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?**

➔ I have analyzed the categorical variables using boxplot and barplot. Here are some points we can get from visualizations:

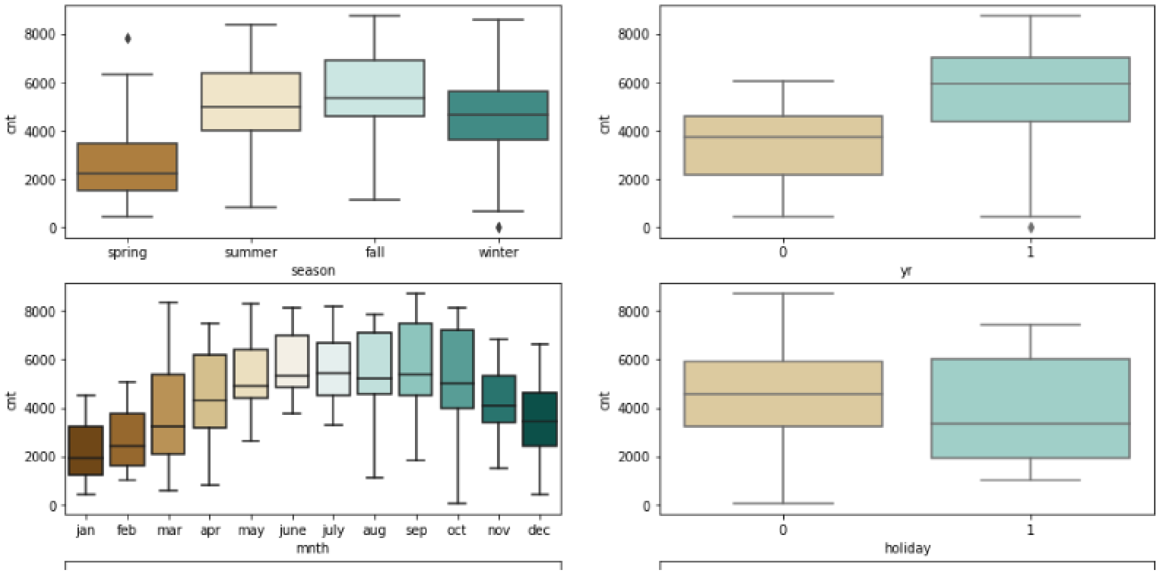
➔ Bookings appear to have increased throughout the fall season. And, from 2018 to 2019, the number of bookings in each season significantly grew.

➔ Most reservations were made in the months of May, June, July, August, September, and October. Beginning in January and continuing through mid-year, the trend grew before beginning to decline as the year came to a close.

➔ It appears obvious that more bookings were lured by clear weather.

➔ There are more reservations on Thursday, Friday, Saturday, and Sunday than at the beginning of the week.

➔ Bookings appear to be lower when it's not a holiday, which makes sense given that during holidays, people would want to stay home and enjoy time with their families.

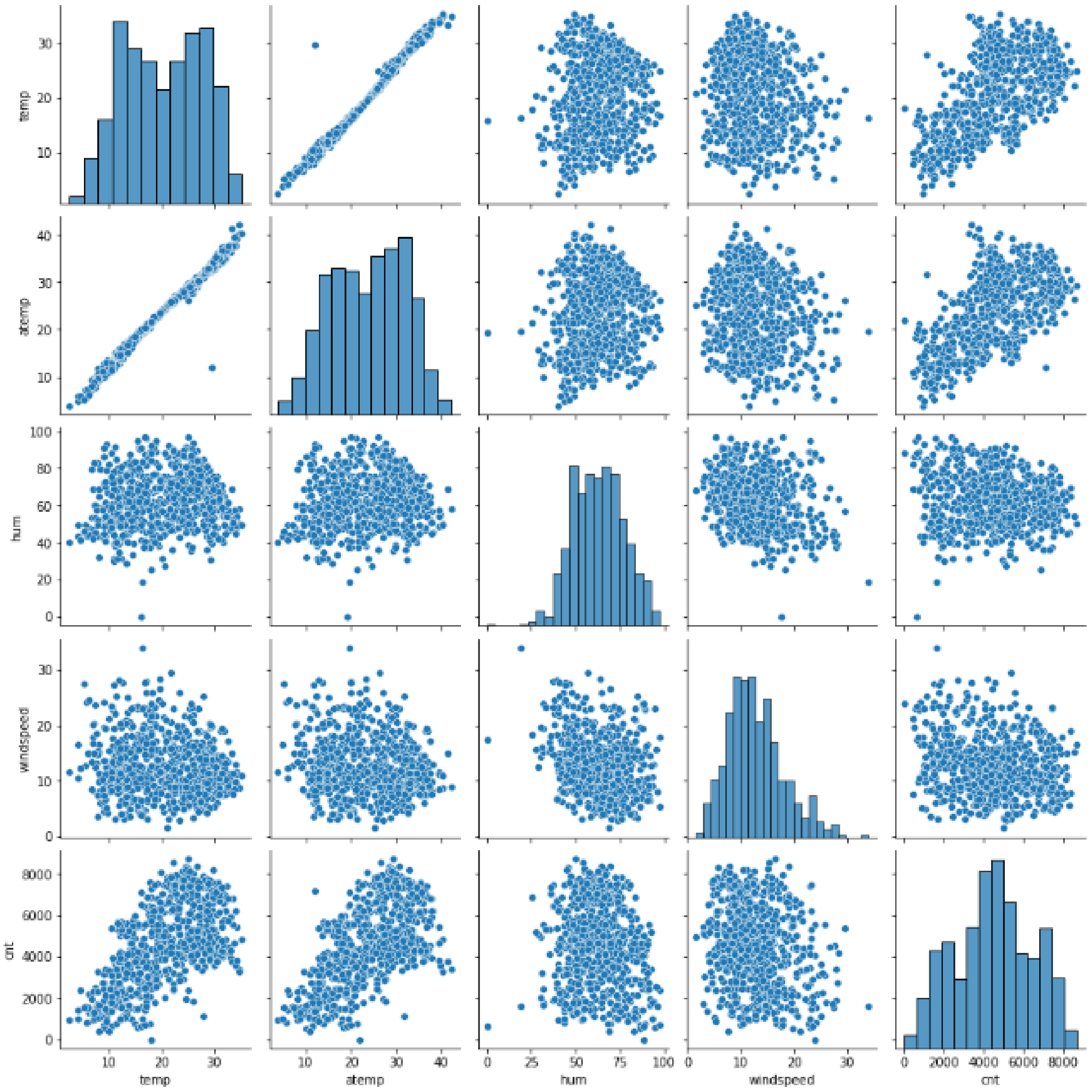


2. **. Why is it important to use drop_first=True during dummy variable creation?**

drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation.

Hence it reduces the correlations created among dummy variables. Let's imagine we want to build a dummy variable for a categorical column that has three different types of data. If one factor is neither A nor B, then it is clear that C. Thus, we do not require the third variable to locate the C.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Tramp and ateamp are the two variables which are highly correlated to each other.
   Attaching the pair-plot snip here:



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

The following criteria are used to validate linear regression models:

- Linearity
- No autocorrelation
- 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?

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

- ➔ temp
- ➔ winter
- ➔ sep

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A type of predictive modeling method called linear regression explains the relationship between the dependent (goal/target variable) and independent factors (predictors). Given that linear regression demonstrates a linear relationship, it may be used to determine how the dependent variable's value changes as a function of the independent variable's value. Such linear regression is known as simple linear regression if there is only one input variable (x). Additionally, this type of linear regression is known as multiple linear regression if there are many input variables. The link between the variables is depicted by a sloping straight line in the linear regression model.

Regression curves can have either positive or negative linear relationships. The best results possible is the purpose of the linear regression procedure.

Mathematically the relationship can be represented with the help of the following equation − $Y = mX + c$

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.

There are two different types of linear regression:

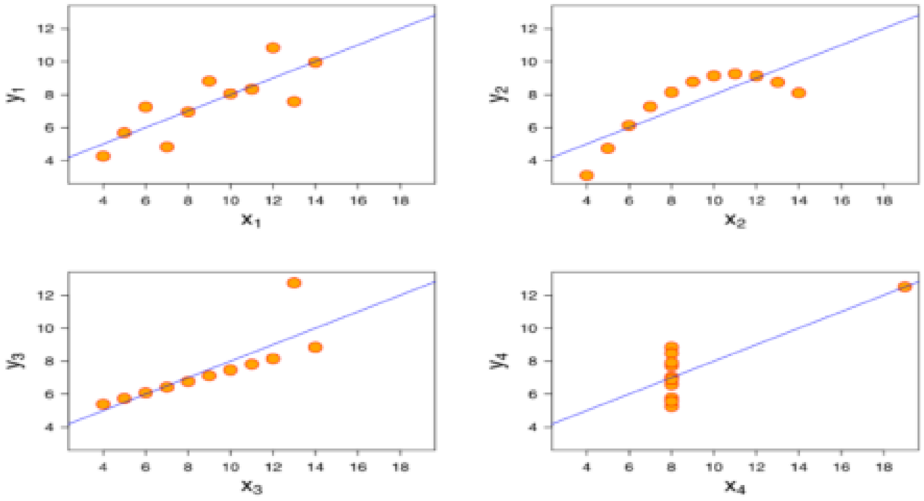
- simple linear regression
- multiple linear regression.

Assumptions - The linear regression model makes the following assumptions on the dataset:

- Multi-collinearity– The linear regression model is predicated on the premise that the data exhibit very little to no multicollinearity. Multicollinearity basically happens when independent variables or features depend on one another.
- Auto-correlation -- Added presumption The linear regression model presupposes that the data exhibits either very little or no autocorrelation. In essence, auto-correlation happens when residual errors are dependent on one another.
- Relationship between variables – The linear regression model presupposes a linear relationship between the response and feature variables.
- Normality of error terms – Error terms should be normally distributed.
- Homoscedasticity – There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail?

Anscombe's Quartet is a set of four data sets that, while they appear to be almost equal in simple descriptive statistics, have certain anomalies that, if a regression model is developed, would deceive it. When displayed on scatter plots, they have significantly different distributions and show up differently. It was designed to demonstrate the significance of graphing data prior to analysis and model creation, as well as the impact of additional observations on statistical features. These four data set plots all contain almost identical statistical observations and offer the same statistical information, including variance and mean values for all x and y points for all four datasets



- Since there appears to be a linear relationship between X and Y, the first data set matches the linear regression model.
- The second data set does not demonstrate a linear relationship between X and Y, hence it does not satisfy the requirements of the linear regression model.
- The third data set reveals several outliers that are present in the dataset but which a linear regression model cannot account for.
- The fourth data set's high leverage point results in a strong correlation coefficient.
- Regression methods are susceptible to manipulation, thus data visualization is crucial when developing a machine learning model.

3. What is Pearson's R?

A numerical evaluation of the strength of the linear connection between the variables is provided by Pearson's r. The correlation coefficient will be positive if the variables tend to rise and fall together. The correlation coefficient will be negative if the variables have a tendency to rise and fall in opposition, with low values of one variable correlated with high values of the other. Between +1 and -1 are the possible values for the Pearson correlation coefficient, or r. There is no link between the two variables, as indicated by a value of 0. Positive associations have values greater than 0, meaning that if one variable's value rises, so does the value of the other. A negative value is one with a value less than 0.

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

Scaling is the process of converting data to match a particular scale. It is a particular kind of data preprocessing step where we scale data to a given size and quicken algorithmic calculations. The features of the collected data vary in size, unit, and range. If scaling is not done, the algorithm has a tendency to overlook other parameters and weigh high magnitude values, which leads to inaccurate modeling.

Differences between standardizing scaling and normalized scaling include:

1. In contrast to standardized scaling, which uses mean and standard deviation for scaling, normalized scaling uses the minimum and maximum value of features.

2. Standardized scaling is used to maintain zero mean and unit standard, whereas normalized scaling is utilized when features are of different scales.

3. Standardized scaling does not have or is not constrained to a specific range, whereas normalized scaling scales values between (0,1) and (-1,1).

4. Standardized scaling is unaffected by outliers, whereas normalized scaling is influenced by them.

5. Standardized scaling is used when the distribution is normal and normalized scaling is used when we don't know the distribution.

6. Standardized scaling is referred to as Z Score Normalization whereas normalized scaling is referred to as scaling normalization.

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

VIF = infinity if there is perfect correlation. A high VIF score denotes a strong connection between the variables. The presence of multicollinearity causes the variance of the model coefficient to be exaggerated by a factor of 4 if the VIF is 4. VIF displays a complete correlation between two independent variables when its value is infinite. If the correlation is perfect, we have R-squared (R2) = 1, which results in 1/ (1-R2) infinite. To fix this, we must remove the variable from the dataset that is the exact multicollinearity's cause.

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

A probability plot, or Q-Q plot, is a graphical technique for contrasting the quantiles of two probability distributions. A graphical technique known as a quantile-quantile (Q-Q) plot can be used to determine whether a collection of data may have originated from a theoretical distribution such as a normal, exponential, or uniform distribution. Determine whether or not two distributions are similar using the QQ plot. You can anticipate the QQ plot to be more linear if they are relatively comparable. Scatter plots are the most effective way to test the linearity assumption. Second, all variables must be multivariate normal in order to do a linear regression analysis. The best way to verify this assumption is via a histogram or Q-Q-Plot.

Importance of QQ Plot in Linear Regression :

It is frequently desirable to determine whether the assumption of a common distribution is supported when there are two data samples. If so, location and scale estimators can combine the two sets of data to derive estimates for the shared position and scale. If two samples do differ, it is also helpful to comprehend the variations. More information about the nature of the difference can be gleaned from the q-q plot than from analytical techniques like the chi-square and Kolmogorov-Smirnov 2-sample tests.