Abalone – Uma breve análise de dados

Aline Mateus da Silva
Departamento de Teleinformática line
Universidade Federal do Ceará
Fortaleza, Brasil
alicemateus13@hotmail.com

Andrea Sampaio Elias
Departamento de Teleinformática line
Universidade Federal do Ceará
Fortaleza, Brasil
andrease@alu.ufc.br

Arnaldo Aguiar Portela Filho
Departamento de Teleinformática line
Universidade Federal do Ceará
Fortaleza, Brasil
arnaldoaguiarp@gmail.com

Italo Viana Severo
Departamento de Teleinformática line
Universidade Federal do Ceará
Fortaleza, Brasil
italo_1002@live.com

***Resumo* — Este artigo trata da análise de dados do molusco abalone. Foram analisados dados como sexo (tipo), medida da concha, diâmetro e peso (total e de componentes).**

Palavras-chave – abalone, análise monovariada, análise bivariada, análise multivariada, python, peso, altura, histogramas.

I. INTRODUÇÃO

Nas diversas áreas de conhecimento tem sido cada vez mais importante a análise de dados, tanto para avaliar as condições presentes, identificar padrões e tirar conclusões que tragam mais conhecimento sobre aquela área ou melhorem e automatizem processos, quanto até mesmo para fazer previsões.

Nas pesquisas de biologia sobre as diversas espécies do reino animal, buscando conhecê-las mais à fundo, é imprescindível a coleta de grande volume de dados.

O abalone é uma espécie de molusco gastrópode e no presente estudo examinou-se dados de mais de 4 mil amostras dessa espécie, objetivando-se identificar características gerais através da análise e cruzamento desses dados.

II. MÉTODOS

O método utilizado para fazer a análise dos dados foi o quantitativo, utilizando a linguagem Python. Primeiramente inseriu-se o arquivo com os dados obtidos no link: http://archive.ics.uci.edu/ml/datasets/Abalone. Após os dados estarem carregados na variável "**tabela**", começou-se a análise.

Inicialmente, podemos ver que toda a pesquisa de dados em relação ao tema discutido foi realizada a partir da tabela disponibilizada com as 4177 amostras da planilha inicial. Com a função "head()" do python pandas tivemos a amostragem da tabela abaixo, onde podemos observar uma análise de distribuição dos cinco primeiros dados da tabela contendo as seguintes características: tipo (**Type**), tamanho da concha (**LongestShell**), diâmetro (**Diamenter**), altura (**Height**), peso Total (**WholeWeight**), peso sem casca (**ShuckedWeight**), peso das vísceras (**VisceraWeight**), peso da concha (**ShellWeight**), anéis (**Rings**).

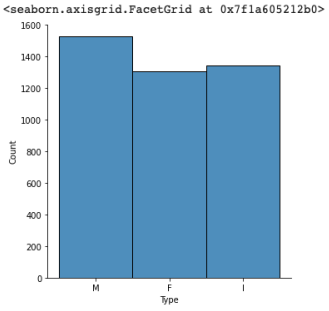
	Type	LongestShell	Diameter	Height	WholeWeight	ShuckedWeight	VisceraWeight	ShellWeight	Rings
0	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7

A primeira abordagem foi para a descrição dos dados quanto ao número de observações, número de variáveis preditivas, número de classes e distribuição de classes. Para tanto, usou-se a função "describe()", com o seguinte comando: "tabela.describe()". Observou-se que foi exibida uma tabela com diversas informações sobre os dados. Foram os seguintes: count, mean, std, min, 25%, 50%, 75%, max, ou seja, respectivamente, a quantidade de valores, a média, o desvio padrão, o valor mínimo, os quartis da distribuição e o valor máximo.

