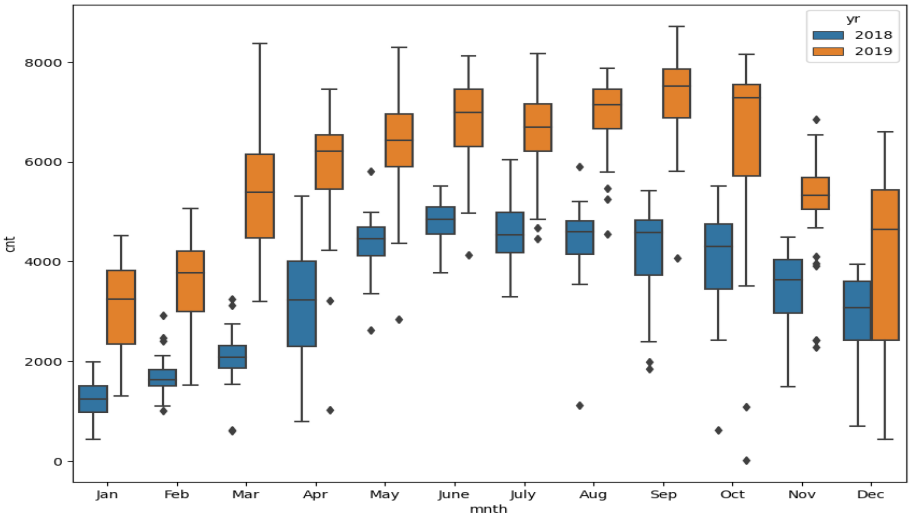
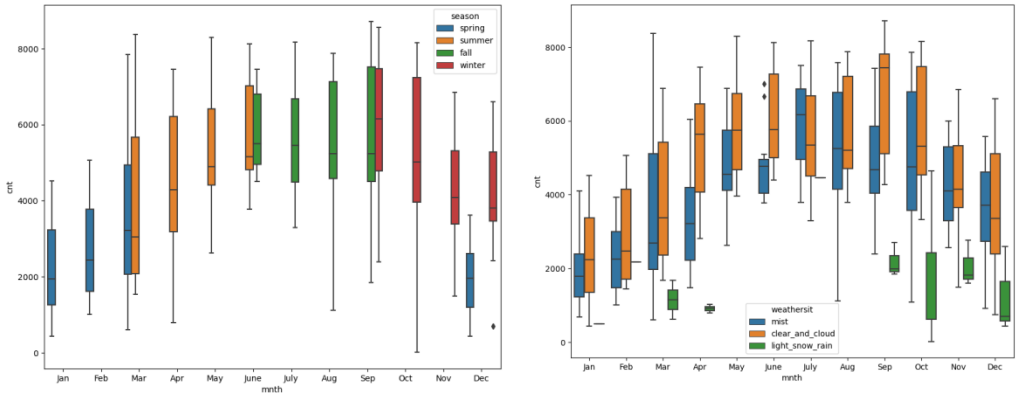
Assignment-based Subjective Questions

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

From the analysis it clear that,



- From the above image, it is evident that, when comparing 2018 and 2019 demand, the demand was much less in 2018 than 2019, as 2018 might be the starting point of BoomBikes



- Especially, the median count tend to increase from Apr to Sep, and continuously falls through the following months, similar pattern was found in both years, as it falls under summer and fall season, which also accounts to clear clouds with no or very less rainfall and/ or low mist
- Rain has a negatively affected the demand for bikes
- Weekdays does not have a much effect, as median remain almost same in all the categories, when coming to holidays it does not follow much difference, but non holiday count overtakes a holiday count by a small margin

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

- drop_first=True, eliminates the first category that is available according to the dataset while creating a dummy variables
- It is used to remove redundant column in the data, which allows us to increase the efficiency of the model
- It also reduces a correlation created among the dummy variables (multicollineraity)

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

Among all the variables given in the dataset, the highest correlated one is atemp (63.07%), also temp which is approximately equal to 63% (62.70%).

