Autora do Estudo: ©Carla G. B. Teixeira
carla.g.b.teixeiral@gmail.com

macrovector - Freepik.com

Projeto de Análise de Dados

Avaliações de Vinhos

Histórico de Revisões deste Documento		
Autor	Data	Alterações
Carla G. B. Teixeira	19/09/2023	Elaboração
Carla G. B. Teixeira	20/09/2023	Revisão
Carla G. B. Teixeira	25/09/2023	Revisão
Carla G. B. Teixeira	26/09/2023	Revisão
Carla G. B. Teixeira	04/10/2023	Revisão
Carla G. B. Teixeira	06/10/2023	Revisão
Carla G. B. Teixeira	07/10/2023	Revisão
Carla G. B. Teixeira	08/10/2023	Revisão
Carla G. B. Teixeira	09/10/2023	Revisão
Carla G. B. Teixeira	13/10/2023	Revisão

Entendimento do Negócio

Contexto do Negócio

Críticas sobre vinhos com conteúdo de 129971 mil observações de avaliações de vinhos com país (*country*), região (*region*), variedade (*variety*), título (*title*), vinícola (*winery*), preço (*price*), pontuação (*points*), descrição (*description*), designação (*designation*). Os dados foram extraídos do *WineEnthusiast* durante a semana de 15 de junho de 2017.

Objetivos

Realizar o resumo e interpretação de um conjunto de observações proporcionando o conhecimento e entendimento dos dados.

Premissas

- Origem das observações:

Dados coletados em: https://www.kaggle.com/datasets/zynicide/wine-reviews
Arquivo: **winemag-data-130k-v2.csv**(52.91 MB)

Os dados foram extraídos do *WineEnthusiast* durante a semana de 15 de junho de 2017.

- Valores Nulos

O campo preço (*price*) tem valores nulos, dos foi preenchido com o valor da sua mediana (US$ 25.00).

Riscos Envolvidos

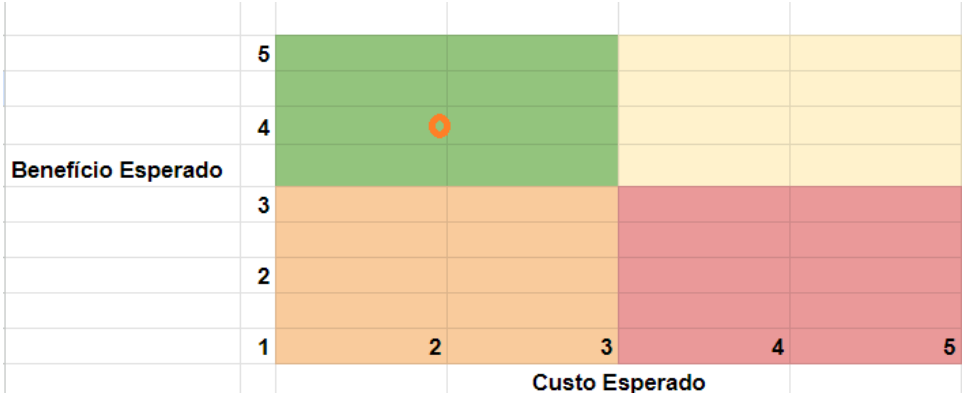
Não determinado.

Custo x Benefício

Planilha Levantamento

Dados	Custo Tangível	Custo Intangível	Benefícios	Nota para o Custo (1 a 5)	Nota para o Benefício (1 a 5)
Avaliação dos Vinhos	0	Tempo de coleta e análise	Aprendizado e Práticas Entendimento dos dados.	2	4

Matriz Custo vs Benefícios



Glossário de Termos

Notas/Pontuação dos Vinhos:
Notas entre 96–100: vinhos considerados extraordinários, profundos e complexos. Apresentam todos os atributos clássicos esperados para essa categoria de bebidas.
Notas entre 90–95: vinhos excelentes, têm complexidade, caráter excepcional e são considerados bebidas de muita qualidade;
Notas entre 80–89: vinhos “acima da média”, finos em diversos aspectos, como sabor e aromas. Por mais que não sejam bebidas de destaque, não apresentam nenhuma falha aparente.
Notas entre 70–79: categoria de vinhos medianos e inócuos.
Notas entre 60–69: vinhos abaixo da média, apresentam falhas evidentes, como presença alta de acidez e taninos, ausência de sabores e aromas.
Notas entre 50–59: vinhos considerados inaceitáveis e inapropriados para consumo.
(Referência: https://www.divvino.com.br/blog/pontuacao-de-vinhos/)

Critérios de Sucesso

Nº	Item
1.	Identificar a concentração e distribuição dos dados com relação ao seu preço.
2.	Identificar a concentração e distribuição dos dados com relação a sua pontuação.
3	Obter as frequências das observações pelos seus atributos de país, província, vinícola, classificação, pontuação, avaliador/testador do vinho, variedade e preço.
4.	Estabelecer a correlação entre as variáveis numéricas e determinar se as variáveis categóricas possuem influência na precificação e pontuação dos vinhos.
6.	Verificar como o Brasil está posicionado com relação às avaliações que foram submetidas.
7.	Verificar se as técnicas foram corretamente aplicadas
8.	Concluir se o estudo foi válido para obter o resumo, interpretação e conhecimento dos dados.
9.	Obter, no mínimo, 5 revisões do estudo realizado. (Reviews)
10	Publicar estudo final no Linkedin e GitHub no Portfólio com as revisões

Entendimento dos Dados

Descrição dos Dados

Amostra dos Dados:

country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	winery	Classificacao
Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	Nicosia	Acima da Média
Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos	Acima da Média
US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris	Rainstorm	Acima da Média
US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	NaN	Alexander Peartree	NaN	St. Julian 2013 Reserve Late Harvest Riesling ...	Riesling	St. Julian	Acima da Média
US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87	65.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Sweet Cheeks 2012 Vintner's Reserve Wild Child...	Pinot Noir	Sweet Cheeks	Acima da Média
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Germany	Notes of honeysuckle and cantaloupe	Brauneberger Juffer-Sonnenuhr	90	28.0	Mosel	NaN	NaN	Anna Lee C. Iijima	NaN	Dr. H. Thanisch (Erben Müller-Burggraef)	Riesling	Dr. H. Thanisch (Erben Müller-	Excelente

Coluna	Descrição
Country	País de origem do vinho
Description	Descrição das características do vinho
Designation	A vinha dentro da adega de onde provêm as uvas que dão origem ao vinho
Points	O número de pontos que a *WineEnthusiast* classificou o vinho em uma escala de 1 a 100
Price	O custo de uma garrafa de vinho em US$
Province	A província ou estado de origem do vinho
Region_1	A área vitivinícola em uma província ou estado (exemplo: Napa)
Region_2	Às vezes, há regiões mais específicas especificadas dentro de uma área vitivinícola (ou seja, Rutherford dentro do Vale de Napa)
Taster_name	Nome do testador/avaliador do vinho (enófilo)

Coluna	Descrição
Taster_twitter_handle	Twitter do testador
Variety	Tipo da uva utilizada na preparação do vinho e responsável por terminar de moldar o sabor de cada bebida, sendo uma das principais responsáveis pela suavidade ou amargor de rótulos específicos.
Winery	Vinícola onde o vinho foi produzido
Classificacao	Criada coluna que atribui uma classificação da qualidade do vinho, conforme a faixa de pontuação: 50-59 -> Inapropriado para Consumo 60-69-> Abaixo da Média 70-79-> Mediano e Inócuo 80-89-> Acima da Média 90-95-> Excelente 96-100-> Extraordinário

Análise Exploratória e Qualidade dos Dados (Sanity Check)

Medidas Resumo

(count, mean, std, min, Q25%, Q50%,Q75%, max, median)

	points	price
count	129971.000000	129971.000000
mean	88.447138	34.646083
std	3.039730	39.664385
min	80.000000	4.000000
25%	86.000000	18.000000
50%	88.000000	25.000000
75%	91.000000	40.000000
max	100.000000	3300.000000
median	88.0	25.0

Tabelas de Frequência

Frequência e Porcentagem de Avaliações Realizadas por País

Número de países englobados: 43 países

	Frequência	Porcentagem(%)
US	54504	41.955846
France	22093	17.006651
Italy	19540	15.041414
Spain	6645	5.115158
Portugal	5691	4.380793
Chile	4472	3.442436
Argentina	3800	2.925147
Austria	3345	2.574899
Australia	2329	1.792807
Germany	2165	1.666564
New Zealand	1419	1.092311
South Africa	1401	1.078456
Israel	505	0.388737
Greece	466	0.358715
Canada	257	0.197832
Hungary	146	0.112387
Bulgaria	141	0.108538
Romania	120	0.092373
Uruguay	109	0.083906
Turkey	90	0.069280
Slovenia	87	0.066970
Georgia	86	0.066201
England	74	0.056963
Croatia	73	0.056194
Mexico	70	0.053884
Moldova	59	0.045417
Brazil	52	0.040028
Lebanon	35	0.026942
Morocco	28	0.021554
Peru	16	0.012316
Ukraine	14	0.010777
Serbia	12	0.009237
Czech Republic	12	0.009237
Macedonia	12	0.009237

Brasil - representa 0.04% do total de avaliações (52 avaliações realizadas)

Distribuição da Frequência Absoluta, Relativa e Acumulada das Classificações dos Vinhos

	Classificacao	Frequência Absoluta	Frequência Relativa	Frequência Acumulada
2	Extraordinário	881	0.006778	129971
1	Excelente	48164	0.370575	129090
0	Acima da Média	80926	0.622647	80926
4	Abaixo da Média	0	0.000000	129971
5	Mediano e Inócuo	0	0.000000	129971
3	Inapropriado para Consumo	0	0.000000	129971

Distribuição da Frequência e Porcentagem por Pontuação (*points*)

	Frequência	Porcentagem(%)
88	17207	13.239107
87	16933	13.028291
90	15410	11.856491
86	12600	9.694470
89	12226	9.406714
91	11359	8.739642
92	9613	7.396265
85	9530	7.332405
93	6489	4.992652
84	6480	4.985728
94	3758	2.891414
83	3025	2.327442
82	1836	1.412623
95	1535	1.181033
81	692	0.532426
96	523	0.402397
80	397	0.305453
97	229	0.176193
98	77	0.059244
99	33	0.025390
100	19	0.014619

