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?

From the boxplot plotted categorised by each categorical variable we could infer the following:

1. More people are taking bikes in clear weather.
2. From Jul to Oct there is a considerable increase in number of people taking bikes
3. By combining the results of above point and from plot of cnt and season. We can conclude the Fall and summer are important months for people taking bikes.
4. There is a considerable increase in people taking bikes in 2019 when compared to 2018

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

The drop_first=True parameter is used to delete the first column when creating a dummy variable. This will reduce the number of dummy variables created (n-1 dummy variables will get created).

If we didn't drop_first=True the number of varaibles will be created more (n dummy variables). This could cause redundancy in the dummy variables.

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

By checking the pairplot between the numerical variable we could see that temp and atemp has the highest correlation with cnt (target variable)

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

We checked the following criteria to validate the model

- Error terms are centered around 0 and normally distributed.
- Error terms do not follow any pattern.
- The VIF off all variables in final model is less than 5.
- Plotted the scatter plot of Y test vs Y pred and verified its linearity.
- And checked the R2 and Adjusted R2 to ensure model is not overfitted.

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

1. atemp

2. weathersit light

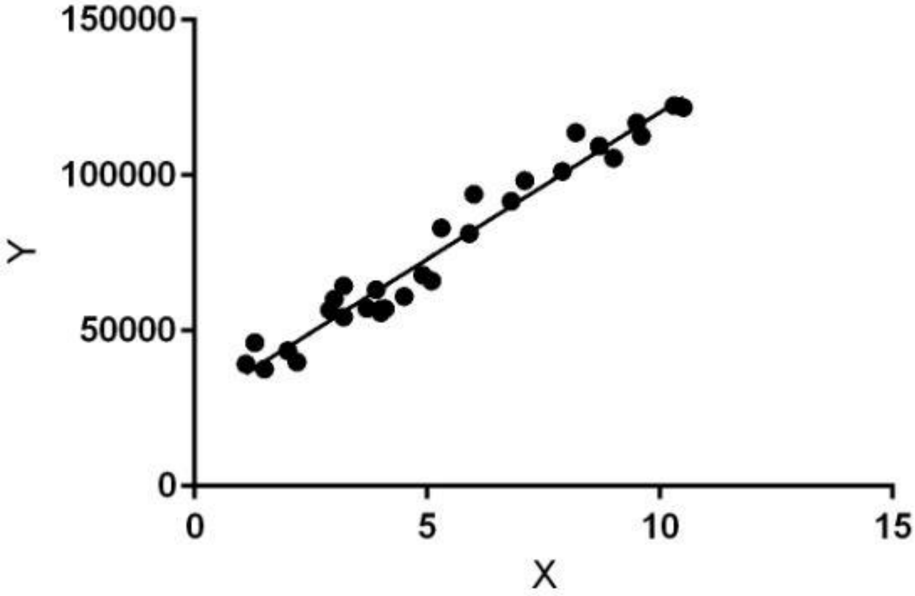
3. yr_2019

General Subjective Questions

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

Linear Regression is a machine learning algorithm based on supervised learning. Regression models a target prediction value based on independent variables. Regression is the statistical approach to find the relationship between variables. Hence, the Linear Regression assumes a linear relationship between variables.

Example:



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Depending on the number of input variables, the regression problem classified into

1) Simple linear regression

2) Multiple linear regression

Simple Linear Regression

Simple linear is an approach for predicting the quantitative response Y based on single predictor variable X.

$$Y = \underset{\text{Intercept}}{\beta_0} + \underset{\text{Coefficient / Slope}}{\beta_1} * X$$

This is the equation of straight-line having slope β1 and intercept β0.

Multiple Linear Regression

If we have p predictor variables, then a multiple linear regression model takes the form:

Y = β0 + β1X1 + β2X2 + … + βpXp + ε

where:

Y: The response variable

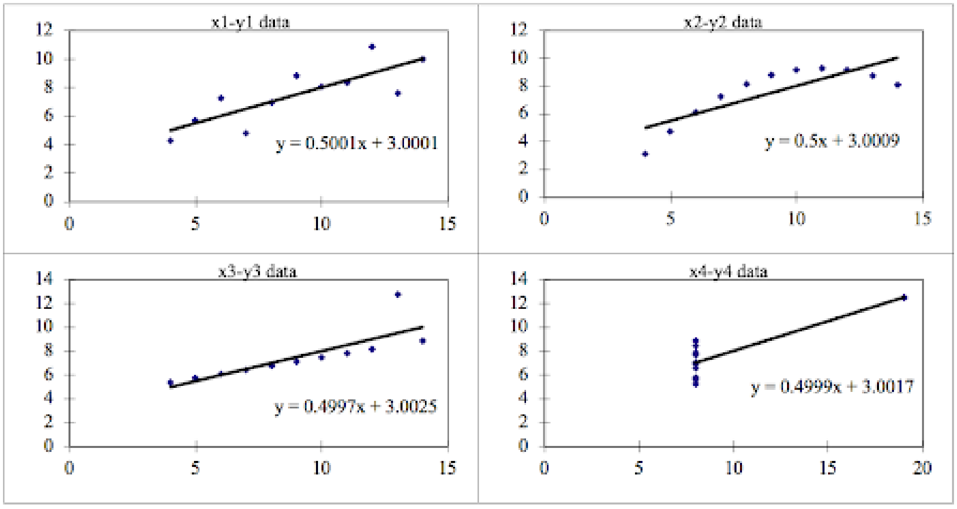
Xj: The jth predictor variable

βj: The average effect on Y of a one unit increase in Xj, holding all other predictors fixed ε: The error term

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



ANSCOMBE'S QUARTET FOUR DATASETS

Data Set 1: fits the linear regression model pretty well.

