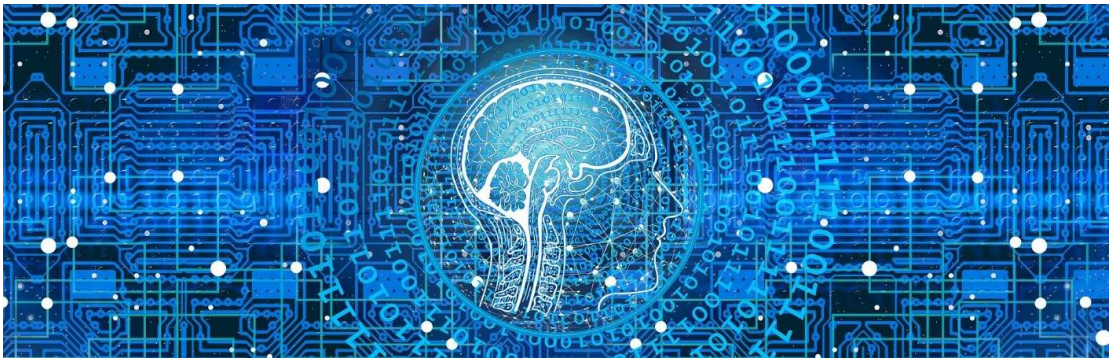
Data Science
2. Teil – Business Understanding, Begriffe, Statistik

Vorlesung an der DHBW Stuttgart, Prof. Dr. Monika Kochanowski



1

Kahoot

2

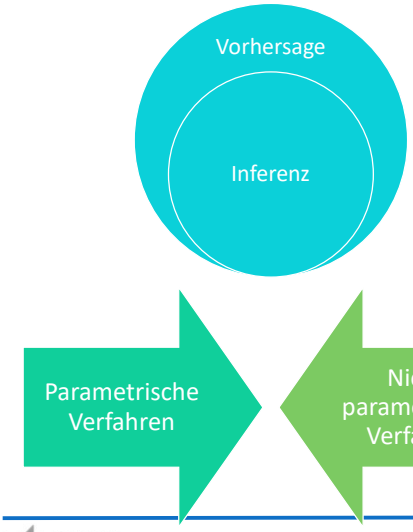
Inhalte der heutigen Vorlesung

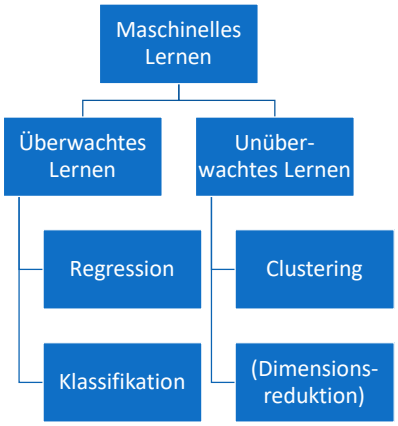
- Business Understanding
- Grundlagen maschinelles Lernen
- Statistik – Wiederholung
- Python – Einstieg

Übungen - Vorstellung

- Gruppenübung: Sie sind Dienstleister für einen kleinen Online-Spezialversandhandel für [**WÄHLEN SIE EIN THEMA AUS**]. Aktuell hat dieser 100 Kunden, die regelmäßig bestellen, und 1000, die einmal bestellt haben. Das sind die vorliegenden Daten. Sie haben von Data Science gehört, und möchten die Daten **»gewinnbringend«** einsetzen.
- **Erster Teil (max. 5 Minuten)**
  - Überlegen Sie sich ein Szenario.
- **Zweiter Teil (20 Minuten, in Gruppen von 4 Personen)**
  - Definieren Sie die **Ziele** Ihres Data Science Vorhabens.
    - Worum geht es? Was wollen wir lernen? Wer will was lernen?
  - Skizzieren Sie Ihr Vorgehen. Ordnen Sie die Schritte den CRISP-DM Phasen zu. Beschreiben Sie wichtige Meilensteine Ihres Projektes.
  - Gerne können Sie bereits Toolvorschläge, Algorithmen, Bewertungsmethoden etc. vorschlagen.
  - Spielen Sie die Fragestellungen zukünftig zu erfassender Daten durch.
- **Skizzieren** Sie Ihre Ergebnisse auf einem Flipchart. Stellen Sie diese vor (max. 2 – 3 Minuten).

Maschinelles Lernen
Motivation





6

Maschinelles Lernen
Begriffe

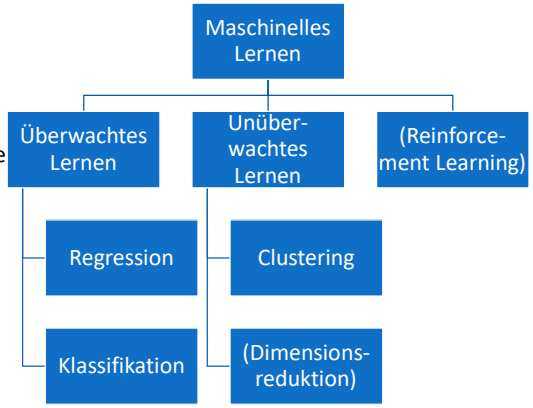
- Maschinelles Lernen (»Statistical Learning«) ist..
  - »Das Erstellen und Verwenden von Modellen, die mit Daten trainiert werden« [Joel Grues 2016]
  - Ein **Datensatz** hat N **Datenpunkte**
    - Diese haben p **Merkmale**, Variablen, Features, Attribute
    - Jeder dieser Merkmale hat eine Ausprägung / einen **Wert**
    - Wir sagen dann: der Datensatz hat p **Dimensionen**
- Ziele
  - **Vorhersage**: Vorhersage von Werten (wieviel ist das Fahrzeug wert?)
  - **Inferenz**: Verständnis von Zusammenhängen (wenn ein Fahrzeug ein Jahr altert, verliert es soviel Wert)

Automarke	Neuwert	Baujahr	Erwerber	Datum	Wohnort	Preis
DAIMLER	50.000 EUR	2014	Meier	12.03.2018	Stuttgart	15.000 EUR
Mercedes	60.000 EUR	2015	Müller	07.02.2018	Hamburg	10.000 EUR
..	..	..	..	..	..	..

7

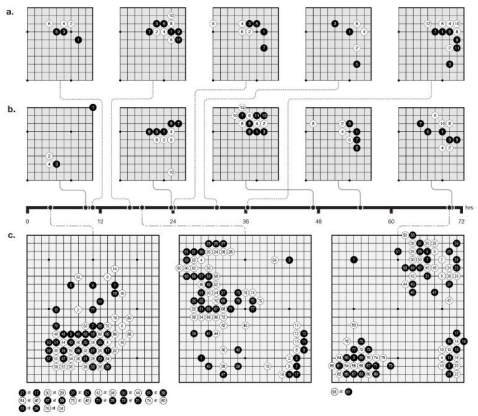
Maschinelles Lernen
Einordnung

- **Überwachtes Lernen (supervised)**
  - Für Datenpunkte liegen für die Merkmale (oder Variablen) sowohl Eingaben aus auch Ausgaben vor
    - **Eingabe**: Eingangsvariablen, Input, Prädiktor, unabhängige Variablen
    - **Ausgabe**: Reaktionsvariable, Output, abhängige Variablen
  - »For each **observation** of the **predictor** measurement(s) there is an associated **response** measurement.« [G. James et. al 2013]
- **Unüberwachtes** Lernen (unsupervised)
  - Clustering
  - (Dimensionsreduktion)
- **Halbüberwachtes** Lernen (semi-supervised)
  - **Motivation?**



Reinforcement Learning

- **Überwachte Lernverfahren**
  - Richtige Antwort ist “bekannt” und kann als Feedback zurückgegeben warden (Beispiel: Bildklassifikation)
  - “supervised learning systems that are trained to replicate the decisions of human experts”
- **Reinforcement Learning**
  - Kein direktes Feedback (Beispiel Go: Zug war „falsch“), es ist am Agenten zu entscheiden ob dieser Zug für den Spielverlust verantwortlich war
  - “reinforcement learning systems are trained from their own experience”
- Die Trennlinie ist nicht immer scharf
  - Es gibt ein **Kontinuum** zwischen überwachten und unüberwachten Lernverfahren
  - z. B. Reinforcement Learning („schwach überwacht“)
  - z. B. Generative Adversarial Networks (unüberwacht, aber es gibt Beispiele)



Silver, David; Schrittwieser, Julian; Simonyan, Karen; Antonoglou, Ioannis; Huang, Aja; Guez, Arthur et al. (2017): Mastering the game of Go without human knowledge. In: *Nature* 550 (7676), S. 354–359. DOI: 10.1038/nature24270.
Russell, Stuart J.; Norvig, Peter; Davis, Ernest; Edwards, Douglas (2016): Artificial intelligence. A modern approach. Third edition, Global edition. Boston, Columbus, Indianapolis, New York, San Francisco, Upper Saddle River, Amsterdam, Cape Town, Dubai, London, Madrid, Milan, Munich, Paris, Montreal, Toronto, Delhi, Mexico City, Sao Paulo, Sydney, Hong Kong, Seoul, Singapore, Taipei, Tokyo: Pearson (Always learning).
Mehralian, Mehran; Karasfi, Babak (2018 - 2018): RDCGAN: Unsupervised Representation Learning With Regularized Deep Convolutional Generative Adversarial Networks. In: 2018 9th Conference on Artificial Intelligence and Robotics and 2nd Asia-Pacific International Symposium. 2018 9th Conference on Artificial Intelligence and Robotics and 2nd Asia-Pacific International Symposium. Kish Island, Iran, 10.12.2018 - 10.12.2018: IEEE, S. 31–38.

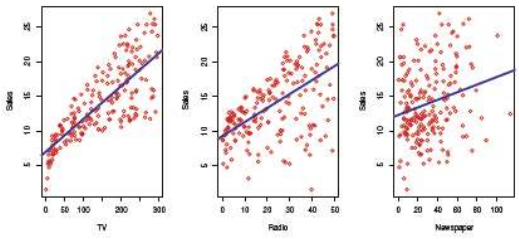
Maschinelles Lernen
Überwachtes Lernen

- Wiederholung: Beim überwachten Lernen liegen für die Datenpunkte für die Merkmale (oder Variablen) sowohl **Eingaben (predictor measurement) X** aus auch **Ausgaben (response measurement) Y** vor
- **Annahme**: es gibt eine Abbildung (oder eine) Funktion der Form $Y = f(X) + \varepsilon$
  - ε: stochastischer Fehlerterm ε
  - 15.000 EUR = f(DAIMLER, Rot, 2014, Meier, 12.03.2018, Stuttgart) + ε
- Um f zu finden ohne Vorwissen, können zwei Arten von Methoden eingesetzt werden
  - **Parametrische** Methoden (Annahme über die Form von f wird getroffen)
  - **Nicht-parametrische** Methoden (keine Annahme über die Form von f)

Automarke	Neuwert	Baujahr	Erwerber	Datum	Wohnort	Preis
DAIMLER	50.000 EUR	2014	Meier	12.03.2018	Stuttgart	15.000 EUR
Mercedes	60.000 EUR	2015	Müller	07.02.2018	Hamburg	10.000 EUR
..	..	..	..	..	..	..

Von der Wichtigkeit der Linearen Regression
Eine bewährte Methode

- Eines der ältesten Verfahren (wenn nicht das älteste Verfahren) des maschinellen Lernens
- Überwachtes Lernen
- Parametrische Methode
- **Einfache lineare Regression**
  - »simple linear regression« $y = ax + b$
- In »jeder« Software enthalten
- Liefert oft »erstaunlich gute« Ergebnisse
- Braucht »relativ wenige« Datenpunkte um dies zu tun
- Um viele Konzepte einzuführen, eignet sich die lineare Regression

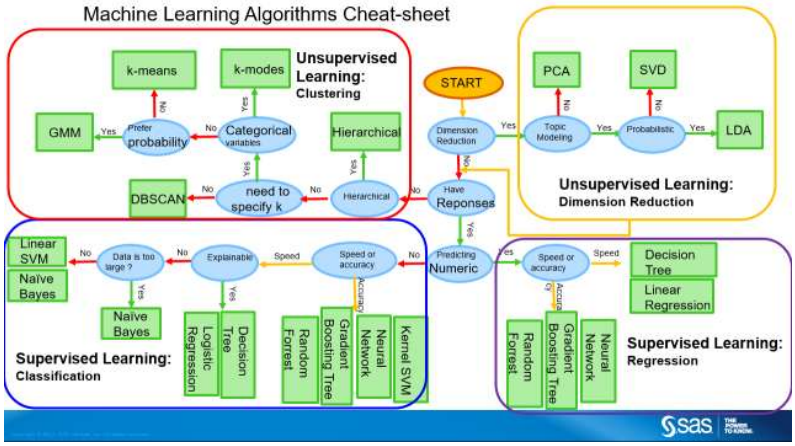


Bildquelle: [G. James et al, 2013]

Algorithmen für Maschinelles Lernen
Als Beispiel eines von (sehr) vielen Cheat Sheets

- Regression
- KNN
- Naive Bayes
- SVM
- ..

