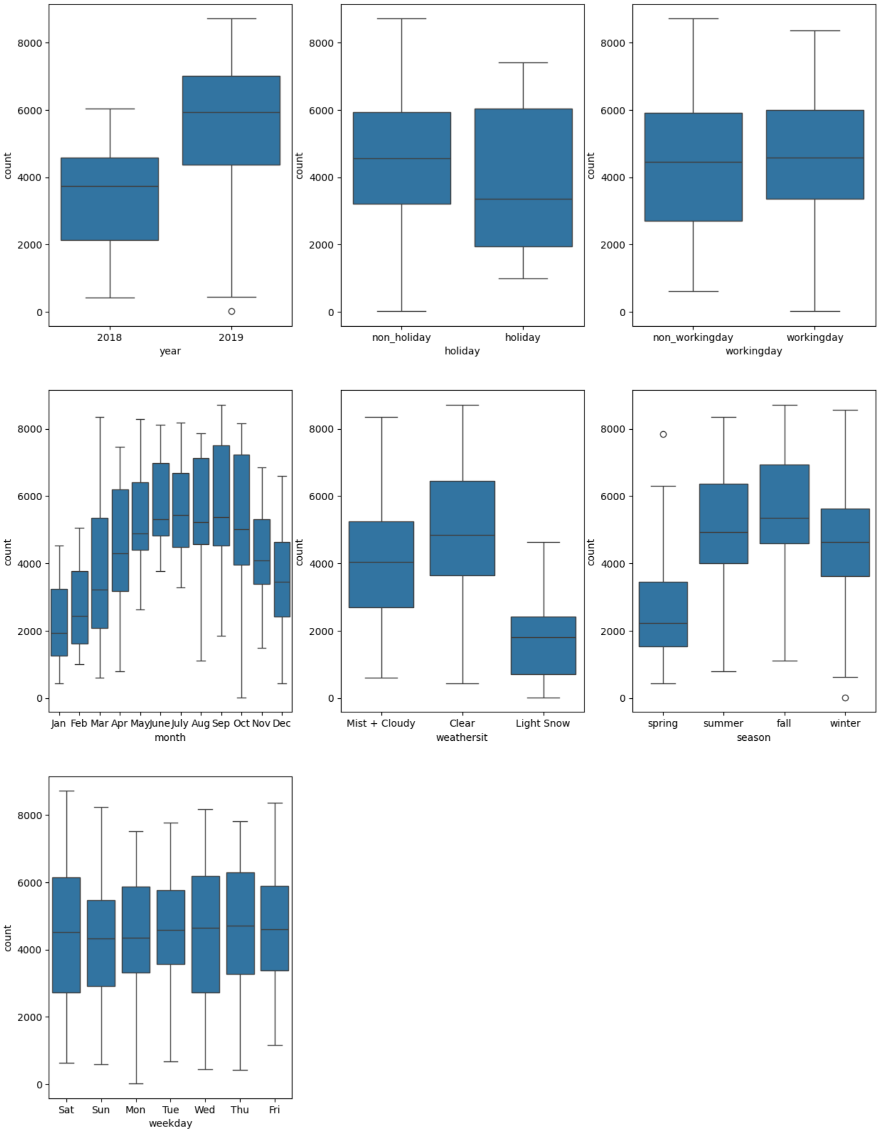
Linear Regression Assignment

Subjective question

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?**

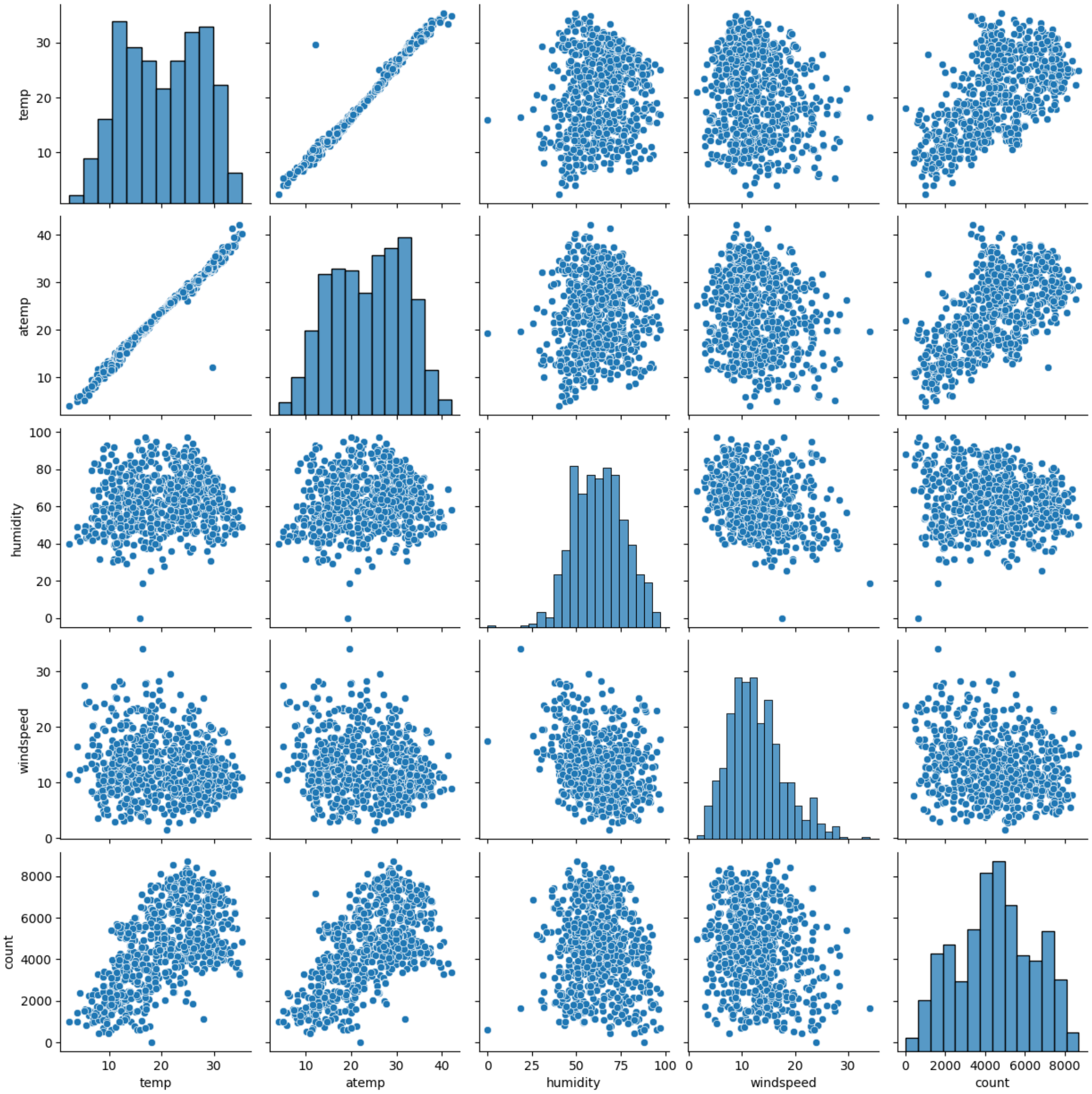


As above figure, we see more bike rentals on clear weather, on fall, from August to October and in 2019 and less rentals on Sunday.

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

If the categorical variable has k level, cretae k-1 dummy variables, so we need to use drop_first=True. If not I get trouble with multicollinearity.

