UNIVERSIDADE DE SÃO PAULO – USP

INSTITUTO DE QUÍMICA DE SÃO CARLOS - IQSC

A UTILIZAÇÃO DE CONCEITOS DAS CIÊNCIAS DE DADOS: UMA ANÁLISE SENSORIAL DA BEBIDA DE CAFÉ E SUAS CORRELAÇÕES

BRUNO PEREIRA DE OLIVEIRA

São Carlos

2022

BRUNO PEREIRA DE OLIVEIRA

A UTILIZAÇÃO DE CONCEITOS DAS CIÊNCIAS DE DADOS: UMA ANÁLISE SENSORIAL DA BEBIDA DE CAFÉ E SUAS CORRELAÇÕES

Monografia apresentada ao Instituto de Química de São Carlos, da Universidade de São Paulo, como um dos requisitos para obtenção do título de Bacharel em Química.

Orientador: Prof. Dr. Álvaro José dos Santos Neto

São Carlos

2022

Dedicatória

Dedico primeiramente à Deus e Arcanjo Miguel, por me propiciar a vida e aos meus pais, Selma Lopes e Adilson, e aos meus avós, Adalgiza e Abrão, que sempre me apoiaram em todos os meus objetivos e sonhos, e até nos momentos mais complexos da minha vida eles me incentivavam: - “Continue, pois está quase chegando o sucesso”.

Sempre fizemos da nossa base o zero absoluto, todavia carregados de amor e compreensão, mesmo eles não entendendo exatamente o que busco, acreditaram em mim e me apoiaram. Sem eles, eu não sei mesmo se teria conseguido. Minha eterna gratidão!

Agradecimentos

Ao Prof. Dr. Álvaro José dos Santos Neto, pela orientação, apoio, suporte e por oferecer a oportunidade deste trabalho ser realizado.

A minha esposa Jade, pela ajuda, incentivo, motivação e por todas as conversas, palavras de conforto e amizade com amor oferecida.

Aos técnicos, professores, bibliotecárias e demais funcionários do Instituto de Química de São Carlos que, de alguma forma, contribuíram para que esse trabalho pudesse ser realizado.

A minha família que desde pequeno me ensinaram a simplicidade e o esforço para conquistar os objetivos planejados.

Aos amigos mais próximos feitos durante a graduação, gratidão pelo apoio e por oferecerem sua amizade sempre que necessário.

E a todos que, de alguma forma, torceram e contribuíram para a elaboração desse trabalho.

Siga sempre em frente com fé e estoicismo,

pois quem te protege não dorme,

o que se imagina se cria,

o que acredita se faz!

São Miguel Arcanjo

Resumo

O café é consumido por 98% dos brasileiros e é considerado a segunda bebida mais consumida no mundo e o produto mais negociado na bolsa (o Brasil é um dos maiores e mais importante produtores de café, tanto no quesito consumo como exportação do grão cru beneficiado).

Há uma complexidade na cadeia produtiva, pois o consumidor busca bebidas que remetem lembranças e boas sensações durante o consumo de um xícara de café. O fato é que para a manutenção e busca de uma padronização dos *blends* comercializados no mercado é necessária uma avaliação degustativa sensorial e a utilização do protocolo desenvolvido por *SCAA protocol method*, avaliando onze requisitos qualitativos e quantitativos da bebida café, na soma final dos quesitos, quanto maior for a pontuação, é considerado melhor o café.

As avaliações são realizadas por degustadores e a conjectura palatável para esta análise é extremamente pessoal, mesmo para profissionais muito experientes. Com isso o presente trabalho avaliou dados sensoriais utilizando de ferramentas na linguagem *python* em um conjunto de amostras sensoriais do café de 36 países diferentes e 1.113 linhas de dados das mais diversas origens, regiões, perfis etc.

Contudo estabeleci uma correlação estatística das variáveis sensoriais, e neste caso se definiu cinco requisitos que impactam diretamente a pontuação final do café que foram avaliados nas seguintes exigências: aroma, equilíbrio, acidez, corpo e doçura.

É ressaltado que a altitude não apresentou influência de correlação direta no teste de *heatmap*. A outra análise realizada –ANOVA- para a comparação das amostras frente a sua origem, constatou que os cafés variam a pontuação de maneira significativa conforme sua origem. Por exemplo: o café brasileiro é diferente do café da etiópia.

Em suma, um fato importante é que o consumidor tem uma análise de qualidade diferente, no quesito aroma, do que se é aplicada no SCAA, e para os especialistas é o quarto item no grau de importância na avaliação profissional.

Palavras chave: Café, ANOVA, Arábica, Conilon

ABSTRACT

Coffee is consumed by 98% of Brazilians and is considered the second most consumed beverage in the world and one of the most traded on the stock exchange, and Brazil is one of the most important coffee producers, both in terms of consumption and exportation of the processed raw bean. There is a complexity in the production chain, as the consumer looks for beverages that bring back memories and good feelings during the consumption of a cup of coffee.

The fact is that for the maintenance and standardization of the blends sold in the markets, a sensorial evaluation is required, which is a protocol applied by the SCAA protocol method, where 11 qualitative and quantitative coffee requirements are evaluated and a score is generated. The higher the score, the better the coffee is considered.

However, the evaluations are carried out by tasters, and the palatable conjecture for this analysis is extremely personal, even for very experienced professionals. With this in mind, the present study aims to apply chemometric tools and data science concepts to a set of sensory samples of coffee in which 36 countries have 1,113 samples of the most diverse varieties of coffee.

Thus, through the statistical correlation of variables showed that 5 requirements directly impact the final score of the coffee which were, Aroma, Balance, Acidity, Flavor, Body and sweetness presented as a neutral requirement.

Another statistical analysis performed is the comparison of the samples in relation to their origin, and it was found that the coffees from each country have a statistically significant variance, that is, it is noticeable when it is structured in the overall score of the product.

Finally, it was understood that the Brazilian consumer seeks as the main quality factor the aroma, and for the specialists this is the fourth most important item in the professional evaluation.

Keywords: Coffee, ANOVA, Arabica, Conilon

Lista de Figuras

Lista de Tabelas

Sumário

1. Introdução

No século XIX o café inicia a sua importância no desenvolvimento econômico brasileiro e durante a evolução temporal ocorreram diversas buscas e mudanças nas percepções do consumidor. O fato é que esta bebida hoje tem 98% dos brasileiros adeptos ao consumo.

A bebida está inserida em nosso cotidiano mesmo com a maior crise e elevação dos preços no ano de 2021 (Figura 1) e o consumo cresceu 1,7% tendo como quantificação 21,5 milhões de sacas de 60 Kg. (Agro, 2022)

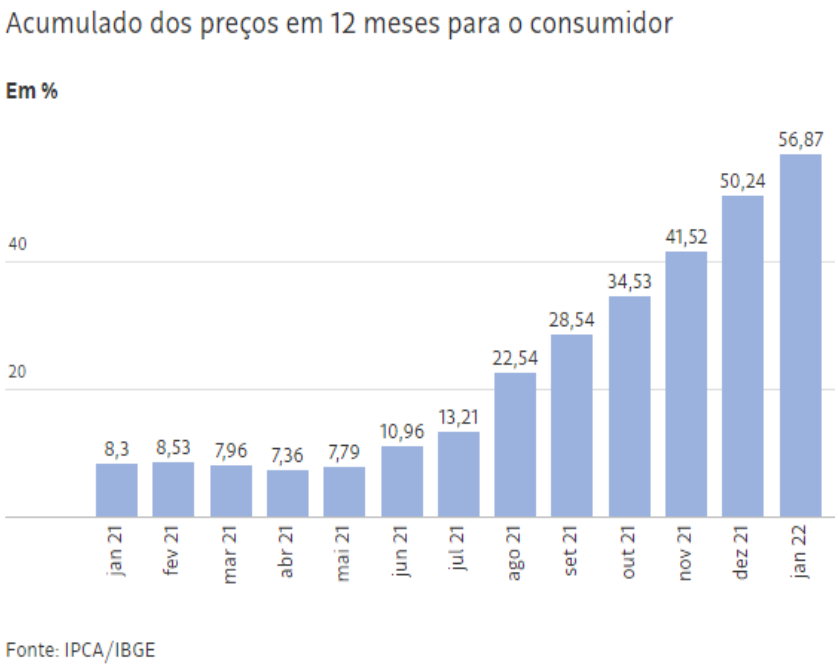


Figura 1: Gráfico dos aumentos de preços do café no ano de 2021 até janeiro de 2022, é considerado um dos piores cenários de todos os tempos.

Fonte: *(Vieceli, 2022)*

Para um breve entendimento a composição da bebida, segundo a portaria DAS nº 570 de 9 de maio de 2022, descreve que a constituição da bebida é pela utilização do grão beneficiado das espécies do gênero *coffea*

(Figura 2) e dentro deste gênero as espécies produzidas se dividem em duas: Arábica e Robusta/conilon *(MAPA, 2022)*.

Figura 2: Café cru já beneficiado em fase final de limpeza e retirada de impurezas mais leves.

Fonte: Próprio autor

Um dos pilares centrais é um rígido controle de qualidade para a garantia de padronização da bebida e uniformidade do produto, visto que é de origem vegetal. Assim, para uma uniformidade no fator sensorial se utiliza a metodologia de *SCAA CUPPING*, que consiste em um formulário de degustação que contêm 11 atributos avaliativos do café: Fragrância/Aroma, Uniformidade, Ausência de defeitos (xícara limpa), Doçura, Sabor, Acidez, Corpo, Finalização, Equilíbrio, Defeitos e Avaliação global (SCAA Cupping protocols, 2008).

