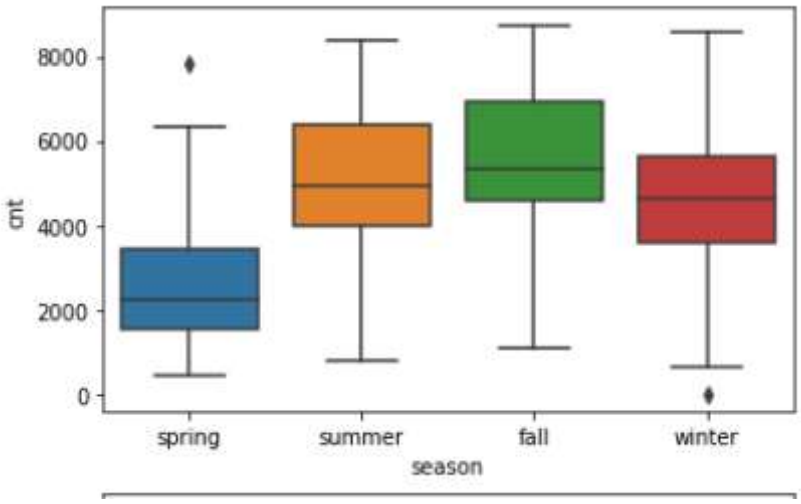
Assignment-based Subjective Questions

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

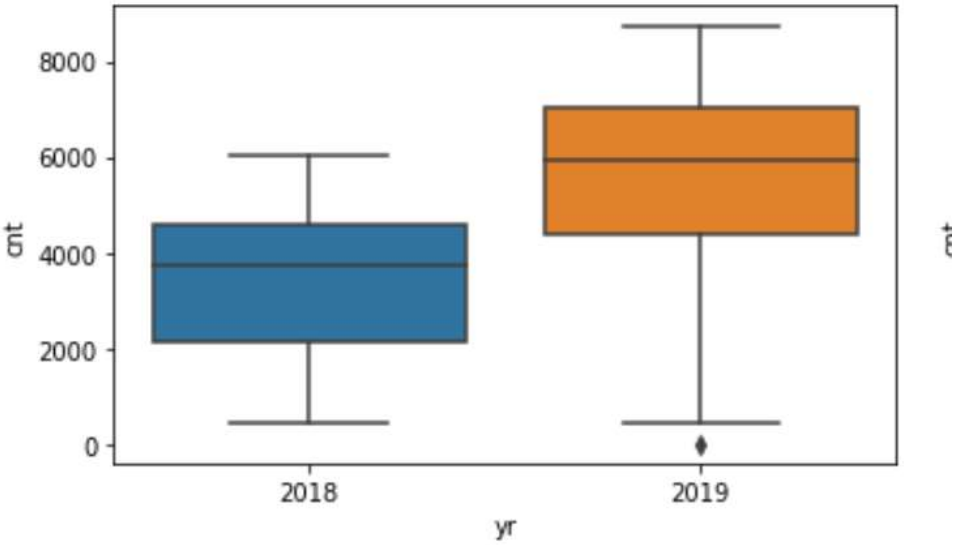
   Answer:

   From EDA, below are few primary observations with respect to categorical data.

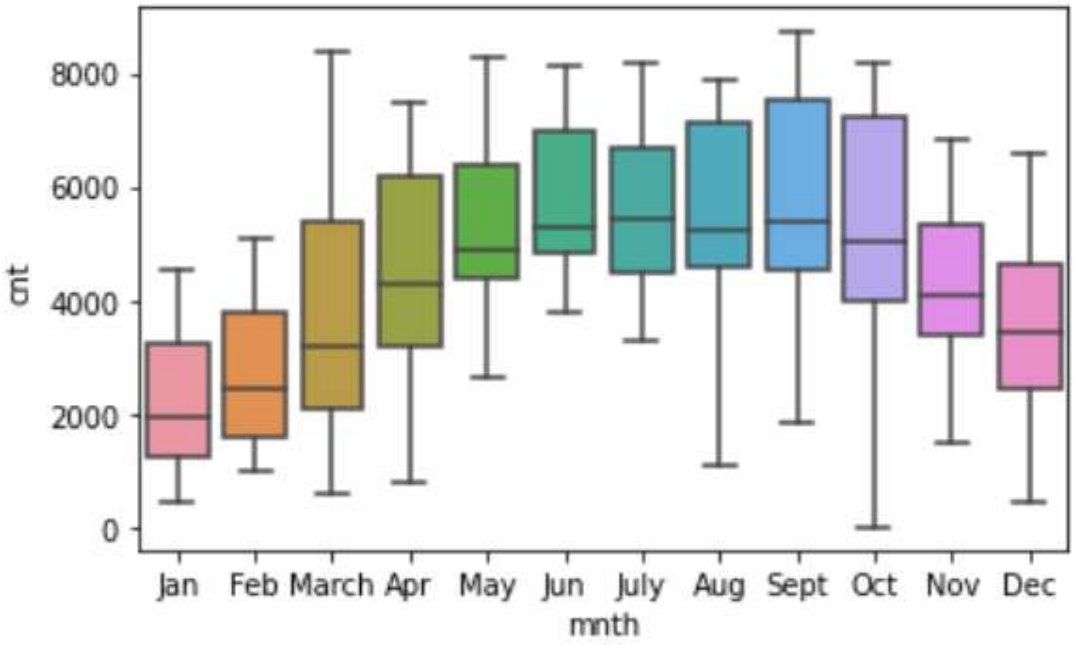
   1. Bike sharing was high during Fall season



