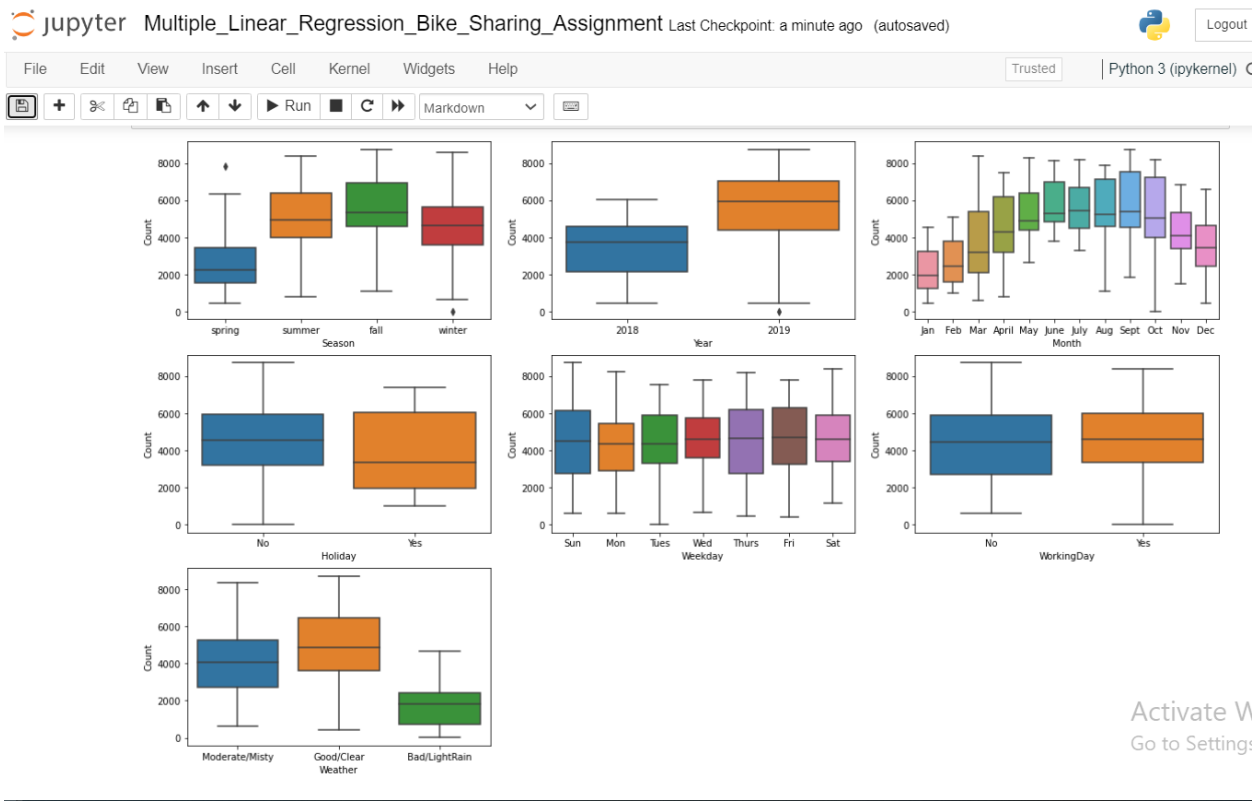
Assignment-based Subjective Questions

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

- There are total 7 Categorical variables in the dataset as follows: Season, Year, Month, Holiday, Weekday, WorkingDay, Weather.
- We assigned the actual values to the Categorical variables as per the data dictionary. Then we plot the boxplot of all these variables against the target variable i.e., Count, the result looks like:



- We can infer the following data after analyzing the above boxplots:
  1. In fall, there seems to be highest demand of rented the bikes, followed by Summer and Winter
  2. Spring seems to be the least season where people rent bikes
  3. Average rented bikes have increased in 2019 almost double that of 2018
  4. There is almost similar average count of rented bikes in August, June, September, July followed by May, October. Company should make sure they prepare with high availability during these months

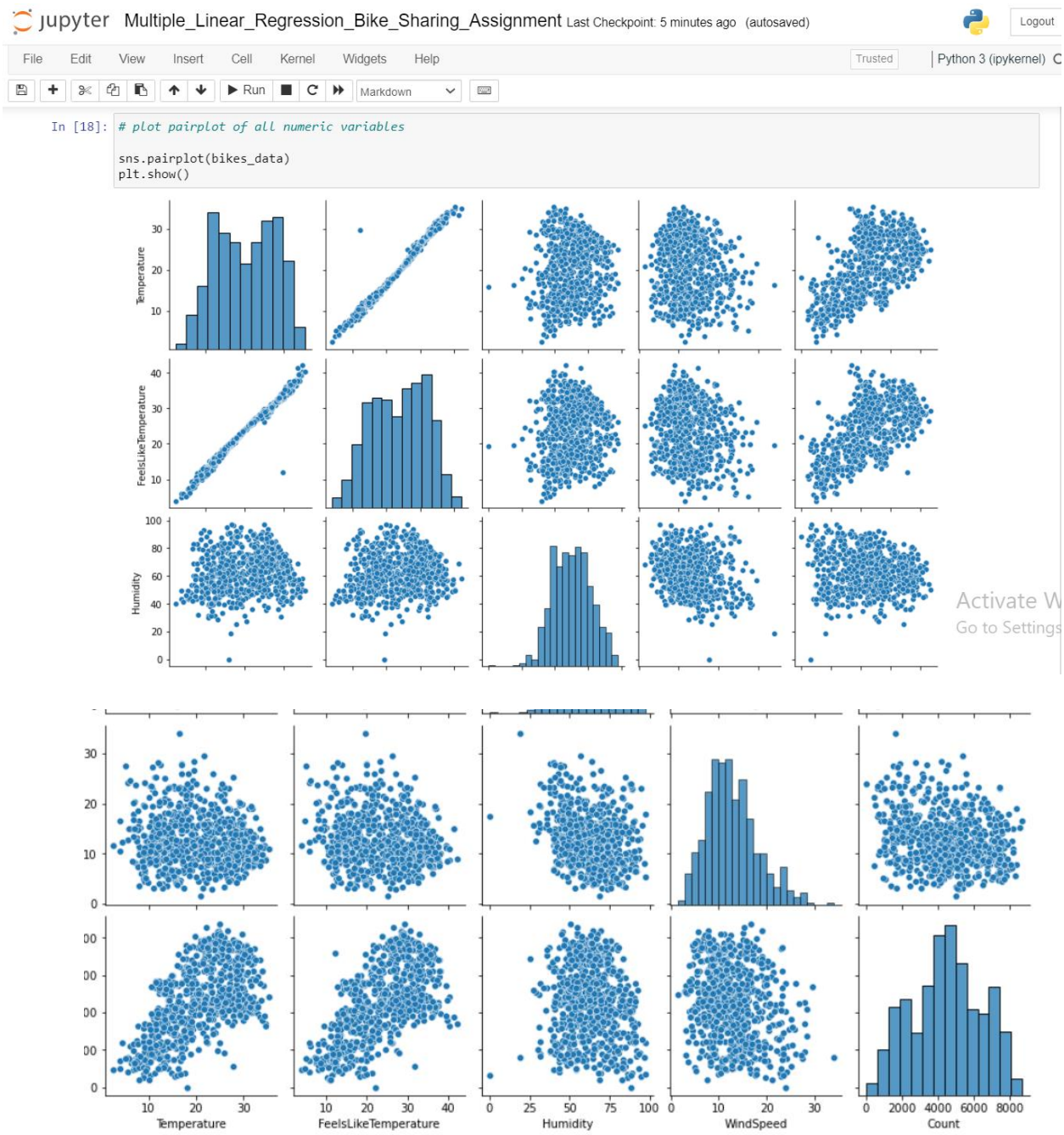
5. December, January, February has the least demand probably due to winter season
6. It looks like all days have similar demands
7. There are almost similar demands whether it's a working day or not
8. It clearly shows that if the weather is clear, the demand is more
9. If the weather is bad, demand decreases drastically
10. Company should leverage and look up for weather forecast to fulfill demands

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

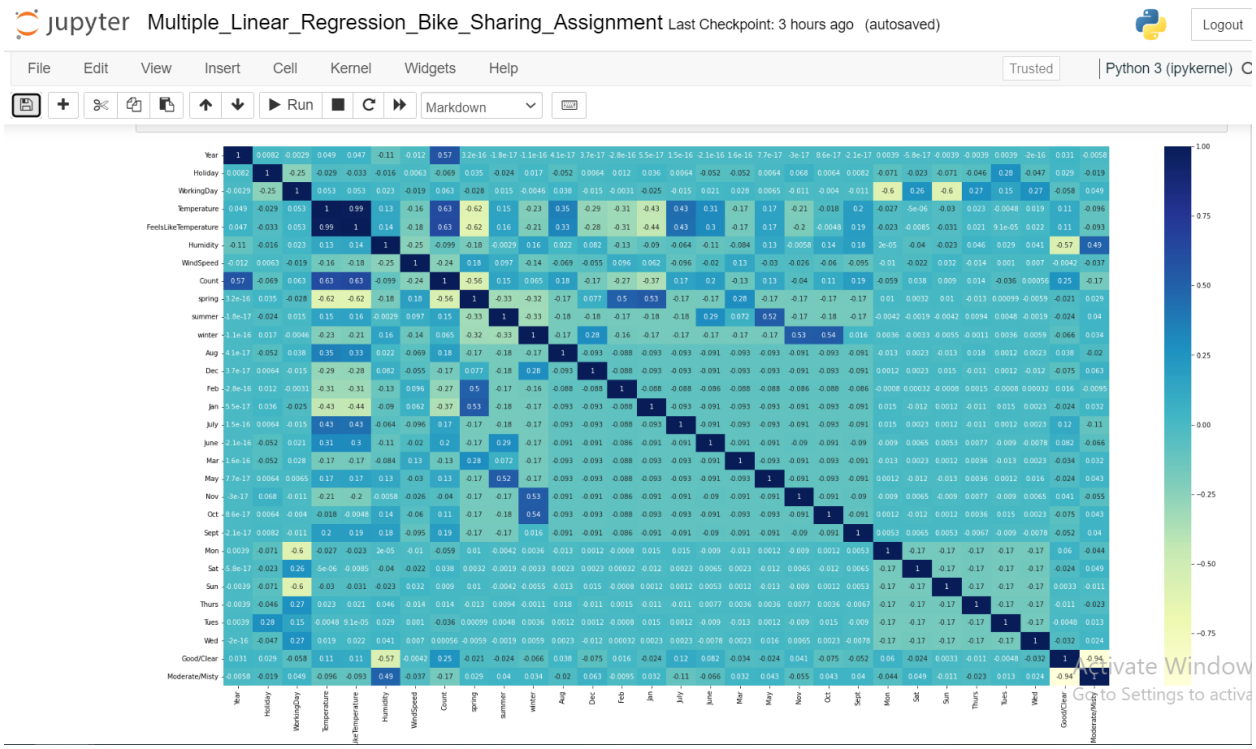
- When we create dummy variables for the Categorical variables, it creates the same number of dummy variables as the levels of that categorical variable.
- Suppose for a categorical variable with n levels, if we do not use drop_first=True, it creates n dummy variables. But we can also represent those values with n-1 dummy variables.
- e.g., Suppose a Categorical variable is 'Furnishing Status' then if one variable is not semi-furnished and furnished, then it is obviously unfurnished.
- So **Categorical variable with n levels** can be represented with **n-1 dummy variables**.
- So, we need to use *drop_first=True,* as it **reduces the extra column created during dummy variable creation**. Hence, it **reduces the correlation created between the dummy variables**.

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

- There are total 5 numeric variables in dataset – temp (Temperature), atemp (FeelsLikeTemperature), Hum (Humidity), windspeed (WindSpeed), cnt (Count)
- Please find below the pair-plots of all these variables:



- Please find below the Correlation Matrix of all the variables (Numeric and Categorical):



- From Above pair-plots and Correlation Matrix, we can see that both columns **FeelsLikeTemperature (atemp) and Temperature (temp)** have same and the highest correlation with target variable Count which is **0.63**.

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

- After the final module is built on training set, for which –
  - P-value of all the variables is within acceptable range i.e., < 0.05
  - VIF of all the variables is within acceptable range i.e., < 5

  We need to find the residual
- **Residual = actual_y – predicted_y**
- This is also known as 'Error Terms'
- These Error terms needs to be Normally distributed and centered at 0.
- To verify this, we need to plot the histogram of error terms and check how it looks.
- Please fins below the histogram of error terms of my final Model:

jupyter Multiple_Linear_Regression_Bike_Sharing_Assignment Last Checkpoint: 11 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Markdown

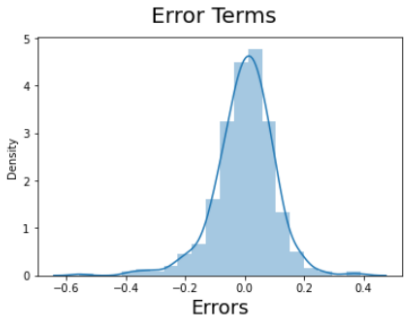
Step 6: Residual Analysis of the train data

So, now to check if the error terms are also normally distributed (which is one of the major assumptions of linear regression), let's plot the histogram of the error terms and see what it looks like.

In [72]:
```
y_train_pred = lm.predict(X_train_lm9)
```

In [73]:
```
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_pred), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
plt.xlabel('Errors', fontsize = 18)                         # X-label
```

Out[73]: Text(0.5, 0, 'Errors')



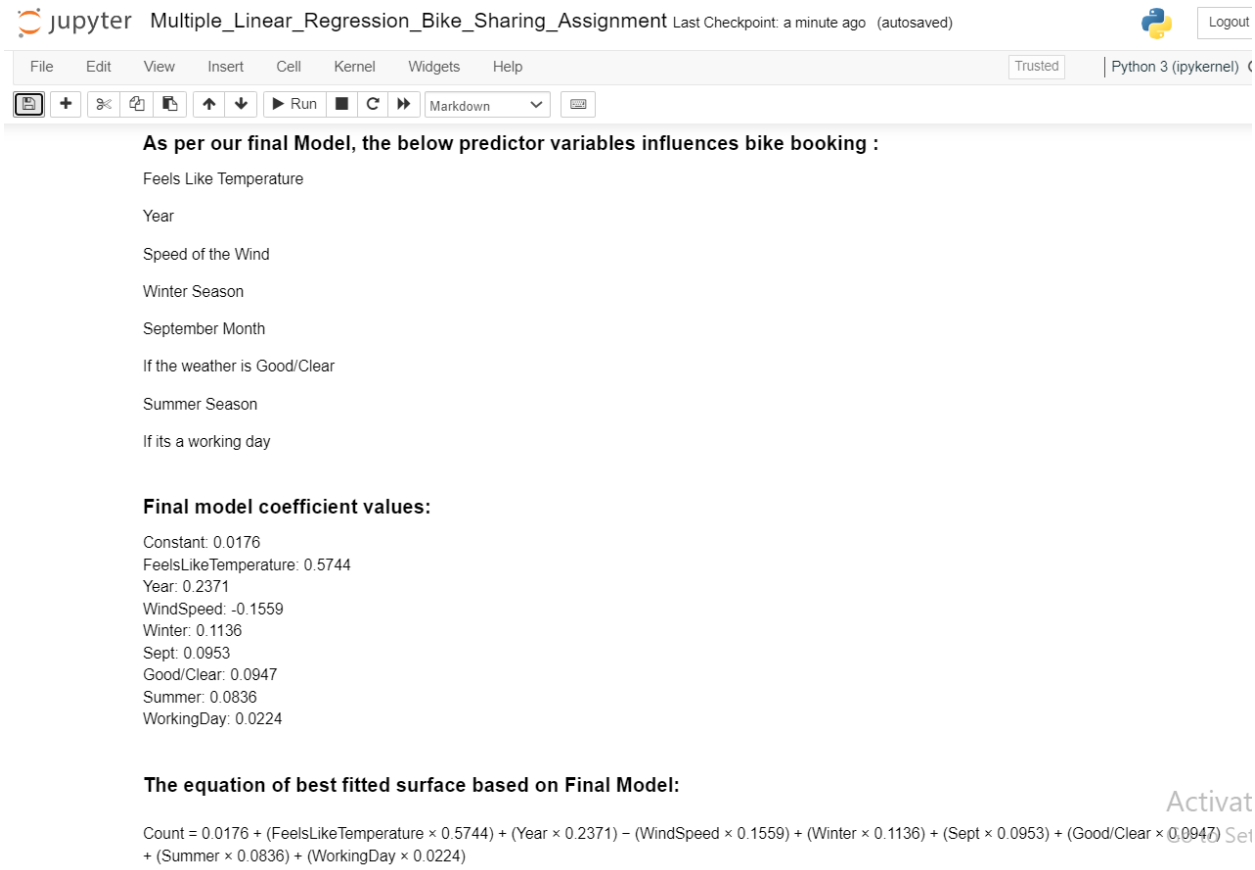
Observations:

- Error terms is normally distributed with mean 0. So assumption of Linear Regression is valid.

- From the above histogram, we can see that the Error terms is normally distributed with mean 0.
- Hence, Assumption of Linear regression is valid.

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

- Based on the final Model, below are the top 3 features which contributes significantly towards the demands of shared bikes:
    1. **Feels like Temperature**
    2. **Year**
    3. **Speed of Wind**
- Please find below the Coefficient values of the features from final Model:



General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear Regression in machine learning algorithm based on Supervised Learning. In Linear regression, we train a model to predict the behavior of the data based on some variables.
- Variables on x-axis and y-axis should be linearly correlated. Mathematically, we can write a linear regression equation as:

  $Y = mx + c$

  Where,

  m = Slope of the line

  c = y intercept of the line

  x = Independent variable from dataset

  y = Dependent variable from dataset
- Find the best fit line for the given scatter plot.
- Residual: It is a difference of actual value of y for a given value of x and predicted value of y calculated from the best fit line for the same value of x.

- Calculate the residuals for each point and take sum of squares of all residuals. It is called **'Residual Sums of Squares' (RSS).**
- Cost Function: By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum.
- Cost function of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).

2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed.
- It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.
- There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x, y points in all four datasets.
- This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.
- Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

3. What is Pearson's R?

- The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation.
- It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.
- Please find below the values and interpretation of it:

Pearson correlation coefficient (r)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is no relationship between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

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

- Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
- It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

- Normalization/Min-Max Scaling:

  It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

- Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

- If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables.
- In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.
- To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.
- A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.
- The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.
- A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
- A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.