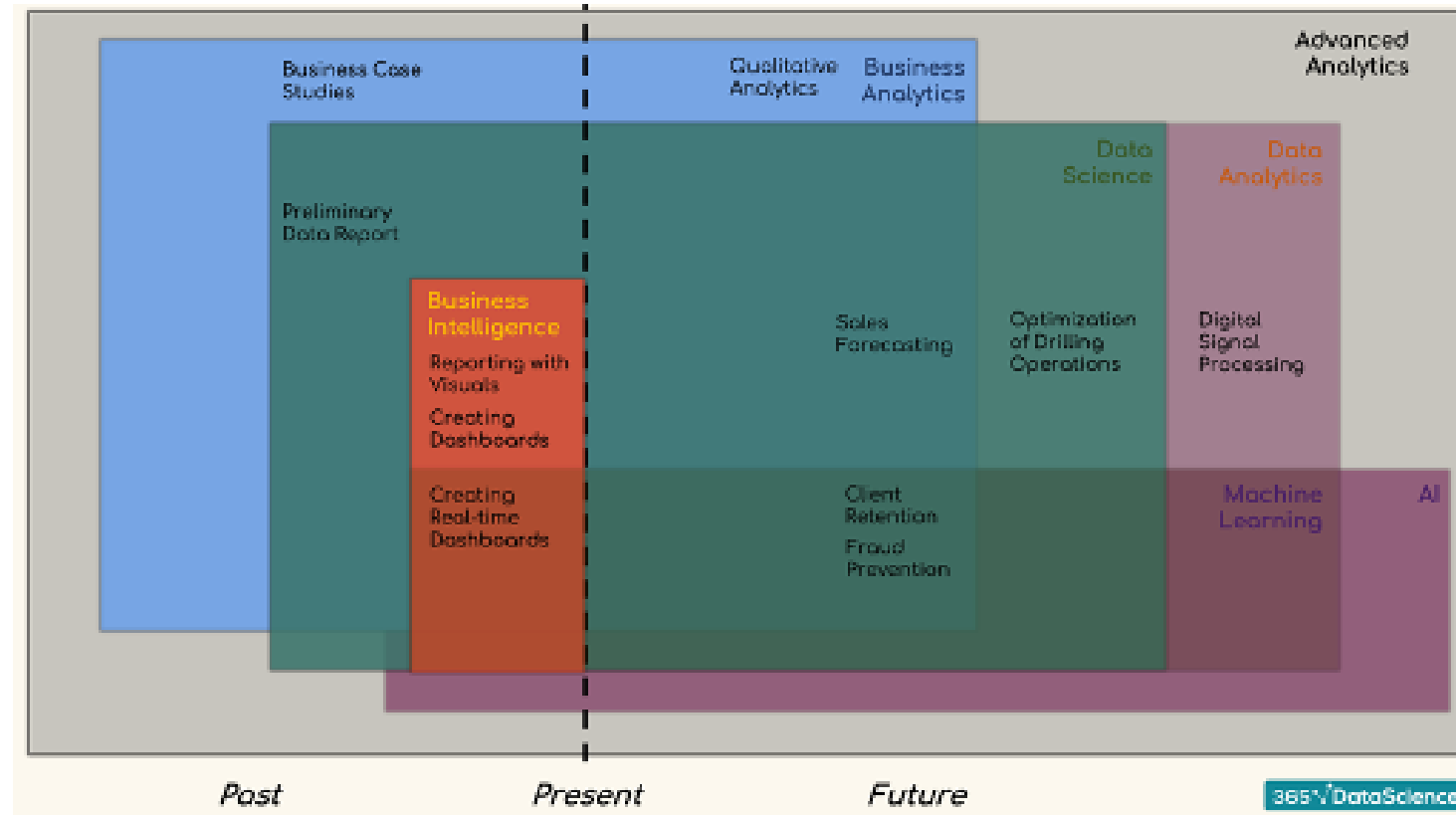


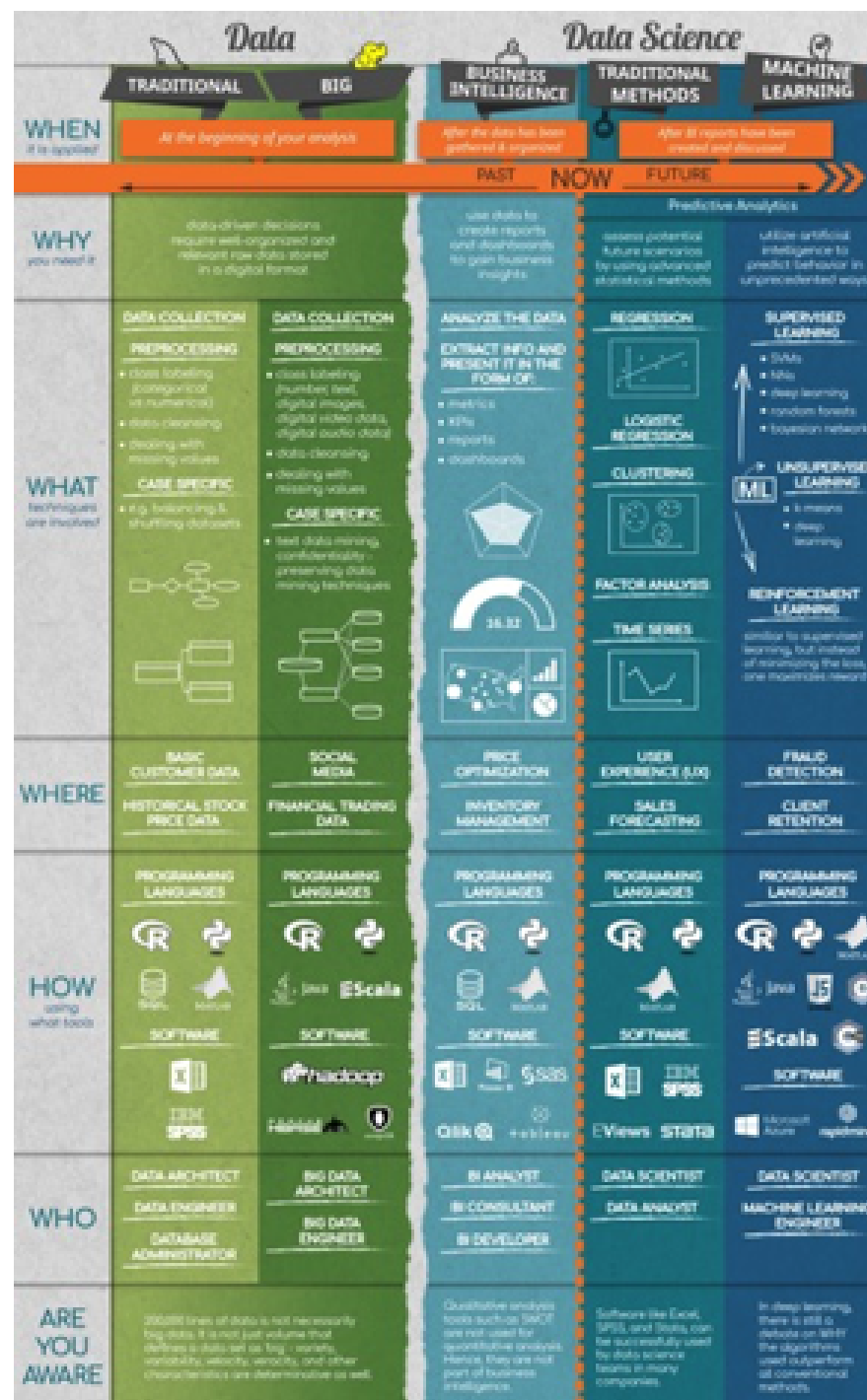
Visualisierung in Python

Visualisierungen

Datascience?



Datascience!



Unterschiede (DS / ML)

- Data Science
 - focuses on statistics and algorithms
 - unsupervised and supervised algorithms
 - regression and classification
 - interprets results
 - presents and communicates results
- Machine Learning
 - focus on software engineering and programming
 - automation
 - scaling
 - scheduling
 - incorporating model results into a table/warehouse/UI

(Grober) Ablauf

Steps in a full machine learning project



(Grober) Ablauf



Für eine Visualisierung kann man dies ebenfalls so angehen:

1. Problemdefinition
2. Daten
3. Evaluation (wie definiere ich Erfolg)
4. Features (was brauche ich genau)
5. Visualisieren
6. Testen (Experimentieren)

Visualisieren sollte ein iterativer Prozess sein

1. Problemdefinition
2. Daten
3. Evaluation (wie definiere ich Erfolg)
4. Features (was brauche ich genau)
5. Visualisieren
6. Testen (Experimentieren)

Ein Schritt zurück ist jederzeit aufgrund neuer Erkenntnisse möglich (und nicht unbedingt schlecht)!

Dogs-Science

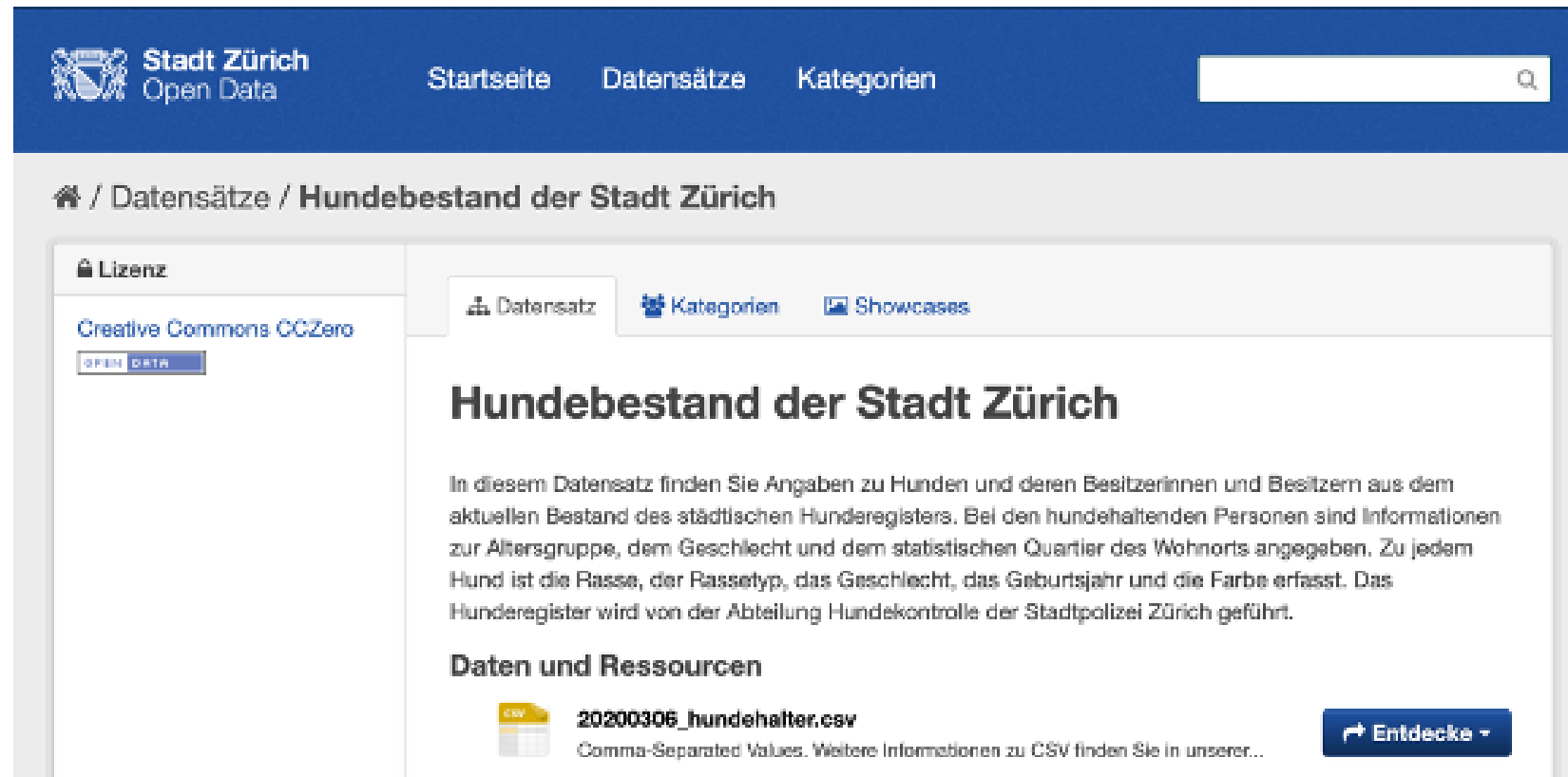
Die Trendmarke für Ihren Hund



Problemstellung (Problemdefinition)

- Dog-Science ist eine Trendmarke aus dem asiatischen Raum und möchte gerne nach Zürich expandieren.
- Im Heimatland von Dog-Science ist genau bekannt wo die meisten Hundehalter wohnen und welche Produkte bevorzugt werden. Die Schweiz und insbesondere Zürich ist aber unbekannt
- Das Budget von Dog-Science ist limitiert, die Stadt Zürich hat aber (fiktiv) Daten und Unterstützung für einen Pop-Up Store zugesichert
- Können wir die Ausgangslage von Dog-Science verbessern?

Ein Erster «Blick» in die Daten



The screenshot shows the 'Stadt Zürich Open Data' website. The header is dark blue with the city logo and name on the left, and navigation links 'Startseite', 'Datensätze', and 'Kategorien' in the center. A search bar is on the right. Below the header, a breadcrumb trail reads 'Home / Datensätze / Hundebestand der Stadt Zürich'. The main content area is divided into a left sidebar and a right main panel. The sidebar contains a 'Lizenz' section with 'Creative Commons CCZero' and an 'OPEN DATA' button. The main panel has tabs for 'Datensatz', 'Kategorien', and 'Showcases'. The 'Datensatz' tab is active, displaying the title 'Hundebestand der Stadt Zürich'. Below the title is a descriptive paragraph about the dog registry data. Further down is a section 'Daten und Ressourcen' featuring a CSV file icon, the filename '20200306_hundehalter.csv', a brief description, and a blue 'Entdecke' button with a right arrow.

Stadt Zürich
Open Data

[Startseite](#) [Datensätze](#) [Kategorien](#)

[Home](#) / [Datensätze](#) / **Hundebestand der Stadt Zürich**

Lizenz


Creative Commons CCZero

OPEN DATA

Hundebestand der Stadt Zürich

In diesem Datensatz finden Sie Angaben zu Hunden und deren Besitzerinnen und Besitzern aus dem aktuellen Bestand des städtischen Hunderegisters. Bei den hundehaltenden Personen sind Informationen zur Altersgruppe, dem Geschlecht und dem statistischen Quartier des Wohnorts angegeben. Zu jedem Hund ist die Rasse, der Rassetyp, das Geschlecht, das Geburtsjahr und die Farbe erfasst. Das Hunderegister wird von der Abteilung Hundekontrolle der Stadtpolizei Zürich geführt.

Daten und Ressourcen

 **20200306_hundehalter.csv**
Comma-Separated Values. Weitere Informationen zu CSV finden Sie in unserer...

[Entdecke](#)

Ein Erster «Blick» in die Daten

HALTER_ID	ALTER	GESCHLECHT	STADTKREIS	STADTQUARTIER	RASSE1	GEBURTSJAHR_HUND	GESCHLECHT_HUND	HUNDEFARBE
574	61-70	w	2	23	Mischling gross	2013	w	schwarz
695	41-50	m	6	63	Labrador Retriever	2012	w	braun
893	71-80	w	7	71	Mittelschnauzer	2010	w	schwarz
916	41-50	m	3	34	Mischling klein	2015	w	hellbraun
1177	51-60	m	10	102	Shih Tzu	2011	m	schwarz/weiss
4054	51-60	w	11	111	Lagotto Romagnolo	2016	w	weiss/beige
4135	41-50	w	9	91	Mischling klein	2016	w	schwarz
4206	71-80	w	8	82	Havanese	2016	w	hellbraun/weiss
4281	61-70	w	9	91	Chihuahua	2011	w	hellbraun
4388	61-70	w	11	115	Mops	2006	m	beige
4726	51-60	m	5	52	Mischling gross	2007	m	schwarz
4726	51-60	m	5	52	Golden Retriever	2013	w	creme
4747	61-70	m	2	24	Chihuahua	2013	m	weiss/braun
4850	51-60	m	4	42	Chihuahua	2013	w	beige
4862	51-60	m	4	42	Mops	2006	m	braun
5040	61-70	m	10	102	Labrador Retriever	2016	w	gelb
5088	61-70	m	7	72	Labrador Retriever	2014	w	schwarz
5113	61-70	m	11	119	Beagle	2010	w	tricolor
5113	61-70	m	11	119	Chihuahua	2017	w	beige
5225	71-80	m	3	34	Lagotto Romagnolo	2007	m	braun
5227	71-80	m	10	101	Border Terrier	2011	w	tricolor

Ein Erster «Blick» in die Daten

- Die Daten haben eine gute aber nicht perfekte Qualität
- Die Rassen wurden leider sehr ungenau definiert
 - Oft als «Mischling gross» oder «Mischling klein» definiert aber nicht weiter ausdefiniert
 - Oft vertreten
 - Chihuahua 573
 - Labrador Retriever 426
 - Selten vertreten
 - Oesterreichischer Pinscher 1
 - Daisy-Dog 1

