Datasheet for California Housing Prices Dataset*

Shuyuan Zheng

April 22, 2024

For the purpose of this research, the datasheet is constructed to provide a detailed view of the California Housing Market dataset, motivated by questions from Gebru et al. (2021) and analyzed using the open source statistical programming language R (R Core Team 2023). This dataset focuses on socio-economic and geographical variables influencing housing prices in California, particularly utilizing data from the 1990 census(“Key2STATS” 2024).

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.*
    - Created for educational purposes, this dataset provides an introduction to machine learning with practical data cleaning requirements. It fills the gap for a real-world dataset that is complex enough for meaningful analysis yet manageable for beginners.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was featured by R. Kelley Pace and Ronald Barry in their paper and further popularized by Aurélien Géron in his machine learning book.
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.*
    - There is no specific funding information provided for the creation of the dataset as it was compiled for academic and educational use.
4. *Any other comments?*
    - N/A

[*] Code and data are available at: https://github.com/EAsUpluckYX/House-Prices.git.

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.*

   - The instances represent individual records of housing data from various California districts as captured in the 1990 census. Variables include geographical coordinates, house age, room counts, and economic demographics.

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

   - The dataset contains 20,433 instances, each representing a housing unit within California.

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).*

   - It is a comprehensive dataset derived from a specific subset of the 1990 California census data, aimed at covering diverse geographic and socio-economic conditions across California.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

   - Each instance consists of structured data with features including longitude, latitude, housing median age, total rooms, total bedrooms, population, households, median income, median house value, and ocean proximity.

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

   - The target variable in the dataset is 'medianhousevalue', which is used for predicting housing prices based on other features.

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.*

   - Some instances have missing values, particularly in the 'total_bedrooms' feature, requiring preprocessing steps like imputation before use in machine learning models.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

- The dataset primarily focuses on individual housing data without explicit relationships between instances, although spatial relationships can be inferred through geographical features.

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.*

   - N/A

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

   - N/A

10. *Any other comments?*

    - N/A

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.*

- The data were directly collected from the 1990 California census, ensuring accuracy and relevance to the geographic and economic contexts of the time. Given its source, the data is considered reliable and was likely validated by the census bureau through their standard processes for data collection and verification.

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

- The data collection was executed through census surveys, which involve both manual data entry by census workers and automated data processing systems. These systems undergo rigorous testing and validation to ensure accuracy and reliability of the data.

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

- As a subset of the larger California census data, the sampling was probabilistic, designed to accurately represent the demographic and economic conditions across various districts within the state.

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)?*

   - Americans in general were involved in the data collection process. The participants were not paid for the survey.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

   - The data is collected from the year 1990.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? 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.*

   - N/A

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

   - The data is obtained from key2stats.com.

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.*

   - Yes.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

   - Yes.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - N/A

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.*

- N/A

2. *Any other comments?*

   - It requires rudimentary data cleaning, has an easily understandable list of variables and sits at an optimal size between being to toyish and too cumbersome.

Uses

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

   - This dataset has primarily been used for educational purposes, particularly in teaching basic concepts of machine learning and data science. It serves as a practical example in the book 'Hands-On Machine Learning with Scikit-Learn and TensorFlow' by Aurélien Géron, where it is employed to demonstrate how to implement various machine learning algorithms.

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.*

   - https://www.key2stats.com/data-set/view/1597

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

   - Legislation, policy-making, researching, and educating.

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?*

   - N/A

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

   - N/A

6. *Any other comments?*

   - N/A

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.*

- The dataset is publicly available and can be distributed freely as it is under the Public Domain (CC0) license.

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

   - The dataset is distributed on key2stats. https://www.key2stats.com/data-set/view/1597

3. *When will the dataset be distributed?*

   - N/A

Maintenance

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

   - Key2STATS was built using technology and resources developed by National Science Foundation funding under NSF Award Numbers 0937989, 1418163, and 1742083.

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

   - NA

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

   - No.

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)?*

   - The dataset will be updated when research and new knowledge of labeling and instances are introduced at the time.

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.*

   - All the data collected in 1990

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.*

   - Old version of the California Housing dataset available from:Luís Torgo's page (University of Porto)

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.*

    - No.

8. *Any other comments?*

    - N/A

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

"Key2STATS." 2024. *Key2STATS*. National Science Foundation. https://www.key2stats.com/data-set/view/1597.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.