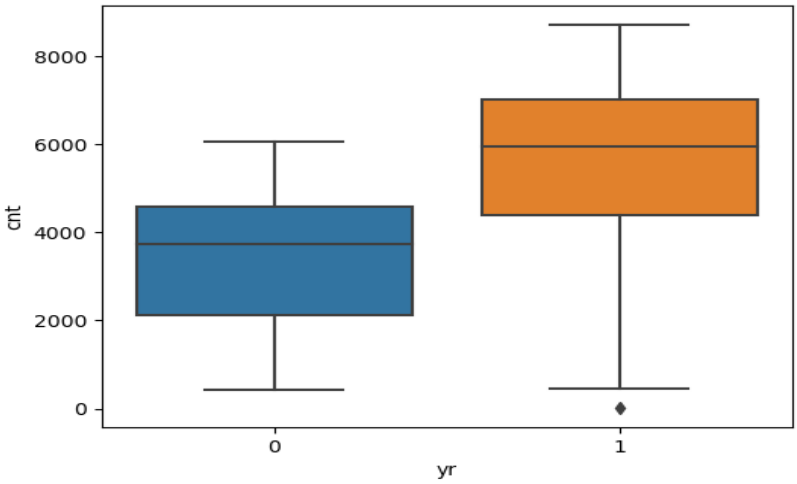
Assignment-based Subjective Questions:

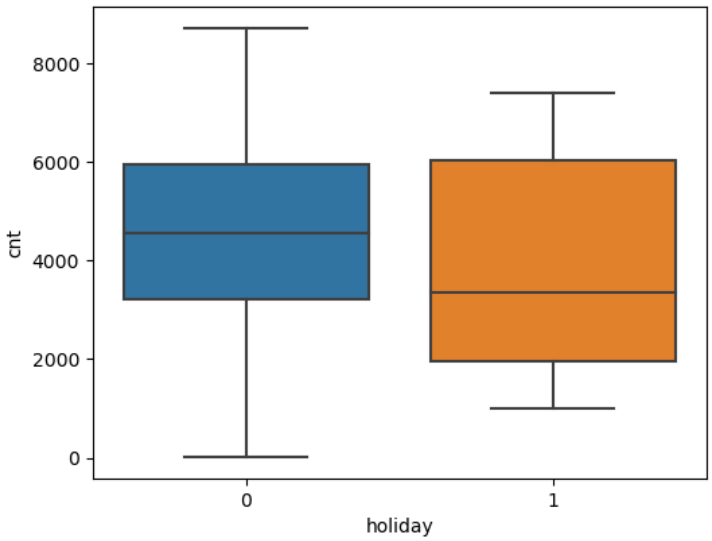
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

The given categorical variables in the data are:

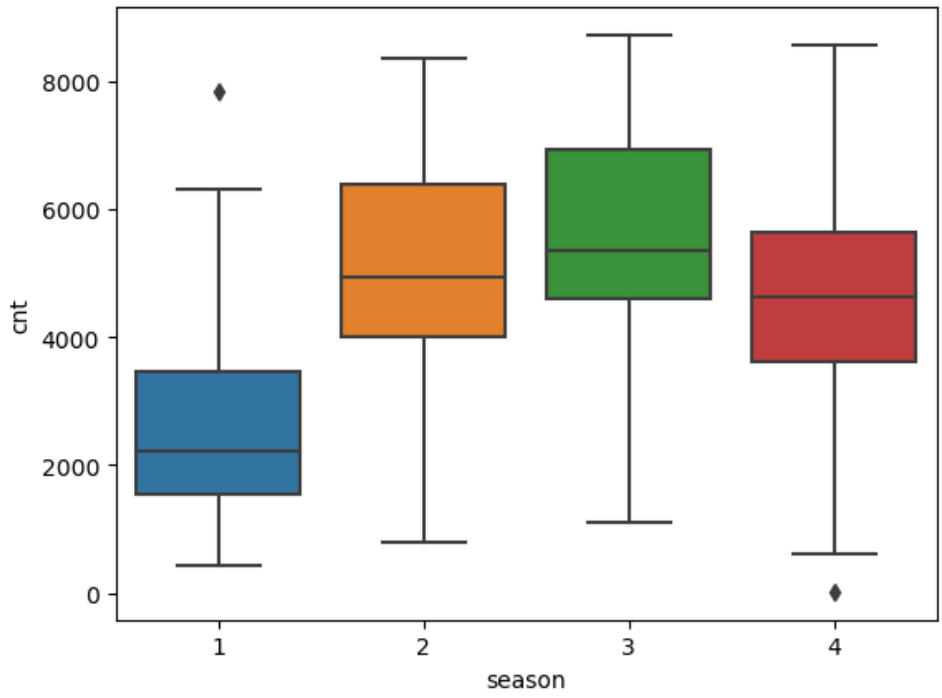
a. Year: The demand is higher in 2019 compared to 2018.



