

## Data Science 2

Sommersemester 2021

### Übungsblatt 5

Aus Krankheitsgründen müssen die Vorlesung und die Übung diese Woche leider ausfallen. Zur Vertiefung gibt es daher nochmal ein Wochenübungsblatt zum Thema Zeitreihen und Pandas.

#### Aufgabe 1 \* (RKI Zahlen)

Wir hatten im ersten Teil des Kurses ja bereits den Datensatz des RKI zu den Corona Fallzahlen betrachtet. Darin enthalten sind die Fälle von Neuinfektionen, u.a. nach Altersgruppen, Bundesland, usw. Einen recht aktuellen (nicht tagesaktuellen) Export der RKI-Zahlen finden Sie unter

`Kurse/DataScience2/data/rki-covid19-2021-05-01.csv`

In dieser Aufgabe geht es darum, dass Sie sich nochmal etwas mit der Zeitreihen-Darstellung von Pandas auseinandersetzen. Dazu betrachten wir vereinfacht lediglich die Spalten **AnzahlFall** und **Meldedatum**.

1. Der Datensatz ist relativ groß (viele Zeilen), weil die Aufteilung in Bundesländer, Landkreise, usw. natürlich sehr weit auffächert. Wir sind an den Fallzahlen in der Spalte **AnzahlFall** interessiert.

Summieren Sie die Anzahl der Faelle nach Meldedatum auf!  
Wie groß ist der daraus entstandene Datensatz?

2. Auf dem gruppierten Datensatz wollen wir jetzt mit den Zeitreihen-Funktionen arbeiten. Wandeln sie dazu die Spalte **Meldedatum** in ein **datetime64**-Typ um.
3. Erstellen Sie einen Plot, der den Verlauf der Neuinfektionen tageweise angibt.
4. Benutzen Sie die Funktion **rolling** von **Series**-Objekten um einen 7-Tagesdurchschnitt der Neuinfektionen zu plotten.
5. Erstellen mehrere Line-Plots mit dem Verlauf der Neuinfektionen - einen Plot je Altersgruppe (dafür müssen Sie ggf. nochmal vom ursprünglichen Datensatz aus starten.).

Die Plots sollen in *einem* Chart dargestellt werden.