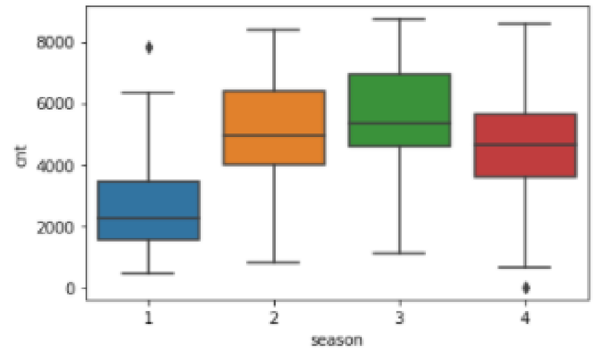
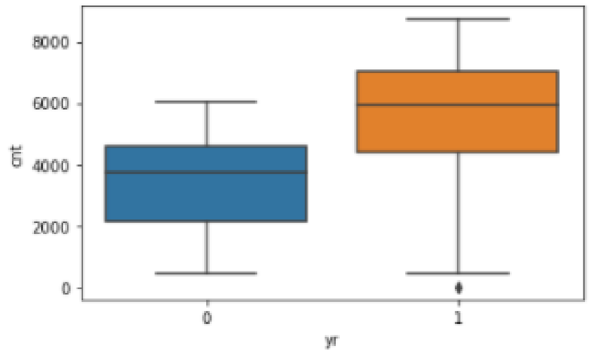
Neeraj Kumar Shah (MLC34) Date: 12-Jan-2022

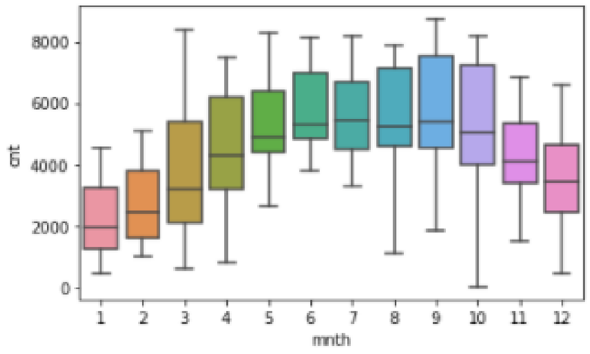
Assignment-based Subjective Questions

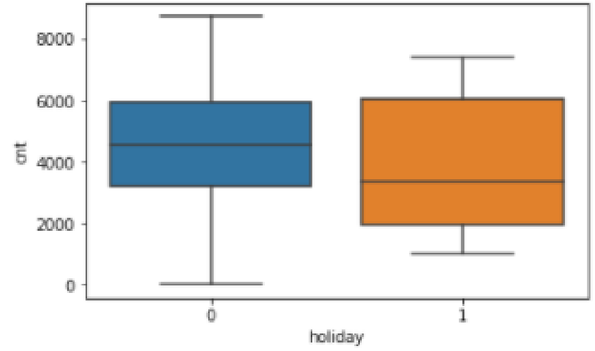
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

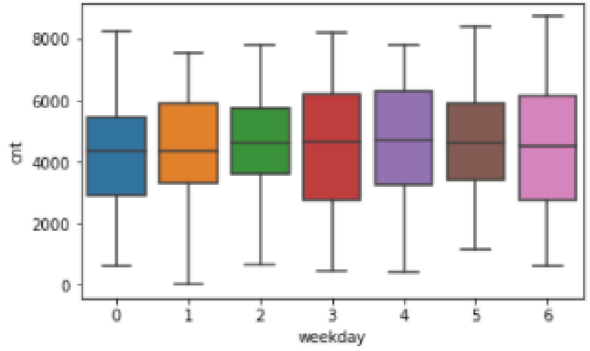
Ans1. Please find below box plots of categorical variables with the target variable cnt.

