Instituto de Ciências Matemáticas e de Computação
Departamento de Sistemas de Computação
SSC0800 - Introdução a Ciência da Computação I

Análise de Dados do Times World University Rankings 2024

Alunos: Arthur Martins Ferreira de Sousa, Gabriel Carbinato, Gabriel Ligabô Baba
Professor: Leonardo Tórtoro Pereira

Dezembro
2023

Conteúdo

1 Introdução

O advento da análise de dados e modelagem estatística tornou-se um grande expoente na conjectura tecnológica mundial. Com isso, surge a necessidade de praticar a aplicação de técnicas que contribuam à extração, limpeza, desenvolvimento, análise, classificação, inferência e modelagem de dados. Assim, o trabalho desenvolvido e apresentado a seguir serve como fonte de introdução e aprendizado às diversas propostas de data science e computação em python, com a meta de extrair informações sobre qualidade de ensino e educação ao nível superior mundial.

Dessa maneira, vistas as problemáticas enfrentadas mundialmente ao acesso à educação superior e curiosidade pela diversidade de ensino, o grupo escolheu a base de dados: Times World University Rankings 2024 (THE WUR2024). Tal database é uma coletânea bruta, categórica e numérica de 2673 instituições de ensino superior.

A The Times Higher Education (THE) tem reconhecimento por financiar coletâneas de informações essenciais sobre o cenário universitário global desde 2004 e o nosso objetivo é evidenciar desempenho em 5 panoramas: ensino, ambiente de pesquisa, qualidade de pesquisa, indústria e abordagem internacional (internacionalização). Das 2763 universidades participantes, somente 1904 foram rankeadas nos parâmetros da THE, enquanto cerca de 769 são classificadas como "*reporter*"(relatoras); ou seja, tais institutos providenciaram dados e respostas, mas não satisfizeram critérios de elegibilidade para ganhar um ranking. Os chamados calibradores de performance e a metodologia da pesquisa podem ser encontrados com mais detalhes no link a seguir:

"Methodology for overall and subject rankings for the THE WUR2024"

Nas seções seguintes, serão introduzidos os resultados e metodologias do trabalho, cujo aprendizado reflete as habilidades aprendidas nas matérias de introdução à ciência da computação e visualização e exploração de dados.

2 Desenvolvimento

2.1 Visão Geral Base de Dados

O grupo iniciou o andamento do projeto utilizando o Kaggle para achar a database adequada aos nossos objetivos. Em seguida, iniciou-se um projeto colaborativo no Google Colab, pois este ambiente de computação em nuvem permite ampla manipulação em grupo nos projetos Python. Clique aqui para acessar os códigos

Introdução a DataBase:

Como a base de dados possui dimensões grandes, está disponível abaixo, em formatos html, csv e xlsx para download, consulta e interação, um link-repositório no GitHub temporário criado somente para o trabalho: clique aqui.

O primeiro passo da análise consiste em importar as bibliotecas utilizadas no projeto, elas são encontradas no link acima do colab. Em seguida, realiza-se a leitura do arquivo .csv da data base com a biblioteca *pandas* e atribui-se o nome *df* à base.

O primeiro comando *df.info()* gera um relatório geral e imprescindível ao conhecimento prévio dos dados; visto que há, como nota-se na Tabela 1, um resumo do nome de cada coluna, o número de dados faltantes e os *Dtypes*, tipos de dados que o *pandas* categoriza. Esse passo é essencial para filtrarmos as colunas que não serão utilizadas no decorrer da análise exploratória.

As colunas nomeadas *['url', 'nid', 'disabled','aliases','closed', 'website_url', 'member_level', 'unaccredited']* foram removidas da análise, pois não revelavam informações úteis à completude do trabalho ou produção gráfica, já que seria mais dinâmico trabalhar com a base de dados de maneira simplificada.

Além disso, a coluna *'stats_pc_intl_students'*, que indica o percentual (x%) de estudantes internacionais em cada universidade, foi remodelada a partir de uma função criada que transforma os dados da coluna, antes *"object"*, em *"float64"*. Isso facilitará a criação gráfica de intervalos numéricos futuramente.

Por fim, a nova base de dados um pouco mais limpa foi nomeada como *cleaned_df* e assim será chamada daqui em diante.

Por ser uma coletânea fundamentada, a Database tem poucos (ou quase nulos) dados faltantes em colunas essenciais. Veja que o tratamento de dados faltantes e outras ocorrências, como filtragem e agrupamento, serão feitos no decorrer das análises gráficas e estatísticas, visto que tratá-los de maneira imediata seria um pouco imprevisível, já que o objetivo inicial é conhecer a database apenas.

Column	Non-Null Count	Dtype
rank	2673	obj
name	2673	obj
scores_overall	1904	obj
scores_overall_rank	2673	int64
scores_teaching	1904	float64
scores_teaching_rank	2673	int64
scores_research	1904	float64
scores_research_rank	2673	int64
scores_citations	1904	float64
scores_citations_rank	2673	int64
scores_industry_income	1904	float64
scores_industry_income_rank	2673	int64
scores_intl_outlook	1904	float64
scores_intl_outlook_rank	2673	int64
record_type	2673	obj
member_level	2673	int64
url	2673	obj
nid	2673	int64
location	2673	obj
stats_number_students	2673	obj
stats_student_staff_ratio	2673	float64
stats_pc_intl_students	2673	obj
stats_female_male_ratio	2580	obj
aliases	2673	obj
subjects_offered	2669	obj
closed	2673	bool
unaccredited	2673	bool
disabled	2673	bool
website_url	344	obj

Tabela 1: Descrição de data columns no df não-limpo

2.2 Análise Exploratória dos Dados

Nossa análise exploratória consiste em diversas células no Google colaboratory que descrevem trechos do *cleaned_df*. A seguir, serão apresentados alguns *highlights* do desenvolvimento que serviram como inspiração para a criação da parte gráfica. Note que os códigos criados no Google colab na "Parte 1 - Análise Exploratória" foram feitos como guia mental dos caminhos que o grupo poderia seguir. Não necessariamente todas as informações nela extraídas fizeram parte do resultado que apresentaremos agora:

2.2.1 Os líderes

Rank	Name	Scores Overall	Scores Overall Rank
1	University of Oxford	98.5	10
2	Stanford University	98.0	20
3	Massachusetts Institute of Technology	97.9	30
4	Harvard University	97.8	40
5	University of Cambridge	97.5	50
6	Princeton University	96.9	60
7	California Institute of Technology	96.5	70
8	Imperial College London	95.1	80
9	University of California, Berkeley	94.6	90
10	Yale University	94.2	100

Tabela 2: Top 10 Universidades Mundiais

Na Tabela 2, estão as 10 Universidades consideradas as melhores do mundo pelo ranking. Todas elas obtiveram uma pontuação (*"Scores Overall"*) acima de 94. Isso significa que essas instituições atingiram quase a nota máxima nos métodos avaliativos do WUR2024. Nota-se, também, a ausência de uma pontuação 100, visto que atingir a total perfeição em uma instituição de ensino é algo extremamente difícil, mesmo nos melhores cenários mundiais. Além disso, é interessante que as 10 universidades estão localizadas no Reino Unido e nos Estados Unidos - países com alto desenvolvimento tecnológico e notável renda bruta e renda *per capita*

O código utilizado para criar a tabela acima foi a linha Python:

```
cleaned_df.iloc[0:10, 0:4]
```

Nele, selecionamos o comando iloc, que faz busca por indexação na base de dados. Seleciona-se, no primeiro argumento, as linhas de 1 a 10, e no segundo parâmetro, as colunas de 1 a 4

2.2.2 Universidade de São Paulo

Nessa subseção, faremos uma análise rápida de onde encontra-se a USP no ranking. A motivação por trás dessa parte surgiu devido ao interesse em localizar a USP no cenário mundial, e servirá na leitura de gráficos apresentados na próxima seção.

```
cleaned_df.iloc[237, [0,1,2,5,7,9,11,13]]
```

number	237
rank	201-250
name	University of São Paulo
scores_overall	55.9-58.6
scores_teaching_rank	75
scores_research_rank	82
scores_citations_rank	840
scores_industry_income_rank	477
scores_international_outlook_rank	1031

Tabela 3: Visão USP

A tabela acima serve como importante medidor de excelência da USP. Nela, percebe-se que a Universidade, indexada na posição 237 da DataBase, pertence ao intervalo de posições 201-250, visto que a WUR2024 utiliza esses intervalos de classificação.

Ademais, a pontuação obtida pela USP no ranking foi em média 57.25, o que representa, aproximadamente, apenas 58% do total da pontuação obtida pela "melhor"Universidade do mundo, a *University of Oxford*.

Em contrapartida com o que fora apresentado acima, verifica-se que a USP possui ranking 75 no quesito ensino, e posição 82 no *"scores_research_rank"* ambos bem acima da sua classificação geral. Assim, se a USP possui uma qualidade de ensino muito acima da média e alto número de publicações, o que será que poderia justificar sua classificação abaixo do "normal"?

