

# **COVID-19:**

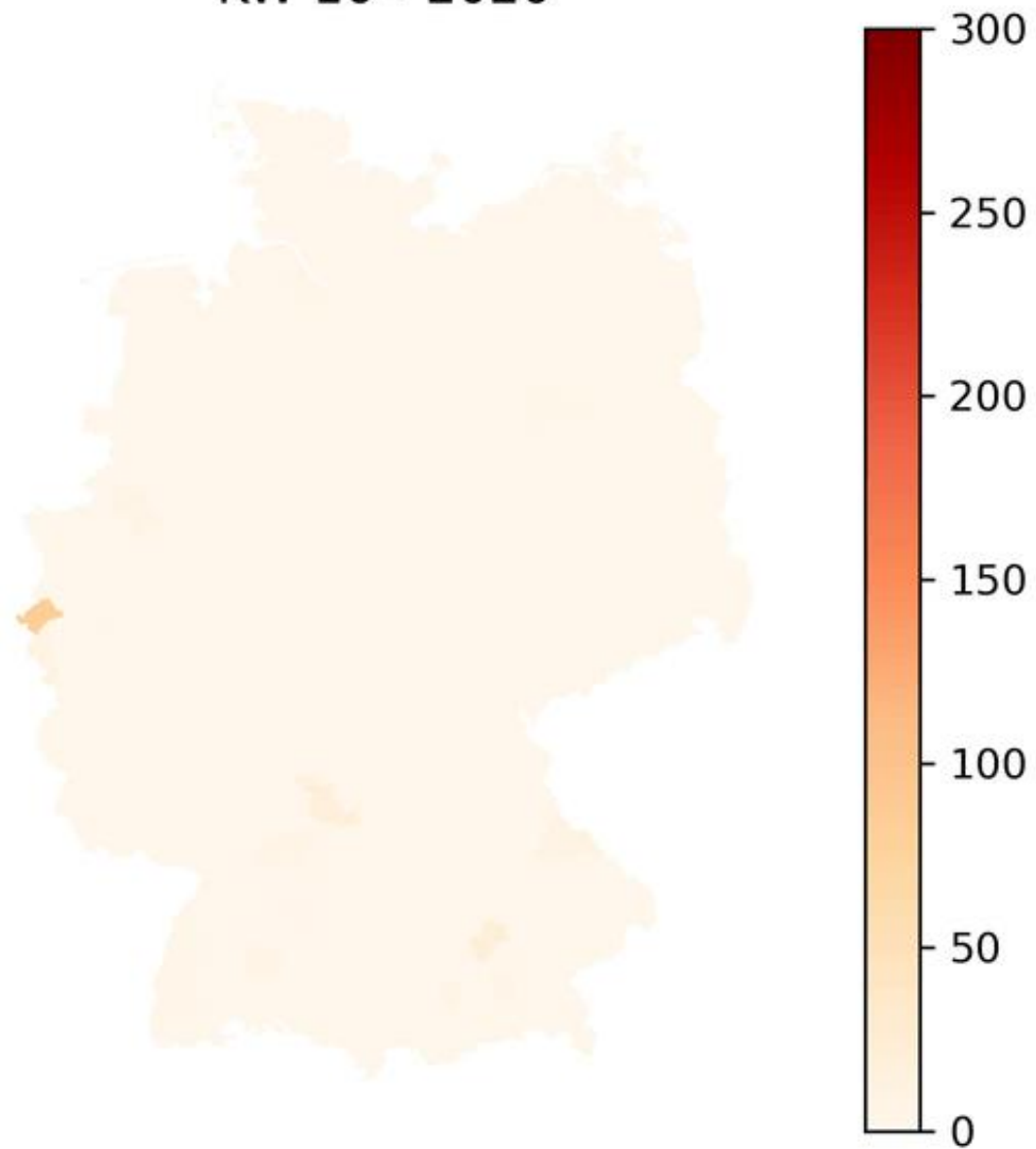
## **Prognosen auf kommunaler Ebene**

*Projekt Data Science – Sommersemester 2021*

# 7-Tage-Inzidenz

März 2020 bis April 2021

KW 10 - 2020



# Ziel

Vorhersage von COVID-19-Fallzahlen

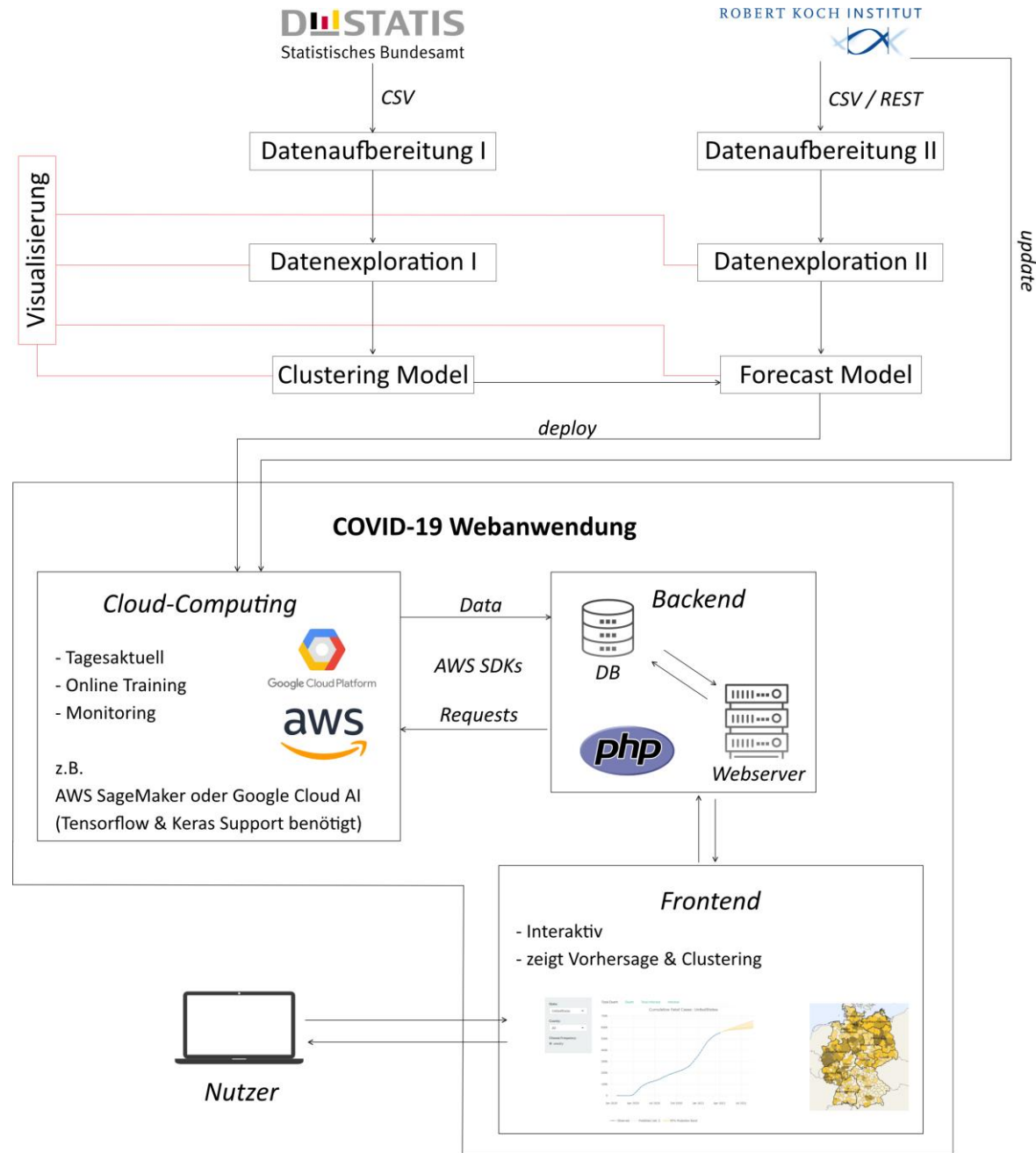
- Für alle 401 Landkreise
- Für die nächsten sieben Tage
- In einer Webanwendung

# Motivation

- Gesellschaftliche Relevanz
- Existenzielle Bedrohung
- Ausnutzen von Föderalismus-Strukturen
- Vermeidung künftiger Massenquarantänemaßnahmen
- Interesse an Zeitreihendaten

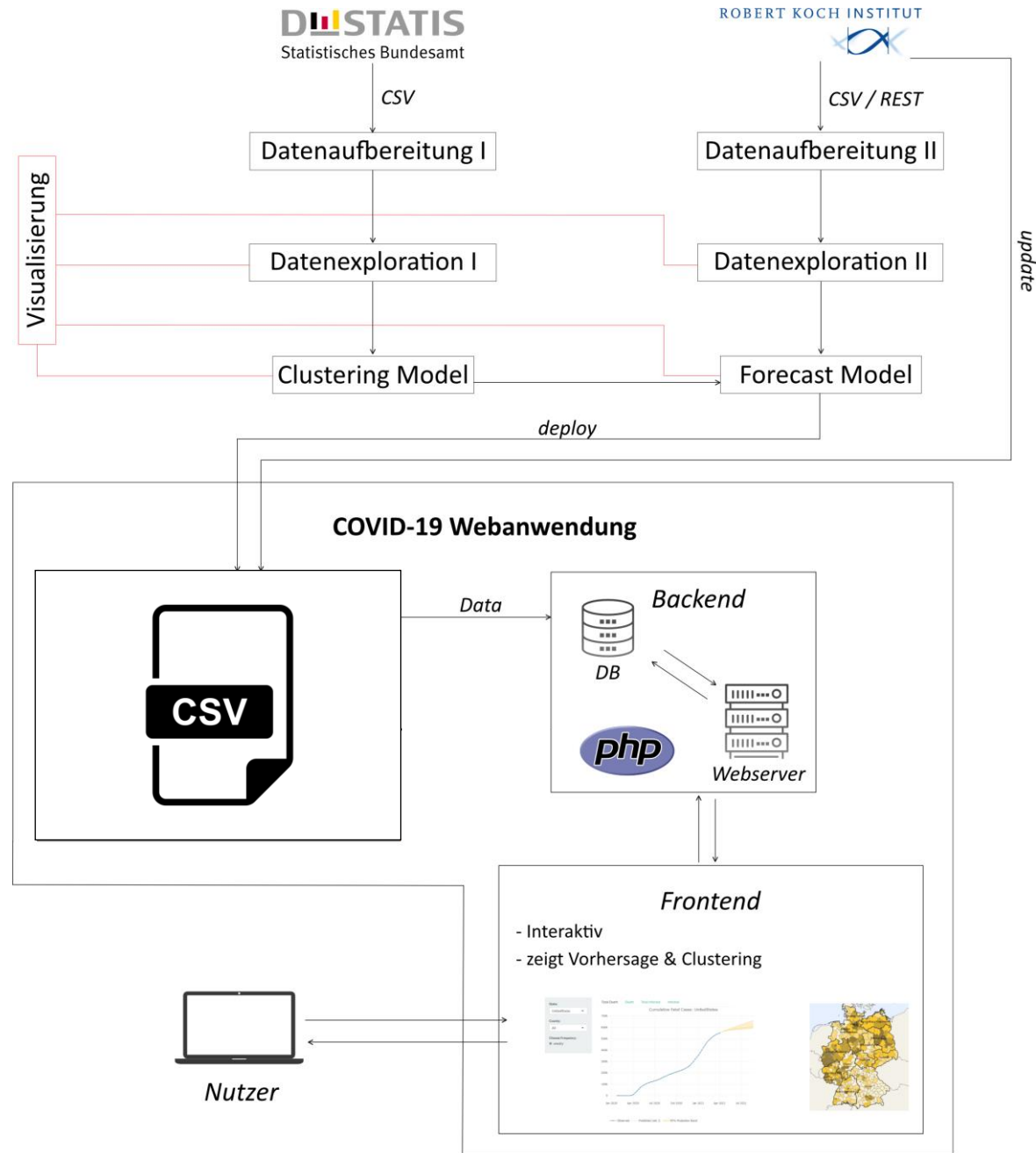
# Implementierung

Ursprüngliches Konzept



# Implementierung




Tatsächliche Umsetzung



# Daten

## 1. Fallzahlen

ROBERT KOCH INSTITUT

 1.695.186 x 18		Fallzahlen
 250 MB		Sterblichkeit/Genesung
 Robert Koch Institut		Alter und Geschlecht
 Veröffentlicht seit März 2020		
		 Öffentlich zugänglich



## 2. Regionalatlas und Einwohnerzahlen

STATIS  
Statistisches Bundesamt







 401 Zeilen x 40 Spalten		Demographie
 89 KB		Sozioökonomie
 Statistisches Bundesamt		
 Veröffentlicht in: 2017- 2020		
		 Öffentlich zugänglich

## 3. Koordinaten / Grenzen der Kreise


 Bundesministerium für Gesundheit






 402 Zeilen x 7 Spalten		
 973 KB		
 BKG		
 Veröffentlicht in: 2011		
		 Öffentlich zugänglich

## 4. Koordinaten / Grenzen der Kreise

 401 Zeilen x 7 Spalten		
 19.5 KB		
 Esri		
 Veröffentlicht in: 2021		
		 Öffentlich zugänglich

## 5. Impffortschritt

 Bundesministerium für Gesundheit

 136 Zeilen x 33 Spalten		
 31 KB		
 Statistisches Bundesamt		
 Veröffentlicht in: 2017- 2020		
		 Öffentlich zugänglich

## K-Means

Ähnlichkeitsmaß: Euklidischer Distanz

$K = 8$

## Agglomeratives Clustering

Ähnlichkeitsmaß: Euklidischer Distanz

Hierarchischer Algorithmus

- DBSCAN hat sehr empfindlich auf Dichteparameter reagiert
- Auswahl geeigneter Attribute (Statistisches Bundesamt – Regionalatlas)
  - Demographisch
  - Sozioökonomisch
  - Infrastruktur
- Probe, ob Clusterverfahren anwendbar sind (anhand Kreisart als Benchmark)
- Ellenbogenmethode zur Parameterbestimmung
- Clusterevaluierung

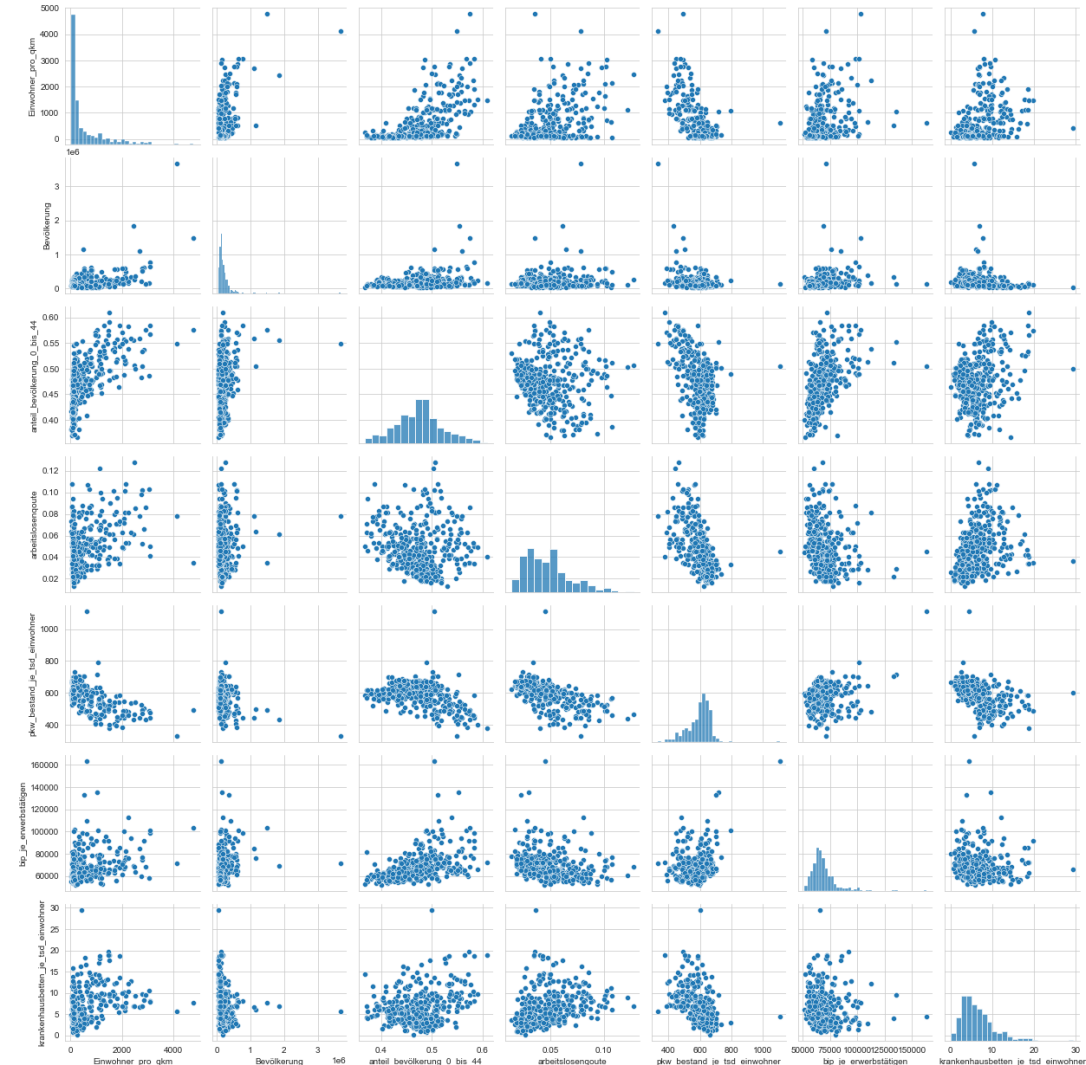
# Clustering

## Feature Selection

1. Bevölkerungsdichte  
(Einwohner pro Quadratkilometer)
2. Bevölkerungsanzahl
3. Altersverteilung  
(Anteil der 0 bis 44-jährigen)
4. Arbeitslosenquote
5. PKW-Bestand  
(Je Tsd. – Einwohner)
6. BIP je Erwerbstätigen
7. Medizinische Infrastruktur  
(Krankenhausbetten je Tsd. – Einwohner)

→ **MinMax-Skalierung** aller Attribute

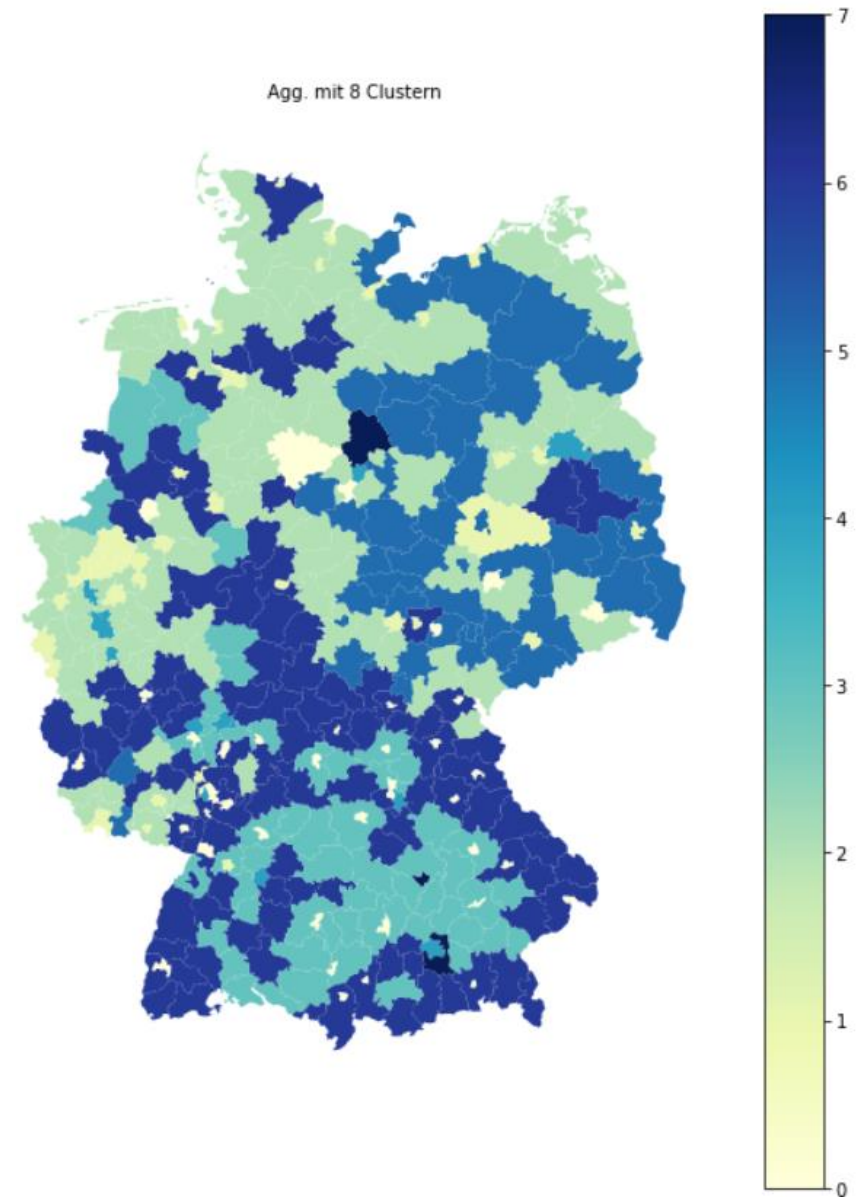
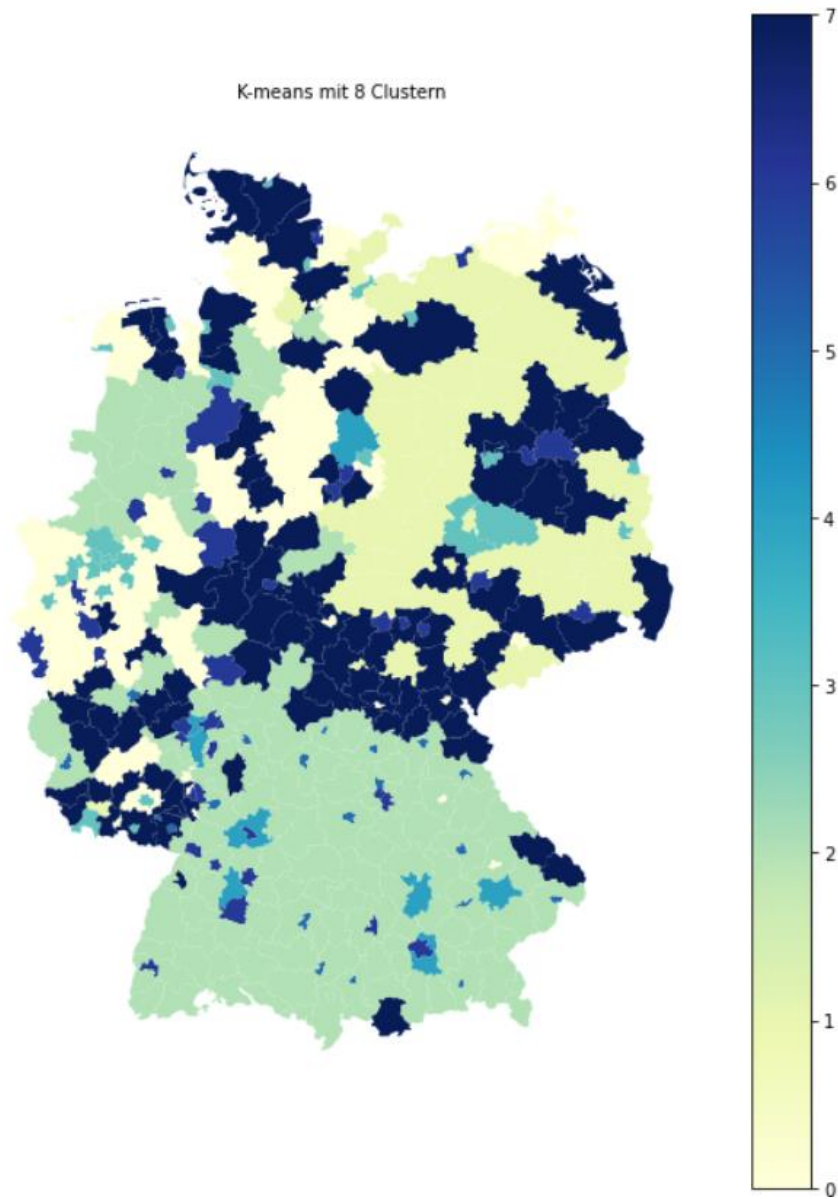
## Korrelation der Attribute





