House Sales in King County, USA

Author: Noor Syed
Date: February 2026

Project Overview

Title: Data Analysis with Python – House Sales in King County, USA

This project, conducted by Noor Yadullahi, focuses on analyzing house sale data for King County, including Seattle, covering homes sold between May 2014 and May 2015. The goal is to extract insights into factors affecting house prices, visualize trends, build predictive models, and prepare the project for deployment and integration into real-world systems.

Dataset Description

The dataset includes key features:

- **id, date, price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, yr_built, yr_renovated, zipcode, lat, long, sqft_living15, sqft_lot15**

It provides a comprehensive view of King County housing transactions over a one-year period, making it suitable for both exploratory analysis and predictive modeling.

Project Objectives

The project aims to prepare and clean the dataset, explore and visualize trends, identify features most correlated with price, build accurate predictive models, and enable the integration of insights into business intelligence tools and automated pipelines for real-world use.

1) Prepare and migrate the dataset for analysis.
2) Clean and preprocess the data.
3) Perform exploratory data analysis (EDA) to identify trends.
4) Build predictive models for house prices.
5) Data Integration with business intelligence tools and pipelines.
6) Generate actionable insights.

Data Migration

Before analysis, the project was prepared for seamless use and potential deployment:

Data Migration:

- Cleaned dataset exported to CSV for consistency.
- Ensured schema is standardized for all features to avoid errors during analysis or modeling.

And scripts were modularized **(data_cleaning.py, eda.py, modeling.py)** to ensure reusability. All dependencies were captured in **requirements.txt**, allowing smooth execution in any environment.

This migration ensures that the project code are fully ready for analysis, modeling, and integration with other systems.

Tools and Libraries Used

*Python 3.x pandas * NumPy * matplotlib * seaborn * scikit-learn * statsmodels*

Data Cleaning and Preprocessing

The preprocessing step involved loading the dataset, handling missing values, converting data types, engineering new features like house age, and removing outliers. These steps ensured data consistency and quality, providing a reliable foundation for analysis and modeling.

- ❖ Loaded data using pandas.
- ❖ Checked for and handled missing values.
- ❖ Converted types to numeric or datetime where needed.
- ❖ Feature engineering (house_age = yr_sold - yr_built).
- ❖ Removed unrealistic outliers to improve model accuracy.

Exploratory Data Analysis (EDA)

EDA revealed that price is strongly influenced by features such as square footage, grade, and bathrooms, while waterfront and view also significantly affect sale prices. Visualizations including **histograms, scatterplots, and heatmaps** helped identify trends, correlations, and distribution patterns across key variables.

Key insights:

- Price distribution is skewed; most houses are under $1.2M.
- Strong correlation between price and features like sqft_living, grade, and bathrooms.
- Waterfront and views significantly increase price.
- Houses in certain zip codes show higher pricing trends.

Visualizations included:

***Histograms * scatterplots * boxplots * correlation heatmaps.**

Modeling and Analysis

Predictive modeling included both simple and multiple linear regression techniques. The analysis confirmed that living space, grade, and above-ground square footage are the strongest predictors of house price, while waterfront properties command a significant premium. Evaluation metrics indicated strong model performance after feature selection and outlier removal.

- Simple Linear Regression: Price vs sqft_living.
- Multiple Linear Regression: Including bedrooms, bathrooms, grade, and waterfront.
- Evaluation Metrics: R^2 score, MAE, MSE.

Findings:

- sqft_living, grade, and sqft_above are the strongest predictors.
- Waterfront adds significant premium.
- Feature selection improves model performance.

Data Integration

- The project is designed to integrate seamlessly with business systems. Analysis results can be exported to **Tableau** or **Power BI** dashboards to provide live insights on pricing trends.

- Additionally, new data can be loaded automatically from **APIs** or **CSVs**, and predictions can be updated using pipelines such as Airflow, ensuring continuous, actionable insights.

This ensures that insights can be continuously applied and visualized in decision-making systems.

Conclusion and Outcome

The project successfully provides a full workflow for analysis. It identifies key drivers of house prices, builds accurate predictive models, and demonstrates readiness for both migration and integration into real-world business systems, offering practical insights for buyers, sellers, and analysts.

Summary

This project demonstrates a complete data analysis workflow on King County housing data. Starting from migration of raw data and modularized scripts, the project includes data cleaning, exploratory analysis, modeling, and concludes with integration into BI dashboards and automated pipelines.