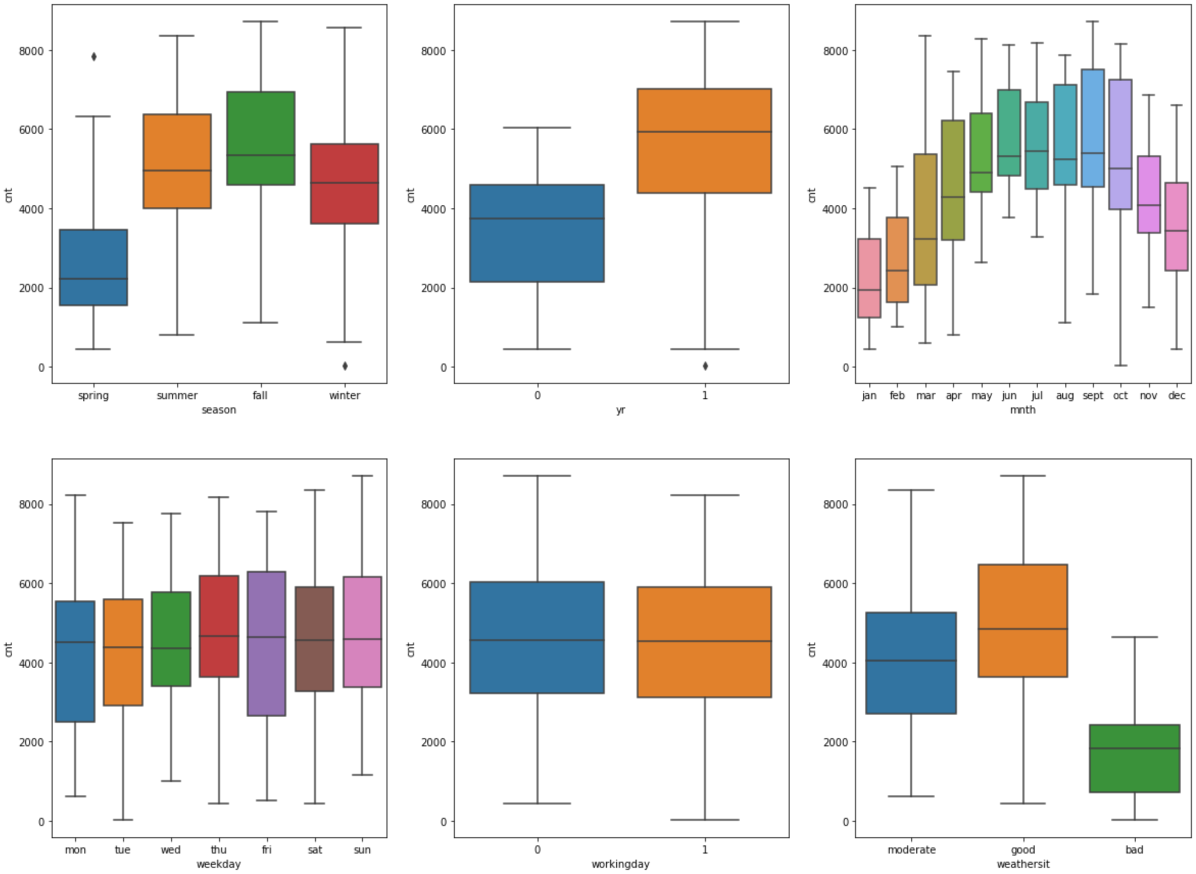
Assignment-based Subjective Answers

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

There are several categorical variables that significantly influence the dependent variable 'cnt', namely season, mnth, yr, weekday, working day, and weathersit. The correlation among these variables is illustrated in the following figure.



