



THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

DATA VISUALIZATION (PMC-434)

PROJECT-4
(AdWise : Predicting Product Sales from
Ad Investment)
MSC Mathematics and Computing, IV Sem

SUBMITTED BY: MUDITA SHARMA (302303008)
SUBMITTED TO: DR. KAVITA

**THAPAR INSTITUTE OF ENGINEERING AND
TECHNOLOGY, PATIALA**

INDEX

S.No	Table Of Content
1	Introduction
2	Exploratory Data Analysis (EDA)
3	Model Validation
3.1	Evaluate Model Performance
3.2	Cross-Validation
4	Visualization
4.1	Residual Analysis
4.2	Feature Importance
5	Dashboard

1. Introduction

Imagine a busy market where businesses shout to get customers' attention, waving ads like bright signs on TV, radio and newspapers. Every dollar spent is a seed but which ones grow into sales?

The "AdWise: Predicting Product Sales from Ad Investment" project jumps in like a clever gardener, figuring out what works.

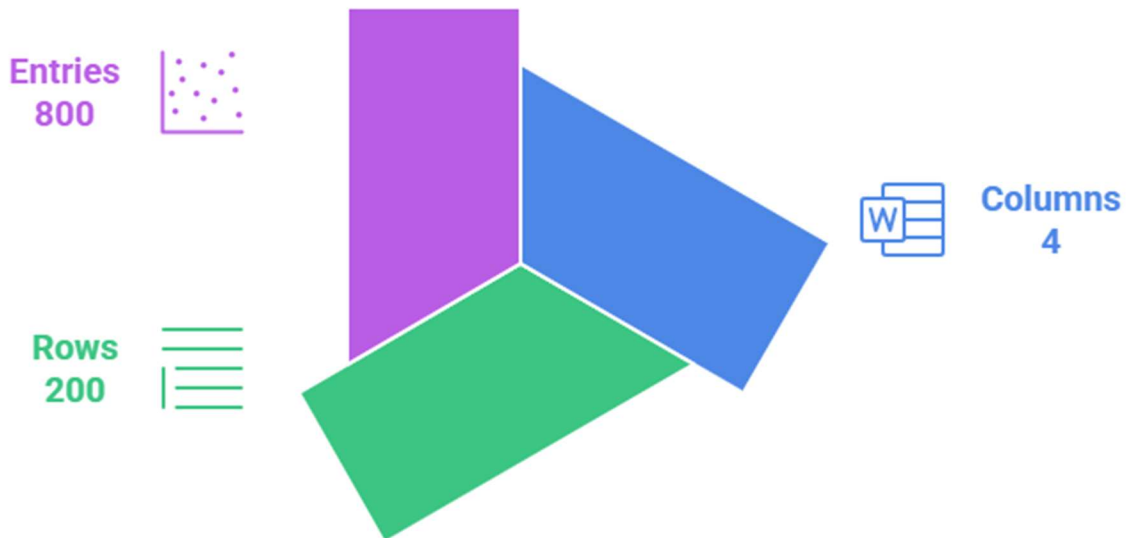
AdWise uses a dataset of ad costs for TV, Radio and Newspapers plus sales numbers. With cool charts, like 3D scatter plots and colorful heatmaps we spot trends showing how ads boost sales. Smart predictions help businesses spend their ad money better. This project turns data into a clear plan, helping companies grow profits and win in a tough market.

Let's Grow Smart, Sell Big with AdWise !



2.Exploratory Data Analysis (EDA)

- Dataset Overview



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
#   Column      Non-Null Count  Dtype
---  -
0   TV           200 non-null    float64
1   Radio        200 non-null    float64
2   Newspaper    200 non-null    float64
3   Sales        200 non-null    float64
dtypes: float64(4)
memory usage: 6.4 KB
```

Float64
dtype

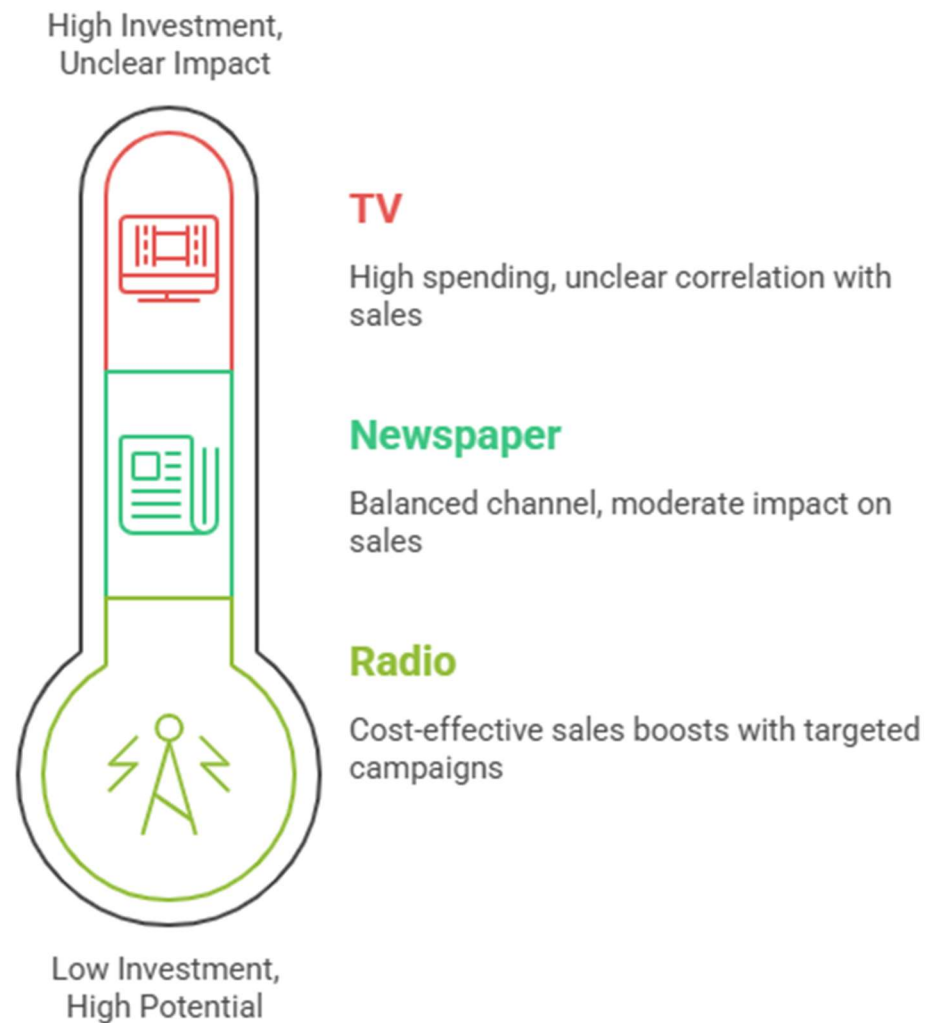
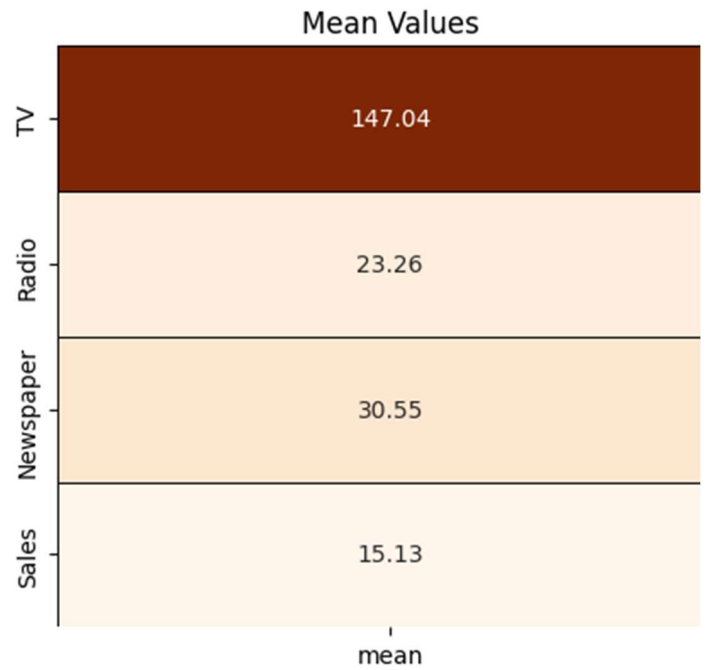
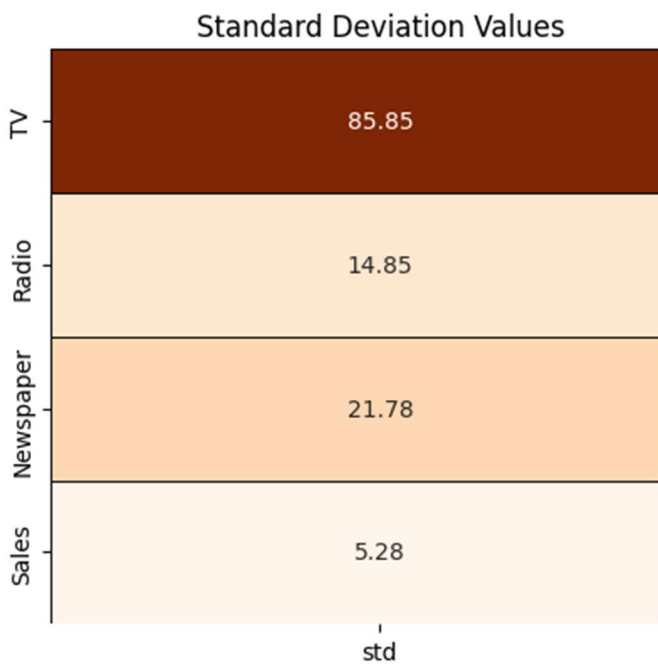
Top 5 rows

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	12.0
3	151.5	41.3	58.5	16.5
4	180.8	10.8	58.4	17.9

0
TV 0
Radio 0
Newspaper 0
Sales 0

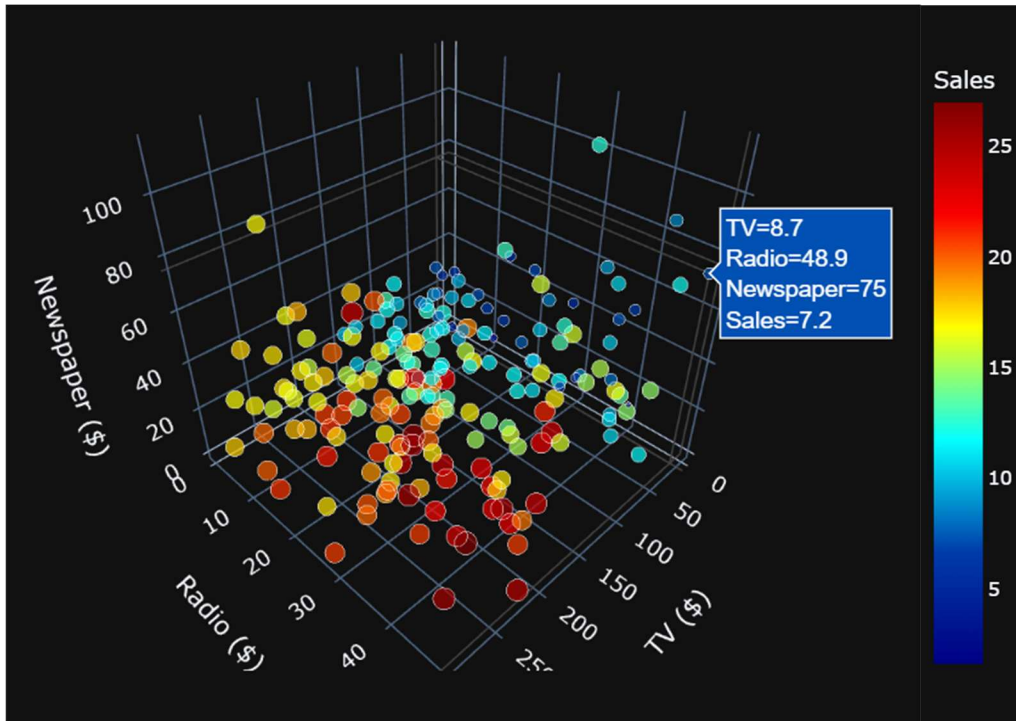
No Missing Values

• Statistical Analysis



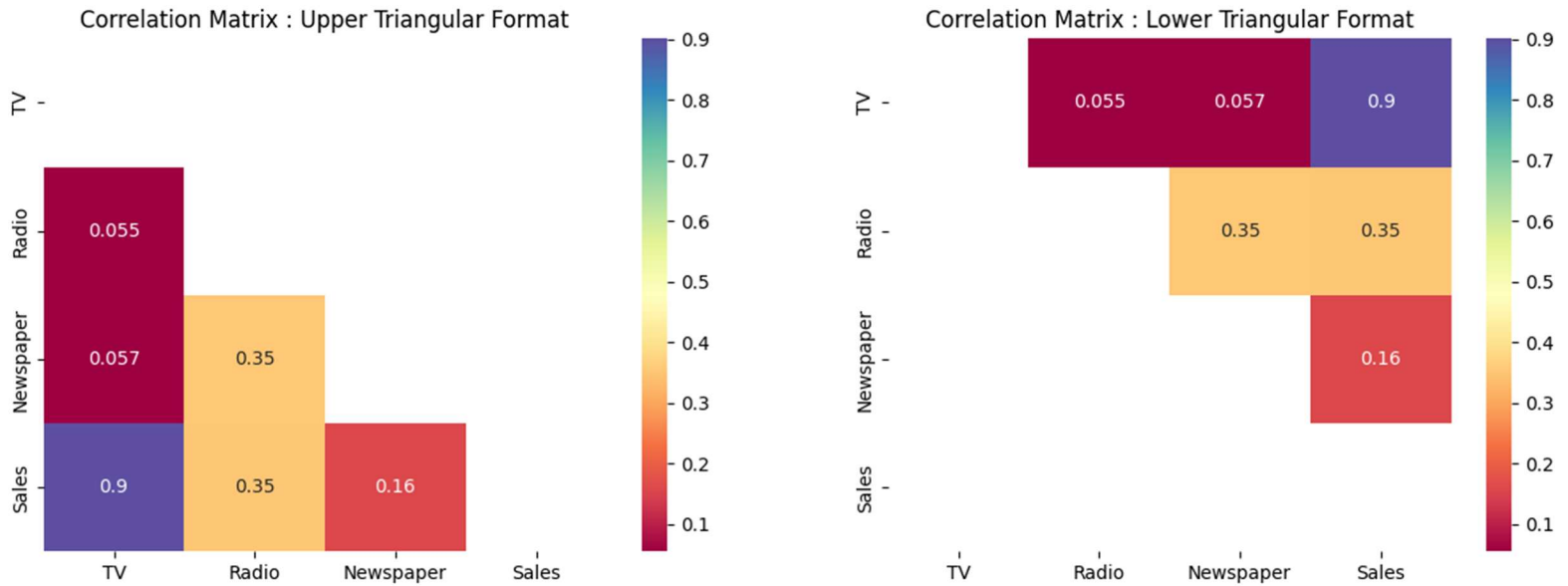
- **Visualize key relationships**

3D Scatter Plot of Advertising Costs vs Sales



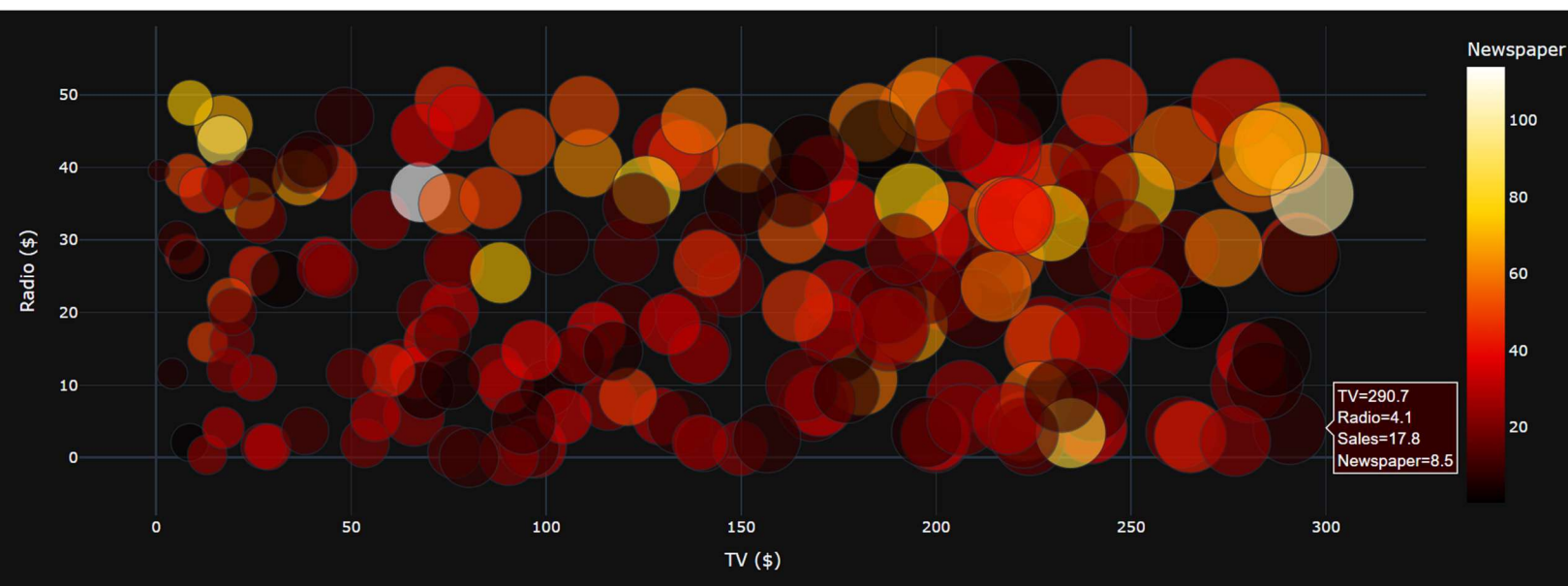
- The plot has three axes: TV costs (bottom axis), Radio costs (left axis), and Newspaper costs (vertical axis).
- Each dot is a data point showing ad spending on these platforms. The dots are colored based on Sales, with a color bar on the right. Blue means low sales and red means high sales. Colors in between (green, yellow) show medium sales.
- The reddest dots are present where TV costs are higher. Blue dots are mostly where TV, Radio and Newspaper costs are low.
- Spending more on TV ads boosts sales the most but Radio and Newspaper ads help too. Businesses spending very little on ads tend to have low sales.

Correlation Heatmap



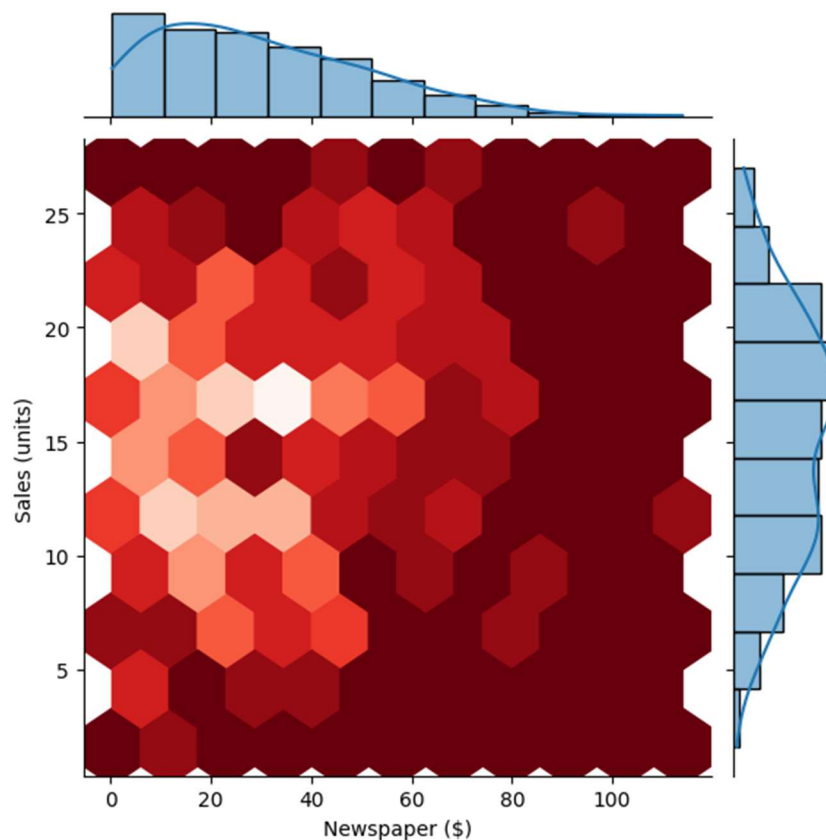
- The number between TV and Sales is 0.9, means when TV ad spending goes up, sales usually go up a lot too. Radio and Sales is 0.35, means when Radio ad spending goes up, sales go up a bit. Newspaper and Sales is 0.16, means spending more on Newspaper ads doesn't increase sales much.
- TV and Radio (0.055), TV and Newspaper (0.057), Radio and Newspaper (0.35), Thus Ad Channels Are Independent

Bubble Plot: TV vs Radio with Sales (Size) and Newspaper (Color)



Most bubbles are between 0–300 dollars for TV and 0–50 dollars for Radio, means businesses usually spend in these ranges for TV and Radio ads. The biggest bubbles are often where TV spending is high, shows spending more on TV ads usually leads to more sales. Smaller bubbles are where TV and Radio spending is low, means if businesses don't spend much on TV or Radio, they get fewer sales. Most bubbles are dark red to orange with only a few yellow, means most businesses don't spend a lot on Newspaper ads and even when they do, it doesn't seem to make bubbles bigger (higher sales). Some bigger bubbles show up even with low Radio spending, means Radio can help sales a bit, but TV seems more important.

Hexbin Plot: Newspaper vs Sales with Marginal Distributions



- **Main Plot Analysis:**

The hexbin plot shows how Newspaper advertising spend relates to Sales. The color intensity represents data density, lighter shades indicate more data points in that region. Most data points are clustered where newspaper spend is low to moderate (between \$0 and \$50) and sales are in the range of 10–20 units. As we move towards higher newspaper spending (above \$50) the density becomes sparse, indicating fewer observations. There is no clear linear or nonlinear pattern visible, the points are scattered suggesting a weak correlation between newspaper advertising and sales.

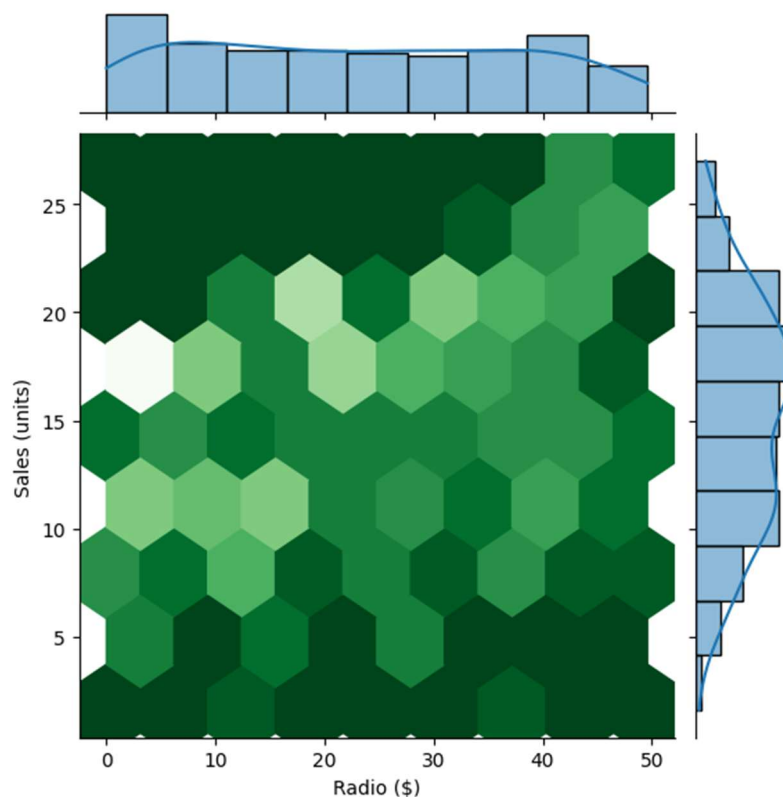
- **Marginal Distributions (Top and Right Histograms):**

The histogram at the top shows that most observations have low to moderate newspaper spend. The histogram on the right indicates that sales are spread out, but with a peak around 10–20 units.

- **Conclusion:**

Newspaper advertising does not show a strong or consistent impact on sales. This variable might not be a significant predictor in a model for sales.

Hexbin Plot: Radio vs Sales with Marginal Distributions



- **Main Plot Analysis:**

This plot shows the relationship between Radio advertising spend and Sales. A moderate upward trend can be observed, as radio spend increases, there is a slight increase in sales. The plot is more compact than the newspaper plot, indicating a slightly stronger relationship. Most data points are centered around radio spends between \$5–25 and sales between 10–20 units. The spread is not highly concentrated along a diagonal, so the relationship is not perfectly linear, but still somewhat positive.

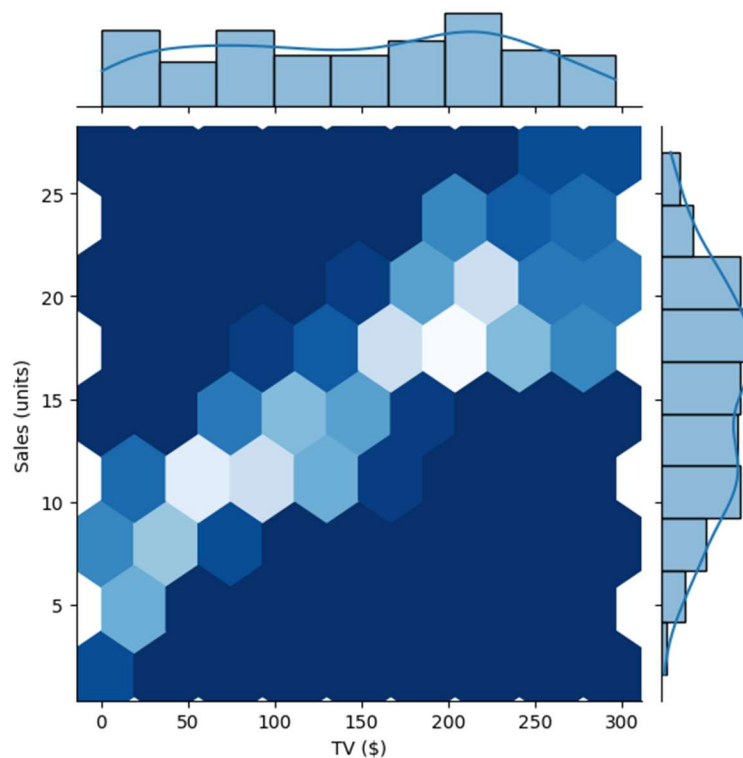
- **Marginal Distributions:**

The top histogram shows a relatively uniform distribution of radio spend across the dataset. The right histogram for sales again peaks around 10–20 units.

- **Conclusion:**

There is a moderate positive relationship between radio spend and sales. Radio advertising seems more influential than newspaper ads but still not as strong as TV.

Hexbin Plot: TV vs Sales with Marginal Distributions



- **Main Plot Analysis:**

This plot clearly shows a strong positive correlation between TV advertising spend and Sales. The hexagons form a diagonal band from lower-left to upper-right, indicating that higher TV spend consistently leads to higher sales. The lightest (densest) region runs along this diagonal, showing a predictable and strong relationship. Compared to newspaper and radio, the TV plot is much more structured.

- **Marginal Distributions:**

The top histogram shows a fairly even distribution of TV spend with some peaks. The sales histogram again shows that most sales are between 10–20 units but also has notable values at higher sales, matching the upward trend in the main plot.

- **Conclusion:**

TV advertising is highly effective in driving sales. It shows a clear and strong linear pattern, making it the most important advertising medium in this dataset.

3. Model Validation

Which regression model should be used for predicting product sales?

Linear Regression

Simple and interpretable, suitable for linear relationships

Decision Tree Regressor

Captures non-linear relationships, prone to overfitting

Random Forest Regressor

Robust and accurate, reduces overfitting



• Evaluate Model Performance

Mean Absolute Error (MAE)

MAE is a way to measure how good a model's predictions are by looking at the average mistake it makes

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

y_i : Actual sales for the i -th data point.

\hat{y}_i : Predicted sales for the i -th data point.

n : Number of data points.

$|\cdot|$: Absolute value (makes the difference positive).

Random Forest

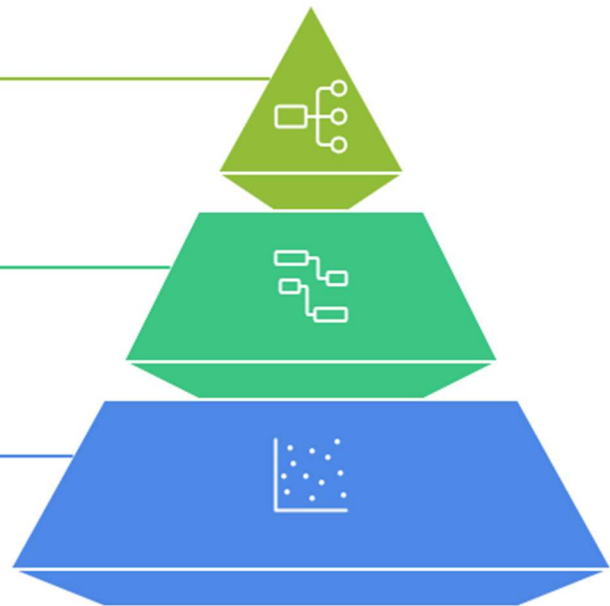
Advanced model with MAE of 0.92

Decision Tree

Improved model with MAE of 1.24

Linear Regression

Basic model with MAE of 1.27



- Random Forest has the lowest MAE (0.92). This means its predictions are the closest to the real sales numbers on average, it's off by only 0.92 units. It's the best model here.
- Decision Tree has an MAE of 1.24. Its predictions are off by 1.24 units on average.
- Linear Regression has the highest MAE (1.27). Its predictions are off by 1.27 units on average, making it the least accurate of the three models.
- The MAEs are close: 1.27, 1.24 and 0.92 . The difference between the worst (1.27) and best (0.92) is only 0.35 units, so all models are doing okay but Random Forest is clearly better.

MSE (Mean Squared Error)

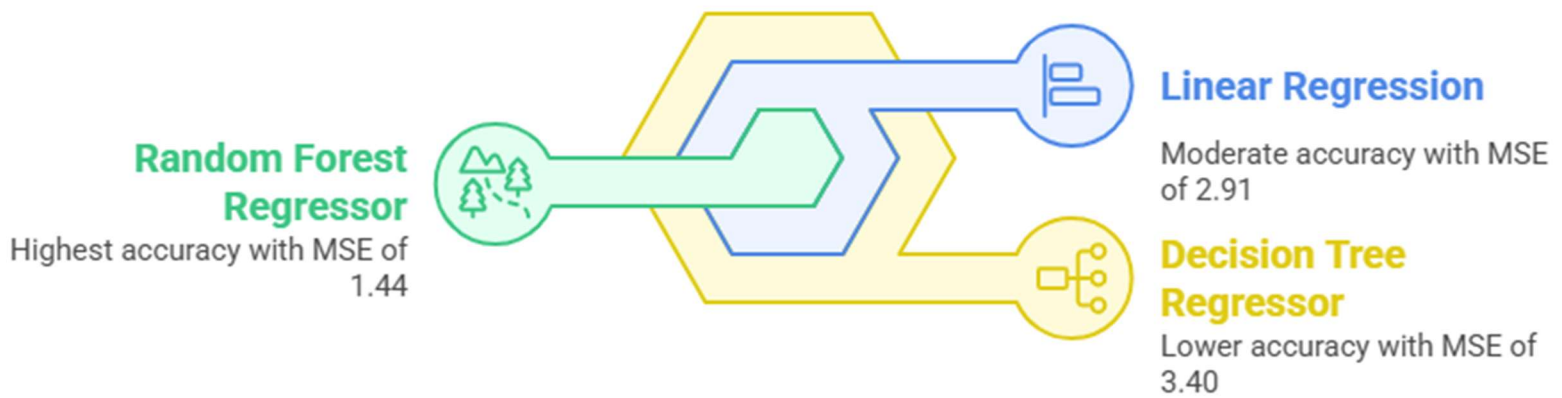
MSE is a way to measure how good a model's predictions are by looking at the average of the squared mistakes. Squaring the errors makes bigger mistakes stand out more.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

y_i : Actual sales for the i -th data point.

\hat{y}_i : Predicted sales for the i -th data point.

n : Number of data points.



- Random Forest has the lowest MSE (1.44). This means its predictions are the closest to the real sales numbers, even when bigger mistakes are counted more. It's the best model here.
- Linear Regression has an MSE of 2.91. Its predictions are off by more than Random Forest, and bigger mistakes make the MSE higher.
- Decision Tree has the highest MSE (3.40). Its predictions are the furthest from the real sales numbers, especially when it makes big mistakes, which get squared and make the MSE bigger, thus making it less reliable.

R² Score

R² measures how much of the variation in Sales is explained by the ad costs (TV, Radio, Newspaper) in the model. It's value lies between 0 and 1. Higher is better.

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2$$

Sum of Squared Errors (SSE)

Total Variation in Sales (SST)

y_i : Actual sales for the i -th data point.

\hat{y}_i : Predicted sales for the i -th data point.

\bar{y} : Average of all actual sales.

n : Number of data points.



- Random Forest has the highest R^2 score (0.9535). This means it explains about 95% of the changes in Sales using the ad costs.
- Linear Regression has an R^2 score of 0.9059. It explains about 91% of the changes in Sales.
- Decision Tree has the lowest R^2 score (0.8899). It explains about 89% of the changes in Sales. It's the weakest of the three, but still pretty good since it's close to 90%.
- The R^2 scores are all high, meaning all models do a good job explaining how ad costs lead to Sales.

Overall Best Model



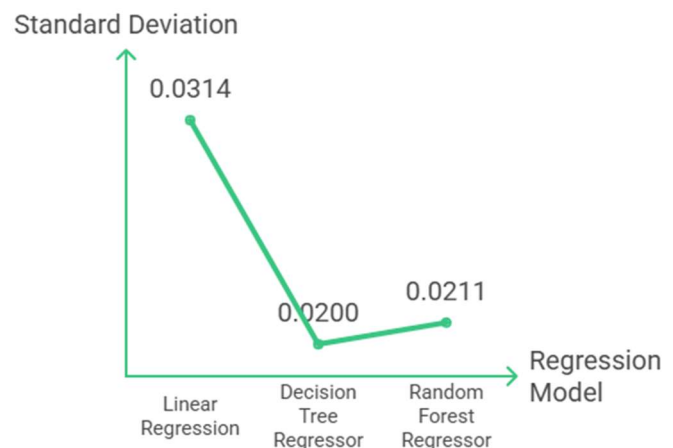
- It has the lowest MAE (0.92), its predictions are the closest to real Sales on average.
- It has the lowest MSE (1.44), it makes the smallest mistakes, even when bigger errors are counted more.
- It has the highest R^2 (0.9535), it explains 95% of the changes in Sales, better than the others at understanding how TV, Radio and Newspaper ad costs affect Sales.

Random Forest is the best choice for businesses to predict Sales from ad spending. It can help them plan their ad budgets (TV, Radio, Newspaper) more accurately, leading to better sales and profits giving the most trustworthy results.

Cross-Validation

Cross-validation is a way to test how well a model works by splitting the data into smaller parts (called folds). It helps businesses see how well these models will work on new data, not just the data they were trained on. I've used 5-fold cross-validation, the data is split into 5 parts. The model trains on 4 parts and tests on the 1 part left out, then repeats this 5 times (each part gets a turn being the test set). This gives a more reliable idea of how the model will perform on new data.

- Mean R^2 Score: How well each model explains Sales changes (higher is better, 1 is perfect).
- Std Dev: How much the R^2 scores change across folds (lower is better, means more consistency)

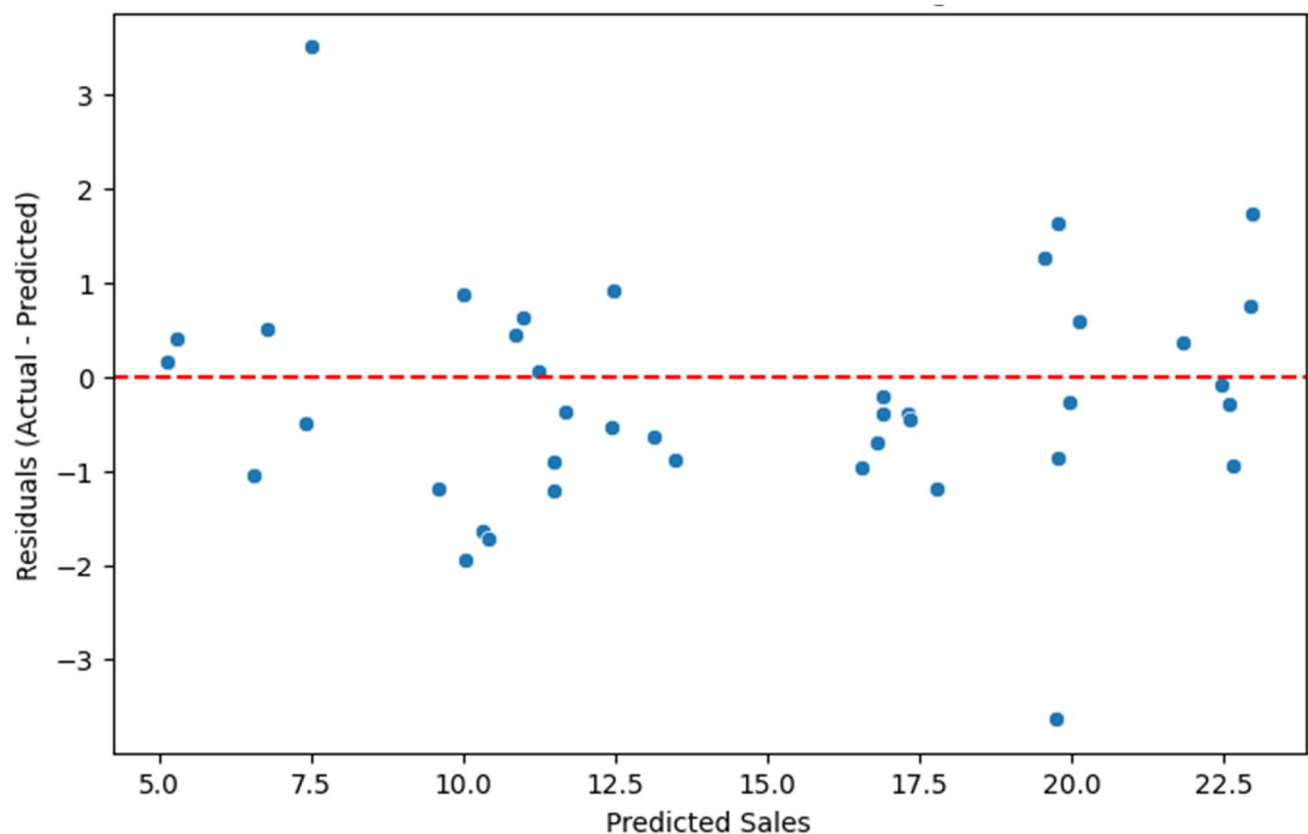


- Random Forest is the Best Overall: With the highest Mean R^2 (0.9417), Random Forest is the best at explaining how ad costs affect Sales. Its Std Dev (0.0211) shows it's also pretty consistent, so businesses can trust it on new data.
- Decision Tree is Reliable: Even though its Mean R^2 (0.9058) isn't the highest, its low Std Dev (0.0200) means it performs steadily across different data parts. It's a good choice if consistency matters more than top performance.
- Linear Regression is Okay but Less Stable: With a Mean R^2 of 0.8954, it's the weakest at explaining Sales changes, and its higher Std Dev (0.0314) means its performance varies more, making it less reliable on new data.

4. Visualization

Residual Analysis

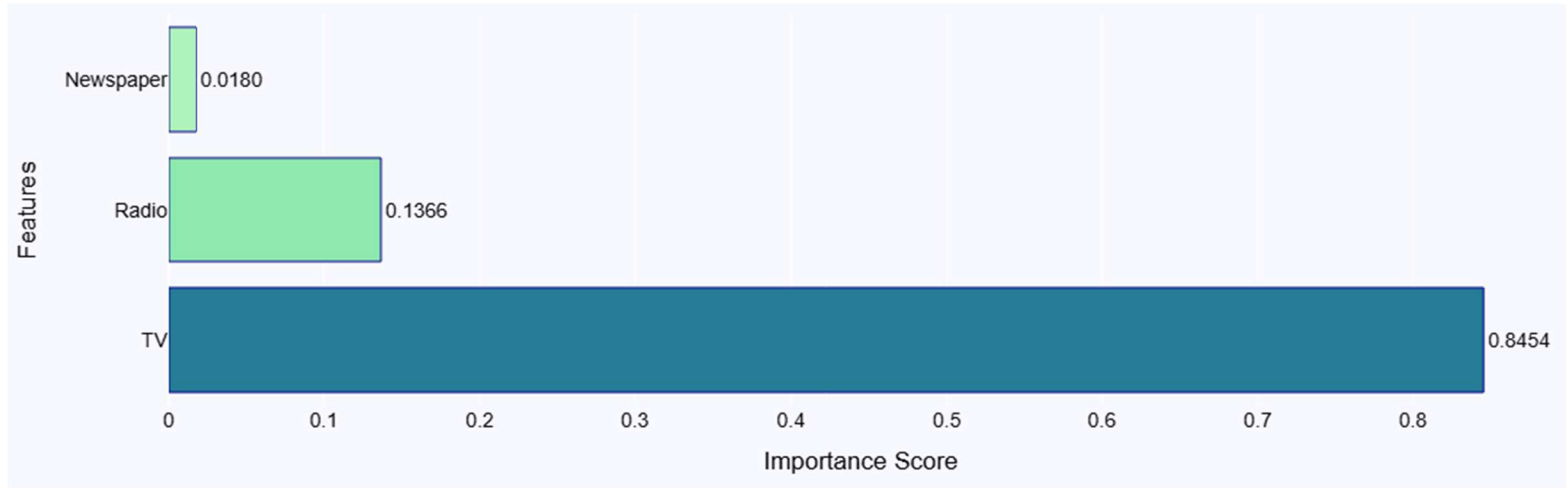
What's a Residual Plot?: It shows the mistakes (called residuals) the model makes when predicting Sales. A residual is the difference between the actual Sales and the predicted Sales. X-Axis: Predicted Sales (what the model guesses, in units). Y-Axis: Residuals (Actual Sales - Predicted Sales). If the residual is 0, the prediction was perfect. Positive residuals mean the model underpredicted (actual Sales were higher), and negative residuals mean it overpredicted (actual Sales were lower). Red Line: The line at 0 on the Y-axis. If all dots were on this line, the model's predictions would be perfect. Dots: Each dot is a data point showing the residual for a predicted Sales value.



Most dots are close to the red line , meaning the model's predictions are close to the actual Sales. The dots go from about -3 to +3 on the Y-axis, means the model's predictions are off by up to 3 units. The dots look scattered randomly around the red line and don't form a clear shape, means the model isn't making the same kind of mistake over and over. There are about the same number of dots above the red line (positive residuals, underpredicted) as below it (negative residuals, overpredicted). This means the model isn't consistently guessing too high or too low it's balanced.


Feature Importance

What's Feature Importance?: In a Random Forest model, feature importance tells us how much each feature helps the model make good predictions. Higher scores mean the feature is more important for predicting Sales. **Y-Axis:** The features (TV, Radio, Newspaper). **X-Axis:** Importance Score (from 0 to 1). The total of all scores adds up to 1. **Bars:** Each bar shows the importance score for a feature.



TV has the highest importance score, means TV ad costs are the biggest factor in predicting Sales. The model relies on TV the most to make accurate guesses. So, Businesses should focus on TV ads to boost sales as the model says it's the biggest factor. The importance scores match what we saw in earlier analyses (like correlation and hexbin plots) confirming that TV is the key to predicting Sales, while Radio and Newspaper play smaller roles.

5. Dashboard

 **Sales Prediction**


TV Advertising (\$)

Radio Advertising (\$)

Newspaper Advertising (\$)

Predict Sales

Dashboard Example


 **Sales Prediction**

230.1

37.8

69.2

Predict Sales

 **Predicted Sales: 22.02 units**

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1

