Analysis

January 13, 2019

1 Descriptive Statistics For "CVE Details" Site

1.1 Parts of the work

1.1.1 1) run script scrapProducts.py for scraping products for each year in the range 1999-2019. This script creates many xslx-files (1999_31_5.xlsx, 1999_36_23.xlsx, ...) for each products:

```
Number  Product           Product ID Vendor ID
1   Debian Linux Debian      36     23
2   Linux Kernel Linux       47     33
4   Mac Os X Apple           156    49
8   Ubuntu Linux Canonical   20550  4781
12  Windows 7 Microsoft      17153  26
15  Opensuse Opensuse        14195  8184
17  Windows Vista Microsoft  9591   26
22  Windows 10 Microsoft     32238  26
23  Windows Xp Microsoft     739    26
25  Windows 8.1 Microsoft    26434  26
28  Mac Os X Server Apple    2274   49
34  Enterprise Linux Redhat  78     25
43  Windows 2000 Microsoft   107    26
48  Solaris SUN              31     5
```

2) run script createGeneralDF.py. This script creates All_Data_for_Analysis.xlsx from xslx-files (1999_31_5.xlsx, 1999_36_23.xlsx, ...). ### 3) run this jupyter notebook for the analysis. The file All_Data_for_Analysis.xlsx from previous step is input file. ### 4) the script userDialog.py produces 1 ouput file for the single product_id, vendor_id, year. For example, this is user dialog during the script running:

```
Please enter the year: 2008
Please enter product_id: 156
Please enter vender_id: 49
The file 2008_156_49.xlsx has created
>>
```

This script gives opportunity to add the new file before combine all xlsx-files to All_Data_for_Analysis.xlsx.

1.2 Import modules

```
In [1]: # -*- coding: utf-8 -*-
        %matplotlib inline
        import pandas as pd
```

```
import numpy as np
from pandas import ExcelWriter
import matplotlib.pyplot as plt
import matplotlib.ticker as mtick
```

1.3 Read data and create dataframe

```
In [2]: df=pd.read_excel('All_Data_for_Analysis.xlsx', "CVE Details")
        print(len(df.index))
```

```
13888
```

1.4 Descriptive Statistics

```
In [3]: df.head()
```

Out[3]:

	CVE ID Number	Publish Date	Update Date	Software Type	Vendor	Product
0	CVE-1999-0914	1999-01-03	2008-09-09	OS	Debian	Debian Linux
1	CVE-1999-0389	1999-01-03	2008-09-09	OS	Debian	Debian Linux
2	CVE-1999-0678	1999-01-17	2008-09-09	OS	Debian	Debian Linux
3	CVE-1999-0457	1999-01-17	2008-09-09	OS	Debian	Debian Linux
4	CVE-1999-0373	1999-02-01	2008-09-09	OS	Debian	Debian Linux

	Version	CVSS Score	Confidentiality Impact	Integrity Impact	...
0	2.0	7.2	Complete	Complete	...
1	2.0	7.2	Complete	Complete	...
2	4.0	5.0	Partial	None	...
3	2.0	7.2	Complete	Complete	...
4	2.0	7.2	Complete	Complete	...

	Access Complexity	Authentication	Gained Access	Vulnerability Type
0	Low	Not required	Admin	Overflow
1	Low	Not required	Admin	Overflow
2	Low	Not required	None	NaN
3	Low	Not required	Admin	Gain privileges
4	Low	Not required	Admin	Execute CodeOverflow

	Summary Text	filename
0	\tBuffer overflow in the FTP client in the Deb...	1999_36_23.xlsx
1	\tBuffer overflow in the bootp server in the D...	1999_36_23.xlsx
2	\tA default configuration of Apache on Debian ...	1999_36_23.xlsx
3	\tLinux ftpwatch program allows local users to...	1999_36_23.xlsx
4	\tBuffer overflow in the "Super" utility in De...	1999_36_23.xlsx

	product_vender	days	score level	year
0	36_23	3538	high	1999
1	36_23	3538	high	1999
2	36_23	3524	medium	1999
3	36_23	3524	high	1999
4	36_23	3509	high	1999

[5 rows x 21 columns]

```
In [4]: df.tail()
```

Out[4]:

	CVE ID Number	Publish Date	Update Date	Software Type	Vendor
13883	CVE-2018-19208	2018-11-12	2018-12-13	OS	Suse
13884	CVE-2018-19214	2018-11-12	2018-12-13	OS	Redhat
13885	CVE-2018-19215	2018-11-12	2018-12-13	OS	Redhat
13886	CVE-2018-16850	2018-11-13	2018-12-18	OS	Redhat
13887	CVE-2018-19478	2019-01-02	2019-01-11	OS	Debian

	Product	Version	CVSS Score
13883	Suse Linux Enterprise Server	11	4.3
13884	Enterprise Linux	7.0	6.8
13885	Enterprise Linux	7.0	6.8
13886	Enterprise Linux	7.6	7.5
13887	Debian Linux	8	4.3

	Confidentiality Impact	Integrity Impact	...	Access Complexity
13883	None	None	...	Medium
13884	Partial	Partial	...	Medium
13885	Partial	Partial	...	Medium
13886	Partial	Partial	...	Low
13887	None	None	...	Medium

	Authentication	Gained Access	Vulnerability Type
13883	Not required	None	Denial Of Service
13884	Not required	None	NaN
13885	Not required	None	NaN
13886	Not required	None	Sql Injection
13887	Not required	None	NaN

	Summary Text	filename
13883	\tIn libwpd 0.10.2, there is a NULL pointer de...	2018_78_25.xlsx
13884	\tNetwide Assembler (NASM) 2.14rc15 has a heap...	2018_78_25.xlsx
13885	\tNetwide Assembler (NASM) 2.14rc16 has a heap...	2018_78_25.xlsx
13886	\tpostgresql before versions 11.1, 10.6 is vul...	2018_78_25.xlsx
13887	\tIn Artifex Ghostscript before 9.26, a carefu...	2019_36_23.xlsx

	product_vender	days	score level	year
13883	78_25	31	medium	2018
13884	78_25	31	medium	2018
13885	78_25	31	medium	2018
13886	78_25	35	high	2018
13887	36_23	9	medium	2019

[5 rows x 21 columns]

In [5]:
```
# summary statistics of character columns only
df.describe(include=['object'])
```

Out[5]:

	CVE ID Number	Software Type	Vendor	Product	Version
count	13888	13888	13888	13888	10536
unique	9604	3	335	522	1332
top	CVE-2015-1819	OS	Microsoft	Linux Kernel	-
freq	5	11824	4623	2058	794

	Confidentiality Impact	Integrity Impact	Availibility Impact
count	13888	13888	13888

```
unique                     3               3                 3
top                 Complete            None          Complete
freq                    5009            5202              6116

       Access Complexity Authentication Gained Access Vulnerability Type  \
count              13888          13888         13888               11787
unique                 3              3             3                  91
top                  Low   Not required          None  Denial Of Service
freq                7799          13170         12331                2467

                                                Summary Text  \
count                                                  13888
unique                                                  9442
top     \tRace condition in win32k.sys in the kernel-m...
freq                                                      84

                    filename product_vender score level
count                  13888          13888       13888
unique                   191             14           3
top     2018_20550_4781.xlsx          47_33      medium
freq                     484           2158        6227
```

In [6]:
```
# summary statistics of all columns
df.describe(include='all')
```

Out[6]:
```
        CVE ID Number         Publish Date          Update Date Software Type  \
count           13888                13888                13888         13888
unique           9604                 2030                  750             3
top     CVE-2015-1819  2017-03-16 00:00:00  2018-10-30 00:00:00            OS
freq                5                  189                 2672         11824
first             NaN  1999-01-01 00:00:00  2008-09-05 00:00:00           NaN
last              NaN  2019-01-02 00:00:00  2019-01-12 00:00:00           NaN
mean              NaN                  NaN                  NaN           NaN
std               NaN                  NaN                  NaN           NaN
min               NaN                  NaN                  NaN           NaN
25%               NaN                  NaN                  NaN           NaN
50%               NaN                  NaN                  NaN           NaN
75%               NaN                  NaN                  NaN           NaN
max               NaN                  NaN                  NaN           NaN

           Vendor       Product Version    CVSS Score Confidentiality Impact  \
count       13888         13888   10536  13888.000000                  13888
unique        335           522    1332           NaN                      3
top     Microsoft  Linux Kernel       -           NaN               Complete
freq         4623          2058     794           NaN                   5009
first         NaN           NaN     NaN           NaN                    NaN
last          NaN           NaN     NaN           NaN                    NaN
mean          NaN           NaN     NaN      6.095536                    NaN
std           NaN           NaN     NaN      2.231697                    NaN
min           NaN           NaN     NaN      1.200000                    NaN
25%           NaN           NaN     NaN      4.600000                    NaN
50%           NaN           NaN     NaN      6.800000                    NaN
75%           NaN           NaN     NaN      7.500000                    NaN
max           NaN           NaN     NaN     10.000000                    NaN
```

```
       Integrity Impact      ...      Access Complexity Authentication  \
count             13888      ...                 13888          13888
unique                3      ...                     3              3
top                None      ...                   Low   Not required
freq               5202      ...                  7799          13170
first               NaN      ...                   NaN            NaN
last                NaN      ...                   NaN            NaN
mean                NaN      ...                   NaN            NaN
std                 NaN      ...                   NaN            NaN
min                 NaN      ...                   NaN            NaN
25%                 NaN      ...                   NaN            NaN
50%                 NaN      ...                   NaN            NaN
75%                 NaN      ...                   NaN            NaN
max                 NaN      ...                   NaN            NaN

       Gained Access Vulnerability Type  \
count          13888              11787
unique             3                 91
top             None  Denial Of Service
freq           12331               2467
first            NaN                NaN
last             NaN                NaN
mean             NaN                NaN
std              NaN                NaN
min              NaN                NaN
25%              NaN                NaN
50%              NaN                NaN
75%              NaN                NaN
max              NaN                NaN

                                           Summary Text  \
count                                             13888
unique                                             9442
top     \tRace condition in win32k.sys in the kernel-m...
freq                                                 84
first                                               NaN
last                                                NaN
mean                                                NaN
std                                                 NaN
min                                                 NaN
25%                                                 NaN
50%                                                 NaN
75%                                                 NaN
max                                                 NaN

                  filename product_vender          days  score level  \
count                13888          13888  13888.000000        13888
unique                 191             14           NaN            3
top     2018_20550_4781.xlsx          47_33           NaN       medium
freq                   484           2158           NaN         6227
first                  NaN            NaN           NaN          NaN
last                   NaN            NaN           NaN          NaN
mean                   NaN            NaN   1580.664099          NaN
std                    NaN            NaN   1675.213523          NaN
```

```
min                     NaN            NaN       0.000000          NaN
25%                     NaN            NaN     153.000000          NaN
50%                     NaN            NaN     948.000000          NaN
75%                     NaN            NaN    2758.000000          NaN
max                     NaN            NaN    7237.000000          NaN

                year
count   13888.000000
unique           NaN
top              NaN
freq             NaN
first            NaN
last             NaN
mean     2012.703413
std         4.683774
min      1999.000000
25%      2009.000000
50%      2015.000000
75%      2016.000000
max      2019.000000

[13 rows x 21 columns]
```

```
In [7]: df['score level'].describe()
```

```
Out[7]: count      13888
        unique         3
        top       medium
        freq        6227
        Name: score level, dtype: object
```

```
In [8]: df['days'].describe()
```

```
Out[8]: count    13888.000000
        mean      1580.664099
        std       1675.213523
        min          0.000000
        25%        153.000000
        50%        948.000000
        75%       2758.000000
        max       7237.000000
        Name: days, dtype: float64
```

df['CVSS Score'].describe()

```
In [9]: #List unique values in the df['Vendo'] column
        df.Vendor.unique()
```

```
Out[9]: array(['Debian', 'Washington University', 'Suse', 'Linux', 'Todd Miller',
               'Earl Hood', 'Redhat', 'SGI', 'Paul Kranenburg', 'Apple',
               'Microsoft', 'SUN', 'Turbolinux', 'Freebsd', 'Trustix',
               'Slackware', 'Sam Lantinga', 'Mandrakesoft', 'Zope', 'Proftpd',
               'Progeny', 'University Of Cambridge', 'Immunix', 'John Bovey',
               'Oracle', 'Netbsd', 'Easy Software Products', 'Winzip',
               'Xi Graphics', 'Xfree86 Project', 'Sendmail', 'Michael Jennings',
               'Semi', 'Openpkg', 'Openbsd', 'Xpdf', 'Windriver', 'FTE', 'Perl',
```

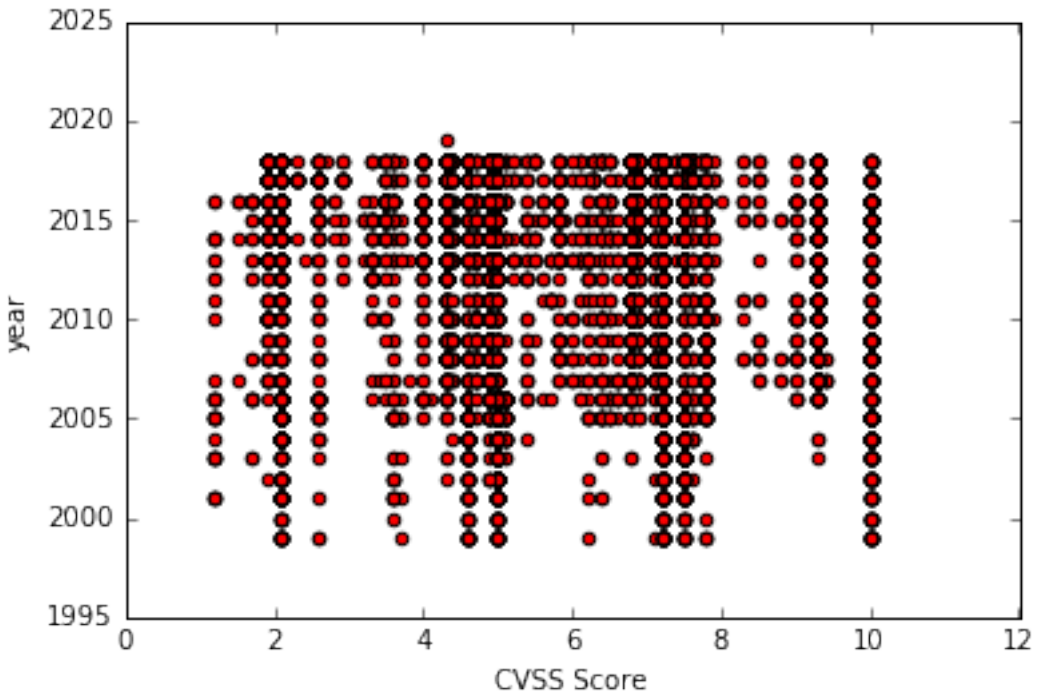
'Gallery Project', 'William Deich', 'Webmin', 'Mysql', 'Pavuk',
'SUP', 'Mailreader.com', 'Roaring Penguin', 'Samba', 'Ubuntu',
'Openldap', 'Vmware', 'Wxgtk2', 'Nortel', 'SOX', 'Tinysofa',
'Dgen', 'Viewcvs', 'ZGV', 'SCO', 'Gentoo', 'Zinf', 'Squid-cache',
'Sukria', 'Mantis', 'GNU', 'Vserver', 'Opendarwin',
'Perry Kiehtreiber', 'Yamaha', 'Sylpheed-claws', 'Rob Flynn',
'Wietse Venema', 'Mozilla', 'Openssl', 'L2tpns', 'Thomas Lange',
'Next', 'Windows', 'X.org', 'Phpmyadmin', 'Rpath', 'Postgresql',
'PHP', 'Digium', 'Omnigroup', 'Pcre', 'Canonical', 'Opensuse',
'Openafs', 'Yassl', 'Lighttpd', 'Ruby-lang', 'Moodle', 'Opengroup',
'Cisco', 'VIM', 'Xfig', 'Google', 'Sebastian Heinlein',
'Eucalyptus', 'Mahara', 'Wireshark', 'Viewvc', 'Novell',
'Libexpat', 'Gnome', 'ISC', 'XEN', 'Drupal', 'Pizzashack',
'Rubyonrails', 'X', 'Winscp', 'Libtiff', 'Konstanty Bialkowski',
'Ubuntu Developers', 'Wouter Verhelst', 'Ffmpeg', 'Haxx',
'Phil Schwartz', 'Qualcomm', 'Net-snmp', 'Openstack', 'Digia',
'Puppetlabs', 'Transmissionbt', 'Xmlsoft', 'Djangoproject',
'Python', 'Perlmonks', 'Spice Project', 'Jean-paul Calderone',
'David King', 'Marc Deslauriers', 'Michael Vogt', 'Evan Dandrea',
'Martin Pitt', 'Radscan', 'Gnupg', 'Quassel-irc', 'Freedesktop',
'Robert Ancell', 'Opus-codec', 'Plataformatec', 'Strongswan',
'Openvpn', 'Percona', 'Squirrelmail', 'Xinetd', 'Rubygems',
'Scientificlinux', 'Linuxfoundation', 'Opensuse Project',
'Mantisbt', 'Polarssl', 'Wordpress', 'Libvncserver', 'Mageia',
'W1.fi', 'Pidgin', 'Mageia Project', 'Jquery', 'Libreoffice',
'Graphviz', 'Lsyncd Project', 'Unrtf Project', 'Nlnetlabs',
'Powerdns', 'Sixapart', 'RPM', 'Nvidia', 'Httplib2 Project',