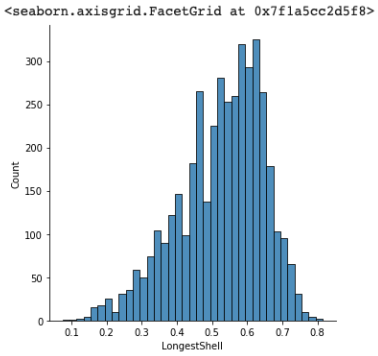
	LongestShell	Diameter	Height	WholeWeight	ShuckedWeight	VisceraWeight	ShellWeight	Rings
count	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000
mean	0.523992	0.407881	0.139516	0.828742	0.359367	0.180594	0.238831	9.933684
std	0.120093	0.099240	0.041827	0.490389	0.221963	0.109614	0.139203	3.224169
min	0.075000	0.055000	0.000000	0.002000	0.001000	0.000500	0.001500	1.000000
25%	0.450000	0.350000	0.115000	0.441500	0.186000	0.093500	0.130000	8.000000
50%	0.545000	0.425000	0.140000	0.799500	0.336000	0.171000	0.234000	9.000000
75%	0.615000	0.480000	0.165000	1.153000	0.502000	0.253000	0.329000	11.000000
max	0.815000	0.650000	1.130000	2.825500	1.488000	0.760000	1.005000	29.000000

As próximas imagens abaixo são os histogramas do conjunto de dados tabulados. Para a plotagem dos gráficos foi usado o seguinte comando do python pandas sea.displot(tabela, x= 'o dado a ser plotado').

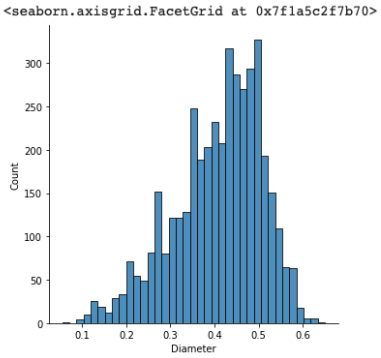
No histograma abaixo, temos a distribuição de frequência de acordo com o tipo de abalone: masculino, feminino e infantil. Podemos observar que o abalone predominante é o masculino, seguido pelo infantil e por último, o menos predominante, o feminino. Outro ponto a se examinar é que a diferença da frequência entre o feminino e o infantil é pequena, comparada à disparidade da predominância masculina.



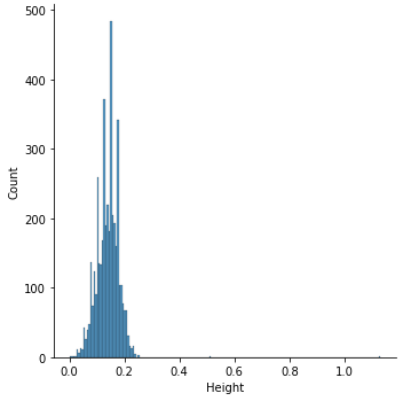
No histograma abaixo, temos a distribuição de frequência de acordo com o tamanho da concha do abalone; nesse gráfico podemos observar que existe uma grande variação de tamanhos e que as mais predominantes são as que ficam na faixa de 0,45 a 0,65.



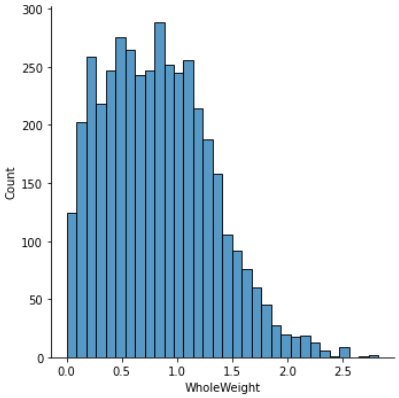
No histograma abaixo, temos a distribuição de frequência de acordo com o diâmetro da concha do abalone, onde podemos inferir que a maior quantidade de abalones tabulados possui uma predominância de diâmetro na faixa de 0,35 a 0,55.



