UNIVERSITÄT BAYREUTH

Data Mining im Marketing mit R

Team 2a der Kohorte 2:
Tat Cheong Chu, Yinchi Luo

Clustering

“Byrdes of on kynde and color flok and flye allwayes together.”

— William Turner

Die Clusteranalyse bietet einen Einblick in die Daten, indem die Objekte in Gruppen (Cluster) von Objekten unterteilt werden, sodass Objekte in einem Cluster einander ähnlicher sind als Objekte in anderen Clustern (Jain et al. 1988).

Im Allgemeinen gibt es zwei Zwecke für die Verwendung der Clusteranalyse: Verständnis und Nutzen (Tan et al. 2005).

Clustering für das Verständnis ist die Verwendung von Clusteranalyse, um automatisch begrifflich sinnvolle Gruppen von Objekten mit gemeinsamen Eigenschaften zu finden. Es spielt eine wichtige Rolle dabei, die in den Gruppen verborgenen wertvollen Informationen zu analysieren, zu beschreiben und zu nutzen (Wu 2012), was auch das Ziel dieses Projekts ist.

Algorithmus von Clustering

Die frühesten Forschungen zur Clusteranalyse: Moment-Matching-Methode von Karl Pearson (Pearson, 1894).

Die zahlreichen Algorithmen unterscheiden sich vor allem in ihrem Ähnlichkeits- und Gruppenbegriff, ihrem Cluster-Modell, ihrem algorithmischen Vorgehen (und damit ihrer Komplexität) und der Toleranz gegenüber Störungen in den Daten.

In diesem Projekt fokussieren uns wir jedoch auf den k-means Algorithmus, was zu dem Prototype-basierten Algorithmen gehört. Der Algorithmus lernt für jeden Cluster einen Prototyp und bildet Cluster nach Datenobjekten um die Prototypen (MacQueen 1967).

Methoden

“The method of science is logical and rational; the method of the humanities is one of imagination, sympathetic understanding, ‘indwelling'.”

— Andrew Louth

Überblick

In diesem Projekt fokussieren uns wir auf k-means Methode.

Während des „Gehversuchs“ behandeln wir den berühmten Iris Datensatz. Der Prozess ist ziemlich einfach, denn in diesem Datensatz gibt es nur vier Variablen und alle vier Variablen sind im gleichen Maßstab.

Aber in unserem selbst gewählten Datensatz, PUBG Spieler Datensatz, ist es trickreich, denn es gibt große Menge von Variablen, deren Maßstäben unterschiedlich sind und deren Abhängigkeiten unklar sind. Deshalb müssen wir den Datensatz vorbearbeitet haben, bevor wir die k-means Methode auf den Datensatz anwenden, dadurch das weitere Clustering richtig und effizient durchgeführt werden kann. In diesem Fall wird auch die sogenannte „Principal Component Analysis“ eingeführt. Danach setzen wir die k-means Methode ein.

UNIVERSITÄT BAYREUTH

Skalierung und mittlere Normalisierung:

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)}$$

Hier jedes $x_j^{(i)}$ mit wird $x_j - \mu_j$ ersetzt.

Und für verschiedene Variablen auf unterschiedlichen Skalen müssen die Variablen skaliert werden, damit ein vergleichbarer Wertebereich aufgewiesen wird.

$$x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{s_j}$$

PCA wurde von Karl Pearson 1901 eingeführt (Pearson 1901). Es wurde in den 1930er Jahren von Harold Hotelling weiterentwickelt. Diese Analyse wird als Mittel zum Zweck der Dimensionsreduktion betrachtet (Ng 2013).

Reduktion von n-Dimension zum k-Dimension: k Vektoren finden, auf die die Daten projiziert werden, um die Projektionsfehler zu minimieren.

Kovarianzmatrix Σ berechnen:

$$\Sigma = \frac{1}{m} \sum_{i=1}^{n} (x^{(i)})(x^{(i)})^T$$

Eigenvektor von Kovarianzmatrix Σ berechnen:

```
[U, S, V] = svd(Sigma)

Ureduce = U(: ,1 : k)

z = Ureduce′ * X
```

K-means Algorithmus Pseudocode

K-means Algorithmus von MacQueen (1967)

K Clusterzentren initialisieren

Wiederholung {

for i = 1 to m

$c^{(i)}$:= Index (von 1 zu K) des Clusterzentrums nächsten zu $x^{(i)}$

for k = 1 to K

μ_k := durchschnittlicher Wert von Punkten, die dem Cluster zugewiesen sind

}

$c^{(i)}$ = index des Clusterzentrums(1, 2, …, K), das x(i) ist derzeit zugewiesen

R als Programmiersprache

In unserem Projekt werden die Programmiersprache R benutzt, damit die Clustering Analyse durchgeführt werden kann.

R wurde 1992 von Ross Ihaka und Robert Gentleman als ein Schema-ähnlicher Interpreter für den Anwender mit statistischen Aufgaben entwickelt (Ihaka 1998), was jetzt die Standardsprache für statistische Problemstellungen sowohl in der Wirtschaft als auch in der Wissenschaft ist.

R ist seit 1995 ein freies Software-Projekt unter der Bezeichnung GNU-Lizenz, was die am weitesten verbreitete Software Lizenz ist, die einem gewährt, die Software auszuführen, zu studieren, zu ändern und zu verbreiten (FSF 2018).

Die Arbeit (beinhaltet den gesamten Quellcode) in diesem Projekt ist unter einer Creative Commons Attribution 4.0 International Lizenz lizenziert (Creative Commons 2017).

Gehversuche

“Only those who dare to fail greatly can ever achieve greatly.”

— Robert F. Kennedy

Überblick

Variablen	Beschreibung
Sepal.Length	Länge von Kelchblatt
Sepal.Width	Breite von Kelchblatt
Petal.Length	Länge von Blütenblatt
Petal.Width	Breite von Blütenblatt

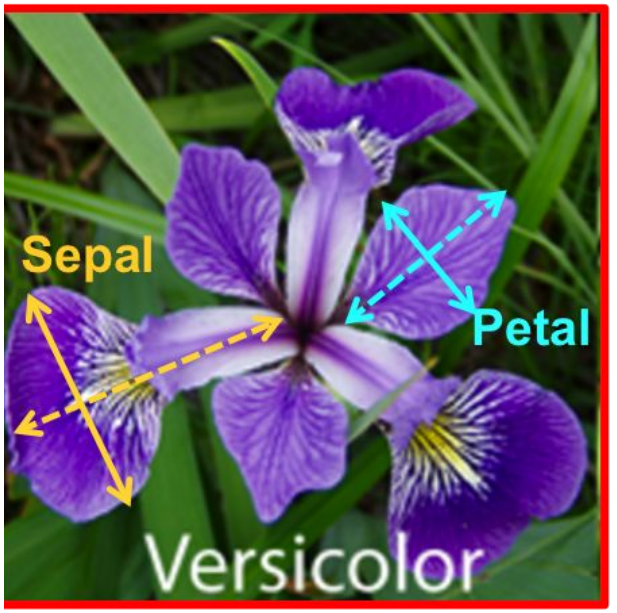
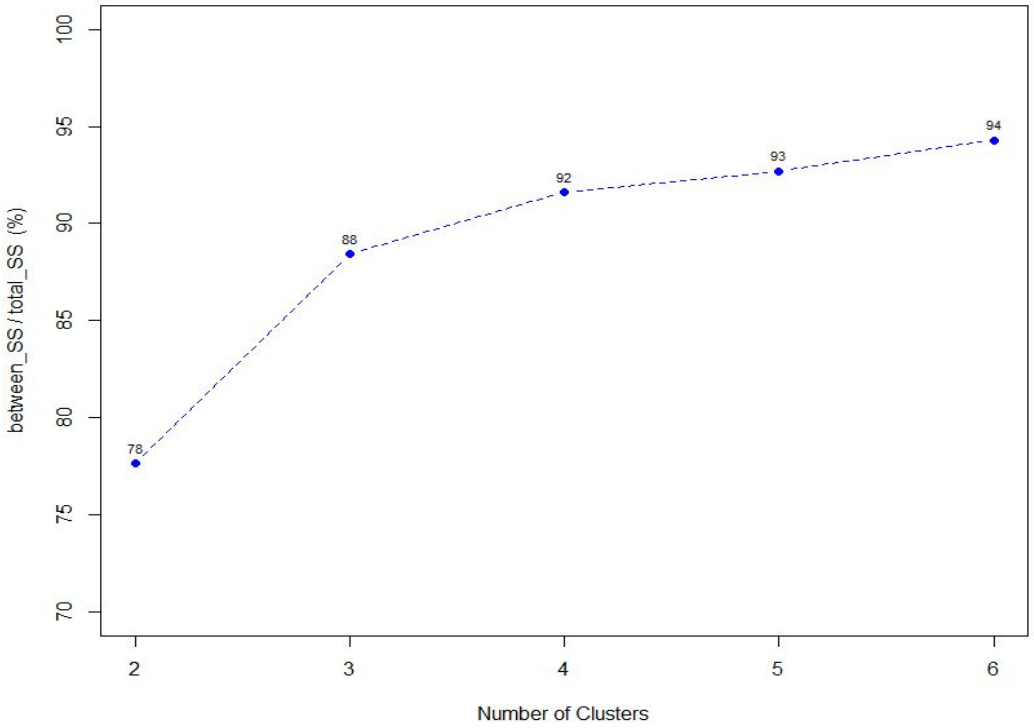


Abb 1: Fialoke, S. (2016). Classification of Iris Varieties.

http://suruchifialoke.com/2016-10-13-machine-learning-tutorial-iris-classification/

Gehversuch
Visualisierung von BCSS/CSS-Veränderung

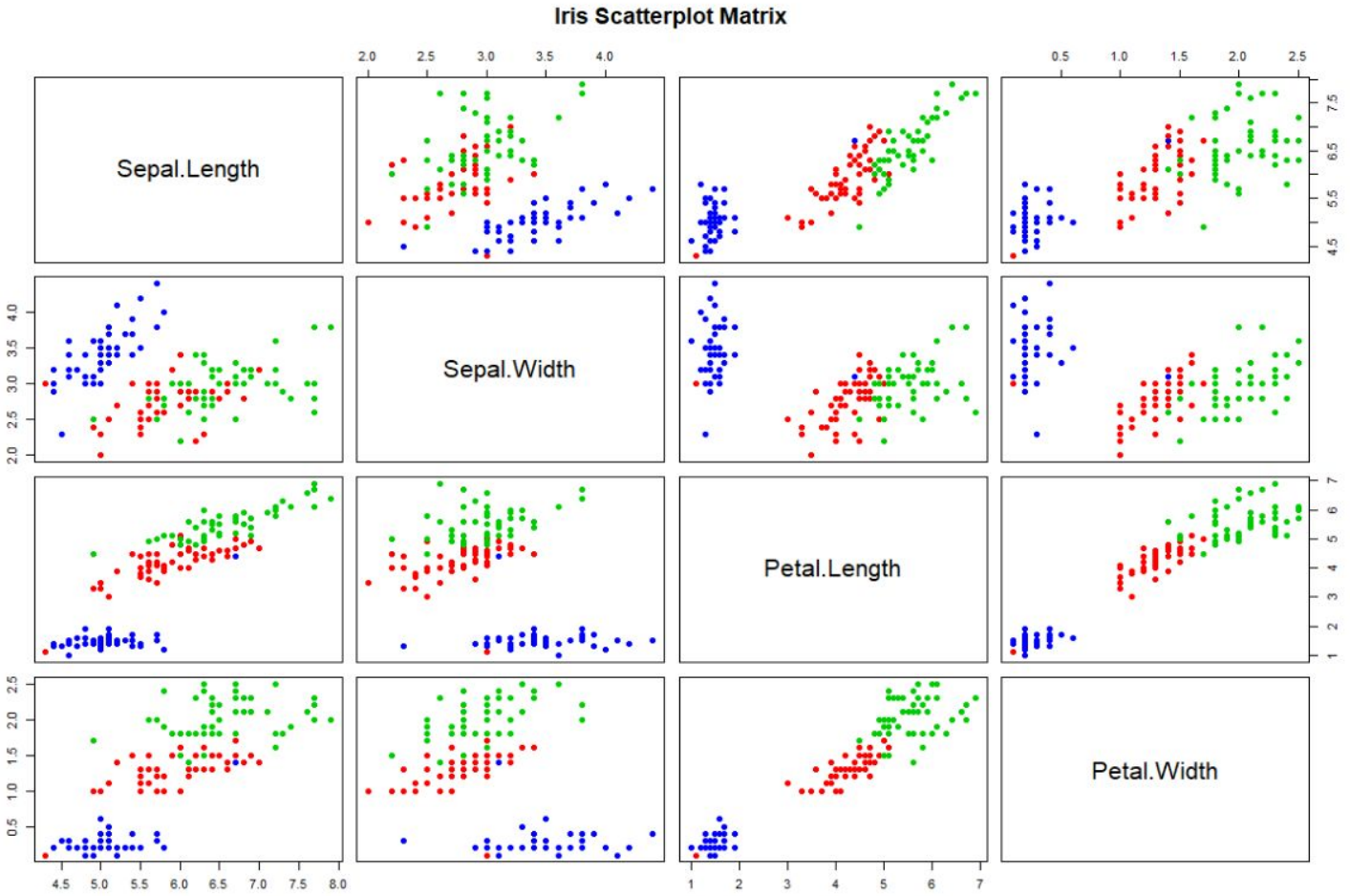


Je mehr Clusterzentren gewählt wird, desto höher wird der Anteil der Varianten von dem Ergebnis der Clusterings erklärt.

Wenn die Anzahl zu hoch ist, führt es zu Problemen wie “Overfitting” und komplexer Interpretation.

Es scheint, mit der erhöhten Anzahl von Clusters ist ab 4 Clusters nicht mehr sinnvoll, weil die Erhöhung von BCSS/CSS danach nicht merkwürdig ist.

Gehversuch
Matrix von Verteilungen



Obwohl die Datensätzen von Iris sehr oft als ein Beispiel für Clustering verwendet werden, aber man kann sehen, dass die grünen Punkte sich optisch nicht separat mit den roten in den meisten Dimensionen unterscheiden, was anders als unsere Vorstellung von dem idealen K-Means Algorithmus mit der klar separaten Gruppierung ist.

