

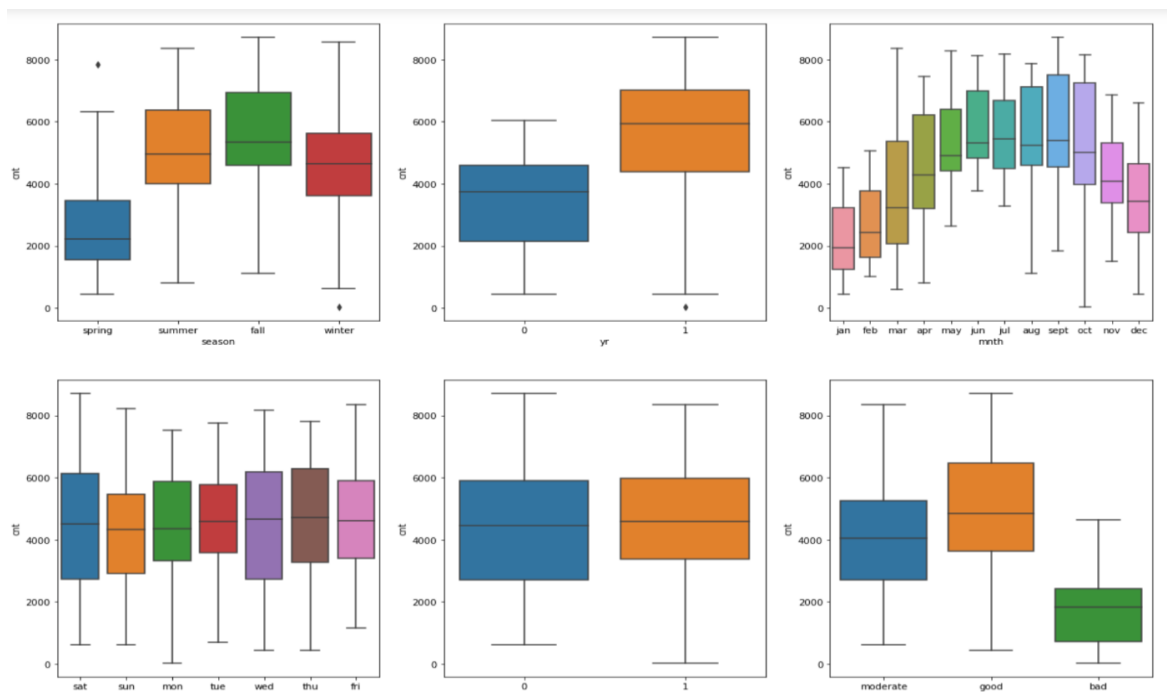
Assignment-Based Subjective Questions

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

ANSWER:

According to my analysis:

1. Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
2. Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
3. Clear weather attracted more booking which seems obvious.
4. Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
5. When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
6. Booking seemed to be almost equal either on working day or non-working day.
7. 2019 attracted more number of booking from the previous year, which shows good progress in terms of business.



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.

Syntax -

drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

The intention behind the dummy variable is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one. Hence drop_first=True is used so that the resultant can match up n-1 levels. Hence it reduces the correlation among the dummy variables.

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

Answer:

The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'.

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

Answer:

Linear Regression models are validated based on :

- a. Normality of error terms
 - i. Error terms should be normally distributed
- b. Multicollinearity check
 - i. There should be insignificant multicollinearity among variables.
- c. Linear relationship validation
 - i. Linearity should be visible among variables
- d. Homoscedasticity
 - i. There should be no visible pattern in residual values.
- e. Independence of residuals
 - i. No auto-correlation

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

Answer:

Top 3 features that has significant impact are temperature, month of year and season.

General Subjective Questions

1. Explain the linear regression algorithm in detail ?

Answer:

Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation –

$$Y = mX + c$$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slope of the regression line which represents the effect X has

on Y. c is a constant, known as the Y-intercept. If $X = 0$, Y would be equal to c.

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for a_0 and a_1 to find the best fit line and the best fit line should have the least error.

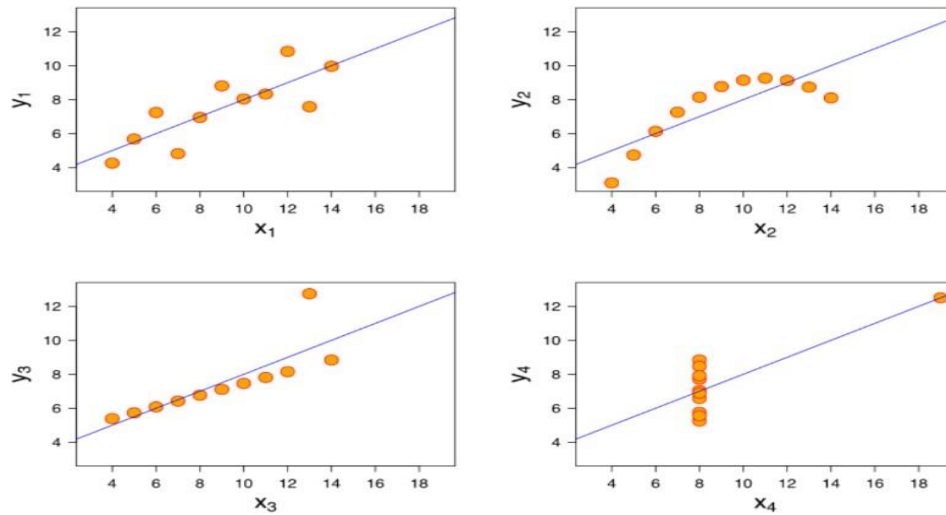
In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for a_0 and a_1 , which provides the best fit line for the data points.

2. Explain the Anscombe's quartet in detail?

Answer:

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 with very different distributions and appear differently when plotted on scatter plots.

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.



1. 1st data set fits linear regression model as it seems to be linear relationship between X and y
2. 2nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.
3. 3rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.
4. 4th data set has a high leverage point means it produces a high correlation coefficient.

Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before build machine learning model.

3. What is Pearson's R?

Answer:

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.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

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

Answer:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed, then algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.

Difference between Normalizing Scaling and Standardize Scaling:

1. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
5. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.
6. Normalized scaling is called scaling normalization whereas standardized scaling is called Z Score Normalization.
7. For Normalized scaling Scikit-Learn provides a transformer called MinMaxScaler while Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Answer:

VIF (Variance Inflation Factor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below: A VIF value of greater than 10 is high, a VIF of greater than 5 should also not be ignored and inspected appropriately.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which leads to $1/(1-R^2)$ infinity. To solve this, we need to drop one of the variables from the dataset which is

causing this perfect multicollinearity.

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

Answer:

Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Quantile-Quantile (Q-Q) plot is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

QQ plot can also be used to determine whether two distributions are similar or not. If they are quite similar you can expect the Q-Q plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

Importance of QQ Plot in Linear Regression:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both datasets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.

Advantages:

- It can be used with sample size also.
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.