Autor: Leonardo Vinicius Wanderley Jatobá da Silva
Mini projeto de ciência de dados

Proposta do curso: Análise e interpretação de dados em python

Dataset escolhido: cidades com o custo de vida saudável 2021.
Link: https://www.kaggle.com/datasets/prasertk/healthy-lifestyle-cities-report-2021

Objetivo: analisar as principais semelhanças e diferenças entre as cidades com o custo de vida mais saudável, a influência do seu país, das suas regiões e dos temas tratados no dataset em questão.

Análise inicial:
A base de dados tem para análise os seguintes atributos de cada cidade: nome da cidade, rank, horas de sol, custo de uma garrafa de água, nível de poluição, horas trabalhadas anuais, horas ao ar livre, número de lugares para visitar, custo de uma mensalidade de academia, e do país: nível de felicidade, nível de obesidade e expectativa de vida. Analisando os dez primeiros dados notamos que grande parte desses países estando nessa colocação tem equilíbrio em alguns aspectos, citando por exemplo a cidade de Copenhagen, 5° colocada do rank, entre os 10 primeiros ela tem o maior custo de uma garrafa de água, mas em compensação entre eles é o país com menor horas trabalhadas anuais, há frequentemente esse equilíbrio entre os dados do dataset, o que falta em algo é completado em outro e o país com que mais se destaca nesse quesito é Amsterdam, ocupando a 1° posição. Mas para analisar os dados em si esse "equilíbrio" e o fato de que cada cidade tem as suas peculiaridades, torna a análise visualmente difícil se vista cada coluna separadamente, os dados não seguirão uma distribuição normal, por exemplo se analisarmos o custo de 1 garrafa de água, esse custo não vai começar do maior valor na cidade de último rank e vai ter o menor valor na cidade de 1° rank, apenas alguns atributos analisados juntos é que se pode tirar conclusões de uma análise mais assertiva.

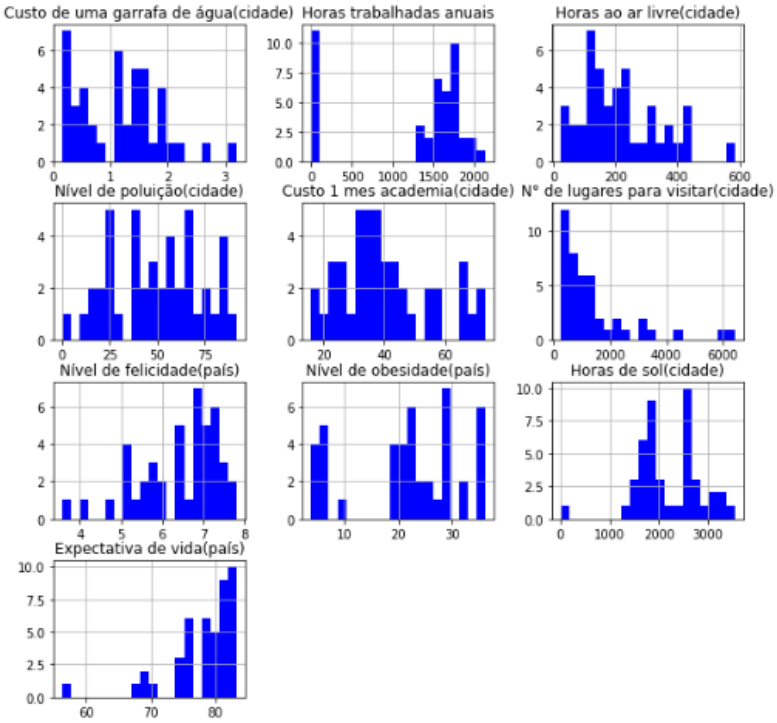
	Cidade	Rank	Horas de sol(cidade)	Custo de uma garrafa de água(cidade)	Nível de obesidade(país)	Expectativa de vida(país)	Nível de poluição(cidade)	Horas trabalhadas anuais	Nível de felicidade(país)	Horas ao ar livre(cidade)	Nº de lugares para visitar(cidade)	Custo 1 mes academia(cidade)
0	Amsterdam	1	1858	£1.92	20.40%	81.2	30.93	1434	7.44	422	1048	£34.90
1	Sydney	2	2636	£1.48	29.00%	82.1	26.86	1712	7.22	406	1103	£41.66
2	Vienna	3	1884	£1.94	20.10%	81.0	17.33	1501	7.29	132	1008	£25.74
3	Stockholm	4	1821	£1.72	20.60%	81.8	19.63	1452	7.35	129	598	£37.31
4	Copenhagen	5	1630	£2.19	19.70%	79.8	21.24	1380	7.64	154	523	£32.53
5	Helsinki	6	1662	£1.60	22.20%	80.4	13.08	1540	7.80	113	309	£35.23
6	Fukuoka	7	2769	£0.78	4.30%	83.2	-	1644	5.87	35	539	£55.87
7	Berlin	8	1626	£1.55	22.30%	80.6	39.41	1386	7.07	254	1729	£26.11
8	Barcelona	9	2591	£1.19	23.80%	82.2	65.19	1686	6.40	585	2344	£37.80
9	Vancouver	10	1938	£1.08	29.40%	81.7	24.26	1670	7.23	218	788	£31.04

