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?
    a. Below are the inferences drawn from the analysis of categorical variables (i.e., season, year, month, holiday, weekday, working day, weathersit
        i. There isn't significant change of the number of rides based on weekday, workingday, holiday.
        ii. Season-3 (fall) has the highest demand for rental bikes
        iii. Demand for bikes has grown from 2018 to 2019 significantly
        iv. Demand for bikes is increasing from jan till sept and peaks in Sept and gradually decreases thereafter.
        v. Bikes have more demand during the weathersit=1 (i.e., Clear, Few clouds, Partly cloudy, Partly cloudy)

2. Why is it important to use **drop_first=True** during dummy variable creation?
    a. Dummy variables take the value 1 if a particular condition is met and a value of 0 otherwise. If there are n different classes, if none of the conditions are met, that is all the values for those **n-1** classes are 0 that ultimately makes the value of **n^{th}** class to be 1. Therefore the number of dummy variables for n different classes must equal n-1 and this is performed by using the argument **drop_first=True** in get_dummies method.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    a. Atemp/temp – feeling temperature/temperature

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    a. I have checked for the following assumptions
        i. Linearity check
        ii. Error terms are normally distributed with mean 0
        iii. Error terms are not following any pattern.
        iv. Error terms have 0 variance homoscedascity

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    a. Since it was asked as influencing and its not mentioned weather a positive or negative influence, I am listing the top 3 features in terms of magnitude in decreasing order.
        i. Atemp/temp (+ve influence)
        ii. Weathersit_3 (poor weather situation) (-ve influence)
        iii. Year (+ve influence)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Ans: Linear regression is a statistical method that models the **linear relationship** between a dependent variable and one or more independent variables. It is a way of finding out how two or more things are related to each other. It can be used to **predict** the value of the dependent variable based on the values of the independent variables.

   The formula for a simple linear regression between a dependent variable y and independent variable x is:

   $$y = \beta_0 + \beta_1 x$$

   Where β_0 is the intercept, β_1 is the slope/rate of change of y wrt x/ beta co-efficient

   The formula for a multi linear regression is:

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \in$$

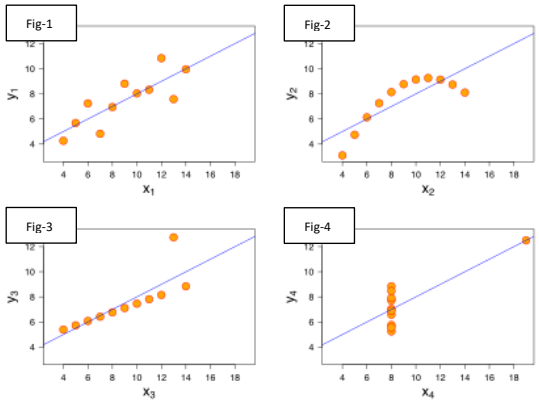
   where y is the predicted value, β_0 is the intercept, $\beta_1, \beta_2, \ldots \beta_p$ are the beta co-efficients and $\in$ is the error term

   The beta co-efficients can be interpreted as change in mean response (E(y)) per unit change in that variable(feature) when other predictors are held constant.

2. Explain the Anscombe's quartet in detail

   Ans: Anscombe's quartet is a group of four datasets that have the same mean, standard deviation, and regression line, but look very different when graphed. It shows the importance of visualizing data before applying various algorithms to build models on it.

   The importance of data visualization and how simple it is to trick a regression algorithm are both made clear by Anscombe's quartet. Therefore, we must first visualize the data set in order to help in the development of a well-fit model before attempting to analyze, model, or apply any machine learning method.



Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

(source: Wikipedia)

As you can see from the above graphs, the same linear regression line was output by the model for different datasets with almost exact or similar mean, variance, r-squared. However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm.

We can describe the four data sets as: **ANSCOMBE'S QUARTET FOUR DATASETS**

- **Data Set 1**(fig-1)**:** fits the linear regression model pretty well.
- **Data Set 2**(fig-2)**:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3**(fig-3)**:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4**(fig-4)**:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

3. What is Pearson's R?

   Ans: Pearson's r or also called as Pearson correlation coefficient or Bivariate correlation or also know The correlation coefficient. Although it has many names, Pearson's r is as a measure of how close the observations are to the line of best fit.
   Pearson's R can have values between -1 and 1. Coefficient of -1 means perfect negative correlation with the target variable & coefficient of 1 means perfect positive correlation and 0 means no correlation with target variable.
   Pearson's r is calculated by dividing the covariance of the two variables by the product of their standard deviations.

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

   Ans: Scaling is a process to normalize the data within a particular range. In multi linear regression, the features might be of different magnitude and once we train a linear regression model on them, the Beta coefficients that we get will also vary in high magnitudes based on the influence of each variable. And so we cannot determine which feature has more influence on the

target variable. To mitigate this issue, we perform scaling to get features into a similar range. There are 2 types of scaling Normalized & Standardized.
In **Normalized** scaling we map the min to 0 and max to 1. Hence all the values of that feature will lie in between [0,1]. This is also called MinMaxScaling.

$$z = \frac{x - x_{min}}{x_{max} - x_{min}}$$

In **Standardized** scaling, we transform the data to have mean of 0 and standard deviation of 1. We technically don't enforce any range in this.

$$z = \frac{x - \mu}{\sigma}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Ans: VIF means Variance Inflation Factor is a measure of multicollinearity between the predictor variables in Linear Regression. Multicollinearity occurs when two or more features considered for linear regression are highly correlated with each other so that they don't provide unique or independent information in the regression model.
   Usually, the VIF starts from 1 and has no upper limit. Heuristically, VIF threshold value is considered as 5 (and some other times 2)
   - VIF of 1 means that there is no correlation between that feature and other predictor variables.
   - A VIF value between 1 and 5 means there is moderate correlation between that feature and other predictor variables.

   Usually we repeat the linear regression by dropping the features with high VIF and p-value and repeat the same process until the VIF goes below 5 (heuristic value).

$$VIF = \frac{1}{1 - R^2}$$

   A VIF of infinite means that the particular feature can be perfectly predicted by other predictor variables considered. In other words, that particular feature has a perfect linear regression with other predictor variables. That means its R^2 would be almost equal to 1. So, from the equation above as R^2 approaches 1, VIF approaches infinite.

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

   Ans: Q-Q plot, also known as Quantile-Quantile plot, is a graphical tool that helps us compare two sets of data and see if they come from same distribution or not. It helps us asses, if a set of data plausibly came from some theoretical distribution like Normal, exponential or uniform. It plots the quantiles of one data set against the quantiles of another data set.
   This helps us in the scenario of linear regression to confirm if both the data sets are from populations with same distributions, when we receive the training and test data set **separately**.
   Also, in linear regression, we use Q-Q plots to see if the residuals (the differences between the real and estimated values) are normally distributed or not. This is one of the conditions we need

to meet before we can use the model to make predictions. If the residuals are normal, then the Q-Q plot should be close to a straight line with a 45-degree slope. If the points are away from this line, then it means that the residuals are not normal and there might be some issues with the model.