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? (Do not edit)

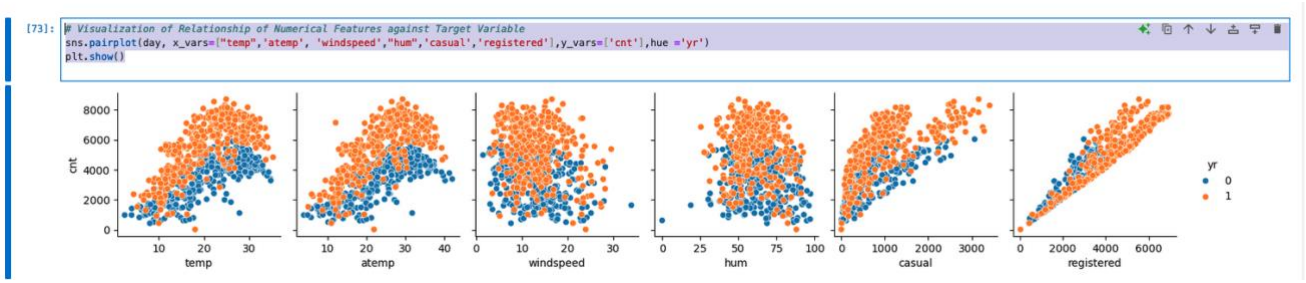
Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Visualization of Relationship of Numerical Features against Target Variable

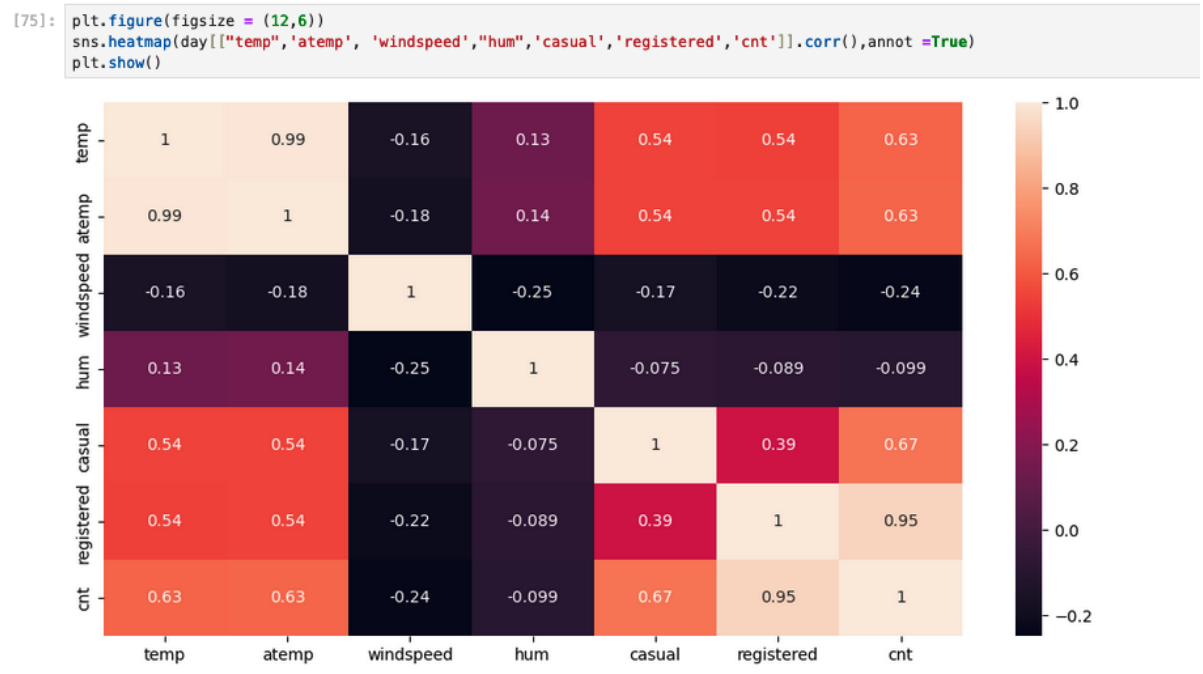
We will use pairplot using seaborn library to visualize relationships between various features like temp, atemp, windspeed, hum, casual, registered and dependent variable count (cnt) by year.

```
# Visualization of Relationship of Numerical Features against Target Variable
sns.pairplot(day, x_vars=["temp",'atemp', 'windspeed',"hum",'casual','registered'],y_vars=['cnt'],hue ='yr')
plt.show()
```



Use Heatmap to show the correlation between variables

```
plt.figure(figsize = (12,6))
sns.heatmap(day[["temp",'atemp', 'windspeed',"hum",'casual','registered','cnt']].corr(),annot =True)
plt.show()
```

