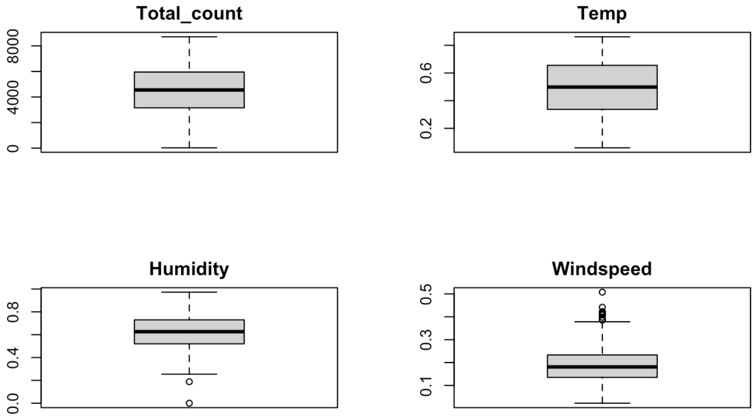
Weather-wise distribution of counts

The last plot showed the weather-wise distribution of counts. It was observed that the demand for bike rentals was high during clear weather as compared to other weather conditions, which makes complete sense in this situation. Again, the demand is the highest in the fall and summer compared to other seasons.



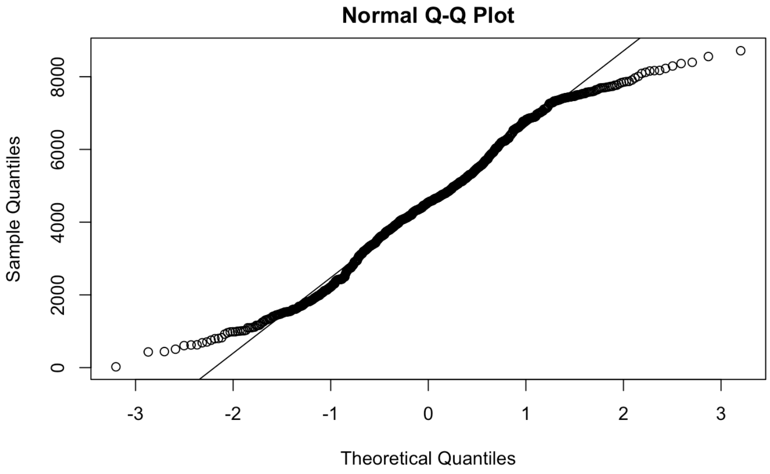
Outlier Analysis

Finally, outlier analysis was performed using box plots. From the box plot, it can be observed that no outliers are present in the total count and normalized temp variables. However, a few outliers are present in the normalized windspeed and humidity variables and these outliers are removed before further analysis.



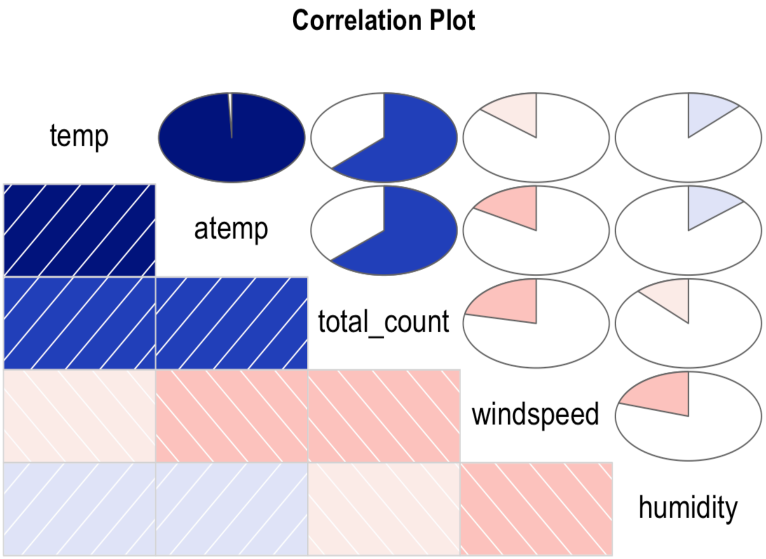
Normal Probability Plot

A normal QQ plot, also known as a quantile-quantile plot or a Q-Q plot, is a graphical tool used to assess whether a given data set follows a normal distribution. It compares the quantiles of the observed data against the quantiles of a theoretical normal distribution. The plot basically tells us about the goodness of the fit. From the Q-Q plot, we can observe that some target variable data points are deviating from the normal. However, most of the points seem to be on or round the normal.



