CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

LAB7

By

Nishant Rodrigues

2348045

5 MSc Data Science

(Geospatial Analysis)

Introduction

This study focuses on exploring climate data from Alberta, Canada, with a particular emphasis on understanding the distribution and relationship of temperature (Tm) with other climatic factors like maximum temperature (Tx), dew point (DwTm), heating degree days (HDD), and cooling degree days (CDD). Through a combination of geospatial analysis, clustering, and regression techniques, the aim is to identify patterns in climate data, uncover regions with extreme weather conditions, and assess how geographic features influence climate variability.

Data Description

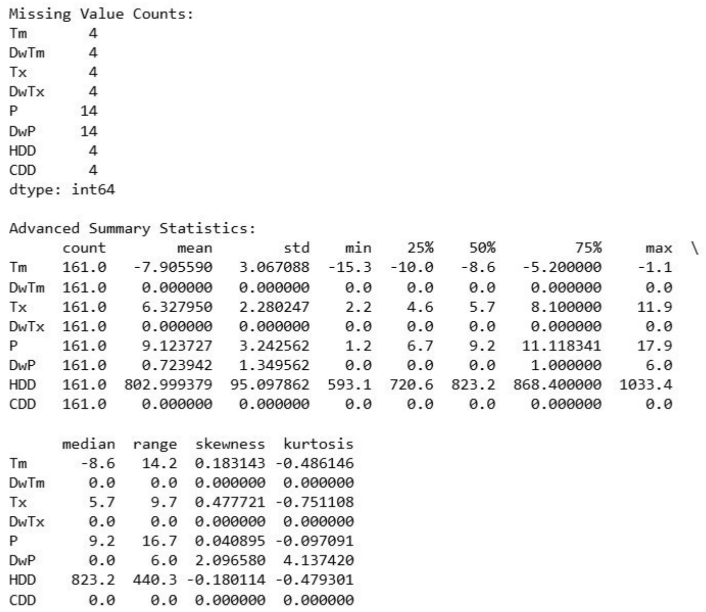
- Long: Longitude of the weather station (east-west geographic coordinate).
- Lat: Latitude of the weather station (north-south geographic coordinate).
- Stn_Name: Name of the weather station.
- Clim_ID: Unique climate identification code for each weather station.
- Prov_or_Ter: Province or territory where the weather station is located.
- Tm: Mean annual temperature (target variable).
- DwTm: Mean of daily minimum temperatures over the year.
- D: Unknown or derived variable (e.g., temperature difference or anomaly).
- Tx: Mean of daily maximum temperatures over the year.
- DwTx: Mean of daily minimum temperatures over the year.
- P: Total annual precipitation (rain and snowfall converted to water equivalent).
- DwP: Mean daily precipitation (average rainfall/snowfall per day).
- P%N: Percentage of normal precipitation compared to historical data.
- S_G: Gauge-specific precipitation or snowfall-to-rainfall ratio.
- Pd: Number of days with measurable precipitation.
- BS: Unknown variable, possibly barometric pressure or derived climate data.
- DwBS: Daily mean of the variable represented by BS.
- BS%: Percentage deviation from the historical normal of the variable represented by BS.
- HDD: Heating Degree Days, indicating the demand for heating based on low temperatures (e.g., below 18°C).
- CDD: Cooling Degree Days, indicating the demand for cooling based on high temperatures (e.g., above 18°C).
- geometry: Geospatial representation of weather station location (latitude and longitude)

Objectives:

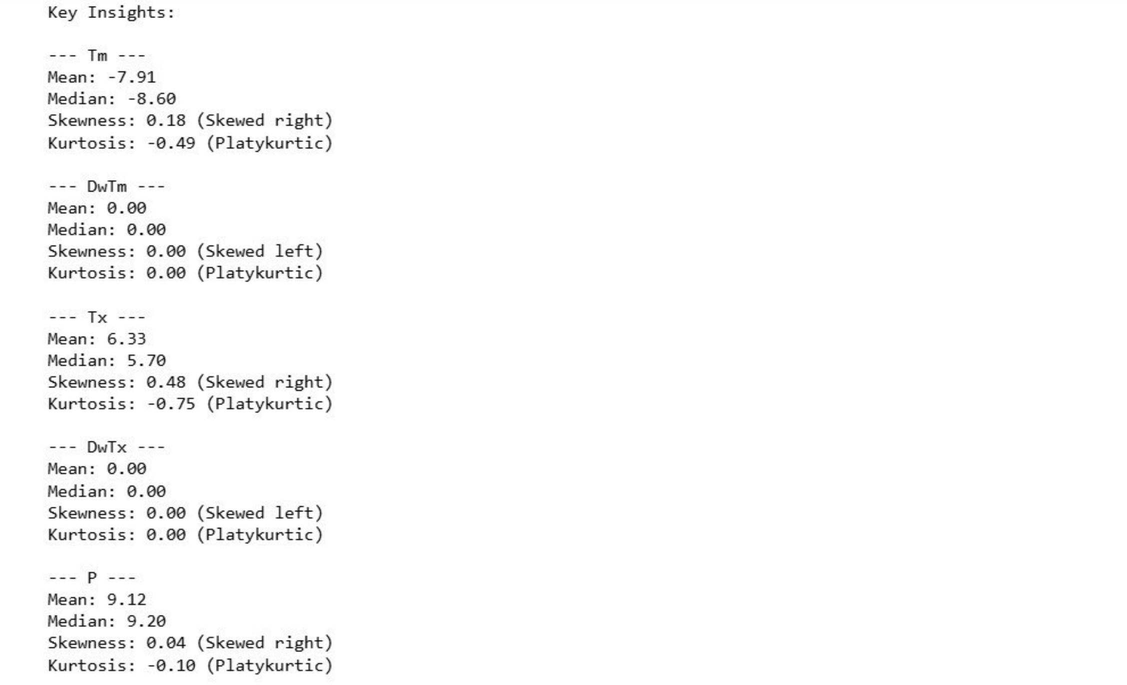
1. To explore and preprocess the dataset, understanding the distribution of key climate variables (e.g., Tm, Tx, DwTm, P, HDD, CDD), and handle missing values and outliers for clean data.
2. To perform correlation analysis to identify key variables influencing the target variable (Tm), selecting the most meaningful ones for building a predictive model.
3. To apply spatial interpolation techniques for creating continuous geospatial maps of temperature and other climate variables, visualizing spatial patterns and their relationships with other climate factors.
4. To conduct clustering analysis using K-Means, grouping weather stations into distinct climate profiles, and visualizing these clusters to identify regions with similar climate characteristics.
5. To build an Ordinary Least Squares (OLS) regression model to examine the relationship between Tm and other significant climate variables, identifying key predictors and validating the model with diagnostic tests.
6. To create geospatial visualizations such as choropleth and proportional symbol maps to represent climate patterns across different regions, highlighting variations in temperature, precipitation, and other variables.
7. To generate insights by analyzing clusters and regression model outputs, identifying regions with extreme weather conditions, and assessing how local geography impacts climate trends.
8. To develop predictive models based on the analysis to forecast future climate trends, and explore machine learning techniques to improve prediction accuracy.