Logo em seguida, notamos que a coluna *"scores_citations_rank"* está extremamente deslocado para baixo, ocupando posição 840. Apesar de diversificado o número de publicações *"Uspianas"*, tal posição indicaria uma relevância bem abaixo da média para a maioria das publicações, mas ainda é pouca informação e pouco contexto para tirar essas conclusões, outras fontes e análises cuidadosas deveriam ser feitas.

A coluna *"scores_industry_income_rank"* indica uma posição relativamente considerável no assunto renda, visto que o atributo mede o nível de patrimônio médio que os formandos ao nível bacharel acumulam ao decorrer dos anos e a criação de patentes.

Por fim, o nível de internacionalização da USP, medido por *"scores_international_outlook_rank"*, requer uma melhora extrema, já que a instituição está mal classificada (1031ª) nessa categoria. Políticas de incentivo ao estrangeiro e imigrante são, indiscutivelmente, urgentes ao meio acadêmico brasileiro.

2.2.3 Visão Cursos

```
all_subjects = cleaned_df['subjects_offered'].str.split(',').explode()
subject_counts = all_subjects.str.strip().value_counts()
subjects_df = pd.DataFrame({'Subject': subject_counts.index, 'Count': subject_counts.values})
subjects_df = subjects_df.sort_values(by='Count', ascending=False)
subjects_df.head(10)
```

	Subject	Count
0	Computer Science	2277
1	Business & Management	2267
2	Accounting & Finance	1965
3	Biological Sciences	1946
4	Mathematics & Statistics	1914
5	Literature & Linguistics	1913
6	Languages	1913
7	Economics & Econometrics	1891
8	Electrical & Electronic Engineering	1858
10	General Engineering	1769

Tabela 4: 10 Cursos mais oferecidos

A tabela acima é uma contagem bruta dos 10 cursos mais oferecidos no geral de todas as instituições, onde os números abaixo de *"Count"* são a ocorrência de cada curso na database. Veja, que o bacharelado em Ciência da Computação é, segundo o ranking 2024, o curso mais ministrado por essas universidades, com 2277 instituições oferecendo-o.

É marcante, também, o número de instituições que oferecem o curso de Matemática e Estatística, essencial ao desenvolvimento de Data Science e às demandas mundiais.

A tabela 4 e a Seção 2.2.3 foram justamente introduzidas para notarmos que a maioria dos cursos oferecidos a nível superior são da área de exatas e tecnologia, como será evidenciado visualmente na seção de gráficos e imagens.

2.2.4 Medidas Descritivas

	scores_teaching	scores_research	scores_citations	scores_industry_income	scores_international_outlook
count	1904.000000	1904.000000	1904.000000	1904.000000	1904.000000
mean	29.060662	23.416176	52.189706	47.057405	49.907143
std	13.967201	16.697923	25.071342	26.150454	21.828249
min	9.400000	4.600000	3.400000	15.600000	16.100000
25%	19.400000	11.800000	30.400000	22.275000	31.975000
50%	25.800000	17.250000	52.400000	41.050000	45.300000
75%	34.500000	29.900000	73.225000	68.200000	65.400000
max	99.000000	100.000000	99.700000	100.000000	98.800000

Tabela 5: Descrição dos Scores

Em estatística, medidas de tendência central, como a média, moda e mediana, e medidas de dispersão, como desvio padrão e variância, e quantis são de extrema importância para análise.

O intuito da tabela acima é analisar como variam as pontuações (*"scores"*) da maioria das universidades do ranking; as principais instituições globais de ensino segundo o WUR2024. Uma maneira fácil e precisa de avaliar essas medidas é pelo uso de boxplots, que serão introduzidos nessa categoria de pontuações na próxima seção

A tabela 5 fora criada com o comando *".describe()"* e, por default, a tabela com as medidas é feita eliminando-se automaticamente todos os dados faltantes das colunas. Veja que na linha *"count"* todas as categorias de pontuação estão em 1904, número de universidades classificadas com dados no ranking geral, como fora visto nas informações da database no começo da análise.

Porém, de modo geral, nota-se que a qualidade de ensino e pesquisa estão incrivelmente desqualificadas numa análise média das Universidades, obtendo pontuações, respectivamente, de 29.06 e 23.42 de um total de 100.0. Já as categorias, citações, renda e mercado industrial e internacionalização, giram em torno de 49.72 pontos, ainda desejável quanto a qualidade, entretanto próximas dos 50 pontos

O desvio padrão (*"std"*), expressão que mede dispersão entre os dados, é relativamente considerável em todas as categorias, o que pode indicar uma heterogeneidade nas amostras coletadas. Essa informação já era esperada, dado que as Universidades encontram-se em países de distintos níveis de desenvolvimento, fato que oscila a pontuação universitária obtida categoricamente.

As análises dos quantis será feita em breve pelos boxplots, já que facilitam a visualização interquartil.

Percentual de países

A tabela 6 abaixo mostra, percentualmente, os 11 países com maiores incidências na database. Observe que isso não indica o número de países com maior quantidade de universidades no mundo, mas sim no escopo da pesquisa. Nota-se o Brasil com 2.51% do total

Países	Percentagem %
United States	6.546951
Japan	6.322484
United Kingdom	6.098017
India	4.638982
Russian Federation	4.040404
Turkey	3.628881
Pakistan	3.292181
China	3.217359
Algeria	3.217359
Iran	2.805836
Brazil	2.506547

Tabela 6: Países com mais universidades no ranking

A tabela 6 foi feita como motivação da construção do gráfico de barras.

Matriz de correlações

Em nossa última parte de medidas descritivas, realizamos a montagem da matriz de correlações através do comando:

```
cleaned_df.loc[:, col].corr()
```

	ensino	**pesquisa**	**citações**	**indústria**	**internacional**
ensino	1.000000	0.871223	0.527623	0.678078	0.421367
pesquisa	0.871223	1.000000	0.652535	0.786352	0.567576
citações	0.527623	0.652535	1.000000	0.588662	0.624062
indústria	0.678078	0.786352	0.588662	1.000000	0.483999
internacional	0.421367	0.567576	0.624062	0.483999	1.000000

Tabela 7: Matriz de correlação dos *"Scores"*

A matriz de correlação é uma ferramenta imprescindível na observação de como as variáveis se comportam em conjunto. Servem como base na montagem de regressões e visualizações como o heatmap, variando numericamente de -1.0 até 1.0.

Na diagonal principal, os valores são todos 1, já que a correlação entre a própria variável é 1.0. Nas demais entradas da matriz, um número diferente aparece. Se o valor se aproxima de 1.0 e é positivo, dizemos que as variáveis são fortemente correlacionadas e de tendência positiva; ou seja, enquanto uma variável aumenta, a outra tende a aumentar junto e de alguma forma (linear, polinomial, exponencialmente, "etc"). Caso o valor se aproxime de -1.0, vale o caminho contrário do que fora mencionado anteriormente, e as variáveis são inversamente proporcionais.

A correlação entre cada variável será aprofundada e interpretada na análise do heatmap

3 Resultados

3.1 Considerações Iniciais

Na seção de Resultados, serão apresentados gráficos e imagens que compõem uma visualização essencial na leitura e interpretação da base de dados. A motivação por trás da produção de charts e plots com ferramentas Python vem do fato de que tabelas não são suficientes para uma interpretação completa e certas nuances só são observáveis com figuras. A seguir, tentaremos extrair o máximo de informações possíveis de cada tópico apresentado anteriormente na seção de análise exploratória e de assuntos inéditos. Ou seja, fortalecer e assegurar hipóteses antes vistas nas tabelas. As subseções serão agrupadas por tipo de gráfico e todas as imagens, quando necessárias, estarão referenciadas às tabelas.

Infelizmente, algumas análises visuais são muito grandes para o formato pdf e por isso precisam ser reduzidas em números de observações gerais para formatação no documento. Entretanto, podem ser vistas por completo no colab do grupo.

