

Assignment-based Subjective Questions

Q1-From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans-

- 1.The demand of bikes is less in the month of spring when compared with other seasons. The demand bike decreased with wind speed,
- 2.The demand of bikes is less for weathersit conditions are (Light Snow/Light Rain + Thunderstorm + Scattered clouds/Light Rain + Scattered clouds), (Mist + Cloudy/Mist + Broken clouds/Mist + Few clouds/Mist)
- 3.The demand of bikes is more on Sunday & workingday
- 4.Demand is high in the months 4,5,6,7,8,9,10

Q2-Why is it important to use drop_first=True during dummy variable creation?

Ans-

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.

Q3-Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans-

temp and atemp variables are having highest correlation.

Q4- How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans-

- looking at the Variance Inflation Factors (VIF). It is calculated by regressing each independent variable on all the others and calculating a score as follows:

$$VIF = 1/(1-R^2).$$

Hence, if there exists a linear relationship between an independent variable and the others, it will imply a large R-squared for the regression and thus a larger VIF. As a rule of thumb, VIFs scores above 5 are generally indicators of multi co-linearity (above 10 it can definitely be considered an issue).

- One more method is checking the P value, if P value is high we can say that it is statistically less significant.

Q5-Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans-

The Top 3 features contributing significantly towards the demands of share bikes are:

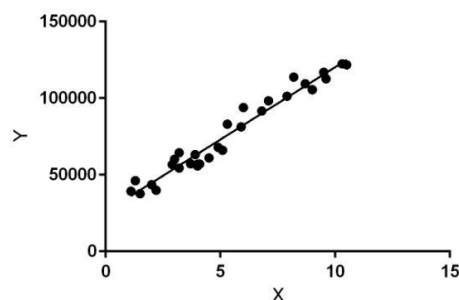
- Season(spring) has negative correlation,
- Workingday has positive correlation,
- Weathersit ((Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) & (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)) has negative correlation.
- weekday (Sunday) has positive correlation.

General Subjective Questions

1. Explain the linear regression algorithm in detail ?

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. 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 being 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). 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.

$$y = a_1 + a_2 \cdot x$$

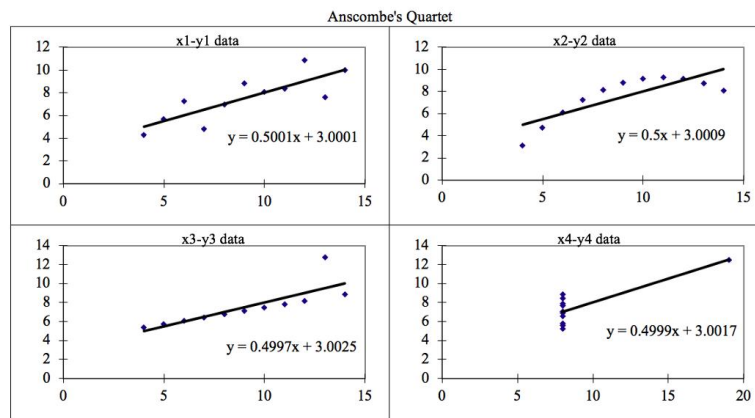
here, a_1 is intercept, a_2 is the coefficient of x , x : input training data, y : labels to data

2) Explain the Anscombe's quartet in detail ?

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.

Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets. 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 in order 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

The four datasets can be described as



Dataset 1: this fits the linear regression model pretty well.

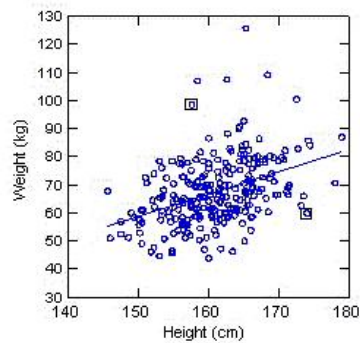
Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

3) What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.



"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrates.

