**Projeto UC – Inteligência Artificial**

**Professores Responsáveis:**

● Cleber Silva

● Vinicius

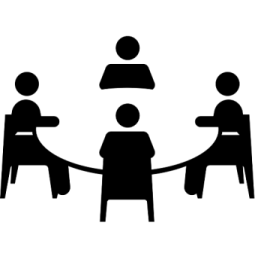
**Grupo:**

Aluno1: Jorge Leandro Piva RA: 820268722

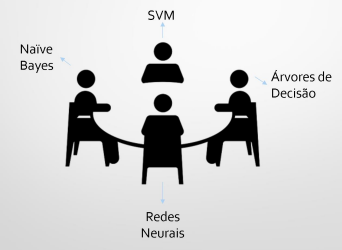
Aluno2: Valério Vanuzi Pereira RA: 32213195

**O que será o projeto?**

O projeto da UC de Inteligência Artificial será fazer um Comitê de Classificadores



Ou seja, cada grupo analisará uma base de dados e aplicará métodos de IA para chegar a algum resultado, após a aplicação desses métodos, deverá compará-los para chegar à conclusão de qual teve o melhor resultado, melhor custo benefício, etc.



**Quais os métodos que poderão ser aplicados?**

● Árvores de Decisão

● Naive Bayes

● Redes Neurais

● SVM

● KNN

● etc

**Base de dados**

Foi escolhida um dataset no Kaggle sobre Fraudes bancárias. Esta UC está sendo ministrada simultaneamente com a UC de Análise de dados e Big Data, desta forma utilizamos o mesmo dataset para as duas atividades A3, porém, estamos utilizando técnicas de análise exploratória com Pandas e Numpy além de criação de alguns gráficos com Matplotlib e Seaborn para facilitar algumas visualizações, estas técnicas nos permite conhecer melhor os dados que serão utilizados nos algoritmos.

O dataset pode ser encontrado no site do Kaggle no link.

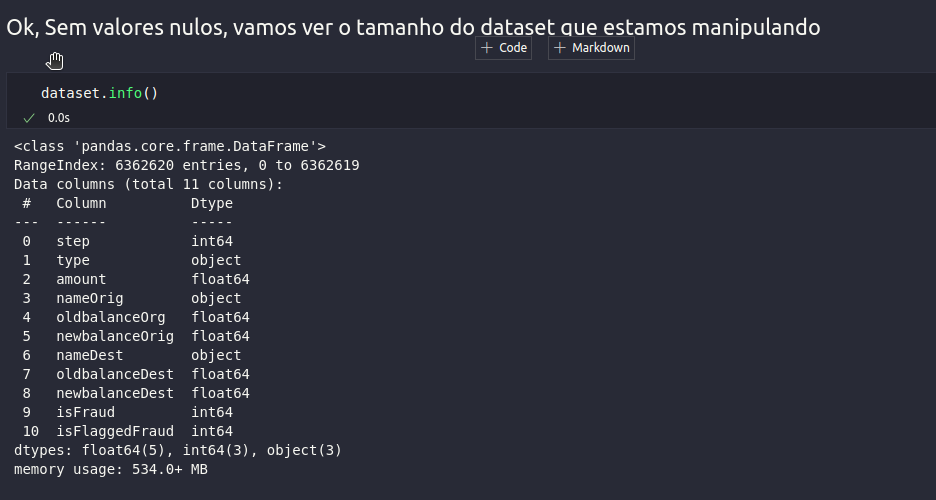
<https://www.kaggle.com/datasets/chitwanmanchanda/fraudulent-transactions-data>

Se por algum motivo o Kaggle remover o dataset do ar fizemos um backup no google drive que pode ser baixado por qualquer pessoa através do link:

<https://drive.google.com/file/d/1Uwtd9_sW51Au0M1RtDBCGHL7Q9Ar14Gu/view?usp=sharing>

Este dataset é muito grande, já que estamos falando de IA e Análise de Dados e “Big Data” tentamos pelo menos trazer um V desse big data com um conjunto de dados que não seria possível nem abri-lo no excel, ou seja, teriamos que analisar com técnicas mais avançadas.

As dimesões desse relatório são 6362620 linhas e 11 colunas, após subir o CSV em um dataset do pandas foi possível utilizar a função info para descobrir essas dimensões.

Este dataset possuí 11 colunas sendo elas

**step** - mapeia uma unidade de tempo no mundo real. Neste caso, 1 passo é 1 hora de tempo. Total de etapas 744 (simulação de 30 dias).

**type**- CASH-IN, CASH-OUT, DEBIT, PAYMENT e TRANSFER.

**amount**- valor da transação em moeda local.

**nameOrig** - cliente que iniciou a transação

**oldbalanceOrg** - saldo inicial antes da transação

**newbalanceOrig** - novo saldo após a transação

**nameDest** - cliente que é o destinatário da transação

**oldbalanceDest** - destinatário do saldo inicial antes da transação. Observe que não há informações para clientes que começam com M (Comerciantes).

**newbalanceDest** - novo destinatário do saldo após a transação. Observe que não há informações para clientes que começam com M (Comerciantes).

**isFraud** - São as transações feitas pelos agentes fraudulentos dentro da simulação. Neste conjunto de dados específico, o comportamento fraudulento dos agentes visa lucrar tomando controle ou contas de clientes e tentando esvaziar os fundos transferindo para outra conta e depois sacando do sistema.

**isFlaggedFraud** - O modelo de negócios visa controlar transferências massivas de uma conta para outra e sinaliza tentativas ilegais. É uma flag criada pelo próprio modelo.

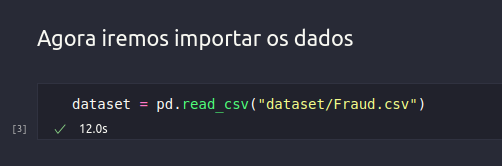
**Como chegar aos resultados?**

● Estudar a base para entende-la

Toda a análise exploratória, os gráficos gerados além dos algoritmos utilizados estão disponíveis no github no repositório: <https://github.com/jorgelpiva/A3_IAvsBigData>

A análise está sendo conduzida em um jupyter notebook [analise\_exploratoria.ipynb](https://github.com/jorgelpiva/A3_IAvsBigData/blob/main/analise_exploratoria.ipynb)

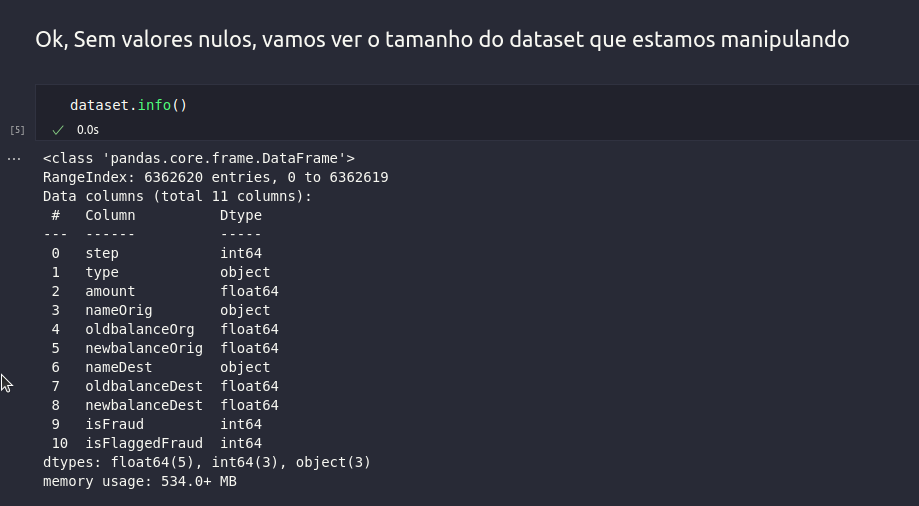
Primeiramente importamos o arquivo.



