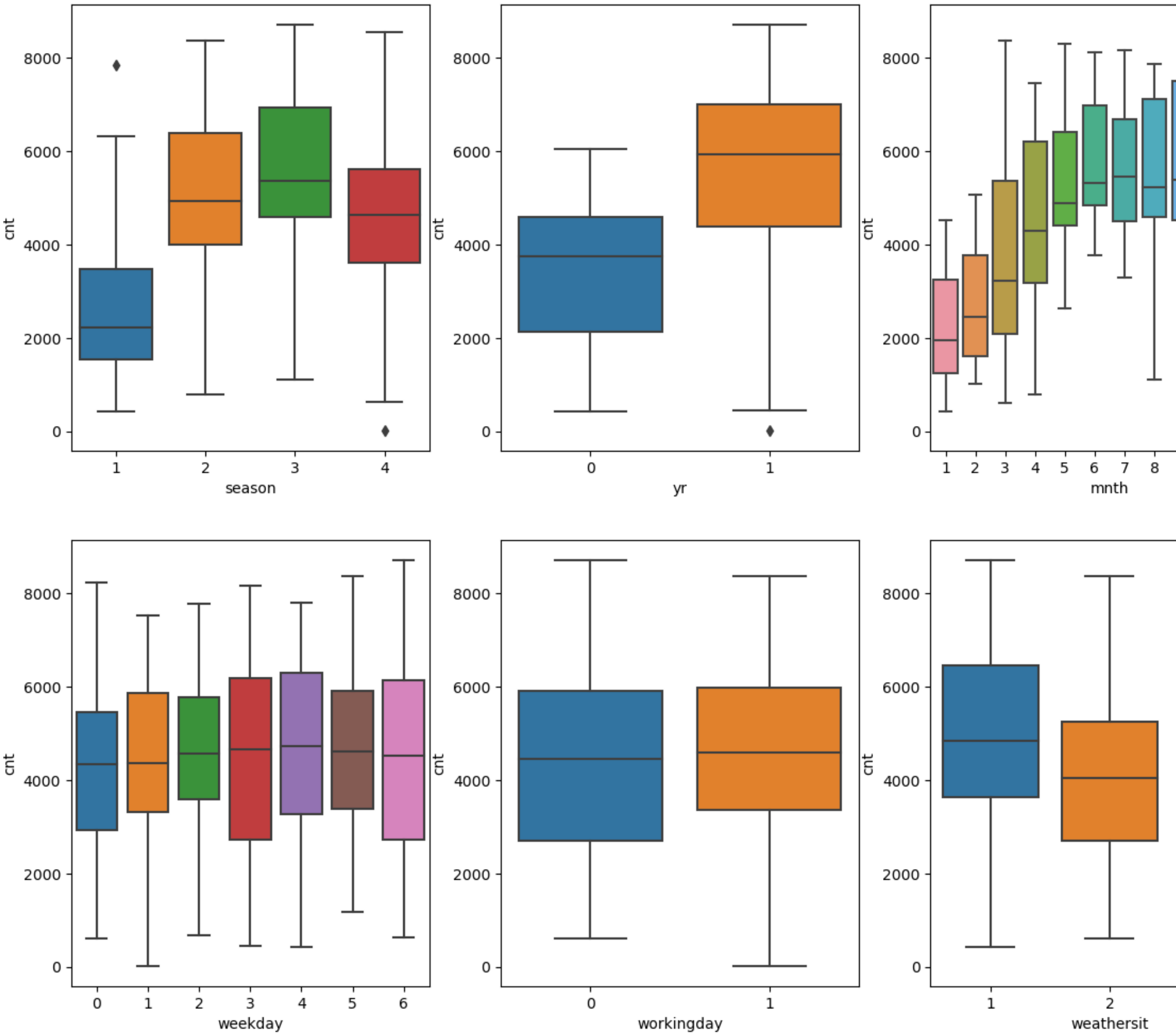
Assignment-based Subjective Questions

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

I have drawn set of box plots to check how **cnt** varies with different set of categorical variables. From the visualization of plot set what I can conclude



1. Season wise we can see median is much higher for 3 (fall) compared to 1 (spring). This indicates that season is a potential predictor variable.
2. year wise a higher trend in 2019 compared to 2018, i.e. there is a year-to-year growth.

3. month wise we see the pick at around 6,7,8,9 with median is more than 5000. Whereas month 1,2,3 and 11,12 saw median less than 4000. This indicates the possibility of month wise booking trend.
4. demand is higher on weekday compared to holidays.
5. A clear day (where weathersit = 1) shows much higher demand compared to rainy/foggy (weathersit = 2) or snowy day/heavy rainy day (weathersit = 3). So weathersit is a potential predictor.

