Accessing Databases from R for the Purpose of Data Mining

Ing. Jan Manďák
VŠB-Technical University of Ostrava
Faculty of Economics
Department of Systems Engineering

Agenda

- Introduction
- R programming language
- Data Mining
- Clustering
- Conclusion

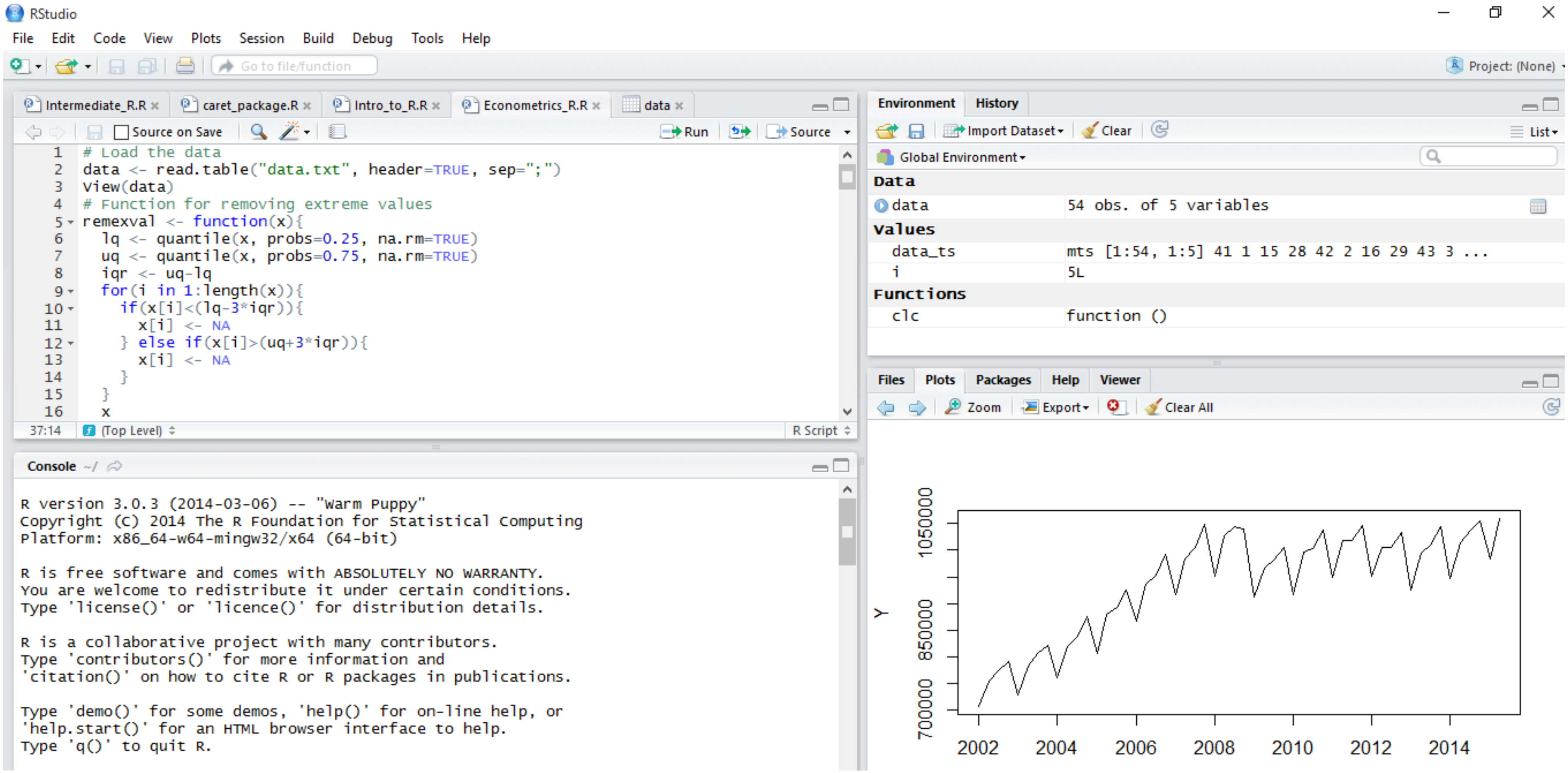
Introduction

- companies are storing huge amount of data in databases
- data are used for creating reports in reporting tools – e.g. Business Objects, Cognos
- using R it is possible to use this data for advanced analytics