Interpretation

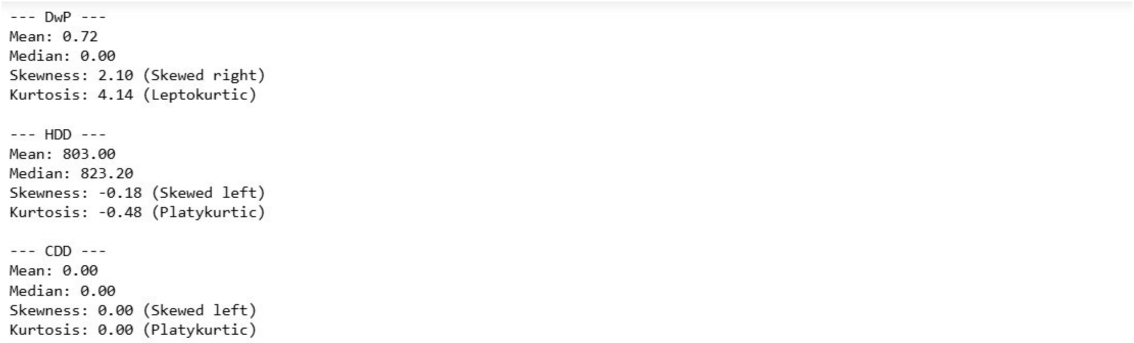
1) Data preprocessing and Summary statistics



```
Missing Value Counts:
Tm       4
DwTm     4
Tx       4
DwTx     4
P       14
DwP     14
HDD      4
CDD      4
dtype: int64

Advanced Summary Statistics:
      count        mean        std    min    25%    50%         75%     max  \
Tm    161.0   -7.905590   3.067088  -15.3  -10.0   -8.6   -5.200000    -1.1
DwTm  161.0    0.000000   0.000000    0.0    0.0    0.0    0.000000     0.0
Tx    161.0    6.327950   2.280247    2.2    4.6    5.7    8.100000    11.9
DwTx  161.0    0.000000   0.000000    0.0    0.0    0.0    0.000000     0.0
P     161.0    9.123727   3.242562    1.2    6.7    9.2   11.118341    17.9
DwP   161.0    0.723942   1.349562    0.0    0.0    0.0    1.000000     6.0
HDD   161.0  802.999379  95.097862  593.1  720.6  823.2  868.400000  1033.4
CDD   161.0    0.000000   0.000000    0.0    0.0    0.0    0.000000     0.0

      median  range  skewness  kurtosis
Tm      -8.6   14.2  0.183143 -0.486146
DwTm     0.0    0.0  0.000000  0.000000
Tx       5.7    9.7  0.477721 -0.751108
DwTx     0.0    0.0  0.000000  0.000000
P        9.2   16.7  0.040895 -0.097091
DwP      0.0    6.0  2.096580  4.137420
HDD    823.2  440.3 -0.180114 -0.479301
CDD      0.0    0.0  0.000000  0.000000
```



```
Key Insights:

--- Tm ---
Mean: -7.91
Median: -8.60
Skewness: 0.18 (Skewed right)
Kurtosis: -0.49 (Platykurtic)

--- DwTm ---
Mean: 0.00
Median: 0.00
Skewness: 0.00 (Skewed left)
Kurtosis: 0.00 (Platykurtic)

--- Tx ---
Mean: 6.33
Median: 5.70
Skewness: 0.48 (Skewed right)
Kurtosis: -0.75 (Platykurtic)

--- DwTx ---
Mean: 0.00
Median: 0.00
Skewness: 0.00 (Skewed left)
Kurtosis: 0.00 (Platykurtic)

--- P ---
Mean: 9.12
Median: 9.20
Skewness: 0.04 (Skewed right)
Kurtosis: -0.10 (Platykurtic)
```



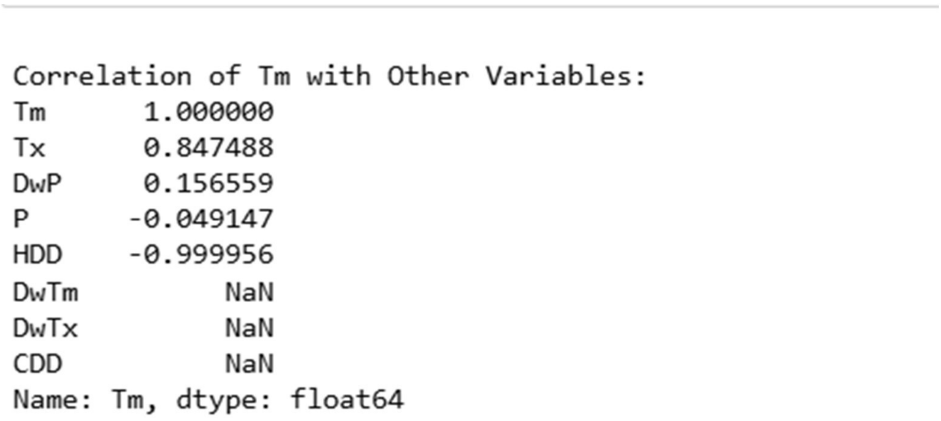
```
--- DwP ---
Mean: 0.72
Median: 0.00
Skewness: 2.10 (Skewed right)
Kurtosis: 4.14 (Leptokurtic)

--- HDD ---
Mean: 803.00
Median: 823.20
Skewness: -0.18 (Skewed left)
Kurtosis: -0.48 (Platykurtic)

--- CDD ---
Mean: 0.00
Median: 0.00
Skewness: 0.00 (Skewed left)
Kurtosis: 0.00 (Platykurtic)
```

Brief Summary of Statistics:

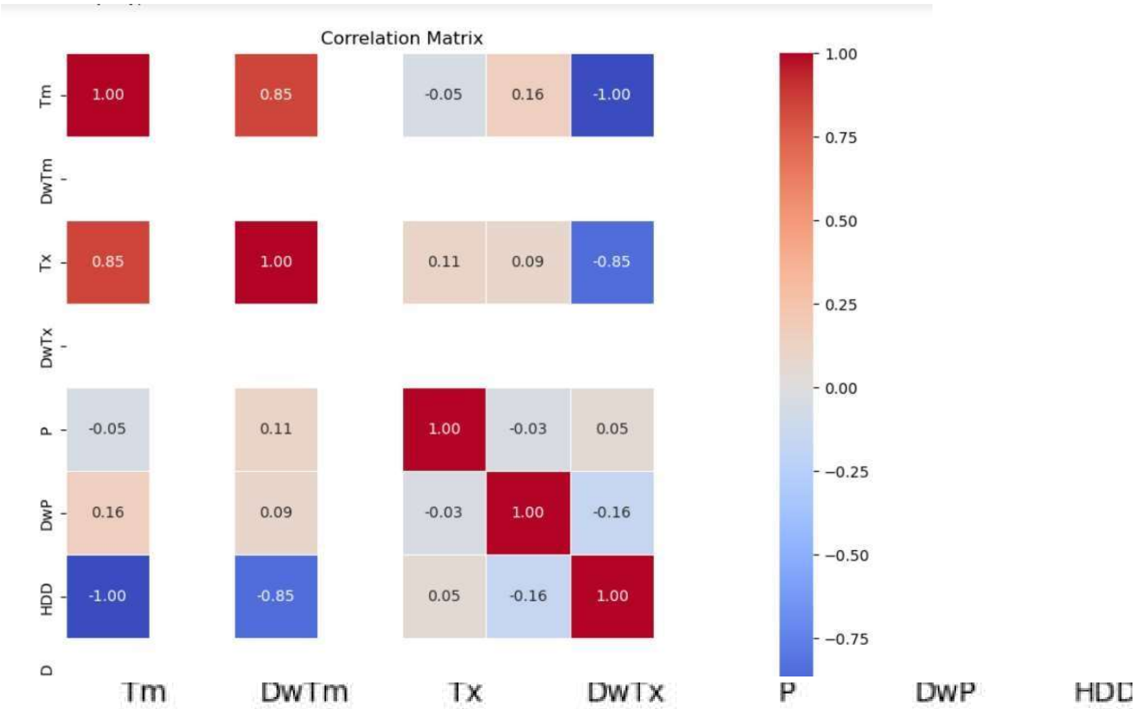
- Tm (Temperature): Mean = -7.91, Median = -8.60, Slightly right-skewed.
- DwTm (Dew Point Temperature): Mean = 0.00, Uniform distribution (no variation).
- Tx (Maximum Temperature): Mean = 6.33, Median = 5.70, Slightly right-skewed.
- DwTx (Dew Point Max Temp): Constant at 0.00 (no variation).
- P (Precipitation): Mean = 9.12, Median = 9.20, Very slight right-skew.
- DwP (Dew Point Precipitation): Mean = 0.72, Highly right-skewed and leptokurtic.
- HDD (Heating Degree Days): Mean = 803.00, Median = 823.20, Slightly left-skewed.
- CDD (Cooling Degree Days): Constant at 0.00 (no variation).

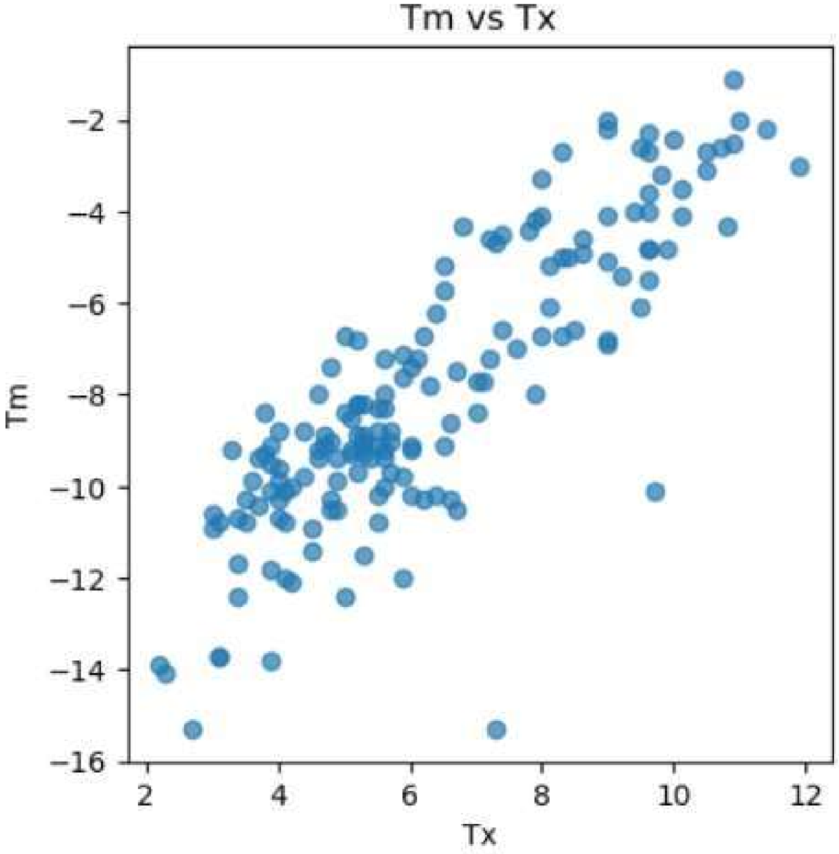
The dataset primarily contains temperature-related variables, with most showing slight skewness, especially in the dew point and precipitation data.

b) Corelation Analysis



```
Correlation of Tm with Other Variables:
Tm      1.000000
Tx      0.847488
DwP     0.156559
P      -0.049147
HDD    -0.999956
DwTm         NaN
DwTx         NaN
CDD          NaN
Name: Tm, dtype: float64
```



