Changelog - July 5th, 2023

Dane:

- Visualization
  - Ergot
  - Station distribution
  - Station data (hly / dly)
  - Soil moisture
  - Soil
- Updated importErgot script (rejects if incomplete)
- Merged districts 4840 and 4841/persisted changes
- Soil data (condensed and added to readme)
- Aggregated soil, ergot and weather data
- Feature engineering
- Started research/modeling

Daniel:

- Peer review
- Ergot visualization
- Ergot statistics
- Research on models
- Continue implementing autoencoder and ML models
- Cross validation techniques

Dharmit:

- Ergot Visualisation
- Soil Data Aggregation
- Satellite Soil Moisture Data
- Started research on models

Jay:

- Process soil moisture data
- Aggregate soil moisture data
- Soil moisture Visualization
- Soil moisture statistics
- Did some research and build a basic MLP model
- Tried different variation to merge different datasets

Joseff:

- Copernicus satellite data retrieval / aggregation
- Copernicus skip existing data
- Data Synchronization between databases
- Peer review
- Git/Python mentoring
- Data aggregation strategy
- Repo Linter Actions
  - Code format consistency - Black
  - Code type consistency - Mypy
  - ~~Code PEP8 conformance - PyLint~~

Aggregation strategy

Primary aggregation attributes

- Year
- District

Ergot:

- Future predictors
- Districts:
    - MB = crop_district + 4600
    - SK = crop_district + 4700 - 1
    - AB = (crop_district)(10) + 4800

Soil:

- Very granular and many different soils per polygon
- Using weighted averages based on the percentage they occupy within their polygon
    - Districts have multiple polygons
- Non numerical values become booleans, currently ignoring them

Weather: 365 days x features

Soil Moisture:

- Replaced null values
- Using year and district to aggregate and generated separate column for min, max and mean

Removed:

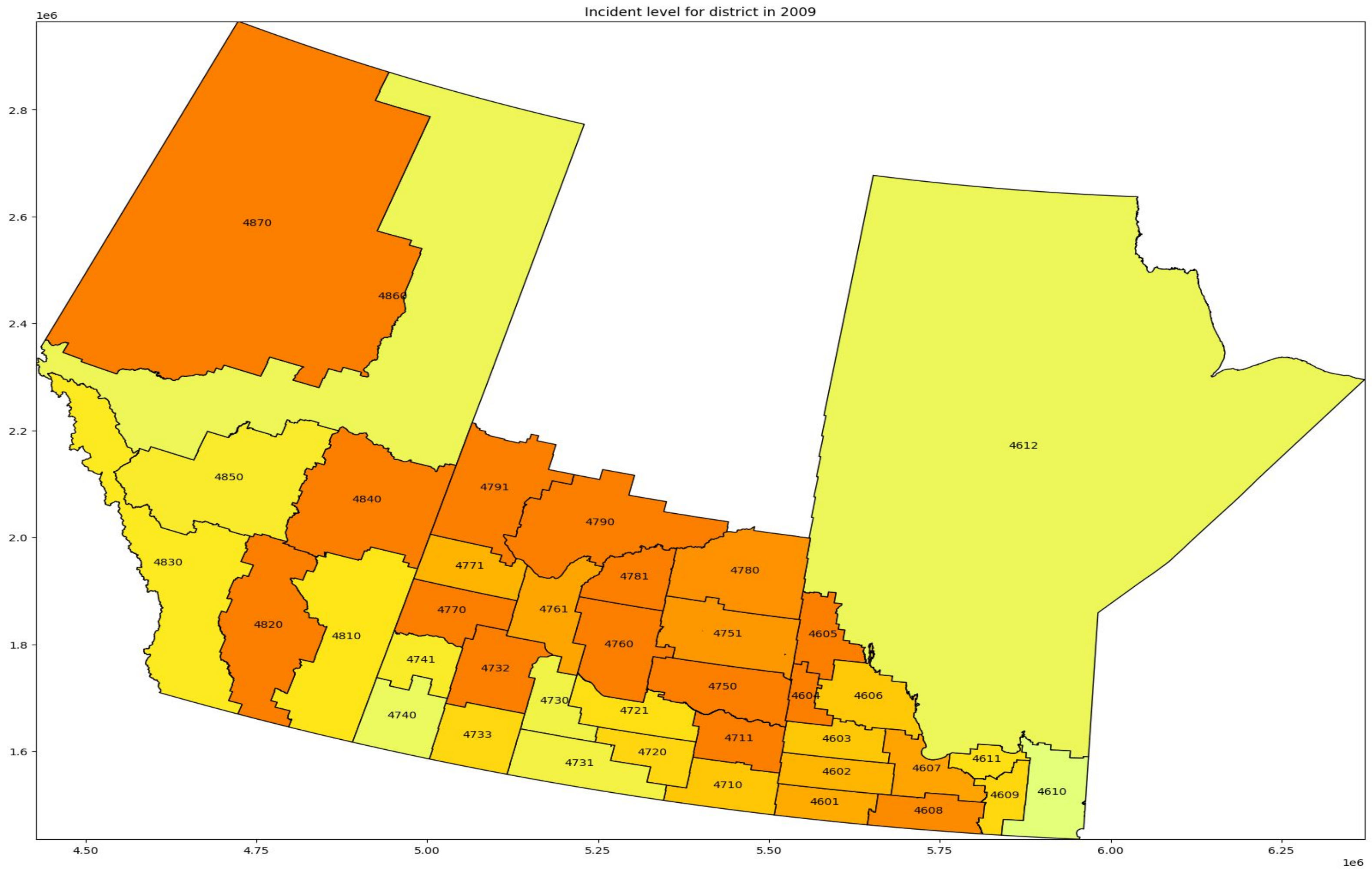
Water table (groundwater)
Root restriction types
Drainage
Parent material textures
Parent material chemicals
Mode of depositions
Sand texture

Visualization - Ergot

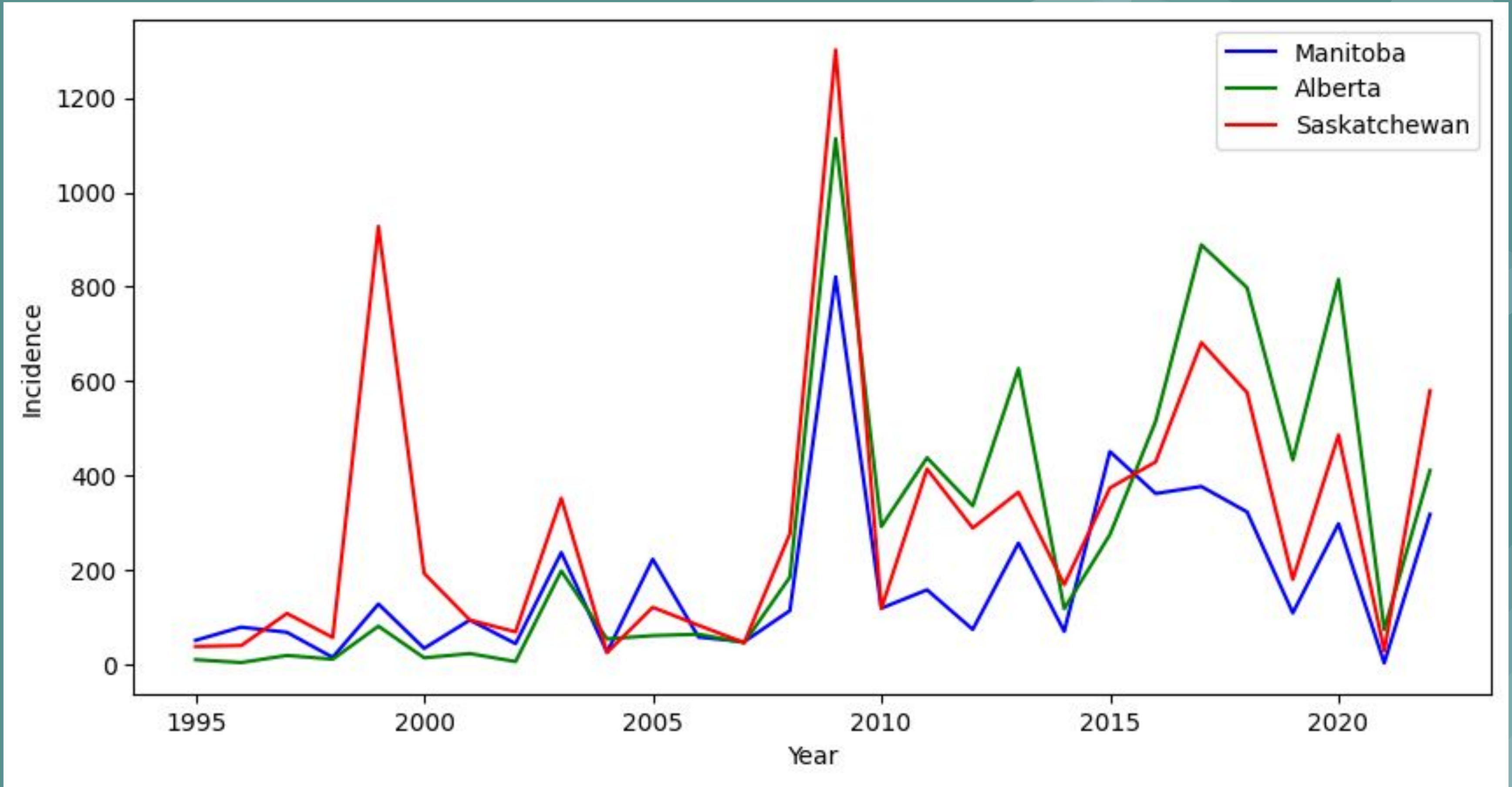
1. What kind of statistics we have done?
   a. Correlation
   b. Incidence ratio

2. What visualization we have created?
   a. Correlation plots
   b. Pair plots
   c. Region plot (show the incidence level)
   d. Line plot (to show the number of ergot over the year)

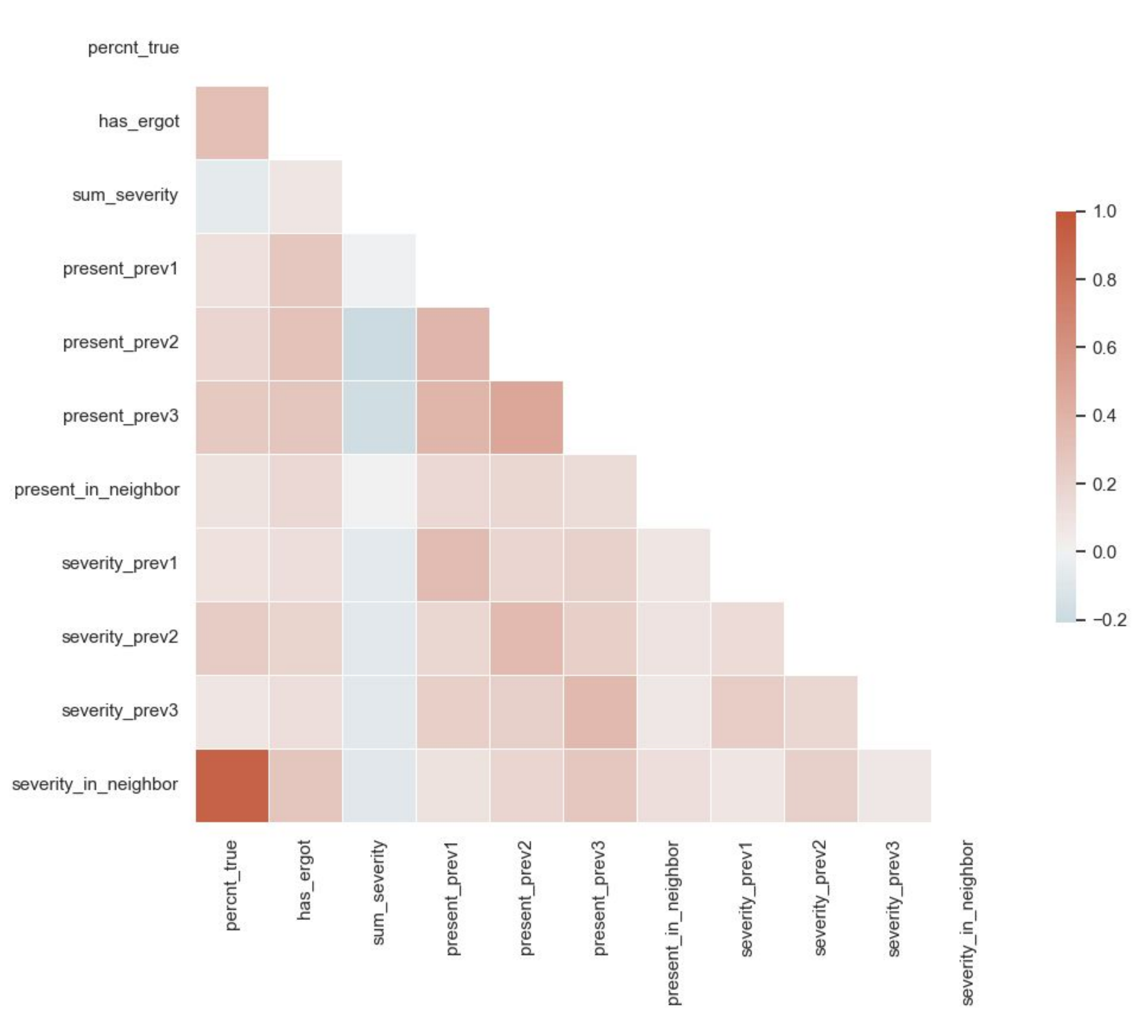
3. From the visualization, what we have learned about the dataset?
   - 2009 was when ergot happened significantly.
   - We have 139 outliers out of 1092 data points with q1, q2 (median), q3 are 0.02, 0.13, 0.55, respectively
   - Percent of having ergot when prev year had ergot: 0.6813186813186813
   - Percent of having ergot when prev 2 year had ergot: 0.673992673992674
   - Percent of having ergot when prev 3 year had ergot: 0.6446886446886447
   - Percent of having ergot when neighbor is having ergot: 0.7976190476190477



Incident level for district in 2009



Manitoba
Alberta
Saskatchewan
Incidence
0
200
400
600
800
1000
1200
Year
1995
2000
2005
2010
2015
2020

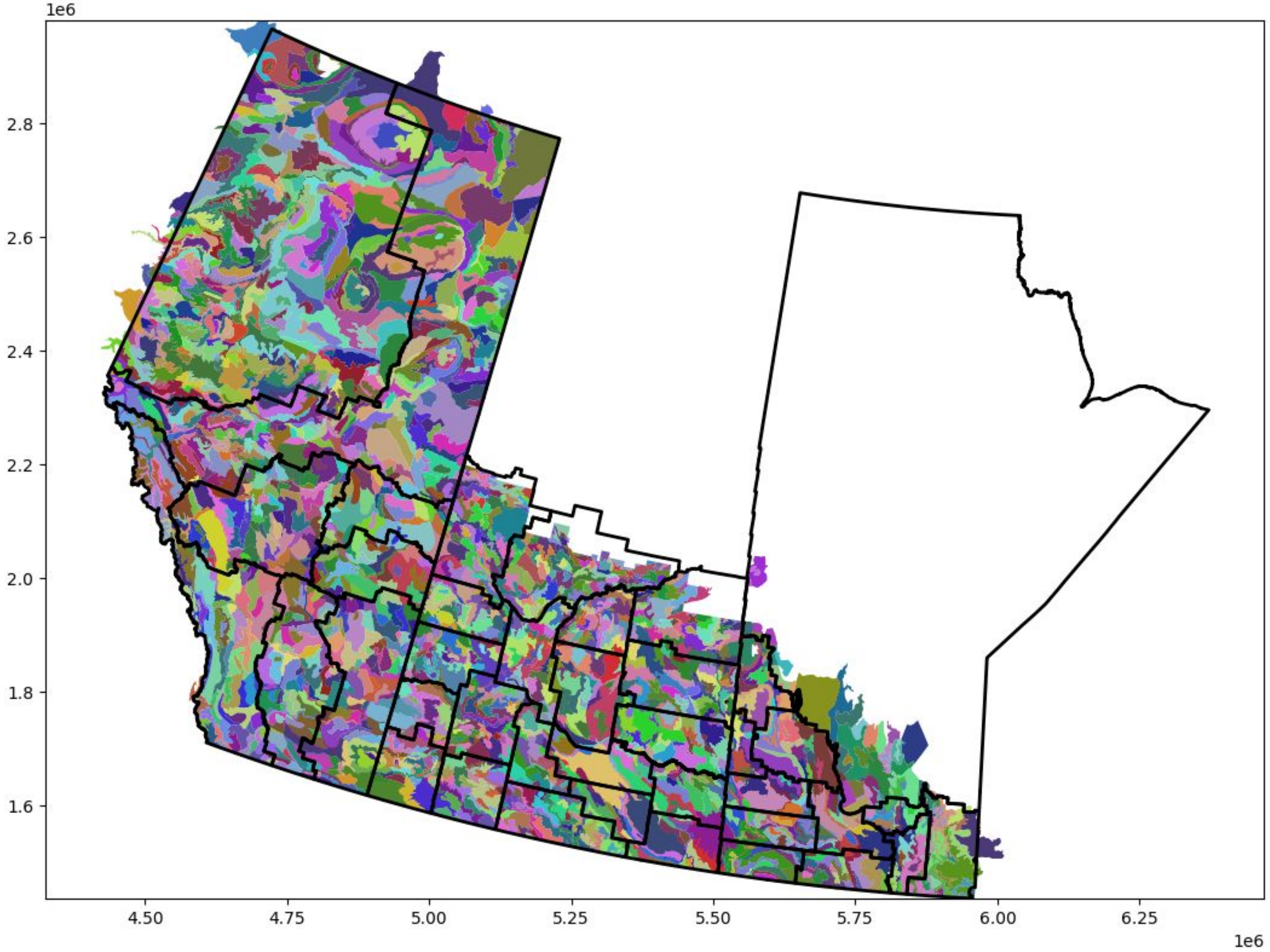

percnt_true
has_ergot
sum_severity
present_prev1
present_prev2
present_prev3
present_in_neighbor
severity_prev1
severity_prev2
severity_prev3
severity_in_neighbor
percnt_true
has_ergot
sum_severity
present_prev1
present_prev2
present_prev3
present_in_neighbor
severity_prev1
severity_prev2
severity_prev3
severity_in_neighbor
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

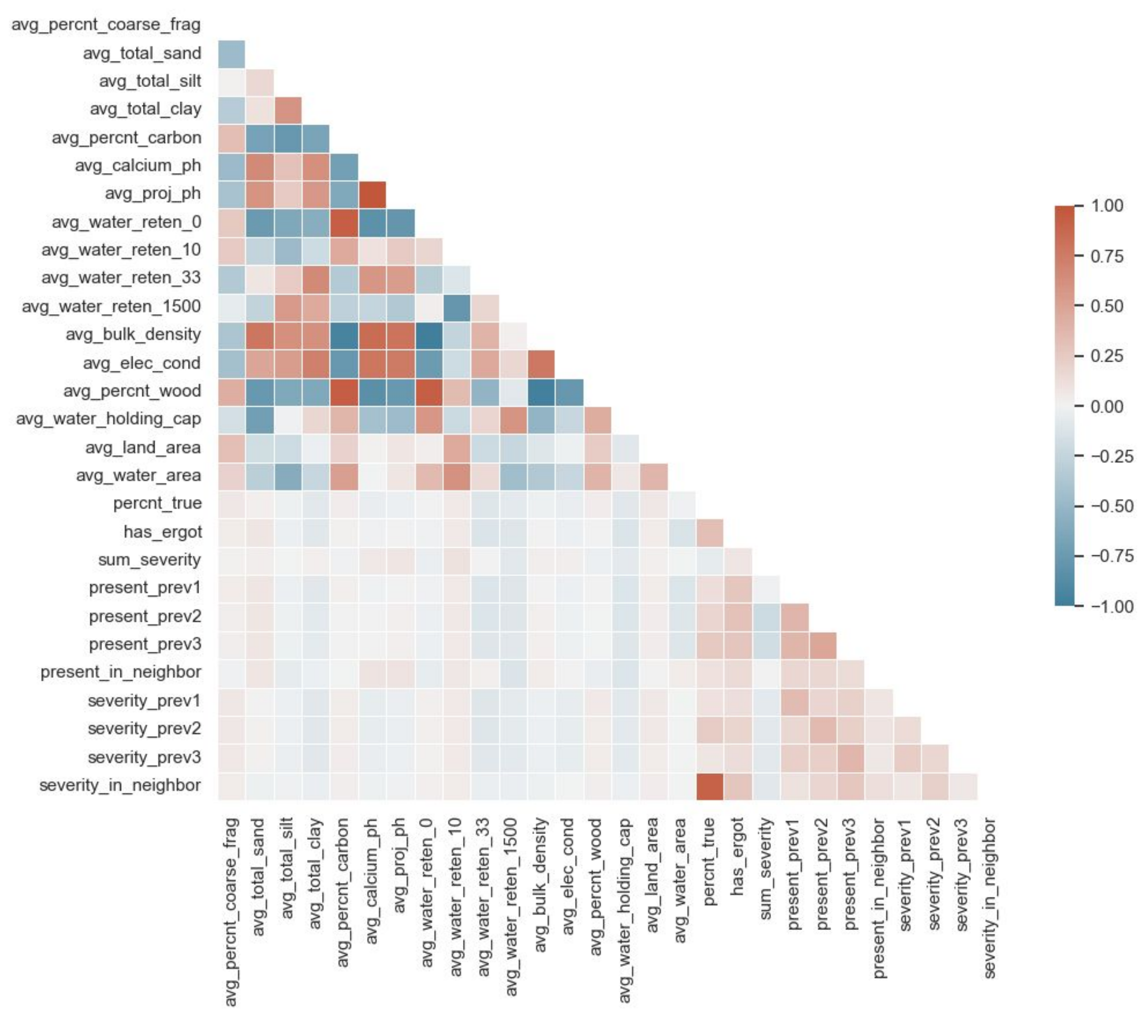
Visualization - Soil Data

1. What kind of statistics we have done?
   a. Correlation to ergot

2. What visualization we have created?
   a. Correlation plot
   b. Pair plots
   c. Region plot of soils across Canada (as many as 23 different soil types represented by a color in some cases)

3. From the visualization, what we have learned about the dataset?
   a. Complex relationships and complex data
      i. Exploring data (processing/features) and more complicated models will be important
      ii. Important to consider alternative aggregation strategies (possibly through combining factors together and feature engineering)


1e6
2.8
2.6
2.4
2.2
2.0
1.8
1.6
4.50
4.75
5.00
5.25
5.50
5.75
6.00
6.25
1e6

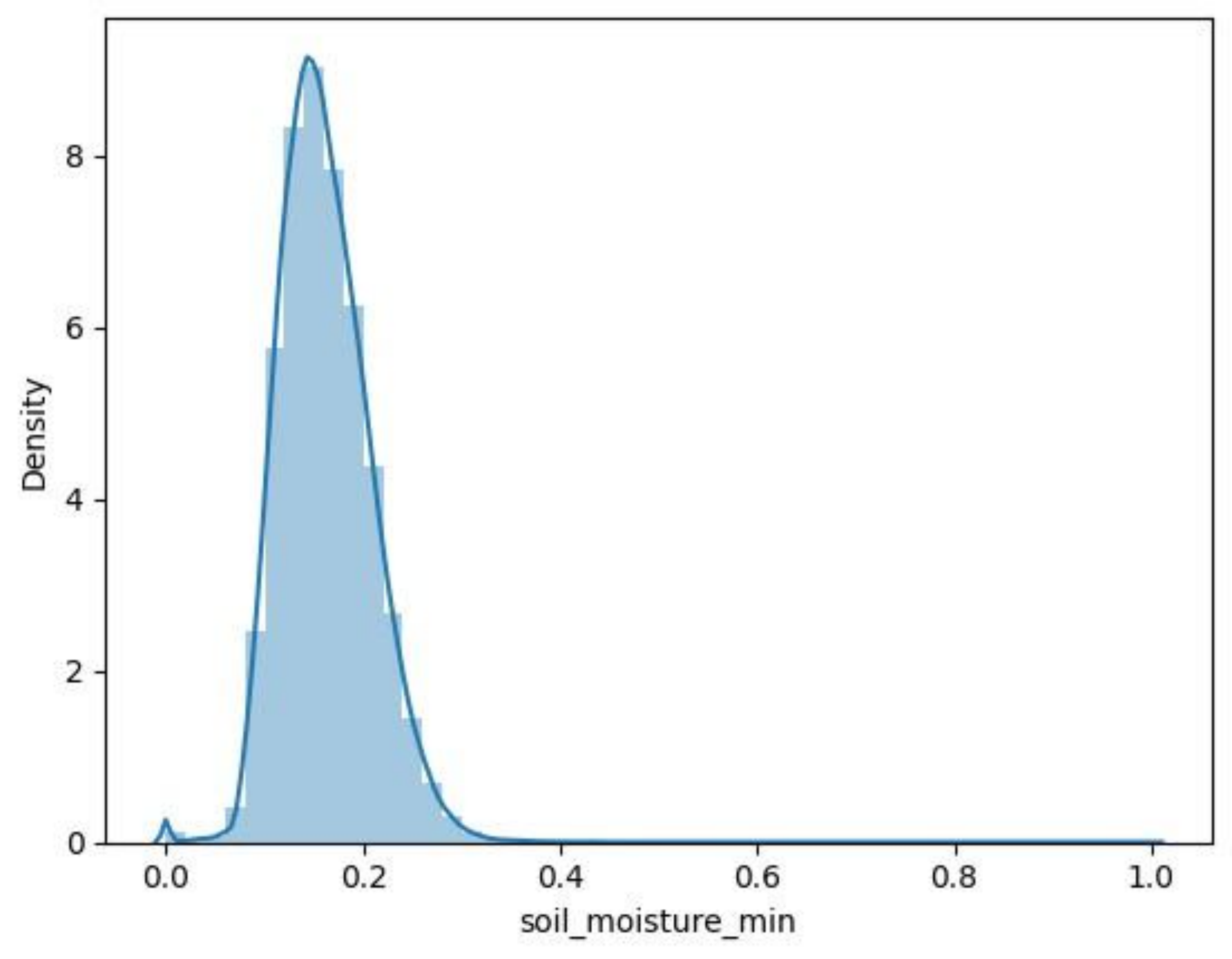

avg_percnt_coarse_frag
avg_total_sand
avg_total_silt
avg_total_clay
avg_percnt_carbon
avg_calcium_ph
avg_proj_ph
avg_water_reten_0
avg_water_reten_10
avg_water_reten_33
avg_water_reten_1500
avg_bulk_density
avg_elec_cond
avg_percnt_wood
avg_water_holding_cap
avg_land_area
avg_water_area
percnt_true
has_ergot
sum_severity
present_prev1
present_prev2
present_prev3
present_in_neighbor
severity_prev1
severity_prev2
severity_prev3
severity_in_neighbor
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

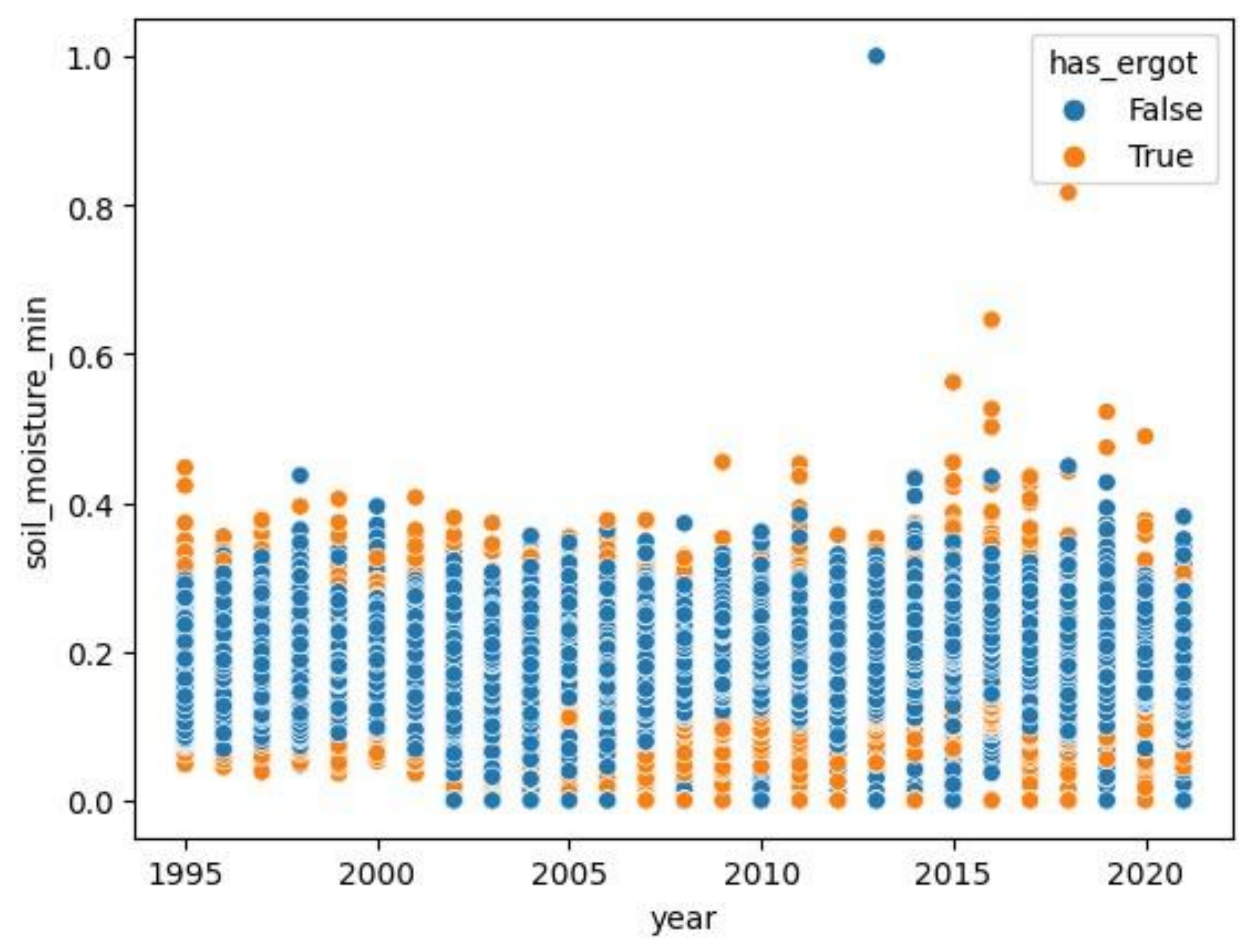
Visualization - Soil Moisture

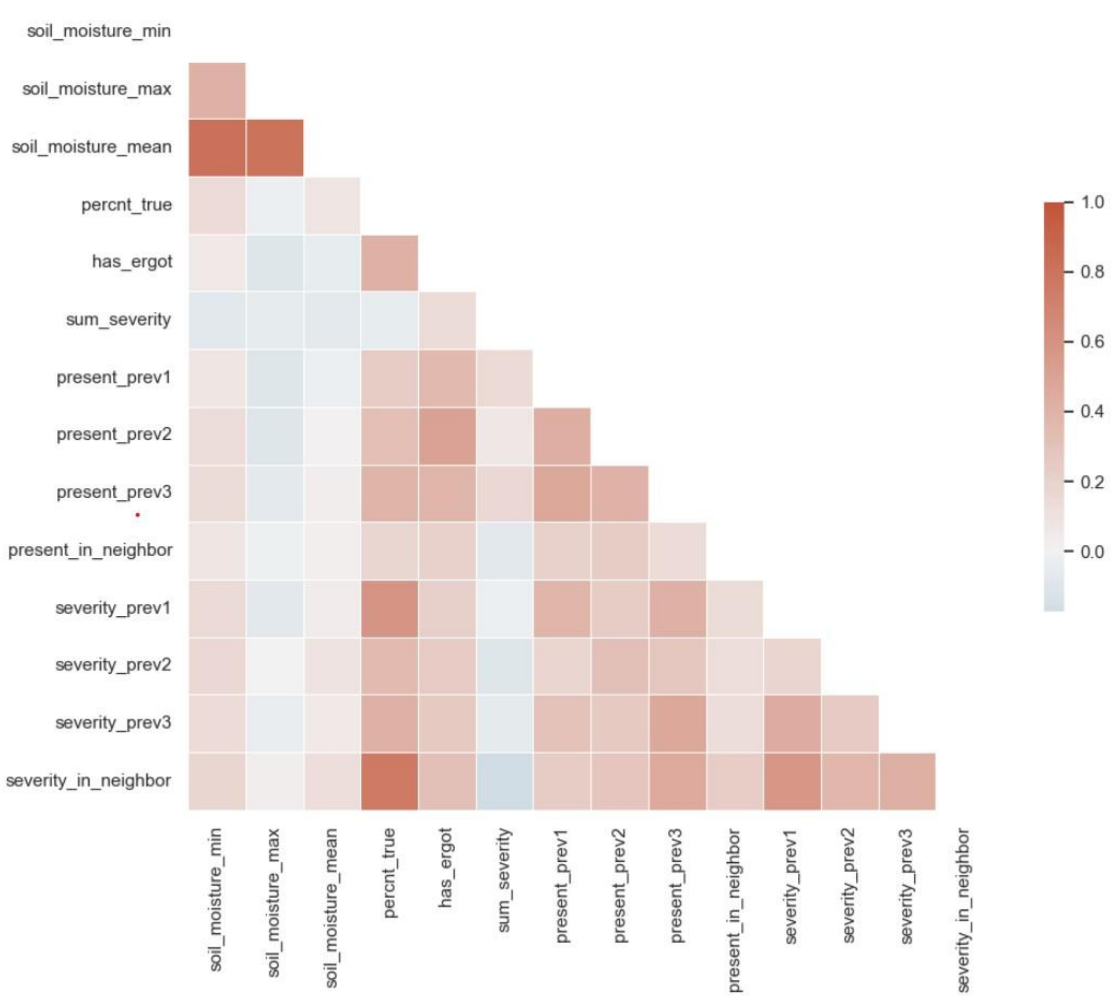
1. What kind of statistics we have done?
    a. Correlation to ergot
    b. Data distributions - min, mean, max

2. What visualization we have created?
    a. Correlation plot
    b. Distribution plot (histogram + kde)
    c. Scatter plot

3. From the visualization, what we have learned about the dataset?
    a. The values of soil moisture is skewed
    b. Complex relationships
        i. Exploring data (processing/features) and more complicated models will be important
        ii. There is some kind of reaction between soil moisture and ergot


Density
8
6
4
2
0
0.0
0.2
0.4
0.6
0.8
1.0
soil_moisture_min


has_ergot
● False
● True
soil_moisture_min
1.0
0.8
0.6
0.4
0.2
0.0
1995
2000
2005
2010
2015
2020
year

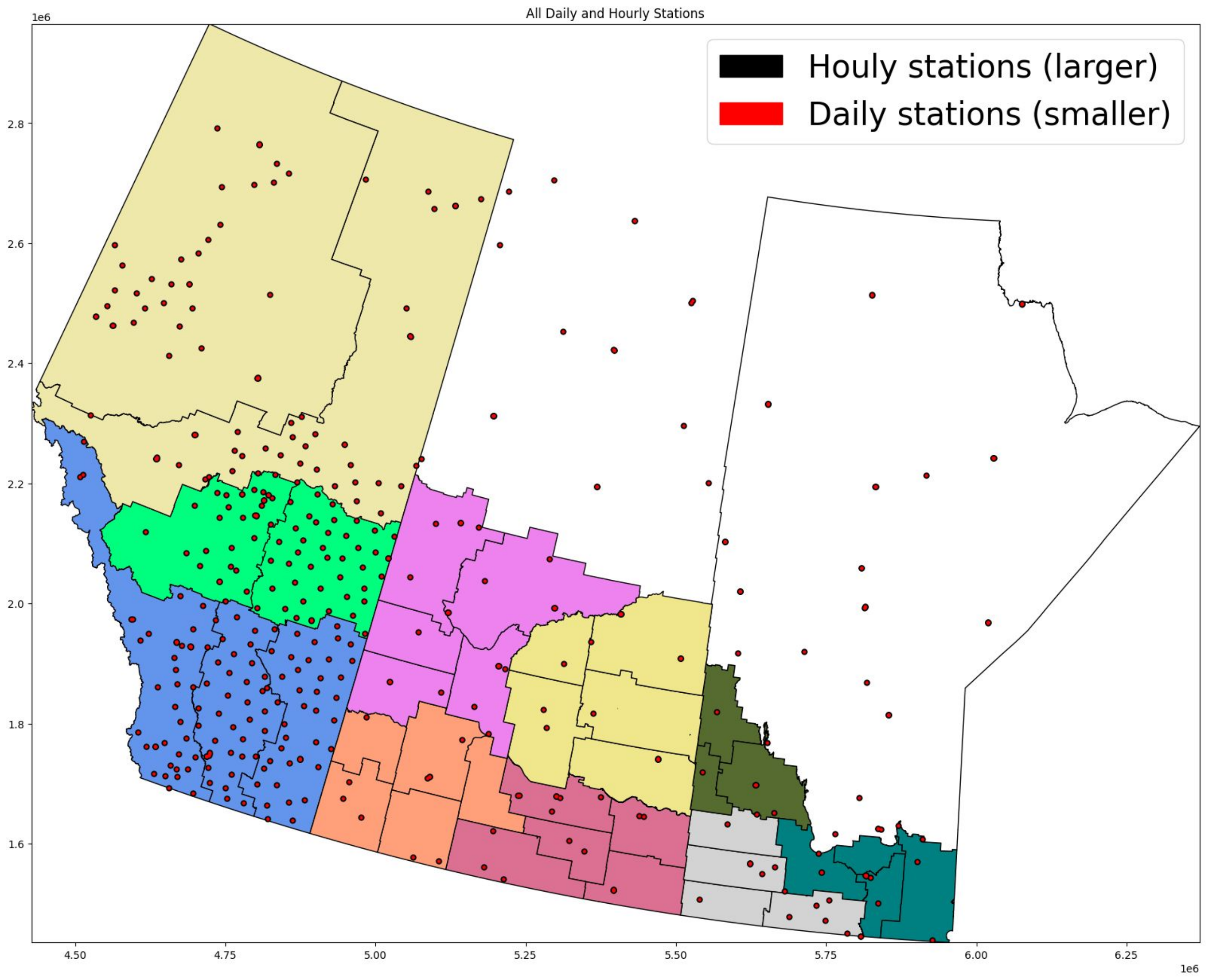

soil_moisture_min
soil_moisture_max
soil_moisture_mean
percnt_true
has_ergot
sum_severity
present_prev1
present_prev2
present_prev3
present_in_neighbor
severity_prev1
severity_prev2
severity_prev3
severity_in_neighbor
soil_moisture_min
soil_moisture_max
soil_moisture_mean
percnt_true
has_ergot
sum_severity
present_prev1
present_prev2
present_prev3
present_in_neighbor
severity_prev1
severity_prev2
severity_prev3
severity_in_neighbor
1.0
0.8
0.6
0.4
0.2
0.0

