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?**

The final Multiple Linear Regression model contains many predictor variables that are categorical in nature and some of them have been encoded to dummy variables.

```
                           OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.798
Model:                            OLS   Adj. R-squared:                  0.793
Method:                 Least Squares   F-statistic:                     151.0
Date:                Wed, 29 May 2024   Prob (F-statistic):          3.79e-163
Time:                        20:16:36   Log-Likelihood:                 446.95
No. Observations:                 511   AIC:                            -865.9
Df Residuals:                     497   BIC:                            -806.6
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4506      0.019     24.203      0.000       0.414       0.487
year           0.2463      0.009     27.070      0.000       0.228       0.264
workingday     0.0569      0.012      4.575      0.000       0.032       0.081
windspeed     -0.1918      0.028     -6.803      0.000      -0.247      -0.136
spring        -0.2433      0.016    -15.312      0.000      -0.275      -0.212
summer        -0.0433      0.014     -3.164      0.002      -0.070      -0.016
winter        -0.0123      0.016     -0.766      0.444      -0.044       0.019
dec           -0.1126      0.019     -5.966      0.000      -0.150      -0.076
jan           -0.1222      0.020     -6.249      0.000      -0.161      -0.084
nov           -0.1046      0.020     -5.107      0.000      -0.145      -0.064
sep            0.0542      0.018      2.986      0.003       0.019       0.090
sat            0.0662      0.016      4.128      0.000       0.035       0.098
rainy         -0.2293      0.028     -8.186      0.000      -0.284      -0.174
sunny          0.0886      0.010      9.112      0.000       0.069       0.108
==============================================================================
Omnibus:                       67.982   Durbin-Watson:                   1.968
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              173.760
Skew:                          -0.675   Prob(JB):                     1.86e-38
Kurtosis:                       5.518   Cond. No.                         10.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
    Features   VIF
```

```
2    windspeed  4.29
1   workingday  3.39
5       winter  2.79
3       spring  2.77
12       sunny  2.39
4       summer  1.97
0         year  1.92
8          nov  1.76
7          jan  1.64
10         sat  1.58
6          dec  1.46
9          sep  1.19
11       rainy  1.12
```

Categorical Variable list:

winter
spring
sunny
summer
year
nov
jan
sat
dec
sep
rainy

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

The drop_first=True parameter in the get_dummies function of pandas library is used to avoid the "dummy variable trap".

This trap is a state of perfect multicollinearity, where there is a perfect correlation between the dummy variables. When you have N categories for a feature, creating N dummy variables introduces perfect multicollinearity because the information of the N-1 variables completely determines the Nth one.

This can disrupt calculations in algorithms that require matrix inversion, such as Linear Regression, making the estimates of your model's parameters highly sensitive to changes in the model specification. By using drop_first=True, you create N-1 binary variables, thereby omitting the first one and ensuring that your dummy variables are not multicollinear.

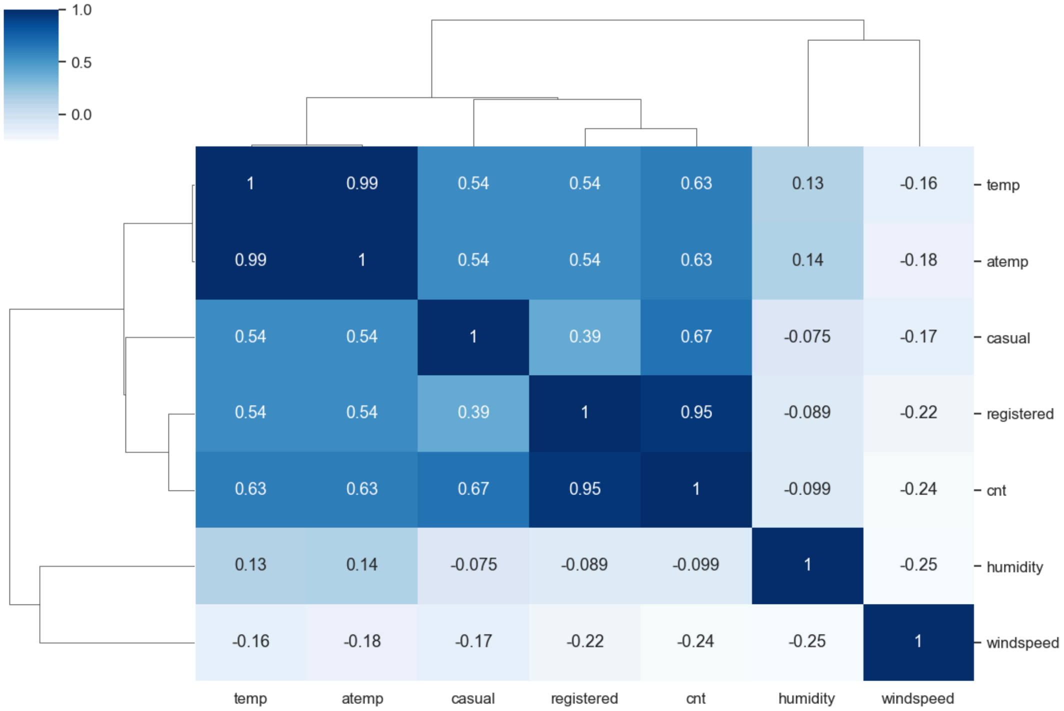
This technique is also known as "one-hot encoding". Note that the necessity of this approach depends on the algorithm you're using, as some can handle multicollinearity without issues.

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

So, before model building and training, the pair plot shows highest correlation for registered variable having correlation 0.95. But we are not using casual and registered in our pre-processed training data for model training. casual + registered = cnt.

This might leak out the crucial information and model might get overfit. So, excluding these two variables atemp is having highest correlation with target variable cnt which is followed by temp.

As per the correlation heatmap, correlation coefficient between atemp and cnt is 0.63. And correlation coefficient between temp and cnt is 0.63.

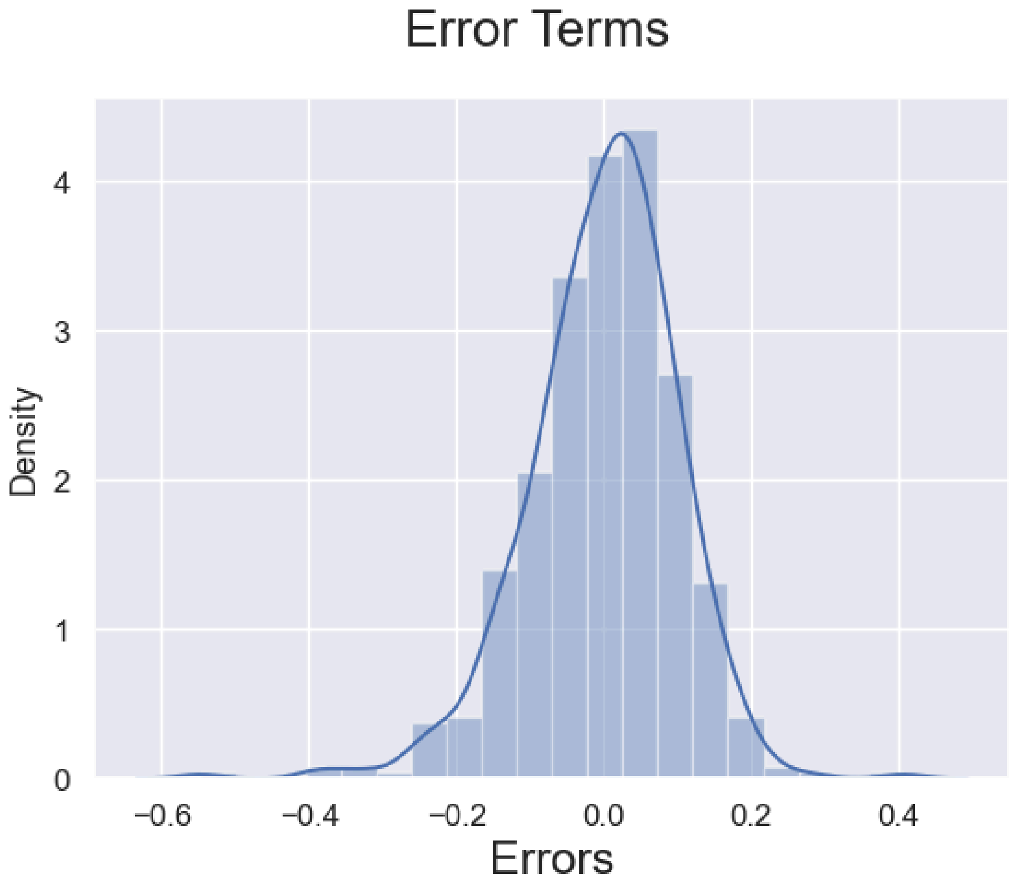


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

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 relationship between predictor variables and target variable:

This is happening because all the predictor variables are statistically significant (p-values are less than 0.05). Also, R-Squared value on training set is 0.798 and adjusted R-Squared value on training set is 0.793. This means that variance in data is being explained by all these predictor 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 and P-Value is less than the 0.05