'Gpsd Project', 'Freetype', 'Qemu', 'Gdm-guest-session Project',
'Fedoraproject', 'KDE', 'Serf Project', 'Procmail', 'Libvirt',
'Chkrootkit', 'Python Bugzilla Project',
'Standards Based Linux Instrumentation Project', 'Pyyaml', 'Otrs',
'Travis Shirk', 'Phppgadmin Project', 'Openvas', 'Urs Wolfer',
'Redcloth', 'Websvn', 'Privoxy', 'Info-zip', 'Typo3',
'E2fsprogs Project', 'Pfsense', 'Libssh2', 'Tcpdump',
'Simon Tatham', 'Rxspencer Project', 'Gaia-gis', 'Dulwich Project',
'Shibboleth', 'Tuxfamily', 'Sqlite', 'HP', 'Xiph', 'Owncloud',
'Xml-libxml Project', 'Openinfosecfoundation', 'Redislabs',
'Fuse Project', 'Haproxy', 'Rack Project', 'Htacg',
'Libevent Project', 'Netfilter', 'Rpcbind Project',
'Phpmailer Project', 'Libpng', 'Icu-project', 'Linuxcontainers',
'Opentype Sanitiser Project', 'Python-requests', 'Oxide Project',
'Clamav', 'Module-signature Project', 'T1utils Project', 'Libav',
'Mediawiki', 'Libvdpau Project', 'Simpestreams Project',
'Xscreensaver Project', 'Wvware',
'Standards Based Linux Instrumentation', 'Polkit Project',
'Roundcube', 'Prosody', 'SIL', 'Didiwiki Project', 'Quagga',
'Fuseiso Project', 'Spip', 'Dhcpcd Project', 'Oar Project',
'Kamailio', 'Redmine', 'Inspire Ircd', 'Libssh', 'Remotesensing',
'Roundup-tracker', 'Xymon', 'Python Imaging Project', 'Tryton',
'Horde', 'Optipng Project', 'Optipng', 'Xdelta', 'Varnish-cache',
'Libgd', 'Libpam-sshauth', 'Tardiff Project', 'Mercurial',
'Ikiwiki', 'Enlightenment', 'Xstream Project', 'Sensiolabs',
'Zend', 'Nginx', 'Videolan', 'Fontconfig Project', 'Libarchive',
'Flex Project', 'Openjpeg', 'Irssi', 'Unadf Project', 'Xrdp',

```
 'Nghttp2', 'Sophos', 'Filemaker', 'Pygments', 'GTK',
 'Jasper Project', 'Pixman', 'Xchat', 'Imagemagick',
 'Libksba Project', 'Thekelleys', 'Gnu Wget Project', 'Ecryptfs',
 'Paolo Bacchilega', 'Systemd Project', 'Wolfssl', 'NTP',
 'Shotwell Project', 'Moinmo', 'Exim', 'Click Project', 'Muscle',
 'Lightdm Project', 'Nasm', 'Libjpeg-turbo', 'Sosreport Project',
 'Vgough', 'Simplejson Project', 'Torproject', 'Dovecot',
 'Memcached', 'Rsyslog', 'ZSH', 'Webkitgtk', 'Qpdf Project',
 'Libraw', 'Liblouis', 'Git-scm', 'File Project', 'Exiv2',
 'Perl-archive-zip Project', 'Libsoup Project', 'Mutt',
 'Lftp Project', 'Xkbcommon', 'Fig2dev Project', 'Orcamo',
 'Pdfinfo Project', 'Netapp', 'Freerdp', 'Lxml', 'Wavpack', 'Tivo',
 'Nomachine', 'Selinux Project', 'Sprockets Project', 'Zmanda'],
dtype=object)
```

