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?**
   Answer:
   - Season-wise **Fall season** have the highest demand.
   - In the Month-wise category **Sept month** seems to be the highest out of all the months.
   - On the day when the **weather is clear** seems to be the most favorable day for bike sharing.
   - Clearly on the **holiday** the demand is decreased for bike rentals it may be because people may want to spend time at home and enjoy with family.
   - Year-wise trend shows a clear growth in demand from the **year- 2018 to year- 2019**.
   - Most of the bookings has been done during the month of May, June, July, Aug, Sep and Oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.

2. **Why is it important to use drop_first = True during dummy variable creation?**
   Answer:
   drop_first = True helps in reducing extra columns during dummy variable creation. Thus, reducing the collinearity between variables.

   Syntax:
   `pandas.get_dummies(data, prefix=None, prefix_sep='_', dummy_na=False, columns=None, sparse=False, drop_first=False, dtype=None)`

   where, drop_first parameter takes Boolean value, by default its false.
   Whether to get k-1 dummies out of k categorical levels by removing the first level.
   For Ex: If we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Answer:
   By looking at the pair-plot between the numerical variables we can clearly see that ‘temp’ and ‘atemp’ variables have highest correlation with target variable ‘cnt’.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Answer:
   I've validated the assumptions of linear regression by verifying these following assumptions on the final model:
   - Linear relationship between X and y.
   - Error terms are normally distributed with mean zero.
   - Error terms are independent of each other.
   - Error terms have constant variance(homoscedasticity) and there is no visible pattern between residuals.
   - There should be insignificant multicollinearity among variables.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Answer:
   Top 3 features contributing significantly are:
   - yr
   - temp
   - light_snowrain

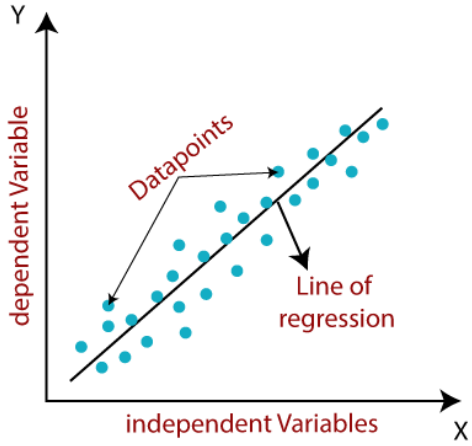
General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   Answer:
   Linear Regression algorithm is a machine learning algorithm based on supervised learning method. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous variables.

   Mathematically the relationship can be represented with the help of following equation:

   $y = mX + c$

   Where,
   y = Dependent variable (variable to be predicted)
   X = Independent variable (variable using to predict)
   m = Slope of the regression line
   c = Intercept of the regression line



Linear regression algorithm is of the following two types:

- Simple Linear Regression: In this there is only one independent variable.
- Multi Linear Regression: In this there is more than one independent variable.

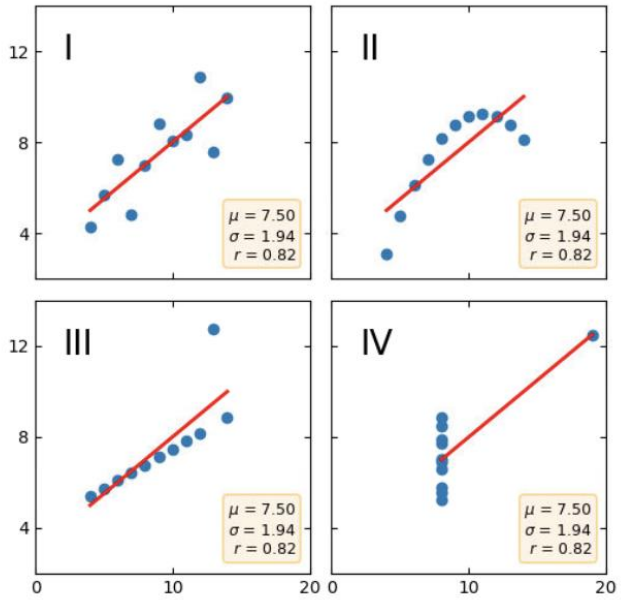
There are followings assumptions of linear regression algorithm:

- Linear relationship between X and y.
- Error terms are normally distributed with mean zero.
- Error terms are independent of each other.
- Error terms have constant variance(homoscedasticity) and there is no visible pattern between residuals.
- There should be insignificant multicollinearity among variables.

2. **Explain the Anscombe's quartet in detail.**
   Answer:
   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.
   Those 4 sets of 11 data points are given below.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The statistical properties and visual representation of the above data points is given below:



Application of the Anscombe's quartet: The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R?

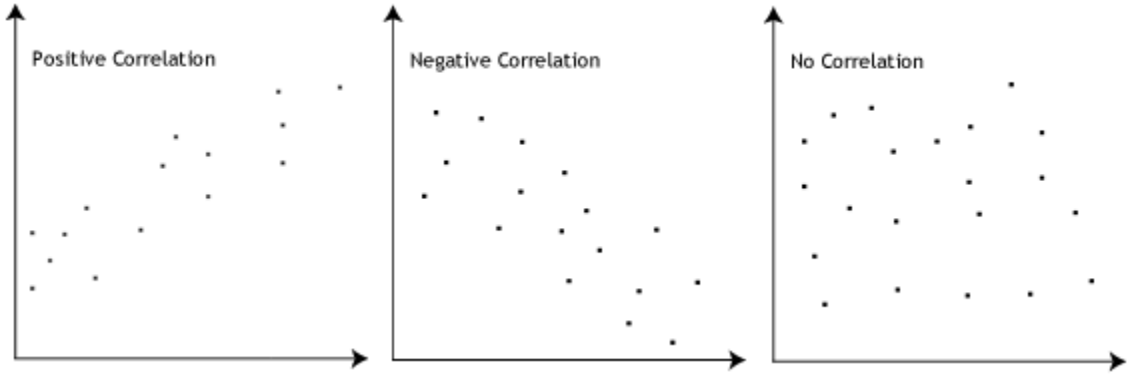
Answer:

Pearson's correlation coefficient also known as Pearson's R is the covariance of the two variables divided by the product of their standard deviations. it is essentially a normalized measurement of the covariance, such that the result always has a value

between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables.

Where,

0		= No Association
1	& -1	= Linear relation (positive or negative slope)
0 to -0.3	& 0 to 0.3	= Small Association
-0.3 to -0.5	& 0.3 to 0.5	= Medium association
-0.5 to -1.0	& 0.5 to 1.0	= Large Association



Pearson's R is given by:

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}$$

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Answer:

Scaling is a technique to standardize the independent features present in the data in a fixed range.

It is performed during the data pre-processing to handle highly varying magnitudes or values or units. 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, regardless of the unit of the values.

Difference between Normalized Scaling and Standardized Scaling:

- In Normalized scaling Minimum and Maximum values of the feature is used for the scaling while in Standardized scaling Mean and Standard deviation of the feature is used for the purpose of scaling.
- Normalized scaling is used when features are of different scale and Standardized scaling is used when we want to ensure feature is scaled with zero mean and unit standard deviation.
- Normalized scaling is affected by outliers while Standardized scaling is not.
- Normalization rescales the value in range of [0, 1] and standardization rescales the feature to have zero mean and unit standard deviation.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   Answer:
   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. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   Answer:
   The Q-Q (quantile-quantile) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

   Use of Q-Q plot:
   A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

   Importance of Q-Q plot:
   A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.