

**PROJETO: ANALISE EXPLORATÓRIA DOS DADOS – AVALIAÇÃO DE SERIE TEMPORAL DO PREÇO DE COMMODITIES**

SPRINT 1

Requerente: COOPERATIVA AGRÁRIA AGROINDUSTRIAL

Residentes: Bruno de Oliveira Pereira

Tiago Dutra Galvão

Londrina, Janeiro de 2023

Sumário

[Sumário 3](#_Toc124144629)

[1 INTRODUÇÃO 4](#_Toc124144630)

[1.1 OBJETIVO 4](#_Toc124144631)

[1.2 ESCOPO 4](#_Toc124144632)

[1.3 CONTRA-ESCOPO 4](#_Toc124144633)

[1.4 RISCOS 4](#_Toc124144634)

[1.5 METAS DO PROJETO 4](#_Toc124144635)

[1.5.1 Entregas da Meta 1 5](#_Toc124144636)

[2 RESULTADOS ALCANÇADOS 6](#_Toc124144637)

[2.1 SPRINT 1 6](#_Toc124144638)

[2.1.1 Limpeza e organização dos dados 6](#_Toc124144639)

[2.1.2 Estudo de ferramentas estatísticas que pudessem ser aplicadas aos dados 6](#_Toc124144640)

[2.1.3 Analise exploratória dos dados usando Python 7](#_Toc124144641)

[2.1.4 Documentação dos resultados 7](#_Toc124144642)

# INTRODUÇÃO

A análise exploratória dos dados tem por si só, uma grande quantidade de informações que são extraídas olhando para o comportamento que os dados tem ou tendem a ter atrás de um viés estatístico. Nessa sprint foi possível trabalhar com esses conceitos com objetivo de ao longo das 4 sprints, ser construído um projeto mais completo, em associação aos conhecimentos desenvolvidos pelo treinamento especializado oferecido pelo AI2 em conjunto com o SENAI.

## OBJETIVO

Este projeto tem como objetivo o estudo de tecnologias e proposição de uma POC (Prova de Conceito *– Proof of Concept*) de uma ferramenta interativa online para consulta de dados por gestores quanto à performance de commodities no mercado, com expectativa de entrega ao final da *Sprint* 4.

## ESCOPO

O escopo deste projeto compreende o estudo de ferramentas para desenvolvimento de ferramentas que auxiliam na tomada de decisão para venda de commodities, o estudo primário foi feito com base em dados entregues pela própria Agrária.

## CONTRA-ESCOPO

Não faz parte do escopo deste projeto o desenvolvimento/entrega de um MVP (Mínimo Produto Viável – *Minimum Valuable Product*).

## RISCOS

O principal risco identificado ao projeto seria a não viabilidade de um projeto que entregue uma predição dos valores da commodities visto que isso é impossível, caso os dados não tenham características favoráveis para tal.

## METAS DO PROJETO

Considerando a subdivisão do projeto de um ano, vão ser elaboradas metas para cada uma das sprints. Para a primeira sprint, foi elaborada e alcançada. em 4 sprints foram propostas, a partir do objetivo do projeto:

* Meta 1: Limpeza e organização dos dados e estudo da base de dados com analise exploratória dos dados;

### Entregas da Meta 1

Para o atingimento da Meta 1 foram propostas 5 entregas listadas a seguir:

* Limpeza dos dados;
* Organização dos dados;
* Estudo de ferramentas estatísticas que pudessem ser aplicadas aos dados;
* Analise exploratória dos dados usando Python;
* Documentação dos resultados.

