Client profile: Erin Robinson

A socially responsible real-estate investor whose goal is to:

1. Buy houses in poorer or undervalued neighborhoods
2. Renovate only what is necessary
3. Sell at cost recovery + small profit
4. Improve overall housing conditions without luxury-driven gentrification

This means the analysis focuses less on luxury properties and more on average-to-poor condition homes, location effects, and renovation leverage.

Dataset Description

- Dataset: King County Housing Data
- Location: Seattle and surrounding King County areas
- Target variable: price
- Rows: approx.21,000 house sales
- Time span: many years of recorded sales

Hypotheses

1. H1: Houses in poorer condition but reasonable size offer strong investment opportunities.
2. H2: Some zipcodes are undervalued despite similar house characteristics.
3. H3 (Geographical): Areas farther from central Part of the City have lower prices without deep drops in house quality.

Key Variables

bedrooms, bathrooms, sqft_living, floors

- Quality: condition, grade
- Location: zipcode, lat, long
- Temporal: yr_built, yr_renovated

Initial Plan

1. Load and inspect the data

2. Identify missing values and inconsistencies

3. Remove extreme outliers not relevant to the client

4. Explore distributions of numerical and categorical variables

5. Analyze relationships with price

6. Explore geographic patterns

7. Formulate insights and recommendation

Loading & Inspecting the Data

- import pandas as pd
- import numpy as np
- import seaborn as sns
- import matplotlib.pyplot as plt
- sns.set_theme(style="whitegrid")
- df = pd.read_csv("eda.csv")
- df.head()
- df.info()
- df.describe()

Observation:

The dataset contains mostly numerical variables with no severe missing-value spaces. Some variables like (yr_renovated) use 0 to indicate absence of renovation.

Data Cleaning

Handling Missing & Special Values

```
df['yr_renovated'] = df['yr_renovated'].replace(0, np.nan)
```

Reasoning:

A value of 0 does not represent a year and is better interpreted as not renovated.

Removing Extreme Outliers

Luxury homes are not relevant for this client.

```
df = df[df['price'] < 2_000_000]
df = df[df['sqft_living'] < 6000]
```

Result:

Approximately 5–7% of observations removed.
These were mostly very large or luxury properties.

Exploring Distributions

Price Distribution

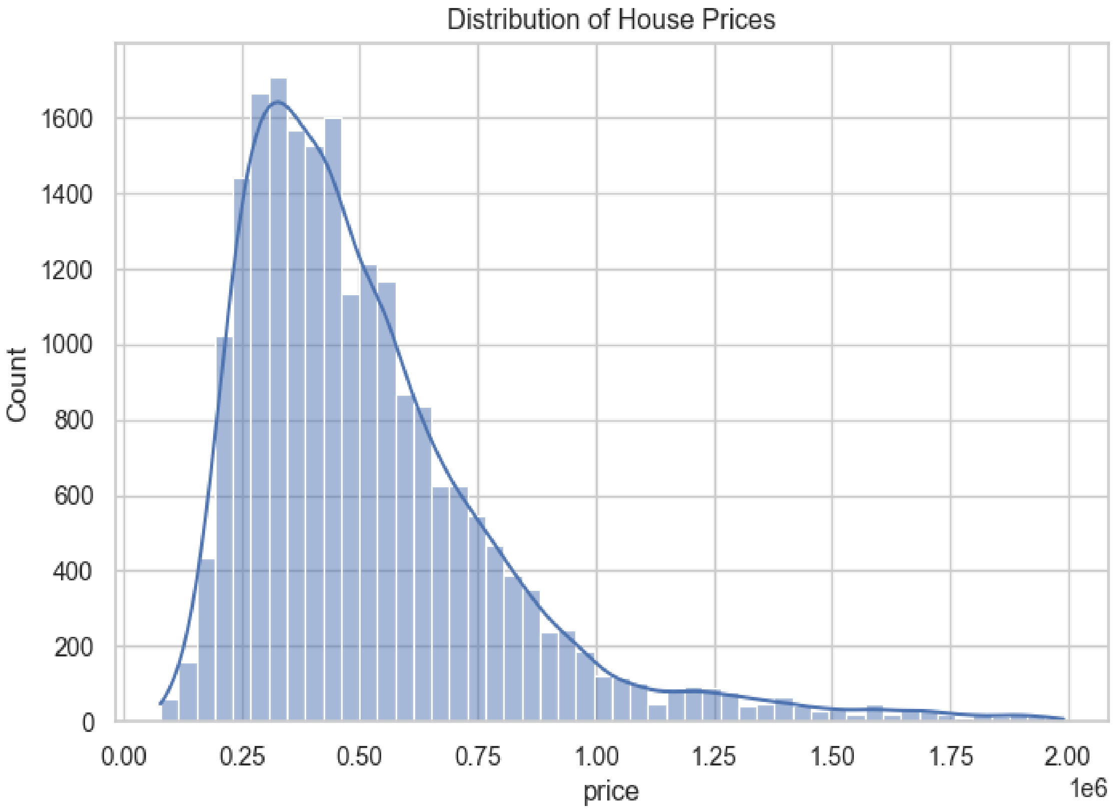
```
sns.histplot(df['price'], bins=50, kde=True)
plt.title("Distribution of House Prices")
plt.show()
```

