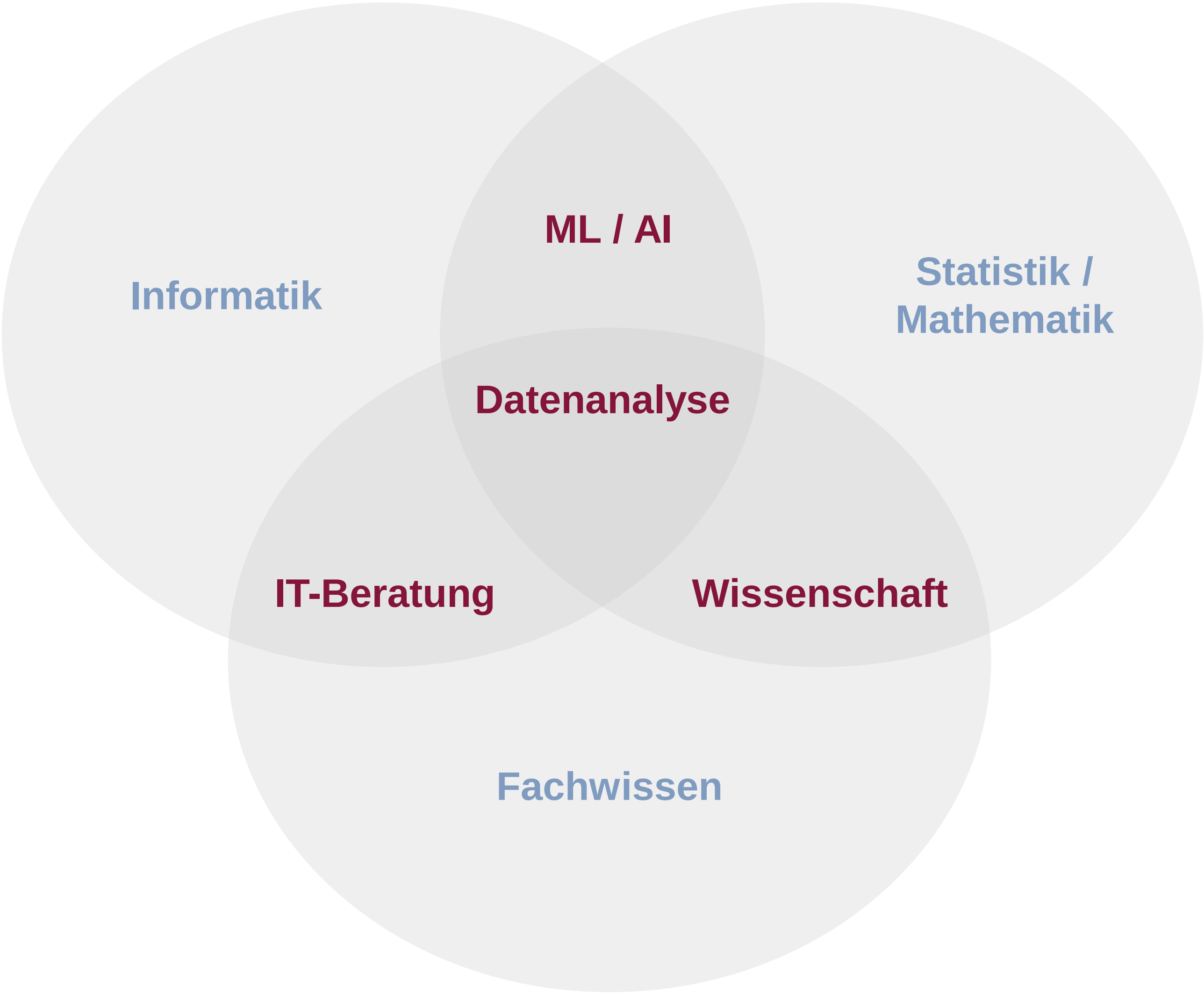
Datenanalyse

Automotive Communications Financial Services Government Insurance Life Science & Healthcare Travel & Logistics Utilities Automotive Communications Financial Services Government
Insurance Life Science & Healthcare Travel & Logistics Utilities Automotive Communications Financial Services Government Insurance Life Science & Healthcare Travel & Logistics
Utilities Automotive Communications Financial Services Government Insurance Life Science & Healthcare Travel & Logistics Utilities Automotive Communications Financial Services
Government Insurance Life Science & Healthcare Travel & Logistics Utilities Automotive Communications Financial Services Government Insurance Life Science & Healthcare Travel
& Logistics Utilities Automotive Communications Financial Services Government Insurance Life Science & Healthcare Travel & Logistics Utilities Automotive Communications
Financial Services Government Insurance Life Science & Healthcare Travel & Logistics Utilities Automotive Communications Financial Services Government Insurance Life
Science & Healthcare Travel & Logistics Utilities Automotive Communications Financial Services Government Insurance Life Science & Healthcare Travel & Logistics Utilities
Automotive Communications Financial Services Government Insurance Life Science & Healthcare Travel & Logistics Utilities Automotive Communications Financial Services
Government Insurance Life Science & Healthcare **Travel & Logistics** Utilities Automotive Communications Financial Services Insurance Life Science & Healthcare
Travel & Logistics Utilities Automotive Communications Financial Services Government Insurance Life Science & Healthcare Travel & Logistics Utilities Automotive
Communications Financial Services Government Insurance Life Science & Healthcare Travel & Logistics Utilities Automotive Communications Financial Services Government
Insurance Life Science & Healthcare Travel & Logistics Utilities Automotive Communications Financial Services Government Insurance Life Science & Healthcare Travel & Logistics
Utilities Automotive Communications Financial Services Government Insurance Life Science & Healthcare Travel & Logistics Utilities Automotive Communications Financial
Services Government Insurance Life Science & Healthcare Travel & Logistics Utilities Automotive Communications Financial Services Government Insurance Life Science &
Healthcare Travel & Logistics Utilities Automotive Communications Financial Services Government Insurance Life Science & Healthcare Travel & Logistics Utilities Automotive
Communications Financial Services Government Insurance Life Science & Healthcare Travel & Logistics Utilities Automotive Communications Financial Services Government Insurance
Life Science & Healthcare Travel & Logistics Utilities Automotive Communications Financial Services Government Insurance Life Science & Healthcare Travel & Logistics Automotive Utilities

.consulting .solutions .partnership

Datenanalyse nach Drew Conway



US Census Income Data

Explorative Datenanalyse

Daten

adult.data

```
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male,
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White
38, Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners, Not-in-family, White, Male, 0, 0,
53, Private, 234721, 11th, 7, Married-civ-spouse, Handlers-cleaners, Husband, Black, Male, 0,
28, Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-specialty, Wife, Black, Female, 0,
37, Private, 284582, Masters, 14, Married-civ-spouse, Exec-managerial, Wife, White, Female, 0,
49, Private, 160187, 9th, 5, Married-spouse-absent, Other-service, Not-in-family, Black, Female
52, Self-emp-not-inc, 209642, HS-grad, 9, Married-civ-spouse, Exec-managerial, Husband, White,
31, Private, 45781, Masters, 14, Never-married, Prof-specialty, Not-in-family, White, Female,
42, Private, 159449, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male,
37, Private, 280464, Some-college, 10, Married-civ-spouse, Exec-managerial, Husband, Black, Mal
30, State-gov, 141297, Bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, Asian-Pac-Is
```

- Erste 12 Instanzen mit
- 15 Variablen

Daten

CensusData.R | rawData

32,561 observations of 15 variables

	id	employerKind	fnlwgt	degree	yearsOfEd	maritalStatus	occupation	relationshipRole
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family
2	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband
3	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family
4	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband
5	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife
6	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife
7	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family
8	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband
9	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family
10	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband
11	37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband
12	30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband

- Erste 12 Instanzen als Data-Frame mit
- 8 von 15 Variablen

Daten

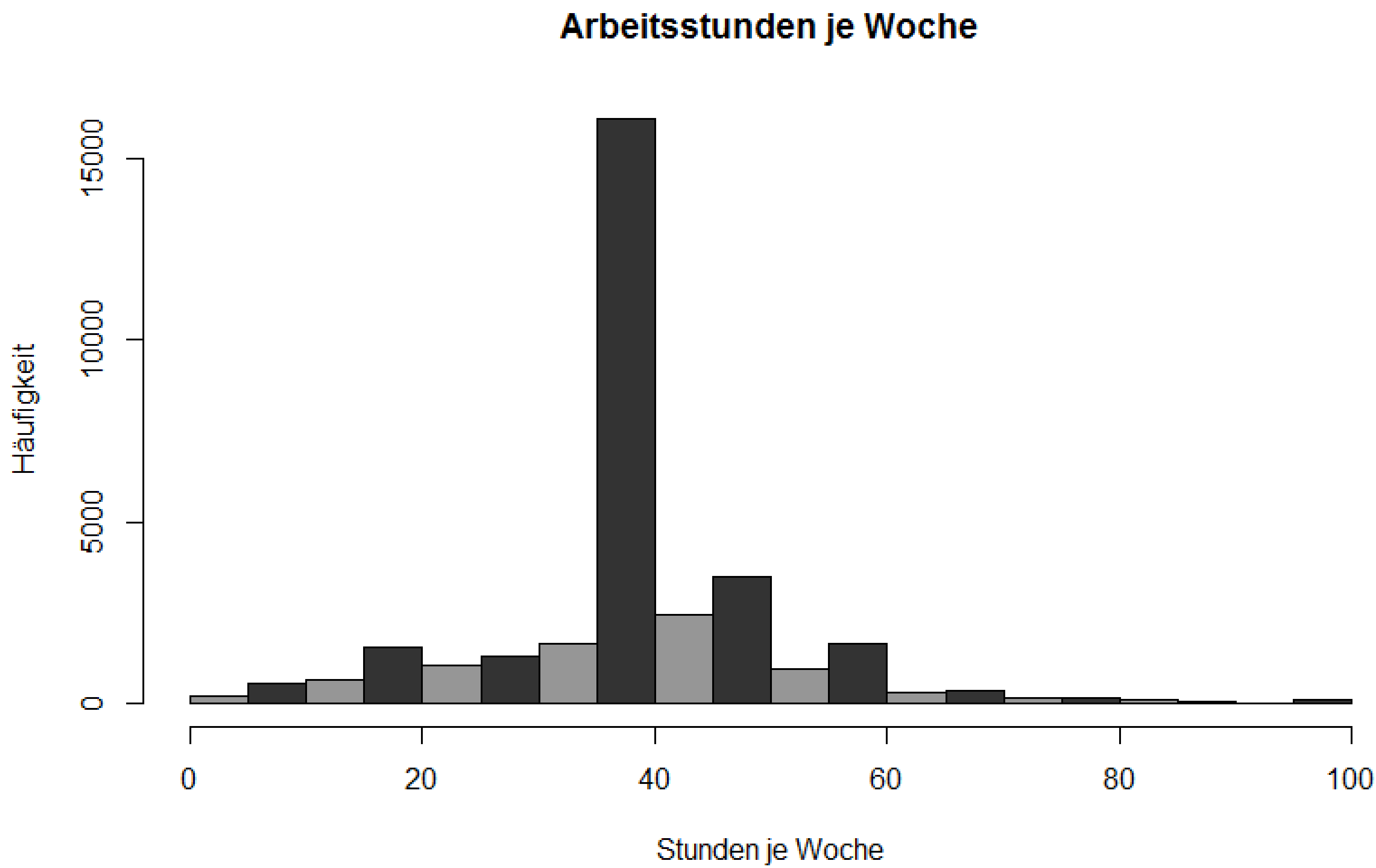
CensusData.R rawData

32,561 observations of 20 variables

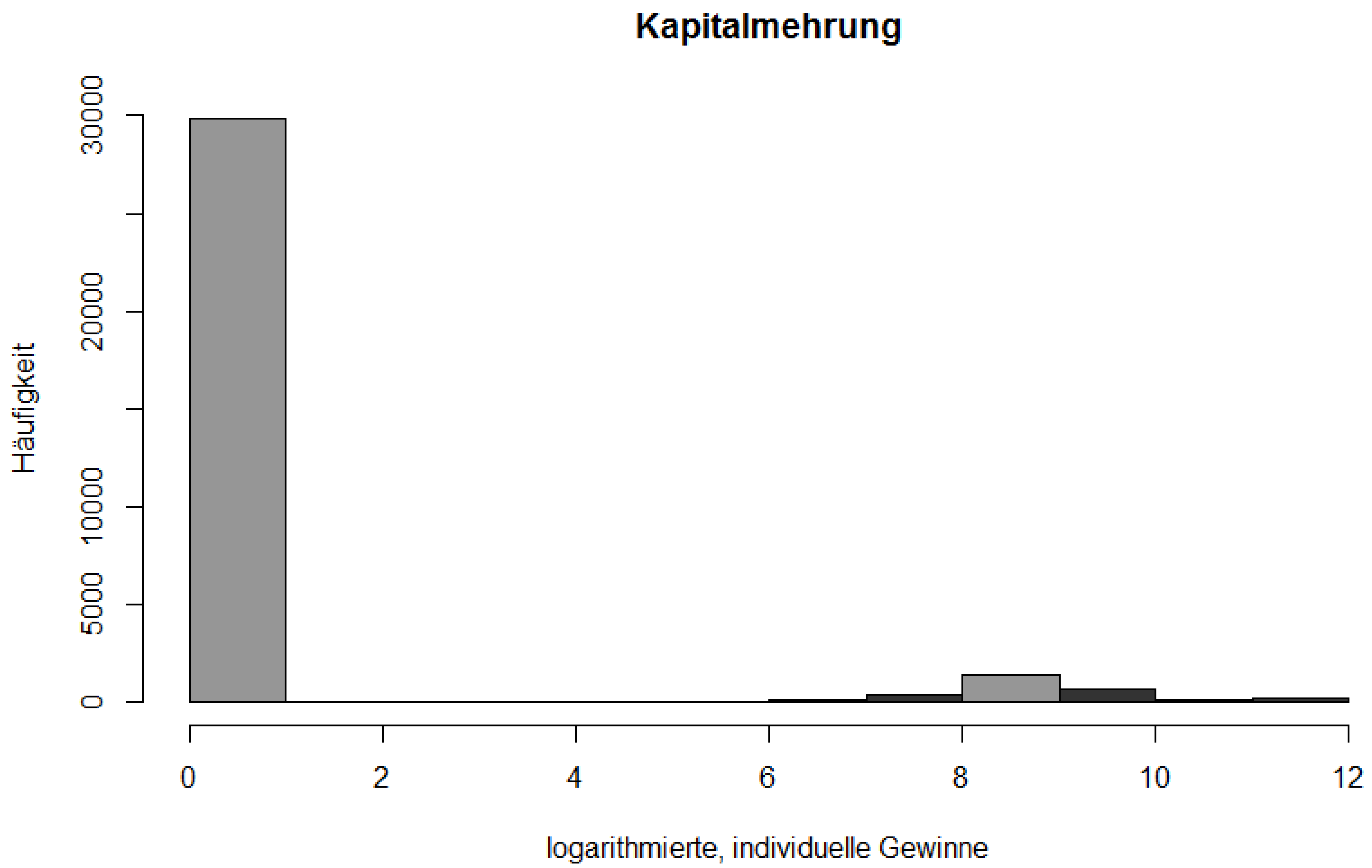
incomeGroup	academicLvl	incomeMoreThan50K	capitalDeviation	yearsOfEdStdUnits	workingHoursAWeekStdUnits
<=50K	Bachelor	TRUE	2.7562412	1.13472134	-0.03542890
<=50K	Bachelor	TRUE	-0.3071748	1.13472134	-2.22211900
<=50K	Highschool	TRUE	-0.3071748	-0.42005317	-0.03542890
<=50K	1	TRUE	-0.3071748	-1.19744043	-0.03542890
<=50K	Bachelor	TRUE	-0.3071748	1.13472134	-0.03542890
<=50K	Master	TRUE	-0.3071748	1.52341497	-0.03542890
<=50K	1	TRUE	-0.3071748	-1.97482769	-1.97915343
>50K	Highschool	FALSE	-0.3071748	-0.42005317	0.36951371
>50K	Master	FALSE	3.5009221	1.52341497	0.77445632
>50K	Bachelor	FALSE	3.1020895	1.13472134	-0.03542890
>50K	College	FALSE	-0.3071748	-0.03135955	3.20411198
>50K	Bachelor	FALSE	-0.3071748	1.13472134	-0.03542890

