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? (3marks)

Ans:

a) The demand for the bikes is more in summer and fall.

b) The demand for the bikes is less in bad weather conditions such as mist and light snow/rain and highest on clear weather conditions.

c) The demand for the bikes is more on working days and less on holidays.

d) Weekdays have little/almost no effect on demand of bikes.

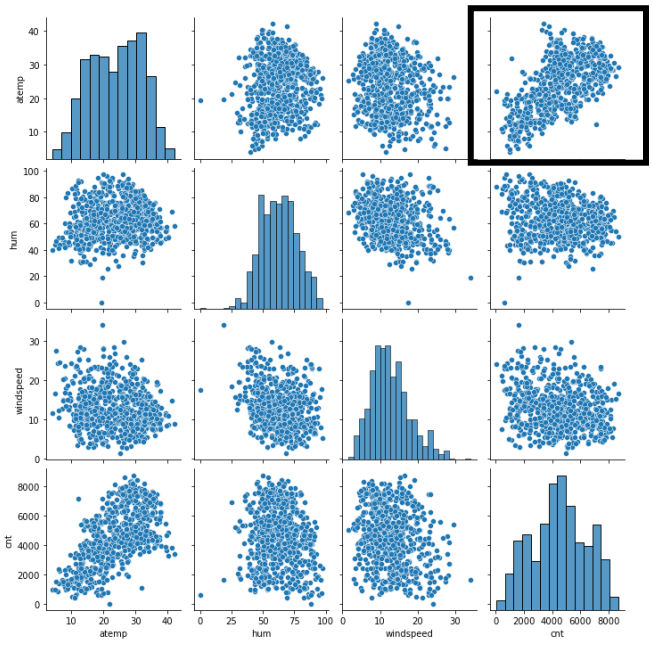
The same was even inferred from EDA analysis conducted in the notebook.

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

Ans: Let us say that we have a categorical feature with three values say A, B and C. A is already defined by not having B and C i.e ***A is defined by B=0 and C=0***. ***So if we do not drop the first column we are creating the features that are having multi collinearity***. Also some more advantages of drop_first is it reduces the computational time for model having more input features.

3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: From the heat map and scatter plots in the notebook we can infer that variables atemp and temp have equal correlation (0.63) with cnt. The same can be inferred from the model coefficients obtained at the end.

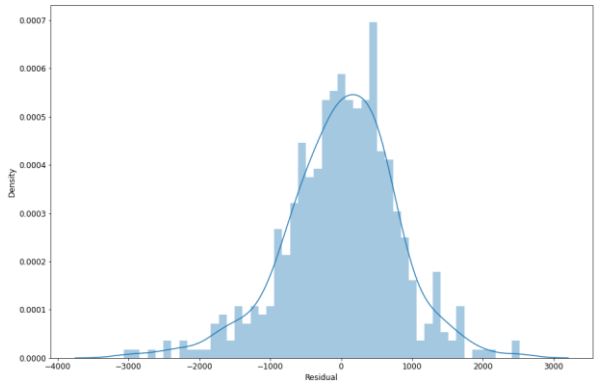


4) How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: The assumptions of linear regression can be validated by plotting residual plot on the train set i.e y_train_actual – y_train_predicted. The assumptions of linear regression are

a) Mean of the residual should be 0
b) The residual should follow a normal distribution
c) There should not be multi collinearity between input features
d) Homoscedasticity – i.e variance of residual is same for all values of y

The same can be inferred from the image below.



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: The top three features explaining the demand of bikes are

a) atemp
b) weathersit – Light snow rain (3)
c) yr

General Subjective questions:

1) Explain the linear regression algorithm in detail. (4 marks)

Ans: The linear regression algorithm attempts to build a model by assuming linear relationship between the input features and target variables. The equation for the same can be written as

$$y = b0 + b1*x1 + b2*x2 + \ldots. + bn*xn$$

Here b0,b1,b2 etc are the parameters which will be found by the algorithm and x1,x2,x3 etc are the input features to the model.

The algorithm uses gradient descent in order to minimize the least squared error between ypred and yactual.