Acima da Média
Excelente
Extraordinário

Distribuição da Frequência da Classificação por País

Classificacao country	Inapropriado para Consumo	Abaixo da Média	Mediano e Inócuo	Acima da Média	Excelente	Extraordinário
Argentina	0.0	0.0	0.0	0.795789	0.203421	0.000789
Armenia	0.0	0.0	0.0	1.000000	0.000000	0.000000
Australia	0.0	0.0	0.0	0.593388	0.397166	0.009446
Austria	0.0	0.0	0.0	0.396712	0.594021	0.009268
Bosnia and Herzegovina	0.0	0.0	0.0	1.000000	0.000000	0.000000
Brazil	0.0	0.0	0.0	1.000000	0.000000	0.000000
Bulgaria	0.0	0.0	0.0	0.765957	0.234043	0.000000
Canada	0.0	0.0	0.0	0.466926	0.533074	0.000000
Chile	0.0	0.0	0.0	0.854651	0.145349	0.000000
China	0.0	0.0	0.0	1.000000	0.000000	0.000000
Croatia	0.0	0.0	0.0	0.835616	0.164384	0.000000
Cyprus	0.0	0.0	0.0	1.000000	0.000000	0.000000
Czech Republic	0.0	0.0	0.0	1.000000	0.000000	0.000000
Egypt	0.0	0.0	0.0	1.000000	0.000000	0.000000
England	0.0	0.0	0.0	0.189189	0.810811	0.000000
France	0.0	0.0	0.0	0.583941	0.404653	0.011406
Georgia	0.0	0.0	0.0	0.790698	0.209302	0.000000
Germany	0.0	0.0	0.0	0.437413	0.550577	0.012009
Greece	0.0	0.0	0.0	0.839056	0.160944	0.000000
Hungary	0.0	0.0	0.0	0.554795	0.417808	0.027397
India	0.0	0.0	0.0	0.222222	0.777778	0.000000
Israel	0.0	0.0	0.0	0.586139	0.413861	0.000000
Italy	0.0	0.0	0.0	0.659775	0.333521	0.006704
Lebanon	0.0	0.0	0.0	0.714286	0.285714	0.000000
Luxembourg	0.0	0.0	0.0	0.833333	0.166667	0.000000
Macedonia	0.0	0.0	0.0	1.000000	0.000000	0.000000
Mexico	0.0	0.0	0.0	0.928571	0.071429	0.000000
Moldova	0.0	0.0	0.0	0.813559	0.186441	0.000000
Morocco	0.0	0.0	0.0	0.750000	0.250000	0.000000
New Zealand	0.0	0.0	0.0	0.653982	0.346018	0.000000
Peru	0.0	0.0	0.0	1.000000	0.000000	0.000000
Portugal	0.0	0.0	0.0	0.646108	0.347039	0.006853
Romania	0.0	0.0	0.0	0.966667	0.033333	0.000000
Serbia	0.0	0.0	0.0	1.000000	0.000000	0.000000
Slovakia	0.0	0.0	0.0	1.000000	0.000000	0.000000
Slovenia	0.0	0.0	0.0	0.816092	0.183908	0.000000
South Africa	0.0	0.0	0.0	0.688794	0.311206	0.000000
Spain	0.0	0.0	0.0	0.759819	0.236870	0.003311
Switzerland	0.0	0.0	0.0	0.571429	0.428571	0.000000
Turkey	0.0	0.0	0.0	0.733333	0.266667	0.000000
US	0.0	0.0	0.0	0.590397	0.403163	0.006440
Ukraine	0.0	0.0	0.0	1.000000	0.000000	0.000000
Uruguay	0.0	0.0	0.0	0.844037	0.155963	0.000000

Brasil: representa 0.04% do total de avaliações -> 52 avaliações realizadas. 100% dos vinhos avaliados como “Acima da Média”

Frequência e Porcentagem de Vinhos Extraordinários por País (*country*)

Total: 881 avaliações com vinhos classificados como extraordinários (Pontuação entre 96-100)

	Frequência	Porcentagem(%)
US	351	39.841090
France	252	28.603859
Italy	131	14.869467
Portugal	39	4.426788
Austria	31	3.518729
Germany	26	2.951192
Australia	22	2.497162
Spain	22	2.497162
Hungary	4	0.454030
Argentina	3	0.340522

Frequência e Porcentagem por Avaliador/Testador (*Taster_Name*)

	Frequência	Porcentagem(%)
Roger Voss	25514	24.597260
Michael Schachner	15134	14.590222
Kerin O'Keefe	10776	10.388809
Virginie Boone	9537	9.194327
Paul Gregutt	9532	9.189507
Matt Kettmann	6332	6.104486
Joe Czerwinski	5147	4.962064
Sean P. Sullivan	4966	4.787567
Anna Lee C. Iijima	4415	4.256365
Jim Gordon	4177	4.026917
Anne Krebiehl MW	3685	3.552595
Lauren Buzzeo	1835	1.769067
Susan Kostrzewa	1085	1.046015
Mike DeSimone	514	0.495532
Jeff Jenssen	491	0.473358
Alexander Peartree	415	0.400089
Carrie Dykes	139	0.134006
Fiona Adams	27	0.026030
Christina Pickard	6	0.005784

Frequência e Porcentagem por Variedade (Variety)

	Frequência	Porcentagem(%)
Pinot Noir	13272	10.211587
Chardonnay	11753	9.042856
Cabernet Sauvignon	9472	7.287836
Red Blend	8946	6.883127
Bordeaux-style Red Blend	6915	5.320459
...	...	...
Cabernet Sauvignon-Barbera	1	0.000769
Sauvignonasse	1	0.000769
Forcallà	1	0.000769
Meseguera	1	0.000769
Bobal-Cabernet Sauvignon	1	0.000769

707 rows × 2 columns

macrovector - Freepik.com

Frequência Absoluta Variedade (Variety) e Testador (Taster_Name)

taster_name	Alexander Peartree	Anna Lee C. Iijima	Anne Krebiehl MW	Carrie Dykes	Christina Pickard	Fiona Adams	Jeff Jenssen	Jim Gordon	Joe Czerwinski	Kerin O'Keefe	Lauren Buzzeo	Matt Kettmann	Michael Schachner	Mike DeSimone
variety														
Bordeaux-style Red Blend	17	79	0	20	0	1	8	29	69	0	86	142	139	27
Chardonnay	52	291	67	15	1	1	29	421	456	123	197	924	877	23
Pinot Noir	12	201	170	0	1	0	30	560	636	0	61	1570	524	3
Red Blend	27	70	76	9	0	1	60	430	85	2507	116	337	1496	91
Portuguese Red	0	0	0	0	0	0	0	0	3	0	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Manzoni	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Maria Gomes-Bical	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Marsanne-Viognier	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Marzemino	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Žilavka	0	0	0	0	0	0	1	0	0	0	0	0	0	0

670 rows × 19 columns

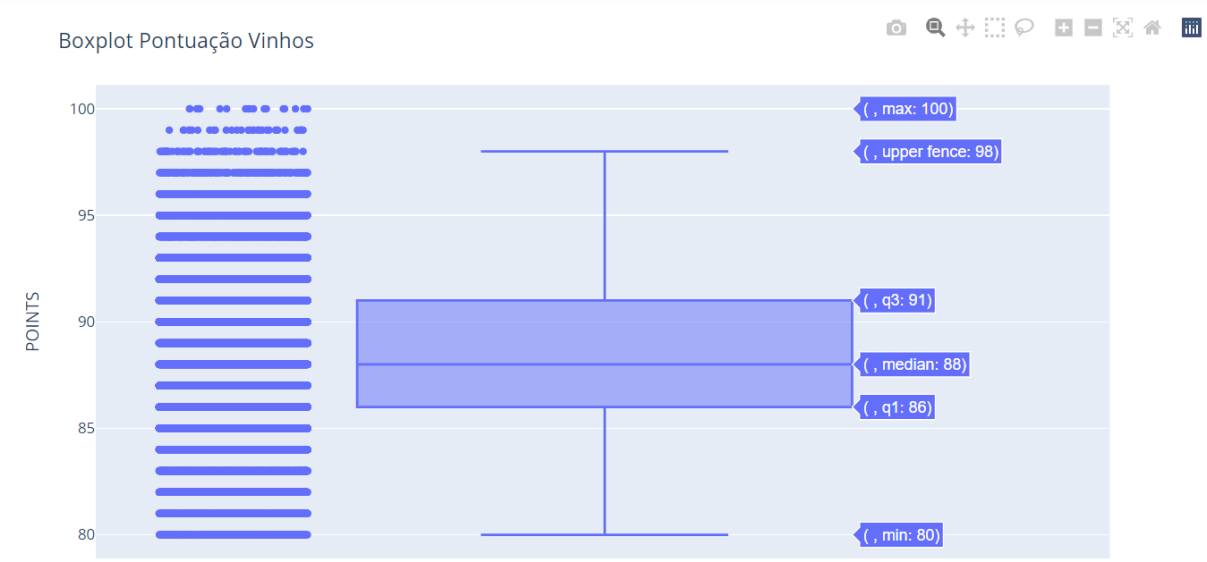
taster_name	Paul Gregutt	Roger Voss	Sean P. Sullivan	Susan Kostrzewa	Virginie Boone
variety					
Bordeaux-style Red Blend	319	4710	469	34	241
Chardonnay	797	2786	349	105	1429
Pinot Noir	2721	1872	56	11	1891
Red Blend	576	288	424	116	400
Portuguese Red	0	2462	0	0	0
...	...	...	...	...	...
Manzoni	0	0	0	0	0
Maria Gomes-Bical	0	1	0	0	0
Marsanne-Viognier	0	0	1	0	0
Marzemino	0	0	0	0	0
Žilavka	0	0	0	0	0

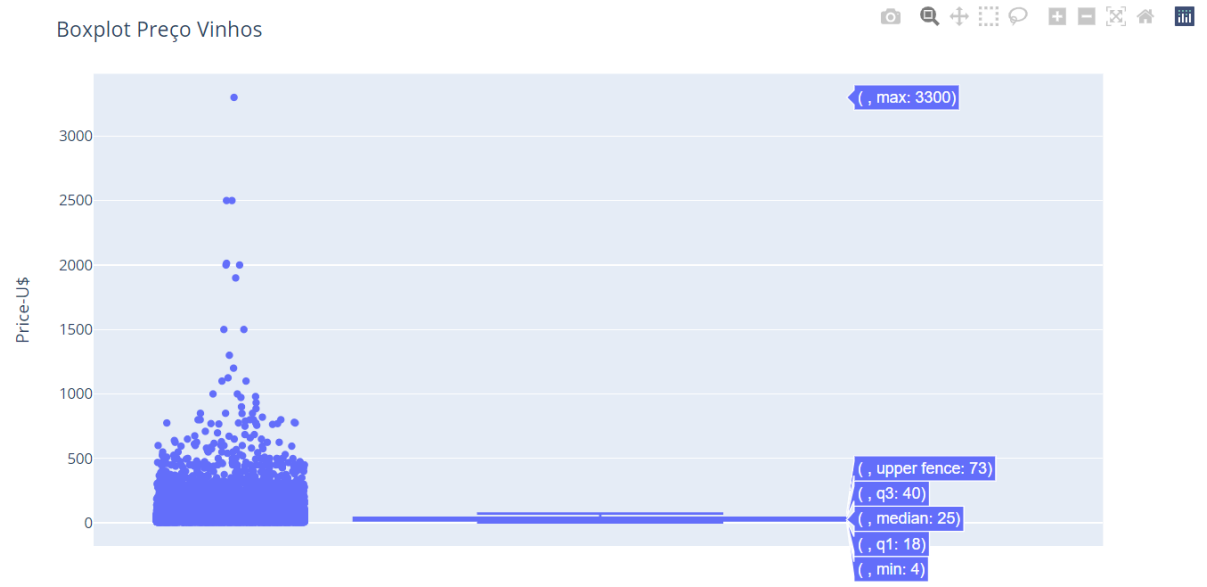
Frequência e Porcentagem por Preço (Price)

	Frequência	Porcentagem(%)
25.0	14801	11.387925
20.0	6940	5.339653
15.0	6066	4.667195
30.0	4951	3.809311
18.0	4883	3.756992
...	...	...
574.0	1	0.000769
630.0	1	0.000769
764.0	1	0.000769
319.0	1	0.000769
848.0	1	0.000769

