Getting data from PDFs

We need to install a few tools for this. And the truth is, Linux is a masterpiece in installing and managing dependencies.

Since I guess you don't all have a Linux box, we can use Google Colab:

https://colab.research.google.com/ (https://colab.research.google.com/)

Camelot (https://camelot-py.readthedocs.io (https://camelot-py.readthedocs.io)) is a fanstastic tool to extract data from PDFs.

An important thing to note is that camelot is only able to process text-based PDFs (the ones that can be selected - not scanned). Scanned PDFs require a preprocessing that we will see later.

In [1]:

```
# dependencies needed by camelot (this works in Google Colab)
#!apt install python-tk python3-tk ghostscript
```

In [2]:

```
# install camelot
#!pip install camelot-py
```

In [3]:

```
import camelot
```

Let's start with a table in a sample pdf: https://camelot-py.readthedocs.io/en/master/_static/pdf/foo.pdf (https://camelot-py.readthedocs.io/en/master/_static/pdf/foo.pdf)

In [4]:

```
foo_url = 'https://cs.unibg.it/data/pdf1.pdf'
tables = camelot.read_pdf(foo_url)
tables
```

Out[4]:

```
<TableList n=1>
```

In [5]:

```
tables[0].parsing_report
```

Out[5]:

```
{'accuracy': 99.02, 'whitespace': 12.24, 'order': 1, 'pag
e': 1}
```

In [6]:

```
tables[0].df
```

Out[6]:

	0	1	2	3	4	5	6
0	Cycle \nName	KI \n(1/km)	Distance \n(mi)	Percent Fuel Savings			
1				Improved \nSpeed	Decreased \nAccel	Eliminate \nStops	Decreased \nIdle
2	2012_2	3.30	1.3	5.9%	9.5%	29.2%	17.4%
3	2145_1	0.68	11.2	2.4%	0.1%	9.5%	2.7%
4	4234_1	0.59	58.7	8.5%	1.3%	8.5%	3.3%
5	2032_2	0.17	57.8	21.7%	0.3%	2.7%	1.2%
6	4171_1	0.07	173.9	58.1%	1.6%	2.1%	0.5%

By default, camelot uses the lattice method, which is based on lines.

Another method is based on spaces (`stream`) and can be selected like this:

In [7]:

```
tables = camelot.read_pdf(foo_url, flavor='stream')
tables
```

Out[7]:

```
<TableList n=1>
```

In [8]:

```
tables[0].parsing_report
```

Out[8]:

```
{'accuracy': 95.87, 'whitespace': 38.1, 'order': 1, 'pag
e': 1}
```

In [9]:

```
tables[0].df
```

Out[9]:

	0	1	2	3	4	5	6
0	reducing the number of stops in high KI cycles...						
1		Table 2-1. Simulated fuel savings from isolate...					
2						Percent Fuel Savings	
3	Cycle	KI	Distance				
4				Improved	Decreased	Eliminate	Decreased
5	Name	(1/km)	(mi)				
6				Speed	Accel	Stops	Idle
7	2012_2	3.30	1.3	5.9%	9.5%	29.2%	17.4%
8	2145_1	0.68	11.2	2.4%	0.1%	9.5%	2.7%
9	4234_1	0.59	58.7	8.5%	1.3%	8.5%	3.3%
10	2032_2	0.17	57.8	21.7%	0.3%	2.7%	1.2%
11	4171_1	0.07	173.9	58.1%	1.6%	2.1%	0.5%

Camelot automatically manage rotated PDFs.

https://camelot-py.readthedocs.io/en/master/_static/pdf/rotated.pdf (https://camelot-py.readthedocs.io/en/master/_static/pdf/rotated.pdf)

In [10]:

```
rotated_url = 'https://cs.unibg.it/data/pdf2.pdf'
tables = camelot.read_pdf(rotated_url)
tables
```

Out[10]:

```
<TableList n=1>
```

In [11]:

```
tables[0].parsing_report
```

Out[11]:

```
{'accuracy': 96.99, 'whitespace': 5.38, 'order': 1, 'pag
e': 1}
```

In [12]:

```
tables[0].df
```

Out[12]:

	0	1	2	3	4	5	6	7
0	State	Nutritional Assessment \n(No. of individuals)				IYCF Practices \n(No. of mothers: \n2011-12)	Blood Pressure \n(No. of adults: \n2011-12)	
1		1975-79	1988-90	1996-97	2011-12		Men	Women
2	Kerala	5738	6633	8864	8297	245	2161	3195
3	Tamil Nadu	7387	10217	5813	7851	413	2134	2858
4	Karnataka	6453	8138	12606	8958	428	2467	2894
5	Andhra Pradesh	5844	9920	9545	8300	557	1899	2493
6	Maharashtra	5161	7796	6883	9525	467	2368	2648
7	Gujarat	4403	5374	4866	9645	477	2687	3021
8	Madhya Pradesh	*	*	*	7942	470	1965	2150
9	Orissa	3756	5540	12024	8473	398	2040	2624
10	West Bengal	*	*	*	8047	423	2058	2743
11	Uttar Pradesh	*	*	*	9860	581	2139	2415
12	Pooled	38742	53618	60601	86898	4459	21918	27041

In [13]:

```
other_url = 'https://cs.unibg.it/data/pdf3.pdf'
tables = camelot.read_pdf(other_url)
tables
```

Out[13]:

```
<TableList n=2>
```

In [14]:

```
tables[0].df
```

Out[14]:

	0	1	2	3
0	Offense charged	Total	Male	Female
1		Under 18 \n18 years \nTotal\nyears\nand over	Under 18 \n18 years \nTotal\nyears\nand over	Under 18 \n18 years \nTotal\nyears\nand over
2	Total .\n .\n . . . . .\n . .\n	11,062 .6\n1,540 .0\n9,522 .6\n467 .9\n69 .1\n...	8,263 .3\n1,071 .6\n7,191 .7\n380 .2\n56 .5\n3...	2,799 .2\n468 .3\n2,330 .9\n87 .7\n12 .6\n75 ....

In [15]:

```
tables = camelot.read_pdf(other_url, flavor='stream')
tables
```

Out[15]:

```
<TableList n=2>
```

In [16]:

```
tables[0].df
```

Out[16]:

	0	1	2	3	4	5	6	7
0	[In thousands (11,062.6 represents 11,062,600)...							
1	Program. Represents arrests reported (not char...							
2	by the FBI. Some persons may be arrested more ...							
3	could represent multiple arrests of the same p...							
4			Total			Male		
5	Offense charged		Under 18	18 years		Under 18	18 years	
6		Total	years	and over	Total	years	and over	Total
7	Total .\n .\n . . . . .\n . .\n	11,062 .6	1,540 .0	9,522 .6	8,263 .3	1,071 .6	7,191 .7	2,799 .2
8	Violent crime . . . . . .\n . .\n...	467 .9	69 .1	398 .8	380 .2	56 .5	323 .7	87 .7
9	Murder and nonnegligent							
10	manslaughter . . . . . .\n. .\n. .\n. .\n....	10.0	0.9	9.1	9.0	0.9	8.1	1.1
11	Forcible rape . . . . . .\n. .\n. .\n. .\n...	17.5	2.6	14.9	17.2	2.5	14.7	–
12	Robbery . . . .\n. .\n. . .\n. . .\n.\n. . .\n...	102.1	25.5	76.6	90.0	22.9	67.1	12.1
13	Aggravated assault . . . . . .\n. .\n. .\n.	338.4	40.1	298.3	264.0	30.2	233.8	74.4
14	Property crime . . . .\n . .\n . . .\n ...	1,396 .4	338 .7	1,057 .7	875 .9	210 .8	665 .1	608 .2

	0	1	2	3	4	5	6	7
15	Burglary . .\n. . . . .\n. .\n.\n. .\n.\n.\n...	240.9	60.3	180.6	205.0	53.4	151.7	35.9
16	Larceny-theft . . . . . .\n. .\n. .\n. .\n...	1,080.1	258.1	822.0	608.8	140.5	468.3	471.3
17	Motor vehicle theft . . . . .\n. .\n. . .\n.\n...	65.6	16.0	49.6	53.9	13.3	40.7	11.7
18	Arson .\n. . . .\n. . .\n.\n. .\n.\n.\n.\n. .\n...	9.8	4.3	5.5	8.1	3.7	4.4	1.7
19	Other assaults .\n. . . . . .\n. . .\n.\n.	1,061.3	175.3	886.1	785.4	115.4	670.0	276.0
20	Forgery and counterfeiting .\n. . . . . .\n.	68.9	1.7	67.2	42.9	1.2	41.7	26.0
21	Fraud .\n.\n.\n. .\n. . . .\n. .\n.\n. .\n.\n....	173.7	5.1	168.5	98.4	3.3	95.0	75.3
22	Embezzlement . . .\n. . . . .\n. . .\n.\n.	14.6	–	14.1	7.2	–	6.9	7.4
23	Stolen property 1 . . . . . .\n. . .\n. .\n...	84.3	15.1	69.2	66.7	12.2	54.5	17.6
24	Vandalism . . . . . .\n. .\n. .\n. .\n.	217.4	72.7	144.7	178.1	62.8	115.3	39.3
25	Weapons; carrying, possessing, etc. .	132.9	27.1	105.8	122.1	24.3	97.8	10.8
26	Prostitution and commercialized vice	56.9	1.1	55.8	17.3	–	17.1	39.6
27	Sex offenses 2 . . . . .\n. . . . .\n. .\n.	61.5	10.7	50.7	56.1	9.6	46.5	5.4
28	Drug abuse violations . . . . . .\n. .\n.\n.	1,333.0	136.6	1,196.4	1,084.3	115.2	969.1	248.7
29	Gambling .\n. . . . .\n. .\n.\n. . .\n.\n.	8.2	1.4	6.8	7.2	1.4	5.9	0.9
30	Offenses against the family and							