Tratamento de dados:

Sobre o tratamento de dados, algumas colunas estavam em um formato equivocado e precisaram ser transformadas, algum símbolo como por exemplo o símbolo de euro “£” precisou ser removido para dada transformação, e foi removida a coluna de Rank, pois as cidades já estavam na ordem não seria muito interessante analisar o rank, por exemplo se um histograma fora feito utilizando a coluna rank não dava para se chegar a nenhuma conclusão, tendo em vista que é apenas uma sequência de números de 1 até 44(quantidade total de cidades).

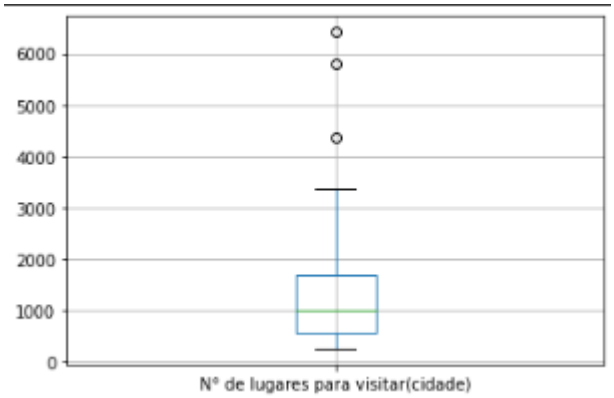
Visualização de dados:

Sobre a visualização de dados, foram escolhidas algumas formas de visualização, primeiramente há um histograma geral para cada atributo separadamente, apenas para uma análise e observação prévia.

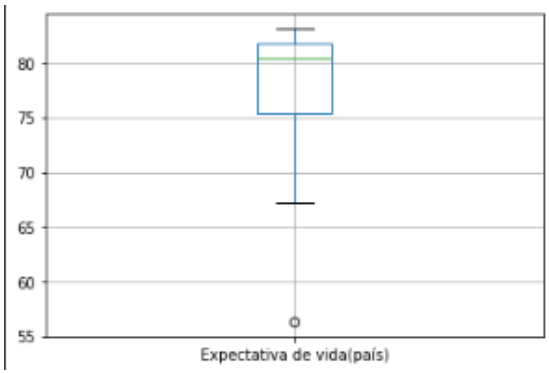


Em seguida temos uma confecção de três gráficos no estilo boxplot(diagrama de caixa), e nele podemos notar alguns outliers, alguns valores que apresentam uma certa discrepância quanto aos dados separados em quartis.

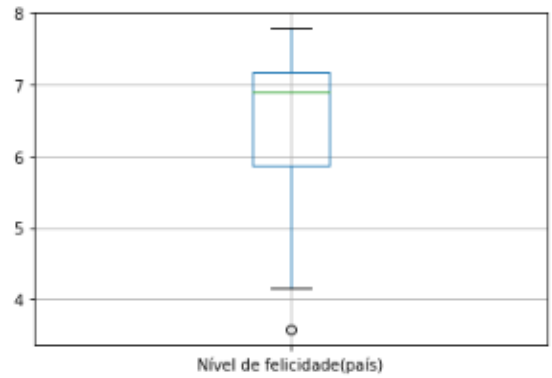
O primeiro boxplot representa a quantidade de lugares para visitar por cidade com a presença de 3 outliers, os quartis com valores entre 500 e 2000, com o valor máximo um pouco acima de 3000, o outlier com maior valor é a cidade de Londres possuindo 6.417 lugares para visitar.



O segundo boxplot é o de expectativa de vida do país onde possuímos o outlier de Johanesburg (África do Sul) com a faixa etária entre 56 e 57 anos.

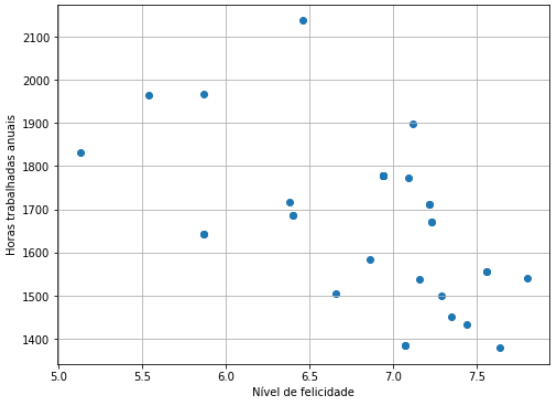


E no último boxplot analisando o nível de felicidade dos países temos mais um outlier com Johanesburg (África do Sul) aparecendo novamente dessa vez com a nota de 4.81 para o nível de felicidade do país.

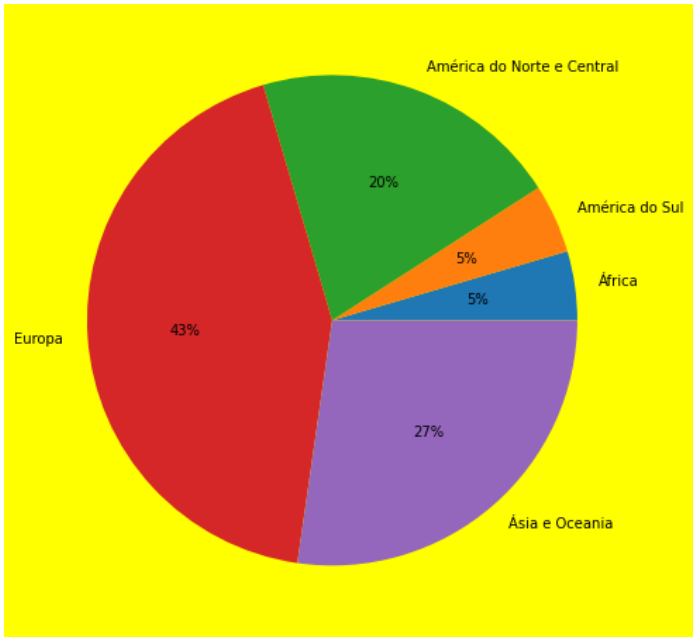


Após a análise dos boxplot, há a confecção de um gráfico de dispersão fazendo uma relação entre duas colunas, são elas: Horas trabalhadas anuais e Nível de felicidade, onde percebe-se

que as duas são intimamente ligadas, sendo perceptível que, quanto maior o nível de felicidade do país, menores são as horas trabalhadas anualmente.



E por fim, separando os países em seus respectivos continentes e criando um gráfico de pizza, há de se saber que, 43% dos países são do continente europeu, 27% são da Ásia e Oceania, 20% da América do Norte e Central, 5% para África, e 5% também para a América do Sul.



O que se pode observar é a grande quantidade de presença de cidades da Europa e Ásia e Oceania, só a Europa tem quase a metade das cidades no dataset, América do Sul e África estão bem abaixo dos outros continentes, possuindo apenas 5% cada, e destaque na África para Johanesburg que possui presença como outlier em dois boxplot o de nível de felicidade e de expectativa de vida.

Análise final:

Analisando a base de dados e por meio de conhecimentos gerais, há de se notar que, vários fatores são importantes para a adoção de um estilo de vida saudável, e podemos ver analisando essa base de dados com 44 países, cada um com suas peculiaridades, sua quantidade de lugares para visitar, seus custos, nível de poluição, expectativa de vida, entre outros. Nota-se a relação entre as horas trabalhadas pela população relacionada com o nível de felicidade, quanto menos as pessoas trabalham mais elas são felizes com esses dados, entre todas essas cidades há uma quantidade muito grande de lugares para se visitar, é observável pelo boxplot que os números ficam numa concentração muito boa entre 500 e 2000. Analisando os 10 primeiros vemos a semelhança entre eles como o custo de uma garrafa de água menor que 2.19 euros, a taxa de obesidade sempre abaixo de 30%, mas o destaque fica mesmo para o equilíbrio entre os atributos que faz algumas cidades terem esse destaque e serem consideradas ótimos e saudáveis lugares para se viver no ano de 2021.