Data Science Course - HIT

FBI Crime Data Statistics

Presented By: Yonatan Avizov, Asaf Dangoor

The Project:

The project we have decided to do was to handle the Crime Stats Data given by the United States of America, by a service called the NIBRS.

National Incident-Based Reporting System (NIBRS) is an incident-based reporting system used by law enforcement agencies in the United States for collecting and reporting data on crimes. Local, state and federal agencies generate NIBRS data from their records management systems.

Back in 2014 they have created a suitable Crime Data API that lets you traverse their data.

The Project: (Continued)

We took that API and utilized it to our best efforts to answer Data Science related questions. The questions we went with in the end are the following:

1. Pie Graphs:
    1.1. How many and what is the rate of crimes for incidents involving people under the age of 18 per their crime location (Airport, nightclub, etc.).
    1.2. How many and what is the rate of crimes for incidents delegating per their crime location (Airport, nightclub, etc.).
2. Stacked Bar Graphs:
    2.1. Rate of crime per region versus the Race involved in the incidents.
3. Machine Learning: Predicting what gender would the criminal be per the crime reported.

While working on the project we found that some of our initial questions were not applicable, and so after combining the datasets in our project and finding what data we actually have access to we created the questions seen above.

The Project: Flow

The Crime Stat API has indeed a nice API that you can hook up to and eventually just run with, per commands, and build your own needed service. However, there is also a way for none devs to download the data fully in .zip files. And so, we used those downloads instead of the API itself, due to cleaner results.

And so, we have created a dynamic script that lets you place any range of years that the NIBRS has data on, along with the States you wish, and it will download the files (and eventually run the whole project with that data, creating the graphs).

In general we did:

- Downloaded the data from the website.
- Extracted the data into proper folders.
- Combined the extracted data into a new csv in a new folder per state.
- Ran different methods of graph creations per main csv file per year per state.
- Ran the ML question per the main csv file.

The Project: Data

The data provided by the NIBRS is .zip with basically lists of CSVs along with other lists of CSVs as “translators” to the coded data.

This is per

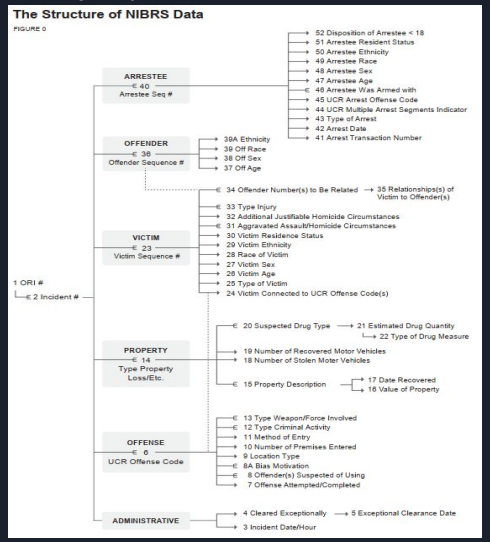
State per year.

WA-2013.zip

- nibrs_using_list.csv
- nibrs_victim.csv
- nibrs_victim_circumstances.csv
- nibrs_victim_injury.csv
- nibrs_victim_offender_rel.csv
- nibrs_victim_offense.csv
- nibrs_victim_type.csv
- nibrs_weapon.csv
- nibrs_weapon_type.csv

- agency_participation.csv
- cde_agencies.csv
- nibrs_activity_type.csv
- nibrs_age.csv
- nibrs_arrest_type.csv
- nibrs_arrestee.csv
- nibrs_arrestee_weapon.csv
- nibrs_assignment_type.csv
- nibrs_bias_list.csv
- nibrs_bias_motivation.csv
- nibrs_circumstances.csv
- nibrs_cleared_except.csv
- nibrs_criminal_act.csv
- nibrs_criminal_act_type.csv
- nibrs_diagram.pdf
- nibrs_drug_measure_type.csv
- nibrs_ethnicity.csv
- nibrs_incident.csv
- nibrs_injury.csv
- nibrs_justifiable_force.csv
- nibrs_location_type.csv
- nibrs_month.csv
- nibrs_offender.csv
- nibrs_offense.csv
- nibrs_offense_type.csv

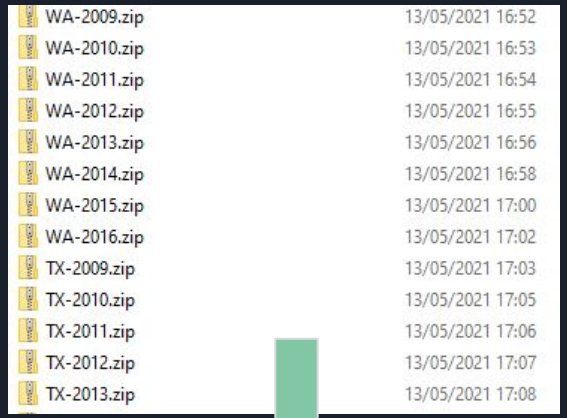
http://s3-us-gov-west-1.amazonaws.com/cg-d4b776d0-d898-4153-90c8-8336f86bdfec/nibrs_diagram.pdf

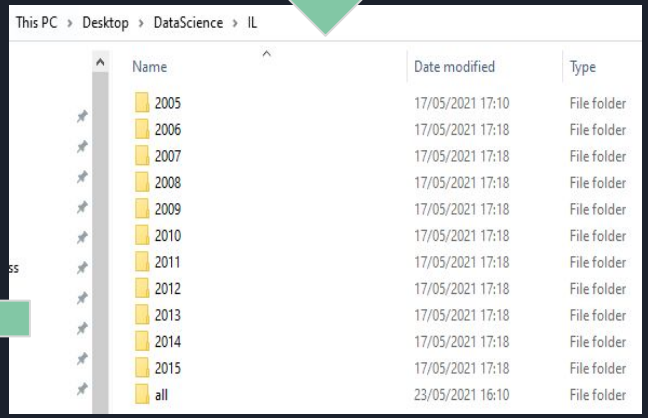


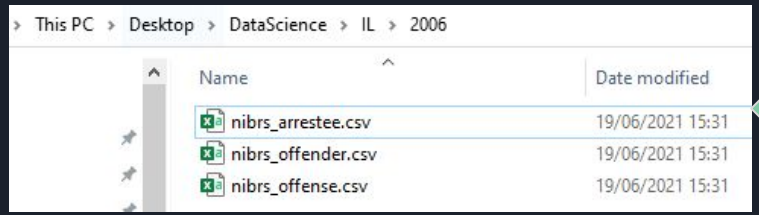
- nibrs_offense_type.csv
- nibrs_prop_desc_type.csv
- nibrs_prop_loss_type.csv
- nibrs_property.csv
- nibrs_property_desc.csv
- nibrs_relationship.csv
- nibrs_suspect_using.csv
- nibrs_suspected_drug.csv
- nibrs_suspected_drug_type.csv

Stage 1: Getting & Unpacking the Data

- We first found that we can download the .zip from the website itself - https://crime-data-explorer.fr.cloud.gov/pages/docApi
- However that wasn't sufficient for us, we wanted a dynamic system, so we created a way that took the download links of the amazon-cloud and would iterate over the given values of state and year.
- After we downloaded the zips we uncompressed individual files from the long list of csvs given to us by the download, and saved them to neat DIRS.







Stage 2: Processing the Data

- The data was far from ready, we removed everything that was not useful, and placed it in a new CSV that combined the 3 we extracted from the .zip and placed it in a “all” folder.
- We did this per each state and year.
- We then added more layers of information, translating the Coded columns to defined columns we used for graphs later on.

Nibrs_arresstee.csv (Given)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	arrestee_	incident_i	arrestee_s	arrest_nu	arrest_date	arrest_typ	multiple_	offense_t	age_id	age_num	sex_code	race_id	ethnicity_	resident_	under_18_	clearance_	ff_line_nu	age_range	age_range_high_num	
2	9632786	36400683	1	06-180410	20/11/2006 0:00	3	N	7	5	34	M	1	2	R						

Nibrs_offender.csv (Given)

	A	B	C	D	E	F	G	H	I	J	K	L
1	offender_	incident_i	offender_	age_id	age_num	sex_code	race_id	ethnicity_	ff_line_nu	age_range	age_range_high_num	
2	37839802	33724612	0									

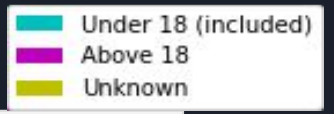
Nibrs_offence.csv (Given)

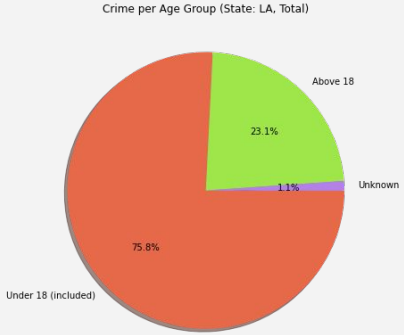
	A	B	C	D	E	F	G	H	I
1	offense_i	incident_i	offense_t	attempt_c	location_i	num_prer	method_e	ff_line_number	
2	35222665	33724612	45	C	20				

IL_2006.csv (Combined and Created)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	incident_id	age_num	sex_code	race_id	offense_t	location_i	arrest_date	ethnicity_	ethnicity_name	location_name	location_region_type	race_desc	offense_name	offense_category_name
2	33742048	19	F	2	45	24	19/04/2006 0:00	2	Not Hispanic or Latino	Specialty Store	Store	Black or African American	All Other Larceny	Larceny/Theft Offenses
3	33742096	13	M	2	39	22	20/04/2006 0:00	2	Not Hispanic or Latino	Educational Grounds/College	Educational Grounds	Black or African American	Weapon Law Violations	Weapon Law Violations

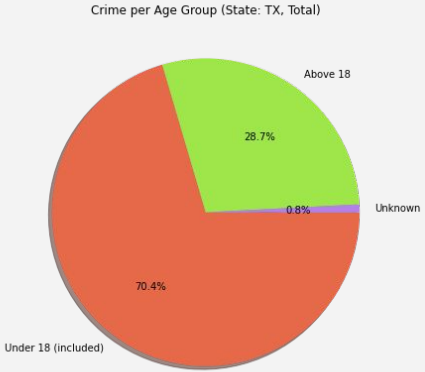
Results Q1.1: **Teenagers (age 18 or below) per their crime location**





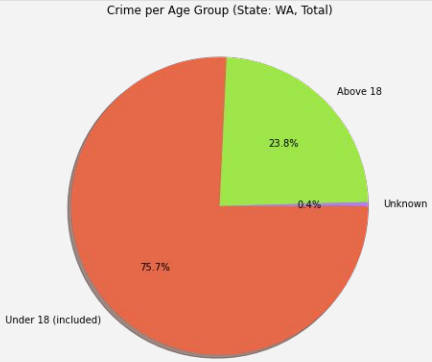
LA average

2005-2015



TX average

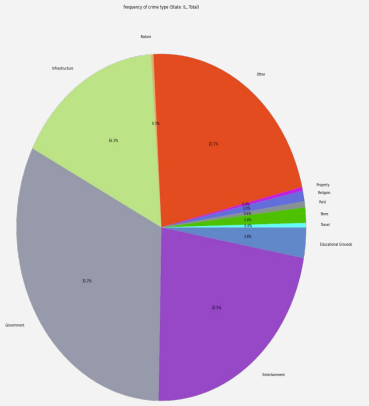
2005-2015



WA average

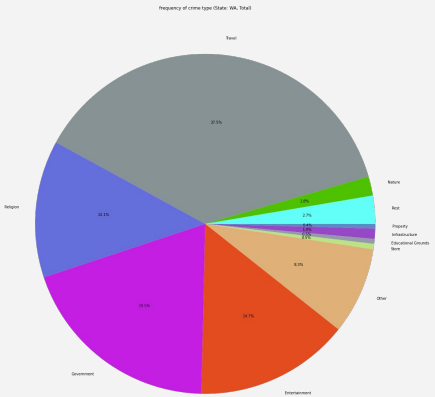
2005-2013

Results Q1.2 : **Rate of crimes per the Region**



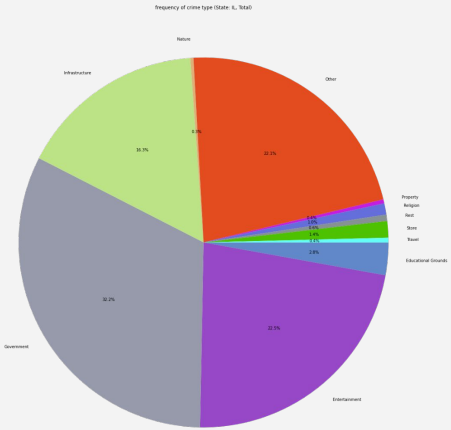
LA_freq

2005-2015



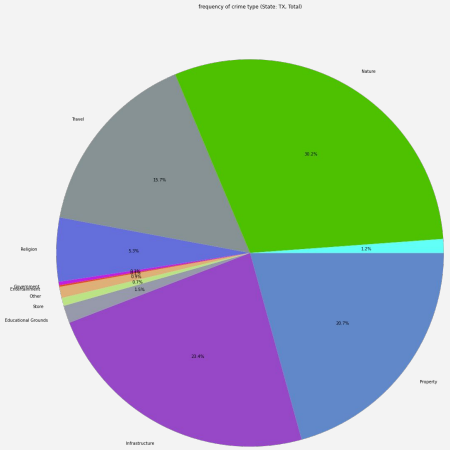
WA _freq

2005-2015



IL_freq

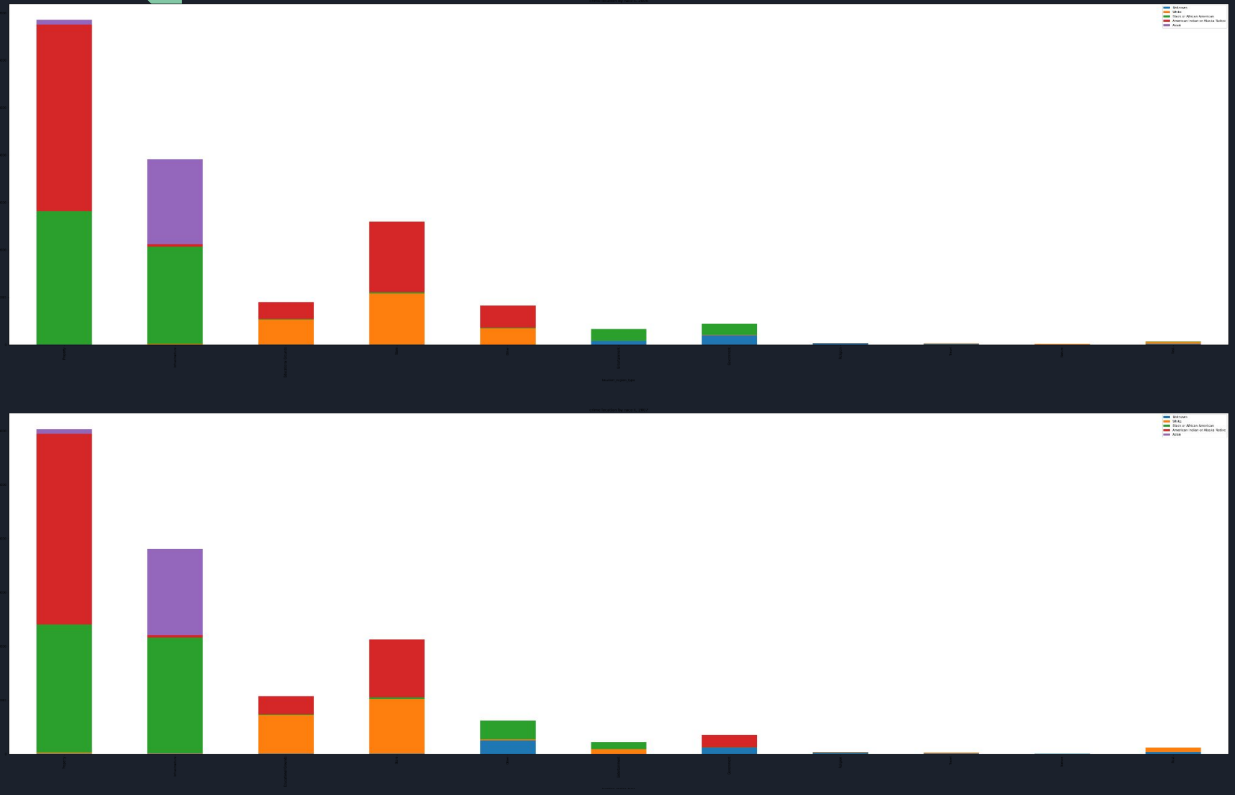
2006-2015

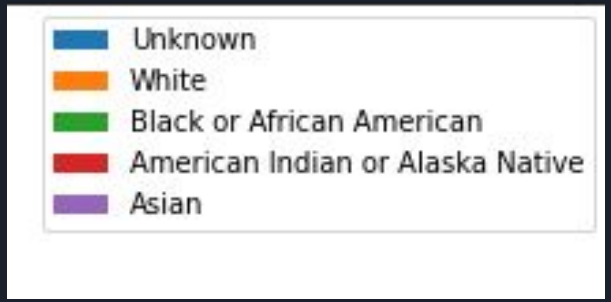


TX_freq

2005-2015

Results Q2: **The rate of crime per Region per Race**





The following is the example of TX in years 2005-2006.

For more examples see the attached images.

Results [ML]: Gender crime prediction

We tried to make a Logistic Regression ML that given the data from each state and year, would then try (with its given model) to predict if the crime was done by a Male or a Female.

We adjusted the data prior to running this for each state, but at the end we reached that our model is ineffective.

Most likely the reason is that there is no real correlation between what gender does the crime itself.

```
[[3585    0]
 [1123    0]]
[[3193    1]
 [1021    0]]
```

```
[0.4322922947063788, 0.4310205183585313, 0.42550540016615895, 0.42165709598031176, 0.42314679463268573, 0.42875508892221986, 0.4207068148
372766, 0.4272046589018303, 0.4167383820998279, 0.4320583794665324]
```

Summery

In general we learned a lot from this project, we had multiple issues at each step of the way but we managed to solve each of them on our own.

If we had more time we'd investigate more about the type of data and create a better ML and other niche things as said by this Crime Data Analyst: https://jacobdkaplan.com/data.html

Links we used:

https://www.fbi.gov/file-repository/ucr/ucr-2019-1-nibrs-user-manua-093020.pdf/view

https://crime-data-explorer.fr.cloud.gov/pages/home

https://www.icpsr.umich.edu/web/pages/NACJD/NIBRS/varlist.html

https://github.com/fbi-cde/crime-data-frontend

https://www.fbi.gov/file-repository/ucr/ucr-2019-1-nibrs-user-manua-093020.pdf/view