

Correlation analysis on spatiotemporal data

Heejong Kim(hk2451), Jun Yuan(jy50), Jiin Nam(jn1664)

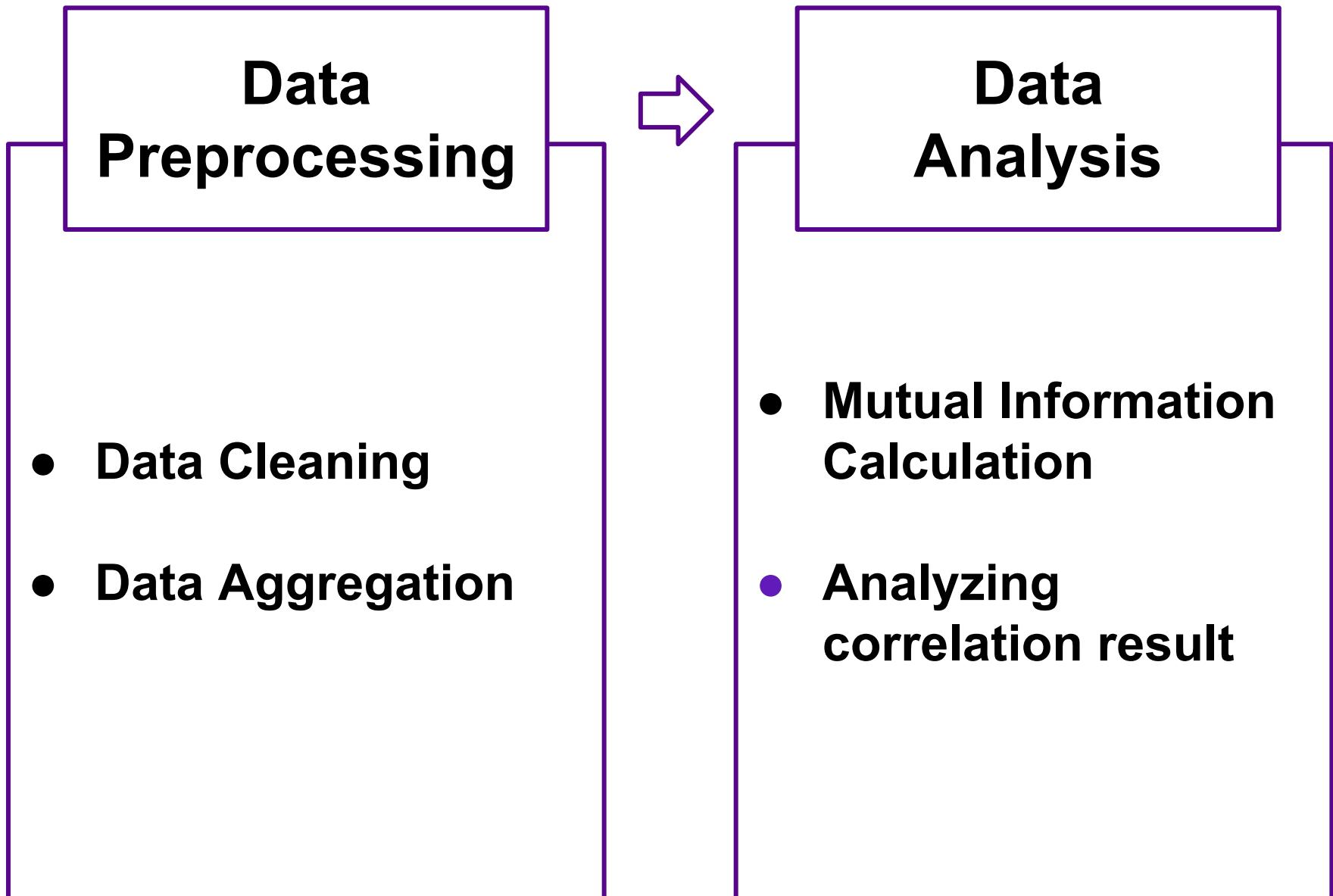
Cool Name Pending Team

Big Data (DS-GA 1004) 18' Spring

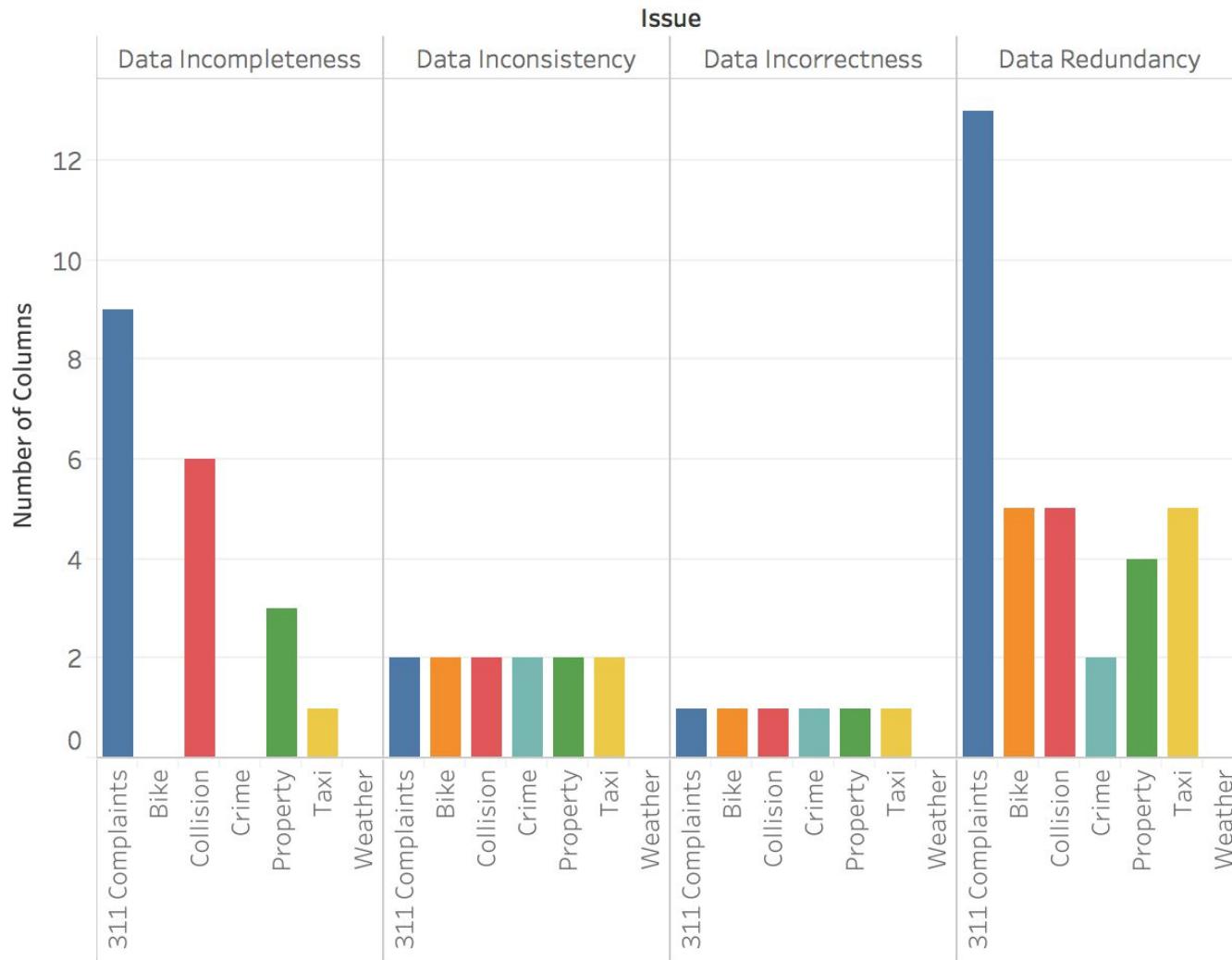


NEW YORK UNIVERSITY

Pipeline



Data cleaning



Data Quality Issues

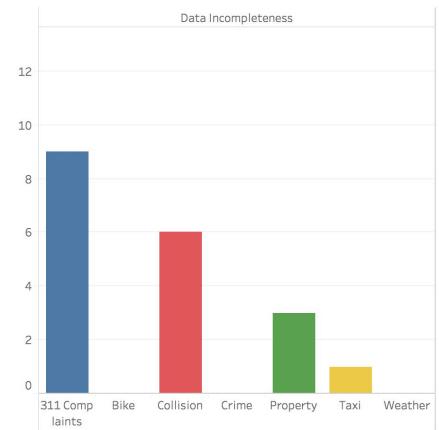
1. Incompleteness
2. Inconsistency
3. Incorrectness
4. Redundancy

Number of columns for each issue in each dataset

