RA272746_Tarefa01

April 5, 2024

1 IA376I – Tópicos em Engenharia de Computação VII

1.1 Tópico: Análise de Dados Visual (Visual Analytics)

Professora: Wu, Shin - Ting Aluno: Luiz Roberto Albano Junior RA: 272746 Tarefa 1 - 05/04/2024

Base de dados utilizada: #### World University Rankings 2023 Url: https://www.kaggle.com/datasets/alitaqi000/world-university-rankings-2023

Pergunta: Como as universidades de diferentes países se comparam em termos de classificação mundial e quais são os principais fatores que contribuem para suas classificações?

Importação dos dados

[2]:
```
import pandas as pd
import numpy as np
from plotnine import *

#carregamento do arquivo de dados
university_ranking = pd.read_csv("World-University-Rankings-2023.csv")
#university_ranking.head()
university_ranking
```

[2]:

	University Rank	Name of University	Location	\
0	1	University of Oxford	United Kingdom	
1	2	Harvard University	United States	
2	3	University of Cambridge	United Kingdom	
3	3	Stanford University	United States	
4	5	Massachusetts Institute of Technology	United States	
..	…	…	…	
194	196	University of Erlangen-Nuremberg	Germany	
195	196	Sichuan University	China	
196	198	Durham University	United Kingdom	
197	198	Queen’s University Belfast	United Kingdom	
198	198	University of Reading	United Kingdom	

	No of student	No of student per staff	International Student	\

```
0          20,965                    10.6                    42%
1          21,887                     9.6                    25%
2          20,185                    11.3                    39%
3          16,164                     7.1                    24%
4          11,415                     8.2                    33%
..            …                       …                      …
194        30,303                    43.4                    13%
195        49,543                    15.8                     6%
196        18,425                    14.1                    35%
197        19,060                    15.8                    39%
198        15,720                    16.4                    32%

    Female:Male Ratio  OverAll Score  Teaching Score  Research Score  \
0             48 : 52           96.4            92.3            99.7
1             50 : 50           95.2            94.8            99.0
2             47 : 53           94.8            90.9            99.5
3             46 : 54           94.8            94.2            96.7
4             40 : 60           94.2            90.7            93.6
..               …              …               …               …
194           51 : 49           54.5            44.6            47.5
195               NaN           54.5            57.1            58.6
196           54 : 46           54.4            40.0            44.6
197           57 : 43           54.4            31.1            37.9
198           56 : 44           54.4            36.5            39.6

     Citations Score  Industry Income Score  International Outlook Score
0               99.0                   74.9                         96.2
1               99.3                   49.5                         80.5
2               97.0                   54.2                         95.8
3               99.8                   65.0                         79.8
4               99.8                   90.9                         89.3
..               …                     …                            …
194             68.8                   90.7                         53.5
195             48.6                   93.4                         38.7
196             70.0                   39.4                         94.3
197             84.4                   41.6                         97.4
198             78.5                   42.2                         93.3

[199 rows x 13 columns]
```

Observações importantes sobre a base de dados: - Escolhi esta base por ser a base mais atualizada que encontrei. Outras fontes de dados possuíam dados de mais de 10 anos atrás. - A base possui algumas colunas com dados nulos. Para a coluna Location, pesquisei o país de algumas universidades que estavam sem esta informação

---

1.1.1 Comparação da classificação das universidades agrupadas por país

Neste gráfico desejo representar um comparativo por país da classificação de suas universidades. Para tal visualização estou agrupando as universidades por país e obtendo a médida da avaliação geral para ordenação.

Tratamento dos dados

[3]:
```
rank_by_country = university_ranking[['Name of University', 'Location',␣
 ↪'OverAll Score']]

#Converto os dados da coluna para valores numéricos pois o DataFrame apresenta␣
 ↪valores agrupados ou não numéricos
rank_by_country["OverAll Score"] = pd.to_numeric(rank_by_country["OverAll␣
 ↪Score"], errors='coerce')

#Filtrando os registros, excluindo as colunas NaN
rank_by_country = rank_by_country[rank_by_country['OverAll Score'].notna() &␣
 ↪rank_by_country['Location'].notna()]

#Agrupa as universidades por país e sua média
rank_by_country = rank_by_country.groupby('Location')['OverAll Score'].mean().
 ↪sort_values(ascending=False).reset_index()

#Ordena os valores do Eixo Vertical
countries = rank_by_country.groupby('Location')['OverAll Score'].mean().
 ↪sort_values(ascending=False).reset_index()['Location'].tolist()
countries_cat = pd.Categorical(rank_by_country['Location'],␣
 ↪categories=countries)
rank_by_country = rank_by_country.assign(Country=countries_cat)
```

