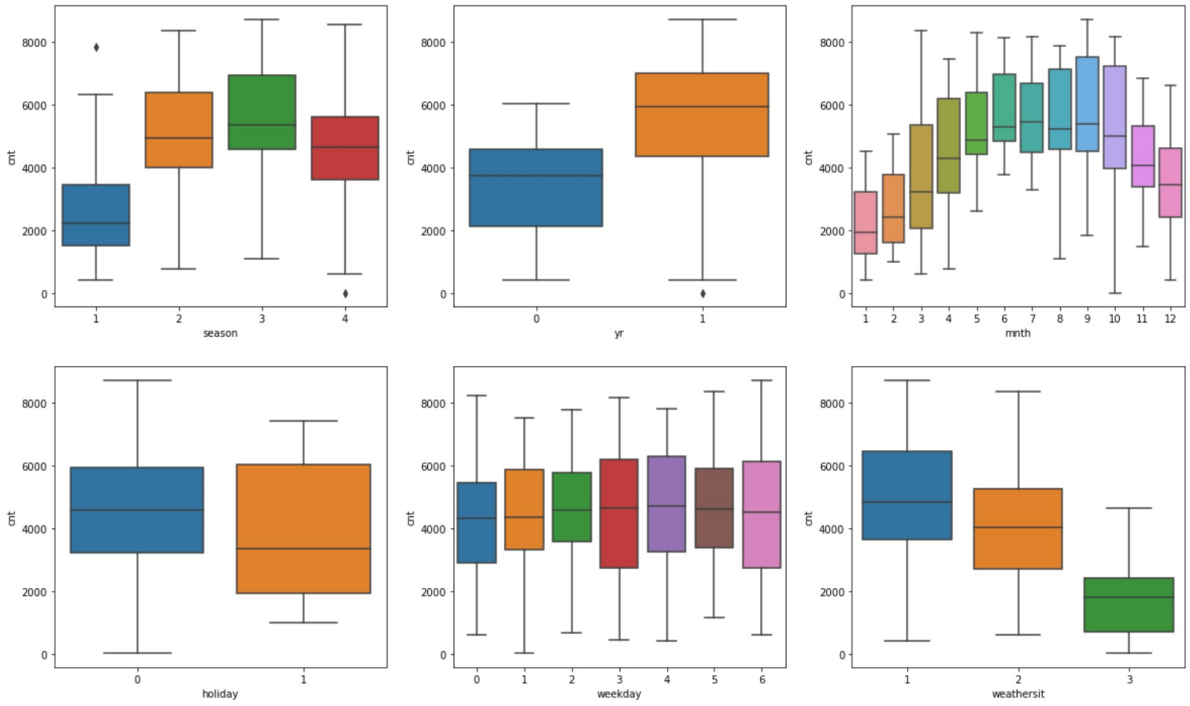
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Inference of Categorical variables:

- Good cnt value is observed during the season-3(Autumn/Fall Season)
- Good cnt value is observed in the year 2019 compare to the previous year 2018. So, business is growing.
- Good cnt value is observed in the months of September and October.
- Holiday is having (97% of value is 0 (Not a Holiday); Only 2.88% of data belong to a holiday.
- Good cnt value is observed on Wednesdays and Thursdays.
- Good cnt value is observed when there is Clear or Partly cloud. No business is observed during Heavy rain.

Screenshot from the study:



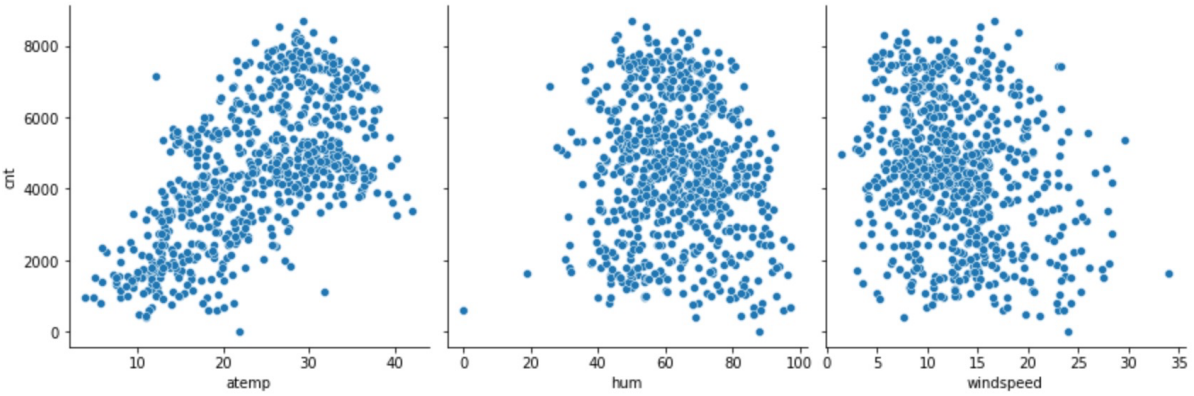
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

