TBD*

TBD

Ziheng Zhong

April 8, 2024

TBD

Table of contents

*Code and data are available at: https://github.com/iJustinn/House_Price.git

Table 1: Summary statistics of the California housing dataset

Table 2: Count of missing values for each variable

1 Introduction

2 Data

Data used in this paper was cleaned, processed and tested with the programming language R (R Core Team 2022). Also with support of additional packages in R: `tidyverse` (Wickham et al. 2019), `ggplot2` (Wickham 2016), `janitor` (Firke 2023), `readr` (Wickham, Hester, and Bryan 2023), `knitr` (Xie 2014), `modelsummary` (Arel-Bundock 2023), `testthat` (Wickham Year of publication), `KableExtra` (Zhu 2023), `viridis` (Garnier et al. 2018), `lubridate` (Grolemund and Wickham 2021), `maps` (Deckmyn et al. 2021), `mgcv` (Wood 2021).

Table 3: Count of missing values for each variable after cleaning

Table 4: Modeling Results for Linear Models

	Multiple Regression	Polynomial Regression
(Intercept)	−13 969 628.019	212 976.031
	(395 357.616)	(1359.110)
Year	6958.769	
	(196.466)	
NumBedroom	59 897.327	
	(993.637)	
poly(Year, 2)1		3 866 485.706
		(105 994.679)
poly(Year, 2)2		1 601 816.793
		(105 993.208)
poly(NumBedroom, 2)1		6 591 537.312
		(105 993.313)
poly(NumBedroom, 2)2		1 365 656.076
		(105 994.574)
Num.Obs.	6082	6082
R2	0.445	0.479
R2 Adj.	0.445	0.479
AIC	158 396.4	158 018.2
BIC	158 423.3	158 058.5
Log.Lik.	−79 194.199	−79 003.096
F	2440.891	1397.712
RMSE	109 331.51	105 949.60

Table 5

	GAM Regression
(Intercept)	212 976.031
	(1322.680)
Num.Obs.	6082
R2	0.506
AIC	157 697.2
BIC	157 801.7
RMSE	103 028.50