# RESULTADOS ALCANÇADOS

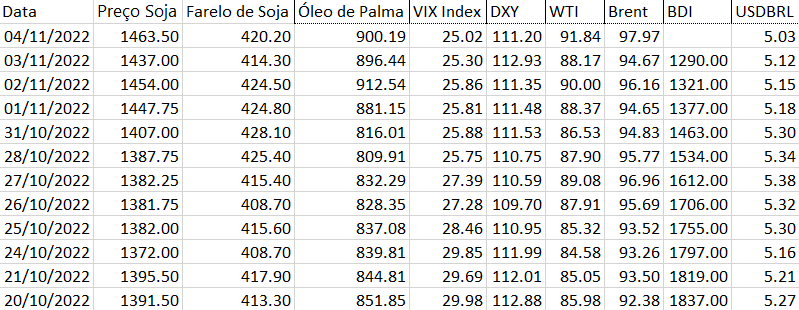
## SPRINT 1

Abaixo resultados sprint 1

### Limpeza e organização dos dados

Apesar de serem tarefas distintas, para fins didáticos, fica mais adequado descreve-las em conjunto. Os dados recebidos pelo HUB, tinham formatos diferentes e quantidades diferentes. É de conhecimento para a ciência de dados que, o formato dos dados que tem interesse em fomentar o *target* precisam ter mesmo formato (dia, mês e ano) e mesma quantidade de linhas, para que em cada resultado que seja validade como um alvo, todas as outras colunas devem estar com informações igualmente validas. Sendo assim, o primeiro passo foi organizar e retirar valores faltantes, dados que não são compatíveis com os dados alvo.

Já em segundo plano, a organização desses dados em um novo documento, onde se tivesse todas as informações dispostas em colunas, com o intuito de “maior compreensão do algoritmo que viria a ser feito”.



Como pode ser visto, apesar do extenso valor de colunas que foram passados nos dados e dicionário de dados, nem todos puderam ser utilizados por falta de informação ou simetria com os dados do alvo, que nesse caso é “Preço Soja”, o primeiro comodities que estudamos.

### Estudo de ferramentas estatísticas que pudessem ser aplicadas aos dados

Para essa parte do processo, contamos com o acompanhamento do consultor técnico Mateus Roder do AI2, que nos forneceu o auxílio para entender a construção logica das informações que poderíamos fazer com os dados, começando com um *Heatmap* das correlações de Pearson e de Spearman, seguido por *PaitPlots* e algumas plotagens de Curva de tendência desses dados.

### Analise exploratória dos dados usando Python

Nessa parte do processo, construímos o aparato técnico usando a plataforma Google Colab, afim de agilizar a disponibilidade do código para acesso da Agrária. Nele usamos a linguagem de programação Python e as bibliotecas Pandas, Numpy, Matplotlib, Seaborn e Stats Model.

