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:

* season : season (1:spring, 2:summer, 3:fall, 4:winter)

Fall has more bike rentals followed by summer and winter. So it should be one of the good predictor
* Seems bikes rentals increases year by year. So it should be one of the good predictor
* Months 5 to 10 has higher bike rentals compared to remain months.
* Seems non-holidays have more bike rentals compared to holdiays.
* There seems to be no significance difference on weekdays. So it might not be significance variable.
* Working days have higher bike rentals.
* Clear weather have higher bike rentals. So it should be one of the good predictor.

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

Answer:
Drop_first=True helps to reduce the redundant dummy column creation and which in turn helps to reduce the correlation among the dummy variables. So it is important to use this option.

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

Answer: atemp

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

Answer:

1) Error terms or Residuals we plot using Histogram and it should have normal distribution with 0 mean
2) For Multicollinearity we can check VIF values.

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:

3) temp : 0.5499
4) weathersit_3 : -0.2880
5) yr : 0.2331

General Subjective Questions

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

Answer:

Linear regression is one of the very basic forms of machine learning where 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 a is intercept and b is slope

X is independent variable

Y is target/ dependent variable

In case multiple linear regression where more than one independent variable influence that target variable we can write the equation as

Y= a+b1x1+b2x2+…….+bnxn

b1, b2, b3 are slopes/ coefficients of independent variables

a is constant or intercept

We can interpret this as one unit change in x1 variable cause to change b1 units of Y.

Similarly one unit change in any independent variables cause their respective coefficients times of change in Y.

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

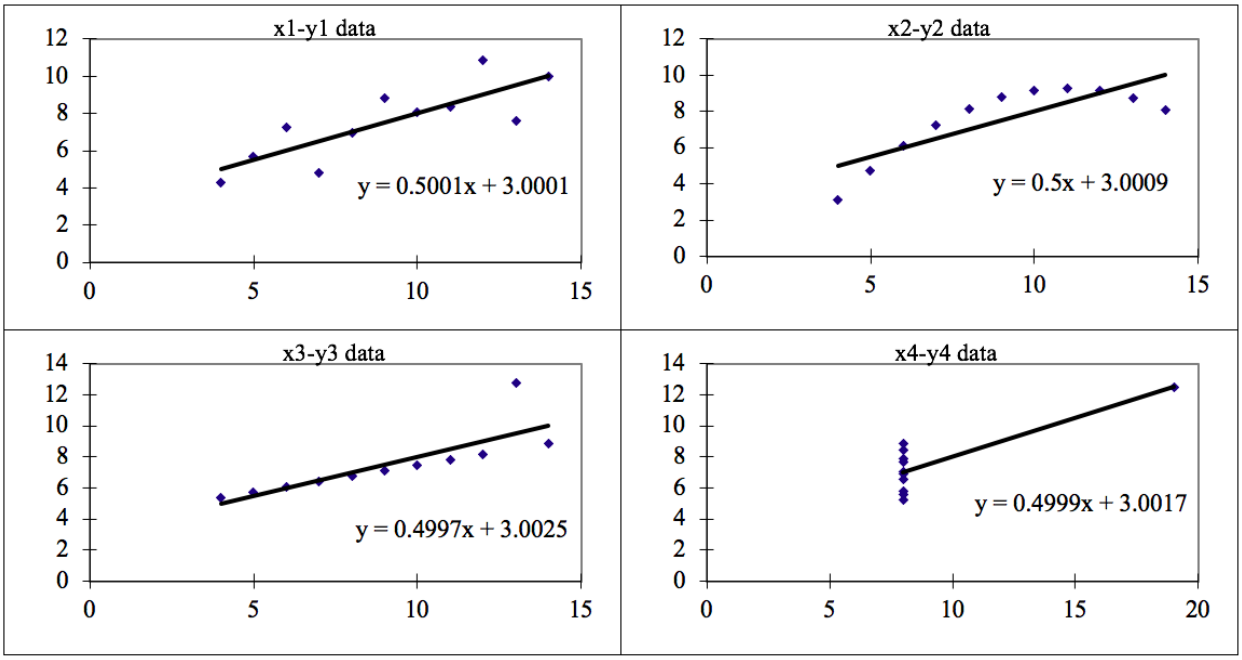
Answer:

Anscombe's quarter can be defined as four datasets which nearly identical in simple descriptive statistics (mean, standard deviation, coefficient and constant) but when we plot the data all four have very different distributions such that linear regression can't be applied to some of the datasets.

