UNESA - Universidade Estácio de Sá - Rio de Janeiro
Campus Niterói

TÍTULO DO PROJETO DE EXTENSÃO

Docentes: Matheus Sant'Ana Rosa & Thiago Alves Fraga
Orientadora: Simone Ingrid Monteiro Gama

2025
Niterói / Rio de Janeiro

Sumário

1. DIAGNÓSTICO E TEORIZAÇÃO

1.1. Problemática

O elevado número de pessoas desaparecidas no estado do Rio de Janeiro ao longo das últimas décadas representa um problema social de grande impacto. Apesar da relevância do tema, a população possui pouco acesso a análises claras e acessíveis sobre a evolução desses dados.

Durante conversas e trocas com membros da instituição de ensino, foi identificada a falta de informações consolidadas e à ausência de estudos que indiquem tendências ou padrões ao longo do tempo. Diante dessa demanda sociocomunitária, o projeto de extensão busca analisar, por meio de técnicas de Big Data, os registros mensais de desaparecimentos entre 2003 e 2023.

Para isso, foram utilizados Python no Google Colab, regressão linear, frequência absoluta e relativa, além de representações gráficas com a biblioteca Matplotlib. O objetivo é oferecer uma leitura objetiva do fenômeno, contribuindo para a compreensão pública e possíveis ações preventivas.

1.2. Objetivos

O objetivo deste trabalho de extensão é apresentar uma análise de dados Big Data dos dados das pessoas desaparecidas do município do Rio de Janeiro.

1.3. Referencial teórico (subsídio teórico para propositura de ações da extensão)

O projeto de Big Data sobre Pessoas Desaparecidas no Rio de Janeiro é orientado por referenciais teóricos que subsidiam tanto a metodologia de análise (o *como fazer*) quanto a justificativa para as ações de extensão propostas (o *porquê fazer*). A situação-problema a alta e persistente taxa de desaparecimentos é abordada sob as lentes da Ciência de Dados e do Impacto Social da Tecnologia.

1. Big Data e Análise Preditiva como Ferramenta Cívica

A primeira base teórica justifica a metodologia do projeto: a utilização de grandes volumes de dados (Big Data) e métodos analíticos para resolver problemas complexos da sociedade.

- **Boyd e Crawford (2012):** Estes autores definem e discutem as implicações do Big Data, enfatizando que a coleta e o processamento de dados em grande escala não

são apenas sobre volume, mas sobre a capacidade de encontrar novos *insights* e padrões que seriam invisíveis em amostras menores. No nosso projeto, o volume de 20 anos de dados de desaparecimentos (Jan/2003 a Out/2023) exige ferramentas como o PySpark, alinhando-se à necessidade de processamento distribuído de Big Data. A teoria deles valida o uso de técnicas de ponta para a extração de valor cívico.

- **A Regressão Linear na Projeção de Tendências:** O uso de Regressão Linear se justifica teoricamente como uma técnica de modelagem preditiva. Conforme a estatística clássica, a regressão permite estabelecer uma relação funcional entre a variável tempo (Ano) e a variável social (Taxa de Desaparecidos), dando subsídio teórico para projetar cenários futuros. Isso transforma o dado histórico em uma ferramenta para *planejamento estratégico* de ações preventivas.

2. A Inserção de Dados no Debate Social (Extensão Universitária)

O segundo pilar teórico estabelece a ponte entre a análise técnica e a propositura de ações de extensão, garantindo que o projeto não seja apenas um exercício acadêmico.

- **Demo (2000)**: O conceito de Extensão Universitária por Demo exige que o conhecimento produzido na academia retorne à sociedade como instrumento de transformação social. A análise dos dados de desaparecidos se torna o *subsídio teórico-científico* que ele menciona, orientando a propositura de ações. Ou seja, as descobertas sobre a sazonalidade (pela Frequência Relativa) e as áreas de maior incidência (se analisado) devem informar diretamente a criação de campanhas de conscientização ou o desenvolvimento de *dashboards* públicos. A teoria de Demo valida o impacto social do projeto.
- **Habermas (1984)**: Embora complexo, o conceito de Ação Comunicativa de Habermas se aplica ao projeto de extensão. Ele propõe que a validade das proposições deve ser alcançada pelo diálogo racional. Ao transformar os dados brutos em informações claras e acessíveis (por meio das distribuições de frequência e visualizações feitas com MatPlotLib), o projeto busca inserir fatos objetivos no debate público. Isso é crucial para que as ações de extensão propostas sejam aceitas e implementadas de forma legítima pelos órgãos competentes e pela comunidade.

Conclusão Teórica:

A articulação entre a capacidade de processamento do Big Data (Boyd & Crawford) e o imperativo ético da extensão universitária (Demo) fornece a base teórica completa para o projeto. O detalhamento técnico (Pandas, PySpark, Frequência, Regressão) é o meio científico para cumprir o fim social: usar a análise preditiva para justificar e orientar a criação de ações de extensão que visam reduzir a situação-problema dos desaparecimentos no Rio de Janeiro.

2. PLANEJAMENTO E DESENVOLVIMENTO DO PROJETO

2.1. Detalhamento técnico do projeto

No inicio da disciplina extensionista de tópicos de Big Data 2025.2, foram disponibilizados diversas bases de dados (Big Data) para escolher aquelas que melhor atendem ao impacto social exigido na disciplina.

A equipe deste trabalho escolheu a base de dados : **Pessoas desaparecidas, por mês, total, diferenças com relação ao ano anterior, população e taxa por cem mil habitantes, segundo o ano, no Município do Rio de Janeiro entre Janeiro/2003 e Outubro/2023**

[https://www.data.rio/documents/ebef6fa9e5dd4301a0903679af8acccf/about] e foram realizados:

1. Limpeza e Preparação da Base de Dados

- **Python:** Linguagem de programação central utilizada para todas as etapas de manipulação e análise.
- **Pandas:** Empregada para a leitura, manipulação e limpeza estrutural dos dados, incluindo o tratamento de valores ausentes , conversão de tipos de dados e a organização das colunas.
- **PySpark:** Utilizado para o processamento distribuído da base, abordando o volume de dados e o requisito de Big Data de forma eficiente.
- **Scikit-learn (Sklearn):** Aplicado na fase de pré-processamento, realizando a padronização ou normalização de variáveis para garantir a adequação dos dados à modelagem.
- **MatPlotLib:** Biblioteca essencial para a visualização gráfica dos dados, como histogramas e gráficos de linha, facilitando a identificação de sazonalidade e tendências de longo prazo.
- **Distribuição de Frequência:** Técnica central para entender a ocorrência dos desaparecimentos. Foi calculada para identificar a concentração dos casos por mês e por ano.
- **Frequência Absoluta:** Demonstra o número total e bruto de desaparecimentos em cada período.
- **Frequência Relativa:** Apresenta a proporção (%) dos desaparecimentos, permitindo comparações justas de impacto entre diferentes períodos.
- **Regressão Linear:** Modelo estatístico implementado para quantificar a relação entre as variáveis temporais e os desaparecimentos.

3. ENCERRAMENTO DO PROJETO

3.1. Relato de Experiência Individual (Pontuação específica para o relato individual)

3.1.1. RESULTADOS E DISCUSSÃO:

Matheus Sant'ana (202208858422):

Ao iniciar o projeto, eu esperava apenas aplicar técnicas de análise de dados, mas a experiência acabou me mostrando o quanto a tecnologia pode contribuir para compreender problemas sociais reais. Durante o desenvolvimento, observei que a maior dificuldade foi integrar diferentes linguagens, ferramentas e processos, principalmente na parte de higienização e padronização dos dados. Em contraste, a modelagem e a análise estatística foram mais simples do que eu imaginava, graças às ferramentas utilizadas.

A vivência me trouxe um sentimento de responsabilidade ao perceber que muitos dados importantes não estão facilmente acessíveis, embora sejam essenciais para entendermos a dimensão dos desaparecimentos na cidade. O trabalho resultou em um aprendizado significativo, tanto técnico quanto pessoal, reforçando a importância de projetos que aproximem dados, tecnologia e sociedade.

Entre minhas principais descobertas, destaco o valor da organização dos dados e a necessidade de tornar informações públicas mais acessíveis. Recomendo que futuros projetos mantenham um cuidado especial com essas etapas iniciais, pois elas influenciam diretamente na qualidade da análise final. No geral, saio dessa experiência mais consciente e motivado a desenvolver soluções que gerem impacto social.

Thiago Alves Fraga (202309135159):

Minha experiência neste projeto de Big Data, focado em Pessoas Desaparecidas, foi de grande aprendizado prático.

Minha expectativa inicial era pular logo para a modelagem (Regressão Linear), mas a realidade vivida me ensinou que o PySpark e o Pandas gastam a maior parte do tempo na limpeza de dados. Descobri que a dificuldade principal não era o código, mas sim garantir a qualidade do *dataset*.

A experiência resultou em uma análise confiável, onde a Distribuição de Frequência e os gráficos de MatPlotLib revelaram padrões importantes de sazonalidade.

A principal descoberta é que a limpeza de dados é o fator mais crítico de sucesso. Minha recomendação é priorizar a configuração do ambiente (como o PySpark) e a limpeza no início, pois a modelagem é a parte mais rápida. Senti-me realizado ao transformar dados brutos em *insights* de impacto social.

3.1.2. CONSIDERAÇÕES FINAIS

Matheus Sant’ana (202208858422):

A partir da análise realizada e do desenvolvimento contínuo do projeto, foram identificados diversos aspectos que podem ser aprofundados em trabalhos futuros, tanto na perspectiva extensionista quanto na pesquisa acadêmica. Embora este projeto tenha se concentrado na análise estatística básica dos desaparecimentos na cidade do Rio de Janeiro entre 2003 e 2023, surgiram demandas e possibilidades que ampliaram significativamente o campo de atuação.

Em primeiro lugar, destaca-se a necessidade de criar mecanismos mais acessíveis de visualização e interpretação desses dados pela população. A comunidade demonstrou interesse em ferramentas que facilitem o acompanhamento de oscilações mensais e anuais, permitindo que famílias, organizações civis e gestores públicos compreendam melhor o comportamento desse fenômeno. Nesse sentido, um trabalho futuro poderia incluir o desenvolvimento de um dashboard interativo, que reúna gráficos, filtros, mapas e indicadores atualizados automaticamente. Plataformas como *Tableau*, *Power BI*, *Looker Studio* ou mesmo aplicações web em *Python* (usando *Flask* ou *Django*) poderiam ser exploradas para atender essa demanda.

Do ponto de vista extensionista, também foi evidenciada a importância de desenvolver ações educativas junto à comunidade, abordando temas como prevenção ao desaparecimento, canais oficiais de denúncia e orientações práticas sobre como proceder em casos emergenciais. Tais ações podem ser realizadas por meio de oficinas, materiais informativos e parcerias com instituições que já trabalham com direitos humanos e suporte às famílias.

Por fim, cabe mencionar que outras soluções tecnológicas poderiam ter sido implementadas no projeto atual, caso houvesse maior disponibilidade de tempo ou recursos. Entre elas, destaca-se a aplicação de técnicas para atualização automática das bases de dados, integração com APIs públicas de segurança, uso de bancos de dados relacionais para armazenamento estruturado e até o desenvolvimento de um aplicativo móvel dedicado à consulta de estatísticas e informações de apoio para a comunidade.

Em síntese, embora o presente trabalho tenha cumprido seu objetivo inicial de realizar uma análise estatística clara e acessível, ele abre caminho para múltiplas frentes de desenvolvimento futuramente — ampliando o impacto social, tecnológico e científico do estudo e fortalecendo a relação entre universidade e comunidade.

Thiago Alves Fraga (202309135159):

O nosso projeto, depois de toda a análise de dados, não termina aqui. Ele é, na verdade, um ponto de partida para um trabalho de maior alcance, tanto na Extensão quanto na Pesquisa.

Em termos de Extensão, a primeira coisa que podemos fazer junto à parte interessada é transformar nossos resultados em algo prático e acessível. Pensando nisso, a perspectiva mais imediata é o desenvolvimento de um Painel de Dados (Dashboard) que mostre em tempo real, ou com atualizações frequentes, a sazonalidade e os picos de desaparecimento que descobrimos com a Distribuição de Frequência. Isso pega o nosso conhecimento técnico (pesquisa) e o transforma em uma ferramenta que a comunidade ou o órgão responsável pode usar para planejar ações preventivas.

Na área de Pesquisa, podemos ir muito além da Regressão Linear. O trabalho futuro envolve a migração para modelos de Séries Temporais mais avançados, como ARIMA ou Prophet. Esses modelos são desenhados especificamente para fazer previsões mais acuradas de desaparecimentos nos próximos meses, o que daria um subsídio ainda mais robusto para a segurança pública.

E olhando para as soluções tecnológicas alternativas, percebemos que poderíamos ter implementado abordagens de Machine Learning para Classificação. Em vez de apenas prever a taxa, poderíamos tentar classificar os casos, por exemplo, por probabilidade de serem resolvidos ou por tipo de vulnerabilidade. Além disso, para a escalabilidade do nosso Big Data, a migração do processamento local (PySpark) para plataformas de nuvem, como o Google Cloud Platform (GCP), teria tornado o projeto final mais robusto e pronto para uso contínuo.

No fim das contas, o valor do projeto está na continuidade: garantir que o nosso *insight* acadêmico se transforme em uma ferramenta de transformação social e em pesquisa de fronteira.