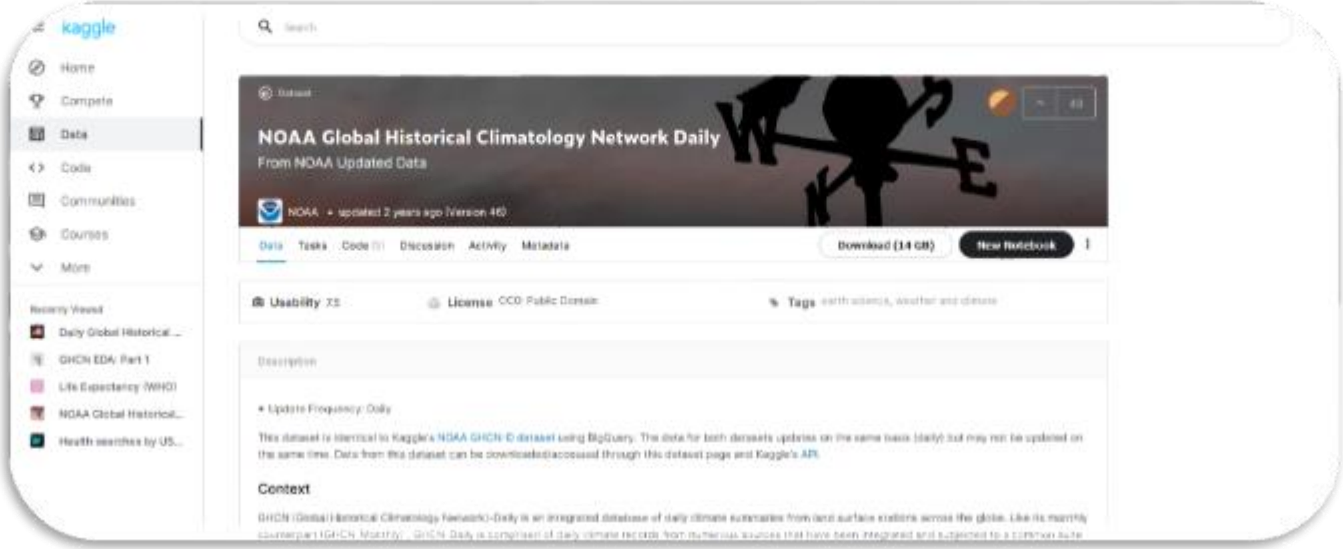
Term Project Group 2

By Riker Santivong, Francisco Cortes, Kevin Danao, Kevin De La Torre

DATA SET LOCATION



NOAA Global Historical Climatology Network Daily

From NOAA Updated Data

NOAA • updated 2 years ago (Version 46)

- Update Frequency: Daily

This dataset is identical to Kaggle's NOAA GHCN-D dataset using BigQuery. The data for both datasets updates on the same basis (daily) but may not be updated on the same time. Data from this dataset can be downloaded/accessed through this dataset page and Kaggle's API.

Context

GHCN (Global Historical Climatology Network)-Daily is an integrated database of daily climate summaries from land surface stations across the globe. Like its monthly counterpart (GHCN-Monthly) , GHCN-Daily is comprised of daily climate records from numerous sources that have been integrated and subjected to a common suite

- **FILESIZE**
- COMPRESSED: 14GB
- UNCOMPRESSED: 93GB
- Over 253 CSV files
- Files varied in size: 24kb – 1.3GB

https://www.kaggle.com/noaa/noaa-global-historical-climatology-network-daily

Github Link: https://github.com/fcortes19/CIS-3200-Project

OVERVIEW

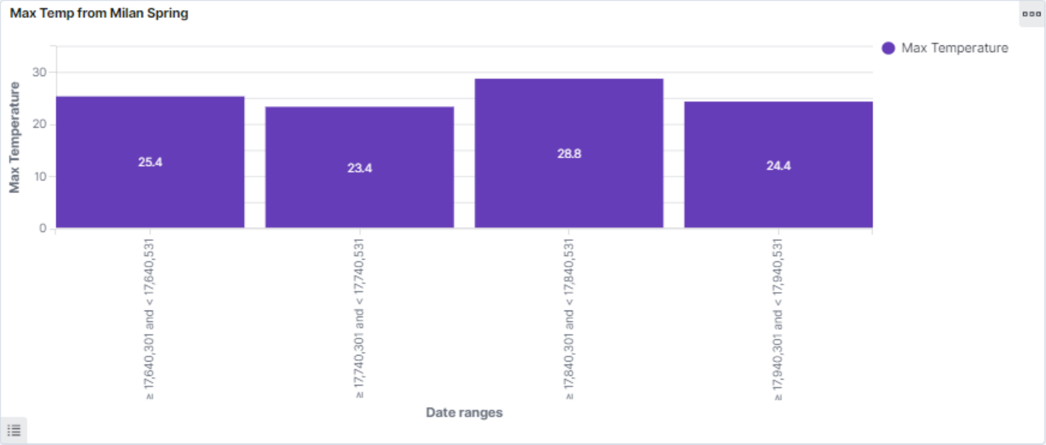
- To visualize and analyze data to help identify possible signs of global warning and rising temperatures during a given time period
- Utilizing data collected from 1764 to 1894 in intervals of 10 years
- Visualize minimum, maximum and average temperatures in a single location

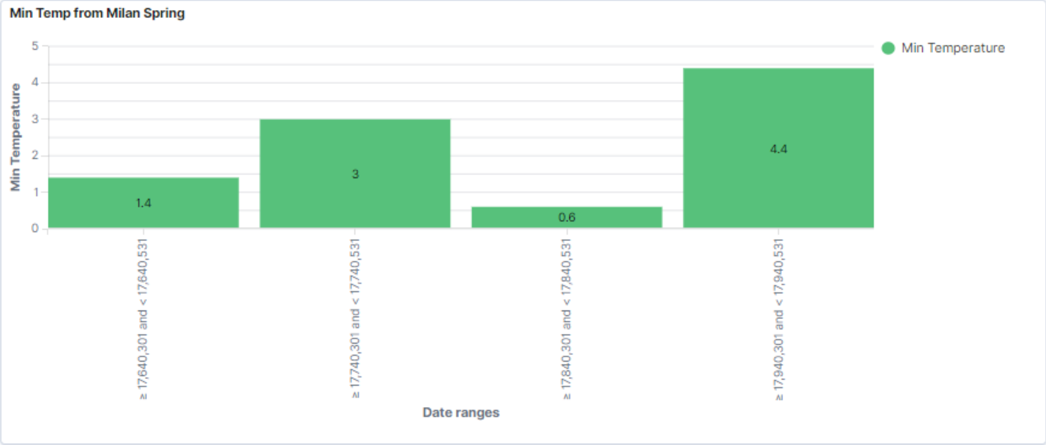
Hardware Specification

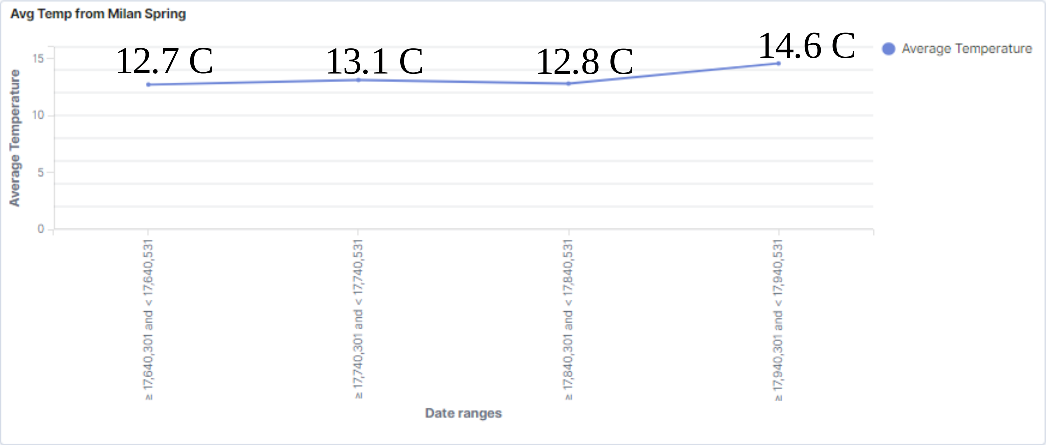
ELASTICSEARCH	
Storage	240 GB
Memory	8 GB
Master memory	1 GB
KIBANA	
Memory	1 GB
ML	
Memory	1 GB
APM	
Memory	512 MB
TOTAL	
Total Storage	240 GB
Total Memory	11.5 GB

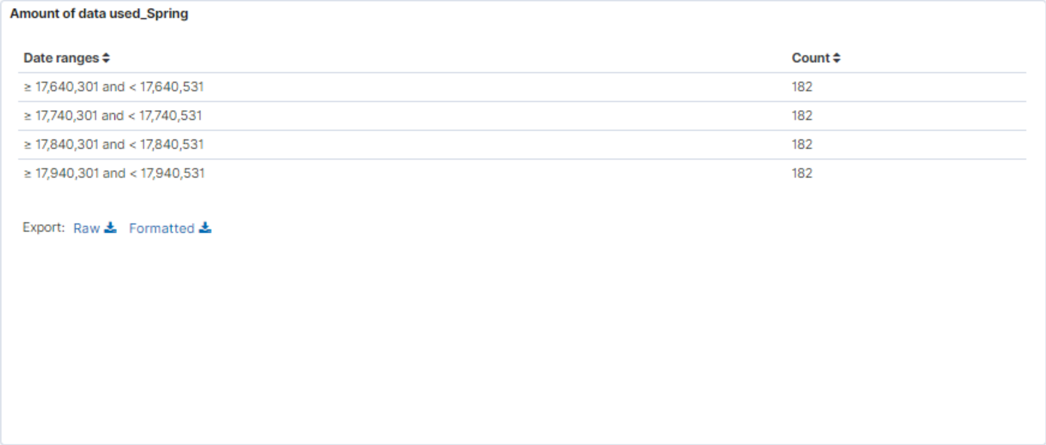
Hardware Specification (cont)

Microsoft Azure Machine Learning Studio	
Max Storage Space	10 GB
Execution/performance	Single Node
Max number of modules per experiment	100
Max experiment duration	1 hour per experiment


