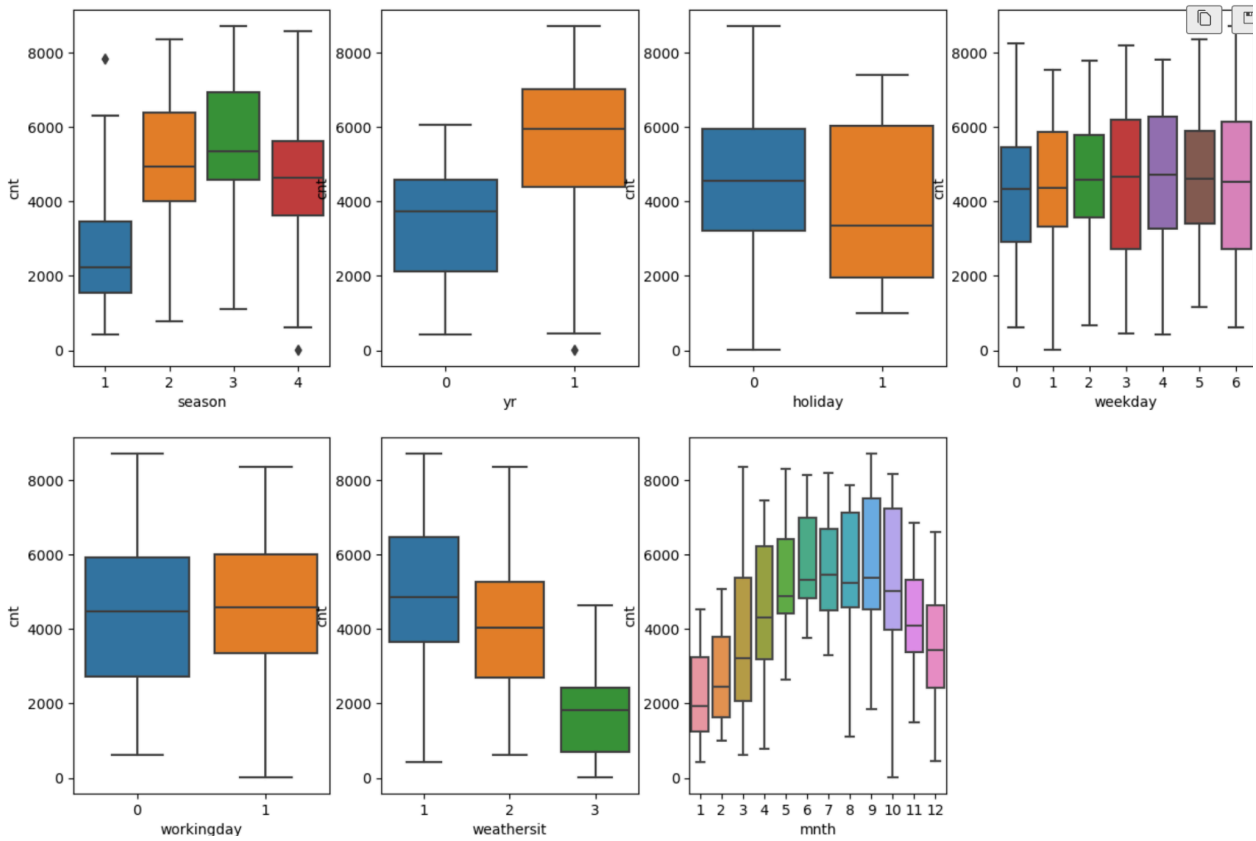
Boom-Bikes-Q&A

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



The categorical variable used in the dataset: season , yr(year) , holiday, weekday ,workingday, and weathersit(weather situa□on) and mnth(month) . These were visualized using a boxplot.

These variables had the following effect on our dependant variable: -

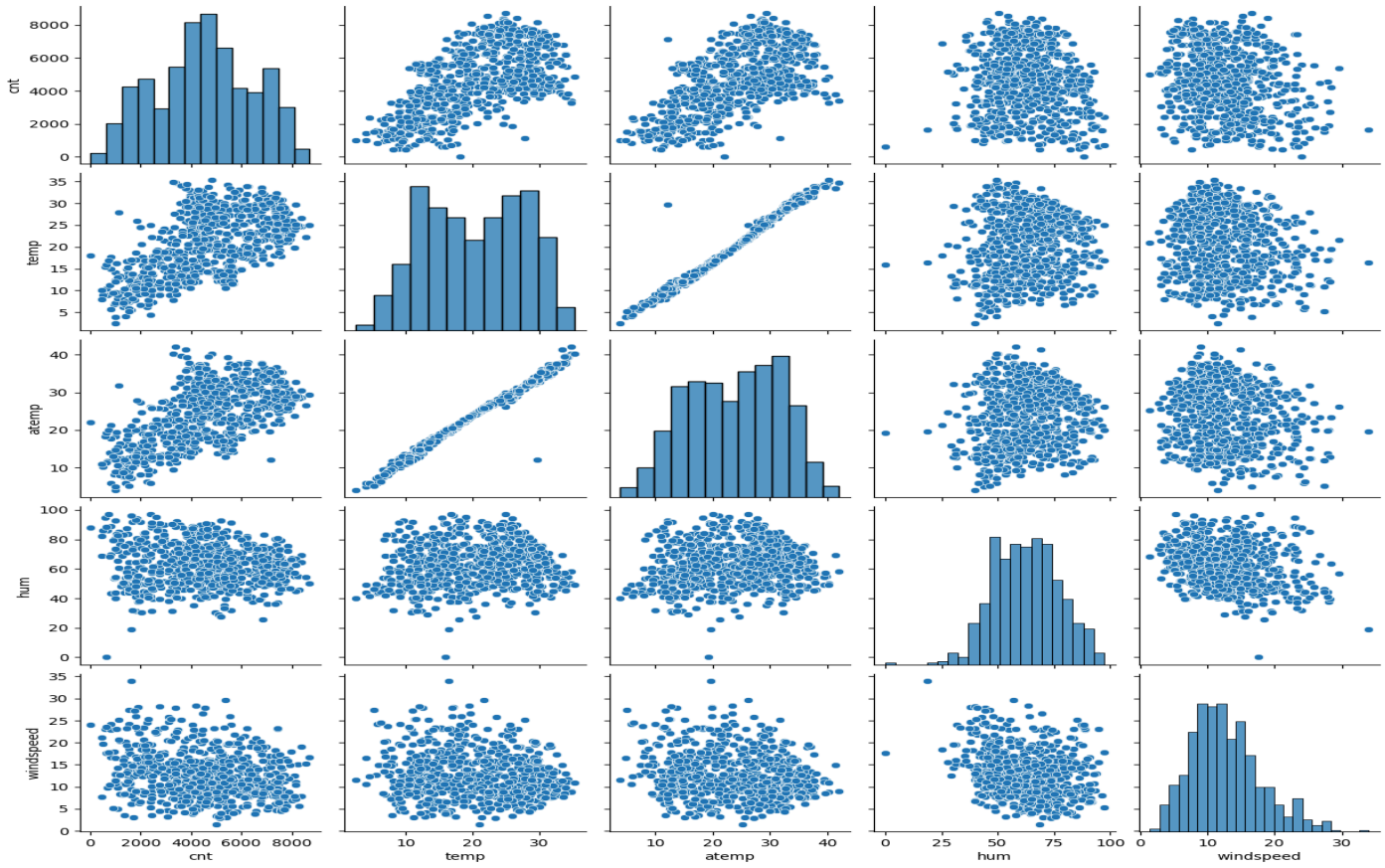
- Season - For the variable season, we can clearly see that the category 3: Fall, has the highest median, which shows that the demand was high during this season. It is least for 1: spring.
- Yr - The year 2019 had a higher count of users as compared to the year 2018.
- Holiday - rentals reduced during holiday.
- Weekday - The bike demand is almost constant throughout the week.
- Workingday – From the "Workingday" boxplot we can see those maximum bookings happening between 4000 and 6000, that is the median count of users is constant almost throughout the week. There is not much of difference in booking whether its working day or not.
- Weathersit - There are no users when there is heavy rain/ snow indica□ng that this weather quite adverse. Highest count was seen when the weather situa□on was Clear, Partly Cloudy.
- Mnth - The number of rentals peaked in September, whereas they peaked in December. This

observa□on is consistent with the observations made regarding the weather. As a result of the typical substan□al snowfall in December, rentals may have declined

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

It is important to reduce the no. of columns, as this helps in reducing the complexity of the model and therefore becomes easier to interpret the variables that are having a significant impact on the dependent variable. Including a dummy variable for every category of a categorical variable can lead to multicollinearity, this can be avoided by using 'drop_first=True'.

