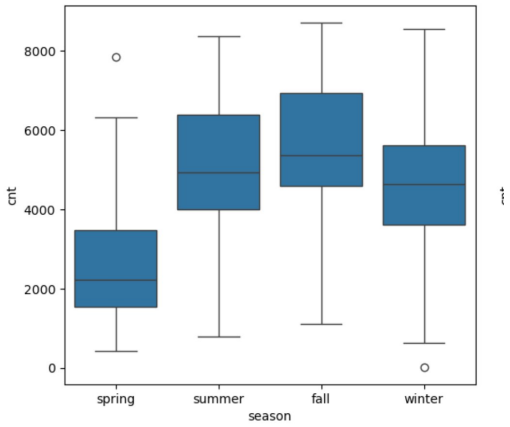
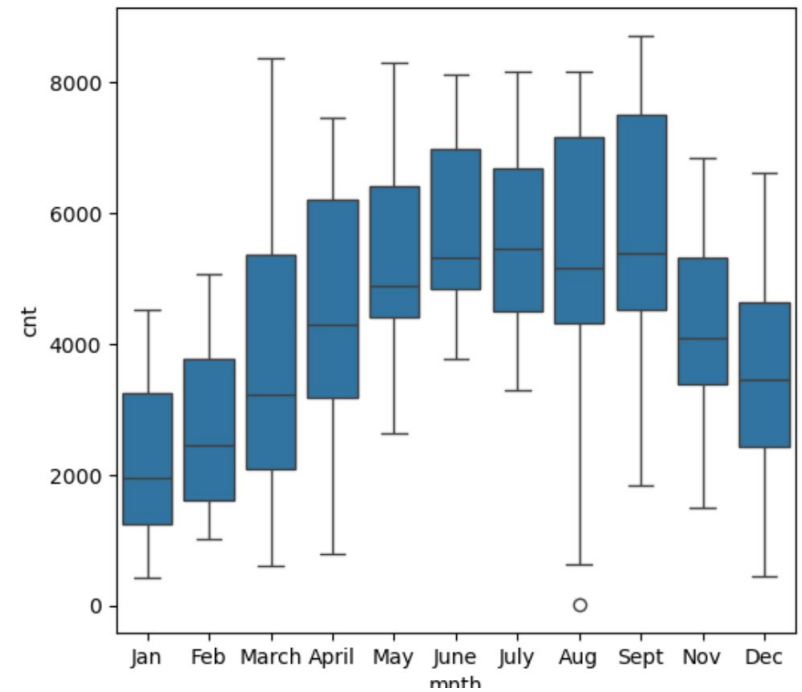
Assignment-based Subjective Questions

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



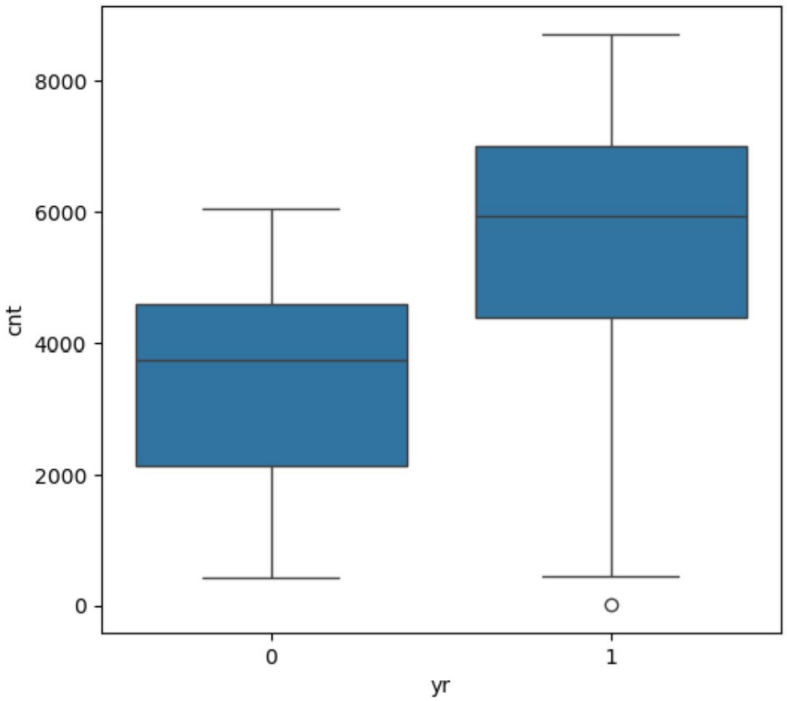
This indicated that services are availed more in fall and summer seasons compared to winter season with total count going further down in spring season.

- **Month**



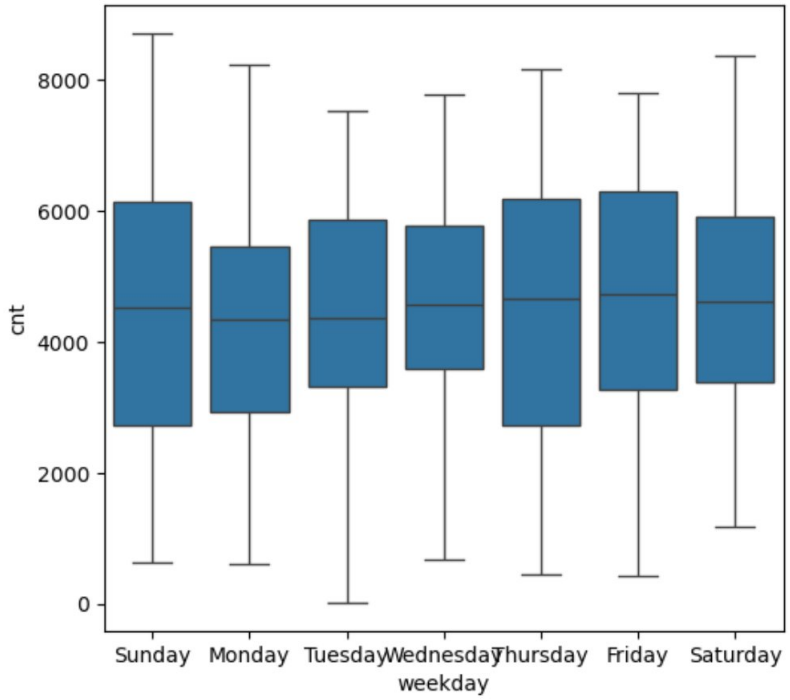
Total counts for each month also shows that the months of summer and fall are the months where people are using shared bike more compared to other months.

- **Year**



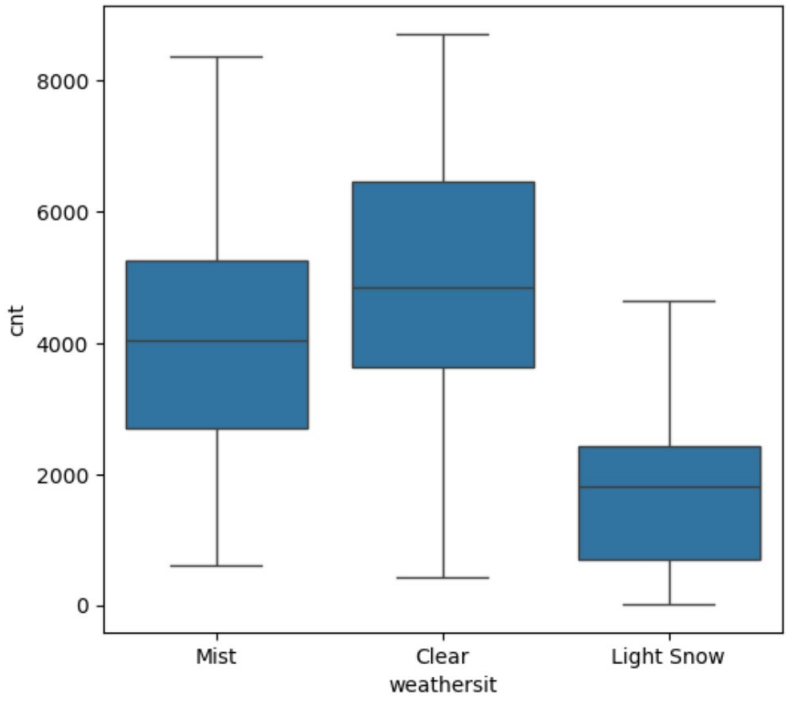
Year 2019 shows significant increase in total no of counts.

- **Weekdays**



No significant difference with all weekdays.

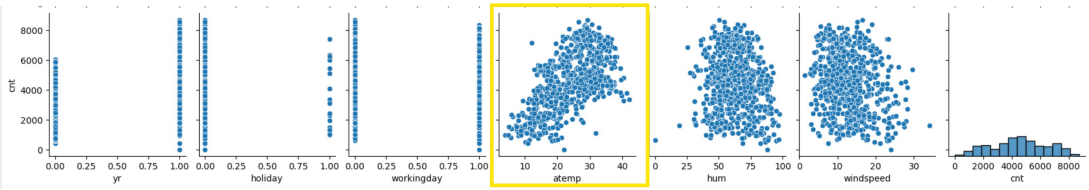
- **Weather**



People are using bikes when there is clear weather or have mist. People are not using bikes if there is snow because of danger of slipping.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
    - Using drop_first=True during dummy variable creation is important because it avoids the "dummy variable trap," a problem of perfect multicollinearity. When a categorical variable with n categories is converted into dummy variables, normally n binary columns are created. However, these columns are linearly dependent (one can be exactly predicted from the others), which can confuse regression models.
    - By setting drop_first=True, one dummy variable is dropped (usually the first category). This reduces redundancy and prevents multicollinearity, improving model stability and interpretability without losing information since the dropped category can be inferred from the others.
    - In summary, drop_first=True creates n-1 dummy variables instead of n, avoids multicollinearity issues, and leads to better, more reliable modelling results.

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



4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
    - **No Multicollinearity (for multiple regression) –** VIF is checked after each iterative model to see multicollinearity and removed one by one.

```
[237]: # Create a dataframe that will contain the names of all the feature variables and their respective VIFs
vif = pd.DataFrame()
vif['Features'] = X_train_m6.columns
vif['VIF'] = [variance_inflation_factor(X_train_m6.values, i) for i in range(X_train_m6.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```

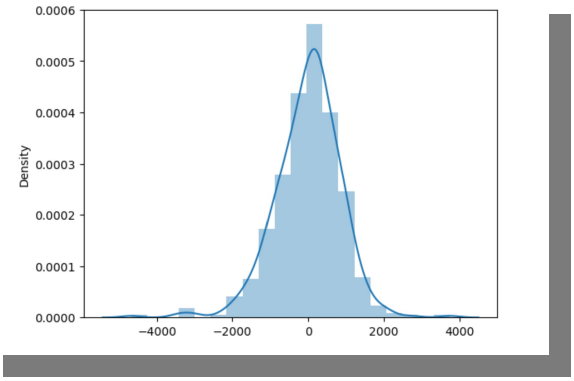
[237]:

	Features	VIF
2	windspeed	3.57
1	workingday	3.44
3	spring	2.43
0	yr	1.90
6	Jan	1.60
10	Sunday	1.59
12	Mist	1.57
4	winter	1.55
5	Aug	1.44
8	May	1.26
7	June	1.21
9	Sept	1.19
11	Light Snow	1.09

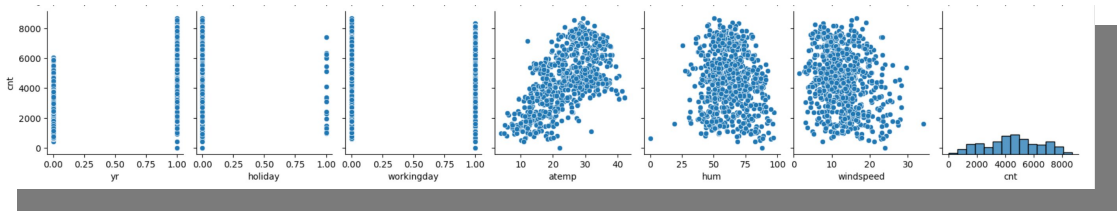
- **Normality of Errors –** After final model (Model 6), residues were found and then plotted using seaborn distplot.

Residual Analysis of the train data

```
[238]: y_train_pred = lr.predict(X_train_lm)

[239]: # Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_pred), bins = 20)
```

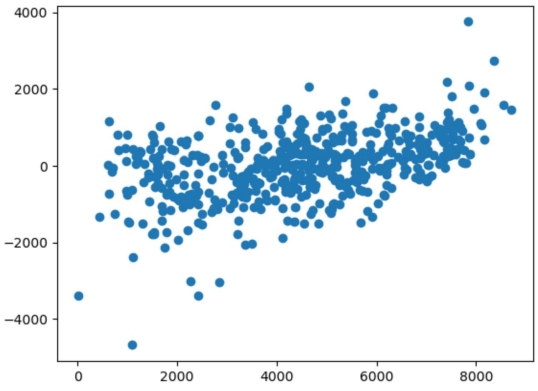


- **Linearity –** Plotted pair plot for all numeric variables with ‘cnt’ which provided the relationship between the variables.



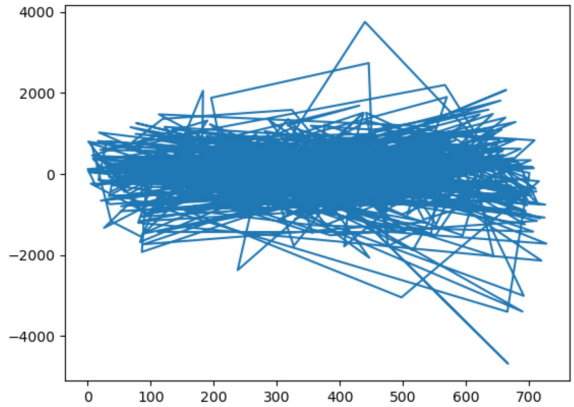
- **Homoscedasticity**

