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:

There are 6 categorical variables in the dataset. Box plot is used to study their effect on the dependent variable (‘cnt’).

The inference that could be derive were:

season:
Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

mnth:
Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

weathersit:
Almost 67% of the bike booking were happening during ‘good weathersit' with a median of close to 5000 booking (for the period of 2 years). This was followed by 'moderate' with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

holiday:
Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

weekday:
weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

workingday:
Almost 69% of the bike booking were happening in ‘workingday’ with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

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

Answer:

one of the columns can be generated completely from the others, and hence retaining this extra column does not add any new information for the modelling process, it would be good practice to always drop the first column after performing One Hot encoding, regardless of the algorithm of choice

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

Answer:

Using the pair plot, it is observed that there is a linear relation between temp and atemp variable with the predictor ‘cnt’

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

Answer:

- Using Residual Analysis

- Error terms

- Linearity check using R-Squared

- Homoscedacity

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

Answer:

Significant variables to predict the demand for shared bikes

- holiday
- temp
- hum

<u>General Subjective Questions</u>

1. Explain the linear regression algorithm in detail.

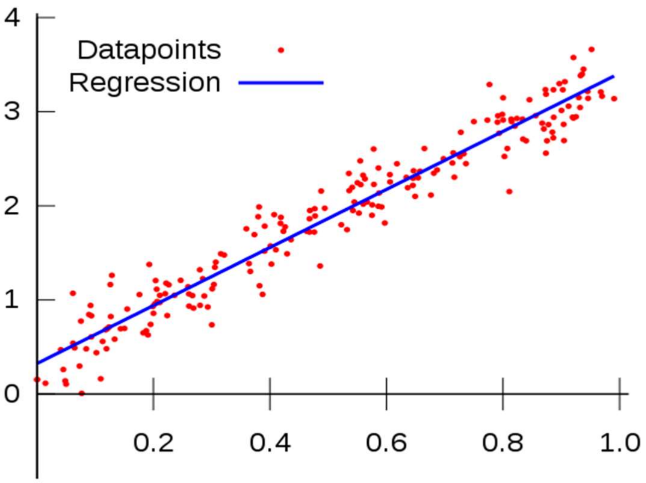
Answer:

Linear Regression is a machine learning algorithm based on supervised learning.

A simple linear regression model is an attempt to explain the relationship between a dependent and an independent variable using a straight line.The independent variable is also known as the predictor variable. And the dependent variables are also known as the output variables.

Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).

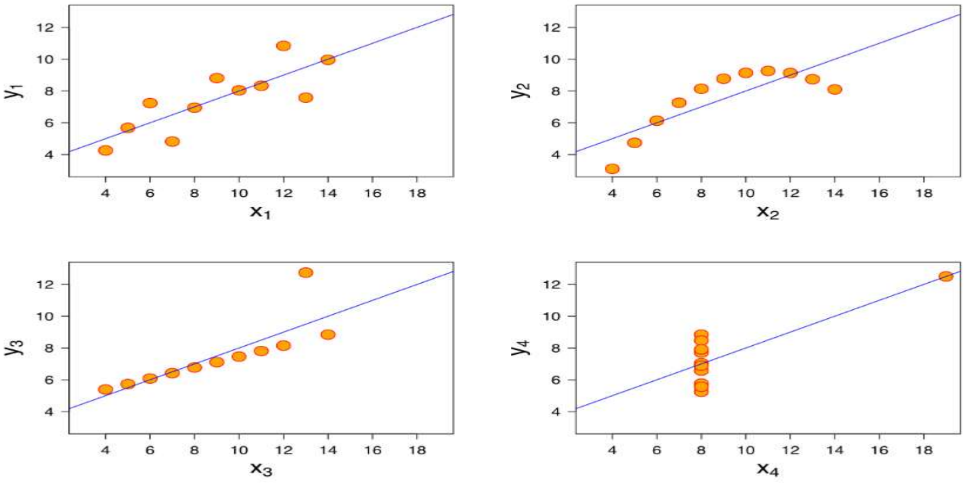


2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help to identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.)



3. What is Pearson's R?

Answer:

Pearson's *R* (also called Pearson's correlation) is a **correlation coefficient** commonly used in linear regression to shows the linear relationship between two sets of data.

The Pearson’s R method is the most common method to use for numerical variables; it assigns a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and − 1 is total negative correlation.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

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

Answer:

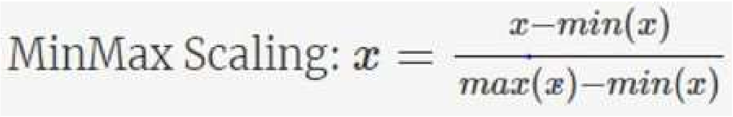
Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

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 modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
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-squared, etc.

Normalization:

It is also called as MinMax Scaling
It brings all of the data in the range of 0 and 1.

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

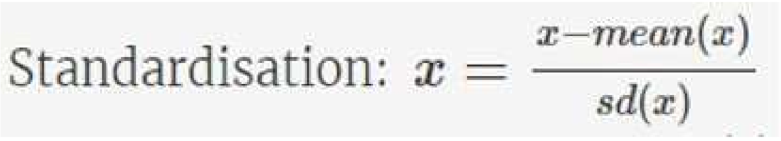


$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

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).

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



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

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

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

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 Q-Q plot in linear regression.

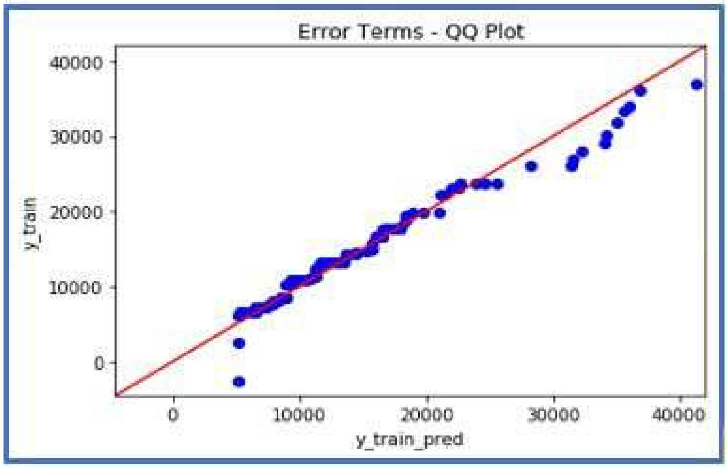
Answer:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

use and importance of a Q-Q plot in linear regression

It 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.


Error Terms - QQ Plot
y_train
y_train_pred