

Relatório do Projeto de Data Science

Análise das Perdas Militares da Rússia na Guerra da Ucrânia

Curso: CCC269 - Data Science

Aluno: Welliton Slaviero

Matrícula: 178342

Professor: Prof. Dr. Carlos Amaral Hölbig

Trabalho: Projeto de Data Science

Data: Junho de 2025

1. Tema do Projeto

Este projeto tem como objetivo analisar as perdas militares da Rússia durante o conflito na Ucrânia, utilizando dados oficiais coletados entre 2022 e 2025. A análise visa identificar padrões nas perdas de equipamentos e pessoal militar ao longo do tempo, aplicando técnicas de Data Science para extrair insights relevantes sobre a intensidade e evolução do conflito.

2. URL do Projeto no GitHub

Repositório: <https://github.com/wellsla/analise-guerra-na-ucrania-ds-upf>

3. Dataset Utilizado

3.1 Origem dos Dados

- **Fonte:** Kaggle - "2022 Ukraine Russian War Dataset"
- **URL:** <https://www.kaggle.com/datasets/piterfm/2022-ukraine-russian-war>
- **Período:** Fevereiro de 2022 a Junho de 2025
- **Frequência:** Dados diários

3.2 Estrutura dos Dados

O dataset é composto por três arquivos principais:

3.2.1 `russia_losses_equipment.csv`

- **Registros:** 1.207 linhas

- **Colunas:** 19 variáveis
- **Principais variáveis:**
 - date: Data do registro
 - day: Dia do conflito
 - aircraft: Aeronaves perdidas (acumulado)
 - helicopter: Helicópteros perdidos (acumulado)
 - tank: Tanques perdidos (acumulado)
 - APC: Veículos blindados perdidos (acumulado)
 - field artillery: Artilharia perdida (acumulado)
 - drone: Drones perdidos (acumulado)

3.2.2 russia_losses_personnel.csv

- **Registros:** 1.207 linhas
- **Colunas:** 5 variáveis
- **Principais variáveis:**
 - date: Data do registro
 - personnel: Pessoal perdido (acumulado)
 - POW: Prisioneiros de guerra

3.3 Transformações Realizadas

1. **Limpeza de Dados:** - Conversão de datas para formato datetime - Tratamento de valores nulos - Renomeação de colunas para português
2. **Criação de Novas Variáveis:** - Cálculo de perdas diárias por diferenciação - Criação de variáveis temporais (ano, mês) - Correção de valores negativos (ajustes nos dados originais)
3. **Integração de Dados:** - Merge entre datasets de equipamentos e pessoal - Seleção de variáveis principais para análise

4. Metodologia e Workflow

O projeto seguiu o workflow de Data Science apresentado em aula, implementando os seguintes elementos:

4.1 Coleta e Organização de Dados (pandas)

- Importação dos arquivos CSV
- Estruturação em DataFrames
- Verificação da qualidade dos dados

4.2 Análise Exploratória - Transformação (pandas)

- Limpeza e padronização dos dados
- Cálculo de estatísticas descritivas
- Criação de variáveis derivadas

4.3 Visualização de Dados (matplotlib, seaborn)

- Gráficos de evolução temporal das perdas
- Análise de perdas diárias médias por equipamento
- Visualizações dos clusters identificados

4.4 Machine Learning - Clustering (scikit-learn)

- Aplicação do algoritmo K-Means
- Identificação de padrões de intensidade de combate
- Análise de 3 clusters distintos de atividade militar

5. Principais Resultados

5.1 Estatísticas Gerais

- **Período analisado:** 1.207 dias de conflito
- **Perdas totais de pessoal:** Aproximadamente 1.000.000 baixas
- **Perdas de equipamentos:** Mais de 10.000 tanques, 400+ aeronaves

5.2 Padrões Identificados

5.2.1 Evolução Temporal

- Crescimento contínuo das perdas acumuladas
- Aceleração das perdas em períodos específicos
- Drones representam o maior volume de perdas de equipamentos

5.2.2 Análise de Clustering

A análise identificou 3 clusters principais: - **Cluster 0:** Dias de baixa intensidade (perdas menores) - **Cluster 1:** Dias de intensidade moderada - **Cluster 2:** Dias de alta intensidade (perdas elevadas)

5.3 Insights Principais

1. As perdas de drones superam significativamente outros equipamentos
2. Existe uma correlação entre perdas de pessoal e equipamentos
3. Períodos de alta intensidade de combate são identificáveis pelos clusters

6. Tecnologias Utilizadas

6.1 Linguagem e Ambiente

- **Python 3.x** no Google Colab
- **Pandas** para manipulação de dados
- **NumPy** para cálculos numéricos

6.2 Visualização

- **Matplotlib** para gráficos básicos
- **Seaborn** para visualizações estatísticas

6.3 Machine Learning

- **Scikit-learn** para algoritmos de clustering
- **K-Means** para identificação de padrões

7. Limitações e Considerações

7.1 Limitações dos Dados

- Dados baseados em fontes oficiais ucranianas
- Possíveis inconsistências em correções retrospectivas
- Ausência de dados de perdas ucranianas para comparação

7.2 Limitações da Análise

- Análise focada apenas em aspectos quantitativos
- Não considera fatores geopolíticos ou estratégicos
- Clustering baseado apenas em perdas diárias

8. Conclusões

Este projeto demonstrou com sucesso a aplicação do workflow de Data Science em um contexto real e relevante. As principais conquistas incluem:

1. **Organização eficiente** de dados complexos de múltiplas fontes
2. **Transformação adequada** dos dados para análise temporal
3. **Visualizações claras** que revelam padrões importantes
4. **Aplicação bem-sucedida** de algoritmos de Machine Learning

A análise revelou padrões interessantes sobre a intensidade do conflito e a evolução das perdas militares, fornecendo uma base sólida para futuras análises mais aprofundadas.

9. Trabalhos Futuros

Possíveis extensões deste projeto incluem: - Análise preditiva das perdas futuras - Correlação com eventos geopolíticos específicos - Análise de dados georreferenciados por região - Comparação com dados históricos de outros conflitos

Referências

1. Dataset utilizado: <https://www.kaggle.com/datasets/piterfm/2022-ukraine-russian-war>
2. Documentação do Pandas: <https://pandas.pydata.org/docs/>
3. Documentação do Scikit-learn: <https://scikit-learn.org/stable/>
4. Documentação do Matplotlib: <https://matplotlib.org/stable/contents.html>
5. Documentação do Seaborn: <https://seaborn.pydata.org/>

*Trabalho desenvolvido para a disciplina CCC269 - Data Science
Instituto de Tecnologia - Universidade de Passo Fundo*