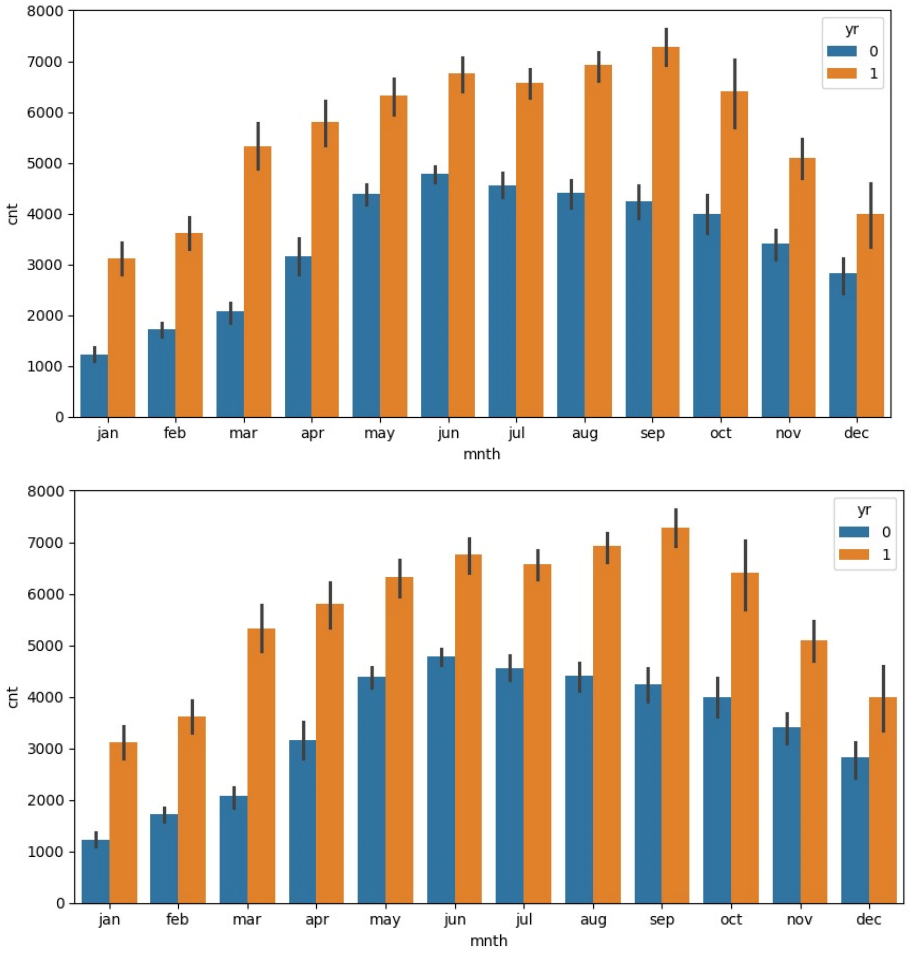
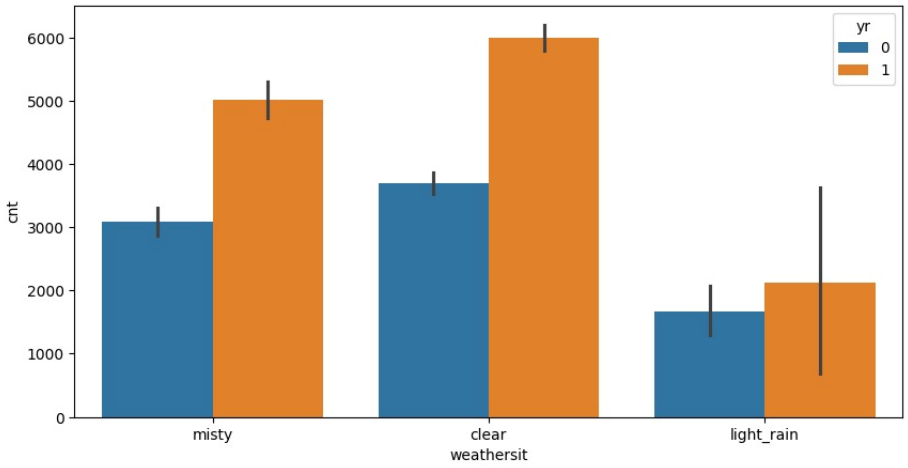
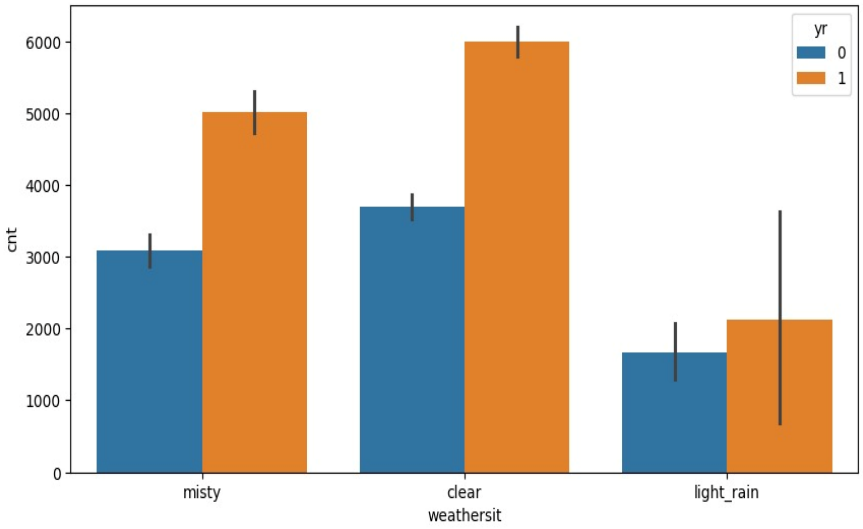
Assignment-Based Subjective Questions

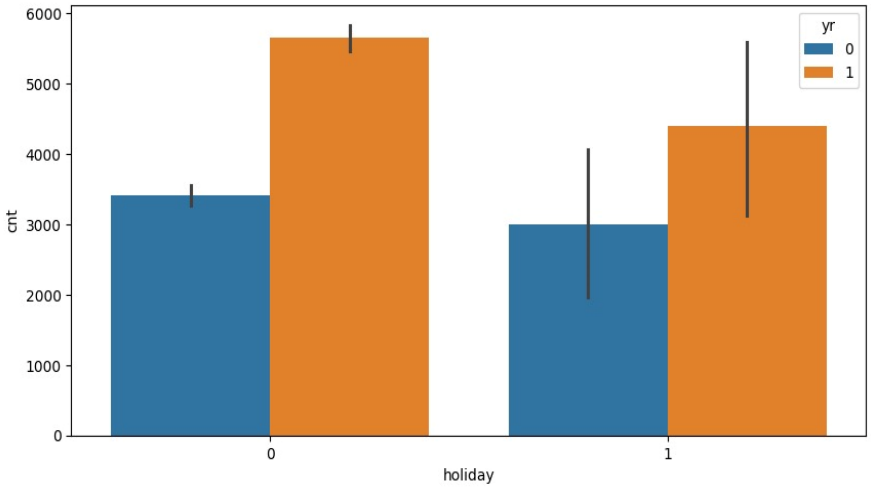
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

   - Several categorical variables—**season**, **month**, **year**, **weekday**, **working day**, and **weather situation**—play a key role in influencing the dependent variable **'cnt'**. The chart below highlights how these variables correlate with one another and with **'cnt'**.







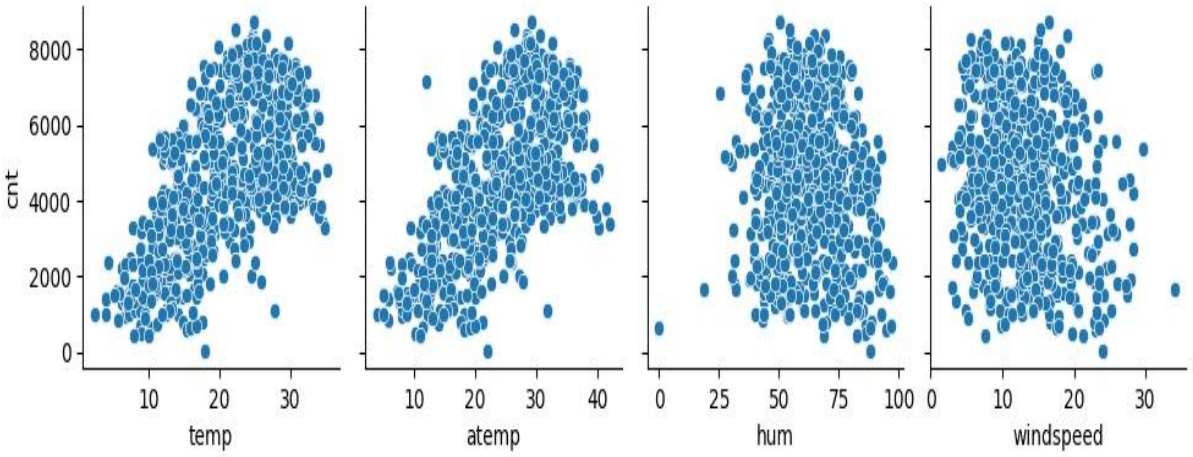


- These variables are visualized using bar plot and Box plot both.

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

- Dummy variables are created to represent categorical variables with multiple levels. For a categorical variable with **n** categories, **n-1** binary columns are generated—each indicating the presence (1) or absence (0) of a specific category.
- The parameter drop_first=True is used to drop the first category, helping to avoid multicollinearity by ensuring only **n-1** indicators are used.
  Example: If a variable has 3 levels, setting drop_first=True will exclude the first one and create dummy columns for the remaining two.

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



- The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'.

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

- Linear Regression models are validated based on Linearity, No auto-correlation, Normality of error, Homoscedasticity, Multicollinearity.

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

- Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature, year and season

General Subjective Questions

1. **Explain the linear regression algorithm in detail ?**

- Linear regression is a widely used predictive modelling technique that examines the relationship between a **dependent variable** and one or more **independent variables**. It assumes a **linear correlation**, where changes in the predictor(s) correspond to proportional changes in the target variable.
- With a single predictor, it's known as **Simple Linear Regression**.
- With multiple predictors, it's called **Multiple Linear Regression**.
- The model fits a straight line through the data to describe the relationship, which can be either **positive** or **negative**. The goal is to determine the best-fit line by minimizing the **error** between actual and predicted values using metrics like **Mean Squared Error (MSE)** or techniques like **Recursive Feature Elimination (RFE)**.

2. Explain the Anscombe's quartet in detail?

- **Anscombe's Quartet** refers to a set of four datasets that share nearly identical **descriptive statistics**—such as the mean, variance, correlation, and linear regression line—but reveal **strikingly different patterns** when visualized using scatter plots. Despite their statistical similarities, each dataset has unique characteristics that can mislead regression models if the data is not properly visualized.
- The purpose of Anscombe's Quartet is to emphasize the **importance of data visualization** in exploratory data analysis. It serves as a reminder that relying solely on summary statistics can obscure meaningful patterns, outliers, or trends in the data. Although the datasets yield similar statistical summaries for both the **x** and **y** variables, their plots demonstrate that the underlying distributions and relationships are significantly different.

3. What is Pearson's R?

- In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

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

- ❖ Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modelling.

    - **Difference between Normalizing Scaling and Standardize Scaling:**

- ➢ In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
- ➢ Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
- ➢ Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
- ➢ Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
- ➢ Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.
- ➢ Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

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

- ❖ VIF(VarianceInflationFactor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below: A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately. A very high VIF value shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

- ❖ Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a

Normal, exponential or Uniform distribution. QQ plot can also be used to determine whether or not two distributions are similar or not.

- If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

▪ Importance of QQ Plot in Linear Regression :

- In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- It can be used with sample size also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot Q-Q plot use on two datasets to check
- If both datasets came from population with common distribution
- If both datasets have common location and common scale
- If both datasets have similar type of distribution shape
- If both datasets have tail behavior.