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)**

Ans:

- Season – Lesser demand in Spring and Winter, High demand in Summer and Fall
- Yr. – There is significant growth from 2018 to 2019
- Month - cnt keep increasing from Jan – Jul then falls until December
- Holiday – cnt is less on holiday compared to non-holidays
- Weekday – cnt doesn't vary that much by weekday basis
- Workingday – cnt doesn't vary a lot between working and non-working day
- Weather – cnt is lowest on a snowy day and highest on a clear day

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

Ans:

- For a categorical variable with n levels, if we do not use drop_first= True we will get n dummy variables, which leads to collinearity between the variables. To avoid that we need to remove one level

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

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

- Used residuals vs fitted charts to check if it has no pattern
- Used VIF to check that there are no multicollinear variables
- Plotted distribution of errors terms to check if they are normally distributed and centred around 0

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

Ans: Temperature, Season- Winter, Weather- LightSnow

General Subjective Questions

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

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear regression is based on the popular equation "y = mx + c". It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x).

In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error. In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression. 1. Simple Linear Regression: SLR is used when the dependent variable is predicted using only one independent variable. 2. Multiple Linear Regression: MLR is used when the dependent variable is predicted using multiple independent variables.

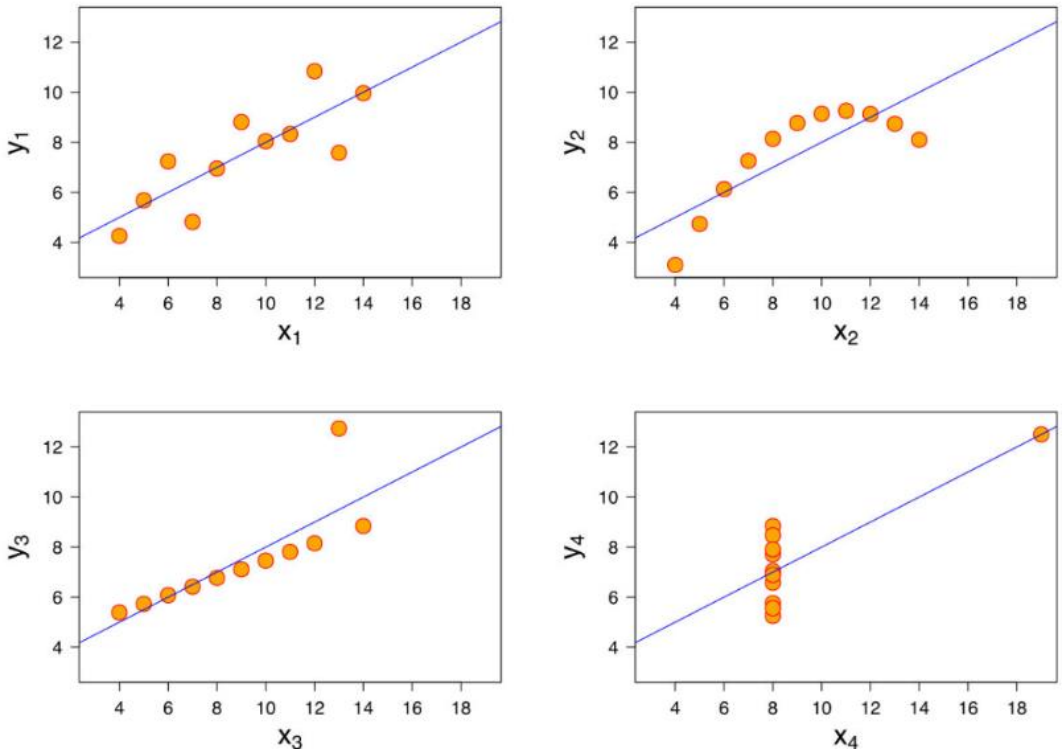
The equation for MLR will be:

Y = b0 + b1*X1 + b2*X2 + …….

b1 = coefficient for X1 variable b2 = coefficient for X2 variable β3 = coefficient for X3 variable and so on… b0 is the intercept (constant term).

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

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties.



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

Pearson's r is a numerical summary of the strength of the linear association between the variables. Its value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data? r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a negative slope r = 0 means there is no linear association

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 or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

Normalization brings all data between 0 and 1 including outliers. Standardization brings all the data into a distribution with mean (mu) and standard deviation (sigma). Thereby outliers will not be restricted in terms of range

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

When linear combination of some the variables can perfectly explain another variable then we get vif = inf. As such a model will have R^2 = 1 and there by VIF = 1/(1-R^2) = 1/ (1-1) = 1/0 =inf

The resolution is to identify these multi collinear variables and remove one of them

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 is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.