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?**

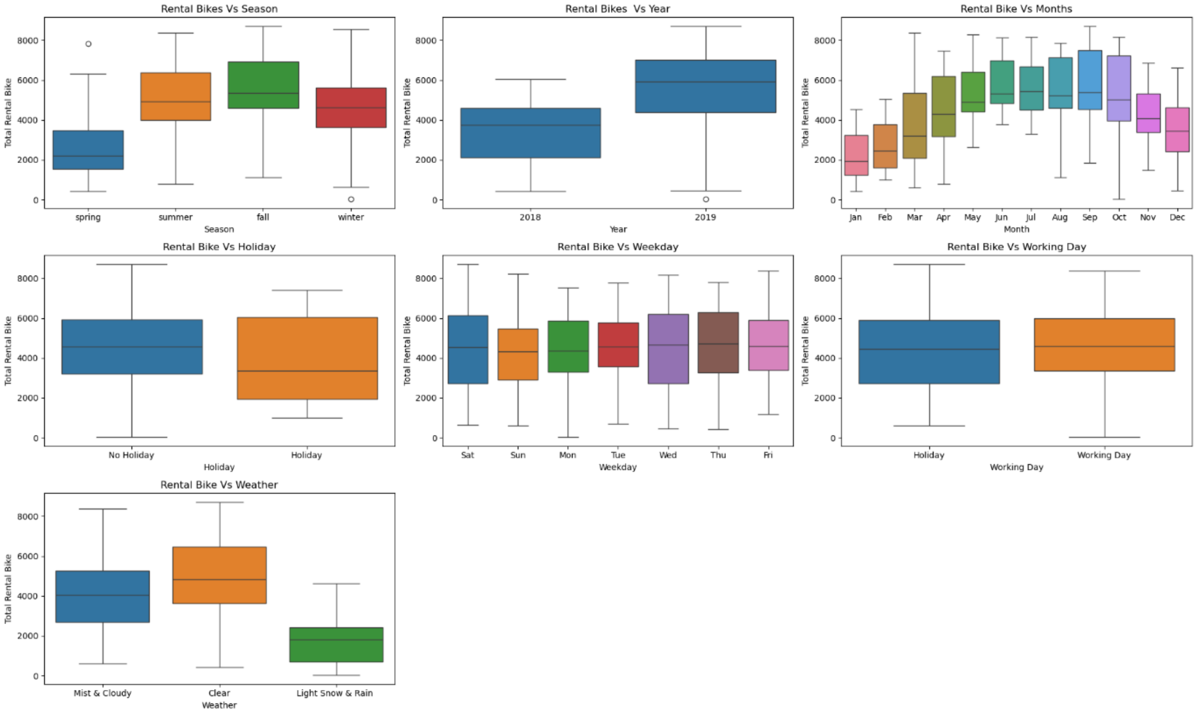
   There are more Rental bikes rides in summer and fall season compared to winter and lowest in spring

   Total Rides have increased significantly in 2019 compared to 2018

   Bike rides peaks in month from Aug-Sep-Oct

   There are no Bike Rides in Heavy Rain season

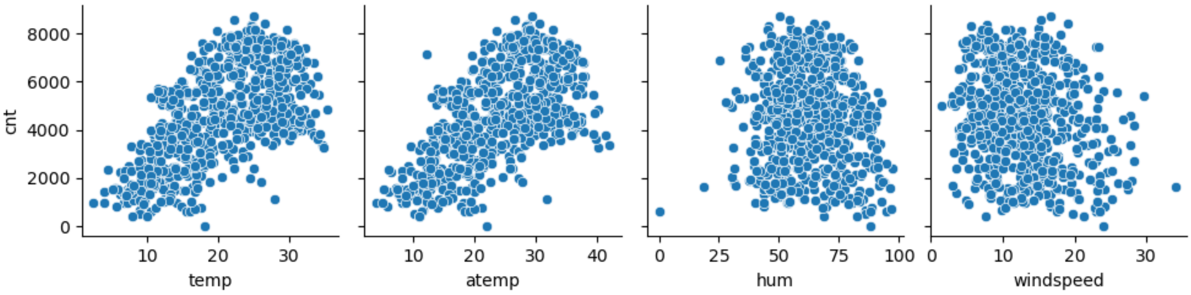
   Median of bike rides remain same across weekday



