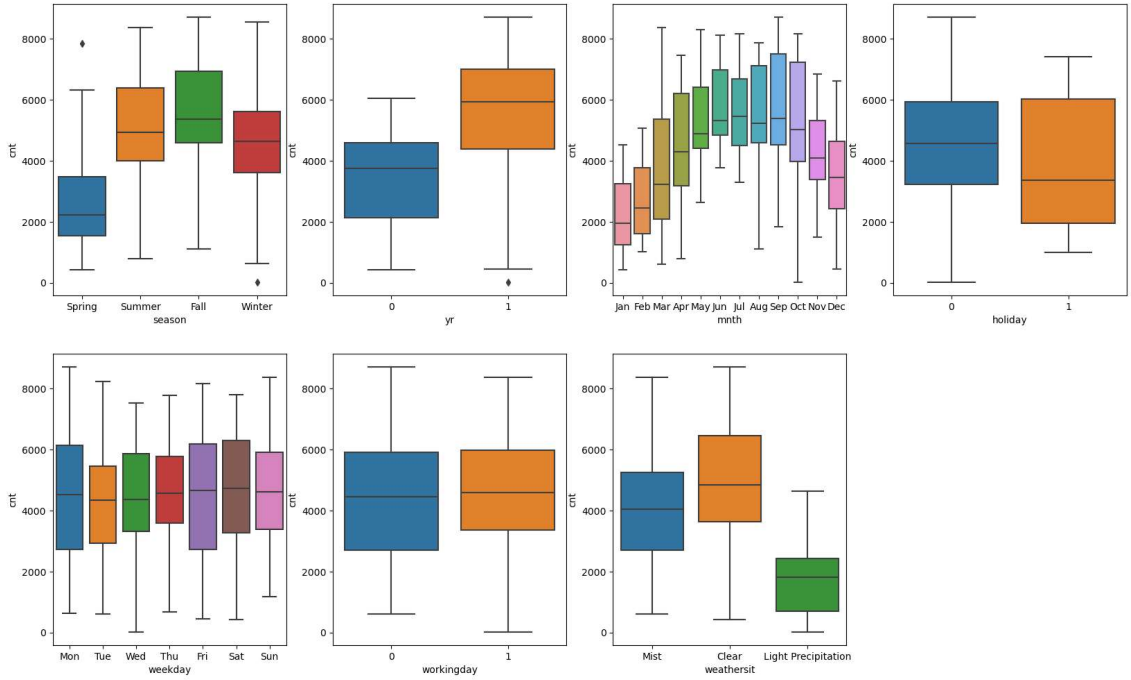
Assignment-based Subjective Questions

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

   Basing on the data the following categorical variables are identified, viz.,

   Season: Season
   Yr: Year
   Mnth: Month
   Holiday: Holiday
   Weekday: Weekday
   Workingday: Working Day
   Weathersit: Weather Situation

   Boxplot categorical_variable vs count:



   Observations:

   a. Season: Highest bike demand is in Fall and lowest is in Spring
   b. Year: 2019 has seen more demand than 2018
   c. Month: Demand has varied much across months, with highest in September, while lowest in January. With an consistent demand from May to October
   d. Holiday: Demand on non-holiday vs holiday
   e. Weekday: No such noticeable difference
   f. Working Day: No such noticeable difference

g. Weather Situation: High demand in Clear weather, with the lowest on Light Precipitation and no demand at all on High Precipitation

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

It is important because it helps to avoid multicollinearity issues and prevents the "dummy variable trap."

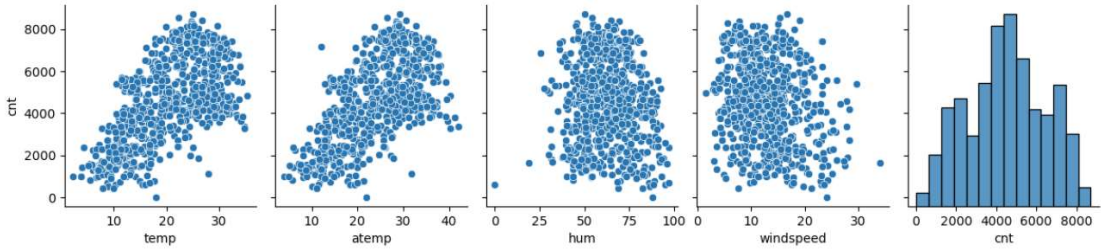
Multicollinearity: Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated. If the dummy variables are created without dropping one-level, these dummy variables will be perfectly correlated because they can be derived from each other, which can cause problems in regression analysis, including unstable coefficient estimates.

The Dummy Variable Trap: The dummy variable trap is a specific case of multicollinearity. It occurs when dummy variables aren't dropped one-level. Because of which the model cannot distinguish the effect of the omitted category from the intercept term. As a result, the coefficients for the remaining dummy variables become collinear, and it becomes impossible to isolate the unique contribution of each category.

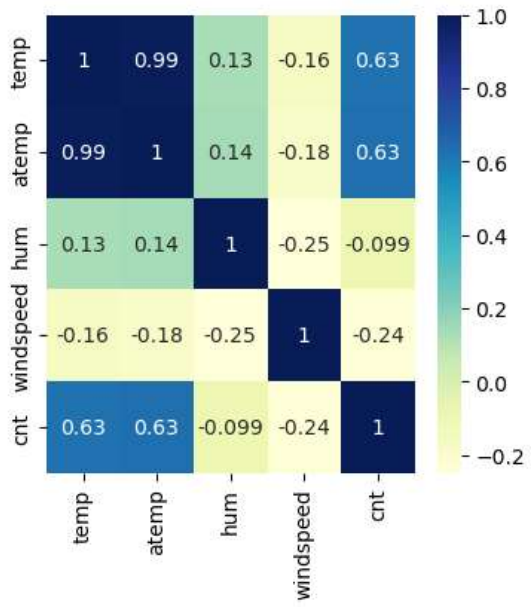
By setting drop_first=True, we are dropping dummy variables to one-level less, i.e., n-1 levels.

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

Looking at the pair-plot temp and atemp have higher correlation with cnt (count, which is the target variable).



To be more precise we need to look at correlation heatmap, which tell us temp and atemp are the most correlated with a value of 0.63.



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

   Validating the assumptions of linear regression after building the model can be done as follows,

   1. **Residual Analysis:**
      Computing the residuals, i.e., the differences between the actual and predicted values for the training data and then plotting a dist-plot with KDE curve to see a roughly symmetrical bell-shaped distribution exists.

   2. **Linearity:**
      Creating a scatterplot of the predicted values against the actual values to visually check for linearity. The points should be roughly aligned around a 45-degree line.

   3. **Homoscedasticity (Constant Variance):**
      Create a plot of residuals vs. fitted values. The spread of residuals should be roughly constant across all levels of the predicted values. If the spread widens or narrows systematically, it's an indication of heteroscedasticity.

   4. **Independence of Errors:**
      Examine autocorrelation plots of residuals (e.g., ACF and PACF plots) to check for any patterns or correlations in the residuals over time, especially if the data has a time series component.

5. **Collinearity:**
   Calculating the Variance Inflation Factor (VIF) for each predictor variable to check for multicollinearity. High VIF values (typically greater than 5) may indicate collinearity issues.

6. **Cross-Validation:**
   Perform cross-validation on the training set to assess the model's predictive performance on unseen data. Cross-validation helps identify overfitting and generalization issues.

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

The below are the top 3 contributing factors explaining the demand of shared bikes,

#	Feature	Coefficient	Remarks
1	Light Precipitation	-0.3108	Negative Impact
2	yr	0.2474	Positive Impact
3	Spring	-0.1906	Negative Impact

General Subjective Questions

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

Linear Regression is a machine learning algorithm, which comes under Supervised Machine Learning, to model the relationship between dependent (target) and one or more independent variables (features) by fitting a linear equation.

Model Assumptions: It assumes a linear relationship between the independent and dependent variable.

a. **Simple Linear Regression (SLR):** If there is a single independent variable, then its called Simple Linear Regression.

Equation: Y = β0 + β1 * X

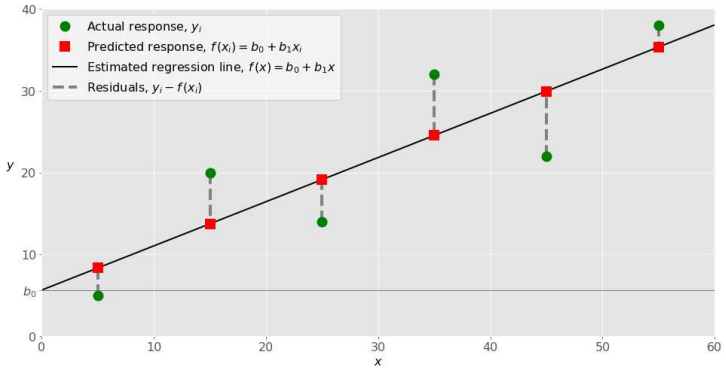
Where,

Y = Target

X = Single feature

β0 = Intercept

β1 = Slope



Metrics:

i. **R-squared:** It is the coefficient of determination that measure the proportion of the variance in the dependent variable. It can range from 0 – 1.

R^2 = 1 – (RSS / TSS)

Where,

RSS = Residual Sum of Squares

TSS = Total Sum of Squares

ii. **Residual Standard Error:**

It quantifies the average amount by which the ovserved values deviate from the predicted values.

$$RSE = \sqrt{\frac{1}{\text{n-2}}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where,

n = number of data points

y_i = actual observed value of i-th data point

$\hat{y}_i$ = predicted value of i-th data point

b. Multiple Linear Regression (MLR): If there are more than one independent variables, then its called Multiple Linear Regression.

Equation: Y = β0 + β1 * X1 + β2 * X2 + ... + βn * Xn + ε

Where,

Y = Target

X = Single feature

β0 = Intercept

β1 = Coefficient of 1st variable

βn = Coefficient of nth variable

Metrics:

i. **Adjusted R-squared:**

It's a modified version of R-squared, penalizing the inclusion of irrelevant variables.

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Where,

N = count of data points

P = predictors count

j. **Variance Inflation Factor (VIF):**

It measures the extent of multicollinearity among the independent variables. Typically >5 is considered bad and could be a potential candidate for dropping.

$$VIF = \frac{1}{1 - R^2}$$

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

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

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

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

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

7.