International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)
ISSN: 2456-236X
Vol. 05 Issue 01 | 2020

Prediction of Bike Rental Count

Shweta Singh
Department of Computer Application Jain University, Bangalore, India

ABSTRACT

In recent years, bike-sharing programs have become more prevalent. Bicycle usage can be affected by different factors, such as nearby events, road closures, and on-campus traffic policies. The research presented here analyzed the effect of weather (average temperature, total daily precipitation, average wind speed, and weather outlook), day of the week, holiday/workday, month, and season on the use of the Great Rides Bike Share program in Fargo, North Dakota, U.S.A. This study also focused on predicting the 2016 rental demand for the Great Rides Bike Share program using Bayesian methods and decision trees. Further, the order of importance among the causal attributes was assessed. It was found that decision trees worked well to predict the 2016 demand

1. INTRODUCTION

Today, more than 500 cities in 49 countries host bike-sharing programs. Urban transport advisor Peter Midgley notes that "bike sharing has experienced the fastest growth of any mode of transport in the history of the planet" [1]. Modern bike-sharing systems have greatly reduced the theft and vandalism that hindered earlier programs by using easily identified specialty bicycles with unique parts that would have little value to a thief, by monitoring the cycles' locations with radio frequency or GPS, and by requiring credit-card payment or smart-card-based membership to check out bikes. With most systems, after paying a daily, weekly, monthly, or annual membership fee, riders can pick up a bicycle that is locked to a well-marked bike rack or electronic docking station for a short ride (typically an hour or less) at no additional cost and can return it to any station in the system. Riding longer than the program's specified amount of time generally incurs additional fees to maximize the number of available bikes. Bike-sharing programs are becoming popular for the following reasons [2]: • They decrease greenhouse gases and improve public health. • They increase transit use due to the new bike transit trips, the improved connectivity to other modes of transit because of the first-mile/last-mile solution that bike-sharing helps solve, and the decreased number of personal vehicle trips. Due to the increased popularity of these bike-sharing programs across the world, it is increasingly becoming important to analyze these systems from different perspectives. Figure 1 shows the growth of these bike-sharing programs over the last decade. In this paper, I focus on predicting the 2016 bike-rental demand for the Great Rides Bike Share system based in Fargo, North Dakota. Fargo's Great Rides is an 11-station, 101-bicycle seasonal system. In 2015, there were 143,000 trips and an average of 6-7 rides per bike per day, more usage per bike than in 2 New York; Washington, D.C.; or Paris [3]. The main reason for the program's success is the integration with student IDs; the Great Rides seasonal pass is included as part of the mandatory student-activity fees at North Dakota State University (NDSU).

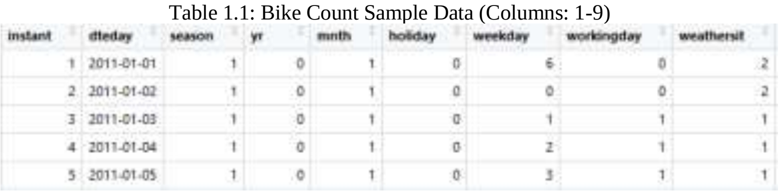
1.1 Problem Statement

The point of this undertaking is to foresee the check of bicycle rentals dependent on the occasional and ecological settings. By foreseeing the check, it is conceivable to help oblige in dealing with the quantity of bicycles needed consistently, and being ready for appeal of bicycles during top periods.

1.2 Data

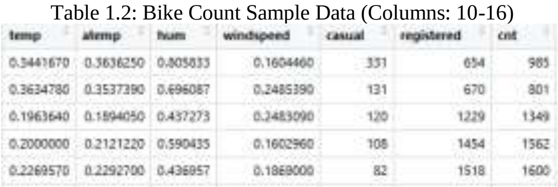
The objective is to manufacture relapse models which will anticipate the quantity of bicycles utilized dependent on the ecological and season conduct. Given underneath is an example of the informational collection that we are utilizing to anticipate the quantity of bicycles

Table 1.1: Bike Count Sample Data (Columns: 1-9)



instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit
1	2011-01-01	1	0	1	0	6	0	2
2	2011-01-02	1	0	1	0	0	0	2
3	2011-01-03	1	0	1	0	1	1	1
4	2011-01-04	1	0	1	0	2	1	1
5	2011-01-05	1	0	1	0	3	1	1

Table 1.2: Bike Count Sample Data (Columns: 10-16)



temp	atemp	hum	windspeed	casual	registered	cnt
0.3441670	0.3636250	0.805833	0.1604460	331	654	985
0.3634780	0.3537390	0.696087	0.2485390	131	670	801
0.1963640	0.1894050	0.437273	0.2483090	120	1229	1349
0.2000000	0.2121220	0.590435	0.1602960	108	1454	1562
0.2269570	0.2292700	0.436957	0.1869000	82	1518	1600

2. METHODOLOGY

Pre-Processing A prescient model necessitates that we take a gander at the information before we begin to make a model. Be that as it may, in information mining, seeing information alludes to investigating the information, cleaning the information just as envisioning the information through diagrams and plots. This is known as Exploratory Data Analysis.