Ergebnis des Clusterings

Mittelwerte von Variablen mit K-means Clustering

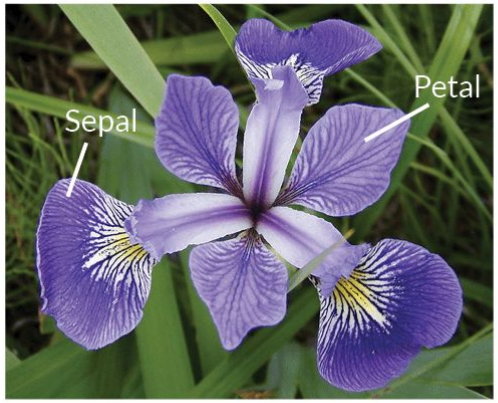
	Group 1	Group 2	Group 3
Sepal.Length	5.901613	6.850000	5.006
Sepal.Width	2.748387	3.073684	3.428
Petal.Length	4.393548	5.742105	1.462
Petal.Width	1.433871	2.071053	0.246

Der Vergleich zu “echter” Gruppierung in dem ursprünglichen Datensatz

	setosa	versicolor	virginica
1	0	48	14
2	0	2	36
3	50	0	0

Iris Setosa

Mit kleinen und kurzen Blütenblättern



Iris Versicolor

Im Vergleich von Setosa: Blüten- und Kelchblätter sehen groß und “dick” aus

Iris Virginica

Mit langen Kelch- und Blütenblättern

Abb. 2 : Santos, R. (2018). Data Science Example - Iris dataset. In Computação e Matemática Aplicada. http://www.lac.inpe.br/~rafael.santos/Docs/R/CAP394/WholeStory-Iris.html

Datensatz

“Data is the new science. Big data holds the answers.”

— Pat Gelsinge

UNIVERSITÄT BAYREUTH

Datenbank aus Kaggle

https://www.kaggle.com/lazyjustin/pubgplayerstats

Spielerstatistiken für ungefähr 85.000 der bestplatzierten PUBG-Spieler

150 numerische Merkmale geteilt durch den Server-Typ (Solo, Duo und Squad) Alle Statistiken sind über alle Regionen hinweg aggregiert.

Abb. 3 : Madam, A. (2018). PlayerUnknown's Battlegrounds dethrones Call of Duty: WWII in weekly U.K. games sales. https://www.windowscentral.com/playerunknowns-battlegrounds-pubg-dethrones-call-duty-wwii-weekly-uk-xbox-one-games-sales

Ziel

Gruppierung der Spieler nach unterschiedlichen Spielverhalten

Iteration und Evaluation

Innerhalb dieser Zeitraum haben wir viel mal versucht.

Die zwei repräsentative Iterationen werden in folgenden Folien gezeigt.

Prozess der Iteration

- Selektion von Variablen
- Extraktion von Variablen mittels PCA
- Clustering mittels Kmeans

Damit Vegleich zwischen Iterationen und Verbesserung der zukünftigen Clusteranalyse ermöglichen. Die folgenden Indikatoren werden bewertet.

Bei PCA:

- Wie groß ist der kumulativer Anteil des erklärten Varianz?
  - ΣAnteil des PCs (zur Clustering)

Bei Clustering:

- Wie viele Prozent der Datenvariation hat der Algorithmus erklärt?
  - BCSS/CSS * 100%

Erste Iteration

“Fortune does favor the bold and you’ll never know what you’re capable of if you don’t try.”

— Sheryl Sandberg

Erste Iteration
Selektion von Variablen

UNIVERSITÄT BAYREUTH

1. nur Solo Spielstatistik
2. nur Spielverhalten relevante Variablen

```
data <- read.csv("PUBG_Player_Statistics.csv")
data1 <- data
```

```
#pick the neccessary variables
solo_data <- data1[c("solo_KillDeathRatio", "solo_WinRatio",
                     "solo_Top10Ratio", "solo_DamagePg",
                     "solo_HealsPg",
                     "solo_KillsPg", "solo_MoveDistancePg",
                     "solo_TimeSurvivedPg",
                     "solo_AvgSurvivalTime", "solo_AvgWalkDistance")]
```

Variablen	Beschreibung
solo_KillDeathRatio	Verhältnis von den erspielten Kills und den erlittenen Toden
solo_WinRatio	Gewinn-Verhältnis
solo_Top10Ratio	Top-10-Verhältnis
solo_DamagePg	verursachter Schaden pro Spiel
solo_HealsPg	wiederhergestellte Gesundheitspunkte pro Spiel
solo_KillsPg	erspielte Kills pro Spiel
solo_MoveDistancePg	Laufleistung vom Spieler pro Spiel
solo_TimeSurvivedPg	durchschnittliche Überlebenszeit pro Spiel
solo_AvgSurvivalTime	durchschnittliche Überlebenszeit pro Spiel
solo_AvgWalkDistance	durchscnittliche Laufleistung pro Spiel

Erste Iteration
PCA Analyse und Transformation

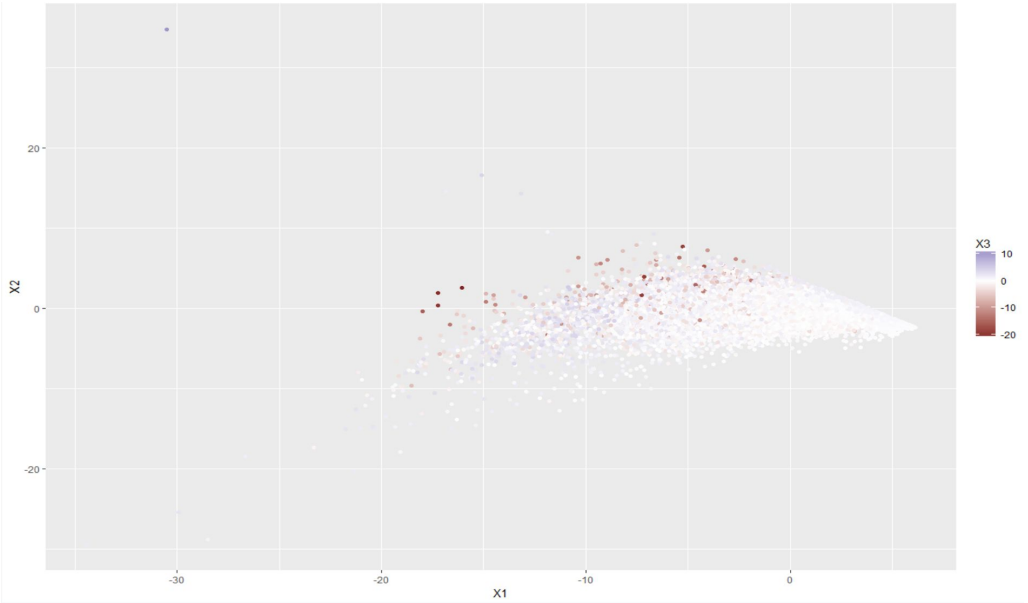
```
> fit1 <- prcomp(solo_data, scale=TRUE)
> summary(fit1)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     2.3981 1.2612 0.91739 0.76303 0.65287 0.59606 0.46675 0.36625 0.29271 0.12441
Proportion of Variance 0.5751 0.1591 0.08416 0.05822 0.04262 0.03553 0.02179 0.01341 0.00857 0.00155
Cumulative Proportion  0.5751 0.7341 0.81831 0.87653 0.91916 0.95468 0.97647 0.98988 0.99845 1.00000
```

PC1, PC2, PC3 zusammen können 81.8% der Variation erklären

PC1 PC2 PC3

	X1	X2	X3
1	-2.15575149	-0.07042678	-0.103461480
2	-3.50512786	-0.92017868	-0.125268782
3	-1.02183406	-1.54045564	0.183207086
4	-10.98342734	-4.42282512	1.816868371
5	-10.15642863	-2.26373320	0.657360088
6	-5.72550406	-0.77987338	0.998553050
7	-1.88821796	0.53782974	-0.327674599
8	-3.24279958	0.77044738	0.778966889
9	-1.43304540	-1.89400746	-0.306278314
10	-4.53355233	0.02638778	1.291346042

