LINEAR REGRESSION ASSIGNMENT

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?
    - I have done analysis on categorical columns using the boxplot. Below are the few points we can infer from the visualization.
    - Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
    - Most of the bookings has been done during the month of May, June, July, August, September, and October. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
    - Clear weather attracted more booking which seems obvious.
    - Thursday, Friday, Saturday, and Sunday have more number of bookings as compared to the start of the week.
    - When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
    - Booking seemed to be almost equal either on working day or non-working day.
    - 2019 attracted more number of booking from the previous year, which shows good progress in terms of business.

2. Why is it important to use drop_first=True during dummy variable creation?
    - The intention behind the dummy variable is that for a categorical variable with 'n ' levels, you create 'n-1' new columns each indicating whether that level exists or not using a Zero or one.

- Hence drop_first=True is used so that the resultant can match up n-1 levels.
- Hence it reduces the correlation among the dummy variables.
- If there are 3 levels, the drop_first will drop the first column.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    - "Temp" is the variable which has the highest correlation with target variable.
    - Similarly, 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    - Normality of error terms
    - Multicollinearity check
    - Linear relationship validation
    - Homoscedasticity
    - Independence of residuals

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    - Top 5 features that has significant impact towards explaining the demand of the shared bikes are Summer, Winter, September, Temperature and year

General Subjective Questions:

1. Explain the linear regression algorithm in detail
    - Linear Regression is a form of predictive modelling technique which tells us the relationship between the independent and dependent variables.
    - Linear Regression between variables means that when the value of one or more independent variables will change, the value of dependent variable will also change accordingly.
    - There are two types of linear regression - simple linear regression and multiple linear regression.
    - Simple Linear Regression – One independent variable is used.
    - Simple Linear Regression uses traditional slop-intercept from, where m and b are the variable our algorithm will try to " learn" to produce the most accurate predictions. X represents our input data and y represents our prediction.
        - $Y = mx + b$
    - Multiple Linear Regression – one or more independent variables are used.
    - A more complex, multi-variable linear equation might look like this, where W represents the coefficients, or weights, our model will try to learn
        - $F(x, y, z) = W1x + W2y + W3z$
    - The variables x, y, z represent the attributes, or distinct pieces of information, we have about each observation.
    - A regression line can be a positive Linear Relationship or a Negative Linear Relationship.
    - Positive Linear Relationship:

- A linear relationship will be called positive if both independent and dependent variable increases.
- A positive linear relationship is when the dependent variable on the Y-axis along with the independent variable in the X-axis.

➢ Negative Linear Relationship:

- A linear relationship will be called positive if independent increases and dependent variable decreases.
- If dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship.

2. Explain the Anscombe's quartet in detail.

➢ Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyse it and build your model.

➢ Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very difference distributions and appear very different when presented graphically.

➢ 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 dataset.

3. What is Pearson's R?

The Pearson's R also known as Pearson's correlation coefficient is used to establish a linear relationship between two quantities. It gives an indication of the measure of strength two variables and the value of the coefficient can be between -1 and +1.

- -1 coefficient indicates strong inversely proportional relationship.
- 0 coefficient indicates no relationship.
- 1 coefficient indicates strong proportional relationship.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   - The Scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.
   - In simpler terms, in machine learning algorithms we need to bring all features in the same standing, so that one significant number doesn't impact the model just because of their large magnitude. This is called scaling or Feature scaling.
   - The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.
   - Normalization/min-max scaling – The min max scaling normalized the data within the range of 0 and 1. The min max scaling helps to normalize the outliers as well.
     - Min Max-Scaling:
       - $X = x\text{-}min(x) / max(x) – min(x)$
   - Standardization converges all the data points into a standard normal distribution where means is 0 and standard deviation is 1.
     - Standardization:
       - $X = x – mean(x) / sd(x)$

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

$$VIF = 1 / 1 – R\ 2$$

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) = 1, Which lead to 1 / (1 – R2) infinity. To solve this 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.
    - Quantile – Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, Exponential or Uniform distribution.
    - Also, it helps to determine if two data sets come from populations with a common distribution.
    - Advantages:
        - It can be used with sample size also
        - The sample size do not need to be equal.
    - Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
    - Q-Q plot is very useful to determine:
        - If two datasets came from population with common distribution
        - If two datasets have common location and common scale
        - If two datasets have similar type of distribution shape
        - If two datasets have tail behaviour.

- If residuals follow a normal distribution. Having a normal error terms is an assumption in regression and we can verify if it’s met using this.

- Skewness of distribution.