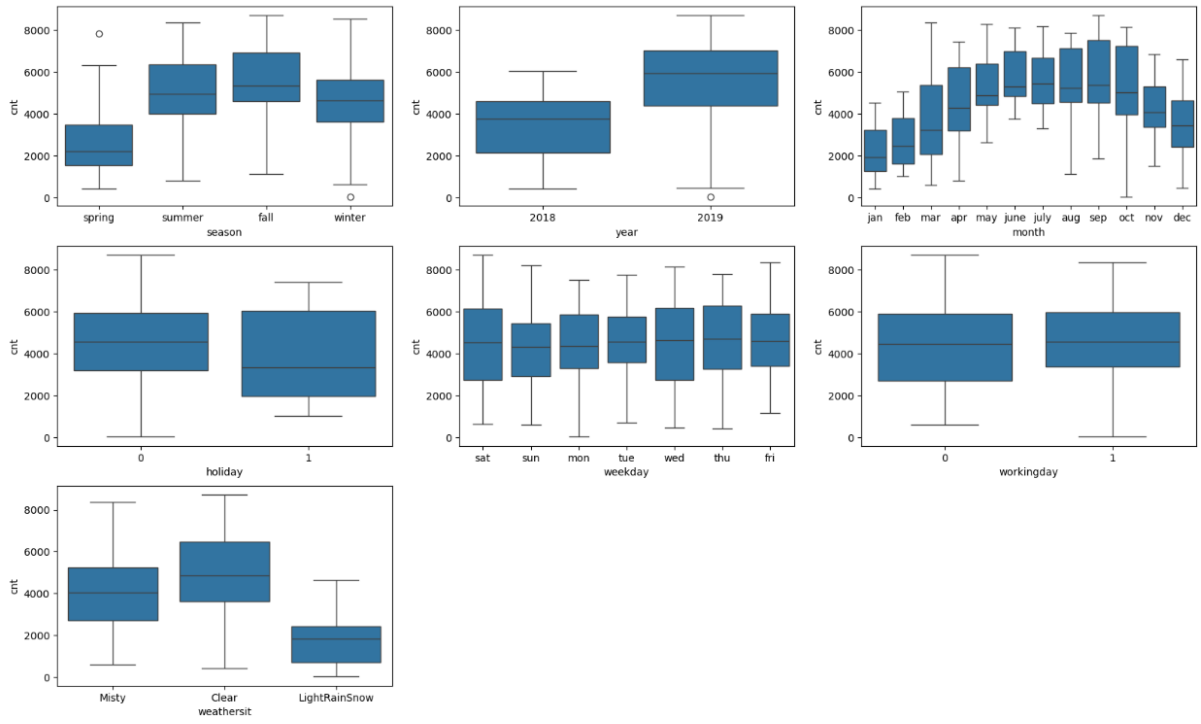
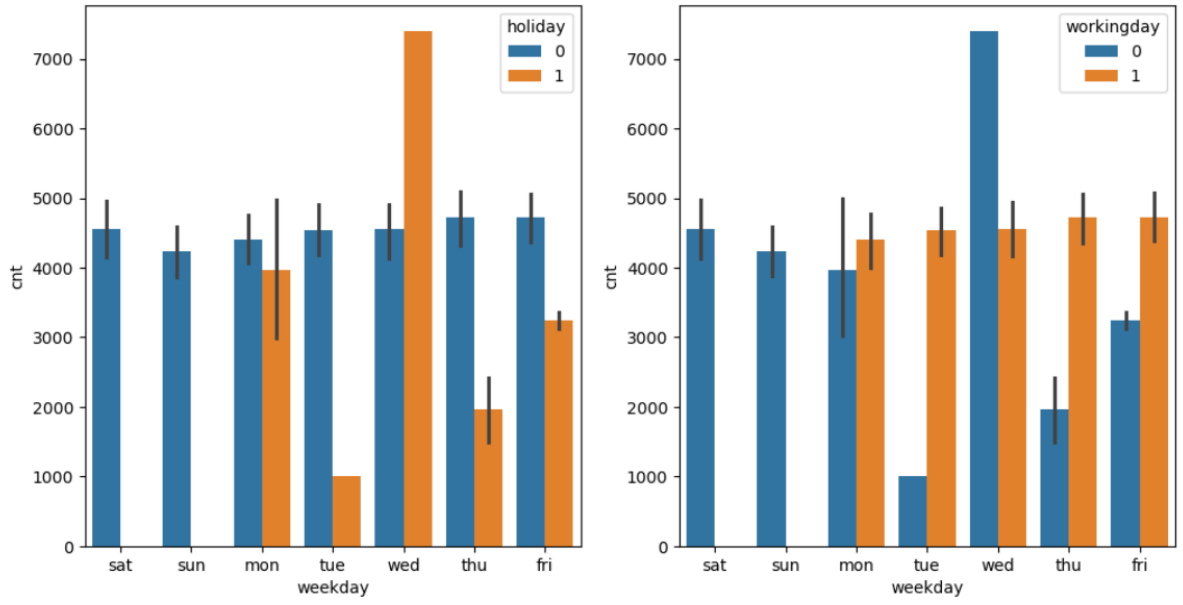
Assignment-based Subjective Questions

