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 like year, holiday, workingday,summer , fall , winter , mar, july, sept, thu, partly_cloudy, mist, and light_rain are the most important categorical feature variable for the prediction.

Removing of any of these categorical variable results in decrease in the R2 value.

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

During the creation of dummy variable, if there are m categorical values for a variable , then there are m columns get created each with value 0 or 1.

Now, we drop a column so that when all the value will be 0 for each of the column the resulting value will then be for the dropped column value.

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

Temp variable shows highest correlation with the target variable.

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

The error term graph between the predicted y value and actual y value comes out to be a normally distributed graph.

The coefficients values are also not weird and are between 0 and 1.

This proves that there is a linear relationship between the independent variables and y.

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

The temp, yr and windspeed are the top 3 features that contributes towards the demand prediction of the shared bikes.

These have the highest coefficient values.

1. Explain the linear regression algorithm in detail.

Linear Regression Algorithm

- The Regression algorithm predicts a continuous output variable.
- Linear Regression algorithm is an algorithm which tries to best fit a line passing through most of the points.
- The linear regression algorithm is part of the supervised machine learning algorithm, i.e. it tries to predict the output based on the given data.

The linear regression equation is -

$Y = b0 + b1X1 + b2X2 + b3X3 + \ldots$

Here,

Y is the output variable

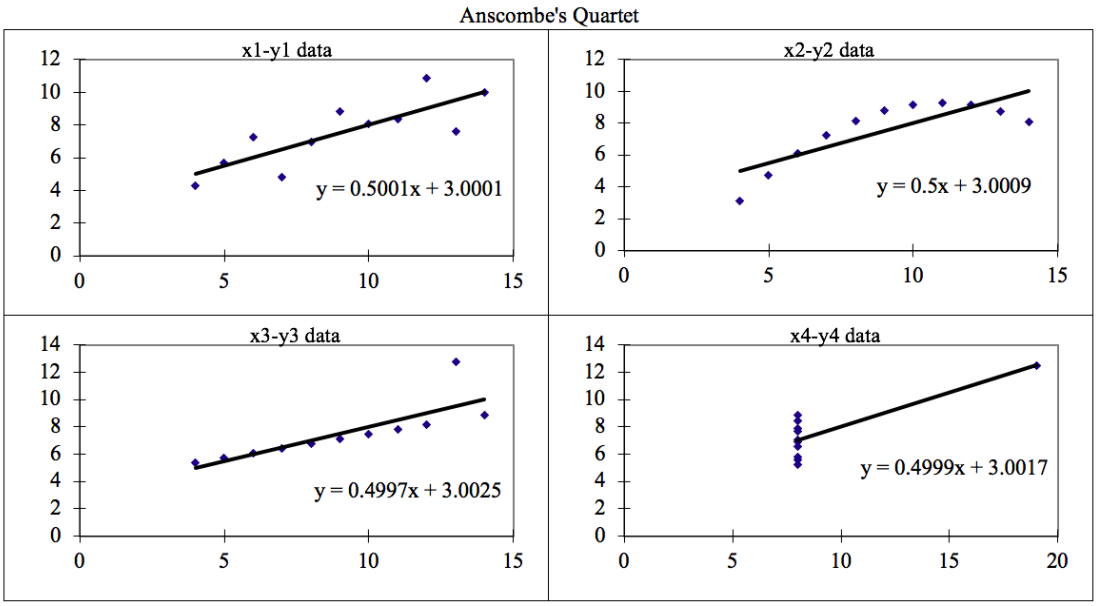
X1, X2, X3 variables are the independent variables which are used to determine the output.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet

Anscombe's Quartet defines the four data sets that when visualized have a complete different data points pattern and the regression line fitted .

All the four regression line fitted have a very similar equation but they differ in actual equation.



Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

Hence, data visualization is important before fitting any regression line which will help in creating a better fit of the data points.

3 . What is Pearson's R?

Pearson's R

The Pearson correlation coefficient (r) is the most widely used correlation coefficient.
It describes the strength and direction of the linear relationship between two quantitative variables.
The numerical values of the correlation coefficient lies between -1.0 and +1.0.
Using the formula proposed by Karl Pearson, we can calculate a linear relationship between the two given variables.

There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

The formula given is:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

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

Scaling

Scaling is the conversion of the variable value to a particular range like between 0 and 1 so that there is no dependency between the variables.

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret.

So we need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

You can scale the features using two very popular method:

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

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

- If all the independent variables are orthogonal to each other, then VIF = 1.0.
- If there is perfect correlation, then VIF = infinity.
- A large value of VIF indicates that there is a correlation between the variables.
- If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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

Q-Q plots are also known as Quantile-Quantile plots.

In Statistics, Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other.

If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

Q-Q plots are also used to find the Skewness (a measure of “asymmetry”) of a distribution.

When we plot theoretical quantiles on the x-axis and the sample quantiles whose distribution we want to know on the y-axis then we see a very peculiar shape of a Normally distributed Q-Q plot for skewness.

We can talk about the Kurtosis (a measure of “Tailedness”) of the distribution by simply looking at its Q-Q plot.

The distribution with a fat tail will have both the ends of the Q-Q plot to deviate from the straight line and its center follows a straight line, whereas a thin-tailed distribution will form a Q-Q plot with a very less or negligible deviation at the ends thus making it a perfect fit for the Normal Distribution.

Q - Q plots 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.