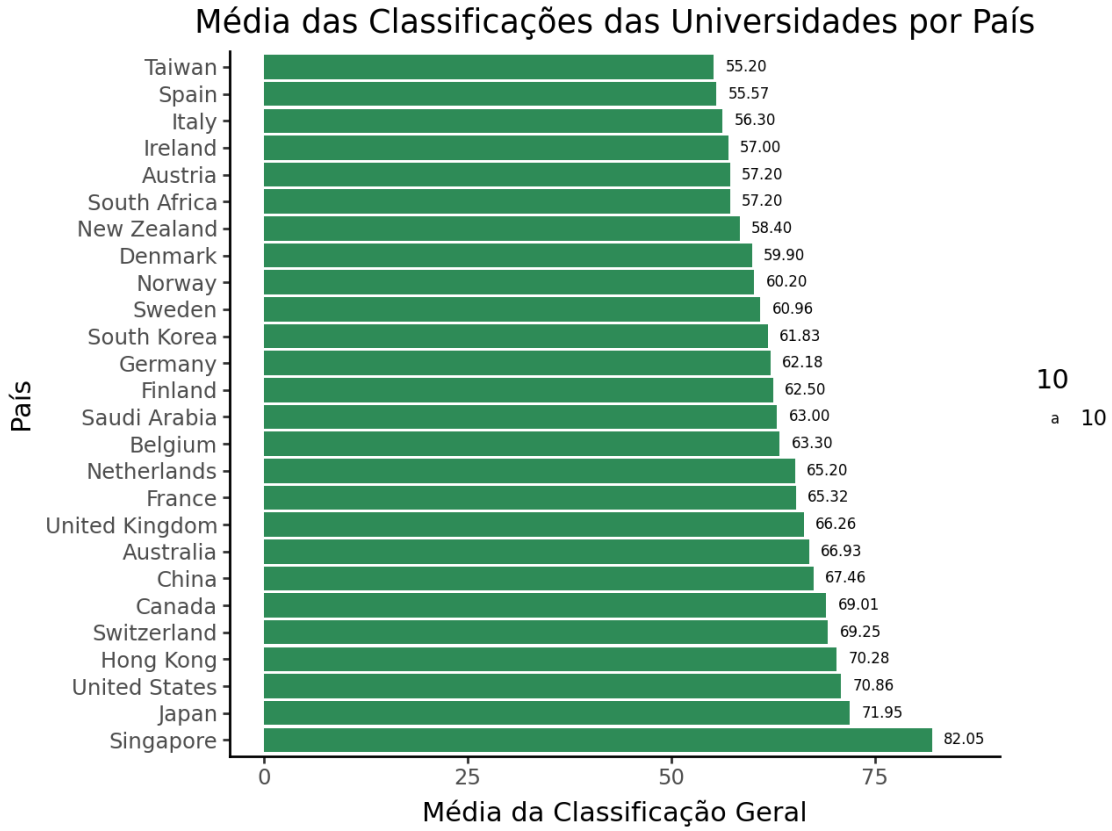
```
C:\Users\junio\AppData\Local\Temp\ipykernel_16980\2137760829.py:4:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

Plotagem do gráfico Gráfico em barras para visualização do comparativo

[4]:
```
(
    ggplot(rank_by_country)
    + aes(x='Country', y='OverAll Score', label='OverAll Score', size=10)
    + geom_bar(stat='identity', fill='seagreen', show_legend=False)
    + geom_text(format_string="{:.2f}", nudge_y=4)
    + theme(axis_text_x=element_text(rotation=45, hjust=1))
    + labs(title='Média das Classificações das Universidades por País',␣
 ↪x='País', y='Média da Classificação Geral')
```

```
    + coord_flip()
    + theme_classic()
)
```



1.1.2 Visualização de coorelações entre parâmetros

Tratamento dos dados

[5]:
```
# Filtrar as colunas necessárias e remover valores nulos
university_reputation = university_ranking[['Name of University', 'Location',␣
 ↪'University Rank', 'OverAll Score', 'International Outlook Score', 'No of␣
 ↪student', 'No of student per staff']].dropna()
university_reputation["International Outlook Score"] = pd.
 ↪to_numeric(university_reputation["International Outlook Score"],␣
 ↪errors='coerce')
university_reputation["OverAll Score"] = pd.
 ↪to_numeric(university_reputation["OverAll Score"], errors='coerce')
university_reputation["University Rank"] = pd.
 ↪to_numeric(university_reputation["University Rank"], errors='coerce')
```

```
university_reputation.head()
```

[5]:

	Name of University	Location	University Rank	OverAll Score	International Outlook Score	No of student	No of student per staff
0	University of Oxford	United Kingdom	1	96.4	96.2	20,965	10.6
1	Harvard University	United States	2	95.2	80.5	21,887	9.6
2	University of Cambridge	United Kingdom	3	94.8	95.8	20,185	11.3
3	Stanford University	United States	3	94.8	79.8	16,164	7.1
4	Massachusetts Institute of Technology	United States	5	94.2	89.3	11,415	8.2

Correlação entre Reputação Acadêmica (OverAll Score) e Classificação Mundial (University Rank) Para facilitar a leitura e entendimento dos fatores, estou limitando o DataFrame para as primeiras 50 universidades.

[6]:
```
university_reputation_50 = university_reputation.iloc[:50]
university_reputation_50
```

[6]:

	Name of University	Location
0	University of Oxford	United Kingdom
1	Harvard University	United States
2	University of Cambridge	United Kingdom
3	Stanford University	United States
4	Massachusetts Institute of Technology	United States
5	California Institute of Technology	United States
6	Princeton University	United States
7	University of California, Berkeley	United States
8	Yale University	United States
9	Imperial College London	United Kingdom
10	Columbia University	United States
11	ETH Zurich	Switzerland
12	The University of Chicago	United States
13	University of Pennsylvania	United States
14	Johns Hopkins University	United States
15	Tsinghua University	China

16	Peking University	China
17	University of Toronto	Canada
18	National University of Singapore	Singapore
19	Cornell University	United States
20	University of California, Los Angeles	United States
21	UCL	United Kingdom
22	University of Michigan-Ann Arbor	United States
23	New York University	United States
24	Duke University	United States
25	Northwestern University	United States
26	University of Washington	United States
27	Carnegie Mellon University	United States
28	University of Edinburgh	United Kingdom
29	Technical University of Munich	Germany
30	University of Hong Kong	Hong Kong
31	University of California, San Diego	United States
32	LMU Munich	Germany
33	University of Melbourne	Australia
34	King’s College London	United Kingdom
35	Nanyang Technological University, Singapore	Singapore
36	London School of Economics and Political Science	United Kingdom
37	Georgia Institute of Technology	United States
38	The University of Tokyo	Japan
39	University of British Columbia	Canada
40	École Polytechnique Fédérale de Lausanne	Switzerland
41	KU Leuven	Belgium
42	Universität Heidelberg	Germany
43	Monash University	Australia
44	Chinese University of Hong Kong	Hong Kong
45	McGill University	Canada
46	Paris Sciences et Lettres - PSL Research Unive…	France
47	University of Illinois at Urbana-Champaign	United States
48	Karolinska Institute	Sweden