Correlation Plot

A correlation plot tells about linear relationship between attributes and help us to build better models. From the correlation plot here, we can observe that some features are positively correlated and some are negatively correlated to each other. The temp and normalized temperature are highly positively correlated to each other, it means that both are carrying similar information. So, we are going to ignore the normalized temperature variable for further analysis.



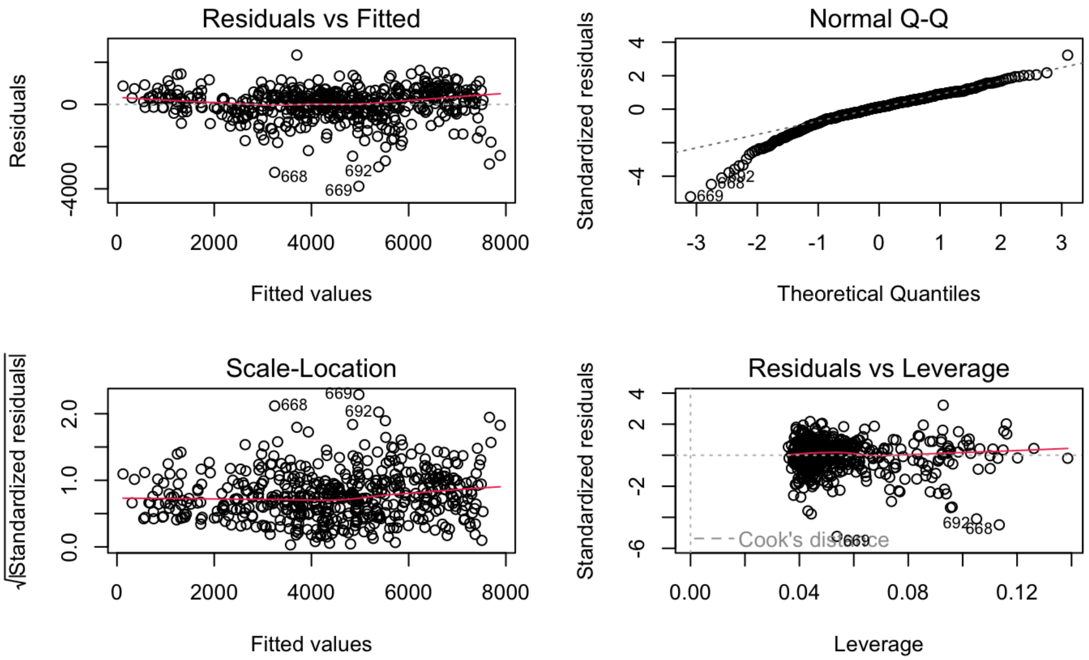
Machine Leaning Models

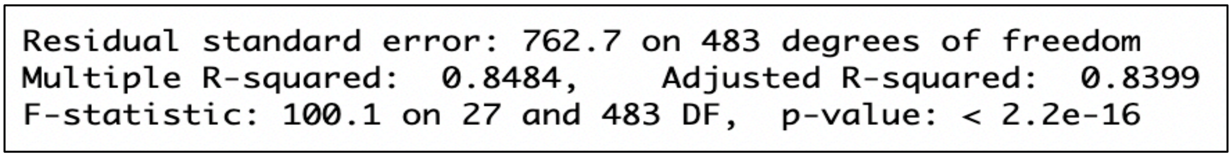
Train and Test data: The data was split into a training set containing 70% of the data from the selected variables and a test set containing the remaining 30% in order to apply machine learning models.

Then, subsets of specific columns containing the train data and test data that are considered independent variables (features) and the dependent variable (target). The selected columns include attributes such as season, year, month, holiday, weekday, working day, weather condition, temp, humidity, windspeed, and total count.

The data was then divided into categorical and numerical subsets, and encoding was performed to change the categorical variables into numerical values using dummy variables. The resulting datasets with encoded attributes are then used for further analysis or modeling tasks.

