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

There is definitely a significant increase from year 0 (2018) to year 1 (2019) and a pattern of a marked increase in the use of shared bikes between the months of March to November.

Additionally, the spring season sees a marked reduction in the number of shared bike rentals as compared to any other season.

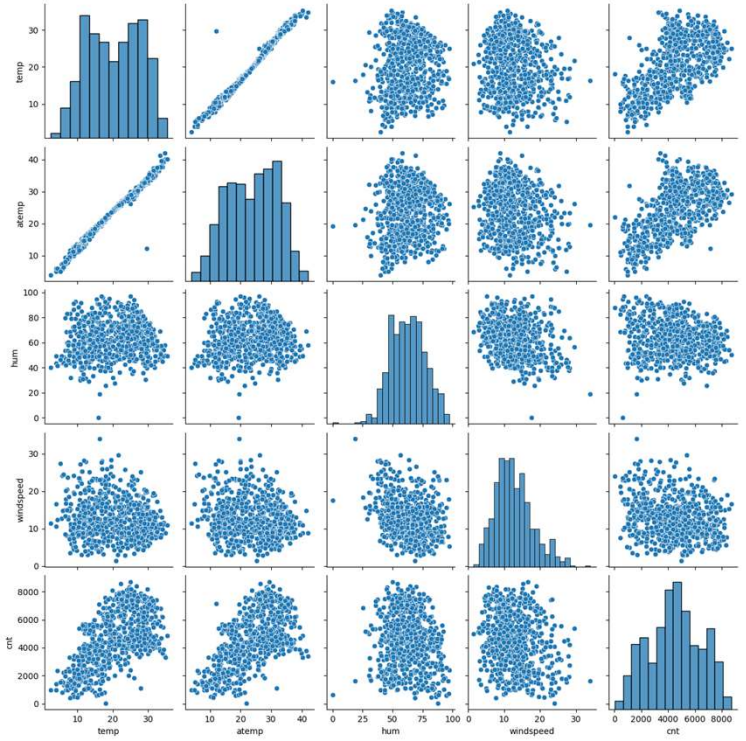
People tend to rent bikes more on holidays than on non-holidays

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

This is important since this will firstly, reduce the overall number of columns being used for analysis and, secondly, remove redundancy since one of the categorical values can be determined by setting the value of all the remaining categories to 0.

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

Upon inspecting the pairplot, it appears that temp and atemp have the highest correlation with the target variable among the numerical variables. These two are also extremely highly correlated with each other.

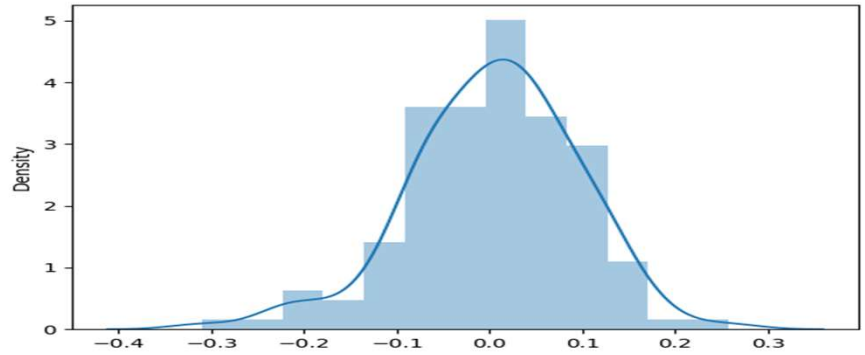


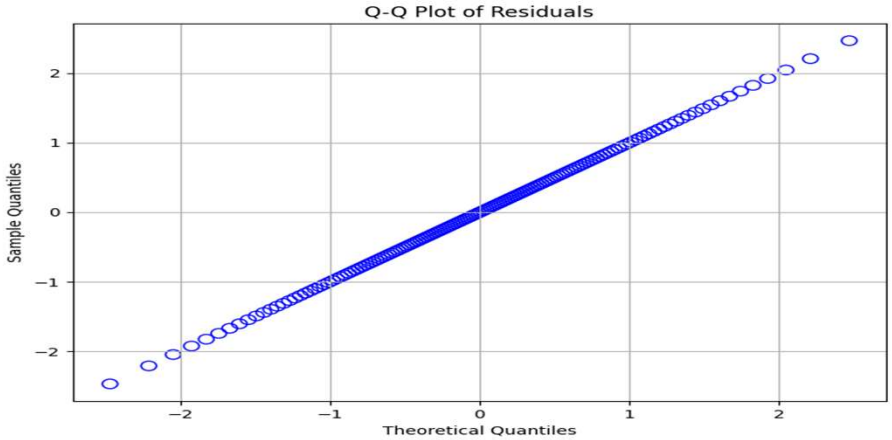
How did you validate the assumptions of Linear Regression after building the model on the training set?

To assess the distribution of the residuals, a distplot was used. They were found to be normally distributed.

A Q-Q plot was also used to test the same.

Both indicated that the residuals were distributed normally.





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

1. Temperature
2. Humidity
3. Year
4. Special mention – Constant which captured the effect of the missing dummy columns

Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm which is used to predict continuous, numerical values based on a set of input features.

At its most basic level, it creates a best-fit line between two variables. The equation of this line will be in the form of $y = b_0 + b_1x$. Here, y is the target variable (the one we are attempting to predict) and x is the input variable. For any given value of x, the algorithm will predict a y-value based on this formula.