O protocolo de teste SCAA segue os passos (Figura 3):

1 – Padronização da amostra;

2 – Torra do café: Torra média clara;

3 – Medição da Escala de cor na AGTRON-SCAA;

4 – Perfil granulométrico: Café em pó passa na peneira de *mesh* 20;

5 – Água: Mineral (pH<7, Sais < 200 ppm, T~ 93ºC);

6 – Avaliação sensorial: pós-hidratação.

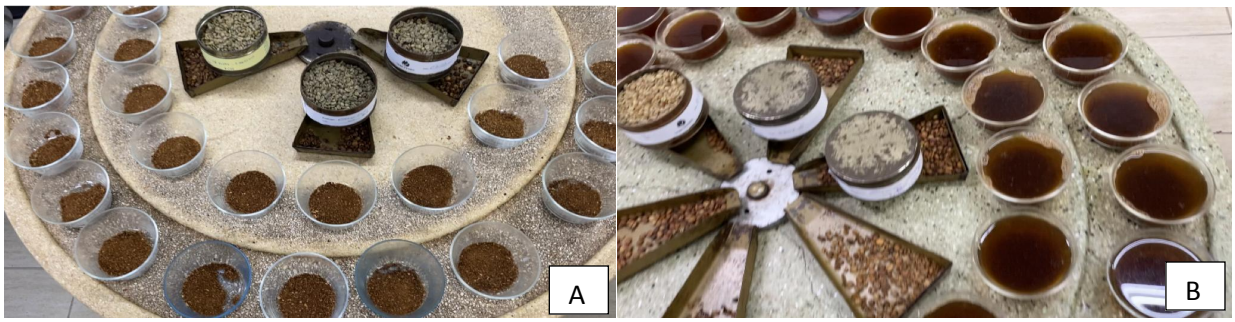


Figura 3: As amostras depois de torradas são moídas para análise. A-) Elas são inseridas em 5 xícaras de testes em porções igualmente pesadas para a difusão de água; B-) É adicionada água a temperatura de aproximadamente 93ºC e espera de 4-5 min para infusão do café em pó, torrado e moído, seguido de teste sensoriais.

Fonte: Próprio autor.

Uma exemplificação da ficha utilizada na análise sensorial do café (Figura 4), descrevendo os pontos a serem avaliados.

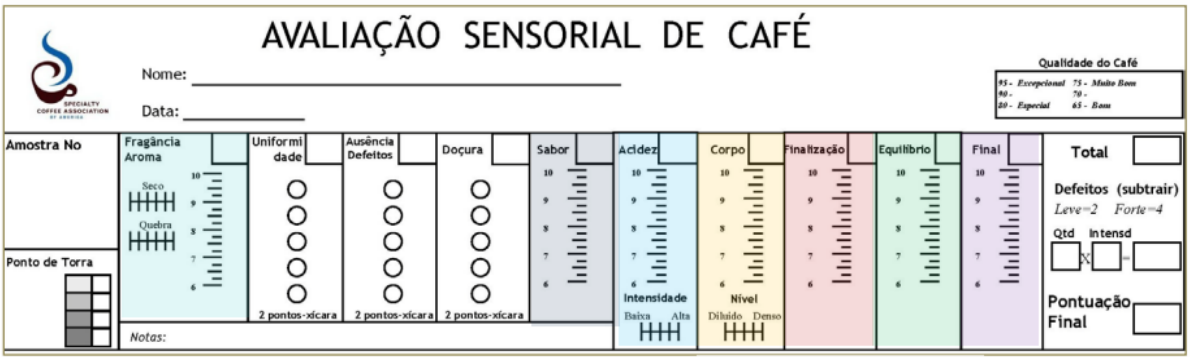


Figura 4: Exemplo de ficha de preenchimento da análise sensorial do café em avaliação de qualidade e sabores para a composição da pontuação final.

Fonte: *(SCAA Cupping protocols, 2008)*

Após o preenchimento da avaliação sensorial em triplicata, normalmente com a presença de 3 diferentes degustadores, faz a somatória da pontuação do café, portanto definindo a sua posição de qualidade e classificando como: cafés especiais, cafés comerciais finos, cafés comerciais, cafés inferiores (Figura 5).

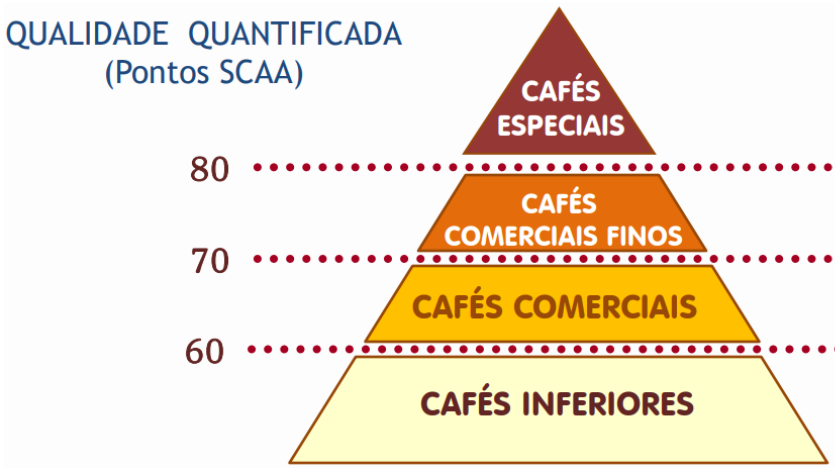


Figura 5: A pontuação adquirida pelo lote faz com que este tenha o seu posicionamento na escala de pontos SCAA da pirâmide de qualidade.

Fonte: (*SCAA Cupping protocols,* 2008*)*

Desta forma se classifica amplamente as bebidas produzidas e a sua forte variação na percepção dos sabores e sensações devido aos degustadores que obtém as experiências, conforme as percepções construídas no tempo. Dessa forma o mesmo lote de amostra de café, avaliado em regiões diferentes e com grupos de especialistas diferentes, tenderão a obtenção de resultados diferentes. Uma vez que a compreensão da qualidade está diretamente correlacionada com a experiência do degustador e também susceptíveis as variações dos mais diversos aspectos degustáveis e palatável.

2. Objetivo

A obtenção de um conjunto de dados e a utilização de testes estatísticos de correlação entre as variáveis (*heatmap*) e também de teste de significância através da ANOVA para comparação sensorial dos cafés de diferentes países.

3. Materiais e Métodos

Buscou-se um conjunto de dados no *Kaggle Dataset, The World Bank*, em que contêm medidas de qualidades tais como: aroma, sabor, retrogosto, acidez, corpo, equilíbrio, uniformidade, limpeza do copo, doçura, umidade e defeitos. Os meta dados presentes são: método de processo, cor, espécie, proprietário, origem, norma da fazenda, número de lote, companhia, altitude e região do café e o formato dos documentos são valores separados por vírgulas (CSV).

Os arquivos foram importados e interpretados, consequentemente os termos tiveram a tradução dos dados da língua inglesa para portuguesa. Após está etapa foi processado de maneira independente os dados e se fez o teste *heatmap* e ANOVA.

Os dados foram obtidos do *Coffee Quality Institute (CQI)* ano de 2018, com a licença de utilização *Open Data Commons*, que é uma ferramenta de dados legalizada e com confiabilidade garantida pela *Open Knowledge Foundation*.

As análises foram desenvolvidas na plataforma *Anaconda*, utilizando o compilador *Spyder* (3.8) na linguagem de programação *python 3.8*. Todos os códigos desenvolvidos e utilizados estão nos anexos.

Na aplicação da linguagem *python* foi utilizado a biblioteca de cálculo pandas e também as funções estatísticas *seaborn*, que são pacotes com constituições matemáticas validadas de diversos métodos estatísticos, tendo como aplicação o desenvolvimento da estruturação do programa a ser compilado, sem a necessidade de desenvolvimento de todas as equações do zero.

4. Resultados e Discussão

Inicialmente é de senso comum que o desenvolvimento de um produto a base de café se tem como pilar central a sua qualidade e padronização do produto a ser fabricado, dessa forma são analisados 11 fatores sensoriais das bebidas de cafés e estas são avaliadas já torradas e infundidas em água a temperatura de aproximadamente 93 ºC.

Para uma compreensão referente as percepções do consumidor, analisando a qualidade sensorial do café e do direcionamento computacional,

é necessário a utilização de uma pesquisa desenvolvida e publicada pela Associação Brasileira de Indústria de Café – ABIC, para que se tenha como base o direcionamento e análises do desenvolvimento do software em questão, pesquisa está realizada em 2020 com um total de 5.460 pessoas mais de 180 mil respostas em 14 munícipios distribuídos nos estados PA, SP, SC, RS, MG, DF, PE, RJ, PR (ABIC, 2021) :

Público Geral: Apreciadores do café comum, não buscam e/ou tenham conhecimentos além do básico do produto.

Entusiastas: Apreciadores de café que tem e/ou busquem mais informações sobre produtos e tem curiosidade no assunto.

Especialistas: Apreciadores de café, que visam estudos formais e/ou trabalham diretamente com o produto e já inicialmente dominam aspectos técnicos.

