LR Subjective questions

Sunday, April 10, 2022 8:44 PM

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)

Categorical values did not play as important role as numeric values in the data set
Of the categorical value," Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist" had negative correlation and Fall season had the positive correlation with dependent variable.

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

If we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.
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.

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

Temperature had the highest correlation

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

1. Error terms or The distribution of error or (values of Y at any X) is normal distribution, ie, ϵ has , mean= 0, SD= σ (ie, variance =σ2). This can be validated by plotting a distribution plot and checking the nature of distribution.
2. Error terms have constant variance ,should have same sigma (homoscedasticity) ie no patterns seen: This can be shown by scatter plot of the residuals.
3. There is a linear correlation between dependent and independent variables : This can be checked from the coefficients and their corresponding p-values being less than 0.05

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

- Temperature has hightest influnce on the bike demand
- Fall season has the next positive influence on the bike demand
- Windspeed has the next negative crrelation/influence on the bike demand

General Subjective Questions

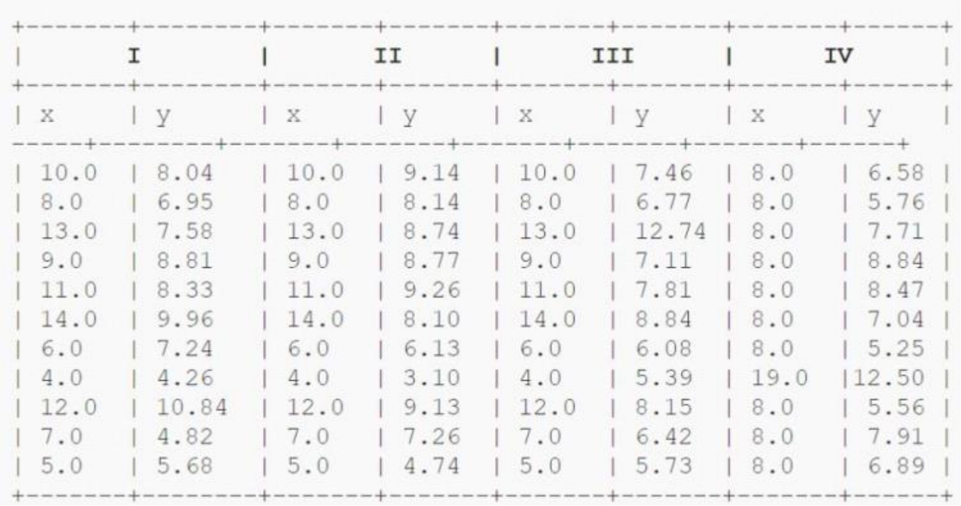
1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression algorithms are a category of Machine learning algorithms which are based on supervised learning.
The type of problems where linear regression can be applied are when there is one dependent variable and one or more independent variable(s) and as the independent variable(s) increases or decreases the dependent variable also increases or decreases by some factor.
In this case, the Linear Regression models predict the values of dependent variable based on independent variables.

When there is one independent variable, it is called Single linear regression model and when there are multiple independent variables it is called Multi Linear Regression.
The line formula ie, y=mx+c is used in this case where y = dependent variable, x= independent variable . m=slope and c=constant
The generalized formula for LR is :
y=B0 + B1X1 + B2X2 +B3X3 …… +BnXn
Where X1, X2,X3….Xn are independent factors
B0,B1,B2 …..Bn are the coefficients.

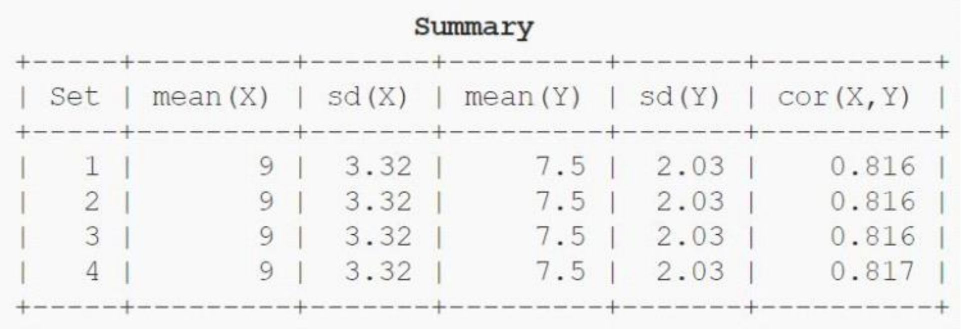
2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

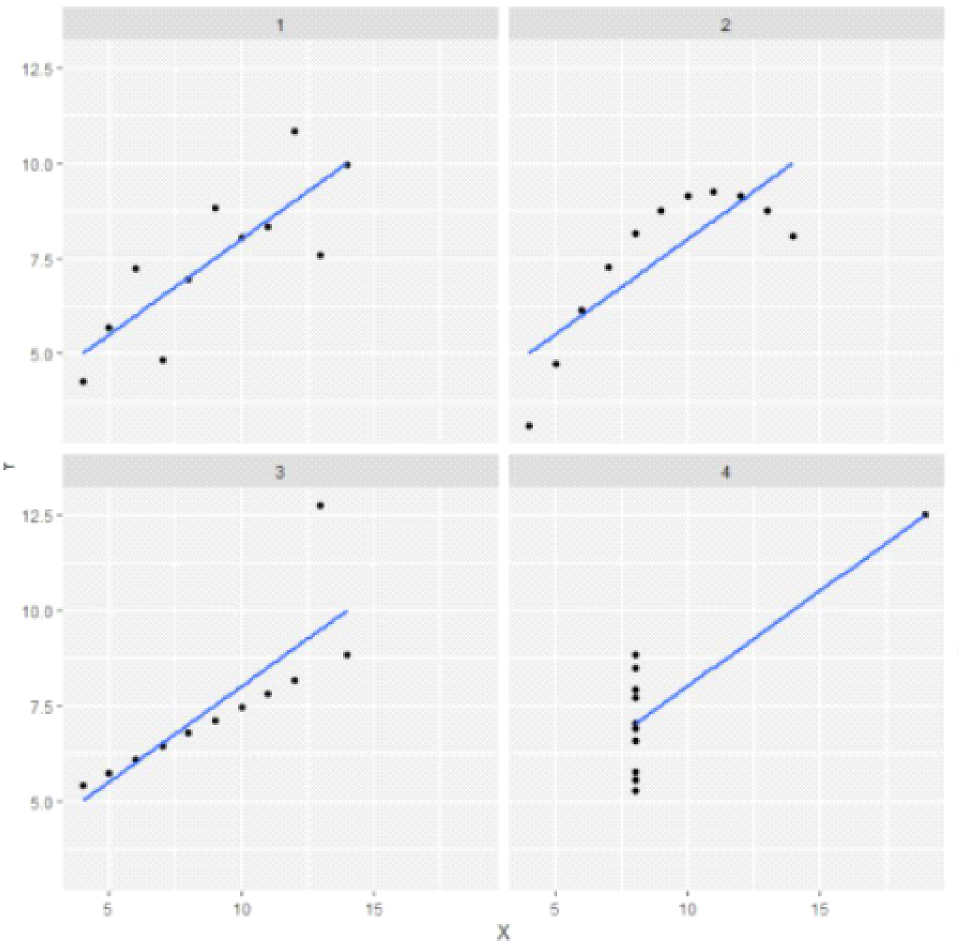
When statistics is applied, all four types of data produce the same mean , Standard Deviation and same correlation between X and Y



Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

However, when plotted they show totally different characteristics



1st Linear correlation exists
2nd - no linear correlation exists
3rd - Linear correlation exists initially but later the values drift apart
4th - one outlier can heavily impact the co-efficients
This shows the importance of visualizing the data before starting to analyze it.

3. What is Pearson's R? (3 marks)
The Person's R is used to find out the relationship or co-relation between two variables.
Hence this is used in Bivariate analysis. The Pearson's R is also called as the Pearson product-moment correlation coefficient (PPMCC) .
The value of Person's R can be between -1 to +1
Closer the value towards 1, higher is the correlation between two variables.
Values closer to 0 indicate no correlation between the two variables.
+ve value indicate that change is one variable, changes the other variable in same direction
-ve value indicate that change is one variable, changes the other variable in opposite direction

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

Scaling is a technique to make the data points closer to each other ie, it is used for making data points generalized so that the distance between them will be lower.
Scaling is important as most of the machine learning models learn from the data . If the distribution of the data points is large then this increase the uncertainty in the results of the model.
Machine learning algorithms like linear regression, logistic regression use Gradient descent algorithms to learn values. It is an iterative optimization algorithm for finding a local minimum of a different function. The gradient descent function slides through the data set while applied to the data set, step by step. So if the distance between the data points increases the size of the step will change and the movement of

the function will not be smooth.

Normalization is the method of scaling data where we try to fit all the data points between the range of 0 to 1 so that the data points can become closer to each other. Min Max technique is commonly used to normalize the data in 0 to 1 scale.

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

Standardization is a function is to make data points centered about the mean of all the data points presented in a feature with a unit standard deviation. This means the mean of the data point will be zero and the standard deviation will be 1.

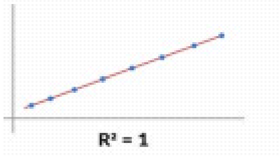
$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation }(x)}$$

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

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 or impossible to assess accurately the contribution of predictors to a model.

$$VIF = \frac{1}{1 - R^2}$$

Rsquared is high, then the variables are highly correlated.



When Rsquared is 1, the variables are perfectly correlated to each other
In this case VIF = 1/ (1-1) = 1/0 = Infinity.

Hence VIF can be infinity if the two predictors are perfectly correlated and the model cannot indentify which predictor has higher influence on the model

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

When the quantiles of two variables are plotted against each other, then the plot obtained is known as quantile – quantile plot or qqplot. This plot provides a summary of whether the distributions of two variables are similar or not.
QQ plots are useful in linear regression model to check if the points lie approximately on the line,ie to check the fit of the model. If the points are on the line, then the error distributions are normal . If they don’t fit closer to the line, then the error distributions, ie , the residuals aren’t Gaussian hence violating the assumption of Linear regression, indicating the model does not fit well.