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)
   - The dependent variable 'cnt' is affected by the categorical variables
     - workingday
     - temp
     - Saturday
     - spring
     - hum
     - windspeed
     - Mild Weather
     - yr

This model gives an r-squared score **0.80 - 0.81** w.r.t test data - training data

There is a positive correlation for the following features

- temp
- workingday
- Saturday (53 -> 52)

We see the variables increased from 2018 and 2019

There is a negative correlation for the following features

- Spring
- Hum
- Windspeed
- Mild Weather

We also see the variables decreased / same from 2018 and 2019
These could also explain why there has been an increase in bikes used in 2019

It looks like favorable climate or environment is the main driving factor for using bikes
There might be people using bikes for their daily work as `workingday` is one of the features with a positive correlation and an equal amount of people using it on `Saturday` for leisure

Business should look to see if people working from home is one of the factors in reducing sales

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   - It is not important but rather efficient to have a smaller set of data to perform the regression analysis.
   - This will reduce the amount of time required by the algorithm to find the best fit line with the given data points

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   - The registered variable has the highest correlation with the target variable but as these were removed as part of the model creation the next highest correlation was the temp variable
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   - The model created was then assessed with the test set that was separated from the initial dataset. This gave a similar r2 value as the training set.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   - Workingday
   - Temp
   - Weathersit (3)
   - yr

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   - Linear regression is based on supervised learning algorithm in machine learning. The basic idea is to find a relationship between a target variable to a set of independent variables.
     For a target variable y and the set of independent variables $X_1 \ldots X_p$
     The linear regression follows the equation:

$$y = \boldsymbol{\beta_{0+}}\boldsymbol{\beta_1 X_1} + \boldsymbol{\beta_2 X_2} + \boldsymbol{\beta_3 X_3} + \cdots + \boldsymbol{\beta_p X_p} + \boldsymbol{\varepsilon}$$

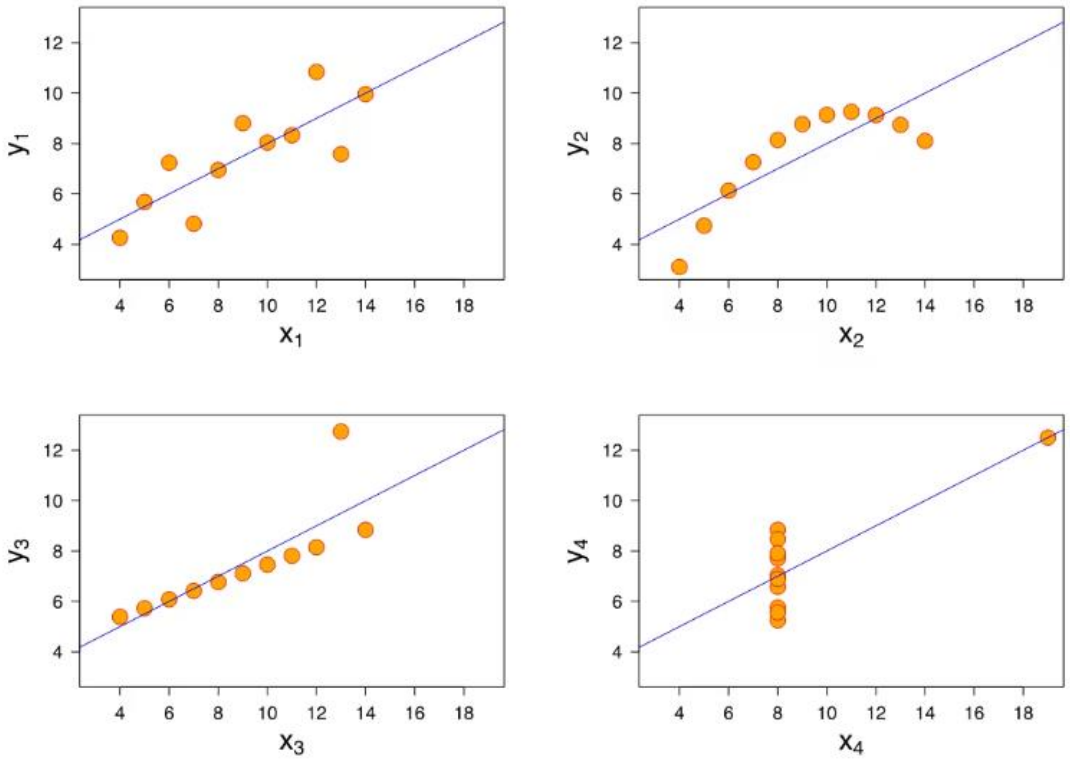
     where,
     β_0 – is the intercept
     β_1 – is the slope or gradient also called the coefficient for the independent variable X_1
     These values are derived by using the cost function like gradient descent which identifies the optimum value of β from a set of data points.
     Once completed a line is identified in the plane (hyperplane for multivariable regression) that best represents the data.
     There are caveats in doing this and not all data can be regressed. The main requirements are
     a There is linear relationship between the target and independent variable
     b The error terms are normally distributed around the mean 0
     c Error terms don't show a pattern and are independent of each other
     d Error terms should show consistent variance

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

- This follows the idea that different datasets might have similar statistical output however when plotted in the graph, they may come to be different

  The four (quartet) common variations are:

  1. Linear relationship between x and y
  2. Non-linear relationship between x and y
  3. Perfect linear relationship except for a few outliers
  4. Most of the values are on a constant for x except for one outlier



This quartet illustrates the importance of plotting data because of the inadequacy to describe the dataset just by the statistical properties

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

- Is an algorithm to identify correlation and describes the strength and direction of the linear relationship between two variables.
  It ranges is between -1 and 1 to indicate whether the relation between two variables x and y, is negatively correlated (closer to -1) or positively correlated (closer to 1) or not correlated (closer to 0)
  The Pearson's R can be used when the data is:
  1. Variables are quantitative
  2. Normally distributed
  3. Data has no outliers
  4. Relationship is linear

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

- Scaling is the process where independent variables with different scales are normalized to a similar range of values.
  If we have one variable indicating the annual salaries which can range from 100,000 and above and another variable age which is between 23-65, the scales between these two variables are vastly different and inferring any sort of significance will be overshadowed by the larger variable. In order to avoid this the larger variable – i.e. Annual Salary is scaled to a smaller range.

- **Normalized scaling** is given by the below:

$$x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

  Also called MinMax scaling, the variable is reduced to range between 0 and 1. The variable x is reduced by the minmum value of x and divided by the distance between max and min of x. The advantage in this method is that it removes the outliers

- **Standardization scaling** is given by the below:

$$x = \frac{x - mean(x)}{sd(x)}$$

Standardization basically brings all the data into a standard normal distribution with mean zero and standard deviation one. The advantage of this method is that it doesn't compress the data and is useful if there are extreme data-points

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

- VIF tells the relationship of one variable with all other variables.
  It's defined by the formular below:

$$VIF_i = \frac{1}{(1-R_i^2)}$$

  Essentially when the R_i^2 approaches 1 the VIF value becomes inf

This indicates a variable can be perfectly predicted by another variable in the model

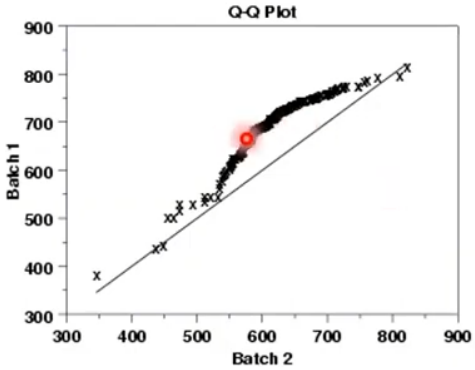
This is resolved by removing this or the other variable from the model

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

(3 marks)

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two datasets come from populations with a common distribution

In a Q-Q plot there is 45% reference line between the x-y plot.



If two sets of data come from a population with the same distribution, the points will fall approximately along the 45% line. Alternatively, if the datasets are further from this reference line, then the data-sets don't come from the same distribution set.

This is useful when we have training and test data set separately. In such scenarios we can confirm that both come from population with similar distributions by investigating the q-q plot. If the points fall closer to the 45% reference point then we can consider them to be from the same distribution