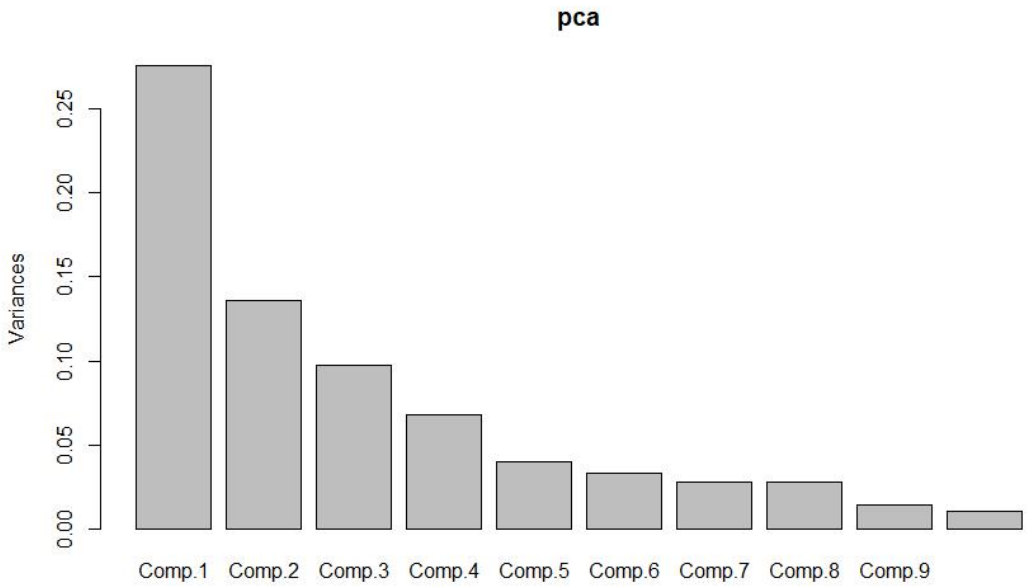
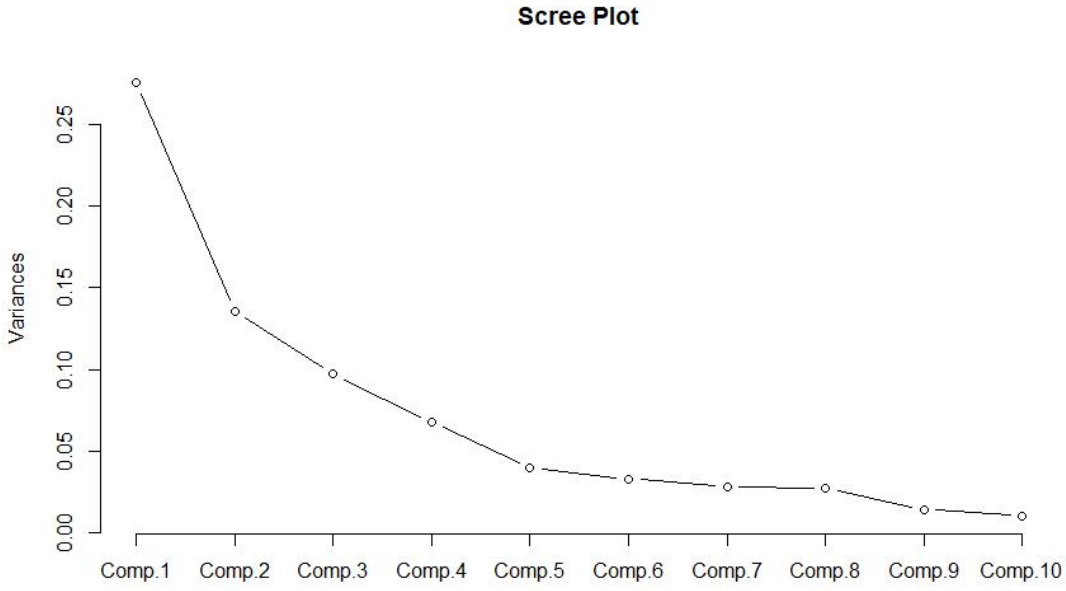
Class assignment 7

Yizhen Chen and Rayming Wang

The result of PCA

We first normalized the VPAL data so that the some columns with high variance will not affect the result.The we have some pictures like this:





In these two pictures above,We found there is an obvious turning point on component 9.So it means the whole VPAL data can reduce in 9 dimensions. Then we load the score of PCA to confirm,We find on component 9 there is a cumulative proportion over 95% percent.Thus we can

say all major components have been calculated.

```
> summary(pca)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9
Standard deviation     0.5249621 0.3683111 0.3116869 0.2608160 0.20064758 0.18241165 0.16759284 0.16622586 0.12048526
Proportion of Variance 0.3666966 0.1805015 0.1292672 0.0905148 0.05356967 0.04427476 0.03737335 0.03676616 0.01931608
Cumulative Proportion  0.3666966 0.5471981 0.6764653 0.7669801 0.82054979 0.86482455 0.90219790 0.93896406 0.95828014
                          Comp.10     Comp.11     Comp.12     Comp.13     Comp.14 Comp.15
Standard deviation     0.10240609 0.086436239 0.082494252 0.076003855 0.028527067       0
Proportion of Variance 0.01395413 0.009941292 0.009055209 0.007686389 0.001082843       0
Cumulative Proportion  0.97223427 0.982175560 0.991230769 0.998917157 1.000000000       1
```

After that we load the PCA analysis and find which column consist of each component.

```
> loadings(pca)

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
Quest.Taken            -0.312 -0.228         0.165 -0.457                  0.256  0.571         -0.284  -0.191   0.318
Quest.Completed        -0.250 -0.128 -0.109  0.231                                0.125 -0.715   0.445   0.328
Dialogues              -0.302 -0.112         0.274 -0.239  0.170         0.177 -0.135  0.341           0.302  -0.676
Loots                  -0.510  0.174  0.238 -0.384  0.122  0.166        -0.146
Loots.Items            -0.489         0.319 -0.321        -0.280  0.162 -0.182                                -0.121
Loots.Dead             -0.103        -0.138 -0.194  0.329  0.647 -0.373  0.172  0.265          0.116   0.101
Shots                         0.347         0.306                0.431                0.308           0.529   0.425
Kills                  -0.131  0.387 -0.179  0.127 -0.158  0.194 -0.178 -0.229 -0.305 -0.357  -0.629
Interaction.NPC        -0.309        -0.135  0.275 -0.163  0.118 -0.170        -0.503  0.212   0.417  -0.398   0.322
Interaction.Container  -0.277  0.113         0.341  0.619 -0.379 -0.134  0.459                 -0.168
Attacks                       0.483               -0.206 -0.119 -0.161  0.153  0.124          0.192          -0.116
Attacks...Quest.Related       0.401  0.244  0.265         0.310  0.374         0.198 -0.151   0.131  -0.491  -0.291
Attacks...Friendly.NPC  0.100 0.422        -0.147 -0.318 -0.233 -0.398  0.272                 0.185   0.172
Attacks...UnMotivated  -0.102  0.115 -0.365  0.170  0.143 -0.270 -0.277 -0.607  0.397  0.251          -0.136  -0.152
Attacks...SelfDefense  -0.158  0.129 -0.738 -0.364                0.409  0.274                         -0.125
                       Comp.14 Comp.15
Quest.Taken
Quest.Completed
Dialogues
Loots                  -0.654
Loots.Items             0.631
Loots.Dead              0.369
Shots                   0.107
Kills                   0.113
Interaction.NPC
Interaction.Container
Attacks                          0.765
Attacks...Quest.Related         -0.258
Attacks...Friendly.NPC          -0.572
Attacks...UnMotivated           -0.108
Attacks...SelfDefense

               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
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000   1.000   1.000   1.000   1.000   1.000
Proportion Var  0.067  0.067  0.067  0.067  0.067  0.067  0.067  0.067  0.067   0.067   0.067   0.067   0.067   0.067
Cumulative Var  0.067  0.133  0.200  0.267  0.333  0.400  0.467  0.533  0.600   0.667   0.733   0.800   0.867   0.933
               Comp.15
SS loadings      1.000
Proportion Var   0.067
Cumulative Var   1.000
>
```

The result of Feature selection

After discussion,we summarized three different kind of person in this game.They are Explorers,quest_takers and bandits.For each personality we defined the target value as follows.

