Case Study: Bike rental

This dataset contains the information of a rental bike firm that rents different bikes throughput the year. The firm completely depends upon the rental service to run and depends upon different types of predictors like weather Condition, Temperature and other predictors.

1.

Motivation:

Due to Pandemic the demand of rental bikes went down as the lockdown was imposed all over the country and thus had a huge impact on tourism and business related with that. As country is now moving back towards normal the Bike rental firm in Goa now wants to do a fresh start altogether with moving back to business based on rental bikes, but this time they want to predict their future scope of business in the field and how they can lookout for more opportunities at different parts of the state.

Objective:

The main objective here is to predicting the demand for rental bikes depending on various factors throughout a year.

Outcome:

By using regression analysis here, we can predict the important factors those have an impact over business cycle and their revenue streams. Regression model here will help us to predict the predictors those are actually impacting business and remove those reductant predictors those do not affect business or have less impact on business and are negligeable. This will help us in building a model which is both cost effective and helps on future forecasting of the business as per changes in coefficients of predictors is helpful for business in the long run. This analysis will let business know when and how they need to react on different point of time to maintain a growth and profit generation altogether.

2. Data gathered from:

https://www.kaggle.com/imakash3011/rental-bike-sharing

3. Data Set Description

Name	Type	Details
instant	Categorical Variable	Serial Number, Index
dteday	Categorical Variable	Date of data entry
season	Categorical Variable	Season-
yr	Categorical Variable	Year
mnth	Categorical Variable	Month
holiday	Categorical Variable	Listed as Holiday or not
weekday	Categorical Variable	A regular day except Sunday
workingday	Categorical Variable	Not a holiday or weekend
weathersit	Categorical Variable	Weather Situation
temp	Continuous Variable	Temperature in Celsius
atemp	Continuous Variable	Felt Temperature in Celsius
hum	Continuous Variable	Humidity
windspeed	Continuous Variable	Windspeed

casual	Continuous Variable	Count of casual users
registered	Continuous Variable	Count of registered users
count	Continuous Variable (Response Variable)	Count of total rental bikes including both casual and registered

Categorical Variables-

Season	1-Spring 2-Summer 3 Autumn 4- Autumn
Year	0-2018 1-2019
Month	0-12 (As per calendar)
Weather Situation	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

4. Descriptive analysis of the response Variable

We have considered count as the response variable in the bike rental dataset, and did descriptive analysis on it. We plotted Histogram, Box Plot and calculated Mean Median Mode on the response variable with the help R(RStudio)

