House Price Analysis

- Sarvag Dillikar (51189345)

About Dataset

- This dataset contains house sale prices for **King County**, which includes Seattle.
- It includes homes sold information and column names such as : 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 with **21,613** rows and **21** columns.

Out[2]:

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0	...	7	1180	0	1955	0	98178	47.5112	-122.257	
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0	...	7	2170	400	1951	1991	98125	47.7210	-122.319	
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0	...	6	770	0	1933	0	98028	47.7379	-122.233	
3	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0	...	7	1050	910	1965	0	98136	47.5208	-122.393	
4	1954400510	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0	...	8	1680	0	1987	0	98074	47.6168	-122.045	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
21608	263000018	20140521T000000	360000.0	3	2.50	1530	1131	3.0	0	0	...	8	1530	0	2009	0	98103	47.6993	-122.346	
21609	6600060120	20150223T000000	400000.0	4	2.50	2310	5813	2.0	0	0	...	8	2310	0	2014	0	98146	47.5107	-122.362	
21610	1523300141	20140623T000000	402101.0	2	0.75	1020	1350	2.0	0	0	...	7	1020	0	2009	0	98144	47.5944	-122.299	
21611	291310100	20150116T000000	400000.0	3	2.50	1600	2388	2.0	0	0	...	8	1600	0	2004	0	98027	47.5345	-122.069	
21612	1523300157	20141015T000000	325000.0	2	0.75	1020	1076	2.0	0	0	...	7	1020	0	2008	0	98144	47.5941	-122.299	

21613 rows × 21 columns

Checking for Null Values, Missing values and Duplicates:

- There were no Null values,
- no Missing values
- neither duplicates were found in the dataset.

```
In [6]: # Check for null values
        df.isnull().sum()
```

```
Out[6]: id              0
        date            0
        price           0
        bedrooms        0
        bathrooms       0
        sqft_living     0
        sqft_lot        0
        floors          0
        waterfront      0
        view            0
        condition       0
        grade           0
        sqft_above      0
        sqft_basement   0
        yr_built        0
        yr_renovated    0
        zipcode         0
        lat             0
        long            0
        sqft_living15   0
        sqft_lot15      0
        dtype: int64
```

```
In [7]: # Check for duplicate rows
        df.duplicated().sum()
```

```
Out[7]: 0
```

```
In [8]: # Check for any missing values in the entire DataFrame
        if df.isnull().values.any():
            print("Missing values found in the dataset.")
        else:
            print("No missing values found.")
```

```
No missing values found.
```

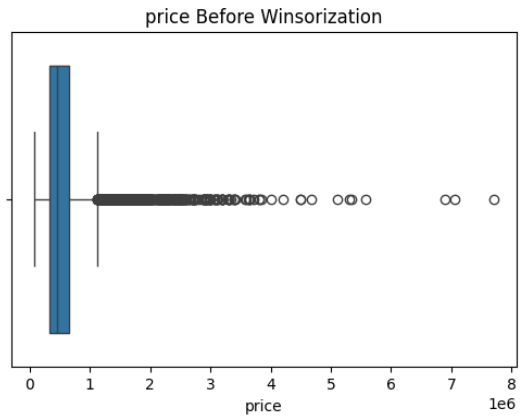
Outlier Detection and Handling:

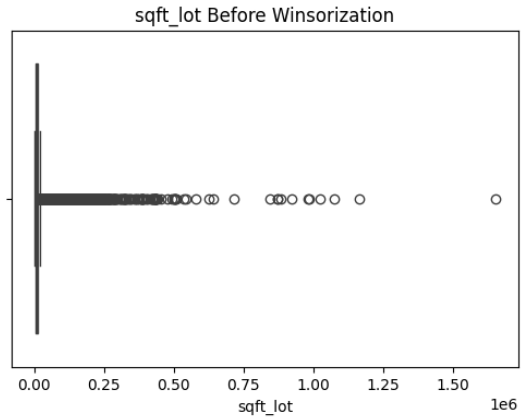
- Identify outliers in numerical columns of the dataset.
- Demonstrate outlier detection using box plots.
- Implement outlier handling using the IQR method and winsorization.

