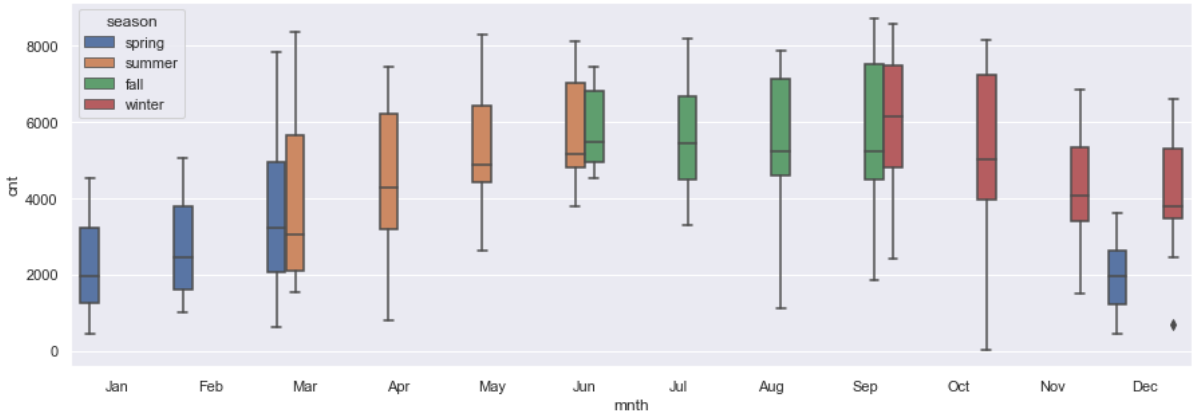
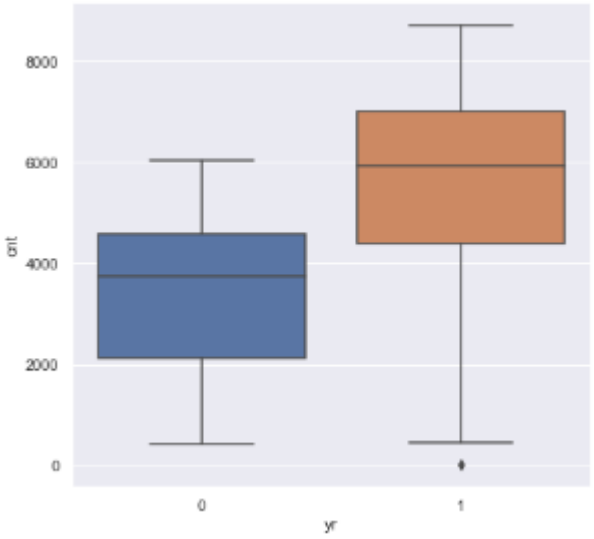
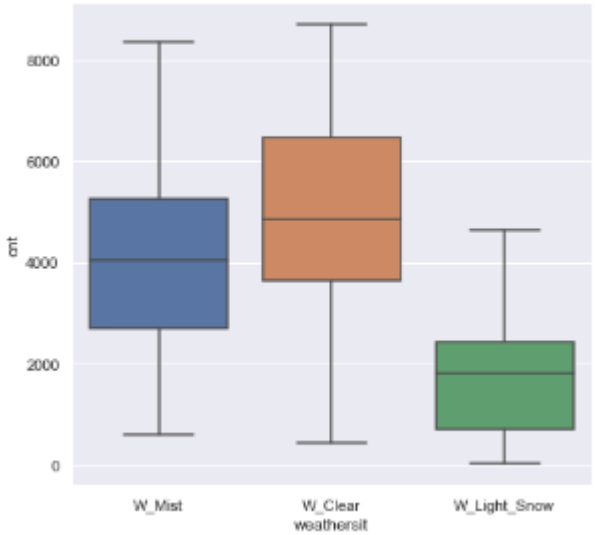
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans :







From the visualisation of categorical variables with 'cnt', shows certain trends like :-

- During Fall Season, Bike Rental is high when compared to other seasons and spring being the lowest.
- Its visible that Bike Rental became popular in 2019 more than when it started in 2018.

