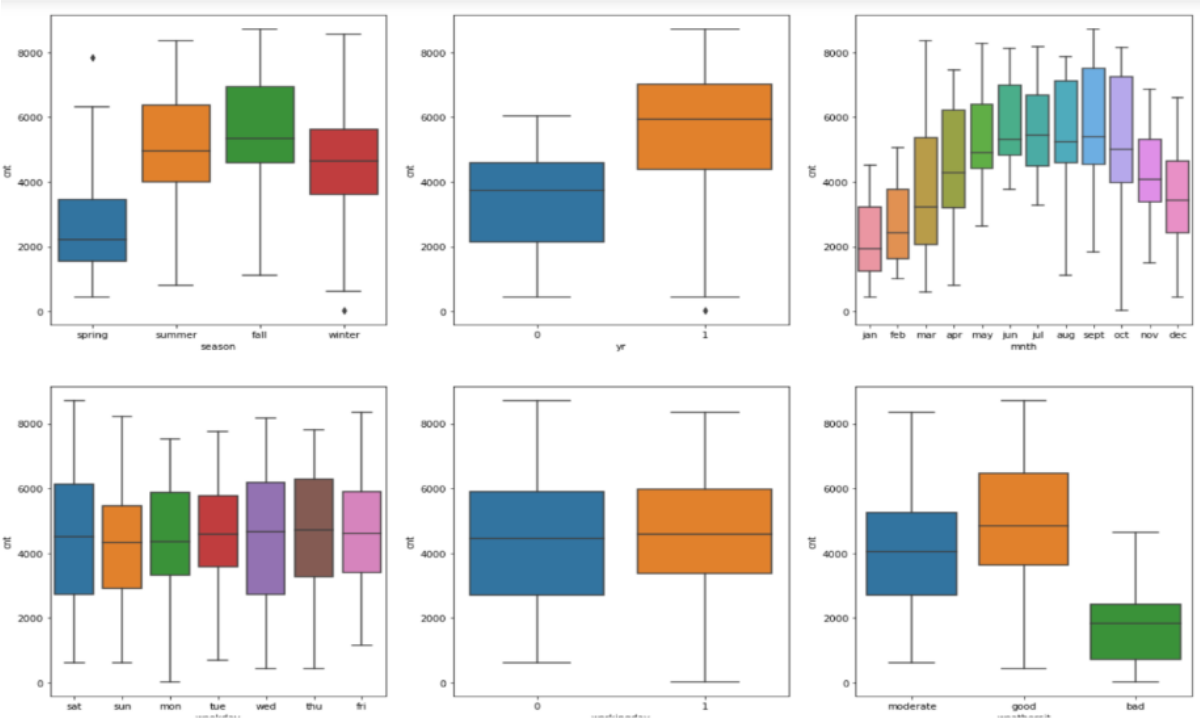
Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

- There are categorical variables like season, month, weekday, weathersit, holiday, yr, workingday which has moderate to significant impact on our target variable cnt.
- Please find the below box plot for better understanding.



- At the time of final model we found that some of the categorical variable have more significant impact than other after removing multicollinearity between them.
    'yr'
    'workingday'
    'season_spring'
    'season_winter',
    'mnth_dec'
    'mnth_nov'
    'weekday_sat'
    'weathersit_bad',
    'weathersit_moderate'

- After removing the dummy variable and calculating VIF between the above categorical variable were left. There were having more significant impact on the target variable cnt.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

The intention behind the dummy variable is that for a categorical variable with 'n' levels, you create new columns each indicating whether that level exists or not using a zero or one.
The n levels in categorical variable can be easily explained by n-1 dummy variables, hence that 1 category in dummy variable become redundant and can cause dummy variable trap due to multicollinearity.

Using drop_first=True when creating dummy variables is important because it helps avoid the dummy variable trap, which occurs when there is multicollinearity between the dummy variables.

Hence drop_first=True is used so that the resultant can match up n-1 levels. Hence it reduces the correlation among the dummy variables.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable 'cnt'. There is almost a linear relationship between them.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Linear Regression models are validated based on Linearity, Normality of error, Homoscedasticity, there should be no Multicollinearity within target variables.

One of the assumptions that we checked on training set is when check the normality yr by plotting histogram on residuals after we have done predictions on training data. It checked out and residual were showing normality.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

We could see that atemp, yr and season(winter season) were the most significant features

contributing towards explaining the demand of the shared bikes. Please find below features along with their coefficients values.

yr (0.241933) , atemp(0.420321) , season_winter(0.096814)

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>
Linear regression is a form of predictive modeling technique which tells us the relationship between the dependent (target variable) and independent variables(predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables (y = Bo + B1 X).

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for coefficients and intercept to find the best fit line and the best fit line should have the least error.

In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for coefficients and intercept, which provides the best fit line for the data points

---

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

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. It was constructed to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

Anscombe's quartet consists of four small datasets, each with 11 data points. The four datasets are:

Dataset I (Linear Relationship)
Dataset II (Non-linear Relationship)
Dataset III (Outliers)
Dataset IV (Vertical Line)
Even though the statistical summaries (mean, variance, correlation, regression line) of all four datasets are identical, their visual characteristics are quite different.

Anscombe's quartet is a classic example used in statistical education to teach the importance of exploratory data analysis, especially the need to visualize data before applying statistical methods. It emphasizes that numerical summaries alone do not give a complete picture of the data and that the structure of the data matters greatly in interpreting results accurately.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>
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, giving a value between -1 and +1.

Interpretation of Pearson's r:

- +1: A perfect positive linear relationship, meaning as one variable increases, the other also increases in a perfectly linear manner.
- 0: No linear correlation, meaning there is no linear relationship between the variables. (Note: This does not necessarily mean there's no relationship at all, as non-linear relationships could still exist.)
- -1: A perfect negative linear relationship, meaning as one variable increases, the other decreases in a perfectly linear manner.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>
Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.

Difference between Normalizing Scaling and Standardize Scaling:
1. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
3. Normalized scaling scales values between (0,1) whereas standardized scaling is not having or is not bounded in a certain range.
4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
5. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.
6. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

VIF (Variance Inflation Factor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:
A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.

A very high VIF value shows a perfect correlation between two independent variables. 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.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a specific theoretical distribution, most commonly the normal distribution. It compares the quantiles of the data against the quantiles of a chosen theoretical distribution. In a Q-Q plot, if the data is from the theoretical distribution (e.g., normal distribution), the points will approximately lie on a straight line.

X-axis: The quantiles of the theoretical distribution.
Y-axis: The quantiles of the sample data

In the context of linear regression, the Q-Q plot is primarily used to check the residuals (the differences between the observed and predicted values) for normality. This is important because one of the assumptions of linear regression is that the residuals are normally distributed. Violating this assumption can lead to inefficient estimates and affect the reliability of hypothesis tests and confidence intervals.
