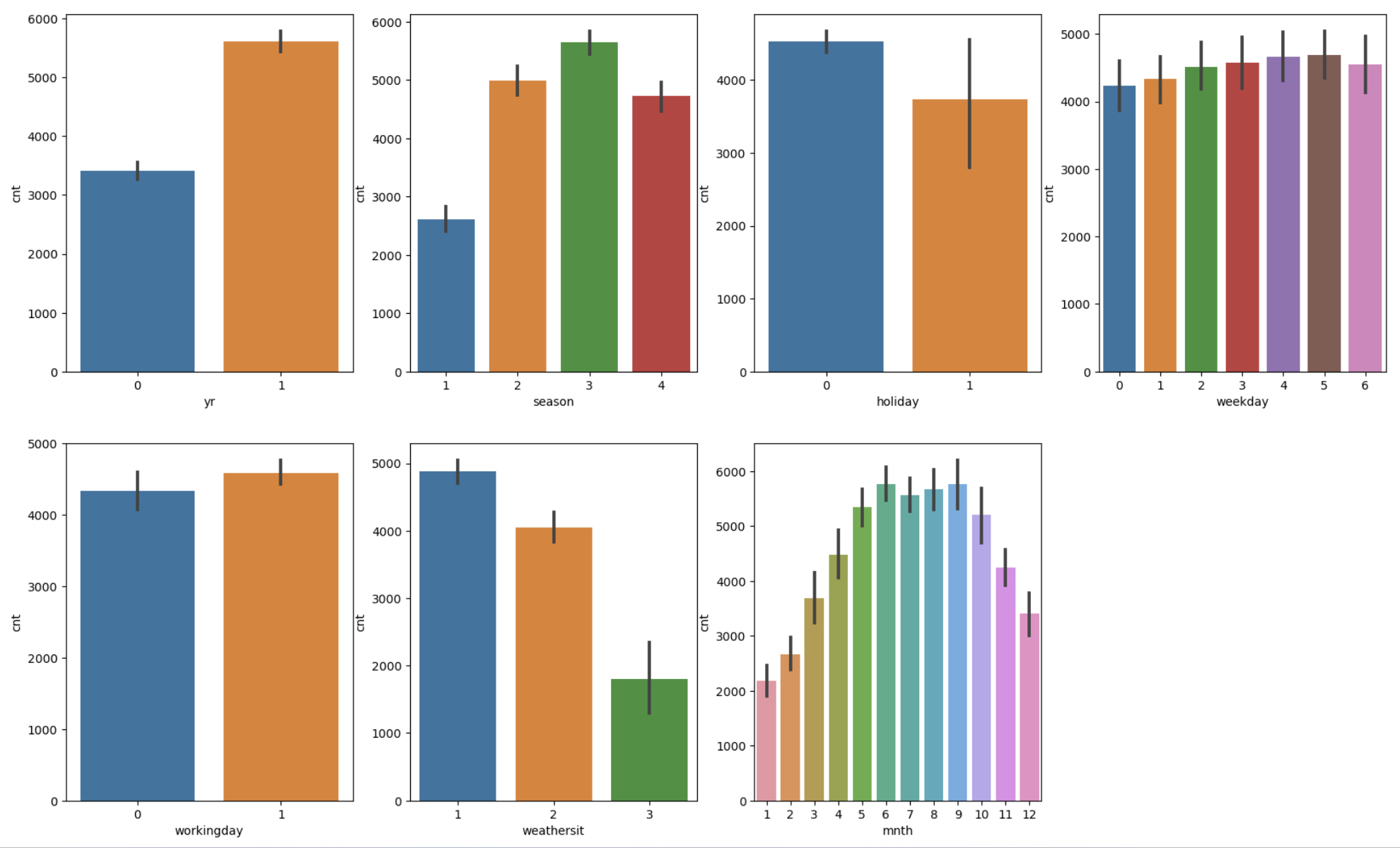
## Assignment-based Subjective Questions

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:** Categorical variables like Season, holiday, month, weather situation has significant effect on the dependent variable.

