Lisa Matzner

House Sales in King County

2023-01

OVERVIEW & PURPOSE

The goal is to conduct an exploratory visual analysis in Python and find connections between variables that seem worth exploring. After developing hypotheses, various advanced analytical approaches will be used to help test the hypotheses. The results of the analyses will be presented in a Tableau dashboard/storyboard.
Note that not all of the results will fit into the dashboard. Any additional analyses conducted as part of the project will be included in a GitHub repository.

CONTEXT

King County is located in Washington, USA. It is the most populous county in the state and is located in the western part of Washington, surrounding the city of Seattle. The county is known for its diverse geography and long coastline along the Puget Sound. The county is a hub for technology, commerce, and tourism and has a thriving arts and cultural scene, being home to several museums, theaters, and festivals.
King County is one of three Washington counties that are included in the Seattle-Tacoma-Bellevue metropolitan statistical area. About two-thirds of King County's 2.3 Mio population lives in Seattle's suburbs.

OBJECTIVES

To build an interactive dashboard that will visually showcase well-curated results of an advanced exploratory analysis conducted in Python.

DATA SOURCE

Data Content

The data set contains house online sales in King County USA between May 2014 and May 2015. All houses are listed with their unique ID as well as their features, such as number of bedrooms, number of bathrooms, their size in square feet, etc..

Data Sourcing

The data set comes from an external data source and is publicly available open-source data.
The data was downloaded from: https://www.kaggle.com/datasets/harlfoxem/housesalesprediction
Last accessed on 2023-01-25

Data Collection

The data is administrative data that comes from the official public records of home sales in the King County area, Washington State.
https://info.kingcounty.gov/assessor/esales/Residential.aspx

Data Relevance

The data set contains the main data of this analysis and is crucial to answer relevant questions and form hypotheses.

DATA CLEANING

Data Cleaning and Cosistency Checks

Data shape before cleaning: 21613 rows, 21 columns

1. Changing data type for id variable to String
2. Check for mixed data types in variables
3. Check for missing values
4. Check for duplicants
5. Removing entries with absurd values that:
    a. Had an absurd amount of bedrooms (33): 1 entry
    b. Had 0 bedrooms: 13 entries
    c. Had 0 bathrooms: 3 entries

Data shape cleaned data set: 21596 rows, 21 column

DATA LIMITATIONS & ETHICS

The data contains multiple variables that contain subjective values (view, condition, grade). This leaves room for potential bias. It is not specified how the grading is done or what factors play into the score. Potential sellers may have influence on the grading system to improve or worsen the selling price. As the data set is older, there is no additional information on the exact collection method. It could have been scraped from the Kings County website or could have been freely available. As such there is potential for measurement bias.

Additionally there are multiple instances were houses have 0 bed- or bathrooms. Which seems to be an error.

DATA PROFILE

		Data Types			
Variables	**Description**	**time-variant/ -invariant**	**structured/ unstructured**	**qualitative/ quantitative**	**qualitative: nominal/ordinal quantitative: discrete/continuous**
id	Unique ID for each home sold	time-invariant	structured	qualitative	nominal
date	Date of the home sale	time-variant	structured	quantitative	discrete
price	Price of each home sold	time-variant	structured	quantitative	continuous
bedrooms	Number of bedrooms	time-invariant	structured	quantitative	discrete
bathrooms	Number of bathrooms (0.25 = room with only toilet, 0.5 = room with toilet & sink, no shower or bathtub, 0.75 = room with toilet, sink & shower **or** bathtub, 1 = room with toilet, sink, shower **and** bathtub)	time-invariant	structured	quantitative	continuous
sqft_living	Square footage of the apartments interior living space	time-invariant	structured	quantitative	continuous
sqft_lot	Square footage of the land space	time-invariant	structured	quantitative	continuous
floors	Number of floors (0.5 = home with a partial (second) floor added)	time-invariant	structured	quantitative	discrete
waterfront	Wether or not the house was overlooking the waterfront (0 = no waterfront view, 1 = waterfront view)	time-invariant	structured	qualitative	binary
view	An index from 0 to 4 of how good the view of the property was (0 = No view, 1 = Fair 2 = Average, 3 = Good, 4 = Excellent)	time-invariant	structured	qualitative	ordinal

condition	An index from 1 to 5 on the condition of the apartment (1 = Poor- Worn out, 2 = Fair- Badly worn, 3 = Average, 4 = Good, 5= Very Good)	time-variant	structured	qualitative	ordinal
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.	time-invariant	structured	qualitative	ordinal
sqft_above	The square footage of the interior housing space that is above ground level	time-invariant	structured	quantitative	continuous
sqft_basement	The square footage of the interior housing space that is below ground level	time-invariant	structured	quantitative	continuous
yr_built	The year the house was initially built	time-invariant	structured	quantitative	discrete
yr_renovated	The year of the house's last renovation (0 = not renovated)	time-variant	structured	quantitative	discrete
zipcode	What zipcode area the house is in	time-invariant	structured	quantitative	discrete
lat	Lattitude	time-invariant	structured	quantitative	continuous
long	Longitude	time-invariant	structured	quantitative	continuous
sqft_living15	The square footage of interior housing living space for the nearest 15 neighbors	time-variant	structured	quantitative	continuous
sqft_lot15	The square footage of the land lots of the nearest 15 neighbors	time-variant	structured	quantitative	continuous

QUESTIONS TO EXPLORE

- Does the sale price of houses vary with location?
- Are there any patterns in the time of year when houses are sold?
- How does the size of a house relate to its sale price?
- Are houses with more bedrooms and bathrooms generally more expensive?
- Does having a waterfront property affect a house's sale price?
- How does the view of the property affect a house's sale price?
- Does the condition and grade of the house affect its sale price?
- Does the year the house was built and last renovated affect its sale price?
- How does the square footage of the interior living space of a home compare to that of its nearest 15 neighbors?
- How does the grade of a home relate to the square footage above and below ground level?
- How does the number of floors in a home relate to the grade level?

ADDITIONAL LINKS

https://info.kingcounty.gov/assessor/esales/Glossary.aspx#v
King County Online Glossary for Property Sales

https://info.kingcounty.gov/assessor/esales/Residential.aspx
King County eSales Property Search

https://www.kaggle.com/datasets/harlfoxem/housesalesprediction
Kaggle - House Sales in King County, USA by harlfoxem

https://gis-kingcounty.opendata.arcgis.com/datasets/zipcodes-for-king-county-and-surrounding-area-shorelines-zipcode-shore-area/explore?location=47.505388%2C-121.477600%2C8.81
Zipcodes for King County and Surrounding Area (Shorelines) / zipcode shore area
Data last accessed on 2023-02-01