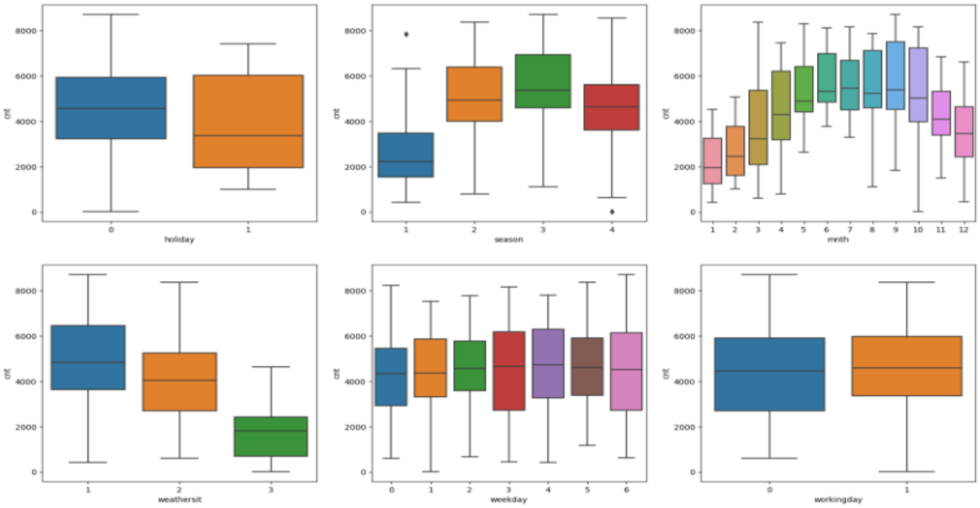
Question: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



For Categorical Variables

Holiday

- Median of the holiday is higher than non holiday days
- 75 percentile for holiday & non holiday is equal
- Minimum number of ridership is more for holiday than non holiday days
- Maximum number of ridership is less for holiday than non holiday days

Session:

- The highest ridership is in season fall.
- The lowest ridership is in the season spring.

Month:

- The highest ridership is in month 9.
- The lowest ridership is in month 1.
- The highest median is in month 7.

Weathersit:

- There is no ridership for weathersit 4.
- There is highest ridership when weather is clear.
- Depending upon weather ridership goes up and down.

Working day:

- No Major deference between the median of working & non working days
- 75 percentile for both type of days are equal
- Minimum number of ridership is more for non-working days than working days.
- Maximum number of ridership is less for non-working days than working days.

✧ The categorical variable workingday and holiday representing the very similar data.
✧ The categorical variable season and mnth representing the very similar data.

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

bike_df['season'] = bike_df['season'].map({1: 'spring', 2:'summer', 3:'fall', 4:'winter'})
season = pd.get_dummies(bike_df['season'],dtype='int', drop_first=True)
This will create 4 variables: spring, summer, fall, winter

spring	summer	fall	winter	Comment
1	0	0	0	This combination represent spring
0	1	0	0	This combination represent summer
0	0	1	0	This combination represent fall
0	0	0	1	This combination represent winter

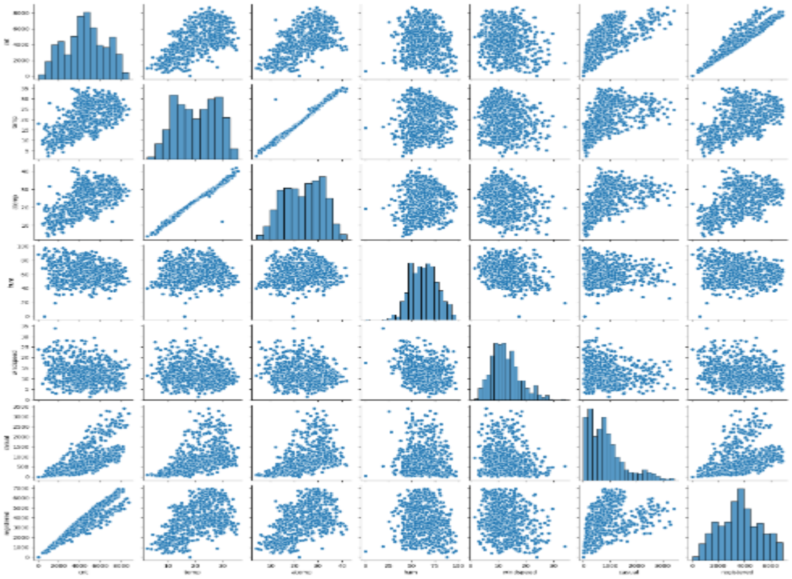
Now if I drop variable spring, even we can identify spring from the following combination
season = pd.get_dummies (bike_df['season'],dtype='int', drop_first=True)

summer	fall	winter	Comment
0	0	0	This combination represent spring
1	0	0	This combination represent summer
0	1	0	This combination represent fall
0	0	1	This combination represent winter

This way, removing the first variable helps to reduce the number of variables without losing any information.

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

- The highest correlation between cnt & temp , cnt & atemp .
- Temp & atemp have the highest correlation



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

Answer:

- The model should give very similar R square to the unseen test data set as it giving to the training data set.
- Distribution plot of the residual should be near to standard distribution.

