6.1: Sourcing Open Data

1. Data source

1.1 Summary of data source

Click here to view: Dataset of House Sales in King County, USA

This dataset includes house sale prices for King County, encompassing Seattle, and contains data on homes sold between May 2014 and May 2015. It is an excellent resource for assessing simple regression models. The dataset contains columns that represent important details of the housing sales market, as shown in the table below.

1.2 The table shows the details from the original dataset.

Check the raw dataset in the attached Excel file on the **"Original Dataset"** worksheet.

Column Name	Description	Variable Type
ID	Unique ID for each home sold	Categorical
Date	Date of the home sale	Categorical
Price	Price of each home sold	Continuous
Bedrooms	Number of bedrooms	Continuous
Bathrooms	Number of bathrooms, where .5 accounts for a room with a toilet but no shower	Continuous
Sqft_living	Square footage of the apartments interior living space	Continuous
Sqft_lot	Square footage of the land space	Continuous
Floors	Number of floors	Continuous
Waterfront	A dummy variable for whether the apartment was overlooking the waterfront or not	Categorical
View	An index from 0 to 4 of how good the view of the property was	Categorical
Condition	An index from 1 to 5 on the condition of the apartment	Categorical
Grade	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high-quality level of construction and design	Categorical
Sqft_above	The square footage of the interior housing space that is above ground level	Continuous
Sqft_basement	The square footage of the interior housing space that is below ground level	Continuous
Yr_built	The year the house was initially bult	Categorical
Yr_renovated	The year of the house's last renovation	Categorical

Column Name	Description	Variable Type
Zipcode	What zipcode area the house is in	Categorical
Lat	Latitude	Continuous
Long	Longitude	Continuous
Sqft_living15	The square footage of interior housing living space for the nearest 15 neighbors	Continuous
Sqft_lot15	The square footage of the land lots of the nearest 15 neighbors	Continuous

1.3 The reason for choosing this dataset

For the past five years, I have been considering whether to buy a house. During this time, I have been researching online and learning about the components and details necessary to make an informed home purchase. After reviewing various datasets from the project brief, I found that this house price dataset suits me well. I am familiar with the details of each column, which gives me confidence that I can complete this task with a better understanding than with other datasets.

Furthermore, since I started researching house prices across the country, I became curious about the factors that cause house prices to vary by region and state. I am interested in understanding what drives these differences and why they exist. I hope to uncover something interesting through this analysis that could help answer my questions.

2. Data Profile

2.1 Data cleaning and consistency checks

Check the **Excel file (name: 1. Cleaned in Excel → Worksheet: Cleaned Dataset)**.

At this step, the data has been cleaned in **Excel**. **"Zipcode Dataset"** worksheet is provided as an additional resource as well for the city/town column.

Check the **Excel file (name: 1. Cleaned in Excel → Worksheet: Zipcode Dataset)**.

Column Name	Action	Function Used
Bedrooms	Removed 1 row with 33 bedrooms	
Date	Changed format from 20141013T000000 → 10/13/2014	DATE()
City/Town	Added a column to map with the provided zipcode column for further clarification for people who are not familiar with the zip codes	VLOOKUP()

2.2 Summary Statistics

This step involves descriptive analysis. Only continuous (numerical) columns will be used in this analysis to perform statistical calculations. These columns include price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, sqft_above, sqft_basement, lat, long, sqft_living15, and sqft_lot15.

Check the **Excel file (name: 1. Cleaned in Excel → Worksheet: Summary Statistics)**.