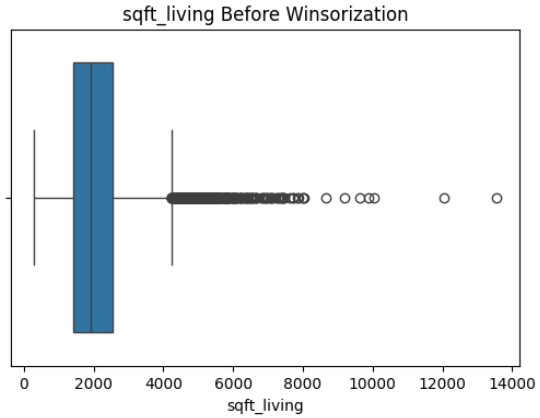
Visualizing Outliers with Box Plots

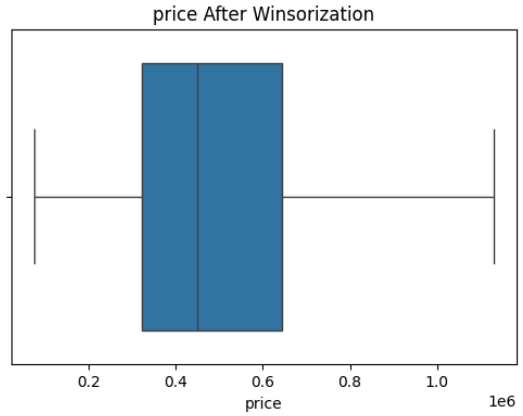
- Box plots visually represent the distribution of data, including quartiles, median, and outliers.
- Outliers are typically represented as individual points beyond the whiskers of the box plot.

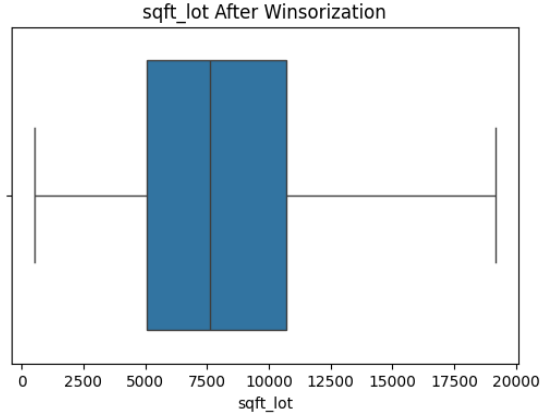
Box Plots:

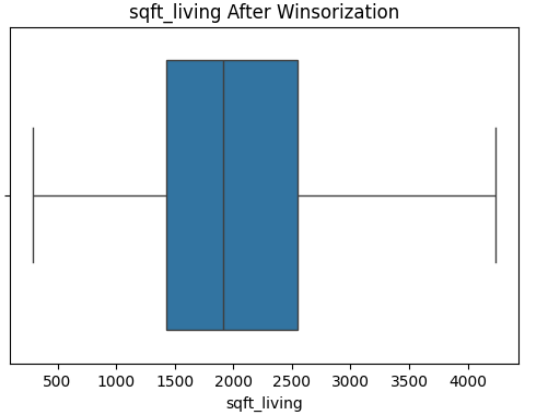

price Before Winsorization
0
1
2
3
4
5
6
7
8
price
1e6


sqft_lot Before Winsorization
0.00
0.25
0.50
0.75
1.00
1.25
1.50
sqft_lot
1e6

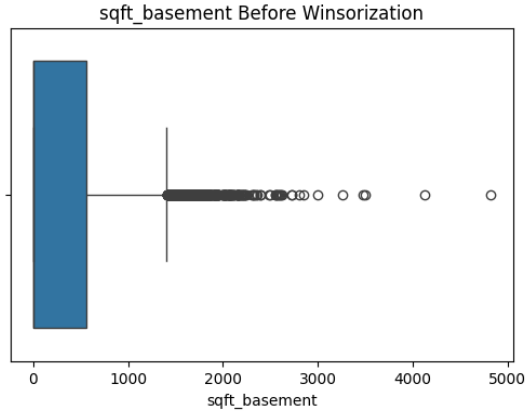

sqft_living Before Winsorization
0
2000
4000
6000
8000
10000
12000
14000
sqft_living

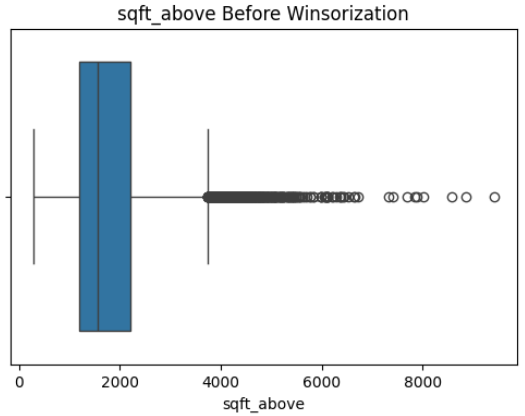

price After Winsorization
0.2
0.4
0.6
0.8
1.0
price
1e6

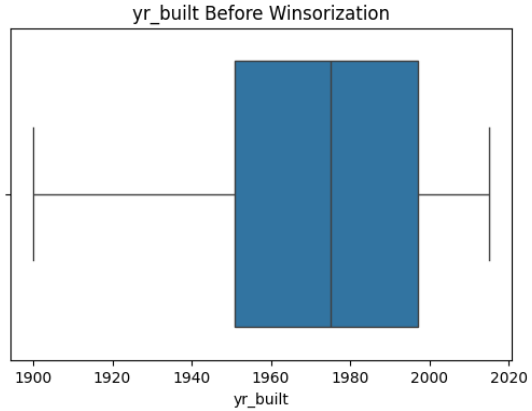

sqft_lot After Winsorization
0
2500
5000
7500
10000
12500
15000
17500
20000
sqft_lot

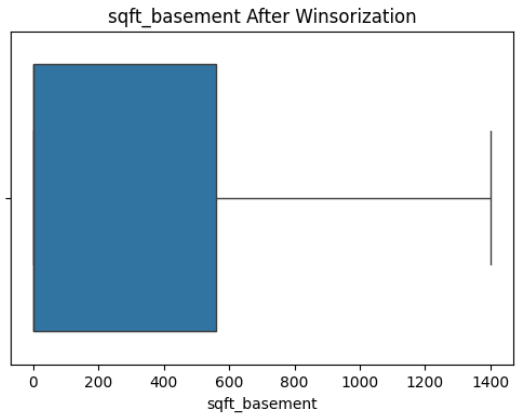

sqft_living After Winsorization
500
1000
1500
2000
2500
3000
3500
4000
sqft_living

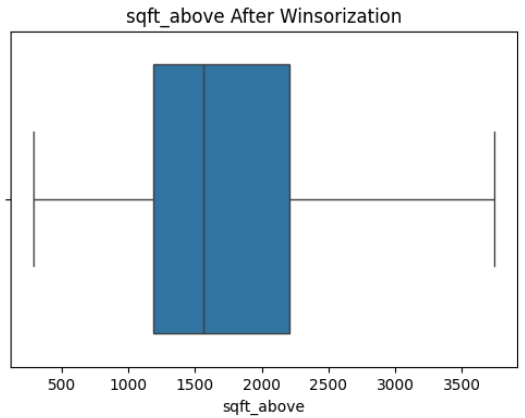
Box Plots:

