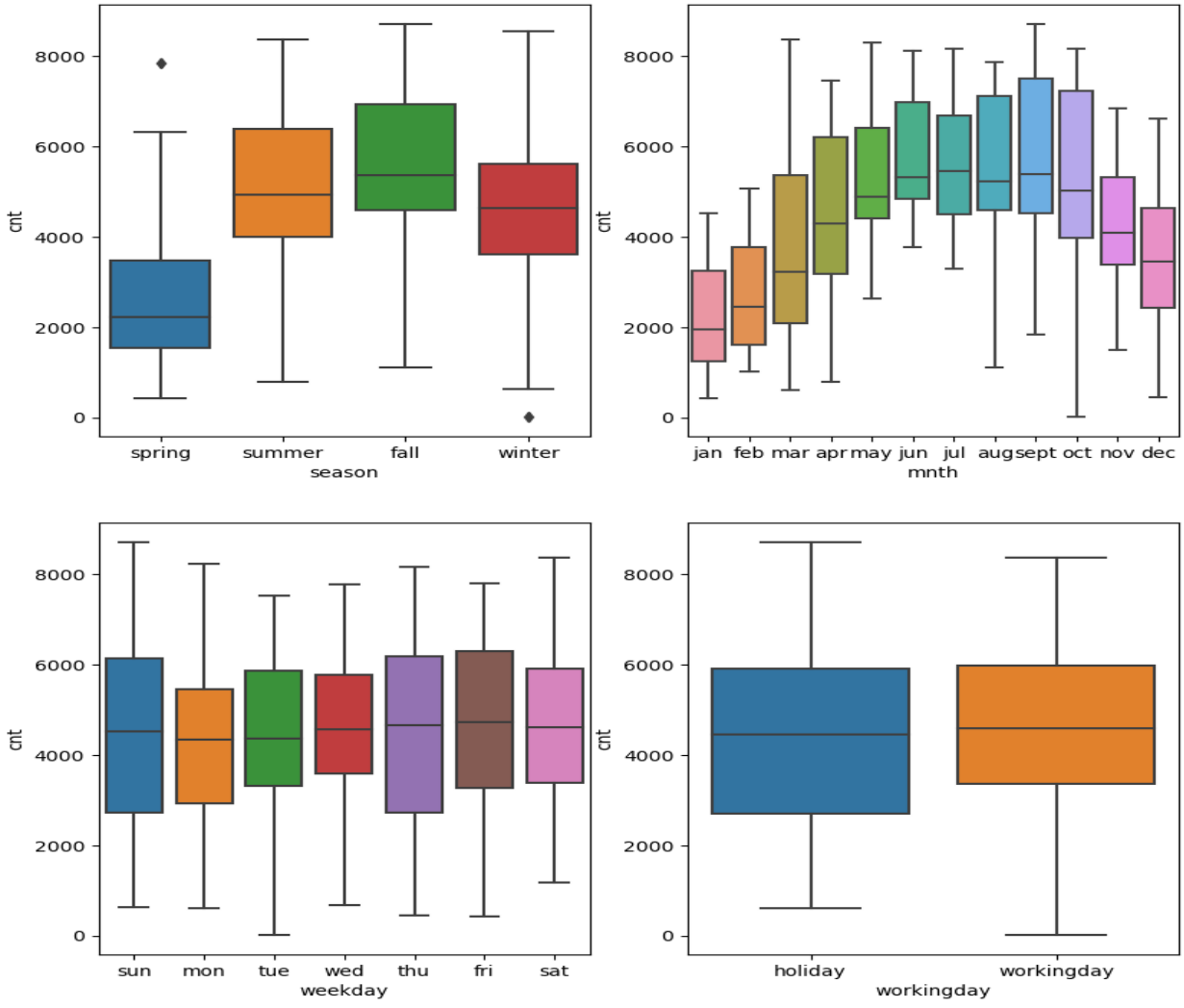
