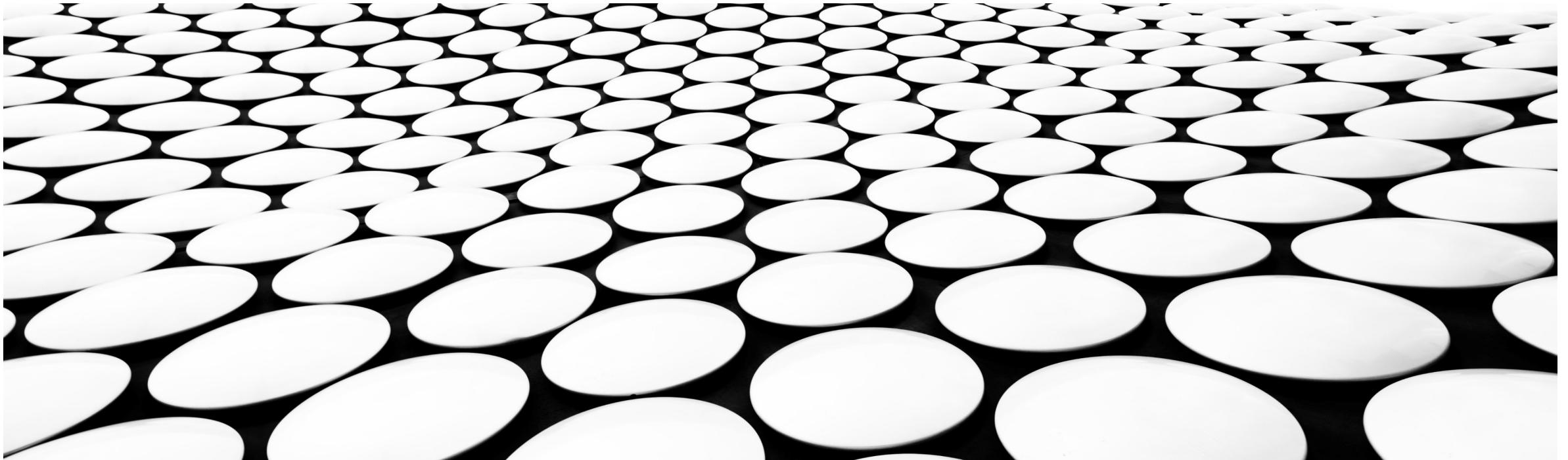

HANDS ON DATASCIENCE

A «SHORT» INTRODUCTION



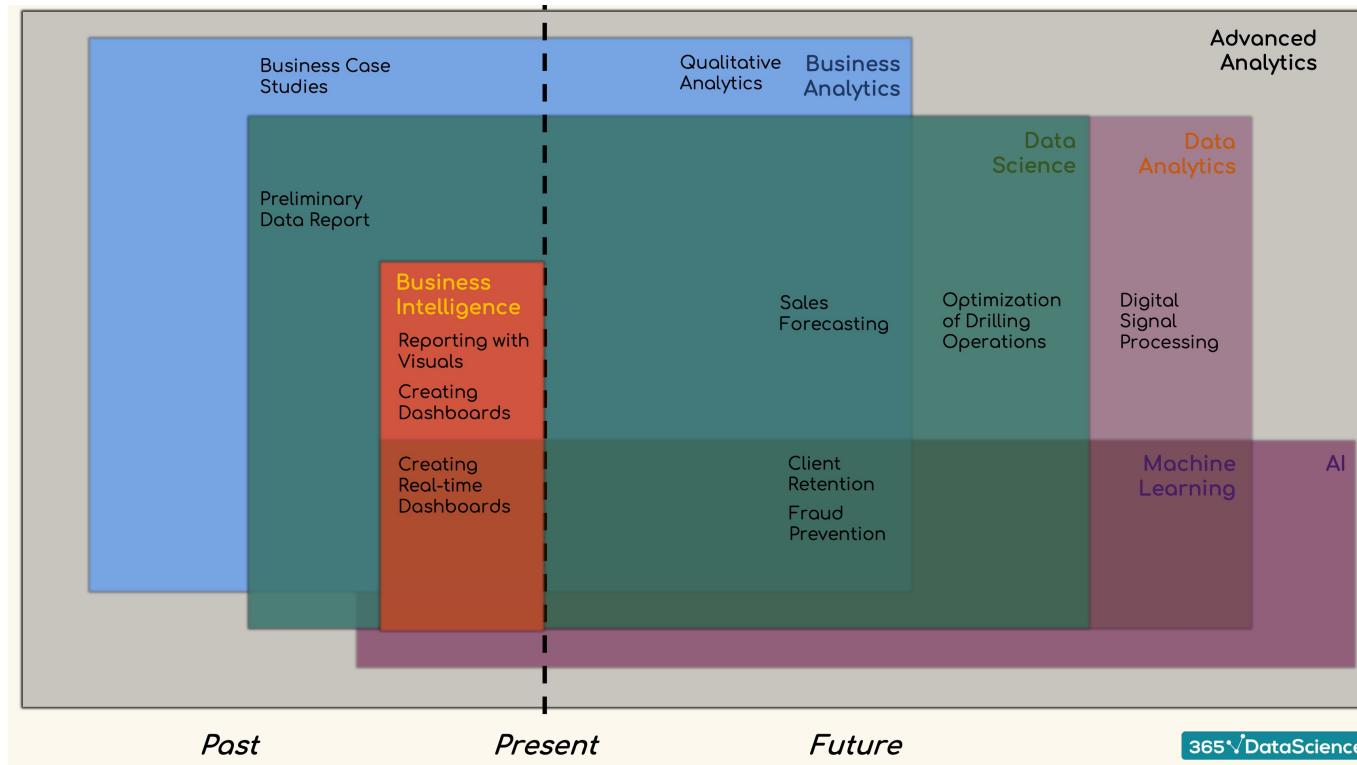
DAVID PINEZICH

- **Ausbildung**
 - Informatiker Applikationsentwicklung EFZ (BMS / Passerelle)
 - Bachelor of Informatics at UZH
 - Master of Informatics at UZH
 - *Lehrdiplom für Maturitätsschulen (Informatik)*
- **Arbeitserfahrung**
 - Paul Scherrer Institut (PSI)
 - Architonic
 - ti&m
 - Helsana (Fachlicher Leiter Webentwicklung)
- david.pinezich@gmail.com / david.pinezich@uzh.ch

