Bike Sharing Assignment – Multiple Linear Regression
SUBMISSION

Anurag Bombarde

Problem Statement

BoomBikes have contracted a consulting company to understand the factors on which the demand for these shared bikes depends. Specifically, they want to understand the factors affecting the demand for these shared bikes in the American market.

The company wants to know:

- Which variables are significant in predicting the demand for shared bikes.
- How well those variables describe the bike demands

Assignment-based Subjective Questions

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer –

We have 2 categorical variables

- Season (1:spring, 2:summer, 3:fall, 4:winter)
- Weathersit (1: Clear,2 :Cloudy,3 : Light Snow or Rain,4 : Heavy Rain)
- Year

Inference –

Season Variable Effect

- From the boxplot of season It is evident that count of the total rental bikes are gradually increasing from Spring , Summer , Fall and then slightly decreasing to winter
- Median of count for Summer , Fall and Winter lies between 4000 to 6000
- During Fall maximum number of people are opting for renting the bike from BoomBikes
- Lowest rent is on Spring

Weathersit Variable Effect

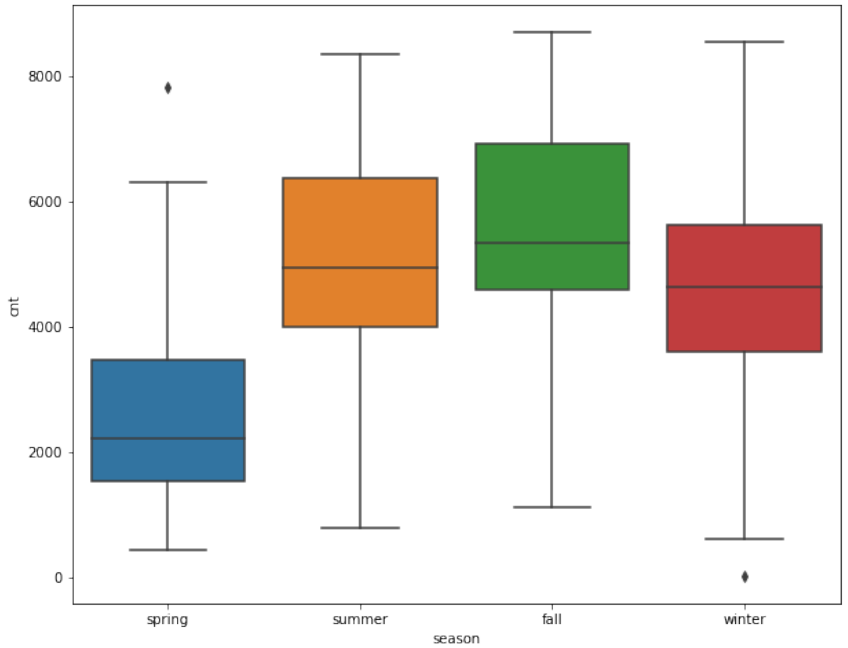
- Median of count for Clear and Cloudy weather lies between 4000 to 6000
- In clear weather maximum number of people are opting for renting the bike from BoomBikes
- Less rent is when weather is with light snow or light rain median is close to 2000
- Zero rent is when heavy rain or heavy snow is there

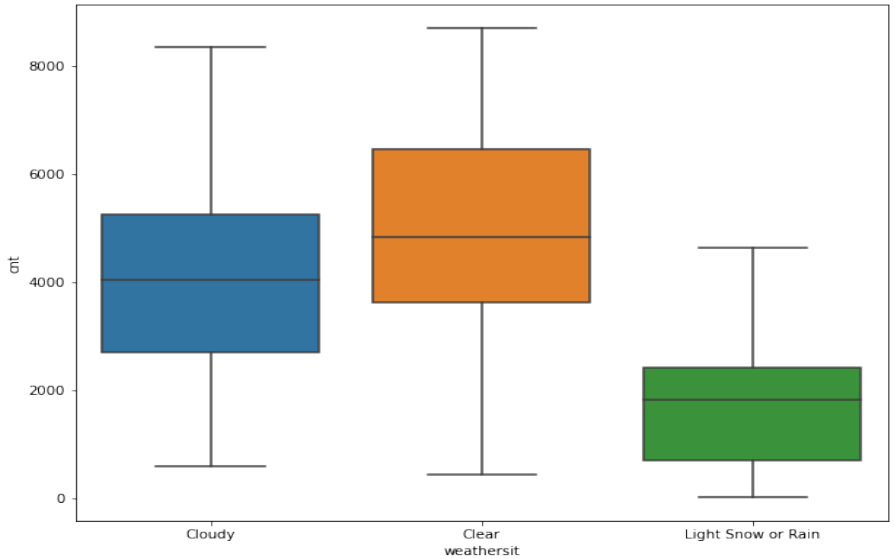
Year

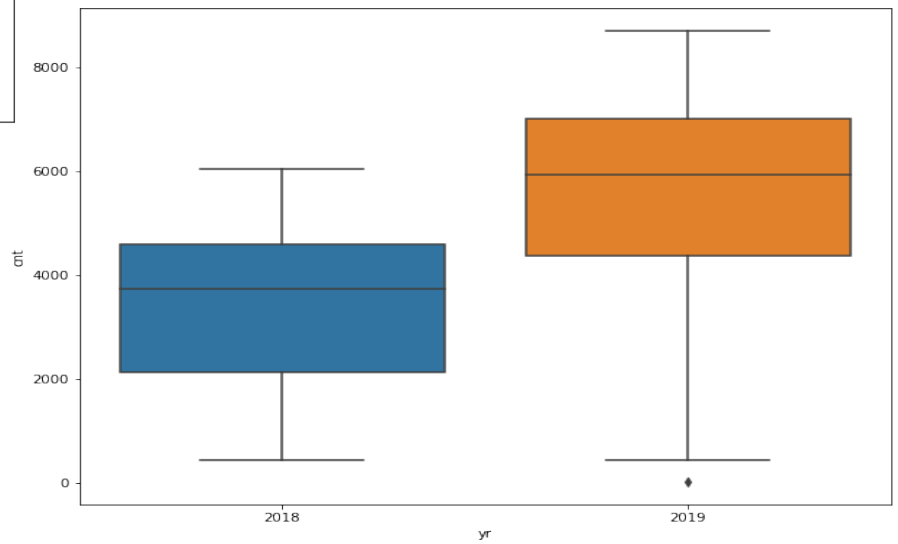
- Boombikes did good business in Year 2019 average was 6000 Rentals and in Year 2018 average was around 5000

In next slide please refer the Boxplots for these categorical variables

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?





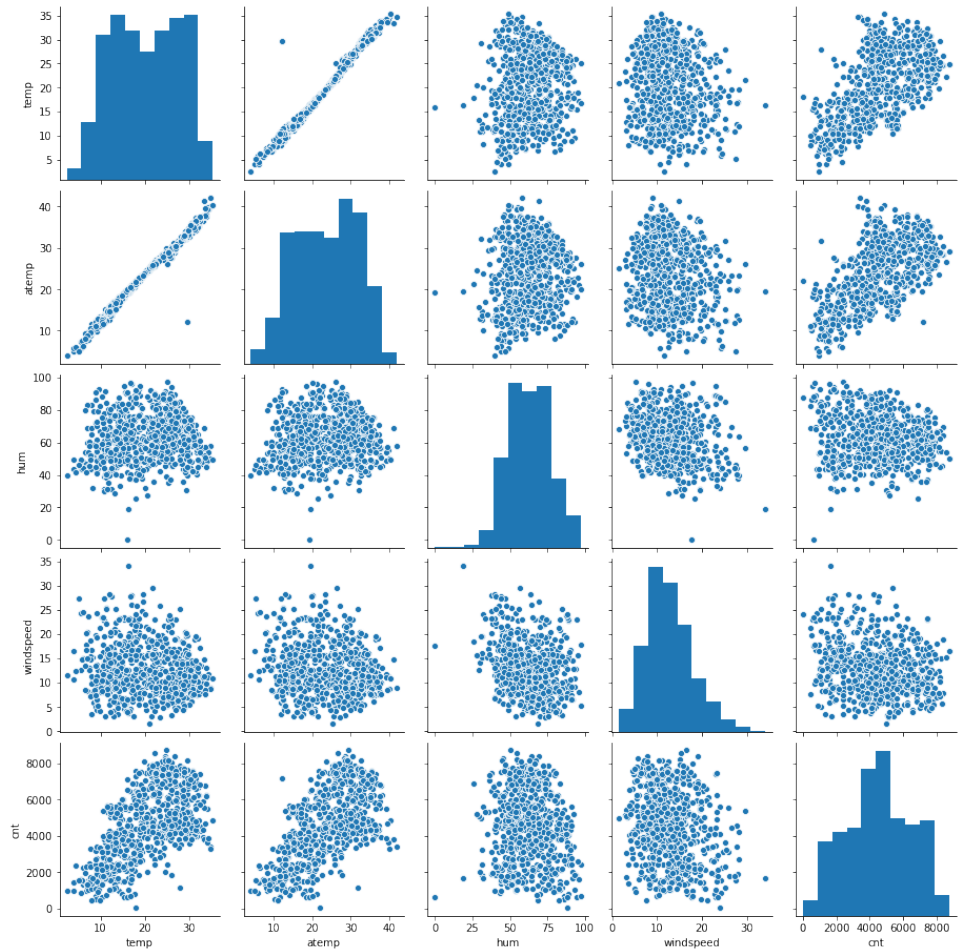


Why is it important to use drop_first=True during dummy variable creation?

