Sales Prediction Case Project

Yashmith Raj

SRM AP University

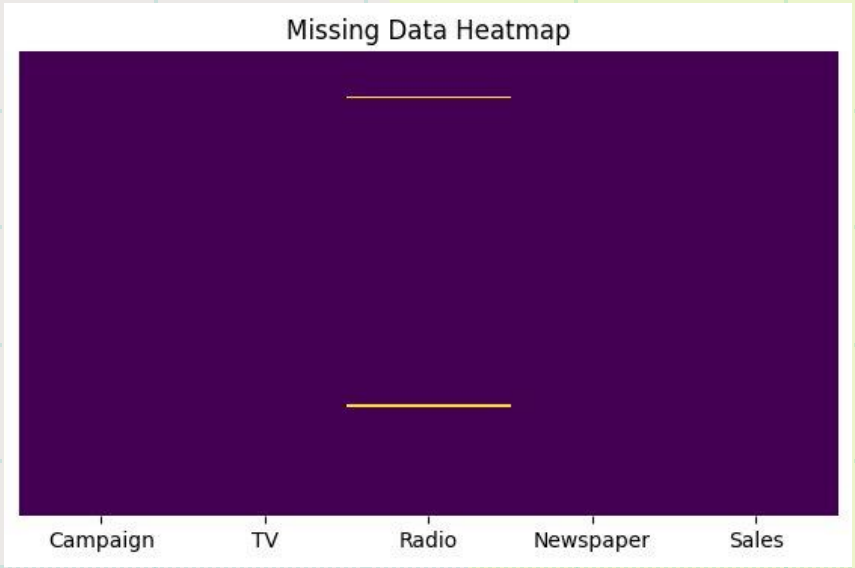
Missing Values Analysis

- The code output reveals missing values in the 'Radio' columns with 2 missing entries.
- Given the small number of missing values relative to the overall dataset size and their random distribution without any specific pattern, we can drop these missing values without significantly affecting the analysis.
- To validate that these missing values are Missing Completely at Random (MCAR), a t-test was conducted comparing the 'Sales' values of rows with missing 'Radio' data against those with non-missing 'Radio' data.
- T-test Results:
  - **T-statistic:** 0.1260
  - **P-value:** 0.8999
  - The high P-value indicates no significant difference between the sales values for rows with and without missing 'Radio' data. Thus, we can confidently drop these missing values.

Handling MissingValues

```
[ ] print(df.isnull().sum())

Campaign     0
TV           0
Radio        2
Newspaper    0
Sales        0
dtype: int64
```



Plot 1:
What is the average amount spent on TV advertising in the dataset?

- The average amount spent on TV advertising is **146.79** (two decimal places).
- TV advertising has the highest mean expenditure (**146.79**) compared to Radio (**23.26**) and Newspaper (**30.69**), indicating that companies prioritize TV for their advertising efforts.

```
print(df.describe())
```

	TV	Radio	Newspaper	Sales
count	198.000000	198.000000	198.000000	198.000000
mean	146.785859	23.260606	30.694444	15.125758
std	86.213342	14.921914	21.842166	5.309478
min	0.700000	0.000000	0.300000	1.600000
25%	73.725000	9.925000	12.650000	11.000000
50%	149.750000	22.400000	26.050000	16.000000
75%	219.475000	36.575000	45.100000	19.150000
max	296.400000	49.600000	114.000000	27.000000

Questions:

1. What is the average amount spent on TV advertising in the dataset?

```
[45] print("average amaount spent on TV advertising: ",df['TV'].mean())
```

```
average amount spent on TV advertising:  146.78585858585862
```

Plot 2:
What is the correlation between radio advertising expenditure and product sales?

- Correlation of 0.35 indicates a moderate positive relationship between radio advertising expenditure and product sales.
- This means that, generally, as radio advertising expenditure increases, product sales also tend to increase, though the relationship is not very strong.
- While there is a positive correlation, the strength is moderate. This suggests that radio advertising does have an effect on sales, but it is not the only factor influencing sales. Other factors and variables may also play significant roles.

Questions:

2. What is the correlation between radio advertising expenditure and product sales?

```
[46] crr = df[['Radio', 'Sales']].corr().loc['Radio', 'Sales']
     print("correlation between radio advertising expenditure and product sales : ",crr)
```

```
correlation between radio advertising expenditure and product sales :  0.3497277129207838
```

Plot 3:
Which advertising medium has the highest impact on sales based on the dataset?

- **TV Advertising:**
  - Regression Coefficient: 0.054494
  - Correlation with Sales: 0.901372
  - **TV advertising has the strongest impact on sales based on both the high correlation and positive regression coefficient.**
- **Radio Advertising:**
  - Regression Coefficient: 0.107180
  - Correlation with Sales: 0.349728
  - **Radio advertising has a moderate impact, with a positive correlation and coefficient.**
- **Newspaper Advertising:**
  - Regression Coefficient: -0.000019
  - Correlation with Sales: 0.159125
  - **Newspaper advertising has the least impact on sales, as evidenced by both its very low correlation and near-zero regression coefficient**.

Regression Coefficients for analysis of highest impact

```
[50] from sklearn.linear_model import LinearRegression

     x = df[['TV', 'Radio', 'Newspaper']]
     y = df['Sales']

     lr = LinearRegression().fit(x, y)
     coefs = pd.DataFrame(lr.coef_, x.columns, columns=['Coeff'])

     print("Regression Coefficients:")
     print(coefs)
```

```
Regression Coefficients:
              Coeff
TV         0.054494
Radio      0.107180
Newspaper -0.000019
```

```
[49] crr = df.corr()
     print(crr)
```

	TV	Radio	Newspaper	Sales
TV	1.000000	0.054848	0.059075	0.901372
Radio	0.054848	1.000000	0.354987	0.349728
Newspaper	0.059075	0.354987	1.000000	0.159125
Sales	0.901372	0.349728	0.159125	1.000000

4. Plot a linear regression line that includes all variables to predict Sales, and visualize the model's predictions against the actual sales values.

- The scatter plot visualizes the relationship between actual sales and predicted sales based on a linear regression model that includes TV, Radio, and Newspaper advertising expenditures as predictors.
- **Positive Correlation:** The overall trend suggests a positive correlation between actual and predicted sales, indicating that the model is able to capture some of the underlying relationship between advertising expenditure and sales.
- **Scatter:** The points are scattered around the regression line, indicating that the model is not perfect in predicting sales. There is some variability in the data that the model is unable to explain.
- **Regression Line:** The red line represents the linear regression line, which represents the model's best estimate of the relationship between actual and predicted sales.

Questions:

5. How would sales be predicted for a new set of advertising expenditures: $200 on TV$, 40 on Radio, and $50 on Newspaper?

```
[ ]  sample = pd.DataFrame({
         'TV': [200],
         'Radio': [40],
         'Newspaper': [50]
     })

     y_pred_sm = lr.predict(sample)
     print(f"Predicted Sales for TV=$200, Radio=$40, Newspaper=$50: $",y_pred_sm)
```

```
Predicted Sales for TV=$200, Radio=$40, Newspaper=$50: $ [19.81937816]
```

5. How would sales be predicted for a new set of advertising expenditures: $200 on TV, $40 on Radio, and $50 on Newspaper?

- The prediction utilizes a linear regression model to estimate sales based on specified advertising expenditures for TV, Radio, and Newspaper.
- **Advertising Expenditures:**
  - TV: $200
  - Radio: $40
- Newspaper: $50
- **Predicted Sales:** The model predicts that with these advertising expenditures, the sales would be **approximately $19.82.**

6. How does the performance of the linear regression model change when the dataset is normalized?

- **Normalization Impact:** Normalizing the data using the MinMaxScaler has significantly reduced MAE, MSE, and RMSE values, indicating that normalization improves the model's accuracy by making predictions closer to the actual values.
- **Performance Improvement:** The reduction in MAE, MSE, and RMSE after normalization suggests that scaling features can enhance model performance. This is because normalization helps to ensure that all features contribute equally to the model, making the optimization process more effective.
- **R-squared Consistency:** The R-squared value remains constant, implying that normalization improves prediction accuracy without altering the model's explanatory power.
- **Enhanced Model Accuracy:** The significant reduction in error metrics after normalization demonstrates the importance of data preprocessing. Normalizing data leads to more precise and reliable predictions.

```
Actual data:
Mean Absolute Error (MAE): 1.2397
Mean Squared Error (MSE): 2.7230
Root Mean Squared Error (RMSE): 1.6502
R-squared (R2): 0.9029

Normalized data:
Mean Absolute Error (MAE): 0.0488
Mean Squared Error (MSE): 0.0042
Root Mean Squared Error (RMSE): 0.0650
R-squared (R2): 0.9029
```

7. What is the impact on the sales prediction when only radio and newspaper advertising expenditures are used as predictors?

```
Actual data:
Mean Absolute Error (MAE): 1.2397
Mean Squared Error (MSE): 2.7230
Root Mean Squared Error (RMSE): 1.6502
R-squared (R2): 0.9029

Reduced data:
Mean Absolute Error (MAE): 4.2631
Mean Squared Error (MSE): 24.5784
Root Mean Squared Error (RMSE): 4.9577
R-squared (R2): 0.1237
```

- **Model with All Features (TV, Radio, Newspaper):**
  - **MAE:** 1.2397
  - **MSE:** 2.7230
  - **RMSE:** 1.6502
  - **R-squared (R^2):** 0.9029
- **Model with Reduced Features (Radio and Newspaper only):**
  - **MAE:** 4.2631
  - **MSE:** 24.5784
  - **RMSE:** 4.9577
  - **R-squared (R^2):** 0.1237

- **Increased Errors:**
  - **MAE, MSE, and RMSE:** All error metrics are significantly higher when TV advertising expenditure is excluded. This indicates that the accuracy of sales predictions deteriorates without including TV data.
- **R-squared Value:**
  - **R-squared:** The R-squared value drops drastically from 0.9029 to 0.1237 when TV is excluded. This demonstrates that the model with only Radio and Newspaper expenditures explains far less of the variance in sales compared to the model that includes TV advertising.

Final Insights of the dataset:

- **Impact of Normalization:**

- **Observation:** Normalizing the data using MinMaxScaler significantly reduced MAE, MSE, and RMSE values.
- **Insight:** Normalization enhances model accuracy by making predictions closer to actual sales values and improves performance by ensuring that all features contribute equally to the model. The R-squared value remains unchanged, indicating that normalization improves accuracy without affecting the model's explanatory power.

- **Effect of Feature Selection:**

- **Observation:** Excluding TV advertising expenditure led to a substantial increase in MAE, MSE, and RMSE, and a significant drop in R-squared from 0.9029 to 0.1237.
- **Insight:** TV advertising expenditure plays a critical role in predicting sales. The inclusion of TV data provides a better model fit and more accurate predictions compared to using only Radio and Newspaper expenditures. The substantial increase in error metrics and decrease in R-squared highlight the importance of including TV data for effective sales prediction.

My Experience :

- **Normalization and Feature Scaling:** Explored the effects of Min-Max normalization on model accuracy. I learned how scaling features to a uniform range can significantly enhance the performance of regression models, making predictions more precise and reliable.
- **Handling Missing Data:** Gained insight into the mechanics of Missing Completely At Random (MCAR) and its implications for data analysis. Understood strategies for managing missing data, including imputation and exclusion, to ensure data quality and improve model performance.
- **Correlation Analysis:** Analyzed the relationships between advertising expenditures and sales to determine their impact. This involved using correlation metrics to understand how different advertising channels contribute to sales outcomes and refining feature selection accordingly.
- **Model Evaluation and Comparison:** Enhanced my skills in evaluating regression models by comparing performance metrics such as MAE, MSE, RMSE, and R-squared. This allowed me to assess the effectiveness of different features and preprocessing techniques in predicting sales.