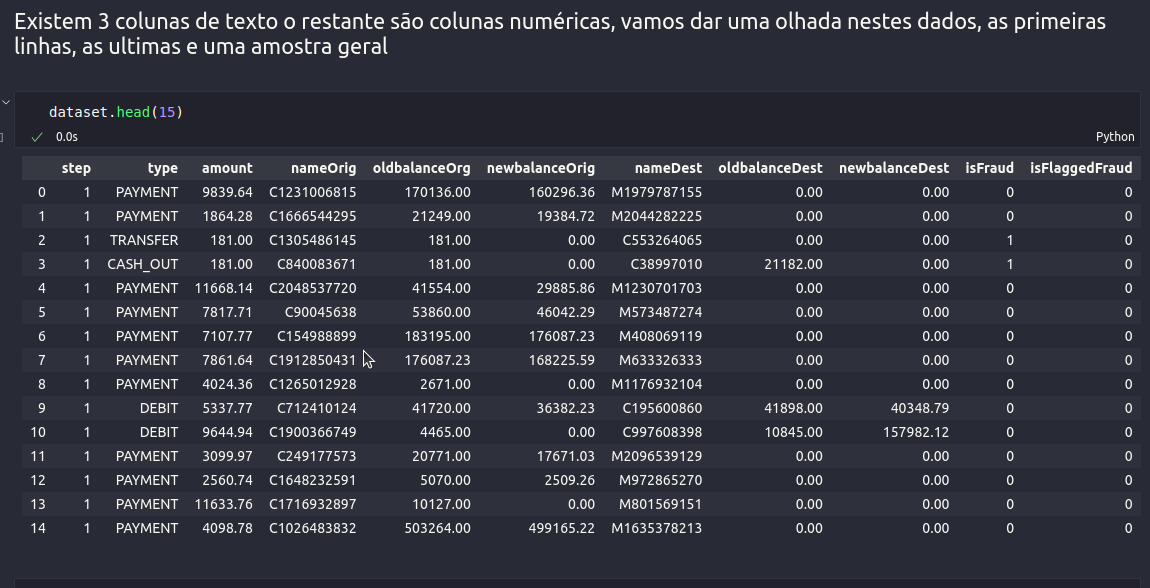
●

Depois verificamos se haviam valores nulos:

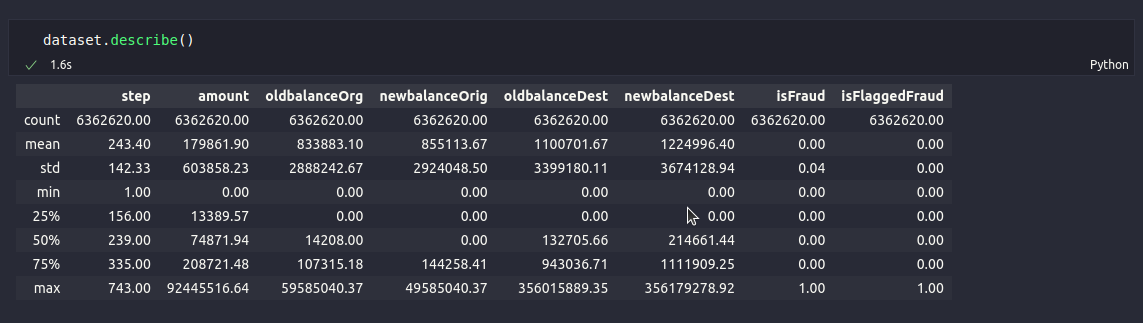
Depois coletamos maiores informações sobre a base de dados.



