assignment 6 (own)

March 8, 2022

1 Univariate analysis on 2 columns of my own dataset

[1]:
```
import pandas as pd
```

[2]:
```
import seaborn as sns
```

Steam data set is a dataset with data from the digital games store "Steam".

[3]:
```
df = pd.read_csv('steam.csv', sep=',')
```

[4]:
```
df.pop('appid')
# Convert english too boolean
df['english'] = df['english'].astype('bool')
# set release date to datetime
df['release_date'] = pd.to_datetime(df['release_date'])
# create 3 seperate platform fields instead of 1
df['windows'], df['mac'], df['linux'] = df['platforms'].apply(lambda x:␣
↪'windows' in x),df['platforms'].apply(lambda x: 'mac' in x),df['platforms'].
↪apply(lambda x: 'linux' in x)
# Split owners categorical value in two numerical values
df['owners_low'] = df['owners'].apply(lambda x: x.split('-')[0]).astype('int')
df['owners_high'] = df['owners'].apply(lambda x: x.split('-')[1]).astype('int')
# Create int out of data column
df['release_year'] = df['release_date'].dt.year
genres = df['genres'].apply(lambda x: x.split(';')[0])
```

[5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27075 entries, 0 to 27074
Data columns (total 23 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   name              27075 non-null  object
 1   release_date      27075 non-null  datetime64[ns]
 2   english           27075 non-null  bool
 3   developer         27075 non-null  object
 4   publisher         27075 non-null  object
```

```
 5   platforms         27075 non-null  object
 6   required_age      27075 non-null  int64
 7   categories        27075 non-null  object
 8   genres            27075 non-null  object
 9   steamspy_tags     27075 non-null  object
 10  achievements      27075 non-null  int64
 11  positive_ratings  27075 non-null  int64
 12  negative_ratings  27075 non-null  int64
 13  average_playtime  27075 non-null  int64
 14  median_playtime   27075 non-null  int64
 15  owners            27075 non-null  object
 16  price             27075 non-null  float64
 17  windows           27075 non-null  bool
 18  mac               27075 non-null  bool
 19  linux             27075 non-null  bool
 20  owners_low        27075 non-null  int32
 21  owners_high       27075 non-null  int32
 22  release_year      27075 non-null  int64
dtypes: bool(4), datetime64[ns](1), float64(1), int32(2), int64(7), object(8)
memory usage: 3.8+ MB
```

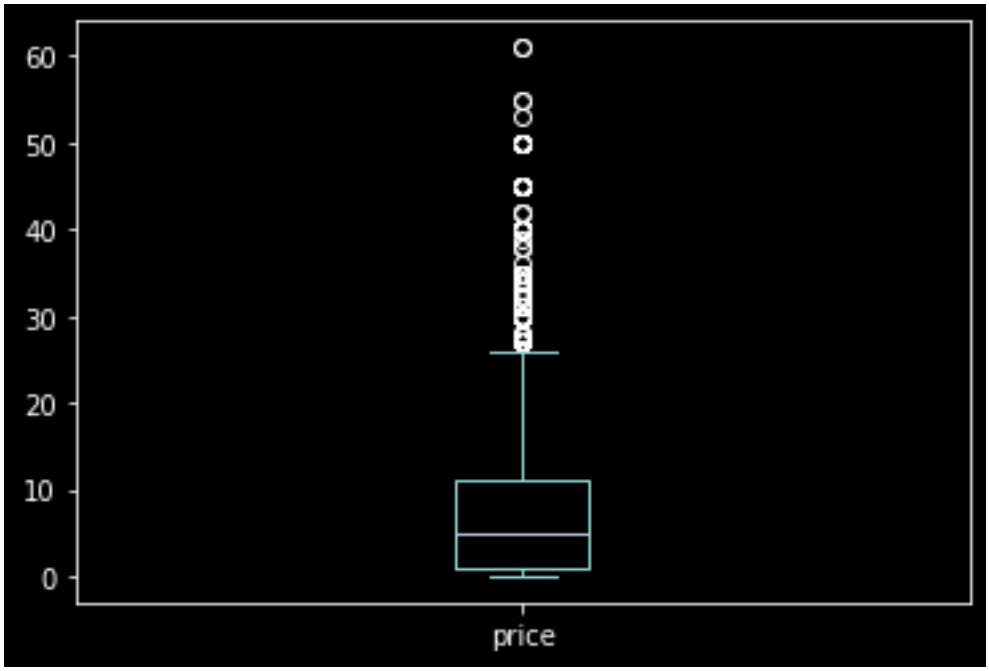
The median and owners filter are to make sure that games that nobody playes are excluded to get a more accurate representation of games

[6]:
```
medianPlaytimeFilter = df['median_playtime']> 0.5
ownersFilter = df['owners_low'] > 20000 #lowest range above 0
noFreeGameFilter = df['price'] > 0.1
reviewFilter = df['positive_ratings'] > 5
```

Boxplot doesn't work, as a lot of games are free

[7]:
```
df['price'][ownersFilter][df['positive_ratings'] > 1][medianPlaytimeFilter].
↪plot(kind='box')
```

[7]:
```
<AxesSubplot:>
```

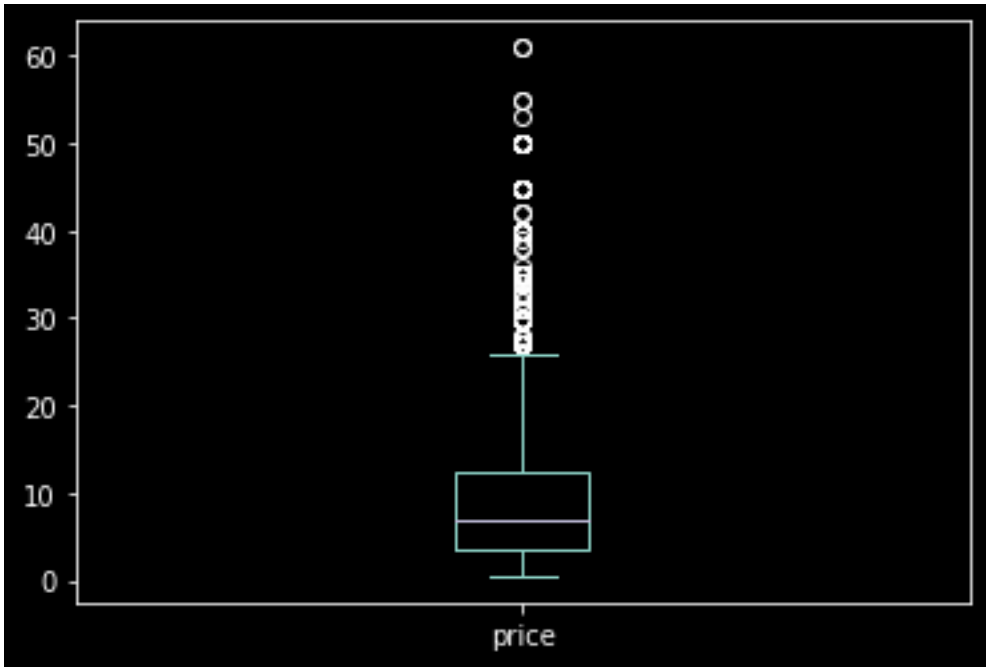


Filter out the free games

[8]:
```
df[noFreeGameFilter][ownersFilter][reviewFilter][medianPlaytimeFilter]['price'].
↪plot(kind='box')
```

```
C:\Users\Stijn\AppData\Local\Temp/ipykernel_7872/1341901083.py:1: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  df[noFreeGameFilter][ownersFilter][reviewFilter][medianPlaytimeFilter]['price'
].plot(kind='box')
```

[8]:
```
<AxesSubplot:>
```



2 The most expensive games

[26]:
```
df[medianPlaytimeFilter][['name','price']].sort_values('price',ascending=False).
↪head(10)
```

[26]:

	price	name
13061	114.99	GameMaker Studio 2 Web
5384	60.99	RPG Maker MV
3950	60.99	AppGameKit: Easy Game Development
2435	54.99	X-Plane 11
21371	54.99	WARRIORS OROCHI 4 - OROCHI
21571	54.99	DEAD OR ALIVE 6
1498	52.99	RPG Maker VX Ace
10029	49.99	ACE COMBAT 7: SKIES UNKNOWN
10016	49.99	BERSERK and the Band of the Hawk
6785	49.99	ARSLAN: THE WARRIORS OF LEGEND

3 Prices of games

(filtered out outliers and the games that nearly nobody owns/playes)

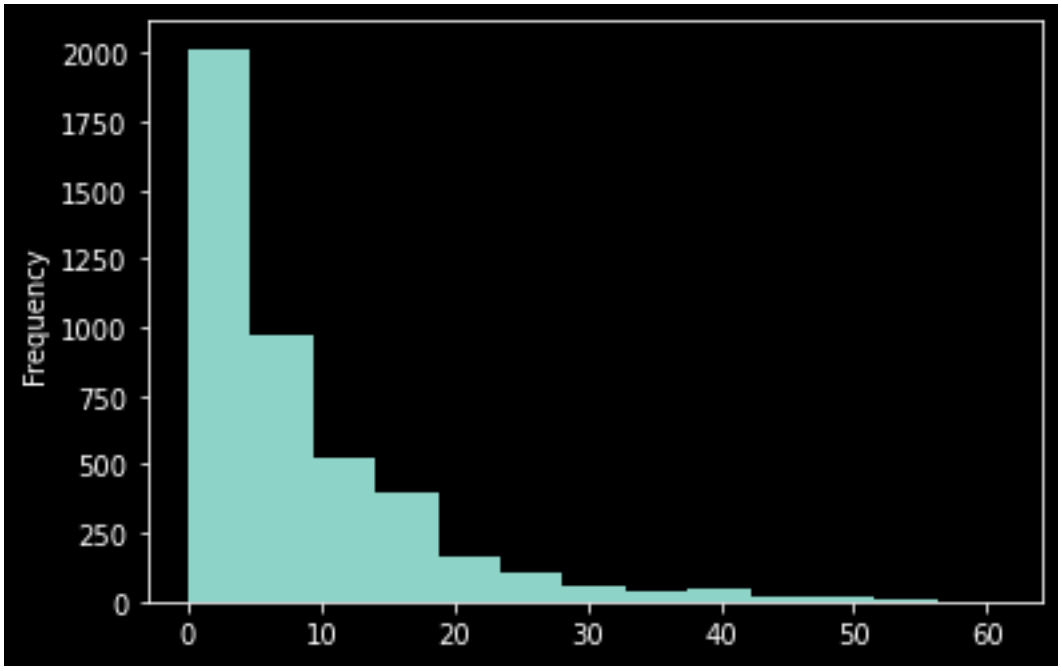
It seems there are many free games

[10]:
```

```