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

Top 3 features significantly contributing towards demand of shared bikes are:

1) ***year***

2) ***sept***

3) rainy

```
                        OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.798
Model:                            OLS   Adj. R-squared:                  0.793
Method:                 Least Squares   F-statistic:                     151.0
Date:                Wed, 29 May 2024   Prob (F-statistic):          3.79e-163
Time:                        20:16:36   Log-Likelihood:                 446.95
No. Observations:                 511   AIC:                            -865.9
Df Residuals:                     497   BIC:                            -806.6
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4506      0.019     24.203      0.000       0.414       0.487
year           0.2463      0.009     27.070      0.000       0.228       0.264
workingday     0.0569      0.012      4.575      0.000       0.032       0.081
windspeed     -0.1918      0.028     -6.803      0.000      -0.247      -0.136
spring        -0.2433      0.016    -15.312      0.000      -0.275      -0.212
summer        -0.0433      0.014     -3.164      0.002      -0.070      -0.016
winter        -0.0123      0.016     -0.766      0.444      -0.044       0.019
```

```
dec          -0.1126      0.019     -5.966      0.000      -0.150      -0.076
jan         -0.1222      0.020     -6.249      0.000      -0.161      -0.084
nov          -0.1046      0.020     -5.107      0.000      -0.145      -0.064
sep           0.0542      0.018      2.986      0.003       0.019       0.090
sat          0.0662      0.016      4.128      0.000      0.035       0.098
rainy        -0.2293      0.028      -8.186      0.000      -0.284      -0.174
sunny         0.0886      0.010      9.112      0.000       0.069       0.108
==============================================================================
Omnibus:                      67.982   Durbin-Watson:                   1.968
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              173.760
Skew:                        -0.675   Prob(JB):                    1.86e-38
Kurtosis:                     5.518   Cond. No.                        10.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
     Features   VIF
2    windspeed  4.29
1   workingday  3.39
5       winter  2.79
3       spring  2.77
12       sunny  2.39
4       summer  1.97
0         year  1.92
8          nov  1.76
7          jan  1.64
10         sat  1.58
6          dec  1.46
9          sep  1.19
11       rainy  1.12
```

General Subjective Questions

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

Linear regression serves as a predictive statistical tool, primarily employed when the target variable is continuous. In its simplest form, known as simple linear regression, it establishes a relationship between a dependent variable 'y' and a single independent variable 'x' through a straight line.

The model equation, y = bx + a + e, delineates this relationship, where 'b' signifies the slope (the impact of X on Y), 'a' represents the y-intercept (the value of y when x=0), and 'e' embodies the error term (the variance between observed and predicted values).

With multiple linear regression, the model accommodates more than one independent variable. The algorithm's objective entails determining the most optimal fitting line by minimizing the sum of squared residuals, thereby enhancing predictive accuracy.

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It's a type of regression analysis where the goal is to find the best-fitting straight line (or hyperplane in higher dimensions) to describe the relationship between the variables.

Here's a detailed breakdown of the algorithm:

Purpose: Linear regression predicts the relationship between a dependent variable and independent variables.

Data Collection: Gather data with dependent and independent variables.

Preprocessing: Clean data, handle missing values, and split into training/testing sets.

Model Selection: Choose linear regression assuming a linear relationship.

Model Equation: Y=β0+β1X+εY = \beta_0 + \beta_1 X + \varepsilonY=β0+β1X+ε for simple linear regression.

Fitting: Estimate coefficients minimizing the difference between predicted and actual values.

Evaluation: Assess performance using metrics like MSE, MAE, R2R^2R2.

Interpretation: Understand relationships via coefficients; positive/negative indicates the direction.

Prediction: Use the model for predictions on new data.

Validation: Ensure model generalizes well on unseen data.

Improvement: Refine model based on performance, feature selection, or transformations

In multiple linear regression, the model includes more than one independent variable. The goal of the algorithm is to find the best fitting line by minimizing the sum of squared residuals.

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

Anscombe's quartet is a set of four datasets that have nearly identical descriptive statistics but exhibit significantly different relationships when plotted. This demonstration highlights the importance of visualizing data before drawing conclusions.

For instance, consider four datasets with the following properties:

1. Dataset 1: Linear relationship
2. Dataset 2: Non-linear relationship
3. Dataset 3: Outlier significantly affecting regression
4. Dataset 4: Perfectly linear except for one outlier pair

Despite having the same mean, variance, correlation, and regression line parameters, when graphed, each dataset displays unique patterns. Dataset 1 shows a clear linear relationship, Dataset 2 has a non-linear pattern, Dataset 3's outlier distorts the regression line, and Dataset 4's outlier pair significantly influences the regression line.