O resultado final compilado (Figura 6) descreve os atributos percebidos pelos consumidores brasileiros:

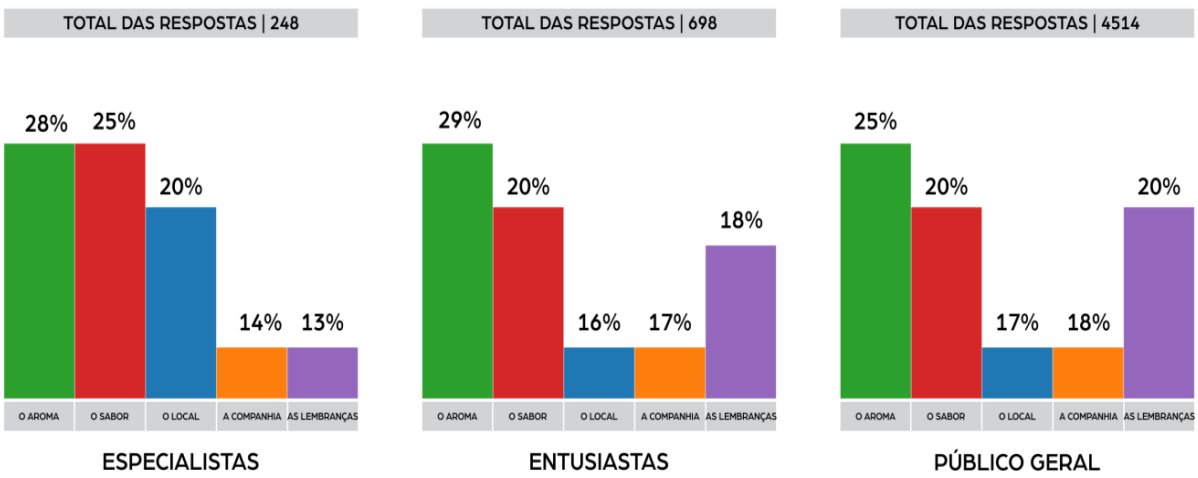


Figura 6: A Percepção do consumidor final de café no quesito percepção de qualidade e o que é avaliado por estes na sua escolha por uma marcar e/ou uma degustação.

Fonte: (ABIC, 2021)

Observamos (Figura 6) que a percepção do consumidor em qualidade da bebida de café são: aroma, em primeira colocação como item de maior

percepção, e em segundo ponto é o sabor, ficando na segunda colocação geral. Isto é, o ponto de vista do consumidor está dentro do que é utilizado para avaliação dos cafés pelos especialistas que aplicam o método SSCA *cupping*.

A constituição do sabor é através do uso da língua que se tem terminações nervosas ligadas ao cerebelo e o aroma através do uso do nariz que está ligado as terminações nervosas no córtex olfatório (Figura 7).

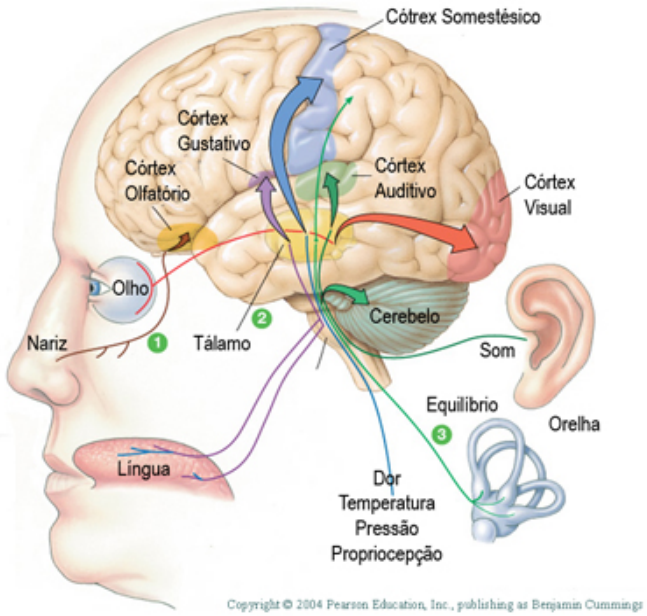


Figura 7: Configuração dos sentidos sensoriais do corpo humano.

Fonte: (Nishida, 2022)

O fato é que o primeiro contato de qualquer alimento que o consumidor percebe é no sentido olfativo, devido os fenólicos, flavonoides e antioxidantes voláteis imersos em bebidas quentes tendem a se volatilizar e fisicamente o nariz é o primeiro por posicionamento frontalmente a face do ser humano.

A produção de matéria prima do café verde e cru são complexas e consequentemente as suas cadeias de beneficiamentos químicos também são, devido a constituição física do fruto e aspectos sensoriais, a composição antes do processo de torração contêm, já identificados, 228 substâncias químicas e outro comportamento após este processo de torração descreve que há 971 substâncias químicas (Tabela 1), formando o aroma do café já torrado (Moon, 2009) (Ruosi, 2021) (Baltes & Mevissen, 1988) (Parliment, 2000).

Tabela 1: Descrição dos compostos químicos que compõem os cafés verdes e após o processo de torração.

Fonte: Adaptado de Parliment 2000.

Classe química	Café Verde	Café Torrado
Hidrocarboneto	41	74
Álcoois	24	20
Aldeídos	32	30
Cetonas	21	73
Ácidos	3	25
Ésteres	26	31
Lactonas	4	3
Fenóis e Éteres	10	48
Furanos	17	127
Tiofenos		26
Pirroles		71
Oxazoles		35
Tiazóis		27
Compostos nitrogenados	38	208
Piridinas		19
Pirazinas		86
Aminas e mistura de compostos nitrogenados		32
Compostos Sulfúricos	7	47
Diversos	5	17
Total	228	791

Um outro viés é que no Brasil as indústrias de café são permanentemente proibidas da utilização de *blend* (liga de café verde, mistura de origens e gênero) de origem de outros países, assim o café produzido no Brasil é produção 100% nacional, colhidos e beneficiados em fazendas e cooperativas, se tornando comum a mistura do gênero Arábica junto ao gênero Conilon nas mais diversas composições de porcentagens entre eles.

Contudo, o mercado está evoluindo e marcas de posicionamento mundial estão focalizando a distribuição de seus produtos em solos brasileiros e apresentado os mais diversos *blends* do mundo, tais como cafés originados da Colômbia, Quênia, Etiópia, Guatemala, EUA, dentre outras.

Existe um segundo movimento que é a apresentação de *blend* de cafés únicos da mesma espécie, como exemplo a espécie 100 % arábica de qualquer região brasileira e está sendo posicionado como o café da mais alta qualidade.

Para a compreensão de toda a pesquisa desenvolvida nestes textos, é fundamental o entendimento dos dados iniciais, sem pré-tratamento de processamento e obtido no *Coffee Quality Institute (CQI)*, no qual é constituído das mais diferentes localidades e com informações obtidas até o ano de 2018.

E as espécies Arábica e Conilon e distribuídas (Figura 8).

Figura 8: As amostras se distribuem em duas espécies Arábica e Conilon

Os dados brutos apresentam em sua totalidade 2,13 % das amostras da espécie Robusta/Conilon e 97,86% são da espécie Arábica. Este é um indicativo de que o consumo mundial tende a maior utilização da espécie Arábica. No Brasil não ocorre essa tendência, pois grandes marcas que estão no mercado trabalham com *blends* contendo acima de 50% da espécie Conilon, e existem poucas marcas em escala nacional que tenham em sua composição 100 % Arábica.

. O ano de colheita dos dados está distribuído da seguinte forma (Figura 9).

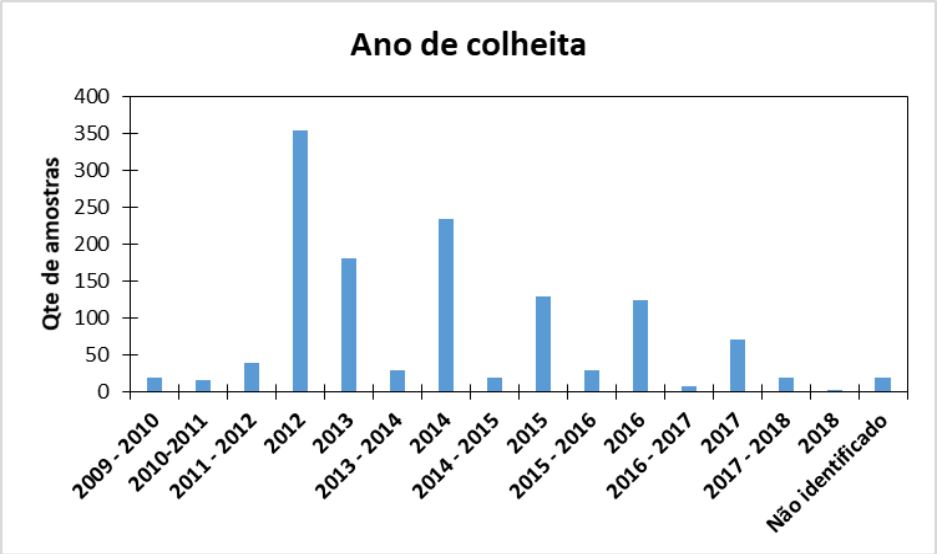


Figura 9: Os dados foram obtidos entre as colheitas de 2009 até 2018 e as amostras analisadas no ano corrente.

Uma vez que os dados estão presentes em quase uma década de informações sensoriais, é remetido que há uma homogeneidade no quesito variação de sazonalidade, pois diminui a influência de um ano considerado verbalmente bom com um ano considerado verbalmente ruim.

As origens das amostras são compostas (Figura 10) por diferentes origens.

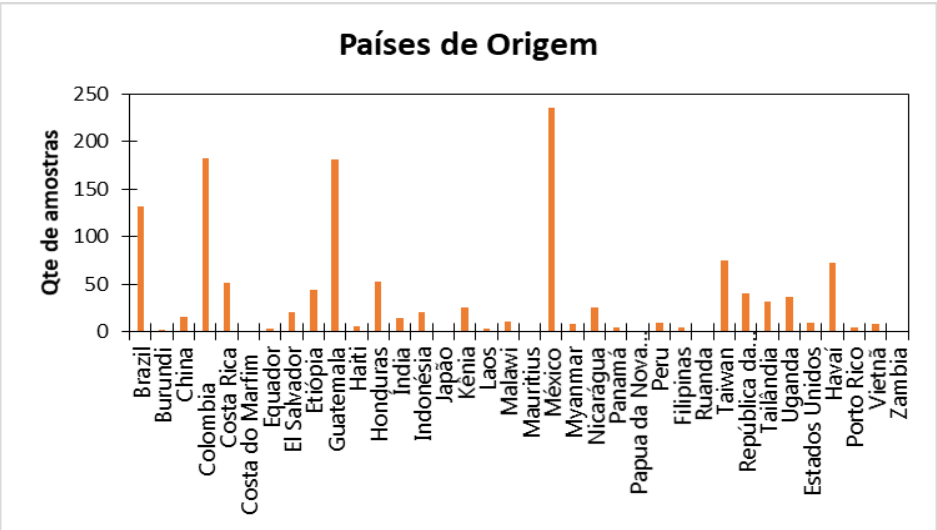


Figura 10: Origem das amostras de cafés avaliadas pelo *Coffee Quality Institute* (CQI).

Assim como resultados temos que o México apresenta o maior volume de 236 amostras, seguido da Colômbia que apresenta 183 amostras,

Guatemala apresenta 181 amostras e por fim o Brasil apresenta 132 amostras, permanecendo na posição 4º dos países com maior número de amostras avaliadas pelo CQI.

É ressaltado nesta discussão que não está se confrontando os dados de exportação fornecida pelo órgão oficial, mas sim que neste conjunto de dados disponibilizados o café brasileiro está na posição supracitada de amostras.

Consequentemente se compilou (Figura 11) as variedades presentes dentro das espécies Arábica e Conilon, sem distinção das espécies.

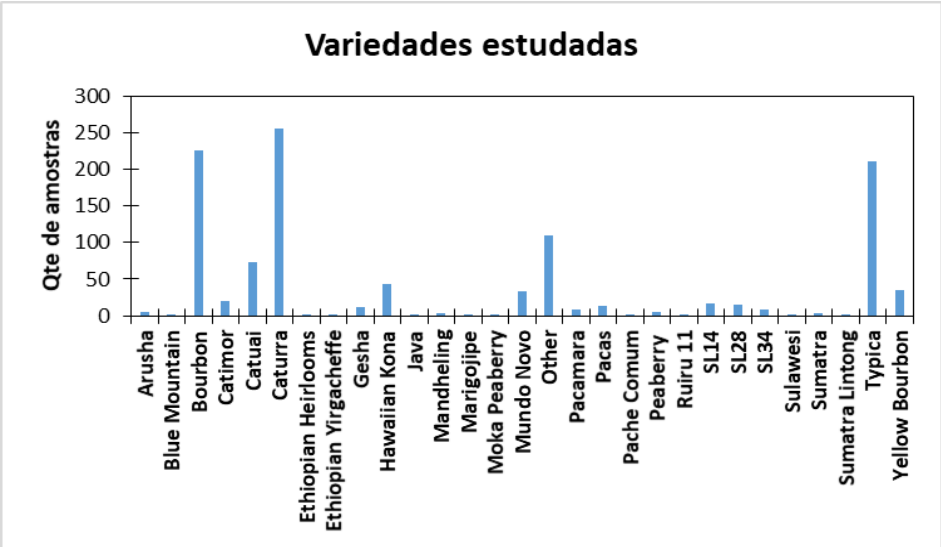


Figura 11: Variedades de cafés presentes nos dados do CQI.

As três variedades de cafés de maior expressão são:

- Caturra: É um variedade de cultivares de porte baixo (Figura 12) e com frutos vermelhos e amarelos, provavelmente originadas de mutações naturais do *Bourbon Vermelho* de alto porte, introduzida e utilizada no Brasil desde 1937 e o Registro Nacional de Cultivares (RNC) em 1999 com a sigla *IAC 477*, ou seja se trata de um café do espécie Arábica (BSCA, 2022).

Figura 12: Á esquerda a imagem do pé de café da variedade caturra amarela, com o fruto desenvolvido. Á direita imagem da mesma variedade com o pé em floração.

Fonte: (Coffee Valore, 2022)

- Bourbon: É uma variedade originada do caturra de porte alto, com sabor adocicado, equilibrado e apresenta dois tipos de frutos: o vermelho e o amarelo (maior valor agregado), e a sua introdução no Brasil foi em 1859 e o aprimoramento genético na década de 1930 (Paiva, 2021).

Figura 13: Imagem de um pé de café da variedade Bourbon vermelha, com fruto já desenvolvido.

Fonte: (Grão Gourmet, 2019)

- Typica: É amplamente conhecida dentre os gêneros do café Arábica do mundo e sua origem se dá na Jamaica na América Central. É uma variedade pai da variedade Mundo Novo (Produzida no Brasil na região mogiana) e Pacamara são plantas de grandes porte e folhagens grandes e a mesma pode ser rastreada até a Etiópia (The Coffee Traveler, 2011).

Figura 14: Imagem da variedade de café typica com frutos vermelhos já maduros

Fonte: *(Perfect Daily grind, 2021)*

Aplicando o tratamento dos dados utilizando a biblioteca pandas e *matplolib* do python com o estilo de visualização dos dados em *boxplot*, teremos a composição total do conjunto de pontos.

Para uma percepção qualitativa, em que os gráficos no eixo Y é a pontuação geral da bebida e no eixo X os requisitos sensoriais avaliados.

Cada box no gráfico contém pontuação (y) e nota especifica (x) de cada quesito e é composto por valor mínimo, valor máximo, mediana (segundo quartil – 50% da amostra), primeiro quartil (25% da amostra inferior) e terceiro quartil (25% da amostra superior) e o exemplo de aplicação é demonstrado na Figura 15.

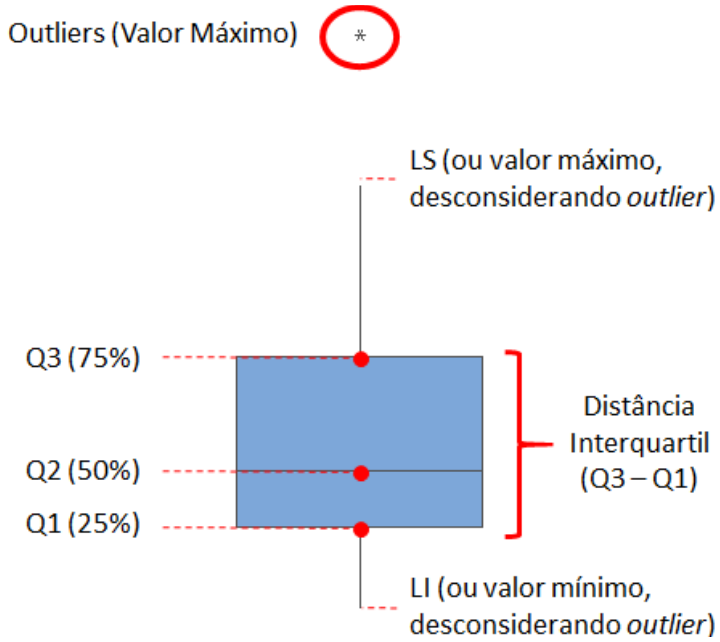


Figura 15: Exemplo explicativo de um gráfico de boxplot e como realizar a leitura dos dados

Fonte: (EDTI be better, 2019)

Os resultados são compilados no decorrer do presente texto e é baseado em análise de gráficos *boxplot*, no qual se perfaz deduções qualitativas e deduções quantitativas.

Neste primeiro momento é apresentado (Figura 16) um conjunto de gráficos com 6 diferentes requisitos de qualidade (Sabor, Aroma, Acidez, Doçura, Equilíbrio). No eixo Y está o valor da pontuação e no eixo X temos as pontuações dos requisitos sabor, aroma, corpo, acidez, doçura e equilíbrio.

As cores estão presentes para diferenciar o conjunto de dados em questão, ou seja, os requisitos entre si. Não foi inserido uma análise separando gêneros, espécie, variedade ou origem e sim uma configuração geral dos comportamentos das amostras.

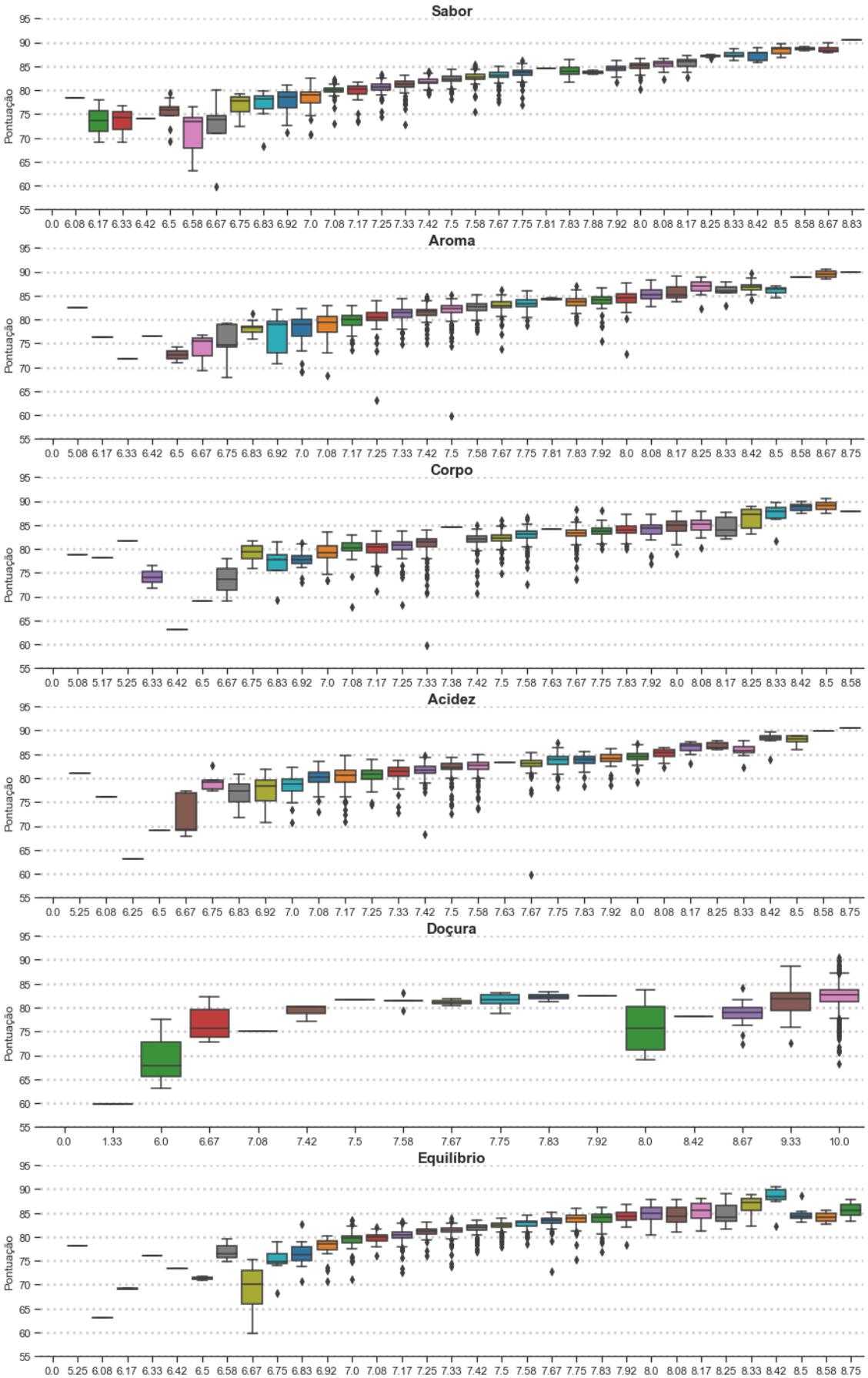


Figura 16: No eixo y está a pontuação geral do café e no eixo x as notas de cada requisito

De maneira empírica se nota que há uma grande variação dos dados sabor, aroma e acidez que são maiores quando comparados aos gráficos nos quais tem os quesitos corpo, doçura e equilíbrio (Figura 16).

Para um entendimento da real correlação entre as variáveis e utilizando os conhecimentos de ciências de dados, foi desenvolvido um código escrito em *python* de mapa de calor denominado como correlação de *heatmap,* utilizando a biblioteca *seaborn*, presente no pandas e o compilador *spyder* (Delft Stack, 2021)

Este método contém um conjunto de equações que compara a correlação da linearidade das variáveis entre si e como elas se movem em relação umas às outras, neste caso foi definido por normalização os valores entre 0 a +1.

A compreensão é que o 0 ou tendendo a ele indica que as variáveis são independentes entre si, dominantemente, o aumento deste valor e a sua tendência a 1 se denota que as variáveis possuem correlações entre si.

Com isso se determinou a correlação, isto é, a dependência ou independência da variáveis entre si (Figura 17), neste caso o comportamento do requisitos de pontuação do café.

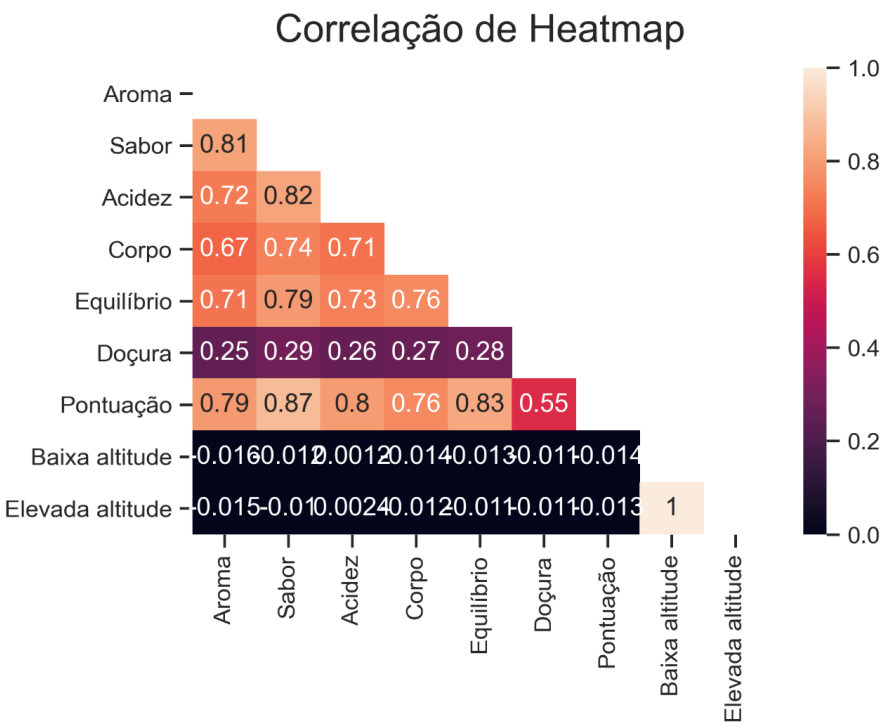


Figura 17: Análise estatística de correlação *heatmap* entre os requisitos de qualidade e a respectiva correlação entre si.

Os resultados acima mostra que um dos fatores amplamente visto e dito no mercado de café é que a altitude tem direta influência na pontuação, por fato comum, que cafés de maiores altitudes são os mais elevados aspectos de qualidade.

O fato é que as correlações de altitudes se apresentaram independentes com relação a pontuação. Isto é, não se tem influência na pontuação final dos cafés, mas as amostras são constituídas de 98% de uma mesma espécie de café, que é Arábica (Figura 8), e está é uma espécie de cultivar de altitude.

Praticamente os dados compreendem somente um tipo de espécie, que consequentemente indica uma homogeneidade nas amostras no quesito espécie. Todavia influencia (valor de 1) e aqui não foi estudado composições diferentes se variando a porcentagem das espécies de Conilon na amostragem portanto alterará o fator de correlação da altitude.

Os resultados indicam que as variáveis analisadas têm maiores correlações e influenciam de maneira direta a pontuação são: sabor (0,87), equilíbrio (0,83), acidez (0,8) e aroma (0,89) conforme é supra mostrado.

No entanto, a primeira percepção no olhar do consumidor referente a qualidade é o aroma, em seguida o sabor. Por outro lado os especialistas (caso dos dados aplicados neste texto) afirmam que o fator prioritário é o sabor e depois o equilíbrio.

Dessa forma, o atributo de qualidade sensorial sabor se torna a conexão de qualidade com a conexão dos especialistas ao público que consome, e também é ressaltado que os cafés analisados são praticamente 100% Arábica (98%), o seu aroma é característico por natureza.

A bebida de café a base de 100 % Conilon, é considerada uma bebida neutra, o seu aroma remete a lembrança de terra (levemente) e a sua bebida apresenta um aspecto mais escuro e sem aroma característico de cítrico, floral dentre outros, fazendo com que o consumidor brasileiro (não acostumado com

arábica), que habitualmente consome café com porcentagem maior que 50% de Conilon, indicarem que quando há produtos com maior porcentagem no *blend* da espécie Arábica, se tem a sensação de maior qualidade.

O que pouco se discute publicamente é que Conilon é um café, que por característica de desenvolvimento, a sua palatabilidade é de uma bebida que remete pouca sensação de impressionar sensorialmente o consumidor final e a sua utilização é para a redução de custo nos cafés vendidos comercialmente.

Um outro aspecto de senso comum para o mercado de café é em relação ao beneficiamento do café (não discutido os métodos neste texto), e o seu comportamento frente a pontuação do mesmo.

Os grupos foram separados nos tipos de secagem do fruto e o seu despolpamento (retirada da casca que envolve a semente – fruto) são processos que controlam a porcentagem em torno de 13 – 15% da humidade relativa da semente, que consequentemente os grãos são levados para o armazenamento separados por tamanhos e vendidos no mercado.

O gráfico abaixo não foi separado por gênero (Arábica e Conilon), mas através do método utilizado de secagem (Figura 18).

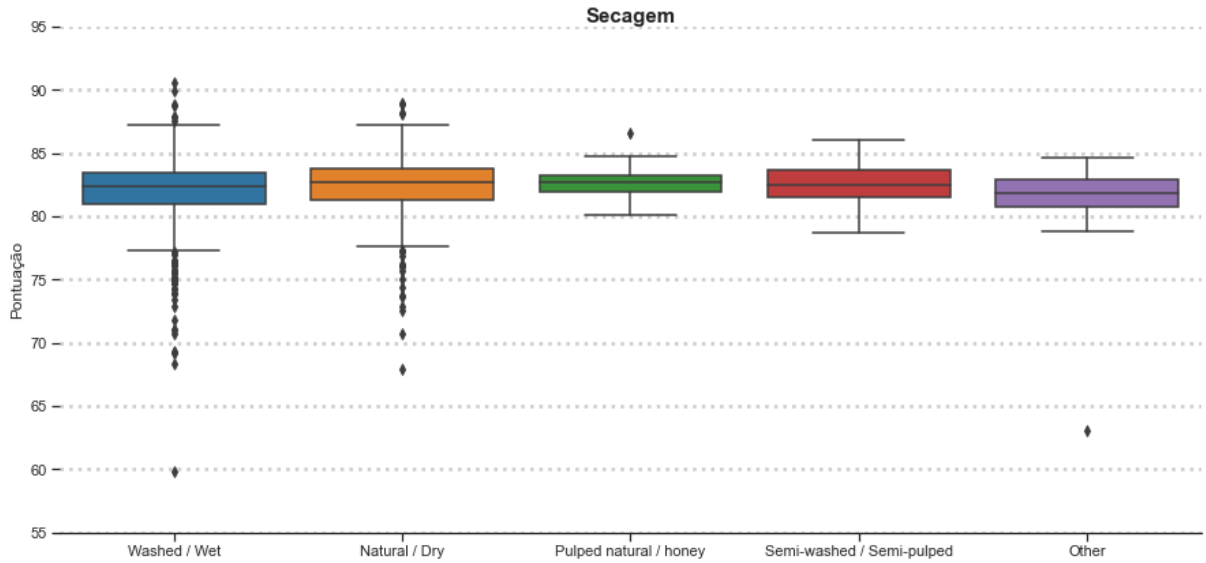


Figura 18: Boxplot de dados separados por tipo de secagem e a sua relação com a pontuação.

O segundo gráfico (Figura 19) separou os grupos através da origem do café (eixo x) *versus* a pontuação (eixo y). A pontuação do mesmo frente ao país de origem e como estão distribuídos os dados para cada país, salientando que não está sendo levado em consideração o gênero (Arábica e Conilon).

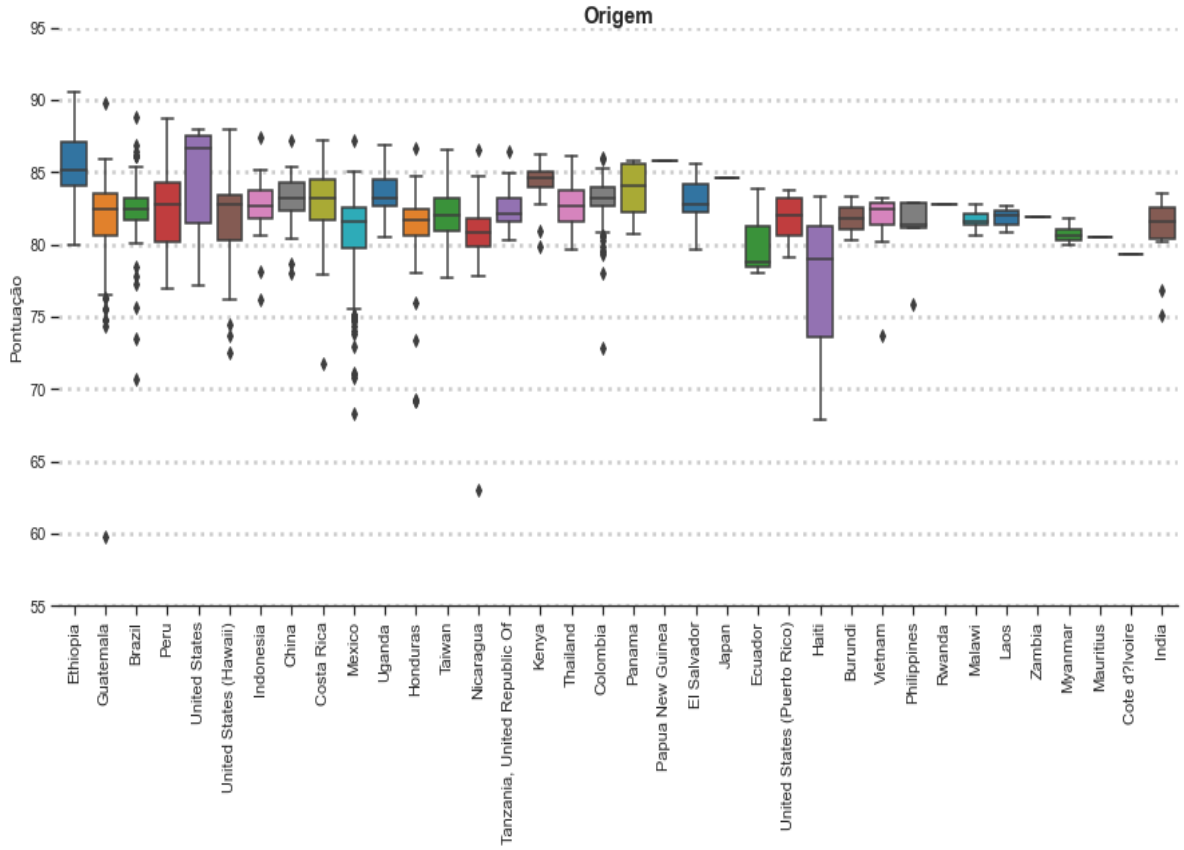


Figura 19: *Boxplot* dos dados de pontuação *versus* origem de produção.

As variações de pontuação final do conjunto de dados se concentra entre as faixas de 80 a 85 pontos, tendo alguns com uma grande variação que é o caso do Haiti, Estados Unidos. Já os cafés de origem na Etiópia é visto que 75% dos seus resultados estão acima de 85 pontos e o café brasileiro 75% dos resultados estão em aproximadamente 82 pontos.

É visto que os dados contém um total de 1.113 colunas e 29 tipos das variedades de cafés, tanto do gênero Arábica como Conilon, e as origens estão distribuídas por 36 países diferentes.

Deste modo o presente estudo se aplicou o teste de análise de variância (ANOVA), tendo como característica determinar se as médias dos grupos são

ou não diferentes entre si, isto foi realizado se aplicando a biblioteca do pandas da linguagem de programação do *python*, onde foi desenvolvido um código que compila os dados e exprime o resultado de comparação entre as médias da pontuação geral em relação aos países que produzem (Tabela 2).

Tabela 2: Resultado da aplicação do teste de variância entre as médias (ANOVA).

	Sum_sq	**df**	**F**	**PR (>F)**
Pontuação	162.36	1	4337,00	0.0
Residual	50.05	1337.0	*NaN*	*NaN*

Enfatizado que este teste desenvolvido se aplica uma análise em todos os conjuntos de dados (Figura 23) e foi definido que a análise de comparação dos requisitos foram origens (entre si) versus pontuação. Ou seja, o protocolo ANOVA compara todo o conjunto de dados separando por origem e compara todos entre si, para validar se a amostra varia entre si ou se elas não variam de forma estatística.

Normalmente, em análises feitas em softwares mais comumente utilizados, como Excel, por exemplo, compara somente 2 grupos. Neste caso foi comparado os 36 países entre si e verificado se há diferenças estatísticas dentro deste grupo de amostras.

O fator F é a variância entre os grupos divididos pela variância dentro dos grupos, e temos que este número está 4.337 (Santos, Data Hackers, 2021). Isto é, os grupos variam mais entre eles. O grupo 75% dos dados estão 81-83 e a variação global 80-85 do Brasil, no entanto, os cafés da Etiópia 75% dos dados estão entre 84-87 e a variação global entre 80 - 90 aproximadamente (Figura 19). Portanto, as variações globais estão inseridas uma dentro da outra.

O resultado supracitado nos conta o fator, no qual avalia a pontuação e a sua independência, se o PR(>F) tem um valor menor que 0,05 assim é denotado estatisticamente que as pontuações dos países analisados têm diferenças, caso este valor seja menor é denotado que as amostras não apresentam diferenças estatísticas.

É visto que o PR (>F) tende a 0, estatisticamente as origens tem diferença significativa entre si.

5. Conclusão

O presente trabalho teve como objetivo a aplicação de técnicas matemáticas para análise de correlação das varáveis sensoriais do café. Como é de senso comum há uma grande variabilidade nas análises quando se tem diversos analistas (degustadores) em questão, mesmo que os protocolos tenham reprodutibilidade científica há algo subjetivo nas análises.

O fato é que) foi implementado recentemente (2022) no mercado uma legislação via Ministério da Agricultura (MAPA), em que se define especificamente cafés dentro do tipo e fora do tipo, e as porcentagens de aceitabilidade para o produto final. Todavia, o café tem por características a passagem de conhecimento de geração para geração.

Dessa forma, o sensorial se torna um dos métodos de filtragem para a qualidade e padronização, há aplicação do *SCAA protocol method*, para a escolha e definição de matéria-prima a ser utilizada no processo industrial.

Mesmo com a aplicação deste método, há uma grande variação na análise da amostra, por exemplo, a susceptibilidade de qualquer não conformidade na saúde bucal do degustador impacta diretamente o resultado final. Devido a este fato, está apresentado neste texto uma compilação estatística de um conjunto de dados sensoriais do café, em que se interpretou as suas correlações nos requisitos para a formação da pontuação do café (definição de qualidade).

Porém, atualmente se aplica 11 variáveis, como conjunto formador dos métodos e matematicamente 5 destas mostraram correlação direta a pontuação e influência de qualidade da bebida café, e mais, neste estudo foi visto que a altitude e o método de secagem não influenciaram diretamente nestes fatores.

Em conclusão, os cafés, das mais diferentes origens do mundo, estatisticamente variam entre si e as suas respectivas pontuações e qualidades.