5 Sekundärvariablen

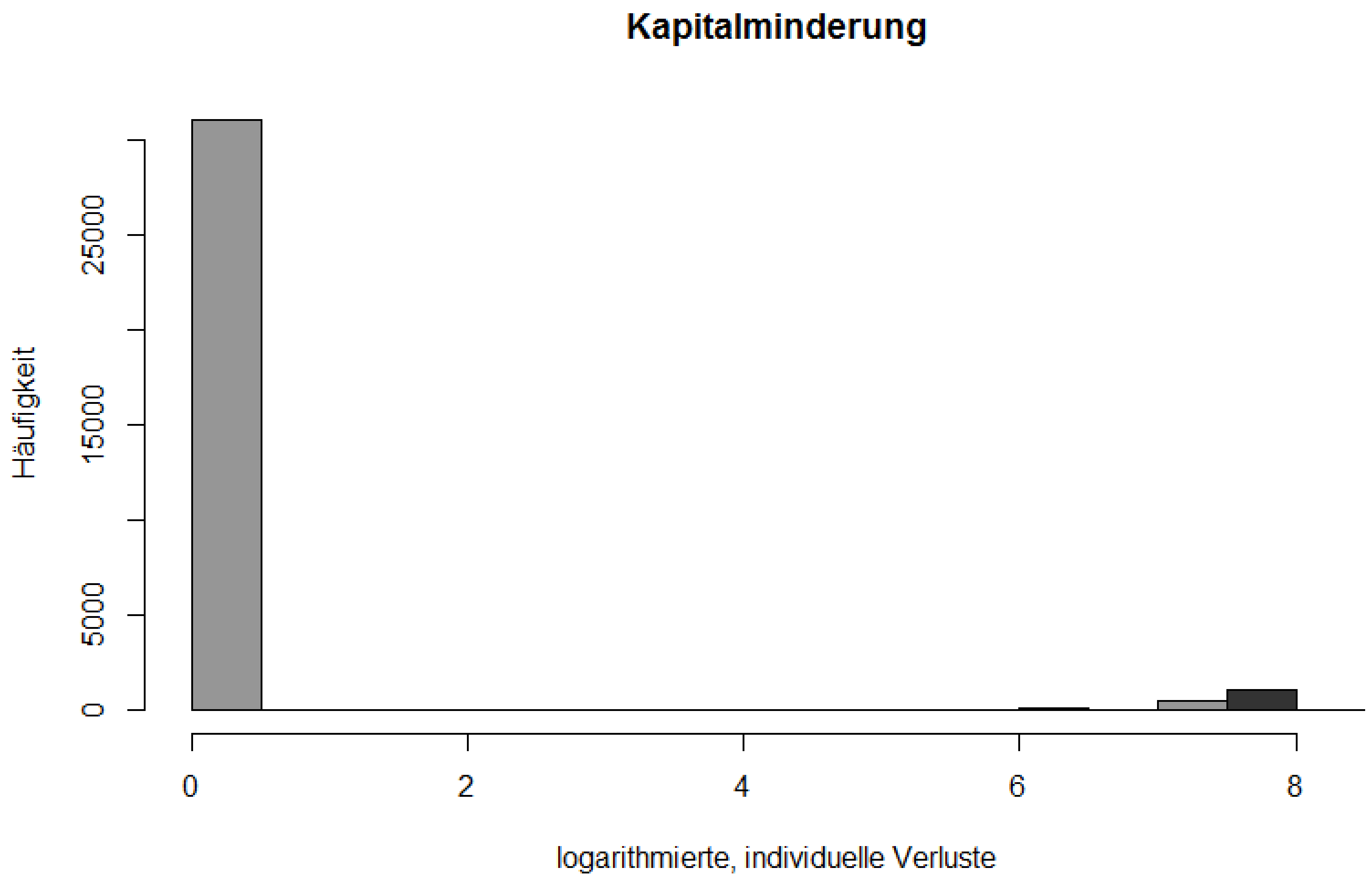
Visualisierungen



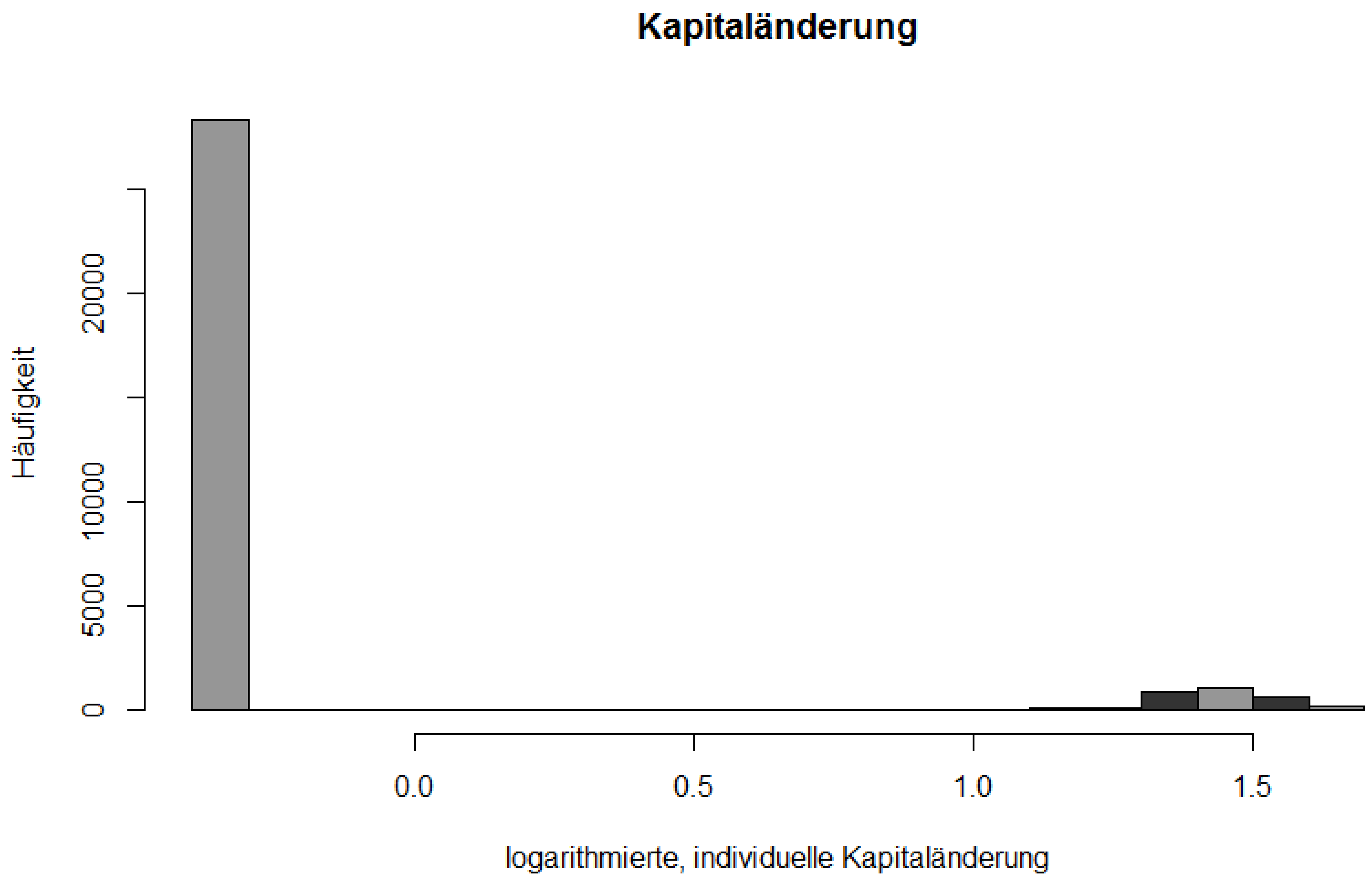
Visualisierungen



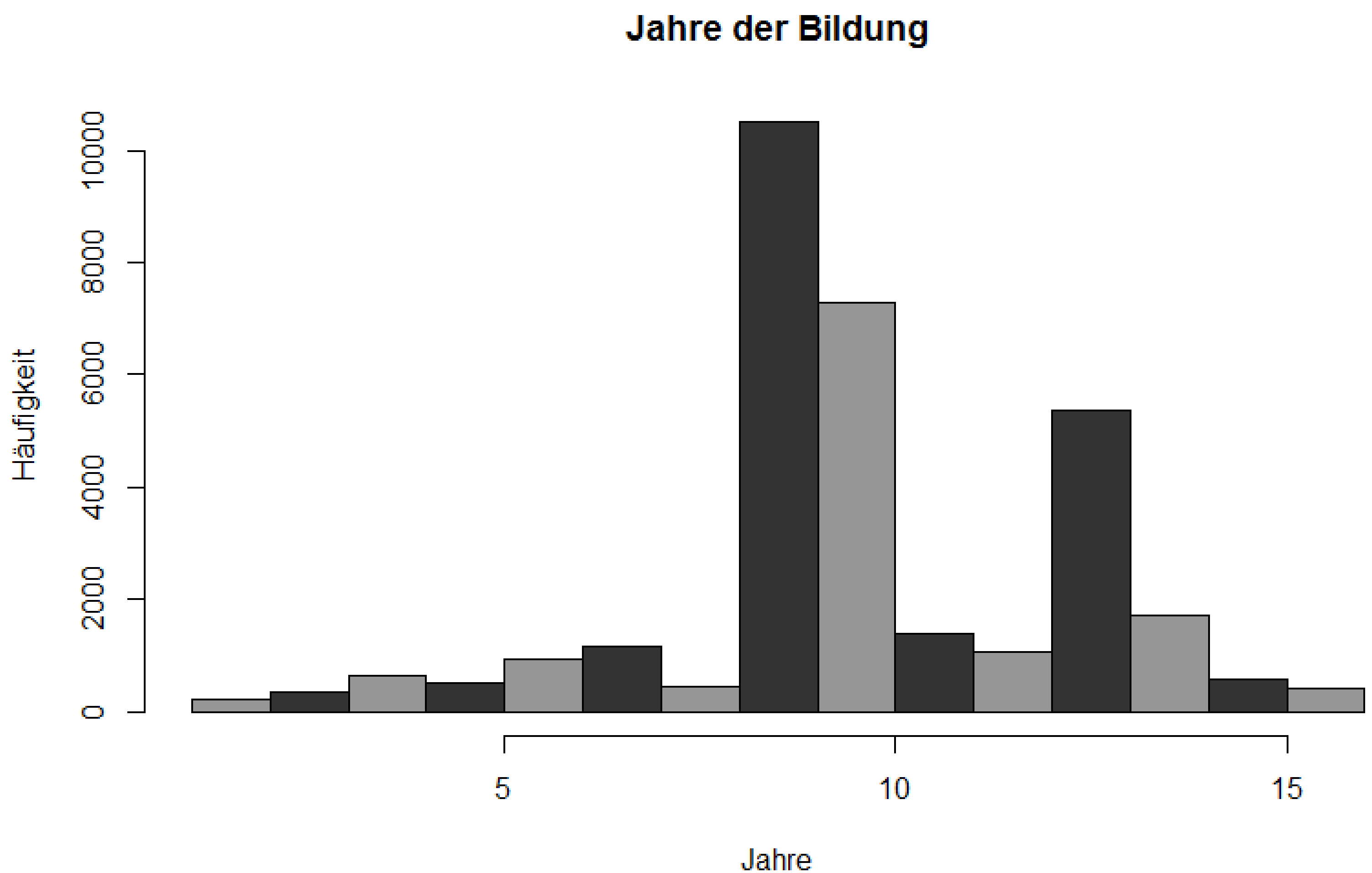
Visualisierungen



