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 – For the categorical variable 'season' the fall season has maximum number of cnt values after that comes summer, winter and spring in the order written.

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

Ans - It is required to drop_first column in dummy variable creation because its requirement can be fulfilled by other columns that are created. Hence it only results in increment of a column while model building.

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

Ans- atemp has the highest correlation with cnt of 0.65

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

Ans – We checked that the error terms are normally distributed by plotting the distplot.

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- year, temp and winter(season) are the top three features.

General Subjective Questions

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

Ans- Linear regression is a supervised machine learning algorithm that is used to predict continuous values. It is one of the most popular and widely used machine learning algorithms. Linear regression works by finding the best-fit line through a set of data points. The line is then used to predict the value of the dependent variable for new data points.

The linear regression algorithm can be summarized in the following steps:

1. Choose the dependent variable.: The dependent variable is the variable that you are trying to predict.
2. Choose the independent variables.: The independent variables are the variables that you are using to predict the dependent variable.
3. Fit the model.: The model is the line that you will use to predict the dependent variable.
4. Evaluate the model.: The model is evaluated by how well it predicts the dependent variable for new data points.

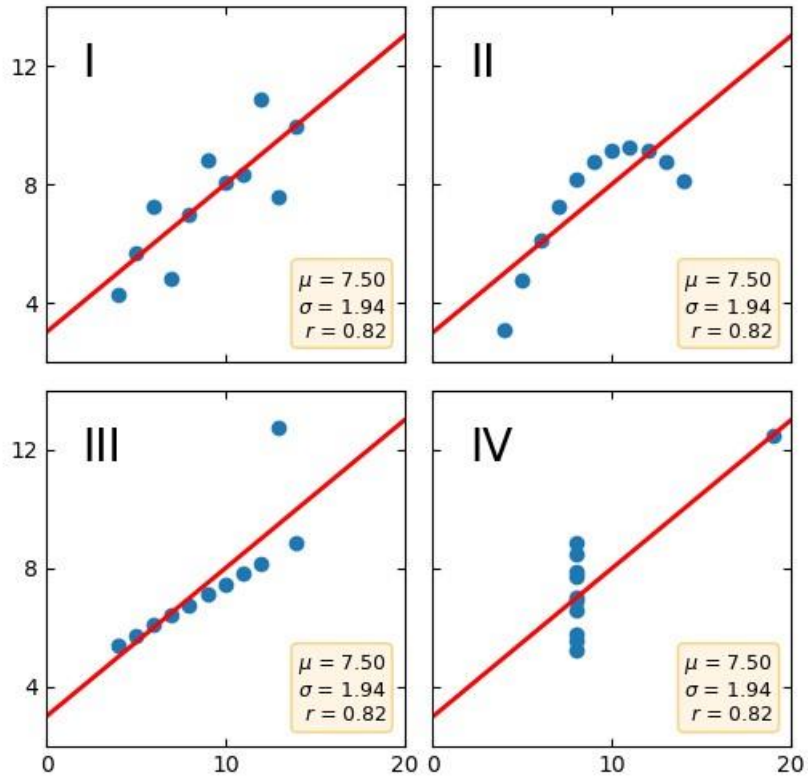
Linear regression can be used for a variety of tasks, such as:

- Predicting sales
- Predicting customer churn
- Predicting the price of a house
- Predicting the risk of heart disease

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.



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

Ans – The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Another way to think of the Pearson correlation coefficient (r) is as a measure of how close the observations are to a line of best fit. The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive.1

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

Ans - It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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 are commonly used to compare a data set to a theoretical model. This can provide an assessment of goodness of fit that is graphical, rather than reducing to a numerical summary statistic. Q–Q plots are also used to compare two theoretical distributions to each other.