```
df[df['price'] <␣
↪200][ownersFilter][medianPlaytimeFilter][reviewFilter]['price'].
↪plot(kind='hist',bins=13)
```

```
C:\Users\Stijn\AppData\Local\Temp/ipykernel_7872/3573847891.py:1: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  df[df['price'] < 200][ownersFilter][medianPlaytimeFilter][reviewFilter]['price
'].plot(kind='hist',bins=13)
```

[10]: <AxesSubplot:ylabel='Frequency'>

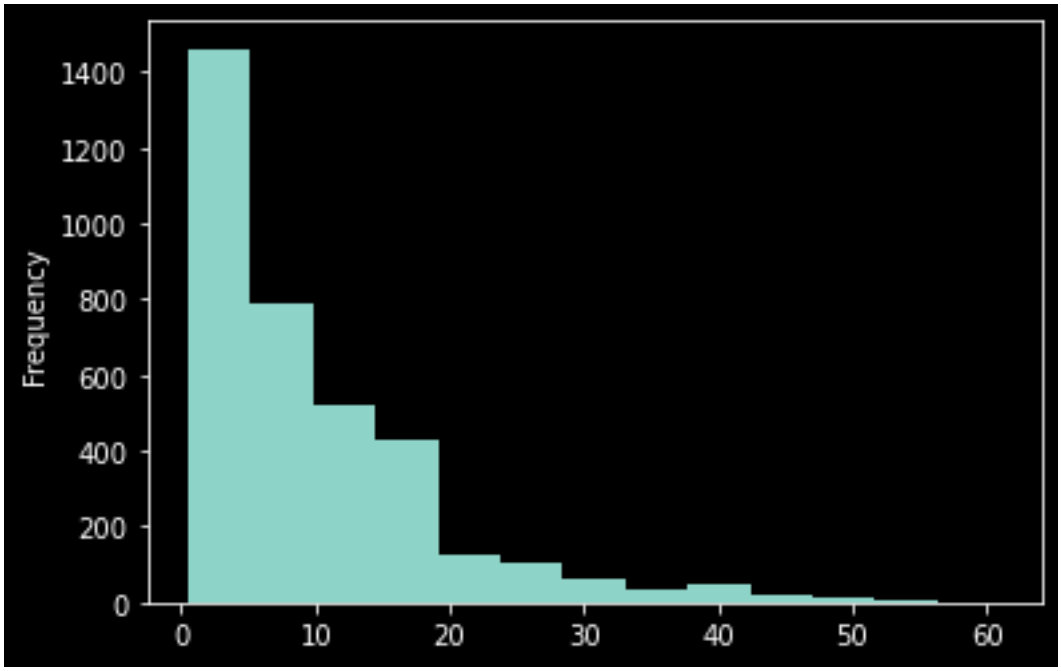


4 Price of payed games

[11]:
```
df[df['price']> 0][df['price'] <␣
↪200][ownersFilter][medianPlaytimeFilter]['price'].plot(kind='hist',bins=13)
```

```
C:\Users\Stijn\AppData\Local\Temp/ipykernel_7872/791919376.py:1: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  df[df['price']> 0][df['price'] <
200][ownersFilter][medianPlaytimeFilter]['price'].plot(kind='hist',bins=13)
```

[11]: <AxesSubplot:ylabel='Frequency'>



Seems there are also a lot of cheap games

5 Number of paid and free games

[12]:
```
df[noFreeGameFilter][df['price'] <␣
↪200][ownersFilter][medianPlaytimeFilter]['price'].count() , df[df['price'] <␣
↪200][ownersFilter][medianPlaytimeFilter]['price'].count()
```