Modeling Results for Non-linear Models

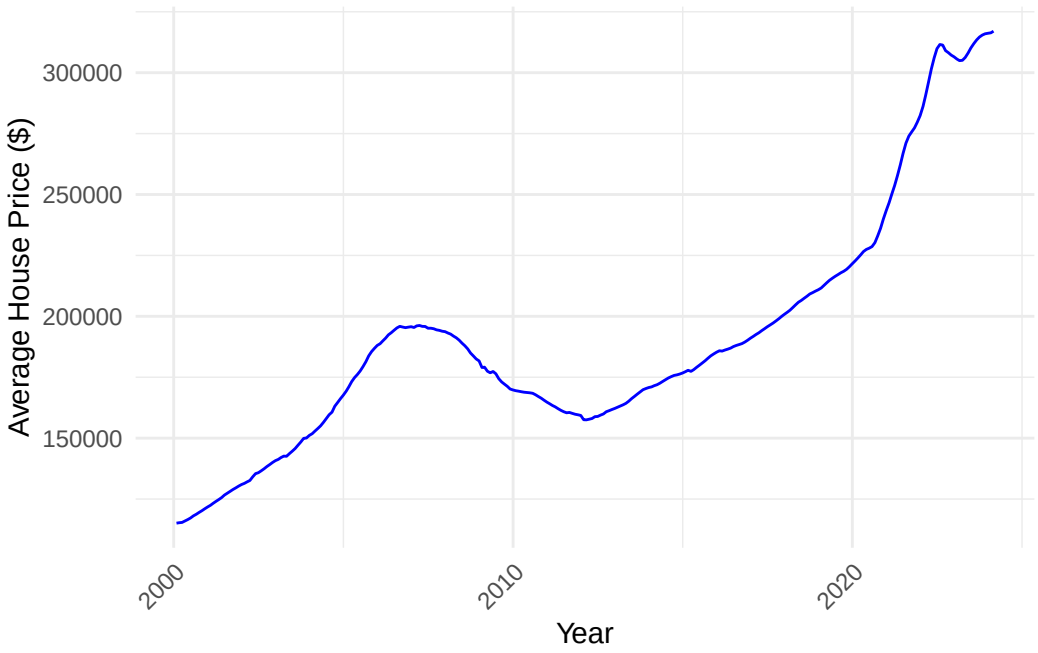


Figure 1: Trend of Average House Price from 2000 to 2024

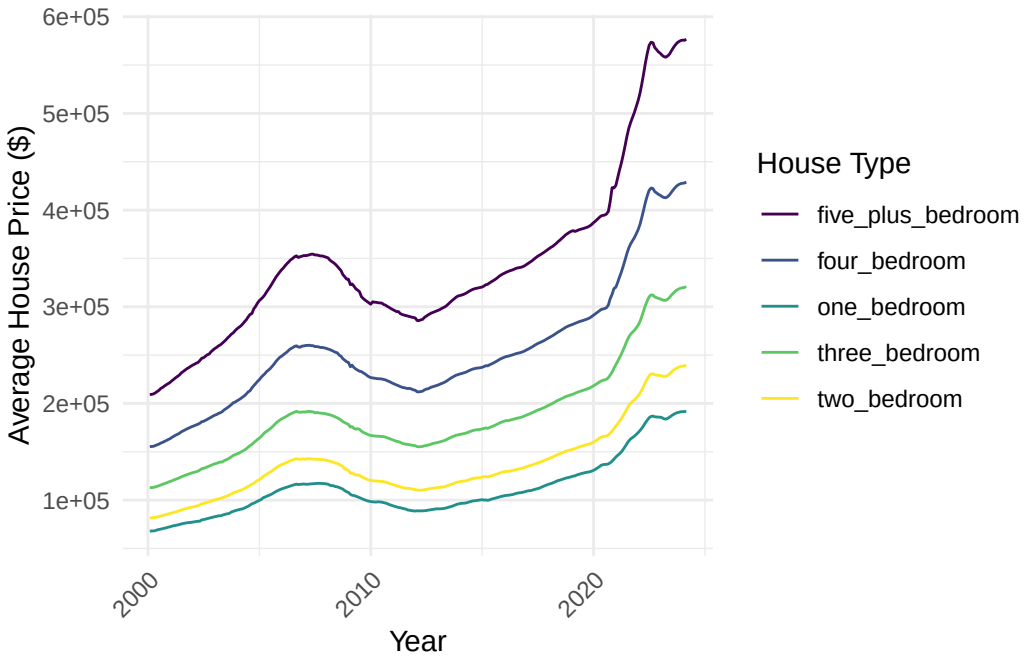


Figure 2: Trend of Average House Price from 2000 to 2024 by House Type

2.1 Source

2.2 Method

3 Results

3.1 Data Trend

3.2 Heat Maps

3.3 Modeling

4 Discussion

4.1 Demographic Shifts

4.2 Health-related Behaviors

4.3 Government Policies

4.4 Environmental Changes

4.5 Possible Improvements

5 Conclusion

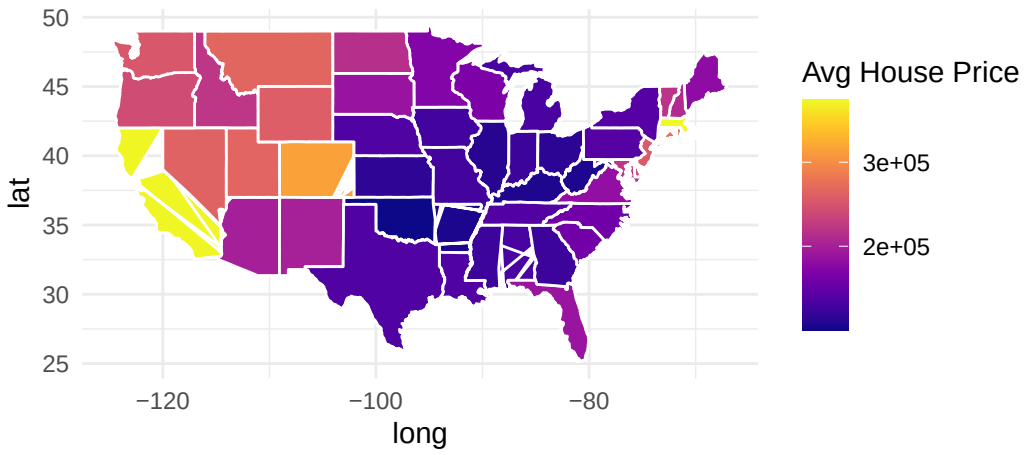


Figure 3: Average Price by State in the US for All House Types

A Appendix

A.1 Datasheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

    - The datasets are created to support a wide range of real estate analysis and research by Zillow Research group. These datasets typically serve purposes such as market trend analysis, price prediction, housing supply studies, and economic impact assessments. They are designed to fill the gap for comprehensive, accurate, and accessible real estate data for researchers, policymakers, and the general public interested in the housing market dynamics.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The datasets available on the research page of Zillow, a leading real estate and rental marketplace, are typically created by Zillow's own economic research team. This team focuses on providing insights into the housing market and economic trends.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

    - The datasets on Zillow's research page are generally created and funded internally by Zillow Group, Inc. itself, without specific external grants. As a commercial entity with a vested interest in real estate markets, Zillow utilizes its resources to compile and analyze these datasets for public and internal use.

