探索性数据降维分析

> 本报告主要包含以下内容：
>
> 1. 数据介绍
> 2. 基本原理介绍
> 3. 结合案例数据进行分析
> 4. 最后总结
> 5. 附上代码和参考

数据介绍

本报告所使用的是洛杉矶街区数据，其中包含每个街区的名字、收入中位数、公立学校API中位数、种族多样性、年龄中位数、有房家庭占比等14项字段，共有110个观测数据。本报告的主要目的是对这个数据的字段（变量）进行分析，并且探索性地尝试使用主成分分析和因子分析等降维方法来对数据进行降维分析。

基本原理介绍

主成分分析

主成分分析是一种降维方法，通过原始数据一系列的线性变换找到对数组总体变异性（信息量）贡献比较大的主成分，这些线性变换保留了原始变量的大部分信息。其中，总体的变异性是指总体中所包含的信息，保留了原始变量的大部分信息是指保留了大部分的变异（PCA中使用的是方差）信息。

主成分分析步骤：

设$X_1,\ldots,X_p$为p个随机变量。

1. 将原始数据标准化，得到标准化数据矩阵

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

2. 计算变量X的相关系数矩阵：$R=(r_{ij})_{p\times p}=X^{'}X$
3. 求R的特征值$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p>0$及相应的单位特征向量。

$$u_1=\begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1p} \end{bmatrix},u_2=\begin{bmatrix} u_{21} \\ u_{22} \\ \vdots \\ u_{2p} \end{bmatrix},\cdots,u_1=\begin{bmatrix} u_{p1} \\ u_{p2} \\ \vdots \\ u_{pp} \end{bmatrix}$$

4. 将特征值排序，计算方差贡献率和累计方差贡献率，找到达到累计贡献率的前几个特征值$\lambda_1,\lambda_2,\cdots,\lambda_q$所对应的特征向量$u_1,u_2,\cdots,u_q$，不妨设我们所需要的累积贡献率为80%，即$\sum\limits_{i=1}^{q}\lambda_i\geq0.8$。
5. 确定主成分，用原指标的线性组合来计算各主成分得分：以各主成分对原指标的相关系数为权，将各主成分表示远原指标的线性组合，即：

$$C_j=u_{j1}x_1+u_{j2}x_2+\cdots+u_{jp}x_p,\quad j=1,2,\cdots,1$$

主成分分析的主要理论结果：

1. 对任意$1 \leq i \leq p$，第i各主成分的系数向量a_i取为u_i，因此第i个主成分是$Y_i = u_{i1}x_1 + u_{i2}x_2 + \cdots + u_{ip}x_p$.
2. 对任意$1 \leq i \leq p, Var(Y_i) = \lambda_i$.
3. 对任意$1 \leq i \neq j \leq p, Cov(Y_i, Y_j) = 0$，即$Y_i$与$Y_j$不线性相关。
4. 数据经过以上的标准化之后，总方差为$\sum_{i=0}^{p} \lambda_i = p$，，第$i$个主成解释的方差比例为$\lambda_i/p$，前$q$个主成分解释总方差的比例为$\sum_{i=1}^{q} \lambda_i/p$.
5. 对任意$1 \leq i, k \leq p$, Y_i与X_k的相关系数为$Corr(Y_i, X_k) = e_{ik}\sqrt{\lambda_i}$.

主成分个数的选择：

1. Kaiser准则：保留那些对应特征值大于所有特征值的平均之的主成分，解释总方差比例大于平均解释比例的主成分，在这里数据标准化周就保留对应的特征值大于1的主成分。
2. 总方差中被前q个主成分解释的比例达到一定大小**（常用）**。
3. 保留的主成分在实际应用中具有可解释性。
4. 使用崖底碎石图绘出特征值与其顺序的关系，找到一个拐点，使得此节点后对应的特征向量都比较小，然后选择拐点之前的一点。

主成分的含义：

1. 对第i个主成分，选择主成分中原变量系数绝对值比较大的变量，作为主成分的主要解释。
2. 根据相关系数来解释，计算第i个主成分与各个原始变量之间的相关系数，根据相关系数比较大的来作为主成分的解释。

探索性因子分析

不同主成分分析直接将将一些对总体变异贡献率比较高的特征值所对应的特征向量，来对原始变量进行线性组合，得到新的主成分。因子分析主要是将变量的变异性解释为由两种类型的因子的变异，分别是潜在的、不可观测的公共因子和只与该变量有关的特殊因子。这里因子分析主要是寻找少量公共因子（对应于主成分分析的主成分）来解释一组输入变量的共同的变异性。

一些符号：

假设$X = (x_1, x_2, \cdots, x_n)$是一个$p$维随机向量，它的均值向量为$\mu = (\mu_1, \mu_2, \cdots, \mu_p)^T$，协方差矩阵为$\Sigma$，其对角线上的值$\sigma_k^2$表示$X_k$的方差$(k = 1, \cdots, p)$，令$F_1, \cdots, F_q (q \leq p)$表示q个潜在的公共因子，令$\epsilon_1, \cdots, \epsilon_q$表示特殊因子。正交矩阵模型：

$$\begin{aligned} X_1 - \mu_1 =& l_{11}F_1 + l_{12}F_2 + \cdots + l_{1q}F_q + \epsilon_1 \\ X_2 - \mu_2 =& l_{21}F_1 + l_{22}F_2 + \cdots + l_{2q}F_q + \epsilon_2 \\ \cdots & \quad \cdots \\ X_p - \mu_p =& l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pq}F_q + \epsilon_q \end{aligned}$$

写成矩阵形式$X - \mu = LF + \epsilon, F = (F_1, \cdots, F_q)$是共因子， $\epsilon = (\epsilon_1, \cdots, \epsilon_q)$是特殊因子，$L$是载荷矩阵，其中第$k$行第$i$列的值$l_{ki}, k = 1, \cdots, p, i = 1, \cdots, q$表示$X_k$在因子$F_i$上的载荷。

由于公共因子和特殊因子是不可观测的，所以需要作出**一些假定**：

1. $E(F_i) = 0, Var(F_i) = 1, \forall 1 \leq i \leq q.$

   $Cov(X_i, X_k) = 0, \forall i, k = 1, \cdots, q$
2. $E(\epsilon_k) = 0, Var(\epsilon_k) = \Psi_k$

   $Cov(\epsilon_k, \epsilon_m) = 0, \forall k, m = 1, \cdots, q$
3. $Cov(F_i, \epsilon_k) = 0, \forall 1 \leq i \leq q, 1 \leq k \leq p$

主要结论：

1. $Var(X_k)=l_{k1}^2+l_{k2}^2+\cdots+l_{kq}^2+\Psi_k$

   $l_{k1}^2+l_{k2}^2+\cdots+l_{kq}^2$平方和是方差中能被公共因子解释的部分，称为共性方差。

   Ψ_k是不能被公共因子解释的部分，成为X_k的特殊方差。
2. $Cov(X_k,X_m)=l_{k1}l_{m1}+l_{k2}l_{m2}+\cdots+l_{kq}l_{mq},\forall 1\leq k\neq m\leq p$
3. $Cov(X_k,F_i)=l_ki,\forall k=1,\cdots,p,i=1,\cdots,q$

对**公共因子的解释**，同主成分分析，对F_i载荷系数的绝对值较大的输入的变量来解释。

载荷矩阵估计：

主要有主成分法（主因子法）和最大似然估计法。

主成分法将协方差Σ（标准化后为相关系数矩阵R）拆分为$\Sigma=\lambda_1e_1e_1^T+\lambda_2e_2e_2^T+\cdots+\lambda_pe_pe_p^T$，将式子中前$q$项归结为主因子解释的部分，后面$p-q$项归结为特殊因子，得到$L$和$\Psi$的估计$\tilde{L}$和$\tilde{\Psi}$。

最大似然估计法：

假定公共因子F和特殊因子ϵ服从正态分布，可以得因子载荷的最大似然估计。

得到载荷矩阵之后，如果得到的因子对原始变量的可解释性不强，则还需要进行因子旋转，以得到较为明显的实际含义。

旋转后的载荷矩阵满足以下几个条件：

（1）对于任意因子$F_i,i=1,\cdots,q$而言，只有少数输入变量在该因子上的绝对值较大，其他都接近于0.

（2）对任意输入变量$X_1,\cdots,X_p$而言，它只在少数因子上的载荷$l_{ki},k=1,\cdots,p,i=1,\cdots,q$的绝对值较大，在其他因子上的载荷接近于0.

（3）任何两个因子对应的载荷呈现不同的模式，因为在解释时这两个因子具有不同的含义。

比如对于因子F_1,F_2,对于第一个变量X_1，因子F_1的载荷是l_{11}，因子F_2的载荷是l_{12}

$$X_1-\mu_1=l_{11}F_1+l_{12}F_2+\cdots+l_{1q}F_q+\epsilon_1$$

多维标度分析

介绍：它是一个在低维空间展现高维数据的可视化方法，使得低维空间中观测点之间的距离与高维空间中观测点之间的距离大致匹配。

多维度标度根据度量的形式分为**度量**和**非度量**，度量形式直接采用观测点之间的距离，非度量形式采用观测点之间距离的排序。这里介绍非度量形式。

假设一共有N个观测，他们之间两两匹配的对数为$M=N(N-1)/2$，计算高维空间中M对观测数据的距离，对其近似排序可以得到

$$d_{i_1k_1}^*<d_{i_2k_2}^*<\cdots<d_{i_Mk_M}^*$$

设低维空间的维度为q，将N个观测点放置在q维空间中，即每个观测点用一个q维坐标向量代表。令$d_{i_jk_j}^{(q)}$来表示q维空间中观测点之间的距离。

给出一些定义：

应力函数：

$$Stress(q)=\{\frac{\sum_{j=1}^M(d_{i_jk_j}^{(q)}-\hat{d}_{i_jk_j}^{(q)})^2}{\sum_{j=1}^M[d_{i_jk_j}^{(q)}]^2}\}^{1/2}$$

或者：

$$SStress(q)=\{\frac{\sum_{j=1}^{M}[(d_{i_jk_j}^{(q)})^2-(\hat{d}_{i_jk_j}^{(q)})^2]^2}{\sum_{j=1}^{M}[d_{i_jk_j}^{(q)}]^4}\}^{1/2}$$

应力函数的值越小，低维空间中观测点之间距离的排序与原来高维空间中观测点之间的距离的排序的一致性越高。

算法步骤：

1. 初始化N个观测点在q维空间的坐标向量，并据此计算$d_{i_jk_j}^{(q)},j=1,\cdots,M$。
2. 在每次循环中：

   （1）固定$d_{i_jk_j}^{(q)},j=1,\cdots,M$，寻找M个数值$\hat{d}_{i_jk_j}^{(q)},j=1,\cdots,M$，使得他们的排序与$d_{i_jk_j}^{*},j=1,\cdots,M$的排序完全一致，并且应力函数的值达到最小；

   （2）固定$\hat{d}_{i_jk_j}^{(q)},j=1,\cdots,M$，寻找N个观测点在q维空间的坐标向量，使得应力函数的值达到最小。

持续循环直到应力函数的值无法减小为止。

案例分析

尝试对数据进行主成分降维

在进行主成分分析之前，先对数据标准化，这里R语言中的princomp函数就在将原始数据输入进去之后就已经是标准化数据了，所以我们不需要再对数据进行标准化，直接将原始数据输入到princomp函数中去即可。

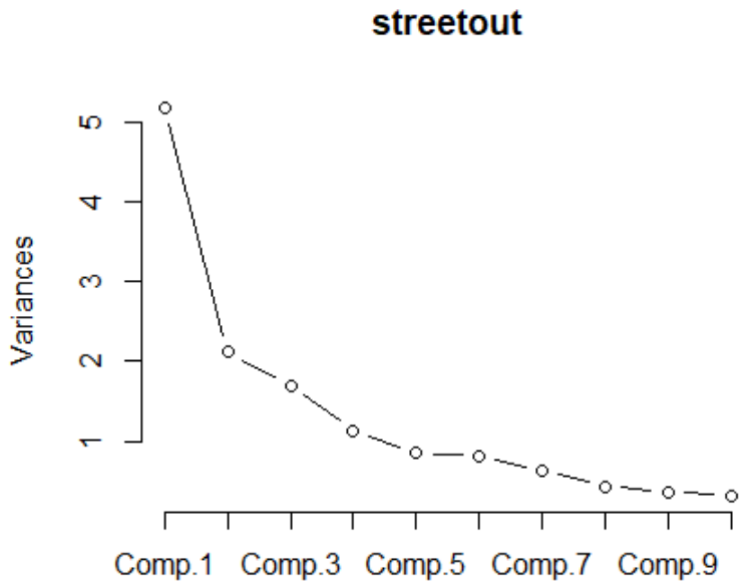
在进行主成分分析之后，查看数据的因子载荷，找到对总体变异性贡献比较大的q个主成分。

```
> summary(streetout,loadings=T)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4    Comp.5
Standard deviation     2.276290 1.4526977 1.3046265 1.06283297 0.9180742
Proportion of Variance 0.370107 0.1507379 0.1215750 0.08068671 0.0602043
Cumulative Proportion  0.370107 0.5208449 0.6424199 0.72310664 0.7833109
                           Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     0.90433923 0.79680013 0.65673588 0.60545708 0.55773624
Proportion of Variance 0.05841639 0.04534932 0.03080729 0.02618416 0.02221927
Cumulative Proportion  0.84172733 0.88707664 0.91788393 0.94406809 0.96628736
                          Comp.11    Comp.12     Comp.13      Comp.14
Standard deviation     0.47952347 0.40015285 0.285223173 2.365740e-02
Proportion of Variance 0.01642448 0.01143731 0.005810876 3.997663e-05
Cumulative Proportion  0.98271184 0.99414915 0.999960023 1.000000e+00
```

如上图所示，第一个主成分对总体数据变异的贡献率为0.370107，第二个主成分对总体贡献率为0.1507379，此时的累计贡献率为0.5208449.

这里我们选择使用**Kaiser准则**来找到前q个主成分：

首先画图：

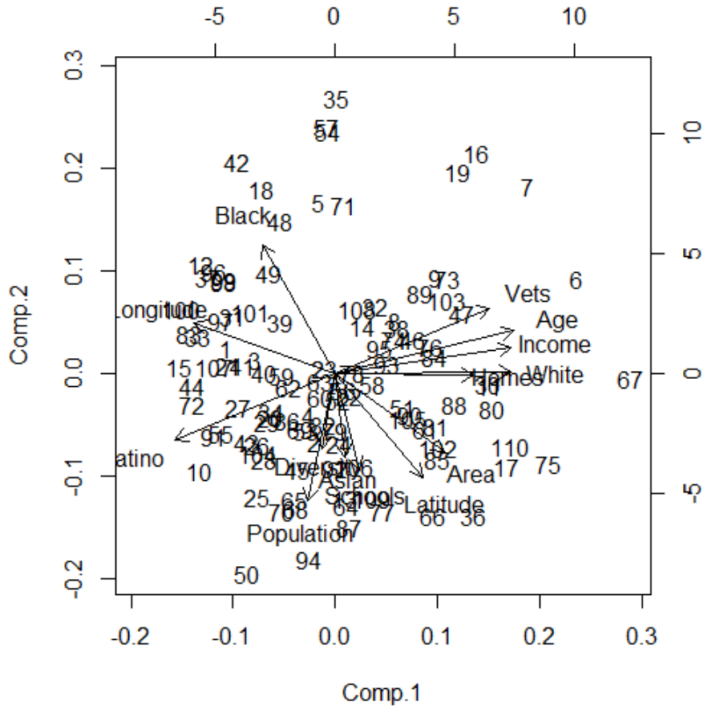


由图可知，**第五个主成分是碎石图的一个拐点**，在此点之前，特征值都比较大，在此点之后，特征值都比较小，所以根据Kaiser准则，我们选择前4个主成分来作为代表原始变量，进而达到降维的目的。

```
Loadings:
           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14
Income      0.385                0.209                0.179                 0.476   0.720
Schools            -0.330        -0.327  0.816  0.121 -0.284
Diversity          -0.257 -0.487 -0.477 -0.183 -0.262         0.268  0.116 -0.440          0.273
Age         0.391   0.146 -0.177                              -0.302                        -0.144  -0.809
Homes       0.302                -0.199         0.556  0.367         0.552  0.125  -0.317
Vets        0.338   0.224        -0.337                0.218 -0.115 -0.119 -0.310  0.481  -0.409   0.380
Asian              -0.286 -0.617                       0.253 -0.187 -0.295  0.496   0.131          0.200  -0.195
Black      -0.160   0.438  0.134 -0.558                       -0.230 -0.145  0.285          0.308          -0.427
Latino     -0.351  -0.228  0.159                0.354  0.188  0.235        -0.165  0.362  -0.124  -0.280  -0.563
White       0.386                0.282  0.116 -0.283 -0.179         0.160 -0.173  -0.279          0.216  -0.680
Population         -0.438  0.351               -0.439  0.243 -0.469  0.376  0.123   0.171   0.110
Area        0.238  -0.270  0.385                       0.433  0.183 -0.583         -0.378
Longitude  -0.309   0.174 -0.146  0.237  0.263  0.147  0.317 -0.492 -0.104 -0.504  -0.135   0.256   0.125
Latitude    0.191  -0.355               -0.431  0.415 -0.491 -0.388 -0.182 -0.178          0.116
```

再对这里的主成分的含义进行解释，由上图所示，对于第一个主成分，它代表是是总体情况，第一个主成分主要由Income, Age, White三个变量来解释；第二个主成分主要有Black, Population, Latitude来解释；第三个主成分主要由Diversit, Asian, Area来解释。依次类推可以对剩下的一个主成分进行解释。

尝试使用双标图来将各个观测和各个变量绘制的同一张图上。得到



由图可知，第19，16，7观测为异常观测。图中第一个主成分在Latitude和Area等变量上的系数为正，在Latino和Black等变量上的系数为负。

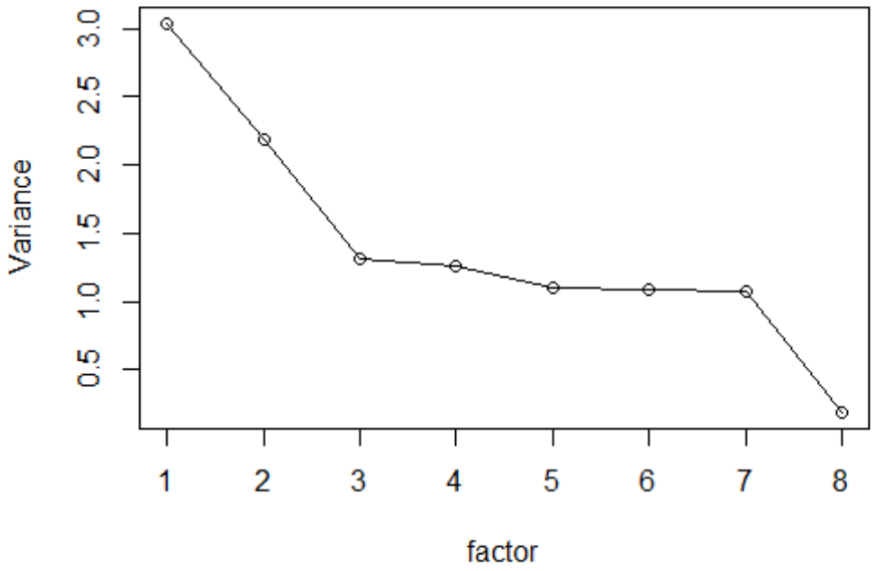
最终降维结果：取成分的个数为4，最后将原始14个变量降维至4个变量。

尝试使用因子分析来降维

再接着尝试使用因子分析来降维，再R语言中的factanal函数中的代表因子个数的参数factors设置为8，得到因子分析的结果：

```
                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7 Factor8
SS loadings       3.037   2.183   1.311   1.265   1.099   1.082   1.081   0.188
Proportion Var    0.217   0.156   0.094   0.090   0.079   0.077   0.077   0.013
Cumulative Var    0.217   0.373   0.467   0.557   0.635   0.713   0.790   0.803
```

这暂时还不能够直观地看出在探索性因子分析中，选的多少个因子才是合适的，因此我们根据以上因子分析的因子方差来画图：



由上图可知，因子个数取3是一个拐点，结合上面的累计误差知道当取因子个数为2时，累计误差仅为0.373，这两个因子并不能够很好地代表整个总体。所以我们结合Kaiser准则和总方差中的累计贡献率，假设我们要求在80%左右就可以了，这里我们**选择因子个数为7**，这时7个因子在总方差中的累计贡献率近似达到80%。而且上图也表明，当因子个数超过7时，累计贡献率会从1.081急剧下降到0.188.