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

Answer

Following are the top 3 features of the model is

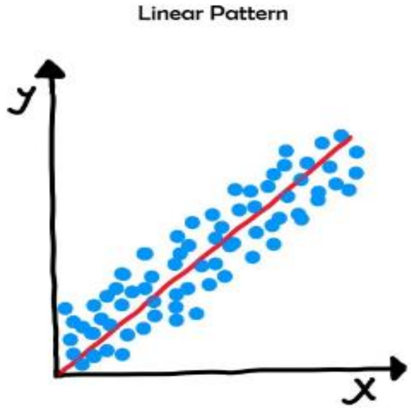
1. atemp
2. yr
3. Holiday

Question: Explain the linear regression algorithm in detail.

Answer

Linear regression algorithm: it tells linear relationship between one or more independent(x) variable with one dependent (y) variable. Therefore it is called linear regression. Since linear regression shows the linear relationship

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



For single Linear regression: y=mx+c

For multiple linear regression: y =m1x1+ m2x2+m3x3….+c

y= Dependent Variable (Target Variable)

x= Independent Variable (predictor Variable)

c= intercept of the line (Gives an additional degree of freedom)

m= Linear regression coefficient (scale factor to each input value).

For example if sales of any product have a linear relation with marketing then

Sales = coefficient * Marketing + intercept

Question: . Explain the Anscombe's quartet in detail

Answer

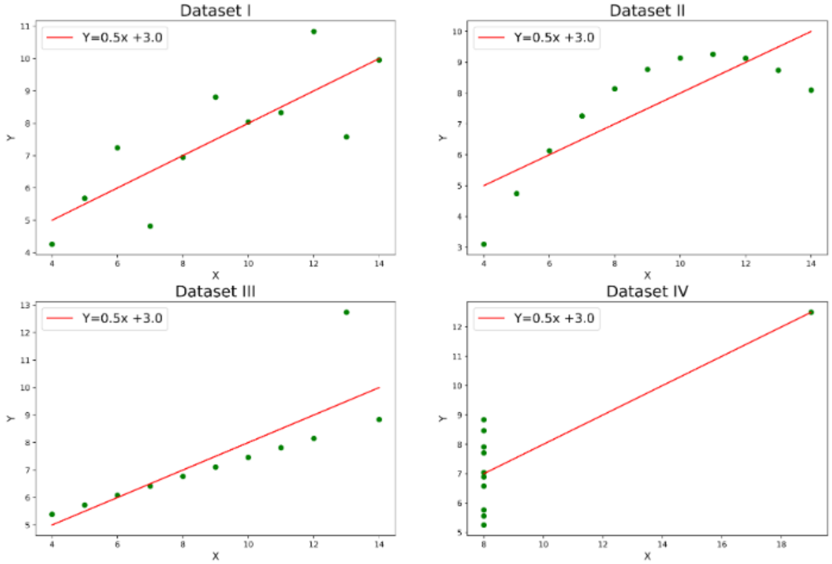
Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Following dataset have the same summary statistics also have very similar line too but have totally different data visualization to spot trends, outliers, and other crucial details.



Question: What is Pearson's R?

Answer

Pearson's r or Pearson correlation coefficient or Pearson's correlation coefficient is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.

In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

For example: Up till a certain age (in most cases), a child’s height will keep increasing as his/her age increases. Of course, his/her growth depends upon various factors like genes, location, diet, lifestyle, etc.
This approach is based on covariance and, thus, is the best method to measure the relationship between two variables.
The Pearson coefficient correlation has a high statistical significance. It looks at the relationship between two variables. It seeks to draw a line through the data of two variables to show their relationship. The relationship of the variables is measured with the help Pearson correlation coefficient calculator. This linear relationship can be positive or negative.
Pearson correlation coefficient formula:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

Question: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer:
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.

Data Normalization
One of the most popular methods for preparing data is Normalization, which enables us to alter the values of numerical columns in the dataset to a standard scale.
Normalization is the method used to arrange the data in a database. It is a scaling method that reduces duplication in which the numbers are scaled and moved between 0 and 1. When there are no outliers since it can't handle them, normalization is employed to remove the undesirable characteristics from the dataset.
One technique to process data to produce easily comparable findings within and across several data sets is the normalization procedure. Anyone reading data can benefit from it, but those using machine learning and significant amounts of data may find it most regularly helpful. Understanding the normalization formula will help you decide if it is the best way to handle your data set.

Data Standardization Standardization, often referred to as z-score Normalization, occasionally is a method for rescaling the values that meet the characteristics of the standard normal distribution while being similar to normalizing.

Standardization is crucial because it enables reliable data transmission across various systems. It would be easier for computers to exchange data and communicate with one another with standardization. Additionally, standardization makes it simpler to process, analyze, and store data in a database. Businesses can use their data to make better judgments with this method. Companies can more readily compare and evaluate data when standardized, allowing them to gain insights into how to run their businesses better.

When the data is distributed Gaussianly, standardization can be helpful. But it's okay for this to be the case. Standardization also lacks a bounding range, in contrast to normalizing. Therefore, normalization will have no effect on any outliers you may have in your data

Normalization	**Standardization**
This method scales the model using minimum and maximum values.	This method scales the model using the mean and standard deviation.
When features are on various scales, it is functional.	When a variable's mean and standard deviation are both set to 0, it is beneficial.
Values on the scale fall between [0, 1] and [-1, 1].	Values on a scale are not constrained to a particular range.
Additionally known as scaling normalization.	This process is called Z-score normalization.
When the feature distribution is unclear, it is helpful.	When the feature distribution is consistent, it is helpful.

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

Answer

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.

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

Answer:

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.

Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variable(s)

In summary, A Q-Q plot helps you compare the sample distribution of the variable at hand against any other possible distributions graphically

Q-Q plots can help identify outliers by revealing data points that fall far from the expected pattern of the distribution. Outliers may appear as points that deviate from the expected straight line in the plot.