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? (3 marks)

Categorical variables (few examples year,atemp,Season) have huge effect on dependent variable cnt. Influence of these variables can be positive co-relation or negative co-relation as well. Lets consider an example of positive co-relation where from derived equation we can clearly understand increase in actual temperature also lead to increase in bike sharing count where we on a given day if windspeed is more its affecting bike sharing count negatively

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

When we have non-Boolean categorical variable having 'n' different values in the list, dummy variable creation returns n number of new columns. Since its possible to represent one type without additional column we can use drop_first=True

Spring	Summer	Winter
1	0	0
0	1	0
0	0	1
0	0	0

In above table we can still represent Season type "Fall" for last record without having additional column. Hence, we use drop_first=True to reduce the extra column. Along with that it helps to reduce correlation between dummy columns.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

atemp is having highest correlation with Target variable cnt

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

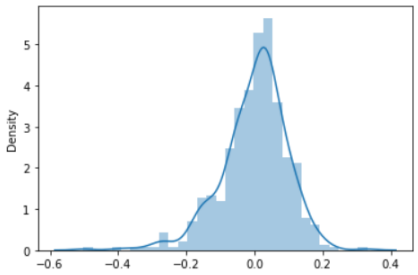
- Calculated VIF to ensure there is no multicollinearity
- Calculated Y_Pred and derived residual. Plotted distribution graph to check if errors are normally distributed

```
In [45]:  #Calculate residuals
          res = y_train - y_train_pred
```

```
In [46]:  sns.distplot(res)
```

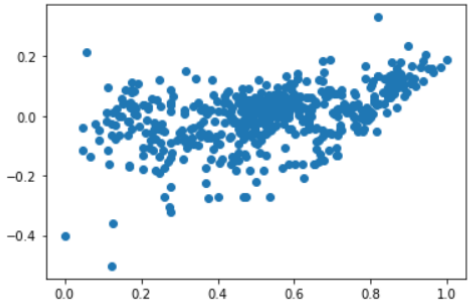
```
C:\Users\i320807\Anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated functi
on and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with sim
ilar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

```
Out[46]: <AxesSubplot:ylabel='Density'>
```



- Additional validation with scattered plot between y_train and residuals

```
#Scatter plot for y-train and residuals
plt.scatter(y_train,res)
plt.show()
```



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Top variables contributing count of shared bikes

- atemp - Positive co-relation
- yr - Positive co-relation
- Light snow - Negative co-relation

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

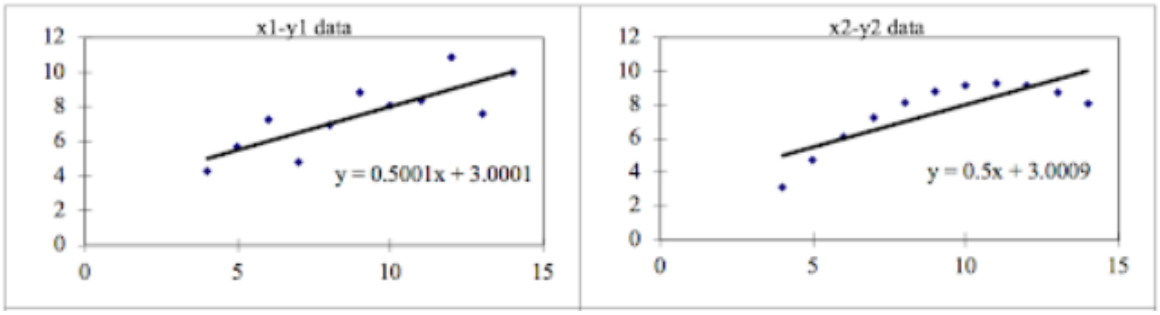
Linear Regression algorithm is a process of deriving relationship between variables. In this algorithm we consider one of the variables as target variable (Dependent) and will try to establish relationship between other variables (Independent). The linear regression algorithm explains how dependent variable values change with each unit change in the values of independent variables. In simple linear

regression there one independent variable determining value of target variables and In a simple linear regression, we can change one. But in case of multiple linear regression multiple independent variables determine the outcome of a single dependent variable.

Examples of linear regression include forecasting product sales in a business based on past data and student grades in a topic based on historical data.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet emphasizes the value of data visualization before using various algorithms to create models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.



In above example by plotting the data we can conclude linear regression is better suited for first one. Hence, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3. What is Pearson's R? (3 marks)

Pearson correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations which is a normalized measurement of the covariance, result always has a value between – 1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope
- r = -1 means the data is perfectly linear with a negative slope
- r = 0 means there is no linear association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

In any model we are building if we need to interpret coefficients, it is extremely important all variables are comparable. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale.

There are 2 ways of scaling:

- Normalized Scaling – Min-Max scaling values are shifted and rescaled so that values are between 0 and 1. Helps for data which have outliers.
- Standardized Scaling – Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

The value of VIF is infinite when there is a perfect correlation. If the correlation is perfect, we have R2 = 1, which results in 1/(1-R2) infinite. The variable that is producing this perfect multicollinearity must be removed from the dataset in order to remedy the issue.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.