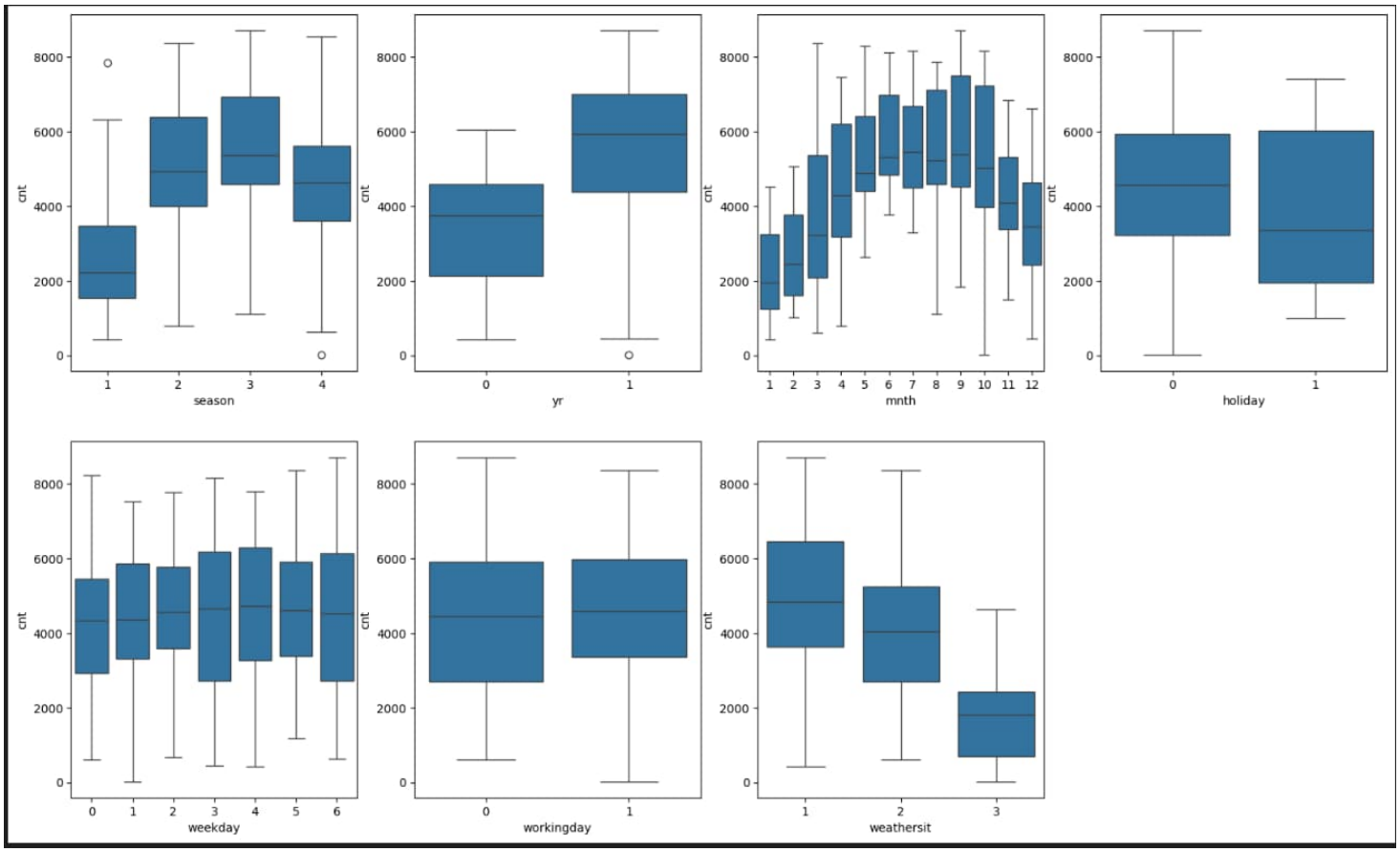
ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

1. FROM YOUR ANALYSIS OF THE CATEGORICAL VARIABLES FROM THE DATASET, WHAT COULD YOU INFER ABOUT THEIR EFFECT ON THE DEPENDENT VARIABLE?

I have drawn set of box plots to check how cnt varies with different set of categorical variables.



1. Season-wise, we observe that the median demand is significantly higher during fall (season 3) compared to spring (season 1). This suggests that the season could be a promising predictor variable.
2. Year-wise, there is a noticeable upward trend in 2019 compared to 2018, indicating year-to-year growth.
3. Month-wise, we notice demand peaks around June, July, August, and September, with median values exceeding 5000. Conversely, months January, February, March, November, and December exhibit medians below 4000, hinting at monthly booking trends.
4. Demand tends to be higher on weekdays compared to holidays.
5. On clear days (where weathersit = 1), demand is substantially higher than on rainy/foggy days (weathersit = 2) or snowy/heavy rainy days (weathersit = 3). Therefore, weathersit appears to be a potential predictor.

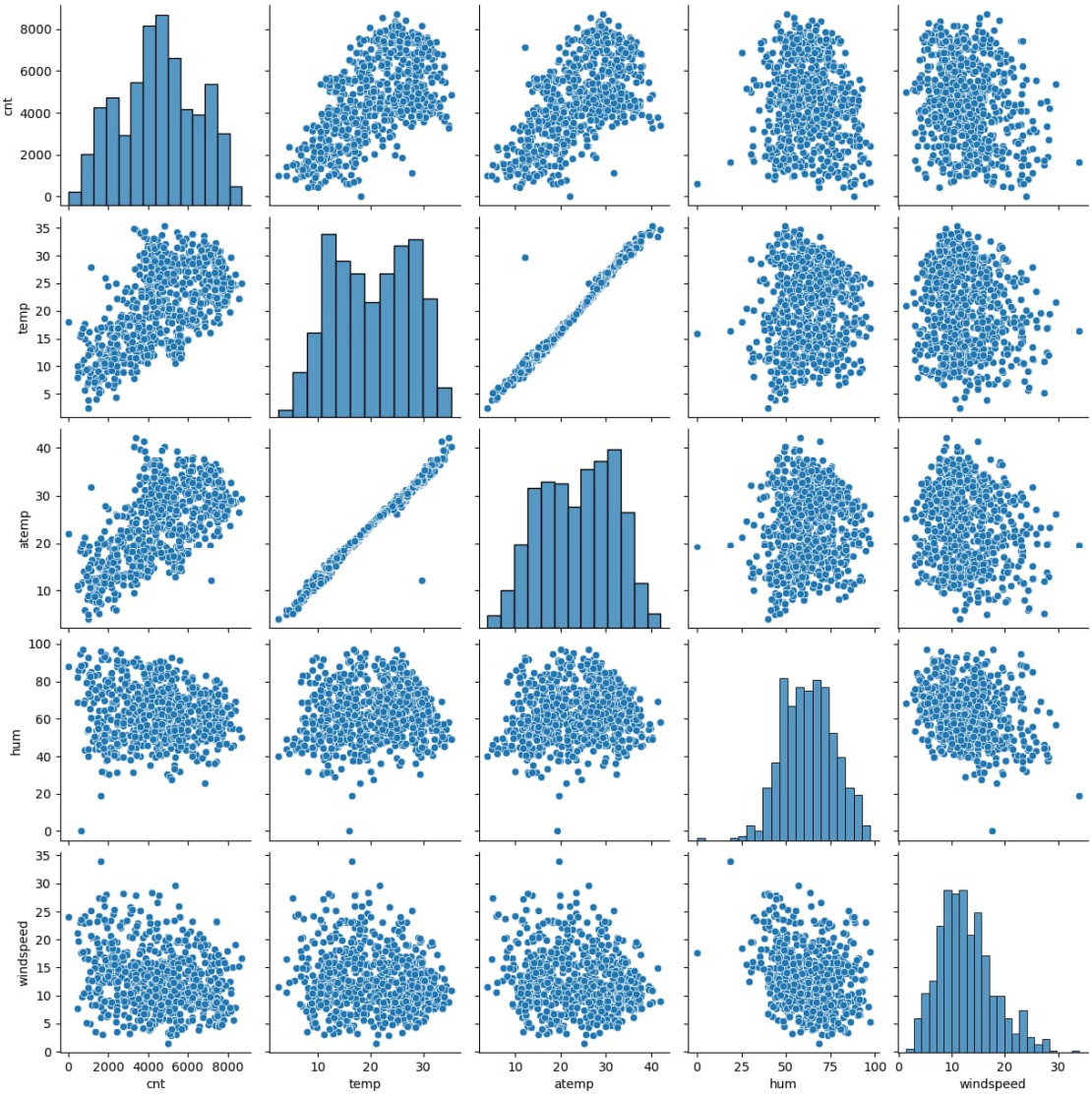
2. WHY IS IT IMPORTANT TO USE DROP_FIRST=TRUE DURING DUMMY VARIABLE CREATION?

Since dummy variable values are represented in Boolean n-1 number of dummy variables represented by n number of categories. By setting drop first=True, we avoid creating an extra column for the first category. By dropping the first column, we reduce the risk of hih correlations between dummy variables.

bike = pd.get_dummies(bike, dtype=int,drop_first=True)

3. LOOKING AT THE PAIR-PLOT AMONG THE NUMERICAL VARIABLES, WHICH ONE HAS THE HIGHEST CORRELATION WITH THE TARGET VARIABLE?

we can see huge degree of collinearity between temp and atemp.



