DATA420 – Assignment 1

75958138 – NDU31

Processing

Question 1a:

Directory Tree:

```
In [11]: !hdfs dfs -ls /data/ghcnd

         Found 5 items
         drwxr-xr-x   - jsw93 supergroup          0 2022-08-08 01:20 /data/ghcnd/daily
         -rw-r--r--   8 jsw93 supergroup       3659 2022-08-08 01:21 /data/ghcnd/ghcnd-countries.txt
         -rw-r--r--   8 jsw93 supergroup   33384684 2022-08-08 01:21 /data/ghcnd/ghcnd-inventory.txt
         -rw-r--r--   8 jsw93 supergroup       1086 2022-08-08 01:21 /data/ghcnd/ghcnd-states.txt
         -rw-r--r--   8 jsw93 supergroup   10496042 2022-08-08 01:21 /data/ghcnd/ghcnd-stations.txt
```

ghcnd/

├─ daily/

| ├─ 1763.csv.gz

| ├─ 1764.csv.gz

| ├─ …

| ├─ 2021.csv.gz

| ├─ 2022.csv.gz

├─ ghcnd-countries.txt

├─ ghcnd-inventory.txt

├─ ghcnd-states.txt

├─ ghcnd-stations.txt

This displays that there are four files and a directory located under ghcnd. Daily is the only directory and the rest are all files.

Question 1b:

I can see 260 files from 1763 to 2022, there are no gaps so 160 continuous years are contained in daily. The size of the data increases year by year – presumably due to more observations around the world and more advanced technology being used. 2022 file is smaller than 2021 as the year is not yet complete and hence doesn't contain a full dataset in comparison to prior years.

```
!hdfs dfs -ls /data/ghcnd/daily
```

```
Found 260 items
-rw-r--r--   8 jsw93 supergroup       3358 2021-08-09 15:08 /data/ghcnd/daily/1763.csv.gz
-rw-r--r--   8 jsw93 supergroup       3327 2021-08-09 15:03 /data/ghcnd/daily/1764.csv.gz
-rw-r--r--   8 jsw93 supergroup       3335 2021-08-09 15:03 /data/ghcnd/daily/1765.csv.gz
-rw-r--r--   8 jsw93 supergroup       3344 2021-08-09 14:56 /data/ghcnd/daily/1766.csv.gz
-rw-r--r--   8 jsw93 supergroup       3356 2021-08-09 15:06 /data/ghcnd/daily/1767.csv.gz
-rw-r--r--   8 jsw93 supergroup       3325 2021-08-09 15:02 /data/ghcnd/daily/1768.csv.gz
-rw-r--r--   8 jsw93 supergroup       3418 2021-08-09 15:03 /data/ghcnd/daily/1769.csv.gz
```

```
-rw-r--r--   8 jsw93 supergroup  207618101 2021-08-09 15:00 /data/ghcnd/daily/2015.csv.gz
-rw-r--r--   8 jsw93 supergroup  209584081 2021-08-09 15:04 /data/ghcnd/daily/2016.csv.gz
-rw-r--r--   8 jsw93 supergroup  207342349 2021-08-09 15:06 /data/ghcnd/daily/2017.csv.gz
-rw-r--r--   8 jsw93 supergroup  163202973 2022-08-08 01:19 /data/ghcnd/daily/2018.csv.gz
-rw-r--r--   8 jsw93 supergroup  161945469 2022-08-08 01:19 /data/ghcnd/daily/2019.csv.gz
-rw-r--r--   8 jsw93 supergroup  161995783 2022-08-08 01:19 /data/ghcnd/daily/2020.csv.gz
-rw-r--r--   8 jsw93 supergroup  159598394 2022-08-08 01:19 /data/ghcnd/daily/2021.csv.gz
-rw-r--r--   8 jsw93 supergroup   88195367 2022-08-08 01:20 /data/ghcnd/daily/2022.csv.gz
```

```
In [19]: !hdfs dfs -du -h /data/ghcnd
```

Question 1c:

The total size of the data is 15.8 G, this size reflects mainly daily as the smaller text files only consist of a few kilobytes or megabytes.

```
: !hdfs dfs -du -s -h /data/ghcnd
```

```
15.8 G  126.5 G  /data/ghcnd
```

```
!hdfs dfs -du -h -v /data/ghcnd
```

```
SIZE    DISK_SPACE_CONSUMED_WITH_ALL_REPLICAS  FULL_PATH_NAME
15.8 G  126.1 G                                /data/ghcnd/daily
3.6 K   28.6 K                                 /data/ghcnd/ghcnd-countries.txt
31.8 M  254.7 M                                /data/ghcnd/ghcnd-inventory.txt
1.1 K   8.5 K                                  /data/ghcnd/ghcnd-states.txt
10.0 M  80.1 M                                 /data/ghcnd/ghcnd-stations.txt
```

Question 2:

The description of the dataset was correct, some unexpected errors were that DateTime and Timestamp types were not able to be used as they don't work in the schema which meant that "DATE" and "OBSERVATION_TIME" were kept as string type.

The amount of rows per metadata table are as follows:

Stations	122047
States	74
Countries	219
Inventories	725754

The number of stations that do not have a WMO ID was 113953, which means that the majority of the stations do not have one.

Question 3:

To gather the information for question 3, I developed a datatable that aggregated the minimum first year, maximum last year and total element count for each unique station.