The goal is to try and compute the optimal b_1 so as to minimize the squared difference between the actual and predicted values of y (known as residuals).

In a multiple regression scenario, the equation will be of the form $y = b_0+x_1*b_1+x_2*b_2+\ldots+x_n*bn$, with $x_1,x_2,x_3,\ldots x_n$ being the input variables and b1,b2,b3,…bn being the coefficients of each.

The assumptions of linear regression are as follows:

1. That a linear relationship exists between the dependent and independent variables.

2. That the variables are independent of each other. There is no relationship between the residuals.

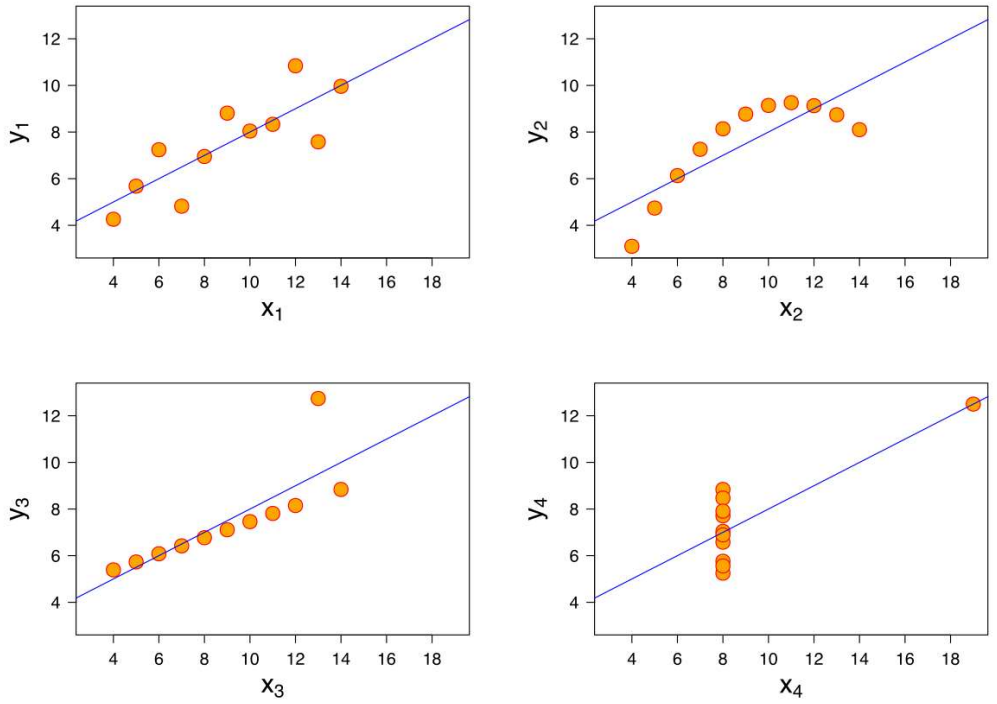
3. That the variance of the residuals is uniform across all the independent variables at all levels.

4. That the residuals are normally distributed around a mean of 0.

5. That there is no multicollinearity between the independent variables.

Explain the Anscombe's quartet in detail.

As shown in the adjoining image, Anscombe's quartet refers to 4 datasets that share similar properties (mean, variance) but have very different patterns when graphed.
The first (x1,y1) is a roughly linear relationship
The second (x2,y2) is a curvilinear relationship which would not be appropriate for linear regression
The third (x3,y3) is a near perfectly correlation with one outlier
The fourth (x4,y4) contains an outlier which would significantly affect the regression line.
The purpose of this quartet is to explain the importance of handling outliers



Original image source:https://upload.wikimedia.org/wikipedia/commons/thumb/e/ec/Anscombe%27s_quartet_3.svg/1200px-Anscombe%27s_quartet_3.svg.png

What is Pearson's R?

Pearson's R or Pearson's correlation coefficient is a measure which ranges between -1 and +1 and is used to determine the strength of the linear relationship between two variables X and Y.

The formula is as follows:

R = covariance(X,Y)/(Standard deviation(X)*Standard deviation(Y))

Where covariance is calculated as the product of the summations of (X-Mean(X)) and (Y-Mean(Y)) divided by the number of observations – 1.

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

Scaling is the process of transforming numerical variables into a fixed range or specific type of distribution. It is done in order to reduce the influence of variables in a model which have a significantly larger range of values relative to others. For example; age is usually in the range of 0-100 years whereas height in cm may go beyond even 200.

Normalization or min-max scaling is done based on the following formula: $x\_scaled = (X-X_{min})/(X_{max}-X_{min})$. Here all the values are brought to within the range of 0 to 1 based on the minimum and maximum values among the observations.

Standard scaling is done based on the mean and standard deviation of the values in question. This scaling preserves the shape of the distribution and uses the formula x_scaled = (X-Mean(X))/Standard deviation(X)

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

This refers to a situation called perfect multicollinearity. This arises when one variable can be perfectly predicted by a combination of other variables. This can happen when variables are v=created using linear combinations of variables or due to dummy variables being used, the values of which can perfectly predict other dummy variables.

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

A Q-Q or Quantile-Quantile plot is a graph which is used to assess whether there is a common distribution between two datasets. This can be extremely useful in a linear regression to determine whether the residuals are normally distributed. This can be used to test the goodness of fit of multiple models.