Bike Sharing Assignment
Submitted By
Moumita Ghosh

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)

Ans.

- Bike demand is the highest in the fall .
- Bike demand is the lower in spring.
- Bike demand in year 2019 is higher as compared to 2018.
- Bike demand is high in the months from May to October. In Winter season, its demand lowers due to wind speed.
- Bike demand is high if weather is partly cloudy while it is low when its rainy.
- The demand of bike is almost similar throughout the weekdays.
- Bike demand doesn't change whether day is working day or holiday.

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

Ans.
In order to create k-1 dummy variables, it is crucial since it may be used to erase superfluous columns while generating dummy variables.

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

Ans.
atemp and temp both have same correlation with target variable of 0.63 which isthe highest among all numerical variables.

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

Ans.
By plotting the residual distribution where it is normal distribution with mean value 0.

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

Ans.

The Top 3 features contributing significantly towards the demands of share bikes are:
weathersit_Light_Snow(negative correlation).
yr_2019(Positive correlation).
temp(Positive correlation).

General Subjective Questions

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

Ans.

Linear regression strives to show the relationship between two variables by applying a linear equation to observed data. One variable is supposed to be an independent variable, and the other is to be a dependent variable.

For example, the weight of the person is linearly related to his height. Hence this shows a linear relationship between the height and weight of the person. As the height is increased, the weight of the person also gets increased.

The measure of the extent of the relationship between two variables is shown by the **correlation coefficient**. The range of this coefficient lies between -1 to +1. This coefficient shows the strength of the association of the observed data for two variables.

A linear regression line equation is written in the form of:

Y = a + bX

where X is the independent variable and plotted along the x-axis

Y is the dependent variable and plotted along the y-axis

The slope of the line is b, and a is the intercept (the value of y when x = 0)

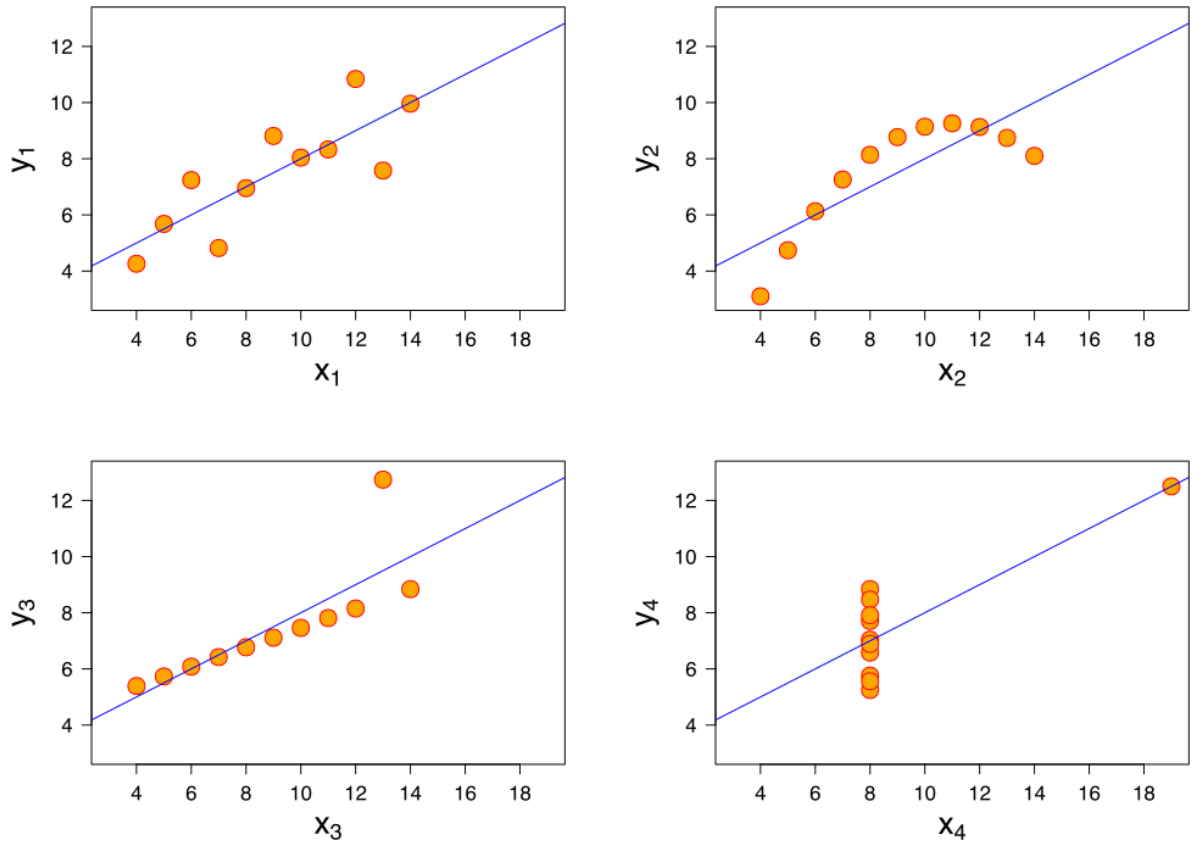
2. Explain the Anscombe's quartet in detail. (3 marks)

