# Project_based learning 1:
# Làm sạch và phân tích tập dữ liệu nông nghiệp

**Đặng Văn Nam**

dangvannam@humg.edu.vn

namdv@aiacademy.edu.vn

# Làm sạch và Phân tích tập dữ liệu nông nghiệp

**(Tiếp cận từ bài toán với dữ liệu thực tế)**

# 1.1 Mô tả bài toán

+ Tập dữ liệu về giá bán một số mặt hàng nông sản của Mỹ theo từng tháng từ 04/1990 đến 04/2020. Bao gồm 8 loại sản phẩm:

1. Coarse wool (Len thô)
2. Copra (Cùi dừa)
3. Cotton (Bông)
4. Fine wool (Len min)
5. Hard wood (Gỗ cứng)
6. Hard sawnwood (Gỗ xẻ cứng)
7. Hide (Da thú)
8. Rubber (Cao su)




# 1.2 Tập dữ liệu

- Dữ liệu lữu trữ trong file agricultural_raw_material.csv:

| A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|
| Month | Coarse wool Price | Coarse wool price % Change | Copra Price | Copra price % Change | Cotton Price | Cotton price % Change | Fine wool Price | Fine wool price % Change | Hard log Price | Hard log price |
| Apr-90 | 482.34 | - | 236 | - | 1.83 | - | 1,071.63 | - | 161.2 | - |
| May-90 | 447.26 | -7.27% | 234 | -0.85% | 1.89 | 3.28% | 1,057.18 | -1.35% | 172.86 | |
| Jun-90 | 440.99 | -1.40% | 216 | -7.69% | 1.99 | 5.29% | 898.24 | -15.03% | 181.67 | |
| Jul-90 | 418.44 | -5.11% | 205 | -5.09% | 2.01 | 1.01% | 895.83 | -0.27% | 187.96 | |
| Aug-90 | 418.44 | 0.00% | 198 | -3.41% | 1.79 | -10.95% | 951.22 | 6.18% | 186.13 | |
| Sep-90 | 412.18 | -1.50% | 196 | -1.01% | 1.79 | 0.00% | 936.77 | -1.52% | 185.33 | |
| Oct-90 | 394.64 | -4.26% | 198 | 1.02% | 1.79 | 0.00% | 901.85 | -3.73% | 189.76 | |
| Nov-90 | 334.5 | -15.24% | 236 | 19.19% | 1.82 | 1.68% | 888.61 | -1.47% | 179.02 | |
| Dec-90 | 328.24 | -1.87% | 237 | 0.42% | 1.85 | 1.65% | 870.55 | -2.03% | 171.13 | |
| Jan-91 | 319.47 | -2.67% | 233 | -1.69% | 1.85 | 0.00% | 887.41 | 1.94% | 169.19 | |
| Feb-91 | 323.23 | 1.18% | 226 | -3.00% | 1.87 | 1.08% | 596.02 | -32.84% | 176.93 | |
| Mar-91 | 328.24 | 1.55% | 236 | 4.42% | 1.86 | -0.53% | 586.39 | -1.62% | 162.57 | |
| Apr-91 | 365.82 | 11.45% | 224 | -5.08% | 1.83 | -1.61% | 596.02 | 1.64% | 175.59 | |
| May-91 | 371.88 | 1.66% | 226 | 0.89% | 1.82 | -0.55% | 721 | 20.97% | 174.04 | |
| Jun-91 | 340.6 | -8.41% | 245 | 8.41% | 1.78 | -2.20% | 777.51 | 7.84% | 200.15 | |
| Jul-91 | 337.48 | -0.92% | 303 | 23.67% | 1.7 | -4.49% | 723.48 | -6.95% | 207.82 | |
| Aug-91 | 337.22 | -0.08% | 299 | -1.32% | 1.62 | -4.71% | 680.64 | -5.92% | 210.57 | |
| Sep-91 | 313.96 | -6.90% | 296 | -1.00% | 1.55 | -4.32% | 613.45 | -9.87% | 210.68 | |
| Oct-91 | 308.39 | -1.77% | 353 | 19.26% | 1.5 | -3.23% | 558.18 | -9.01% | 214.44 | |
| Nov-91 | 307.57 | -0.27% | 385 | 9.07% | 1.4 | -6.67% | 641.49 | 14.93% | 195.5 | |
| Dec-91 | 295.4 | -3.96% | 411 | 6.75% | 1.36 | -2.86% | 652.19 | 1.67% | 200.33 | |
| Jan-92 | 297.04 | 0.56% | 488 | 18.73% | 1.31 | -3.68% | 619.37 | -5.03% | 202.85 | |
| Feb-92 | 341.89 | 15.10% | 471 | -3.48% | 1.24 | -5.34% | 655.83 | 5.89% | 214.04 | |
| Mar-92 | 341.18 | -0.21% | 429 | -8.92% | 1.22 | -1.61% | 667.93 | 1.84% | 201.96 | |
| Apr-92 | 352.12 | 3.21% | 425 | -0.93% | 1.28 | 4.92% | 667.7 | -0.03% | 199.67 | |