390 rows × 2 columns

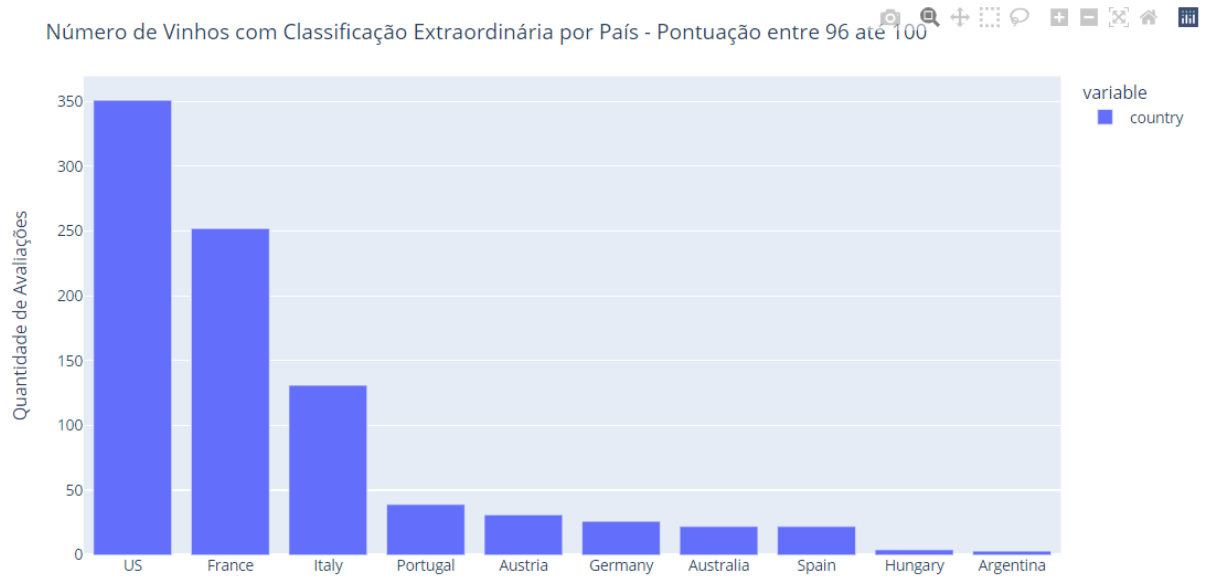
BoxPlot


Boxplot Pontuação Vinhos
POINTS
100
95
90
85
80
(, max: 100)
(, upper fence: 98)
(, q3: 91)
(, median: 88)
(, q1: 86)
(, min: 80)


Boxplot Preço Vinhos
Price-U$
3000
2500
2000
1500
1000
500
0
(, max: 3300)
(, upper fence: 73)
(, q3: 40)
(, median: 25)
(, q1: 18)
(, min: 4)

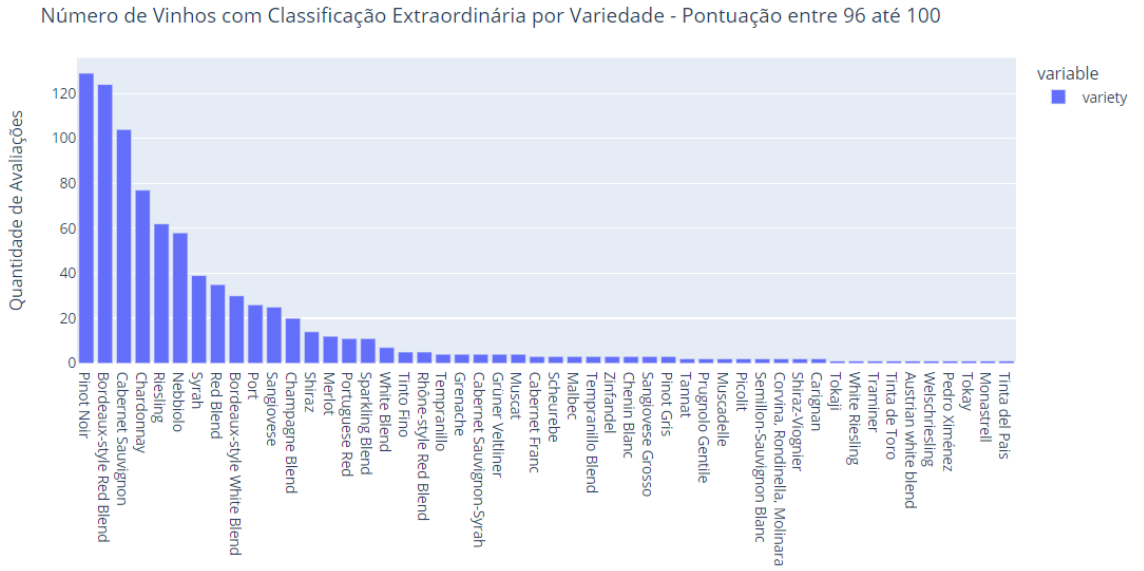
BarPlot (Gráficos de Barras)

Número de Vinhos com Classificação Extraordinária por País - Pontuação entre 96 até 100

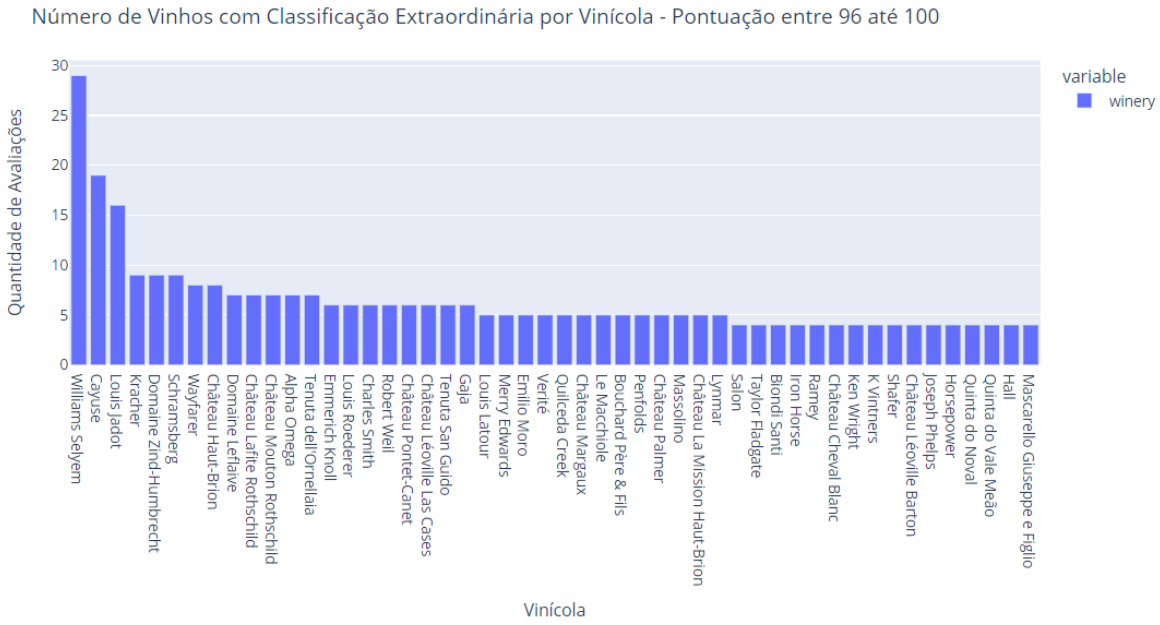


Número de Vinhos com Classificação Extraordinária por Variedade - Pontuação entre 96 até 100

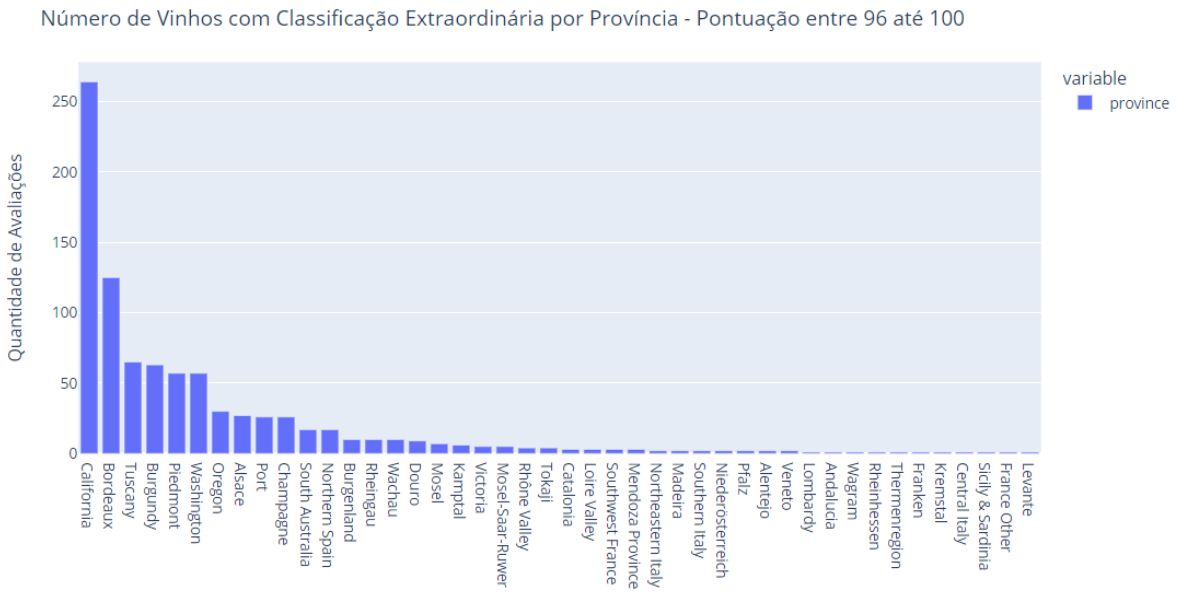
Apresenta as 50 primeiras variedades com maior número de avaliações extraordinárias.



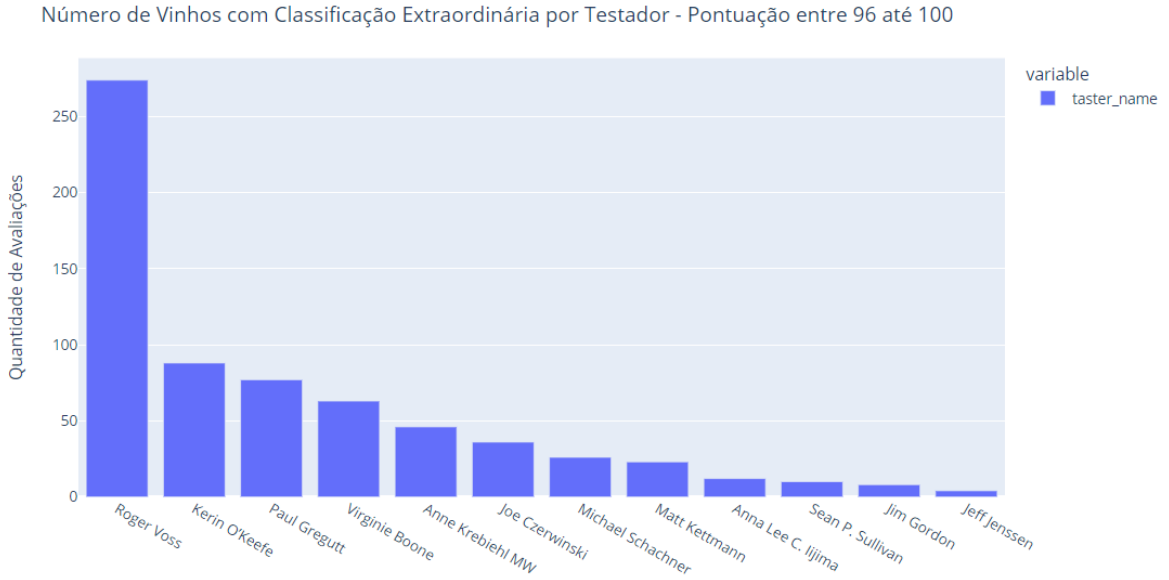
Número de Vinhos com Classificação Extraordinária por Vinícola - Pontuação entre 96 até 100



Número de Vinhos com Classificação Extraordinária por Província - Pontuação entre 96 até 100

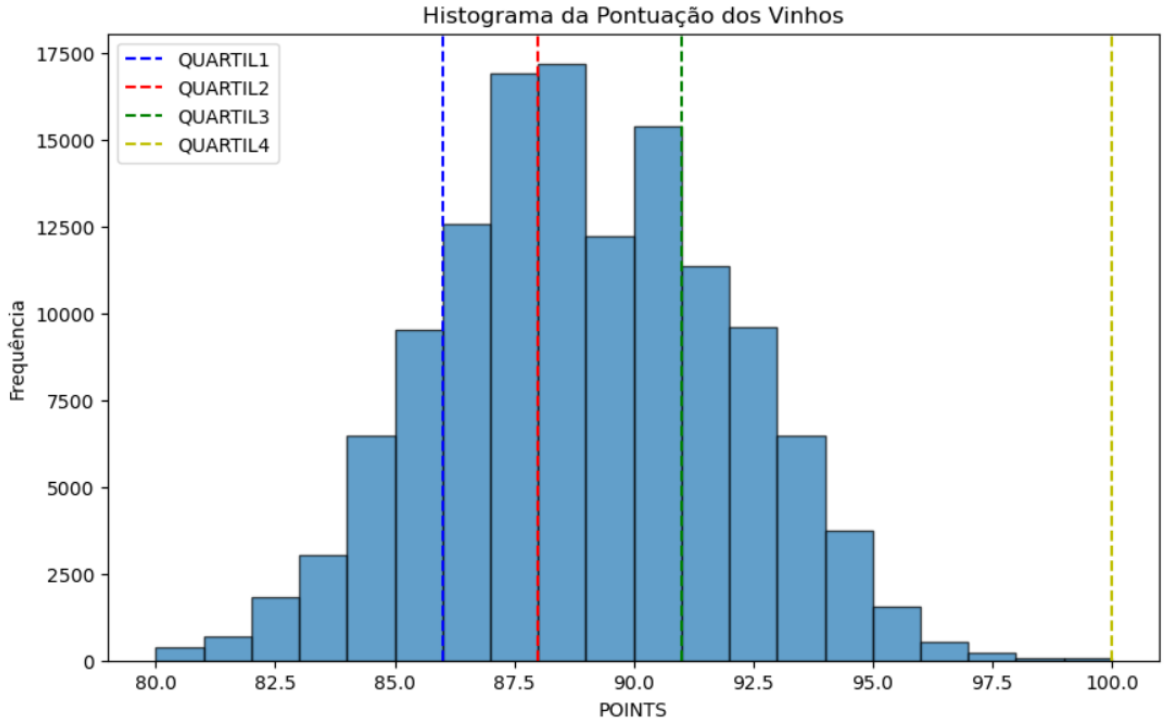


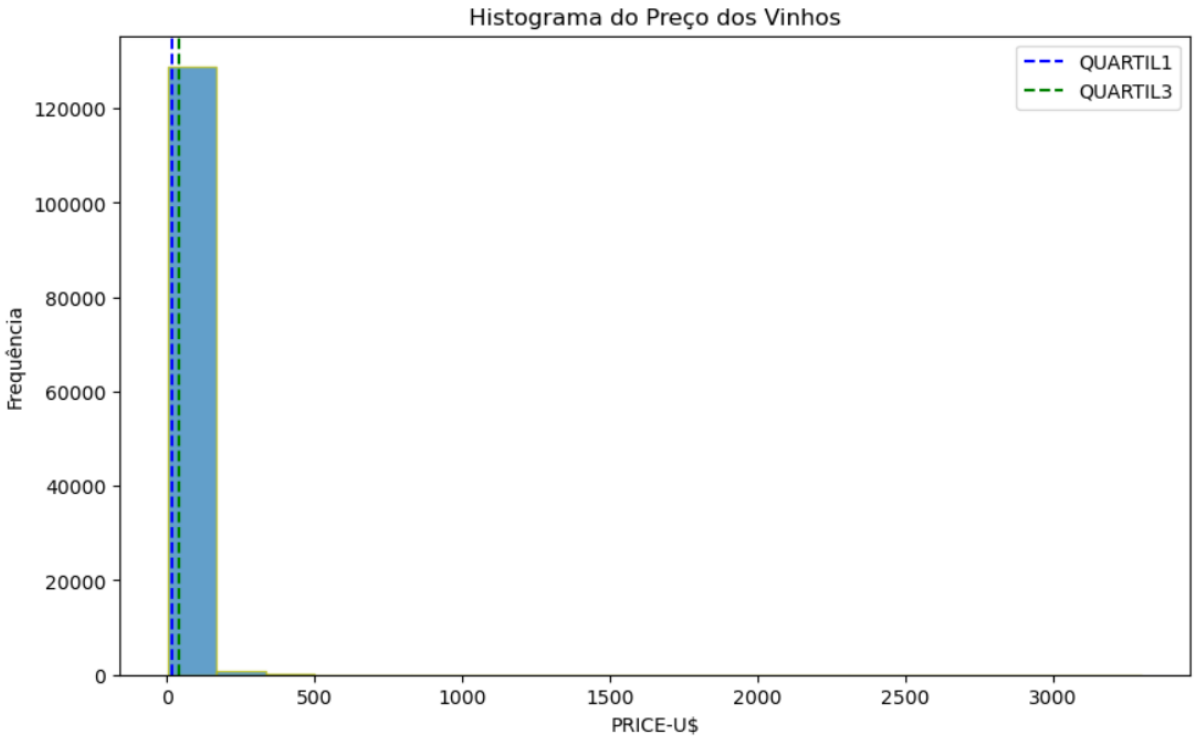
Número de Vinhos com Classificação Extraordinária por Testador - Pontuação entre 96 até 100



Histograma

Como está a distribuição das informações.

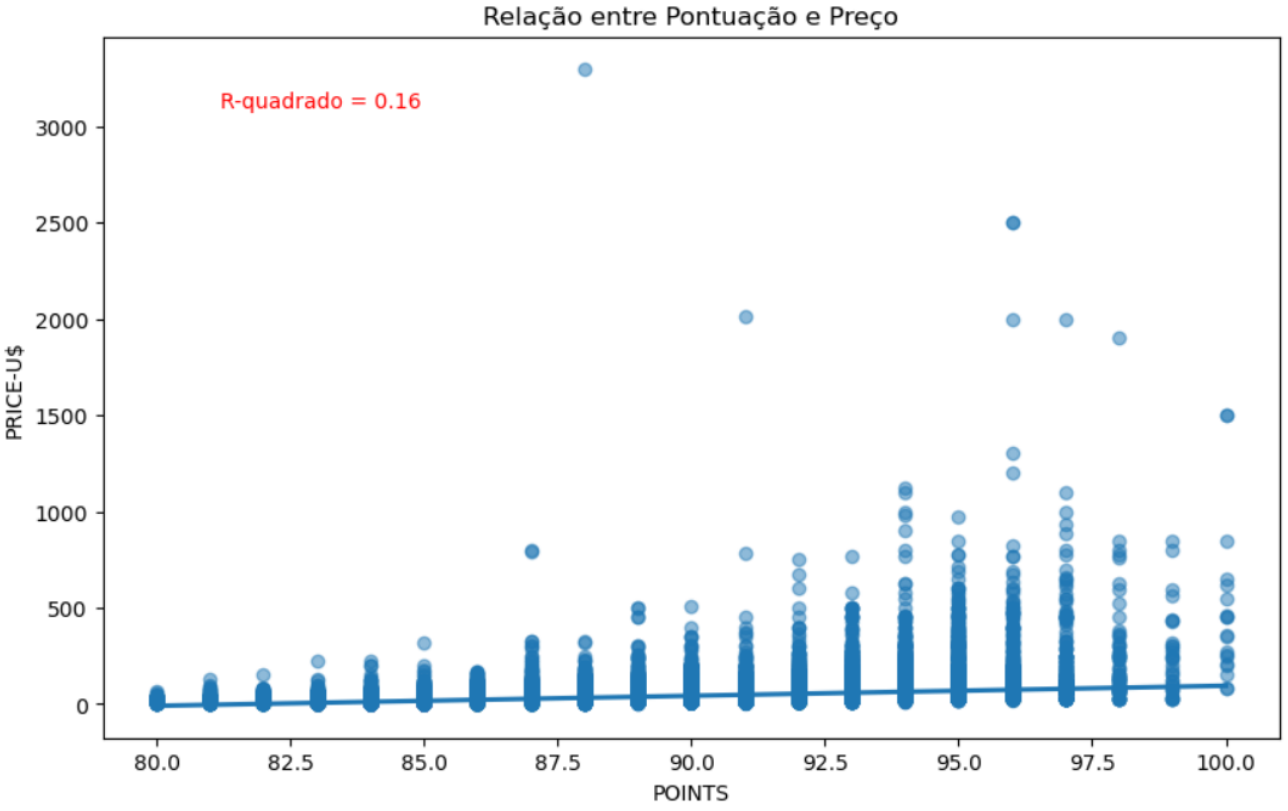




Correlação R^2

Considerando que quanto mais o valor calculado do R^2 se aproxima de 1, mais forte será a correlação das variáveis analisadas (no caso Pontuação x Preço), conclui-se que a correlação entre essas variáveis é FRACA. Desta forma, é importante considerar a existência de outros fatores (outras variáveis) que possam influenciar na precificação do vinho.

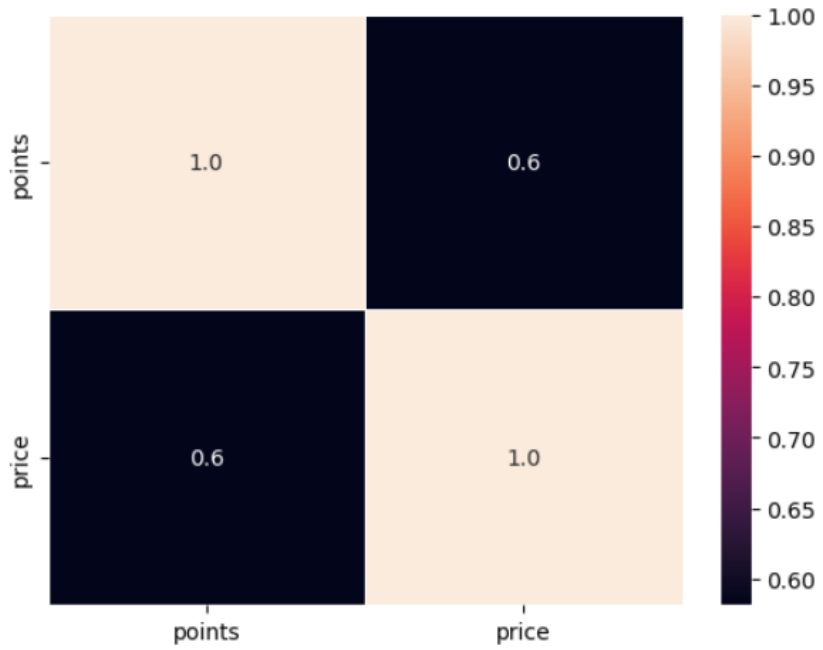
Abaixo gráfico de dispersão representativo com linha de tendência.