Distribution of consistent factors It can be seen from the beneath histograms is that temperature and feel temperature are typically disseminated, whereas the factors windspeed and stickiness are marginally slanted. The skewness is likely a result of the presence of anomalies and extraordinary information in those factors.

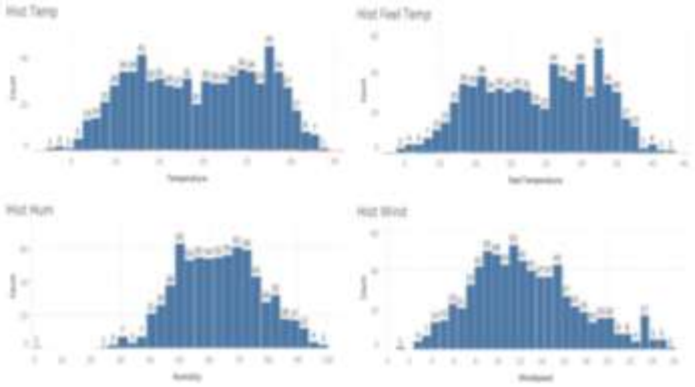


Fig 2.1: Distribution of continuous variables using Histograms

Distribution of clear-cut factors the circulation of all out factors is as appeared in the beneath figure:

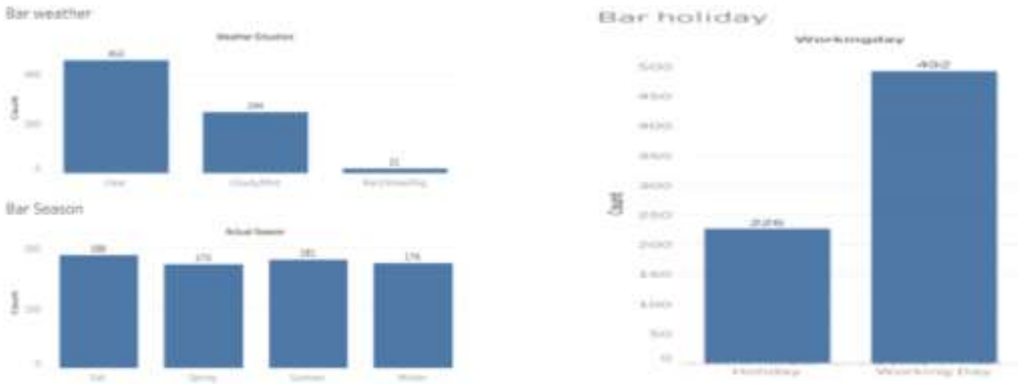


Fig 2.2: Distribution of categorical variables using bar plots

Relationship of Continuous factors against bicycle check the underneath figure shows the connection between nonstop factors and the objective variable utilizing disperse plot. It very well may be seen that there exists a straight good connection between the factors temperature and feel temperature with the bicycle rental check. There additionally exists a negative straight connection between the variable's dampness and windspeed with the bicycle rental check.

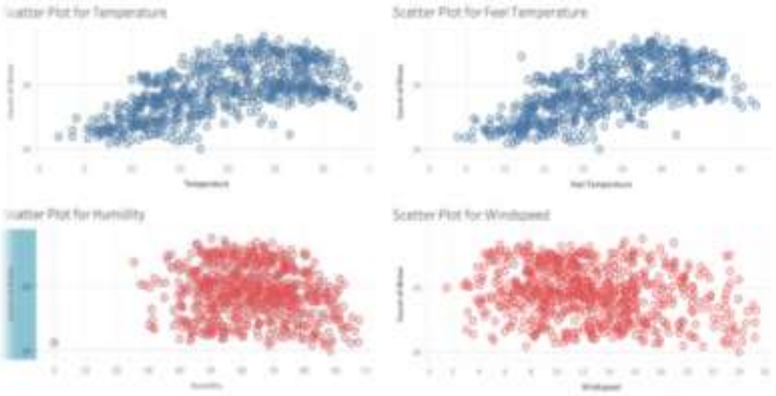


Fig 2.3: Scatter plot for continuous variables

Detection of exceptions: Outliers are distinguished utilizing boxplots. Beneath figure delineates the boxplots for all the ceaseless factors.

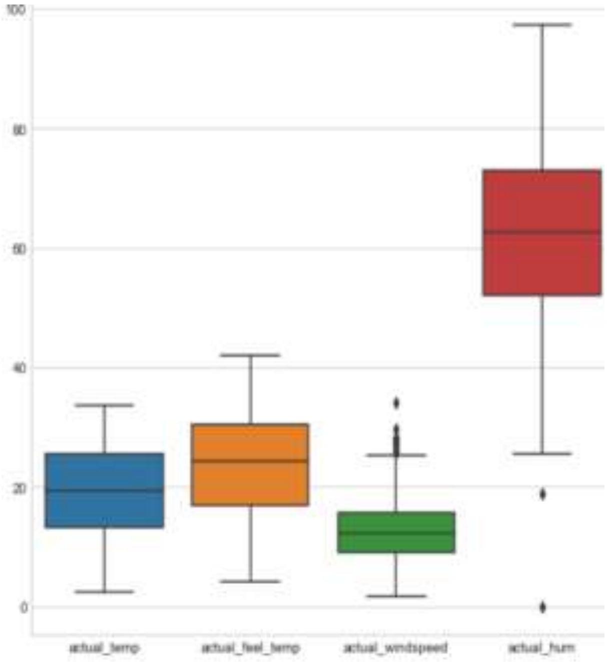


Fig 2.4: Boxplot of continuous variables

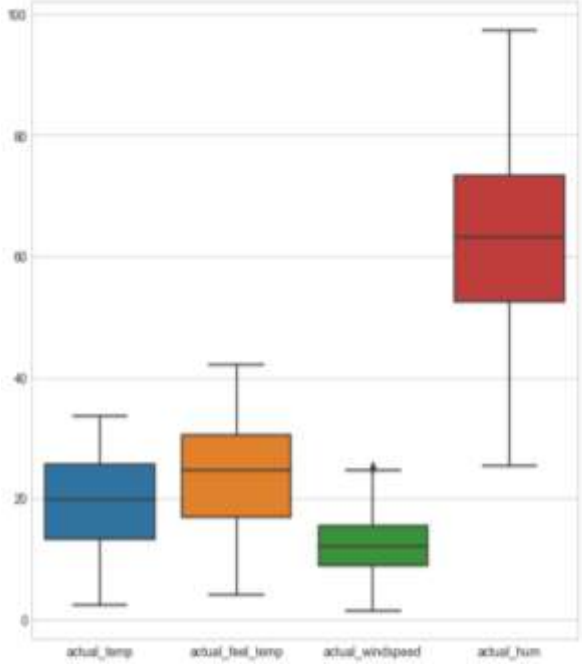


Fig 2.5: Boxplot of continuous variables after removal of outliers

Exceptions can be taken out utilizing the Boxplot details technique, wherein the Inter Quartile Range (IQR) is determined and the base and most extreme worth are determined for the factors. Any worth going external the base and most extreme worth are disposed of. The boxplot of the persistent factors in the wake of eliminating the exceptions is appeared in the beneath figure

It very well may be seen from the dissemination of Windspeed and stickiness after expulsion of anomalies, is that information isn't slanted as much as before the evacuation of exceptions. The figure appeared underneath shows the dispersion of constant factors utilizing histograms.

Feature Selection: Feature Selection diminishes the unpredictability of a model and makes it simpler to decipher. It likewise diminishes overfitting. Highlights are chosen dependent on their scores in different factual tests for their connection with the result variable. Connection plot is utilized to see whether there is any multicollinearity between factors. The exceptionally collinear factors are dropped and afterward the model is executed