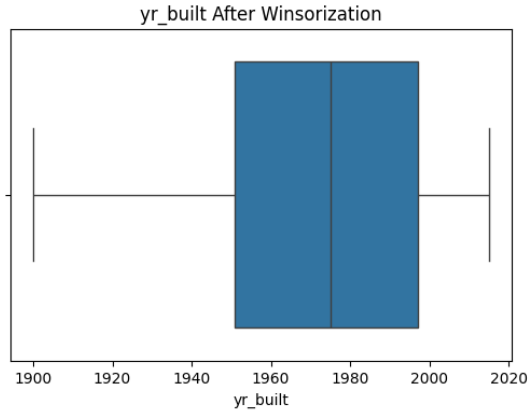

sqft_basement Before Winsorization
0
1000
2000
3000
4000
5000
sqft_basement

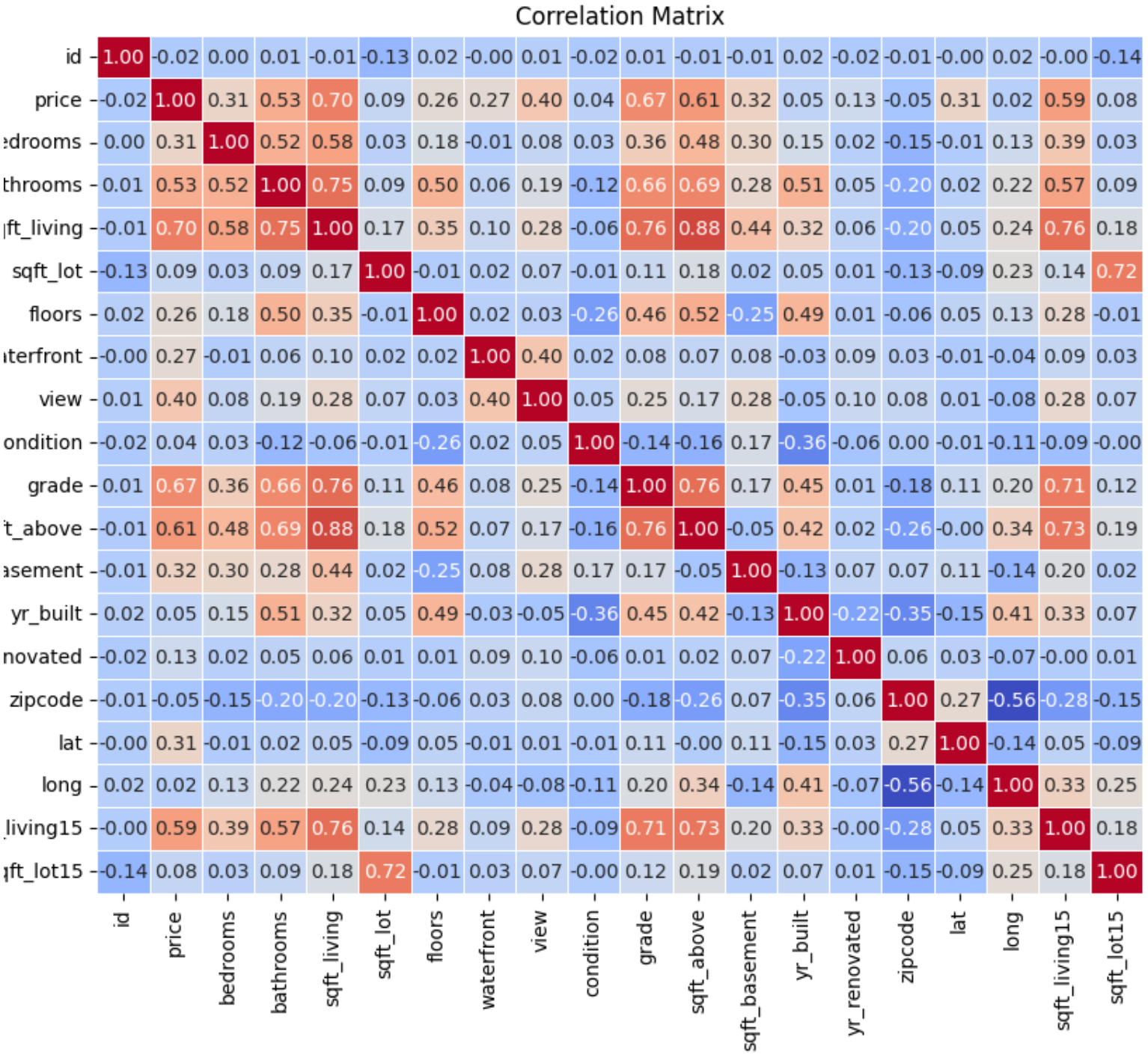

sqft_above Before Winsorization
0
2000
4000
6000
8000
sqft_above


yr_built Before Winsorization
1900
1920
1940
1960
1980
2000
2020
yr_built


sqft_basement After Winsorization
0
200
400
600
800
1000
1200
1400
sqft_basement


sqft_above After Winsorization
500
1000
1500
2000
2500
3000
3500
sqft_above


yr_built After Winsorization
1900
1920
1940
1960
1980
2000
2020
yr_built



Correlation Matrix

- **Key Observations:**

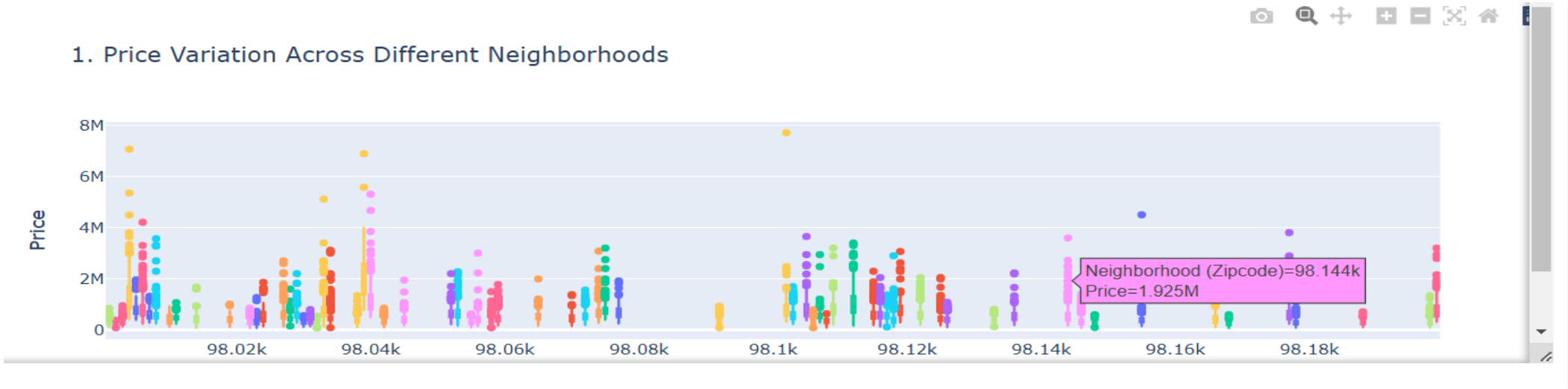
Strong Positive Correlations:

- **Price vs. Square Footage (sqft_living, sqft_above):** There is a strong positive correlation between house price and living space.
- **Bedrooms and Bathrooms vs. Price:** More bedrooms and bathrooms generally lead to higher prices.
- **Grade vs. Price:** Higher-grade houses tend to have higher prices.

Strong Negative Correlations:

- **Condition vs. Price:** Interestingly, there seems to be a slight negative correlation between condition and price.

1. How do house prices vary across different neighborhoods (zipcodes)?



Analysis:

• The boxplot displays the median, IQR, and outliers for house prices by neighborhood.

• Larger boxes indicate more price variation, while whiskers show the price range.

• Outliers point to neighborhoods with extreme price values.

