DOKUMENTATION – STUDIENARBEIT INFORMATIONSVISUALISIERUNG

1. Wahl des Datensatzes:

Um einen geeigneten Datensatz zu finden, der für eine Visualisierung passend wäre, suchte ich zuerst auf dem Destatis-Portal, also der Website der Datenbank des statistischen Bundesamtes. Nachdem ich hier nicht fündig wurde, weitete ich die Suche nach Daten auf globale Datensätze aus. Hier fand ich einige github Repositories, die vielversprechend waren. Ich fand einen Datensatz mit Angaben über die Längen- und Breitengrade der jeweiligen Länder über die die Daten mit Infizierten und Todesfällen aufgeführt waren, aber nicht ihre ISO Codes. Hier kam mir die Idee, die Längen- und Breitengrade auf eine Weltkarte zu mappen und auf der Weltkarte die Daten der Länder anzeigen zu lassen. Darum fragte ich in der letzten Vorlesung nach, wie so ein Visualisierung zu bewerkstelligen wäre. Nachdem ich das Feedback erhielt, dass die einfachste Lösung hierfür eine svg-Datei einer Weltkarte mit ISO Codes zu benutzen und über die ISO Codes des Datensatzes die Daten auf die Länder zu mappen. Also suchte ich nach einem Datensatz mit ISO Codes und wurde auf dem github Profil von „Our World in Data“ fündig, die einen täglich aktualisierten Datensatz führen, der die ISO Codes der Länder über die die Daten stammen enthält. Im master branch dieses Repositories (https://github.com/owid/covid-19-data/tree/master/public/data) gab es drei Dateitypen. Einmal eine csv-Datei, eine json-Datei und eine xlsx-Datei. Ich entschied mich für die csv-Datei, da ich gewöhnt war mit csv-Dateien zu arbeiten.
Ich nahm die Datei speziell vom 02.07.2020, da diese Dateien täglich aktualisiert werden und ich deswegen mein notebook und die website dynamisch an die Daten anpassen müsste. Deshalb entschied ich mich für einen statischen Datensatz. Ich entschied mich für diesen Datensatz, da er einerseits die ISO Codes der jeweiligen Länder enthält und ich diese brauchte, um meine Visualisierung durch eine Weltkarte zu verwirklichen; andererseits weil der Datensatz sehr gut strukturiert ist für eine Visualisierung, da jeder Tag eines Landes vom 31.12.2019 bis zum 02.07.2020 eine eigene Reihe hat, die die Daten der gesamt Infizierten, der gesamten Todesfälle, usw. beinhaltet.

```
iso_code,continent,location,date,total_cases,new_cases,total_deaths,new_deaths,total_cases_per_million,new_cases_per_million,total_deaths_per_mi
AFG,Asia,Afghanistan,2019-12-31,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-01,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-02,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-03,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-04,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-05,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-06,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-07,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-08,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-09,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-10,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-11,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-12,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-13,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-14,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-15,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-16,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-17,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-18,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-19,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-20,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-21,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-22,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-23,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-24,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-25,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-26,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-27,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
AFG,Asia,Afghanistan,2020-01-28,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,,,,,,0.0,38928341.0,54.422,18.6,2.581,1.337,1803.987,,597.029,9.59,,,37.746,0.5,64.83
```

Abbildung 1: Ausschnitt des Datensatzes

2. Data Transformation mit Jupyter Notebooks:

Nach dem Einbinden der benötigten Bibliothekten für die Visualisierung mit Jupyter Notebooks und dem Einbinden des Datensatzes erstellte ich ein erstes Diagramm das mir die mittleren gesamten Todesfälle pro Land mit Varianz anzeigt. Hier kann man sehen, dass an letzter Stelle im Datensatz eine Kategorie der kumulierten Daten aller Länder steht. Diese Kategorie hat den ISO Code „OWID_WRL" und als Landesnamen „World", also werden hier Daten der ganzen Welt gezeigt. Ich entschied mich, mich im Verlauf des Notebooks auf diese Kategorie zu beschränken, da hier allgemein für die ganze Welt kumulierte Daten sind, und ich keine einzelnen Länder auswählen müsste, um sie zu visualisieren, die möglicherweise ein verfälschtes Bild über die ganze Lage geben.

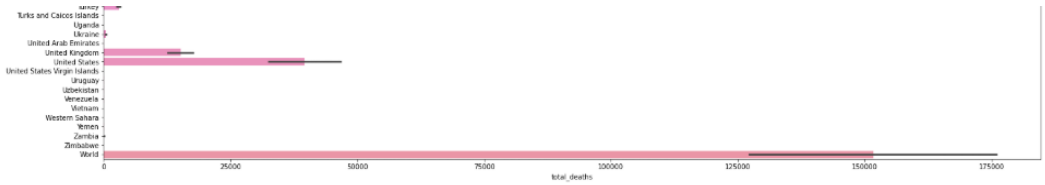

Abbildung 2: Letzte Reihen des Diagramms

Als nächstes visualisierte ich die ersten Spalten des Datensatzes in Bezug auf diese Weltdaten mit Lineplots. Die ersten Spalten sind die insgesamten Todesfälle (kumuliert), die insgesamten Infizierten (kumuliert), die neuen Todesfälle pro Tag und die neu Infizierten pro Tag. Nach diesen Graphen suchte ich nach neuen Spalten, die visualisierbar sein könnten, und entschied mich dafür den Datensatz dafür anzupassen, da es zu viele Spalten gab mit nicht relevanten Informationen, wie zum Beispiel Krankenhaus Betten pro Infiziertem pro Land. Ich löschte alle Spalten mit nicht relevanten Informationen und füllte alle Angaben von NaN mit 0 auf. Übrig blieben mir die oben genannten schon visualisierten Spalten und die gesamt Infizierten pro Millionen Menschen, die neu Infizierten pro Millionen Menschen pro Tag, die gesamten Todesfälle pro Millionen Menschen, die neuen Todesfälle pro Millionen Menschen pro Tag und die Bevölkerungsdichte. Von diesen Spalten erhoffte ich mir, dass die ersten vier einen Einblick in den Anteil der Bevölkerung geben, die Infiziert sind, und von der letzten, dass sie zeigt ob die Bevölkerungsdichte eine Rolle in der Verbreitung des Virus spielt.

	isocode	continent	location	date	total_cases	new_cases	total_deaths	new_deaths	total_cases_per_million	new_cases_per_million	total_deatl
0	AFG	Asia	Afghanistan	2019-12-31	0.0	0.0	0.0	0.0	0.000	0.000	
1	AFG	Asia	Afghanistan	2020-01-01	0.0	0.0	0.0	0.0	0.000	0.000	
2	AFG	Asia	Afghanistan	2020-01-02	0.0	0.0	0.0	0.0	0.000	0.000	
3	AFG	Asia	Afghanistan	2020-01-03	0.0	0.0	0.0	0.0	0.000	0.000	
4	AFG	Asia	Afghanistan	2020-01-04	0.0	0.0	0.0	0.0	0.000	0.000	
...	...	...	...	...	...	...	...	...	...	...	
26058	OWID_WRL	0	World	2020-06-28	9953229.0	181386.0	498550.0	4633.0	1276.906	23.270	
26059	OWID_WRL	0	World	2020-06-29	10113462.0	160233.0	501597.0	3047.0	1297.463	20.556	
26060	OWID_WRL	0	World	2020-06-30	10273424.0	159962.0	505309.0	3712.0	1317.985	20.522	
26061	OWID_WRL	0	World	2020-07-01	10465987.0	192563.0	511045.0	5736.0	1342.689	24.704	
26062	OWID_WRL	0	World	2020-07-02	10665758.0	199771.0	515973.0	4928.0	1368.317	25.629	

26063 rows × 13 columns

Abbildung 3: Angepasster Datensatz

Die neuen Spalten visualisierte ich wieder mit Lineplots, da das die beste Wahl in Bezug auf Datumsangaben ist. Die Bevölkerungsdichte kombinierte ich mit den vorherigen Spalten durch Scatterplots, um einen möglichen Zusammenhang zwischen der Bevölkerungsdichte und den acht vorherigen Spalten zu zeigen.
Die Merkmale, die aus diesen Graphen hervorgingen sind einerseits, dass es besser ist in die Visualisierung mit d3 Datumsangaben miteinzubeziehen und andererseits, dass die Bevölkerungsdichte nicht genug Informationen

beinhaltet, um sie in der Visualisierung mit d3 zu benutzen. Somit wird die Visualisierung mit d3 eine Weltkarte mit gemappten Daten aus dem veränderten Datensatz sein, der als Anhaltspunkt das Datum des jeweiligen Tages benutzt.

3. Visual Mapping und Visualisierung mit d3.js:

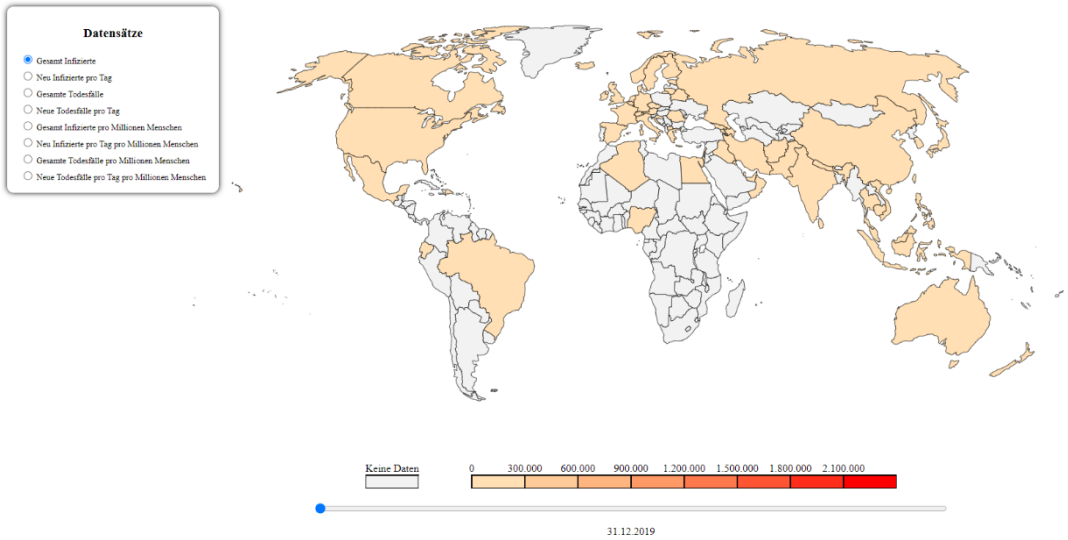

Abbildung 4: Fertige Visualisierung mit d3

3.1 Visual Mapping:

Das Datum wurde in einen Slider gemappt, der das Datum vom 31.12.2019 bis zum 02.07.2020 anzeigt und entsprechend eingestellt werden kann. Dies ist die effizienteste Lösung, dem Benutzer die Daten des jeweiligen Tages zu zeigen, da der Benutzer selbst den gewünschten Tag auswählen kann, zu dem er Daten sehen möchte. Die Spalten wurde in ein Auswahlmenü umgewandelt, in dem man wählen kann, welche Spalte auf die Weltkarte gemappt werden soll. Durch das Auswahlmenü wurde mehr Interaktivität mit der Website ermöglicht, und außerdem der komplette Datensatz, außer die Bevölkerungsdichte, mit wenig Aufwand auf der Website untergebracht. Die Daten der einzelnen Länder wurden auf die Länder durch Farben gemappt und unterhalb der svg Graphik durch eine Legende erklärt. Je dunkler die Farbe desto höher

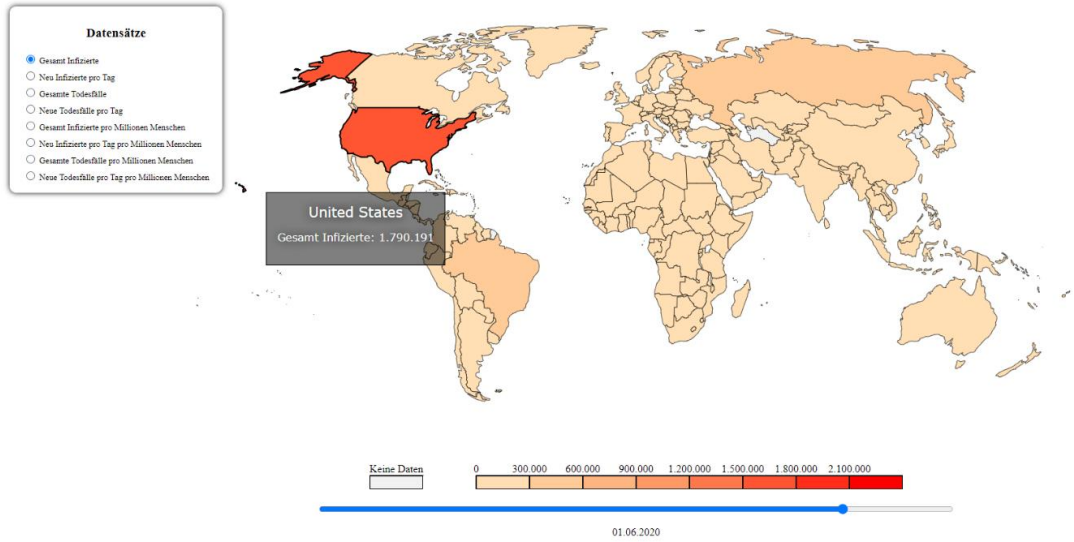

Abbildung 5: Genaue Informationen über die Länder durch Hovern

ist die Zahl, die in der bestimmten Reihe und Spalte des bestimmten Landes steht. So wurde ein einfaches und schnelles Verständnis der Daten des jeweiligen Landes gewährleistet. Außerdem können immer noch die genauen Daten des Datensatzes aufgerufen werden, indem der Benutzer über das bestimmte Land hovert, über das er Informationen möchte.

3.2 Visualisierung mit d3:

Die Visualisierung beginnt, indem einer der Radio Buttons des Auswahlmenüs vom Benutzer gedrückt wird. Standardmäßig wird der erste Button ausgewählt, der die Spalte der Daten der gesamten Infizierten beinhaltet. Wenn einer dieses Buttons ausgewählt wurde, startet eine Funktion, die die Farbe der Visualisierung und die Skalierung der Legende unterhalb der svg Graphik festlegt. Außerdem wird die ausgewählte Spalte in einer Variablen festgehalten. Dann startet die d3 Visualisierung. Zuerst wird der Datensatz eingebunden und in einer globalen Variablen gespeichert, um den Datensatz für andere Funktionen außerhalb des Bereichs der d3 Funktion zugänglich zu machen. Danach werden für jede Spalte des Datensatzes, der visualisierbare Daten enthält das Minimum und Maximum berechnet und in globalen Variablen gespeichert. Dann werden durch d3 zehn Rechtecke an ein div unterhalb der svg Datei angehängt. Die Rechtecke bekommen die Färbung der vorher festgelegten Spalte wobei die Sättigung und Färbung bei jedem Rechteck verringert wird anhand des HSL Modells, damit eine farbliche Abstufung in der Legende entsteht. Die Rechtecke bekommen Event Listener, sodass, wenn man mit der Maus über eines dieser Rechtecke hovert, alle Länder gezeigt werden, die dieselbe Färbung wie dieses Rechteck haben, der Rest wird ausgegraut.

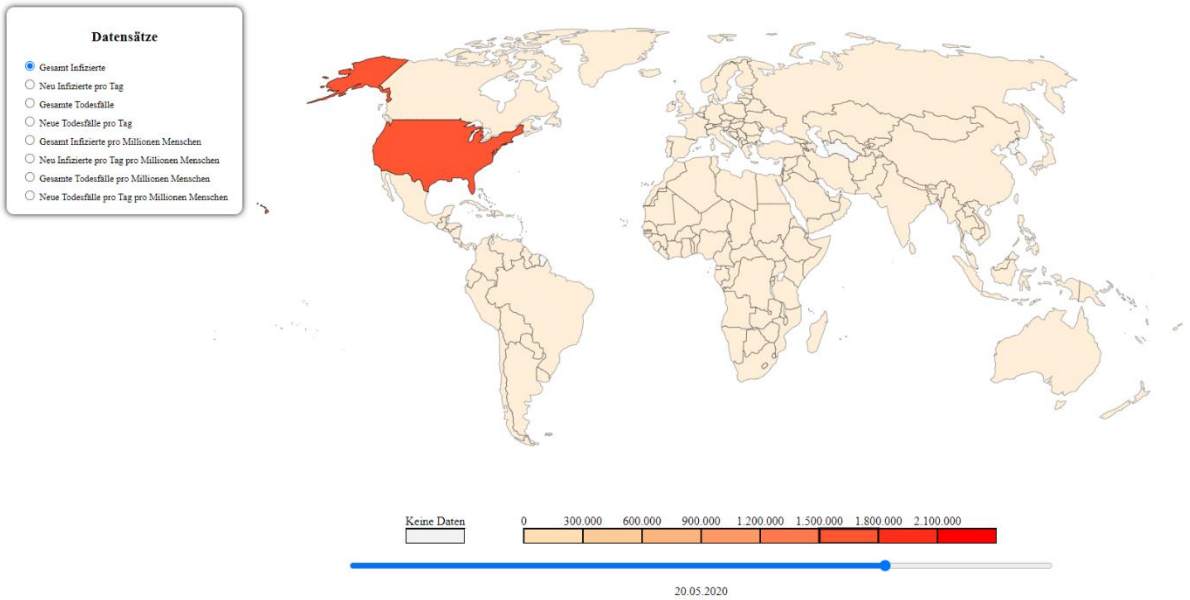

Abbildung 6: Anzeige beim Hovern über das Rechteck mit gleicher Farbe

Auch jedes Land bekommt einen Event Listener, der die Umrandung des Landes breiter darstellt und genaue Informationen des Landes anzeigt (Abbildung 5). Danach wird eine weitere Funktion aufgerufen, die den ganzen Datensatz nach dem Datum durchsucht das unter dem Slider steht. Wenn die Funktion ein Datum findet, speichert es die Daten der gewollten Spalte dieser Reihe in eine Variable und vergleicht in welchem Bereich der Legende der Wert der vorliegenden Variablen liegt und färbt das Land in der dementsprechenden Farbe der Legende. Jedes mal wenn der Slider verändert wird, wird diese Funktion aktualisiert. Somit hat jedes Land immer die richtige Farbe für den richtigen Bereich und der Benutzer hat den kompletten Datensatz auf einer Website visualisiert.

4. Satz von Tufte:

„Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space."

Greatest number of ideas in the shortest time: Durch die svg Graphik sieht der Benutzer sofort, dass es sich um eine Visualisierung handelt, die sich um die ganze Welt dreht. Die Legende unten gibt Anhaltspunkte in welchem Zahlenbereich sich jedes Land aufhält, da jedes Land dieselbe Farbe wie das dazu gehörende Rechteck besitzt. Den Namen des Datensatzes der visualisiert wird kann der Benutzer im Auswahlmenü herauslesen oder über die Länder hovern, die selbst Informationen über den Namen des benutzten Datensatzes (oder besser gesagt der Spalte des Datensatzes) geben und die Daten des dazu gehörenden Landes ausgeben. Auch kann der Benutzer am Slider erkennen in welchem Tag er sich befindet, da das Datum unten geschrieben steht.

With the least ink: Die vier Bestandteile der Visualisierung sind die svg Graphik, die Legende, das Auswahlmenü und der Slider zum Einstellen des Datums. Daraus lassen sich alle Informationen herauslesen mit sehr wenigen Bestandteilen der eigentlichen Visualisierung.

In the smallest space: Die gesamte Visualisierung nimmt nur eine Seite einer Website ein, da sich beim Wechseln der zu Visualisierenden Spalten nur die Farbe und die Legende ändert. Deshalb braucht diese Visualisierung nur den mindestens nötigen Platz, um den kompletten Datensatz auf eine Seite zu bringen.