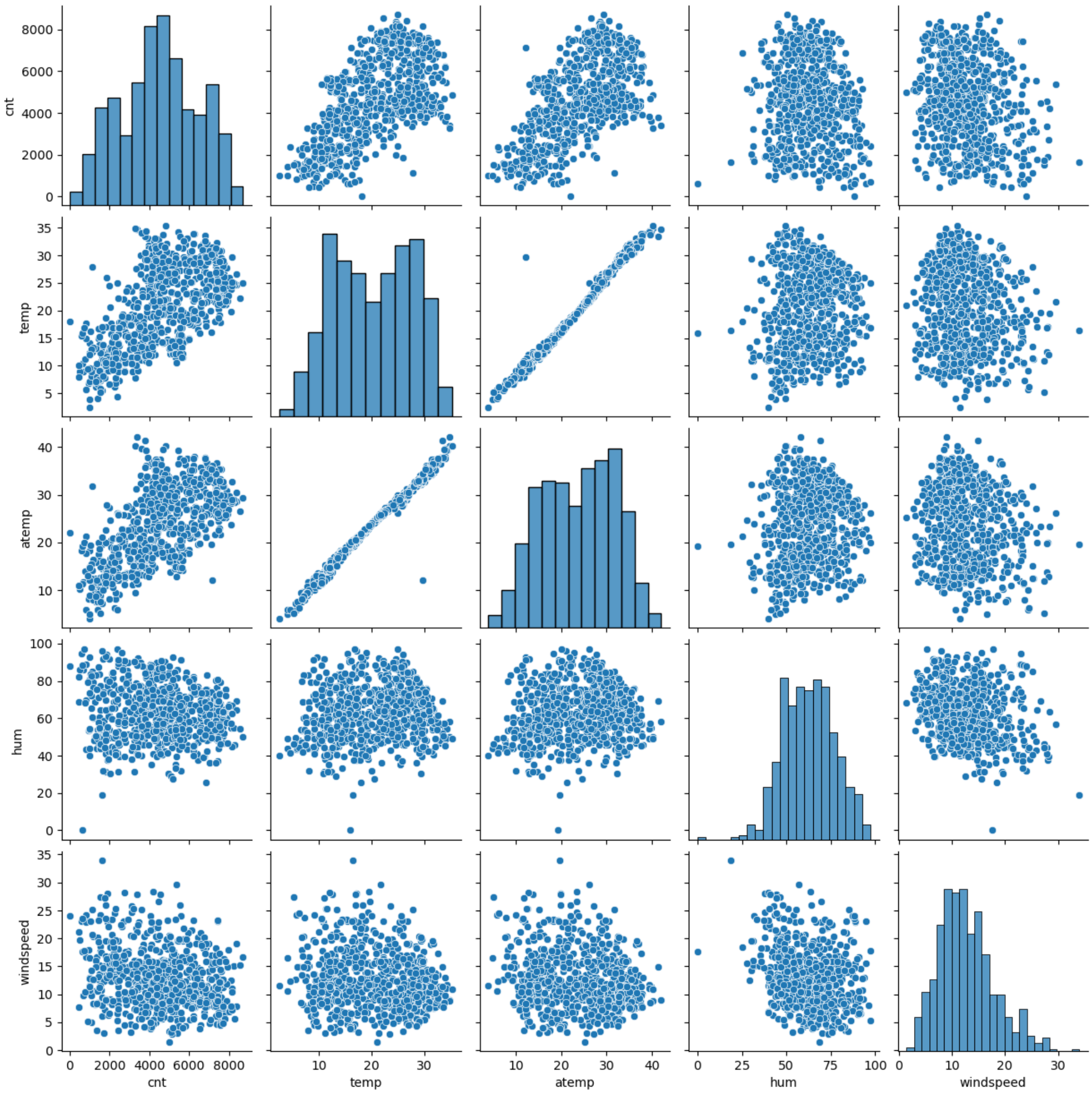
2. WHY IS IT IMPORTANT TO USE DROP_FIRST=TRUE DURING DUMMY VARIABLE CREATION?

Since dummy variable values are represented in 1/0 (Boolean) k-1 number of dummy variables are enough to represent k number of categories. Pandas function get_dummies() create k number of dummy variables for k number of categories. Thus, an explicit parameter drop_First = True must be passed to reduce the number of variables which inevitably makes the model building process faster and simpler.

bike = pd.get_dummies(bike, dtype=int,drop_first=True)

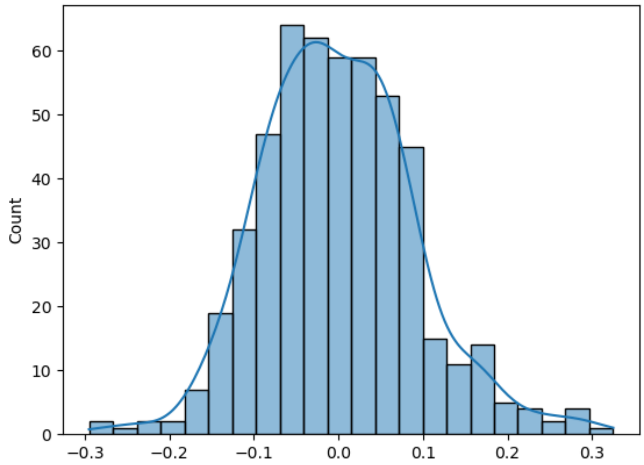
3. LOOKING AT THE PAIR-PLOT AMONG THE NUMERICAL VARIABLES, WHICH ONE HAS THE HIGHEST CORRELATION WITH THE TARGET VARIABLE?