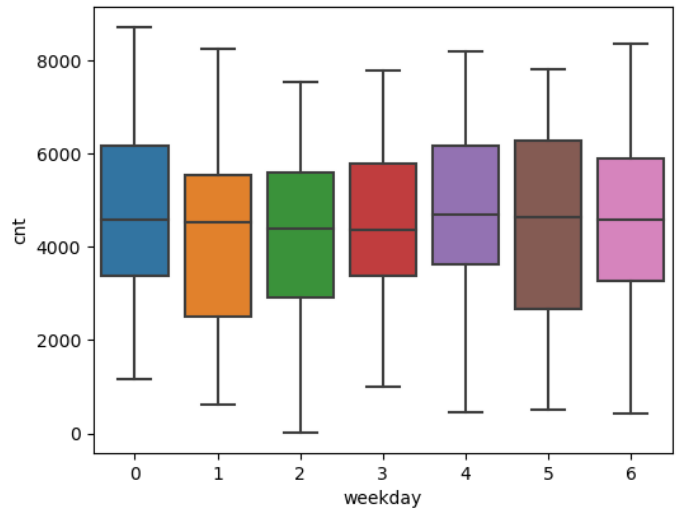
b. Holiday: It can be noted that on the spread of demand was more on the working days than holidays as people may use these bikes as a public transport



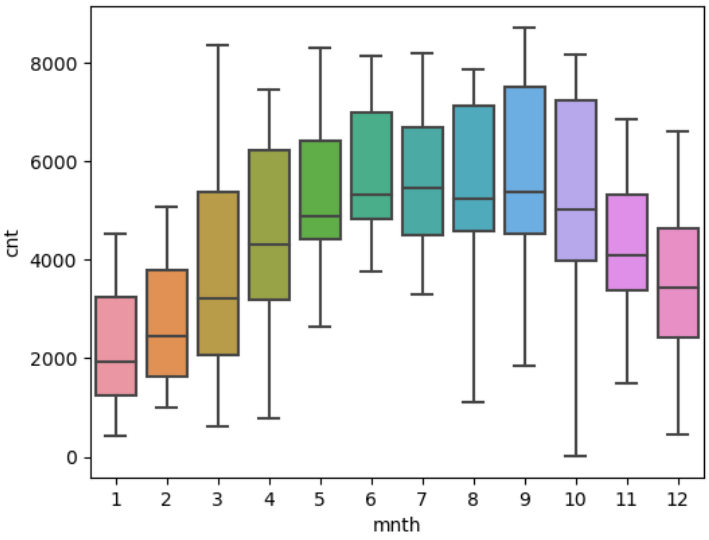
c. Season: The median demand is greater during seasons 2 and 3, which are summer and fall.



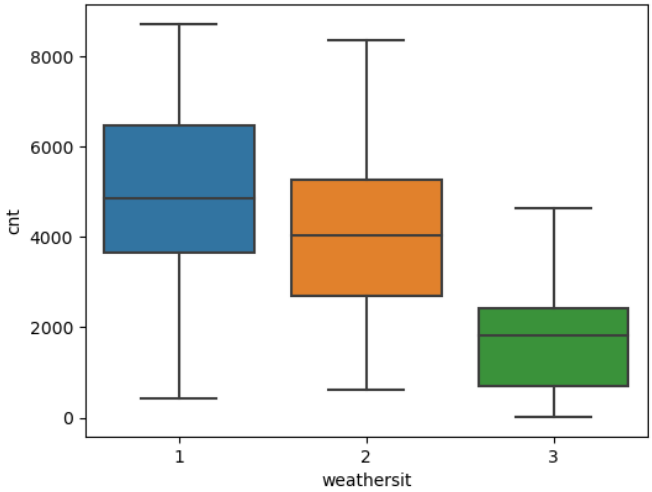
d. Weekday: The median demand was almost same during all days of the week, while the maximum demand can be observed on Sunday and minimum on Tuesday.



e. Month: The demand gradually increased from January till June and started declining from September. This is like the season's box plot.



f. Weathersit: The median demand was high for weather type 1(Clear, Few clouds, partly cloudy) and the least for Weather type 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)



2) Why is it important to use drop first=True during dummy variable creation?

