COVID-19-Pandemie

Projekt in

Introduction to Data Science

Eingereicht bei:

Prof. Dr. Thomas Kopinski

Fachhochschule Südwestfalen

Standort Meschede

Von
Julian Bornemann
Schöffenwiese 8
julian.bornemann@gmx.net / 02992 979123
30021747

Marsberg, den 11.10.2021

Die Daten

Diese Auswertung hat sich vor allem auf Deutschland und die USA als Vergleichsland bezogen. Somit wurden die täglichen Infektionszahlen, die vom Robert-Koch-Institut bereitgestellt wurden, genutzt. Sie wurden selbst gescraped.[1] Hierbei wurde, mit den erhobenen Daten bis zum 20.06.2020, mithilfe der linearen Regression ein Prognosemodell erstellt. Mit den prognostizierten Daten wurde dann die prognostizierte mittlere Wachstumsrate jedes Bundeslandes berechnet.

Des Weiteren wurden bereits von verschiedenen Gesundheitsämtern und Landkreisen erhobene Daten genutzt, um so eine Auswertung aller deutscher Landkreise zu schaffen.[2] Diese Daten wurden auch Vorverarbeitet um die mittlere Wachstumsrate der Covid-Infektionen und Tode in Folge durch Covid bis zum 20.06.2020 jedes Landkreises zu haben. Außerdem wurde jedem Landkreis, als zweite Kennzahl, die Bevölkerungsdichte (Personen pro km²) zugeordnet.[3] Bei einigen Landkreisen, wie der Hochsauerlandkreis wurden noch die täglichen Neuinfektionen und die täglichen Todeszahlen bestimmt. Danut sind diese Daten bereit für weitere Analyse.

Der letzte Datensatz besteht aus US-amerikanischen Counties, diese sind mit deutschen Landkreisen vergleichbar. Die bereits aufbereiteten Daten werden kostenlos von einer amerikanischen Non-Profit-Organisation namens „USAFacts“ bereitgestellt.[4] Auch hier wurde die mittlere Wachstumsrate der Covid-Infektionen für jedes County berechnet. Um diese Daten mit denen der Landkreise zu vergleichen wurde, als zweite Kennzahl, auch die Bevölkerungsdichte berechnet. Dies stellte sich jedoch als schwieriger heraus, da es dazu keine aktuellen verlässlichen Daten gibt. Somit wurde die Bevölkerungszahl von der letzten Volkszählung aus 2010 nehmen. Da in den USA kein Einwohnermeldeamt existiert, gibt es keine aktuelleren Daten. Damit wurde die Bevölkerungsdichte berechnet.

Weitere Vorverarbeitungsschritte wurden nicht gemacht, somit konnte die richtige Auswertung der Daten beginnen.

[1] https://de.wikipedia.org/wiki/COVID-19-Pandemie_in_Deutschland/Statistik#Tages-Statistik_der_gemeldeten_kumulierten_Inzidenz_nach_Bundesl%C3%A4ndern
[2] https://public.fusionbase.com/explore/covid19-germany/data
[3] https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/Administrativ/04-kreise.html
[4] https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

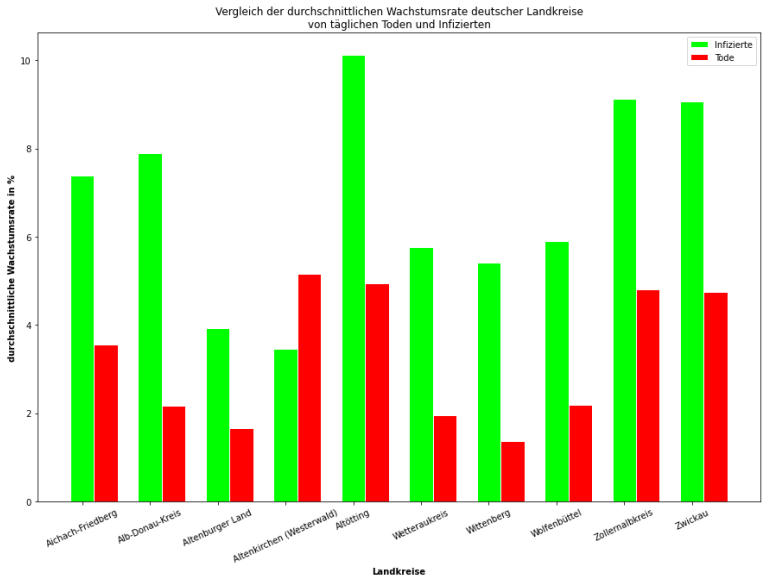
Auswertung:

Die Prognosen für die Bundesländer liegen im Durchschnitt bei 0,579%. Die Standardabweichung liegt bei 0,016. Der kleinste Wert ist 0,551% Schleswig-Holstein und der größte 0,622% Bremen. Das untere Quartil liegt bei 0,52 und das obere bei 0,584. Man kann Bremen als eine Art Ausreißer sehen. Hier sollte nochmal betont werden, dass diese Daten mithilfe der linearen Regression prognostiziert wurden.

Bei den Landkreisen wurden 280 Landkreise in Betracht gezogen. In jedem Landkreis durchschnittlich jeden Tag die Infizierten Zahl um 9,66 % an. Dabei liegt die Standardabweichung bei 16,1. Da der maximale Wert bei 152,11 % liegt und das obere Quartil bei 7,51 kann man davon ausgehen, dass die Daten durch Ausreißer verzerrt sind. Denselben Schluss kann man bei dem Wachstum der Toten ziehen.

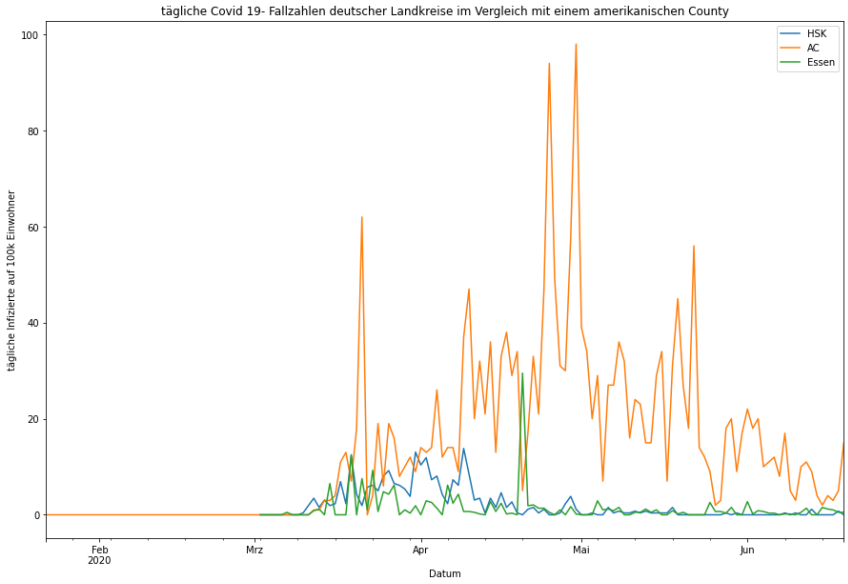
Für einen Vergleich mit den USA wurden 1607 Counties zum Vergleich gezogen. Hier lag im Durchschnitt die durchschnittliche Wachstumsrate nur bei 3,68 %. Was sich aber auch auf die Bevölkerungsdichte zurückführen lassen kann. Wo in Deutschland die Landkreise im Durchschnitt eine Bevölkerungsdichte von 341 Einwohner/km², haben Counties eine Bevölkerungsdichte von 91,15 Einwohner/km². Die Charakteristika des durchschnittlichen Wachstums ähneln sich hingegen wieder sehr mit den deutschen Pendants.

