Task 6.1

Advanced Analytics & Dashboard Design

Euikyu Kim

Oct 21st, 2023

DATA SOURCE: House Sales in King County, USA

Data Source

- This is an external data source.
- The dataset, 'House Sales in King County, USA' contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

Data Collection

- It is provided by the "Kaggle" (https://www.kaggle.com/harlfoxem/housesalesprediction) which is a data science competition platform and online community of data scientists and machine learning practitioners under Google LLC.
- I would say this was manually collected because, unlike the usage data collecting method, it is necessary to collect, combine, and classify data collected through a number of agents from different real estate companies. In addition, incomplete data such as errors, non-responses, etc. are cases that can't be handled automatically.

Data Contents

- The dataset, 'House Sales in King County, USA' contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.
- Brief descriptions of columns.

Column Name	Description
id	Unique identification per house sale
date	Date the house sale
price	Price of house sale in currency of USD
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms
sqft_living	Square footage of the house's interior living space
sqft_lot	Square footage of the land space
floors	Number of floors(level) of the house

waterfront	An index to indicate if the house was overlooking the waterfront or not. 0 represents no waterfront, and 1 represents waterfront.
view	An index from 0 to 4 of how good the view of the property was. 0 represents no good view, and 4 represents a very good view.
condition	An index from 1 to 5 on the condition of the house. 1 represents the poorer condition, and 5 represents the superb condition.
grade	An index from 1 to 13. 1 to 3 falls short of building construction and design, 7 has an average level of construction and design, and 11 to 13 has a higher quality level of construction and design.
sqft_above	The square footage of the interior housing space that is above the ground level
sqft_basement	The square footage of the interior housing space that is below the ground level
yr_built	The year of house built
yr_renovated	The year of the house's last renovation
zipcode	The zip code is the postal code to indicate the area the house is in
lat	Latitude
long	Longitude
sqft_living15	The average square footage of interior housing living space for the nearest 15 neighboring houses
sqft_lot15	The average square footage of land space for the nearest 15 neighboring houses

Data Relevance

- This data is useful to compare prices of houses with waterfront and without waterfront and the views of the house with waterfront and without waterfront.
- It's a great dataset for evaluating simple regression models which fits the best with predicting the prices.

Data Limitation

- Some of the variables needed to be log-transformed to satisfy regression assumptions, and any new data used with the model would have to undergo similar preprocessing.
- Additionally, given regional differences in housing prices, the model's applicability to data from other counties may be limited.

DATA PROFILE

Data Cleaning

Column	Type of Inconsistency	Action
date	Unnecessary texts 'T000000' (e.g. '20141013T000000')	Removed only 'T000000' after the date.
sqft_above sqft_basement lat long sqft_living15 sqft_lot15	Unnecessary variables for this analysis.	Dropped the columns.
bedrooms ('id' = 2402100895)	An extreme outlier suspected of typos.	Replaced the value '33' to '3'

Understanding Data

Column Name	Qualitative/Quantitative	Discrete/Continuous	Nominal/Ordinal/Binary
id	Qualitative	Discrete	Nominal
date	Quantitative	Discrete	Nominal
price	Quantitative	Continuous	Nominal
bedrooms	Quantitative	Discrete	Nominal
bathrooms	Quantitative	Continuous	
sqft_living	Quantitative	Continuous	
sqft_lot	Quantitative	Continuous	
floors	Quantitative		
waterfront	Quantitative	Discrete	Binary
view	Quantitative	Discrete	Ordinal
condition	Quantitative	Discrete	Ordinal
grade	Quantitative	Discrete	Ordinal
yr_built	Quantitative	Continuous	
yr_renovated	Quantitative	Continuous	
zipcode	Qualitative	Discrete	Nominal

