DATATHON 1.0

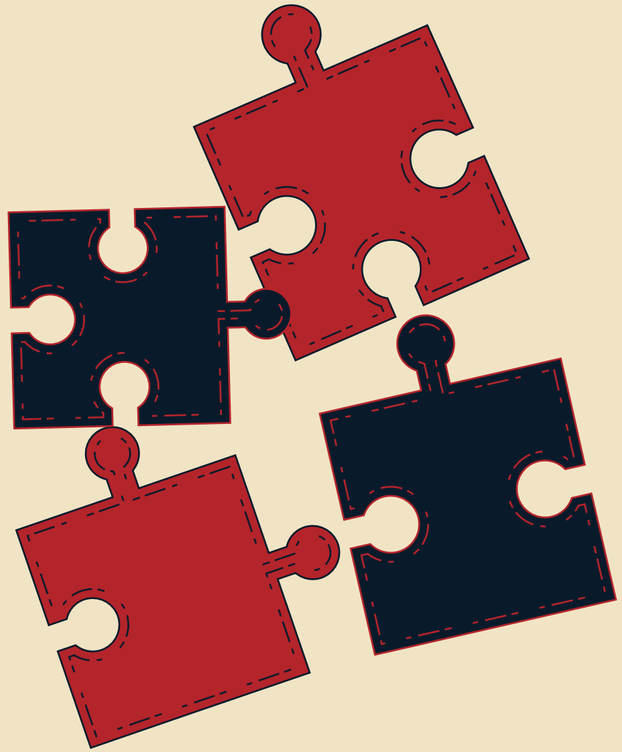
GRYFFINDOR _CATIANS

TABLE OF CONTENTS

DATASET

OVERVIEW

- Data was provided as 4 files: train_subset_1 → train_subset_4
- Subsets had different shapes → harder to merge
- Column names were not consistent across subsets
- Some features were split/merged differently in each file
- Some columns were unlabeled



CHALLENGE unify all subsets into a single dataset

DATA CLEANING

TRANSFORMATION

1. Country names: corrected misspellings using fuzzy matching
2. Unified formats across dataset: units, and currencies.
3. Extracted new columns from existing ones to enhance analysis
4. Standardized data types for consistency

OUTLIERS

1. Reviewed extreme values to check if they were reasonable.
2. Kept valid outliers (e.g., large stores) and handled illogical ones (e.g., negative weights).

MISSING VALUES

Used different imputation methods depending on the case:

- Median for numeric columns
- Mode for categorical columns
- Group-based imputation to keep distributions realistic

EDA

INSIGHTS

01

RELATIONSHIPS OVERVIEW

- Certain features show strong interactions (e.g., promotion, brand, product, department).
- Store area (binned) and income categories also show clear patterns.
- Gender, marital status, review score, recyclable status, and cities show very weak relationships

02

DRIVERS OF COST

- Categorical features (promotion name, brand, product, department), Store area (binned into categories), Income categories, work, and number of children → strong relationships
- Gender, marital status, review score, recyclable status, Cities → very weak relationships

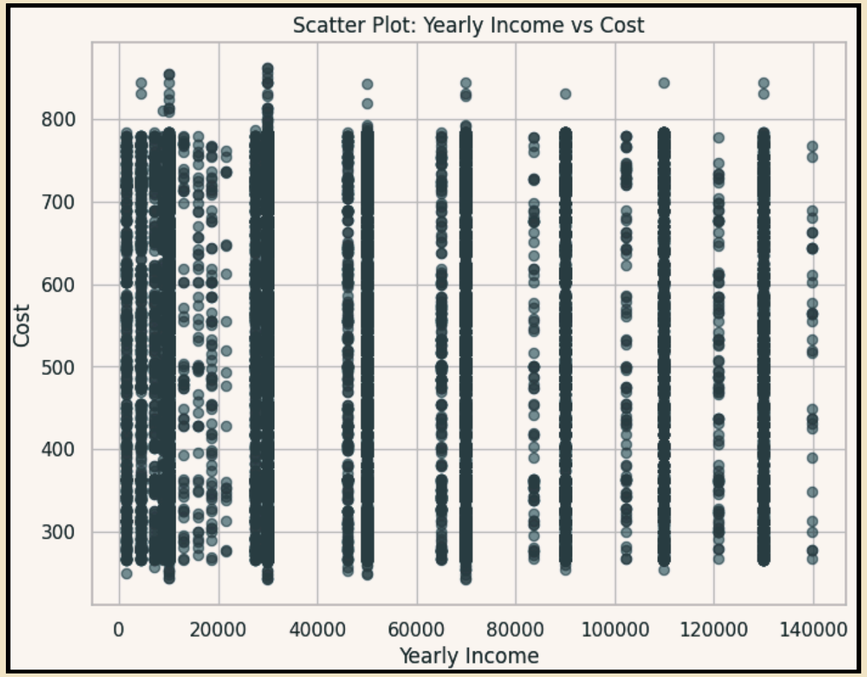
03

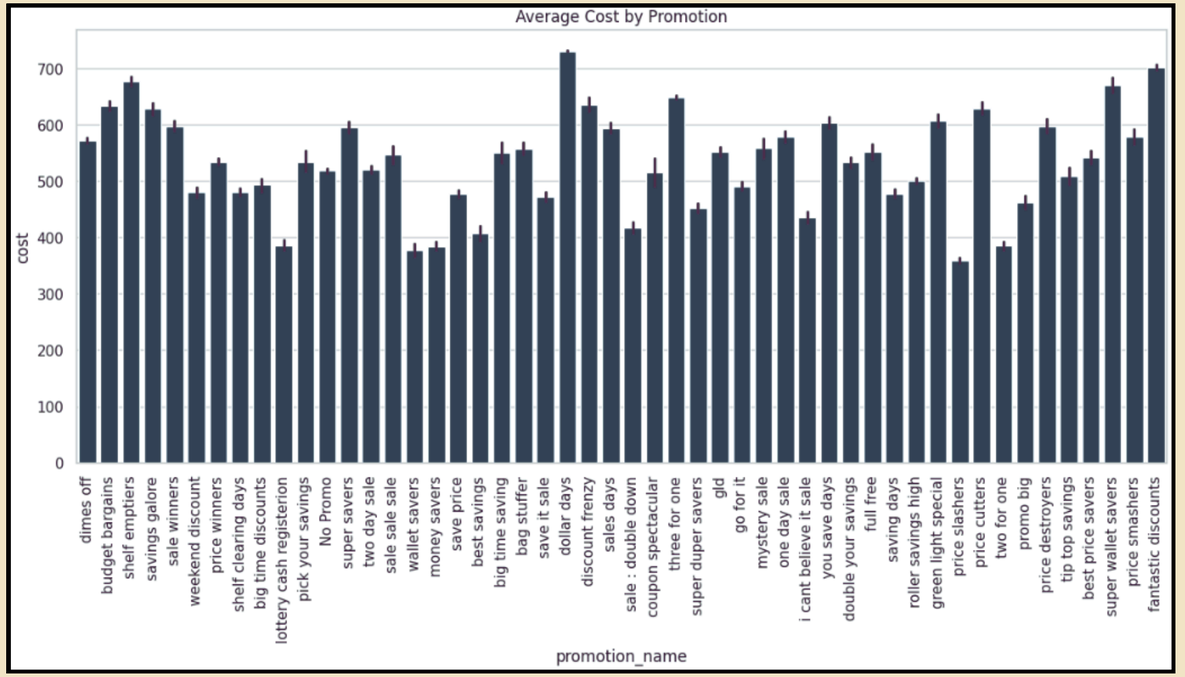
KEY ACTIONS

- Kept promotion, product, brand, department, store area, income category
- Dropped irrelevant columns (is_recyclable, gender, status, review_score, distance_km, cities)

E D A

INSIGHTS





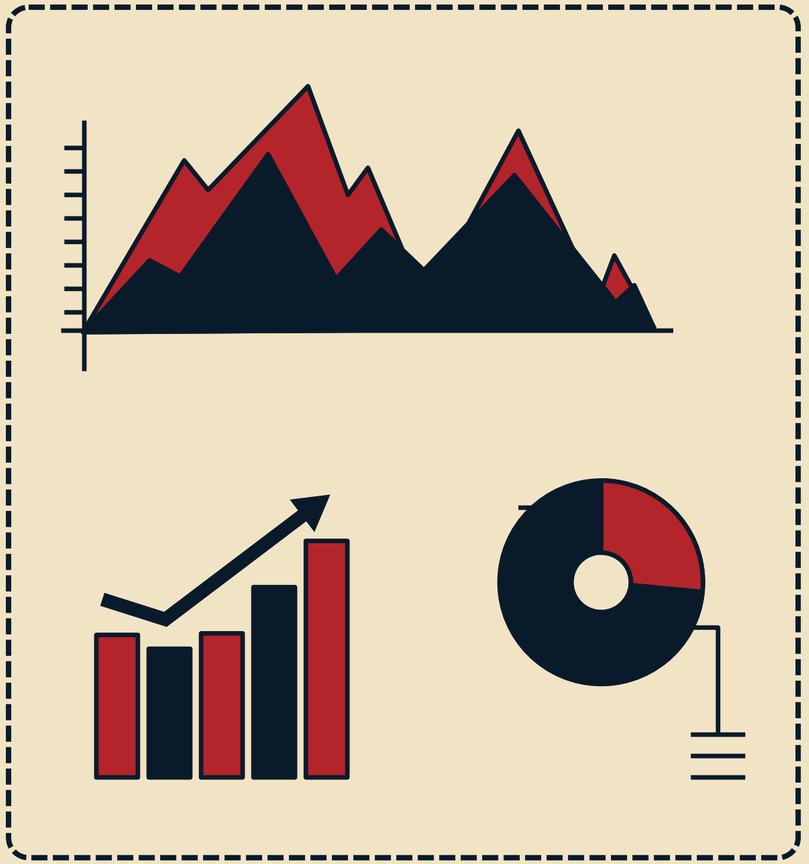
POWERBI

DASHBOARD

- Overview of clients, products, and brands
- Visual insights from geography and sales
- Key business takeaways

KEY TAKEAWAY

From messy raw data → business insights



D A T A

PREPROCESSING

- Applied One-Hot Encoding for small/clear categories e.g., children, education, work, store_kind
- Applied Frequency Encoding for high-cardinality columns e.g., brand, promotion_name, product

ENCODING CATEGORICAL FEATURES

- Scaling not required (Random Forest is scale-invariant)

HANDLING NUMERIC FEATURES

- Ensured consistent feature set between train & test
- Dropped unused raw categorical columns after encoding

DATASET ALIGNMENT

R F

MODELING

1. A Random Forest Regressor was built to predict product cost from the cleaned dataset.
2. RandomizedSearchCV was used to tune hyperparameters
3. Model performance was evaluated with 5-fold cross-validation.
4. Cross-validation results show an average RMSE of 78.167
.
5. This approach establishes a reliable pipeline from raw, messy data to actionable predictions.

FUTURE WORK

1. Handling missing values differently in some columns.
2. More feature engineering & feature selection
3. Try different encodings for categorical features
4. Hyperparameter tuning other than Randomized search
5. Add model explainability (SHAP, LIME)

TIMELINE

TEAMWORK

DATA CLEANING

- Unified columns & formats
- Fixed missing values & outliers

EDA & DASHBOARD

- Identified key cost drivers
- Dropped irrelevant features
- Built initial Power BI reports

PREPROCESSING

- Encoding (One-Hot & Frequency)
- Train/test alignment
- Skewed columns handled

MODELING

- Tried multiple models
- Selected best-performing model after tuning

DATATHON 1.0

EXPLORE

NOTEBOOK

For full code, preprocessing steps, and detailed analysis, please check the notebook:

Click Here

OPEN

DISCUSSION

Thank you for your attention