



# DataLens

Deterministic EDA & Insight Engine

Transform raw datasets into clean, structured, and insight-ready data with transparent, deterministic analytics.



# The Problem with Current EDA Tools

## Static Reports

Limited interactivity and flexibility for deeper exploration.

## AI Hallucination Risk

Heavy reliance on AI-generated insights that may lack transparency or accuracy.

DataLens bridges this gap by separating computation from explanation, ensuring credibility and reliability.



# Core Objectives



## Zero Hallucination

Deterministic end-to-end EDA ensures reproducible, trustworthy insights.



## ML Readiness

Comprehensive assessment of dataset preparedness for modeling.



## Natural Language

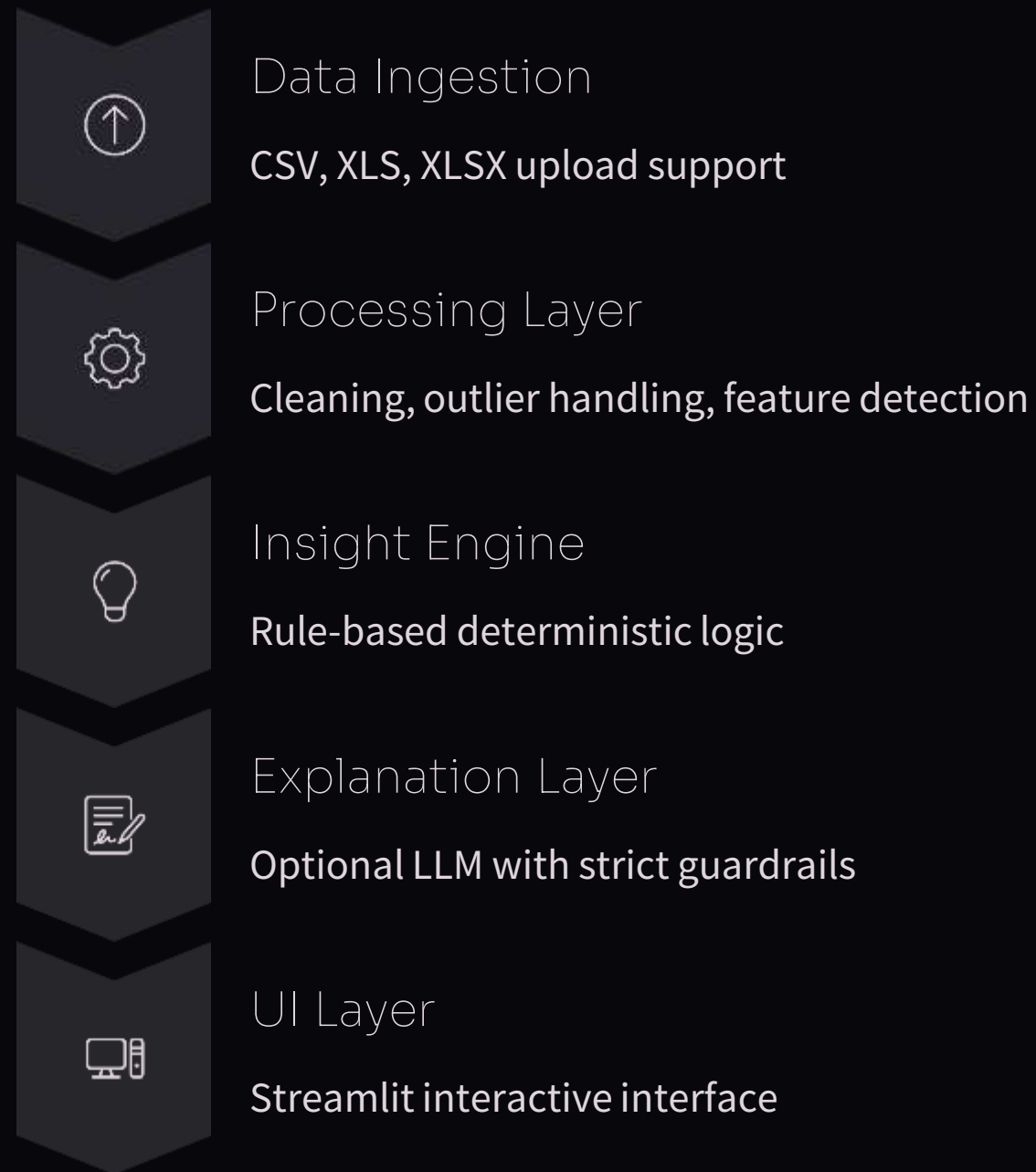
Optional explanations in plain English without compromising accuracy.



