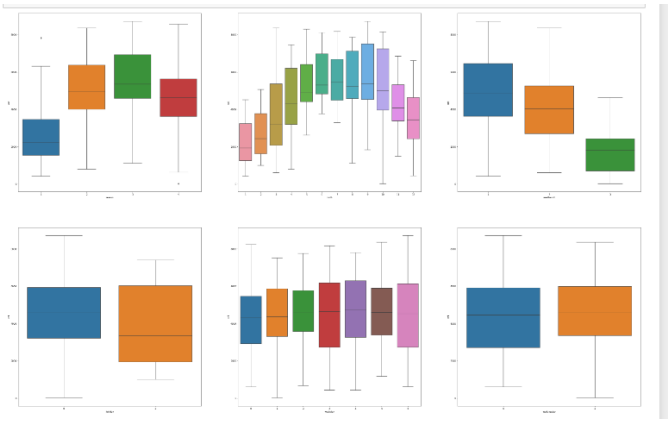
Assignment-based Subjective

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

Answer 1 : I have considered Season, Weather, Month and Weekday as the category variables as per my final analysis and based on below EDA I can conclude that Season and weather is having the most significant impact on the sales or we can say dependent variable Total Sale (cnt) is impact with independent variables Season,Weather and Month but not much on Weekdays.

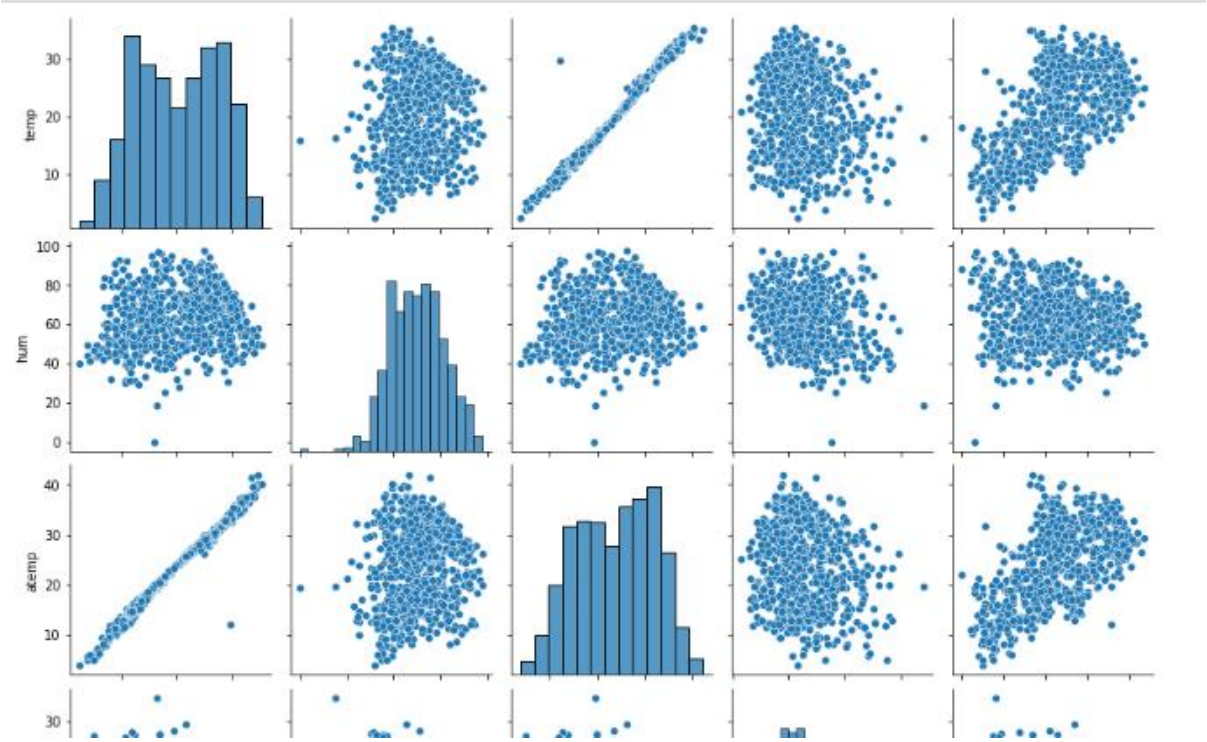


2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans: The main reason to including the drop_first = TRUE is to avoid creating a multicollinearity issue between the variables I think it reduces the space.

