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

Here are some inferences the drawn from the categorical variables

1. Fall has the highest median,which is expected as weather conditions are most optimal to ride bike followed by summer
2. Median bike rents are increasing year on as year 2019 has a higher median then 2018,might be they are getting noticed by more people
3. Spring season has higher median and the months falling in spring has high median than the others with sept has top in rentals
4. Overall median across all days is same could not able to draw much insights from it
5. Workinday median is higher than the holiday which indicates there are more bookings on working day than holiday

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

A variable with n levels can be represented by n-1 dummy variables.Drop first removes the first column so we are remained with n-1 values

Example if a status has three values married,divorced,single

X1	X2	Value
0	0	single
0	1	married
1	0	divorced

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

From the pair plot temp and atemp has a correlation 0=of 0.63 with cnt variable

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

By plotting the residual distribution >it come out to be a normal distribution with mean 0

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

- **Temperature(temp)**:A coefficient value of ‘0.5373’ indicated that a unit increase in temp variable increases the bike hire numbers by '0.5373' units
- **Weather situation 3(weathersit3)**:A coefficient value of ‘-0.2788’ indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.2788 units.
- **year(yr)**:A coefficient value of ‘0.2297’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units

Explain the linear regression algorithm in detail

A linear regression algorithm tries to explain the relationship between the independent variable and dependent variable by using straight line .It is applicable to numerical variables only

Following are the steps performed while doing linear regression

1.The data is divided into test and train

2.Trained data is divided to dependent and independent features

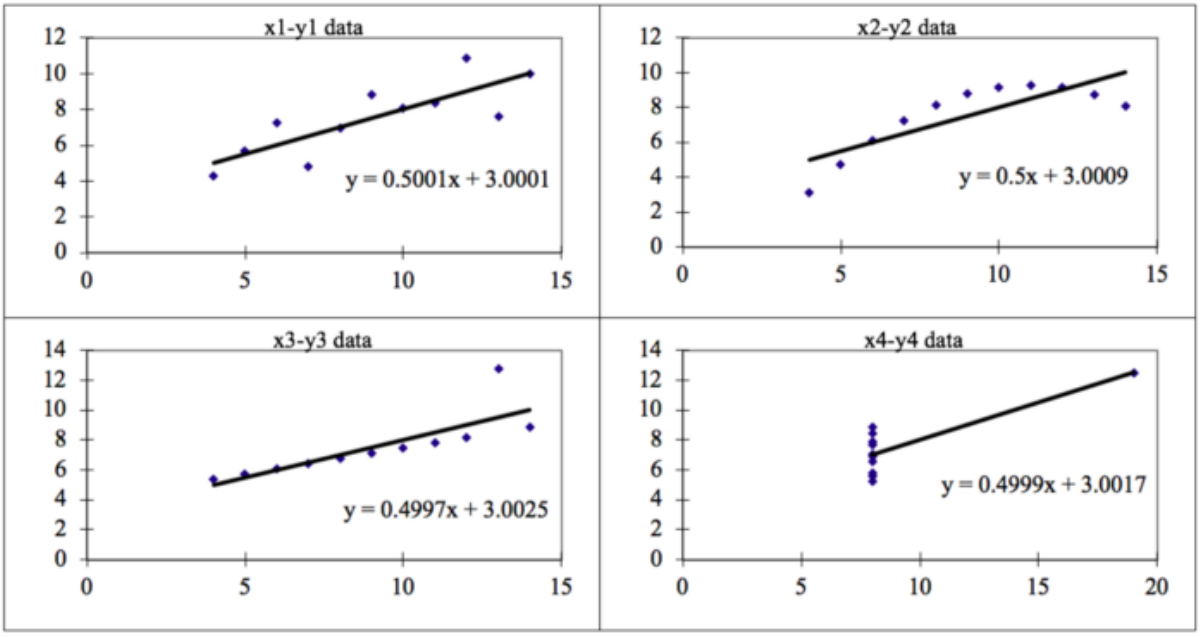
3. A linear model is fitted using the training dataset.Gradient descent algorithm can be used to find the best fit coefficients

4.Incase of multiple variables ,the predicted variable is a hyperplane instead of line.

5.The predicted variable is compared with the test data and assumptions are checked

Explain the Anscombe's quartet in detail.

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.



The four datasets can be described as:

1. Dataset 1: this fits the linear regression model pretty well.
2. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
3. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

4. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

What is Pearson's R?
Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r

Assumptions

1. For the Pearson r correlation, both variables should be normally distributed. i.e the normal distribution describes how the values of a variable are distributed.

