Title Datasheet for Toronto Monthly Weather Normals Dataset*

Kuiyao Qiao

March 27, 2024

The datasheet summarizes weather data for Toronto over the past three decades. It documents the motivation for the dataset, its composition, the collection process and the proposed use.(R Core Team 2023)

Motivation

The dataset appears to have been designed to provide detailed historical climate data for Toronto, possibly to aid climatological research and urban planning and to make historical weather patterns available to the public. Its mission may have been to fill gaps in Toronto's long-term weather data to facilitate analysis of climate trends, weather prediction modelling, and educational purposes.

The creator of the dataset is Environment and Climate Change Canada. The dataset does not provide details on funding. Still, it would have been supported by a government grant or the operating budget of a public agency dedicated to environmental monitoring and public information dissemination.

The dataset is a valuable resource for understanding weather patterns and climate change in the Toronto area. It helps fill a significant gap in historical weather data by providing a structured and comprehensive record that can support a wide range of analyses and practical applications.("Amateur Weather Statistics for Toronto, Ontario," n.d.) (Canada 2024)

Composition

The dataset consists of instances representing monthly weather statistics for Toronto. The cases do not differ in type but are uniform. Each instance reflects weather data for a particular month, such as temperature, precipitation, and possibly other meteorological variables, such as wind speed and humidity.The total number of instances in the dataset is equal to the number of months covered, with 12 cases per year for 30 years.

*Code and data are available at: https://github.com/QPP123/Toronto-Monthly-Weather-Dataset

The dataset may represent a sample of a larger dataset of daily weather observations. It has been carefully selected to provide monthly averages or totals that contribute to understanding long-term climate patterns rather than short-term weather variations. The representativeness of a dataset depends on its length and the consistency of the record; a dataset is representative if it consistently covers a long period with no apparent gaps.

Each instance in the dataset may include features extracted from raw weather data, such as average monthly temperature, total rainfall, and snowfall. These are not raw data points but aggregated statistics that provide a monthly view of weather conditions.

In terms of labels or targets, each instance can be associated with the month and year it represents, allowing the user to identify a specific temporal context. There may be missing information if there are gaps in the data collection process due to equipment malfunction or other problems, although this needs to be explicitly stated.

The dataset may need well-defined relationships between instances because each month is an independent observation of weather conditions. Since the primary purpose of the dataset is not to be used for machine learning but to analyze historical weather patterns, no suggested method for partitioning the data is provided.

With a rigorous data collection and quality control process, errors, noise, or redundancy will likely be minimal. The dataset is self-contained and not directly dependent on external sources, which ensures its long-term stability and reliability.

No confidentiality issues are associated with this publicly available, non-personal data. Given the nature of this dataset as a collection of weather statistics, it is unlikely to contain offensive or sensitive content. Subpopulations or personal identification are not relevant.Dataset is a structured collection of historical monthly weather statistics for Toronto intended for climate analysis and long-term trend observation.

Collection process

The data in the dataset were most likely collected by automated meteorological monitoring systems that use various hardware devices or sensors to measure meteorological elements such as temperature, precipitation, and wind speed. These data are directly observable and are collected continuously over time by these instruments, ensuring a systematic and consistent data collection process.

Weather stations equipped with these sensors are usually part of a more extensive network operated by national or local meteorological services to ensure standardization of data collection methods. Validation of these mechanisms is often rigorous, including regular equipment calibration and maintenance and cross-validation with nearby stations to ensure accuracy and consistency.

The data set is a sample of a more considerable continuous weather record, organized monthly for easy analysis and interpretation. The sampling strategy here is systematic, with each

month's data representing a summary or average of a series of daily observations to provide an overview of the weather conditions for that month.

Those involved in the data collection will likely be meteorologists and technicians specializing in climatology and atmospheric science employed by weather monitoring agencies. Their compensation will be part of their regular duties in these organizations.

The data collection time frame indicates long-term weather monitoring, which can be decades. This duration is necessary to determine climate trends and averages. The data associated with each instance (monthly summaries) are produced over this extended period, corresponding to the collection period.

Given the nature of meteorological data, an ethical review process is generally not applicable as the data do not involve human subjects or personal information. The data collection is from environmental observations, not individuals, and is therefore irrelevant to issues of notice, consent, and impact on the data subject.

Finally, the potential impact of using this dataset is primarily in the context of climate research and weather pattern analysis, with little risk to individuals or communities. The use of the dataset is intended to benefit the public, improve understanding of climate trends, and support weather-related planning and decision-making.

Preprocessing/cleaning/labeling

To ensure the accuracy and usability of the data, the dataset may have been preprocessed and cleaned. This may include averaging daily weather measurements to create monthly normals, handling missing values using statistical methods or interpolation, and ensuring consistency in data format (e.g., units of measurement for temperature and precipitation).

Raw data, typically daily weather observations, may be stored separately by the WMO that collects the data. However, the dataset information does not provide access to this raw data. Researchers or analysts must contact the data provider directly if they need daily data for more detailed analysis. No mention of the availability of software for pre-processing/cleaning/tagging data.

Uses

The dataset can be used for climate analysis tasks, such as studying long-term weather patterns, urban planning, and environmental studies. Specific uses of this dataset in published papers or systems need to be cited.

In addition to current applications, the dataset can be used for educational purposes, to inform decisions related to climate adaptation, or for modelling to predict future climate conditions in Toronto. The historical nature of the data makes it a valuable resource for understanding climate change and its impacts on the urban environment.

Given the composition of the dataset and the collection process, potential users need to recognize that it represents historical averages and may not be suitable for predicting short-term

weather events. The lack of daily granularity means it should not be used for purposes requiring high temporal resolution, such as immediate weather forecasting or emergency planning.

This weather dataset has no apparent immediate risk or harm as it does not contain sensitive or personal data. However, users should be careful not to misinterpret the data as indicative of specific future weather conditions. Such misuse could lead to unpreparedness for weather-related events or a misleading understanding of climate change.

The dataset should not be used for applications requiring real-time data or single event analysis, as it contains monthly averages and lacks the temporal resolution required for such tasks. Misuse may lead to accurate conclusions or appropriate responses to weather-related needs.

Distribution

The information provided is synthesized from multiple Environment and Climate Change Canada data sources and is available to all.

Due to its educational and informational nature, the dataset already be publicly available, allowing users to download it in formats like CSV. This information will not be included in the dataset details.

Datasets are released continuously and may be updated monthly or annually to reflect new meteorological data. Regarding licensing, the datasets may be distributed under licenses that allow free use, distribution, and modification but with certain conditions to prevent misuse, such as attribution requirements or restrictions on commercial use. There are very few intellectual property or other restrictions on such data, as meteorological data are usually considered in the public domain or made available under open data licenses. Export controls or other regulatory restrictions are unlikely to apply to meteorological data due to its non-sensitive nature and the public interest in its availability.

The dataset will likely be widely distributed with minimal restrictions to support its use in weather-related research, education, and planning.

Maintenance

The dataset is supported, hosted and maintained by the volunteer-run weatherstats.ca. They can be contacted via email.

As meteorological data is collected on an ongoing basis, it is expected that the datasets will be updated periodically. Updates may include corrections to past data, the addition of new monthly statistics, and the removal of erroneous entries. The frequency of updates may be monthly, yearly.

Since the dataset is meteorological statistics and not personal data, there are no data retention issues related to individuals. Therefore, privacy-related data retention restrictions do not apply in this case.

The policy for supporting older versions of the datasets depends on Environment and Climate Change Canada. Ideally, older versions should remain accessible for historical comparisons and research.

Since the data is provided directly by Environment and Climate Change Canada, no third party can extend/edit/build/contribute.

In summary, maintenance of the dataset includes regular updates and communication with users to ensure the data's accuracy, relevance, and usefulness for various applications. The maintenance entity should provide clear and accessible channels for users to participate and contribute to the development and refinement of the dataset.

Acknowledgments

Thanks to Zijun Meng for all the helpful feedback while I was working on this paper.

References

"Amateur Weather Statistics for Toronto, Ontario." n.d. https://toronto.weatherstats.ca/download.html.

Canada, Environment. 2024. "Weather Information." https://weather.gc.ca/.

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