Intro to Big Data Science: Project

Due Date: May 28, 2021

Problem(Data Science for DC Crime)

The data "DC_Crime.csv" and their description can also be found at:

https://dcatlas.dcgis.dc.gov/crimecards/

Crime is a common social problem in the modern human societies. It has a lot to do with economy, culture, politics, technology, and people's happiness. In this project, you will be provided the crime data in the DC area from 2008 to 2017. More complete data can be downloaded from the above website. You can play with these data and uncover much underlying information using data mining techniques. In the following, let us give you some explanations about the data.

All statistics presented here are based on preliminary DC criminal code offense definitions. All preliminary offenses are coded based on DC criminal code and not the FBI offense classifications.

On February 1 2020, the methodology of geography assignments of crime data was modified to increase accuracy. From January 1 2020 going forward, all crime data will have Ward, ANC, SMD, BID, Neighborhood Cluster, Voting Precinct, Block Group and Census Tract values calculated prior to, rather than after, anonymization to the block level. This change impacts approximately one percent of Ward assignments.

Feature description:

- 1) NEIGHBORHOOD_CLUSTER
 - what neighborhood cluster the case belongs to
 - Example: cluster 21
- 2) CENSUS_TRACT

- part of block group index
- Example: 008702
- 3) offensegroup
 - what offense group the case belongs to
 - Example: property
- 4) LONGITUDE
 - longitude
 - Example: -77.0035742966363
- 5) END_DATE
 - what date the case ended at
 - Example: 2017-04-29T08:00:23.000
- 6) offense-text
 - text form offense info
 - Example: theft f/auto
- 7) SHIFT
 - the shift of the case report time
 - Example: day, evening, midnight
- 8) YBLOCK
 - block y index
 - Example: 138139
- 9) DISTRICT
 - district index
 - Example: 5
- 10) WARD
 - one kind of geographic info
 - Example: 5
- 11) YEAR
 - year
 - Example: 2017
- 12) offensekey
 - offense group | offense
 - Example: property|theft f/auto
- 13) BID

- one kind of geographic info
- Examples: noma, adams morgan, downtown

14) sector

- sector index
- Example: 5D1

15) PSA

- Police Station Area index
- Example: 502

16) ucr-rank

- UCR-Rank (crime severity rank) of 1-9
- Example: 7

17) BLOCK_GROUP

- block group index
- Example: 0087022

18) VOTING_PRECINCT

- one kind of geographic info
- Example: precinct 75

19) XBLOCK

- block x index
- Example: 399690

20) BLOCK

- block info of the case
- Example: 150 299 block of q street ne

21) START_DATE

- case start date
- Example: 2017-04-29T01:30:14.000

22) CCN

- Criminal Case Number
- Example: 17070672

23) OFFENSE

- what kind of offense
- Example: theft f/auto
- 24) OCTO_RECORD_ID

- Office of the Chief Technology Officer (OCTO) record id
- Example: 17070672-01
- 25) ANC
 - one kind of geographic info
 - Example: 5E
- 26) REPORT_DAT
 - case report date
 - Example: 2017-04-29T13:49:31.000Z
- 27) METHOD
 - what method is used in the case
 - Examples: gun, others
- 28) location
 - (latitude, longitude)
 - Example: "38.911121322949178,-77.003576581965632"
- 29) LATITUDE
 - latitude
 - Example: 38.9111135327066

Supplementary materials:

You may also want to know the relationship between the criminal circumstances and the economics in DC. Here we also provide you the housing data in DC with geographic information and other housing related information. You can combine the two datasets by connecting their geographic information and time information. Then you will find the relationship between the crimes and the housing prices. This may help you to dig into more details about the economic behavior and the social behavior. In the following, we will show you the feature description of the housing data.

'DC_Properties.csv':

- 1) BATHRM
 - Number of Full Bathrooms
 - Example: 4
- 2) HF_BATHRM
 - Number of Half Bathrooms (no bathtub or shower)
 - Example: 0
- 3) HEAT
 - Heating

- Example: Warm Cool
- 4) AC
 - Cooling
 - Example: Y
- 5) NUM_UNITS
 - Number of Units
 - Example: 2.0
- 6) ROOMS
 - Number of Rooms
 - Example: 8
- 7) BEDRM
 - Number of Bedrooms
 - Example: 4
- 8) AYB
 - The earliest time the main portion of the building was built
 - Example: 1910.0
- 9) YR_RMDL
 - Year structure was remodeled
 - Example: 1988.0
- 10) EYB
 - The year an improvement was built more recent than actual year built
 - Example: 1972
- 11) STORIES
 - Number of stories in primary dwelling
 - Example: 3.0
- 12) SALEDATE
 - Date of most recent sale
 - Example: 2003-11-25 00:00:00
- 13) PRICE
 - Price of most recent sale
 - Example: 1095000.0
- 14) QUALIFIED
 - Qualified

- Example: Q
- 15) SALE_NUM
 - Sale Number
 - Example: 1
- 16) GBA
 - Gross building area in square feet
 - Example: 2522.0
- 17) BLDG_NUM
 - Building Number on Property
 - Example: 1
- 18) STYLE
 - Style
 - Example: 3 Story
- 19) STRUCT
 - Structure
 - Example: Row Inside
- 20) GRADE
 - Grade
 - Example: Very Good
- 21) CNDTN
 - Condition
 - Example: Good
- 22) EXTWALL
 - Extrerior wall
 - Example: Common Brick
- 23) ROOF
 - Roof type
 - Example: Built Up
- 24) INTWALL
 - Interior wall
 - Example: Hardwood
- 25) KITCHENS
 - Number of kitchens

- Example: 2.0

26) FIREPLACES

- Number of fireplaces
- Example: 5

27) USECODE

- Property use code
- Example: 24

28) LANDAREA

- Land area of property in square feet
- Example: 1680

29) GIS_LAST_MOD_DTTM

- Last Modified Date
- Example: 2018-07-22 18:01:43

30) SOURCE

- Raw Data Source
- Example: Residential

31) CMPLX_NUM

- Complex number
- Example: 1066.0

32) LIVING_GBA

- Gross building area in square feet
- Example: 888.0

33) FULLADDRESS

- Full Street Address
- Example: 1748 SWANN STREET NW

34) CITY

- City
- Example: WASHINGTON

35) STATE

- State
- Example: DC

36) ZIPCODE

- Zip Code

- Example: 20009.0

37) NATIONALGRID

- Address location national grid coordinate spatial address
- Example: 18S UJ 23061 09289

38) LATITUDE

- Latitude
- Example: 38.91468021

39) LONGITUDE

- Longitude
- Example: -77.04083204

40) ASSESSMENT_NBHD

- Neighborhood ID
- Example: Old City 2

41) ASSESSMENT_SUBNBHD

- Subneighborhood ID
- Example: 040 D Old City 2

42) CENSUS_TRACT

- Census tract
- Example: 4201.0

43) CENSUS_BLOCK

- Census block
- Example: 004201 2006

44) WARD

- Ward (District is divided into eight wards, each with approximately 75,000 residents)
- Example: Ward 2

45) SQUARE

- Square (from SSL)
- Example: 0152

46) X

- longitude
- Example: -77.04042907495098
- 47) Y

- latitude

- Example: 38.914881109044266

48) QUADRANT

- City quadrant (NE,SE,SW,NW)

- Example: NW