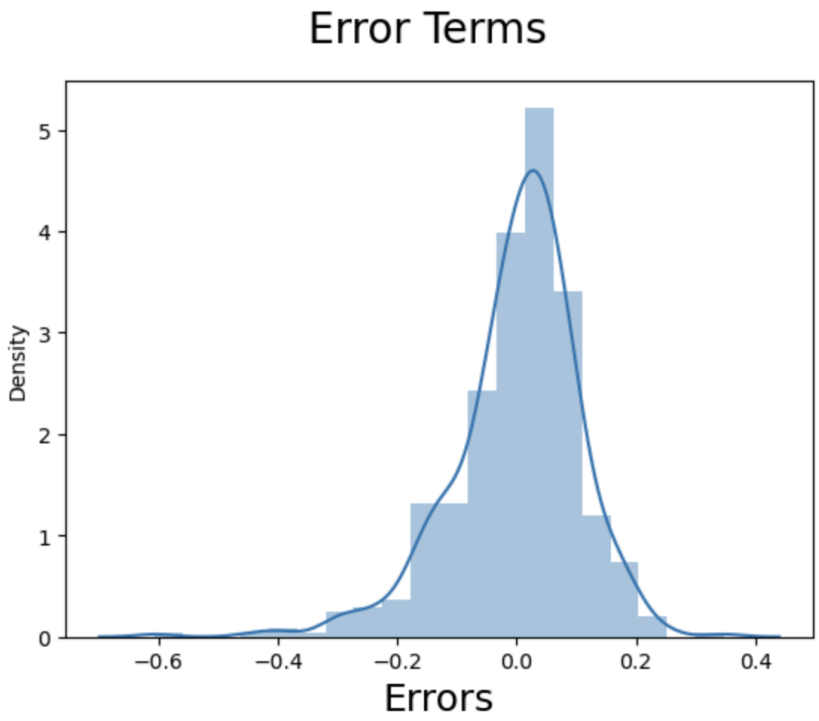
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



"temp" and "atemp" are the two numerical variables which are highly correlated with the target variable (cnt).

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

We have done following tests to validate assumptions of Linear Regression:
->There should be a linear relationship between independent and dependent variables. We visualized the numeric variables using a pairplot to see if the variables are linearly related or not. (ref. see above ques□on's pairplot)
->Residuals distribution should follow normal distribution and centered around 0 (mean = 0). We validated this assumption about residuals by plo□ng a distplot of residuals and saw if residuals are following normal distribution or not.



->linear regression assumes that there is litile or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to quantify how strongly the feature variables in the new model are associated with one another.

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

1. temp - coefficient : 0.51562
2. yr - coefficient : 0.240787
3. Weather- winter/summer & windspeed

General Subjective Questions

5.Explain the linear regression algorithm in detail.

Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task, which means it predicts a con□nuous output variable (y) based on one or more input variables (x). It is mostly used for finding out the linear relationship between variables and forecasting.

The basic idea of linear regression is to find a line that best fits the data points, such that the distance between the line and the data points is minimized. The line can be represented by an equation of the form:
$y = \theta 0 + \theta 1x$

where θ0 is the intercept (the value of y when x is zero) and θ1 is the slope (the change in y for a unit change in x). These are called the parameters or coefficients of the linear model.

To find the best values of θ0 and θ1, we need to define a cost function that measures how well the line fits the data. A common choice is the mean squared error (MSE), which is the average of the squared differences between the actual y values and the predicted y values:

$MSE = (1/n) * \sum(y - y')^2$
where n is the number of data points, y is the actual value, and y' is the predicted value.

The goal is to minimize the MSE by adjusting θ0 and θ1. There are different methods to do this, such as gradient descent, normal equation, or using libraries like scikit-learn.

Linear regression can also be extended to mul□ple input variables (x1, x2, ..., xn), in which case the equation becomes:

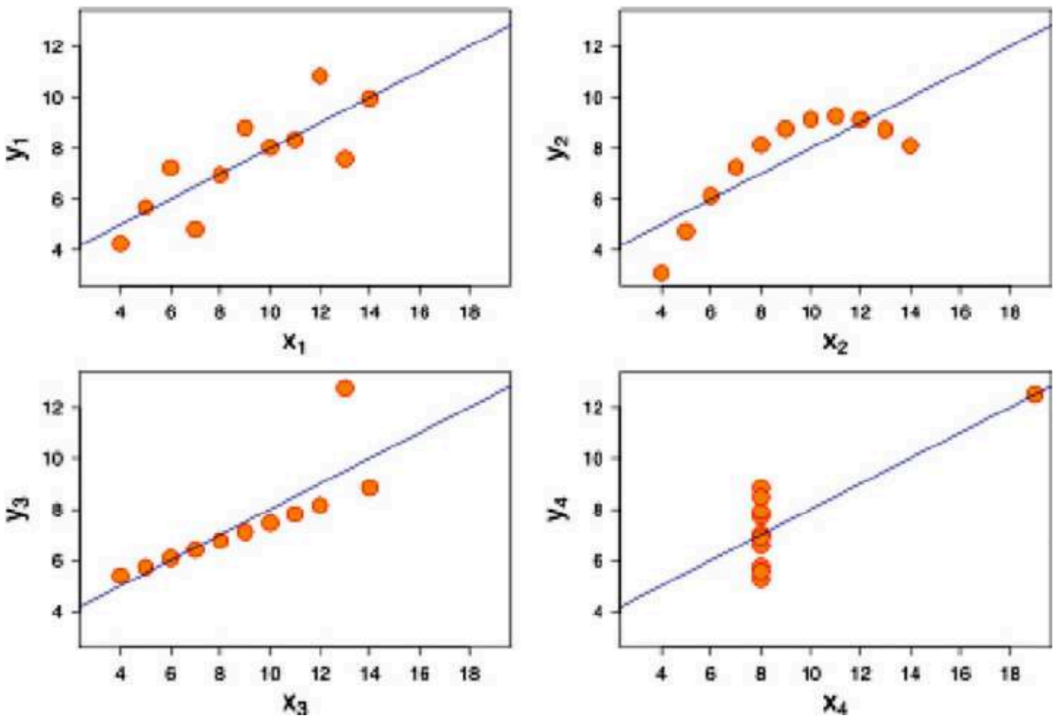
$y = \theta 0 + \theta 1x1 + \theta 2x2 + ... + \theta nxn$

Limitations are: it assumes a linear relationship between the input variables and the output variable, which may not always be the case. Another limitation is that it may be sensitive to outliers or multicollinearity.

6. Explain the Anscombe's quartet in detail

Anscombe's quartet consists of 4 different plots which are similar in terms of descriptive statistics (mean and variance), however their distributions are very different from each other. Anscombe's quartet illustrates the importance of first plotting the features before analysis or building a model. Following figure shows the linear regression model fitted for these 4 datasets-
Dataset 1: Here the linear regression model is a good fit
Dataset 2: The linear regression model cannot capture non-linear distribution
Dataset 3: Linear regression model is sensitive to the outlier, resulting in an incorrect result
Dataset 4: Here again the sensitivity towards outliers of the linear regression model is demonstrated

To summarize, anscombe's quartet depict how easy it is to build an incorrect regression model without due diligence of first visualizing the relationships. Fig next page



7. What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us "can we draw a line graph to represent the data?"

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

The pearson's correlation coefficient can be primarily used, when all of the following points are true-

->There is linear relationship between the quantitative variables
->There are no outliers
->The errors are normally distributed
->The pearson's coefficient also provides a measure of how well the observations are distributed around the best-fit line. A value of -1 or 1 indicates that all the observations are distributed on the line.

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

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespec□ve of the units of the values.

● Normalization is generally used when you know that the distribu□on of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribu□on of the data like K-Nearest Neighbours and Neural Networks.
● Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

10. You might have observed that some□mes the value of VIF is infinite. Why does this happen?
The VIF (Variance Inflation Factor) gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model. Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have

perfect correla□on and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity" The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a par□cular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correla□on with other predictors.

A rule of thumb for interpreting the variance infla□on factor: 1 = not correlated.
Between 1 and 5 = moderately correlated.
Greater than 5 = highly correlated.

$$VIF = \frac{1}{1 - R^2}$$

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

A q-q plot is a plot of the quantiles of the first data set against the quan□les of the second data set. It is used to compare the shapes of distributions .A Q-Q plot is a scatterplot created by ploting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. The q-q plot is used to answer the following questions:

● Do two data sets come from populations with a common distribution ● Do two data sets have common location and scale?
● Do two data sets have similar distributional shapes?
● Do two data sets have similar tail behavior?