Assignment Based Subjective Question

Q-1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer

- Demand is high in clear and mist weather conditions. Highest demand is in clear weather condition.
- Demand is high on working days.
- Demand is high in spring, summer and fall season. Highest demand is in fall season.
- Demand is high in non holidays.
- Demand is hight from april to October.
- Demand is almost same on every day of the week.
- Demand is overall increases from 2018 to 2019.

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

Answer

- It is used to remove the redundency or colinearity in the dataset.

- When you have a categorical variable with say 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels.

- For a variable say, 'Relationship' with three levels namely, 'Single', 'In a relationship', and 'Married'. No need to define three different levels. 2 will be sufficient. e.g. 'In a relationship' and 'Married' can together define the Relationship wheather it is Single or not. When both the values are zero. we can say that Relationship is Single.

- For a variable say, 'Gender' with levels 'Male' and 'Female'. No need to define both levels. One level will be sufficient. e.g. When Male is zero that means Gender is Female.

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

Answer

By looking at the pairplot among numeric variable, It's clear that **atemp** is positively correlated with target variable **cnt**.

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

Answer

- Model will fit a hyperplane e.g. Y = a + b X + c Z + d K + E
- Coefficient have been calculated using OLS method from statsmodel api. i.e. by minimising the sum of squared error.
- Residual analysis have been performed. It's clearly zero-mean, independent and normally distributed error terms with constant variance (homoscedasticity).
- Multicollinearity and Feature selection have been performed using VIF and RFE.

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

Answer

Top three features contributing significantly towards explaining the demand of shared bikes

- **atemp**: feeling temperature in Celsius with coefficient **0.57**
- **weathersit_Light**: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds with coefficient **-0.25**
- **yr**: Year with coefficient **0.23**

General Subjective Questions

Q-1 Explain the linear regression algorithm in detail.

Answer

Linear regression is an algorithm that provides a linear relationship between an independent variables and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

The independent variable is also known as the predictor variable. And the dependent variables are also known as the output variables.

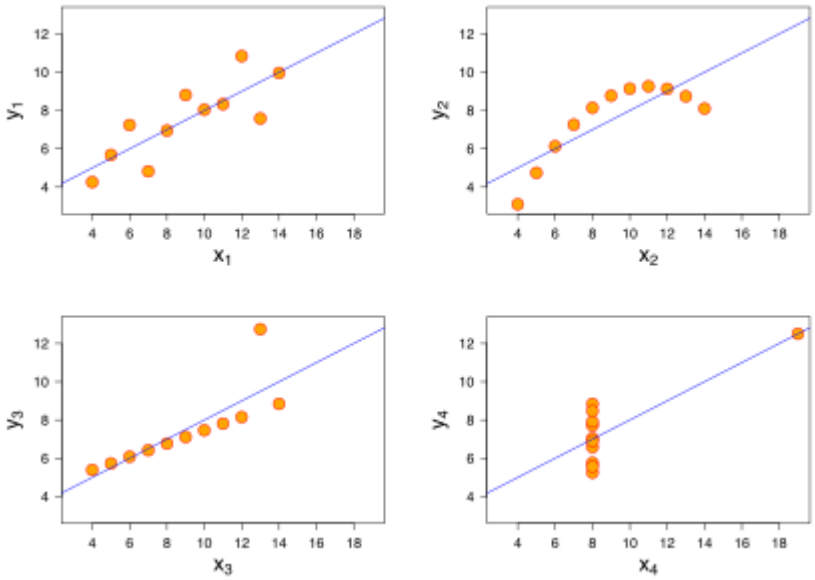
Following steps are involved in linear regression algorithm:-

1. Find residuals and RSS for any given line passing through the scatter plot.
2. Find the equation of the best-fit line by minimising the RSS and found the optimal value of coefficients. **Gradient Descent optimisation algorithm** is used to obtain the best fit line.

Q-2 Explain the Anscombe’s quartet in detail.

Answer

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and 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 when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Q-3 What is Pearson’s R?

Answer

In statistics, the Pearson correlation coefficient (PCC) — also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient — is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a

linear correlation of variables, and ignores many other types of relationships or correlations. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

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

Answer

Scaling is a technique to standardize the independent features present in the data in a fixed range. It helps in

- Ease of Interpretation
- Faster convergence of Gradient Descent Method.

There are two major methods to scale the variables:

- **standardisation**: Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one.

$$x' = \frac{x - \bar{x}}{\sigma}$$

- **MinMax scaling**: brings all of the data in the range of 0 and 1.

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

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

Answer

VIF = 1 / (1 - R-Sq)

and

R-Sq = 1 - (RSS/TSS)

VIF can be infinite only when value of R-Sq is 1 and R-Sq can be 1 only when RSS is 0. i.e. resedual sum of squares is zero. All the data point lies perfectly on the staright line.

Higher the value of VIF, means multicollinearity is very high. We can get rid of this variable from the model.

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

Answer

In statistics, a Q–Q plot (quantile-quantile plot) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). This defines a parametric curve where the parameter is the index of the quantile interval.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the identity line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q–Q plots can be used to compare collections of data, or theoretical distributions. The use of Q–Q plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions. A Q–Q plot is generally more diagnostic than comparing the samples' histograms, but is less widely known. Q–Q plots are commonly used to compare a data set to a theoretical model. This can provide an assessment of goodness of fit that is graphical, rather than reducing to a numerical summary statistic. Q–Q plots are also used to compare two theoretical distributions to each other. Since Q–Q plots compare distributions, there is no need for the values to be observed as pairs, as in a scatter plot, or even for the numbers of values in the two groups being compared to be equal.