SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Intro to Big Data Science: Project

Due Date: May 28, 2021

✏ 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?
    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.
    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.
    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.

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.

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.

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.