

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:

From data file we found that:

- The demand of bike is less in the month of spring when compared with other seasons
- The demand bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with working day and non working day.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Lightsnow and light rainfall. We do not have any data for Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog, so we can not derive any conclusion. Maybe the company is not operating on those days or there is no demand of bike.

After building the model we found that:

- Demand for the BoomBikes company is dependent on whether it is a working day or not.
- Additionally more rentals seem to be demanded in the winters as compared to the summer and spring.
- We had observed that the months of September and October had higher use of rentals.
- In terms of days the maximum focus was on days like Wed, Thurs and Sat and more on holidays.

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

Answer:

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 i.e. avoids redundancy of any kind.

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

Answer:

atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

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

Answer:

According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship:

- This was validated by plotting a scatter plot between the features and the target.
- Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable.
- The error distribution was observed and it was a normal distribution.

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

Answer:

The top 3 features contributing significantly towards the demand of the shared bikes are the temperature, the year and the holiday variables.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear Regression is an ML algorithm used for supervised learning. It helps in predicting a dependent variable(target) based on the given independent variable(s). The regression technique tends to establish a linear relationship between a dependent variable and the other given independent variables.

There are two types of linear regression- simple linear regression and multiple linear regression.

Simple linear regression is used when a single independent variable is used to predict the value of the target variable.

Multiple Linear Regression is when multiple independent variables are used to predict the numerical value of the target variable. A linear line showing the relationship between the dependent and independent variables is called a regression line.

A positive linear relationship is when the dependent variable on the Y-axis along with the independent variable in the X-axis. However, if dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship.

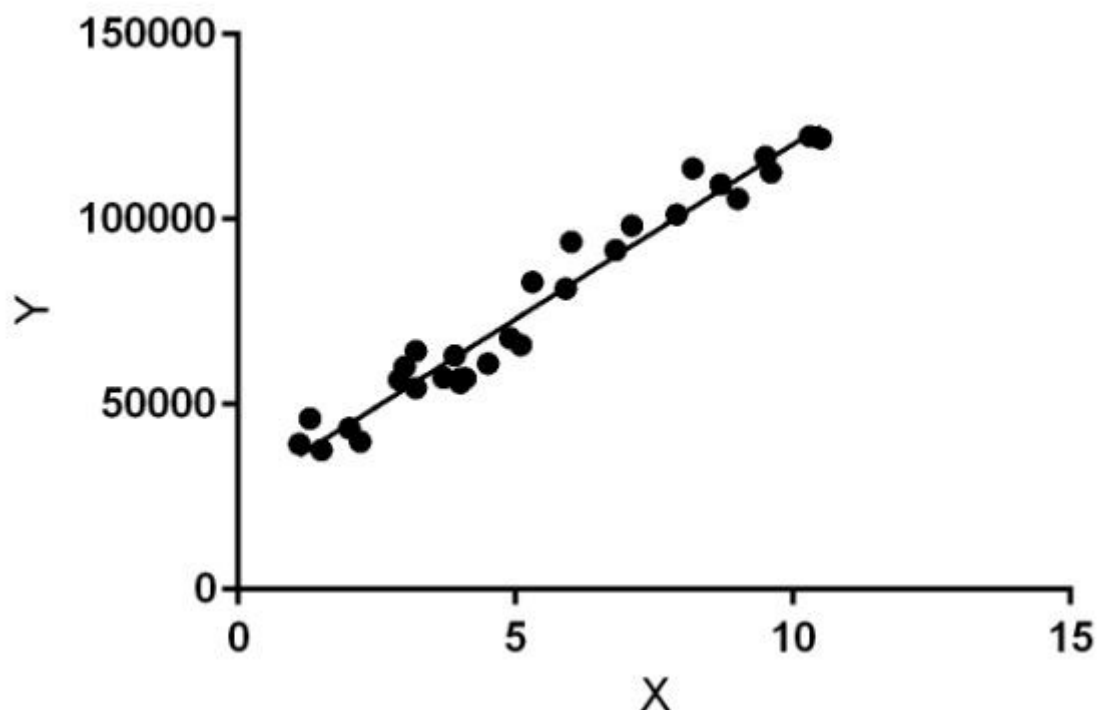


Diagram description: Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Example:

While training the model we are given : x: input training data (univariate – one input variable(parameter)) y: labels to data (Supervised learning) When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values. θ_1 : intercept θ_2 : coefficient of x Once we find the best θ_1 and θ_2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each dataset consists of eleven points.

The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give the an accurate representation of two datasets being compared.

3. What is Pearson's R?

Answer:

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).

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

Answer:

Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range.

The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.

The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1 , standardization replaces the values with their Z scores.

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

Answer:

If there is perfect correlation, then $VIF = \text{infinity}$. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which lead 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.

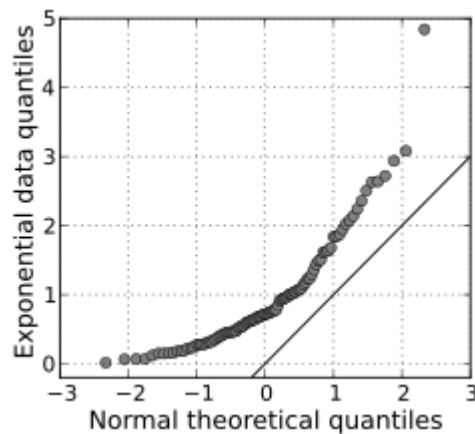
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

For example here one of my variable dteday variable was forgot to drop and the model showed r^2 as 1, the variable was having high correlation with some variables, so i dropped it.

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

Answer:

The quantile-quantile (Q-Q) plot are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.



A Q Q plot showing the 45 degree reference line:

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.