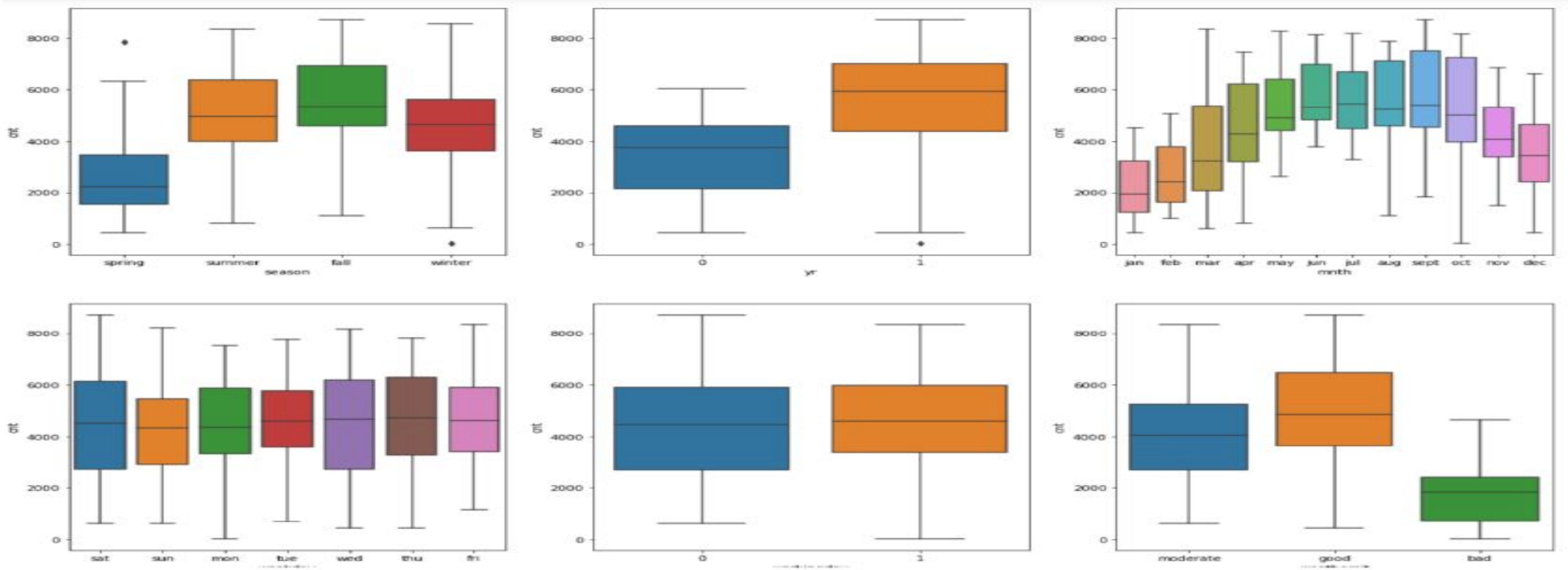
Assignment Based - Subjective Questions

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

There are a number of categorical variables, namely Season, Month, Year, Day of the Week, Day of the Week and Weather. These categorical variables have a large effect on the dependent variable "cnt". The figure below shows the correlation



These variables are visualized using bar plot and Box plot both.