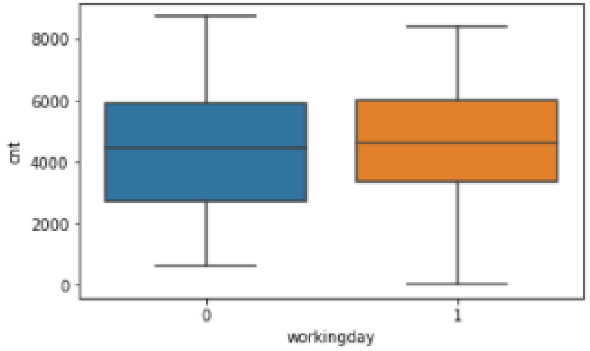


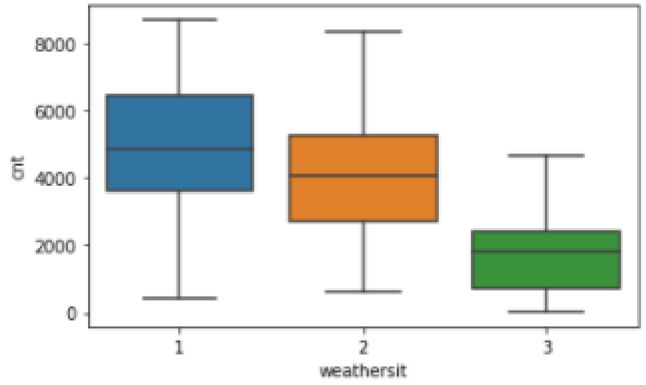












a. There is high demand of bikes in the fall season and lowest demand in the spring season. The same demand is expressed by the month variable that is the demand of bikes is high in the 3rd quarter and lowest in the first quarter. This similar relation of month/quarter and season is due to their similar date time frame.
b. The demand of bikes is increasing significantly in the next year.
c. The holidays and weekends have low effect in the change in demand of bikes.
d. There is high demand of bikes in the clear weather and lowest demand in the bad weather.

---

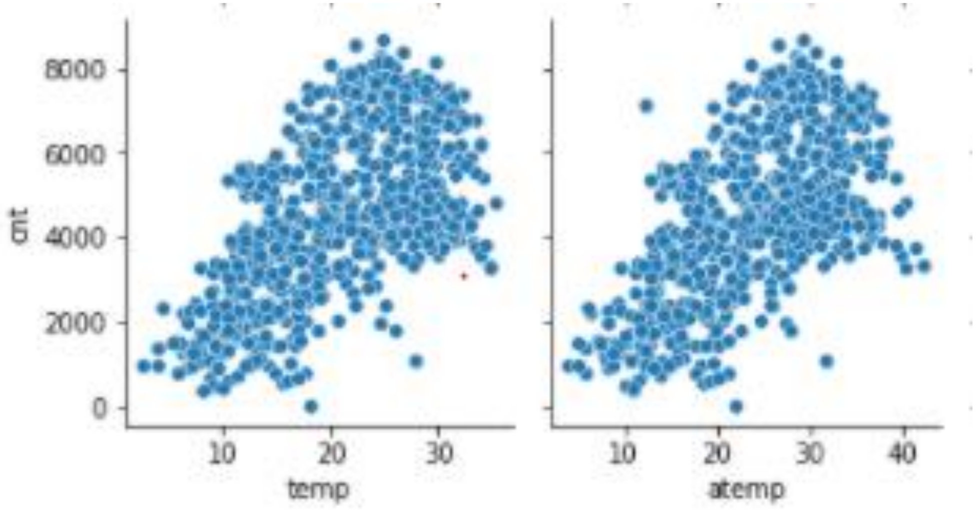
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans2. The first the column is dropped using the above method while creating dummy variables since we need only (n-1) variables to define/express the various levels (n) of the categorical variable.

---

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans3. Temp/aTemp has the highest correaltion with the target variable cnt. Please find below pairplot screenshot of temp/atemp.

