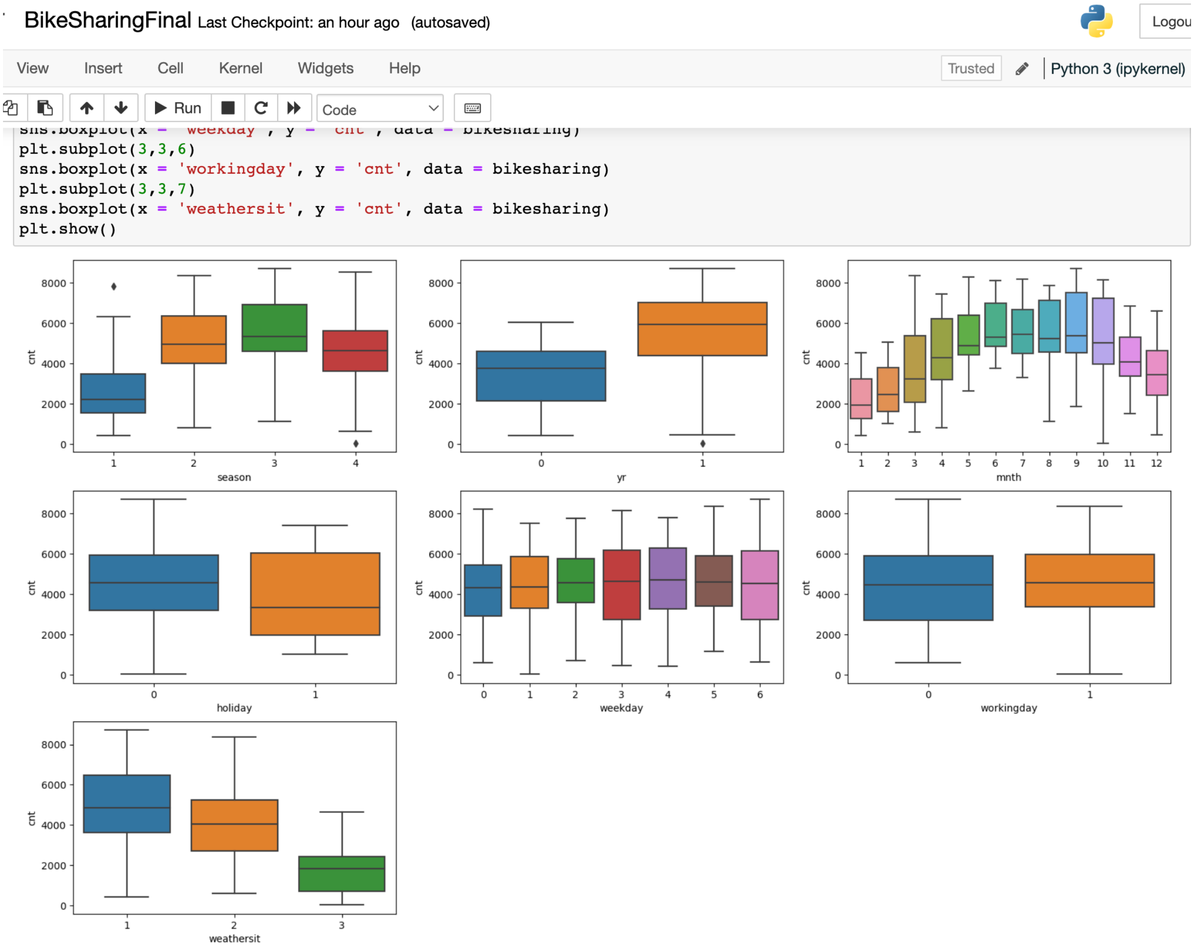
Assignment-based Subjective Questions

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

We had 7 categorical variables in the dataset. We used box plots to determine their effect on dependent variable ‘cnt’.



```
sns.boxplot(x = 'weekday', y = 'cnt', data = bikesharing)
plt.subplot(3,3,6)
sns.boxplot(x = 'workingday', y = 'cnt', data = bikesharing)
plt.subplot(3,3,7)
sns.boxplot(x = 'weathersit', y = 'cnt', data = bikesharing)
plt.show()
```

Inferences drawn:

- The demand for shared bikes seem to peak in the season of fall.
- The bike sharing systems are slowly gaining popularity as we can see cnt has increased from 2018 to 2019
- Demand also seems to spike at the end of Quarter 3 and beginning of Quarter 4 i.e., months of Aug, Sep and Oct
- The quartiles seem to have higher values on non-holidays compared to holiday
- Demand seems to peak on Wednesdays and Saturdays
- Demand seems to be higher on non-working days
- As weather changes from 1:Clear Skies to 2:Hazy to 3:Showers demand seems to decrease as expected. People don't go out when its cloudy or raining outside.

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

When we have a categorical variable with say 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels. For a variable say, 'Relationship' with three levels namely, 'Single', 'In a relationship', and 'Married', we would create a dummy table like the following:

Relationship Status	Single	In a relationship	Married
Single	1	0	0
In a relationship	0	1	0
Married	0	0	1

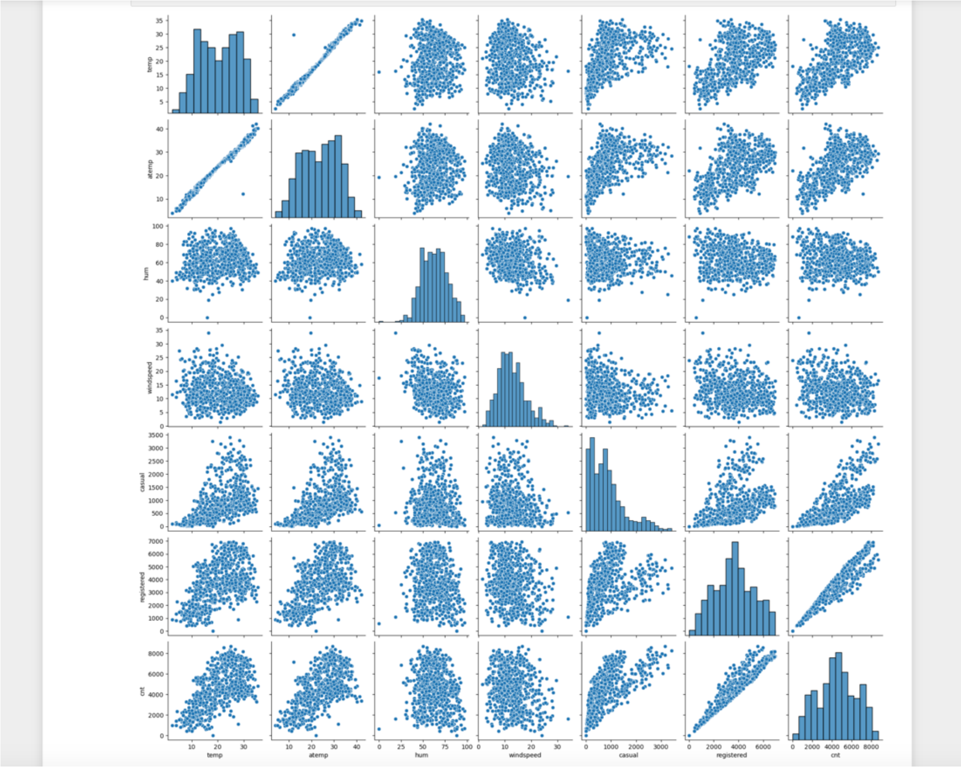
But we can clearly see that there is no need of defining **3** different levels. If we drop a level, say 'Single', we would still be able to explain the three levels, it's just redundant and causes multicollinearity.

Relationship Status	In a relationship	Married
Single	0	0
In a relationship	1	0
Married	0	1

Here Single becomes the base state and the interpretation becomes effect of state vs base state.