Linear Regression Model





```
Residual standard error: 762.7 on 483 degrees of freedom
Multiple R-squared:  0.8484,    Adjusted R-squared:  0.8399
F-statistic: 100.1 on 27 and 483 DF,  p-value: < 2.2e-16
```

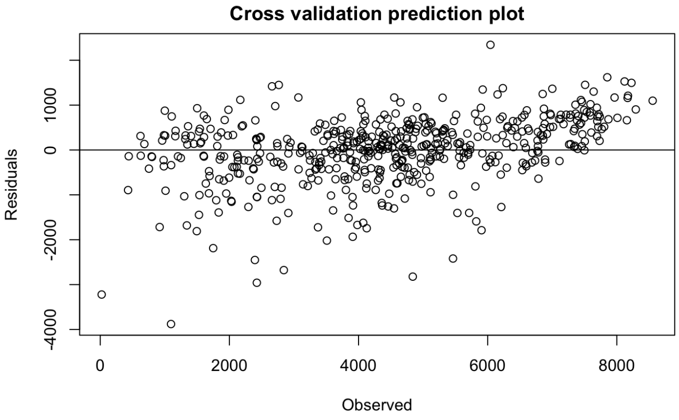
Cross-validation is a technique used to assess the performance of a model on unseen data. In this case, a 3-fold cross-validation has been performed, which means that the data has been divided into three parts, and the model has been trained and tested three times, each time using a different part as the test set and the other two parts as the training set.

The adjusted R-squared or coefficient of determination is a measure of how well the regression model fits the data, and represents the proportion of the variance in the target variable that is explained by the independent variables. A value of 0.8399 means that the model explains 83.99% of the variance in the target variable.

The p-value is a measure of the statistical significance of the results. A value less than 0.05 means that the results are unlikely to have occurred by chance, and the null hypothesis that there is no relationship between the independent and dependent variables can be rejected.

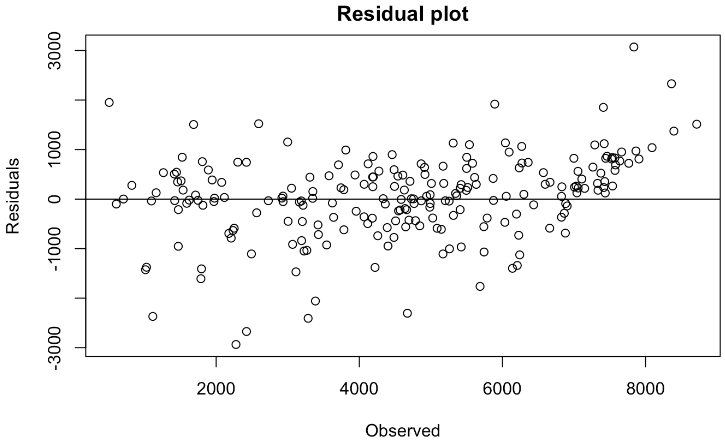
Cross Validation Prediction

A cross-validation prediction plot shows the relationship between actual and predicted values of a model across multiple cross-validation folds. In this plot, some data points may have similar variance between the predicted and actual values, resulting in a tight cluster of points on the plot, while others may have greater variance, resulting in a more scattered cluster of points. This can indicate that the model is better at predicting certain types of data points than others but is good overall.



Residual Plot

A residual plot is a graphical representation of the difference between the predicted value and the actual value in a regression analysis. In general, a good residual plot should show no clear pattern, with the points scattered randomly around the x-axis. This plot shows somewhat of a pattern but also has quite a few data points that are scattered all over. It is a good plot overall.