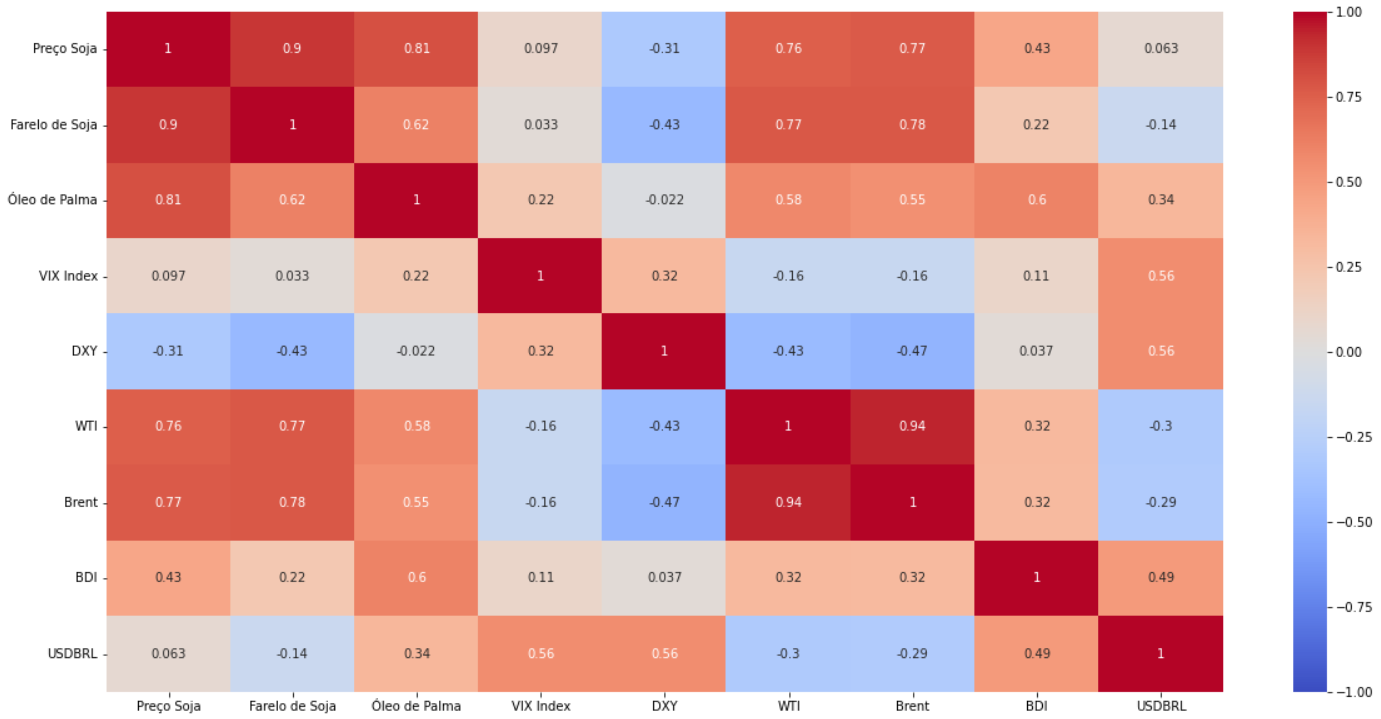
### Documentação dos resultados

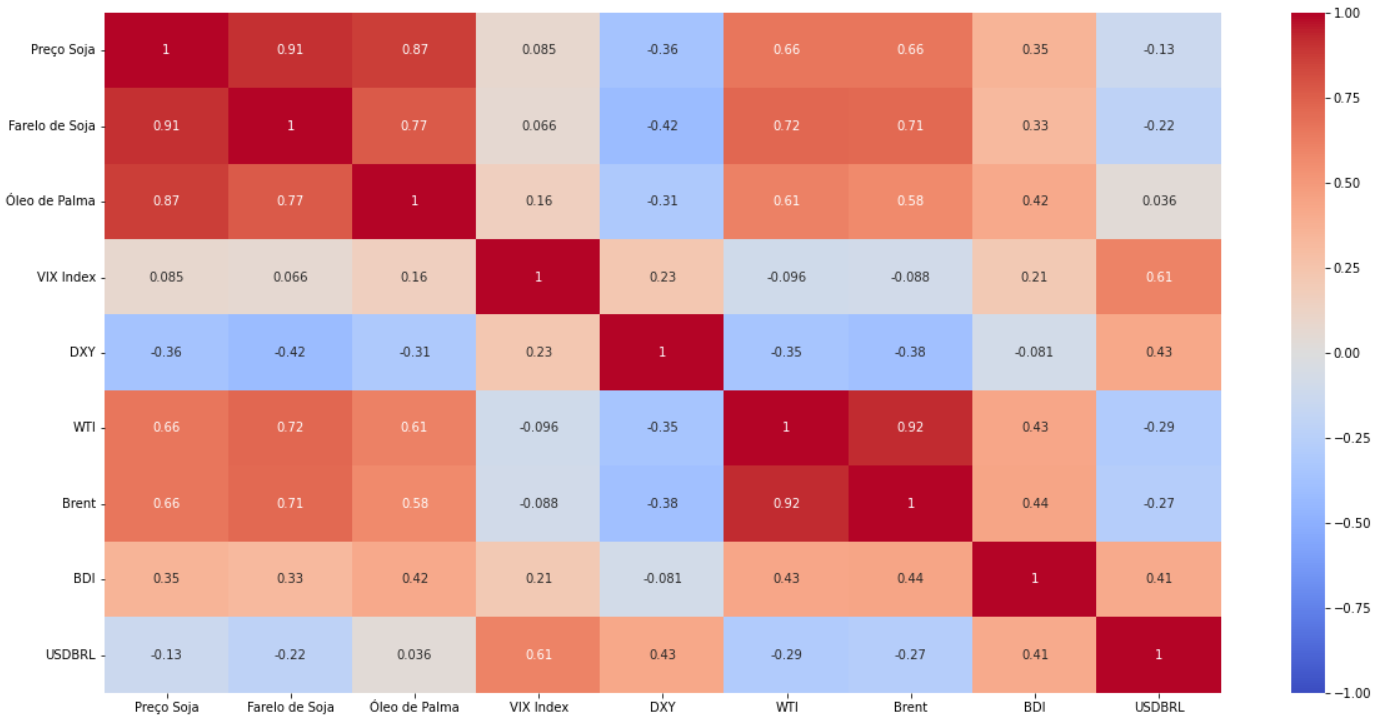
Para a documentação dos resultados foram elaborados dois documentos: Uma apresentação em powerpoint e um repositório do projeto no GitHub.

