ANÁLISE DE CORRELAÇÕES NO DATASET "STACK OVERFLOW 2018 DEVELOPER SURVEY": UMA ABORDAGEM COM PYTHON

Davi de França Vasconcelos Nunes[1]
Henrique Gabriel Gasparelo[2]
Isaías Gouvêa Gonçalves[3]
José Thevez Gomes Guedes[4]
Rafael de Pinho André[5]

RESUMO

Este paper visa analisar as relações entre os dados do conjunto intitulado "*Stack Overflow 2018 Developer Survey*". O objetivo principal é identificar correlações entre diferentes colunas do dataset, proporcionando uma base para análises mais aprofundadas. Para alcançar esse objetivo, utilizou-se *Python*, empregando as bibliotecas *Pandas*, *Numpy* e *Seaborn*, que permitem a manipulação eficaz dos dados e a criação de visualizações informativas, possibilitando uma análise visual das relações identificadas. Com este objetivo, foram elaboradas quatro hipóteses, que são construídas e analisadas a fim de evidenciar possíveis relações entre as variáveis do dataset selecionado. Como resultado, obtêm-se análises que oferecem insights sobre padrões e tendências, além de sugerir ideias para pesquisas futuras e estudos mais focados em algumas relações expostas.
Palavras-chave: Análise de dados; Desenvolvedores; Stack Overflow; Visualização de dados; Python.

1 INTRODUÇÃO

No âmbito da programação, o desenvolvimento de software e a experiência dos desenvolvedores são frequentemente associados e correlacionados por uma variedade de fatores, como habilidades técnicas, ambiente de trabalho e preferências pessoais. O "*Stack Overflow 2018 Developer Survey*"fornece uma base de dados que permite examinar essas relações. Esta pesquisa, que coletou informações de milhares de desenvolvedores em 2018, é uma valiosa fonte para investigar as correlações entre diferentes variáveis, como linguagem de programação, nível de experiência e satisfação no trabalho.

Além disso, compreender as relações entre as variáveis pode fornecer insights relevantes para empresas de tecnologia, instituições educacionais e os próprios desenvolvedores, ajudando na tomada de decisões estratégicas e na identificação de tendências na indústria. Por exemplo, será que desenvolvedores que utilizam linguagens de programação mais recentes estão associados com os salários mais altos? Ou existe uma correlação entre a experiência em desenvolvimento e a empregabilidade?

Este paper busca então realizar uma análise de associação e não uma análise de causalidade entre os fatores explicitados no dataset e está estruturado da seguinte maneira: na seção

[1] FGV - Rio de Janeiro/RJ - Ciência de Dados e Inteligência Artificial - davifvnunes1@gmail.com
[2] FGV - Rio de Janeiro/RJ - Ciência de Dados e Inteligência Artificial - henriquegasparelo@gmail.com
[3] FGV - Rio de Janeiro/RJ - Ciência de Dados e Inteligência Artificial - isaias.ggoncalves@gmail.com
[4] FGV - Rio de Janeiro/RJ - Ciência de Dados e Inteligência Artificial - josethevez@gmail.com
[5] FGV - Rio de Janeiro/RJ - rafael.pinho@fgv.br

2, é discutido a metodologia utilizada na análise, incluindo a descrição do dataset e as técnicas de manipulação de dados. Na seção 3, é apresentado os resultados obtidos com a análise de relação, seguidos por uma discussão sobre os resultados e como eles poderiam ser estudados mais a fundo. E por final, a seção 4 concluirá o paper, ressaltando as limitações do estudo, além de sugerir direções para futuras pesquisas.

2 DESENVOLVIMENTO

O dataset "*Stack Overflow 2018 Developer Survey*"disponível gratuitamente na plataforma *Kaggle*[5] que pode ser acessado pelo hiperlink (link), contém uma variedade de informações sobre desenvolvedores, incluindo, dados individuais, linguagens de programação utilizadas, ferramentas e tecnologias preferidas, além de percepções sobre o mercado de trabalho. Para a presente análise, foi selecionado um subconjunto de colunas que foram julgadas ser mais relevantes para investigar as relações.

