3/16/2022

Subjective Questions

1. Assignment Based
2. General Based

Naveen Kumar Jagadeesan
UPGRAD STUDENT

1. 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?

1. The Bike Sharing Bookings are high during FALL Season.
2. The Bike Sharing Bookings are high in the year 2019 and low in 2018.
3. The Bike Sharing Bookings are high in the middle of the year (may to oct).
4. At start of the year tends to increase and at the end of the year it decreases
5. The Bike Sharing Bookings are high in Weekends and holidays
6. The Bike Sharing Bookings are high when they slightly moving towards weekends.
7. The Bike Sharing Bookings are high in the clear weather compared to mist and light snow.

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

If we have all dummy variables without dropping means it will lead to multicollinearity between the dummy variables. So, this is one of the main reasons we are following (n-1) concept for n dummy variables. Also, we will get more no of information's with the (n-1) itself. Dropping that one column will not be a major impact on the whole datasets. It will further help us to drill down the process and make it easier.

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

we can say temp and atemp are the two variables which are highly correlated with each other and correlated more against the target variable. its almost 0.99 percent correlated with cnt. It can be checked using pair plot and heat map.

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

1. Error terms are normally distributed with mean zero.
2. The Probability of distribution of the errors has constant variance.
3. Error values are statistically independent and not depend on any other values.
4. Independent variables having linear relationships with the dependent variables

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

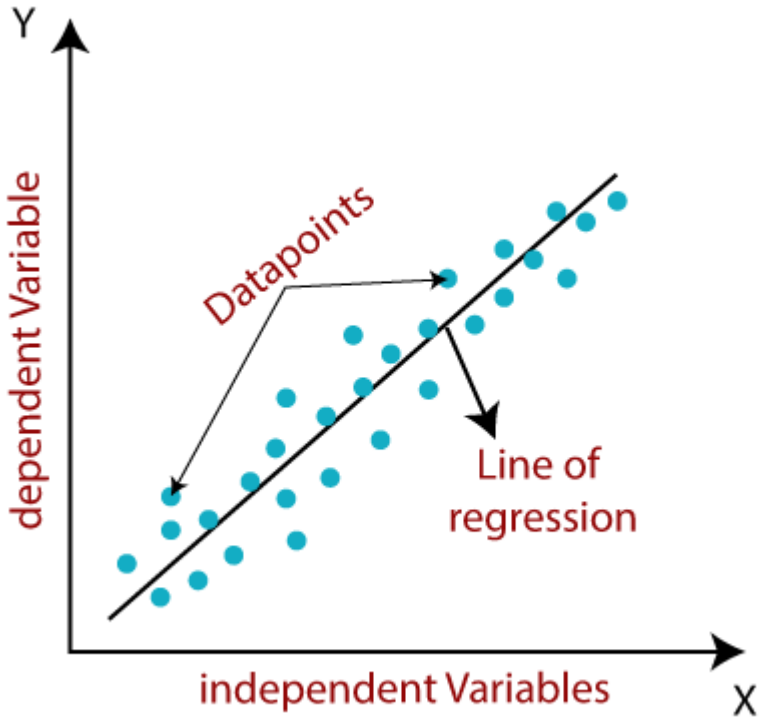
1. Temp - 0.548
2. Year - 0.233
3. Winter - 0.131

These are the 3 variables contributed significantly towards explaining the demand of shared bikes.

2. General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is one of the static methods that is used for predictive analysis. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables. It finds how the value of the dependent variable is changing according to the value of the independent variable.



$$y= a0+a1x+ \varepsilon$$

Y = Dependent Variable (Target Variable)
X = Independent Variable (predictor Variable)
a0 = intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

Simple Linear Regression:

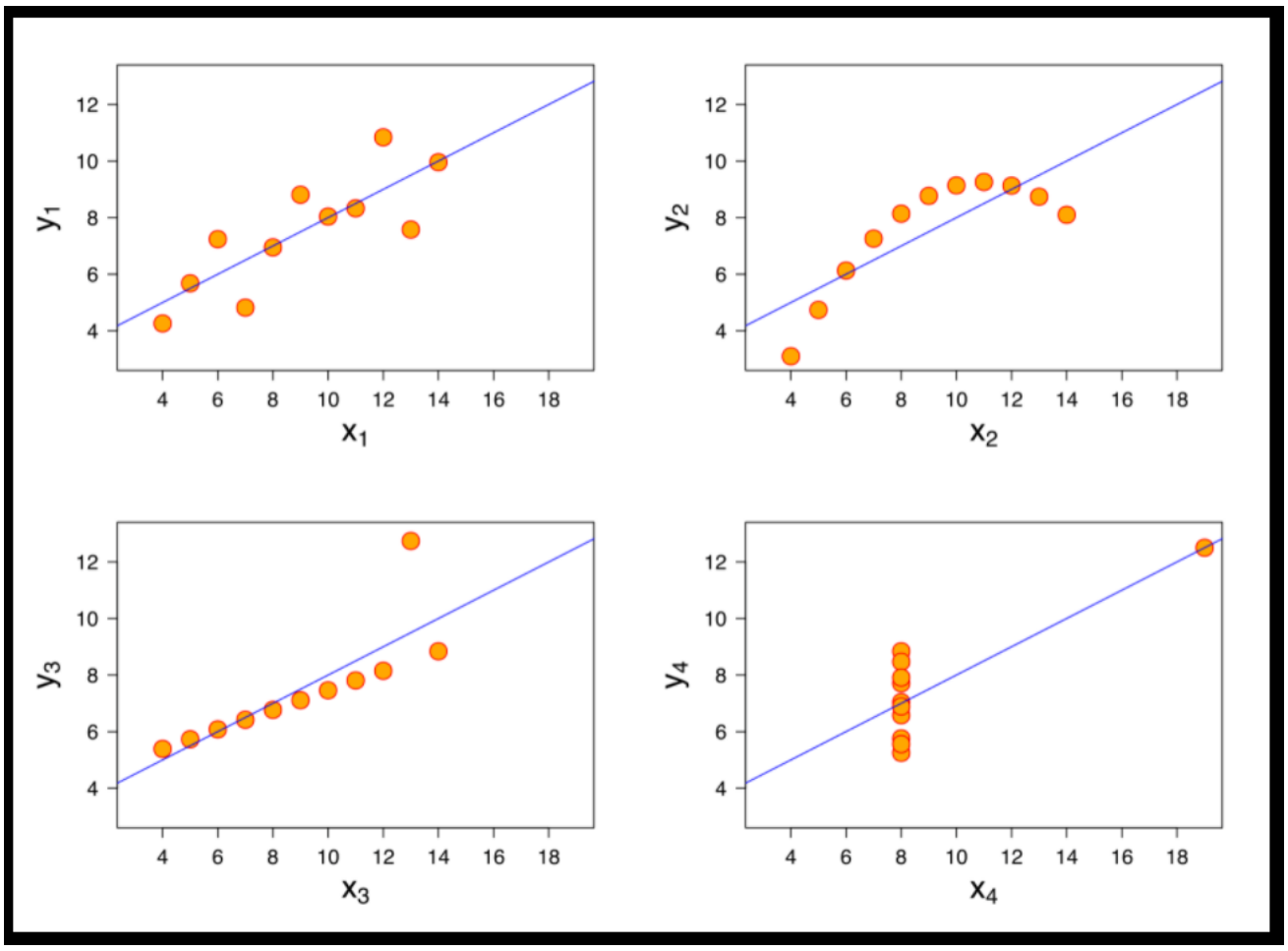
If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

Multiple Linear Regression:

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, but very different distributions and appear very different in graph. It was developed by Francis Anscombe.