- drop_first=True basically drops the extra column which gets created because of the dummy variable creation
- If we do not drop the first column there will be unnecessary increased multicollinearity
- In my case when I didn't do drop_first = True I was getting VIF = INF i.e infinity because of the perfect collinearity with R2 as 1

Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Temperature has the highest correlation
With Target variable



How did you validate the assumptions of Linear Regression after building the model on the training set?

We can validate assumptions of LR by based on following steps

1. by coming to the final trained model based on P Values and Lowest VIF which should have good RSquared and Adjusted R Sqaured values
2. By Referring to the residual analysis and a distribution plot which should be centered at zero and normally distributed
3. Finally evaluating the model with test data by r2_score(y_true=y_test , y_pred=y_test_pred)

    1. For my analysis RSquared value is coming as 83% and adjusted RSquared value is coming as 82.6% which is good for trained model – Please refer slid number 10
    2. For my analysis Distribution plot is centered to zero and normally distributed of residual analysis please refer slide number – 10
    3. In my analysis R2Score for test data is 80.5% which is very close to what train data has given

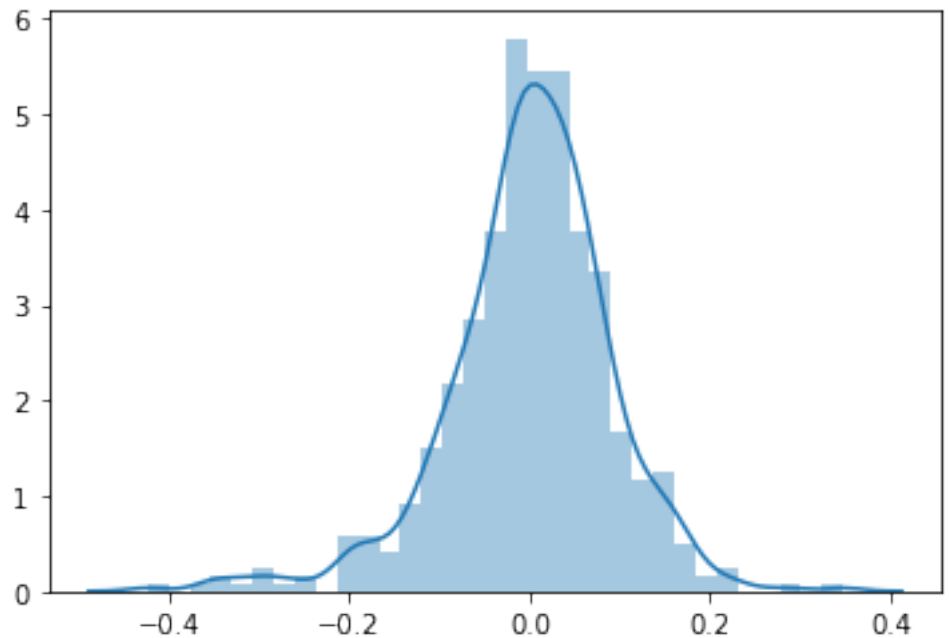
OLS Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1813	0.031	5.887	0.000	0.121	0.242
holiday	-0.0638	0.027	-2.335	0.020	-0.117	-0.010
weekday	0.0087	0.002	4.197	0.000	0.005	0.013
workingday	0.0190	0.009	2.073	0.039	0.001	0.037
temp	0.4669	0.033	13.978	0.000	0.401	0.532
windspeed	-0.1552	0.026	-6.073	0.000	-0.205	-0.105
season_spring	-0.0813	0.020	-3.968	0.000	-0.122	-0.041
season_summer	0.0402	0.014	2.919	0.004	0.013	0.067
season_winter	0.0782	0.017	4.718	0.000	0.046	0.111
yr_2019	0.2351	0.008	28.032	0.000	0.219	0.252
weathersit_Cloudy	-0.0775	0.009	-8.714	0.000	-0.095	-0.060
weathersit_Light Snow or Rain	-0.2828	0.025	-11.244	0.000	-0.332	-0.233

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.830
Model:	OLS	**Adj. R-squared:**	0.826
Method:	Least Squares	**F-statistic:**	220.5
Date:	Sun, 07 Mar 2021	**Prob (F-statistic):**	2.17e-183
Time:	22:23:00	**Log-Likelihood:**	490.20
No. Observations:	510	**AIC:**	-956.4
Df Residuals:	498	**BIC:**	-905.6
Df Model:	11		
Covariance Type:	nonrobust		

Below is the distribution plot coming for my analysis which is centered at zero and normally distributed



Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

As per my final model below are the top 3 features contributing significantly towards explaining the demand of shared bikes

1. Temp with positive Coefficient of 0.4669
2. yr_2019 with positive Coefficient of 0.2359
3. season_winter with positive Coefficient of 0.0782

General Subjective Questions

Explain the linear regression algorithm in detail.