Data Cleaning

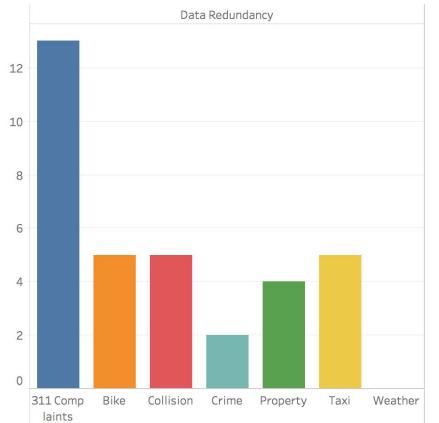
Data Incompleteness

- Performance: A lot of “empty” values (e.g.: null, N/A, Unspecified, ect.)
- Solution: Delete columns with over 90% empty values.



Data Redundancy

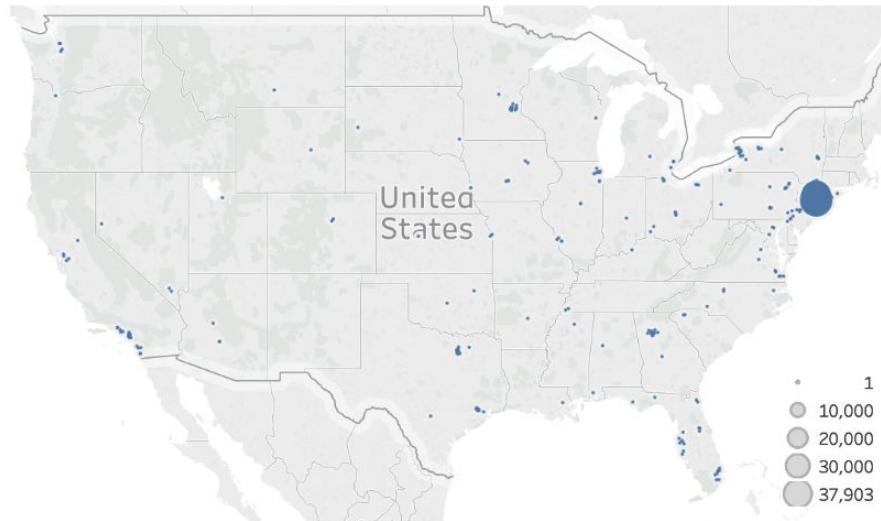
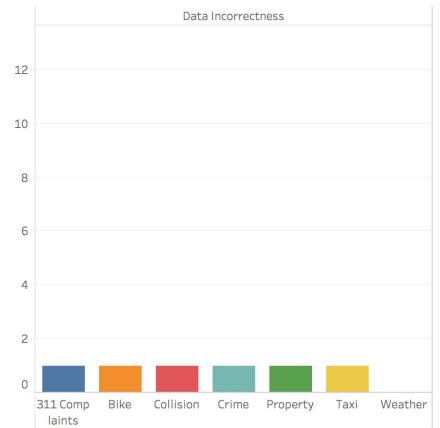
- Performance: Multi-columns describe the same / similar information
- Solution: Keep one column among the similar columns



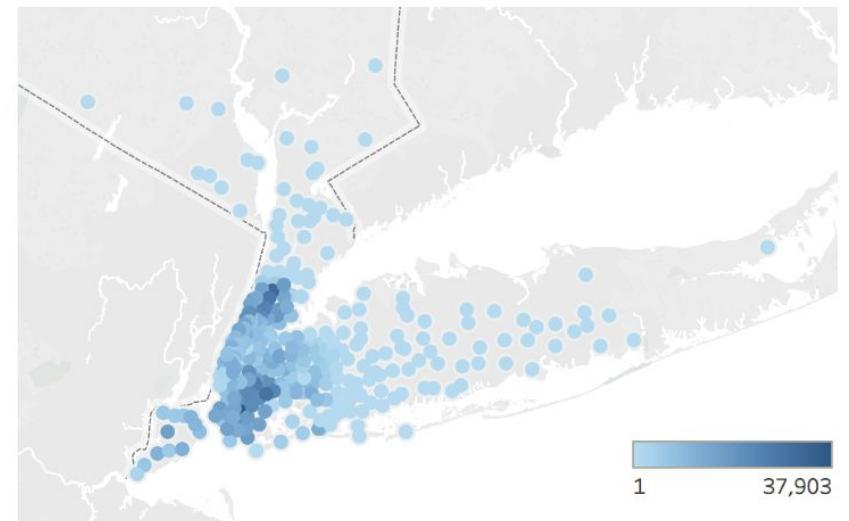
Data Cleaning

Data Incorrectness

- Performance: location not in New York City
- Solution: NYC zip code validation



Locations **Before** Cleaning

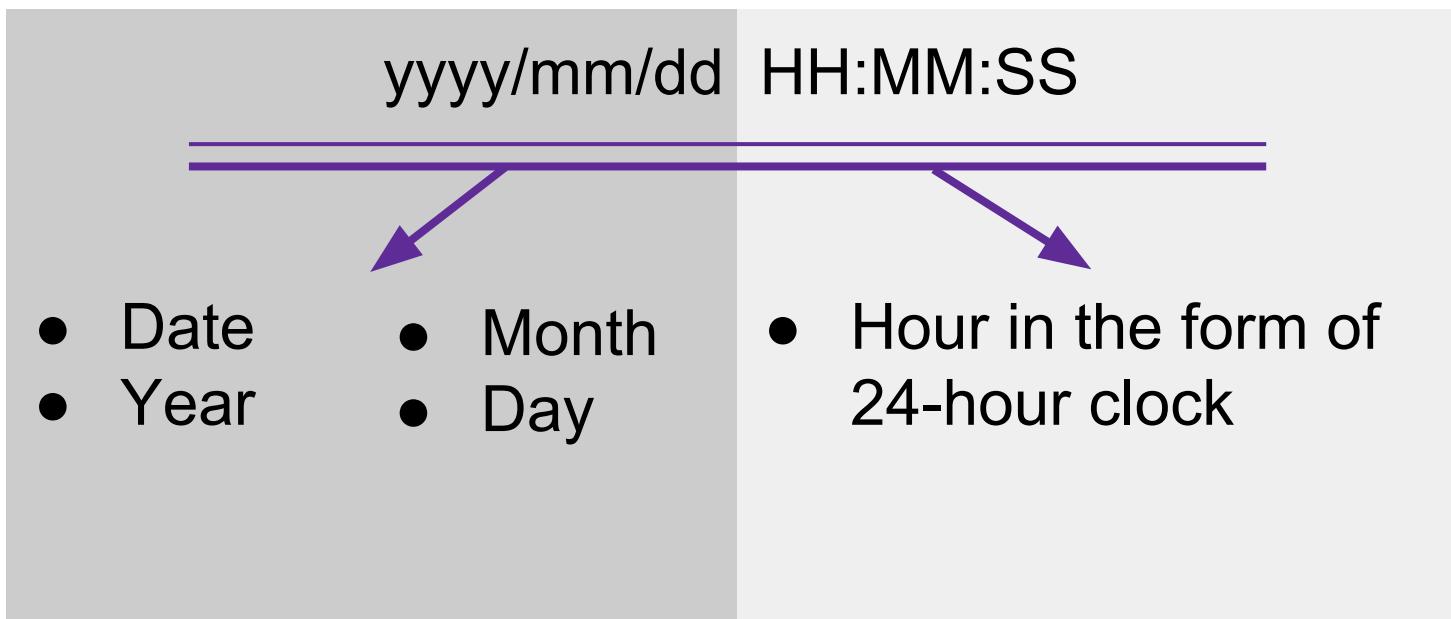
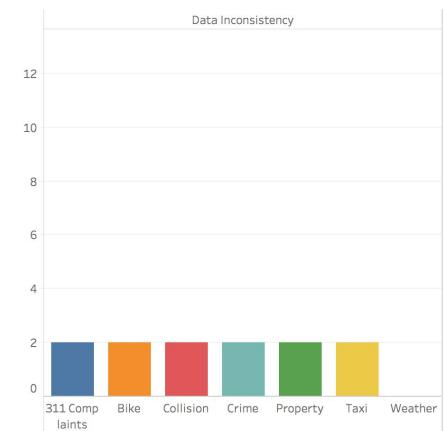


Locations **After** Cleaning

Data Cleaning

Data Inconsistency

- Performance: Different format and granularity
- Solution
 - For zipcode: keep first 5 digits
 - For time: Formatter and new attrs mapping

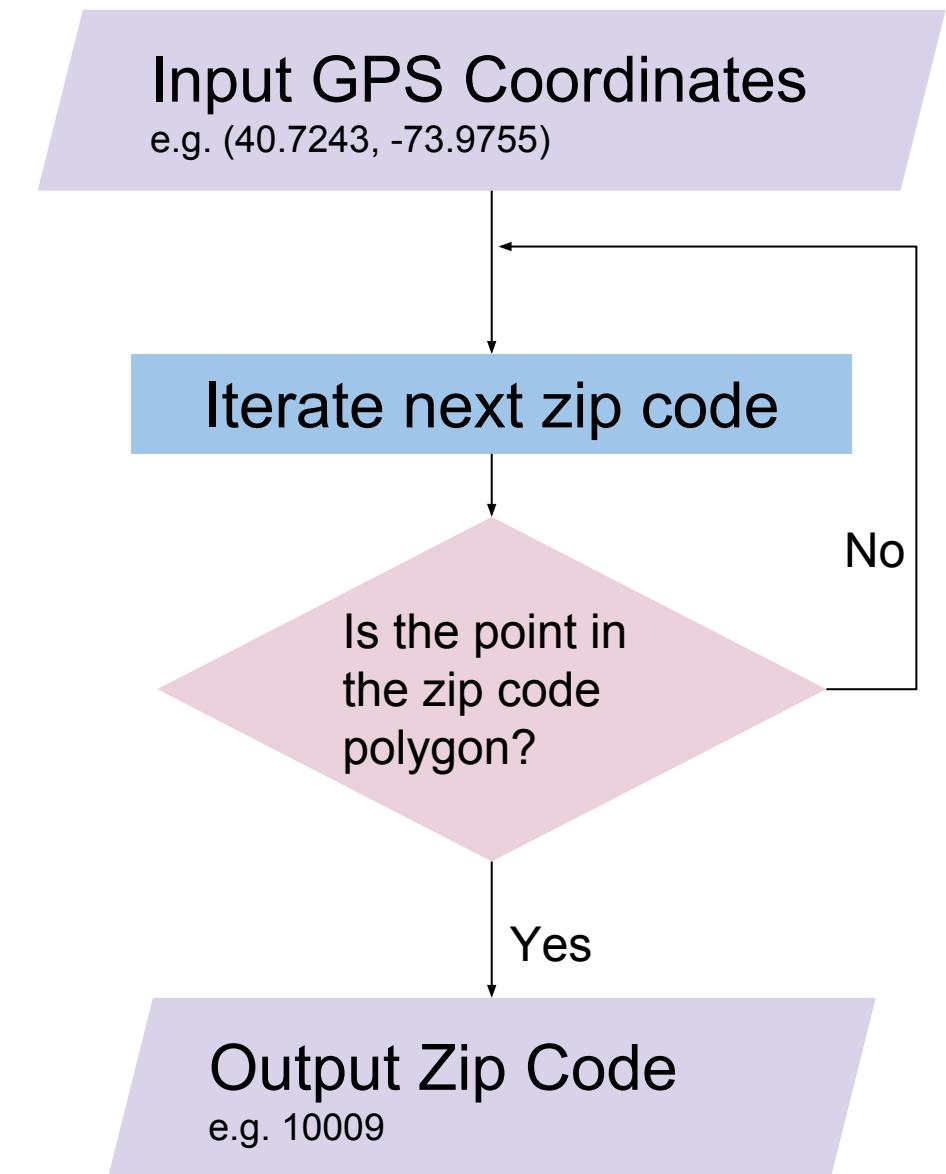
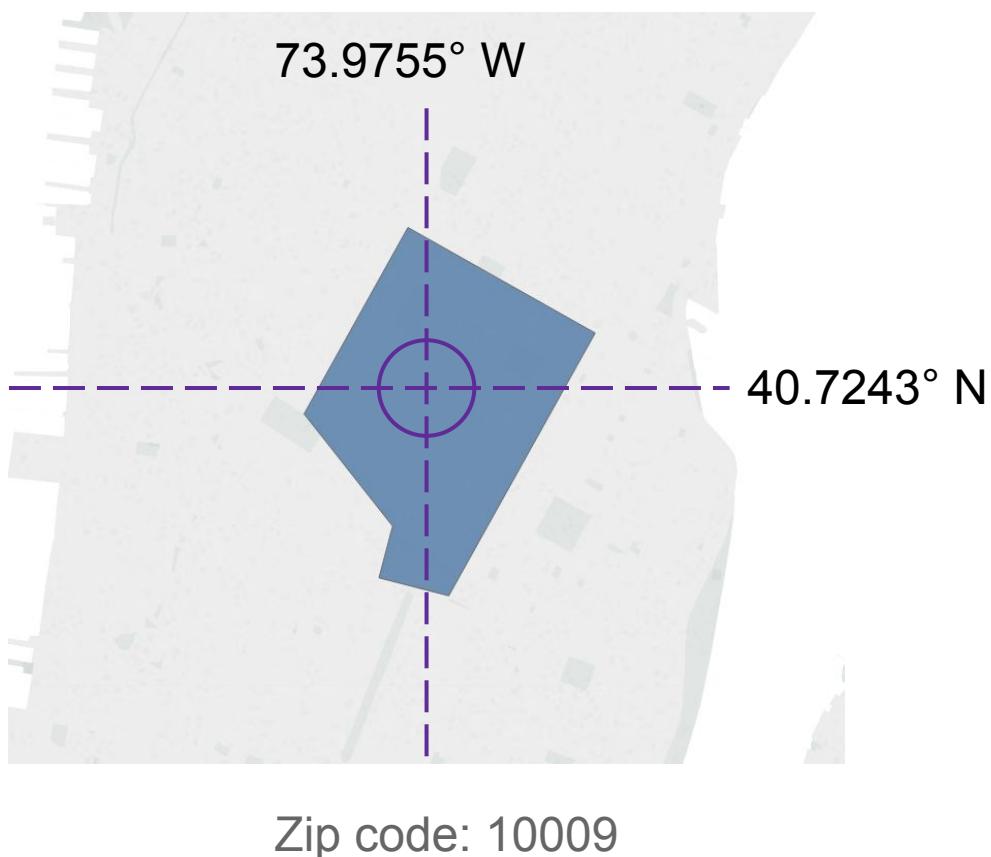


Data Aggregation

- Zip Code Mapping
- Attribute Transformation
- Attribute Categorization

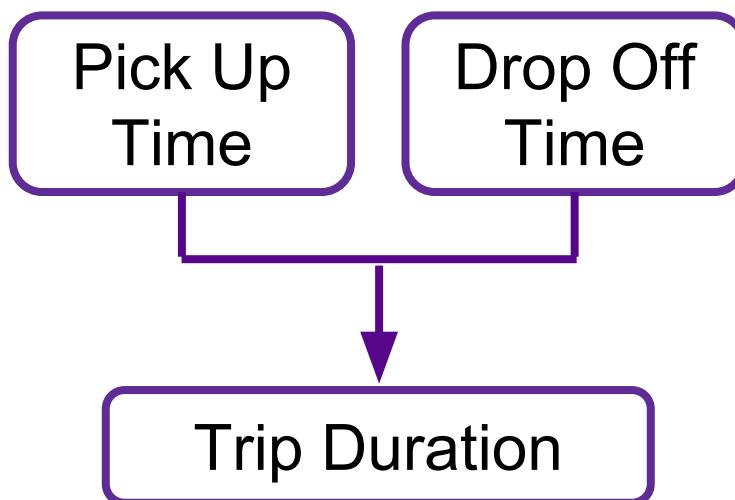
Data Aggregation

Zip Code Mapping



Data Aggregation

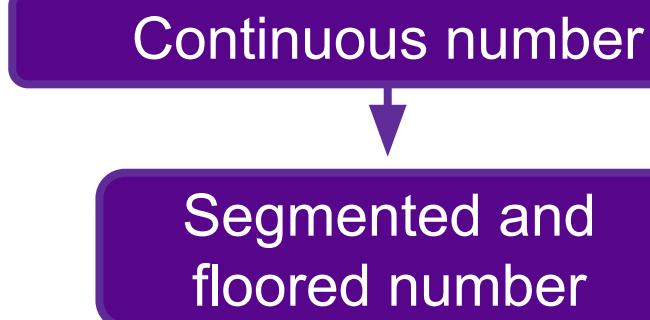
Transformation



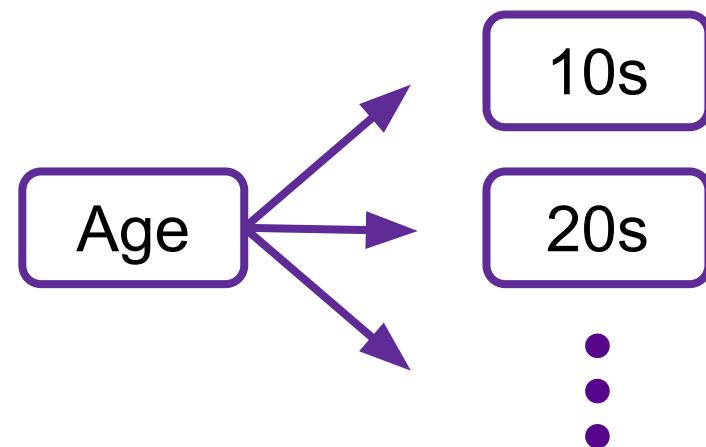
Birth Year

Age

Categorization

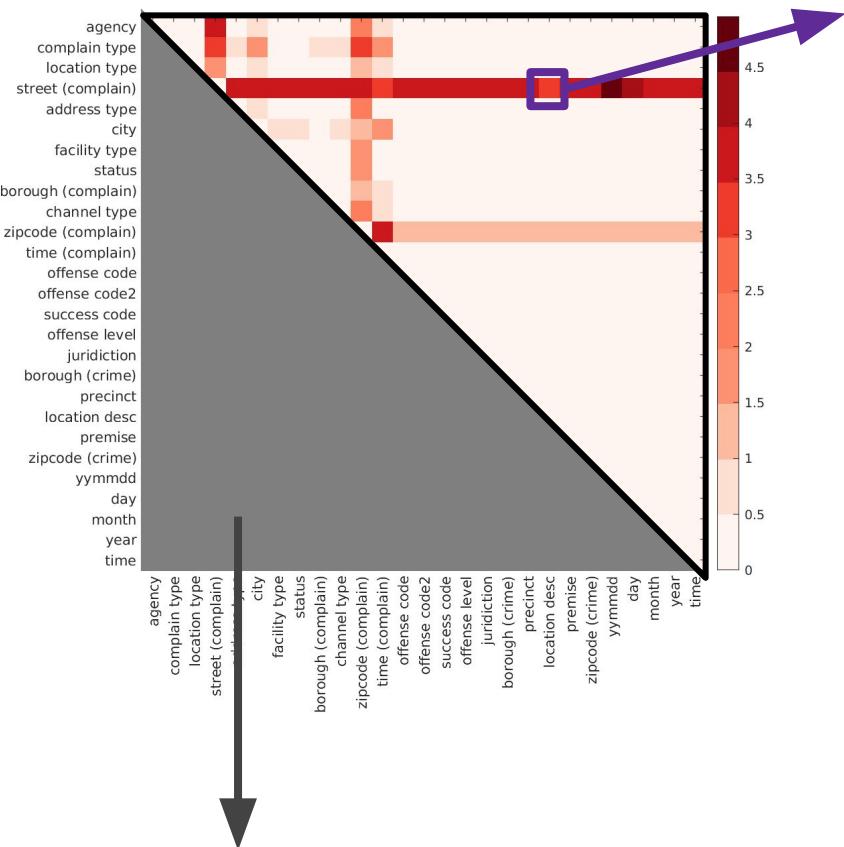


Example: speed, temperature, etc



Mutual Information

Calculating “Mutual Information” with example



Mutual information between
“Street name (X) from complain dataset”
& “Location description (Y) from crime dataset”

