DAMG7245 Big Data Systems and Intelligent Analytics - Spring 2022 Assignment -1

| **Summary** | Analyzing SEVIR and Storm Events data using Google BigQuery and Google Data Studio. |
| --- | --- |
| **Status** | Published |
| **Category** | Web |
| **Environment** | web, kiosk |
| **Author** | Aditi Krishna, Abhishek Jaiswal, Sushrut Prashant Mujumdar |

# **Part I:**

Analysis of CATALOG.csv (the metadata file for SEVIR dataset) and Storm Events datasets for the years 2018 and 2019.

## **Steps for analyzing SEVIR metadata in Big Query and Data Studio**

1. Download SEVIR Catalog dataset and Storm Events dataset for 2018 and 2019.
2. In the Google Cloud Console, create a new project in BigQuery.
3. Enable the BigQuery API.
4. Select the project name on the left side of the UI, then create a new dataset.
5. Create a new table within the newly created dataset

* The following widget allows uploading files from your local system, Google Drive, and Google Cloud Storage.
* Since we already have the CSV files for our project we use the upload functionality. It allows uploading files up to 100MB of size.

1. Name the table and choose a schema for the table

* BigQuery offers a functionality to auto-detect the schema. However, it can also be done manually.

1. A load job is then created

* Once the loading is complete we can view the table details where you can check the schema, details, and preview the dataset.

1. In the query editor, add a query that you wish to run on the table.

## **Creating a data view in BigQuery for Analysis**

1. Loading data directly to Google Data Studio for analysis has certain limitations. Data blending in Data Studio uses left outer join.
2. As per our use case, we need to combine Storm-Event details files for the year 2018 and 2019. Since Google Data Studio does not allow union operation we perform this union in BigQuery. Run the following query in the query editor.

| SELECT EPISODE\_ID, EVENT\_ID, STATE, YEAR, MONTH\_NAME, EVENT\_TYPE, CZ\_NAME, CZ\_TIMEZONE, INJURIES\_DIRECT, DEATHS\_DIRECT, DAMAGE\_CROPS, DAMAGE\_PROPERTY, MAGNITUDE, BEGIN\_LAT, BEGIN\_LON, STATE\_FIPS  FROM `assignment-1-340501.storm2018.storm\_details2018`  UNION ALL  SELECT EPISODE\_ID, EVENT\_ID, STATE, YEAR, MONTH\_NAME, EVENT\_TYPE, CZ\_NAME, CZ\_TIMEZONE, INJURIES\_DIRECT, DEATHS\_DIRECT, DAMAGE\_CROPS, DAMAGE\_PROPERTY, MAGNITUDE, BEGIN\_LAT, BEGIN\_LON, STATE\_FIPS  FROM `assignment-1-340501.storm2019.storm\_details2019` |
| --- |

1. Now that we have a combined view of the Storm-Event dataset for2018 and 2019, we perform a left join on the new table with SEVIR's metadata file. Run the following query in the query editor on BigQuery.

| SELECT E.EPISODE\_ID, E.EVENT\_ID, E.STATE, E.YEAR, E.MONTH\_NAME, E.EVENT\_TYPE, E.CZ\_NAME, E.CZ\_TIMEZONE, E.INJURIES\_DIRECT, E.DEATHS\_DIRECT, E.DAMAGE\_CROPS, E.DAMAGE\_PROPERTY, E.MAGNITUDE, E.BEGIN\_LAT, E.BEGIN\_LON, E.STATE\_FIPS, S.id, S.file\_name, S.file\_index, S.img\_type  FROM `assignment-1-340501.storm\_union.storm\_details\_all` as E LEFT JOIN `assignment-1-340501.catalog.sevir-catalog` as S ON E.EVENT\_ID = S.event\_id |
| --- |

1. Save the above view into a new table, this will be added as a source to Google Data Studio for analysis.

## 

## 

## **Adding a data source and creating report using Google Data Studio and BigQuery connector**

1. Open Google Data Studio. Click on the Blank Report template to create a new report.
2. Click on Add data on the toolbar on top of the window. Select BigQuery from Add data to report popup.
3. From My Projects, select the name of your project under which you are creating the report. Then select the name of the source dataset and the table name. Click on Add to Report. The data source is now ready to be used.
4. Add charts on the report. Using the data tab on the window set dimensions, metrics, filter and other functionalities, for the information you would like to represent on the report.
5. We can also use customer queries for reports. In the data tab select BigQuery as the data source. A window will pop-up where you can select Custom Query as the source. A query editor will open up, run the query and save the view for the chart to utilize.
6. The style tab allows formatting of the charts and their properties.

## **Dashboard**

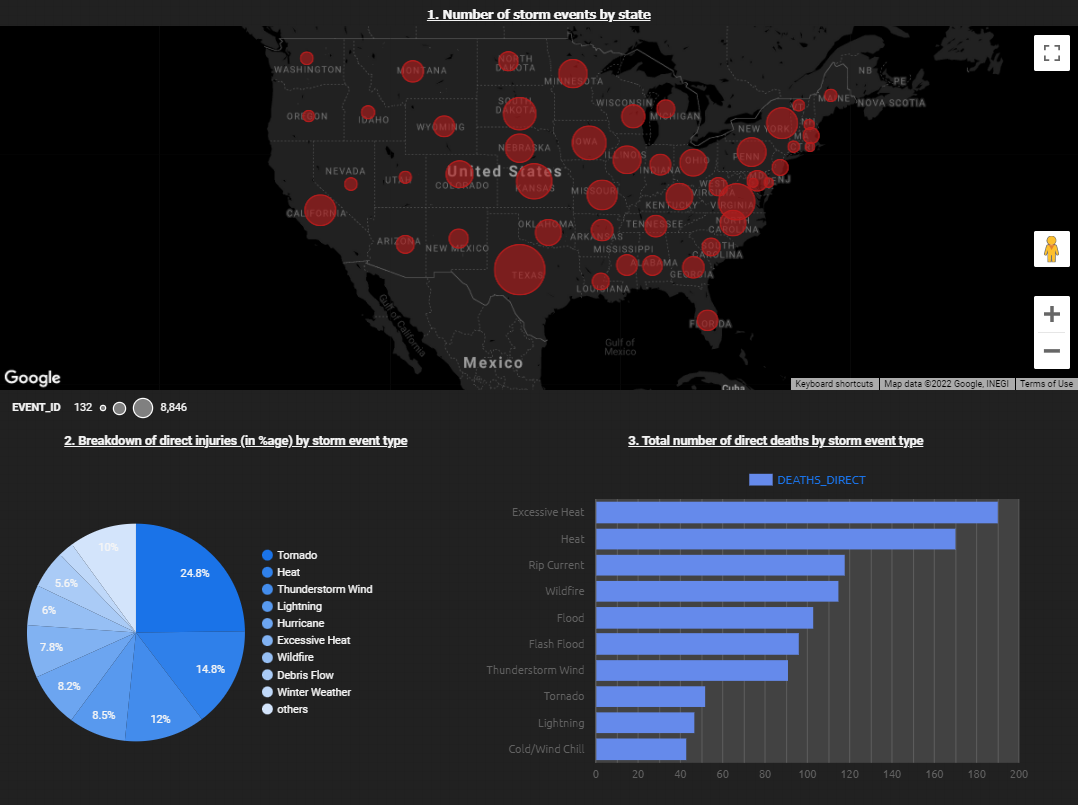
SEVIR and Storm-Event datasets analysis answers the following queries for the years 2018 and 2019.

1. Number of unique storm events (event\_id) by state using a Geo-chart.
2. Breakdown of direct injuries (in percentage) by storm event type.
3. Total number of direct deaths by storm event type.
4. Count of images by image type and storm event type.
5. Highest magnitude of Thunderstorm-wind and Hail-storm by state.

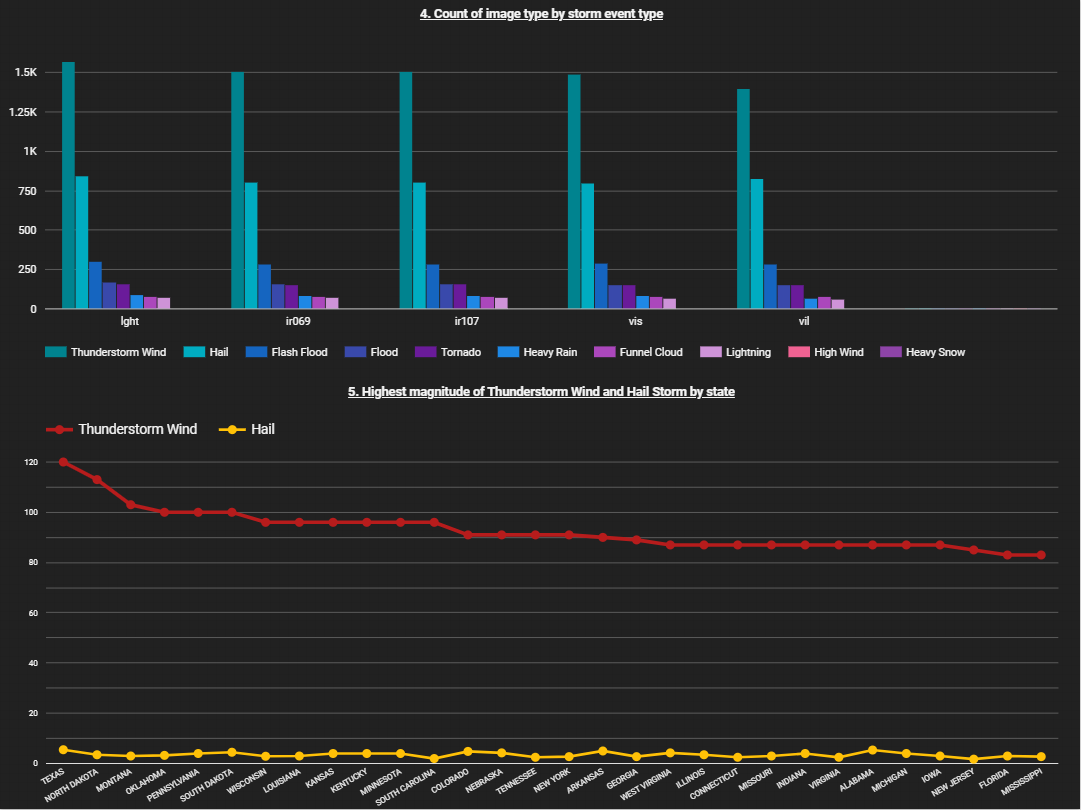
* For this we used a Custom Query

| SELECT STATE, MAGNITUDE, EVENT\_TYPE  FROM `assignment-1-340501.sevir\_storm\_leftjoin.storm\_sevir\_left`  WHERE EVENT\_TYPE IN ('Hail', 'Thunderstorm Wind')  ORDER BY MAGNITUDE DESC |
| --- |

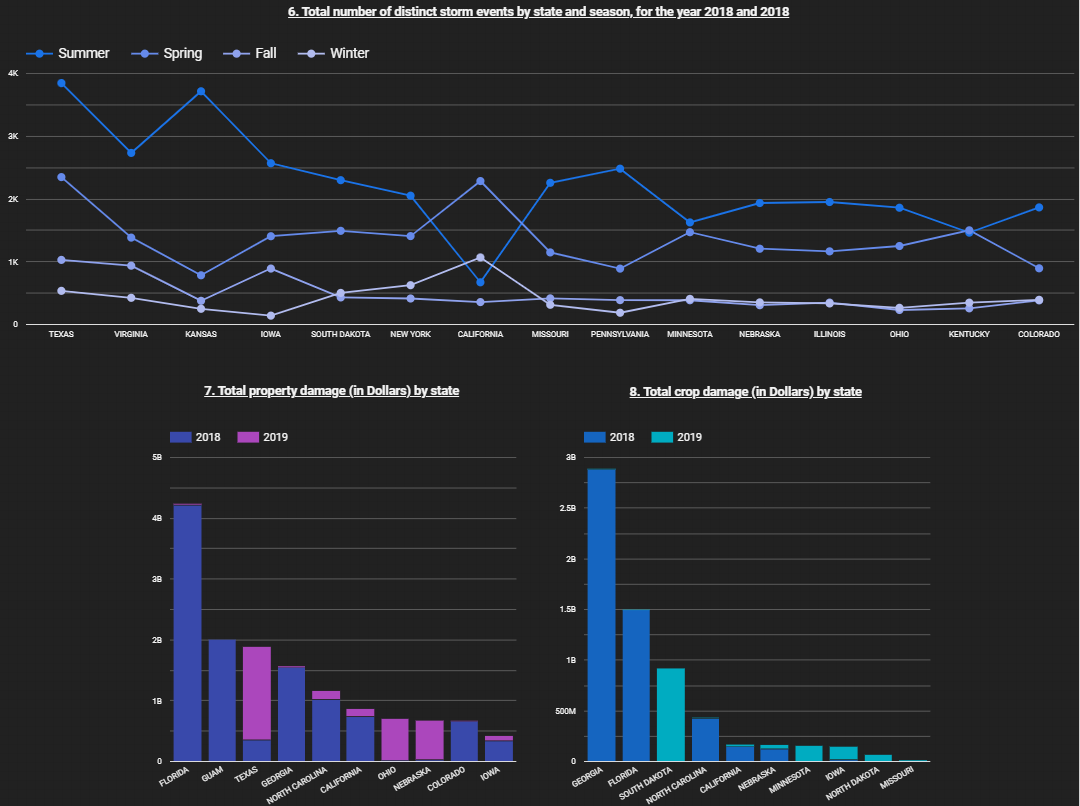
1. Total number of distinct storm events by state and season.
2. Total property damage (in Dollars) by state
3. Total crop damage (in Dollars) by state



**Fig. 1: Dashboard page # 1**



**Fig. 2: Dashboard page # 2**



**Fig. 3: Dashboard page # 3**

# **Part II:**

Storm EVent ImagRy (SEVIR) dataset for EventID: 835047 contains images of storm events captured by satellite and radar. SEVIR is a collection of thousands of "storm events", which are 4-hour sequences of weather recorded by five separate sensors. The dataset provides five sensing modalities.

The data collected by the National Weather Service (NWS) can augment many of the events in SEVIR. This database includes the type of severe weather (high winds, tornado, hail), storm impacts (damage due to crop and properties, injuries, and deaths), and a summary of the event.

The two main components are Catalog and data files. Catalog primarily contains the metadata of the event and data files contain events for a certain sensor in an hierarchical format, where the data is stored as an integer type. Depending on the sensor type, these integers can be decoded into floating type values, which represent the actual values captured by the sensor. Decoding is performed either using linear scaling or an exponential transformation.

# **References**

1. <https://github.com/googlecodelabs/tools>
2. <https://www.ncdc.noaa.gov/stormevents/ftp.jsp>
3. [https://nbviewer.jupyter.org/github/MIT-AI-Accelerator/eie-sevir/blob/master/examples/SEVIR\_Tutorial.ipynb](https://nbviewer.jupyter.org/github/MIT-AI-Accelerator/eie-sevir/blob/master/examples/SE)
4. [Visualizing BigQuery data using Data Studio | Google Cloud](https://cloud.google.com/bigquery/docs/visualize-data-studio)
5. [GitHub - MIT-AI-Accelerator/sevir\_challenges: AI Challenges based on the SEVIR weather dataset](https://github.com/MIT-AI-Accelerator/sevir_challenges)
6. [Loading data into BigQuery](https://www.youtube.com/watch?v=Abzj-Vyhi74&ab_channel=GoogleCloudTech)