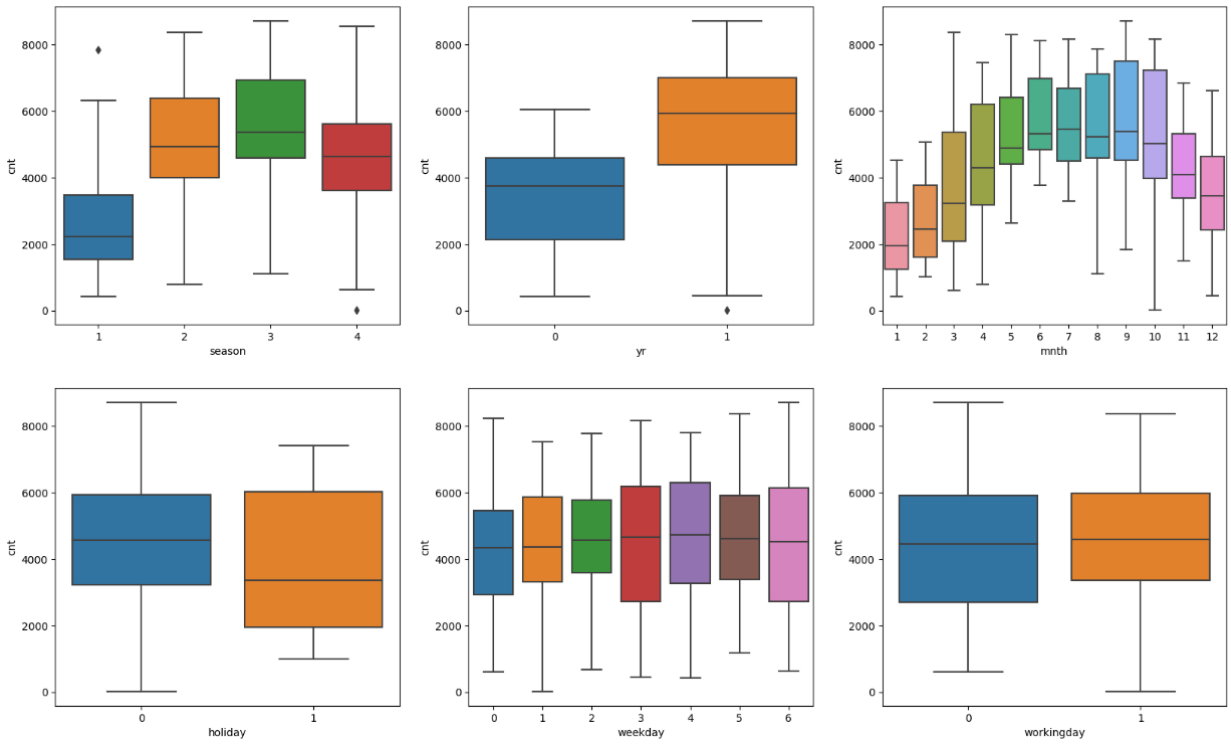
Assignment-based Subjective Questions

Q.No.1 - From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans.1 Plotting categorial variable with respect to cnt (demand) variable generated the following box plots:-



Analysis

- The demand for bike is more during Cloudy/misty(category 2) and light(category 3) rainy weather compared to heavy rain(category 4) and least during clear/partly cloud(category 1) weather.
- There has been increase in demand of bikes over time i.e. there was significantly more demand during the year 2019 compared to the year 2018
- There is more variation in demand of bikes(more min-max difference) during Holidays compared to non-holidays with on an average less demand of bikes during holidays compared to non-holidays.
- Month of Jan has lowest demand of bikes and September being the highest.

Q.No 2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans 2. It is important to use drop_first=true during dummy variable creation because n-1 columns are sufficient to understand n values on any categorical clumn. The first columns just add redundancy to the data which can be removed without introducing any impact on the model.

Q.No. 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. 3. The column temp/atemp has the highest correlation with the target variable

Q.No. 4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans. 4. I validated the assumption of linear regression by looking at the summary of the model. The R-squared value is .84 and Adj. R Squared is .839 which signifies that around 84% of the demand of the shared bike can be explained using the selected variables for training the model. The residual analysis i.e. plotting of error terms of both training and test data shows a normal distribution with the mean centered on 0.

Q.No. 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans. 5. The top three features contributing significantly towards explaining the demand of the shared bikes are:-

atemp, year and season

General Subjective Questions

Q.No 1. Explain the linear regression algorithm in detail.