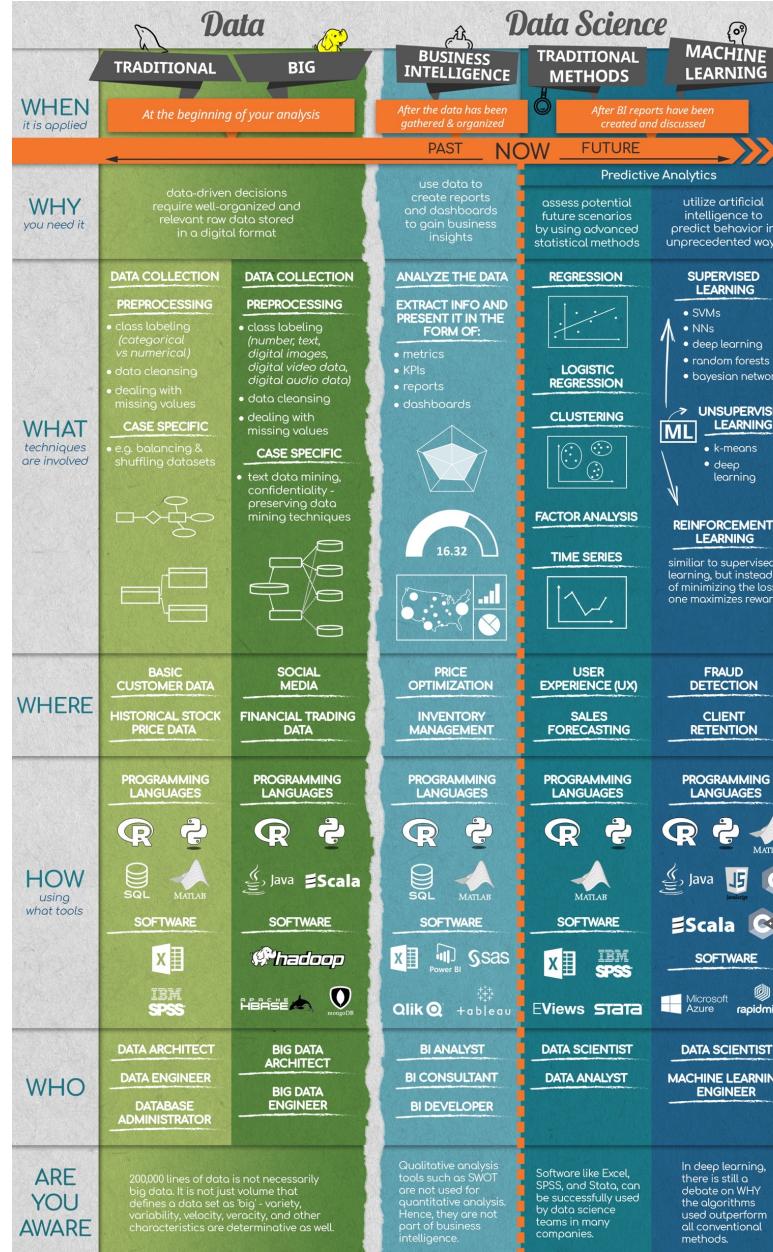
TERMINPLAN

- Tag 1
 - Session 1: Einführung und Ablauf erklären / Datascience-Prozess
 - Session 2: Einführung GIT & Visualisierung / Gruppen / Dataset auswählen / «Frage» bestimmen
 - Session 3: «Frage» finalisieren / Arbeit beginnen
 - Session 4: Weiterarbeit Teilnehmer
- Tag 2
 - Session 5: Breakout-Präsentation / Weiterarbeit
 - Session 6: Weiterarbeit Teilnehmer
 - Session 7: «Portfolio» / Abschlussarbeiten

DATASCIENCE?



DATASCIENCE!

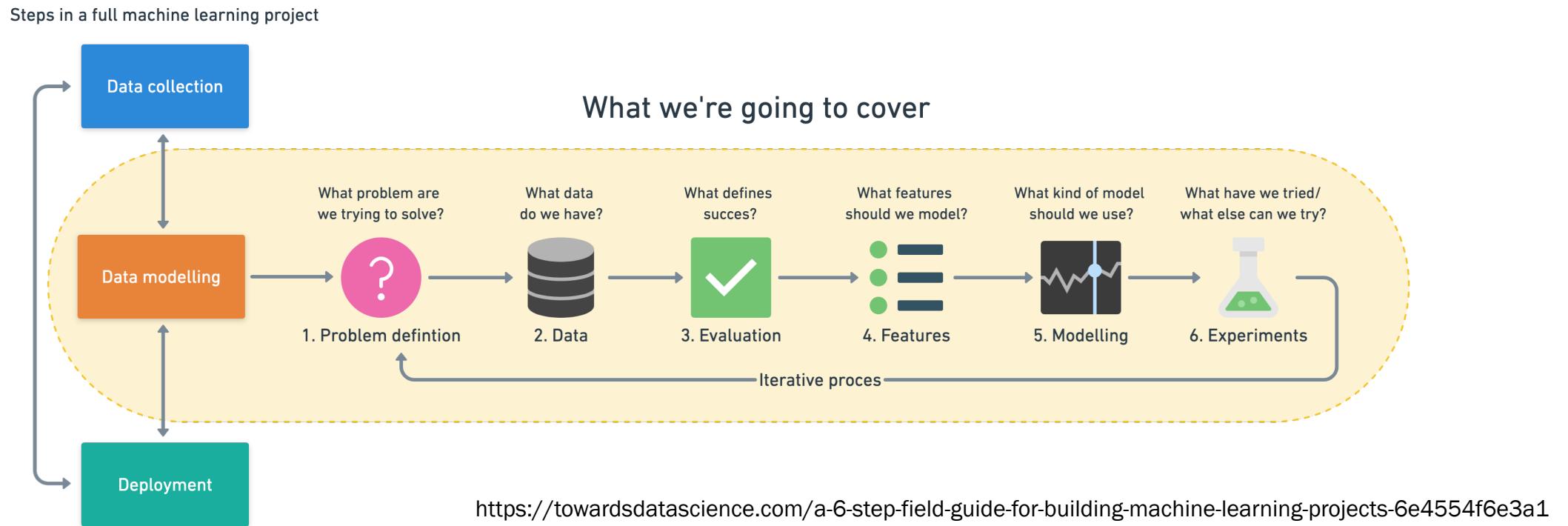


DIFFERENCES (DS / ML)

- 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

<https://towardsdatascience.com/data-science-vs-machine-learning-heres-the-difference-530883d6de3a>

(GROBER) ABLAUF



DATASCIENCE (UND ML) IST EIN ITERATIVER PROZESS

- Schritt 1: Problemdefinition
- Schritt 2: Daten sichten / Reinigung / Transformation / erste Darstellungen
- Schritt 3: Das Problem weiter eingrenzen, beschreiben was «Erfolg» bedeutet
- Schritt 4: Welche Features sollen modelliert werden (und wie)
- Schritt 5+6: Modellierung + Experimente

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

DOG-SCIENCE

DIE TRENDMARKE FÜR IHREN HUND



<https://i0.wp.com/cosmodoggyland.com/wp-content/uploads/2018/10/jamie-street-804224-unsplash-773145922-1539500541636.jpg?fit=3586%2C3100&ssl=1>

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 a screenshot of the Stadt Zürich Open Data website. The header is blue with the logo 'Stadt Zürich Open Data'. Below the header, there is a breadcrumb navigation: 'Startseite / Datensätze / Hundebestand der Stadt Zürich'. The main content area has a title 'Hundebestand der Stadt Zürich' and a detailed description of the dataset. On the left, there is a sidebar with a license section showing 'Creative Commons CCZero' and an 'OPEN DATA' button. At the bottom, there is a download link for '20200306_hundehalter.csv' and a 'Entdecke' button.

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 Havaneser	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

- Siehe Step 0 in den Python-Unterlagen

EIN ERSTER «BLICK» IN DIE DATEN - ERKENNTNISSE

- 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



<https://i.pinimg.com/originals/4e/ae/65/4eae65b560cfad428d1874af80835451.jpg>

EIN ERSTER «BLICK» IN DIE DATEN - ERKENNTNISSE

- 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 Hundalter 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
- Die Verteilung der Hundebesitzer zu den Stadtteilen sehen wir in den Visualisierungen

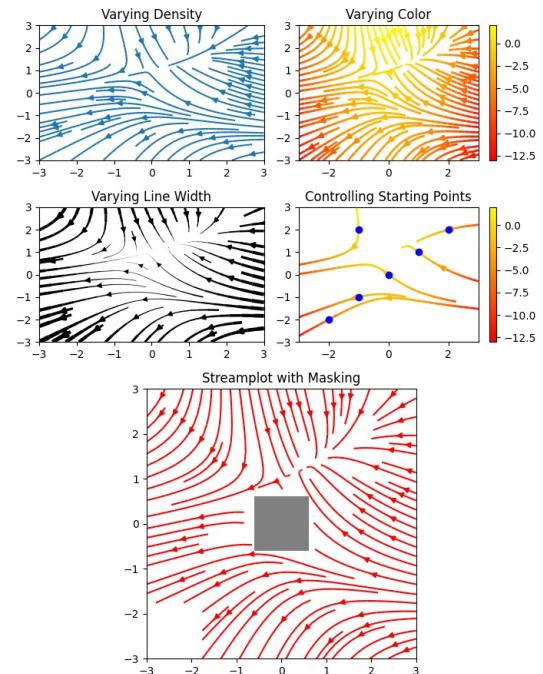
PANDAS

- `read_csv()` – Liest CSV Dateien ein
- `head()` – Gibt die ersten fünf Reihen aus
- `describe()` – Deskriptive Statistik
- `astype()` – Casting In eine anderen Datentyp
- `loc[:]` – Ausschnitte erstellen
- `value_counts()` – Zählt die Anzahl
- `drop_duplicates()` – Entfernt doppelte Einträge (Achtung!)
- `groupby()` – Gruppierungsfunktion
- `merge()` – Mehrere Dataframes zusammenführen
- `sort_values()` – Im Dataframe sortieren (Achtung!)
- `fillna()` – Ausfüllen von NaN values

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

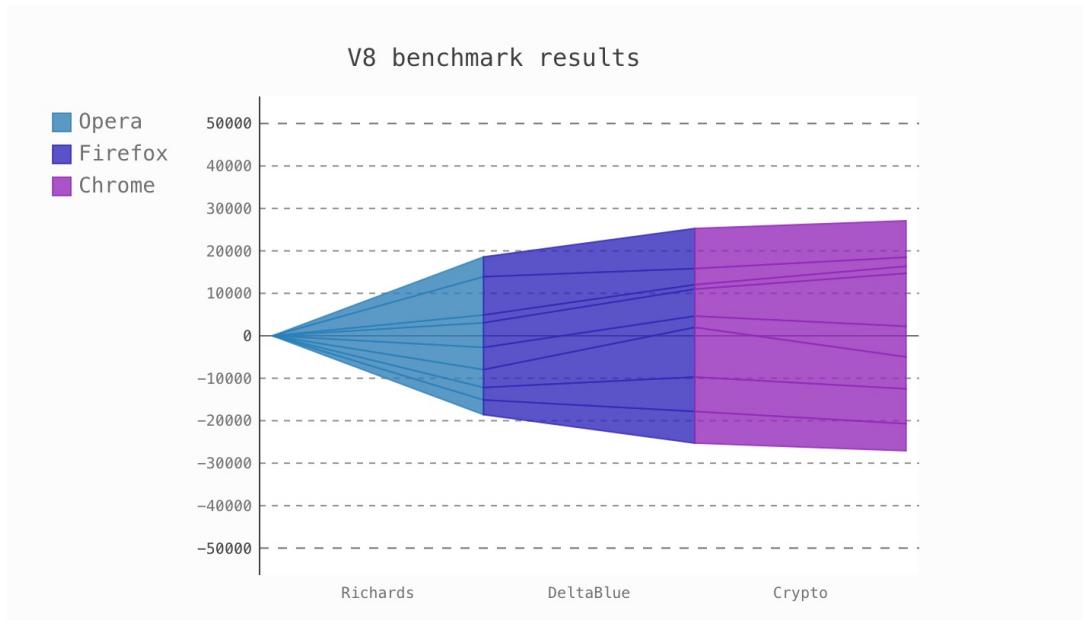
MATPLOTLIB



Streamplot with various plotting options.

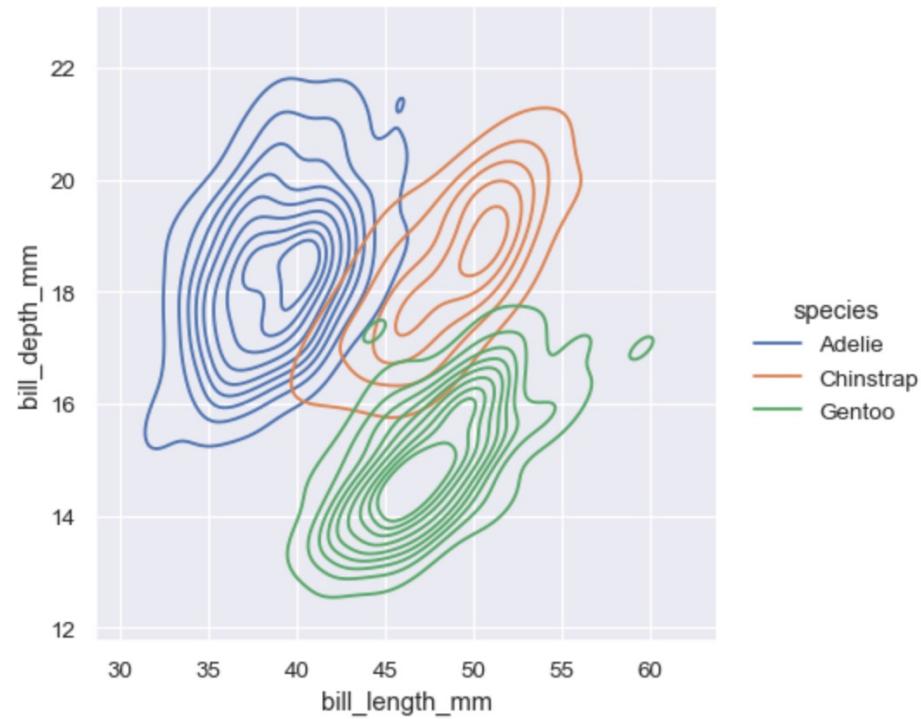
- Geeignet für einfache und komplexe Darstellungen
- Mehrere Darstellungen via Subplots möglich
- Bietet das grösste und allgemeinste Spektrum

