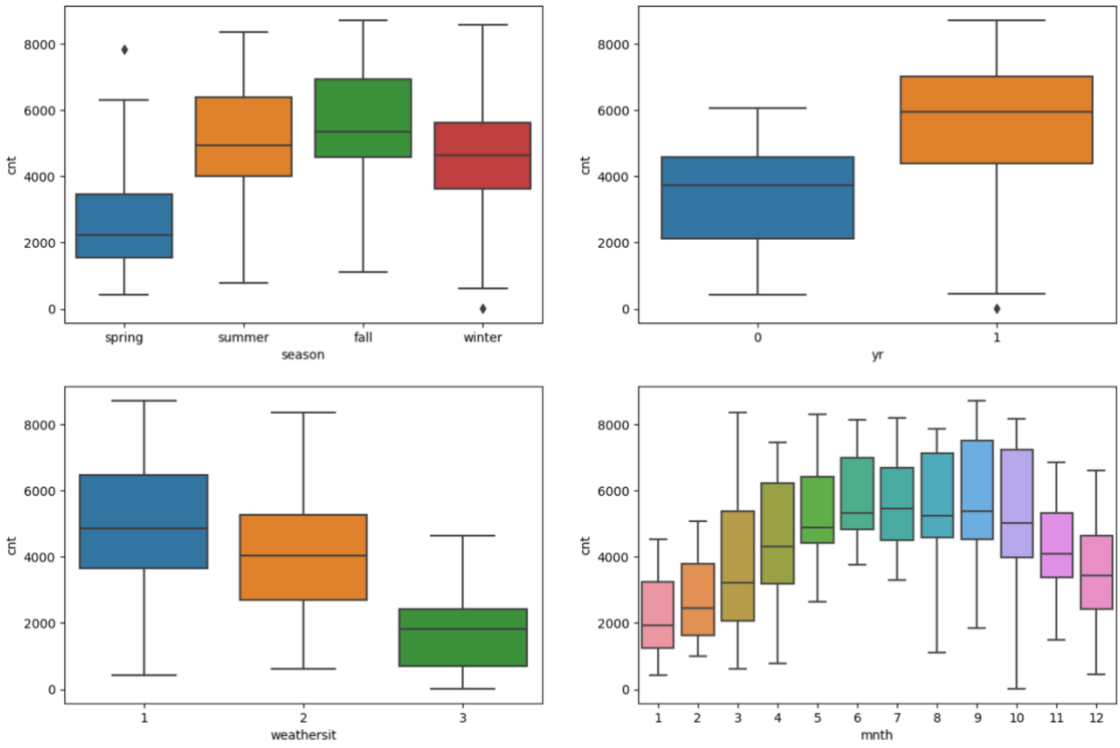
Assignment-based Subjective Questions

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

    a. Summer and Fall seasons log more demand of bikes.
    b. Weather situation affects dependent variable. Clear weather leads to high demand.
    c. As the year passes, demand increase. In 2019 there were rise in demand from previous year.
    d. Months of the year also affects target variable. From March onwards demand increases, peaks and then start coming down from November.



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

    To get k-1 dummy variables out of k categorical levels of the original variable removing the first level. Essentially a categorical variable with k level can be represented well with k-1 dummy variables.

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

    temp and atemp both have high correlation with target variable

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

    a. By looking at pair-plot, I established the linear relationship between temp and atemp (X) and cnt(Y).
    b. Plotting the residual distribution plot (Residual analysis), concluded that error terms are normally distributed and centred around mean zero
    c. Low VIFs indicated there are no multicollinearity among independent variables

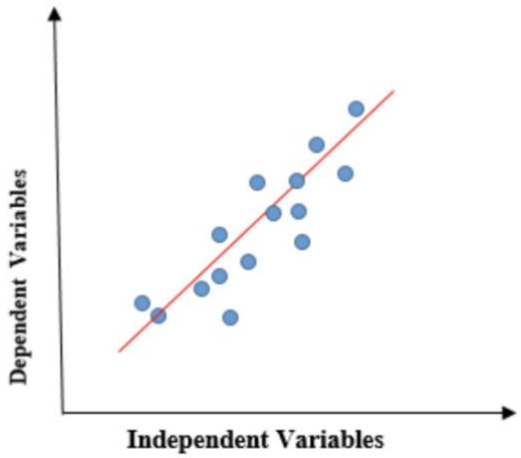
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

    a. **Temperature:** This is the strongest predictor of bike sharing demand with a positive coefficient of 0.4777. As the temperature rises, there is a very high probability of increase in demand.
    b. **Weather situation 3**(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds): This kind of weather situation with a high negative coefficient of -0.2850 indicates a drop in bike sharing demand.
    c. **Year**: With a positive coefficient of 0.2341, this variable suggests that as the year passes with consistent presence of business, the service will become popular and hence demand would grow