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

Ans There is a linear relationship between temp and atemp with target variable.



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

Ans: There are couple of factors we have considered :

- A Very Low Multicollinearity between the predictors and the p-values for all the predictors in our case all are 0.
- Rsquare and R value is good and very close approx .766.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: The demand of the bike first year wise, second factor is Season and third is the month.

General Subjective Questions

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

Linear regression is a part of Statistical Analysis here we can run multiple regression to predict the outcome of dependent variable based on independent variable. The simplest form of the regression equation(Simple Linear Regression) with one dependent and one independent variable is defined by the formula y = c + m*x, where y = estimated dependent variable, c = constant, m = regression coefficient, and x = score on the independent variable.

There are 2 types of Algorithm we practice in ML

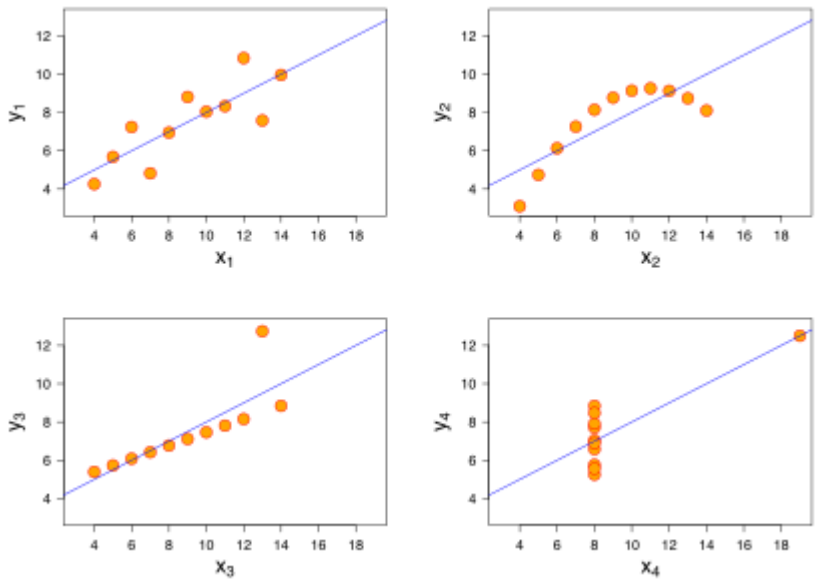
1. Simple Liear Regression where one dependent variable and one independent variable

2. Multiple Linear Regression where 1one dependent variable and n independent variable deriving the factor.

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

Ans: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.

Below Image SOURCE WIKIPEDIA



The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R? (3 marks)

Ans: It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

A negative value of the correlation coefficient means that when there is a change in one variable, the other changes in a proportion but in the opposite direction, and if the value of the correlation coefficient is positive, both the variables change in a proportion and the same direction.

4. What is scaling? Why is scaling performed?

What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: It is a very efficient way to bring down the data set in the range and easy to treat the outliars. The different independent variables mught having different magnitude and based on this if we perform linear regression the model will not fit in hence with the help of scaling we bring all the different magnitude values within the range of 0 to 1 only.

Normalized Scaling brings down the value within the score of 0-1 with below formula.

Min Max Scaling: x= (x- min(x))/(max(x) –min(x))

Standardization Scaling: It replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

Formula x = (x-mean(x))/sd(x)

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

Ans; Whenever there is there is perfect co=-relation between two variable the VIF is infinity. In this case R2= 1 hence 1/(1-R2) will become infinity. This is a example of perfect linear combination so we can say dependent variable is linear or exact related to independent variable.

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

Ans. Q-Q plots are also known as Quantile-Quantile plots. 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.