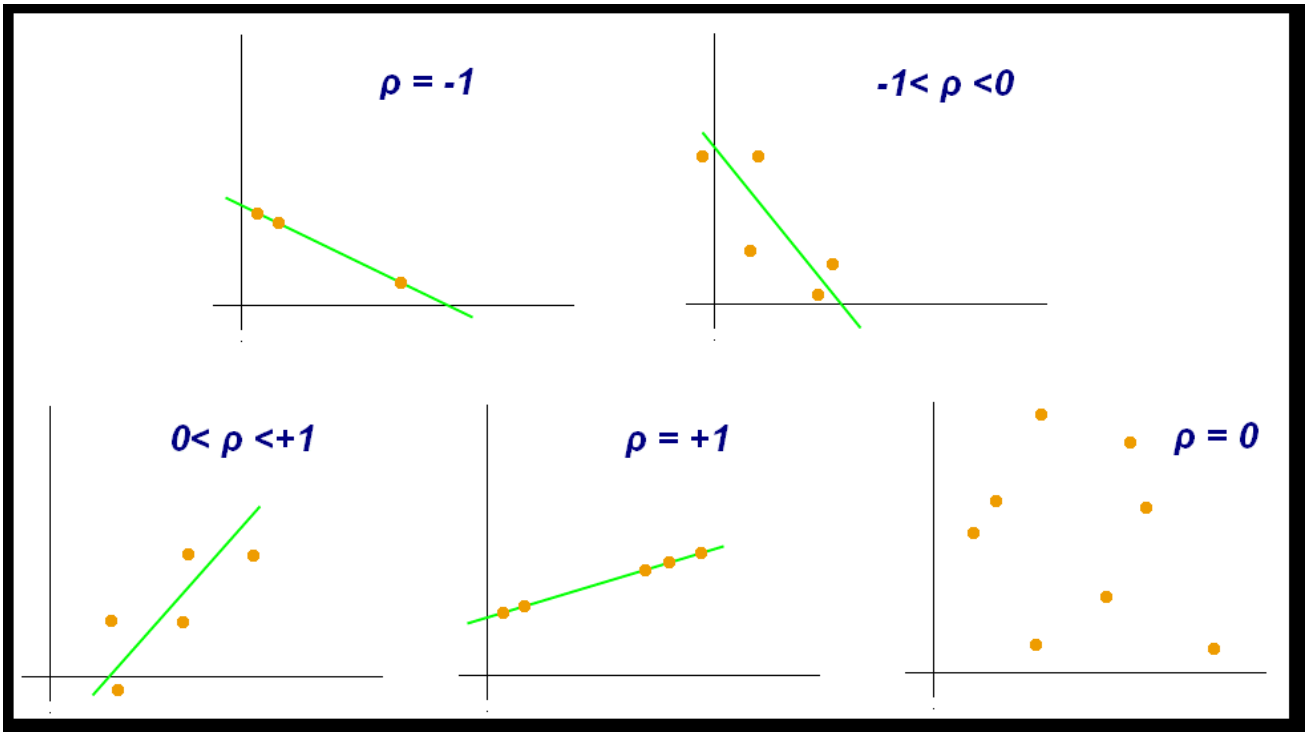


- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) has a relationship between the two variables, but it is not linear.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line .Ex we can say the regression line is about 0.78 instead of 1.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Note: The Diagrams are taken from the Wikipedia

3. What is Pearson's R?

Pearson R also Known as Pearson's relationship Co-efficient, value measures the strength of the linear relationship between 2 variables. It lies between -1 to 1. It is used to find the co relation between the variables.



Note: The Diagrams are taken from the Wikipedia

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

Feature Scaling is a method used to normalize or standardize the range of independent variables or features of data. 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.

Standardization and Normalization formulas

Standardization = X = x-mean(x) / sd(x)
Normalization = X = x-min(x) / (max(x) – min(x))

The result of standardization (or Z-score normalization) is that the features will be rescaled to ensure the mean and the standard deviation are 0 and 1.

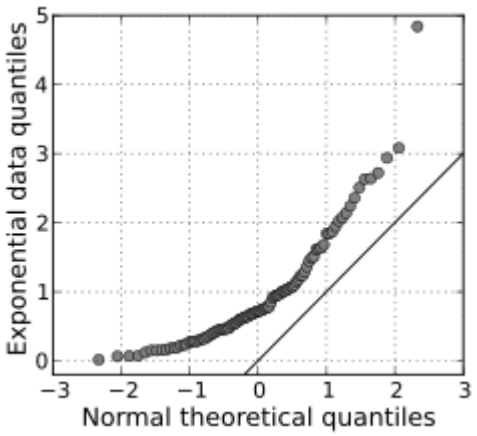
Another common approach is the so-called max/min normalization (min/max scaling). This technique is to re-scales features with a distribution value between 0 and 1. For every feature, the minimum value of that feature gets transformed into 0 and the maximum value gets transformed into 1.

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

1. For perfect correlation, VIF = infinity. This shows a perfect correlation between two independent variables.
2. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.
3. To Overcome we need to drop one of the variables from the dataset which is causing this multicollinearity.
4. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

1. Q-Q Plots are plots of two quantiles against each other.
2. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.
3. 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



Note: The Diagrams are taken from the Wikipedia

4. 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.
5. 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.