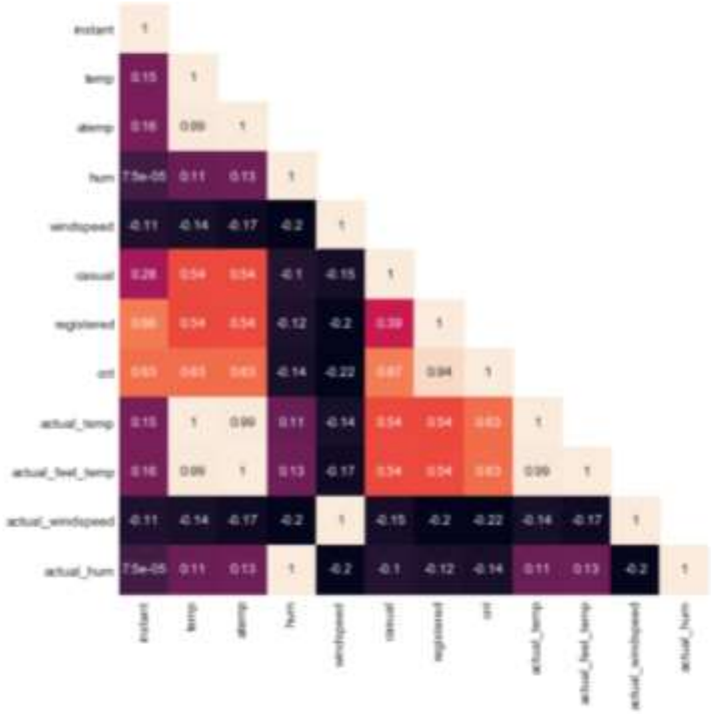
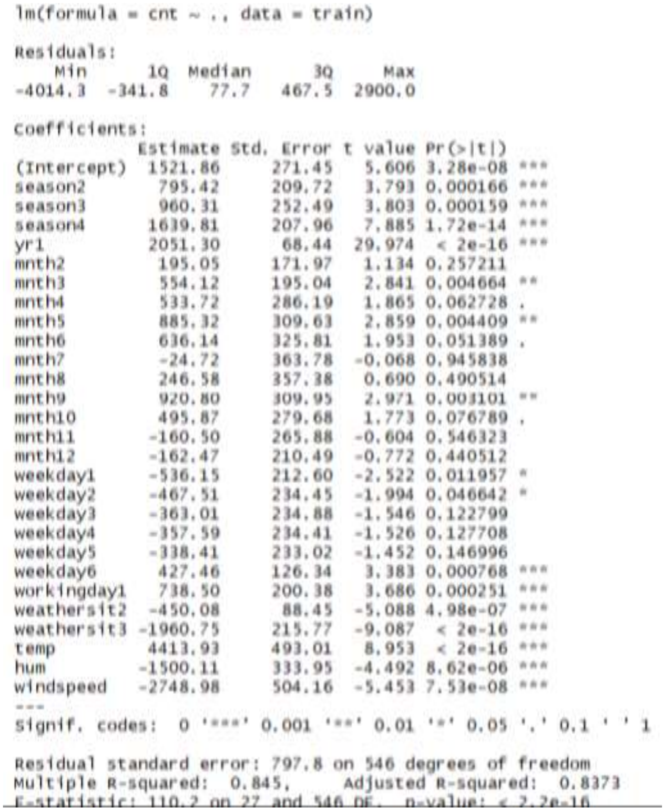


Fig 2.7: Correlation plot of all the variables

3. MODELING

Model Selection The reliant variable in our model is a persistent variable i.e., Count of bicycle rentals. Henceforth the models that we pick are Linear Regression, Decision Tree and Random Forest. The mistake metric picked for the difficult articulation is Mean Absolute Error (MAE).

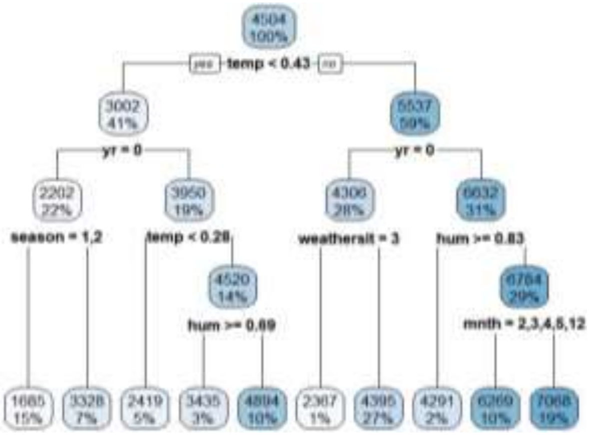
Multiple Linear Regression Multiple straight relapse is the most well-known type of direct relapse examination. Numerous direct relapse is utilized to clarify the connection between one ceaseless ward variable and at least two autonomous factors. The free factors can be constant or absolute



```
lm(formula = cnt ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-4014.3  -341.8    77.7   467.5  2900.0

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1521.86     271.45   5.606 3.28e-08 ***
season2          795.42     209.72   3.793 0.000166 ***
season3          960.31     252.49   3.803 0.000159 ***
season4         1639.81     207.96   7.885 1.72e-14 ***
yr1             2051.30      68.44  29.974  < 2e-16 ***
mnth2            195.05     171.97   1.134 0.257211
mnth3            554.12     195.04   2.841 0.004664 **
mnth4            533.72     286.19   1.865 0.062728 .
mnth5            885.32     309.63   2.859 0.004409 **
mnth6            636.14     325.81   1.953 0.051389 .
mnth7            -24.72     363.78  -0.068 0.945838
mnth8            246.58     357.38   0.690 0.490514
mnth9            920.80     309.95   2.971 0.003101 **
mnth10           495.87     279.68   1.773 0.076789 .
mnth11          -160.50     265.88  -0.604 0.546323
mnth12          -162.47     210.49  -0.772 0.440512
weekday1        -536.15     212.60  -2.522 0.011957 *
weekday2        -467.51     234.45  -1.994 0.046642 *
weekday3        -363.01     234.88  -1.546 0.122799
weekday4        -357.59     234.41  -1.526 0.127708
weekday5        -338.41     233.02  -1.452 0.146996
weekday6         427.46     126.34   3.383 0.000768 ***
workingday1      738.50     200.38   3.686 0.000251 ***
weathersit2     -450.08      88.45  -5.088 4.98e-07 ***
weathersit3    -1960.75     215.77  -9.087  < 2e-16 ***
temp            4413.93     493.01   8.953  < 2e-16 ***
hum            -1500.11     333.95  -4.492 8.62e-06 ***
windspeed      -2748.98     504.16  -5.453 7.53e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 797.8 on 546 degrees of freedom
Multiple R-squared:  0.845,     Adjusted R-squared:  0.8373
F-statistic: 110.2 on 27 and 546 DF,  p-value: < 2.2e-16
```