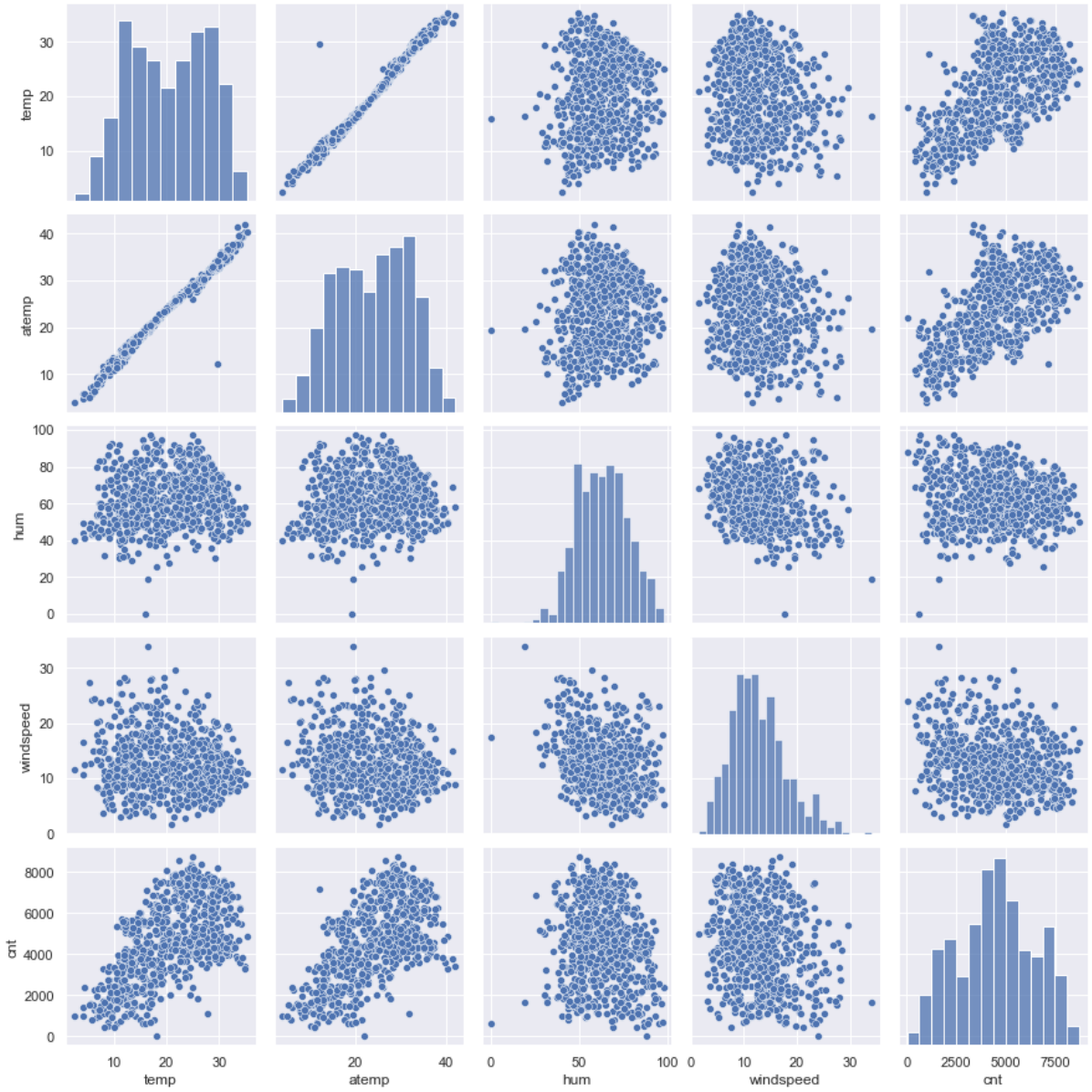
- It’s also visible that demand of Bike Rental increase from Jan steadily and peaks around at mid of the year and again decreases.
- It’s also give indication that Bike Rental is widely used when there is a Clear Weather condition and Snow and Rain has the lowest usage.
- June to Sept are the months where Bike Rentals are widely popular than other months, and most of the Fall season falls during this time, confirming the usage in Fall Season.

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

Ans: It is used to avoid redundancy in the data, as for example if the dummy variable is created for a categorical variable involving 3 categories, only 2 columns are enough to provide information on the column than 3 column which causes redundancy of the same data and contributes to multicollinearity. Dropping unwanted columns can help in contributing to the ease of modelling and the speed of mathematical operations.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans:



- Obviously temp and atemp are the variable which has highest correlation with cnt.
- Also through correlation matrix or heatmap it can be seen that temp and atemp are highly correlated.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:

- Conducted Residual Analysis on the train data to see if the difference of y_train and y_train_predicted follows normal distribution. As per the assumption of Linear Regression Error Terms follow Normal distribution.