```
C:\Users\Stijn\AppData\Local\Temp/ipykernel_7872/1283793119.py:1: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  df[noFreeGameFilter][df['price'] <
200][ownersFilter][medianPlaytimeFilter]['price'].count() , df[df['price'] <
200][ownersFilter][medianPlaytimeFilter]['price'].count()
```

[12]:
```
(3616, 4351)
```

6 Games that are played the most hours (median)

[13]:
```
df[ownersFilter][medianPlaytimeFilter][reviewFilter].
↪sort_values('median_playtime',ascending=False).head(10)
```

```
C:\Users\Stijn\AppData\Local\Temp/ipykernel_7872/2057969152.py:1: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  df[ownersFilter][medianPlaytimeFilter][reviewFilter].sort_values('median_playt
ime',ascending=False).head(10)
```

[13]:

	name	release_date	english	\
9201	The Abbey of Crime Extensum	2016-05-19	True	
1478	The Banner Saga: Factions	2013-02-25	True	
6014	The Secret of Tremendous Corporation	2015-10-12	True	
8969	PRICE	2016-09-15	True	
3969	Shroud of the Avatar: Forsaken Virtues	2018-03-27	True	
2435	X-Plane 11	2017-03-30	True	
12195	The Price of Freedom	2016-12-22	True	
8796	MANDAGON	2016-08-03	True	
2737	Heroine's Quest: The Herald of Ragnarok	2014-03-20	True	
3152	The Desolate Hope	2014-05-05	True	

	developer	\
9201	Manuel Pazos;Daniel Celemín	
1478	Stoic	
6014	Sebastian Krzyszkowiak;Konrad Burandt;Paweł Radej	
8969	YETU GAME	
3969	Portalarium	
2435	Laminar Research	
12195	Construct Studio	
8796	Blind Sky Studios	
2737	Crystal Shard	
3152	Scott Cawthon	

	publisher	platforms	required_age	\
9201	Manuel Pazos;Daniel Celemín	windows;mac;linux	0	
1478	Versus Evil	windows;mac	0	
6014	dosowisko.net	windows;linux	0	
8969	YETU GAME	windows	0	
3969	Portalarium	windows;mac;linux	0	
2435	Laminar Research	windows;mac;linux	0	
12195	Construct Studio Inc.	windows	0	
8796	Blind Sky Studios	windows;mac	0	
2737	Crystal Shard	windows;linux	0	
3152	Scott Cawthon	windows	0	

	categories	\
9201	Single-player	
1478	Multi-player;Cross-Platform Multiplayer	
6014	Single-player;Captions available;Steam Cloud	
8969	Single-player;Steam Achievements;Steam Trading…	
3969	Single-player;Multi-player;MMO;Co-op;Cross-Pla…	
2435	Single-player;Local Multi-Player;Partial Contr…	
12195	Single-player	
8796	Single-player;Steam Achievements;Partial Contr…	
2737	Single-player;Steam Achievements;Steam Trading…	
3152	Single-player	

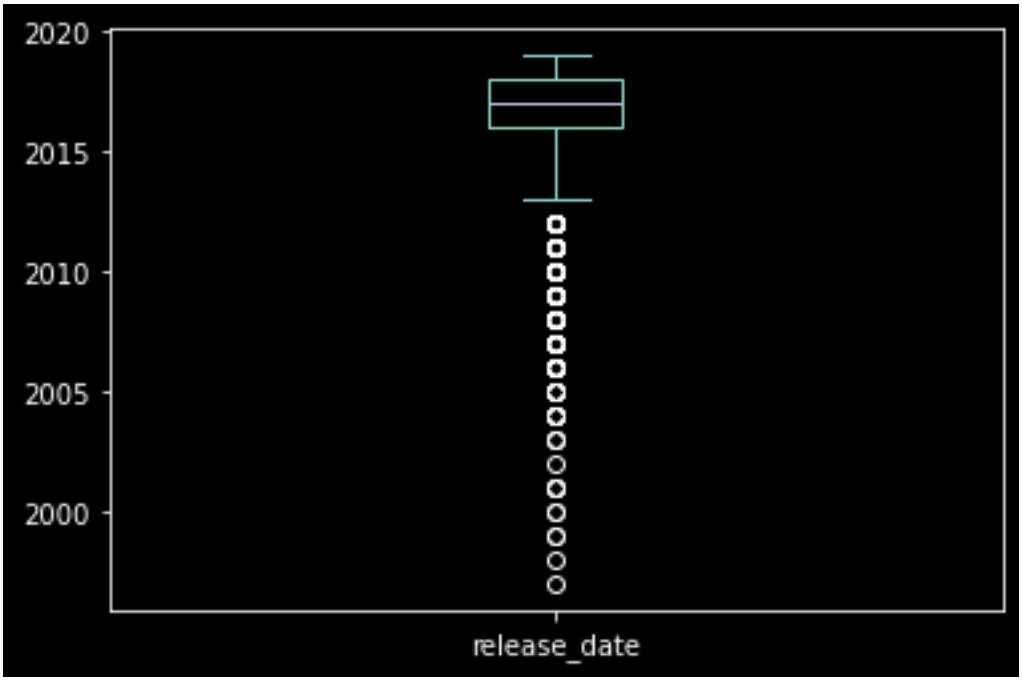
```
                                      genres  \
9201                  Adventure;Free to Play
1478          Free to Play;Indie;RPG;Strategy
6014      Adventure;Casual;Free to Play;Indie
8969                  Adventure;Casual;Indie
3969   Free to Play;Massively Multiplayer;RPG
2435                              Simulation
12195                         Adventure;Indie
8796             Adventure;Free to Play;Indie
2737         Adventure;Free to Play;Indie;RPG
3152               Action;Adventure;Indie;RPG

                               steamspy_tags  …  average_playtime  \
9201             Free to Play;Adventure;Retro  …             190625
1478                Free to Play;Strategy;RPG  …              95245
6014             Free to Play;Adventure;Indie  …              95242
8969                Puzzle;Free to Play;Anime  …              63481
3969   RPG;Massively Multiplayer;Free to Play  …              54618
2435              Simulation;Flight;Realistic  …              44169
12195              Indie;Adventure;Story Rich  …              36029
8796    Free to Play;Pixel Graphics;Adventure  …              21233
2737              Adventure;RPG;Point & Click  …              21247
3152               RPG;Free to Play;Adventure  …              21168

       median_playtime          owners  price  windows    mac  linux  \
9201            190625    50000-100000   0.00     True   True   True
1478            190489   200000-500000   0.00     True   True  False
6014            190445   100000-200000   0.00     True  False   True
8969             63490   200000-500000   0.00     True  False  False
3969             54618    50000-100000   0.00     True   True   True
2435             44169   100000-200000  54.99     True   True   True
12195            36029    50000-100000   0.00     True  False  False
8796             31845   200000-500000   0.00     True   True  False
2737             31835  500000-1000000   0.00     True  False   True
3152             31751   200000-500000   0.00     True  False  False

       owners_low  owners_high  release_year
9201        50000       100000          2016
1478       200000       500000          2013
6014       100000       200000          2015
8969       200000       500000          2016
3969        50000       100000          2018
2435       100000       200000          2017
12195       50000       100000          2016
8796       200000       500000          2016
2737       500000      1000000          2014
```

```
3152       200000       500000          2014

[10 rows x 23 columns]
```

7 Release year of a game

[14]:
```
df['release_date'].dt.year.plot(kind='box')
```

[14]:
```
<AxesSubplot:>
```



Boxplot has many outliers, but why?

The lower values are due to Valves own games and the few games that were allowed on early that were quite rare

[15]:
```
df.sort_values(by='release_year').head(5)
```

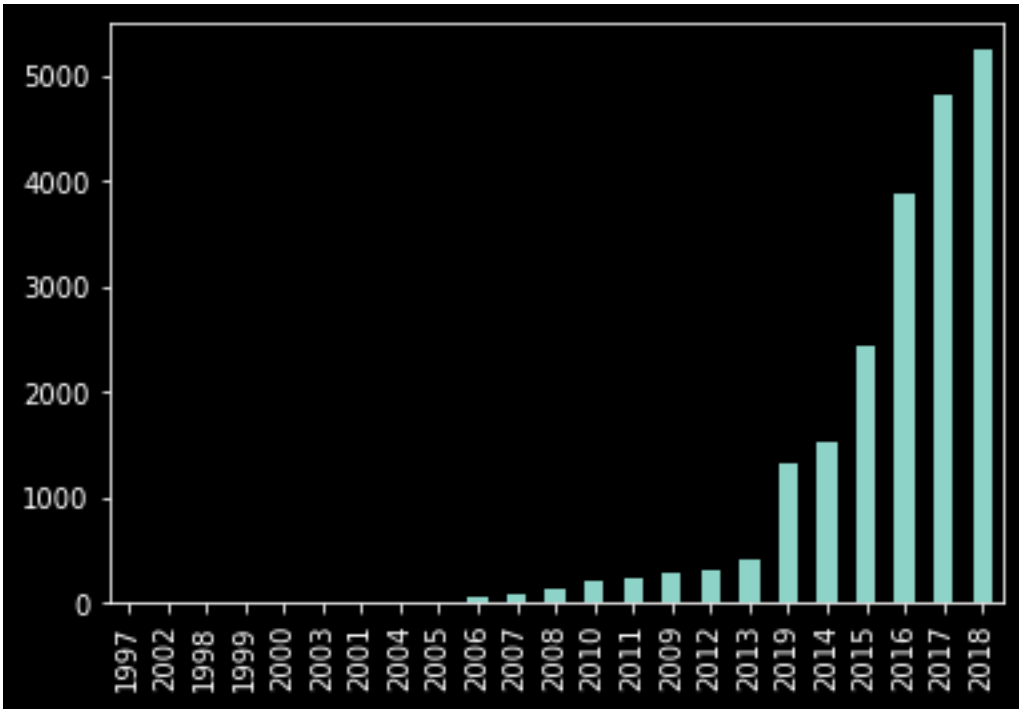
[15]:
```
                          name release_date  english            developer  \
2685       Carmageddon Max Pack   1997-06-30     True  Stainless Games Ltd
6                     Half-Life   1998-11-08     True                Valve
1         Team Fortress Classic   1999-04-01     True                Valve
4     Half-Life: Opposing Force   1999-11-01     True     Gearbox Software
5                      Ricochet   2000-11-01     True                Valve

        publisher          platforms  required_age  \
```

```
2685  THQ Nordic            windows             0
6          Valve  windows;mac;linux             0
1          Valve  windows;mac;linux             0
4          Valve  windows;mac;linux             0
5          Valve  windows;mac;linux             0

                                          categories                genres  \
2685    Single-player;Multi-player;Steam Trading Cards  Action;Indie;Racing
6     Single-player;Multi-player;Online Multi-Player…                Action
1     Multi-player;Online Multi-Player;Local Multi-P…                Action
4     Single-player;Multi-player;Valve Anti-Cheat en…                Action
5     Multi-player;Online Multi-Player;Valve Anti-Ch…                Action

               steamspy_tags  …  average_playtime  median_playtime  \
2685   Racing;Action;Classic  …                13               13
6         FPS;Classic;Action  …              1300               83
1     Action;FPS;Multiplayer  …               277               62
4          FPS;Action;Sci-fi  …               624              415
5     Action;FPS;Multiplayer  …               175               10

                owners  price  windows    mac  linux  owners_low  owners_high  \
2685     50000-100000   5.99     True  False  False       50000       100000
6     5000000-10000000   7.19     True   True   True     5000000     10000000
1     5000000-10000000   3.99     True   True   True     5000000     10000000
4     5000000-10000000   3.99     True   True   True     5000000     10000000
5     5000000-10000000   3.99     True   True   True     5000000     10000000

      release_year
2685          1997
6             1998
1             1999
4             1999
5             2000

[5 rows x 23 columns]
```

8 Amount of games per year

[16]:
```
df['release_year'][df['positive_ratings'] > 5].value_counts().sort_values().
↪plot(kind='bar')
```

[16]:
```
<AxesSubplot:>
```

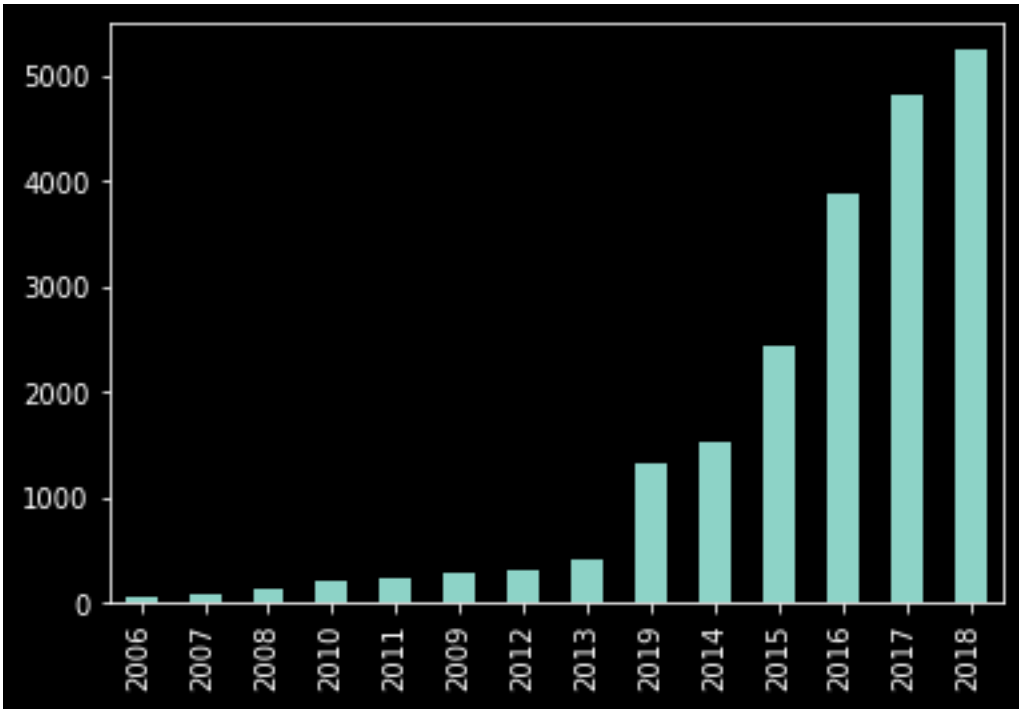


9 Amount of games per year after 2005

[17]:
```
df[df['release_year'] > 2005][df['positive_ratings'] > 5]['release_year'].
↪value_counts().sort_values().plot(kind='bar')
```