In python, we create dummy variables from categorical using one-hot encoding and drop_first=true says whether one of the dummy variables representing a category must be dropped. **It helps to avoid multicollinearity issues and ensures that remaining dummy variables are linearly independent.**

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



'Temp', 'atemp', 'casual' and 'registered' all of these have good positive correlation with target variable.

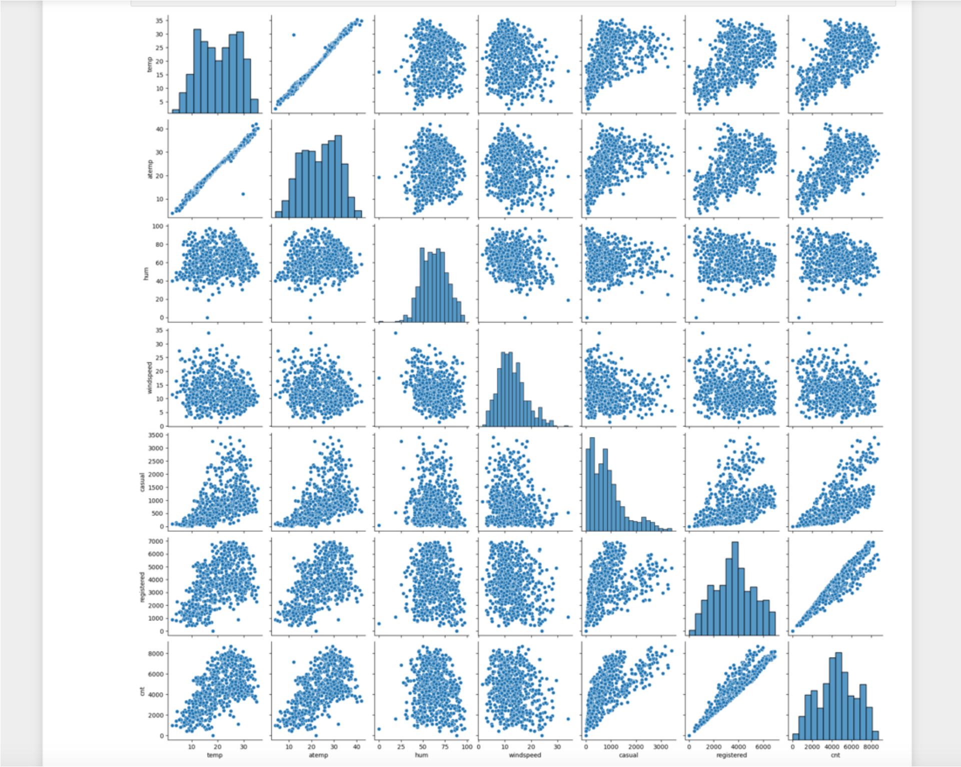
But 'casual' and 'registered' these two directly contribute to 'cnt' and hence they cause multicollinearity. So, excluding them ' **temp'** and **'atemp'** seems to have highest correlation with 'cnt' target variable.

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

In linear regression, we have 4 assumptions:

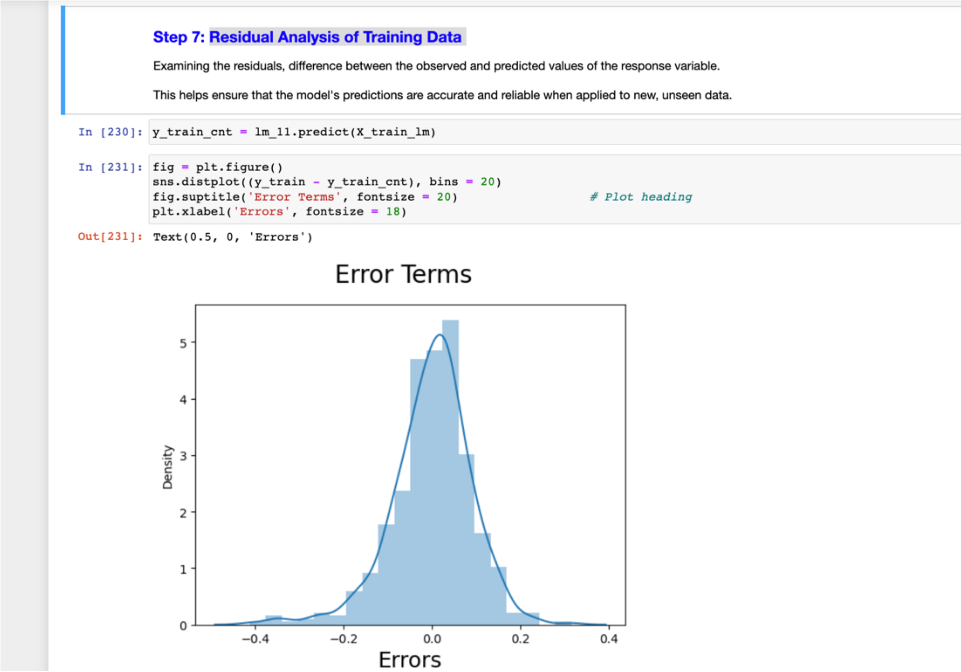
- **Linear relationship between X and Y**

We created a pair plot and checked if there are any linear relationships found between dependent and independent variables. We found temp and atemp, hence concluded opting for linear regression model is appropriate.



- **Error terms are normally distributed with mean 0.**

We performed residual analysis. We used the final model to predict values for target variable and compared it against actual target variable values to examine the differences. These differences are nothing but 'residuals' or 'residual error terms'. we plotted these residuals as a histogram and verified that error terms are indeed distributed normally.



Step 7: Residual Analysis of Training Data

Examining the residuals, difference between the observed and predicted values of the response variable.

This helps ensure that the model's predictions are accurate and reliable when applied to new, unseen data.

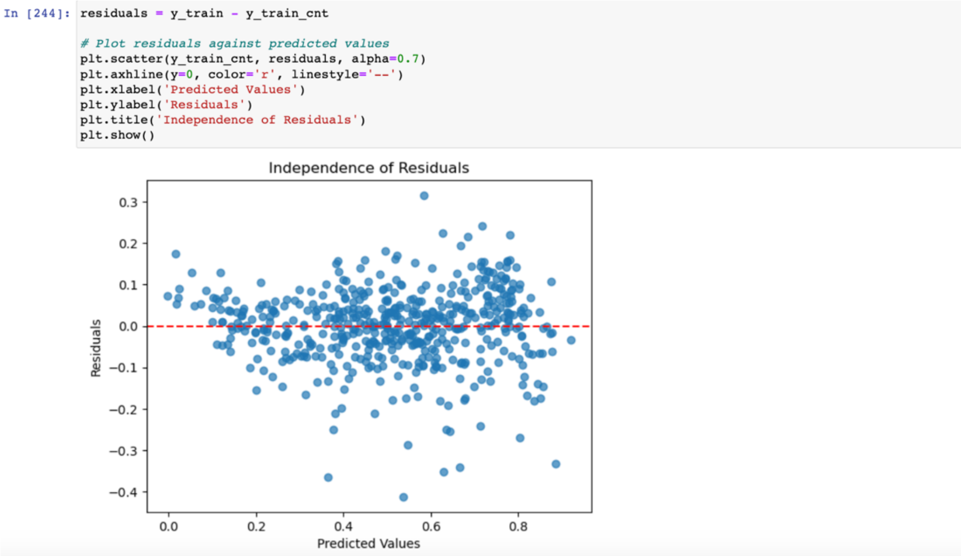
```
In [230]: y_train_cnt = lm_11.predict(X_train_lm)

In [231]: 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)

Out[231]: Text(0.5, 0, 'Errors')
```

- **Error terms are independent of each other.**

  We plotted a scatter plot with residuals against predicted values.
  Since error terms are independent, we see random scatter of points around the zero line.



```
In [244]: residuals = y_train - y_train_cnt

          # Plot residuals against predicted values
          plt.scatter(y_train_cnt, residuals, alpha=0.7)
          plt.axhline(y=0, color='r', linestyle='--')
          plt.xlabel('Predicted Values')
          plt.ylabel('Residuals')
          plt.title('Independence of Residuals')
          plt.show()
```

- **Error terms have constant variance or constant deviation (homoscedasticity).**

  In the scatter plot above that shows residual vs predicted value, the data points are spread across equally without a prominent pattern. It means the residuals have constant variance (homoscedasticity).
  No funnel-shaped pattern is seen, hence no depiction of non-constant variance (heteroscedasticity).

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

As per our final model, **top 3 features** that are significant in predicting demand for shared bikes are:

1. **Temperature (temp)** : One unit increase in temp variable increases the bike rental count by 0.5636 units. Higher temperatures are associated with higher bike rental counts.
2. **Weather (weathersit 3: weather_3_showers)** : Compared to reference category weathersit 1, one unit increase in weathersit 3 variable decreases bike rental count by 0.288 units. Presence of showers (weathersit 3) is associated with lower bike rental counts.
3. **Year (yr)** : One unit increase in year variable increases the bike rental count by 0.233 units. Bike rental counts have increased over the years.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is one of the fundamental algorithms in machine learning used to predict continuous target variables based one or more input features. It models the relationship between input features and target by fitting a linear equation.

1. In Simple Linear Regression, with only one input feature the model is represented as:
   $y = \beta0 + \beta1x$
   where:
   - **y** is the predicted target variable.
   - **x** is the input feature.
   - **β1** is the slope (weight) of the line.
   - **β0** is the intercept.

   In Multiple Linear Regression this becomes:
   $y = \beta0 + \beta1X1 + \beta2X2 + ... + \beta pXp + \epsilon$

   In SLR, we say if we increase X by 1 unit it increases Y by β1. Now in MLR, we say "Change in the mean response E(y), per unit increase in the variable when other predictors are held constant."
2. The goal is to find coefficients that minimize difference between actual and predicted values. One of the most common measures is Mean Squared Error (MSE).
3. We find such coefficients by using techniques such as Ordinary Least Squares (OLS) method. It aims to find the coefficients that minimize the sum of squared differences between the observed and predicted values.
4. Linear Regression assumes that:
   - There is Linear relationship between X and Y.
   - Error terms are normally distributed with mean 0.
   - Error terms are independent of each other.
   - Error terms have constant variance or constant deviation (homoscedasticity).
5. Below are the steps we follow when building a model:
   i) Split data into training and test data sets
   ii) Scale numerical variables. It helps improve the performance, convergence, and interpretability of linear regression models by ensuring that the features are all on a similar scale.
   iii) We divide into X(predictors) and Y(target) for model building
   iv) Either we use manual or automated approach like RFE for feature selection
   v) We build the Linear Regression model using techniques like OLS
   vi) We try to improve the model:
   We do this by looking at p-value that measures the significance of individual predictor variables and their relationship with the response variable.
      - **p-value <= 0.05** - implies that the predictor variable is likely to be significant and can be **included** in the model.

- **p-value > 0.05** - implies that the predictor variable is not statistically significant in explaining the variability in the response variable and can be **dropped**.

One more measure we look at is VIF or Variance Inflation that detects multicollinearity between 2 or more predictor variables. Because highly correlated predictors can lead to issues in interpreting the model's coefficients and making accurate predictions.

- A VIF value of **1** indicates no multicollinearity and can be **included** in the model.
- VIF values between **1 and 5** are generally considered acceptable and can be therefore **included** in the model.
- VIF values above **5 or 10** suggest high multicollinearity and hence needs to be **dropped** from model.

Therefore, we will build the model as below:
1. **Low p-value and Low VIF** : we include in the model
2. **High p-value and High VIF** : we should definitely drop this
3. **High P, Low VIF** : we will drop these first
   Low P, High VIF : we will remove these next

vii) Once we have the final model, we evaluate the model using techniques like Residual Analysis and validate if the linear regression assumptions still hold

viii) We then make predictions using the final model, evaluate model by plotting between actuals and predicted values. Evaluate its performance using measures like R-Squared score, Mean Squared Errors and Root Mean Squared Error.

ix) Finally we present the model with our interpretation.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets in statistics that have the same simple statistical properties, but when plotted, they reveal different and sometimes surprising patterns. It was first created by the statistician Francis Anscombe in 1973.

The four datasets in Anscombe's quartet have nearly identical mean, variance, correlation, and linear regression parameters. However, when we graph them, we will see that they have distinct shapes and patterns. Some show a straight line, some have a curve, and others have a pattern that doesn't follow the usual rules. This is surprising because we expect that similar numbers should give similar graphs.

The point of Anscombe's quartet is to show that just looking at the numbers isn't enough. We need to actually draw the graphs to really understand what's happening with the data. This teaches us an important lesson in statistics: while numbers can tell us part of the story, seeing the data visually can reveal unexpected things and help us make better conclusions.

3. What is Pearson’s R?

Pearson's correlation coefficient or Pearson’s R is a statistical measure that tells how strongly two variables are related to each other. It quantifies the linear relationship between two continuous variables.

Pearson's R value ranges between -1 and +1.

1. Value of +1 indicates a perfect positive linear relationship. As one variable increases, the other also increases proportionally.
2. Value of -1 indicates a perfect negative linear relationship. As one variable increases, the other decreases proportionally.
3. Value around 0 suggests a weak or no linear relationship between the variables.

The formula involves calculating the covariance i.e, how the variables change together and dividing it by the product of their standard deviations i.e, how spread out they are.
In simple words, it measures how the variables move together while considering their individual spreads.

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

Scaling refers to the process of transforming the values of variables to a specific range or distribution. It is performed to ensure that all variables are on a similar scale.

This helps improve the convergence of optimization algorithms and ensure fair treatment of variables during coefficient estimation. It can also make the coefficients in linear regression models more interpretable, as they represent the change in the target variable per unit change in the scaled feature.

There are two common scaling techniques: normalization and standardization.

Normalization / Min-Max scaling scales the variables to a specific range, typically [0, 1].

X_normalized = (X - X_min) / (X_max - X_min)

For eg: Let's say we have a feature 'Age' with values ranging from 20 to 60. After normalization, the values could be transformed to lie between 0 and 1.

Standardisation / z-score scaling scales the variables to have a mean of 0 and a standard deviation of 1.

X_standardized = (X - X_mean) / X_std

For eg: If we have a feature 'Income' with a mean of $50,000 and a standard deviation of $20,000, after standardization, the values could be transformed to have a mean of 0 and a standard deviation of 1.

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

This happens when there is perfect multicollinearity among the predictor variables in a multiple linear regression model

Perfect multicollinearity can arise when

1. We have duplicate or linearly dependent predictor variables, where one of them can be expressed as a linear combination of others
2. Data Entry errors are present leading to repeated or identical values in predictor variables.
3. Creating dummy variables. If one category's dummy variable is a perfect linear combination of other category dummy variables.

It's important to detect and address multicollinearity issues before building regression models. Common solutions include dropping one of the correlated variables, combining correlated variables, or using dimensionality reduction techniques.

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

Q-Q plots are also known as Quantile-Quantile plots. They basically plot quantiles of a sample distribution against quantiles of a theoretical distribution. This helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential etc.

In the context of linear regression, a Q-Q plot is often used to check the assumption of normality for the residuals of the regression model.

The first step is to obtain the residuals from the linear regression model. Residuals are the differences between the observed values and the predicted values
The Q-Q plot then compares the quantiles of the residuals against the quantiles of a theoretical normal distribution. If the residuals closely follow the normal distribution, the points on the Q-Q plot will fall roughly along a straight line

Use and Importance in Linear Regression:

1. If the points in the Q-Q plot deviate significantly from the straight line, it suggests that the residuals do not follow a normal distribution. This could indicate potential issues in the regression model.
2. Many statistical tests and confidence intervals in linear regression assume normality of residuals. By confirming this assumption through a Q-Q plot, we ensure the validity of the statistical inferences drawn from the model.
3. If the Q-Q plot reveals anomaly in the residuals, it prompts further investigation.