- The other Assumptions of Linear Regression are
    - Linear relationship between X and Y
    - Error terms are independent of each other
    - Error terms have constant variance (homoscedasticity)

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans:

- Validated by testing the model on the Test set as well, and comparing the R^2 and adjusted R^2 of both training and test data set.

Conclusions

Comparing Train and Test Dataset :

- Train Dataset R^2 : 0.836
- Train Dataset Adjusted R^2 : 0.832
- Test Dataset R^2 : 0.805
- Test Dataset Adjusted R^2 : 0.793

- We can see that the equation of our best fitted line is:

$$cnt = 0.2335 \times yr + -0.0980 \times holiday + 0.4915 \times temp + -0.1480 \times windspeed + -0.0669 \times spring + 0.0453 \times summer + 0.0831 \times winter + -0.2852 \times W_Light_Snow + -0.0816 \times W_Mist + -0.0524 \times Jul + 0.0767 \times Sep + 0.1996 \times const$$

- The main features which contribute significantly are
    - Temp
    - Year
    - Seasons – Fall season had high demand
    - Weather Situation – During Clear Sky Bikes had high demand

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

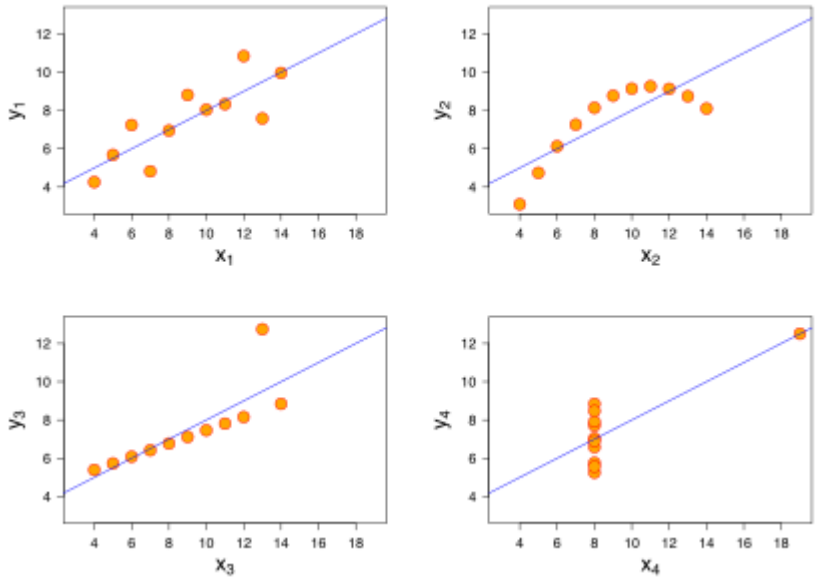
Ans :

- Linear Regression is technique used for Modelling/Predicting relationship btw a dependent variable and one or more independent variables and fitting a line/linear equation btw them.
- The line is expressed by
    - $Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$
    - Where Y is the dependent variable
    - X1, X2, …, Xn are the independent variables
    - β_o is the intercept
    - β_1 … βn are the coefficients
- The best fit line or linear equation is found by minimizing the Residual Sum of Errors (RSS).
- This is achieved by following methods
    - Differentiation
    - Gradient Descent Method
- The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$
- R^2 value ranges from 0 – 1. 1 indicates all the variance is being explained by the fit

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans :

- Anscombe's quartet is an example which shows the importance of visualizing data's or having graphical representation than to just rely on statistical evidences.
- It comprises 4 datasets which gives identical or closely related statistical result ie, mean, variance, correlation coefficient and slope of linear regression line but they are very different when visualised.



- So this example proves that it is as equally important to visualize data or perform EDA on the data while building models.

