Processo Seletivo 2024.2 - UFRJ Analytica

Equipes de Desenvolvimento e Competição - Tarefa de Análise de Dados - A2

1 Tarefa

A **segunda tarefa** para os participantes do **Grupo A2** da terceira fase do Processo Seletivo para a equipe de Desenvolvimento da UFRJ Analytica consistirá em um projeto de análise a partir de dados relativos a diferentes temas de importância socioeconômica: Educação, Saúde, Segurança Pública, Infraestrutura, Saneamento Básico, entre outros.

Para essa análise, o tema deverá ser relacionado a **"Desigualdades no Brasil"**. Dentro desse amplo tema, o(a) candidato(a) deverá definir um foco e hipóteses de pesquisa, escolher as bases de dados adequadas dentre as disponibilizadas, tratar os dados e escrever um relatório sumarizando as principais descobertas e conclusões.

Nesse texto, ele deverá indicar também quais estratégias/ferramentas foram utilizadas no tratamento de dados. Sugerimos desenvolver um esquema de questionamentos a serem respondidos. Para isso, determine uma questão central e uma sequência de sub-perguntas a serem respondidas para a análise. Isso facilita a manipulação dos dados e evita enviesamentos.

O prosseguimento dos projetos dos(as) candidatos(as) também poderá ser acompanhado por meio de reuniões individuais com os coordenadores pelo Discord, a serem marcadas em caso de necessidade. Essas reuniões poderão ser utilizadas para tirar eventuais dúvidas dos candidatos sobre seus projetos e/ou pedir encaminhamentos de focos de análise, ferramentas ou estudos a serem considerados.

Como referência, recomendamos a leitura do relatório realizado pela equipe no âmbito do Datathon Open Data Day da Base dos Dados, disponível no seguinte link:

https://medium.com/ufrj-analytica/datathon-open-data-day-base-dos-dados

2 Entregas

Para ambos os grupos, os **entregáveis** serão:

1. Documento PDF com o relatório da análise produzida.
2. Repositório no GitHub com scripts de coleta, tratamento e avaliação dos dados.

O documento PDF e o link para o seu repositório do GitHub devem ser entregue até o dia **01/10/2024**, **às 23h59**, via e-mail oficial da equipe (*analytica@labnet.nce.ufrj.br*).

3 Propostas de Trilha

Recomendamos o uso do datalake da Base dos Dados por ser de fácil acesso, contando com uma API dedicada, mas qualquer conjunto pode ser usado, desde que

documentado no relatório. Recomendamos fortemente o uso agregado de mais de uma base para a análise.

Para as trilhas, também damos liberdade escolher livremente um tema e bases de dados do seu interesse, mas para facilitar a ideação sugerimos algumas ideias de trilha:

3.1 Análise de desigualdade na educação brasileira

Para esta trilha o foco será pensar na desigualdade na educação do nosso país. Este tipo de análise é de extrema importância e nos ajuda a melhor visualizar a diferença de desempenho entre os diferentes grupos sociais, regiões e genêros e assim entender melhor o problema.

Bases Sugeridas:

- `https://basedosdados.org/dataset/indicador-de-diferenca-entre-os-desempenhos-ob`
- `https://basedosdados.org/dataset/sinopses-estatisticas-do-exame-nacional-do-ens`
- `https://basedosdados.org/dataset/br-inep-indicador-nivel-socioeconomico`
- `https://basedosdados.org/dataset/br-inep-ideb`
- `https://basedosdados.org/dataset/br-inep-censo-escolar`

3.2 Análise de indicadores econômicos, educacionais e sociais

O objetivo desta trilha é entender como os diferentes indicadores econômicos, educacionais e sociais se relacionam entre si, e entender como determinada tendência econômica de determinada região influencia indicadores educacionais. Assim como na trilha anterior, fica a cargo do candidato se ele deseja se ater a um contexto específico (determinado município, estado, etc.) ou a todo o país.

Bases Sugeridas:

- `https://basedosdados.org/dataset/br-inep-indicadores-educacionais`
- `https://basedosdados.org/dataset/br-rj-rio-de-janeiro-ipp-ips`
- `https://basedosdados.org/dataset/mundo-onu-adh`
- `https://basedosdados.org/dataset/br-ibge-pib`

3.3 Análise de investimentos em transporte/infraestrutura e seus impactos

Os que optarem por esta trilha devem realizar uma análise acerca de investimentos em transporte e infraestrutura em cidades e países, e avaliar os resultados/impactos desses investimentos. Assim como em outras trilhas é opcional o contexto ao qual a análise será aplicada, se ela contemplará o Brasil inteiro ou somente municípios.

Bases Sugeridas:

- `https://basedosdados.org/dataset/br-ana-atlas-esgotos`
- `https://basedosdados.org/dataset/br-mobilidados-indicadores`

- https://basedosdados.org/dataset/br-denatran-frota
- https://basedosdados.org/dataset/br-ipea-acesso-oportunidades

3.4 Análise de inclusão e diversidade em ambientes profissionais/acadêmicos

Para esta trilha o foco será pensar na inclusão e diversidade, ou falta deles, nos ambientes profissionais e acadêmicos. Esse tipo de análise é importante para perceber a falta de representatividade nos diferentes ambientes profissionais e entender como solucionar tal problemática.