Decision Tree Regression

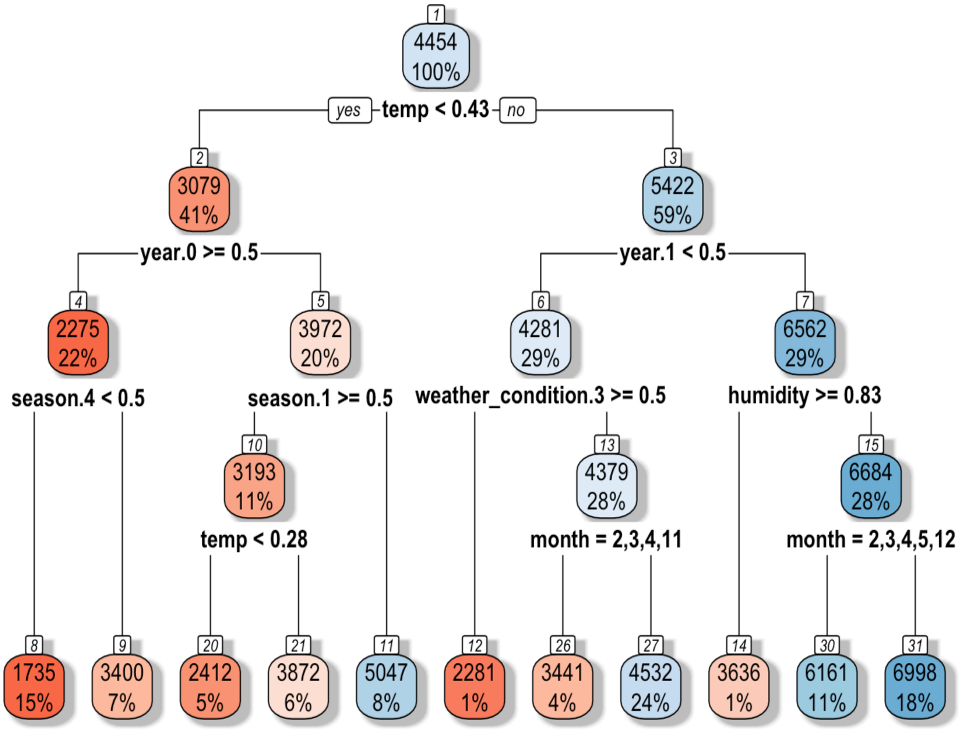
A decision tree model helps in predicting the value of a response variable based on the values of several predictor variables.

The output shows how the model splits the data based on the values of the predictor variables to create subgroups that have similar response variable values, and provides predicted values for the response variable for each subgroup.

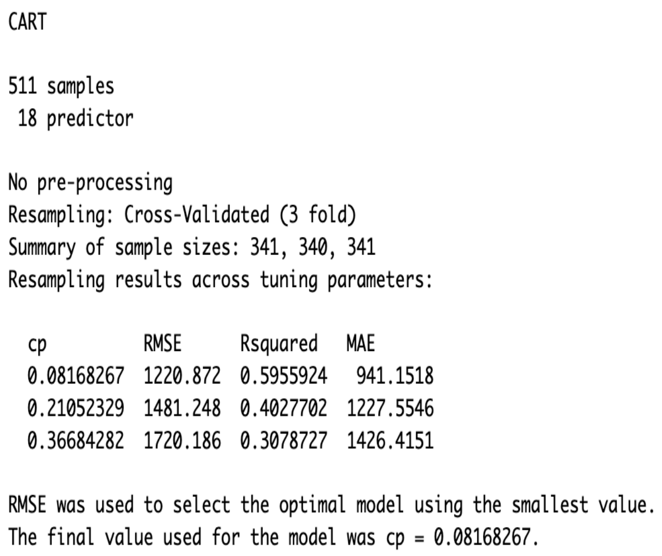
The decision tree continues to split the data based on various variables and their values, ultimately resulting in 31 terminal nodes, denoted by the asterisk (*).

```
n= 511

node), split, n, deviance, yval
      * denotes terminal node

 1) root 511 1853354000.0 4454.487
   2) temp< 0.432373 211  467984900.0 3079.090
     4) year.0>=0.5 111  120342700.0 2275.117
       8) season.4< 0.5 75   28759450.0 1735.120 *
       9) season.4>=0.5 36   24151500.0 3400.111 *
     5) year.0< 0.5 100  196255200.0 3971.500
      10) season.1>=0.5 58   63186110.0 3192.517
        20) temp< 0.2804165 27   18497810.0 2412.148 *
        21) temp>=0.2804165 31   13925200.0 3872.194 *
      11) season.1< 0.5 42   49270980.0 5047.238 *
   3) temp>=0.432373 300  705479500.0 5421.850
     6) year.1< 0.5 150  105297000.0 4281.420
      12) weather_condition.3>=0.5 7      946211.7 2281.429 *
      13) weather_condition.3< 0.5 143   74980390.0 4379.322
        26) month=2,3,4,11 20     8559473.0 3441.450 *
        27) month=5,6,7,8,9,10 123   45968350.0 4531.821 *
     7) year.1>=0.5 150  210008300.0 6562.280
      14) humidity>=0.8322915 6   17506720.0 3635.500 *
      15) humidity< 0.8322915 144  138963900.0 6684.229
        30) month=2,3,4,5,12 54   44746260.0 6161.315 *
        31) month=6,7,8,9,10 90   70592440.0 6997.978 *
```

