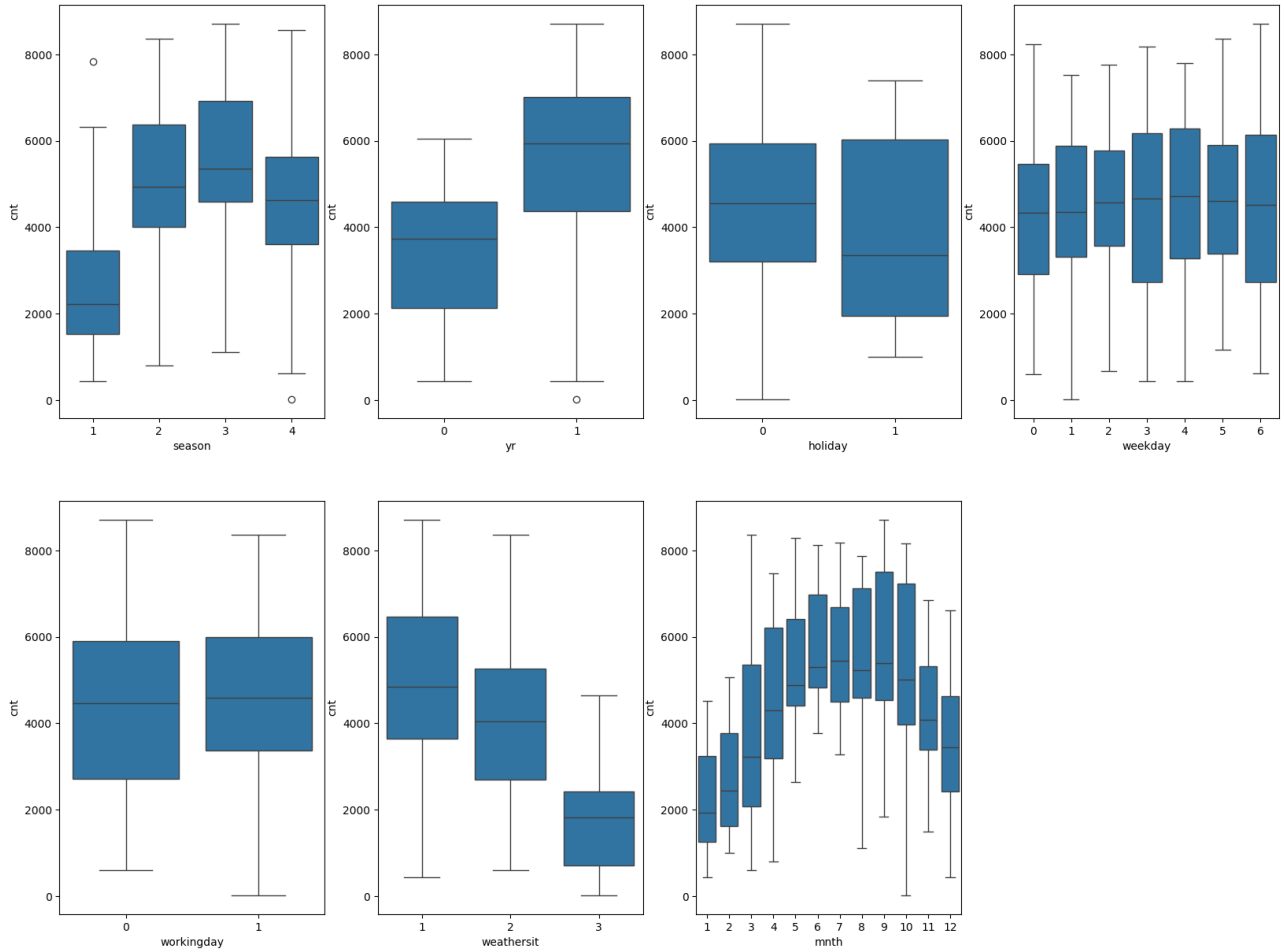
Assignment-based Subjective Questions

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



I visualized the categorical variables using box plot to get the relation

- Season: Demand is highest in Fall (3) and lowest in Spring (1).
- Year: 2019 shows higher rentals than 2018.
- Weekday: Rentals remain fairly consistent across the week.
- Weather: Heavy rain/snow results in almost no rentals; clear/partly cloudy days show the highest demand.
- Month: Rentals peak in September and drop significantly in December due to snowfall.
- Holiday: Fewer users rent bikes on holidays.
- Working day: Median rentals stay similar on working and non-working days, indicating minimal