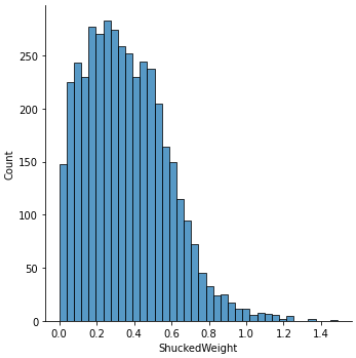
No histograma abaixo, temos a distribuição de frequência de acordo com a altura do abalone. A partir do histograma, podemos observar que não existe uma grande variação na altura dos abalones, já que a altura varia em uma pequena faixa.



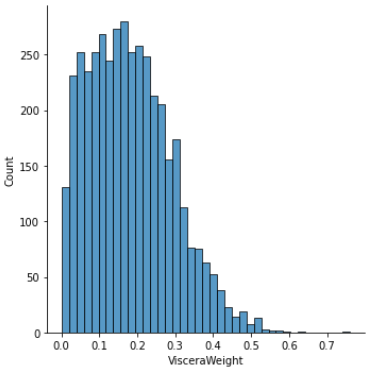
No histograma a seguir, temos a distribuição de frequência de acordo com o peso total do abalone, onde podemos inferir que a maior quantidade de abalones tabulados possui uma predominância de peso na faixa de 0,25 a 1,2.



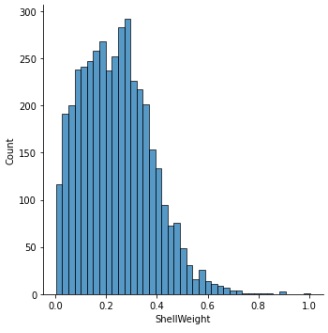
No histograma abaixo, temos a distribuição de frequência de acordo com o peso sem casca do abalone. A partir do histograma podemos observar que os abalones sem casca possuem uma variedade de pesos, os abalones que foram analisados para o gráfico abaixo possuem uma predominância de peso na faixa de 0,2 a 0,3.



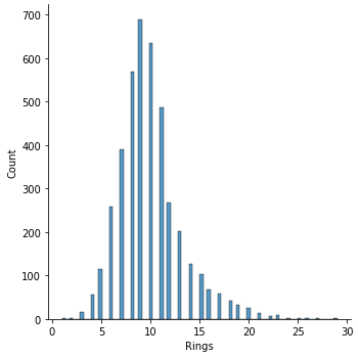
No histograma abaixo temos a distribuição de frequência de acordo com o peso das vísceras do abalone. É notório que o peso de vísceras varia bastante entre os abalones analisados, cuja a maioria tem seu peso na faixa de 0,0 a 0,3.



No histograma abaixo, temos a distribuição de frequência de acordo com o peso da concha do abalone. Podemos inferir que o peso da concha dos abalones analisados é bastante variada.



No histograma abaixo temos a distribuição de frequência de acordo com o número de anéis de abalone. Podemos observar que a maioria dos abalones possui entre 8 e 11 anéis, que existem poucos abalones com mais de 25 anéis e que também tem poucos abalones com um número reduzido de anéis.



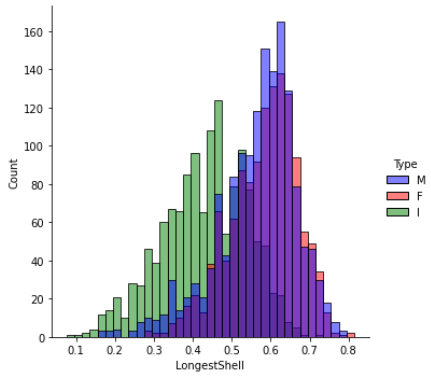
Na imagem a seguir temos o cálculo da média, do desvio padrão e da assimetria das características dos abalones.

```
[ ] tabela.mean()

LongestShell     0.523992
Diameter         0.407881
Height           0.139516
WholeWeight      0.828742
ShuckedWeight    0.359367
VisceraWeight    0.180594
ShellWeight      0.238831
Rings            9.933684
dtype: float64

[ ] tabela.std()

LongestShell     0.120093
Diameter         0.099240
Height           0.041827
WholeWeight      0.490389
ShuckedWeight    0.221963
VisceraWeight    0.109614
ShellWeight      0.139203
Rings            3.224169
dtype: float64

[ ] tabela.skew()

LongestShell    -0.639873
Diameter        -0.609198
Height           3.128817
WholeWeight      0.530959
ShuckedWeight    0.719098
VisceraWeight    0.591852
ShellWeight      0.620927
Rings            1.114102
dtype: float64
```

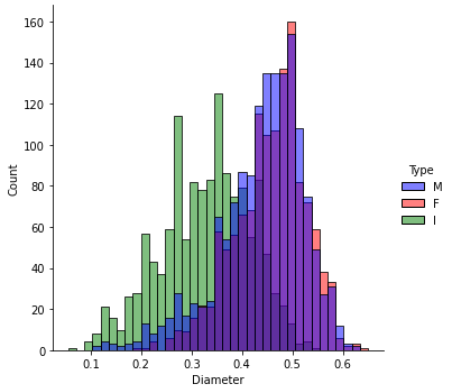
A tabela da média geral (tabela.mean()) leva em consideração todas as amostras de abalones. Na tabela de desvio padrão (tabela.std()) podemos observar um grande desvio padrão nos dados dos anéis (rings), Na tabela de assimetria (tabela.skew()) podemos ver que a maioria das características tem assimetria de distribuição moderada, já as características altura (height) e anéis (rings) têm assimetria de distribuição altamente distorcida.

Para fazer uma análise mais refinada dos gráficos e ter acesso a novas informações, fizemos uma análise univariável condicional, para agrupar nossos dados da análise anterior a partir das classes M, F e I e para isso utilizamos a função displot() da biblioteca Seaborn novamente dessa vez utilizando o parâmetro hue com a nossa coluna de 'Type' e com isso ele gera os gráficos fazendo o agrupamento pelas classes definidas nesta coluna.

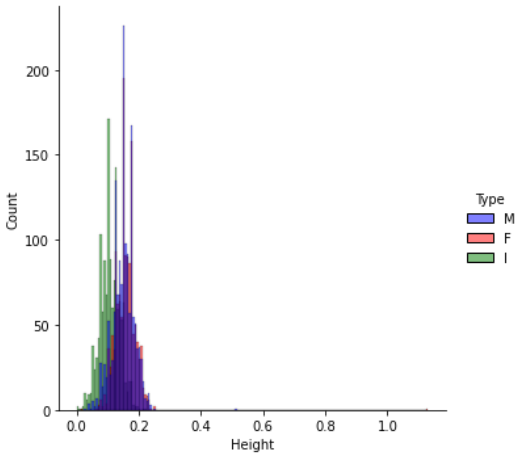
A seguir histograma de frequência condicional quanto ao tamanho da concha; pode-se observar que os abalones do tipo M com tamanho de casca entre 0.5 e 0.65 têm uma maior frequência em relação aos outros tipos e que os abalones com o tamanho de 0.1 a 0.45 têm sua maior parte formada por indivíduos do tipo I:



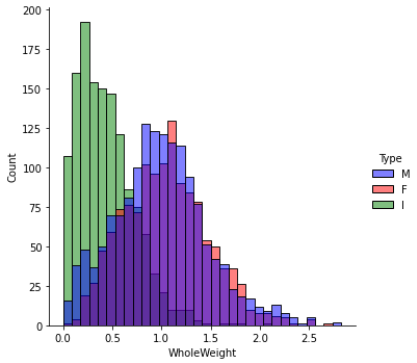
Histograma de frequência condicional quanto ao diâmetro; a maior parte dos indivíduos com diâmetro entre 0 e 0.38 são do tipo I; os abalones do tipo F têm maior variação no seu diâmetro:



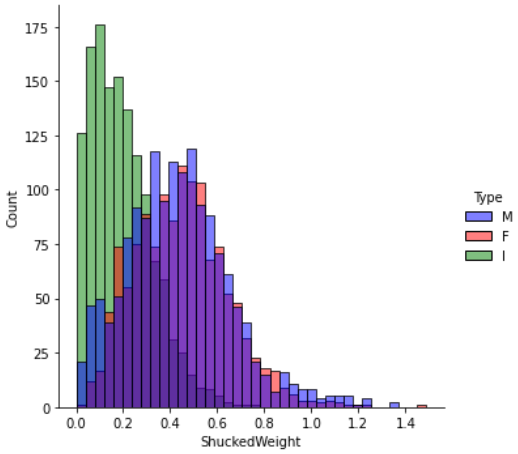
Histograma de frequência condicional quanto à altura; observável que a faixa de tamanho dos abalones varia de próximo de 0 até 0.25, tendo os infantis um tamanho próximo a 0.1:



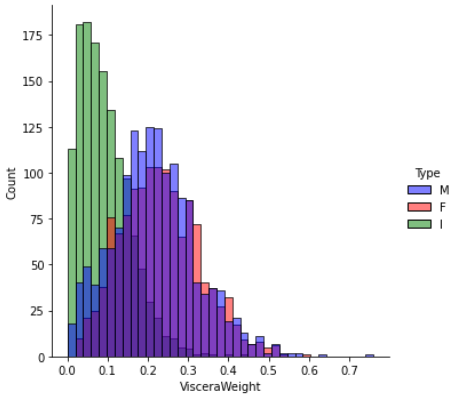
A seguir histograma de frequência condicional quanto ao peso total; o peso total dos indivíduos infantis varia de 0 a 1.5; indivíduos M e F possuem um peso muito aproximado, porém foram observadas mais amostras do tipo M:



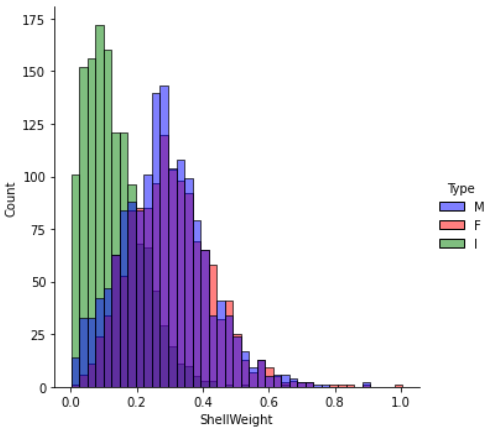
Histograma de frequência condicional quanto ao peso sem casca; o peso dos infantis varia de 0 até 0.8 e os abalones com peso acima de 0.85 têm em sua maior composição indivíduos do tipo M:



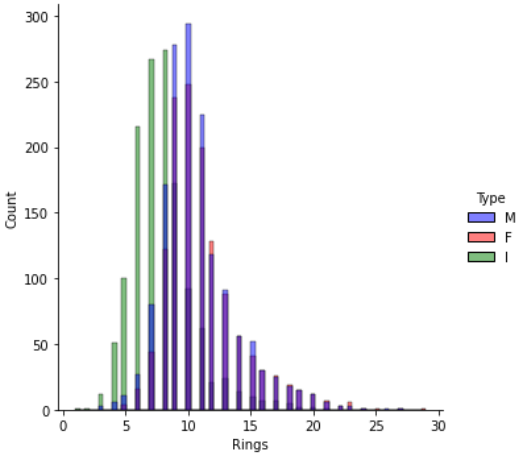
A seguir histograma de frequência condicional quanto ao peso das vísceras, os indivíduos do tipo infantil têm o peso das vísceras em maior parte variando de 0 a 0.3, indivíduos do tipo M e F variam entre 0 e 0.6:



A seguir histograma de frequência condicional do peso da concha; indivíduos infantis possuem peso de casca variando de próximo de 0 até 0.4, os demais tipos possuem o peso da concha aproximado entre si e variando de 0 a 0.8:



A seguir histograma de frequência condicional quanto aos anéis; os indivíduos do tipo infantil possuem predominância nas amostras até a contagem de 8 anéis; após isso a predominância fica com os indivíduos do tipo M, seguidos pelos do tipo F; é visível que a maioria dos abalones possuem uma faixa de anéis que vai de 5 até 15, tendo uma pequena quantidade que excede a contagem de 15 anéis:



Aqui, um compilado das médias dos gráficos agrupados por tipo do abalone (M, F e I); é possível notar que a média dos atributos entre M e F são muito próximos, porém possuem uma média bem acima se comparada com os do tipo I:

Type	LongestShell	Diameter	Height	WholeWeight	ShuckedWeight	VisceraWeight	ShellWeight	Rings
F	0.579093	0.454732	0.158011	1.046532	0.446188	0.230689	0.302010	11.129304
I	0.427746	0.326494	0.107996	0.431363	0.191035	0.092010	0.128182	7.890462
M	0.561391	0.439287	0.151381	0.991459	0.432946	0.215545	0.281969	10.705497

Desvio padrão das amostras, os maiores desvios padrões estão relacionados aos indivíduos do tipo M, indicando que estes têm maior variação quanto a sua média nesses atributos:

Type	LongestShell	Diameter	Height	WholeWeight	ShuckedWeight	VisceraWeight	ShellWeight	Rings
F	0.086160	0.070954	0.039984	0.430316	0.198663	0.097617	0.125649	3.104256
I	0.108858	0.088109	0.031995	0.286275	0.128405	0.062536	0.084927	2.511554
M	0.102697	0.084398	0.034804	0.470581	0.223000	0.104919	0.130834	3.026349

A assimetria dos dados agrupados pelos tipos M, F e I, as amostras dos tipo I em relação aos atributos LongestShell, Diameter e Height apresentaram simetria nos dados, já os indivíduos do tipo M apresentaram simetria nos atributos WholeWeight, ShellWeight, VisceraWeight e Height e os do

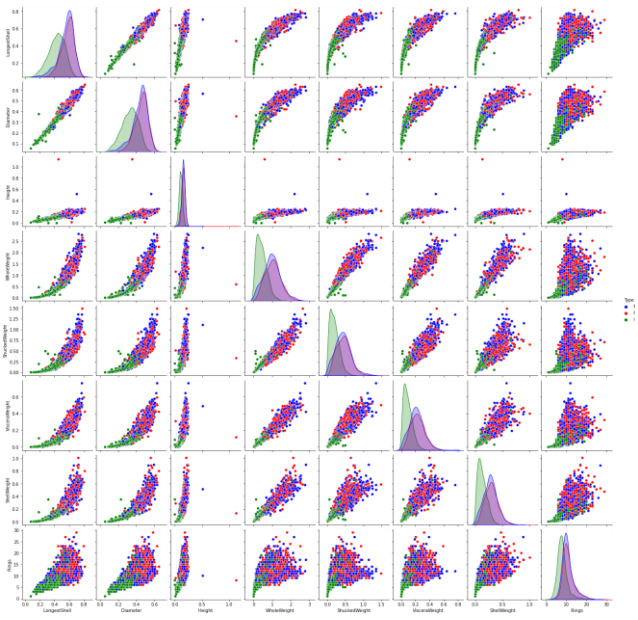
tipo F a simetria ficou nos atributos VisceraWeight, WholeWeight e Diameter.

Type	LongestShell	Diameter	Height	WholeWeight	ShuckedWeight	VisceraWeight	ShellWeight	Rings
F	-0.528735	-0.506289	10.925682	0.368498	0.546770	0.393427	0.691757	1.474022
I	-0.346951	-0.292925	-0.058515	0.974459	0.865294	1.066459	1.001923	1.326831
M	-0.913565	-0.923321	0.417547	0.406007	0.632451	0.506076	0.487628	1.255072

Na quarta parte do estudo foi feita uma análise bivariada incondicional dos dataset de modo a traçar gráficos entre todos os pares de dados. Nos gráficos foram colocadas a identificação das classes (masculino, feminino e infantil) através de diferentes cores. A análise bivariada consiste em um método de análise de duas variáveis. Foi utilizada a função "*pairplot"* da biblioteca *seaborn* do Python, que é uma biblioteca responsável pela visualização de dados, ou seja, a criação dos gráficos.

O comando dado, especificamente foi: "GridBiVar = sea.pairplot(tabela, hue='Type', diag_kind='hist')". Os parâmetros usados na função seaborn.pairplot foram *data* e *hue*, sendo "data" os dados armazenados na variável "tabela" e "hue" a variável escolhida entre os dados pra serem mapeadas em diferentes cores, no caso a variável "type", representando as classes dos abalones (masculino, feminino e infantil).

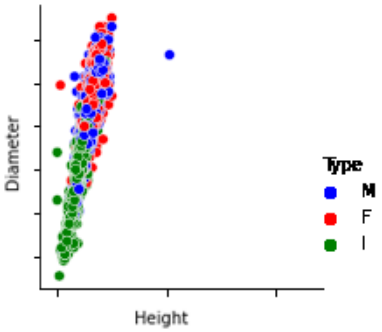
O resultado do comando foi a plotagem de diversos gráficos de pares de dados, como mostrado a seguir:



Os gráficos acima representam os pares de dados como na tabela a seguir (gerada a partir da função "*corr",* função esta que acha a correlação de pares de todas as colunas no dataframe):

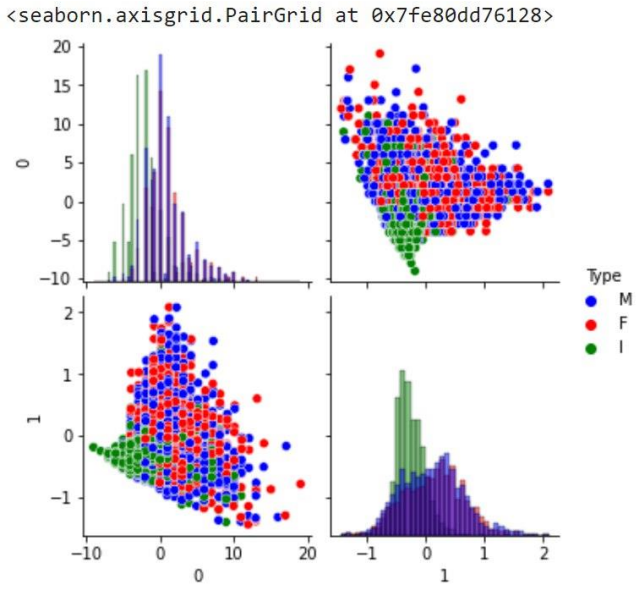
	LongestShell	Diameter	Height	WholeWeight	ShuckedWeight	VisceraWeight	ShellWeight	Rings
LongestShell	1.000	0.987	0.828	0.925	0.898	0.903	0.898	0.557
Diameter	0.987	1.000	0.834	0.925	0.893	0.900	0.905	0.575
Height	0.828	0.834	1.000	0.819	0.775	0.798	0.817	0.557
WholeWeight	0.925	0.925	0.819	1.000	0.969	0.966	0.955	0.540
ShuckedWeight	0.898	0.893	0.775	0.969	1.000	0.932	0.883	0.421
VisceraWeight	0.903	0.900	0.798	0.966	0.932	1.000	0.908	0.504
ShellWeight	0.898	0.905	0.817	0.955	0.883	0.908	1.000	0.628
Rings	0.557	0.575	0.557	0.540	0.421	0.504	0.628	1.000

