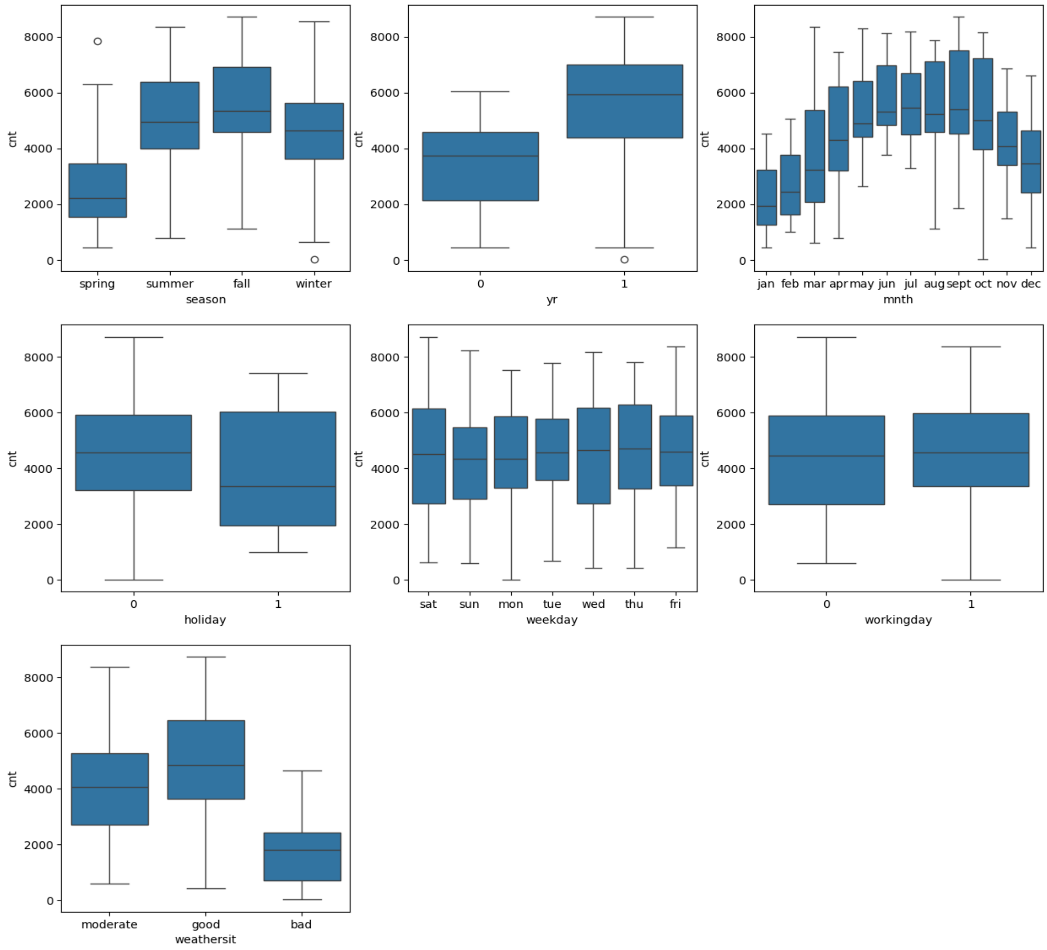
Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans: From our analysis of the categorical variables from dataset, below insights can be drawn.

- Season: 3: fall has highest demand for rental bikes.
- Demand for next year has grown.
- Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing.
- When there is a holiday, demand has decreased. Weekday is not giving clear picture about demand.
- The clear weathersit has highest demand.
- During September, bike sharing is more. During the year end and beginning, it is less, could be due to extereme weather conditions.



Q2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans: If we do not use drop_first = True, then n dummy variables will be created, and these predictors (n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap, so drop first=True is very important to use, as it helps in reducing multicollinearity.

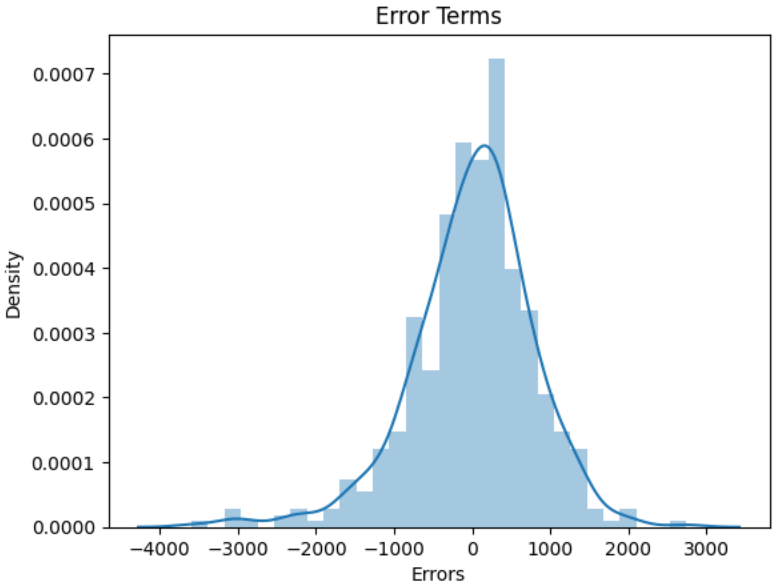
When we create dummy variable, one column is always obsolete as its can be derived based on other available variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: *atemp* and *temp* has the highest correlation with the target variable, as is evident from the pair plot.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

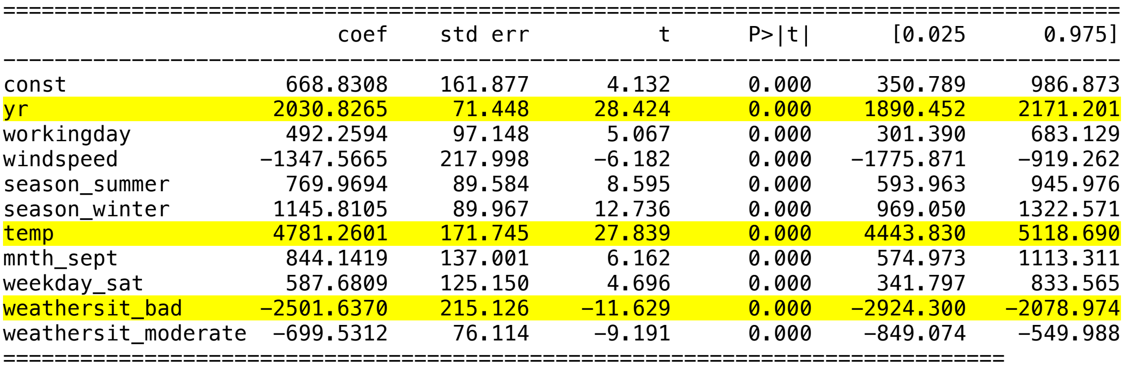
Ans. One of the fundamental assumptions of a linear regression model is that the error terms should correspond to a normal curve, when plotted on histogram and spotted the same. Hence, this assumption is validated.



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans. The top 3 features directly influencing the count are the features with highest coefficients. These are: -

- Temp (Positively influencing)
- Year (Positively influencing)
- snowy and rainy weather (negatively influencing)



```
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                  668.8308    161.877      4.132      0.000     350.789     986.873
yr                    2030.8265     71.448     28.424      0.000    1890.452    2171.201
workingday             492.2594     97.148      5.067      0.000     301.390     683.129
windspeed            -1347.5665    217.998     -6.182      0.000   -1775.871    -919.262
season_summer          769.9694     89.584      8.595      0.000     593.963     945.976
season_winter         1145.8105     89.967     12.736      0.000     969.050    1322.571
temp                  4781.2601    171.745     27.839      0.000    4443.830    5118.690
mnth_sept              844.1419    137.001      6.162      0.000     574.973    1113.311
weekday_sat            587.6809    125.150      4.696      0.000     341.797     833.565
weathersit_bad       -2501.6370    215.126    -11.629      0.000   -2924.300   -2078.974
weathersit_moderate   -699.5312     76.114     -9.191      0.000    -849.074    -549.988
==============================================================================
```

General Subjective Questions

Q1. Explain the linear regression algorithm in detail. (4 marks)

An interpolation technique used to predict correlation between variables and how an independent variable is influenced by the dependent variable(s), is linear regression.

After looking into the data and cleaning it with EDA(exploratory data analysis), we split the dataset into training set (which would be used to train a model) and the testing set (which would be used to check how close is our model to the actual output). After checking the collinearity of variables and using the requisite variables to train the model and checking the R-square value of the model and the p-values of dependent variables, after dealing/dropping the necessary columns and reiterating the steps (feature elimination), we come to a final model.

According to the conditions of linear regression which states that the error curve must be a normal one, we proceed to testing the model with the test dataset. The conclusion hence drawn on the model would be used to provide valuable insights/predictions on datapoints in the range of the model.

Q2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe’s quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyse it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

A regression model is not always necessarily an exact one, it can also be fooled by some (smart) data! In certain cases, there are multiple datasets which are completely different but after training, the regression model looks the same. A group of four such datasets having identical descriptive statistics but with some peculiarities, is the Anscombe's quartet.

Anscombe’s Quartet Four Datasets
Data Set 1: Fits the linear regression model pretty well.
Data Set 2: Cannot fit the linear regression model because the data is non-linear.
Data Set 3: Shows the outliers involved in the data set, which cannot be handled by the linear regression model.
Data Set 4: Shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

As you can see, Anscombe’s quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

Q3. What is Pearson’s R? (3 marks)

Pearson's correlation coefficient, also known as Pearson's R, is a measure of the strength of correlation between two variables. It is commonly used in linear regression. The value of Pearson's R always lie between -1 and +1, the latter indicating a perfectly positive and linear correlation and the former indicating a perfectly linear negative regression. The values in between denotes the relative collinearity of two variables.

When r is 1 or –1, all the points fall exactly on the line of best fit:
When r is greater than .5 or less than –.5, the points are close to the line of best fit:
When r is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit:
When r is 0, a line of best fit is not helpful in describing the relationship between the variables:
Below is a formula for calculating the Pearson correlation coefficient (r):

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

We can also use software such as R or Excel to calculate the Pearson correlation coefficient.
The Pearson correlation coefficient is a good choice when all of the following are true:
Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.
The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is necessary for a model to be functional with the appropriate range of coefficients. For e.g., if there were two independent variables named price and months on which the sale of car depended, the price range would be far too high because there are only 12 months in a year. In that case, scaling the variable price appropriately won't allow decimal errors to happen in the model. There are two types of scaling:

Normalized scaling: This scaling is done to make the distribution of data into a Gaussian one. It doesn't have a preset range. Typically used in Neural networks broadly.

Standardized scaling: The example given above is of standardized scaling. Here, the values of variable(s) is/are compressed into a specific range to suit the model.

The values of a normalized dataset will always fall between 0 and 1. A standardized dataset will have a mean of 0 and a standard deviation of 1, but the maximum and minimum values are not constrained by any specified upper or lower bounds.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

In the context of VIF, infinity represents perfect correlation, and as VIF values increase, the reliability of the regression results decreases. On the flip side, a low VIF indicates that the variable is relatively independent and doesn't suffer from multicollinearity concerns.

If there is a perfect correlation between the dependent variable and independent variable(s), the R-squared value comes out to be 1. Hence VIF, which is (1/(1-R^2)) turns out to approach infinity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans. Q-Q plots are also known as Quantile-Quantile plots. It is a graphical tool to assess if sets of data come from the same statistical distribution. It is particularly helpful in linear regression when we are given testing and training datasets differently. In this scenario, it becomes important to check whether both the data comes from the same background, in order to maintain the sanity of the model.

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian distribution, uniform distribution, exponential distribution or even a Pareto distribution. We can tell the type of distribution using the power of the Q-Q plot by looking at it.