Data Set 2: cannot fit the linear regression model because the data is non-linear.

Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.

Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model

Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm.

3. What is Pearson's R?

The Pearson product-moment correlation coefficient depicts the extent that a change in one variable affects another variable. This relationship is measured by calculating the slope of the variables' linear regression.

The value of Person r can only take values ranging from +1 to -1 (both values inclusive). If the value of r is zero, there is no correlation between the variables.

If the value of r is greater than zero, there is a positive or direct correlation between the variables. Thus, a decrease in first variable will result in a decrease in the second variable.

If the value of r is less than zero, there is a negative or inverse correlation. Thus, a decrease in the first variable will result in an increase in the second variable.

When plotted on a diagram, a positive correlation will see a line which slopes downwards from left to right and a negative correlation will see a line which slopes downwards from right to left.

Example

A classic case of two variables affecting one another is demand and supply in an economy when the price of the product and the quantity demanded and supplied is known. The values are represented using a simple linear regression.

Pearson R shows that demand and supply have a positive correlation. As more consumers demand products, the amount suppliers are will to produce increases as well.

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

Scaling 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.

Normalisation

Normalisation, also known as min-max scaling, is a scaling technique whereby the values in a column are shifted so that they are bounded between a fixed range of 0 and 1.

Formula of Normalized scaling:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardisation

On the other hand, standardisation or Z-score normalisation is another scaling technique whereby the values in a column are rescaled so that they demonstrate the properties of a standard Gaussian distribution, that is mean = 0 and variance = 1.

Formula of Standardized scaling:

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

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

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.

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.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

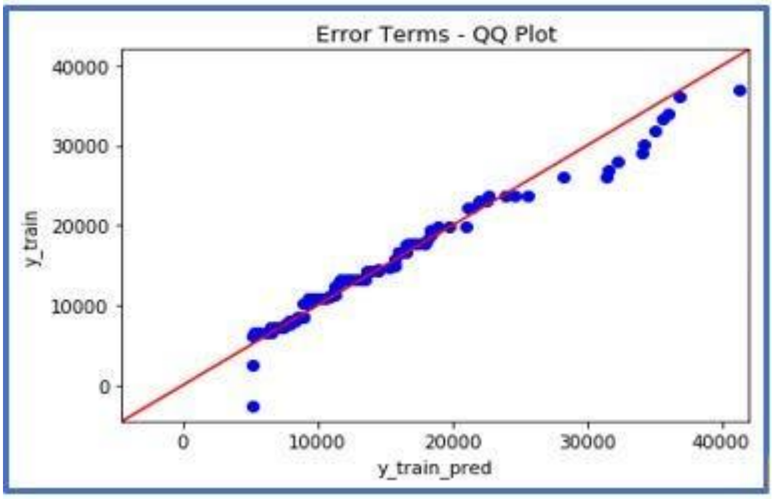
Interpretation:

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

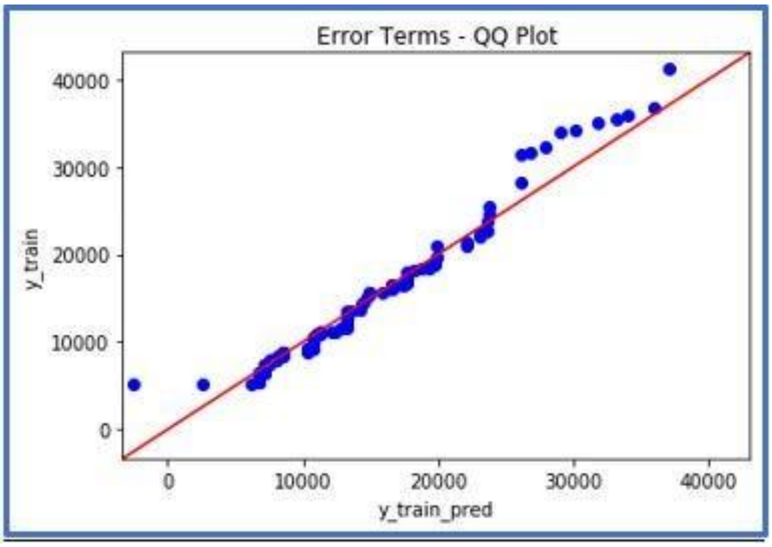
Below are the possible interpretations for two data sets.

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