Visualizing Housing Market Trends: An Analysis of Sale Prices and Features

1. High-Level Architecture Overview

Data Sources → Data Ingestion → Data Storage → Data Processing → Analytics Layer → Visualization Layer → End Users

Architecture Style:

- Modular Data Pipeline

- Cloud-ready (scalable)

- Batch + Optional Real-time support

- BI-driven visualization

2. Architecture Components

2.1 Data Sources

Structured Data:

- Real estate transaction datasets (CSV, Excel)

- Government open data portals

- MLS exports

- Kaggle housing datasets

External APIs (Optional):

- Zillow API

- Redfin API

- Census demographic data

2.2 Data Ingestion Layer

Batch Processing:

- Python ETL scripts

- Apache Airflow (workflow orchestration)

- Scheduled ingestion (daily/weekly)

Real-Time (Optional):

- Apache Kafka (streaming ingestion)

- REST API ingestion service

Responsibilities:

- Data validation

- Schema enforcement

- Deduplication

- Timestamp standardization

2.3 Data Storage Layer

Raw Storage (Data Lake):

- AWS S3 / Azure Blob / Google Cloud Storage

Processed Storage (Data Warehouse):

- Amazon Redshift

- Google BigQuery

- Snowflake

- PostgreSQL

Schema Design:

Fact Table: Sales

Dimension Tables:

- Property Features

- Location

- Date

- Seller/Buyer

2.4 Data Processing Layer

Tools:

- Python (Pandas, NumPy)

- Apache Spark

- dbt

Processes:

- Missing value handling

- Feature engineering

- Outlier detection

- Price normalization

- Inflation adjustment

- Time series aggregation

Derived Metrics:

- Price per square foot

- Median price by location

- Year-over-year growth

- Feature-price correlation

2.5 Analytics & Modeling Layer

- Linear Regression

- Ridge/Lasso Regression

- Random Forest Regression

- XGBoost

Use Cases:

- Predict housing prices

- Identify influential features

- Trend forecasting

2.6 Visualization Layer

BI Tools:

- Tableau

- Power BI

- Looker

- Apache Superset

Custom Dashboard (Optional):

Frontend: React.js, Next.js, D3.js

Backend: FastAPI / Flask

Visualization Examples:

- Price trend over time

- Geographic heat maps

- Feature impact analysis

- Interactive filtering

3. Deployment Architecture (Example: AWS)

- S3 → Raw data

- AWS Glue → ETL

- Redshift → Warehouse

- EC2 / Lambda → APIs

- CloudFront → Frontend

- IAM → Access control

- CloudWatch → Monitoring

4. Non-Functional Requirements

- Scalability

- Performance optimization

- Security (RBAC, encryption)

- Reliability (backups, monitoring)

5. Recommended Tech Stack

- Ingestion: Python + Airflow

- Storage: PostgreSQL / Snowflake

- Processing: Pandas / Spark

- Modeling: Scikit-learn

- Visualization: Power BI / Tableau

- Backend API: FastAPI

- Deployment: Docker + AWS