A. Histogram

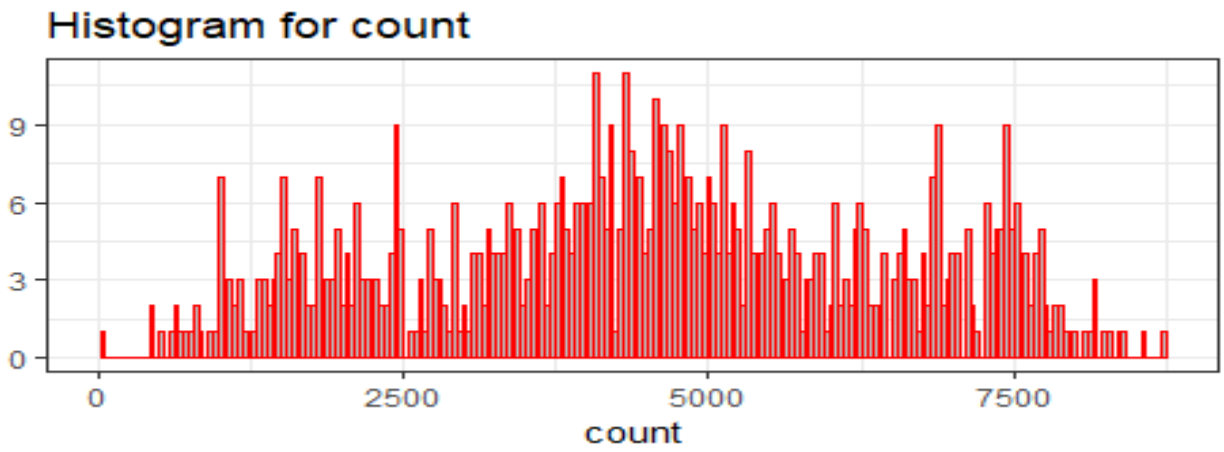

Figure 1

Interpretation – The Histogram clearly indicates that it is more or less normally distributed. The width of the histogram is just above 8500 and attains a peak at around 4500-4800 which falls at centre of distribution.

B. Box Plot

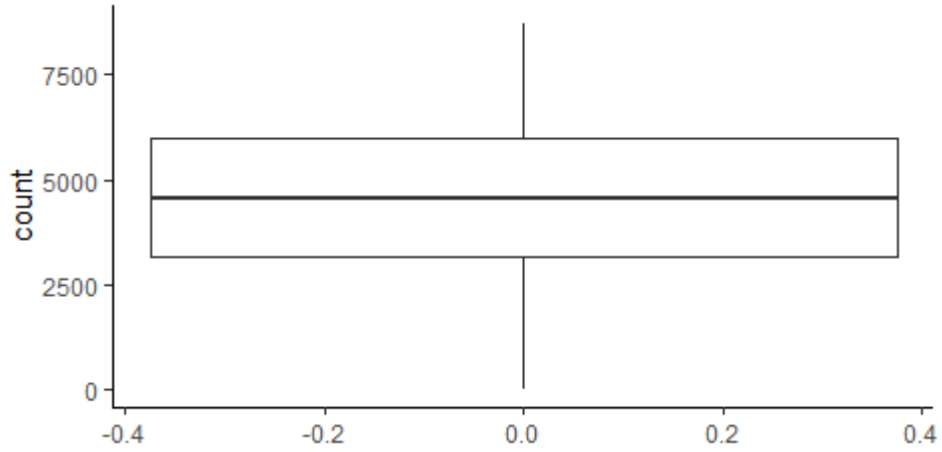

Figure 2

Interpretation – It clearly says that data has no outliers and both the whiskers are of the same size, thus indicating that the distribution is normal.

The percentile is as follows-

25th percentile 3170

50th percentile 4548

75th percentile 5966.

```
> quantile(data2$count)
    0%    25%    50%    75%   100%
 22.00 3169.75 4548.50 5966.00 8714.00
```

The interquartile ranges lie between 3170 and 5966. There are no extreme outliers.

```
> summary(data2$count)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     22    3170    4548    4508    5966    8714
```

C. Mean Median Mode

```
      nbr.val      nbr.null       nbr.na
7.300000e+02 0.000000e+00 0.000000e+00
         min          max         range
2.200000e+01 8.714000e+03 8.692000e+03
         sum       median          mean
3.290845e+06 4.548500e+03 4.508007e+03
     SE.mean CI.mean.0.95           var
7.165501e+01 1.406748e+02 3.748141e+06
     std.dev     coef.var
1.936012e+03 4.294607e-01
```

Figure 3

- ✓ Sample Size is 730
- ✓ As Median is 4548 and Mean is 4508 as both are almost equal hence it is a central tendency
- ✓ Dispersion: Range is 8692 (min- 22 and max -8714). Standard Deviation is 1936
- ✓ co-efficient of variation is 0.429. Co-efficient of variation is 42.9%,
- ✓ SD is around 42.9% relative to the mean. Since SD is less than mean which stats that the data is under-dispersed.

5. Fitting the MLR Model.

A. Applying the MLR

Step 1:

Divided dataset into 2 parts, 80% of the data set i.e., 584 observations are loaded

in train data and 20% of the data set i.e., 146 observations are loaded into test data.

