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?
   Ans:

   - The demand for bike is less in the month of `spring` when compared with other seasons. It has a negative co-efficient.
   - The demand bike increased in the year 2019 when compared with year 2018.
   - Bike demand is less in holidays in comparison to non-holidays. It has a negative co-efficient
   - There is no significant change in bike demand with working day and non-working day

2. Why is it important to use **drop_first=True** during dummy variable creation?
   Ans:
   drop_first=True **helps in reducing the extra column created during dummy variable creation** and it reduces the correlations created among dummy variables

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Ans: from the pair-plot we could observe that, `temp` has highest positive correlation with target variable `cnt`

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Ans: Residual Errors Have a Mean Value of Zero and Residual Errors Have Constant Variance

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans: temp(positive correlation), year(positive correlation) and wind-speed(negative correlation)

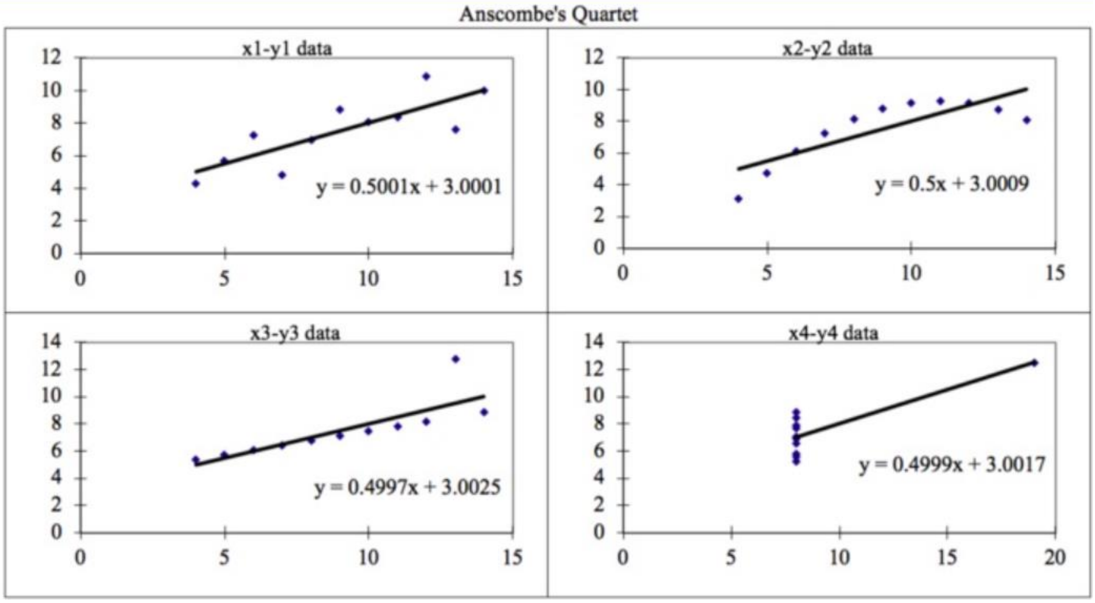
General Subjective Questions

1. Explain the linear regression algorithm in detail

   a) Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables.
   b) Linear regression assumes a linear or straight line relationship between the input variables (X) and the single output variable (y)
   c) In simple linear regression we can use statistics on the training data to estimate the coefficients required by the model to make predictions on new data.
   d) In linear regression tasks, there are two kinds of variables being examined: the dependent variable and the independent variable. The independent variable is the variable that stands by itself, not impacted by the other variable. As the independent variable is adjusted, the levels of the dependent variable will fluctuate. The dependent variable is the variable that is being studied, and it is what the regression model solves for/attempts to predict.

2. Explain the Anscombe's quartet in detail

   Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.



It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties.

3. What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

   Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x