Holly Ringsak
Task 6.1 Sourcing Open Data

Data Source

This dataset was acquired through Kaggle.com. It is an example of external data as the data was originally sourced from King County's local government. The data is reliable as it is provided by an official government website.

Data Collection

This dataset is an example of administrative data. It was not obtained by usage or survey data collection. It was obtained by the local government to track local housing prices.

Data Contents

The original data set contains 21 columns of data:

1. Id – Unique identification number given to each home sold
2. Date – Date the house was sold
3. Price – Price the home was sold for
4. Bedrooms – Number of bedrooms in the home
5. Bathrooms – Number of bathrooms in the home
6. Sqft_living – Square footage of the home
7. Sqft_lot – Square footage of the lot
8. Floors – Number of floors in the home
9. Waterfront – Boolean, describes whether or not it is a waterfront property
10. View – Scale from 0-4 of how good the view is
11. Condition – Condition of house scaled 1-5
12. Grade – Scale 1-13 of construction quality, higher grade the better
13. Sqft_above- Square footage above ground level
14. Sqft_basement – Square footage below ground level

15. Yr_built – Year house was built
16. Yr_renov – Year the house was renovated, 0 if never
17. Zipcode – zipcode of house
18. Lat – latitude of address
19. Long – longitude of address
20. Squft_liv15 – average square footage of houses within the closest 15 houses
21. Squft_lot15- Average square footage of the lots within the closest 15 houses

Why Data was Chosen

I chose this dataset because I felt it had an adequate number of columns and rows for analysis. It has a large amount of continuous and categorical variables, as well as a geographical variable, and a time-dependent variable.

Data Profile

Summary Statistics:

Variable	**Time-variant/ invariant**	**Structured/ Unstructured**	**Qualitative/ Quantitative**	**Nominal/ Ordinal/ Discrete/ Continuous**
Id	Invariant	Structured	Qualitative	Nominal
Date	Invariant	Structured	Qualitative	Nominal
Price	invariant	Structured	Quantitative	Continuous
Bedrooms	Invariant	Structured	Quantitative	Discrete
Bathrooms	Invariant	Structured	Quantitative	Discrete
Sqft_liv	Invariant	Structured	Quantitative	Continuous
Sqft_lot	Invariant	Structured	Quantitative	Continuous
Floors	Invariant	Structured	Quantitative	Discrete

Waterfront	Invariant	Structured	Qualitative	Ordinal
View	Invariant	Structured	Qualitative	Ordinal
Condition	Invariant	Structured	Qualitative	Ordinal
Grade	Invariant	Structured	Qualitative	Ordinal
Sqft_above	Invariant	Structured	Quantitative	Continuous
Sqft_basmt	Invariant	Structured	Quantitative	Continuous
Yr_built	Invariant	Structured	Qualitative	Nominal
Yr_renov	Invariant	Structured	Qualitative	Nominal
Zipcode	Invariant	Structured	Qualitative	Nominal
Lat	Invariant	Structured	Qualitative	Nominal
Long	Invariant	Structured	Qualitative	Nominal
Squft_liv15	Invariant	Structured	Quantitative	Continuous
Squft_lot15	invariant	Structured	Quantitative	Continuous

Cleaned Data

No missing or duplicate values in dataset. No mixed data types. The only cleaning that had to be done is an entry error in a row that said a house had 33 bedrooms, so I removed the row.

Limitations and Ethical Considerations

The data only contains information from the years 2014-2015 which is slightly limiting when trying to make future predictions. It could also be insightful to know what the house was listed at versus what it was sold for, and how long the home was on the market.

Questions to Explore

- Are there specific areas/neighborhoods with more expensive housing?
- Are there specific times/seasons when homes sell more?

- Are there specific times/seasons when homes sell for higher/lower than their estimated value?
- What variables have a higher/lower impact on determining the value?
- Does the squft_liv15 variable have a relation to each home's sqft_liv?
- What is the comparison in value of whether or a not homes built around the same time have been renovated?
- What are the most desirable areas to buy a home?