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: The categorical variables in the dataset were season, holiday, weathersit, mnth, yr & weekday. These were visualized using boxplot. These variables had the following effects on our dependant variable:

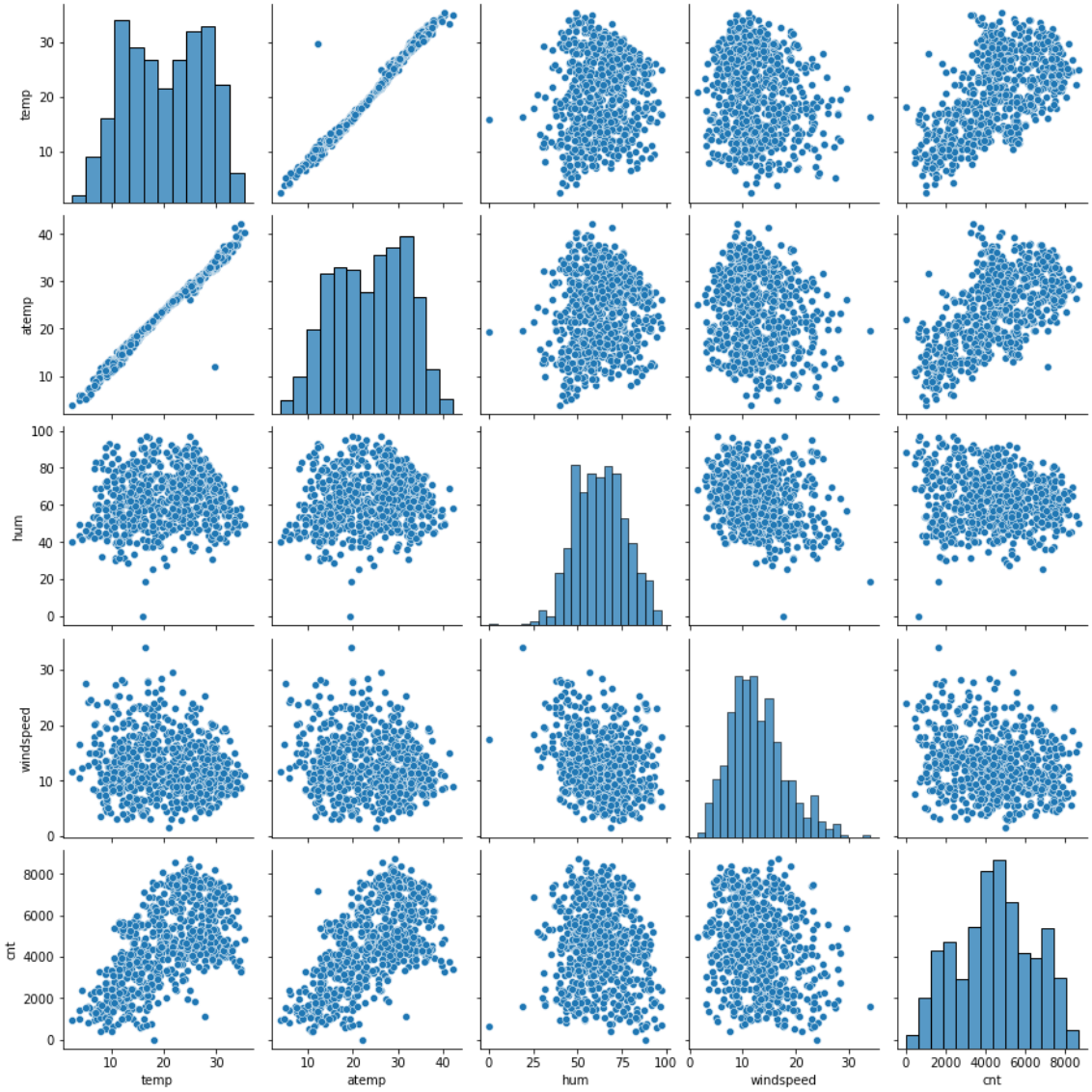
a) season: The boxplot showed that spring season has least value of cnt whereas fall had maximum value of cnt. Summer and winter had intermediate value of cnt.
b) weathersit: There was no users during heavy rain & snow indicating that this weather is extremely unfavourable. Highest count was seen when weather was clear & misty.
c) holiday: Rentals reduced during holiday
d) mnth: September saw the highest number of rentals while December saw the least. This observation is on par with the observation made with weathersit. The weather situation in December is usually heavy snow.
e) yr: The number of rentals in 2019 was more than 2018.

