ADVERTISING SALES PREDICTION ANALYSIS

ASSIGNMENT INSIGHTS & MODEL EVALUATION

DATASET OVERVIEW

- Dataset contains ad spending on TV, Radio, Newspaper
- Target variable: Product Sales
- Objective: Predict Sales from advertising expenditure

CONCEPT OF CORRELATION AND STEPS TO FIND CORRELATION MATRICES

- First, I converted the 'Status Published' column from text format to datetime format and created separate columns for date, hour, and year using methods from the datetime library
- Next, I removed all unnecessary and null columns to make the dataset more concise and suitable for modeling and analysis. I then checked whether any null values were still present in the remaining columns.Then correlation matrix between different columns is shown in the following slides.

Following this, a correlation matrix between different columns is presented in the subsequent slides.
The concept behind the correlation values is as follows: a correlation value between 0 and 1 indicates a positive correlation, a value between -1 and 0 indicates a negative correlation, and a value of 0 indicates no correlation.

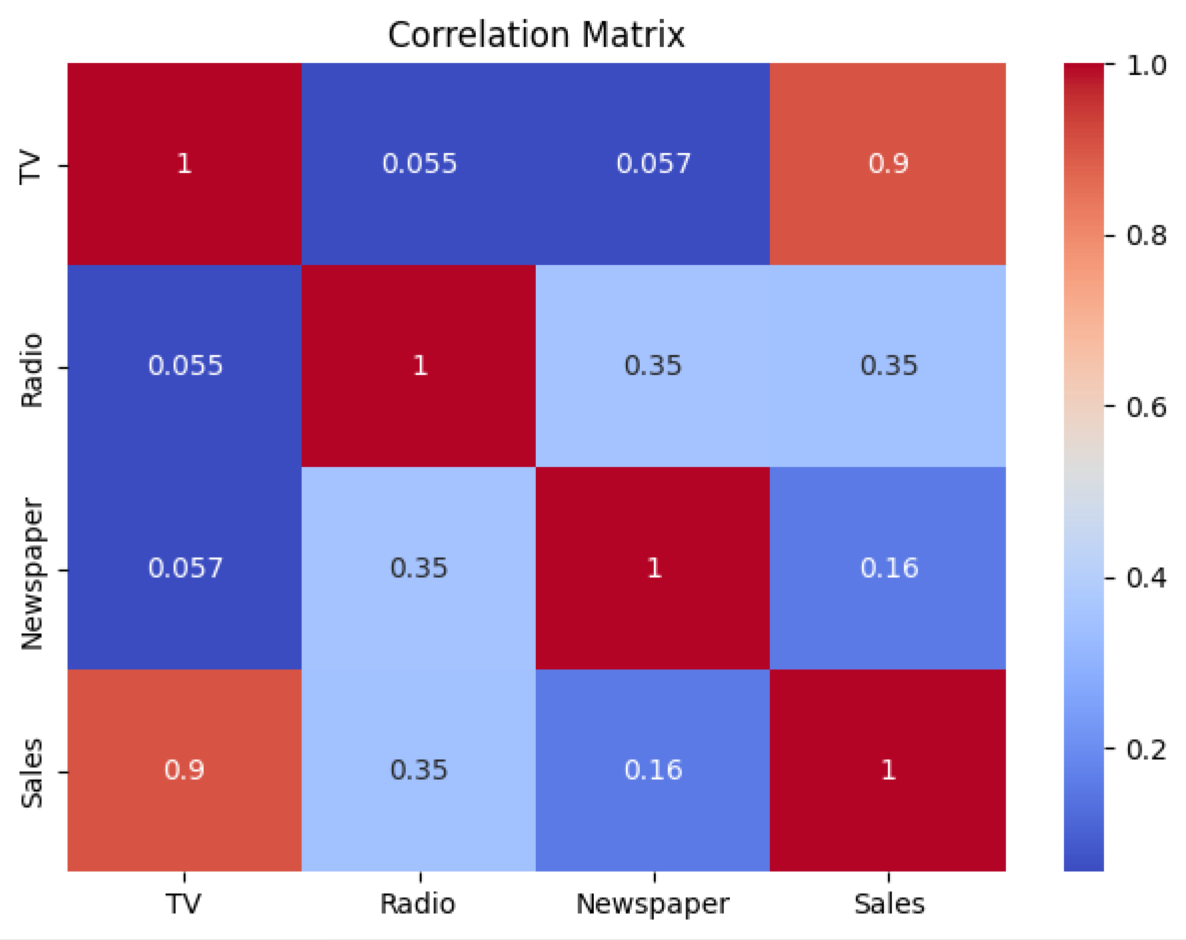
CORRELATION AND VISUALIZATION OF DIFFERENT FEATURES ON SALES LABEL

- The average spend on TV advertising is: 147.0425
- the correlation between radio advertising expenditure and product sales is :

	Sales
TV	0.901208
Radio	0.349728
Newspaper	0.157960

- The advertising medium has the highest impact on sales based on the dataset is : TV i.e **0.901208**

Correlation value close to 1 represents higher positive correlation. The diagonal value of the matrix always 1 because of same column in x and y axis.
And the colour bar represents the weightage of the correlation in the heatmap representation.

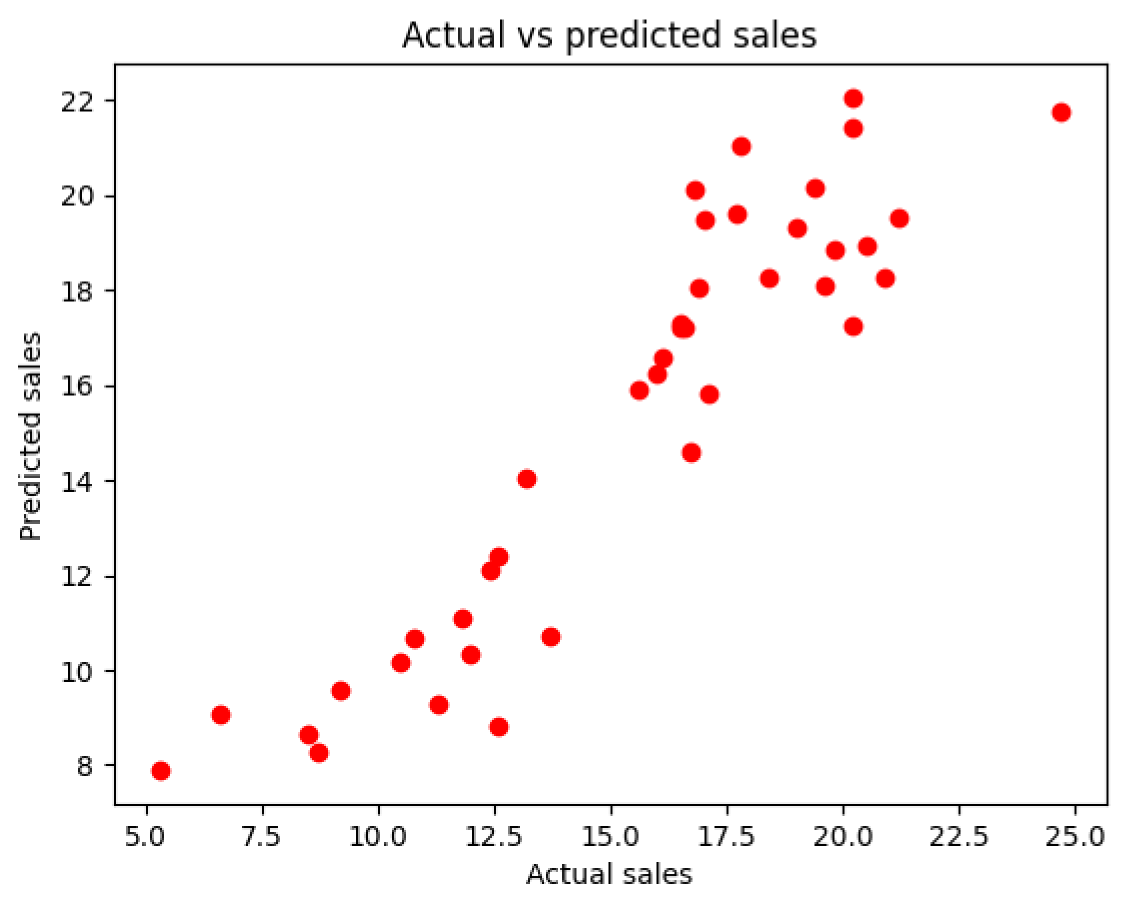


DATA PREPROCESSING

- First I checked if any columns contains null values because null values has high impact on the regression model. Here I got radio column has 2 null values. To remove null values we can delete the entire row or we can use SimpleImputer class from sklearn library to replace numerial column null values with mean or median and the categorical columns with mode(frequently occure).
- I replaced null values with mean of the column.
- Then I decide the feature matrix columns and target label and save with x and y variables respectively.
- After that I divided data into train(x_train,y_train) and test(x_test,y_test) dataset where train part we use in the training the model and test part we use to check the models accuracy .
- Then I used the train data to train the model using the LinearRegression class of sklearn library and make the prediction using x_test and save it in y_pred.

PREDICTION ACCURACY CHECK

- The visualization of the model's predictions(y_pred) against the actual sales values(y_test) is plotted in the graph.
- I used r2_score module from sklear to check the accuracy of the model and I found r2 score around 0.84072805177607 and this shows prediction is better.
- And the sales prediction for a new set of advertising expenditures using the model: $200 on TV, $40 on Radio, and $50 on Newspaper is **19.80605821.**

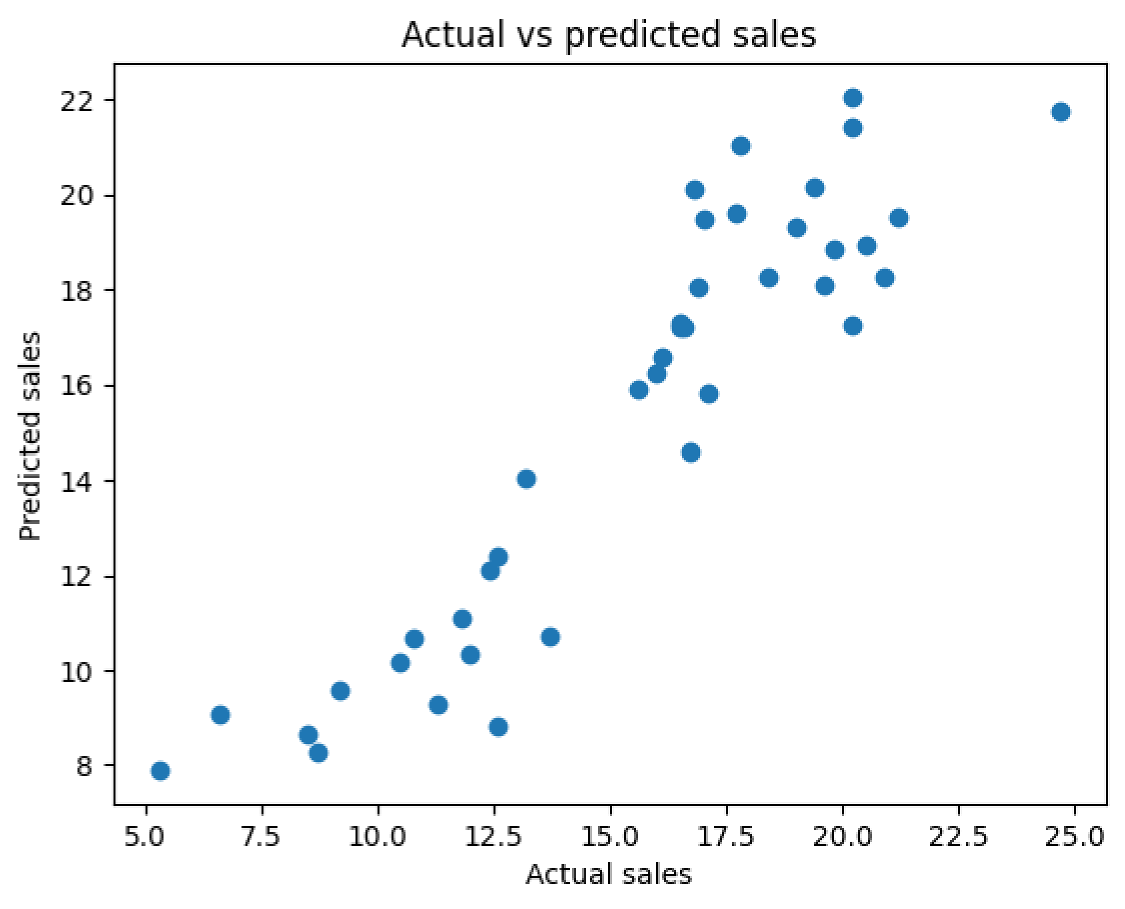


PREDICTION MODEL AFTER NORMALIZATION THE COLUMNS

- I applied feature scaling using the MinMaxScaler class, which normalizes the feature values to a common scale. This step is crucial to eliminate bias caused by features with larger ranges dominating the learning process.
- IMP: Use MinMaxScaler on x_train to fit and transfer and only transfer to X_test and Then used the regression procedure to make the model.

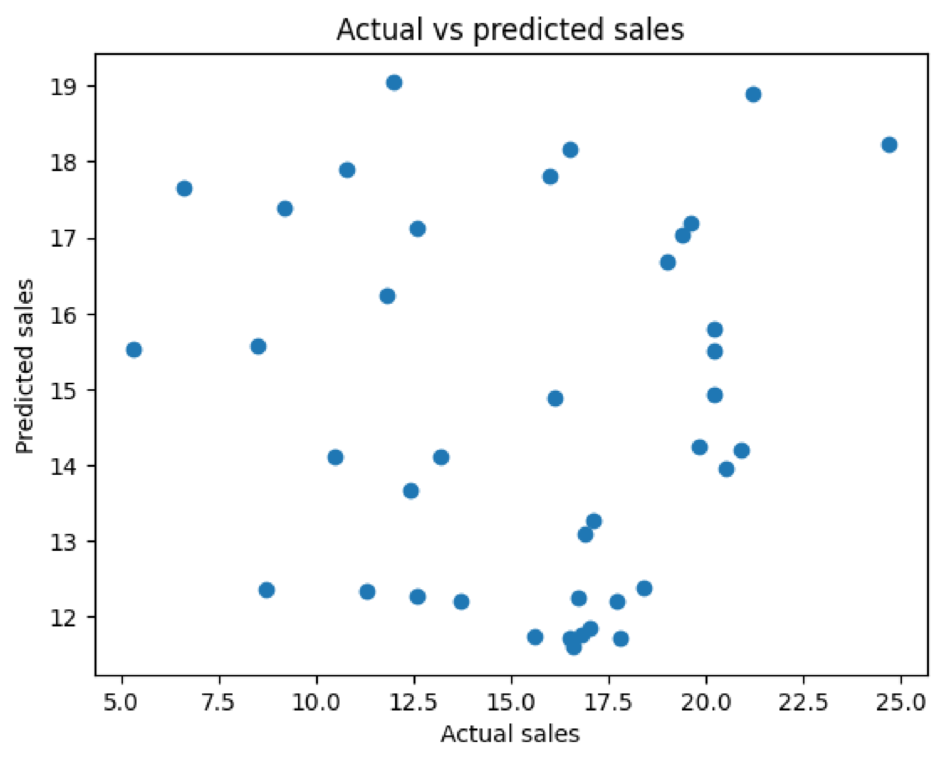
PREDICTION ACCURACY CHECK

- The visualization of the model's predictions(y_pred) against the actual sales values(y_test) is plotted in the graph.
- I used r2_score module from sklear to check the accuracy of the model and I found r2 score around 0.84072805177607 and this shows prediction is better.
- **I got around same r2 score as before normalization, So normalization has not much effect in this model i.e scales of all features in the same range before also.**



MODEL USING ONLY RADIO AND NEWSPAPER

- All the previous steps are same except change the feature matrix i.e only I took radio and newspaper columns. The graph represents the actual and predicted sales based on radio and newspaper feature.
- r2 score: -0.343308946798755I
- And the r2 score represents the model is bad and can not be used for prediction.


Actual vs predicted sales
Predicted sales
Actual sales