R Code for setting up traindata and testdata:

```
set.seed(123)

indset<- sample(2,nrow(data2),replace=T,prob=c(0.8, 0.2))

traindata<-data2[indset==1,]

testdata<-data2[indset==2,]

head(traindata)
```

Once we divide the dataset, we are ready to run MLR

R Code For MLR-

```
model1=lm((count)~as.factor(season)+as.factor(yr)+as.factor(mnth)+as.factor(holiday)+as.factor(weekday)+as.factor(workingday)+as.factor(weathersit)+temp+atemp+hum+windspeed,data=traindata)

summary(model1)
```

We have omitted instant and dteday as both are reductant with respect to Multiple

Regression model analysis.

- ❖ The Result for MLR is shown below:

Residuals:

Min	1Q	Median	3Q	Max
-3883.5	-359.7	54.3	449.4	2773.1

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1847.950	267.161	6.917	1.28e-11	***
as.factor(season)2	873.197	201.360	4.336	1.72e-05	***
as.factor(season)3	976.221	243.138	4.015	6.76e-05	***
as.factor(season)4	1740.137	205.439	8.470	2.23e-16	***
as.factor(yr)1	2016.983	65.820	30.644	< 2e-16	***
as.factor(mnth)2	195.155	160.690	1.214	0.225083	
as.factor(mnth)3	644.123	187.484	3.436	0.000636	***
as.factor(mnth)4	353.945	277.630	1.275	0.202889	
as.factor(mnth)5	692.081	299.602	2.310	0.021256	*
as.factor(mnth)6	483.392	318.869	1.516	0.130102	
as.factor(mnth)7	-165.162	356.527	-0.463	0.643366	
as.factor(mnth)8	357.262	344.278	1.038	0.299858	
as.factor(mnth)9	705.038	303.516	2.323	0.020547	*
as.factor(mnth)10	373.926	272.732	1.371	0.170920	
as.factor(mnth)11	-308.158	260.984	-1.181	0.238209	
as.factor(mnth)12	-193.271	208.724	-0.926	0.354870	
as.factor(holiday)1	-859.607	460.286	-1.868	0.062354	.
as.factor(weekday)1	131.544	512.626	0.257	0.797576	
as.factor(weekday)2	156.294	509.096	0.307	0.758957	
as.factor(weekday)3	249.734	504.839	0.495	0.621022	
as.factor(weekday)4	386.315	512.650	0.754	0.451432	
as.factor(weekday)5	409.637	512.796	0.799	0.424732	
as.factor(weekday)6	-130.281	119.501	-1.090	0.276098	
as.factor(workingday)1	-487.551	508.875	-0.958	0.338434	
as.factor(weathersit)2	-497.564	85.942	-5.790	1.19e-08	***
as.factor(weathersit)3	-1875.403	222.424	-8.432	2.98e-16	***
temp	71.634	36.388	1.969	0.049501	*
atemp	36.397	30.954	1.176	0.240161	
hum	-14.130	3.260	-4.334	1.74e-05	***
windspeed	-41.776	7.073	-5.907	6.10e-09	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 773.2 on 552 degrees of freedom

Multiple R-squared: 0.8425, Adjusted R-squared: 0.8343

F-statistic: 101.9 on 29 and 552 DF, p-value: < 2.2e-16

Interpretation From MLR-

We have 7 categorical variables and 4 continuous variables.

List of Categorical Variables

For Categorical variables the predictors are one less than the actual database. For every categorical variable one base variable is considered. With respect to the base variable (missing one) the intercept is being calculated.

1. Season = (4-1) =3 Predictors
2. Year= (2-1) =1 Predictor
3. Month= (12-1) =11 Predictors
4. Holiday= (2-1) =1 Predictor
5. Weekday= (7-1) =6 Predictors
6. Working Day= (2-1) = 1 Predictor
7. Weathersit = (4-1) = 3 Predictors

List of Continuous Variables

1. temp = 1 predictor
2. atemp =1 predictor
3. humidity = 1 predictor
4. windspeed=1 predictor

Total number of predictors =29+1 (intercept) =30 predictors.

