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 can be explained as the statistical method used for analyzing linear behavior within a dependent variable and the given set of independent variables. Linear behavior within variables meaning that when one or more variables value changes the dependent variable value will also change simultaneously.

Mathematics representation of relationship:

$Y = mX + c$

Here is the meaning –

Y – Dependent variable which we want to predict

X – Independent variables used for predictions

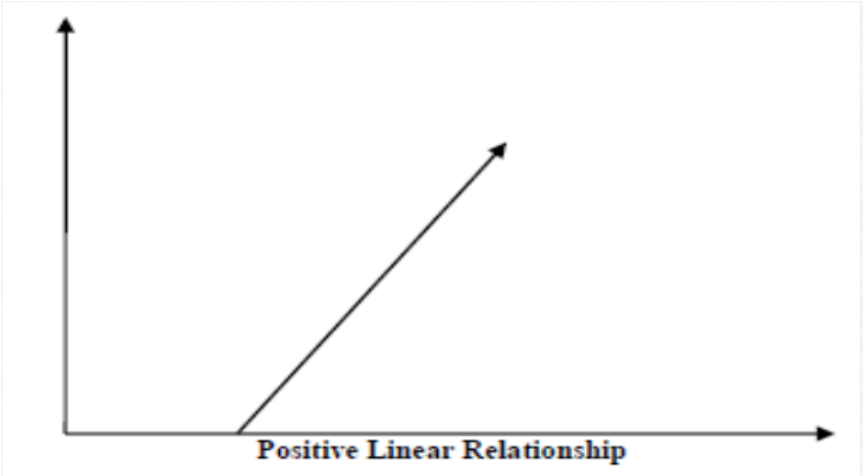
m – Slope of the regression line, showing effect of X has on the Y

c – Constant, known as Y-Intercept. If X = 0, then Y = c

Linear Relationship may be a positive or negative which are briefly stated as –

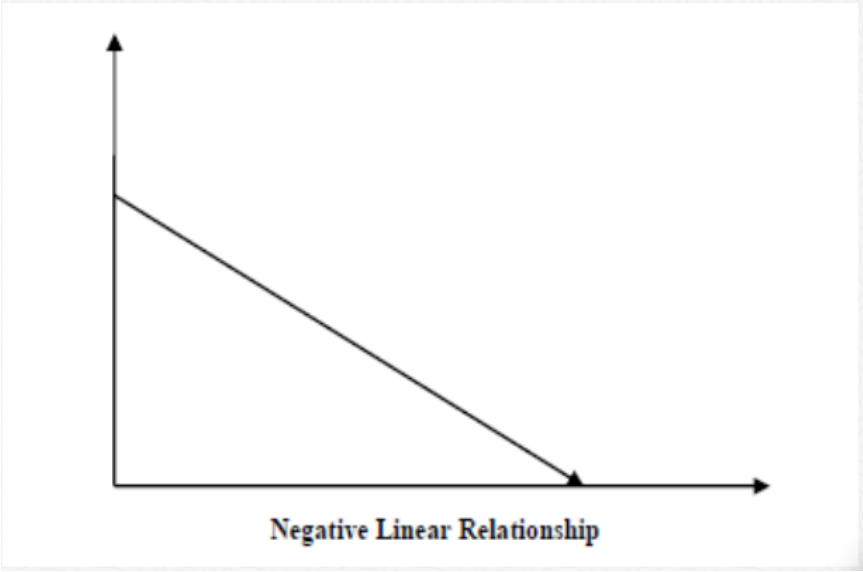
Positive Linear Relationship –

A linear relationship is called as Positive if the independent and dependent variable increase simultaneously, through the graph we can see if one value is Increasing then corresponding value is also increasing.



Negative Linear Relationship –

A linear relationship is called as Negative if the Independent variable increases and dependent variable Decreases, graph shows how the one value is increasing and the corresponding value is decreasing.



There are following two types of Linear Regression –

Simple Linear Regression – Single independent and single dependent variables

Multiple Linear Regression – Many independent and single dependent variables

Assumptions :

Multi-colinearity – It is assumed that the linear regression model will have very less or no multi-colinearity within the data. Multi-colinearity exists if there is dependency present among the independent variables.

Auto-correlation – It is assumed that model will have very less or no auto-correlation within the data. As auto-correlation exists when residual errors posses dependency within them.

Relationship within variables – Linear Regression model consider that the relationship within the independent and dependent variable will be linear only.

Normality of error terms – There should be Normal distribution of error terms.

Homoscedasticity - Residual values doesn’t shows any visible pattern

2. Explain the Anscombe’s quartet in detail.

Answer –

Anscombe's quartet was invented by statistician Francis Anscombe. It is sample modal which determines the significances of data visualization. It consist of 4 data-set of 11 (x, y) coordinates. Importance of this dataset is that they have equal numbers for their mean, variance, standard deviation etc, only differs when represented in terms of graph. Each and every graph explains own story though have similar stats

Calculating values :

Avg val of x = 9

Avg val of y = 7.50

Variance of x = 11

Variance of y = 4.12

Correlation Coefficient = 0.816

Linear Regression Eq – y = 0.5x+3

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

Despite of having similar statistics values for the above dataset, following is the graphical representation by plotting x & y coordinates.

Dataset I – Showing linear relationship along

With little variance

Dataset II – Showing curve with no linear

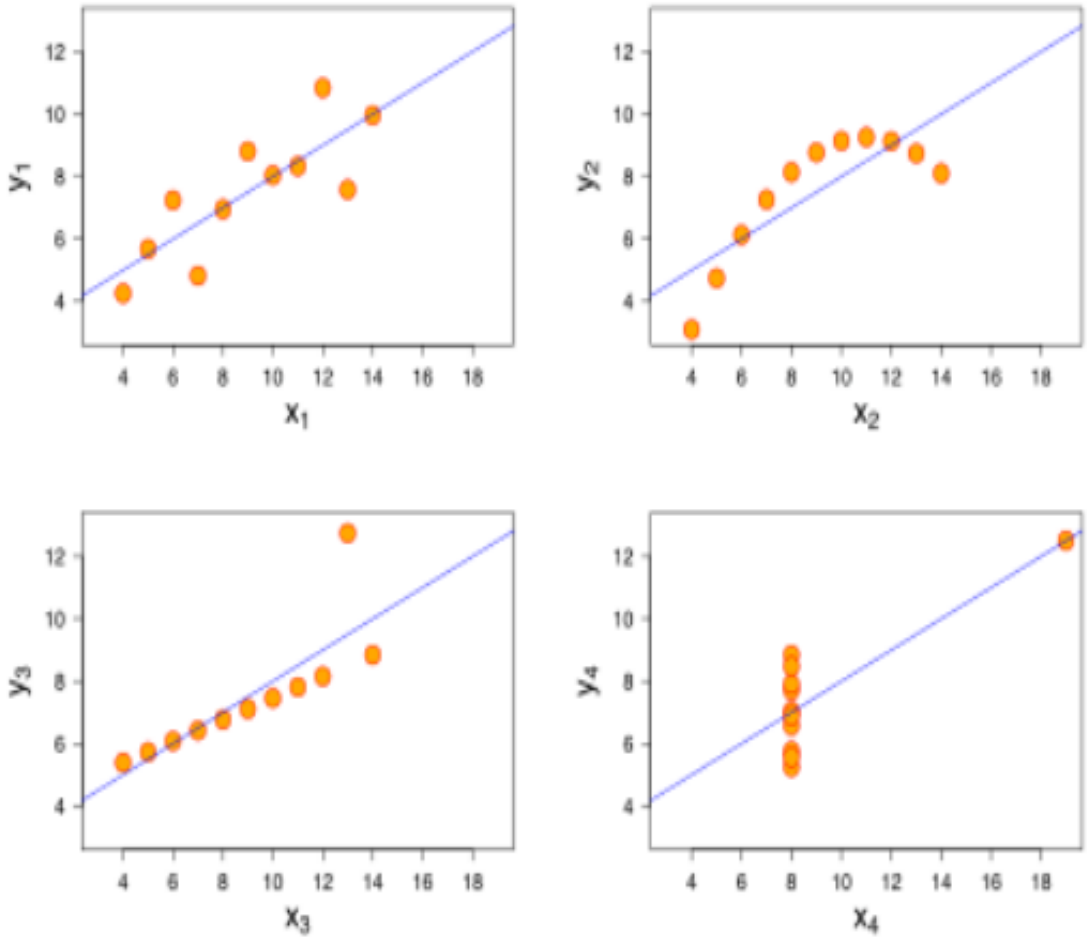
Relationship

Dataset III – Showing strong linear relationship

With just 1 outlier

Dataset IV – Value of x remaining same with just

1 outlier



.

3. What is Pearson’s R?

Answer –

Pearson’s R is a numerical summary showing the strength for the linear regression model with the dependent variable and independent variables. It will be Positive when the variables increases or decreases in parallel. It will be Negative when the variables increases or decreases in parallel.

R value will have value within from +1 to -1.