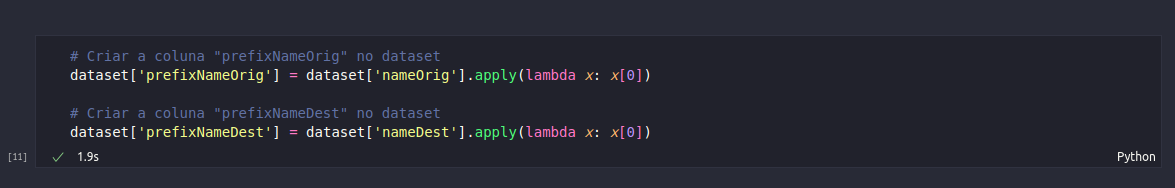
Depois demos uma olhada nos dados



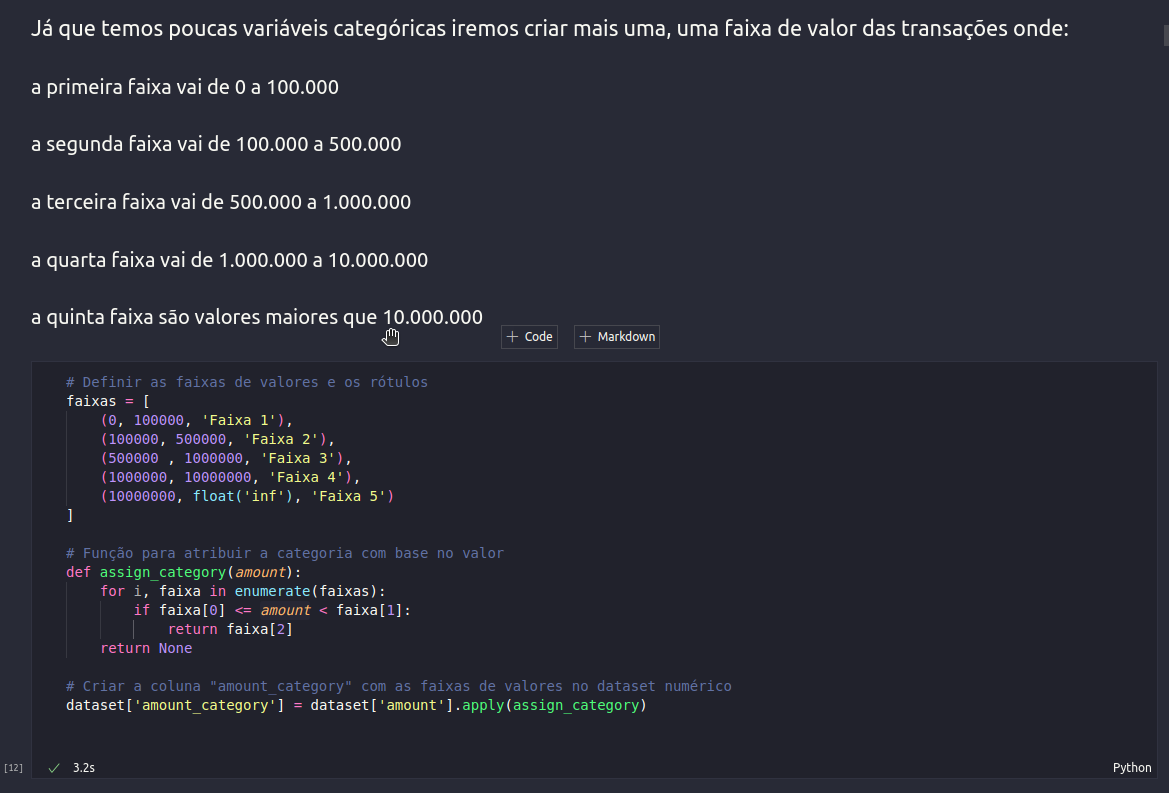
Também utilizamos a função describe para entender os valores.



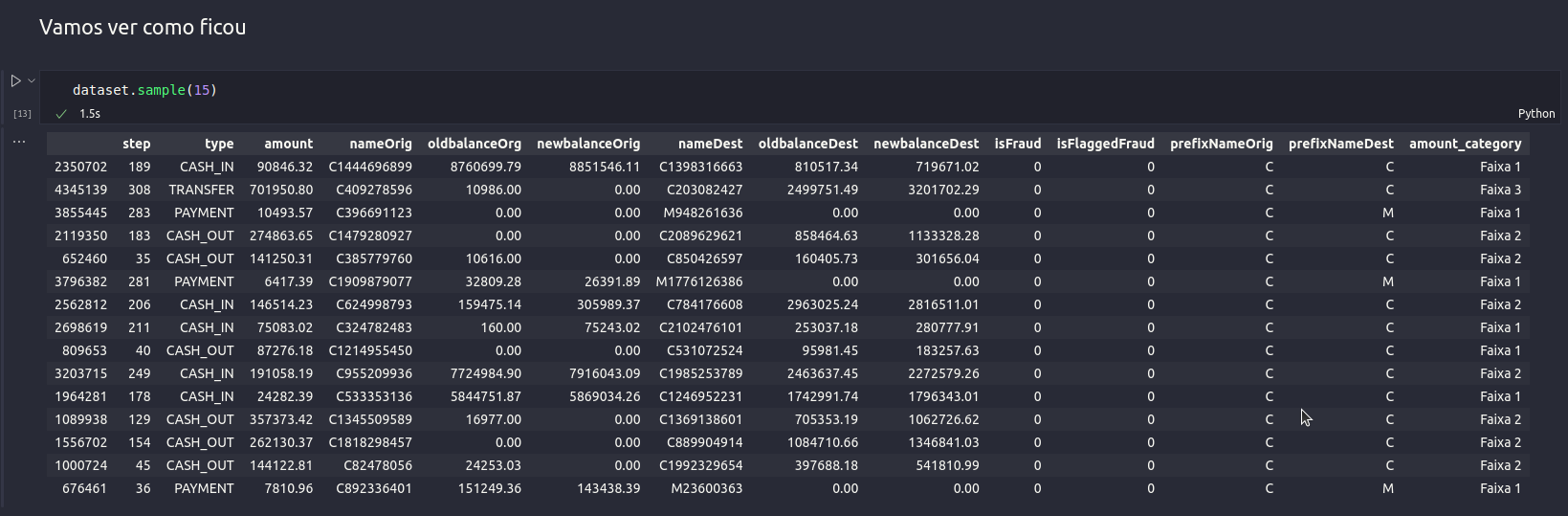
Depois criamos mais 3 colunas no dataset o prefixo do nameDest, o Prefixo do nameOri que seria a primeira letra de cada nome.



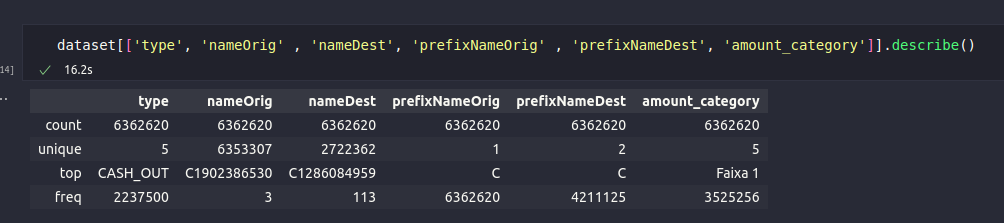
Depois criamos uma faixa para os valores das transações.



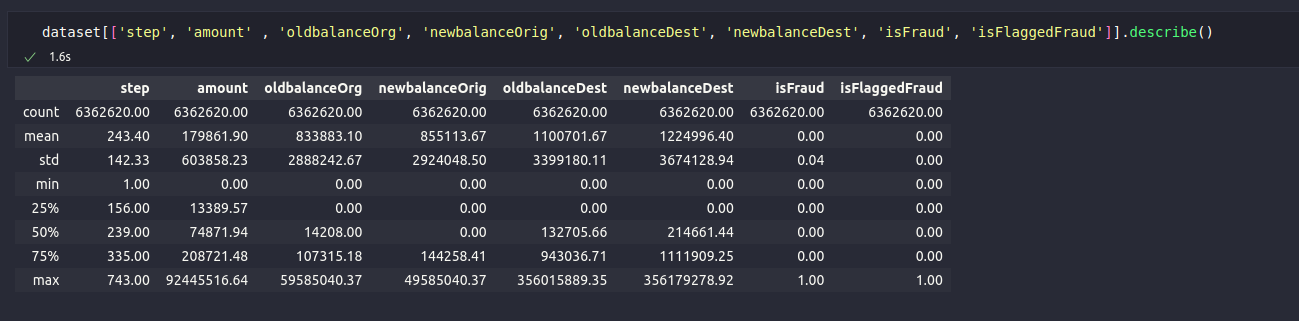
Depois demos mais uma olhadinha de como ficou:



Aí fizemos o describe só nas colunas de texto.



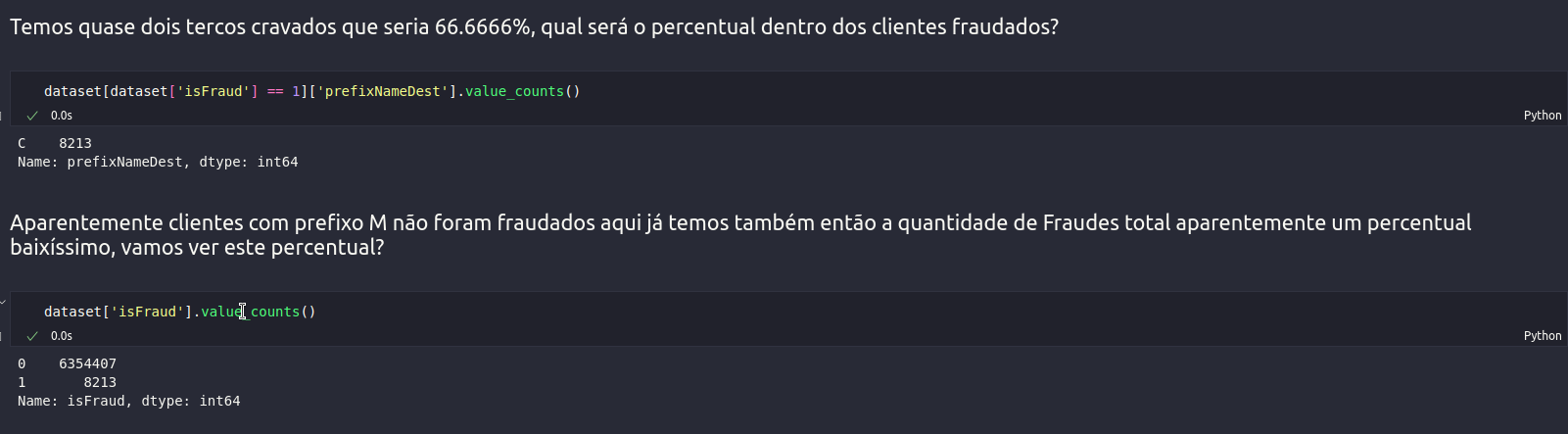
E somente nas colunas numéricas também.