Nun wurden 10 zufällig ausgewählte deutsche Landkreise, mithilfe eines Säulendiagramms, verglichen. Bei den ausgewählten Landkreisen sieht man, dass die tägliche Wachstumsrate der Infizierten und Tode sich ähneln und es keinen wirklichen Ausreißer gibt.

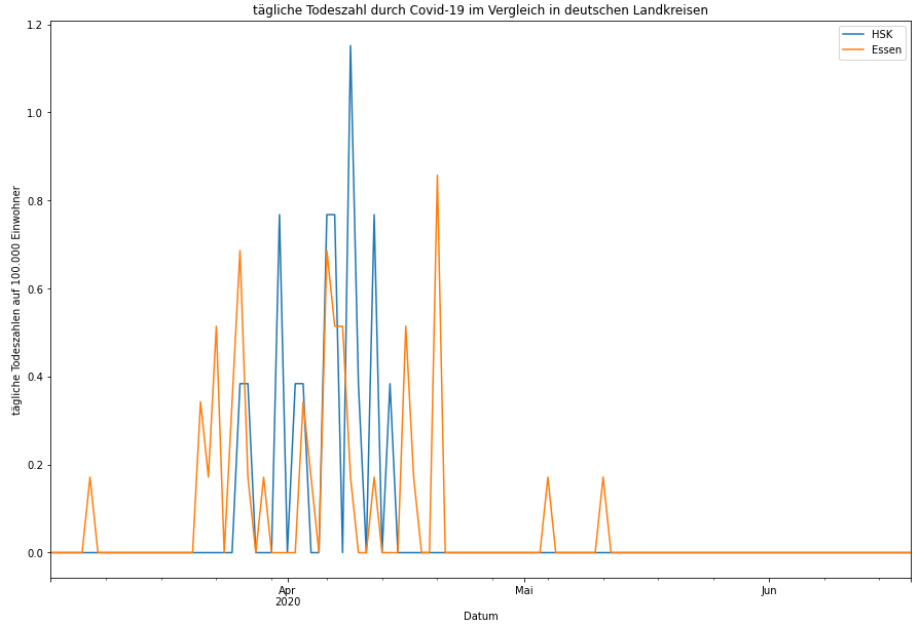


Mithilfe eines Liniendiagramms wurden die täglichen Covid 19 Fallzahlen auf 100.000 Einwohner visualisiert. Hier wurden der Hochsauerlandkreis und Essen als deutsche Beispiele genutzt und das Albany County als amerikanisches Beispiel. Das Albany County ist ein, im

Staat New York liegendes ländliche County. Die Datenerhebung fängt erst Anfang März richtig an. Dort kann man sehen, dass die Kurven von Essen und dem HSK sehr ähnlich verlaufen, nur dass Essen Ende März und der HSK erst Anfang April einen Corona Ausbruch hatten. Die Zahlen vom Albany County sind im Vergleich sehr hoch. Dies könnte möglicherweise darauf zurückführen, dass sie durch die Aufnahme von Infizierten, New York City entlastet haben. Da zu diesen Zeitpunkt New York City der größte Herd in den USA war.[5]