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

These are the assumptions that are used to validate Linear regression algorithms

1. Simple Linear Regression
    a. There is a linear relationship exists between independent variable and dependent variable
    b. Error terms (residuals) are normally distributed
    c. Error terms (residuals) are independent to other independent variables
    d. Error terms (residuals) have common variance
2. Multiple Linear Regression
    a. All the assumptions that are validated in Simple Linear Regression also used here (1.a, 1.b, 1.c, 1.d)
    b. Model fits a hyper plane instead of a line
    c. Multicollinearity
    d. Model may overfit, by becoming over complex (Feature Selection is used)

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

From the final model below are the ones with the highest relationship,

1. Temperature
2. Year
3. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds ==> light_snow_rain (Dataset notation) – is negatively correlated with the demand

General Subjective Questions

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

Linear regression predicts the relationship between two variables by identifying the linear relationship between independent variable and dependent variable. The best fit is identified by minimising the sum of squared errors between the dependent variable and predicted variable. Multiple variables can also be used to predict a dependent variable which is known as Multiple Linear Regression.

- Simple Linear Regression:
    - In Simple linear regression, there will be only one independent variable and dependent variable. This model estimates the slope and intercept of the line of the nest fit,
    - $\hat{Y} = Y_i = \beta_0 + \beta_1 X_i$
    - Where β_0 is the intercept of the line, β_1 slope of the variable along the line,
    - The goal of the linear regression is to find the best values for β_0, β_1 to find the best find of the line. The difference between the dependent variable and predicted variable is called the residuals,
    - Residuals = $\varepsilon_i = y_{pred} - y_i$, Which can also be called as loss function
    - In Regression, Mean Squared Error Cost function helps to get the best values for β_0, β_1, where MSE is the mean sum of squared errors,
    - MSE = J = $1/n \sum_{i=1}^{n}(Y_i - Y_{pred})^2$, Gradient Descent is used to minimise the cost function J

    - Assumptions that made in the SLR are
        - There is a linear relationship exists between independent variable and dependent variable
        - Error terms (residuals) are normally distributed
        - Error terms (residuals) are independent to other independent variables
        - Error terms (residuals) have common variance
- Multiple Linear Regression:
    - In this regression, a dependent variable is predicted by multiple predictors
    - $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n$
    - MLR, follows the same steps as in Linear Regression to find the optimized line, but it add more assumptions that needs to be validated while creating the model,
    - Assumptions that made in the MLR are
        - All the assumptions that are validated in Simple Linear Regression also used here
        - Model fits a hyper plane instead of a line
        - Multicollinearity
        - Model may overfit, by becoming over complex (Feature Selection is used)
- Evaluation: R-Squared is the best measure that is used to validate the best fir for the model, higher the value of R-squared, the best fit is the model,
    - R-squared = 1-(RSS/TSS)
    - Where RSS is the Residual sum of squares, TSS is the total sum of squared

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

Anscombe's quartet comprises of four set of datasets, having same mean, standard deviation, R-squared and variance. It is used to illustrate the significance of exploratory data analysis and limitations of depending only on summary statistics. This suggests the data features must be visualized to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

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

Pearson's correlation coefficient measures the linear relationship between a dependent and independent variable(s). It is a number between -1 and 1, which tells us about the direction and also measures how strongly they are correlated. If the value is between -1 and <0, the variables are negatively correlated and moving in the negative direction and vice versa. If it is 0, there is no correlation exists between the variables.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

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

Scaling is a method that is used to scale the values from to a specific range (changing the range of values). It is also a important step in a machine learning process, to improve the speed/ efficiency of the model by making the values smaller to compute.

The main difference between normalized scaling and standardized scaling is that, the former changes the both the shape and range of the data [0, 1], whereas in the latter only the range of the values are changed by making mean to 0, and standard deviation to 1 and the shape remains the same.

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

- Higher the VIF values, higher the correlation exists among independent variables,
- If there is a perfect correlation between the independent variables, the VIF values becomes infinity, thereby introducing a concept of Multicollinearity,
- If VIF is equal to 1, no relationship exists among predictors, if greater than 1, they are correlated to each other

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

Q-Q plot is used to compare two distributions, one distribution is usually the observed set of quantiles in the observed data and other is any one of the distribution usually normal distribution. To use Q-Q plot, we need to check the shape and pattern of the points that are plotted. If the points are distributed along the line, then it follows the reference distribution closely, else it is deviated from the reference distribution. Q-Q plots are used in the regression models to check whether the assumptions are true, i.e. for example to check whether the residuals are normally distributed and have a constant variance (homoscedasticity)