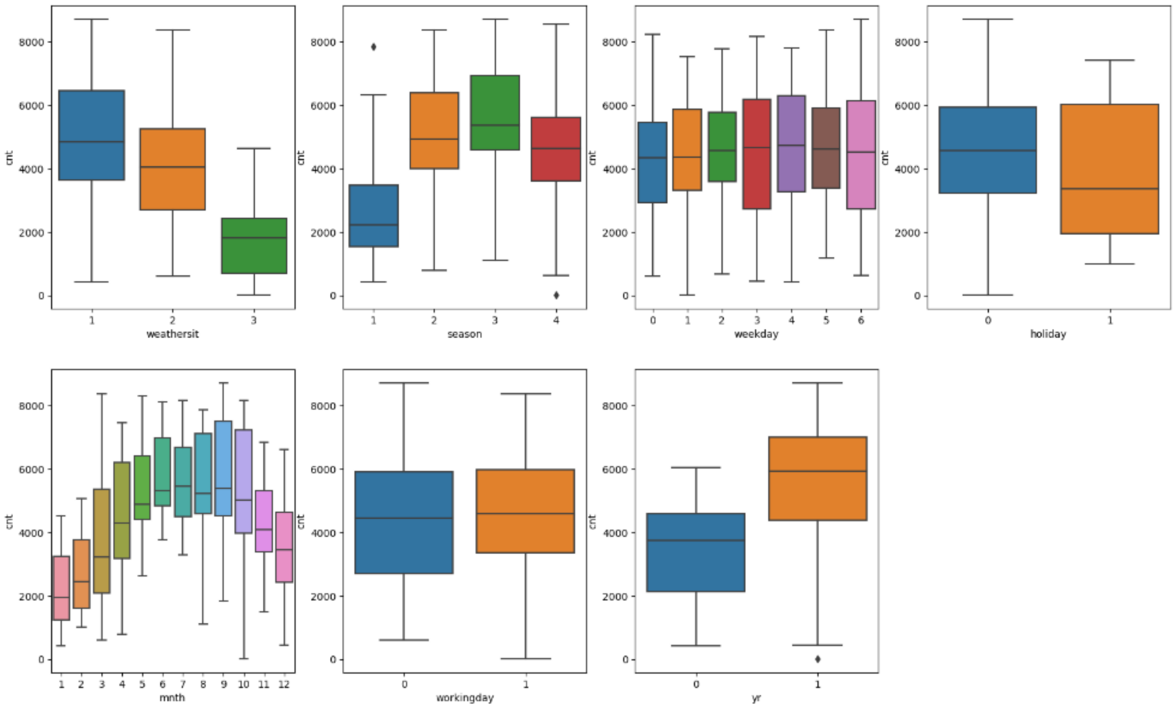
1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer..



season: In season 2 and season 3 with a median of over 5000 booking. This shows, season can be a good predictor variable for the dependent variable.

mnth: in the months 5,6,7,8,9 & 10 with a median of over 4000 booking per month. This shows, mnth has some information for bookings and can be a good predictor for the dependent variable.

weathersit: for value weathersit1 the median value of number of bikes booked in near to 5000 and weathersit2 mean is above 4000 this shows it can be good predictor variable.

holiday: bike booking were happening more when it is not a holiday. This indicates, holiday can't be a good predictor for the dependent variable for increasing bookings.

weekday: has almost same mean so no influence towards the predictor.

workingday: median of close to 5000 booking. This indicates, workingday can be a good predictor for the dependent variable

yr: 2019 shows higher and better year the bike booking happened. This indicates it could be useful.

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

Its important drop the first as once we convert the categorical variable to dummy variable then we can predict the values with one less dummy variable.

For example lets if we have 3 categories where data is distributed. Then with 00,01,10 with this only we can predict third is true when first two dummy variable value are 00. Its important to drop.

if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

Cnt-atemp and cnt-temp has highest correlation with target variable

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

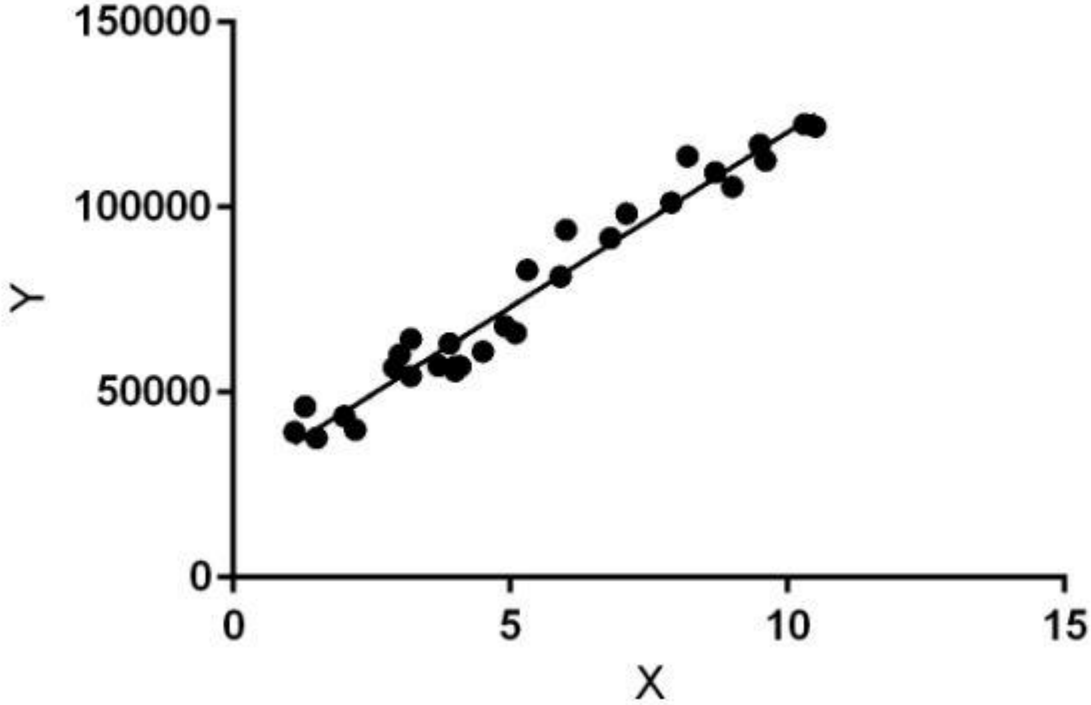
Validated by looking residuals are normally distributed or not after building the model for linear regression

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

Temperature (temp), yr, mnth_9

1. 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. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.



Linear regression is used in many different fields, including finance, economics, and psychology, to understand and predict the behavior of a particular variable. For example, in finance, linear regression might be used to understand the relationship between a company's stock price and its earnings, or to predict the future value of a currency based on its past performance.

Y=mx+C

Where c is constant where exactly the regression line corsses y axis and m is slope.

2.

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.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

3.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.

3.

Feature scaling is one of the most important data preprocessing step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled.

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

X_new = (X - mean)/Std

Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Geometrically speaking, it translates the data to the mean vector of original data to the origin and squishes or expands the points if std is 1 respectively. We can see that we are just changing mean and standard deviation to a standard normal distribution which is still normal thus the shape of the distribution is not affected.

4. This shows perfect correlation between two independent variables and VIF becomes infinity

In perfect correlation we get R2 =1 which lead to 1/(1-R2) to infinity to solve this need to drop one of the variable.

5. Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc.

QQ plot used to check whether the data is normally distributed or not

2.
3.