MACHINE LEARNING E VISÃO COMPUTACIONAL: ANÁLISE DE UMA BASE DE DADOS UTILIZANDO MÉTODOS ESTATÍSTICOS

Daniel Cepluki Junior
Donathan Ramalho Gonçalves
Joyce Santos Mendes

RESUMO
O presente artigo tem como objetivo central apresentar uma análise aprofundada abrangendo um conjunto de informações provenientes de uma base de dados de vagas de emprego da plataforma de recrutamento e seleção Glassdoor referente ao ano de 2017, focando no setor de tecnologia. O conjunto de dados fornece informações sobre posições de trabalho e salários. A pesquisa utiliza análise descritiva de dados, além dos métodos estatísticos média aritmética para criar gráficos informativos e aprofundar a análise dos dados. Ao desenvolver esta pesquisa foi possível compreender as diferenças significativas entre os setores, destacando as áreas com as médias salariais mais altas e mais baixas. Além disso, foram aplicados princípios de aprendizado de máquina para extrair insights e embasar decisões de profissionais, empresas e pesquisadores em Engenharia de Software e áreas relacionadas.

Palavras-chave: análise; métodos estatísticos;conjunto de dados.

MACHINE LEARNING AND COMPUTER VISION: ANALYSIS OF A DATABASE USING STATISTICAL METHODS

ABSTRACT
This article conducts a detailed analysis of a job vacancy dataset from the Glassdoor recruitment and selection platform for the year 2017, with a focus on the technology sector. The dataset provides information about job positions and salaries. The research employs descriptive data analysis, as well as statistical methods like arithmetic mean and K-NN(K-Nearest Neighbors) to create informative graphs and delve into data analysis. Stemming from the Machine Learning and Computer Vision discipline of the 6th semester of Software Engineering at UniSenai - CIC, conducting this research allowed for an understanding of significant differences between sectors, highlighting areas with the highest and lowest salary averages. Additionally, machine learning and computer vision principles were applied to extract insights and support decision-making for professionals, companies, and researchers in Software Engineering and related fields.

Key words: Analysis; Statistical Methods; Dataset.

1 INTRODUÇÃO

A análise de dados é o processo de examinar, limpar, transformar e modelar conjuntos de dados para obter insights valiosos e informação útil. Ela envolve o uso de métodos estatísticos, ferramentas de visualização e técnicas de mineração de dados para descobrir padrões, tendências e relações nos dados. A análise ajuda a compreender o contexto, identificar problemas, oportunidades e tomar decisões embasadas em evidências (ZENDESK, 2022).

O presente artigo tem como objetivo central apresentar uma análise aprofundada abrangendo um conjunto de informações provenientes de uma base de dados de vagas de emprego da plataforma de recrutamento e seleção Glassdoor referente ao ano de 2017, focando no setor de tecnologia. O conjunto de dados analisados oferece uma perspectiva sobre as tendências salariais. A análise de dados, respaldada por métodos estatísticos, desempenha um papel importante nesse processo, permitindo a identificação de fatores determinantes nos padrões de remuneração.

Este estudo busca aplicar métodos analíticos para aprofundar a compreensão dos dados salariais e suas implicações. Através da análise detalhada dos salários em diferentes cenários, o intuito é não apenas identificar as influências cruciais nos ganhos, mas também fornecer informações valiosas para profissionais, empresas e pesquisadores que atuam no campo da Engenharia de Software e áreas afins. Ao longo deste artigo, explorou-se o potencial da análise de dados para desvendar padrões e tendências salariais, fornecendo insights significativos que podem orientar a tomada de decisões.

2 FUNDAMENTAÇÃO

Neste capítulo, será apresentado uma breve exposição sobre a estatística e a análise de dados, enfatizando as ferramentas essenciais empregadas ao longo do processo. A compreensão de padrões em conjuntos de dados é vital, e a análise de dados desempenha um papel central na interpretação dessas informações. São destacadas ferramentas como Glassdoor, SQL Server, Kaggle e Tableau, que colaboram de maneira sinérgica, desde a coleta até a interpretação final dos dados. Essa combinação de recursos proporciona uma

base robusta para explorar as nuances e complexidades inerentes aos conjuntos de dados analisados.

2.1 Estatística e Análise de dados

A estatística é uma disciplina matemática que se concentra na coleta, análise, interpretação, apresentação e organização de dados. Ela desempenha um papel crucial na tomada de decisões em diversas áreas, desde a ciência e a pesquisa acadêmica até os negócios e a política. A estatística envolve a coleta de dados, seja por meio de observações diretas, pesquisas ou experimentos, e ajuda a resumir e descrever esses dados por meio de medidas como média, mediana, moda, desvio padrão, entre outras. A estatística é usada para fazer inferências ou previsões com base em dados amostrais, extrapolando para uma população maior. Gráficos e visualizações estatísticas são frequentemente usados para representar dados de forma compreensível. Ela é fundamental na identificação de padrões e tendências nos dados, avaliação de hipóteses, tomada de decisões, previsões, avaliação de riscos e incertezas, além de desempenhar um papel fundamental na pesquisa científica, permitindo a validação deresultados e a generalização de conclusões com base em amostras de dados. Em resumo, a estatística é uma ferramenta essencial na análise de dados, capacitando profissionais a tomar decisões embasadas em evidências sólidas (MEDRI, 2011).

2.2 Glassdoor

O Glassdoor é uma das maiores plataformas internacionais de recrutamento e seleção de vagas de emprego. Para as empresas que utilizam o serviço, o foco está na construção de sua marca empregadora (employer branding) para atrair os melhores talentos. Por outro lado, para os candidatos, o Glassdor oferece não apenas vagas, mas também avaliações de funcionários e ex-funcionários das empresas, proporcionando insights sobre a rotina no ambiente de trabalho. Além disso, os candidatos têm a opção de seguir empresas para receber atualizações de conteúdo relacionado a elas (B2BSTACK, 2022).

2.3 Kaggle

O Kaggle é uma plataforma abrangente para aprendizado de ciência de dados e a maior comunidade online dedicada a assuntos relacionados à Data Science. O Kaggle é conhecido por oferecer tutoriais, competições premiadas, cursos, rankings, fóruns, datasets e uma rica gama de recursos para profissionais e entusiastas da área. Além de seu aspecto comunitário, o software se destaca por suas competições que contribuem para a profissionalização das práticas em ciência de dados (HASHTAG, 2022). Neste contexto, a Kaggle foi a plataforma escolhida para a extração de dados da base da Glassdoor, utilizada como material de pesquisa.

2.4 SQL Server

O SQL Server é um sistema de gerenciamento de banco de dados relacional desenvolvido pela Microsoft. O software, escolhido para o tratamento dos dados que serão apresentados, é projetado para armazenar, recuperar, gerenciar e manipular grandesquantidades de dados de forma eficiente. O SQL Server é amplamente utilizado por organizações de todos os tamanhos, desde pequenas empresas até grandes corporações, devido à sua confiabilidade, desempenho e recursos avançados (PACIEVITCH, 2023).