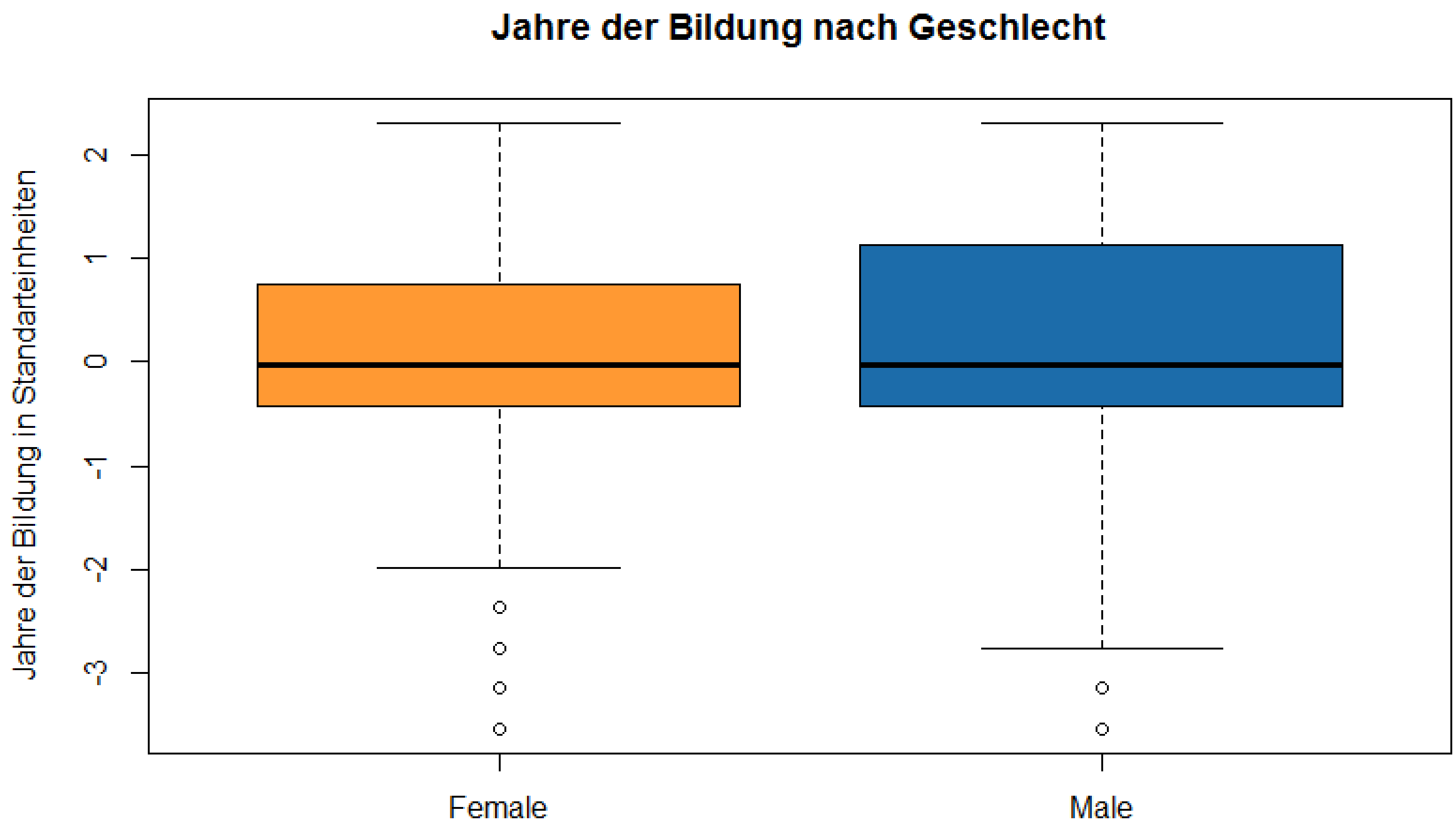
Visualisierungen



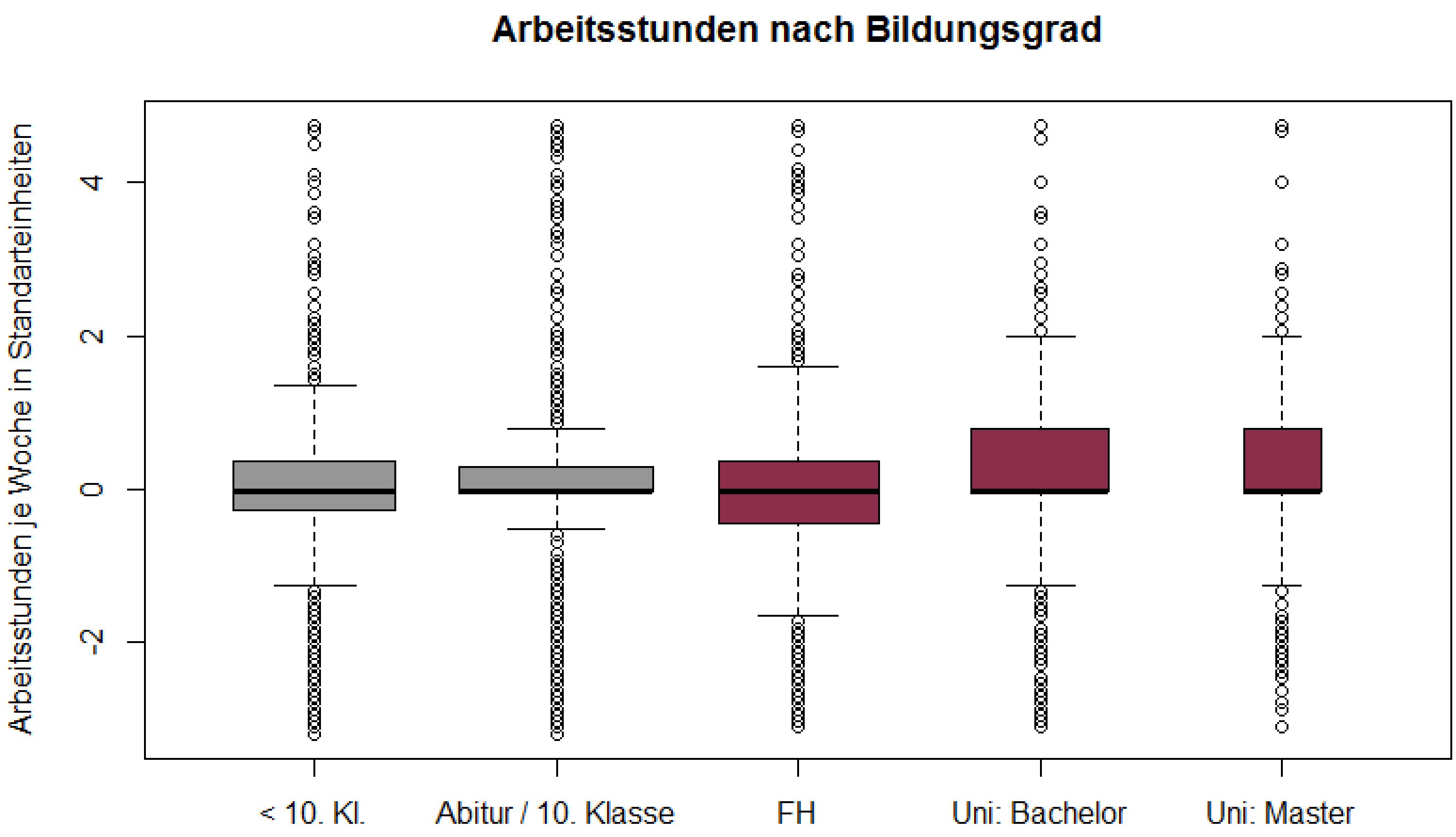
Visualisierungen



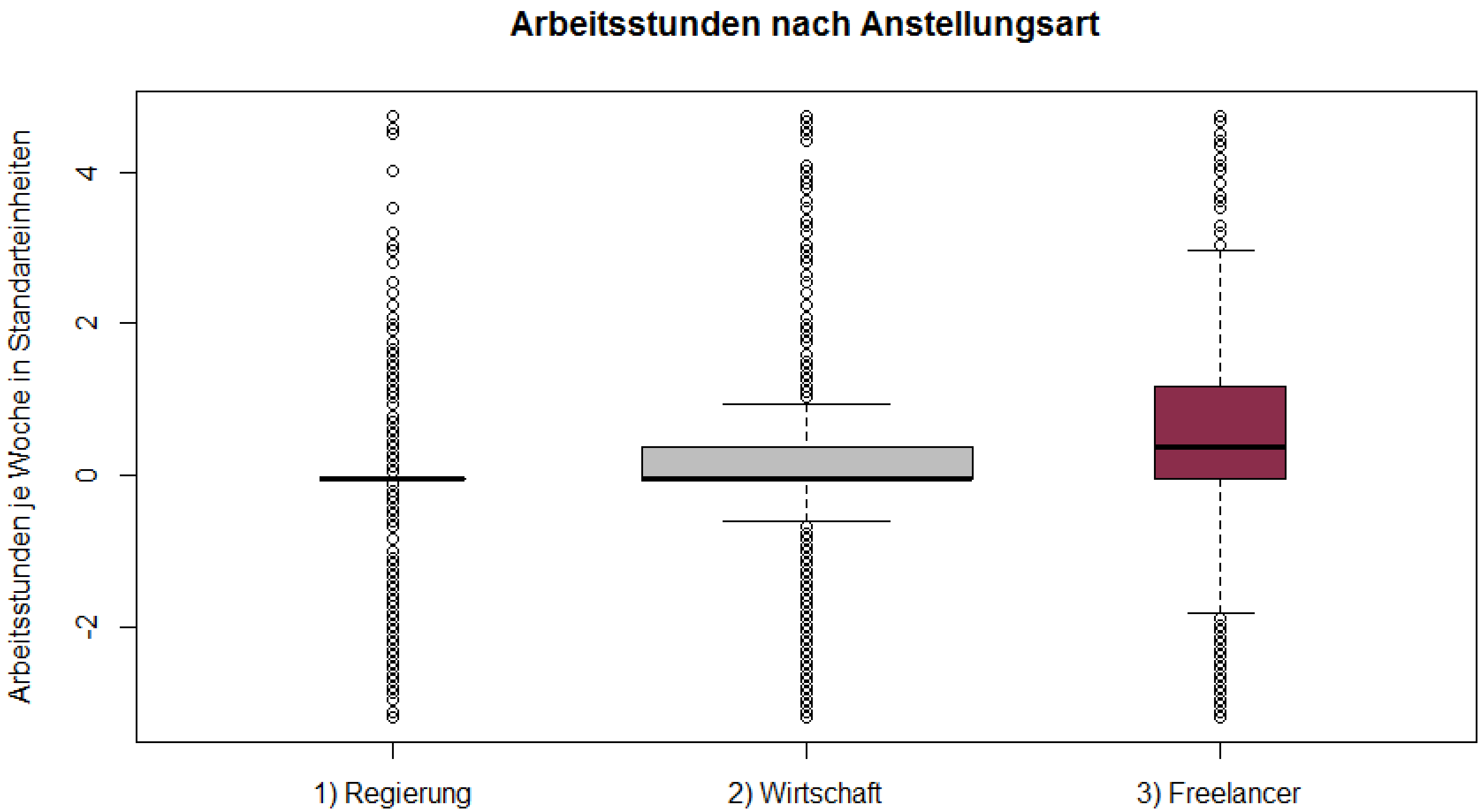
Visualisierungen



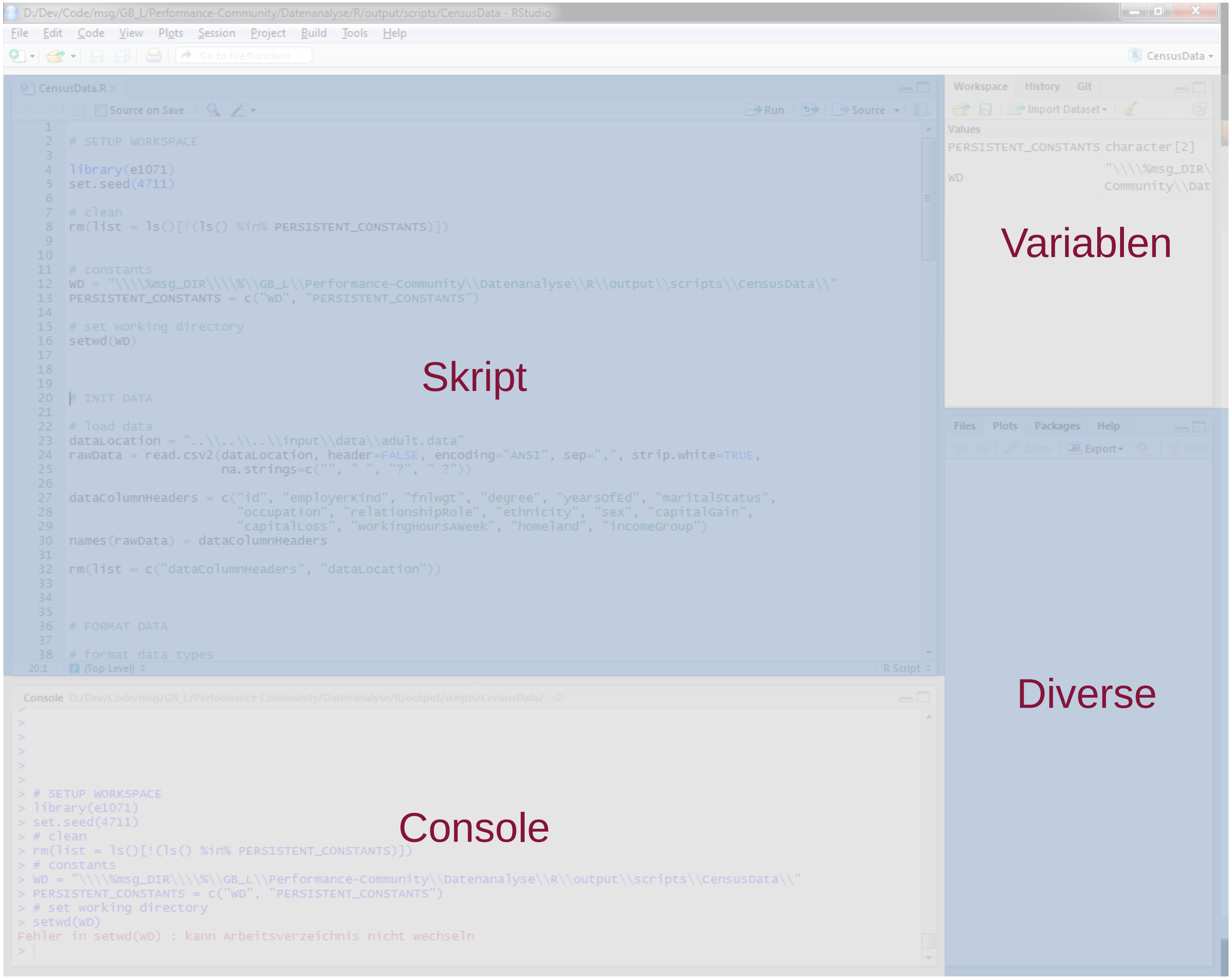
