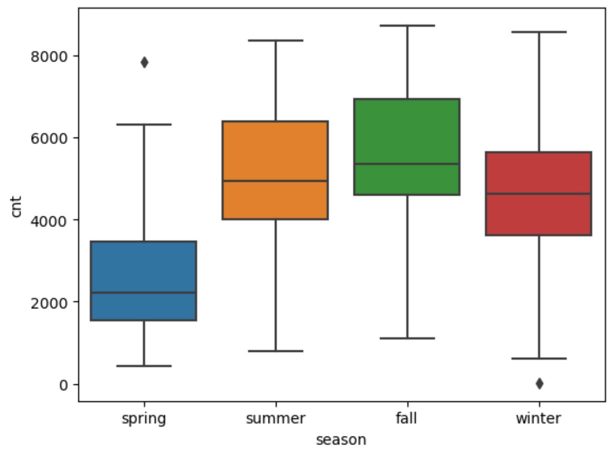
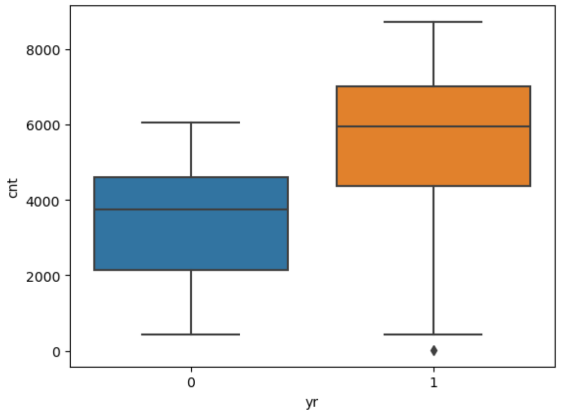
Assignment-based Subjective Questions

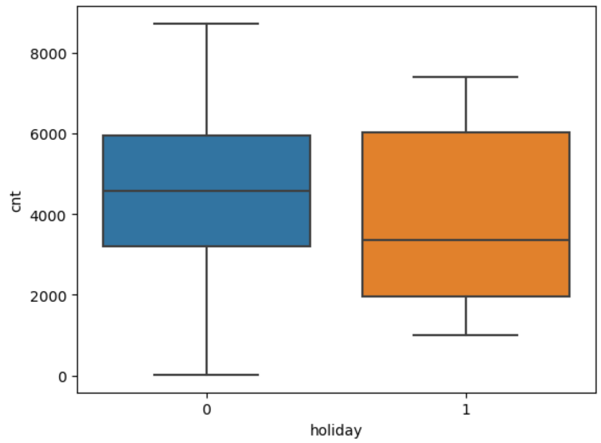
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

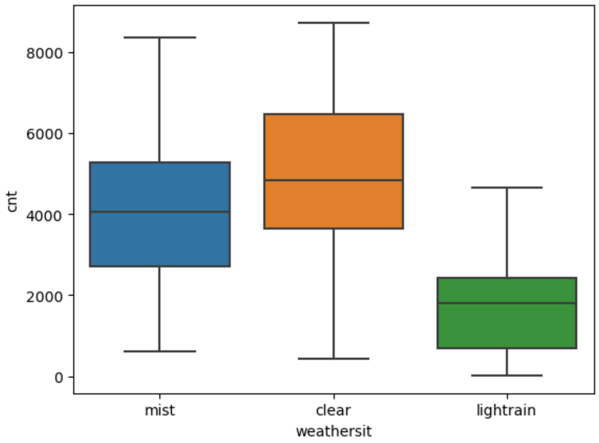
The categorical variables analyzed from the dataset are - 'season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday', 'weathersit'. The observations are:

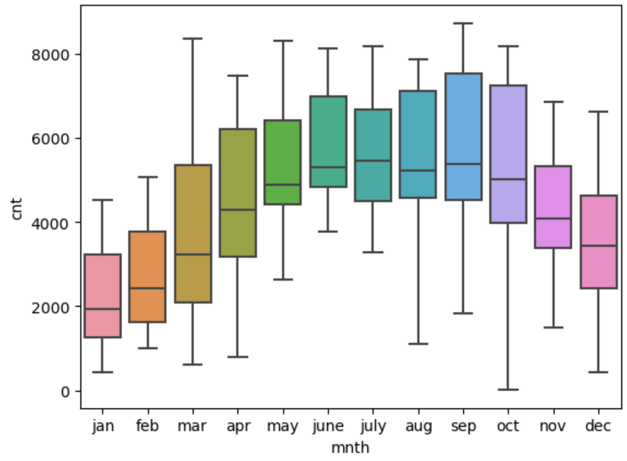
- The demand is very less in Spring season.
- The demand increased from year 2018 to 2019.
- The demand is less on a holiday.
- There is no demand for weathersit value 4(heavy rain), and very less for value 3 (light rain).
- The demand is very less in the months of Jan and Feb. This also correlates with the observation made based on season.
- The demand is constant throughout the week (monday to friday).

