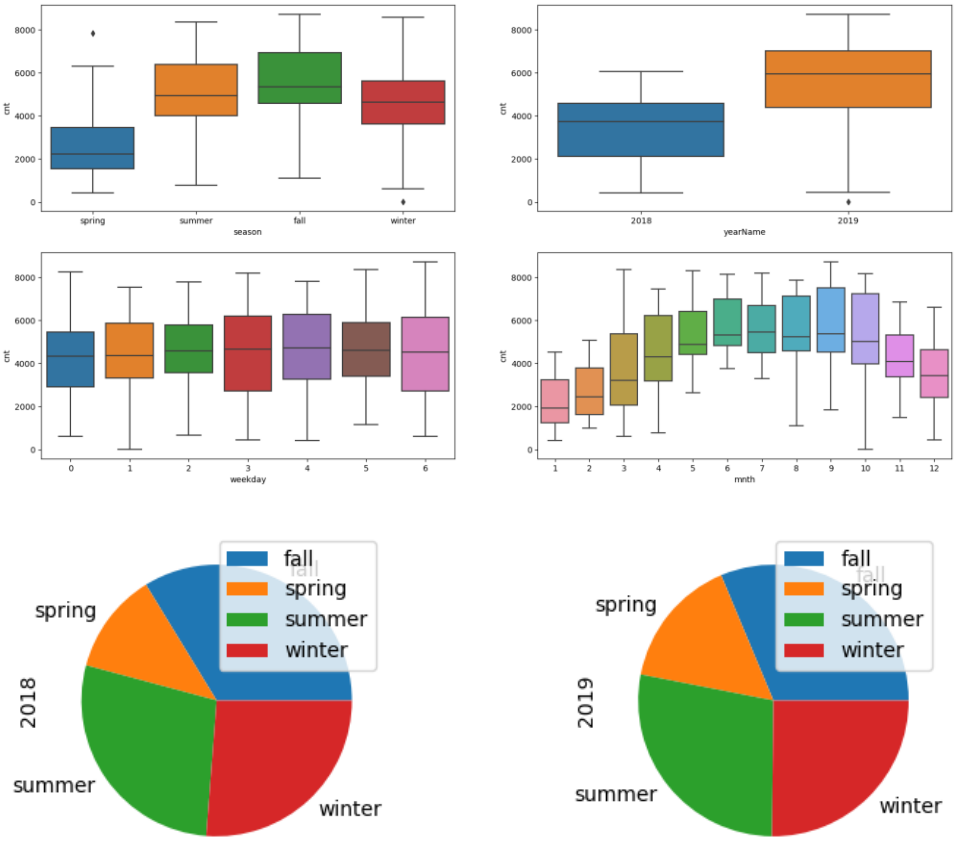
Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans. *From my analysis of the categorical variables in the Dataset, I can infer:*

a. *The median count of bike users/customers increase during the season of Summer & Fall, while it was little less during winters and least during spring*
b. *The number of customers increased from 2018 to 2019*
c. *The pattern of sales across the seasons and across the year are same.*



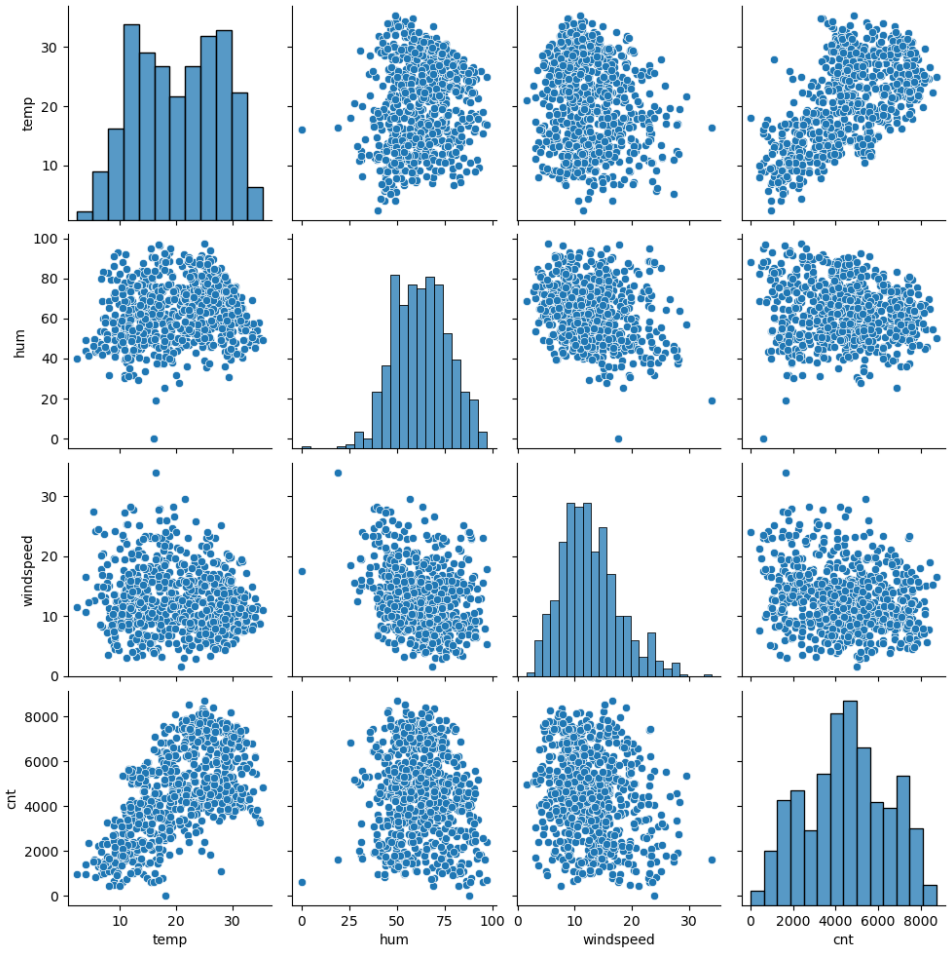
Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans. *The drop_first = True is important to use during dummy variable creation, as it will create redundancy in the data. For example, if we have a categorical variable: Gender as*

Male & Female. Upon converting it into dummy variables, it will result into a column having data 1,0. However, if we create one column for male & one for female. It will add redundancy in the model as 1 already stands for Male and 0 for female.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. *Temp has the highest Correlation with the Cnt (Target) column.*

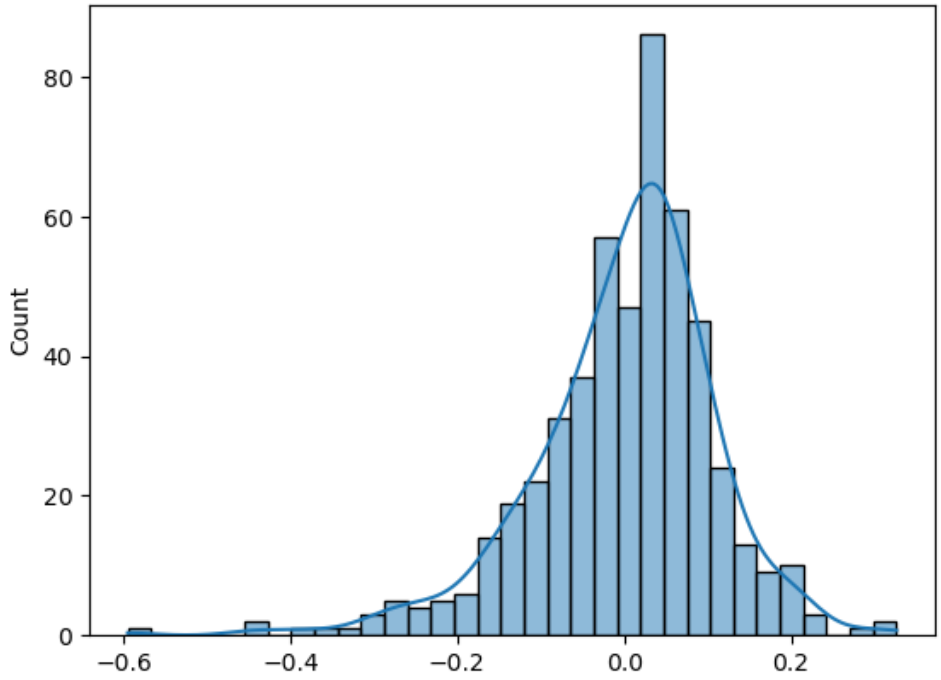


Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans. *The following validations were performed after creating the LR model.*

a. *Checked the P Value of the Coefficients to ensure they are significant ($p < 0.05$)*

b. *Check the VIF among the features to ensure none of the features have dependencies.*
c. *Checked residual distribution to see if it follows normal distribution*



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans. *Below 3 features contribute significantly to demand of the shared bikes:*

a. *Temp (Temperature)*
b. *Yr (Year)*
c. *Hum (Humidity)*

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans. *The linear regression algorithm is a machine learning algorithm that makes use of the concept of vector algebra to create a best fitting line (or plane) to a set of given points.*

Linear Regression are of two types:

a. *Simple Linear Regression (Single independent Variable)*
b. *Multiple Linear Regression (More than one independent variable).*

As part of creating a model using the Linear Regression model, few assumptions are to be considered.

- ***Linear relationship****: Meaning the pattern do not follow a non-linear pattern*
- ***Multivariate normality****: All the variables are normally distributed*
- ***No or little multicollinearity****: All variables are independent*
- ***No autocorrelation****: The value of one variable does not depend on the value of the variable at a previous time.*
- ***Homoscedasticity****: Variance is constant throughout.*

To validate the correctness of the model, certain metrics are used:

- ***R2: aka R Square:*** *This value ranges between 0 & 1. The higher the value of R2, the better the model. However, too large a value may suggest overfitting.*
- ***Adjusted R2:*** *While, R2 always increases with addition of new variables, the adjusted R2 was introduced that penalizes the model for adding more variables.*
- ***AIC***
- ***BIC***

Q2. Explain the Anscombe's quartet in detail.

Ans: *The Anscombe's quartet is a set of dataset that have identical statistical values like mean, median, mode, std dev etc. However, when plotted in a change show completely different patterns. These can be linear, non linear or random patterns. Anscombe's quartet main idea is to stress on the importance of visualizing the data.*

Q3. What is Pearson's R?

Ans. *Pearson's R is a correlation metric used to compute the correlation factor and the direction of correlation between two continuous variables. The range of the Pearson's R value is between –1 & 1 a –ve correlation means inversely proportional and a +ve correlation refers to direct proportionality.*

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. *Scaling in the data is performed to ensure that the data falls within the same range as the other variables. If scaling is not performed, it can have drastic impact on the model. For example, if our model is designed on Rupees where 80 Rupees = 1 Dollar and the model is run against a dataset where the price is mentioned in dollars it will have a huge impact on the predicted output.*

Normalized Scaling: *This type of scaling bring the data in the range 0-1.*

Formula: $\boldsymbol{(x - x_{min}) / (x_{max} - x_{min})}$

Standardized Scaling: *This type of scaling bring the data in the range –1 to 1. This is also similar to the z-score*

Formula: $\boldsymbol{(x_i - u) / (sigma)}$

Where sigma is the std dev. Of the variable.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. *The formula for VIF is:*

$\boldsymbol{1/(1 - R2)}$

Based on this formula, if the value of R2 becomes 1, the value of VIF will reach infinite. This happens in an overfit scenario, or in the case when one variable can completely help in explaining the variance in the other variable(s).

Q6. *W*hat is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans. *The Q-Q plot stands for Quantile-Quantile plot. It is a graph that can be leveraged to understand if the dataset follows a certain distribution such as Normal distribution. The quantiles are calculated based on the chosen distribution (ex: Normal distribution). Using these quantile values we can get an idea about the distribution of the data. If the dataset fall in approx. straight line, then the dataset follow the suggested distribution.*