Assignment-based Subjective Questions

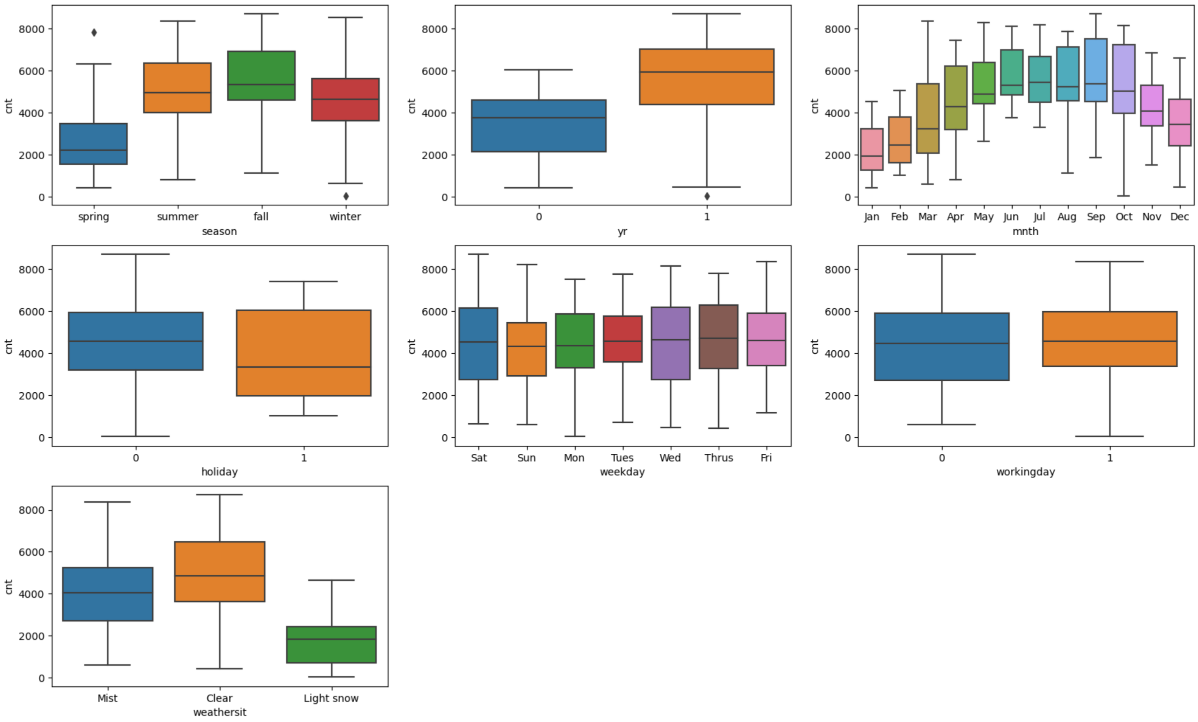
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: *Equation of best fitted line:*

$cnt=0.1907+0.2296\ast yr+0.0526\ast workingday+0.5684\ast temp-0.1643\ast hum-0.1943\ast windspeed+0.0629\ast Sat-0.0401\ast Jan-0.0429\ast Jul+0.0909\ast Sep+0.0765\ast summer+0.1251\ast winter-0.2425\ast Lightsnow-0.0538\ast Mist$

As per the above equation, categorical variables like year, workingday, Sat, Jan and other features are important factor while doing the analysis of dependent variable cnt.

For year, 2019 there were more bikes on rent as compared to 2018. Similarly, we can see trend of bike count based on season, month, weather situation from the below box plot.



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

Ans: When we create dummies using get_dummies method of pandas it creates n number of columns for a categorical column having n different values,

example: For season column there is 4 values (1:spring, 2:summer, 3:fall, 4:winter) and when we run “pd.get_dummies(bikes['season'])” it creates four columns “fall”, “spring”, “summer”, “winter”.

To understand and build the model we don’t need the 4 columns we can have “spring”, “summer”, “winter” columns and it will still make same sense as if all the column values are 0 it means that season is “fall”. Hence, we use **drop_first=True** during dummy variable creation.

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

Ans: “temp” and “atemp” both are showing the same correlation with the target variable “cnt”.

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

   Ans: To validate the assumptions of Linear Regression after building the model on the training set, we do Residual analysis of the train data. Plot the histogram of the error terms and check if the plot is normally distributed around zero mean value.

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

   Ans: Top 3 features significantly towards explaining the demand of the shared bikes are:

   1. Temp
   2. Year
   3. September

General Subjective Questions

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

   Ans: Linear regression algorithm provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. The dependent variable to be predicted is a continuous variable. This comes under supervised learning method.

   There are two types of linear Regression:

   1. **Simple Linear Regression:** A simple linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line.

      Equation: y = m*X + c

      Where, m = slope (change in y/change in x) and c = intercept (y value when x=0)

   There are below assumptions for simple linear regression:

   - There will be linear relationship between X and y
   - Error terms are normally distributed.
   - Error terms are independent of each other.
   - Error terms have constant variance.

   2. **Multiple Linear Regression:** Multiple linear regression is an extension of simple linear regression. It attempts to explain the relationship between a dependent variable with multiple independent variables using a straight line.

      Equation: y = m1x1 + m2x2 + m3x3 + …. + mnxn + c

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

   Ans: Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

   The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. What is Pearson’s R?

Ans: Pearson’s correlation measures the strength of the linear relationship between two variables. The value of Pearson’s R lies between -1 to 1.

Where,

- -1 meaning a total negative linear correlation.
- 0 meaning no correlation.
- +1 meaning a total positive correlation.

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

Ans: Scaling is a process of normalizing the range of features in the given dataset. To bring different scale data to a small range, scaling is performed.

Difference between normalized scaling and standardized scaling:

Normalized scaling: Normalized scaling brings all the data in the range of 0 and 1.

$$x = (x - \text{mean}(x)) / (\max(x) - \min(x))$$

Standardized scaling: Standardized scaling brings all the data into a standard normal distribution with mean zero and standard deviation one.

$$x = (x - \text{mean}(x)) / \text{sd}(x)$$

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

Ans: Variance Inflation factor (VIF) calculates how well one independent variable is explained by all other independent variables combined. To consider an independent variable in final liner regression model VIF value should be lower. Sometimes we might observer that the value of VIF is infinite, it means there is perfect correlation between the given independent variable with all other independent variable.

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

Ans: Quantile-Quantile (Q-Q) plot is a graphical tool which helps us in analysing if 2 given datasets, came from the populations with a common distribution.

In linear regression, when we have received train and test dataset separately, then we can use Q-Q plot to confirm if both the datasets are from population with same distribution.