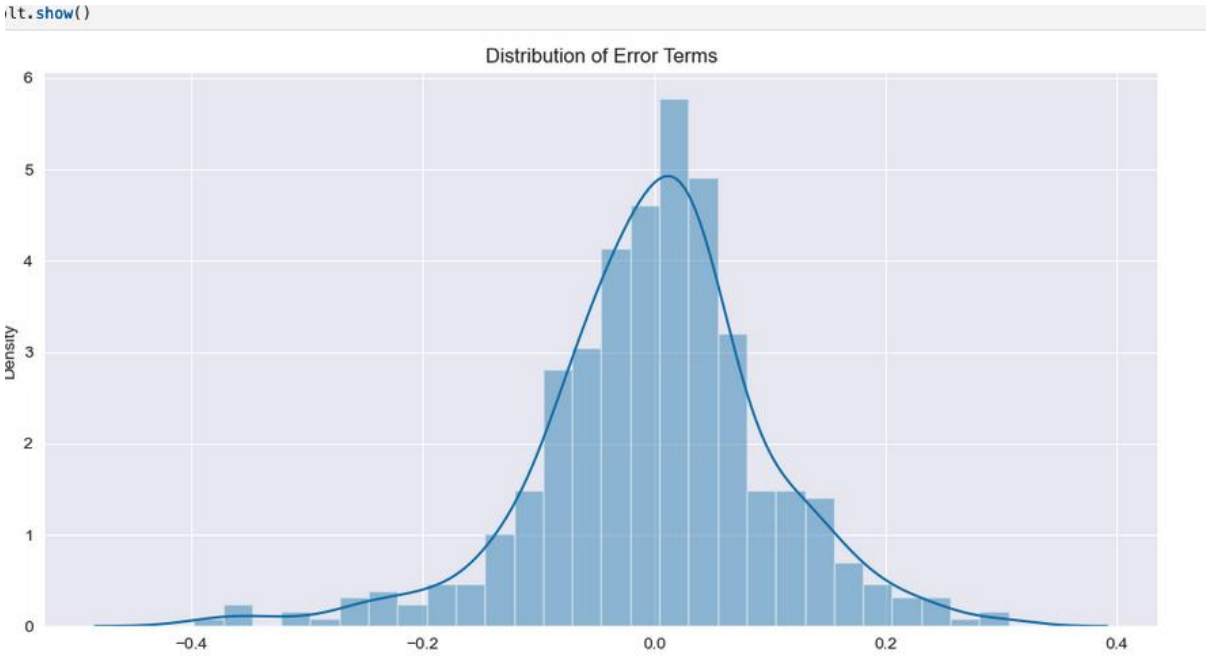


Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Using drop_first=True when creating dummy variables prevents multicollinearity and ensures that your regression model estimates coefficients correctly, providing meaningful insights into how each category influences the dependent variable.
Multicollinearity makes it difficult for regression models to estimate coefficients properly, as the variables are not truly independent. It can lead to unstable and unreliable estimates of the regression coefficients, making it hard to interpret the effect of each variable.
This makes sure Dummy dataset avoid Multicollinearity situation by starting by n-1 for a set of n values.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)
Summer season has higher count of bikes sold

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)
This is validated using Residual Analysis
# Calculation of Error in Prediction for Training Data
y_train_pred = lr_model.predict(X_train_sm)
res = (y_train - y_train_pred)
O/P I following Normal Distribution pattern



Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Temperature is the Most Significant Feature which affects the Business positively,
Whereas the other Environmental condition such as Raining, Humidity, Windspeed and Cloudy affects the Business negatively.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>
Components of the Equation

1. Dependent Variable (not explicitly mentioned): This is the value you're trying to predict or explain. It could be something like energy consumption, sales, or any other measurable outcome.
2. Independent Variables: These are the features or predictors in your equation, each with an associated coefficient:

•temp: This likely represents temperature. The coefficient 0.6 suggests that as temperature increases by one unit, the dependent variable is expected to increase by 0.6 units, holding all other variables constant.

•yr: This probably represents the year. A coefficient of 0.23 means that each additional year is associated with an increase of 0.23 in the dependent variable.

•const: This likely represents a constant or intercept term in the model. It's a baseline value when all other predictors are zero.

•winter: This could indicate whether it's winter (binary variable, e.g., 1 if winter, 0 otherwise). A coefficient of 0.11 indicates a positive association with the dependent variable when it's winter.

•workingday: This indicates whether it's a working day (again, likely binary). A coefficient of 0.03 suggests a slight positive effect on the dependent variable.

•Wednesday, Friday, Thursday, Monday, Tuesday: These variables likely represent the days of the week (binary). The negative coefficients suggest a decrease in the dependent variable on these specific days compared to a reference day (often Sunday or Saturday).