The reason behind creation of dummy variables is to convert the non-binary categorical variables into binary variables. If there is a non-binary categorical variable having k-values, it is enough to have k-1 dummy variables each having value of either 0 or 1. If all given dummy variables have the value 0 then it indicates the value for which dummy variable is not created.

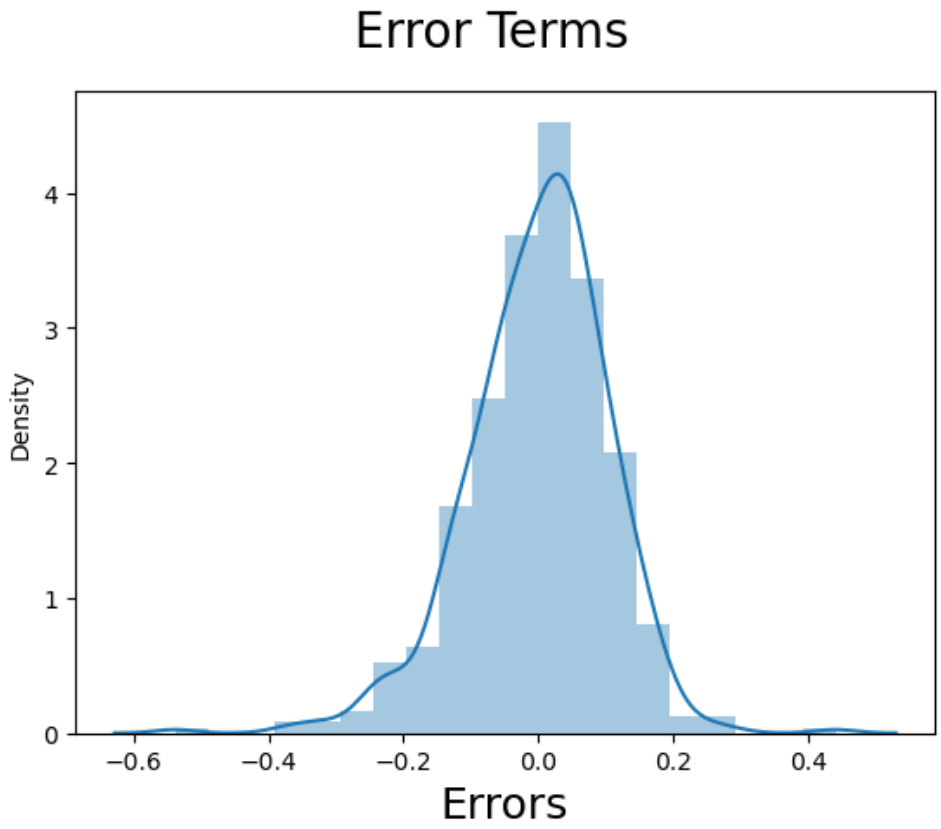
So, while creating dummy variables in pandas it is essential to use drop_first = True to get k-1 dummy variables.

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

Here “cnt” representing demand is the target variable. Among all numeric variables, temperature has the highest correlation, I.e., 0.627044. This is followed by humidity and windspeed.

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

After building the model on training set, we ensure that the error terms obtained are in normal distribution, having 0 as mean by plotting a histogram.



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

The final model equation obtained will be as follows:

demand = 0.5819 + (0.2479)*year + (-0.0946)*holiday + (-0.1898)*windspeed + (-0.2551)*spring + (-0.0382)*summer + (-0.1039)*winter + (-0.1021)*Jan + (0.0941)*Oct + (0.0845)*Sep + (-0.0883)*Weather_2 + (-0.3142)*Weather_3

From this equation, the top 3 features are year, followed by Sep(indicating if the month is September or not) and Oct.