4. *Any other comments?*

    - None.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - In the dataset used for this paper, each instance representing a geographic region. These regions mainly include the country as a whole (United States), metropolitan statistical areas (e.g., New York, NY; Los Angeles, CA). The dataset contains a series of monthly average house prices, spanning from January 2000 to February 2024, as shown by the date-formatted columns. The data does not mix different types of instances (like movies, users, and ratings or people and interactions between them) but focuses solely on regional housing price data over time. Each row includes identifiers and names for the regions, the type of region (such as 'country' or 'msa' for metropolitan statistical area), and state names for regions within specific states.

2. *How many instances are there in total (of each type, if appropriate)?*

    - 50 States
    - 895 Regions
    - 290 Dates

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - The dataset is likely a sample from a larger set of all possible geographic areas in the United States for which housing data could be collected. This larger set would include a comprehensive collection of all geographic regions in the United States, encompassing every city, town, rural area, and Metropolitan Statistical Area (MSA) with available housing market data. A clear example indicating that this is just a

sample is the absence of data for Hawaii. Regarding representativeness, this data is representative, mainly because Zillow is one of the leading real estate trading companies in the United States, and its data covers almost all cities. Additionally, there is no special focus on regions that provide the most insights into the national housing market trends.

4. *What data does each instance consist of? "Raw" data or features? In either case, please provide a description.*

    - RegionID: A unique identifier for each geographic region.

- SizeRank: A rank based on the size or significance of the region, presumably in terms of population or housing market activity.
- RegionName: The name of the region, which can be a country (like the United States), a metropolitan statistical area (MSA), or possibly other region types.
- RegionType: The type of region, such as 'country' or 'msa', indicating the scope or level of the geographic area.
- StateName: The name of the state in which the region is located, applicable to regions within specific states.
- Monthly Average House Prices: A series of columns representing the average house prices for each month, spanning from January 2000 to February 2024. These columns are named by date (e.g., '2000-01-31', '2000-02-29', etc.), with each column representing the average housing price in that region for the given month.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - RegionID: Unique ID associated with each region in RegionName.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

    - Information such as the latitude and longitude for each region is not included. The absence of geographic coordinates (latitude and longitude) means that while the dataset provides a comprehensive temporal view of housing price trends across different regions, it does not directly offer spatial data that would allow for mapping or spatial analysis of these trends. The reason for missing information is because of the dataset's primary focus on economic trends and prioritizes time-series analysis of housing prices.

7. *Are relationships between individual instances made explicit? If so, please describe how these relationships are made explicit.*

    - The relationships between individual instances (regions) are not explicitly defined in terms of direct interactions or connections between them. The dataset primarily focuses on time-series data for housing prices within various geographic regions

without detailing explicit relationships like proximity, hierarchical structures (e.g., how states are composed of multiple cities), or economic interdependencies among these regions.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - There are detailed datasets for houses with specific number of bedrooms. Thus no need to split anything.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - None.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - This is a self-contained dataset.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - None.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - None.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified.*

    - The data are labled with StateName and RegionName. Therefore, data from each state or region coule be view as a subpopulations of this dataset.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - None.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - None.

16. *Any other comments?*

    - None.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how. How were these mechanisms or procedures validated?*

    - The data in the Zillow dataset, specifically the housing prices for various regions, is primarily derived from Zillow's own listings and transactions data, along with public records and assessments. Here's a breakdown of the data acquisition methods:
        - Direct Observation and Public Records: Zillow aggregates housing price data from several sources, including direct listings on their platform, real estate transactions, and public property records such as sales and assessments. This means that much of the data is directly observable or comes from official records, making it a robust and reliable source of housing market information.
        - Zestimate: Some of the housing price data, particularly for times or areas where direct transaction data may be sparse, could be supplemented by Zillow's Zestimate® home values. The Zestimate is an estimated market value calculated using proprietary models that analyze public and user-submitted data. This would fall under data indirectly inferred/derived from other data.
        - Validation and Verification: For directly observed or recorded data (listings, transactions, public records), the validity comes from the data's official or commercial nature. These are factual records of housing sales and listings. For data like the Zestimate, Zillow continuously updates and refines its models based on new data, market trends, and feedback. The accuracy of Zestimates is evaluated by comparing estimated values with actual sale prices when they become available, and Zillow publishes accuracy metrics for its Zestimates, providing a form of validation.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)?*

- Mainly direct obervation (using data from their platform, as mentioned in previous question).