6. Referências Bibliográficas

ABIC. (10 de Setembro de 2021). *Associação Brasileira da Indústria e Café*. Fonte: ABIC: https://estatisticas.abic.com.br/wp-content/uploads/2022/04/pesq_cafe_superior_abic_spch.pdf

Agro, F. (6 de Abril de 2022). *Forbes*. Fonte: https://forbes.com.br/forbesagro/2022/04/consumo-de-cafe-no-brasil-cresce-17-em-2021/

Baltes, W., & Mevissen, L. (1988). Volatile reaction products from the reaction of phenylalanine glucose during cooking and roasting. *Original Beiten*, 209 - 214.

BSCA. (10 de Abril de 2022). *Caturra Vermelho*. Fonte: BSCA - Cafés Especiais do Brasil: https://brazilcoffeenation.com.br/variety/show/id/9

Coffee Valore. (15 de Abril de 2022). *Coffee Valore*. Fonte: https://www.coffeevalore.com.br/o-cafe-caturra/

Delft Stack. (20 de Novembro de 2021). Fonte: https://www.delftstack.com/pt/howto/seaborn/correlation-heatplot-seaborn-python/

EDTI be better. (2019). Fonte: https://edtisensei.zendesk.com/hc/pt-br/articles/360028876412-Gr%C3%A1fico-de-Box-Plot-no-Minitab

Grão Gourmet. (23 de Abril de 2019). *Grão Gourmet*. Fonte: https://www.graogourmet.com/blog/post1-o-cafeeiro-e-seu-ciclo/

MAPA. (9 de Maio de 2022). *Cafeicultura*. Fonte: https://revistacafeicultura.com.br/index.php?tipo=ler&mat=71928&portaria-sda-n---570--de-9-de-maio-de-2022----estabelece-o-padr--o-oficial-de-classifica----o-do-caf---torrado-.html

Moon, J. K. (2009). Role of Roasting Conditions in the Profile of Volatile Flavor. *Agricultura and Food Chemistry Article*, pp. 5823 - 5831.

Nishida, S. L. (20 de Abril de 2022). *UNESP*. Fonte: https://www2.ibb.unesp.br/Museu_Escola/2_qualidade_vida_humana/Museu2_qualidade_corpo_sensorial_introducao.htm

Paiva, L. (21 de Junho de 2021). *Review Café*. Fonte: https://reviewcafe.com.br/dicas-e-receitas/cafe-bourbon/

Parliment, T. H. (2000). An Overview of Coffee Roasting. Em T. H. Parliment. Nova Iorque: American Chemical Society.

Perfect Daily grind. (10 de Março de 2021). Fonte: https://perfectdailygrind.com/pt/2021/03/10/variedade-de-cafe-typica-o-que-e-e-por-que-e-tao-importante/

Ruosi, M. (19 de Outubro de 2021). A Further tool to monitor the coffee roasting process: Aroma Composition and Chemical Indices. *Agricultural and Food Chemistry*, pp. 11283 - 11291.

Santos, G. (3 de Janeiro de 2021). *Data Hackers*. Fonte: https://medium.com/data-hackers/estat%C3%ADstica-para-sele%C3%A7%C3%A3o-de-atributos-81bdc274dd2c

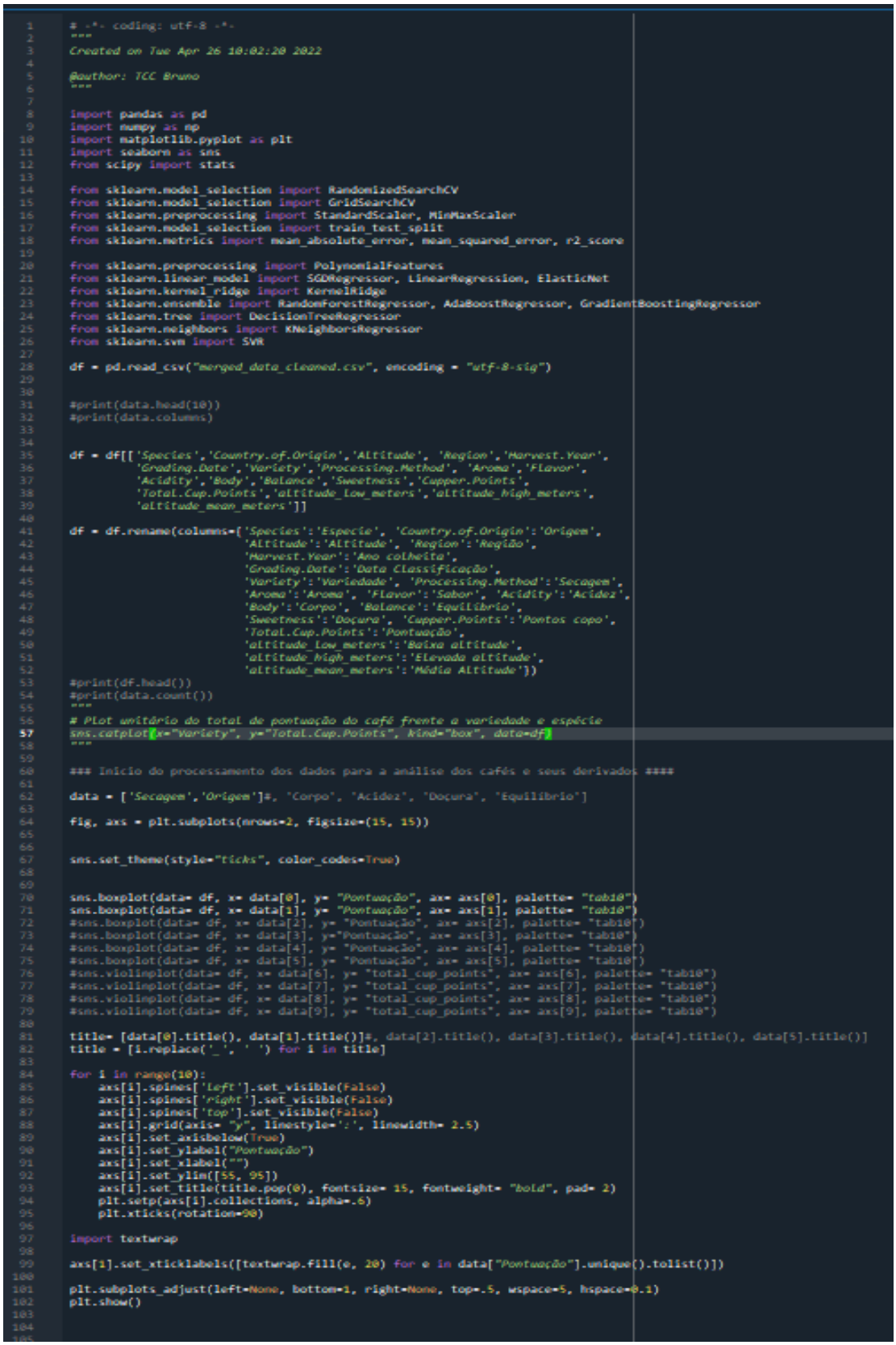
Santos, G. (s.d.). *Medium - Data Hackers*. Fonte: https://medium.com/data-hackers/estat%C3%ADstica-para-sele%C3%A7%C3%A3o-de-atributos-81bdc274dd2c

SCAA Cupping protocols. (Novembro de 2008). *Coffee traveler*. Fonte: http://coffeetraveler.net/wp-content/files/901-SCAA_CuppingProtocols_TSC_DocV_RevDec08_Portuguese.pdf

The Coffee Traveler. (21 de Agosto de 2011). Fonte: https://www.thecoffeetraveler.net/home-1/tag/Typica

Vieceli, L. (21 de fevereiro de 2022). *Folha de São Paulo*. Fonte: Folha de São Paulo: https://www1.folha.uol.com.br/mercado/2022/02/precos-do-cafe-sobem-mais-de-50-e-alteram-consumo-do-brasileiro.shtml

