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)**

From the scatter plots the categorical variables do not seem to have too much influence on the target variable except year.

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

We use drop_first=True to reduce one column for the purpose of efficiency. If we can reduce one column then the processing expense gets reduced.

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

Atemp has the highest correlation of 0.5973

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

Did an error plot

Checked correlation heat map

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

Atemp, Light_snow_rain and year

General Subjective Questions

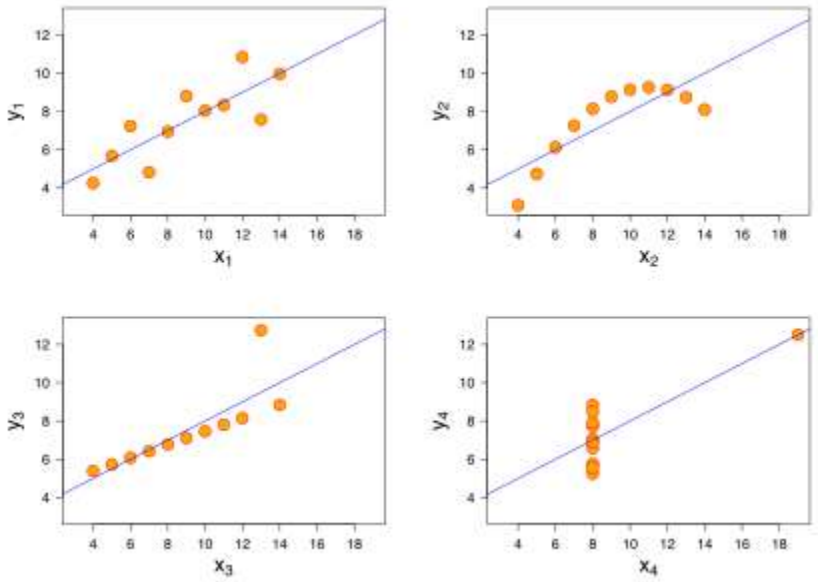
1. Explain the linear regression algorithm in detail. (4 marks)

1. Data preparation
    a. Drop unnecessary columns
    b. Graph all the features and check visually
    c. Create dummy variables

2. Split into train and test
3. Rescaling the Features
4. Training the model using RFE
    a. Chosing number of features to start with
    b. Dropping features based on what RFE says is not useful (False flag)
    c. Iterate with p-value and VIF to drop further features
5. Plotting correlation between remaining terms and checking for distribution of error terms
6. Using model to make predictions

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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. The graphs can be seen below and it's quite obvious that the relation between the independent and target variable is different in all 4 cases.



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

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

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

Scaling is the process of converting all features to have uniform effect on the target variable. This is done so that some features don't have outsize effect because they are distributed in a higher numeric range.

Normalization Scaling is used to transform all features to be same scale.

$$X\_new = (X - X\_min)/(X\_max - X\_min)$$

This scales the range to [0, 1] or sometimes [-1, 1]. Normalization is useful when there are no outliers as it cannot cope up with them.

Standardization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

$$X\_new = (X - mean)/Std$$

In Standardization, outliers have lesser effect.

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

If there is perfect correlation, then VIF = infinity. This shows 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)

Statistics, Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc. You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot.