多元-07-2020270026

2020270026 王姿文

4/20/2021

1. 数据

- 数据叙述：数据为国家的一些特征，一共是rows = 50, columns = 8，故共有8个维度。
- 目标： 进行主成分分析

下表为其中几笔数据，以及数据的结构：

```
data(state)
kbl(head(state.x77)) %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, font_si
ze = 7)
```

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad	Frost	Area
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417
Arkansas	2110	3378	1.9	70.66	10.1	39.9	65	51945
California	21198	5114	1.1	71.71	10.3	62.6	20	156361
Colorado	2541	4884	0.7	72.06	6.8	63.9	166	103766

2. PCA

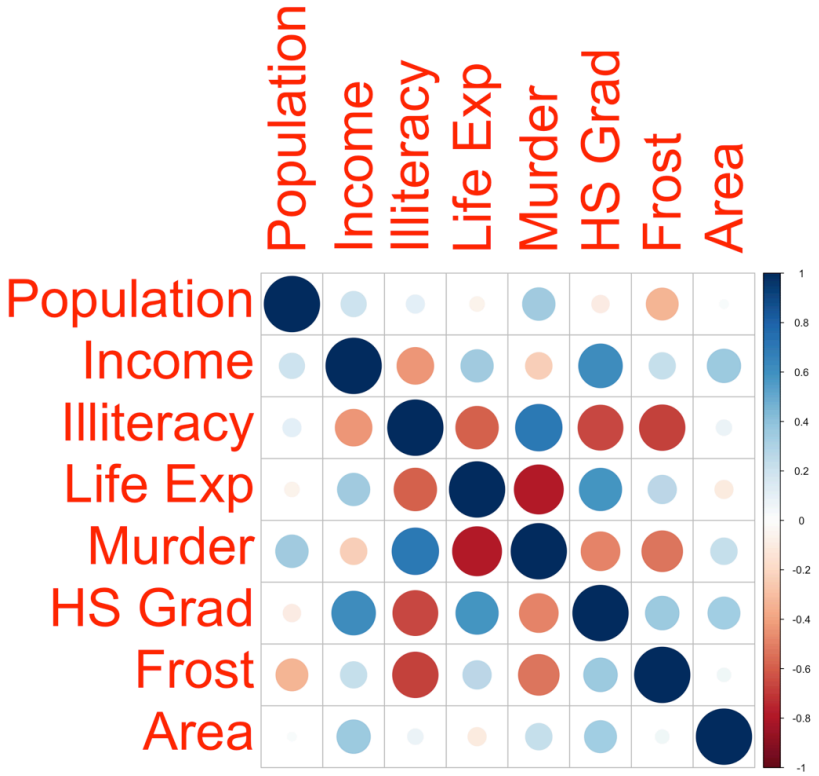
因为各列不可比，所以后续的pca分析需要用样本相关阵。
其中只有 `Murder vs Illiteracy 、 Murder vs Life Exp` 的绝对值相关性>0.7：

```
kbl(round( cor(state.x77), 3)) %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, font_si
ze = 7)
```

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad	Frost	Area
Population	1.000	0.208	0.108	-0.068	0.344	-0.098	-0.332	0.023
Income	0.208	1.000	-0.437	0.340	-0.230	0.620	0.226	0.363
Illiteracy	0.108	-0.437	1.000	-0.588	0.703	-0.657	-0.672	0.077
Life Exp	-0.068	0.340	-0.588	1.000	-0.781	0.582	0.262	-0.107
Murder	0.344	-0.230	0.703	-0.781	1.000	-0.488	-0.539	0.228
HS Grad	-0.098	0.620	-0.657	0.582	-0.488	1.000	0.367	0.334

Frost	-0.332	0.226	-0.672	0.262	-0.539	0.367	1.000	0.059
Area	0.023	0.363	0.077	-0.107	0.228	0.334	0.059	1.000

```
corrplot(cor(state.x77),tl.cex=4)
```



开始做主成份分析，可以看到在Comp=5时，可解释累积变异就>90%，因此选择五个维度变可解释93%的原始数据。此外，若使用两个主成分的话，其可解释累积变异为65%，后续可就前两维来画图看一些insight。

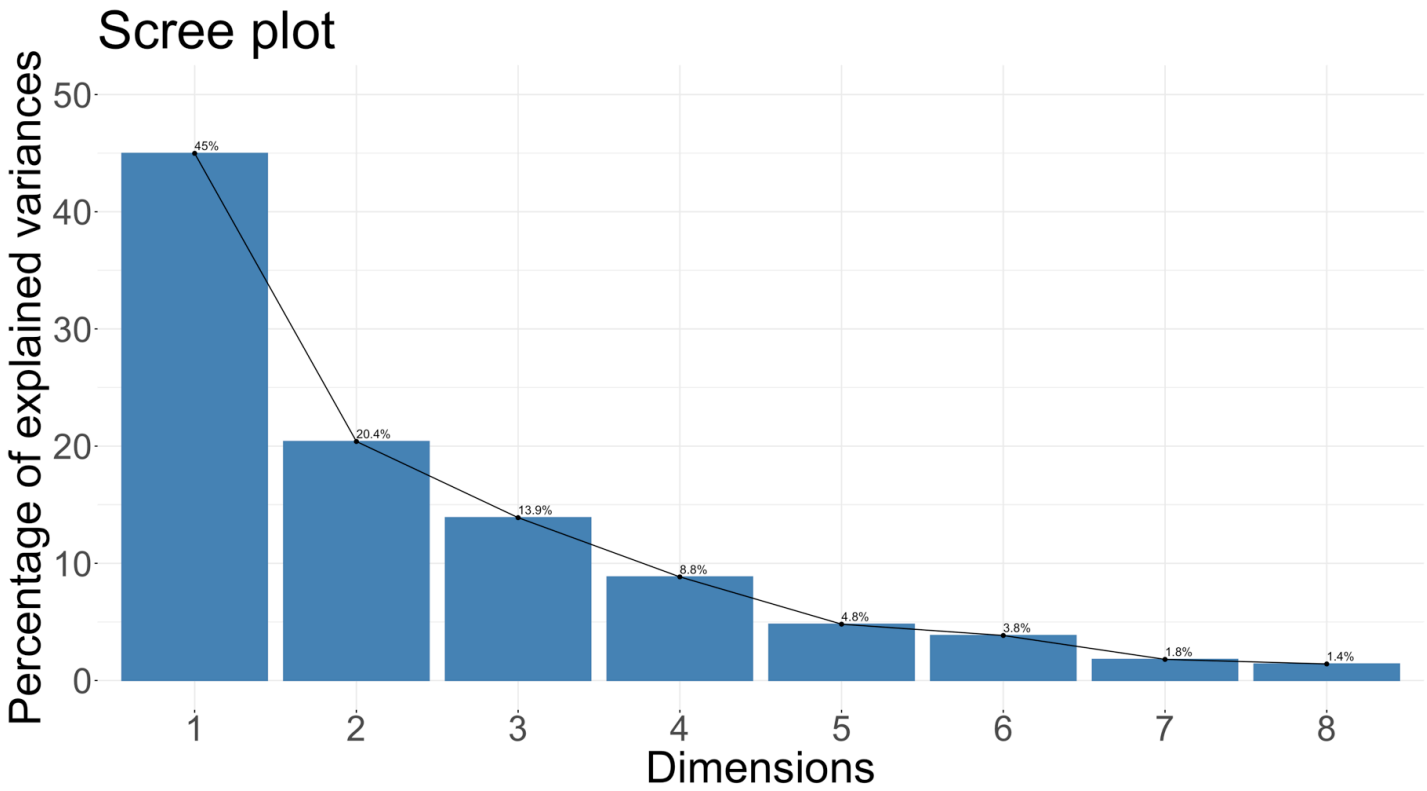
而我们也能从loadings来判断不同主成分的性质，例如：Comp.1主要是 `Illiteracy` 、 `Life Exp` 、 `Murder` 、 `HS Grad` 的值较大，故第一主成分是生活和犯罪成分（而且基本上就是上述说说到相关系数较大的变量）；Comp.2主要是 `Population` 、 `Income` 、 `Area` 的值较大，故第二主成分是地理资讯和收入成分...等以此类推。

```
pca1 <- princomp(state.x77, cor=TRUE)
summary(pca1, loadings=TRUE)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     1.8970755 1.2774659 1.0544862 0.84113269 0.62019488
## Proportion of Variance 0.4498619 0.2039899 0.1389926 0.08843803 0.04808021
## Cumulative Proportion  0.4498619 0.6538519 0.7928445 0.88128252 0.92936273
##                            Comp.6    Comp.7     Comp.8
## Standard deviation     0.55449226 0.3800642 0.33643379
## Proportion of Variance 0.03843271 0.0180561 0.01414846
## Cumulative Proportion  0.96779544 0.9858515 1.00000000
##
## Loadings:
##            Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## Population  0.126  0.411  0.656  0.409  0.406                 0.219
## Income     -0.299  0.519  0.100        -0.638 -0.462
## Illiteracy  0.468                -0.353        -0.387  0.620  0.339
## Life Exp   -0.412         0.360 -0.443  0.327 -0.219  0.256 -0.527
## Murder      0.444  0.307 -0.108  0.166 -0.128  0.325  0.295 -0.678
## HS Grad    -0.425  0.299        -0.232         0.645  0.393  0.307
## Frost      -0.357 -0.154 -0.387  0.619  0.217 -0.213  0.472
## Area               0.588 -0.510 -0.201  0.499 -0.148 -0.286
```

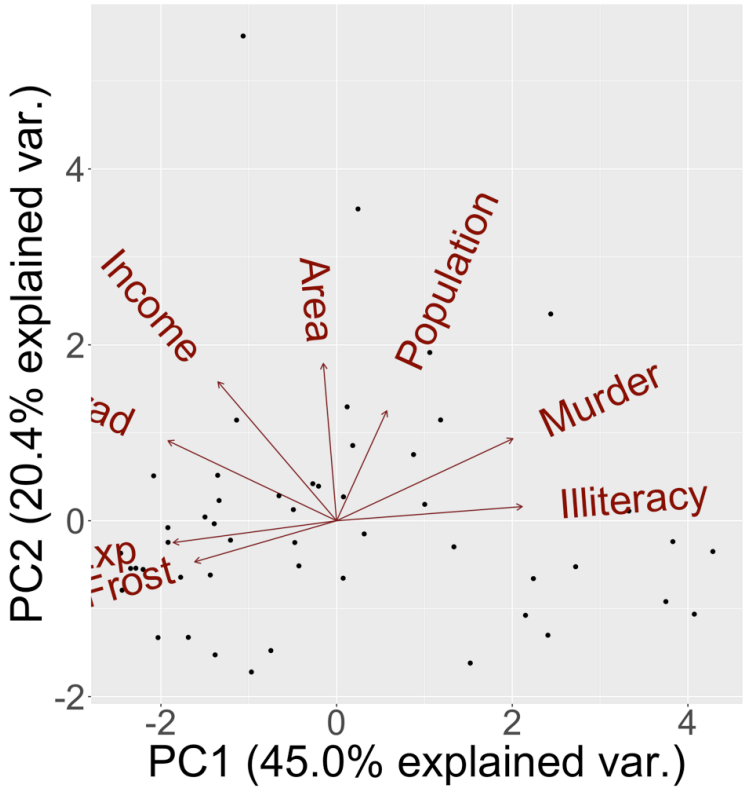
可以从**Scree plot&**简单看出每一个主成分的可解释变异占比：

```
state.x77.pca <- PCA(state.x77, graph = FALSE,scale.unit = TRUE)
fviz_eig(state.x77.pca, addlabels = TRUE, ylim = c(0, 50)) +
  theme(text = element_text(size = 40))
```



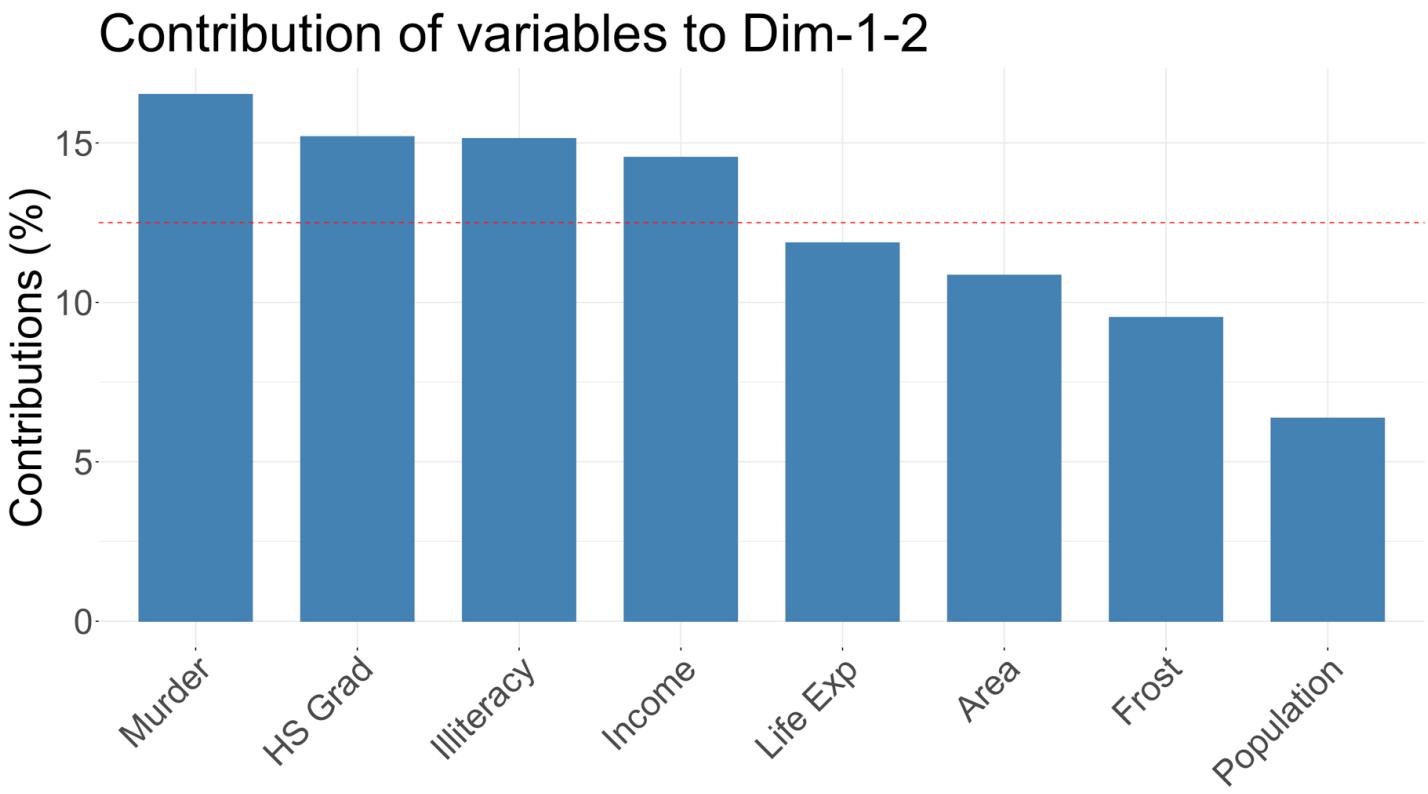
下图为前两个主成分的loadings所画出来的原始变量表现，可以看出在第一个主成分是 `Illiteracy` 、 `Life Exp` 、 `Murder` 、 `HS Grad` 的值较大，并且可以看出正负；在第二个主成分是 `Population` 、 `Income` 、 `Area` 的值较大，且几乎每个原始变量都是正值：

```
ggbiplot(pca1, obs.scale = 1, var.scale = 1,labels = NULL,varname.size = 13) +
  theme(text = element_text(size = 40))
```



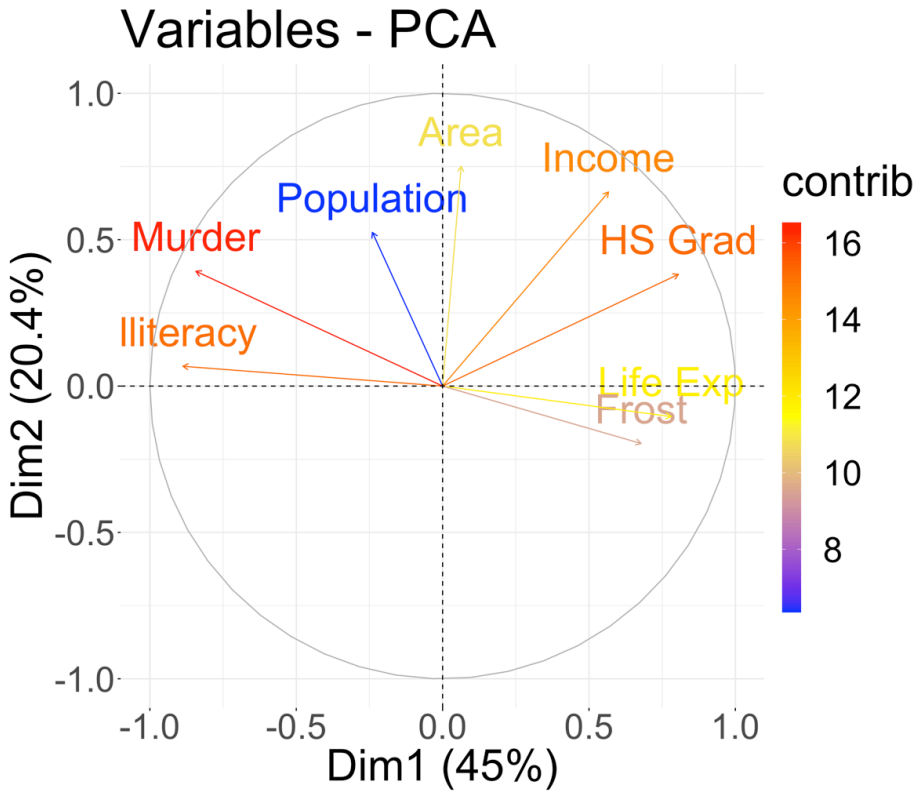
综合来看原始八个变量在第一主成分和第二主成分的贡献占比为下图，前五个占比为 Illiteracy 、 Life Exp 、 Murder 、 HS Grad 、 Income ，这五者的corr只少都有和其中一者高达0.6以上：

```
fviz_contrib(state.x77.pca, choice = "var", axes = 1:2, top = 10) +
  theme(text = element_text(size = 40))
```

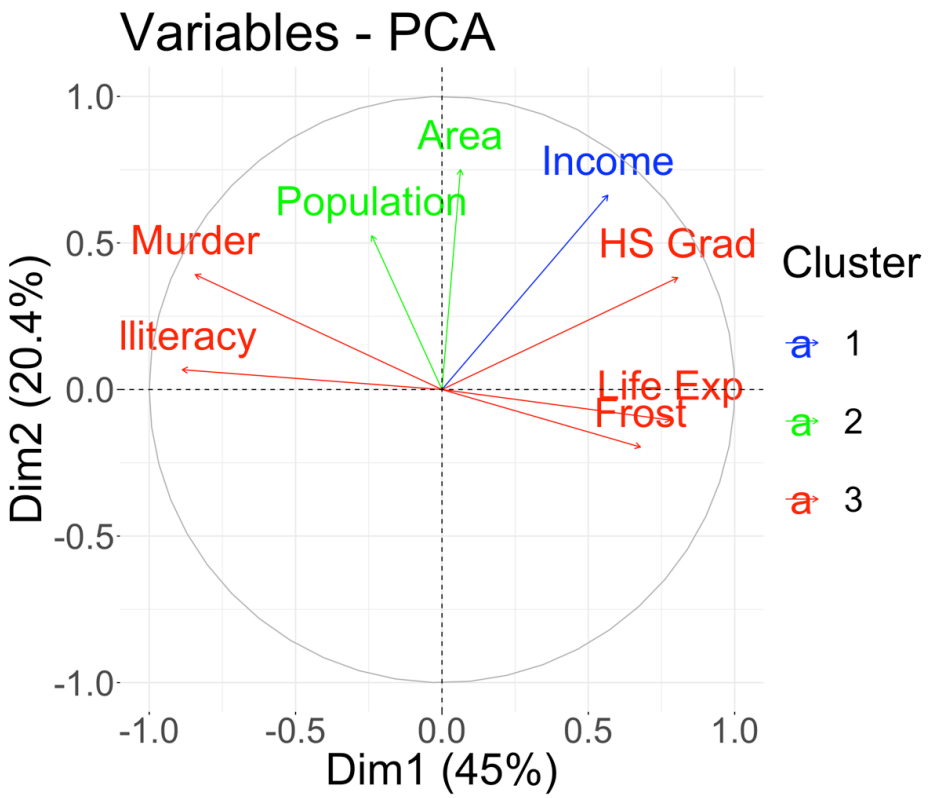


接着来看原始八个变量在第一在第一主成分和第二主成分的贡献占比和k mean分类，特别的是由两个主成分，可以用pca完成降维跟聚类的结果，此处我设定共有三个类别，结果为第一类： Income ；第二类： Population 、 Area ；第三类： Illiteracy``、Life Exp 、 Murder 、 HS Grad 、 Frost ：

```
fviz_pca_var(state.x77.pca, col.var = "contrib",labelsize = 13,
             gradient.cols = c("blue", "yellow", "red"))+
    theme(text = element_text(size = 40),
          legend.key.height = unit(2.5, "cm"))
```

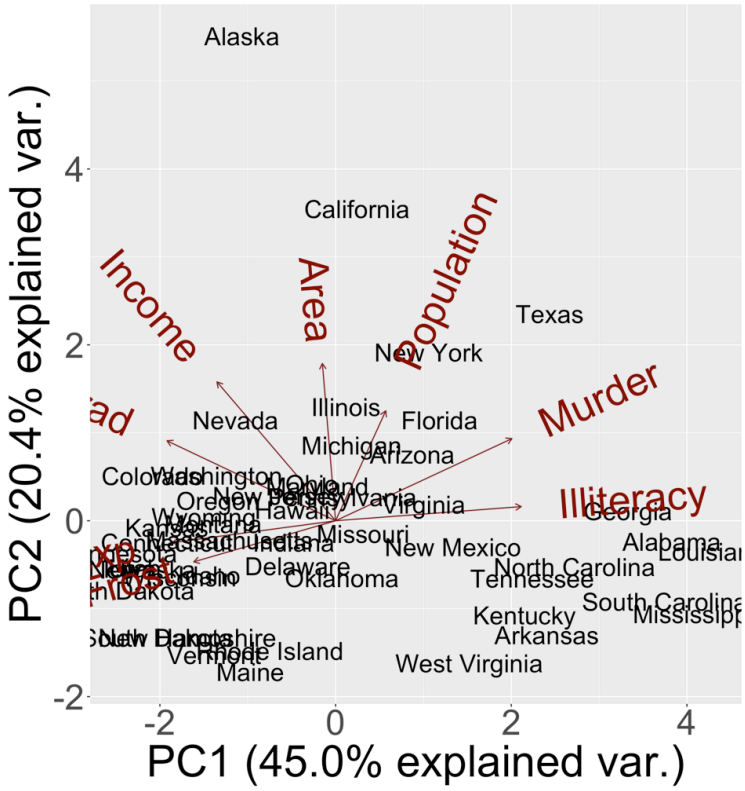



```
set.seed(123)
var <- get_pca_var(state.x77.pca)
var.kms <- kmeans(var$contrib, centers = 3, nstart = 25)
kms.grp <- as.factor(var.kms$cluster)
fviz_pca_var(state.x77.pca, col.var = kms.grp, palette = c("blue", "green", "red")
,legend.title = "Cluster",labelsize = 13)+
    theme(text = element_text(size = 40),
          legend.key.height = unit(2.5, "cm"))
```



最后将国家与前两个主成分的loadings所画出来的原始变量画在一起，可以看出不同国家在两个主成分和原始变量间的分布：

```
ggbiplot(pca1, obs.scale = 1, var.scale = 1,labels = row.names(state.x77)[1:50],va
rname.size = 13, labels.size = 8) +
  theme(text = element_text(size = 40))
```



3. 小結

通常降维的原因是因为，数据在维度高时会相对稀疏，导致特征没这么明显，因此希望可以降维（但希望解释变异不至于损失太多）。降维后可以以主成分取代原始变量成为新的特征，并以此来做后续的聚类、分类，或预测，此次作业仅做了主成分分析并展现各个主成分的insight，若后续继续建模，则可以依照此次内容结合模型。