Assignment – Based Subjective Questions

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

The analysis of the bookings data reveals several interesting trends:

1. Seasonal Trend: Bookings were most popular during the months of May, June, July, August, September, and October. The trend showed an increase from the beginning of the year until the middle of the year and then started decreasing as the year approached its end.
2. Yearly Growth: There was an overall increase in the number of bookings from 2018 to 2019. Each month in 2019 witnessed more bookings compared to the same month in 2018.
3. Weather Impact: Clear weather conditions attracted more bookings, which is quite understandable. Additionally, bookings increased for all weather situations in 2019 compared to 2018.
4. Day of the Week: Thursdays, Fridays, Saturdays, and Sundays recorded higher numbers of bookings compared to the start of the week (Mondays and Tuesdays).
5. Holidays vs. Non-Holidays: Bookings were generally lower on non-holiday days, which is expected as people tend to stay home and spend time with family during holidays.
6. Working Days vs. Non-Working Days: The number of bookings seemed almost equal between working days and non-working days. However, there was an increase in bookings from 2018 to 2019, regardless of the type of day.
7. Business Growth: The data indicates that the business experienced significant growth in 2019, as it attracted more bookings compared to the previous year (2018).

Question –2: Why is it important to use drop_first=True during dummy variable creation?

Using `drop_first=True` when creating dummy variables is essential to avoid multicollinearity, ensure model interpretability, and improve efficiency.

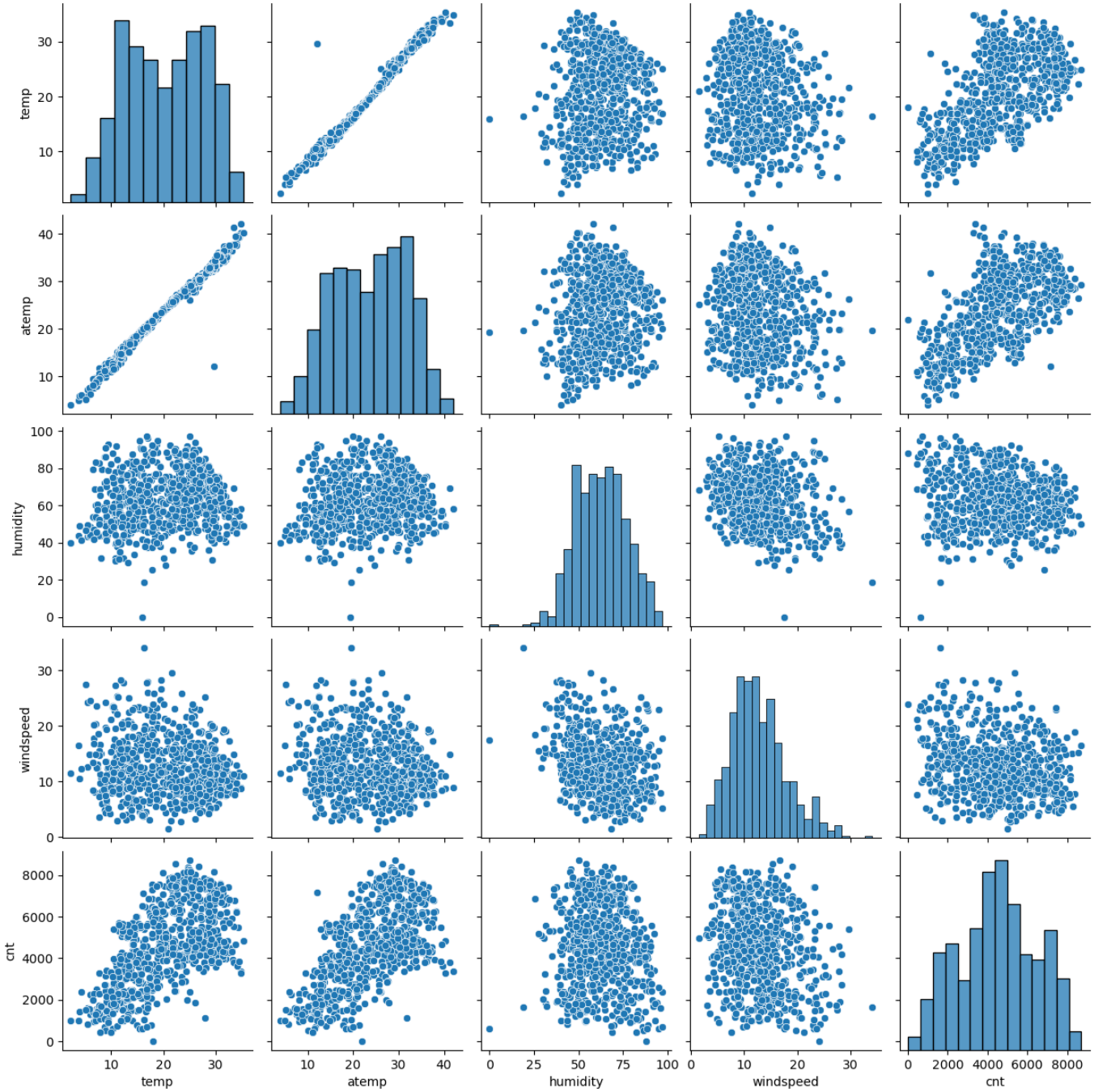
When dealing with categorical variables, it creates N-1 dummy variables for N categories, using one as a reference category. This prevents perfect multicollinearity between dummy variables, leading to stable coefficient estimates and better model performance.