# Clustering

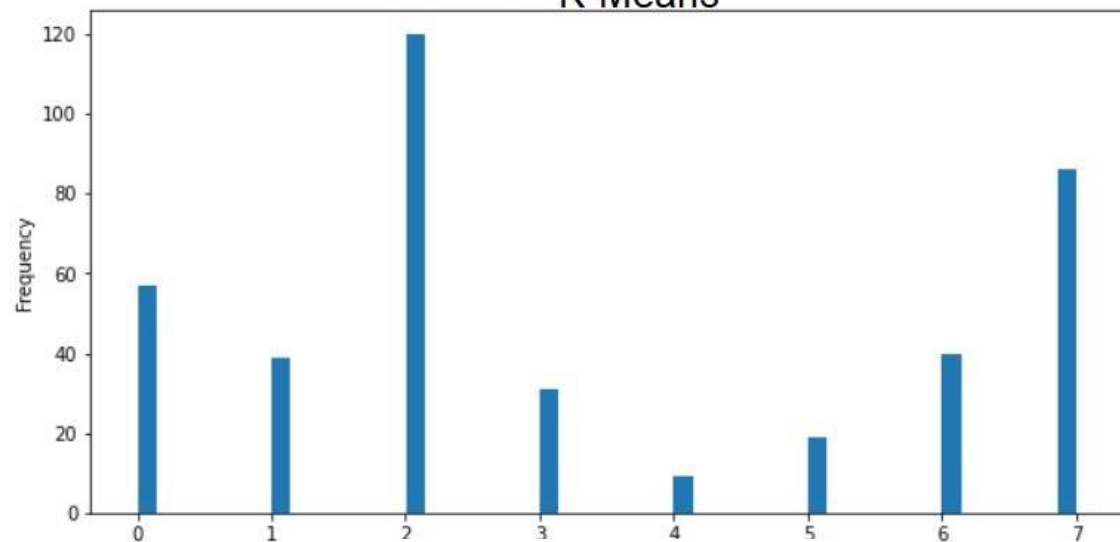
Ergebnisse



# Clustering

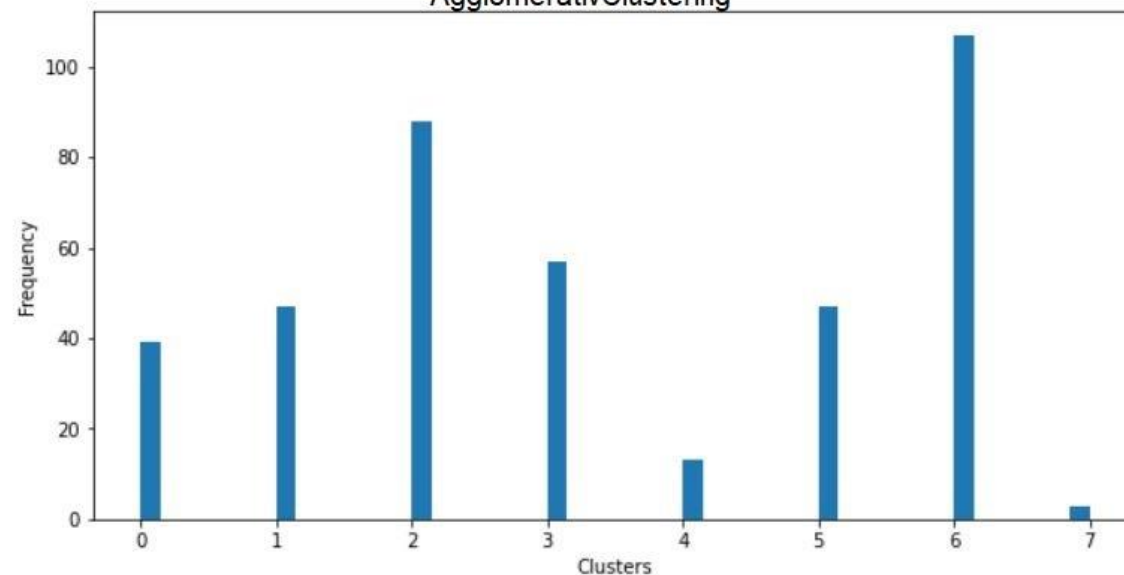
## Ergebnisse

K-Means



	id	Landkreis	Cluster
18	3103	Wolfsburg, Stadt	4
122	6433	Groß-Gerau	4
125	6436	Main-Taunus-Kreis	4
179	8115	Böblingen	4
185	8125	Heilbronn	4
222	9161	Ingolstadt	4
238	9184	München	4
240	9186	Pfaffenhofen a.d.Ilm	4
256	9279	Dingolfing-Landau	4

AgglomerativClustering



	id	Landkreis	Cluster
18	3103	Wolfsburg, Stadt	7
222	9161	Ingolstadt	7
238	9184	München	7

# Clustering

## Ergebnisse

### Cluster 4 – Kmeans Clustering

Landkreis	Einwohner_pro_qkm	Bevölkerung	anteil_bevölkerung_0_bis_44	arbeitslosenquote	pkw_bestand_je_tsd_einwohner	bip_je_erwerbstätigen	krankenhausbetten_je_tsd_einwohner
Wolfsburg, Stadt	0.120650	0.024806	0.570248	0.278261	1.000000	1.000000	0.143836
Groß-Gerau	0.120861	0.066441	0.632231	0.295652	0.387176	0.410813	0.123288
Main-Taunus-Kreis	0.218519	0.056217	0.504132	0.173913	0.593035	0.441638	0.092466
Böblingen	0.126556	0.098648	0.603306	0.130435	0.403238	0.514053	0.089041
Heilbronn	0.058426	0.085347	0.570248	0.130435	0.477512	0.392344	0.116438
Ingolstadt	0.209660	0.028388	0.764463	0.139130	0.492033	0.751143	0.321918
München	0.103776	0.087002	0.595041	0.078261	0.477512	0.729221	0.126712
Pfaffenhofen a.d.Ilm	0.027842	0.025867	0.623967	0.026087	0.403238	0.444877	0.082192
Dingolfing-Landau	0.015609	0.017190	0.553719	0.139130	0.472758	0.436001	0.037671

### Cluster 7 - Hierarchisches Clustering

Landkreis	Einwohner_pro_qkm	Bevölkerung	anteil_bevölkerung_0_bis_44	arbeitslosenquote	pkw_bestand_je_tsd_einwohner	bip_je_erwerbstätigen	krankenhausbetten_je_tsd_einwohner
Wolfsburg, Stadt	0.120650	0.024806	0.570248	0.278261	1.000000	1.000000	0.143836
Ingolstadt	0.209660	0.028388	0.764463	0.139130	0.492033	0.751143	0.321918
München	0.103776	0.087002	0.595041	0.078261	0.477512	0.729221	0.126712

**COVID-19**

Daten

ROBERT KOCH INSTITUT



## RKI COVID19



Privates Mitglied   
Private Organisation

## Zusammenfassung

Tabelle mit den aktuellen Covid-19 Infektionen pro Tag (Zeitreihe).

[Vollständige Details anzeigen](#)

**Dataset**  
Table



**8. Juli 2021**  
Informationen aktualisiert



**8. Juli 2021**  
Datenaktualisierung



**18. März 2020**  
Veröffentlichungsdatum



**2.086.381 Datensätze**  
[Datentabelle anzeigen](#)



**Öffentlich**  
Inhalt ist für alle Benutzer sichtbar




**Benutzerdefinierte Lizenz**  
[Anzeigen von Lizenzdetails](#)

25 von 2.086.381 Zeilen werden angezeigt

	IdBundesland	Bundesland	Landkreis	Altersgruppe	Geschlecht	AnzahlFall	AnzahlTodesfall	Meldedatum	Landkreis ID	Neu
	1	Schleswig-Holstein	SK Flensburg	A00-A04	M	1	0	30.9.2020, 02:00	01001	0
	1	Schleswig-Holstein	SK Flensburg	A00-A04	M	1	0	29.10.2020, 01:00	01001	0
	1	Schleswig-Holstein	SK Flensburg	A00-A04	M	1	0	3.11.2020, 01:00	01001	0
	1	Schleswig-Holstein	SK Flensburg	A00-A04	M	1	0	20.11.2020, 01:00	01001	0
	1	Schleswig-Holstein	SK Flensburg	A00-A04	M	1	0	23.11.2020, 01:00	01001	0
	1	Schleswig-Holstein	SK Flensburg	A00-A04	M	1	0	18.12.2020, 01:00	01001	0
	1	Schleswig-Holstein	SK Flensburg	A00-A04	M	2	0	6.1.2021, 01:00	01001	0
	1	Schleswig-Holstein	SK Flensburg	A00-A04	M	1	0	8.1.2021, 01:00	01001	0
	1	Schleswig-Holstein	SK Flensburg	A00-A04	M	1	0	9.1.2021, 01:00	01001	0
	1	Schleswig-Holstein	SK Flensburg	A00-A04	M	1	0	16.1.2021, 01:00	01001	0
	1	Schleswig-Holstein	SK Flensburg	A00-A04	M	1	0	22.1.2021, 01:00	01001	0
	1	Schleswig-Holstein	SK Flensburg	A00-A04	M	1	0	22.1.2021, 01:00	01001	0

RKI COVID19



Privates Mitglied

Private Organisation

Zusammenfassung

Tabelle mit den aktuellen Covid-19 Infektionen pro Tag (Zeitreihe).

Vollständige Details anzeigen

- 

Dataset

Table
- 

8. Juli 2021

Informationen aktualisiert
- 

8. Juli 2021

Datenaktualisierung
- 

18. März 2020

Veröffentlichungsdatum
- 

2.086.381 Datensätze

Datentabelle anzeigen
- 

Öffentlich

Inhalt ist für alle Benutzer sichtbar
- 

Benutzerdefinierte Lizenz

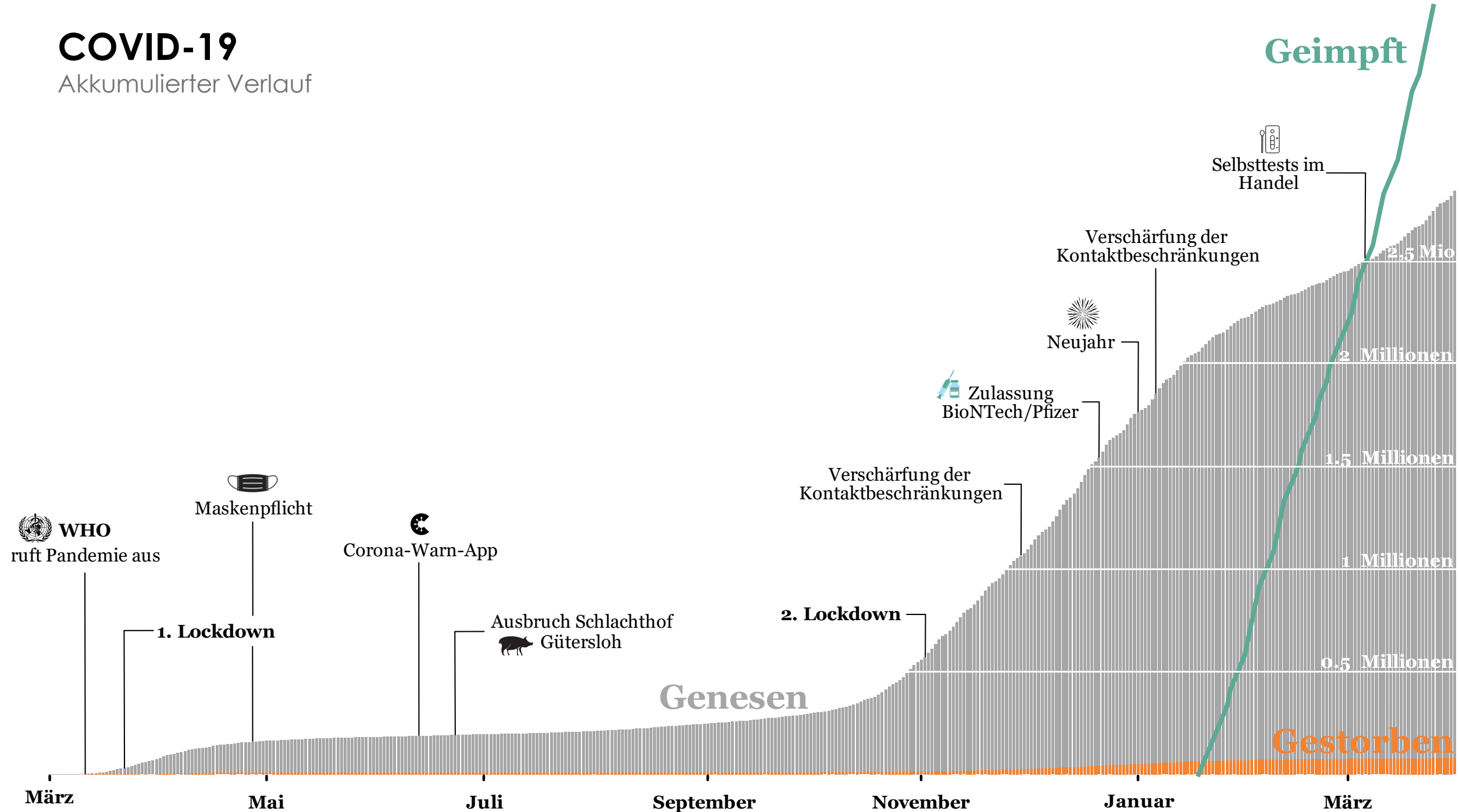
Anzeigen von Lizenzdetails

25 von 2.086.381 Zeilen werden angezeigt

Recht	AnzahlFall	AnzahlTodesfall	Meldedatum	Landkreis ID	Neuer Fall	Neuer Todesfall	Referenzdatum	Neu Genesen	Anzahl Genesen	
	1	0	30.9.2020, 02:00	01001	0	-9	30.9.2020, 02:00	0	1	
	1	0	29.10.2020, 01:00	01001	0	-9	29.10.2020, 01:00	0	1	
	1	0	3.11.2020, 01:00	01001	0	-9	3.11.2020, 01:00	0	1	
	1	0	20.11.2020, 01:00	01001	0	-9	19.11.2020, 01:00	0	1	
	1	0	23.11.2020, 01:00	01001	0	-9	18.11.2020, 01:00	0	1	
	1	0	18.12.2020, 01:00	01001	0	-9	14.12.2020, 01:00	0	1	
	2	0	6.1.2021, 01:00	01001	0	-9	6.1.2021, 01:00	0	2	
	1	0	8.1.2021, 01:00	01001	0	-9	6.1.2021, 01:00	0	1	
	1	0	9.1.2021, 01:00	01001	0	-9	9.1.2021, 01:00	0	1	
	1	0	16.1.2021, 01:00	01001	0	-9	15.1.2021, 01:00	0	1	
	1	0	22.1.2021, 01:00	01001	0	-9	21.1.2021, 01:00	0	1	
	1	0	22.1.2021, 01:00	01001	0	-9	22.1.2021, 01:00	0	1	

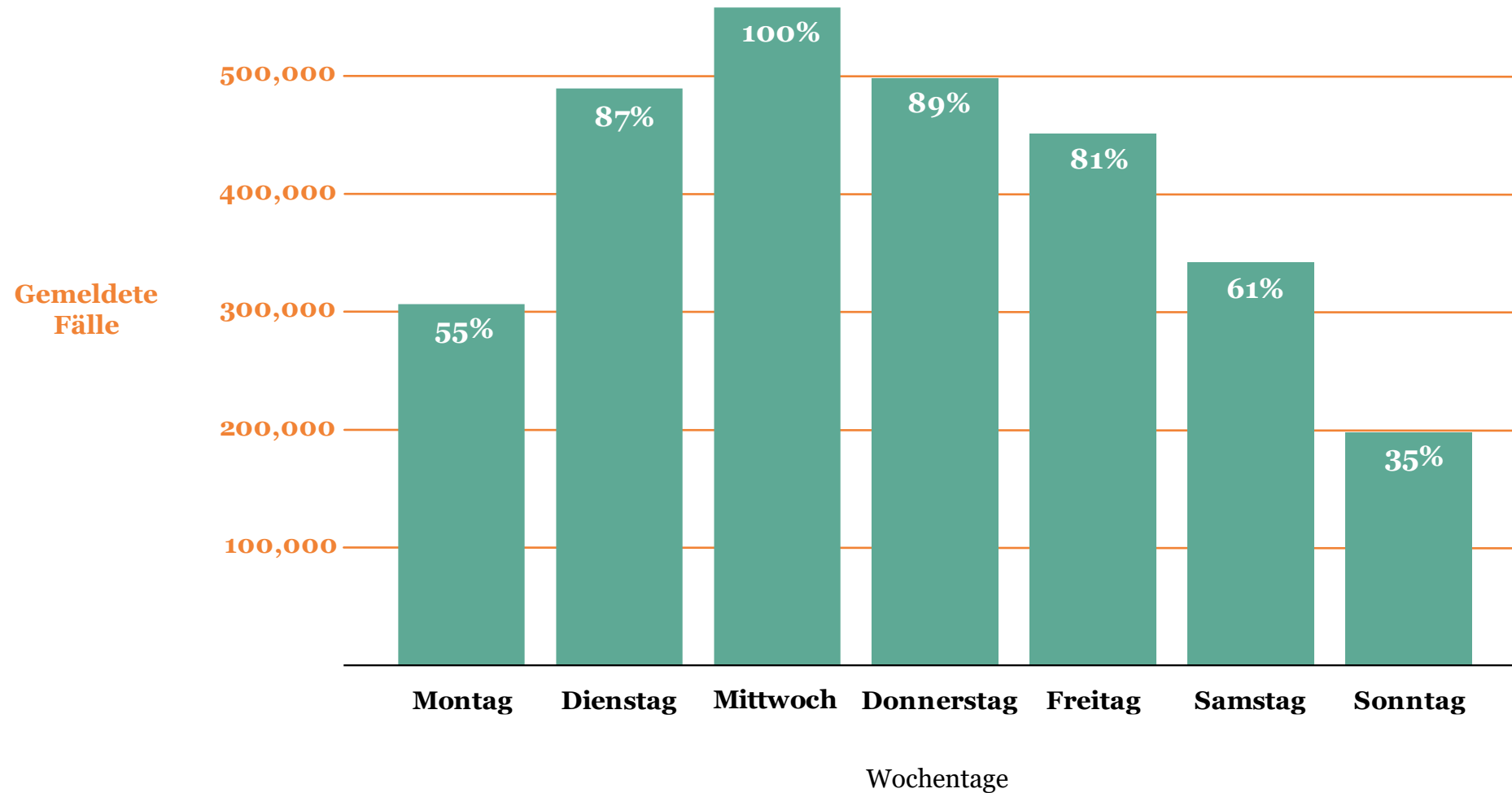
# COVID-19

Akkumulierter Verlauf



# COVID-19

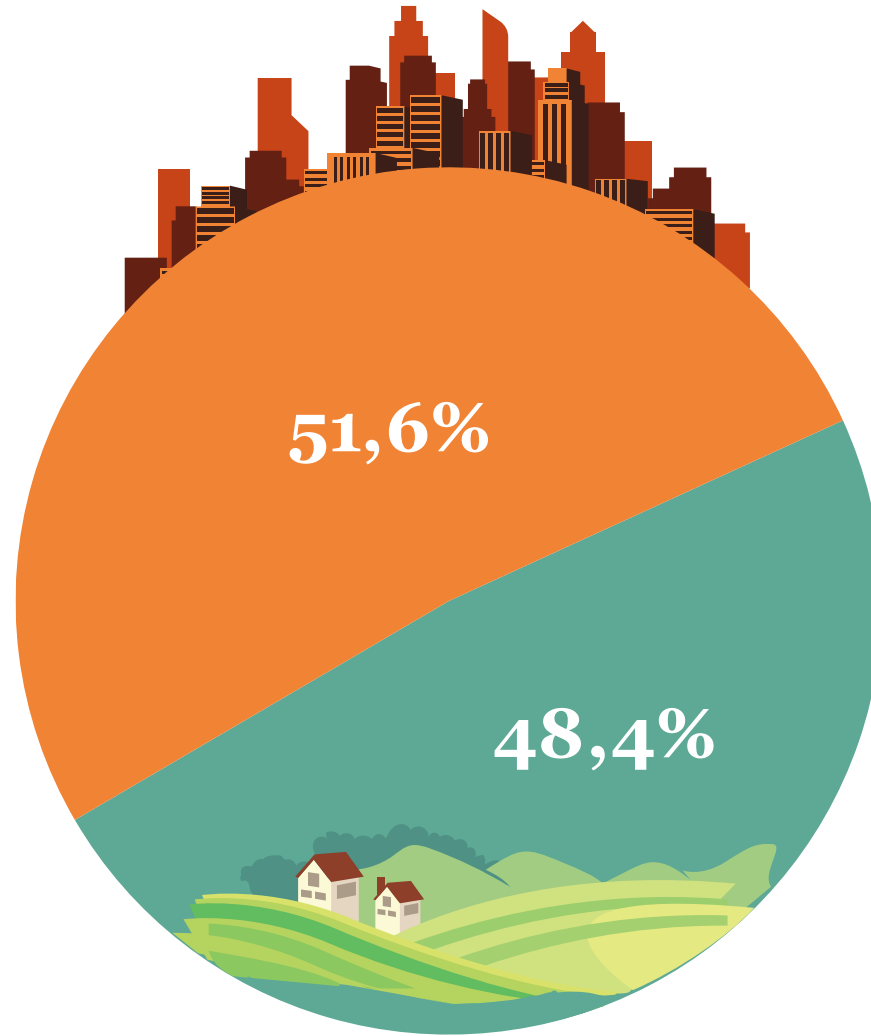
Wochentage





# COVID-19

Stadt vs. Land



# Zeitreihenanalyse

## Einführung

### 1) Stationarität

- Konstanz der Eigenschaften im Zeitverlauf

### 2) Trend

- Langfristige und nachhaltige Veränderung der Zeitreihenvariable
- Unidirektional
- Unabhängig von kurzfristigen und allgemeinen Schwankungen

### 3) Strukturbruch

- Signifikante Veränderung der Regressionsparameter
- Einmalig

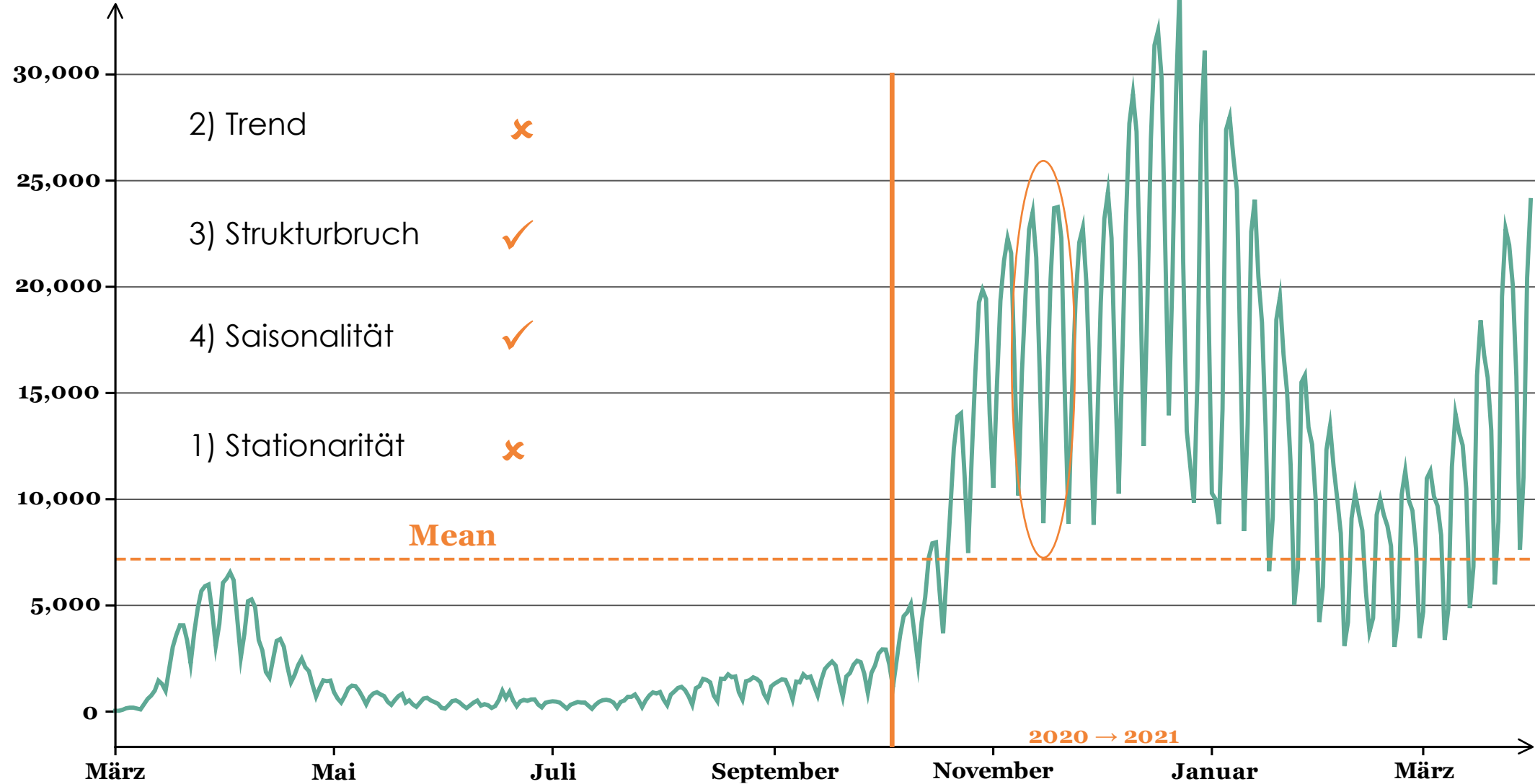
### 4) Saisonalität

- Wiederkehrende Variation
- Innerhalb spezifischer zeitlicher Intervalle
- Simultan möglich

