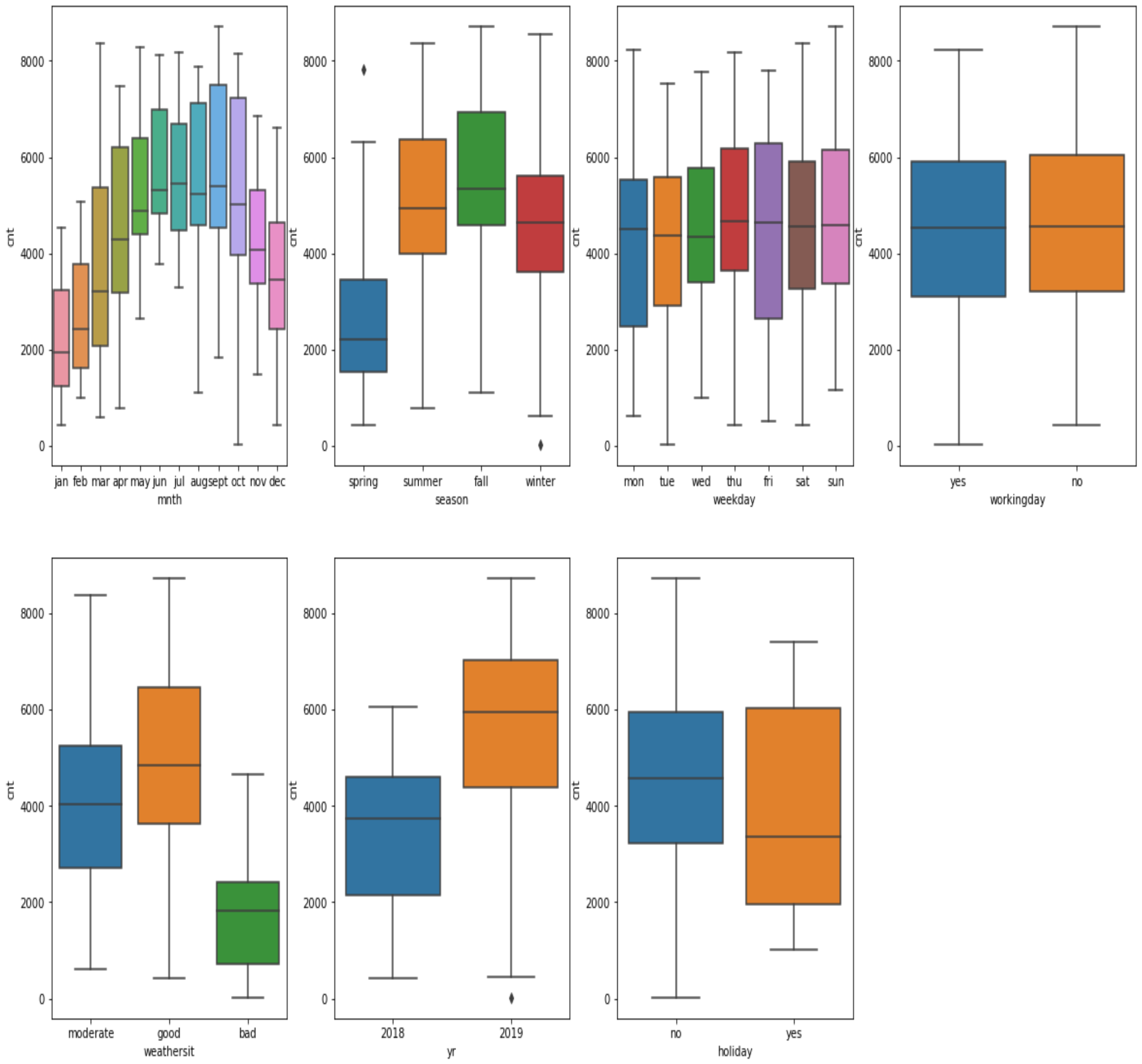
- # Assignment-based Subjective Questions

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

- i)Number of bike users (‘cnt’) increase steadily from the month of January till September, with maximum (median) users observed in month of July. From October onwards ‘cnt’ begins to decrease till December.
- ii)Summer and Fall observes highest number of ‘cnt’ followed by winter and spring.
- iii)All weekdays have almost equal number of ‘cnt’ .
- Iv)Whether its working day or holiday the number of ‘cnt’ is almost same.
- Vi)’ cnt’ is highest for good weather situation, followed by moderate and bad weather situation.
- Vii)Compared to 2018 more ‘cnt’ are observed in 2019.
- Viii)’ cnt’ is found to be decreased if it is holiday.

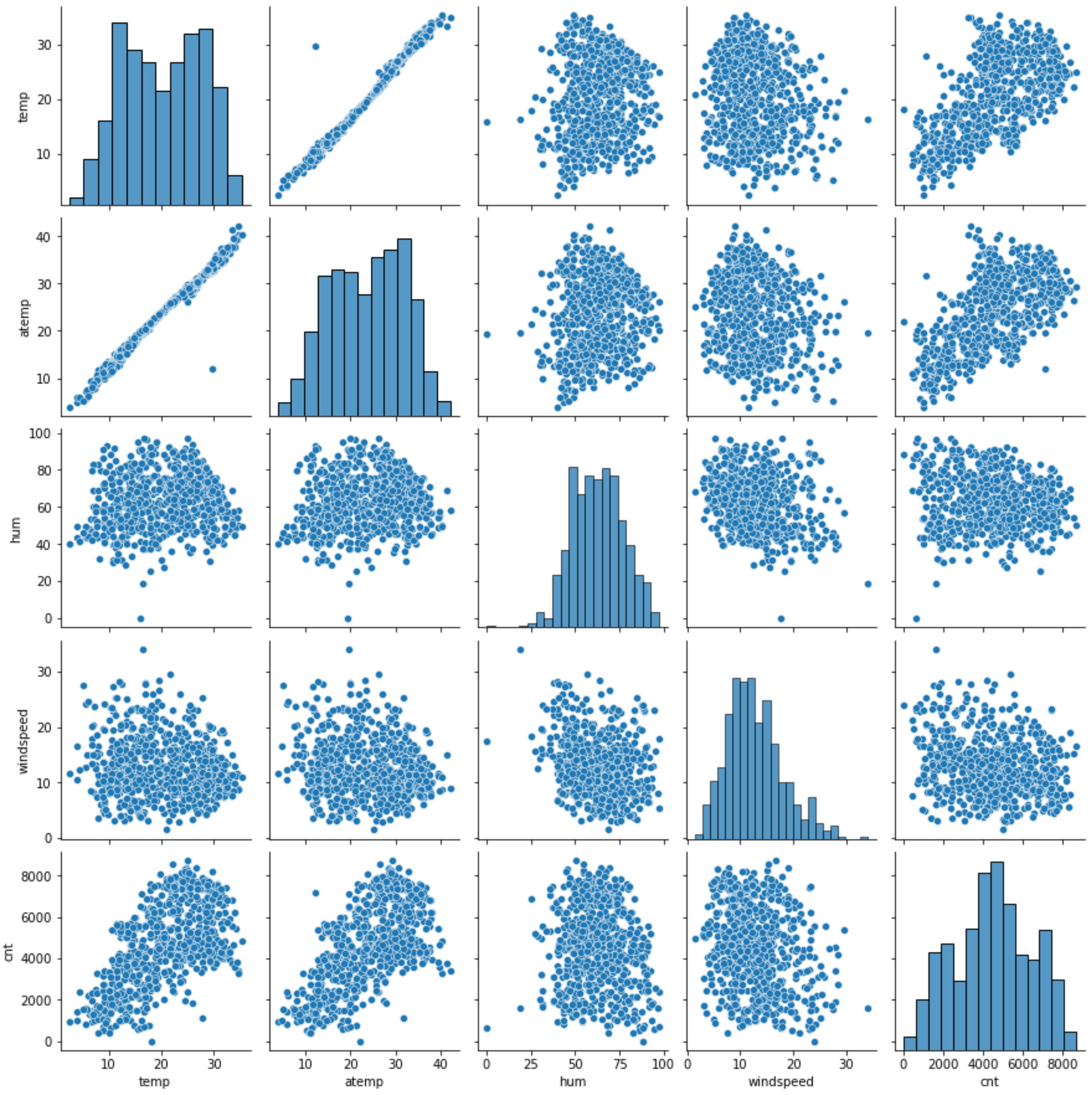


- **2. Why is it important to use drop_first=True during dummy variable creation?**

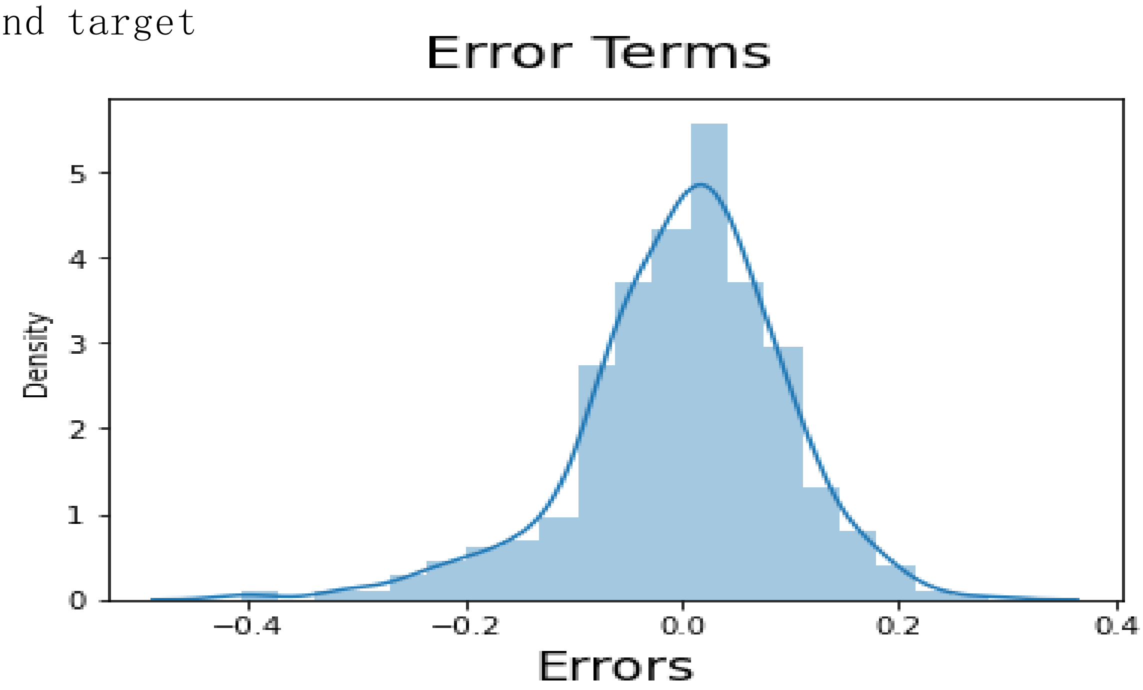
- When we create dummy variables for a particular column let’s say with 3 values, then 3 new columns with either ’0’ or ’1’ will be created for each value. By using drop_first=True, it helps us remove an unnecessary column as information about 1st column can be still inferred by information present in remaining 2 columns. This may have a negative impact on some models, and the effect is amplified when the cardinality is low. Iterative models, for example, may have difficulty convergent, and lists of variable importance may be distorted.More importantly it helps in reducing multicollinearity among predictor variables.

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

- temp and atemp had same and highest correlation with ’cnt’ variable.



- **4. How did you validate the assumptions of Linear Regression after building the model on the training set?**
- We performed residual analysis, to check whether the residuals were normally distributed or not. (It was normally distributed around mean 0.0)
- We checked for VIF (Variance Inflation Factor) and made sure to keep it below threshold value of 5 for the selected predictor variables.
- We checked whether or not there is linear relationship between predictor variables and target



- **5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
- [yr] - 0.230358
- [temp]- 0.574865
- [weathersit_3(light snow & Rain)] - -0.309405
- Following are the top 3 predictor variables with their respective coefficients.

- General Subjective Questions
- **1. Explain the linear regression algorithm in detail.**
- Linear Regression is a supervised Machine Learning approach for numeric value prediction. The most fundamental type of regression analysis is linear regression. The most widely used predictive analysis model is regression.
- The popular equation ″y = mx + c″ underpins linear regression.
- It presupposes that the dependent variable(y) and the predictor(s)/independent variable have a linear relationship (x). The best fit line, which defines the relationship between the independent and dependent variables, is calculated in regression.
- When the dependent variable is a continuous data type, regression is used, and the predictors or independent variables can be of any data type, such as continuous, nominal, or categorical. The regression method seeks out the best-fit line that depicts the connection between the dependent and predictor variables with least errors.
- The output/dependent variable is a function of the independent variable, the coefficient, and the error term in regression.
- Simple linear regression and multivariate linear regression are two types of regression.
- SLR (Simple Linear Regression) is utilised when only one independent variable is used to predict the dependent variable.
- When the dependent variable is predicted using numerous independent variables, Multiple Linear Regression (MLR) is utilised.
- The equation for MLR will be:

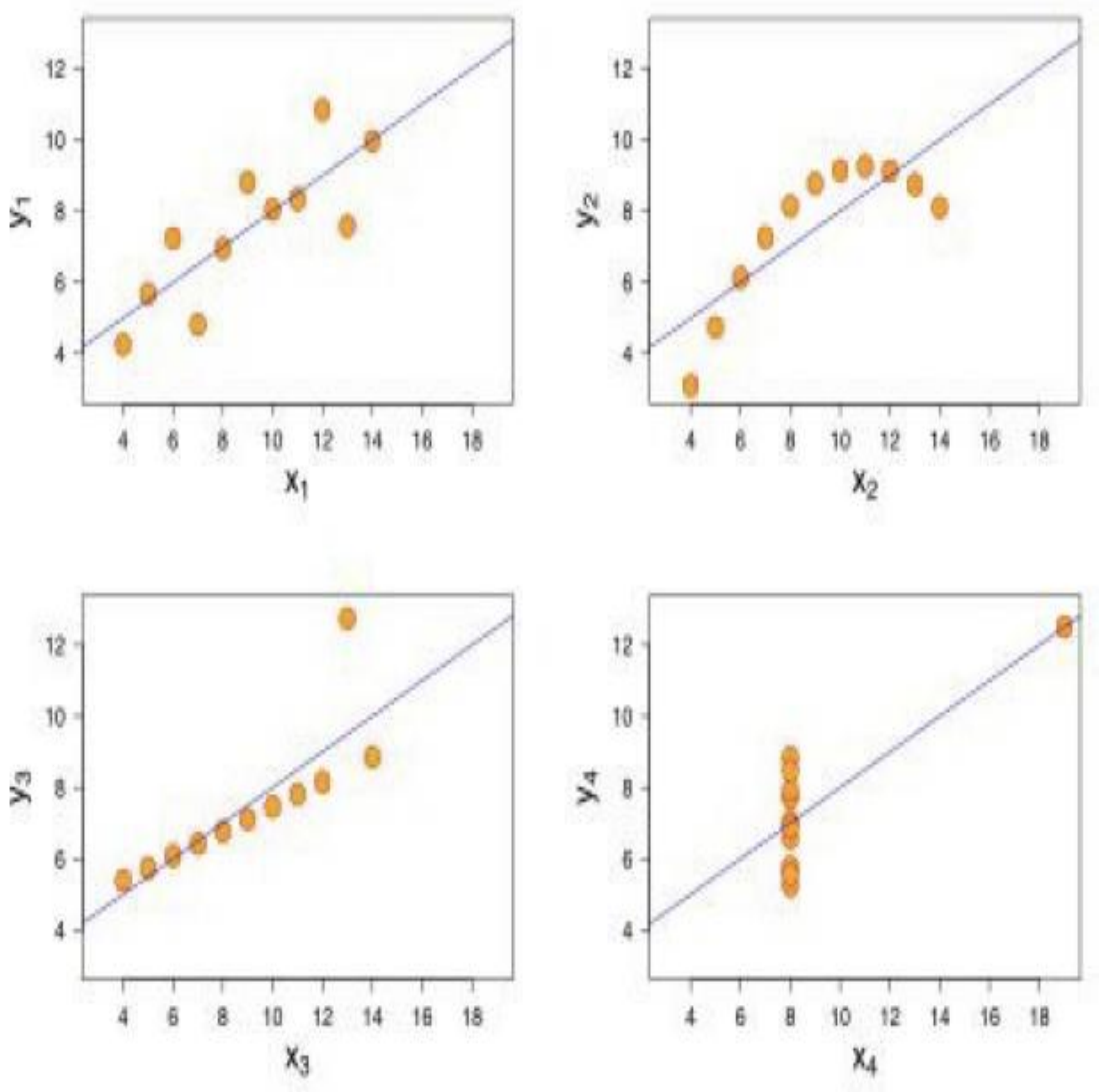
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

- β1 = coefficient for X1 variable β2 = coefficient for X2 variable
- β3 = coefficient for X3 variable and so on…

- β0 is the intercept (constant term).
- t variables.

2. Explain the Anscombe’ s quartet in detail.

- Francis Anscombe, a statistician, devised Anscombe' s Quartet. It has four data sets with nearly equal statistical characteristics, but each has a drastically distinct distribution and appears on a graph in a completely different way. It was created to emphasise the significance of charting data before analysing it, as well as the impact of outliers and other significant observations on statistical features.
- The top left scatter plot looks to show a straightforward linear relationship.
- The second graph (top right) is not normally distributed; while a relationship exists between them, it is not linear.
- The distribution in the third graph (bottom left) is linear, but the regression line should be different.
- The estimated regression is thrown off by one outlier, which has a large enough impact to reduce the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) illustrates how one high-leverage point can yield a high correlation coefficient even when the other data points show no association between the variables.



3. What is Pearson’s R?

- Pearson’s r is a numerical representation of the strength of the linear relationship between the variables. Its value ranges from -1 to +1. It depicts the linear relationship of two sets of data. In layman’s terms, it asks if we can draw a line graph to represent the data.
- r = 1 indicates that the data is perfectly linear and has a positive slope. r = -1 denotes that the data is perfectly linear and has a negative slope. r = 0 indicates that there is no linear relationship.

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

- Feature scaling is a technique for normalising or standardising the range of independent variables or data features. To deal with varying values in the dataset, it is performed during the data preprocessing stage. If feature scaling is not performed, a machine learning algorithm will tend to weight greater values as higher and consider smaller values as lower, regardless of the units of the values.
- Normalization is commonly used when you know your data’s distribution does not follow a Gaussian distribution. This can be useful in algorithms such as K-Nearest Neighbors and Neural Networks that do not assume any data distribution.
- In circumstances where the data follows a Gaussian distribution, on the other hand, standardisation can be beneficial. This, however, does not have to be the case. Standardization, unlike normalisation, does not have a bounding range. As a result, even if your data contains outliers, normalisation will have no effect on them.

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

- The variance inflation factor (VIF) indicates how much collinearity inflates the variance of the coefficient estimate. (VIF) is equal to 1/(1-R 12). VIF = infinity if there is perfect correlation. Where R-1 is the R-square value of the independent variable for which we want to see how well it is explained by other independent variables. If that independent variable can be perfectly explained by other independent variables, then it has perfect correlation and its R-squared value is 1. As a result, VIF = 1/(1-1) provides VIF = 1/0, which is ″infinity.″

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

- The quantiles of the first data set are plotted against the quantiles of the second data set in a q-q graphic. It's a tool for comparing the shapes of different distributions. A scatterplot generated by plotting two sets of quantiles against each other is known as a Q-Q plot. If both sets of quantiles came from the same distribution, the points should form a relatively straight line.
- The following questions are answered using the q-q plot:
- Are the two data sets from populations with similar distributions?
- Is there a common location and scale between two data sets?
- Do the distributional forms of two data sets resemble each other?
- Do the tails of two data sets behave similarly?