PYGAL



- Wird leider nicht mehr (aktiv) weiterentwickelt
- Guter Funktionsumfang und dabei relativ einfach gelernt
- Einfach als interaktives HTML auszugeben

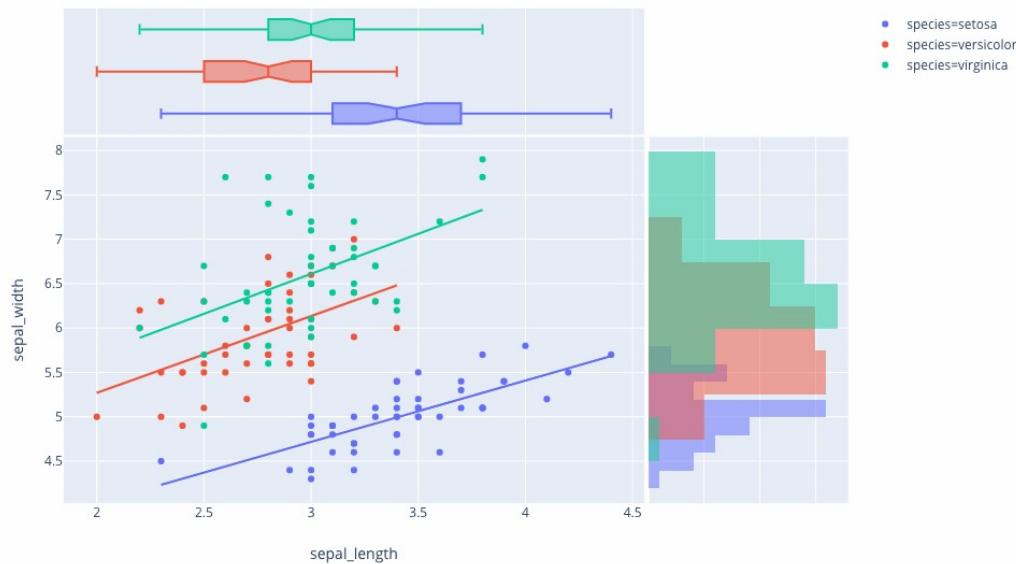
SEABORN



- Basiert auf Matplotlib
- Vor allem für statistische Visualisierungen geeignet

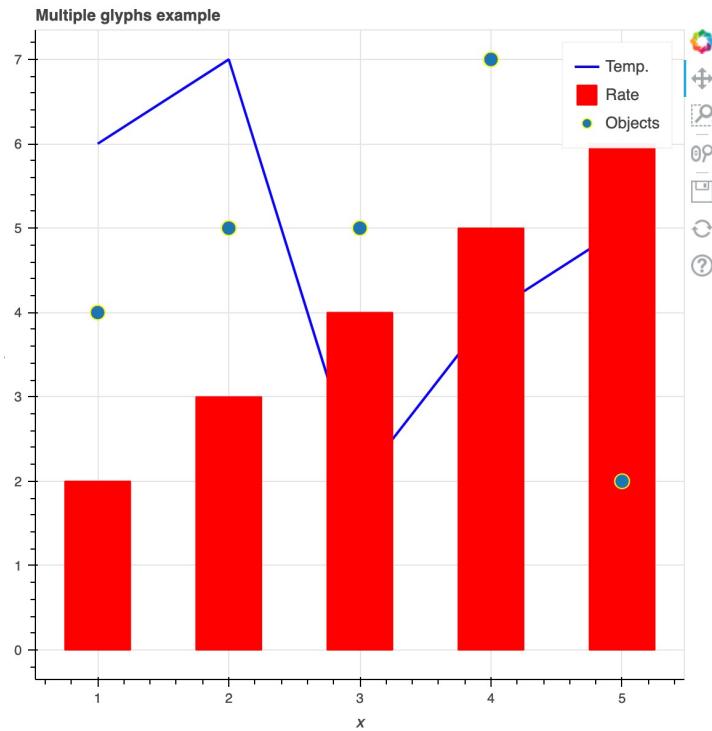
PLOTLY

```
px.scatter(iris, x="sepal_width", y="sepal_length", color="species", marginal_y="histogram",
           marginal_x="box", trendline="ols")
```



- 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



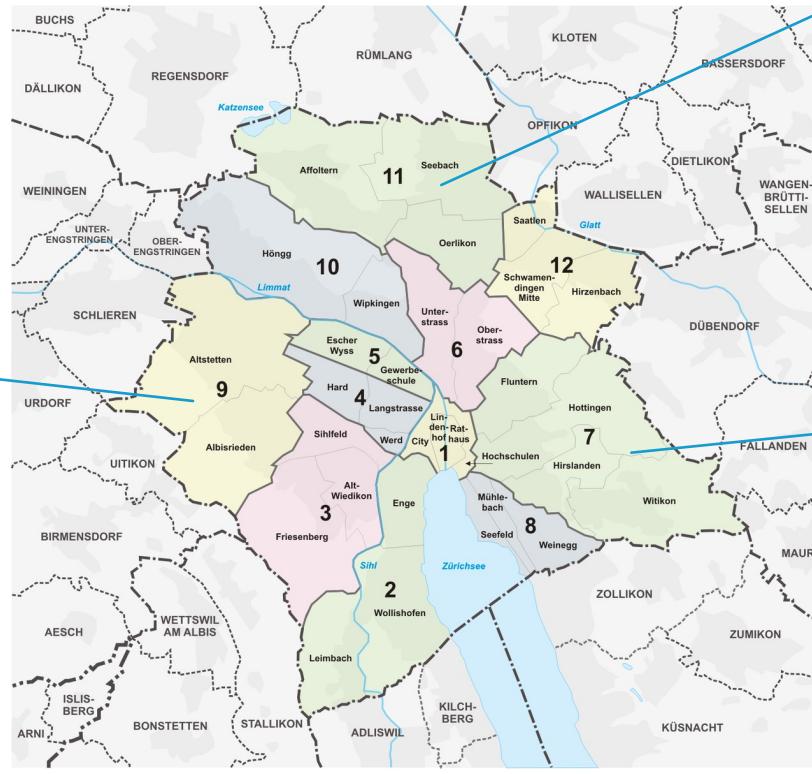
VISUALISIERUNG FÜR DOG-SCIENCE

- Siehe Step 1 in den Python-Unterlagen

DER ORT IST WICHTIG

Platz 1: 984

Platz 1: 1352



Platz 2: 1087

HUNDEHALTER & BEVÖLKERUNG?

 Stadt Zürich
Open Data

Startseite Datensätze Kategorien

🔍

🏠 / Datensätze / Hundebestand der Stadt Zürich

Lizenz
Creative Commons CCZero
OPEN DATA

 **200306_hundehalter.csv**
Comma-Separated Values. Weitere Informationen zu CSV finden Sie in unserer...

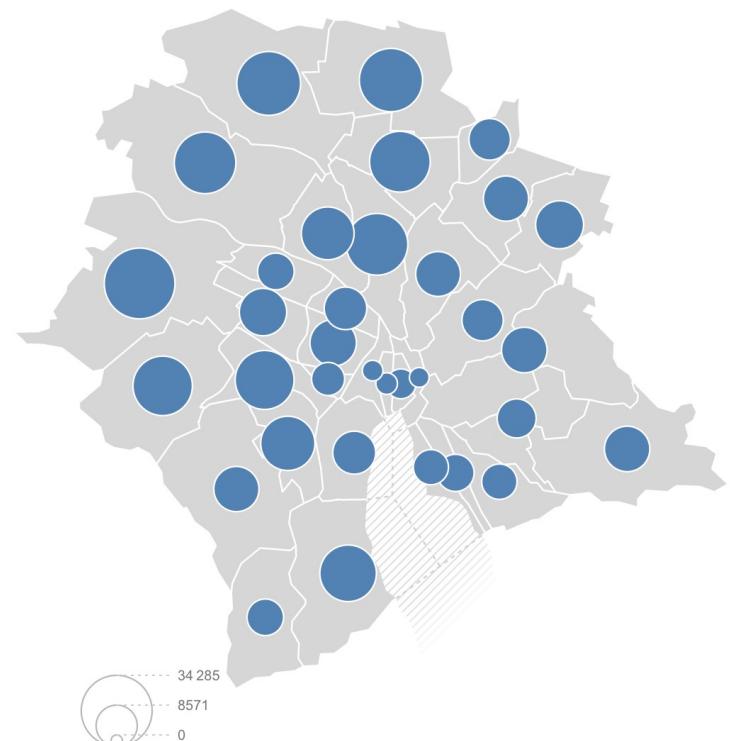


Hundebestand der Stadt Zürich

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 Rassentyp, das Geschlecht, das Geburtsjahr und die Farbe erfasst. Das Hunderegister wird von der Abteilung Hundekontrolle der Stadtpolizei Zürich geführt.

