# 基于网站问卷调查数据的生活方式与生活工作平衡分析

沈宇捷 致理-物22 2022012290

## 摘要

现代快节奏生活下，生活方式与生活工作平衡成为重要话题。本文基于多元统计分析的方法，对于网站上关于生活方式与状态的数据进行分析。利用主成分分析和因子分析进行降维和解读，利用 K-means 聚类的方法对生活状态进行聚类得到类别标签，再利用 Fisher 判别法利用生活方式变量进行分类。研究发现人与人之间的亲密交流与帮助是生活方式的主要因子，压力的大小是生活状态的主要影响因子。

**关键词：生活方式 生活工作平衡 因子分析 K-means 聚类 Fisher 判别法**

## 目录

# 1 研究问题与背景描述

现代快节奏生活之下，当加班熬夜成为年轻人生活的常态，身心健康成为人们关心的重要话题。健康的生活方式对于身心健康非常重要。关于生活方式（lifestyle）与幸福（well-being）之间的关系，已经有不少研究提供了不同的视角和发现。一项全国性的研究通过对28,138名中国成年人的调查数据进行分析，发现生活方式因素与心理健康（如抑郁、焦虑、孤独、感知压力和自评健康状况）及幸福之间存在显著关联（Xue Wang et al.,2023）。在加拿大进行的一项观察性研究调查了 147 名护理专业本科生的生活方式因素与心理健康之间的关系。研究发现，久坐时间、睡眠质量差和乳制品摄入量低与抑郁、焦虑和心理压力得分较高有关,这表明不良生活方式行为可能会降低心理健康(Charlotte T Lee et al.,2022)。

本文聚焦于对生活方式本身的分析和生活方式对幸福度或生活工作平衡程度的影响。采用问卷调查数据，将变量进行分类，包括生活方式变量和生活状态变量。对生活方式变量进行PCA降维，用因子分析解释潜在影响因子。对生活状态变量进行聚类分析，分为不同的类别。最后采用判别分析，用生活方式变量对生活状态进行判别、分类。

# 2 数据介绍与描述

## 2.1 数据集描述

本研究所用数据集为来自Kaggle平台的“Lifestyle and Wellbeing Data”，包含了来自 Authentic-Happiness.com 网站的一项 2015 至 2021 年的关于生活方式的问卷调查的15,977份调查结果，共有24项指标，包括生活方式、生活状况或幸福度衡量等，具体名称及含义见下表1所示。其中前23项为问卷结果，为间隔为1，范围为0~5或0~10的离散变量，“work life balance score”由网站系统通过算法得到。

表1 各变量名称及含义

| 变量名称 | 含义 | 变量名称 | 含义 |
|---|---|---|---|
| Timestamp | 调查时间 | DAILY_STRESS | 每天感受到的压力 |
| FRUITS_VEGGIES | 每天吃蔬菜量 | CORE_CIRCLE | 亲近人数 |
| PLACES_VISITED | 每周拜访新地方量 | WEEKLY_MEDITATION | 每周冥想次数 |
| SOCIAL_NETWORK | 每天和多少人交流 | ACHIEVEMENT | 有多少成就 |
| DONATION | 参与过几次捐款 | BMI_RANGE | BMI指数范围 |
| TODO_COMPLETED | 待办事项完成情况 | FLOW | 每天“沉浸”时间 |
| DAILY_STEPS | 每日步数 | LIVE_VISION | 对几年后有规划 |
| SLEEP_HOURS | 每天睡眠时间 | LOST_VACATION | 每年未休假期数 |
| DAILY_SHOUTING | 吼叫频繁程度 | SUFFICIENT_INCOME | 收入支出是否平衡 |

| | | | |
|---|---|---|---|
| PERSONAL_AWARDS | 一生成就数 | AGE | 年龄范围 |
| SUPPORTING_OTHERS | 帮助几人过上好生活 | TIME_FOR_PASSION | 每天做热爱事情小时数 |
| GENDER | 性别 | WORK_LIFE_BALANCE_SCORE | 生活工作平衡指数 |

## 2.2 探索性数据分析

首先我们观察全体数据本身的尺度等基本信息，画出所有变量的箱线图观察变量的分布，如图 1 所示。

由图 1 可知，数据中异常值较少，BMI_RANGE 和 SUFFICIENT_INCOME 为二值变量。变量尺度不同，且值含义为打分或次数，没有明确单位，因此后续可以考虑标准化处理。

由 2.1 数据集，AGE 分为 4 个范围，分别为小于 20，21 至 35，36 至 50，51 以上。画出年龄、性别同工作生活平衡指数的关系如图 2 所示。可知，21 至 35 岁的青年女性工作生活最不平衡，51 岁以上的老年人生活相对最平衡。