agricultural_raw_material

# 1.2 Tập dữ liệu

- Tập dữ liệu bao gồm 17 cột:
  - **Month:** Tháng của năm (Tháng - Năm): Apr – 90 (Tháng: 3 ký tự đầu tiên – Năm: 2 số cuối của năm)
  - Mỗi sản phẩm nông nghiệp bao gồm 2 thông tin, ví dụ:
    - **Coarse wool Price:** Giá bán Len thô của tháng ($USD): 482.34
    - **Coarse wool price % Change:** Tỷ lệ % thay đổi mức giá bán len thô của tháng so với tháng liền trước đó: -7.27% (Mức giá giảm so với tháng trước là 7.27%) | 1.01% (Mức giá tang so với tháng trước là 1.01 %) | 0.00% (Mức giá tháng này và tháng trước đó như nhau, không thay đổi giá)

***(Tương tự với 7 mặt hàng còn lại)***

# 1.3 Mục tiêu

- Tập dữ liệu thu thập được là tập dữ liệu thô, cần phải được chuẩn bị và tiền xử lý trước khi sử dụng cho bất kỳ mục đích nào.
- Thực hiện phân tích tập dữ liệu:
  - Xác định mối tương quan về giá giữa các mặt hàng.
  - Trực quan hóa, quan sát và rút ra các nhận xét có được từ dữ liệu

INSIGHT

# 1.4 Kết Quả

- **1) Tiền xử lý tập dữ liệu:**
  - Chuẩn hóa dữ liệu thời gian cột Month → Về dữ liệu Datetime
  - Chuẩn hóa dữ liệu cột thay đổi mức giá so với tháng trước → Dạng số (float)
  - Chuẩn hóa dữ liệu các cột giá bán có giá trị lớn hơn 1000 USD → Dạng số (float)
  - Lưu dữ liệu đã làm sạch ra file