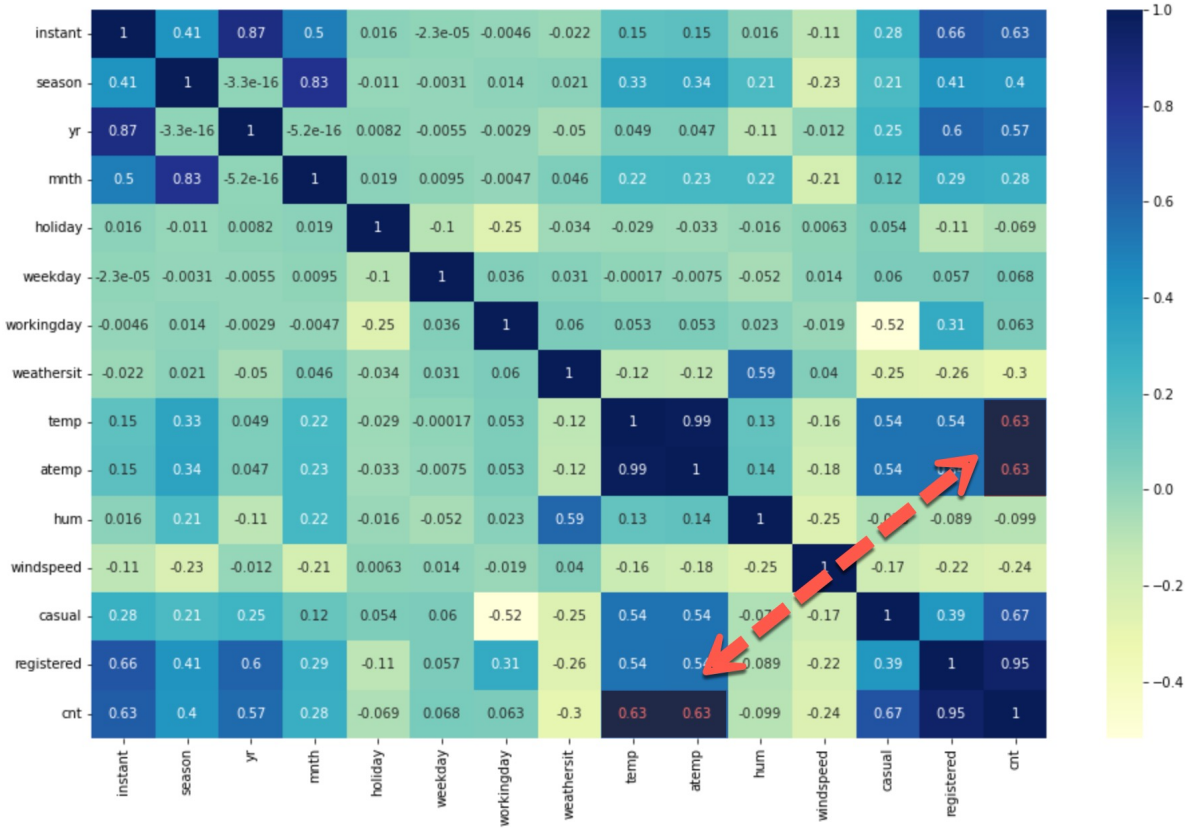
n-1 dummy variables are sufficient to study the *n* dummy variables. Hence the parameter drop_first=True can be set.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Among other continuous variables, *atemp* or *temp* has the highest correlation with the target variable. Both the columns are around 63% correlation with the target variable.

Screenshot from the study:





4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Assumptions of Linear regressions are validated by the following

1) Check the VIF(Variance Inflation Factor)
2) Error distribution of Residuals should be normalised
3) Linear relation between the dependent variable and the target variable.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Top 3 variables which significantly contribute and explain the demand of the shared bikes are temperature or (feel like temperature atemp), year and holiday variables.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression is an ML algorithm used for supervised learning. It helps in predicting a dependent variable(target) based on the given independent variable(s). The regression technique tends to establish a linear relationship between a dependent variable and the other given independent variables.

There are two types of linear regression

(i) Simple Linear Regression
(ii) Multiple Linear Regression.

Simple linear regression is used when a single independent variable is used to predict the value of the target variable. The formula for a simple linear regression is:

$$y = \beta_0 + \beta_1 X + \epsilon$$

y is the predicted value of the dependent variable (y) for any given value of the independent variable (x).
B0 is the intercept, the predicted value of y when the x is 0.
B1 is the regression coefficient – how much we expect y to change as x increases.
X is the independent variable (the variable we expect is influencing y).
e is the error of the estimate, or how much variation there is in our estimate of the regression coefficient.

Multiple Linear Regression is when multiple independent variables are used to predict the numerical value of the target variable. The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

y = the predicted value of the dependent variable
B_0 = the y-intercept (value of y when all other parameters are set to 0)
$B_1 X_1$ = the regression coefficient (B_1) of the first independent variable (X_1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)
… = do the same for however many independent variables you are testing
$B_n X_n$ = the regression coefficient of the last independent variable
ϵ = model error (a.k.a. how much variation there is in our estimate of y)

A linear line showing the relationship between the dependent and independent variables is called a regression line. A positive linear relationship is when the dependent variable on the Y-axis along with the independent variable in the X-axis. However, if dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician *Francis Anscombe* in 1973 to signify both the importance of plotting data before analysing it with statistical properties. It comprises of four data-set that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. The basic thing to analyse about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behaviour irrespective of statistical analysis.

3. What is Pearson's R? (3 marks)

Pearson's Correlation Coefficient (R) is used to establish a linear relationship between two quantities. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between -1 and +1

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

Pearson correlation coefficient (*r*)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

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

Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range.

The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.

The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

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

The value of VIF is infinite when there is a perfect correlation between the two independent variables. The R^2 value is 1 in this case. This leads to VIF infinity as VIF equals to $1/(1- R^2)$. This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Quantile-Quantile (Q-Q) plot are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.