	0	1	2	3	4	5	6	7
31	children . . . .\n. . . .\n. .\n. .\n. .\n. .\...	92.4	3.7	88.7	68.9	2.4	66.6	23.4
32	Driving under the influence . . . . . .\n. .	1,158.5	109.2	1,147.5	895.8	8.2	887.6	262.7
33	Liquor laws . . . . . .\n. .\n. .\n. .\n....	48.2	90.2	368.0	326.8	55.4	271.4	131.4
34	Drunkenness . . .\n. . . . .\n. . .\n.\n. . .\...	488.1	11.4	476.8	406.8	8.5	398.3	81.3
35	Disorderly conduct . .\n. . . . . .\n. .\n. ...	529.5	136.1	393.3	387.1	90.8	296.2	142.4
36	Vagrancy . . . .\n. . . . .\n.\n. .\n.\n.\n.	26.6	2.2	24.4	20.9	1.6	19.3	5.7
37	All other offenses (except traffic) . . .\n.	306.1	263.4	2,800.8	2,337.1	194.2	2,142.9	727.0
38	Suspicion . . . .\n. . . .\n. .\n. .\n. .\n.	1.6	–	1.4	1.2	–	1.0	–
39	Curfew and loitering law violations .\n.	91.0	91.0	(X)	63.1	63.1	(X)	28.0
40	Runaways . . . . . .\n. .\n. .\n. .\n. .\...	75.8	75.8	(X)	34.0	34.0	(X)	41.8
41		– Represents zero. X Not applicable. 1 Buying,...						
42		Source: U.S. Department of Justice, Federal Bu...						

In [17]:

```
tables = camelot.read_pdf(other_url, flavor='stream', strip_text='.\n
')
tables
```

Out[17]:

```
<TableList n=2>
```

In [18]:

```
tables[0].df
```

Out[18]:

	0	1	2	3	4	5	6	7
0	[In thousands (11,0626 represents 11,062,600) ...							
1	Program Represents arrests reported (not charg...							
2	by the FBI Some persons may be arrested more t...							
3	could represent multiple arrests of the same p...							
4			Total			Male		F
5	Offense charged		Under 18	18 years		Under 18	18 years	
6		Total	years	and over	Total	years	and over	Total
7	Total	11,062 6	1,540 0	9,522 6	8,263 3	1,071 6	7,191 7	2,799 2
8	Violent crime	467 9	69 1	398 8	380 2	56 5	323 7	87 7
9	Murder and nonnegligent							
10	manslaughter	100	09	91	90	09	81	11
11	Forcible rape	175	26	149	172	25	147	–
12	Robbery	1021	255	766	900	229	671	121
13	Aggravated assault	3384	401	2983	2640	302	2338	744
14	Property crime	1,396 4	338 7	1,057 7	875 9	210 8	665 1	608 2
15	Burglary	2409	603	1806	2050	534	1517	359
16	Larceny-theft	1,0801	2581	8220	6088	1405	4683	4713
17	Motor vehicle theft	656	160	496	539	133	407	117
18	Arson	98	43	55	81	37	44	17
19	Other assaults	1,0613	1753	8861	7854	1154	6700	2760

	0	1	2	3	4	5	6	7
20	Forgery and counterfeiting	689	17	672	429	12	417	260
21	Fraud	1737	51	1685	984	33	950	753
22	Embezzlement	146	–	141	72	–	69	74
23	Stolen property 1	843	151	692	667	122	545	176
24	Vandalism	2174	727	1447	1781	628	1153	393
25	Weapons; carrying, possessing, etc	1329	271	1058	1221	243	978	108
26	Prostitution and commercialized vice	569	11	558	173	–	171	396
27	Sex offenses 2	615	107	507	561	96	465	54
28	Drug abuse violations	1,3330	1366	1,1964	1,0843	1152	9691	2487
29	Gambling	82	14	68	72	14	59	09
30	Offenses against the family and							
31	children	924	37	887	689	24	666	234
32	Driving under the influence	1,1585	1092	1,1475	8958	82	8876	2627
33	Liquor laws	482	902	3680	3268	554	2714	1314
34	Drunkenness	4881	114	4768	4068	85	3983	813
35	Disorderly conduct	5295	1361	3933	3871	908	2962	1424
36	Vagrancy	266	22	244	209	16	193	57
37	All other offenses (except traffic)	3061	2634	2,8008	2,3371	1942	2,1429	7270
38	Suspicion	16	–	14	12	–	10	–
39	Curfew and loitering law violations	910	910	(X)	631	631	(X)	280
40	Runaways	758	758	(X)	340	340	(X)	418
41		– Represents zero X Not applicable 1 Buying, r...						

	0	1	2	3	4	5	6	7
42		Source: US Department of Justice, Federal Bure...						

In [19]:

```
# download the file.
#!wget 'http://cs.unibg.it/data/scan.pdf'
```

In [20]:

```
#!apt install ocrmypdf
```

In [21]:

```
#!ocrmypdf
```

In [22]:

```
#!ocrmypdf scan.pdf output.pdf
```

In [23]:

```
import camelot

import pandas as pd
import seaborn as sns
pd.options.mode.chained_assignment = None
```

In [24]:

```
tables = camelot.read_pdf('output.pdf', flavor='stream')
tables
```

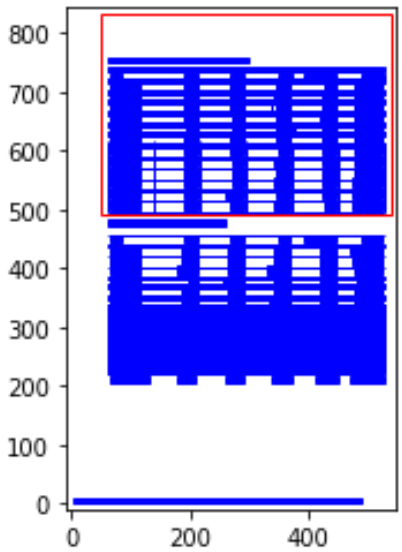
Out[24]:

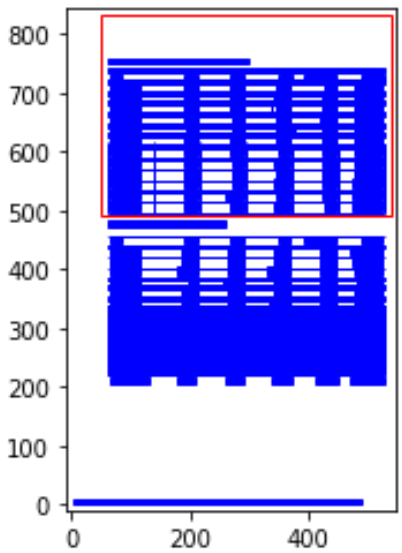
```
<TableList n=1>
```

In [25]:

```
camelot.plot(tables[0], kind='contour')
```

Out[25]:





In [26]:

```
# table_areas accepts strings of the form x1,y1,x2,y2 where (x1, y1)
 -> top-left
# and (x2, y2) -> bottom-right in PDF coordinate space. In PDF coordi
nate space,
# the bottom-left corner of the page is the origin, with coordinates
 (0, 0).
```

In [27]:

```
tables = camelot.read_pdf('output.pdf', flavor='stream',
                          table_areas=['50,800,550,490', '50,490,550,
200'])
tables
```

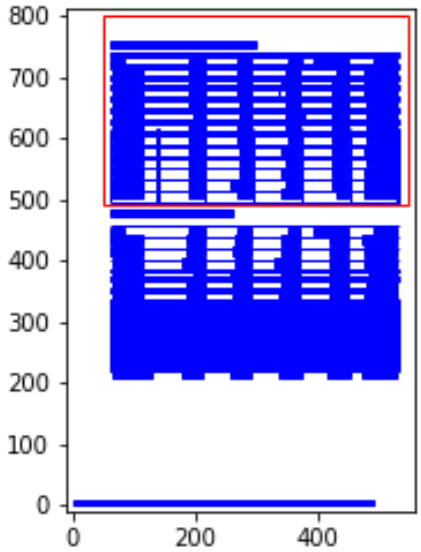
Out[27]:

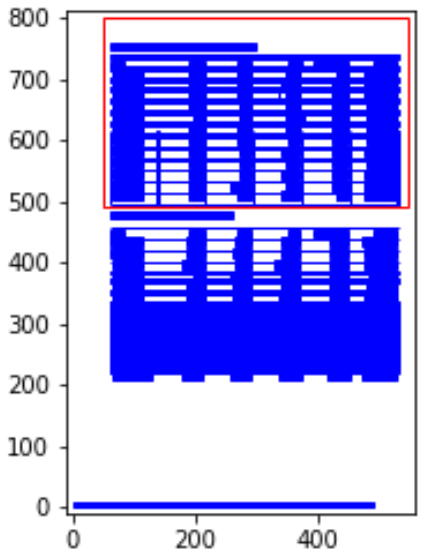
```
<TableList n=2>
```

In [28]:

```
camelot.plot(tables[0], kind='contour')
```

Out[28]:

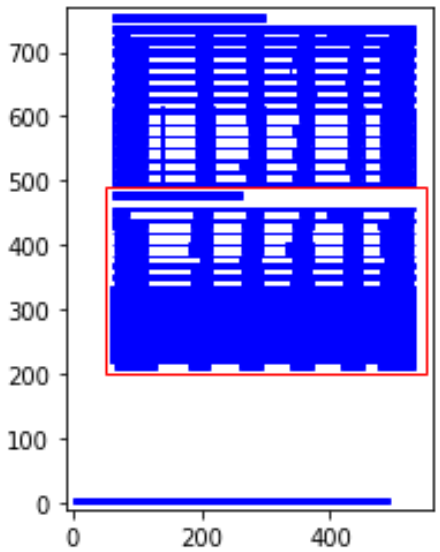


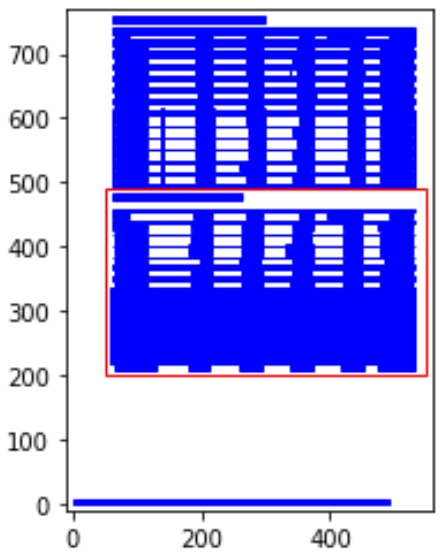


In [29]:

```
camelot.plot(tables[1], kind='contour')
```

Out[29]:





In [30]:

```
tables[0].parsing_report
```

Out[30]:

```
{'accuracy': 88.05, 'whitespace': 18.37, 'order': 1, 'pag
e': 1}
```

In [31]:

```
tables[1].parsing_report
```

Out[31]:

```
{'accuracy': 88.43, 'whitespace': 5.56, 'order': 2, 'pag
e': 1}
```

In [32]:

```
df = tables[0].df
df
```

Out[32]:

	0	1	2	3	4	5	6
0		Table name: daily historical stock prices & vo...					
1	Date		Open	High		Low)	Close Last	Volume
2	01/04/2017		62.48	62.75		62.12	62.3)	21,325,140
3	01/03/2017		62.79	62.84		62.125	62.58}	20,655,190
4	12/30/2016		62.96	62.99	~	62.03	6244|	—- -25)575,720
5	12/29/2016		62.86	63.2		62.73	62.9	10,248,460
6	12/28/2016		63.4	63.4		62.83)	62.99,	14,348,340
7	12/27/2016		63.21	64.07		63.21	63.28)	11,743,650
8	12/23/2016		63.45	63.54		62.8	63.24,	12,399,540
9	12/22/2016		63.84	64.1		63.405	63.55)	22,175,270
10	12/21/2016		63.43	63.7		63.12	63.541	17,084,370
11	12/20/2016		63.69	63.8		63.025	63.54)	26,017,470
12	12/19/2016		62.56	63.77		62.42	63.62}	34,318,500
13	12/16/2016		62.95	62.95		62.115	62.3|	42,452,660

In [33]:

```
df = df.drop(0, axis=0).reset_index(drop=True)
df
```

Out[33]:

	0	1	2	3	4	5	6
0	Date	Open	High		Low)	Close Last	Volume
1	01/04/2017	62.48	62.75		62.12	62.3)	21,325,140
2	01/03/2017	62.79	62.84		62.125	62.58}	20,655,190
3	12/30/2016	62.96	62.99	~	62.03	6244|	—--25)575,720
4	12/29/2016	62.86	63.2		62.73	62.9	10,248,460
5	12/28/2016	63.4	63.4		62.83)	62.99,	14,348,340
6	12/27/2016	63.21	64.07		63.21	63.28)	11,743,650
7	12/23/2016	63.45	63.54		62.8	63.24,	12,399,540
8	12/22/2016	63.84	64.1		63.405	63.55)	22,175,270
9	12/21/2016	63.43	63.7		63.12	63.541	17,084,370
10	12/20/2016	63.69	63.8		63.025	63.54)	26,017,470
11	12/19/2016	62.56	63.77		62.42	63.62}	34,318,500
12	12/16/2016	62.95	62.95		62.115	62.3|	42,452,660

In [34]:

```
df = df.replace(r'[^!-~]+', ' ', regex=True)
df
```

Out[34]:

	0	1	2	3	4	5	6
0	Date	Open	High		Low)	Close Last	Volume
1	01/04/2017	62.48	62.75		62.12	62.3)	21,325,140
2	01/03/2017	62.79	62.84		62.125	62.58}	20,655,190
3	12/30/2016	62.96	62.99	~	62.03	6244|	--25)575,720
4	12/29/2016	62.86	63.2		62.73	62.9	10,248,460
5	12/28/2016	63.4	63.4		62.83)	62.99,	14,348,340
6	12/27/2016	63.21	64.07		63.21	63.28)	11,743,650
7	12/23/2016	63.45	63.54		62.8	63.24,	12,399,540
8	12/22/2016	63.84	64.1		63.405	63.55)	22,175,270
9	12/21/2016	63.43	63.7		63.12	63.541	17,084,370
10	12/20/2016	63.69	63.8		63.025	63.54)	26,017,470
11	12/19/2016	62.56	63.77		62.42	63.62}	34,318,500
12	12/16/2016	62.95	62.95		62.115	62.3|	42,452,660

In [35]:

```
df = df.replace(to_replace=r'[(){}\[\]~|-]', value='', regex=True)
df
```