Então descobrimos que existiam dois prefixos de nameDest, C e M



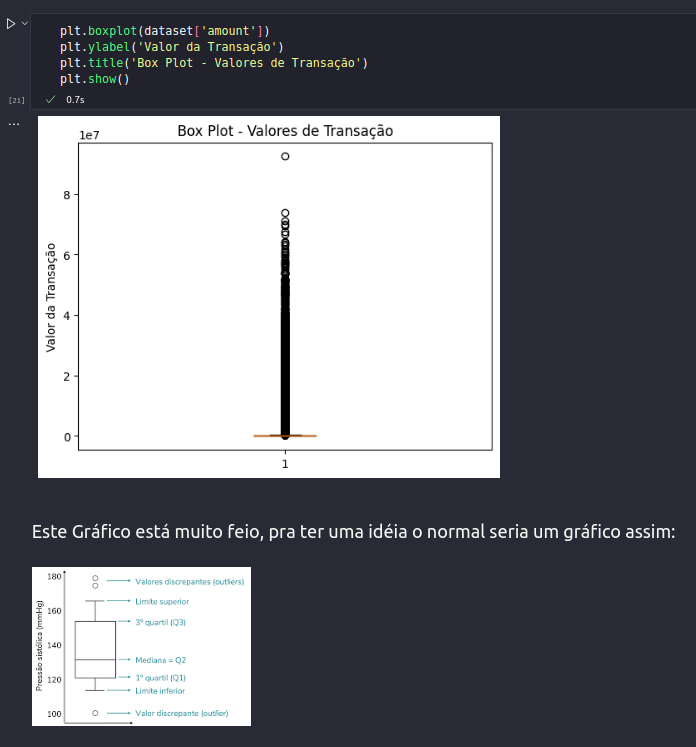
Descobrimos que 1/3 dos prefixos era M mas nenhum deles foi fraudado.



E junto com esta descoberta veio uma informação bombástica, a quantidade de fraudes era muito pequena, dentro de uma base de 6 milhões de registros apenas 8 mil eram fraudulentas, o que dificulta muito para os algoritmos de IA e para nós mesmo conduzirmos análises, a partir deste momento, deveríamos trabalhar sempre olhando para o geral e para as operações fraudulentas de forma apartada.

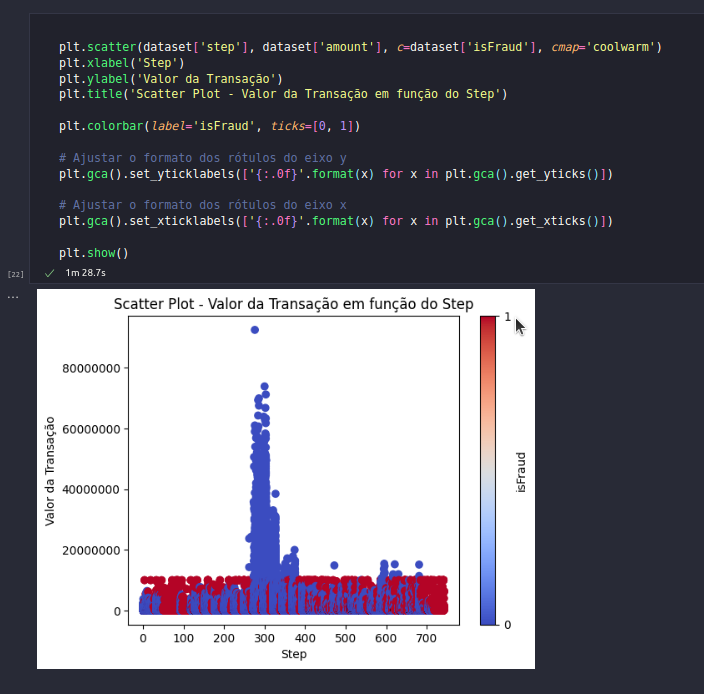


Tentamos identificar outliers.



E descobrimos que tinhamos muitos outliers.

Então plotamos um scatterplot e ficou evidente que tínhamos um outlier, um momento de sazonalidade entre as transações ok e havia regularidade entre as transações fraudadas.



Para este estudo a nossa variável target foi o **isFraud.**