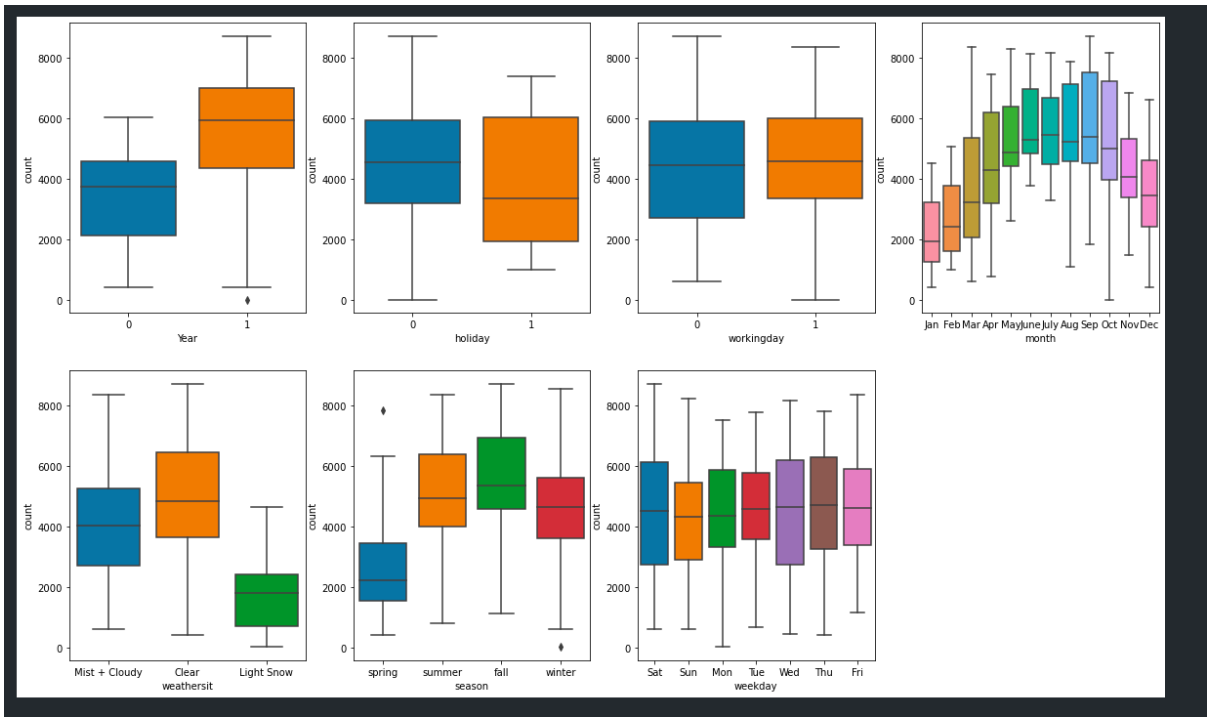
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)



Based on the above analysis, following Observations are observed

- Season: In fall season, majority of the people used bike sharing service.
- Year: Compared to 2018, in 2019 People used the Bike sharing service more.
- Month:
  - In alignment with fall season, the bike sharing system was used more in the mid year from May to July.
  - Bike sharing service was used more in March, April, Aug, Sep and Oct.
- Weekday: There is no significant pattern in weekdays, Mean of Bike sharing service is same almost all the days. This states that Bike sharing service is almost same every day.
- Weathersit: Bike sharing service was used least when the weather situation is of lightsnow.
- Holiday: Service is more used on holiday, compared to Non-holiday.
- Workingday: The bike sharing service was slightly high on weekend vs weekday. The difference is negligible.

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

Based on the other columns which have values 0's and 1's. The initial column's value can be inferred. Hence, first column wouldn't be required in the analysis.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Registered column has the highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

There are multiple components that were taken into consideration along building the model. Using RFE model, the 15 most correlated variables to the target variable are identified and then getting VIF and observing the Summary details of R square, Adjusted Rsquare, P value, multiple variables are dropped and the model is built.

Similar process is performed on Test set too and the Trained set R square, Trained Adjusted R square, Test set R square, Test set Adjusted R square values are compared to build the model

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Year, September and Saturday are the positively related coefficients to the model.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

While training the model we are given :
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.
θ1: intercept
θ2: coefficient of x

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

3. What is Pearson's R? (3 marks)

In Statistics, the Pearson's Correlation Coefficient is also referred to as **Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation**. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

- Min max scaling: It brings all of the data in the range of 0 and 1

- Standardized Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

If VIF = infinity. This states perfect correlation between two variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

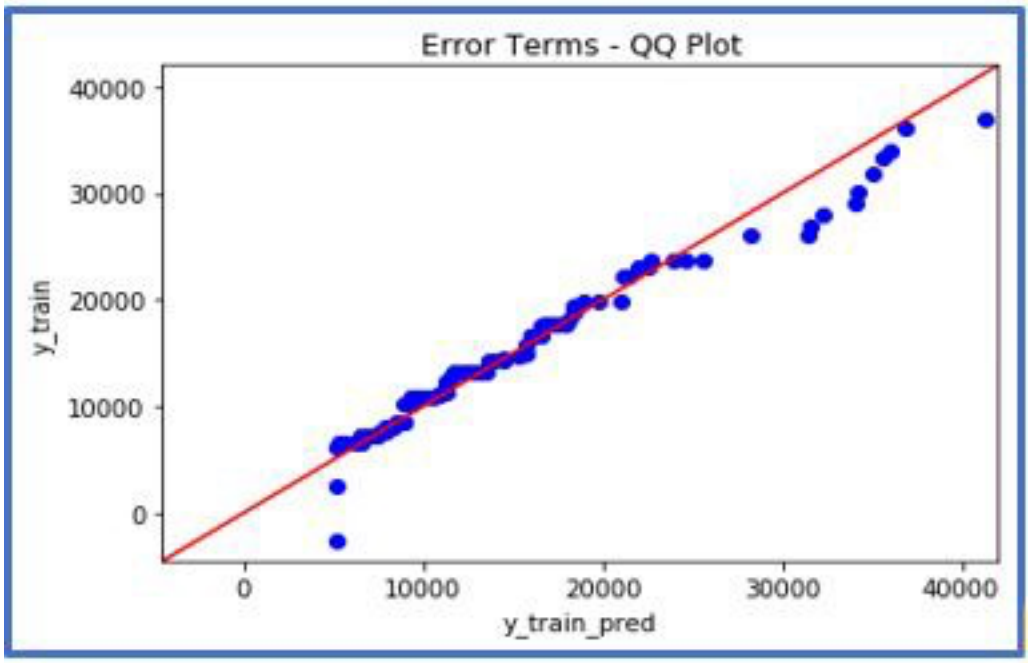
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.