SAFETY INDEXING USING BIG DATA TECHNIQUES IN NEW YORK CITY

Muhammad Shihab Rashid, Mehnaz Tabassum Mahin, Madhurima Chakraborty, Shamali Shinde, Urja Parekh
Group - 16

INTRODUCTION

Safety Index: The Safety Index is a weighted index measurement which takes into consideration all kind of threats such as mugging, robbery, road death toll and occurrence of terrorist attacks, quantifying the relative state of safety across all neighborhoods of a city.

$$SI_i = (1 - \frac{1}{k_i} \times \sum_{n=1}^{k} \frac{X_{n,i} - \min X_n}{\max X_n - \min X_n} \times W_n) \times 100$$

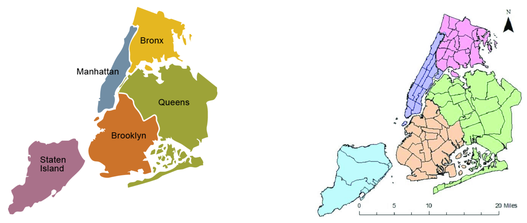
$X_{n,i}$ = No. of Crimes of type n in area i

An index of 100 means the neighborhood is perfectly safe while 0 means it is extremely dangerous.

Objective:

- To provide an effective indicator of an area to be safe for accommodation and trips
- To reduce the crime rates by tracking continuous change of crime regions
- To visualize the safety indicator of an area at a glance

DATA COLLECTION



Required Crime Features:

◈ Type ◈ Seriousness ◈ Location

Types of NYC Boundaries:

◈ Borough ◈ Neighborhood ◈ ZIP Code

NYC Datasets:

- NYPD Crime Dataset [1]
  - ✔6.5M crime entries with 35 features
- NYC Neighborhood and ZIP Code Dataset [2]
  - ✔Neighborhoods of boroughs with zip codes
- NYC ZIP Code Dataset [3]
  - ✔Latitude, longitude of ZIP codes

DATA PREPARATION

We have 3 datasets that contain different informations required to process. In this step, we merge the datasets together to create one complete dataset that contains crime types, zip codes and neighborhoods

- NYC Crime Dataset

Crime	Crime Type	Borough	Latitude	Longitude	Timestamp	ZIP code	Neighborhood
348	VEHICLE AND TRAFFIC	MANHATTAN	40.81077	-73.9526	20:10:00	10027	Central Harlem

- NYC Location Dataset

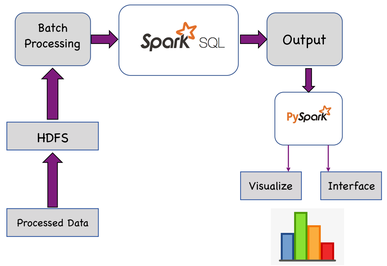
ZIP code	Latitude	Longitude	Neighborhood	Borough
10026	40.802381	-73.952681	Central Harlem	MANHATTAN
10027	40.811407	-73.953060	Central Harlem	MANHATTAN
10029	40.791763	-73.943970	East Harlem	MANHATTAN
10030	40.818267	-73.942836	Central Harlem	MANHATTAN
10128	40.781432	-73.950013	Upper East Side	MANHATTAN
10280	40.708538	-74.016650	Lower Manhattan	MANHATTAN

Closest ZIP Code within the same Borough

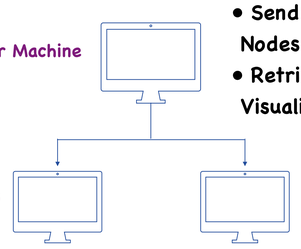
PROPOSED METHODOLOGY

We are working with Big Data. Hence we have used distributed environment, namely Spark Engine to process our data and Python to visualize it.

Overview:



Distributed Environment:



- Send Job to Slave Nodes
- Retrieve Result and Visualize
- Batch Processing
- Spark SQL Query Processing
- Calculate Indexes
- Send Back Results

Spark SQL:

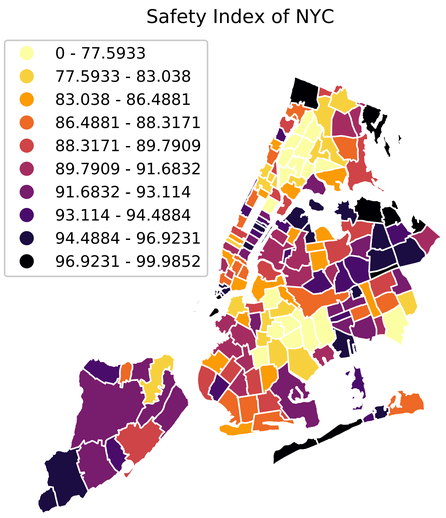
- Process clean data using Spark SQL engine
- Aggregate, Count
- Calculate Index
- The final output consists of zipcode, neighborhood and the value
- The output csv is now ready to be processed and visualize

Crime Type	Latitude	Longitude	...
Burglary	40.73	-73.93	
Murder	3.75	165.0	
Theft	5.0	180.0	

Spark SQL → Aggregate → Count → Weighted Index Calculation

Zipcode	Neighborhood	Latitude	Longitude	Safety Index
10001	East Harlem	40.7506	-73.9917	89.92
10002	Upper East	40.7157	-73.9862	79.02
10003	Central Bronx	40.7318	-73.9891	82.18
10004	Sunset Park	40.6883	-74.0182	57.19

VISUALIZATION



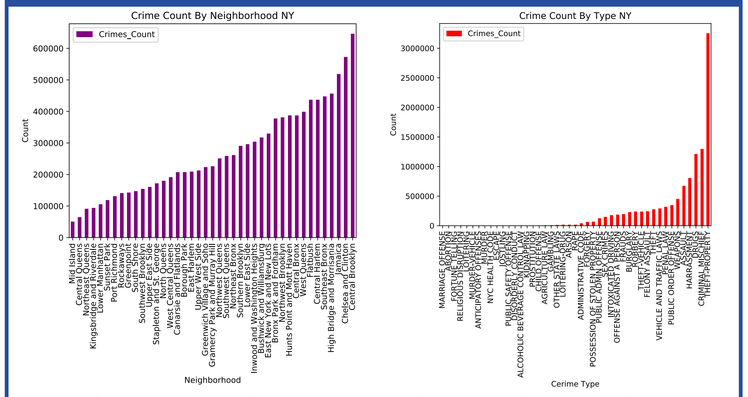
Process:

- Convert csv to geojson
- Python
- geoplot, geopandas, mapclassify

Output:

- A heatmap of NYC according to safety index divided by zipcodes
- The darker the color, the safer the region

STATISTICAL ANALYSIS



Statistical analysis help us to find out patterns and frequencies in crime types.
We can see that Brooklyn area has lots of crimes
Most frequent crime is Property Theft

FUTURE WORK

✓Build a web UI for interactive visualization
✓Include more features like crime density and population
✓Calculate safety index of any country

REFERENCES

[1] https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i
[2] https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/?refine.state=NY&q=New+York
[3] https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm