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 –

Observations from the analysis done on the categorical variables of the dataset it can be inferred as follows -

- People are more likely to take bikes on rent during Summer and the Fall of the Season.
- September & October months shows high number of renting bikes.
- Saturday, Wednesday & Thursday showing high number of bikes renting.
- Clear weather shows high number for taking bikes on rent.
- More Renting of bikes happened in 2019.
- No change in bike renting whether it is working day or not.
- Bikes rental rates are higher on holiday

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

Answer –

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?

Answer –

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?

Answer –

Validating assumptions of the linear regression by checking **the VIF, Multicolinearity Check, Error Distribution of the Residuals, Homoscedasticity and the Linear Relationship** between the Dependent Variable and the Feature Variable.

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

Answer –

The top 3 features contributing significantly towards the demand of the shared bikes are the **temp, Winter and the Sep** variables.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer –

Linear Regression is an ML algorithm used for supervised learning. It helps in predicting a dependent

variable(target) based on the given independent variable(s). The regression technique tends to establish a linear relationship between a dependent variable and the other independent variables.

There are two types of linear regression- **Simple Linear Regression** and **Multiple Linear Regression**.

Simple Linear Regression is used when a single independent variable is used to predict the value of the target variable.

Multiple Linear Regression is when multiple independent variables are used to predict the numerical value of the target variable.

A linear line showing the relationship between the dependent and independent variables is called a regression line.

A positive linear relationship is when the dependent variable on the Y-axis along with the independent variable in the X-axis.

However, if dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship.

Mathematically, we can write a linear regression equation as:

y = a + bx

Where a and b given by the formulas:

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

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

- 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

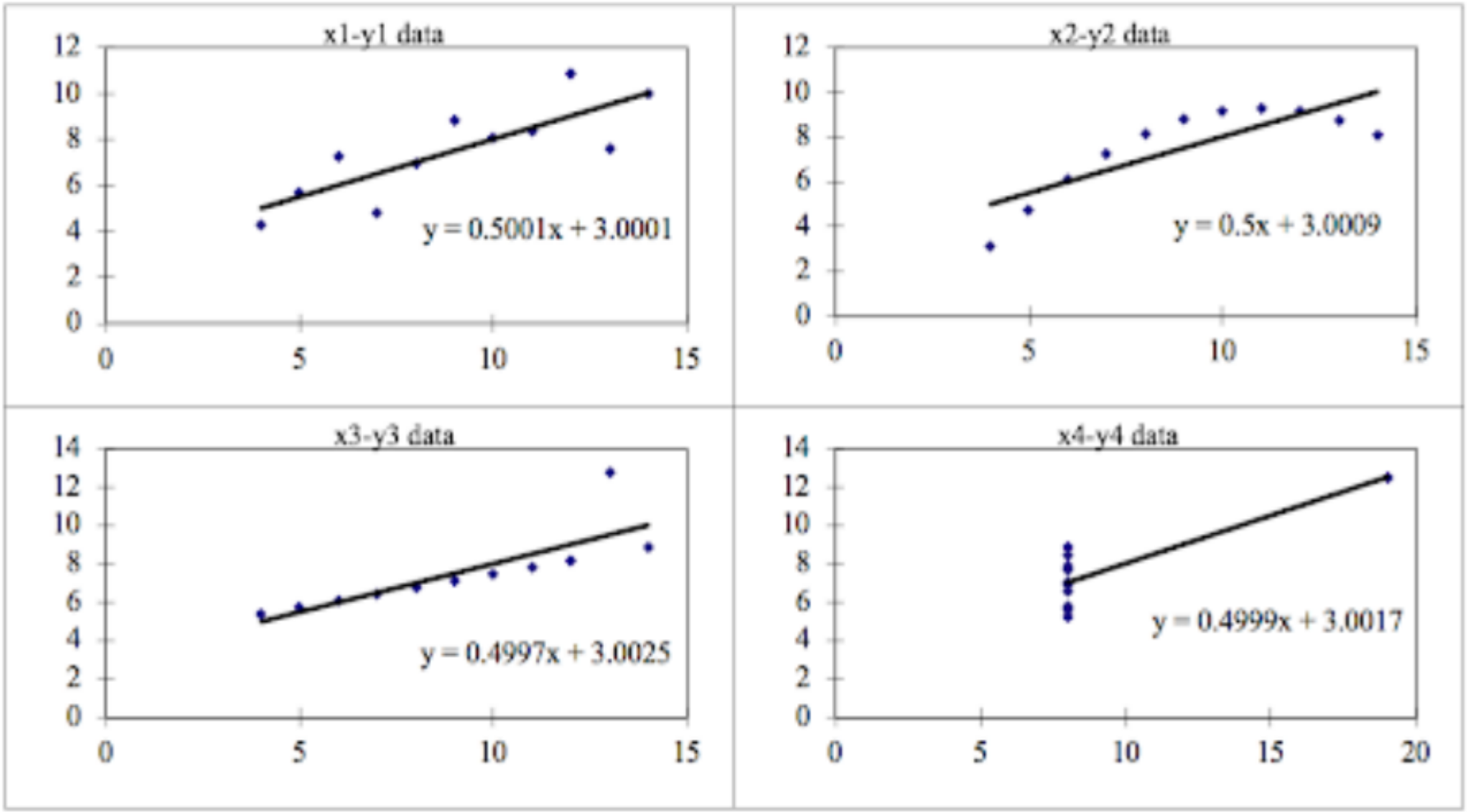
2. Explain the Anscombe’s quartet in detail.

