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)**

- The month of September has the highest demand, and months of December and January have the lowest demand. The demand rises gradually at the beginning of the year, and peaks in September, before falling steeply towards the end of the year.
- Weekdays see a higher number of bookings when compared to weekends and holidays.
- Weather situation also has a high effect on the demand of bikes. Highest number of bikes are booked when the weather situation is good.
- Year 2019 has a higher overall demand for bikes.
- Fall season has the highest demand for bikes.

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

- It is important to drop the first dummy variable, during their creation because they would be redundant. It is easy to infer looking at other columns that, if in a particular set of dummy variable columns, if all the values in that row are zero, then then it would be the one which we dropped previously. It might also create the problem of multicollinearity between the predictor variables and might cause important features to be dropped because of this issue.

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

- The variables of 'temp' and 'atemp' have the highest correlation with the target variable. The value is approximately +1.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

(3 marks)

- By plotting the pairplots of numerical variables, I could find linear relationships between the predictor variables and the target variable.
- By plotting the residual terms, I found that the error terms are normally distributes around the value of zero.
- By checking the VIF values, I found that there is very little multicollinearity between the predictor variables.
- The ß1, ß2, ß3,… ßn values in our regression equation are not equal to zero, hence our model is significant.
- There is homoscedasticity in our model, as the variance is normal throughout the data.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

- Temperature is a factor which significantly contributes towards the demand of bikes. This has a positive relation with demand. This can be explained by its coefficient of (0.4364).
- Year has a positive relationship with demand. Its coefficient is (0.2494).
- The demand is low, when the weather situation is bad. There is a negative correlation between these two variables. The coefficient is $-(0.2445)$

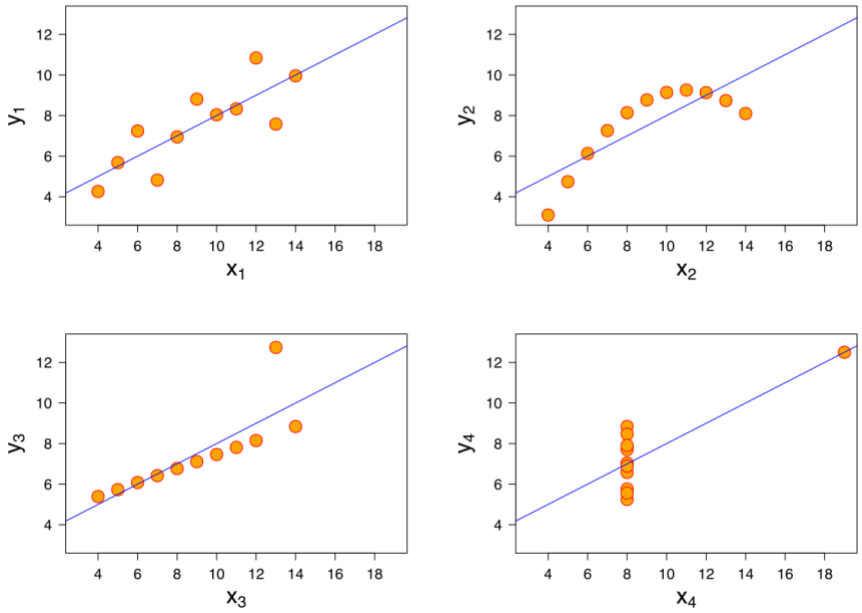
General Subjective Questions

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

- Linear Regression is a machine learning algorithm based on supervised learning.
- Linear Regression can be explained as a linear function, which attempts to predict one numeric variable depending on one or more numeric or categorical variables.
- The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.
- The following are the assumptions of Linear Regression:
    a) There is a linear relationship between the independent (predictor) variables and the dependent (target variable).
    b) The error terms of the predicted values are normally distributed.
    c) The multicollinearity between independent variables is low.
    d) There is homoscedasticity in the error terms, which means that variance throughout the distribution is similar.
- There are mainly two types of Linear Regression.
    a) **Simple Linear Regression**: It is represented by the function **$y = ßx + C$**, where ß is the slope of the regression line and C is its intercept.
    It attempts to predict the value of a numeric variable based on another single independent numeric variable.
    The null hypothesis (H_0) is that the slope of the line (ß) is zero. Which means there is no relationship between the predictor and target variable.
    If we reject he null hypothesis, then our prediction model would have significance.
    b) **Multiple Linear Regression:** It is represented by the function **$y = ß_1x_1 + ß_2x_2 + ß_3x_3 + … + ß_nx_n + C$,** where ß is the slope in each dimension of the plane, and n is the number of dependent variables.
    It attempts to predict the value of a target variable, based on multiple other predictor variables. The term R squared tells us how well our function is fitting in our given data.
    The null hypothesis (H_0) is that the slope of the lines in all the planes ($ß_1 = ß_2 = ß_3 = ß_n$) is zero. Which means there is no relationship between the predictor variables and the target variable.

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

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another. It tells us the importance of visualizing our data before fitting a model in the data.



- The first plot y1 appears to have a simple linear relationship between the variables.
- The second plot is not normally distributed, and the relationship between the variables is not linear.
- Our model is usually quite sensitive to outliers. The presence of one outlier which is quite far from the rest of the data can cause our model to be off the actual value by a fair bit.
- In the fourth plot, the presence of one extreme outlier gives us a correlation between the variables. In reality, there is no relationship between the variables.

3. **What is Pearson's R? (3 marks)**
   - Pearson's R also known as Pearson's Correlation Coefficient is the measure of linear correlation between two sets of data. Its value ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation and 1 indicates a perfect positive correlation between two variables.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   - Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. It reduces the problem of high standard deviation and skewness of data.
   - Often times, our data is made up of a variety of units, and we input dummy variables as well for categoric variables. It is important to scale our data for easy interpretation of coefficients, prior to building our model. T
   - There are two types of scaling:
     a) **Normalization**: In this type of scaling, the minimum value of the column is subtracted from each value and is divided by the difference between maximum and minimum. It is represented by: (X – min)/(max – min). All the values in the column are now in between 0 and 1.
     b) **Standardization**: In this type of scaling, we ensure the standard deviation of the data is 1. It is represented by: (X – mean)/ Standard Deviation. It is not usually affected by outliers.

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

   - VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.(VIF) =1/(1-R_1^2). If there is perfect correlation, then VIF = infinity. Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1.So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity"

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

- A Q-Q plot stands for Quantile Quantile plot. The quantiles of two sets of data are plotted against each other. It explains the distribution of a dataset, whether it is normally distributed or not. They can also be comparable with other datasets.
- 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.