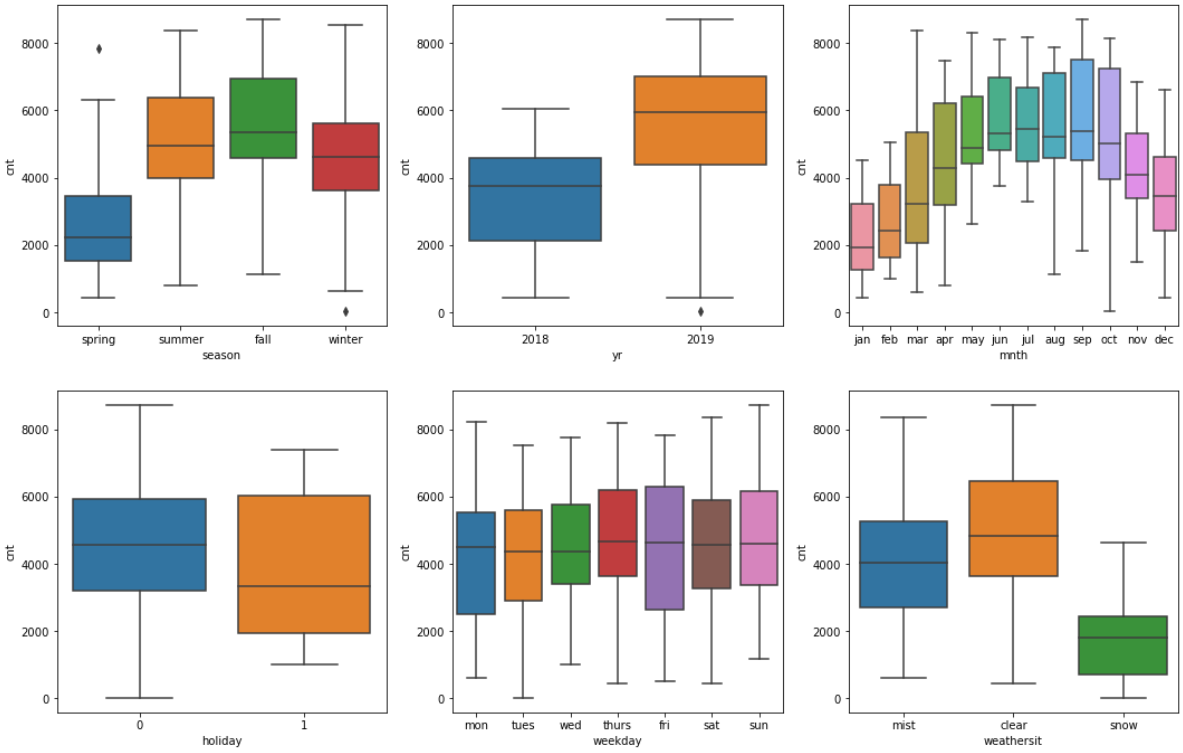
Assignment-based Subjective Questions

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



A: We can infer the following effects from our analysis of the categorical variables:

- Seasons have some effect on the demand of bikes as we can see the fall have highest median comparing to spring which is having lowest.
- Similarly, we can observe from month variable how season is showing its effect. Thus, we can assume that in the months starting from May usually students have vacations in school & colleges which is affecting the demand in a positive way.
- From Year variable we can clearly observed that median came to 6000 in 2019 comparing to 2018 which is having median below 4000.
- In working and non-working, we can see a large spread for non-working days.
- Lastly, the weather being an important aspect, we can observe demand increases on clear weather.

Q-2: Why is it important to use drop_first=True during dummy variable creation?

A: As we create dummy values the variable from which it is extracted will no longer be required and can tend to increase confusion later by having extra column. Thus, to decrease the correlation between the dummy variables we eliminate it.

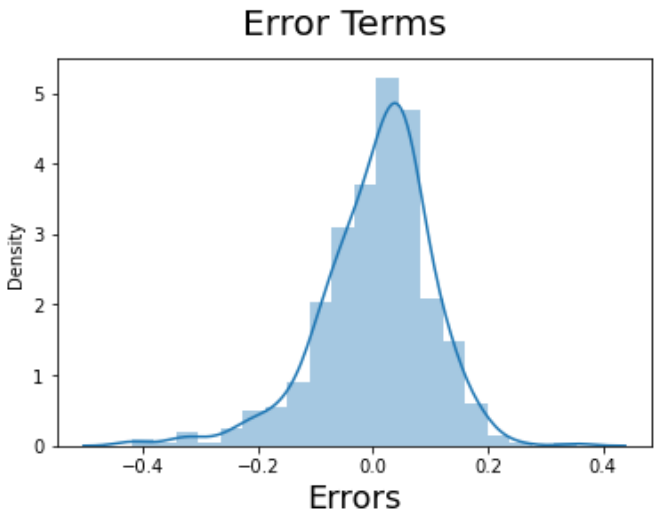
For example: When we created dummy variables in weather where we have 4 values 1-clear, 2-mist, 3-snow, 4-rainy. For representing a clear day all values of other variables will be 0. So even without the first column we will be able to identify the value as clear weather.

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

A: We can observe that temp & atemp is having the highest correlation of 0.63 with the target variable.

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

	Features	VIF
2	temp	2.85
0	yr	1.94
3	season_summer	1.52
6	weathersit_mist	1.44
4	season_winter	1.34
5	mnth_sep	1.19
7	weathersit_snow	1.06
1	holiday	1.03



A: As observed in our final model of training,

- atemp & temp showed high VIF values. But as temp is an important variable we dropped atemp.
- Along with this, humidity which is directly connected to seasons affected the VIF.
- In the end we had our VIF value under 5 which is essential in linear regression.
- When we plotted the error terms we can visualize the normal distribution among them. Thus, validating our assumptions
- Lastly the r2 score metric of y_test and y_testpred is equivalent to 0.79.

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

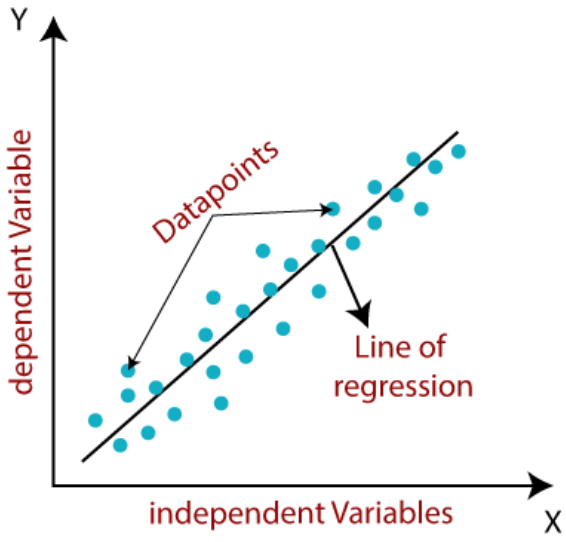
A: The top three features are

- Temp: being the most important feature as we visualized how it affected with coefficient 0.64
- Years: As people are getting more aware about the bikes by marketing or any means with coefficient of 0.24
- weathersit_mist with coefficient -0.06 and weathersit_snow with coefficient -0.28.

General Subjective Questions

Q-1: Explain the linear regression algorithm in detail.

A: Linear regression is an algorithm which comes under supervised learning. It is used to predict the value of the dependent variable. It is based on supervised learning where we perform a regression task using independent variables.



As shown in the figure, our dependent variable(y)'s value is predicted based on independent variables(x). Here, as we find a linear relation and generate the line of regression which is the best fit line. By using the cost function we minimizing the error between predicted value of y and the original value of y. We use Root Mean Square Error for predicting y. To reach the minimized cost Gradient Descent is used to update theta1 and theta2 till it draws the best fit line.

Q-2: Explain the Anscombe's quartet in detail.

A: In Anscombe's quartet, there are basically four datasets which are similar in the manner of descriptive statistics i.e. the mean, standard deviation, etc. yet,

when we visualize them they show completely different patterns. This property of variables shows us how important it is to visualize our data before building the model. As for linear regression model it is essential for the data to be in linear relation. Thus, we should always proceed after data visualization.

Q-3: What is Pearson's R?

A: The Pearson's R also known as bivariate correlation or Pearson Correlation Coefficient (PCC) is a term used to measure linear correlation between two datasets. We can calculate it by firstly finding the covariance of two variables and secondly dividing them by the product of their standard deviations so the measure of the covariance is normalized between -1 and 1.

- r = 0 represents no linear relation
- r = 1 represents linear data and a positive slope
- r = -1 represents linear data but a negative slope

Q-4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A: In general, scaling means to transform or convert data into a specific range of values. It can 0 to 1, 0 to 100 or changing millimetre to meter, thousands to hundreds, dollars to rupees etc. Thus, by applying scaling to data helps us to equally compare different variables. It is essential to apply scaling for reducing the distance between data points.

- In normalized scaling, we define a range in which a data is to be scaled. Here, outliers can be issue so it is not recommended for data like salaries of all employees where the salary of CEO can be an outlier but we cannot remove it.
- Standardisation scaling, also known as Z-score in this scaling there is no requirement for range. It is less affected by outliers as we use mean and standard deviation to transform data into mean vector.

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

A: VIF (Variance inflation factor) represents the correlation between independent variables in regression. Thus, when there is a correlation equal to 1 or -1 which is a perfect positive and negative correlation the VIF goes to infinite. For example when we have perfect correlation the R2 will be equal to 1. Furthermore, the value of VIF will become 1/(1-R2).

Q-6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A: The Q-Q plots or Quantile-Quantile Plots are used to graphically compare two quantiles by plotting them against each other for visualizing the distribution.

They are mainly used for finding the type of distribution of variables. We can find if tow datasets belong to the same distribution by plotting Q-Q. For example if datasets have common distribution the angle of reference line would be 45 degrees.