Commercial Property Data Processing:

|  |  |
| --- | --- |
| **Stakeholder** | Commercial Property Team |
| **Business contact** |  |
| **Duration** | 1 Day |

# Sprint 01: Structuring Data

## Key findings:

1. Bullet list summary of key findings.

## Sprint Aim:

To research all possible resources that can be used for an extraction and sorting pipeline. Furthermore developing a prototype organising system to prep the data for extraction.

## Sprint Objectives:

1. Research all possible resources and explain their role in helping to create a solution
2. Sort and structure the data in preparation for modelling

## Research:

**Data to be used**

CMS.gov

CMS provides data and statistics on various Medical practices that are signed with the organisation. The data I am using is information on prescription drugs prescribed by individual physicians and other health care providers that has been paid for under the Medicare Part D Prescription Drug Program.

Both files “PartD\_Prescriber\_PUF\_NPI\_DRUG\_17” and “PartD\_Prescriber\_PUF\_NPI\_17” – we will call this one Meta data - will be used for the prediction system. One containing prescriber information, the other containing Metadata.

NPPES API

This API can be used to provide data on the registered Physician allowing us to check if they have been registered with us before. This data can be used to check if we have clients who we may expect to receive a claim if they are likely to prescribe Opioids.

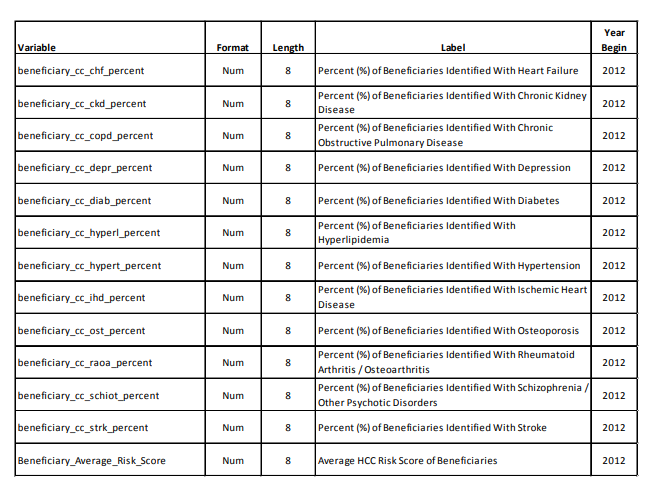
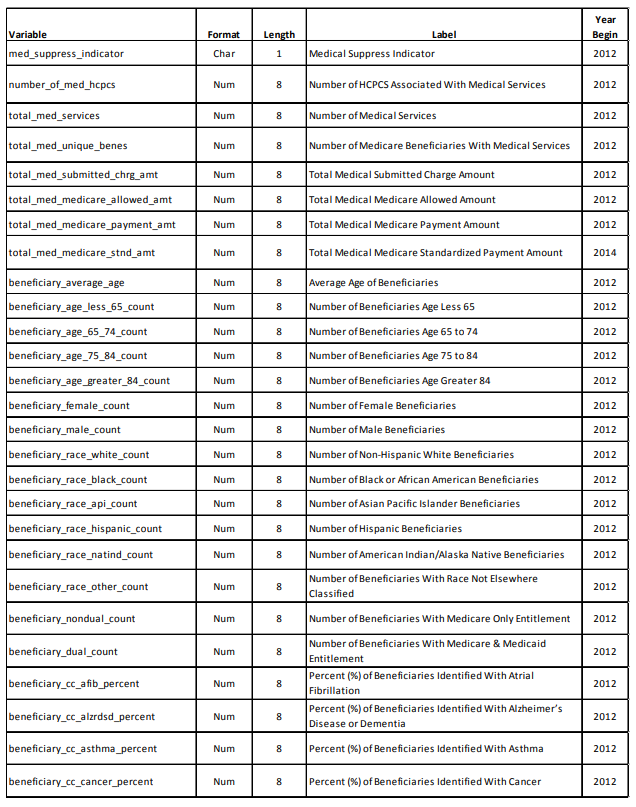
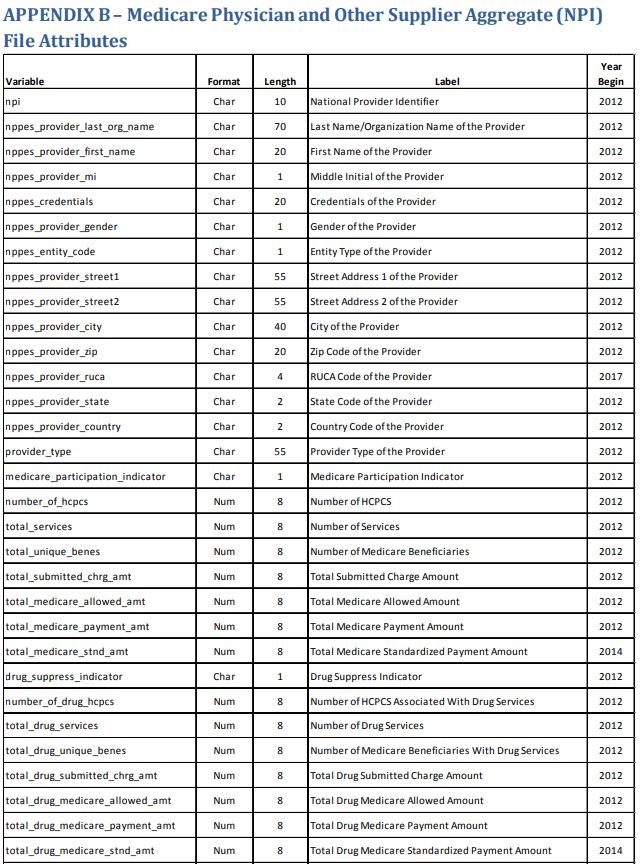
**Factors to consider**

CMS provides data that has been collected over a large period of time. This will no doubt take a long time to process which is why the latest release of data is 2017’s information which was released in 2019.

New medication is not put into markets on a frequent basis (due to the medication testing stages required before being sold) therefore a gap of 2 years in data is not so much of a problem.

Prescriber and Meta Data Variables

Below is a list of data that is within the Prescriber and Meta file data



**Wider Applications**

* + The data includes NPI numbers which may be in our systems if we have insured the practice/individual

## Data Prep - System

## Method:

1. Import all libraries needed (pandas)
2. Load data
3. Replace all dashes in drug\_names with periods
4. Group data by npi, last name, first name and drug name
5. Run through chosen directory and look at file extensions, depending on their extension, create a new directory based on the extension key
   * If the key is already created, just move it into the file
6. If the file type isn’t in the directory dictionary, them create a nw directory called other ad store the file in there
7. Let the above function run through every subfolder within the chosen directory

## Results:

1. Research carried has been out. we will be using;
   1. Tika to extract information from pdf files
   2. Pytesseract to extract data from image based files
   3. Extract\_msg to extract data and attachments from emails
2. The “File organisation” system has been created and is working

**Sources**

Text Detection: <https://www.pyimagesearch.com/2018/08/20/opencv-text-detection-east-text-detector/>

Text Recognition: <https://www.learnopencv.com/deep-learning-based-text-recognition-ocr-using-tesseract-and-opencv/>

OpenCV vs other deep learning frameworks: <https://www.learnopencv.com/cpu-performance-comparison-of-opencv-and-other-deep-learning-frameworks/>