Einzelne Datenobjekten und
ihre Koordinaten in PC1/PC2/PC3



Visualisierung von PC1, PC2 und PC3

Erste Iteration
Ergebnis des Clusterings

UNIVERSITÄT BAYREUTH

```
for(i in 2:5){
  fit_cluster <- kmeans(l, i, nstart = 5)
  result[i-1] <- fit_cluster$betweenss/fit_cluster$totss*100
  print(fit_cluster)
}
```

Verwendung des k-means Algorithmus anhand PC1, PC2, PC3 Koordinaten

Das jeweilige Ergebnis:

```
Within cluster sum of squares by cluster:
[1] 240620.7 198637.9
 (between_SS / total_SS =  38.9 %)
Within cluster sum of squares by cluster:
[1] 132221.59 104197.40  85691.35
 (between_SS / total_SS =  55.2 %)
Within cluster sum of squares by cluster:
[1] 61290.04 82472.28 46966.15 77114.95
 (between_SS / total_SS =  62.8 %)
Within cluster sum of squares by cluster:
[1] 53120.88 44579.66 64002.28 46497.66 31555.90
 (between_SS / total_SS =  66.7 %)
```

Zweite Iteration

“If at first you do not succeed, try something harder”

— Proverb

Zweite Iteration
Änderungen gegenüber der ersten Iteration

UNIVERSITÄT BAYREUTH

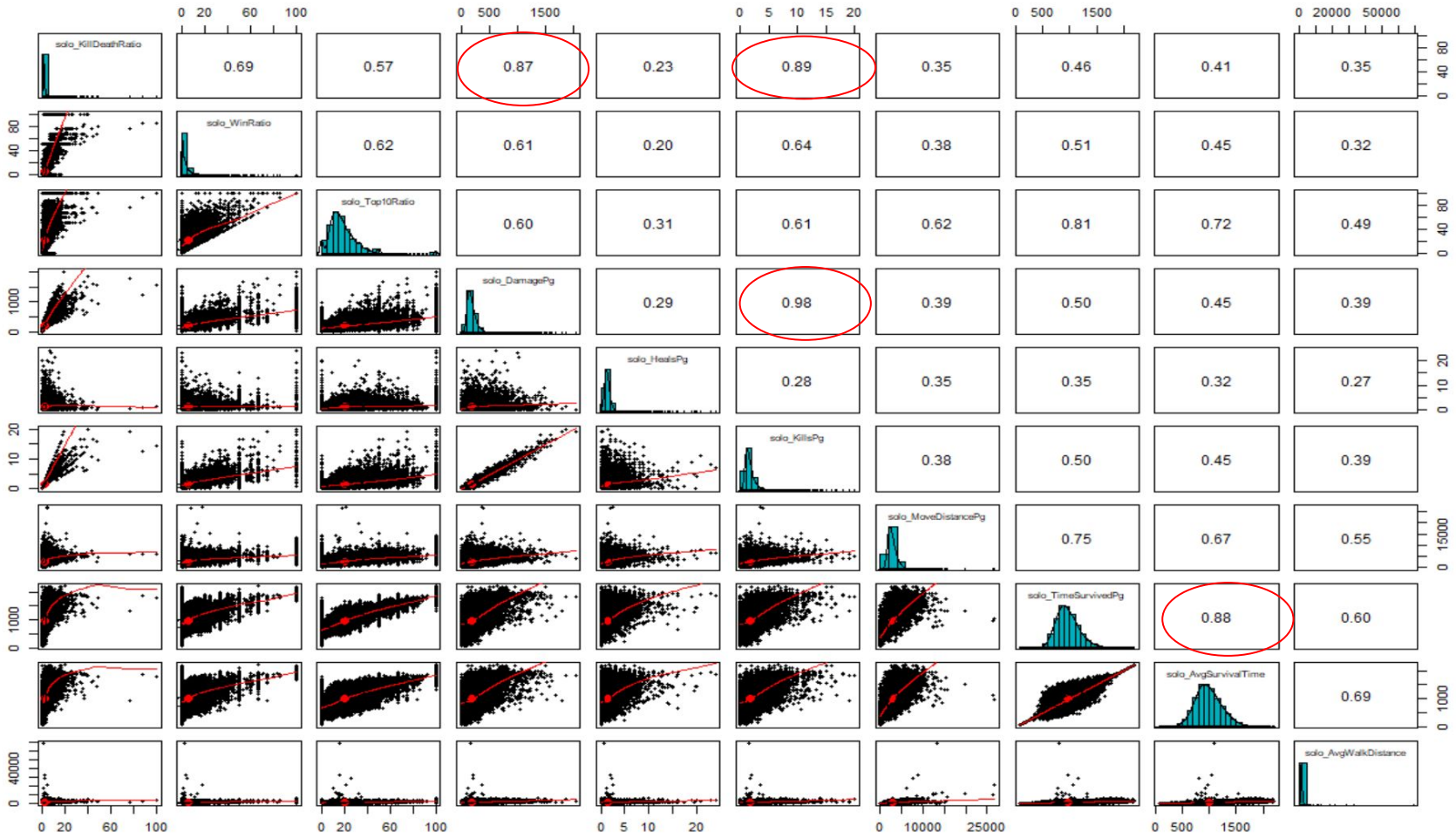
Im Ergebnis von der Clusteranalyse bei der ersten Iteration sehen wir noch große Verbesserungsmöglichkeiten.

Wir denken uns darüber nach, dass wir bei der ersten Iteration keine geeignete Variablen gewählt.

Deshalb gehen wir zu der Selektionsphase and Extraktionsphase zurück. Wir wollen danach das PCA wieder durchführen, um die Leistung der Gruppierung zu verbessern.

Um die Korrelation intuitiv zu zeigen, benutzen wir Streudiagramm-Matrix von Bibliothek “psych”.