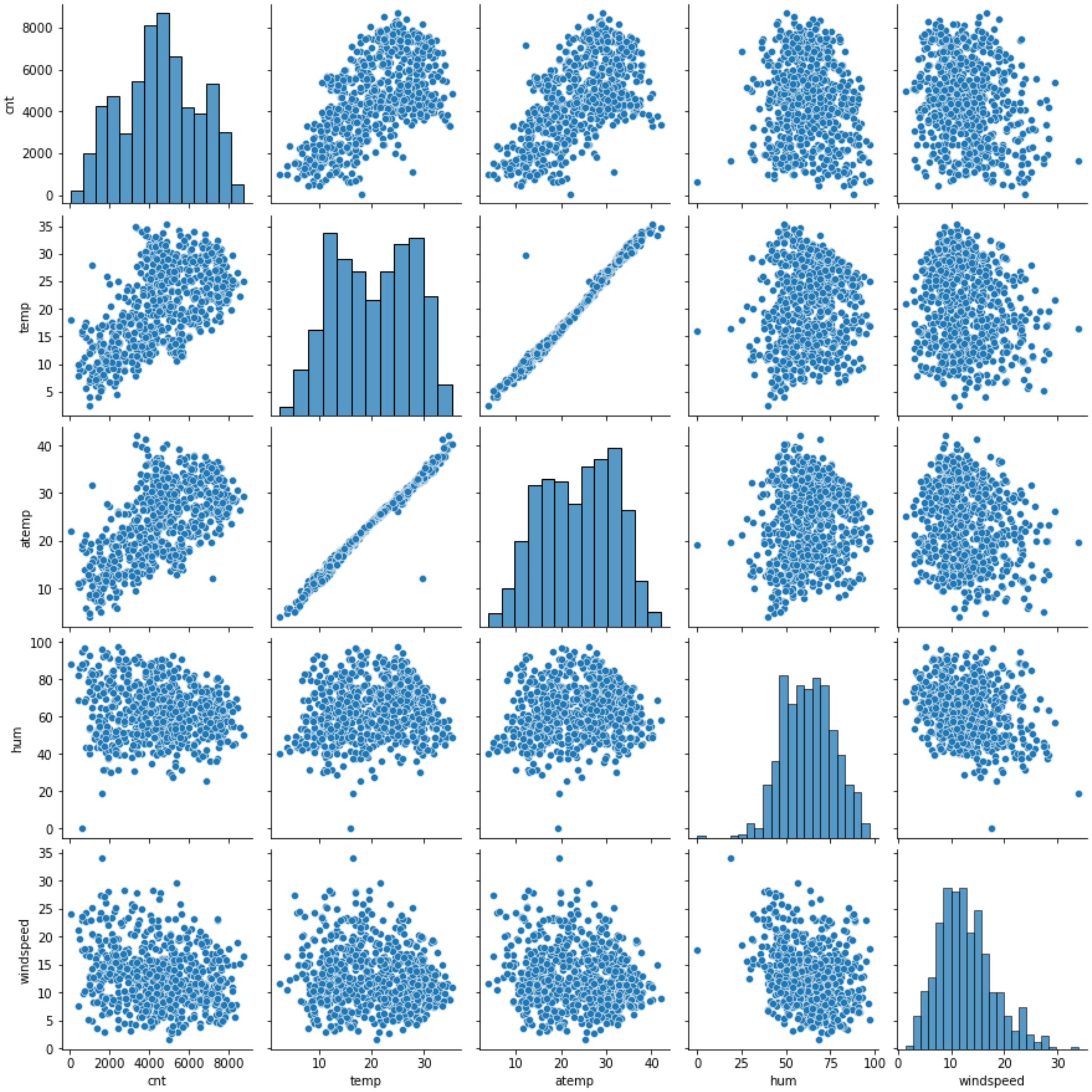
2. **Why is it important to use drop_first=True during dummy variable creation?**

The purpose of using dummy variables is to represent categorical variables with 'n' levels by creating 'n-1' new columns. Each of these new columns indicates whether a specific level exists or not, using binary values (0 or 1). By setting drop_first=True, we ensure that the resulting dummy variables match up with 'n-1' levels, effectively dropping one of the levels. This approach reduces the correlation among the dummy variables.

For example, if there are 3 levels in the categorical variable, setting drop_first=True will drop the first column representing one of the levels.

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

Among all the variables with the target variable 'cnt,' the 'temp' and 'atemp' variables exhibit the highest correlation.



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

The validation of Linear Regression models relies on checking for several key assumptions, which include linearity, absence of auto-correlation, normality of error, homoscedasticity, and multicollinearity.

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

The three most influential features that effectively explain the demand for shared bikes are temperature, year, and season.

General Subjective Answers

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

Linear regression is a predictive modeling technique that allows us to understand the relationship between a dependent variable (target variable) and one or more independent variables (predictors). It captures a linear relationship, meaning it explores how the dependent variable's value changes in response to changes in the independent variable(s). If there is only one input variable (x), this is referred to as simple linear regression, while multiple input variables give rise to multiple linear regression. The linear regression model generates a straight line with a specific slope that describes the relationship between the variables.

This regression line can exhibit either a positive linear relationship or a negative linear relationship. The primary objective of the linear regression algorithm is to determine the optimal values for a0 and a1, enabling the identification of the best-fit line with the least error.

