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?**

   For the casual, the dependent categorical variables are holiday, working day, season, year, month and weather. So, we can say that it has high dependance of categorical variables with most likely having features such as summer from season, September, October from months, light snow from the weather.

   For the registered, everything is similar except for few features such as no influence of October and holiday is required.

   For the cnt, it is similar to registered but October is influential.

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

   Let us consider that we have n values for a categorical label. So, that means we require n-1 variables, which are known as dummy variables. So, the drop_first = True drops the first value and considers from the second.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   Considering cnt as the target variable, cnt has high correlation with “temp” and “atemp” and then “year (2019)”

   Considering registered as the target variable, registered has high correlation with “Year (2019)” then “temp” and “atemp”

   Considering casual as the target variable, casual has high correlation with “temp” and “atemp”

   But, if we consider the data aggregated and cnt as the target variable, it has the highest correlation with “registered”. If we consider registered as the target variable, it has the highest correlation with “cnt” and for casual, the highest correlation is with “cnt”.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   After building the model on the training set, we evaluate the model based on the r-squared value with its test model. The test model will be created by transforming the training model with the test set. We then compare the r-squared values of both. The difference should be minimum.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

   For the analysis that I’ve made, the top three features would be “holiday”, “workingday” and “temp”.

GENERAL SUBJECTIVE QUESTIONS

1. **Explain the linear regression algorithm in detail.**

   Linear Regression Algorithm is a machine learning algorithm based on supervised learning. In this algorithm, we basically train a model based on one or more variables to predict the behaviour of the data. If we use a single variable, then it will be called as a simple linear regression. If we use more than one variable, we should call it as a multi linear regression.

   The equation for a simple linear regression will be y = aX+b+E where b is the intercept and a will be the slope, y is the target variable or dependent variable where as the x will be the independent variable or simply a feature and E is the error.

   For multi linear regression, the equation will have multiple features with multiple slopes or coefficients. The equation is y = a0X0 + a1X1+ … + anXn + E.

   In order to achieve a better model, the cost has to be reduced. This can be done with the help of residual sum of squares method (RSS). This has to be done iteratively until the cost reaches global minima. This can be done with the help of Gradient Descent Algorithm (GSD). Sklearn library uses gsd in the backend and the linear model can be aquired easily.

2. **Explain the Anscombe's quartet in detail.**

   Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. So, it's always better to visualize data in order to get a clear picture of the dataset.

3. **What is Pearson's R?**

   Pearson's correlation coefficient is an effective tool to find the correlation among the data. It's known as Pearson Product Moment Correlation (PPMC). It shows the linear relationship of two sets of data. It is represented by rho or "r".

   The problem with PPMC is that it cannot differentiate dependent variables with the independent one. It will not give you any information about the slope the line. It just provides the relationship.

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

   In order to get the data under the same scale, scaling should be done. This should be done to a numerical data and not for categorical ones. If the scaling didn't take place, there will an issue in the model as the features have to be multiplied with the coefficients and if the scale is different, the values will be far more different from the previous ones and hence there will be a huge difference with the r-squared and prediction.

There are two types of scaling. Normalized Scaling (Min-Max Scaling) and Standard Scaling. The Standard scaler scales the values between -1 and 1 where as Normalized scaler scales the values between 0 and 1. The Normalized scaler is pretty popular.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

   This happens because of the high correlation. 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.**

   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.

   In order to visualize residuals vs the original and visualize the test and training set features simultaneously, a Q-Q plot is an efficient one in Linear Regression.