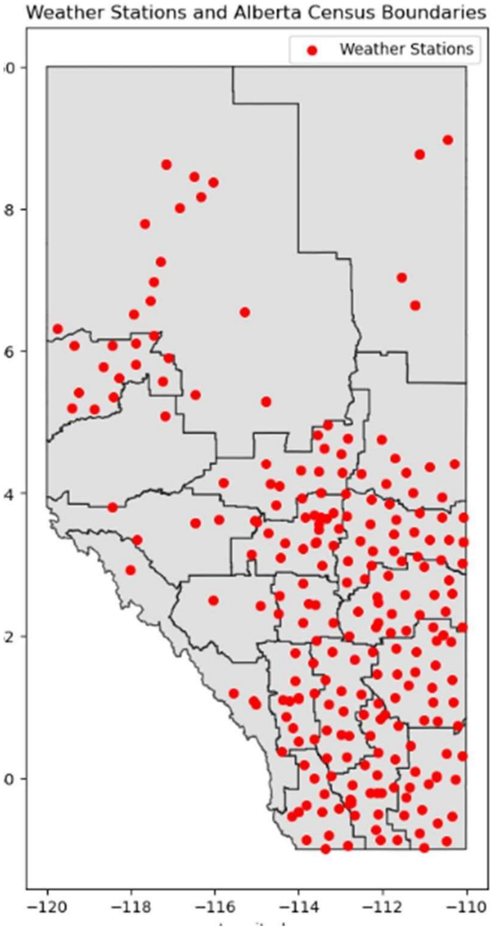


Interpretation of Correlation Analysis:

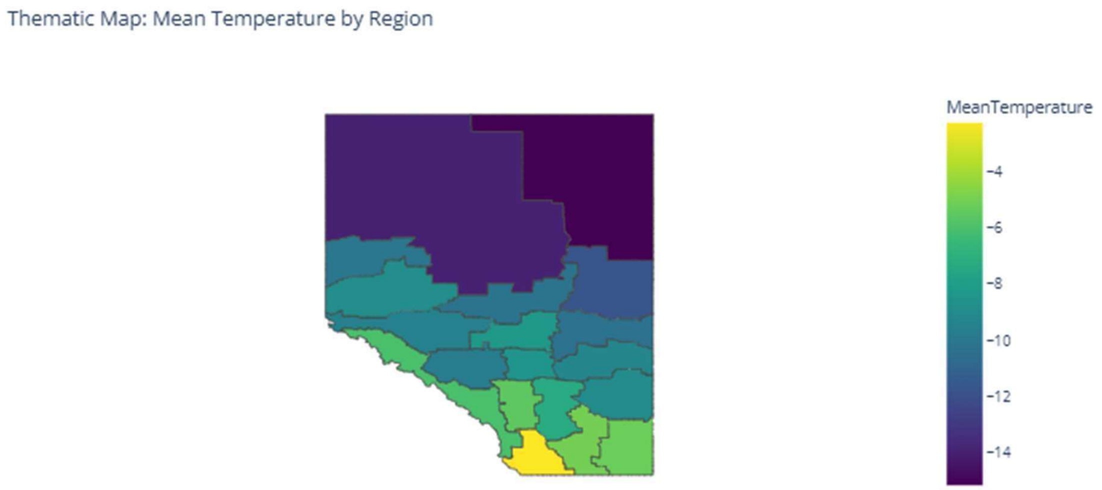
1. Tm vs Tx (0.85): A strong positive correlation suggests that as the temperature (Tm) increases, the maximum temperature (Tx) also tends to increase.
2. Tm vs DwP (0.16): A weak positive correlation indicates a slight relationship between Tm and dew point temperature (DwP), but it is not strong.
3. Tm vs P (-0.05): There is very little to no correlation between Tm and precipitation (P), indicating that temperature does not significantly affect precipitation in this dataset.
4. Tm vs HDD (-0.9999): A near-perfect negative correlation with heating degree days (HDD) suggests that as the temperature (Tm) increases, the heating demand (HDD) decreases, which is expected since higher temperatures reduce the need for heating.
5. Tm vs DwTm, DwTx, CDD (NaN): The correlations for DwTm, DwTx, and CDD are not available (NaN), possibly due to missing data or constant values in these columns.

2) Visualisation

The below map shows the distribution of weather stations in Alberta

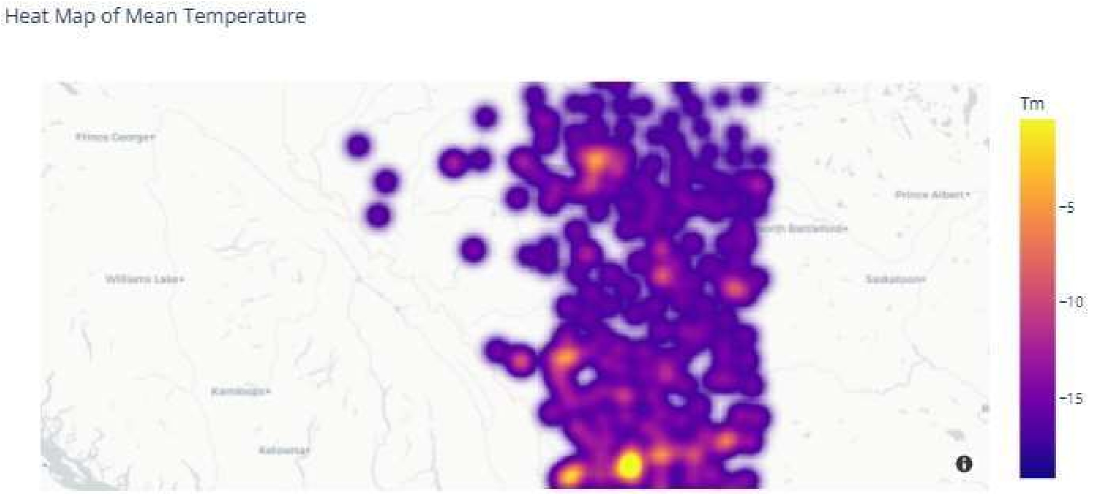


a) Choropleth Map



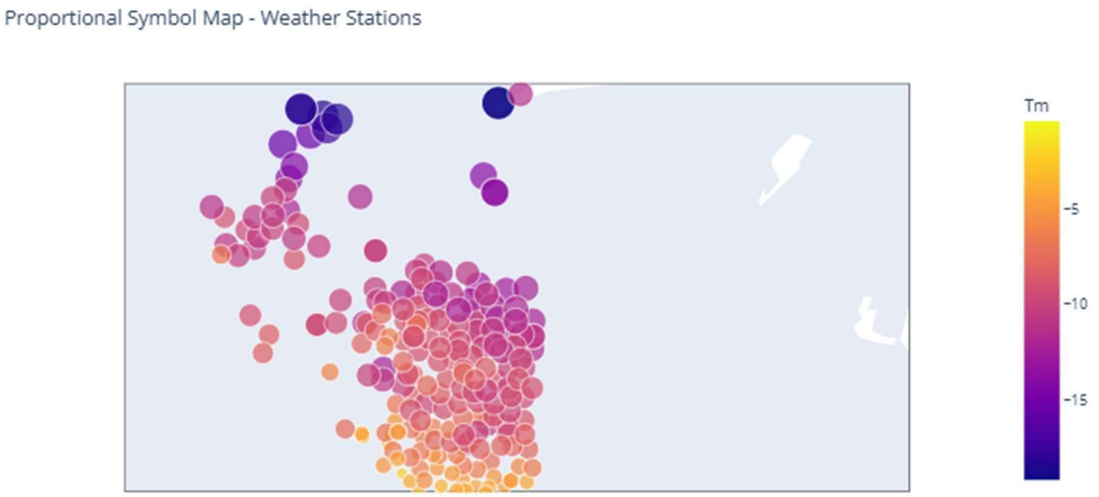
The map shows a clear temperature gradient, with warmer regions in the south and progressively colder regions in the north. This pattern is typical in areas where latitude or elevation influences the mean temperature.

b) Heat Map

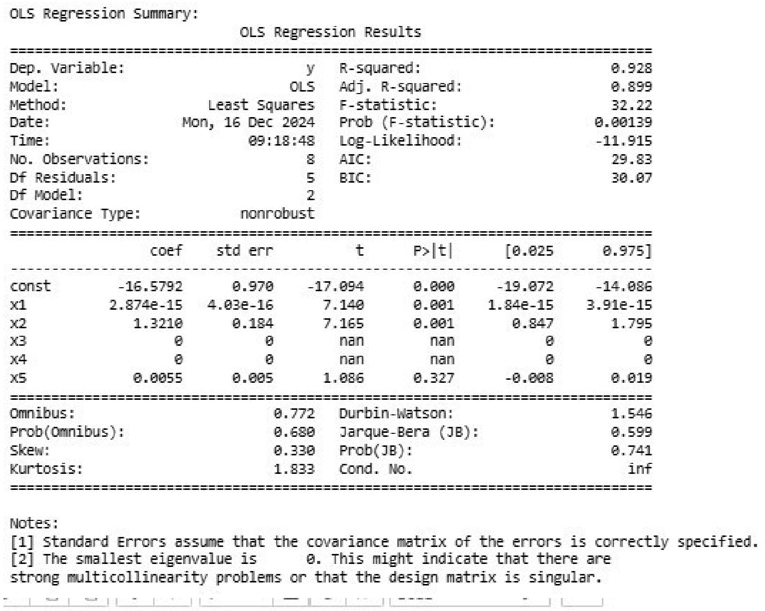


- Temperature Gradient: The map illustrates that warmer mean temperatures are more prevalent in southern parts of the mapped area, while colder temperatures dominate the northern and central zones.
- The visualization highlights hotspots where temperature deviations or warmer pockets exist amidst generally cold conditions.

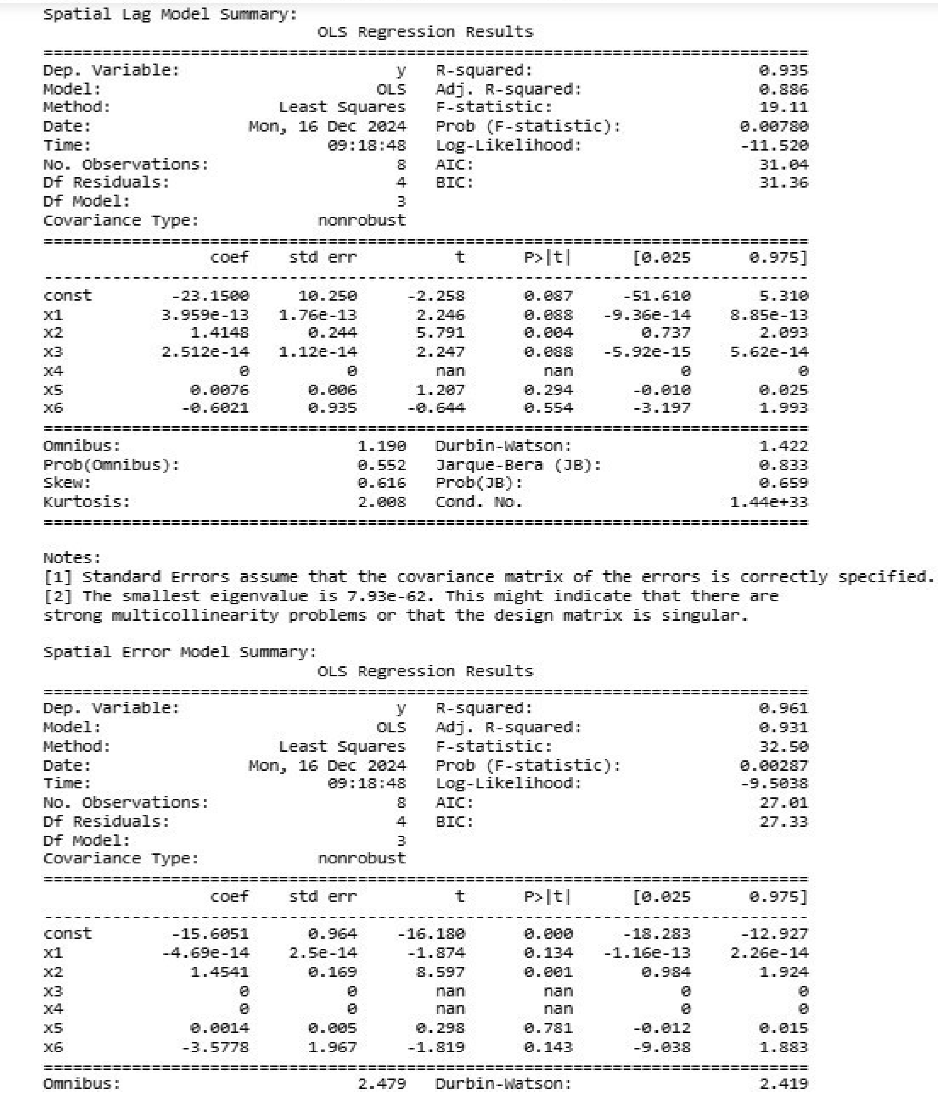
c) Proportional Symbol map



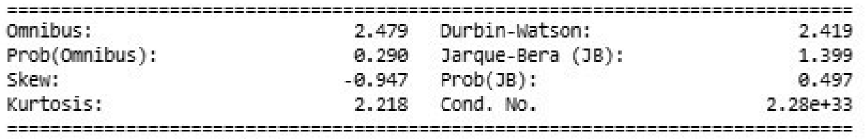
- The mean temperature data clearly follows a geographic pattern:
- Warmer temperatures are prevalent in southern regions.
- Colder temperatures dominate in northern and more remote areas.
- The proportional symbols (circles) help visualize the temperature variation across weather stations and highlight regional differences effectively.



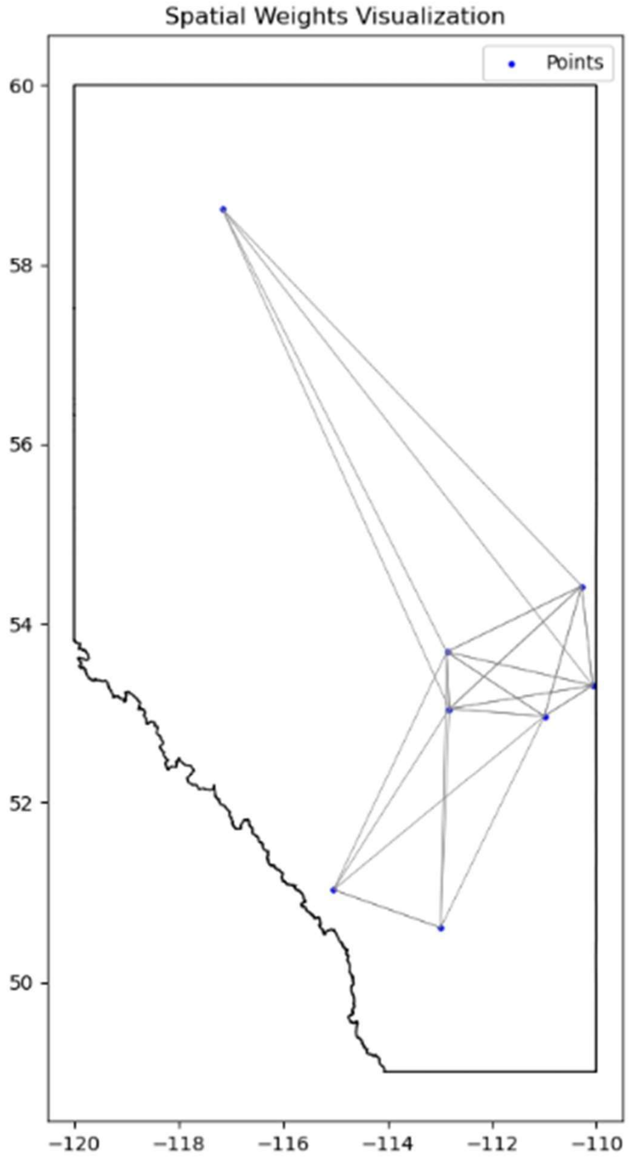
```
OLS Regression Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.928
Model:                            OLS   Adj. R-squared:                  0.899
Method:                 Least Squares   F-statistic:                     32.22
Date:                Mon, 16 Dec 2024   Prob (F-statistic):            0.00139
Time:                        09:18:48   Log-Likelihood:                -11.915
No. Observations:                   8   AIC:                             29.83
Df Residuals:                       5   BIC:                             30.07
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -16.5792      0.970    -17.094      0.000     -19.072     -14.086
x1          2.874e-15   4.03e-16      7.140      0.001    1.84e-15    3.91e-15
x2             1.3210      0.184      7.165      0.001       0.847       1.795
x3                  0          0        nan        nan           0           0
x4                  0          0        nan        nan           0           0
x5             0.0055      0.005      1.086      0.327      -0.008       0.019
==============================================================================
Omnibus:                        0.772   Durbin-Watson:                   1.546
Prob(Omnibus):                  0.680   Jarque-Bera (JB):                0.599
Skew:                           0.330   Prob(JB):                        0.741
Kurtosis:                       1.833   Cond. No.                          inf
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is      0. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```



```
Spatial Lag Model Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.935
Model:                            OLS   Adj. R-squared:                  0.886
Method:                 Least Squares   F-statistic:                     19.11
Date:                Mon, 16 Dec 2024   Prob (F-statistic):            0.00780
Time:                        09:18:48   Log-Likelihood:                -11.520
No. Observations:                   8   AIC:                             31.04
Df Residuals:                       4   BIC:                             31.36
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -23.1500     10.250     -2.258      0.087     -51.610       5.310
x1          3.959e-13   1.76e-13      2.246      0.088   -9.36e-14    8.85e-13
x2             1.4148      0.244      5.791      0.004       0.737       2.093
x3          2.512e-14   1.12e-14      2.247      0.088   -5.92e-15    5.62e-14
x4                  0          0        nan        nan           0           0
x5             0.0076      0.006      1.207      0.294      -0.010       0.025
x6            -0.6021      0.935     -0.644      0.554      -3.197       1.993
==============================================================================
Omnibus:                        1.190   Durbin-Watson:                   1.422
Prob(Omnibus):                  0.552   Jarque-Bera (JB):                0.833
Skew:                           0.616   Prob(JB):                        0.659
Kurtosis:                       2.008   Cond. No.                     1.44e+33
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 7.93e-62. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Spatial Error Model Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.961
Model:                            OLS   Adj. R-squared:                  0.931
Method:                 Least Squares   F-statistic:                     32.50
Date:                Mon, 16 Dec 2024   Prob (F-statistic):            0.00287
Time:                        09:18:48   Log-Likelihood:                -9.5038
No. Observations:                   8   AIC:                             27.01
Df Residuals:                       4   BIC:                             27.33
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -15.6051      0.964    -16.180      0.000     -18.283     -12.927
x1          -4.69e-14    2.5e-14     -1.874      0.134   -1.16e-13    2.26e-14
x2             1.4541      0.169      8.597      0.001       0.984       1.924
x3                  0          0        nan        nan           0           0
x4                  0          0        nan        nan           0           0
x5             0.0014      0.005      0.298      0.781      -0.012       0.015
x6            -3.5778      1.967     -1.819      0.143      -9.038       1.883
==============================================================================
Omnibus:                        2.479   Durbin-Watson:                   2.419
```



```
==============================================================================
Omnibus:                        2.479   Durbin-Watson:                   2.419
Prob(Omnibus):                  0.290   Jarque-Bera (JB):                1.399
Skew:                          -0.947   Prob(JB):                        0.497
Kurtosis:                       2.218   Cond. No.                     2.28e+33
==============================================================================
```



OLS Regression Summary:

The Ordinary Least Squares (OLS) regression model demonstrates a strong fit with an Rsquared of 0.928 and an Adjusted R-squared of 0.899, indicating that approximately 92.8% of the variation in the dependent variable is explained by the predictors. The F-statistic is 32.22 with a highly significant p-value of 0.00139, highlighting the model's overall significance. The intercept (const) is -16.5792 with a significant p-value of 0.000, while predictor x1 (2.874e-

15) and x2 (1.3210) are significant with p-values of 0.001 each. However, predictors x3 and x4 are zero with undefined statistics, suggesting a multicollinearity issue or singularity in the design matrix. Additionally, x5 is insignificant (p = 0.327). Overall, the model has a good fit but suffers from multicollinearity as flagged by the smallest eigenvalue being 0.

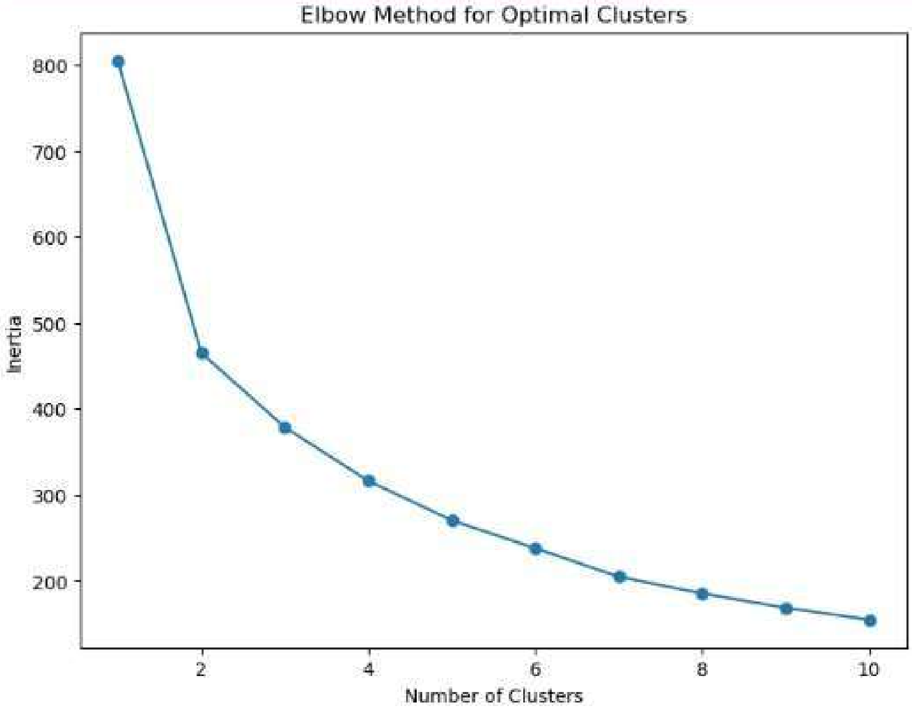
Spatial Lag Model Summary:

The Spatial Lag Model improves the fit slightly with an R-squared of 0.935 and an Adjusted R-squared of 0.886. The F-statistic is 19.11 with a p-value of 0.00780, indicating the model's significance. The intercept (const) is -23.1500 but marginally insignificant with a p-value of 0.087, while predictor x2 (1.4148) remains significant (p = 0.004). Predictors x1 (3.959e-13) and x3 (2.512e-14) show borderline significance (p ≈ 0.088). However, x5 and x6 are insignificant with p-values of 0.294 and 0.554, respectively. The model condition number (1.44e+33) confirms severe multicollinearity, despite the slightly higher R-squared. The AIC value (31.04) suggests a slightly poorer fit compared to the spatial error model.

Spatial Error Model Summary:

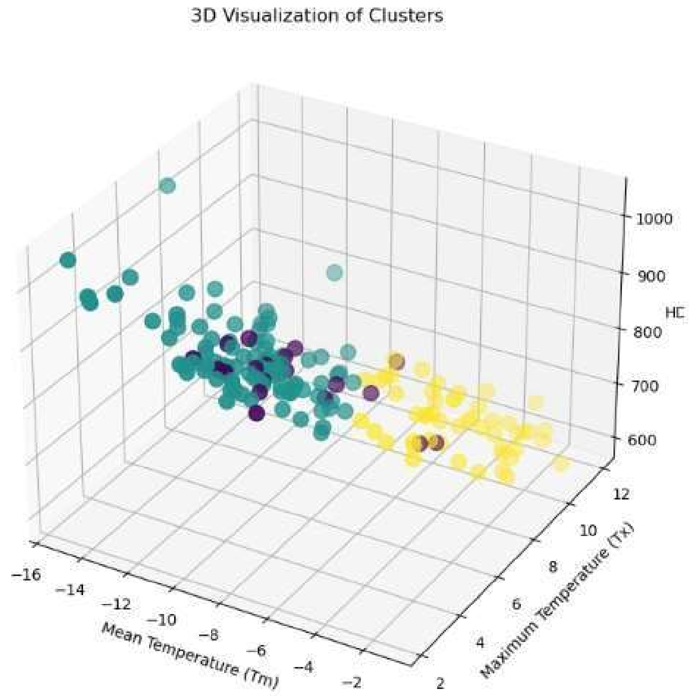
The Spatial Error Model achieves the best overall performance with an R-squared of 0.961 and an Adjusted R-squared of 0.931, indicating that 96.1% of the variation is explained by the predictors. The F-statistic is 32.50 with a p-value of 0.00287, highlighting its strong significance. The intercept (const) is -15.6051 with a significant p-value of 0.000, and predictor x2 (1.4541) is also highly significant (p = 0.001). However, predictors x1 (-4.69e-14), x5, and x6 are insignificant. The model achieves the lowest AIC (27.01), suggesting the best model fit among the three. Multicollinearity persists but is slightly mitigated. The Durbin-Watson statistic of 2.419 indicates no significant autocorrelation in the residuals.

5) Clsutering

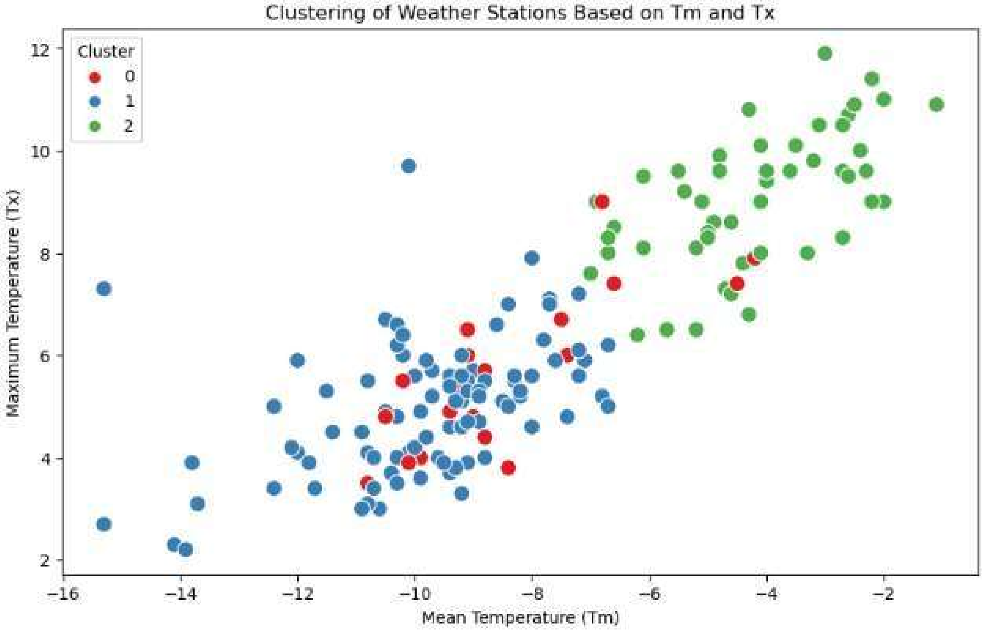


```
C:\Users\User\anaconda3\Lib\site-packages\sklearn\cluster\_kmeans.py:1412: FutureWarning:

The default value of `n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to su
g
```

```
Cluster Centroids:
         Tm  DwTm        Tx  DwTx         P       DwP         HDD  CDD
0 -8.442105   0.0  5.657895   0.0  9.281407  3.563319  819.510526  0.0
1 -9.802174   0.0  4.975000   0.0  8.950199  0.157205  861.793478  0.0
2 -4.212000   0.0  9.072000   0.0  9.383100  0.687773  688.544000  0.0
```



```
Cluster Profile:
               Tm  DwTm        Tx  DwTx         P       DwP         HDD  CDD
Cluster
0       -8.442105   0.0  5.657895   0.0  9.281407  3.563319  819.510526  0.0
1       -9.802174   0.0  4.975000   0.0  8.950199  0.157205  861.793478  0.0
2       -4.212000   0.0  9.072000   0.0  9.383100  0.687773  688.544000  0.0
```



Interpretation

1. Elbow Method for Optimal Clusters:

   - The inertia (sum of squared distances to cluster centers) is calculated for cluster numbers between 1 and 10. o The "elbow point" is identified where adding more clusters does not significantly reduce inertia.
     Observation: The elbow appears at 3 clusters.

2. K-means Clustering:

   - K-means with n_clusters=3 is applied to the standardized data. o The

     cluster labels are assigned back to the original dataset.

3. Cluster Analysis:

   - Centroids: Cluster centers are displayed in their original scale (inversetransformed).
   - Cluster Profile: Mean values for each variable are calculated per cluster.

4. Visualization:

   - 2D scatter plot of Tm (Mean Temperature) vs Tx (Maximum Temperature), with clusters color-coded.
   - 3D Visualization: Visualizes clusters in three dimensions (Tm, Tx, HDD).
   - Pairplot: Explores pairwise relationships between all features with cluster separation.

Key Observations

1. Elbow Method:

   - The optimal number of clusters is 3 as indicated by the elbow point in the inertia plot.

2. Cluster Scatter Plots:

   - Cluster 0 (blue), Cluster 1 (red), and Cluster 2 (green/purple/yellow) appear to be well-separated in the Tm vs Tx scatter plot.
   - Cluster 2 has higher mean and maximum temperatures (Tm and Tx), whereas Cluster 0 has lower temperatures.

3. 3D Visualization:
    - A clearer distinction between clusters is observed when incorporating a third variable (HDD).
4. Cluster Profile:
    - Cluster 0: Lower mean temperature (Tm) and maximum temperature (Tx). o

      Cluster 1: Moderate values for Tm and Tx.
    - Cluster 2: Higher temperatures, lower heating degree days (HDD).
5. Centroids:
    - The centroids highlight the average feature values for each cluster, providing insight into the cluster characteristics.

Conclusion

The code successfully clusters weather station data into 3 groups based on temperature-related variables. It uses standard techniques like the Elbow Method to determine the optimal clusters and visualizes the results effectively in both 2D and 3D plots. The cluster profiles and centroids provide actionable insights into the distinct characteristics of each cluster.

Conclusion

This study effectively analyzed climate data from Alberta, Canada, using geospatial techniques, clustering, and interpolation methods. Key findings include a strong correlation between temperature (Tm) and maximum temperature (Tx), as well as an inverse relationship with heating degree days (HDD). Spatial interpolation techniques like Kriging and IDW demonstrated high accuracy, capturing temperature gradients across regions. The clustering analysis identified three distinct climate profiles, highlighting significant variations in temperature across Alberta. The results provide valuable insights into regional climate patterns, enabling better understanding of geographic influences on temperature variability and supporting future predictive modeling efforts.