Diamond Price Analysis and Prediction using Big Data Analytics Lifecycle

Team Members

- **Nouran Elsayed** **23P0006**
- **Reetaj Ahmed** **23P0114**
- **Abdelmoneim Mahmoud** **23P0015**
- **Hossam Osama** **23P0010**
- **Jana Ehab** **23P0105**
- **Hasan Sherif** **23P0017**

DISCOVERY → DATA PREPARATION → MODEL PLANNING → MODEL BUILDING → COMMUNICATE RESULTS → OPERATIONALIZE

Dataset Overview

- Dataset: Diamonds Prices 2022
- Source: Real-world diamond pricing dataset
- Size: Thousands of records
- Attributes:

price, carat, cut, color, clarity, depth, table, length, width, height
(10+ attributes ✓)

Data Cleaning Steps

- Checked missing values → **No significant missing data**
- Removed **outliers** using **IQR method**
- Removed invalid values:
  - carat > 0
  - price > 0
  - depth ∈ [40, 80]
  - dimensions > 0
- Removed duplicate rows

Numerical Hypothesis Tests (Pearson Correlation)

- H_0: No relationship between feature and price
- **H_1: Significant relationship exists**

Feature	Result
Carat	→ Strong positive correlation
Volume	→ Strong positive correlation
Depth	→ Weak correlation
Table	→ Weak correlation

Models Applied

- Linear Regression
- Support Vector Regression (SVR – RBF kernel)
- Decision Tree Regression

Performance Metrics

- R^2 (Accuracy)
- MAE
- RMSE

Model	R^2
Linear Regression	→ Good
Decision Tree	→ Moderate
SVR	→ Best Performance

Key Insights

- Carat and Volume are the strongest predictors of price
- Categorical attributes significantly influence pricing
- SVR is highly effective for nonlinear price prediction

Problem Statement

Diamond prices depend on multiple physical and categorical attributes.

The challenge is to **understand the key factors affecting price and build predictive models** that accurately estimate diamond prices.

Feature Engineering

- Created new feature:
- Volume = length × width × height
- Converted categorical variables to factors:
  - cut
  - color
  - clarity

Depth and Table excluded from prediction models

Categorical Hypothesis Tests (ANOVA):

- Cut → Significant effect on price
- Color → Significant effect on price
- Clarity → Significant effect on price

Features Used

carat, volume, cut, color, clarity

Train/Test Split

80% Training
20% Testing

Key Result

SVR achieved the highest R^2, indicating superior prediction accuracy

Model	R^2 (Accuracy)
Linear Regression	~0.92
Decision Tree Regression	~0.90
Support Vector Regression (SVR)	**~0.97**

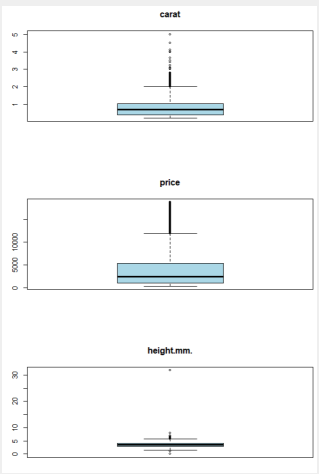
Possible Applications

- Diamond price estimation systems
- Jewelry market analysis
- Decision support for buyers and sellers

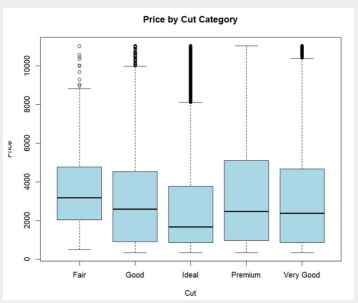
Project Objectives

- Analyze diamond price behavior using statistical and visual analysis
- Clean and preprocess raw data
- Apply hypothesis testing to validate relationships
- Build and compare predictive models
- Identify the most influential features affecting price

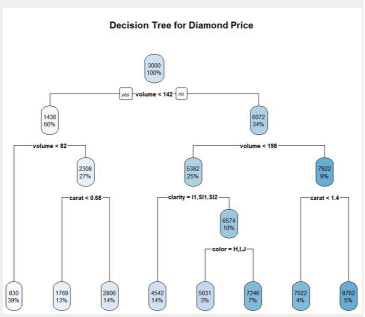
Outlier Detection (RAW Data)



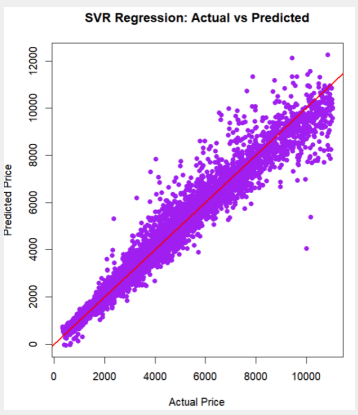
Visualizations



Visualizations



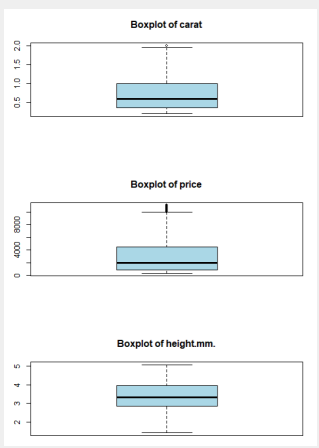
Visualizations



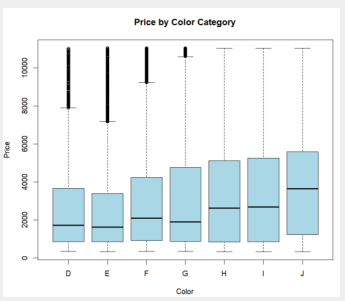
Future Work

- Apply ensemble learning techniques (e.g., Random Forest)
- Explore neural network–based regression models
- Extend the dataset with larger and more diverse market data

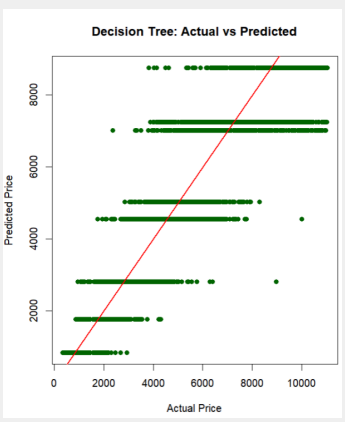
Outlier Detection (Cleaned Data)



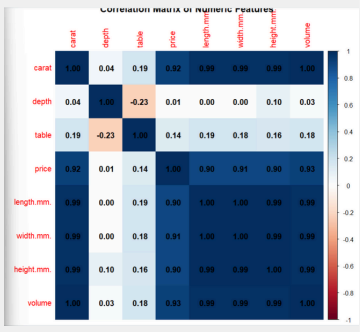
Visualizations



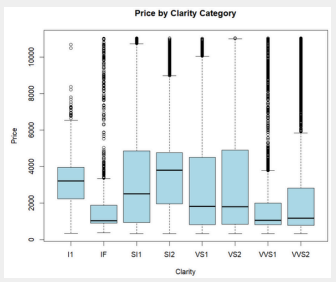
Visualizations



Exploratory Data Analysis (EDA)



Visualizations



Visualizations