Ein Erster «Blick» in die Daten

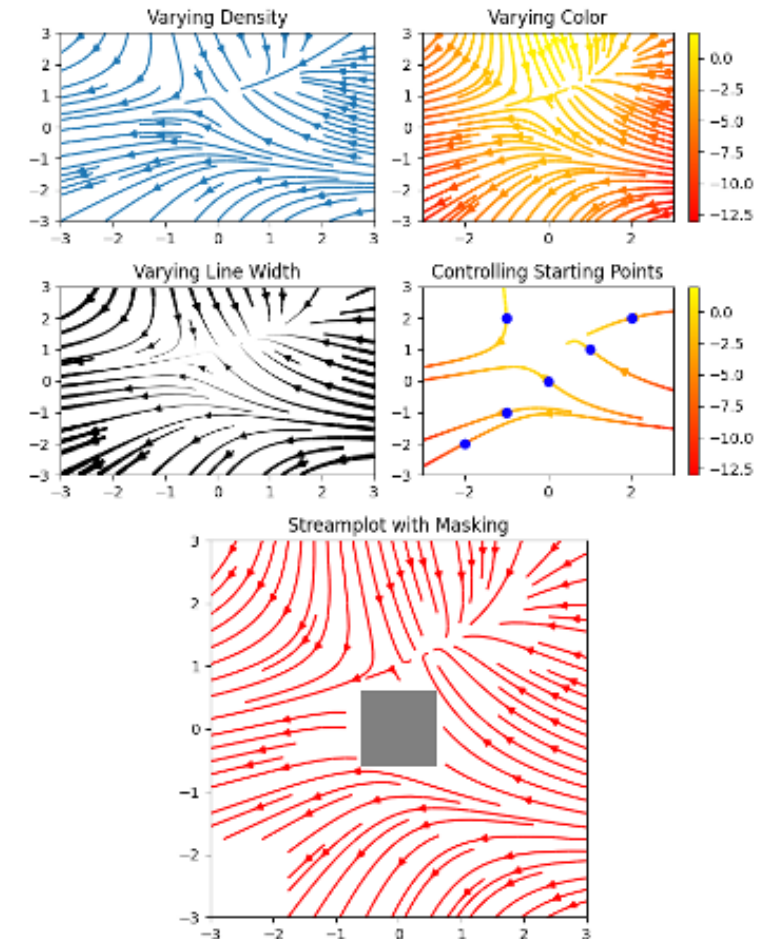
- Deutlich mehr weibliche als männliche Hunde (5402 zu 2439)
- Bei den Fell-Farben ist schwarz am meisten vertreten (800), danach tricolor (725) und weiss 634), seltener «schwarz melliert» oder «gelb / schwarz) je ein mal
- Das Hundesalter schauen wir uns in den Visualisierungen an (bitte Ausreisser beachten)
- Wie ist die Hundehalter / Hund Ratio?
 - 6562 Hundehalter haben einen gemeldeten Hund
 - 581 Hundehalter haben 2 oder mehrere gemeldete Hunde
 - Der Top-Hundehalter besitzt 14 gemeldete Hunde

Visualisierung

- Viele Libraries mit unterschiedlichen Spezialitäten
 - Matplotlib: Für statische, animierte und interaktive Visualisierungen (eine der ältesten Bibliotheken)
 - Pygal: Dynamische SVG-Charting Bibliothek
 - Seaborn: Basiert auf Matplotlib und bietet ein high-level Interface für statistische Grafiken
 - Altair: Basiert auf Vega/Vega Lite und ist eine deklarative statistische Visualisierungs-Bibliothek
 - Ggplot2: System zur Erstellung von deklarativen Grafiken
 - Plotly: Interaktiv und vom User analysierbar
 - Bokeh: Bibliothek für interaktive Visualisierungen
 - Geoplotlib: Hauptsächlich für Karten

Matplotlib

- Geeignet für einfache und komplexe Darstellungen
- Mehrere Darstellungen via Subplots möglich
- Bietet das grösste und allgemeinste Spektrum



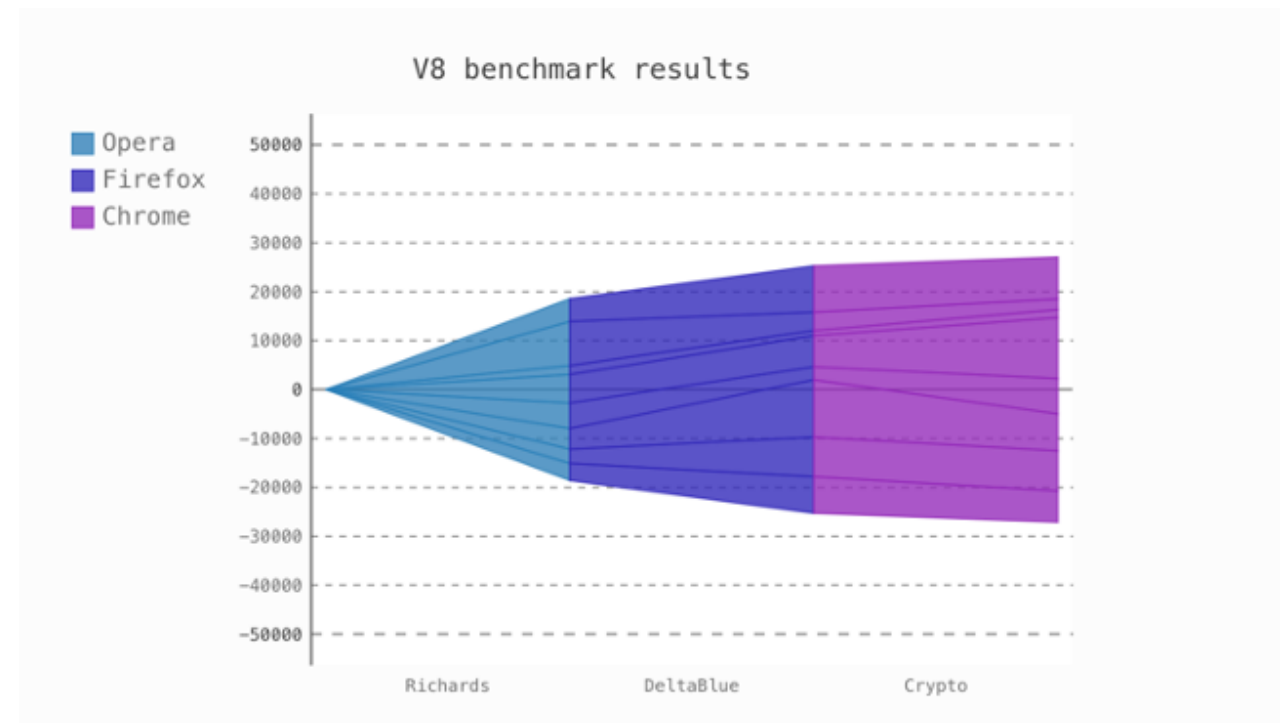
Streamplot with various plotting options.

