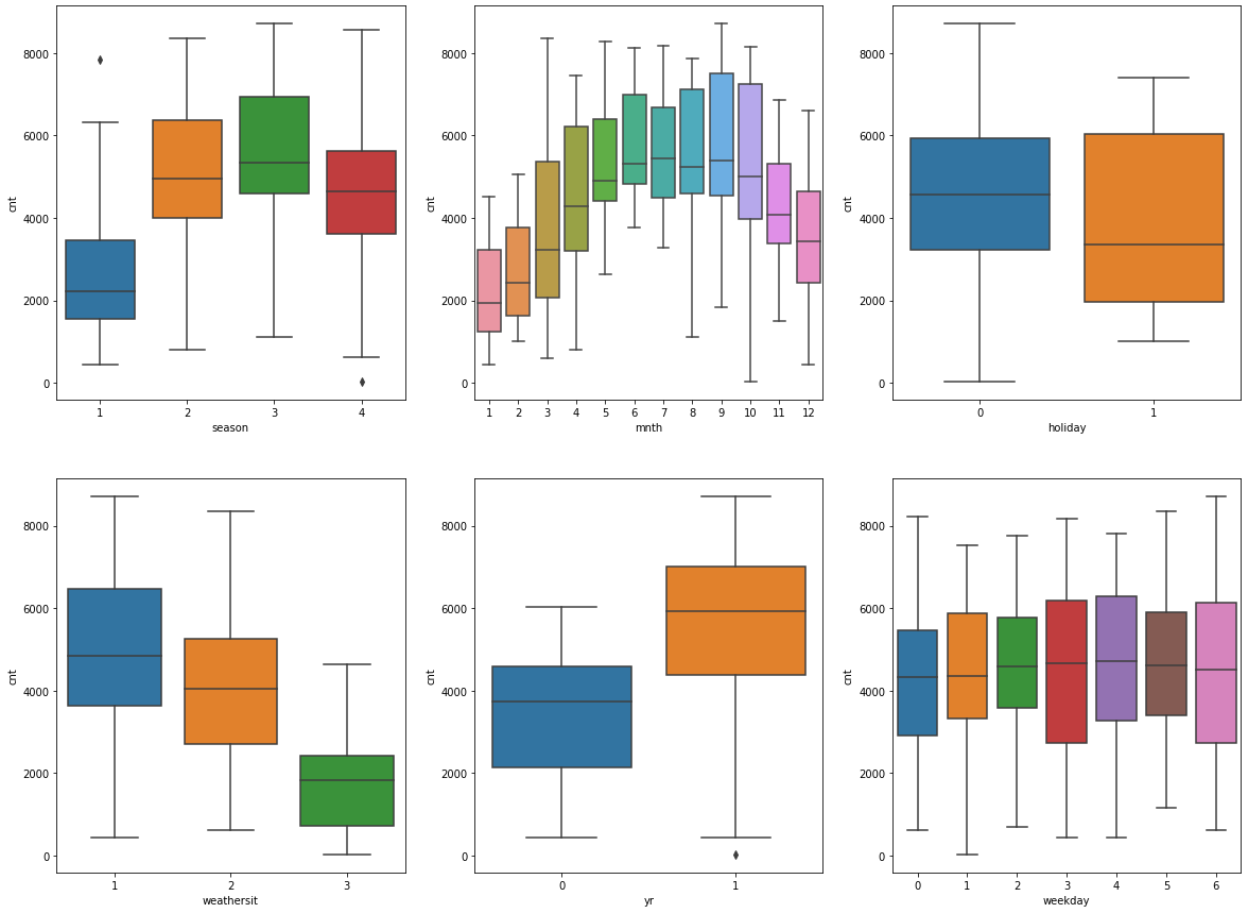
Assignment-based Subjective Questions

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



Ans: - In our dataset we have **Season, yr., mnth, weathershit, weekday** these are our categorical

