

CIS3200 Term Project Tutorial CIS3200 Te Project Tutorial



Authors: Arne Muller, Erik Pacheco, Michael Do, Shadman Sayef, & Steven Vi

Instructor: [Jongwook Woo](#)

Date: 12/16/2023

Lab Tutorial

Storm Events Data Analysis using Elastic Cloud

Objectives

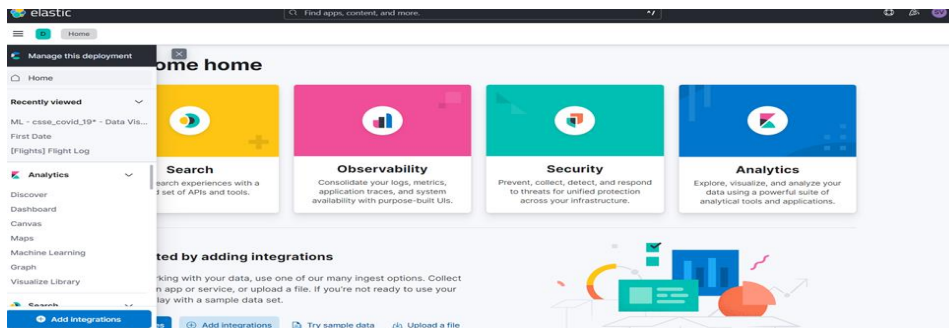
List what your objectives are. In this hands-on lab, you will learn how to:

- Get data manually
- Create prediction analysis, geo-map, & visual charts
- Using Elasticsearch cloud system
- Using Elasticsearch cloud to perform the analysis.
- Visualization

Platform Spec

- Windows 11
- CPU Speed: 1.35 GHz
- Number of CPU cores: 16 cores
- Number of nodes: 4
- Total Memory Size: 31 GB

Machine Learning



1. After logging into the cloud server, go to Machine Learning.

Data Visualizer

File

Data View

Data Comparison

2. Go to Data visualizer and go to file.
3. From there, upload the CSV files. In this case, the storm detail files from either 2010 and 2011.

StormEvents_details-ftp_v1.0_d2011_c20230417.csv [Select a different file](#)

File contents
 First 1,000 lines

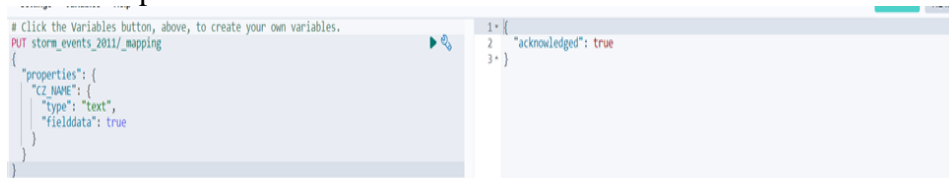
```

1 BEGIN_YEAR, BEGIN_MONTH, BEGIN_DAY, BEGIN_TIME, END_YEAR, END_MONTH, END_DAY, END_TIME, EPISODE_ID, EVENT_ID, STATE, STATE_FIPS, YEAR, MONTH_NAME, EVENT_TYPE, CZ_TYPE, CZ_FIPS, CZ_NAME, WFO, BEGIN_DATE_TIME, CZ_TIMEZONE, END_DATE_TIME, INJURIES_DIRECT, INJURIES_INDIRECT, DEATHS_DIRECT, DEATHS_INDIRECT, DAMAGE_PROPERTY, DAMAGE_CROPS, SOURCE, MAGNITUDE, MAGNITUDE_TYPE, FLOOD_CAUSE, CATEGORY, TOR_F_SCALE, TOR_LENGTH, TOR_WIDTH, TOR_OTHER_WFO, TOR_OTHER_CZ_STATE, TOR_OTHER_CZ_FIPS, TOR_OTHER_CZ_NAME, BEGIN_RANGE, BEGIN_AZIMUTH, BEGIN_LOCATION, END_RANGE, END_AZIMUTH, END_LOCATION, BEGIN_LAT, BEGIN_LON, END_LAT, END_LON, EPISODE_NARRATIVE, EVENT_NARRATIVE, DATA_SOURCE
2 201102,27,300,201102,27,1500,49358,289076,"NEW HAMPSHIRE",33,2011,"February","Winter Storm","Z",12,"EASTERN HILLSBOROUGH","BOX","27-FEB-11 03:00:00","EST-5","27-FEB-11 15:00:00","0","0","0","0","0.00K","0.00K","Trained Spotter",,,,,,,,,,,,,,"A weak area of low pressure moved eastward out of the Great Lakes Region and off the New England shoreline gradually pushing out to sea. Mainly advisory level snows were experienced over much of the forecast region as a very cold air mass was in place over the region with slightly drier conditions.",,"A total of 6 to 8 inches of snow fell
        
```

Summary

Number of lines analyzed	1000
Format	delimited
Delimiter	,
Has header row	true

4. Import the files as is.
5. Make the index name: *storm_events_2011*
6. Import the file.



7. From Kibana menu, go to Dev Tools. Type in **PUT storm_events_2011/_mapping**

```

{
  "properties": {
    "CZ_NAME": {
      "type": "text",
      "fielddata": true
    }
  }
}

```

}
}
}

- This is necessary for both storm_events_2010 and storm_events_2011.

8. From Kibana menu, go back to Machine Learning.

9. Go back to jobs and create job.

10. From there, pick the dataset storm_events_2010 or storm_events_2011.

11. Select Regression.

12. From dependent variable, select MAGNITUDE.

13. Training Percent can be 80 percent.

2 Additional options

Advanced configuration

Feature importance values

Specify the maximum number of feature importance values per document to return.

Model memory limit

☒ Use estimated model memory limit

The approximate maximum amount of memory resources that are permitted for analytical processing.

> Hyperparameters

Continue

Prediction field name

Define the name of the prediction field in the results. The default is <dependent_variable>_prediction.

Maximum number of threads

The maximum number of threads to be used by the analysis. The default value is 1.

Randomize seed

The seed for the random generator used to pick training data.

14. Put 5 in feature importance values and 150mb in model memory limit.

3 Job details

Job ID

An analytics job with this ID already exists.

Job description

☒ Destination index same as job ID

An index with this name already exists. Be aware that running this analytics job will modify this destination index.

☒ Use results field default value: "ml"

Continue

15. For job id, type in regression_model_storm_events_2010 or 2011.

16. After validation, press continue.

17. Finally, just wait for the data to be uploaded into the cloud.

18. You must do these same steps for both storm events 2010 and storm events 2011. From steps 2-17 to have them successfully uploaded to the cloud server.

19. Go back to jobs from the Kibana menu.

20. On the right, go to the models, click view.

×

Training
Testing

ation index **regression_model_storm_events_2010**

[Regression evaluation docs](#)

Training error
 22,596 docs evaluated

7.38
 Mean squared error [?](#)

0.989
 R squared [?](#)

NaN
 Mean squared logarithmic error [?](#)

1.06
[Pseudo Huber loss function](#) [?](#)

21. Press the training button and you'll find MSE and R-squared.

Model evaluation
Job status **stopped**

Generalization error
 40,210 docs evaluated

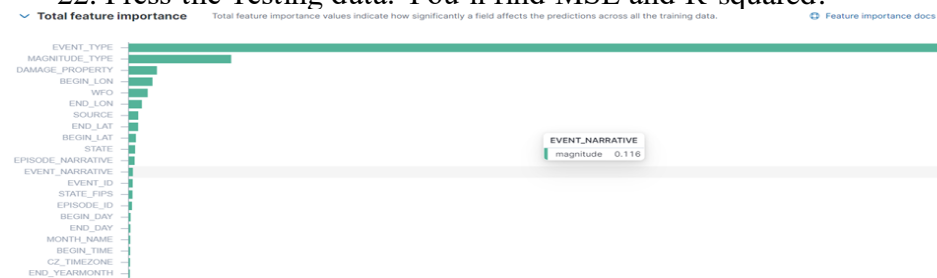
33.3
 Mean squared error [?](#)

0.938
 R squared [?](#)

0.0471
 Mean squared logarithmic error [?](#)

2.69
[Pseudo Huber loss function](#) [?](#)

22. Press the Testing data. You'll find MSE and R-squared.



23. Scroll down and you'll find total feature importance. You'll find that EVENT_TYPE has the highest magnitude.

47 columns hidden Sort fields Histogram charts

ml.is_training	ml.MAGNITUDE_pred...	MAGNITUDE	ml.feature_importance	BEGIN_AZIMUTH	BEGIN_DATE_TIME	BEGIN_DAY	BEGIN_LAT
false		33.785	[{"feature_name": "EVENT...		07-JUL-10 12:51:00		7
false		34.584	[{"feature_name": "EVENT...		17-JAN-10 23:00:00		17
false		41.707	[{"feature_name": "EVENT...		01-OCT-10 08:30:00		1
false		32.955	[{"feature_name": "EVENT...		08-JUL-10 09:51:00		6
false		35.617	[{"feature_name": "EVENT...		26-DEC-10 17:00:00		26
false		60.918	[{"feature_name": "EVENT...		25-FEB-10 23:05:00		25
false		32.482	[{"feature_name": "EVENT...		16-FEB-10 12:00:00		16
false		39.652	[{"feature_name": "EVENT...		14-MAR-10 13:45:00		14
false		34.378	[{"feature_name": "EVENT...	ENE	21-JAN-10 03:00:00		21
false		42.639	[{"feature_name": "EVENT...	ESE	11-MAR-10 18:16:00		11
false		0.924	[{"feature_name": "EVENT...	W	06-APR-10 16:42:00		6
false		32.835	[{"feature_name": "EVENT...		20-NOV-10 13:00:00		20
false		32.835	[{"feature_name": "EVENT...		20-NOV-10 13:00:00		20
false		42.516	[{"feature_name": "EVENT...	NRW	12-JUL-10 14:54:00		12
false		42.182	[{"feature_name": "EVENT...	NRW	12-JUL-10 15:10:00		12
false		42.435	[{"feature_name": "EVENT...	N	25-JUL-10 15:45:00		25
false		40.93	[{"feature_name": "EVENT...	SSW	25-JUL-10 16:00:00		25

24. Scroll down further and you'll find ML_feature importance.

TUDE	ml.feature_importance	BEGIN_AZIMUTH
	[{"feature_name": "EV...	

Decision plot JSON

```

name": "EVENT_TYPE", "importance": 9.9573857454566
name": "MAGNITUDE_TYPE", "importance": -6.83082202220297
name": "DAMAGE_PROPERTY", "importance": -0.71302022204838

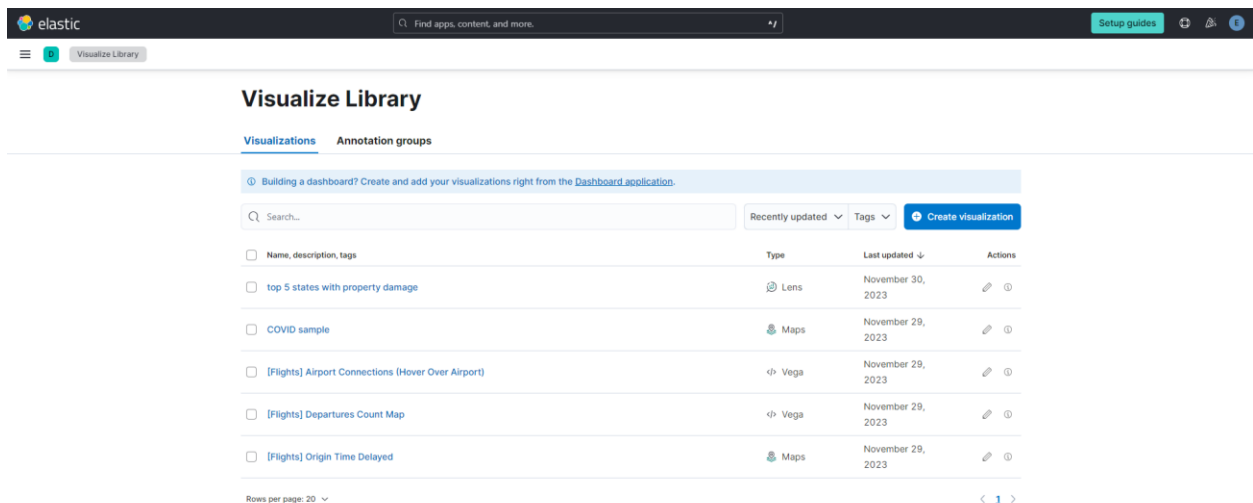
```

25. From here, click on one of the labels, and you'll see how much of an impact each feature has with the dataset. We learn that EVENT_TYPE has the highest magnitude of 9.95 from the 2010 dataset.

26. That's it for the tutorial on how to use machine learning for these CSV files.


Create Visualizations

1. Open the Visualize Library in Elastic.
2. Click Create Visualization



3. Select Lens.

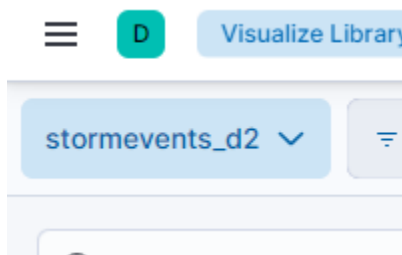
New visualization



Lens

Create visualizations with our drag and drop editor. Switch between visualization types at any time. *Recommended for most users.*

4. Make sure the stormevents selection is made:



5. Drag and drop the TOP 5 VALUES OF STATE and UNIQUE COUNT OF PROPERTY DAMAGE

Bar horizontal

▼

☰

stormevents_d2

▼

Vertical axis

Optional

⚠️

Top 5 values of STATE

Horizontal axis

■

Unique count of DAMAGE_PROPERTY

+

Add or drag-and-drop a field

Breakdown

Optional

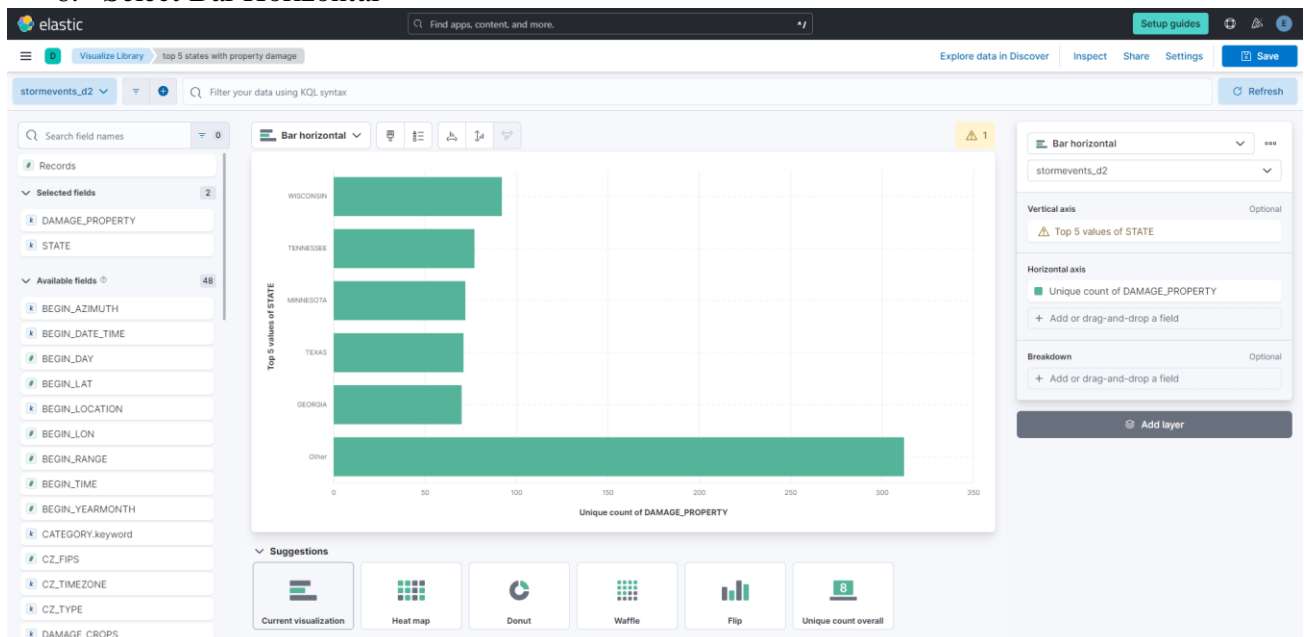
+

Add or drag-and-drop a field

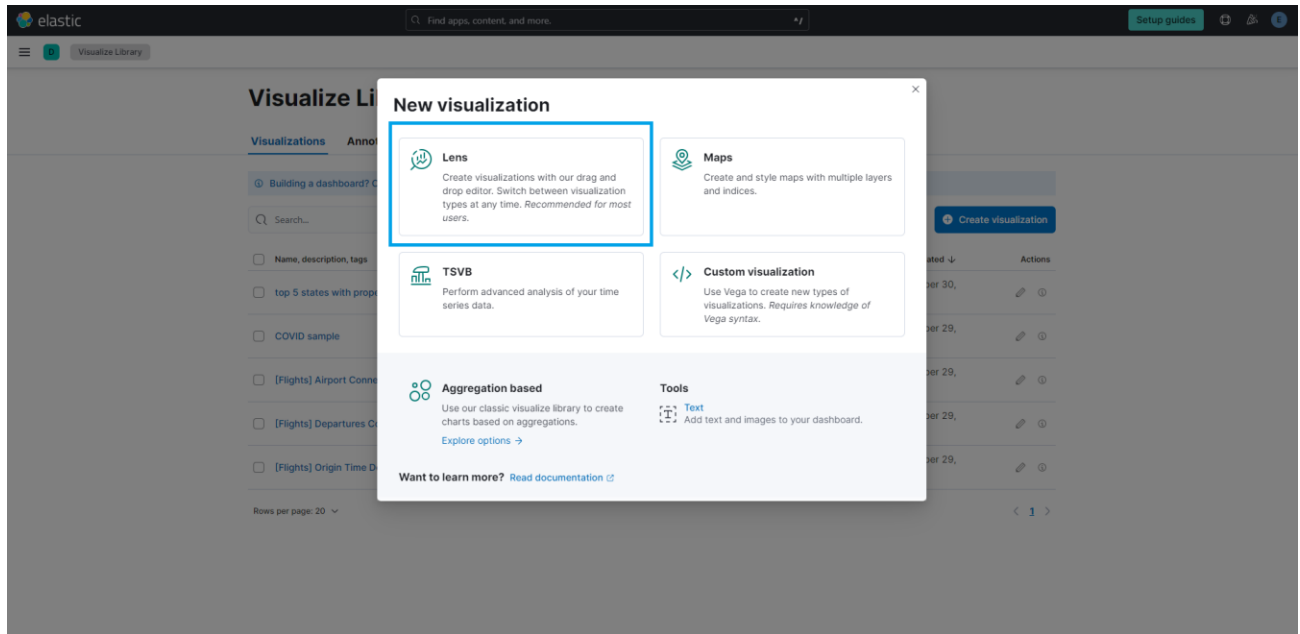
☰

Add layer

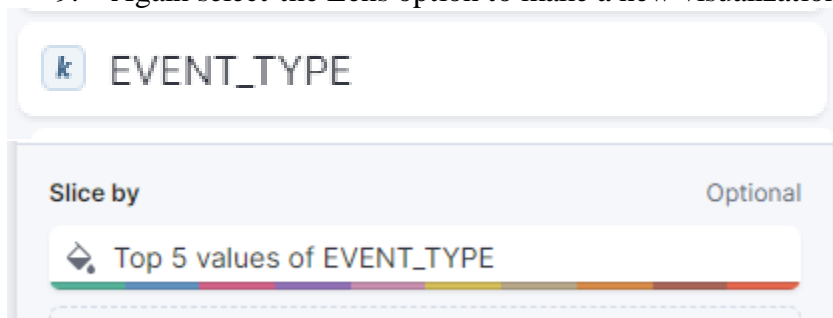
6. Select Bar Horizontal



7. Save and name the Bar Chart Visualization.
8. Open the Visualize Library in Elastic to make another visualization.



9. Again select the Lens option to make a new visualization.



10. Drag event type into the visualization space.

11. Select pie chart.

12. Click on Top 5 Values of EVENT TYPE, and click on Advanced settings.

Slice

×

5

Rank by

Count of records

Rank direction

Ascending

Descending

Collapse by

None

Advanced

Include documents without the selected field

Group remaining values as "Other"

Enable accuracy mode

Include values

Use regular expression

Select values or create a new one

Exclude values

Use regular expression

Select values or create a new one

Appearance

Name

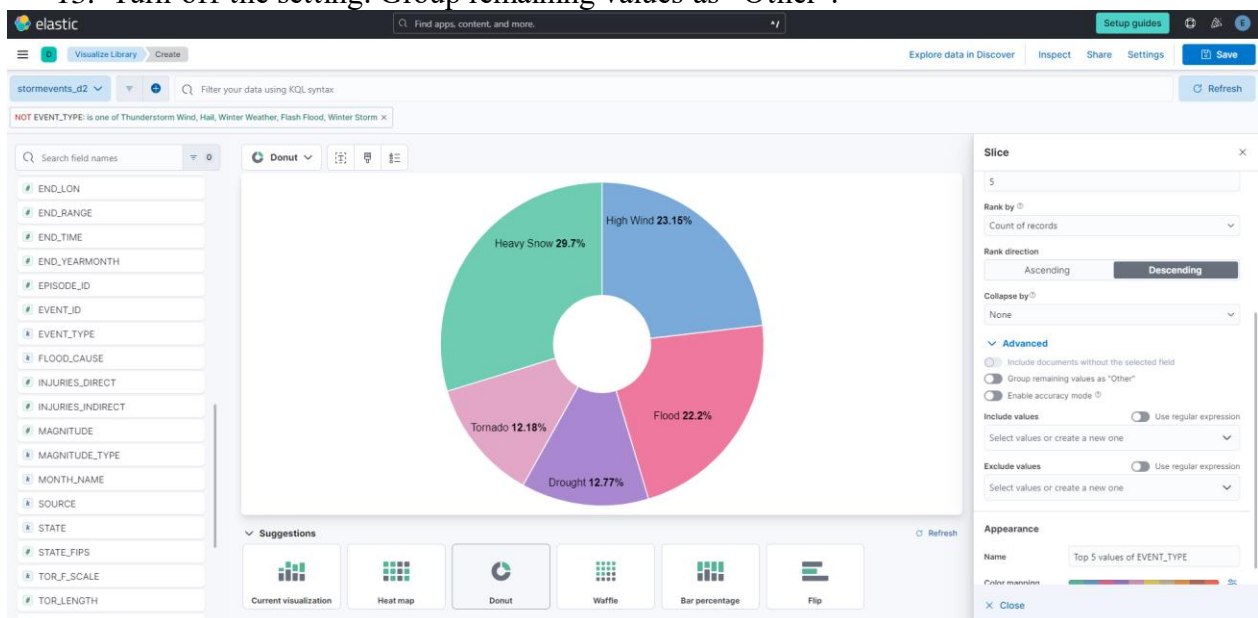
Top 5 values of EVENT_TYPE

Color mapping

×

Close

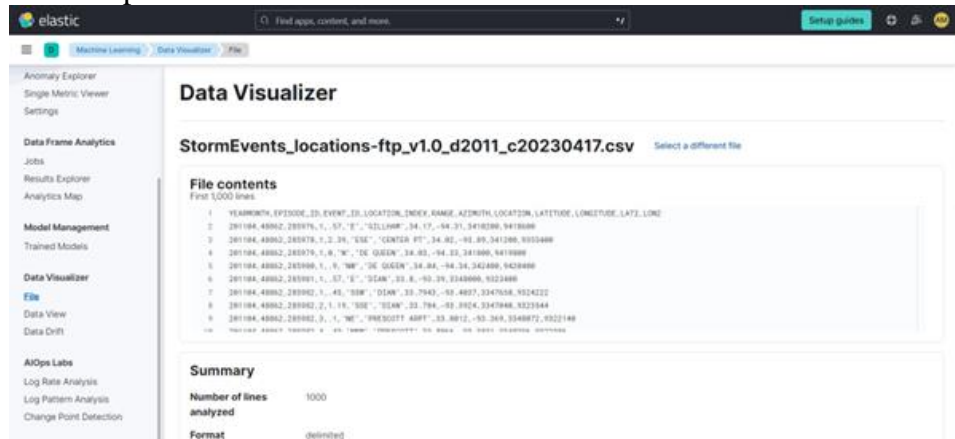
13. Turn off the setting: Group remaining values as "Other".



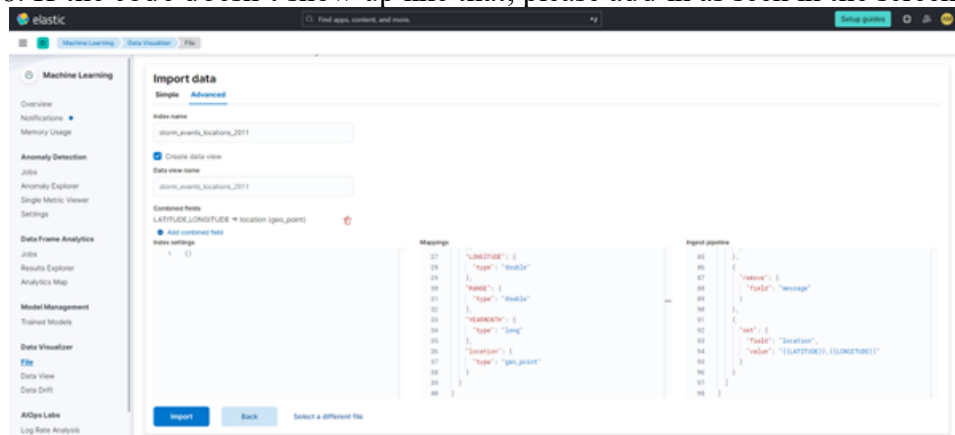
14. Save the Visualization as TOP 5 TYPES of STORM EVENTS. Save and exit.

Uploading the location data sets for 2010 and 2011:

1. Go to: Top left menu -> Analytics -> Machine Learning -> Data Visualizer -> File
2. Then upload the downloaded Storm Events Locations files.



3. Check the indexing and mapping of the uploaded files.
4. Now check if latitude, longitude and location is indexed and mapped right in the advanced view of the imported data.
5. The lines need to show up for “latitude” and “longitude” as shown in the screenshot on the Mappings window on line 27. Also “location” needs to show up with the type of “geo_point” as in line 36 and 37 of the Mappings window.
6. Additionally, in the Ingest pipeline window, the code needs to be shown as in line 92 to 94.
7. Set needs to have “location” for the field type and “{{LATITUDE}}, {{LONGITUDE}}” for value type.
8. If the code doesn’t show up like that, please add in as seen in the screenshot.



9. Check if the map values show up as intended
10. After the upload is finished, scroll down and have a look at the data structure of the uploaded file, to check if latitude, longitude and location is indexed and mapped correctly as in the screenshot.

Type	Name	Documents (%)	Distinct values	Distributions
keyword	AZIMUTH	999 (100%)	16	min, median, max
keyword	EPISODE_ID	999 (100%)	72	min, median, max
keyword	EVENT_ID	999 (100%)	604	min, median, max
keyword	LAT2	999 (100%)	851	min, median, max
keyword	LATITUDE	999 (100%)	851	min, median, max
keyword	LOCATION	999 (100%)	626	min, median, max
keyword	LOCATION_INDEX	999 (100%)	8	min, median, max
keyword	LONG2	999 (100%)	897	min, median, max
keyword	LONGITUDE	999 (100%)	897	min, median, max
keyword	RANGE	999 (100%)	338	min, median, max
keyword	YEARMONTH	999 (100%)	4	min, median, max

11. Check if the values show up in Kibana Data View

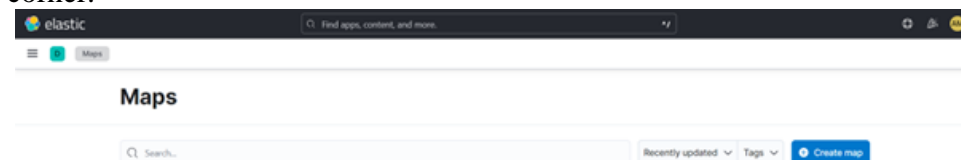
12. Go to: Top left menu -> Stack Management -> Kibana -> Data Views -> select your file

13. Check if the values that got mapped and indexed before show up correctly in the data view.

Name	Type	Format	Searchable	Aggregatable	Excluded
AZIMUTH	keyword		✓	✓	✗
EPISODE_ID	keyword		✓	✓	✗
EVENT_ID	keyword		✓	✓	✗
LAT2	keyword		✓	✓	✗
LATITUDE	double		✓	✓	✗
LOCATION	keyword		✓	✓	✗
LOCATION_INDEX	keyword		✓	✓	✗
LONG2	keyword		✓	✓	✗
LONGITUDE	double		✓	✓	✗
RANGE	double		✓	✓	✗

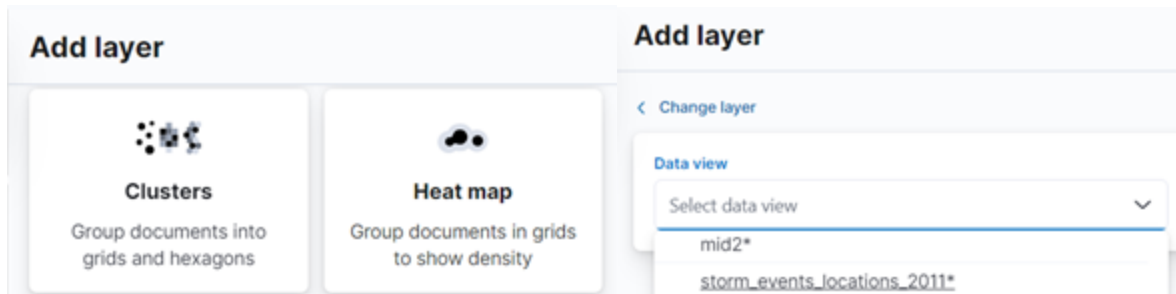
Creating the Storm Events Locations map:

1. Now go to Elastic main page -> Analytics -> Maps, and click on create map in the top right corner.

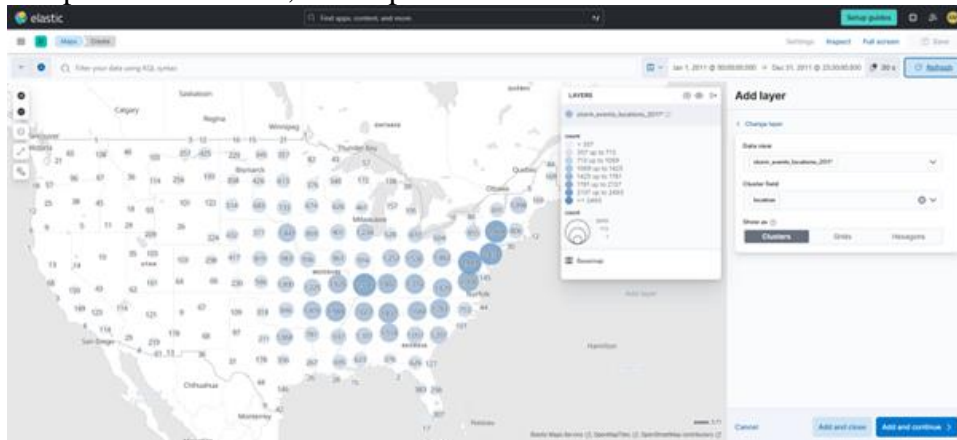


2. Then in the new view, click on “add layer” at the right side, scroll down and select clusters.

3. For the data view now select the uploaded mapped and indexed locations file.



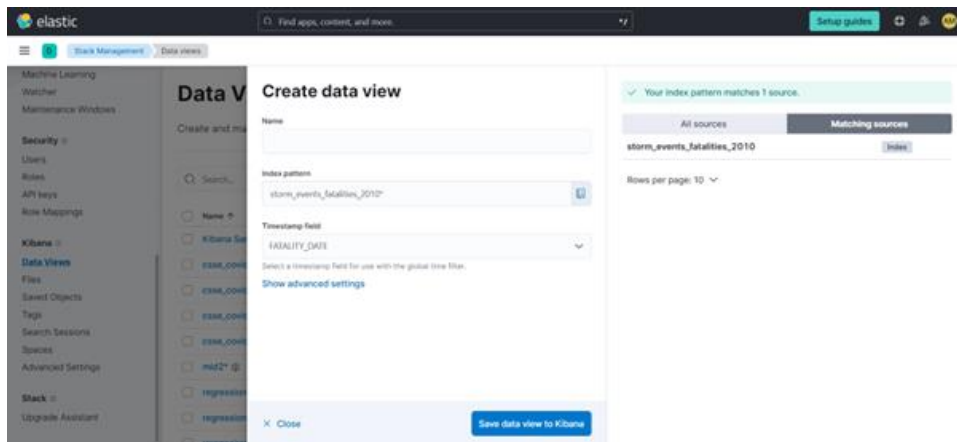
4. Now select “location” for cluster field and adjust the time span on the top right corner.
5. Select January 1st with a time of 00:00 am to December 31st with a time of 23:30 pm and either for the whole year of 2010 or 2011 depending on the file that you working with right now.
6. Update and refresh, the map should look like shown in the screenshot now. Save the view.



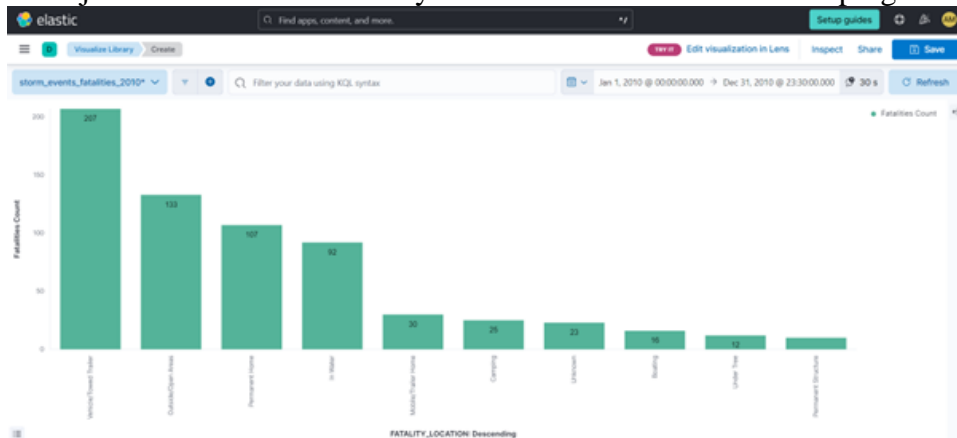
7. Save the map and exit.
8. Repeat the steps above to have the maps for both years.

Creating the storm events fatalities data view:

1. Upload the data sets for the fatalities as described in earlier sections.
2. Create a visualization as described in earlier sections.
3. Go to: Top left menu -> Stack Management -> Kibana -> Data Views -> Create Data View



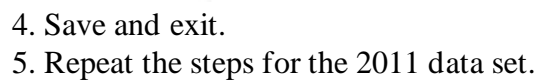
4. Create a bar chart in the same way as you already before in the visualizations section, but this time use the storm_events_fatalities_2010* data set that you uploaded.
5. Select Bar Vertical.
6. Select descending order.
7. Select show numbers in bar.
8. Adjust the date to the whole year with absolute dates in the top right corner.



9. Save the view and exit.
10. Repeat the steps for the 2011 data set.

Creating the storm events fatalities discovery view:

1. Go to: Elastic main page -> analytics -> discover
2. Select the storm_events_fatalities_2010* dataset that you uploaded earlier.
3. Adjust the date to the whole year with absolute dates in the top right corner.



1. URL of Data Source, <https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/>

-StormEvents_details-ftp_v1.0_d2011_c20230417.csv.gz

-StormEvents fatalities-ftp v1.0 d2010 c20220425.csv.gz

-StormEvents fatalities-ftp v1.0 d2011 c20230417.csv.gz

-StormEvents locations-ftp v1.0 d2010 c20220425.csv.gz

-StormEvents locations-ftp v1.0 d2011 c20230417.csv.gz

2. URL of your Github: https://github.com/MichaelCIS/3200_Stormevent.git