- Ziel: Grundlagen anhand bekannter Algorithmen legen und die »wichtigsten« kennenlernen



Bildquelle: https://whatsthebigdata.com/2017/05/02/types-of-machine-learning-algorithms-and-when-to-use-them/, am 10.02.2018

12

Gruppenübung

- **10 Minuten Zeit – 2er Gruppen – Diskussion in der großen Runde**
- Entscheiden Sie für die folgenden drei Szenarien (1) ob es sich um eine **Klassifikations-** oder **Regressionsaufgabe** handelt und geben Sie an, ob wir mehr in (2) **Inferenz** oder **Vorhersage** interessiert sind. Weiterhin geben Sie die (3) Anzahl der **Dimensionen** p und die Anzahl der **Datenpunkte** N an.
- **(a)** Wir haben eine Liste der Top-500 Unternehmen in Deutschland. Für jede Firma liegt der Gewinn, Umsatz, die Anzahl der Mitarbeiter, die Branche und das Gehalt des CEOs vor. Wir interessieren uns für die Faktoren welche das CEO-Gehalt beeinflussen.
- **(b)** Wir möchten in neues Produkt im Markt einführen und wollen wissen, ob es ein Erfolg oder ein Ladenhüter wird. Wir haben Daten von 20 ähnlichen Produkten welche eingeführt wurden. Wir wissen ob es ein Erfolg oder ein Ladenhüter war, den Preis, das Budget für Marketing, die Preise der Konkurrenz und zehn weitere Variablen.
- **(c)** Wir möchten gerne die prozentuale Veränderung des Euros in Zusammenhang mit den Börsen weltweit vorhersagen. Dafür liegen Daten aus ganz 2018 vor. Für jede Woche ist bekannt wie sich der Euro prozentual verändert hat, sowie die Börsenwerte er USA, GB und Deutschland.

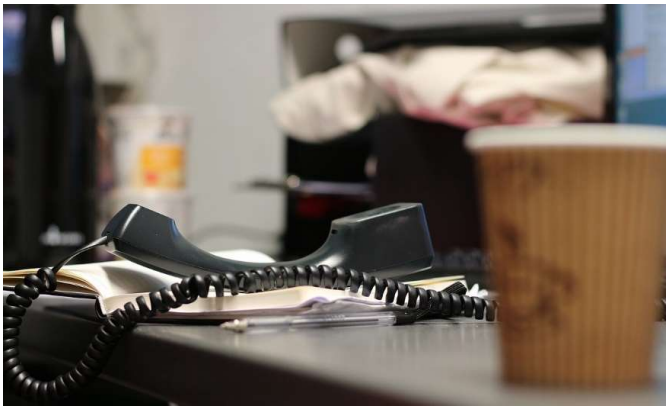
Quelle: Frei nach [James et al. 2013]

13

Business Understanding –
Vorschau Bewertungsmetriken für Data Science / Machine Learning Projekte

PROZESS 1

Anzahl Kunden	100.000
Kosten pro Anruf	6 €
Anteil der Verkäufe pro Anruf	0,15
Gewinn pro Verkauf	170 €
Bisheriger Prozess	
Alle anrufen	
Anzahl Anrufe	100.000
Kosten Anrufe	600.000 €
Gewinn Verkauf	2.550.000 €
Summe	**1.950.000 €**



Business Understanding –
Vorschau Bewertungsmetriken für Data Science / Machine Learning Projekte

PROZESS 2

Kaufende Kunden	15.000
Nicht kaufende Kunden	85.000
Genauigkeit	0,8
Erkannte Kaufende	12.000
Erkannt nicht Kaufende	68.000
Neuer Prozess: Data Science optimiert	
Vorhersage mit o.g. Genauigkeit	
Anzahl Anrufe	29.000
Kosten Anrufe	174.000 €
Gewinn Verkauf	2.040.000 €
Summe	**1.866.000 €**

PROZESS 3

Provision	22 €
Gewinn	148 €
Kosten pro Anruf	2 €
Genauigkeit	0,635
Erkannte Kaufende	13.500
Erkannte nicht Kaufende	50.000
Alternativer Prozess: Outsourcing	
Vorhersage mit o.g. Genauigkeit	
Anzahl Anrufe	48.500
Kosten Anrufe	0 €
Gewinn Verkauf	1.998.000 €
Summe	**1.998.000 €**
Dienstleistersicht:	
Kosten	97.000 €
Gewinn Verkauf	297.000 €
Summe	**200.000 €**

Aufgabe: Beantworten Sie folgende Fragen

- **15 Minuten Zeit – 2er Gruppen – Diskussion in der großen Runde**
- Betrachten Sie die drei Prozesse:
  - Prozess 1: ursprünglicher Prozess (alle anrufen)
  - Prozess 2: Verbesserter Prozess (Vorauswahl mit Vorhersageverfahren)
  - Prozess 3: Outgesourcter Prozess (Vorauswahl durch Dienstleister)
- Welche von den drei Lösungen ist für das Unternehmen am Besten?
- Welche von den drei Lösungen ist für den Dienstleister am Besten?
- Was hat die Data Science Abteilung des Unternehmens falsch gemacht?
- Wie müsste es eigentlich „richtig“ gemacht werden?
- Was hat das an verschiedenen Stellen für einen Einfluss auf den Data Science Prozess?

Statistische Inferenz

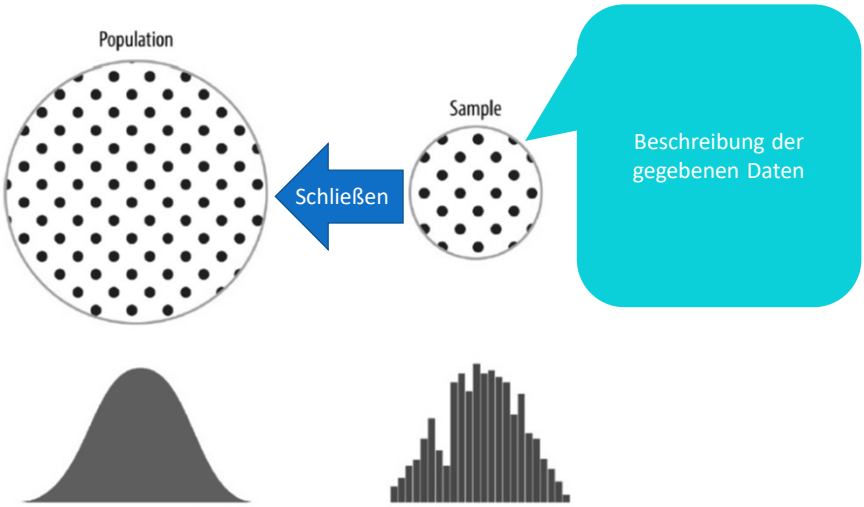
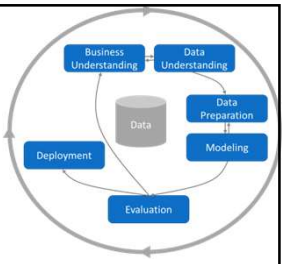


Figure 2-1. Population versus sample [Bruce et al. 2020]:

18

Vorgehen bei statistischer Inferenz



Hypothesen aufstellen → Experiment designen → Daten sammeln → Inferenz / Schlüsse ziehen

[Bruce et al. 2020]:

19

Datentypen und –skalen: Erster Durchlauf

Skala	Alias	Mögliche Operationen	Beispiele und Aussagen
Nominalskala	Kategoriale Daten, qualitatives Merkmal	Gleichheit, Ungleichheit (== / !=), Häufigkeit (Modus)	Zweitstimme bei der Bundestagswahl, Geschlecht (gleiche Wahl wie ..)
Ordinalskala	Rangordnung	Ordnen möglich (==, !=, >, <), Häufigkeit, Reihenfolge, Median	Likert-Skala (stimme zu, stimme eher zu, teils / teils, stimme eher nicht zu, lehne ab), Schulnoten (mehr / weniger als..)
Intervallskala (Kardinalskala)	Quantitative Merkmale, metrische Daten, numerical data	Abstände (Intervalle) besitzen eine Bedeutung (==, !=, >, <, +, -, *, %), Häufigkeit, Reihenfolge, Abstand, arith. Mittel	Temperatur (Celsius), Geburtsjahr (Unterschied ist..), IQ
Verhältnisskala (Kardinalskala)		Mit absolutem Nullpunkt, Häufigkeit, Reihenfolge, Abstand	Einkommen, Alter (doppelt so viel), Geschwindigkeit, Längen, Zeiten, ..

Datenexploration: Wiederholung Deskriptive Statistik

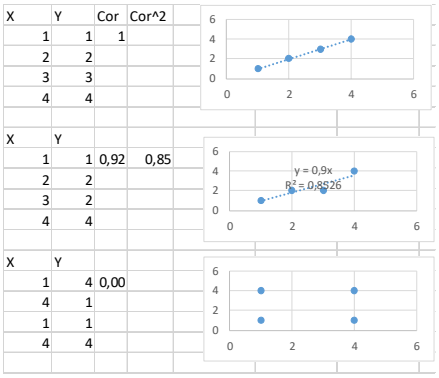
- Berechnete Maßzahlen, Parameter, Kennzahlen und Grafiken, die charakteristische Eigenschaften einer Datenmenge kennzeichnen bzw. darstellen.

- Vor allem wichtig bei sehr **großen, unübersichtlichen** Datenmengen.
- Arithmetisches Mittel (Mean)
- Modus (Mode)
- Median
- Stichprobenvarianz (Sample Variance)
- Standardabweichung (Sample Standard Deviation)
- Interquantilbereich (IQR = Inter Quartile Range)
- Korrelation

Vor- und Nachteile der Maße

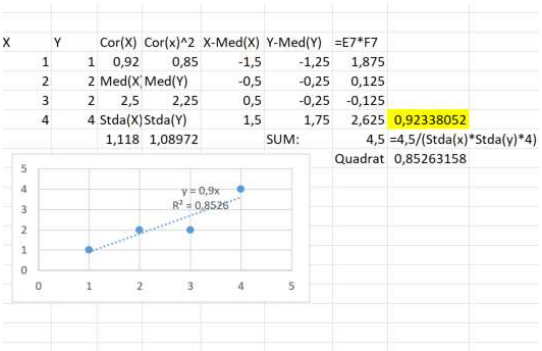
Hintergrund: Statistik-Grundlagen
Lageparameter

- **Mittelwert**: $\bar{x} = \text{mean}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Im Beispiel rechts: $\bar{x} = (1 + 2 + 3 + 4)/4 = 2{,}5$
  - Mittleres Beispiel: $\bar{y} = 2{,}25$
- **Median**: $\tilde{x} = \text{med}(X) \begin{cases} x_{\frac{n+1}{2}} & \text{n ungerade} \\ (x_{\frac{n}{2}} + x_{\frac{n}{2}+1})/2 & \text{n gerade} \end{cases}$
  - Im Beispiel: $\frac{2+3}{2} = 2{,}5$ oder $\frac{1+4}{2} = 2{,}5$
  - **Quantile**: $: \tilde{x} = \begin{cases} x_{(\lfloor np+1 \rfloor} & \text{np nicht ganzzahlig} \\ (x_{np} + x_{np+1})/2 & \text{np ganzzahlig} \end{cases}$
  - Wichtig in der Literatur: Quartile, dann ist p = 0,25 und 0,75, d. h. jeweils 25 % aller Werte bzw. 75 % aller Werte liegen unter diesem Wert
    auch: **25 % oder 75 % Perzentile**
  - Im Beispiel rechts oben: $\frac{1+2}{2} = 1{,}5$ sowie 3,5
- **Modus**: die häufigste Zahl (mittleres Bsp.: $\text{mod}(Y) = 2$)
- **Diskussion: Was sagen uns diese Werte?**

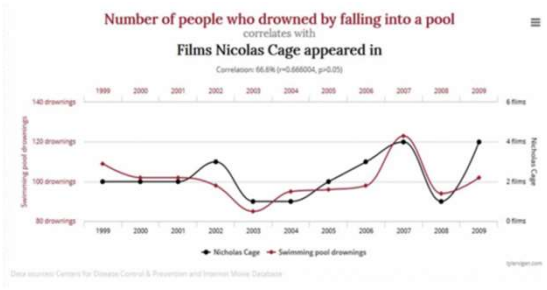


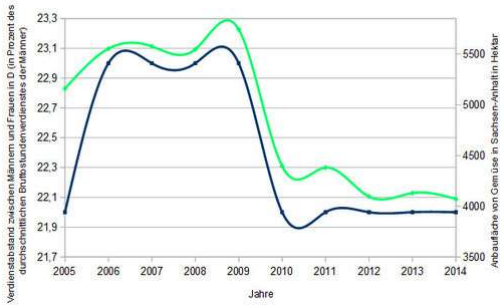
Hintergrund: Statistik Grundlagen
Streuungsparameter und Zusammenhangmaße

- Varianz: $Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - Im Beispiel: $\frac{(1{,}5^2 + 0{,}5^2 + 0{,}5^2 + 1{,}5^2)}{4} = 1{,}25$
  - Standardabweichung: $\sigma = \sqrt[2]{Var(x)}$
    - Im Beispiel: 1,12
- Korrelation: $Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$
  - Für den mittleren Fall:
  - $Cor(X,Y) = \frac{-1{,}5*-1{,}25+(-0{,}5*-0{,}25)+0{,}5*-0{,}25+1{,}5*1{,}75}{1{,}12*1{,}09*4} = 0{,}92$
  - Interessant: $0{,}92^2 = 0{,}85 = R^2$ = Bestimmtheitsmaß
    (ACHTUNG: Dieser Zusammenhang gilt nur für einfache lineare Regression!)