The visual representation:



Cross Validation Prediction

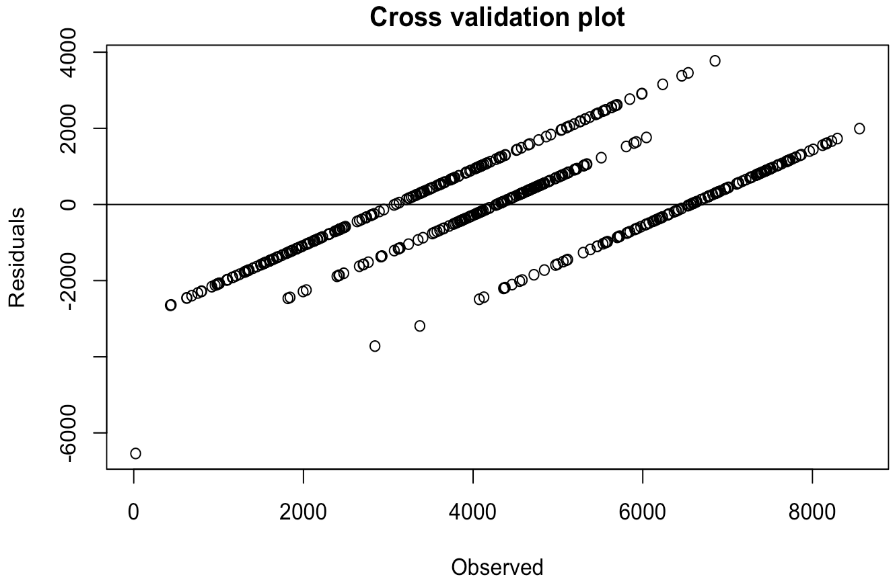


```
CART

511 samples
 18 predictor

No pre-processing
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 341, 340, 341
Resampling results across tuning parameters:

  cp          RMSE      Rsquared   MAE
  0.08168267  1220.872  0.5955924   941.1518
  0.21052329  1481.248  0.4027702  1227.5546
  0.36684282  1720.186  0.3078727  1426.4151

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.08168267.
```

The model was evaluated using cross-validation with 3 folds, and the results are shown for different values of the cost complexity parameter. The results suggest that the optimal value for the cost complexity parameter is 0.0816, as it yielded the smallest RMSE among all tested values of cp.

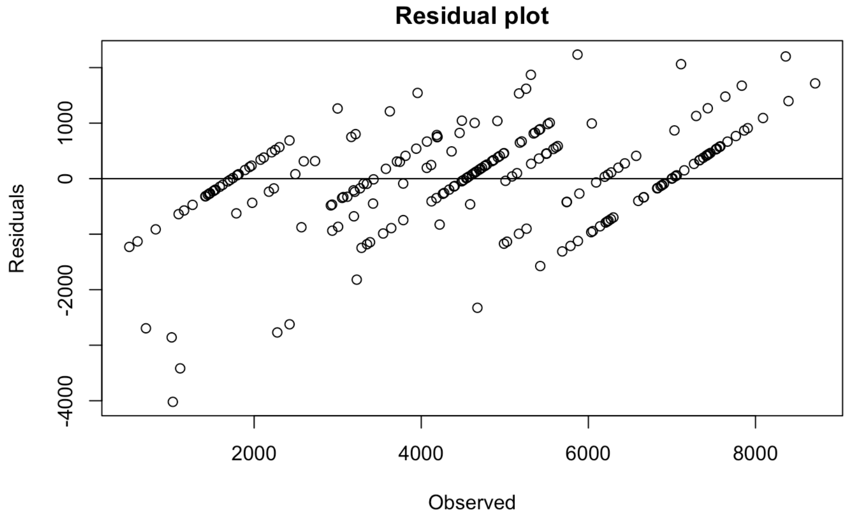
The performance metrics for this optimal model were: RMSE = 1220.872, R-squared = 0.5955924, and MAE = 941.1518. R-squared value indicates that the predictor was able to explain only 59.5% of the variance in the target variable.