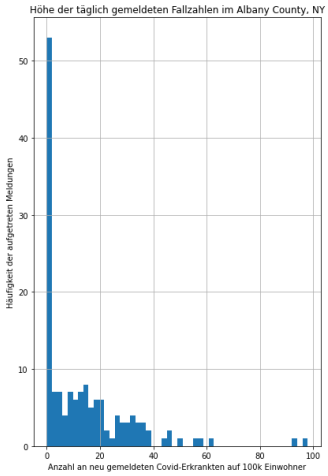


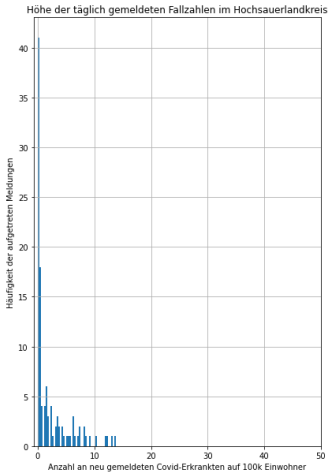
Auch die täglichen Tode auf 100.000 Einwohner wurden für den HSK und Essen visualisiert. Dies war für das Albany County nicht möglich, da die herangezogenen Daten keine Todesopfer beinhalteten. Man kann sehen, dass die herangezogenen Daten vergleichbar sind und es bei beiden Landkreisen hin und wieder nur ein paar Todesmeldungen gab.



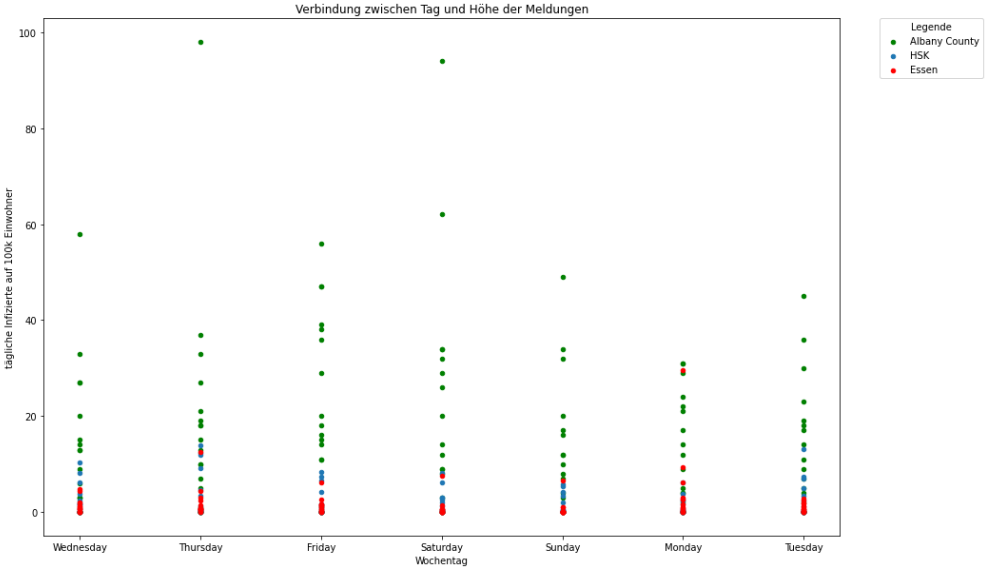
[5] https://www.nytimes.com/2020/05/27/opinion/coronavirus-morgue-trucks-nyc.html

Als nächstes wurde mithilfe eines Histogramms gezeigt, dass die meisten täglichen Meldungen im Albany County und im Hochsauerlandkreis eine sehr geringe Anzahl an neuen Infizierten hatten. Außerdem zeigt der Vergleich, dass im Albany County täglich höhere neue Fallzahlen gemeldet wurden als im HSK. Wobei es im HSK maximal zu Tagen mit 15 neuen Fällen kam gibt es im Albany County schon Tage, an denen bis zu 40 neue Infizierte gemeldet wurden.