3.2 Gráficos de Barra

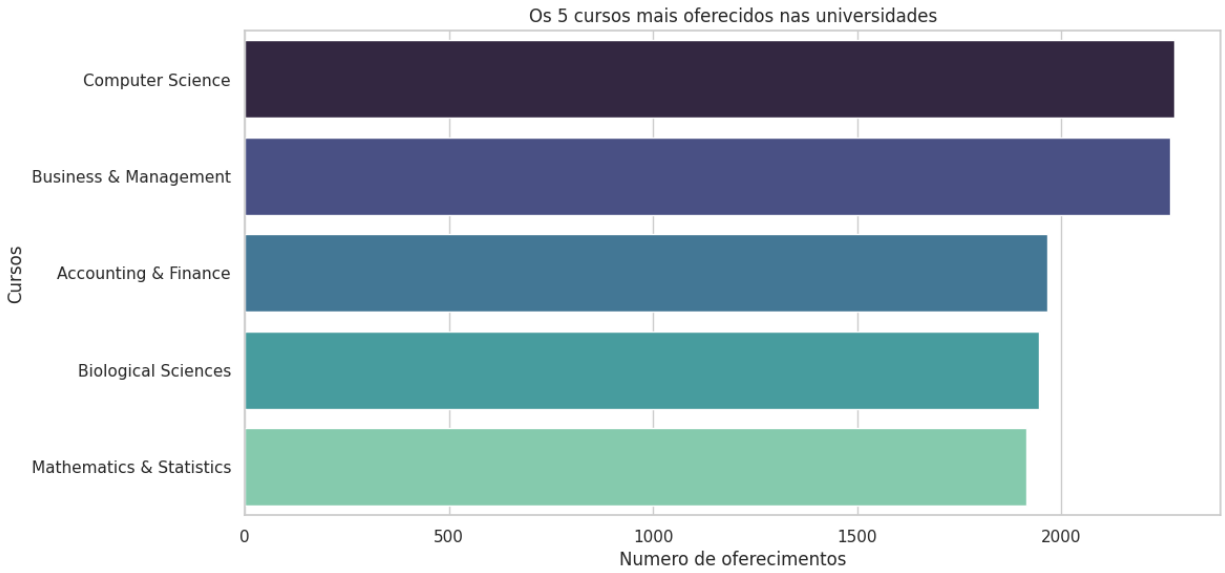

Figura 1: Os 5 cursos mais oferecidos nas universidades

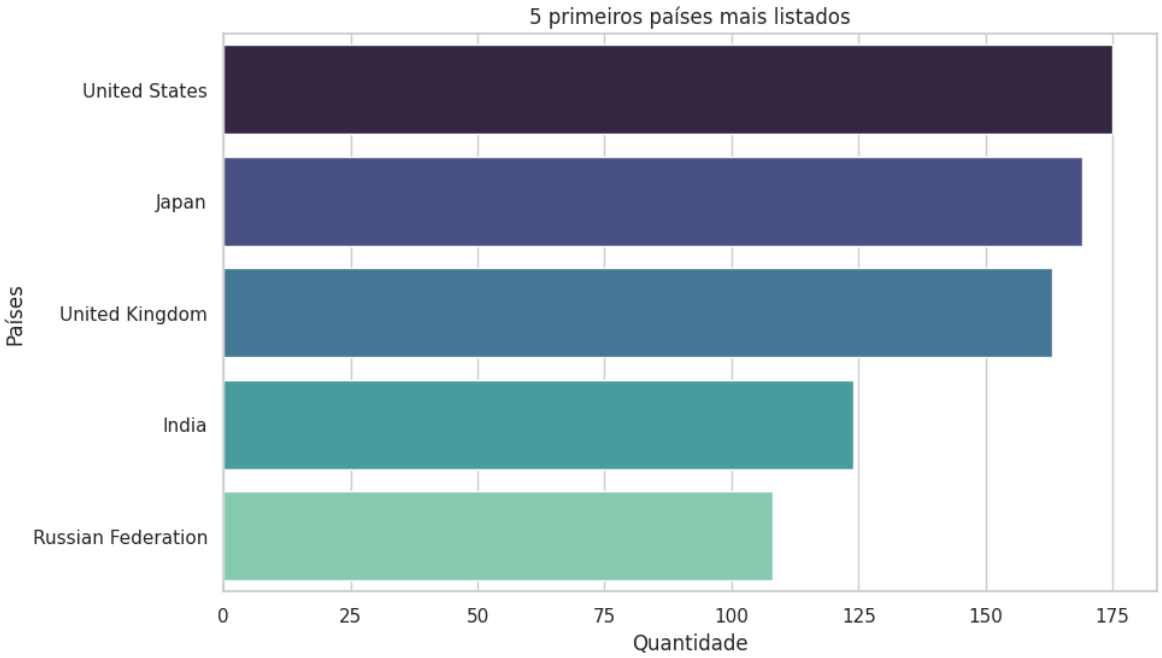

Figura 2: Os 5 países com maior incidência na WUR2024

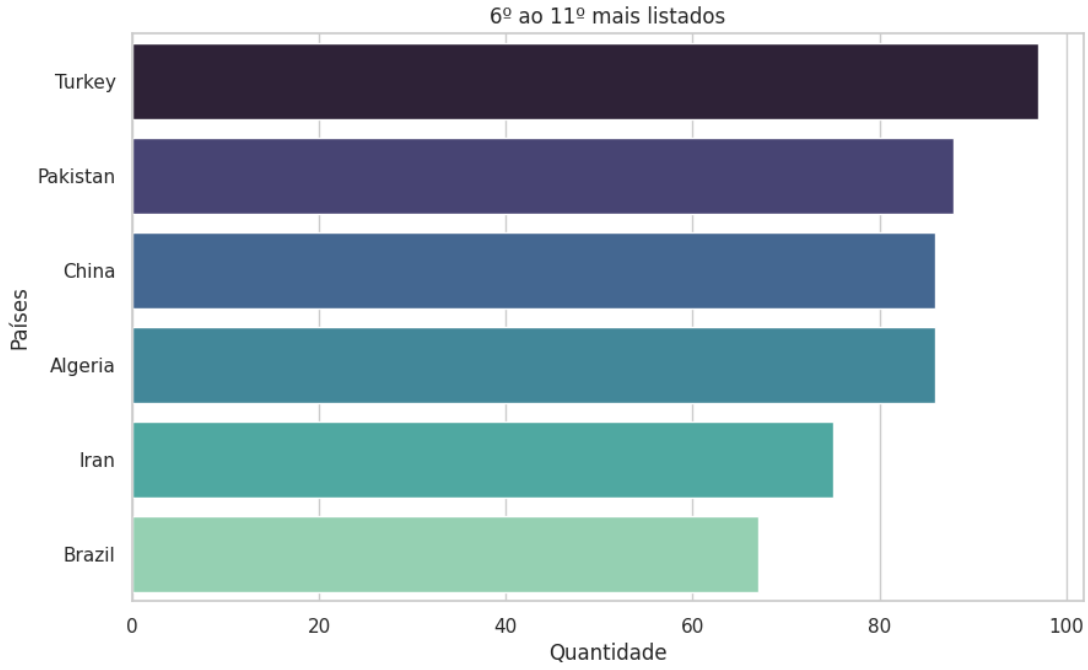

Figura 3: sexto ao décimo-primeiro país mais listado na WUR2024

Iniciaremos nossa análise gráfica com visualizações simples: os gráficos de barra.

Na figura 1, foi feita a exposição gráfica dos 5 cursos mais oferecidos pelas universidades. O número de oferecimentos no eixo x e os nomes de cursos no eixo y são a síntese perfeita de uma variável categórica e uma numérica simples representadas conjuntamente. O intuito da figura 1 é evidenciar, como visto na subseção dos cursos, que grande parte dos bacharelados procurados são da área de exatas e tecnologia para suprir as demandas globais de desenvolvimento. Nota-se no gráfico que aproximadamente 80% dos 5 primeiros cursos mais oferecidos são da área de exatas.

Já a figura 2 e a figura 3 representam o mesmo assunto abordado na tabela 6. Entretanto, o eixo x agora representa a frequência absoluta (número bruto) de universidades por país.

Na liderança da tabela, estão os Estados Unidos, com exatas 175 universidades na pesquisa. Tal frequência é justificada pelo fato de que o país é um grande produtor de graduandos e mão de obra devido à alta população. Destaca-se, também, o Brasil na 11ª posição, com 67 universidades no ranking, apesar das classificações gerais das universidades brasileiras não ter sido a mais alta dentre as participantes.

Os gráficos de barras, de qualquer forma, são visualizações simples, mas tiveram sua importância na compreensão da database do grupo.

3.3 Boxplots, Histogramas e Outras Visualizações

Boxplots são incríveis representações de dados por sua completude visual e diversidade de informação. Geralmente são usados para visualização de variáveis numéricas e categóricas ordinais (hierarquizadas) em sua maioria, ou senão, representam categóricas nominais. Também são excelentes para retratar distribuições que não são bem comportadas pela média (não-normais), e por isso, histogramas e plots de densidade foram estudados.

A motivação por trás da construção de boxplots pelo grupo vem da análise das medidas descritivas em observar tendências nos *"scores"* (pontuações) e, possivelmente, ver as simetrias das distribuições das pontuações.

Definidas as motivações, utilizamos os 11 países que mais aparecem na pesquisa (como visto na seção 3.2) e construímos 6 boxplots no Google colab; um para cada score. Porém, serão apresentados somente alguns dos resultados a seguir:

3.3.1 *Scores*

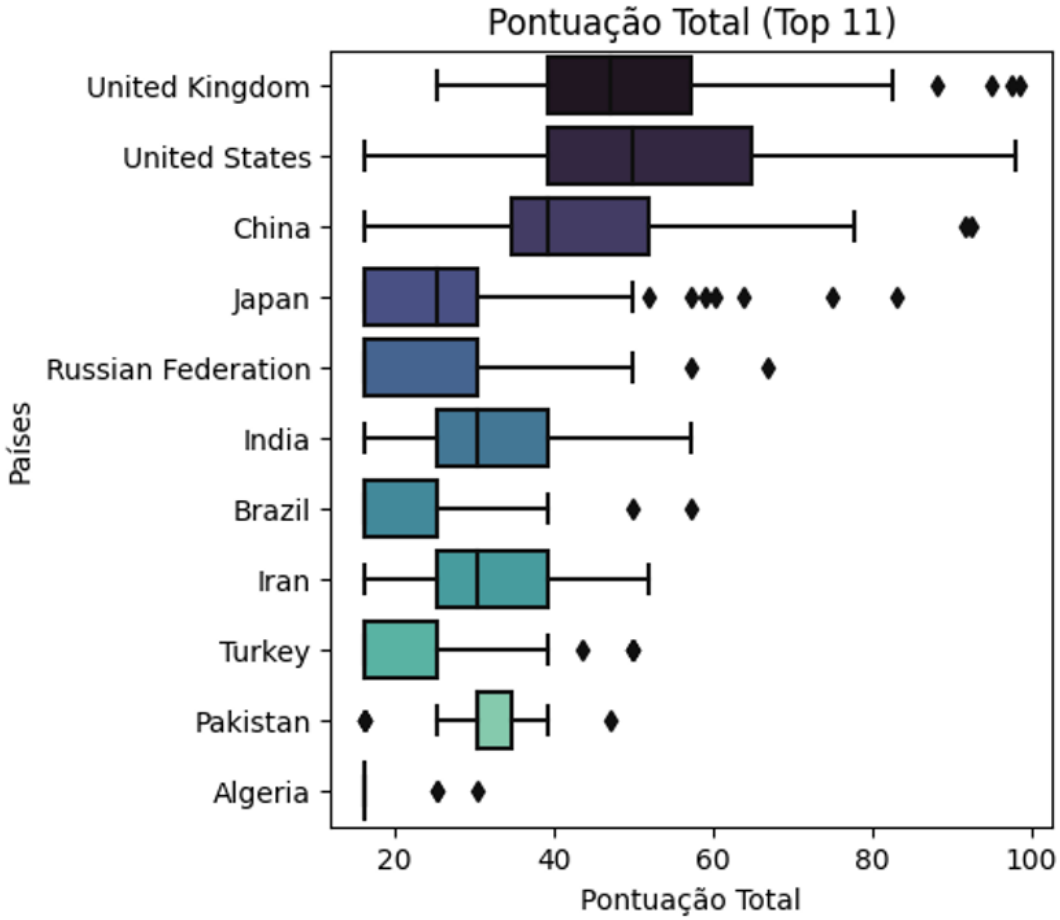

Figura 4: Boxplot da pontuação total de cada país

Na figura 4, nota-se que a pontuação total obtida por cada universidade de cada país é diversa. Em países como Reino Unido e Estados Unidos, as caixas dos boxplots estão situadas, aproximadamente e em média (do 1º ao 3º quartil), entre 40 a 60 pontos. Isso significa que basicamente 50% das universidades dos 2 países tem pontuações naquele intervalo predominantemente. Já as caudas longas indicam uma dispersão considerável nas observações coletadas, ou seja, a diferença entre os limites superiores e inferiores, a chamada amplitude, é alta. Esse fato pode eventualmente indicar uma fácil ruptura da pontuação média do ranking das universidades inglesas e americanas; já que valores muito afastados daqueles centrais podem facilmente influenciar o valor final da média. Porém, o intervalo interquartil (IQR) - definido como a diferença entre o 3º e 1º quartil - é uma

estatística mais robusta do que a amplitude para medição de dispersão, já que não sofre influência de outliers; e, diferentemente do desvio padrão, o IQR considera a organização dos dados em rol crescente.

Outra inferência importante dos 2 boxplots é a mediana, a linha dentro da caixa do boxplot. A mediana indica o valor numérico que divide os dados na metade, ou seja, a partir dela, 50% dos dados estão concentrados acima e os outros 50% abaixo. Um resultado interessante dos boxplots é procurar por caixas em que a linha da mediana é centralizada no retângulo. Isso pode indicar uma distribuição simétrica, onde a mediana, a moda e a média assumem valores iguais, mesmo a moda e a média não sendo explicitamente observáveis em boxplots. Distribuições em que a mediana é igual ou muito próxima à média podem ser distribuições aproximadamente normais e isso servirá de motivação à construção dos gráficos de densidade apresentados em breve. Além disso, caso a mediana esteja próxima ao $1^{\underline{o}}$ quartil, os dados são assimétricos positivos; e caso esteja próxima ao $3^{\underline{o}}$ quartil, os dados estão potencialmente distribuídos de maneira assimétrica negativa.

Ainda observando a figura 4, nota-se que a Argélia é o país em que quase todas partes do boxplot (à exceção dos outliers) são coincidentes, indício forte de que os dados devem seguir uma distribuição gaussiana. Na seção de testes estatísticos que será introduzida em breve, testaremos hipóteses e iremos analisar normalidade para certos pontos da DataBase.

Os outliers, pontos discrepantes fora das caudas, indicam valores que se afastam significativamente dos outros pontos amostrais. Na figura 4, notam-se muitas universidades com outliers no ranking, talvez provenientes das amostras heterogêneas de cada país, visto que no Brasil, por exemplo, a universidade com melhor colocação, a USP, está pontuada muito acima das outras 66 universidades brasileiras pesquisadas.

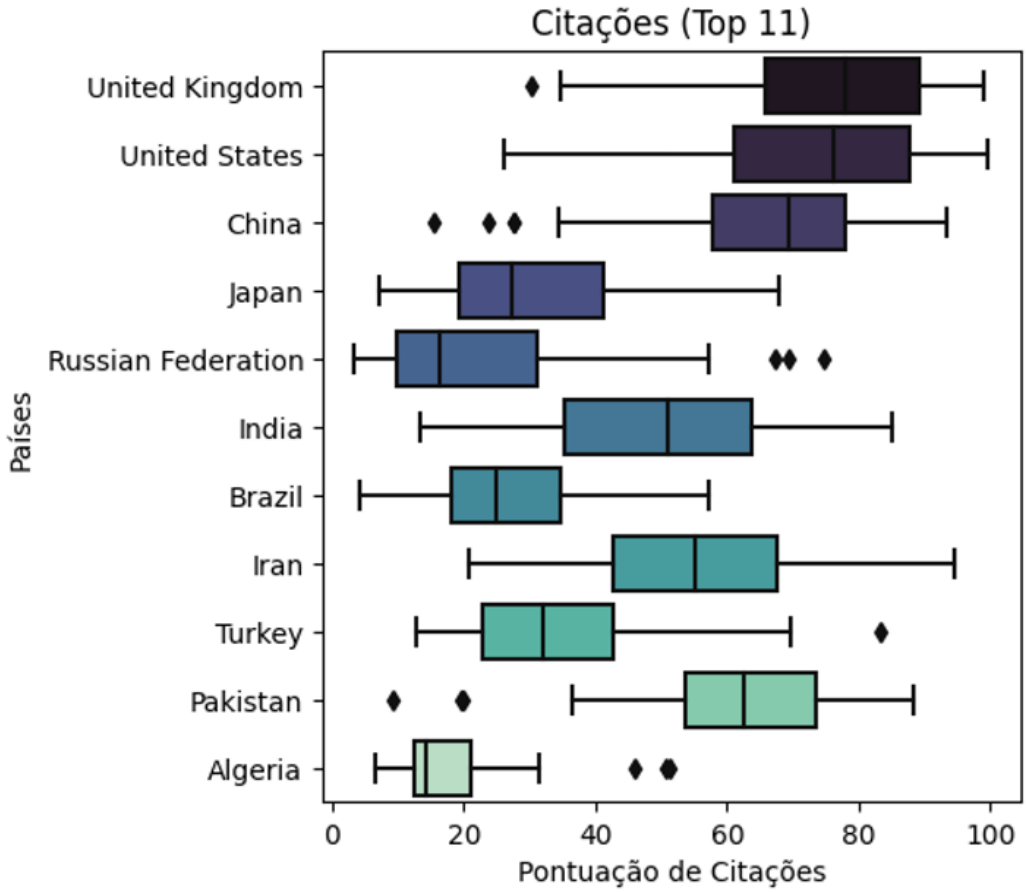

Figura 5: Boxplot das citações entre os 11 países considerados

Na figura 5, a interpretação dos dados segue a mesma lógica do que fora apresentado para a figura 4, mas o foco está na pontuação de citações em pesquisas.

Resumidamente, observa-se uma heterogeneidade na maioria dos dados de cada país. Reino Unido, Estados Unidos e China se destacam em relevância nas pontuações, já que suas caixas estão deslocadas em direção às maiores pontuações. No Brasil, percebe-se que 75% das pontuações em citações encontram-se, aproximadamente, abaixo dos 35 pontos, sendo o maior valor alcançado em torno de 60 pontos. Fato este preocupante quanto à relevância dos artigos científicos brasileiros no meio acadêmico internacional.

Porém, a maior motivação à menção do boxplot acima, após visualizarmos os 6 boxplots gerados pelos *"scores"*, fundamenta-se na possível hipótese de que as citações, em média, têm pontuações maiores do que as outras categorias de todas as universidades. Ou seja, as citações em pesquisas e artigos científicos para cada

universidade costumam ser mais bem avaliadas do que as outras classificações. Vale ressaltar que nas figuras 4 e 5 estudamos uma amostra pequena dentro da população total de universidades, mas os testes estatísticos serão feitos com todas as 2673 amostras (possivelmente filtradas).

A possível conclusão mencionada no parágrafo antecessor vem do fato de que na figura 5, a maioria das caixas encontra-se mais "à direita"do que nas outras pontuações. Na seção de testes estatísticos, essa hipótese será avaliada e observaremos se as distribuições de citações são realmente próximas de uma Gaussiana.

Gráfico de Violino

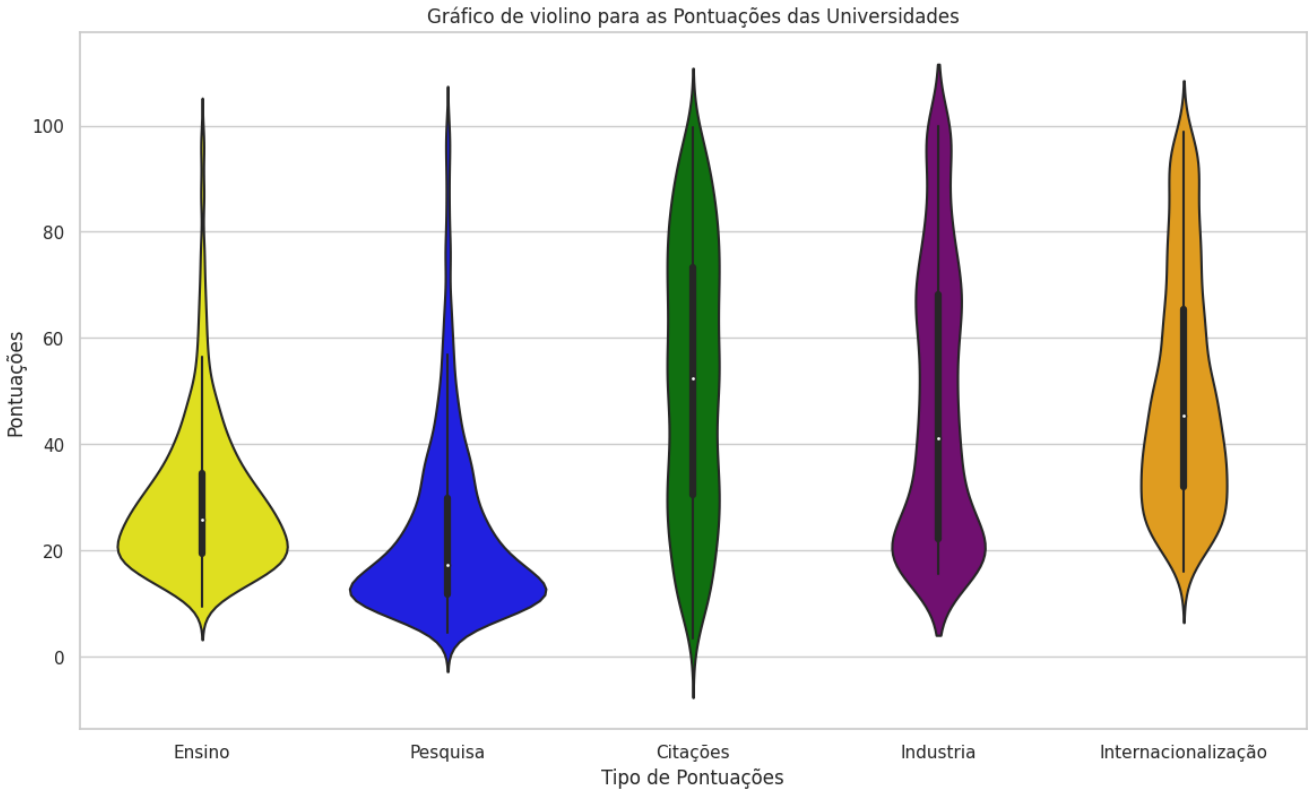

Figura 6: Gráfico de violino para todos *"scores"*

O gráfico de violino é uma visualização interessante, já que consolida features do boxplot com plot de densidade. Na figura 6, as linhas dentro da área de cada violino é a representação do boxplot. Em destaque (parte mais grossa), está a caixa do box plot com os 3 quartis, e em seguida, nas linhas contínuas e mais finas, os limites. A extensão de área colorida logo após os limites de cada gráfico são os outliers, então quando mais "agudos"os extremos, maior o número de outliers.

Logicamente, as áreas maiores (mais "inchadas"), apresentam os intervalos com maior concentração de dados, já as áreas menores (mais finas), indicam menor concentração. Assim, é possível estudar como variáveis numéricas se comportam dentro de classes.

A vantagem do violin plot é condensar visualizações, o que poderia eventualmente dispensar a criação de histogramas ou gráficos de densidade, por exemplo. Mas logo abaixo, introduziremos um plot de densidade para confirmar o que fora relatado na figura 6.

Com isso, observa-se que o gráfico de violino acima fora feito levando em conta todas as 2673 universidades da WUR2024 e serviu como ferramenta auxiliar na comparação de resultados de densidade do grupo.

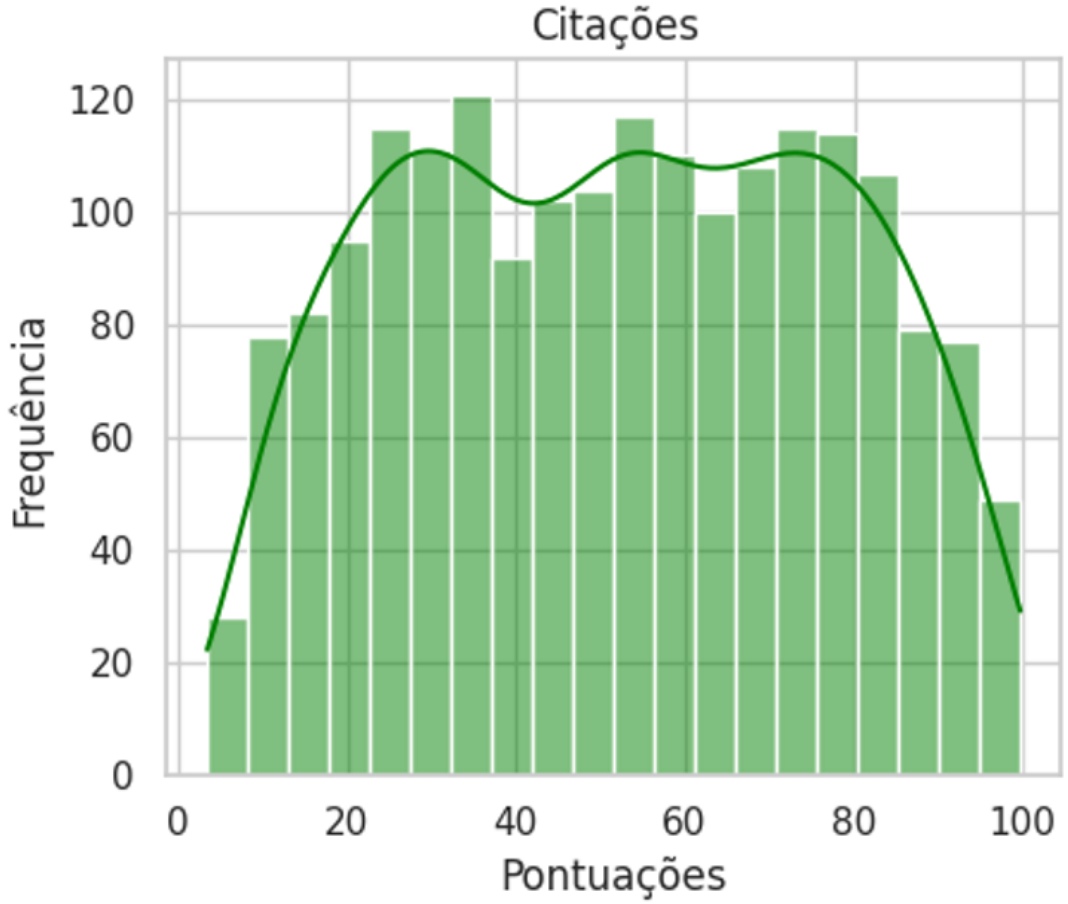

Figura 7: Plot de densidade e histograma para a categoria de citações acadêmicas

A partir desse ponto, é notável o foco do grupo na categoria de citações acadêmicas, dada que nossa hipótese na seção de testes estatísticos será em torno da categoria.

Visualmente, a figura 7 é um histograma, perfeito para retratar variáveis numéricas contínuas (assumem qualquer valor real intervalar) e útil para estudar como as citações se distribuem ao longo das universidades. É notório que a moda das pontuações é de 35 pontos, ou seja, a maioria das universidades alcançou 35 pontos na categoria; mas mesmo assim, a dispersão de pontos amostrais parece ser homogênea e relativamente consistente ao longo dos intervalos construídos.

Agora iniciaremos a discussão de outras abordagens e na próxima seção voltamos com os resultados a respeito do teste de hipóteses para citações.

Heatmap: Matriz de Correlações

Por fim da subseção de *"scores"*, apresentaremos uma releitura rápida da matriz de correlações da parte medidas descritivas, o gráfico de calor (ver figura 8 abaixo). Porém, a diferença é que dessa vez restringiremos nossa população e coletaremos uma amostra de todas as universidades dos 11 países que mais foram citados na WUR2024.

Visualizando-se a figura 8 pela primeira vez, notam-se tons mais escuros de azul (correlação mais baixa) em basicamente duas comparações:

- internacionalização X ensino: 0.40
- internacionalização X indústria: 0.45

Ambos itens refletem possivelmente o porquê a internacionalização possui as menores correlações, já que mentes estrangeiras numa universidade, por mais que enriqueçam imensamente o pensamento acadêmico, não necessariamente geram maior riqueza e inserção industrial no mercado ou afetam a qualidade de ensino das instituições. Como o heatmap fora construído levando em conta países que costumam receber poucos imigrantes, talvez o impacto da mão de obra estrangeira formada em universidades nacionais não se reflita economicamente nos países.

De qualquer forma, do que fora explicado no parágrafo acima, vale relembrar a frase *correlação não implica causa*, ou seja, é um erro deduzir causa e efeito puramente da correlação entre duas variáveis; e, portanto, as conclusões anteriores podem ser falhas. Porém, notar que a razão por trás de um argumento é falha não necessariamente indica que a conclusão é falsa sempre, por isso testes estatísticos são necessários.

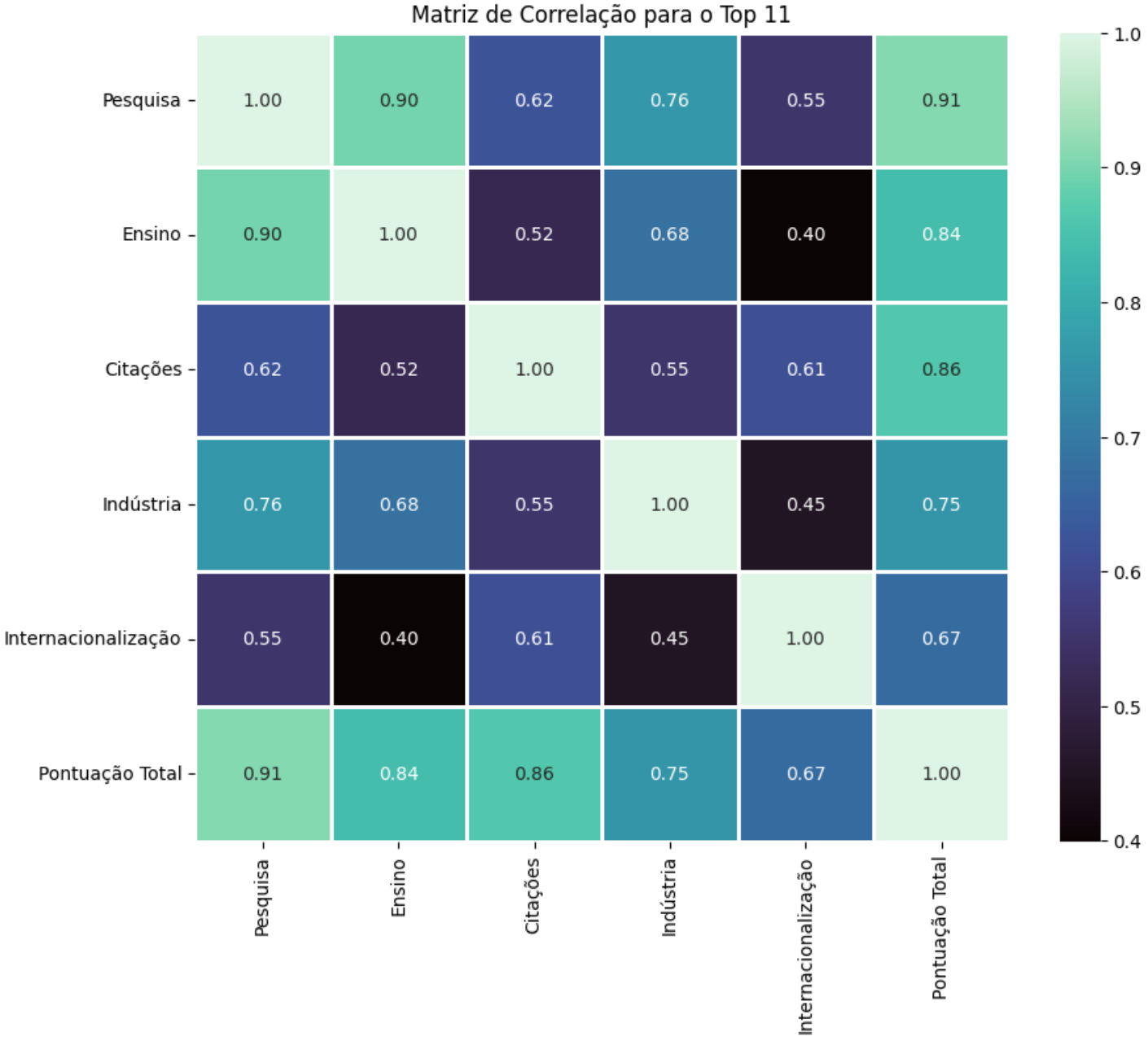

Figura 8: Matriz de correlações dos scores em forma de heatmap

3.3.2 Análise – sexo

Scatter plot inicial

Nessa parte de análises, iremos avaliar a proporção do número de homens e mulheres nas universidades.

Basicamente, a motivação por trás dos plots de análise de sexo está fundamentada na coluna *"stats_female_male_ratio "* da base de dados. Um dos desafios encontrados nessa parte era retrabalhar a coluna, já que os dados estavam em forma de *"object"* do jeito *y:x*; proporção de mulheres para homens. Assim, precisaríamos tratar os dados da coluna como *floats*, já que esse tipo de variável permitiria a construção de análises numéricas.

Com algumas sintaxes, localizamos os caracteres *":"* e quebramo-os de cada linha, separando os valores que vêm antes dele dos valores que vêm depois. Então, cada valor foi atribuído às novas colunas no dataframe, separadamente.

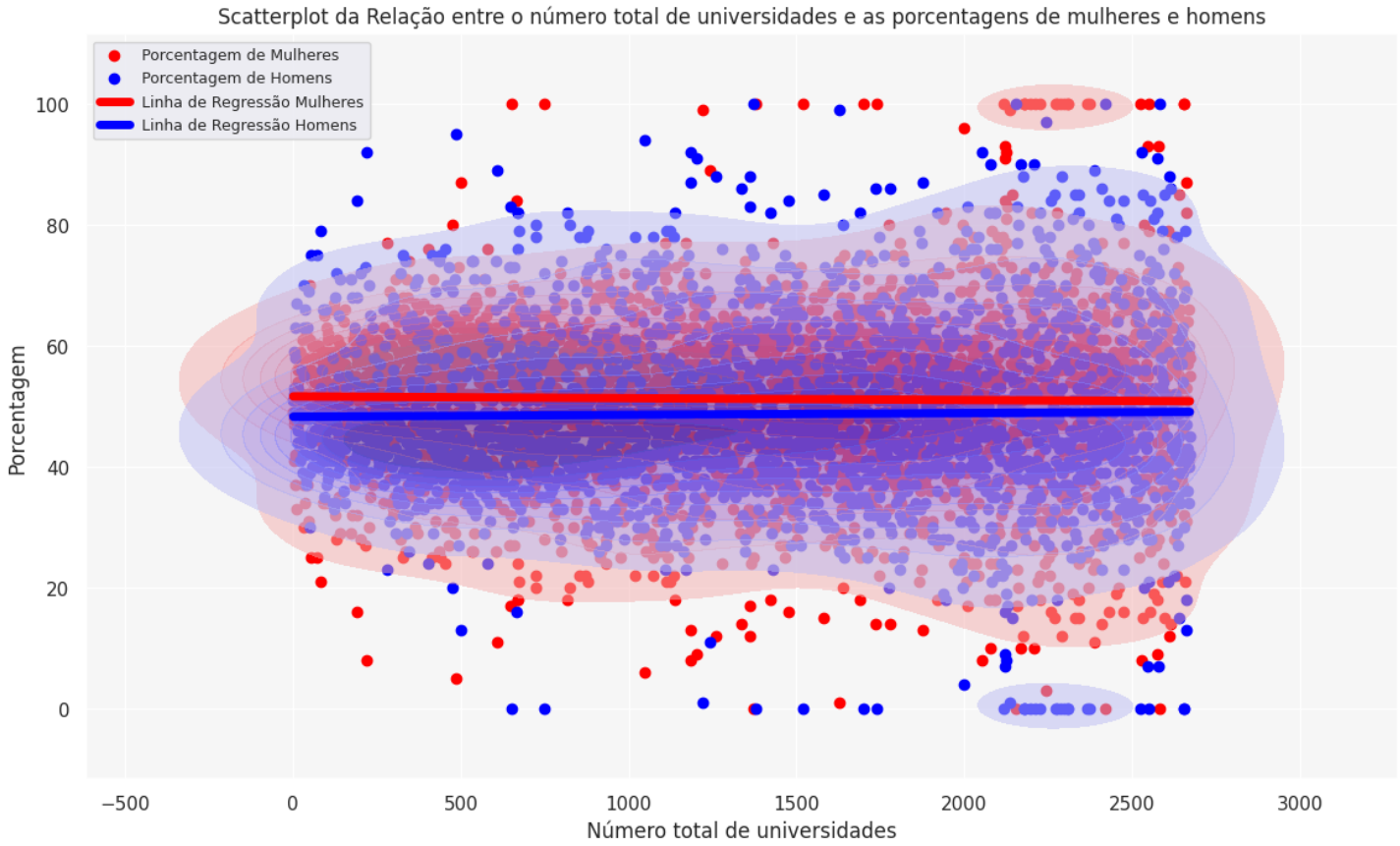

Figura 9: scatter plot da proporção entre homens e mulheres

Feita a manipulação da coluna, criamos o primeiro plot de dispersão. Nele, acrescentamos as linhas de contorno (uma técnica chamada *"Kernel Density Estimation- KDE*) e duas regressões lineares simples.

Os milhares de pontos dispersos na figura 9 formam o scatter plot e indicam um alto indício de concentração de dados ao longo de todas as universidades, ou seja, a tendência observada é de que a proporção entre homens (azul) na maioria das universidades, é parecida (homogênea). O mesmo é válido para mulheres.

As linhas de contorno no gráfico (KDE) servem para observar agrupadores numa amostra, o que é visto pela densidade dos contornos na figura 9. Essas linhas conectam pontos de igual densidade, revelando áreas em que os dados estão mais concentrados ou mais dispersos. Veja que as cores e os contornos vermelhos e azuis estão mais densos (sobrepostos) no centro do gráfico.

A consistência dos dados é observada na quase que constante linha de regressão para ambos os sexos. A discreta inclinação negativa da linha vermelha poderia indicar que as primeiras universidades colocadas na lista são predominantemente de público feminino, mas existe uma tendência dessa predominância diminuir quando o número de universidades aumenta no eixo x (universidades piores rankeadas) por exemplo. Porém, devido à considerável altura vertical entre os pontos amostrais vermelhos e a linha vermelha, muito provavelmente não há nenhuma relação entre a porcentagem de mulheres em todas as universidades; já que o fato de uma universidade na Arábia Saudita ter menos mulheres não influenciaria a proporção do gênero feminino na USP, por instância. Comprovações como essa não serão o foco desse relatório.

Histograma de frequência absoluta

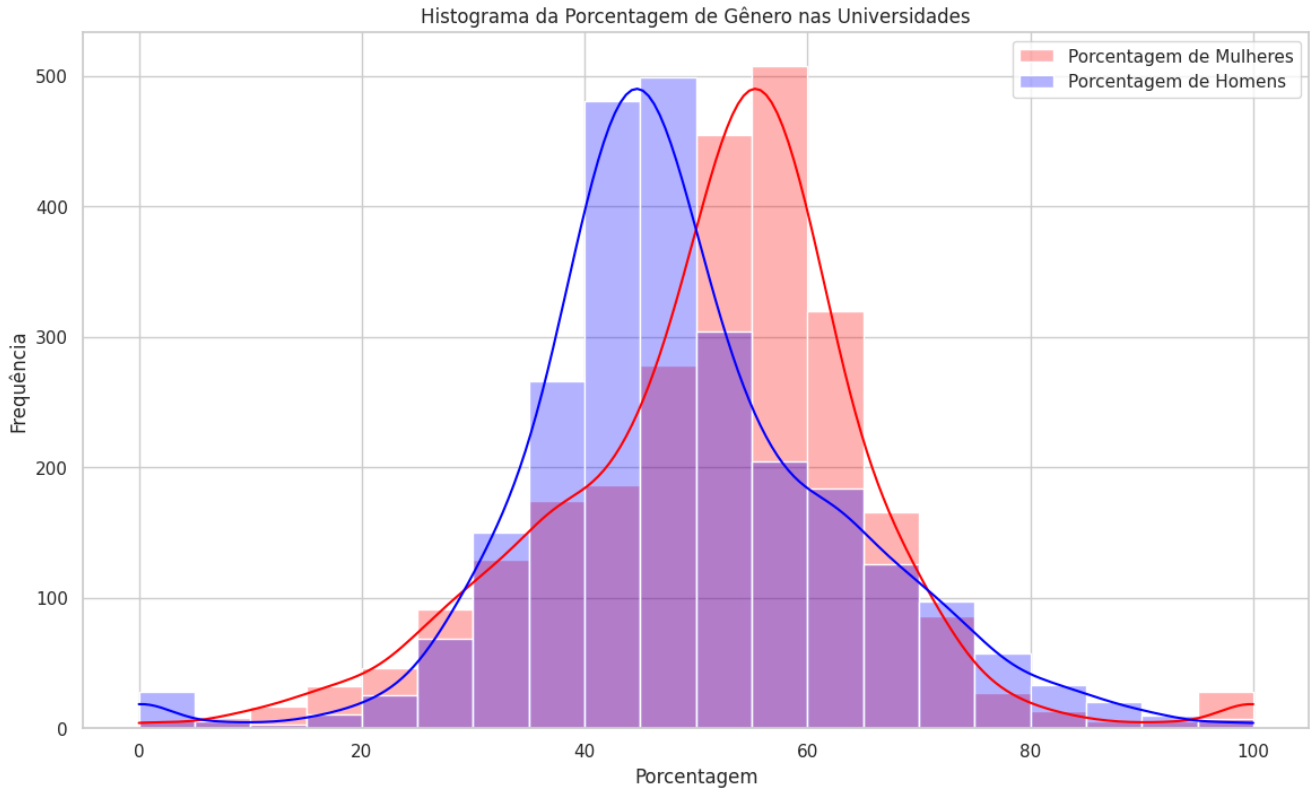

Figura 10: Histogramas

Os histogramas acima, adicionados às linhas de densidade, revelam que as amostras de sexo masculino e feminino para todas as universidades são unimodais, com modas, respectivamente, de 45% e 55% visto que há somente um pico para cada histograma. Nas barras azuis, 500 universidades têm entre 45% e 50% de homens estudantes. Por outro lado, um pouco mais de 500 universidades têm de 55 a 60% de mulheres estudantes.

Outra inferência interessante é o histograma da porcentagem de mulheres mais à direita do que o histograma de homens. Esse fato indica que as estatísticas de tendência central na porcentagem de mulheres em meios universitários são maiores do que aquelas dos homens. É importante notar que como não temos certeza que as distribuições são completamente Gaussianas, inferir média, moda e mediana diretamente da figura 9 não seria uma tarefa extremamente precisa, já que somente mediante de cálculos os histogramas fornecem medidas centrais **aproximadas**.

3.4 Visualizações USP

A última subseção das análises gráficas será dedicada puramente à Universidade de São Paulo e complementará a parte da análise exploratória. Assim, o grupo criou *radar plots*:

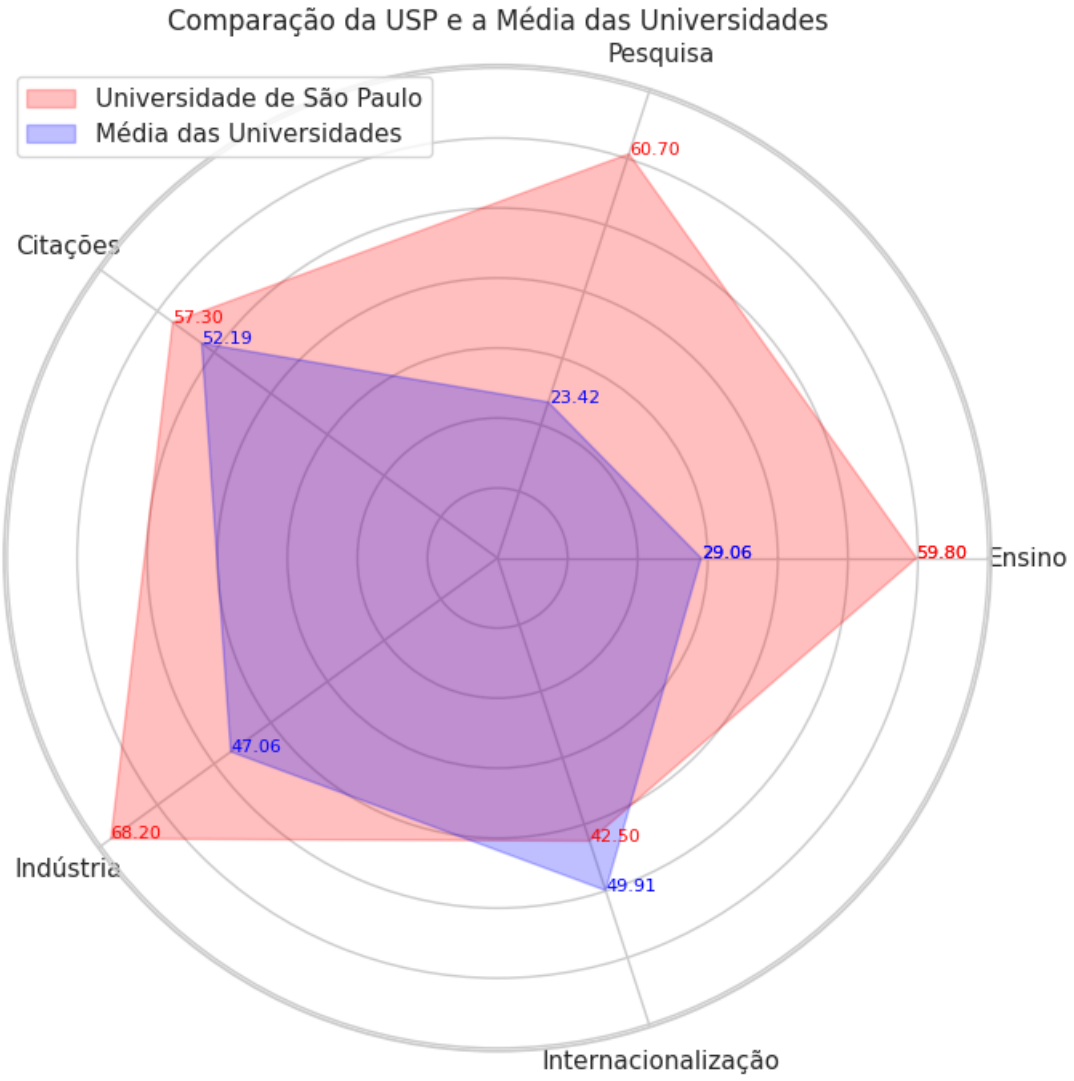

Figura 11: Gráfico de radar

Uma utilidade dos gráficos de radar está na supervisão de qualidade, permitindo a visualização das métricas de desempenho da USP perante a média global universitária. Também, o centro é 0% e quanto mais distante as extremidades das

linhas estiverem do centro, maior será o percentual atingido na categoria, e por isso, os gráficos de radar são úteis para retratar observações multivariadas. No caso da figura 11, estamos analisando desempenho médio da universidade com 5 características diferentes.

É intuitivo do gráfico observar que em todas as categorias, à exceção da internacionalização, a USP apresenta desempenho melhor que a média de todas as universidades da pesquisa.

Com essa breve conclusão, a USP, de fato, necessita ampliar suas políticas de apoio à internacionalização.

4 Testes Estatísticos

4.1 Introdução

Durante a realização do trabalho, o grupo percebeu por gráficos e tabelas que, em média, a pontuação na relevância das citações acadêmicas são mais altas do que as pontuações das outras categorias para a maioria das Universidades. Com isso, apresentaremos testes estatísticos que realmente comprovem, ou pelo menos fortaleçam, nossa hipótese. Seguimos alguns passos vistos em sala de aula com o Professor Leo e, a partir deles, tentamos reproduzir com nossas análises.

4.2 Hipótese

As pontuações de citações, em média, são maiores do que as demais pontuações avaliativas

Na primeira etapa de código e a mais longa, fizeram-se as configurações do teste estatístico. Foi utilizada como base a comparação feita pelo professor em sala de aula para os Pokémons. Dessa forma, nessa primeira etapa, definimos os 3 tipos de teste de hipótese, além do nível de significância de 5%.

Adicionalmente, foram feitas limpezas e configurações dos dados, criando listas para cada uma das colunas de interesse e removendo os dados ausentes. Além disso, foi feita a contagem da composição de cada coluna para que, posteriormente, pudéssemos calcular o valor de *"ratio"* no poder estatístico.

A partir desses valores, foi possível aplicar o teste de poder estatístico. O valor de *"ratio"* é 1 para qualquer um dos casos de comparação entre dois grupos, já que todos os grupos têm o número de observações e sempre iguais a 1904. O valor *"alpha"*, como dito anteriormente, é de 0.05. Por fim, o *"effect_size"*, que mede a magnitude padrão da diferença entre as médias de dois grupos, foi atribuído o valor 0.5, que por padrão, indicaria uma magnitude moderada.

```
power_analysis = statsmodelspower.TTestIndPower()
power = power_analysis.power(effect_size=0.5, ratio=1,
    alpha=alpha, nobs1=1904, alternative="larger")
print(power)
    1
```

Obtivemos o valor 1 para o poder estatístico, e suspeitamos de algum erro. Visto que, embora seja conceitualmente possível um poder estatístico 1, na prática, é raro alcançar esse valor devido às complexidades e variabilidades dos dados observados.

Um dos possíveis motivos para encontrarmos 1 no valor do teste de poder estatístico é que o tamanho da amostra é grande o suficiente para detectar até mesmo efeitos pequenos com uma probabilidade extremamente alta.

Uma alternativa para obtermos um valor mais real, na prática, seria diminuir o tamanho do efeito. Isso indicaria uma diferença ou efeito menor entre os grupos, ou variáveis em estudo. Então reduziu-se o valor do *"effect_size"* para 0.2 e notamos um poder estatístico praticamente igual a 1 novamente.

Posteriormente, seguimos com o teste de hipótese. Nesse caso, queremos mostrar que, em média, a coluna de citações é numericamente maior que as demais. Há uma distinção fundamental entre métodos paramétricos, aplicados apenas em distribuições gaussianas, e métodos não paramétricos, que não assumem uma distribuição normal. Caso a distribuição seja gaussiana, é mais indicado utilizar testes paramétricos, pois apresentam um desempenho superior. Para verificar se a distribuição é ou não gaussiana, pode-se realizar o teste de Shapiro-Wilk.

Definimos duas variáveis, a *"stat"*, ou melhor, teste estatístico, e *"p"*, o famoso *p-value*. Após aplicado o teste de Shapiro-Wilk, a variável *stat* resultou aproximadamente em 0.9165 e *p-value* em 0. Assim, um indício do teste estatístico numericamente próximo de 1 e um *p-value* nulo, **concluiria** que todos os **dados de *scores* não são Gaussianos**.

Vale a observação de que no começo da análise estatística, o grupo concatenou as colunas dos 5 scores existentes em somente uma, a *"stacked_column"*. Essa operação permitiu a realização do teste de Shapiro para a análise do parágrafo acima.

Devido ao elevado número de observações na *stacked_column*, talvez o valor-p não seja o mais preciso, e por essa razão, o grupo realizou 5 testes individuais dos 5 *scores*, onde os valores-p de cada resultaram em números bem próximos de 0, mas não totalmente nulos.

A observação acima a respeito dos valores-p baixos, ainda pode ser comparada com o valor $alpha = 0.05$:

Caso o valor de p seja inferior a alpha, podemos rejeitar a hipótese de que a distribuição é gaussiana. Portanto, é possível utilizar um teste-t se o valor de p for superior a alfa, enquanto, caso contrário, podemos recorrer ao teste U de Mann-Whitney. Neste caso, refutamos a normalidade e optamos pelo teste U de Mann-Whitney.

Assim, o p-valor de qualquer um dos dois testes retornados vai indicar se rejeitamos ou não nossa **hipótese nula** de que as médias das pontuações de uma primeira coluna selecionada são maiores do que a coluna de citações.

Foram feitos testes comparando as colunas duas a duas; a coluna de citações e alguma outra das 4 colunas de scores restantes.

Sendo assim, os códigos escritos realizam um teste U de Mann-Whitney para

comparar duas amostras, usando um teste de normalidade. Como concluímos através de códigos de que o valor-p é menor que alfa para todas as comparações dois a dois amostras não seguem uma distribuição gaussiana, é realizado um teste U de Mann-Whitney. Os valores-p foram sempre menores que alpha e consideravelmente pequenos, indicando uma diferença significativa entre as amostras, sendo que a coluna de *score* analisada sempre é menor que a coluna de citações, conforme a hipótese alternativa especificada. **Em resumo, rejeita-se a hipótese nula, concluindo com grande chance de certeza que as amostras não vêm da mesma distribuição, e as colunas de pesquisa, ensino, indústria e internacionalização tem pontuações significativamente menores que a coluna de citações.**

Referências

"Methodology for overall and subject rankings for the THE WUR2024"

"Link Kaggle da DataBase"

"From data to Viz"

"Análise Pokémon Professor Leo"

5 Anexo

Colab de desenvolvimento oficial do grupo

GitHub do trabalho para download de arquivos