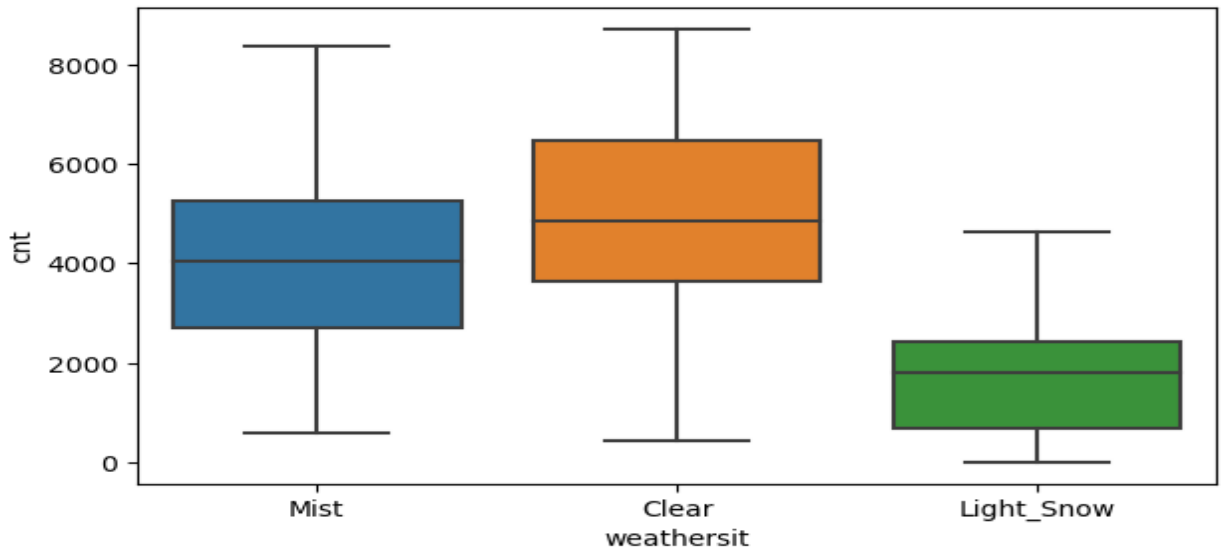
Assignment-based Subjective Questions

