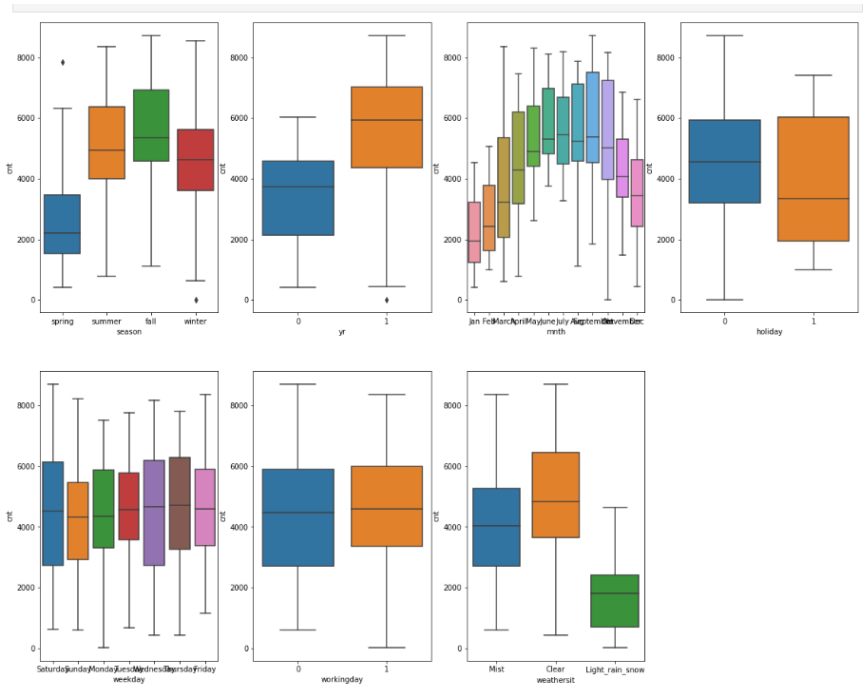
Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- The cnt of ride is between 4100 – 6500 in fall season as shown below.
- The cnt was more in the year 2019.
- The demand of bike sharing was more in weathersit as “Clear, Few clouds, Partly cloudy, Partly cloudy”



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

Answer:

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Suppose, you have a column for gender that contains 4 variables- "Male", "Female", "Other", "Unknown". So, a person is either "Male", or "Female", or "Other". If they are not either of these 3, their gender is "Unknown".

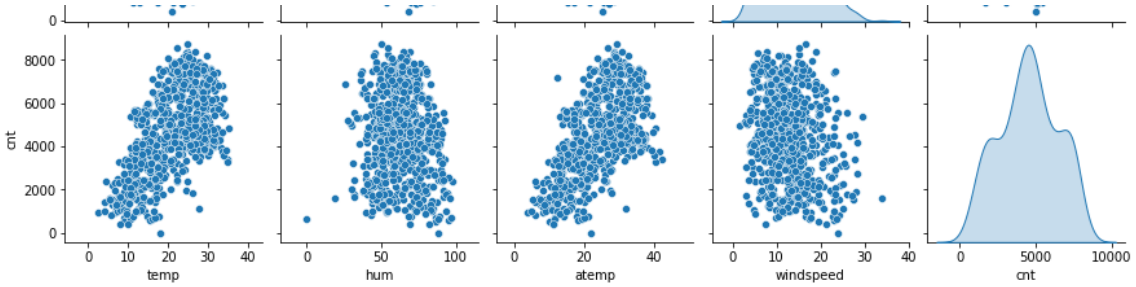
In below screenshot the season_1 is dropped.

```
In [72]: bike_ds_prep = pd.get_dummies(bike_ds_prep, drop_first=True)
         bike_ds_prep.info()

         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 730 entries, 0 to 729
         Data columns (total 30 columns):
          #   Column        Non-Null Count  Dtype
         ---  ------        --------------  -----
          0   yr            730 non-null    int64
          1   holiday       730 non-null    int64
          2   workingday    730 non-null    int64
          3   temp          730 non-null    float64
          4   atemp         730 non-null    float64
          5   hum           730 non-null    float64
          6   windspeed     730 non-null    float64
          7   cnt           730 non-null    int64
          8   season_2      730 non-null    bool
          9   season_3      730 non-null    bool
          10  season_4      730 non-null    bool
```

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

Answer:

Temp has highest correlation with the target variable.



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

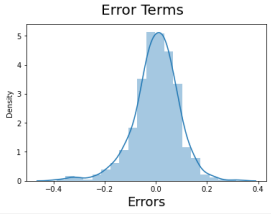
Answer:

There exists a Linear Relationship between X and y.

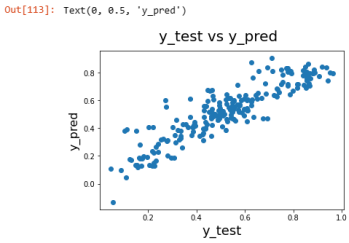
The features like temp, yr etc are having linear relationship with target variable cnt.

The error terms are following normal distribution.

```
In [381…
# Importing the required libraries for plots.
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

In [382…
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_cnt), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
plt.xlabel('Errors', fontsize = 18)                         # X-label

C:\Users\akhil\AppData\Local\Temp/ipykernel_20660/2756863869.py:3: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot((y_train - y_train_cnt), bins = 20)

Out[382…
Text(0.5, 0, 'Errors')
```



The error terms are independent of each other.



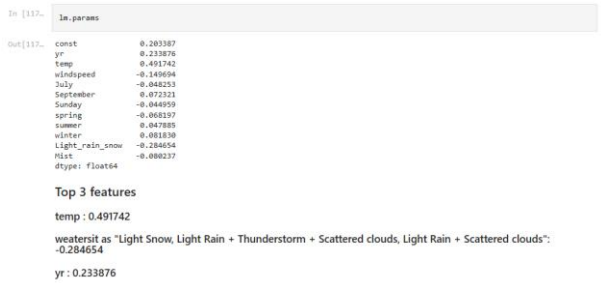
The error terms have constant variance, that is they exhibit homoscedasticity.

Question 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 contributing towards explaining the demand of shared bike are below:

1) temp : 0.49
2) weatersit as "Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds": -0.28
3) yr : 0.23



Question 1: Explain the linear regression algorithm in detail.

Regression is defined as a statistical method that attempts to determine a relationship between two or more correlated variables. It is used to predict value of a variable (also called dependent variable), given the values of other variable/s (also called predictor variable/s).

The regression algorithm falls under Supervised Learning method where historic data is labelled and used to determine the value of the output variable.

If two numerical variables are linearly correlated, we will have their correlation coefficient value that falls between -1 and 1. When we want to use this relationship to predict the value of one variable(dependent variable) based on the value of other variable(predictor variable), we use Linear Regression.

Simple Linear Regression: Only one predictor variable is used to predict the values of dependent variable.

Equation of the line : y = c + mx (only one predictor variable **x** with co-efficient **m)**

2. Multiple Linear Regression: More than one predictor variables are used to predict the values of dependent variable.

Equation of the line : y = c + m1x1+ m2x2 + m3x3 … + mixi (many predictor variables x1, x2 … xi). (**m1, m2 … mi** are respective co-efficients**)**

Best Fit Line:

The scatter plot is used to see how two numeric variables are related to each other and often if there is a linear relationship, we try to fit a line. But we cannot call any line as the Best Fit Line.

The data contains a set of values for dependent variable (denoted by y) and independent variable/s (denoted by X) and a best fit line is the one for which the Residual Sum of Squares of error terms is minimum.

The error term may be positive or negative. So we take square or all the error terms and sum it. This is called **Residual Sum of Squares** of error terms.

So the aim of linear regression is to find the best fit line for given X and y variables such that the Residual Sum of Squares of errors is minimum. In mathematical terms this RSS is our cost function which we need to minimize. There are several minimization techniques, but the most used is: **Gradient Descent method.**

Assumptions of Linear Regression:

The assumptions are below:

- There exists a Linear Relationship between X and y.
- The error terms are following normal distribution.
- The error terms are independent of each other.
- The error terms have constant variance, that is they exhibit homoscedasticity.

Example: Predicting house price in a city base on area, floor, city etc.

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

Answer: Anscombe's Quartet was developed by statistician Francis Anscombe. This is a method which keeps four datasets, each containing eleven (x, y) pairs. The important thing to note about these datasets is that they share the same descriptive statistics. Each graph tells a different story

irrespective of their similar summary statistics. Below is the glimpse of the statistics of the 4 datasets:

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

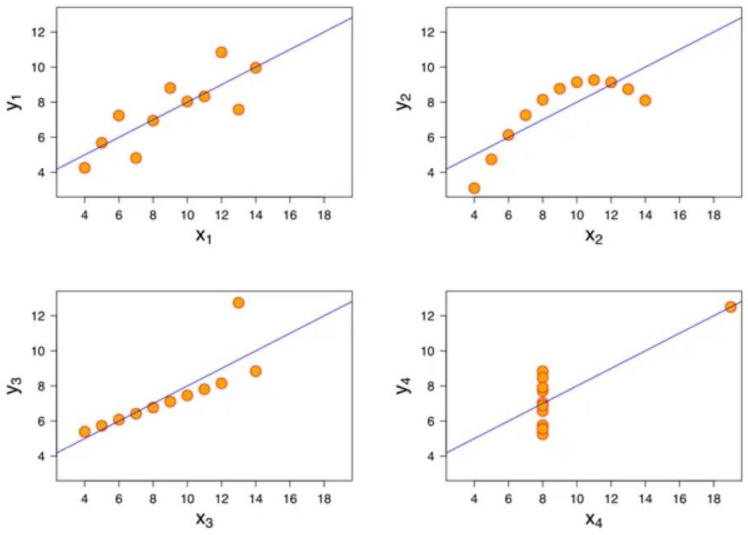
The summary statistics show that the means and the variances were identical for x and y across the groups:

· Mean of x is 9 and mean of y is 7.50 for each dataset.

· The variance of x is 11 and variance of y is 4.13 for each dataset

· The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



· Dataset I appears to have clean and well-fitting linear models.

· Dataset II is not distributed normally.

· In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

· Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

Also, Anscombe's Quartet warns of the dangers of outliers in data sets.

Question 3: What is Pearson's R?

Answer: Pearson's R was developed by **Karl Pearson** and it is a correlation coefficient which is a measure of the strength of a linear association between two variables and it is denoted by 'r'. it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

Mathematically, Pearson's correlation coefficient is denoted as the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Formula:

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

- N = number of pairs of scores
- Σxy = sum of the products of paired scores
- Σx = sum of x scores
- Σy = sum of y scores
- Σx^2 = sum of squared x scores
- Σy^2 = sum of squared y scores

Example:

· Statistically significant relationship between age and height.

· Relationship between temperature and ice cream sales.

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

Answer: **Scaling** is the process to normalize the data within a particular range. Many times, in our dataset we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single range.

The two most discussed scaling methods are Normalization and Standardization. Normalization typically scales the values into a range of [0,1]. Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Formula of Normalized scaling:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Formula of Standardized scaling:

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

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

Answer: The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1 - {R_i}^2}$$

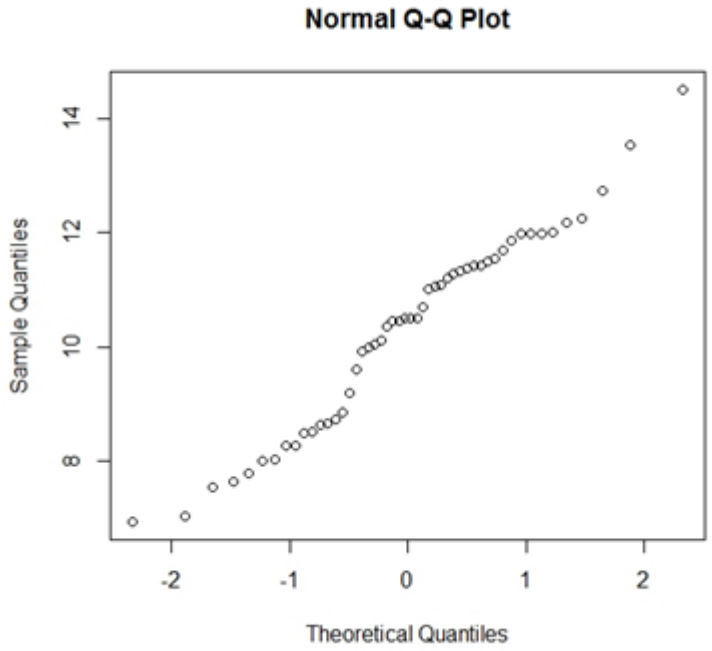
Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

Answer: The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression:

The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.