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:

- Categorical variables like season, weather situation, year, holiday has both positive and negative impact on Dependent variable.
- In the given two years given in data set, the shared bikes count increased by 23% from 2018 to 2019
- Weather situation when becomes extreme (storm, rain, etc.,), can reduce the shared bikes count. As weather worsens, it leaves negative impact on count.
- Holiday also has negative effect on the count. During holiday, the shared bikes count could reduce
- In Winter season, the count of shared bikes might increase by nearly 12%, which is a positive effect

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

Answer:

- 'drop_first' drops first variable of the dummy variables.

Example: Let us take an example of a variable with three categories, CarType – A, B, C

A	B	C
0	0	1
0	1	0
1	0	0

Now, if we remove column A, then B & C values are sufficient to define all 3 types.
00 – defines type A
10 – defines type B
01 – defines type C, using all 3 columns would result in redundancy.

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

Answer:

- As per the pair-plot of numerical variables, temperature (both 'temp' and 'atemp' depict temperature of the day) has the highest correlation with target variable

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

Answer:

1. *Linear Relationship b/w X and Y:* Based on the scatter plots for numeric variables, we can see a linear relationship seen between predictors and count variable.
2. *Residuals are normally distributed*: The distribution plot of residual terms in the notebook clearly shows the assumption is true for our built model.

3. *Zero mean Assumption:* From the graph of Error distribution, the residuals are centered around 'zero' which implies mean of residuals is zero.
4. *Error terms have constant variance*: Graph showing the distribution between fitted values and residuals, there is not much variance in the residuals.

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:

- **temperature** is first feature with 0.5914 coefficient value explaining the demand of shared bikes
- **weathersituation** with coefficient value of -0.2471, is also contributing to the demand for shared bikes
- **year** (coeff-value of 0.2327) has significant effect on demand for shared bikes

General Subjective Questions

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

Answer:

- Linear Regression is a process of estimating the relation of one variable against the other variable. We estimate the relationship between dependent variable and multiple independent variables. These independent variables are also called 'Predictors'.
- Mathematically, Simple Linear Regression equation is given as

  $y = mx + c$ Here,
  - y is the dependent variable
  - x is the independent variable
  - m is the coefficient of x variables or slope of the line
  - c is the constant or y-intercept
- When more than one independent variable is used, it is called Multiple Linear Regression.
- This is given by equation, $y = c + m_1x_1 + m_2x_2 + .. + m_nx_n$, for 'n' independent variables. Where, m_1, m_2, … are coefficients of each of x variables

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

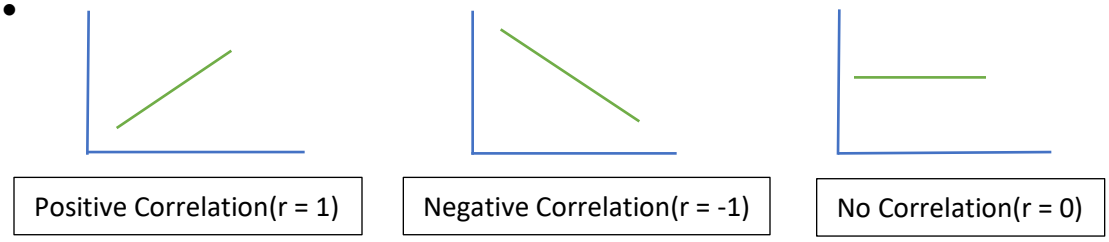
Answer:

- Anscombe's quartet is used to tell the importance of plotting the data set, before we proceed with the analysis and building the model.
- Anscombe's quartet consists of four data sets, each with eleven data points, with nearly identical statistics.
- The x value of the first three datasets is same and fourth data set has almost same values in X variable.
- However, when these 4 data sets are plotted on the graph, they are completely different. This explains the importance of plotting the dataset and observe the distribution of data set before we proceed with model building.

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

Answer:

- Pearson's R is also called as Pearson Correlation coefficient. It is defined as the correlation between two variables or the two data sets, with value always ranging between -1 and 1
- For any given two data sets, if r is **0 to 1**, then two sets are strongly correlated. When one variable changes, other changes in same direction.
- If r is **-1 to 0**, then two sets are negatively correlated. When one variable changes, the other changes in negative direction.
- If **r = 0**, then there is no correlation between the two data sets. When one set changes, other remain constant.
-



Positive Correlation(r = 1) | Negative Correlation(r = -1) | No Correlation(r = 0)

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

Answer:

- Scaling is defined as the transformation of the variables to fit the data into a particular defined range. For example, all the data is fit between 0 and 1.

- If the variables within the data set have huge variation and different ranges, then scaling is performed to transform the variables and bring all these variables into the same range.
- This process is used to ease the interpretation of the model.
- *Normalization Scaling*: This is also called as Min-Max Scaling. This scaling transforms the variables to fit the values between the range of 0 and 1.
    - x – min(x) / max(x) – min(x)
- *Standardization Scaling*: Standardization transforms the data based on the standard mean and normal standard deviation.
    - x – mean(x) / sd(x)

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

Answer:

- *Variance Inflation Factor (VIF*) is used to determine the multicollinearity between the independent variables.
    - VIF = $1/(1-R^2)$
- VIF value of a variable is determined based on the R^2 value of the model.
- If VIF is infinite for a variable, then it implies that the R^2 value of the model is 1.
- This means that the given variable is dependent on the other variables and the other independent variables can explain that variable.
- Model with R^2 as 1 is considered a too perfect fit model and variables causing this should be removed to build a proper model.

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

Answer:

- Q-Q plot called as Quantile-Quantile plot is the graphical representation of distribution of quantiles of two sets against each other.
- For example, median is a quantile, where 50% of the data lies below the data point and 50% of data lies above the data point.
- Q-Q plot is mainly used to verify the distribution of the two sets.
- If the distribution of two sets is similar, the all the data points will fall on the y=x line plotting in Q-Q plot.
- If data points don't fall on the line, then it is understood that the distribution of two data sets is completely different.