

# Data Science Job Salaries 2024

Gabriel Santos de Oliveira Arruda<sup>1</sup>, Paulo

Henrique Souza Camargo<sup>2</sup>

FCI – Universidade Presbiteriana Mackenzie

Ciências da Computação - Inteligência Artificial  
Universidade Presbiteriana Mackenzie – São Paulo, SP – Brasil

{10388025, 10389453}@mackenzista.com.br

**Resumo.** Este artigo apresenta uma análise do dataset "Data Science Job Salaries 2024", que inclui dados sobre ano, nível de experiência, título, salário, moeda, salário em dólar, localização, porcentagem de trabalho remoto e tamanho da empresa. O foco é prever o salário na área da TI nos EUA em empregos de período integral.

## 1. Introdução

A área de Data Science tem testemunhado um crescimento significativo nos últimos anos, com a demanda por profissionais qualificados em constante ascensão. Nesse contexto, compreender os padrões salariais e fatores determinantes torna-se crucial para candidatos e empresas. Este estudo se propõe a analisar o conjunto de dados "Data Science Job Salaries 2024", focalizando-se em registros nos Estados Unidos e contratos de tempo integral. A predição do nível de senioridade e localização é o cerne deste projeto, visando fornecer insights valiosos para profissionais e organizações do setor.

### 1.1. Contextualização

Com o avanço da tecnologia e a crescente importância dos dados para tomada de decisões estratégicas, a área de Data Science tem se destacado como uma das mais promissoras no mercado de trabalho. A demanda por profissionais capacitados em ciência de dados continua a aumentar, impulsionada pela necessidade de interpretar e extrair valor de grandes volumes de dados.

### 1.2. Justificativa

Diante desse cenário dinâmico e competitivo, compreender os padrões salariais e os fatores que influenciam nas remunerações torna-se essencial para profissionais que buscam ingressar ou progredir na carreira de Data Science. Além disso, empresas que buscam atrair e reter talentos precisam entender as expectativas salariais dos candidatos e as tendências do mercado para permanecerem competitivas.

### 1.3. Objetivo

O principal objetivo deste estudo é realizar uma análise preditiva dos níveis de senioridade e localização para cargos de Data Science nos Estados Unidos, considerando apenas contratos em tempo integral. Por meio da aplicação de técnicas de machine learning, pretendemos fornecer insights valiosos para profissionais em busca de oportunidades de emprego e para empresas que desejam tomar decisões informadas sobre contratações e estratégias de remuneração.

## 1.4. Opção do Projeto

A opção por focar exclusivamente em registros nos Estados Unidos e em contratos de tempo integral foi feita para delimitar o escopo do projeto e garantir a relevância dos resultados para o contexto específico do mercado de trabalho nesse país e para o perfil de profissionais interessados em oportunidades de emprego em tempo integral na área de Data Science.

## 2. Descrição do problema

O cenário atual do mercado de trabalho em tecnologia, especialmente no campo da ciência de dados, é marcado por uma alta demanda por talentos qualificados. No entanto, tanto candidatos quanto empresas enfrentam desafios ao tentar encontrar o melhor ajuste entre habilidades, expectativas salariais e localização geográfica. Essas questões são complexas, pois estão sujeitas a uma variedade de fatores, como experiência profissional, título do cargo, tamanho da empresa e localização. O conjunto de dados "Data Science Job Salaries 2024" oferece uma oportunidade única para abordar essas preocupações, fornecendo insights detalhados sobre os padrões salariais e a distribuição geográfica das oportunidades de emprego nos Estados Unidos. Ao realizar análises preditivas sobre os níveis de senioridade e localização, este projeto visa fornecer informações valiosas para profissionais em busca de oportunidades de carreira e empresas procurando atrair e reter talentos no competitivo mercado de tecnologia dos EUA.

## 3. Dataset

Utilizamos o dataset disponibilizado por [Shaw 2024], como fonte primária de dados para nossa análise. A análise exploratória dos dados do dataset revelou insights importantes sobre a natureza e a qualidade dos dados disponíveis. Inicialmente, ao examinar a distribuição dos atributos, observou-se que alguns atributos continham uma quantidade significativa de dados ausentes ou inconsistentes. Isso exigiu a aplicação de técnicas de limpeza de dados, como remoção de registros duplicados, tratamento de valores nulos e correção de inconsistências nos dados.

Além disso, ao explorar a distribuição dos salários, ficou evidente que os dados estavam dispersos de forma irregular, o que dificultava a identificação de padrões ou tendências claras. A falta de uma distribuição clara dos salários também dificultou a identificação de potenciais outliers ou valores discrepantes que poderiam distorcer as análises posteriores.

Outra observação importante foi a análise da correlação entre os atributos preditores e o atributo alvo (salário). Descobriu-se que muitos dos atributos preditores apresentavam correlações fracas ou insignificantes com o salário, o que sugeria que eles poderiam ter pouco poder preditivo em relação aos salários dos funcionários. Essa falta de correlação representou um desafio adicional na construção de um modelo preciso de previsão de salários.

Em resumo, a análise exploratória dos dados revelou várias questões importantes relacionadas à qualidade, distribuição e correlação dos dados. Essas descobertas destacaram a necessidade de um cuidadoso pré-processamento dos dados e uma consideração cuidadosa na seleção de atributos para a construção de modelos de previsão de salários mais precisos e confiáveis.

## 4. Metodologia

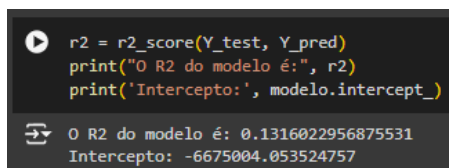
A metodologia aplicada no projeto envolve uma série de etapas detalhadas para explorar, preparar e analisar os dados de salários. Inicialmente, foram importadas as bibliotecas necessárias para a análise de dados, visualização e manipulação, como pandas, numpy, seaborn, matplotlib. Em seguida, os dados foram carregados a partir de um arquivo CSV e inspecionados para verificar a presença de valores nulos e duplicados. As variáveis categóricas (experience level) e (employment type) foram transformadas em valores numéricos para facilitar a análise. Foi realizada uma análise descritiva dos dados, verificando estatísticas básicas, valores nulos e registros duplicados. Para a limpeza dos dados, registros duplicados e nulos foram removidos. Na análise exploratória, boxplots foram criados para visualizar a distribuição dos dados e, em seguida, os dados foram filtrados para incluir apenas registros dos Estados Unidos, com emprego em tempo integral e excluindo salários outliers acima de 300.000. Colunas irrelevantes foram removidas, e um heatmap foi criado para visualizar as correlações entre as variáveis. A distribuição dos dados foi analisada com histogramas e a assimetria foi calculada. Após a preparação dos dados, o dataset tratado foi salvo e carregado novamente para a etapa de modelagem preditiva.

Na modelagem preditiva, começamos dividindo os dados em conjuntos de treino. Com isso, instanciamos um modelo de regressão linear e o treinamos com os dados de treino. Em seguida, usamos o modelo treinado para fazer previsões sobre o conjunto de teste e comparamos esses valores previstos com os valores reais através de um gráfico de dispersão. Para avaliar a performance do modelo, calculamos o coeficiente de determinação ( $R^2$ ), que mede o quanto o modelo consegue explicar a variabilidade dos dados. Além disso, exibimos o intercepto do modelo. Resumindo, a metodologia aplicada envolveu uma preparação cuidadosa dos dados, desde a limpeza e transformação das variáveis até a análise exploratória com visualizações, e finalmente a criação e avaliação de um modelo de regressão linear para prever os salários.

## 5. Resultados

### 5.1. Coeficiente de determinação - $R^2$

Os resultados obtidos a partir do modelo de regressão linear foram analisados detalhadamente. Após a divisão dos dados em conjuntos de treino e teste (75% para treino e 25% para teste), o modelo foi treinado e as previsões foram realizadas para o conjunto de teste. A precisão do modelo foi avaliada pelo coeficiente de determinação ( $R^2$ ), que foi calculado em 0.1316. Este valor indica que aproximadamente 13,16% da variância nos salários pode ser explicada pelas variáveis independentes incluídas no modelo. Embora este valor de  $R^2$  seja relativamente baixo, sugerindo uma capacidade limitada do modelo em prever os salários com precisão, ele fornece uma base inicial para melhorias futuras.



```
r2 = r2_score(Y_test, Y_pred)
print("O R2 do modelo é:", r2)
print('Intercepto:', modelo.intercept_)

O R2 do modelo é: 0.1316022956875531
Intercepto: -6675004.053524757
```

Figura 1. Coeficiente de determinação

## 5.2. Gráfico de dispersão

O gráfico de dispersão gerado comparando os valores observados com os valores previstos pelo modelo mostra uma dispersão significativa, indicando que o modelo não está prevendo com alta precisão. Os pontos no gráfico representam a relação entre os salários reais e os salários previstos, com a linha preta tracejada indicando onde os valores observados seriam iguais aos valores previstos. A dispersão dos pontos ao longo desta linha sugere que as previsões do modelo têm um nível considerável de erro.

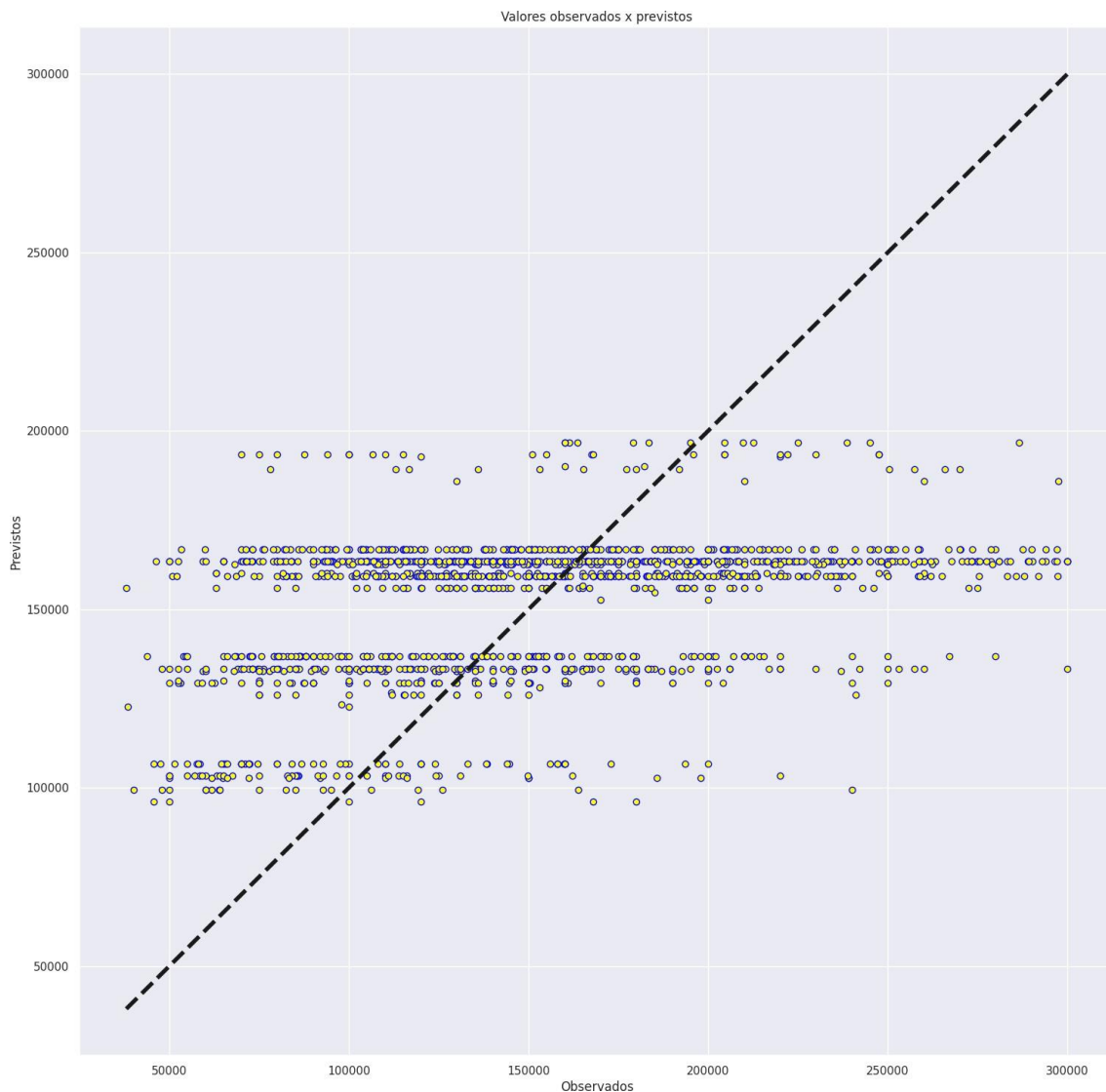
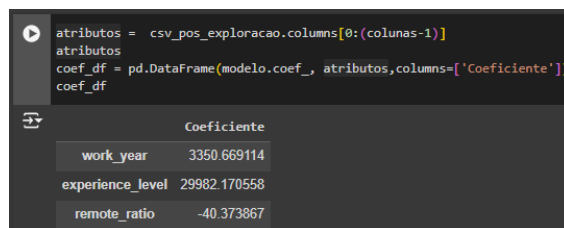


Figura 2. Coeficiente de determinação

## 5.3. Coeficientes

Além disso, os coeficientes das variáveis independentes no modelo foram analisados. O intercepto do modelo foi de -6675004.05, enquanto os coeficientes das variáveis foram os seguintes: para *work\_year* foi 3350.67, para *experience\_level* foi 29982.17, e para *remote\_ratio* foi -40.37. Esses coeficientes indicam que, para cada unidade de aumento em *work\_year*, o salário aumenta em média \$3350.67, para cada nível de experiência

adicional, o salário aumenta em média \$29982.17, e para cada unidade de aumento na proporção de trabalho remoto, o salário diminui em média \$40.37.



**Figura 3. Coeficientes**

## 6. Conclusão

O modelo gerado não se mostrou adequado para uso, uma vez que não aprendeu o suficiente para prever os salários de maneira precisa. Vários fatores podem ter contribuído para esse resultado insatisfatório. Primeiramente, a quantidade limitada de atributos preditores pode ter restringido a capacidade do modelo de capturar as variáveis relevantes. Além disso, a dispersão excessiva dos dados sugere uma grande variabilidade que dificulta a identificação de padrões claros. Por fim, a ausência de correlação significativa entre os atributos utilizados e os salários prejudica a eficácia das previsões, evidenciando a necessidade de uma seleção mais criteriosa dos atributos a serem incluídos no modelo.

## 7. Endereço GitHub

<https://github.com/gabriel07099/ProjetoIA>

## Referências

- Bilal, M. (2024). Data scientist salary: Ultimate 2024 guide. In Bilal, M., editor, <https://www.iu.org/blog/salaries/data-scientist-salary-expectations/>.
- Rehman, F. (2024). Data science job salary prediction (glassdoor). In Rehman, F., editor, <https://www.kaggle.com/code/fahadrehman07/data-science-job-salary-prediction-glassdoor>.
- Shaw, A. (2024a). Data science job salaries 2024. In Shaw, A., editor, <https://www.kaggle.com/datasets/abhinavshaw09/data-science-job-salaries-2024>.
- Shaw, A. (2024b). Data science job salaries eda. In Shaw, A., editor, <https://365datascience.com/career-advice/data-science-salaries-around-the-world/>.
- Yosifova, A. (2024). Data science salaries around the world in 2024. In Yosifova, A., editor, <https://365datascience.com/career-advice/data-science-salaries-around-the-world/>.