Environmental Health Project Part 1 - Problem 2 (Extract linear relationship by computing Pearson Correlation Coefficient)

March 28, 2023

0.0.1 Environmental Health Project Part 1 - Problem 2 (Extract linear relationship by computing Pearson Correlation Coefficient)

Author: Kenneth Cochran

Date written : 3/24/2023

[1]:
```
import os
import numpy as np
import pandas as pd
from scipy import stats
import seaborn as sns
%matplotlib inline

result_dir = '~/Environmental Health project - part 1/result'
data_dir = '~/Environmental Health project - part 1/data'

in_file_name = 'Clean PFAS_J.xlsx'
out_file_name = "Filtered_Pearson_Correlation_Coefficient.csv"

in_file_full_name = os.path.join(result_dir, in_file_name)
out_file_full_name = os.path.join(result_dir, out_file_name)

data_in = pd.read_excel(in_file_full_name) #Reads the Clean PFAS_J excel sheet␣
 ↪into a data frame object
data = data_in.drop(['Unnamed: 0', 'SEQN'], axis = 1) #Drops the index column␣
 ↪and the seqn label column.

column_points = data.columns #Gets the labels for the final data frame
length = len(column_points) #Gets the amount of labels of the dataframe for the␣
 ↪column and index so the final dataframe can be shaped correctly
values = [] #Creates an empty list that the pearson correlation coefficients␣
 ↪will be stored in

for h in column_points: #Calculates the pearson correlation coefficient for␣
 ↪every pair of columns
    for k in column_points:
```

```
        res = stats.pearsonr(data[h], data[k])
        values.append(res.statistic) #Stores the pearson correlation␣
 ↪coefficient for each pair of columns in the values list

reshaped_values = np.array(values).reshape(length,length) #Creates a reshaped␣
 ↪numpy array using the values list
matrix = pd.DataFrame(reshaped_values, index = column_points, columns =␣
 ↪column_points) #Creates a dataframe of the pearson correlation coefficient␣
 ↪with the variables as the index and column
print("DataFrame of the pearson correlation coefficient values ")
display(matrix)

filtered_matrix = matrix[(matrix > 0.50) & (matrix < 0.999)] #Filters the␣
 ↪matrix to only show the values greater than 0.50 and less than 0.999
print("Filtered Pearson Correlation Coefficient Values")
display(filtered_matrix)

print("Heatmap of the filtered pearson correlation coefficient values")
print(sns.heatmap(filtered_matrix)) #Creates a heatmap of the filtered values
filtered_matrix.to_csv(out_file_full_name) #Outputs the filtered matrix to a␣
 ↪csv file
```

```
DataFrame of the pearson correlation coefficient values
```

	WTSB2YR	LBXPFDE	LBDPFDEL	LBXPFHS	LBDPFHSL	LBXMPAH	\
WTSB2YR	1.000000	-0.057078	-0.038925	-0.009589	-0.026496	-0.011288	
LBXPFDE	-0.057078	1.000000	-0.201507	0.133904	-0.036122	0.105230	
LBDPFDEL	-0.038925	-0.201507	1.000000	-0.114021	0.170727	-0.061466	
LBXPFHS	-0.009589	0.133904	-0.114021	1.000000	-0.052724	0.071096	
LBDPFHSL	-0.026496	-0.036122	0.170727	-0.052724	1.000000	-0.025762	
LBXMPAH	-0.011288	0.105230	-0.061466	0.071096	-0.025762	1.000000	
LBDMPAHL	-0.042757	-0.069084	0.165554	-0.061631	0.073136	-0.338707	
LBXPFNA	-0.008649	0.667573	-0.263935	0.213898	-0.067991	0.187789	
LBDPFNAL	-0.022725	-0.143093	0.494405	-0.113879	0.192800	-0.088334	
LBXPFUA	-0.094165	0.831445	-0.154302	0.192058	-0.026555	0.075416	
LBDPFUAL	-0.009103	-0.305095	0.452866	-0.083236	0.061358	-0.060042	
LBXNFOA	0.018409	0.336893	-0.166541	0.440998	-0.064774	0.101672	
LBDNFOAL	-0.027582	-0.034069	0.169071	-0.038520	0.627278	-0.023667	
LBXBFOA	-0.002378	-0.000130	-0.033849	0.009723	0.011692	0.066845	
LBDBFOAL	0.012419	-0.000433	0.003819	0.018145	-0.057273	-0.015102	
LBXNFOS	-0.091077	0.689337	-0.185722	0.300094	-0.055458	0.230324	
LBDNFOSL	-0.026600	-0.031534	0.156489	-0.035753	0.678129	-0.021770	
LBXMFOS	-0.021651	0.391998	-0.208064	0.344314	-0.076492	0.314731	
LBDMFOSL	-0.032105	-0.050676	0.220105	-0.056217	0.481506	-0.026530	