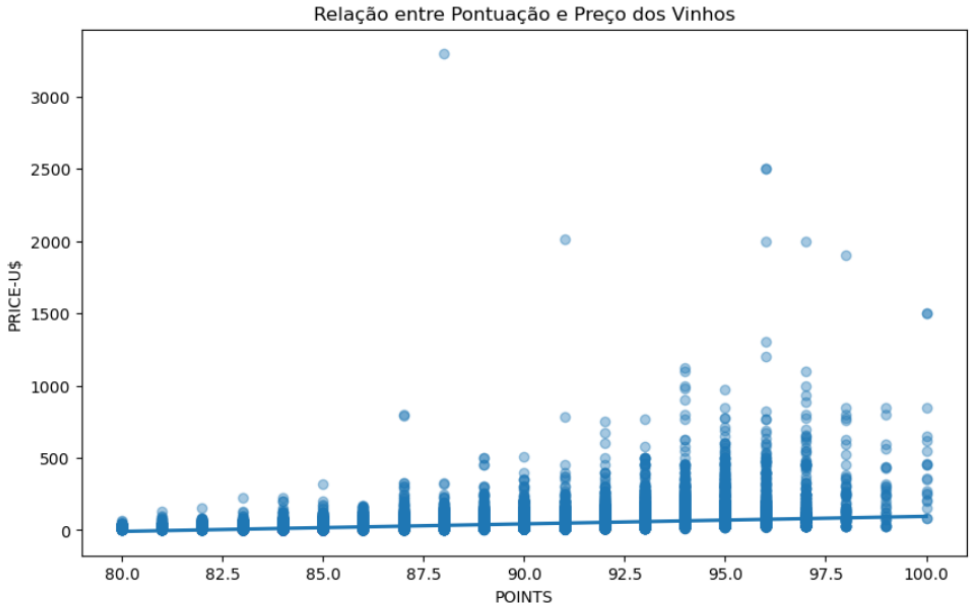
Correlação de Spearman

Verificamos que existem valores muito extremos de preço (*price*), causando uma variância muito grande. Desta forma vamos verificar a correlação com o método Spearman e representação do gráfico de calor e dispersão.

	points	price
points	1.000000	0.581179
price	0.581179	1.000000




```
Text(0.5, 1.0, 'Relação entre Pontuação e Preço dos Vinhos')
```

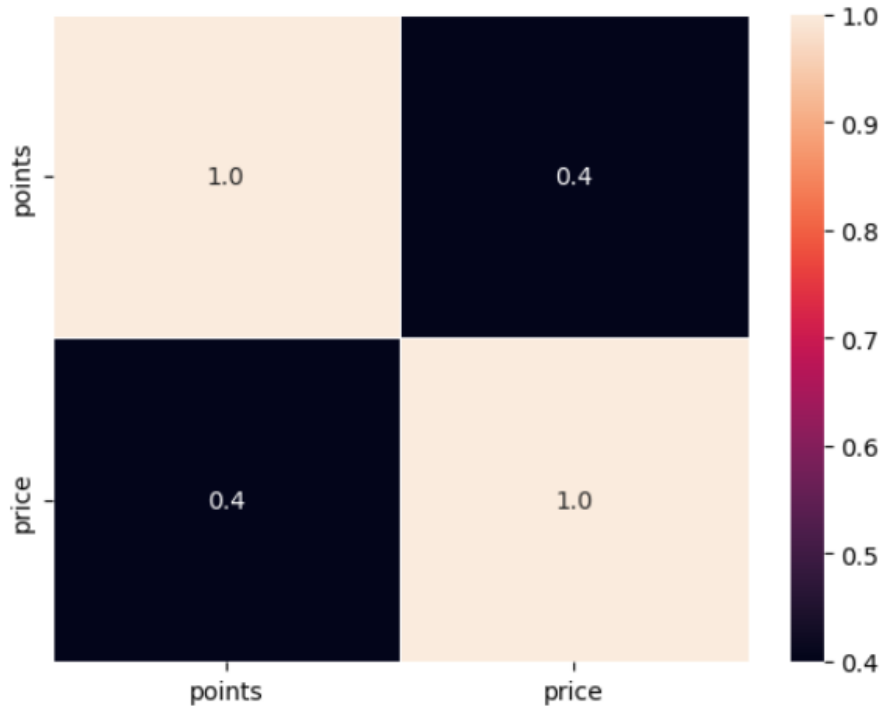


Correlação de Pearson

```
In [67]: correlation_pearson
```

Out[67]:

	points	price
points	1.000000	0.399231
price	0.399231	1.000000



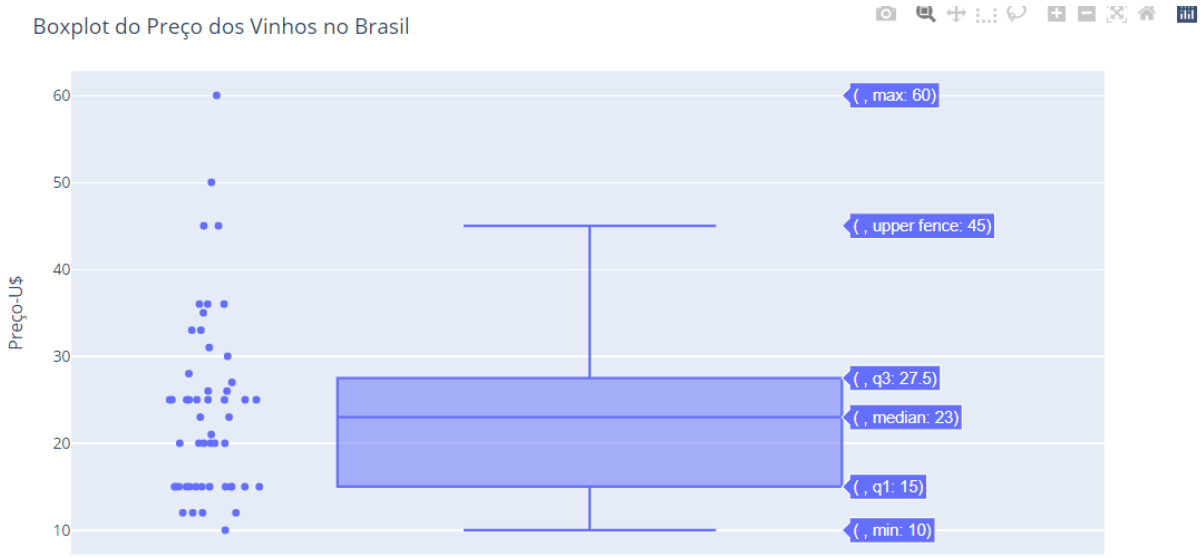
O Brasil

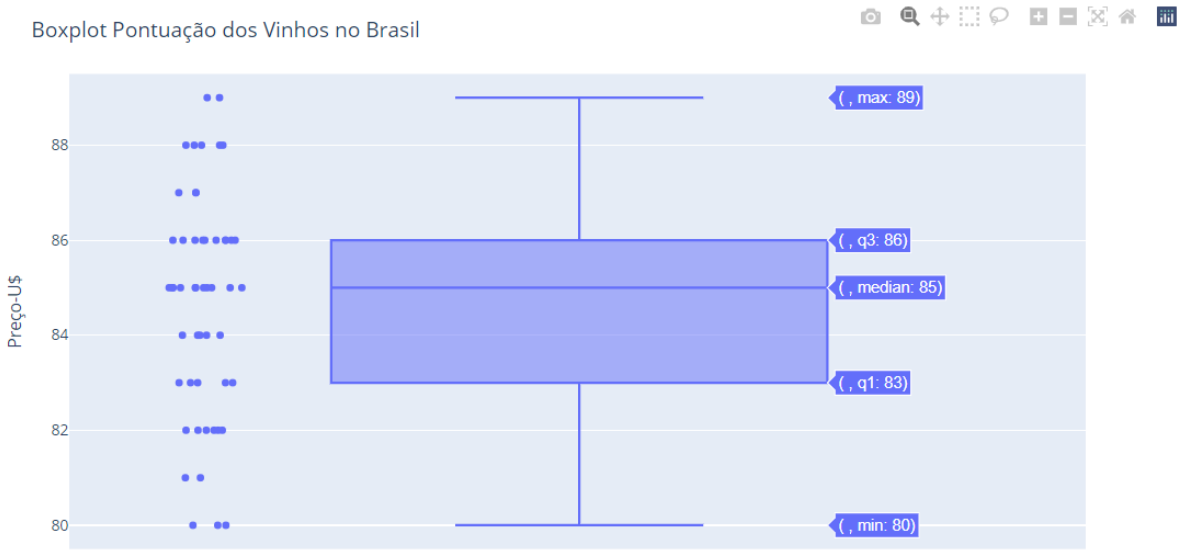
Medidas Resumo - Brasil x Medidas Resumo - Global

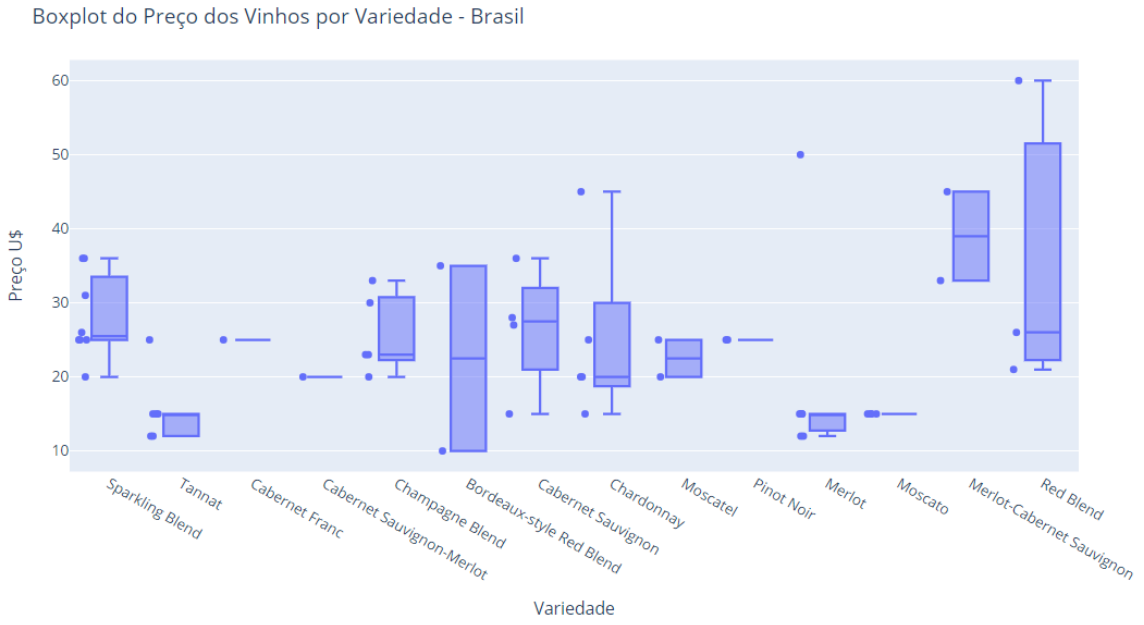
	points	price
count	52.000000	52.000000
mean	84.673077	23.884615
std	2.340782	10.504255
min	80.000000	10.000000
25%	83.000000	15.000000
50%	85.000000	23.000000
75%	86.000000	27.250000
max	89.000000	60.000000
median	85.0000	23.00000

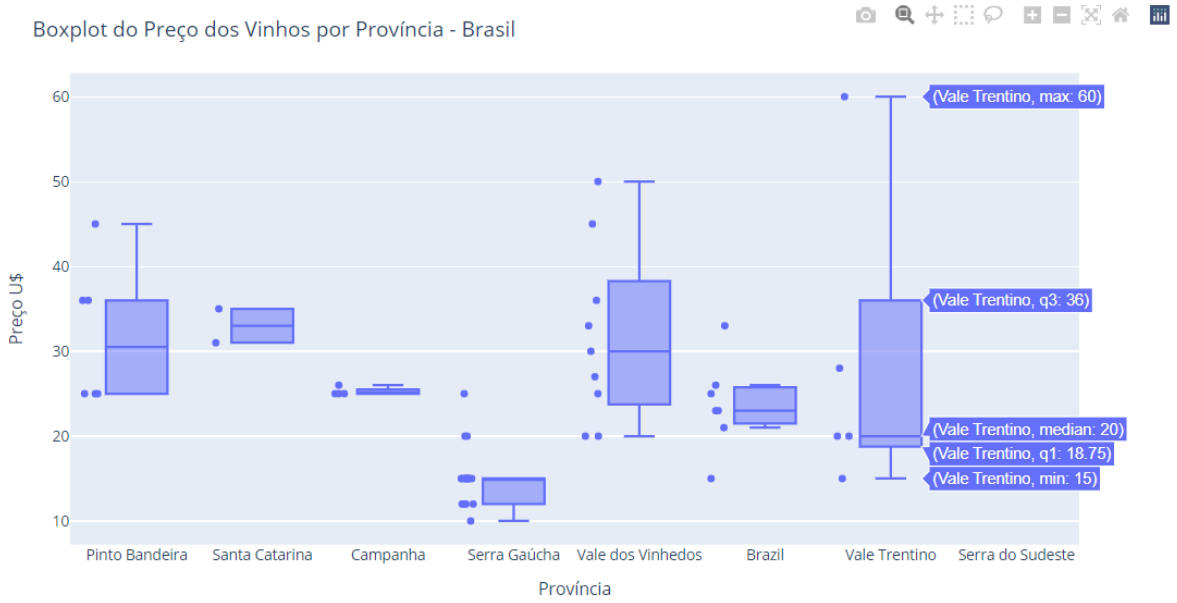
	points	price
count	129971.000000	129971.000000
mean	88.447138	34.646083
std	3.039730	39.664385
min	80.000000	4.000000
25%	86.000000	18.000000
50%	88.000000	25.000000
75%	91.000000	40.000000
max	100.000000	3300.000000
median	88.0000	25.00000

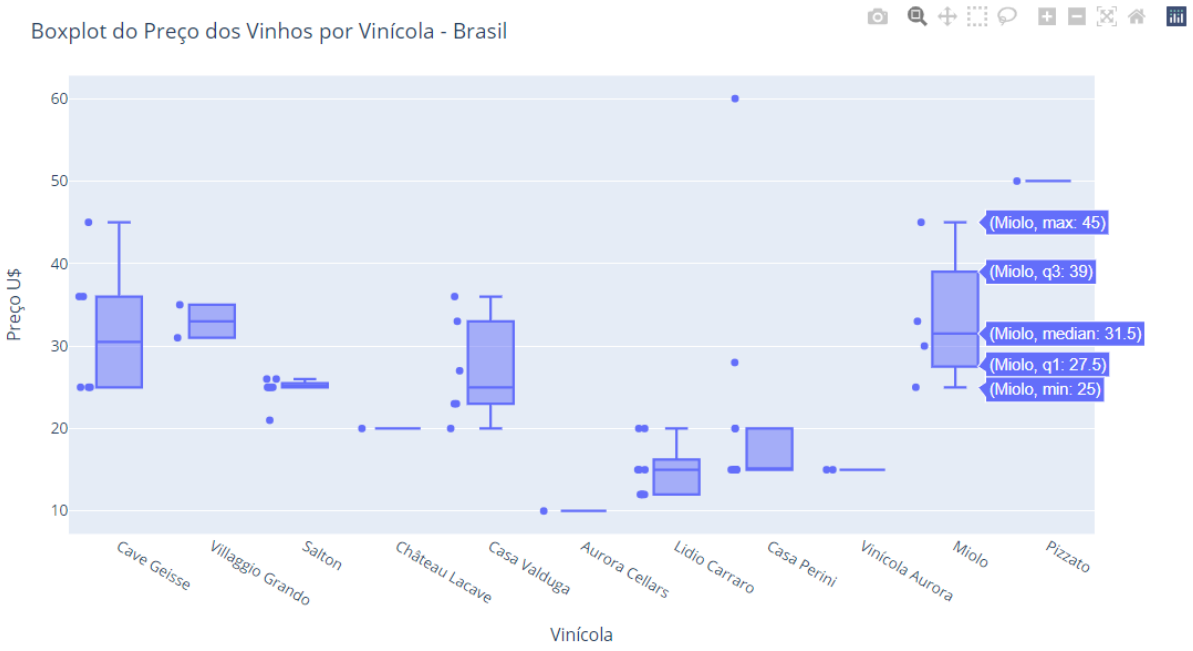
Boxplot Brasil



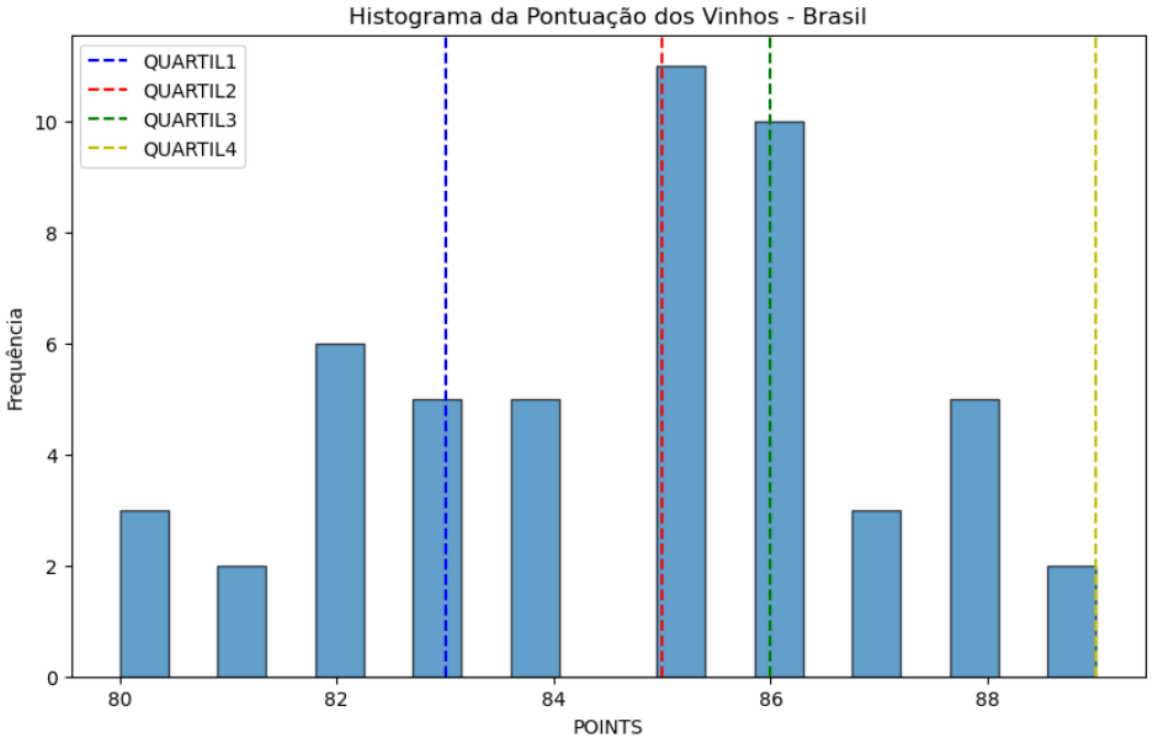



Boxplot do Preço dos Vinhos por Variedade - Brasil
Preço U$
60
50
40
30
20
10
Sparkling Blend
Tannat
Cabernet Franc
Cabernet Sauvignon-Merlot
Champagne Blend
Bordeaux-style Red Blend
Cabernet Sauvignon
Chardonnay
Moscatel
Pinot Noir
Merlot
Moscato
Merlot-Cabernet Sauvignon
Red Blend
Variedade

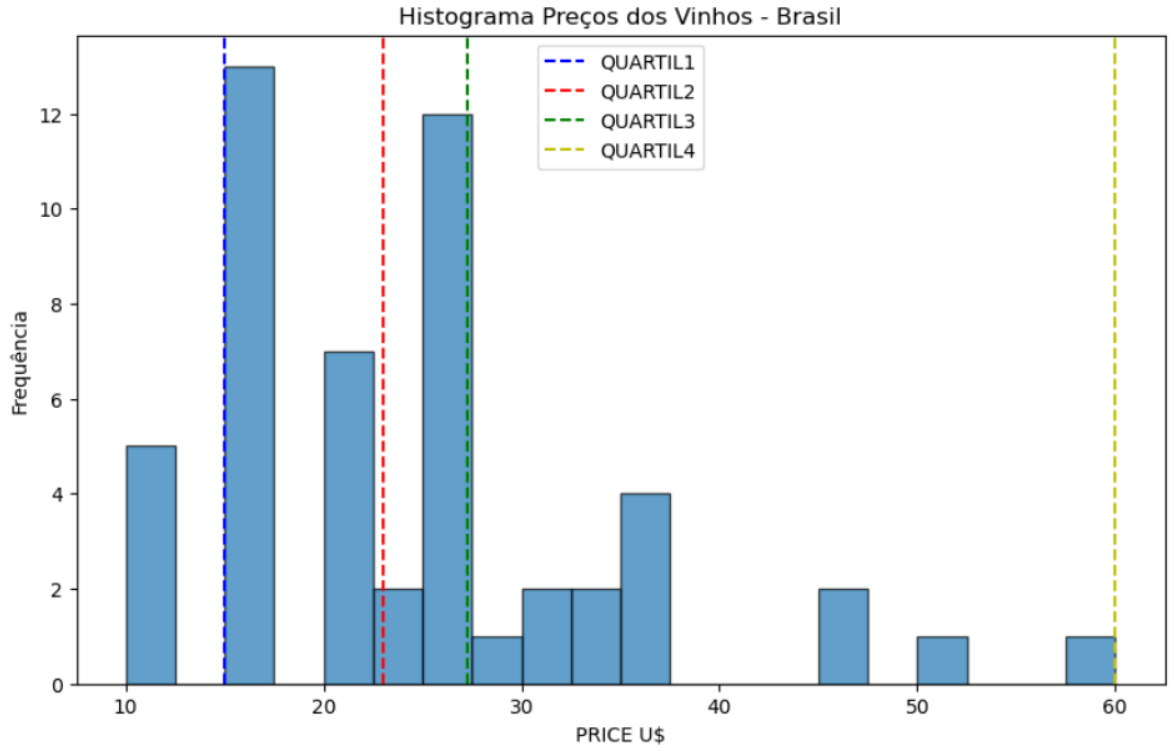

Boxplot do Preço dos Vinhos por Província - Brasil
Preço U$
60
50
40
30
20
10
(Vale Trentino, max: 60)
(Vale Trentino, q3: 36)
(Vale Trentino, median: 20)
(Vale Trentino, q1: 18.75)
(Vale Trentino, min: 15)
Pinto Bandeira
Santa Catarina
Campanha
Serra Gaúcha
Vale dos Vinhedos
Brazil
Vale Trentino
Serra do Sudeste
Província



