Datasheet for the City of Austin, Texas '2020 Racial Profiling (RP) dataset' (Austin Police Department 2023a)*

Benny Rochwerg

April 18, 2024

These questions are 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 purpose for which this dataset was created is unclear. However, the dataset's website states that "the data provided are for informational use" (Austin Police Department 2023a). It is unclear as to whether a specific task was in mind or whether a specific gap needed to be filled. Despite this, the dataset comprises 68,552 entries (Austin Police Department 2023b) on 2020 Austin Police Department traffic stop-based arrests, warnings or field observations, and citations (Austin Police Department 2023a).

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

- It is unclear as to who created the dataset, but the owner of the dataset is the Austin Police Department (Austin Police Department 2023a). It is unclear as to whether this dataset was created on behalf of an entity. However, this dataset was made available on the City of Austin's Open Data Portal, which is run by the Communications and Technology Management Department's Open Data Team (City of Austin, Texas, n.d.a). The City of Austin's Open Data Portal's website indicates that public data usage is supported in part to make the government more transparent (City of Austin, Texas, n.d.a).

*Code employed in the creation of this datasheet can be found here: https://github.com/bennyrochwerg/profiling

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.*

- It is unclear as to whether the creation of the dataset was funded based on the dataset's website (Austin Police Department 2023a).

4. *Any other comments?*

- No.

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.*

- According to the dataset's website, "Each row is a Motor Vehicle Stop" (Austin Police Department 2023a). This suggests that there are not multiple types of instances.

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

- In total, the dataset comprises 68,552 entries (Austin Police Department 2023b) on 2020 Austin Police Department traffic stop-based arrests, warnings or field observations, and citations (Austin Police Department 2023a).

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 unclear as to whether the dataset contains all instances or if it is a sample; the dataset's website states that "The data provided are for informational use only and may differ from official APD crime data" (Austin Police Department 2023a).

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 data that includes the motor vehicle stop type, sex, race, stop reason, street form, whether a search was performed, the foundation of the search if applicable, what the search located, the stop outcome, the foundation of the arrest if applicable, the stop census tract, the stop council district, the stop county, whether the arrest was non-custody or custody in nature if applicable, where the stop occurred, the stop sector, the stop date and time, x- and y-coordinates, and the stop zip code (based on Austin Police Department (2023c)).

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

- Every stop has a distinct number called a "Stop Key" (based on Austin Police Department (2023c)).

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 appears to be only infrequent missing information from individual instances (based on Austin Police Department (2023b)), but it is unclear as to why it is missing.

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.*

- It is unclear as to whether relationships between individual instances are made explicit (based on Austin Police Department (2023b)).

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 do not appear to be recommended data splits (based on Austin Police Department (2023a)).

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

- The dataset's website states that the data "may differ from official APD crime data" (Austin Police Department 2023a).

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 relies on a "2020 Racial Profiling Guide for a description of fields" (Austin Police Department 2023a). There does not appear to be a guarantee that the guide will exist and remain constant over time (based on Austin Police Department (2023d)). It is unclear as to whether there are official archival versions of the complete dataset and

the guide (based on Austin Police Department (2023a) and Austin Police Department (2023d)). The license of the dataset is "Public Domain" (Austin Police Department 2023a), but the guide does not appear to have a license (based on Austin Police Department (2023d)). Neither the dataset nor the guide appear to have any restrictions (based on Austin Police Department (2023a) and Austin Police Department (2023d)), and links to these are https://data.austintexas.gov/Public-Safety/2020-Racial-Profiling-RP-dataset/c65h-gw3m/about_data and https://data.austintexas.gov/Public-Safety/2020-Racial-Profiling-RP-Guide/64yt-89ub/about_data, respectively.

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.*

- The dataset does not appear to contain data that might be considered confidential (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

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.*

- No (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

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 identifies sub-populations such as sex and race (based on Austin Police Department (2023c)). With respect to the individuals who were arrested, approximately 76% of individuals arrested based on traffic stops by the Austin Police Department in 2020 were male, while approximately 24% were female (based on an analysis of the dataset from Austin Police Department (2023a)). In addition, based on an analysis of the dataset from Austin Police Department (2023a), approximately 46% of individuals arrested were Hispanic or Latino, followed by White (approximately 32%), Black (approximately 21%), Asian (approximately 1%), and Middle Eastern, Hawaiian/Pacific Islander, or American Indian/Alaskan Native (less than 0.5% combined).

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 does not appear to be possible to identify individuals from the dataset (based on Austin Police Department (2023b)).

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 data such as sex, race, where the stop occurred, x- and y-coordinates, and the stop zip code (based on Austin Police Department (2023c)).

16. *Any other comments?*

- No.

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.*

- This dataset comprises 68,552 entries (Austin Police Department 2023b) on 2020 Austin Police Department traffic stop-based arrests, warnings or field observations, and citations (Austin Police Department 2023a). This indicates how the variables were measured for storage in this dataset. This suggests that the data was directly observable.

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 collected based on traffic stops (based on Austin Police Department (2023a)). It is unclear as to whether this procedure was validated (based on Austin Police Department (2023a)).

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

- It is unclear as to whether this dataset is a sample from a larger set (based on Austin Police Department (2023a)).

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)?*

- The Austin Police Department appears to have been involved in the data collection process (based on Austin Police Department (2023a)). The nature of any compensation is unclear (based on Austin Police Department (2023a)).

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 appears to have been collected in 2020 (based on Austin Police Department (2023a)). However, the dataset was created on "September 10, 2021" (Austin Police Department 2023a).