So which tells us that how importance is to first visualize the data and plot it before proceeding to build the model.

If you see below graphs all 4 has similar coefficient and constant but plots are completely different.

1) Plot1 has linear relation ship but where as remain 3 plots doesn't have linear relation and linear regression model is not right choice for these three datasets.



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

Answer:

Pearson's R is also called Pearson's correlation coefficient commonly used in linear regression. It is measure of linear relation between two sets of data.

It will lie between -1 to 1. -1 means they both negatively correlated and 1 positively correlated and 0 mean no correlation. Between 0 to -1 or 1 tells how much they are correlated each other.

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

Answer:

Scaling is data pre-processing step which is applied to features to normalize the data with in particular range. It helps to speed up the calculations in an algorithm.

Most of the times our dataset contains a features with highly varying in magnitude, units and range. If scaling is not done then algorithm would consider higher magnitude values as dominance ones. In order to avoid this dominance and speeding up the convergence of model we bring wider range of values into common scale.

There are two types of scaling. Normalize/Min-Max scaling and Standardize scaling.
The difference between these two is, in case of normalized scaling values will be typically scaled in range of [0,1]. And in case of standardization scaling values will be scaled to have mean of zero and standard deviation of 1 unit.

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

Answer:

The VIF formula is 1/1-R*R. It comes infinite when R- Square is 1. R-Square 1 means it is perfectly correlated. That means one of the variable in our model is perfectly correlated with others and so we need to drop that variable.

(3 marks)

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

(3 marks)

Q-Q is Quantile – Quantile plot which helps to determine if two datasets are from populations with common distributions or not.

In linear regression when we have training and test dataset received separately we can confirm using Q-Q plot that both datasets are from populations with same distributions.

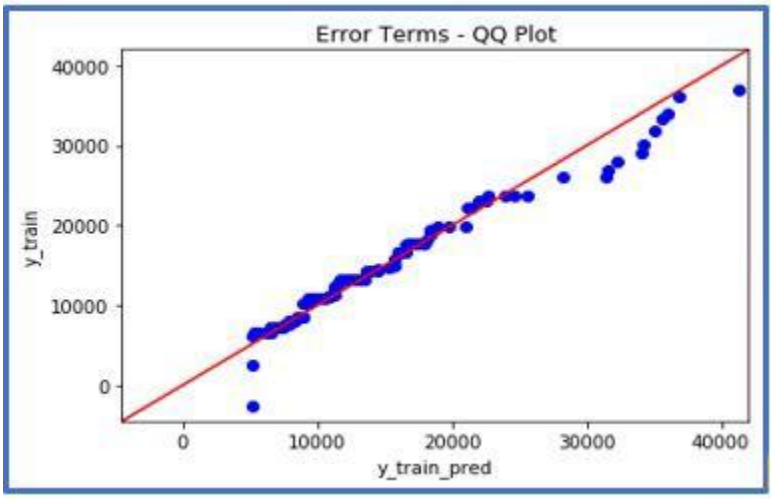
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Interpretation:

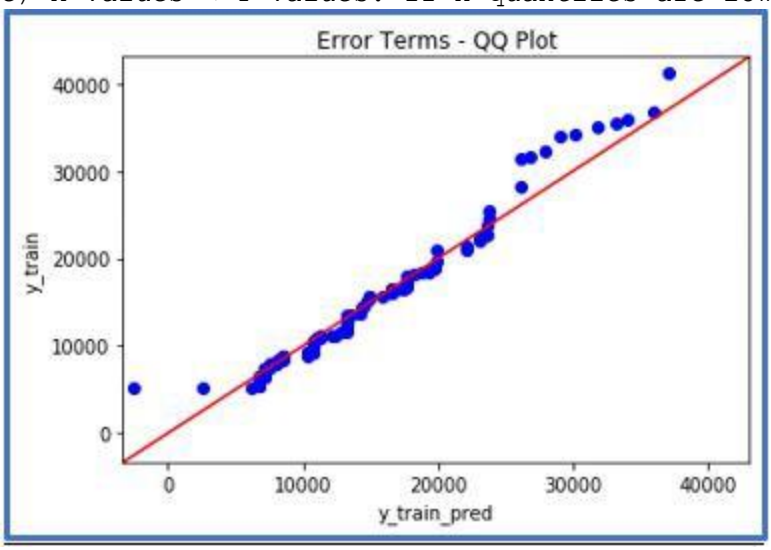
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Python:

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.