2.5 Tableau

Originado em 2003 como um desdobramento de um projeto de ciência da computação na Universidade Stanford, o Tableau é uma plataforma de análise visual amplamente adotada no campo de Business Intelligence. Com essa poderosa ferramenta, é possível extrair dados brutos e convertê-los em análises acessíveis e de fácil compreensão, simplificando o processo de interpretação dos dados (ALURA, 2022). Por este motivo, e além da familiaridade dos autores ao utilizar o tableau diariamente, os gráficos apresentados posteriormente foram desenvolvidos utilizando essa ferramenta.

3. METODOLOGIA

Nesta seção, será apresentada a metodologia utilizada neste artigo, detalhando as abordagens e técnicas empregadas na coleta, análise e interpretação dos dados.

3.1 Coleta de dados

Para aplicar os conceitos de análise de dados utilizando métodos estatísticos, o primeiro passo a ser realizado foi a coleta de dados. De acordo com Grus (2016), a coleta de dados envolve a obtenção de informações brutas de diversas fontes, como bancos de dados, APIs, arquivos, raspagem da web e outros métodos. O autor enfatiza que a qualidade e a integridade dos dados coletados são cruciais para garantir resultados precisos e confiáveis em projetos de ciência de dados.

Por intermédio da ferramenta kaggle, que oferece uma ampla variedade de conjuntos de dados, cuja escolhida para este processo foi a de predição salarial, tal base contém dados de vagas de emprego do *Dataset* (conjunto de dados) da plataforma de recrutamento e seleção Glassdoor. Os referidos dados contêm informações como cargo, descrição do cargo, setor, salário, média salarial, ano etc.

3.2 Tratamento dos dados

Para transformar os dados obtidos em informações relevantes, foi necessária uma limpeza. A limpeza de dados é um processo fundamental na preparação e manutenção de conjuntos de dados. Ela envolve a identificação e correção de inconsistências, erros, duplicações e valores ausentes nos dados, tornando-os mais precisos e confiáveis. A limpeza de dados é essencial para garantir que as análises e os modelos de dados sejam baseados em informações de qualidade (OLIVEIRA, 2004).

A utilização de dados limpos e de qualidade, sem apontamentos discrepantes, permite aumentar a acuracidade e eficiência das decisões por parte dos gestores e tomadores de decisão, principalmente a partir da

consciência de que a maior precisão dos dados demonstra de maneira mais efetiva a realidade dos fatos por eles representados. (MACHADO; WILDAUER; 2021).

Nas figuras 1-1 e 1-2, prints tirados do banco de dados utilizando o software SQL server antes do tratamento, é possível visualizar como os dados estão originalmente dispostos na base de dados.

Figura 1-1 - Dados brutos

	Job_Title	Salary_Estimate	min_salary	max_salary	avg_salary	Industry	Sector
1	Data Scientist	$53K-$91K (Glassdoor est.)	53	91	720	Aerospace & Defense	Aerospace & Defense
2	Healthcare Data Scientist	$63K-$112K (Glassdoor est.)	63	112	875	Health Care Services & Hospitals	Health Care
3	Data Scientist	$120K-$160K (Glassdoor est.)	120	160	1400	Internet	Information Technology
4	Data Scientist	$80K-$90K (Glassdoor est.)	80	90	850	Security Services	Business Services
5	Data Scientist	$56K-$97K (Glassdoor est.)	56	97	765	Energy	Oil, Gas, Energy & Utilities
6	Data Scientist	$86K-$143K (Glassdoor est.)	86	143	1145	Advertising & Marketing	Business Services
7	Data Analyst	$46K-$85K (Glassdoor est.)	46	85	655	Advertising & Marketing	Business Services
8	Customer Data Scientist	$118K-$189K (Glassdoor est.)	118	189	1535	Enterprise Software & Network Solutions	Information Technology
9	Data Scientist	$71K-$119K (Glassdoor est.)	71	119	950	Real Estate	Real Estate
10	Data Scientist	$54K-$93K (Glassdoor est.)	54	93	735	Banks & Credit Unions	Finance
11	Research Scientist	$38K-$84K (Glassdoor est.)	38	84	610	Health Care Services & Hospitals	Health Care
12	Data Scientist	$86K-$142K (Glassdoor est.)	86	142	1140	Consulting	Business Services
13	Data Scientist	$126K-$201K (Glassdoor est.)	126	201	1635	Other Retail Stores	Retail
14	Data Scientist	$75K-$124K (Glassdoor est.)	75	124	995	Banks & Credit Unions	Finance
15	Associate Data Analyst	$34K-$61K (Glassdoor est.)	34	61	475	Insurance Carriers	Insurance
16	Clinical Data Scientist	$63K-$105K (Glassdoor est.)	63	105	840	Biotech & Pharmaceuticals	Biotech & Pharmaceuticals
17	Data Scientist	$64K-$106K (Glassdoor est.)	64	106	850	Research & Development	Business Services
18	Staff Data Scientist - Technology	$106K-$172K (Glassdoor est.)	106	172	1390	Department, Clothing, & Shoe Stores	Retail
19	Data Scientist	$83K-$144K (Glassdoor est.)	83	144	1135	Biotech & Pharmaceuticals	Biotech & Pharmaceuticals
20	Data Engineer I	$102K-$190K (Glassdoor est.)	102	190	1460	Motion Picture Production & Distribution	Media
21	Senior Data Scientist	$110K-$150K(Employer est.)	110	150	1300	Real Estate	Real Estate
22	Senior Data Scientist	$80K-$129K (Glassdoor est.)	80	129	1045	Transportation Equipment Manufacturing	Manufacturing

Fonte: Os Autores

Figura 1- 2 - Dados brutos

Job_Description	Rating	Company_Name	Location	Headquarters	Size
Data Scientist Location: Albuquerque, NM Education R...	38	Tecolote Research 3.8	Albuquerque, NM	Goleta, CA	501 to 1000 emp
What You Will Do: I. General Summary The Healthcar...	34	University of Maryland Medical System 3.4	Linthicum, MD	Baltimore, MD	10000+ employe
isn't your usual company. Our work is powered by the p...	46	<intent> 4.6	New York, NY	New York, NY	51 to 200 emplo
KnowBe4, Inc. is a high growth information security com...	48	KnowBe4 4.8	Clearwater, FL	Clearwater, FL	501 to 1000 emp
"Organization and Job ID"" Job ID: 310709 Directorate:...	38	PNNL 3.8	Richland, WA	Richland, WA	1001 to 5000 em
Data Scientist Affinity Solutions / Marketing Cloud seeks...	29	Affinity Solutions 2.9	New York, NY	New York, NY	51 to 200 emplo
Are you an experienced Data Analyst, skilled at providin...	41	Yesler 4.1	Seattle, WA	Seattle, WA	201 to 500 empl
Company Overview H2O.ai is the open source leader i...	43	h2o.ai 4.3	Mountain View, CA	Mountain View, CA	201 to 500 empl
CyrusOne is seeking a talented Data Scientist who hold...	34	CyrusOne 3.4	Dallas, TX	Dallas, TX	201 to 500 empl
Job Description **Please only local candidates apply - t...	41	ClearOne Advantage 4.1	Baltimore, MD	Baltimore, MD	501 to 1000 emp
SUMMARY The Research Scientist I will be tasked with ...	33	Rochester Regional Health 3.3	Rochester, NY	Rochester, NY	10000+ employe
Advanced Analytics - Lead Data Scientist Overview W...	38	Logic20/20 3.8	San Jose, CA	Seattle, WA	201 to 500 empl
At Wish, our Data Science & Engineering team is compr...	35	Wish 3.5	San Jose, CA	San Francisco, CA	501 to 1000 emp
As the Enterprise Data and Analytics Center of Excellen...	35	First Tech Federal Credit Union 3.5	Hillsboro, OR	San Jose, CA	1001 to 5000 em
Friday, January 17, 2020 Our Enterprise Data and Ana...	37	The Hanover Insurance Group 3.7	Worcester, MA	Worcester, MA	5001 to 10000 e
THIS ROLE MUST BE BASED IN SAN DIEGO As part o...	40	Pfizer 4.0	Groton, CT	New York, NY	10000+ employe
Secure our Nation, Ignite your Future Summary The s...	41	ManTech 4.1	Chantilly, VA	Herndon, VA	5001 to 10000 e
Position Summary... Drives the execution of multiple bu...	32	Walmart 3.2	Plano, TX	Bentonville, AR	10000+ employe
Job Description Takeda is looking for a Data Scientist ...	37	Takeda Pharmaceuticals 3.7	Cambridge, MA	OSAKA, Japan	10000+ employe
This opportunity is within Audibles Data Engineering gr...	36	Audible 3.6	Newark, NJ	Newark, NJ	1001 to 5000 em
Working for Equity Residential (EQR), a leading multi-fa...	43	Equity Residential 4.3	Chicago, IL	Chicago, IL	1001 to 5000 em
We are AAM. We have the POWER to move the world. ...	33	American Axle & Manufacturing 3.3	Southfield, MI	Detroit, MI	10000+ employe

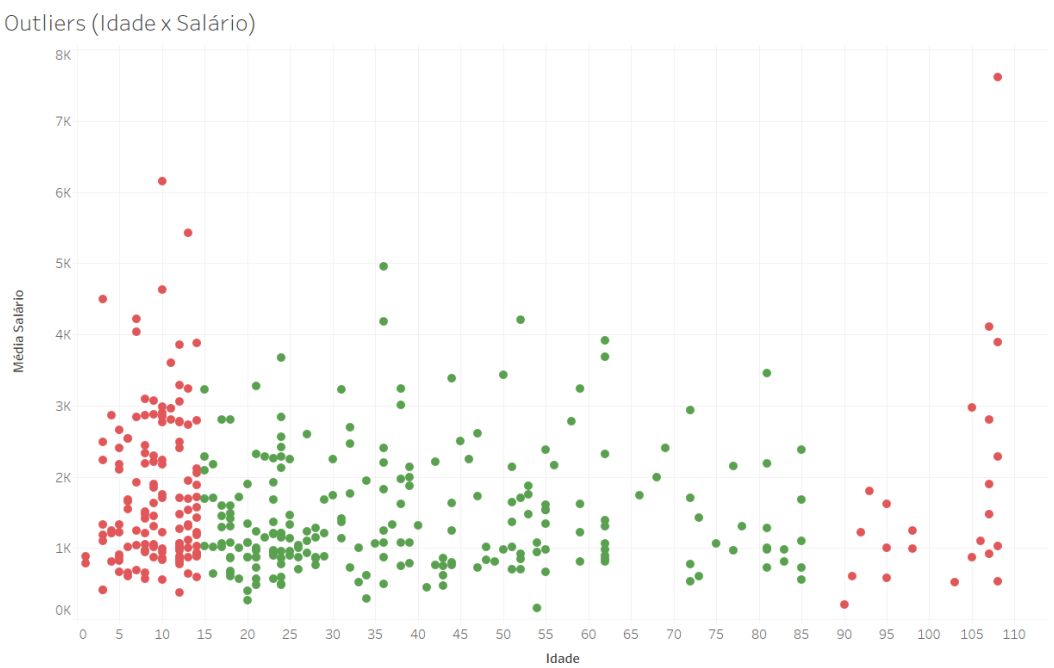
Fonte: Os Autores

Para dividir os dados em mais campos a fim de facilitar a leitura das informações, bem como identificar "*Outliers - valores discrepantes*", foi realizada uma análise descritiva.

A análise descritiva é uma etapa fundamental na exploração de dados que se concentra em resumir e descrever as principais características de um conjunto de dados. Isso inclui o cálculo de estatísticas resumidas, como média, mediana, moda, desvio padrão e quartis, além da criação de visualizações gráficas, como histogramas e gráficos de dispersão. A análise descritiva permite compreender a distribuição dos dados, identificar padrões, tendências e discrepâncias, fornecendo uma visão geral dos dados que ajuda na tomada de decisões informadas e na formulação de hipóteses para análises mais avançadas (REIS, 2002).

Na figura 2, é possível observar os "*Outliers*" em uma comparação entre Idade x Salário, onde os dados em vermelho representam os valores discrepantes e os dados em verde os dados corretos para análise.

Figura 2 - Outliers



Fonte: O Autor

Neste gráfico, os *outliers* são dados que não refletem a realidade, uma vez que inclui crianças de 0 a 14 anos e idosos de 90 a 110 anos em uma representação de salários. Isso é improvável, pois assume-se que crianças não trabalham de forma remunerada e idosos em uma faixa etária tão avançada também não são a principal força de trabalho (CORSEUIL, 2014).

Visualmente o gráfico não aponta uma relação direta entre a idade e o

salário. Para confirmar essa informação foi utilizado o método de correlação de Pearson. De acordo com MORE, McCABE e CRAIG (2015), a Correlação de Pearson é um método estatístico usado para avaliar a relação linear entre duas variáveis contínuas. Ela mede a força e a direção da associação entre essas variáveis, indicando se elas estão positivamente correlacionadas (aumentam juntas), negativamente correlacionadas (quando uma aumenta, a outra diminui) ounão têm correlação.