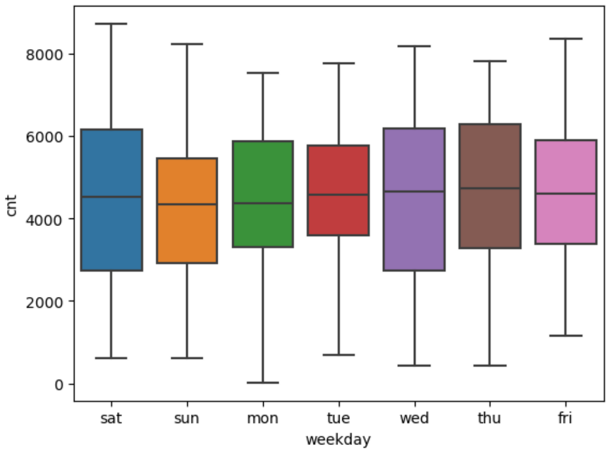












2. Why is it important to use drop_first=True during dummy variable creation?

To represent a categorical variable and use it in model building, we create dummy numeric variables. To represent M values(levels) of a variable, (M-1) dummy variables are sufficient. The additional variable becomes redundant, will increase time and complexity in building the model.

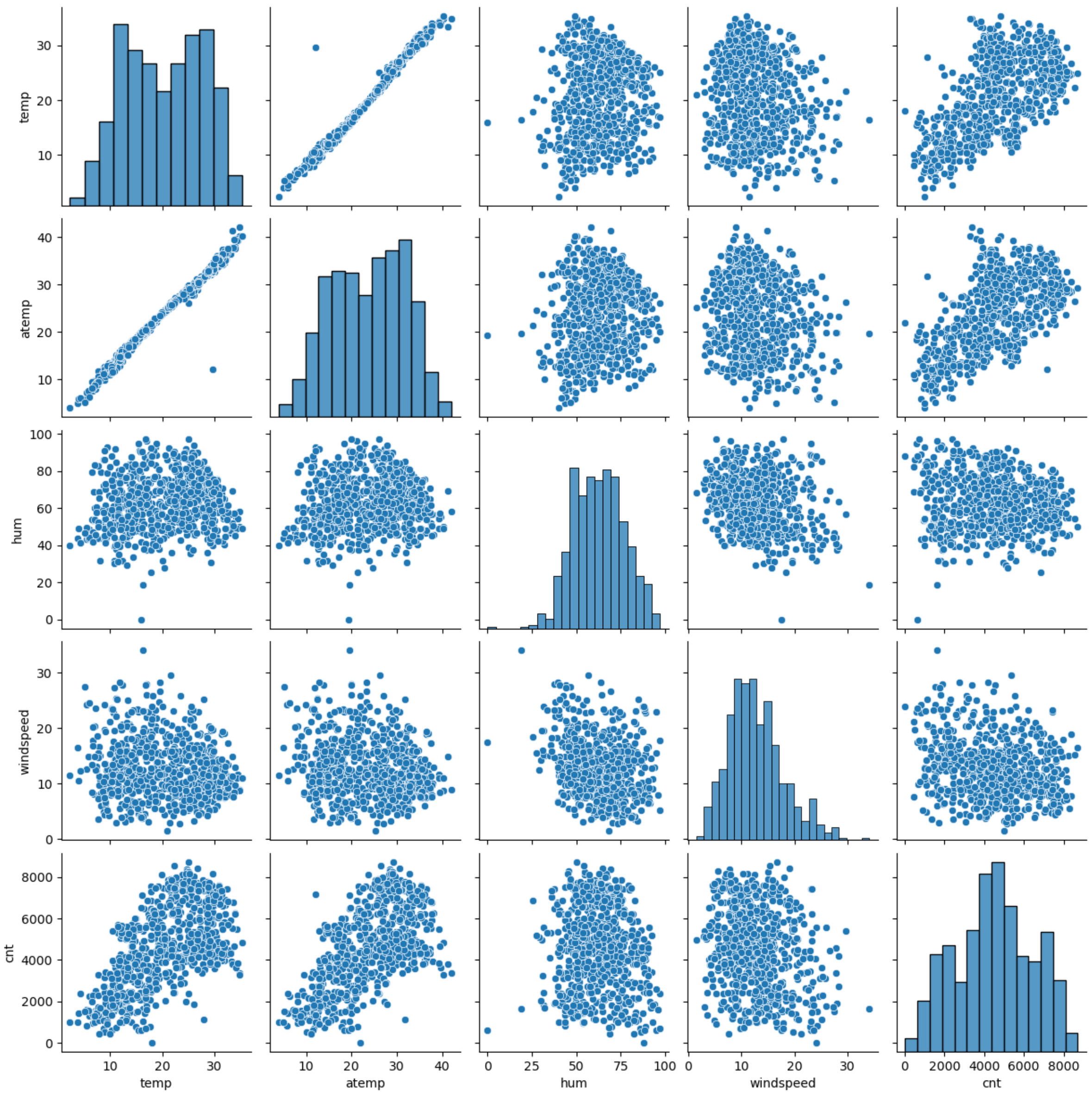
For example, consider a categorical variable ‘furnishing_status’ with three values – un-furnished, semi-furnished, fully-furnished.