As should be obvious the Adjusted R-squared worth, we can clarify 83.73% of the information utilizing our various straight relapse model. By taking a gander at the F-measurement and consolidated p-esteem we can reject the invalid speculation that target variable doesn't rely upon any of the indicator factors. This model clarifies the information well indeed and is viewed as acceptable. Even in the wake of eliminating the non-critical factors, the exactness, Adjusted R-squared and Fstatistic don't change by a lot, consequently the precision of this model is picked to be conclusive. Mean Absolute Error (MAE) is determined and discovered to be 494. MAPE of this numerous direct relapse model is 12.17%. Henceforth the exactness of this model is 87.83%. This model performs very well for this test information.

Decision Tree: A choice tree can be utilized to outwardly and expressly speak to choices and dynamic. As the name goes, it utilizes a tree-like model of choices Utilizing choice tree, we can anticipate the estimation of bicycle tally. MAE for this model is 684. The MAPE for this choice tree is 17.47%. Subsequently the exactness for this model is 82.53%.



Random Forest: Using Classification for forecast examination for this situation isn't typical, however it tends to be finished. The quantity of choice trees utilized for forecast in the timberland is 500. MAE for this model is 392. Utilizing arbitrary woodland, the MAPE was discovered to be 10.68%. Thus the exactness is 89.32%.

4. CONCLUSION

Now that we have a couple of models for foreseeing the objective variable, we have to choose which one to pick. There are a few standards that exist for assessing and contrasting models. We can look at the models utilizing any of the accompanying rules:

- Prescient Performance
- Interpretability
- Computational Efficiency For our situation of Bike check forecast Data, Interpretability and Computation Efficiency, don't hold a lot of criticalness.

Hence, we will utilize Predictive execution as the standards to look at and assess models. Prescient execution can be estimated by contrasting Predictions of the models and genuine estimations of the objective factors, and figuring some normal mistake measure.

4.1 Mean Absolute Error (MAE)

MAE is one of the mistake estimates used to ascertain the prescient exhibition of the model. We will apply this measure to our models that we have created in the past segment.

MAE <-work (real, pred) { print(mean (abs (real - pred))) }

Linear Regression Model: MAE = 494 Decision Tree: MAE = 684. Arbitrary Forest: MAE = 392

Based on the above mistake measurements, Random Forest is the better model for our investigation. Henceforth Random Forest is picked as the model for forecast of bicycle rental check.

5. APPENDIX

5.1 Figures

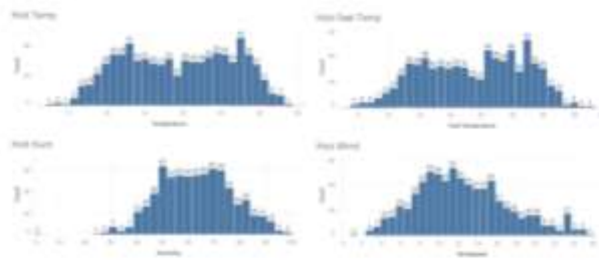


Fig 2.1: Distribution of continuous variables using Histograms

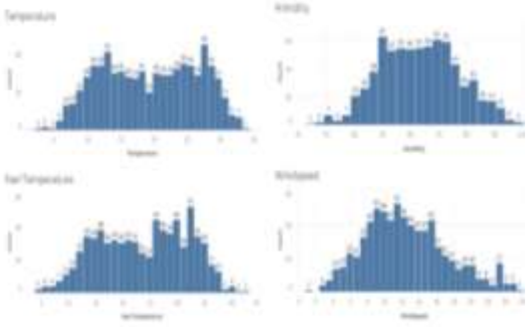


Fig 2.6: Distribution of numerical data using histograms after removal of outliers