O coeficiente de correlação de Pearson, frequentemente representado como "r," varia de -1 a 1, onde:

- r = 1: Correlação positiva perfeita.
- r = -1: Correlação negativa perfeita.
- r = 0: Ausência de correlação

Passo a passo para a realização do cálculo:

- Colocar os dados "Idade" e "Salário" em uma lista
- Calcular a média total de cada medida.
- Calcular o desvio (Valor n - Média) para cada medida.
- Elevar o desvio ao quadrado para cada medida.
- Realizar a soma do quadrado dos desvios.
- Pegar a soma do quadrado dos desvios e dividir pela raiz quadrada do produto dos desvios ao quadrado.

O resultado obtido, Rxy = 0,08, revela um valor extremamente próximo de zero. Com base nesse resultado, é possível afirmar que não existe uma relação direta ou indireta significativa entre essas duas variáveis. O coeficiente de correlação próximo de zero indica uma associação muito fraca entre as variáveis, sugerindo que as variações em uma variável não estão fortemente relacionadas às variações na outra. Essa falta de correlação destaca a independência ou a ausência de influência mútua substancial entre as variáveis em questão.

3.3 Análise multivariada

Neste capítulo, serão abordadas técnicas estatísticas fundamentais para a análise da base de dados em questão. Entre as metodologias exploradas, destaca-se o algoritmo K-Nearest Neighbors (KNN), amplamente utilizado em

aprendizado de máquina para classificação e regressão, sendo aplicado aqui para compreender a formação de grupos na base de dados. Além disso, será examinada a aplicação da média aritmética, uma medida estatística central, para avaliar disparidades salariais entre setores e cargos. A correlação de Pearson será empregada para investigar possíveis associações lineares entre variáveis, com foco na relação entre idade e salário. Por fim, será introduzido o teste de Análise de Variância (ANOVA), uma ferramenta crucial para avaliar diferenças significativas entre as médias de grupos, proporcionando uma compreensão mais profunda da dinâmica salarial na base de dados analisada. Essas técnicas combinadas oferecerão insights valiosos sobre padrões erelações presentes nos dados, contribuindo para uma interpretação mais abrangente e informada por parte do pesquisador.k-nearest neighbors.

O K-Nearest Neighbors (K-NN) é um algoritmo de aprendizado de máquina utilizado para classificação e regressão. Ele funciona com base na ideia de que objetos semelhantes tendem a estar próximos no espaço de características (PETERSON. 2009).

Ainda de acordo com Peterson, para classificar um novo ponto de dados, o algoritmo encontra os K vizinhos mais próximos no conjunto de treinamento e determina a classe com base na maioria dos vizinhos. A escolha do valor de K é crucial, pois afeta diretamente o desempenho do modelo.

Por exemplo, um conjunto de dados com várias classes (como gatos, cachorros e pássaros) e um novo ponto cuja classe é desconhecida, o K-NN analisa os K pontos mais próximos e classifica o novo ponto com base nas classes desses vizinhos. A classe mais comum entre os vizinhos será atribuída ao novo ponto.

Como o objetivo era avaliar a correlação entre as medidas "Rating" e "Salário", foram definidos 4 grupos para se ter uma visão de quadrante e validar se há alguma relação entre satisfação dos funcionários e salários. A escolha do número de grupos foi realizada com base no desenvolvimento de um código em Python para realizar uma análise de variância (ANOVA) de um fator.

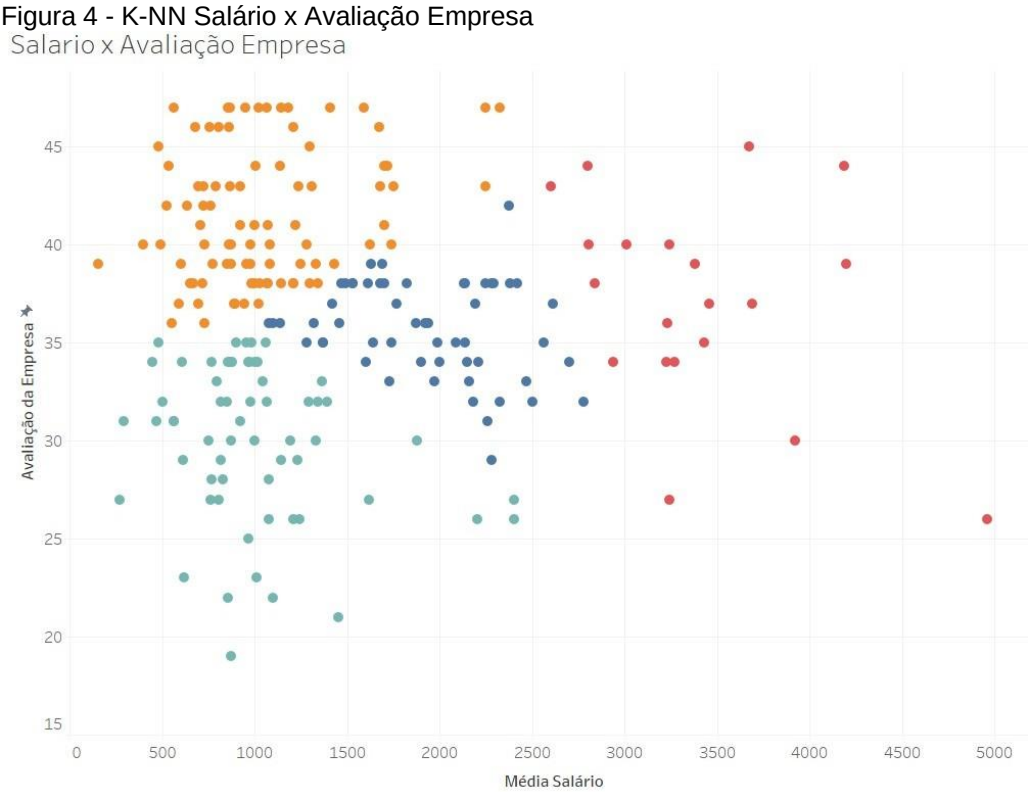
A Anova refere-se a uma análise de variância, uma técnica estatística utilizada para avaliar se há diferenças significativas entre as médias de três ou mais grupos (BUSSAB; MORETTIN, 2017). O objetivo principal é avaliar a variabilidade na separação dos grupos gerados pelo algoritmo KNN (K-Nearest

Neighbors). O código utiliza a coluna "Grupo" para atribuir cada dado a um cluster específico identificado pelo KNN.

Dado que se trata de uma ANOVA de um fator, a análise foi conduzida de forma separada para as variáveis "Rating" e "Salary". Isso permite avaliar a variação dessas variáveis entre os grupos formados com base no critério Calinski-Harabasz no KNN. Essa abordagem visa compreender como as variáveis respondem à separação dos dados em clusters pelo algoritmo KNN, proporcionando insights sobre a eficácia do método de agrupamento.