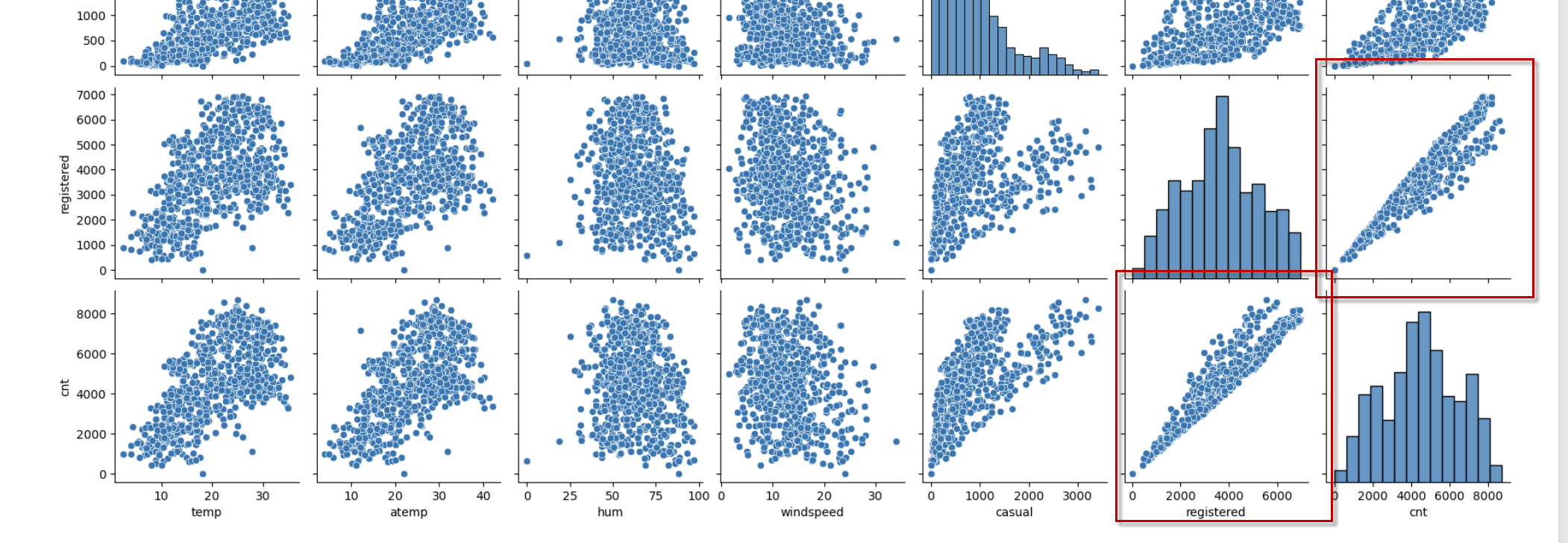


1. **Why is it important to use drop\_first=True during dummy variable creation? (2 mark)**

**Answer:** Including all variables into dummy variables will leads to multicollinearity making the sum of coefficients of all the dummy variables equals to 1, which indeed refers to have high multicollinearity. For this reason it is advised to drop first variable.

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

**Answer:** ‘registered’ variable has the highest correlation with target variable ‘cnt’.



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

**Answer:** Doing residual analysis and making sure the residual mean is ZERO, residuals are normally distributed and variance of errors is constant across all levels.

A picture containing diagram, plot, text, screenshot

Description automatically generated A picture containing screenshot

Description automatically generated

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

**Answer:** Temperature, humidity and windspeed are the three significant features.

A screenshot of a computer

Description automatically generated with medium confidence

## General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   1. Linear regression is used when we have a dataset consisting of pairs of input variables (independent variables) and corresponding output variables (dependent variables). The goal is to find the coefficients of the independent variable that would predict the dependent variable.
   2. We have two types of linear regression, namely
      1. Simple linear regression – We use only one variable to predict the dependent variable.
      2. Multiple linear regression – We use multiple variables to predict the dependent variable.
   3. The goal of linear regression is to find the best-fitting line that minimizes the sum of squared errors between the predicted and actual values.
2. **Explain the Anscombe’s quartet in detail. (3 marks)**

**Answer:** Anscombe emphasizes the importance of visualizing data to gain a deeper understanding of its underlying patterns and relationships instead of relying solely on summary statistics. Though the statistical parameters like mean, mode, median of two datasets might be same, but might exhibit different patterns when visualized. Visualizing the data helps the user to view it in different perspectives and helps to avoid drawing misleading conclusions.

1. **What is Pearson’s R? (3 marks)**

**Answer:** Pearson's R is a measure of the linear relationship between two variables, ranging from -1 to +1. It quantifies the strength and direction of the correlation.

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

**Answer:**

1. Scaling is a process of transforming data to a common scale or range.
2. Variables with different units or scales can introduce bias in the analysis. Scaling brings them to a common scale, ensuring fair comparisons.
3. There are two types of scaling process, namely
   1. Normalization:
      1. Normalized scaling (min-max scaling) rescales the data to a specific range, typically between 0 and 1.
      2. Normalized scaling preserves the shape of the distribution while bringing the values within a specific range.
   2. Standardization:
      1. Standardized scaling (or z-score scaling) transforms the data to have a mean of 0 and a standard deviation of 1.
      2. Standardized scaling centers the data around the mean and adjusts the scale based on the standard deviation.
4. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

**Answer:** When one variable being a constant multiple of another variable, it creates perfect multicollinearity and infinite VIF values. For example, when a values in two variables are [2, 4, 6] and [4, 8, 12], the relation between these two is constant that would product infinite VIF.

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

**Answer:** Q-Q plots provide a visual assessment of the fit between observed data and an assumed distribution, typically the normal distribution in linear regression. They help validate the assumptions of the regression model, detect outliers, identify departures from normality, and guide potential data transformations.