Zweite Iteration
Streudiagramm-Matrix von 10 Variablen bei der ersten Iteration



Zweite Iteration
Variablenreduktion

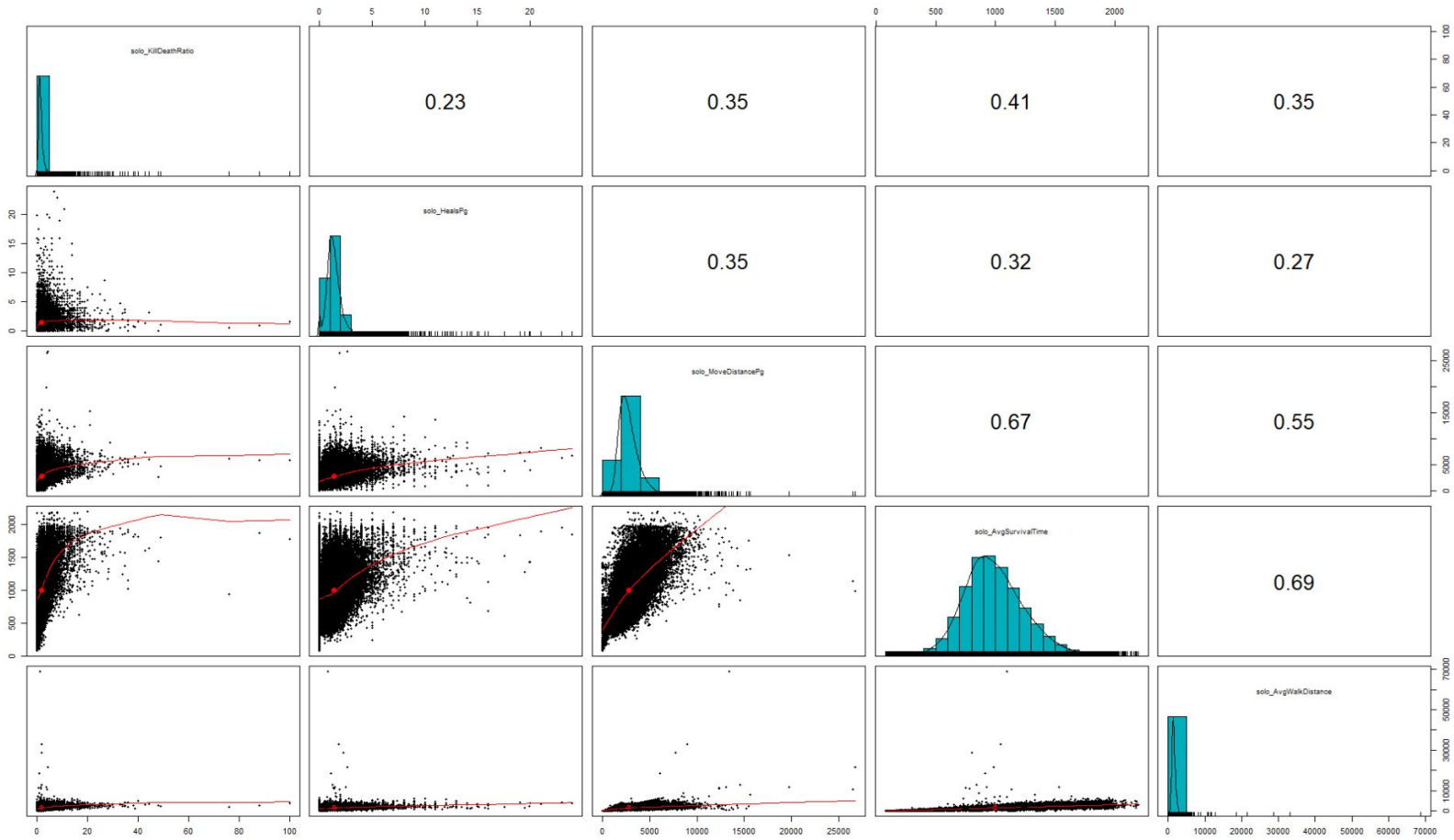
```
solo_data <- data[c("solo_KillDeathRatio", "solo_HealsPg",
                    "solo_MoveDistancePg",
                    "solo_AvgSurvivalTime",
                    "solo_AvgWalkDistance")]
```

Verbesserungen:

- "solo_KillDeathRatio" hat hohe Korrelation mit "solo_DamagePg" und "solo_KillsPg" → nur das erste wählen
- "solo_AvgSurvivalTime" hat hohe Korrelation mit "solo_TimeSurvivedPg" → nur das erste wählen
- "solo_WinRatio" und "solo_Top10Ratio" stellen dar, wie häufig der Spieler gewonnen hat, was eigentlich keine enge Beziehung zu dem Spielverhalten ist. Deshalb werden die beiden Variablen nicht mehr betrachtet.

Variablen	Beschreibung
solo_KillDeathRatio	Verhältnis von den erspielten Kills und den erlittenen Toden
solo_HealsPg	wiederhergestellte Gesundheitspunkte pro Spiel
solo_MoveDistancePg	Laufleistung vom Spieler pro Spiel
solo_AvgSurvivalTime	durchschnittliche Überlebenszeit pro Spiel
solo_AvgWalkDistance	durchschnittliche Laufleistung pro Spiel

Zweite Iteration
Streudiagramm-Matrix von 5 Variablen



5 Variablen ohne hohe Korrelation

Zweite Iteration
PCA Analyse

UNIVERSITÄT BAYREUTH

```
> fit1 <- prcomp(solo_data, scale=TRUE)
> summary(fit1)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5
Standard deviation     1.6542 0.9009 0.8569 0.6656 0.52396
Proportion of Variance 0.5473 0.1623 0.1469 0.0886 0.05491
Cumulative Proportion  0.5473 0.7096 0.8565 0.9451 1.00000
```

0.81831 -> 0.8565

größere kumulative Anteil bei der zweiten Iteration als bei der ersten

```
fit_cluster <- kmeans(l, 5, nstart = 5)
fit_cluster
```

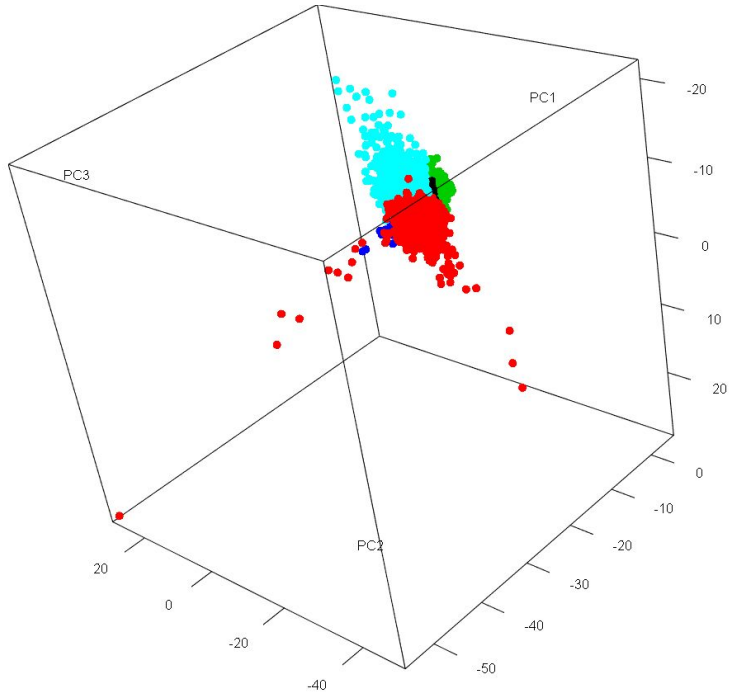
Das Ergebnis von 5 Gruppen und deren Note.

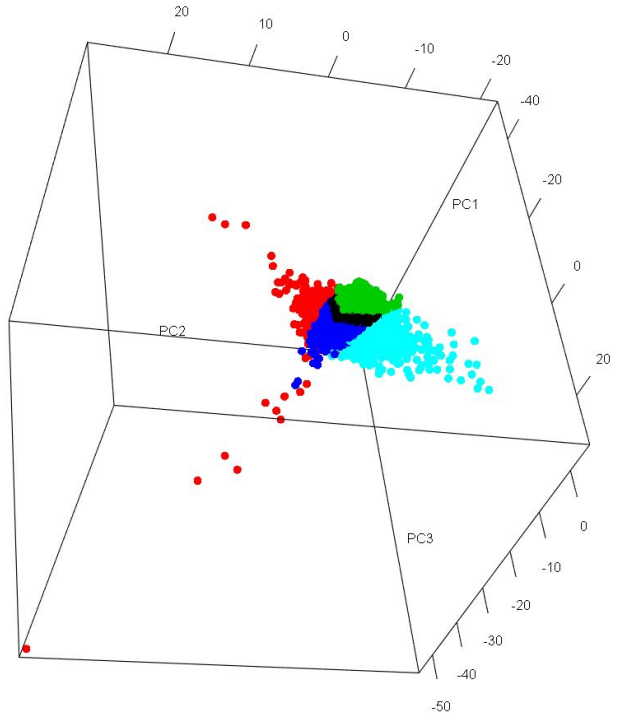
```
Within cluster sum of squares by cluster:
[1] 20541.17 26700.39 32365.54 35415.94 31057.23
 (between_SS / total_SS =  61.2 %)
```

Zweite Iteration
3D Visualisierung der K-means Clustering in 5 Gruppen

```
install.packages("rgl")
library(rgl)
l$groups <- fit_cluster$cluster
attach(l)
plot3d(X1, X2, X3, size = 10, col=l$groups)
```

Weil wir drei Variablen haben, verwenden wir die packages “rgl”, um die Verteilungen in drei Dimensionen zu visualisieren.





