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:

The categorical variables are:

Season, Month, Weekday, Weathersit, Holiday, Workingday.

A boxplot was created to check the effects of these variables and below are the inferences:

1. **Season** 3: "Fall" has the highest demand for rental bikes

2. **Month**: Demand is growing each month till June. September month (which is also related to season "fall"), has the highest demand after September, demand is decreasing.

3. **Weekday**: do not have much differences, so it cannot be concluded.

4. **Weathersit** = 'Clear' has highest demand which implies, people prefer shared bikes on "Clear weather days".

5. **Holiday**: On "Holidays" i.e., Holiday = 0, the demand decreases.

6. **Workingday**: shows that, usually the demand is more on workingday.

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

Answer: drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

If we have categorical variables with n-levels, then we need to use n-1 columns to represent the dummy variables.

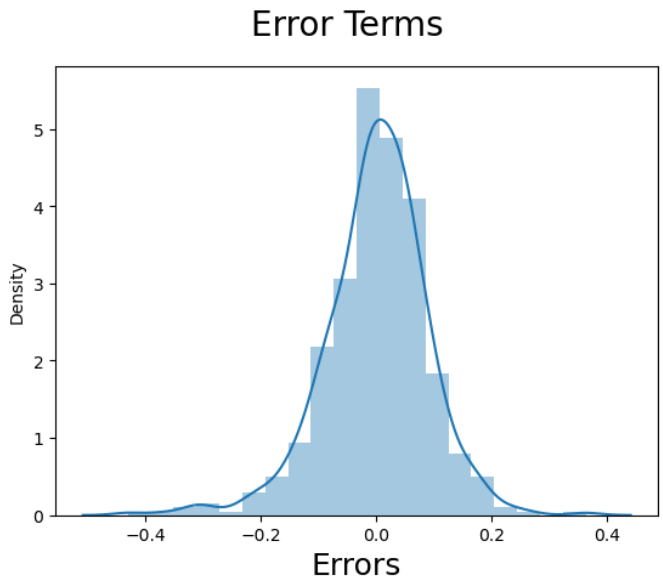
Hence drop_first = True ensures to achieve this goal.

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

Answer: The ‘temp’ and ‘atemp’ variables, has the highest correlation with the target variable ‘cnt’ and has linear relationship.

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

Answer: Once the linear regression model is built (which in my case was at model 5), I analysed the error terms



From the above figure, one can understand that,

The error terms follow normal distribution and is centred at mean = 0

They are independent.

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

Answer: According to the final model,

Year: With coefficient = 0.2342, unit increase in the "Year" variable increases no. of bikes shared by 0.2342

Month = "Sep": With coefficient = 0.0561, unit increase in the variable of "Sep" (which represents "Spring" season also), there is a increase in no. of bikes shared.

Weathersit = 3 (Light_snowrain): With coefficient = -0.2907, decrease in the variable value decreases the no. of bikes shared.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a type of machine-learning algorithm more specifically a supervised machine-learning algorithm that learns from the labelled datasets and maps the data points to the most optimized linear functions, which can be used for prediction on new datasets.

Machine learning, basically uses 3 types of algorithms:

1. Regression
2. Classification
3. Clustering

Regression: To simply put, trying to predict future analysis based on past experiences is what we do in regression. The output variable to be predicted is a continuous variable.

Classification: Able to classify whether a particular variable fall under which area is classification. The output variable to be predicted is a categorical variable.

Clustering: Lots of data that we can cluster as and when required, and group them, and then discover, it belongs to which category.

To do this, we have supervised and unsupervised learning methods.

Supervised: Past data with labels used for building the models. Regression and classification algorithms fall under this category.

Unsupervised: No predefined labels are assigned to input data. Clustering falls under this category.

Linear regression is based on the popular equation:

$$\text{“}Y = mx + c\text{”}$$

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

Linear regression models can be classified into two types depending upon the number of independent variables:

Simple linear regression: When the number of independent variables is 1.

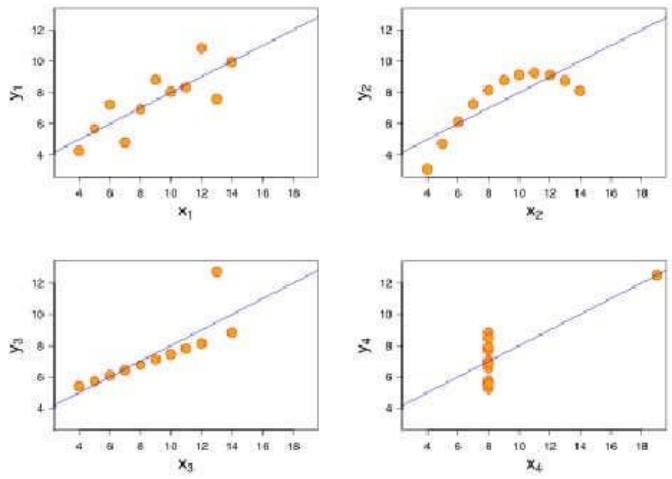
$$Y = \beta_0 + \beta_1 X + e$$

Multiple linear regression: When the number of independent variables is more than 1.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

2. Explain the Anscombe’s quartet in detail.

Answer: Anscombe’s quarter comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.



3.What is Pearson's R?

Answer: The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

r = 1: Perfectly linear with positive slope

r= -1: perfectly linear with negative slope

r= 0: no linear association

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

Answer:

Scaling: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why scaling: Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling.

Difference:

Normalization: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization: It replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

Answer: This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.\

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

Answer: Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.