Bases Sugeridas:

- https://basedosdados.org/dataset/painel-dinamico-da-fiscalizacao-do-abastecimen
- https://basedosdados.org/dataset/eu-fra-lgbt
- https://basedosdados.org/dataset/the-geographic-diversity-project

3.5 Tema Livre

Os candidatos também podem optar por fazer uma análise abordando outro tema que envolva **"Desigualdades no Brasil"**. Para isso nós também recomendamos o uso de outras possíveis bases de dados. É possível que os candidatos também usem qualquer outra base a escolha deles.

Notamos que a temática dos dados escolhidos é **LIVRE** ao(à) candidato(a) dentre as bases disponíveis, não havendo preferência da comissão de processo seletivo por um tema específico. Além disso, não é necessário utilizar todas as bases, mas apenas aquelas que forem pertinentes ao(s) contexto(s) de análise determinados pelo(a) participante.

No entanto, exigimos que ao menos UM dos conjuntos de dados utilizados ao longo do relatório seja proveniente do repositório Base dos Dados Mais (BD+). https://basedosdados.org/

Outras Origens Relevantes:

ISP/RJ - Dados de Segurança Pública do Rio de Janeiro (*ispdados.rj.gov.br*)

- http://www.ispdados.rj.gov.br/estatistica.html

MEC - Dados do PROUNI (*brasil.io*)

- https://brasil.io/dataset/cursos-prouni/cursos/
- https://dados.gov.br/dados/conjuntos-dados/mec-prouni

Câmara dos Deputados do Brasi (*brasil.io*)

- https://brasil.io/dataset/gastos-deputados/cota_parlamentar/

Ministério da Saúde - Preço de Medicamentos no Brasil (*dados.gov.br*)

- https://dados.gov.br/dados/conjuntos-dados/preco-de-medicamentos-no-brasil-cons

INEP - Dados dos inscritos no ENEM (*dados.gov.br*)

- https://dados.gov.br/dados/conjuntos-dados/inep-microdados-do-enem

4 Quais ferramentas podem ser utilizadas?

Para o grupo A2, as linguagens **Python**, **R** e **SQL** podem ser utilizadas. Caso o(a) candidato(a) deseje utilizar outras linguagens ou softwares, deverá entrar em contato com a coordenação para que o pedido seja analisado.

5 Como realizar minha análise?

Deixamos aqui uma sugestão de "caminho" que você pode tomar para realizar sua análise de forma efetiva. Lembrando que essa ordem de ações é apenas uma sugestão, podendo ser adaptada conforme o seu desejo e/ou forma de trabalho.

1. Aproximação inicial aos conjuntos de dados, de forma geral, objetivando entender quais informações estão presentes em cada tabela.

2. A partir do primeiro momento, pensar em um problema geral de pesquisa e estabelecer algumas hipóteses. Por exemplo, para um problema geral relacionado a "Desigualdade e Desastres Naturais", poderia se pensar em perguntas como "Quais cidades são mais e menos afetadas?", "Alguma região recebeu mais dinheiro do que outra para prevenção de desastres?", entre outros. Quanto mais perguntas e possibilidades de análise, melhor.

3. Selecionar, dentre os conjuntos de dados fornecidos, aquele(s) que serão mais pertinente(s) para responder sua pergunta de pesquisa.

4. Analisar de forma mais específica os conjuntos de dados escolhidos. Quais colunas vão ser usadas para responder às suas hipóteses? Quais podem ser ignoradas e/ou retiradas? É interessante criar novas colunas sumarizando dados? Há dados faltantes? Algum dado surpreendente?

5. Realizar o tratamento de dados. Isso pode incluir, por exemplo, a deleção ou adição de linhas, colunas ou elementos, o cálculo de estatísticas, máximos e mínimos, regressões, distribuições... O céu (e a criatividade) é o limite!
   Atenção: não esqueça de registrar quais tratamentos estão sendo feitos nos dados, para que possam ser explicados posteriormente no relatório.

6. É hora de tornar os dados visualizáveis: construa gráficos, tabelas e outras formas de representação que permitam que seus dados possam ser compreendidos pelos futuros leitores de sua análise!

7. Analise os resultados obtidos. Como você interpreta os dados? Eles fazem sentido de acordo com suas hipóteses? Você tem dados suficientes ou ao menos indícios para confirmá-la ou refutá-la?

8. Enfim, chegou a hora de escrever o relatório. Não esqueça de descrever bem sua pergunta de pesquisa, suas hipóteses, os procedimentos realizados e resultados encontrados.

Por fim: é claro que esperamos uma EXCELENTE análise de vocês, mas fiquem tranquilos que não estamos exigindo nenhuma tese de Doutorado. Então não é pra se desesperar, hein? Pega um tema que você se interessa, se joga nos dados e tudo vai rolar no final das contas!

6 Como registrar meu código?

O código deve ser versionado por meio de seu GitHub pessoal, de forma que possa ser acessado posteriormente pela coordenação do Processo Seletivo. **Dessa forma, pedimos ao(à) candidato(a) que coloque o link para seu GitHub ao final do relatório.**

7 Tenho uma dúvida sobre o projeto. Como fazer?

Pedimos que quaisquer dúvidas gerais sejam tiradas pelo Discord, pois elas podem ser comuns a outros(as) candidatos(as). Além disso, teremos os espaços das reuniões gerais e das reuniões complementares para esclarecer outras questões. Em último caso, também pode mandar um e-mail que a gente responde :)