Insper

Ciência dos dados - ENG

2024 – 2º semestre

Professora: Maria Kelly Venezuela
Professor: Marcio Fernando Stabile Junior

PROJETO 2: ANÁLISE EXPLORATÓRIA DE DADOS

Este documento apresenta as premissas do Projeto 2 de Ciência dos Dados.

Conjunto de dados

Neste projeto, será utilizado **exclusivamente** os microdados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua), conduzida pelo IBGE. O principal objetivo da PNAD Contínua é fornecer informações detalhadas sobre a inserção da população no mercado de trabalho, além de características como idade, sexo e nível de escolaridade.

A PNAD Contínua também possibilita a análise do desenvolvimento socioeconômico do país, oferecendo dados anuais sobre outros aspectos, como formas alternativas de trabalho, trabalho infantil, migração, entre outros. Esta pesquisa é realizada com uma amostra probabilística de aproximadamente 211 mil domicílios a cada trimestre, e os dados são divulgados mensalmente. Essa pesquisa foi implementada experimentalmente em outubro de 2011 e, a partir de janeiro de 2012, passou a ser realizada de forma definitiva em todo o território nacional.

Além dos temas principais, a PNAD Contínua também investiga tópicos suplementares em trimestres específicos:

- **Primeiro Trimestre:** Enfoque nos dados gerais sobre mercado de trabalho, como taxa de desocupação, nível de ocupação, informalidade e características da população ativa. Os dados coletados aqui servem como base de acompanhamento anual.
- **Segundo Trimestre:** Além dos temas gerais, há um módulo suplementar sobre Educação. Este módulo coleta informações detalhadas sobre o nível de escolaridade da população, o acesso ao sistema educacional e outras características relacionadas à formação educacional.
- **Terceiro Trimestre:** Mantém o foco no mercado de trabalho, com dados acumulados do ano até o momento. Não há temas suplementares específicos nesse trimestre.
- **Quarto Trimestre:** Além dos temas principais, um módulo suplementar sobre Acesso à Televisão e à Internet e Posse de Telefone Celular é incluído. Este levantamento investiga a penetração das tecnologias de comunicação nos domicílios brasileiros e sua relação com o nível socioeconômico.

Microdados: Disponibilizado no Blackboard um arquivo Jupyter Notebook ensinando como fazer a leitura dos dados da PNAD Contínua dos quatro trimestres de 2023. **NOTA:** Leia esse código completo antes de utilizar uma das partes que seja mais interessante para adaptar no seu Projeto 2.

Insper

Ciência dos dados - ENG

2024 – 2º semestre

Professora: Maria Kelly Venezuela
Professor: Marcio Fernando Stabile Junior

ATENÇÃO: Qualquer mínima parte do código de um grupo que seja parecida com de um outro grupo, será considerado caso de plágio. Não importa qual seja a motivação que deixou os códigos iguais (ou quase iguais), e nem mesmo se todos entre os grupos sejam muito amigos.

Objetivo

O principal objetivo do Projeto 2 (que irá se estender ao Projeto 3) é **prever uma variável principal em função de demais outras variáveis que podem influenciar em seu comportamento**. Para seu conhecimento, a tabela abaixo mostra como essas variáveis são nomeadas nas áreas de Ciência dos dados e Estatística.

	Ciência dos dados	**Estatística**
Variável principal	*Target*	Variável resposta ou dependente
Demais variáveis	*Features*	Variáveis explicativas ou independentes

O **tema deverá ser proposto pelo grupo**, **considerando uma seleção de colunas microdados da PNAD** que permita responder alguns interesses levantados no tema escolhido.

O **tema deve deixar claro uma pergunta** e o objetivo deve contemplar obrigatoriamente um dos casos abaixo:

- **Previsão de um rótulo** (nesse caso, o *target* é qualitativo e trata-se de uma classificação). Por exemplo, considerando uma *playlist* de uma pessoa, o *Spotify* deve recomendar uma nova música a essa pessoa.
- **Previsão de uma informação numérica** (nesse caso, o *target* é quantitativo). Por exemplo, considerando o lançamento de um empreendimento imobiliário em uma determinada região, qual o preço ideal de venda desse imóvel a partir de suas características e localização?

Temas (sugestões feita por IA):

A PNAD Contínua realizada pelo IBGE, coletada trimestralmente, oferece um vasto conjunto de dados que possibilita investigar diversos temas socioeconômicos e demográficos no Brasil. Dependendo do trimestre, alguns tópicos específicos são abordados, permitindo estudos detalhados em áreas complementares ao mercado de trabalho. Veja **alguns exemplos de temas** (seu grupo pode pensar em

outro tema. Apenas inspiração!) que podem ser investigados com os dados dos quatro trimestres de 2023:

1. **Desigualdade de Gênero no Mercado de Trabalho:**
Dados coletados em todos os trimestres podem ser utilizados para analisar a diferença de inserção entre homens e mulheres no mercado de trabalho. Isso inclui o nível de ocupação, taxa de desemprego, rendimentos médios, e a distribuição setorial por gênero. A desigualdade salarial entre gêneros também pode ser estudada.

2. **Educação e Empregabilidade:**
Segundo trimestre: O módulo suplementar sobre educação oferece informações detalhadas sobre o nível de escolaridade da população. A partir desses dados, é possível investigar como o nível educacional impacta a empregabilidade, os tipos de ocupações e os rendimentos. Pode-se também analisar as relações entre a escolarização e a informalidade no mercado de trabalho.

3. Impacto da Tecnologia no Mercado de Trabalho:
Quarto trimestre: O suplemento sobre Acesso à televisão, Internet e posse de telefone celular pode ser utilizado para investigar a relação entre o uso de tecnologias e a inserção no mercado de trabalho. Estudar o impacto da inclusão digital, como o acesso à Internet, pode revelar sua influência nas oportunidades de emprego e trabalho remoto.

4. Trabalho Informal e Subocupação:
Em todos os trimestres, a PNAD Contínua permite explorar o fenômeno da informalidade no Brasil. É possível analisar a evolução do trabalho informal, comparando a formalização do emprego em diferentes regiões e entre diferentes grupos populacionais (por exemplo, por nível educacional, idade e gênero). Além disso, pode-se investigar o número de pessoas subocupadas, que trabalham menos horas do que gostariam ou precisam.

5. Mobilidade Social e Migração:
Terceiro trimestre: Com os dados sobre migração e o desenvolvimento do mercado de trabalho, é possível estudar como a migração (interna e externa) influencia a mobilidade social. Comparar regiões com maior fluxo migratório permite analisar como os migrantes se integram ao mercado de trabalho e se suas condições de vida melhoram após a mudança.

6. Desemprego Juvenil:
A análise focada nos jovens pode ser feita em qualquer trimestre, usando variáveis relacionadas à faixa etária. A alta taxa de desemprego entre os jovens, que é uma questão relevante no Brasil, pode ser analisada sob diferentes perspectivas, como nível de escolaridade, capacitação profissional e região.

Esses temas mostram como os dados da PNAD Contínua permitem uma análise abrangente das condições de vida e do mercado de trabalho no Brasil, refletindo a evolução econômica e social do país ao longo do ano.

Insper

Ciência dos dados - ENG

2024 – 2º semestre

Professora: Maria Kelly Venezuela
Professor: Marcio Fernando Stabile Junior

Habilidades a serem desenvolvidas no projeto

A condução da análise de dados deve refletir um elevado grau de autonomia dos integrantes do grupo, proporcionando liberdade na escolha do tema e incentivando o aprendizado das técnicas mais apropriadas para uma análise exploratória eficaz.

Essa abordagem permitirá a criação de visualizações que possibilitem a interpretação e comparação das variáveis no conjunto de dados em relação ao tema proposto.

