Find The Right Cigar

You struggle to find a cigar that fits your budget but also your quality expectations ?

In this project, I propose to build a cigars dataset from scratch. Then I will save it as an Excel file that you will be able to filter by your own in order to find cigars that will fit your expectations.

In this notebook I will use :

- Web scraping for collecting data on a website.
- Object oriented programming.
- Data cleaning for making the datas standardized and usable.
- Data analysis for checking data quality and getting insights.

1) Import Python libraries

In [1]:
```
from bs4 import BeautifulSoup
import requests
import re
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt

import warnings;
warnings.filterwarnings('ignore')

%matplotlib inline
```

2) Web Scraping

Cigars are made in almost all latin american and caribbean countries but there are four main producers :

- Cuba.
- Dominican Republic.
- Nicaragua.
- Honduras.

In [2]:
```
url = 'https://www.maison-du-cigare.be/cigares/'
terroirs = ['cubains','dominicains','honduriens','nicaraguayens']
```

In [3]:
```
all_brands = []


for terroir in terroirs:


    page = requests.get(url + terroir + '/')
    soup = BeautifulSoup(page.content,'html.parser')
    tableau = soup.find_all(class_='flex_column_table av-equal-height-column-flextable -flextable')


    for i in range(len(tableau)):

        for j in tableau[i].find_all('a'):
            text = j.text
            text = text.replace("\xa0","").replace("\xab"," ")
            all_brands.append([text,terroir])
```

In [7]:
```
print(all_brands[5:18:2])
```

```
[['Edicion Limitada & Reserva', 'cubains'], ['M. Upmann', 'cubains'], ['Jose L. Piedra', 'cubains'], ['La Flor de Cano', 'cubains'], ['Montecristo Linea 1935', 'cubains'], ['Partagas', 'cubains'], ['Punch', 'cubains']]
```

In [8]:
```
len(all_brands)
```

Out[8]:
```
130
```

In [9]:
```
all_cigars = []
for liste in all_brands:


    page = requests.get("https://www.maison-du-cigare.be/cigares/"+liste[1]+'/'+liste[0]+'/')
    soup = BeautifulSoup(page.content,'html.parser')
    tableau = soup.find_all(class_='av-catalogue-item-inner')


    for i in range(len(tableau)):

        for j in tableau[i].find_all('div',{'class':'av-catalogue-title av-cart-update-title'}):
            l1 = []
            l1.append(liste[0])
            l1.append(liste[1])
            text = j.text
            text = text.replace("\xa0","").replace("\xab","")
            l1.append(text)

            for k in tableau[i].find_all('span',{'class':'woocommerce-Price-amount amount'}):
                text = k.text
                text = text.replace("€","").replace(",",".")
                l1.append(text)

            for n in tableau[i].find_all('div',{'class':'av-catalogue-content'}):
                text = n.text
                text = text.replace(",",".")
                l2 = re.findall(r"[-+]?\d*\.\d+|\d+",text)
                l1 += l2



    all_cigars.append(l1)
```

3) Data quality check & data cleaning

We can observe in each list the following ranking :

- list[0] : Unitary Price
- list[1] : Box Price
- list[2] : Units per Box
- list[3] : Cigar Diameter
- list[4] : Cigar Length
- list[5] : Cigar Name
- list[6] : Brand Name
- list[7] : Origin

In [10]:
```
print(all_cigars[:5])
```

```
[['Bolivar', 'cubains', 'Coronas Junior', '7.48', '185.00', '25', '1.7', '11.0'], ['Bolivar', 'cubains', 'Royal Coronas', '12.50', '312.50', '25', '2.0', '12.4'], ['Bolivar', 'cubains', 'Belicosos Finos', '14.60', '365.00', '25', '2.0', '14.0'], ['Cohiba', 'cubains', 'Siglo I', '12.40', '310.00', '25', '1.6', '10.2'], ['Cohiba', 'cubains', 'Siglo II', '15.50', '387.50', '25', '1.7', '12.9']]
```

In [11]:
```
len(all_cigars[1])
```

Out[11]:
```
8
```

In [12]:
```
lenB = 0
other = 0

for sublist in all_cigars:

    if len(sublist) == 8:
        lenB += 1

    else:
        other += 1

print(other*100/lenB)
```

```
0.0
```

4) Dataset Creation

In [13]:
```
column_name = ['brand', 'origin', 'name','unit_price_eur', 'box_price_eur', 'cig_per_box', 'diameter_cm', 'length_cm']

df = pd.DataFrame(all_cigars, columns=column_name)
```

In [28]:
```
df
```

Out[28]:

	brand	origin	name	unit_price_eur	box_price_eur	cig_per_box	diameter_cm	length_cm
0	Bolivar	cubains	Coronas Junior	7.40	185.00	25	1.7	11.0
1	Bolivar	cubains	Royal Coronas	12.50	312.50	25	2.0	12.4
2	Bolivar	cubains	Belicosos Finos	14.60	365.00	25	2.0	14.0
3	Cohiba	cubains	Siglo I	12.40	310.00	25	1.6	10.2
4	Cohiba	cubains	Siglo II	15.50	387.50	25	1.7	12.9
...								
495	Tatuaje	nicaraguayens	Tattoo Advierno (Gordo)	7.50	375.00	50	2.3	13.8
496	Tatuaje	nicaraguayens	Cojonu 2003	16.00	400.00	25	2.1	16.5
497	Tatuaje	nicaraguayens	Gran Cojonu	18.50	222.00	12	2.5	16.5
498	The T	nicaraguayens	Robusto	15.50	310.00	20	2.0	12.9
499	The T	nicaraguayens	Toro	16.50	330.00	20	2.0	15.4

500 rows × 8 columns

In [22]:
```
for col in  ["unit_price_eur","box_price_eur","cig_per_box","diameter_cm","length_cm"]:
    df[col] = pd.to_numeric(df[col], errors='coerce')
```

In [24]:
```
df.describe()
```

Out[24]:

	unit_price_eur	box_price_eur	cig_per_box	diameter_cm	length_cm
count	500.000000	500.000000	500.000000	500.000000	500.000000
mean	14.063400	238.079800	20.370000	1.996000	13.773700
std	9.299636	139.904798	7.263824	0.267543	2.227806
min	2.300000	23.000000	4.000000	1.000000	8.250000
25%	8.675000	152.375000	20.000000	1.900000	12.500000
50%	11.500000	214.800000	20.000000	2.000000	13.500000
75%	15.925000	300.000000	25.000000	2.200000	15.200000
max	65.000000	995.000000	55.000000	2.600000	21.500000

In [25]:
```
terroirs_dict = {'cubains':'Cuba','dominicains':'Dominican Rep.',
                 'honduriens':'Honduras','nicaraguayens':'Nicaragua'}

for i in range(len(df)):
    if df.iat[i,1] in terroirs_dict:
        df.iat[i,1] = terroirs_dict[df.iat[i,1]]
    else:
        pass
```

In [26]:
```
df.head()
```

Out[26]:

	brand	origin	name	unit_price_eur	box_price_eur	cig_per_box	diameter_cm	length_cm
0	Bolivar	Cuba	Coronas Junior	7.4	185.0	25	1.7	11.0
1	Bolivar	Cuba	Royal Coronas	12.5	312.5	25	2.0	12.4
2	Bolivar	Cuba	Belicosos Finos	14.6	365.0	25	2.0	14.0
3	Cohiba	Cuba	Siglo I	12.4	310.0	25	1.6	10.2
4	Cohiba	Cuba	Siglo II	15.5	387.5	25	1.7	12.9

In [27]:
```
df.to_excel('cigars_dataset.xlsx', index=False)
```

5) Data Analysis

In [29]:
```
df[['origin','unit_price_eur','diameter_cm','length_cm']].groupby(['origin']).describe().transpose()
```

Out[29]:

	origin	Cuba	Dominican Rep.	Honduras	Nicaragua
unit_price_eur	**count**	119.000000	166.000000	58.000000	157.000000
	mean	16.094538	15.237048	18.770690	12.508873
	std	12.332080	9.375013	5.308484	6.843253
	min	2.300000	2.600000	4.500000	4.700000
	25%	8.600000	9.500000	8.125000	8.500000
	50%	13.600000	12.900000	18.000000	10.500000
	75%	17.500000	17.000000	11.000000	14.500000
	max	65.000000	54.000000	33.000000	45.000000
diameter_cm	**count**	119.000000	166.000000	58.000000	157.000000
	mean	1.867227	1.987952	2.044828	2.090446
	std	0.256453	0.277826	0.206196	0.242261
	min	1.000000	1.000000	1.500000	1.500000
	25%	1.700000	1.900000	2.000000	2.000000
	50%	1.900000	2.000000	2.000000	2.100000
	75%	2.050000	2.200000	2.100000	2.200000
	max	2.300000	2.500000	2.400000	2.600000
length_cm	**count**	119.000000	166.000000	58.000000	157.000000
	mean	13.348739	13.783675	13.962069	14.005096
	std	2.338327	2.323589	2.064217	2.065678
	min	9.000000	8.250000	9.000000	8.800000
	25%	11.800000	12.500000	12.700000	12.700000
	50%	12.900000	13.400000	14.000000	14.100000
	75%	14.450000	15.200000	15.300000	15.300000
	max	19.400000	21.500000	18.700000	19.000000

In [30]:
```
dmean = df[['origin','unit_price_eur','length_cm','diameter_cm']].groupby(['origin']).mean()

fig = dmean.plot.bar(edgecolor='black',figsize=(15,4))
fig.tick_params(axis='x', labelrotation = 0)
```

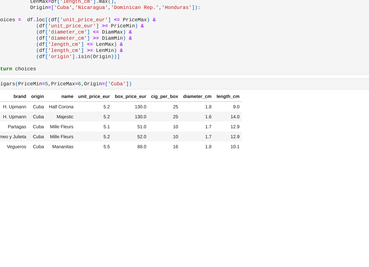
In [76]:
```
fig, (ax1, ax2, ax3) = plt.subplots(1,3,figsize=(15,4))

ax1.hist(df['diameter_cm'])
ax1.set_title('Diameter Repartition')

ax2.hist(df['length_cm'])
ax2.set_title('Length Repartition')

ax3.hist(df['unit_price_eur'])
ax3.set_title('Price Repartition')
```

In [76]:
```
Text(0.5, 1.0, 'Price Repartition')
```



Top 5 : Less and more expensive brands

In [82]:
```
company = df[['brand','unit_price_eur']].groupby(['brand']).mean()
```

In [83]:
```
col = ['unit_price_eur']
print("5 most expensive brands :",company.nlargest(5, col))
print("")
print("5 less expensive brands :",company.nsmallest(5, col))
```

```
5 most expensive brands :                unit_price_eur
brand
Jose L. Piedra           56.562500
Cohiba Behike            56.500000
Cuaba                    44.363333
La Aldea                 35.750000
La Estancia              31.000000

5 less expensive brands :                unit_price_eur
brand
Quorum                    3.000000
San Pedro de Macoris      4.814286
Casa Fernandez            5.250000
Vega Fina                 6.107143
Quesada                   6.500000
```

6) Application

In [86]:
```
def find_cigars(PriceMin=df['unit_price_eur'].min(),
                PriceMax=df['unit_price_eur'].max(),
                DiaMin=df['diameter_cm'].min(),
                DiaMax=df['diameter_cm'].max(),
                LenMin=df['length_cm'].min(),
                LenMax=df['length_cm'].max(),
                Origin=['Cuba','Nicaragua','Dominican Rep.','Honduras']):


    choices = df.loc[(df['unit_price_eur'] >= PriceMax) &
                     (df['unit_price_eur'] <= PriceMin) &
                     (df['diameter_cm'] >= DiaMin) &
                     (df['diameter_cm'] <= DiaMax) &
                     (df['length_cm'] >= LenMin) &
                     (df['length_cm'] <= LenMax) &
                     (df['origin'].isin(Origin))]


    return choices
```

In [89]:
```
find_cigars(PriceMin=5,PriceMax=6,Origin=['Cuba'])
```

Out[89]:

	brand	origin	name	unit_price_eur	box_price_eur	cig_per_box	diameter_cm	length_cm
23	H. Upmann	Cuba	Half Corona	5.2	130.0	25	1.8	9.0
24	H. Upmann	Cuba	Magnolo	5.2	130.0	25	1.6	14.0
67	Partagas	Cuba	Mille Fleurs	5.1	51.0	10	1.7	12.9
94	Romeo y Julieta	Cuba	Mille Fleurs	5.2	52.0	10	1.7	12.9
115	Veguenos	Cuba	Mananitas	5.5	88.0	16	1.8	10.1