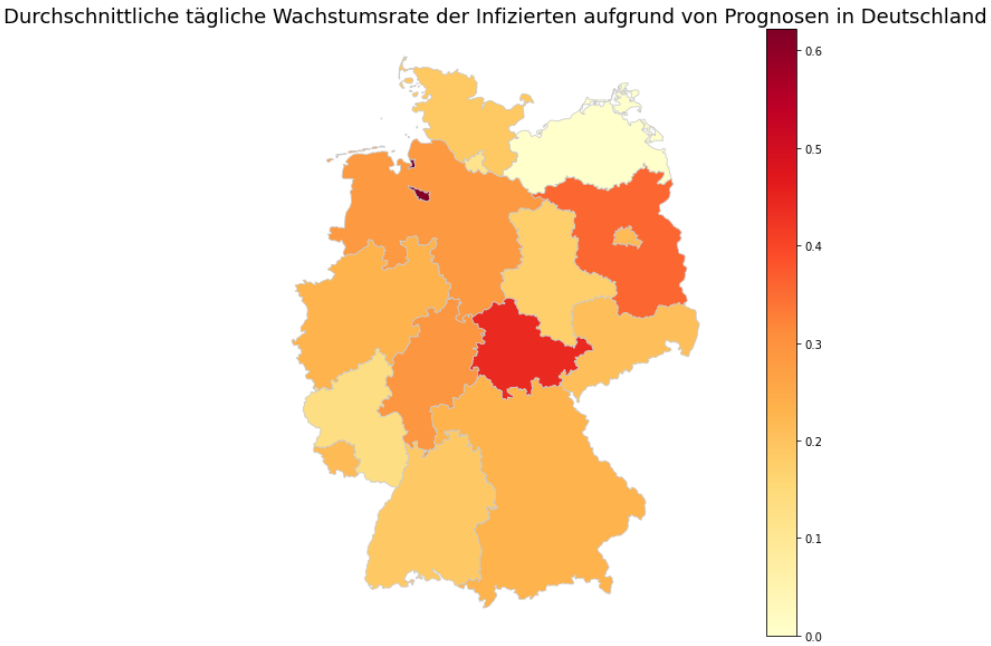


Mithilfe eines Streudiagramms wurde versucht einen Zusammenhang zwischen den täglich gemeldeten Infizierten auf 100.000 Einwohner und dem gemeldeten Wochentag nachzuweisen. Hierfür wurden wieder der HSK, Essen und das Albany County herangezogen. Jedoch kann man aus dem Streudiagramm keinen Zusammenhand erkennen, da die Punkte sich an jedem Tag gleich verhalten.

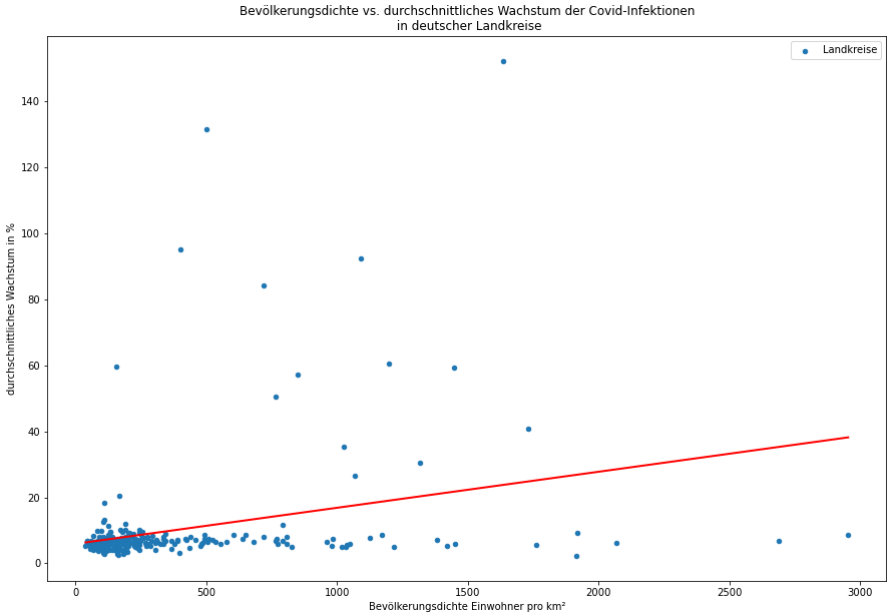


Des Weiteren wurde eine Deutschlandkarte angefertigt, in der die zuvor beschriebenen Daten zu den täglichen Wachstumsraten der Infizierten in den deutschen Bundesländern zum

Ausdruck kommen. Es ist kein Muster zwischen Ost und West oder regierenden Parteien erkennbar.

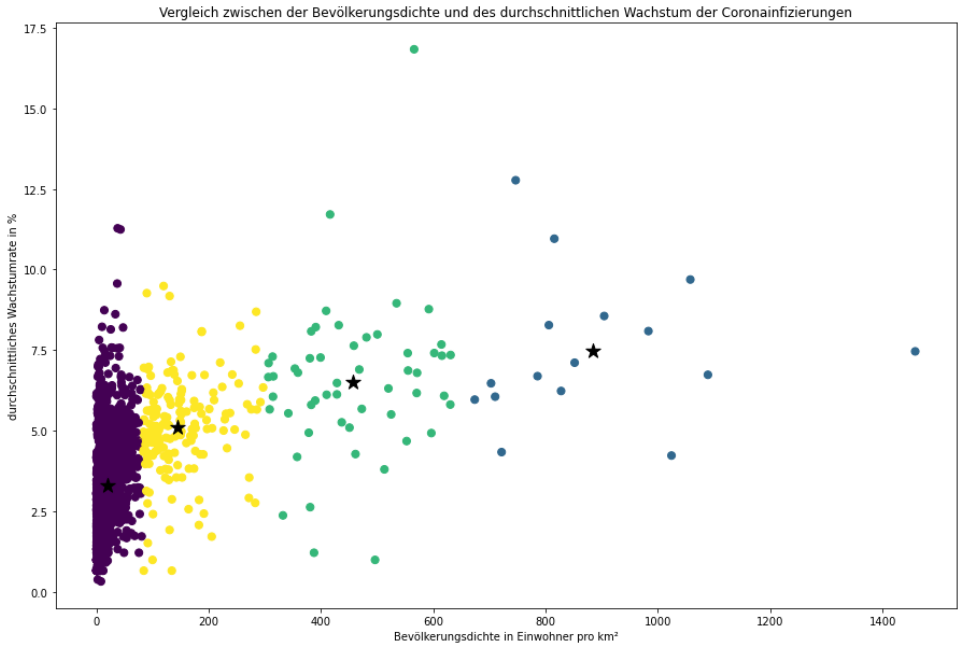


Als nächstes wurden in einem Streudiagramm die durchschnittliche Wachstumsrate und die Bevölkerungsdichte gegenübergestellt, um so einen Zusammenhang zu zeigen. Des Weiteren wurde eine Regressionsanalyse gemacht. Der p-Wert von 0,008 und die Korrelation von 0,2996 unterstützen die Hypothese nicht, jedoch ist bei der Trendlinie (5,93+0,011x) eine leichte Steigung erkennbar. Für die Counties lässt sich derselbe Schluss ziehen.

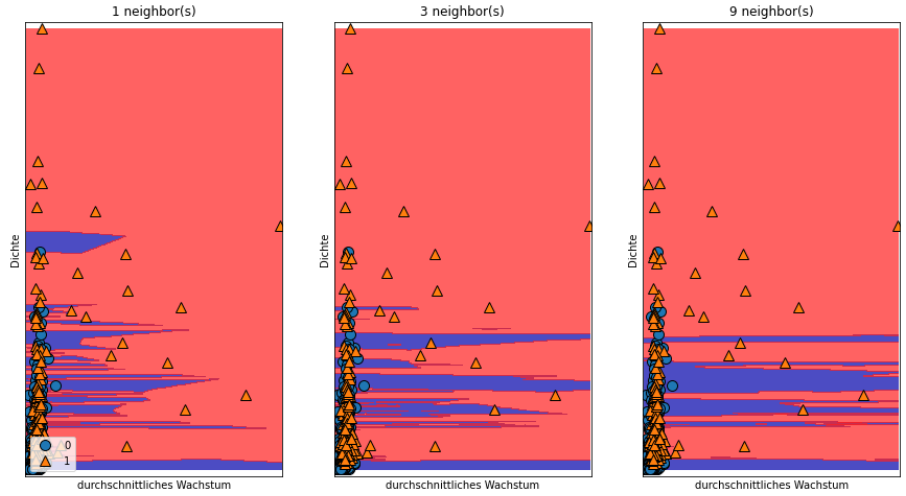


Danach wurde mit dem Cluster-Algorithmus KMeans versuch, bei den amerikanischen Counties, Cluster herauszubilden. Dies ist jedoch nicht geglückt, da die Daten für Clustering

