

Who Am I & Who Is the Client?

- Role: Data Science Undergraduate Student
- Client: Erin Robinson
- Socially responsible real-estate investor
- Goal: Improve housing quality without luxury-driven gentrification

Dataset Description

- Dataset: King County Housing Data
- Location: Seattle & surrounding King County areas
- Size: ~21,000 house sales
- Time span: Multiple years of recorded transactions

Data Overview

- Target variable: House price
- Key features: size, condition, location, renovation status
- Mostly numerical variables with minimal missing data
- 0 values in renovation year treated as not renovated

Research Questions & Hypotheses

- H1: Poor-condition houses offer strong renovation leverage
- H2: Some zipcodes are undervalued despite similar house characteristics
- H3: Latitude & longitude reveal undervalued housing areas away from Seattle center.

Data Cleaning
Missing Values

df.isna().sum()
yr_renovated == 0 → Not renovated (keep)

No big reason to drop rows
unless critical fields missing

id	0
bedrooms	0
bathrooms	0
sqft_living	0
sqft_lo	0
floors	0
waterfront	2391
view	63
condition	0
grade	0
sqft_above	0
sqft_basement	452
yr_built	0
yr_renovated	
3848	
zipcode	0
lat	0
long	0
sqft_living15	0
sqft_lot15	0
date	0
price	0
dtype:	int64

Remove Obvious Outliers

Extremely large houses are not relevant for this client.

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

```
df = df[df['sqft_living'] < 6000]
```

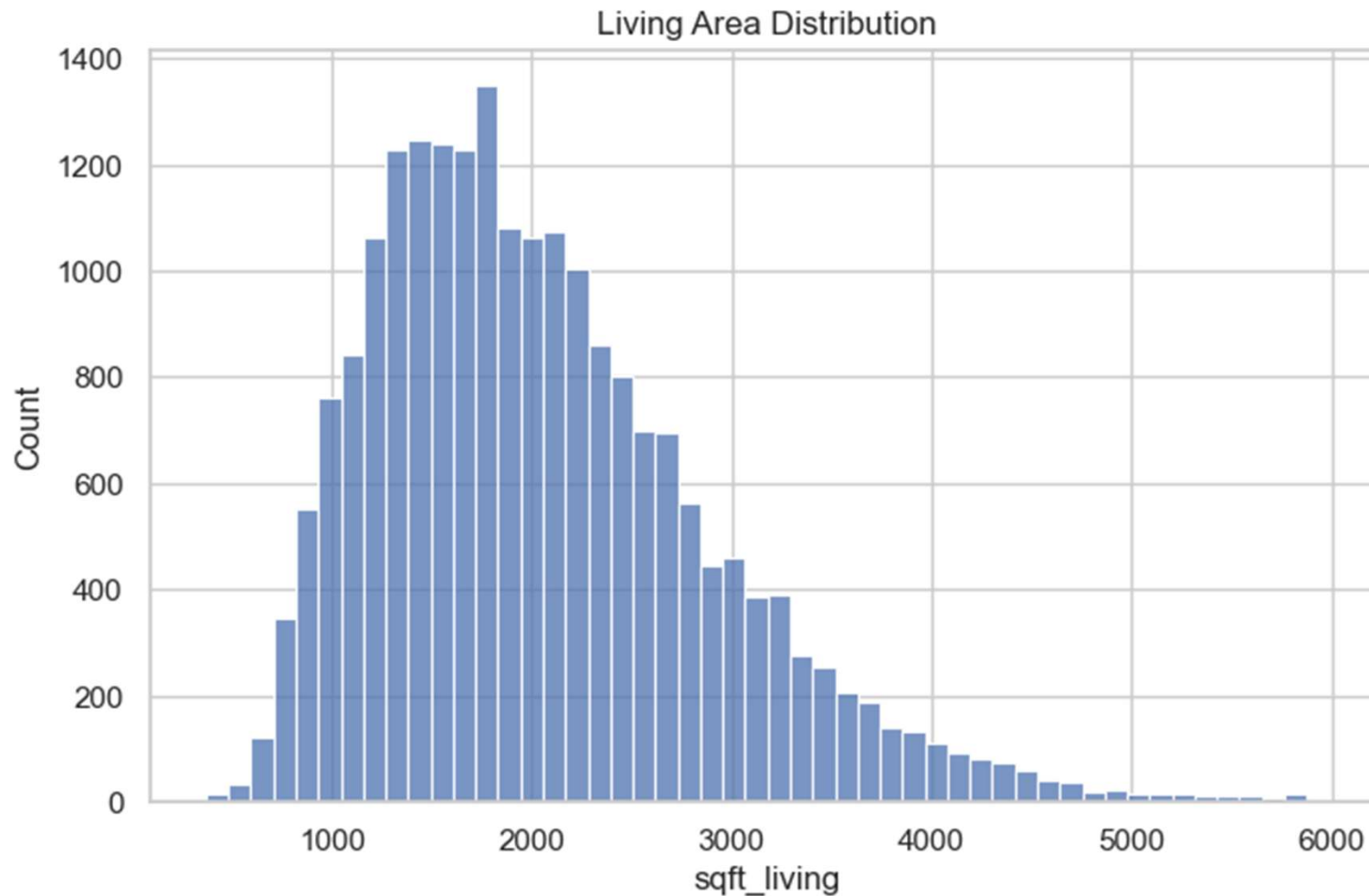
```
df['yr_renovated'] = df['yr_renovated'].replace(0, np.nan)
```

5–7% of data removed (mostly luxury properties)

Observation:

Typical homes: 1,000–2,500 sqft

Large mansions are mostly excluded → filtered out



Exploring Distributions

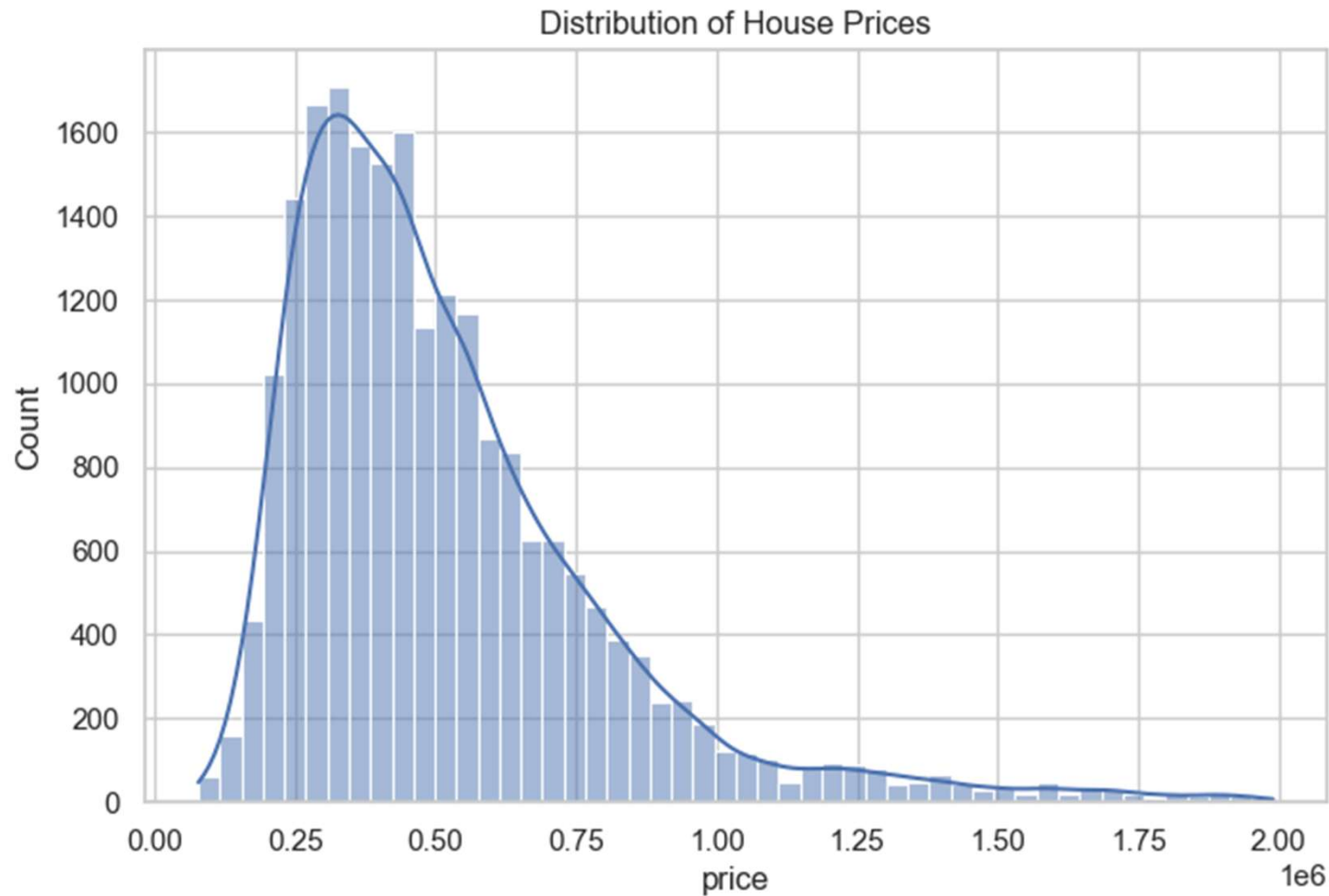
Target Variable – Price

Observation:

Strong right skew

Majority of homes priced below approx.\$750 000

Suitable for value-based investing



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()
```

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

Renovated vs Non-Renovated Houses



Renovated vs Non-Renovated Houses

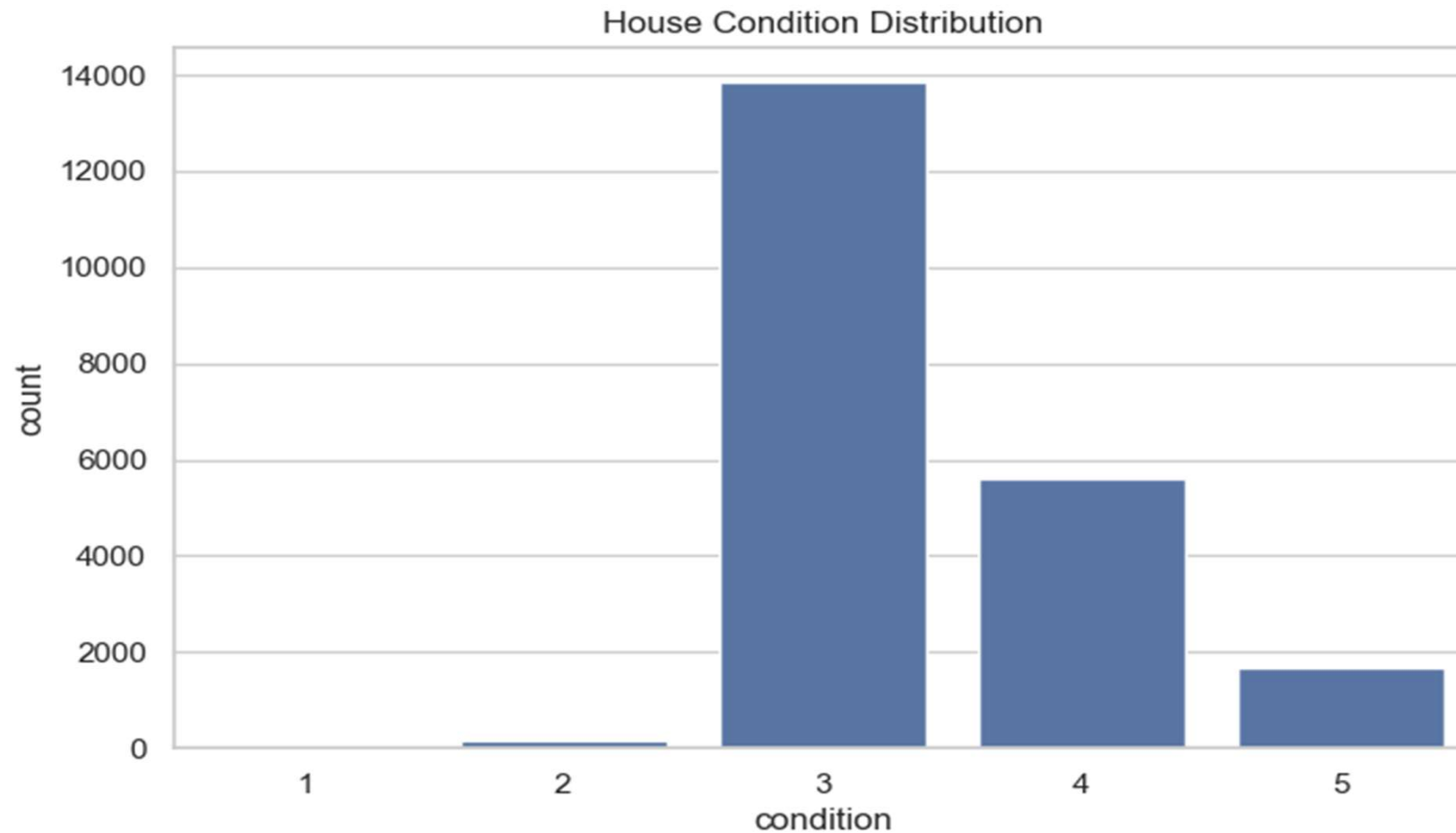
- Renovated houses sell for substantially higher prices
- Distance-based affordability does not imply poor housing quality

Observation:

Most houses rated 3 (average)

Few of them in very poor or excellent condition

one can see a good Opportunity to upgrade house condition cheaply



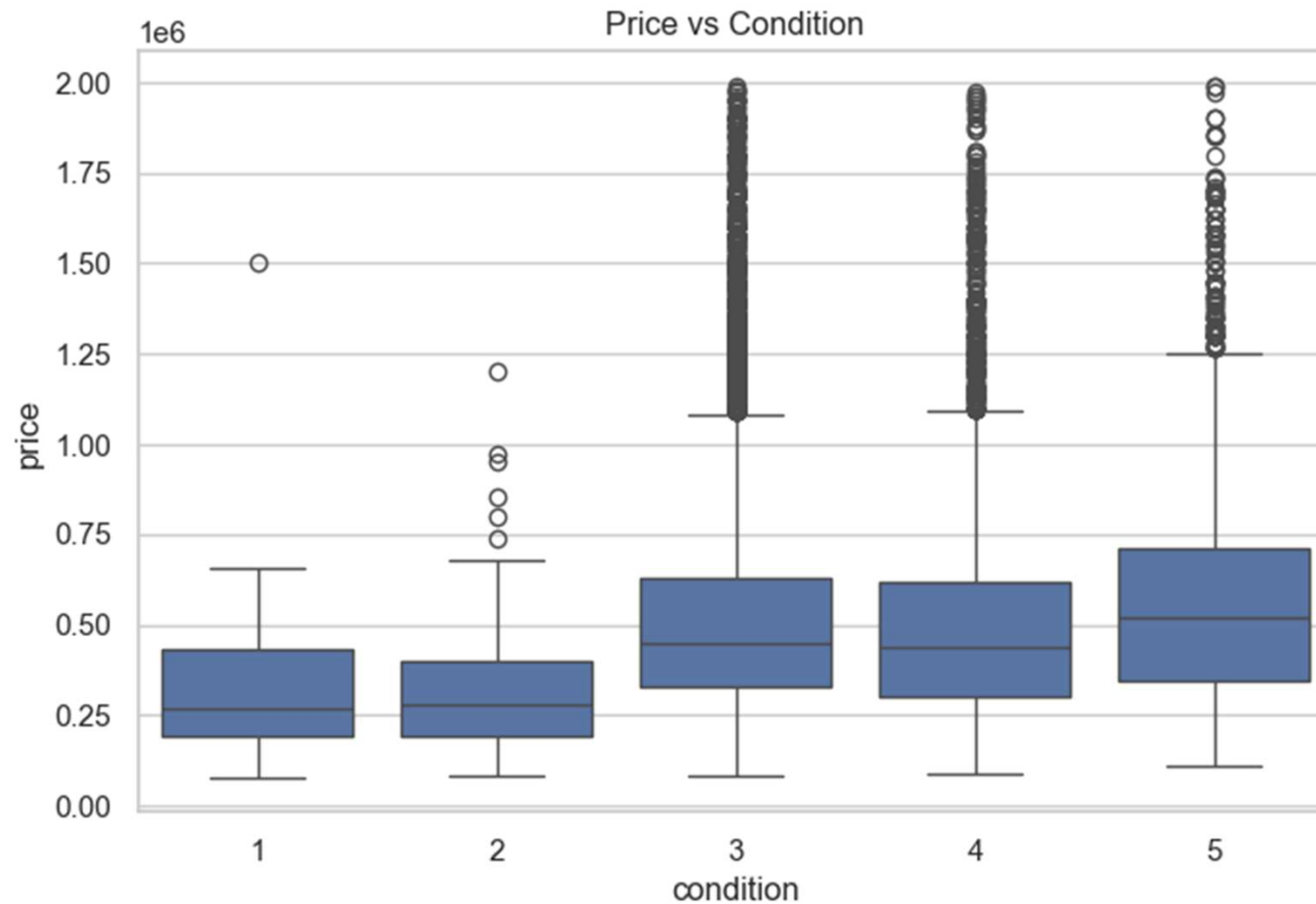
Hypothesis Analysis

Condition vs Price

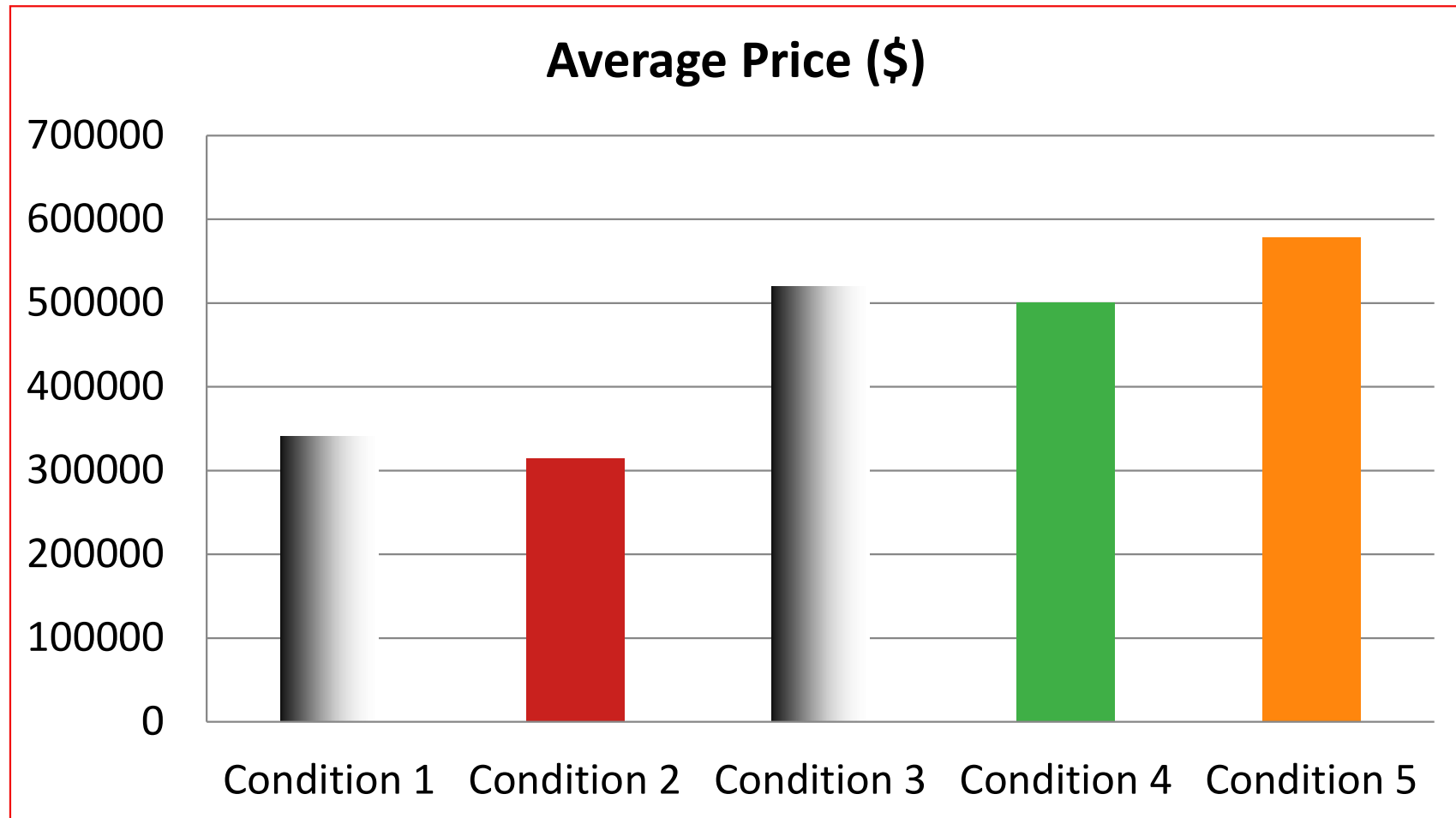
Insight:

Price does not rise sharply with condition

Renovations can increase resale value more than cost



H1 Result: Average Price by House Condition



H1 Interpretation

