Satej Soman
CAPP30254: Machine Learning for Public Policy
Spring 2019

HW 1
DIAGNOSTIC ASSIGNMENT

1 Data Acquisition & Analysis

1.1 Chicago Open Data Portal

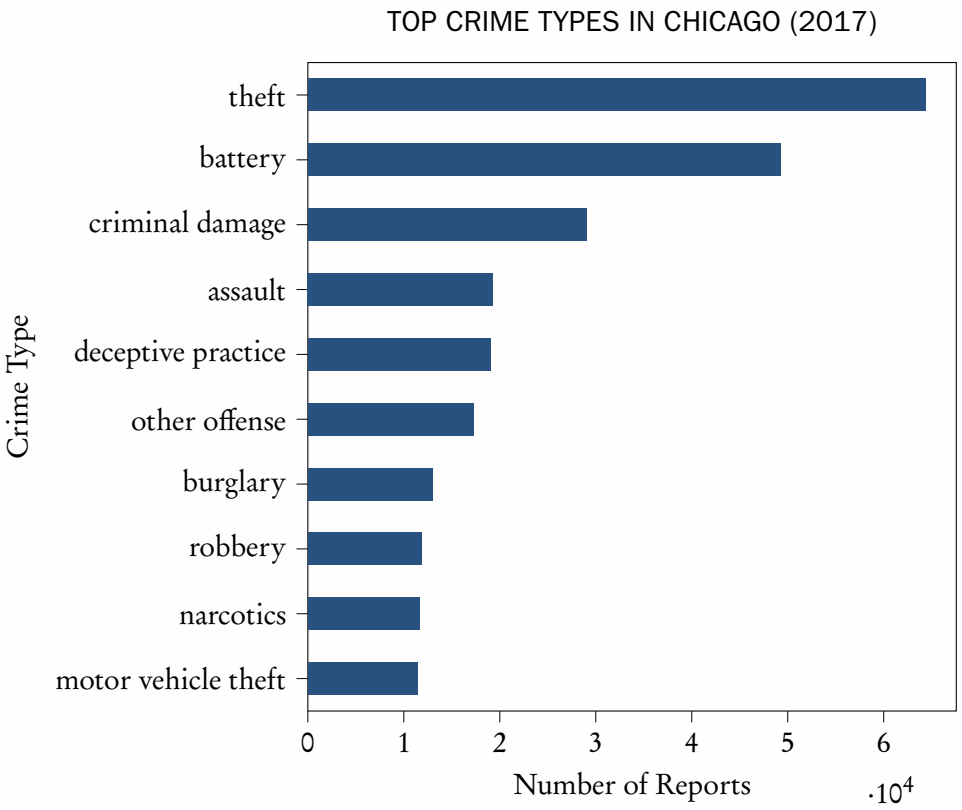
Chicago crime data is available, filtered by year, from the Chicago Data Portal (`https://data.cityofchicago.org/browse?category=Public%20Safety`). We can download this data and load it into a Pandas `DataFrame`:

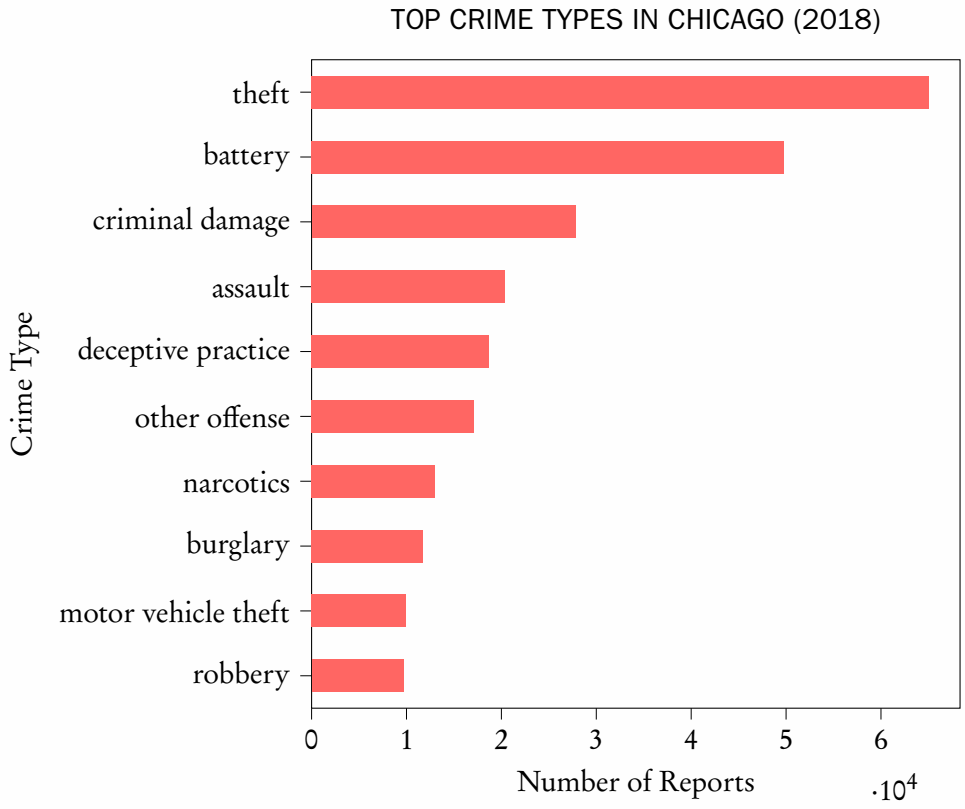
```
from pathlib import Path

import pandas as pd
import requests

# download crime data if we don't have it locally
base_url = "https://data.cityofchicago.org/api/views/{}/rows.csv?accessType=DOWNLOAD"
crime_resources = {
    2017: (Path("./crime_data_2017.csv"), "3i3m-jwuy"),
    2018: (Path("./crime_data_2018.csv"), "d62x-nvdr"),
}

for (year, (path, identifier)) in crime_resources.items():
    if not path.exists():
        url = base_url.format(identifier)
        print("{} data not found locally, downloading from {}".format(year, url))
        response = requests.get(url)
        with path.open("wb") as f:
            f.write(response.content)

crime_stats = pd.concat([
    pd.read_csv(crime_resources[2017][0]),
    pd.read_csv(crime_resources[2018][0])
])
```

1.2 Summary Statistics for Crime Report Data, 2017-2018

year	2017	2018	AVG
number of reported crimes	268094	266246	267170

year	2017	2018	OVERALL
crimes involving an arrest	19.53%	19.75%	19.64%
crimes considered domestic	15.90%	16.39%	16.14%


TOP CRIME TYPES IN CHICAGO (2017)
theft
battery
criminal damage
assault
deceptive practice
other offense
burglary
robbery
narcotics
motor vehicle theft
Crime Type
0
1
2
3
4
5
6
Number of Reports
·10^4


TOP CRIME TYPES IN CHICAGO (2018)
theft
battery
criminal damage
assault
deceptive practice
other offense
narcotics
burglary
motor vehicle theft
robbery
Crime Type
0
1
2
3
4
5
6
Number of Reports
·10^4

2 Data Augmentation & APIs

2.1 Chicago Crime Reports, Augmented with ACS Demographic Information

To pull in data from the American Community Survey, we need to identify which census tract each crime report corresponds to. This correspondence can be found by performing a *spatial join*: with shapefiles representing the geometry of Chicago-area census tracts, each crime report's latitude/longitude pair can be assigned to a census tract based on which polygon contains the report's coordinates. Census tract shapefiles are available from the City of Chicago's Data Portal.

```
from shapely.geometry import Point
import geopandas as gpd

def assign_census_tracts(crime_stats):
    boundary_shp = "./Boundaries - Census Blocks -
        2000/geo_export_8e9f6d85-3c5b-429f-b625-25afcc3dea85.shp"
    census_tracts = gpd.read_file(boundary_shp).drop(columns=["perimeter", "shape_area",
        "shape_len"])
    # restrict geocoding to valid locations
    crime_stats = crime_stats[crime_stats["Location"].notna()]
    crime_stats["geometry"] = crime_stats.apply(lambda row: Point(row["Longitude"],
        row["Latitude"]), axis = 1)
    return gpd.tools.sjoin(gpd.GeoDataFrame(crime_stats), census_tracts, how="inner")
```

2.1.1 What types of blocks have reports of "Battery"?

2.1.2 What types of blocks get "Homicide"?

2.1.3 Does that change over time in the data you collected?

2.1.4 What is the difference in blocks that get "Deceptive Practice" vs "Sex Offense"?

3 Analysis & Communication

3.1 Changes in Crime, 2017-2018

3.2 Analysis of Jacob Ringer's Claims

To evaluate Jacob Ringer's claims, we can filter down the relevant ward (the 43rd ward), and isolate the time periods he analyzes.

```
crime_stats_w43 = crime_stats[crime_stats["Ward"] == 43]
```

3.3 Key Findings

3.4 Caveats & Limitations

4 Probability Exercise

a)

b)

c)

Sources