# Stappenplan integratie externe apotheekdata

## Voorbereidingen (30’):

**Kaggle API voorbereiden:**

1. Installeer Kaggle lokaal (opgelet – installeer dit binnen de juiste omgeving – was bij mij eerst niet in mijn Anaconda-omgeving geïnstalleerd zodat het niet gevonden werd)
2. Maak Kaggle-account aan, haal persoonlijke token op voor API (JSON-file), en sla op in de map .Kaggle (onzichtbare mappen hiervoor zichtbaar maken)

*Ik heb hier wat moeten zoeken, ik ben meer gewend om op Windows zo’n zaken te doen, mijn Mac gebruikte ik tot nu toe meer voor persoonlijke zaken en daar moet ik niet zo diep duiken in de computer. Ik kende ook Kaggle niet dus moest dat ook even uitzoeken.*

## Bekijken file (10’)

Omzetten van .csv naar .xlsx en kort scannen van de file (handmatig)

## Validaties

## Script voor ophalen en archivering (1u30’)

**01\_download\_pharmacies.py**Python-script dat maandelijks geïnitieerd wordt via Cron – file wordt gedownload en in ruwe versie opgeslaan.

* Doellocatie: ~/Testje\_Baldwin/OUTPUT\_Pharmacies
* Naam: Pharmacies\_mostrecent.csv

**02\_save\_ruw\_YYYYMMDD.py**

Dupliceer Pharmacies\_mostrecent.csv naar een nieuwe file met bestandsnaam Pharmacies\_ruw\_YYYYDDMM. Ik wis of vervang \_Pharmacies\_mostrecent niet, beide files worden behouden omdat dit later soms van pas komt als er fouten optreden, dan heb je nog zicht op de ruwe export.

**03\_make\_masterfile.py**

Sommige apotheken kunnen **verdwijnen** uit de dataset, maar we willen ze **toch bewaren** in onze databank – ik gebruik hier ID als key.   
*Side note – ik ga ervan uit dat de gegevens in de file zoals die online staan, correct zijn. Ik heb geen stappen ingebouwd om inhoudelijke inconsistenties te corrigeren, bv. ontbrekend adres, veranderd adres. Ik ga ervan uit dat wat meest recent op Kaggle staat, correct is, of op zijn minst als dusdanig verwerkt moet worden.*

Verwerking - Longitude en Latitude toevoegen (20’+10’):   
  
~~Ik doe dit eenmalig voor alle adressen op de allereerste file – via OpenStreetMap, lijkt me financieel de beste keuze. Daarna doe ik dit obv de meest recente file, en vul enkel de ontbrekende waarden in, om kosten te besparen voor opvragen long/lat data en om uitvoering script minder zwaar te maken.~~

Omdat het gevraagd werd dit toe te voegen ging ik er van uit dat dit er nog niet in stond, maar longitude en latitude bleken er al in te staan zag ik daarna pas… Ik heb voor deze variabelen de variabelennamen aangepast (jullie noemen dit wellicht anders dus ik ga mijn woordenschat nog wat moeten aanpassen) en ze ook aangepast naar numerieke latitude/longitude zodat je er meteen mee kunt rekenen of plotten.

Hier evt nog stukje script bijschrijven voor wat te doen met long/lat gegevens die niet gelezen en getransformeerd konden worden, zodat die ergens in een aparte file komen te staan gelinkt aan ID en we iets kunnen schrijven om ook deze te corrigeren via apart script.

## Voorkomen fouten door gelijktijdig schrijven (15’)

Hou rekening met het feit dat **meerdere imports** en **meerdere bijdragers** in dezelfde codebase ook apotheekdata kunnen schrijven.

* Atomic-write stap ingevoegd zodat er geen half geschreven bestanden achterblijven
* De verwijderstap is ingekapseld in try/except, voor als twee processen tegelijk de verwijder-check zouden uitvoeren.
* File lock ingevoerd.

Dit is nieuw voor mij, in het verleden hadden wij gewoon tussen collega’s afspraken wie wanneer wat deed.

## Validaties (10’)

Mogelijke stap die ik nog zou kunnen toevoegen is om het bestand te gaan cleanen zodat er enkel nog de info inzit die we echt nodig hebben. Persoonlijk doe ik dit liever pas na het binnenhalen en archiveren. Zo kunnen we altijd terug naar de brondata.   
  
Aanpassingen die ik zou doen voor finale cleaning om tot een werkbaar bestand te komen:

* Overbodige, lege variabelen wissen (oa NAISDESCR tem PROVID\_105)
* Heel wat variabelen hebben geen naam – zou ik een naam geven zodat we ze kunnen benoemen/gebruiken.

## Verplaatsen files en uploaden GIT (20’)

Ik ben hier nog onvoldoende vertrouwd mee om dit nu nog verder uit te zoeken. Ik begrijp hoe het werkt maar ik veronderstel dat dit niet de plek is om ook de output-files in te (laten) zetten. Ik deel daarom enkel de python-scripts hier en deel daarnaast ook mijn folder via Finder.

## Nakijken en aanpassen van dit document (15’)