I will say it is variable temp which has the highest corelation with cnt. On the other hand, we can also see huge degree of collinearity between temp and atemp.



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

One of the assumptions of linear regression is that the residual (difference between the actual values and the predicted values) should be normally distributed with a mean at 0. After creating the model, we can calculate the residual and then plot a scatter plot with kde = true using seaborn. If this results a normal distribution, we can successfully claim that the assumption is passed.



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

a. Temp = 0.5568

b. weathersit_3 = -0.2577 [It shows a negative linearity]

c. season_4 = 0.1776

General Subjective Questions

1. Explain the linear regression algorithm in detail.

In linear regression machine learning algorithm, we target to fit a straight line that passes through the maximum number of data points on a target variable.
A best fit linear line equation is y = c + m0 x0 + m1 x1 + m2 x2 + ……..mn xn

Here c is called the intercept and m0,m1,m2 ….these are the slopes or co-efficient for predictor variables x0,x1,x2…..
Y is the target variable.
In linear regression we are supposed to calculate the co-efficient and as well as predict the value of y and the independent variables X.

Step 1 : Understanding of Data
Assumptions of linear regression must be true at every phase of model building phase. In this phase the key assumption is –
Linearity: The relationship between the independent and dependent variables is linear.
If we don't see any linearity, then we can discard the use case from linear regression-based modelling.

Another important assumption at this phase is the independence of observations. Under no circumstances the two values of X should have dependency between them or time driven dependency is not going to work out for linear regression.

Step 2: Model Building

In the model building phase train and test data is splitted and a model is built and trained on the train data set. On train data we can apply different algorithms like ordinary least square or gradient descent methos to obtain the optimal values of the co-efficient and the constant. With the target model it is important to assure that the assumption of normal distribution of errors are validated. During model building we need to assure that the variables must not show a trend of multi collinearity.

Once the model is built, we can apply that on the train data to predict the target variable value and the computed values can be evaluated against different statistical instruments like R-square, adjusted R-square and F statistics. We must watch the p-values of each variable to assess the significance of them on the model.

Step 3: Evaluate model against the test data.

At the final step we apply the model against the test data and predict the values of the target variables. We must check if the model is not overfit by comparing the R-square value on test and train data.

2. EXPLAIN THE ANSCOMBE’S QUARTET IN DETAIL.

Anscombe's quartet is a famous example in statistics that demonstrates the importance of graphing data before analysing it and highlights the limitations of relying solely on summary statistics. It consists of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, and regression line), yet they look very different when plotted.

Anscombe’s quartet indicates that relying on summary statistics can be misleading and thus summary statistics must be compensated by exploratory data analysis and plotting of data. Data visualization can reveal the actual characteristic of data which otherwise is unknown from summary statistics.

3. WHAT IS PEARSON’S R?

Pearson correlation, also known as Pearson correlation coefficient, is a measure of the linear relationship between two continuous variables. It quantifies the degree to which two variables change together in a linear fashion. The Pearson correlation coefficient is denoted by the symbol r.

R value can be from -1 to 1, where -1 shows a negative linearity, 0 means no correlation and 1 says positive collinearity

Formula to calculate r is

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

n is the number of data points

x and y are the data points

4. WHAT IS SCALING? WHY IS SCALING PERFORMED? WHAT IS THE DIFFERENCE BETWEEN NORMALIZED SCALING AND STANDARDIZED SCALING?

Scaling is the process to bring the values of all the independent variables in an unified range.

Scaling is performed to have a clear understanding on the weightage of the independent variables on the target vale or on the model. Without scaling the model may predict the correct value of the target variable but unable to find the weightage of the predictors.

Normalized Scaling

Normalization, also known as Min-Max scaling, rescales the features to a fixed range, typically between 0 and 1. It subtracts the minimum value of the feature and then divides by the range of the feature (maximum value minus minimum value).

Normalized value of X = (x – Min X) /(Max X -Min X)

Standardized Scaling

Standardization, also known as Z-score normalization, rescales the features so that they have a mean of 0 and a standard deviation of 1. It subtracts the mean of the feature and then divides by the standard deviation of the feature.

Standardized value of X = (x – Mean of the data) / standard deviation of the data

5. YOU MIGHT HAVE OBSERVED THAT SOMETIMES THE VALUE OF VIF IS INFINITE. WHY DOES THIS HAPPEN?

The formula to calculate VIF is

VIF$_i$ = 1/(1-R$_i^2$)

Where R$_i$ = R$_i^2$ is the R-square value keeping the ith variable as target variable. For a very strong collinearity R=square can be 1 or very close to 1 and thus VIF can be infinity

6. WHAT IS A Q-Q PLOT? EXPLAIN THE USE AND IMPORTANCE OF A Q-Q PLOT IN LINEAR REGRESSION.

The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.