Neste trabalho, utilizou-se o *Python* como principal ferramenta de manipulação e análise de dados, apoiando-se em bibliotecas amplamente reconhecidas, como *Pandas*[6], *Numpy*[7], *Matplotlib*[8] e *Seaborn*[9]. Utilizando a biblioteca Pandas, é realizada a limpeza e o pré-processamento dos dados, tratando valores ausentes e convertendo esses tipos de dados conforme necessário. O uso do Numpy possibilitou operações matemáticas que auxiliaram na análise exploratória dos dados, enquanto a biblioteca Seaborn é utilizada para criar visualizações que ilustram essas relações de forma clara e acessível (WASKOM, 2021).

Para a análise dos dados, foi selecionada como abordagem principal: a análise visual por meio de gráficos. A análise visual, conforme descrita por Morettin e Bussab (2017) no livro Estatística Básica, é uma poderosa ferramenta para a interpretação de dados, pois permite: “(a) identificar padrões e relações visíveis; (b) confirmar ou refutar expectativas prévias em relação aos dados; (c) descobrir novos fenômenos e tendências que poderiam passar despercebidos em análises puramente numéricas; (d) validar ou questionar as suposições feitas sobre os procedimentos estatísticos adotados; e (e) apresentar os resultados de maneira mais rápida e acessível, facilitando a compreensão dos achados.” (MORETTIN; BUSSAB, 2017, pág. 6, apud CHAMBERS et al., 1983). Essas visualizações não apenas facilitam a comunicação dos resultados, mas também ampliam a capacidade de interpretação ao permitir uma visão mais intuitiva das relações entre variáveis.

Neste estudo, foram elaboradas quatro hipóteses que serão discutidas na seção de resultados. A verificação dessas hipóteses envolveu a prévia manipulação dos dados e a geração de visualizações gráficas com o auxílio da biblioteca *Seaborn*. A partir dessas visualizações, foi possível realizar uma análise de correlação, que forneceu insights importantes sobre as relações entre as variáveis, permitindo uma interpretação mais clara e fundamentada dos dados.

---

[5] Plataforma que disponibiliza gratuitamente datasets, promove competições de ciência de dados, oferece notebooks colaborativos e fomenta uma comunidade ativa de profissionais e entusiastas de machine learning.

[6] Biblioteca Python usada para manipulação e análise de dados, oferecendo estruturas de dados como DataFrames, que facilitam o trabalho com dados tabulares.

[7] Biblioteca fundamental para computação científica em Python, oferecendo suporte a arrays multidimensionais e operações matemáticas eficientes.

[8] Biblioteca para visualização de dados em Python, que permite a criação de gráficos de forma flexível em 2d e 3d.

[9] Biblioteca de visualização de dados em Python que permite a criação de gráficos estatísticos, integrando-se de maneira eficaz com Pandas e construindo sobre a funcionalidade da Matplotlib.

3 RESULTADOS E DISCUSSÕES

Nesta seção, é apresentado as quatro hipóteses formuladas, acompanhadas de suas respectivas visualizações e análises de associação e relação.

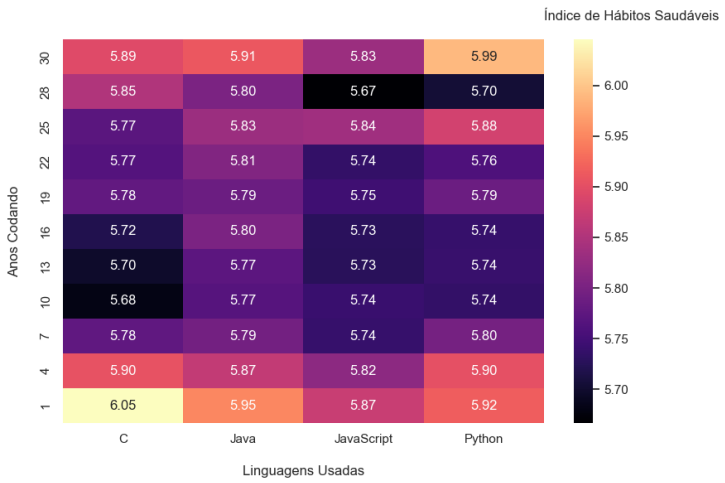
3.1 Hipótese 1: Índices De Hábitos Saudáveis Entre Programadores