最终选择因子个数为7，得到的载荷矩阵为：

```
Loadings:
           Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
Income      0.548   0.537   0.273           0.277  -0.317
Schools                             0.325
Diversity                                           0.734   0.221
Age         0.728   0.565   0.135                           0.116
Homes       0.238   0.618                   0.212
Vets        0.455   0.781  -0.167           0.130          -0.198
Asian                       0.253                   0.463   0.845
Black      -0.127           -0.926  -0.258  -0.118          -0.202
Latino     -0.941  -0.260   0.126           -0.133
White       0.846   0.222   0.408           0.196  -0.113
Population -0.189  -0.177           0.636                  -0.111
Area        0.136   0.410           0.698   0.209  -0.242
Longitude  -0.342  -0.255          -0.273  -0.850
Latitude            0.211   0.264   0.227   0.396
```

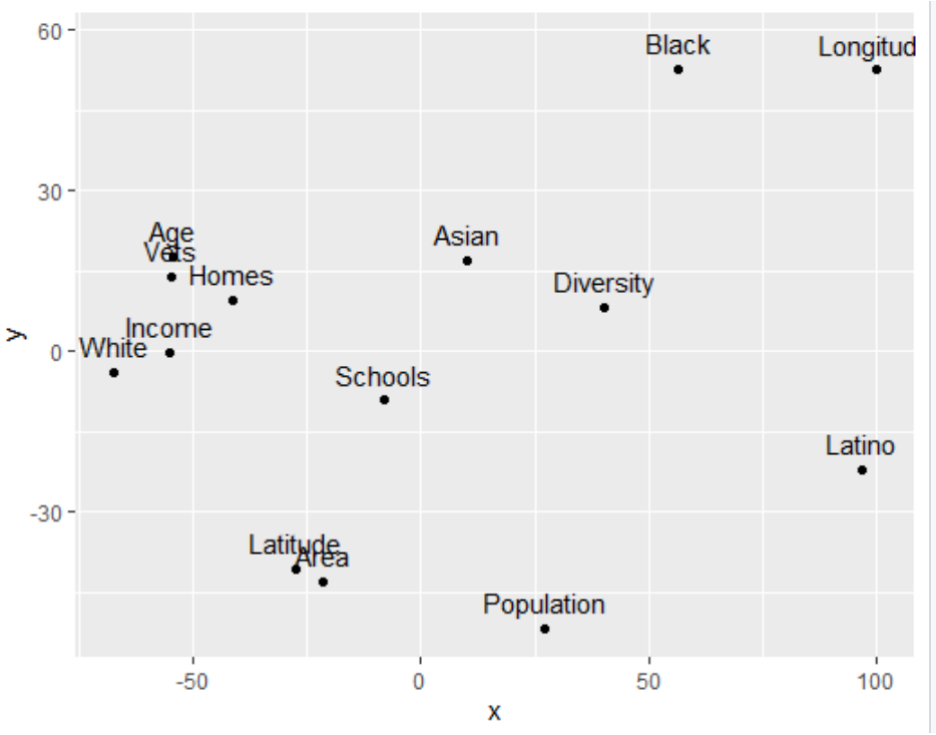
由图可知，第一个因子可以很好地由Lantino, White来解释；第二个因子可以由第Vets来解释；第三个因子可以由Black解释。此外，黑人和白人是不相关的，这符合因子分析假设，各个因子之间是不相关的假设。

最终降维结果：取因子的个数为7，最后将原始14个变量降维至7个变量。

尝试使用多维标度分析来降维查看数据

分析的流程：首先，将原始110行，14列的数据转置，得到14列，110行的数据。再求出利用曼哈顿的度量方法来找到数据之间的距离。将数据转换成矩阵形式，使用非度量的方式来分析原始数据，得到应力值为14.12094. 这说明使用2维空间对原来高维空间拟合的效果一般。

当低维维度设置为2时，进行排序之后的数据进行排序得到：



由图所示，在低维空间中，Income收入和White白种人距离比较接近，Latitude和Area比较接近，Longitude以及Latino距离其他变量都比较远。

因为当将低维维度设置为2时，效果不好，所以这里我们再尝试使用其他的较低的维度来对原来高维空间进行拟合，发现将**低维的维度设置为7，应力值才能够降到1以下。**效果都不太好，所以最后**不考虑使用多维标度分析来进行降维。**

