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?

➔ The categorical variables identified in the dataset are Season, Weather Situation (weathersit), Holiday , month, working day and weekday.
Box plot is used for visualisation.

The influence of variables on the dependent variable (cnt) is listed below:-

- ***Season:*** Spring season has the lowest value of cnt and fall season has the highest value of cnt. Summer and winter had cnt values in the middle.

- ***Weather Situation (weathersit):*** The highest value of cnt was observed when the weather is clear / partly cloudy. During heavy rain/ snow there was a very significant drop in the number of users.

- ***Holiday:*** The number of users was very low during holidays.

- ***Month:*** September had higher number of rentals

- ***Weekday:*** Weekends had a significant increase in the number of users

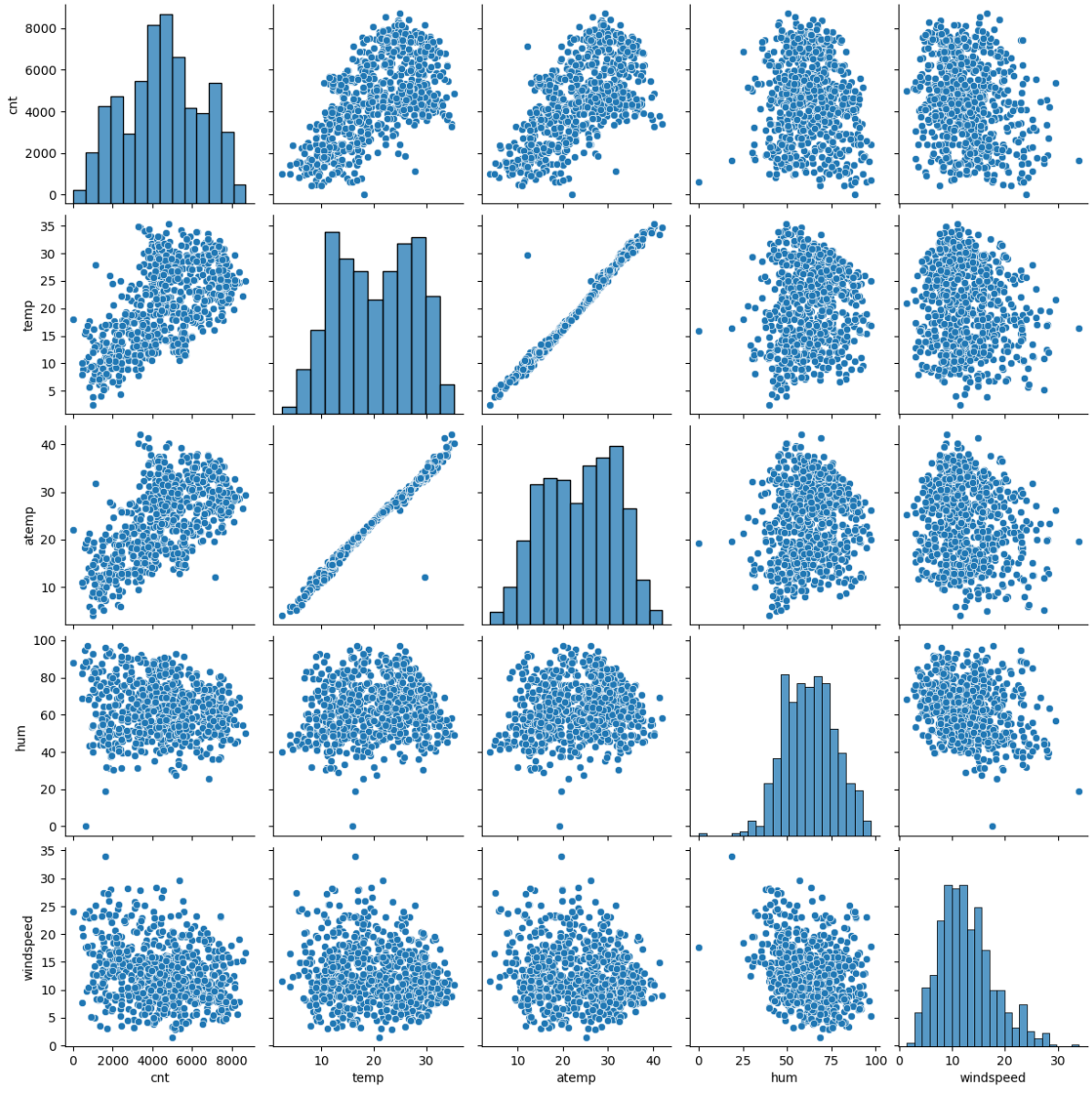
- ***Working day:*** had very little impact on the number of users renting the bike

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

➔ Dummy variables will be correlated if the first columns (redundant) are not removed. This may have negative impact on some models and the effect will be amplified when cardinality is low.
Another argument is that having all dummy variables results in multicollinearity between them.

