Project Report

RADIATION & NUCLEAR DATA ANALYSIS

Yashashwini Gupta(yg1568)
Udita Gupta(ung200)
Hitarthi Shah(hus206)
Mohammed Ilyas Habeeb(mih278)

December 15, 2017

Instructor: Professor Juan Rodriguez

ABSTRACT

Radiation and its environmental health effects are issues which are fraught with deep-seated controversy. Following the meltdown of the Fukushima Daiichi Nuclear Power Plant in 2011, the amount of radiation water is growing 150 tons per day. It has become imperative to analyze the radiation levels and its correlation between the types of reactors, their location, population exposed, etc.

This project examines nuclear plant and radiation level data for the years 1990 till 2017. We attempt to find a correlation between the radiation levels, population, type of reactor, and the nuclear power plant's location (like Inland near a river), etc. We also analyzed:
a) the total power generated per country by reactors,
b) population exposed at a particular distance, and
c) yearwise radiation level change.

We have analyzed the data with the help of technologies like PySpark and Spark MLib. In this report, we have addressed the steps that we performed to analyze the data including challenges that we faced while doing that.

ACKNOWLEDGEMENT

We owe a debt of deepest gratitude to our professor Mr. Juan C Rodriguez, Department of Computer Science for his constant guidance, support, motivation, and encouragement throughout the period of this work carried out. His readiness for consultation at all times, his educative comments, his concern and assistance even with practical things have been invaluable. Mr. Juan C Rodriguez has been a great source of inspiration for us and we thank him from the bottom of our hearts. We are grateful to him for providing us with the necessary opportunities and for the completion of our project. We also thank the other staff members for their invaluable help and guidance.

Not to forget, we also owe our Teaching Assistants, Prashant Patel and Sneha Arkery, for their continuous support and helping us whenever required.

–

Hitarthi Shah
Ilyas Habeeb
Yashashwini Gupta
Udita Gupta

CONTENTS

1 Introduction

Radiation and its environmental and health effects are issues which are fraught with deep-seated controversy. Unfortunately, it has been difficult until now to find radiation data which truly has been free of bias, or of the perception of bias in favor of one ideological position or another.

As of April 2017, 30 countries worldwide are operating 449 nuclear reactors for electricity generation and 60 new nuclear plants are under construction in 15 countries. In this project, we use big data analytics to identify the correlations among population exposed, number of reactors, types of reactors etc.

2 Methodology

2.1 Data Capture

DataSet List:

1. Population Exposure Estimates in Proximity to Nuclear Power Plants, Locations.
2. Population Exposure Estimates in Proximity to Nuclear Power Locations, Reactors.
3. Radiation Level Data(~10 GB) (~ 70 million observations)

Data was obtained from the Socioeconomic Data and Applications Center (sedac): A Data Center in NASA's Earth Observing System Data and Information System (EOSDIS). The Data gives the overview of the nuclear power plants and radiation levels around them.

Links for each of these datasets:
http://sedac.ciesin.columbia.edu/data/set/energy-pop-exposure-nuclear-plants-locations
https://api.safecast.org/
https://www.epa.gov/enviro/data-downloads

NASA has published the open source data to the people to explore further and analyze their data. The data includes the below fields:

- Power Plant Locations
- Number of Reactors
- Type of Reactors
- Population exposed at 1200kms
- Population exposed at 600kms
- Population exposed at 300kms
- Population exposed at 150kms
- Population exposed at 30kms
- Total Power Generated by each reactor
- Power Generated Yearwise

The safecast API provided us with the radiation levels across the world from 1970 to the present day. The data was collected and cleaned using PySpark. It includes the fields:

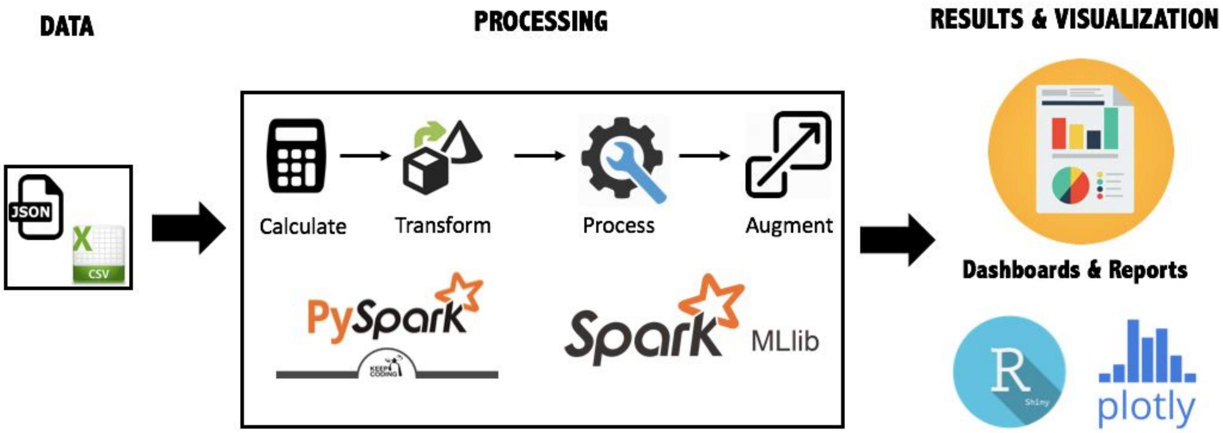
- Radiation value
- Measurement Unit
- Latitude and Longitude of location
- Timestamp of when it was recorded
- Device id

2.2 Tech Stack

We implemented the project using the following technologies:

- PySpark
- SparkML
- Python Matplot lib
- RShiny
- Python pandas
- Safecast API

The following figure shows the pipeling and architecture of the final system.



2.3 Data Cleaning

The radiation data obtained from SafeCast API was about 10 GB of data and required a lot of data preprocessing. Since the dataset was collected from a variety of users and places, there was a lot of discrepancy in the data. The process of data cleaning was hence, very tedious and involved several steps:

1. **Collecting the data**: The data was collected from the API was in a json format and had to be transformed to a dataframe.

2. **Changing column names**: All column names has certain spaces in them, which could potentially create problems for us in the future when we would have analyzed the data. Just removing them would not make any sense, so they were replaced by “ ”.

3. **Changing data formats**: Since there were multiple files in the data set that we obtained, we found that there are three different formats that are used for representing data. Again, while implementing, we could have faced problems if the data was not converted to a particular format.

4. **Removing NA values**: NA values are critical since they can affect the analysis of our data, depending upon the number of NAs that are present in the data set. In our case, NAs were present in values and unit. So had to filter out those column to avoid manipulating the results.

5. **Changing unit values**: The dataset contained different notations for the same unit of measurement. For example, Microsievert was represents as ‘uSv/hr’, ‘usv/hr’, ‘‘uSv/h’, ‘uSv’ and ‘microsievert’. We had to identify the same units of measurement and then finally represented the units as ‘usv’.

6. **Changing unit format**: The dataset contained the radiation in various unit such as Microsievert and Counts per minute(cpm). This was due to different standard in different countries to measure the level. We used a User Defined function to convert the values to cpm and then dropped the unit column to save time in processing and transformation.

7. **Parsed timestamp to extract data**: We had to parse the timestamp in the dataset so as to extract the year and month of the row for further analysis.

8. **Filtered by year values**: On extraction we found the dataset to contain values for the years before 1970 which were removed.

9. **User and device id columns**: The radiation dataset contained two other columns Username and device id columns which were not required in our analysis and were dropped.

2.4 Data Analysis

Since we have obtained the complete data in section 2.3, we can go ahead and start performing some analytics on the data. Some of the major analytics that we performed are:

1. Operational Status of Reactors currently all over the world.
2. Reactor locations all over the world.
3. Number of Reactors built per year

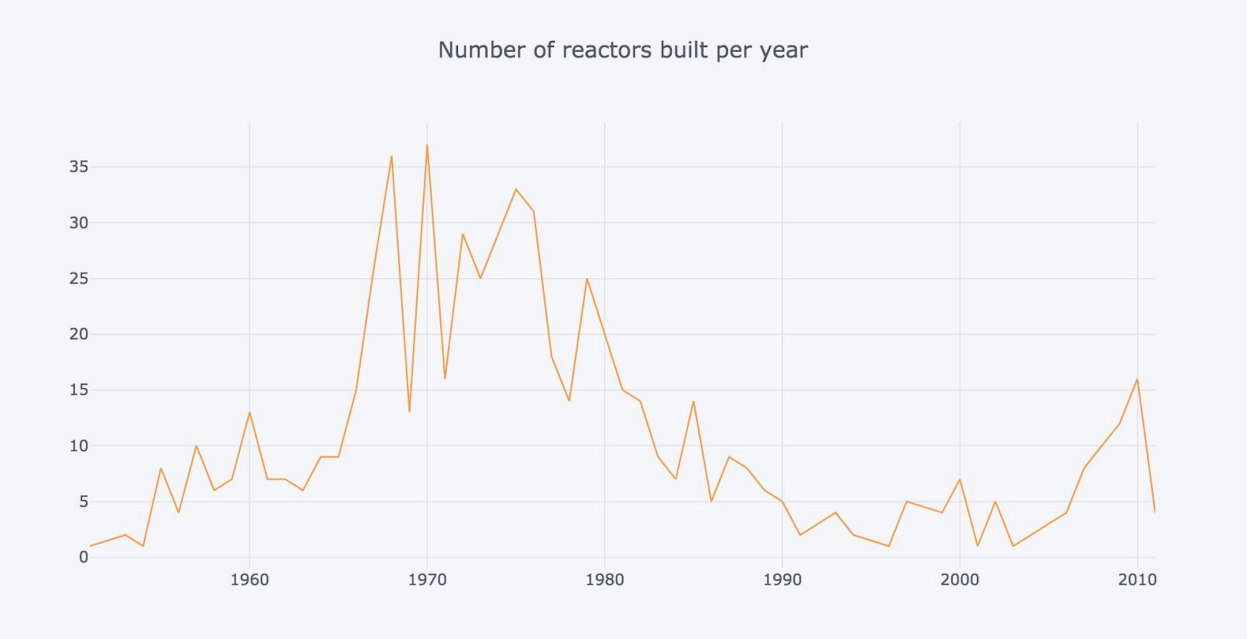
4. Power generated per year all over the world.
5. Power generated Countrywise and Reactor Type-wise.
6. A correlation between Type of nuclear power plant and the power generated.
7. An interactive R-shiny Dashboard where we show the population exposed to radiation near a certain nuclear power plant.
8. An interactive R-shiny Dashboard where we plot all the nuclear power plant locations.
9. A machine learning model, using SparkML which predicts the radiation level for the coming months based on features like year, month, latitude, longitude etc.

The model used was Linear Regression with Radiation Level as the label and Year, Month, Latitude, Longitude etc as the features.

3 Results and Discussion

3.1 Results obtained with plotly and matplotlib-

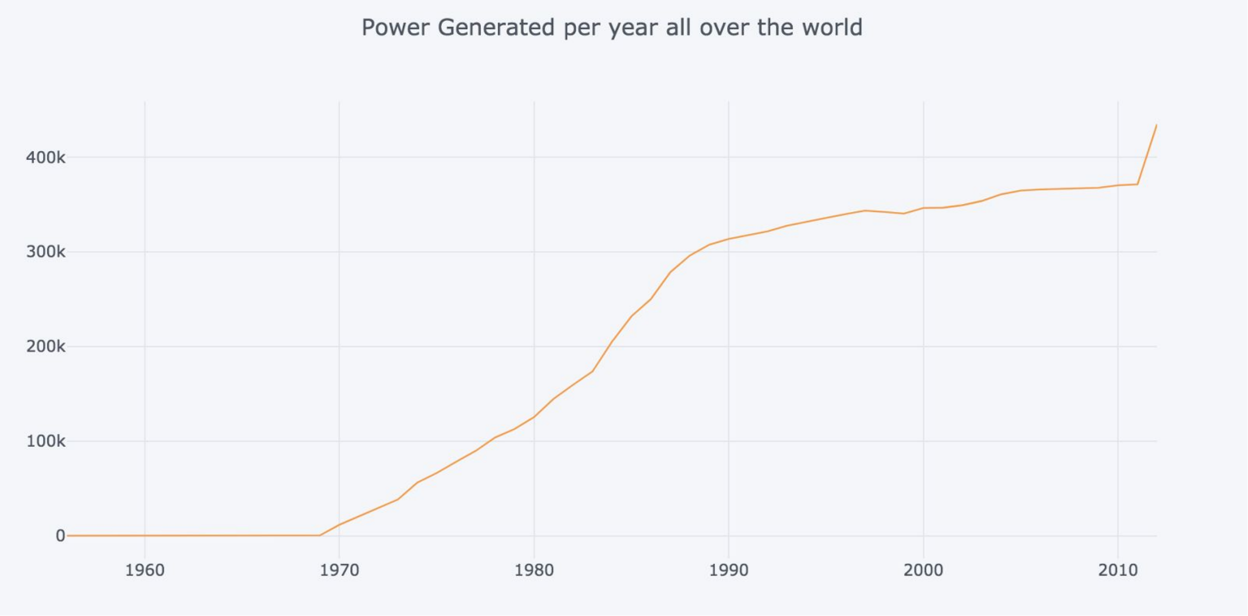
3.1.1 Reactors build per year



The figure above shows that from the late 1970s to about 2002, the nuclear power industry suffered some decline and stagnation. Few new reactors were ordered, the number coming on line from mid 1980s little more than matched retirements, though capacity increased by nearly one third and output increased 60% due to capacity plus improved load factors.