总结

本报告首先介绍了原始数据，再比较详细地分别介绍了主成分分析，因子分析以及多标度分析的原理，并且还结合案例数据来探索进行数据降维。

在因子分析选择主要因子个数时，仅凭借个人直觉判定应该首先使用碎石图来进行初步地判断，再结合因子对总方差的累计贡献率，最终选择因子个数为7，感觉有些多理论性又不强，是否正确我觉得在后续的学习中还需要进一步地学习去验证。

模型最后多标度分析并只尝试了使用使用二维空间的数据来对高维空间进行拟合，这里得到的应力函数的值维14.12094，使用课本上的评价方法来说，是很一般的拟合，在后续的探索中我发现，随着维度的增大，对原始数据的拟合的应力值也是相应增大的，但是考虑到篇幅的问题就没有添加上去，后续要需要继续探索学习。

参考

[1]高惠璇.应用多元统计分析[M] 北京:北京大学出版社.2008.

[2]王斌会.多元统计分析及R语言建模[M]广州:暨南大学出版社.2014.

[3] 如何理解“方差越大信息量就越多”?

[4] 机器学习实战之PCA

[5] 一篇深入剖析PCA的好文

[6] PCA：详细解释主成分分析

[7] R语言 | 典型相关分析、多维标度法以及综合评价方法及R使用

[8] R语言 3.14 多维标度法MDS

代码

```
##加载程序包
library(dplyr)
library(ggplot2)

# ---------读入数据---------
setwd("D:/lagua/CODING/R-learn/R-code/Chap5_DimensionReduction")
street <- read.csv("LANeighborhoods.csv", header=T, skip=1)
colnames(street)
# 去除掉第一列，因为后面需要计算相关系数矩阵
street = street[, -1]
summary(street)


# ---------主成分分析---------
help("princomp")
streetout <- princomp(street,cor = T,scores = T)
summary(streetout,loadings=T)

streetout <- princomp(scale(street),cor = T,scores = T)
#streetout$scores记录了每个观测的主成分得分。
streetout$scores
##显示分析结果
summary(streetout,loadings=T)

##画崖底碎石图
screeplot(streetout,type = "lines")

##画前两个主成分的双标图
biplot(streetout,choices = 1:2,col="black")
help(biplot)


# ---------因子分析---------
# help("factanal")
# 使用极大似然估计，
# 因子旋转：方差最大
streetout <- factanal(scale(street),fm="ml",factors = 8,rotation = "varimax")
streetout # 查看结果
# 画碎石图，查看方差变化情况，以便选择因子个数
plot(c(3.037,   2.183,   1.311,   1.265,   1.099,   1.082,   1.081,   0.188),
     type="o",
     xlab="factor",
     ylab="Variance")
```

```r
# 选择因子个数为7
streetout <- factanal(scale(street),fm="ml",factors = 4,rotation = "varimax")
streetout
plot(c(3.037,   2.183,   1.311,   1.265),
     type="o",
     xlab="factor",
     ylab="Variance")


# 显示分析结果
streetout
#streetout数据集包含两个公共因子的载荷矩阵Loadings,
summary(streetout)


# ---------多维标度分析---------
tmpstreet <- t(scale(street))
help(dist)
# 求出距离矩阵，  使用L_1范数--曼哈顿距离
diststreet <- dist(tmpstreet,method = "manhattan")


# ---------度量形式
# help("cmdscale")
out <- cmdscale(diststreet) %>% as.data.frame()
#使用cmdscale函数进行多维标度分析。

ggplot(out,aes(x=V1,y=V2))+
  geom_point()+
  geom_text(aes(y=V2+5,label=row.names(out)))

# ---------非度量形式
library(MASS)
D = as.matrix(diststreet) # 转化为矩阵形式
help("isoMDS")
MDS = isoMDS(D, k=2) # 非度量形式的多维标度分析
# MDS
MDS$stress # 查看应力值
x = MDS$points[, 1]
y = MDS$points[, 2]
# 对代表高维空间数据的低维空间数据画图
ggplot(out,aes(x=x,y=y))+
  geom_point()+
  geom_text(aes(y=y+5,label=row.names(D)))


# ----------查看其他更低维度的应力值
for (i in 3:9){
  MDS = isoMDS(D, k=i) # 非度量形式的多维标度分析
  # MDS
  print(MDS$stress) # 查看应力值
}
```