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 : The categorical variables in the data set are

“season","yr","mnth","holiday","weekday","workingday","weathersit".

The categorical variable season , weatherfit and month have siginificant impact on the dependent variable.

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

Ans : drop_first= helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.For example , suppose we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

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

Ans : The variable temp has the highest correlation with temp variable .

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

Ans : The assumptions of Linear regression are

i. The core premise of multiple linear regression is the existence of a linear relationship between the dependent (outcome) variable and the independent variables. This linearity can be visually inspected using scatterplots, which should reveal a straight-line relationship rather than a curvilinear one
ii. The analysis assumes that the residuals (the differences between observed and predicted values) are normally distributed with 0 mean . This s validated by having distributed plot of residual for both test and train data.
iii. It is essential that the independent variables are not too highly correlated with each other, a condition known as multicollinearity. This is verified by the VIF which is observed to be < 5 for all variables.

iv. Homoscedasticity: The variance of error terms (residuals) should be consistent across all levels of the independent variables. A scatterplot of residuals versus predicted values confirms this.

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 are Temperature, Season,year

General Subjective Questions

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

Ans : Linear Regression is a machine learning algorithm where the output to be predicted is a continuous variable .eg rent of a house , score of a student etc . It is a supervised learning method where the labels are predefined. It can be classified as simple linear regression where the output is dependent on only one predictor variable or multiple linear regression where the output is dependent on multiple predictor variables. Simple linear regression explains the relationship between the dependent and the independent variable using a straight line with intercept and coefficient. The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable. The strength of the linear regression model can be assessed using 2 metrics:
1. R^2 or Coefficient of Determination
2. Residual Standard Error (RSE)
Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. The formulation for multiple linear regression is also similar to simple linear linear regression with the small change that instead of having beta for just one variable, you will now have betas for all the variables used.

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

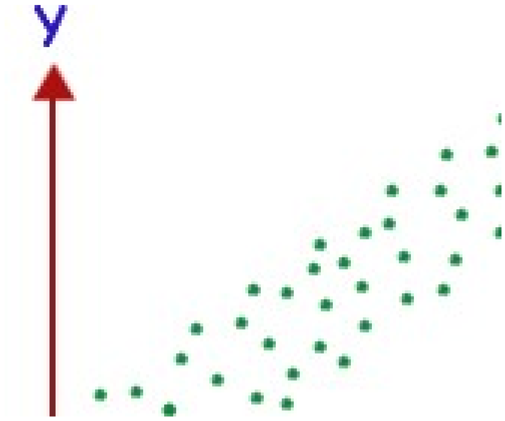
Ans : Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties. Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph. The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

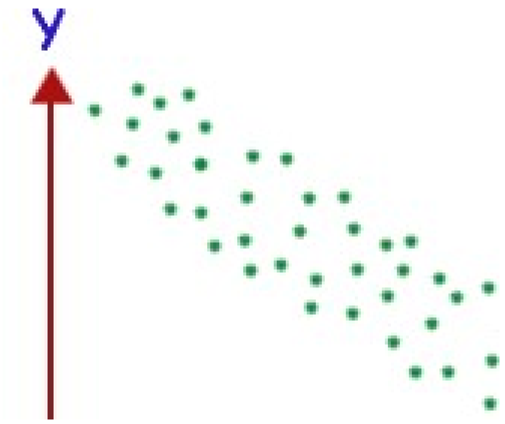
3. What is Pearson's R? (3 marks)

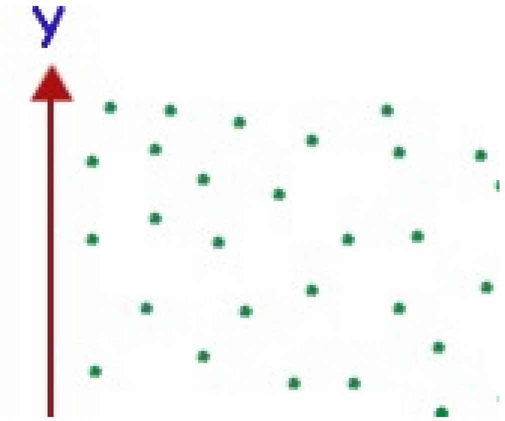
Ans : The name correlation suggests the relationship between two variables as their Co-relation. The correlation coefficient is the measurement of correlation. To see how the two sets of data are connected, we make use of this formula. The linear dependency between the data set is done by the Pearson Correlation coefficient. It is also known as the Pearson product-moment correlation coefficient. The value of the Pearson correlation coefficient product is between -1 to +1. When the correlation coefficient comes down to zero, then the data is said to be not related. While, if we are getting the value of +1, then the data are positively correlated and -1 has a negative correlation.

The graphical representation of positive, negative and no correlation is shown below:



Positive Correlation





The Pearson correlation coefficient is denoted by the letter “r”. The formula for Pearson correlation coefficient r is given by:

$r=n(\sum xy)-(\sum x)(\sum y)[n\sum x2-(\sum x)2][n\sum y2-(\sum y)2]$

Where,
r = Pearson correlation coefficient
x = Values in the first set of data
y = Values in the second set of data
n = Total number of values.

----

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

Ans : Scaling is the process of transforming the values of the features of a dataset till they are within a specific range, e.g. 0 to 1 or -1 to 1. This is to ensure that no single feature dominates the distance calculations in an algorithm, and can help to improve the performance of the algorithm.

In Normalized scaling , values are shifted and rescaled so that they end up ranging from 0 to 1. We do this by subtracting the min value and dividing by the max minus the min.

In standardization, first it subtracts the mean value (so standardized values always have a zero mean), and then it divides by the standard deviation so that the resulting distribution has unit variance. Unlike min-max scaling, standardization does not bound values to a specific range, which may be a problem for some algorithms . However, standardization is much less affected by outliers

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

Ans : If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   Ans : 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 a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal or exponential. 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 a Pareto distribution. You can tell the type of distribution using the power of the Q-Q plot just by looking at it