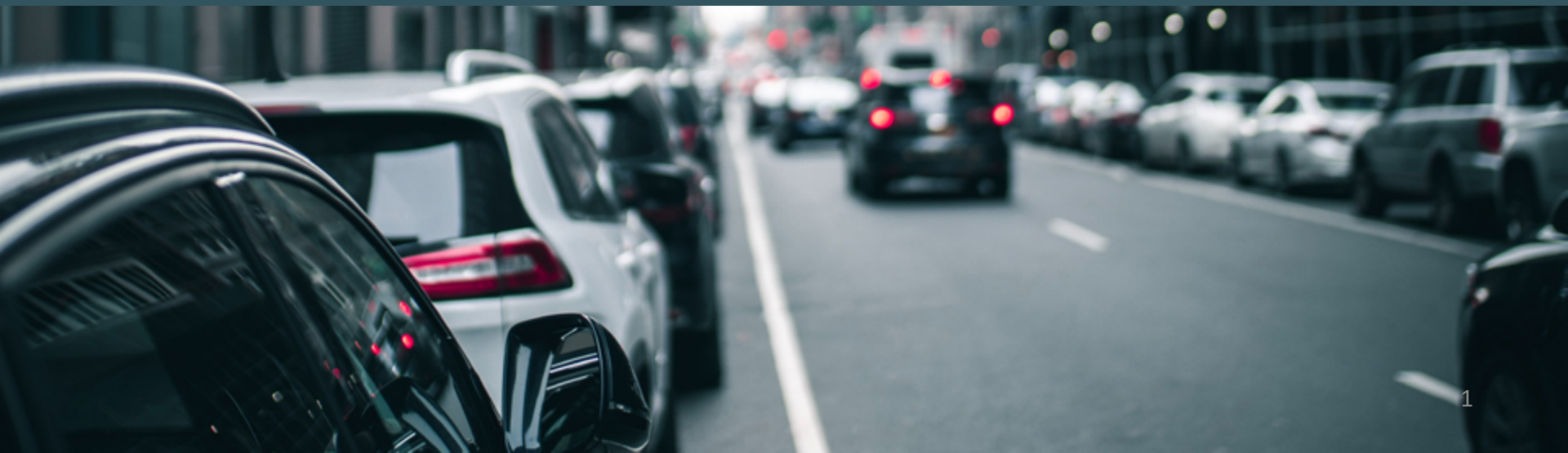
Analyze Crimes In Boston

Team5: Hao Cui, PinHo Wang, Tianju Zhou



Use Cases

1. Users input queries (e.g. date) and receive lists of matching records

2. The system will make clusters to filter crimes and select crimes which happens around holidays

3. The system will analyze the dangerous place for each holiday, then create the related hot-spot chart

4. Users can customize their hot-spot chart by making some inputs

Methodology

1. Digest data from csv files
2. Use Spark to process data
3. Utilize TextRank algorithm to make clusters for crimes related to personal safety
4. Analyze data mathematically and make the hot-spot image

Data Source

Crimes in Boston

https://www.kaggle.com/AnalyzeBoston/crimes-in-boston

Provided by Analyze Boston.

319073 pieces of data in the dataset.

The records begin in June 14, 2015 and continue to September 3, 2018.

Milestones

Data Processing:
- Deal with missing data
- Remove unnecessary data
- Build the program frame-work to analyze data

11.15

Data Visualization

11.22

Analyze:
- Analyze data with Spark SQL
- Get necessary information

11.29

Optimization:
- Optimize the program
- Prepare for the final presentation

12.6

Code

The program will be coded in Scala and Spark.

Repository:
https://github.com/CSYE7200/Analyze-Crimes-Boston

Acceptance Criteria

1. All Spark SQLs should be executed within 5 seconds

2. ~~All data (319073 rows) need to be analyzed~~

3. ~~Include at least 10 test cases~~

4. ~~Should solve all 4 goals posted as our goals~~

Goals

01

Figure out:

- What types of crimes are most common?
- Does the frequency of crimes change over the day?
- Where does the crime most frequently occur?
- Is crime rate relate to other factors such as housing price and wage?

02

Provide information:

High crime rate locations during holidays or weekends in Boston

Thank you