Insights:

• High-price neighborhoods show large IQRs and outliers, while affordable areas have smaller spreads.

2. How does the number of bedrooms affect house prices?

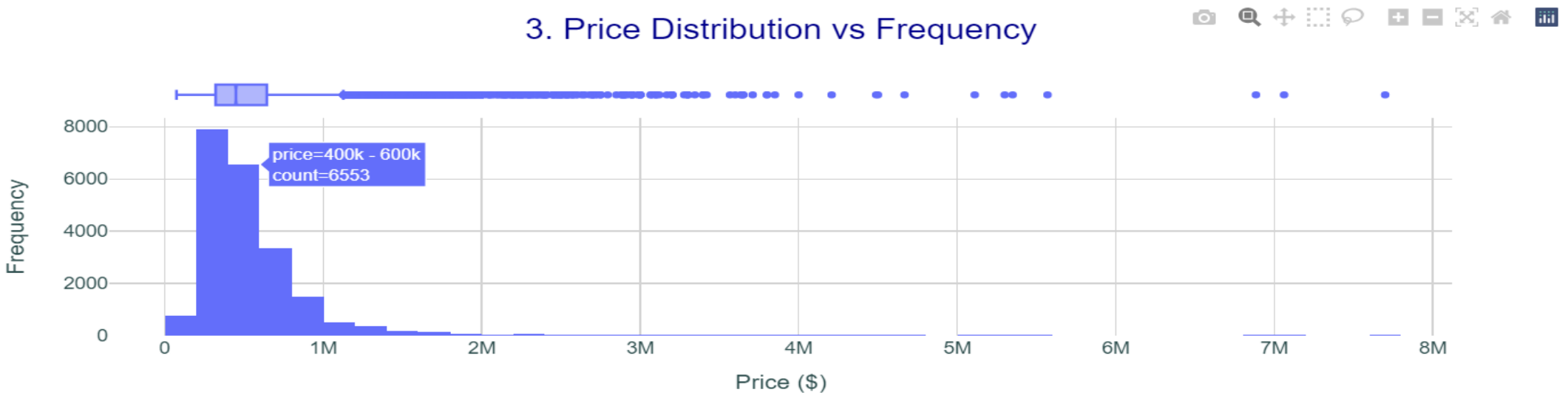
Analysis:

- The scatter plot with a trendline reveals the relationship between the number of bedrooms and price.
- The trendline (OLS) indicates a positive correlation, meaning that as the number of bedrooms increases, so does the price.
- Coloring by number of bedrooms helps highlight the variation in prices across different bedroom counts.

Insights:

- More bedrooms generally lead to higher house prices.

3. What is the distribution of house prices in the dataset?



Analysis:

• The histogram provides a clear view of house price distribution, with marginal box plots for additional insights into the spread and central tendency.

• The price distribution shows a skewed distribution, indicating most houses fall within a certain price range with a few high-priced outliers.

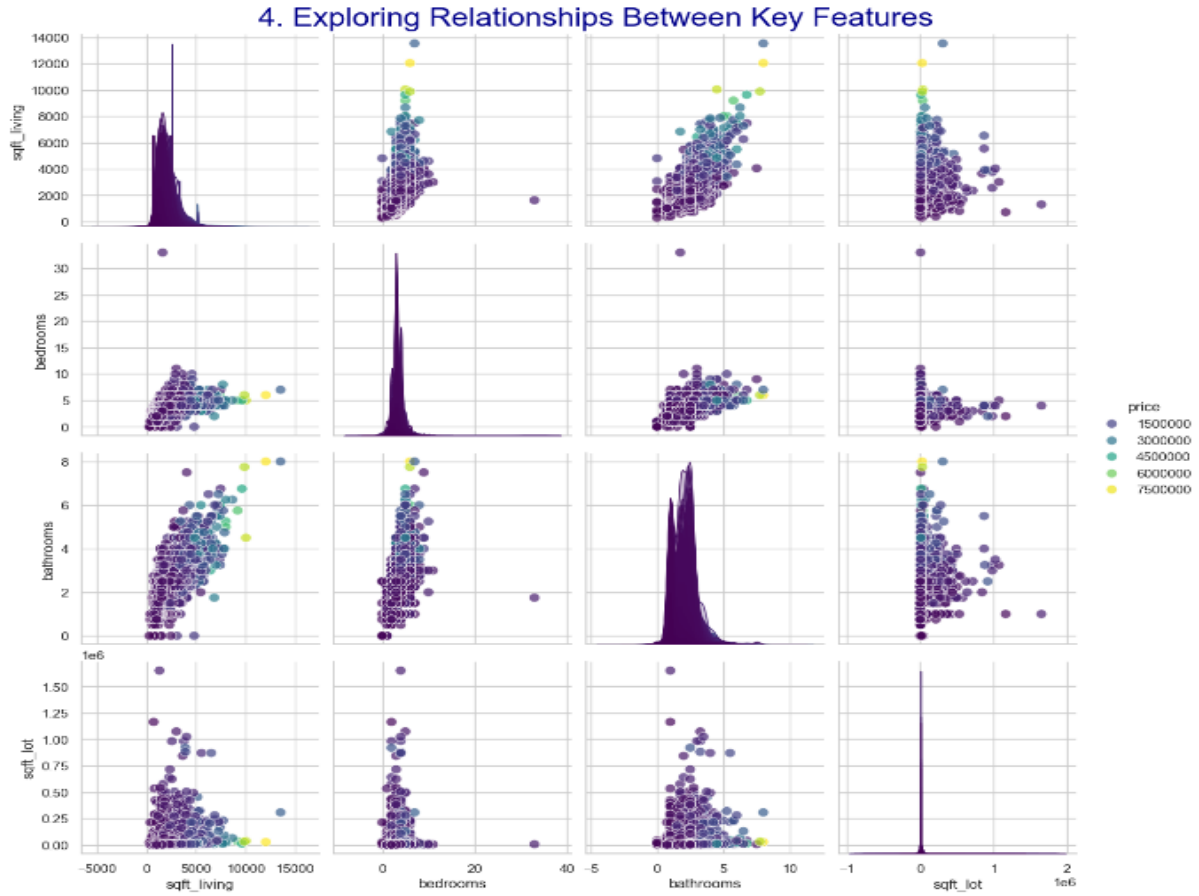
• The use of the marginal box plot provides information on the median price, interquartile range, and presence of outliers.