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.*

- It is unclear as to whether any ethical review processes were conducted (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

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

- Since the data pertains to 2020 Austin Police Department traffic stop-based arrests, warnings or field observations, and citations (Austin Police Department 2023a), the data appears to have been collected from the individuals in question directly.

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.*

- It is unclear as to whether the individuals in question were notified about the data collection (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

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.*

- It is unclear as to whether the individuals in question consented to the collection and use of their data (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

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).*

- It is unclear as to whether the individuals in question consented to the collection and use of their data (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

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.*

- It is unclear as to whether an analysis of the potential impact of the dataset and its use on data subjects has been conducted (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

12. *Any other comments?*

- No.

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.*

- The nature of any potential preprocessing/cleaning/labeling of the data is uncertain (based on Austin Police Department (2023a) and Austin Police Department (2023d)). Despite this, every stop has a distinct number called a "Stop Key" (based on Austin Police Department (2023c)).

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.*

- It is unclear as to whether the "raw" data was saved in addition to any potentially preprocessed/cleaned/labeled data (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

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

- It is unclear as to whether software was used to preprocess/clean/label the data (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

4. *Any other comments?*

- No.

Uses

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

- It is unclear as to whether this dataset has been used for any tasks already (based on Internet searches).

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.*

- It is unclear as to whether there is a repository that links to papers or systems that use the dataset (based on Internet searches).

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

- The dataset (Austin Police Department 2023a) could be used to assess whether there is a relationship between the prevalence of custody arrests and socioeconomic factors such as income. In addition, the dataset (Austin Police Department 2023a) could be used for a geographical analysis of custody arrests, such as by zip code or county, for purposes such as examining whether certain areas are over-policed.

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?*

- It is unclear as to whether there is anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

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

- It is unclear as to whether there are tasks for which the dataset should not be used (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

6. *Any other comments?*

- No.

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 was uploaded to a GitHub Repository (https://github.com/bennyrochwerg/profiling), which is outside of the entity on behalf of which the dataset was created.

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 available on a website: https://data.austintexas.gov/Public-Safety/2020-Racial-Profiling-RP-dataset/c65h-gw3m/about_data (Austin Police Department 2023a). The dataset does not appear to have a DOI (based on Austin Police Department (2023a)).

3. *When will the dataset be distributed?*

- The dataset was created on "September 10, 2021" (Austin Police Department 2023a).

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.*

- The license of the dataset is "Public Domain" as listed here: https://data.austintexas.gov/Public-Safety/2020-Racial-Profiling-RP-dataset/c65h-gw3m/about_data (Austin Police Department 2023a). The Terms of Use includes information related to areas such as attribution and is available here: https://data.austintexas.gov/stories/s/ranj-cccq (City of Austin, Texas, n.d.b).

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.*

- It is unclear as to whether any third parties have imposed IP-based or other restrictions on the data associated with the instances (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

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.*

- It is unclear as to whether any export controls or other regulatory restrictions apply to the dataset or to individual instances (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

7. *Any other comments?*

- No.

Maintenance

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

- The City of Austin, Texas Open Data Portal is hosting the dataset (based on Austin Police Department (2023a)).

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

- The owner of the dataset is Sinah Kang (Austin Police Department 2023a). The owner of the dataset can be contacted by accessing the dataset's website (https://data.austintexas.gov/Public-Safety/2020-Racial-Profiling-RP-dataset/c65h-gw3m/about_data; Austin Police Department (2023a)), clicking on "Actions", and clicking on "Contact dataset owner" (Austin Police Department 2023a).

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

- It is unclear as to whether there is an erratum (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

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)?*

- It is unclear as to whether the dataset will be updated (based on Austin Police Department (2023a) and Austin Police Department (2023d)). The dataset was most recently updated on "April 10, 2023" (Austin Police Department 2023a).

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.*

- It is unclear as to whether there are applicable limits on the retention of the data associated with the instances (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

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.*

- It is unclear as to whether older versions of the dataset will continue to be supported/hosted/maintained (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

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.*

- It is unclear if there is a mechanism for others to extend/augment/build on/contribute to the dataset (based on Austin Police Department (2023a) and Austin Police Department (2023d)).

8. *Any other comments?*

- No.

References

Austin Police Department. 2023a. "2020 Racial Profiling (RP) Dataset." City of Austin, Texas Open Data Portal. https://data.austintexas.gov/Public-Safety/2020-Racial-Profiling-RP-dataset/c65h-gw3m/about_data.

———. 2023b. "2020 Racial Profiling (RP) Dataset." City of Austin, Texas Open Data Portal. https://data.austintexas.gov/Public-Safety/2020-Racial-Profiling-RP-dataset/c65h-gw3m/data_preview.

———. 2023c. "2020 Racial Profiling (RP) Guide." City of Austin, Texas Open Data Portal. https://data.austintexas.gov/Public-Safety/2020-Racial-Profiling-RP-Guide/64yt-89ub/data_preview.

———. 2023d. "2020 Racial Profiling (RP) Guide." City of Austin, Texas Open Data Portal. https://data.austintexas.gov/Public-Safety/2020-Racial-Profiling-RP-Guide/64yt-89ub/about_data.

City of Austin, Texas. n.d.a. "About Us." https://data.austintexas.gov/stories/s/g7cx-4kmk.

———. n.d.b. "Terms of Use for the City of Austin Open Data Portal." https://data.austintexas.gov/stories/s/ranj-cccq.

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.