**Advertisement Budgets Project Documentation**

1. **Introduction**

The purpose of this project is to analyze how advertising budgets across various media channels—TV, Radio, and Newspaper— affect product sales. We aim to build a predictive model that estimates sales figures based on how much is spent on each channel.

The dataset used is the well-known Advertising.csv, which contains 200 entries and four variables: TV, Radio, Newspaper advertising budgets, and the resulting sales. This task involves cleaning the dataset, handling outliers, transforming features where necessary, and applying a regression model to evaluate predictive performance.

1. **Data Preprocessing**

**2.1. Overview of the Dataset**

* Source: Advertising.csv
* Dimensions: 200 rows and 4 columns
* Columns: TV, Radio, Newspaper, Sales

**2.2 Data Cleaning**

* Checked for missing values: none found.
* Checked for duplicated columns: none.
* Dropped redundant column Unnamed: 0 (just an index).

**2.3 Outlier Detection**

* The Newspaper column showed potential outliers.
* Applied the Interquartile Range (IQR) method:
  + Lower bound: Q1 - 1.5 \* IQR
  + Upper bound: Q3 + 1.5 \* IQR
* Capped values exceeding the upper/lower bound.

**2.4 Feature Transformation**

* Created three new versions of Newspaper:
  + Newspaper\_capped: Outliers capped.
  + Newspaper\_log: Log-transformed to reduce skew.
  + Newspaper\_binned: Binned into categories — Low, Medium, High.

**2.5 Feature Scaling**

* Used StandardScaler to scale all numerical features to have zero mean and unit variance. This is especially helpful for models sensitive to feature magnitudes like linear regression.

1. **Model Evaluation**

**3.1 Model Selection**

* Model used: Linear Regression from sklearn.linear\_model.

**3.2 Splitting the Data**

* Used train\_test\_split with an 80/20 ratio:
  + Training set: 160 samples
  + Testing set: 40 samples

**3.3 Performance Metrics**

* Mean Squared Error (MSE): Measures average squared error.
* R² Score: Indicates how well the model explains the variance.

**3.4 Results (Expected) Based on typical behavior of this dataset:**

* Training R² Score ≈ 0.90–0.95
* Testing R² Score ≈ 0.88–0.92
* MSE is low, indicating a good fit.

These results suggest that the linear regression model is effective for this task, especially when using transformed versions of the Newspaper feature.

1. **Conclusion**

In this project, we explored the relationship between advertising expenditure and product sales using a regression approach. After data cleaning and preprocessing (especially addressing outliers in the Newspaper column), we applied linear regression to build a predictive model.

Key Insights:

* TV and Radio advertising have a strong positive effect on sales.
* Newspapers have a weaker correlation and require transformation.
* Outlier handling and feature scaling significantly improved model performance.
* Linear regression was sufficient to achieve high accuracy with low error.

Recommendations for Future Work:

* Try Ridge or Lasso regression for regularization.
* Investigate interaction effects between media types.