Daten und Ressourcen

Bevölkerung nach Stadtquartier

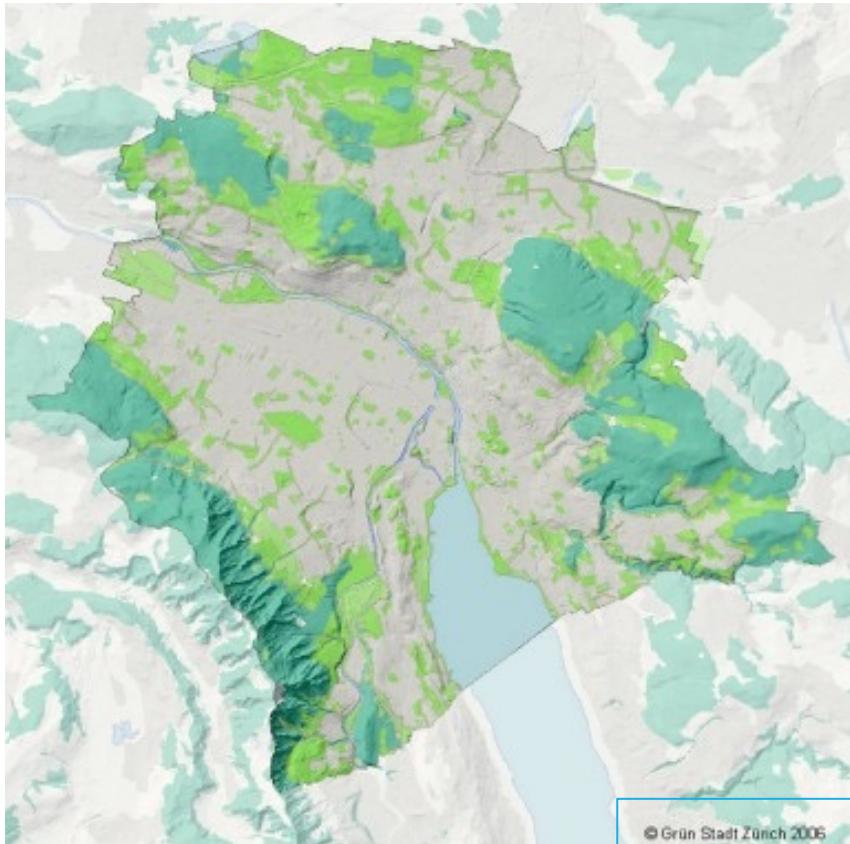


Hier besonders auf zeitliche Unterschiede achten:
Hundedaten: März 2020 / Bevölkerungsdaten: 1993-2020

IST DIE ANZAHL PERSONEN WICHTIG?

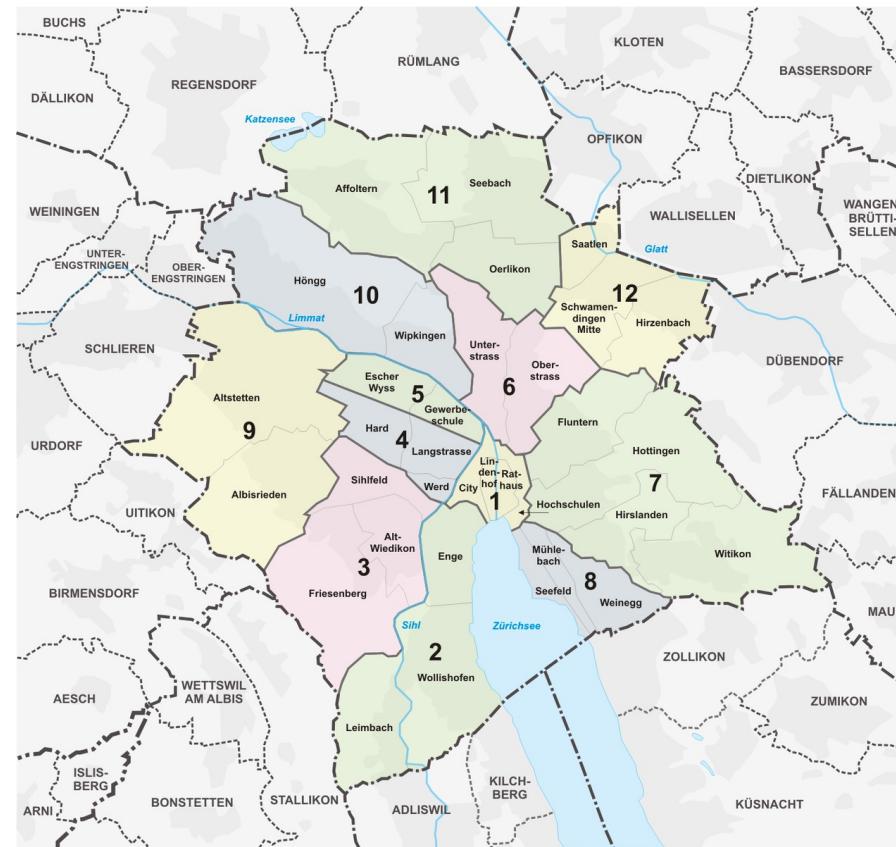
- Kreis 1; 5 831
- Kreis 2; 35 552
- Kreis 3; 50 756 => viele Menschen, weniger Hunde (697)
- Kreis 4; 29 034
- Kreis 5; 15 622
- Kreis 6; 35 317
- Kreis 7; 38 629
- Kreis 8; 17 456
- Kreis 9; 56 462
- Kreis 10; 41 044
- Kreis 11; 76 188
- Kreis 12; 32 845

SIND GRÜNFLÄCHEN WICHTIG?