General Subjective Questions:

1) Explain the linear regression algorithm in detail

Linear regression is a predictive algorithm used for predicting the value of one dependent variable based on the values of one or more independent variables. The regression algorithm will be as follows:

1) First, we would read and understand the data. In the process, we visualize the data, check for outliers, null values, missing values etc. And make appropriate changes.
2) Data preparation:
    - We convert the categorical variables to numeric variables.
    - We convert categorical non-binary variables to binary variables using the concept of dummy variables.

3)

After these steps of data preparation, we split the data into training and testing data sets. In general, 70% of the data is converted to training set. We also rescale the features essential to reduce the error in calculating p-values, r square values of the models.

4)

We then build a linear model either using Bottom top approach where we start model with one variable and slowly build on it. In Top-Bottom approach we build the model by taking all variables and eliminate the variables which are unnecessary.

5)

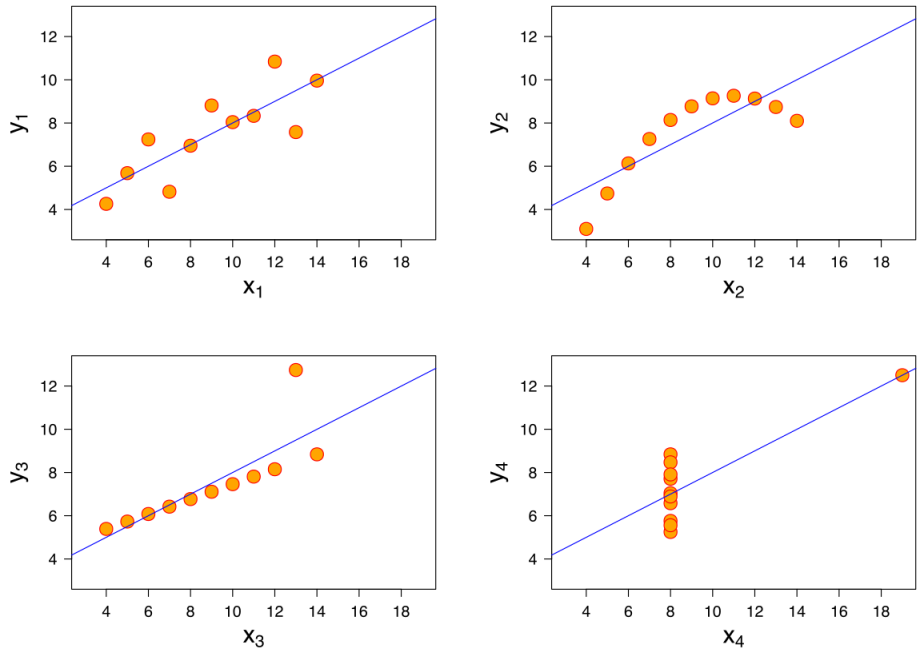
We then check the assumptions of liner regression by plotting the histogram of difference between train data and predicted data. We check for distribution to be normal having mean 0 and uniform variance.

6)

We then start making predictions on the test data using the model we made above. We then check the r square value of the model on the test set and compare it with train set.

2) Explain the Anscombe's quartet in detail.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Anscombe Quartet is a case comparing 4 sets of data having identical statistical parameters I.e., all four of the data sets have same values of mean, median, mode, standard deviation. But on scatter plotting each set of values we observe different correlation between the two variables. The main takeaway is that visualizing the data is as important as calculating the statistical parameters.

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

Scaling is the process of standardizing the values of the features in a dataset. The main purpose is for the ease of interpretation of data, and faster convenience while applying gradient descent methods during model evaluation.

Methods for scaling:

1) Standardization: This method converts all the data into normal distribution with mean 0 and standard deviation of 1. The formula is

   X = (x-mean(x))/ (standard deviation of x)

2) Min Max Scaling: This method converts all the data into the range of 0 and 1. The formula is

   X = x-(minimum of x)/(max(x) - min(x))

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

VIF stands for Variance inflation factor. It calculates how well one independent variable is explained by all other independent variables combined. It is calculated by

VIF = 1/ (1-R_Square)

Here R_Square is the value of model consisting of the variables. If VIF is infinite, it means that R_Square is 1. This implies that the given variable is completely explained by the remaining variables considered.