Histograma - Brasil



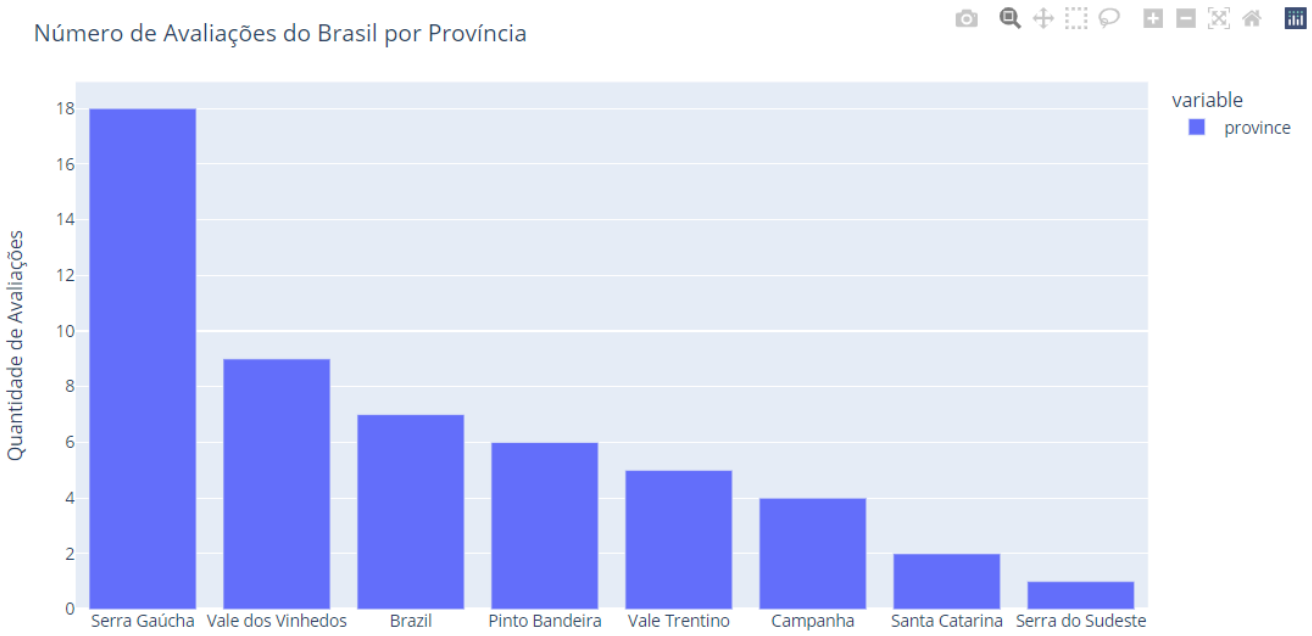

```
Primeiro Quartil (QUARTIL1): 83.0
Segundo Quartil (QUARTIL2): 85.0
Terceiro Quartil (QUARTIL3): 86.0
Quarto Quartil (QUARTIL4): 89.0
```



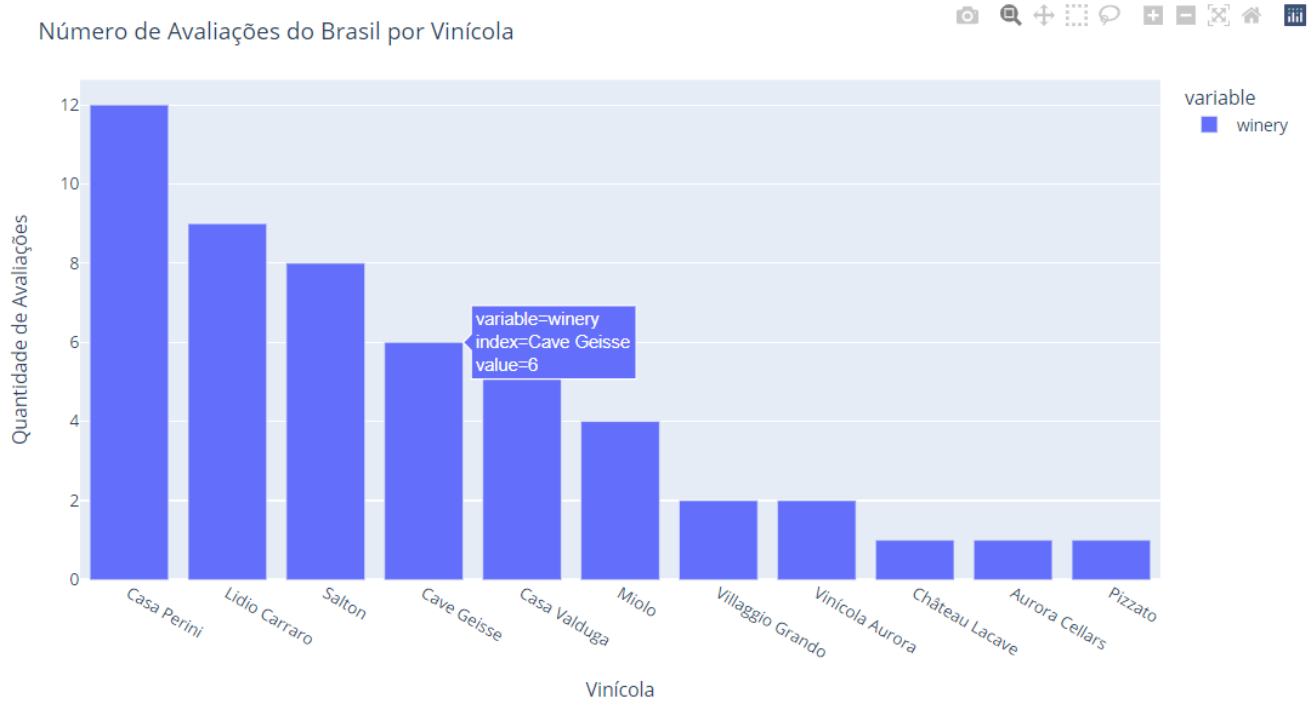

```
Primeiro Quartil (QUARTIL1): 15.0
Segundo Quartil (QUARTIL2): 23.0
Terceiro Quartil (QUARTIL3): 27.25
Quarto Quartil (QUARTIL4): 60.0
```

BarPlot (Gráfico de Barras)

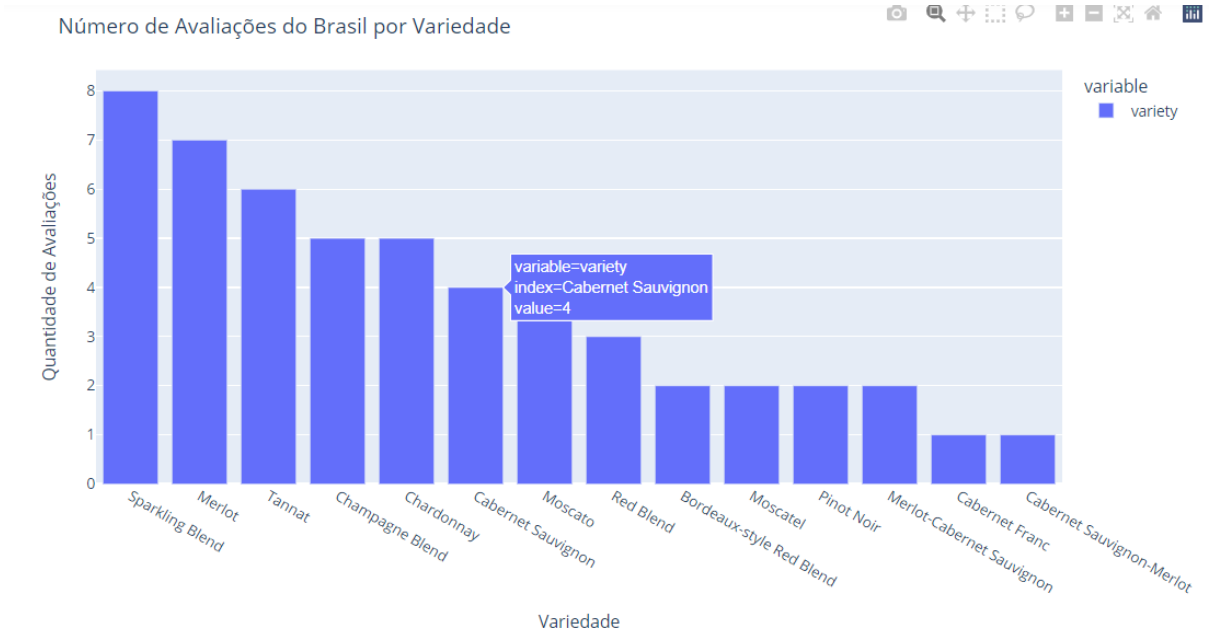
Número de Avaliações por Província - Brasil



Número de Avaliações por Vinícola - Brasil



Número de Avaliações por Variedade - Brasil



Número de Avaliações por Testador - Brasil

Preparação dos Dados

Limpeza e Formatação dos Dados

Valores Nulos (Faltantes)

Para a coluna *price* foi imputado o valor da mediana para as linhas com valores nulos.

Coluna	**Nº de Linhas com Valores Nulos**
country	63
description	0
designation	37465
points	0
price	8996
province	63
region_1	21247
region_2	79460
taster_name	26244
taster_twitter_handle	31213
title	0
variety	1
winery	0
Classificacao	0

Categorização de Variáveis Numéricas

Foi criada coluna que atribui uma **classificação** da qualidade do vinho, conforme a faixa de pontuação (*points*):

50-59 -> Inapropriado para Consumo
60-69 -> Abaixo da Média
70-79 -> Mediano e Inócuo
80-89 -> Acima da Média
90-95 -> Excelente
96-100 -> Extraordinário

Seleção das Variáveis

Information Value (IV)

O cálculo do IV considerou todas as variáveis < 0.02 (0.0). Desta forma não seriam válidas para utilização em modelos preditivos para o preço ou pontuação dos vinhos *considerando a amostra analisada de 129971 registros.*

```
< 0.02, não deve ser usado para previsão
0.02 - 0.1, preditor fraco
0.1 - 0.3, preditor médio
0.3 - 0.5, preditor forte
> 0.5, parece bom demais para ser verdade
```

Target: price

```
iv, woe = iv_woe(data = df_filtered, target = 'price')
```

```
Information value of country is 0.0
Information value of description is 0.0
Information value of designation is 0.0
Information value of points is 0.0
Information value of province is 0.0
Information value of region_1 is 0.0
Information value of region_2 is 0.0
Information value of taster_name is 0.0
Information value of taster_twitter_handle is 0.0
Information value of title is 0.0
Information value of variety is 0.0
Information value of winery is 0.0
Information value of Classificacao is 0.0
```