The most significant factors (***) are windspeed, humidity, weathersit, month, season.

- In this analysis have considered spring as season1 .and as per the MLR outcome season2 which is summer has a standard error of 897.642 that implies that with spring being considered the base season, bikes being rented in summer(season2) is going to increase by 897 units.
- Predictor temp is a Quantitative value which has a value of -18.32, which implies that if we keep all other factors constant, and increase temperature by 1 degree Celsius, the bike rent count is going to decrease by 18.32 units.

Relation between multiple R square and Adjusted R square

- Multiple R square value as 0.8452, it shows that 84.52% of the total variability in the response variable (count) is being explained by the full MLR model.

- ❖ The multiple R square and Adjusted R square values are around 84% which means that the model is a good fit.
- ❖ The difference between R square and Adjusted R square is not much, that means not many reductant predictors are present in the model.

- F test statistic= 118.1 with 29 numerator degrees of freedom and 627denominator degrees of Freedom and is distributed with F distribution of 29 and 627 degrees of freedom.
- As per MLr model p-value: < 2.2e-16 which is much less than 0.05 , thus we reject the Null Hypothesis and can conclude that full MLR model is significant at a 5% level of significance.

B. QQ-plot /Shapiro-Wilk's Test

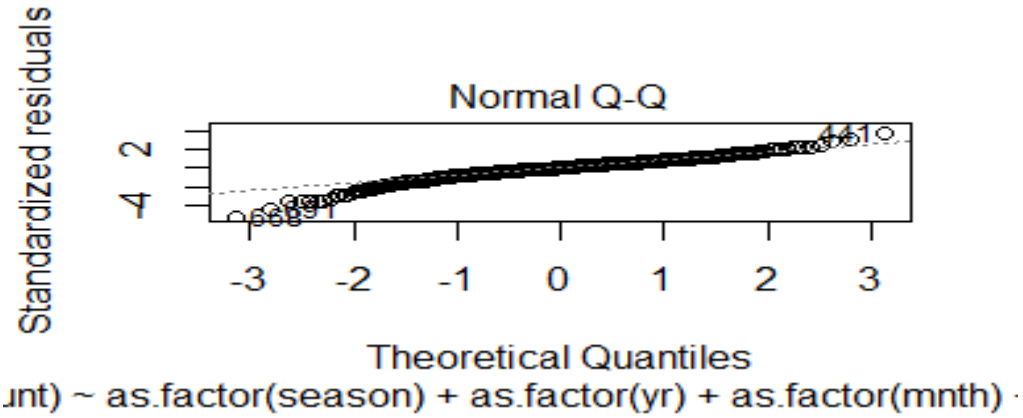

Figure 4

Interpretation-

The QQ plot of residuals is used to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line. In our model, all the points are not following a straight line and we can see some divergence from Straight line and according to that we can say that by observing QQ plot we can see some divergence from straight line and normality is not fulfilled

Shapiro-Wilk's Test -It is a formal hypothesis test to check whether the normality assumption is being satisfied. Here we consider null hypothesis for this test as that the data is normally distributed. Thus, if the p-value is greater than 0.05, then the null hypothesis is not rejected, and we can conclude that normality is satisfied.

Shapiro-Wilk's Test- Code and output

Shapiro-Wilk normality test

data: model1$residuals

W = 0.9536, p-value = 1.454e-12

As per above result here p value is less than 0.05, thus indicating normality is not satisfied.

As P < 0.05 in Shapiro Wilks test, therefore performing **box cox transformation** to find the ideal lambda

For normality we conduct a Box Cox Transformation of the data.

> bc = boxcox(model1, lambda= seq(-5,5))

> best.lam = bc$x[which(bc$y==max(bc$y))]

> best.lam

= 0.7575758

Once we get lambda, we adjust model by taking the response variable to the power of lambda

adjusted_mod1=lm((count)^0.75~as.factor(mnth)+as.factor(season)+as.factor(yr)+as.factor(workingday)+as.factor(weathersit)+temp+atemp+windspeed,data=traindata)

plot(adjusted_mod1)

Once we run the MLR with the lambda we got from box cox and checked with the QQ plots, we were not able to see that all the points are not following a straight line, thus normality is not fulfilled and we performed Shapiro Wilkson test on the adjusted model.

Shapiro-Wilk normality test

data: adjusted_mod1$residuals

W = 0.94051, p-value = 1.687e-14

As p value < 0.05 even after box cox transformation. hence continue with the original model model1.

Homoscedasticity-

The variance of residuals is constant across all the observations. residual plot outputs a scatter

plot between the fitted response values on the x-axis and the residuals on the y-axis.

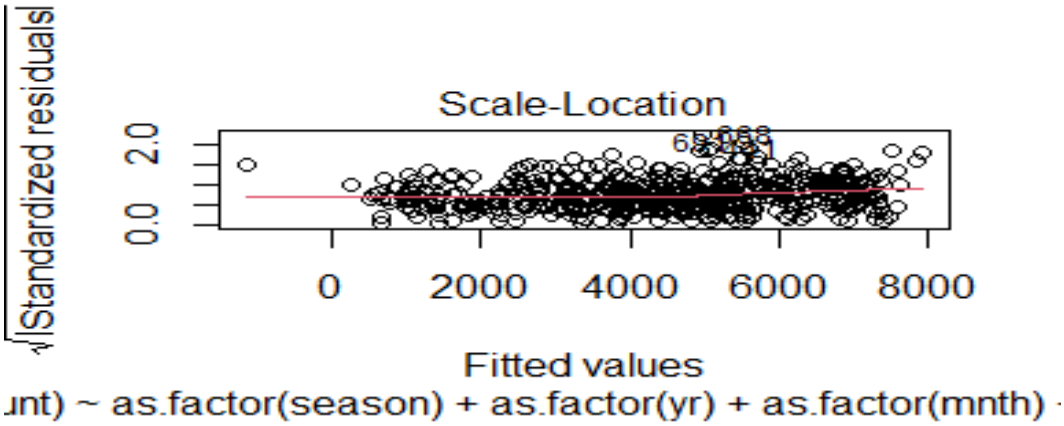

Figure 5

Interpretation-

By studying the plot generated from the fitted model we can clearly see that arandom scatter with no visible pattern is obtained and this indicates that Homoscedasticity is being depicted throughout the residuals.

Durbin-Watson test (Independence of Error Test)

This test helps us in checking independence of errors that is it checks whether the random errors are auto-correlated or not.

The null and alternative hypothesis for this test is given by:

H0: The residuals from the linear regression are uncorrelated (independence)

H1: There exists some form of autocorrelation (dependence) among the residuals.

Result:

Durbin-Watson test

data: model1

DW = 1.3429, p-value < 2.2e-16

alternative hypothesis: true autocorrelation is greater than 0

Here we can see that the p-value is greater than 0.05 indicating that the assumption of independence of errors is satisfied.

Multicollinearity-

It can be defined as the occurrence of high intercorrelations among two or more independent variables in a MLR model. This test helps us to determine how well each independent variable can be used effectively to predict and understand the dependent variable in a model presented.

```
> vif(model1)
                        GVIF       Df     GVIF^(1/(2*Df))
as.factor(season)     184.154825  3       2.385231
as.factor(yr)           1.053983  1       1.026637
as.factor(mnth)       472.412033 11       1.323001
as.factor(holiday)      5.848048  1       2.418274
as.factor(weekday)     57.584059  6       1.401819
as.factor(workingday)  54.813694  1       7.403627
as.factor(weathersit)   1.820212  2       1.161530
temp                   71.485709  1       8.454922
atemp                  60.863898  1       7.801532
hum                     2.062655  1       1.436195
windspeed               1.287078  1       1.134494
```

Interpretation

As per results above we can see that, VIF of Month is the highest and we need to remove it and reframe the model and try again for multicollinearity test.

Testing Model after reframing -

```
model2=lm((count)~as.factor(season)+as.factor(yr)+as.factor(holiday)+as.factor(weekday)+as.factor(workingday)+as.factor(weathersit)+temp+atemp+hum+windspeed,data=traindata)

summary(model2)
```

vif(model2)

> vif(model2)

	GVIF	Df	GVIF^(1/(2*Df))
as.factor(season)	3.644414	3	1.240522
as.factor(yr)	1.029511	1	1.014648
as.factor(holiday)	5.739572	1	2.395740
as.factor(weekday)	54.479725	6	1.395360
as.factor(workingday)	54.033370	1	7.350739
as.factor(weathersit)	1.730363	2	1.146923
temp	61.960385	1	7.871492
atemp	58.196945	1	7.628692
hum	1.818708	1	1.348595
windspeed	1.237428	1	1.112397

Interpretation

As now we can check that temp is showing high value, we will now remove temp from the data as it possesses multicollinearity and reframe and will test again.

model3=lm((count)~as.factor(season)+as.factor(yr)+as.factor(holiday)+as.factor(weekday)+as.factor(workingday)+as.factor(weathersit)+atemp+hum+windspeed,data=traindata)

summary(model3)

vif(model3)

> vif(model3)

	GVIF	Df	GVIF^(1/(2*Df))
as.factor(season)	3.239860	3	1.216432
as.factor(yr)	1.028651	1	1.014224
as.factor(holiday)	5.738252	1	2.395465
as.factor(weekday)	53.997348	6	1.394326
as.factor(workingday)	53.959580	1	7.345718
as.factor(weathersit)	1.725608	2	1.146134
atemp	3.069708	1	1.752058
hum	1.818705	1	1.348594
windspeed	1.193995	1	1.092701

Interpretation

As now we can check that Weekday is showing high value, we will now remove temp from the data as it possesses multicollinearity and reframe and will test again.

model4=lm((count)~as.factor(season)+as.factor(yr)+as.factor(holiday)+as.factor(workingday)+as.factor(weathersit)+atemp+hum+windspeed,data=traindata)

vif(model4)

> vif(model4)

	GVIF	Df	GVIF^(1/(2*Df))
as.factor(season)	3.188163	3	1.213175
as.factor(yr)	1.020178	1	1.010039
as.factor(holiday)	1.077255	1	1.037909
as.factor(workingday)	1.090183	1	1.044118
as.factor(weathersit)	1.690125	2	1.140197
atemp	3.042882	1	1.744386
hum	1.808602	1	1.344843
windspeed	1.188498	1	1.090183

Interpretation

So now all the values are less than 10 we will go for Variable selection for the model.

Variable Selection:

Final MLR Model:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1526.154	236.442	6.455	2.13e-10	***
as.factor(season)2	1221.608	118.051	10.348	< 2e-16	***
as.factor(season)3	984.870	152.017	6.479	1.84e-10	***
as.factor(season)4	1582.053	101.612	15.570	< 2e-16	***
as.factor(year)1	2025.680	64.245	31.530	< 2e-16	***
as.factor(holiday)1	-704.286	212.182	-3.319	0.000953	***
as.factor(workingday)1	-193.644	71.067	-2.725	0.006608	**
as.factor(weather_condition)2	-417.280	84.485	-4.939	1.00e-06	***
as.factor(weather_condition)3	-1855.903	234.856	-7.902	1.19e-14	***
atemp	110.371	6.933	15.920	< 2e-16	***
humidity	-12.368	3.078	-4.018	6.56e-05	***
windspeed	-38.133	6.651	-5.733	1.52e-08	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 811.7 on 645 degrees of freedom

Multiple R-squared: 0.823, Adjusted R-squared: 0.8199

F-statistic: 272.6 on 11 and 645 DF, p-value: < 2.2e-16

Interpretation

From the above we can conclude that all the factors are important predictors in predicting the count of the rental bikes and as we can see p value is way less than 0.05 which means overall model is significant with 5% significance level.