



Automated Feature Extraction

Module 10: Enriching Pandas DataFrames with Intelligent Variables

The Shift in Feature Engineering

From Manual RegEx to Semantic Intelligence

NLP Evolution: A Comparison

Traditional Methods

- **TF-IDF & BoW:** Heavy on frequency, light on meaning.
- **Regex:** Hardcoded patterns; brittle against variation.
- **Manual Labeling:** Costly, slow, and prone to inconsistency.

LLM-Based Extraction

- **Semantic Awareness:** Understands context and nuance.
- **Zero-Shot:** Ready to categorize without training data.
- **Scalable Logic:** Automates logic that previously required humans.

Common Extraction Tasks



Sentiment

Continuous or categorical metrics derived from complex reviews.



Tagging

Automated classification tags for tickets, logs, or feedback.



Entities

Extracting names, brands, or dates into structured columns.



Intent

Identifying the "why" behind a user's textual input.

Technical Orchestration



Powering Pandas with LangChain

The Python Implementation

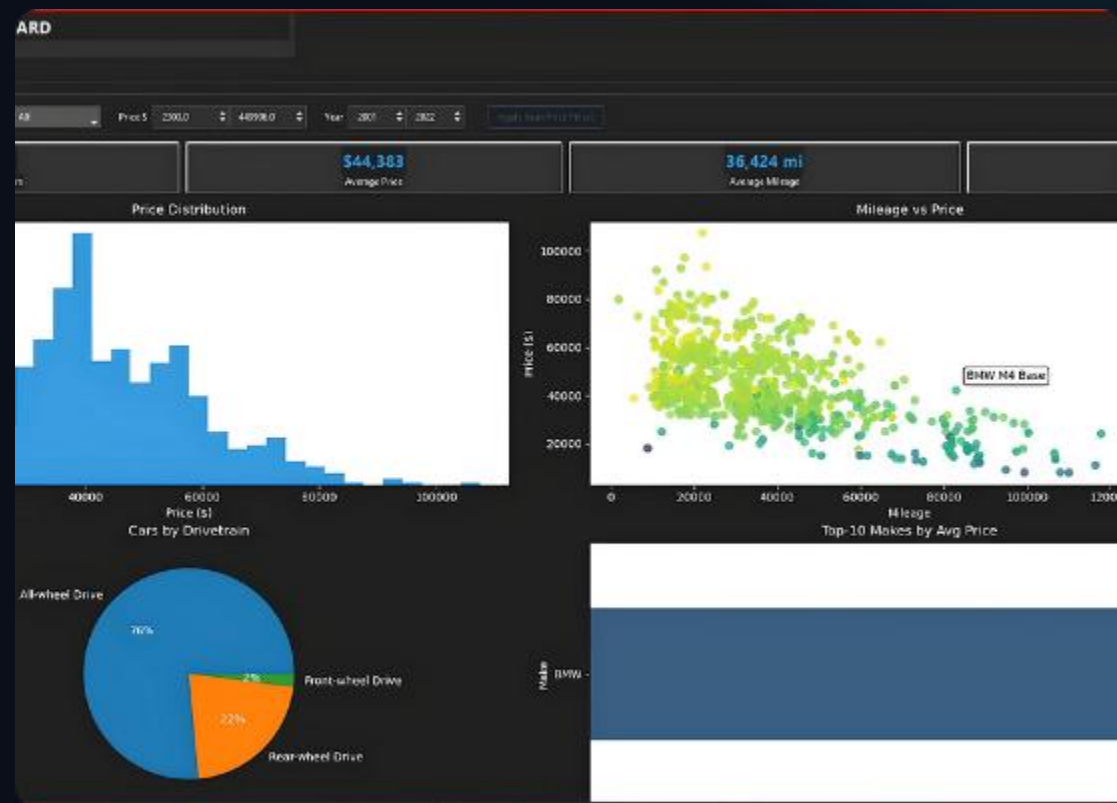
The core of the workflow relies on piping text data from a DataFrame through a LangChain operation.

`</>` Define a structured chain.

`⋮` Use list comprehensions for speed.

`⌘` Directly assign results to `df['new_col']`.

```
df['sentiment'] = [chain.invoke(x) for x in
df['raw_text']]
```



| Unlocking Productivity

95%

Reduction in Manual Effort

Faster Feature Engineering

LLMs allow data scientists to turn unformatted text into thousands of structured features in minutes rather than weeks of manual labeling or regex tuning.

This accelerates the entire ML model development lifecycle.

Execution Best Practices

-  **Rate Limiting:** Use sleep timers or batching to stay within API tier limits.
-  **Deterministic Output:** Keep temperature at 0.0 for consistent feature values.
-  **Error Handling:** Wrap invokes in try-except blocks to prevent partial failures.
-  **Pydantic Parsing:** Enforce schemas to ensure outputs are valid for Pandas ingestion.

Transformation: Text to Features

Raw Customer Feedback	Sentiment	Primary Tag	Urgency
"The battery dies after 2 hours. Terrible."	Negative	Hardware	High
"Love the design, but shipping took weeks."	Mixed	Logistics	Medium
"Is there a dark mode for the dashboard?"	Neutral	Feature Request	Low
"Fixed my issue instantly! Great support."	Positive	Customer Support	Low

Accuracy vs Implementation Speed



LLMs bridge the gap between human precision and machine speed.

Scaling Data Enrichment

Building intelligent pipelines that automatically translate raw customer voices into structured, predictive features for downstream machine learning models.

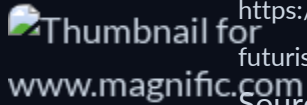
Questions?

Mastering Module 10: Feature
Extraction

✉ instructor@datascience.ai

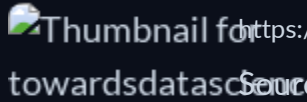
🌐 www.llm-bootcamp.com

Image Sources



https://img.magnific.com/premium-photo/abstract-digital-network-background-with-glowing-nodes-lines-representing-data-flow-connectivity-technology-blue-orange-colors-create-futuristic-dynamic-visual-effect_256259-10953.jpg

Source: www.magnific.com



<https://towardsdatascience.com/wp-content/uploads/2025/08/tkinter.png>

Source: towardsdatascience.com



https://images.stockcake.com/public/c/0/b/c0bb964e-607f-49fe-9ed6-0367fa9b1c7e_large/futuristic-data-center-stockcake.jpg

Source: stockcake.com