

# Dynamic Price Optimization

-Vinod Ghanchi  
-Ashay Katre  
-Yashvi Mehta



# Table of contents

**01**

**Objective and value proposition**

**02**

**Dataset Used**

**03**

**Learnings from EDA**

**04**

**Modeling &  
Hypothesis Test  
results**

**05**

**Conclusion and  
insights**

**06**

**Challenges &  
Future Scope**

# Objective and value proposition

- The primary objective of this project is to develop an advanced predictive model for accurately estimating the selling price of products listed on Mercari.
- By analyzing various product features such as item condition, brand, category, and description, the project aims to create a model that enables sellers to determine optimal pricing strategies



# Dataset Used

- The project utilizes the Mercari Price Suggestion dataset, which is publicly available on Kaggle.

train_id		name	item_condition_id	category_name	brand_name	price	shipping	item_description
0	0	MLB Cincinnati Reds T Shirt Size XL	3	Men/Tops/T-shirts	NaN	10.0	1	No description yet
1	1	Razer BlackWidow Chroma Keyboard	3	Electronics/Computers & Tablets/Components & P...	Razer	52.0	0	This keyboard is in great condition and works ...
2	2	AVA-VIV Blouse	1	Women/Tops & Blouses/Blouse	Target	10.0	1	Adorable top with a hint of lace and a key hol...
3	3	Leather Horse Statues	1	Home/Home Décor/Home Décor Accents	NaN	35.0	1	New with tags. Leather horses. Retail for [rm]...
4	4	24K GOLD plated rose	1	Women/Jewelry/Necklaces	NaN	44.0	0	Complete with certificate of authenticity

# Preprocessing

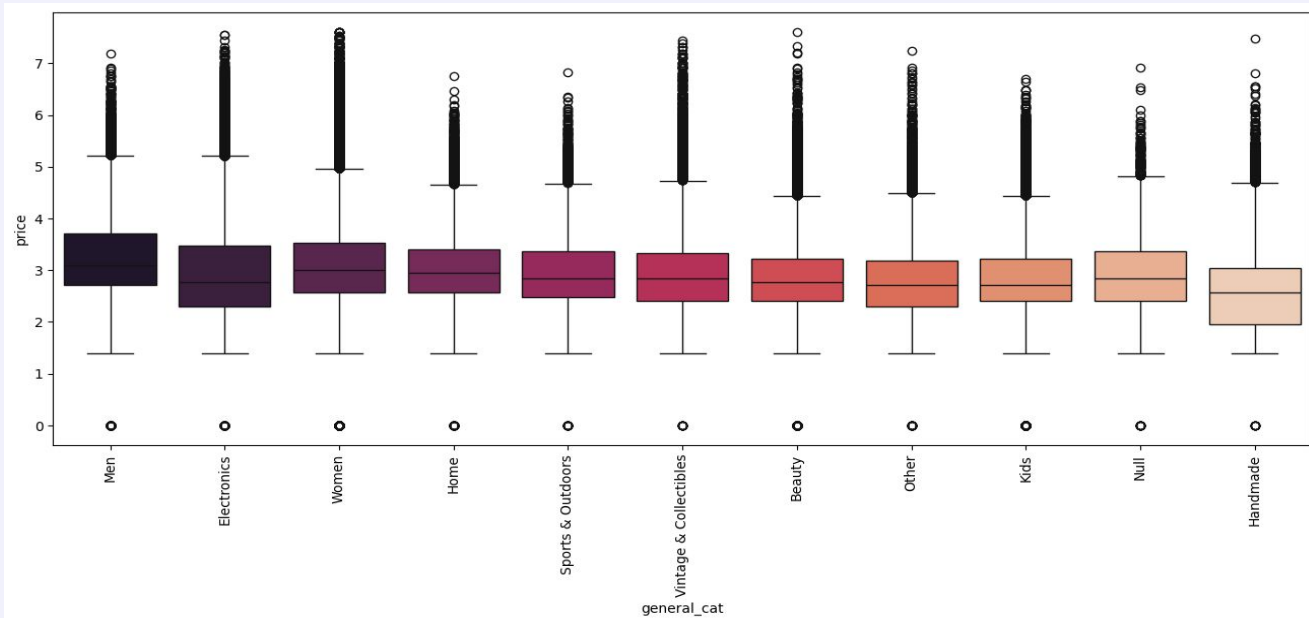
- Text features like brand\_name and item\_description were cleaned using basic NLP techniques, including tokenization, stopword removal, and stemming. Numerical features, such as price, were log-transformed to address skewness, and all features were standardized for consistency across models.
- The dataset was preprocessed by splitting hierarchical categories into three levels: general category, sub-category 1, and sub-category 2.

