Submitted by: Samin Ratna Shakya

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?

   Answer:

Many of the predictor variables in the final multiple regression model are categorical in nature, and some of them have been converted to dummy variables.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2772	0.020	14.073	0.000	0.238	0.316
yr	0.2371	0.009	25.519	0.000	0.219	0.255
temp	0.3857	0.026	14.831	0.000	0.335	0.437
windspeed	-0.1502	0.028	-5.390	0.000	-0.205	-0.095
spring	-0.1476	0.014	-10.842	0.000	-0.174	-0.121
weathersit_3	-0.2433	0.028	-8.834	0.000	-0.297	-0.189

Spring falls under season category and have been encoded.
weathersit_3 falls under weathersit category and have been encoded.
we can infer from the above image that these variables are statistically significant and explain the variance in model very well.

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

Answer:

To evade the trap of dummy variables. Multicollinearity problems between dummy variables could result from the dummy variable trap. This could result in a violation of the linear regression's assumptions. Therefore, we only employ k-1 levels for dummy variable encoding if we have k levels where k >= 3. Intercept handles the lowered level as a default scenario.

Assume that there is a class of colors that includes Red, Green, and Blue. Nominal categorical variables include this category. There isn't a relationship or order between Red, Green, and Blue. Simply labeling them as 1, 2, or 3 won't work because it will confuse our model.
It could result in prejudice based on order, such as Red < Green < Blue. We dummy encode

scenarios such as this to prevent this. Additionally, as the model is unable to comprehend text or string data, it is required to translate these to numerical data.

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

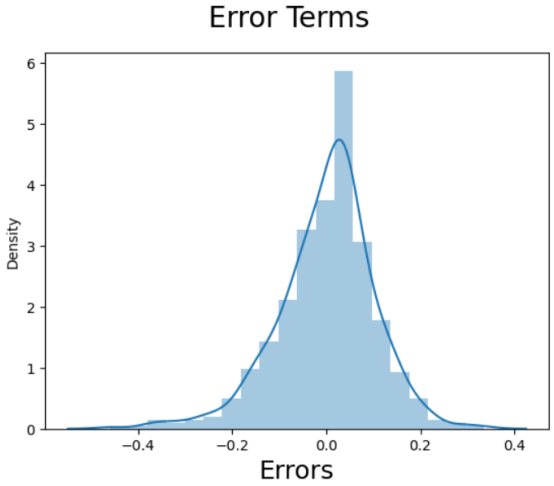
Answer:
Therefore, the pair plot displays the strongest correlation for the registered variable (correlation 0.945) prior to model creation and training. However, we are not training the model with random and registered data from our pre-processed training set. Registered + casual = CNT. This could cause the model to become overfit and leak important information.
After removing these two factors, the target variable, cnt, has the lowest association with atemp, which is followed by temp.
The correlation coefficient between atemp and cnt is 0.631, according to the correlation heatmap. Additionally, there is 0.627 link between temp and cnt.

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

Answer:

To validate assumptions of the model, and hence the reliability for inference, we go with the following procedures:

- Residual Analysis:
  We need to check if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression).I have plotted the histogram of the error terms and this is what it looks like:



  The residuals are following the normally distribution with a mean 0. All good!

- Linear relation between predictor variables and target variable:
  This happening because all the predictor variables are statistically significant (p-values are less than 0.05). Also, R-squared value on training set is 0.787 and adjusted R-Squared value on training set is 0.785. This means that variance in data is being explained by these predator variables.
- Error terms are independent of each other:
  Handled properly in the model. The predictor variables are independent of each other. Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5.

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

- temp (coef: 0.3857)
- yr (coef: 0.2371)
- windspeed (coef: -0.1502)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Answer:
   A key machine learning approach for supervised learning tasks, namely continuous prediction, is called linear regression. It creates a linear relationship between one or more independent variables (properties that affect the prediction) and a dependent variable (what you want to predict). Below is a summary of the salient features:

   Objective:

   Finding the linear equation that best fits a set of data points is the main goal of linear regression; for basic linear regression, this equation is a straight line. Next, predictions for fresh data points can be derived from this equation.

Types of Linear Regression:

- **Simple Linear Regression:** Involves a single independent variable and a continuous dependent variable.
- **Multiple Linear Regression:** Employs multiple independent variables to predict the dependent variable.

Steps Involved:

1. **Data Preparation:**

- Gather a dataset containing the independent and dependent variables you want to analyze.
- Ensure data quality by cleaning and handling missing values if necessary.

2. **Model Representation:**
   - The linear regression model is typically represented by the equation:
   - $y = \beta_0 + \beta_1 x + \varepsilon$

     where:

     - y: Dependent variable (what you want to predict)
     - β_0: Intercept (y-axis value where the line crosses)
     - β_1: Slope of the line (indicates the change in y for a unit change in x)
     - x: Independent variable (feature used for prediction)
     - ε: Error term (represents the difference between the actual y value and the predicted y value)
3. **Parameter Estimation (Finding the Best-Fit Line):**
   - Linear regression algorithms aim to minimize the error term (ε) across all data points. This essentially means finding the values for β_0 and β_1 that produce the line with the closest fit to the actual data points.
   - Common methods for parameter estimation include:
     - **Least Squares Method:** Calculates the sum of squared errors (SSE) between the predicted y values and the actual y values. The algorithm iteratively adjusts β_0 and β_1 to minimize SSE.
4. **Model Evaluation:**
   - Once the model parameters (β_0 and β_1) are estimated, it's crucial to evaluate the model's performance. Common metrics include:
     - **Mean Squared Error (MSE):** Average squared difference between predicted and actual y values. Lower MSE indicates a better fit.
     - **R-squared (coefficient of determination):** Represents the proportion of variance in the dependent variable explained by the independent variable(s). A value closer to 1 indicates a stronger linear relationship.
5. **Prediction:**
   - After model evaluation, you can use the estimated equation (with the determined β_0 and β_1) to predict the dependent variable (y) for new unseen data points (x values).

Underlying Assumptions:

- Linear regression relies on some key assumptions about the data:
  - **Linear Relationship:** The relationship between the independent and dependent variables should be close to linear.
  - **Homoscedasticity:** The variance of the error terms should be constant across all independent variable values.
  - **Normality:** The error terms should be normally distributed.

Applications:

Linear regression is widely used in various domains for tasks like:

- **Sales forecasting:** Predicting future sales based on historical data and market trends.
- **Stock price prediction:** Estimating future stock prices based on factors like past performance and economic indicators.
- **Customer churn prediction:** Identifying customers at risk of leaving a service based on their past behavior.
- **Medical diagnosis:** Supporting medical professionals in diagnosing diseases by analyzing patient data.

2. Explain the Anscombe's quartet in detail.
   Answer:

   Anscombe's quartet is a set of four datasets constructed by statistician Francis Anscombe in 1973. These datasets are particularly interesting because they all share the same basic statistical properties:

- **Mean (average) of x:** 9
- **Mean (average) of y:** 7.5
- **Variance of x:** 11
- **Standard deviation of x:** 3.3
- **Correlation coefficient between x and y:** 0.816 (strong positive correlation)
- **Linear regression line:** Approximately the same slope and intercept

Despite these identical summary statistics, the four datasets visually appear very different when plotted as scatter plots. This highlights a crucial point in data analysis: **descriptive statistics can be deceiving, and visualization is essential for understanding the underlying relationships in your data.**

Here's a breakdown of the four datasets:

1. **Linear:** This dataset has a clear linear relationship between x and y.
2. **Curved:** This dataset has a curved, non-linear relationship between x and y.
3. **Outlier:** This dataset has a single outlier that significantly impacts the linear regression line.
4. **Clustered:** This dataset has two distinct clusters of points, indicating a potentially more complex relationship.

The key takeaway from Anscombe's quartet is that relying solely on summary statistics can lead to misinterpretations of data. Visualization techniques like scatter plots are crucial for uncovering non-linear trends, outliers, and other patterns that might not be evident in numerical summaries.