Visualization - Weather Station

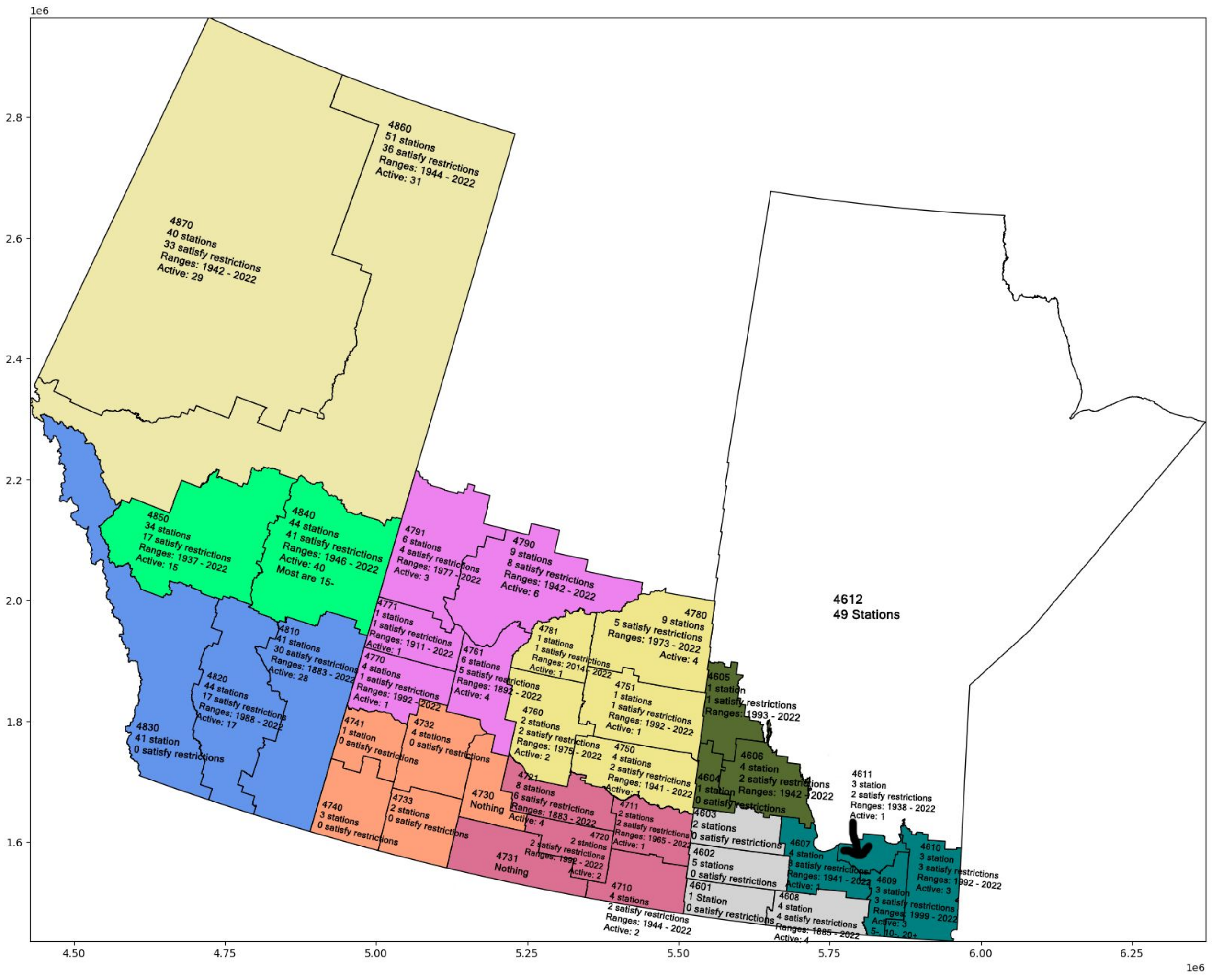
1. What kind of statistics we have done?
   a. Station elevation
   b. Which stations are still active?
   c. Which stations are hourly and which are daily?
   d. Amount of data collected

2. What visualization we have created?
   a. Station summaries for each district
   b. Region plots for stations

3. From the visualization, what we have learned about the dataset?
   a. Usually multiple stations located at the same coordinates
   b. All locations collect both hourly and daily data



