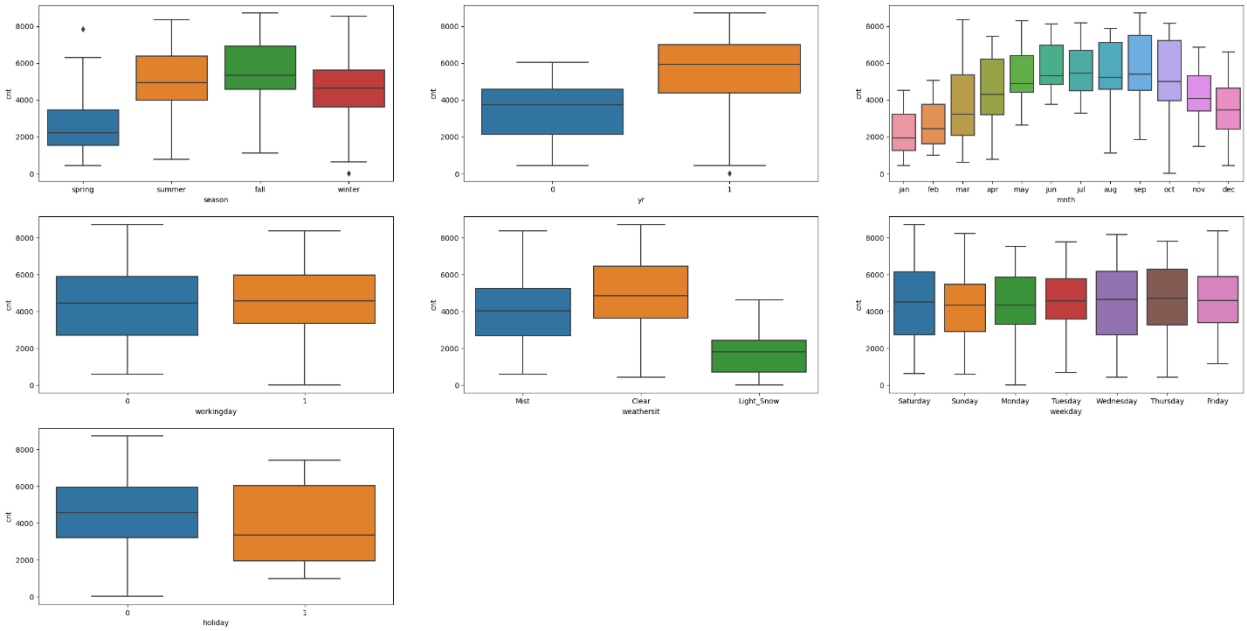
Assignment-based Subjective Questions

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



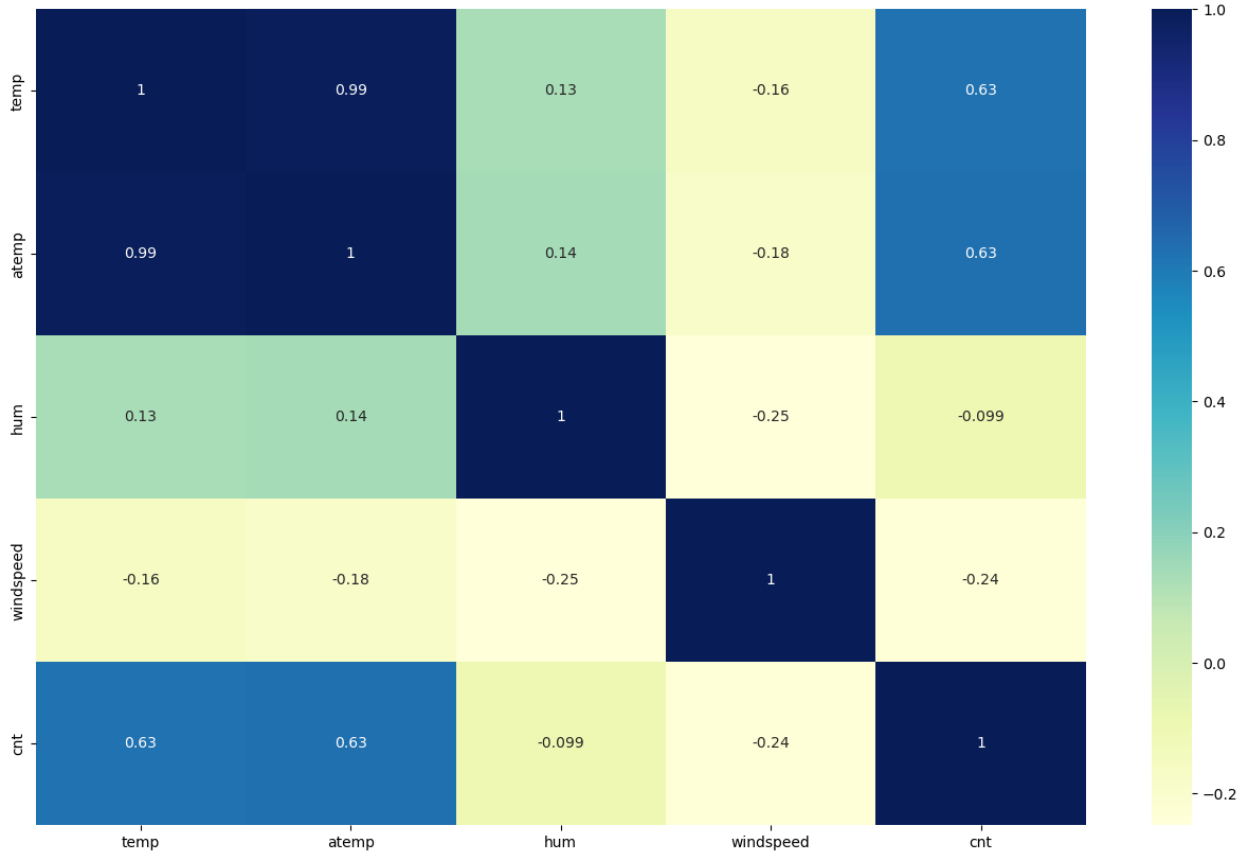
While studying the categorical variable, following inferences were made:

- Bike demand is higher during summer and fall
- Bike demand was higher for 2019 than 2018
- Average Bike demand is higher when it is not a holiday
- Average bike demand is higher during months from May to Oct

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

This is used to reduce co-relation among dummy variables. When drop_first is True, it drops the first column to avoid redundancy, so for n possible values only n-1 columns are created

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



Here, temp and atemp has the highest correlation with the cnt variable

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

A: The assumptions of Linear Regression include:

- Linear Relationship: This was validated through correlation and pair plot
- Multivariate Normality: This was done through plotting of residuals which followed a normal curve
- No Multicollinearity: VIF was checked and multicollinear data points were removed
- Homoscedasticity: We plotted normal distribution of error teams

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

A: Top 3 features that significantly impacted are:

a. Year
b. Weather situation (Light snow)
c. Season (Spring)

General Subjective Questions

1. *Explain the linear regression algorithm in detail. (4 marks)*
   Linear regression is supervised machine learning model which used to predict the dependent variable using independent variable assuming correlation between variables such multicollinearity does not exist among variables. It follows the equation:
   Y = mx +c
   Where , y is dependent
   x is independent, c is constant and m is the slope
   The linear regression utilizes feature prioritization by identification of signification independent variables
   Here, the residuals also follow the normal curve

2. *Explain the Anscombe's quartet in detail. (3 marks)*
   Anscombe's quartet is a data visualization technique by plotting the descriptive statistics like mean, mode, median, range std dev etc. This is done to avoid situations in which the data set is completely different but produce similar kind on result on plotting regression etc
3. *What is Pearson's R? (3 marks)*
   This is used to depict the strength of correlation whose value lies between +1 to -1 where +1 means strongly positively corelated, 0 means no correlation and -1 means strongly negatively corelated. This does not necessarily indicate causality

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is done to normalize the data in terms of variation in magnitude, units or range to the independent variables.
If scaling is not done then, our model might get biased towards a magnitude and incorrectly model the data. It is preferred as it helps in accelerating algorithmic calculations.
Normalized scaling :The values of a normalized data will always fall between 0 and 1.
Standardized scaling: A standardized data will have a mean of 0 and a standard deviation of 1, but the maximum and minimum values are not constrained by any specified upper or lower bounds.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

A: When multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), or the correlation is perfect such that the R-sq =1,then, the VIF tends to infinity.

6. *What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression*.
   It is also known as quantile plot that helps us understand if the two data sets have come from same population or not. This is used in case of Linear regression to identify if the training and test data belong to the same population or not