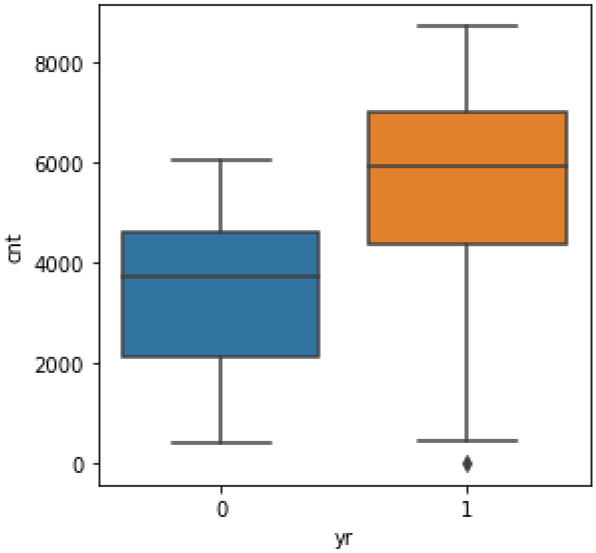
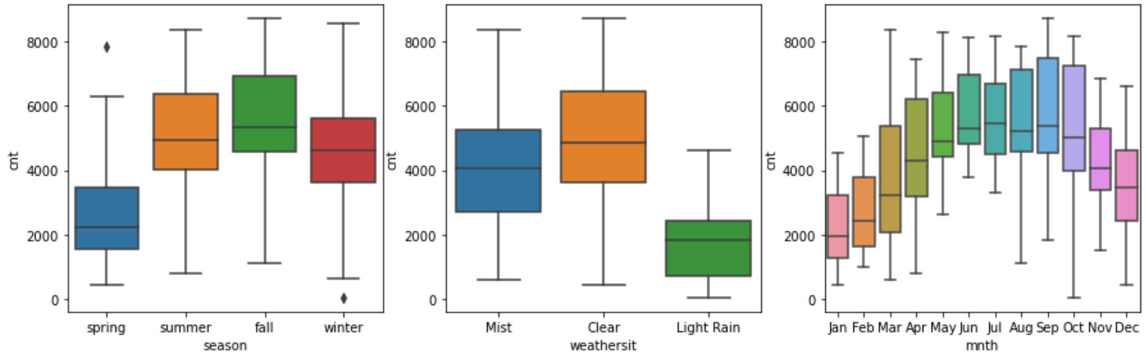
Assignment-based Subjective Questions

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

A: The categorical variables yr, weathersit, season and month have an impact on the dependent variable. yr variable have highest impact.





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

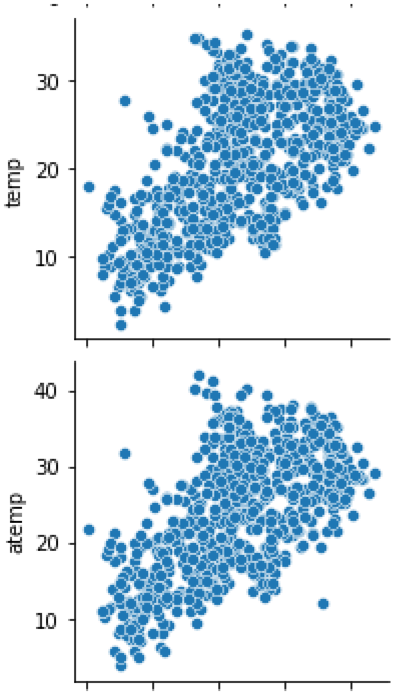
A:

Pandas get_dummies function generates one variable per label.

For n labels, we only require n-1 columns. Therefore, we have to drop first column corresponding to first label.

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

A: Variable atemp (temp) has highest correlation with target variable.



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

A:

Assumption validation of linear regression model are:

- Linear Relationship : Using scatter plot as shown in python notebook.
- Absence of Multicollinearity: Heatmap and VIF.
- Residual analysis: The residuals follow the normal distribution with a mean 0.

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

A:

Based on the final model, top 3 features are

1. `temp, 0.59`
2. `weathersit_Light Rain, -0.2318`
3. `yr, 0.23`

General Subjective Questions

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

A:

There is a linear relationship between independent and dependent variable.

Assumption validation of linear regression model are:

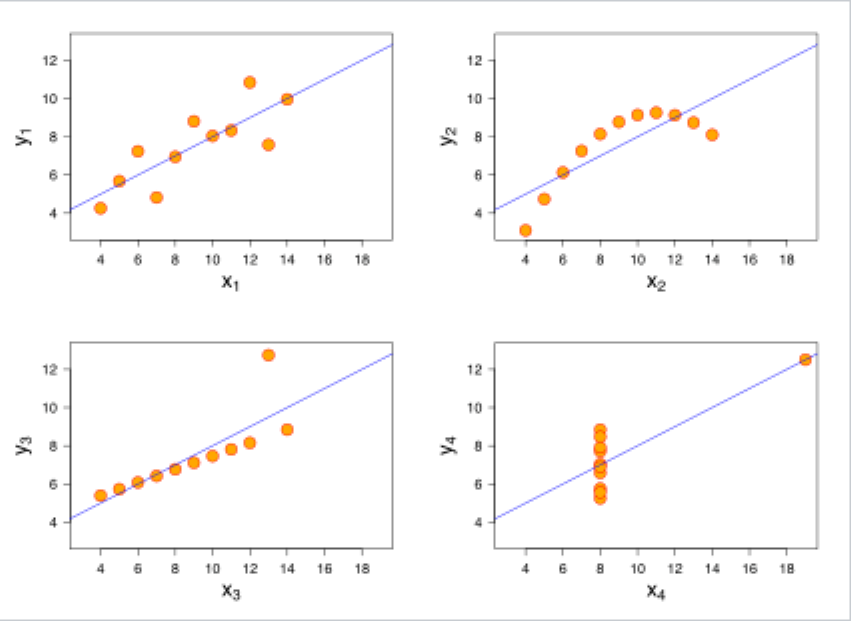
- Linear Relationship between independent and dependent variable.
- Absence of Multicollinearity.
- The residuals (error terms) follow the normal distribution with a mean 0.
- no correlation between error terms

Linear regression model steps:

1) Understanding data
2) Data prepartion
3) univariate and bivariate analysis
4) splitting the test and train set
5) scale data wherever necessary
6) build model iteratively automated and manual for best results
7) Evaluation: check assumptions are followed.
8) make predictions and adjusted R2 of test data

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

A: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

3. What is Pearson’s R? (3 marks)

A:

It measures the strength of the linear relationship between two variables and is always between -1 and 1.

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

A:

Scaling changes the range of values, but without changing the shape of distribution.

It makes model interpretation easier.

1) Normalization: range 0 to 1: where 1 represent max value and 0 min value of training set.
   $x=x-min(x)/(max(x)-min(x))$
2) Standardization:
   $x=x-mean(x)/sd(x)$

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

A:

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

(3 marks)

A: