

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

Seasons and Months are the categorical variables and we could infer that Bikes has got less demand in spring season amongst other seasons
Bikes demand increased in 2019 over 2018

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

This option will drop the first column that has been created to uniquely identify the variables. If there are k-levels, we just need k-1 dummy variables to uniquely identify value.
It also reduces the co-relation amongst the variables.

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

Registered variable has got highest co-relation with target variable cnt

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

Calculate the VIF (Variance Inflation Factor). $VIF \leq 5$ implies no multicollinearity, whereas $VIF > 5$ implies serious multicollinearity.

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

Bikes are rented less on holiday and there is loss during snow season. Bikes are rented less in September and more in Jan, so its dependent on mainly Season, Month, year

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 getting 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.

Explain the Anscombe's quartet in detail.

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.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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.

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

It 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.

Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1.

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).

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

If there is perfect correlation, then $VIF = \infty$. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which leads to $1/(1-R^2)$ 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).

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.

It can be used with sample sizes also. 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.