# Data-Driven Pricing Model for Refurbished Smartphones

**Project: Foundations - ReCell**

Sanjana Addanki
10/23/2025

AGENDA

- Executive Summary
- Business Problem & Solution Approach
- Data Overview
- Exploratory Data Analysis (EDA)
- Data Preprocessing
- Model Performance Summary

# Business Problem Overview

- ReCell is struggling with setting accurate resale prices for these refurbished phones.
  - Ultimately leads to lost revenue or unsold inventory
- Prices vary by:
  - Brand
  - Condition
  - Features
- Manual pricing inconsistent, so need a dynamic pricing strategy

# Solution Approach

1. Exploratory Data Analysis.
    a. Will show any data issues and relationships

2. Data Preprocessing.

3. Build and Refine Linear Regression Models.
    a. With VIF checks and p-value driven selection

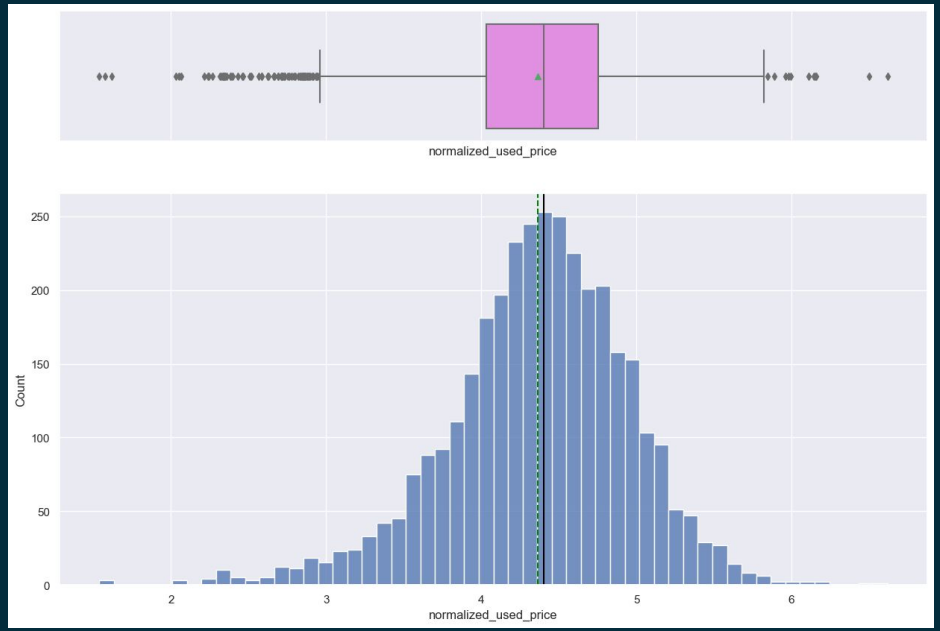
4. Evaluate model on test/train splits.
    a. Validate assumptions with residuals, heteroscedasticity, multicollinearity.

# Data Overview

- There are 3454 rows and 15 columns.
- Key features:
  - brand_name, os, screen_size, 4g, 5g, main_camera_mp, selfie_camera_mp, int_memory, ram, battery, weight, release_year, days_used, normalized_new_price, normalized_used_price
- Missing values in columns:
  - main_camera_mp: 179
  - selfie_cameria_mp: 2
  - int_memory/ram: 4 each
  - battery: 6
  - weight: 7
- There are no duplicates.

# EDA - Univariate Observations

## Target Distribution: Normalized_used_price




Boxplot:
Shows outliers on both sides, and shows a balanced unimodal distribution with central tendency.

Histogram:
The distribution is approximately symmetric, with a possible slight left skew. The mean and median close to identical, and the spread is centered. There are some outliers on both tails.

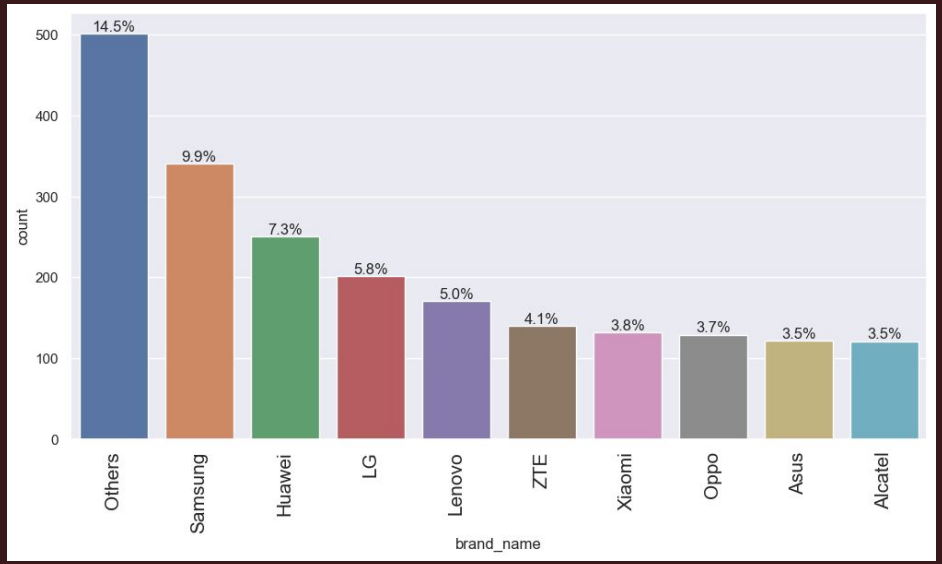
# EDA

## Patterns in Correlation Heatmap


- Normalized_new_price has the strongest positive correlation with normalized_used_price
- Camera specs, RAM, and screen time all positively correlated.
- Years_since_release negatively correlated: older → cheaper

# EDA

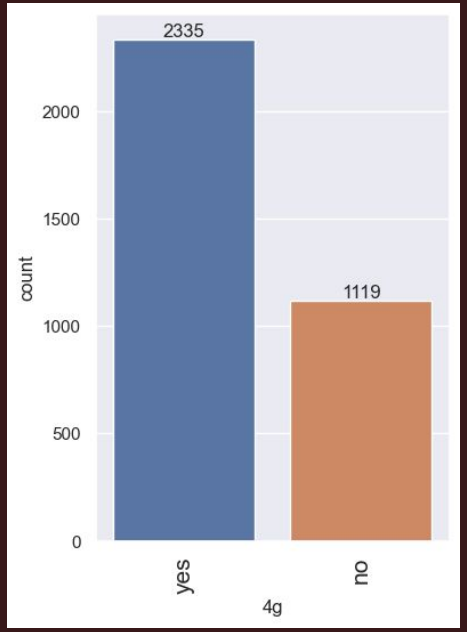
## Barplots for categorical variables
## brand_name

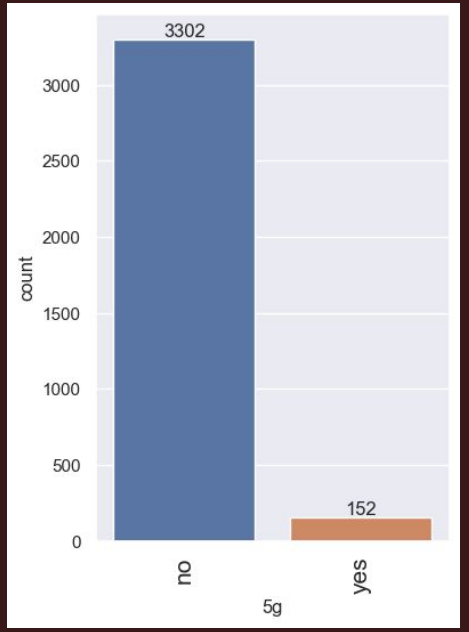



There is a rather moderate imbalance across brands, as others is overrepresented.

# EDA - Bivariate Analysis

## 4g vs 5g




From these graphs, we can see that the majority of phones were 4g versus 5g.

# Data Preprocessing

Missing-value imputation:

- Median within groups

Encoding:

- Dummy variables with drop_first = True

Outlier detection:

- Boxplots were used to flag extreme values in battery and weight. We chose not to remove the outliers as the model was strong enough.

Multicollinearity

- After calculating VIF, dropped all variables with VIF > 10.

Feature Selection

- Used iterative p-val elimination
  - Remove variable with highest p value > 0.05 until all are significant

★ Created years_since_release and dropped release_year to avoid redundancy.

# Final Selected Features

- screen_size
- main_camera_mp
- selfie_camera_mp
- int_memory
- ram
- battery
- weight
- normalized_new_price
- normalized_used_price
- years_since_release
- brand_name_Celkon
- brand_name_Nokia
- brand_name_Xiaomi
- os_Others
- 4g_yes
- 5g_yes

# Model Performance Summary

| Metric | Train | Test |
|---|---|---|
| R^2 | 0.847 | 0.833 |
| Adjusted R^2 | 0.846 | 0.831 |
| RMSE | 0.23 | 0.24 |
| MAE | 0.18 | 0.19 |
| MAPE | 4.32% | 4.51% |

# Final Model Insights

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | normalized_used_price | R-squared: | 0.847 |
| Model: | OLS | Adj. R-squared: | 0.846 |
| Method: | Least Squares | F-statistic: | 886.8 |
| Date: | Wed, 22 Oct 2025 | Prob (F-statistic): | 0.00 |
| Time: | 18:14:34 | Log-Likelihood: | 110.96 |
| No. Observations: | 2417 | AIC: | -189.9 |
| Df Residuals: | 2401 | BIC: | -97.27 |
| Df Model: | 15 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 1.3715 | 0.052 | 26.565 | 0.000 | 1.270 | 1.473 |
| screen_size | 0.0291 | 0.003 | 8.473 | 0.000 | 0.022 | 0.036 |
| main_camera_mp | 0.0234 | 0.001 | 16.151 | 0.000 | 0.021 | 0.026 |
| selfie_camera_mp | 0.0119 | 0.001 | 10.644 | 0.000 | 0.010 | 0.014 |
| int_memory | 0.0002 | 6.66e-05 | 2.836 | 0.005 | 5.83e-05 | 0.000 |
| ram | 0.0293 | 0.005 | 5.686 | 0.000 | 0.019 | 0.039 |
| battery | -1.46e-05 | 7.19e-06 | -2.030 | 0.043 | -2.87e-05 | -4.94e-07 |
| weight | 0.0008 | 0.000 | 6.200 | 0.000 | 0.001 | 0.001 |
| normalized_new_price | 0.4092 | 0.011 | 36.760 | 0.000 | 0.387 | 0.431 |
| years_since_release | -0.0218 | 0.004 | -5.847 | 0.000 | -0.029 | -0.014 |
| brand_name_Celkon | -0.1905 | 0.053 | -3.571 | 0.000 | -0.295 | -0.086 |

...

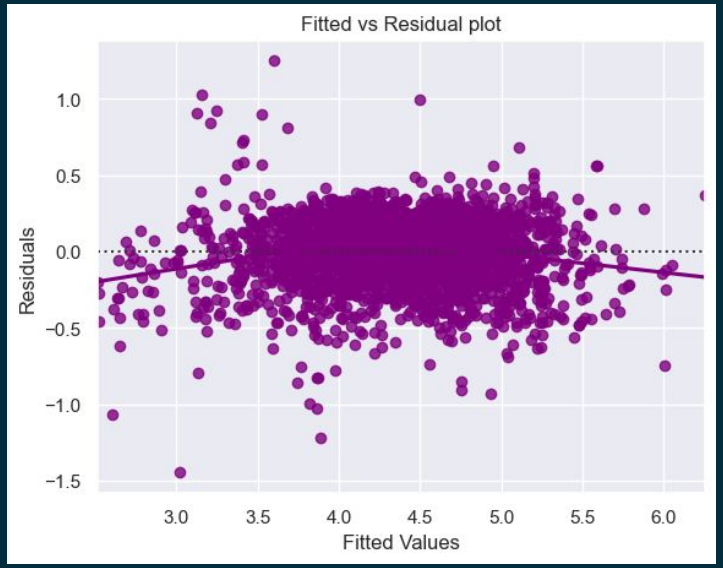
Notes:

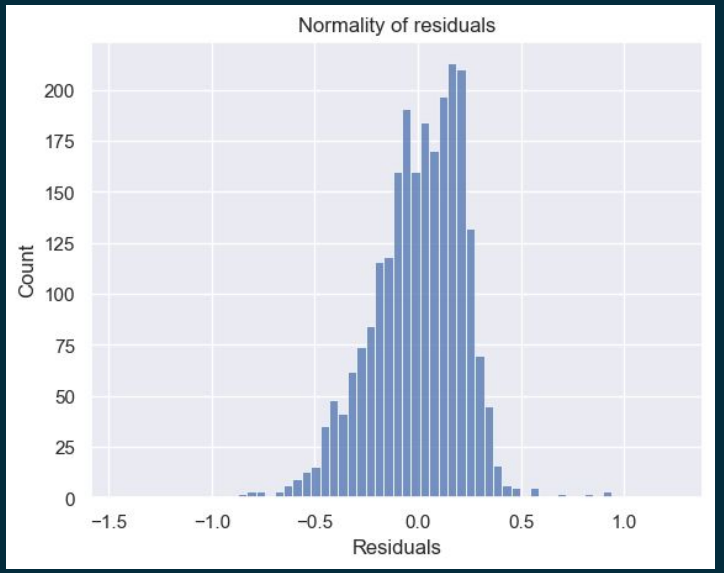
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 4.08e+04. This might indicate that there are strong multicollinearity or other numerical problems.

- normalized_new_price
  - Largest driver of used price (+0.41)
- RAM, screen_size, camera MP all have positive impacts on used price
- Battery
  - Small negative impact or could just be specific to this dataset
- Brand Name
  - Celkon → negative impact
  - Nokia, Xioami → positive impact
- 4g → positive impact
- 5g → small negative impact
  - Could be due to limited data

# Diagnostic Checks: Residuals




Residuals seem to be roughly symmetric with a rather minor deviation from normality, still very much acceptable

# Diagnostic Checks Continued

Shapiro-Wilks Test Result:

```
ShapiroResult(statistic=0.9634838700294495,
pvalue=2.995824519245496e-24)
```

➢ The p-value is less than 0.05, we reject the null hypothesis

➢ This model residuals are not normally distributed.

Homoscedasticity Result:

```
[('F statistic', 0.9366544370124994), ('p-value', 0.8706713505249848)]
```

➢ P-value is greater than 0.05, we fail to reject the null hypothesis.

➢ Variance of errors is roughly constant across the fitted values

Multicollinearity:

- We already controlled for this when we removed VIF values if they exceeded the threshold.

# Executive Summary

## Model Performance:

- Linear Regression (final model)
  - R^2: 0.847 (train) / 0.833 (test)
  - RMSE: 0.23-0.24
  - MAPE: 4.4%
- Diagnostics:
  - No evidence of heteroscedasticity
  - Not an apparent amount of multicollinearity
  - Residuals are non-normal

# Executive Summary

## Key Insights:

- Positive: normalized_new_price, RAM, screen_size, camera_mp, 4g
- Negative: years since release, Celkon brand, 5g (limited data effect)
- Nokia and Xiaomi add value

## Recommendations:

- We adopt the model for dynamic and consistent pricing
- Use the insights for inventory valuation and further marketing strategy
- We should explore nonlinear models to capture more complex effects

# Thank you