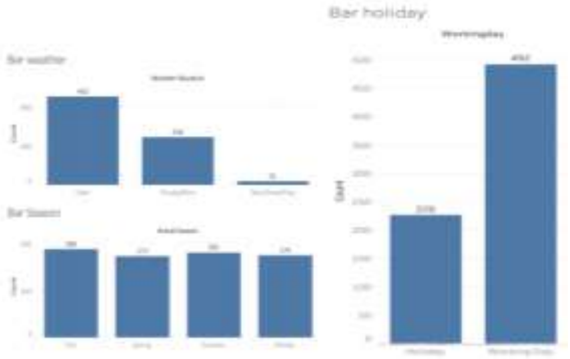


Fig 2.2: Distribution of categorical variables using bar plots

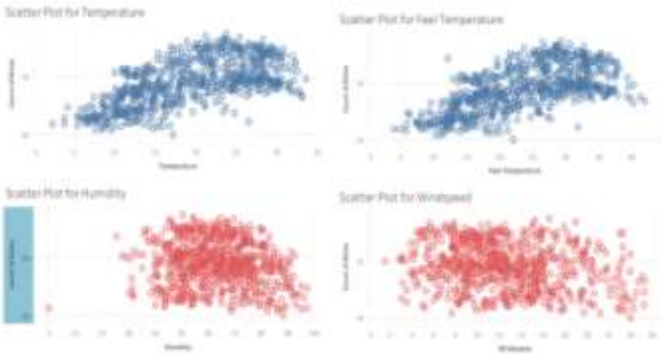


Fig 2.3: Scatter plot for continuous variables

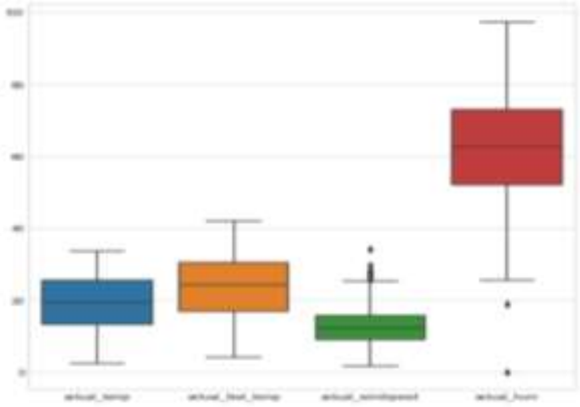


Fig 2.4: Boxplot of continuous variables

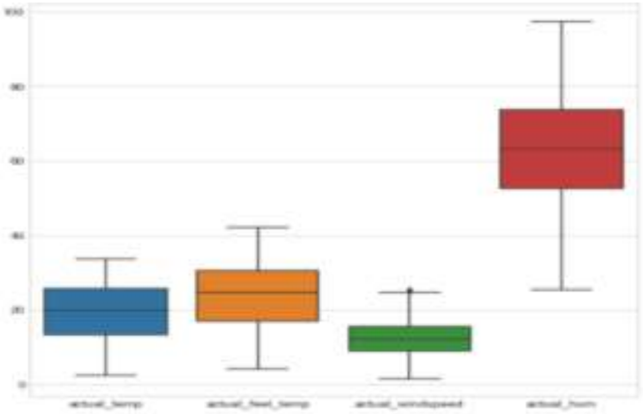
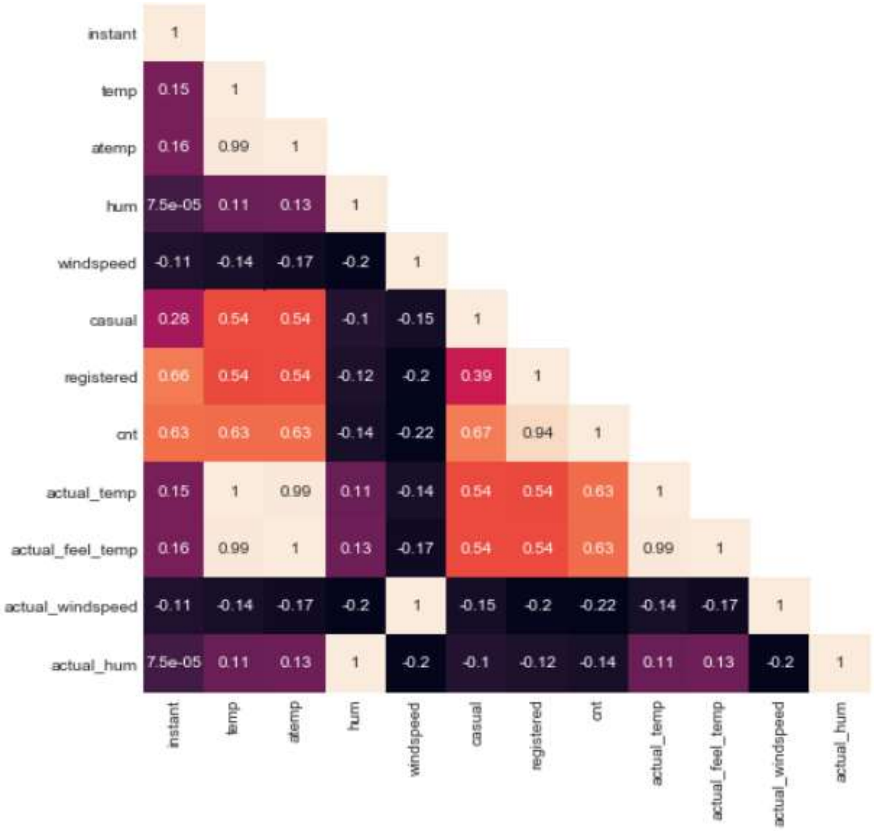


