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)**

Answer: The dataset contained following categorical variables:

- season - 1: spring, 2: summer, 3: fall, 4: winter
- yr - 0: 2018, 1:2019
- mnth – 1 through 12 representing January through December.
- Holiday - Whether the day is a holiday or not.
- Weekday – Day of the week
- weathersit –
    - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

By analysing these variables, it can be inferred that:

- There has been an increase in customers in year 2019 as compared to year 2018.
- The season Fall get maximum customers. September month has highest count of customers.
- Weather impacts the demand and we see maximum count during the partly cloudy/clear sky, while heavy rain doesn't have any data.
- The count or demand drops during holidays.

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

Answer: Using drop_first=True in dummy variable creation helps in preventing Multicollinearity. It avoids perfect correlation between dummy variables by dropping one category, which makes the model more stable.

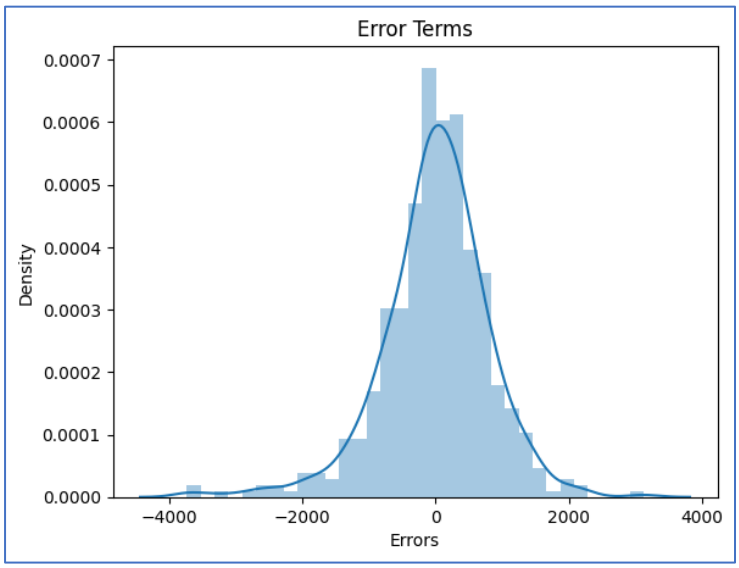
Also, it simplifies the model by reducing the number of features by using one category as a reference group, without losing any important information.

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

Answer: The variables for temperature (temp) and Feeling Temperature (atemp) has the highest correlation 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)**

Answer: The fundamental assumption is that the Error Terms follow the normal distribution with a mean 0. The Distplot displays the same.



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

Answer: Based on the final model output, the top 3 features contributing significantly to the demand for shared bikes are:

1. Temperature (temp): Coefficient = 4051.31, indicating that an increase in temperature has the strongest positive impact on demand.
2. Year (yr): Coefficient = 2038.54, showing that bike demand increases significantly in the later year (likely indicating a growing trend in bike sharing).
3. Weather (weathersit_Bad): Coefficient = -2451.42, a strong negative impact, meaning bad weather conditions lead to a significant reduction in bike demand.

General Subjective Questions

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

Answer: Linear regression is a method used to predict a continuous target variable based on one or more input variables. It assumes that there is a straight-line relationship between the inputs (independent variables) and the output (dependent variable).

The goal is to find a linear equation that best predicts the target variable Y using the input variables X. The equation for one input (simple linear regression) looks like:

$$Y = mX + c$$

- Y - dependent variable we are trying to predict.
- X - independent variable we are using to make predictions.

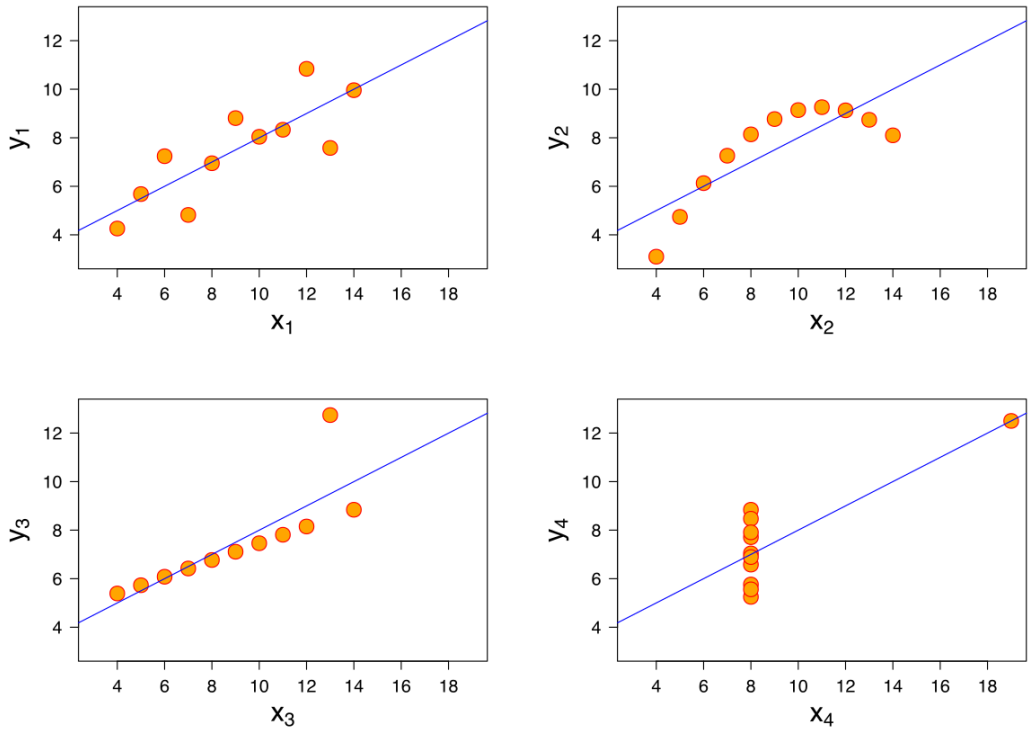
- m - slope of the regression line which represents the effect X has on Y
- c - constant, known as the Y-intercept.

If X = 0, Y would be equal to c.

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

Answer:

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It was constructed to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x, y points in all four datasets.



- 1st data set fits linear regression model as it seems to be linear relationship between X and y
- 2nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.
- 3rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.

- 4th data set has a high leverage point means it produces a high correlation coeff. Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before build machine learning model.

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

Answer:

Pearson's r is a numerical 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 correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases

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

Answer:

Scaling is necessary for a model to be functional with the appropriate range of coefficients. For e.g., if there were two independent variables named price and months on which the sale of car depended, the price range would be far too high because there are only 12 months in a year. In that case, scaling the variable price appropriately won't allow decimal errors to happen in the model. There are two types of scaling: Normalized scaling: This scaling is done to make the distribution of data into a Gaussian one. It doesn't have a preset range. Typically used in Neural networks broadly. Standardized scaling: The example given above is of standardized scaling. Here, the values of variable(s) is/are compressed into a specific range to suit the model.

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

Answer:

If there is a perfect correlation between the dependent variable and independent variable(s), the R-squared value comes out to be 1. Hence VIF, which is (1/(1-R^2)) turns out to approach infinity.

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

Answer:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests