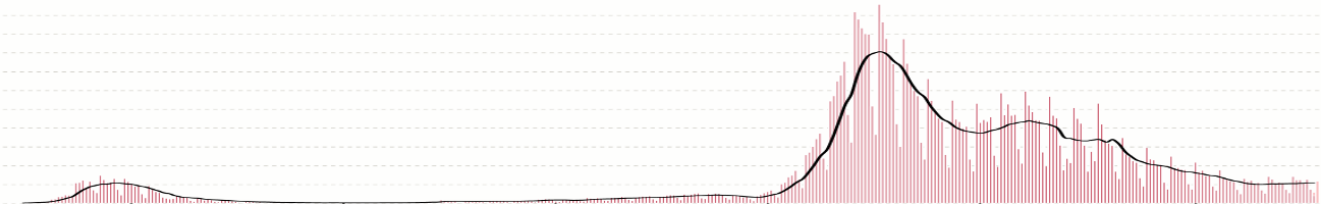
Data Wrangling (daw)
Cédric Huwyler
FS2021

Fachhochschule
Nordwestschweiz

Minichallenge: Covid-19 - Datenaufbereitung



Du bist Praktikantin bzw. Praktikant bei einer internationalen Grossfirma, die stark von den wirtschaftlichen Folgen von Covid-19 betroffen ist. Da die Möglichkeit eines weitereren Soft-Lockdowns in der Schweiz und Restriktionen in anderen Ländern bzw. auch Einreisebeschränkungen die gewählte Unternehmensstrategie stark beeinflussen, soll ein Dashboard der täglichen neuen bestätigten Fälle und der Anzahl Verstorbenen aufgebaut werden, damit die Geschäftsrisiken abgeschätzt werden können. Insbesondere sollen damit auch die Daten für weitergehende Modellierungen bereits bereitstehen. Es sollen konkret die Daten der John Hopkins - Universität verwendet werden. Da ein grosser Teil des Geschäfts deiner Firma auch in der Schweiz stattfindet, sollen die bestätigten Fälle und Todesfälle aus einer lokalen Quelle ebenfalls pro Kanton visualisiert werden. Du wirst angefragt, das dafür nötige Data Wrangling zu erledigen und für die Entwickler ein Jupyter-Notebook mit einem Proof-of-Concept abzugeben. Konkret bekommst du folgende Vorgaben:

Datenquellen:

- Die Daten der John Hopkins - Universität sollen aus den täglichen Rohdaten aus Github (Link) entnommen werden (**Achtung:** Auf dem Github-Repository finden sich auch die bereits aggregierten Daten; hier sollen die hier verlinkten Rohdaten verwendet werden).
- Die Daten der einzelnen Kantone sollen statt vom BAG von den offenen Behördendaten des Kantons Zürich bezogen werden, da dort die Daten oft genauer sind. Ein entsprechendes CSV-File ist über Github (Link) verfügbar.

Zahlen, die täglich pro Land geliefert werden sollen:

- total bestätigte Fälle pro Land seit Anfang der Epidemie
- neue bestätigte Fälle pro Land seit Anfang der Epidemie
- total bestätigte Todesfälle pro Land seit Anfang der Epidemie
- neue bestätigte Todesfälle pro Land seit Anfang der Epidemie

Zahlen, die täglich pro Kanton geliefert werden sollen:

- total bestätigte Fälle pro Kanton seit 1. Juni 2020 (Beginn 2. Welle)
- neue bestätigte Fälle pro Kanton seit 1. Juni 2020
- total bestätigte Todesfälle pro Kanton seit 1. Juni 2020
- neue bestätigte Todesfälle pro Kanton seit 1. Juni 2020

Die Historie der Kennzahlen der einzelnen Länder soll graphisch dargestellt werden können und ebenfalls die Kennzahlen der einzelnen Kantone (aber erst seit Beginn der 2. Welle). Ausserdem sollen die Plots mit den neuen bestätigten Fällen noch gestrichelt eine geglättete Variante der Kurve anzeigen, damit der weitere Verlauf etwas besser abgeschätzt werden kann.

Du sollst dazu eine Data Preparation Pipeline bauen, die jeden Tag neu automatisiert ausgeführt werden kann. Da die Datenmenge klein ist und die Daten manchmal rückwirkend korrigiert werden, soll sie jeweils alle Daten noch einmal von den besagten Quellen herunterladen.

Der Senior Data Scientist gibt dir ausserdem einige Tipps für das Vorgehen:

- Die CSV-Files der John Hopkins - Universität ziehst du am besten von Github, indem du entweder ein Skript schreibst, das dir das Repository klont oder indem du einfach die Struktur der Dateinamen erfasst und `pd.read_csv` und `pd.date_range` benutzst, um die einzelnen Files herunterzuladen, zu öffnen und als grosses Data Frame aneinanderzuhängen.
- Die Daten von openzh enthalten noch zusätzlich Daten für das Fürstentum Lichtenstein, die kannst du entfernen.
- Für die Software-Engineers reicht ein Proof-of-Concept der Plots für jeweils ein Land / einen Kanton, dann können sie für das Dashboard alles selbst automatisieren.
- Der Output deiner Data Preparation Pipeline soll aus den zwei sauber aufbereiteten Data Frames der beiden Datenquellen bestehen.

Du schätzst den Aufwand dafür auf einen halben Tag und machst dich sofort ans Wranglen.