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: The categorical variables can further be categorized as either nominal categorical variables. Nominal variables are variables that have two or more categories, but which do not have an intrinsic order. For example, a real estate agent could classify their types of property into distinct categories.

   Each level of the categorical variable to the mean of the **dependent** variable at for the reference group, it makes sense with a nominal variable. However, it may not make as much sense to use a coding scheme that tests the linear effect of race. As we describe each type of coding system, we note those coding systems with which it does not make as much sense to use a nominal variable.

   Ref: https://statistics.laerd.com/statistical-guides/types-of-variable.php

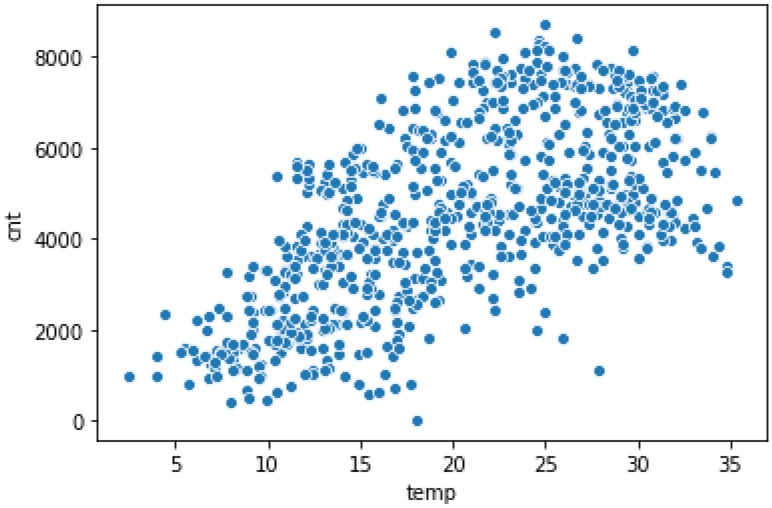
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

   Ans: Because one level of your categorical feature becomes the reference group during dummy encoding for regression and it is redundant.

   So a categorical variable of K categories we remove the first variable to remove the redundant as k-1 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: The variable which has strong correlation with cnt, which is temp.



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

   Ans: Once we run assumptions on the training set then weI need to evaluate the test set by dropping the same variables.

- After that we need to predict the model using the last train model and X_test dropped variable data set.
- Use of r2_score method and pass the test and predicted test value and check the response.
- Evaluated valued should not have much difference of you last predicted train model R2
- Then it will be your final evaluated model.

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: 0.5499*temp + 0.2331 * yr + 0.1318 * winter (So these are more significant towards explaining the demand of shared bikes.).

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Ans: Linear Regression is a machine learning algorithm. It performs a regression task. Regression models a target prediction value based on independent variables.
   So this regression technique finds out a linear relationship between x (input) and y(output)

   And predict the output values based on input features from the data given in the system. This algorithm builds a model on the features of training data and using the model to predict the value for new data.
2. Explain the Anscombe’s quartet in detail. (3 marks)

   Ans: I have referred to this link for the answer:
   Ref: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

3. What is Pearson’s R? (3 marks)

   Ans: Pearsons R is a measure of the strength of the association between the two variables. To determine the correlation between variables is significant, compare the p-value to your significance level.
   To know more about the strong correlation between two continuous variables is to draw a scatter plot of the variables to check for linearity.

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

   Ans: The feature of scaling means data normalization. It is a step of Data Pre Processing which is applied to independent variables or features of data.
   Why: It basically helps to normalise the data within a particular range. It also helps in speeding up the calculations in an algorithm.
   Difference: Normalization and standardization are used interchangeably, but they usually refer to different things. Normalization usually means to scale a variable to have a value **between 0 and 1**. But the standardization transforms data to have a **mean of zero and a standard deviation of 1**

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Ans: When we start with all variables, and proceeds by repeatedly deselecting variables showing a high VIF. 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. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   Ans: A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. An example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.
   Ref: https://data.library.virginia.edu/understanding-q-q-plots/