A primeira hipótese proposta neste estudo examina a seguinte questão: há uma diferença nos índices de hábitos saudáveis entre programadores que utilizam linguagens de programação mais modernas (Python, JavaScript) e aqueles que utilizam linguagens tradicionais (Java, C)?

Para testar essa hipótese, foi adicionada uma nova coluna no dataset, que reflete o índice de hábitos saudáveis de cada desenvolvedor. Esse índice foi construído a partir de uma função que atribui uma pontuação baseada em variáveis como: prática de exercícios físicos, frequência com que pulam refeições, horas dedicadas a atividades ao ar livre e tempo gasto no computador. Embora o número de variáveis seja limitado, esse índice permite obter uma primeira visão sobre as tendências de hábitos saudáveis, abrindo espaço para a identificação de possíveis padrões e correlações.

Para visualizar esses dados, foi criado um gráfico de calor, onde o eixo X representa os anos de experiência, divididos em faixas, e o eixo Y as diferentes linguagens de programação. A variação de cor ilustra o índice de hábitos saudáveis, permitindo comparar como esse índice se comporta em função do tempo de experiência e da linguagem utilizada. A seguir, apresenta-se a visualização (Figura 1):

Figura 1 — Gráfico de calor da relação entre o índice de hábitos saudáveis e linguagens utilizadas.



Fonte: Autoria própria (2024).

Para a observação do gráfico de calor vale notar que a amplitude do índice de hábitos saudáveis é relativamente baixa, o que significa que as variações observadas não são grandes, mas ainda permitem identificar tendências interessantes, isso ocorre pois foi utilizado a média para chegar nos valores.

Mesmo com essa baixa amplitude, é possível perceber algumas tendências ao longo dos anos de experiência e entre linguagens. Por exemplo, na coluna referente à linguagem C, há uma queda acentuada nos índices nos primeiros anos de carreira, seguida de uma recuperação lenta ao longo do tempo. Esse comportamento pode levantar questões sobre o que acontece nos estágios iniciais da carreira dos programadores que usam C.

Outro ponto importante é a análise das linguagens mais modernas, como Python e JavaScript. Entre os 25 e 30 anos de experiência, observa-se uma variação significativa no índice de saúde, com períodos de melhora seguidos por quedas e, em seguida, recuperação. Isso é diferente das linguagens mais tradicionais como C e Java, onde as variações no índice de saúde são mais suaves.

Além disso, é interessante observar que, independentemente da linguagem, há um período comum de baixos índices de hábitos saudáveis entre 10 e 22 anos de carreira, o que poderiam estar refletindo algum fator relacionado ao tempo de experiência, ao equilíbrio entre vida pessoal e profissional ou a outras variáveis que afetam os hábitos nesse intervalo de tempo.

Na comparação geral, programadores que utilizam Java tendem a apresentar índices consistentemente mais elevados ao longo dos anos, mesmo com as diferenças de amplitude sendo pequenas. As linguagens mais modernas, por outro lado, mostram variações mais marcantes em fases avançadas da carreira, o que pode abrir caminho para estudos mais aprofundados.

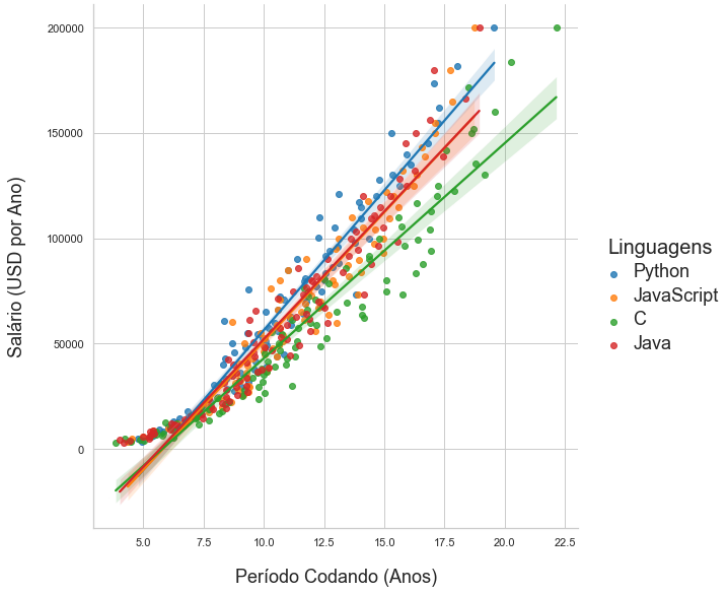
3.2 Hipótese 2: Crescimento Salarial Entre Programadores De *Python*, *JavaScript* E *Java/C*

A segunda hipótese levanta a seguinte questão: o crescimento salarial é mais acentuado para programadores que utilizam *Python* e *JavaScript* em comparação com aqueles que utilizam *Java* e *C*, ao longo dos anos de experiência (years coding)?

Para essa análise, foram utilizados os dados salariais dos programadores, em dólares americanos e se referindo ao salário anual, divididos em grupos conforme as linguagens de programação em seu repertório: aqueles que utilizam *Python*, *Java*, *C* e *JavaScript*. Foram removidos os valores salariais atípicos, para evitar distorções na visualização e nas conclusões. Dada a grande quantidade de dados, foi selecionada uma amostra representativa, na qual os dados de cada categoria foram divididos em cem quantis. Para cada quantil, foi calculada a média do tempo de experiência, que serviu como base para os pontos no gráfico de dispersão.

Com os dados tratados, foi construído um gráfico de dispersão onde o eixo X representa os anos de experiência e o eixo Y o salário, com as cores diferenciando os grupos de programadores. Adicionalmente, foi incluída uma linha de tendência para cada grupo, facilitando a comparação do crescimento salarial ao longo do tempo. A seguir, a visualização elaborada (Figura 2):

Figura 2 — Gráfico de dispersão da relação entre anos de experiência e salário em cada linguagem.



Fonte: Autoria própria (2024).

Apesar da proximidade dos dados, a linguagem *Python* (azul) parece exibir uma tendência de crescimento salarial um pouco mais acentuada, especialmente a partir dos 10 anos de experiência. Isso pode sugerir que programadores que utilizam *Python* tendem a atingir salários maiores mais rapidamente em comparação com outras linguagens, embora essa diferença seja pouca.

Por outro lado, a linha de tendência para *C* (verde) sugere um crescimento salarial mais moderado, o que poderia indicar que programadores que utilizam essa linguagem experimentam um aumento mais lento nos salários ao longo do tempo. No caso de *Java* (vermelho) e *JavaScript* (laranja), as linhas de tendência permanecem muito próximas à de *Python*, o que poderia estar indicando que programadores que utilizam essas linguagens tendem a ter um avanço salarial parecido.

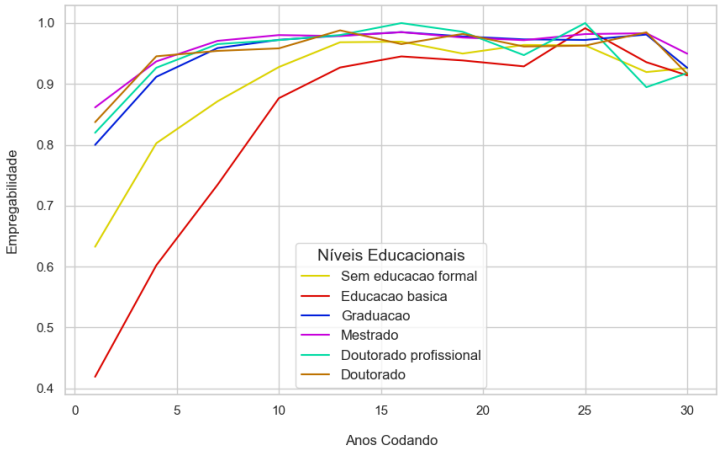
É importante ressaltar que os dados se mantêm relativamente próximos, indicando que, em geral, há progressão salarial em relação ao tempo relativamente semelhante para todos os programadores, independentemente da linguagem, nos dados selecionados. As observações levantam hipóteses sobre possíveis diferenças no ritmo de crescimento salarial para cada linguagem, mas não são conclusivas deixando espaço para futuros estudos e análises.