Nos resultados obtidos da primeira sprint, ressalto o *heatmap* ou mapa de calor como um grande resultado positivo, um mapa de calor é uma técnica de visualização de dados que mostra a magnitude de um fenômeno por meio de cor em duas dimensões.

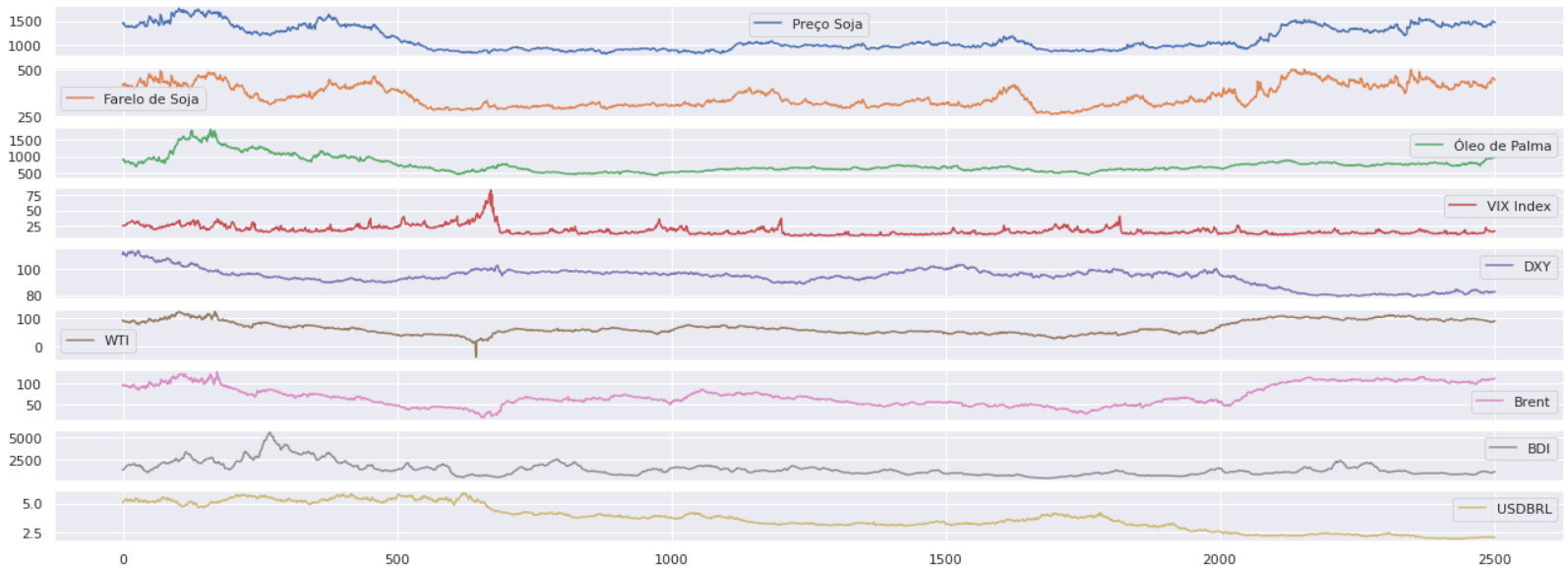
E nesse mapa, nos trabalhamos o conceito de correlação, que é qualquer relação estatística entre duas variáveis. Então cada quadrado em nosso mapa, corresponde a correlação de duas variáveis, indicadas na lateral esquerda e inferior do mapa.

