数据可视化（CSE578）  
王毅Arizona State University[ywan1369@asu.edu](mailto:ywan1369@asu.edu)

1、 项目目标这是针对 2023 年夏季学期数据可视化的（CSE511）的项目报告。  
在该项目中，我们以小组为单位，总计需要完成12组针对不同场景的的数据可视化组合，从而以支持完成项目目标（详见xyz公司的项目目标说明）。

具体而言，则是小组成员分工合作，从课程提供的数据文件种选择九种数据（共计提供15种数据—其中收入（income）为必须选择的数据），既是需要另外选取八种数据，基于这些数据，再结合课程中所学的可视化的具体知识，将这些数据经过合理的处理后，采用所学习的可视化图形将其展示出来，从而达成通过这些数据可视化能够预测人员的收入的项目终极目的。

2、 数据选择

通过对原始数据文件的简单初略的查阅和分析，可以发现，数据文件一共包含15种数据，分别为"age"（年龄）, "workclass"（工作类型）, "fnlwgt"（不理解）, "education"(教育), "education\_num"（教育年限）, "marital\_status"（婚姻状况）,

"occupation"（职业）, "relationship"（婚姻中的角色）, "race"（种族）, "sex"（性别）, "capital\_gain"（资本收益）, "capital\_loss"（资本损失）,"hours\_per\_week"（每周工作时间）, "native\_country"（出生国籍）, 以及"income"（收入），经过简单分析后，我们选择以下数据类型来进行数据分析和可视化展示：

1. 组员1，采用以下3类型数据
2. 组员4（YiWang）,选用了以下3组数据，并基于此3组数据继续分析
3. 基于所学的词云图形展示职业（occupation）和收入（income）之间的关系
4. 基于所学的堆叠条形图，展示sex（性别）,education（教育）,hours\_ per week（每周工作时间） 之间的关系
5. 基于所学的气泡图（一种变形的散点图），展示婚姻状况（marital\_status），工作类型（workclass）,年龄（age）之间的关系

3、 用户场景

1. 场景
2. 通过数据的简单探索，认为基于以上所选的3种数据组合类型，通过对其分析和数据可视化分析后；应当可以有如下所示的结论：

不同的职业，会有不同的收入分配，高收入者分布于一种或者某几种确定的职业；

具有不同教育程度，以及不同的性别的每周工作时间长度应当是不同的，必然有一些教育背景下的人群，其周工作时间长度是比较短的；

不同的年龄，不同的工作类型的人群的婚姻状况，具有显著的较大差异，根据经验，一般认为，工作类型不太稳定（比如在私有企业工作）的年轻人群一般属于未婚状态。

4、 可视化实现

基于以上用户场景的设计和数据类型的选取，这里具体加以实现。对于具体的的可视化的实现，我们基于python语言进行实现。数据的基本处理由xx已经完成，主要将数据集中的无线数据剔除，并为数据集添加数据类型名。

1、

2、

3、

4、组员4（YiWang）,对于选取的3组数据的具体实现如下：

a. 词云展示职业（occupation）和收入（income）。将数据按照>50k和≤50K区分后，通过词云图形，分别统计不同的职业的数据总量，然后将数据整理后，再通过词云进行展示出来，其效果如下图所示

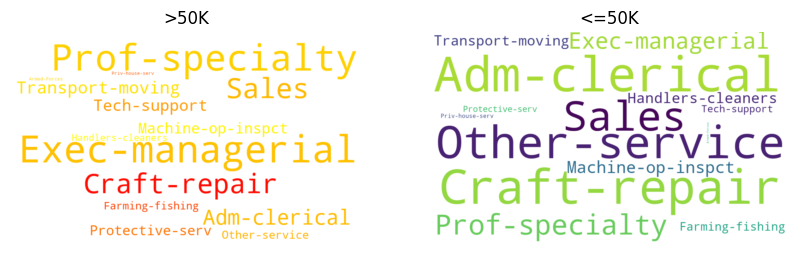


图1

如图1所示，我们可以明显地看到EXEC-MANAGERRIAL,PROF-SPECIALTY等职业在>50k的人群中的占比是比较高的，说明这几种职业的高收入人群比例最大。

b. 堆叠条形图，展示sex(性别), education(教育), hours\_ per week(每周工作时间)。将数据按照

男性、女性进行划分，然后分别统计不同教育的人群的周平均工作时间，然后按照教育，将男性和女性周平均工作时间合并为一条条形图进行展示。其所示的效果图如下

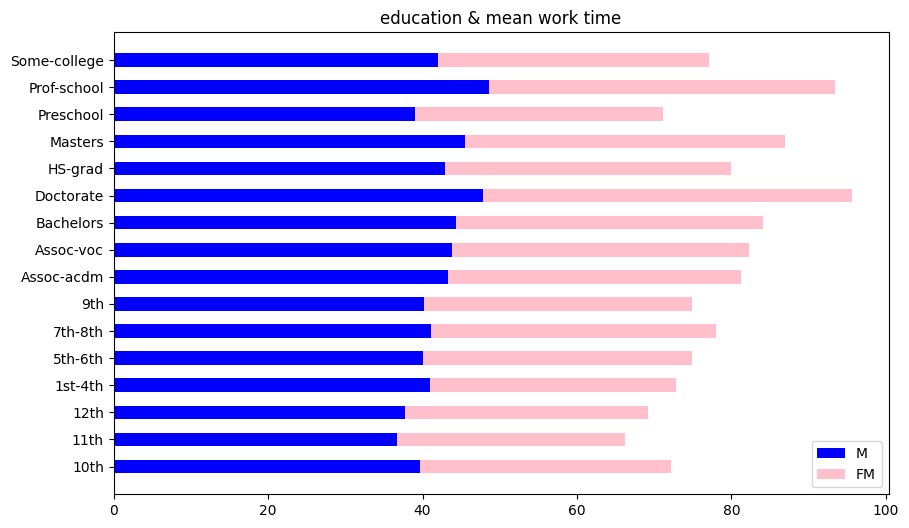


图2

如图2所示，我们可以明显地看到11th,12th的人群的工作时间最短，doctorate的人群的工作时间最长，总体而言，男性的工作时间比女性较长。

c. 气泡图（一种变形的散点图），展示婚姻状况（marital\_status），工作类型（workclass）, 年龄（age）。将数据按照婚姻状况和工作类型进行分类，并计算各类人群的平均年龄，之后将整体人群按照年龄划分为3个大的群体，年轻，中等，偏大。然后，将其通过气泡图进行展示，其展示的效果图如下所示