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', 'mnth', 'weekday', 'workingday', 'weathersit' are the categorical variables used for analysis from the given dataset.

Dependent variable is 'cnt' which gives the count of bike rentals.





This dataset was visualized using a boxplot and below are the inferences made on their effect on the dependent variable.

- Season: From the boxplot we can clearly understand that in the spring season the number of bike rentals is less when compared to other seasons. Fall has the highest demand for bike rentals.
- Month: Every month starting from January to June the count of rental increases.September has the highest count of bike rentals and then the count drops in the month of October to December.
- For Weekday and Workingday, the variation is not very significant.
- Weathersit: Count of bike rental is high for clear weather.

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

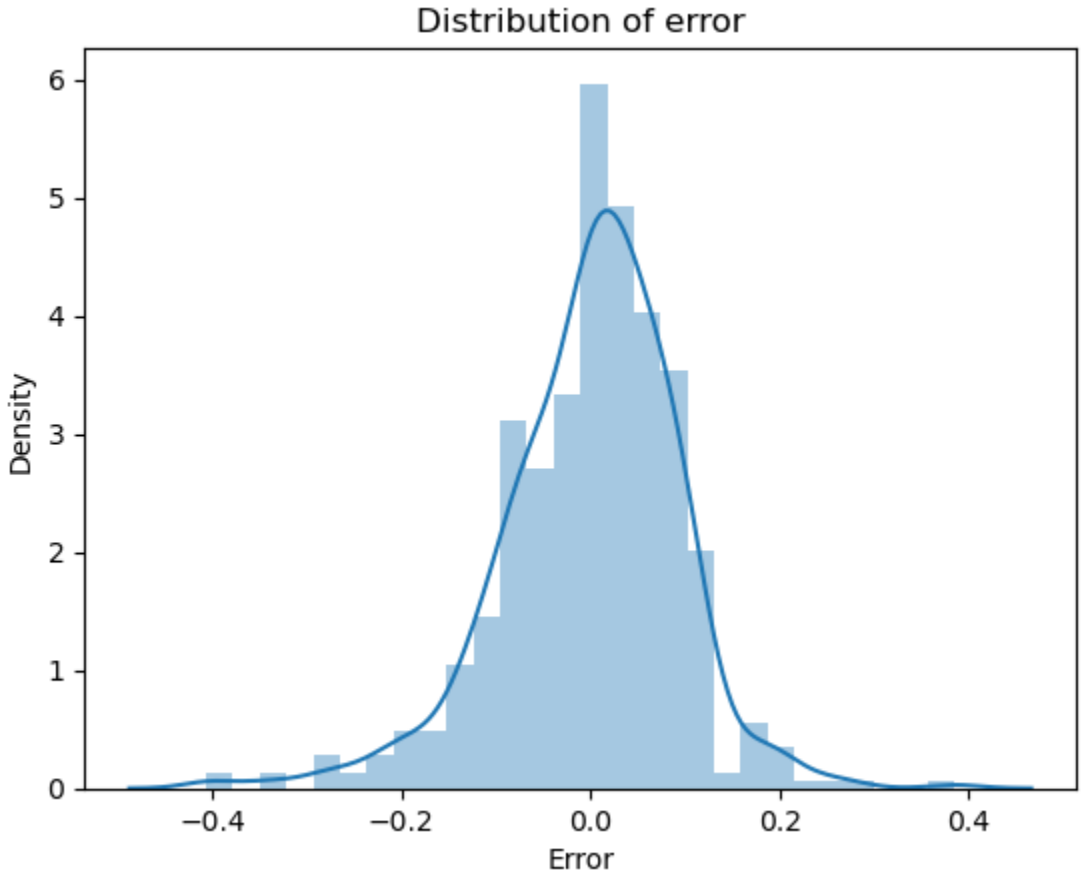
'drop_first=True' is used to eliminate the extra columns created during dummy variable creation and thus reducing the correlation among the dummy variables. If we are having categorical variables with n categories, we only need n-1 dummy variables. Failing to drop the extra dummy variable created will cause issues when building the model.

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

'temp' and 'atemp' has the highest correlation with target variable 'cnt'

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

The distribution of residuals should be normally distributed and mean should be zero.To validate this, after calculating the residuals a distplot of residuals is used to check if the error terms are normally distributed or not.
Found that the error term is normally distributed.



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

The top three predictor variables contributing significantly towards
explaining the demand of the shared bikes are:

'temp' : Temperature has a coefficient of '0.3937' which means that a unit increase in temp variable increases the count of bike rentals by '0.3937' units.

'Weathersit': weathersit_Light_Snow has a coefficient of '-0.2748' which means they are negatively correlated. This suggests that a unit increase in 'weathersit_Light_Snow' decrease the count of bike rentals by '-0.2748' units.

'yr': year has a coefficient of '0.2356' which means that a unit increase in yr increase the number of bike rentals by '0.2356' units.

General Subjective Questions

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

Linear regression is a statistical model that analyzes the linear relation between two variables i.e dependent and independent variables.
Mathematically the relation can be represent as

$$y = mx + c$$

Where y is the dependent variable and x is the independent variable.
m is the slope of the regression line and c is a constant which give the value y-intercept.
Linear regression models can be classified into two categories based on the number of independent variables.
Simple linear regression: When the independent variable is equal to 1
Equation for simple linear regression:

$$y = \beta 0 + \beta 1X$$

Multiple linear regression: When the independent variable is more than 1
Equation for Multiple linear regression:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \ldots + \beta iXi + \varepsilon$$

Assumptions of linear regression:

1. There is a linear linear relation between X and Y
2. Error should be normally distributed with mean zero
3. Error terms are independent of each other
4. Homoscedasticity, error should have a constant variance.

