Boombikes Assignment Submission

Rupesh Kumar V

Assignment-based Subjective Questions

- Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
  - *We can infer that the Categorical variable Season has second most effect on the dependent variable while the holiday has negative impact (considering a day is holiday positive)*

- Q 2. Why is it important to use drop_first=True during dummy variable creation?
- *Beacause there is a redundancy of one level, which comes in a separate column. since one of the combination will be uniquely representing this redundant column*

- Q 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
  - *Based on the pairplot the correlation between Temp and the dependent variable is highest followed by temperature and season*

- Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?
  - *The First data to look for the Rsquared value which came at 0.799*
  - *Second to look for and independent variables which have P Value greater than 0.05 there by making them insignificant and further to drop them to improve the Rsquare on the train data further closer to value of 1*
  - *Finally perform residual analysis on the difference between trained data and test data and getting data that is closer to the P Value found in the Trained data set*

- Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
  - *Based on the final model the three features contributing to high demand is Temperature , season and weekday which has greater say in the demand for any given day based on the data set we have analysed*

General Subjective Questions

- Q1 Explain the linear regression algorithm in detail
  - *A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is.*
    - *The algorithm starts by analyzing the correlation of the data among the various independent variable on the dependent variable*
    - *Estimating the best line fitting and validating the same interms of correctness of the model*
    - *On estimating the line equation its validity is tested by using the method of least square to minimize the residual*
    - *The final step is the test of significance using ftest*

- Q2 Explain the Anscombe's quartet in detail
  - ***Anscombe's Quartet** can be defined when 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. They have very different distributions and **appear differently** when plotted on scatter plots.*

- Q3 What is Pearson's R?
  - It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

Contd…..

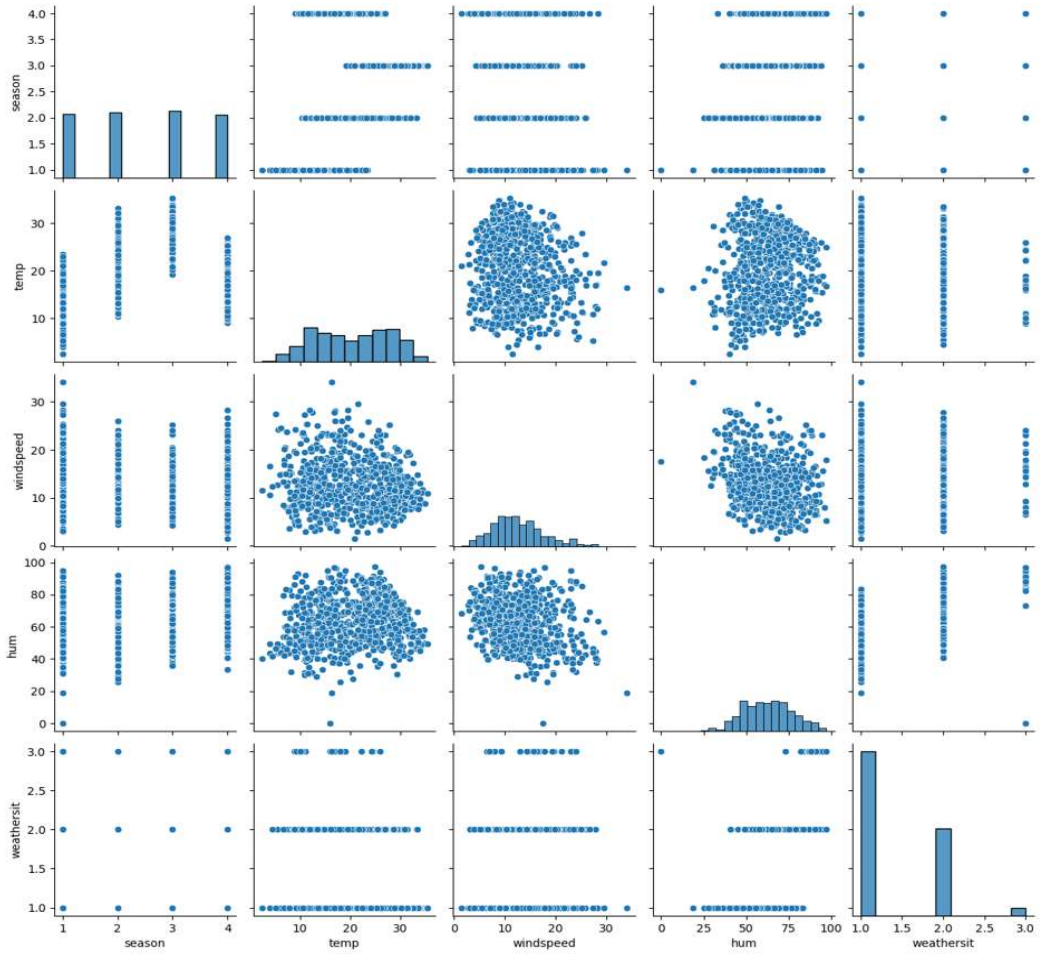
- Q4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
  - *It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range*
  - *Normalization is the method used to arrange the data in a database. It is a scaling method that reduces duplication in which the numbers are scaled and moved between 0 and 1.normalization is employed to remove the undesirable characteristics from the dataset.*
  - *Standardization is subtracting the mean and dividing by the standard deviation of a feature from all of its values. This is usually preferred because it is less outlier sensitive*