Pygal

- Wird leider nicht mehr (aktiv) weiterentwickelt

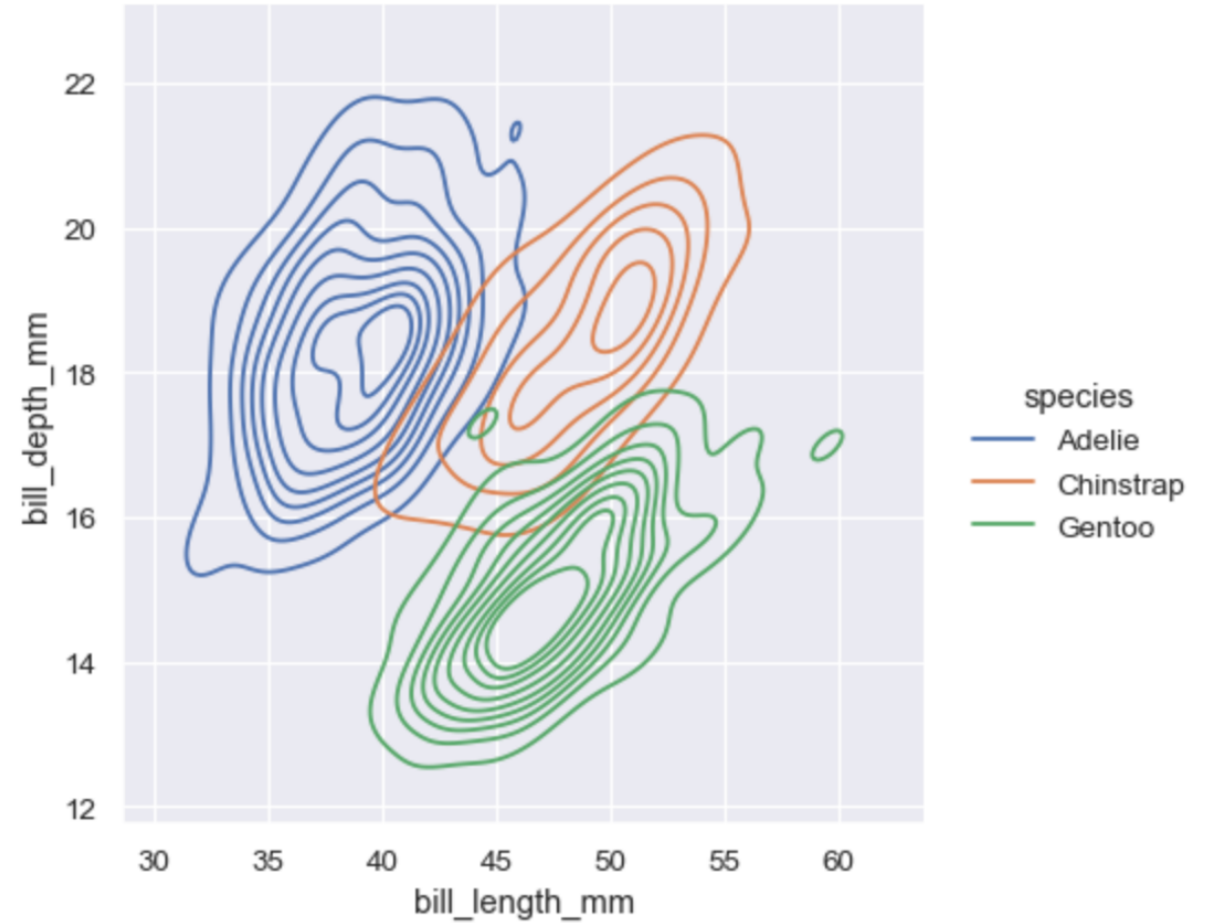
Guter Funktionsumfang und dabei relativ einfach gelernt

Einfach als interaktives HTML auszugeben



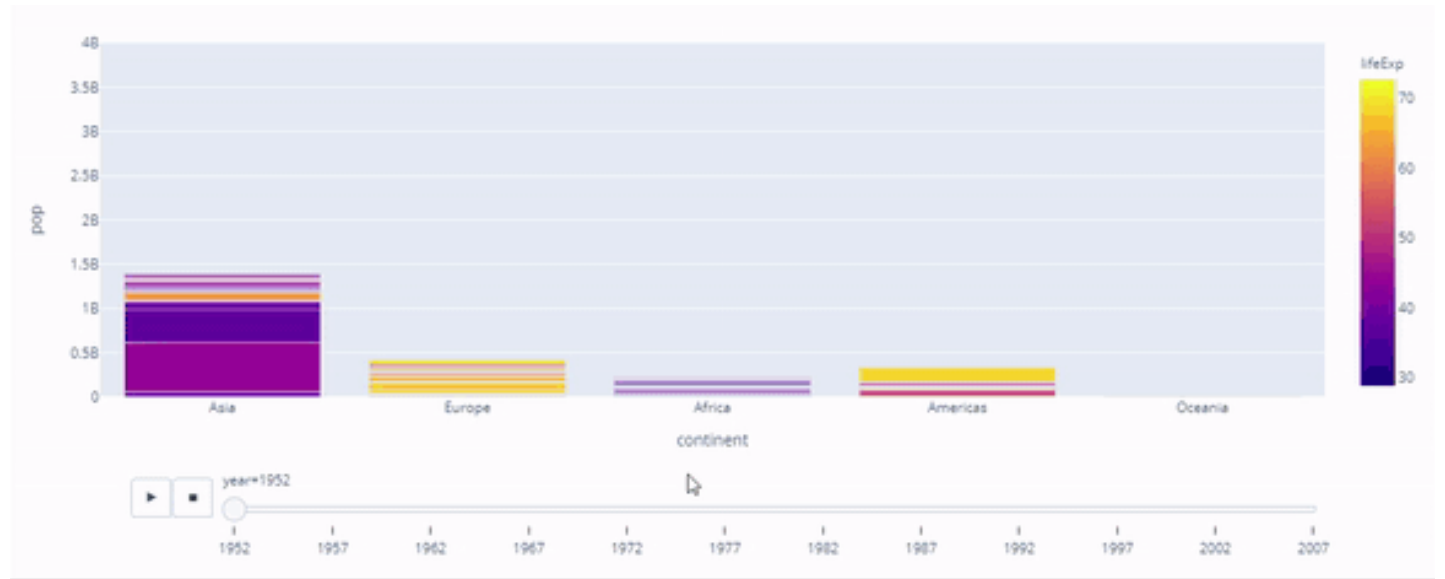
Seaborn

- Basiert auf Matplotlib
- Vor allem für statistische Visualisierungen geeignet
- Wird unser erstes Beispiel sein



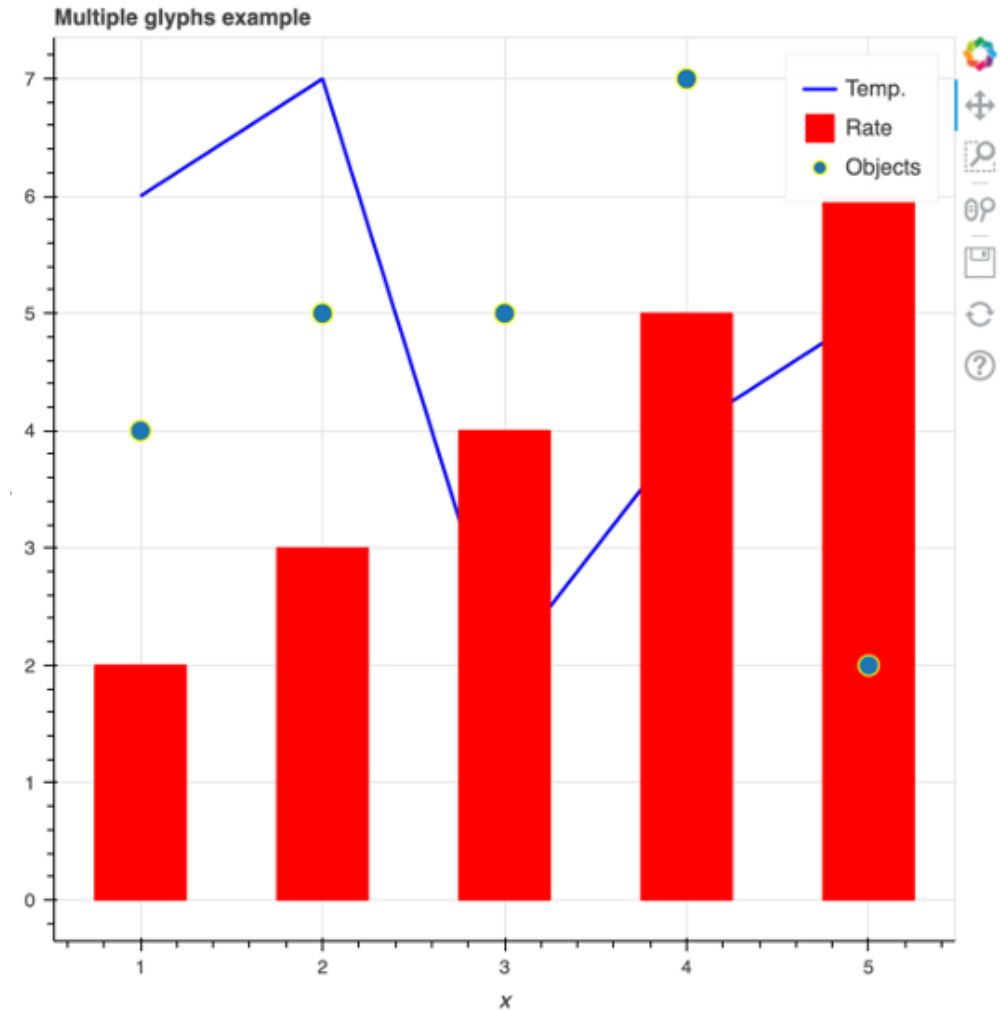
Plotly

- Bietet eine grosse Bandbreite an Charts
- Lässt sich animieren
- Besitzt «out of the box» Manipulationstools