49	University of Texas at Austin	United States

	University Rank	OverAll Score	International Outlook Score	No of student	\
0	1	96.4	96.2	20,965	
1	2	95.2	80.5	21,887	
2	3	94.8	95.8	20,185	
3	3	94.8	79.8	16,164	
4	5	94.2	89.3	11,415	
5	6	94.1	83.6	2,237	
6	7	92.4	80.3	8,279	
7	8	92.1	78.4	40,921	
8	9	91.4	70.9	13,482	
9	10	90.4	97.5	18,545	
10	11	89.4	79.9	21,781	

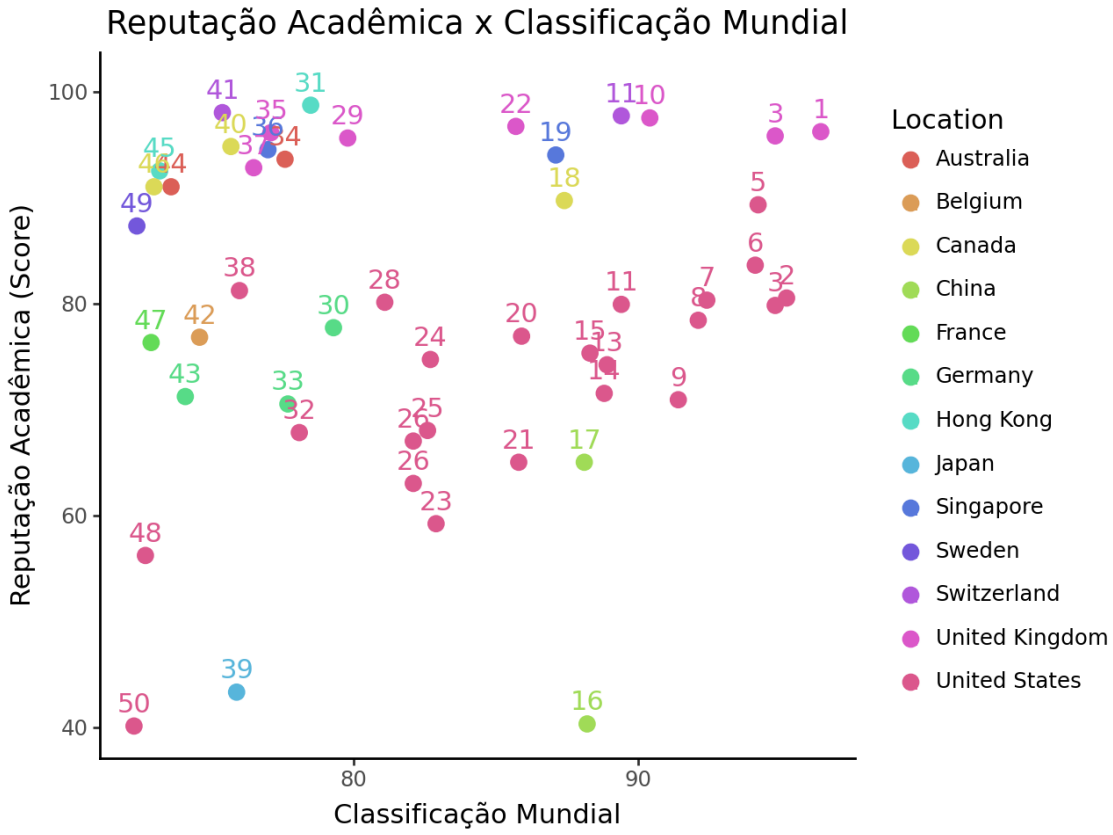
11	11	89.4	97.7	21,665
12	13	88.9	74.2	15,366
13	14	88.8	71.5	21,453
14	15	88.3	75.3	17,584
15	16	88.2	40.3	38,324
16	17	88.1	65.0	31,994
17	18	87.4	89.7	77,468
18	19	87.1	94.0	32,337
19	20	85.9	76.9	24,027
20	21	85.8	65.0	42,434
21	22	85.7	96.7	36,790
22	23	82.9	59.2	45,912
23	24	82.7	74.7	36,337
24	25	82.6	68.0	16,091
25	26	82.1	67.0	19,175
26	26	82.1	63.0	47,727
27	28	81.1	80.1	14,305
28	29	79.8	95.6	32,845
29	30	79.3	77.7	33,960
30	31	78.5	98.7	18,087
31	32	78.1	67.8	37,030
32	33	77.7	70.5	35,003
33	34	77.6	93.6	49,588
34	35	77.1	96.1	28,965
35	36	77.0	94.5	24,651
36	37	76.5	92.8	11,120
37	38	76.0	81.2	28,826
38	39	75.9	43.3	26,112
39	40	75.7	94.8	56,452
40	41	75.4	98.0	11,641
41	42	74.6	76.8	47,663
42	43	74.1	71.2	19,347
43	44	73.6	91.0	58,725
44	45	73.2	92.5	18,468
45	46	73.0	91.0	32,309
46	47	72.9	76.3	16,218
47	48	72.7	56.2	48,674
48	49	72.4	87.3	8,021
49	50	72.3	40.1	49,171

	No of student per staff
0	10.6
1	9.6
2	11.3
3	7.1
4	8.2
5	6.2

6	8.0
7	18.4
8	5.9
9	11.2
10	4.5
11	14.8
12	6.0
13	6.3
14	4.7
15	11.6
16	10.3
17	25.8
18	19.8
19	10.3
20	9.7
21	10.3
22	8.2
23	9.8
24	4.2
25	13.2
26	10.8
27	12.9
28	11.8
29	40.6
30	18.2
31	11.3
32	34.2
33	23.8
34	11.8
35	15.1
36	11.9
37	24.7
38	10.5
39	18.9
40	12.3
41	36.5
42	13.5
43	42.5
44	19.0
45	12.5
46	15.0
47	17.4
48	9.9
49	17.4

1.1.3 Reputação Acadêmica x Classificação

[7]:
```
# Criar o gráfico de dispersão
(
    ggplot(university_reputation_50)
    + aes(y='International Outlook Score', x='OverAll Score', color='Location',␣
 ↪label='University Rank')
    + labs(title='Reputação Acadêmica x Classificação Mundial', y='Reputação␣
 ↪Acadêmica (Score)', x='Classificação Mundial')
    + geom_abline(intercept=university_reputation_50['International Outlook␣
 ↪Score'].max(),
          slope=(1- np.log10(university_reputation_50['International Outlook␣
 ↪Score'].mean())), linetype="dashed", color="darkgrey")
    + geom_point(show_legend=True, size=3)
    + geom_text(nudge_y=2)
    + theme()
    + theme_classic()
)
```



1.1.4 Estudantes por Professor x Pontuação Geral

[14]:
```
(
    ggplot(university_reputation_50)
    + aes(y='No of student per staff', x='OverAll Score', color='Location',␣
 ↪label='OverAll Score')
    + labs(title='Estudantes por Professor x Pontuação Geral', y='Estudantes/
 ↪Professor', x='Pontuação Geral')
    + geom_point(show_legend=True, size=3)
    + geom_text(nudge_y=1.5)
    + theme()
    + theme_classic()
)
```