7. Anexos



```python
# -*- coding: utf-8 -*-
"""
Created on Tue Apr 26 10:02:20 2022

@author: TCC Bruno
"""

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats

from sklearn.model_selection import RandomizedSearchCV
from sklearn.model_selection import GridSearchCV
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import SGDRegressor, LinearRegression, ElasticNet
from sklearn.kernel_ridge import KernelRidge
from sklearn.ensemble import RandomForestRegressor, AdaBoostRegressor, GradientBoostingRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.svm import SVR

df = pd.read_csv("merged_data_cleaned.csv", encoding = "utf-8-sig")


#print(data.head(10))
#print(data.columns)


df = df[['Species','Country.of.Origin','Altitude', 'Region','Harvest.Year',
         'Grading.Date','Variety','Processing.Method', 'Aroma','Flavor',
         'Acidity','Body','Balance','Sweetness','Cupper.Points',
         'Total.Cup.Points','altitude_low_meters','altitude_high_meters',
         'altitude_mean_meters']]

df = df.rename(columns={'Species':'Especie', 'Country.of.Origin':'Origem',
                        'Altitude':'Altitude', 'Region':'Região',
                        'Harvest.Year':'Ano colheita',
                        'Grading.Date':'Data Classificação',
                        'Variety':'Variedade', 'Processing.Method':'Secagem',
                        'Aroma':'Aroma', 'Flavor':'Sabor', 'Acidity':'Acidez',
                        'Body':'Corpo', 'Balance':'Equilíbrio',
                        'Sweetness':'Doçura', 'Cupper.Points':'Pontos copo',
                        'Total.Cup.Points':'Pontuação',
                        'altitude_low_meters':'Baixa altitude',
                        'altitude_high_meters':'Elevada altitude',
                        'altitude_mean_meters':'Média Altitude'})
#print(df.head())
#print(data.count())
"""
# Plot unitário do total de pontuação do café frente a variedade e espécie
sns.catplot(x="Variety", y="Total.Cup.Points", kind="box", data=df)
"""

### Início do processamento dos dados para a análise dos cafés e seus derivados ####

data = ['Secagem','Origem']#, 'Corpo', 'Acidez', 'Doçura', 'Equilíbrio']

fig, axs = plt.subplots(nrows=2, figsize=(15, 15))


sns.set_theme(style="ticks", color_codes=True)


sns.boxplot(data= df, x= data[0], y= "Pontuação", ax= axs[0], palette= "tab10")
sns.boxplot(data= df, x= data[1], y= "Pontuação", ax= axs[1], palette= "tab10")
#sns.boxplot(data= df, x= data[2], y= "Pontuação", ax= axs[2], palette= "tab10")
#sns.boxplot(data= df, x= data[3], y="Pontuação", ax= axs[3], palette= "tab10")
#sns.boxplot(data= df, x= data[4], y= "Pontuação", ax= axs[4], palette= "tab10")
#sns.boxplot(data= df, x= data[5], y= "Pontuação", ax= axs[5], palette= "tab10")
#sns.violinplot(data= df, x= data[6], y= "total_cup_points", ax= axs[6], palette= "tab10")
#sns.violinplot(data= df, x= data[7], y= "total_cup_points", ax= axs[7], palette= "tab10")
#sns.violinplot(data= df, x= data[8], y= "total_cup_points", ax= axs[8], palette= "tab10")
#sns.violinplot(data= df, x= data[9], y= "total_cup_points", ax= axs[9], palette= "tab10")

title= [data[0].title(), data[1].title()]#, data[2].title(), data[3].title(), data[4].title(), data[5].title()]
title = [i.replace('_', ' ') for i in title]

for i in range(10):
    axs[i].spines['left'].set_visible(False)
    axs[i].spines['right'].set_visible(False)
    axs[i].spines['top'].set_visible(False)
    axs[i].grid(axis= "y", linestyle=':', linewidth= 2.5)
    axs[i].set_axisbelow(True)
    axs[i].set_ylabel("Pontuação")
    axs[i].set_xlabel("")
    axs[i].set_ylim([55, 95])
    axs[i].set_title(title.pop(0), fontsize= 15, fontweight= "bold", pad= 2)
    plt.setp(axs[i].collections, alpha=.6)
    plt.xticks(rotation=90)

import textwrap

axs[1].set_xticklabels([textwrap.fill(e, 20) for e in data["Pontuação"].unique().tolist()])

plt.subplots_adjust(left=None, bottom=1, right=None, top=.5, wspace=5, hspace=0.1)
plt.show()
```

Figura 20: Leitura e tratamento iniciais dos dados *plotting* iniciais.

```python
# -*- coding: utf-8 -*-
"""
Created on Wed Apr 27 09:21:52 2022

@author: BPO - TCC QUimica 2022
"""
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np




data = pd.read_csv("merged_data_cleaned.csv", encoding = "utf-8-sig")

data = data[['Species','Country.of.Origin','Altitude', 'Region',
             'Harvest.Year','Grading.Date','Variety','Processing.Method',
             'Aroma','Flavor','Acidity','Body','Balance','Sweetness',
             'Cupper.Points','Total.Cup.Points','altitude_low_meters',
             'altitude_high_meters','altitude_mean_meters']]

data = data.rename(columns={'Species':'Especie', 'Country.of.Origin':'Origem',
                            'Altitude':'Altitude', 'Region':'Região',
                            'Harvest.Year':'Ano colheita',
                            'Grading.Date':'Data Classificação',
                            'Variety':'Variedade',
                            'Processing.Method':'Secagem',
                            'Aroma':'Aroma', 'Flavor':'Sabor',
                            'Acidity':'Acidificação', 'Body':'Corpo',
                            'Balance':'Equilíbrio', 'Sweetness':'Doçura',
                            'Cupper.Points':'Pontos copo',
                            'Total.Cup.Points':'Pontuação',
                            'altitude_low_meters':'Baixa altitude',
                            'altitude_high_meters':'Elevada altitude',
                            'altitude_mean_meters':'Média Altitude'})


#print(data.info())

quantidade = data[['Especie', 'Origem']].groupby('Especie').count()
paises = data[['Origem', 'Especie']].groupby('Origem').count()
variedade = data[['Variedade', 'Especie']].groupby('Variedade').count()
ano = data[['Ano colheita', 'Especie']].groupby('Ano colheita').count()
ano = data[['Acidez', 'Especie']].groupby('Acidez').count()
#altitude = data[['Média Altitude', 'Origem']].groupby('Média Altitude').count()
#quantidade = compras[['entrada', 'entr-pedido']].groupby('entrada').count()

#plot dos dados em questão
#print(quantidade)
#print(paises)
#print(variedade)
#print(ano)
#print(altitude)
print(Acidez)
```

Figura 21: Código para criação de gráficos separando e agrupando grupos de interesse.

```python
# -*- coding: utf-8 -*-
"""
Created on Wed Apr 27 21:33:40 2022

@author: BPO TCC de química
"""
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

df = pd.read_csv("merged_data_cleaned.csv", encoding = "utf-8-sig")

df = df[['Species','Country.of.Origin','Altitude', 'Region','Harvest.Year',
         'Grading.Date','Variety','Processing.Method', 'Aroma','Flavor',
         'Acidity','Body','Balance','Sweetness','Cupper.Points',
         'Total.Cup.Points','altitude_low_meters','altitude_high_meters',
         'altitude_mean_meters']]

df = df.rename(columns={'Species':'Especie', 'Country.of.Origin':'Origem',
                        'Altitude':'Altitude', 'Region':'Região',
                        'Harvest.Year':'Ano colheita',
                        'Grading.Date':'Data Classificação',
                        'Variety':'Variedade', 'Processing.Method':'Secagem',
                        'Aroma':'Aroma', 'Flavor':'Sabor', 'Acidity':'Acidez',
                        'Body':'Corpo', 'Balance':'Equilíbrio',
                        'Sweetness':'Doçura', 'Cupper.Points':'Pontos copo',
                        'Total.Cup.Points':'Pontuação',
                        'altitude_low_meters':'Baixa altitude',
                        'altitude_high_meters':'Elevada altitude',
                        'altitude_mean_meters':'Média Altitude'})


df = df.drop(columns={'Ano colheita', 'Data Classificação',
                      'Pontos copo','Média Altitude'})

df.corr()

#sns.heatmap(df.corr());
mask = np.triu(np.ones_like(df.corr(), dtype=np.bool))
heatmap = sns.heatmap(df.corr(), vmin=0, mask=mask,  vmax=1, annot=True)
heatmap.set_title('Correlação de Heatmap', fontdict={'fontsize':18}, pad=12);
plt.savefig('heatmap.png', dpi=300, bbox_inches='tight')

```

Figura 22: Código de leitura e desenvolvimento da análise de heatmap e definições de correlações.

```python

@author: BPO - TCC Química - USP/IQSC
"""
import pandas as pd
# load data file
df = pd.read_csv("merged_data_cleaned.csv")
df = df[['Species','Country.of.Origin','Altitude', 'Region','Harvest.Year',
         'Grading.Date','Variety','Processing.Method', 'Aroma','Flavor',
         'Acidity','Body','Balance','Sweetness','Cupper.Points',
         'Total.Cup.Points','altitude_low_meters','altitude_high_meters',
         'altitude_mean_meters']]
df = df.rename(columns={'Species':'Especie', 'Country.of.Origin':'Origem',
                        'Altitude':'Altitude', 'Region':'Região',
                        'Harvest.Year':'Ano colheita', 'Grading.Date':
                            'Data Classificação', 'Variety':'Variedade',
                            'Processing.Method':'Secagem', 'Aroma':'Aroma',
                            'Flavor':'Sabor', 'Acidity':'Acidez', 'Body':'Corpo',
                            'Balance':'Equilíbrio', 'Sweetness':'Doçura',
                            'Cupper.Points':'Pontos copo',
                            'Total.Cup.Points':'Pontuação',
                            'altitude_low_meters':'Baixa altitude',
                            'altitude_high_meters':'Elevada altitude',
                            'altitude_mean_meters':'Média Altitude'})
df = df.drop(columns={'Especie', 'Altitude', 'Região', 'Ano colheita',
                      'Data Classificação', 'Variedade', 'Secagem','Aroma',
                      'Acidez', 'Corpo', 'Equilíbrio','Doçura', 'Pontos copo',
                      'Baixa altitude', 'Elevada altitude', 'Média Altitude'})

origem = df['Origem']

#print(origem)
#print(df)
#Desenvolvendo o plot e teste anova para cada origem em questão e verificação de similiaridade.
#import matplotlib.pyplot as plt
#import seaborn as sns
Data = df

#ax = sns.boxplot(x='Sabor', y='Pontuação', data=df, color='#99c2a2')
#ax = sns.swarmplot(x="Sabor", y="Pontuação", data=df, color='#7d0013')
#plt.xticks(rotation=90)
#plt.show()
## Teste ANOVA
import scipy.stats as stats
fvalue, pvalue = stats.f_oneway(df['Sabor'], df['Pontuação']) # , df['Ethiopia'], df['Costa Rica'], df['Guatemala'])
print(fvalue, pvalue)


# Desenvolvendo o teste ANOVA EM QUESTÃO para avaliação das significância estátisticas
import statsmodels.api as sm
from statsmodels.formula.api import ols

# Ordinary Least Squares (OLS) model
model = ols('Sabor ~ Pontuação', data=df).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
anova_table
anova = anova_table
print(anova)

```

Figura 23:Código de leitura e aplicação do teste estatístico de ANOVA