Homework 4 Critique Component of Diamond Sales

Stat436

Peiyuan Li

Word Count: 398

Project Introduction

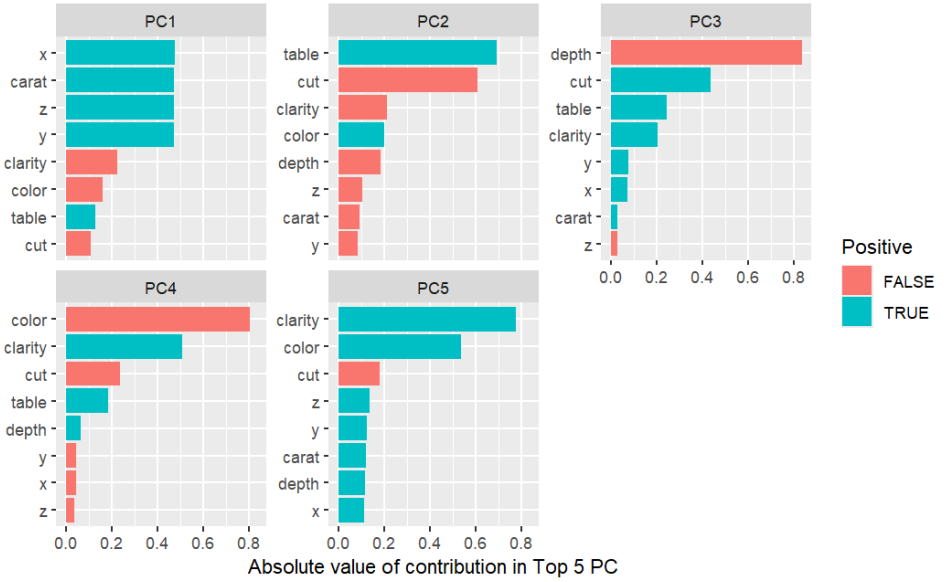
"A diamond is forever" – diamonds are timeless treasures whose value often transcends simple appraisal. Understanding the multifaceted determinants of diamond prices has long been an art within the trading sphere. In this project, I leverage a Kaggle dataset on diamond sales to predict pricing using minimal data points, aiming to distill complex valuation into comprehensible models.

Data Processing

The dataset contains categorical labels such as cut, color, and clarity—integral attributes in the jewelry industry that cannot be quantified without context. I numerically encoded these qualities to align with industry standards—cut quality (from Fair to Ideal), diamond color (from J to D), and clarity (from I1 to IF)—transforming subjective assessments into a quantitative scale suitable for Principal Component Analysis (PCA), a prerequisite for the modeling technique employed.

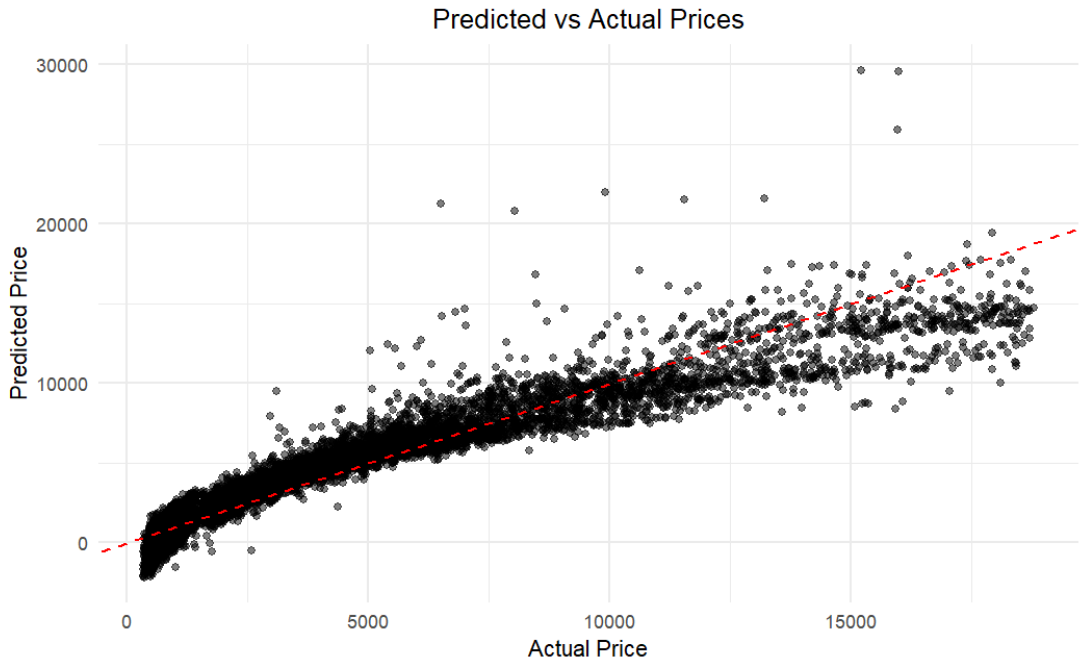
Visualization Design and Implementation

My analysis unfolds through two pivotal visualizations. The first exposes the principal elements that affect pricing, excluding the price itself. Post-PCA, I visualized the coefficients of the top five principal components and arranged them by their absolute values. Notably, the dimensions x, y, z, and carat shared strikingly similar contributions in PC1, hinting at a correlation among these features. PC2 and PC3 highlighted significant loadings for table and cut, and depth and cut, respectively, suggesting potential interactions between these attributes. The strong loadings for color and clarity in PC4 and PC5 indicated their importance and possible interrelation.



Selecting the strongest contributor from each PC—carat, table, depth, color, and clarity—I then constructed a linear model to gauge their collective predictive power over diamond prices. The result is a strikingly accurate representation of price dispersion when applied to a split dataset of training and test groups. The model's efficacy suggests that these five attributes not only encapsulate a comprehensive

understanding of pricing but can also surrogate for the dataset's broader



Therefore, my visualization address the core question: "Can a minimal subset of features derived from PCA represent the intricate pricing structure of diamonds?"

Conclusion

The project underscores the feasibility of distilling complex valuation into manageable models that effectively capture essential pricing factors in the diamond industry. The exploratory journey from raw data to informative visualizations has demonstrated PCA's strength in uncovering the latent structure within high-dimensional data, affirming the chosen methodology's relevance and effectiveness.

In coding, the data is hosted on github and can be loaded directly, comments are also detaily attached.