1.5 Plots

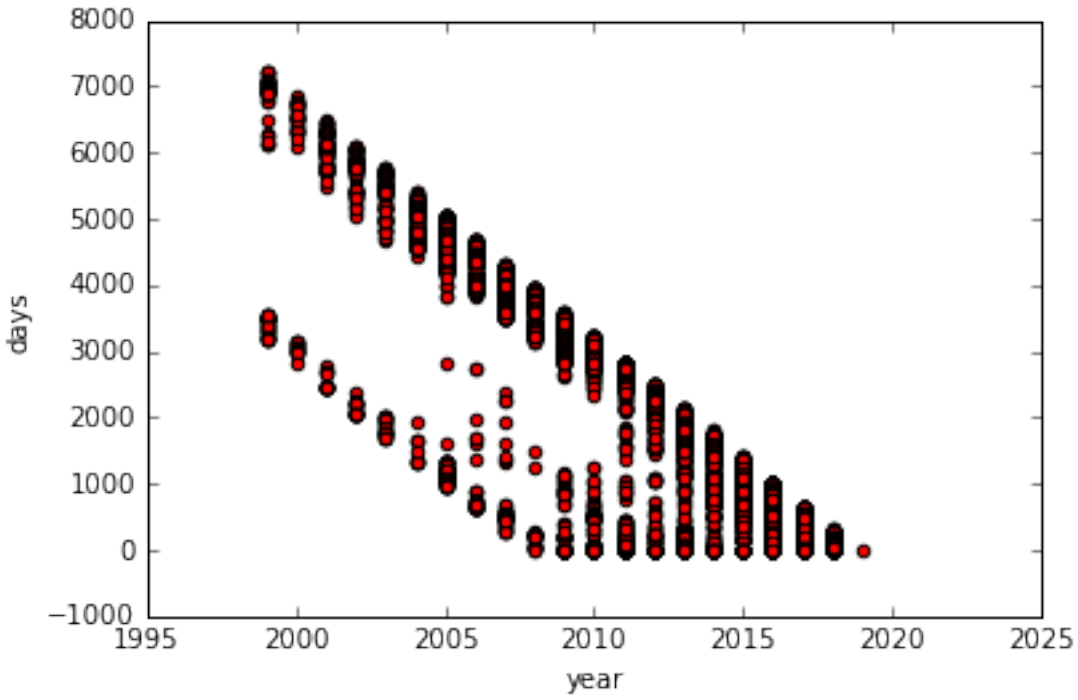
```
In [10]: df.plot(kind='scatter',x='CVSS Score',y='year',color='red')
```

```
Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x124909cc0>
```



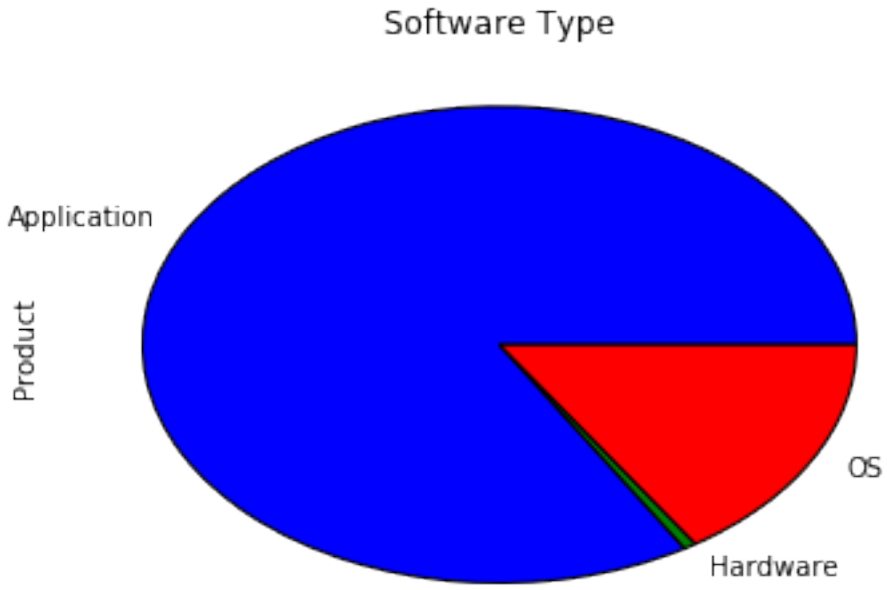

```
In [11]: df.plot(kind='scatter',x='year',y='days',color='red')
```

```
Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x1257621d0>
```



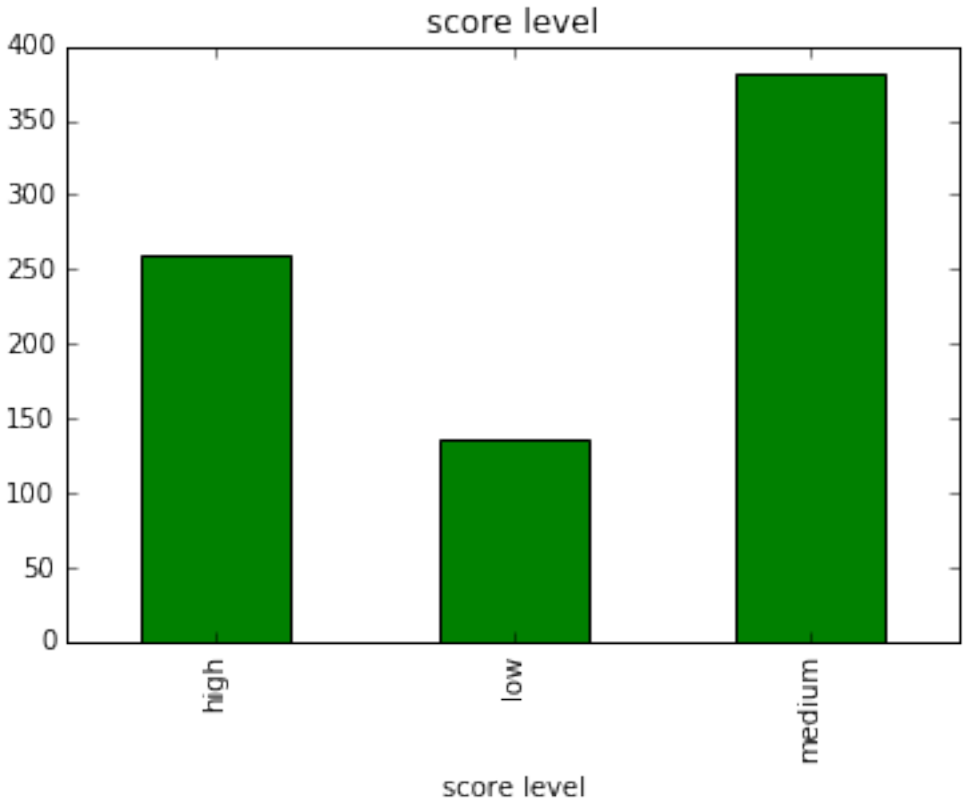

```
In [12]: df.groupby('Software Type')['Product'].nunique().plot(kind='pie')
         plt.title('Software Type')
```

```
Out[12]: <matplotlib.text.Text at 0x1262d20b8>
```



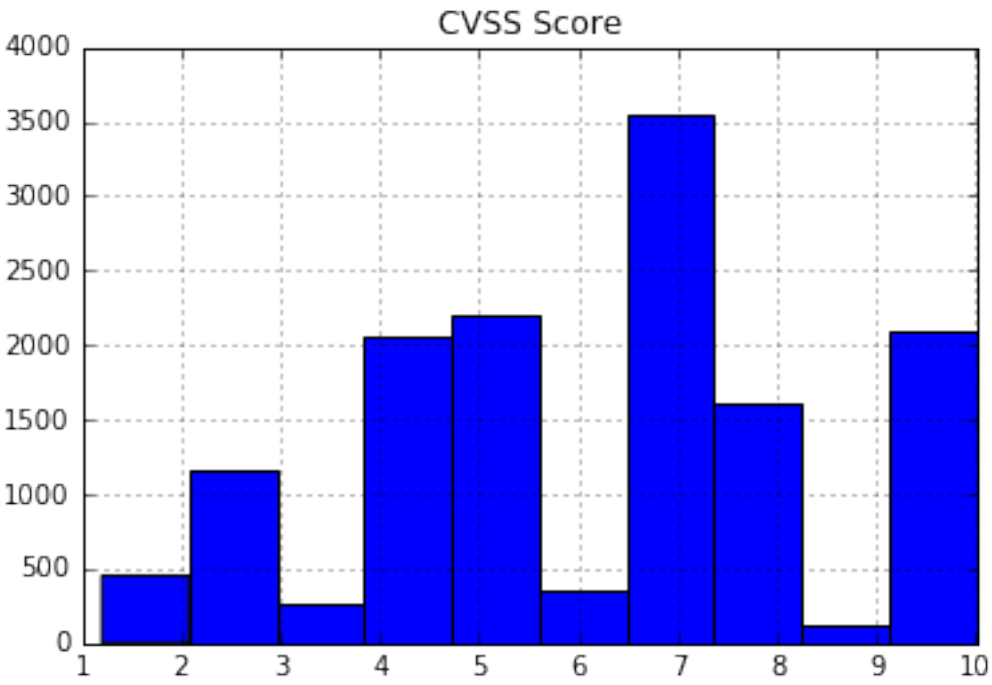
In [13]:
```
df.groupby('score level')['Product'].nunique().plot(kind='bar',color='green')
plt.title('score level')
```