Anscombe's quartet underscores the importance of not solely relying on summary statistics and emphasizes the need for visual exploration to understand the true nature of data relationships. It serves as a cautionary example against drawing conclusions solely based on numerical summaries.

3. What is Pearson's R?

Pearson's correlation coefficient, denoted as r, quantifies the strength and direction of the linear relationship between two continuous variables. Ranging from -1 to 1, r signifies:

- r=1 implies a perfect positive linear relationship,
- r=−1 indicates a perfect negative linear relationship, and
- r=0 suggests no linear relationship.

Calculated by dividing the covariance of the variables by the product of their standard deviations, r assesses how much one variable changes concerning another. While widely used in fields like psychology and economics, rrr assumes linearity between variables and may not capture non-linear associations or be robust to outliers. Its interpretation provides valuable insights into data patterns, guiding decision-making processes and research endeavors.

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

Scaling is a preprocessing step in data analysis and machine learning to transform features to a common scale. It ensures equal contribution of features, improves algorithm performance, and aids convergence. Normalized scaling (Min-Max) rescales features to a fixed range (typically 0 to 1), while standardized scaling (z-score) standardizes features to have a mean of 0 and a standard deviation of 1. Normalization maintains the original distribution shape, while standardization centers data around 0. The choice depends on the problem and data characteristics.

Normalized scaling compresses or stretches data to a fixed range, maintaining original distribution shape. Standardized scaling centers data around 0, with a standard deviation of 1, aiding interpretability. Both methods address algorithm sensitivity and ensure fair feature contribution in machine learning models.

Normalized scaling (Min-Max) is suitable when features have bounded ranges, while standardized scaling (z-score) is preferable for algorithms sensitive to feature distributions. Selecting the appropriate scaling method depends on the specific requirements of the analysis and the characteristics of the dataset.

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

The phenomenon of VIF (Variance Inflation Factor) reaching infinity occurs due to perfect multicollinearity, where one independent variable in a regression model is perfectly predictable from others. This condition arises when variables are linearly dependent, such as one variable being a constant multiple of another or being a linear combination of others. When calculating VIF, a key step involves inverting the correlation matrix of the independent variables. However, in cases of perfect multicollinearity, this matrix becomes singular, meaning it cannot be inverted, leading to an undefined result or infinity.

Perfect multicollinearity leading to infinite VIF values can be illustrated with a simple example. Suppose we have a dataset with two independent variables, X1 and X2, where X2 is exactly twice the value of X1 for every observation. In this scenario, there is perfect multicollinearity between X1 and X2 because one variable can be expressed as a perfect linear combination of the other.

When calculating the VIF for X2, the formula involves the inversion of the correlation matrix of all independent variables. However, because X2 is a perfect multiple of X1, the correlation matrix becomes singular, meaning it cannot be inverted. As a result, the VIF calculation for X2 breaks down, yielding an infinite value.

This situation has significant implications for regression analysis. The infinite VIF indicates that the variance of the regression coefficient for X2 cannot be estimated accurately due to perfect multicollinearity. Consequently, coefficient estimates become unreliable, and the model's interpretability is compromised.

To address this issue, one approach is to remove either X1 or X2 from the model. Alternatively, if theoretically justified, the two variables could be combined into a single variable. Understanding and managing multicollinearity is essential for ensuring the robustness and validity of regression models in data analysis.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a given dataset follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution, typically a normal distribution.

In linear regression, Q-Q plots are crucial for evaluating the assumption of normality of residuals, which is essential for the validity of statistical inference. Here's how it works:

1. **Construction**: The Q-Q plot compares the quantiles of the observed residuals (vertical axis) against the quantiles of a theoretical normal distribution (horizontal axis).

2. **Interpretation**: If the residuals are normally distributed, the points on the Q-Q plot will fall approximately along a straight line. Any deviations from a straight line indicate departures from normality.
3. **Importance**: A Q-Q plot helps identify patterns in the residuals that violate the assumption of normality. Non-linear patterns or significant deviations from the straight line suggest that the residuals may not be normally distributed.
4. **Example**: Suppose we fit a linear regression model to predict house prices based on various features. After obtaining the residuals (observed minus predicted values), we create a Q-Q plot. If the plot shows a reasonably straight line, it indicates that the residuals follow a normal distribution, supporting the assumption of normality in linear regression analysis.

In summary, Q-Q plots provide a visual assessment of the normality assumption of residuals in linear regression, aiding in model validation and ensuring the reliability of statistical inference.