

### **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?

Ans: From the analysis of the categorical variables from the dataset it could be inferred the bike rental rates are likely to be higher in summer and the fall season, are more prominent in the months of September and October, more so in the days of Sat, Wed and Thurs and in the year of 2019. Additionally we could discern that bike rental are higher on holidays.

2. Why is it important to use drop\_first=True during dummy variable creation?

Ans: drop\_first=True helps in reducing the extra column created during the dummy variable creation and hence avoid redundancy of any kind.

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

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

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

Ans: Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable.

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

Ans: The top 3 features contributing significantly towards the demand of the shared bikes are the temperature, the year and the holiday variables.

## **General Subjective Questions**

1. Explain the linear regression algorithm in detail.

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(\text{slope}) = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

$$a(\text{intercept}) = \frac{n \sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

As mentioned above, Linear regression estimates the relationship between a dependent variable and an independent variable. Let's understand this with an easy example:

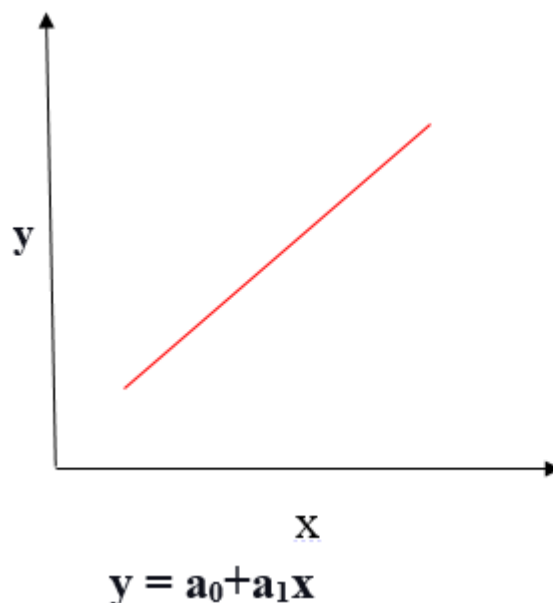
Let's say we want to estimate the salary of an employee based on year of experience. You have the recent company data, which indicates that the relationship between experience and salary. Here year of experience is an

independent variable, and the salary of an employee is a dependent variable, as the salary of an employee is dependent on the experience of an employee. Using this insight, we can predict the future salary of the employee based on current & past information.

*A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.*

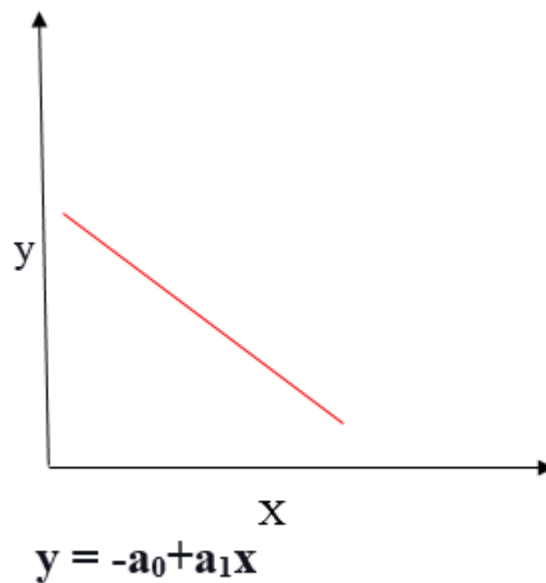
#### **Positive Linear Relationship**

If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.



#### **Negative Linear Relationship**

If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship.



The goal of the linear regression algorithm is to get the best values for  $a_0$  and  $a_1$  to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.

## 2. Explain the Anscombe's quartet in detail.

***Anscombe's Quartet*** is the modal example to demonstrate the importance of data visualization which was developed by the statistician ***Francis Anscombe*** in 1973 to signify both the importance of plotting data before analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

x1	y1	x2	y2	x3	y3	x4	y4
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

*Four Data-sets*

Apply the statistical formula on the above data-set,

Average Value of x = 9

Average Value of y = 7.50

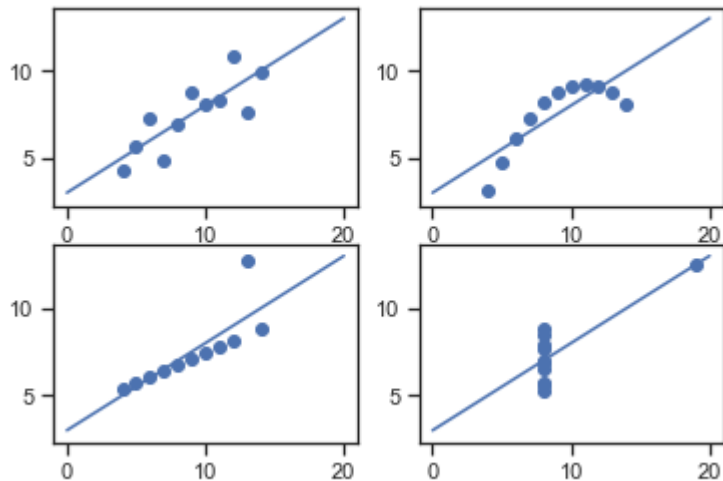
Variance of x = 11

Variance of y =4.12

Correlation Coefficient = 0.816

Linear Regression Equation :  $y = 0.5x + 3$

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behavior.



*Graphical Representation of Anscombe's Quartet*

- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
- Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.
- Data-set IV — looks like the value of x remains constant, except for one outlier as well.

### 3. What is Pearson's R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's  $r$ , the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between  $-1$  and  $1$ .

The Pearson's correlation coefficient varies between  $-1$  and  $+1$  where:

- $r = 1$  means the data is perfectly linear with a positive slope ( i.e., both variables tend to change in the same direction)
- $r = -1$  means the data is perfectly linear with a negative slope ( i.e., both variables tend to change in different directions)
- $r = 0$  means there is no linear association
- $r > 0 < 5$  means there is a weak association
- $r > 5 < 8$  means there is a moderate association
- $r > 8$  means there is a strong association

#### Pearson r Formula

Here,

- =correlation coefficient
- =values of the x-variable in a sample
- =mean of the values of the x-variable
- =values of the y-variable in a sample
- =mean of the values of the y-variable

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

- 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 typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S.NO	Normalisation	Standardisation
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.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.



7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is often called as Scaling Normalization	It is often called as Z-Score Normalization.

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

If there is perfect correlation, then  $VIF = \infty$ . This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get  $R^2 = 1$ , which leads to  $1/(1-R^2)$  infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

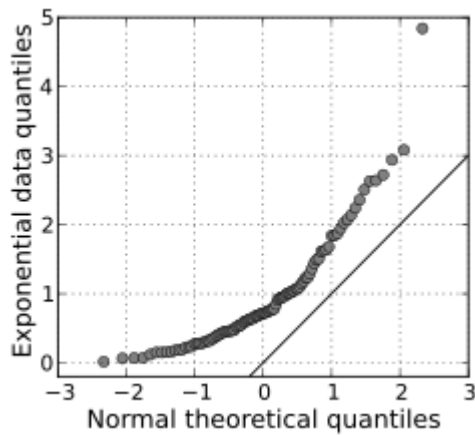
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. 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.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line  $y = x$ . If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line  $y = x$ . Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

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.