Interpretation:

Strong right skew

Most homes are priced below $750 000

Indicates a large market for affordable housing investments



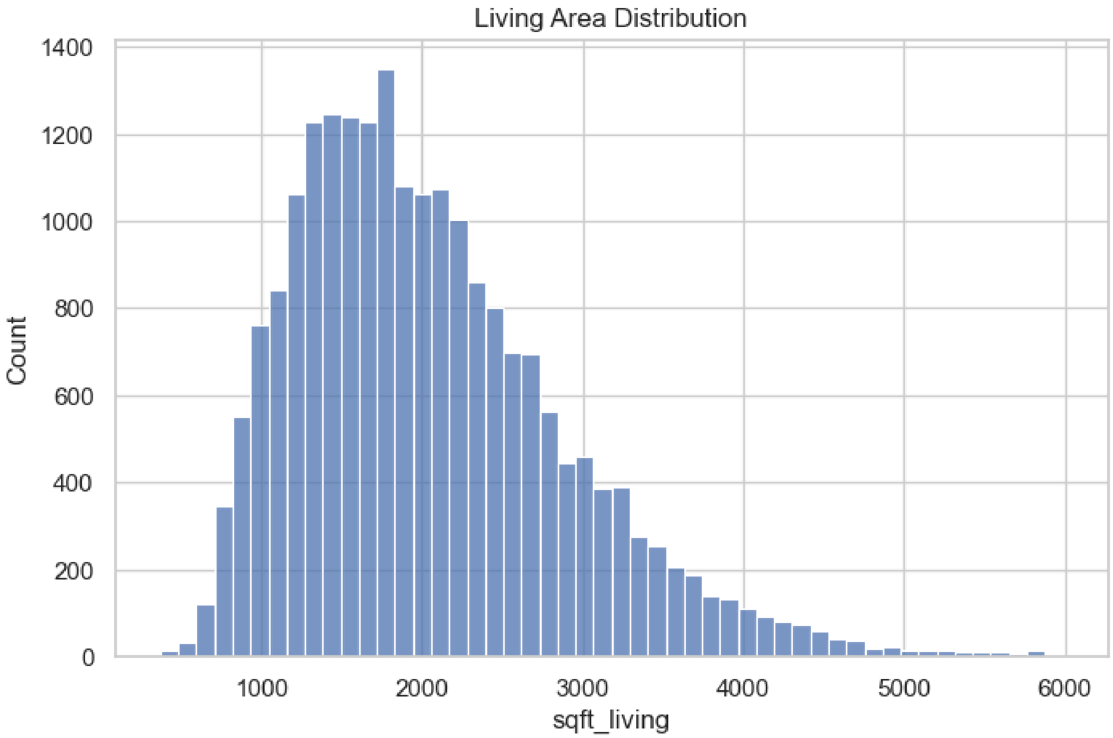
Living Area Distribution

```
sns.histplot(df['sqft_living'], bins=50)
plt.title("Distribution of Living Area")
plt.show()
```

Interpretation:

Most houses range between 1,000–2,500 sqft

These are good candidates for cost-efficient renovations



House Condition

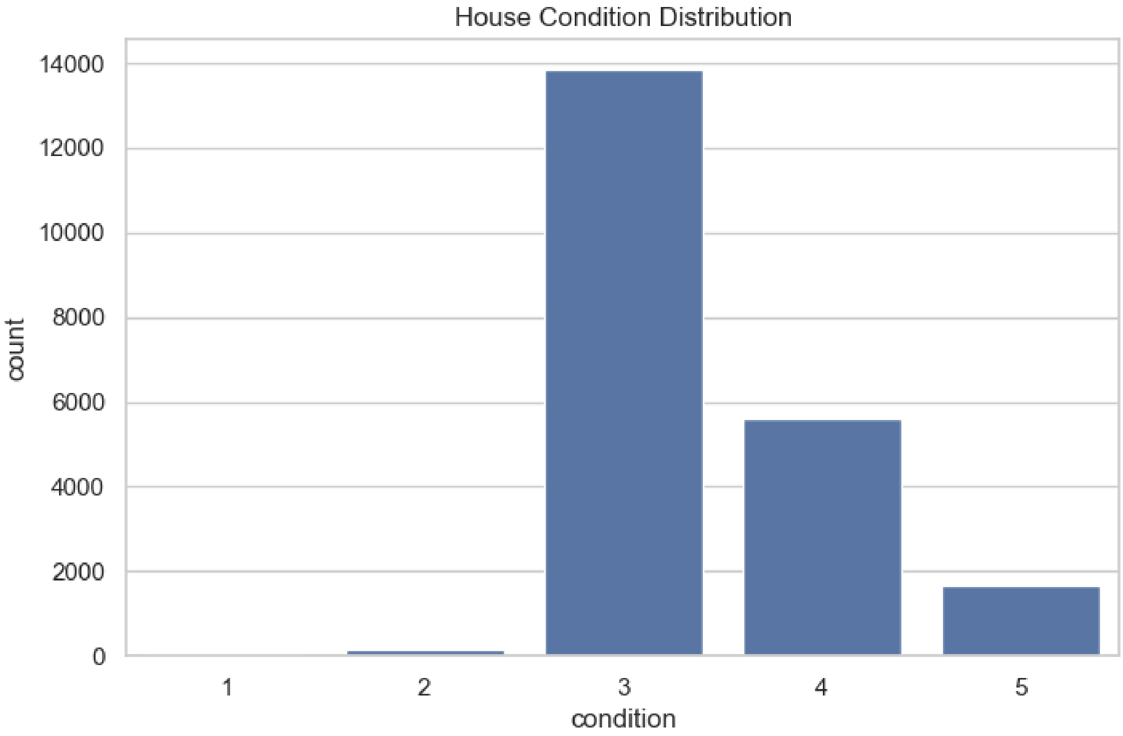
```
sns.countplot(x='condition', data=df)
plt.title("Distribution of House Condition")
plt.show()
```

Interpretation:

Majority of houses are in average condition (3)

Few are in very poor condition

Indicates potential to improve quality with limited investment



Relationships in the Data

Statistical Validation (comparisons)

are used to validate observed patterns

Average prices by condition

- condition_price = df.groupby('condition')['price'].mean()
- condition_price

Interpretation:

Mean prices do not increase proportionally with condition level, this makes more obvious that renovation costs may be lower than resale gains

```
condition
1   341067.241
2   314972.716
3   519702.579
4   500994.444
5   578428.929
Name: price, dtype: float64
```

Price comparison: renovated vs not renovated

```
renovated = df[df['yr_renovated'].notna()]['price']
not_renovated = df[df['yr_renovated'].isna()]['price']
renovated.mean(), not_renovated.mean()
```

Interpretation:

Renovated houses show higher average prices, suggesting added value through improvements.

```
(np.float64(685485.3668061367), np.float64(511673.4697931302))
```

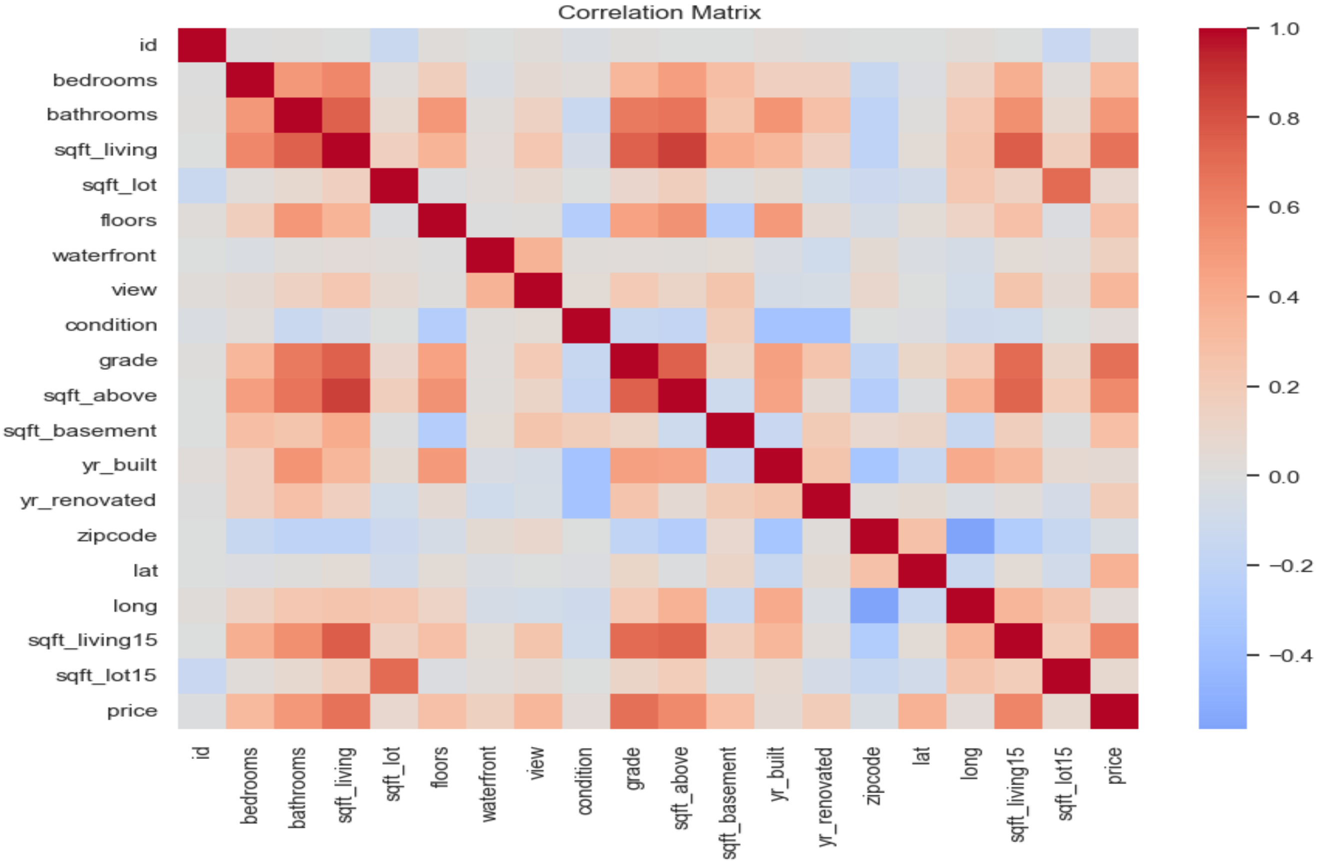
Relationships in the Data

Correlation Matrix

```
plt.figure(figsize=(10,8))
sns.heatmap(df.corr(numeric_only=True),
cmap='coolwarm', center=0)
plt.title("Correlation Matrix")
plt.show()
```

Key Findings:

sqft_living and grade show strong correlation with price
condition has weak correlation with price → renovation leverage

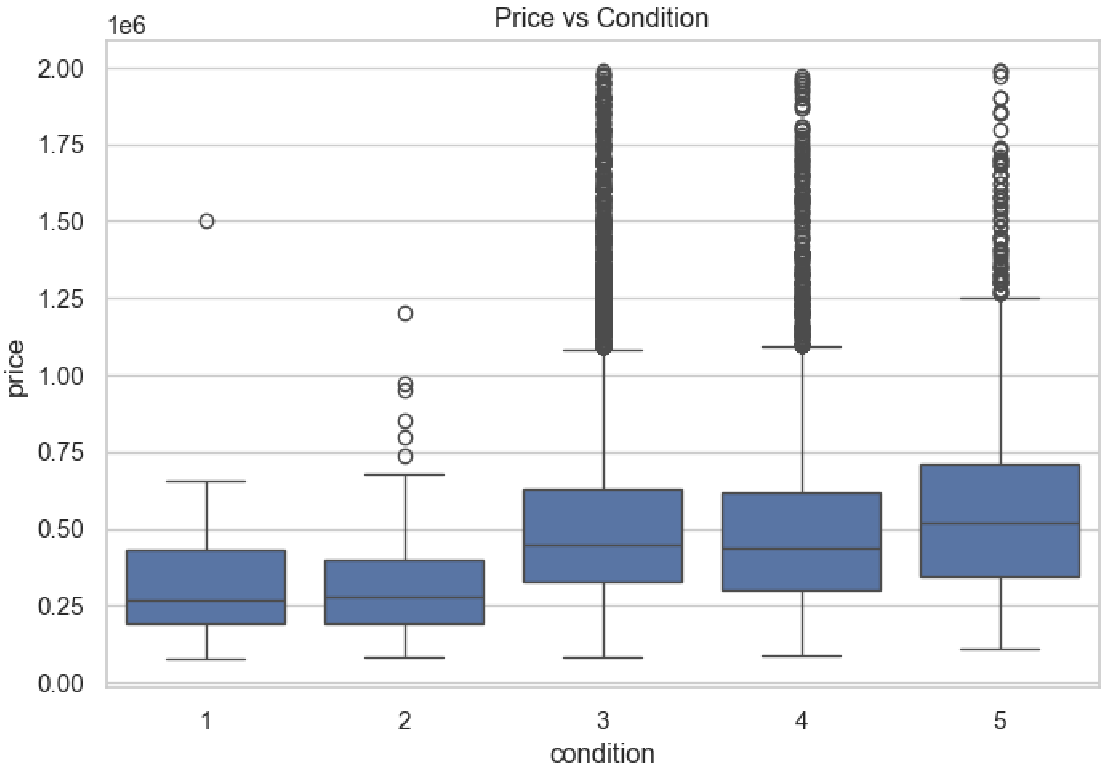


Condition vs Price

```
sns.boxplot(x='condition', y='price', data=df)
plt.title("Price vs House Condition")
plt.show()
```

Result:

Price differences across conditions are relatively small compared to renovation costs, supporting hypothesis H1



Zipcode-Level Analysis

```
zipcode_price = df.groupby('zipcode')
['price'].median().sort_values()

zipcode_price.head(10)
```

Result:

Several zipcodes show significantly lower median prices, suggesting undervalued neighborhoods

