# Datasheet for 'raw_data.csv'*

Sameeck Bhatia

December 3, 2024

Extract of the questions from Gebru et al. (2021).

**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 dataset was created to allow real estate agents and the public to view current property listings in the city of Seattle.
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 created by the real estate agents of Seattle on behalf of their brokerage firms, who, in turn, created it for the National Association of Realtors.
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 creation of the dataset has been funded by real estate brokerage firms who pay for appraisal services.
4. *Any other comments?*
    - The observations in the dataset were collected by appraisers and reported to real estate agents and brokers, who have then entered this data into the Multiple Listing Service.

**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.*

*Code and data are available at https://github.com/SameeckBhatia/Seattle-Real-Estate

- Each instance represents a unique property that has been listed for sale.

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

   - There are 1403 instances/listings.

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 a sample of a larger set. The population is all current property listings of all types (e.g. townhouse, land, multi-family, etc.) in Seattle.

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 features that define the property. There features range from the number of days the property has been up on the market to geographic coordinates.

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

   - There is a label associated for every instance. The labels (variables in the data) have been outlined in `paper.qmd` in the appendix.

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.*

   - There are some instances missing that are labeled by `ADDRESS`, `LOCATION`, `SQUARE FEET`, `YEAR BUILT`, `HOA/MONTH`. The instances in `ADDRESS` and `LOCATION` are related and are missing due to the address of the listings being kept private. One instance of `SQUARE FEET` is missing and is likely due to no appraiser measuring the size yet. A few instances in `YEAR BUILT` are missing as the information has likely not been recorded. Lastly, many instances of `HOA/MONTH` are missing as most single-family homes do not require homeowner's association fees.

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.*

   - There are no explicit relationships between individual instances.

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.*

- One recommended split it having training data cover the outliers in price so that a model trained on it does not underfit the testing data. There is no specific proportion for a train-test split.

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

    - There are no errors and no duplicates in the dataset.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? 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.*

    - The dataset is not self-contained. All instances have a link to the individual listing's website. There is a guarantee that the link will exist in the future as the data source (Redfin) ensures the listing is updating to reflect its status (e.g. active, sold, sold conditional). There are no official archives of the 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.*

    - There is no data that is confidential as there is no personal identifiable information published onto the Multiple Listing Service.

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.*

    - The dataset contains no instances that might be offensive of threatening.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - The dataset does not contain data on a sub-population.

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.*

    - It is not possible to identify individuals with any instance in the dataset as there is no PII (Personal Identifiable Information).

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.*

    - The dataset contains no instances that are sensitive.

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.*

    - The data was directly observable by the appraiser who reported each instance.

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 was manually collected by an appraiser, a human that measures each feature 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)?*

    - The dataset is a sample of a larger set and it was collected using non-probabilistic sampling. The dataset was sampled by the most recent listings on the MLS.

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)?*

    - Appraisers, real estate agents/brokers, and the seller of each property was involved. The appraiser was employed by either the seller of the broker and collected the data with consent of the seller. There is no information on the exact amount each appraiser was paid, however, in Seattle they are typically paid $600 - $800 per appraisal.

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 timeframe for data collection is unknown as the appraisal could have been conducted on any day and the information could have been entered in the MLS database on another.

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.*

   - The American Society of Appraisers usually conducts an ethical review for the data collection methods and reviews each appraiser's compliance with the regulations.

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 was collected via a third party that collects it from the Multiple Listing Service database (Redfin).

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.*

   - There are no individuals in question for the dataset.

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.*

   - There are no individuals in question for the dataset.

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).*

    - There are no individuals in question for the dataset.

11. *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.*

    - Analysis on this particular dataset and its impact has not been conducted as there are no particular subjects in question.

12. *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.*
    - There was no preprocessing done on the data as it is a raw dataset.
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.*
    - Not applicable.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - Not applicable.
4. *Any other comments?*
    - None.

**Uses**

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - This particular dataset has not been used for any tasks. However, similar versions of the dataset (from Realtor.com and Zillow) has been used for time-series price prediction analysis.
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.*
    - There is no one repository that links to the system that uses the dataset as it is consumed by multiple users.
3. *What (other) tasks could the dataset be used for?*
    - The dataset can be used to identify areas where a potential buyer would want to purchase their property.
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?*

- The dataset contains some information regarding the terms of use that is not directly relevant to the instance of the data for analysis.

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

   - The dataset should not be used for advertising campaigns and non real-estate activities such as reposting active listings and overloading the MLS system.

6. *Any other comments?*

   - None.

**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 will be distributed to real estate management and brokerage companies as a form of collaboration.

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

   - There is no information on how the dataset will be distributed.

3. *When will the dataset be distributed?*

   - There is no information on when the dataset will be distributed.

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.*

   - There is no information on the copyright that the dataset will be distributed under.

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.*

   - There are no IP-based restrictions on the data.

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.*

- There is one export control. The control is the condition that the individual extracting the data has a Redfin account.

7. *Any other comments?*
    - None.

**Maintenance**

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The dataset will be mainted by Redfin as the company is constantly extracting data from the MLS database.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - The owner of the dataset cannot be contacted as the contact information is not publicly available.
3. *Is there an erratum? If so, please provide a link or other access point.*
    - There is no erratum.
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. It is updated every day on the Redfin website.
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.*
    - The dataset does not relate to people.
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.*
    - Older versions of the dataset will be kept in a separate database and can be accessed by request from the source (Redfin).
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.*

- There is definitely a way to build on the dataset. The dataset has been collected at a given point in time (cross-sectional data) and any data collected after that point is a natural addition to the dataset.

8. *Any other comments?*

   - None.

## 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.