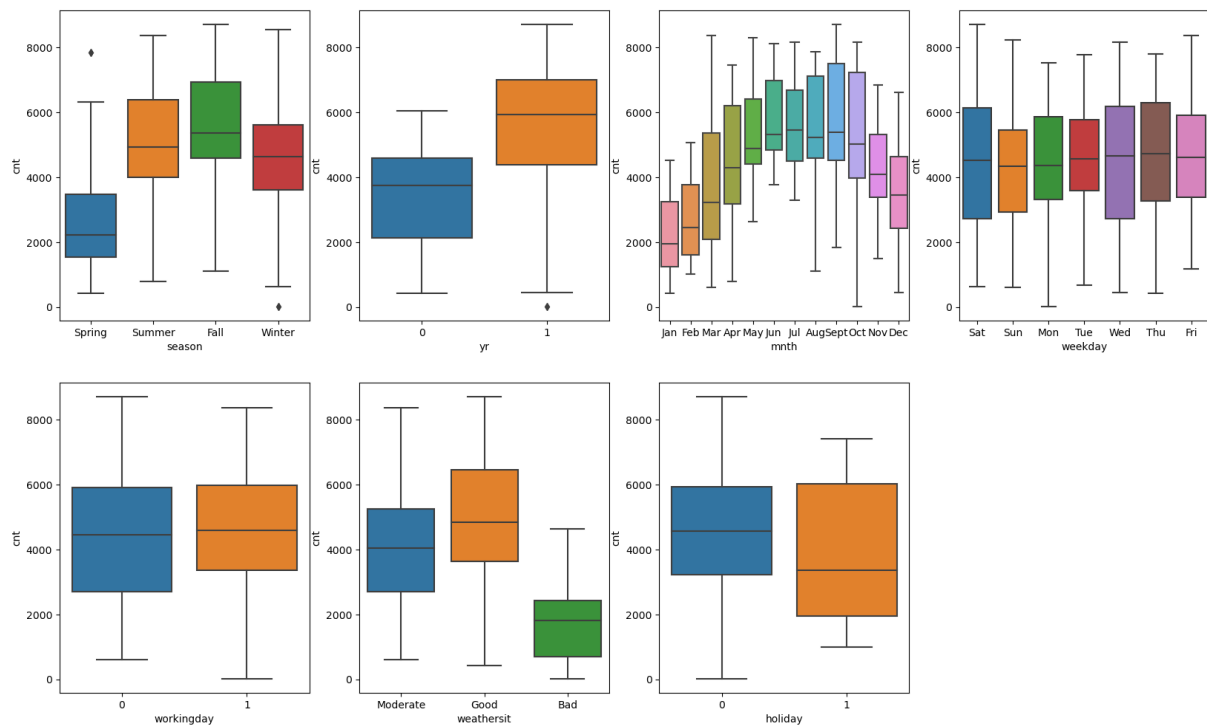


Assignment-based Subjective Questions

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

Answer: Categorical variables **season**, **yr**, **mnth**, **weekday**, **workingday** and **weathersit**. These categorical variables have a major effect on the dependent variable **cnt**.



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

Answer: It helps in reducing the extra column created during dummy variable creation. It creates n-1 new columns each indicating whether that level exists or not using a zero or one. If there are 3 levels, the drop_first will drop the first column.

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

Answer: **atemp** and **temp** variables have the highest correlation with the target variable.

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

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

- Normality of error terms: Error terms are following normal distribution
- Multi Collinearity: VIF values are less than 5 which is good

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

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

- temp
- season
- yr

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

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 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.

$Y = mX + c$

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

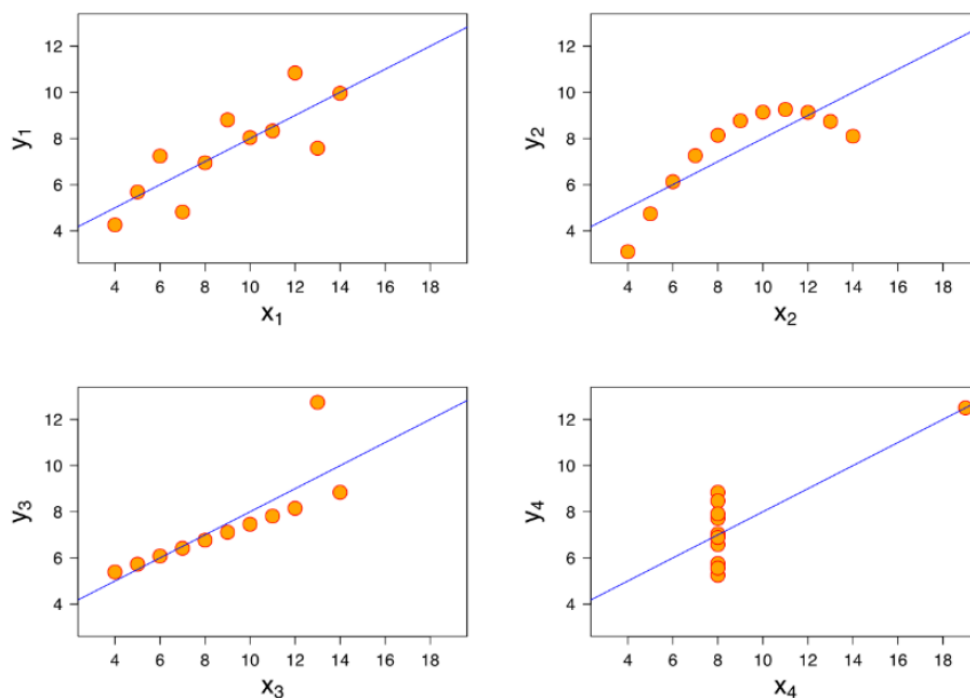
X is the independent variable we are using to make predictions.

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.

Question 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. 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.

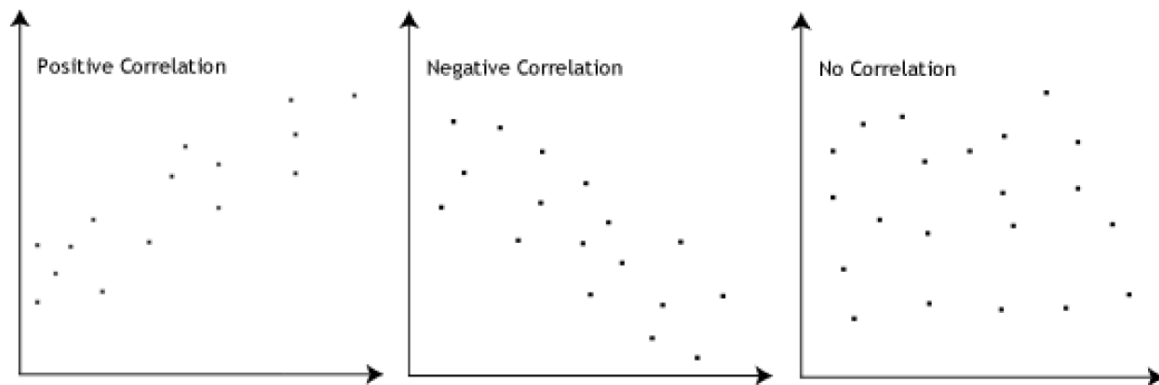


- Dataset 1 appears to have clean and well-fitting linear models.
- Dataset 2 is not distributed normally.
- In Dataset 3 the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset 4 shows that one outlier is enough to produce a high correlation coefficient.

Question 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. This is shown in the diagram below:



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

Answer: Scaling is a technique to transforming the data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in a 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:

- In normalized scaling minimum and maximum values of features being used whereas in standardised scaling mean and standard deviation are 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 does not have 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 scaling normalization whereas standardized scaling is

called as Z Score Normalization.

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

Answer: If there is perfect correlation, then $VIF = \infty$. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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 $1/(1-R^2) = \infty$. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

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 per cent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% per cent of the data falls 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.