***Dữ liệu đã được làm sạch:***

Data_Agriculture_OK - Excel

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do... Dang Van Nam Share

Date

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Date | Coarse wo | Coarse wo | Copra Pric | Copra pric | Cotton Pric | Cotton pric | Fine wool F | Fine wool p | Hard log Pr | Hard log pr | Hard sawn | Hard sawn | Hide Price | Hide price ' | Rubber Pri | Rubber price | % Change |
| 2 | 1990-04-01 | 482.34 | 0 | 236 | 0 | 1.83 | 0 | 1071.63 | 0 | 161.2 | 0 | 549.91 | 0 | 100 | 0 | 0.84 | 0 | |
| 3 | 1990-05-01 | 447.26 | -7.27 | 234 | -0.85 | 1.89 | 3.28 | 1057.18 | -1.35 | 172.86 | 7.23 | 491.88 | -10.55 | 99.46 | -0.54 | 0.85 | 1.19 | |
| 4 | 1990-06-01 | 440.99 | -1.4 | 216 | -7.69 | 1.99 | 5.29 | 898.24 | -15.03 | 181.67 | 5.1 | 495.39 | 0.71 | 97.9 | -1.57 | 0.85 | 0 | |
| 5 | 1990-07-01 | 418.44 | -5.11 | 205 | -5.09 | 2.01 | 1.01 | 895.83 | -0.27 | 187.96 | 3.46 | 485.86 | -1.92 | 96.75 | -1.17 | 0.86 | 1.18 | |
| 6 | 1990-08-01 | 418.44 | 0 | 198 | -3.41 | 1.79 | -10.95 | 951.22 | 6.18 | 186.13 | -0.97 | 487.52 | 0.34 | 91.89 | -5.02 | 0.88 | 2.33 | |
| 7 | 1990-09-01 | 412.18 | -1.5 | 196 | -1.01 | 1.79 | 0 | 936.77 | -1.52 | 185.33 | -0.43 | 487.75 | 0.05 | 87.66 | -4.6 | 0.9 | 2.27 | |
| 8 | 1990-10-01 | 394.64 | -4.26 | 198 | 1.02 | 1.79 | 0 | 901.85 | -3.73 | 189.76 | 2.39 | 505.24 | 3.59 | 84.59 | -3.5 | 0.9 | 0 | |
| 9 | 1990-11-01 | 334.5 | -15.24 | 236 | 19.19 | 1.82 | 1.68 | 888.61 | -1.47 | 179.02 | -5.66 | 511.33 | 1.21 | 82.86 | -2.05 | 0.9 | 0 | |
| 10 | 1990-12-01 | 328.24 | -1.87 | 237 | 0.42 | 1.85 | 1.65 | 870.55 | -2.03 | 171.13 | -4.41 | 499.2 | -2.37 | 84.85 | 2.4 | 0.88 | -2.22 | |
| 11 | 1991-01-01 | 319.47 | -2.67 | 233 | -1.69 | 1.85 | 0 | 887.41 | 1.94 | 169.19 | -1.13 | 497.05 | -0.43 | 81.64 | -3.78 | 0.87 | -1.14 | |
| 12 | 1991-02-01 | 323.23 | 1.18 | 226 | -3 | 1.87 | 1.08 | 596.02 | -32.84 | 176.93 | 4.57 | 492.56 | -0.9 | 75.9 | -7.03 | 0.85 | -2.3 | |
| 13 | 1991-03-01 | 328.24 | 1.55 | 236 | 4.42 | 1.86 | -0.53 | 586.39 | -1.62 | 162.57 | -8.12 | 491.12 | -0.29 | 76.6 | 0.92 | 0.83 | -2.35 | |
| 14 | 1991-04-01 | 365.82 | 11.45 | 224 | -5.08 | 1.83 | -1.61 | 596.02 | 1.64 | 175.59 | 8.01 | 495.44 | 0.88 | 86.45 | 12.86 | 0.82 | -1.2 | |
| 15 | 1991-05-01 | 371.88 | 1.66 | 226 | 0.89 | 1.82 | -0.55 | 721 | 20.97 | 174.04 | -0.88 | 509.75 | 2.89 | 88.72 | 2.63 | 0.82 | 0 | |
| 16 | 1991-06-01 | 340.6 | -8.41 | 245 | 8.41 | 1.78 | -2.2 | 777.51 | 7.84 | 200.15 | 15 | 480.3 | -5.78 | 86.7 | -2.28 | 0.84 | 2.44 | |
| 17 | 1991-07-01 | 337.48 | -0.92 | 303 | 23.67 | 1.7 | -4.49 | 723.48 | -6.95 | 207.82 | 3.83 | 480.3 | 0 | 83.91 | -3.22 | 0.82 | -2.38 | |
| 18 | 1991-08-01 | 337.22 | -0.08 | 299 | -1.32 | 1.62 | -4.71 | 680.64 | -5.92 | 210.57 | 1.32 | 553.48 | 15.24 | 77.52 | -7.62 | 0.82 | 0 | |
| 19 | 1991-09-01 | 313.96 | -6.9 | 296 | -1 | 1.55 | -4.32 | 613.45 | -9.87 | 210.68 | 0.05 | 582.79 | 5.3 | 74.05 | -4.48 | 0.81 | -1.22 | |
| 20 | 1991-10-01 | 308.39 | -1.77 | 353 | 19.26 | 1.5 | -3.23 | 558.18 | -9.01 | 214.44 | 1.78 | 573.05 | -1.67 | 75.36 | 1.77 | 0.83 | 2.47 | |
| 21 | 1991-11-01 | 307.57 | -0.27 | 385 | 9.07 | 1.4 | -6.67 | 641.49 | 14.93 | 195.5 | -8.83 | 573.89 | 0.15 | 75.26 | -0.13 | 0.81 | -2.41 | |
| 22 | 1991-12-01 | 295.4 | -3.96 | 411 | 6.75 | 1.36 | -2.86 | 652.19 | 1.67 | 200.33 | 2.47 | 567.39 | -1.13 | 71.29 | -5.28 | 0.79 | -2.47 | |
| 23 | 1992-01-01 | 297.04 | 0.56 | 488 | 18.73 | 1.31 | -3.68 | 619.37 | -5.03 | 202.85 | 1.26 | 577.77 | 1.83 | 70.7 | -0.83 | 0.8 | 1.27 | |
| 24 | 1992-02-01 | 341.89 | 15.1 | 471 | -3.48 | 1.24 | -5.34 | 655.83 | 5.89 | 214.04 | 5.52 | 599.19 | 3.71 | 67.84 | -4.05 | 0.81 | 1.25 | |
| 25 | 1992-03-01 | 341.18 | -0.21 | 429 | -8.92 | 1.22 | -1.61 | 667.93 | 1.84 | 201.96 | -5.64 | 602.24 | 0.51 | 69.77 | 2.84 | 0.83 | 2.47 | |
| 26 | 1992-04-01 | 352.12 | 3.21 | 425 | -0.93 | 1.28 | 4.92 | 667.7 | -0.03 | 199.67 | -1.13 | 616.38 | 2.35 | 75.95 | 8.86 | 0.86 | 3.61 | |