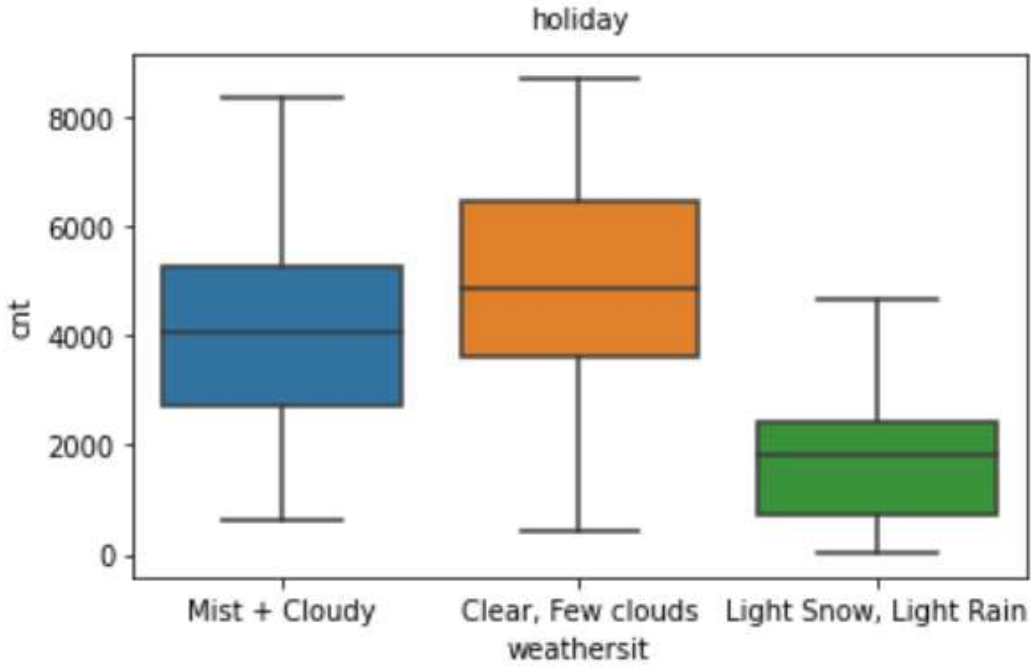
   2. Yr : No. of bike share was high during year 2019



3. Month : Bike sharing was high between months June – October



4. Weathersit : No. of bike sharing is high when weather is clear or with few clouds



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

   Answer:

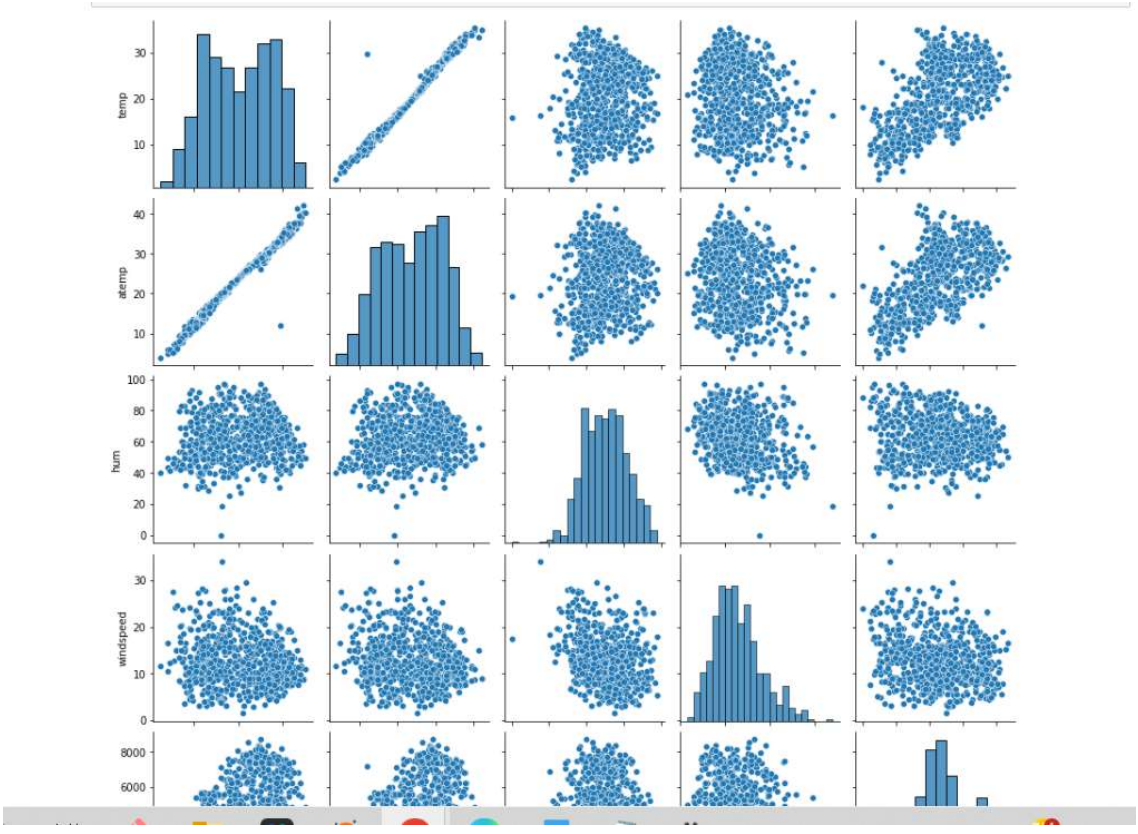
   drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

---

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

Answer :

From pairplot, we can see that variables “temp” and “attempt” have 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)

Answer:

We can validate the assumptions of Linear Regression after building the model on the following training set by below method:

1)Fitted regression line is linear.

2)Error terms came out normally distributed with mean as 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)

Answer:

The changes of increasing the number of bikes being rented increases during the working day.

The demand for bikes on rent if negatively affected by windspeed

The demand for bikes on rent is high in Fall season

The demand of bikes on rent is high in clear weather.

General Subjective Questions

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

Answer:

Linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables.

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog).

There are two main types: Simple regression: -

1. Simple linear regression uses traditional slope-intercept form, where m and b are the variables our algorithm will try to "learn" to produce the most accurate predictions. x represents our input data and y represents our prediction. y=mx+by=mx+b
2. Multivariable regression: - A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn. f (x, y, z) =w1x+w2y+w3z
   The variables x,y,z represent the attributes, or distinct pieces of information, we have about each observation.

---

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

Answer:

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

these datasets share identical summary statistics:

- Same means and variances for both x and y.
- Same correlation coefficients.
- Same linear regression lines.

Created by the statistician Francis Anscombe in 1973, this quartet serves as a powerful reminder of the importance of data visualization.

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

Answer:
Pearson's r is a numerical summary of the strength of the linear association between the variables.

If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's r takes value between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
0 < r < 5 means there is a weak association
5 < r < 8 means there is a moderate association
r > 8 means there is a strong association

---

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

Answer:

Scaling is a crucial step in data pre-processing that is applied to independent variables. Its purpose is to normalize the data within a specific range.

Why Scaling?

Data sets often contain features with varying magnitudes, units, and ranges. If we don't scale the data, algorithms might only consider magnitude and ignore units, leading to incorrect modeling.

Scaling ensures that all variables are on the same magnitude level, which is essential for accurate modeling.

Normalization (Min-Max Scaling):

- Brings data within the range of 0 to 1.
- Useful when we want to maintain the original distribution.
- Implemented using sklearn.preprocessing.MinMaxScaler in Python.

Standardization Scaling:

- Replaces values with their Z scores.
- Transforms data into a standard normal distribution (mean = 0, standard deviation = 1).
- Implemented using sklearn.preprocessing.scale in Python.
- Retains more information about outliers compared to normalization.

---

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

Answer:

The VIF is a statistical measure used to assess collinearity (correlation) between predictor variables in a regression model.

VIF becomes infinite, when two or more predictor variables are perfectly correlated.

An infinite VIF indicates that the predictor variable is linearly dependent on other variables in the model. This can lead to numerical instability and unreliable results.

Infinite CIF can be handled by Removing one of the correlated variables or Combining correlated variables into a composite variable.

---

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

Answer:

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a set of data plausibly follows a theoretical distribution (such as Normal, exponential, or Uniform).

It compares the quantiles (ordered values) of the observed data with the quantiles of a theoretical distribution.

We can use a Q-Q plot to check if the residuals of the Linear regression model are normally distributed