SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Intro to Big Data Science: Project

Due Date: May 30, 2024

The students can choose one from the two problems as their course project topics.

✏ **Problem 2** (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: 008702 2
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

Tasks:

1. Task 1: Besides the datasets we provide you here, you can also find other relevant datasets by yourself, if you think these datasets are helpful for your analysis.

2. Task 2 (Mandatory): You need to first get familiar with the data by data statistics and visualization.

   1) You are asked to first do the data preprocessing for all the data (include the relevant data you found). There may be redundancy in the data so that you have to process it at the beginning. Then visualize the data using python package 'matplotlib' and 'seaborn'. For instance, the histograms across the attributes 'offense', 'year', etc. You can also plot the histograms based on the geographic info. Based on these plots, can you find any interesting relations and draw any conclusions?

      数据预处理

   2) Please also analyze the correlation between different features, such as the correlation between 'shift' and 'offense', and the correlation between 'offense' and 'method'. By dividing the time series into several parts, judge whether the correlations change are evident with the time changes.

      属性相关性分析

   3) Moreover, you are asked to find the correlation between the crime events and geographic locations. You are also encouraged to analyze the changes of crime events in both time and space.

      犯罪事件在时间和空间上

   4) How does the number of crimes vary geographically and temporally? As a warm-up for the next task, we suggest you to visualize the number of crimes according to the geographical districts. You can first divide the DC area into several disjoint subareas according to either administrative district or other types of geographical districts defined by yourselves. Then you may count the number of crimes in each subarea in a certain time period and plot them on the geographical map (e.g., you may plot heatmap). Please also show your plot for some different time, seasons, years, etc.

      地理区域与犯罪数量的关联

3. Task 3 (Mandatory): classify (or cluster) the geography by the crime events.

   According to the distance function defined by yourself, divide the block/location information into several categories. You are asked to use classification or clustering methods such as kNN/KMeans. Encourage multiple methods apply to this task.

   将地区分类

4. Task 4 (Mandatory): predict the housing price.

   Is there any correlation between the crime and the housing price? We provide you an additional dataset named “DC_Properties.csv” which records the housing price and related information in the DC area. In particular, sufficient geographical information (such as latitude, longitude, zipcode, census_tract, census_block, ward, etc) and temporal information (such as ayb, eyb, saledata, etc.) are provided so that you can make a correspondence between the two datasets. Please combine the two datasets to make a new dataset with new features that you can define by yourselves. Remember to keep the most important information in the Crime data and the Properties data so that you can get everything to predict the housing price. The prediction can be made by using regression methods. Try what you learned in the course to make the regression as good as possible.

   犯罪率和房价

5. Task 5: Extra data analysis.

   Hereby we give you some suggestions. For instance, you may predict the number of crimes in the future (or for certain types of crimes). You can divide different prediction models according to the space-time grid. Either standard regression or other methods based on time series analysis can be used.

   You can also find the correlation between crime situation and local economic status/housing price (find datasets by yourself). We encourage you do this task by using PCA to reduce dimensionality or boosting method to reduce bias, etc.

When you do the above tasks, you should visualize them appropriately, show the results, and summarize your conclusions. We hope you can get some interesting conclusions in this project.

Submission policy: This project can be finished either individually or in groups of two partners. You can find your partners by yourselves. Each student should demonstrate which role he/she plays in the team.

Your report should be in PDF format. It is suggested to use jupyter notebook to include both your code and your main text, and convert it to PDF file. Your main text should include the following several aspects:

- Background introduction
- Data exploration: data statistics and data visualization;
- Data preprocessing: detecting missing values and outlier samples (if any), data discretization, concatenation, and normalization (if necessary), etc.;
- Model construction: you could use any model you prefer, even the model we did not cover in class;
- Feature selection and model selection if necessary;
- Model evaluation;
- Conclusion.

DO NOT just submit the code file. Necessary statements, analysis, formula, figures, and tables should be included in your report. You should also have a complete set of codes. Please pack all your necessary documents in a compressed file (e.g., zip file), and **use your student ID and name to rename your zip file**, e.g., “12000000_ 张三”. Then the zip file shall be sent to the TA or the course instructor.

There will a presentation about the project at the end of this semester.

Last but not the least, we hope you really get familiar with whole data science procedure and discover the new world of your own.

Now enjoy your journey of data science!