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

   - Not mentioned by data provider.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

   - Zillow's Platform and Automated Systems: Much of the data collection is carried out through Zillow's own technological infrastructure, which aggregates and processes listings, sales data, and property information from across the United States. This process is automated, involving sophisticated software systems designed to handle large volumes of data.
   - Real Estate Professionals: Realtors and other real estate professionals often use Zillow to list properties. While their primary motivation is to market properties to potential buyers, their contributions add to the dataset's comprehensiveness. Compensation for these professionals comes through the real estate transactions facilitated by their listings, not directly from Zillow for the data per se.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances? If not, please describe the timeframe in which the data associated with the instances was created.*

   - The dataset provided spans from January 2000 to February 2024, indicating that the data collection covers this specific timeframe. This period reflects the creation timeframe of the data associated with each instance, which means each instance's housing price data was collected or estimated for these specific monthly intervals.

6. *Were any ethical review processes conducted? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - For datasets like the one provided by Zillow, which compiles housing market data across various regions in the United States, specific ethical review processes might not be as prominently documented or required in the same manner as they would be for research involving human subjects directly. However, Zillow, like many data-providing entities, operates within a framework of legal and ethical considerations, especially regarding privacy, data accuracy, and transparency.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - Data was obtained from Zillow.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - This is a publicly available information on housing prices and transactions.

9. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - None

10. *Any other comments?*

    - None.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Cleaning:
        - Columns (SizeRank, RegionID, RegionType, RegionName) are removed from the dataset, likely because they are not necessary for the subsequent analysis focused on state-level price trends.
    - Further Processing:
        - Monthly Aggregation: The data is grouped by StateName, and the mean house price for each state is calculated for each month. This step involves aggregating all numeric columns (assumed to be monthly house price data) and computing their means, excluding missing values.
        - Yearly Aggregation: The monthly data is then transformed to calculate the average house price by year for each state. This involves converting month columns to a long format, extracting the year from each date, and then calculating the yearly mean house price for each state.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - The raw data was saved in the project folder [https://github.com/iJustinn/House_Price.git], specific location identified in the README section.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - Programming langugae R [https://cran.r-project.org/] in the IDE RStudio [https://www.rstudio.com/products/rstudio/download/].

4. *Any other comments?*

    - None.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - xxx

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

    - xxx

3. *What (other) tasks could the dataset be used for?*

    - xxx

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

    - xxx

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

    - xxx

6. *Any other comments?*

    - xxx

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - xxx

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - xxx

3. *When will the dataset be distributed?*

    - xxx

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - xxx

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - xxx

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - xxx

7. *Any other comments?*

    - xxx

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - xxx

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - xxx

3. *Is there an erratum? If so, please provide a link or other access point.*

    - xxx

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

- xxx

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

   - xxx

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

   - xxx

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

   - xxx

8. *Any other comments?*

   - xxx

References

Arel-Bundock, Vincent. 2023. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. https://vincentarelbundock.github.io/modelsummary/.

Deckmyn, Alex, Original S code by Richard A. Becker, Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka, and Alex Deckmyn. 2021. *Maps: Draw Geographical Maps*. https://CRAN.R-project.org/package=maps.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Garnier, Simon, Noam Ross, Bob Rudis, and Marco Sciaini. 2018. *Viridis: Default Color Maps from 'Matplotlib'*. https://CRAN.R-project.org/package=viridis.