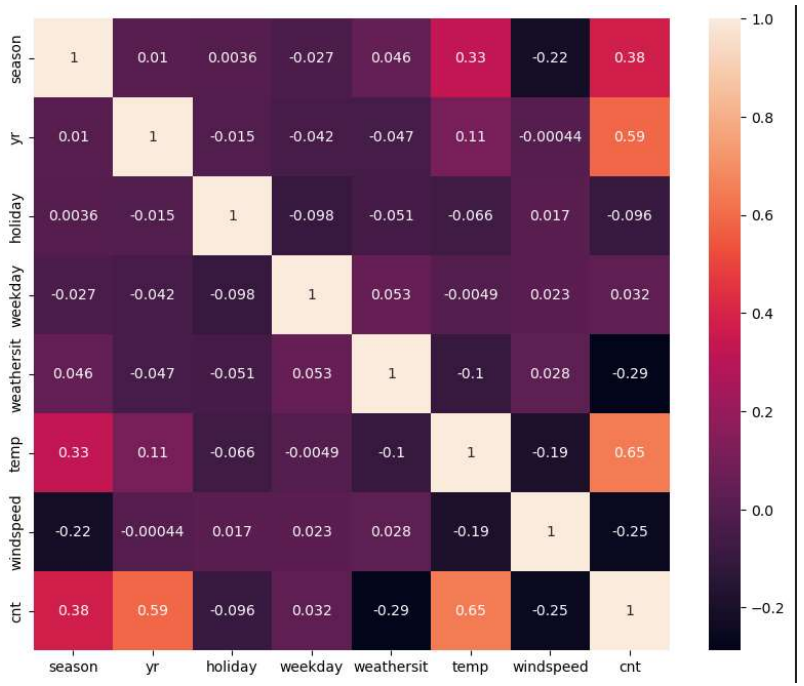
Q5 You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- *An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables*
- *To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity*.