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)

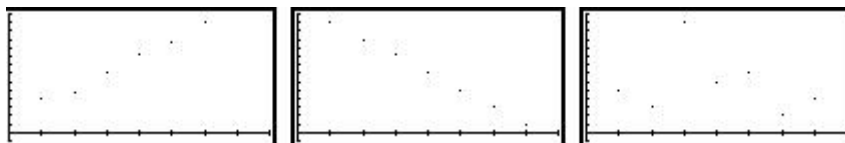
$r = 0$ means there is no linear association

$r > 0 < 5$ means there is a weak association

$r > 5 < 8$ means there is a moderate association

$r > 8$ means there is a strong association

The figure below shows some data sets and their correlation coefficients. The first data set has an $r=0.996$, the second has an $r = -0.999$ and the third has an $r= -0.233$



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

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.

1- Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. `sklearn.preprocessing.MinMaxScaler` helps to implement normalization in python.

2- Standardization Scaling:

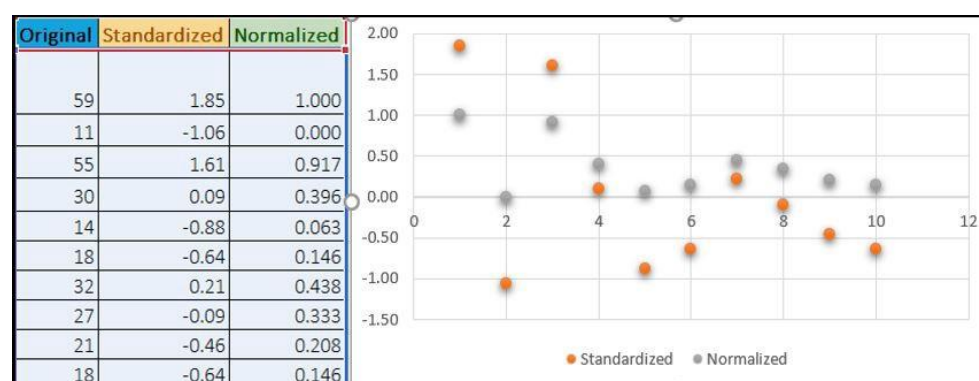
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

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

Example:

Below shows example of Standardized and Normalized scaling on original values.



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

The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to assess accurately the contribution of predictors to a model.

$$VIF = 1/(1-R^2)$$

If there is perfect correlation, then $VIF = \text{infinity}$. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that the standard error of this coefficient is inflated by a factor of 2. The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.

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

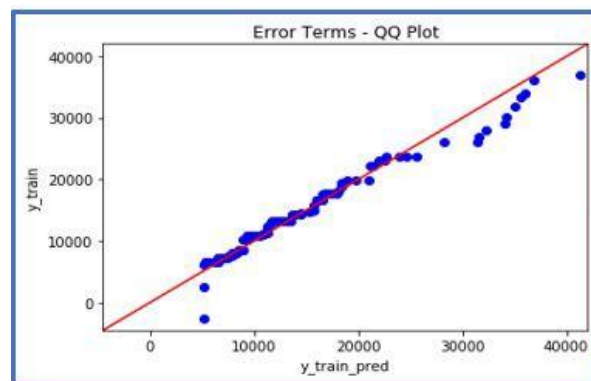
Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if 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 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.

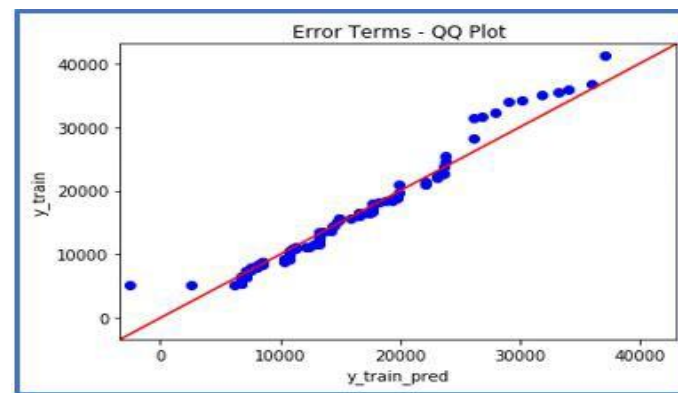
Below are the possible interpretations for two data sets.

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