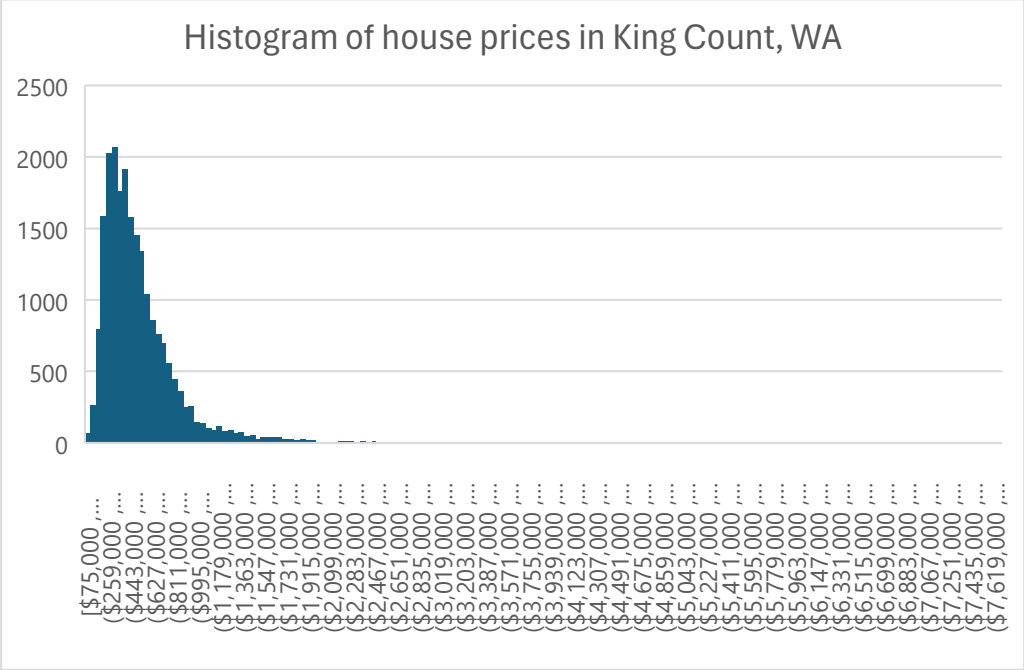

Figure 1: The histogram chart shows house prices in King County, WA

From the figure 1, we can see that the dataset is spread out and difficult to interpret just by looking at it. That is why we need descriptive analysis to summarize and help us understand the main characteristics of the data. Additionally, this will guide us in organizing the dataset, such as eliminating some outliers.

2.3 Identifying Data Types

Some variables need to have their data type changed. This step is starting from **Excel file** and executed in **Python**. I imported the cleaned dataset in **CSV file (name: 2. Cleaned Dataset)** and run the file in **Python**. Check the attached file (**name: 3. Python Source Code**).

Some variables have had their types changed appropriately, with most being converted from integer to string to avoid unnecessary calculations. The yellow highlighted are the columns got changed their datatypes in Python.

Column Name	Original Data Type	Changed to be
ID	Int64	String
Date	String	Datetime
Price	Float64	Int64
Bedrooms	Int64	Int64
Bathrooms	Float64	Float64
Sqft_living	Int64	Int64
Sqft_lot	Int64	Int64
Floors	Float64	Float64
Waterfront	Int64	Int64
View	Int64	Int64
Condition	Int64	Int64
Grade	Int64	Int64
Sqft_above	Int64	Int64
Sqft_basement	Int64	Int64
Yr_built	Int64	Int64
Yr_renovated	Int64	Int64
Zipcode	Int64	String
Lat	Float64	Float64
Long	Float64	Float64
Sqft_living15	Int64	Int64
Sqft_lot15	Int64	Int64
City/Town	String	String

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21612 entries, 0 to 21611
Data columns (total 22 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id             21612 non-null  int64
 1   date           21612 non-null  object
 2   price          21612 non-null  float64
 3   bedrooms       21612 non-null  int64
 4   bathrooms      21612 non-null  float64
 5   sqft_living    21612 non-null  int64
 6   sqft_lot       21612 non-null  int64
 7   floors         21612 non-null  float64
 8   waterfront     21612 non-null  int64
 9   view           21612 non-null  int64
 10  condition      21612 non-null  int64
 11  grade          21612 non-null  int64
 12  sqft_above     21612 non-null  int64
 13  sqft_basement  21612 non-null  int64
 14  yr_built       21612 non-null  int64
 15  yr_renovated   21612 non-null  int64
 16  zipcode        21612 non-null  int64
 17  lat            21612 non-null  float64
 18  long           21612 non-null  float64
 19  sqft_living15  21612 non-null  int64
 20  sqft_lot15     21612 non-null  int64
 21  City/Town      21612 non-null  object
```

Figure 2: Data types from the raw dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21612 entries, 0 to 21611
Data columns (total 22 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id             21612 non-null  object
 1   date           21612 non-null  datetime64[ns]
 2   price          21612 non-null  int64
 3   bedrooms       21612 non-null  int64
 4   bathrooms      21612 non-null  float64
 5   sqft_living    21612 non-null  int64
 6   sqft_lot       21612 non-null  int64
 7   floors         21612 non-null  float64
 8   waterfront     21612 non-null  int64
 9   view           21612 non-null  int64
 10  condition      21612 non-null  int64
 11  grade          21612 non-null  int64
 12  sqft_above     21612 non-null  int64
 13  sqft_basement  21612 non-null  int64
 14  yr_built       21612 non-null  int64
 15  yr_renovated   21612 non-null  int64
 16  zipcode        21612 non-null  object
 17  lat            21612 non-null  float64
 18  long           21612 non-null  float64
 19  sqft_living15  21612 non-null  int64
 20  sqft_lot15     21612 non-null  int64
 21  City/Town      21612 non-null  object
```

