Instituto Federal do Maranhão (IFMA)

Curso: Bacharelado em Sistemas de Informação

Disciplina: Tópicos Avançados em Visualização de Dados

Professor: Josenildo Costa Da Silva

Equipe: Edilson Marques (20221SI0033), Gabriel Vinícius Silveira (20221SI0006), Jorge Lucas (20221SI0004) e Reinaldo Dias (20221SI0029)

Relatório de Análise Exploratória - Pesquisa Data Hackers 2024/2025

Dataset utilizado: [State of Data Brazil 2024/2025 – Data Hackers](https://www.kaggle.com/datasets/datahackers/state-of-data-brazil-20242025/data)

1. Introdução

O presente relatório apresenta os principais resultados da análise exploratória aplicada à base de dados da pesquisa **Data Hackers 2024/2025**. O objetivo da atividade foi compreender melhor o perfil dos profissionais da área de dados no Brasil, identificando padrões, tendências e eventuais desigualdades.

O dataset foi previamente tratado e se encontra com colunas padronizadas, valores inconsistentes removidos e variáveis adequadamente transformadas. A análise foi conduzida com o uso de bibliotecas como **Pandas**, **Matplotlib**, **Seaborn** e **Scipy**.

2. Etapas de Preparação (Resumo)

O dataset original possuía centenas de colunas, sendo muitas derivadas de perguntas de múltipla escolha convertidas por one-hot encoding. A equipe fez uma partição temática dos dados em grupos funcionais, facilitando a visualização e manipulação.

Utilizou-se o método `info()` para entender o tipo de dados e a existência de valores nulos. Colunas irrelevantes, duplicadas ou com baixa variabilidade foram descartadas. Além disso, criaram-se cópias dos subconjuntos com `df.copy()` para preservar o dataset original.

- **Leitura dos dados:** A base foi carregada a partir de um arquivo `.csv` previamente tratado.

- **Tratamento prévio:** Foram removidos valores nulos, padronizados os nomes das colunas e convertidos os tipos das variáveis categóricas.

- **Bibliotecas:** Utilizou-se `pandas` para manipulação dos dados, `seaborn` e `matplotlib` para visualização e `scipy.stats` para testes estatísticos.

Foi aplicada a verificação de nulos com `isna().sum()`. As colunas com grande proporção de nulos foram eliminadas. Em outras, os valores faltantes foram tratados com imputação por valores padrão ou mais frequentes.

A distribuição de atributos numéricos como **idade** foi verificada por meio de boxplots e histogramas. Utilizou-se `skew()` e `kurtosis()` para verificar assimetria e curtose. Detectaram-se valores extremos (e.g., idades implausíveis) que foram filtrados ou ajustados.

Variáveis qualitativas (ex: modelo de trabalho ideal) foram transformadas com codificação numérica para análises de correlação. Foram criadas tabelas cruzadas com `pd.crosstab` e aplicadas transformações com `melt` ou `pivot`. Garantiu-se que colunas numéricas representassem grandezas reais e não códigos.

3. Principais Achados da Análise

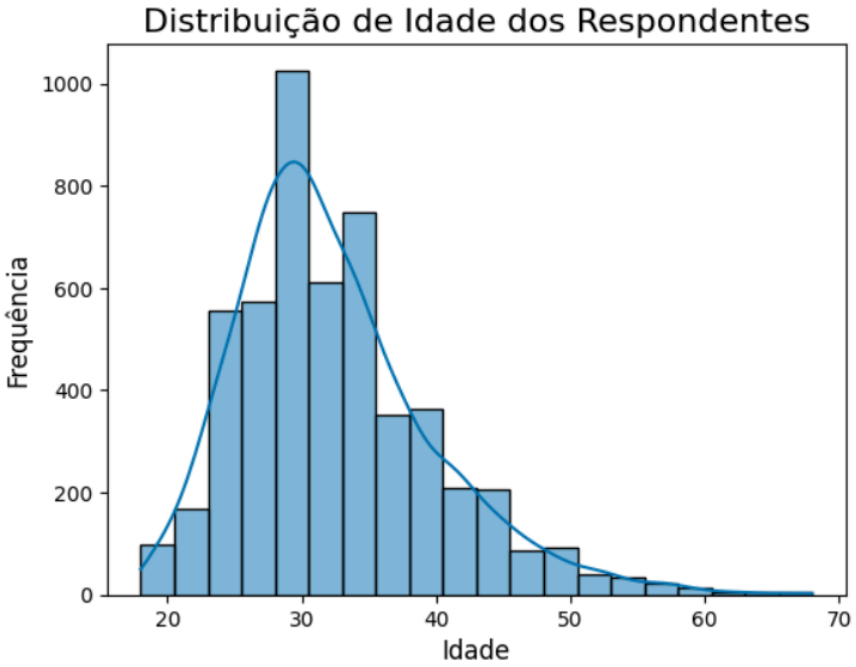
No que concerne à análise univariada, identificou-se que a maioria dos respondentes está entre 20 e 35 anos. Há predominância de pessoas com pós-graduação. A faixa salarial mais comum está entre R$ 4.000 e R$ 8.000 mensais. São Paulo se destaca como a UF com maior número de respondentes.

Para a análise bivariada, foi elaborado Heatmap com Correlação positiva (ainda que fraca) entre idade e preferência por modelo remoto de trabalho e avaliada a concentração de profissionais com pós-graduação nas faixas salariais mais altas. Outros cruzamentos mostraram como variáveis

como nível profissional, escolaridade e localização afetam a remuneração e preferências de trabalho.

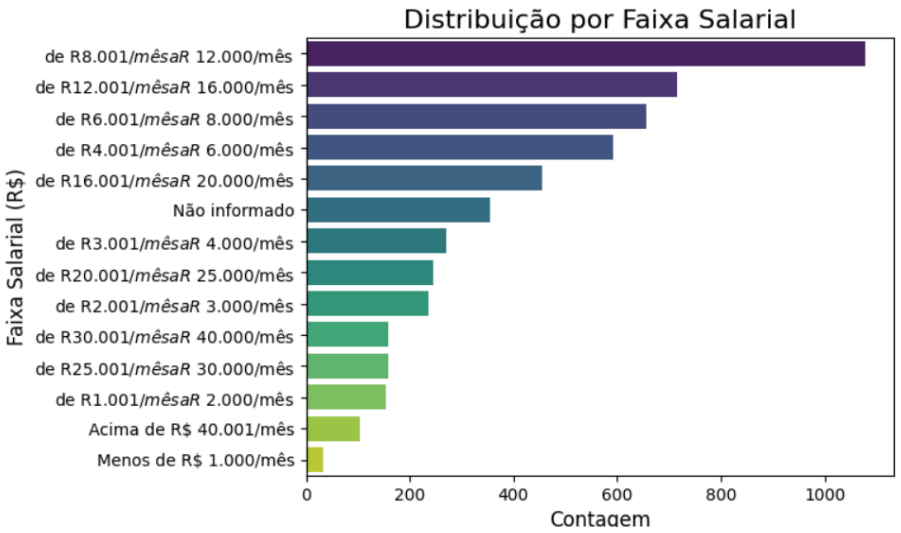
a) Perfil Etário

A distribuição da idade dos respondentes mostra uma **concentração expressiva entre 25 e 35 anos**, evidenciando que a maioria dos profissionais de dados ainda está em início ou meio de carreira. Este padrão indica um setor jovem e em ascensão.



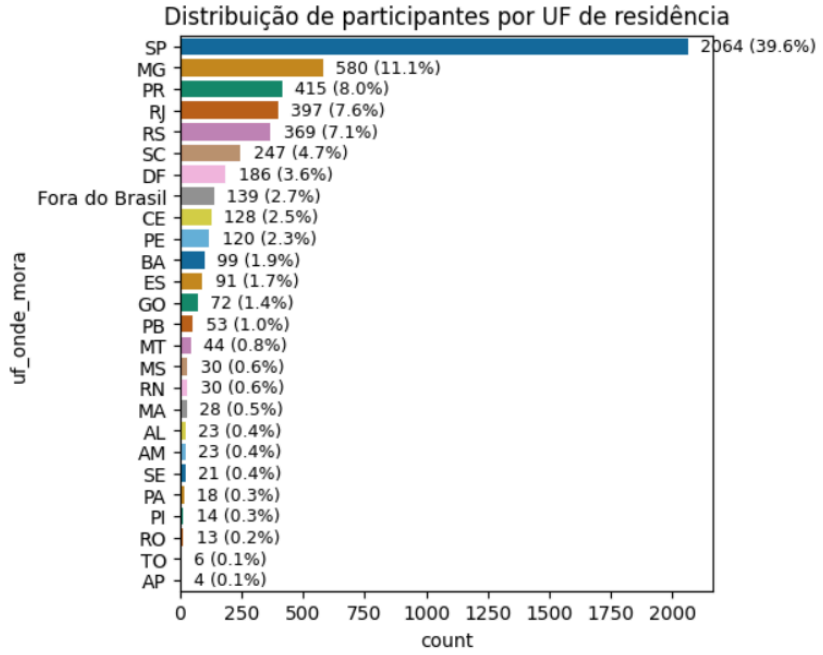
b) Faixa Salarial

A maior parte dos profissionais se concentra nas faixas **de R$ 4.000 a R$ 8.000 mensais**, com uma leve predominância na faixa intermediária. Os salários são significativamente influenciados por fatores como **nível de experiência** e **região de atuação**.



c) Localidade

A distribuição geográfica revela uma maior concentração de profissionais nos estados do **Sudeste**, especialmente **São Paulo**, o que é consistente com a concentração de empresas de tecnologia e centros urbanos na região.



d) Testes de Normalidade

Foram aplicados os testes **Anderson-Darling** e **D'Agostino K²** para avaliar a normalidade da variável idade. Ambos os testes rejeitaram a hipótese de normalidade, indicando que a distribuição da idade é assimétrica (possivelmente enviesada à esquerda).

4. Conclusão

A análise revelou que o mercado de dados mostra-se inclusivo em relação à formação, mas ainda privilegia quem possui pós-graduação. São Paulo concentra as melhores oportunidades em termos de salário e vagas. A valorização do modelo remoto é sutilmente maior entre os profissionais mais velhos.

Com o crescimento contínuo da área de dados, é interessante compreender e agir sobre esses padrões para promover um mercado mais equilibrado e acessível tanto para empresas e instituições como para os profissionais da área.