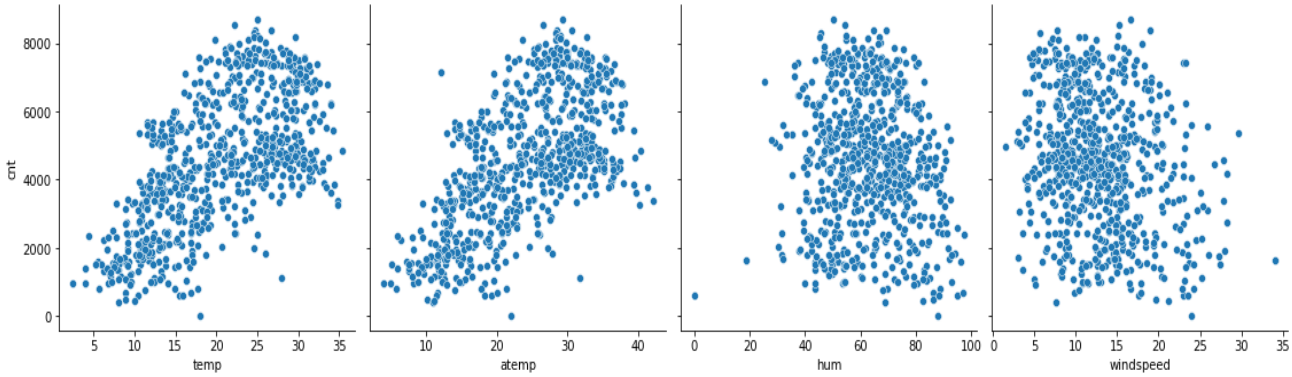
variables as we see in our visualizing the data section we create some boxplot and also we create some barplot in data preparation section We can say these following things as mention below:

1. **Mnth: -** In **August, September, October** month bike share demand is high but **January, February** month bike share demand is low.
2. **Season: -** In **Fall** season bike share demand is high but **Spring** season bike share demand is low.
3. **Yr.: -** in **2019** year bike share demand is high but **2018** year bike share demand is low.
4. **Weathershit: -** In **Clear** weather bike share demand is high but **Low Snow, Heavy Rain** weather bike share demand is low.
5. **Weekday: -** In **Thursday** bike share demand is high but **Sunday** bike share demand is low.

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

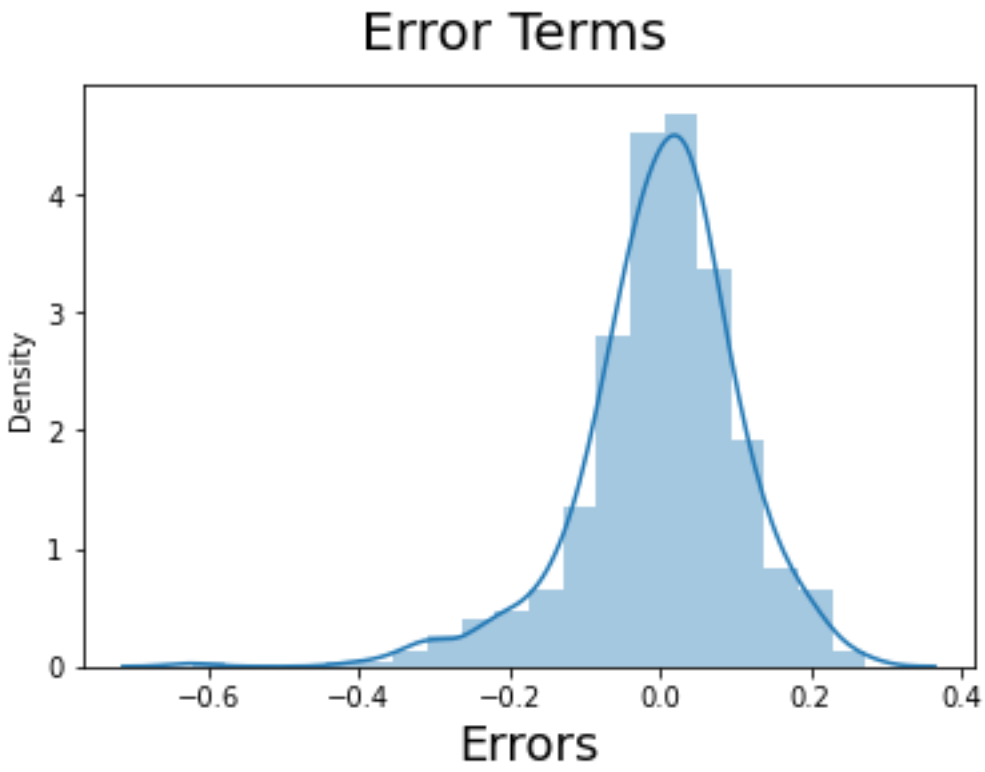
   Ans: - Because **drop_first** allows you whether to keep or remove the reference (whether to keep k or k-1 dummies out of k categorical levels). Here **drop_first = False** meaning that the reference is not dropped and k dummies created out of k categorical levels! You set **drop_first = True**, then it will drop the reference column after encoding. If you don't drop the first column then your dummy variables will be correlated and affect some models. That's the reason it is important to use.

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



Temp, atemp is highest correlation with our target variable **cnt** .

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



Ans: - Because we need to check if the error terms are also normally distributed or not here, we plot the histogram of the errors terms and see how it's looks like.

As per the graph we can say the residuals are following the normally distributed with a mean 0.