Max Temp from Milan Spring


Min Temp from Milan Spring


Avg Temp from Milan Spring

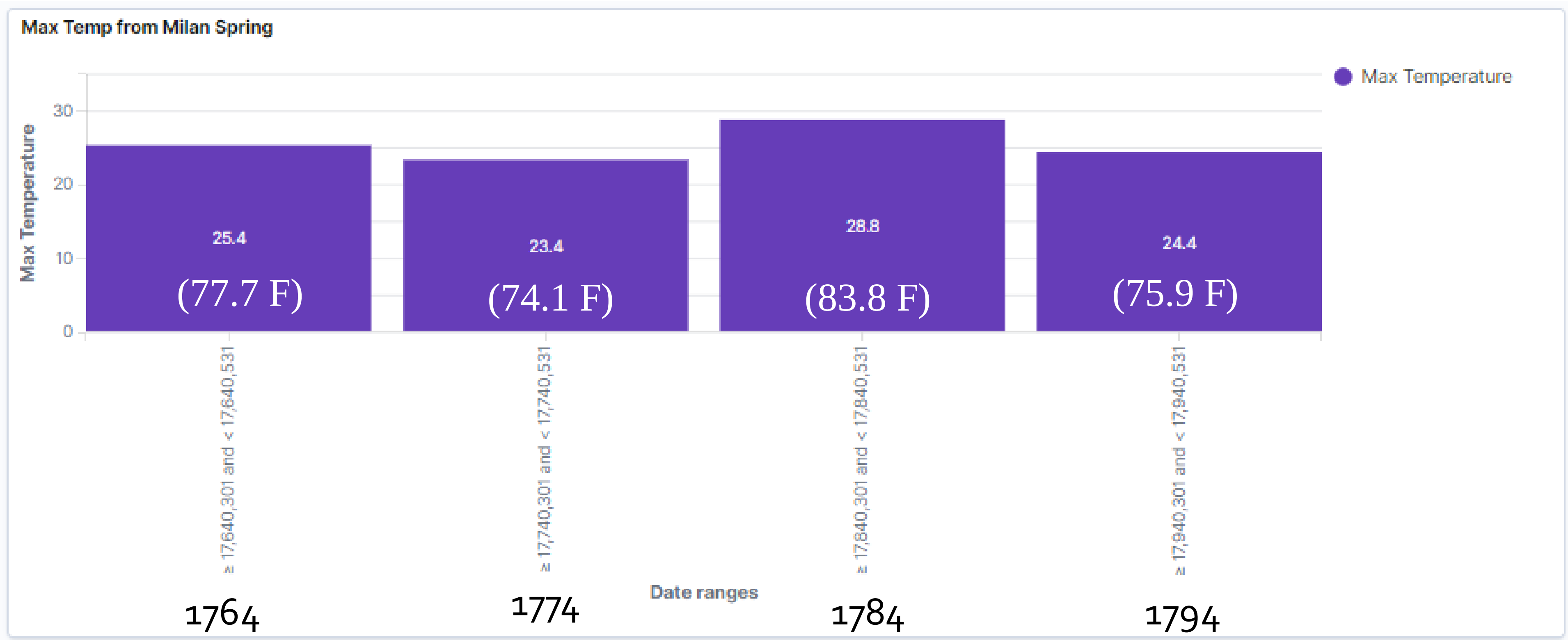

Amount of data used_Spring

Date ranges	Count
≥ 17,640,301 and < 17,640,531	182
≥ 17,740,301 and < 17,740,531	182
≥ 17,840,301 and < 17,840,531	182
≥ 17,940,301 and < 17,940,531	182

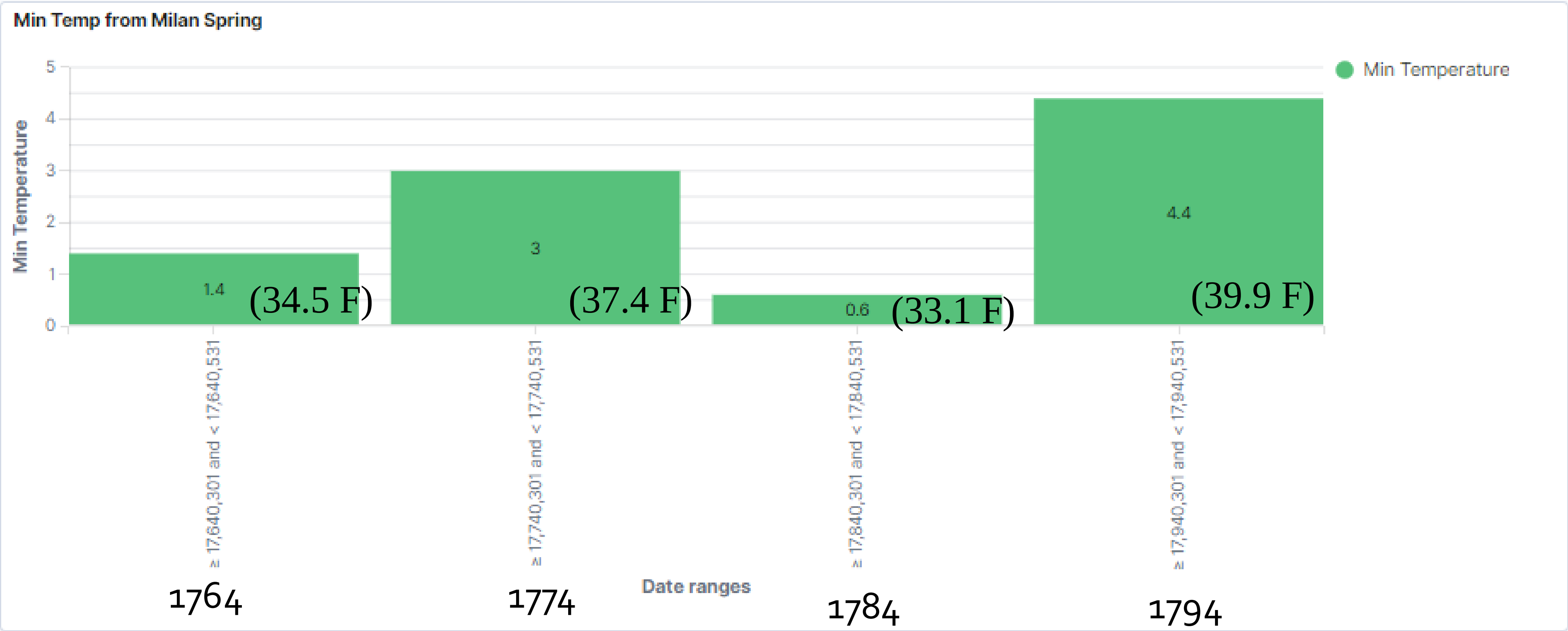
Export: Raw Formatted

Dashboard

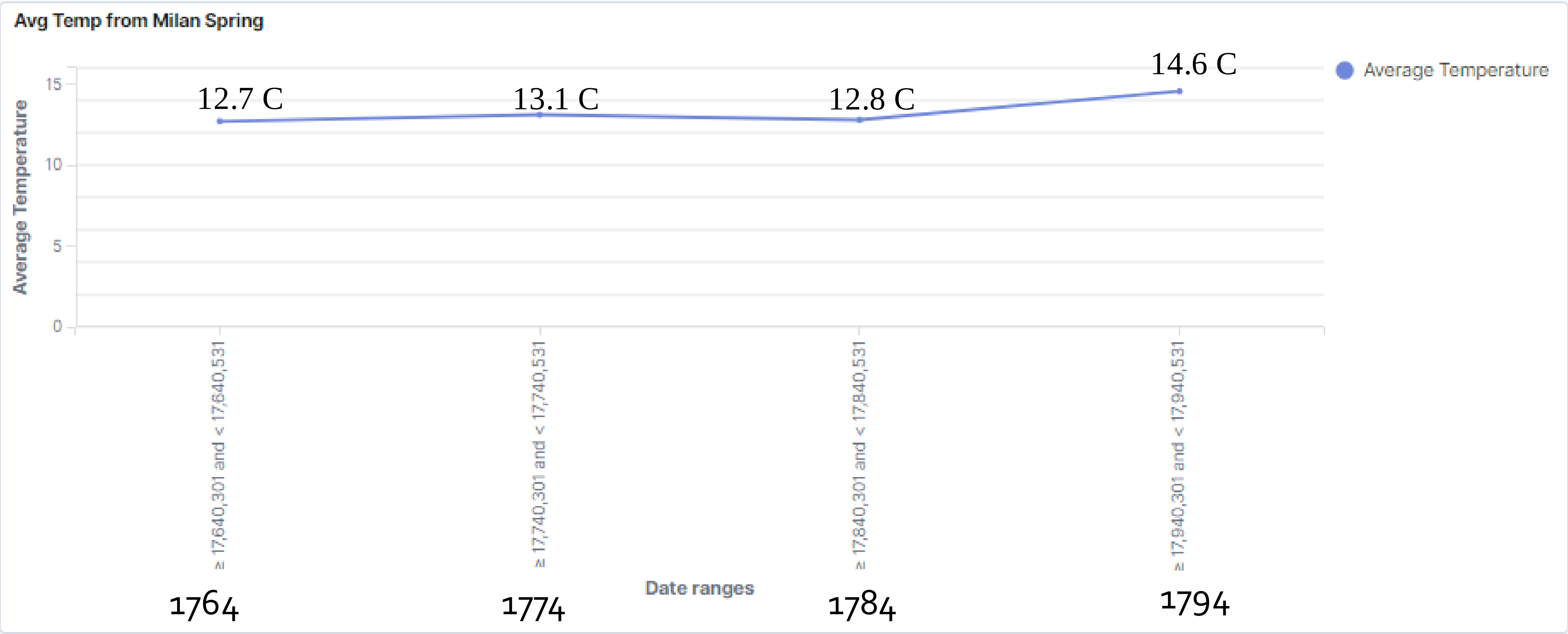
Spring Seasons Temps from Year: 1764, 1774, 1784, 1794



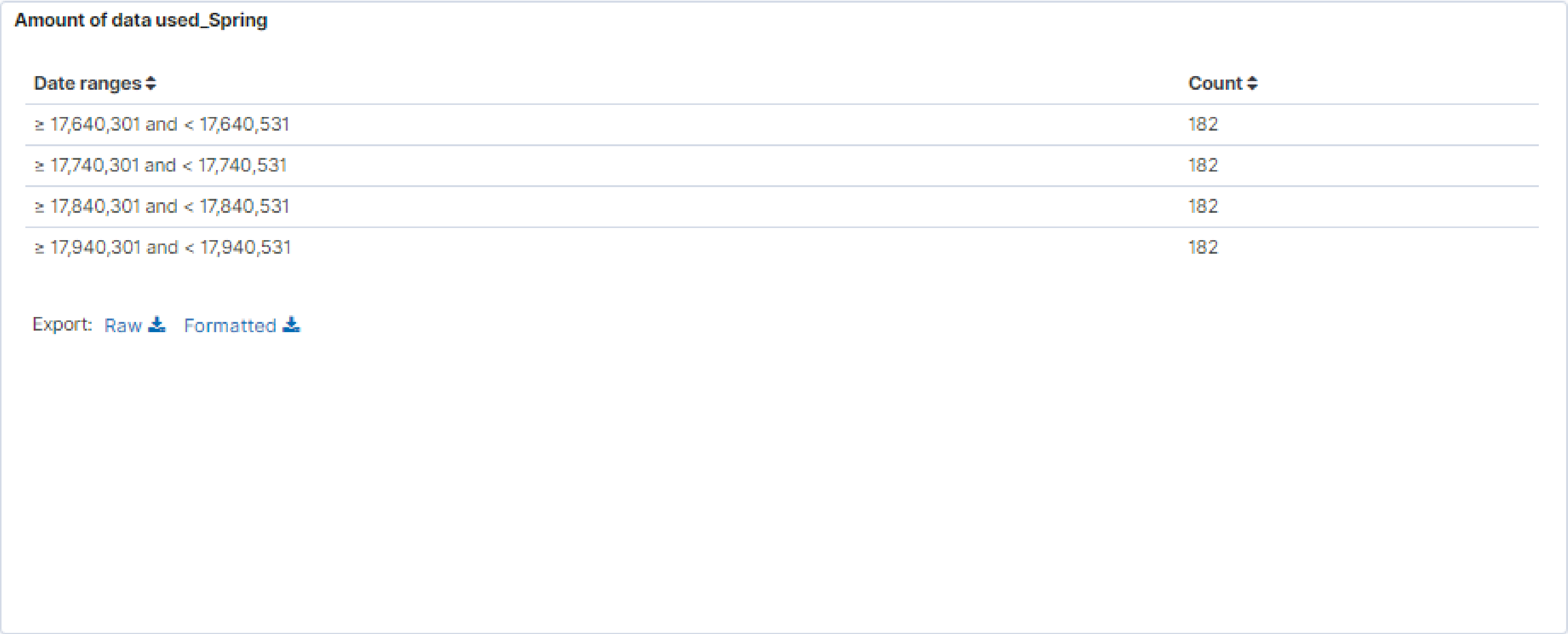
Max Spring Temperatures from Year: 1764, 1774, 1784, 1794



Spring Temperatures from Year: 1764, 1774, 1784, 1794



Spring Temperatures from Year: 1764, 1774, 1784, 1794

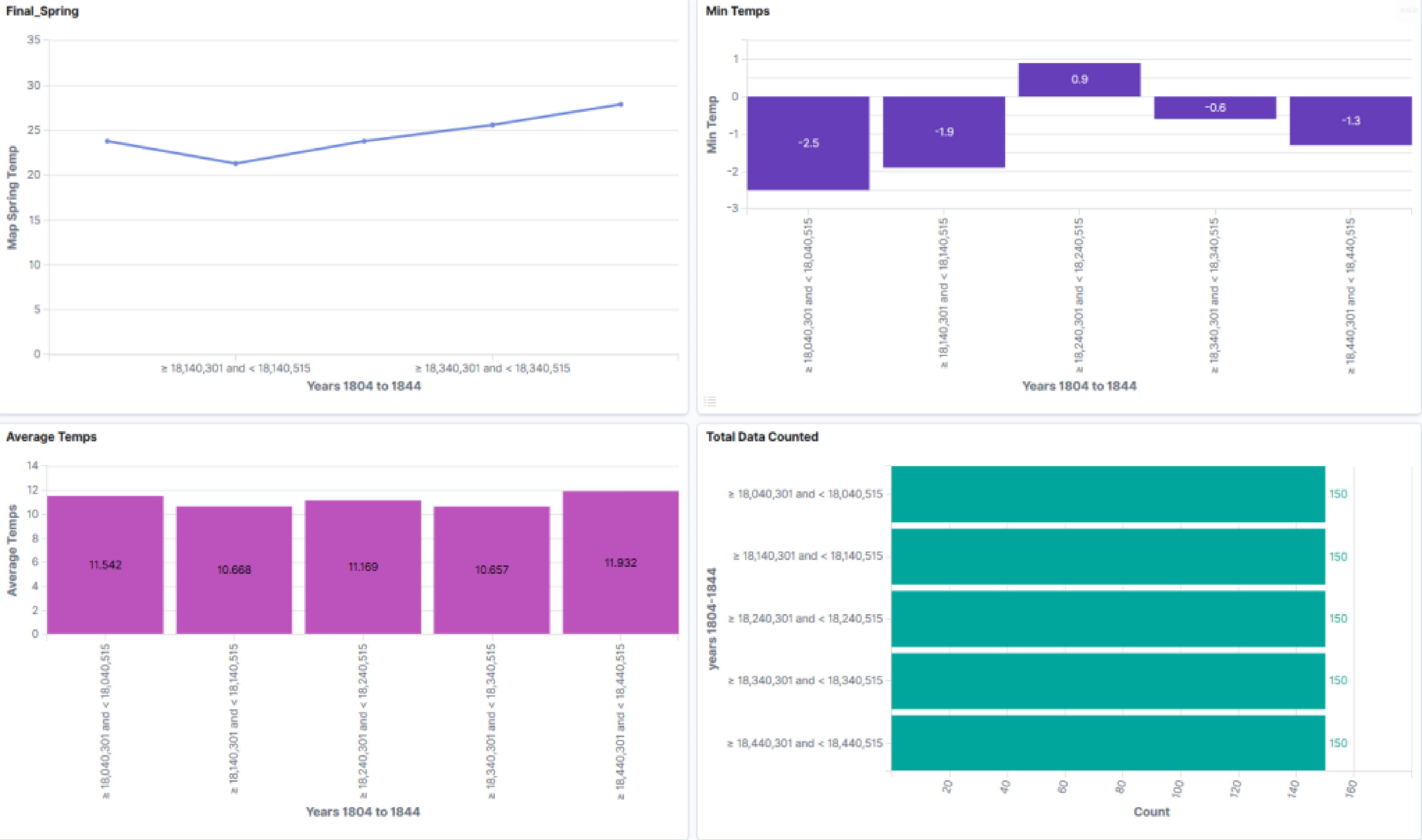


Amount of data used_Spring

Date ranges ⬍	Count ⬍
≥ 17,640,301 and < 17,640,531	182
≥ 17,740,301 and < 17,740,531	182
≥ 17,840,301 and < 17,840,531	182
≥ 17,940,301 and < 17,940,531	182

Export: Raw ⭳ Formatted ⭳

Spring Temperatures from Year: 1764, 1774, 1784, 1784



Years 1804 to 1844

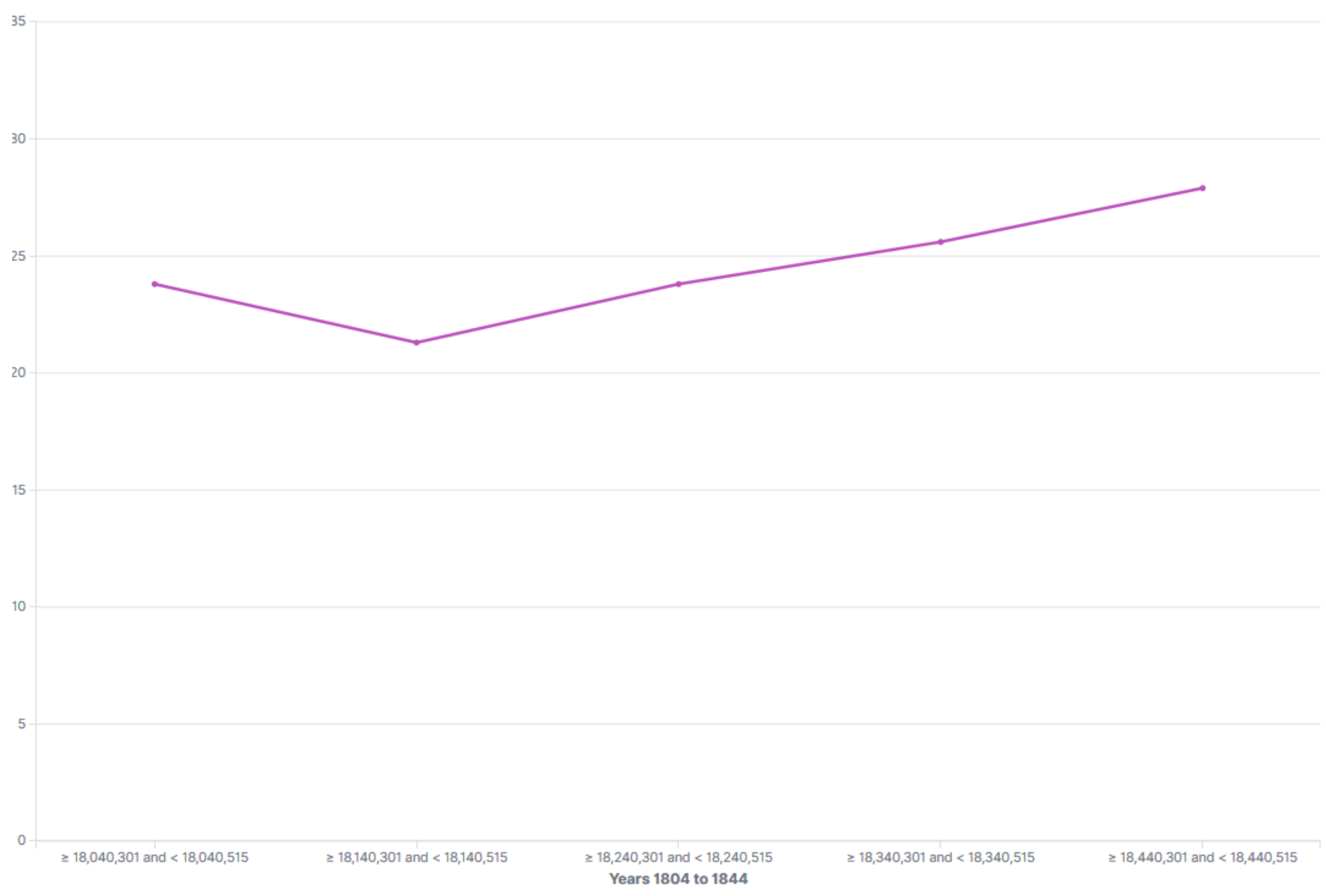
Max Spring Temp

Lowest Max Temp was for 1814, 21.3C (70.3F)

Temps rose by 10%

Highest temp in 1844, 27.9C(82.2F)

An increase of 30% from the lowest temp



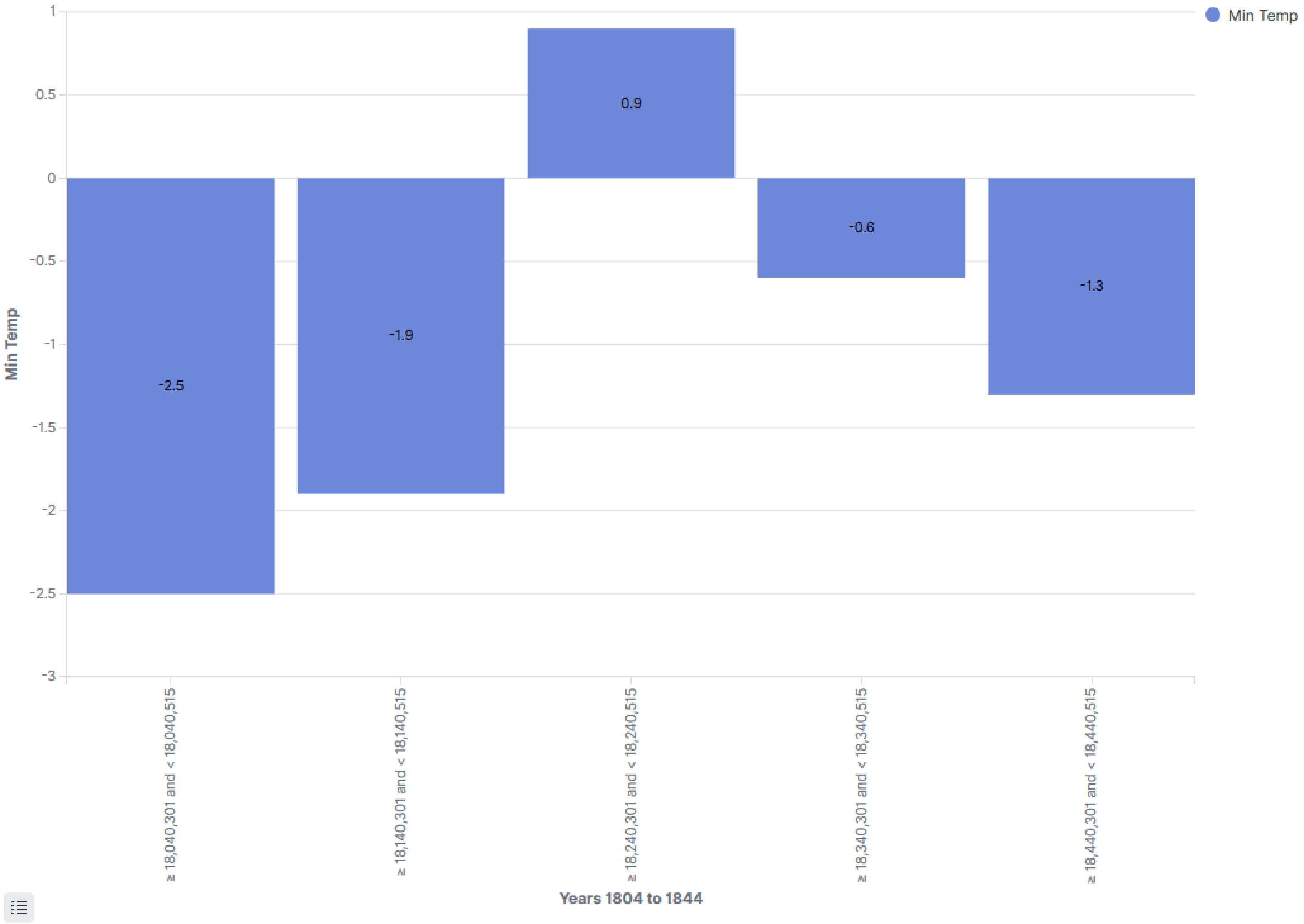
Min Spring Temp

Lowest temp recorded, 1804 –2.5C (27.5F)

Highest low temp is 1824, 0.9C (33.6F)

1824 only year above 0, all others below

Compared to 1764-1794 these are the lowest temps recorded for spring

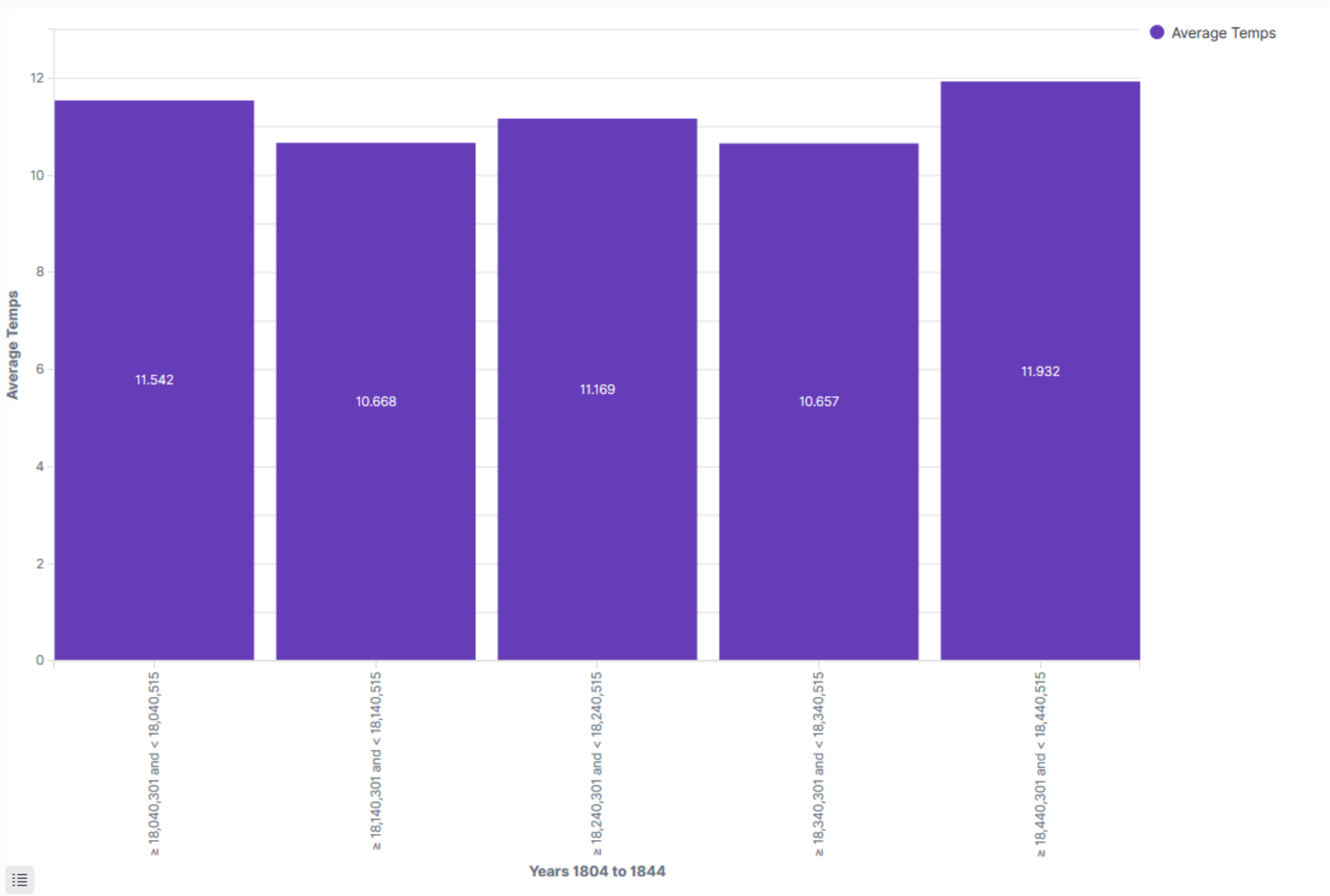


Average Spring Temps

The highest year with avg temps 1844, 11.9C (53.4F)

1814-1834 average temp were approx 51.4F

Average temps from 1764-1794 seem to be higher

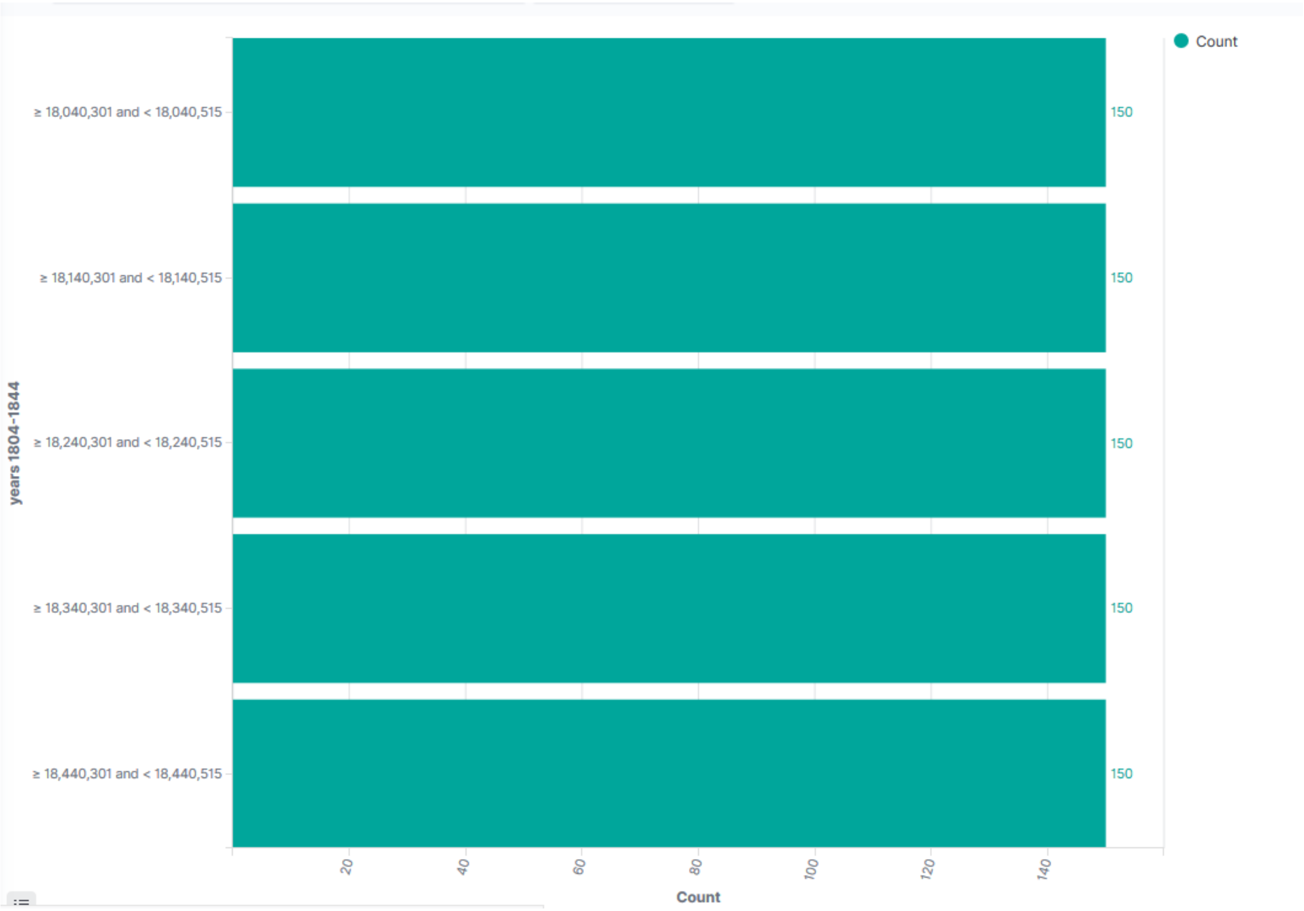


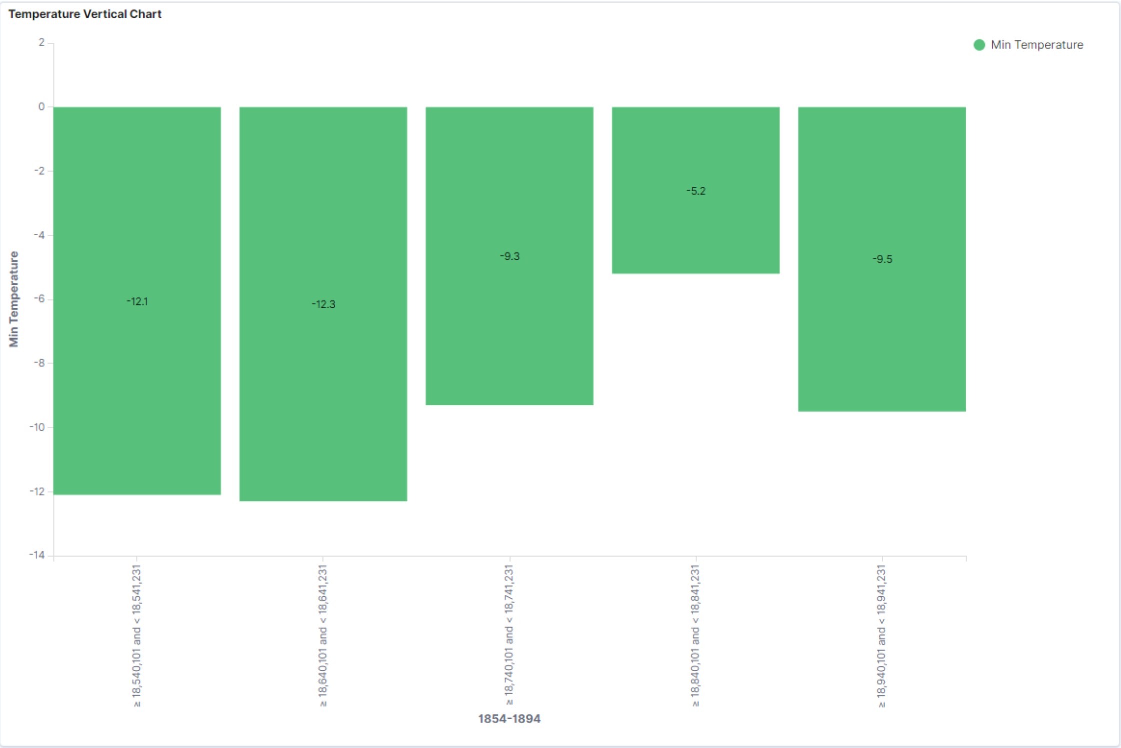
Total Number of Rows

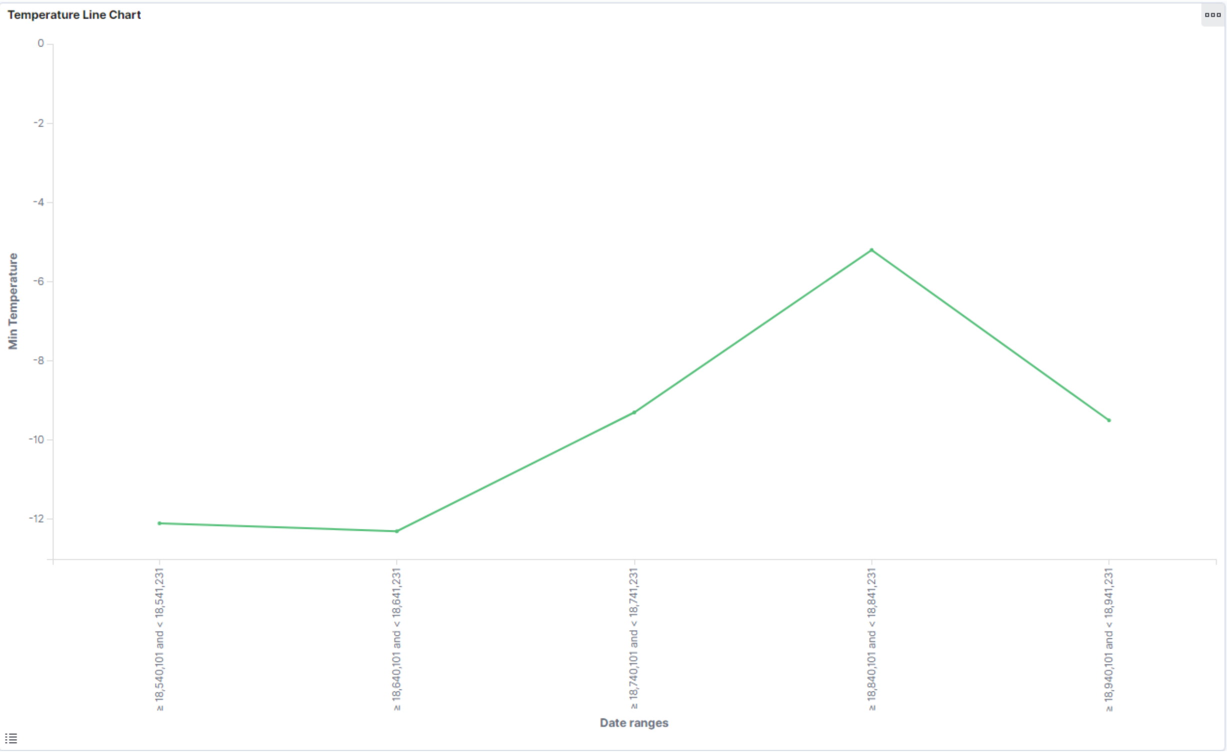
150 rows of data were viewed

All from a single location/same location (Milan)

Different than previous years, due to possibly less important data being collected







Years 1854 to 1894

Max Spring Temp

1854: 33.6°C (92.48°F)

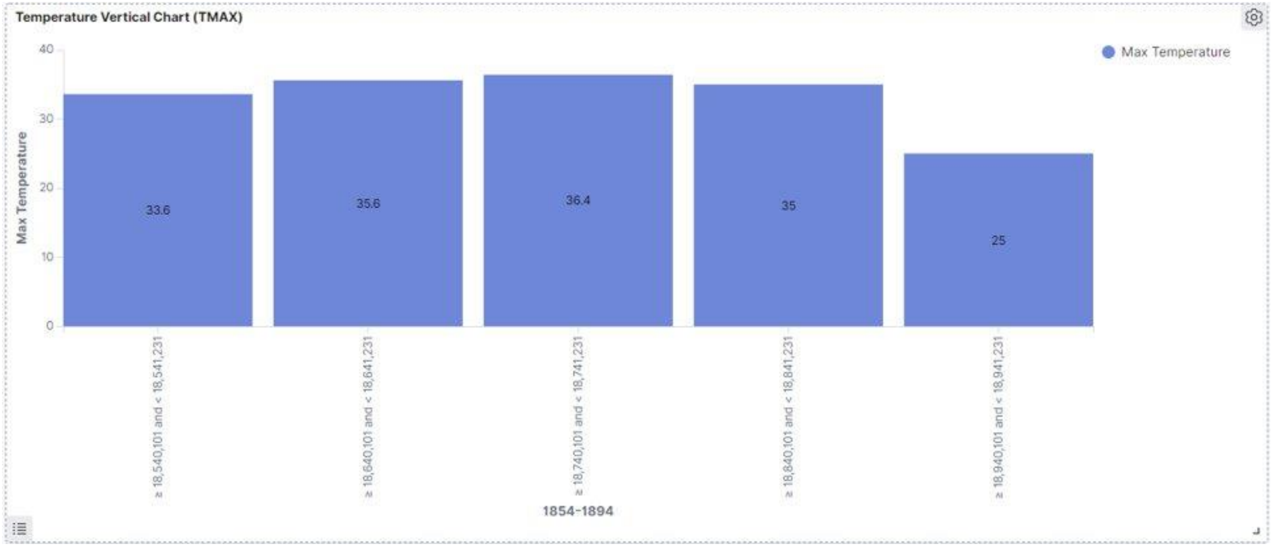
1864: 35.6°C (96.08°F)

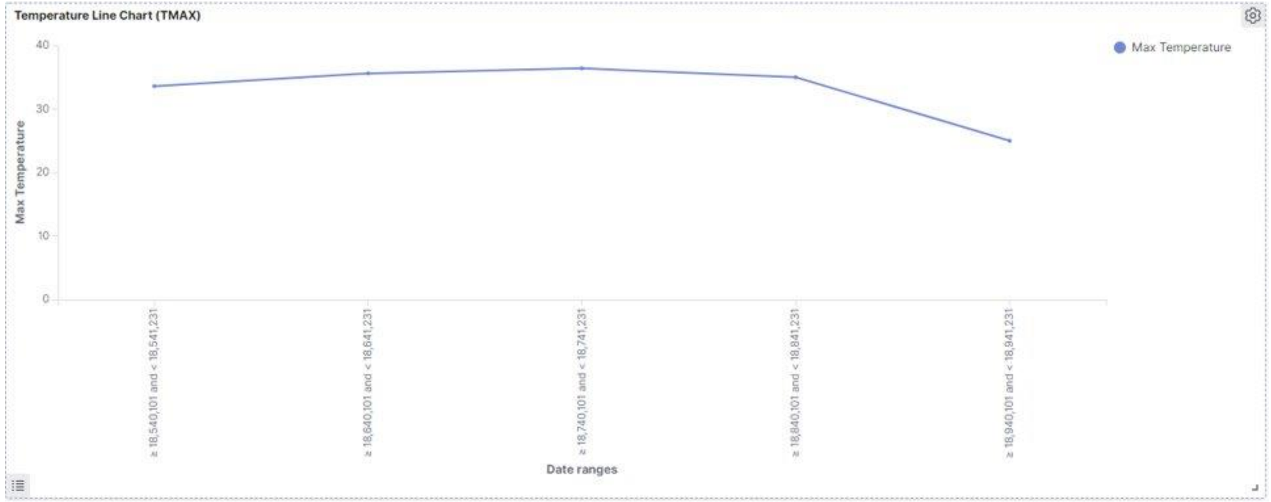
1874: 36.4°C (97.52°F)

1884: 35°C (95°F)

1894: 25°C (77°F)

The Maximum Temperatures at first seemed to be stable until 1894 where it suddenly dropped.





Min Spring Temp

1854: -12.1°C (10.22°F)

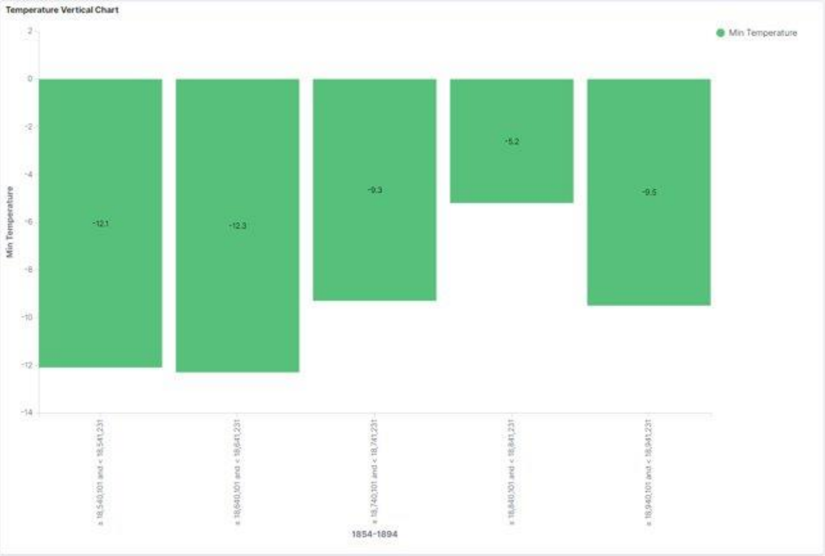
1864: -12.3°C (9.86°F)

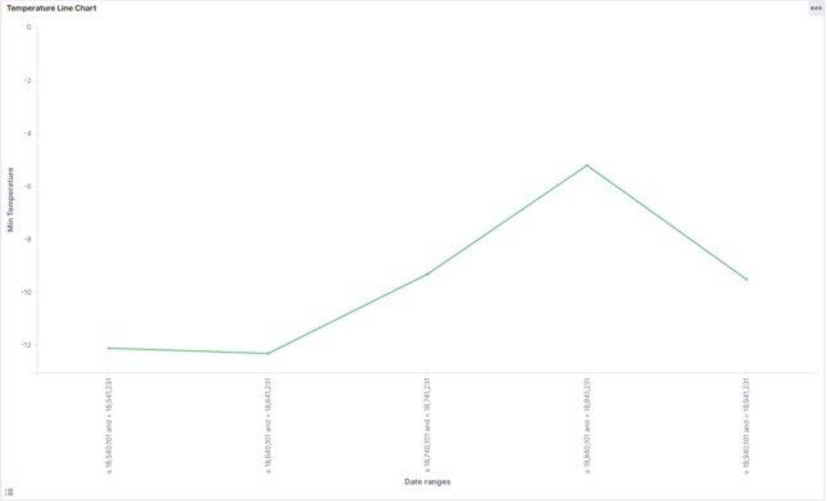
1874: -9.3°C (15.26°F)

1884: -5.2°C (22.64°F)

1894: -9.5°C (14.9°F)

An almost opposite event happens where a sudden spike appears. However, instead of it occuring in 1894, it occurs in 1874 and 1884. It then starts to lower again.





Average Spring Temps

1854: 12.513°c (54.523°F)

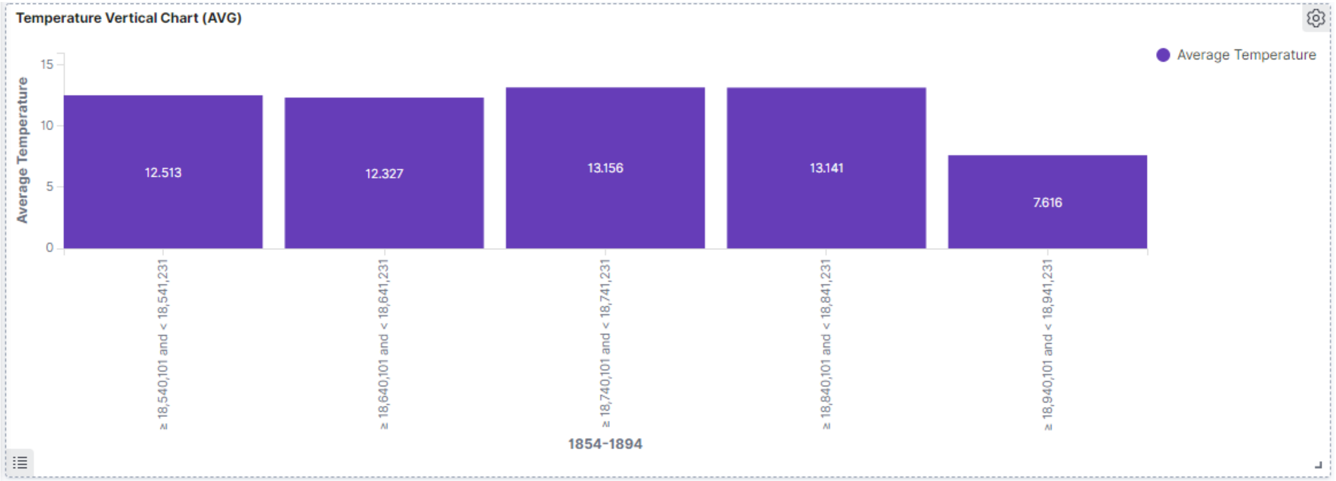
1864: 12.327°c (54.188°F)

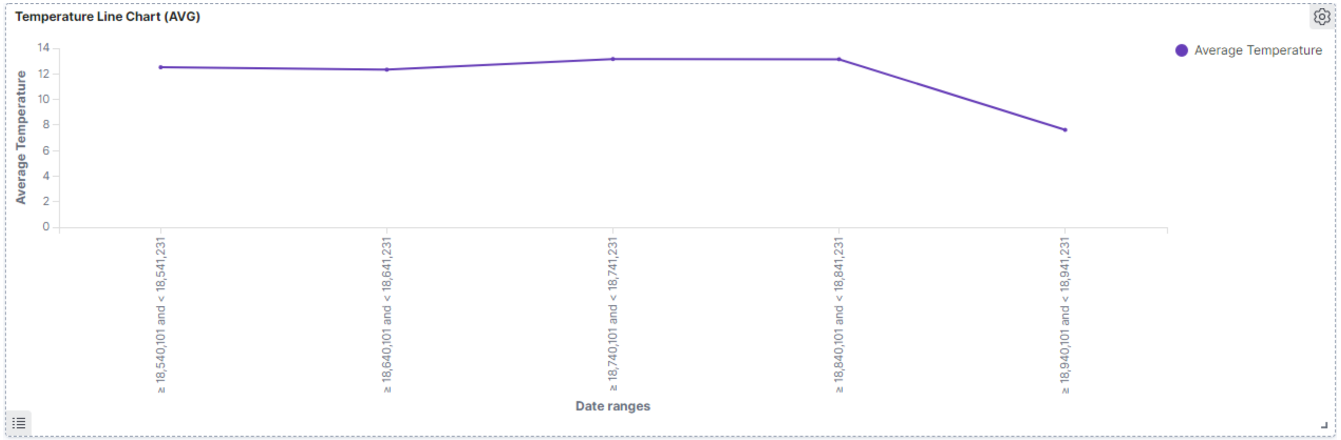
1874: 13.156°c (55.680°F)

1884: 13.141°c (55.653°F)

1894: 7.616°c (45.708°F)

There is a simillar downwards trend when comparing the Average Spring Temps and Max Spring Temps


Temperature Vertical Chart (AVG)


Temperature Line Chart (AVG)

Total Count for Milan

1854: 728 recorded temperatures

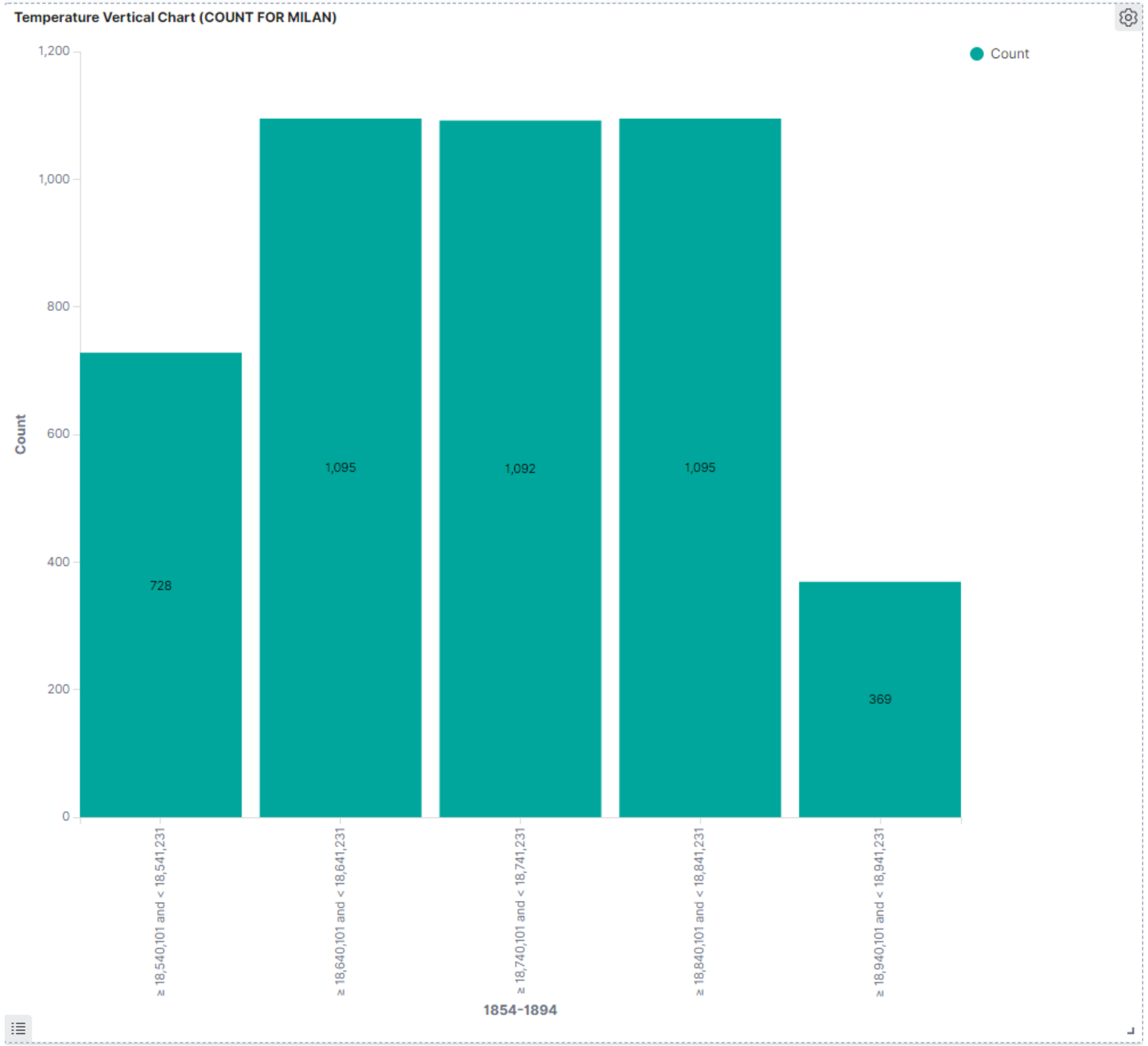
1864: 1095 recorded temperatures

1874: 1092 recorded temperatures

1884: 1095 recorded temperatures

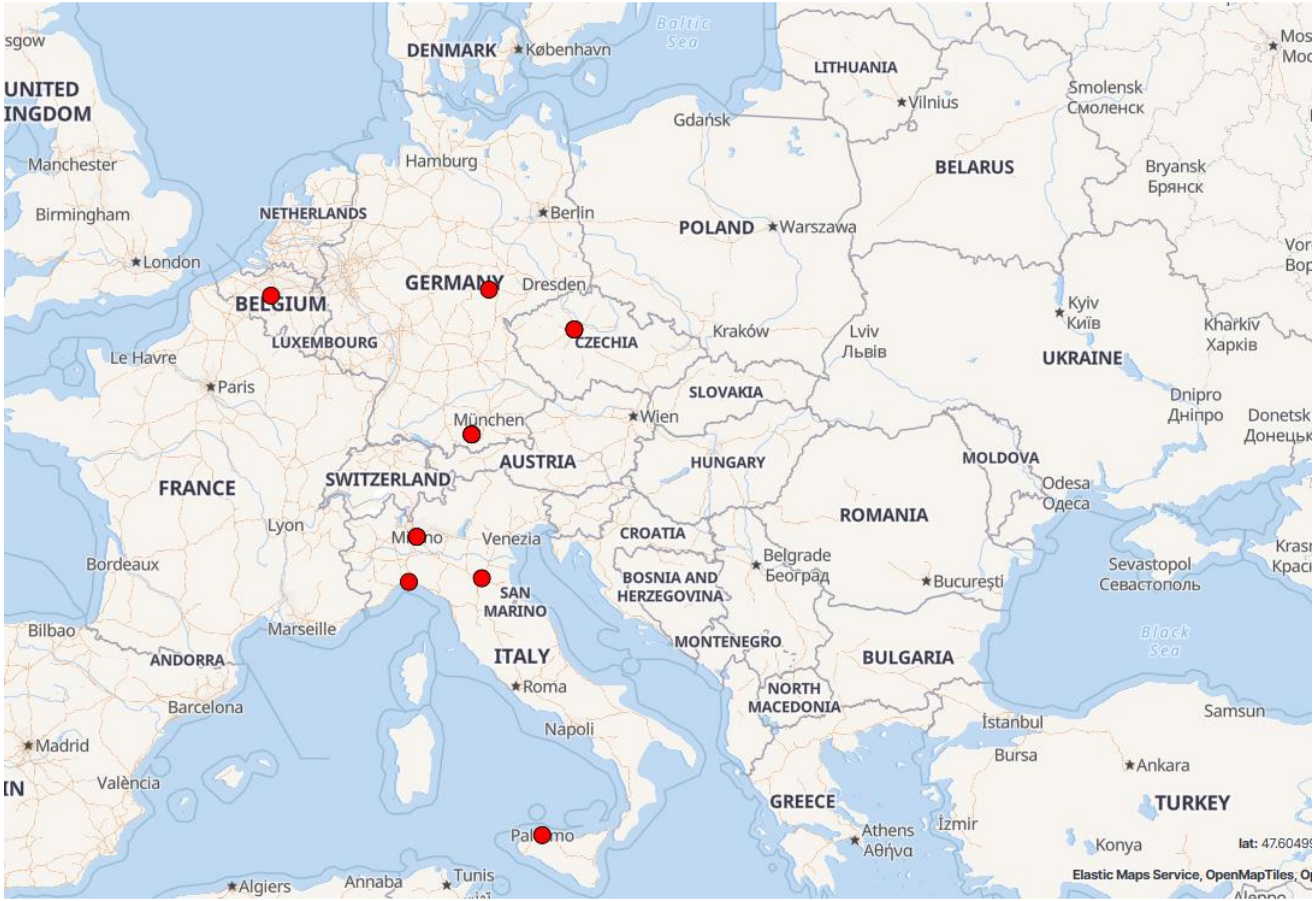
1894: 369 recorded temperatures

The most recorded temperature data was between 1864 and 1884.



Geomapping

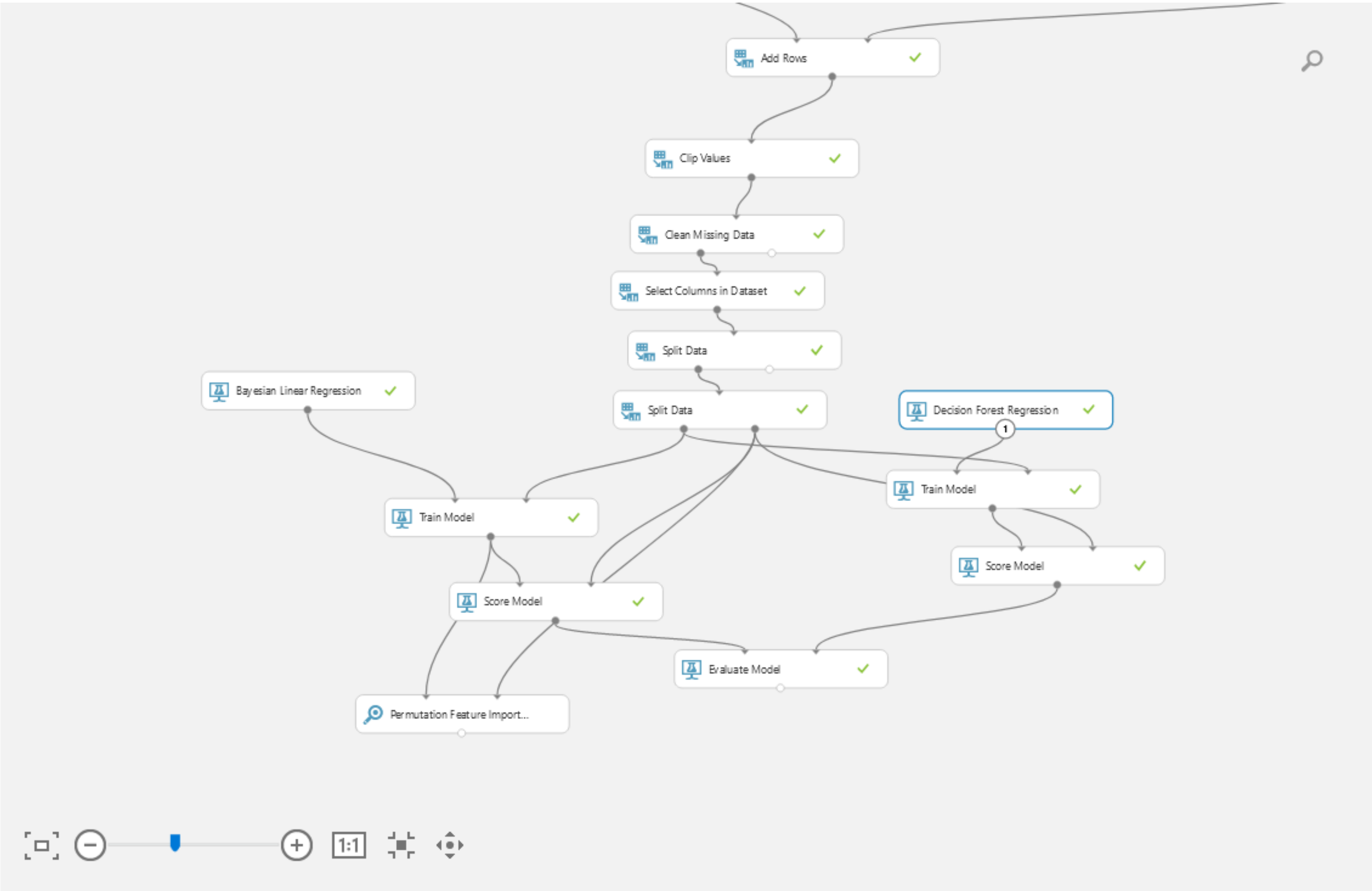
- Mainly focused on Milan, Italy
- As years were added on, more locations began recording data
- Up until 1844 data was mainly collected in Europe



Modeling

We used Bayesian Linear Regression and Decision Forest Regression to model the data in the temperature column.

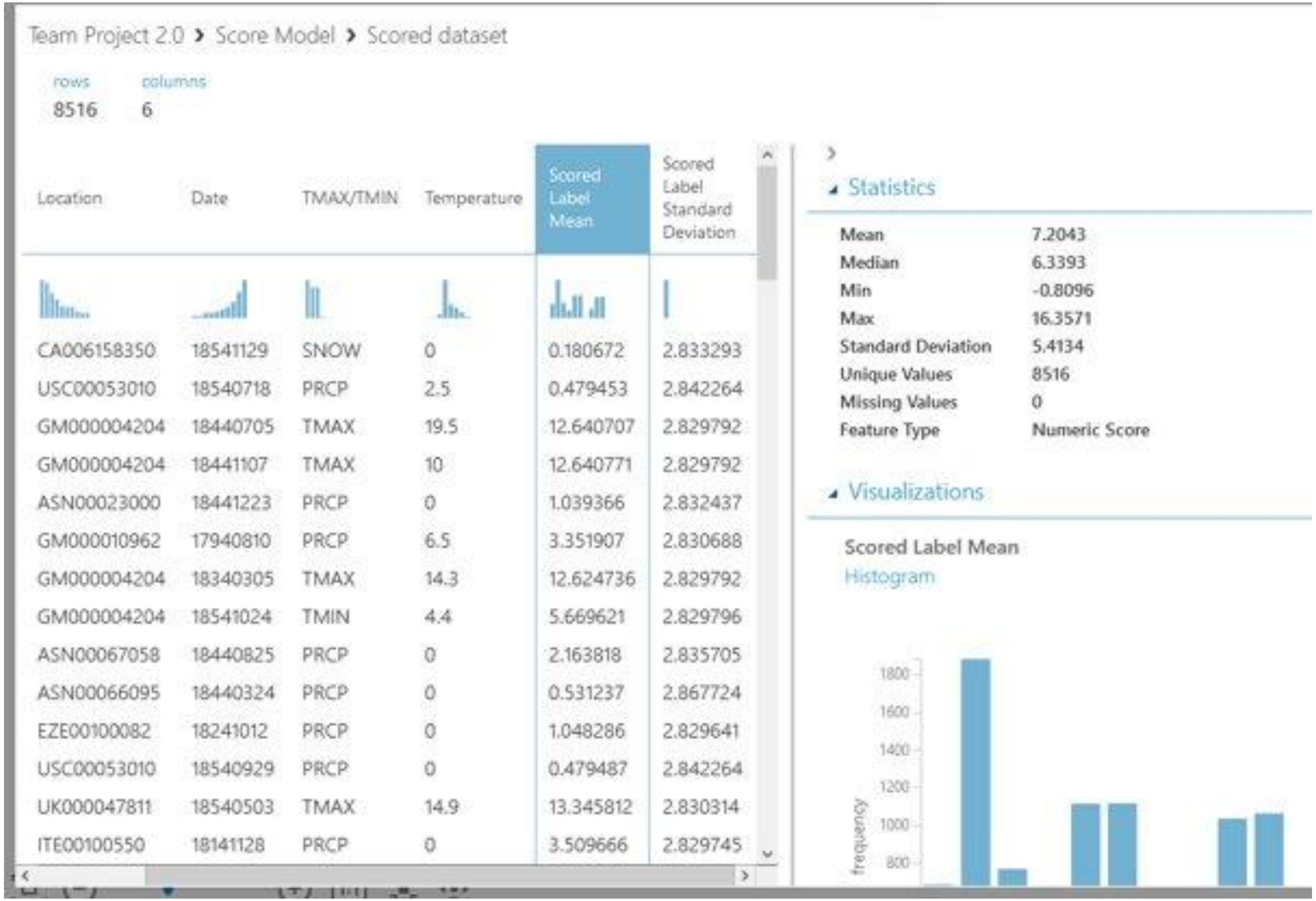
We ran the two models separately and then compared the results with the evaluate model function to see which model worked best.



Modeling -Bayesian Linear Regression

Compare the Scored Label Mean column to the Temperature column. The former represents the predicted values for the temperature column.

Predictions are somewhat close, but some are completely off.