Answer –

Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each dataset consists of eleven points. The primary purpose of Anscombe’s quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give accurate representation of two datasets being compared.

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
Summary Statistics								
N	11	11	11	11	11	11	11	11
Mean	9	7.5	9	7.500909	9	7.5	8	7.5
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

However, when these models are plotted on a scatter plot, each dataset shows a different kind of plot that isn't interpretable by any regression algorithm, as you can see below –



We can describe the four data sets as:

- **Data Set 1:** fits the linear regression model pretty well.
- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

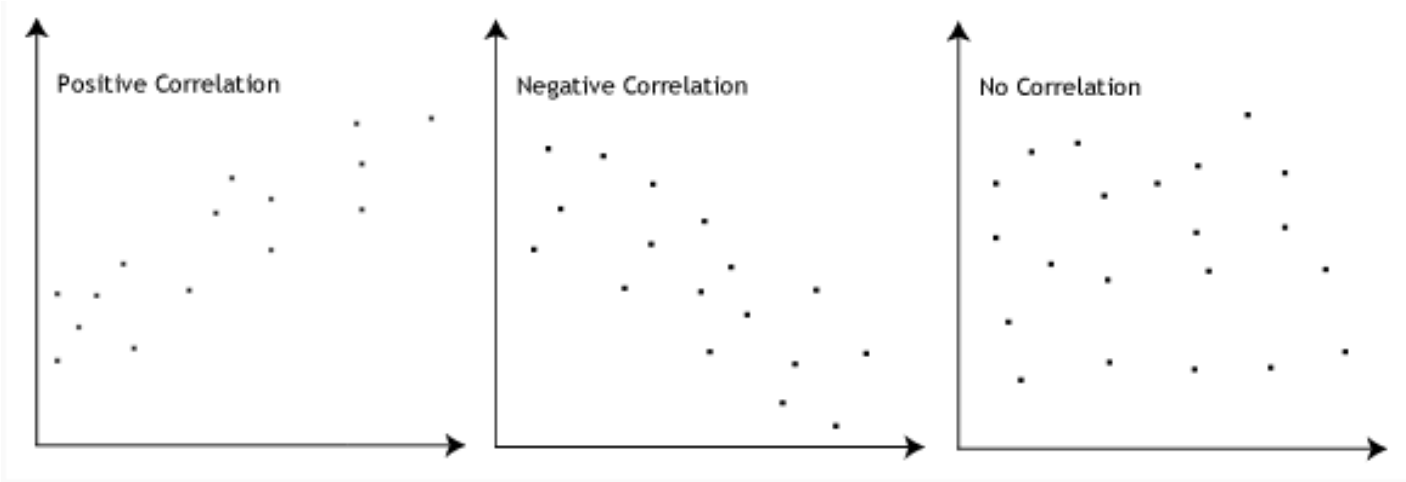
3. What is Pearson's R?

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 opposite direction with low values of one variable associated with the high values of the other, then the correlation coefficient will be Negative.

The Pearson's 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 as below –



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

Answer -

Scaling is a technique used to for in pre-processing during building a ML model to standardize the independent feature variables in the dataset in a fixed range. The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.

The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

Example –

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

S. No	Normalized Scaling	Standardized Scaling
1	Minimum & Maximum value of features are used for Scaling	Mean & 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?

Answer –

The value of VIF is infinite when there is a perfect correlation between the two independent variables. The R Squared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that, there is a problem of Multicollinearity and one of these variables need to be dropped in order to define a working model for regression.

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

Answer –

The quantile-quantile (Q-Q) plot is a graphical technique for determining if two datasets come from populations with common distribution.

Use of (Q-Q) plot:

A q-q plot is a plot of the quantitles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction(or percent) of points below the given value. That is, the 0.3(or 30%) quantile is the point at which 30% of the data fall below and 70% fall above the value. A45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line.

The greater the evidence for the conclusion that the two datasets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumptions of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insights into the nature of the difference than analytical methods such as the chi-square and kolmogorov-Smirnov 2-sample tests.