To consider this variable for linear regression model, 2 dummy variables are sufficient. So, we can use drop_first=True to create only 2 dummy variables instead of 3.

semi_furnished	fully_furnished	
1	0	=> Indicates semi-furnished
0	1	=> Indicates fully-furnished
0	0	=> Indicates un-furnished

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

The following is the pair-plot for the numerical variables. And based on this, the variables ‘temp’ and ‘atemp’ have the highest correlation with the target variable ‘cnt’.

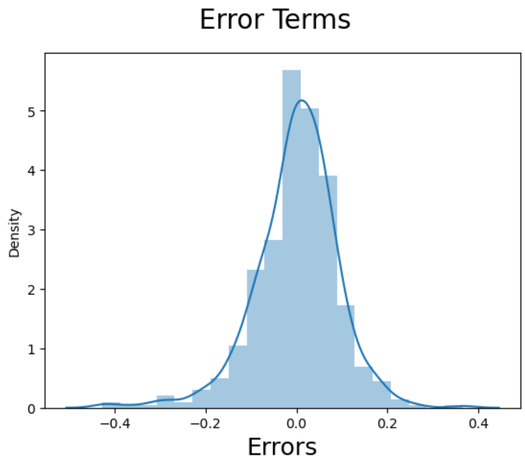


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

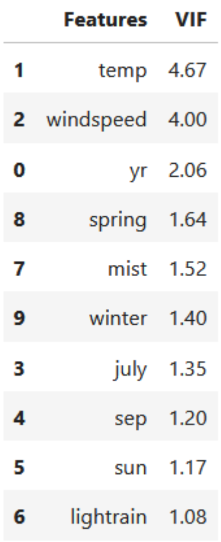
Before building the model, plots are drawn for continuous and categorical variables to verify if any of the variables have linear relation.

After the linear regression model is built, we looked at the F-statistic which has high value, and Prob (F-statistic) is close to zero indicating the model is stable and is significant. And looked at the R-squared and adjusted R-Squared values which had good results.

The equation obtained for the best fit line is
cnt = 0.2586 + 0.2346yr + 0.4493temp - 0.1410windspeed - 0.0698july + 0.0522sep - 0.0443sun - 0.2859lightrain - 0.0797mist - 0.1147spring + 0.0438winter
We looked at the error terms to confirm our assumptions of the model - The errors are normally distributed, the mean of errors is zero, the errors are independent, and their variance is constant.



Also, the variable selection is taken care to ensure there is no multi collinearity. The VIF of variables present in the final model are all less than 5 indicating the variables are all independent.



	Features	VIF
1	temp	4.67
2	windspeed	4.00
0	yr	2.06
8	spring	1.64
7	mist	1.52
9	winter	1.40
3	july	1.35
4	sep	1.20
5	sun	1.17
6	lightrain	1.08

5. 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 that is built, the equation for the best fit line is
cnt = 0.2586 + 0.2346yr + 0.4493temp - 0.1410windspeed - 0.0698july + 0.0522sep - 0.0443sun - 0.2859lightrain - 0.0797mist - 0.1147spring + 0.0438winter
The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- temp(Temperature) – with a positive coefficient of 0.4493
- yr(Year) – with a positive coefficient of 0.2346
- weathersit(value=3 light rain) – with a negative coefficient of 0.2859

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a statistical approach to establish linear relation between a set of input variables (independent/predictor variables) and one output variable (dependent/target variable). This is used to derive a formula that can define this relation in a mathematical form like

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

The formula or the best fit line is arrived at by minimizing the cost function which in this case is the RSS (Residual Sum of Squares).

There are certain assumptions made while building the linear regression model – There exists a linear relation between the predictor variables and the target variable. Also, the residuals/error terms are normally distributed, have a mean of zero, have a constant variance and are independent from the data.

Factors like F-statistic and Prob(F-statistic) are used to check if the overall model is significant or not.

