1. Research objectives

Research questions:

1. What are the trends in weather conditions in Greece and how do they correlate with wildfire occurrences?
2. How do specific weather variables, such as temperature and wind speed, influence the frequency and severity of wildfires?

2. Data sources

2.1 Daily weather data

Description: This dataset offers weather data for Greece, including variables like temperature, wind speed. It was filtered to cover from week 41 of 2018 to week 41 of 2022.

- **Source**: Kaggle - Historical Weather Data of All Country Capitals
- **Data type**: CSV
- **License**: CC BY-SA 3.0
- **Data structure and quality**: The dataset includes columns for date, country, city, latitude, longitude, avg temp., min temp., max.temp., wind direction, wind speed. The data is structured in tabular format and cleaned to removee any missing entries.
- **Licensing obligations**: The dataset is under the Creative Commons BY-SA 3.0 license, allowing for sharing and adaptation with proper attribution. The obligations will be fulfilled by crediting the original source in all publications and derived works.

2.2 Wildfire occurrences

Description: This dataset contains records of wildfire occurrences, specifically the number of fire alerts per week. It was filtered to align with the time frame of the weather data.

- **Source**: Global Forest Watch - VIIRS Fire Alerts
- **Data type**: CSV
- **License**: Not specified
- **Data structure and quality**: The dataset includes columns for year, week, count of alerts, and confidence category. The data is structured in a tabular format and cleaned to focus on high-confidence alerts for the specified time frame.
- **Licensing obligations**: The dataset does not specify a license. Usage will comply with data usage terms stated on the website, ensuring proper attribution.

3. Data pipeline

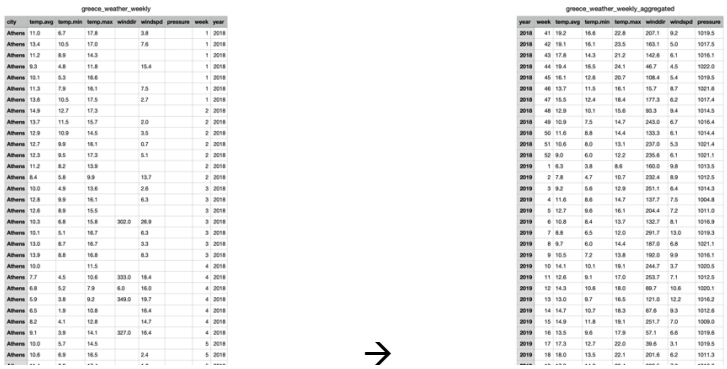
3.1 Overview

The data pipeline is designed to automate the extraction, transformation, and loading of data from the two sources to prepare it for analysis. The pipeline was implemented using Python, leveraging libraries such as Pandas for data manipulation and SQL for data storage.

3.2 Transformation and cleaning steps

Weather data:

- Filtered to include only records from Greece
- Converted date column to datetime format
- Extracted week number and year from the date
- Aggregated data to provide weekly averages for temperature and wind speed
- Dropped unnecessary columns such as latitude, longitude and city
- Renamed columns for clarity



greece_weather_weekly

city	temp.avg	temp.min	temp.max	winddir	windspd	pressure	week	year
Athens	11.0	6.7	17.8		3.8		1	2018
Athens	13.4	10.5	17.0		7.6		1	2018
Athens	11.2	8.9	14.3				1	2018
Athens	9.3	4.8	11.8		15.4		1	2018
Athens	10.1	5.3	16.6				1	2018
Athens	11.3	7.9	16.1		7.5		1	2018
Athens	13.6	10.5	17.5		2.7		1	2018
Athens	14.9	12.7	17.3				2	2018
Athens	13.7	11.5	15.7		2.0		2	2018
Athens	12.9	10.9	14.5		3.5		2	2018
Athens	12.7	9.9	16.1		0.7		2	2018
Athens	12.3	9.5	17.3		5.1		2	2018
Athens	11.2	8.2	13.9				2	2018
Athens	8.4	5.8	9.9		13.7		2	2018
Athens	10.0	4.9	13.6		2.6		3	2018
Athens	12.8	9.9	16.1		6.3		3	2018
Athens	12.6	8.9	15.5				3	2018
Athens	15.3	6.8	15.8	302.0	26.9		3	2018
Athens	10.1	5.1	16.7		6.3		3	2018
Athens	13.0	8.7	16.7		3.3		3	2018
Athens	13.9	8.8	16.8		8.3		3	2018
Athens	13.0		11.5				4	2018
Athens	7.7	4.5	10.6	333.0	18.4		4	2018
Athens	6.8	5.2	7.9	6.0	16.0		4	2018
Athens	5.9	3.8	9.2	349.0	19.7		4	2018
Athens	6.5	1.9	10.8		16.4		4	2018
Athens	8.2	4.1	12.8		14.7		4	2018
Athens	9.1	3.9	14.1	327.0	16.4		4	2018
Athens	10.0	5.7	14.5				5	2018
Athens	10.6	6.9	16.5		2.4		5	2018