Assignment Based - Subjective Questions

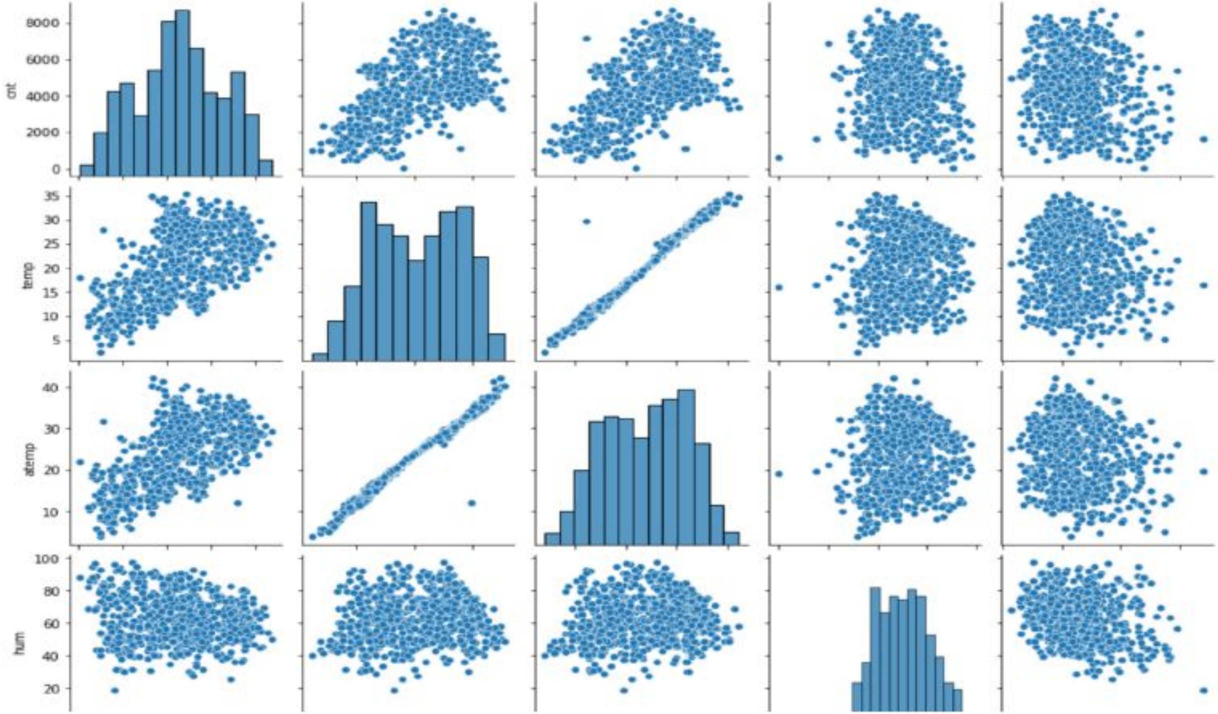
2.Why is it important to use drop_first=True during dummy variable creation?

The purpose of a dummy variable is to create "n-1" new columns for a categorical variable with "n" levels, each indicating whether that level exists or not, use zero or one. Therefore drop_first=True is used so that the result corresponds to n-1 levels. Therefore, it reduces the correlation between dummy variables.

For example: if there are 3 levels, drop_first drops the first column.

Assignment Based - Subjective Questions

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



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

Assignment Based - Subjective Questions

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, absence of autocorrelation, normality of error, homoscedasticity and multicollinearity.

5.Based on the final model, what are the three main characteristics that significantly explain the demand for shared bicycles?

The three main characteristics that have a significant impact on explaining the demand for shared bicycles are temperature, year and season.

General Subjective Questions

1.Explain the linear regression algorithm in detail.

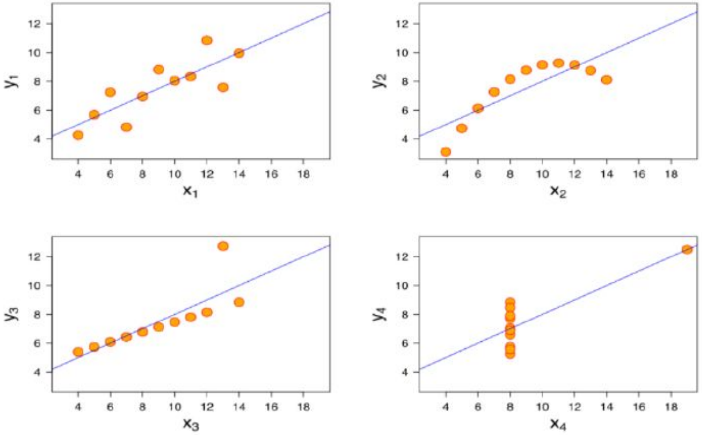
Linear regression is a form of predictive modeling that describes the relationship between dependent (target variables) and independent variables (predictors). Because linear regression shows a linear relationship, which means it finds out how the value of the dependent variable changes according to the value of the independent variable. When there is a single input variable (x), such linear regression is called simple linear regression. And when there are more than one input variables, such linear regression is called multiple linear regression. A linear regression model provides a sloping line that describes the relationship between variables. The regression line can be a positive linear relationship or a negative linear relationship. The goal of the linear regression algorithm is to obtain the best values for a0 and a1 to find the best-fit line, and the best-fit line should have the smallest error. Linear regression uses the RFE or mean squared error (MSE) or cost function to help find the best possible values for a0 and a1 that provide the best fit to the data points.