2. There should be no significant outliers. We all know what outliers are but we don't know the effect of outliers on Pearson's correlation coefficient, r.Pearson's correlation coefficient, r, is very sensitive to outliers

3. Each variable should be continuous i.e. interval or ratios for example weight, time, height, age etc. If one or both of the variables are ordinal in measurement, then a Spearman correlation could be conducted instead.

4. The two variables have a linear relationship. Scatter plots will help you tell whether the variables have a linear relationship. If the data points have a straight line (and not a curve), then the data satisfies the linearity assumption. If the data you have is not linearly related you might have to run a non-parametric

5. The observations are paired observations. That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable.If there are 12 observations of weight, you should have 12 observations of age. i.e. no blanks.

6.Homoscedastic :Homoscedasticity describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables. A scatter-plot makes it easy to

check for this. If the points lie equally on both sides of the line of best fit, then the data is homoscedastic

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

machine learning algorithms, to bring all features in the same standing, we need to do scaling so that one significant number doesn't impact the model just because of their large magnitude.scaling is where all the values of features are in between o and 1 or [-1,1]

Why do we need

The machine learning algorithm works on numbers and does not know what that number represents. A weight of 10 grams and a price of 10 dollars represents completely two different things — which is a no brainer for humans, but for a model as a feature, it treats both as same.Thus feature scaling is needed to bring every feature in the same footing without any upfront importance

Difference between standard and normalizer scaler

StandardScaler() standardizes features (such as the features of the person data i.e height, weight)by removing the mean and scaling to unit variance.

Normalizer() rescales each sample. For example rescaling each company's stock price independently of the other.

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

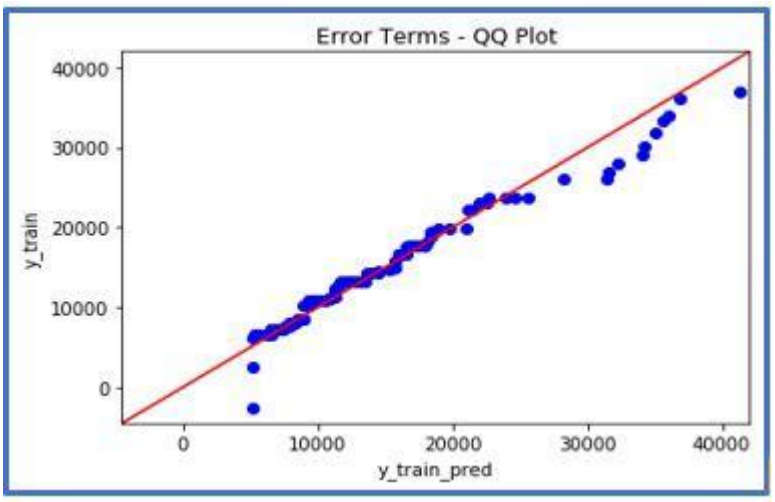
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

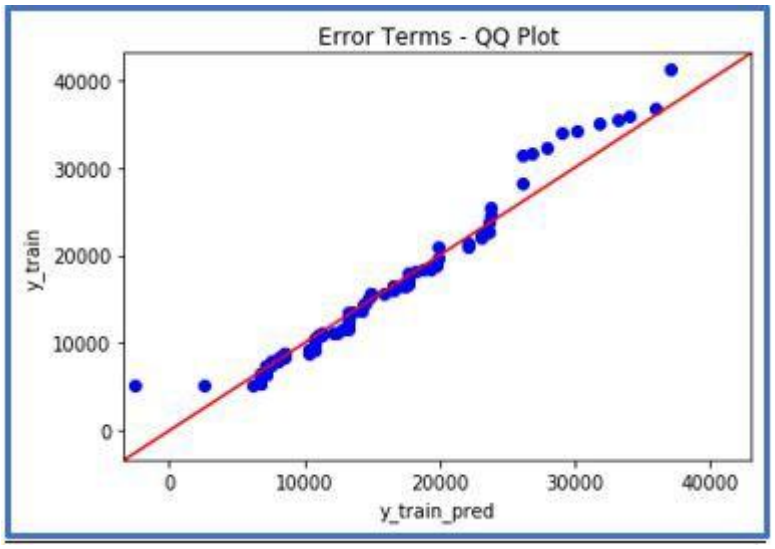
Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:

Y values < x-values :If y-quantiles are lower than the x-quantiles



X-values < y-values:If x-quantile are lower than the y-quantities



if the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the 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.