2. **Why is it important to use drop_first=True during dummy variable creation?**
   The categorical variable has n levels it can be represented by n-1 , if a categorially variable has 3 levels {A,B,C} we can define it by using only 2 dummy variables {B,C} when value for this two is 0 across row then it means the value is A

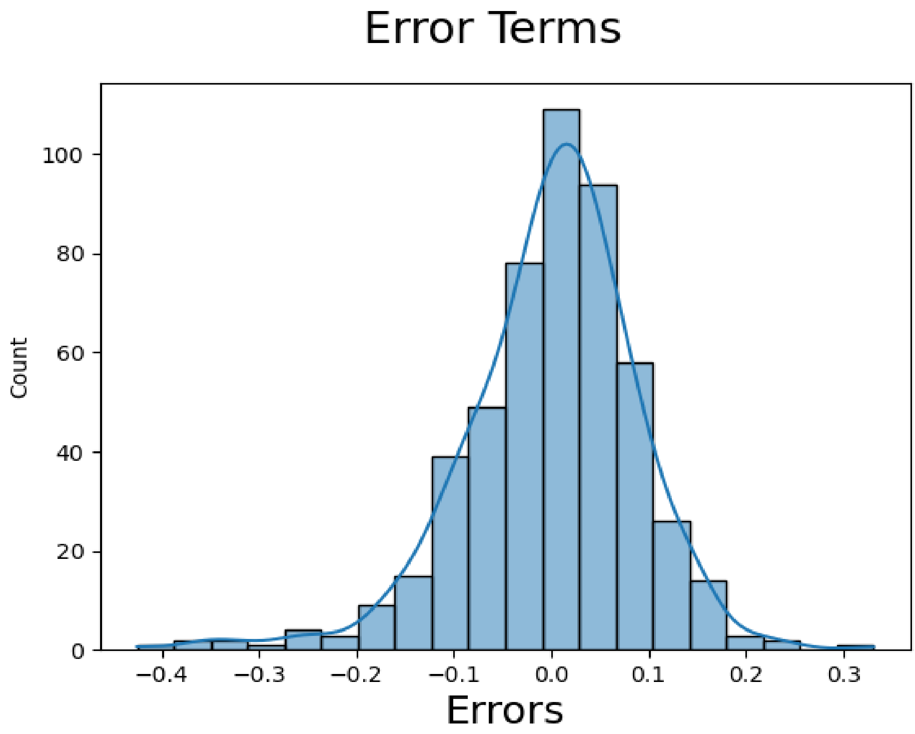
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   By Looking at plot we can see target variable has highest Correlation with temp and atemp fields



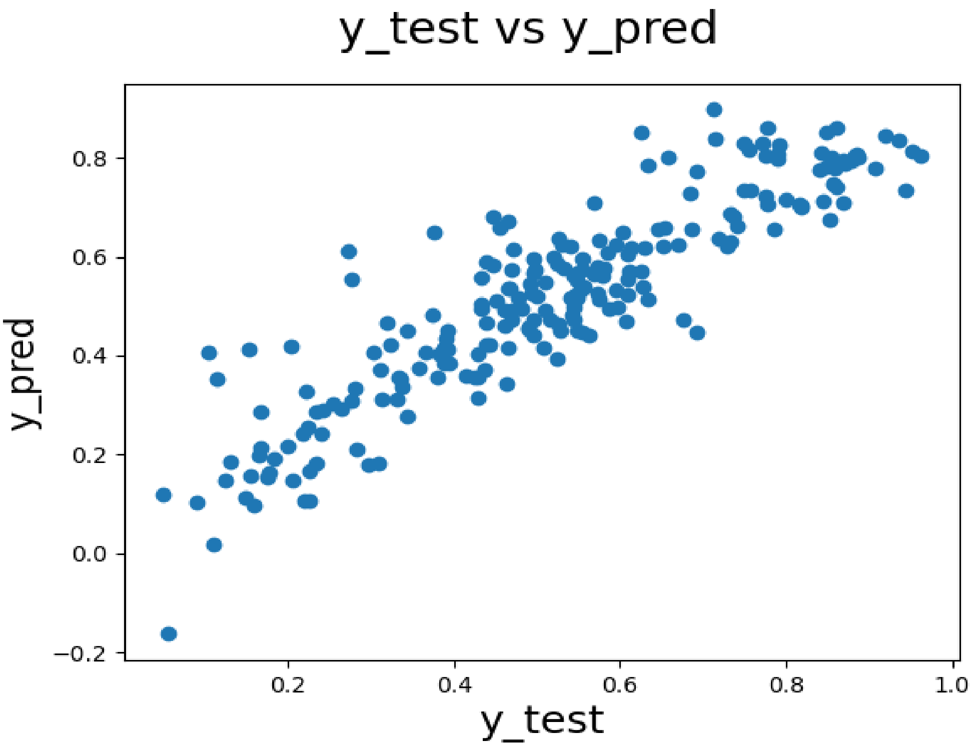
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