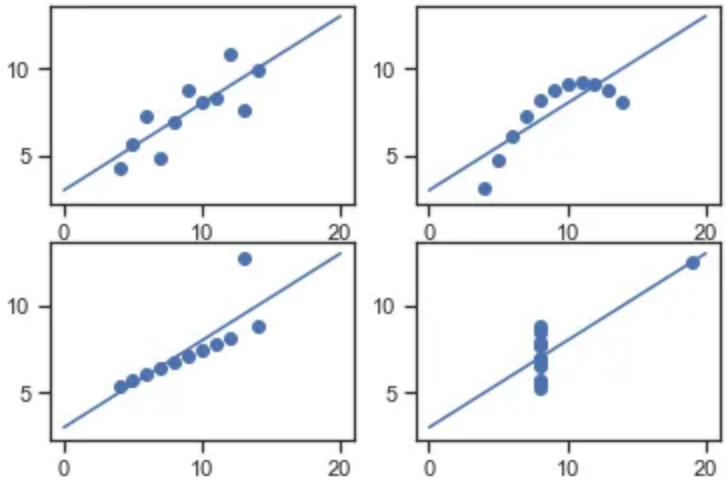
2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a model that signifies the importance of data visualization before analyzing it with statistical properties. It was developed by the statistician Francis Anscomb in 1973. It consists of four data sets and each data set consists of 11 x,y points. When analyzing the data set it is found that all the data sets have the same descriptive statistics i.e mean, variance standard deviation but have different graphical representations.

The data set given below have 4 sets of x and y values and each set have average of x = 9 , average of y = 7.50, variance of x = 11 & variance of y = 4.12

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

But when we plot these four data sets across x and y coordinate plane, we get different behavior even though they have similar statical values.



For the first data set the x and y points have linear relationship
For the second data set x and y does not have a linear relationship
For the third data set, x and y are linearly related with one outlier point.
For the fourth data set the x remains constant for the increase in y.

Conclusion:
With the Anscombe's quartet model we can conclude that the data sets which are identical over statistical analysis produce different graphs when graphically represented. So we can say that this model illustrates the importance of graphical representation when exploring the data. It is an important tool which can be used to analyze large data sets.

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

Pearson's R is used for measuring linear correlation i.e it is a numerical representation of the strength and direction of the linear relationship between two variables. It is a number between -1 and 1.
R value between 0 and 1 shows a positive correlation. When one variable changes the other variable also changes in the same direction
R value between 0 and -1 shows a negative correlation. When one variable changes the other variable also changes in the opposite direction.
R value 0 shows that there is no correlation.

When to use Pearson's correlation coefficient:

1. Both the variables are quantitative
2. Both the variables are normally distributed.
3. Data should not have any outliers
4. Variables in data should have a linear relation

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

Feature scaling is a method used to normalize the range of independent variables so that we can analyze the data accurately. It is performed to ensure that the features are having the same scale which will be beneficial for certain machine learning algorithms.If scaling is not done, the machine learning algorithm tends to weigh greater values as higher and consider smaller values as the lower regardless of the unit of values.

slno	Normalized scaling	Standardized scaling
1.	It used when data does not have gaussian distribution	It is used when data follows gaussian distribution
2.	It is affected by outliers	It is slightly affected by outliers
3.	It recales values into a range of [-1,1]	It rescales the data to have a mean of 0 and standard deviation of 1
4.	It is preferred when the algorithm does not make any assumptions about the data distribution	It is preferred when the algorithm make assumptions about the data distribution

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

Variance inflation factor (VIF) becomes infinite when there is a perfect correlation. A larger value of VIF indicates that there is a correlation between the variables. This

happens when some variables are able to create perfect correlation between the variables and a VIF greater than 10 indicates multicollinearity and that needs to be fixed. Generally, we can say that VIF ranging between 1 and 5 indicates that variable is having moderate correlation with other variables. To solve this, we need to drop the variable from the dataset which is creating perfect correlation.

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

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.

Use of Q-Q plot:

- It is a graphical technique used to compare the sample distribution of a variable against a theoretical distribution, such as the normal distribution.
- It is used for assessing the assumptions of normality of residuals in linear regression.
- If a point in the Q-Q plot approximately lies on a straight line, it indicates that the variable is normally distributed.
- It helps to assess the assumption of normally distributed errors in linear regression.

Importance of a Q-Q plot in linear regression

- As mentioned above it helps to assess the normality of residuals.
- In post deployment scenarios, Q-Q plots are used to identify covariate shift or dataset distribution changes, which is important for model performance and generalization.