3.3 Hipótese 3: Grau De Formação E Empregabilidade

A terceira hipótese examina a possibilidade de que: Profissionais da área de programação com grau de formação superior encontram-se mais empregados no mercado de trabalho ao longo de suas carreiras?

Para essa análise, utilizou-se a coluna de emprego presente no dataset, onde foi definido se o programador estava empregado ou não no momento da pesquisa. A partir disso, calculou-se a proporção de programadores empregados em relação ao total, agrupando-os de acordo com seu grau de formação (sem formação, ensino médio, graduação, mestrado, doutorado, etc.). Com essas informações, foi criado um gráfico de linhas onde o eixo X representa o tempo de experiência do programador (em anos) e o eixo Y mostra a proporção de empregabilidade em cada nível de formação. As diferentes linhas no gráfico correspondem aos distintos graus de formação, permitindo a comparação do comportamento da empregabilidade ao longo da carreira para cada grupo. A seguir, é apresentada a visualização (Figura 3) desenvolvida:

Figura 3 — Gráfico de linha da relação entre anos de programação e empregabilidade em relação ao grau de formação.



Fonte: Autoria própria (2024).

A partir do gráfico construído é possível observar que no início da carreira (até aproxima-

damente os 5 primeiros anos de experiência), a empregabilidade varia significativamente entre os diferentes níveis de educação. Os profissionais com ensino superior (graduação, mestrado, doutorado) tendem a ter uma empregabilidade alta, ultrapassando 90% em grande parte dos casos. Em contraste, aqueles sem educação formal apresentam uma curva mais acentuada, começando com uma empregabilidade mais baixa e alcançando uma estabilização ao redor de 10 anos de experiência.

Já os programadores com educação básica apresentam um crescimento na empregabilidade nos primeiros 10 anos, chegando a uma estabilização ao redor de 90%, o que também é observado em níveis de formação mais altos.

Diferentemente do início da carreira onde é observado uma diferença na empregabilidade em relação ao grau de formação, entre 15 e 25 anos de carreira é observado uma tendência de que indiferentemente do grau de formação mais de 90% dos profissionais estejam empregados.

Isso levanta a questão, será que o mercado de trabalho tende a priorizar programadores com ensino superior em comparação com aqueles sem formação no início da carreira? Vale ressaltar que, embora o gráfico ofereça uma visão desta diferença, não é possível a partir dela concluir uma causalidade, deixando essa questão como sugestão para estudos futuros.

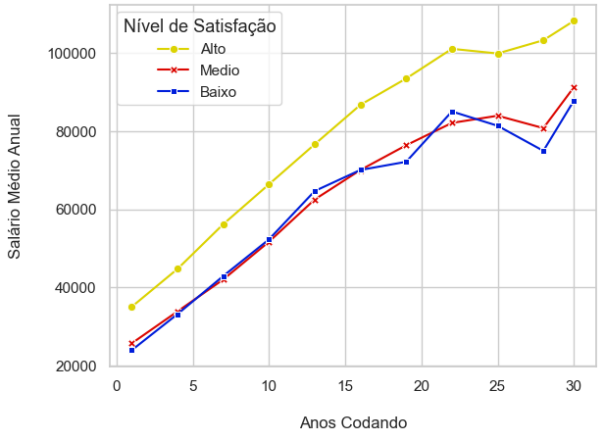
Ao observar como a empregabilidade se comporta é possível notar que, ao chegar perto dos 30 anos de carreira, há uma queda na empregabilidade para todos os graus de formação, mas em especial para os sem educação formal e com educação básica, o que poderia estar refletindo algum padrão de comportamento dos profissionais ao chegarem perto do final da carreira.

3.4 Hipótese 4: Satisfação Na Carreira, Experiência E Crescimento Salarial

A quarta hipótese propõe a investigação de uma possível relação: Programadores com maior satisfação na carreira atingem níveis salariais mais altos em um tempo menor, em comparação com programadores com níveis médios ou baixos de satisfação?

Para testar essa hipótese, foram utilizados os dados sobre satisfação na carreira e salários, organizados de acordo com os anos de experiência em programação. Os programadores foram divididos em três grupos, com base em sua satisfação profissional (alta, média e baixa). Para visualizar os resultados, foi criado um gráfico de linhas no qual o eixo X representa os anos de experiência em programação e o eixo Y representa o salário médio. As três linhas correspondem aos níveis de satisfação profissional (alta, média e baixa), permitindo uma análise comparativa de como os salários evoluem ao longo da carreira em função da satisfação. A seguir, apresenta-se a visualização (Figura 4):

Figura 4 — Gráfico de linhas empilhadas com marcadores sobre a relação entre anos de experiência e salários em relação ao nível de satisfação.



Fonte: Autoria própria (2024).

O gráfico revela que programadores com alto nível de satisfação tendem a atingir salários mais altos com menos tempo de experiência, em comparação com os grupos de satisfação média e baixa. A linha correspondente à alta satisfação (amarela) mostra um crescimento salarial mais rápido e que logo de início tem remunerações maiores, fato esse que tende a se manter ao decorrer do tempo, evidenciado pelo gráfico, onde a linha da alta satisfação está constantemente acima das outras.

Em contrapartida, as linhas representando os grupos de satisfação média (vermelho) e baixa (azul) apresentam um atraso relativo em comparação com o nível de satisfação alto. Embora o crescimento salarial parece ocorrer com intensidade semelhante para todos os grupos, os programadores com níveis de satisfação média e baixa levam mais tempo para atingir patamares salariais elevados. Além disso, esses grupos experimentam um padrão salarial semelhante, com uma diferença mínima entre eles ao longo de grande parte da carreira, sugerindo que a tendência é que a satisfação média ou baixa tende a resultar em trajetórias salariais parecidas.

No entanto, ao se aproximar dos 22 anos de carreira, o gráfico indica uma queda geral nos salários médios para todos os níveis de satisfação. Embora esse e outros comportamentos sejam muito similares entre todos estes níveis, não é possível afirmar nem uma causalidade ou influência direta entre categorias de dados. Entretanto, esse declínio abre margem para questionamentos e hipóteses sobre possíveis fatores associados.

4 CONSIDERAÇÕES FINAIS

Este estudo demonstrou que é possível identificar relações interessantes entre diferentes variáveis no "*Stack Overflow 2018 Developer Survey*". Através de análises e visualizações, é mostrado algumas das relações possíveis de serem feitas e o que elas estão mostrando de tendências. No entanto, é importante reconhecer as limitações deste estudo, incluindo a possibilidade de viés nos dados coletados e análises que buscam apenas identificar relações. Futuros trabalhos podem explorar outros datasets ou expandir a análise para incluir outras variáveis de interesse, contribuindo para uma compreensão ainda mais profunda do ambiente de trabalho dos desenvolvedores.

REFERÊNCIAS

MORETTIN, Pedro Alberto; BUSSAB, Wilton de Oliveira. **Estatística Básica**. 9. ed. [S.l.]: Saraiva Uni, 2017. P. 568.

WASKOM, Michael. seaborn: statistical data visualization. **Journal of Open Source Software**, v. 6, abr. 2021. Disponível em: <https://joss.theoj.org/papers/10.21105/joss.03021>. Acesso em: 10 out. 2024.