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



As above figure, the temperature 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?**

I validate the assumption of Linear Regression after building the model on training set:

- Check the various assumption: p_value, VIF, normality of error
- Check the Adjusted R_square and R_square

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

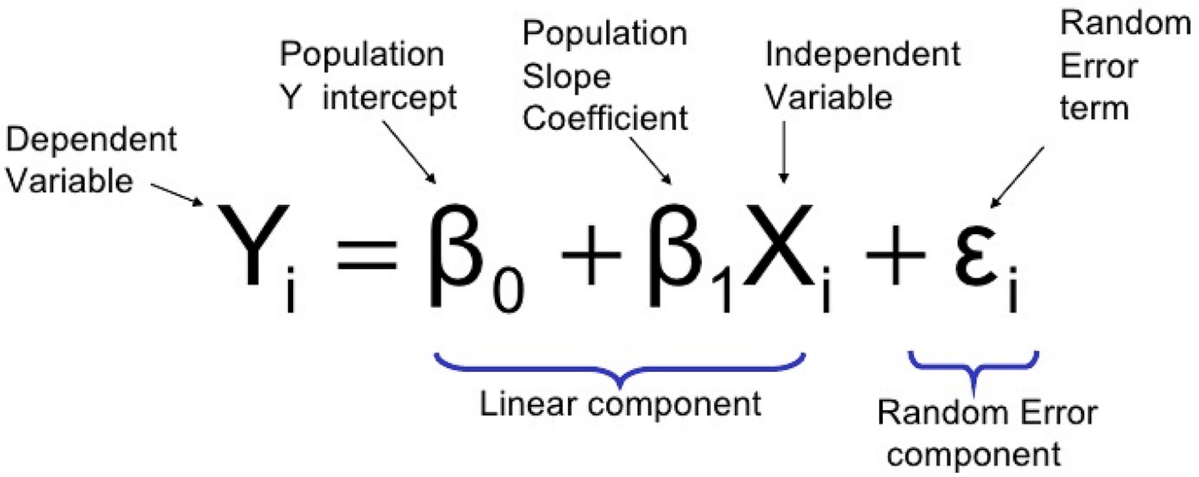
Top 3 features contributing significantly towards explaining the demand of the shared bike are: temperature(temp), year and season (particularly, winter)

General Subjective Questions

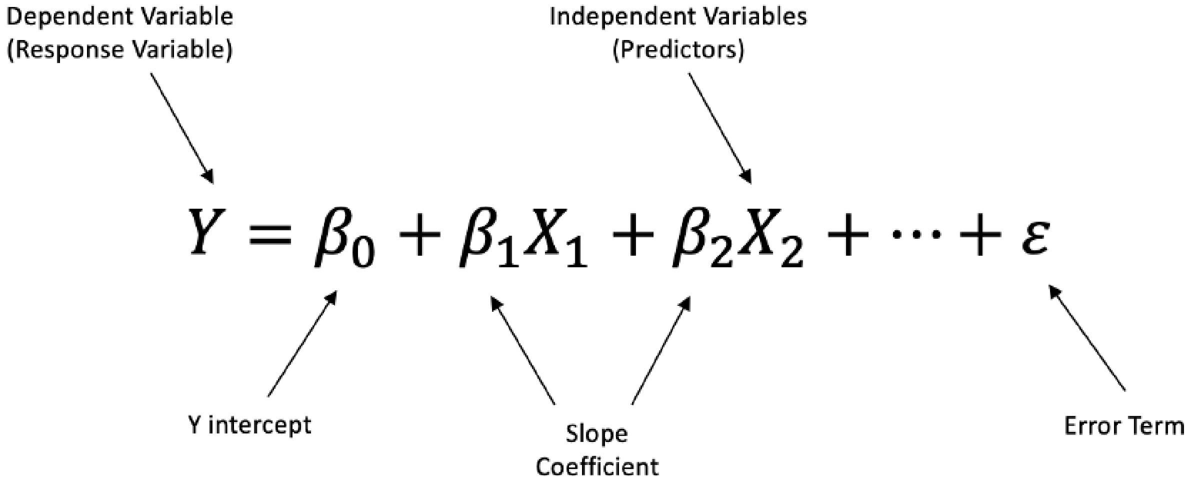
1. **Explain the linear regression algorithm in detail**

