SUTD 2021 50.034 1D Project

Ragul Balaji, James R T, Shoham Chakraborty

Non-Transitivity Property of Pearson's Correlation Coefficient

Approach

- Deriving conclusions from real-world data, instead of the other way around (i.e., [cherry-]picking or generating simulated data to fit theoretical proposals or hypotheses).
- Not the most time-saving, but grounded and based on reality.

Process

- Spurious correlations: https://tylervigen.com/spurious-correlations
- Built a correlation finder script to ~~mine~~ detect the needed relationship ~~for the memes~~.*
- Processed a lot of random datasets (from Pokemon to GitHub repositories and cryptocurrencies).*
- Some correlations were not that convincingly strong.
- Some random variables were derived/calculated using formulas (“fake” strong correlations since they are latent - functions of observable R.V.’s).
- Learned/discovered random correlations (some might be valid, others might be a little bit shady).

* Full source code and explored datasets are available at: https://github.com/jamestiotio/pns

It's coding time, baby!

We're in ISTD. Less talk, more code.

Where's the (pseudo)code, young man?

```
Obligatory cool-looking code snippet incoming...
```

Where's the (pseudo)code, young man?

```
dataframe = pandas.read_csv(CSV_DATASET_FILE)
correlation_matrix = dataframe.corr(method="pearson")

for i in correlation_matrix:
    for j in correlation_matrix[i]:
        # we skip to the next iteration of the loop since the correlation of a random variable with itself is always 1
        if i == j: continue
        if correlation_matrix[i][j] < -THRESHOLD:
            for k in correlation_matrix:
                if i == k or j == k: continue  # same reasoning as above
                if correlation_matrix[i][k] > THRESHOLD and correlation_matrix[k][j] > THRESHOLD:
                    # we print the desired random variables and their relevant correlation coefficients
                    print(i, k, j, correlation_matrix[i][k], correlation_matrix[k][j], correlation_matrix[i][j])
```

On the hunt for data!

Data and more data...

- Our World in Data
- Data.world
- Statista
- Kaggle
- GitHub
- IEEE Research Papers
- Cryptowatch
- Hacker Noon

At last, we found a good one!

COVID-19 Dataset

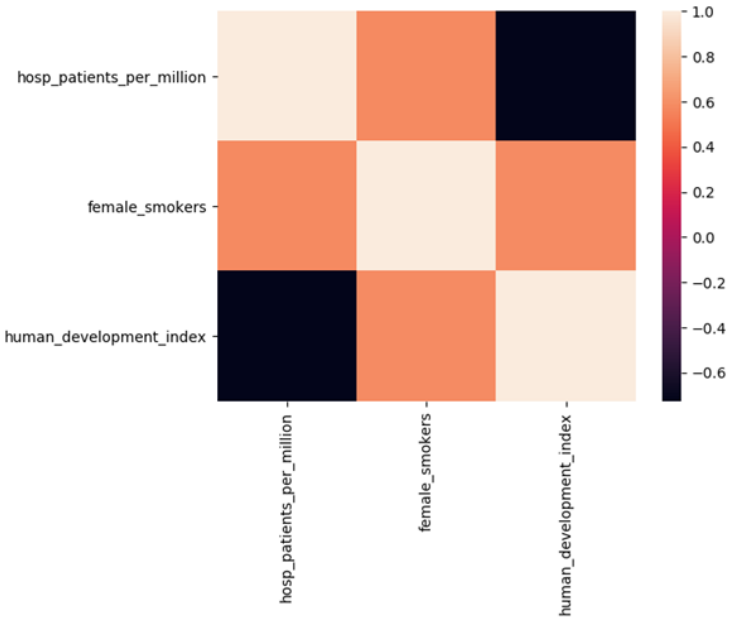
For **a specific country**, we define the following random variables:

- Let **X** be **hosp_patients_per_million**: the number of COVID-19 patients in hospital on a given day per 1,000,000 people.
- Let **Y** be **female_smokers**: the percentage or share of adult women who smoke in percent (ranging between 0.0 and 100.0) out of all adult women (ages 15 and over), at the most recent year available.
- Let **Z** be **human_development_index**: the composite index (ranging between 0.0 and 1.0) measuring average achievement in three basic dimensions of human development — a long and healthy life, knowledge and a decent standard of living. Values for 2019, imported from http://hdr.undp.org/en/indicators/137506.

COVID-19 Dataset

Using the df.corr() function from the pandas library in Python:

- ρ(X, Y) ≈ 0.575291 (> 0)
- ρ(X, Z) ≈ -0.726692 (< 0)
- ρ(Y, Z) ≈ 0.582708 (> 0)



	hosp_patients_per_million	female_smokers	human_development_index
hosp_patients_per_million	1.000000	0.575291	-0.726692
female_smokers	0.575291	1.000000	0.582708
human_development_index	-0.726692	0.582708	1.000000

Explanation Time!

Strange at first glance!?

From COVID-19 to gender equality/equity and female empowerment!

Question to ponder: do people care more about/advocate more for gender equality and “free will” than health?

Thank You!
Any Questions?