# COVID-19

Zeitreihe März 2020 bis April 2021

Gemeldete Fälle  
pro Tag



# Modellierung

## Überblick

### 1) Naïve / Snaïve

- $Vorhersage_t = Fallzahlen_{t-7}$

### 2) Snaïve mit Trendkomponente

- $Vorhersage_t = Fallzahlen_{t-7} * \frac{Fallzahlen_{t-1}}{Fallzahlen_{t-8}}$

### 3) Exponentielle Glättung

- $Vorhersage_t = Fallzahlen_{t-7} * \alpha(Fallzahlen_{t-7} - Vorhersage_{t-7})$
- $\alpha = 0,3$

### 4) ARIMA

### 5) & 6) MLP

- Features:  $Fallzahlen_{t-1}, Fallzahlen_{t-2}, \dots, Fallzahlen_{t-7}$
- Wochentag One-Hot-Encoded
- Ohne und mit Ergebnissen des Clusterings

### 7) RNN

- LSTM (Long short-term memory)
- Features:  $Fallzahlen_{t-1}, Fallzahlen_{t-2}, \dots, Fallzahlen_{t-7}$

## Testzeitraum

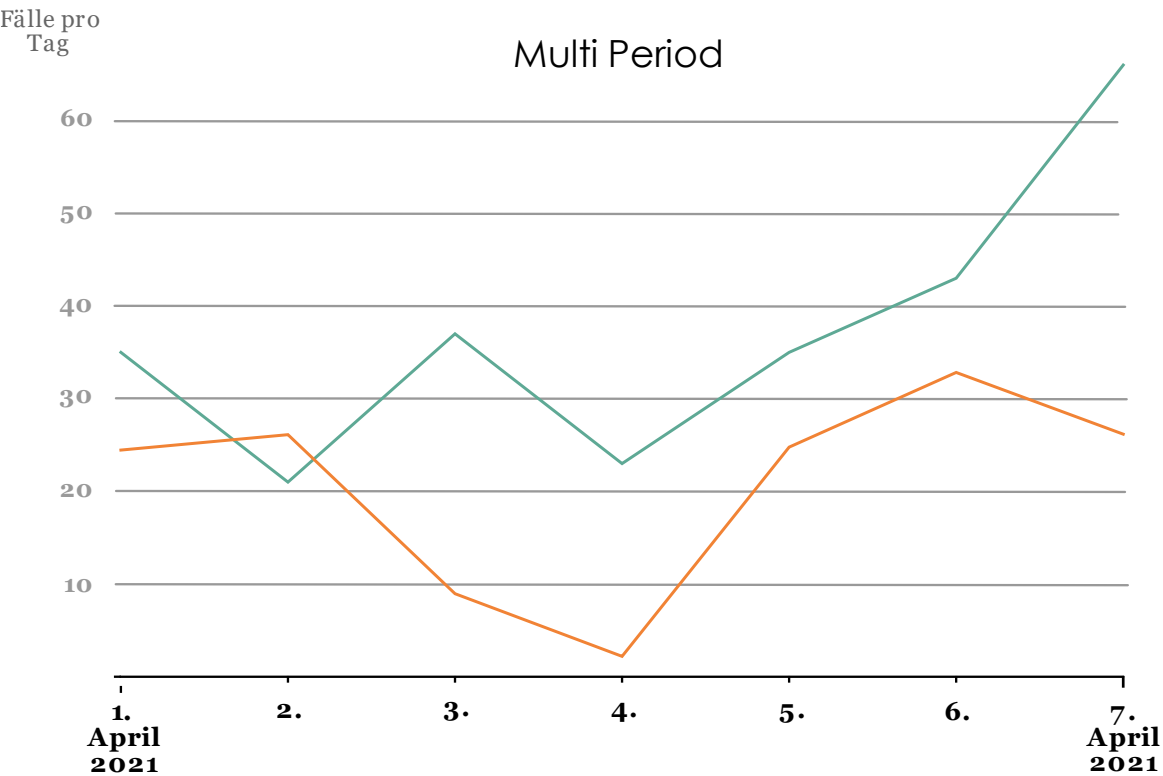
01. April 2021 bis 20. Mai 2021

## Trainingszeitraum

01. März 2020 bis 31. März 2021

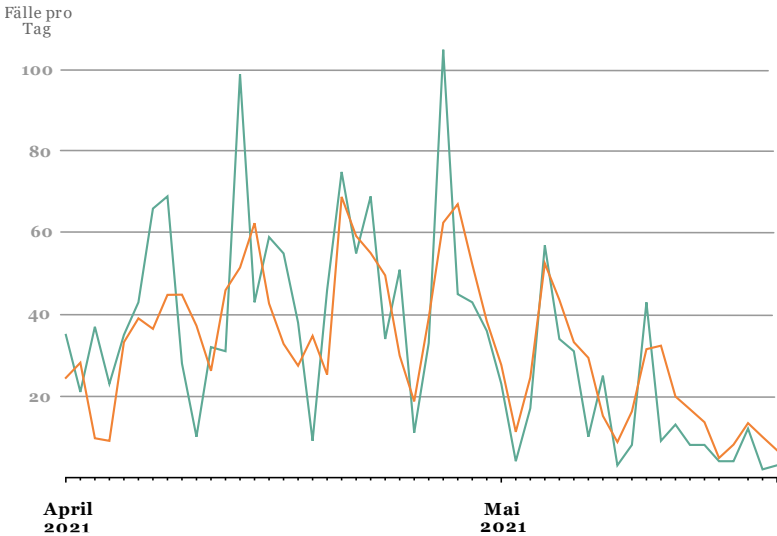
# Evaluierung

Anhalt - Bitterfeld

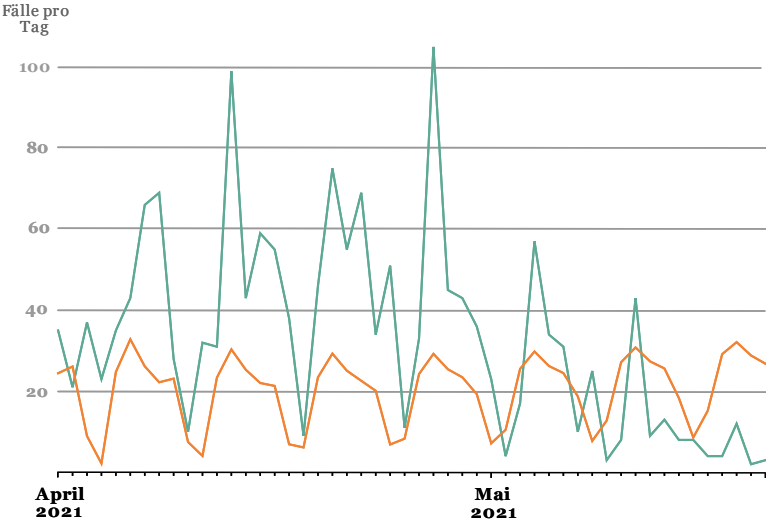


Tatsächlich  
Vorhersage

Single Period



Total Period



# Evaluierung

## Testperformanz

Kreis	Zeitraum	Snaïve		SNaïve mit Trend		Exponentielle Glättung		ARIMA		MLP		MLP mit Clustering		RNN	
		MAE	R <sup>2</sup>	MAE	R <sup>2</sup>	MAE	R <sup>2</sup>	MAE	R <sup>2</sup>	MAE	R <sup>2</sup>	MAE	R <sup>2</sup>	MAE	R <sup>2</sup>
Hamburg	Single Period	63.7	0.53	52.4	0.64	68.8	0.45	38.3	0.82	47.3	0.74	48.3	0.74	47.6	0.77
	Multi Period	90.0	-0.94	94.3	-0.77	120.3	-2.02	86.3	-0.24	91.4	-0.48	114.0	-0.79	54.8	0.45
	Total Period	180.9	-2.71	299.7	-8.76	210.0	-3.81	173.5	-1.97	245.3	-5.33	445.0	-19.36	103.5	-0.14
Ortenau	Single Period	30.9	0.39	37.7	-0.76	37.8	0.11	29.1	0.55	29.1	0.46	27.4	0.53	28.3	0.51
	Multi Period	38.7	0.10	36.1	-0.01	45.0	-0.16	35.9	0.37	40.0	0.30	40.7	0.22	35.8	0.44
	Total Period	29.5	0.40	59.3	-0.90	33.5	0.23	35.8	0.34	37.7	0.20	34.8	0.36	36.9	0.26
Anhalt-Bitterfeld	Single Period	16.6	0.28	34.7	-15.20	18.7	0.03	14.9	0.42	14.5	0.42	13.8	0.47	13.0	0.54
	Multi Period	16.7	-1.34	110.1	-322.26	18.7	-1.90	16.5	-1.14	14.1	-0.59	13.4	-0.26	17.8	-1.36
	Total Period	18.7	0.04	67.4	-16.68	20.0	-0.02	20.6	-0.19	23.2	-0.40	22.0	-0.28	20.8	-0.17
Sächsische Schweiz- Osterzgebirge	Single Period	27.7	0.21	56.7	-11.62	34.0	-0.19	29.9	0.24	25.1	0.39	23.7	0.49	23.4	0.44
	Multi Period	35.0	-0.04	32.4	-0.13	33.3	0.05	42.3	-0.03	32.9	0.30	33.9	0.21	26.7	0.51
	Total Period	23.0	0.55	327.6	-117.34	24.0	0.54	37.8	0.04	28.0	0.27	25.7	0.37	34.1	0.14

# Evaluierung

## Testperformanz

Kreis	Zeitraum	Snaïve		SNaïve mit Trend		Exponentielle Glättung		ARIMA		MLP		MLP mit Clustering		RNN	
		MAE	R <sup>2</sup>	MAE	R <sup>2</sup>	MAE	R <sup>2</sup>	MAE	R <sup>2</sup>	MAE	R <sup>2</sup>	MAE	R <sup>2</sup>	MAE	R <sup>2</sup>
Hamburg	Single Period	63.7	0.53	52.4	0.64	68.8	0.45	38.3	0.82	47.3	0.74	48.3	0.74	47.6	0.77
Ortenau	Single Period	30.9	0.39	37.7	-0.76	37.8	0.11	29.1	0.55	29.1	0.46	27.4	0.53	28.3	0.51
Anhalt-Bitterfeld	Single Period	16.6	0.28	34.7	-15.20	18.7	0.03	14.9	0.42	14.5	0.42	13.8	0.47	13.0	0.54
Sächsische Schweiz-Osterzgebirge	Single Period	27.7	0.21	56.7	-11.62	34.0	-0.19	29.9	0.24	25.1	0.39	23.7	0.49	23.4	0.44