Linear regression a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors).

Two types of linear regression are -

- Simple linear regression
- Multiple linear regression

1. **Simple Linear Regression**

   Explains the relationship between a dependent variable and one independent variable using a straight-line.

2. **Multiple Linear Regression**

   Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

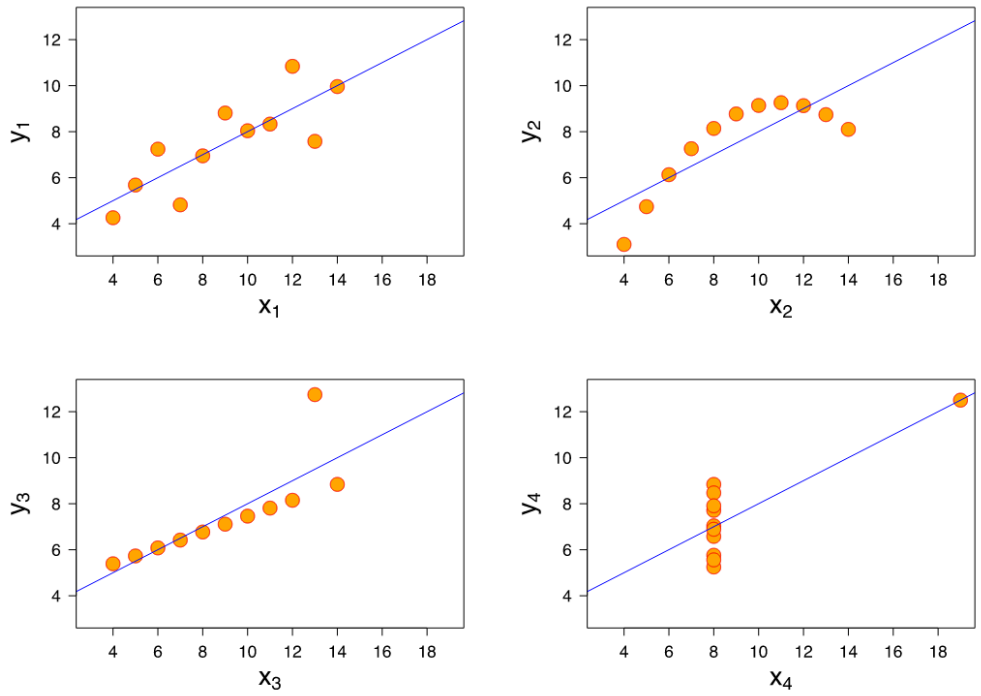
The strength of the linear regression model can be assessed using 2 metrics:

1. R^2 or Coefficient of Determination - R2 is a number which explains what portion of the given data variation is explained by the developed model
2. 2. Residual Standard Error (RSE) -it is defined as the total sum of error across the whole sample

Explain the Anscombe's quartet in detail.

- Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built
- It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties
- Refer These plots -
  - Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance.
  - Dataset II fits a neat curve but doesn't follow a linear relationship (maybe it's quadratic?).
  - Dataset III looks like a tight linear relationship between *x* and *y*, except for one large outlier.
  - Dataset IV looks like *x* remains constant, except for one outlier as well.

Thus, Computing summary statistics or staring at the data may give wrong idea . Hence, it's important to visualize the data to get a clear picture of what's going on



Anscombe's Quartet Plots

What is Pearson's R?

Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r.

Here are the assumptions for Person's R

1. Normal Distribution – Both the variables must be normally distributed (Bell Curve or Gaussian Curve)
2. There should be no significant outliers
3. Each variable should be continuous
4. Both variables should have linear relationship . This can be found using scatter plot
5. Paired observation – For every observation of independent variable there must be corresponding observation for dependent variable
6. Homoscedacity - If the points lie equally on both sides of the line of best fit, then the data is homoscedastic.

What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

What is scaling?

- It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm

Why Scaling is performed -

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret.
So we need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

Difference -

1. Standardized Scaling: The variables are scaled in such a way that their mean is zero and standard deviation is one.
2. MinMax /Normalised Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data. I

You might have observed that sometimes the value of VIF is infinite. Why does this happen?

This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity

What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

What is a Q_Q Plot –

A **Q–Q plot** is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions

Use and Importance- in linear regression -

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions

Below are the possible interpretations for two data sets.

*a) **Similar distribution**: axisIf all point of quantiles lies on or close to straight line at an angle of 45 degree from x -*
*b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles*
***c) X-values < Y-values:** If x-quantiles are lower than the y-quantiles*
*d) **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis*