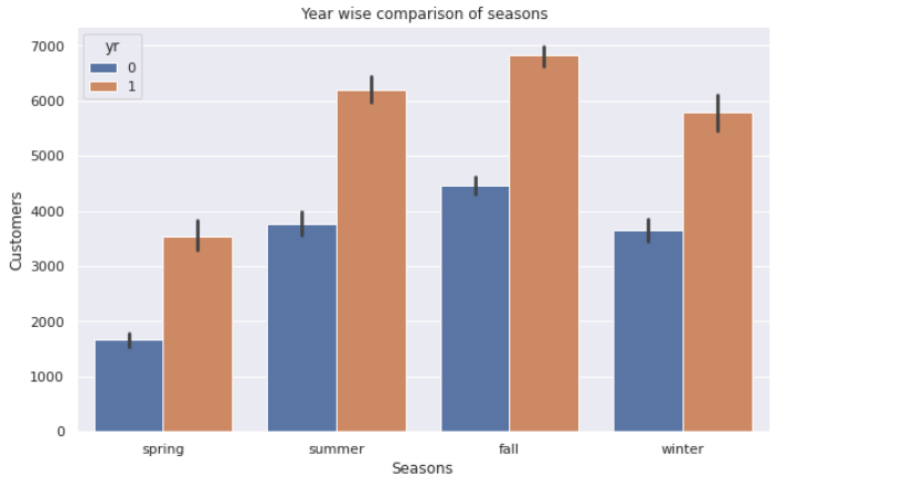
NIDHIN RAJ

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?

Season

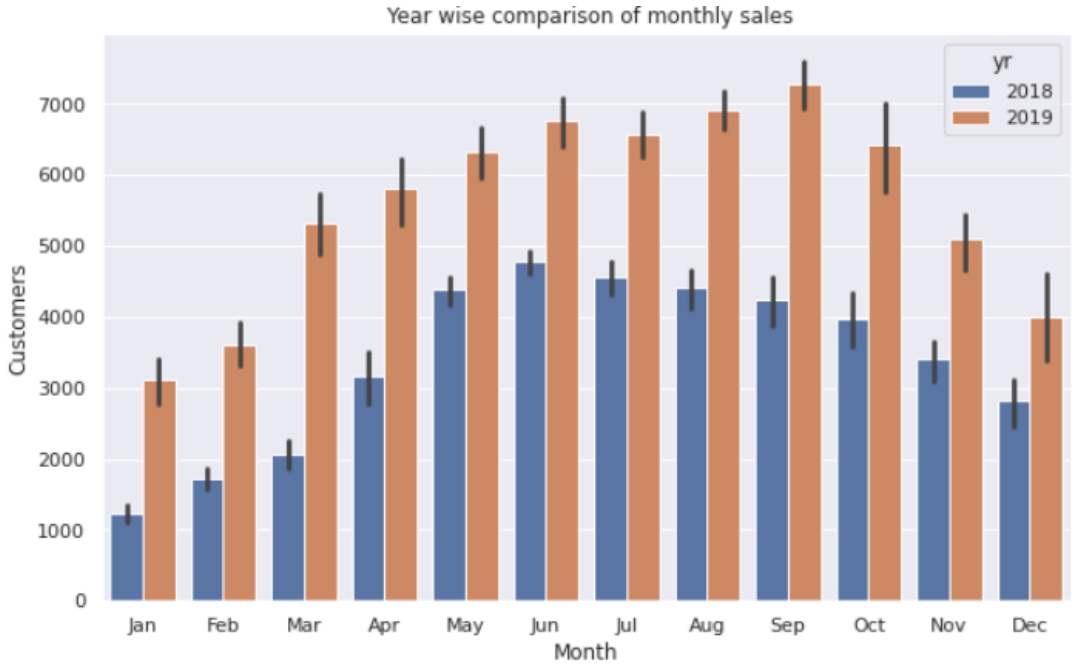
During spring there is less customers and more during fall



Year & Month

As observed above during fall months there is a general increase in demand

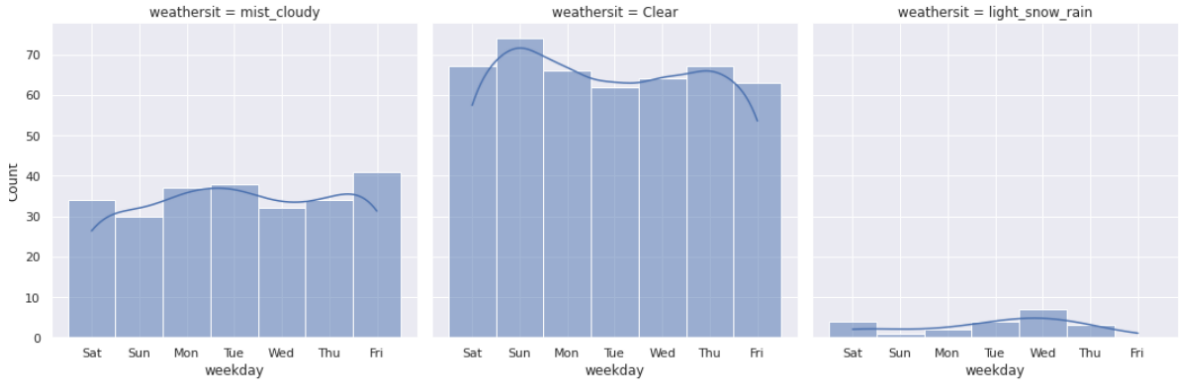
Also we can observe how there is an increase in demand after covid lockdown eased.



Weather Situation & Weekdays

On a clear day obviously the sales is high and on a cloudy day it gets reduced by half

and if there is rain or snow then sales is almost nil



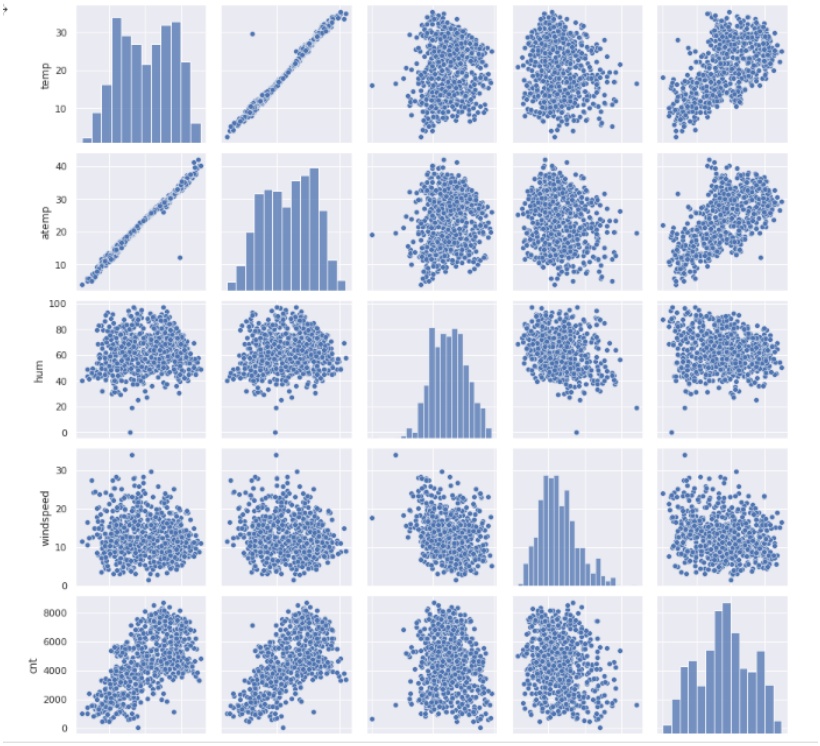
2. Why is it important to use drop_first=True during dummy variable creation?

   During one hot encoding we need only one less variable to actually handle all scenarios.

   Dropping one column will reduce that much processing in memory and thus increasing performance.

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

   Temperature variable had the most linear correlation with target variable “Cnt”



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

   Checked the distribution of error terms to do a residual analysis to make sure its normally distributed.

- Error terms are normally distributed , so linear regression assumption on error terms is correct

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

- Temperature
- Light Snow and Rain
- Windspeed
- Holiday

```
const              0.2597
yr                 0.2340
holiday           -0.1062
temp               0.4502
windspeed         -0.1396
light_snow_rain   -0.2916
mist_cloudy       -0.0831
spring            -0.1102
winter             0.0494
Sun               -0.0479
Jul               -0.0704
Sep                0.0564
--------------------------------
```

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   linear regression is a method to identify relationship between two variables.

   It assumes there's a direct correlation between the two variables and that this relationship can be represented with a straight line.

   It enables us to predict the dependent variable if we know the independent variable.

   The governing equation for linear regression is : y = B* x + A

   Here y is the dependent variable, x is the independent variable, and A and B are coefficients determining the slope and intercept of the equation

   We can use the linear regression equation, with values for A and B, to calculate predictions for each value of x. we also calculate the error for each value of x by subtracting the prediction for that x from the actual, known data.

   Sum the error of all of the points to identify the total error from a linear regression equation using values for A and B.

2. Explain the Anscombe's quartet in detail.
   Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. These datasets is that they share the same descriptive statistics but when we graph them they are represented differently
   - Dataset I is clean and well-fitting linear models.
   - Dataset II is not distributed normally.

- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis.

3. What is Pearson's R?

   The Pearson product-moment correlation coefficient is a measure of the strength of the linear relationship between two variables. It is referred to as Pearson's correlation or simply as the correlation coefficient. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.

   The symbol for Pearson's correlation is "ρ" when it is measured in the population and "r" when it is measured in a sample.

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

   it is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale,

   Normalization rescales the values into a range of [0,1]. This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost.

   Standardization rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1 (unit variance).

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

   If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

   Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the

Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.