The interpretable coefficients represent changes concerning the reference category, aiding model understanding.

Moreover, reducing the number of features improves computational efficiency and mitigates overfitting risks, particularly with numerous categories. In summary, `drop_first=True` is crucial for a reliable and interpretable model with improved overall performance.

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

Based on the pair plot analysis, it is evident that the variables "temp," "atemp," and "windspeed" exhibit the strongest correlation with the target variable.



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

The assumptions of the linear regression model have been validated as follows:

1. The error terms follow a normal distribution.

2. Variance Inflation Factor (VIF) values are all below 5, indicating no issue of multicollinearity among predictor variables.

3. Linearity can be observed from visualizations of the relationships between predictors and the target variable.

4. There is no visible pattern observed in the plot of residuals, indicating homoscedasticity.

5. The Durbin-Watson value of the final model (lr_6) is 2.085, suggesting no autocorrelation in the residuals.

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

1. **Temp:** With a coefficient of 0.5096, a one-unit increase in the "temp" variable leads to an increase of 0.5096 units in the number of bike rentals.
2. **Weather Situation:** The coefficient of -0.2475 indicates that a one-unit increase in the "Light_snowrain" variable results in a reduction of 0.2475 units in the number of bike hires.
3. **Year:** With a coefficient of 0.2305, a one-unit increase in the "year" variable leads to an increase of 0.2305 units in the number of bike rentals.

General Subjective Questions

Question – 1: Explain the linear regression algorithm in detail.

Linear Regression is a fundamental statistical and machine learning technique employed for predictive modeling and understanding the relationship between a dependent variable and one or more independent variables. It is particularly useful when predicting continuous numeric outcomes.

The algorithm makes certain assumptions:

- The relationship between the dependent and independent variables should be linear
- Observations must be independent.
- The variance of errors should be consistent across all independent variables.
- The errors should follow a normal distribution with zero mean.

Types of linear regression:

- Simple linear regression deals with one independent variable and one dependent variable, represented as a straight-line equation.
- Multiple linear regression involves multiple independent variables and one dependent variable, with a hyperplane equation to establish relationships among the variables.

The standard equation of the regression line is given by the following expression: $Y = \beta_0 + \beta_1.X$

Here β_0 is the intercept and β_1 is the slope. They are collectively called as regression coefficients. X represents the independent variable and Y represents the dependent variable.