Data_Agriculture_OK

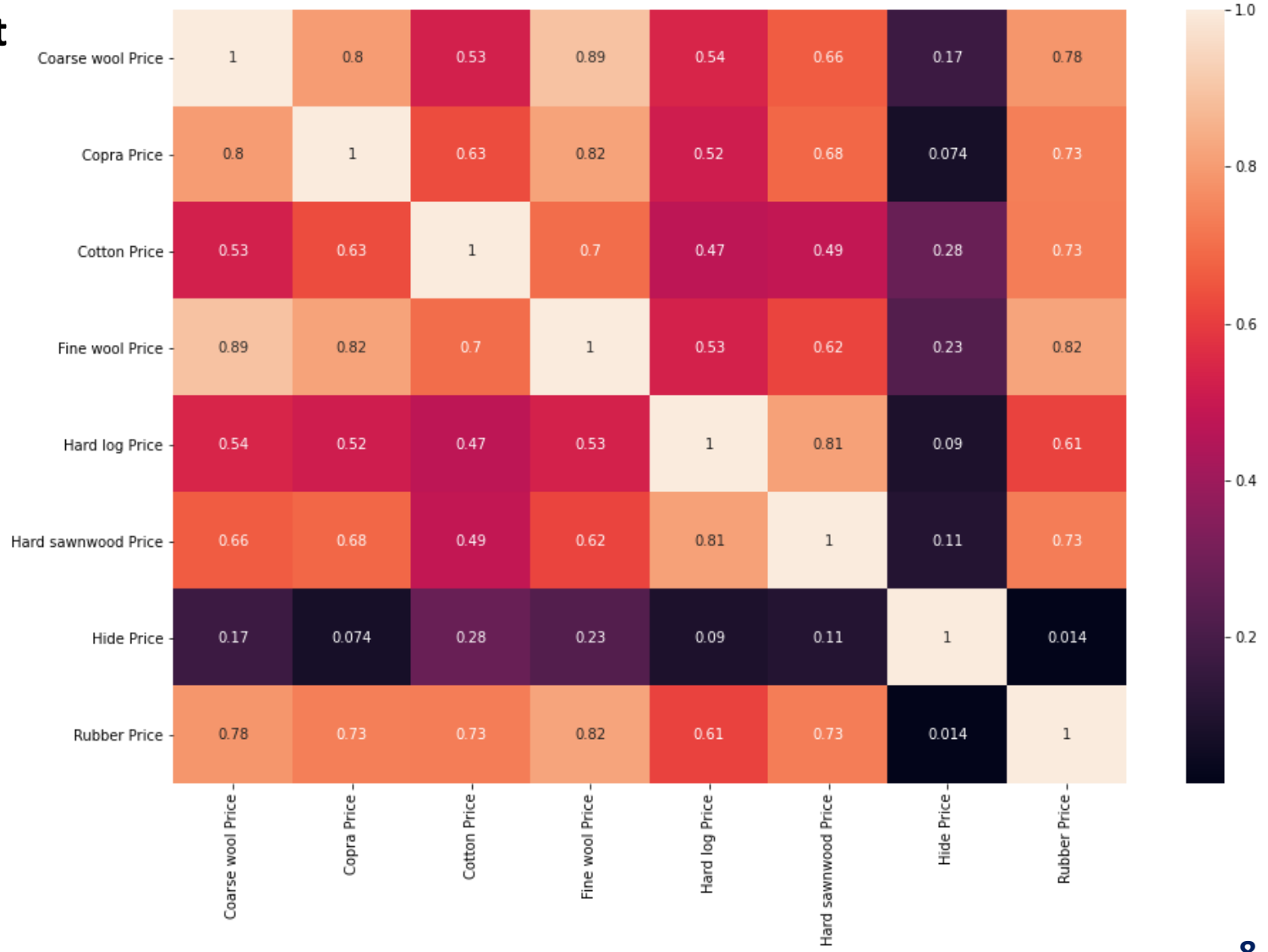
Ready 100%

*Chi tiết các bước thực hiện trong file jupyter notebook...*

# 1.4 Kết Quả

**2) Xác định mối tương quan giữa các mặt hàng nông sản:**

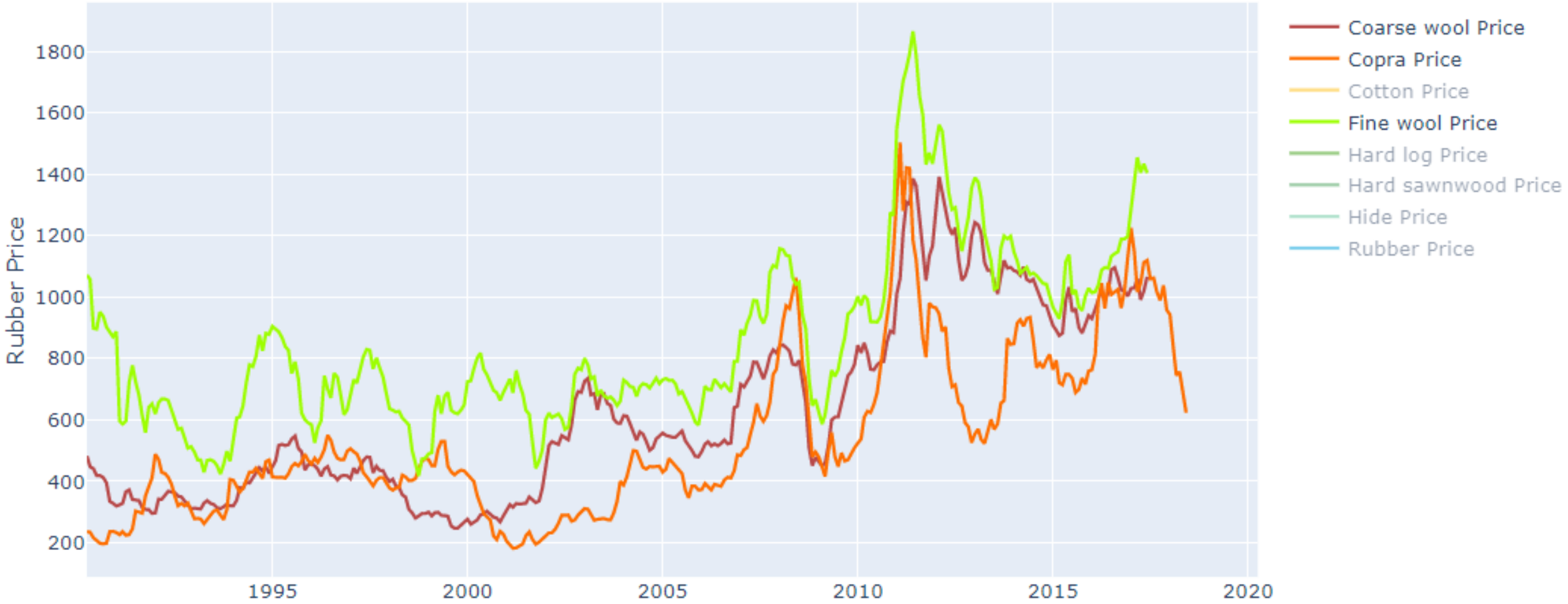
- Giá len thô (Coarse wool): Tương quan thuận với giá bán Fine wool (len mịn): 0.89, Copra (cùi dừa): 0.8
- Giá cùi dừa (Copra) Tương quan với giá len mịn (Fine wool): 0.82
- Giá Cotton tương quan thuận với giá cao su (Rubber): 0.73
- Giá len min (Fine Wool) tương quan thuận mạnh với giá len thô (Coarse wool): 0.89
- Giá gỗ cứng (Hard log) tương quan thuận với giá gỗ xẻ cứng (Hard sawnwood): 0.81
- Giá da thú (hide) ko bị ảnh hưởng nhiều với các mặt hàng khác
- Giá cao su (Rubber) tương quan mạnh với giá len min (Fine Wool):0.82