Figure 3: Data types after being changed to appropriate types

```
print(df[['id', 'zipcode']].applymap(type))
```

```
                 id        zipcode
0      <class 'str'>  <class 'str'>
1      <class 'str'>  <class 'str'>
2      <class 'str'>  <class 'str'>
3      <class 'str'>  <class 'str'>
4      <class 'str'>  <class 'str'>
...              ...            ...
21608  <class 'str'>  <class 'str'>
21609  <class 'str'>  <class 'str'>
21610  <class 'str'>  <class 'str'>
21611  <class 'str'>  <class 'str'>
21612  <class 'str'>  <class 'str'>
```

Figure 4: Checking if the object datatype is a string datatype

Even though we converted the values in two columns (‘id’ and ‘zipcode’) to strings, the dtype in Python still appears as 'object' because the string datatype is considered a subset of 'object'. Figure 4 above shows that after converting these two columns, the datatype has been successfully changed to string.

2.4 Missing Data Analysis

There are no missing values (null values) in all columns as shown in the Python result after executing the function .info(). This step is executed using **Python**. Check the attached file (**name: 3. Python Source Code**).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21612 entries, 0 to 21611
Data columns (total 22 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id             21612 non-null  object
 1   date           21612 non-null  datetime64[ns]
 2   price          21612 non-null  int64
 3   bedrooms       21612 non-null  int64
 4   bathrooms      21612 non-null  float64
 5   sqft_living    21612 non-null  int64
 6   sqft_lot       21612 non-null  int64
 7   floors         21612 non-null  float64
 8   waterfront     21612 non-null  int64
 9   view           21612 non-null  int64
 10  condition      21612 non-null  int64
 11  grade          21612 non-null  int64
 12  sqft_above     21612 non-null  int64
 13  sqft_basement  21612 non-null  int64
 14  yr_built       21612 non-null  int64
 15  yr_renovated   21612 non-null  int64
 16  zipcode        21612 non-null  object
 17  lat            21612 non-null  float64
 18  long           21612 non-null  float64
 19  sqft_living15  21612 non-null  int64
 20  sqft_lot15     21612 non-null  int64
 21  City/Town      21612 non-null  object
```

Figure 5: No missing values shown in Python

2.5 Duplicate Data Detection

There are no duplicates found. This step is executed using **Python**. Check the attached file (**name: 3. Python Source Code**).

Figure 6: There are no duplicates found in Python

2.6 Data limitations

The data is from 2014-2015, meaning that insights derived from this dataset may not accurately reflect current housing price trends or the realistic housing market conditions in 2024. Additionally, the dataset is specific to King County, Washington, which represents a localized area within the county. While this focuses on a smaller geographic region, to enhance the geographical context, an additional resource has been integrated into the dataset. I have mapped town names from the supplementary resource to the zip codes provided in the original dataset. This allows users to better understand the towns associated with the data, especially for those who may not be familiar with the zip codes.

2.7 Data ethical considerations

- **Data Privacy and Anonymity**

Although the dataset doesn't contain personally identifiable information (PII), the inclusion of zip codes, latitudes, and longitudes could potentially reveal specific property locations when combined with other data, raising privacy concerns if shared publicly.

- **Bias in Data Representation**

The dataset focuses on King County, including Seattle, and may not represent housing markets in other areas. Generalizing results could introduce geographic bias. Additionally, with data from 2014-2015, it may not accurately reflect current market trends, introducing time-based bias in any analysis.

3. Define questions to explore

Key questions

- How does the price vary by zipcode (city/town) in King County, WA?
- What are the top 10 neighborhoods in King County with the highest average home prices?
- What are the top 10 neighborhoods in King County with the lowest average home prices?
- Does the year the house was built impact its price?
- What might the future trend of house prices look like?
- What other factors/variables, aside from those in the dataset, really affect house prices?

Reference

Convert Zipcode to City/Town in King County, WA