Assignment-based Subjective Questions

1. Effect of Categorical Variables on the Dependent Variable

```
In [222]: #Let's see the summary of our linear model
          print(lm.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.770
Model:                            OLS   Adj. R-squared:                  0.765
Method:                 Least Squares   F-statistic:                     151.2
Date:                Wed, 28 Aug 2024   Prob (F-statistic):          6.62e-151
Time:                        10:07:16   Log-Likelihood:                 388.31
No. Observations:                 510   AIC:                            -752.6
Df Residuals:                     498   BIC:                            -701.8
Df Model:                          11
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                            0.4694      0.018     26.363      0.000       0.434       0.504
windspeed                       -0.1324      0.027     -4.993      0.000      -0.184      -0.080
season_spring                   -0.2861      0.013    -21.369      0.000      -0.312      -0.260
season_winter                   -0.0871      0.015     -5.903      0.000      -0.116      -0.058
yr_1                             0.2578      0.010     25.244      0.000       0.238       0.278
mnth_5                           0.0215      0.019      1.137      0.256      -0.016       0.059
mnth_9                           0.1136      0.019      6.115      0.000       0.077       0.150
mnth_10                          0.1068      0.022      4.877      0.000       0.064       0.150
weekday_6                        0.0626      0.018      3.467      0.001       0.027       0.098
workingday_1                     0.0672      0.013      5.041      0.000       0.041       0.093
weathersit_Light_Snow_Rain      -0.2883      0.034     -8.453      0.000      -0.355      -0.221
weathersit_Mist_Cloudy          -0.0948      0.011     -8.807      0.000      -0.116      -0.074
==============================================================================
Omnibus:                       51.341   Durbin-Watson:                   2.118
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              128.173
Skew:                          -0.518   Prob(JB):                     1.47e-28
Kurtosis:                       5.227   Cond. No.                         10.2
==============================================================================

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

- There are a number of categorical variables in the final model.
- Categorical variables like season, weather conditions, and whether it's a working day have a significant impact on bike rentals.
- For instance, bike demand drops notably during the spring season and on days with poor weather conditions such as light snow or rain.
- Months (mnth_5, mnth_9 and mnth_10), also have an impact on bike rentals.
- Additionally, the year variable shows that demand has increased over time, suggesting a growth trend in bike-sharing popularity.

2. Importance of Using `drop_first=True` in Dummy Variable Creation

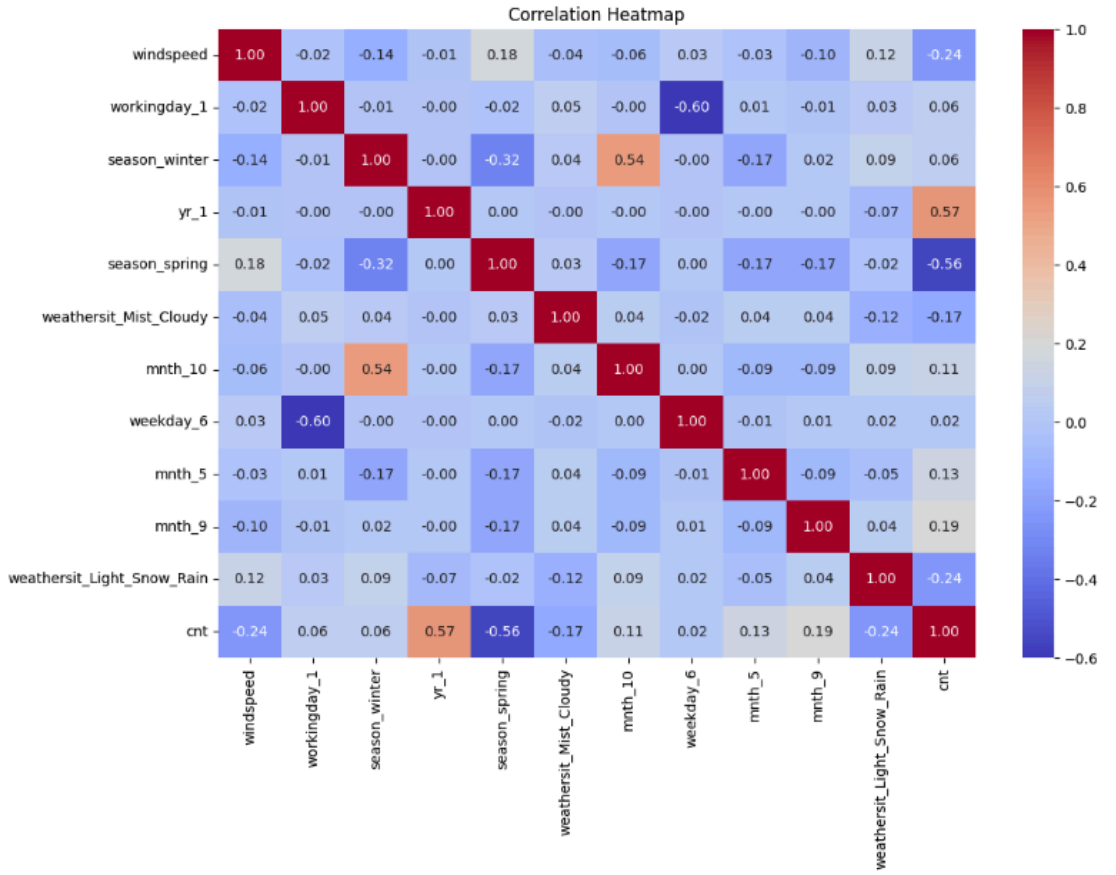
`Drop_first=True` in Dummy Variable Creation results in k-1 dummies out of k categorical levels by removing the first level. (https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html)

Having k-1 dummy variables helps to avoid multicollinearity. By dropping the first category, we prevent the dummy variables from being perfectly collinear, ensuring that the model remains stable and interpretable.

3. Highest Correlation with the Target Variable

```
In [75]: # Create a correlation matrix
         corr_matrix = df[final_columns].corr()

         # Plot the heatmap
         plt.figure(figsize=(12, 8))
         sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f')
         plt.title('Correlation Heatmap')
         plt.show()
```



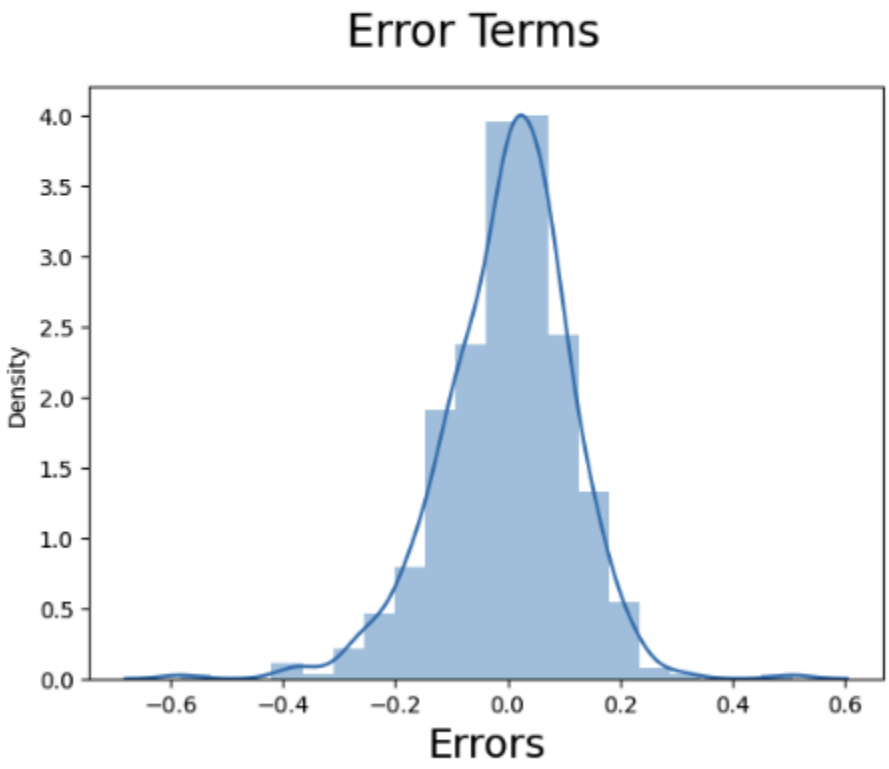
From the heatmap, it appear that tr_1 has the highest correlation with cnt at 0.57.

4. Validation of Linear Regression Assumptions

In [65]:
```
# Importing the required libraries for plots.
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [66]:
```
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_price), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
plt.xlabel('Errors', fontsize = 18)                         # X-label
```

Out[66]: Text(0.5, 0, 'Errors')



Conclusion: The errors look roughly normally distributed, and acceptable to move on to the next step

To validate the assumptions of Linear Regression, I checked for linearity by plotting residuals against predicted values, ensuring that they were randomly distributed. The error terms appear

to be normally distributed. Lastly, multicollinearity was assessed using Variance Inflation Factor (VIF) to ensure that the predictors were not highly correlated with each other.

5. Top 3 Features Contributing to Bike Demand

- yr_1: This has the highest positive coefficient (0.2578) and is highly significant (p-value = 0.000). This suggests that the year 2021 had a substantial positive impact on bike demand.
- season_spring: This has a large negative coefficient (-0.2861) with high significance (p-value = 0.000), indicating that spring significantly reduces bike demand.
- weathersit_Light_Snow_Rain: This also has a notable negative coefficient (-0.2883) and is highly significant (p-value = 0.000), showing that light snow or rain significantly decreases bike demand.

General Subjective Questions

1. Linear Regression Algorithm Linear regression is a statistical method used to model the relationship between a dependent variable and independent variables. We have learned two types of linear regression so far, with one independent variable (simple linear regression) and multiple independent variables (multiple linear regression). The algorithm seeks to find the best-fitting straight line (or hyperplane in multiple dimensions) through the data points by minimizing the sum of the squares of the vertical distances of the points from the line (least squares method). The resulting line, or equation, allows us to predict the dependent variable based on new values of the independent variables.

2. Anscombe's Quartet Anscombe's Quartet is a group of four datasets that have nearly identical simple descriptive statistics but appear very different when graphed. It illustrates the importance of graphing data before analyzing it and shows that statistical properties like mean, variance, and correlation don't provide a full picture without visual context.

3. Pearson's R Pearson's R, also known as the Pearson correlation coefficient, measures the linear correlation between two variables, giving a value between -1 and 1. A value closer to 1 indicates a strong positive correlation, while a value closer to -1 indicates a strong negative correlation. A value around 0 suggests no linear correlation.

4. Scaling Scaling is the process of adjusting the range of data features. It's performed to ensure that all features contribute equally to the model. Normalized scaling adjusts the values to a 0-1 range, whereas standardized scaling adjusts them to have a mean of 0 and a standard deviation of 1. The choice between the two depends on the specific needs of the model and the dataset.

5. Infinite VIF Values An infinite VIF value typically occurs when there's perfect multicollinearity in the dataset, meaning one predictor variable can be perfectly predicted by a linear combination of other predictors. This situation causes instability in the model and suggests that the variable in question may need to be removed or combined with others.

6. Q-Q Plot in Linear Regression A Q-Q plot, or Quantile-Quantile plot, compares the distribution of residuals to a normal distribution. It's used to check the normality assumption of residuals in linear regression. If the residuals lie along the reference line in the Q-Q plot, they are normally distributed. This is important because normality of residuals is one of the key assumptions of linear regression, affecting the reliability of the model's estimates and confidence intervals.