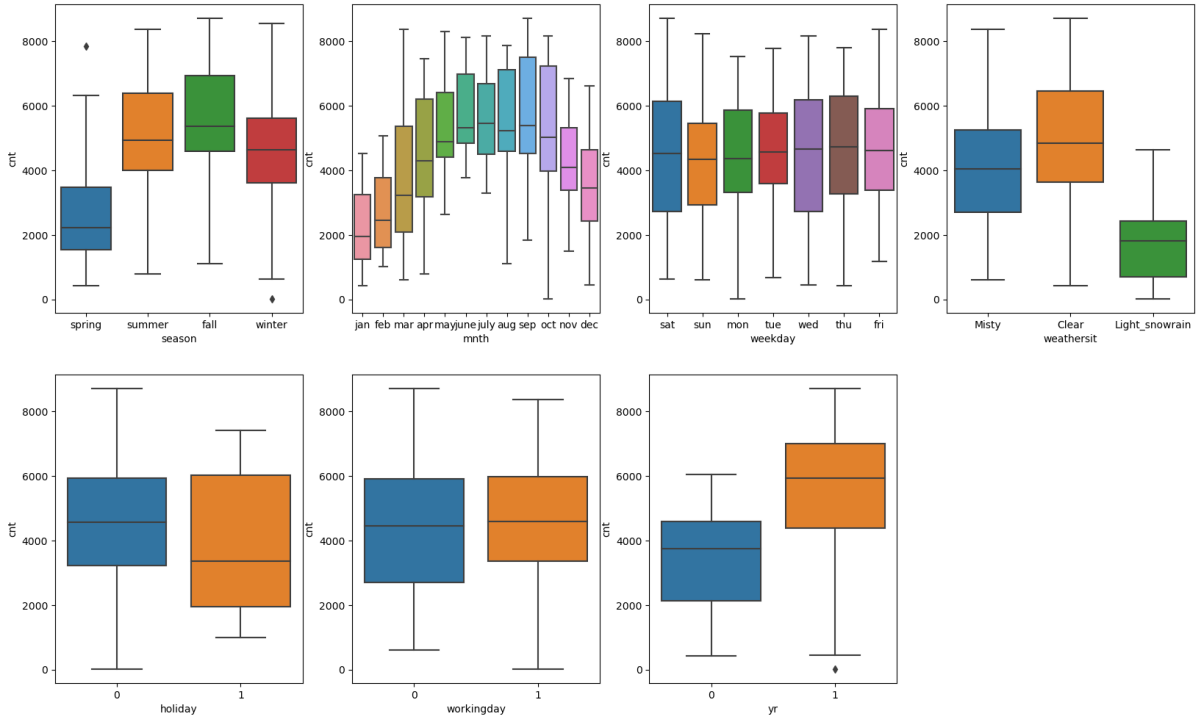
Assignment-based Subjective Questions

1. ***From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable?*** (3 marks)

Answer:

There are a couple of categorical variables namely season,mnth,yr,weekday, working day and weathersit. These categorical variables have a major effect on the dependent variable 'cnt'. The below fig shows the correlation among the same

Fig :



Using both Barplot and boxplot , these variables are analyzed.

- Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019
- Clear weather attracted more booking which seems obvious.
- 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.
- 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?*** ***(2 mark)***

Answer:

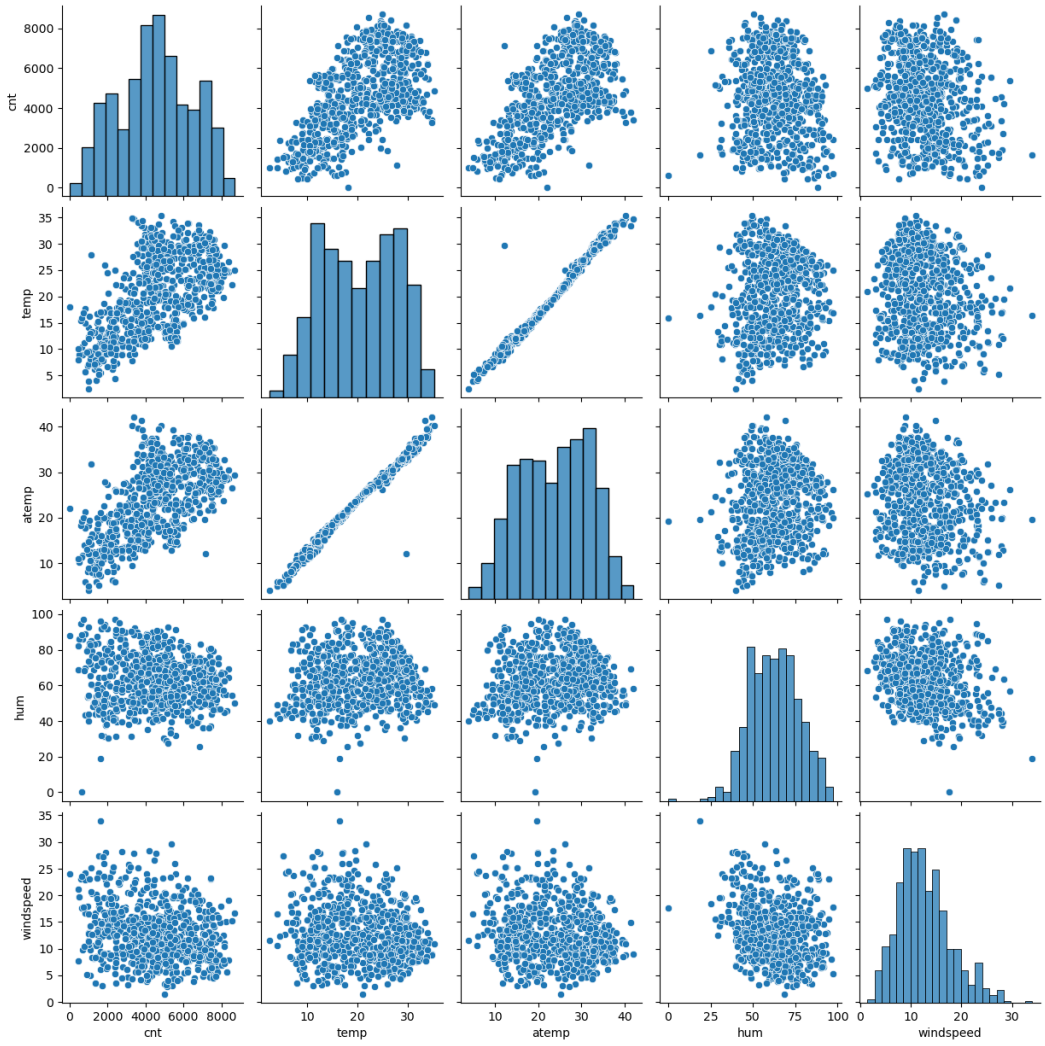
The intention behind the dummy variable is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one. Hence drop_first=True is used so that the resultant can match up n-1 levels. Hence it reduces the correlation among the dummy variables. Eg: If there are 3 levels, the drop_first will drop the first column.

Syntax –

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

3. ***Looking at the pair-plot among the numerical variables, which one has the highest correlationwith the target variable?*** (1 mark)

Answer:



from the above plots we can clearly conclude that temp and atemp are having high correlation. from the plots we can also say that there is a linear relationship between TEMP and ATEMP

4. ***How did you validate the assumptions of Linear Regression after building the model on the training set?*** (3 marks)

Answer:

I have validated the assumption of Linear Regression Model based on below assumptions –

- Normality of error terms (Error terms should be normally distributed).
- Multicollinearity check (There should be insignificant multicollinearity among variables).
- Linear relationship validation (Linearity should be visible among variables).
- Homoscedasticity (There should be no visible pattern in residual values).
- Independence of residuals (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?*** (2 marks)

Answer:

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- Temperature
- Year
- Season

General Subjective Questions

1. ***Explain the linear regression algorithm in detail.*** (4 marks)

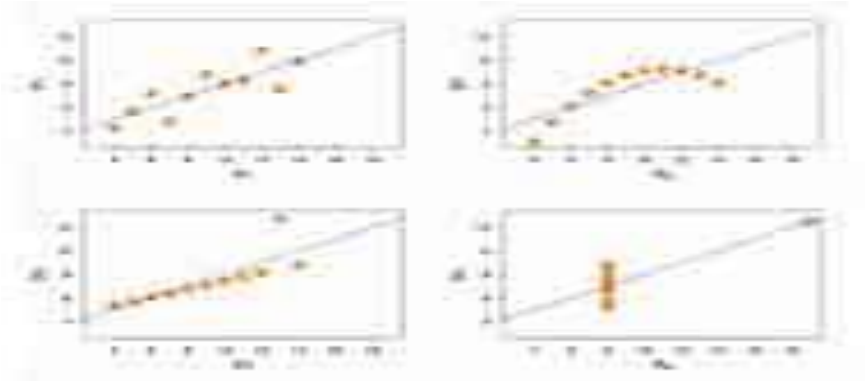
Answer:

Linear regression is a form of predictive modeling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables. 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 a0 and a1 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 a0 and a1, which provides the best fit line for the data points.

2. ***Explain the Anscombe's quartet in detail.*** (3 marks)

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. They have very different distributions and appear differently when plotted on scatter plots. It was constructed 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.



- 1st data set fit linear regression model as it seems to be linear relationship between X and y
- 2nd dataset does not show a linear relationship between X and Y, which means it does not fit the linear regression model.
- 3rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.
- 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?*** (3 marks)

Answers:

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.

Formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

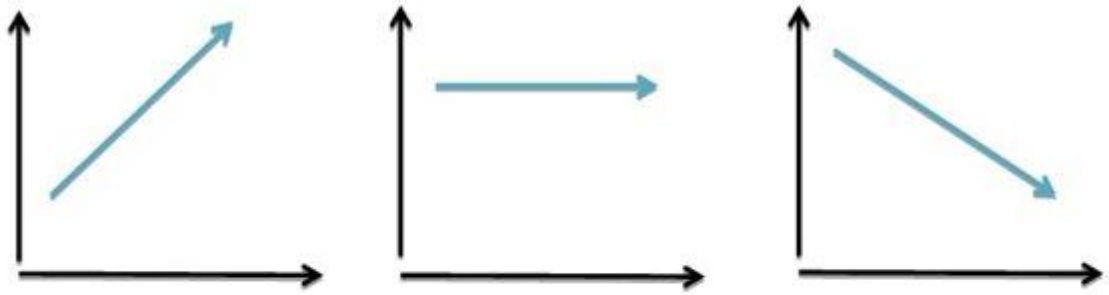
x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

Figure :



POSITIVE CORRELATIOON ZERO CORRELATION NEGATIVE CORRELATION

Pearson's correlation coefficient **"r"** is **defined** as the **covariance** (it's a measure of how much two variables change together) of the two variables **divided** by the **product of their standard deviations.**

r = Covariance / (S.D.1* S.D.2)

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

4. ***What is scaling? Why is scaling performed? What is the difference between normalized scalingand standardized scaling?*** (3 marks)

Answer:

Scaling means you're transforming your data so that it fits within a specific scale. 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 than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling. Difference between Normalizing Scaling and Standardize Scaling:

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

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

VIF(VarianceInflationFactor) 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 definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately. A very high VIF value 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.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. Importance of Q-Q plot:

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 data sets 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.