Como é possível observar na figura 4, não existe uma relação direta ou indireta entreas medidas, uma vez que ambas variam de forma independente.

Figura 4 - K-NN Salário x Avaliação Empresa



Fonte: Os Autores

Categorização com base em Média Salarial e Avaliação:

Em Laranja, média salarial mais baixa - avaliação mais alta;

Em Verde, média salarial mais baixa - avaliação mais baixa;

Em Azul, média salarial média - avaliação média;

Em Vermelho, média salarial alta - avaliação variável;

3.3.1 Média Aritmética

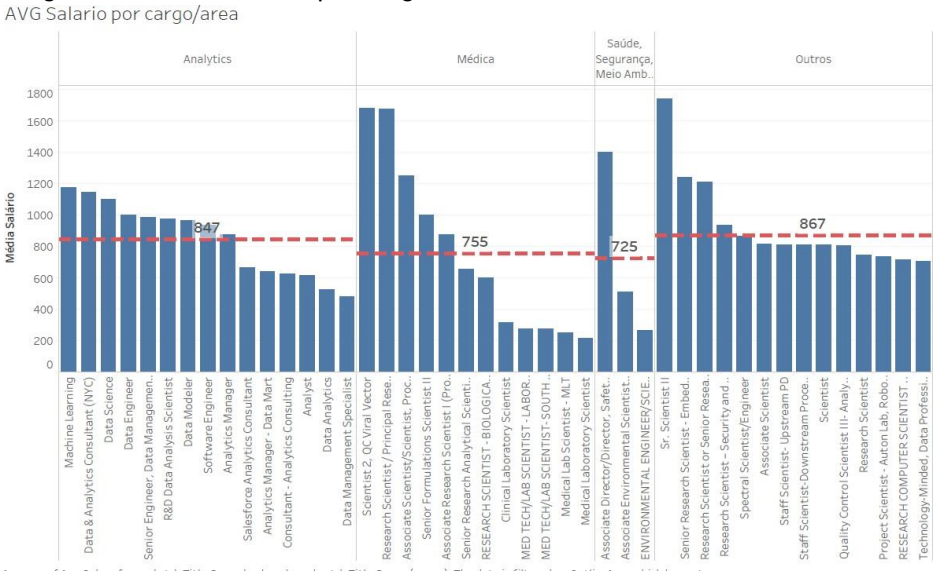
A média aritmética é um conceito fundamental em Estatística e na prática experimental, com aplicação significativa tanto no ambiente escolar quanto na vida cotidiana (Gal, 1995).

Conforme observado por Strauss e Bichler (1988), a média aritmética é caracterizada por sete propriedades, cuja compreensão por parte do indivíduo indica o domínio do conceito:

3.3.1.1 a média está localizada entre os valores extremos (mínimo ≤ média ≤ máximo);
3.3.1.2 a soma dos desvios a partir da média é zero (Σ (Xi – média) = 0);
3.3.1.3 a média é influenciada por cada um e por todos os valores (média = Σ Xi / n);a média não necessariamente tem que coincidir com um dos valores;
3.3.1.4 a média pode ser um uma fração que não tem uma contrapartida na realidade física(por exemplo, o número médio de filhos por mulher é igual a 2,3);
3.3.1.5 o cálculo da média leva em consideração todos os valores, inclusive os nulos e osnegativos, e
3.3.1.6 a média é um valor representativo dos dados a partir dos quais ela foi calculada. Emtermos espaciais, a média é aquela que está mais próxima de todos os valores.

A média aritmética foi um dos métodos estatísticos utilizados no estudo. Na Figura 5,é visível a média salarial de cargos diversos no setor de tecnologia, abrangendo áreas como análises, medicina, segurança, meio ambiente e outras, classificadas como diárias. A linha vermelha representa a média por área.

Figura 5 - Média Salarial por cargo/área



Fonte: Os Autores (2023)

Utilizando o teste Anova, foi realizado o experimento da variação da média dos salários separados por grupos de cargo, ou seja, dentro da base de vagas, à diversas posições, a partir da descrição do cargo, a posição era adicionada à um dos 3 grupos, sendo eles analista, engenheiro de dados ou cientista de dados.

De acordo com BUSSAB; MORETTIN (2017), o valor p (p-value) é uma medida estatística que ajuda a quantificar a evidência contra uma hipótese nula em um teste estatístico. Por convenção o valor de alfa, frequentemente estabelecido em 5% (0,05), é uma escolha comum para o nível de significância em testes estatísticos. O teste apresentou um resultado de 0.145, isso significa que o valor está fora dos 5% do valor de alfa, o que sugere uma maior probabilidade de um valor de um grupo de cargos estar em outro grupo.

Isso indica que independentemente de você escolher uma função de analista, engenheiro de dados ou cientista de dados, a variação do salário desses grupos, dentro da área de tecnologia, não tem relevância estatística.

4. RESULTADOS

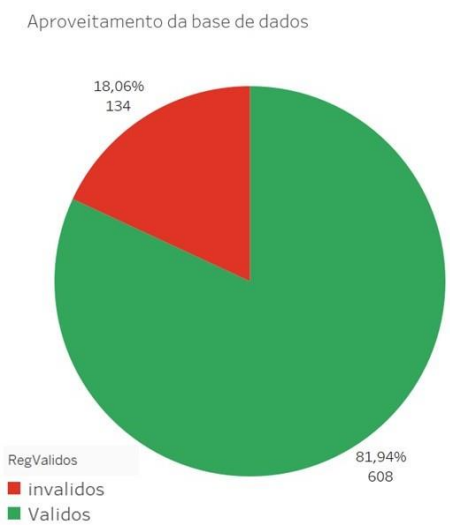
A análise de dados é uma ferramenta valiosa para obter insights e informações úteis a partir de conjuntos de dados. Este estudo concentrou-se na análise da base de dados de salários, fornecendo uma visão abrangente das tendências salariais em diferentes setores e segmentos de mercado.

Observou-se que a análise de dados é fundamental para identificar fatores determinantes nos padrões de remuneração. No contexto da disciplina de Machine Learning e Visão Computacional, foram aplicados métodos analíticos para compreender os dados salariais e suas implicações. Explorou-se o vasto potencial da análise de dados para desvendar padrões e tendências salariais, fornecendo insights significativos para profissionais, empresas e pesquisadores no campo da Engenharia de Software e áreas afins.

