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?

   The categorical variables in the dataset are: seasons, yr, mnth, holiday, weekday, workingday, weathersit

   a. More people are using the bikes on non-holiday days than holidays and the median of non-holiday is higher than the holiday records
   b. There are lot of data on the working day but the median for working day and non-working day are the same
   c. We see that there is record for every month and year
   d. Other than the season 'Spring', we see all other seasons the usage of the bike range from 4000-7000 cnt
   e. In 2019, more people rented the bike than 2018
   f. We can see that in the month range from May to Oct there is a spike in the usage and the Jan is recorded as the lowest

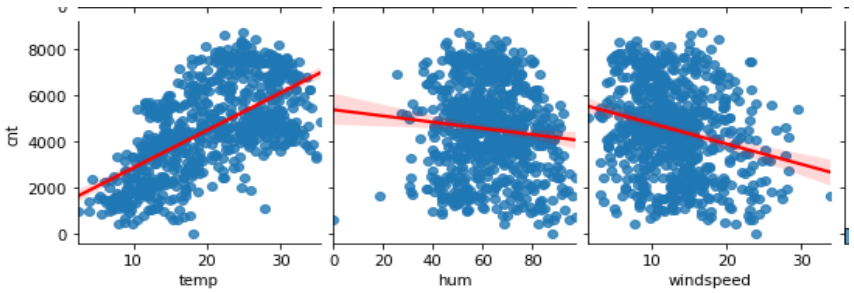
2. Why is it important to use drop_first=True during dummy variable creation?

   drop_first=True will reduce one column in the dummy variable creation (ie first column in the dummy variable created will be removed). We create dummy variable for the categorical value. For example, we have the furniture status as Semi-furnished, furnished, unfurnished categorical value. We create column for each of the category value and assign the value as 1 for the column if matching is present. If we remove unfurnished column and if the value for both the semi-furnished and furnished are 0 it means it is unfurnished. Without this column also we can infer value for unfurnished data.

This reduces the number of columns in the data set for analysis without affecting the required information for the analysis.

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

   We can see that “temp” variable is highly correlated with the target variable “cnt”



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

   The Linear Regression assumptions are:

   a. High correlation between the dependent and the independent variables. This can be verified using the scatter plot, Pearson correlation. The relationship between the dependent and independent variable should be linear.
   b. No or minimal multicollinearity data - This can be obtained by the pearson correlation matrix, VIF. The VIF of all the data should not be more than 5. Else there is multicollinearity in the data. This can be rectified either
      i. by removing the multicollinear data having high VIF or
      ii. by removing other data which can reduce the collinearity if in case, the independent variable is highly correlated with the dependent variable than any other variables in the list.

c. The R2 and Adjusted R2 obtained should not be of much difference. Prob F-statistics should be less than 0. The p-value of all the variables should be less than 0.05. The VIF of all the variable is less than 5
d. Verifying the error term to be normally distributed with mean as 0, homoscedastic with the constant variance

These all verification can be done while building the model with the train set. After the data is trained, the same prediction applied to the test data set. There should not be much difference in R2 value between the test and train data set.

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

The top 3 features of the final model are:

a. Temperature has the coefficient value as 0.5480 being the highest record
b. Weather situation as Light Rain has the negative coefficient value as 0.2838 being the second highest parameter. Even though the negative symbol is provided, the coefficient value takes precedence
c. Year has the coefficient value as 0.2328 as the third best parameter Temperature, Weather situation as Light Rain and Year features are the most required for explaining the demand of the shared bikes.