The cross-validation prediction plot shows a pattern but most of the data points are leaning away from the diagonal line. This means that the model is not making accurate predictions. When a cross-validation plot shows a pattern, it usually indicates that the model is not performing well on the validation set. A pattern may suggest that the model is overfitting or underfitting the data. Overfitting occurs when the model is too complex and captures noise in the training data, leading to poor performance on new data.



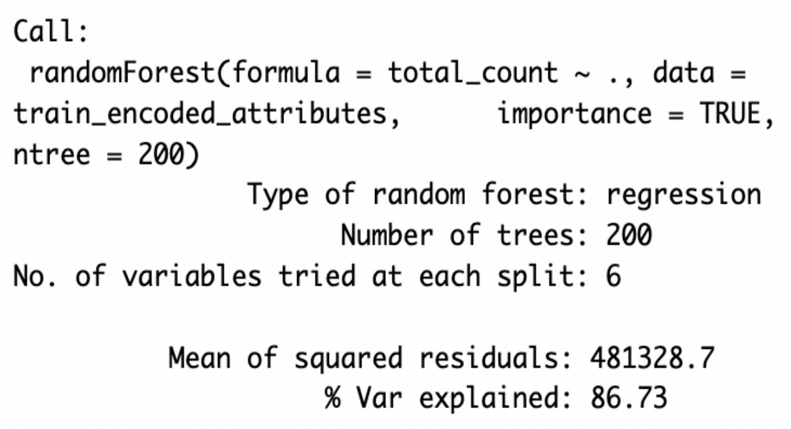
Residual Plot

This plot shows somewhat of a pattern but also has quite a few data points that are scattered all over. This means that the model is not very accurate for the fitted data. When a residual plot shows a pattern, it suggests that there is a relationship between the independent variables and the dependent variable that is not being captured by the model. The pattern could indicate that the model is mis-specified or that there is a nonlinear relationship between the variables.



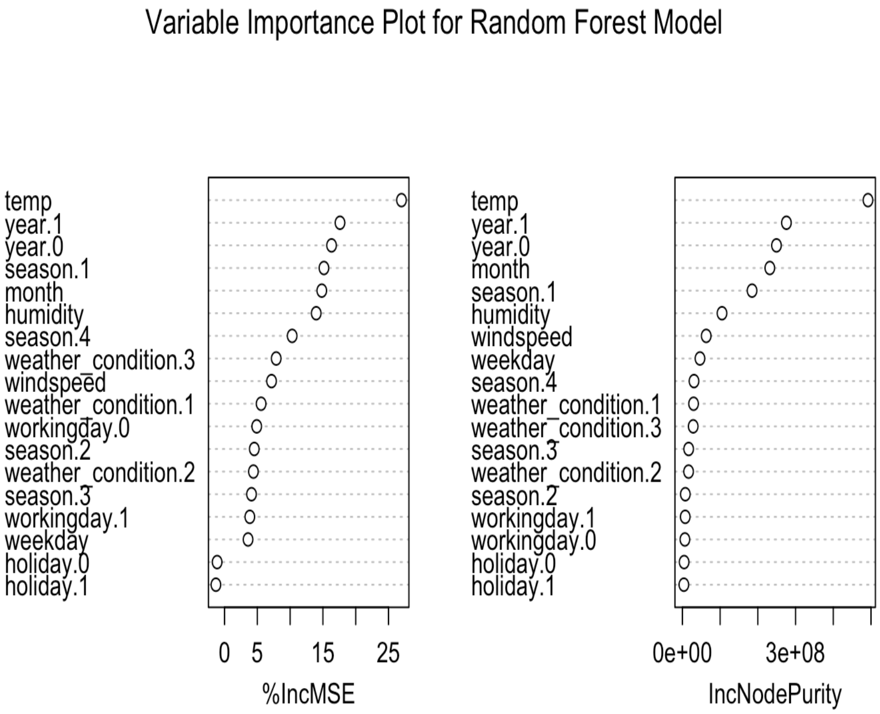
Random Forest Regression

Random Forest Regression combines multiple decision trees to create an ensemble model that can predict continuous numerical values. The R-squared or coefficient of determination is 0.8673 on average for 3-fold cross validation, it means that predictor is only able to predict 86.73% of the variance in the target variable which is contributed by independent variables.