The primary goal of linear regression is to minimize the difference between the predicted and actual values. Model evaluation is crucial to assess the performance of the trained model.

Question – 2: Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, but they look very different when plotted and analyzed visually. These datasets were created by the British statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not relying solely on summary statistics.
The four datasets in Anscombe's quartet share the following characteristics:

1. Same Mean: All four datasets have the same mean for both the x and y variables.

2. Same Variance: The variance of the y variable is the same or very close for each dataset.

3. Same Correlation: The correlation between x and y is approximately the same for all four datasets.

4. Same Linear Regression Line: The linear regression line for each dataset (y on x) is nearly identical.

Despite these shared statistical properties, the datasets are visually distinct.

Descriptive statistics alone may not be sufficient to fully understand the relationship between variables. Visualizing the data in different ways can provide valuable insights and help in selecting appropriate models for analysis.

Question – 3: What is Pearson's R?

Pearson's correlation coefficient, commonly denoted as "r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson and is one of the most widely used correlation coefficients in statistics.

Pearson's correlation coefficient ranges from -1 to +1:

- When "r" is close to +1, it indicates a strong positive linear correlation, meaning that as one variable increases, the other variable also tends to increase, and vice versa.

- When "r" is close to -1, it indicates a strong negative linear correlation, meaning that as one variable increases, the other variable tends to decrease, and vice versa.

- When "r" is close to 0, it suggests a weak or no linear correlation, indicating that there is no consistent linear relationship between the variables.

Pearson's correlation measures linear relationships, not accurate for non-linear. Considers alternatives. Correlation ≠ causation; strong correlation doesn't imply causative connection.

Question – 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling, refers to the process of transforming the variables of a dataset to a specific range or distribution. It is a crucial step in data preparation, particularly when dealing with machine learning algorithms that are sensitive to the scale of features.

Scaling is essential for:

1. Equalizing Impact: Prevents bias in distance-based algorithms by giving all features equal weight.

2. Accelerating Convergence: Optimizes convergence speed for algorithms like gradient descent.

3. Regularization: Ensures uniform application of regularization penalties to features.

Difference between normalized scaling and standardized scaling:

1. Normalized Scaling (Min-Max Scaling): Normalization, also known as Min-Max scaling, transforms the data into a specific range, typically between 0 and 1. Normalization is useful when we want to preserve the original distribution of the data while ensuring that all features are on the same scale. It works well when the data has a bounded range and is not heavily affected by outliers.

2. Standardized Scaling (Z-score Scaling): Standardization, also known as Z-score scaling, transforms the data so that it has a mean of 0 and a standard deviation of 1. Standardization is more appropriate when the data has a Gaussian (normal) distribution and when dealing with algorithms that assume the data to be centered around 0 and have a standard deviation of 1. Standardization is less affected by outliers compared to normalization.

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

In some cases, the Variance Inflation Factor (VIF) can become infinite. This occurs when there is perfect multicollinearity among the predictor variables in a multiple linear regression model.

VIF detects multicollinearity, where independent variables are highly correlated, causing issues in regression analysis.

Perfect multicollinearity renders some variables redundant, making coefficient estimates unsolvable. To address this, identify correlated variables and remove/combine them. Use feature selection, create composite variables, or apply regularization techniques like ridge/LASSO regression to mitigate multicollinearity's impact.

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a specific theoretical distribution, such as a normal distribution. It is particularly useful for checking the normality assumption in linear regression and other statistical analyses.

Q-Q plot in linear regression is important for:

1. Checking Normality Assumption: Verifies if residuals follow a normal distribution.

2. Valid Inference: Normal residuals ensure reliable hypothesis tests and confidence intervals.

3. Model Fit: Detects potential misspecification by examining deviations from the straight line.

4. Outlier Detection: Identifies outliers deviating from the expected line.