	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view
count	2.161300e+04	2.161300e+04	21613.000000	21613.000000	21613.000000	2.161300e+04	21613.000000	21613.000000	21613.000000
mean	4.580302e+09	5.400881e+05	3.369454	2.114757	2079.899736	1.510697e+04	1.494309	0.007542	0.234303
std	2.876566e+09	3.671272e+05	0.907964	0.770163	918.440897	4.142051e+04	0.539989	0.086517	0.766318
min	1.000102e+06	7.500000e+04	0.000000	0.000000	290.000000	5.200000e+02	1.000000	0.000000	0.000000
25%	2.123049e+09	3.219500e+05	3.000000	1.750000	1427.000000	5.040000e+03	1.000000	0.000000	0.000000
50%	3.904930e+09	4.500000e+05	3.000000	2.250000	1910.000000	7.618000e+03	1.500000	0.000000	0.000000
75%	7.308900e+09	6.450000e+05	4.000000	2.500000	2550.000000	1.068800e+04	2.000000	0.000000	0.000000
max	9.900000e+09	7.700000e+06	11.000000	8.000000	13540.000000	1.651359e+06	3.500000	1.000000	4.000000

	condition	grade	yr_built	yr_renovated	zipcode
count	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000
mean	3.409430	7.656873	1971.005136	84.402258	98077.939805
std	0.650743	1.175459	29.373411	401.679240	53.505026
min	1.000000	1.000000	1900.000000	0.000000	98001.000000
25%	3.000000	7.000000	1951.000000	0.000000	98033.000000
50%	3.000000	7.000000	1975.000000	0.000000	98065.000000
75%	4.000000	8.000000	1997.000000	0.000000	98118.000000
max	5.000000	13.000000	2015.000000	2015.000000	98199.000000

DATA LIMITATION AND ETHICS

Limitations

Sample Bias	The dataset represents house sales in King County, USA. It may not be representative of the entire real estate market, as it only covers a specific geographical area. This could limit the generalizability of insights drawn from the data.
Data Completeness	The dataset may have missing or incomplete data, which could affect the quality of analysis and modeling. It's important to address missing values appropriately.
Data Quality	The accuracy and reliability of the data are crucial. Errors or inaccuracies in the dataset can lead to incorrect conclusions or predictions.
Temporal Relevance	Real estate data can quickly become outdated, and the dataset may not reflect current market conditions. It's important to consider the dataset's temporal relevance when making predictions or drawing insights.
Privacy Concerns	The dataset may contain sensitive or personally identifiable information (PII) about property owners and buyers. Care must be taken to ensure that privacy laws and ethical standards are followed when handling and sharing such data.

Ethical Considerations

Privacy and Anonymity	When using real estate data, it's important to protect the privacy and anonymity of individuals involved in the transactions. Researchers and data analysts should not attempt to identify specific property owners or buyers.
Bias and Fair Housing	Analyzing real estate data should be done in a way that avoids perpetuating or amplifying biases related to race, gender, or other protected characteristics. Ethical considerations include ensuring fair housing practices are upheld.
Transparency	It's essential to be transparent about data sources and methodologies. Sharing details about how the data was collected and any potential limitations is important for ethical data usage.
Consent	When using data related to property sales, it's important to ensure that data usage and sharing comply with consent and data protection laws. Property owners and buyers may not have consented to their data being used for research or analysis.
Data Security	Data security is crucial to prevent unauthorized access and data breaches. Ensuring the security of the dataset is an ethical responsibility.
Equitable Use	The use of real estate data should be equitable and not harm vulnerable populations. Ethical considerations include avoiding predatory practices or discriminatory behavior in real estate transactions.
Legal Compliance	Ensure compliance with all applicable laws and regulations, including those related to real estate, data protection, and privacy.

DEFINING QUESTIONS TO EXPLORE

1) What is the distribution of house prices in King County, USA?
2) How do types of houses impact the prices?
3) How does the value of the waterfront or view affect house prices?
4) Are any specific ZIP codes in King County known for higher or lower house prices?
5) What is the average size of living space and lot and how does it correlate with the price?
6) How have house prices been influenced by grades or conditions?
7) What is the average age of houses in the dataset, and does that affect prices?