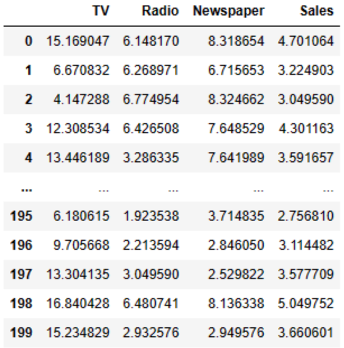
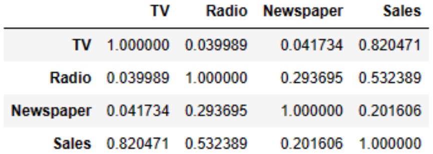
The advertising sales dataset was obtained from Kaggle on 5/5/2023. The dataset captures sales revenue generated with respect to advertisement costs across multiple channels like radio, TV, and newspapers. This dataset was chosen because I wanted to review the fundamentals of linear regression analysis, and this dataset appeared to be a good candidate for that analysis. For the purposes of this analysis, the sales variable is the dependent variable; radio, TV, and newspaper are the independent variables. Refer below for a snapshot of this data.



	TV	Radio	Newspaper	Sales
0	15.169047	6.148170	8.318654	4.701064
1	6.670832	6.268971	6.715653	3.224903
2	4.147288	6.774954	8.324662	3.049590
3	12.308534	6.426508	7.648529	4.301163
4	13.446189	3.286335	7.641989	3.591657
...	...	...	...	...
195	6.180615	1.923538	3.714835	2.756810
196	9.705668	2.213594	2.846050	3.114482
197	13.304135	3.049590	2.529822	3.577709
198	16.840428	6.480741	8.136338	5.049752
199	15.234829	2.932576	2.949576	3.660601

I began my analysis by checking for missing values within the dataset; additionally, I checked for correlation between variables to better understand the relationship between them. There were no missing values identified. Also, no independent variables are correlated with each other. Note: TV is highly correlated with sales – that is, sales will increase as expenditure on TV ads increases. The strength of this relationship is considered strong because the correlation coefficient is close to 1.0. This correlation does not imply causation. There could be other factors influencing the relationship between the two variables; however, this relationship is worth noting. Refer below for a screenshot for the correlation matrix.



	TV	Radio	Newspaper	Sales
TV	1.000000	0.039989	0.041734	0.820471
Radio	0.039989	1.000000	0.293695	0.532389
Newspaper	0.041734	0.293695	1.000000	0.201606
Sales	0.820471	0.532389	0.201606	1.000000

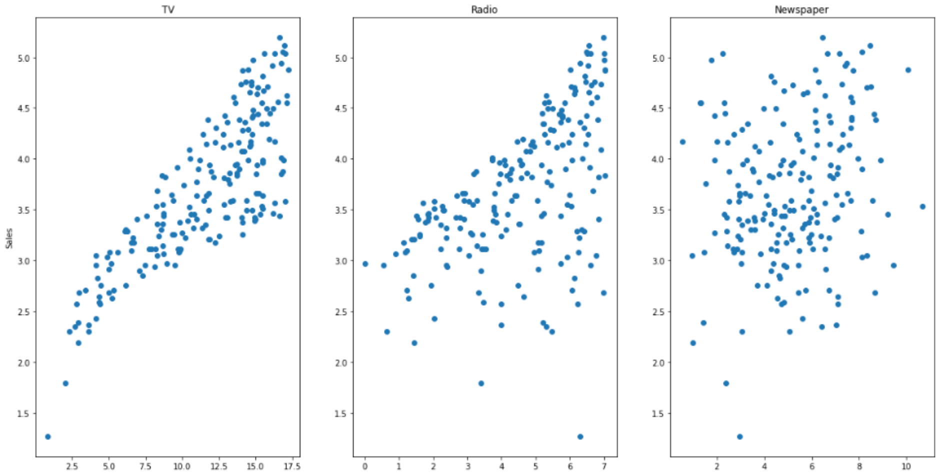
As stated previously, linear regression will be applied to this dataset. Linear regression is a supervised, parametric model. Supervised models are models where the dependent and independent variables are known. The dependent variable is the variable we want to predict; the independent variables are used to predict the value of the dependent variable. Multivariate linear regression analyzes the relationship between a single dependent variable and multiple independent variables. The multivariate linear regression model will be used for this report; however, a good understanding of linear regression is needed before continuing.

The multivariate linear regression model is denoted by the following formula: Y = B0 + B1x1 + B2x2 + … + Bp*xp. B0 is the intercept, which is the value of the dependent variable when X is zero. B1 is the slope, which gives the change of dependent variable Y for a one-unit change of independent variable X. To solve for data points that do not fit on a straight line, use the following formula: Y = B0 + B1x1 + B2x2 + … + Bp*xp + E. In this modified formula, E represents the error term. The error term identifies the variation of Y that cannot be explained by the linear regression model. The error term should be identically and independently distributed with a normal distribution of the zero mean. More specifically, the error term identifies the assumptions that the model must meet to be implemented. Those assumptions are as follows:

1. Zero Mean: For any observation of i, the corresponding error terms will have an expected value of zero.
2. Constant Variance: Variance of residuals is constant for all data points of the independent variable.
3. Mutually Independent: One instance of variance, or data point, should not have an impact on another data point.
4. Normally Distributed: Data should be normally distributed. Meaning, as the sample size increases, the data should represent a bell curve when displayed on a histogram chart.

The first three assumptions are called the Gauss-Markov assumptions. Those assumptions guarantee the least square estimator to be the best linear unbiased estimator. This guarantees the most efficient model with minimum variance (best), and an estimation of coefficients that are accurate (unbiased).

To assess the error term assumptions, I began by creating a scatterplot of independent variables. There are three independent variables, so three scatterplots were created. Those scatterplots can be found below. The independent variables can be found on the X axis and the dependent variable can be found on the Y axis. The TV feature shows a positive, linear relationship; the Radio feature shows a weak, positive relationship; and finally, the Newspaper feature does not show any relationship. To improve the linear relationship between independent and dependent variables, I calculated the square root of each feature. Note: Polynomial regression, nonlinear transformation of the response variable, nonlinear transformation of predictors, and Box-Cox transformation did not improve the linear relationship between independent and dependent variables.

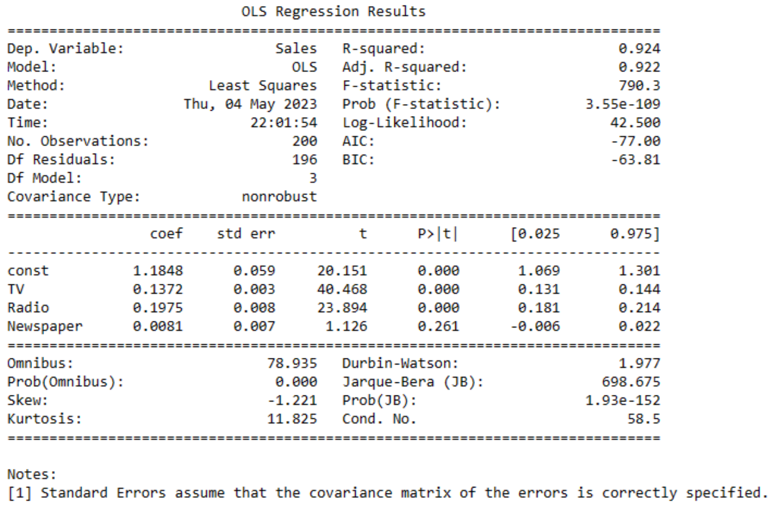


After analyzing the relationships of independent and dependent variables, I built a linear regression model using the statsmodel.api library. Refer below for a screenshot of the model. The R-square is 0.924; however, since my analysis uses a multivariate linear regression model, adjusted R-square will provide a more accurate estimate for goodness-of-fit. The adjusted R-square value is 0.922. This means that 92.2% of the variability in Sales can be explained by the model.

The model results provided by statsmodel.api list each independent variable and their respective p-value results. For a linear regression model, the p-value will help determine whether a variable is statistically significant for the model. Refer below for the null and alternative hypotheses for a linear regression model.

- Ho: Bp*xp = 0
- Ha: Bp*xp != 0

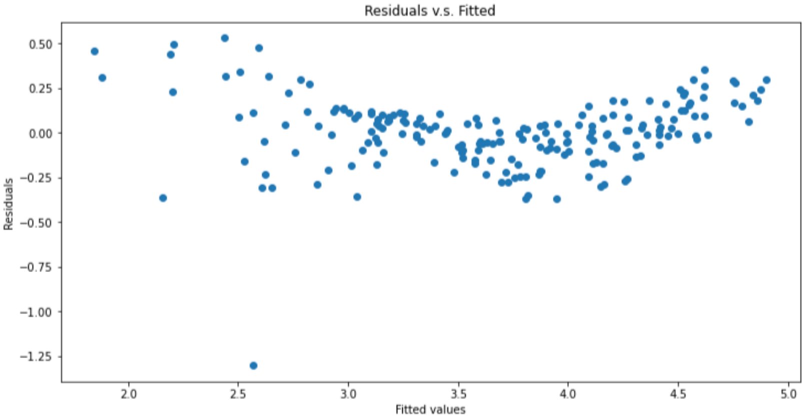
The significance level (alpha level) used for hypothesis testing is set to 0.05. If the p-value is less than 0.05, then we reject the null hypothesis and conclude that the independent variable is significantly affecting Sales; if the p-value is greater than 0.05, then we fail to reject the null hypothesis and conclude that the independent variable is not significantly affecting Sales. TV and Radio have p-values of 0.00, indicating that there is sufficient evidence of a statistically significant relationship between the dependent and independent variables. Newspaper has a p-value of 0.261, indicating that there is insufficient evidence of a statistically significant relationship between the dependent and independent variables.



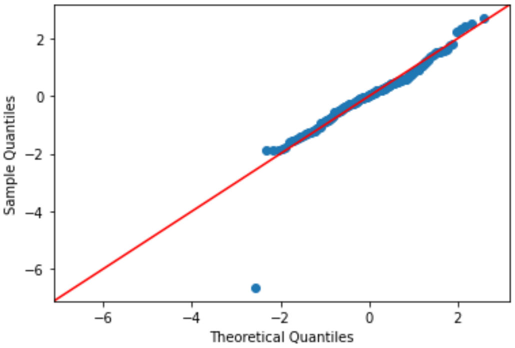
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.924
Model:                            OLS   Adj. R-squared:                  0.922
Method:                 Least Squares   F-statistic:                     790.3
Date:                Thu, 04 May 2023   Prob (F-statistic):          3.55e-109
Time:                        22:01:54   Log-Likelihood:                 42.500
No. Observations:                 200   AIC:                            -77.00
Df Residuals:                     196   BIC:                            -63.81
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.1848      0.059     20.151      0.000       1.069       1.301
TV             0.1372      0.003     40.468      0.000       0.131       0.144
Radio          0.1975      0.008     23.894      0.000       0.181       0.214
Newspaper      0.0081      0.007      1.126      0.261      -0.006       0.022
==============================================================================
Omnibus:                       78.935   Durbin-Watson:                   1.977
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              698.675
Skew:                          -1.221   Prob(JB):                    1.93e-152
Kurtosis:                      11.825   Cond. No.                         58.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

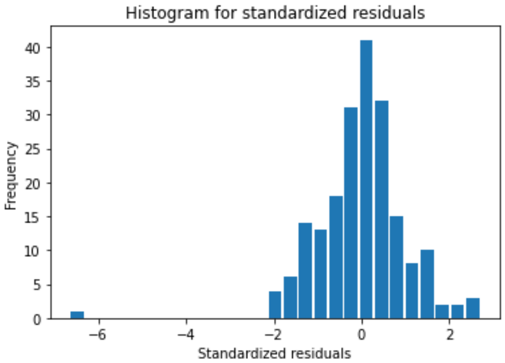
After obtaining the model results, I began assessing the error term assumptions. In the above screenshot, the Durbin-Watson test has a value of 1.977. Durbin-Watson values close to 2.00 indicate that data points in the model are mutually independent; therefore, the data points within the advertising sales dataset are mutually independent of each other.

To assess the constant variance and zero mean assumptions, I built a scatterplot comparing residuals to fitted response values. Refer to the screenshot below to view the scatterplot. Variance does not increase or decrease as the fitted response value increases; as a result, the constant variance assumption is not violated. There is a downward curve to the scatterplot, indicating a potential violation of the zero mean assumption. As stated previously, I took the square root of the original dataset values to improve the linear relationship between independent and dependent variables; however, it appears the transformation of data is not enough for the zero mean assumption to be satisfied. Before implementing a final model, other models need to be explored. Nonparametric models, like random forest, may perform better for this dataset. The initial goal of my analysis was to investigate the advertising sales dataset using a multivariate linear regression model; therefore, I continued my analysis while acknowledging that multivariate linear regression may not be the optimal model for this dataset.



Next, I investigated the normality assumption of the error term. To do this, I used a quantile-quantile plot (Q-Q plot) and a histogram plot. Those plots can be found below, respectively.

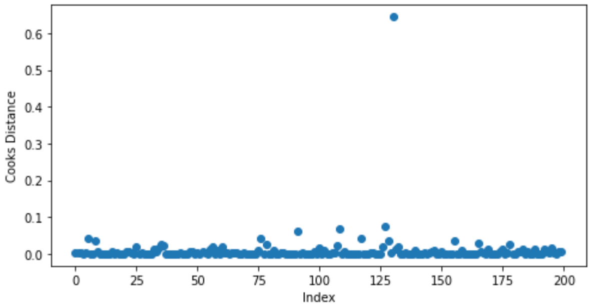




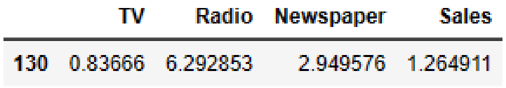
The Q-Q plot places all residuals in ascending order. The plot identifies the corresponding percentiles for each value. The vertical axis measures the observed residual values; the horizontal axis measures the theoretical residual values. If the observed values equal the theoretical values, then the normality assumption is satisfied. Alternatively, if either tail of the data points is off the diagonal line, then the normality assumption cannot be satisfied. Most data points align with the diagonal line. Notably, the lower end of the data points is slightly off the diagonal line; additionally, there appears to be an outlier for the observed residual values. The normality assumption is mostly satisfied given that most data points align along the diagonal line.

The histogram plot is used to test the normality assumption of the linear regression model. Frequency of the standardized residuals can be found on the vertical axis; standardized residuals can be found on the horizontal axis. For the normality assumption to be met, the data points should be roughly symmetric. The data in the histogram plot is roughly symmetric with a notable outlier. This indicates that the normality assumption is mostly satisfied.

The Q-Q plot and the histogram plot both identified outliers. To assess outliers, I began by using Cook's distance. Cook's distance is a method of measuring the influence of individual observations on the coefficients of a linear regression model. Cook's distance incorporates both residual and leverage for each observation. Large values of Cook's distance (Di) indicate a strong influence on the fitted model. Generally, Di > 1 is of great concern, and Di > 0.5 should have the outlier evaluated. Refer below for a scatterplot showing Cook's distance applied to the dataset. As seen below, there is a single observation with a Cook's distance of approximately 0.65.

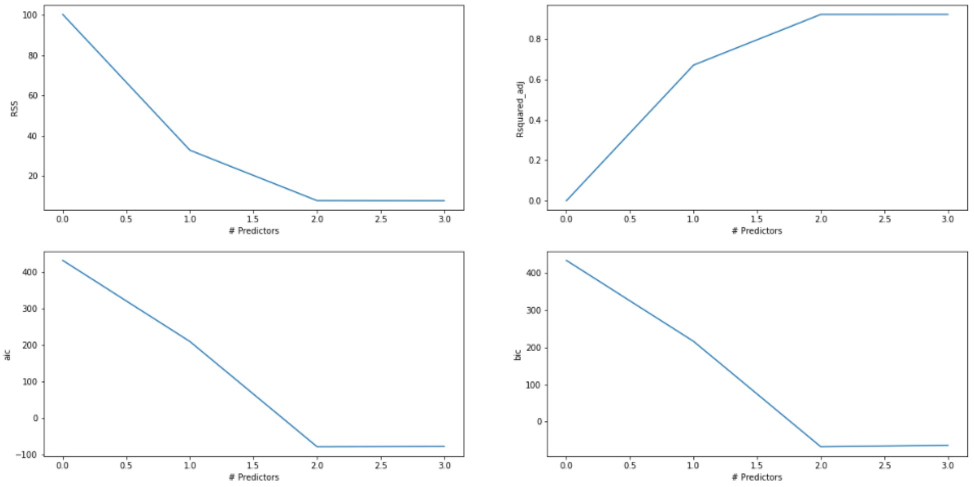


To identify the outlier, I filtered the advertising sales dataset for data points with a z-score greater than 3. Filtering the dataset based on z-score returned the results shown below. Considering that the dataset only has 200 rows of data, I decided to include the outlier in my analysis. This was done so that the statistical power of the dataset would not be diminished.

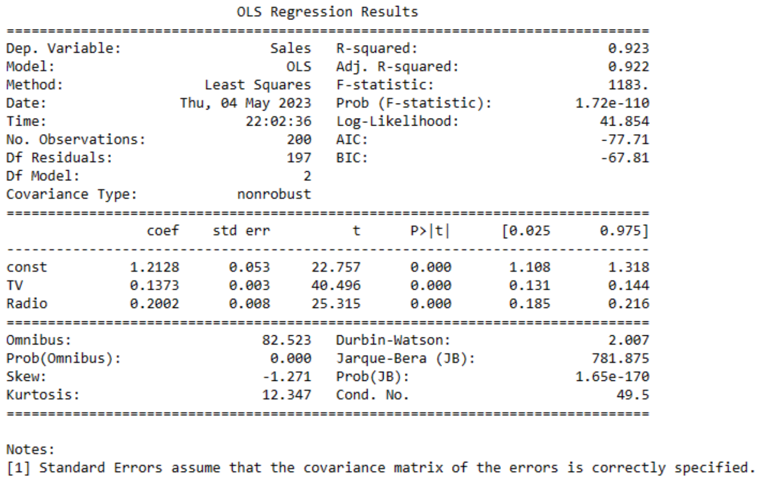


	TV	Radio	Newspaper	Sales
130	0.83666	6.292853	2.949576	1.264911

After assessing outliers, I began feature selection. Given that there are only three potential features to choose from, I decided to implement the best subset selection algorithm. The best subset selection algorithm is an exhaustive search algorithm. The algorithm evaluates all possible subsets of features and returns evaluation metrics for each set of features. The set of features with the best performance metrics are chosen for the final model. This method of feature selection guarantees a global optimal output; however, it is computationally expensive. This method of feature selection is only feasible for datasets that are small, with a limited number of columns. The performance metrics I chose for this algorithm are Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), residual sum of squares (RSS), and adjusted R-square. AIC and BIC are similar metrics – that is, they both use the estimated likelihood function to evaluate the goodness-of-fit for a statistical model. RSS measures the sum of squared differences between the predicted values and the actual values of the response variable. Adjusted R-square, as previously mentioned, measures the amount of variability in the response variable explained by the predictors. For AIC, BIC, and RSS, the smaller the number, the better the model. For adjusted R-square, the larger the number, the better the model. After using best subset selection, I plotted each set of predictors with their respective performance metrics. Refer below to view a chart showing the results of each performance metric.



As shown above, the optimal number of predictors for the model is 2.0. Running the model with the optimal number of predictors gives the results shown below. Adjusted R-square for the optimal model is 0.922. Meaning, 92.2% of the variability in Sales can be explained by the variables TV and Radio. The fitted model for the results shown below is the following: Sales = TV(0.1373) + Radio(0.2002). Note: since I transformed the variables to be the square root of the original dataset, Sales must be squared for the fitted model. This gives the ability to obtain a prediction of Sales in millions of dollars, which is the original unit of measurement given in the dataset.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.923
Model:                            OLS   Adj. R-squared:                  0.922
Method:                 Least Squares   F-statistic:                     1183.
Date:                Thu, 04 May 2023   Prob (F-statistic):          1.72e-110
Time:                        22:02:36   Log-Likelihood:                 41.854
No. Observations:                 200   AIC:                            -77.71
Df Residuals:                     197   BIC:                            -67.81
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.2128      0.053     22.757      0.000       1.108       1.318
TV             0.1373      0.003     40.496      0.000       0.131       0.144
Radio          0.2002      0.008     25.315      0.000       0.185       0.216
==============================================================================
Omnibus:                       82.523   Durbin-Watson:                   2.007
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              781.875
Skew:                          -1.271   Prob(JB):                    1.65e-170
Kurtosis:                      12.347   Cond. No.                         49.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

To conclude my analysis, I performed k-fold cross validation. K-fold cross validation is when the dataset is split into K number of folds and is used to evaluate the model's performance. For each fold, a subset of data is used as the test set and the other parts are used as the training sets. A visual representation of this process can be found below.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Test Data	Training Data	Training Data	Training Data	Training Data
Split 2	Training Data	Test Data	Training Data	Training Data	Training Data
Split 3	Training Data	Training Data	Test Data	Training Data	Training Data
Split 4	Training Data	Training Data	Training Data	Test Data	Training Data
Split 5	Training Data	Training Data	Training Data	Training Data	Test Data

Using k-fold cross validation, I obtained an average mean square error (MSE) of 1.6296. Note: I had to square the predicted values for Sales and the actual values for Sales to obtain an MSE in the original measuring unit (millions of dollars). An MSE of 1.6296 means that, on average, the squared difference between the predicted and actual values of the dependent variable is 1.6296 million dollars. The square root of the MSE can be obtained to provide the average error for the model, which is 1.276 million dollars. An average error of 1.276 million dollars is likely too high for this model to be useful; however, domain knowledge of the advertising business is needed to make that determination. As stated previously in this report, other models may perform better for the advertising dataset. More work will need to be performed before deciding on the best model.