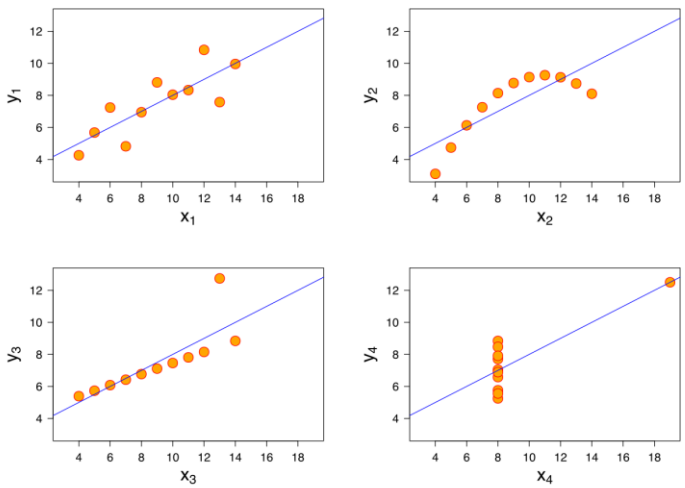
The algorithm can be summarized as

$$\textit{Min } \sum(yact - ypred)^2$$

$$\textit{where y pred} = b0 + b1*x1 + b2*x2 + \ldots. + bn*xn$$

2) Explain the Anscombe's quartet in detail. (3 marks)

Ans: ***Anscombe's quartet as the name suggests consists of four data sets having x and y for which the statistical measures such as mean, variance, correlation etc are same whereas the data is entirely different when plotted***. This gives us a very valuable learning that in order to understand the distribution of data, the data should not be simply compared by statistical measures but should be plotted visually. The image below shows the same. The image is sourced from Anscombe's quartet - Wikipedia. We can also quote the famous quote i.e ***Corelation is not causation.***



3) What is Pearson's R? (3 marks)

Ans: Pearson's R or commonly called as correlation coefficient is a statistical measure that determines the ***strength of linear relationship between two variables of interest***. Its value ranges from -1 to 1. A sign indicated positive or negative correlation whereas the magnitude of value indicates the strength of correlation. The formula is given below

$$R = \frac{\sum (xi - xmean)\,(yi - ymean)}{\sqrt{\sum (xi - xmean)^2 \sum (yi - ymean)^2}}$$

When both y and x are perfectly linearly related the numeration = denominator making

R value to be +1 or -1.

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

Ans: Scaling is mathematical transformation of the features to other values. It is done for better interpretability of results / for the faster training of the model / better performance of the model. ***Some of the machine learning algorithms that are based on distance measures are sensitive to scaling (if the distance of one feature is more than the other during model training, model assigns less weight to that feature) i.e their performance increases if the features are scaled***. The most commonly used scaling techniques are normalization and standardization. In normalization the entire features are scaled to range from 0 to 1 whereas in standardization the mean of the transformed variables is 0 and their standard deviation is 1.

$$\text{Normalization: } \frac{xi - xmin}{xmax - xmin}$$

$$\text{Standardization: } \frac{xi - xmean}{\sigma}$$

Both of the techniques have their advantages and disadvantages. There is no rule to tell which one is the best. It depends on the data distribution.

5) You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans: We know that Variance Inflation Factor (VIF) is given by the formula.

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

For calculation VIF for an input feature, we try to build a linear model by taking the remaining input features. So a higher values of VIF (usually > 5) indicates that the feature has a collinearity with one/more input features. ***A values of infinite indicates that the feature is perfectly related linearly with one/more input features (excluding the one for which the calculation is done)***. So that feature can be dropped while model building as it is explained by other features.

6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans: ***Quantile – Quantile plot commonly called as Q-Q plot is a plot between two variables of interest. If the points fall on y=x line then we can conclude that the distribution of both of the variables is similar. Normal Q-Q plot is a Q-Q plot between standardized normally distributed variable and the variable of interest. The z scores of both the variables are taken and plotted on x and y axis respectively***. If the points of interest are almost lying on the y=x line in the plot then we can say that the variable is normally distributed. If we have good number of points not lying on the line then our distribution is not normal. Q-Q plots are generally used to validate the assumptions of linear regression i.e to check if the residuals are normally distributed or not. The image below (sourced from Q-Q Plots Explained. Explore the powers of Q-Q plots. | by Paras Varshney | Towards Data Science) shows Q-Q plot for perfect normal distribution.