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

➔ “temp and “atemp” are the two numerical variable that has highest correlation with the target variables.

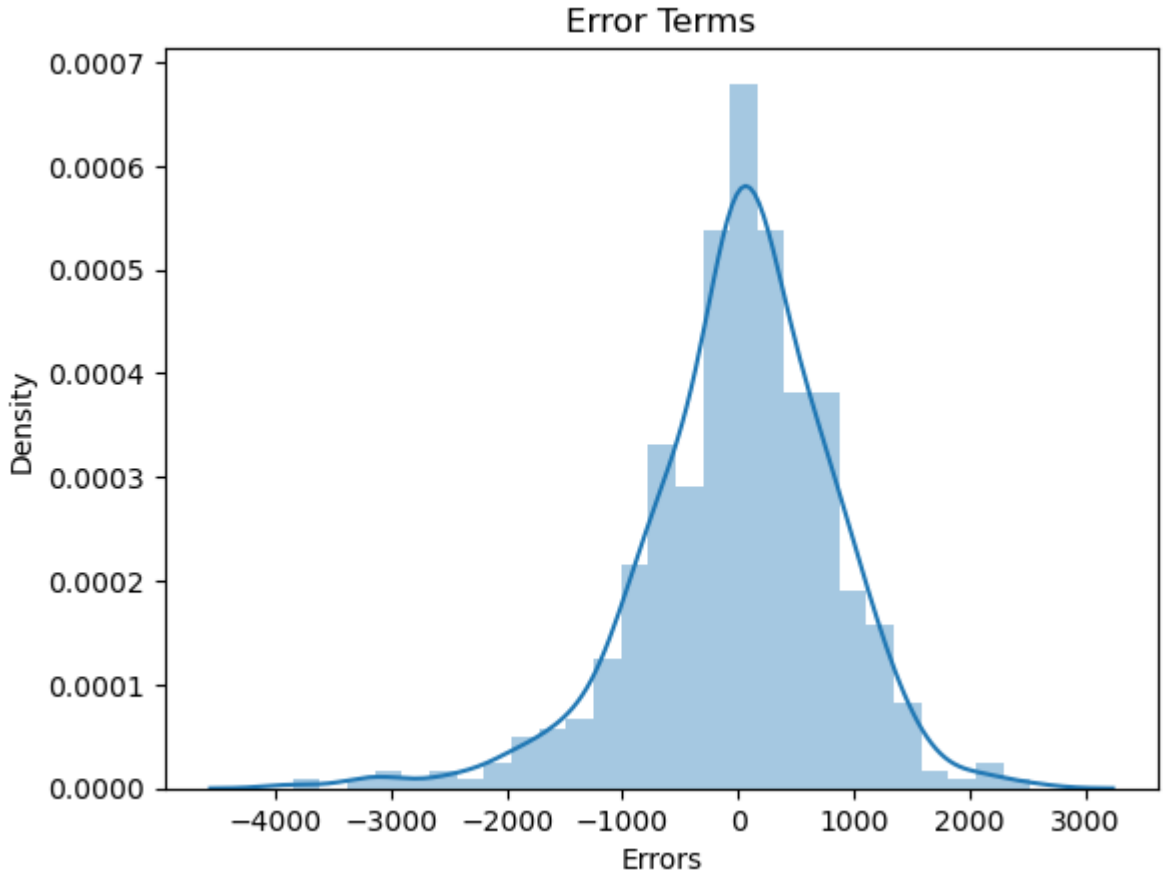


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

   ➔ The distribution of residuals should be normal and around 0 (the mean is zero)

   We test this residuals assumption by producing a distplot of residuals to see if they follow a normal distribution or not

   The residuals are scattered around 0 (shown in the diagram).



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

   ➔ The top 3 predictor variables that influence the bike rental are:-

   - ***Temperature (temp)*** : with a coefficient of 0.5173, a unit increase in the temp variable increases the number of rentals by 0.5173 units.

   - ***Weather situation (weathersit_3)*** : with a coefficient of -0.2828, a unit increase in the weathersit_3 variable reduces the number of rentals by 0.2828 units as compared to weathersit_3 refers to light snow, light rain + thunderstorm + scattered clouds , light rain + scattered clouds.

   - ***Year(yr)*** with a coefficient of 0.2324, a unit increase in the year variable increases the number of rentals by 0.2324 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

➔ Linear Regression is a type of supervised Machine Learning algorithm that is used for prediction of numeric values. Linear regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model. Linear regression is based on the popular equation:-

$$y = b_o + b_1 x$$

where b_o is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.

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

In Regression the output/ dependent variable is the function of the independent variable and the coefficient and the error term.