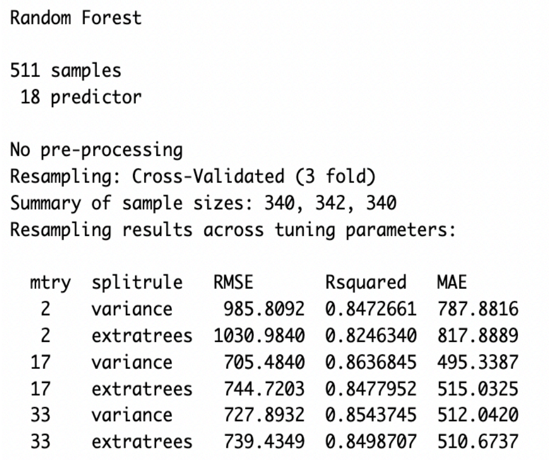
```
Call:
 randomForest(formula = total_count ~ ., data =
train_encoded_attributes,      importance = TRUE,
ntree = 200)
               Type of random forest: regression
                     Number of trees: 200
No. of variables tried at each split: 6

          Mean of squared residuals: 481328.7
                    % Var explained: 86.73
```

From the variable importance plot, we can clearly see that the temperature and both the years (2011 and 2012) are the top three variables that are strongly correlated to the total bike users.



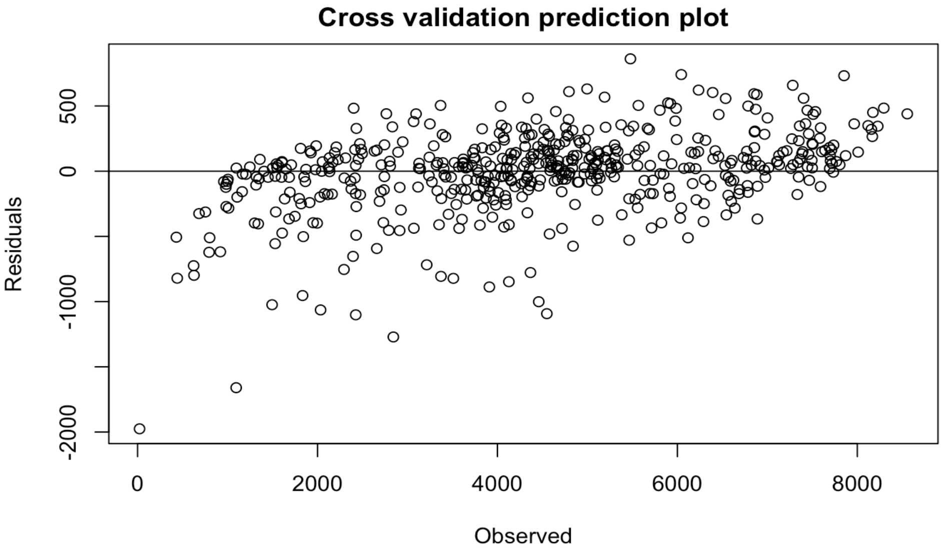
Cross Validation Prediction

The optimal model was selected based on the smallest RMSE value, and it was found to have mtry = 17, splitrule = variance, and min.node.size = 5.



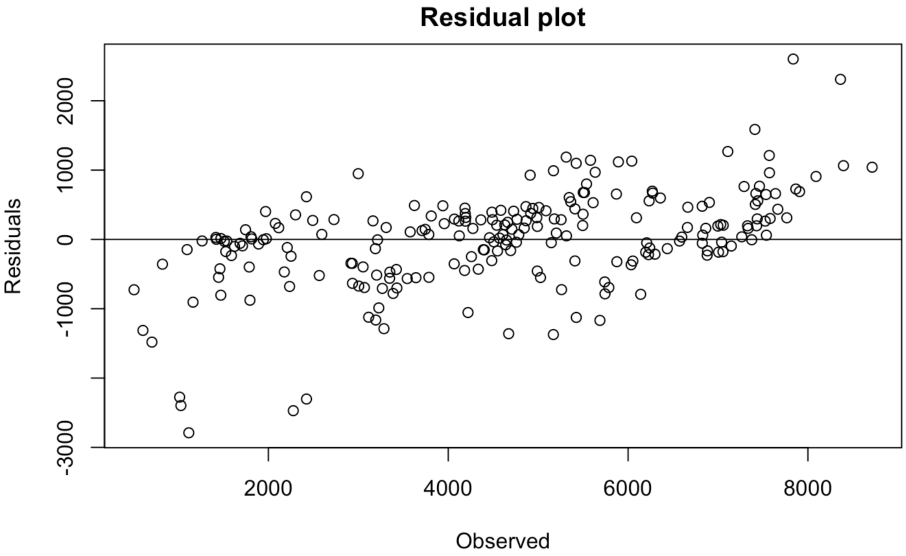
```
Random Forest

511 samples
 18 predictor

No pre-processing
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 340, 342, 340
Resampling results across tuning parameters:

  mtry  splitrule   RMSE       Rsquared   MAE
   2    variance     985.8092  0.8472661  787.8816
   2    extratrees  1030.9840  0.8246340  817.8889
  17    variance     705.4840  0.8636845  495.3387
  17    extratrees   744.7203  0.8477952  515.0325
  33    variance     727.8932  0.8543745  512.0420
  33    extratrees   739.4349  0.8498707  510.6737
```