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

Ans: If you won't drop the first column then your dummy variable will be correlated(redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example, iterative models may trouble converging and lists of variable importance may be distorted. Another reason is, if we have all dummy variables it leads to multicollinearity between the dummy variables. To keep this under control, we drop one column.

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

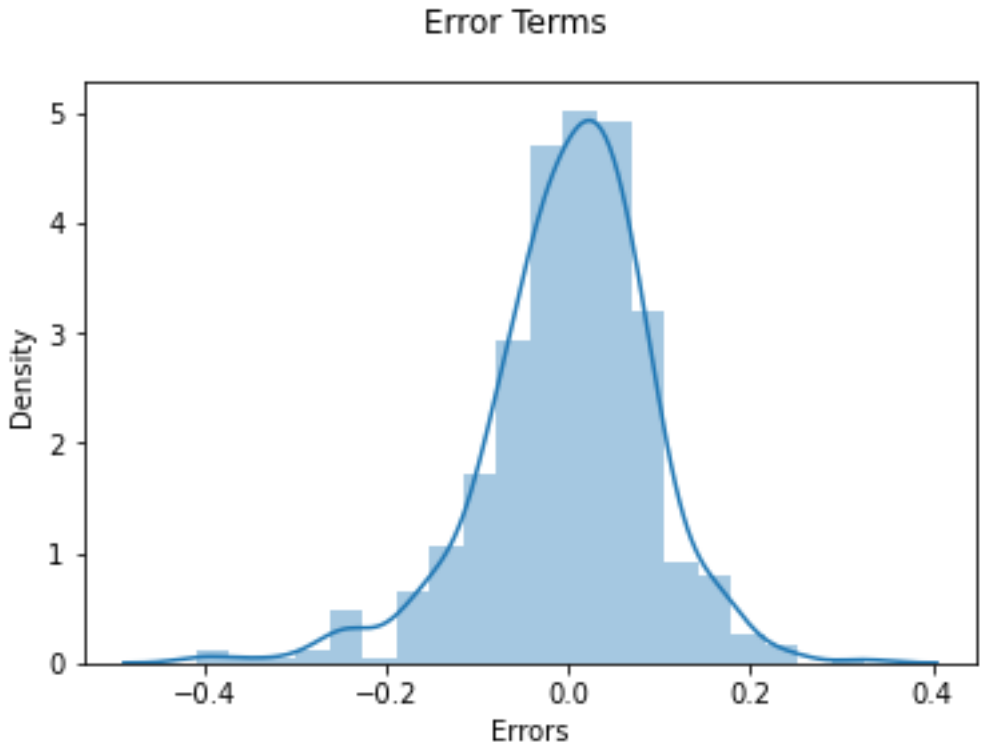
Ans:



- 'temp' and 'atemp' are the two variables which are highly correlated with the target variable(cnt)

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

Ans:



Residual distribution should follow normal distribution and centred around 0(mean = 0). We validate this assumption about residuals by plotting a distplot of residual and see if residuals are following normal distribution or not. The above plot shows that the residuals are distributed about mean = 0.

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: As per the final model, predictor variables that influences bike booking are:

a) Temperature(temp): A coefficient value of '0.564763' indicates that temperature has significant importance on bike rentals.
b) Light Rain & Snow (weathersit = 3): A coefficient value of '- 0.260625' indicates that the light snow & rain deters people from renting out bikes.
c) Year(yr): A coefficient value of '0.234441' indicates that year wise bike rental numbers are increasing.

General Subjective Questions:

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

Ans: Linear regression is a type of supervised machine learning algorithm that is used for the prediction of numeric values. Linear regression is the most basic form of regression analysis. Regression is most commonly used predictive analysis model.

Linear regression is based on the popular equation $y = mx + c$

It assumes that there is a linear relationship between the dependent variable (y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.

Regression is performed when the dependent variable is continuous data type and predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error,

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

a) Simple Linear Regression: SLR is used when the dependent variable is predicted using one independent variable.
b) Multiple Linear Regression: MLR is used when the dependent variable is predicted using multiple independent variables.