Regression is broadly divided into 2 types Simple Linear Regression and multiple linear regression

Simple Linear Regression: SLR is used when the dependent variable is predicted using only one independent variable

Multiple Linear Regression: MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR:-

$$y = b_o + b_1x_1 + b_2x_2 + b_3x_3 \ldots . + b_nx_n$$

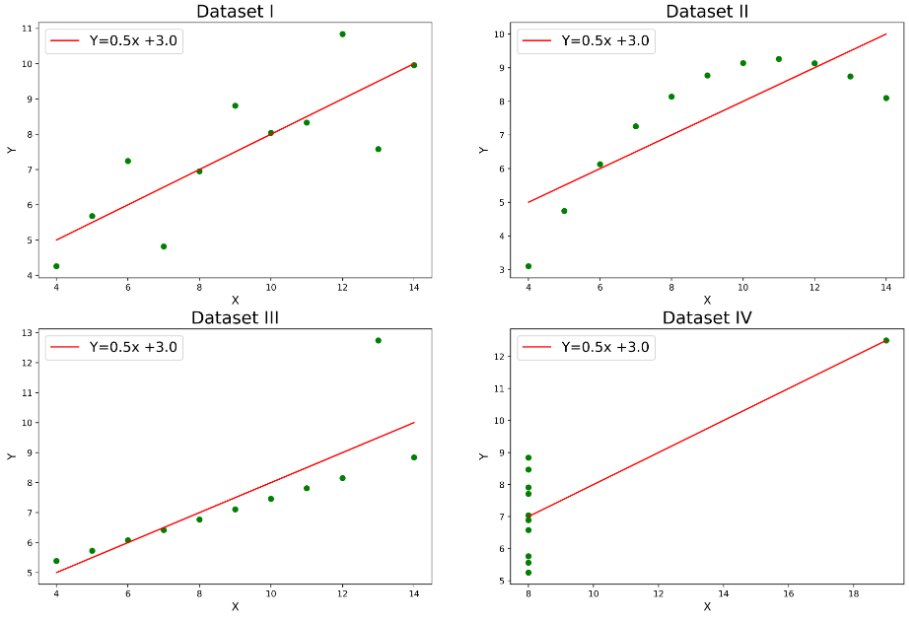
(where b_o is the intercept, $b_1, b_2, b_3, b_4 \ldots, b_n$ are coefficients or slopes of the independent variables $x_1, x_2, x_3, x_4 \ldots, x_n$ and y is the dependent variable.)

In this formula:

- Y stands for the predictive value or dependent variable.
- The variables (X1), (X2) and so on through (Xp) represent the predictive values, or independent variables, causing a change in Y. It's important to note that each X factor represents a distinct predictive value.
- The variable (b_0) represents the Y-value when all the independent variables (X1 through Xp) are equal to zero.
- The variables (b1) through (bp) represent the regression coefficients.

2. Explain the Anscombe's quartet in detail.

➔ Anscombe's quartet was developed by statistician Francis Anscombe. It includes 4 datasets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on the graph. It was developed to emphasize both the importance of graphing data before analysing it and the effect of outliers and other influential observations.



Statistical properties:-

- The first scatter plot (top left) appears to be simple linear relationship
- The second graph (top right) is not distributed normally; while there is a relation between them, it is not linear.
- In the third graph (bottom left) the distribution is linear , but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to the lower corelation coefficient from 1 to 0.816
- Finally the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient even thought the other data points do not indicate any relationship between the variables.

3. What is Pearson's R?

➔ Pearson's R is a numerical representation of the strength of the linear relationship between the variables. Its value ranges from -1 to +1. It depicts the linear relationship of 2 sets of data. In layman's terms, it asks if we can draw a line graph to represent the data.

r=1 means the data is perfectly linear with a positive slope
r= -1 means the data is perfectly linear with a negative slope
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?

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

- ***Normalisation*** 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 neighbours and neural networks.

- ***Standardisation*** 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 normalisation, standardisation does not have a bounding range. So, even if you have outliers in the data, they will not be affected by the standardisation.

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

➔ ***VIF: Variance inflation Factor:*** - indicates how much collinearity has increased the variance of coefficient estimate. VIF is equal to 1/ (1- Ri^2). VIF = infinity if there is perfect correlation. Where R-1 denotes the R-square value of the independent variable which we want to see how well it is explained by other independent variables. – if an independent variable can be completely described by other independent variables, it has perfect correlation and has a R-Squared value of 1. As result VIF = 1/ (1-1) provides VIF = 1/0 which is infinity.

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

➔ The quantiles of the first data set are plotted against the quantiles of the second data set in a q-q graphic. It's a tool for comparing the shapes of different distributions. A scatterplot generated by plotting 2 sets of quantiles against each other is know as Q-Q plot.

Because both sets of quantiles come from the same distribution, the plots should form a line. That's fairly a straight line.

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

Do 2 data sets come from populations with a common distribution ?

Do 2 data sets have common location and scale?

Do 2 data sets have similar distributional shapes?

Do 2 data sets have similar tail behaviour?

****End of the document****