Consonance EKG Task

Michal Pietruszka

03.12.2017

1. Scraping/Parsing code description

- To gather and parse the data provided in the NFZ webpages, I used the following libraries, dependent on Python 3:

  BeautifulSoup from bs4
  urlopen from urllib.request
  unicodedata

- The main data-scraping procedure consisted of iterating over the ids on the webpage, where the 'id=' was appended accordingly:

  https://prog.nfz.gov.pl/app-jgp/AnalizaPrzekrojowaSzczegoly.aspx?id=

- BeautifulSoup was used to gather data from the provided webpages. On every single page of every possible illness, my algorithm searched for links to specific annual data. I iterated over those links and opened the necessary sub-pages. From each sub-page, I extracted the following data:

  year, code, code description, hospitalization count, procedure name and procedure performed count.

- For every procedure that contained specific substrings I exported data to a .txt file using the above format. Specifying the right substrings helped detect the following procedures which correspond to:

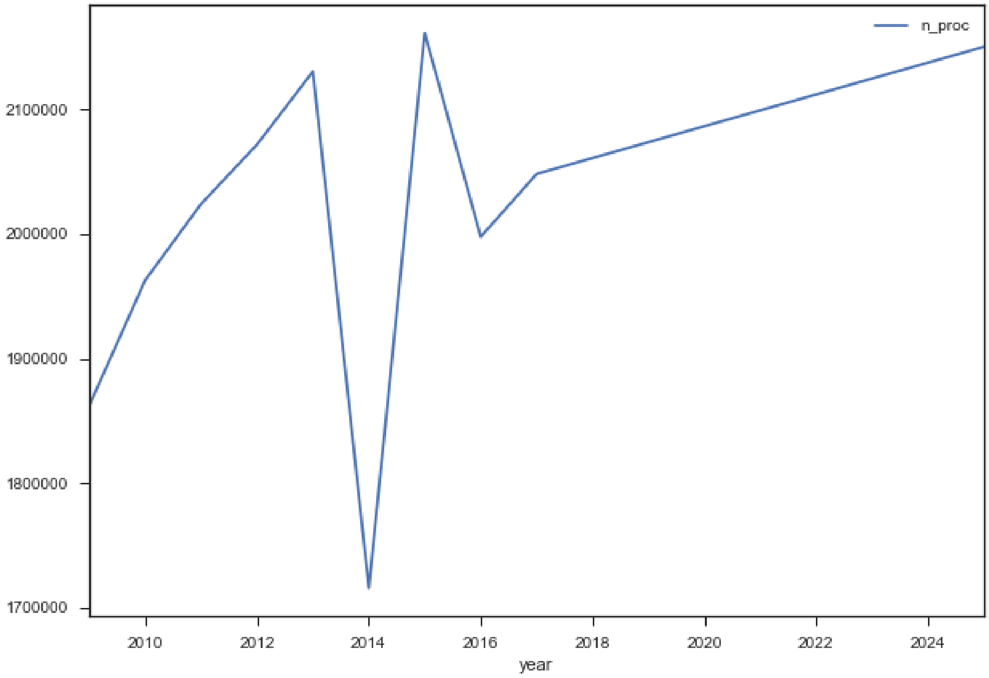
  Elektrokardiogram (ocena rytmu serca);
  Monitorowanie elektrokardiograficzne – inne;
  Elektrokardiogram z 1-3 odprowadzeniami;
  Elektrokardiogram nieokreślony;
  Monitorowanie elektrokardiograficzne;
  Elektrokardiogram;

2. Results

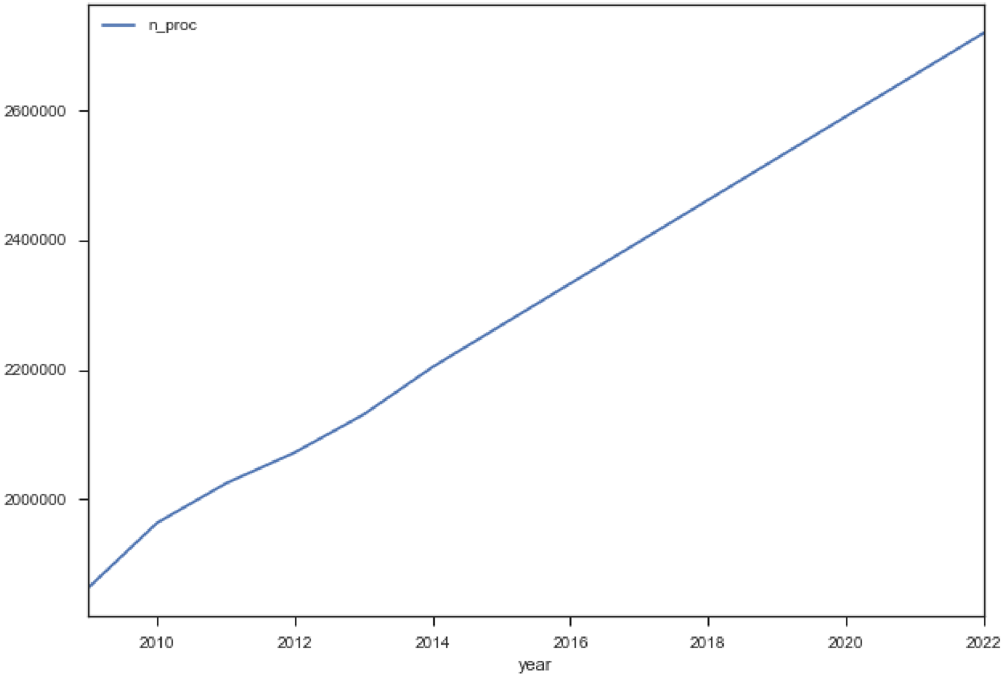
- Once the data was parsed into a .csv format, I read the gathered data into a Python pandas data frame. I then summed up the EKG procedures annually.

Annual EKG procedures (est)	
year	**count**
2009	1 862 393
2010	1 962 989
2011	2 024 214
2012	2 071 940
2013	2 130 648
2014	1 715 625
2015	2 161 318
2016	1 997 804

- Since there was a change in the number of EKG performed annually, also built a simple prediction model and fitted it to the data.



However, some of the data seems to be missing for 2014. It may be due to the medical law's changes or some internal NFZ's problem with data. For the years 2009-2013, a stable trend can be observed, which was used as a base for the second model that takes only those years' values as inputs.



- For data modelling simple regression technique was used. Since our data does not have a long history, it suggests that a time series analysis for prediction would not be fitting. However, without any seasonal changes, a regression model is very useful and has high interpretability.

- In order to assess the predictive linear model, I calculated the R^2 metric to show the goodness of fit. When only the years 2009-1013 were used as predictive data, the R^2 coefficient was 0.978. When the years 2009-2016 were used, the R^2 coefficient was 0.041

- Additionally, I created a list of the top illnesses for which EKG is being used most often.

Most popular usage of EKG		
Illness code in NFZ database	**Illness name in NFZ database**	**count**
E53	Niewydolność krążenia > 69 r.ż. lub z pw	1101269
D46	POChP i inne obturacyjne choroby układu oddechowego	478604
E88	Nadciśnienie tętnicze > 17 r.ż.	456864
E61	Zaburzenia rytmu serca > 69 r.ż. lub z pw	447455
E77	Inne choroby układu krążenia > 17 r.ż.	397261
E56	Choroba niedokrwienna serca > 69 r.ż. lub z pw	387613
D28	Choroby nowotworowe układu oddechowego i klatki piersiowej	361536
E62	Zaburzenia rytmu serca > 17 r.ż. < 70 r.ż. bez pw	353617
E73	Choroby zastawek serca > 17 r.ż.	344309

- For future work, these results could be made more accurate by building a prediction model for every illness in the dataset, and then aggregating the predicted results.

Thank you for reading!