Os alunos são incentivados a utilizar bibliotecas como Matplotlib para a visualização de dados, mas também são encorajados a explorar outras bibliotecas do Python, como Seaborn, a fim de aprimorar a apresentação visual de suas análises. Essa combinação permitirá criar gráficos mais sofisticados e esteticamente agradáveis, facilitando a interpretação dos resultados.

Grupos

O projeto pode ser realizado em grupos de **até três alunos** (pode ser individual, dupla ou trio) ou **quatro alunos com exigência maior em termos de quantidade de variáveis (colunas) na base de dados**.

- Até TRIO: ter pelo menos 7 variáveis explicativas e 1 variável target.
- QUARTETO: ter pelo menos 10 variáveis explicativas e 1 variável target.

Atenção: Se o *dataframe* tiver a coluna com a informação do estado (UF) e for construída a coluna Região a partir do estado, ambas contam como uma variável explicativa, por exemplo.

Estrutura do Projeto

É esperado que o seu projeto seja **autocontido**, ou seja, um leitor que não saiba sobre o que esse projeto se trata deve ser capaz de entender a sua linha de raciocínio. Escreva para um leitor que não possua os mesmos conhecimentos técnicos que você (por exemplo: um aluno do primeiro semestre, que ainda não cursou Ciência dos Dados). Abaixo, apresentamos uma sugestão de estrutura para organizar o seu documento. Se quiser seguir uma estrutura diferente, valide-a primeiro com sua professora.

IMPORTANTE: Independente da estrutura adotada, a qualidade do texto produzido é tão importante quanto a análise em si e também será avaliada. Não adianta obter resultados excelentes se eles não forem comunicados de maneira clara.

Professora: Maria Kelly Venezuela
Professor: Marcio Fernando Stabile Junior

Tópicos para organizar as seções em seu Jupyter Notebook:

A. Introdução

- Detalhar objetivo escolhido para trabalhar neste projeto juntamente com descrição da base de dados (obrigatoriamente tema da PNAD Contínua como dito na página 1). Pesquise trabalhos na literatura que discutam o tema escolhido.
- Para trabalhos acadêmicos, acesse https://scholar.google.com.br/. Guarde as referências estudadas para citá-las no seu projeto.
- No site do G1, é possível encontrar várias matérias feitas apenas com estudos do IBGE para enriquecer a introdução do projeto e escolha de várias no conjunto de dados. Acesse https://g1.globo.com/tudo-sobre/ibge/

B. Minerando Dados e Características do Dataset

- Se necessário (e tenho quase 100% de que será), faça filtro na base de dados tanto de linhas como de colunas em prol do objetivo traçado anteriormente.
- Descreva as variáveis finais que serão utilizadas a partir deste ponto.

C. Análise Exploratória dos Dados

- Realizar uma análise descritiva detalhada das variáveis é fundamental para entender o comportamento da variável *target* em relação a cada *feature* do conjunto de dados, sempre alinhado ao objetivo do problema proposto pelo grupo. É importante investigar como a variável *target* se comporta em cruzamentos com outras variáveis, levando em consideração diferentes combinações: duas variáveis quantitativas, duas variáveis qualitativas ou uma de cada tipo, a depender das informações contidas no conjunto de dados. Como vimos no curso, cada tipo de cruzamento exigirá ferramentas descritivas específicas, como tabelas de frequências relativas para variáveis qualitativas ou gráficos de barras (Aula 03), gráficos de dispersão para variáveis quantitativas (Aula 10) e histogramas ou boxplots para cruzar uma variável quantitativa com uma ou mais qualitativas (Aulas 5 e 7).
- Além disso, será exigido a criação de um *dashboard* (por exemplo, com uso do plt.suplot) que apresente visualizações claras e informativas, permitindo que as análises sejam acessíveis e facilmente interpretáveis. Um *dashboard* bem projetado não apenas facilita a visualização dos dados analisados, mas também suporta a tomada de decisões informadas e a comunicação eficaz dos resultados obtidos.