All Daily and Hourly Stations

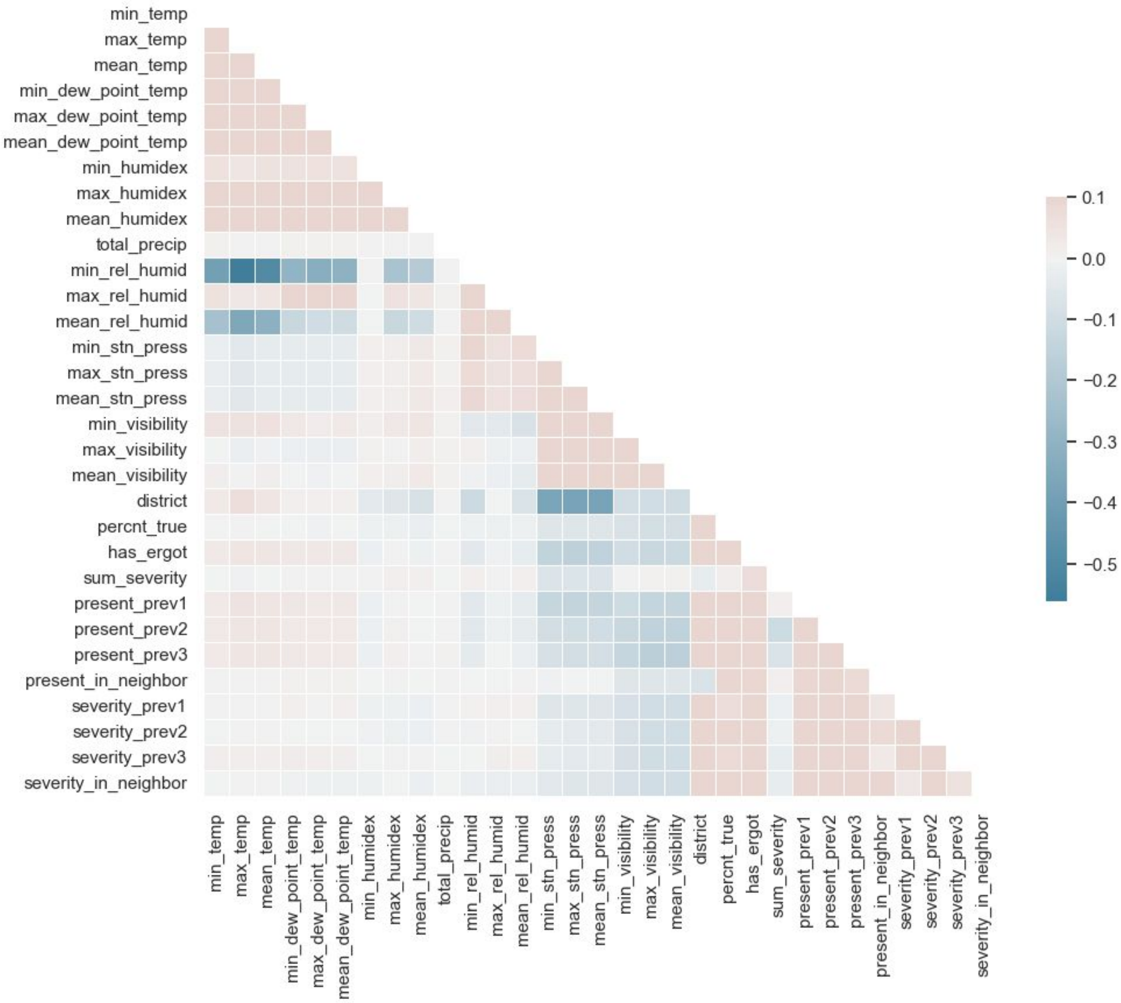

1e6
2.8
2.6
2.4
2.2
2.0
1.8
1.6
4.50
4.75
5.00
5.25
5.50
5.75
6.00
6.25
1e6
4860
51 stations
36 satisfy restrictions
Ranges: 1944 - 2022
Active: 31
4870
40 stations
33 satisfy restrictions
Ranges: 1942 - 2022
Active: 29
4850
34 stations
17 satisfy restrictions
Ranges: 1937 - 2022
Active: 15
4840
44 stations
41 satisfy restrictions
Ranges: 1946 - 2022
Active: 40
Most are 15-
4810
41 stations
30 satisfy restrictions
Ranges: 1883 - 2022
Active: 28
4820
44 stations
17 satisfy restrictions
Ranges: 1988 - 2022
Active: 17
4830
41 station
0 satisfy restrictions
4791
6 stations
4 satisfy restrictions
Ranges: 1977 - 2022
Active: 3
4790
9 stations
8 satisfy restrictions
Ranges: 1942 - 2022
Active: 6
4771
1 stations
1 satisfy restrictions
Ranges: 1911 - 2022
Active: 1
4770
4 stations
1 satisfy restrictions
Ranges: 1992 - 2022
Active: 1
4761
6 stations
5 satisfy restrictions
Ranges: 1892 - 2022
Active: 4
4781
1 stations
1 satisfy restrictions
Ranges: 2014 - 2022
Active: 1
4780
9 stations
5 satisfy restrictions
Ranges: 1973 - 2022
Active: 4
4751
1 stations
1 satisfy restrictions
Ranges: 1992 - 2022
Active: 1
4760
2 stations
2 satisfy restrictions
Ranges: 1975 - 2022
Active: 2
4750
4 stations
2 satisfy restrictions
Ranges: 1941 - 2022
Active: 1
4741
1 station
0 satisfy restrictions
4732
4 stations
0 satisfy restrictions
4730
Nothing
4721
8 stations
6 satisfy restrictions
Ranges: 1883 - 2022
Active: 4
4711
2 stations
2 satisfy restrictions
Ranges: 1965 - 2022
Active: 1
4740
3 stations
0 satisfy restrictions
4733
2 stations
0 satisfy restrictions
4720
2 stations
2 satisfy restrictions
Ranges: 1992 - 2022
Active: 2
4731
Nothing
4710
4 stations
2 satisfy restrictions
Ranges: 1944 - 2022
Active: 2
4612
49 Stations
4605
1 station
1 satisfy restrictions
Ranges: 1993 - 2022
4606
4 station
2 satisfy restrictions
Ranges: 1942 - 2022
4604
1 station
0 satisfy restrictions
4611
3 station
2 satisfy restrictions
Ranges: 1938 - 2022
Active: 1
4603
2 stations
0 satisfy restrictions
4602
5 stations
0 satisfy restrictions
4601
1 Station
0 satisfy restrictions
4607
4 station
3 satisfy restrictions
Ranges: 1941 - 2022
Active: 1
4608
4 station
4 satisfy restrictions
Ranges: 1885 - 2022
Active: 4
4609
3 station
3 satisfy restrictions
Ranges: 1999 - 2022
Active: 3
5-, 10-, 20+
4610
3 station
3 satisfy restrictions
Ranges: 1992 - 2022
Active: 3

Visualization - Weather Station Data

1. What kind of statistics we have done?
    a. Correlation to ergot
    b. Data distributions - min, mean, max

2. What visualization we have created?
    a. Correlation plot
    b. Histograms
    c. Pair plots
    d. Box plots (subsetting caused issues)

3. From the visualization, what we have learned about the dataset?
    a. Most exciting correlation was between max_temperature and ergot
    b. Outliers
    c. Complex relationships
        i. Exploring data (processing/features) and more complicated models will be important


min_temp
max_temp
mean_temp
min_dew_point_temp
max_dew_point_temp
mean_dew_point_temp
min_humidex
max_humidex
mean_humidex
total_precip
min_rel_humid
max_rel_humid
mean_rel_humid
min_stn_press
max_stn_press
mean_stn_press
min_visibility
max_visibility
mean_visibility
district
percnt_true
has_ergot
sum_severity
present_prev1
present_prev2
present_prev3
present_in_neighbor
severity_prev1
severity_prev2
severity_prev3
severity_in_neighbor
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5

Engineered Features

Note that checks for previous years
are not accumulative

percnt_true	FLOAT
has_ergot	BOOL
median_severity	FLOAT
sum_severity	FLOAT
present_in_neighbor	BOOL
sum_severity_in_neighbor	FLOAT
present_prev1	BOOL
present_prev2	BOOL
present_prev3	BOOL
percnt_true_prev1	FLOAT
percnt_true_prev2	FLOAT
percnt_true_prev3	FLOAT
sum_severity_prev1	FLOAT
sum_severity_prev2	FLOAT
sum_severity_prev3	FLOAT

median_prev1	FLOAT
median_prev2	FLOAT
median_prev3	FLOAT
percnt_true_in_q1	BOOL
percnt_true_in_q2	BOOL
percnt_true_in_q3	BOOL
percnt_true_in_q4	BOOL
sum_severity_in_q1	BOOL
sum_severity_in_q2	BOOL
sum_severity_in_q3	BOOL
sum_severity_in_q4	BOOL

Regression models - ML

- Types of models used:
  - Logistic Regression
  - Random Forest
  - Decision Tree
  - Gradient Boosting
  - Support Vector Classifier (took long)
  - Linear Support Vector Classifier (took long)

- How it performed so far:
  - Current features used are from the soil moisture table (including min, max, mean of the moisture)
  - Most of them has good r-square score (>80%) and f1 score (>90%) using different cross validation techniques (kfold and kfold stratified)
  - Outperform so far: Random Forest

Models - MLP

- Preparing data for the model means converting categorical columns into numeric column with one-hot encoding like approach. And dealing with missing data and merging different data set.
- We are building Multi-layer perceptron with 1 hidden layer using sigmoid function for the output layer and relu function for the hidden layer.
- As this a binary classification problem we are using binary cross entropy as a loss function and Adam as a optimizer
- Precision and accuracy are too good to be true, which is likely >95% of the time. The suspect is that the number of true samples dominate the number of false samples.
- Metrics are used to measure the models: accuracy, precision, f1 score

Cross-validation techniques

- Normal train test split
- KFold
- KFold Stratified
- Leave one out

Regularization strategy

Normal distribution parameters

- Normalization scaling

Skewed distribution parameters

- Log scaling
- Then normalization

Biased data

- Random sampling with
    - Over sampling (of less dominant attribute)
    - Under sampling (of more dominant attribute)

Other options?

Goals for next 2 weeks

Dane:

- Models
- Dimensionality Reduction
- Update documentation i.e missing tables
- Interacting with the system?
  - Front-end?
  - Improved pipeline?

Daniel:

- Merge dataset - looking into what kind of datasets we should experiment on
- Continue modelling
- Write script to run all the models and comparing model on different of kind metrics (worth looking the custom metric?)

Dharmit:

- Research into data and models
- Data Visualisation on correlated data
- Look into aggregation methods
- Models

Jay:

- Figure out the best way to merge dataset with leaking information to model
- Understand data to decide which attribute to keep
- Create/improve more model (complex) if MLP starts giving promising results
- More research on data and models

Joseff:

- Data visualization not already done (e.g. yearly corr plots)
- Validate aggregation methods
- Models