```
C:\Users\Stijn\AppData\Local\Temp/ipykernel_7872/4157551761.py:1: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  df[df['release_year'] > 2005][df['positive_ratings'] >
5]['release_year'].value_counts().sort_values().plot(kind='bar')
```

[17]:
```
<AxesSubplot:>
```

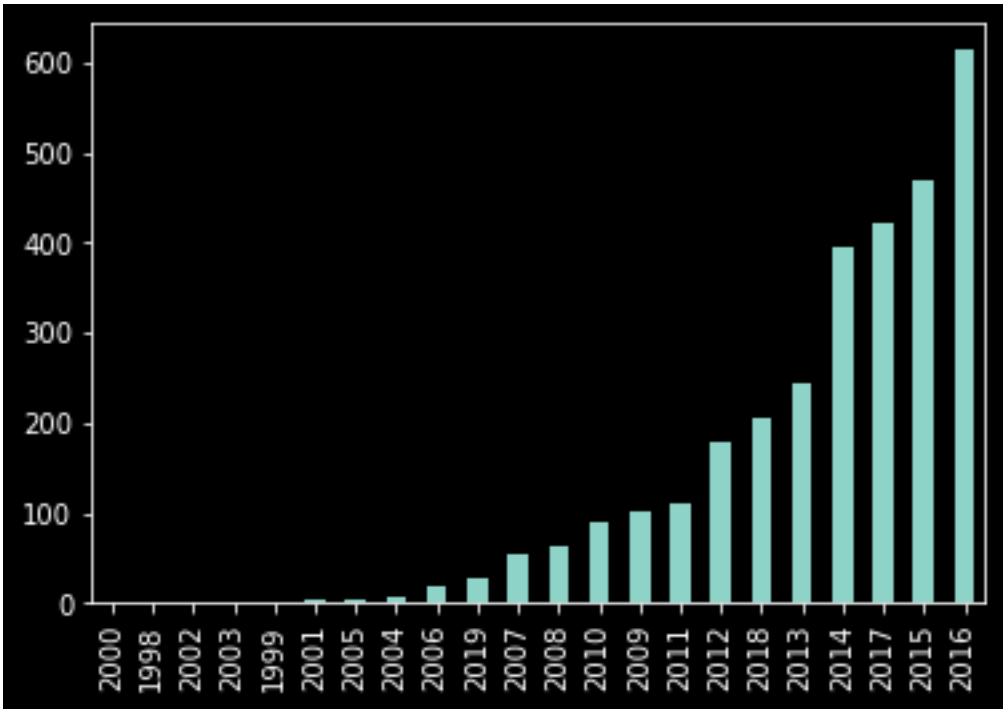


10 Amount of games that were played for longer than 30 hours

[18]:
```
df[df['owners_low'] > 50000][df['median_playtime']> 30]['release_year'].
↪value_counts().sort_values().plot(kind='bar')
```

```
C:\Users\Stijn\AppData\Local\Temp/ipykernel_7872/3507090343.py:1: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  df[df['owners_low'] > 50000][df['median_playtime']>
30]['release_year'].value_counts().sort_values().plot(kind='bar')
```

[18]:
```
<AxesSubplot:>
```



[19]:
```
df[df['positive_ratings'] > 5]['developer'].value_counts().head(10)
```

[19]:
```
Choice of Games               69
KOEI TECMO GAMES CO., LTD.    61
Ripknot Systems               42
RewindApp                     38
Humongous Entertainment       36
Nikita "Ghost_RUS"            34
For Kids                      32
Hosted Games                  31
EnsenaSoft                    31
MumboJumbo                    29
Name: developer, dtype: int64
```

[20]:
```
df.sort_values('positive_ratings',ascending = False).head(10)
```

[20]:

	name	release_date	english	\
25	Counter-Strike: Global Offensive	2012-08-21	True	
22	Dota 2	2013-07-09	True	
19	Team Fortress 2	2007-10-10	True	
12836	PLAYERUNKNOWN'S BATTLEGROUNDS	2017-12-21	True	
121	Garry's Mod	2006-11-29	True	
2478	Grand Theft Auto V	2015-04-13	True	
1467	PAYDAY 2	2013-08-13	True	

```
3362                          Unturned   2017-07-07     True
1120                          Terraria   2011-05-16     True
21                       Left 4 Dead 2   2009-11-19     True

