CS - GY 6513 Big Data Project

Group 6

Minghao Shao
Haozhong Zheng
Hanqing Zhang

Introduction

1. Start from "Citywide Payroll Data (Fiscal Year)" dataset
2. Hunting for datasets from NYC OpenData with overlap columns
3. Profiling and clean the datasets
4. Refining our strategies
5. Visualisation

Overlap detection

- Handling large scale of dataset on PEEL HPC server
- We use spark to measure overlap similarity
- K-shingle with jaccard similarity based on column names
- Problems about fuzzy-matching:

  “Work Location Borough” vs “Borough”?

Agency Name

- Original data clean strategy

  Capital letter integration, mark missing values, KNN clusters.

- Result of effectiveness and problems found

  Quite effective, still has wrong form problems.

- Reference data regarding agency names and the department they belongs to

	Agency	Agency Name
1	HPD	Department of Housing Preservation and Development
2	DOT	Department of Transportation
3	DEP	Department of Environmental Protection
4	NYPD	New York City Police Department
5	DOB	Department of Buildings
6	DPR	Department of Parks and Recreation
7	DOHMH	Department of Health and Mental Hygiene
8	DCA	Department of Consumer Affairs

Agency reference data example

Agency Name

- Refined strategy

  Create reference data, map abbreviation to fix form issue.

- Challenges and limitations

  KNN cluster reliability, can not visualize the data because of the scale.

	Agency	Agency Name
2	HPD	Department of Housing Preservation and Development
3	DOT	Department of Transportation
4	DEP	Department of Environmental Protection
5	NYPD	New York City Police Department
6	DOB	Department of Buildings
7	DPR	Department of Parks and Recreation
8	DOHMH	Department of Health and Mental Hygiene
9	DCA	Department of Consumer Affairs

Agency reference data example

Borough

- Data formatting

  Missing values, Conjunctive columns and abbreviations

- Typo issues

  Handled manually supported with KNN cluster

- Reference data

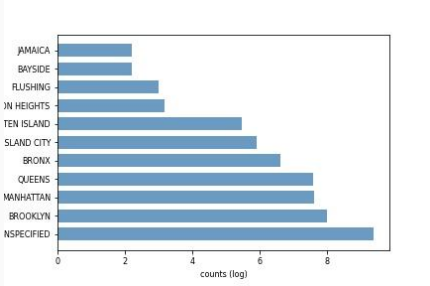
  Mapping between full borough name and their potential abbreviations

- Visualisation

  Treemap and bar chart

	Abbreviation	Borough
1	MN	MANHATTAN
2	BK	BROOKLYN
3	QN	QUEENS
4	BX	BRONX
5	S.I.	STATEN ISLAND
6	LIC	LONG ISLAND CITY
7	SI	STATEN ISLAND

Borough reference data formatting



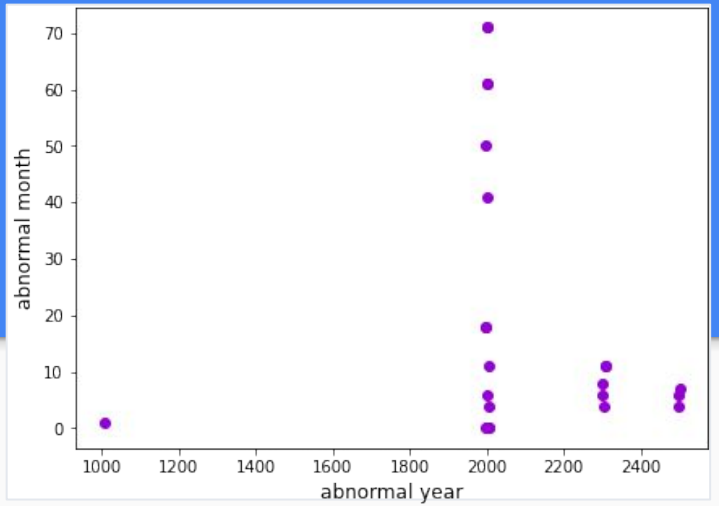
One example of visualisation

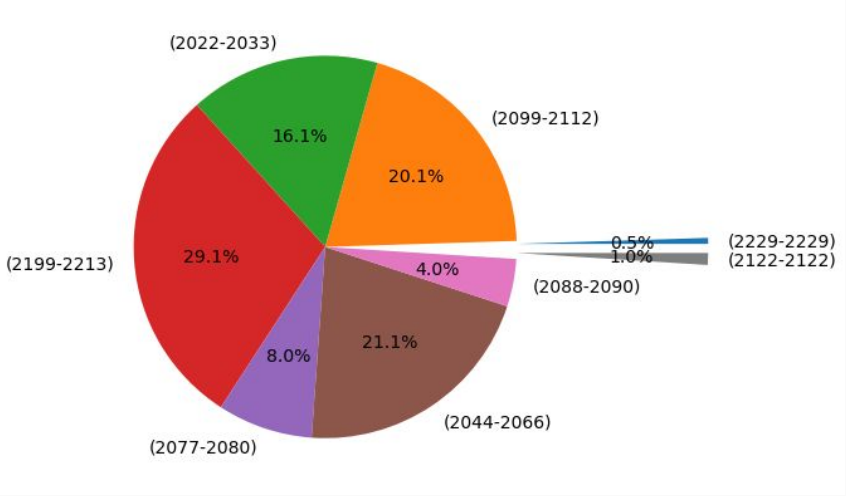
Date

- Original dataset: column “Agency Start Date”
- strictly follows format “%m/%d/%Y”
- Original profiling and cleaning strategy: split by slash character
- Delete rows: years > 2021
- ***Not going to work on other datasets!***
- Years > 2021 is possible
- Validation: “13/41/2020”

Date

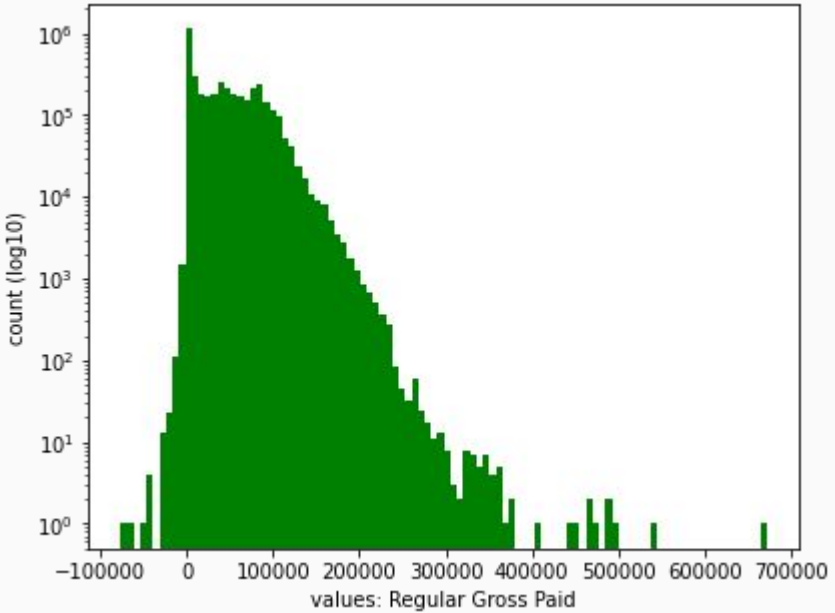
- Datetime-parsing methods to validate
- Multiple datetime formats: %m/%d/%Y", "%Y-%m-%dT%H:%M:%S" etc.
- DBSCAN: Clustering possible outliers (year > 2021)





Salary

- Many kinds of unusual format:
- “ 123,456 ”, “2/3”
- Analyzing the distribution via histogram
- Negative numbers are guaranteed to be outliers
- Very hard to find other outliers: a person’s salary is possible to go from 0 to millions



Summary - gaining and challenging

- For some column, even the fields and values are overlap, they may recorded in very different format, which should be inspected manually
- Generating reference data from a large scale of dataset collection is very helpful to extend data clean strategy to other datasets (Reproducibility and Replicability)
- Investigation of dataset background is important, such as their expected format, and their meaning, which could help us to identify the problem in datasets
- Measuring the effectiveness of data clean is not a easy job, sometimes we may find the naive data clean strategy is hard to handle data with different format, even they are overlap
- Data balance is another problem, sometimes unbalanced data could be confuse

Reference links

- Github repository link:

  https://github.com/NickNameInvalid/Big_Data_Report

- Google Drive:

  https://drive.google.com/drive/u/1/folders/1Gmjduu2zaeupyAYgQPRpstCdkOa6LwVy

Thanks!
Q & A

Contact us:

Minghao Shao
ms12416 @nyu.edu

Haozhong Zheng
hz2675@nyu.edu

Hanqing Zhang
hz2758@nyu.edu