Linear Regression Subjective Questions Answer Sheet

Assignment-based Subjective Questions

1. Available categorical variables were – mnth, weekday, weathersit, season. There are few observations as per boxplot plotted for –
    a. Mnth: During months of May – Oct for the given two years data, there were more booking with the median of 5000 bookings or more per month.
    b. Season: More booking in seasons 2,3, and 4 with the median of 5000 bookings or more.
    c. Weekday: this will have on the average booking almost same through-out all the days with the median of around 4500 bookings.
    d. Weathersit: weathersit 1 have maximum booking of median around 5000 bookings.

2. It is important to use **drop_first=True** during dummy variable creation because n-1 provide the discrete data combination where n is the number of dummy variables to be created. In case adding all the n dummy variables then it will introduce multicollinearity among themselves.
3. Looking at the pair-plot among the numerical variables, temp has the highest correlation with the target variable.
4. To validate the assumptions of Linear Regression after building the model on the training set, we can follow below steps:
    a. Assumption 1: Error terms are normally distributed with mean zero (not X, Y) – we can draw a histogram with training data and check that it is normally distributed.
    b. Assumption 2: There is a linear relationship between X and Y – With plotting pairplot we can check if there is a linear correlation between independent and dependent variables.
    c. Assumption 3: There is No Multicollinearity between the predictor variables – Using heat map and VIF technique, once can check the Multicollinearity between independent variables.
5. Top 3 features contributing significantly towards explaining the demand of the shared bikes are - temp, yr and weathersit

General Subjective Questions

1. Linear regression algorithm: It a technique to machine learning modeling where it try to find out the best fit linear line between target variable and independent variables. There are two types:
    - Simple Linear Regression: where there is only one independent variable present.
      Equation: y = b0 + b1*x
    - Multiple Linear Regression: where there are more than one independent variables exist in the data.
      Equation: y = b0 + b1*x1 + b2*x2 …
      Where y is the target variable, x1, x2.. are the independent variables, b0 is the constant interceptor and b1,b2… are the slope of the line
2. **Anscombe's quartet** comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines

but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading. **Anscombe's quartet** is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. **Pearson's R**: It's nothing but correlation coefficient that has value between -1 and 1. 1 means there is a positive linear correlation between two variables and -1 means it's a negative linear correlation. 0 means there is no linear correlation in between.
4. Sometimes variables are at different scale e.g. turnover is in crores and sales units is in units so that will calculate coefficients which are difficult to interpret. In that case it is required to scale all the required variables values. There are two popular ways of scaling:
    a. Standardizing: scales such that mean is 0 and standard deviation is 1.
    x_scaled = (x - mean(x)) / std(x)
    x – original feature, mean(x) – feature's mean, std(x) – standard deviation
    b. MinMaxScaling: scales in such a way that value lies between 0 and 1
    x_scaled = (x - min(x)) / (max(x) - min(x))
5. VIFi = 1/ (1- Ri-square) for i-th variable
   In case R-square value become 1 then denominator become 0 which led to VIF value as infinity. R-square = 1 means perfect correlation between the variables that should not be a case and we need to drop one of those variables causing perfect multicollinearity.
6. The Q-Q plot is quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution. It's sort of scattered plot where id both quantiles belong to same distribution then it appears as almost a straight line in the scattered plot.