Fig 2.5: Boxplot of continuous variables after removal of outliers



	instant	temp	atemp	hum	windspeed	casual	registered	cnt	actual_temp	actual_feel_temp	actual_windspeed	actual_hum
instant	1											
temp	0.15	1										
atemp	0.16	0.99	1									
hum	7.5e-05	0.11	0.13	1								
windspeed	-0.11	-0.14	-0.17	-0.2	1							
casual	0.28	0.54	0.54	-0.1	-0.15	1						
registered	0.66	0.54	0.54	-0.12	-0.2	0.39	1					
cnt	0.63	0.63	0.63	-0.14	-0.22	0.67	0.94	1				
actual_temp	0.15	1	0.99	0.11	-0.14	0.54	0.54	0.63	1			
actual_feel_temp	0.16	0.99	1	0.13	-0.17	0.54	0.54	0.63	0.99	1		
actual_windspeed	-0.11	-0.14	-0.17	-0.2	1	-0.15	-0.2	-0.22	-0.14	-0.17	1	
actual_hum	7.5e-05	0.11	0.13	1	-0.2	-0.1	-0.12	-0.14	0.11	0.13	-0.2	1

Fig 2.7: Correlation plot of all the variables

6. R CODE

```
####################EXPLORE USING GRAPHS##################### #CHECK THE
DISTRIBUTION OF CATEGORICAL DATA USING BAR GRAPH BAR1 = GGPLOT(DATA = DAY, AES(X =
ACTUAL_SEASON)) + GEOM_BAR() + GGTITLE("COUNT OF SEASON") BAR2 = GGPLOT(DATA = DAY,
AES(X = ACTUAL_WEATHERSIT)) + GEOM_BAR() + GGTITLE("COUNT OF WEATHER") BAR3 =
GGPLOT(DATA = DAY, AES(X = ACTUAL_HOLIDAY)) + GEOM_BAR() + GGTITLE("COUNT OF
HOLIDAY") BAR4 = GGPLOT(DATA = DAY, AES(X = WORKINGDAY)) + GEOM_BAR() +
GGTITLE("COUNT OF WORKING DAY")
GRIDEXTRA::GRID.ARRANGE(BAR1,BAR2,BAR3,BAR4,NCOL=2) #CHECK THE DISTRIBUTION OF
NUMERICAL DATA USING HISTOGRAM HIST1 = GGPLOT(DATA = DAY, AES(X =ACTUAL_TEMP)) +
GGTITLE("DISTRIBUTION OF TEMPERATURE") + GEOM_HISTOGRAM(BINS = 25) HIST2 =
GGPLOT(DATA = DAY, AES(X =ACTUAL_HUM)) + GGTITLE("DISTRIBUTION OF HUMIDITY") +
GEOM_HISTOGRAM(BINS = 25) HIST3 = GGPLOT(DATA = DAY, AES(X =ACTUAL_FEEL_TEMP)) +
GGTITLE("DISTRIBUTION OF FEEL TEMPERATURE") + GEOM_HISTOGRAM(BINS = 25) HIST4 =
```

