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)

   Ans: i) Season variable value 3 (Fall season) results in more bike booking as compared to other seasons ii) Months 6, 7, 8 (Jun, Jul, Aug) has more bookings iii) Holidays have more number of bookings

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

   Ans: it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables

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

   Ans: variable temp has highest corelation

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

   Ans: It was validated using Residual Analysis of (y_train - y_train_cnt) by plotting histogram. Error terms are normally distributed for the given model which is one of the assumption of the linear regression

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

   Ans:

   i) Temperature – higher the temperature good booking number

   ii) LightRain

   iii) Windspeed

General Subjective Questions

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

   Ans: While building a linear model, you assume that the target variable and the input variables are linearly dependent.

   X (input variable) and Y (target variable) should display some sort of a linear relationship.

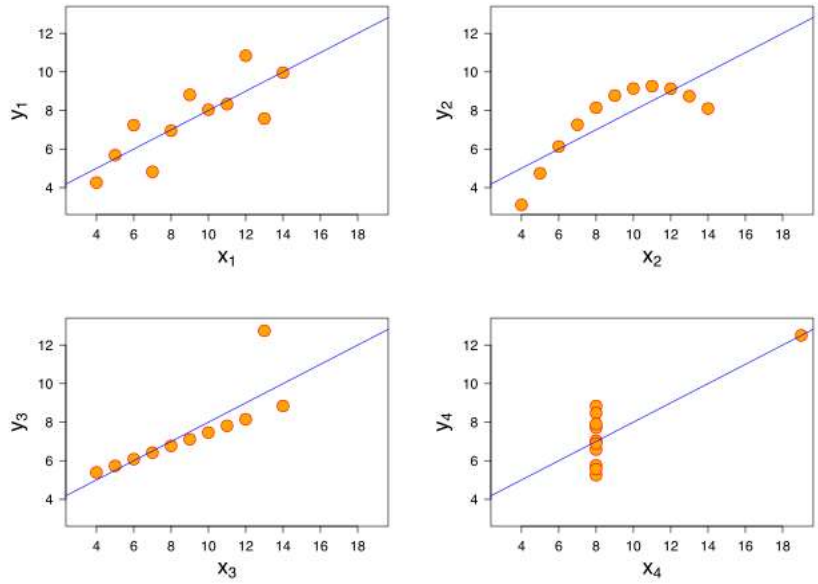
   $Y = \beta0 + \beta1X1 + \beta2X2 + ... + \beta pXp + \epsilon$. Here Y is target variable, β0 is constant, β1 is slope for X1…

   P values of the coefficients tell us whether the coefficient is significant or not. To generate the model, dataset is divided into test, train dataset, numerical features are scaled, categorial converted into dummy variables, model is trained on train dataset with different features before finalising the one having significant coefficients.

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

   Study Reference: **Wikipedia**:

   Anscombe's quartet comprises four data sets that have nearly **identical simple descriptive statistics**, yet have **very different distributions and appear very different when graphed**. Each dataset consists of eleven (x,y) points. Below figure shows what it means:



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

   Study Reference: **Wikipedia**:

   Ans:is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation

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

   What: Scaling is a technique to standardize the independent features present in the data in a fixed range.

   Why: easy interpretation of the coefficients, It is performed during the data pre-processing to handle highly varying magnitudes or values or units

   Normalization

   Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]

   Standardization

   Feature standardization makes the values of each feature in the data have zero mean and unit variance.

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

Ans: It indicates a perfect corelation with other variable. In this case the other variable can be removed and VIF can be calculated again to see its impact of the removal of variable

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

Ans: Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.[1] First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.

If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.

Uses:

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q–Q plots can be used to compare collections of data, or theoretical distributions. The use of Q–Q plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions. A Q–Q plot is generally a more powerful approach to do this than the common technique of comparing histograms of the two samples, but requires more skill to interpret. Q–Q plots are commonly used to compare a data set to a theoretical model.[2][3] This can provide an assessment of "goodness of fit" that is graphical, rather than reducing to a numerical summary.