4. HOW DID YOU VALIDATE THE ASSUMPTIONS OF LINEAR REGRESSION AFTER BUILDING THE MODEL ON THE TRAINING SET?

One of the assumptions of linear regression is that they should be normally distributed. After creating the model, we can calculate the residual and then plot a scatter plot. If this results a normal distribution, we can claim that the assumption holds good.

5. BASED ON THE FINAL MODEL, WHICH ARE THE TOP 3 FEATURES CONTRIBUTING SIGNIFICANTLY TOWARDS EXPLAINING THE DEMAND OF THE SHARED BIKES?

- Temp = 0.5568
- weathersit_3 = -0.2577
- season_4 = 0.1776

GENERAL SUBJECTIVE QUESTIONS

1. EXPLAIN THE LINEAR REGRESSION ALGORITHM IN DETAIL.

Linear Regression in Machine Learning aims to find a straight line that best fits a set of data points related to a target variable. The equation for this best-fit line is given by:

y = c + m0 x0 +.....mn xn .

In linear regression, we calculate these coefficients and use them to predict the value of (y) based on independent variables (X).

Let's break down the process:

1. Understanding the Data (Step 1):
    - Assumptions related to linear regression must hold true throughout the model-building process.
    - A crucial assumption is linearity: The relationship between independent and dependent variables should be linear.
    - If linearity is absent, the use case may not be suitable for linear regression modeling.
    - Another important assumption is the independence of observations: The values of (X) should not exhibit dependency between each other. Time-driven dependencies are also problematic for linear regression.
2. Model Building (Step 2):
    - We split the data into training and test sets.
    - Using algorithms like ordinary least squares or gradient descent, we build and train a model on the training data.
    - During model building, we validate the assumption of normal distribution of errors.
    - We also check for multicollinearity among variables.
    - Once the model is ready, we apply it to the training data to predict the target variable.
    - Statistical instruments such as R-squared, adjusted R-squared, and F statistics help evaluate the model's performance.
    - Monitoring the p-values of each variable helps assess their significance in the model.
3. Evaluate against the test data (step 3)
    - We compare the R-squared value (coefficient of determination) between the test data and the training data.
    - If the R-squared value on the test data is significantly lower than that on the training data, it could indicate overfitting.
    - Overfitting occurs when the model performs well on the training data but poorly on unseen data (test data).

2. EXPLAIN THE ANSCOMBE'S QUARTET IN DETAIL.

Anscombe's quartet emphasizes the significance of visualizing data before diving into analysis. This quartet comprises four distinct datasets, each sharing nearly identical summary statistics (such as mean, variance, correlation, and regression line). However, when these datasets are graphed, they reveal striking differences. To gain a comprehensive understanding, we must complement these metrics with exploratory data analysis and the powerful tool of data visualization.

3. WHAT IS PEARSON'S R?

Pearson correlation, also known as Pearson correlation coefficient, is a measure of the linear relationship between two continuous variables and quantifies the degree to which two variables change together in a linear fashion. The Pearson correlation coefficient is denoted by the symbol r.

r value can be from -1 to 1,

Formula to calculate r is

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

n is the number of data points

x and y are the data points

4. WHAT IS SCALING? WHY IS SCALING PERFORMED? WHAT IS THE DIFFERENCE BETWEEN NORMALIZED SCALING AND STANDARDIZED SCALING?

Scaling is the technique brings data points that are far from each other closer in order to increase the algorithm effectiveness and speed up the Machine Learning processing.

- Normalized Scaling is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling

  Normalized value of X = (x – Min X) /(Max X -Min X)

- Standardized Scaling

  Standardization (also called, Z-score normalization) is a scaling technique such that when it is applied the features will be rescaled so that they'll have the properties of a standard normal distribution with mean,$\mu=0$ and standard deviation, $\sigma=1$; where μ is the mean (average) and σ is the standard deviation from the mean..

  Standardized value of X = (x – Mean of the data) / standard deviation of the data

5. YOU MIGHT HAVE OBSERVED THAT SOMETIMES THE VALUE OF VIF IS INFINITE. WHY DOES THIS HAPPEN?

The formula to calculate VIF is

$VIFi = 1/(1-Ri^2)$

Where $Ri = Ri^2$ is the R-square value keeping the ith variable as target variable. For a very strong collinearity R=square can be 1 or very close to 1 and thus VIF can be infinity

6. WHAT IS A Q-Q PLOT? EXPLAIN THE USE AND IMPORTANCE OF A Q-Q PLOT IN LINEAR REGRESSION.

The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.