Out[35]:

	0	1	2	3	4	5	6
0	Date	Open	High		Low	Close Last	Volume
1	01/04/2017	62.48	62.75		62.12	62.3	21,325,140
2	01/03/2017	62.79	62.84		62.125	62.58	20,655,190
3	12/30/2016	62.96	62.99		62.03	6244	25575,720
4	12/29/2016	62.86	63.2		62.73	62.9	10,248,460
5	12/28/2016	63.4	63.4		62.83	62.99,	14,348,340
6	12/27/2016	63.21	64.07		63.21	63.28	11,743,650
7	12/23/2016	63.45	63.54		62.8	63.24,	12,399,540
8	12/22/2016	63.84	64.1		63.405	63.55	22,175,270
9	12/21/2016	63.43	63.7		63.12	63.541	17,084,370
10	12/20/2016	63.69	63.8		63.025	63.54	26,017,470
11	12/19/2016	62.56	63.77		62.42	63.62	34,318,500
12	12/16/2016	62.95	62.95		62.115	62.3	42,452,660

In [36]:

```
df = df.apply(lambda x: x.str.strip())
df
```

Out[36]:

	0	1	2	3	4	5	6
0	Date	Open	High		Low	Close Last	Volume
1	01/04/2017	62.48	62.75		62.12	62.3	21,325,140
2	01/03/2017	62.79	62.84		62.125	62.58	20,655,190
3	12/30/2016	62.96	62.99		62.03	6244	25575,720
4	12/29/2016	62.86	63.2		62.73	62.9	10,248,460
5	12/28/2016	63.4	63.4		62.83	62.99,	14,348,340
6	12/27/2016	63.21	64.07		63.21	63.28	11,743,650
7	12/23/2016	63.45	63.54		62.8	63.24,	12,399,540
8	12/22/2016	63.84	64.1		63.405	63.55	22,175,270
9	12/21/2016	63.43	63.7		63.12	63.541	17,084,370
10	12/20/2016	63.69	63.8		63.025	63.54	26,017,470
11	12/19/2016	62.56	63.77		62.42	63.62	34,318,500
12	12/16/2016	62.95	62.95		62.115	62.3	42,452,660

In [37]:

```
df = df.replace(r',$', '', regex=True)
df
```

Out[37]:

	0	1	2	3	4	5	6
0	Date	Open	High		Low	Close Last	Volume
1	01/04/2017	62.48	62.75		62.12	62.3	21,325,140
2	01/03/2017	62.79	62.84		62.125	62.58	20,655,190
3	12/30/2016	62.96	62.99		62.03	6244	25575,720
4	12/29/2016	62.86	63.2		62.73	62.9	10,248,460
5	12/28/2016	63.4	63.4		62.83	62.99	14,348,340
6	12/27/2016	63.21	64.07		63.21	63.28	11,743,650
7	12/23/2016	63.45	63.54		62.8	63.24	12,399,540
8	12/22/2016	63.84	64.1		63.405	63.55	22,175,270
9	12/21/2016	63.43	63.7		63.12	63.541	17,084,370
10	12/20/2016	63.69	63.8		63.025	63.54	26,017,470
11	12/19/2016	62.56	63.77		62.42	63.62	34,318,500
12	12/16/2016	62.95	62.95		62.115	62.3	42,452,660

In [38]:

```
df = df.replace(r',', '', regex=True)
df
```

Out[38]:

	0	1	2	3	4	5	6
0	Date	Open	High		Low	Close Last	Volume
1	01/04/2017	62.48	62.75		62.12	62.3	21325140
2	01/03/2017	62.79	62.84		62.125	62.58	20655190
3	12/30/2016	62.96	62.99		62.03	6244	25575720
4	12/29/2016	62.86	63.2		62.73	62.9	10248460
5	12/28/2016	63.4	63.4		62.83	62.99	14348340
6	12/27/2016	63.21	64.07		63.21	63.28	11743650
7	12/23/2016	63.45	63.54		62.8	63.24	12399540
8	12/22/2016	63.84	64.1		63.405	63.55	22175270
9	12/21/2016	63.43	63.7		63.12	63.541	17084370
10	12/20/2016	63.69	63.8		63.025	63.54	26017470
11	12/19/2016	62.56	63.77		62.42	63.62	34318500
12	12/16/2016	62.95	62.95		62.115	62.3	42452660

In [39]:

```
df = df.drop(3, axis=1)
df
```

Out[39]:

	0	1	2	4	5	6
0	Date	Open	High	Low	Close Last	Volume
1	01/04/2017	62.48	62.75	62.12	62.3	21325140
2	01/03/2017	62.79	62.84	62.125	62.58	20655190
3	12/30/2016	62.96	62.99	62.03	6244	25575720
4	12/29/2016	62.86	63.2	62.73	62.9	10248460
5	12/28/2016	63.4	63.4	62.83	62.99	14348340
6	12/27/2016	63.21	64.07	63.21	63.28	11743650
7	12/23/2016	63.45	63.54	62.8	63.24	12399540
8	12/22/2016	63.84	64.1	63.405	63.55	22175270
9	12/21/2016	63.43	63.7	63.12	63.541	17084370
10	12/20/2016	63.69	63.8	63.025	63.54	26017470
11	12/19/2016	62.56	63.77	62.42	63.62	34318500
12	12/16/2016	62.95	62.95	62.115	62.3	42452660