O processo de análise incluiu a coleta de dados por meio da plataforma Kaggle, que ofereceu uma ampla variedade de conjuntos de dados, incluindo o de predição salarial da plataforma Glassdoor. Foi realizada a limpeza e tratamento dos dados para garantir sua qualidade e integridade, identificando inclusive valores discrepantes que poderiam afetar as análises. Neste contexto,

ainda utilizando a ferramenta Tableau, em sua linguagem própria foi realizada uma análise quantitativa dos dados aproveitados na pesquisa, bem como os descartados. Esta análise consistiu no desenvolvimento de uma condição utilizando lógica de programação. A condição explorada para este tratamento foi a faixa etária, distribuída entre idades menores que 14 anos e maiores que 90 anos, e apresentado utilizando o gráfico de setores ou pizza como comumente chamado. Este gráfico originou-se quando o matemático e engenheiro escocês William Playfair começou a desenvolver várias técnicas de visualização de dados, e tem sido bastante utilizados em análises quantitativas, cujo objetivo é dispor em setores ou fatias, os valores encontrados (SPENCE, 2005). A figura 6, apresenta o resultadodesta análise:

Figura 6 - Análise quantitativa de aproveitamento de dados



Fonte: Os Autores (2023).

A análise disposta no gráfico da figura 6, demonstra que dentre 742 registros, 18,06% são classificados como dados discrepantes e 81,94% de dados aproveitados, ou seja, quantitativamente falando, uma boa métrica de dados foi proveitosa para o estudo.

Os resultados da análise de correlação de Pearson entre a idade e o salário revelaram um coeficiente (Rxy) de 0,08, indicando uma associação extremamente fraca entre essas variáveis. Essa pontuação próxima de zero sugere a ausência de uma relação significativa e linear entre a idade e o salário dos indivíduos no conjunto de dados analisados. O gráfico, previamente identificado com outliers representando faixas etárias improváveis para

participação no mercado de trabalho remunerado, foi corroborado pela fraca correlação. Assim, os dados sugerem que, dentro do escopo dessa análise, a variação na idade não está substancialmente ligada à variação nos salários, fortalecendo a conclusão de independência entre essas duas variáveis no contexto considerado.

A análise utilizando o algoritmo K-Nearest Neighbors (K-NN) destacou a importância da escolha do valor de K para o desempenho do modelo. A definição de 4 grupos foi guiada pela necessidade de compreender a relação entre satisfação dos funcionários e salários. A aplicação da análise de variância (ANOVA) de um fator, conforme Bussab e Morettin (2017), permitiu avaliar diferenças significativas entre as médias das variáveis "Rating" e "Salary" em relação à formação de clusters pelo K-NN. Os resultados sugerem que essas variáveis variam de forma independente, não apresentando uma influência significativa uma sobre a outra dentro do contexto da formação de clusters pelo algoritmo K-Nearest Neighbors (K-NN). Essa constatação reforça a importância de compreender a dinâmica específica das variáveis e como elas respondem à separação em grupos, proporcionando uma perspectiva clara sobre a falta de associação entre satisfação dos funcionários e salários nesse cenário específico.

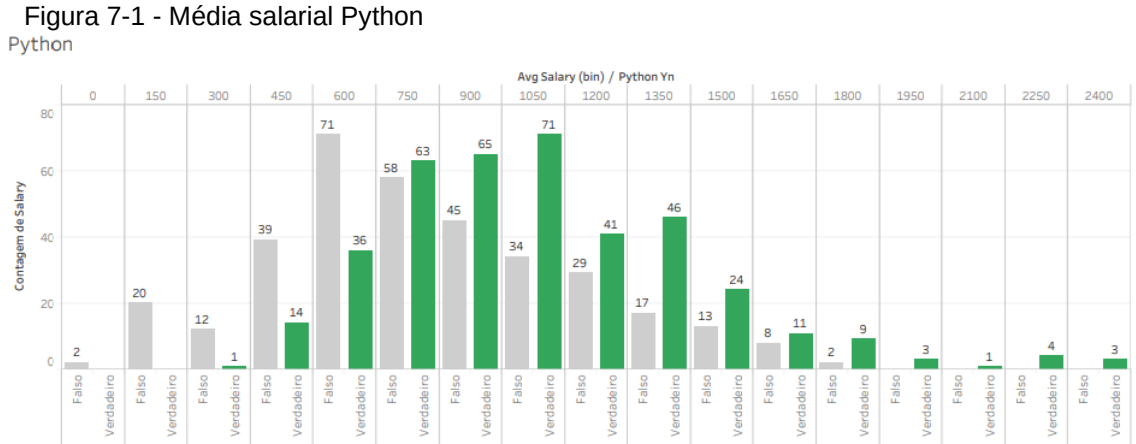
Além disso, explorou-se o conceito de média aritmética e suas propriedades, conforme destacado por Strauss e Bichler (1988), demonstrando a importância dessa medida estatística na análise de dados. Através do gráfico de barras/colunas foi apresentado a disparidade salarial entre setores e cargos da tecnologia no ano de 2017, dispostos na base de dados da plataforma Glassdoor.

Outro ponto apurado também utilizando a média, foi análise salarial com base na ferramenta utilizada. Nesse caso foi realizada a média entre as linguagens Python, que de acordo com Borges (2014), é uma linguagem de programação de alto nível, interpretada e multiparadigma. Ela é conhecida por sua sintaxe clara e legível, o que a torna uma excelente escolha para programadores, desde iniciantes até profissionais experientes. E SQL, que significa "*Structured Query Language*" (Linguagem de Consulta Estruturada), é uma linguagem de programação especializada usada para gerenciar e manipular bancos de dados relacionais. Bancos de dados relacionais são sistemas que

armazenam dados em tabelas, com relações definidas entre essas tabelas (MELTON; SIMON.1999).

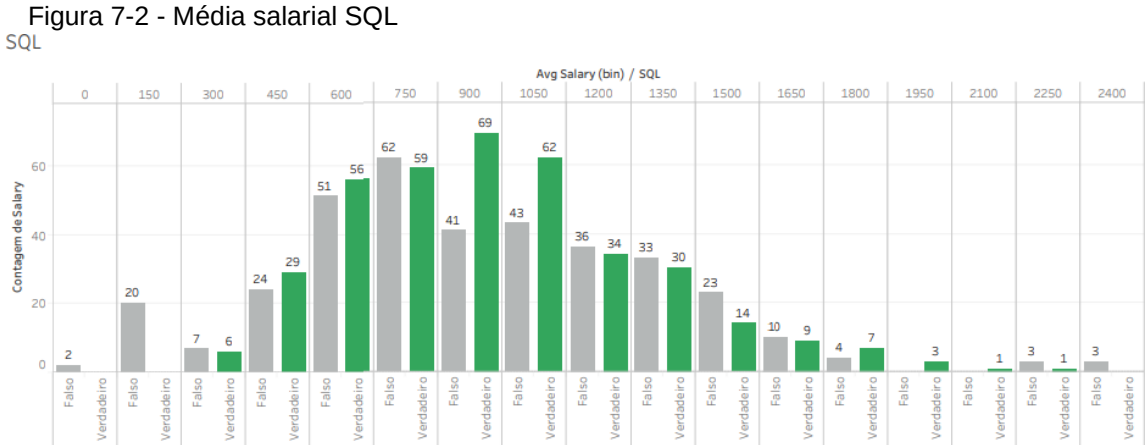
Dispostos na figura 7-1 e 7-1, estão os gráficos da média salarial entre as duaslinguagens mencionadas, sendo em verde a representação dos cargos que possuíam a ferramenta como requisito e em cinza os cargos que não.

Figura 7-1 - Média salarial Python



Fonte: Os Autores (2023)

Figura 7-2 - Média salarial SQL



Fonte: Os Autores (2023)

O gráfico de histograma revela que, em termos de média salarial, naquele ano de 2017, a habilidade com Python era mais necessária em faixas salariais entre $600 e $1500, mas a demanda por essa habilidade diminuiu em faixas salariais mais elevadas. Isso sugere que Python era mais requisitado em posições com salários intermediários. Para fazer uma comparação com a análise de SQL, o padrão é semelhante, com uma alta demanda em faixas salariais intermediárias, mas uma diminuição nas faixas salariais superiores, indicando

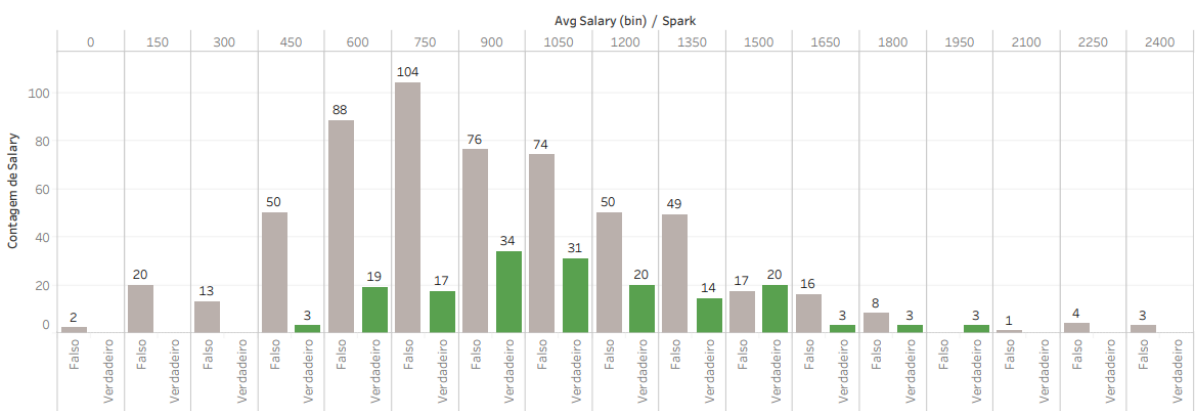
que SQL também era uma habilidade relevante, embora não tão crucial em posições de salário mais alto. Ambas as habilidades são valiosas, mas sua importância pode variar de acordo com o nível salarial da posição.

Outra comparação realizada dentre as ferramentas foram as habilidades em Excel, que é uma planilha eletrônica da Microsoft amplamente utilizada na estatística. Ele auxilia na organização e visualização de dados, oferece funções para cálculos estatísticos simples, criação de gráficos e realização de testes estatísticos básicos. Além disso, é útil na análise de regressão, amostragem, simulação e na documentação de resultados (FÁVERO, 2017). A ferramenta Spark, uma plataforma de processamento de dados em código aberto que é altamente escalável e eficiente para o processamento de grandes volumes de dados, e que embora não seja estritamente uma ferramenta estatística, ele desempenha um papel significativo na análise de dados e na estatística devido à sua capacidade de processamento rápido e distribuído (MENG, 2016).

Na figura 8-1 e 8-2, é possível analisar os gráficos comparativos entre as duas ferramentas.

Figura 8-1 - Média salarial Excel

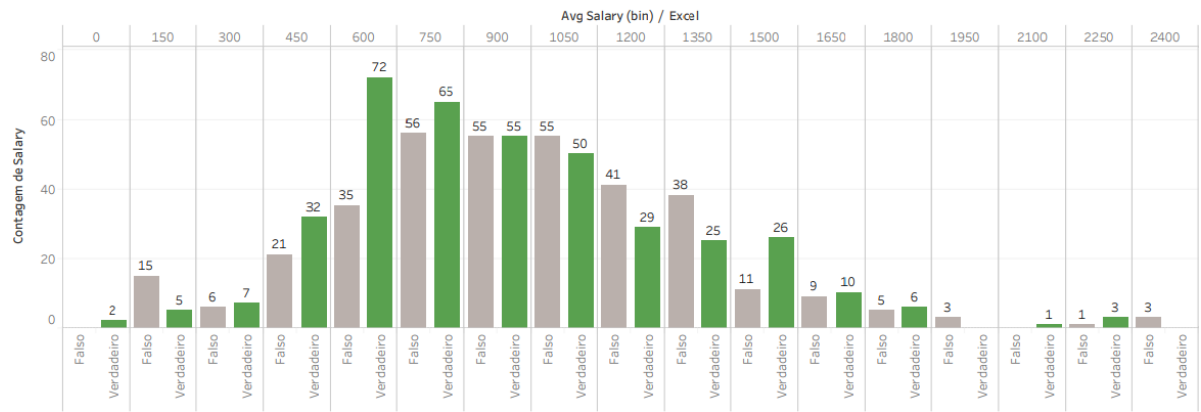
Spark



Fonte: Os Autores (2024)

Figura 8-2 - Média salarial Spark

Excel



Fonte: Os Autores (2024)

Com base no gráfico de histograma, a necessidade de habilidades com Spark e Excel é mais proeminente em faixas salariais entre $600 e $1500 e $600 e $1650, respectivamente. À medida que as faixas salariais se elevam, a demanda por ambas as habilidades tende a diminuir. A análise de Spark segue um padrão semelhante ao do Excel em relação à média salarial. Isso sugere que dentro do período analisado ambas as habilidades são mais cruciais em posições de salários intermediários, mas sua importância diminui à medida que os salários aumentam. Essas habilidades são valiosas, mas sua relevância pode variar dependendo do nível salarial da posição.

O teste ANOVA foi aplicado para avaliar a variação média dos salários em diferentes grupos de cargos (analista, engenheiro de dados e cientista de dados). Com um valor p de 0,145, fora do nível de significância de 5%, não há evidência estatística significativa para sugerir que a escolha entre essas funções impacta a variação salarial na área de tecnologia. Em resumo, a análise indica que a variação salarial entre esses grupos de cargos não é estatisticamente relevante, proporcionando insights sobre a independência estatística da escolha entre essas posições e os salários correspondentes.

