プログラミングでゴリ押しデータ分析

愛知県立一宮高等学校　寺西陽菜

1.研究背景と目的

情報の授業でオープンデータを用いてデータ分析をする課題が出た(相関の有無は問わないもの)

◆ **相関の強いデータで分析したいけど**
パッと見で相関係数がわからない
◆ **分析に時間がかかる**
◆ **いざテーマを決めて分析しても**
相関が無いかも…不安

全ての相関係数を算出しよう！

そうすれば…

✓ **相関の強いものでテーマを決められる**
✓ **相関係数算出時間を短縮**
✓ **結論がわかっているので安心して分析できる**

2.相関係数算出

2-1.相関係数を算出する手法

◆ データ内の全ての項目の組み合わせで相関係数を算出するプログラムをChatGPTを使いながら作った

2-2.使用したデータや言語

◆ 使用データ：SSDSE-社会生活
◆ 使用言語：VBA

2-3.プログラムの調整

A.124×124項目の相関係数を計算する
B.総数や合計の項目を除外
C.同項目同士で計算しているものを除外
D.時刻を除外
E.相関係数の絶対値0.5未満の相関の弱いデータを除外

相関係数の推移

A.14641
B.13110
C.11254
D.11130
E.2450

2-4.プログラムの出力

◆ 出力する相関係数の範囲をいじることによって様々なデータが取れた

```
For i = 4 To 115
    Set dataRangeX = sourceSheet.Range(sourceSheet.Cells(4, i), sourceSheet.Cells(50, i))
    colX = i
    colNameX = ColumnLetter(colX)
    For j = 4 To 115
        If i <> j Then
            Set dataRangeY = sourceSheet.Range(source
            colY = j
            colNameY = ColumnLetter(colY)
            If InStr(1, sourceSheet.Cells(2, i).Value, "総数
                InStr(1, sourceSheet.Cells(2, i).Value, "合計
                InStr(1, sourceSheet.Cells(2, j).Value, "総数
                InStr(1, sourceSheet.Cells(2, j).Value, "合計
                correlation = CalculateCorrelation(dataRan
                .
                .
                .
```

3.分析テーマ決め

プログラムの出力から様々な興味深い相関係数を多く得ることができた
以下に一部を載せる

項目1	項目2	相関係数
テニス	楽器の演奏	0.545975
テニス	趣味としての料理・菓子作り	0.504878
テニス	絵画・彫刻の制作	0.560519
テニス	写真の撮影・プリント	0.566900
パソコンなどの情報処理	ボウリング	0.629993
パソコンなどの情報処理	登山・ハイキング	0.583300
パソコンなどの情報処理	パチンコ	-0.526134
パソコンなどの情報処理	サイクリング	0.839007
つり	映画館以外での映画鑑賞	-0.531998
つり	コンサートなどによるクラシック音楽鑑賞	-0.507398
つり	趣味としての読書（マンガを除く）	-0.565804
つり	行楽(日帰り)	-0.501159

実際のデータ分析課題では「釣り」と「ハイキング」、「釣り」と「観光旅行」など、「釣り」とその他の趣味に負の相関があるものが多かったことから、
「釣り人口とその他の趣味人口の負の相関の要因検証」というテーマに決定した

4.まとめ

今後は散布図の作成や外れ値の判断と除外など、データ分析そのものにもプログラミングを用いることで効率化したい
R言語にも挑戦してみたい