Linear Regression Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

   Answer: Median bike rents are increasing year on as year 2019 has a higher median then 2018, it might be due the fact that bike rentals are getting popular and people are becoming more aware about environment.
   - Working and non-working days have almost the same median although spread is bigger for non-working days as people might have plans and do not want to rent bikes because of that.
   - Clear weather is most optimal for bike renting, as temperate is optimal, humidity is less, and temperature is less.
2. Why is it important to use **drop_first=True** during dummy variable creation?

   Answer: A variable with n levels can be represented by n-1 dummy variables. So, if we remove the first column then also, we can represent the data. If the value of variable from 2 to n is 0, it means that the value of 1st variable is 1.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

   Answer: 'temp' has highest correlation coefficient of 0.63.
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

   Answer: By plotting the residuals distribution. It came out to be a normal distribution with a mean value of 0.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

   Answer: The following are the top 3vfeatures contributing significantly towards explaining the demands of the shared bikes:
   - atemp(0.412)
   - yr(0.236)
   - weathersit Light rain(-0.275)

General Subjective Questions

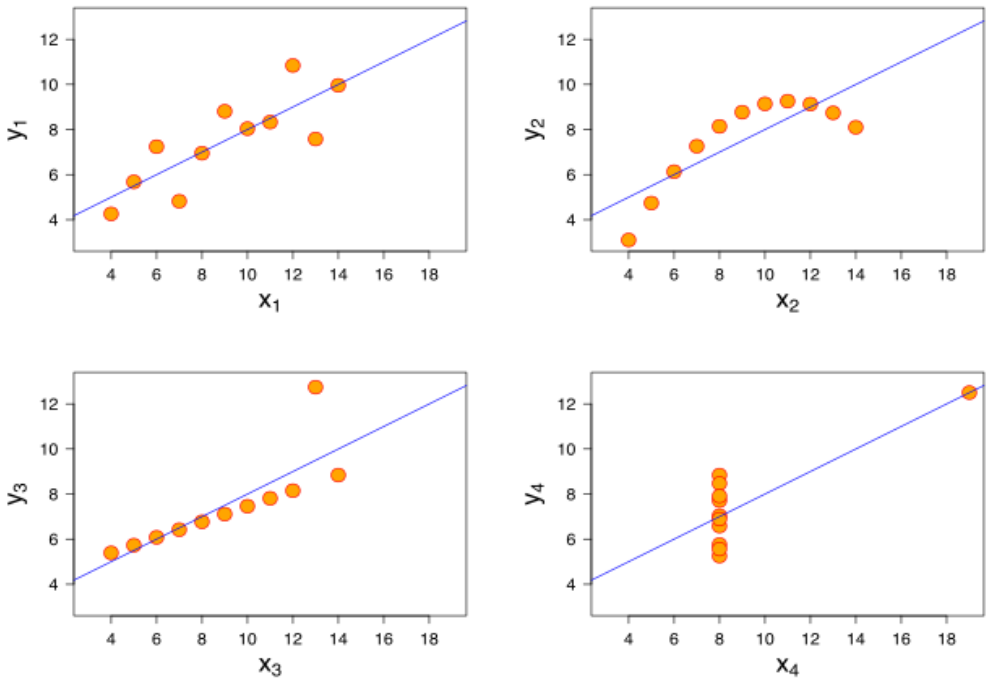
1. Explain the linear regression algorithm in detail.

   Answer: A linear regression algorithm tries to explain the relation between independent and dependent variable using a straight line. It is applicable to numerical variable only. Following steps are performed while doing linear regression.
   - The dataset is divided into test and training data.
   - Train data is divided into feature(independent) and target (dependent) datasets
   - A linear model is fitting using the training dataset. Internally the api's from python uses gradient descent algorithm works by minimising the cost function. A typical example of cost function is residual sum of squares.

- The predicted variable is than compared with test data ans assumptions are checked.

2. Explain the Anscombe's quartet in detail.

   Answer: Anscombe quartet comprises of four data sets that have nearly identical simple descriptive statistics but have quite different distribution when visualized graphically. The simple statistics consist of mean, sample variance of x and y, correlation coefficient, linear regression line and R-Square value. Anscombe's Quartet shiws that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed. The graphs are shown below.



3. What is Pearson's R?

   Answer: Pearson's R measure the strength of association of two variable. It is the covariance of two variables divided by the product of their standard deviation. It has a value from +1 to -1.

   - A value of 1 means a total positive linear correlation. It means that if one variable increase then other will also increase.
   - A value of 0 means no correlation.
   - A value of -1 means a total negative correlation. It means that if one variable increase then other will decrease.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

   Answer: Scaling of a variable is performed to keep a variable in certain range. Scaling is a pre-processing step in linear regression analysis. The reason we scale a variable is to make the computation of gradient descent faster. The step size of gradient descent is generally low for accuracy, if the data has some small variables and some big variables than the time taken by the gradient descent algorithm will be huge.

Normalised Scaling	Standardized Scaling
Called min max scaling, scales the variable such that the range is 0-1	Values are central around mean with a unit standard deviation
Good for non-gaussian distribution	Good for gaussian distribution
Value id bounded between 0 and 1	Value is not bounded
Outliners are also scaled	Does not affect outliners

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Answer: The formula for VIF is

   $$VIFi = 1/1-Ri2$$

   Basically, if R square is 1 than VIF becomes infinite. It means that there is perfect correlation between the features.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Answer: A Q-Q plot is a scatter plot of two sets of quantiles against each other. Its purpose is to check if the two sets of data came from the same direction. It is a visual check of data. If the data is from same source than the plot will appear as a line.