ACHTUNG: Scheinkorrelationen



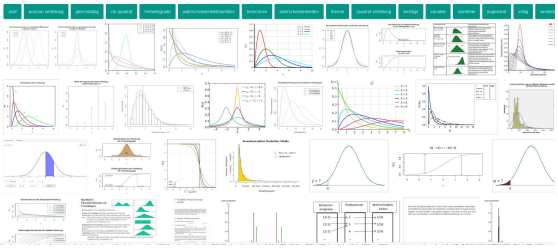


Bildquelle: https://scheinkorrelation.jimdo.com/

- »Scheinkorrelation bezeichnet den scheinbaren kausalen Zusammenhang zwischen korrelierenden Variablen, der zwar statistisch existent ist, aber nicht auf ein Ursache-Wirkungsprinzip zurückgeführt werden kann.
  - Beispielsweise steigt die Häufigkeit von Krankenhausbesuchen mit dem Nettoeinkommen an. Ursache hinter diesem Phänomen ist aber wahrscheinlich nicht das Einkommen, sondern das Alter der Befragten. Ältere Menschen verdienen im Schnitt mehr als jüngere Menschen – und sind gleichzeitig anfälliger für Erkrankungen und Beschwerden.«
  - Quelle: de.statista.com, abgerufen am 12.02.2018

Hintergrund: Verteilungen und Ihre Eigenschaften

- Verteilungen haben bestimmte Eigenschaften, die man für Datenanalysen nutzen kann
  - Aber: »they only have names because someone observed them enough times to think they deserved names.« [O'Neil and Schutt 2013]
- Wiederholung: Modus / Median / Mittelwert / »schief«



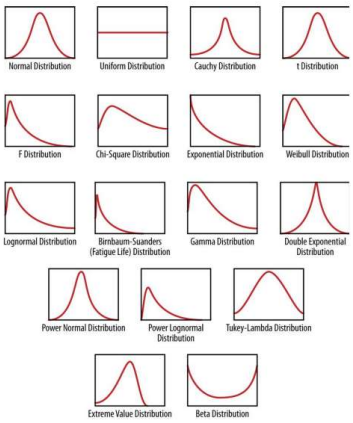


Figure 2-1. A bunch of continuous density functions (aka probability distributions)

Bildquelle: [O'Neil and Schutt 2013]

Unsicherheit und Streuung von Daten in Stichproben

	Streuung (Spread) der Stichprobe	Unsicherheit (Uncertainty) für Rückschlüsse auf Population
Parametrisch	Standardabweichung (Standard Deviation)	Standardfehler (Standard Error)
Nicht-parametrisch (Datenbasiert)	**Perzentilen** (z. B. in Boxplots dargestellt)	**Konfidenzintervall** (Bootstrap)

Oft hilfreich: Grafische Darstellung von Streuung / Unsicherheit

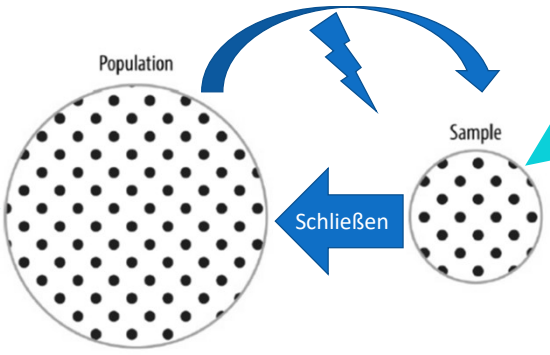
https://seaborn.pydata.org/tutorial/error_bars

27

Statistische Inferenz

Unsicherheit durch Stichprobenwahl (Zufall!)

Abschätzen der Unsicherheit durch den Sampling Prozess, entweder parametrisch oder nicht-parametrisch über Bootstrap

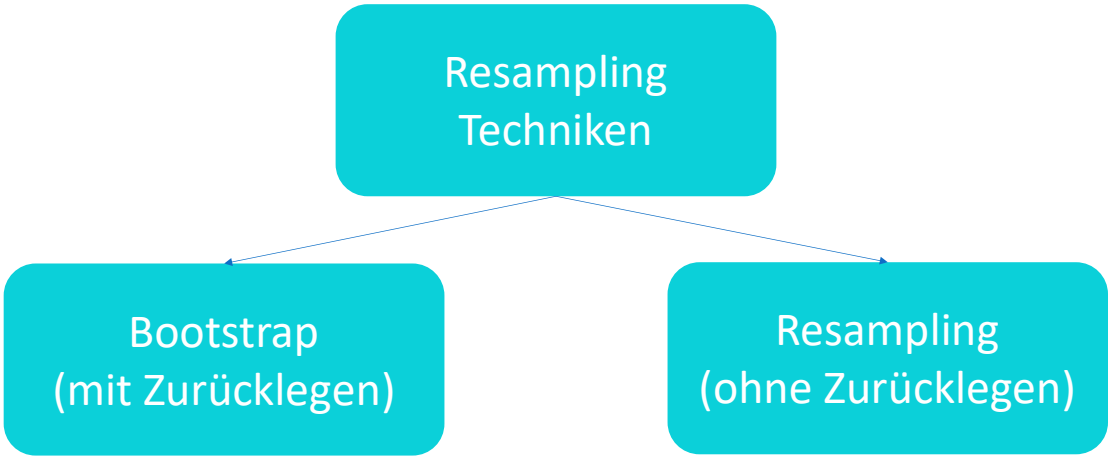


Beschreibung der gegebenen Daten, z. B. über Perzentile oder Standardabweichung

Figure 2-1. Population versus sample [Bruce et al. 2020]:

28

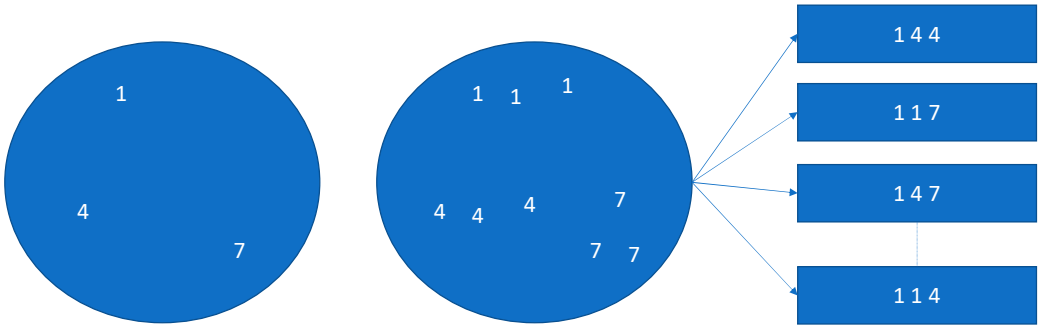
Anwendungen des Bootsraps im Data Science Kontext



29

Basic Bootstrap Theory

Diskussion: ohne Zurücklegen?



Originaldaten „Samples“

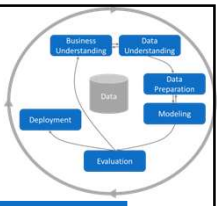
Originaldaten n mal repliziert

n „Resamples“

Nach [Bruce et al. 2020]:

30

Anwendungen von Bootstrap oder Resampling im Data Science Kontext



- Konfidenzintervalle für Stichproben (Data Understanding)
- Signifikanz und Größe eines Unterschiedes (Data Understanding)
- Konfidenzintervalle für Parameter in Verfahren selbst (Modeling) z. B. in linearer Regression
- Einsatz in Lernverfahren selbst (Modeling) z. B. in Random Forest
- Bestimmen des Einflusses von Modellparametern (Evaluation) z. B. über Feature Permutation
- Konfidenzintervalle für Bewertungsmetriken (Evaluation)

31

Boostrap und Konfidenzintervalle

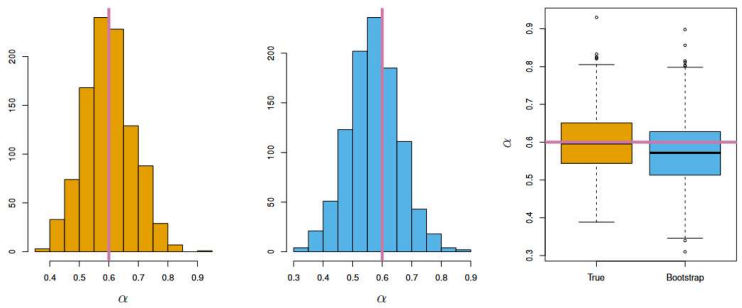


FIGURE 5.10. Left: *A histogram of the estimates of α obtained by generating 1,000 simulated data sets from the true population.* Center: *A histogram of the estimates of α obtained from 1,000 bootstrap samples from a single data set.* Right: *The estimates of α displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of α.*

Bildquelle: [James et al. 2013]

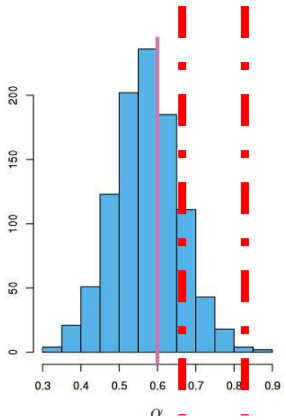
32

Permutationstest für Unterschiede

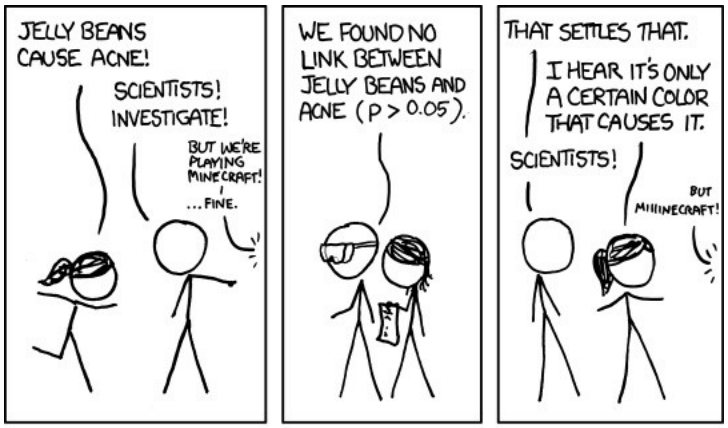
- Seite A: Verweildauer 4 sek im Durchschnitt
- Seite B: Verweildauer 5 sek im Durchschnitt
- Beobachteter Unterschied: 1 sek
- Anzahl Messungen: 30

- Frage: Signifikant?

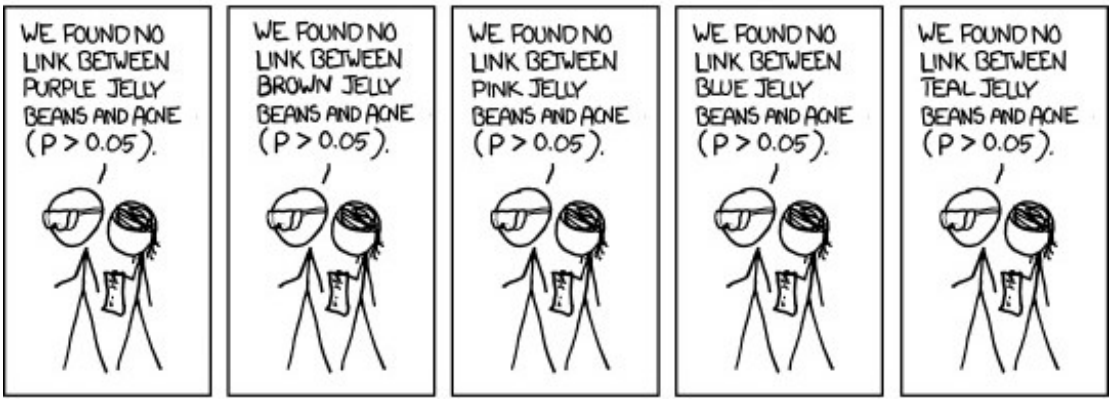
- Idee:
  - Resampling zufällig in den Seiten A und B ohne Zurücklegen
  - Unterschiede aufzeichnen (Histogramm)
  - Beobachteten Unterschied eintragen
- Beispiele links: **1 und 2 Diskussion**



Hypothesentests

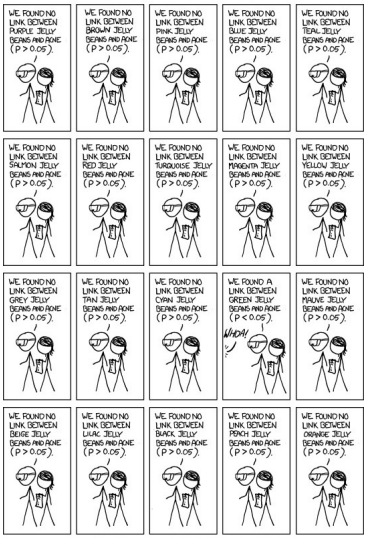


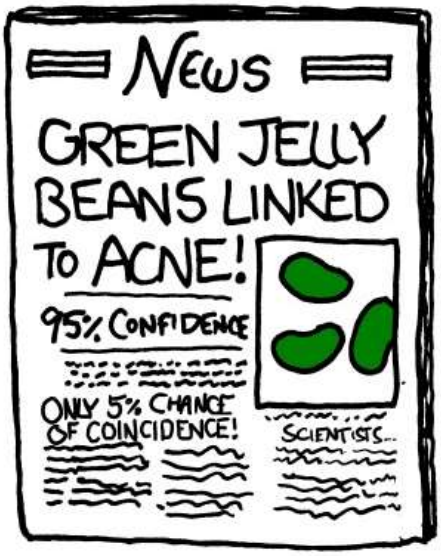
Hypothesentests



37

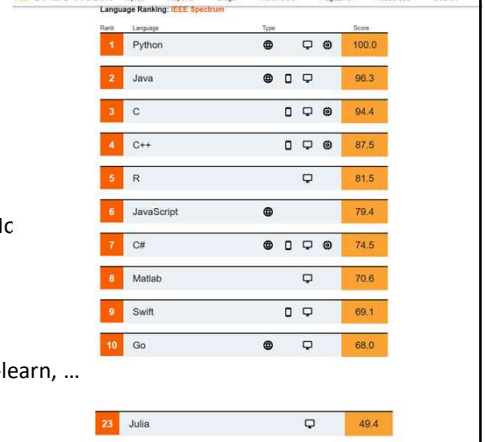
Hypothesentests





38

Python
IEEE Ranking 2019

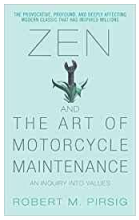


- Interpretierte Hochsprache
- Multiparadigmen-Sprache: objektorientiert, funktional und prozedural
- *„Batteries included“*
- Plattformunabhängig
- Dynamisch bzw. optional typisierte Sprache
- Speichermanagement/Garbage Collection
- Anfang der 90iger Jahre von Guido van Rossum entwickelt und nach Mc
- Version 2.0 ist ab April 2020 endlich endgültig Geschichte
- Aktuelleste Version 3.8.1 – 18.12.2019
- Gut für Glue-Code
- Wichtige Packages für Data Science: numpy, matplotlib, pandas, scikit-learn, …
- Data Engineer‘s sind eher im „Python-Lager“ zu finden …

Python
Zen of Python: The »Pythonic« way

- Beautiful is better than ugly.
- Explicit is better than implicit.
- Simple is better than complex.
- Complex is better than complicated.
- Flat is better than nested.
- Sparse is better than dense.
- **Readability** counts.
- Special cases aren't special enough to break the rules.
- Although practicality beats purity.
- Errors should never pass silently.
- Unless explicitly silenced.
- In the face of ambiguity, refuse the temptation to guess.
- **There should be one-- and preferably only one -- obvious way to do it.**
- Although that way may not be obvious at first unless you're Dutch.
- Now is better than never.
- Although never is often better than *right* now.
- If the implementation is hard to explain, it's a bad idea.
- If the implementation is easy to explain, it may be a good idea.
- **Namespaces** are one honking great idea -- let's do more of those!
  - **Was ist ein Namensraum?**
  - **C/C++/Java/JavaScript/Assembler**

http://legacy.python.org/dev/peps/pep-0020/



Data Science mit Python
Notwendige Pakete und Datenstrukturen

- **Scikit-learn** http://scikit-learn.org/stable/user_guide.html http://scikit-learn.org/stable/documentation
  - Aktuell **DAS** Paket für maschinelles Lernen
- NumPy (notwendig für `scikit-learn`)
  - Paket für wissenschaftliche Berechnungen
  - Besonders relevant: `ndarray` (NumPy Array)
- SciPy (ebenso)
  - Weitere Funktionen für wissenschaftliche Berechnungen
  - Besonders relevant: `scipy.sparse` (dünnbesetzte Matrix)
- **matplotlib**
  - Grafiken erstellen für wissenschaftliche Veröffentlichungen
  - Können durch `%matplotlib inline` (oder `%matplotlib notebook`) verwendet werden (Übung später)
- **pandas**
  - Für Datenvorbereitung, laden, bereinigen, etc.
  - `pandas DataFrame`: eine Art Tabellenblatt (analog Excel, R)
- IPython
  - interaktives Python
  - Skript-artiges Vorgehen
- **Jupyter**
  - Eine Unterstützung dafür (und viele andere Sprachen)
- Spyder
  - ausgewachsene Programmierumgebung
  - Analog z. B. Eclipse
- **Seaborn**
  - häufig erwähnte Bibliothek für schönere Charts
- Gibt auch Alternativen (wie immer☺)

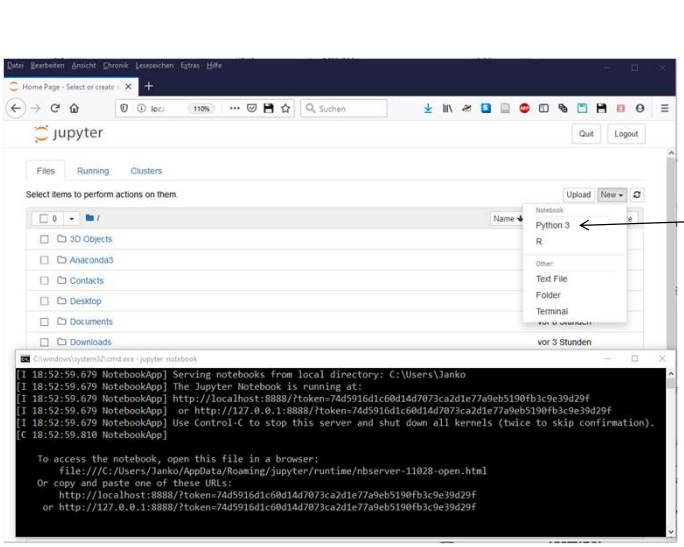
Aus: [Mueller and Guido 2017]

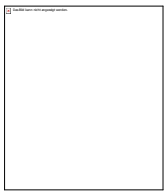
Python Installieren
Let's do it

- Versionsfrage: Uneinigkeit in der Data Science Community
  - Wir nehmen Python Version 3, WEIL wir in der Lehre sind und uns den Luxus »leisten« können und Sie noch sehr viel Zeit in der Praxis vor sich haben..
- **Installieren Sie Python**
  - Bequem: gleich die Version die in der Prüfungsleistung vorgegeben ist
- Starten Sie das Jupyter-Notebook (das reicht in der Regel aus)
- Benutzen Sie Python als Taschenrechner
- Versuchen Sie einige String-Operationen

seaborn

Python ausführen in Jupyter





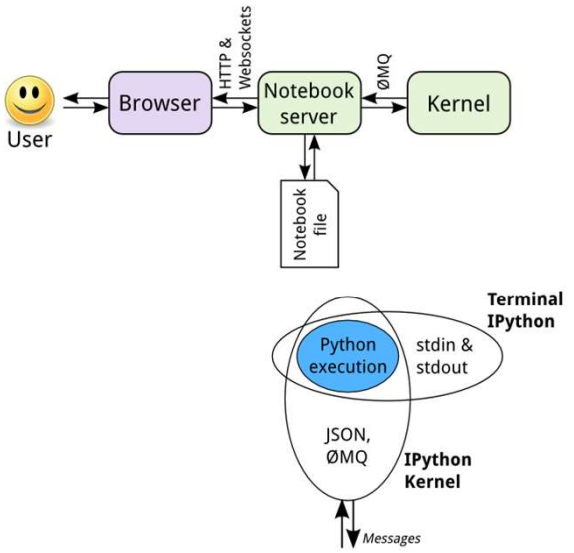
Python 3 Kernel

44

BACKGROUND: Was ist Jupyter?

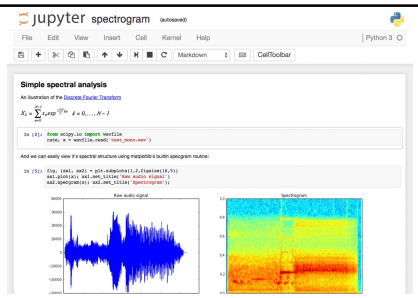
- IPython lange Zeit als REPL (Read-Eval-Print-Loop), d.h. als interaktive Python-Shell beliebt
- 2014 von Fernando Pérez als Spin-Off des IPython-Projekts – IPython-Notebooks
- Kommunikation des Frontends, wie z.B. Notebook oder QT-Konsole über JSON mit dem IPython-Kernel als Backend
- Möglichkeit viele verschiedene Kernel anzubinden, wie R..

Jupyter Documentation



45

Notebooks in Jupyter



Notebook-File:

- Enthält verschiedene Zellen
  - Code-Zellen für die Funktionalität
  - Markdown-Zellen für Kommentare und die Strukturierung des Dokuments
- JSON-basiertes Format
- Ergebnis (Output) einer Code-Zelle wird im Dokument abgelegt
- Extension ist „.ipynb“ (IPYthon-NoteBook)
- Wird nur im Notebook-Server „behandelt“, d.h. man kann Notebooks bearbeiten auch wenn man keinen passenden Kernel hat – hat dann aber keine Ergebnisse
- Potentielle Probleme:
  - Große Result-Datenmengen in Notebooks
  - Code-Versionierung – wird ein Notebook neu ausgeführt ohne eine Zeile an Code oder Markdown zu ändern, entsteht trotzdem ein Diff

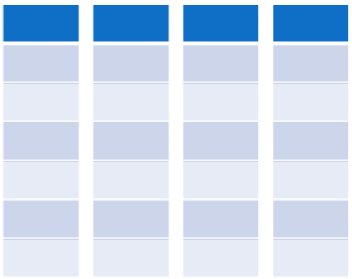
Vorteile von Jupyter

- Explorative Phase einer Datenanalyse:
  - Vieles muss ausprobiert werden
  - Schnelles Feedback, Ergebnisse sind schnell zu erkennen
- ***Literate Programming*** (Donald E. Knuth) – Reporting
  - Vorteil von Beschreibung und Code in einem Dokument
  - Sowieso müssen Ergebnisse eine Data-Science Projektes am Ende Entscheidungsträgern präsentiert und vermittelt werden – oder mindestens dokumentiert
- Reproduzierbare Wissenschaft bzw. Reproduzierbarkeit der Ergebnisse
  - „Ausführbares Dokument“
  - Erleichtert das Nachprüfen, Verständnis und das Aufbauen auf Ergebnissen

Pandas – Datenkonzept

Series: 1-Dimensionale Datenstruktur

Dataframe: 2-Dimensionale Datenstruktur

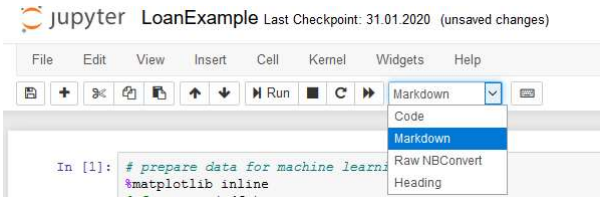


Hintergrund hierzu: https://databasecamp.de/python/pandas-series
https://www.python-kurs.eu/pandas_DataFrame.php

48

Grundlegendes Python
Wiederholung und Anmerkungen

- Zeilenumbrüche sind ohne besondere Zeichen möglich
- # für Code-Kommentare
- Markdown-Funktion (Esc für CommandMode, dann M für Markdown)
- Shortcuts sind wichtig: probieren Sie die rechts genannten aus
- Hinweis auf `%matplotlib inline` und `%matplotlib notebook`



Jupyter-Shortcuts
`Tab`
Autocomplete
`Shift + Tab`
Dokumentation
`Strg+Enter`
Ausführen
`Alt+Enter`
Ausführen+neu
`Strg+Shift+-`
Aufteilen
`;`
Unterdrücken (bei plt)
Run-all (unter Cells)

Command-Mode:
`Esc <-> Enter`
Mehrfachselektion
`Shift+J`
Merge
`Shift+M`
Delete
`D zweimal`

49

Literaturliste

- [James et al. 2013] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani: An introduction to statistical learning
  - Favorit: Sehr gut gemachte Einführung, jedoch Beispiele in R, verständlich mit Mathematik, als pdf frei erhältlich
- [Hastie et al. 2008] Trevor Hastie, Robert Tibshirani, Jerome Friedman: The elements of statistical learning
  - DIE Referenz, für Mathematiker geschrieben, als pdf frei erhältlich
- [O'Neil and Schutt 2013] Cathy O'Neil and Rachel Schutt: Doing Data Science
  - Spannend zu lesen, teilweise Erfahrungsberichte (durch Drittautoren)
- [Mueller and Guido 2017] Andreas C. Müller & Sasha Guido: An Introduction to Machine Learning with Python
  - Interessant da Python 3 tatsächlich genutzt wird für die Einführung inklusive der üblichen Bibliotheken
- [Grues 2016] Joel Grues (übersetzt von Kristian Rother): Einführung in Data Science
  - Auf deutsch gut übersetzt, nutzt Python für grundlegendes Verständnis ohne die üblichen Bibliotheken, extrem leicht lesbar
- [Alpaydin 2008]: Ethem Alpaydin (übersetzt von Simone linke): Maschinelles Lernen
  - Auf deutsch gut übersetzt, relativ viel Mathematik, in Deutschland scheint das weit verbreitet zu sein
- [Bruce et al. 2020]: Peter Bruce, Andrew Bruce, Peter Gedeck: Practical Statistics for Data Scientists
  - Das einzig wahre Statistikbuch was keines ist
- [Reinhart 2016]: Alex Reinhart (übersetzt von Knut Lorenzen): Statistics done wrong
  - Bevor man wirklich Konfidenzint4ervalle oder p-Werte angibt und über „Signifikanz" spricht, sollte man das gelesen haben