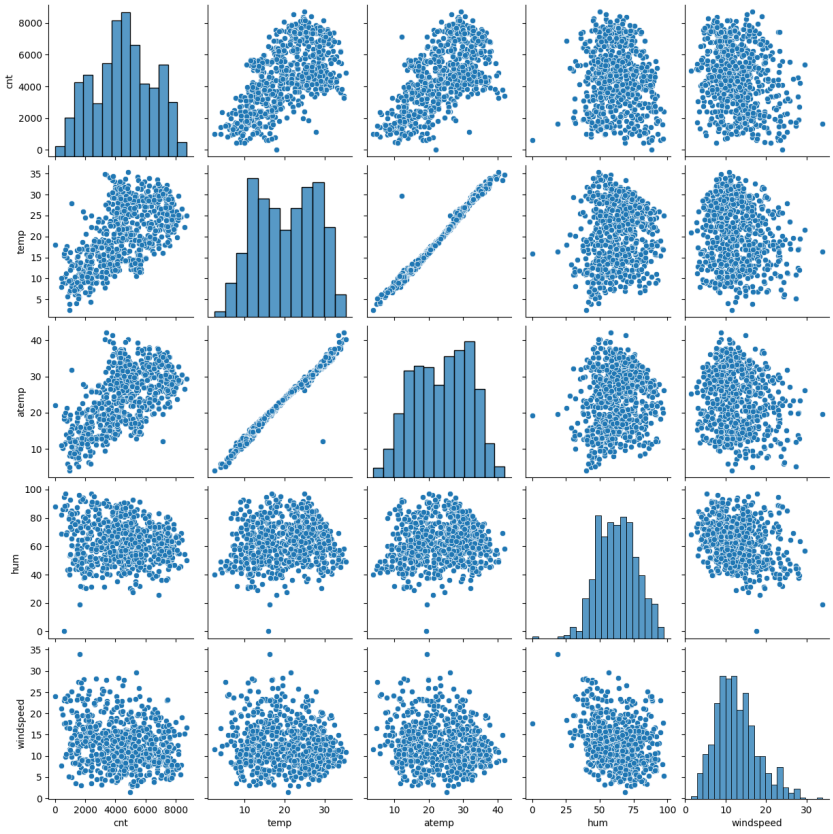
2. Why is it important to use drop_first=True during dummy variable creation?

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.

if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

   Temp and atemp has highest correlation with cnt



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

   I validated the key assumptions of Linear Regression using the following checks:

   - I examined the relationship between the predictors and the target variable using pair plots to confirm that the associations appear linear.
   - I plotted the distribution of residuals to ensure they are roughly normally distributed and centred around zero.
   - I assessed multicollinearity by computing the Variance Inflation Factor (VIF) for each feature to check whether any independent variables were highly correlated with one another.

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

   Based on my evaluation it is
   - Temp
   - weather conditions such as light snow or rain
   - year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear regression is a supervised machine learning method used to predict a continuous value. It works by finding a straight line that best fits the relationship between the input variables (x) and the output variable (y). The idea is to draw a line that stays as close as possible to all the data points.

   This line is written as:

   y = Beta0 + Beta1x

   where Beta0 is the intercept and Beta1 is the slope. These numbers (called coefficients) tell us how much y changes when x changes.

   To figure out the best values for these coefficients, the algorithm tries to reduce the difference between the actual values and the predicted ones. A common way to measure this difference is the Mean Squared Error (MSE), which averages the squared errors. The aim is to make this error as small as possible.

   Linear regression can also handle several input variables, in which case the formula expands to include more terms.

   However, it has some limitations: it assumes the relationship between inputs and output is linear, and it can be affected by outliers or situations where input variables are strongly related to each other (multicollinearity).

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a classic example used in statistics to show why just looking at numbers is not enough. It contains four different datasets that all have almost the same summary statistics—same mean, variance, correlation and even the same regression line.

But when you actually plot them, each dataset looks completely different. One is nicely linear, another is curved, one has a single outlier that pulls the line, and the last one is almost a vertical cluster of points with one odd value creating the trend.

The main idea is simple:
Don't rely only on statistics—always visualize your data, because graphs reveal patterns, outliers, and problems that summary numbers can easily hide.

3. What is Pearson's R?

Pearson's R, also called the Pearson correlation coefficient, is a number that tells us how strongly two variables are related. It measures whether they move together in a straight-line (linear) pattern.
The value of Pearson's R always lies between –1 and +1:

- +1 → perfect positive relationship (when one increases, the other always increases)
- –1 → perfect negative relationship (when one increases, the other always decreases)
- 0 → no linear relationship

For example, temperature and ice-cream sales will likely have a positive Pearson's R, while temperature and heater usage may have a negative one.
In simple words, Pearson's R tells you how strong and in what direction two things are related.

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

Scaling is the process of transforming numerical features so that they fall within a similar range. It does not change the meaning of the data—it only changes the scale of the numbers.

Why ? Scaling is performed because different features often have very different ranges. For example, "temperature" might range from 0–40, while "count of users" might range from 0–6000.
If we don't scale them:

- Models like linear regression, KNN, and SVM may get biased toward the variables with larger values.
- It helps the algorithm learn faster and more accurately.
- It reduces the impact of features that dominate purely because of their magnitude.

Normalized Scaling vs. Standardized Scaling

- Normalization: shrinks values into a fixed range (0–1).
- Standardization: rescales values based on mean and standard deviation, without fixing a specific minimum or maximum.

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

   A VIF value becomes infinite when one feature can be predicted perfectly from one or more other features. In other words, there is perfect multicollinearity.

   This happens when:

   - Two columns are exact duplicates
   - One column is a constant multiple of another
   - One column is created from another (e.g., temp and temp_scaled)
   - Dummy variables are not created correctly (dummy variable trap)

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

   A Q–Q (Quantile–Quantile) plot is a graph used to check whether a dataset follows a particular theoretical distribution—most commonly, the normal distribution. It does this by comparing the quantiles of the actual data with the quantiles of a normal distribution.
   If the points lie roughly on a straight line, it means the data is close to normally distributed.

   A Q–Q plot shows us whether our model's errors behave the way linear regression expects them to. When residuals look normal on a Q–Q plot, we can trust the model's statistical conclusions much more.