Out[13]: <matplotlib.text.Text at 0x1257c5780>

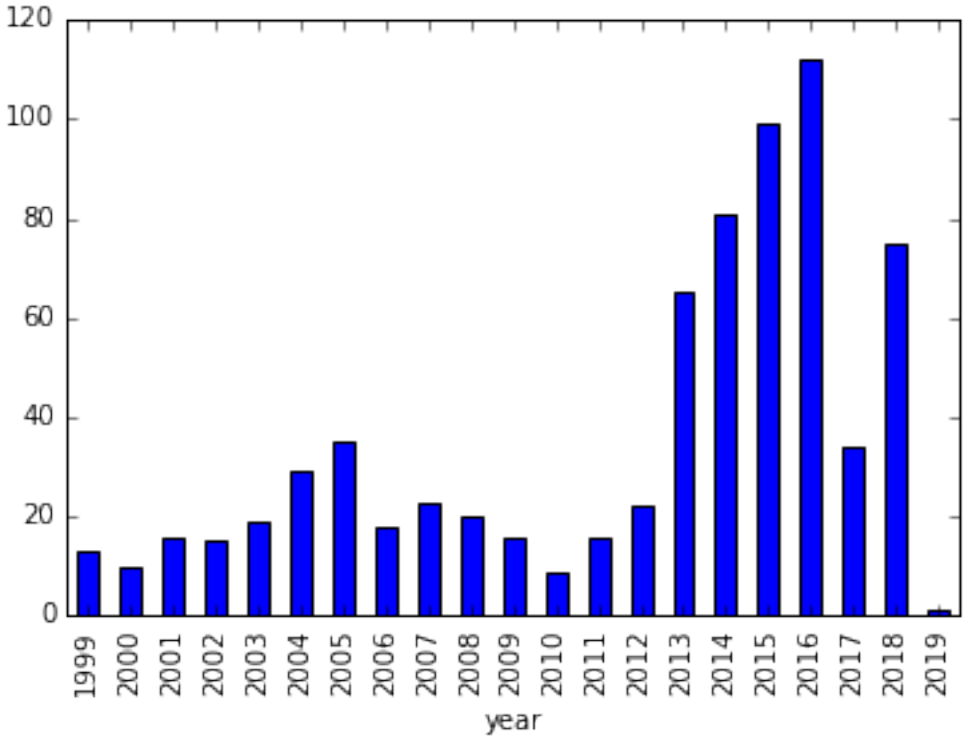


In [14]:
```
df['CVSS Score'].hist()
plt.title('CVSS Score')
```

Out[14]: <matplotlib.text.Text at 0x1257bddd8>

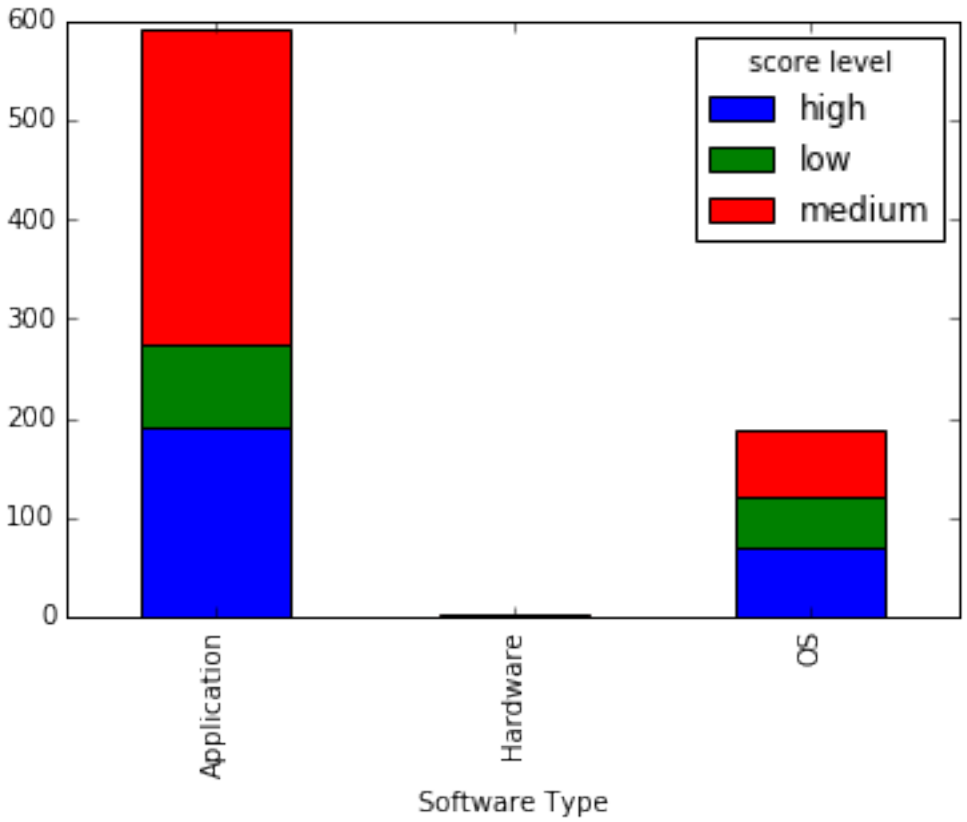



```
In [15]: df['Vendor'].nunique()

Out[15]: 335

In [16]: df.groupby('year')['Vendor'].nunique().plot(kind='bar')

Out[16]: <matplotlib.axes._subplots.AxesSubplot at 0x1257bd940>
```