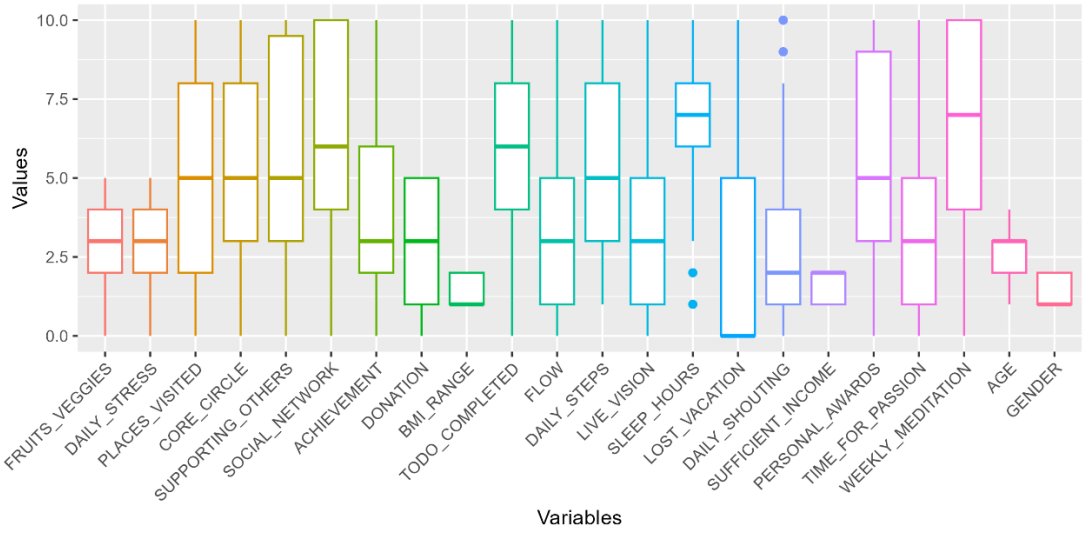


图 1 所有变量箱线图

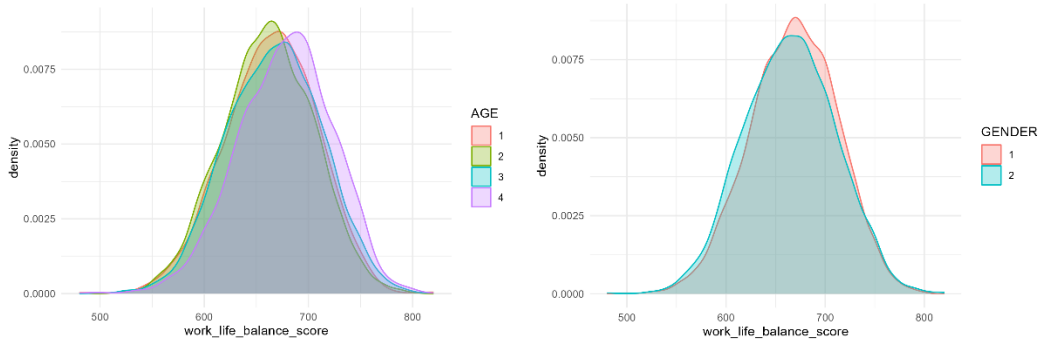


图 2 性别、年龄与工作生活平衡指数的关系

接下来我们考虑不同变量之间的关系，画出相关系数热力图。发现工作生活平衡指数与AGE，GENDER 之外的大部分变量相关系数较大，其中 DAILY_STRESS，BMI_RANGE，LOST_VACATION 和 DAILY_SHOUTING 呈负相关性，其他为正相关，符合逻辑。各变量之间有相关性，后续可以考虑降维处理。

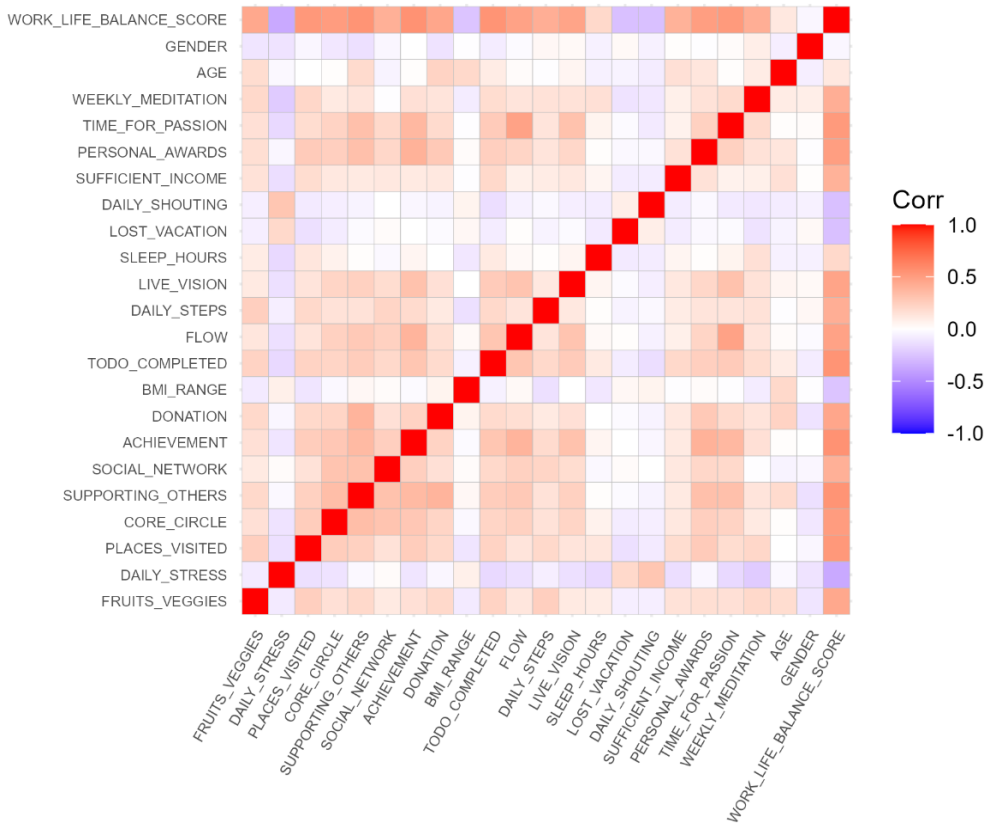


图 3 变量间相关系数热力图

# 3 数据预处理与降维

## 3.1 主成分分析与降维

删去 Timestamp 列，删去含有缺失数据的行。若直接对年龄、性别、得分外的 20 项指标进行 PCA 降维，效果较差，前几主成分占方差比小，同时变量可分为“生活方式”变量和“生活状态”变量，含义不同。因此首先进行手动特征选择，挑选与 WORK_LIFE_BALANCE_SCORE 相关性较高且数据本身分布较好的“生活方式”变量，共 9 个（具体可见表 2）。对数据标准化处理。

再进行 PCA 降维，画出崖底碎石图如图 4。在第二点处出现拐点，但第二点处累积解释方差比例仅为 38%。因此最终决定选择 5 个主成分，累积方差解释比例达 68%。降维效果其

实仍不理想，但 PCA 为后续可视化提供基础。在可解释性上，前五主成分系数分布都较为均匀，可解释性弱，因此考虑使用因子分析。

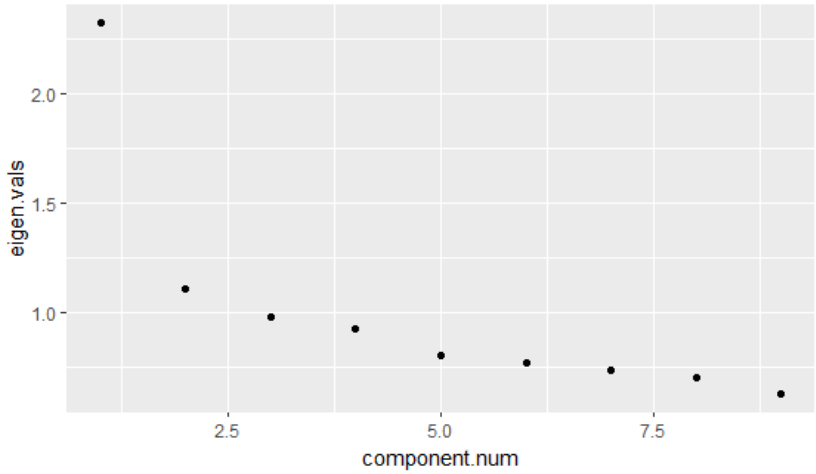


图 4 崖底碎石图

### 3.2 因子分析

经尝试，决定采用 MLE 结合 varimax 旋转的方法，尽可能提高解释性。由 PCA 结果，选取 5 个因子，得到的各因子系数如下表 2 所示。

表 2 各因子系数

| 变量 | 因子 1 | 因子 2 | 因子 3 | 因子 4 | 因子 5 |
|---|---|---|---|---|---|
| FRUITS_VEGGIES | | | | 0.488 | 0.156 |
| PLACES_VISITED | 0.151 | 0.127 | | 0.393 | 0.233 |
| CORE_CIRCLE | 0.965 | 0.151 | | 0.191 | |
| SUPPORTING_OTHERS | 0.154 | 0.959 | | 0.228 | |
| TODO_COMPLETED | 0.120 | 0.157 | | 0.373 | 0.180 |
| DAILY_STEPS | | | | 0.523 | |
| SLEEP_HOURS | | | | | 0.459 |
| DAILY_SHOUTING | | | 0.991 | | |
| WEEKLY_MEDITATION | | | | 0.300 | 0.311 |

由此可知，因子 1 主要由 CORE_CIRCLE 决定，可解释为社交因子；因子 2 主要由 SUPPORTING_OTHERS 决定，可解释为帮助他人因子；因子 3 主要由 DAILY_SHOUTING 决定，可解释为情绪稳定因子；因子 4 主要由 FRUITS_VEGGIES 和 DAILY_STEPS 决定，可解释为生活健康因子；因子 5 主要由 SLEEP_HOURS 决定，可解释为睡眠质量因子。因子分析可解释性较好。我们可以看到，人与人之间的交流和帮助是生活方式的重要影响因素。

## 4 类别信息分析

## 4.1 聚类分析

本研究目标为由生活方式对人的工作生活平衡度或幸福度进行分类。困难在于，数据集所给 WORK_LIFE_BALANCE_SCORE 为连续变量，且不知计算方法，难以根据得分高低确定分类数和分类指标（难以确定分类界线）。因此本研究打算根据数据集中的“生活状态”变量进行聚类分析，为后续判别分析获取分类指标。

首先，挑选“生活状态”变量为 DAILY_STRESS,ACHIEVEMENT,TIME_FOR_PASSION，对数据标准化处理，进行 PCA 降维获取前两个主成分为可视化做准备。由此第一主成分主要代表压力程度，第二主成分主要代表成就程度。

表 3 主成分系数

| 变量 | PC1 | PC2 |
|---|---|---|
| DAILY_STRESS | 0.758 | 0.545 |
| ACHIEVEMENT | -0.601 | 0.823 |
| TIME_FOR_PASSION | 0.254 | -0.157 |

再确定聚类的个数，计算不同聚类个数聚类后的组内方差，画出崖底碎石图（图 5），推荐的聚类个数为 4。于是我们利用 K-means 方法进行聚类。但由图 6 可知，聚类结果并不好，第 3 类和其他 3 类严重重合。

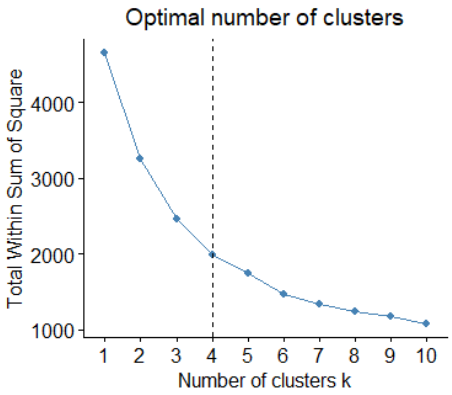


图 5 推荐的聚类数量

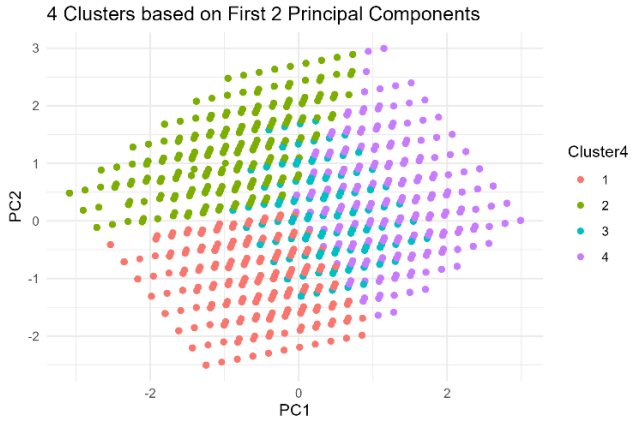


图 6 聚类数为 4 时聚类结果

因此，把聚类数改为 3 和 2，重新进行 K-means 聚类，结果如图 7 和图 8 所示。聚成 3 类时第 2、3 类有部分区域重叠，但整体的聚类的分类效果较好。且具有可解释性：分为 3 类时第一类可解释为压力较小且成就尚可，第二类为压力较小且成就少，第三类为压力较大；分为 2 类时第一类可解释为压力较小，第二类压力较大。结合现实，我们可以形象但粗略地把三类解释为“介于躺平内卷之间”，“躺平”和“内卷”三类人。同时，可以看到压力为区分生活状态的重要因素。

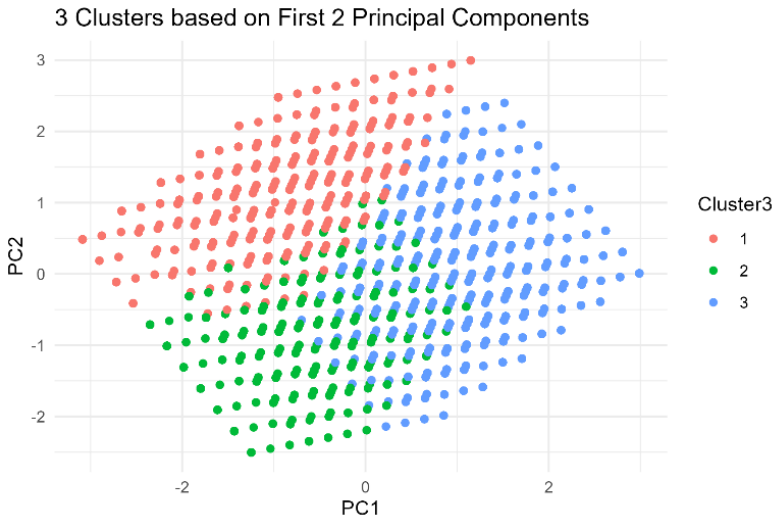


图 7 聚类数为 3 时聚类结果

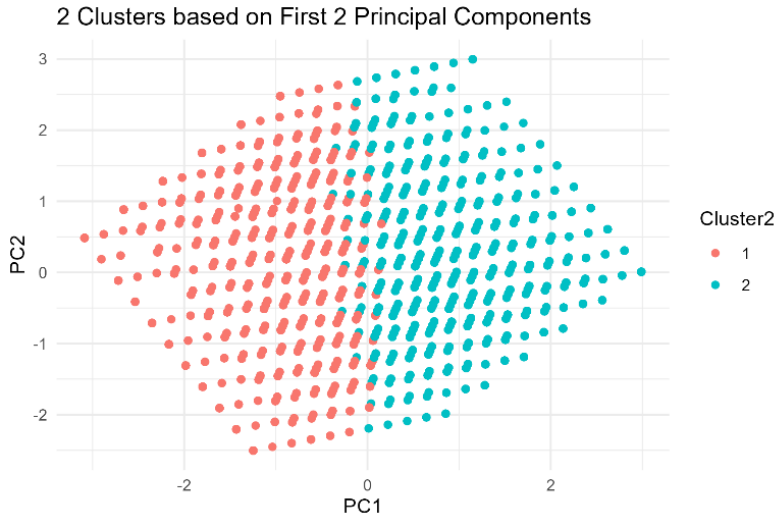


图 8 聚类数为 2 时聚类结果

试图将此聚类结果与 WORK_LIFE_BALANCE_SCORE 列结果进行对比。首先对得分进行非常粗略的分类：得分大致呈正态分布，中位数和均值接近，按均值分为两大类。进行可视化，如图 9 所示。发现两类有较多重叠，且与图 8 差距较大，说明将得分简单二分的分类方法不好，或是与前述自定的分类方法不一致。因此，在后面判别分析时暂时不考虑得分列。

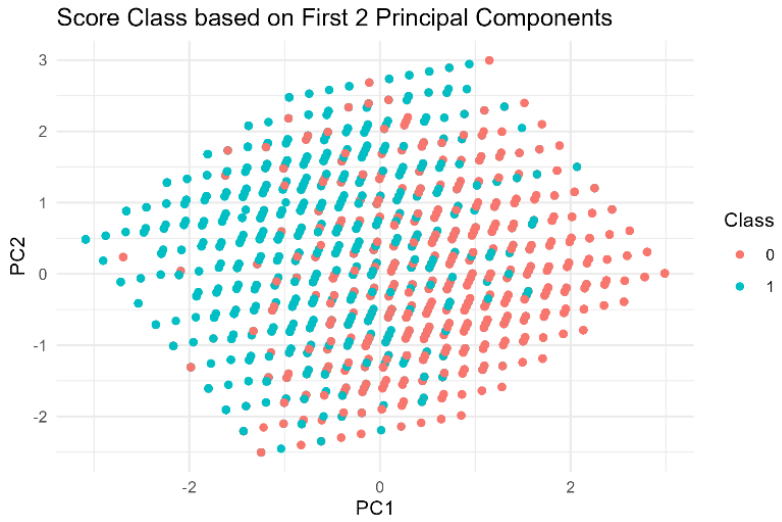


图 9 得分分为两类可视化

## 4.2 判别分析

判别分析的目的是试图通过“生活方式”的 9 个变量把得到每个样本的生活状态分类。把 4.1 中聚类数为 2 和 3 时的分类结果作为两种类别标签，进行 Fisher 线性判别分析即 LDA。得到的混淆矩阵如下表 4、5 所示。分成 3 类和 2 类的 APER 值分别为 23%和 11%。将分类结果进行可视化如图 10、11 所示。

表 4 分为 3 类时的混淆矩阵

| | 真实 1 类 | 真实 2 类 | 真实 3 类 |
|---|---|---|---|
| 预测 1 类 | 6485 | 388 | 547 |
| 预测 2 类 | 257 | 1619 | 954 |
| 预测 3 类 | 44 | 1501 | 4176 |

表 5 分为 2 类时的混淆矩阵

| | 真实 1 类 | 真实 2 类 |
|---|---|---|
| 预测 1 类 | 8018 | 1693 |
| 预测 2 类 | 82 | 6178 |

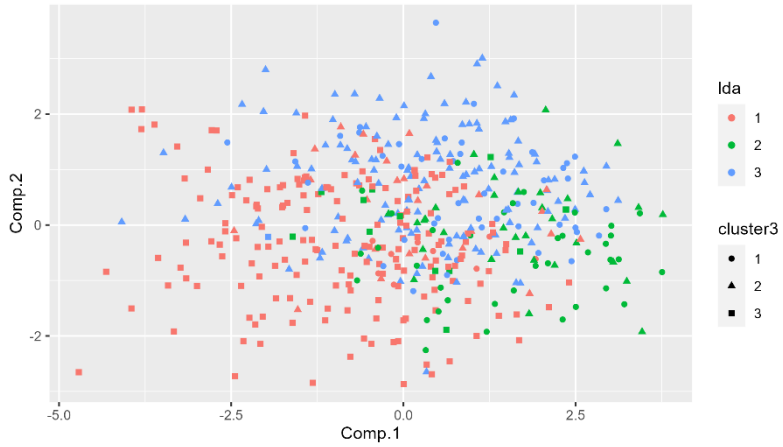


图 10 分为 3 类时的 LDA 结果

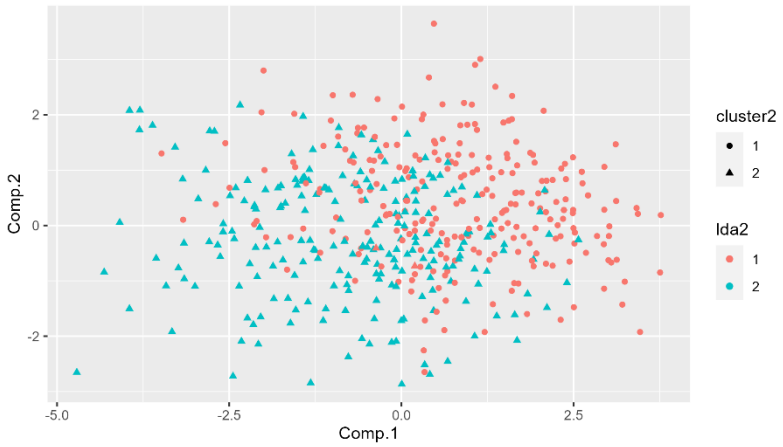


图 11 分为 2 类时的 LDA 结果

对于分成 3 类的情况，APER 值较高。因此考虑采用其他判别分析方法。使用 QDA 方法，得到 APER 为 22%，仍较高。可视化如图 12 所示。可能原因有：（1）变量非正态分布，QDA 方法不适合（2）分类标签不够理想。由图 7，2、3 类标签有部分区域重合；由表 3，较大误差出现在 LDA 分类器把很多 2 类分成了 3 类（也是基于此，本文尝试了分成 2 类的 LDA）。

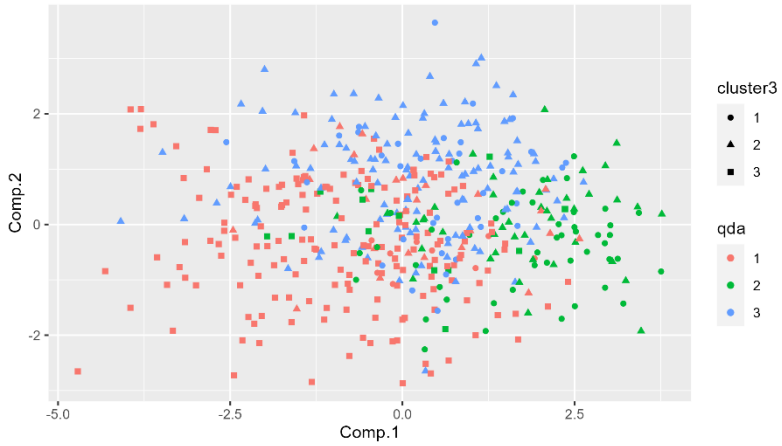


图 12 分为 3 类时的 QDA 结果

对于分成 2 类的情况，为观察过拟合的影响，将数据集按 7：3 比例分为训练集和测试集，只用训练集训练。发现在训练集和测试集上的 APER 分别为 11.04%和 11.13%，说明基本没有过拟合现象。

总体来说，可以在仅仅知道生活方式数据的情况下，得到生活状态的分类信息，二者有较强关联。

# 5 总结与改进

## 5.1 总结

本研究目标为分析生活方式和生活状态数据，并分析二者的关联。

数据集中的变量可分为生活方式变量，生活状态变量，年龄性别和综合得分变量。因此本文先用探索性数据分析的方法对于各变量分布以及相互关系进行初步分析。再通过生活状态变量对每个样本的幸福指数进行分类，并用生活方式变量进行判别分析。其中重要步骤如下：

对于主成分分析，由于生活方式变量个数较多，本文首先进行手动重要特征提取，并采用 PCA 方法进行降维处理，获得主成分得分以便后续可视化。

对于因子分析，由于主成分分析得到的主成分难以解释，所以为提高可解释性，用因子分析方法探究影响生活方式的潜在因子，获得 5 个因子，发现亲密朋友数量和帮助他人程度为生活方式变量的重要解释因子。

对于聚类分析，由于得分变量为连续变量，难以分类，因此通过对生活状态变量进行聚类得到分类标签。分析发现，聚成 2 或 3 类都是较为合理的方案，并对各个类别的特点进行了可解释性分析。

对于判别分析，采用 LDA 和 QDA 方法，发现两种方案对 3 分类方案效果相近，且都不是非常理想。将数据分为训练集和测试集，发现 LDA 的 2 分类方案几乎不存在过拟合问题。

