# Trabalho A1: Análise de Dados

## 1. Objetivo

O propósito deste projeto é aplicar técnicas de análise de dados com **Python**, utilizando **obrigatoriamente** as bibliotecas **NumPy** e **Pandas**. O projeto visa também incentivar boas práticas de desenvolvimento e qualidade de software, como documentação, testes unitários, controle de versão e tratamento de erros.

## 2. Divisão de Grupos e Tarefas

**2.1** Turma de **Matemática Aplicada**: o trabalho deve ser realizado em grupos de **3 alunos**. A divisão de tarefas dentro do grupo deve ser clara e descrita em detalhes no relatório final.

**2.2** Turma de **Ciência de Dados**: o trabalho deve ser realizado em grupos de **4 ou 5 alunos**. A divisão de tarefas dentro do grupo deve ser clara e descrita em detalhes no relatório final.

## 3. Entrega e Utilização do GitHub

Todo o desenvolvimento do projeto deve ser feito em um repositório no GitHub, que será utilizado como meio de entrega. O grupo deverá enviar um e-mail para o respectivo professor com o link do repositório até **10/10 às 7h da manhã**. Não haverá prorrogação.

É obrigatório o uso adequado do GitHub, com commits relevantes e consistentes ao longo do período de desenvolvimento. O histórico do repositório será avaliado em conjunto com a descrição detalhada da divisão de tarefas do grupo. Qualquer commit após a data limite acarretará um desconto mínimo de 2 pontos.

## 4. Critérios de Avaliação

- ☐ Clareza e relevância das hipóteses propostas.
- ☐ Complexidade da base de dados selecionada.
- ☐ Qualidade do pré-processamento dos dados.
- ☐ Profundidade da análise exploratória.
- ☐ Eficácia das visualizações geradas para apoiar as hipóteses propostas.
- ☐ Validação das hipóteses propostas.
- ☐ Qualidade da documentação do código.
- ☐ Organização e estrutura do código.
- ☐ Cobertura e qualidade dos testes unitários e utilização correta de tratamento de erros.

- ☐ Uso adequado de controle de versão no GitHub.

# 5. Instruções do Trabalho

O projeto consiste na análise de um conjunto de dados escolhidos pelo grupo, com o objetivo de aplicar os conceitos discutidos na disciplina. As etapas principais incluem:

A. **Escolha de uma Base de Dados:**
   Selecionar um tema de interesse (filmes, música, games, séries, animes, dados governamentais, etc.) e encontrar uma base de dados pública, evitando bases simples. Sugestão: utilizar o **Kaggle** ou pesquisar por **"<Tema> dataset"** no Google.

B. **Levantamento de Perguntas de Negócio e Hipóteses:**
   Definir uma pergunta de negócio ou hipótese por integrante do grupo para endereçar utilizando a base de dados selecionada. As perguntas devem ser desafiadoras e testáveis, relacionadas a padrões, correlações ou comportamentos dos dados.

C. **Pré-processamento dos Dados:**
   Realizar o pré-processamento dos dados, incluindo limpeza, tratamento de valores ausentes, e possíveis normalizações ou transformações dos dados.

D. **Análise Exploratória e Teste de Hipóteses (EDA):**
   Aplicar estatística descritiva e visualização de dados para explorar a base de dados e testar as hipóteses propostas. A análise deve buscar padrões, correlações e observações relevantes, utilizando bibliotecas como **pandas**, **matplotlib** e **seaborn** para criar visualizações efetivas. Não é necessário aplicar testes estatísticos formais, a análise exploratória pode ser suficiente para validar ou refutar as hipóteses levantadas.

E. **Documentação e Relatório:**
   Criar um relatório detalhado que inclua i) a introdução, ii) as hipóteses propostas, iii) o pré-processamento dos dados, iv) os resultados da análise exploratória, v) a validação das hipóteses, vi) os desafios encontrados e vii) as contribuições de cada integrante do grupo. O relatório deve ter entre **3 a 5 páginas**, dependendo da quantidade de gráficos incluídos. Utilize a ferramenta **Sphinx** para combinar texto, código e visualizações, com saída em **HTML** ou **PDF**.

F. **Milestone – Apresentação dos Datasets (23/09):**
   **No dia 23/09**, haverá uma apresentação em sala de aula para verificar a adequação dos datasets escolhidos. **Cada grupo** deverá preparar de **2 a 3 slides**, mencionando o nome dos integrantes, o tema escolhido, a base de dados e as hipóteses que pretendem explorar. Este checkpoint é essencial para evitar problemas na entrega final.

G. **Sugestão de Organização do Código:**
É importante que o repositório GitHub seja bem organizado e de fácil navegação. Uma sugestão de estrutura de arquivos é a seguinte:

```
repo/
      docs/ # Diretório para documentação e relatórios
      src/ # Código-fonte do projeto
      tests/ # Scripts de testes unitários
      data/ # Diretório para os dados utilizados no projeto
      readme.md # Arquivo README explicando o projeto e como executá-lo
      requirements.txt # Arquivo com dependências do projeto (bibliotecas)
```

O arquivo **`readme.md`** deve conter uma descrição clara do projeto, instruções para instalação das dependências e para executar o código. O arquivo **`requirements.txt`** deve listar todas as bibliotecas necessárias para que o projeto funcione corretamente.

# 6. Exemplo

**Base de dados**
https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

**Descrição**
Base de 100.000 pedidos de produtos feitos no marketplace OList entre os anos de 2016 e 2018

**Hipótese**

1. Houve alterações significativas de faturamento por categoria ao longo dos anos de 2016, 2017 e 2018?

**Desafios**

1. As informações estão quebradas em diferentes tabelas, é necessário uni-las
2. Segmentar os pedidos por período
3. Reunir os produtos por categoria e entender quais tiveram maior faturamento em cada ano

**Outras possibilidades**

- Amazon Cell Phones Cleaned Scraped Data
- Anime Recommendations Database
- BITCOIN Historical Datasets 2018-2024
- Steam Games Dataset
- Netflix Movies and TV Shows
- Sleep Health and Lifestyle Dataset
- Mental Health in Tech Survey