*Chi tiết các bước thực hiện trong file jupyter notebook…*

# 1.4 Kết Quả

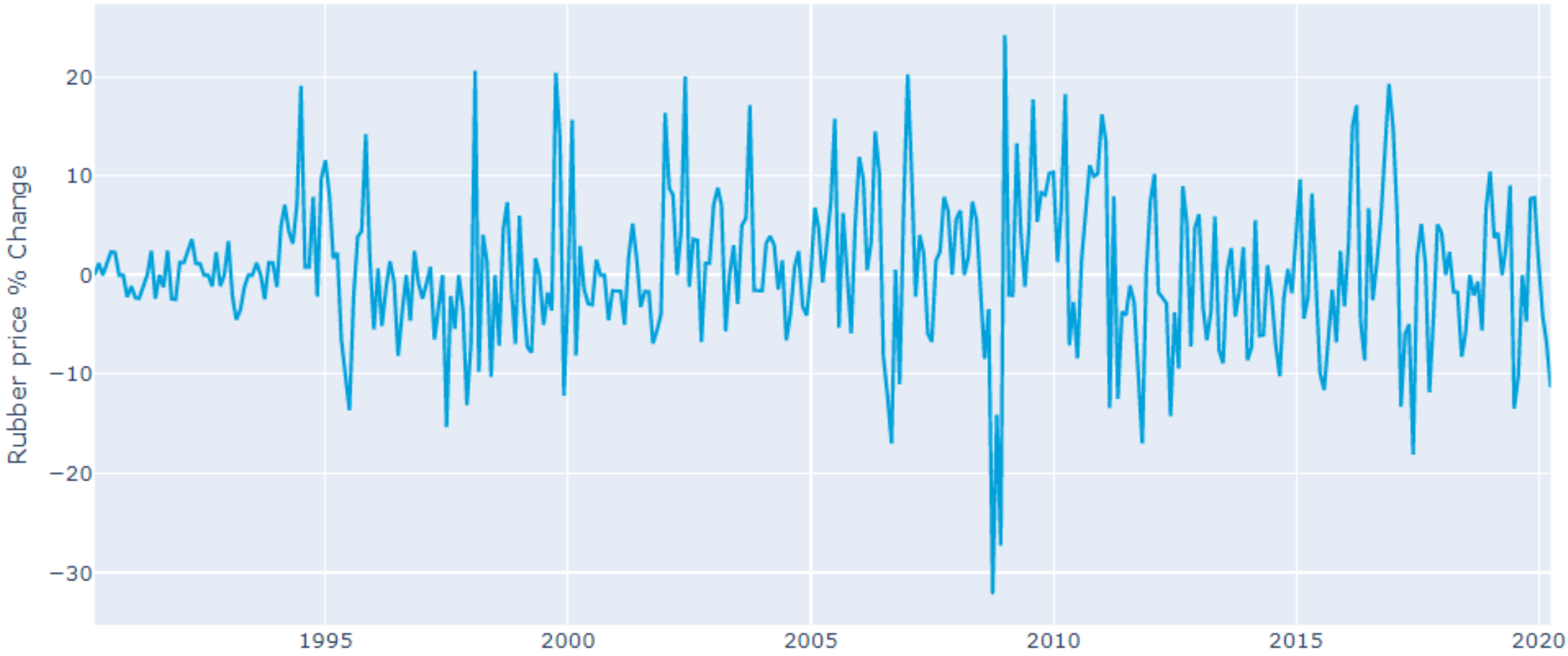
- **3) Trực quan hóa dữ liệu**




*Chi tiết các bước thực hiện trong file jupyter notebook…*

# 1.4 Kết Quả

- **3) Trực quan hóa dữ liệu**




*Chi tiết các bước thực hiện trong file jupyter notebook…*

# Thank you!