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. After plotting the box plot for categorical variable

- Season: 3: fall has highest demand for rental bikes.
- The demand for bike for next year has grown.
- The demand of bike is continuously growing each month till June. September month has highest demand. After September, demand is decreasing.
- When there is a holiday, demand has decreased.
- When there is a working, the demand has increased.
- Weekday is not giving clear picture about demand.
- The clear weathershit has highest demand.

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

Ans: We use **drop_first=True** during dummy variable creation so that we reduce the multicollinearity between dummy variables. To achieve this, we reduce the extra column that is created during the dummy variable creation.

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

Ans: After looking at we can conclude that the variable "**temp"** has the highest correlation with the target variable "**cnt".**

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

Ans: I can validate the assumptions of Linear Regression after building the model on the training set based on below parameters.

- Error terms are normally distributed with mean 0.
- Error Terms do not follow any pattern.
- Multicollinearity check using VIF(s).
- Ensured the overfitting by looking the R2 value and Adjusted R2.

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: The top 3 features contributing significantly towards explaining the demand of the shared bikes are "**holiday**", "**temp**" and season "**hum**".

General Subjective Questions:

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

Ans: The linear regression algorithm is a statistical approach to estimate the relationship between two variables. In this approach we focus two types of variables one is the dependent variable, and the other is independent variable (also called as predictors).

In the linear regression algorithm, our goal is to predict the change in the value of dependent variable with respect to the dependent variable. And to achieve this we can change the one variable at a time.

This algorithm is used to forecast or predict the value of a dependent variable.
We can use two approaches to predict the value of dependent variable.

Interpolation: In this we predict the value of a dependent variable on the independent values that lie within the range of already present data.

Extrapolation: In this we predict the dependent variable on the independent values that lie outside the range of the already present data.

The linear regression shows the corelation between variables.
It is a kind of parametric regression in which data follows fixed number of parameters.

The linear regression model fits a straight line into the summarized data to establish the relationship between two variables.

A simple linear regression aims to find the best relationship between X (independent variable) and Y (dependent variable).

And it is represented as $Y = \beta_o + \beta_1 X$
Here,
β_o = y-intercept of the line
β_1 = Slope of the line
X = Independent variable from dataset
Y = Dependent variable from dataset

We need to consider the below assumption while performing simple linear regression.

1. There must be a linear relationship between X and Y.
2. The error terms should be normally distributed with mean zero (not X, Y).
3. The error terms are independent of each other.
4. The error terms have constant variance (homoscedasticity) i.e., the variance should not change as the error values change.

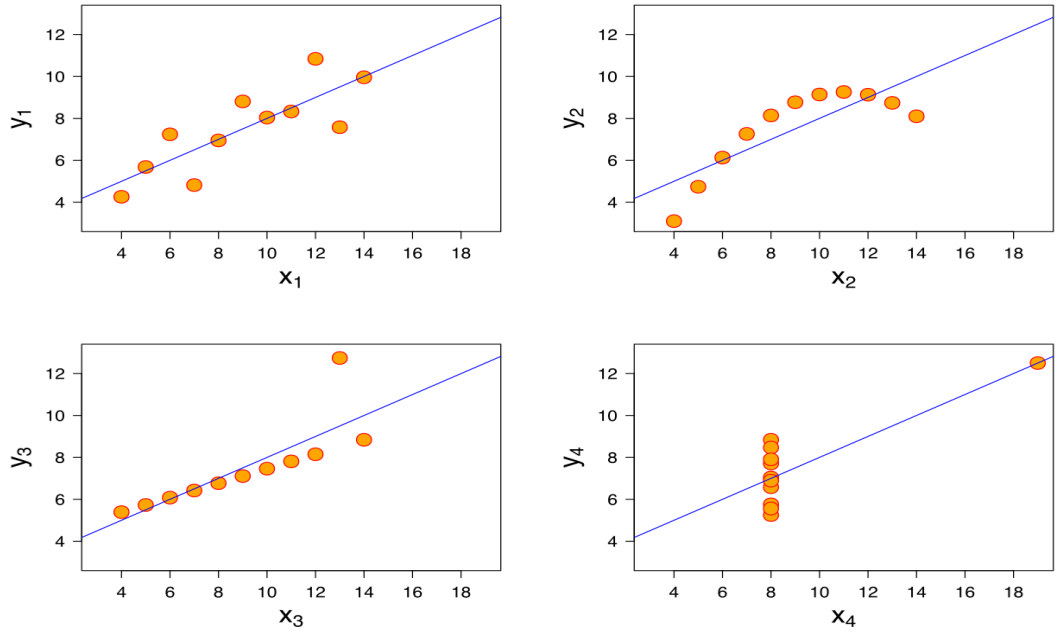
2. Explain the Anscombe's quartet in detail. (3 marks)

Ans. Anscombe's quartet consist of four datasets that have almost similar statistical properties but when plotted using graph they appear differently. Each dataset consists of eleven (X,Y) points.

We use the Anscombe's quartet to demonstrate the importance effect of outliers and graphical data before analysis on statistical properties.

The four datasets in Anscombe's quartet are as follows:

1. **Dataset 1:** This follows linear regression model with some variance.
2. **Dataset 2:** This fits a curve but does not follow linear regression model.
3. **Dataset 3:** This shows a linear relationship between *x* and *y*, except for one large outlier.
4. **Dataset 4:** This shows X remains constant, except for one outlier.



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

Ans. Pearson correlation coefficient is a measure of strength of linear relationship between two variables. It is represented by **r.**

The Pearson's R always lies between -1 & 1 where:

1. r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
2. r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
3. r = 0 means there is no linear association
4. r > 0 < 5 means there is a week association.
5. r > 5 < 8 means there is a moderate association.
6. r > 8 means there is a strong association.

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

Ans. Scaling is used to bring all the independent variables that are on a different scale onto a same scale in a regression to ease the representation of variable.

The scaling only affects the coefficients and not the other parameters like t-statistic, F statistic, p-values, R-square,

1. **Standardizing**: The variables are scaled in such a way that their mean is 0 & standard deviation is 1.

2. **MinMax Scaling**: The variables are scaled in such a way that all the values lie between 0 & 1.

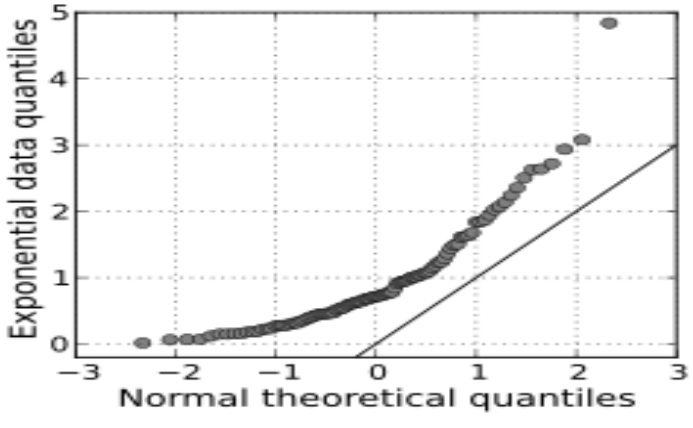
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans. If the value of VIF is infinity. This tells us that there is a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity

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

Ans. Q-Q Plots (Quantile-Quantile plots) are plots of quantiles where we plot the quantiles of first dataset against the quantiles of the second dataset.

- If the points of the quantiles lie approximately on the y = x at an angle of 45 on Q-Q plot, then the distribution is similar
- If the points of the quantiles lie approximately on the y = x at an angle of 45 on Q-Q plot, then the distribution is different.



A Q-Q plot is used to find out if the two datasets are from same population distribution, and have similar properties such as location, scale, distribution shape.

We use the concept of Q-Q plot in linear regression when we receive the training and test dataset separately and so with the help of Q-Q plot we can predict that both the data sets are from populations with same distributions.