A tabela a seguir apresenta algumas ferramentas descritivas vistas no curso:

Ferramentas estatísticas

Duas variáveis qualitativas	Tabela cruzadas (com uso de *normalize* adequado ao problema); Gráficos de barras (empilhados ou *stacked*); entre outras.
Duas variáveis quantitativas	Medidas de associação; Gráficos de dispersão; entre outras
Uma variável de cada	Medidas-resumo da variável quantitativa segmentando por rótulo da variável qualitativa; Histograma (ou boxplot) da variável quantitativa segmentando por rótulo da variável qualitativa; entre outras

- *Storytelling* com dados: encontre uma representação gráfica que descreva bem os seus dados e que também favoreça no *storytelling* que pretende fazer ao explicar sua linha de raciocínio às outras pessoas (seja em formato escrito ou em apresentação). Caso tenham interesse em estudar sobre o assunto, vejam neste link a parte Data Visualization. Um trecho com os links dessa seção:

- "**O que estudar**: aprenda sobre **Teoria das Cores** (tem esse vídeo sensacional que explica um pouco em 2 minutos); **Storytelling with Data**, da Cole Nussbaumer (aproveita pra seguir o blog); recomendo também seguir o **blog Nightingale** e participar da comunidade **Dataviz Society**."

D. Conclusão

- Faça conclusão final com detalhes levando em consideração todas as interpretações realizadas no decorrer do projeto.

E. Referências Bibliográficas

- Todas as pesquisas feitas e estudadas que foram relevantes para o desenvolvimento devem ser citadas no projeto.

Insper

Ciência dos dados - ENG

2024 – 2º semestre

Professora: Maria Kelly Venezuela
Professor: Marcio Fernando Stabile Junior

Cronograma

DATA	Finalização:
17/04 (quinta) **DEADLINE** **Até às 23h59**	Cadastro do grupo no Blackboard (todos integrantes do grupo): ✔ Até Trio ou ✔ Quarteto (com rubrica diferente).
24/04 (quinta) **DEADLINE** **Até às 23h59**	No Blackboard (pelo menos um integrante do grupo): ✔ Ter **dados** e **tema** no escopo do projeto (Leitura das seções **Objetivo** e **Estrutura do Projeto: A-Introdução e B-Minerando Dados e Características do Dataset** descritos acima no enunciado do Projeto 2). **Destacando que se você não cumprir com esse deadline, já estará atrasado com o Projeto 2.**
ENTREGA FINAL: **06/05 (terça)** **DEADLINE** **Até às 23h59**	No Blackboard (pelo menos um integrante do grupo): ✔ **Análise exploratória dos dados** pronta (conteúdo visto no início do semestre) (Leitura de **todo enunciado do Projeto 2**).