ID	FIRSTYEAR	LASTYEAR	TOTAL_ELEMENTS
AEM00041217	1983	2022	4
AGE00147708	1879	2022	5
AGE00147714	1896	1938	3
AGM00060452	1985	2022	4
AGM00060511	1983	2022	5
AJ000037749	1936	2022	5
ALE00100939	1940	2000	2
AQC00914021	1955	1957	10
AQC00914424	1969	1975	5
AQC00914873	1955	1967	12
AR000000002	1981	2000	1
AR000087374	1956	2022	5
AR000870470	1956	2022	5
AR000877500	1956	2022	5
ARM00087480	1965	2022	5
ARM00087509	1973	2022	5
ARM00087679	1973	2022	4
ASN00001006	1951	2022	6
ASN00001020	2004	2022	9
ASN00001021	1941	2005	10

only showing top 20 rows

This was further developed to count the number of core elements vs. the number of other elements that each station collected.

ID	FIRSTYEAR	LASTYEAR	TOTAL_ELEMENTS	PRCP	SNOW	SNWD	TMAX	TMIN	OTHER_ELEMENTS
AEM00041217	1983	2022	4	1	0	0	1	1	4
AGE00147708	1879	2022	5	1	0	1	1	1	5
AGE00147710	1909	2009	4	1	0	0	1	1	4
AGE00147714	1896	1938	3	1	0	0	1	1	3
AGE00147719	1888	2022	4	1	0	0	1	1	4
AGM00060360	1945	2022	4	1	0	0	1	1	4
AGM00060445	1957	2022	5	1	0	1	1	1	5
AGM00060452	1985	2022	4	1	0	0	1	1	4
AGM00060511	1983	2022	5	1	0	1	1	1	5
AGM00060540	1981	2022	5	1	0	1	1	1	5
AJ000037679	1959	1987	1	1	0	0	0	0	1
AJ000037749	1936	2022	5	1	0	1	1	1	5
AJ000037831	1955	1987	1	1	0	0	0	0	1
AJ000037912	1955	1991	1	1	0	0	0	0	1
AJ000037981	1959	1987	1	1	0	0	0	0	1
AJ000037989	1936	2017	5	1	0	1	1	1	5
ALE00100939	1940	2000	2	1	0	0	1	0	2
AM000037683	1936	1988	1	1	0	0	0	0	1
AM000037698	1959	1976	1	1	0	0	0	0	1
AM000037719	1912	1992	5	1	0	1	1	1	5

only showing top 20 rows

In total 20300 stations collected all five core elements and 16159 collected precipitation and no other elements.

After joining the datasets all together, it was shown that all stations were included in the the new stations file.

I think it would be incredibly expensive to join all of daily and stations as just one year of daily contained thousands of observations and there are 260 years of data to then be joined. Using the subtract function allows us to check if there are any stations missing withoug using LEFT JOIN, and this confirms that there aren't any missing.

Analysis:

Question 1:

There are 122047 stations in total, of those 42588 are active in 2021. There are 991 stations in the GSN and 1218 in the HCN, there are none in the CRN for 2021. 14 of these are in more than one of the aforementioned networks.

In the Southern Hemisphere there are 25337 stations. There are also 354 stations that are located in territories of the United States.

Question 2:

Using the Haversine formula we can compute the geographical distance between two stations using their longitude and latitude. In this instance the two stations of the closest distance in New Zealand are Wellington Aero and Paraparaumu.

```
+-----------+------------------+----------+-----------+-----------+---------------+----------+-----------+--------+
|ID_1       |STATION_1         |LATITUDE_1|LONGITUDE_1|ID_2       |STATION_2      |LATITUDE_2|LONGITUDE_2|DISTANCE|
+-----------+------------------+----------+-----------+-----------+---------------+----------+-----------+--------+
|NZM00093439|WELLINGTON AERO AWS|-41.333   |174.8      |NZ000093417|PARAPARAUMU AWS|-40.9     |174.983    |50.54424|
+-----------+------------------+----------+-----------+-----------+---------------+----------+-----------+--------+
only showing top 1 row
```

There are 134217728 blocks within the default blocksize of HDFS. Blocks required for daily 2022 is 1, compared to the blocks required for daily 2021 which is 2 as the individual blocksize averages 79799197 B. As the files are compressed it is not possible for Spark to load and apply transformations in parallel for 2022, this is the same for 2021. This is because the files need to be loaded to be processed which means transformations cannot happen in parallel.

3b)

Job Id (Job Group)	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
69	count at NativeMethodAccessorImpl.java:0 count at NativeMethodAccessorImpl.java:0	2022/09/22 18:58:53	13 s	2/2	2/2
68	count at NativeMethodAccessorImpl.java:0 count at NativeMethodAccessorImpl.java:0	2022/09/22 18:58:45	7 s	2/2	2/2

Stage 68

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
101	count at NativeMethodAccessorImpl.java:0 +details	2022/09/22 18:58:53	25 ms	1/1			59.0 B	
100	count at NativeMethodAccessorImpl.java:0 +details	2022/09/22 18:58:45	7 s	1/1	84.1 MiB			59.0 B

Stage 69

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
103	count at NativeMethodAccessorImpl.java:0 +details	2022/09/22 18:59:06	17 ms	1/1			59.0 B	
102	count at NativeMethodAccessorImpl.java:0 +details	2022/09/22 18:58:53	13 s	1/1	152.2 MiB			59.0 B

As we can see above each load and counting of observations in 2021 and then separately in 2022 both 2 stages with 2 tasks. Each stage only executed on task. The number of tasks executed corresponded with the number of blocks for the 2021 observations but not for the 2022 data as only 1 block was used.

3c)

Job Id (Job Group)	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
74	count at NativeMethodAccessorImpl.java:0 count at NativeMethodAccessorImpl.java:0	2022/09/22 19:47:24	20 s	2/2	10/10

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
113	count at NativeMethodAccessorImpl.java:0 +details	2022/09/22 19:47:44	27 ms	1/1			531.0 B	
112	count at NativeMethodAccessorImpl.java:0 +details	2022/09/22 19:47:24	20 s	9/9	1492.4 MiB			531.0 B

There were 2 stages with 10 tasks for the 2014-2022 data, one stage had 1 task and the other stage had 9 tasks, this corresponds with loading 9 years' worth of data and then performing one count operation.

Each of these parts uses a core so the more splits in the data during the compressed, this would allow more data to be parallel processed. Can't work parallel when they are compressed. It will try to put a compressed file into one partition but if it can't fit it will split part of it off.

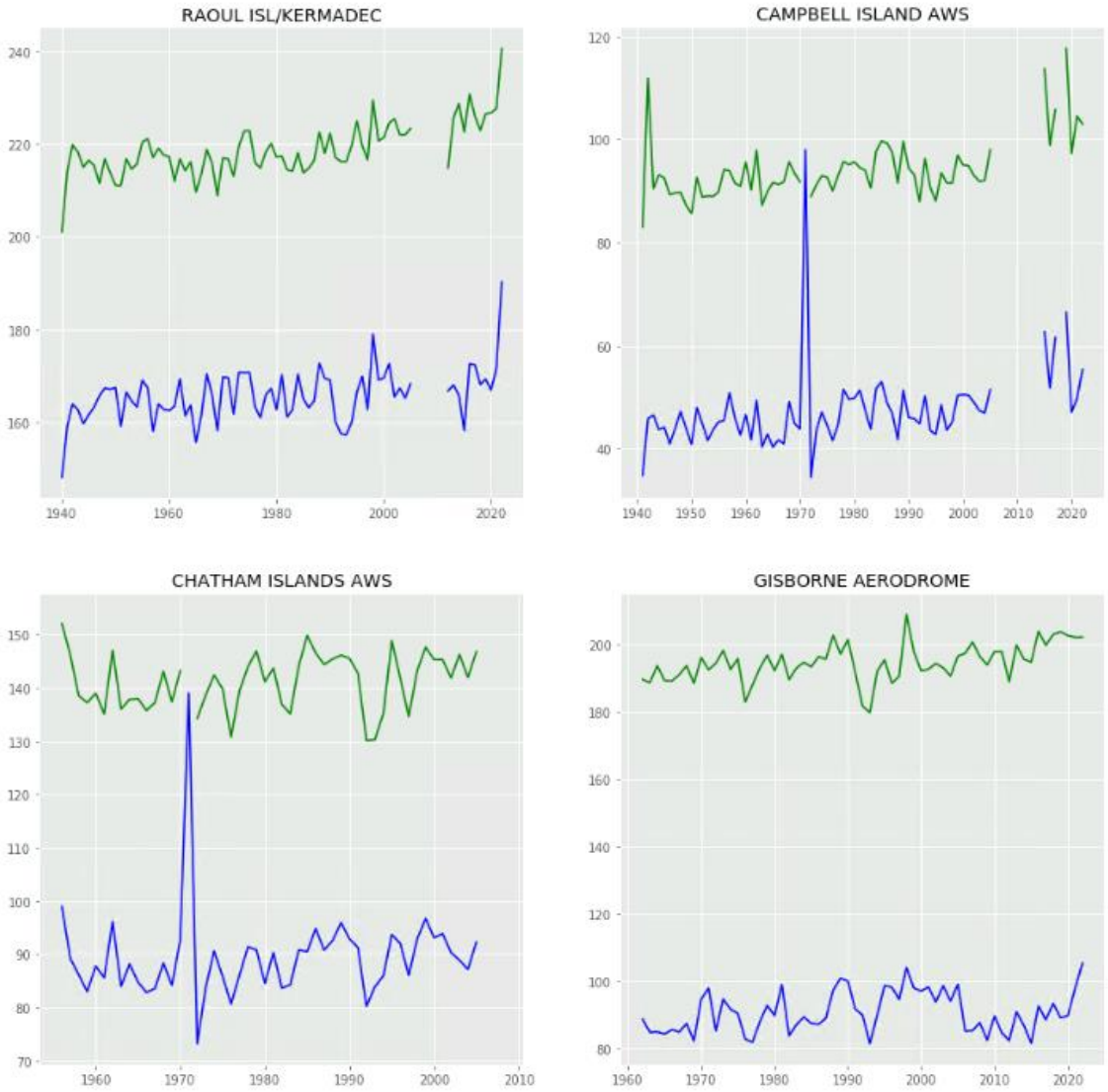
The number of rows in the total daily file is 303501016 and the number of element observations is displayed in the table below.

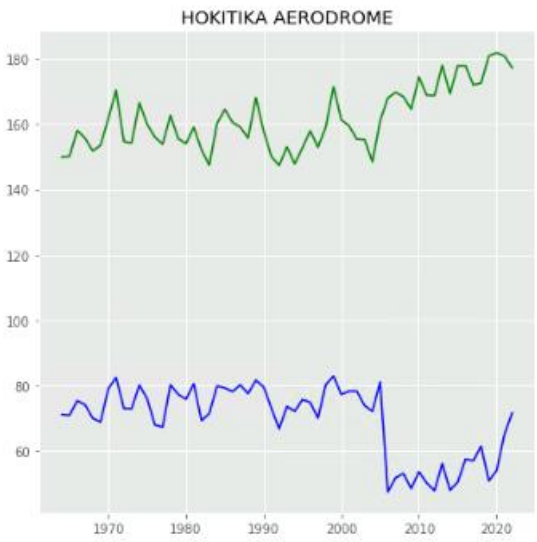
ELEMENT	ELEMENT_COUNT
PRCP	1048156273
TMAX	447084093
TMIN	445687425
SNOW	344268930
SNWD	290998195

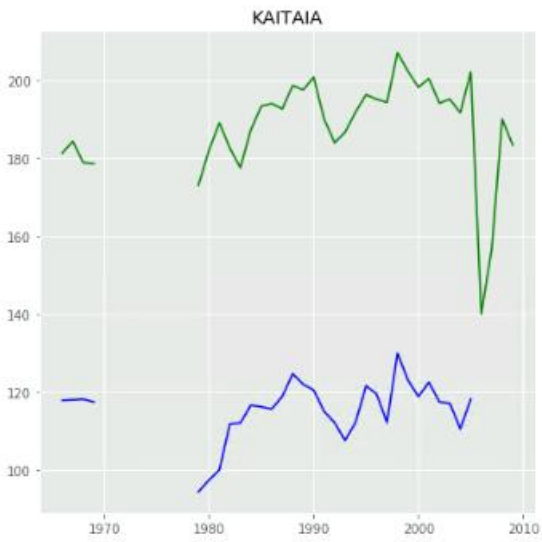
The element with the most observations is PRCP. There were 8848299 stations that had corresponding TMIN and TMACX observations, of which 27678 distinct stations contributed to the observations.

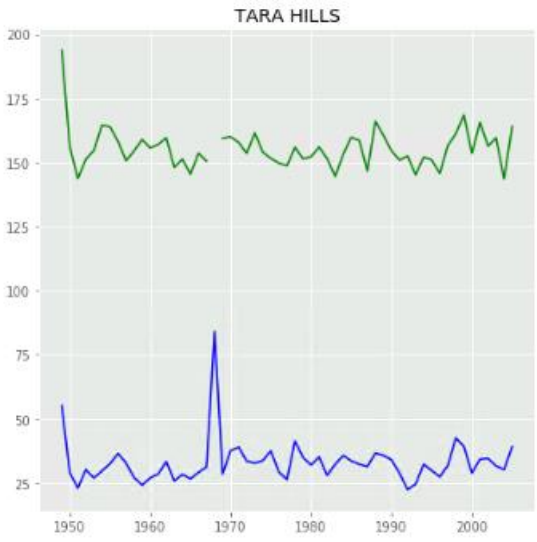
There are 474654 observations of TMIN and TMAX for all stations in New Zealand, these cover the years between 1940 and 2022. The observation count is also confirmed by the use of the wc -l bash command.

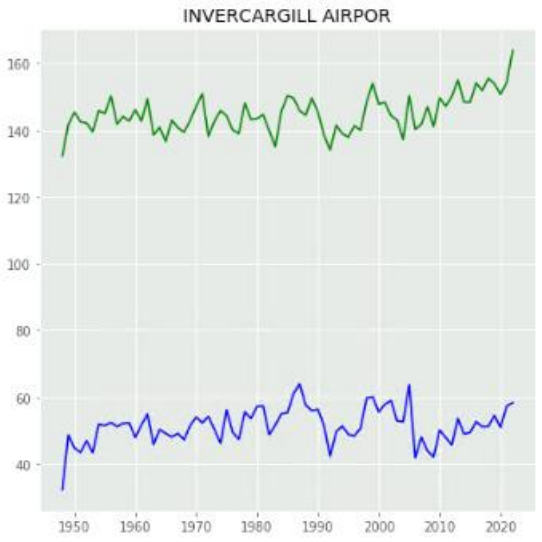
Time-Series for TMIN and TMAX for each station in NZ

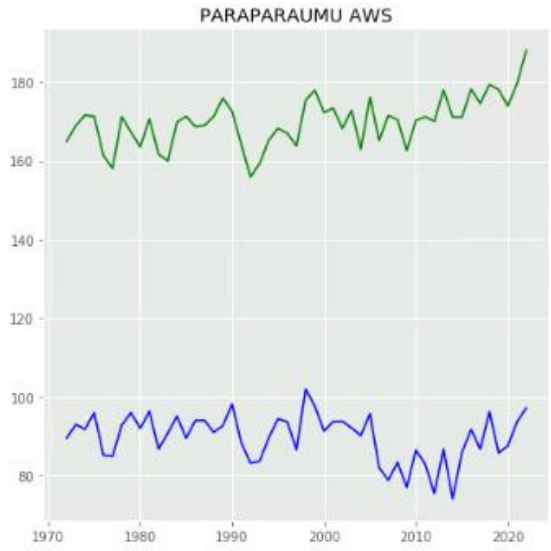


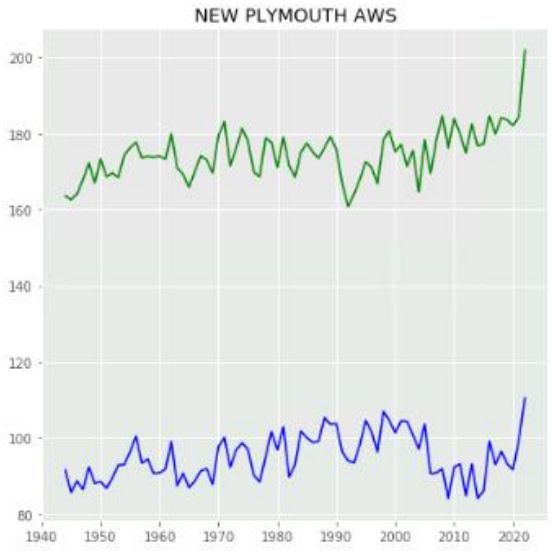

HOKITIKA AERODROME
180
160
140
120
100
80
60
1970
1980
1990
2000
2010
2020

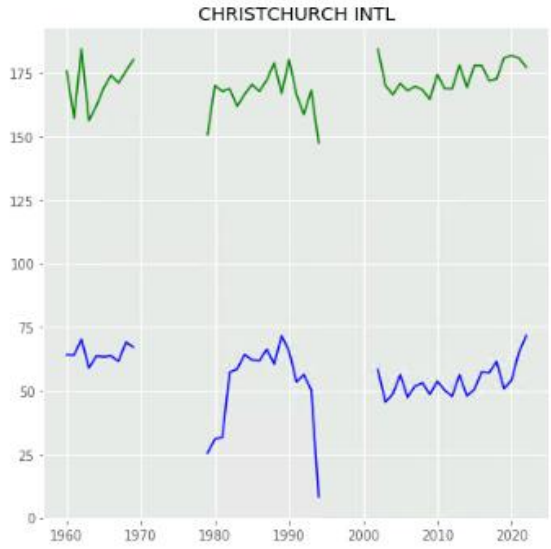

KAITAIA
200
180
160
140
120
100
1970
1980
1990
2000
2010

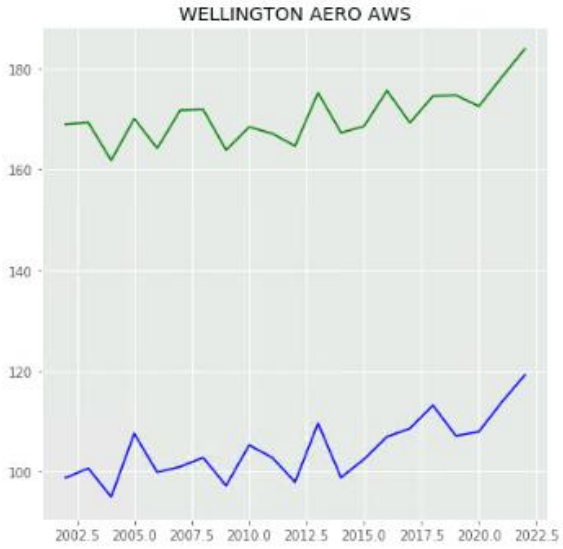

TARA HILLS
200
175
150
125
100
75
50
25
1950
1960
1970
1980
1990
2000

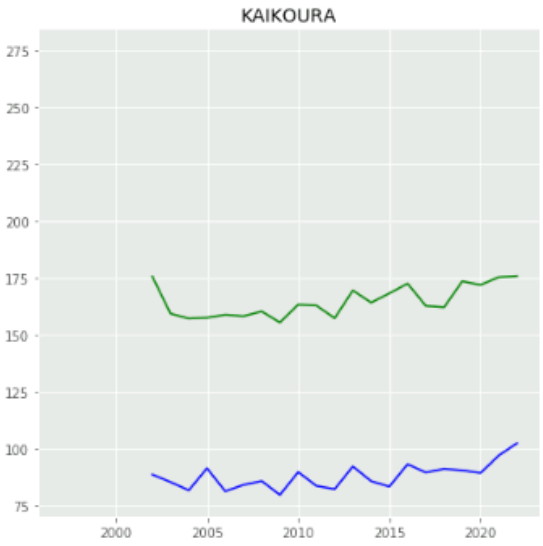

INVERCARGILL AIRPOR
160
140
120
100
80
60
40
1950
1960
1970
1980
1990
2000
2010
2020

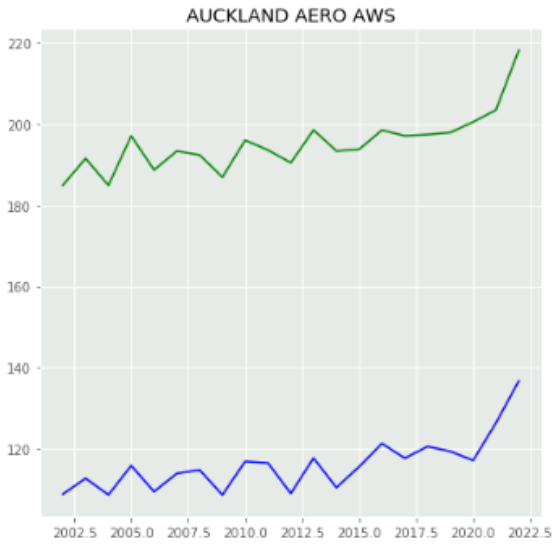

PARAPARAUMU AWS
180
160
140
120
100
80
1970
1980
1990
2000
2010
2020

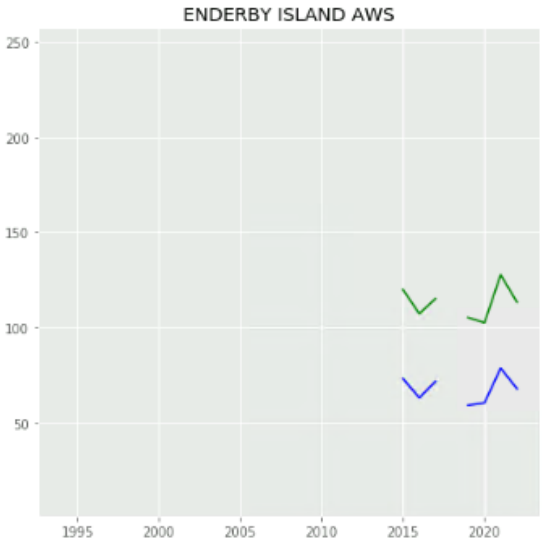

NEW PLYMOUTH AWS
200
180
160
140
120
100
80
1940
1950
1960
1970
1980
1990
2000
2010
2020


CHRISTCHURCH INTL
175
150
125
100
75
50
25
0
1960
1970
1980
1990
2000
2010
2020

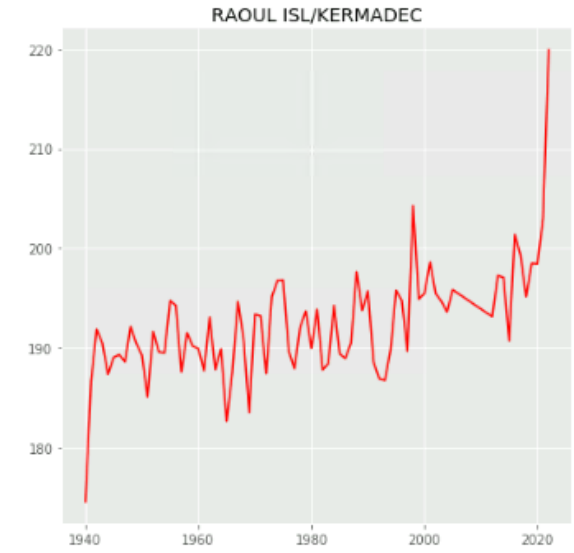

WELLINGTON AERO AWS
180
160
140
120
100
2002.5
2005.0
2007.5
2010.0
2012.5
2015.0
2017.5
2020.0
2022.5

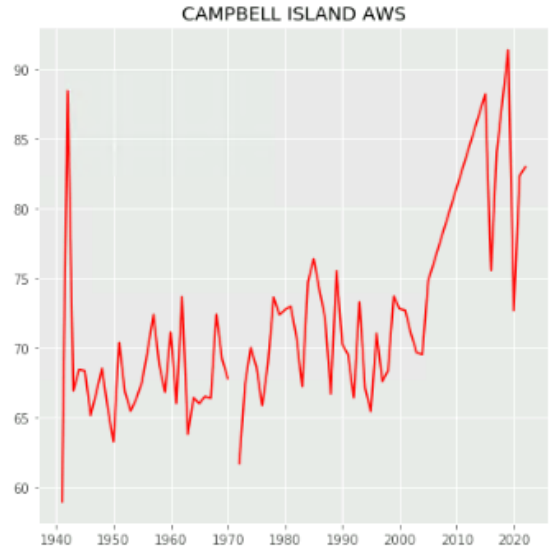

KAIKOURA
275
250
225
200
175
150
125
100
75
2000
2005
2010
2015
2020

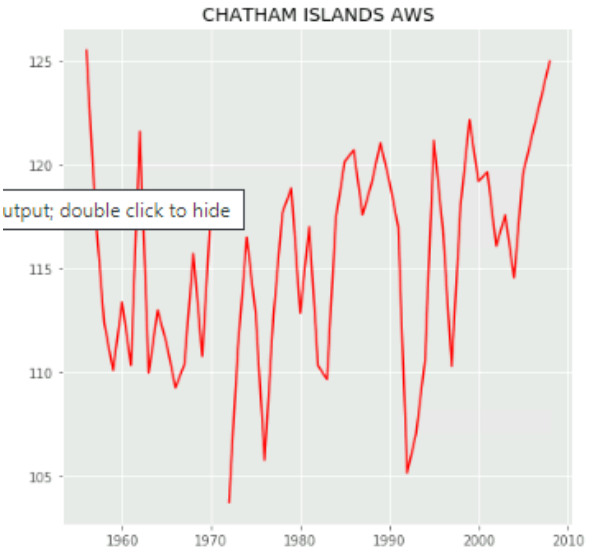

AUCKLAND AERO AWS
220
200
180
160
140
120
2002.5
2005.0
2007.5
2010.0
2012.5
2015.0
2017.5
2020.0
2022.5


ENDERBY ISLAND AWS
250
200
150
100
50
1995
2000
2005
2010
2015
2020

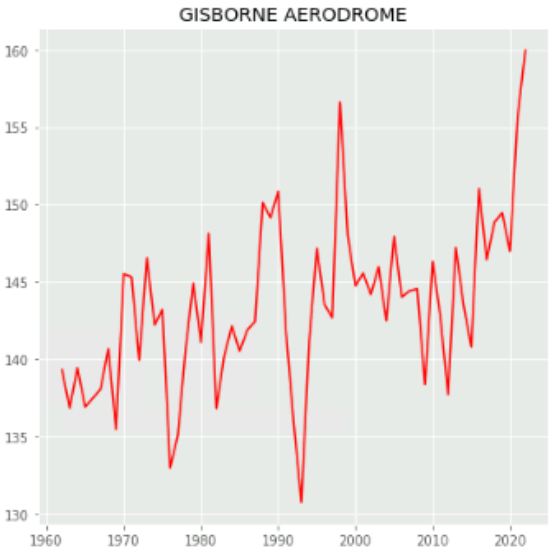
Average Time Series for TMIN and TMAX in NZ:

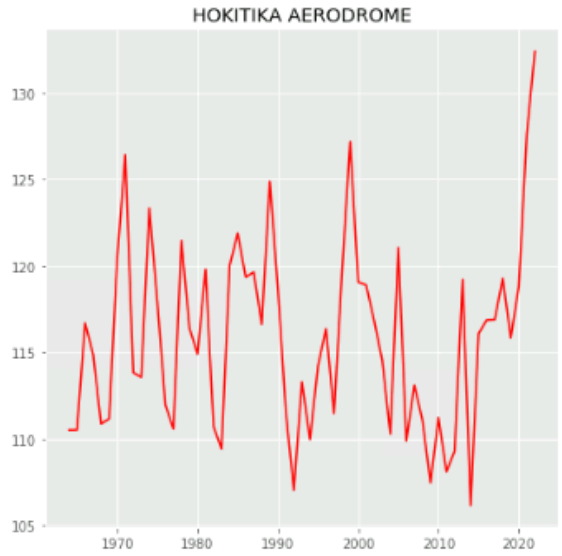


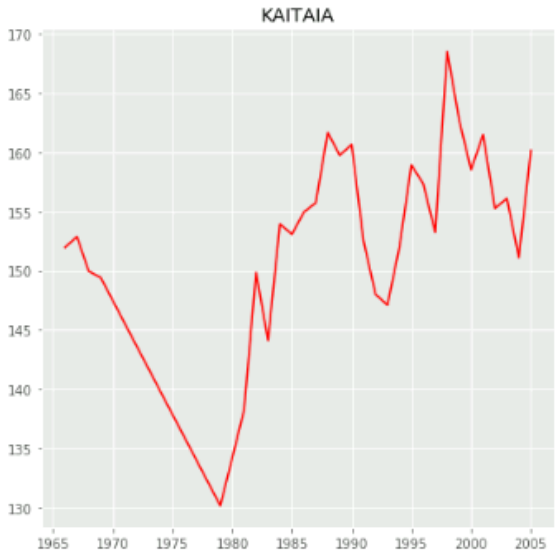


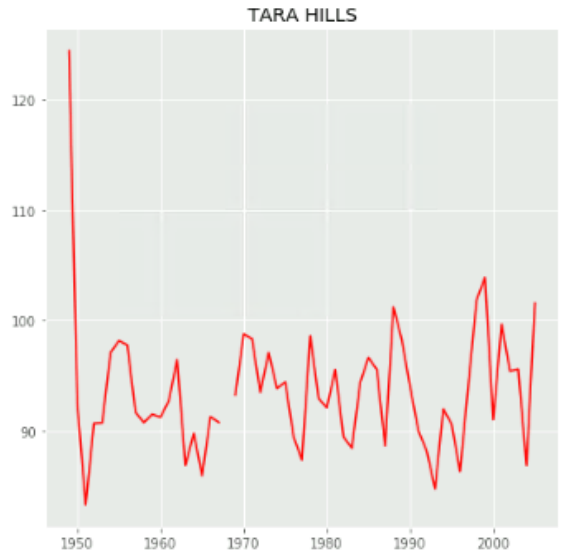


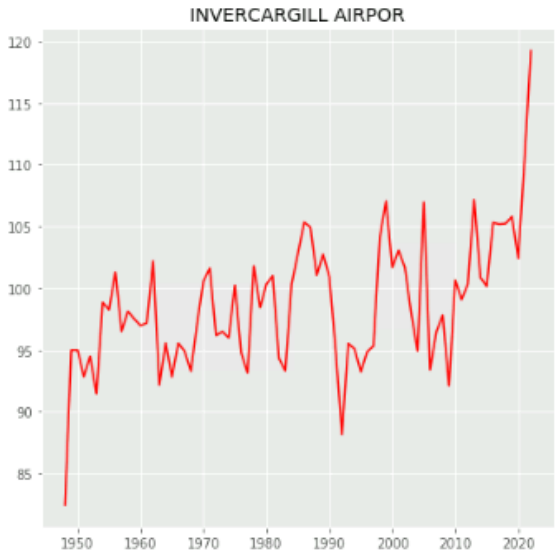
utput; double click to hide

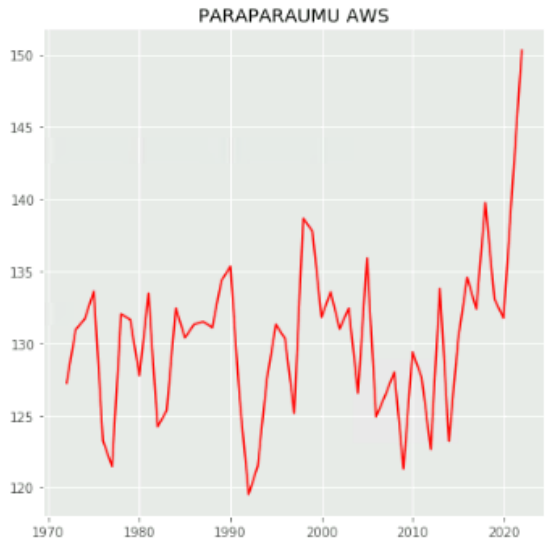


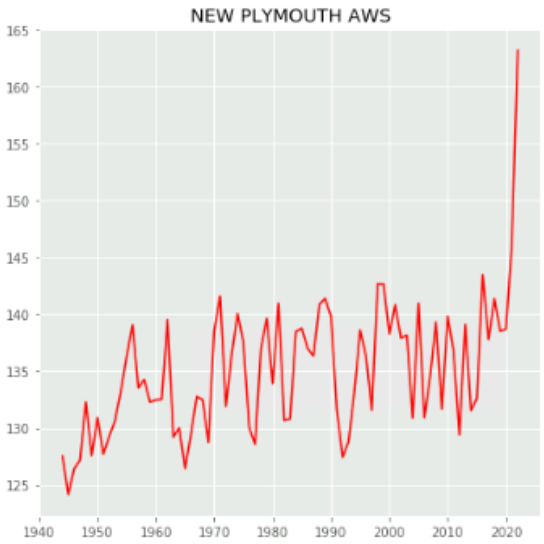

HOKITIKA AERODROME
130
125
120
115
110
105
1970
1980
1990
2000
2010
2020

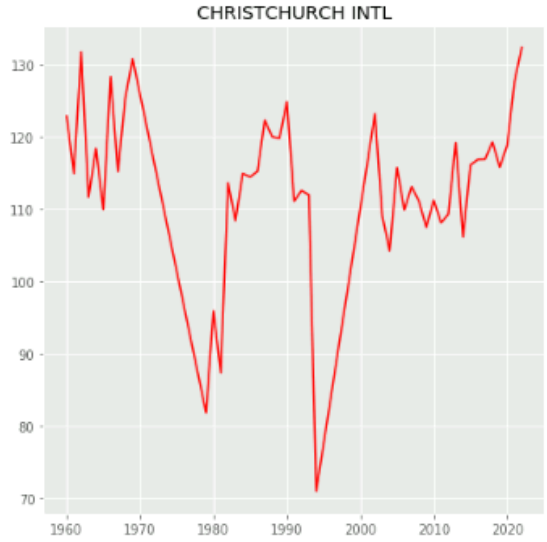

KAITAIA
170
165
160
155
150
145
140
135
130
1965
1970
1975
1980
1985
1990
1995
2000
2005

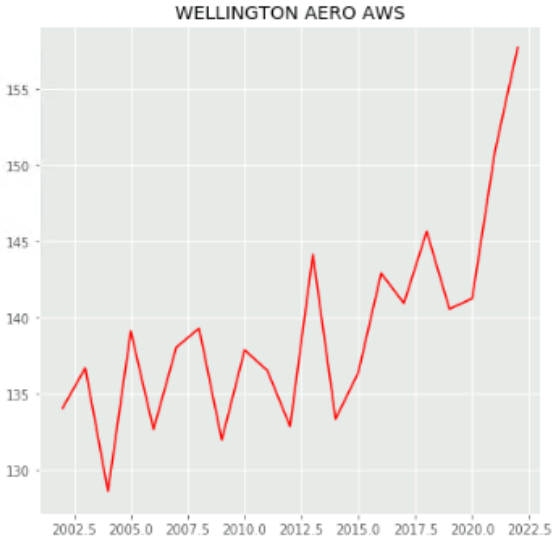

TARA HILLS
120
110
100
90
1950
1960
1970
1980
1990
2000

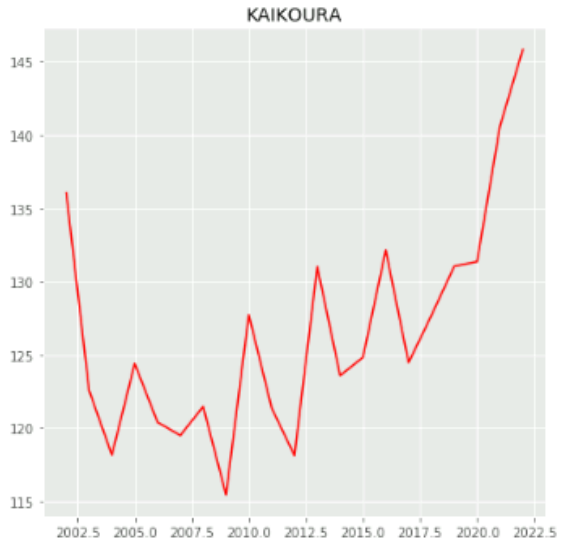

INVERCARGILL AIRPOR
120
115
110
105
100
95
90
85
1950
1960
1970
1980
1990
2000
2010
2020

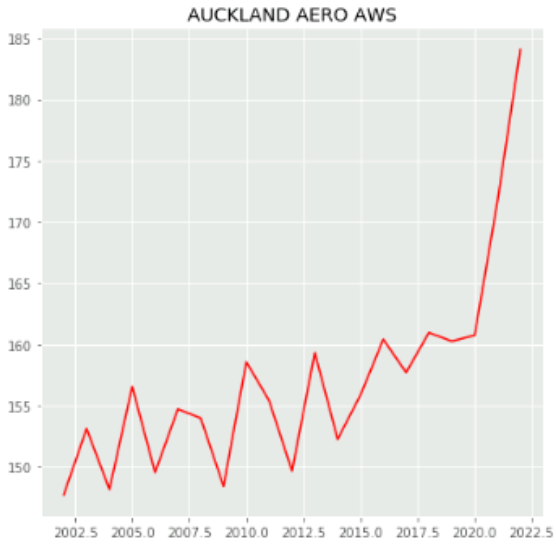

PARAPARAUMU AWS
150
145
140
135
130
125
120
1970
1980
1990
2000
2010
2020

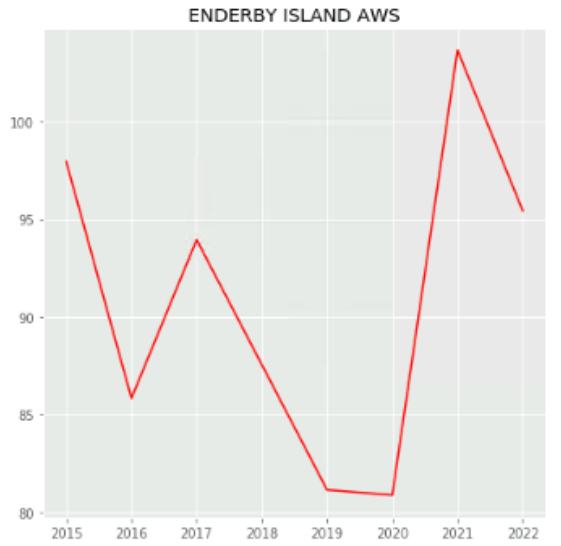

NEW PLYMOUTH AWS
165
160
155
150
145
140
135
130
125
1940
1950
1960
1970
1980
1990
2000
2010
2020


CHRISTCHURCH INTL
130
120
110
100
90
80
70
1960
1970
1980
1990
2000
2010
2020


WELLINGTON AERO AWS
155
150
145
140
135
130
2002.5
2005.0
2007.5
2010.0
2012.5
2015.0
2017.5
2020.0
2022.5







The country with the highest average rainfall in a single year is Equatorial Guinea, with 4361.0. This could be completely sensible as it is a tropical climate which is known for a lot of rain.