The accuracy of the model built is evaluated using R2, with a value between 0 and 1. Higher the value, better the model – it can predict the outcomes well.

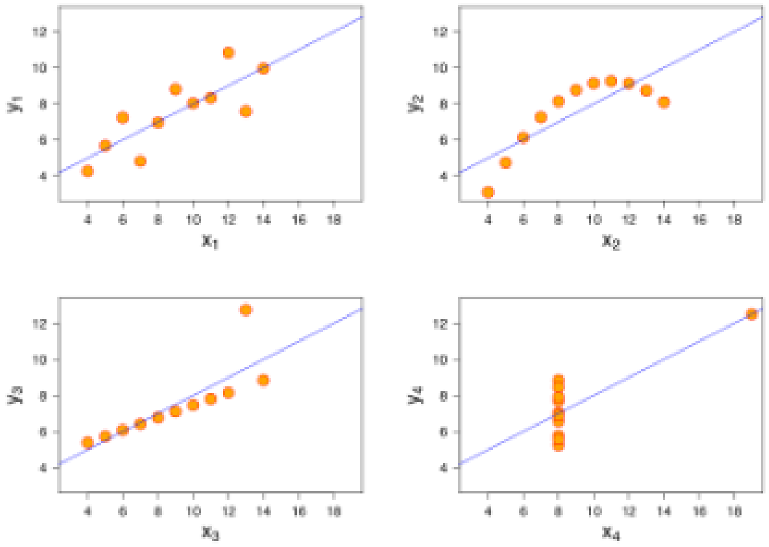
There are two types of linear regression:

- Simple Linear Regression – Used when the target is related to only one dependent variable.
- Multiple Linear Regression – Used when the target is related to multiple dependent variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet was constructed by statistician Francis Anscombe. It is used to explain some short comings of simple linear regression. And why we should not just run a regression model directly but analyze and explore the data via graphs and using Exploratory Data Analysis (EDA) before jumping into building the regression model.

Anscombe's quartet contains graphs plotted for four datasets. All four datasets have identical statistical parameters like mean, R-squared and same linear regression line derived for it. However, the graphs for these datapoints are completely different from each other.



The model looks fine in quarter-1. In quarter-2, the graph shows that linear regression doesn't look a right model for this data. In quarters 3 and 4, the presence of one outlier datapoint has significantly impacted the linear relation equation.

3. What is Pearson's R?

Pearson correlation coefficient (Pearson's R) is a statistical method used to measure linear correlation between two sets of data.

The formula produces a value between -1(perfect negative correlation) and 1(perfect positive correlation). The coefficient provides a measure of both strength of correlation as well as the direction (positive or negative). A value of 0 indicates no correlation.

The formula for Pearson's R is given by:

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

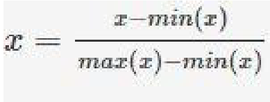
Where,
r = Pearson correlation coefficient
x = Values in the first data set
y = Values in the second data set
n = Total number of values.

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

In case where there are multiple predictor variables with wide range of values, it will lead to complexities with the model and its output. The model will have difficulties with convergence with gradient descent. Also, the derived coefficients will be widely varying which makes it difficult to interpret the impact of a variable on the model. To avoid these challenges, we use the mechanism of feature scaling. Scaling is done on the input dataset before starting the model building.

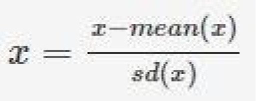
There are two types of scaling techniques:

MinMax (Normalized scaling) – The values of variables are modified such that the values of all are between 0 and 1.



$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardized Scaling – The values of variables are modified such that the mean of new values is zero and their standard deviation is one.



$$x = \frac{x - mean(x)}{sd(x)}$$

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

Variance Inflation Factor (VIF) is an indicator of how well the variance in this variable or the impact of the variable on the model can be explained by all other variables. Basically, it gives an indication of the relation between the variables used to compute the VIF.
VIF is calculated as 1 / (1-R^2).
A value of infinite for VIF means that R-square is 1. Which means that that variable has strong correlation with rest of the features. The total impact of this variable on the model is taken care by the other feature variables. So, this specific variable can be dropped from being considered for the model building. In general, a feature with a VIF>10 can be dropped from the model.

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

Quantile-Quantile plot or Q-Q plot is a scatter plot of quantiles of the first dataset against the quantiles of second dataset. This is used to compare and analyze the distribution of the data in the two sides. In the plots drawn, if the points are aligned on the diagonal line, it means the distribution of data is similar in both datasets. If there is a deviation from the diagonal, it indicates difference in the distributions.
In linear regression, when we do the train test data split, Q-Q plot can be used to observe the data distribution in both the sets.