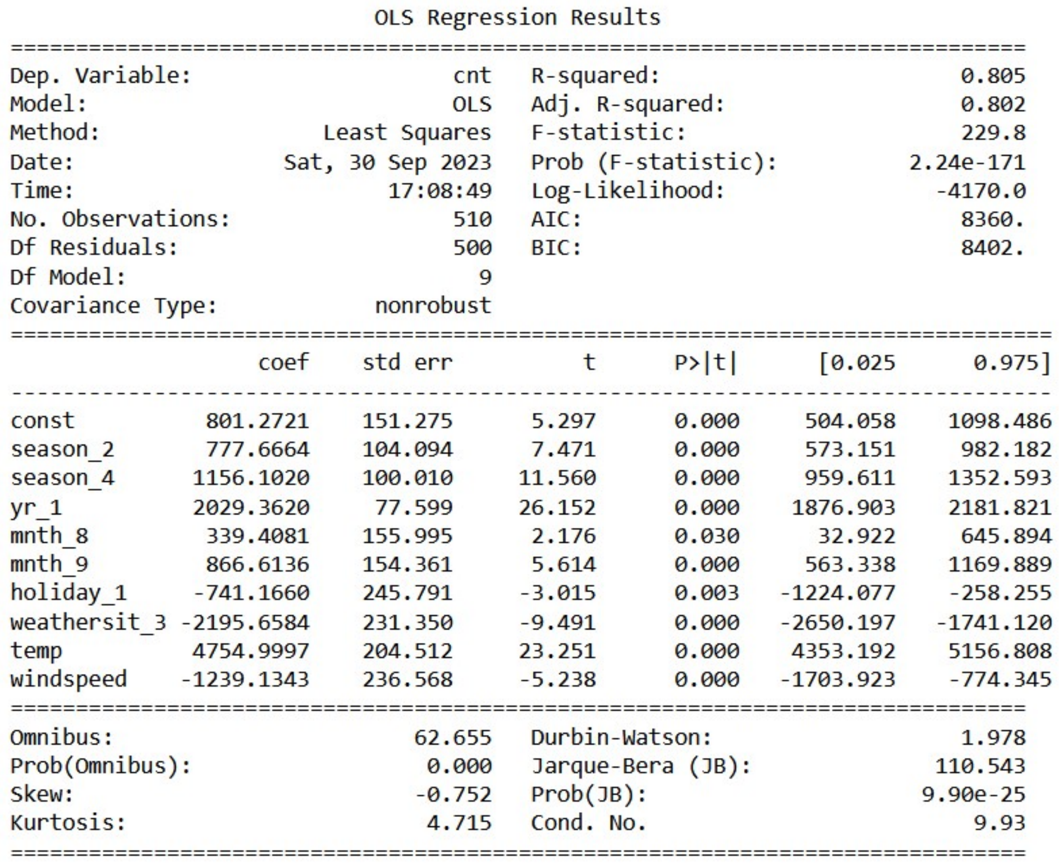
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: As our model has be MinMax scaled the effect of any categorical variables can be easily interpreted from whether it made it to the final regression model, and if it did, its coefficient can be a predictor of how important that variable is for making a prediction.