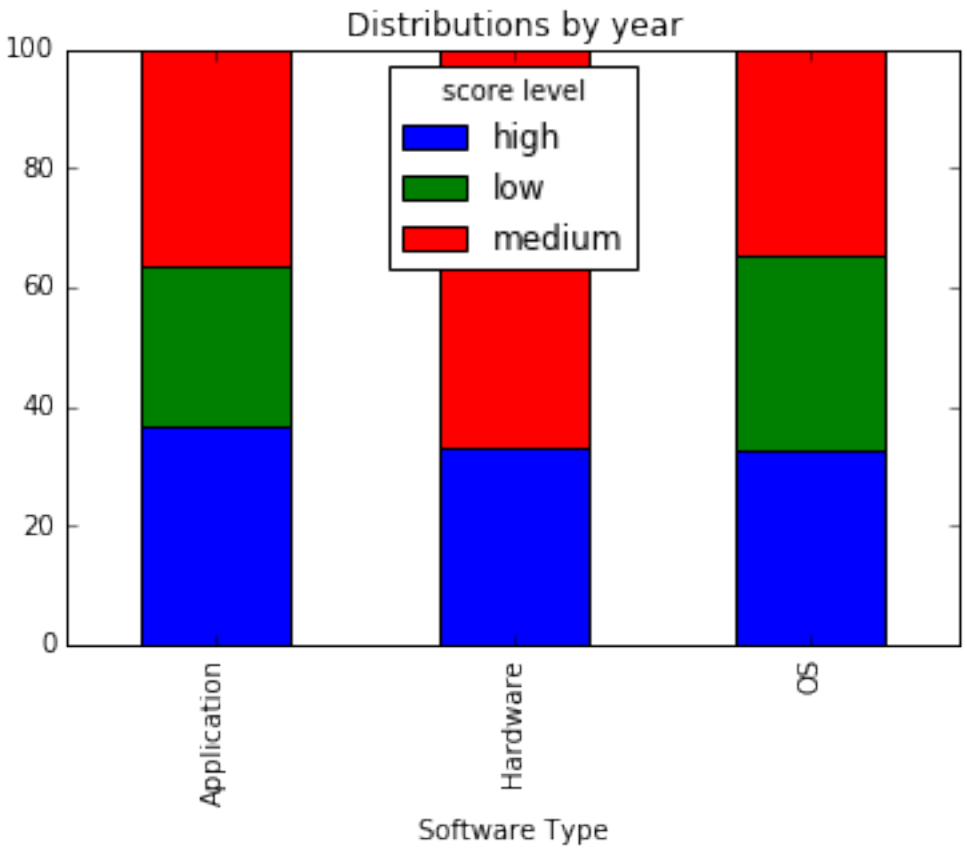

```
In [17]: df.groupby(['Software Type', 'score level'])['Product'].nunique().unstack().plot(kind='bar',sta
```

```
Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0x124cfec88>
```

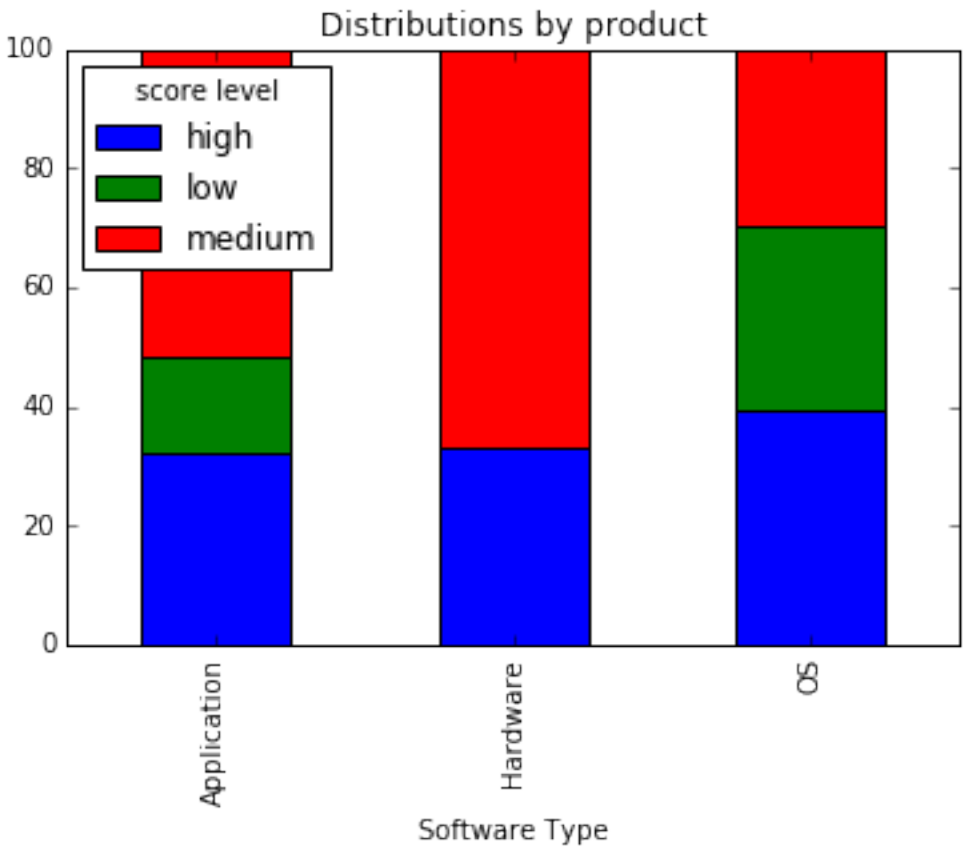



```
In [18]: #the distributions, not raw amounts
         df.groupby(['Software Type','score level'])['year'].nunique().groupby(level=0).apply(
             lambda x: 100 * x / x.sum()
         ).unstack().plot(kind='bar',stacked=True)
         plt.title('Distributions by year')
```

```
Out[18]: <matplotlib.text.Text at 0x124851d30>
```




```
In [19]: #the distributions, not raw amounts
         df.groupby(['Software Type','score level'])['Vendor'].nunique().groupby(level=0).apply(
             lambda x: 100 * x / x.sum()
         ).unstack().plot(kind='bar',stacked=True)
         plt.title('Distributions by product')

Out[19]: <matplotlib.text.Text at 0x12519cdd8>
```


Distributions by product
score level
high
low
medium
100
80
60
40
20
0
Application
Hardware
OS
Software Type