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)
    a. cnt is more in fall season and less in spring season
    b. cnt is more June month. its 25 percentile is greater than all of the months 50 percentile other than July.
    c. cnt is more in clear weather and low in light snow weather

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

    pd.get_dummies() will return the dataframe of dummy variables. So for variable with 3 categories, it will create a dataframe of 3 columns, but 1 column of that will create collinearity because we can represent three variables with 2 columns.

    For ex:

    Let type is a column with values: a,b,c, so pd.get_dummies() will create dataframe like
    A | b | c

    1 0 0 —> this will represent a

    0 1 0 —> this will represent b

    0 0 1 —> this will represent c

    But we can represent it with one less variable, like

    B | c

    1 0 —> this will represent b

    0 1 —> this will represent c

    0 0 —> this will represent a

    Thus we can successfully remove first column and has no loss of information

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

    temp and atemp looks to have highest correlation with cnt.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    a. Errors in prediction should be normally distributed
    b. Errors in prediction should have zero mean
    c. r2_score should be comparable with training set
    d. No multicollinearity
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
    a. If we do RFE with n = 3, we will get 'yr', 'temp', 'windspeed' as the top 3 significant feature

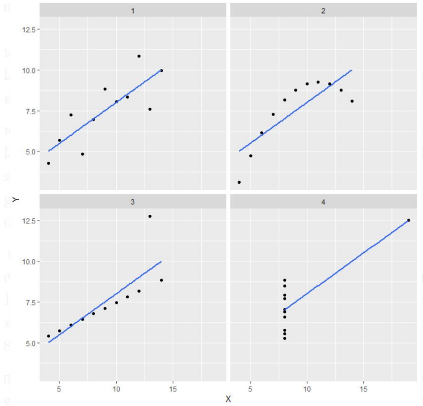
General Subjective Questions

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

   In Linear regression, we try to fit a linear line to explain the relation between target variable and independent variables. In reality, we can't have a line to perfectly fit all the independent variables with the target variable. So we try to fit a line which has minimum least square error, that is the value predicted by the line - the actual value.

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

   Anscombe created 4 set of 11 different datapoints which have same mean, median, standard deviation. But on representing the data on graphs, it looked very different



It proves that even even same statistical properties, the data can be very different and thus can't be compared with one another just on the basis of statistical properties

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

   It is used to define the relationship between two variables and how strong it is. It ranges from -1 to 1. If the value is close to 1, it signifies a strong positive relation, if it is close -1, it signifies strong negative relation. If it is close to 0, it represent week or no relation between variables.

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

   Scaling is re-representing the values in specific range.

   It is performed to create uniformity among all variables, so that we have uniformity among linear coefficients as well. It also help to map outlier in the acceptable limit.

   In Normalized scaling, we scale the values in 0 to 1. The minimum value is mapped to 0, while the max is mapped to 1.

   New_val = (val-min_val)/(max_val - min_val)

   In standardized scaling, we scale the values such that mean of scaled dat is 0 and standard deviation is 1. It is done by the formula below

   New_val = (val-mean)/sd

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

   If vif is infinite, then according to formulae vif = 1/(1-R2), R2 = 1, that is, that variable is perfectly explained by all other variables, so we can safely remove that variable from our model

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

   QQ plot is used to determine if two sets of data are from the same distribution or not.

   A y=x line is drawn on a plot and if the two datasets are from the same distribution, the plot is drawn on that reference line. If they are linearly related, the plots are plotted near the line.