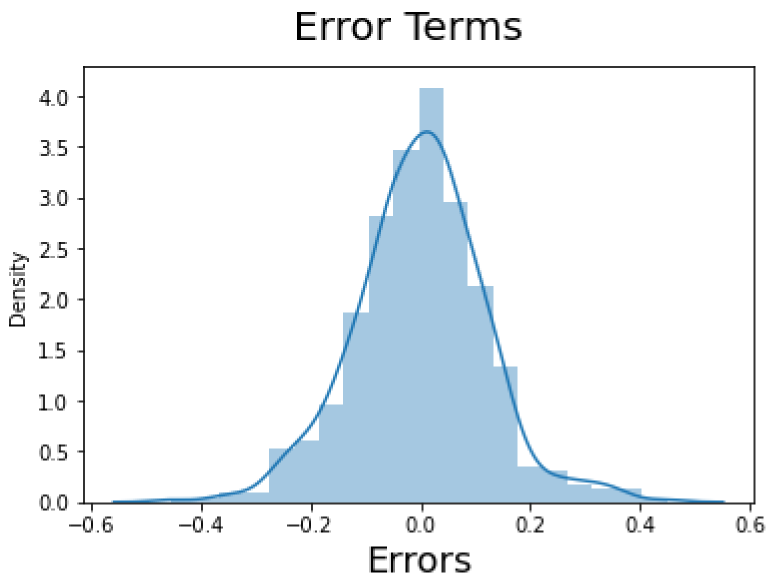


---

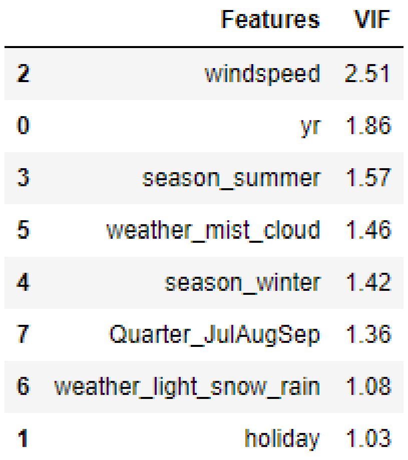
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans4. Please find following for the validation of assumptions of Linear Regression.

a. The error terms are normally distributed and verified by the scatter plot of output predicted (y_pred) and actual output (y_test).



b. The VIF for the predictors are less than 5 and within permissible range and thus validating that there is no multicollinearity between the predictors.



	Features	VIF
2	windspeed	2.51
0	yr	1.86
3	season_summer	1.57
5	weather_mist_cloud	1.46
4	season_winter	1.42
7	Quarter_JulAugSep	1.36
6	weather_light_snow_rain	1.08
1	holiday	1.03

c. The error terms does not follow any pattern.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans5. Thee top three features are windspeed, year & season.

____

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans1. The linear regression algorithm is used to predict a continuous variable based on its linear relation with the predictor variables. The linear regression model is expressed by a best fit line that describes the most appropriate linearity between the target variable and the predictor variable by finding the least sum of residual squares using ordinary least squared method. The residuals are the difference between the predicted values and actual values of the target variable.

____

2. Explain the Anscombe’s quartet in detail. (3 marks)

Ans2. Anscombe's quartet contains four data sets that have almost same statistics parameters such as different quantiles, mean, mode, etc,however when they are plotted, they describe a very different performance parameters.

____

3. What is Pearson’s R? (3 marks)

Ans3. The Pearson’s R is the most commonly used correlation coefficients in the linear regression model. It ranges from -1 to +1, where values close to -1 & 1 represents very high negative and positive collinearity and the values close 0 represents very low collinearity.

____

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans4. We scale the data to express all the variables on the same scale. Scaling helps us the achieve representable beta coefficients of the predictors. The normalized scaling is where the variables are scaled between 0 (min) to 1 (max) and the standardized scaling is where the mean is 0 and the standard deviation is 1 for the variables.

____

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans5. The VIF values is infinite when there is perfect correlation between the independent variables.

____

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans6. A Q-Q plot is a graphical representation of the Quantiles of the values of 2 datasets. It is used to analyse the if two data sets have a common distribution and scale.

____