4. CONSIDERAÇÕES FINAIS

Os resultados desta análise de dados oferecem uma visão abrangente das tendências salariais em diferentes setores e segmentos de mercado, especialmente no

contexto da disciplina de Machine Learning e Visão Computacional. A aplicação de métodos analíticos revelou-se fundamental para identificar fatores determinantes nos padrões de remuneração. A coleta de dados via Kaggle e a subsequente limpeza e tratamento desses dados foram etapas cruciais para garantir a qualidade e integridade das análises.

A análise quantitativa de aproveitamento de dados, apresentada no gráfico da Figura 6, destacou que 81,94% dos dados foram considerados aproveitáveis, reforçando a solidez do estudo. A fraca correlação (Rxy = 0,08) entre idade e salário, corroborada pelo gráfico de setores, indicou a ausência de uma relação linear significativa entre essas variáveis, especialmente considerando faixas etárias improváveis para participação no mercado de trabalho remunerado.

Ao explorar o algoritmo K-Nearest Neighbors (K-NN), a análise de variância (ANOVA) de um fator revelou que as variáveis "Rating" e "Salary" variam de forma independente, destacando a importância de compreender a dinâmica específica das variáveis ecomo elas respondem à formação de clusters pelo K-NN.

A média salarial foi uma ferramenta valiosa na análise, permitindo visualizar disparidades salariais entre setores, cargos e habilidades específicas. A comparação entre as linguagens de programação Python e SQL, assim como as ferramentas Excel e Spark, revelou padrões interessantes, indicando que a importância dessas habilidades pode variar com base no nível salarial da posição.

Finalmente, o teste ANOVA aplicado para avaliar a variação média dos salários entre diferentes grupos de cargos não apresentou evidência estatística significativa para sugerir que a escolha entre analista, engenheiro de dados ou cientista de dados impacta a variação salarial na área de tecnologia. Essa conclusão fortalece a ideia de independência estatística entre as escolhas de posições e os salários correspondentes nesse contexto específico. Em suma, esta análise proporciona insights valiosos para profissionais, empresas e pesquisadores, contribuindo para uma compreensão mais aprofundada das dinâmicas salariais na área detecnologia.

REFERÊNCIAS

BUSSAB, W. O., & MORETTIN, P. A. Estatística Básica. 10ª ed. SP: Editora Saraiva. 2017.

CATUNDA, H. O QUE É KAGGLE? Disponível em https://www.hashtagtreinamentos.com/kaggle. Acesso em: 18 de set. 2023.

CORSEUIL, C. H.L.; BOTELHO, R. U. Desafios à trajetória profissional dos jovens brasileiros. 2014. Disponível em https://www.ipea.gov.br/atlasviolencia/arquivos/artigos/5655-livrodesafioscompleto-web-com pactado.pdf. Acesso em: 20 de set. 2023.

FÁVERO, L. P.; BELFIORE, P. Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®**.** RJ: Elsevier Brasil, 2017.

GAL, I. Statistical tools and statistical literacy: the case of the average. Teaching Statistics, v. 17, n. 3, p. 97-99, 1995.

GRUS, J. Data Science do Zero. Editora Exemplo, 2016.

Kelley, W. M.; Donnelly, R. A.*The Humongous Book of Statistics Problems*. New York, NY: Alpha Books, 2009.

LIRA, M. GLASSDOOR: SITE DE VAGAS E RECRUTAMENTO. APRENDA COMO USAR! Disponível em: https://blog.b2bstack.com.br/glassdoor-aprenda-como-usar/. Acesso em: 18 de set. 2023.

BORGES, L. E. Python para desenvolvedores: aborda Python 3.3. Novatec Editora, 2014.

MACHADO, A. O.; WILDAUER, E. W. AVALIAÇÃO DA QUALIDADE DOS DADOS DO PROCESSO DE CONSUMO E ABASTECIMENTO DE COMBUSTÍVEL DE UMA FROTA DE ÔNIBUS SOB A DIMENSÃO DA PRECISÃO. Disponível em: https//www.even3.com.br/anais/casi2020/328574-AVALIACAO-DA- QUALIDADE-DOS-DADOS-DO-PROCESSO-DE-CONSUMO-E-ABASTECIMENTO-DE COMBUSTIVEL-DE-UMA-FROTA-DE-ONIBUS-SOB. Acesso em: 20 de set. 2023.

MEDRI, Waldyr. ANÁLISE EXPLORATÓRIA DE DADOS. Disponível em https://docs.ufpr.br/~benitoag/apostilamedri.pdf. Acesso em: 18 de set. 2023.

MELTON, J.; SIMON, A. R. SQL: 1999: understanding relational language components.Elsevier, 2001.

MENG, X. Mllib: Machine learning in apache spark**.** The journal of machine learning research, v. 17, n. 1, p. 1235-1241, 2016.

MOORE, D. S.; McCABE, G. P.; CRAIG, B. A. Introduction to the Practice of Statistics. W. H. Freeman and Company, 2015.

OLIVEIRA, P.; RODRIGUES, F.; H. P. Limpeza de dados: Uma visão geral. Data Gadgets, p. 39-51, 2004. Disponível em https://www.researchgate.net/profile/Fatima-Rodrigues-5/publication/266583309_Limpeza_de_Dados_-Uma_Visao_Geral/links/555c4b2d08ae8f66f3ade927/Limpeza-de-Dados-Uma-Visao-Geral.pdf. Acesso em: 20 de set. 2023.

PACIEVITCH, Y. SQL SERVER. Disponível em https://www.infoescola.com/informatica/sql-server/. Acesso em: 18 de set. 2023.

PETERSON, L. E. K-nearest neighbor. Scholarpedia, v. 4, n. 2, p. 1883, 2009. Disponível em http://scholarpedia.org/article/K-nearest_neighbor. Acesso em: 20 de set. 2023.

REIS, E. A.; REIS, I.A. Análise descritiva de dados. Relatório Técnico do Departamento de Estatística da UFMG, v.1, 2002. Disponível em https://www.est.ufmg.br/portal/wp-content/uploads/2023/01/RTE-02-2002.pdf. Acesso em: 20de set. 2023.

RIBEIRO, L. A.; SANTANA, L.C. de. Qualidade de vida no trabalho: fator decisivo para o sucesso organizacional. Revista de Iniciação Científica–RIC Cairu, v. 2, n. 02, p. 75-96, 2015. Disponível em:https://portalidea.com.br/cursos/qualidade-de-vida-no-trabalho-apostila04.pdf. Acesso em: 25 de set. 2023.

SPENCE, I. No humble pie: the origins and usage of a statistical chart. Journal of Educational and Behavioral Statistics Winter, v. 30, n. 4, p. 353-368. 2005.

STRAUSS, S.; Bichler, E. The development of children's concepts of the arithmetic average. Journal for Research in Mathematics Education, 19(1), 64-80, 1988.

[Recebido/Received: 07 Maio 2024; Aceito/Accepted: 10 Junho 2024]