Task 4 Report
Roberto Jackson Baeza, 1861128

For this subtask I utilized Mahalanobis distance paired with a chi squared test threshold to give each data point an Outlier Score.

First I separated the desired attributes from the rest of the data, precipitation, sunshine, temp_mean and humidity. Then, as the given weather data used attributes with different scales and ranges it is pertinent to normalize the data. I used Min-Max Normalization on all of the attributes to ensure all attributes have an equal weight on the Mahalanobis distance calculation.

Once all points have been given an Mahalanobis Distance from the average day I then used 3 chi square thresholds of 1% of all examples are outliers, 5% outliers, and 10% outliers. And gave each point an outlier score based on the result of the distance divided by the threshold. I generated the outlier score based on the min max normalization results but I returned the actual value of each day so that the actual value could be scrutinized and understood. Below are the data points returned by the algorithm that I made.

For all of the Hyperparameters of the algorithm all the days returned are the same. This makes sense as the chi squared threshold only affects the OLS and it would affect every point evenly. Overall the algorithm found that the most unusual days tend to have High amounts of precipitation and humidity as well as low temp_mean and sunshine. This makes sense as the score is based on distance from average and so the days with greatest distance would have attributes that are either very small or very large. For the 2 most normal days they have very low amounts of precipitation with moderate amounts of sunshine, temp_mean, and humidity. This is intuitive as most days don't rain and are mild.

Top 6 Outliers: Alpha = 0.01

Rank	DATE	precipitation	sunshine	temp_mean	humidity	OLS
1	20010203	5.95	0	-17.1	0.9	0.5864
2	20090904	5.68	0.1	3	0.94	0.5458
3	20031101	5.35	0	-3.8	0.98	0.4991
4	20000317	5.08	0	-12.8	0.97	0.4780
5	20060429	4.61	0	-6.5	1	0.4172
6	20031008	4.57	0	-9.4	0.96	0.4168

Bottom 2 Normal: Alpha = 0.01

Rank	DATE	precipitation	sunshine	temp_mean	humidity	OLS
1	20030503	0.59	4.3	-3.9	0.85	0.0327
2	20031117	0.53	4.7	-7.4	0.87	0.0189

Top 6 Outliers: Alpha = 0.05

Rank	DATE	precipitation	sunshine	temp_mean	humidity	OLS

1	20010203	5.95	0	-17.1	0.9	0.8206
2	20090904	5.68	0.1	3	0.94	0.7637
3	20031101	5.35	0	-3.8	0.98	0.6984
4	20000317	5.08	0	-12.8	0.97	0.6689
5	20060429	4.61	0	-6.5	1	0.5838
6	20031008	4.57	0	-9.4	0.96	0.5832

Bottom 2 Normal: Alpha = 0.05

Rank	DATE	precipitation	sunshine	temp_mean	humidity	OLS
1	20030503	0.59	4.3	-3.9	0.85	0.0264
2	20031117	0.53	4.7	-7.4	0.87	0.0458

Results for Hyperparameter Setting Alpha = 0.1

Rank	DATE	precipitation	sunshine	temp_mean	humidity	OLS
1	20010203	5.95	0	-17.1	0.9	1.0008
2	20090904	5.68	0.1	3	0.94	0.9314
3	20031101	5.35	0	-3.8	0.98	0.8518
4	20000317	5.08	0	-12.8	0.97	0.8158
5	20060429	4.61	0	-6.5	1	0.7120
6	20031008	4.57	0	-9.4	0.96	0.7113

Bottom 2 Normal:

Rank	DATE	precipitation	sunshine	temp_mean	humidity	OLS
1	20030503	0.59	4.3	-3.9	0.85	0.0322
2	20031117	0.53	4.7	-7.4	0.87	0.0558

Citations:

OpenAI. (2023, Nov. 20). *ChatGPT* (Nov. 20 version) [Large language model]. https://chat.openai.com/chat

Sergen Cansiz. (2023, Nov. 20). *Multivariate Outlier Detection in Python.* Medium. https://towardsdatascience.com/multivariate-outlier-detection-in-python-e946cfc843b3

Hfahmida Data Science and Business Analytics. (2023, Nov. 20). *Visualizing Mahalanobis Distance using Python*. Medium. https://medium.com/@hfahmida/visualizing-mahalanobis-distance-using-python-1165ac426816#:~:text=Mahalanobis%20distance%20is%20a%20measure,outlier%20detection%20or%20clustering%20analysis.

Darth Espressius. (2023, Nov. 20). *Anomaly Detection I - Distance Based Methods.* Dev Community. https://dev.to/_aadidev/anomaly-detection-i-distance-based-methods-278g