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)

Answer: Categorical variables are season, mnth, holiday, yr, weekday, workingday, weathersit

We can infer the following from these by analyzing boxplots of these variables vs target variable:

Season: fall and summer cover maximum variance followed by winter and spring.

Mnth: The median starts increasing from January till July and then starts to decrease

Holiday: maximum variance is covered by non-holiday days.

Yr: more bikes were rented in 2019 than 2018.

Weekday: medians of weekdays lie at approximately equal levels.

Workingday: medians of workingday lie at approximately equal levels.

Weathersit: most bikes were booked on clear weather days followed by mist and then least in light/rain.

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

Answer:

- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
- Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So, we do not need 3rd variable to identify the unfurnished.

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

Answer: 'atemp' 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? (3 marks)

Answer: We plot residuals versus predicted values, which are a part of standard regression output. The points should be symmetrically distributed around a diagonal line in the former plot or around horizontal line in the latter plot, with a roughly constant variance.

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

Answer: The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. weathersit_Light Snow/Rain
2. yr_2019
3. season_Winter

General Subjective Questions

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

Answer: Linear Regression Algorithm is a machine learning algorithm(a part of regression analysis). Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.

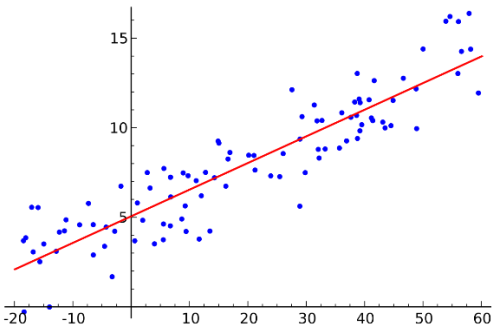
Here, we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

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
Use Cases of Linear Regression:

1. Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.
2. Price Prediction – Using regression to predict the change in price of stock or product.
3. Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.



