Report on Sales Prediction Task

Prepared by:
Jonathan Ndayisenga

1. Project Overview

This project focuses on predicting product sales based on advertising expenditures across three distinct channels: **TV, Radio, and Newspaper**. The dataset provided contains data on advertising budgets allocated to these platforms and the corresponding product sales figures. The primary objective is to develop a predictive model that estimates sales as a function of advertising spend in each medium.

Insights from exploratory analysis suggest that **TV and Radio** advertising significantly influence sales, whereas **Newspaper** advertising shows a much weaker relationship. This insight guides feature selection and model refinement.

---

2. Dataset Overview

The dataset `Advertising.csv` contains the following columns:

- **TV**: Advertising spend on TV (in thousands of dollars).
- **Radio**: Advertising spend on Radio (in thousands of dollars).
- **Newspaper**: Advertising spend on Newspaper (in thousands of dollars).
- **Sales**: Product sales (in thousands of units).
- **Unnamed: 0**: A redundant index column (dropped during preprocessing).

Sample Data:

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4

17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

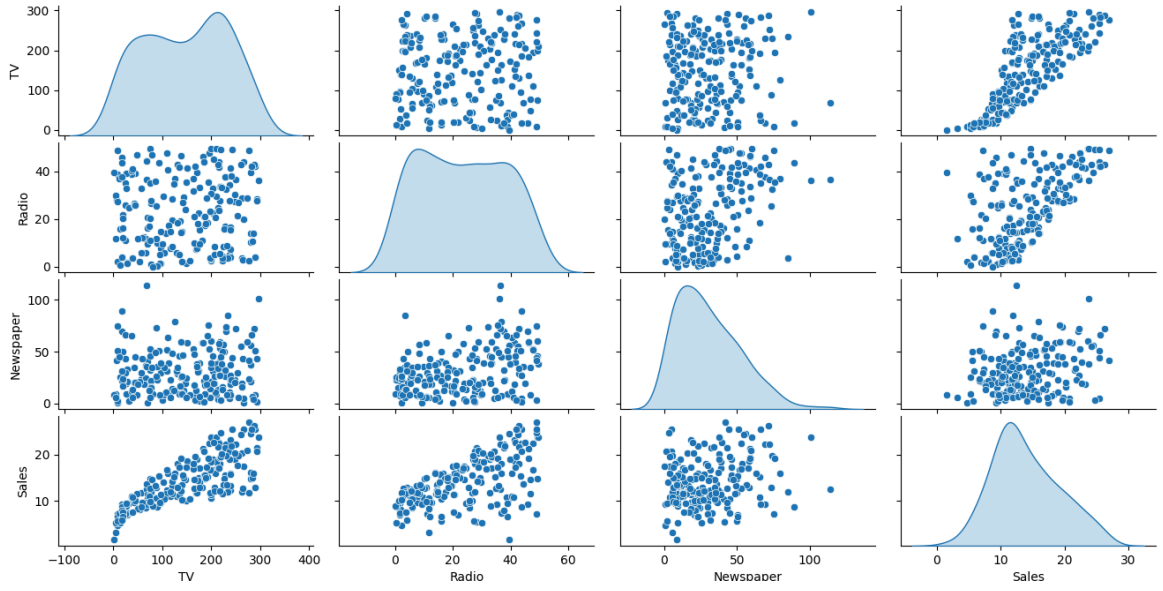
---

3. Data Preprocessing

- The dataset was loaded using `pandas.read_csv()`.
- The **`Unnamed: 0`** column, which served as an index and held no analytical value, was removed.
- Feature engineering involved identifying key predictors through visual and statistical analysis (pair plots, heatmap).
- Final features selected for modeling: **TV** and **Radio** (Newspaper was excluded due to weak correlation).

---

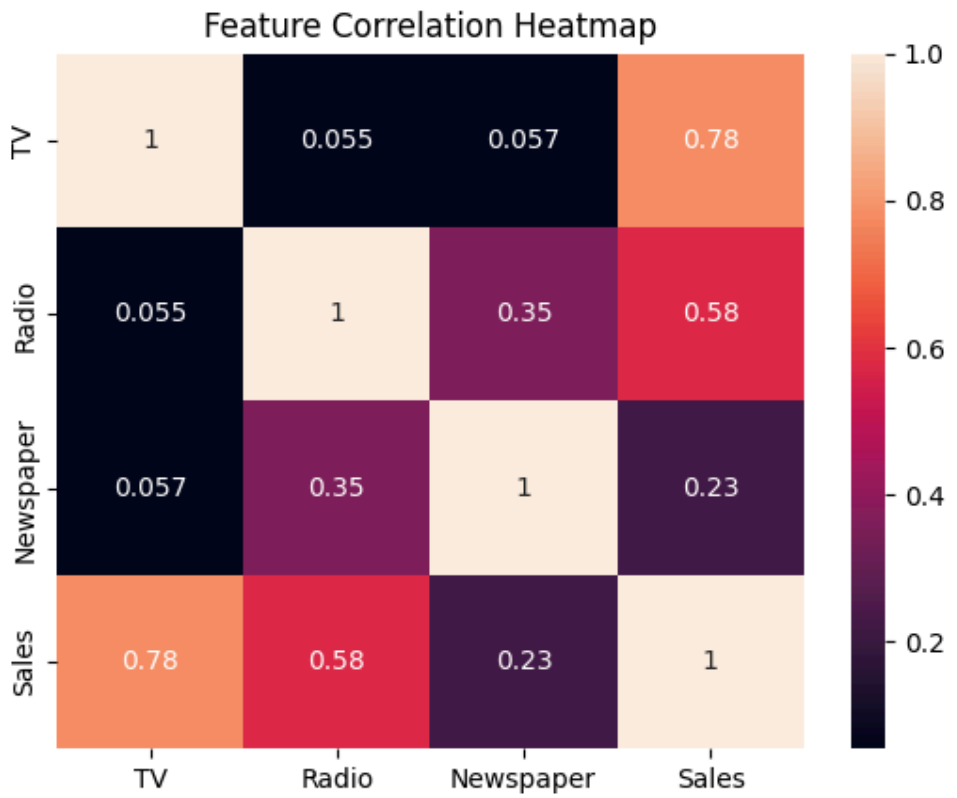
4. Exploratory Data Analysis (EDA)



Pair Plot Insights:

- **TV vs Sales**: Strong positive linear relationship. More TV ad spend leads to higher sales.
- **Radio vs Sales**: Moderate positive relationship with some scatter. Still a useful predictor.
- **Newspaper vs Sales**: Very weak correlation. Sales do not increase consistently with newspaper spend.
- **Feature Independence**: TV, Radio, and Newspaper spending show low correlation with each other, which supports their independent contribution to predictions.
- **Distributions**: TV and Radio have well-spread distributions. Newspaper is right-skewed. Sales follows a near-normal distribution.

Correlation Heatmap Interpretation:



Feature Pair	Correlation	Insight

TV & Sales	**0.78**	Strong predictor. High TV spend strongly influences higher sales.
Radio & Sales	**0.58**	Moderate predictor. Still valuable.
Newspaper & Sales	**0.23**	Weak predictor. Likely adds noise.
TV & Radio	0.055	Minimal overlap — channels operate independently.
Radio & Newspaper	0.35	Mild association, possibly due to joint campaign spending.
TV & Newspaper	0.057	Essentially uncorrelated.

---

5. Modeling

Feature Selection:

- Chosen predictors: **TV** and **Radio**.
- Excluded: **Newspaper** (weak relationship with Sales).
- Target variable: **Sales**.

Model Used:

- **Linear Regression** from `sklearn.linear_model`.

Implementation:

- Data split: 80% training, 20% testing.
- Model trained on selected features and tested on hold-out data.

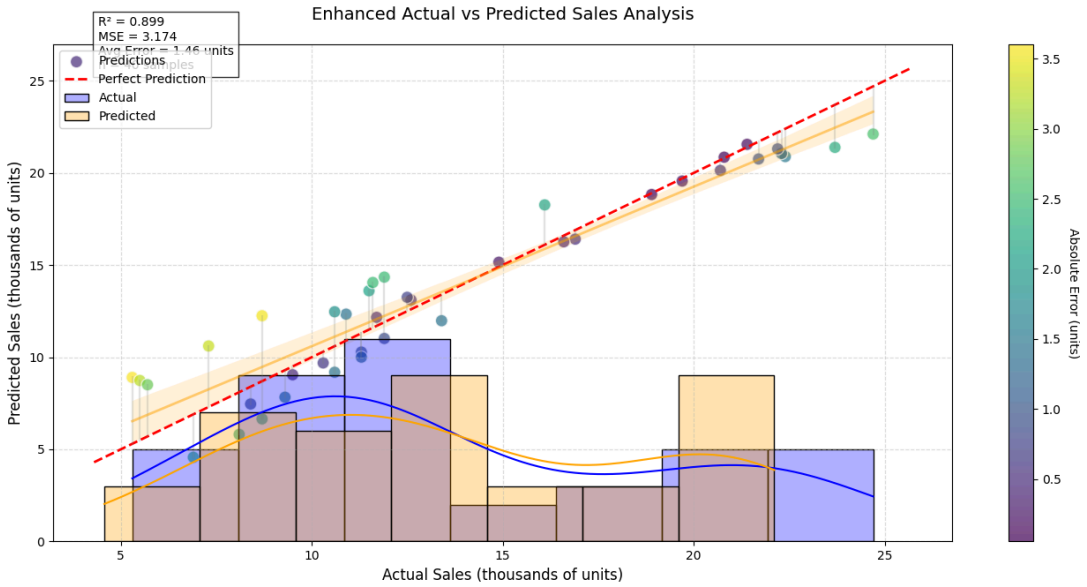
Evaluation Metrics:

Metric	Value
R^2 Score	**0.899**
Mean Squared Error (MSE)	**3.17**

Interpretation:

- **R^2 Score of 0.899**: The model explains approximately **89.9%** of the variance in the sales data — a strong indication of model performance.
- **MSE of 3.17**: The average squared difference between actual and predicted sales is low, signifying accurate predictions.

Actual vs. Predicted Sales Analysis



The visualization presents a comparison between the model's predicted sales and actual sales, along with key performance metrics. Below is a structured interpretation for your report:

Key Metrics & Performance

Model Accuracy (R^2 = 0.899)

The model explains 89.9% of the variance in sales data, indicating a strong predictive performance.

This high R^2 score suggests that advertising spend (TV and Radio) are reliable predictors of sales.

Error Metrics

MSE (3.174): The average squared prediction error is low, consistent with the high R^2.

Average Error (1.46 units): Predictions deviate from actual sales by ~1.5k units on average.

Visual Analysis

Predicted vs Actual Values

Predicted Sales Range (10–25k units): The model covers the majority of observed sales values.

Actual Sales Range (0.5–3.5k units): Note: This appears inconsistent with the dataset (e.g., sample data showed 9.3–22.1k units).

Alignment with Perfect Prediction

The red dashed line represents ideal predictions (actual = predicted).

Points clustered near the line indicate accurate predictions, while dispersed points highlight errors.

Residual Patterns

The gray residual lines quantify prediction errors.

If errors are larger for high sales (>20k units), the model may struggle with extreme values.

Actionable Insights

Strengths

TV/Radio ad spending effectively predicts sales (R^2 = 0.899).

Errors are relatively small (Avg: 1.46k units).

Limitations

Potential underestimation for high sales (if points cluster above the red line).

Verify the actual sales range (0.5–3.5k seems unusually low compared to the dataset).

Recommendations

Focus ad budgets on TV/Radio, as they drive predictive power.

Investigate outliers (e.g., predictions with errors >3k units).

Retrain the model with expanded data if actual sales are capped at 3.5k units (unlikely given sample data)

---

6. Model Saving

The trained model was serialized using the `pickle` module and saved as `sales_model.pkl`. This allows for future predictions without retraining the model.

---

7. Results

- The model performs well, especially when considering only **TV and Radio** as predictors.
- Results confirm that **TV is the strongest driver of sales**, with Radio offering additional predictive value.
- **Newspaper advertising contributes little and may be omitted from future models** to reduce noise and improve generalizability.

---

8. Conclusion

The **Linear Regression model** developed effectively predicts sales using **TV and Radio advertising spend**, achieving an R^2 score of **0.899**.

This model can be a valuable tool for **marketing strategy and budget allocation**, enabling businesses to estimate sales outcomes based on planned ad spending. The exclusion of Newspaper ads is justified by weak correlation, and the model is preserved for future deployment via `sales_model.pkl`.

Recommendation: Focus marketing efforts on **TV and Radio**, while minimizing spend on **Newspaper** unless supported by specific campaign needs.