Ans. 1. When we want to predict an outcome (output variable) which is dependent upon a set of independent variables, the statistical model which helps to analyze this relationship is know a Linear regression algorithm. The model helps to calculate the linear relationship i.e. when value of an independent variable changes then how much value of the dependent variable would change in respect to the independent variable.

Mathematically a simple linear regression(one independent variable) can be represented as

$Y = B0 + B1X$

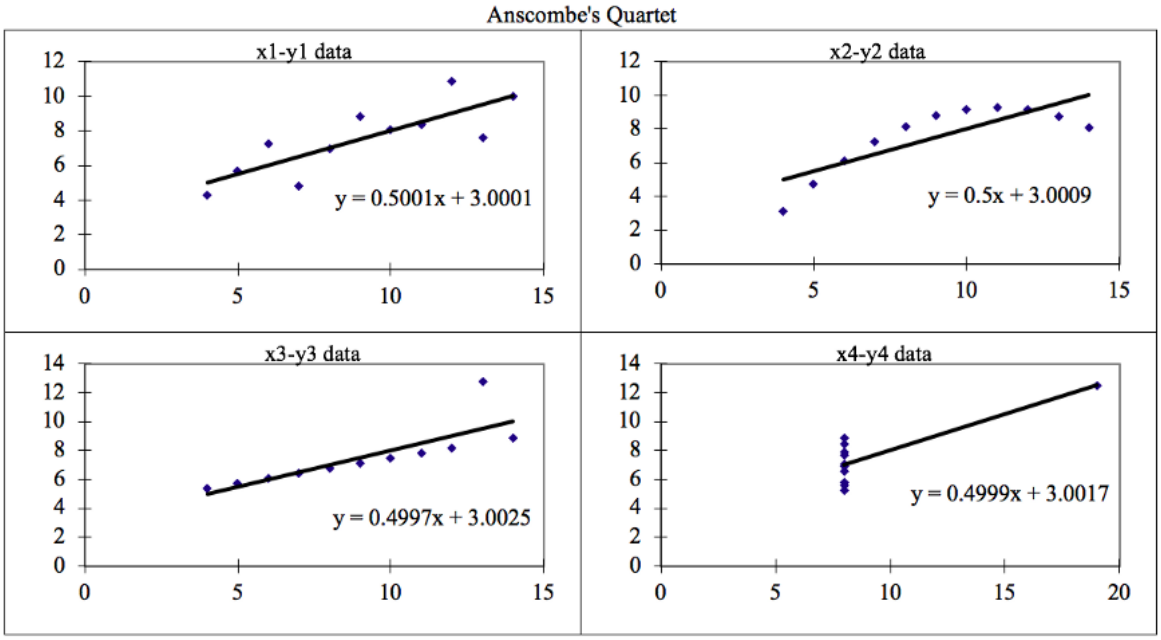
Where B0 is known as the intercept and B1 is known as the slope of the regression line.

For multiple independent variables i.e. relationship between two or more independent input variables and a response variable can be represented as:-

$Y=\beta0+\beta1X1+\beta2X2+...+\beta pXp9$

Q.No. 2 Explain the Anscombe's quartet in detail.

Ans. **Anscombe's quartet** comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties



The four datasets can be described as:

1. **Dataset 1:** this **fits** the linear regression model pretty well.
2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
4. **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

Q.No 3 What is Pearson's R?

Ans.- Pearson's R is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be

positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. It is calculated as

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

Q.No. 4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. Scaling is a step performed on independent variables before fitting/training the model so that the variable values normalize between a defined range. Scaling is performed when we have 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. We need to scale features because of two reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent methods

In standardized scaling the variables are scaled in such a way that their mean is zero and standard deviation is one whereas in normalized 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

Standardized Scaling:- $x = x - \text{mean}(x)/\text{sd}(x)$

Normalized Scaling:- $x = x - \min(x)/\max(x) - \min(x)$

Q.No. 5 You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. VIF represents the correlation between the variables and a very high value (generally > 5) is considered a very high correlation due to presence of multicollinearity. If there is a perfect correlation then the value of VIF tends to be infinite. In this case its recommended to drop the variable from calculations.

Q.No. 6 What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A.No. Q-Q plot is also know as quantile-quantile plot is a probability plot. The purpose of the quantile-quantile (QQ) plot is to show if two data sets come from the same distribution. Plotting the first data set's quantiles along the x-axis and plotting the second data set's quantiles along the y-axis is how the plot is constructed.

In linear regression Q-Q plot helps when we the training data set and test data set is received from two separate sources, then Q-Q plot helps us to determine if both the data sets are from the same distributions.