復旦大學

数据挖掘项目报告

课程名称： 数据挖掘

学　　院： 航空航天系

学　　号： 16300290001

姓　　名： 杨宇锋

针对世界各个国家幸福指数的数据挖掘项目

摘 要

1. 本项目选择了样例数据集中的经济类数据“world happiness”, 针对诸多变量分析幸福指数的制约因素和国家综合实力的聚类关系.

2. 在数据预处理阶段, 我们主要提取了数据集中的六个关键变量“经济、健康、自由、家庭、政府信任度、慷慨度”和两个因变量“幸福指数和幸福指数的标准差”, 希冀从中分析出有用信息. 提取完之后, 我们先对六个变量做一次 PCA 观察下六个变量数据的冗余程度和线性相关程度.

3. 在数据分析阶段, 先考察了各主要变量与幸福指数的相关性, 并针对经济变量做了一元回归的显著性检验, 由此知道幸福指数与经济等变量很强的线性相关性.

4. 由于 Trust 与幸福指数之间的弱相关性, 之后考察 Trust 和 Economy 变量之间的关系. 利用 DBSACAN 聚类算法, 发现 Trust 与幸福指数其实是有正向关系的, 只不过是二分的跃变关系, 即超过某一阈值之后幸福指数基本都处于全球的第一梯队.

5. 之后我们继续考察了幸福标准差与各变量之间的关系, 起初猜测幸福标准差能很大程度上反映一个国家的生活质量差距和贫富差距, 但从画出的散点图来看关系并没有那么显著, 反而不同国家的标准差都趋于某一稳定值. 但值得注意的是, 当一个国家处于某一特定阶段, 如家庭和政府信任度在 1.1 和 0.1 左右时, 幸福指数的标准差突然偏大, 这其实可以从每个国家的不同发展阶段来进行合理解释, 的确一个国家在某一发展阶段时国内公民的贫富差距会处于历史峰值.

6. 最后我们利用 KMEANS 聚类算法对不同国家的经济、健康、自由变量做了聚类分析, 发现这几个关键变量显著地分隔了不同类的国家, 且这些变量的高低几乎直接决定了幸福指数的高低与否, 很少有低综合实力国家幸福指数处于第一梯队.

7. 对于以上的分析, 我们在总结部分做了归纳和细致的讨论.

目录

1 项目选题介绍 **1**

1.1 **研究问题** 1

1.2 **主要方法** 1

1.3 **数据集介绍** 1

2 数据预处理 **2**

2.1 主成分分析 2

2.2 对二维数据的可视化 2

3 数据分析 **3**

3.1 主要变量与幸福指数的相关性 3

3.2 Trust 变量与 Ecnonomy 变量之间的关系 5

3.3 幸福指数标准差与各变量之间的关系 6

3.4 关于国家或地区的聚类 7

4 总结与讨论 **8**

1 项目选题介绍

21 世纪的今天, 越来越多的国家开始关注起公民幸福指数。世界民意调查 Gallup World Poll 自 2012 年起, 每年发布世界各个国家及地区的幸福指数排名并附有该国家各项国家实力指标, 如 GDP、健康指数等等. 这份排名给了许多国家一些启示, 因为从中也可以看出最让人关注的 GDP 指标并不完全与幸福指数相关联. 针对这份排名数据中诸多有趣的数据关联及因果关系，我们对其中的数据进行分析, 揭示国际上公民的幸福到底与哪些因素休戚相关.

1.1 研究问题

我们的本质问题是探究世界公民的幸福感到底与哪些变量最相关, 并且明确它们对于幸福感的贡献. 以此衍生的问题有:

1. 各个变量之间的相关性
2. 不同的变量怎样影响幸福指数的变化?
3. 哪些原因造成了各个国家之间的幸福感指数的方差不同?
4. 发达国家与发展中国家的幸福指数是否有明显的差异，甚至形成两类?

1.2 主要方法

在研究相关性问题的时候我们主要利用主成分分析及一些统计学中的假设检验方法. 之后也利用了回归分析的方法, 探讨了各变量或多变量与幸福指数的关系. 所有的代码都基于 python 编写，主要用到了 numpy、sklearn、matplotlib 等常用的数据分析扩展库.

1.3 数据集介绍

此数据集来源于所给的样例数据集中的经济类数据 “world happiness” . 该数据集一共包含 3 年的信息, 每一年都有 Country,Region,Happiness Rank,Happiness Score,Standard Error,Economy (GDP per Capita),Family,Health (Life Expectancy),Freedom,Trust (Government Corruption),Generosity 等变量.

所有的挖掘任务仅在该数据集上完成.

2 数据预处理

2.1 主成分分析

为了观察数据的冗余性，我们先对整个数据集做了一次主成分分析，采用 python 的 sklearn 扩展库. 在对六个维度的数据做了 PCA 之后我们发现，各维度的方差比例约是

$$[0.71380549 \quad 0.11285944 \quad 0.07646787 \quad 0.04647401 \quad 0.02790172 \quad 0.02249148] \tag{1}$$

由此也可以看出最大的三个维度方差和占比就已经是 90% 左右了, 所以大约只要三个维度的数据就可以很好的描述整个数据集，也由此证明不同维度之间的数据存在着很大程度上的线性相关性.

2.2 对二维数据的可视化

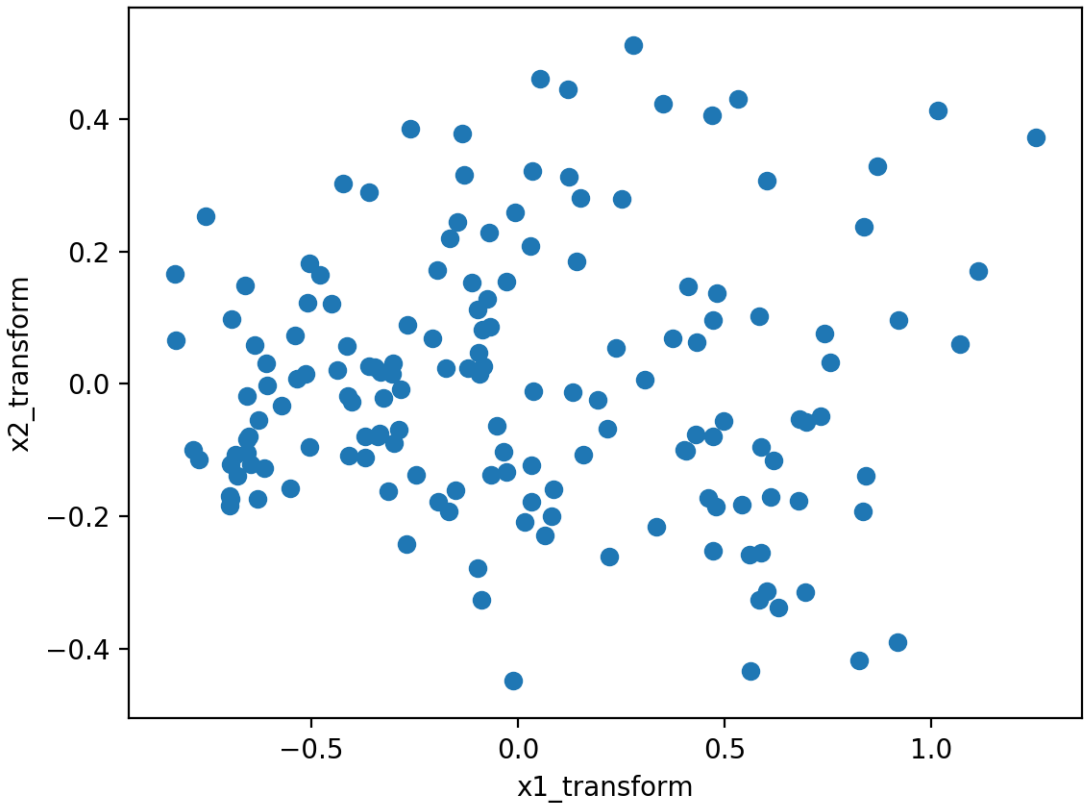

图 1: 从图中我们可以看到方差占比最大的两维数据点分布相对较为均匀, 没有呈现明显的聚类，但还是可以看出左下和右上的数据点密度有着明显差异

3 数据分析

3.1 主要变量与幸福指数的相关性

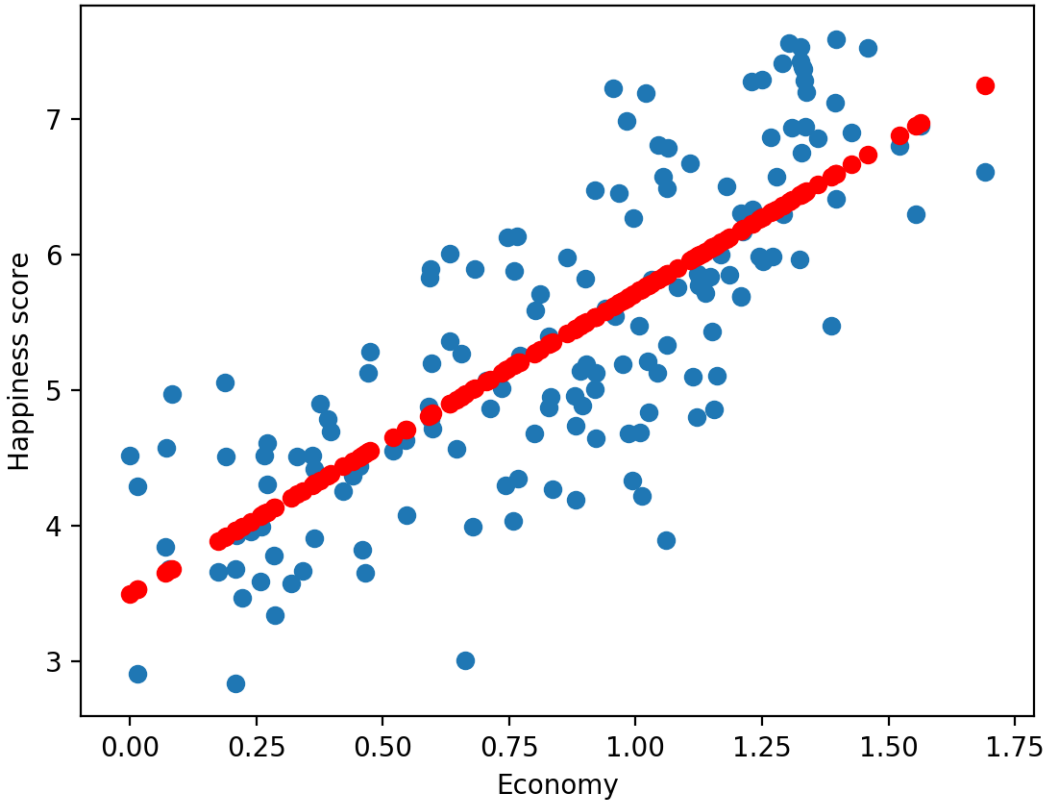

图 2: 经济指数对幸福的影响; 红色直线为线性回归拟合的直线 $y = 3.4988 + 2.2182x$

在我们普罗大众的认知中，金钱对生活幸福的影响较大，所以我们对经济这个变量继续做回归的假设检验以验证回归效果的显著性.

检验假设:

$$\hat{y}_0 = \hat{a} + \hat{b}x \tag{2}$$

$$H_0 : b = 0 \tag{3}$$

$$H_1 : b \neq 0 \tag{4}$$

利用 t 检验法来进行检验:

$$\hat{b} \sim N(b, \sigma^2/S_{xx}) \tag{5}$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 \tag{6}$$

当 H_0 为真时, $b = 0$, 此时,

$$t = \frac{\hat{b}}{\hat{\sigma}}\sqrt{S_{xx}} \sim t(n-2) \tag{7}$$

$$\hat{\sigma^2} = \frac{S_{yy} - \hat{b}S_{xy}}{n-2} \tag{8}$$

其中,

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{9}$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2 \tag{10}$$

将数据带入，可知

$$S_{xx} = 25.5134 \qquad S_{xy} = 56.5947$$

$$S_{yy} = 205.8345 \qquad \sigma^2 = 0.5147$$

$$t = 15.6171$$

取置信水平 $1 - \alpha = 0.95$, 则

$$t_{\frac{\alpha}{2}}(n-2) \approx 1.96 \tag{11}$$

由此可以看出

$$15.6171 = t \quad > \quad t_{\frac{\alpha}{2}}(n-2) \approx 1.96 \tag{12}$$

故可知，t 落在了拒绝域，回归效果是非常显著的。同时也说明经济因素与幸福指数是完全正相关的.

针对其他维度，我们也一并画出二维的散点图，并做一元回归分析.

从图中我们也可以大致看出, 经济、家庭、健康与幸福指数呈现显著的线性相关性. 而自由等变量与幸福指数的散点图呈现不同的分布状态, 有待我们进一步考察其中的具体关系.

如果说 Freedom 变量与幸福指数的散点图大体还能呈现一定的分散性和线性性, 但 Trust 和 Generosity 与幸福指数的散点图则有聚拢在左半图的趋势, 那么这是不是可以侧面反映 Trust 等变量与幸福是独立的, 或者它们只与国家发展情况有关呢?

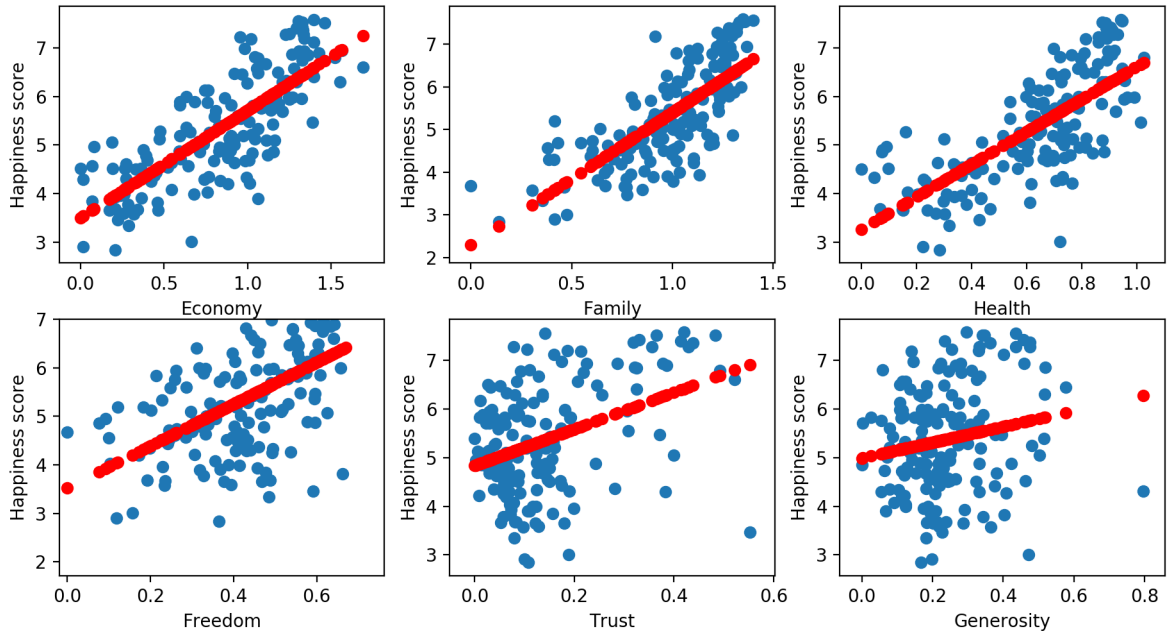

图 3: 六个变量分别与幸福指数的二维散点图

3.2 Trust 变量与 Ecnonomy 变量之间的关系

我们先来画出 Trust 和 Economy 之间的二维散点图.

