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)

Answer: From the MLR the demand for the bike depends on the below

- weather conditions were found to have strong impact on bike demand
- year (2019): the demand for bikes has increased from year 2018 to 2019
- during spring the demand for bikes has decreased
- during winter the demand for bikes has increased
- In the month of September, the demand for bikes has increased
- While in the months of November, December, January, & July the demand for bikes has decreased
- Light-snow & misty weathers were found to decrease the demand of the bikes

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

Answer:

It is very important.

drop_first=True drops the first column during dummy variable creation

If we do not use drop_first = True, then n dummy variables for n categories will be created, and these predictors (n dummy variables) are themselves are highly correlated, will affect the model by leading in to a Dummy Variable Trap.

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

Answer:

The 'temp' variable was found to have a highest linear correlation with the target variable.

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

Answer:

After building the model the four assumptions of Linear Regression Model were tested as below

1. **Error terms are normally distributed with mean zero**
   - The residuals were found to have a normal distribution with mean close to zero
2. **There is a linear relationship between X and Y:**
   - By plotting Y-train & y-train_pred & by plotting Y-test & y-test_pred the linearity was confirmed
3. **Error Terms Being Independent**
   - By plotting the Residual Vs. Predicted Values no trend was observed and the independence of error terms were confirmed
4. **Error terms have constant variance (homoscedasticity)**
   - By plotting Predicted Points Vs. Actual Points, the constant variance from the regression line was confirmed

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

   Answer: Based on the final model below 3 features contributing significantly on prediction of bike demand

   - **Temp,** with a strong positive influence
   - **Year**, 2019 with a strong positive influence
   - **light_snow weather,** with a strong negativee influence

General Subjective Questions

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

Answer:

Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

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

$$Y = mX + c$$

Here, Y is the dependent variable we are trying to predict.

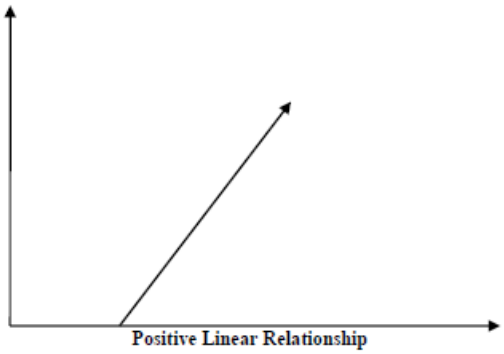
X is the independent variable we are using to make predictions.

m is the slope of the regression line which represents the effect X has on Y

c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Furthermore, the linear relationship can be positive or negative in nature as explained below–

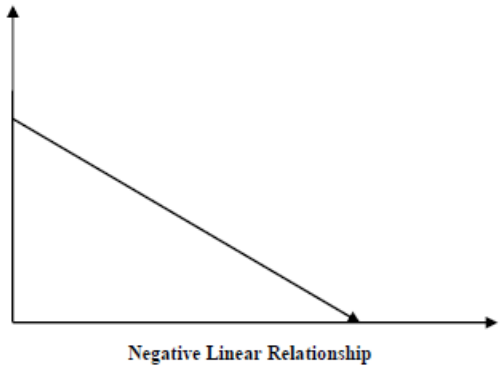
- **Positive Linear Relationship:**
  - A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph –



Positive Linear Relationship

Negative Linear relationship:

- A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph –



Negative Linear Relationship

Linear regression is of the following two types –

- Simple Linear Regression
- Multiple Linear Regression

Assumptions:

1. There is a linear relationship between X and Y:

- X and Y should display some sort of a linear relationship; otherwise, there is no use of fitting a linear model between them.

2. Error terms are normally distributed with mean zero(not X, Y):
   - The assumption of normality is made, as it has been observed that the error terms generally follow a normal distribution with mean equal to zero in most cases
3. Error terms are independent of each other:
   - The error terms should not be dependent on one another (like in a time-series data wherein the next value is dependent on the previous one)
4. Error terms have constant variance (homoscedasticity):
   - The variance should not increase (or decrease) as the error values change
   - Also, the variance should not follow any pattern as the error terms change

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

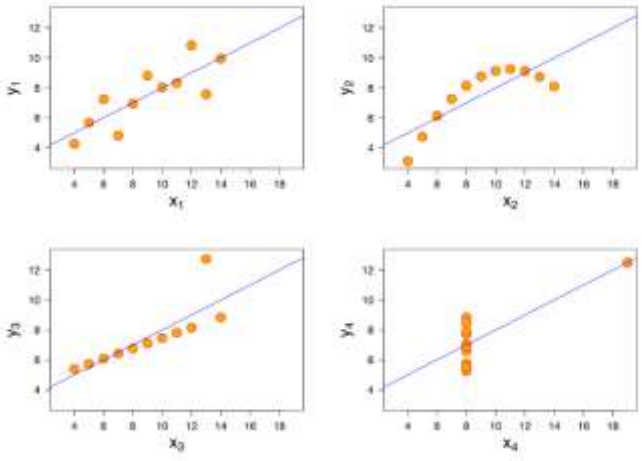
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,1	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,1	4	5,39	19	12,5
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.

• In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

• Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

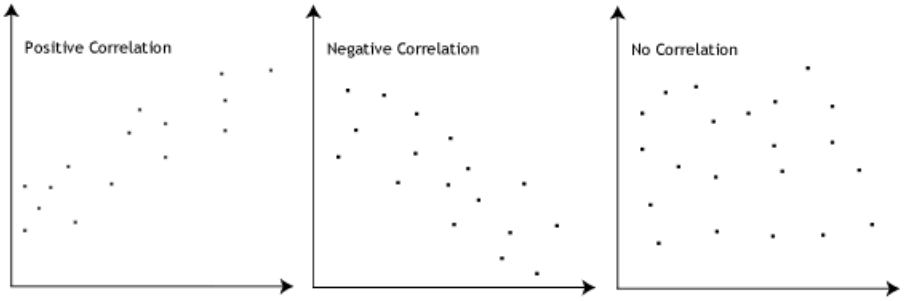
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

Answer:

Pearson's R is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Answer: Feature 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.

Example: If an algorithm is not using feature scaling method, then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Answer:

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.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

The acceptable VIF <5.

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

Answer:

A Q-Q plot, or quantile-quantile plot, is a scatter plot that compares two quantiles. A quantile is the fraction of values that fall below a given value. For example, the median is a quantile where 50% of the data fall below that point.

Q-Q plots are used to determine if two samples of data came from the same population. They can also help determine if a dataset follows a particular type of probability distribution, such as normal, uniform, or exponential.

To create a Q-Q plot, you order each data set in increasing order, then pair off and plot the corresponding values. If the points fall on a straight line, then the distributions have a similar shape. If the points curve off in the extremities, then the data has more extreme values than would be expected from a normal distribution.

Q-Q plots can provide more insight into the nature of differences between samples than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.