Insights:

• Most house prices are concentrated around a mid-range value, but the presence of outliers (extremely high-priced homes) skews the distribution.

• The box plot gives insights into how widely prices are dispersed and where the majority of prices lie.

4. What are the relationships between key features in the dataset and how do they correlate with house prices?



Analysis:

- The pairplot visualizes pairwise relationships between multiple variables, with a smooth KDE (Kernel Density Estimate) on the diagonals to show the distribution of each feature.
- The hue based on price allows us to color-code the data points based on house price, revealing how each feature interacts with price.

Insights:

- Some features like **square footage** and **number of bedrooms** show a stronger relationship with price, while others may be less correlated.
- The KDE diagonal plots provide insights into the distribution of key variables and any skewness or trends within them.

5. How does house condition affect the price distribution across different house conditions?

Analysis:

• The violin plot provides a detailed view of the price distribution across various house conditions.
• The hue based on condition distinguishes between the conditions (e.g., good, fair, poor), allowing for easy comparison of how each condition affects price.
• The Violin plot shapes show the distribution and density of prices.

Insights:

• Houses with better conditions tend to have a higher price distribution, while those in poorer condition may have a wider range of prices but lower median values.
• The plot's distribution and density provide deeper insights into variability within each house condition.

6. How does the size (square footage) of a house affect its price?

Analysis:

• This scatter plot helps to visualize how larger homes tend to have higher prices.

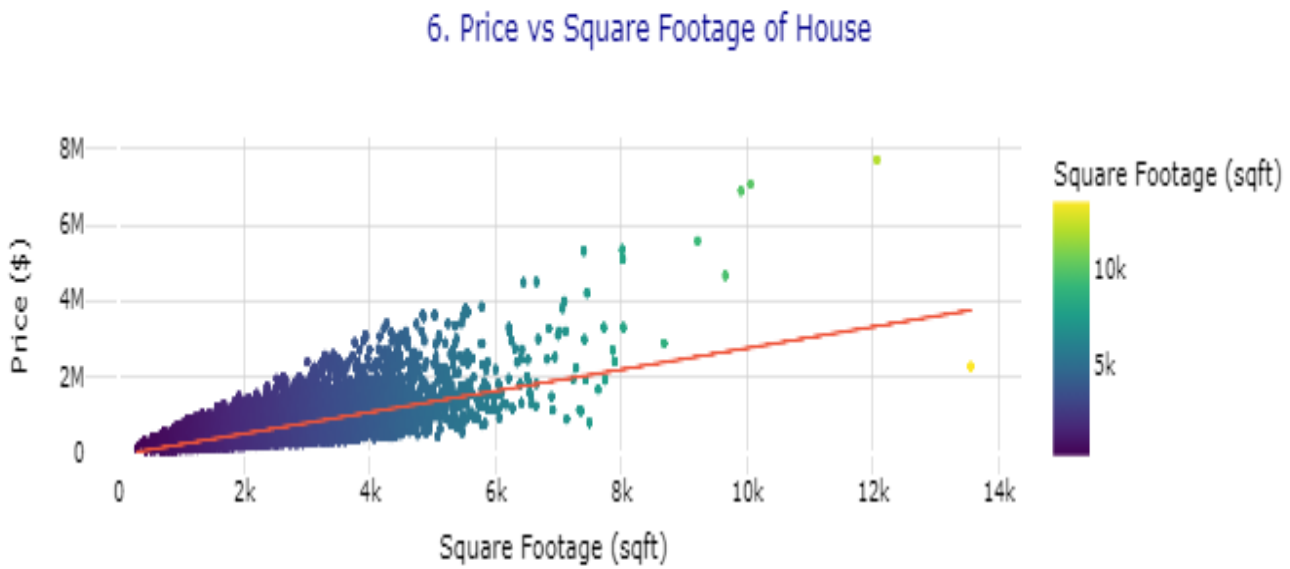
• The trendline provides a clear indication of the positive correlation between square footage and price.

