6.1 Sourcing Open Data

‹Marie Jacobsson›

Data Source

Daily Temperature of Major Cities

The dataset “Daily Temperature of Major Cities” is an external source, which I found on Kaggle. The dataset is collected from files from University of Dayton, and source data is from National Climatic Data Center. I rate the data as trustworthy.

Data Collection

The temperature has been collected by the National Climatic Data Center, and compiled by the University of Dayton. The temperature measured every day between the years 1995 - to present. In the dataset “Daily Temperature of Major Cities” on Kaggle the years are between 1995-2020. I decided to remove 2020 from my analysis though, because it wasn’t complete.

Context

Global warming is the ongoing rise of the average temperature of the Earth's climate system and has been demonstrated by direct temperature measurements and by measurements of various effects of the warming - Wikipedia

So a dataset on the temperature of major cities of the world will help analyze the same. Also weather information is helpful for a lot of data science tasks like sales forecasting, logistics etc.

Content

Daily level average temperature values are present in city_temperature.csv file (Original Data). This site contains files of daily average temperatures for 157 U.S. and 167 international cities. The files are updated on a regular basis and contain data from January 1, 1995 to present (the dataset I used from Kaggle extends to 2020).

Source data for this site are from the National Climatic Data Center. The data is available for research and non-commercial purposes only.

Resources

Average Daily Temperature Archive

Why I chosen this data

I spent several hours on sites searching for an open dataset that fits my project. I found it really hard to find something that combined my interest with the criterias for the task. I guess I checked out 50-70 datasets but either the rows were too few, there were no places/countries or they didn't contain both categorical and continuous columns. At last I found the dataset “Daily temperature of major cities” which contains what I need for my project. And I find it interesting to see how the temperature behaves during time.

Data Profile

The data set contains 2906327 rows and 8 columns.
Columns: Region, Country, State, City, Month, Day, Year, AvgTemperature.

Data wrangling:
The column "State" includes a lot of NaN, and it's needless for this analysis (I have region, country and city), so I decided to drop that column in Jupyter.

```
# Print df after dropping column

df.head()
```

	Region	Country	City	Month	Day	Year	AvgTemperature
0	Africa	Algeria	Algiers	1	1	1995	64.2
1	Africa	Algeria	Algiers	1	2	1995	49.4
2	Africa	Algeria	Algiers	1	3	1995	48.8
3	Africa	Algeria	Algiers	1	4	1995	46.4
4	Africa	Algeria	Algiers	1	5	1995	47.9

After removing column “State” the dataset contains 2906327 rows and 7 columns.

Consistency Checks:

- There are no mixed data types in the dataset.
- There are no missing values in the dataset.
- There are no duplicates in the dataset.
- Due to the information from the University of Dayton there are some values that couldn’t be correctly collected, because of problems with weather station metering equipment.
  They denote missing data using a “–99” flag.
- I change those values to NaN in Jupyter since they affect the basic statistics (minimum AvgTemperature).
- The dataset now includes 79164 NaN values.
- There were two typing errors in the column “Year”, 200 and 201.
  It may be 2000 and 2001, but there are also many other options. Those two years didn’t

contain many values so I decided to remove them.

```
2000    119682
1999    119355
1998    119082
1997    118656
1996    118951
1995    118616
201        351
200         89
```

- I also chose to drop the year 2020 since it's not complete.

Basic descriptive statistics:

2906327 rows

7 columns.

The average temperature (AvgTemperature)

Minimum -50 F (-45,5 C)

Maximum 110 F (43,3 C)

The mean is 60 F (15,5 C).

Limitations and ethical considerations

The limitations in this dataset are only the NaN values. These values were registered when something was wrong with the weather station metering equipment.

I can't find any ethical considerations regarding this dataset.

Questions

- What is happening with temperature over time?
- When, in that case, the temperature begins to behave differently?
- Which Region/Countries are most affected by temperature differences?