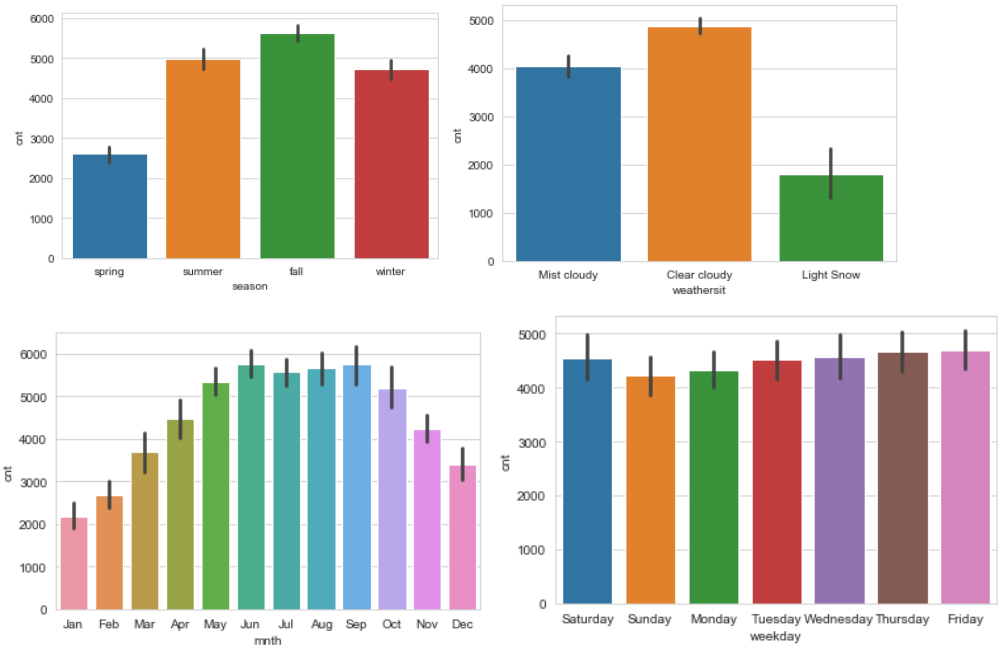
Assignment-based Subjective Questions:

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

Answer:

- After mapping the variables based on the data dictionary, there are four categorical variables.
  - season
  - mnth
  - weekday
  - weathersit



- We could observe that fall season and months of June to September in the weekday of Saturday, Thursday and Friday with the weather condition of Clear, Few clouds, partly cloudy has the significantly high demand of shared bikes and good time for business.

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

Answer:

- To get N-1 dummies out of N categorical levels by removing the first level and the first level becomes the base categorical variable.
- By using drop_first =True, the model becomes less complex by removing a redundant dummy variable.

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

Answer:

- The **temp** variable has the high correlation with the target variable cnt.
- There is a correlation between count of total rental bikes and temperature with the value of 0.627044 correlations.

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

Answer:

- 1. Finding the predicted values for the dependent variable.
- 2. Finding the Residual Error and plot and check for normal distribution.
- 3. Look for patters in residual's error (There should be no patters).

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

Answer:

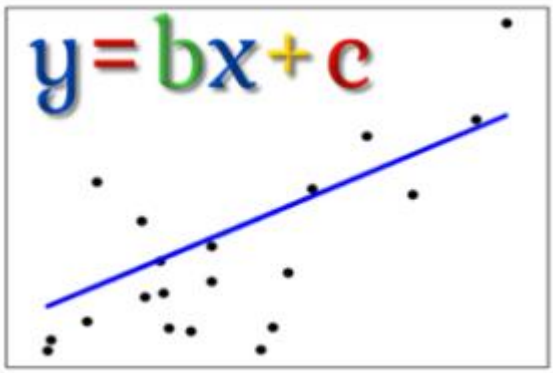
- From the final model the below list of independent variables are the top 3 features contributing significantly towards explaining the demand of the shared bikes.
    - temp (Temperature)
    - yr (Year)
    - mnth_Sep (In Month April is the base month)

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Answer:

- Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task.
- Regression means relation between independent variable and dependent variable. Regression models a target prediction value based on independent variables.
- The regression line is the best fit line for our model.



- So, the equation of regression line is

  Y=BX+C

  Whereas,
  X is independent variable
  Y is dependent variable/Target Variable
  C is the intercept
  B is the Coefficient of X

2. Explain the Anscombe's quartet in detail.

Answer:

- Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.
- Each dataset consists of eleven (x, y) points.
- Each graph tells a different story irrespective of their similar summary statistics.
- The summary statistics show that the means and the standard deviation were identical for x and y across the groups and also the correlation coefficient of each group.

3. What is Pearson's R?

Answer:

- Pearson's R measures the strength of the linear relationship between two variables.
- The best way to check the Pearson's R is visually by using scatter plot.
- If the variables tend to go up and down together, the correlation coefficient will be positive and if the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
- Pearson's R is always between -1 and 1.
- If R = 0 means there is no linear association.

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

Answer:

- Scaling is a technique to standardize the independent features present in the data in a fixed range.
- We do scaling for easy to interpretation and faster convergence for gradient descent methods. So, if scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
- Scaling should be done after the train and test split.

- There are 2 types of scaling,
  - **Standardisation**
    - ✓ Standardisation is scaling is used when we want to transfer the data to have zero mean and standard deviation of 1.
    - ✓ Formula is X =x-mean(x) /SD(x)
  - **Normalization**
    - ✓ Normalization/Min Max Scaling is used when we want to bound our values between two numbers, typically, between min=0 and max=1.
    - ✓ Formula is X = x-min(x) / max(x)-min(x)

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

Answer:

- Generally VIF (Variance Inflation Factor) is for calculating the correlation between the independent variables.
- Formula to find the VIF value is 1 / 1 – R-Squared.
- High VIF says that the variable has multicollinearity and its effects the interpretation and inference.
- An infinite VIF value because that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

Answer:

- Q-Q plot is Quantile-Quantile plot.
- Quantiles means the cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way.
- When the quantiles of two variables are plotted against each other, then the plot obtained is known as Quantile-Quantile plot.
- This Q-Q plot provides a summary of whether the distributions of two variables are similar or not with respect to the locations.