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:

Following are insights drawn from analysis of categorical variable from the dataset:

1) Fall season has highest demand for rental bikes.
2) Demand for next year has grown.
3) Demand is continuously growing till June; September month has highest demand. However, after September, demand is decreasing.
4) During holidays, demand has decreased.
5) The clear weathersit has highest demand.
6) During September, bike sharing is more. During the end and beginning of year, it is less.

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

Answer:

Following drop_first=True is important to use, as it helps in reducing the extra columns that gets created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
Dropping the first columns as (p-1) dummies can explain p categories . In weathersit, first column was not dropped so as not to lose the info about severe weather situation.

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

Answer:

Looking at the pair-plot among the numerical variables, temp and atemp have the highest correlation (0.63) with the target variable (cnt).

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

Answer:

After building Linear regression model, the assumptions were validated by performing below (3) analysis:

1) Residual Analysis:
   a) Errors are normally distributed with a mean of 0.
   b) Actual and predicted result follow the same pattern.
   c) The error terms are independent of each other.
2) R2 score for test predictions: R2 score for predictions on test data (0.807) is similar/close to R2 score of train data(0.822). This is a good R-squared score; hence we can see our model is performing good even on unseen data (test data)
3) Plot Test vs Predicted value test: The prediction for test data is very close to actuals.

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 target variable are:

1) Temp (positive coefficient)
2) weathersit_bad (negative coefficient)
3) Yr (positive coefficient)

General Subjective Questions

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

Answer:

Linear regression is a statistical regression method used for predictive analysis and shows the relationship between the continuous variables. It is used to understand the relationship between one dependent variable and several independent variables. The objective of linear regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

The linear regression model gives a sloped straight line describing the relationship within the variables. The linear regression model can be represented by the following equation:

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

The linear regression model provides a sloped straight line representing the relationship between the variables.

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

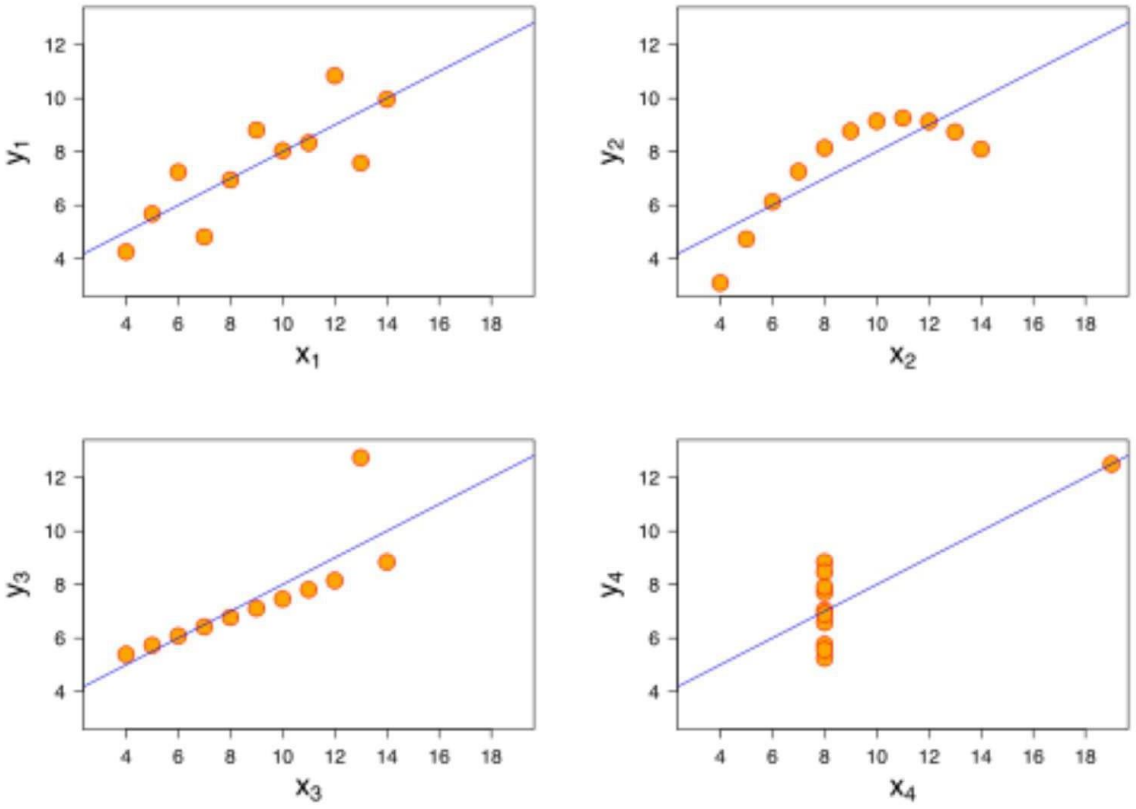
The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.

The cost function helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points. The cost function is used to find the accuracy of the mapping function that maps the input variable to the output variable. This mapping function is also known as the Hypothesis function. In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

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

Answer:

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. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.



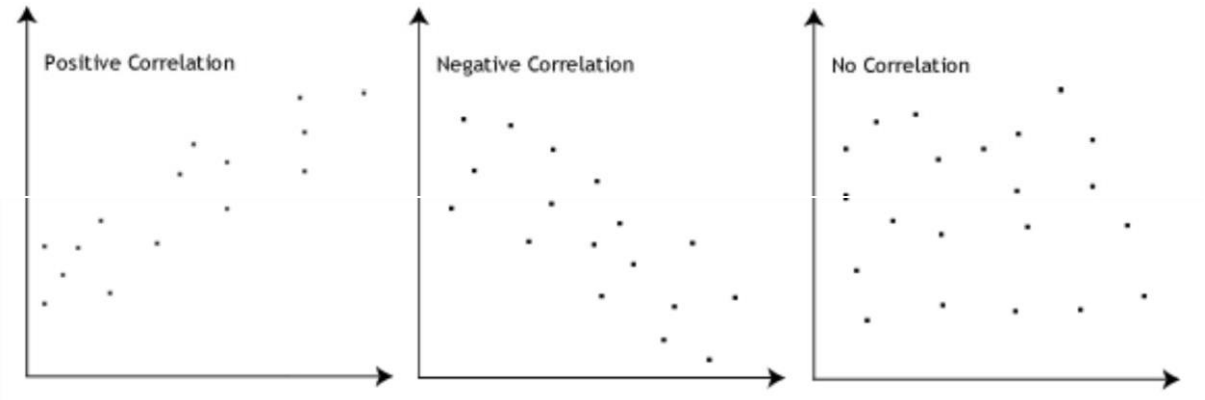
3. What is Pearson's R? (3 marks)

Answer:

Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

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



Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- r =correlation coefficient
- xi =values of the x-variable in a sample
- $\bar{x}$ =mean of the values of the x-variable
- yi=values of the y-variable in a sample
- $\bar{y}$ =mean of the values of the y-variable

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

Answer:

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So, we need to scale features because of two reasons:

- Ease of interpretation
- Faster convergence for gradient descent methods

You can scale the features using two very popular method:

i. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

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

ii. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

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

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

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

Answer:

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 QQ plot in linear regression. (3 marks)

Answer:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

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^{o} 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. 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.