- House condition has a weak relationship with price
- Price differences are small compared to renovation costs
- Supports Hypothesis 1 (renovation leverage exists)

Zipcode-Level Pricing

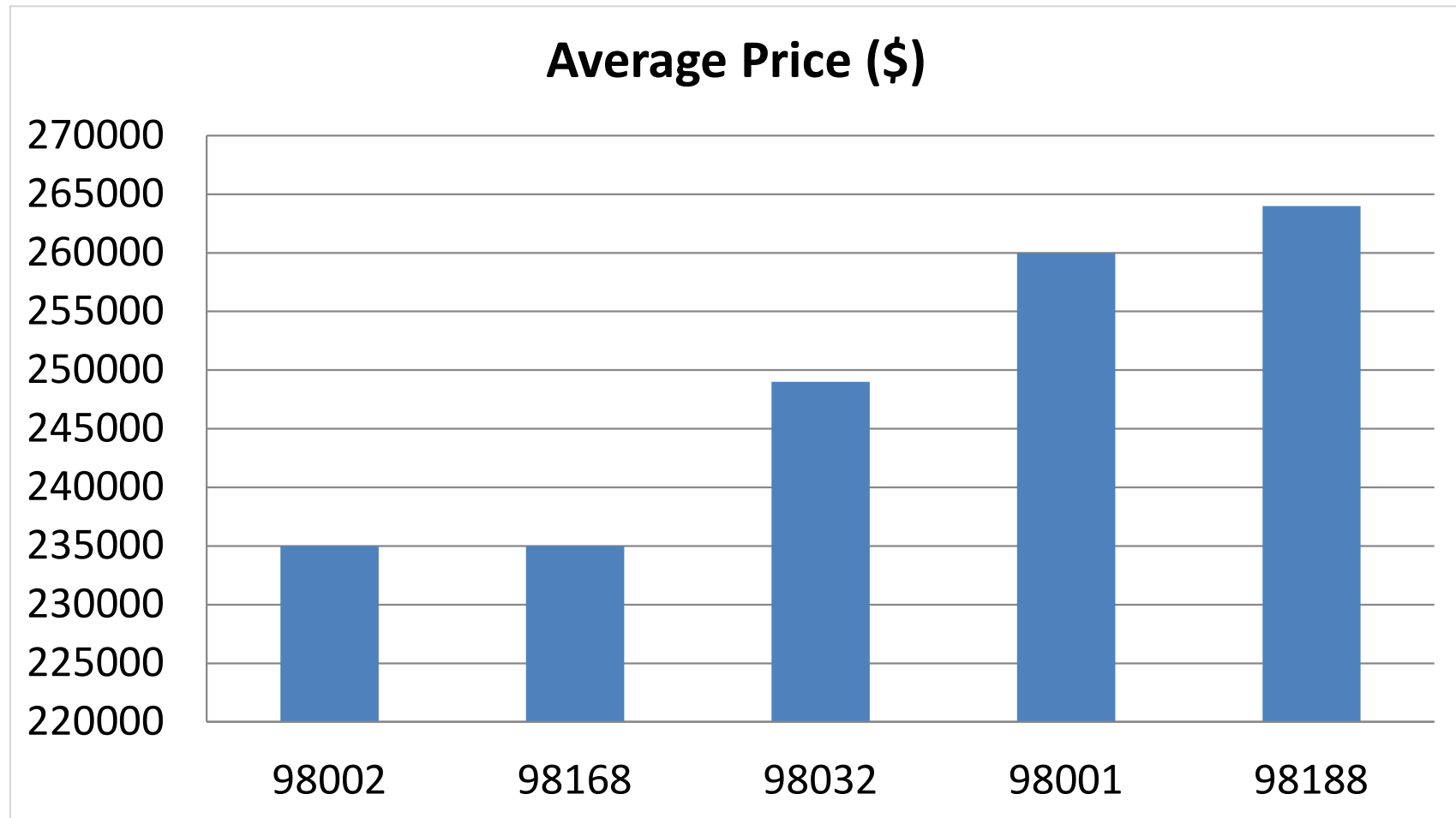
```
zip_price = df.groupby('zipcode')['price'].median().sort_values()  
zip_price.head(10)
```

Insight:

Some zipcodes are consistently undervalued

Ideal for socially responsible investing

H2 Result: Undervalued Zipcodes (Median Price)