	LBDMPAHL	LBXPFNA	LBDPFNAL	LBXPFUA	LBDPFUAL	LBXNFOA	\
WTSB2YR	-0.042757	-0.008649	-0.022725	-0.094165	-0.009103	0.018409	
LBXPFDE	-0.069084	0.667573	-0.143093	0.831445	-0.305095	0.336893	

LBDPFDEL	0.165554	-0.263935	0.494405	-0.154302	0.452866	-0.166541
LBXPFHS	-0.061631	0.213898	-0.113879	0.192058	-0.083236	0.440998
LBDPFHSL	0.073136	-0.067991	0.192800	-0.026555	0.061358	-0.064774
LBXMPAH	-0.338707	0.187789	-0.088334	0.075416	-0.060042	0.101672
LBDMPAHL	1.000000	-0.120590	0.150404	-0.031743	0.099624	-0.063364
LBXPFNA	-0.120590	1.000000	-0.264386	0.628809	-0.355967	0.457874
LBDPFNAL	0.150404	-0.264386	1.000000	-0.116251	0.293807	-0.164761
LBXPFUA	-0.031743	0.628809	-0.116251	1.000000	-0.317164	0.351424
LBDPFUAL	0.099624	-0.355967	0.293807	-0.317164	1.000000	-0.172888
LBXNFOA	-0.063364	0.457874	-0.164761	0.351424	-0.172888	1.000000
LBDNFOAL	0.054932	-0.055966	0.211683	-0.025734	0.065961	-0.051830
LBXBFOA	-0.038321	0.013634	-0.027985	-0.003356	0.004372	0.060506
LBDBFOAL	0.033915	0.022975	0.022542	0.003888	-0.028542	-0.009097
LBXNFOS	-0.130783	0.561441	-0.161985	0.617804	-0.263993	0.251810
LBDNFOSL	0.048139	-0.051801	0.195929	-0.024708	0.077902	-0.047819
LBXMFOS	-0.193334	0.494271	-0.210846	0.306994	-0.226124	0.353510
LBDMFOSL	0.051694	-0.083128	0.277429	-0.036421	0.091336	-0.071616

	LBDNFOAL	LBXBFOA	LBDBFOAL	LBXNFOS	LBDNFOSL	LBXMFOS	LBDMFOSL
WTSB2YR	-0.027582	-0.002378	0.012419	-0.091077	-0.026600	-0.021651	-0.032105
LBXPFDE	-0.034069	-0.000130	-0.000433	0.689337	-0.031534	0.391998	-0.050676
LBDPFDEL	0.169071	-0.033849	0.003819	-0.185722	0.156489	-0.208064	0.220105
LBXPFHS	-0.038520	0.009723	0.018145	0.300094	-0.035753	0.344314	-0.056217
LBDPFHSL	0.627278	0.011692	-0.057273	-0.055458	0.678129	-0.076492	0.481506
LBXMPAH	-0.023667	0.066845	-0.015102	0.230324	-0.021770	0.314731	-0.026530
LBDMPAHL	0.054932	-0.038321	0.033915	-0.130783	0.048139	-0.193334	0.051694
LBXPFNA	-0.055966	0.013634	0.022975	0.561441	-0.051801	0.494271	-0.083128
LBDPFNAL	0.211683	-0.027985	0.022542	-0.161985	0.195929	-0.210846	0.277429
LBXPFUA	-0.025734	-0.003356	0.003888	0.617804	-0.024708	0.306994	-0.036421
LBDPFUAL	0.065961	0.004372	-0.028542	-0.263993	0.077902	-0.226124	0.091336
LBXNFOA	-0.051830	0.060506	-0.009097	0.251810	-0.047819	0.353510	-0.071616
LBDNFOAL	1.000000	0.002960	-0.066328	-0.043746	0.770754	-0.057686	0.564829
LBXBFOA	0.002960	1.000000	-0.735996	-0.012560	-0.000906	0.009764	-0.014543
LBDBFOAL	-0.066328	-0.735996	1.000000	0.033794	-0.043623	0.025606	-0.007780
LBXNFOS	-0.043746	-0.012560	0.033794	1.000000	-0.040689	0.680226	-0.064369
LBDNFOSL	0.770754	-0.000906	-0.043623	-0.040689	1.000000	-0.053370	0.508130
LBXMFOS	-0.057686	0.009764	0.025606	0.680226	-0.053370	1.000000	-0.087640
LBDMFOSL	0.564829	-0.014543	-0.007780	-0.064369	0.508130	-0.087640	1.000000