Significance :

By demonstrating the limitations of relying solely on descriptive statistics, Anscombe's quartet emphasizes the importance of data visualization in exploratory data analysis (EDA). This helps us gain a deeper understanding of the data before applying statistical models or drawing conclusions.

3. What is Pearson’s R?
   Answer:

Pearson's R, also known as the Pearson correlation coefficient (PCC), is a statistical measure that quantifies the **linear relationship** between two continuous variables. It represents the strength and direction of the association between those variables.

Range and Interpretation:

- Pearson's R takes a value between **-1 and +1**.
  - **+1:** Indicates a perfect positive correlation, meaning as the value of one variable increases, the value of the other variable also increases proportionally.
  - **0:** Represents no linear correlation, meaning there's no predictable relationship between the two variables.
  - **-1:** Indicates a perfect negative correlation, meaning as the value of one variable increases, the value of the other variable decreases proportionally.

Interpretation Guidelines:

- While the exact interpretation of the strength of correlation can vary depending on the specific context, here's a general guideline:
  - **0.0 to 0.3:** Weak positive correlation
  - **0.3 to 0.7:** Moderate positive correlation
  - **0.7 to 1.0:** Strong positive correlation
  - **-0.0 to -0.3:** Weak negative correlation
  - **-0.3 to -0.7:** Moderate negative correlation
  - **-0.7 to -1.0:** Strong negative correlation

Important Considerations:

- Pearson's R only measures **linear relationships**. It doesn't capture non-linear relationships, which might exist in the data.
- It's not a measure of causation. Just because two variables are correlated doesn't mean one causes the other.
- The validity of Pearson's R depends on the assumptions of normality (data should be somewhat normally distributed) and homoscedasticity (variance should be constant across the independent variable's range).

Applications:

Pearson's R is widely used in various fields to assess the strength and direction of linear relationships between variables. Examples include:

- **Finance:** Analyzing the correlation between stock prices and economic indicators.
- **Psychology:** Investigating the relationship between personality traits and behaviors.
- **Biology:** Studying the correlation between environmental factors and plant growth.
- **Marketing:** Examining the association between advertising campaigns and sales figures.

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

Answer:

In machine learning, scaling refers to the process of transforming the values of features in your data to a common range. This is often done before training a model to improve its performance and stability.

Here's a breakdown of the key aspects of scaling:

Why Scaling is Performed:

- **Improves Model Performance:** Many machine learning algorithms rely on the distances between data points to learn. Features with vastly different ranges can lead to models giving more weight to features with larger scales, even if they may not be as significant. Scaling ensures all features contribute equally to the learning process.
- **Faster Convergence:** Standardization, a specific type of scaling, can often lead to faster convergence of gradient descent-based algorithms used in training models. This results in the model learning the underlying patterns in the data more efficiently.
- **Reduces Bias:** Scaling can help mitigate biases that may arise from features with different scales. For example, imagine predicting house prices where one feature is "price" (in millions) and another is "number of bedrooms" (an integer). Without scaling, the price would dominate the learning process due to its larger magnitude.

Types of Scaling:

There are two main types of scaling techniques used in machine learning:

1. **Normalization:**
    - Often scales features to a range between 0 and 1 (min-max scaling) or -1 and 1.
    - Useful when the absolute values of features are important. For example, normalizing income data might be appropriate if you're interested in the relative income levels.
2. **Standardization:**

- Transforms features to have a mean of 0 and a standard deviation of 1 (z-score normalization).
- Useful when the distribution of features is important and you want to focus on the relative differences from the mean. For example, standardizing exam scores might be appropriate if you're comparing student performance across different subjects.

Choosing the Right Technique:

The choice between normalization and standardization depends on the specific problem and the machine learning algorithm used. Here are some general guidelines:

- Use normalization if the presence of outliers or the absolute values of features are important.
- Use standardization if the distribution of features matters and you want to focus on relative differences from the mean.

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

A VIF (Variance Inflation Factor) value of infinity in linear regression indicates **perfect multicollinearity** among the independent variables (predictors) in your model. This means that one or more independent variables can be perfectly predicted by a linear combination of the other independent variables.

Here's why perfect multicollinearity leads to an infinite VIF:

VIF Calculation:

The VIF for a particular independent variable is calculated as:

$VIF_i = 1 / (1 - R_i^2)$

- VIF_i represents the VIF for the i-th independent variable.
- R_i^2 represents the R-squared value obtained when you regress the i-th independent variable on all other independent variables in the model.

Perfect Multicollinearity:

In the case of perfect multicollinearity, one independent variable can be entirely predicted by a linear combination of the others. This essentially means that R_i^2 for that variable will be equal to 1.

Infinite VIF:

When R_i^2 becomes 1, the denominator of the VIF formula $(1 - R_i^2)$ becomes 0. Dividing by 0 results in an indeterminate form, which is often represented mathematically as infinity.

Consequences of Perfect Multicollinearity:

- **Unreliable Coefficients:** The regression coefficients for the independent variables become unreliable and difficult to interpret. Coefficients might have high standard errors, making it challenging to assess their statistical significance.
- **Inflated Variance:** The variances of the regression coefficients become artificially inflated, making it appear like the variables have a stronger effect than they actually do.
- **Model Instability:** The model becomes highly sensitive to small changes in the data, leading to poor predictions and unreliable results.

How to Handle Infinite VIF:

Here are some approaches to address infinite VIF and multicollinearity:

- **Identify Collinear Variables:** Techniques like correlation analysis and looking at the variance inflation factors (even non-infinite values can indicate multicollinearity) can help identify problematic variables.
- **Remove Redundant Variables:** If variables are highly correlated and providing redundant information, consider removing one (or more) from the model based on domain knowledge or feature importance analysis.
- **Combine Variables:** If appropriate, explore combining collinear variables into a single feature that captures the shared information.
- **Regularization Techniques:** Techniques like Lasso or Ridge regression can be used to penalize large coefficients, reducing the impact of multicollinearity and improving model stability.

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

   Answer:
   A Q-Q plot (quantile-quantile plot) is a graphical tool used to assess the **distribution of residuals (errors)** in linear regression. It compares the quantiles (percentiles) of your actual residuals with the quantiles of a theoretical distribution, typically the normal distribution.

How it Works:

1. **Quantile Calculation:**
   - Both the residuals from your fitted linear regression model and the quantiles of the theoretical distribution (e.g., normal) are sorted in ascending order.
2. **Plotting the Quantiles:**

    - The quantiles of the residuals are plotted on the x-axis, while the quantiles of the theoretical distribution are plotted on the y-axis.

Interpretation:

- **Ideal Scenario:** If the residuals follow the theoretical distribution (usually normal), the points in the Q-Q plot should fall approximately along a straight diagonal line. This indicates that the errors in your model are well-behaved and meet the assumptions of normality for linear regression.
- **Deviations from the Line:** Deviations from the straight line suggest potential issues with the normality of the errors.
    - **Points below the line:** This might indicate right-skewness (long tail towards positive values).
    - **Points above the line:** This might indicate left-skewness (long tail towards negative values).
    - **Curved patterns:** These can suggest non-linearity in the relationship between the independent and dependent variables.

Importance in Linear Regression:

1. **Evaluating Model Assumptions:** Linear regression relies on the assumption that the errors are normally distributed. A Q-Q plot helps you to visually assess this assumption.
2. **Identifying Potential Issues:** Deviations from the straight line can indicate potential problems like heteroscedasticity (unequal variance of errors) or outliers that might affect the model's accuracy.
3. **Improving Model Performance:** By addressing issues with normality or other problems revealed by the Q-Q plot, you can potentially improve the performance and reliability of your linear regression model.

Additional Considerations:

- Q-Q plots are particularly helpful for visually assessing normality, but they should be used in conjunction with other diagnostic tools like Shapiro-Wilk test to confirm the findings.
- While normality is often preferred, there are cases where linear regression can be relatively robust to violations of normality. However, severe deviations can still impact model performance.

In summary, a Q-Q plot is a valuable tool for understanding the distribution of errors in linear regression. By interpreting the pattern of points, you can gain insights into the validity of model assumptions and identify potential issues that might affect the reliability of your analysis.