category_name	clean_category_name
Men/Tops/T-shirts	men top t-shirts
Electronics/Computers & Tablets/Components & P...	electronic computer & tablet component & parts
Women/Tops & Blouses/Blouse	women top & blouse blouse
Home/Home Décor/Home Décor Accents	home home décor home décor accents
Women/Jewelry/Necklaces	women jewelry necklaces

general_cat	sub_cat1	sub_cat2
Men	Tops	T-shirts
Electronics	Computers & Tablets	Components & Parts
Women	Tops & Blouses	Blouse
Home	Home Décor	Home Décor Accents
Women	Jewelry	Necklaces

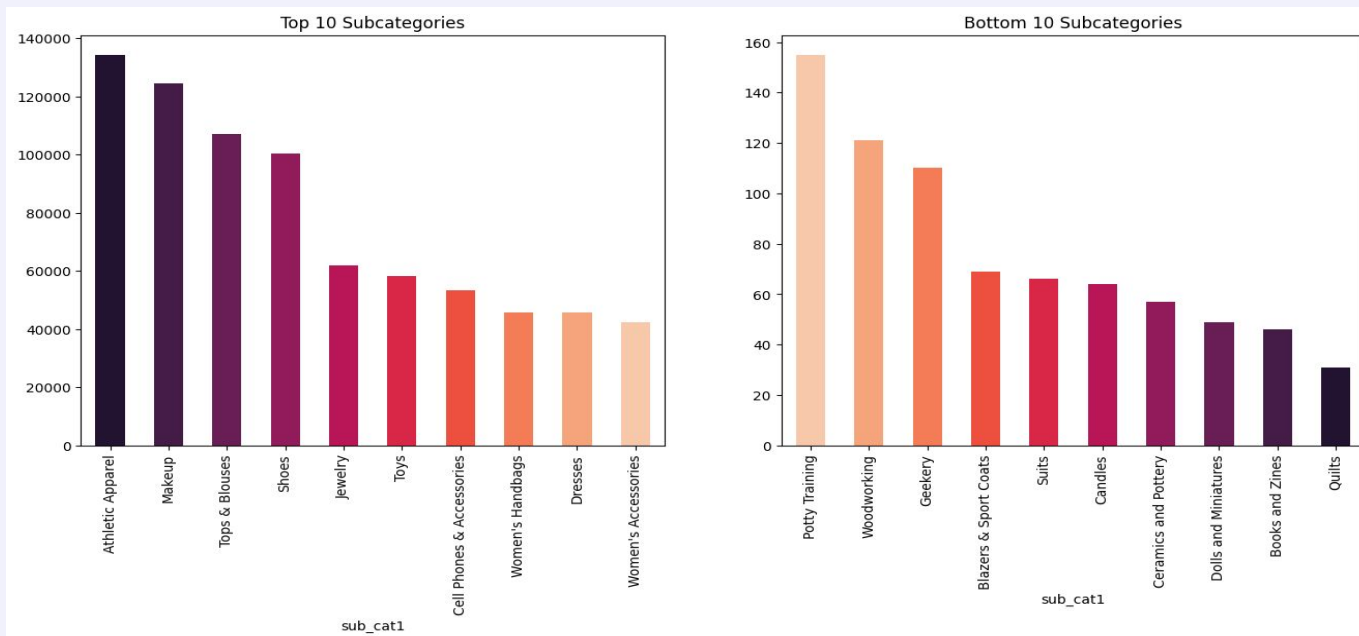
# Major Learnings From EDA

- The box plot shows that categories like Electronics and Vintage & Collectibles have a wide price range, while Beauty and Handmade exhibit more consistent pricing with fewer outliers.



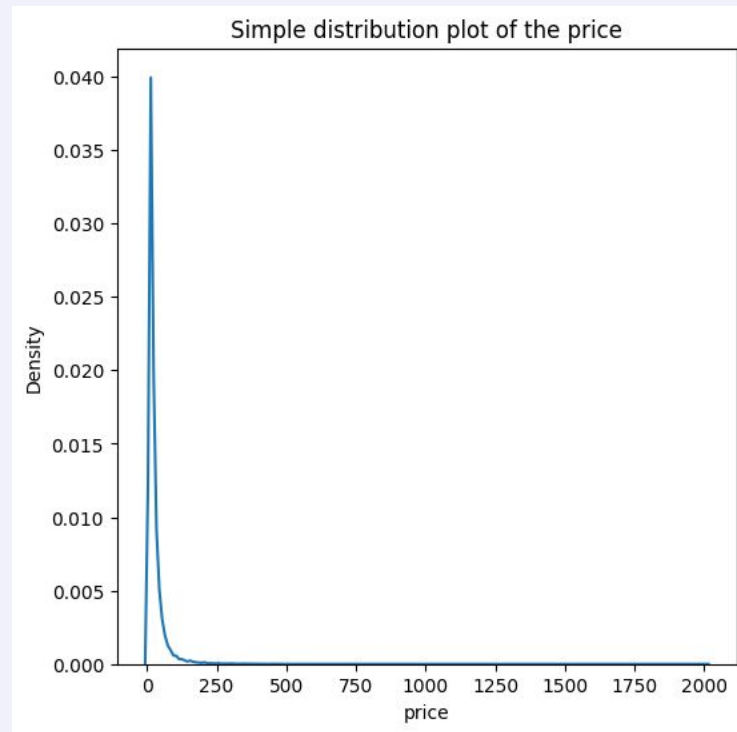
# Major Learnings From EDA

- The bar plots reveal that the Top 10 sub-categories are dominated by a few popular categories like athletic apparel, makeup, tops, etc. while the Bottom 10 sub-categories show relatively lower product counts, indicating niche or less popular product types.



# Major Learnings From EDA

- The distribution plot shows that most items are priced on the lower end, indicating a skew towards affordable products in the dataset.





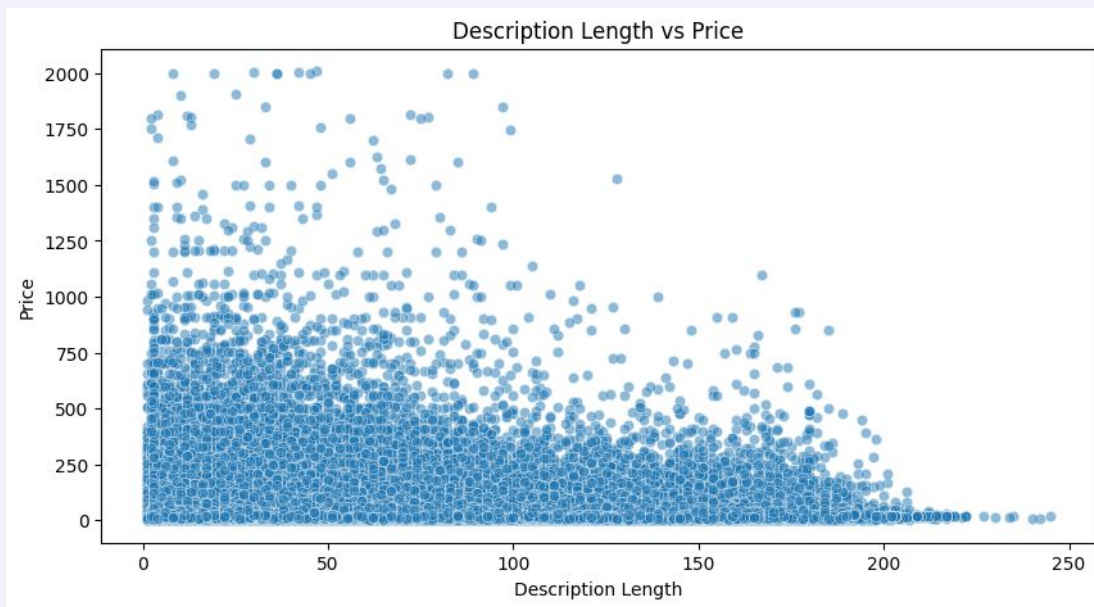
# Major Learnings From EDA



- The box plot reveals that buyers tend to pay shipping fees more frequently for higher-priced items.
- Additionally, the presence of outliers suggests that some expensive items have significantly higher prices compared to the majority, which flattens the distribution of the graph.

# Major Learnings From EDA

- Upon a closer look, the scatter plot seems to show that as the description length increases, the price also tends to increase, though the relationship isn't strictly linear.
- There is a general trend where longer descriptions seem to correspond to higher prices, especially as you move to the right side of the plot (with higher description lengths). This could indicate that products with longer descriptions are generally more expensive, but the data also has significant spread, so there might be other factors influencing the price as well.



# Hypothesis Testing

- To test the difference in the means of prices based on shipping responsibility, we performed an independent two-sample t-test.



## Null Hypothesis ( $H_0$ )

There is no significant difference in the mean prices between items where the buyer pays shipping and items where the seller pays shipping.



## Alternate Hypothesis ( $H_a$ )

There is a significant difference in the mean prices between items where the buyer pays shipping and items where the seller pays shipping.

T-statistic: 119.77499916636837

P-value: 0.0

Reject the Null Hypothesis: There is a significant difference in the mean prices based on shipping responsibility.

# Models Used

## Linear & Ridge Regression

- These baseline models provided an initial benchmark for performance.
- Ridge regression, in particular, was used to mitigate multicollinearity and hyperparameter tuning was conducted to improve model performance.

## Decision Tree

- This model allowed capturing non-linear relationships within the dataset. Hyperparameters like maximum depth and minimum samples per split were fine-tuned to optimize performance.


## LightGBM

- This model outperformed the others in terms of speed and efficiency, particularly for the large dataset.
- LightGBM uses gradient boosting with leaf-wise splitting and handles large datasets with missing values efficiently, making it significantly faster without compromising predictive accuracy compared to the other models.


# Model Results

	MAE	MSE	R <sup>2</sup> Score
Linear Regression	0.56	0.52	0.07
Ridge Regression	0.56	0.52	0.07
Decision Tree Regressor	0.56	0.59	-0.05
Decision Tree with Hyperparameter Tuning	0.50	0.42	0.24
LightGBM	0.46	0.37	0.34


# Conclusion And Insights



In this project, we aimed to optimize pricing prediction for products listed on the Mercari marketplace. Through extensive exploratory data analysis (EDA), we identified key features and relationships that influenced pricing.



After evaluating simpler models like Linear Regression, Ridge, and Decision Tree Regressor, we observed significant improvements through hyperparameter tuning. The Decision Tree model's performance was enhanced, reducing error metrics like MAE and MSE.



LightGBM outperformed all other models, achieving a notable increase in  $R^2$ , demonstrating its ability to capture complex patterns in the data. This work highlights the value of EDA in feature selection and the superior performance of advanced models like LightGBM for large datasets and non-linear relationships.

# Challenges Faced

- **Category Data Complexity:** The category\_name field had nested structures (e.g., Electronics/Computers/Tablets), requiring splitting into subcategories and cleaning to ensure meaningful input for modeling.
- **Handling Large Data:** The dataset size caused computational bottlenecks, making model training and hyperparameter tuning resource-intensive and limiting optimization efforts.





# Challenges Faced

- **Data Cleaning:** Issues like missing brand\_name, inconsistent prices, and outliers required careful handling to ensure data quality without degrading model performance.
- **Hyperparameter Tuning Challenges:** Tuning improved performance but was computationally expensive, requiring simplified approaches to balance accuracy and efficiency.









# Future Work



**Extensive Testing on Unseen Data:** Additional testing on unseen data, including cross-validation and stress-testing, is needed to ensure robustness and generalization.



**Cross-Domain Analysis:** Exploring external factors like market trends or seasonal variations could enhance predictive accuracy.



**Advanced NLP Techniques:** Using advanced NLP models (e.g., BERT or GPT) could improve predictions by capturing nuances in text-based features.

**Thank You!**