General Subjective Questions:

1. Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

   Mathematically, we can write a linear regression equation as:

   y = a + bx

   Where a and b given by the formulas:

   $$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

   $$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

   Here, x and y are two variables on the regression line.

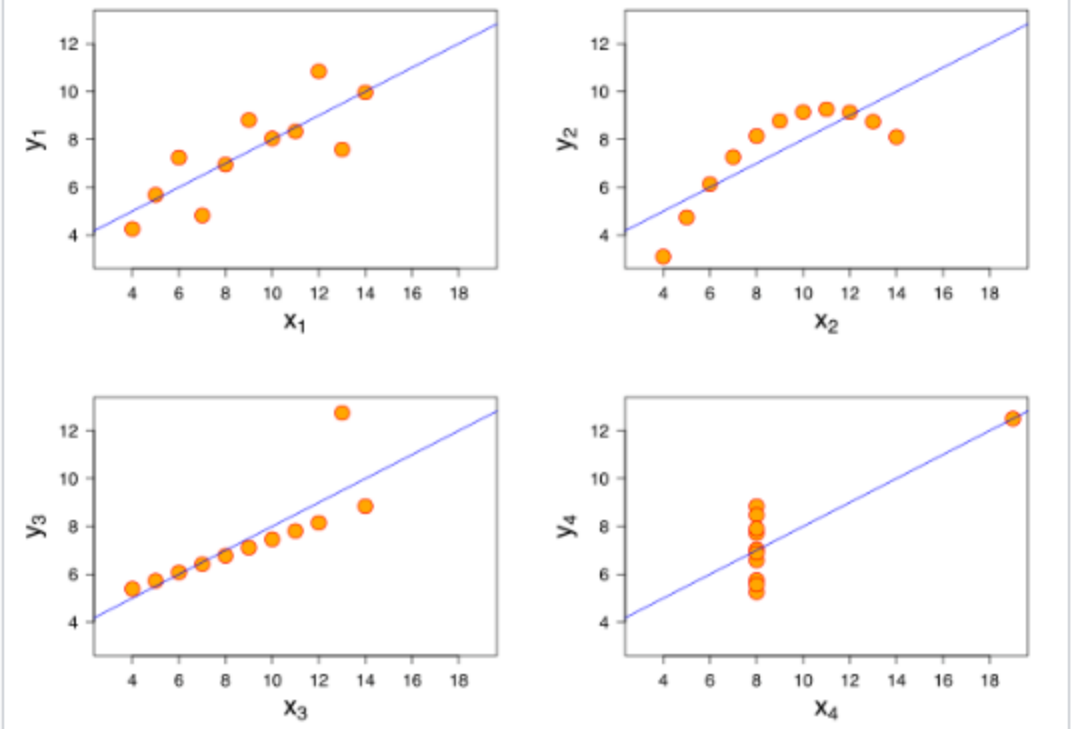
   b = Slope of the line

   a = y-intercept of the line

   x = Independent variable from dataset

   y = Dependent variable from dataset

2. 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. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

3. The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

4. It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

   Most of the times, collected data set contains features highly varying in magnitudes, units, and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:

X_new = (X - X_min)/(X_max - X_min)

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

X_new = (X - mean)/Std

5. If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6. Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. 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. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

   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.

Assignment-based Subjective Questions

1. Categorical variables of season, month, holiday, weathersit have an impact on determining the overall count of bikes rented out. The details are as follows:
   - season_4: In winters people prefer to bike more.
   - mnth_9: From September onwards bike rentals start to soar.
   - season_2: People prefer renting bikes in the summer.
   - weathershit_2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist: Higher conditions as such, lower are the bike rentals.
   - mnth_8: August: Bike rentals in August contribute to higher bike rental numbers.
   - holiday_1: On holidays bike rental is lower.
   - Weekday_6: Saturday's contribute to higher bike rentals.
   - workingday_1: Working days have higher bike rentals.

- mnth_10: Rentals in October are positively correlated with the final bike rental counts.

2. drop_first = True, helps in avoiding another unnecessary variable creation. If there are n categorical values for a given feature, it creates dummy columns for n-1 of them. If all the n-1 columns have the value as 0, this implies that the for the given row, the value is the dropped categorical value.
3. Temperature has the highest correlation with count of bike rentals as observed from the pairplot.
4. Residual analysis of the training data was done to understand the normal distribution of the error terms. P value of the variables and the VIF of the variables were calculated to understand the significance of the predictors.
5. The top three features are:
   - temp: Days when the temperature is warmer, the number of rented bikes is higher.
   - weathershit_3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds. When conditions are such the count of bikes rented decreases.
   - yr: 2019 has witnessed increased number of bike rentals than in 2018.