Equation for MLR is:

$$yi = \beta_o + \beta_1 xi_1 + \beta_2 xi_2 + \beta_p xi_p + \epsilon$$

Where, for i = n observations:

yi *= dependent variable*

xi *= explanatory variable*

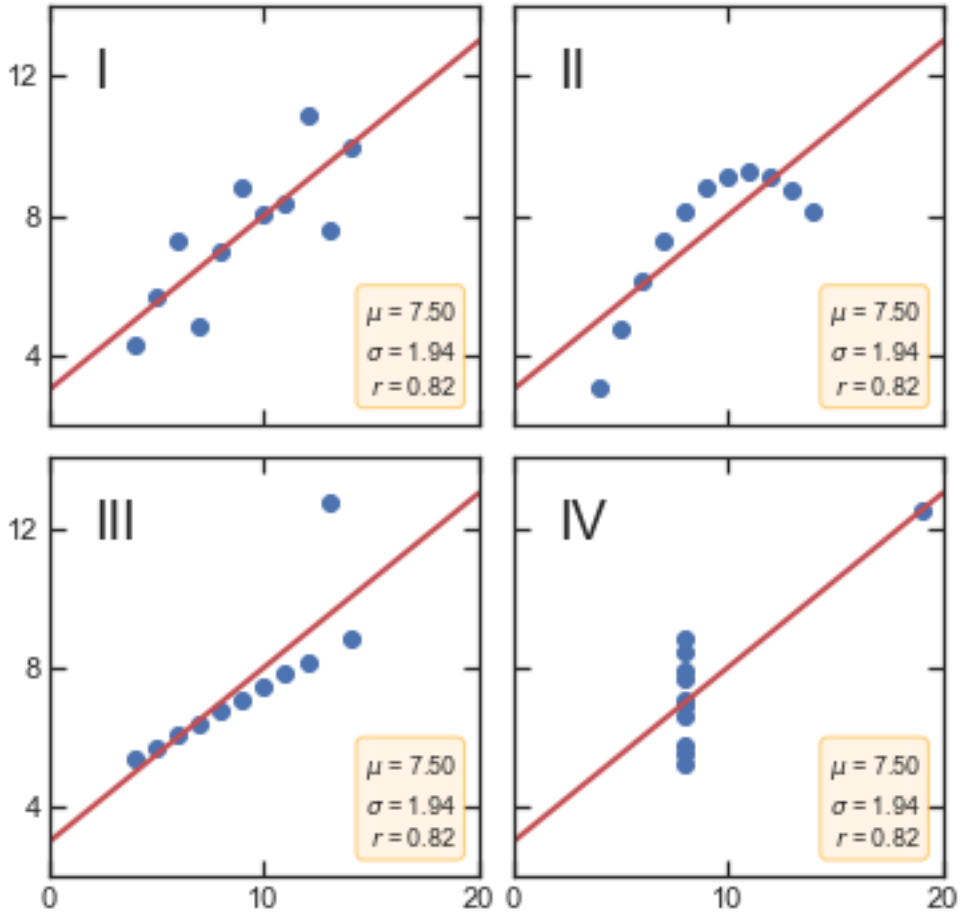
β_o *= y-intercept (constant term)*

β_p *= slope coefficients for each explanatory variable*

ϵ *= the model's error term (also known as residuals)*

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

Ans: Anscombe's quartet compromises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.



Explanation of this output:

a) In the 1st one (top left) if we look at the scatter plot we can observe that there seems to be a linear relationship between x & y.
b) In the 2nd one (top right) if we look at this figure we can conclude that there is a non-linear relationship between x & y.
c) In the 3rd one (bottom left) we can say that when there's a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated to be far away from that line.
d) Finally, the 4th one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Application:

- The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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

Ans: Pearson's R is a numeric summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. 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.

The Pearson's correlation varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction).

r = -1 means the data is perfectly linear with negative slope (i.e., both variables tend to change in different directions).

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? (3 marks)

Ans: Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

- Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest and Neural Network.
- Standardization on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

Ans: VIF: The variance inflation factor gives how much the variance of the coefficient estimate is being inflated by collinearity.

$$VIF = \frac{1}{(1 - R_1^2)}$$

If there is perfect correlation, then VIF = Infinity.

Where R_1 is the R-square value of that independent variable which we want to check how well this independent variable can be explained perfectly by other independent variables, then it will have a perfect correlation and it's R-squared value will be equal to 1. So, $VIF = \frac{1}{(1-1)}$ which gives $VIF = \frac{1}{0}$ which results in 'Infinity'.

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

Ans: A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. It is used to compare the shape of distribution. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another.

If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

The Q-Q plot is used to answer the following questions:

a) Do two datasets come from populations with common distribution?
b) Do two datasets have common location and scale?
c) Do two datasets have similar distributional shapes?
d) Do two data sets have similar tail behaviour?