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

Ans: - Based on Our final model these are the 3 features is given below: -

1. **Temp**: - which have highest coefficient value is 0.579954.
2. **Yr**.: - which have second highest coefficient value is 0.23236.
3. **Windspeed**: - which have highest negatively coefficient value is -0.190149.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Ans: - linear regression is a machine learning algorithm based on supervised learning and it's perform a regression task. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). There are two main types:

   1. **Simple linear regression**: - If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

The equation of a SLR can be written as:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

1. **Y** represents the output or dependent variable.
2. **β0 and β1** are two unknown constants that represent the intercept and coefficient (slope) respectively.
3. **ε** (Epsilon) is the error term.

2. **Multiple linear regression: -** If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

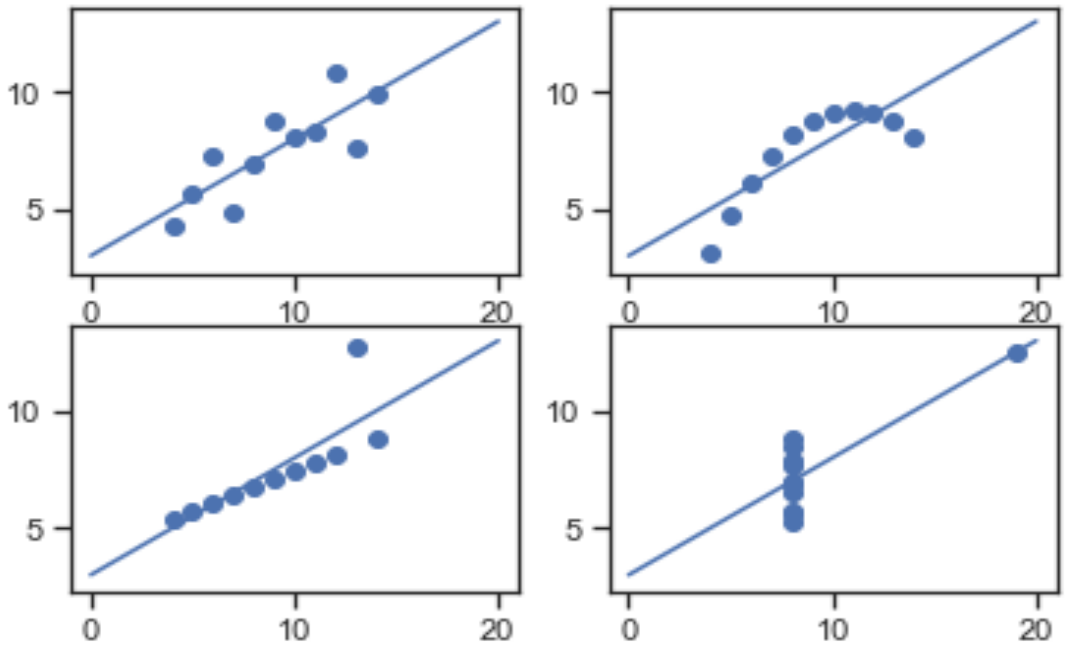
The equation of a SLR can be written as:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

1. $\hat{Y}$ is the predicted or expected value of the dependent variable.
2. X_1 through X_p are **p** distinct independent or predictor variables.
3. b_0 is the value of **Y.**
4. b_1 through b_p are the estimated regression coefficients.

2. Explain the Anscombe's quartet in detail.

Ans: - Anscombe's Quartet can be defined as a group of four data that have nearly identical simple statistical properties, yet appear very different when graphed. They were constructed in 1973 by the **statistician Francis Anscombe** to demonstrate both the importance of graphing data when analyzing it, and the effect of **outliers** and other **influential observations** on statistical properties.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on *x*.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the

corresponding coefficient of determination would be more appropriate.

- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association

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

Ans: - Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. It also helps in speeding up the calculations in an algorithm.

Because of most of the times, collected data set contains features highly varying in magnitudes, units or range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we need to perform scaling to bring all the variables to the same level of magnitude.

Normalization: - Normalization use minimum and maximum value of features are used for scaling and it is really affected by outliers. It is useful

when we don't know about the distribution. Normalization scales values between [0, 1] or [-1, 1].

Standardization: - Standardization use mean and standard deviation is used for scaling and it is much less affected by outliers. It is useful when the feature distribution is Normal or Gaussian. Standardization scales has not certain range.

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

Ans: - VIF shows a perfect correlation between two independent variables if there is perfect correlation, then VIF = infinity. In the case of perfect correlation we get R2 =1 which lead to 1/(1-R2) infinity. To solve this problem

we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans: - Q-Q Plots is known as a Quantile-Quantile plots Q-Q plots is a plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. Q-Q plots is used to find out if two sets of data come from the same distribution and it's also used for compare the shapes of distributions providing a graphical view of properties such as location, scale, and skewness are similar or different in the two distributions.