Linear regression is a supervised machine learning algorithm used for predicting a continuous target variable based on one or more input features. It assumes a linear relationship between the input variables (also called independent variables or features) and the target variable (also called the dependent variable). The aim is to find the best-fitting linear equation to describe the relationship between the input features and the target variable. We have simple linear regression (SLR) and Multiple Linear Regression (MLR):

- SLR: One independent variable



- MLR: More than one independent variable


Dependent Variable (Response Variable)
Independent Variables (Predictors)
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$
Y intercept
Slope Coefficient
Error Term

- Loss function: The primary objective is to find the values of B that minimize the difference between the predicted values (based on the linear equation) and the actual target values. This is typically done by minimizing the sum of squared differences between predicted and actual values, known as the "least squares" method.
- Evaluation: The model's performance is evaluated using various metrics like Mean Squared Error (MSE), Root Mean Squared Error

(RMSE), R-squared (coefficient of determination), Adjusted R_squared, etc., to assess how well it fits the data and generalizes to new unseen data.

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

Anscombe's quartet refers to a set of four small datasets that have nearly identical simple descriptive statistics (such as mean, variance, correlation, and linear regression parameters), yet exhibit markedly different characteristics when graphed.

The quartet consists of four distinct datasets, each containing 11 (x, y) paired observations:

- Dataset I: This dataset forms a perfect linear relationship.
- Dataset II: It looks like Dataset I when using summary statistics, but when plotted, it shows a non-linear relationship between variables.
- Dataset III: It shows a strong linear relationship but is heavily influenced by an outlier, which affects the regression line significantly.
- Dataset IV: It shows no linear relationship between variables but has a high correlation coefficient, demonstrating the impact of a single outlier on the correlation.

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

The Pearson's R (Pearson correlation coefficient) is a statistical measure indicating the degree of linear correlation between two data sets. It's calculated as the ratio between the covariance of the variables and the product of their standard deviations. This normalized measurement always yields a value between -1 and 1, representing the strength and direction of the linear relationship. However, it specifically captures linear correlations and disregards other relationship types. For instance, in a sample of high school teenagers, one would anticipate a PCC between their age and height to be notably higher than 0 but less than 1, as 1 would signify a perfect yet improbable correlation.

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

- What is scaling: Scaling is a preprocessing technique used in machine learning to standardize or normalize the range of independent variables or features in the dataset
- Scaling is performed because it helps in improving the performance and convergence of certain algorithms and ensures that no specific feature dominates solely because of its larger magnitude.
- The difference between normalized scaling and standardized scaling are:
    - Normalized Scaling: Rescales features to a range between 0 and 1, preserving relationships but sensitive to outliers.
    - Standardized Scaling: Centers the data around mean 0 and standard deviation 1, making it less sensitive to outliers and providing a standardized interpretation.

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

$$VIF_i = \frac{1}{1 - R_i^2}$$

When the value of VIF becomes infinite, an extreme case of perfect multicollinearity. In other words, one predictor variable is a perfect linear function of others.

In a perfect fit, R_square=1. Consequently, the denominator in the VIF formula becomes zero , leading to a division by zero and resulting in an infinite VIF value.

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

- A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to assess if a dataset follows a certain theoretical distribution, such as the normal distribution. It compares the quantiles of the dataset's distribution against the quantiles of a theoretical distribution
- How to use:
    - Arrange the data in ascending order.
    - Calculate the quantiles of your dataset.
    - Obtain the theoretical quantiles from the chosen distribution for the same probabilities as the dataset's quantiles.
    - Plot the dataset's quantiles against the theoretical quantiles on a scatter plot.
- Q-Q plot in linear regression helps to visually inspect whether the residuals conform to a normal distribution, which is an essential assumption for valid inference in regression analysis.