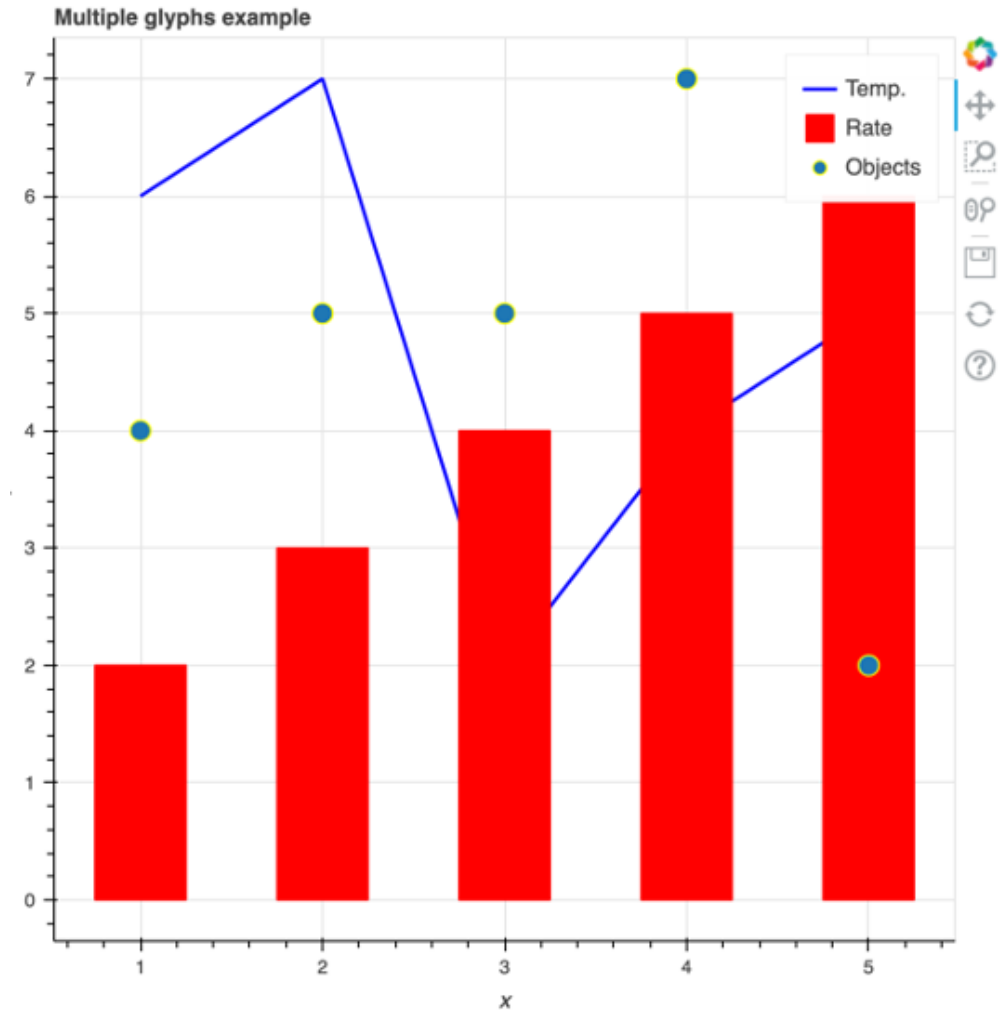
Bokeh

- Bietet eine grosse Bandbreite an interaktiven Charts
- Besitzt «out of the box» Manipulationstools



Bokeh

- Bietet eine grosse Bandbreite an interaktiven Charts
- Besitzt «out of the box» Manipulationstools

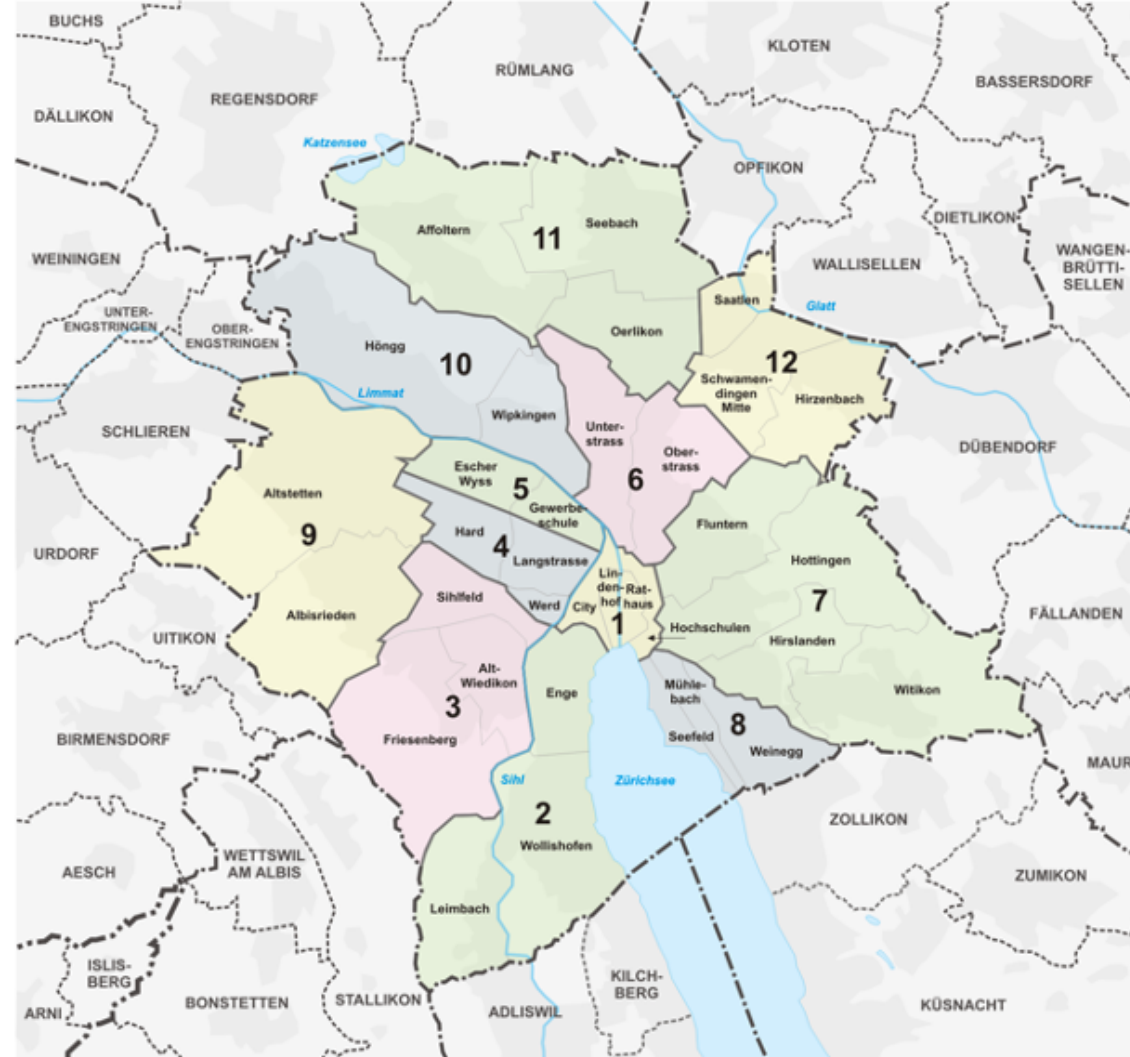


Der Ort ist wichtig...



Der Ort ist wichtig...

Platz 3: 984



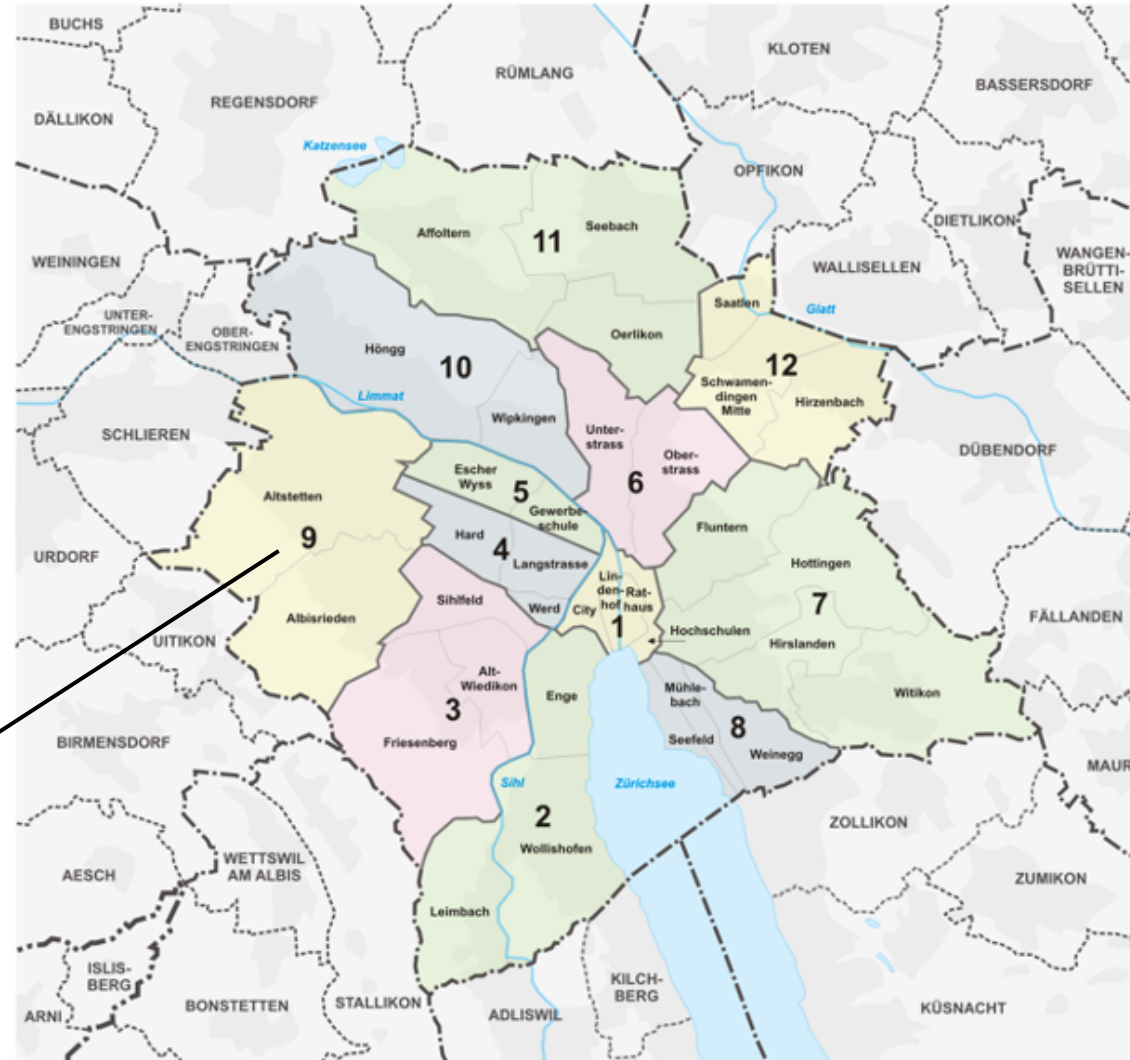
Der Ort ist wichtig...

Platz 3: 984



Der Ort ist wichtig...

Platz 3: 984



Platz 2: 1087

Der Ort ist wichtig...

Platz 3: 984



Platz 2: 1087

Der Ort ist wichtig...



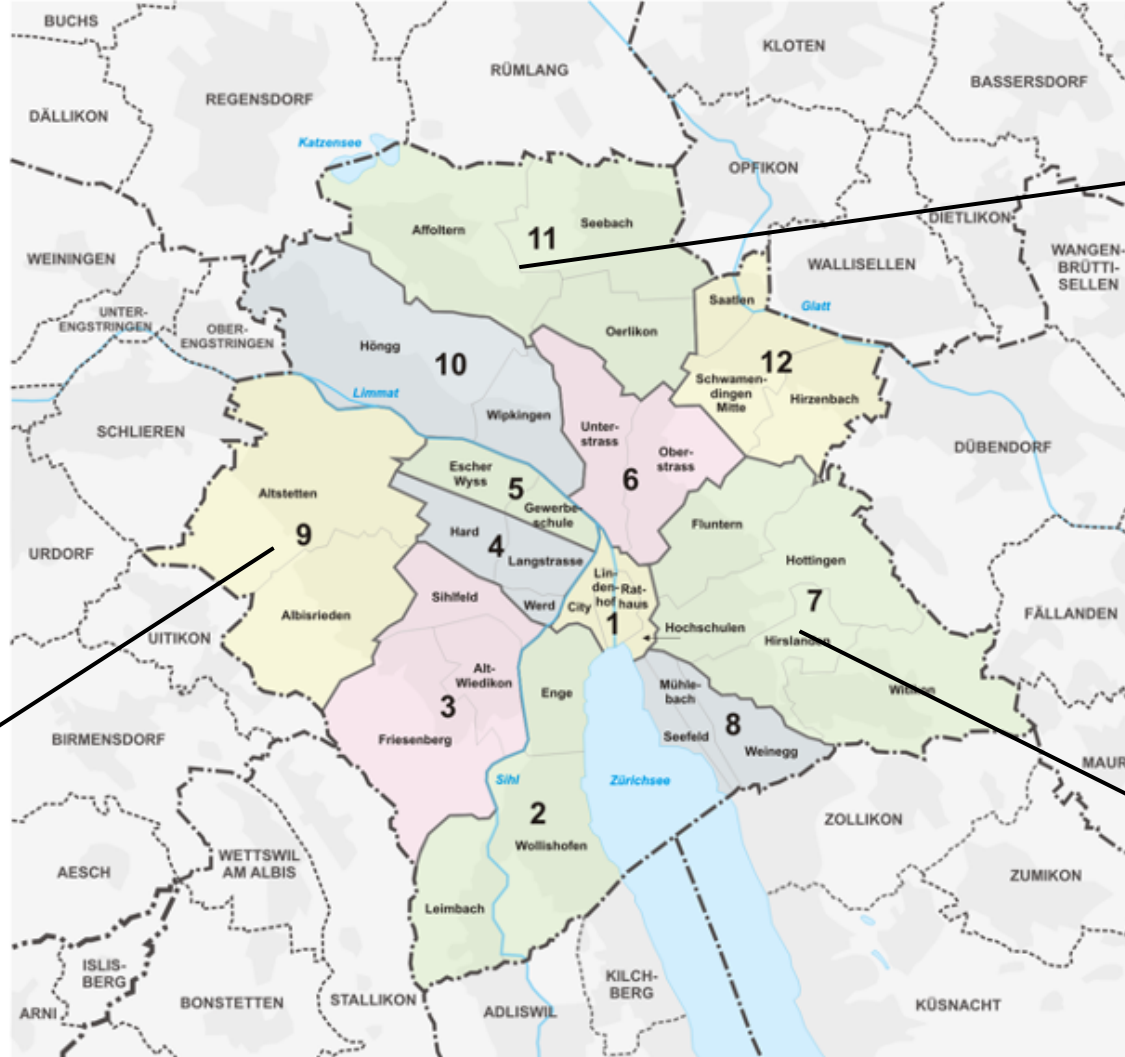
Platz 1: 1352

Platz 3: 984

Platz 2: 1087

Der Ort ist wichtig...

Platz 3: 984

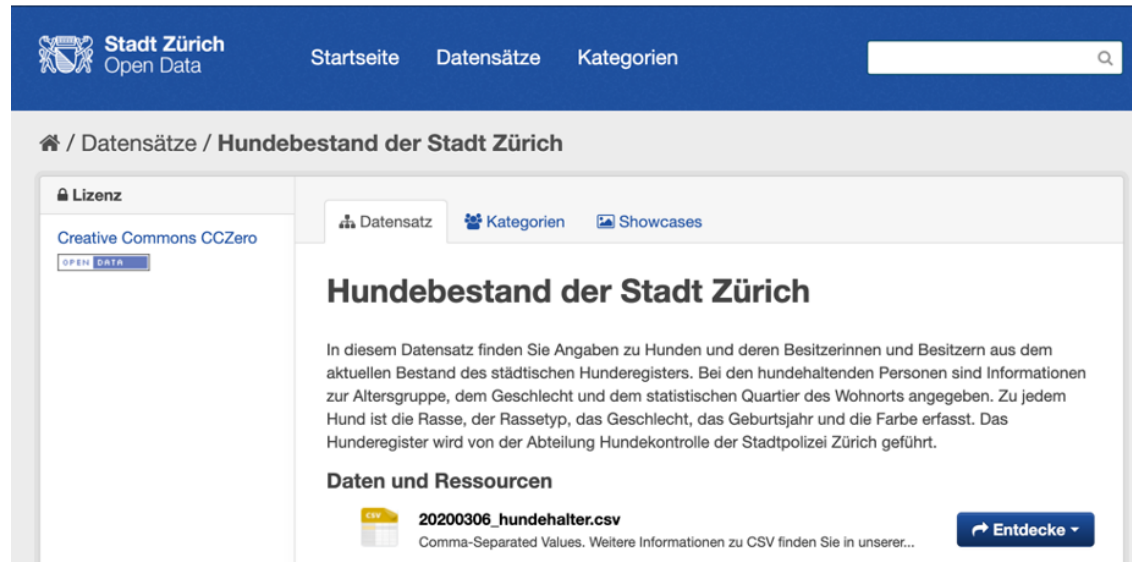


Platz 1: 1352

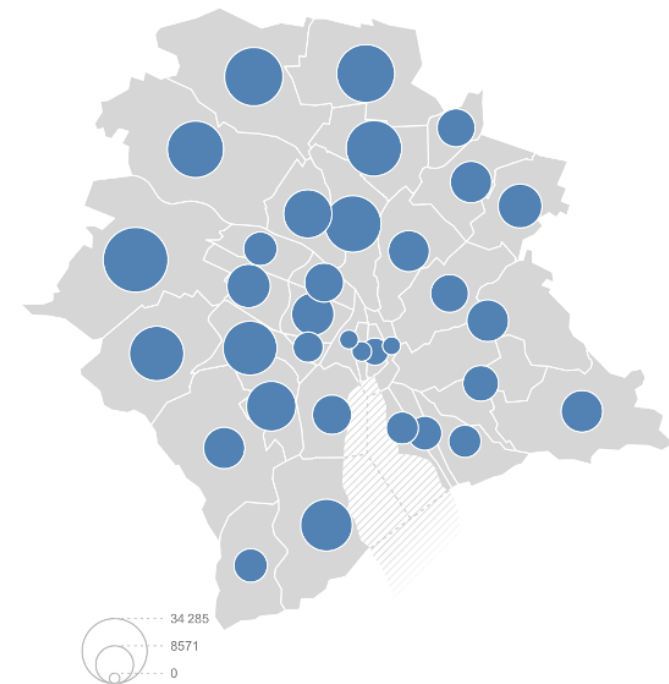
Platz 2: 1087

Der Ort ist wichtig...

Bevölkerung nach Stadtquartier



The screenshot shows the 'Stadt Zürich Open Data' portal. The header includes the city logo, 'Startseite', 'Datensätze', and 'Kategorien' links, along with a search bar. The main content area is titled 'Hundebestand der Stadt Zürich' and includes a description of the dataset, which contains information about dogs and their owners. A sidebar on the left shows the 'Lizenz' (Creative Commons CCZero) and 'OPEN DATA' status. A 'Daten und Ressourcen' section at the bottom lists the dataset '20200306_hundehalter.csv' with a download button and a description of the CSV format.

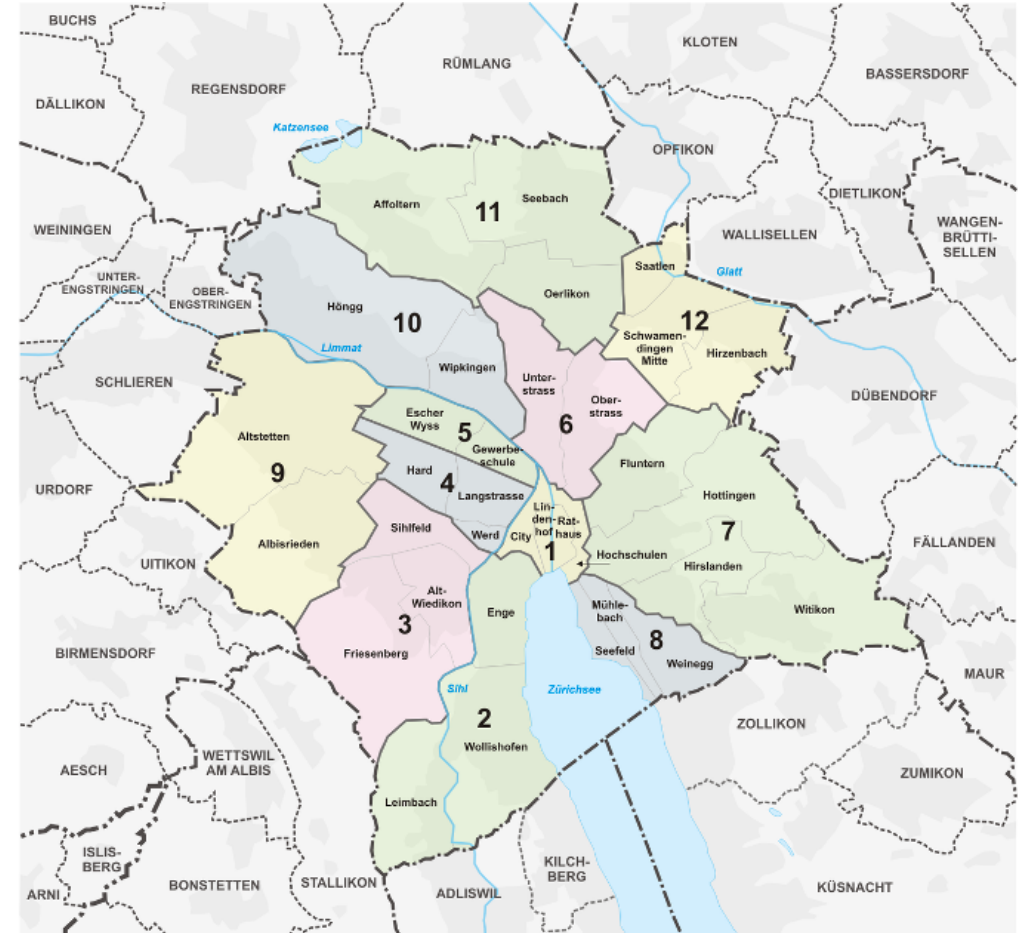
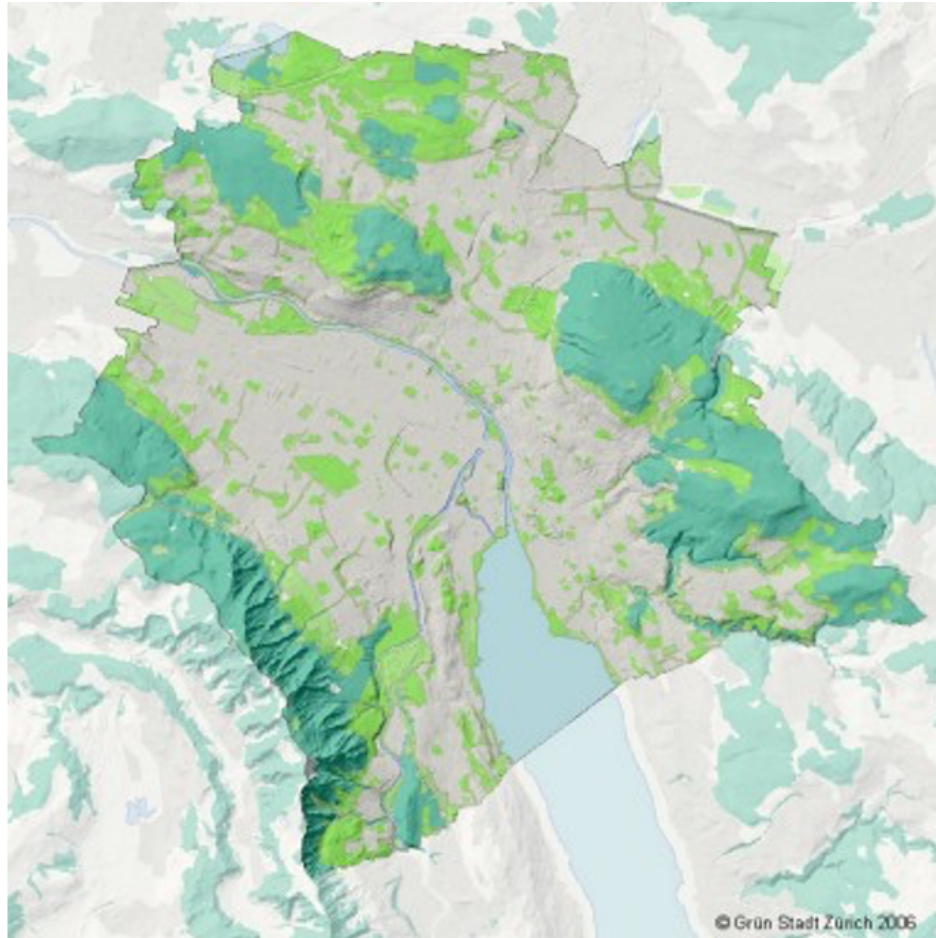


Hier besonders auf zeitliche Unterschiede achten:
Hundedaten: März 2020 / Bevölkerungsdaten: 1993-2020

Ist die Anzahl Personen wichtig?