For the model prepared, the following table lists the variables and the coefficients that each variable has in the linear regression:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.805
Model:                            OLS   Adj. R-squared:                  0.802
Method:                 Least Squares   F-statistic:                     229.8
Date:                Sat, 30 Sep 2023   Prob (F-statistic):          2.24e-171
Time:                        17:08:49   Log-Likelihood:                -4170.0
No. Observations:                 510   AIC:                             8360.
Df Residuals:                     500   BIC:                             8402.
Df Model:                           9
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          801.2721    151.275      5.297      0.000     504.058    1098.486
season_2       777.6664    104.094      7.471      0.000     573.151     982.182
season_4      1156.1020    100.010     11.560      0.000     959.611    1352.593
yr_1          2029.3620     77.599     26.152      0.000    1876.903    2181.821
mnth_8         339.4081    155.995      2.176      0.030      32.922     645.894
mnth_9         866.6136    154.361      5.614      0.000     563.338    1169.889
holiday_1     -741.1660    245.791     -3.015      0.003   -1224.077    -258.255
weathersit_3 -2195.6584    231.350     -9.491      0.000   -2650.197   -1741.120
temp          4754.9997    204.512     23.251      0.000    4353.192    5156.808
windspeed    -1239.1343    236.568     -5.238      0.000   -1703.923    -774.345
==============================================================================
Omnibus:                       62.655   Durbin-Watson:                   1.978
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              110.543
Skew:                          -0.752   Prob(JB):                     9.90e-25
Kurtosis:                       4.715   Cond. No.                         9.93
==============================================================================
```

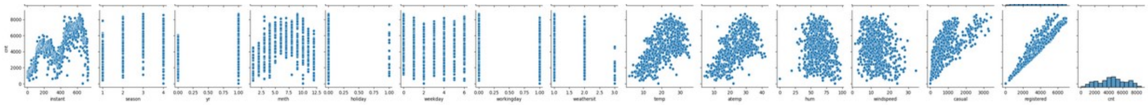
From the above table we can see that the categorical variables weathersit_3, i.e. whether the weather is “Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds” and yr_1, i.e if it is the year 2019 or not, significantly effects the count of bikes rented.

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

While preparing dummy varaibles in certain types of categorical variables such as the ‘season’ variable in our dataset, the sum of all possible values is constant, so when creating a dummy value for all possible values, our dataset creates unnessary redundacy. In such cases, if there are ‘N’ values taken by the categorical variable, all possible combinations of values can be expressed by ‘N-1’ values, where the excluded value can be expressed by the absence of 1 or True in all the N-1 values, hence the option to drop_first is used during dummy variable creation.

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

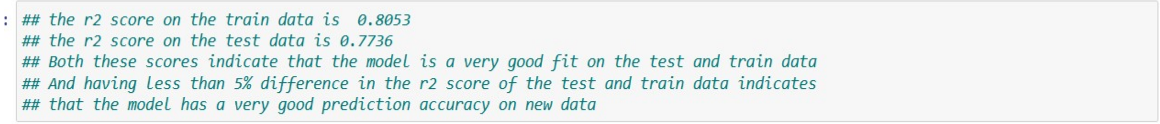
The pair plot is displayed in the output of the line 9 in the python notebook. Extracting only the relevant part of the pair plot, we have below:



From the above plot it is clear the the variables which shows the highest correlation with the target variable ‘cnt’, excluding the variables ‘casual’ and ‘registered’ which are just alternatives of the target variable, the variables ‘temp’ shows the most linear correlation.

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

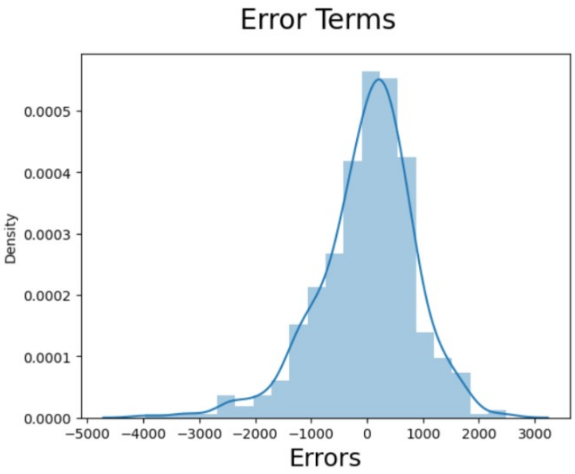
The condition of linearity is assumed initially by looing at the pairplots, and the correlation matrix, but it is later confirmed by checking the values of r2 scores of the training and test datasets.



```
: ## the r2 score on the train data is  0.8053
  ## the r2 score on the test data is 0.7736
  ## Both these scores indicate that the model is a very good fit on the test and train data
  ## And having less than 5% difference in the r2 score of the test and train data indicates
  ## that the model has a very good prediction accuracy on new data
```

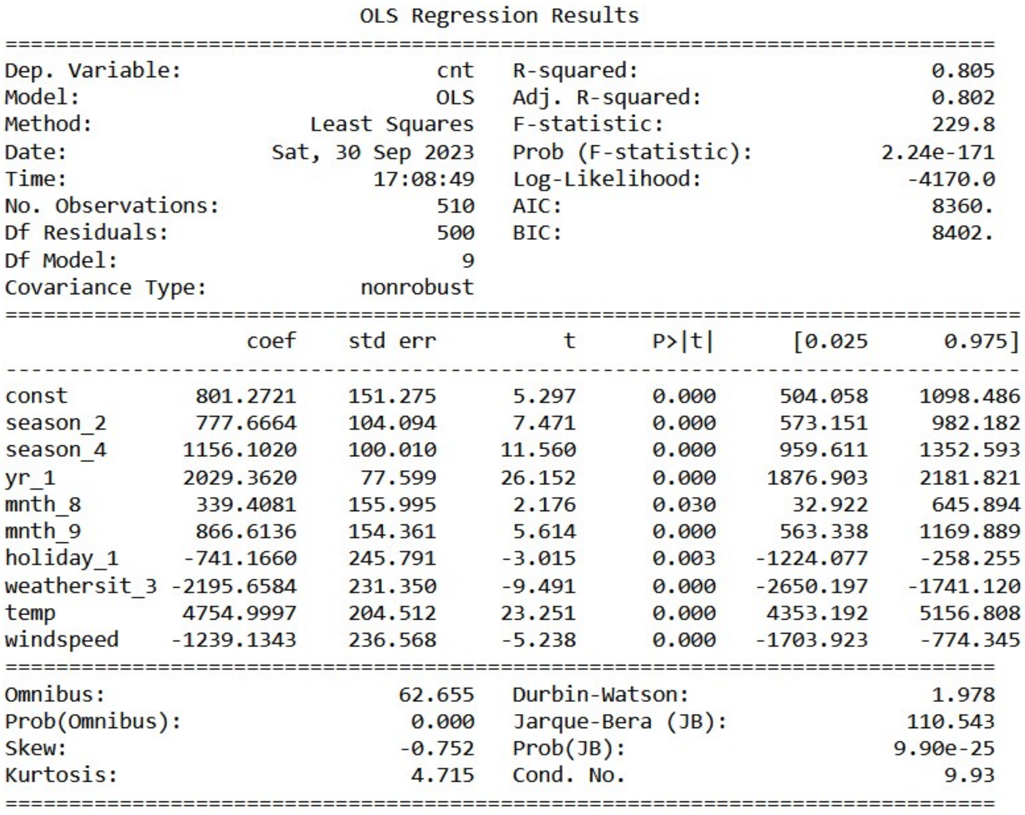
Perfect multicollinearity is eliminated by checking the VIF values of the variables taken in the model, and eliminating variables which show a high VIF values.

Residual analysis confirms that histogram plot of the error terms is indeed close to normal.



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

The coefficients of the final model, as well as the variables can be seen in the below table:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.805
Model:                            OLS   Adj. R-squared:                  0.802
Method:                 Least Squares   F-statistic:                     229.8
Date:                Sat, 30 Sep 2023   Prob (F-statistic):          2.24e-171
Time:                        17:08:49   Log-Likelihood:                -4170.0
No. Observations:                 510   AIC:                             8360.
Df Residuals:                     500   BIC:                             8402.
Df Model:                           9
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          801.2721    151.275      5.297      0.000     504.058    1098.486
season_2       777.6664    104.094      7.471      0.000     573.151     982.182
season_4      1156.1020    100.010     11.560      0.000     959.611    1352.593
yr_1          2029.3620     77.599     26.152      0.000    1876.903    2181.821
mnth_8         339.4081    155.995      2.176      0.030      32.922     645.894
mnth_9         866.6136    154.361      5.614      0.000     563.338    1169.889
holiday_1     -741.1660    245.791     -3.015      0.003   -1224.077    -258.255
weathersit_3 -2195.6584    231.350     -9.491      0.000   -2650.197   -1741.120
temp          4754.9997    204.512     23.251      0.000    4353.192    5156.808
windspeed    -1239.1343    236.568     -5.238      0.000   -1703.923    -774.345
==============================================================================
Omnibus:                       62.655   Durbin-Watson:                   1.978
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              110.543
Skew:                          -0.752   Prob(JB):                     9.90e-25
Kurtosis:                       4.715   Cond. No.                         9.93
==============================================================================
```

From this it can be seen that the variables most significantly effecting our prediction are the following: temp, weathersit_3, and yr_1 (This is obtained by checking the absolute value os their coefficients)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The fundamental idea is to find the best-fitting straight line through the data points that minimizes the sum of the squared differences between the observed and predicted values. This line is often referred to as the "regression line."

The key components and steps involved in linear regression:

Dependent Variable (Y):

The variable that you are trying to predict or explain. It is also known as the response or outcome variable.

Independent Variable(s) (X):

The variable(s) used to predict the dependent variable. These are also known as predictor variables or features.

Linear Equation:

The linear regression model assumes a linear relationship between the independent and dependent variables.

Steps in Linear Regression:

Data Collection:

Gather data on the dependent and independent variables.

Data Exploration:

Analyze and visualize the data to understand the relationships between variables and identify potential outliers.

Model Specification:

Decide on the form of the linear regression equation (simple or multiple linear regression).

Parameter Estimation:

Model Evaluation: Assess the goodness of fit using metrics like R-squared, which measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

Predictions: Once the model is trained, use it to make predictions on new or unseen data.

Assumptions of Linear Regression:

Linear regression makes several assumptions, including:

Linearity: The relationship between variables is linear.

Independence: Residuals (the differences between observed and predicted values) are independent.

Homoscedasticity: Residuals have constant variance.

Normality: Residuals are normally distributed.

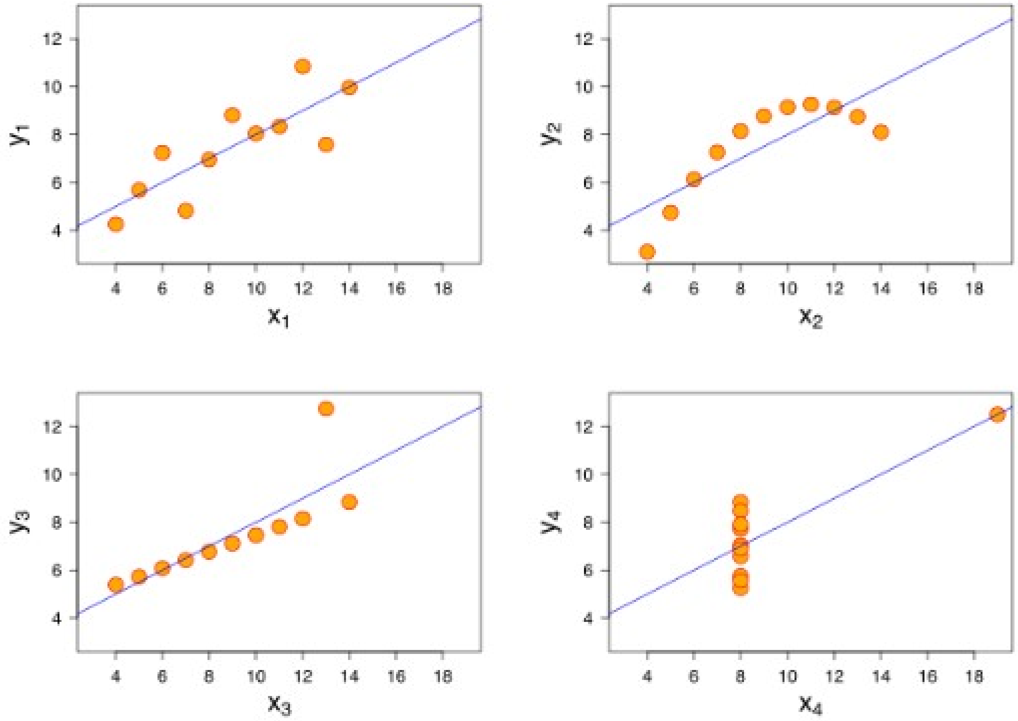
Types of Linear Regression:

Simple Linear Regression:Involves one independent variable.

Multiple Linear Regression:Involves more than one independent variable.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but vary significantly when graphically represented. It was created by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and the limitations of relying solely on summary statistics. The quartet is designed to highlight the impact of outliers and the influence of individual data points on statistical measures.



The four graphs shown above have the exact same summary statistics of mean and variance of both the variables, correlation, regression line, and R2 score.

3. What is Pearson’s R?

Pearson's correlation coefficient, often denoted as r or Pearson's r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is widely used in statistics to assess the degree to which changes in one variable are associated with changes in another variable. Pearson's r ranges from -1 to 1, where:

r=1: Perfect positive linear correlation

r=−1: Perfect negative linear correlation

r=0: No linear correlation

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where:

Xi and Yi are the individual data points and X-bar and Y-bar are the mean of X and Y despectively

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

Scaling is a preprocessing step in data analysis and machine learning where the numerical features of a dataset are transformed to a standardized range. The goal is to bring all features to a similar scale to prevent certain features from dominating the learning process simply because they have larger magnitudes. Scaling is particularly important for algorithms that are distance-based or rely on gradient descent, where feature magnitudes can influence the convergence of the optimization process.

<u>Why Scaling is Performed:</u>

Gradient Descent Convergence:Scaling helps gradient descent algorithms converge more quickly. Features with larger scales might dominate the learning process, causing slower convergence.

Distance-Based Algorithms:Algorithms that use distances between data points, such as k-nearest neighbors or support vector machines, can be sensitive to the scale of features.

Regularization:Regularization methods, like L1 or L2 regularization, penalize large coefficients. Scaling can ensure that features are penalized fairly.

PCA (Principal Component Analysis):Scaling is important when applying PCA, as it is a variance-based method and requires features to have similar scales.

Neural Networks:Deep learning models, especially those using gradient-based optimization, benefit from scaled features for improved convergence.

<u>Types of Scaling:</u>

Min-Max Scaling (Normalization):Involves scaling the features to a specific range, usually between 0 and 1.

Formula: $X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$

This method is sensitive to outliers.

Standardization (Z-score Scaling):Involves scaling features to have a mean of 0 and a standard deviation of 1.

Formula: $X_{\text{standardized}} = \frac{X - \bar{X}}{\sigma}$

Less sensitive to outliers compared to min-max scaling.

Retains the shape of the original distribution.

Differences Between Normalized Scaling and Standardized Scaling:

Range:

Normalized Scaling (Min-Max Scaling): Scales features to a specific range, often between 0 and 1.

Standardized Scaling (Z-score Scaling): Centers features around a mean of 0 with a standard deviation of 1.

Sensitivity to Outliers:

Normalized Scaling: Can be sensitive to outliers, as the range is influenced by the minimum and maximum values.

Standardized Scaling: Is less sensitive to outliers, as it uses the mean and standard deviation.

Formula:

Normalized Scaling: Involves subtracting the minimum value and dividing by the range.

Standardized Scaling: Involves subtracting the mean and dividing by the standard deviation.

Interpretation:

Normalized Scaling: Scales data to a specific, known range, making it interpretable in that context.

Standardized Scaling: Results in a distribution with a mean of 0 and a standard deviation of 1, preserving the shape of the original distribution.

The choice between normalized scaling and standardized scaling depends on the specific requirements of the algorithm and the characteristics of the data. In practice, standardized scaling is often preferred due to its robustness and general applicability.

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

Reasons for Infinite VIF:

Perfect Collinearity:Infinite VIF occurs when there is perfect collinearity, meaning that one predictor variable can be exactly predicted from the others.

Perfect Multicollinearity:The VIF formula involves calculating the inverse of $1 - R_i^2$, and if R_i^2 is exactly 1, the inverse becomes undefined, resulting in an infinite VIF.

Singular Matrix:When the matrix of the predictor variables is singular (i.e., not of full rank), VIF calculations may result in infinite values.

Implications:

Model Instability:Infinite VIF indicates that one or more predictor variables are perfectly predictable from others, leading to instability in the estimation of coefficients.

Unreliable Coefficients:With infinite VIF, the regression coefficients become highly sensitive to small changes in the data, and their estimates become unreliable.

Dealing with Infinite VIF:

Identify and Address Collinearity:Investigate the relationships between predictor variables to identify and address collinearity issues.

Remove Highly Correlated Variables:If two or more variables are highly correlated, consider removing one of them from the model.

Combine Variables:If it makes theoretical sense, consider combining highly correlated variables into a composite variable.

Increase Data Size:In some cases, increasing the sample size may help alleviate multicollinearity issues.

Regularization Techniques:Techniques like Ridge regression can be used to mitigate multicollinearity and stabilize coefficient estimates.

In practice, detecting and addressing multicollinearity, especially when it leads to infinite VIF, is crucial for obtaining reliable and interpretable results from regression analysis.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a given sample or distribution follows a theoretical distribution. In linear regression, Q-Q plots are often used to check the normality assumption of the residuals, which is crucial for making valid inferences and predictions.

Construction of a Q-Q Plot:

Ordered Data:The observed data is first sorted in ascending order.

Theoretical Quantiles:Theoretical quantiles are calculated based on the chosen distribution (e.g., normal distribution).

Plotting: The observed quantiles are plotted against the theoretical quantiles.

Interpretation:

If the Points Follow a Straight Line:If the points on the Q-Q plot roughly follow a straight line, it suggests that the data follows the assumed distribution.

Departure from a Straight Line:Departures from a straight line indicate deviations from the assumed distribution. For normality assessment, deviations may suggest non-normality of the residuals.

Use and Importance in Linear Regression:

Checking Normality of Residuals:Q-Q plots are particularly useful in linear regression to assess whether the residuals (the differences between observed and predicted values) follow a normal distribution. This is important because many statistical tests and confidence intervals in linear regression assume normality of residuals.

Validating Assumptions:Linear regression models assume that the residuals are normally distributed. By examining the Q-Q plot, you can check whether this assumption holds. Deviations from a straight line may indicate non-normality, which can impact the reliability of statistical inferences.

Identifying Outliers:Q-Q plots can also help identify outliers in the data. Outliers can cause deviations from a straight line on the Q-Q plot.

Model Assessment:The Q-Q plot is a graphical way to assess the goodness of fit of the model. If the residuals are close to normally distributed, it suggests that the linear regression model is appropriate for the data.