Filtered Pearson Correlation Coefficient Values

	WTSB2YR	LBXPFDE	LBDPFDEL	LBXPFHS	LBDPFHSL	LBXMPAH	LBDMPAHL	\
WTSB2YR	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBXPFDE	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBDPFDEL	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBXPFHS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBDPFHSL	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBXMPAH	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBDMPAHL	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

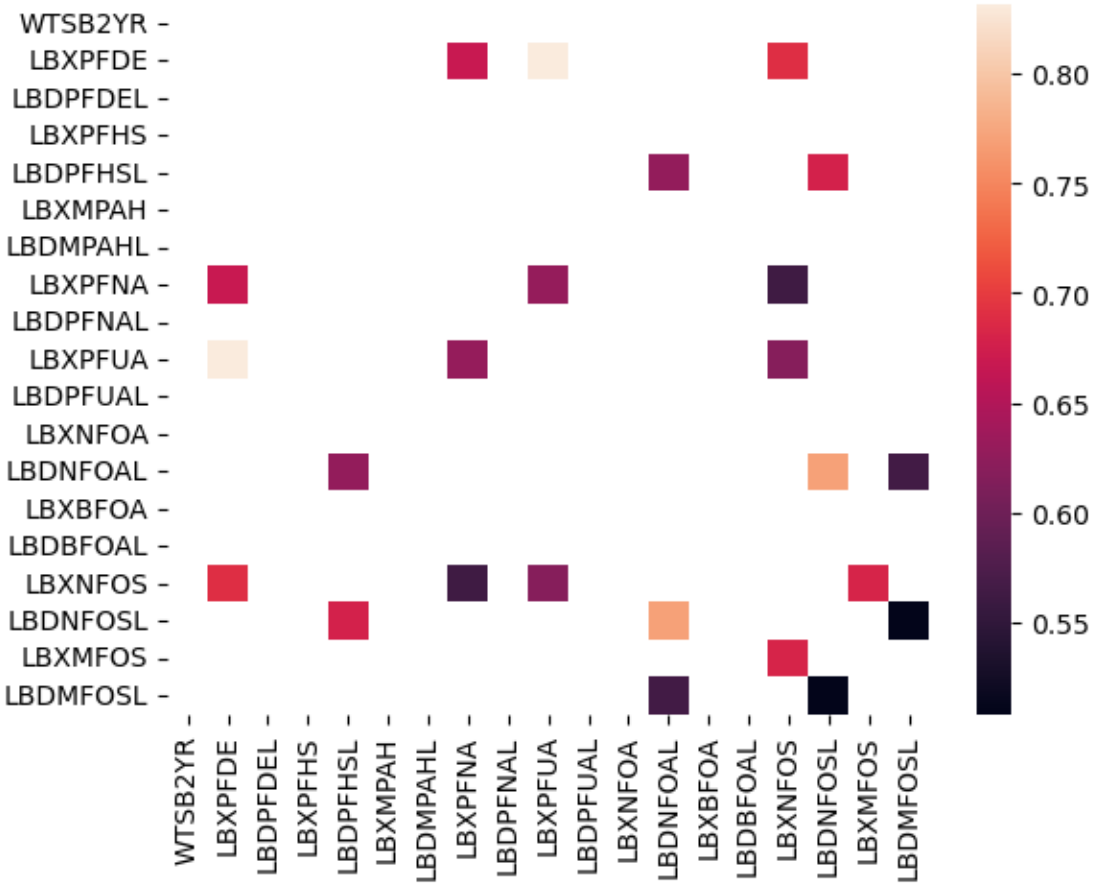
LBXPFNA	NaN	0.667573	NaN	NaN	NaN	NaN	NaN
LBDPFNAL	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LBXPFUA	NaN	0.831445	NaN	NaN	NaN	NaN	NaN
LBDPFUAL	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LBXNFOA	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LBDNFOAL	NaN	NaN	NaN	NaN	0.627278	NaN	NaN
LBXBFOA	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LBDBFOAL	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LBXNFOS	NaN	0.689337	NaN	NaN	NaN	NaN	NaN
LBDNFOSL	NaN	NaN	NaN	NaN	0.678129	NaN	NaN
LBXMFOS	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LBDMFOSL	NaN	NaN	NaN	NaN	NaN	NaN	NaN

	LBXPFNA	LBDPFNAL	LBXPFUA	LBDPFUAL	LBXNFOA	LBDNFOAL	LBXBFOA	\
WTSB2YR	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBXPFDE	0.667573	NaN	0.831445	NaN	NaN	NaN	NaN	
LBDPFDEL	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBXPFHS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBDPFHSL	NaN	NaN	NaN	NaN	NaN	0.627278	NaN	
LBXMPAH	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBDMPAHL	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBXPFNA	NaN	NaN	0.628809	NaN	NaN	NaN	NaN	
LBDPFNAL	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBXPFUA	0.628809	NaN	NaN	NaN	NaN	NaN	NaN	
LBDPFUAL	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBXNFOA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBDNFOAL	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBXBFOA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBDBFOAL	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBXNFOS	0.561441	NaN	0.617804	NaN	NaN	NaN	NaN	
LBDNFOSL	NaN	NaN	NaN	NaN	NaN	0.770754	NaN	
LBXMFOS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
LBDMFOSL	NaN	NaN	NaN	NaN	NaN	0.564829	NaN	

	LBDBFOAL	LBXNFOS	LBDNFOSL	LBXMFOS	LBDMFOSL
WTSB2YR	NaN	NaN	NaN	NaN	NaN
LBXPFDE	NaN	0.689337	NaN	NaN	NaN
LBDPFDEL	NaN	NaN	NaN	NaN	NaN
LBXPFHS	NaN	NaN	NaN	NaN	NaN
LBDPFHSL	NaN	NaN	0.678129	NaN	NaN
LBXMPAH	NaN	NaN	NaN	NaN	NaN
LBDMPAHL	NaN	NaN	NaN	NaN	NaN
LBXPFNA	NaN	0.561441	NaN	NaN	NaN
LBDPFNAL	NaN	NaN	NaN	NaN	NaN
LBXPFUA	NaN	0.617804	NaN	NaN	NaN
LBDPFUAL	NaN	NaN	NaN	NaN	NaN
LBXNFOA	NaN	NaN	NaN	NaN	NaN
LBDNFOAL	NaN	NaN	0.770754	NaN	0.564829

```
LBXBFOA        NaN       NaN       NaN       NaN       NaN
LBDBFOAL       NaN       NaN       NaN       NaN       NaN
LBXNFOS        NaN       NaN       NaN  0.680226       NaN
LBDNFOSL       NaN       NaN       NaN       NaN  0.508130
LBXMFOS        NaN  0.680226       NaN       NaN       NaN
LBDMFOSL       NaN       NaN  0.508130       NaN       NaN
```

```
Heatmap of the filtered pearson correlation coefficient values
AxesSubplot(0.125,0.11;0.62x0.77)
```



This heatmap shows that there is a slightly positive linear correlation between some of the variables in this data. The highest linear relationship is between LBXPFUA and LBXPFDE at 0.831445. The rest of the variable pairs show a slightly positive linear relationship.