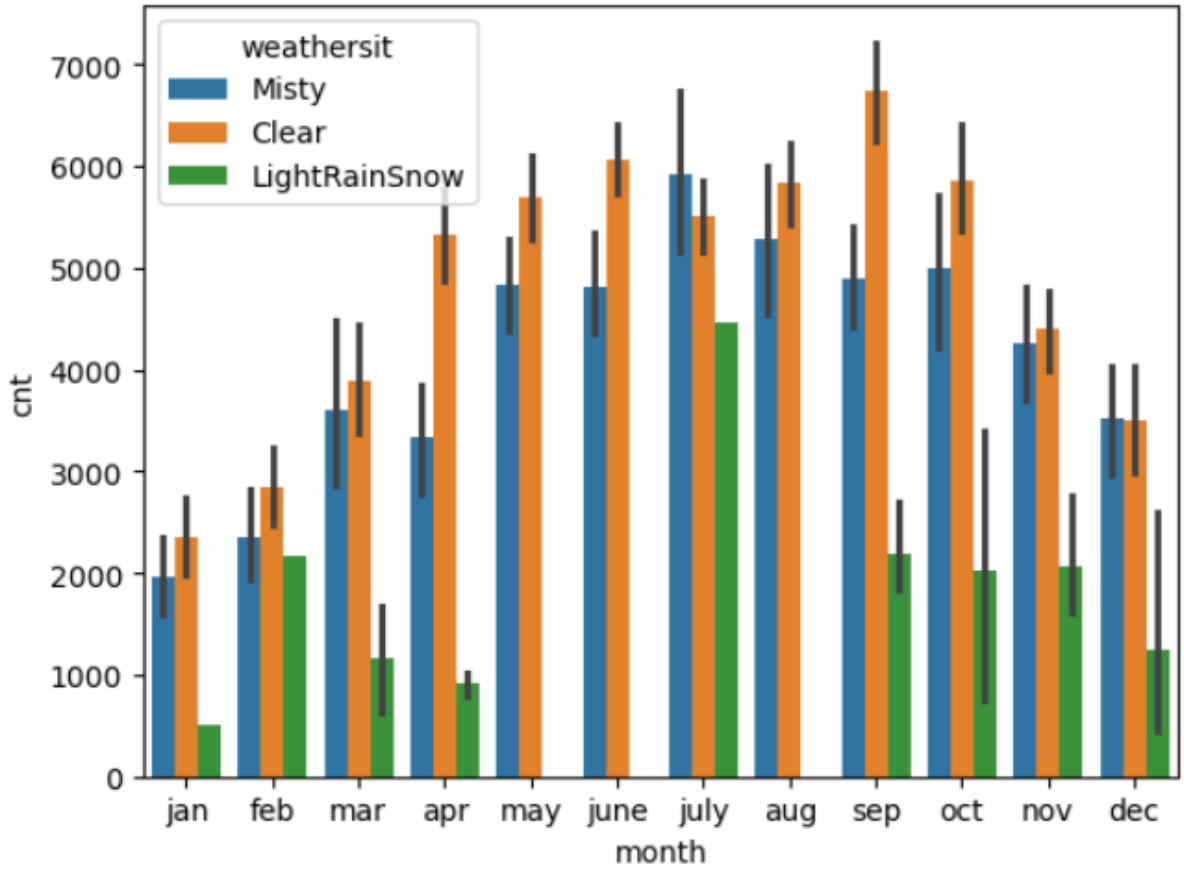
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
   Answer:
   From my EDA analysis on the categorical variables from the dataset, below are the inferences from the charts plotted:

   - Overall, 2019 had more bookings than 2018
   - Fall season had more bike bookings compared to other seasons.
   - More bookings are done during the the months of June, August, September and October months.
   - During “Clear” weather, bookings are more as expected.
   - During weekdays, on “Wednesday” being a “holiday” have more number of bookings than any other “non working day”.
   - “Clear” and “Misty” days are more during July, August , September and October, which has attracted more number of bookings.

   Graphs supporting above statements:



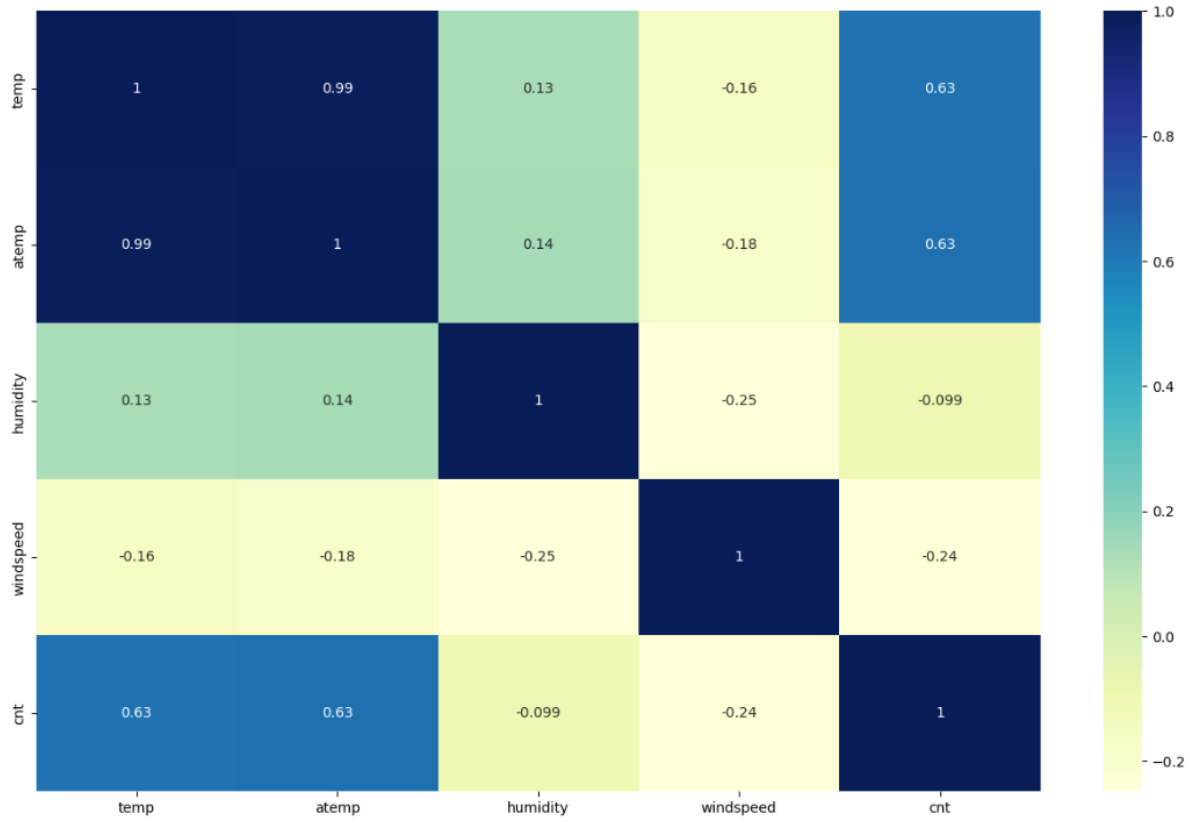




2. **Why is it important to use drop_first=True during dummy variable creation?**
   Answer:
   In case of dummy variables for categorical data with “n” levels, we create “n-1” new columns to depict the data of all levels using 0’s and 1’s in combination.
   drop_first=True helps in reducing the extra column created during dummy variables creation, which reduces the correlations created among dummy variables.

For Example:
If we have 3 types of values in categorical column and we want to create dummy variables for that column. Then we would create two columns as absence of value 1 in both the columns indicate the presence of third value among the categorical values of original column.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Answer:
   "temp" and "atemp" have highest correlation with the target variable "cnt".



4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Answer:
   I have validated the assumptions of Linear Regression after building the model on the training set as below:

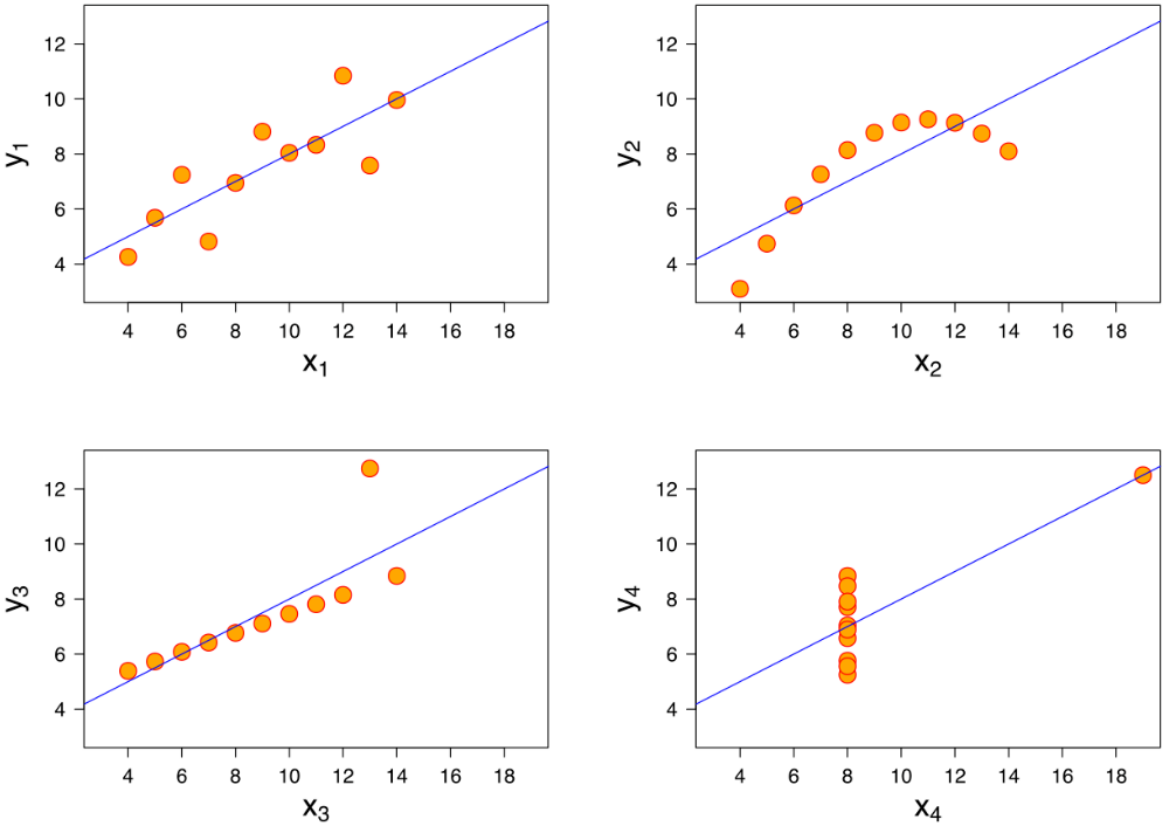
   - Normality of errors → Error terms should be normally distributed.
   - Multicollinearity → There should be insignificant multicollinearity among the variables.
   - Homoscedasticity → There should be no visible pattern among the residuals.
   - Independence of residuals → No auto correlation
   - Linear Relationship validation → Linearity should be visible among variables.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Answer:
   Below are the top 3 features:
   - year
   - temp
   - workingday

