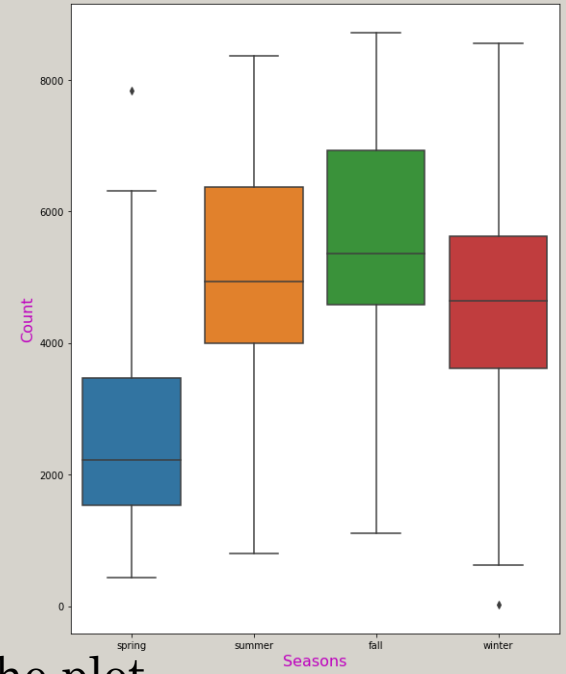
Subjective Questions

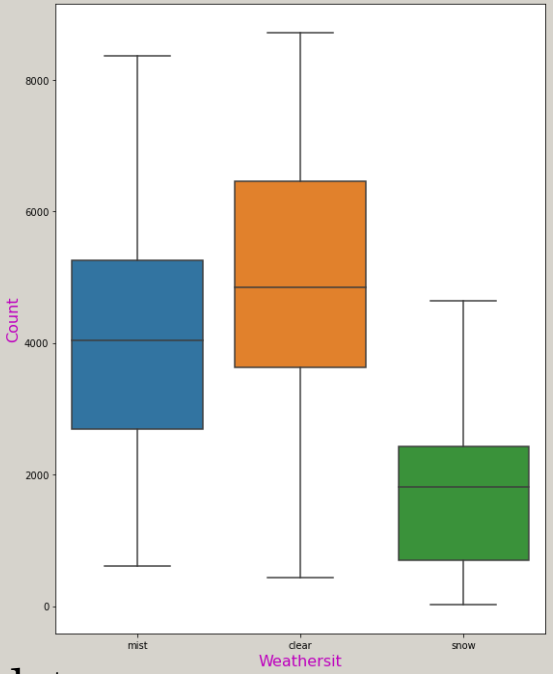
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



As per the plot,
the count of bike rentals depends on the seasons.

The count varies from maximum to minimum in the following order,

1. **Fall**
2. **Summer**
3. **Winter**
4. **Spring**



As per the plot,
the count of bike rentals depends on the seasons.

The count varies from maximum to minimum in the following order,

1. **Clear**
2. **Mist**
3. **Snow**

There is no data related to rain

2. Why is it important to use drop first=True during dummy variable creation?

- The data in the categorical variable, cannot be used as such since the model equation will not be having categorical variable and only numerical value will be used.
- So the categorical variables are to be converted into numerical variable
- For converting the categorical variable to numerical variable, dummy variables are generate in place of existing categorical variables.

- **If n number of levels are present in the categorical variable, (n-1) number of dummy variables are to be added.**
- **The logic behind this is that all the n levels in the variable can be specified with the combination of n-1 dummy variables.**

- For example, for a 2 level categorical variable, male and female, one dummy variable can be used to specify both levels, such that 0 will be used for male and 1 will be used for female.

- **So not removing the first column in the dummy variable will confuse the model, making it considering n+1 levels in categorical variable**
- **That's why it is necessary to remove the first variable in the dummy variables**

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

The variable with highest correlation with the target variable is as below,

1. **Registered**
2. **Casual**
3. **Temp**
4. **Atemp**

As the target variable is the summation of registered and casual, the correlation between both are highest, and as per the above reason, both the data (registered and casual) are to be removed.

Next highest is the correlation of target variable with temp and atemp which are both similar, And as per the correlation between temp and atemp, it is evident that both are the same data, with minor variation between both.

Thus, we have to keep only one of the data, to eliminate multi-collinearity from the data.

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

Assumption 1 :
Linear dependency: The dependent variable should have linear relationship with the independent variables.

Validate: **Scatter plot between X_train and the y_train showing relation between them.**

Assumption 2:
Residuals are independent of each other. No correlation between consecutive error terms

Validate: **Scatter plot between error term and residuals showing no pattern**

Assumption 3:
Independent variables are not correlated with each other

Validate: **Variable Inflation Factor should be less than 5**

Assumption 4 :
Error terms are normally distributed with mean 0

Validate: **Histogram of error values showing normal distribution with mean 0**

Assumption 5 :
Homoscedasticity – Residuals have constant variance

Validate: **Plot residuals Vs predicted values scatter plot showing error values are having constant variance**

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

The following are the top 3 contributors of the model,

1. **Atemp**
    - Major contributor with coefficient of 3801, meaning that target variable is **positively dependent on temperature**
    - This signifies that with the **feeling of high temperature, people prefer to use bike**

2. **Snow**
    - Snow having coefficient of -2437, meaning that target variable is **negatively dependent on snow**
    - This signifies that **with snow people are not preferring to use bike**

3. **Year**
    - Year having coefficient of 2061, meaning that target variable is **positively dependent on year**
    - This signifies that **with increase in the popularity, people prefer to use bikes more**

General Subjective Questions

1) Explain the linear regression algorithm in detail.

- Linear regression is the method of obtaining a statistical model equation with all the contribution variables, in the format of y = c + m1x1 + m2x2 + …. + mnxn
- The logic behind this is the dependent variable is linearly dependent on the independent variables

- **Types:**
  - Single linear regression – Using one independent variable
  - Multiple linear regression – Using more than 1 independent variable

- **Approach:**
  - Forward selection : Considering one variable at a time and adding to the model
  - Backward elimination : Considering all the variables in the model and removing the variables which are not significant

- **Model development :**
  - Model can be developed by the above mentioned methods
  - Train and test data can be divided for verification of the developed model

- **Model:**
  - The developed model can be used for finding the value of the target variable provided that the values of independent variables are given.

2) Explain the Anscombe's quartet in detail

- Anscombe's Quartet is a group of **four data sets** that are nearly identical in simple descriptive statistics
- But there are some differences in the dataset that fool the build regression model.
- In scatter plot, different sets will have different distributions and will appear differently
- The developed model can be only be considered a fit for the data with linear relationships but for the data with outliers will not be handled well by the developed model
- This indicates the importance of visualizing the sets

3) What is Pearson's R?

- Pearson's R will be used for finding the correlation between two data sets.
- This shows the **linear relationship between two sets of data**
- The value of pearson's correlation is between **-1 and 1**
- If the value is 1, the two data sets are positively strongly correlated
- If the value is -1, the two data sets are negatively strongly correlated

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

Scaling:
The process of converting the high range of vales to scaled values is called scaling. This is used when there are multiple data in different ranges

Purpose:
Scaling is done in the linear regression for the following two reasons,

a. Reduction in time for attaining the linear model
b. Obtaining simpler model

Types:

1. Normalizing
2. Standardizing

Standardizing :
Standardizing will Convert the values between (-sigma) to (+sigma) with center as mean

Normalizing:
Normalizing will convert the values between 0 to 1, by using the min and max values in the data

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

The value of VIF will be infinite when there is **perfect correlation between the variables**.

For example, in our data set, the values of temperature was given and the values of feeling temperature was also given.
The value of feeling temperature will be more or less same with the value of temperature, with minor variation.
This will cause the VIF value between them to be infinite.

High value of VIF indicates that there is multi-collinearity in the model, and the resulting model with the multi-collinearity will not be a model capturing all the variations in the dataset.

Thus, the column with high value of VIF should be **removed** from the data set for attaining a better model.

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

- Q-Q plot is called as Quantile – Quantile plot
- This plot is used to plot the quantiles of the sample distribution with the quantiles of population.
- This is used to determine the **distribution followed by the dataset**.
- This is useful in determining the following,
  - If two dataset are of the same distribution
  - Can determine whether the error is following the normal distribution
  - Can also identify the skewness of the distribution