Target: points

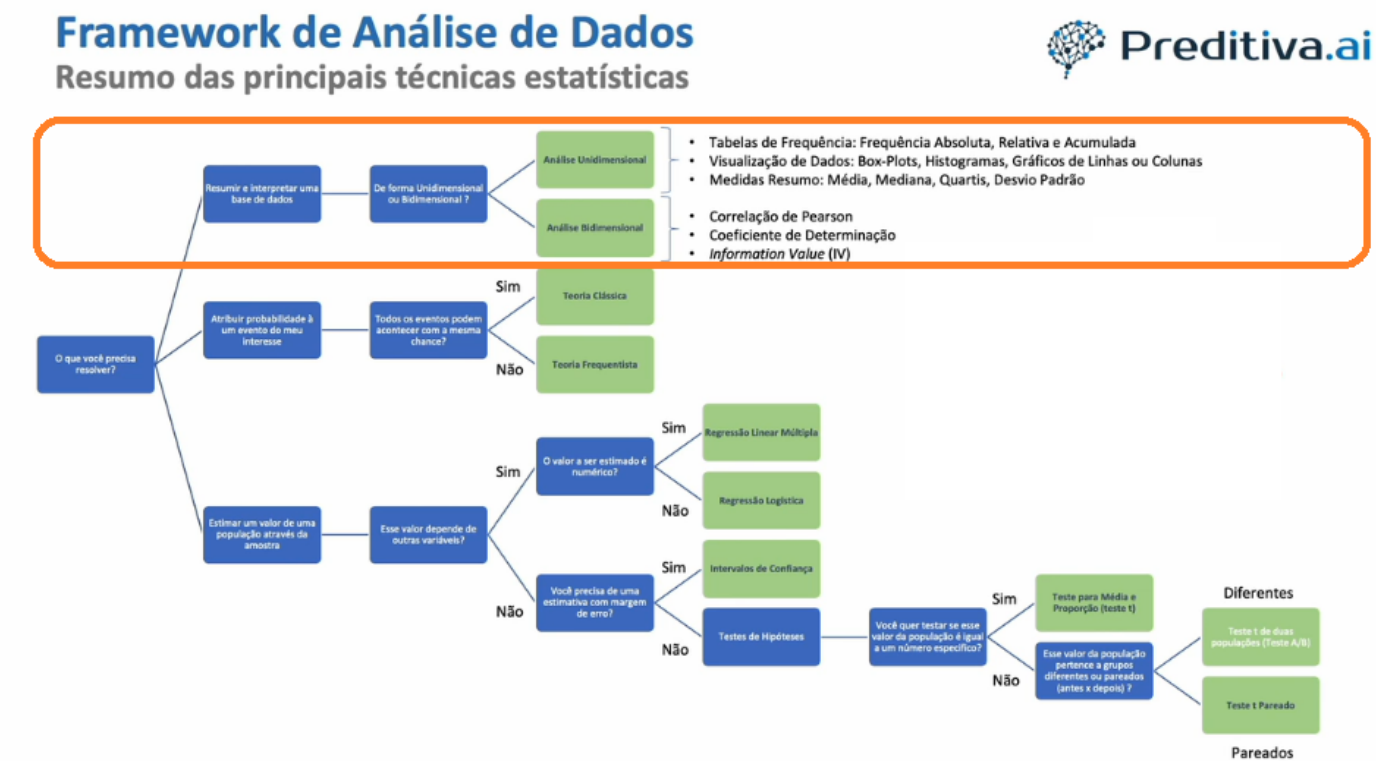
```
: iv, woe = iv_woe(data = df_filtered, target = 'points')
```

```
Information value of country is 0.0
Information value of description is 0.0
Information value of designation is 0.0
Information value of price is 0.0
Information value of province is 0.0
Information value of region_1 is 0.0
Information value of region_2 is 0.0
Information value of taster_name is 0.0
Information value of taster_twitter_handle is 0.0
Information value of title is 0.0
Information value of variety is 0.0
Information value of winery is 0.0
Information value of Classificacao is 0.0
```

Desenvolvimento do Estudo

Técnicas Estatísticas Utilizadas

Considerando que o propósito deste trabalho é resumir e interpretar dados para conhecê-los e entendê-los, utilizamos a estatística descritiva nas análises unidimensionais e bidimensionais. Na imagem abaixo, considerando o *framework* de análise de dados, as técnicas estatísticas utilizadas estão ilustradas no retângulo laranja.



Ferramentas Utilizadas

Linguagem Phyton.

Bibliotecas: pandas, numpy, seaborn, sklearn.linear_model, sklearn.metrics, plotly.express. matplotlib.pyplot

Validação do Estudo

Verificação dos Critérios de Sucessos

Nº	Item	Check
1.	Identificar a concentração e distribuição dos dados com relação ao seu preço.	**OK**
2.	Identificar a concentração e distribuição dos dados com relação ao seu preço.	**OK**
3	Obter as frequências das observações pelos seus atributos de país, província, vinícola, classificação, pontuação, avaliador/testador do vinho, variedade e preço.	**OK**
4.	Estabelecer a correlação entre as variáveis numéricas e determinar se as variáveis categóricas possuem influência na precificação e pontuação dos vinhos.	**OK**
6.	Verificar como o Brasil está posicionado com relação às avaliações que foram submetidas.	**OK**
7.	Verificar se as técnicas foram corretamente aplicadas	**Aguardando Review**
8.	Concluir se o estudo foi válido para obter o resumo, interpretação e conhecimento dos dados.	**Aguardando Review**
9.	Obter, no mínimo, 5 revisões do estudo realizado. (Reviews)	**Aguardando Review**
10	Publicar estudo final no Linkedin e GitHub no Portfólio com as revisões.	**Aguardando Review**

Conclusão

O estudo conclui que a amostra analisada traz uma forte concentração de avaliações nos **US** (41% das avaliações). A seguir temos **França** (17%) e **Itália** (15%). O **Brasil** representa **0.04%** das avaliações, com apenas **52** observações. Todas as 52 avaliações do **Brasil**, tiveram classificação *Acima da Média* (pontuação entre 80 a 89). A **Casa Perini** na **Serra Gaúcha** obteve maior quantidade de avaliações.

Esse viés de seleção, com as observações altamente concentradas em um único país (**US**), faz com que a chance algum tipo de informação seja maior de aparecer neste país. Citamos como exemplo, a frequência da **classificação** do vinho como **extraordinário**, onde obtemos um total de **881** avaliações (0,6%). A concentração de vinhos avaliados como extraordinários está no **US**, seguido da **França** e **Itália**.

As variedades de vinhos mais avaliadas estão concentradas no **Pinot Noir** (10% - 13272 avaliações), seguido do **Chardonnay** (9% - 11753 avaliações) e **Cabernet Sauvignon** (7% - 9472 avaliações) .

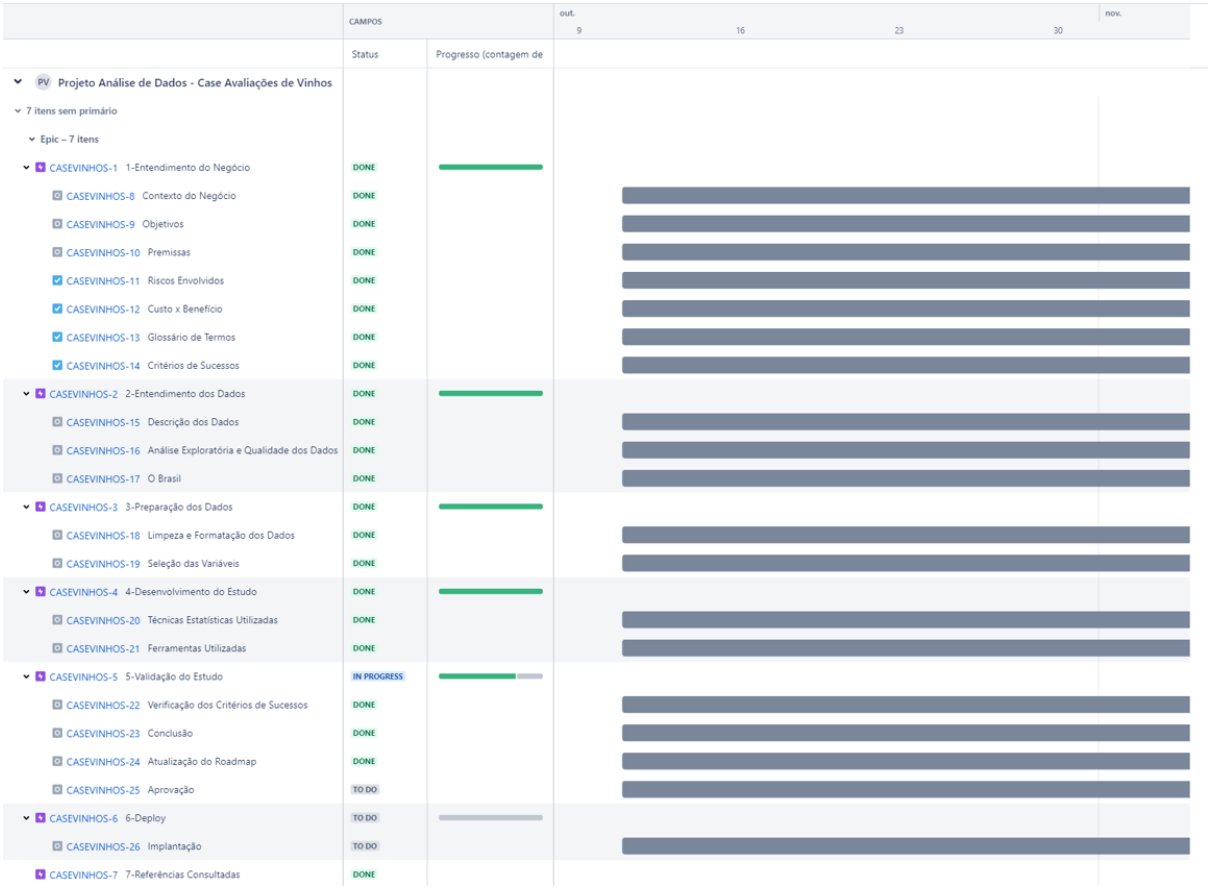
O **preço** dos vinhos possui valores extremos de US$ 4.00 (menor) e US$ (3300) (maior valor). A mediana ficou em US$ 25.00 considerando toda a amostra. A análise de correlação entre a pontuação e preço resultou ser muito fraca nos métodos apurados. Desta forma não podemos afirmar ou inferir, nesta amostra avaliada, que quanto maior a pontuação de um vinho maior será o seu preço ou vice versa, podendo existir outras variáveis que precisam ser analisadas para inferência de preço.

Corroboramos esta informação olhando uma referência citada na WineEnthusiast, apresentando o TOP 100 BEST BY de 2023, onde uma equipe avaliou, as cegas, mais de 30.000 vinhos. *“O vinho nº 1 do TOP 100 é um Syrah de 93 pontos e US$ 15 amplamente disponível da vinícola J. Lohr da Califórnia, sobre o qual o revisor Matt Kettmann diz: “Este é um Syrah muito satisfatório por um preço bastante impressionante””.*

Traços comparativos podem ser melhor aplicados considerando por país ou por conjunto de países que detêm características e frequências similares.

De acordo com o propósito deste trabalho, conclui-se que obtivemos as respostas necessárias para o resumo, interpretação e conhecimento das informações.

Atualização do Roadmap



Deploy

Implantação

Portfólio deste Estudo em:

https://github.com/Carla-G-B-Teixeira/Portfolio/tree/Portfolio/Estudos%20de%20Caso/Vinhos

Referências Consultadas

Imagem da Capa e Cabeçalho:
https://br.freepik.com/vetores-gratis/colecao-de-respingo-de-vinho-com-imagens-realistas-isoladas-de-bando-de-garrafas-de-vinho-tinto-de-ilustracao-de-uvas-e-oculos_6852160.htm

Imagem

Imagem de macrovector no Freepik

WineEnthusiast:
https://www.wineenthusiast.com/toplists/best-buys-2023/

Kaggle: https://www.kaggle.com/datasets/zynicide/wine-reviews

***Divino**: https://www.divvino.com.br/blog/pontuacao-de-vinhos/*

***Preditiva**: https://ead.preditiva.ai/53798-7-metodo-crisp-dm*