In Linear Regression, techniques such as Recursive Feature Elimination (RFE) or Mean Squared Error (MSE), also known as the cost function, are utilized to ascertain the most suitable values for a0 and a1. These values, in turn, provide the best-fit line that aligns with the data points most effectively.

2. **Explain the Anscombe’s quartet in detail.**

Anscombe's Quartet consists of four data sets that exhibit nearly identical simple descriptive statistics. However, despite their apparent similarity, these data sets contain peculiarities that can deceive regression models if applied without careful analysis. Each data set showcases different distributions, and their scatter plots display distinctive patterns. The primary purpose of creating Anscombe's Quartet was to emphasize the significance of graphing and visualizing data before engaging in analysis or model building. It also highlights the impact of certain outliers or extreme observations on statistical properties.

All four data sets yield equivalent statistical information, including variance and mean, for both the x and y values. However, when subjected to regression modeling, they behave differently:

1. The first data set exhibits a linear relationship between X and y, making it suitable for linear regression modeling.

2. The second data set does not demonstrate a linear relationship between X and y, rendering it unsuitable for linear regression modeling.

3. The third data set contains outliers that cannot be effectively accommodated by a linear

regression model.

4. The fourth data set includes a high leverage point, resulting in a significant correlation coefficient.

In conclusion, Anscombe's Quartet serves as a reminder that regression algorithms can be misled, highlighting the importance of data visualization and exploration before constructing machine learning models.

3. What is Pearson’s R?

In the realm of Statistics, the Pearson's Correlation Coefficient is alternatively known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. This statistical measure quantifies the linear correlation between two variables.

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

Scaling involves transforming data to fit within a specific scale, and it is a type of data pre-processing step that speeds up calculations in algorithms. Collected data often contains features with varying magnitudes, units, and ranges. If scaling is not performed, algorithms may give undue importance to high-value magnitudes while neglecting other parameters, leading to inaccurate modeling.

The distinction between Normalizing Scaling and Standardize Scaling can be summarized as follows:

- In Normalized Scaling, features are scaled using their minimum and maximum values, while Standardized Scaling involves using mean and standard deviation for scaling.

- Normalized Scaling is suitable when features have different scales, whereas Standardized Scaling is employed to ensure zero mean and unit standard deviation.

- Normalized Scaling scales values between (0,1) or (-1,1), whereas Standardized Scaling is not constrained to a specific range.

- Normalized Scaling is influenced by outliers, while Standardized Scaling is not affected by outliers.

- Normalized Scaling is preferred when the distribution of data is unknown, while Standardized Scaling is more appropriate when the distribution is normal.

- Normalized Scaling is also known as Scaling Normalization, whereas Standardized Scaling is referred to as Z-Score Normalization.

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

The Variance Inflation Factor (VIF) is a metric used to assess the relationship between one independent variable and all other independent variables in a regression model. The VIF is calculated as follows:

VIF = 1 / (1 - R^2)

A VIF value greater than 10 indicates a significant issue with multicollinearity, and even a VIF value greater than 5 should not be overlooked and should be thoroughly examined.

A remarkably high VIF value implies a perfect correlation between two independent variables. In such cases, the R^2 value approaches 1, leading to a VIF value close to infinity (1/(1-R^2)). To address this problem of perfect multicollinearity, it becomes necessary to remove one of the variables from the dataset that is contributing to this high correlation.

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

A Quantile-Quantile (Q-Q) plot is a graphical method that facilitates the comparison of two probability distributions by plotting their quantiles against each other.

The primary purpose of the Q-Q plot is to help assess whether a given dataset possibly follows a specific theoretical distribution, such as a Normal, exponential, or Uniform distribution. It can also be employed to determine the similarity between two distributions. A more linear Q-Q plot suggests a higher degree of similarity between the compared distributions. The linearity assumption can be further examined using scatter plots, while the requirement for multivariate normality in linear regression analysis can be checked using a histogram or Q-Q plot.

The importance of Q-Q plots in linear regression lies in their ability to verify whether both the training and test datasets originate from the same population with a common distribution.

Advantages of Q-Q plots include their applicability to datasets of varying sample sizes and their effectiveness in detecting various distributional aspects, such as shifts in location, scale, symmetry, and the presence of outliers.

Q-Q plots are used on two datasets to check the following aspects:

- Whether both datasets originated from a population with a common distribution.
- Whether both datasets share a common location and scale.
- Whether both datasets exhibit a similar type of distribution shape.
- Whether both datasets have similar tail behavior.