In the CV plot, some data points may have similar variance between the predicted and actual values, resulting in a tight cluster of points on the plot, while others may have greater variance, resulting in a more scattered cluster of points. This can indicate that the model is better at predicting certain types of data points than others but is good overall.



Residual Plot

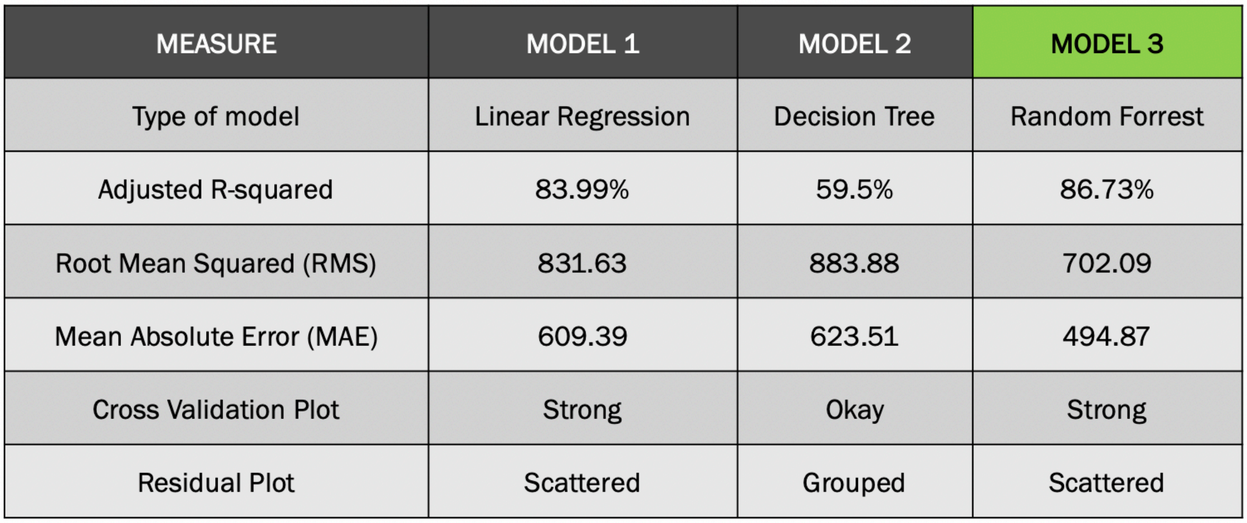
In general, a good residual plot should show no clear pattern, with the points scattered randomly around the x-axis. This plot shows somewhat of a pattern but also has quite a few data points that are scattered all over. It is a good plot overall.



Conclusion

To find out which model is the best for predicting the bike rental demand based on factors such as weather and holiday season, I have compiled all of the statistical data generated using the three models into a table for better visualization.

It is clear that Model 3, which is the Random Forrest Regression works best for the bike rental demand data set. Among all of the variables that were considered for analysis, the temperature, month and year come out on top.



MEASURE	MODEL 1	MODEL 2	MODEL 3
Type of model	Linear Regression	Decision Tree	Random Forrest
Adjusted R-squared	83.99%	59.5%	86.73%
Root Mean Squared (RMS)	831.63	883.88	702.09
Mean Absolute Error (MAE)	609.39	623.51	494.87
Cross Validation Plot	Strong	Okay	Strong
Residual Plot	Scattered	Grouped	Scattered