GGPLOT(DATA = DAY, AES(X =ACTUAL_WINDSPEED)) + GGTITLE("DISTRIBUTION OF WINDSPEED") + GEOM_HISTOGRAM(BINS = 25) GRIDEXTRA::GRID.ARRANGE(HIST1,HIST2,HIST3,HIST4,NCOL=2) #CHECK THE DISTRIBUTION OF NUMERICAL DATA USING SCATTERPLOT SCAT1 = GGPLOT(DATA = DAY, AES(X =ACTUAL_TEMP, Y = CNT)) + GGTITLE("DISTRIBUTION OF TEMPERATURE") + GEOM_POINT() + XLAB("TEMPERATURE") + YLAB("BIKE COUNT") SCAT2 = GGPLOT(DATA = DAY, AES(X =ACTUAL_HUM, Y = CNT)) + GGTITLE("DISTRIBUTION OF HUMIDITY") + GEOM_POINT(COLOR="RED") + XLAB("HUMIDITY") + YLAB("BIKE COUNT") SCAT3 = GGPLOT(DATA = DAY, AES(X =ACTUAL_FEEL_TEMP, Y = CNT)) + GGTITLE("DISTRIBUTION OF FEEL TEMPERATURE") + GEOM_POINT() + XLAB("FEEL TEMPERATURE") + YLAB("BIKE COUNT") SCAT4 = GGPLOT(DATA = DAY, AES(X =ACTUAL_WINDSPEED, Y = CNT)) + GGTITLE("DISTRIBUTION OF WINDSPEED") + GEOM_POINT(COLOR="RED") + XLAB("WINDSPEED") + YLAB("BIKE COUNT") GRIDEXTRA::GRID.ARRANGE(SCAT1,SCAT2,SCAT3,SCAT4,NCOL=2) #CHECK FOR OUTLIERS IN DATA USING BOXPLOT CNAMES = COLNAMES(DAY[,C("ACTUAL_TEMP","ACTUAL_FEEL_TEMP","ACTUAL_WINDSPEED","ACTUAL_HUM")]) FOR (I IN 1:LENGTH(CNAMES)) { ASSIGN(PASTE0("GN",I), GGPLOT(AES_STRING(Y = CNAMES[I]), DATA = DAY)+ STAT_BOXPLOT(GEOM = "ERRORBAR", WIDTH = 0.5) + GEOM_BOXPLOT(OUTLIER.COLOUR="RED", FILL = "GREY" 20 ,OUTLIER.SHAPE=18, OUTLIER.SIZE=1, NOTCH=FALSE) + THEME(LEGEND.POSITION="BOTTOM")+ LABS(Y=CNAMES[I]) + GGTITLE(PASTE("BOX PLOT FOR",CNAMES[I]))) } GRIDEXTRA::GRID.ARRANGE(GN1,GN3,GN2,GN4,NCOL=2) #REMOVE OUTLIERS IN WINDSPEED VAL = DAY[,19][DAY[,19] %IN% BOXPLOT.STATS(DAY[,19])$OUT] DAY = DAY[WHICH(!DAY[,19] %IN% VAL),] #CHECK FOR MULTICOLLINEARITY USING VIF DF = DAY[,C("INSTANT","TEMP","ATEMP","HUM","WINDSPEED")] VIFCOR(DF) #CHECK FOR COLLINEARITY USING CORELATION GRAPH CORRGRAM(DAY, ORDER = F, UPPER.PANEL=PANEL.PIE, TEXT.PANEL=PANEL.TXT, MAIN = "CORRELATION PLOT") #REMOVE THE UNWANTED VARIABLES DAY <- SUBSET(DAY, SELECT = -C(INSTANT,DTEDAY,ATEMP,CASUAL,REGISTERED,ACTUAL_TEMP,ACTUAL_FEEL_TEMP,ACTUAL_WINDSPEED,AC TUAL_HUM,ACTUAL_SEASON,ACTUAL_YR,ACTUAL_HOLIDAY,ACTUAL_WEATHERSIT)) ##########################DECISION TREE######################### #DIVIDE THE DATA INTO TRAIN AND TEST SET.SEED(123) TRAIN_INDEX = SAMPLE(1:NROW(DAY), 0.8 * NROW(DAY)) TRAIN = DAY[TRAIN_INDEX,] TEST = DAY[-TRAIN_INDEX,] #RPART FOR REGRESSION DT_MODEL = RPART(CNT ~ ., DATA = TRAIN, METHOD = "ANOVA") #PREDICT THE TEST CASES DT_PREDICTIONS = PREDICT(DT_MODEL, TEST[,-11]) #CREATE DATAFRAME FOR ACTUAL AND PREDICTED VALUES DF = DATA.FRAME("ACTUAL"=TEST[,11], "PRED"=DT_PREDICTIONS) HEAD(DF) #CALCULATE MAPE REGR.EVAL(TRUES = TEST[,11], PREDS = DT_PREDICTIONS, STATS = C("MAE","MSE","RMSE","MAPE")) 21 ####################RANDOM FOREST################ #TRAIN THE DATA USING RANDOM FOREST RF_MODEL = RANDOMFOREST(CNT~., DATA = TRAIN, NTREE = 500) #PREDICT THE TEST CASES RF_PREDICTIONS = PREDICT(RF_MODEL, TEST[,-11]) #CREATE DATAFRAME FOR ACTUAL AND PREDICTED VALUES DF = CBIND(DF,RF_PREDICTIONS) HEAD(DF) #CALCULATE MAPE REGR.EVAL(TRUES = TEST[,11], PREDS = RF_PREDICTIONS, STATS = C("MAE","MSE","RMSE","MAPE")) #####################LINEAR REGRESSION################ #TRAIN THE DATA USING LINEAR REGRESSION LR_MODEL = LM(FORMULA = CNT~., DATA = TRAIN) #CHECK THE SUMMARY OF THE MODEL SUMMARY(LR_MODEL) #PREDICT THE TEST CASES LR_PREDICTIONS = PREDICT(LR_MODEL, TEST[,-11]) #CREATE DATAFRAME FOR ACTUAL AND PREDICTED VALUES DF = CBIND(DF,LR_PREDICTIONS) HEAD(DF) #CALCULATE MAPE REGR.EVAL(TRUES = TEST[,11], PREDS = LR_PREDICTIONS, STATS = C("MAE","MSE","RMSE","MAPE")) #PREDICT A SAMPLE DATA PREDICT(LR_MODEL,TEST[2,])