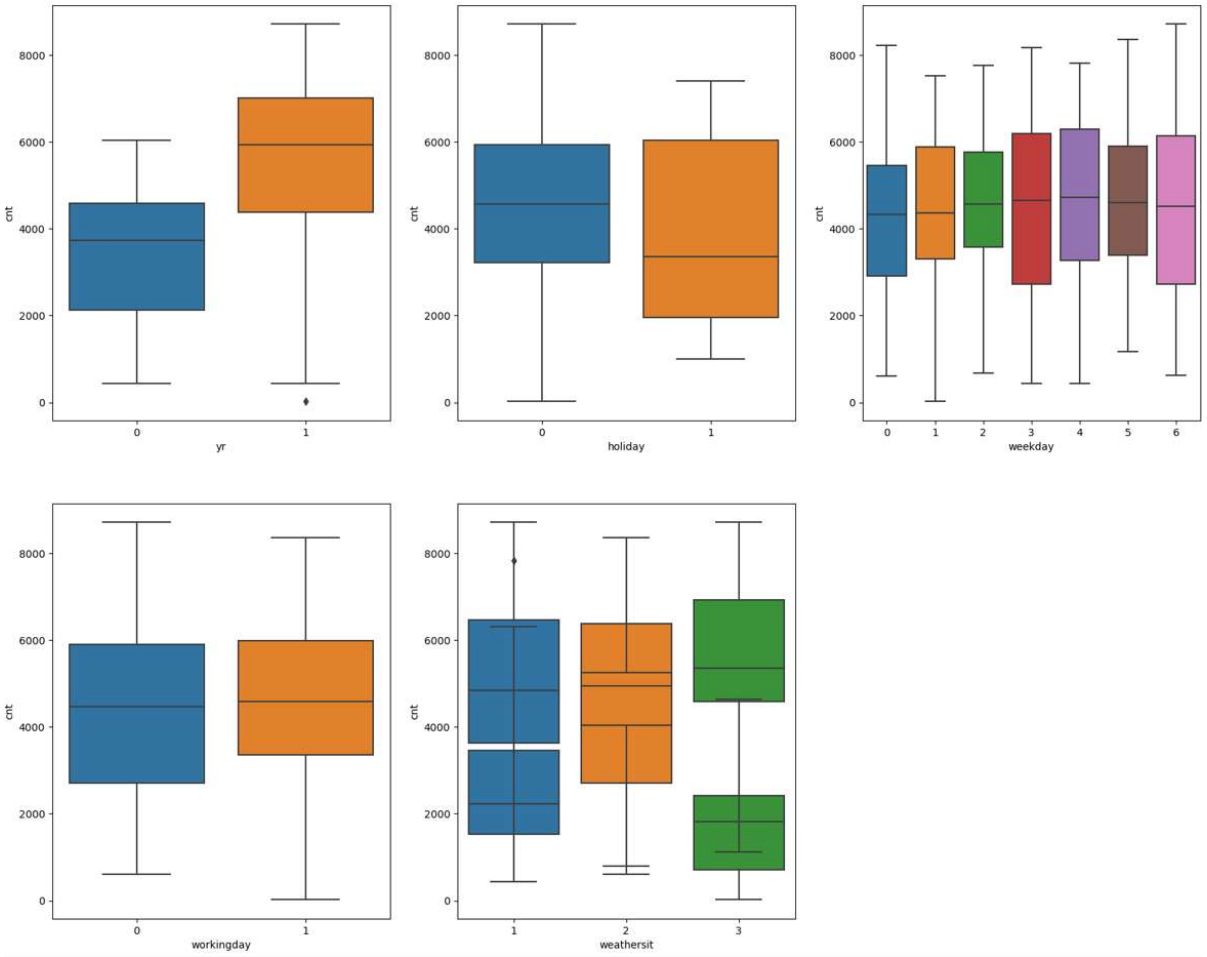
- Q6 What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
  - *The Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. It can be said that it is like plotting quantiles against quantiles.*
  - *A Q-Q plot is **a scatterplot created by plotting two sets of quantiles against one another**. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.*

Pair Plot


season
temp
windspeed
hum
weathersit


season 1 0.01 0.0036 -0.027 0.046 0.33 -0.22 0.38
yr 0.01 1 -0.015 -0.042 -0.047 0.11 -0.00044 0.59
holiday 0.0036 -0.015 1 -0.098 -0.051 -0.066 0.017 -0.096
weekday -0.027 -0.042 -0.098 1 0.053 -0.0049 0.023 0.032
weathersit 0.046 -0.047 -0.051 0.053 1 -0.1 0.028 -0.29
temp 0.33 0.11 -0.066 -0.0049 -0.1 1 -0.19 0.65
windspeed -0.22 -0.00044 0.017 0.023 0.028 -0.19 1 -0.25
cnt 0.38 0.59 -0.096 0.032 -0.29 0.65 -0.25 1
season yr holiday weekday weathersit temp windspeed cnt

Categorical variables



Regression results

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.799
Model:	OLS	**Adj. R-squared:**	0.796
Method:	Least Squares	**F-statistic:**	284.7
Date:	Sun, 12 Mar 2023	**Prob (F-statistic):**	2.90e-170
Time:	11:25:47	**Log-Likelihood:**	-4178.0
No. Observations:	510	**AIC:**	8372.
Df Residuals:	502	**BIC:**	8406.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1255.2621	161.322	7.781	0.000	938.313	1572.212
season	1092.5883	113.917	9.591	0.000	868.776	1316.401
yr	2063.9596	78.660	26.239	0.000	1909.417	2218.502
holiday	-743.4989	249.813	-2.976	0.003	-1234.307	-252.690
weekday	414.0140	117.040	3.537	0.000	184.065	643.963
weathersit	-1625.4228	144.353	-11.260	0.000	-1909.033	-1341.813
temp	4075.5018	187.482	21.738	0.000	3707.155	4443.849
windspeed	-1305.2356	237.685	-5.491	0.000	-1772.216	-838.255

Omnibus:	59.411	**Durbin-Watson:**	1.984
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	121.908
Skew:	-0.665	**Prob(JB):**	3.37e-27
Kurtosis:	4.992	**Cond. No.**	10.6

	features	VIF
5	temp	5.33
6	windspeed	3.17
0	season	3.14
3	weekday	2.83
1	yr	1.98
4	weathersit	1.52
2	holiday	1.03

Thank You!