Foram gerados 64 gráficos ao total. Os pontos em azul representam indivíduos da classe masculino, os pontos de cor vermelha representam indivíduos da classe feminino e os pontos verdes representam indivíduos da classe infantil. Através da observação dos mesmos pode-se tirar várias conclusões, como por exemplo, observando-se o gráfico gerado com os dados Diâmetro (diameter) x Altura (height), observa-se um crescimento linear, ou seja, quanto mais o diâmetro, maior a altura observada. Indivíduos infantis tem menores pesos e alturas e têm os pontos mais à esquerda e mais abaixo, já indivíduos adultos, tanto do sexo masculino quanto feminino, têm seus pontos misturados predominantemente um pouco mais à direita e mais acima do gráfico. Como se pode ver, as alturas dos indivíduos, em geral, não crescem tanto, mesmo o diâmetro crescendo bastante, a altura não cresce tanto, o que resulta numa reta mais inclinada para a reta do eixo y. Demonstra-se o gráfico em análise:



Ao se observar cada um dos gráficos gerado pode-se tirar diversas conclusões sobre a espécie abalone.

Na última e quinta parte do estudo fizemos uma análise dos componentes principais.



A análise PCA - Principal Component Analysis - tem como objetivo encontrar um meio de condensar a informação contida em várias variáveis originais em um conjunto menor de variáveis estatísticas (componentes) com uma perda mínima de informação.

Com o objetivo de realizar a análise, nós fizemos uma transformação linear através de funções do python utilizando as bibliotecas a seguir:

```
from sklearn.decomposition import PCA
from sklearn import preprocessing
```

Através do algoritmo nós conseguimos alcançar uma variância explicada de 97%. Além disso, conseguimos encontrar a tabela de loadings abaixo:

	LongestShell	Diameter	Height	WholeWeight	ShuckedWeight
PC-0	0.021051	0.017940	0.007316	0.083680	0.029735
PC-1	0.183053	0.148804	0.052460	0.834461	0.399952

	VisceraWeight	ShellWeight	Rings
PC-0	0.017460	0.027431	0.995107
PC-1	0.183253	0.205272	-0.097937

Com base na tabela anterior, podemos concluir que o PC-0 pode ser definido como o número de anéis (Rings) e o PC-1 podemos observar que não apresentou uma margem clara de diferença entre seus dados.

III. RESULTADOS

Por fim, através da análise dos dados apresentados no artigo, podemos realizar algumas conclusões. Primeiramente, pudemos observar que a classe que foi melhor separada foi a classe infantil.

Em relação ao limite entre as classes, podemos concluir que ele não é linear e através da análise das variáveis preditoras, concluímos que não é possível se realizar uma diferenciação entre classes masculino e feminino. Por fim, podemos concluir que as classes com alto grau de sobreposição podem ser encontradas entre os gêneros masculino e feminino.

Dessa forma, somente através da análise de dados foi possível chegar à conclusão de que é impossível separar as classes utilizando apenas as variáveis disponíveis.

IV REFERÊNCIAS

[1] G. James, D. Witten, T. Hastie e R. Tibshirani, An Introduction to Statistical Learning with applications in R. Springer, 2013.

[2] https://operdata.com.br/blog/analise-de-componentes-principais/

[3] https://www.youtube.com/watch?v=KqZAC4jyJKc

[4] https://www.youtube.com/watch?v=u0i0S3h8_a0

[5] http://archive.ics.uci.edu/ml/datasets/Abalone

[6] https://pt.wikipedia.org/wiki/Estat%C3%ADstica_descritiva

[7] http://apps.einstein.br/revista/arquivos/PDF/1595-EC_v8n1p1-2.pdf

[8] https://builtin.com/data-science/step-step-explanation-principal-component-analysis