General Subjective Questions

2.Explain the Anscombe's quartet in detail.

An Anscombe Quadruple can be defined as a group of four data sets that are nearly identical in simple descriptive statistics, but the data set has some characteristics that make the regression model look silly. They have very different distributions and look different when plotted on scatter charts. It was designed to illustrate the importance of plotting before analysis and model building, as well as the influence of other observations on statistical properties. These four data set plots are almost the same statistical observations, providing the same statistical data with variance and mean of all x, y points in all four data sets.

- 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 coeff.



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

General Subjective Questions

3.What is Pearson’s R?

In statistics, Pearson's correlation coefficient is also called Pearson's r, Pearson's correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

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

General Subjective Questions

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

Scaling means changing data to fit a specific scale. This is a kind of data processing step where we adjust the data to a certain scale and speed up the calculations in the algorithm. Collected data contains characteristics of different size, unit and scope. If no scaling is done, the algorithm tends to weight large values and ignore other parameters, resulting in incorrect modeling.

Difference between Normalizing Scaling and Standardize Scaling:

1. The normalized scale uses the minimum and maximum value of the characteristics, while the standard scale value uses the scale mean and standard deviation.

2. Normalized scale is used when functions are on different scales, while standardized scale is used to ensure zero mean and unit standard deviation.

General Subjective Questions

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

3. A normalized scale scales values between (0,1) or (-1,1), while a standardized scale does not have a fixed range or is not limited to it.

4. A normalized scale is affected by outliers, while a standardized scale is not affected by outliers.

5. A normalized scale is used when we do not know the distribution, and a standardized scale is used when the distribution is normal.

6. Normalized scaling is called scale normalization, while standardized scaling is called Z-Score normalization.

General Subjective Questions

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 other independent variables. The composition of VIF is given below:

A VIF value greater than 10 is definitely high. A VIF value greater than 5 should also not be ignored and checked properly.

A very high VIF value indicates a perfect correlation between two independent variables. For perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem, we need to drop the one variable from the data set that causes it to be in perfect multicollinearity.

General Subjective Questions

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

A Q-Q plot is a probability plot which is a graphical method of comparing two probability distributions by plotting their quantiles against each other.

A quantile-quantile (Q-Q) plot is a graphical tool that helps us assess whether a set of data can come from some theoretical distribution, such as the normal, exponential, or uniform distribution.

A QQ plot can also be used to determine whether two distributions are similar or not. If they are similar enough, you would expect the QQ plot to be more linear. The assumption of linearity is best tested with scatterplots. Second, linear regression analysis requires all variables to be multivariate normal. This assumption is best checked using a histogram or Q-Q plot.

General Subjective Questions

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

Importance of QQ Plot in Linear Regression: In linear regression, when we have a training and a test data set, we can create a Q-Q plot to check whether both the data train and the test data set come from the population with the same distribution or not.

Advantages:
● Can also be used with sample size.
● Many aspects of the distribution can be observed on this diagram, such as changes in location, changes in scale, changes in symmetry and the presence of anomalies. Using a Q-Q plot with two sets of data to check
● Whether both sets of data come from a population with a common distribution.
● If both datasets have a common location and a common scale.
● If both datasets have the same distribution shape. ● If both datasets have tail behavior.