ungeeignet sind. Es wurde noch versucht die Counties nach ihren Staaten zu gruppieren, aber auch dies führte zu keinem guten Ergebnis.

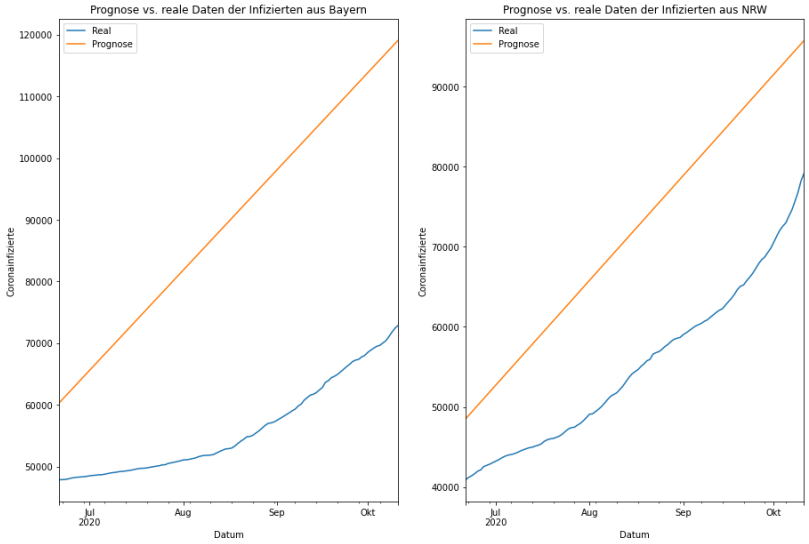


Danach wurde mithilfe des Nearest-Neighbor-Classifiers mit einer verschiedenen Anzahl -von Neighbors geschaut, ob der Algorithmus zwischen deutschen Landkreisen und Counties klassifizieren kann. Hier kam ein Score von 0.88723 für die Testdaten heraus. Dieser Wert ist nicht schlecht, was aber auch daran liegen kann, dass auf jeden Landkreis acht Counties kommen und hier somit ein Ungleichgewicht vorherrscht.



Als nächstes werden NRW und Bayern miteinander verglichen. Mithilfe der Prognosen und den dazugehörigen realen Daten wurden Liniendiagramme gebildet um die Daten miteinander zu vergleichen. Aus der Grafik kann man herauslesen, dass beide Bundesländer unter den Prognosen liegen, jedoch liegen die realen Daten von Bayern viel mehr unter den von NRW,

da die Fläche zwischen den realen und den prognostizierten Daten in Bayern viel größer ist. Dies lässt sich auch mithilfe des Integrals bestimmen, dazu fehlt aber die Funktionsgleichung der realen Daten. Hieraus kann man schließen, dass die Maßnahmen, die ab dem 20.06 von Markus Söder in Bayern ergriffenen wurden, viel erfolgreicher waren, als die Maßnahmen die Armin Laschet für das Land NRW getroffen hat.



Als letztes wurde der Hochsauerlandkreis mit Köln, in Punkten der täglichen Infizierten und Tode auf 100.000 Einwohner bis zum 20.06, verglichen. Die reinen Daten ähneln sich sehr. Mithilfe von Histogrammen zeigten sich auch hier gewisse Ähnlichkeiten, damit lassen sich keine Rückschlüsse auf den Unterschied zwischen ländlichen und urbanen Landkreisen ziehen.