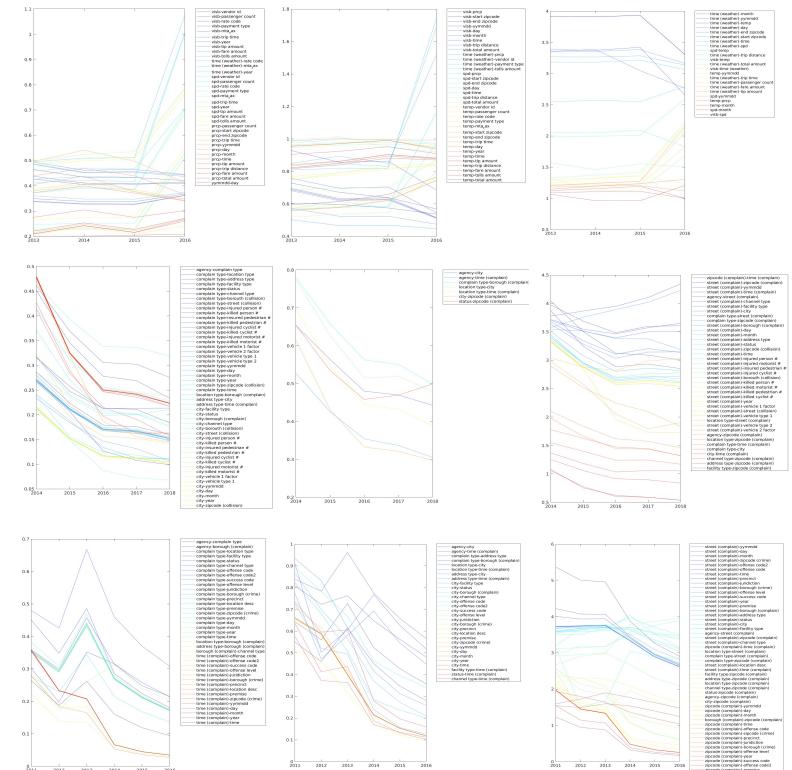
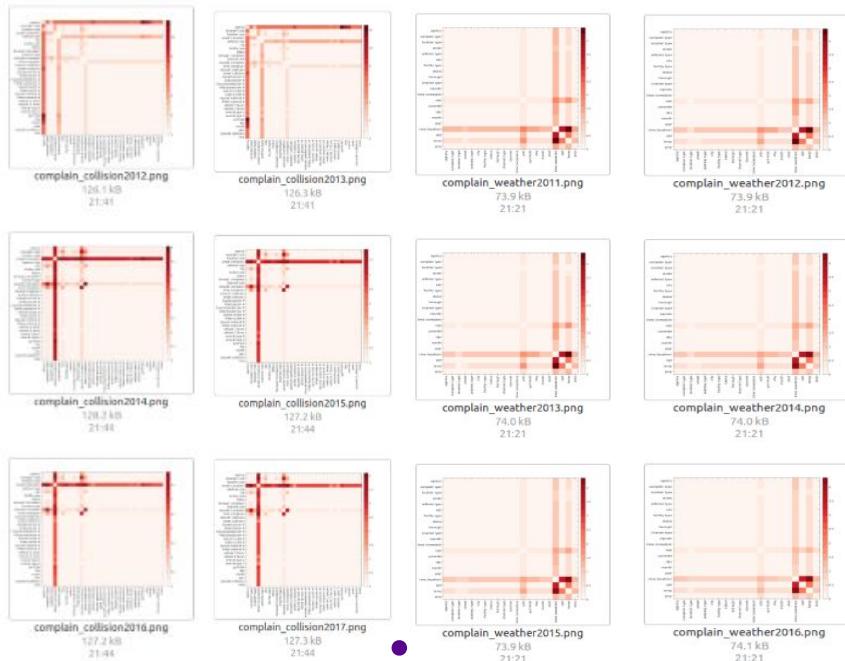
- 1) Join two dataset with dates (“yymmdd”)
- 2) Calculate
 - a) $P(X)$ and $P(Y)$
 - b) $P(XY)$
 - c) Mutual Information =
$$P(XY) * \log(P(XY) / (P(X)*P(Y)))$$

Correlation matrix is symmetric

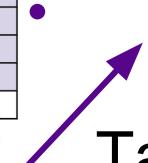
► We calculated upper triangle only for efficiency

Correlation results

Overview

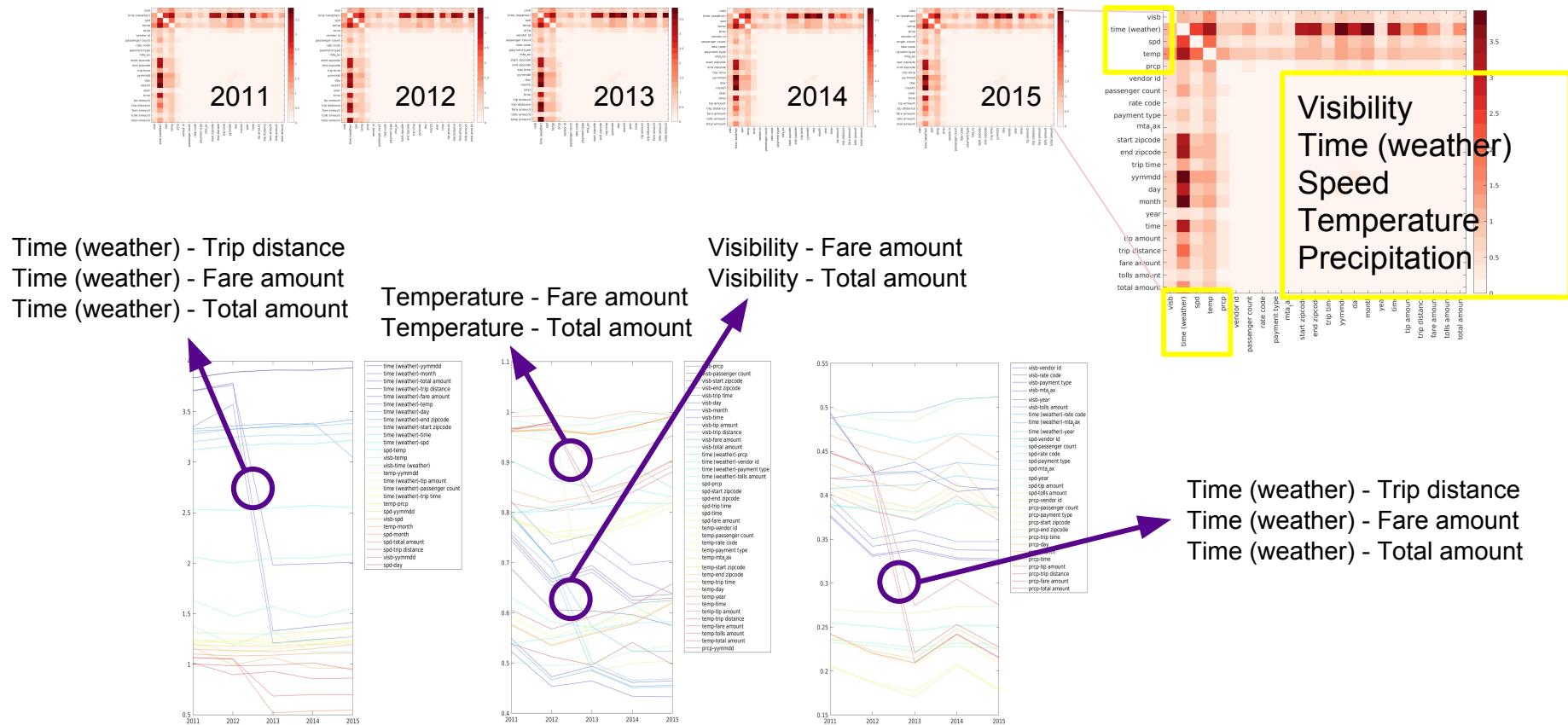


	Collision	Taxi	Complain	Bike	Crime	Property	Weather
Collision	.						
Taxi	.	-					
Complain	.	-	-				
Bike	.	-	-	-			
Crime	.	-	-	-	-		
Property	.	-	-	-	-	-	
Weather	.	-	-	-	-	-	-



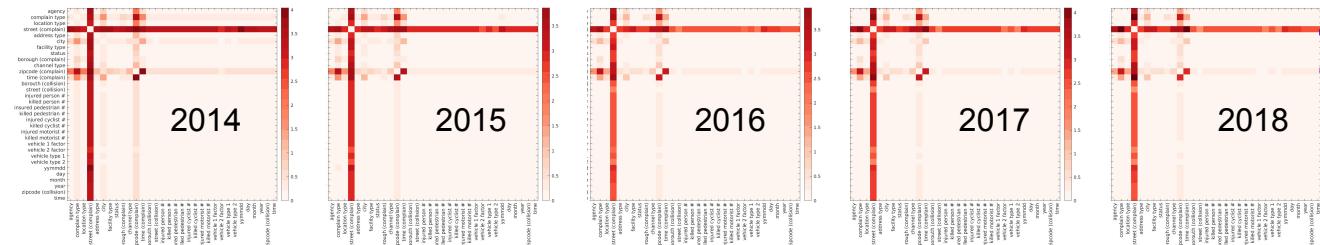
Tables Join By Year

Case study 1: Weather and Taxi



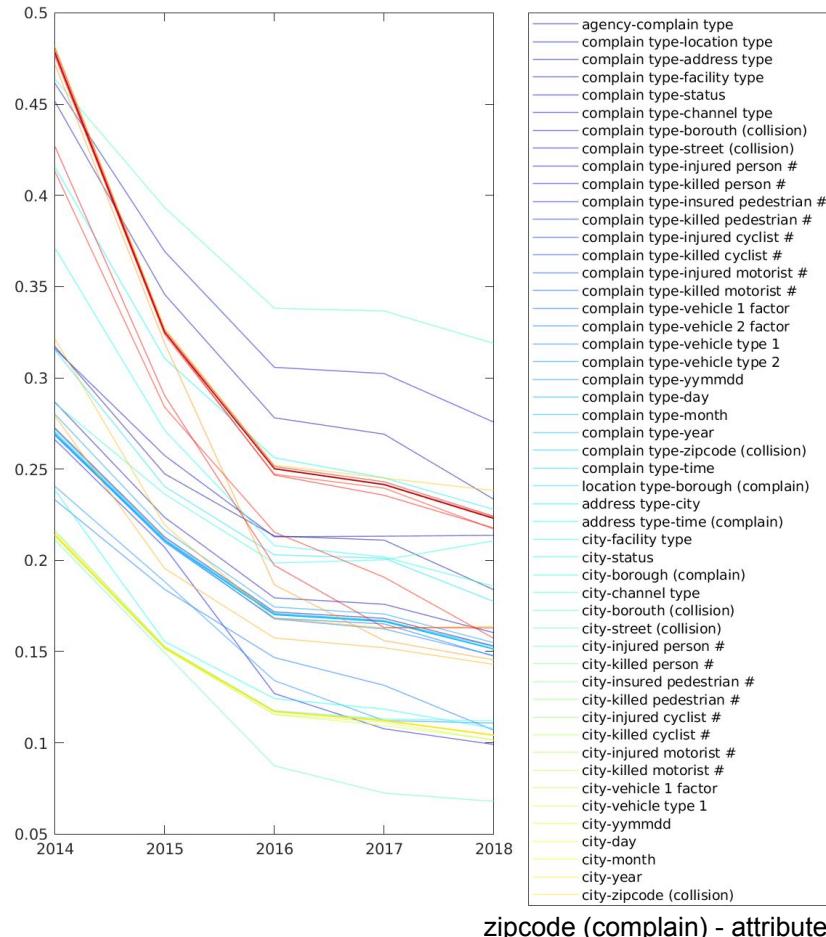
- **Weather attributes** are highly correlated to the taxi tips, passenger count, trip time.
- Correlations between the weather and some taxi trip attrs drop a lot in 2013. Maybe because Uber became popular and got approved in NYC in 2013 [1][2].

Case study 2: Complaint and Collision



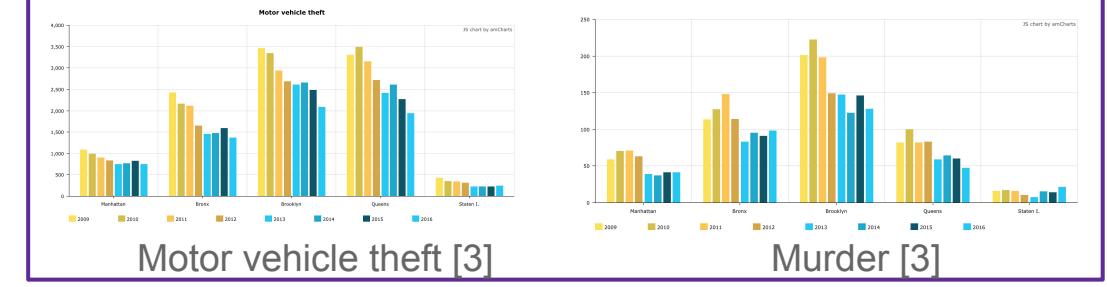
Street attribute in Complaint
Zip code in Complaint

Difference between business area and residential area



- Complain type is correlated to locations, streets
- The correlations are decreasing

- More expensive traffic tickets
- Less Crime in Brooklyn/Bronx => less correlated



References

[1] Uber Closes Yellow Taxi Cab Service In New York City:

<https://www.forbes.com/sites/tomiogeron/2012/10/16/uber-closes-yellow-taxi-cab-service-in-new-york-city/#fe9787051e27>

[2] After long battle, Uber becomes first taxi app to get approved in New York City:

<https://www.theverge.com/2013/4/26/4271490/uber-becomes-first-taxi-app-to-get-approved-in-new-york-city>

[3] Major crime in New York City, 2009-2016:

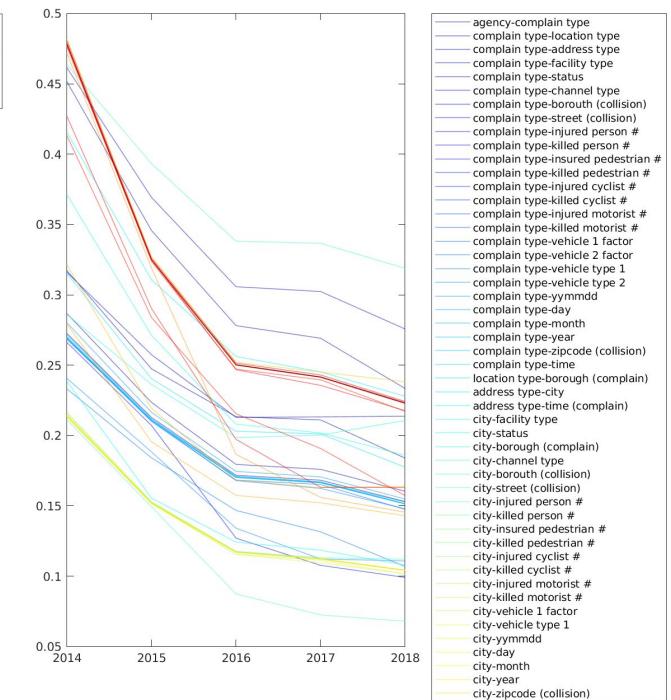
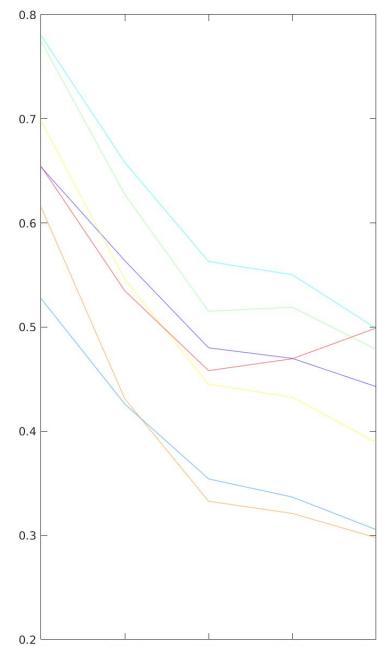
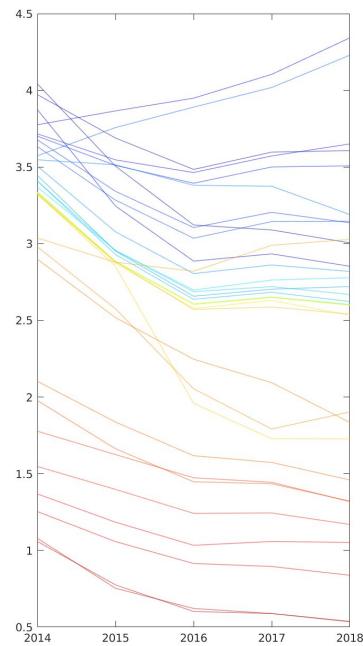
<https://projects.newsday.com/databases/long-island/new-york-city-crime-rate/>

Thanks :)

Cool Name Pending Team

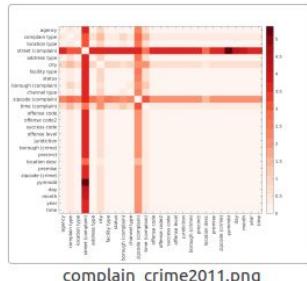
Supplementary

Complain collision 2014-2018

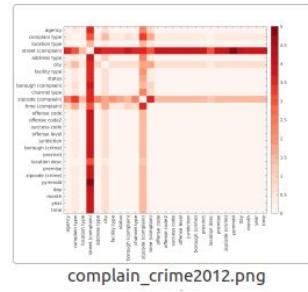


Supplementary

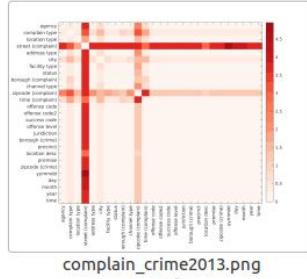
Complain crime 2011-2016



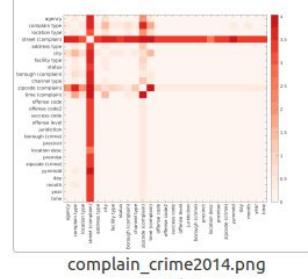
complain_crime2011.png



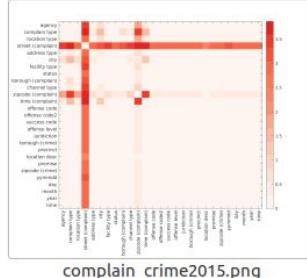
complain_crime2012.png



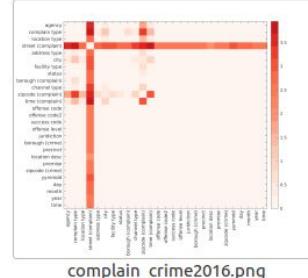
complain_crime2013.png



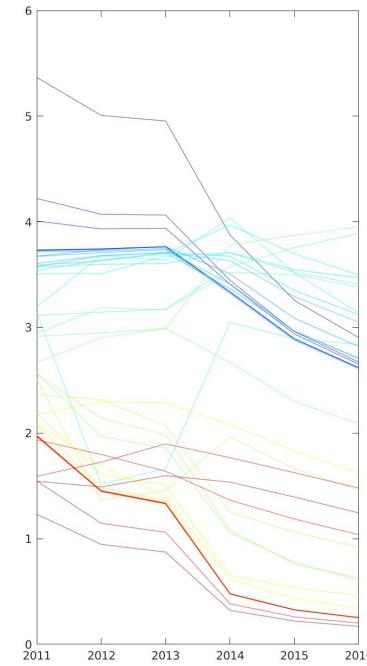
complain_crime2014.png



complain_crime2015.png

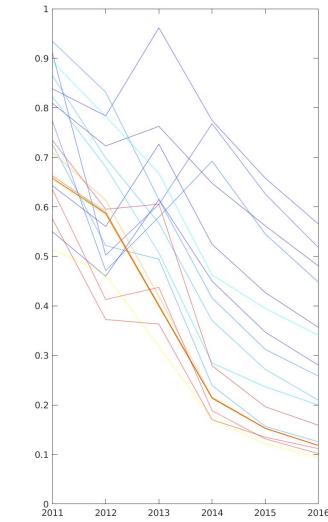


complain_crime2016.png



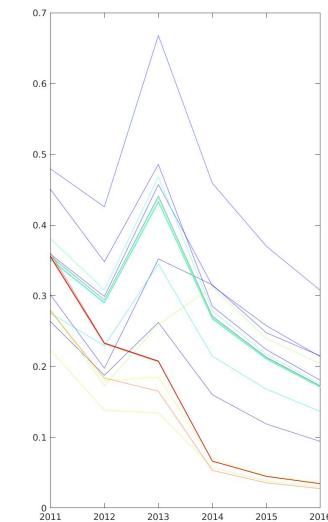
Legend for the line chart:

- street (complain)-ymmdd
- street (complain)-day
- street (complain)-month
- street (complain)-zipcode (crime)
- street (complain)-offense code2
- street (complain)-offense code
- street (complain)-time
- street (complain)-precinct
- street (complain)-jurisdiction
- street (complain)-borough (crime)
- street (complain)-offense level
- street (complain)-success code
- street (complain)-year
- street (complain)-premise
- street (complain)-borough (complain)
- street (complain)-address type
- street (complain)-status
- street (complain)-city
- street (complain)-facility type
- agency-(complain)
- street (complain)-zipcode (complain)
- street (complain)-channel type
- zipcode (complain)-type (complain)
- location type-street (complain)
- complain type-street (complain)
- complain type-zipcode (complain)
- address type-zipcode (complain)
- location type-zipcode (complain)
- channel type-zipcode (complain)
- status-zipcode (complain)
- agency-zipcode (complain)
- zipcode (complain)-ymmdd
- zipcode (complain)-day
- zipcode (complain)-month
- borough (complain)-zipcode (complain)
- zipcode (complain)-type
- zipcode (complain)-offense code
- zipcode (complain)-offense code2
- zipcode (complain)-precinct
- zipcode (complain)-jurisdiction
- zipcode (complain)-borough (crime)
- zipcode (complain)-offense level
- zipcode (complain)-year
- zipcode (complain)-success code
- zipcode (complain)-offense code2
- zipcode (complain)-premise



Legend for the line chart:

- agency-city
- agency-time (complain)
- complain type-address type
- complain type-borough (complain)
- location type-time (complain)
- address type-city
- address type-time (complain)
- city-facility type
- city-store type
- city-offense code
- city-offense code2
- city-offense level
- city-borough (crime)
- city-precinct
- city-location desc
- city-premise
- city-zipcode (crime)
- city-zipcode
- city-day
- city-month
- city-year
- facility type-time (complain)
- status-time (complain)
- channel type-time (complain)

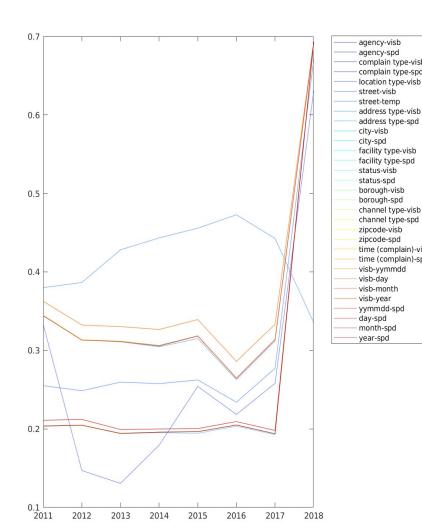
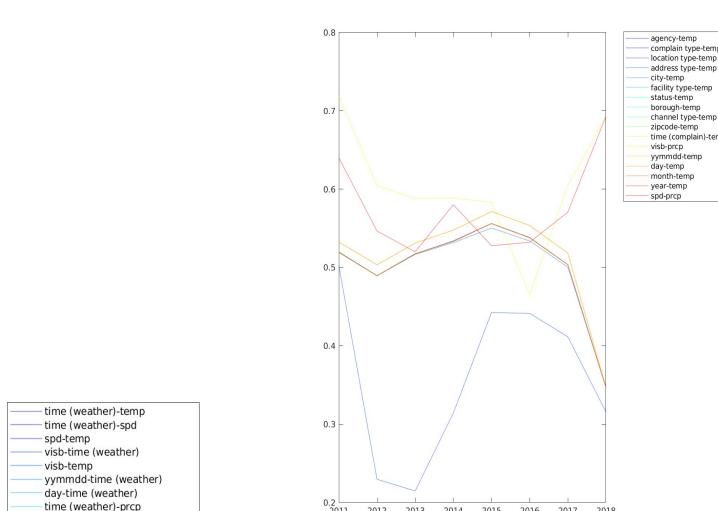
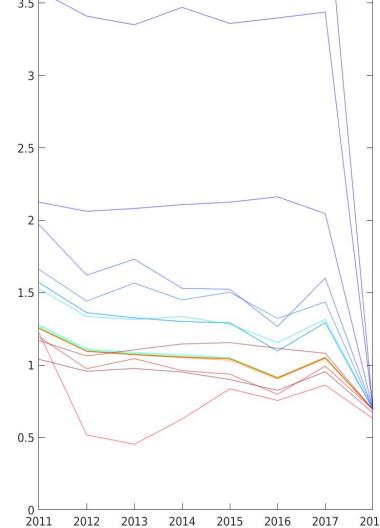
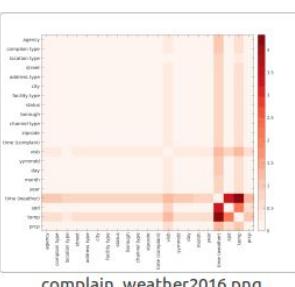
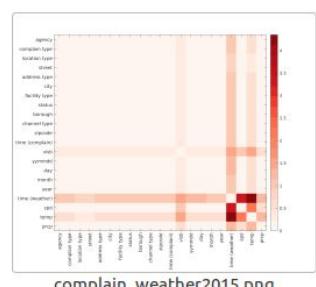
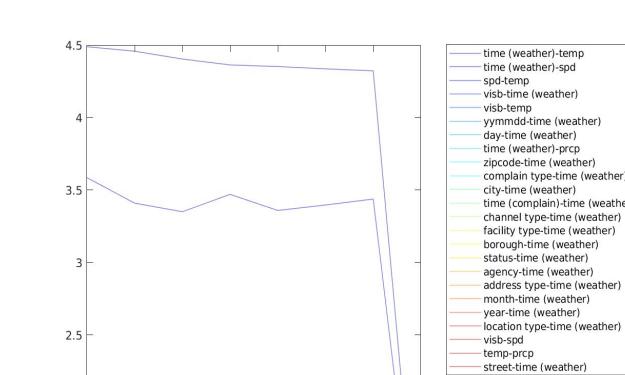
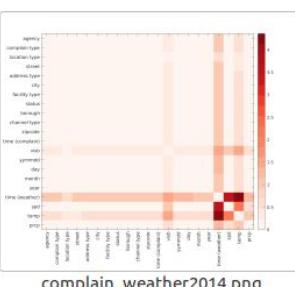
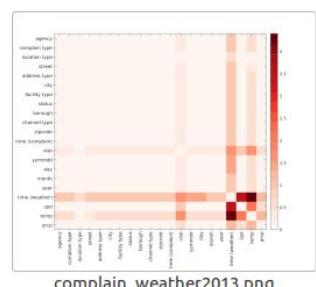
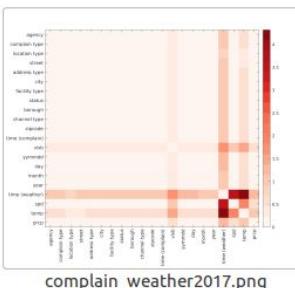
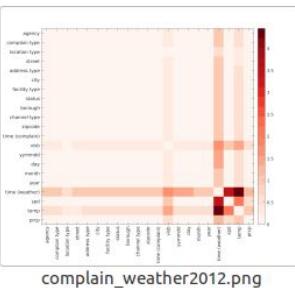
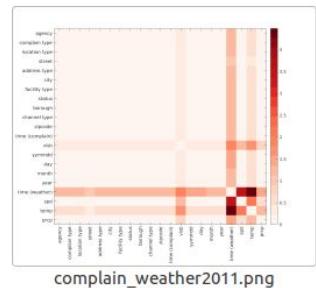


Legend for the line chart:

- agency-complain type
- agency-borough (complain)
- complain type-location type
- complain type-channel type
- complain type-status
- complain type-channel type
- complain type-offense code
- complain type-offense code2
- complain type-success code
- complain type-jurisdiction
- complain type-borough (crime)
- complain type-premise
- complain type-location desc
- complain type-zipcode (crime)
- complain type-ymmdd
- complain type-day
- complain type-month
- complain type-year
- location type-borough (complain)
- borough (complain)-channel type
- time (complain)-offense code
- time (complain)-offense code2
- time (complain)-offense level
- time (complain)-jurisdiction
- time (complain)-borough (crime)
- time (complain)-precinct
- time (complain)-location desc
- time (complain)-premise
- time (complain)-zipcode (crime)
- time (complain)-ymmdd
- time (complain)-day
- time (complain)-month
- time (complain)-year
- time (complain)-time

Supplementary

Complain weather 2011-2017



Supplementary

Weather Bike 2013-2018

