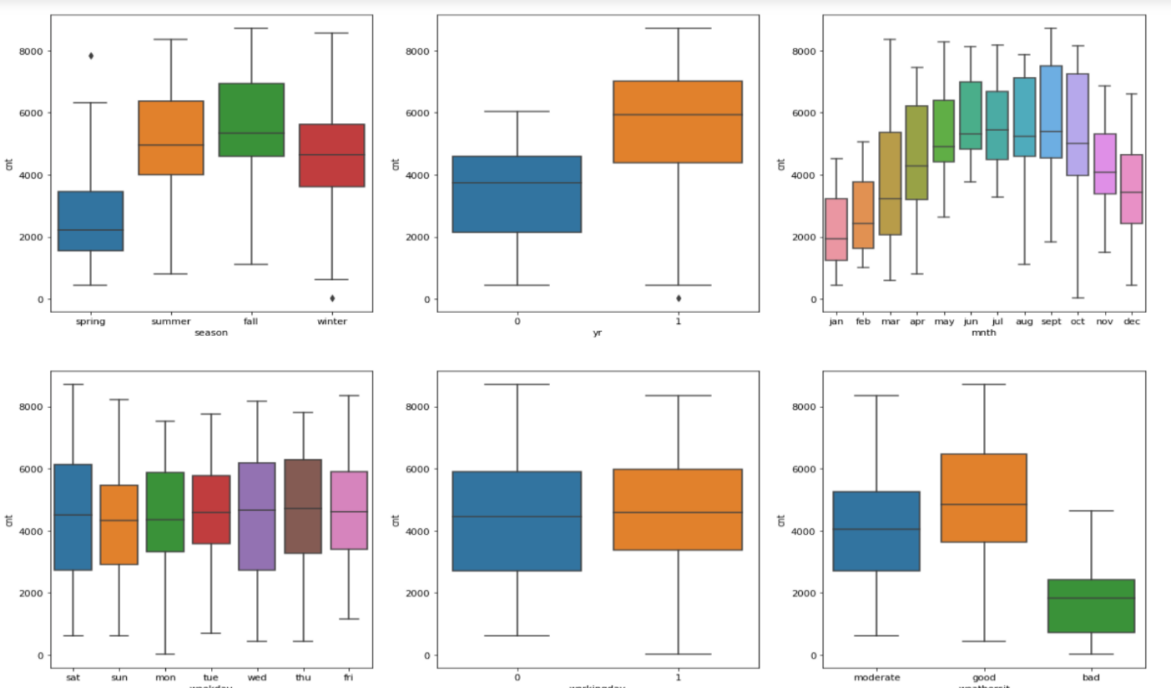
KESHAV ATCHUTUNI

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: Several categorical variables, including season, month, year, weekday, working day, and weather it is significantly influence the dependent variable 'cnt'. The correlation among these variables is depicted in the following figure. Both bar plots and box plots are used to visualize these variables.

