Analysis of Air Quality

Thien Hua, Leon Lin, Matthew Manuel, Ramon Santos

CIS 4560

California State University, Los Angeles

**Abstract:** The data set we have chosen for our project concerns information from Airsofia.info, a website that records data based on air quality trends meant to predict future patterns.. Using AWS’ service, we will upload this data to the class’ Hadoop server and run various queries and views to provide a proper analysis and breakdown of the data with Apache HiveQL. This can help process data based on trends shown in the data set, which can lead to different ways in how it can be applied practically.

**1. Introduction**

This project is an analysis of air quality data obtained from <https://airsofia.info/>, a website designed to record this type of information. This website was created after the country of Bulgaria was taken to task by the European Court of Justice for failing to implement measures against air pollution in the country. The dataset we obtained contains a sample of information taken from the site in 2018, and provides a breakdown of air quality within the city of Sofia to determine the root causes of its pollution and how this can be remedied if possible.

**2. Background**

In 2013, the city of Sofia was singled out in a study conducted by the European Environment Agency as being one of the most polluted countries in the European Union, because of the presence of several major air pollutants in the atmosphere. This is mainly because of the high quantities of particulate matter like gas and carbon monoxide, many of which are related to Bulgarians’ continued reliance on traditional energy methods such as burning wood and coal.

Sofia’s air pollution is contributed to many natural factors such as geographic location and winter seasons creating fog and thermal inversion, with the effect of stagnation of air (Martino). The main of Sofia air pollution is due to human activity. There are over 1,2000,000 people and 550,600 cars with most cars still using diesel fuel causing air pollution (Martino). Another contribution is heating systems used in houses and apartments during the Winter season. There is controversy about whether transportation or heating systems in homes is the main cause of air pollution in Sofia.

For many years air pollution was ignored by the European governments, but recently the people, environmental, and health groups have been suing their government for failure to comply with air quality laws (Taddei). Bulgaria has a major air pollution problem and has some of the dirtiest air in Europe. As of April 2017, the Court of Justice of the European Union (CJEU) declared that Bulgaria failed to fulfil its obligations under European Union Law to reduce the level of particulate matter in the air below the limits (Taddei).

Two groundbreaking decisions known as the Jnecek and ClientEarth cases could point Bulgaria towards cleaner air. These decisions will allow rights such as the right of standing, the right to a legal remedy, and the right to a substantive review. ClientEarth is relying on the CJEU presences on pursuing a substantial amount of clean air cases (Taddei). Towards the end of 2017 municipal authorities created an, “Air, Climate and Energy” directorate, but has yet to be approved. Sofia and other Bulgarian cities immediate solutions would be upgrading heating systems in homes and using energy efficient vehicles.

The European Union has on of the poorest air quality with an estimated 467,000 premature deaths per year (Wooden). Big data is helping win the battle against air pollution by using sensor technology to collect real time information about air quality. By using modern data analytics to create useful information about the data gathered by the sensors such as pollutant levels and other various information (Wooden).

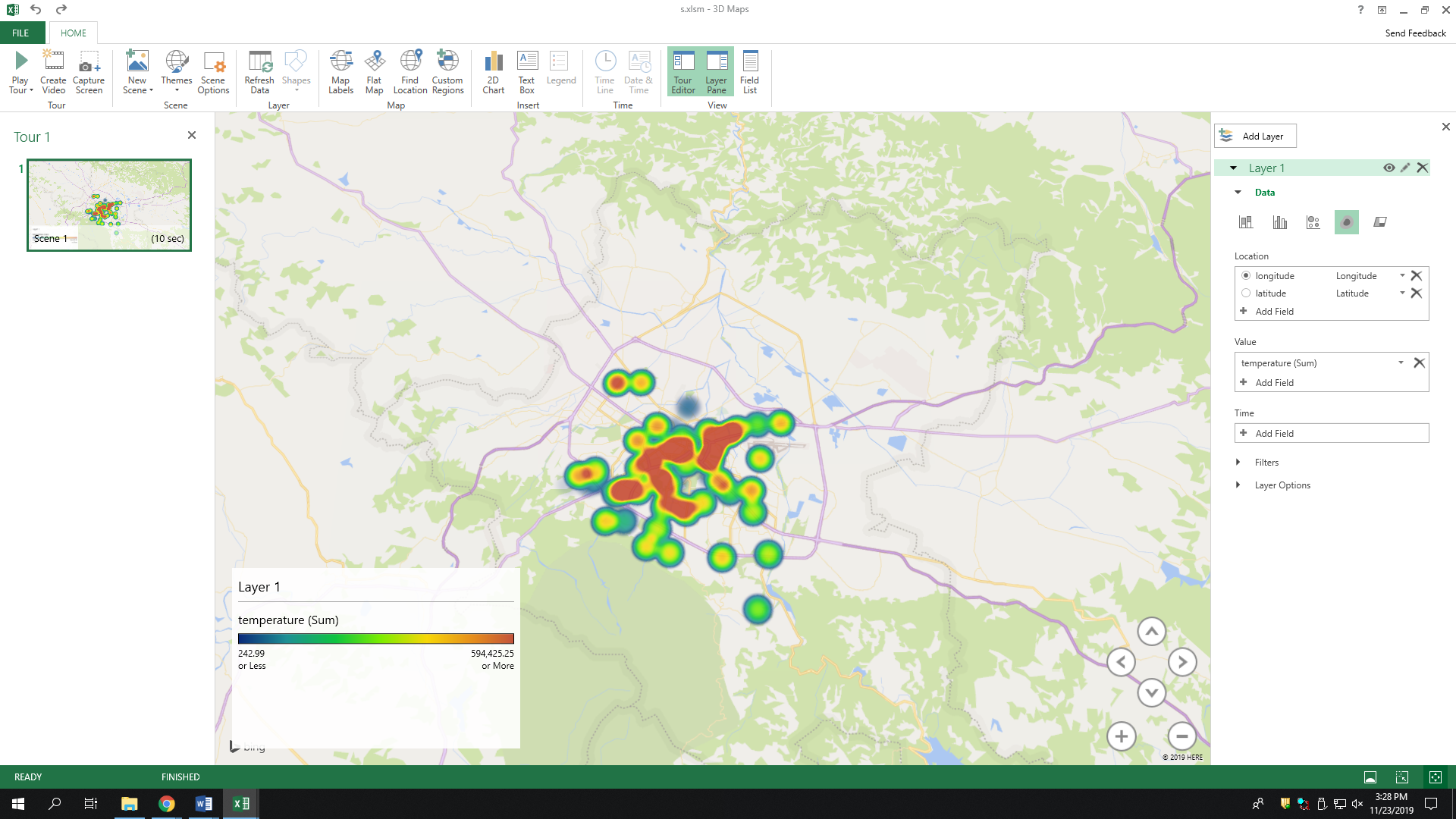
The Future Cities Catapult, which is a United Kingdom backed company, partnered with Intel Collaborative Research Institute, The Royal Parks, The London Borough of Enfield, ScienceScope, and City Insights to create the Sensing London Project. This project is to collect and use real-life city data to observe the impact cities have on human health and the natural environment (Wooden). The Sensing London Project established 5 laboratories all over the UK with sensors measuring local air quality and human activity (Wooden). Intel is leading the way in creating technology to gather information readily available to monitor air pollution.

**2.1 Related Work**

The first major related work comes from the study conducted by the European Environment Agency in 2013, which was cited in the above New York Times article. In it, the agency traced down several causes behind air pollution throughout the continent, and isolated several cities of note that violated certain standards because of their worsening conditions. The second major related work is from the World Health Organization in 2016, where it did a study on ambient air pollution. Both studies focused on finding underlying factors behind the rampant spread of air pollution, including the presence of particulate matter, carbon dioxide and even geographic factors that may influence the climates of those affected regions. Our work utilized big data obtained from a data source from an agency that tracks down air quality. We did this study with a focus on comparing temperatures of a specific region and noting the changes in values over time.

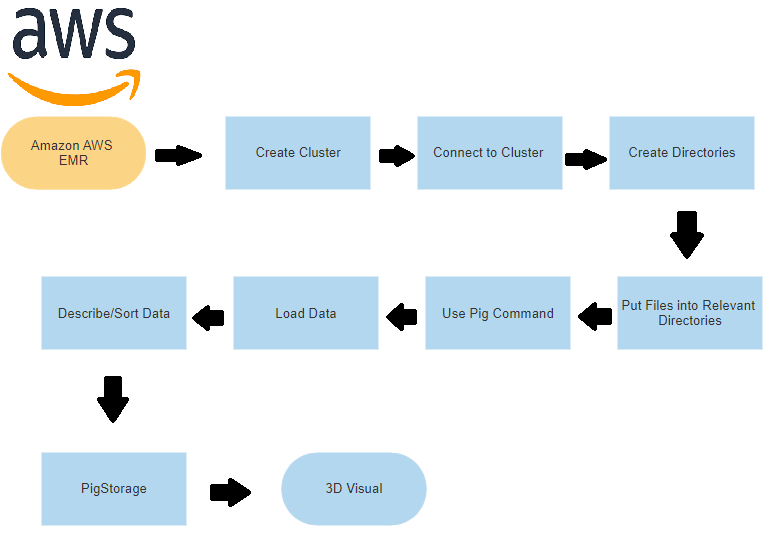
**2.3 Our Work**

For our work, we took the dataset from Kaggle which visualized the air quality of Sofia and other Bulgarian cities. The data was broken down into coordinates that picked apart sections of the country, and the resulting air quality statistics related to them, like temperature, humidity and toxicity. This was later visualized into a heat map that broke down the data into areas based on their air quality, with redder territories being excessively polluted. With SQL queries, we can select specific information based on these factors and isolate them.

****

**(Figure 1: Visualization of temperatures in a specified location due to air quality)**

The main difference our study has from the other related work here is that it focuses mainly on Bulgaria as a country, rather than the entire European continent. This reflects on our smaller dataset, limiting our scope and focus to that region instead. Also, in direct comparison to the other datasets stated above, the information here was polled from 2017 onwards, making it newer by comparison, although by no means meant to reflect the actual data that can be obtained in this current point in time.

****

**(Figure 2: Flowchart of project)**

**3. Conclusion**

With the data presented, we can conclude that Sofia’s air quality has only slightly improved over the course of time. The high presence of contributors to their pollution problem still persist thanks to the city’s continued reliance on older and less environmentally friendly methods of sustaining energy to power their machines. This can be inferred from the queries run on our dataset and the resulting visualization, which could only be collected and properly analyzed with the proper big data techniques and methodologies.

Despite our findings, we believe there can still be more improvements done to this framework. Data sets like these could implement more fields to further clarify the information presented, and could also be updated to be more modern such that the conclusions we can come to are more timely.

**4. Submission Process**

1. Format your paper using this template.

2. Turn the hardcopy by Dec 4th before the lecture starts

### **References**

[1] T.A. Jones, “Writing a good paper,” *IEEE Trans. on General Writing*, Vol. 1, no. 2, pp.1-10, May 2002.

[2] K. Hwang, *Computer Arithmetic*, John Wiley, 1997.