AD DESTINATUM PERSEQUOR

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

Johann Pienaar - u24050505

MIT 805

Big Data: Assignment 1

11 September 2022

SCHOOL OF INFORMATION TECHNOLOGY

MASTER OF INFORMATION TECHNOLOGY

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Faculty of Engineering, Built Environment and Information Technology

INDIVIDUAL ASSIGNMENT COVER PAGE

Name of Student	Johann Pienaar
Student Number	u24050505
Name of Module	Big data
Module Code	MIT 805
Name of Lecturer	Stacey Baror
Date of Submission	11 September 2022
Contact telephone number	(+27) 072 392 1960
E-mail address	jpienaar85@gmail.com / u24050505@tuks.co.za
Declaration:	*I declare that this assignment, submitted by me, is my own work and that I have referenced all the sources that I have used.*
Signature of Student	Johann Pienaar
Date received	
Signature of Administrator	
Mark	
Date	
Signature of Lecturer	

1 Dataset description

The dataset chosen for this assignment is the National Oceanic and Atmospheric Administration Global Surface Summary of the Day (NOAA GSOD) database. The dataset was originally downloaded from Kaggle [1]. However, the initial download only contained data up to and partially inclusive of 2019 and was therefore subsequently enriched from the NOAA FTP server [2]. For this assignment I chose to only enrich the dataset up to 2021, which is the last year for which a full dataset is available.

The dataset contains several daily weather elements for over 9000 stations over a period since 1929 up to the present day, with data since 1973 generally considered to be the most complete [3].

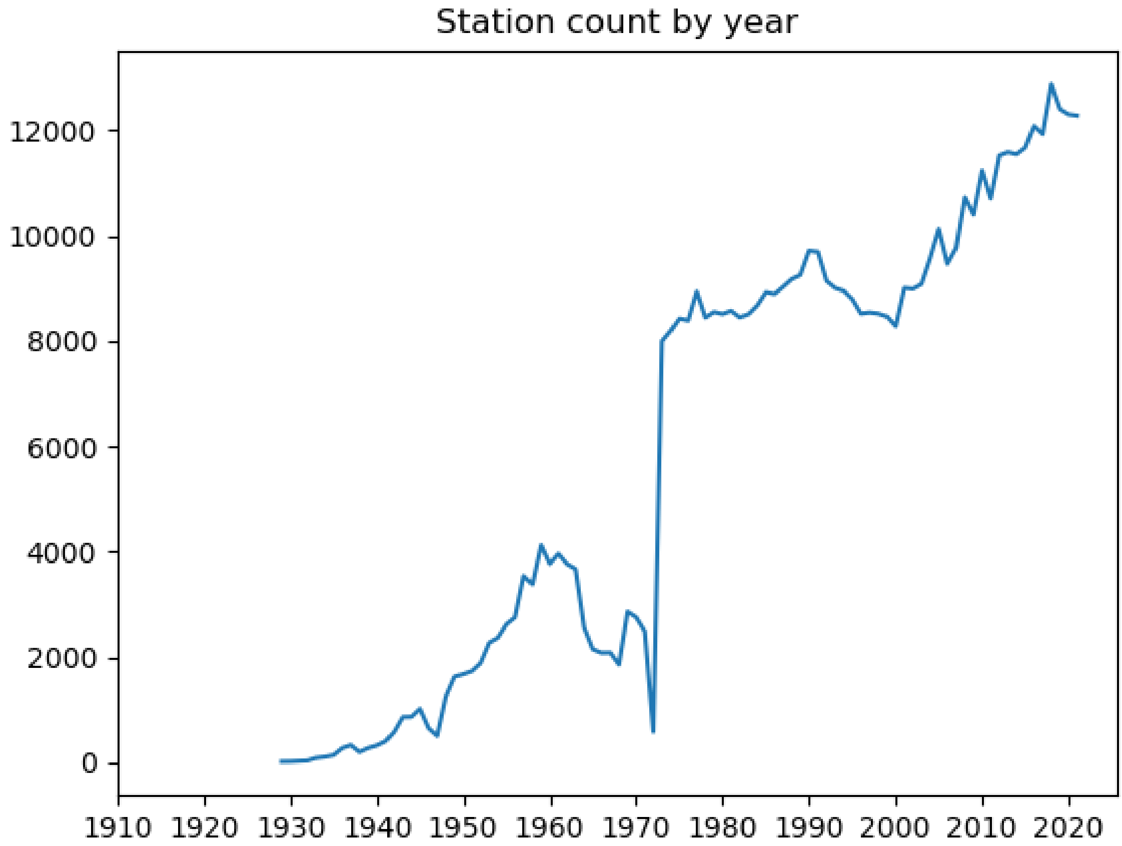

Figure 1 Growth of stations included in the dataset since 1929

Data is contained in a fixed width format in files with a standardized naming convention with 1 file per station per year. Naming is in the format XXXXXX-ZZZZZ-YYYY.op.gz, where XXXXXX is the station's World Meteorological Organisation (WMO) number, ZZZZZ is the station’s Weather Bureau Air Force Navy (WBAN) number (if appropriate) and YYYY is the year of the data. The file format (i.e. starting position and field width of the various fields) as well as quality control measures that was applied is explained in the readme file. Data that wasn’t collected is indicated by all 9’s and will be removed prior to processing to avoid skewing the data. Elements of interest contained includes [3]: mean, minimum and maximum temperatures in Fahrenheit as well as precipitation reported in inches.

Finally, there is additional metadata for each of the stations available in an additional csv file provided with the dataset that contains the station’s:

- location (latitude and longitude) as well as a station name
- elevation
- dates of operation
- ICAO code if the station is part of an ICAO airport
- a country code, for which a further lookup is provided in another file

At present, once the dataset was untarred (but the individual files remained gzipped), it is approximately 3.56 gb and growing at approximately (compressed) 105 mb/y (12275 stations). Note that the dataset is growing both in time as well as due to more stations being added. With the increasing interest in climate science due to global warming and with the increase in internet connected devices (IoT) we can expect that the dataset rate of growth will also increase. Africa in particular lacks significant coverage at present and will likely see increases in the future. [4]

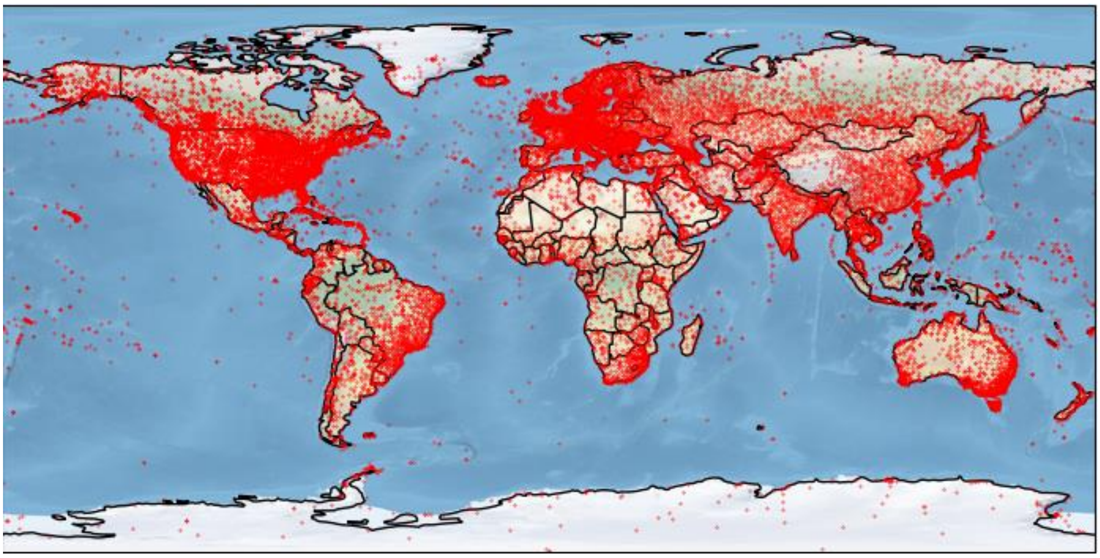

Figure 2 Geospatial distribution of stations included in the dataset

2 Pre-processing

For the purposes of this assignment, I decided to consolidate the different year files per station into a single file i.e. resulting in one csv file per station. Converting the data into a csv format will ease future processing as opposed to working with the fixed width data files that the data was in. While conducting this consolidation I also took the opportunity to:

- convert the units of measure to SI units (i.e. Fahrenheit to Celsius and inches to mm)
- remove null values (9999.9 in the original data)
- process the 8-digit yyyymmdd values into separate columns for year, month and day

Finally, as described in the next section, I will mostly be examining the temperature and precipitation data and therefore dropped some of the other columns.

At the conclusion of this pre-processing I had 26941 uncompressed csv files at approximately 7.68 gb and 159,775,279 individual lines of temperature and precipitation data from across the world.

3 Phenomena to be investigated

3.1 Persistent changes in temperature minima/maxima

A key trend that is expected to be found in the data is a persistent increase in the temperatures experienced across the world. This will affect the both the minimum and maximum daily temperatures and the rate of increase is expected to be in the range 0.1-0.8 °C/decade. Furthermore, the rate of change in morning minimums and afternoon maximums may be different, indicating a narrowing or widening of the diurnal temperature range. Finally, the rate of change is also expected to be different at different latitudes and this will also be investigated.

3.2 Persistent changes in precipitation

As the climate continues to warm across the planet, we’re expecting rainfall patterns to change regionally with some areas receiving more rain and others less. This could potentially alter the farming suitability of large areas of land. In some cases, this may be manageable through a change in crop, but could also render previously arable land hostile to agriculture and decrease food output.

3.3 Frequency of extreme events

Finally, it is often the extreme events (flooding, droughts, cold snaps and heatwaves) that causes the most damage and loss of life. The frequency at which these events occur will therefore also be investigated.

4 References

[1] Kaggle, “NOAA Global Surface Summary of the Day,” 2019. [Online]. Available: https://www.kaggle.com/datasets/noaa/noaa-global-surface-summary-of-the-day. [Accessed 1 9 2022].

[2] NOAA, “NOAA FTP,” [Online]. Available: ftp://ftp.ncdc.noaa.gov. [Accessed 4 9 2022].

[3] NOAA, “NOAA GSOD Readme,” [Online]. Available: ftp://ftp.ncdc.noaa.gov/pub/data/gsod/readme.txt. [Accessed 7 9 2022].

[4] E. Aguilar, A. Aziz Barry, M. Brunet, L. Ekang, A. Fernandes, M. Massoukina, J. Mbah, A. Mhanda, D. J. do Nascimento, T. C. Peterson, O. Thamba Umba, M. Tomou and X. Zhang, “Changes in temperature and precipitation extremes in western central Africa, Guinea Conakry, and Zimbabwe, 1955–2006,” *JOURNAL OF GEOPHYSICAL RESEARCH,* vol. 114, no. D02115, 2009.

5 Appendix

5.1 Data pre-processing to csv files

```
import os, re, gzip, time

def F2C(f): #Fahrenheit 2 Celcius
    return (f-32)*5/9

def inch2mm(i): #inches 2 mm
    return i*25.4

def float_or_none(s, measure):
    r=re.match(r'[\.9 ]+',s)
    if r and r.end() == len(s):
        return ''
    else:
        if measure == 'T':
            return f'{F2C(float(s)):.1f}'
        elif measure == 'I':
            return f'{inch2mm(float(s)):.1f}'

start = time.time()
files={}
regex = re.compile(r'([A0-9]+-\d+)-\d+')
HEADER = 'STN, WBAN, YEAR, MONTH, DAY, MEAN_TEMP, MAX_TEMP, MIN_TEMP,
PRCP\n'
line_data = "{stn}, {wban}, {year}, {month}, {day}, {mean_temp},
{max_temp}, {min_temp}, {prcp}\n"
all_files = sorted(os.listdir(r'.\gsod_all_years'))
file_count = len(all_files)

prev_stn_id=regex.match(all_files[0]).groups()[0]
for i, filename in enumerate(all_files):
    if i%1000==0:
        print(f'{i*100/file_count:.2f}%')
    stn_id = regex.match(filename).groups()[0] #current
    if stn_id != prev_stn_id:
        files[prev_stn_id].close()
    if stn_id not in files:
        files[stn_id] = open(r'.\csv\{}.csv'.format(stn_id),'wt')
        files[stn_id].write(HEADER)
    with gzip.open(os.path.join(os.getcwd(),'gsod_all_years',filename),
'rt') as f:
        f.readline()
        for line in f.readlines():
            stn, wban = stn_id.split('-')
            year = line[14:18]
            month = line[18:20]
            day = line[20:22]
            mean_temp = float_or_none(line[24:30],'T')
            max_temp = float_or_none(line[102:108],'T')
            min_temp = float_or_none(line[110:116],'T')
            prcp = float_or_none(line[118:123],'I')
            files[stn_id].write(line_data.format(**locals()))
    prev_stn_id = stn_id
print(f'That took {time.time()-start:.2f} s')
```

5.2 Geospatial distribution of stations

```python
import pandas as pd
import matplotlib.pyplot as plt
import cartopy.crs as ccrs
import cartopy.feature as cfeature

df = pd.read_csv('isd-history.csv')
df['bogus']=((df['LAT']!=0)|(df['LON']!=0))
df2=df.loc[df['bogus'],:]
ax = plt.axes(projection=ccrs.PlateCarree())
ax.stock_img()
ax.add_feature(cfeature.COASTLINE)
ax.add_feature(cfeature.BORDERS)

plt.scatter(df2['LON'],df2['LAT'], s=0.1, color='red',
            transform=ccrs.PlateCarree())
plt.show()
```

5.3 Growth of station count over time

```python
import os, re
import numpy as np

files = os.listdir('.\gsod_all_years')
stn_str = re.compile(r'([A0-9]+)-(\d+)-(\d+).op.gz')
stns = []
for f in files:
    stns.append(stn_str.match(f).groups())

df = pd.DataFrame(stns, columns=['WMO','WBAN','year'])
counts = df.groupby('year').count()['WMO']
counts.index = pd.to_numeric(counts.index)

plt.plot(counts)
plt.title('Station count by year')
plt.xticks(ticks=np.arange(1910, 2030, 10))
plt.show()
```