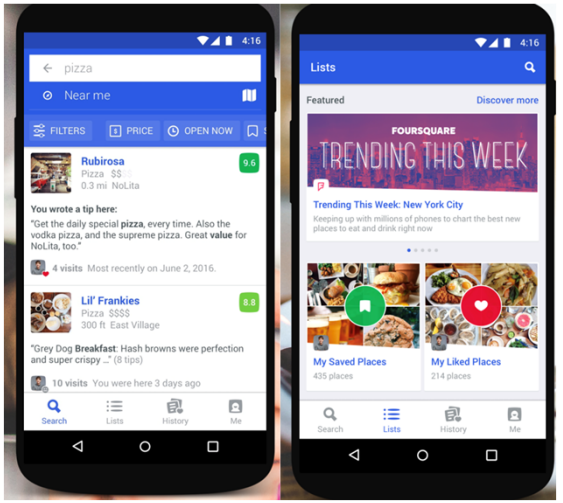
MATLAB程序设计
大作业基本要求及说明

- **数据集**：给大家提供了两个数据集，第一个是Foursquare平台的用户行为数据，第二个是近两年部分国家地区新冠疫情数据。大家可在两个数据集任选其一，利用课程中学习到的相关知识，选择切入点，对数据进行分析。
- **分组**：可单独一人一组，也可两人一组。对于包含两人的小组，两人均需要根据各自的贡献完成相应的小论文。
- **成果**：以小论文形式提交，论文中应包括整体分析思路、数据分析图表及对应文字说明，并附上对应程序的源代码
- 分组和选题情况请在2022年5月21日之前告知任课教师和助教(届时将在课程微信群进行接龙)；大作业需要在2022年5月28日23:59之前通过elearning平台提交。
- 答辩：为了更好地了解大家的工作贡献，将在2022年5月30日组织进行每组5分钟的答辩，同学们会有1-2分钟的陈述时间（可做ppt），之后会由任课教师和助教向同学们提问。

数据集1：Foursquare

简介：

Foursquare是一个基于用户地理位置的手机APP，它支持手机用户同他人进行社交互动并分享其针对某个地点的评论等功能。



本数据集主要由3部分组成——

- 用户信息：users.csv
- 用户评论信息：tips.csv
- 用户好友关系信息：friends.csv

数据集1：Foursquare——用户信息

- 文件名：users.csv
- 每行记录代表一个用户的基本信息（共计999129条），包括12个字段：
  - user_id：用户的唯一标识ID
  - lastname：用户是否公开lastname，若公开为1，否则为0
  - gender：用户性别，若男性为0，女性为1，未公开为2
  - facebook：用户是否公开其关联的Facebook账号，若关联为1，否则为0
  - twitter：用户是否公开其关联的Twitter账号，若关联为1，否则为0
  - bio_len：用户公开的自我介绍中包含的单词个数
  - tips_count:用户历史发表的tips的数量（tips = 评论）
  - lists_count：用户历史发表的lists的数量（lists = 推荐商户列表）
  - checkin_count：用户历史发表的checkins的数量（checkin = 实时签到）
  - friend_count：用户好友数量
  - homecity_public：用户是否公开其homecity,若公开为1，否则为0
  - homecity：若用户公开其homecity，该字段为其公开值，否则为""

数据条目示例：

```
32, 0 ,0 ,1, 1, 116, 624, 116, 12633, 833, 1, "New York, NY"
```

数据集1：Foursquare——用户评价信息

- 文件名：tips.csv
- 每行记录代表一条商户评论(tip)的信息（共计6245413条），包括7个字段：
  - user_id：评论发布者的用户ID
  - venue_name：被评价商户的名称
  - agree_count：该评论收到“like”评价的次数
  - disagree_count：该评论收到“dislike”评价的次数
  - textlen：该评论单词个数
  - date：该评论发布时间(北京时间，UTC+8)
  - photo：该评价是否带有照片，若带有照片为1，否则为0
  - category：地点类别
  - country：地点所在国家或地区缩写

数据条目示例：

```
32, Maiden Lane, 2, 0, 112, 2018-10-11 23:03, 0, Nightlife_Spot, US
```

数据集1：Foursquare——用户好友关系信息

- 文件名：friends.csv
- 每行记录代表两个对应用户之间的双向好友关系（共计19583857条），共2个字段：
  - user1：好友关系中第一个用户的user_id
  - user2：好友关系中第二个用户的user_id

数据条目示例：

32, 75555851

⚠ **friends.csv和users.csv中friend_count的关系：**

- users.csv中的friend_count为用户在整个Foursquare APP中的好友总数量；
- friends.csv所记录的每一条好友关系，都必须满足两个用户都在users.csv当中。

数据集2：新冠疫情信息

简介：

2019冠状病毒病疫情是由严重急性呼吸系统综合征冠状病毒所引发的全球大流行疫情。本数据集是由约翰霍普金斯大学收集的21个国家或地区每日疫情相关信息（截止至2022年4月22日）。

本数据集主要由2部分组成——

- 21个国家或地区的每日疫情信息
- 229个国家或地区的基本情况

数据集2：新冠疫情——每日疫情信息

- 文件名：daily_info.csv
- 包含来自五大洲的21个国家的每日疫情信息（共计16883条），包括13个字段：
  - location：国家
  - continent：所属大陆
  - date：日期
  - new_cases：当日新增病例
  - total_cases：截止至当日的累计病例
  - new_tests：当日新增新冠检测人数
  - total_tests:截止至当日的累计新冠检测人数
  - positive_rate:检测人数中的阳性率（最近七日平均值）
  - new_vaccinations：当日新增疫苗接种
  - total_vaccinations：截止至当日的累计接种人数
  - hosp_patients：新冠患者住院人数
  - icu_patients：新冠患者住院（ICU）人数
  - stringency_index：政府的管控程度，0~100之间的数值，100代表最严格

数据条目示例：

Australia, Oceania, 2021-10-27, 2038, 165904, 195419, 42713811, 0.0121, 217681, 34943366, 1151, 226, 63.43

Malaysia, Asia, 2020-12-15, 1772, 86618, 46323, 3620341, 0.0389, , , 2349, 153, 67.13 (缺漏疫苗接种信息)

数据集2：新冠疫情——国家基本信息

- 文件名：country.csv
- 包含229个国家的基本信息（共计229条），包含13个字段：
  - location：国家
  - continent：所属大陆
  - population：人口
  - population_density：人口密度（/km^2）
  - median_age：国民年龄中位数
  - aged_65_older：大于65岁的人口比例
  - aged_70_older：大于70岁的人口比例
  - gdp_per_capita：国民生产总值（按购买力平价，constant 2011 international dollars）
  - male_smokers：男性中的吸烟比例
  - female_smokers：女性中的吸烟比例
  - cardiovasc_death_rate：心血管疾病死亡率（平均每年、每100,000人）
  - hospital_beds_per_thousand：医院床位数（平均每1,000人）
  - life_expectancy：预期国民寿命

数据条目示例：

Austria, Europe, 9043072, 106.749, 44.40, 19.202, 13.748, 45437, 30.9, 28.4, 145.183, 7.37, 81.54

Jordan, Asia, 10269022, 109.285, 23.2, 3.81, 2.361, 83375, , , 208.257（缺漏男/女吸烟比例信息）

数据集2：数据细节

1. 数据集中存在部分缺漏的数据，表示该数值没有被测量或未被成功记录；

2. 数据由约翰霍普金斯大学统计，由于统计口径等原因，和各国官方公布的每日数据可能存在一定偏差。以中国为例，其病例数据包括了确诊患者和无症状感染者，以及2020年的疑似病例。

数据读取方式（仅供参考）

- 可以使用readtable命令对数据进行读取

```
users=readtable('./users.csv');
```

- readtable提供使用opts即文件导入选项对部分数据进行读取

```
opts = detectImportOptions('./users.csv');
% opts包含控制数据导入过程的属性
opts.DataLines = [100000 100200];
% 控制数据导入为第100000到100200行
opts.SelectedVariableNames = {'user_id'};
% 控制数据导入的VariableNames为'user_id'
% 更多的opts属性可参考MATLAB官网
users=readtable('./users.csv',opts);
% 按照opts导入数据，数据格式为table
```

```
opts =

  DelimitedTextImportOptions - 属性:

  格式 属性:
                    Delimiter: {','}
                   Whitespace: '\b\t '
                   LineEnding: {'\n'  '\r'  '\r\n'}
                CommentStyle: {}
    ConsecutiveDelimitersRule: 'split'
        LeadingDelimitersRule: 'keep'
                EmptyLineRule: 'skip'
                     Encoding: 'GB2312'

  替换 属性:
                  MissingRule: 'fill'
              ImportErrorRule: 'fill'
            ExtraColumnsRule: 'addvars'

  变量导入 属性:
                VariableNames: {'Var1', 'user_id', 'lastname' ... and 10 more}
                VariableTypes: {'double', 'double', 'double' ... and 10 more}
        SelectedVariableNames: {'Var1', 'user_id', 'lastname' ... and 10 more}
              VariableOptions: Show all 13 VariableOptions
        Access VariableOptions sub-properties using setvaropts/getvaropts

  位置 属性:
                    DataLines: [2 Inf]
           VariableNamesLine: 1
              RowNamesColumn: 0
           VariableUnitsLine: 0
     VariableDescriptionsLine: 0
        要显示该表的预览，请使用 preview
```

注意事项

1. 不允许使用非MATLAB的工具或语言（例如C++或Python）进行数据分析，所有的数据分析、作图必须有相应的MATLAB代码。

2. 数据集获取方式：elearning