H2 Interpretation

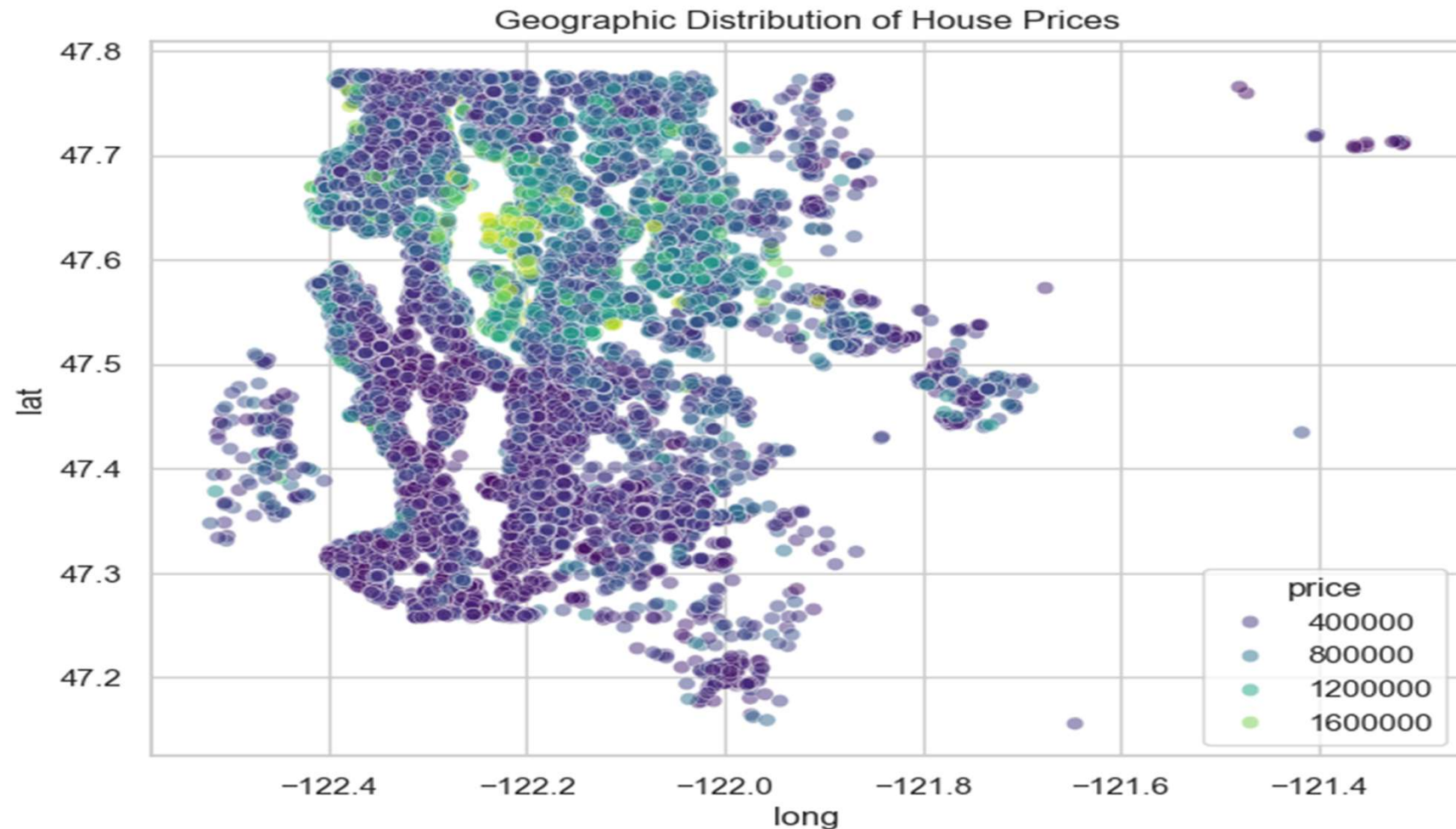
- Several zipcodes show significantly lower median prices
- Price differences not fully explained by house quality
- Supports Hypothesis 2 (location effects are strong)

Geographical Insight: Supports Hypothesis 3

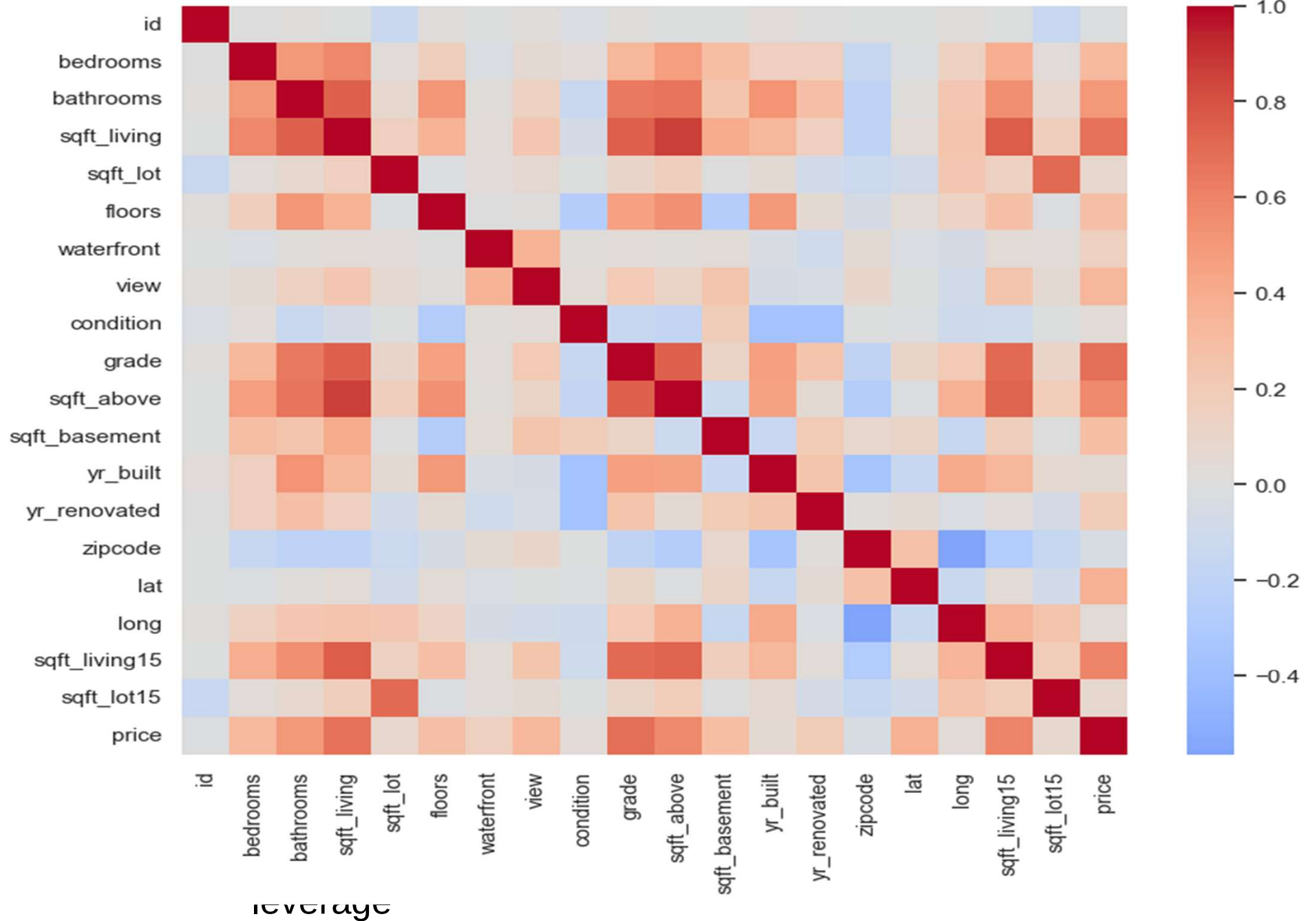
Lower prices cluster south & southeast of Seattle

These areas still have:

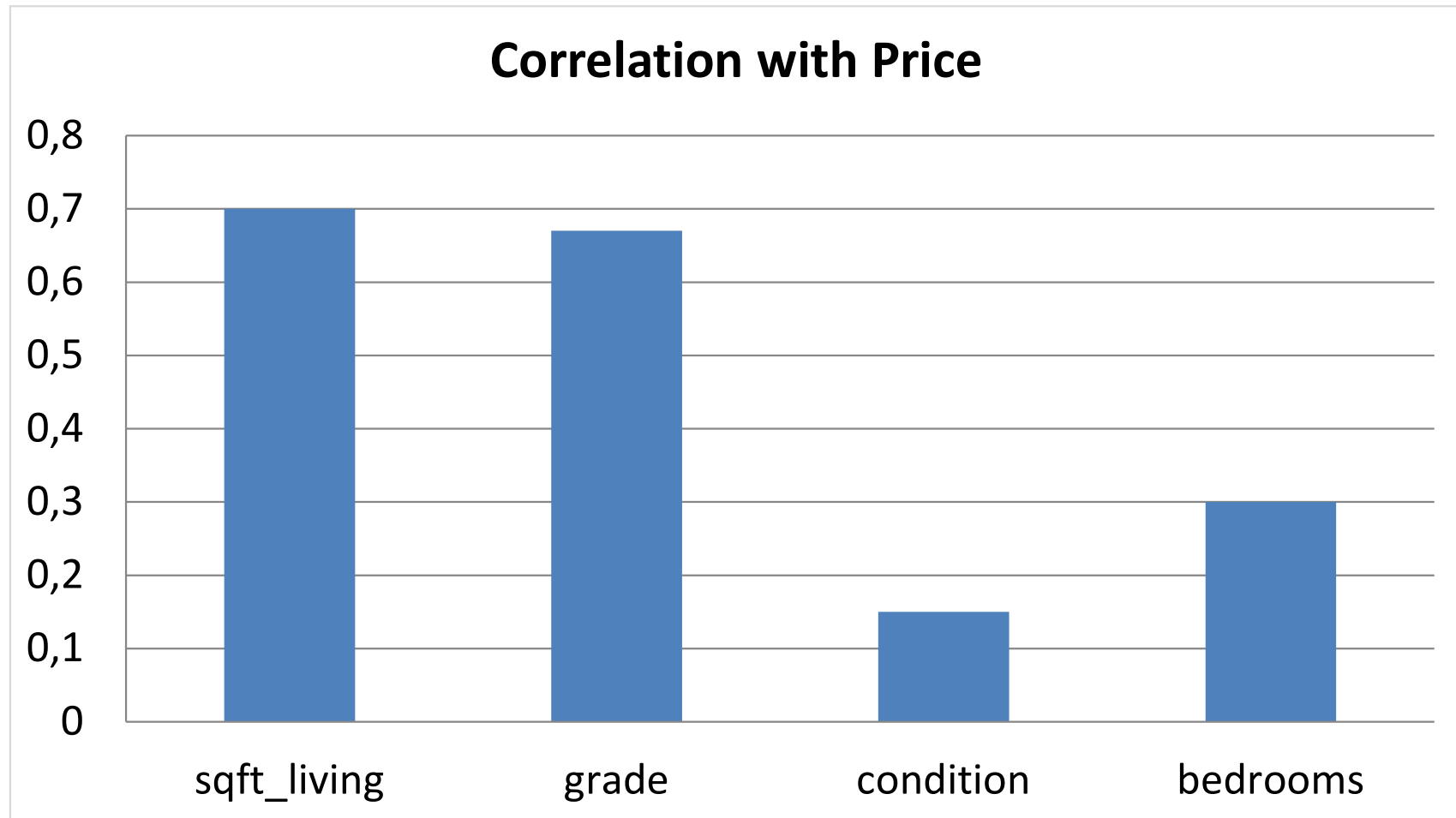
1. Similar house sizes
2. Comparable conditions
3. Best locations for ethical flipping



Correlation Matrix



Correlation Matrix (Key Variables)



Key Findings:

sqft_living \leftrightarrow price \rightarrow strong positive

grade \leftrightarrow price \rightarrow very strong

condition \leftrightarrow price \rightarrow weak

\rightarrow Condition upgrades = cheap leverage

Insights for the Client

- Renovation provides measurable financial uplift
- Zipcode selection is more important than house condition
- Southern King County presents ethical investment opportunities

Recommendations

- Target homes in condition 2–3
- Focus on 1,200–2,500 sqft houses
- Prioritize undervalued southern zipcodes
- Avoid luxury renovations and waterfront properties

Conclusion

- EDA supports all three hypotheses
- Data-driven strategy aligns profitability with social impact
- Further modeling could improve investment predictions