R programming language

- SQL has limited statistical features
- can be used for preparing data for analyses
- R - **free** software environment for statistical computing and graphics
- wide variety of statistical, machine learning and graphical techniques
- recommended IDE - R Studio

R Studio



```r
# Load the data
data <- read.table("data.txt", header=TRUE, sep=";")
View(data)
# Function for removing extreme values
remexval <- function(x){
  lq <- quantile(x, probs=0.25, na.rm=TRUE)
  uq <- quantile(x, probs=0.75, na.rm=TRUE)
  iqr <- uq-lq
  for(i in 1:length(x)){
    if(x[i]<(lq-3*iqr)){
      x[i] <- NA
    } else if(x[i]>(uq+3*iqr)){
      x[i] <- NA
    }
  }
  x
```

Data

data	54 obs. of 5 variables

Values

data_ts	mts [1:54, 1:5] 41 1 15 28 42 2 16 29 43 3 ...
i	5L

Functions

clc	function ()

```
R version 3.0.3 (2014-03-06) -- "Warm Puppy"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

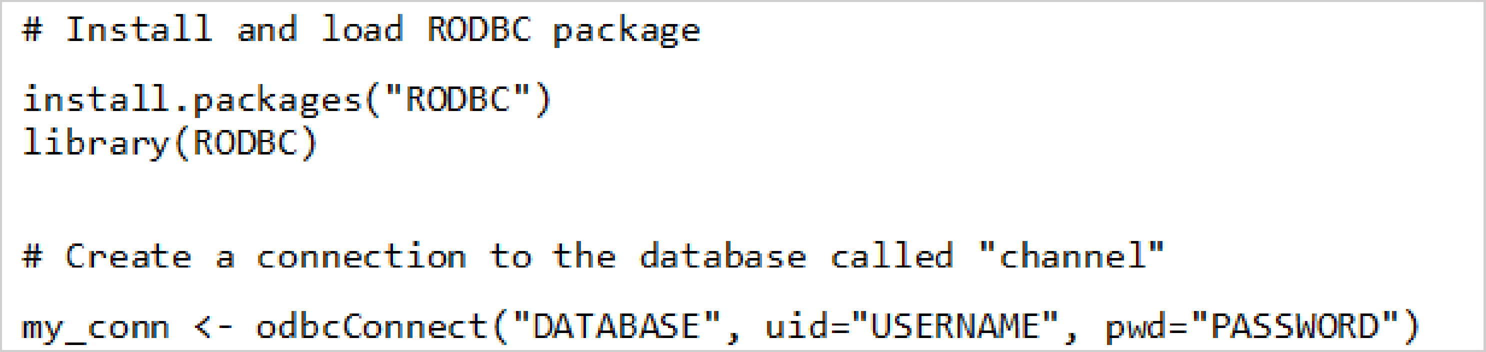
Data Mining

- computational process of discovering patterns in usually large data sets
- methods at the intersection of artificial intelligence, machine learning, statistics and database systems