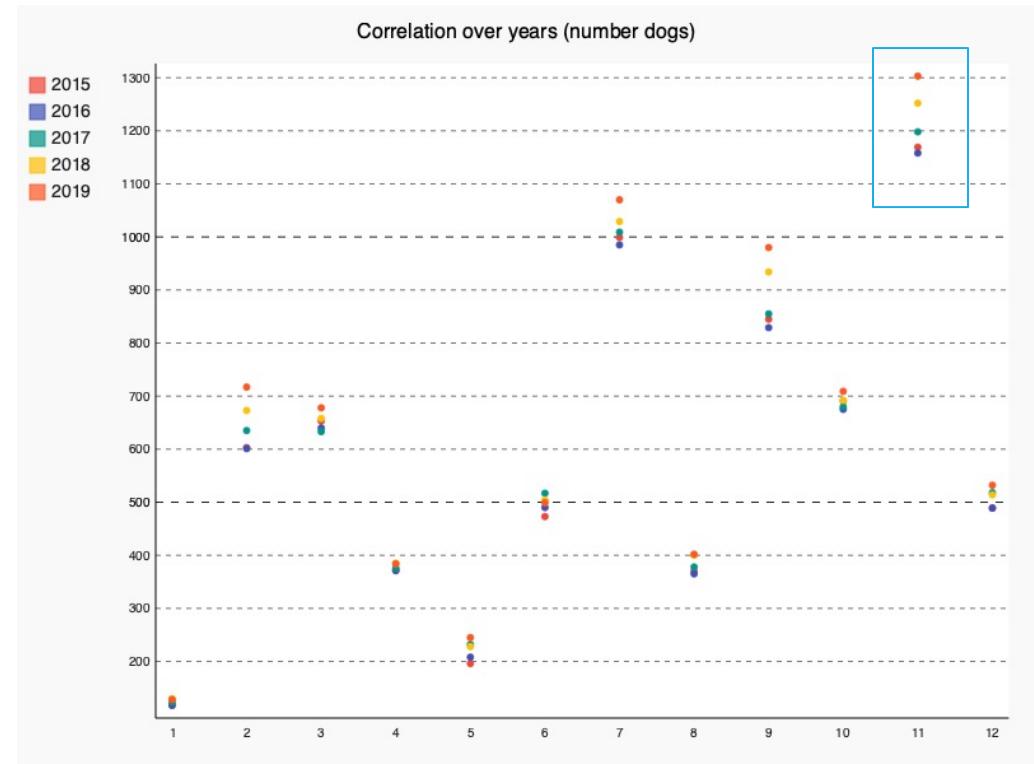
Alter!

https://www.stadt-zuerich.ch/ted/de/index/taz/erhalten/standards_stadtraeume_zuerich/raumtypen/gruenanlagen_gewaesser.html
DAVID PINEZICH, DAVID.PINEZICH@GMAIL.COM

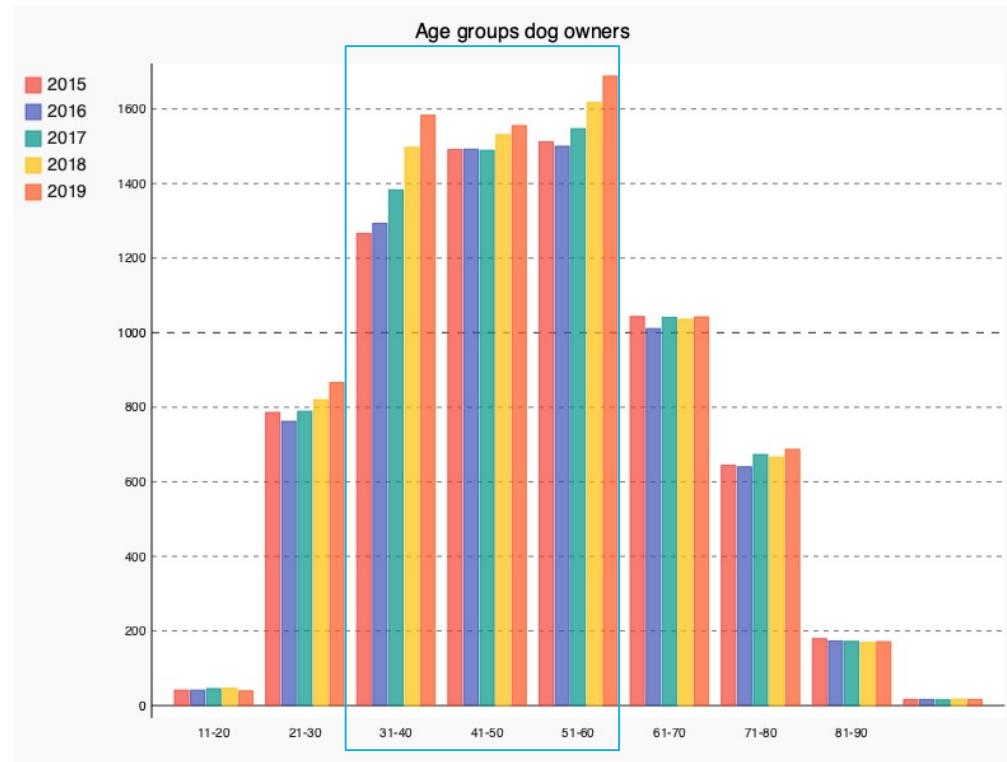


KORRELATION BERECHNEN

- Siehe Step 2 in den Python-Unterlagen



WELCHE ALTERSGRUPPE



- 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 (zugegeben, dafür haben wir eher wenig Daten)
 - Clustering
 - usw.

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 und einfach verständlich sein (Interpretation der Altersgruppe)