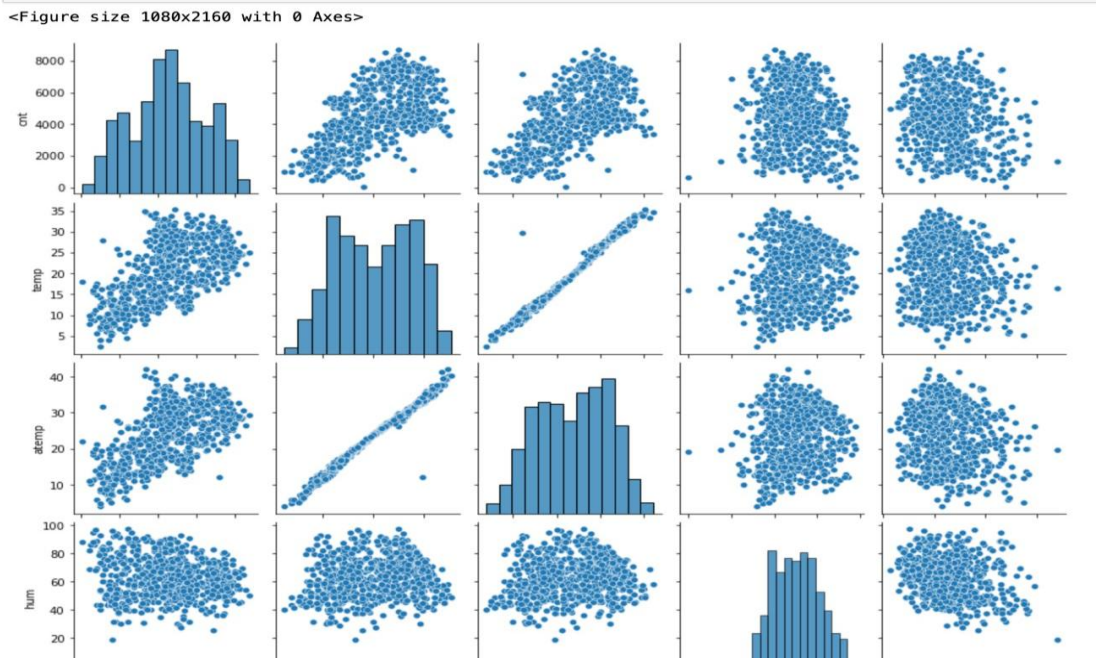


2. **Why is it important to use drop_first=True during dummy variable creation?**

The purpose of a dummy variable is to represent categorical variables with 'n' levels by creating 'n-1' new columns. Each column indicates whether a particular level exists or not using binary values (0 or 1). Setting drop_first=True ensures that the resulting dummy variables match 'n-1' levels, reducing correlation among them. For instance, if there are 3 levels, drop_first will eliminate the first column.

KESHAV ATCHUTUNI

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



The ‘temp’ and ‘temp’ 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?**

   Linear Regression models are validated based on Linearity, No auto-correlation, Normality of error, Homoscedasticity, Multicollinearity.

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

   Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature, year and season

General Subjective Questions

1. Explain the linear regression algorithm in detail?

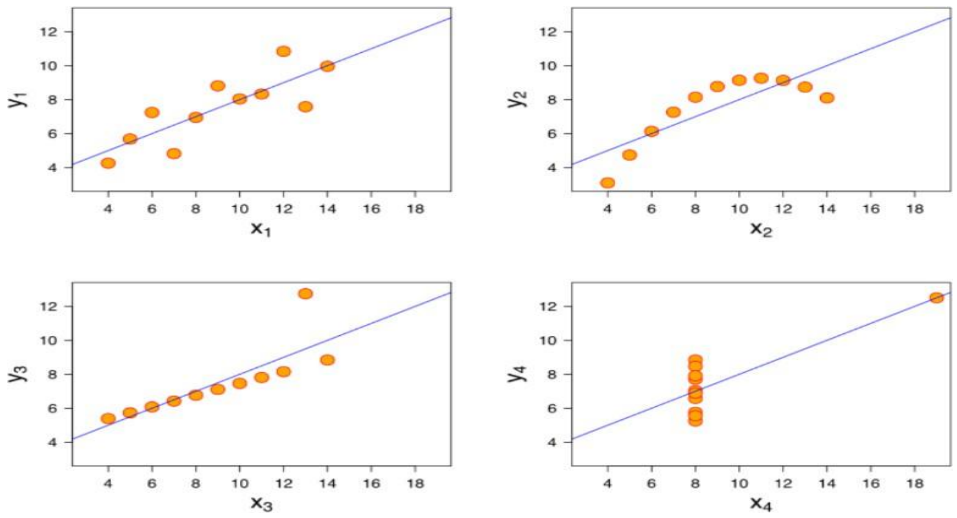
Linear regression is a form of predictive modeling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing 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

KESHAV ATCHUTUNI

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?

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 fits linear regression model as it seems to be linear relationship between X and y
- 2nd data set 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.

KESHAV ATCHUTUNI

3. What is Pearson's R?

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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

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:

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 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?

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 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.

KESHAV ATCHUTUNI

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

The Quantile-Quantile (Q-Q) plot serves as a probability plot, comparing two probability distributions by plotting their quantiles against each other. It helps assess whether a dataset follows a theoretical distribution like Normal, exponential, or Uniform. Additionally, it aids in determining the similarity between two distributions; greater similarity results in a more linear Q-Q plot. This linearity can be tested with scatter plots, while the multivariate normality assumption for linear regression can be assessed with histograms or Q-Q plots.

In linear regression, Q-Q plots are valuable for comparing the distributions of training and test datasets to ensure they are from the same population. Key advantages include applicability to various sample sizes and the ability to detect shifts in location, scale, symmetry, and outliers. Q-Q plots are employed to verify if datasets share a common distribution, location, scale, distribution shape, and tail behavior.