**LEIA-ME**

Este documento aborda como foi feita a Limpeza de dados do Sistema JupiterWeb para se obter a base de dados acerca da Permanência e Evasão na USP. Os detalhes específicos estão no script papfe\_rd.r

**Seguintes Bases de Dados Originais fornecidas pela USP:**

1. PET\_FEA\_V2\_: Cada Observação representa um registro de uma habilitação no sistema (podendo haver mais de um registro por aluno e curso) dos ingressantes da graduação no Campus Butantã entre 2011 e 2018. Cada coluna representa uma característica ligada ao aluno, identificado por um ID anonimizado. Dentre elas, temos i) dados pessoais (raça, gênero, renda) advinda dos questionários socioeconômicos FUVEST/ENEM ii) desempenho no exame de admissão iii) curso e desempenho na graduação.
2. PET\_FEA\_BLS2 > Planilha Pto PAPFE: Cada Observação representa a pontuação socioeconômica obtida por um dado aluno em um dado ano.
3. PET\_FEA\_BLS2 > Planilha Tipo PAPFE: Cada Observação representa um tipo de benefício obtido por um dado aluno em um dado ano, com o intervalo inicial e final da concessão
4. PET\_FEA\_BLS2 > Pesquisa CulturaExt: Cada Observação representa o valor de uma mensalidade de um mês específico de um dado tipo de benefício obtido por um certo aluno

**Metodologia de Limpeza dos dados (dicionário das variáveis nas respectivas bases):**

* A base 2) e 3) são combinadas, juntando a nota socioeconômica dos alunos que solicitaram os benefícios quando calouros e os benefícios efetivamente recebidos nesse primeiro ano. Isso gera a base papfe
* Limpamos a base 1) para obter os desfechos chave de cada aluno (permanência ao n-ésimo ano e/ou evasão/graduação). Para tal, um script calcula com base no tempo inicial e de desligamento do estudante na habilitação e seu desfecho, tal como se o estudante se forma, evade, ou continua matriculado etc. É empregado um script específico para lidar com os casos de alunos que aparecem múltiplas vezes. Isso gera a base df\_final8 que contém uma observação por aluno, abordando TODOS os ingressantes na USP entre 2011 e 2018.
* O resultado de juntar as bases papfe com df\_final8 é a construção da base de dados usp\_all, A base conta com o curso e desfecho acadêmico de todos os estudantes dos quais temos dados pessoais. Para os alunos que solicitaram benefícios, temos colunas também para sua pontuação socioeconômica, Moradia-Auxílio (CRUSP) e Moradia-Financeiro (bolsa).
* A partir da Base df\_final8 e papfe, são selecionados todos que solicitaram PAPFE em 2018 e ingressaram na USP em 2018. A partir disso, usa-se a nota de corte do benefício para calcular a running variable e controles para a Regressão Descontínua (RD). Essa base é papfe\_rd\_reg