I have checked the distribution of residuals and it is a normal distribution centred around 0



I have checked VIF to ensure there is no multicollinearity in predictor variables

I have checked linearity by plotting actual vs predicted values



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

   Based on Final model below are top 3 predictors

   - Temp – Positive correlation
   - Weather Light Snow and Rain – Negative Correlation
   - Yr – Positive Correlation

General Subjective Question

1. **Explain the linear regression algorithm in detail**
   The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight-line

   The standard equation of the regression line is given by the following expression: $Y = \beta_0 + \beta_1 X$
   Y – Dependent Variable (Target Variable)
   X - Independent Variable (Predictor)
   β_0 - Intercept
   β_1 - Slope

   In multiple linear regression (with more than one independent variable), the equation extends to:

   $y=\beta0 + \beta1x1 + \beta2x2 +\cdots+ \beta nxn +\epsilon y$

   The objective of linear regression is to estimate the coefficients β0, β1,…,βn such that the line (or hyperplane) minimizes the difference between the observed and predicted values of y. This difference is known as the **residual** or **error**.

   Assumptions of simple linear regression are:
   1. Linear relationship between X and Y
   2. Error terms are normally distributed (not X, Y)
   3. Error terms are independent of each other
   4. Error terms have constant variance (homoscedasticity)

2. **Explain the Anscombe's quartet in detail.**

   Anscombe's quartet is a collection of four distinct datasets that have nearly identical statistical properties but very different distributions and visual patterns when plotted.

   Key Insights from Anscombe's Quartet

   1. **The Importance of Visualization**: Summary statistics such as the mean, variance, correlation, and regression equations may not fully capture the nature of the data. Visualizing the data through scatter plots reveals patterns and outliers that could be hidden when relying only on numerical summaries.

   2. **Dangers of Relying Solely on Statistics**: By only looking at the identical statistical properties, one might incorrectly assume that the datasets behave similarly. However, visual examination exposes the vastly different patterns present in each dataset.

3. **What is Pearson's R**

Pearson's R, also known as the Pearson correlation coefficient or simply correlation coefficient, is a measure of the linear relationship between two continuous variables. It quantifies how strongly the two variables are related and the direction of the relationship. The value of Pearson's R ranges from -1 to +1

- R = 1: Perfect positive linear correlation (as one variable increases, the other increases proportionally).
- R = -1: Perfect negative linear correlation (as one variable increases, the other decreases proportionally).
- R = 0: No linear correlation (no linear relationship between the variables).

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

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent methods

You can scale the features using two very popular method:

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

   X = X – mean(X) / sd(X)

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

   X = X – min(X) / max(X) - min(X)

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

The Variance Inflation Factor (VIF) is a metric used to detect multicollinearity in regression models. It quantifies how much the variance of a regression coefficient is inflated due to collinearity with other variables.

VIF(Xi) = 1 / 1−Ri^2

If Ri^2 = 1 then:

VIF(Xi)=1 / 1−1 = 1 / 0 = ∞

This situation occurs when there is perfect multicollinearity between a variable and the other independent variables, meaning that one variable is a perfect linear combination of the others.

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It helps visualize whether the data follows a specified distribution by plotting the quantiles of the dataset against the quantiles of the theoretical distribution.

Structure of a Q-Q Plot

- The x-axis represents the theoretical quantiles (e.g., the quantiles from a normal distribution).
- The y-axis represents the sample quantiles (the quantiles of the observed data).

If the data follows the specified distribution (e.g., a normal distribution), the points on the Q-Q plot should lie along a 45-degree line (the line y =x). Deviations from this line indicate departures from the assumed distribution.

Use of Q-Q Plot in Linear Regression

In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. This assumption is important for making valid inferences about the model parameters, constructing confidence intervals, and performing hypothesis testing.

The Q-Q plot helps assess this assumption by comparing the distribution of the residuals to a normal distribution.