In [40]:

```
columns = df.loc[0]
df = df.loc[1:]
df.columns = columns
df
```

Out[40]:

	Date	Open	High	Low	Close Last	Volume
1	01/04/2017	62.48	62.75	62.12	62.3	21325140
2	01/03/2017	62.79	62.84	62.125	62.58	20655190
3	12/30/2016	62.96	62.99	62.03	6244	25575720
4	12/29/2016	62.86	63.2	62.73	62.9	10248460
5	12/28/2016	63.4	63.4	62.83	62.99	14348340
6	12/27/2016	63.21	64.07	63.21	63.28	11743650
7	12/23/2016	63.45	63.54	62.8	63.24	12399540
8	12/22/2016	63.84	64.1	63.405	63.55	22175270
9	12/21/2016	63.43	63.7	63.12	63.541	17084370
10	12/20/2016	63.69	63.8	63.025	63.54	26017470
11	12/19/2016	62.56	63.77	62.42	63.62	34318500
12	12/16/2016	62.95	62.95	62.115	62.3	42452660

In [41]:

```
df.columns
```

Out[41]:

```
Index(['Date', 'Open', 'High', 'Low', 'Close Last', 'Volu
me'], dtype='object', name=0)
```

In [42]:

```
df[['Date']] = df[['Date']].apply(pd.to_datetime)
df
```

Out[42]:

	Date	Open	High	Low	Close Last	Volume
1	2017-01-04	62.48	62.75	62.12	62.3	21325140
2	2017-01-03	62.79	62.84	62.125	62.58	20655190
3	2016-12-30	62.96	62.99	62.03	6244	25575720
4	2016-12-29	62.86	63.2	62.73	62.9	10248460
5	2016-12-28	63.4	63.4	62.83	62.99	14348340
6	2016-12-27	63.21	64.07	63.21	63.28	11743650
7	2016-12-23	63.45	63.54	62.8	63.24	12399540
8	2016-12-22	63.84	64.1	63.405	63.55	22175270
9	2016-12-21	63.43	63.7	63.12	63.541	17084370
10	2016-12-20	63.69	63.8	63.025	63.54	26017470
11	2016-12-19	62.56	63.77	62.42	63.62	34318500
12	2016-12-16	62.95	62.95	62.115	62.3	42452660

In [43]:

```
df[['Open', 'High', 'Low', 'Close Last', 'Volume']] = \
    df[['Open', 'High', 'Low', 'Close Last', 'Volume']].apply(pd.to_n
umeric)
```

In [44]:

```
df
```

Out[44]:

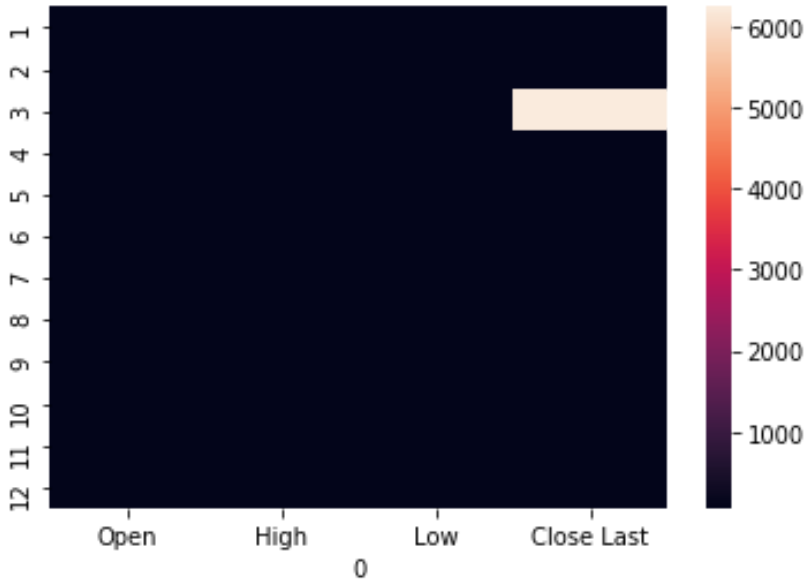
	Date	Open	High	Low	Close Last	Volume
1	2017-01-04	62.48	62.75	62.120	62.300	21325140
2	2017-01-03	62.79	62.84	62.125	62.580	20655190
3	2016-12-30	62.96	62.99	62.030	6244.000	25575720
4	2016-12-29	62.86	63.20	62.730	62.900	10248460
5	2016-12-28	63.40	63.40	62.830	62.990	14348340
6	2016-12-27	63.21	64.07	63.210	63.280	11743650
7	2016-12-23	63.45	63.54	62.800	63.240	12399540
8	2016-12-22	63.84	64.10	63.405	63.550	22175270
9	2016-12-21	63.43	63.70	63.120	63.541	17084370
10	2016-12-20	63.69	63.80	63.025	63.540	26017470
11	2016-12-19	62.56	63.77	62.420	63.620	34318500
12	2016-12-16	62.95	62.95	62.115	62.300	42452660