3. What is Pearson's R? (3 marks)

Ans :

- Statistical measure.
- Pearson's R or Pearson's correlation coefficient.
- It quantifies how strongly and on what direction is the linear relationship btw two continuous variables.
- Pearson's R denoted by r can vary from -1 to 1
    - Where 1 indicates perfect positive correlation
    - -1 indicates perfect negative correlation
    - 0 indicated no correlation
- $$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$
    - Where x and y are the individual data points

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans:

- Scaling is a preprocessing technique used to standardize the range of individual variables in the dataset.
- When a lot of independent variables are present and if they all are in different scales, the model will give a very weird set of coefficients which makes it difficult to interpret.
- In order to make the interpretation easier and faster we scale all the feature variables into common ranges.
- Scaling helps in
    - Easy interpretation
    - Faster convergence for gradient descent method
- This can be achieved using 2 popular methods
    - Standardized Scaling
        - The variables are scaled in such a way that their mean is zero and standard deviation is one.
        - $$x = \frac{x - mean(x)}{sd(x)}$$
    - MinMax/Normalized Scaling
        - The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.
        - $$x = \frac{x - min(x)}{max(x) - min(x)}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

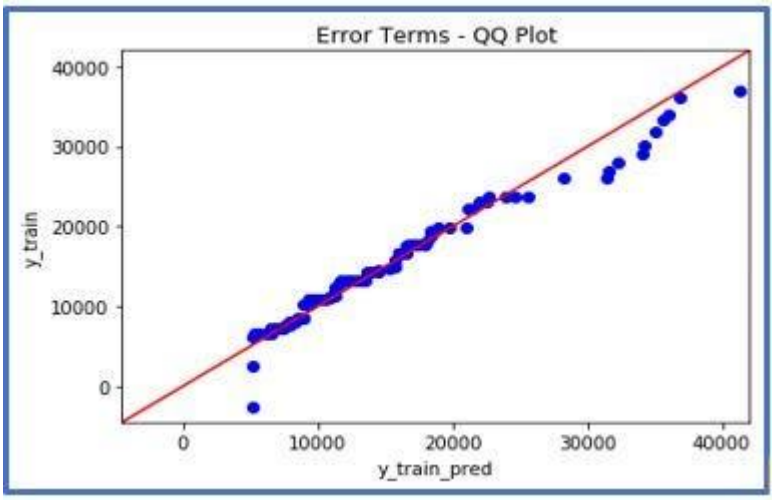
Ans :

- Variance Inflation Factor (VIF) is a measure on how well the variable is explained by all the other variables together.
- It helps in explaining the relationship of one independent variable with all the other independent variables.
- Formula is given by
  - $$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$
    - The common heuristic for VIF is that while a VIF greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.
- The VIF might be infinite as a result of perfect collinearity, that means all the information of that variables is contained in all the other variables
- Its better remove such variables before building models.

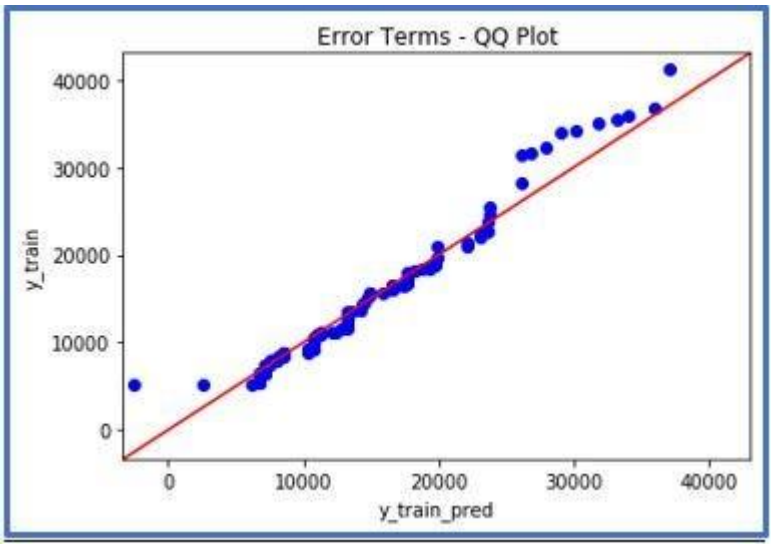
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans :

- Q-Q plot is a graphical representation to determine if a dataset follows a certain probability distribution.
- It is used to check if the two-sample data came from the same source or not.
- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
- Q-Q plots are commonly used in linear regression to assess the normality of residuals, which are the differences between observed and predicted values of the dependent variable.
- Below are the possible interpretations for two data sets.
    - Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
    - Y-values < X-values: If y-quantiles are lower than the x-quantiles.



- X-values < Y-values: If x-quantiles are lower than the y-quantiles.



- Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis