Ask anything

yo chat give me a list of laptops with these specifications bro no something within a reasonable price range ok but which ones better in terms of what i want

we’ve got
a *better*
solution
than
chatgpt

TECH GENIE
AT YOUR SERVICE

PC PART PICKER 3000

AN OVERVIEW

Ever wish you had a genie who could instantly tell you the price of your dream laptop and show you the best match in the market?

We deliver three core ML functions:

- **Descriptive:** K-Means Clustering to segment computers based on price, RAM, and other specs. PCA is a way of showing the clusters of the K-Means
- **Predictive:** LightGBM Regression to estimate **prices** from user inputs
- **Prescriptive:** K-Nearest Neighbors (KNN) to recommend similar listings with ranked similarity

DATA COLLECTION & PREP

CLEANED DATA

- **CSV file:** 8,064 marketplace listings (rows) x **135 raw Spanish-language columns**
- Encoded in **UTF-8-SIG with mixed metrics, units, and labels; .CSV file with 135 columns.**
- **No scraping or APIs; data ingested directly via pandas.read_csv.**

RAW DATA

1. **Dropped Duplicates →**
df.duplicated().sum()

2. **Standardized Column Names with custom slugify function**
   - → removed accents, lowercase, dropped stopwords (e.g., **Pantalla_Tamaño → pantalla_tamano)**

3. **Dropped Unnamed Columns:**
   - df.drop(columns=['unnamed_0'])
   - Full null or >70% null columns

4. Price Normalization

- Parsed "Precio_Rango" (e.g., "1.026,53 € – 2.287,17 €") into:
- `precio_min`, `precio_max`, and `precio_mean`
- **Dropped original string after parsing**

5. Numerical Extraction

- Created functions to **extract float from strings (e.g., RAM, CPU speed)**
- Remove thousands separators.
- Apply apply_cleaning_to_column() across many dirty fields

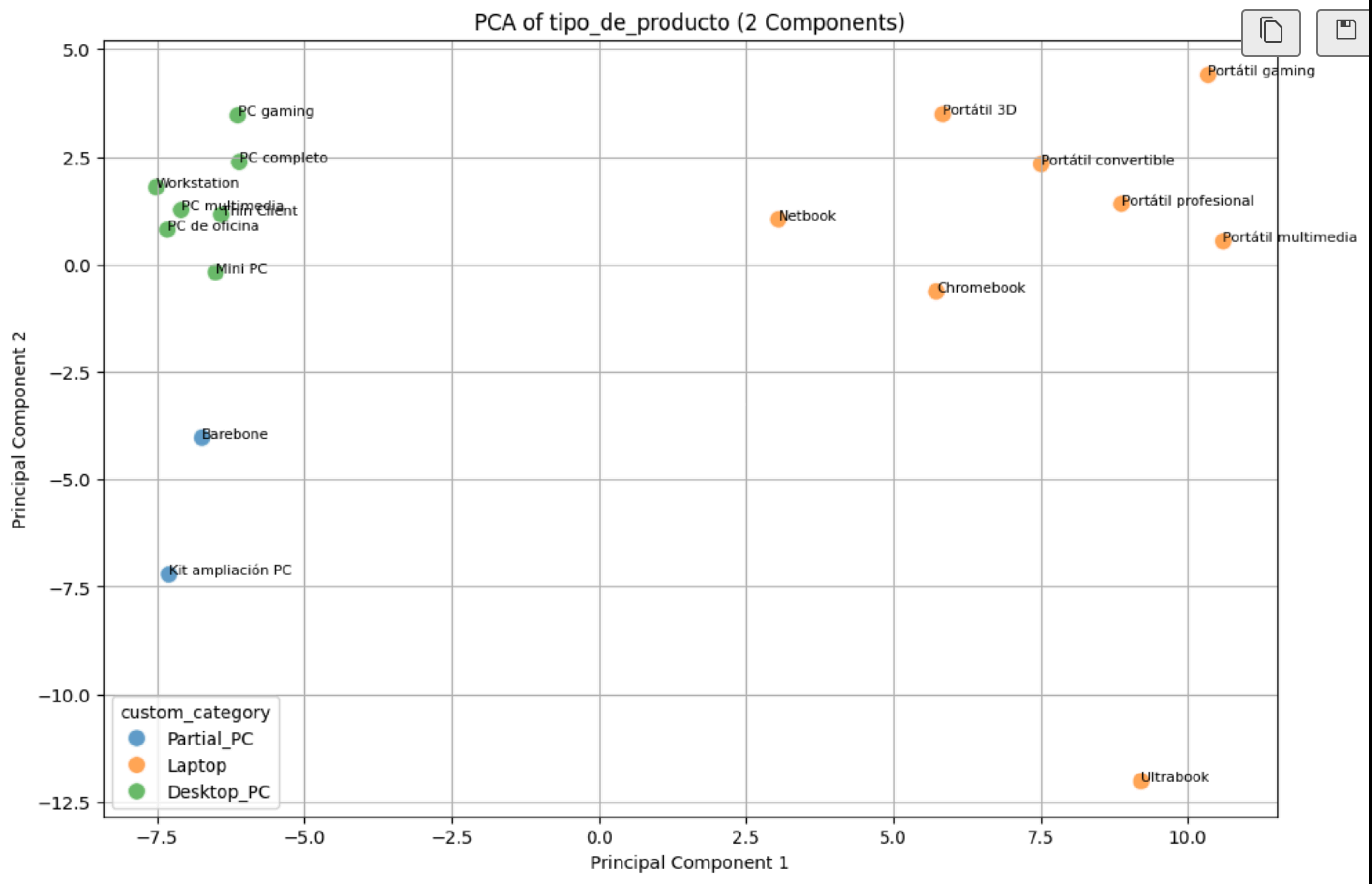
6. Standardized Screen Resolution

- Used regex to convert inconsistent resolution strings to “WIDTHXHEIGHT”

E.g., "4K (3.840 x 2.160)" → "3840x2160"

7. Offers Cleaning

- Convert strings like “200 ofertas” to 200.0 (float) for numeric ops.



PCA of tipo_de_producto (2 Components)
Principal Component 1
Principal Component 2
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
PC gaming
PC completo
Workstation
PC multimedia
Thin Client
PC de oficina
Mini PC
Barebone
Kit ampliación PC
Portátil gaming
Portátil 3D
Portátil convertible
Portátil profesional
Netbook
Portátil multimedia
Chromebook
Ultrabook
custom_category
Partial_PC
Laptop
Desktop_PC

HANDLING MISSING DATA

- **Used** df.isnull().sum() **and** missingno heatmaps
- **Aware of** Missing-Not-At-Random (MNAR) issues (e.g., screens missing in desktops). To solve, we handled it by isolating category-specific structures and then:

STRATEGY?

| 70% missing: dropped

- **30–70%:** conditional imputation or dropped
- **<30%:** imputed by product category using mean/mode

FEATURE ENGINEERING & SELECTION

HALL OF FAME

FEATURE ENGINEERING & SELECTION

1

ONE-HOT ENCODING

for low-cardinality categorical fields.

2

ORDINAL ENCODING

for ordered features *like* *processor generation*

CATEGORICAL HANDLING

3

PCA & CORRELATION ANALYSIS

PCA to retain features explaining 90%+ variance + Removed highly correlated variables (Pearsons).

4

FINAL MATRIX

Final feature matrix optimized for model performance & interpretability.

MEAN PRICE: Extracted from raw price range string

```
def process_price_range(price_str)
```

Volume (cm^3)= height x width x depth

Category Mapping: Mapped devices to **English Classes (Ultrabook, Tower, All-in-One)**

FEATURE ENGINEERING

MODEL TRAINING & VALIDATION

TECHNICAL APPROACH FOR SOLVING FUNCTIONALITIES

DESCRIPTIVE

K-MEANS CLUSTERING

- **Business value:** clustering different product segments.
- **Inputs:** features in the df_engineered.csv dataset.
- **Output:** k=2 clusters.
- **Evaluation:** PCA dimensionality reduction for visual inspection of clusters, plotted clusters and tipo_de_producto feature against same principal component axes.

PREDICTIVE

LIGHTGBM REGRESSION

- **Business value:**
  - Price prediction of computer given user selected specifications.
  - Feature importance on price of computer.
- **Inputs:** features in df_engineered_desktop_pc.csv or df_engineered_laptop.csv.
- **Output:** predicted price and feature importance.
- **Target:** precio_mean feature.
- **Validation:** cross-validation hyperparameter tuning.
- **Evaluation:** RMSE ≈ 520 EUR, R^2 ≈ 0.81.

PRESCRIPTIVE

K-NEAREST NEIGHBORS

- **Business value:**
  - K-similar product offers to user selected specifications.
- Inputs: features in df_engineered_desktop_pc.csv or df_engineered_laptop.csv.
- **Output:** k=5 neighbours.
- **Evaluation:** KNN regressor perdicted versus target price scatter plot and residual plot as baseline performance metrics.

APP ARCHITECTURE & DEPLOYMENT

Google Cloud

FRONTEND STACK

REACT-BASED UI

BACKEND APIS

PYTHON: SCIKIT-LEARN PANDAS, MATPLOTLIB

FOR THE EDA & TRAINING

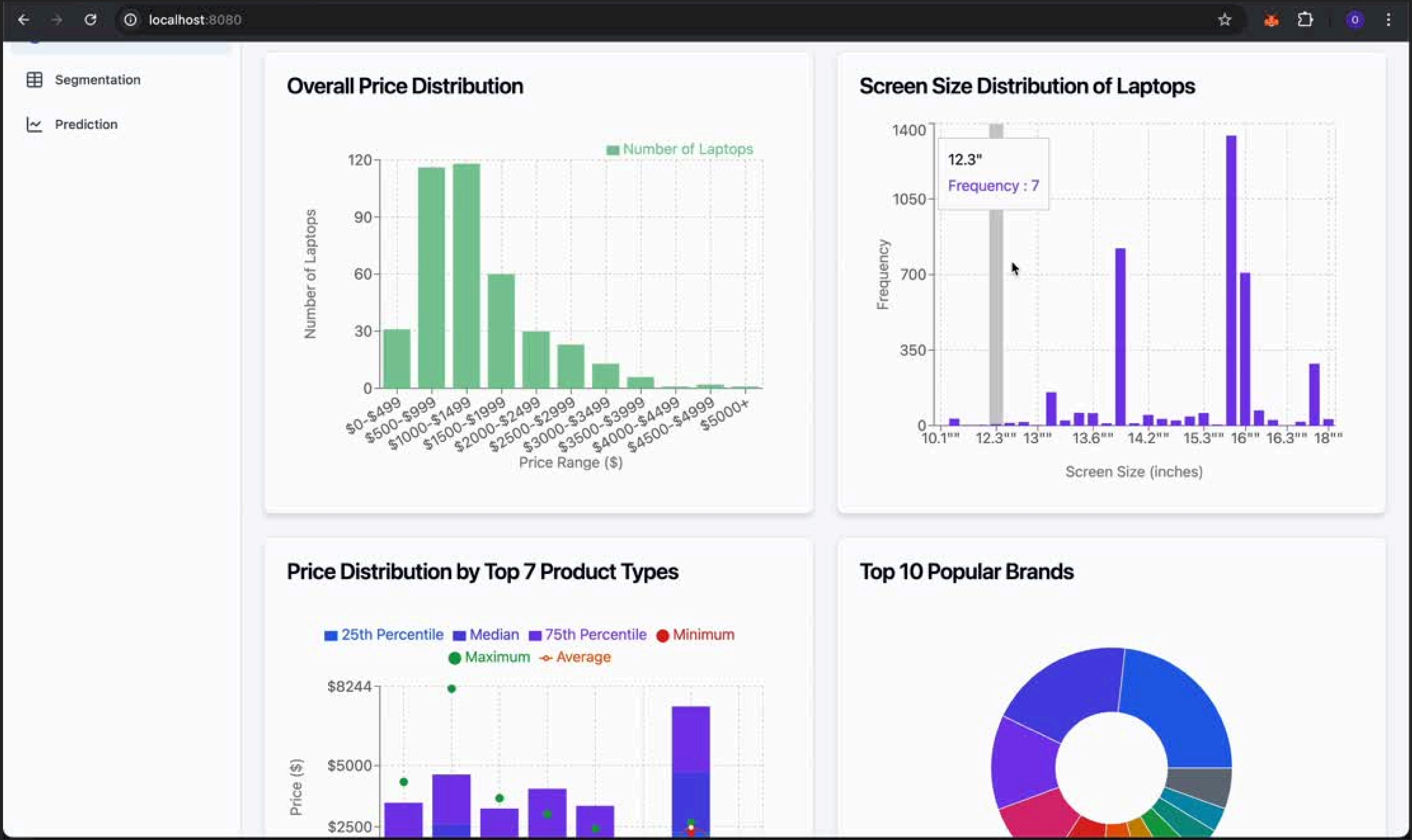
API HOSTING: DEPLOYED VIA GIT HUB → **GOOGLE CLOUD RUN FUNCTIONS**

ML MODELS: LIGHTGBM, KMEANS, KNN IN PYTHON **(JOBLIB SERIALIZED)**

MODEL STORAGE: **GOOGLE CLOUD STORAGE**

CI/CD AUTOMATION: GITHUB ACTIONS - TRIGGERED ON PUSH TO MAIN FOR THE MODELS

LIVE DEMONSTRATION


localhost:8080
Segmentation
Prediction
Overall Price Distribution
Number of Laptops
Number of Laptops
120
90
60
30
0
$0-$499
$500-$999
$1000-$1499
$1500-$1999
$2000-$2499
$2500-$2999
$3000-$3499
$3500-$3999
$4000-$4499
$4500-$4999
$5000+
Price Range ($)
Screen Size Distribution of Laptops
1400
1050
700
350
0
12.3"
Frequency : 7
Frequency
10.1""
12.3""
13""
13.6""
14.2""
15.3""
16""
16.3""
18""
Screen Size (inches)
Price Distribution by Top 7 Product Types
25th Percentile
Median
75th Percentile
Minimum
Maximum
Average
$8244
$5000
$2500
Price ($)
Top 10 Popular Brands

IMPROVEMENTS & NEXT STEPS

GENIE'S NEXT EVOLUTION

- **Live data integration** via APIs to keep listings up to date
- **Prediction confidence intervals** to show uncertainty
- **User-based personalization** using historical preferences
- **Model retraining** via feedback log ingestion
- **Multilingual toggle** to support Spanish/English UIs
- **Domain expansion** to peripherals, monitors, GPUs
- **Feature Feedback** to allow for constant improvements of model & the display of processed data.

THANK YOU