Exercise Set 03: Spurious Correlations

BEE 4850/5850, Fall 2024

Name: Anthony Nicolaides

ID: ajn68

> **Due Date**
>
> Friday, 2/9/24, 9:00pm

Overview

Instructions

The goal of this exercise is for you to find datasets and reason about the relationships (or lack thereof!) between variables.

Load Environment

The following code loads the environment and makes sure all needed packages are installed. This should be at the start of most Julia scripts.

In [ ]:

```julia
import Pkg
Pkg.activate(@__DIR__)
Pkg.instantiate()
```

```
  Activating project at `~/Documents/BEE5850/exercises/ex_week03_BEE5850`
```

The following packages are included in the environment (to help you find other similar packages in other languages). The code below loads these packages for use in the subsequent notebook (the desired functionality for each package is commented next to the package).

In [ ]:

```julia
using DataFrames # tabular data structure
using CSV # reads/writes .csv files
using Plots # plotting library
using StatsBase # statistical quantities like mean, median, etc
using StatsPlots # some additional statistical plotting tools
```

Problem

Find a single or multiple datasets (don’t just pull from Spurious Correlations!!) where two or more variables appear to be correlated, but this correlation is likely spurious. Plot the relevant variable(s) and show they are correlated through any needed quantiative and/or qualitative means. Explain why you think the correlation is spurious.

In [ ]:

```
oranges = CSV.read("statistic_id236882_us-retail-price-of-navel-oranges-1995
```

30×2 DataFrame *5 rows omitted*

Row	U.S. retail price of navel oranges 1995-2022	Column2
	String?	Float64?
1	Retail price of navel oranges in the United States from 1995 to 2022 (in U.S. dollars per pound)	*missing*
2	*missing*	*missing*
3	1995	0.64
4	1996	0.59
5	1997	0.58
6	1998	0.61
7	1999	0.64
8	2000	0.62
9	2001	0.71
10	2002	0.74
11	2003	0.86
12	2004	0.87
13	2005	0.89
⋮	⋮	⋮
19	2011	0.98
20	2012	1.04
21	2013	1.13
22	2014	1.25
23	2015	1.23
24	2016	1.17
25	2017	1.32
26	2018	1.39
27	2019	1.33
28	2020	1.33
29	2021	1.45
30	2022	1.49

In [ ]:

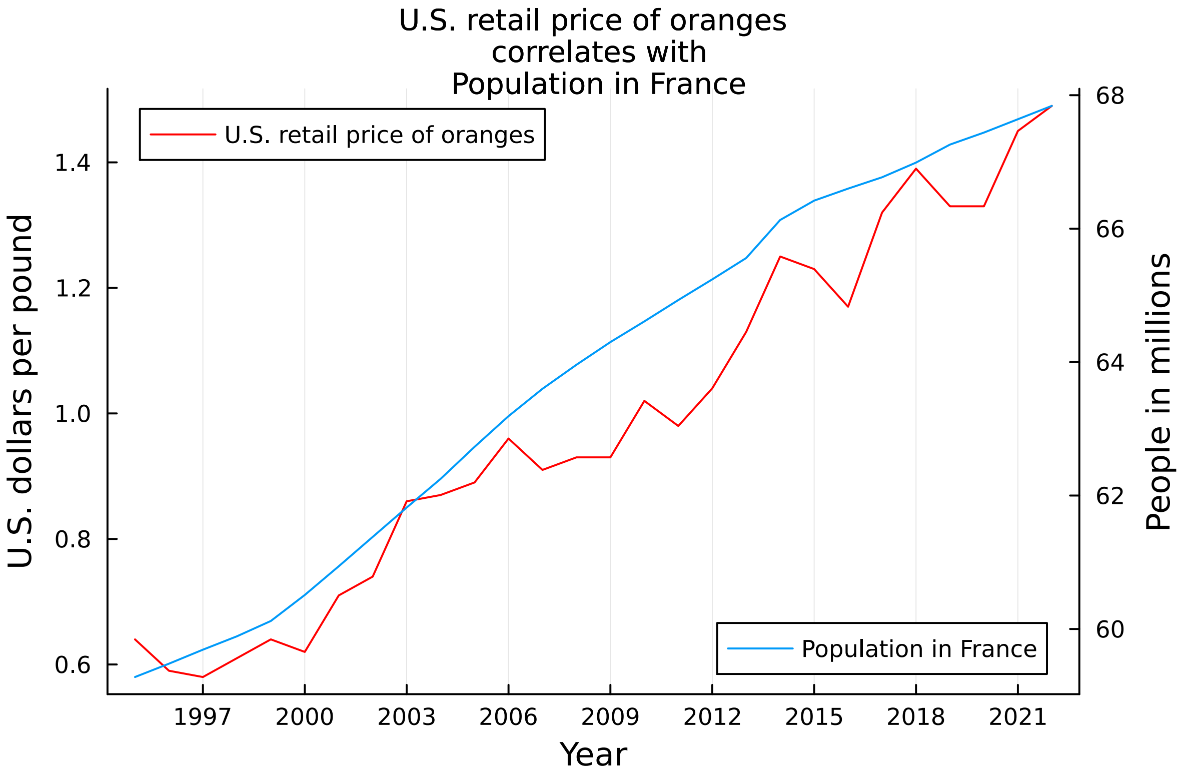
```
france = CSV.read("statistic_id459939_total-population-in-france-1982-2023.c
```

44×2 DataFrame *19 rows omitted*

Row	Total population in France 1982-2023	Column2
	String?	Float64?
1	Total population of France from 1982 to 2023 (in millions)	*missing*
2	*missing*	*missing*
3	1982	55.57
4	1983	55.9
5	1984	56.17
6	1985	56.44
7	1986	56.72
8	1987	57.01
9	1988	57.33
10	1989	57.66
11	1990	57.99
12	1991	58.28
13	1992	58.57
⋮	⋮	⋮
33	2012	65.24
34	2013	65.56
35	2014	66.13
36	2015	66.42
37	2016	66.6
38	2017	66.77
39	2018	66.99
40	2019	67.26
41	2020	67.44
42	2021	67.64
43	2022	67.84
44	2023	68.04

In [ ]:

```
# make sure only the values are from 1995-2022
plot(oranges[:, "U.S. retail price of navel oranges 1995-2022"][3:end],
    oranges[:, 2][3:end],
    ylabel="U.S. dollars per pound",
    label="U.S. retail price of oranges",
    linecolor=:red,
    xlabel="Year")
# twinx() allows for two y-axes
plot!(twinx(), france[:, "Total population in France 1982–2023"][16:end-1],
    france[:, 2][16:end-1],
    ylabel="People in millions",
    label="Population in France",
    legend=:bottomright)
# make sure only the values are from 1995–2022
title!("U.S. retail price of oranges\n correlates with\n Population in Franc
    fontsize=10,
    titlefont=font(10))
```



In [ ]:

```
# calculate the correlation coefficent between the two datasets
correlation = cor(oranges[:, 2][3:end], france[:, 2][16:end-1])
```

```
0.9764065653229481
```

I found two datasets, one is the U.S. retail price of oranges and the other is the population of France. Visually, it can be seen that they both follow an upward trend from 1995-2022, and their correlation coefficent if 0.976, which suggests a very strong positive correlation between these two datasets. However, this is a spurious correlation as it is reasonable to assume the cost of a pound of oranges in the United States has no influnce on the population of France.

References

Both datasets were found through Google's dataset search.

The US prices of oranges dataset is from Statista.

The population in France dataset is also from Statista