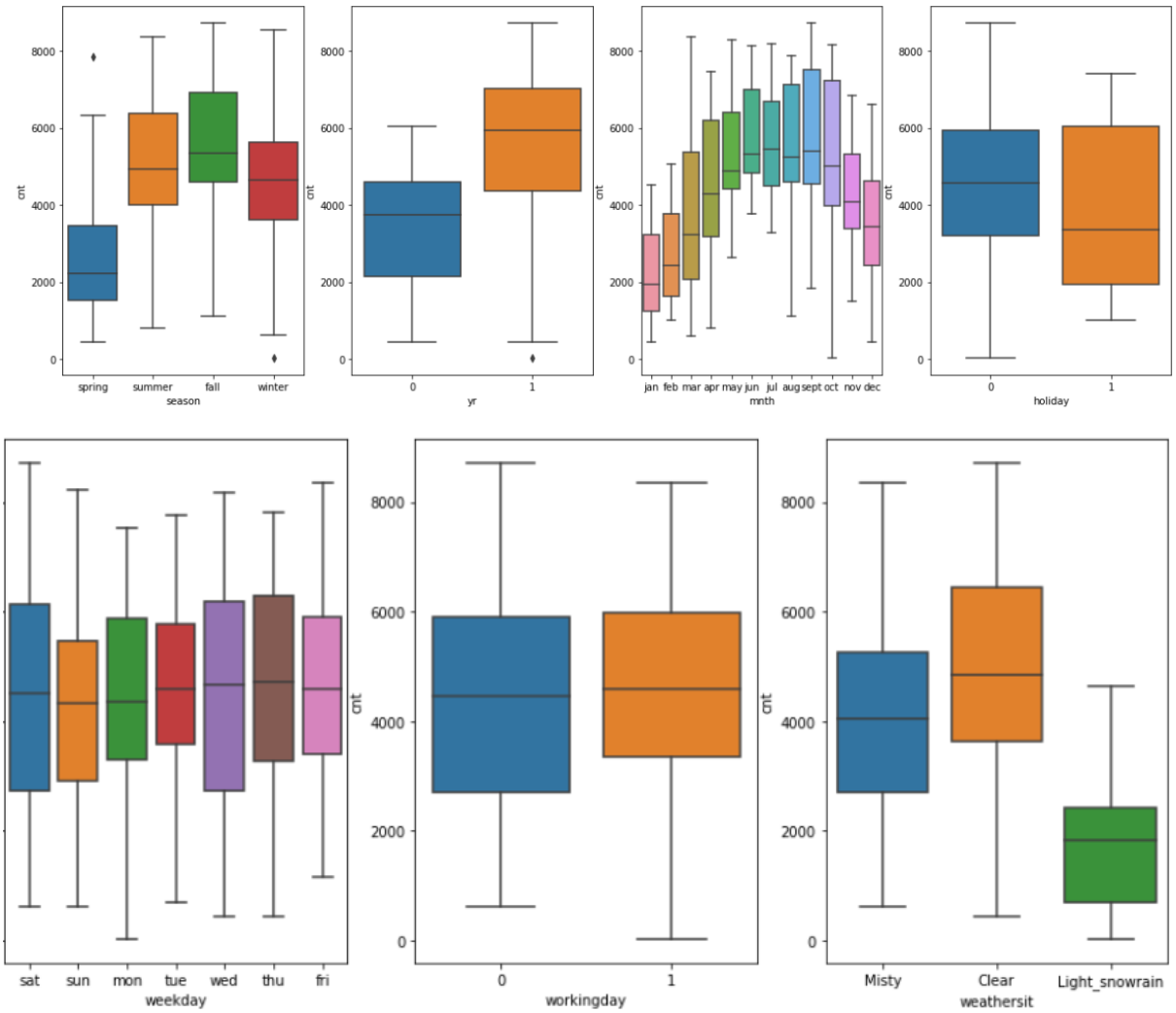
Assignment-based Subjective Questions

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

I have performed the analysis of categorical variables and observed these things:



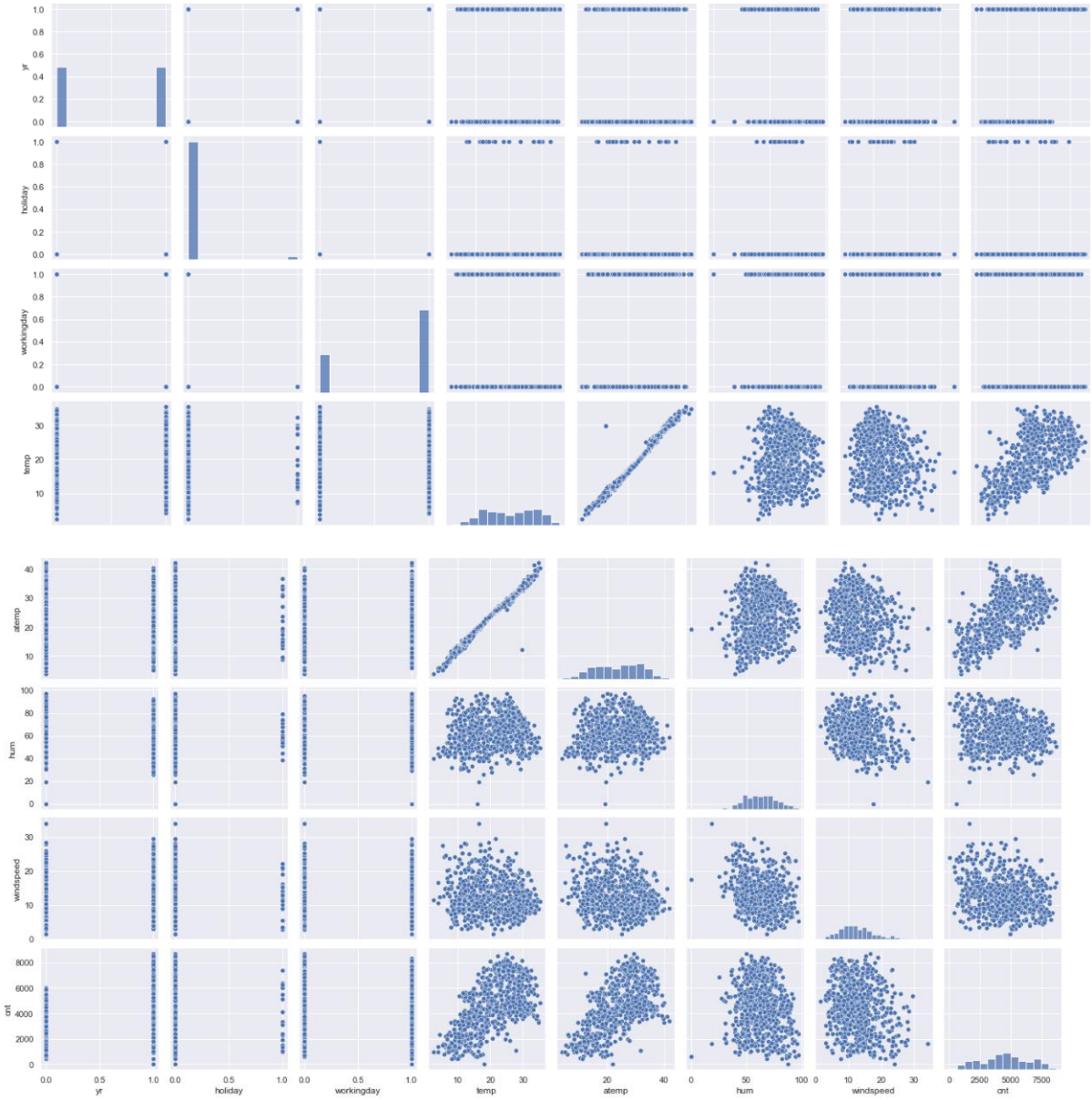
- Fall season has more booking count. The booking count has increased from 2018 to 2019.
- In the months of may, june, july, august, september and october booking count has increased. Booking count increased in the mid of the year and decreased in the end of the year.
- Saturday, wednesday, Thursday and friday has highest booking count. Booking count was low in the start of the week compared to end of the week.
- Clear weather has highest booking count.

- Booking count seems less in holidays. May be people wants to spend holidays at home with family.

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

Drop_first=True will helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables

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



From the pair plot temp has highest correlation with the target variable.

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

   I had validated the assumptions of linear regression after building the model on the training set Linearity, Multicollinearity, Homoscedasticity , Normality of error.

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:
   Temperature, year and season.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   linear regression is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous or numeric variables.

   Linear regression algorithm shows a linear relationship between a dependent (y) variable and one or more independent variables.

   The linear regression model provides a sloped straight line representing the relationship between the variables.

   Y=mx+c

   Y= Target variable

   m=slope

   x=future variable

   c=constant /intercept

   Types of linear regression:

   - Simple linear regression

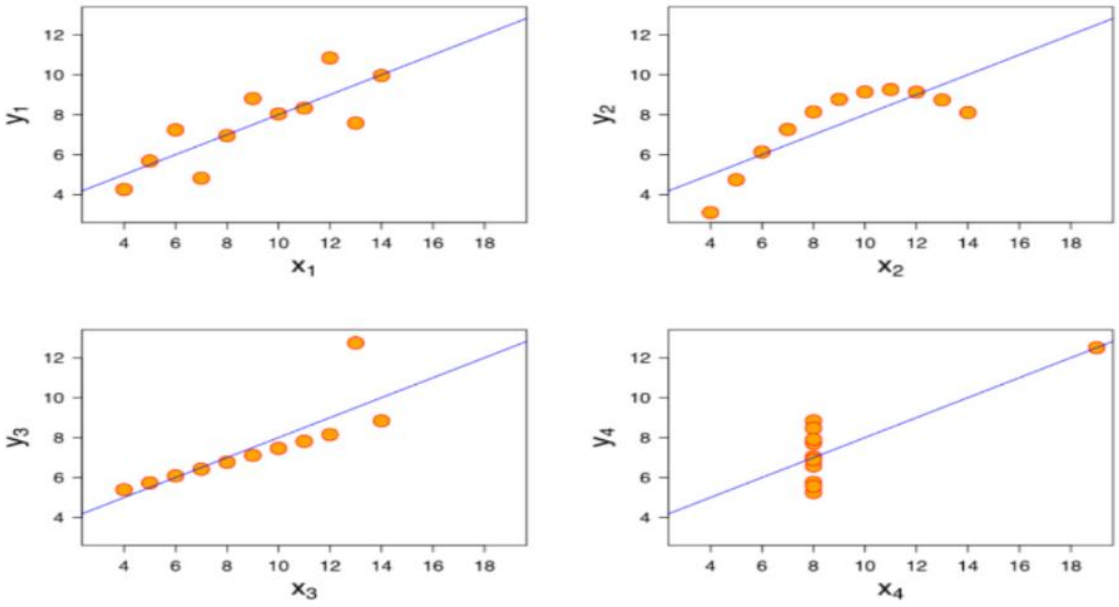
     If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- Multiple Linear regression

  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

2. **Explain the Anscombe's quartet in detail.**

   Anscombe's Quartet described as a collection of 4 statistics units which can be almost same in easy descriptive statistics. They have very unique distributions and seem in a different way while plotted on scatter plots.



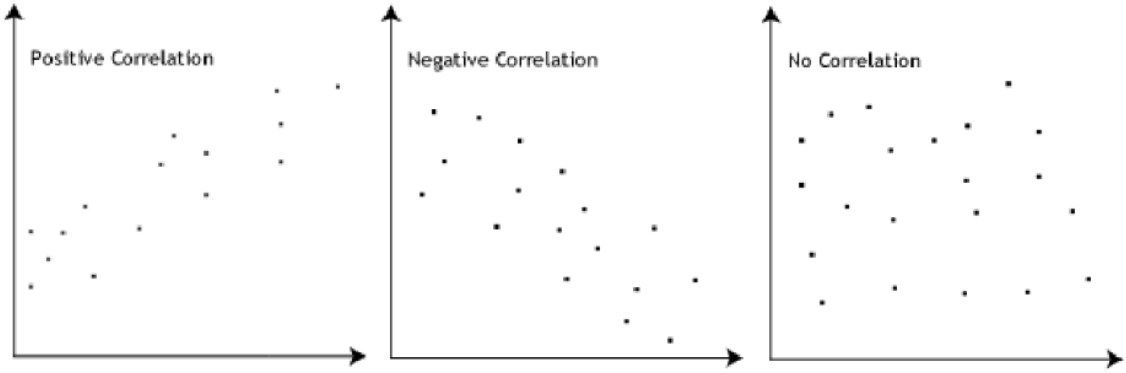
- 1 st dataset is for linear relationship between x and y
- 2nd dataset doesn't have linear relationship between x and y, which means it doesn't fits linear regression model.
- 3 rd dataset was linear but outliers were present.
- 4th dataset represents that outliers produces high relation coefficients.

3. **What is Pearson's R?**

   The Pearson correlation measures the strength of the linear relationship between two variables.

   It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation.

   The below diagram shows different types of correlation:



4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. Scaling also helps in speeding up the calculations in an algorithm.

   Normalization scales the values into a range of [0,1]. Normalization is highly affected by outliers.
   Standardization scales data to have a mean of 0 and a standard deviation of 1 (unit variance). Standardization is less affected by outliers.

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

   A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis.

   - VIF equal to 1 = variables are not correlated
   - VIF between 1 and 5 = variables are moderately correlated
   - VIF greater than 5 = variables are highly correlated

   If there is perfect corelation exists between two variables the VIF will be infinite.

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

   Q-Q plots are known as Quantile-Quantile plot.

Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The purpose of the quantile-quantile (Q-Q) plot is to show if two data sets come from the same distribution.

The purpose of the quantile-quantile (Q-Q) plot is to show whether two data sets come from the same distribution. The plot is created by plotting the quantiles of the first dataset along the x-axis and the quantiles of the second dataset along the y-axis.

A Q-Q plot is used to compare the shapes of the distributions and provide a graphical view of how properties such as position, scale, and skewness are similar or different in the two distributions.

.