In [45]:

```
sns.heatmap(df[['Open', 'High', 'Low', 'Close Last']])
```

Out[45]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x132900a50>
```



In [46]:

```
def fix(n):
    if n > 100:
        return n / 100
    else:
        return n
```

In [47]:

```
df[['Open', 'High', 'Low', 'Close Last']] = \
    df[['Open', 'High', 'Low', 'Close Last']].applymap(fix)
```

In [48]:

```
df
```

Out[48]:

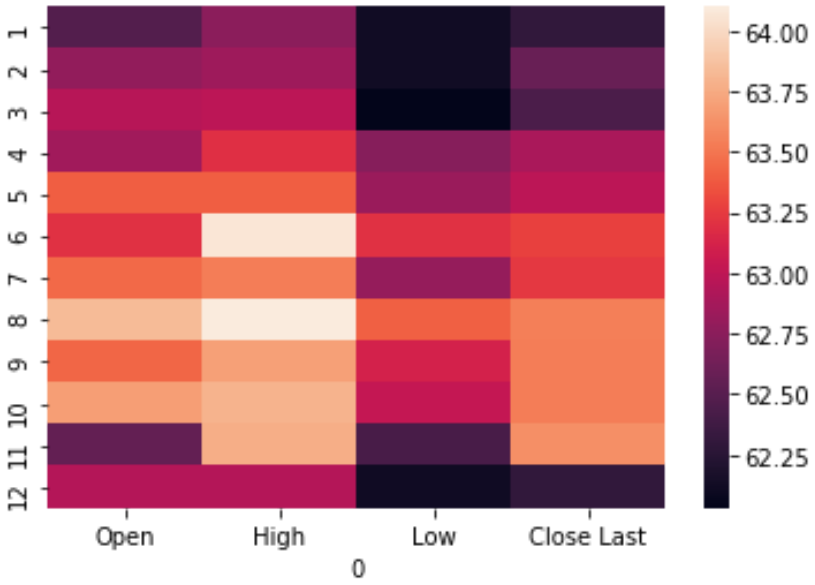
	Date	Open	High	Low	Close Last	Volume
1	2017-01-04	62.48	62.75	62.120	62.300	21325140
2	2017-01-03	62.79	62.84	62.125	62.580	20655190
3	2016-12-30	62.96	62.99	62.030	62.440	25575720
4	2016-12-29	62.86	63.20	62.730	62.900	10248460
5	2016-12-28	63.40	63.40	62.830	62.990	14348340
6	2016-12-27	63.21	64.07	63.210	63.280	11743650
7	2016-12-23	63.45	63.54	62.800	63.240	12399540
8	2016-12-22	63.84	64.10	63.405	63.550	22175270
9	2016-12-21	63.43	63.70	63.120	63.541	17084370
10	2016-12-20	63.69	63.80	63.025	63.540	26017470
11	2016-12-19	62.56	63.77	62.420	63.620	34318500
12	2016-12-16	62.95	62.95	62.115	62.300	42452660

In [49]:

```
sns.heatmap(df[['Open', 'High', 'Low', 'Close Last']])
```

Out[49]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1329ff250>
```



In [50]:

```
df2 = tables[1].df
df2
```

Out[50]:

	0	1	2	3	
0		http://www.nasdaq.com/symbol/fb/historical			
1	Date	Open	High	Low,	(
2	01/04/2017	117.55	119.66	117.29	11
3	01/03/2017	116.03	117.84	115.51	1
4	12/30/2016	116.595	116.83	114.7739	11
5	12/29/2016	117	117.531	116.06	1
6	12/28/2016	118.19	118.25	116.65	11
7	12/27/2016	116.96	118.68	116.864	11
8	12/23/2016	117	117.56	116.3	11
9	12/22/2016	118.86	118.99	116.93	4
10	12/21/2016	118,92	119.2	118.48	11
11	12/20/2016	119.5	119.77	118.8	11
12	12/19/2016	119,85	120.36	118.51	11
13	12/16/2016	120.9	121.5	119.27	11
14	\ 2/20/2016	A24,A	A245	AAS 2	M

In [51]:

```
#from google.colab import drive
#drive.mount('/content/drive')
```

In [52]:

```
#df2.to_excel('/content/drive/My Drive/AEM/df2.xlsx')
```

In [53]:

```
#df2 = pd.read_excel('/content/drive/My Drive/AEM/df2.xlsx')
```

In [54]:

```
#sns.heatmap(df2[['Open', 'High', 'Low', 'Close / Last']])
```