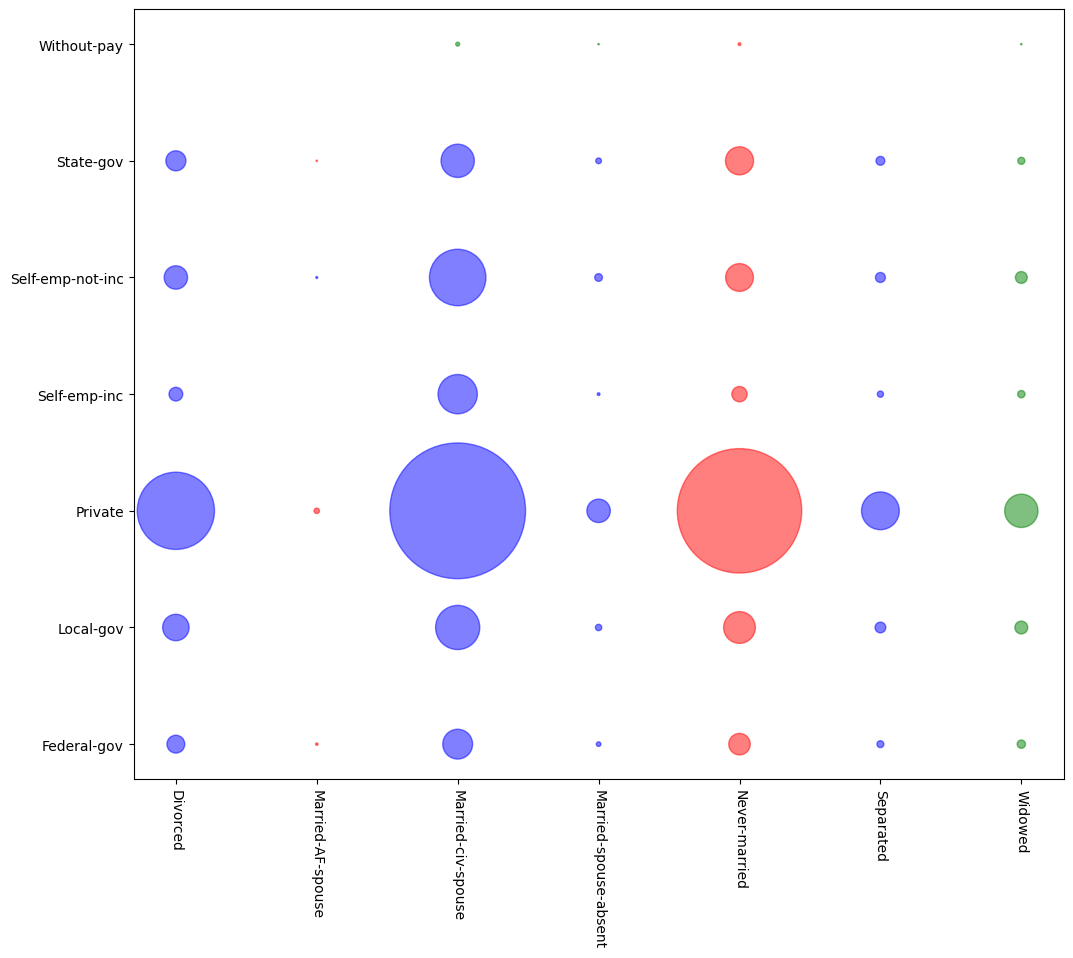


图3

如图3所示（由于办事原因，word种显示不佳，可具体参看代码文件中的图示）。

，其中，横轴所示为婚姻状况，纵轴所示为工作类型，气泡大小代表着人数的多少，颜色代表年龄分布（红色：年轻；蓝色：中等；绿色：偏大（年龄划分较为粗略））。

通过以上图示，我们可以发现，private的人群最多，而这个群体中，年轻人基本为未婚状态，中年人大部分为married-civ-spouse。

基于以上3种数据组合的简单分析和可视化，我简单认为，应当选取职业为

EXEC-MANAGERRIAL,PROF-SPECIALTY的教育程度为11th,12th的工作性质为private中的年轻人群体为最佳目标客户，因为这样的人群具有较高的收入，较多的自由时间，同时又没有家庭的束缚，无疑是满足目标的最佳选择。

5、 挑战

无疑是满足目标的最佳选择。数据集中包含大量的数据类型，应该当选用哪些数据机型数据可视化分析，另外连续性的数据量比较少，大量的数据属于定类数据，对于多变量的数据可视化具有一些困难。

首先，对于数据的选择，首先将数据集中的数据加以简单分析，在去除必要和不理解的数据之后，在结合场景的设定，适当的选取将要使用的数据；

其次，对于数据的可视化的展示的过程中，可能需要首先将数据通过python进行合理的处理之后，然后再根据新的数据的特点，再结合可视化图形的特点，最后才决定将哪些数据类型结合在一起进行展示。

6、 引用[1] Waskom M L. Seaborn: statistical data visualization[J]. Journal of Open Source Software, 2021, 6(60): 3021.  
[2] Brodlie K, Allendes Osorio R, Lopes A. A review of uncertainty in data visualization[J]. Expanding the frontiers of visual analytics and visualization, 2012: 81-109..  
[3] Azzam T, Evergreen S, Germuth A A, et al. Data visualization and evaluation[J]. New Directions for Evaluation, 2013, 2013(139): 7-32.

[4] Ahn Y Y Y Y. Data visualization[J]. 2019.  
[5] Buja A, Cook D, Swayne D F. Interactive high-dimensional data visualization[J]. Journal of computational and graphical statistics, 1996, 5(1): 78-99..