                              developer                 publisher  \
25     Valve;Hidden Path Entertainment                     Valve
22                                Valve                     Valve
19                                Valve                     Valve
12836                 PUBG Corporation          PUBG Corporation
121                  Facepunch Studios                     Valve
2478                    Rockstar North            Rockstar Games
1467   OVERKILL - a Starbreeze Studio.  Starbreeze Publishing AB
3362             Smartly Dressed Games     Smartly Dressed Games
1120                           Re-Logic                  Re-Logic
21                                Valve                     Valve

              platforms  required_age  \
25     windows;mac;linux             0
22     windows;mac;linux             0
19     windows;mac;linux             0
12836            windows             0
121    windows;mac;linux             0
2478             windows            18
1467         windows;linux            18
3362   windows;mac;linux             0
1120   windows;mac;linux             0
21     windows;mac;linux             0

                                            categories  \
25     Multi-player;Steam Achievements;Full controlle…
22     Multi-player;Co-op;Steam Trading Cards;Steam W…
19     Multi-player;Cross-Platform Multiplayer;Steam …
12836            Multi-player;Online Multi-Player;Stats
121    Single-player;Multi-player;Co-op;Cross-Platfor…
2478   Single-player;Multi-player;Steam Achievements;…
1467   Single-player;Multi-player;Co-op;Online Co-op;…
3362   Single-player;Online Multi-Player;Online Co-op…
1120   Single-player;Multi-player;Online Multi-Player…
21     Single-player;Multi-player;Co-op;Steam Achieve…

                                         genres  \
25                          Action;Free to Play
22                 Action;Free to Play;Strategy
19                          Action;Free to Play
12836     Action;Adventure;Massively Multiplayer
121                            Indie;Simulation
2478                           Action;Adventure
```

```
1467                                  Action;RPG
3362   Action;Adventure;Casual;Free to Play;Indie
1120                  Action;Adventure;Indie;RPG
21                                        Action

                      steamspy_tags  …  average_playtime  median_playtime  \
25           FPS;Multiplayer;Shooter  …              22494             6502
22         Free to Play;MOBA;Strategy  …              23944              801
19       Free to Play;Multiplayer;FPS  …               8495              623
12836    Survival;Shooter;Multiplayer  …              22938            12434
121         Sandbox;Multiplayer;Funny  …              12422             1875
2478   Open World;Action;Multiplayer  …               9837             4834
1467                Co-op;Action;FPS  …               3975              890
3362   Free to Play;Survival;Zombies  …               3248              413
1120      Sandbox;Adventure;Survival  …               5585             1840
21                 Zombies;Co-op;FPS  …               1615              566

                    owners  price  windows    mac  linux  owners_low  \
25      50000000-100000000   0.00     True   True   True    50000000
22     100000000-200000000   0.00     True   True   True   100000000
19       20000000-50000000   0.00     True   True   True    20000000
12836   50000000-100000000  26.99     True  False  False    50000000
121      10000000-20000000   6.99     True   True   True    10000000
2478     10000000-20000000  24.99     True  False  False    10000000
1467     10000000-20000000   7.49     True  False   True    10000000
3362     20000000-50000000   0.00     True   True   True    20000000
1120      5000000-10000000   6.99     True   True   True     5000000
21       10000000-20000000   7.19     True   True   True    10000000

       owners_high  release_year
25       100000000          2012
22       200000000          2013
19        50000000          2007
12836    100000000          2017
121       20000000          2006
2478      20000000          2015
1467      20000000          2013
3362      50000000          2017
1120      10000000          2011
21        20000000          2009

[10 rows x 23 columns]
```

[21]:
```
series = df[df['positive_ratings'] > 5][medianPlaytimeFilter]['developer'].
↪value_counts()
series.where(lambda x :  0 < x).dropna()
```

```
C:\Users\Stijn\AppData\Local\Temp/ipykernel_7872/1183673222.py:1: UserWarning:
```

```
Boolean Series key will be reindexed to match DataFrame index.
  series = df[df['positive_ratings'] >
5][medianPlaytimeFilter]['developer'].value_counts()
```

[21]:
```
Valve                                    26
EnsenaSoft                               26
Winged Cloud                             17
Square Enix                              17
Daedalic Entertainment                   17
                                         ..
IMGN.PRO                                  1
Autumn Moon                               1
Lord Kres                                 1
Acido Cinza                               1
Beijing Litchi Culture Media Co., Ltd.    1
Name: developer, Length: 4077, dtype: int64
```

[22]:
```
df[df['positive_ratings'] > 5][medianPlaytimeFilter]['required_age'].
↪value_counts()
```

```
C:\Users\Stijn\AppData\Local\Temp/ipykernel_7872/70254092.py:1: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  df[df['positive_ratings'] >
5][medianPlaytimeFilter]['required_age'].value_counts()
```

[22]:
```
0     5796
18     186
16     116
12      24
7        4
3        3
Name: required_age, dtype: int64
```