## 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:**

   *Few categorical variables are positively correlated with cnt and few are negarively correlated. Although all have some significance below are the one which have high significance*

   1. *yr (Positive coefficient)*
   2. *mnth (Positive coefficient)*
   3. *holiday (Negative coefficient)*
   4. *light_snow (Negative coefficient)*
   5. *mist (Negative coefficient)*

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

   **Answer:**

   *Because n values of a categorical variable can easily be represented by (n-1) dummy values. When all the columns have 0 values, It will represent the first category which is not present in the dummy data*

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

   **Answer:**

   *temp or atemp*

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

   **By performing below steps:**

   1. *Applying the prediction model on the training set and performing residue analysis. Residues should be normally distributed*
   2. *Applying the prediction model on the test set and calculating r2_square. It should be near to the R-squared value*
   3. *Residue analysis of the test set. It should be normally distributed*

*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:**

*1) Month of the Year/Season and*
*2) Temperature*
*3) Weathersit*

## General Subjective Questions

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

**Answer:**

*Linear regression is a method of finding the best straight line fitting to the given data, i.e. finding the best linear relationship between the independent and dependent variables.*
*It can also be a defined as a machine learning approach that finds the best linear-fit relationship on any given data, between independent and dependent variables.*

**Main assumptions in a linear regression model:**

1. *There is a linear relationship between the dependent and independent variables.*
2. *Assumptions about the residuals:*
    a) *Normality assumption: It is assumed that the error terms, $\varepsilon^{(i)}$, are normally distributed.*
    b) *Zero mean assumption: It is assumed that the residuals have a mean value of zero.*
    c) *Constant variance assumption: It is assumed that the residual terms have the same variance, $\sigma^2$ This assumption is also known as the assumption of homogeneity or homoscedasticity.*
    d) *Independent error assumption: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.*
3. *Assumptions about the estimators:*
    a) *The independent variables are measured without error.*

b) *The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data.*

**Explanation:**

*If the residuals are not normally distributed, their randomness is lost, which implies that the model is not able to explain the relation in the data.*
*Also, the mean of the residuals should be zero.*
$Y^{(i)i} = \beta_0 + \beta_1 x^{(i)} + \varepsilon^{(i)}$
*This is the assumed linear model, where ε is the residual term.*
$E(Y) = E(\beta_0 + \beta_1 x^{(i)} + \varepsilon^{(i)})$
$= E(\beta_0 + \beta_1 x^{(i)} + \varepsilon^{(i)})$
*If the expectation(mean) of residuals, $E(\varepsilon^{(i)})$, is zero, the expectations of the target variable and the model become the same, which is one of the targets of the model.*
*The residuals (also known as error terms) should be independent. This means that there is no correlation between the residuals and the predicted values, or among the residuals themselves. If some correlation is present, it implies that there is some relation that the regression model is not able to identify.*

*If the independent variables are not linearly independent of each other, the uniqueness of the least squares solution (or normal equation solution) is lost.*

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

**Answer:**

*It's a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset consists of eleven (x,y) pairs* :

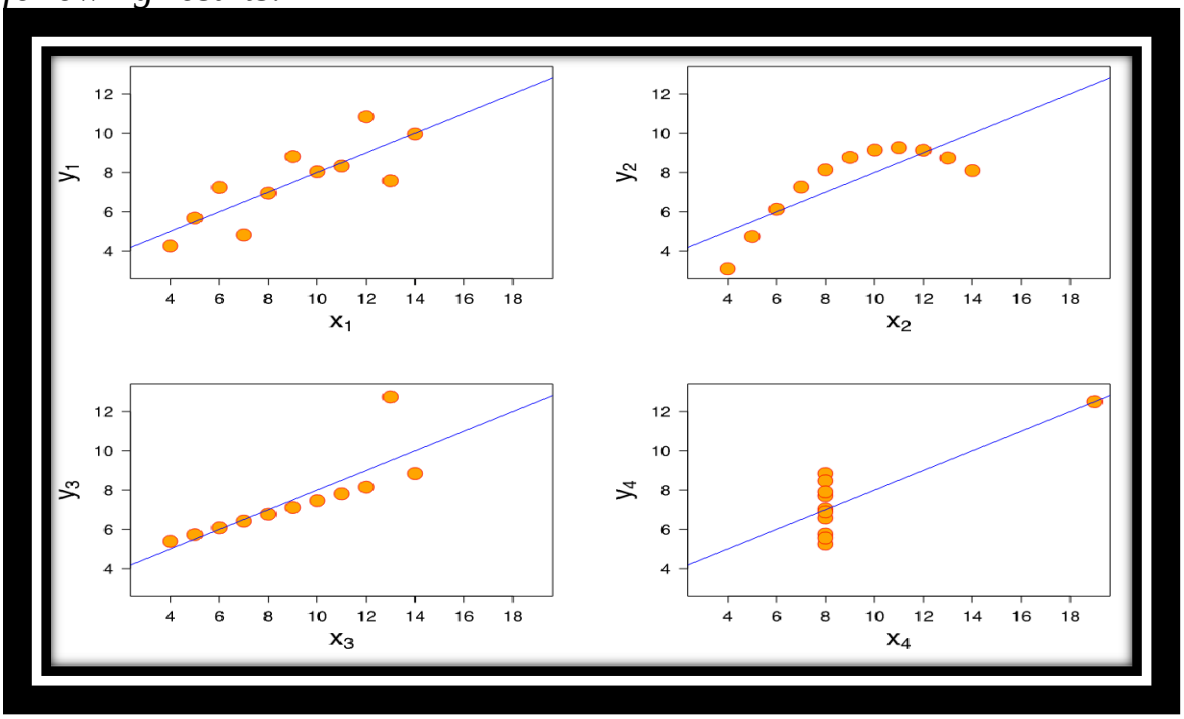
**Example:**

| I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|
| x | y | x | y | x | y | x | y |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |

*All the summary statistics looks identical:*

*The average x value is 9 for each dataset*
*The average y value is 7.50 for each dataset*
*The variance for x is 11 and the variance for y is 4.12*
*The correlation between x and y is 0.816 for each dataset*

*But when we plot these four data sets on an x/y coordinate plane, we get the following results:*



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

**Answer:**

***Pearson's r**, the **Pearson product-moment correlation coefficient** (**PPMCC**), or the **bivariate correlation**,[1] is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation*

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

**Answer:**

*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.*

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

**Answer:**

*VIF will be infinite in case of perfect correlation, because, on that case, R-squared will be 1. This will make VIF infinite as denominator will become 0 (1/(1-1))*

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

**Answer:**

*Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.*

*This 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.*

***Advanteges:***

*a) It can be used with sample sizes also*

*b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.*

*It is used to check following scenarios:*

*If two data sets —*

*I come from populations with a common distribution*

*ii. have common location and scale*

*iii. have similar distributional shapes*

*iv. have similar tail behavior*