General Subjective Questions

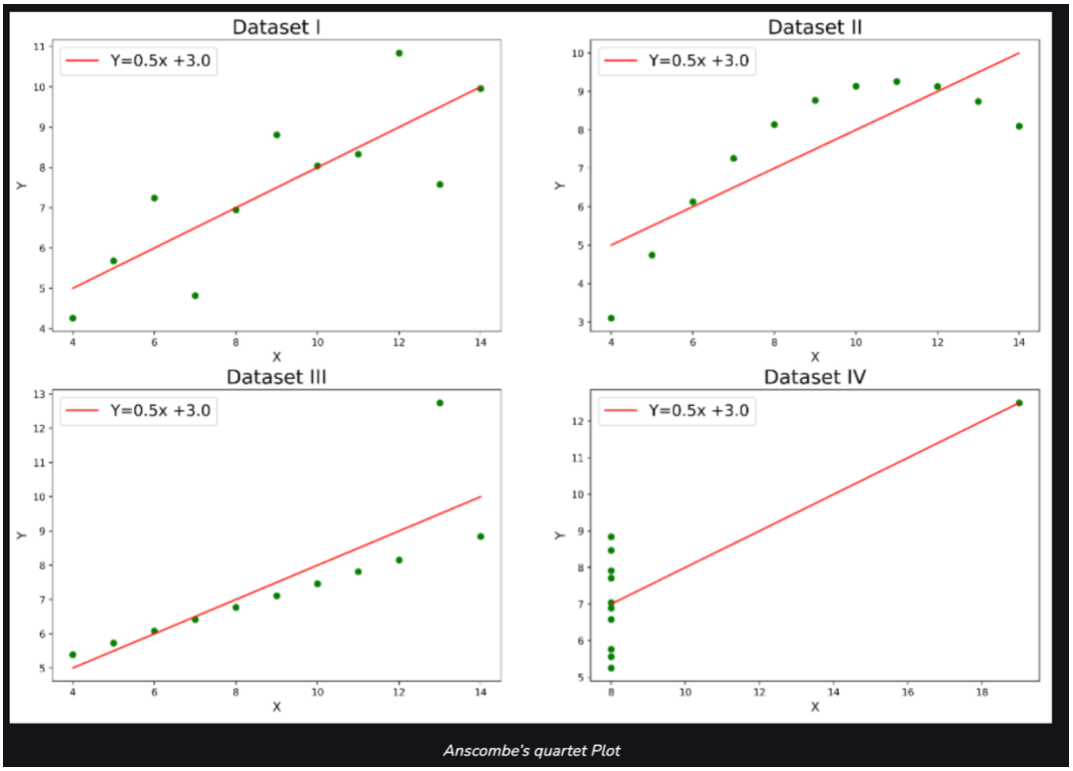
1. **Explain the linear regression algorithm in detail.**

    Linear regression is a supervised learning technique used to predict the target variable and shows the linear relationship between the set of independent variables (X) and target variable (Y). Linear regression depends has a cost function and the algorithm optimises this cost function to get to the optimal solution or a best-fit straight line representing the model.



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

   Anscombe's quartet is a collection of four datasets that have almost same simple descriptive statistics but have very different distributions and appear differently when visualised in graphs. It demonstrates the significance of EDA and drawbacks of relying only on numbers of simple descriptive statistics.



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

   Pearson'r R is the statistical measure of linear correlation between two variables. It is calculated as the ratio between the covariance of two variables and the product of their standard deviations. It is essentially a normalized measurement of covariance where the value lies between -1 and 1. Example: Age and height of students of a class have Pearson's R significantly greater than 0 but less than 1.

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

   Scaling is a technique to put feature values in the same range. The real world data often contain features of varying degree, magnitude, range and units. To enable machine learning models to interpret these features on same scale, we need to perform scaling.
   In normalized scaling, we map the feature values between the range of 0 and 1. E.g MinMaxScaler. While in standardised scaling, the data is not enforced to be in a

range but instead are transformed to have mean = 0 and standard deviation = 1.

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

   Infinite VIF value comes when there is perfect correlation in independent variables, i.e. an independent variable with infinite value of VIF can be expressed exactly by a linear combination of other independent variables.

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

   A Q-Q plot is a plot of quantiles of the first data set against the quantiles of the second data set. Q-Q plot can be used to test distribution amongst two different datasets. This is helpful in machine learning to ensure that model is based on the right distribution