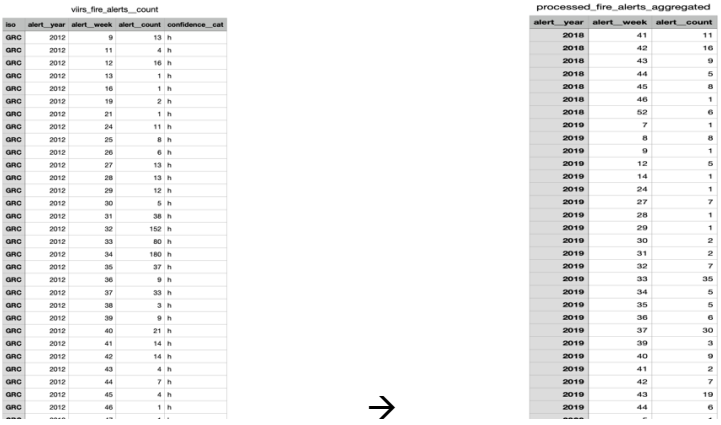
→

greece_weather_weekly_aggregated

year	week	temp.avg	temp.min	temp.max	winddir	windspd	pressure
2018	41	19.2	16.6	22.8	207.1	9.2	1019.5
2018	42	19.1	16.1	23.5	163.1	5.0	1017.5
2018	43	17.8	14.3	21.2	142.6	6.1	1016.1
2018	44	19.4	16.5	24.1	46.7	4.5	1022.0
2018	45	16.1	12.6	20.7	108.4	5.4	1019.5
2018	46	13.7	11.5	16.1	15.7	8.7	1021.8
2018	47	15.5	12.4	18.4	177.3	6.2	1017.4
2018	48	12.9	10.1	15.6	80.3	9.4	1014.5
2018	49	10.9	7.5	14.7	243.0	6.7	1016.4
2018	50	11.6	8.8	14.4	133.3	6.1	1014.4
2018	51	10.6	8.0	13.1	237.0	5.3	1021.4
2018	52	9.0	6.0	12.2	235.6	6.1	1021.1
2019	1	6.3	3.8	8.6	160.0	9.8	1013.5
2019	2	7.8	4.7	10.7	232.4	8.9	1012.5
2019	3	9.2	5.8	12.9	251.1	6.4	1014.3
2019	4	11.6	8.6	14.7	137.7	7.5	1004.8
2019	5	12.7	9.6	16.1	204.4	7.2	1011.0
2019	6	10.8	8.4	13.7	132.7	8.1	1016.9
2019	7	8.8	6.5	12.0	291.7	13.0	1019.3
2019	8	9.7	6.0	14.4	187.0	6.8	1021.1
2019	9	10.5	7.2	13.8	192.0	9.9	1016.1
2019	10	14.1	10.1	19.1	244.7	3.7	1020.5
2019	11	12.6	9.1	17.0	253.7	7.1	1012.5
2019	12	14.3	10.6	18.0	69.7	10.6	1020.1
2019	13	13.0	9.7	16.5	121.0	12.2	1016.2
2019	14	14.7	10.7	18.3	67.8	9.3	1012.6
2019	15	14.9	11.8	19.1	251.7	7.0	1009.0
2019	16	13.5	9.6	17.9	57.1	6.6	1019.6
2019	17	17.3	12.7	22.0	39.6	3.1	1019.5
2019	18	18.0	13.5	22.1	201.6	6.2	1011.3

Wildfire data:

- Filtered to include records from 2018 to 2022 as well
- Dropped columns not relevant to the analysis (e.g., ISO code, confidence category)
- Aggregated data to provide weekly counts of fire alerts
- Renamed columns for clarity



viirs_fire_alerts__count

iso	alert__year	alert__week	alert__count	confidence__cat
GRC	2012	9	13	h
GRC	2012	11	4	h
GRC	2012	12	16	h
GRC	2012	13	1	h
GRC	2012	16	1	h
GRC	2012	19	2	h
GRC	2012	21	1	h
GRC	2012	24	11	h
GRC	2012	25	8	h
GRC	2012	26	6	h
GRC	2012	27	13	h
GRC	2012	28	13	h
GRC	2012	29	12	h
GRC	2012	30	5	h
GRC	2012	31	38	h
GRC	2012	32	152	h
GRC	2012	33	80	h
GRC	2012	34	180	h
GRC	2012	35	37	h
GRC	2012	36	9	h
GRC	2012	37	33	h
GRC	2012	38	3	h
GRC	2012	39	9	h
GRC	2012	40	21	h
GRC	2012	41	14	h
GRC	2012	42	14	h
GRC	2012	43	4	h
GRC	2012	44	7	h
GRC	2012	45	4	h
GRC	2012	46	1	h

→

processed_fire_alerts_aggregated

alert__year	alert__week	alert__count
2018	41	11
2018	42	16
2018	43	9
2018	44	5
2018	45	8
2018	46	1
2018	52	6
2019	7	1
2019	8	8
2019	9	1
2019	12	5
2019	14	1
2019	24	1
2019	27	7
2019	28	1
2019	29	1
2019	30	2
2019	31	2
2019	32	7
2019	33	35
2019	34	5
2019	35	5
2019	36	6
2019	37	30
2019	39	3
2019	40	9
2019	41	2
2019	42	7
2019	43	19
2019	44	6

3.3 Merging datasets

- Merged the weather data and wildfire data on the common columns 'year' and 'week'
- Ensured the merged dataset covered the same time period from week 41 of 2018 to week 41 of 2022
- The merged dataset allows for correlation analysis between weather variables and wildfire occurrences

merged_weather_fire_alerts

year	week	temp.avg	temp.min	temp.max	winddir	windspd	pressure	alert_count
2018	41	19.2	16.6	22.8	207.1	9.2	1019.5	11
2018	42	19.1	16.1	23.5	163.1	5.0	1017.5	16
2018	43	17.8	14.3	21.2	142.6	6.1	1016.1	9
2018	44	19.4	16.5	24.1	46.7	4.5	1022.0	5
2018	45	16.1	12.6	20.7	108.4	5.4	1019.5	8
2018	46	13.7	11.5	16.1	15.7	8.7	1021.6	1
2018	52	9.0	6.0	12.2	235.6	6.1	1021.1	6
2019	7	8.8	6.5	12.0	291.7	13.0	1019.3	1
2019	8	9.7	6.0	14.4	187.0	6.8	1021.1	8
2019	9	10.5	7.2	13.8	192.0	9.9	1016.1	1
2019	12	14.3	10.6	18.0	69.7	10.6	1020.1	5
2019	14	14.7	10.7	18.3	67.6	9.3	1012.6	1
2019	24	27.1	22.5	32.2	151.3	5.9	1012.3	1
2019	27	28.5	23.0	34.0	99.7	7.6	1012.1	7
2019	28	27.4	23.2	32.8	132.4	6.4	1008.7	1
2019	29	26.3	21.8	30.4	123.6	9.1	1010.9	1
2019	30	28.8	24.4	33.0	51.7	9.1	1011.3	2
2019	31	29.7	25.0	34.4	218.9	6.7	1007.1	2
2019	32	29.4	24.9	33.9	58.9	10.2	1011.9	7
2019	33	28.4	24.2	33.7	83.7	7.9	1009.9	35
2019	34	29.2	25.1	32.9	7.4	12.2	1014.8	5
2019	35	28.4	24.1	32.9	2.7	11.6	1013.7	5
2019	36	27.0	22.8	31.6	155.1	8.0	1012.3	6
2019	37	25.6	22.0	29.8	105.6	10.9	1016.8	30
2019	39	23.0	19.0	28.1	231.1	4.3	1014.2	3
2019	40	23.1	19.5	28.2	213.6	5.6	1010.2	9
2019	41	20.9	17.7	25.2	69.9	6.2	1018.1	2
2019	42	21.2	17.4	26.5	11.3	2.9	1018.3	7
2019	43	21.3	17.4	26.5	256.4	8.3	1018.8	19
2019	44	18.1	15.0	22.7	131.9	4.8	1016.8	6
2020	5	12.9	9.1	17.6	248.4	5.8	1017.9	1

3.4 Problems and solutions

- **Data gaps**: Managed by using available data and documenting the missing weeks
- **Dataset disparity**: Acknowledged the difference in geographical scope and adjusted the analysis accordingly
- **Sample size**: Limited number of observations addressed by aggregating data

3.5 Errorhandling

The pipeline includes error handling mechanisms to manage missing data and inconsistencies. It is designed to handle changes in the input data structure by dynamically adjusting the data cleaning steps based on the presence of expected columns.

4. Result and limitations

4.1 Output data

The final output of the data pipeline is a merged dataset that combines the weather and wildfire data on a weekly basis from week 41 of 2018 to week 41 of 2022. This dataset is saved in CSV due to its compatibility and ease of analysis.

4.2 Data structure and format choice

The merged dataset includes columns for year, week, avg. temp., min. temp., max. temp., wind direction, wind speed, pressure, and count of fire alerts. The data is structured in a tabular format and cleaned to ensure consistency.

4.4 Potential issues

- **Geographical scope**: The weather data pertains to the capital city of Greece, while the wildfire data covers the entire country. This discrepancy can limit the comparability of the datasets.
- **Sample size**: The number of observations (N) is relatively low, especially after filtering and merging the datasets, which could impact the robustness of the findings.

References

1. Daily Weather Data: Kaggle Dataset
2. Global Forest Watch: VIIRS Fire Alerts