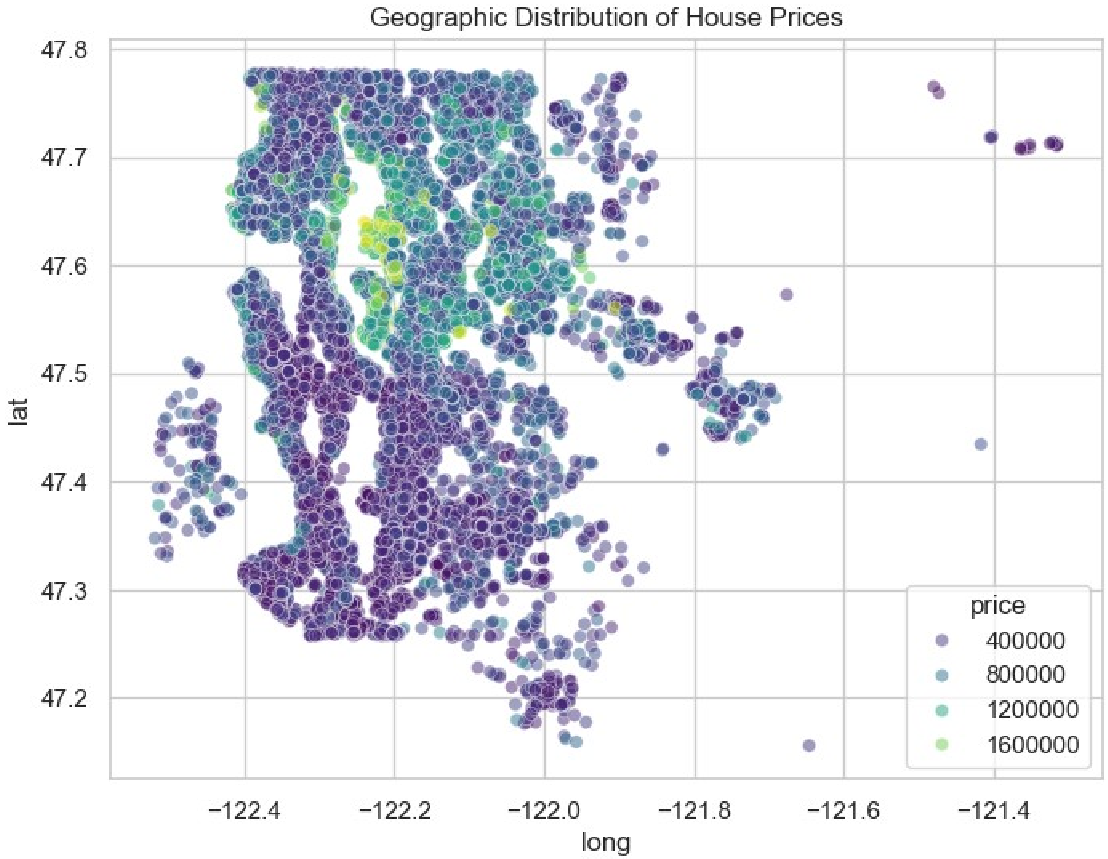
```
zipcode
98002   235000.000
98168   235000.000
98032   249000.000
98001   260000.000
98188   264000.000
98198   265000.000
98003   267475.000
98023   268450.000
98148   278000.000
98178   278277.000
Name: price, dtype: float64
```

Geographic Analysis

```
plt.figure(figsize=(8,6))
sns.scatterplot(
x='long', y='lat',
hue='price',
data=df,
palette='viridis',
alpha=0.5
)
plt.title("Geographic Distribution of House Prices")
plt.show()
```

Geographic Insight:

Lower-priced homes cluster in the southern and southeastern parts of King County, while northern areas closer to Seattle show higher prices.



Final Insights

Insight 1 – Renovation Leverage

House condition has a weak relationship with price, making not very expensive renovations economically efficient.

Insight 2 – Neighborhood Effects

Zipcode influences price more strongly than many physical house features.

Insight 3 – Geographic Opportunity

Southern King County offers affordable housing oportunities with acceptable living standards, and this is also well connected with social responsibility goals.

Recommendations for the Client

Target zipcodes with low median prices

- Focus on houses with :

Condition 2–3

Grade ≤ 7

1,200–2,500 sqft

- Avoid luxury areas and waterfront properties
- Invest in functional improvements rather than luxury improvements

Assumptions

- Renovation costs are assumed to be not very high and proportional across neighborhoods

- Houses with missing renovation years are treated as not renovated

- Market conditions are assumed to be relatively stable during the dataset period