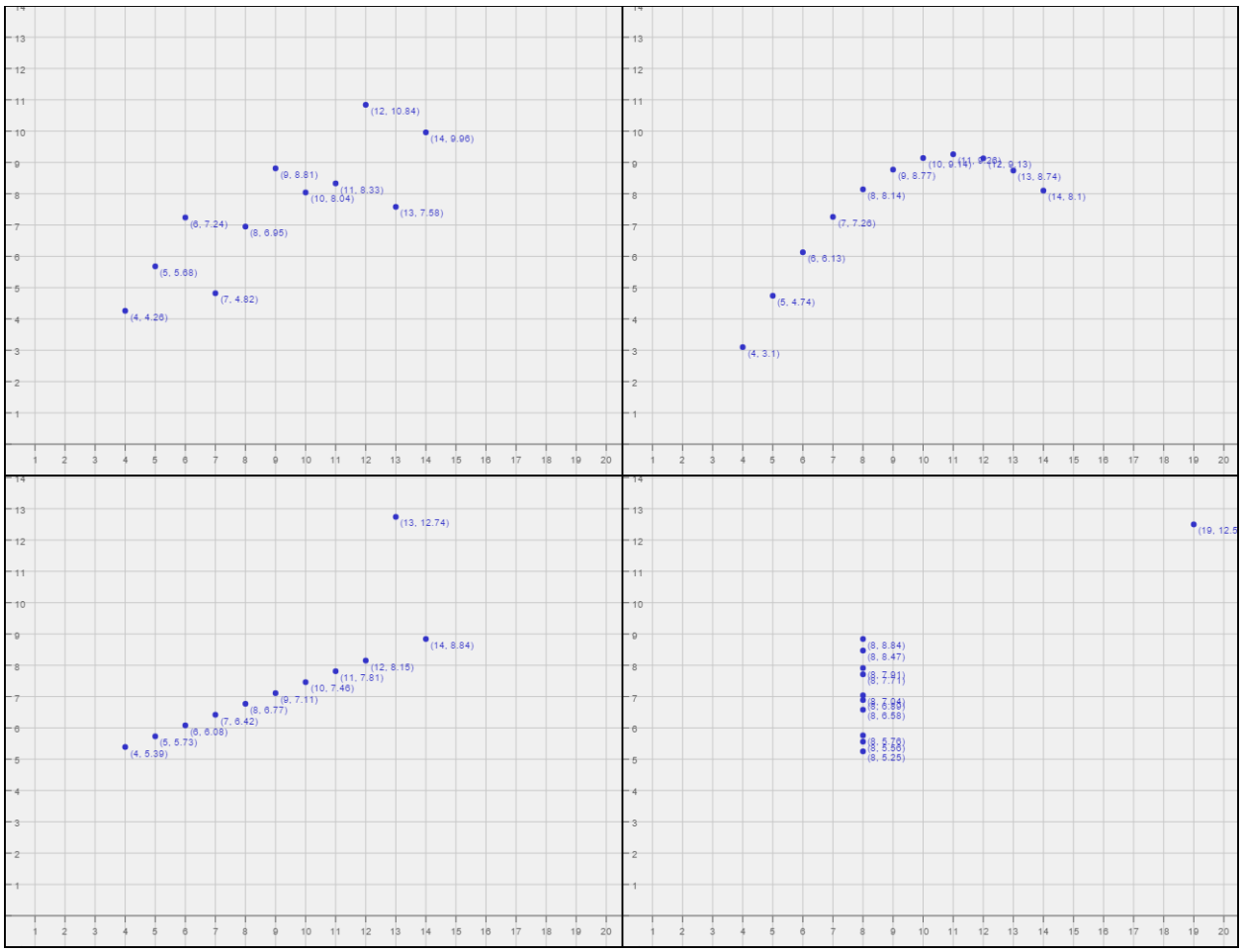
Linear Regression

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

Answer:

Statistics have long been used to describe data in general terms. For example, things like variance and standard deviation allow us to understand how much variation there was in some data without having to look at every data point individually. They give us a rough idea as to how consistent data is. However, knowing variance alone does not give you the full picture as to what the data truly is in its native form.

Statistics are great for describing general trends and aspects of data, but statistics alone can't fully depict any data set. Francis Anscombe realized this in 1973 and created several data sets, all with several identical statistical properties, to illustrate it. These data sets, collectively known as "Anscombe's Quartet," are shown below.



All four of these data sets have the same variance in x, variance in y, mean of x, mean of y, and linear regression. But, as you can clearly tell, they are all quite different from one another. So, what does this mean to you as a statistician?

Well, to start, Anscombe's Quartet is a great demonstration of the importance of graphing data to analyze it. Given simply variance values, means, and even linear regressions can not accurately portray data in its native form. Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed.

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

Answer:

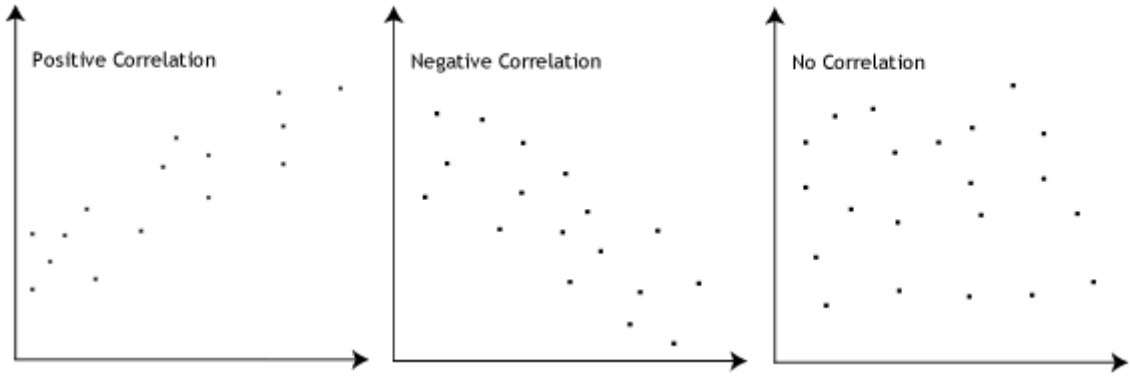
Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables.

The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.

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



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

Answer: Scaling is a technique to standardize the independent features present in the data in a fixed

range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

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, regardless of the unit of the values.

Normalization usually means to scale a variable to have a value between 0 and 1,

while **standardization** transforms data to have a mean of zero and a standard deviation of 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

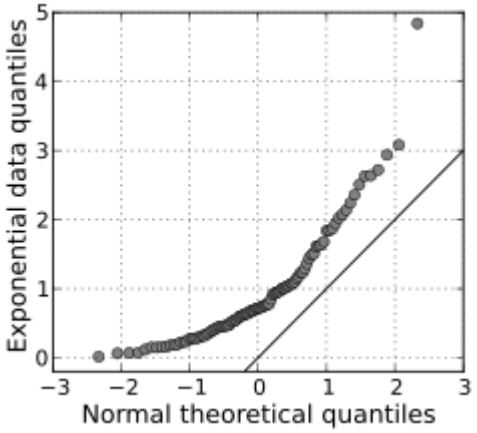
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

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.

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.



The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This Q-Q plot is called a **normal quantile-quantile (QQ) plot.** The points are not clustered on the 45-degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.