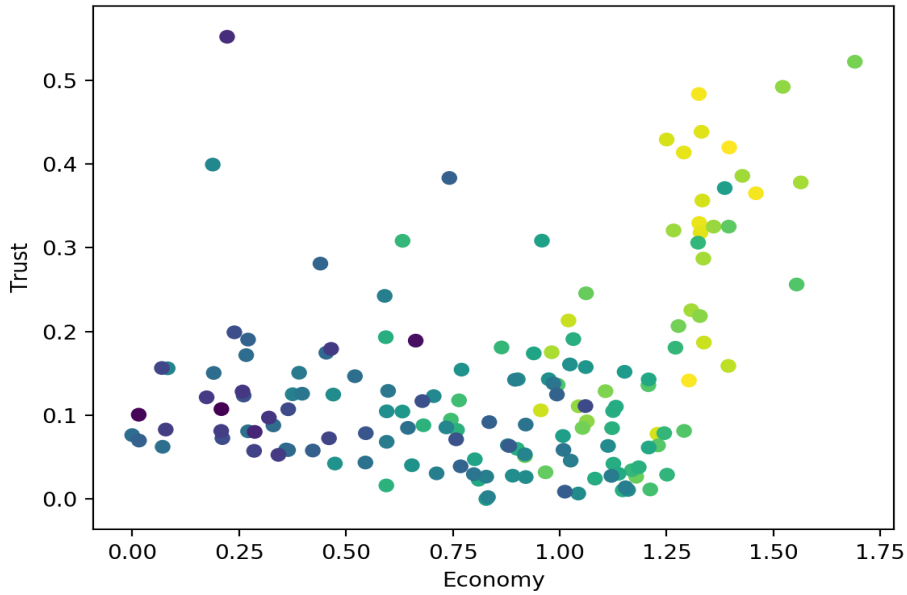

图 4: Economy 和 Trust 的二维散点图; 颜色越浅幸福指数越高

从这张图中我们可以看出 Trust 变量与幸福指数并非没有关系, 而是存在一个跃变, 非常明显地将不同国家分隔开来. 我们利用 DBSCAN 聚类算法尝试将这组数据点进行聚类. 在 DBSCAN 算法中, 我们设置 $\epsilon = 0.06$, 核心对象的 ϵ 邻域内至少有两个点.

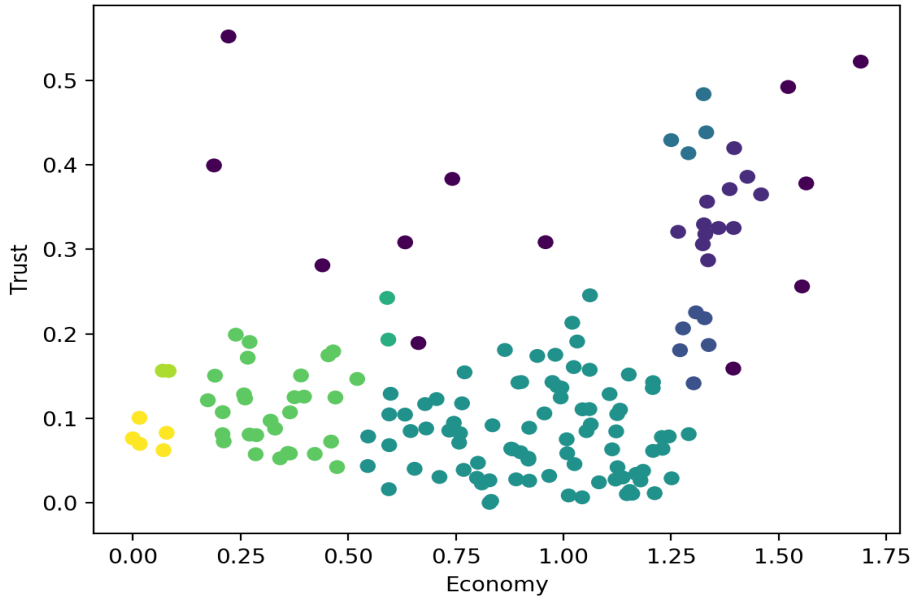

图 5: DBSCAN 聚类结果; 相同颜色代表同一类别

在 DBSCAN 聚类后的结果中可以看到, Trust 与幸福指数是呈正向关系的, 且当一个国家的 Trust 在 0.3 以上时, 幸福指数明显在全球范围内处于第一梯队. 但毫无疑问, 图中左上角存在个位数的离群点, 这些国家虽然 Trust 超过 0.3, 但是由于经济处于下游的原因, 其幸福指数并不高, 这也并不令人感到奇怪.

综上所述, 我们可以归纳出, 某些变量如经济、健康与幸福指数有明显的线性关系, 而另外一些变量如政府信任度或是慷慨度与幸福指数有着跃变的二分关系, 即存在着一个阈值分割了高幸福指数和相对低幸福指数.

3.3 幸福指数标准差与各变量之间的关系

通过正常的推测, 标准差往往能反映一个国家或地区的贫富程度或是生活质量. 为了验证我们的推测, 我们分别来观察下标准差和各主要变量的关系.

从六幅图来看，幸福指数的标准差基本都集中在 0.05 左右, 各变量对幸福指数标准差没有明显的影响. 另外我们还可以发现当家庭指数处在 1.1, 政府信任度在 0.1 左右时, 幸福指数标准差的方差较大, 且绝大多数为低幸福国家或地区. 这种现象可以说明处于某一特定发展阶段的欠发达或发展中国家贫富差距会处于峰值. 从实际现实出发这点同样可以理解, 比如中国在改革开放到现在经济大力发展, 原本的均贫转变成了先富带动后富，贫富差距慢慢开始到达历史最高点. 再以发达国家美国为例, 他们已经经过了高速发展阶段, 整体经济水平处于稳定阶段, 尽管贫富差距依然拉得非常大, 但是大多数人已经达到小康之上，总体的标准差应该是在不断减小的.

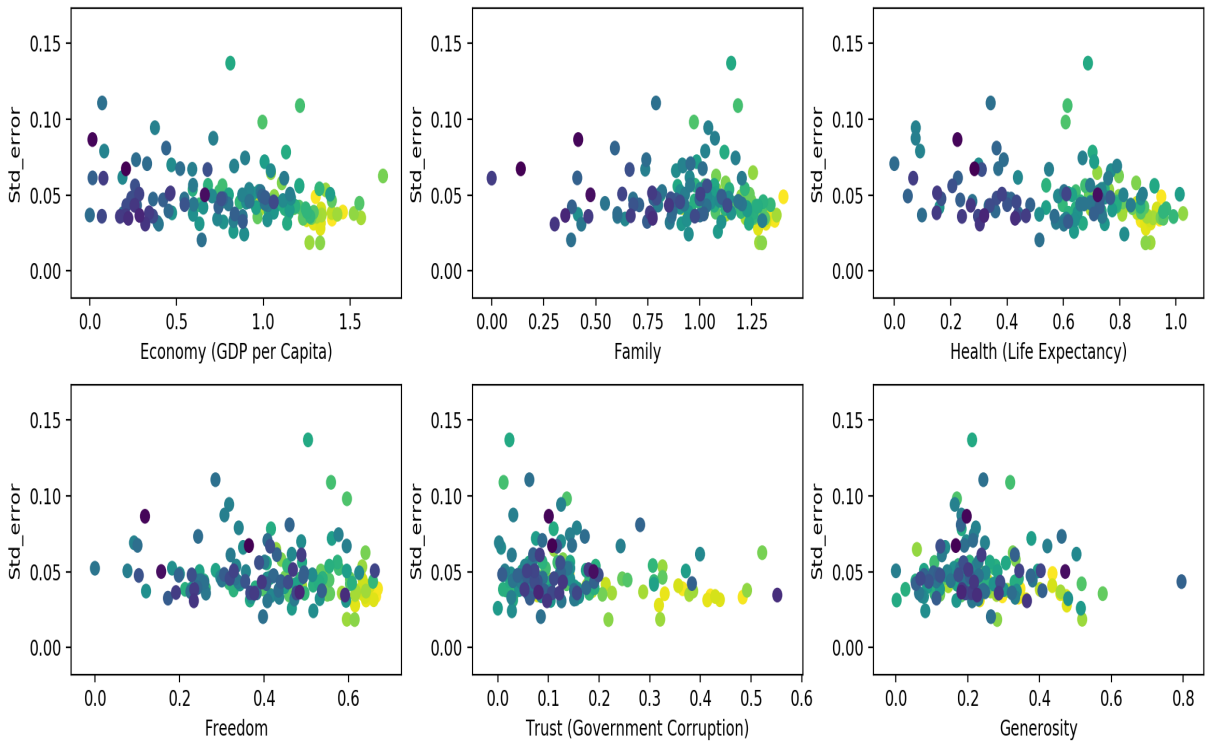

图 6: 幸福标准差和各变量的二维散点图; 颜色越浅幸福指数越高

3.4 关于国家或地区的聚类

在最后的环节, 我们通过 Kmeans 算法讨论一下不同幸福指数的国家在经济、自由、健康的变量下是否有明显的聚类关系.

首先我们将 kmeans 算法中的缺省值 “类别” 设置成 3 类, 初始中心点设为随机. 然后利用 mpltoolkits.mplot3d 扩展库画出三维的散点图. 从图 7 中可以看到, kmeans 的聚类结果和幸福指数的分布非常的相似. 明显的是, 低幸福指数的点与中高幸福指数的点有显著的稀疏间隔. 另外, 我们所选的这三个变量是六个变量中方差最大的三个, 所以更能反映数据集的离散分布.

在图 7 中, 三个变量越高幸福指数越大, 这与我们的认知也大致相同. 并且在分布中我们也可以看到几乎没有高幸福指数国家落在左下角的聚类中, 也就是没有离群异常点. 换句话说, 高幸福的前提是一个国家必须要有大体量的经济、健全的医疗保障制度和公民的自由权利, 缺少任何一个环节都无法有效提高国家公民的幸福指数.

从以上的分析中可以总结出, 不同的国家之间随着国家硬实力的变量的确有着明显的聚类, 而且这种聚类的划分也与幸福指数的高低划分有着极大的趋同性. 国家总体实力的层次的确预示着幸福指数的高低.

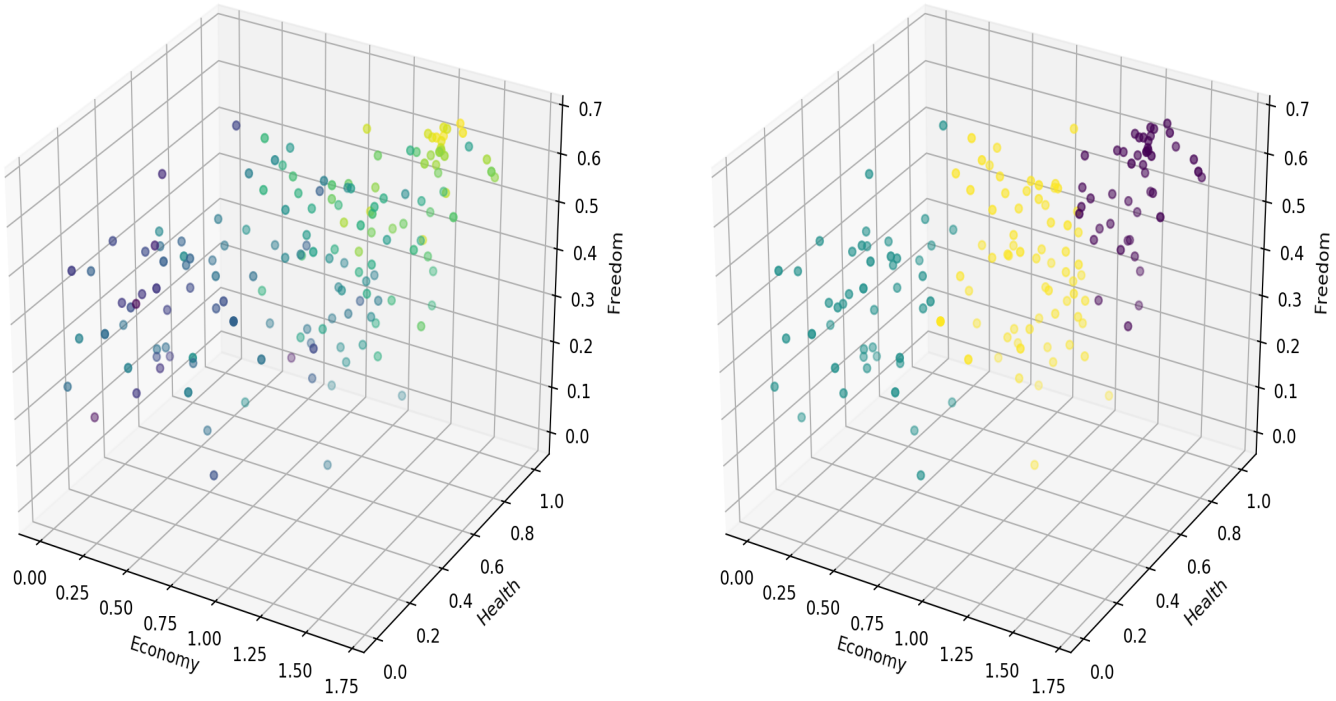

图 7: 左图的颜色越浅代表幸福指数越高; 右图为 kmeans 聚类结果, 不同颜色代表不同类别

4 总结与讨论

在这篇报告中, 我们主要分析了变量与变量之间的相关性、变量与幸福指数的相关性及国家地区之间的聚类关系. 得出了如下结论:

1. 一个国家的经济、健康、信任等变量存在着较强的线性相关性, 意味着大多数国家的各方面实力都是几乎同步提高的.

2. 一个国家的幸福指数主要与经济、健康、自由有较强的线性相关性, 且很少有低经济和自由的国家幸福指数处于中上位置.

3. 某些变量如 Trust 对幸福指数的影响是二分的, 也就是一旦到达某阈值之后, 幸福指数马上处于第一梯队.

4. 幸福指数的标准差不能完全代表贫富差距和生活质量差距, 可能还有一些噪音干扰, 但幸福指数的标准差的确在某个特定的阶段或状态下呈现偏大的趋势, 这可以从某个国家的不同发展阶段来合理解释.

5. 国家或地区的综合实力的确是有聚类划分的明显差异, 高综合实力国家、发达国家的公民幸福指数的确处于第一梯队, 而且少有例外.

参考文献

[1] 盛骤, *概率论与数理统计: 第三版*. 高等教育出版社, 2001.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[3] "world happiness," *https://gitee.com/LightSwitch*, 2018.

[4] Y. Yang, "world happiness datamining minorpj," *https://github.com/yyf710670079*, 2018.