Profiling and analysis of NYC Open Data

Sai Likhith
Kartikeya Negi
Reuben Varghese

Overview

Data has gained prominence in the last few years

Big data problems are the norm

Understanding and solving them is the challenge for this generation of engineers

Key Objectives

Simulate a set of real-world big data problems

Solve these problems meaningfully, using techniques learnt in class

NYC Open data

Generic profiling

Semantic profiling

Data analysis

Understanding the problem

Generic profiling

Data comes with very less description (metadata)

Noise

Unclassified data types

Sparse/wrong/missing data

Semantic profiling

What real world entities (semantics) does our data contain?

Data analysis

What can our data tell us about the real world that it models?

Generic Profiling

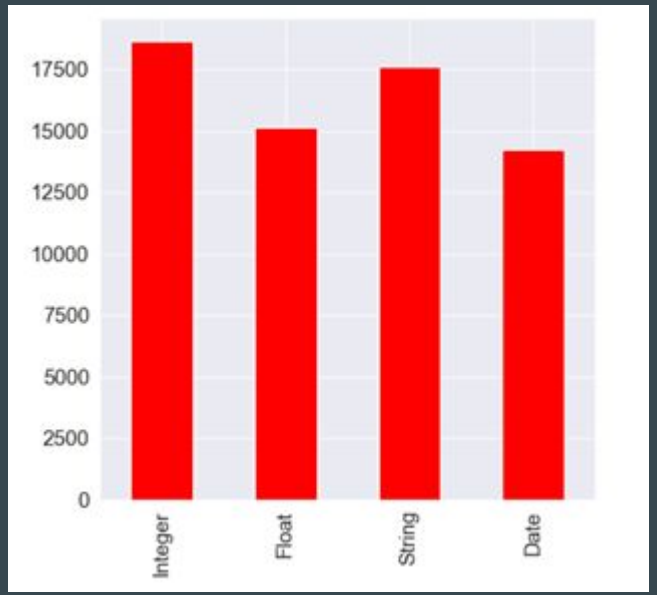
Structure profiling

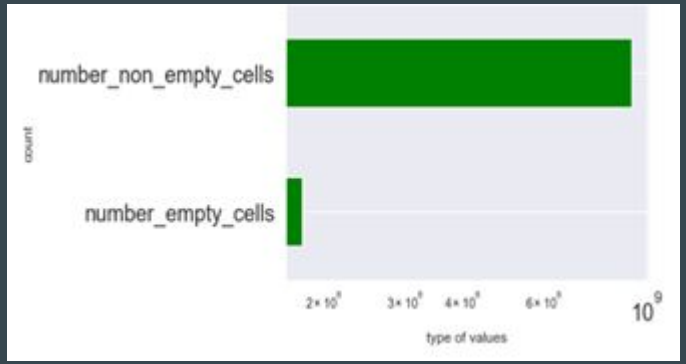
- data types (text, integer(long), time/date, float)

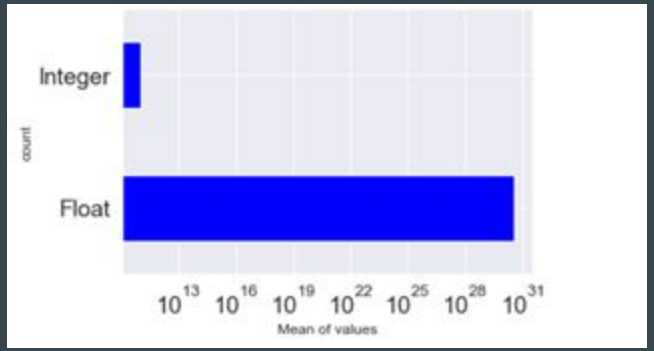
Content profiling

- full/nan values,
- missing values,
- unique values
- frequent values
- max/min/mean/std. dev

Analysis


17500
15000
12500
10000
7500
5000
2500
0
Integer
Float
String
Date


number_non_empty_cells
number_empty_cells
count
type of values


Integer
Float
count
10^{13} 10^{16} 10^{19} 10^{22} 10^{25} 10^{28} 10^{31}
Mean of values

??

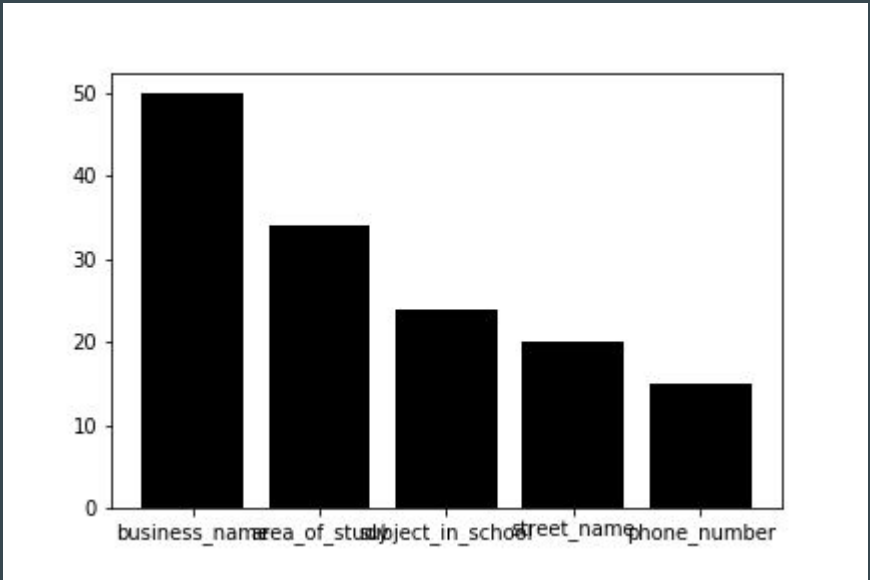
Semantic Profiling

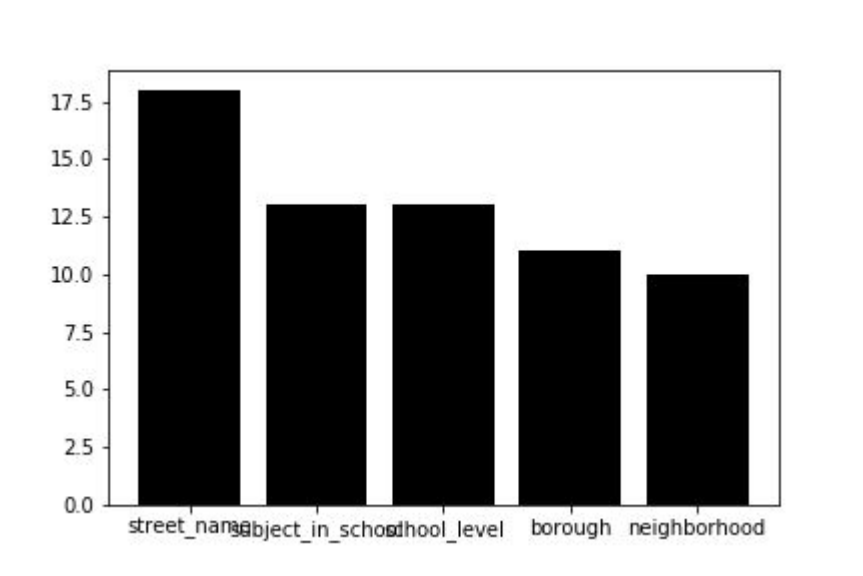
Column semantics using column names: ground truth

Functions defined/ontologies created for each semantic type:

- Regular expressions (phone number, location, etc.)
- NLP model: name entity recognition library (person name, business name, etc.)
- Fuzzy string matching (parks, building classification, vehicle type)
    - Online ontologies

Analysis


50
40
30
20
10
0
business_name
area_of_study
subject_in_school
street_name
phone_number


17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
street_name
subject_in_school
school_level
borough
neighborhood

Data analysis

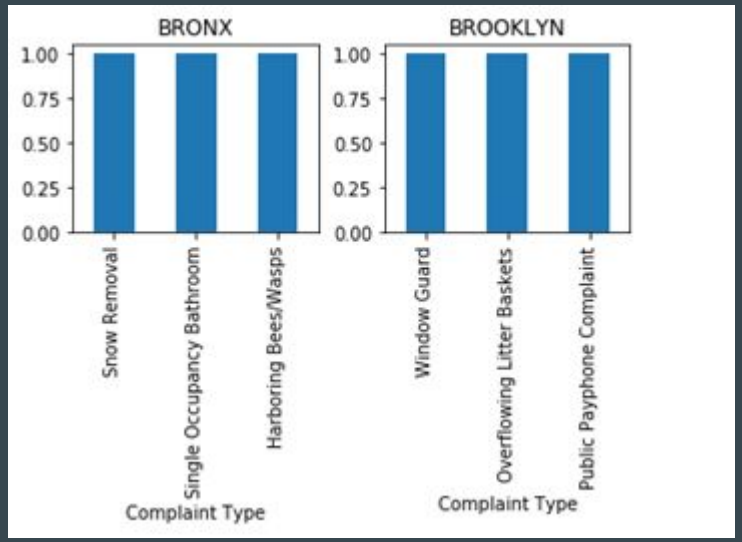
Gaining real world information from our model of the world (data)

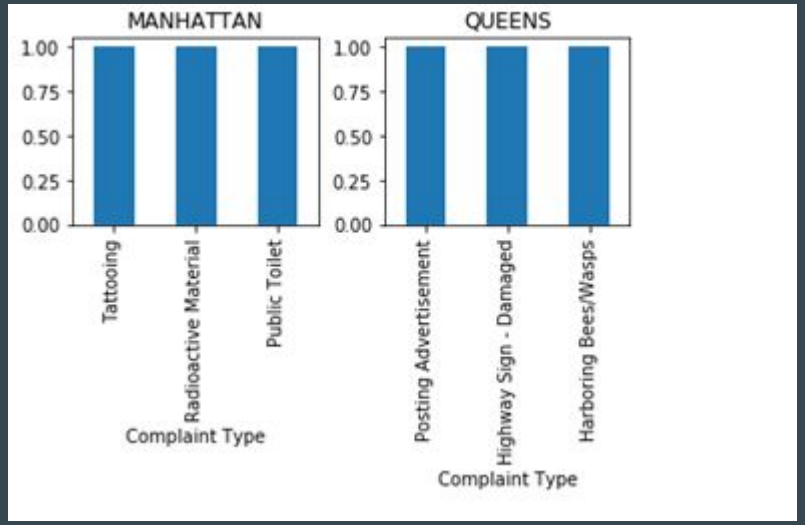
NYC 311 entries (November 2019)

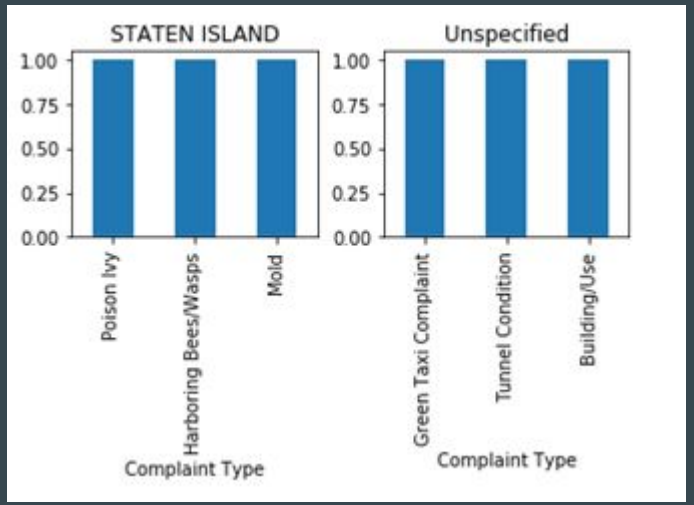
Most frequent complaints by Borough

Why 1 month?

- Most recent/relevant info
- Analyse the effect of thanksgiving on most frequent complaints


BRONX
1.00
0.75
0.50
0.25
0.00
Snow Removal
Single Occupancy Bathroom
Harboring Bees/Wasps
Complaint Type
BROOKLYN
1.00
0.75
0.50
0.25
0.00
Window Guard
Overflowing Litter Baskets
Public Payphone Complaint
Complaint Type


MANHATTAN
1.00
0.75
0.50
0.25
0.00
Tattooing
Radioactive Material
Public Toilet
Complaint Type
QUEENS
1.00
0.75
0.50
0.25
0.00
Posting Advertisement
Highway Sign - Damaged
Harboring Bees/Wasps
Complaint Type


STATEN ISLAND
1.00
0.75
0.50
0.25
0.00
Poison Ivy
Harboring Bees/Wasps
Mold
Complaint Type
Unspecified
1.00
0.75
0.50
0.25
0.00
Green Taxi Complaint
Tunnel Condition
Building/Use
Complaint Type