Visualisierungen



Visualisierungen



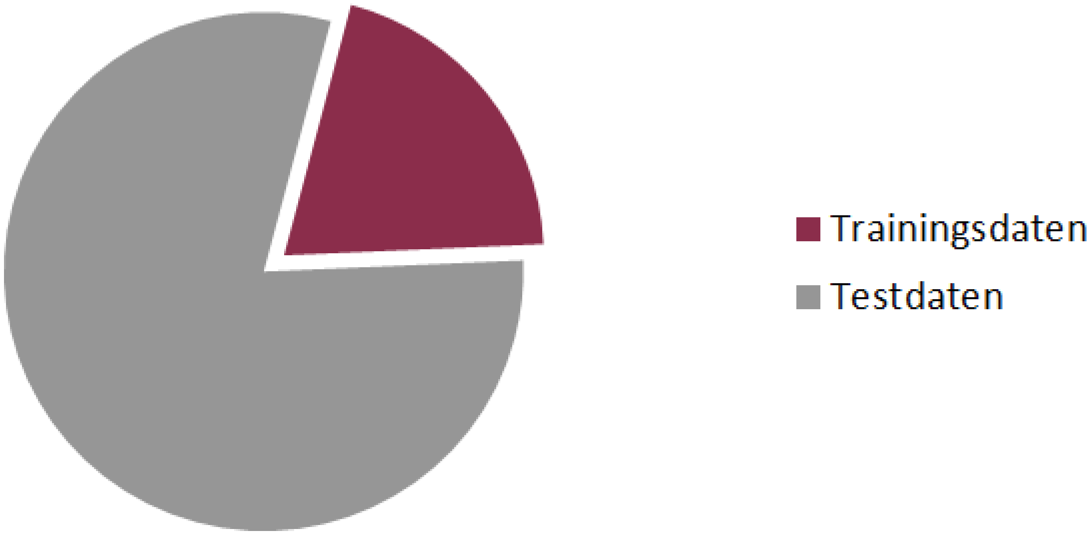
RStudio



US Census Income Data

Klassifikation

Klassifizierung

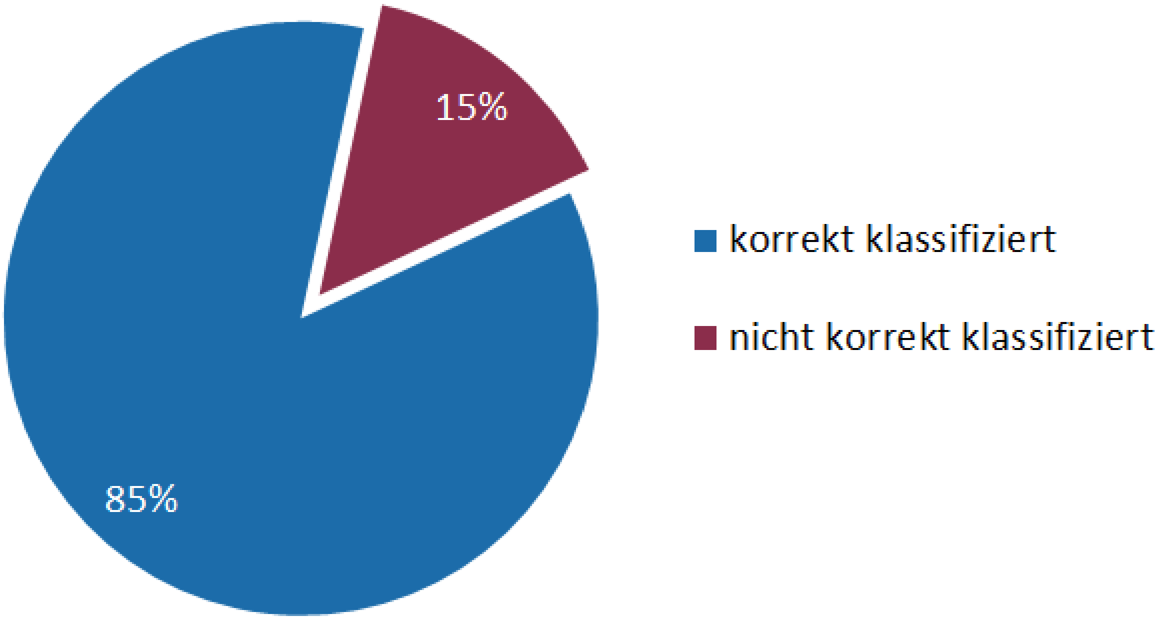


Klassifizierung