## 5.2 改进

本研究存在一些需要改进或进一步研究的地方。

（1） 本研究目标为分类，并自行定义分类指标，为使用得分变量。后续可采用回归分析等方法，从样本的生活方式变量得到具体的幸福度得分。也可探究每个变量对最后得分的影响程度大小，从而为每个样本提出改进哪些方面的生活建议。

（2） 判别分析时只采用 LDA 和 QDA 方法，后续可采用 Logistic 回归或随机森林、支持向量机等判别方法。聚类分析可考虑采用 k-medoids 等方法。

（3） 在理论分析上，由于作者在心理学方面术语知识的缺失，本研究结合常识和已有研究将变量分为生活方式和生活状态变量，具有较大主观性，同时没有具体区分幸福度指数和工作生活平衡得分。

（4） 数据集数据为问卷调查评分，为间隔为 1 的离散变量，数据重复度较高，所以可视化时主成分也呈离散或规律分布。数据的离散性可能会影响统计方法的可靠性与准确度。

# 6 参考文献

[1] Wang X, Wu Y, Shi X, Chen Y, Xu Y, Xu H, Ma Y, Zang S. Associations of lifestyle with mental health and well-being in Chinese adults: a nationwide study. Front Nutr. 2023 Jun 23;10:1198796. doi: 10.3389/fnut.2023.1198796. PMID: 37426182; PMCID: PMC10327438.

[2] Lee, C. T., Ting, G. K., Bellissimo, N., & Khalesi, S. (2022). The associations between lifestyle factors and mental well-being in baccalaureate nursing students: An observational study. Nursing & Health Sciences, 2.7, DOI: 10.1111/nhs.12923.

# 7 附录

数据集来自 Kaggle 网站，网址：https://www.kaggle.com/datasets/ydalat/lifestyle-and-wellbeing-data。