```
[251]: # Homoscadicity
       plt.scatter(y_train,(y_train - y_train_pred))

[251]: <matplotlib.collections.PathCollection at 0x1e90f1f9310>
```



- **Independence (No auto co relation between residuals)**

```
[252]: # No Linearity
       plt.plot(y_train - y_train_pred)

[252]: [<matplotlib.lines.Line2D at 0x1e90f22e210>]
```



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 Positive features**
  - yr 2159.9429
  - Aug 827.8239
  - June 796.4903

- **Top negative features**
  - Light Snow -2677.3036
  - spring -1639.3085
  - windspeed -1578.3049

- **Summary**
  - "yr" (year) has a strong positive influence, increasing "cnt" by about 2159.94 units per year.
  - "workingday" increases "cnt" by 487.13 units, suggesting higher counts on working days.
  - "windspeed" negatively affects "cnt" (coefficient -1578.30), indicating that higher wind speeds reduce "cnt."
  - Seasonal effects like "spring" (-1639.31) and "winter" (-345.08) have negative coefficients, suggesting lower counts during these seasons.
  - The months with positive coefficients
    - August (Aug)
    - June
    - May
    - September (Sept)
  - Weather conditions "Light Snow" (-2677.30) and "Mist" (-813.89) substantially decrease counts.

General Subjective Questions

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

   Ans – Linear regression is a ML model which is

   - Supervised
   - And is used to find relationship between one dependent variable and one or more independent variable

   The core idea is to fit a straight line (best-fit line) through data points in such a way that the sum of the squared differences between the actual values and the predicted values is minimized.

   This line is defined by the equation: y=mx+b

   - *y*: Predicted value (dependent variable)
   - *x*: Independent variable (input)
   - *m*: Slope of the line (effect of x on *y*) also addressed as coefficient
   - *b*: Intercept (value of *y* when x=0)

   For multiple linear regression (more than one independent variable): y = m1x1 + m2x2 + … mixi + b

   Major Assumptions of Linear Regression

   - **Linearity:** There must be a linear relationship between the independent variables and the dependent variable.
   - **Independence (No auto co relation between residuals):** The residuals (errors) should be independent of each other. This means observations are not correlated. Residuals should not be corelated and should not show any form when plotted.
   - **Homoscedasticity:** The variance of the residuals is constant across all levels of the independent variables. This means the spread of errors around the predicted values should be uniform.
   - **Normality of Errors:** The residuals should be normally distributed.
   - **No Multicollinearity (for multiple regression):** Independent variables should not be highly correlated with each other. High multicollinearity can lead to unstable coefficient estimates. This can be found programmatically by calculation VIF.

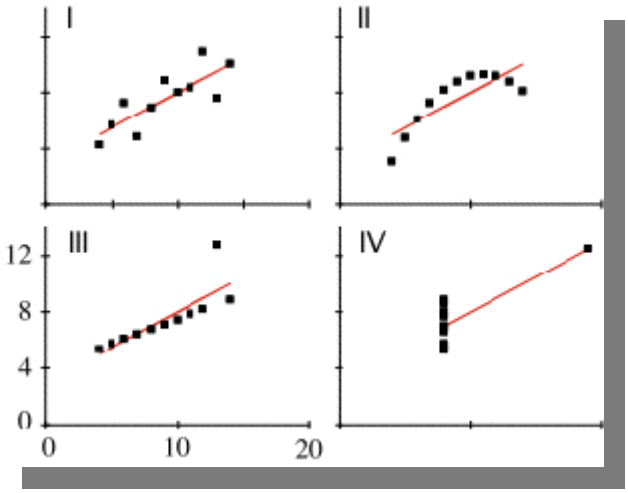
   The steps in the linear regression algorithm are as follows:

   1. **Data Collection:** Gather the dataset with the independent variables (features) and the dependent variable (target) to model.
   2. **Model Building –**
      - Add dummy variables for categorical variables
      - Split data in train and test sets
      - Rescaling of features by min max scaling or standardization
      - Fit a Model using scikit learn or stats model.
   3. **Model Evaluation:** Evaluate the model's performance using metrics such

- **R-squared –** This tells by how much percentage the model explains the variance in dependent variable
- **Coefficients Interpretation –** This tells that how much is each variable significant. It can be positive or negative. P value should be near 0 to keep it in the model.

4. **Residual Analysis –** We can check assumptions related to residues after model is built and residues can be calculated using train data
5. **Prediction:** Use the trained model to predict the output variable for new, unseen data points (test data set).

2. **Explain the Anscombe's quartet in detail. (3 marks)**
    - Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics such as mean, variance, correlation, and regression line, but differ considerably when graphed as shown below.



- despite having the same linear regression equation and statistical properties, the scatter plots show very different data structures, patterns, and outliers.
- This shows why it's important to always look at your data visually and not just rely on summary numbers, because numbers alone can be misleading about what the data really looks like. Each set tells a different story once you see it plotted, even though the simple statistics are the same across all four.
- Developed by statistician Francis Anscombe in 1973

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

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It produces a value between -1 and +1 where:

- +1 indicates a perfect positive linear correlation (as one variable increases, the other increases proportionally),
- -1 indicates a perfect negative linear correlation (as one variable increases, the other decreases proportionally),
- 0 indicates no linear correlation between the variables.

Mathematically, Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations, effectively normalizing their covariance to a unitless value. It assesses only linear associations and is symmetric, meaning the correlation from X to Y is the same as from Y to X. It is widely used in statistics to describe and infer relationships between variables and can also be used for hypothesis testing about correlation significance.

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

   Scaling is the process of transforming data features, so they fall within a specific range or distribution. It is performed to ensure that all features contribute equally to a model. Without scaling, features with larger numeric ranges can disproportionately influence model training, leading to bias models

   There are two common types of scaling:

   1. **Normalized Scaling (Min-Max Scaling):** This rescales the feature values to a fixed range, typically.

      The formula is:

      $X_{norm} = (X - X_{min})/(X_{max} - X_{min})$

      This preserves the shape of the original distribution but compresses values relative to the minimum and maximum. It is useful when you want all values on the same scale but still maintain proportional relationships.
   2. **Standardized Scaling (Z-Score Scaling):** This transforms the data to have a mean of 0 and a standard deviation of 1 by subtracting the mean and dividing by the standard deviation:

      $X_{std} = (X - \mu)/\sigma$

      This centres data around zero and normalizes variance.

      Suitable for data following a normal distribution.

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

   - Variance Inflation Factor (VIF) measures how much the presence of other variables in a regression model makes it harder to estimate the effect of a specific variable.
   - When the VIF value becomes infinite, it means that one of the variables is perfectly predicted by other variables—like having a duplicated variable or one that is a perfect mix of others.
   - In such cases, the coefficient of determination R-squared from regressing that variable on the others approaches 1, making the denominator in the VIF formula $1/1 - R^2$ approach zero, thus causing the VIF to tend to infinity.
   - In simpler terms, an infinite VIF happens when two or more variables are so closely related that the model can't tell their effects apart. This causes confusion in figuring out which variable is really driving the result. To fix this, you usually remove or combine the highly related variables, so the model works better and gives clearer answers.

- If the VIF is 1, it means the variable is completely independent of others. If VIF is higher, it means there is some correlation, and a high VIF (usually above 4 or 5) suggests a warning that variables might be too similar. Very high VIF values (above 10) mean serious multicollinearity. We usually remove variables with VIF value >5 one by one.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
    - A Q-Q plot, or quantile-quantile plot, is a graphical tool used in statistics to compare two probability distributions by plotting their quantiles against each other. In simpler terms, it helps you see if your data follows a particular theoretical distribution, especially the normal distribution.
    - In the context of linear regression, the Q-Q plot is commonly used to check one important assumption: that the residuals (the differences between observed and predicted values) are normally distributed. Normally distributed residuals indicate that the model's errors have predictable behavior, which validates inference like hypothesis testing and confidence intervals.
    - If the points on the Q-Q plot fall roughly along a straight diagonal line, it suggests the residuals are normally distributed. Deviations from this line imply the residuals might be skewed, have heavy tails, or contain outliers, which could affect the accuracy and reliability of the regression model.
    - Thus, Q-Q plots are important diagnostic tools in linear regression to visually assess whether the assumption of normality holds, providing insights into model fit and guiding potential corrective actions.