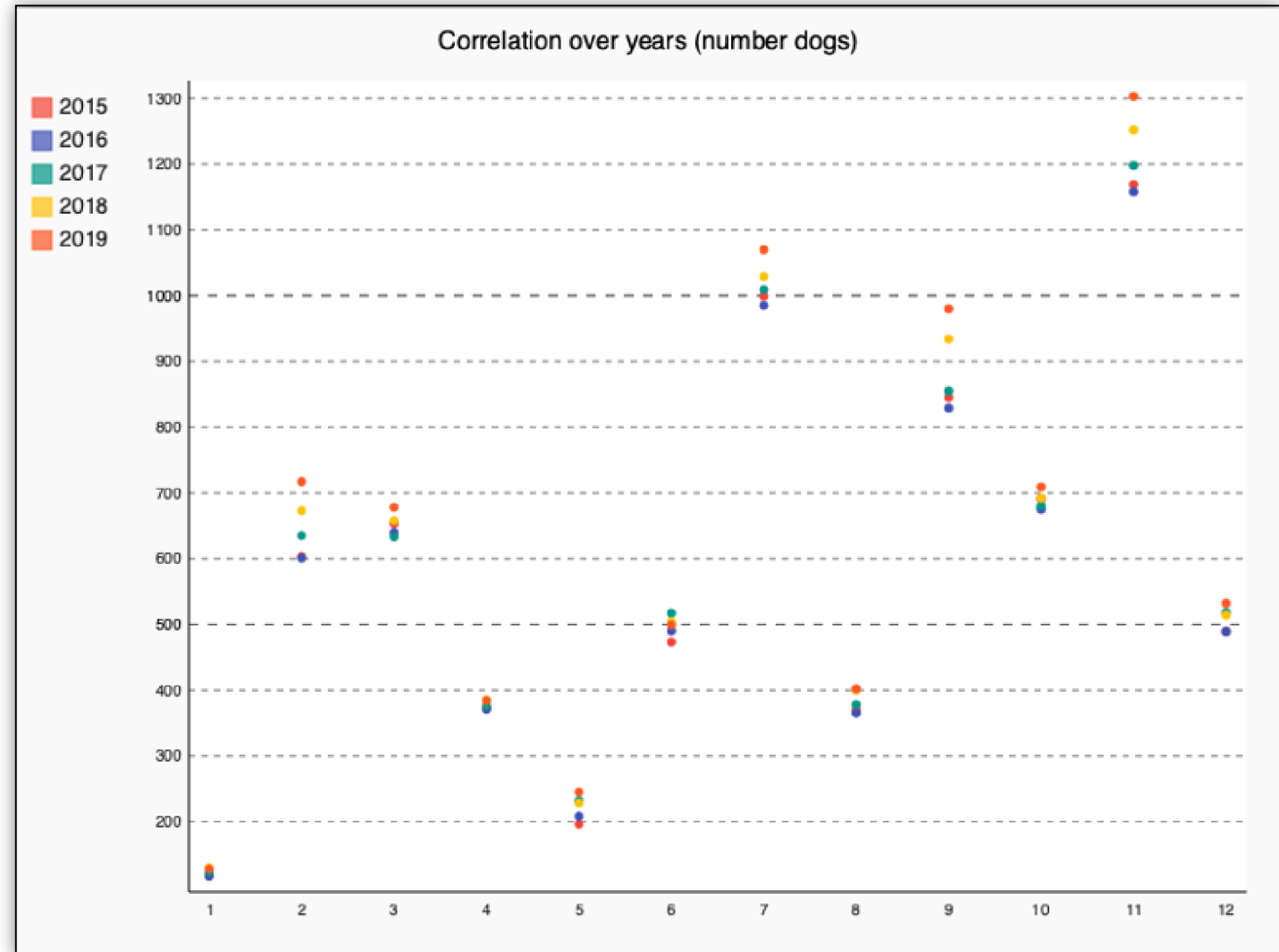
- Kreis 1; 5 831
- Kreis 2; 35 552
- Kreis 3; 50 756 => viele Menschen, weniger Hunde (697)
- Kreis 4; 29 034
- Kreis 5; 15 622
- Kreis 6; 35 317
- Kreis 7; 38 629
- Kreis 8; 17 456
- Kreis 9; 56 462
- Kreis 10; 41 044
- Kreis 11; 76 188
- Kreis 12; 32 845

Sind Grünflächen wichtig?



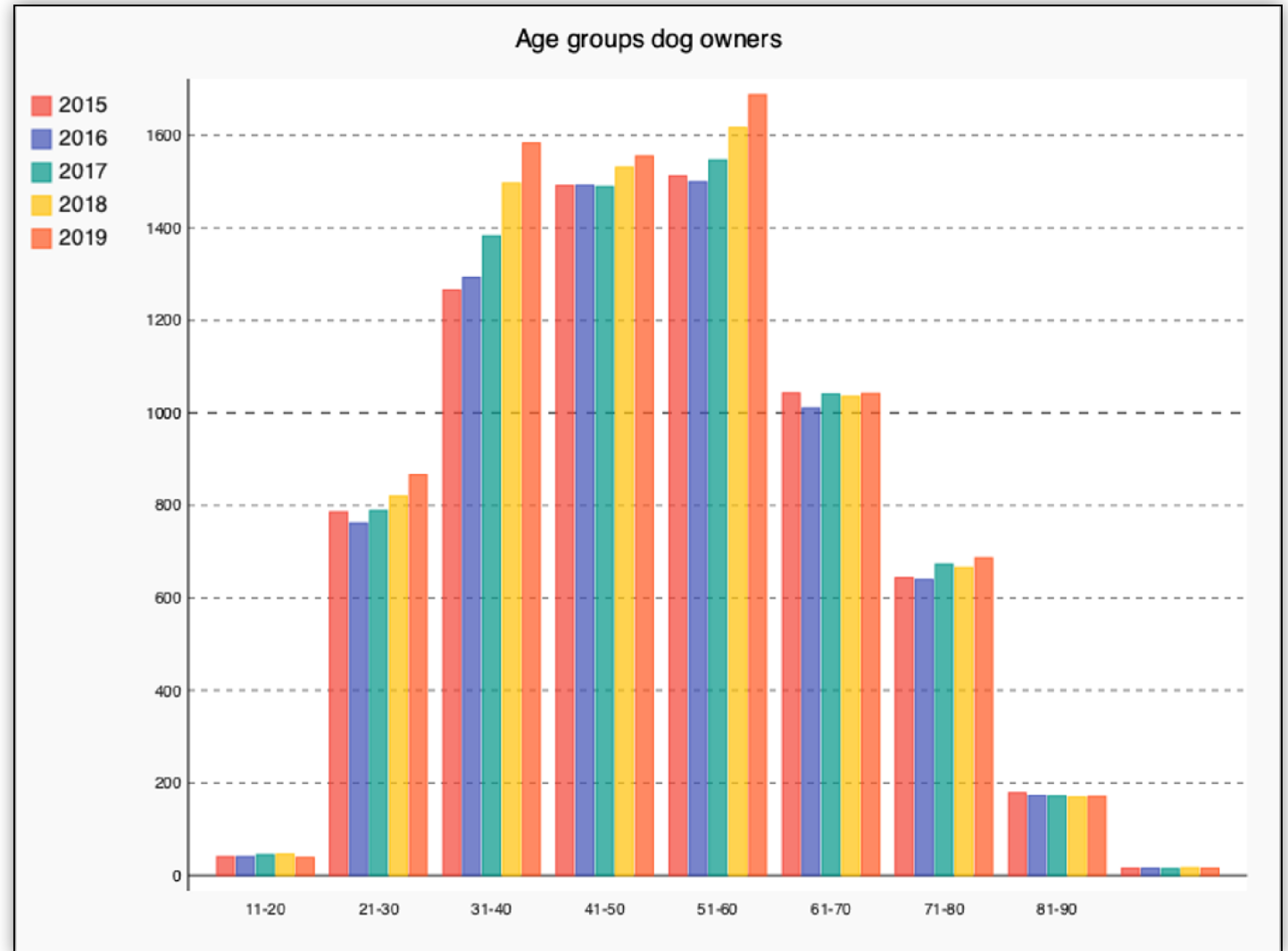
Korrelation berechnen

Siehe Step 2 in den Unterlagen



Altersgruppen

Am stärksten Ausgeprägt
ist die Gruppe von 31-60



Wie weiter?

- Selbstverständlich könnte man hier nun noch weite mehr Verfahren anwenden:
 - K-Nearest Neighbour
 - Naive Bayes
 - Lineare Regression
 - Multiple Regression
 - Logistische Regression
 - Decision-Trees
 - Neuronal-Networks / Deep-Learning
 - Clustering
 - usw.

Visualisieren

- Da für uns aber eher das Visualisieren im Zentrum steht, sollten wir uns die Schritte nochmals vor Augen führen.

(Grober) Ablauf



Für eine Visualisierung kann man dies ebenfalls so angehen:

1. Problemdefinition
2. Daten
3. Evaluation (wie definiere ich Erfolg)
4. Features (was brauche ich genau)
5. Visualisieren
6. Testen (Experimentieren)

Abschluss: Was empfehlen wir Dog-Science?

- Wir empfehlen Dog-Science einen Platz **im Kreis 11** in der **Nähe einer Grünfläche zu beantragen**
- Wir empfehlen das Produkt für 31-61 Jährige Personen zu gestalten
- Es darf etwas kosten (siehe Durchschnittsgehalt Zürich)
- Es darf nicht zu verspielt sein und sollte eher einfach verständlich sein (Interpretation der Altersgruppe)

Ende

Das war alles für dieses Kapitel