General Subjective Questions:

1. **Explain the linear regression algorithm in detail.**
   Answer:
   Linear regression is a form of predictive modelling technique which tells us the relationshop between a dependant (target variable) and independent variables (predictors). Since linear regression shows the linear relationship , which means it finds how the value of the dependant variable is changing according to the value of the independent variable. If there is a single input variable(x), such linear regression is called simple Linear Regression.
   And if there are more than one input variables, such linear regression is called Multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.
   A regression line can be a Positive linear relationship or a Negative linear relationship. The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line having least errors.
   In Linear regression, RFE or Mean Squared Error(MSE) or cost function is used, which helps to figure out the best possible values of a0 and a1, which provides the best fit line for the data points.

2. **Explain the Anscombe's quartet in detail.**
   Answer:
   Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing the data when analysing it , and the effect of outliers and other influential observations on statistical properties.



- The first scatter plot appears to be a simple linear relationship, corresponding to two correlated variables, where y could be modelled as gaussian with mean linearity dependant on x.
- For the second graph, while the relationship between two variables is obvious , its is not linear and Pearsons correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph, the modelled relationship is linear, but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph shows an example when one high leverage point is enough to produce a high correlations coefficient , even though the other data points do not indicate any relationship between them.

3. What is Pearson's R?

Answer:

The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

- Pearson's r
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

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.

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.

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

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

Scaling is a geometric change that linearly enlarges or reduces things. A property of objects or rules known as scale invariance is that they remain unchanged when scales of length, energy, or other variables are multiplied by a common factor.
Scaling law, a law that explains how many natural phenomena exhibit scale invariance.

Scaling is performed because:
It is a data pre-processing procedure used to normalize data within a specific range by applying it to independent variables. Additionally, it aids in accelerating algorithmic calculations. The majority of the time, the obtained data set includes characteristics that vary greatly in magnitudes, units, and range.

The difference between normalized scaling and standardized scaling:
The values of a normalized dataset will always fall between 0 and 1. A standardized dataset 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?**
   Answer:
   If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.

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

   This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

   Few advantages:
   a) It can be used with sample sizes also
   b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

   It is used to check following scenarios:

   If two data sets —
   i. come from populations with a common distribution
   ii. have common location and scale
   iii. have similar distributional shapes
   iv. have similar tail behaviour

   Interpretation:

   A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.

Interpretation of Q-Q plot

- If the points on the plot fall approximately along a straight line, it suggests that your dataset follows the assumed distribution.
- Deviations from the straight line indicate departures from the assumed distribution, requiring further investigation.