- results of these analyses help improve efficiency of business proccesses, increase sales etc.

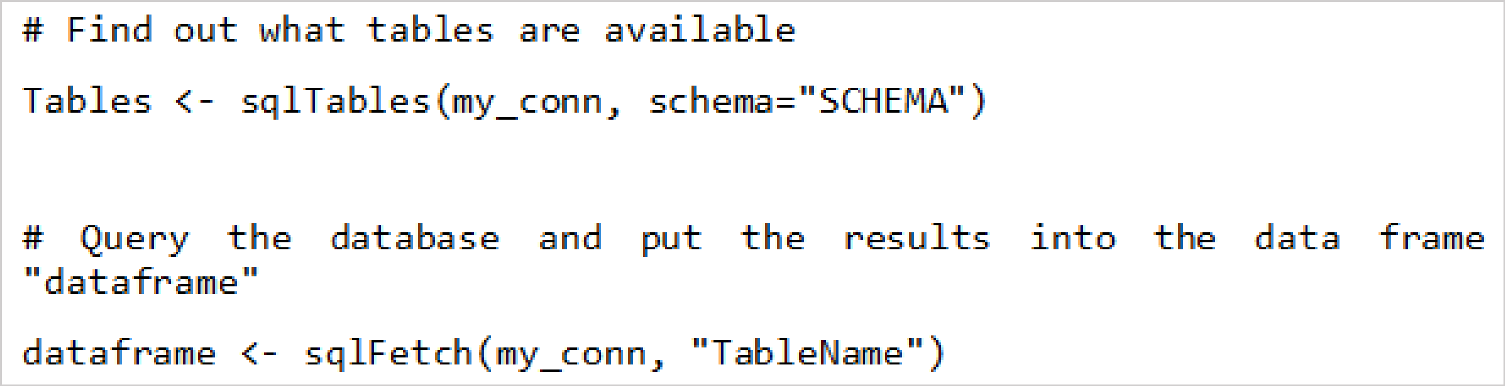
Data Mining Use Cases

- clustering - segmentation of customers
- classification - churn prediction
- regression - prediction of demand
- association rules - market basket analysis
- time series prediction - forecasting of key performance indicators
- text mining - sentiment analysis of social networks
- anomaly detection - fraud detection

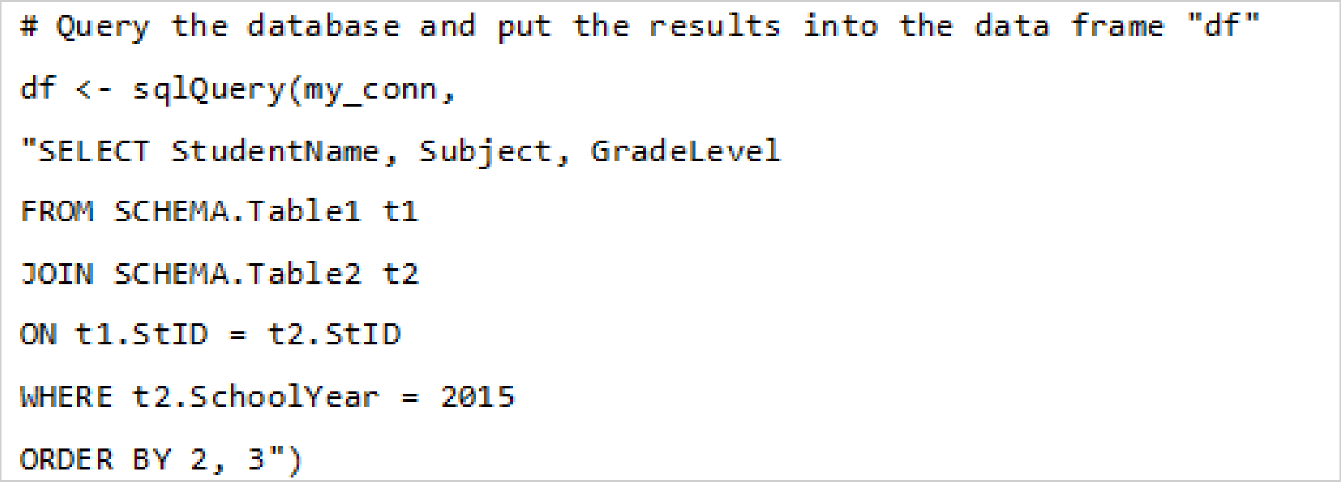
How to get data from DB to R?



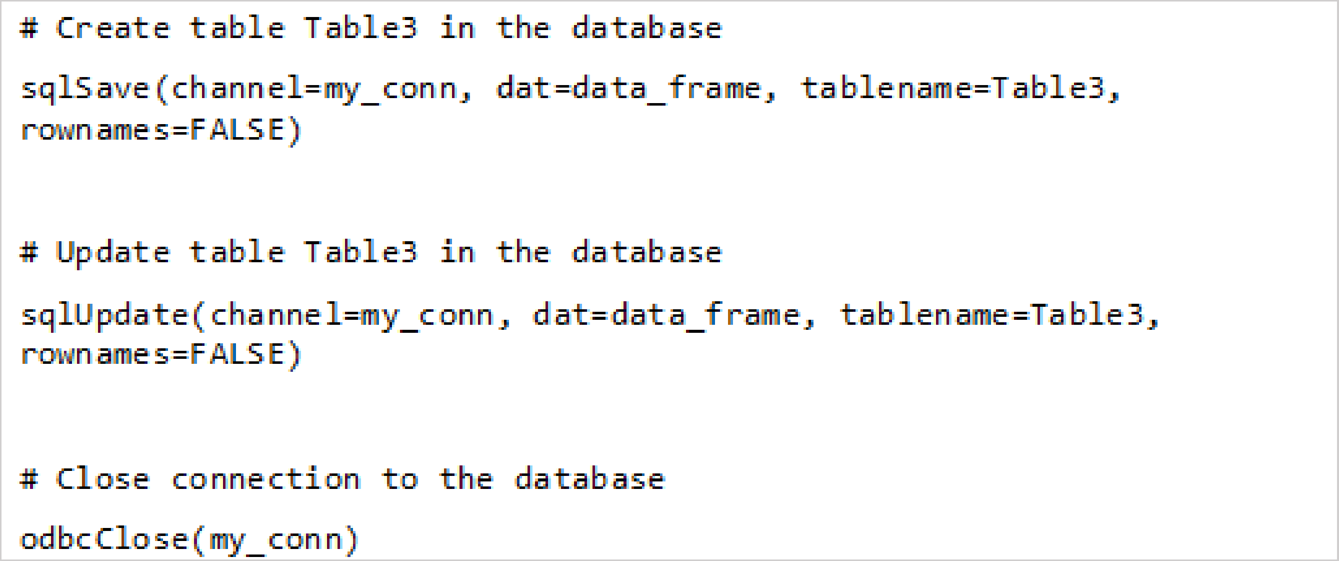
```
# Install and load RODBC package

install.packages("RODBC")
library(RODBC)


# Create a connection to the database called "channel"

my_conn <- odbcConnect("DATABASE", uid="USERNAME", pwd="PASSWORD")
```



```
# Find out what tables are available

Tables <- sqlTables(my_conn, schema="SCHEMA")


# Query the database and put the results into the data frame
"dataframe"

dataframe <- sqlFetch(my_conn, "TableName")
```



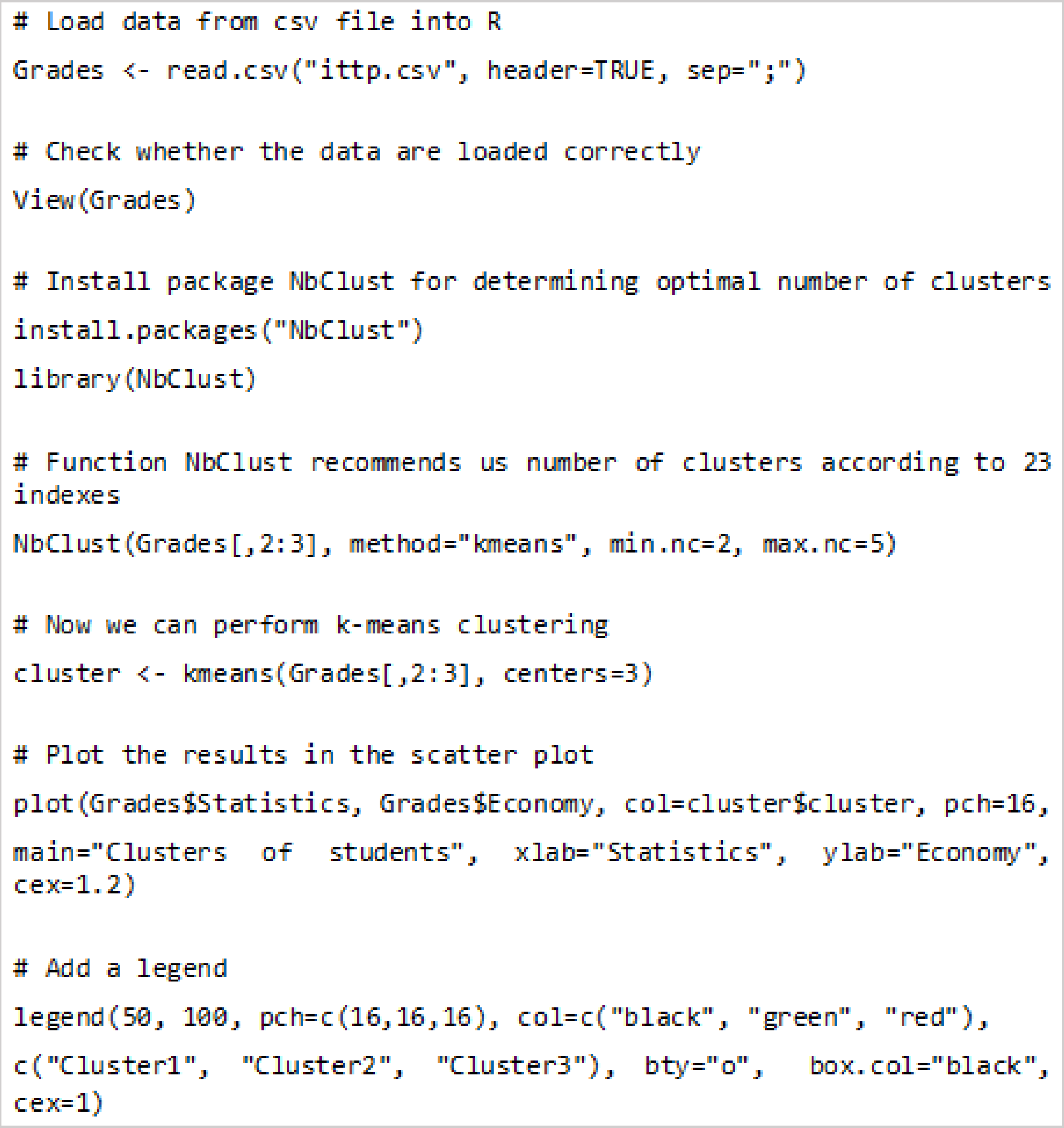
```
# Query the database and put the results into the data frame "df"

df <- sqlQuery(my_conn,

"SELECT StudentName, Subject, GradeLevel

FROM SCHEMA.Table1 t1

JOIN SCHEMA.Table2 t2

ON t1.StID = t2.StID

WHERE t2.SchoolYear = 2015

ORDER BY 2, 3")
```



```
# Create table Table3 in the database

sqlSave(channel=my_conn, dat=data_frame, tablename=Table3,
rownames=FALSE)


# Update table Table3 in the database

sqlUpdate(channel=my_conn, dat=data_frame, tablename=Table3,
rownames=FALSE)


# Close connection to the database

odbcClose(my_conn)
```

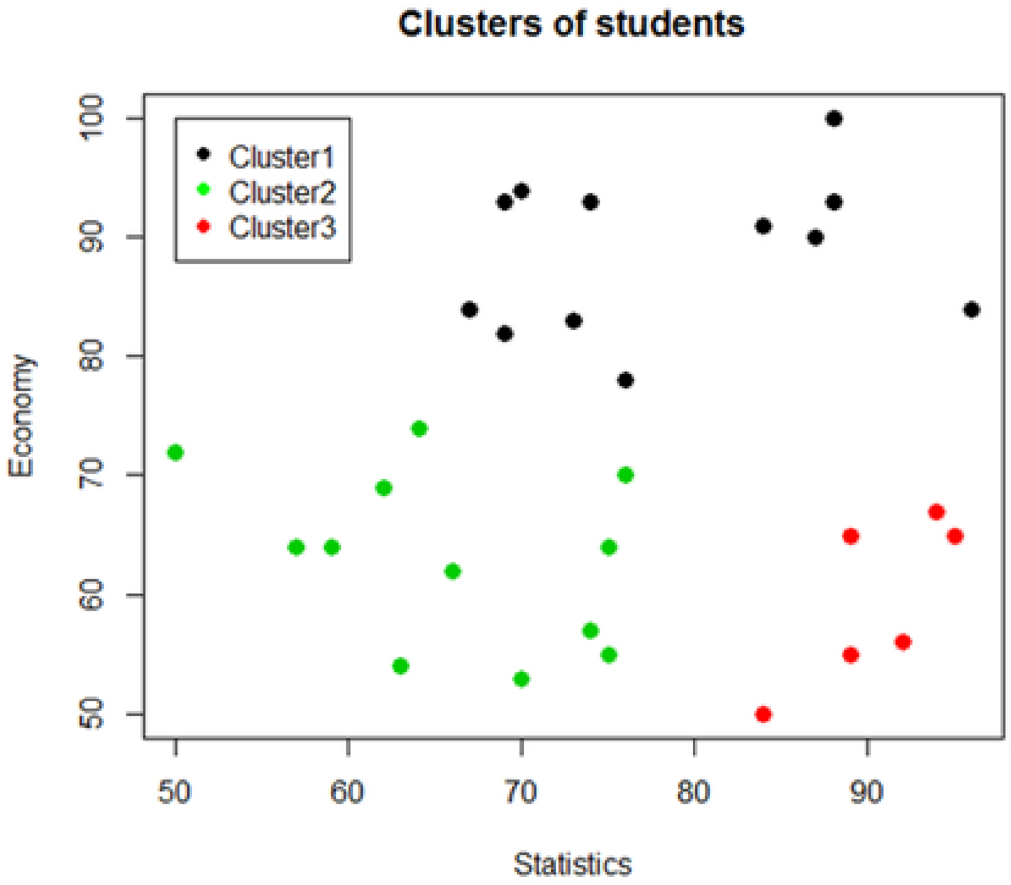
K-means Clustering

1) decide number of clusters
2) initialize the center of the clusters
3) assign each object to the group that has the closest centroid
4) when all objects have been assigned, recalculate the positions of the K centroids
5) repeat steps 3 and 4 until the centroids no longer move



```
# Load data from csv file into R
Grades <- read.csv("ittp.csv", header=TRUE, sep=";")

# Check whether the data are loaded correctly
View(Grades)

# Install package NbClust for determining optimal number of clusters
install.packages("NbClust")
library(NbClust)

# Function NbClust recommends us number of clusters according to 23
indexes
NbClust(Grades[,2:3], method="kmeans", min.nc=2, max.nc=5)

# Now we can perform k-means clustering
cluster <- kmeans(Grades[,2:3], centers=3)

# Plot the results in the scatter plot
plot(Grades$Statistics, Grades$Economy, col=cluster$cluster, pch=16,
main="Clusters of students", xlab="Statistics", ylab="Economy",
cex=1.2)

# Add a legend
legend(50, 100, pch=c(16,16,16), col=c("black", "green", "red"),
c("Cluster1", "Cluster2", "Cluster3"), bty="o", box.col="black",
cex=1)
```

Visualization of Clusters



Conclusion

- data in databases could be used for finding hidden value useful for business
- using **free** statistical programming language R it is possible to perform data mining tasks
- it is necessary to have data analyst/data scientist who is aware of these methods
- one example is e.g. segmentation of customers

Thank you for your attention.