R 代码如下：

```
### 数据预处理
data = read.csv("Wellbeing_and_lifestyle_data_Kaggle.csv",header = T)
data = data[,-1]
data$DAILY_STRESS = as.integer(data$DAILY_STRESS)
data = data[-10006,]#含有 NA
age_dict <- list('Less than 20' = 1, '21 to 35' = 2, '36 to 50' = 3, '51
or more' = 4)
data$AGE <- as.integer(factor(data$AGE, levels = names(age_dict), labels
= age_dict))
gender_dict <- list('Female' = 1, 'Male' = 0)
data$GENDER <- as.integer(factor(data$GENDER, levels = names(gender_dict),
labels = gender_dict))

data$class2 = ifelse(data$WORK_LIFE_BALANCE_SCORE<667,0,1)
data$class2 = as.factor(data$class2)
data$score = as.factor(data$WORK_LIFE_BALANCE_SCORE)

##########################################
### EDA
library(ggplot2)
library(reshape2)
melted_data = melt(data[,-23])
ggplot(melted_data,aes(x =variable, y = value, color = variable)) +
  geom_boxplot() +
  labs(x = "Variables", y = "Values") +
  theme(legend.position = "none") +
```

```
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
ggsave("boxplot.png", plot = last_plot(), width = 8, height = 4, units =
"in", dpi = 300)

library(ggcorrplot)
ggcorrplot(cor(data)) +
  theme_minimal() +
  theme(axis.text.x = element_text(size = 6, angle = 60, hjust = 1),
        axis.text.y = element_text(size = 6)) +
  labs(x = "",y = "")
ggsave("corrplot.png", plot = last_plot(), width = 6, height = 6, units
= "in", dpi = 300, bg = "white")

#不同年龄 score
data$AGE = as.factor(data$AGE)
ggplot(data, aes(x = WORK_LIFE_BALANCE_SCORE, col = AGE,fill=AGE)) +
  geom_density(alpha=0.3) +
  theme_minimal() +
  labs(
    x = "work_life_balance_score",
    y = "density"
  )
ggsave("age-score.png", plot = last_plot(), width = 6, height = 4, units
= "in", dpi = 300, bg = "white")

#不同性别 score 分布
data$GENDER = as.factor(data$GENDER)
ggplot(data, aes(x = WORK_LIFE_BALANCE_SCORE, col = GENDER,fill=GENDER))
+
  geom_density(alpha=0.3) +
  theme_minimal() +
  labs(
    x = "work_life_balance_score",
    y = "density"
  )
ggsave("gender-score.png", plot = last_plot(), width = 6, height = 4,
units = "in", dpi = 300, bg = "white")

##########################################
###PCA
lifedata = data[,c(1,3,4,5,10,12,14,16,20)]
lifedata = scale(lifedata)
welldata = data[,c(2,7,19)]
welldata = scale(welldata)
```

```
library(stats)
lifepca = princomp(lifedata,scores = T,cor = T)

evals<-data.frame(lifepca$sdev^2)
names(evals)<-"eigen.vals"
evals$component.num<-as.integer(seq(nrow(evals)))
ggplot(evals,aes(x=component.num,y=eigen.vals))+geom_point()

pca_scores = lifepca$scores[, 1:5]

wellpca = princomp(welldata)
well_pca_scores = wellpca$scores[,1:2]

##########################################
###FA
fit1 <- factanal(lifedata,factors=5,rotation="varimax")

##########################################
###Clustering
dat = as.data.frame(pca_scores)
dat$comp1 = well_pca_scores[,1]
dat$comp2 = well_pca_scores[,2]
dat$class = as.factor(data$class2)

##k-means
library(factoextra)
fviz_nbclust(welldata, kmeans, method = "wss") + geom_vline(xintercept =
4, linetype = 2) + theme(plot.title = element_text(hjust = 0.5))

#4 Clusters
res4<-kmeans(welldata,4)
dat$cluster4 = as.factor(res4$cluster)
ggplot(dat, aes(x = comp1, y = comp2, color = as.factor(cluster4))) +
  geom_point() +
  labs(title = "4 Clusters based on First 2 Principal Components",
       x = "PC1",
       y = "PC2",
       color = "Cluster4") +
  theme_minimal()
ggsave("cluster4.png", plot = last_plot(), width = 6, height = 4, units
= "in", dpi = 300, bg = "white")

#3 Clusters
```

```
res3 = kmeans(welldata,3)
dat$cluster3 = as.factor(res3$cluster)

ggplot(dat, aes(x = comp1, y = comp2, color = as.factor(cluster3))) +
  geom_point() +
  labs(title = "3 Clusters based on First 2 Principal Components",
       x = "PC1",
       y = "PC2",
       color = "Cluster3") +
  theme_minimal()
ggsave("cluster3.png", plot = last_plot(), width = 6, height = 4, units
= "in", dpi = 300, bg = "white")

#2 clusters
res2 = kmeans(welldata,2)
dat$cluster2 = as.factor(res2$cluster)

ggplot(dat, aes(x = comp1, y = comp2, color = as.factor(cluster2))) +
  geom_point() +
  labs(title = "2 Clusters based on First 2 Principal Components",
       x = "PC1",
       y = "PC2",
       color = "Cluster2") +
  theme_minimal()
ggsave("cluster2.png", plot = last_plot(), width = 6, height = 4, units
= "in", dpi = 300, bg = "white")

#和 score 对比
ggplot(dat, aes(x = comp1, y = comp2, color = class)) +
  geom_point() +
  labs(title = "Score Class based on First 2 Principal Components",
       x = "PC1",
       y = "PC2",
       color = "Class") +
  theme_minimal()
ggsave("score.png", plot = last_plot(), width = 6, height = 4, units =
"in", dpi = 300, bg = "white")

mds=cmdscale(dist(welldata[1:600,],method="euclidean"))
par(mfrow=c(1,2))
plot(mds,col=dat$cluster2,main='kmeans k=2',pch=18)
plot(mds,col=dat$class,main='Original clusters',pch=18)

##########################################
```

```
###DA
#lda
library(MASS)
lifedata = as.data.frame(lifedata)
lifedata$cluster2 = as.factor(dat$cluster2)
lifedata$cluster3 = as.factor(dat$cluster3)
L = lda(cluster3~.,data=lifedata)
yhat = predict(L, lifedata)$class
tab = table(pred=yhat, true=lifedata$cluster3)
dat$lda = as.factor(yhat)
ggplot(dat[1:500,],aes(x=Comp.1,y=Comp.2,col=lda,shape=cluster3))      +
geom_point()
ggsave("lda3.png", plot = last_plot(), width = 7, height = 4, units =
"in", dpi = 300, bg = "white")

L2 = lda(cluster2~.,data=lifedata)
yhat2 = predict(L2, lifedata)$class
tab2 = table(pred=yhat2, true=lifedata$cluster2)
dat$lda2 = as.factor(yhat2)
ggplot(dat[1:500,],aes(x=Comp.1,y=Comp.2,col=lda2,shape=cluster2))     +
geom_point()
ggsave("lda2.png", plot = last_plot(), width = 7, height = 4, units =
"in", dpi = 300, bg = "white")

#CV LDA
train = lifedata[1:11180,]
test = lifedata[11181:15971,]
L3 = lda(cluster2~.,data=train)
yhat3 = predict(L3, test)$class
yhat4 = predict(L3,train)$class
tab4 = table(pred = yhat4,true = train$cluster2)
tab3 = table(pred=yhat3, true=test$cluster2)

#qda
lifedata$cluster3 = dat$cluster3
Q = qda(cluster3~.,data=lifedata)
yhatq = predict(Q, lifedata)$class
tabq = table(pred=yhatq, true=lifedata$cluster3)
dat$qda = as.factor(yhatq)
ggplot(dat[1:500,],aes(x=Comp.1,y=Comp.2,col=qda,shape=cluster3))      +
geom_point()
ggsave("qda3.png", plot = last_plot(), width = 7, height = 4, units =
"in", dpi = 300, bg = "white")
```