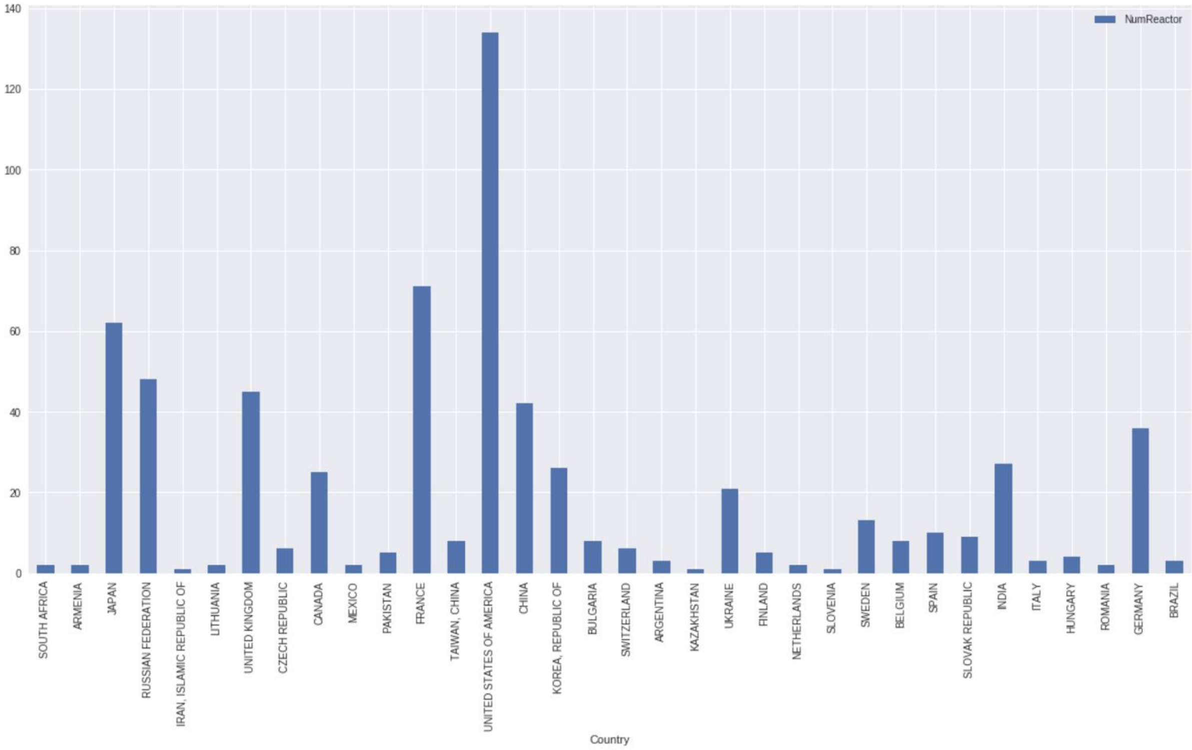
In the new century several factors have combined to revive the prospects for nuclear power. First is realisation of the scale of projected increased electricity demand worldwide, but particularly in rapidly-developing countries. Secondly is awareness of the importance of energy security, and thirdly is the need to limit carbon emissions due to concern about global warming. And last, the availability of a new generation of nuclear power reactors

3.1.2 Power generated using Nuclear Reactors



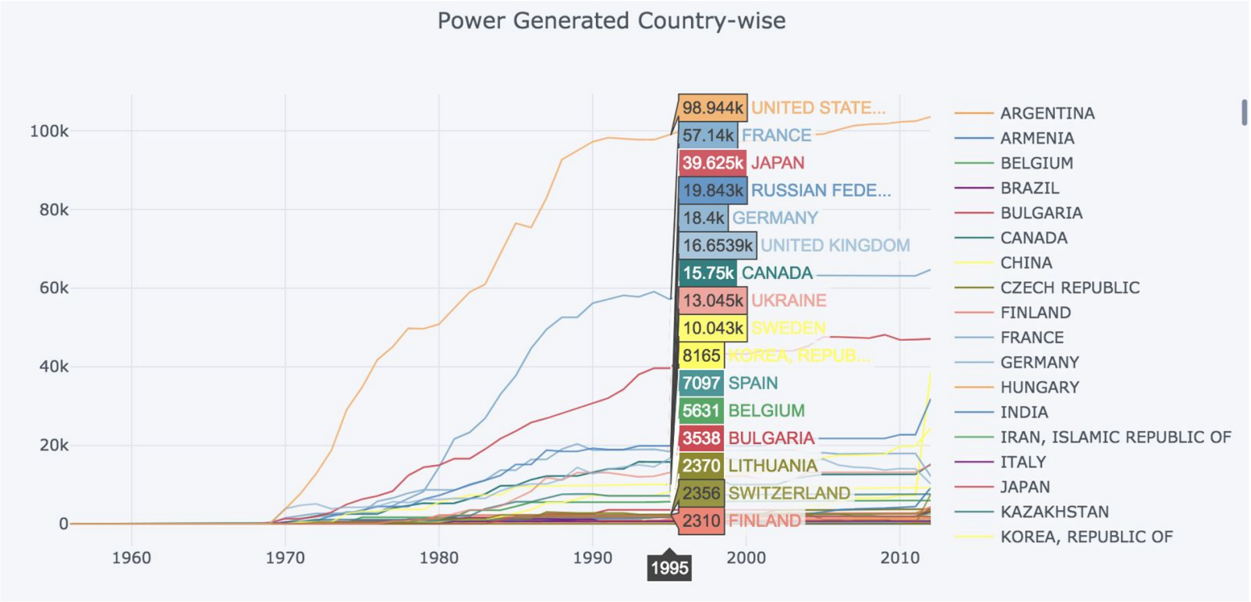
The figure above shows that there was a sudden rise in the power generation since 1970, this coincides with the drop in the Uranium prices.

3.1.3 Total Reactors in each country



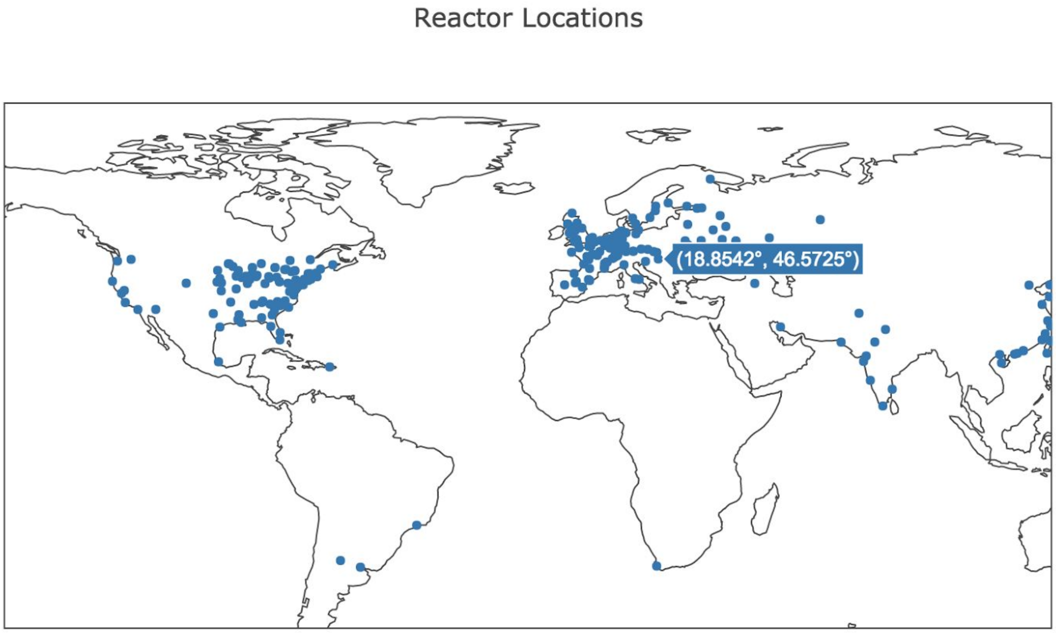
The figure above shows the Total Number of Reactors in each Country. It shows that United States of America has the highest number of reactors and France the second highest.

3.1.4 Power generated in each Country



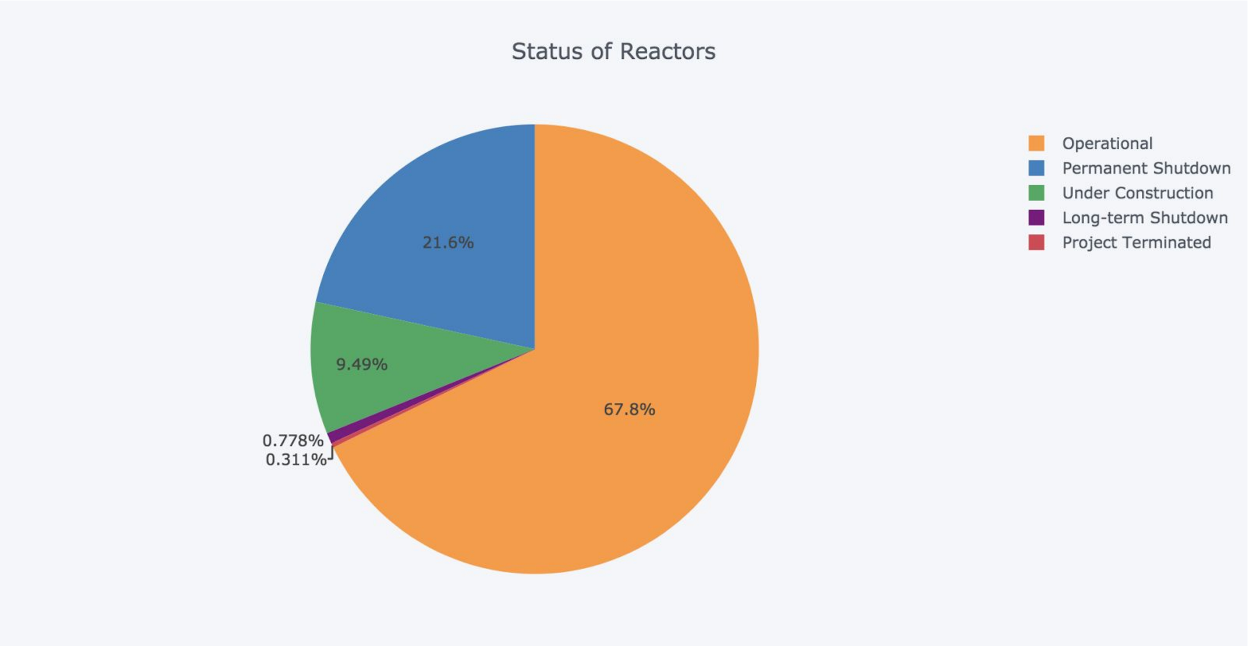
The figure above shows total power generated over the years in each Country with United States of America, France and Japan as the front-runners.

3.1.5 Location of Reactors



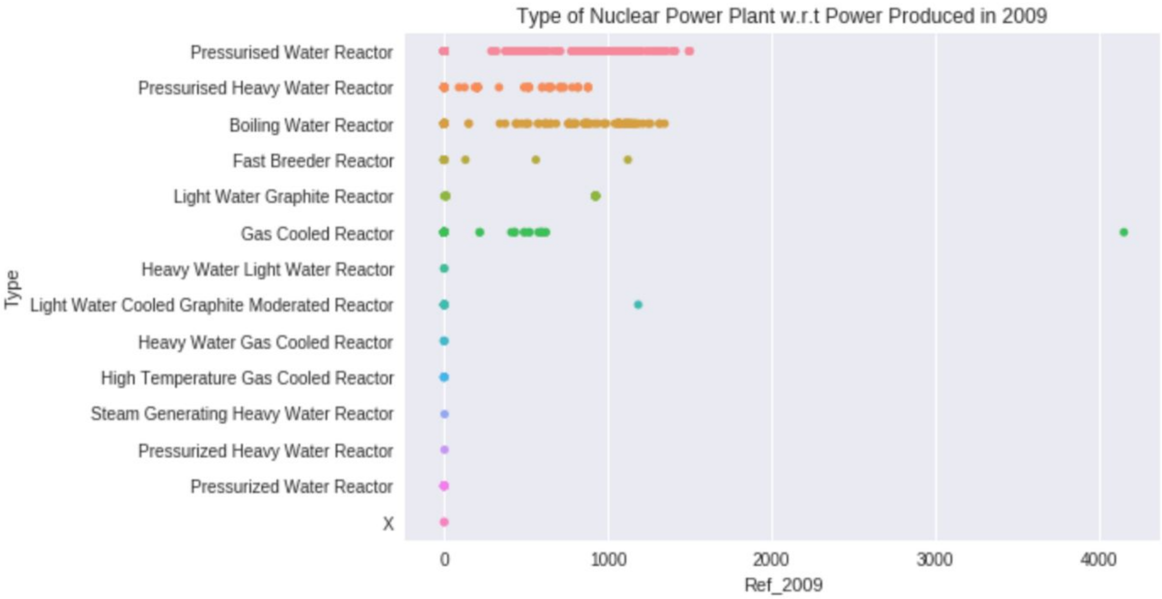
This plot shows the distribution of the reactors across the globe.

3.1.6 Status of Reactors



The figure above shows the status of the various Nuclear Plants and whether they are operational.

3.1.7 Type of Nuclear Power Plants w.r.t. Power Produced

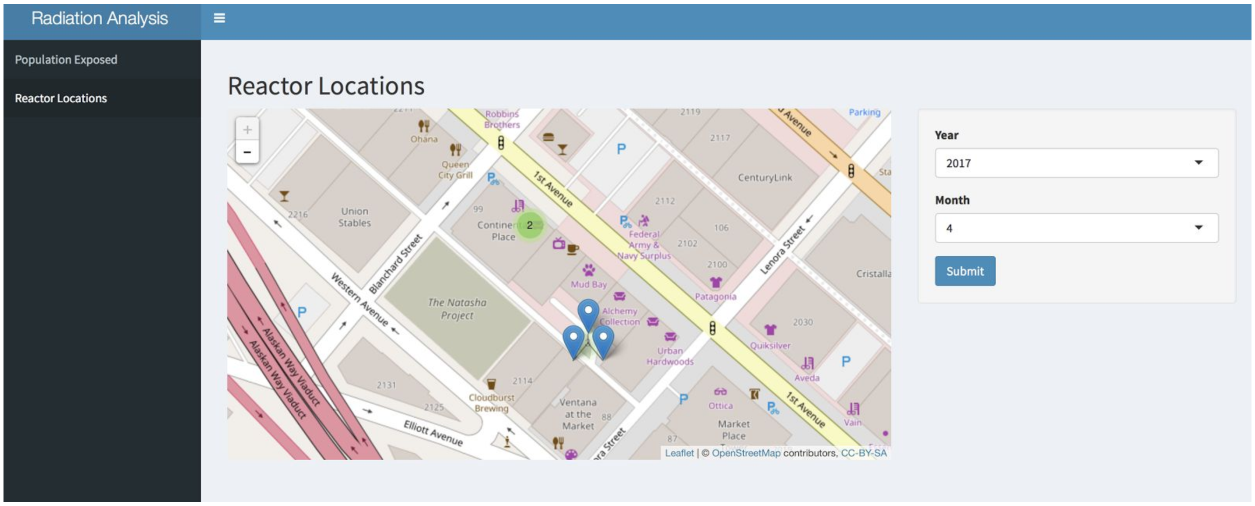


The figure above shows the type of the Nuclear power plant and the respective power produced in the year 2009. It shows that Pressurized Water Reactor, Pressurized Heavy Water Reactor and Boiling Water Reactor produce the

3.2 Results obtained with RShiny

Dashboard to observe the population exposed to radiation and the reactor locations. We designed two dashboards to explore the data further.

3.2.1 Reactors with their locations



Radiation Analysis

Population Exposed

Reactor Locations

Reactor Locations

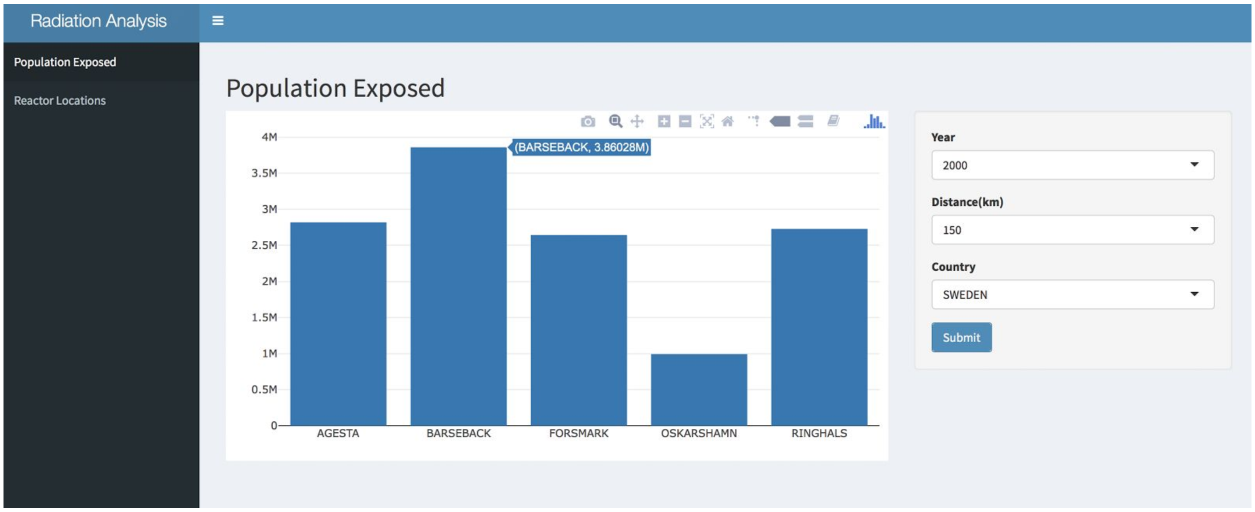
Year: 2017

Month: 4

Submit

In this dashboard the reactor location is visualized and can be filtered using the year and month.

3.2.2 Population exposed to radiation



Radiation Analysis

Population Exposed

Reactor Locations

Population Exposed

Year: 2000

Distance(km): 150

Country: SWEDEN

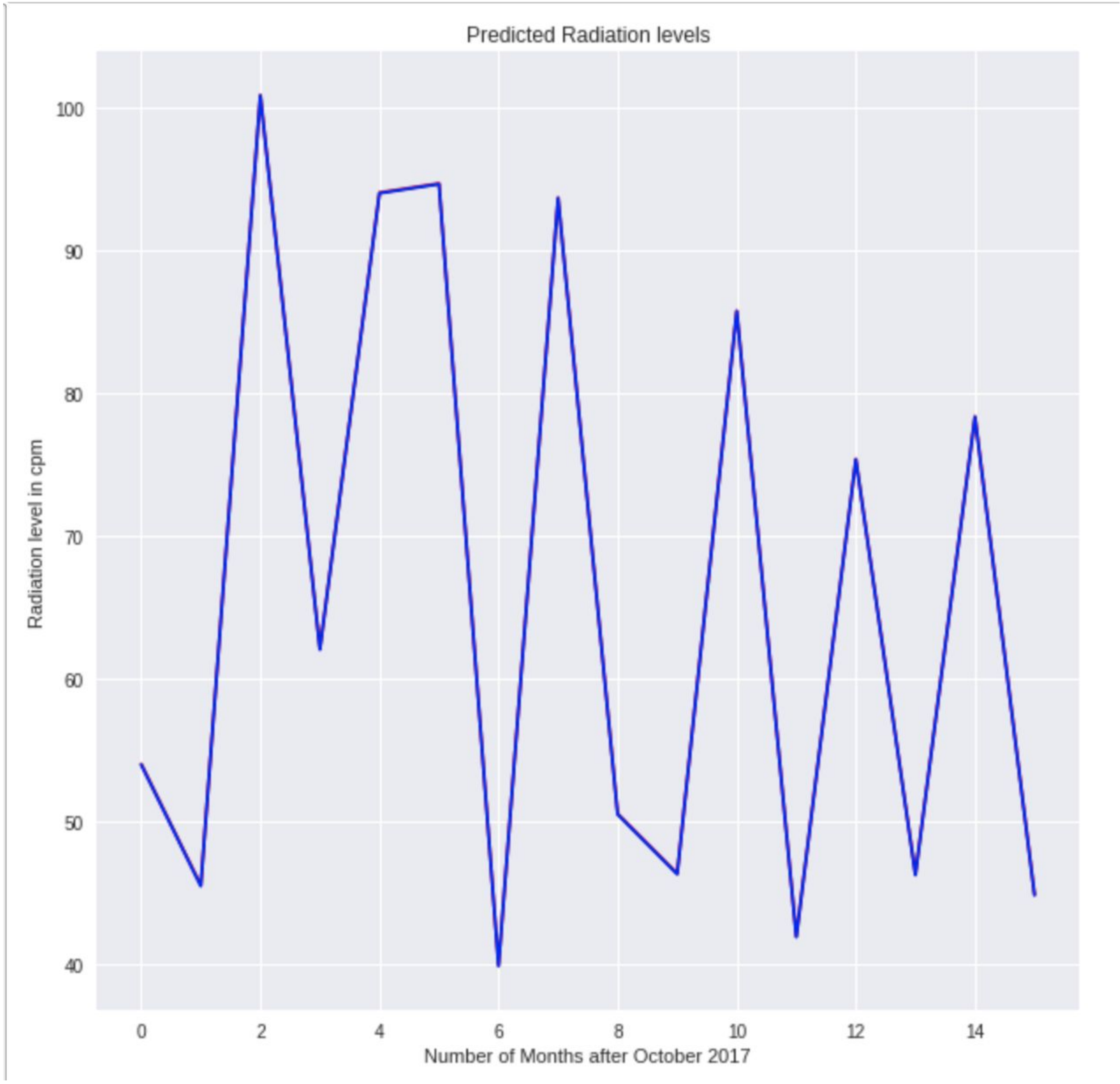
Submit

The figure above shows the population distribution and radiation exposure which can be filtered using a number of parameters like Year, Distance(km) and Country.

3.3 Linear regression on Radiation level

A machine learning model, using SparkML which predicts the radiation level in cpm using the Radiation data. We employed a linear regression model

The results are shown in figure below.



4 Code and References

4.1 Code

The code is present in the folders ‘pyspark’ and ‘RShiny’.

4.2 References

- http://sedac.ciesin.columbia.edu/data/set/energy-pop-exposure-nuclear-plants-locations
- https://api.safecast.org/
- https://www.epa.gov/enviro/data-downloads
- http://spark.apache.org/docs/latest/sparkr.html
- http://spark.apache.org/docs/2.1.0/api/python/pyspark.sql.html
- http://spark.apache.org/docs/2.1.0/api/python/pyspark.html
- http://spark.apache.org/docs/2.1.0/api/python/pyspark.ml.html
- http://spark.apache.org/docs/2.1.0/api/python/pyspark.mllib.html
- http://stackoverflow.com
- http://2knowabout.blogspot.com/2014/02/converting-radiation-readings-from-or.html