

A solução foi construída com python sobre o LangChain e integrado a API do streamlit.

Estrutura da solução

A solução foi organizada em camadas, mantendo separação de responsabilidades:

- 1- Interface: Responsável por toda a interação com o usuário como as perguntas, exibição de resultados, histórico, etc.
- 2- Camada de estado: Guarda **informações persistentes** entre reexecuções automáticas do Streamlit como por exemplo o histórico de perguntas e o arquivo carregado.
- 3- Camada de lógica de negócio: É onde entra a LLM com o Groq e ocorre análise dos dados e geração de respostas ou gráficos.
- 4- Camada de arquivos: É onde ficam salvos os resultados e os gráficos onde são limpos após uma nova pergunta.

A LLM pode receber:

Perguntas livres sobre o dataset.

Solicitações de estatísticas descritivas.

Requisições de geração de gráficos. Se mais de um gráfico gera um arquivo zip.

Interpretação da resposta da LLM

Visualização

Geração de gráficos usando matplotlib e caso tenha mais de um arquivo é gerado um arquivo .zip para download.

Website da aplicação:

<https://phns-csv-agent.streamlit.app/>

O projeto pode ser encontrado no repositório a seguir na branch "feature-streamlit":

<https://github.com/paulohns/phns-csv-agent/tree/feature-streamlit>

Perguntas realizadas a LLM:

P: Qual o tipo dos dados?

R: As variáveis 'Time' e 'Class' são do tipo int64, enquanto as demais variáveis são do tipo float64.

P: Qual o intervalo de cada variável (mínimo, máximo)?

R: O intervalo de cada variável é:

Time: 0.0 a 2.0

V1: -1.35981 a 1.19186

V2: -1.34016 a 0.877737

V3: 0.16648 a 2.53635

V4: -0.863291 a 1.37816

V5: -0.503198 a 0.0600176

V6: -0.0823608 a 1.8005

V7: -0.078803 a 0.791461

V8: -0.270533 a 0.377436

V9: -1.51465 a 0.817739

V10: -0.166974 a 0.753074

V11: -0.5516 a 1.61273

V12: -0.617801 a 1.06524

V13: -0.99139 a 1.34585

V14: -1.11967 a -0.143772

V15: -0.631418 a 2.34586

V16: -2.89008 a 0.463917

V17: -0.684093 a 1.10997

V18: -0.183361 a 1.96578

V19: -2.26186 a 0.803487

V20: -0.208038 a 0.52498

V21: -0.225775 a 0.247998

V22: -0.638672 a 0.798278

V23: -0.190321 a 0.909412

V24: -1.17558 a 0.141267

V25: -0.327642 a 0.16717

V26: -0.221929 a 0.502292

V27: -0.055353 a 0.219422

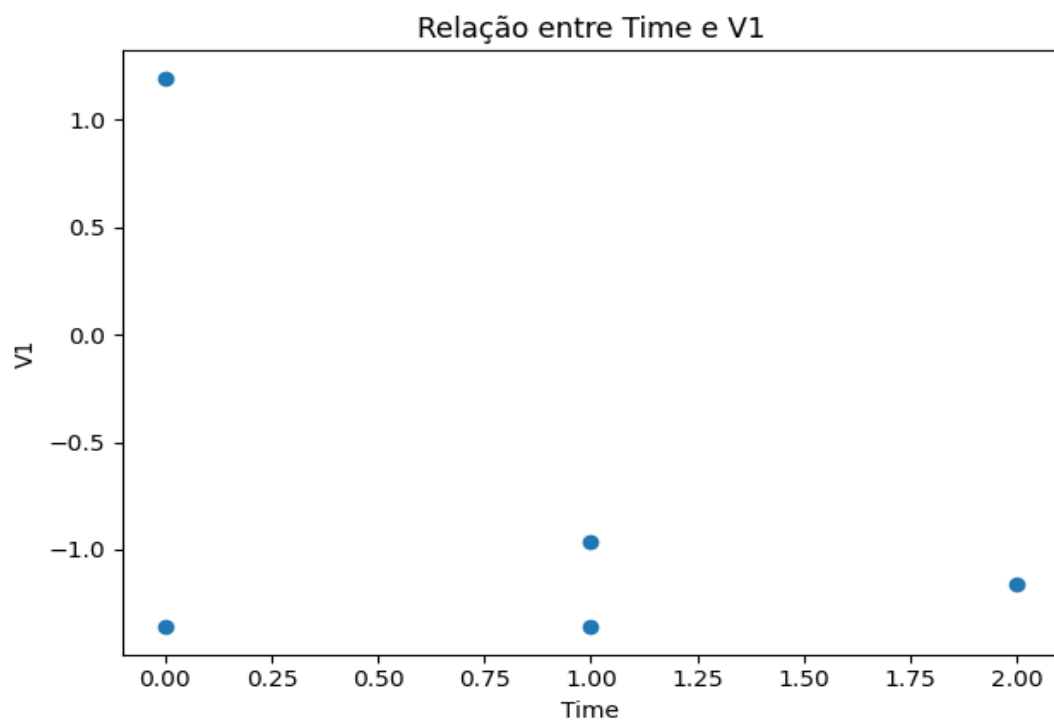
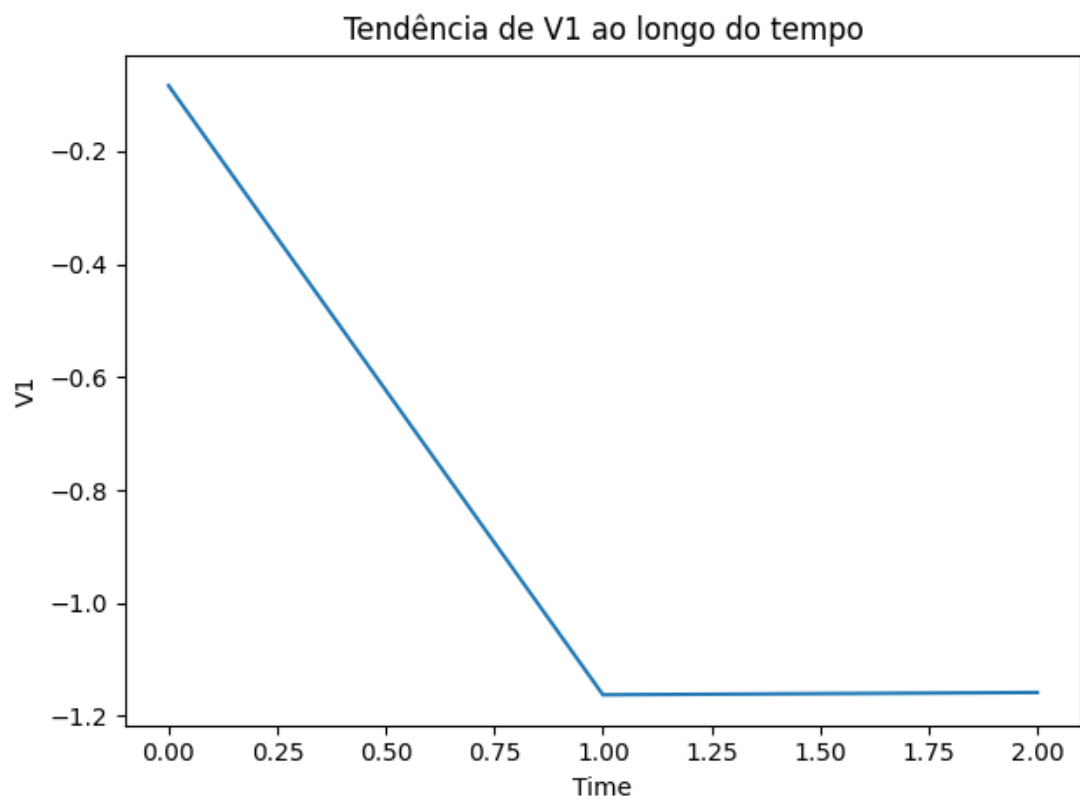
V28: -0.059752 a 0.215153

Amount: 2.69 a 378.66

Class: 0.0 a 0.0

Thought: Esses são os intervalos de cada variável no DataFrame. É importante notar que esses valores podem ser úteis para entender a distribuição dos dados e identificar possíveis outliers.

P: Existem padrões ou tendências temporais?
R: Retornou graficos.zip, segue dois exemplos:



P: Quais são as medidas de tendência central (média, mediana)?

R: As medidas de tendência central (média e mediana) para as colunas numéricas do dataframe são: - Time: média aproximadamente 94813.86, mediana 84692 - V1: média aproximadamente 0, mediana aproximadamente 0.07 - V2: média aproximadamente 0, mediana aproximadamente 0.18 - V3: média aproximadamente 0, mediana aproximadamente 0.18 - V27: média aproximadamente -0.0037, mediana aproximadamente 0.00134 - V28: média aproximadamente -0.00123, mediana aproximadamente 0.01124 - Amount: média aproximadamente 88.35, mediana 22 - Class: média aproximadamente 0.0017, mediana 0

P: Qual a variabilidade dos dados (desvio padrão, variância)?

R: A variabilidade dos dados, medida pelo desvio padrão e pela variância, varia bastante entre as colunas numéricas. Por exemplo, a coluna 'Time' apresenta um desvio padrão de aproximadamente 47.488 e uma variância de cerca de 2.255×10^9 , indicando alta dispersão. As colunas de variáveis como V1 a V28 têm desvios padrão entre aproximadamente 0.33 e 1.96, com variâncias correspondentes, refletindo menor dispersão. A coluna 'Amount' possui um desvio padrão de aproximadamente 250.12 e uma variância de cerca de 62.560, indicando alta variabilidade nos valores de valores monetários. A coluna 'Class' tem um desvio padrão de aproximadamente 0.0415 e uma variância de cerca de 0.00172, indicando baixa dispersão, pois é uma variável de classificação binária.

P: Quais os valores mais frequentes ou menos frequentes?

R: Os valores mais frequentes e menos frequentes no dataset são, respectivamente, '0' e '95.63'.