Explorers:People with high interest in dialogue

Quest takers:People with high interest in quest_taken and quest_complete.

Bandits:Players interested in attacking friendly NPC or attack without motivation.

Then we perform forward and backward feature on each kind of people and this is what we find.

The explorers

This is forward selection

```
Call:
lm(formula = dialogues ~ interactionNPC + quest_taken + shots +
    friendly_attack, data = rawdata)

Coefficients:
    (Intercept)   interactionNPC      quest_taken            shots  friendly_attack
        7.01451          0.90882          2.64209          0.04984         -0.04433
```

This is backward selection

```
Call:
lm(formula = dialogues ~ quest_taken + loots + items + dead +
    shots + interactionNPC, data = rawdata)

Coefficients:
   (Intercept)     quest_taken           loots           items            dead           shots  interactionNPC
       5.37870         2.62615        -0.57498         0.56039         0.83821         0.07085         0.93669
```

The quest_takers

This is forward selection

```
Call:
lm(formula = quest_taken + quest_complete ~ dialogues + items +
    loots + dead + interactionNPC, data = rawdata)

Coefficients:
   (Intercept)       dialogues           items           loots            dead  interactionNPC
       5.27063         0.11145         0.28000        -0.21903         0.16646         0.07822
```

This is backward selection

```
Call:
lm(formula = quest_taken + quest_complete ~ dialogues + loots +
    items + dead + interactionNPC, data = rawdata)

Coefficients:
   (Intercept)       dialogues           loots           items            dead  interactionNPC
       5.27063         0.11145        -0.21903         0.28000         0.16646         0.07822
```

The bandits

This is forward selection

```
Call:
lm(formula = friendly_attack + unmotivated_attack ~ attacks +
    quest_attack + selfdefense_attack, data = rawdata)

Coefficients:
       (Intercept)             attacks        quest_attack  selfdefense_attack
         2.038e-14           1.000e+00          -1.000e+00          -1.000e+00
```

This is backward selection

```
Call:
lm(formula = friendly_attack + unmotivated_attack ~ id + quest_taken +
    quest_complete + dialogues + items + dead + shots + kills +
    container + attacks + quest_attack + `self-defense_attack`,
    data = rawdata)

Coefficients:
         (Intercept)                  id         quest_taken      quest_complete            dialogues
          -1.420e-14           1.503e-16           7.145e-15          -1.711e-14          -7.793e-18
               items                dead               shots               kills           container
          -4.910e-16          -1.963e-16           2.575e-16           1.971e-15          -2.523e-16
             attacks        quest_attack  `self-defense_attack`
           1.000e+00          -1.000e+00          -1.000e+00
```

The comparison between these approaches

Firstly,the PCA and feature selection can both tell us the relationship and structure hidden in the data. But they focus on different aspects of data.In PCA it mainly focus on exploration task.The PCA will focus on project original data into reduce dimension space and it will tells which columns have same structure.However feature selection is based on some kind of hypothesis.We need to first find a(or more) feature as target variable.Then it will keep trying until find the best fit linear model.For forward and backward feature selection.On most time they will return same results.But for the bandits part we found the result return by backward and forward are completely different and the forward feature seems more reasonable.

Perhaps it is an incomplete hypothesis because in VPAL data the attack column consists of quest_attack,friendly npc attack,unmotivated attack and self-defense attack.And we just selected a part of these high related data and this make the backward feature over fitting.