Dritte Iteration

“The third time's the charm.”

— Proverb

Dritte Iteration
Änderungen gegenüber der zweiten Iteration

UNIVERSITÄT BAYREUTH

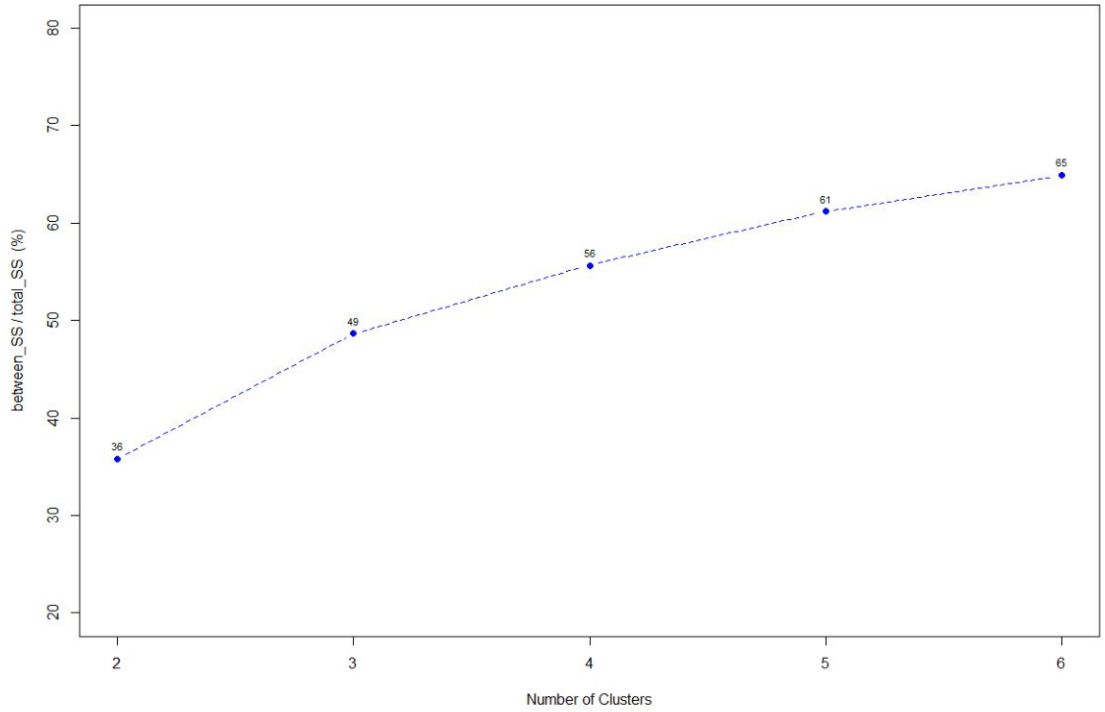
Im Ergebnis von der Clusteranalyse bei der zweiten Iteration sehen wir noch große Verbesserungsmöglichkeiten, besonderes in der Gruppierung Phase mit k-means.

Wir denken uns darüber nach, dass wir bei der zweiten Iteration keine geeignete Nummer von Clusterzentren gewählt. Unsere Entscheidung bei der zweiten Iteration darüber, mit wie vielen Zentren zu arbeiten ist, ist total willkürlich.

In der dritten Iteration wollen wir diese Prozess verändern, indem wir unsere Entscheidung demonstrieren.

Um das zu tun, haben wir die Beziehung zwischen Nummer der Clusterzentren und dazugehörigen BCCS/CSS überprüft, was illustriert, wie viele Prozent der Daten Variation hat jeweilige Ergebnis des Clusterings mit unterschiedlichen Mengen der Clusterzentren erklärt.

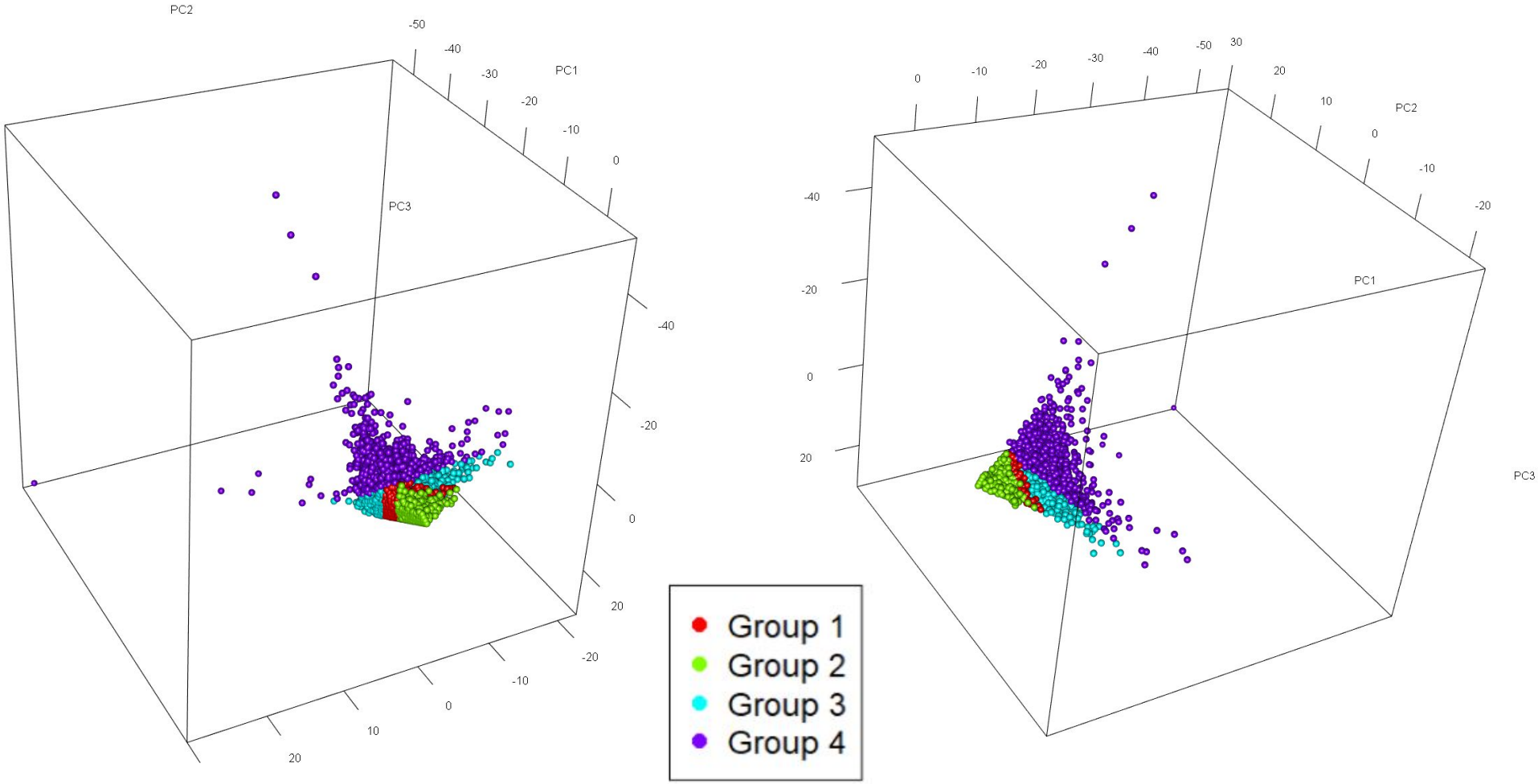
Dritte Iteration
Menge der Clusterzentren und dazugehörige BCSS/CSS



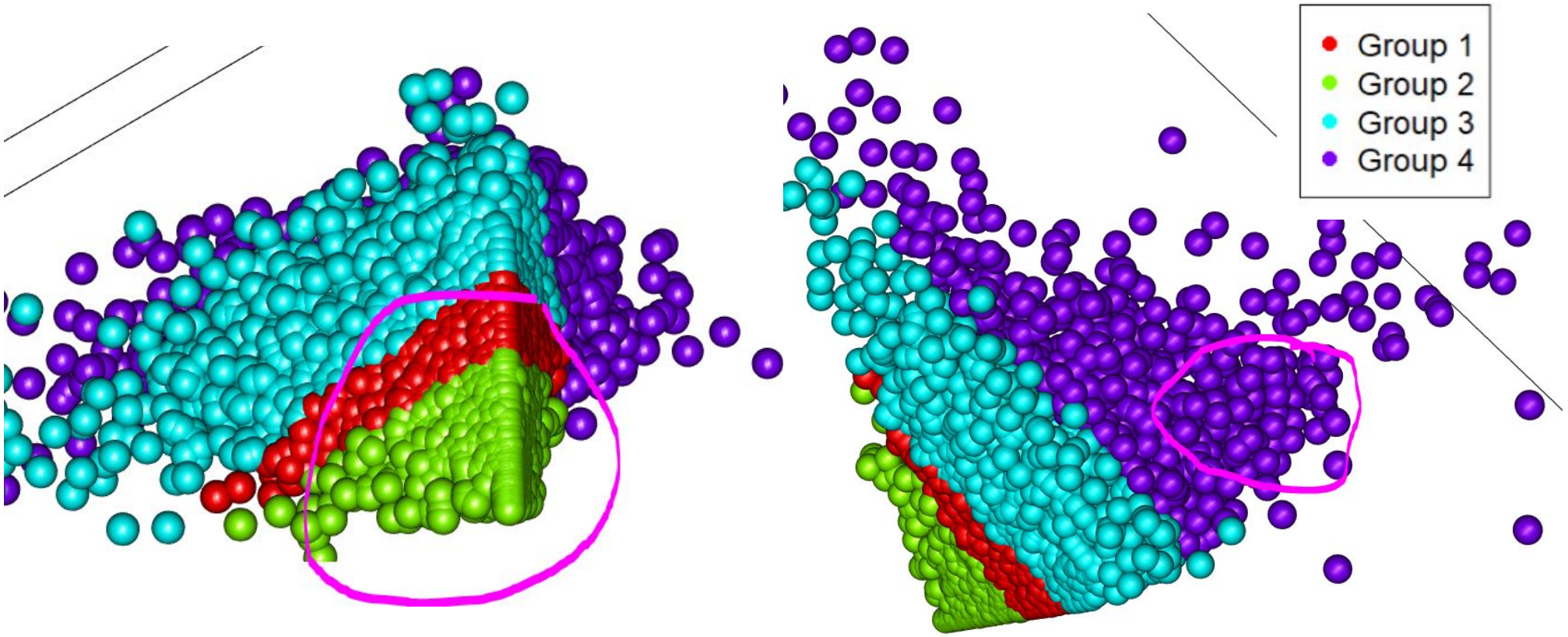
Durch das Diagramm entscheiden wir uns, Clustering mit vier Clusterzentren durchzuführen. Im Vergleich mit Clustering mit drei(36%) und zwei(49%) hat Clustering mit vier Zentren (56%)offensichtlich mehr Varianten erklärt.

Natürlich je mehr Clusterzentren gewählt wird, desto höhere Anteil der Varianten wird von dem Ergebnis der Clusterings erklärt. Aber dann wird das Ergebnis schwer interpretierbar, denn die Unterschiede zwischen Gruppen wird auch mit der steigenden Anzahl der Cluster reduziert.

Dritte Iteration
3D Visualisierung der K-means Clustering in 4 Gruppen



Dritte Iteration
3D Visualisierung der K-means Clustering in 4 Gruppen



Die 3D Visualisierung in der letzten Folie zeigt, als ob die Gruppe 4 (Lila) die meisten Datenpunkte hätte. Aber durch der oberen Visualisierung mit Teilvergrößerung erkennen wir, dass die Dichte des Datenpunkte in Gruppe 1 (rot) und Gruppe 2 (grün) viel höher als in Gruppe 4.

Auswertung

“The best evaluation I can make of a player is to look in his eyes and see how scared they are.”

— Michael Jordan

Überblick und Ergebnis des Clusterings

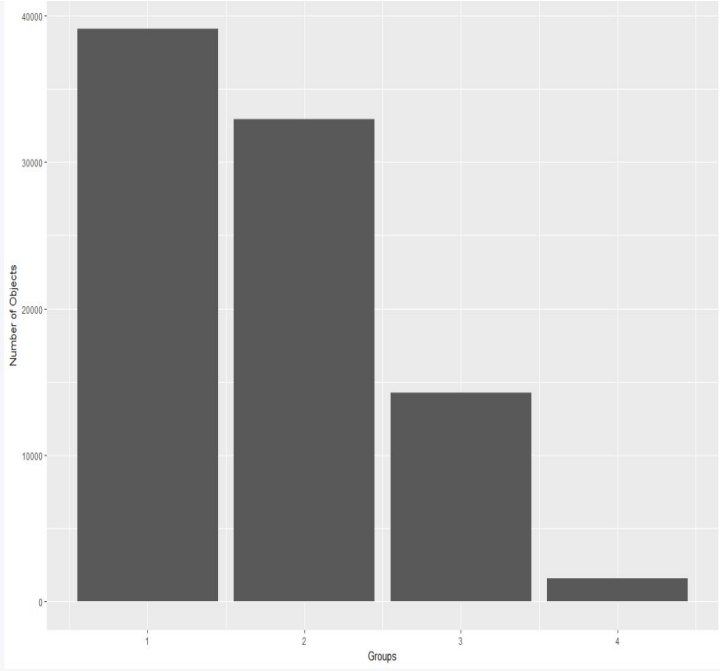
	Group 1	Group 2	Group 3	Group 4
solo_KillDeathRatio	1.703623	1.288655	2.677356	10.446442
solo_HealsPg	1.411028	1.058455	2.096640	2.807322
solo_MoveDistancePg	2934.879238	1936.837242	4328.525021	5252.333502
solo_AvgSurvivalTime	1045.792733	757.822791	1340.881038	1626.971286
solo_AvgWalkDistance	1593.837343	1051.205132	2207.616040	2931.868970

	Groups	Number of Objects
1	1	39106
2	2	32906
3	3	14284
4	4	1602

Mittelwerte von Variablen mit K-means Clustering

Die Spieler in Gruppe 3 und 4 haben offensichtlich bessere Performance als die Spieler in Gruppe 1 und Gruppe 2, weil sie längere Überlebenszeit haben, was direkt widerspiegeln kann, wie gut ein Spieler in solche “Battle Royale” Type Spiele spielt.

Daher können wir behaupten, dass die Spieler in Gruppe 3 und Gruppe 4 fortgeschrittene Spieler sind und die Spieler in Gruppe 1 und Gruppe 2 nicht so gute Leistung haben.



Vergleich zwischen Gruppen

	Group 1	Group 2
solo_KillDeathRatio	1.703709	1.288650
solo_HealsPg	1.411026	1.058559
solo_MoveDistancePg	2935.177092	1936.945773
solo_AvgSurvivalTime	1045.837596	757.872510
solo_AvgWalkDistance	1593.945271	1051.286354

Die Spieler in Gruppe 2 sind **Anfänger,** während die Spieler in Group 1 **Mittlere Spieler** sind**,** denn die Spieler in Gruppe 2 schneiden in allen Aspekten besser ab.

	Group 3	Group 4
solo_KillDeathRatio	2.677356	10.446442
solo_HealsPg	2.096640	2.807322
solo_MoveDistancePg	4328.525021	5252.333502
solo_AvgSurvivalTime	1340.881038	1626.971286
solo_AvgWalkDistance	2207.616040	2931.868970

Die Spieler in Gruppe 3 und Gruppe 4 sind beide **fortgeschrittene Spieler** (wegen der langen durchschnittlichen Überlebenszeit), aber mit ganz unterschiedlichen Spielstile (oder Strategien).

Der größte Unterschied liegt in dem Verhältnis von den erspielten Kills und den erlittenen Toden.

Spieler in Gruppe 4 haben eine sehr gute Treffsicherheit und deswegen haben sie sehr aggressive Spielstil. Ihre Strategie ist: Sie finden alle Feinde heraus und erschießen sie.

Spieler in Gruppe 3 sind nicht so streitlustig, sie versuchen einen Kampf zu vermeiden, der unnötig ist.

Zusammenfassung

“No research is ever quite complete. It is the glory of a good bit of work that it opens the way for something still better, and this repeatedly leads to its own eclipse.”

— Mervin Gordon

Einschränkung und Empfehlungen

UNIVERSITÄT BAYREUTH

Schlussfolgerungen

Umgang mit großartigem Datensatz (87898 Objekten mit 152 Variablen)
Variablenreduktion, von 152 Variablen zu 10 Variablen und 5 Variablen
Anwendung der PCA-Analyse und K-means
„Schöne“ Ergebnis des Clusterings: vier Gruppen von Spieler mit offensichtlichen Unterschieden

Methoden

Wegen der beschränkten Zeit haben wir die anderen Clustering Methoden nicht benutzt, um die Ergebnisse zu vergleichen, was potentiell die tatsächliche Leistung von Gruppierung verbessern kann.

Datensätzen

Obwohl wir denken, dass diese Statistiken unterschiedliche Spielweisen von Spielern darstellen können, basiert die Datensätzen auf die Spielerstatistiken für ungefähr 85.000 der bestplatzierten PUBG-Spieler, während die Nummer der Spieler nach PUBG Stats 875,234 erreicht. Und der Zeitraum von den Datensätzen sind unklar.

Die zukünftige Forschung kann …

Die verschieden Clustering Methoden benutzen und die Ergebnisse unterscheiden, um die Leistung der Gruppierung zu verbessern.
Die Datensätzen selbst sammeln und generieren, um die Ergebnis überzeugender machen.
Die Duo- und Squad Spielstatistiken berücksichtigen, denn die Zusammenspiel kann potenziell eine ganz andere Dimension des Spielstils widerspiegeln.

Literaturen I (für Theorie)

Ihaka, Ross (1998). R : Past and Future History (PDF) (Technical report). Statistics Department, The University of Auckland, Auckland, New Zealand.

Jain, A., Dubes, R.(1988). Algorithms for Clustering Data.Prentice Hall, Englewood Cliffs.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. pp. 281–297.

Pearson,K.. (1894). Contributions to the mathematical theory of evolution. Philos.Trans.Royal Soc. Lond. P.185, 71–110.

Pearson, K. (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space" . Philosophical Magazine. 2 (11): 559–572.

Tan, P.N., Steinbach, M., Kumar, V. (2005). Introduction to Data Mining. Pearson.

Wu, J. (2012). Advances in K-means Clustering: A Data Mining Thinking. Springer.

Literaturen II (für R-Anwendung)

Adler, D., Murdoch. D. (2018). Package ‘rgl’. In CRAN.

https://cran.r-project.org/web/packages/rgl/rgl.pdf

Amoeba. (2015). What is the relation between k-means clustering and PCA?. In Cross Validated.

https://stats.stackexchange.com/questions/183236/what-is-the-relation-between-k-means-clustering-and-pca

Bottoms, C. (2015). Converting rows into columns and columns into rows using R. In Stockoverflow.

https://stackoverflow.com/questions/28680994/converting-rows-into-columns-and-columns-into-rows-using-r

Cdeterman.(2014). r - How to interpret PCA for data reduction?. In Cross Validated.

https://stats.stackexchange.com/questions/118439/how-to-interpret-pca-for-data-reduction

Fanara, C. A Tutorial on Loops in R - Usage and Alternatives. In DataCamp.

https://www.datacamp.com/community/tutorials/tutorial-on-loops-in-r

Fialoke, S. (2016). Classification of Iris Varieties. http://suruchifialoke.com/

http://suruchifialoke.com/2016-10-13-machine-learning-tutorial-iris-classification/

G. Grothendieck (2011). k-means return value in R. In Stackoverflow.

https://stackoverflow.com/questions/8637460/k-means-return-value-in-r

Harrison, M. (2014). PCA and K-means Clustering of Delta Aircraft. In R-bloggers.

https://www.r-bloggers.com/pca-and-k-means-clustering-of-delta-aircraft/

Kabacoff, R.(2017). Cluster Analysis. In Quick-R.

https://www.statmethods.net/advstats/cluster.html

Kabacoff, R.(2017). Scatterplots. In Quick-R.

https://www.statmethods.net/graphs/scatterplot.html

Kassambara, A. Scatter Plot Matrices - R Base Graphs - Easy Guides - Wiki. In STHDA.

http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs

Literaturen III (für R-Anwendung)

koekenbakker. Adding a legend to an rgl 3d plot. In Stackoverflow.
https://stackoverflow.com/questions/27958226/adding-a-legend-to-an-rgl-3d-plot

lukeA (2016). How to get the KMeans Between/Within accuracy percentage in R?. In Stackoverflow.
https://stackoverflow.com/questions/40696614/how-to-get-the-kmeans-between-within-accuracy-percentage-in-r

Ng, A. (2013). Machine Learning. In Coursera.
https://www.coursera.org/learn/machine-learning

Plotting PCA. Principal Component Analysis. In CRAN.
https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html

Revelle, W. pairs.panels function. In R Documentation.
https://www.rdocumentation.org/packages/psych/versions/1.8.4/topics/pairs.panels

Rice, D. (2016). How to Summarize a Data Frame by Groups in R. In r-bloggers.
https://www.r-bloggers.com/how-to-summarize-a-data-frame-by-groups-in-r/

Santos, R. (2018). Data Science Example - Iris dataset. In Computação e Matemática Aplicada.
http://www.lac.inpe.br/~rafael.santos/Docs/R/CAP394/WholeStory-Iris.html

Takahashi, K., Wilke, C., Woo, K.. Create your own discrete scale. In scale_manual • ggplot2.
https://ggplot2.tidyverse.org/reference/scale_manual.html

Wickham, H.. group_by function. In R Documentation.
https://www.rdocumentation.org/packages/dplyr/versions/0.7.6/topics/group_by