Team Project 2.0 › Score Model › Scored dataset

rows 8516 columns 6

Location	Date	TMAX/TMIN	Temperature	Scored Label Mean	Scored Label Standard Deviation
CA006158350	18541129	SNOW	0	0.180672	2.833293
USC00053010	18540718	PRCP	2.5	0.479453	2.842264
GM000004204	18440705	TMAX	19.5	12.640707	2.829792
GM000004204	18441107	TMAX	10	12.640771	2.829792
ASN00023000	18441223	PRCP	0	1.039366	2.832437
GM000010962	17940810	PRCP	6.5	3.351907	2.830688
GM000004204	18340305	TMAX	14.3	12.624736	2.829792
GM000004204	18541024	TMIN	4.4	5.669621	2.829796
ASN00067058	18440825	PRCP	0	2.163818	2.835705
ASN00066095	18440324	PRCP	0	0.531237	2.867724
EZE00100082	18241012	PRCP	0	1.048286	2.829641
USC00053010	18540929	PRCP	0	0.479487	2.842264
UK000047811	18540503	TMAX	14.9	13.345812	2.830314
ITE00100550	18141128	PRCP	0	3.509666	2.829745

Statistics

Mean	7.2043
Median	6.3393
Min	-0.8096
Max	16.3571
Standard Deviation	5.4134
Unique Values	8516
Missing Values	0
Feature Type	Numeric Score

Visualizations

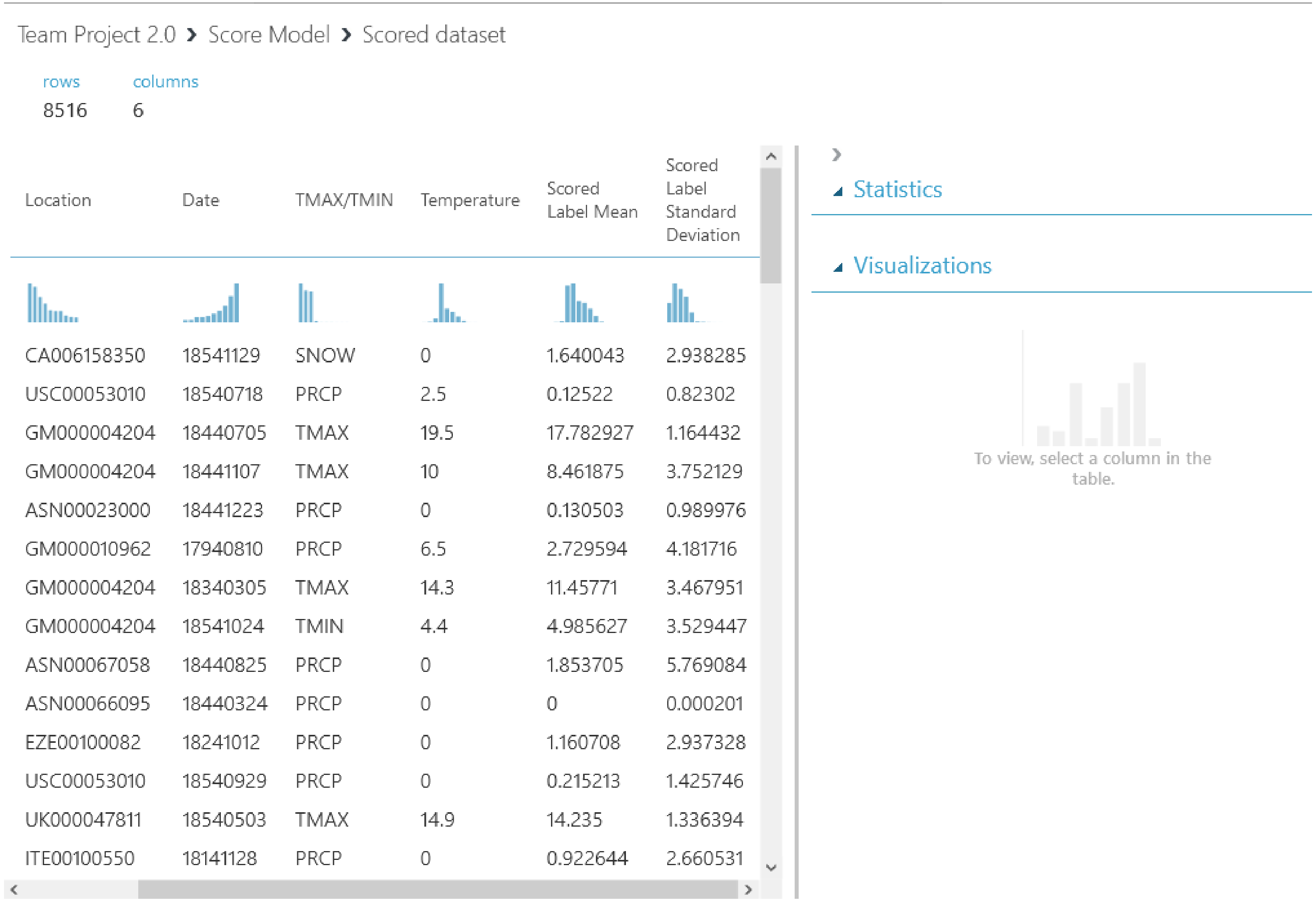
Scored Label Mean

Histogram

Modeling –Decision Forest Regression

Compare the two columns together again.

Predictions are somewhat close.



Team Project 2.0 › Score Model › Scored dataset

rows 8516

columns 6

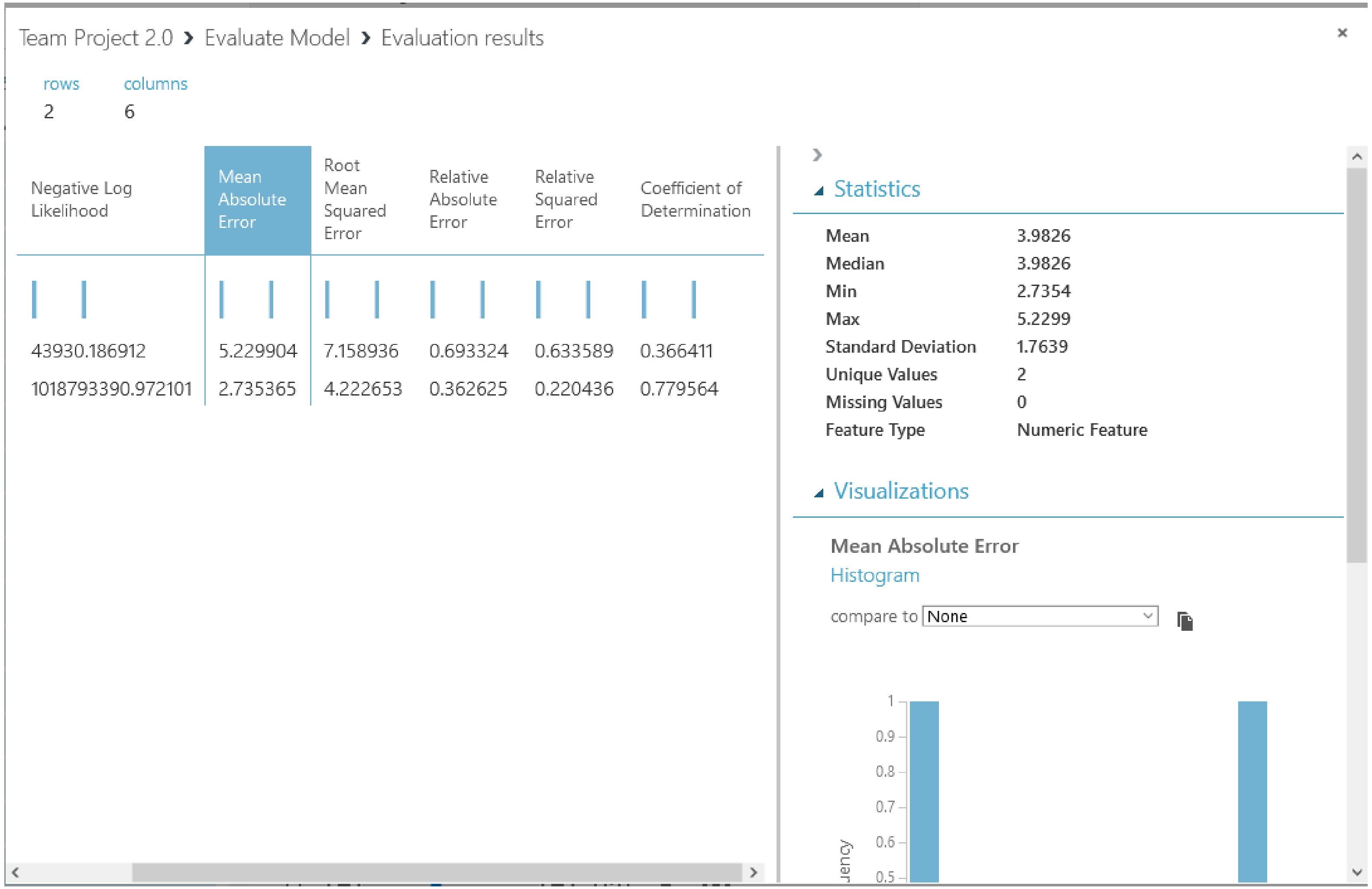
Location	Date	TMAX/TMIN	Temperature	Scored Label Mean	Scored Label Standard Deviation
CA006158350	18541129	SNOW	0	1.640043	2.938285
USC00053010	18540718	PRCP	2.5	0.12522	0.82302
GM000004204	18440705	TMAX	19.5	17.782927	1.164432
GM000004204	18441107	TMAX	10	8.461875	3.752129
ASN00023000	18441223	PRCP	0	0.130503	0.989976
GM000010962	17940810	PRCP	6.5	2.729594	4.181716
GM000004204	18340305	TMAX	14.3	11.45771	3.467951
GM000004204	18541024	TMIN	4.4	4.985627	3.529447
ASN00067058	18440825	PRCP	0	1.853705	5.769084
ASN00066095	18440324	PRCP	0	0	0.000201
EZE00100082	18241012	PRCP	0	1.160708	2.937328
USC00053010	18540929	PRCP	0	0.215213	1.425746
UK000047811	18540503	TMAX	14.9	14.235	1.336394
ITE00100550	18141128	PRCP	0	0.922644	2.660531

To view, select a column in the table.

Modeling – Conclusion

The values are measurements of the accuracy of the models when comparing the label values that it predicts to the known values in the test dataset.

The Error and the Coefficient shows that the Decision Forest Model works the best



Team Project 2.0 › Evaluate Model › Evaluation results

rows	columns
2	6

Negative Log Likelihood	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
43930.186912	5.229904	7.158936	0.693324	0.633589	0.366411
1018793390.972101	2.735365	4.222653	0.362625	0.220436	0.779564

Statistics

Mean	3.9826
Median	3.9826
Min	2.7354
Max	5.2299
Standard Deviation	1.7639
Unique Values	2
Missing Values	0
Feature Type	Numeric Feature

Visualizations

Mean Absolute Error

Histogram

compare to None

The model attached to the left port is presented first (Bayesian), followed by the metrics for the model attached on the right port (Forest)

Questions?