Project 3 Design Document

USDA Food Database-Analyzing Nutrient Information

Panko Aliaksandr

January 11, 2018

Part I
Project Description

The project consists of 7 parts:

1. Get original data in JSON format
2. Load the JSON dataset into Python program
3. Separate the data to provide convenient access and make it easy to analyze
4. Find out, what 'Amino Acids' nutrient group is and create a table showing the different constituents of the group
5. For every Amino Acid find all foods in which they are present
6. For all the different nutrient groups calculate the median Zinc content (median of the zinc content in all the foods that constitute the nutrient group)
7. Plot the distribution of median Zinc Content for different nutrient groups as a bar chart

The document is supposed to be clear for an ordinary person, so python code is beyond of the content. To see implementation details, please use .py file.

Part II
Data Sources

Source of data is USDA National Nutrient Database for Standard Reference, Release 28 (2015) I was provided with JSON file, which contains the mentioned data.

Part III
Python libraries used

The following libraries have been used. They can be installed using Anaconda or using the Unix command pip as follows:
pip install "library name"

- import pandas as pd
- import numpy as np

Part IV
The class implementation

I have created a class called "NutrientsModel", which contains all necessary modules to run the assignment. The class contains next modules:

1. Class initializer (set all parameters)
2. Parsing module (named "ParseFile", which download data into the program and prepare it for analysis)
3. Module, which finds relevant foods for all Amino Acids (named "FindAminoAcidsFood")
4. Module, which calculates median zinc value for a product group (named "FindZincMedianValues")
5. Module, which displays result as bar graph (named: "ShowResults")
6. Main module, which runs all other modules

Part V
Parsing Process

Parsing Process is one of the most tricky part of the whole assignment. The original file has a very complicated structure, so the firs challenge was to understand the structure properly.

The second step was rebuilding the data into more simple and convenient way. This part was done using powerful python functionality.

Part VI
Amino Acids and corresponding foods

Firstly, the whole list of Amino Acids was taken from $http://www.cryst.bbk.ac.uk/education/Am$ website.

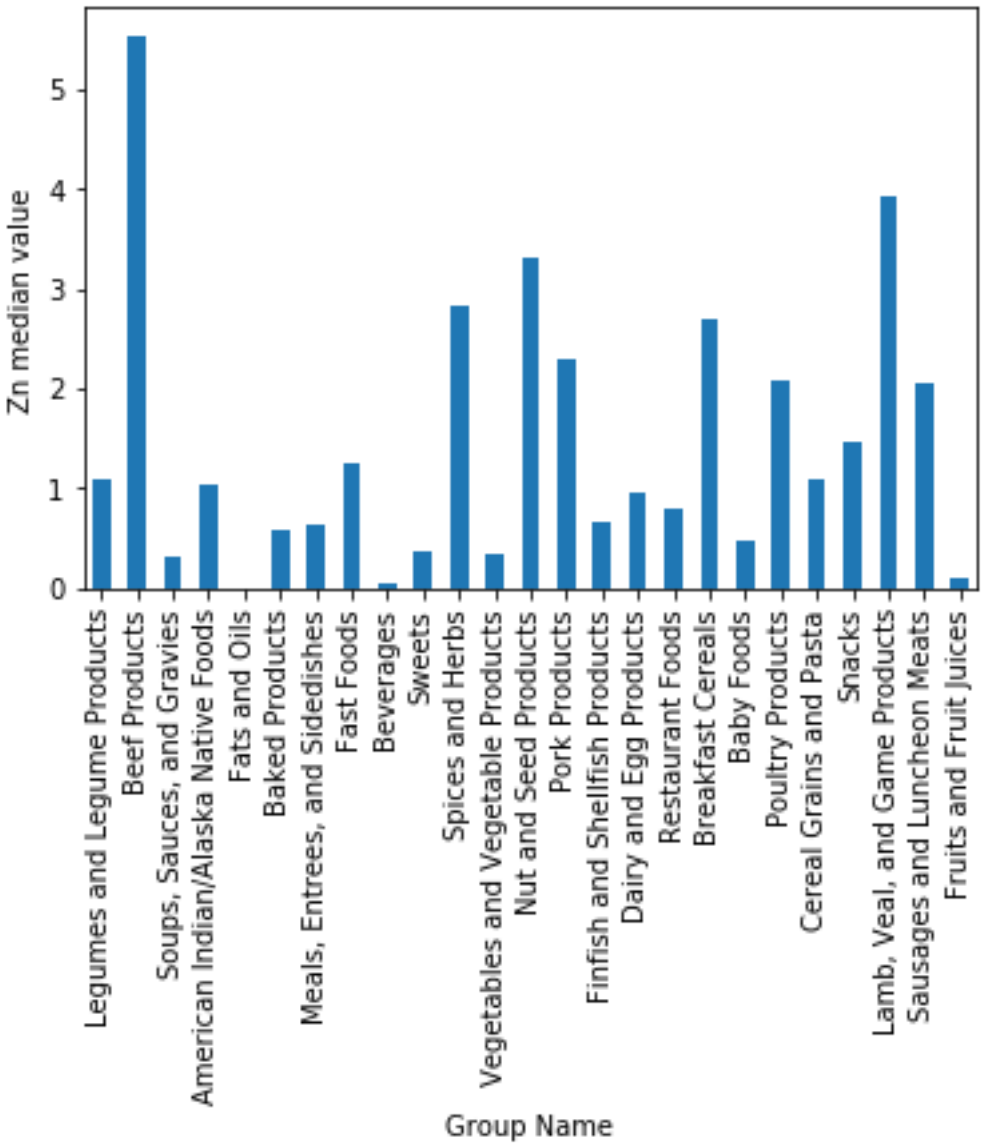
After that, the whole data frame was searched to find one of such acids. When the element was found the appropriate food was added to a list with all foods that contain such acid. That is the was of forming such a list for all acids

Part VII
Zinc Median Value

To find zinc median value I needed to collect for all groups products, which contain zinc and store value for all product. Afterwards, using $numpymedian()$ function I found zinc median value for every group. During this process I faced the problem of converting $Unicode$ data into numeric (float) data. The problem was solved using a python function $float(x)$, which did this conversion.

Part VIII
Visualization

To represent the results *bar* graph was used. $Pandas.DataFrame.plot(kind =' bar')$ function was used to make the chart:



Part IX
Conclusion

The assignment helped me to understand much better how powerful the python is in terms of data analysis. Most important steps have been covered and implemented here.