Datasheet: US Household Income Statistics

Author: Ruge Xu

Link: https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations

Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

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 originally developed for real estate and business investment research. Income is a vital element when determining both the quality and socioeconomic features of a given geographic location.

2. **Who created this dataset (e.g. which team, research group) and on behalf of which entity (e.g. company, institution, organization)**?

Golden Oak Research Group.

3. **What support was needed to make this dataset?** (e.g. who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)

Not mentioned.

4. **Any other comments?**

Composition

Dataset creators should read through the questions in this section prior to any data collection and then provide answers once collection is complete. Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks. The answers to some of these questions reveal information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

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

Each instance represents a geographic location in the U.S. They are in different states, counties, cities, latitudes and longitudes.

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

Totally 348,893 records.

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 (e.g. 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 (e.g. to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset covers all the geographic locations in the U.S. It contains all possible instances.

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

Each instance consists of features. These features describe the location's household and geographic statistics, and Geographic Location.

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

Each location is divided into different types according to the U.S. Census Bureau.

6. **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g. because it was unavailable).

This does not include intentionally removed information, but might include, e.g. redacted text.

Not mentioned.

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

The latitude and longitude show the geographical relationship of the different locations.

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

No, there aren’t.

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

Not clear.

10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g. 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 (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g. licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The data is retrieved from the 2011-2015 ACS 5-Year Documentation by the U.S. Census Reports. The Census Bureau's mission is to serve as the nation's leading provider of quality data about its people and economy. The website will exist and remain constant over time as the restriction is for data prior to 2015.

11. **Does the dataset contain data that might be considered confidential (e.g. 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.

No, all the data are public.

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 household income statistics of each location included in the dataset can be a source of economic anxiety.

13. **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

Yes, as the dataset is about the household income.

14. **Does the dataset identify any subpopulations (e.g. by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No, it doesn’t.

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

No, as the dataset only provides statistics for the entire location.

16. **Does the dataset contain data that might be considered sensitive in any way (e.g. data that reveals racial 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.

No, it doesn’t.

17. **Any other comments?**

Collection

As with the previous section, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. In addition to the goals of the prior section, the answers to questions here may provide information that allow others to reconstruct the dataset without access to it.

1. **How was the data associated with each instance acquired?** Was the data directly observable (e.g. raw text, movie ratings), reported by subjects (e.g. survey responses), or indirectly inferred/derived from other data (e.g. part-of-speech tags, model-based guesses for age or language)? If 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 derived from other data.

2. **What mechanisms or procedures were used to collect the data (e.g. hardware apparatus or sensor, manual human curation, software program, software API)?** How were these mechanisms or procedures validated?

The authors obtained the data from the United State Census Bureau, which provides quality data about its people and economy, and is open for download.

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

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

Not mentioned.

5. **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g. recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.

The dataset was created at 2017-08-08, and last updated in 2018-04-16.

6. **Were any ethical review processes conducted (e.g. 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.

Not mentioned.

7. **Does the dataset relate to people?** If not, you may skip the remainder of the questions in this section.

Yes, as the dataset is about the household income.

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

It was obtained via a website.

9. **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, the U.S. Census Bureau, the original data source, typically sends a formal letter describing the survey an individual's household or business has been selected to participate in and why his or her participation is important. This letter describes what the survey is, what types of questions will be asked, and what you can expect as a survey participant.

The link is https://www.census.gov/programs-surveys/surveyhelp/we-conduct-surveys.html.

10. **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 only says that the Census Bureau will notify the individuals in advance if they are in a survey, and provide different ways to respond.

11. **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).

Not mentioned.

12. **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g. 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.

The Census Bureau is committed to safeguarding the information of survey participants so that we can provide the country with high quality statistics. They claim to have secure technology and cutting-edge safeguards to protect individuals' identities.

The link is https://www.census.gov/programs-surveys/surveyhelp/protect-information.html.

13. **Any other comments?**

Preprocessing / Cleaning / Labeling

Dataset creators should read through these questions prior to any pre-processing, cleaning, or labeling and then provide answers once these tasks are complete. The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

1. **Was any preprocessing/cleaning/labeling of the data done (e.g. 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 remainder of the questions in this section.

They pre-processed the data to translate the results of individual feedback into statistics for each region. This process is recorded in the methodology part.

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

The original data is https://www2.census.gov/programs-surveys/acs/summary_file/2015/data/5_year_by_state/.

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

Not mentioned.

4. **Any other comments?**

Uses

These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

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

Yes, a task of finding predictors of higher percentage of Specialized High School Admissions Test uses this dataset as a supplement to income information.

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.

Some tasks use this dataset can be found in https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations/activity.

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

The dataset is also used for random forest predict means and mapping US household income.

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 future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g. stereotyping, quality of service issues) or other undesirable harms (e.g. financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset may create stereotypes about the economic situation in certain locations in the U.S.

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

Not mentioned.

6. **Any other comments?**

Distribution

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

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

Not mentioned.

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

The dataset is distributed on their website: https://www.goldenoakresearch.com/.

3. **When will the dataset be distributed?**

August 2017.

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 dataset is licensed for use in perpetuity. Anyone can use the product on an ongoing basis with no time limits, or renewal fees. Anyone may display, broadcast, as long as the access to the data is secured and cited appropriately.

The link is https://www.goldenoakresearch.com/terms-of-use.

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.

Not mentioned.

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.

Not mentioned.

7. **Any other comments?**

Maintenance

As with the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

1. **Who is supporting/hosting/maintaining the dataset?**

The members of Golden Oak Research.

2. **How can the owner/curator/manager of the dataset be contacted (e.g. email address)?**

The email address is research_development@goldenoakresearch.com.

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

Not found.

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

Not sure as the data was last updated 4 years ago.

5. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g. were 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.

Not mentioned.

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 users.

Users can send email to them if they find errors. There is also a database feedback form on their own website.

The link is https://www.goldenoakresearch.com/database-feedback.

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 other users? If so, please provide a description.

Not found.

8. **Any other comments?**