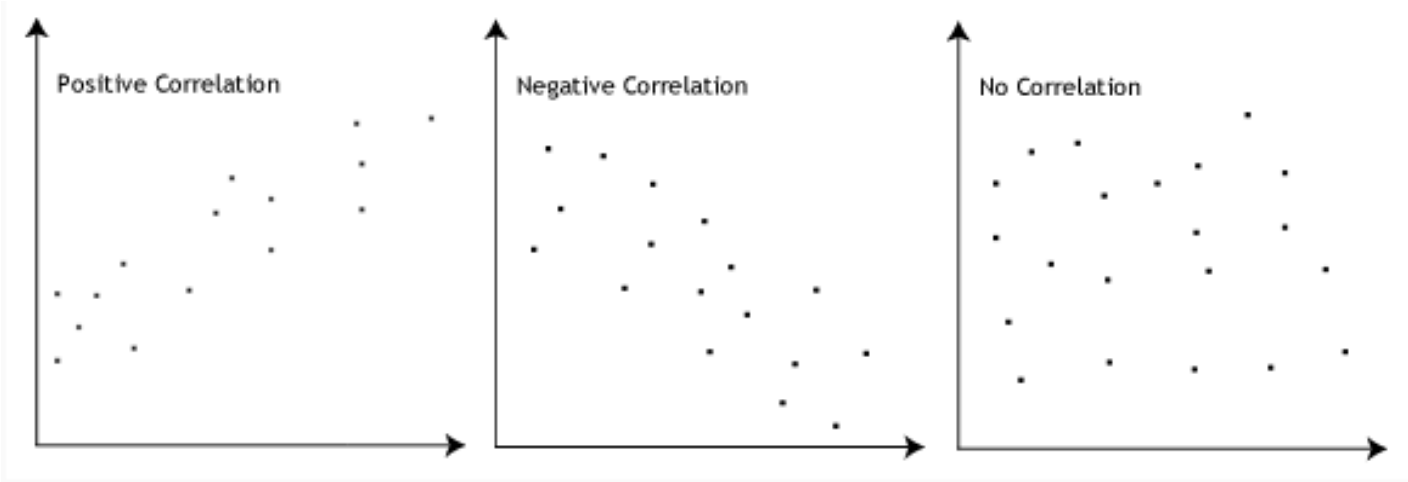
R = 0 meaning that there will be no association between the two variables.

R > 0 meaning a positive association, i.e. both increases or decrease at the same time.

R < 0 meaning a negative association, i.e. when one variable increases another one will decreases and vice-versa.

Following graph show the pictorial representation of the above concepts -

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 methodology which used during pre-processing phase of building a Machine Learning model in order to standardize the independent variables within the dataset to a fixed range. The dataset may have many variables which might be highly ranging within high magnitudes and units. Without scaling there could be modelling issues, leading to incorrect modelling considering mismatch in the units for the various variables available in the model.

The difference between normalized and standardized is that while normalizing bringing all the data points in a range between 0 and 1, and standardizing replaces the values with their Z scores.

Example –

Let say we have a model is not following variable scaling technique then it can be consider that the value of 3K meters to be greater than 5 km however which is incorrect hence the model gives us false predictions. Thus by using Scaling to bring all the values to the equal magnitudes and therefore handle this issues.

S. No	Normalized Scaling	Standardized Scaling
1	Min & Max value of variables are used for Scaling	StdDev & Mean is used for Scaling
2	Should be used during different scales of features	Should be used during the requirement of having zero mean and unit standard deviation
3	Generally Scales values varies within [0,1] or [-1,1]	No range defined
4	Outliers will affect it	Outliers won’t affect it
5	MinMaxScaler for Normalization provided by Scikit-Learn	StandardScaler for Standardization provided by Scikit-Learn

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

Answer –

The perfect correlation between the two independent variables then we have R Squared value is 1. Making VIF infinity, as VIF equals to 1/(1-R2). According to this, it shows the presence of Multicolinearity and which can be removed by dropping one of them in order to re-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 pictorial representation used to understand, if 2 datasets coming from the populations with common distributions.

Use of (Q-Q) plot:

A q-q plot is a plot of the quantiles between first data set plotted against the quantiles of the second dataset. Quantile, is the fraction(or percent) of the points below the given value.

Quantile 0.3(or 30%) - Meaning 30% of the data comes under and 70% comes above value point

A45-degree reference line displaying the distribution of the population of the two sets. Maximum plotting on the lines shows the greater departure. Which act as the proof for the conclusion that the two datasets are being plotted from the population of two different distributions.

Importance of Q-Q plot:

Considering there are two different data sets, we need to justify the assumptions of a common distributions. In this case, the location and scale can pool the data sets to get common location and scale. If it differs, it would be useful for knowing the dissimilarities. Q-Q plot depicts some more granularities of the differences compared to statistics analysis such as kolmogorov-Smirnov 2-sample tests and chi-square.