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)
   Answer: yr and rain (derived from weathersit, where weathersit=3) have a good enough effect on the dependent variable 'cnt'. We can also see this from the heatmap, where it is very clear that yr has a strong relationship with 'cnt'. This basically signifies that demand is expected to rise in future even when not considering other factors. Rain however as -ve co-relation, means, more the rain, less the demand.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   Answer: For n values of a category variable, only n-1 dummy variables are required as the 0 in all these n-1 dummy variables signify the last variable. Hence, we can drop any 1 variable when creating dummy variables for a category with n different values.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Answer: Variable "atemp" seems to have the highest co-relation.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Answer: After building the model we use the test set to calculate the predictions (y-test-pred) and thereafter calculate the R-Squared value for y-test-pred comparing with y-test, whose value should be close enough to R-Squared from the training model to confirm our assumptions.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Answer: atemp, rain (derived from weathersit, where weathersit=3) and yr.

General Subjective Questions

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

   Answer: Linear regression algorithm is a machine learning algorithm which builds a model around 1 dependent variable and 1 or more independent variables. The linear regression model can be represented by:

   $y = \beta0 + \beta1X1 + \beta2X2 + ... + \beta nXn$

   where, y is the dependent variable and β0 is the intercept, β1…βn are coefficient and X1…Xn are independent variables

   When there is only 1 independent variable, it is simple linear regression. For more than 1 independent variables, multiple linear regression comes into picture.

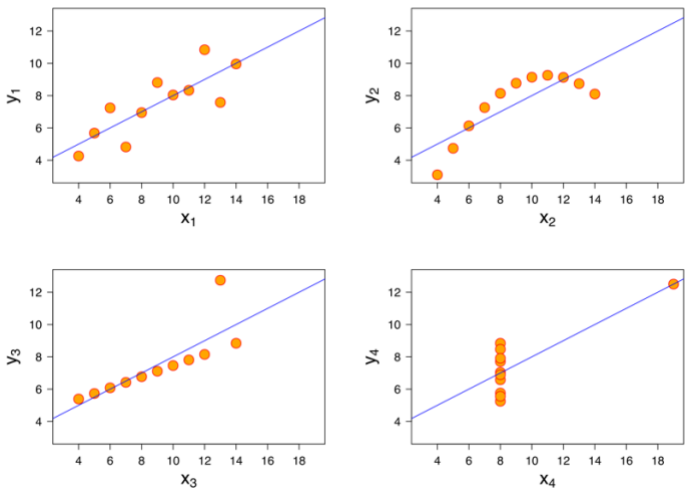
   The basic 6 steps of performing a linear regression are as follows:

   1. Separate X and y
   2. Perform train and test split
   3. Instantiate a model
   4. Train the model
   5. Make predictions on test set
   6. Evaluate the model by MSE, R-Squared

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

   Answer: Anscombe's quartet consists of 4 data sets that have very different distribution but, have same mean, variables and most important, the linear regression line. The aim is to demonstrate the vast difference in data visualization as compared to the numerical properties which are similar for all 4 representations. It shows why study of patterns and outliers affect data analysis.

   The image below shows an Anscombe's quartet.

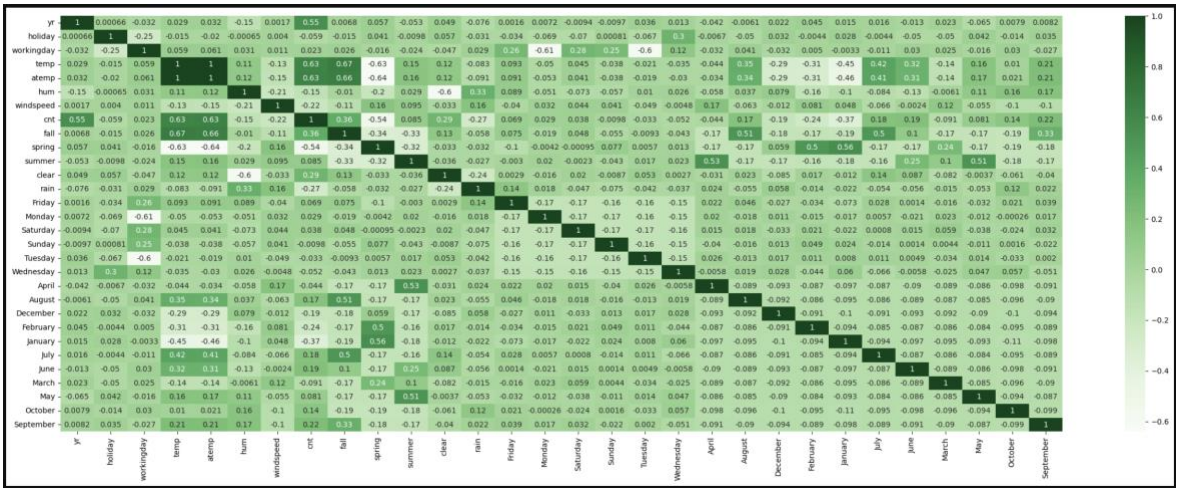


   (source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet#/media/File:Anscombe's_quartet_3.svg)

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

Answer: Pearson R is the measure of corelation between 2 variables ranging between -1 and 1. Here 1 indicates a perfect co-relation and -1 represent perfect negative co-relation (if one variable increases, then the other decreases by in exact same ratio). It assumes the the 2 varialbes have linear relationship and are normally distributed.

The df.corr() function in python gives the pearson co-relation between variables. The following image shows the pearson co-relation on the heatmap generated using sns library for the BikeSharing case study:



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

Answer: Scaling is the process of transforming the values of a variable into a range which is similar for all the other variables in the data-set.

It is performed so that when the model is built, all variables are equally weighted. This is more relevant in case of multiple linear regression models where independent variables can have different ranges, it is important to have all the variables in the same range so that we can evaluate the variables on the same scale.

Normalized scaling or MinMaxScaling scales the data to a specific range, say, 0 and 1. In case of standardized scaling, the data has a mean of 0 when scaled but, no range boundary. Outliers can affect normalized scaling but, not standardized scaling.

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

Answer: In case 2 variables are perfectly co-related, the VIF becomes infinte. VIF is given by:
$1 / (1 - R^2)$, is R^2 becomes 1, the VIF becomes infinte representing perfect co-relation between 2 variables.

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

Answer: Q-Q plot or quantile-quantile plot is used to plot the quantiles of sample distribution and theoretical distribution. It helps in determining, say, whether data is normally distributed.