Single Period  $\triangleq$  Täglich eine neue Vorhersage für den nächsten Tag

# Evaluierung

## Testperformanz

Kreis	Zeitraum	Snaïve		SNaïve mit Trend		Exponentielle Glättung		ARIMA		MLP		MLP mit Clustering		RNN	
		MAE	$R^2$	MAE	$R^2$	MAE	$R^2$	MAE	$R^2$	MAE	$R^2$	MAE	$R^2$	MAE	$R^2$
Hamburg	Multi Period	90.0	-0.94	94.3	-0.77	120.3	-2.02	86.3	-0.24	91.4	-0.48	114.0	-0.79	54.8	0.45
Ortenau	Multi Period	38.7	0.10	36.1	-0.01	45.0	-0.16	35.9	0.37	40.0	0.30	40.7	0.22	35.8	0.44
Anhalt-Bitterfeld	Multi Period	16.7	-1.34	110.1	-322.26	18.7	-1.90	16.5	-1.14	14.1	-0.59	13.4	-0.26	17.8	-1.36
Sächsische Schweiz- Osterzgebirge	Multi Period	35.0	-0.04	32.4	-0.13	33.3	0.05	42.3	-0.03	32.9	0.30	33.9	0.21	26.7	0.51

Multi Period  $\triangleq$  Vorhersage am 01. April 2021 für die nächsten sieben Tage

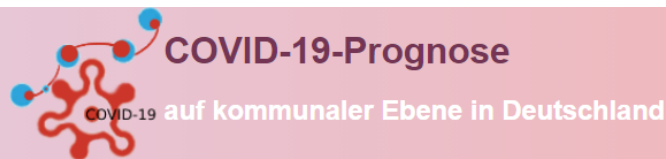


# Evaluierung

## Testperformanz

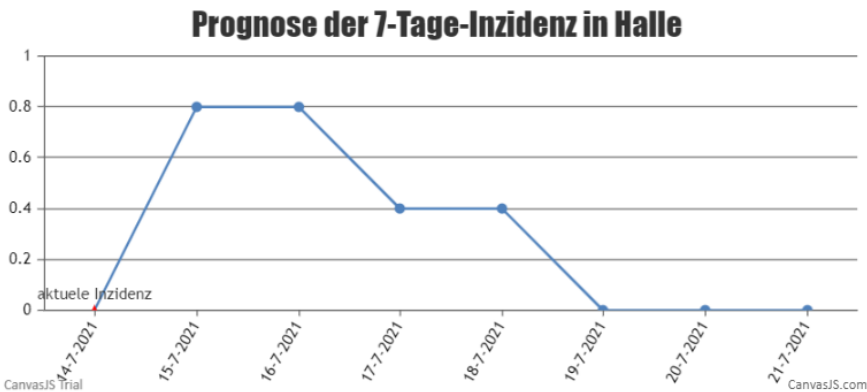
Kreis	Zeitraum	Snaïve		SNaïve mit Trend		Exponentielle Glättung		ARIMA		MLP		MLP mit Clustering		RNN	
		MAE	$R^2$	MAE	$R^2$	MAE	$R^2$	MAE	$R^2$	MAE	$R^2$	MAE	$R^2$	MAE	$R^2$
Hamburg	Total Period	180.9	-2.71	299.7	-8.76	210.0	-3.81	173.5	-1.97	245.3	-5.33	445.0	-19.36	103.5	-0.14
Ortenau	Total Period	29.5	0.40	59.3	-0.90	33.5	0.23	35.8	0.34	37.7	0.20	34.8	0.36	36.9	0.26
Anhalt-Bitterfeld	Total Period	18.7	0.04	67.4	-16.68	20.0	-0.02	20.6	-0.19	23.2	-0.40	22.0	-0.28	20.8	-0.17
Sächsische Schweiz- Osterzgebirge	Total Period	23.0	0.55	327.6	-117.34	24.0	0.54	37.8	0.04	28.0	0.27	25.7	0.37	34.1	0.14

Total Period  $\triangleq$  Vorhersage am 01. April 2021 für die nächsten 50 Tage (bis 20. Mai 2021)



Bundesland Sachsen-Anhalt

Landkreis Halle



Kreis	Halle
Bevölkerung	238762

Bevölkerung in Bundesland Sachsen-Anhalt



# Webanwendung [Telegroum.com](https://telegroum.com)

Umsetzung

## Back-End

- PHP
- MYSQL

## Front-End

- Javascript
- Bootstrap
- Ajax

## Visualisierung

- Javascript-Bibliothek CanvasJS

## Verwendete Frameworks

- Visual Studio Code: Front und Back-End
- Filezilla: Dateitransport
- Plesk: Web Hosting and Server Data Center
- phpMyAdmin: Datenbank bearbeiten
- XAMPP: Webserver Apache und Datenbank MySQL



# Methoden

Aufgabe	Umsetzung
Datenaufbereitung & Datenexploration	NumPy, Pandas
Clustering Model	Scikit-learn, Scipy, GeoPandas
Forecast Model	Scikit-learn, TensorFlow, Keras, Statsmodels
Front-/Backend	PHP, JavaScript, MySQL, phpMyAdmin, FileZilla, Plesk, XAMPP, Visual Studio Code, Photoshop
Visualisierung	Matplotlib, Seaborn, Affinity Designer, CanvasJS
Präsentation	MS PowerPoint

# Fazit

- Einflussfaktoren auf die Dynamiken einer Pandemie schwer messbar
  - **Kausal:** Nachbarschaft, Bevölkerungsmobilität (Lockdowns), Verhalten der Gesellschaft
  - **Korreliert:** Suchanfragen([vgl. Lamos et al. 2021](#)), Verkaufszahlen Selbsttests / Medikamente
- Nicht alle Modelle für alle Vorhersagen geeignet
  - Zeitreihe **zuerst** auf Eigenschaften untersuchen
- Konzeptionelles Verständnis für die Datenvorbereitung essenziell
  - Vorsicht Verzerrung (**Bias**)
- Feature Engineering hat größeren Einfluss als Parameteroptimierung
  - Attribute **skalieren**