General Subjective Questions

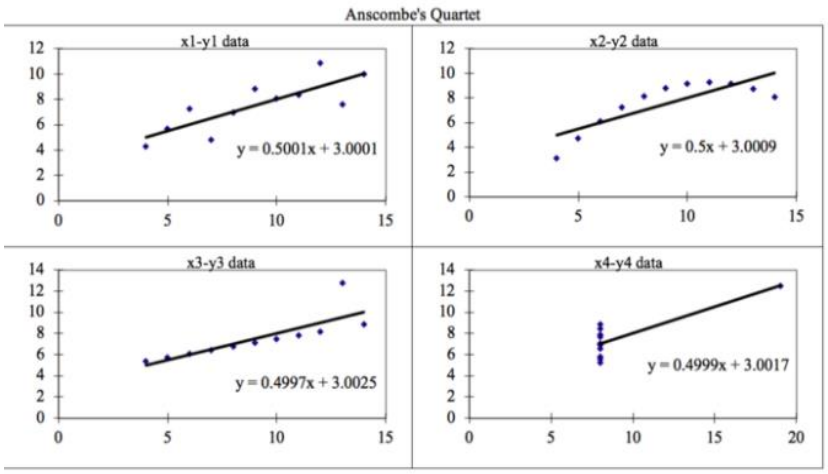
1. Explain the linear regression algorithm in detail.

The linear regression model is the method of finding the best fitting straight line in the data set having the independent and dependent variables. The linear regression is performed on the continuous variable. The relationship between the variables is best explained by

the correlation between the independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.
In the single linear regression, the correlation between one dependent and independent variable is checked. In Multiple linear regression models, the multiple correlated dependent variables and independent variable are predicted. There can be positive or negative relationship between the dependent and independent variables. The best fit line is the taken out of the least squares of errors between the actual and the estimated value.

2. Explain the Anscombe's quartet in detail.

This will illustrate the importance of plotting the graphs before analyzing and model building. Let's take the example of the 4 data set having the nearly same observations as shown in image:



This provides the visual idea of the data distribution before applying the regression model in the data set. The example data taken has the 4 different data having same statistical properties. Let's consider it as same mean and the SD. But when plotted, we can notice that plot1 is linearly related between the dependent and independent variable. While in plot3 we can see some outliers present. Plotting the data gives more visual of the data.

3. What is Pearson's R?

Pearson coefficient is used to find the linear relationship between 2 continuous variables. It is the best method to measure the association between the variables since it is based on the method of covariance. It is the normalized measure of the covariance which always lies in the range of -1 to 1. It gives the correlation between the variables as well as mentions the direction of correlation of the variables.

- The value of 1 displayed represents the data x, y are linearly correlated and all the points lie in the best fit line. Ie. the increase in x with the increase in y.
- Same is for the -1 value. The value x,y are correlated but the increase in x will happen with the decrease in y.
- When the value is 0, it means that the x,y variables are not correlated.

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

Scaling is the technique to standardize the independent features present in the data in fixed range. It is the pre-processing step used to standardize or normalize the data. The values in the dataframe for each column might be of different magnitude. Some columns might have higher range of values; few columns might have values of 0 and 1.

So we need to scale them in a way that all the values are in the same level.

It is easy of interpretation. It uses faster convergence method like gradient descent methods. It doesn't change the p-values or the model accuracy. Distribution is not affected and the outcome is the same.

In short, the scaling methods are used to make sure that the data is internally consistent.

There are 2 types of scaling:

- Standardization
- MinMax scaling

For normalization we use MinMax scaling.

MinMax/Normalisation	Standardisation
Converts the value in the range of 0 to 1	Subtracts the mean and divide by SD. It will be centered at 0 and SD = 1
Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.	Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.
$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$	$X' = \frac{X - \mu}{\sigma}$
The minimum value is 0 and maximum value is 1	μ is mean and σ is SD
Normalization is used when the data doesnot follow the Gaussian distribution	Standardisation is used when the data follows the Gaussian distribution

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

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. If the model has more independent variables which are correlated then the VIF value increases. A large value of VIF indicates that there is a correlation

between the variables. If there is a perfect correlation between the independent variable then the VIF will be displayed as inf. 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.
    - Q-Q plot is used to check the normal distribution of the data.
    - The 2 axes are taken as one being the Data Quantile and other as Normal Quantile. The Data Quantile is taken from the data set and the normal quantile is taken from the normal distribution.
    - If the data were normally distributed then most of the points lie on the line.
    - If all the points of the quantiles lies on or close to the straight line at an angle of 45 degree is called similar distribution
      This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.