## Production-Grade

Recruiter-ready analytics tool with professional architecture.

# System Architecture



# Comprehensive Feature Set

## Data Cleaning

- Missing value analysis and handling
- Duplicate removal
- Outlier capping using IQR
- Dataset health score computation

## Feature Engineering

- Skewness and correlation detection
- High-cardinality identification
- Datetime feature extraction
- One-Hot & Ordinal encoding

Column-wise analysis, ML readiness evaluation, and cleaned dataset export complete the toolkit.





# Deterministic Insight Engine

## Rule-Based Logic

All insights computed using transparent, reproducible algorithms—no AI involved in computation.

## Trustworthy Metrics

Skewed features, correlations, and data quality metrics derived from established statistical methods.

## Zero Assumptions

Every insight is grounded in actual data patterns, ensuring credibility and reliability.

# DataLens Insights

No columns exceed missing value threshold.

⚠ High-cardinality categorical columns

```
▼ {  
  "name" : 47896  
  "host_name" : 11452  
  "neighbourhood" : 221  
  "last_review" : 1764  
}
```

⚠ Skewed numeric features

	0	
minimum_nights		1.29
number_of_reviews		1.23
host_id		1.18

# Responsible AI Integration



AI in Computation



Deterministic Logic

LLMs Used Only For

- Explaining already computed insights
- Answering user questions using dataset summaries

Strict Guardrails Enforce

- No new insight generation
- No assumptions beyond provided data



# ML Readiness Assessment

## Dataset Size

Evaluates sufficient sample size for reliable model training.

## Numeric Feature Ratio

Assesses balance of numeric vs categorical features.

## Missing Value Percentage

Identifies data completeness and potential risks.

Outputs human-readable risk assessments to guide modeling decisions and data preparation strategies.

This ML Readiness Assessment indicates maximum preparedness and zero structural impediments, providing a robust foundation for immediate model development and production deployment.

---

## 1. Dataset Overview

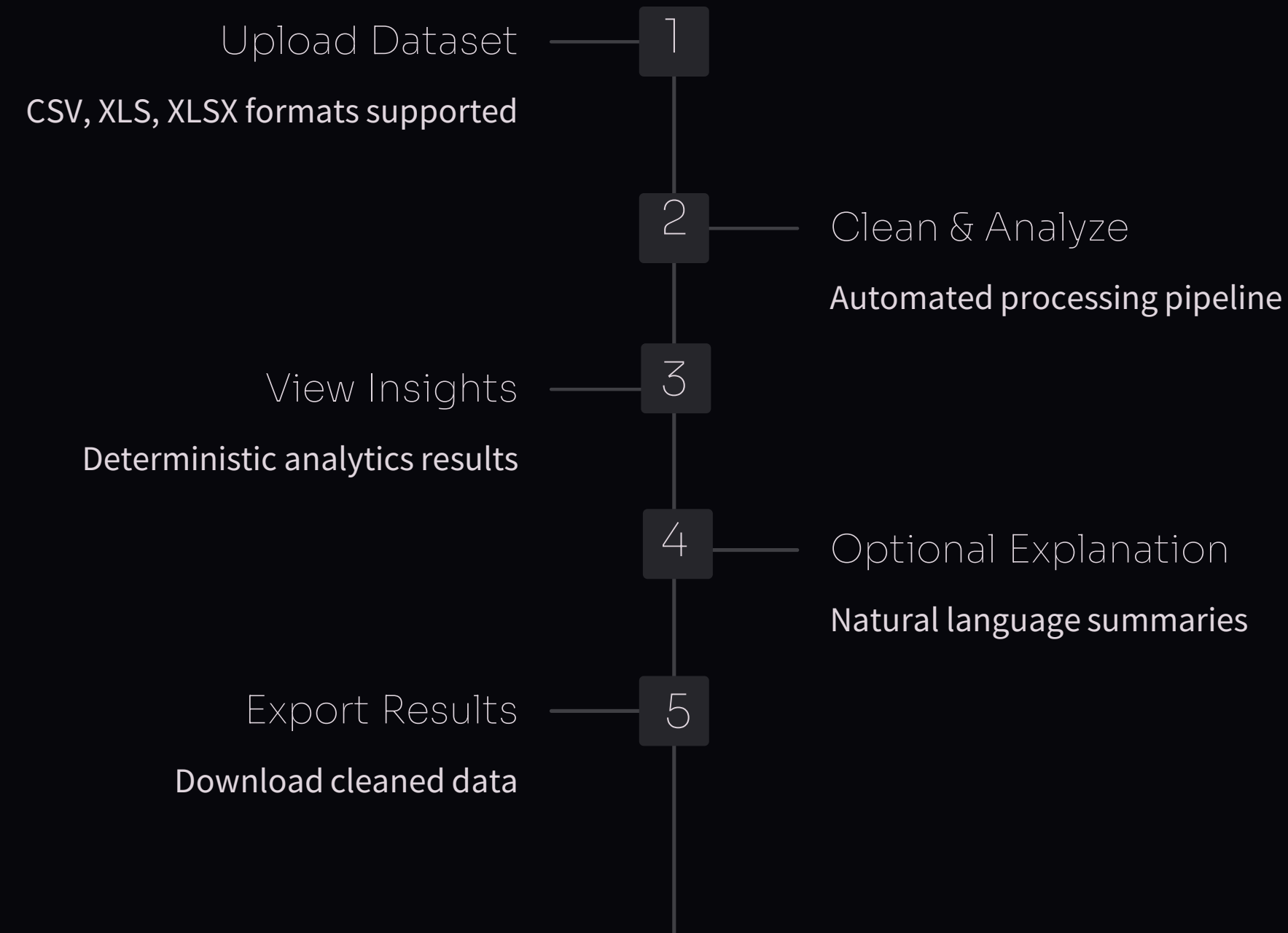
Based strictly on the assessment score of **100**, the dataset is confirmed to be fully compliant and mature across all readiness metrics. This implies successful vetting against necessary criteria related to data volume, feature engineering completeness, labeling accuracy, and accessibility for the Machine Learning lifecycle. The environment is considered *production-ready* from a data perspective.

## 2. Data Quality Issues

**Zero data quality issues were identified.**

The perfect Health Score and the absence of identified risks confirm that all measured data quality dimensions (including completeness, consistency, validity, and timeliness) are met. The underlying data preparation, cleansing pipelines, and schema enforcement processes are operating at optimal efficiency.

# Seamless User Experience





# DATALENS

FROM RAW DATA INTO ACTIONABLE  
KNOWLEDGE

Upload a CSV, xls, xlsx file to explore, clean, visualize, and engineer features dynamically.

Upload a data file



Drag and drop file here

Limit 200MB per file • CSV, XLS, XLSX

Browse files



AB\_NYC\_2019 (1).csv 6.7MB





# Why DataLens Stands Out

1

Transparent Analytics  
Rule-based logic you can trust  
and verify

2

Optional AI  
Not AI-dependent, AI-enhanced

3

ML-Ready Focus  
Built for data science workflows

4

Portfolio-Grade  
Production-style architecture

---

## Technology Stack

Python • Pandas • NumPy • Streamlit • Matplotlib • Scikit-learn • Gemini API (optional, guarded)