Insper

Ciência dos dados - ENG

2024 – 2º semestre

Professora: Maria Kelly Venezuela
Professor: Marcio Fernando Stabile Junior

Rubrica

Conceito	Itens necessários:
I – Insatisfatório	✔ O projeto não apresenta um tema definido ou relevante. ✔ A base de dados utilizada é inadequada e não cumpre os requisitos mínimos de variáveis explicativas. ✔ Não há evidência de uso de ferramentas estatísticas; os gráficos e tabelas são irrelevantes ou ausentes. ✔ Interpretações ausentes ou completamente desconectadas do objetivo. ✔ Conclusões são vagas ou inexistentes.
D Desenvolvi-mento	✔ O tema é muito geral ou pouco focado, dificultando a análise. ✔ A base de dados contém as variáveis exigidas, mas algumas podem ser irrelevantes ou redundantes. ✔ Uso limitado de ferramentas estatísticas que não suportam adequadamente o objetivo do projeto. ✔ Interpretações superficiais das tabelas e gráficos, sem embasamento claro. ✔ Conclusão fraca ou que não sintetiza os principais achados do projeto.
C Suficiente	✔ O projeto tem um tema específico e pertinente. ✔ A base de dados está estruturada com um número restrito de variáveis explicativas, mas pode haver falta de clareza em algumas delas, além da variável *target*. ✔ Ferramentas estatísticas apropriadas foram utilizadas, mas de forma básica. ✔ As tabelas e gráficos são interpretados, mas as análises carecem de profundidade. ✔ Conclusão fraca ou que não sintetiza os principais achados do projeto.
C+	✔ O projeto tem um tema específico e pertinente. ✔ A base de dados está estruturada com um número restrito de variáveis explicativas, mas pode haver falta de clareza em algumas delas, além da variável *target*. ✔ Ferramentas estatísticas apropriadas foram utilizadas, mas de forma básica. ✔ As tabelas e gráficos são interpretados, mas as análises carecem de profundidade. ✔ Conclusão apresenta uma síntese dos achados, mas com limitações em termos de tomada de decisão. **Este item é obrigatorio para conceito C+.**
B Proficiente	✔ O projeto apresenta um tema bem definido e relevante ao contexto da análise. ✔ A estrutura da base de dados é clara, com um bom número de variáveis explicativas adequadas e bem identificadas, além da variável *target*. ✔ Uso adequado de ferramentas estatísticas, mas restritas as vistas na disciplina. ✔ Interpretações das tabelas e gráficos são claras e conectadas ao objetivo do projeto. ✔ Conclusão apresenta uma síntese dos achados, mas com limitações em termos de tomada de decisão.

Insper

Ciência dos dados - ENG

2024 – 2º semestre

Professora: Maria Kelly Venezuela
Professor: Marcio Fernando Stabile Junior

Conceito	Itens necessários:
B +	✔ O projeto apresenta um tema bem definido e relevante ao contexto da análise. ✔ A estrutura da base de dados é clara, com um bom número de variáveis explicativas adequadas e bem identificadas, além da variável *target*. ✔ Uso adequado de ferramentas estatísticas, restritas as vistas na disciplina e com poucas análises mais complexas. ✔ Interpretações das tabelas e gráficos são claras e conectadas ao objetivo do projeto. ✔ A conclusão é bem estruturada e revela pontos importantes a partir dos dados analisados.
A **Avançado**	✔ O projeto demonstra um tema excepcionalmente relevante e inovador. ✔ A base de dados é exemplar, com variáveis explicativas bem contextualizadas e claramente diferenciadas, além da variável *target*. ✔ Excelência no uso de ferramentas estatísticas avançadas (além do uso da biblioteca Matplotlib, por exemplo), com análises detalhadas e complexas que sustentam o objetivo. ✔ Todas as tabelas e gráficos são interpretados de forma incisiva, apresentando clareza absoluta e conexão forte com o objetivo do projeto. ✔ A conclusão é robusta, revela pontos substanciais e recomendações práticas baseadas nos dados analisados. ✔ O Jupyter Notebook deve apresentar boas construções de dashboards, integrando vários gráficos em uma única interface para facilitar a interpretação do conjunto de dados. **Este item é obrigatorio para conceito A.**
A +	✔ O projeto demonstra um tema excepcionalmente relevante e inovador. ✔ A base de dados é exemplar, com variáveis explicativas bem contextualizadas e claramente diferenciadas, além da variável *target*. ✔ Excelência no uso de ferramentas estatísticas avançadas (além do uso da biblioteca Matplotlib, por exemplo), com análises detalhadas e complexas que sustentam o objetivo. ✔ Todas as tabelas e gráficos são interpretados de forma incisiva, apresentando clareza absoluta e conexão forte com o objetivo do projeto. ✔ A conclusão é robusta, revela pontos substanciais e recomendações práticas baseadas nos dados analisados. ✔ O projeto deve estar elaborado de forma limpa, evitando excessos de códigos repetitivos, fazendo bons usos de funções. É essencial que o Jupyter Notebook apresente boas construções de dashboards, integrando vários gráficos em uma única interface para facilitar a interpretação do conjunto de dados. Todas as tabelas e gráficos gerados devem ser claramente interpretados, com *layout* bem organizado, proporcionando uma narrativa coesa e compreensível. **Este item é obrigatorio para conceito A+.**