- Pallavi Patil

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:

In the bike sharing dataset , The demand of bike is low in the month of spring when compared with other seasons. The demand bike boost in the year 2019 compared with year 2018.let's consider the effect of the categorical variable 'weathersit' on the target variable 'cnt'. Perform in EDA, I visualized the relationship between the categorical variables and the target variable As we know y-axis is a dependent variable while x-axis is independent variable so here dependent variable is 'cnt' and independent variable is 'weathersit'. It is analysed that during the weather situation

I analysed the following:-

1] Season- Fall is the high season where the number of bikes rented is high
2] Weather- The number of bikes rented is high when the weather is clear, few clouds
3] Weekdays-The number of bikes rented goes high during mid week.
4] Month-The number of bikes rented goes high during mid year.

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

Answer:

If you do not drop the 1st column then your dummy variables will be linked. During dummy value creation (dummy encoding) it is advisable to use drop_first=True, any other way we will get a redundant feature i.e. dummy variables might be correlated because the first column becomes a reference group during dummy encoding.

This may work on some models adversely and the effect is stronger when the cardinality is smaller.

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

Answer:

The numerical variable ‘registered’ has the highest correlated with the target variable ‘cnt’, if we consider all the features.

When we drop registered due to multicollinearity the numerical variable ‘atemp’ has the highest correlation with the target variable ‘cnt’ after data preparation.

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

Answer:

Pairing plot and checking correlation Pairwise correlated could be the first step to identify potential relation between various independent variables.

Rescaling the Features:-All the columns have small integer values. So it is extremely important to rescale the variables so that they have a comparable scale.

Absence of Multicollinerity:-

A more thorough method, however, would be to look 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 relation 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 multicollinearity [above Ten it can definitely be considered an issue].

Independence of residuals [absence of auto-correlation]:-

To verify that the viewing are not auto-correlated, we can using the Durbin-Watson test. The test will output values between zero and four. The closer it is to two, the less autocorrelation there is between the various variables

[0–2: positive auto-correlation]

[2–4: negative auto-correlation].

Check the P-value and VIF value in the model:-

If significant level is not given then we'll analyze 0.05 as standard significant level and if p-value increases above 0.05 then it becomes insignificant.

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

Answer:

Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are :

1. temp
2. yr
3. weathersit_Light Snow & Rain – coefficient

General Subjective Questions :-

1. Explain the linear regression algorithm in detail.

Answer:

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric value. Linear Regression is the more basic form of regression analysis. Regression is the mostly commonly used in predictive analysis model.

Linear regression is based on the popular equation:

"y = mx + c"

It presume that there is a linear relation between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relation between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal etc.

Regression method tries to find the best fit line which shows the relation between the dependent variable and predictors with least error. In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is divided into simple linear regression and multiple linear regression.

1. Simple Linear Regression : SLR is used when the dependent variable is predicted using only one independent variable.

2. Multiple Linear Regression :MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

$\beta 1$ = coefficient for X1 variable

$\beta 2$ = coefficient for X2 variable

$\beta 3$ = coefficient for X3 variable and so on…

$\beta 0$ is the intercept (constant term).

Equation ::: yi = B0 + B1xi1 +B2xi2 + ... Bpxip + Ei for i = 1,2, ... n.

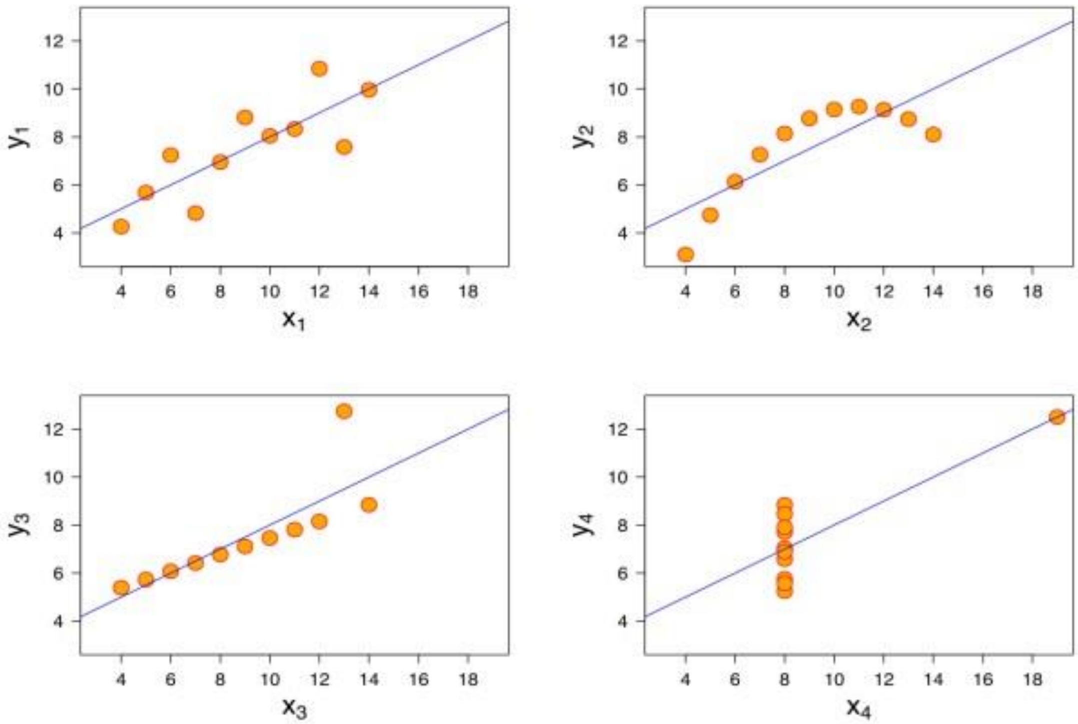
2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet is a very good demo of the importance of graphing data to analyze it. simply variance means, values, and even linear regressions can't accurately portray data in its native form.

Anscombe's Quartet multiple data sets shows that with many similar statistical properties can still be vastly different from one another when graphed.

Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets.



Think about it: if the bottom two graphs didn't have that one point that strayed so far from all the other points, their statistical properties would no longer be identical to the two top graphs.

In fact, their statistical properties would mostly accurately resemble the lines that the graphs seem to depict. How to analyze your data. For example, while all four data sets have the same linear regression, it is obvious that the top right graph really shouldn't be analyzed with a linear regression at all because it's a curvature.

Conversely, the top left graph probably should be analyzed with a linear regression because it's a scatter plot that moves in a roughly linear manner. These observations demonstrate the value in graphing your data before analyzing it.

- The first scatter plot (top left) appears to be a simple linear relationship.

- The second graph (top right) is not distributed normally; while there is a relation between them ,it’s not linear.

3. What is Pearson’s R?

Answer:

Pearson's r is a numerical compendium of the strength of the linear association between the variables. It value ranges between -1 to +1. It shows the linear relationship between two sets of data.

In simple terms, it tells us can we draw a line graph to represent the data? r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a negative slope r = 0 means there is no linear association.

For Example:

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. The opposite is true with regards to price.

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

Answer:

It is performed during the data pre processing stage to deal with varying values in the dataset. Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the low values, irrespective of the units of the values.

- Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks. 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/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.

{ MinMax scaling:x=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).

- Standardisation:

x=x-mean(x)/std 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.

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

Answer:

Full form of VIF → VIF - the variance inflation factor The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.

(VIF) =1/(1-R_1^2)

If there is totally perfect correlation, then VIF = infinity. Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1.
So, VIF = 1/(1-1)

which gives VIF = 1/0 which results in "infinity".

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

Answer:

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.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

- Pallavi Patil