Insights:

• **Positive Correlation:** There's a clear positive relationship between square footage and price, meaning larger homes are typically more expensive.

• **Outliers:** There might be some outliers where small homes are priced higher or large homes are priced lower, possibly due to factors like location, condition, or renovations.

• **Data Distribution:** The color gradient shows a range of square footage values, making it easier to see how larger homes are distributed across the price range.

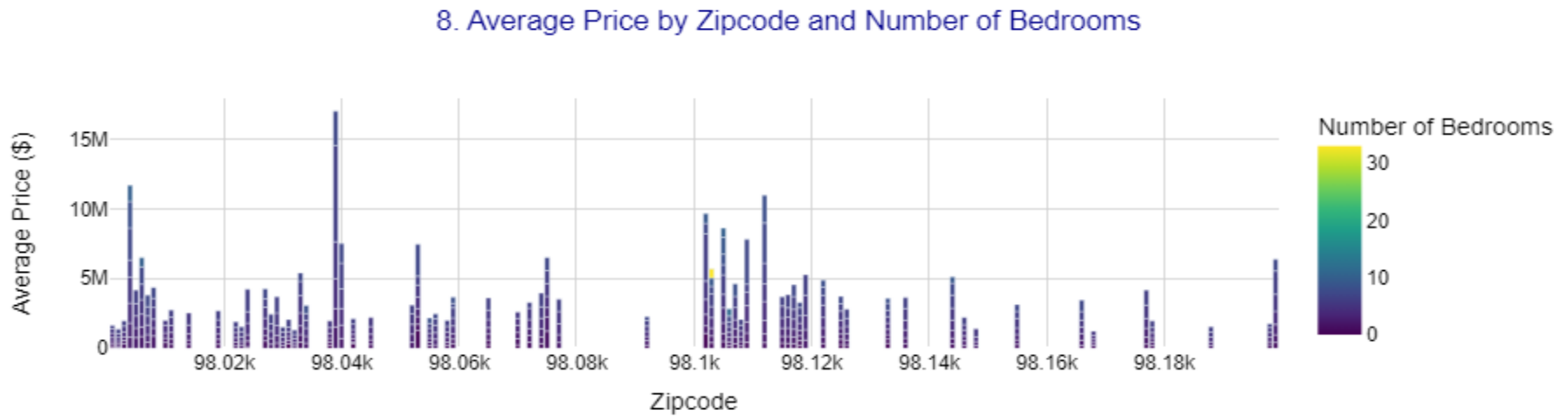


7. How does the year of construction affect house prices?

Analysis:
• Scatter plot shows the relationship between year built and price.
• Color gradient helps differentiate houses by their construction year.
• Modern homes tend to have higher prices due to newer amenities, while older homes are generally priced lower unless renovated.

Insights:
• Newer homes generally fetch higher prices.
• Older homes may be priced lower, with some exceptions (luxury or renovated properties).

8. How does the average house price vary across different zip codes and with respect to the number of bedrooms?



Analysis:

• The bar plot shows how the average price of homes varies by zip code and the number of bedrooms.
• Different colors in the bars represent various bedroom counts, highlighting their impact on home prices across zip codes.
• It helps identify areas where homes with more bedrooms have significantly higher prices.

Insights:

• **Higher Bedroom Count = Higher Price:** Homes with more bedrooms generally have higher average prices.
• **Geographical Influence:** Certain zip codes, likely more affluent or desirable areas, consistently show higher home prices.
• **Price Variation:** Some zip codes show significant price variation, suggesting factors like location and amenities play a major role.
• **Market Trends:** The relationship between price and number of bedrooms indicates demand for larger homes in specific areas.

9. How does proximity to a waterfront affect house prices?

Analysis:

• Waterfront houses (labeled as 'Yes') typically have higher prices compared to non-waterfront houses (labeled 'No').

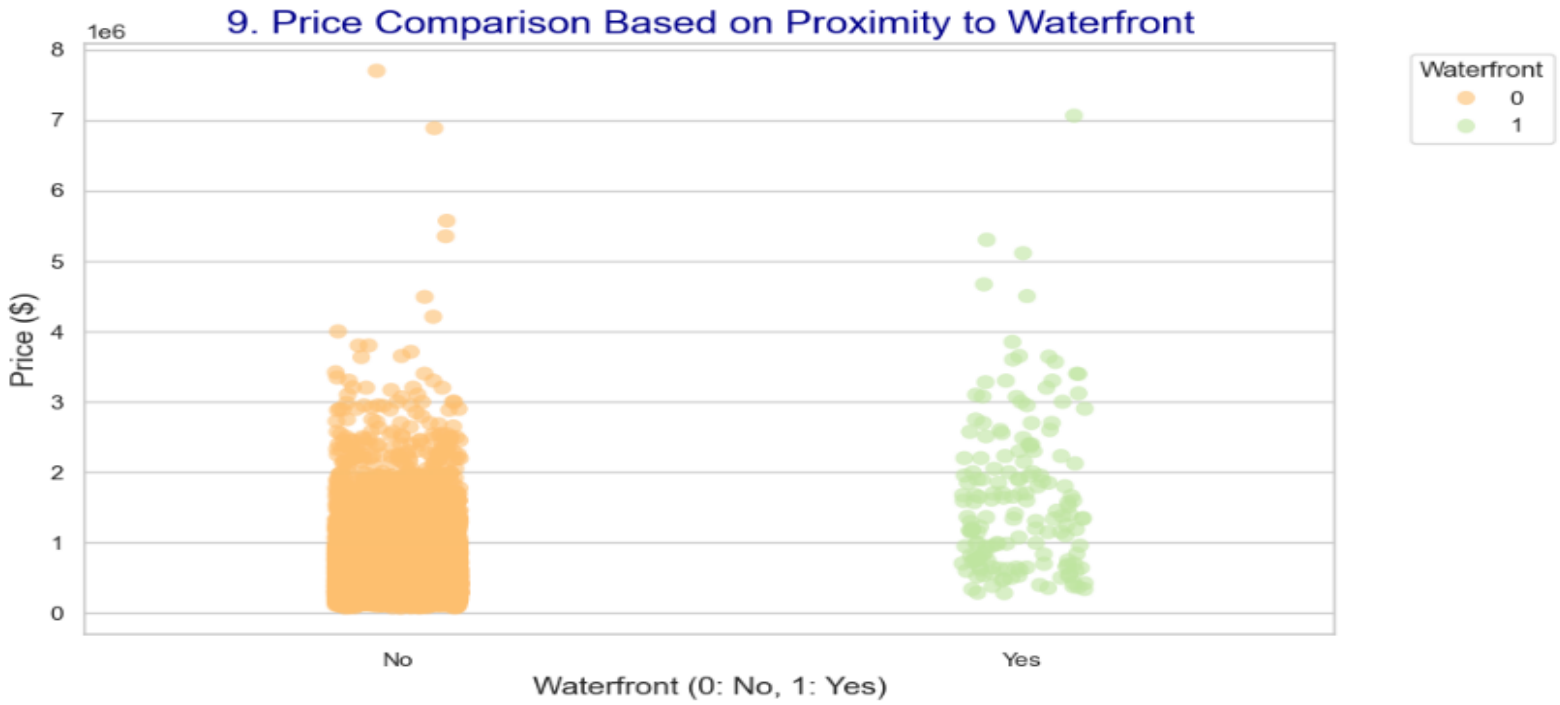
• The plot shows variability in prices, with both waterfront and non-waterfront houses exhibiting a range of values.

Insights:

• Waterfront properties generally command higher prices.

• Price range for waterfront homes is wider, showing both affordable and premium waterfront houses.

• Non-waterfront properties tend to have a lower average price, but can still vary based on other features.



10. How does the number of floors in a house impact its price?

Analysis:

• A heatmap is used to visualize the average price variation across different number of floors in the dataset.

• The pivot table shows the mean price for each number of floors, helping to identify the relationship between floor count and price.

• The color scale indicates price range, with darker shades corresponding to higher prices.

Insights:

• Houses with more floors tend to have higher average prices, though the variation between floors is minimal in some cases.

• Single-floor homes tend to have lower average prices compared to homes with multiple floors.

• The heatmap provides clear visual cues for how floor count is correlated with price.

Machine Learning

- **Model:** Linear Regression
- **Evaluation:** This scatter plot visualizes the model's performance by comparing predicted house prices to their actual values.
- **Ideal Scenario:** Ideally, all data points would fall perfectly on the diagonal line (representing perfect predictions).
- **Key Observations:**
- **Over Fit:** The model appears to over fit - showing a good fit with most points clustering around the ideal line.
- **R-squared:** 0.65 indicates that Interpret R-squared - approximately 65% of the variance in house prices is explained by the model.
- **Mean Squared Error (MSE):** 52585547066.12 quantifies the average squared difference between predicted and actual prices. A lower MSE generally indicates better model accuracy.

```
# Show the plot
plt.legend(loc='upper left')
plt.show()
```

```
Mean Squared Error (MSE): 52585547066.12
R-squared: 0.65
```