Grolemund, Garrett, and Hadley Wickham. 2021. *Lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. Year of publication. *Testthat: Get Started with Testing*. https://CRAN.R-project.org/package=testthat.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

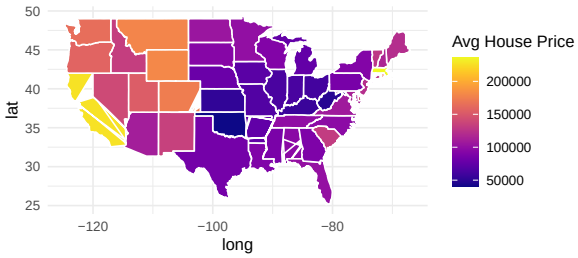
Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

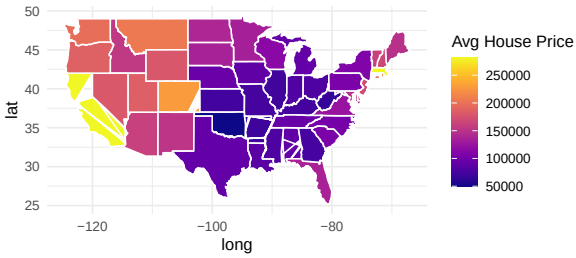
Wood, Simon N. 2021. *Mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*. https://CRAN.R-project.org/package=mgcv.

Xie, Yihui. 2014. *Knitr: A Comprehensive Tool for Reproducible Research in R*. Edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

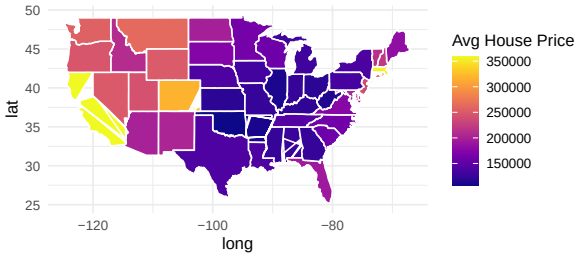
Zhu, Hao. 2023. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.



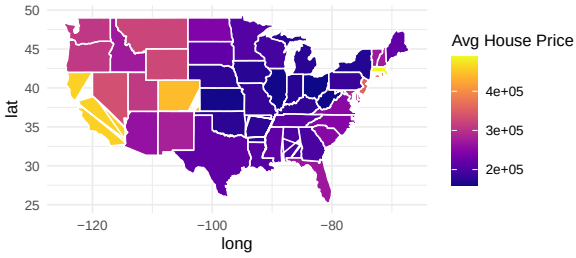
(a) one bedroom



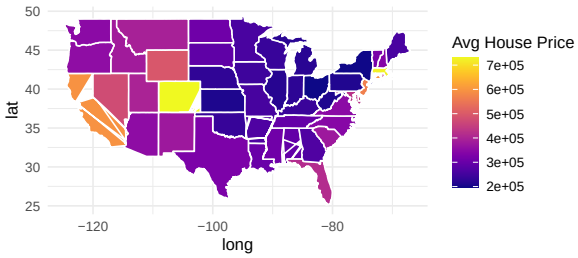
(b) two bedrooms



(c) three bedrooms



(d) four bedrooms



(e) five plus bedrooms

Figure 4: Average Price by State in the US for Different House Types

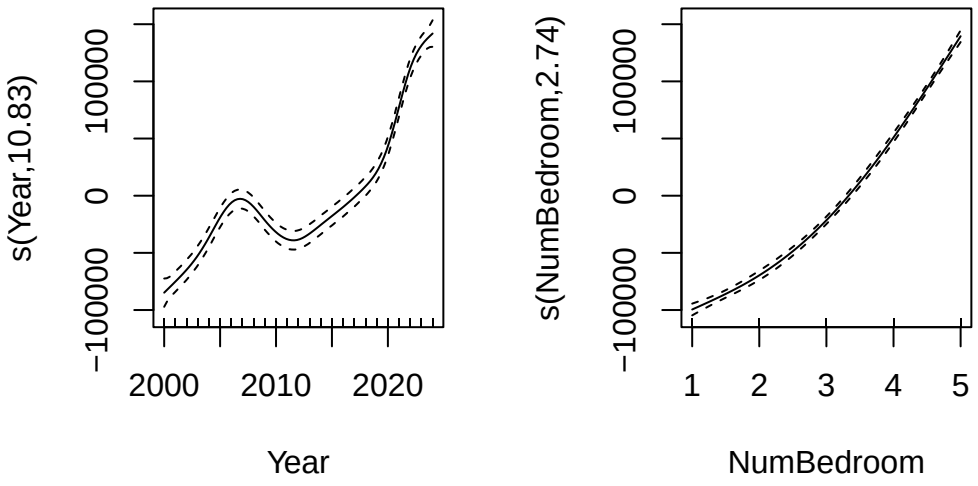

s(Year,10.83)
−100000
0
100000
2000
2010
2020
Year
s(NumBedroom,2.74)
−100000
0
100000
1
2
3
4
5
NumBedroom

Figure 5