Ans.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

The simple statistics consists of mean, sample variance, co-relation coefficient, linear regression and R-Squared value.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed (Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet)

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

Ans.

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

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:

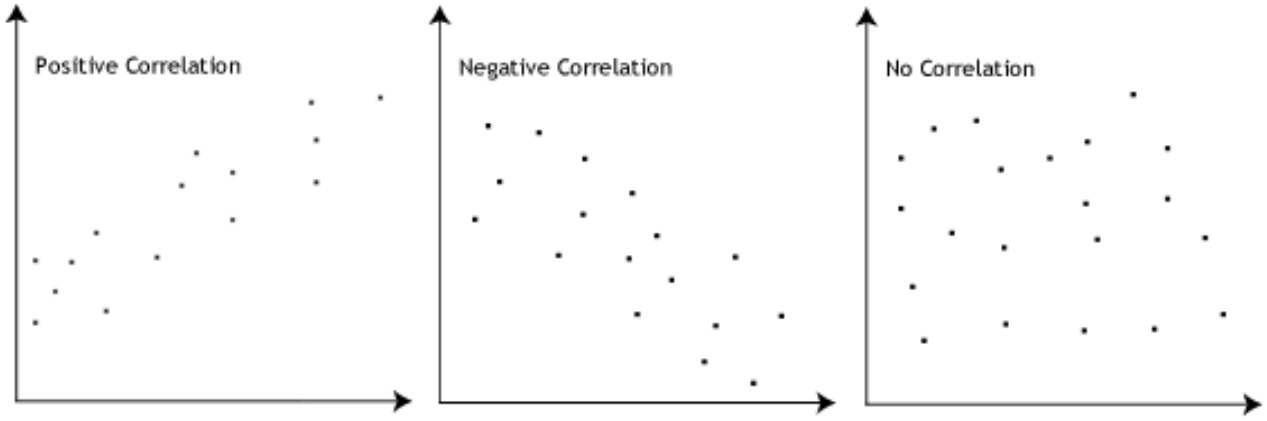


Image source: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

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

Ans.

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

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 modeling.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1.

sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

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

(3 marks)

Ans.

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

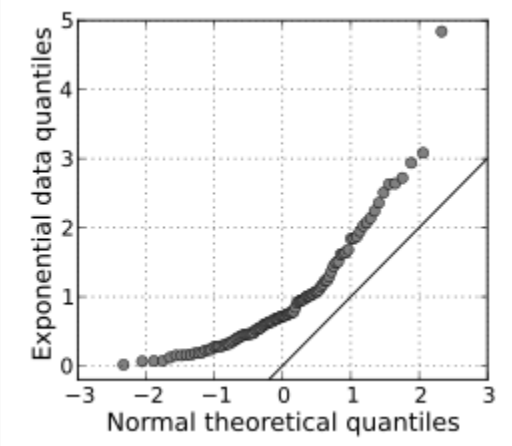
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

(3 marks)

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.