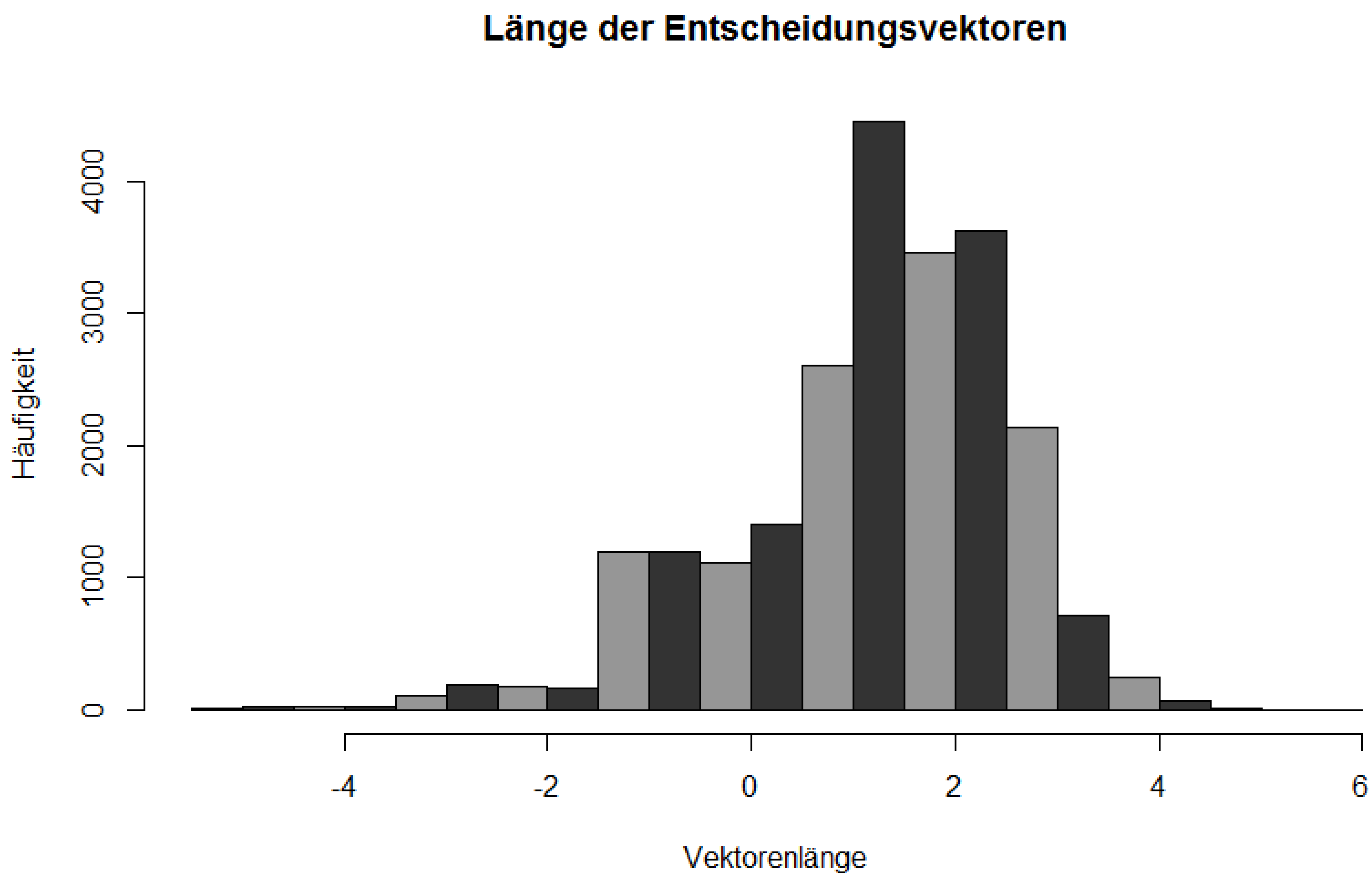
Klassifikation: Sensitivität & Spezifität

	tatsächlich nicht hohes Einkommen	tatsächlich hohes Einkommen
geschätzt nicht hohes Einkommen	13%	5%
geschätzt hohes Einkommen	10%	72%

Klassifizierung



Klassifizierung



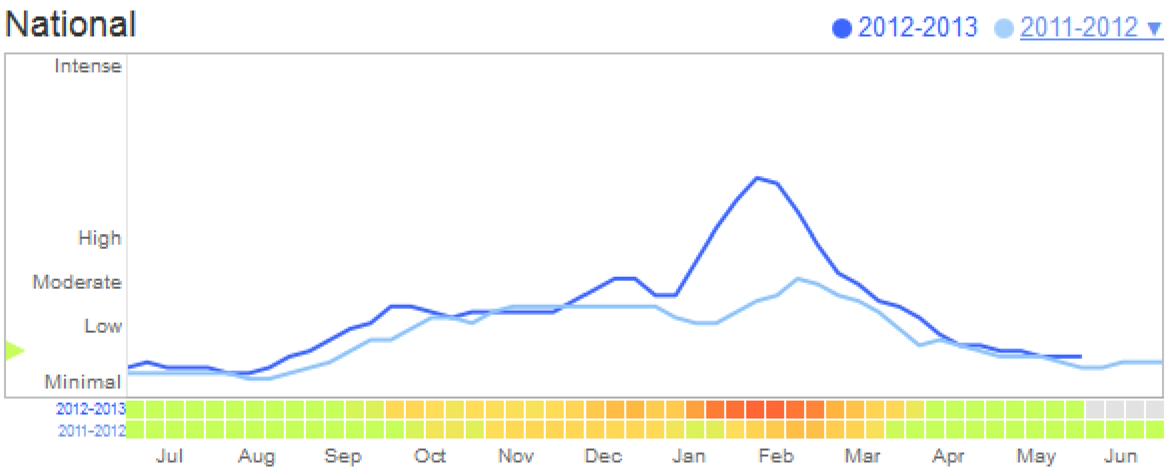
Anwendungsbeispiele

Anwendungsbeispiele

Google

Anwendungsbeispiele

Google



Anwendungsbeispiele

Stanford University

Anwendungsbeispiele

Prof. Sebastian Thrun mit autonom fahrenden VW Touareg „Stanley“

Resourcen

github SOCIAL CODING

LearningR

Training Dataset for R Beginners

Last updated 3 minutes ago

AddHealth-Data-Analysis

The analysis of biases and influencer in attendance of religious services

Last updated a month ago

github.com/danielschulz/LearningR

Datenquellen

UNIVERSITY of CALIFORNIA | IRVINE

UCI Machine Learning Repository
archive.ics.uci.edu/ml

Datenquellen

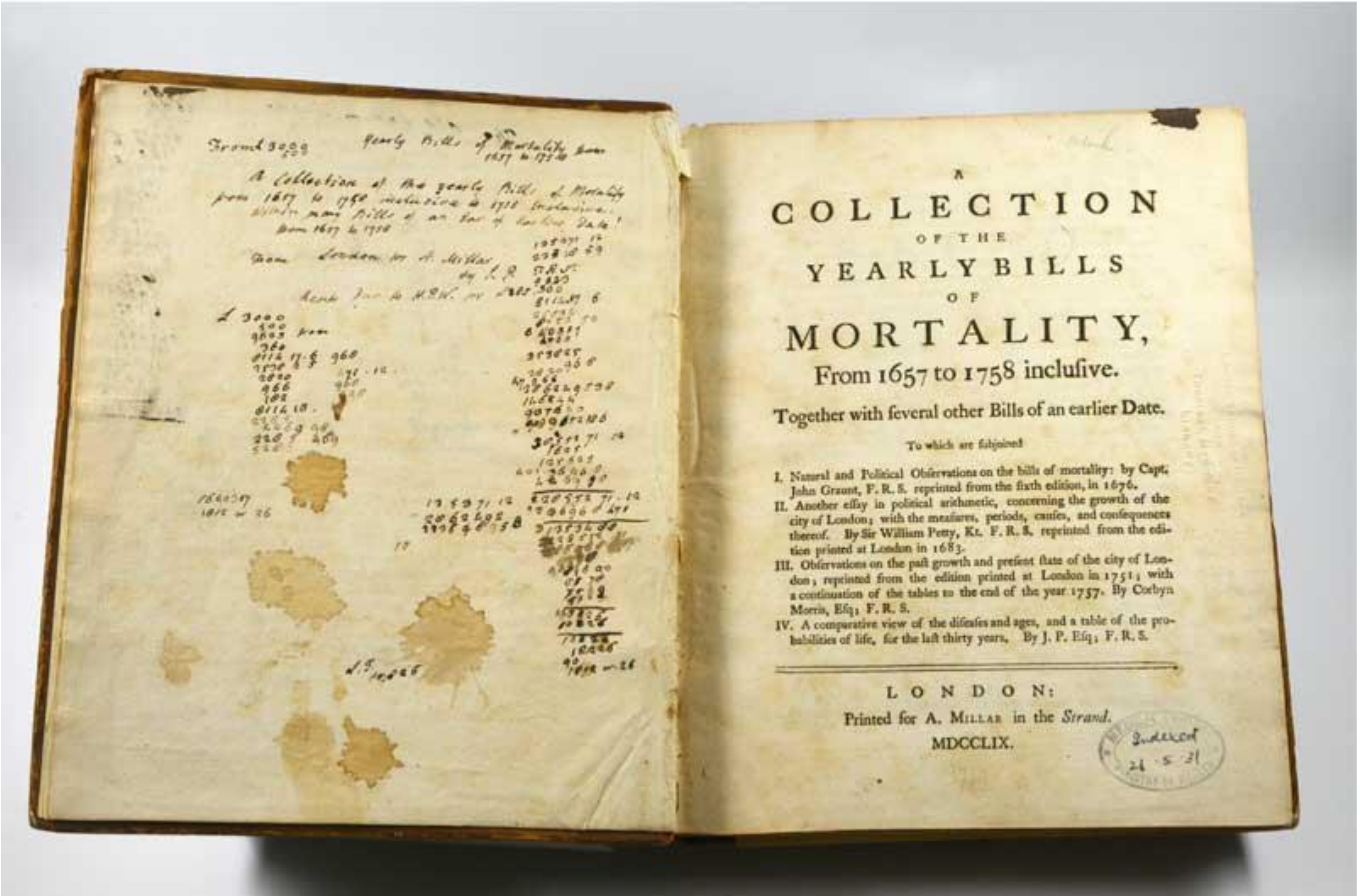
kaggle

kaggle.com

Datenquellen

data.gov

Bills of Mortality



A

COLLECTION

OF THE

YEARLY BILLS

OF

MORTALITY,

From 1657 to 1758 inclufive.

Together with feveral other Bills of an earlier Date.

To which are fubjoined

I. Natural and Political Obfervations on the bills of mortality: by Capt. John Graunt, F. R. S. reprinted from the fixth edition, in 1676.

II. Another effay in political arithmetic, concerning the growth of the city of London; with the meafures, periods, caufes, and confequences thereof. By Sir William Petty, Kt. F. R. S. reprinted from the edition printed at London in 1683.

III. Obfervations on the paft growth and prefent ftate of the city of London; reprinted from the edition printed at London in 1751; with a continuation of the tables to the end of the year 1757. By Corbyn Morris, Efq; F. R. S.

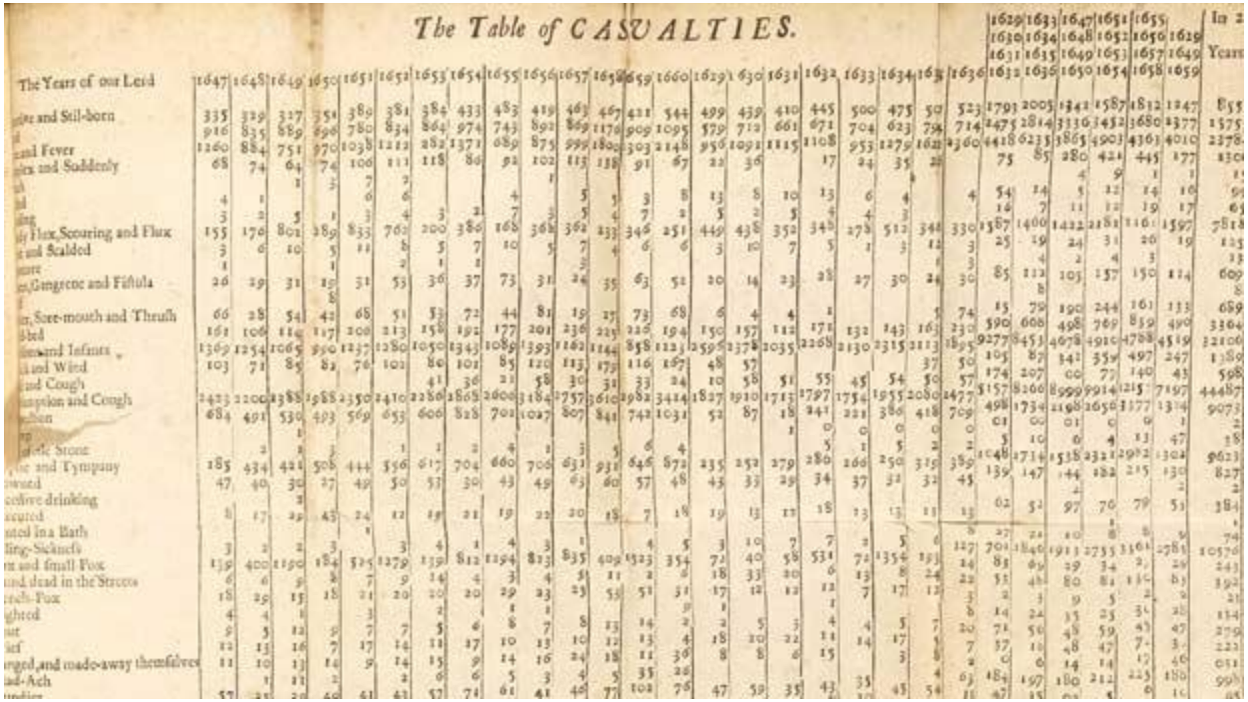
IV. A comparative view of the difeafes and ages, and a table of the probabilities of life, for the laft thirty years, By J. P. Efq; F. R. S.

LONDON:

Printed for A. Millar in the *Strand.*

MDCCLIX.

Bills of Mortality



The Table of CASUALTIES.

Wöchentliche Todesstatistiken

US Census Income Data

Übersicht: Code-Sektionen Schritt für Schritt

US Census Income Data

```r

# SETUP WORKSPACE

library(e1071)
set.seed(4711)

# clean
rm(list = ls()[!(ls() %in% PERSISTENT_CONSTANTS)])

```

Workspace einrichten

US Census Income Data

```r

# INIT DATA

# load data
dataLocation = "..\\..\\..\\input\\data\\adult.data"
rawData = read.csv2(dataLocation, header=FALSE, encoding="ANSI", sep=",", strip.white=TRUE,
                    na.strings=c("", " ", "?", " ?"))

dataColumnHeaders = c("id", "employerKind", "fnlwgt", "degree", "yearsOfEd", "maritalStatus",
                      "occupation", "relationshipRole", "ethnicity", "sex", "capitalGain",
                      "capitalLoss", "workingHoursAWeek", "homeland", "incomeGroup")
names(rawData) = dataColumnHeaders

rm(list = c("dataColumnHeaders", "dataLocation"))

```

Daten laden, Headernamen zuweisen

US Census Income Data

```
# FORMAT DATA

# format data types
rawData$id = as.numeric(rawData$id)
rawData$employerKind = as.factor(rawData$employerKind)
rawData$degree = as.factor(rawData$degree)

# assign secondary variable academic level
rawData$academicLvl = "none"
rawData$academicLvl = as.factor(rawData$academicLvl)

rawData$academicLvl = ifelse ("Doctorate" == rawData$degree || "Prof-school" == rawData$degree,
                              "PhD", rawData$academicLvl)
rawData$academicLvl = ifelse ("Masters" == rawData$degree, "Master", rawData$academicLvl)
rawData$academicLvl = ifelse ("Bachelors" == rawData$degree, "Bachelor", rawData$academicLvl)
rawData$academicLvl = ifelse ("Some-college" == rawData$degree, "College", rawData$academicLvl)
rawData$academicLvl = ifelse ("HS-grad" == rawData$degree, "Highschool", rawData$academicLvl)
```

- Daten-Typen zuweisen
- Sekundärvariablen einfügen

US Census Income Data

```
# assign secondary variable income to be more than 50000 USD / yr
rawData$incomeMoreThan50K = FALSE
rawData$incomeMoreThan50K = as.logical(rawData$incomeMoreThan50K)
rawData$incomeMoreThan50K = ifelse ("<=50K" == rawData$incomeGroup, TRUE, rawData$incomeMoreThan50K)


# assign secondary variable capitial deviation / difference in standard units
rawData$capitalDeviation = rawData$capitalGain - rawData$capitalLoss
rawData$capitalDeviation = scale(log(rawData$capitalDeviation + 1))

# assign secondary variable working hours / wk in standard units
rawData$yearsOfEdStdUnits = scale(rawData$yearsOfEd)
rawData$workingHoursAWeekStdUnits = scale(rawData$workingHoursAWeek)

# format data types
rawData$maritalStatus = as.factor(rawData$maritalStatus)
rawData$occupation = as.factor(rawData$occupation)
rawData$relationshipRole = as.factor(rawData$relationshipRole)
rawData$ethnicity = as.factor(rawData$ethnicity)
rawData$sex = as.factor(rawData$sex)
rawData$homeland = as.factor(rawData$homeland)
rawData$capitalDeviation = as.numeric(rawData$capitalDeviation)
rawData$workingHoursAWeekStdUnits = as.numeric(rawData$workingHoursAWeekStdUnits)
rawData$yearsOfEdStdUnits = as.numeric(rawData$yearsOfEdStdUnits)
```

- Daten-Typen zuweisen
- Sekundärvariablen einfügen

US Census Income Data

```
# DROP COLUMNS
dropColumns = c("id", "fnlwgt", "yearsOfEd", "workingHoursAWeek", "capitalGain",
                "capitalLoss", "incomeGroup")
rawData = rawData[,!(names(rawData) %in% dropColumns)]

# remove dropping column from workspace value list
rm(list = c("dropColumns"))
```

Nicht benötigte Spalten entfernen

US Census Income Data

```r
# SAMPLE TRAINING AND TEST DATA
rawData$clazz = sample(1:5, dim(rawData)[1], replace=TRUE)
rawData$clazz = as.factor(rawData$clazz)

data = rawData
data = na.omit(data) # drop missing value instances

train = subset(data, 1 == data$clazz)
test = subset(data, 1 != data$clazz)

dropColumns = c("clazz")
train = train[,!(names(train) %in% dropColumns)]
test = test[,!(names(test) %in% dropColumns)]
data = data[,!(names(data) %in% dropColumns)]

# remove dropping column from workspace value list
rm(list = c("dropColumns", "rawData"))
```

Trainings- und Testdaten erzeugen

US Census Income Data

```r

# TRAIN CLASSIFICATION MODEL SUPPORT VECTOR MACHINES AND EVALUATE ACCURANCY
svm = svm(train$incomeMoreThan50K ~ ., train, type="C-classification", probability=TRUE,
          gamma=0.0001, cost=100000)
pr = predict(svm, test, probability=TRUE)
# plot(formula=train$capitalDeviation ~ train$workingHoursAWeekStdUnits, data=train)
# plot(formula=test$capitalDeviation ~ test$workingHoursAWeekStdUnits, data=test)

table = table(classifications = pr, test$incomeMoreThan50K)
table

# chisquare = chisq.test(table)
# chisquare
# summary(chisquare)


sumInTable = 0

for (i in c(1:4)) {
  sumInTable = sumInTable + table[i]
}
for (i in c(1:4)) {
  table[i] = table[i] / sumInTable
}

# prediction accurancy is one the main diagonal table[1] + table[4] or for table t: t_11 + t_22
table

rm(list = c("i", "sumInTable", "chisquare"))

```

SVM-Classifizierung trainieren und testen

Resumée

Datenanalyse

- Google´s Chef-Ökonom Hal Varian
  - „The next sexy job“
  - „The ability to take data – to be able to understand it, to process it, to extract value from it, to communicate it – that´s going to be a hugely important skill.“
  - New York Times, 2009

- „Hot new gig in tech“ – Fortune

Vielen Dank für Ihre Aufmerksamkeit

.consulting .solutions .partnership