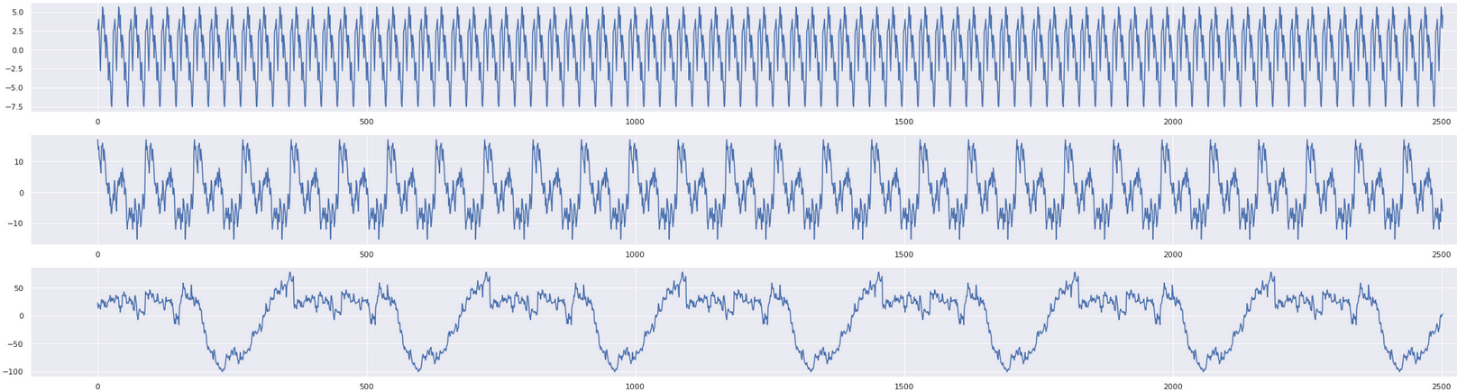
Para construir essas correlações, usamos dois tipos. Correlação de Pearson que mede o grau da correlação entre duas e assume apenas valores entre -1 e 1 e a Correlação de Spearman entre duas variáveis é igual à Correlação de Pearson entre os valores de postos daquelas duas variáveis. Enquanto a correlação de Pearson avalia relações lineares, a correlação de Spearman avalia relações monótonas, sejam elas lineares ou não.

Para Correlação de Pearson, obtivemos:

Para Correlação de Spearman:

Também foram plotados os valores de cada uma das variáveis, para se observar o desenvolvimento delas individualmente com o passar do tempo:



E nosso ultimo resultado foi a decomposição de tendencias, olhando pra sazonalidade. A decomposição de Tendência Sazonal é um método robusto de [decomposição da série de tempo](https://doc.arcgis.com/pt-br/insights/2022.3/analyze/temporal-decomposition-forecasting.htm) frequentemente usado em análises econômicas e ambientais. Esse método usa modelos de regressão ajustados localmente para decompor uma série de tempo em tendência, sazonalidade e componentes remanescentes. Avaliamos o preço da soja e procuramos sazonalidades em períodos de 30, 90 e 365 dias.

Sendo estes, os resultados da primeira Sprint.