•Mist: This might represent weather conditions, where a coefficient of -0.05 indicates a negative impact on the dependent variable when mist is present.

•holiday: This variable indicates whether it's a holiday. A coefficient of -0.07 suggests a decrease in the dependent variable during holidays.

•hum: This represents humidity. The coefficient of -0.14 indicates that higher humidity negatively affects the dependent variable.

•windspeed: A coefficient of -0.17 suggests that as wind speed increases, the dependent variable tends to decrease.

•Light: This likely refers to light levels (e.g., daylight or artificial lighting). A coefficient of -0.24 indicates that higher light levels correlate with a decrease in the dependent variable.

Interpretation of Coefficients

• Positive Coefficient: A positive coefficient (like 0.6 for temperature) means that as the predictor increases, the dependent variable also tends to increase.
• Negative Coefficient: A negative coefficient (like -0.24 for Light) means that as the predictor increases, the dependent variable tends to decrease.

Overall Equation

The equation can be interpreted as a linear combination of these independent variables to predict the dependent variable. Each term contributes to the overall prediction based on its coefficient.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet is a set of four datasets that were constructed by statistician Francis Anscombe in 1973 to illustrate the importance of data visualization and the potential pitfalls of relying solely on statistical summaries. Each dataset consists of 11 pairs of (x, y) values and shares several statistical properties, yet they reveal very different underlying distributions when plotted.

Datasets in Anscombe's Quartet

1. **Dataset I**:
    - x values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
    - y values: 8, 6, 7, 6, 7, 9, 2, 4, 6, 5, 3
    - **Plot**: Shows a clear linear relationship.
2. **Dataset II**:
    - x values: 8 (constant)
    - y values: 6, 5, 7, 6, 5, 6, 8, 7, 9, 6, 5
    - **Plot**: All x values are the same, creating a vertical line with no correlation.

3. **Dataset III**:
    - x values: 13, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
    - y values: 6, 6, 2, 6, 7, 9, 2, 4, 6, 5, 3
    - **Plot**: Similar to Dataset I but contains an outlier that affects the correlation.
4. **Dataset IV**:
    - x values: 8 (constant)
    - y values: 12, 8, 10, 9, 11, 12, 9, 9, 11, 12, 12
    - **Plot**: Similar to Dataset II, showing no correlation with the presence of outliers.

Key Statistical Properties

Despite their differences in appearance, all four datasets have:

- The same mean x value (~9).
- The same mean y value (~7.5).
- The same sample variance for both x and y.
- A similar linear regression line.

Importance of Anscombe's Quartet

The quartet emphasizes the following points:

- **Visualization Matters**: Different datasets can share statistical properties while having vastly different distributions. Visualizing data helps to uncover patterns that numerical summaries might obscure.
- **Misleading Statistics**: The datasets show that statistical measures like correlation coefficients can be misleading without accompanying visualizations.
- **Impact of Outliers**: Outliers can significantly distort statistical results, as seen in Dataset III.

Conclusion

Anscombe's quartet serves as a fundamental teaching tool in statistics, highlighting the necessity of visual analysis in data interpretation. It reminds us that understanding data goes beyond mere numbers, and visualization plays a crucial role in revealing the true story behind the data.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's correlation coefficient is a fundamental statistic in data analysis that helps researchers understand the degree to which two variables are related. While it provides valuable information, it is essential to visualize data and consider its assumptions to draw accurate conclusions.

Pearson's r , also known as Pearson correlation coefficient, is a statistical measure that assesses the strength and direction of the linear relationship between two continuous variables. It is widely used in statistics to determine how well the data fit a linear regression model.

Key Characteristics of Pearson's r

1. Range: The value of Pearson's r ranges from -1 to 1:
   - r = 1 : Perfect positive linear correlation (as one variable increases, the other variable also increases).
   - r = -1 : Perfect negative linear correlation (as one variable increases, the other variable decreases).
   - r = 0 : No linear correlation (no predictable relationship between the variables).
2. Interpretation:
   - Strong Positive Correlation: Values close to 1 (e.g., 0.8 or 0.9).
   - Moderate Positive Correlation: Values around 0.5 to 0.7.
   - Weak Positive Correlation: Values closer to 0 but positive.
   - Strong Negative Correlation: Values close to -1 (e.g., -0.8 or -0.9).
   - Moderate Negative Correlation: Values around -0.5 to -0.7.
   - Weak Negative Correlation: Values closer to 0 but negative.
3. Formula:

The Pearson correlation coefficient is calculated using the formula:

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

Where:

- n = number of pairs of scores
- x = values of the first variable
- y = values of the second variable

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling is a data preprocessing technique used in statistics and machine learning to normalize or transform features (variables) to a common scale without distorting differences in the ranges of values. Scaling is essential because many algorithms perform better or converge faster when features are on a relatively similar scale.

Scaling is performed for several important reasons in data preprocessing, particularly in the context of machine learning and statistical analysis. Here are the key reasons why scaling is essential:

1. Improves Model Performance
2. Facilitates Faster Convergence
3. Enhances Interpretability
4. Improves Numerical Stability
5. Reduces Sensitivity to Outliers
6. Satisfies Algorithm Assumptions
7. Standardizes Data Across Different Features

In summary, scaling is a crucial preprocessing step that enhances model performance, speeds up convergence, improves interpretability, and ensures that different features are treated equally during analysis. By addressing the scale of the data, scaling helps create more robust and effective machine learning models.

Normalized Scaling (Min-Max Scaling)

Normalized scaling, often referred to as Min-Max scaling, rescales the features to a fixed range, typically $[0, 1]$. This method transforms the data based on the minimum and maximum values of each feature.

Standardized Scaling (Z-score Normalization)

Standardized scaling transforms features to have a mean of 0 and a standard deviation of 1. This process, known as Z-score normalization, centers the data around the mean and scales it based on its standard deviation.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in multiple regression models. A high VIF value indicates that a predictor variable is highly correlated with one or more of the other predictor variables, which can inflate the variance of the coefficient estimates and make them unstable.

Reasons for Infinite VIF

1. Perfect Multicollinearity- This occurs when one predictor variable is an exact linear combination of one or more other predictor variables.
2. Insufficient Variation - If a predictor variable has little to no variation (e.g., if all its values are the same), it can create problems in estimating its effect in the regression model.
3. Model Specification Errors - Errors in the way the regression model is specified, such as omitting relevant variables or including irrelevant ones, can lead to perfect multicollinearity.

Addressing Infinite VIF

If you encounter infinite VIF values, consider the following steps:

1. **Identify and Remove Collinear Variables**: Analyze the correlation matrix and remove one of the variables that is causing multicollinearity
2. **Combine Variables**: If appropriate, consider combining collinear variables into a single variable (e.g., using principal component analysis).
3. **Reassess Model Specification**: Ensure that the model is correctly specified and check for omitted variables that might cause multicollinearity.
4. **Use Regularization Techniques**: Consider using regularization methods like Ridge or Lasso regression, which can handle multicollinearity by adding a penalty term to the regression.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>
A Q-Q plot (Quantile-Quantile plot) is a graphical tool used in statistics to assess whether a dataset follows a specific theoretical distribution, most commonly the normal distribution. It compares the quantiles of the sample data against the quantiles of a theoretical distribution. If the data follows the specified distribution, the points on the Q-Q plot will approximately lie along a straight line.
Key Components of a Q-Q Plot

1. Axes:
    - The x-axis represents the theoretical quantiles from the specified distribution (e.g., normal distribution).
    - The y-axis represents the quantiles of the sample data being tested.
2. Data Points:
    - Each point on the plot corresponds to a pair of quantiles, one from the theoretical distribution and one from the sample data.
3. Reference Line:
    - A reference line (often a 45-degree line) is drawn to help visualize how closely the data follows the theoretical distribution. Ideally, if the data follows the distribution, the points will align closely with this line.

Importance of a Q-Q plot in linear regression

A Q-Q plot (Quantile-Quantile plot) plays a crucial role in the diagnostic phase of linear regression analysis. It is primarily used to assess whether the residuals of a regression model are normally distributed, which is one of the key assumptions underlying linear regression.

Importance of Q-Q Plots in Linear Regression

1. Assumption Checking:

    - Linear regression relies on several key assumptions, including linearity, independence, homoscedasticity (constant variance), and normality of residuals. A Q-Q plot specifically addresses the normality assumption. Violations of these assumptions can lead to biased coefficient estimates and unreliable hypothesis tests.

2. Inference Validity:

    - Many statistical tests (e.g., t-tests for coefficients, F-tests for overall model significance) used in conjunction with linear regression rely on the assumption that the residuals are normally distributed. If this assumption is violated, the results of these tests may be invalid, leading to incorrect conclusions about the relationships in the data.

3. Robustness of Predictions:

    - Normality of residuals is important for the reliability of prediction intervals. If the residuals are not normally distributed, the prediction intervals may be too narrow or too wide, impacting the credibility of predictions made by the regression model.

4. Model Improvement:

    - If a Q-Q plot indicates non-normality, it provides insights for model improvement. Analysts can consider data transformations (like logarithmic or square root transformations) to stabilize variance or reduce skewness, which can lead to improved model fit and more accurate predictions.
