

Mini_Project_IMDb

January 30, 2019

Mini Project

IMDb Data Set <https://grouplens.org/datasets/movielens/> ml-20m.zip

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

```
In [2]: # Contents of Data Set Zip
```

```
!ls ./movielens
```

Icon?	genome-scores.csv	links.csv	ratings.csv
README.txt	genome-tags.csv	movies.csv	tags.csv

```
In [3]: #movies = pd.read_csv('./movielens/movies.csv', sep=',')
#print(type(movies))
#movies.head(5)
```

```
In [4]: movies = csv.reader(open('./movielens/movies.csv', 'r'))
```

```
years = []
year_1995 = []

for line in movies:
    line = ''.join(line)
    #print(line)

    if '(1900)' in line:
        year_1995.append(line)
'''
for line in movies:
    if '1995' in line:
        year_1995.append(line)
'''
```

```
Out[4]: "\nfor line in movies:\n    if '1995' in line:\n        year_1995.append(line)\n"
```

```
In [5]: print(year_1995)
```

```
['117909The Kiss (1900)Romance']
```

```
In [6]: print(len(year_1995))
```

```
1
```

```
In [14]: movies = csv.reader(open('./movielens/movies.csv','r'))
```

```
years = []
```

```
for line in movies:
    line = ''.join(line)
    #print(line)

    for i in range(1800, 2020):
        year_parentheses = '(' + str(i) + ')'
        year_list_name = 'years_' + str(i)

        if year_parentheses in line:
            years.append(line)
            #print(i)
```

```
print(len(years))
#print(years)
```

```
27256
```

```
In [8]: # Discrepancy in 27,256 row being read versus 27,279 in the original is probably due to
```

```
with open('./movielens/movies.csv','r') as f:
    reader = csv.reader(f,delimiter = ",")
    data = list(reader)
    row_count = len(data)
    print(row_count)
```

```
27279
```

```
In [9]: # Need to figure out a way to read each line into appropriate dictionary key value
```

```
movies = csv.reader(open('./movielens/movies.csv','r'))
```

```
years = []
```

```
for line in movies:
    line = ''.join(line) # converts lines to String
    #print(line)

    for i in range(1800, 2020):
        year_parentheses = '(' + str(i) + ')'
        year_list_name = 'years_' + str(i)

        if year_parentheses in line:
            years.append(line)
            #print(i)
```

```
x = range(1800,2020)
dct = {}
for i in x:
    dct['year_%s' % i] = []

print(dct)
```

```
{'year_1800': [], 'year_1801': [], 'year_1802': [], 'year_1803': [], 'year_1804': [], 'year_1805': [], 'year_1806': [], 'year_1807': [], 'year_1808': [], 'year_1809': [], 'year_1810': [], 'year_1811': [], 'year_1812': [], 'year_1813': [], 'year_1814': [], 'year_1815': [], 'year_1816': [], 'year_1817': [], 'year_1818': [], 'year_1819': [], 'year_1820': [], 'year_1821': [], 'year_1822': [], 'year_1823': [], 'year_1824': [], 'year_1825': [], 'year_1826': [], 'year_1827': [], 'year_1828': [], 'year_1829': [], 'year_1830': [], 'year_1831': [], 'year_1832': [], 'year_1833': [], 'year_1834': [], 'year_1835': [], 'year_1836': [], 'year_1837': [], 'year_1838': [], 'year_1839': [], 'year_1840': [], 'year_1841': [], 'year_1842': [], 'year_1843': [], 'year_1844': [], 'year_1845': [], 'year_1846': [], 'year_1847': [], 'year_1848': [], 'year_1849': [], 'year_1850': [], 'year_1851': [], 'year_1852': [], 'year_1853': [], 'year_1854': [], 'year_1855': [], 'year_1856': [], 'year_1857': [], 'year_1858': [], 'year_1859': [], 'year_1860': [], 'year_1861': [], 'year_1862': [], 'year_1863': [], 'year_1864': [], 'year_1865': [], 'year_1866': [], 'year_1867': [], 'year_1868': [], 'year_1869': [], 'year_1870': [], 'year_1871': [], 'year_1872': [], 'year_1873': [], 'year_1874': [], 'year_1875': [], 'year_1876': [], 'year_1877': [], 'year_1878': [], 'year_1879': [], 'year_1880': [], 'year_1881': [], 'year_1882': [], 'year_1883': [], 'year_1884': [], 'year_1885': [], 'year_1886': [], 'year_1887': [], 'year_1888': [], 'year_1889': [], 'year_1890': [], 'year_1891': [], 'year_1892': [], 'year_1893': [], 'year_1894': [], 'year_1895': [], 'year_1896': [], 'year_1897': [], 'year_1898': [], 'year_1899': [], 'year_1900': [], 'year_1901': [], 'year_1902': [], 'year_1903': [], 'year_1904': [], 'year_1905': [], 'year_1906': [], 'year_1907': [], 'year_1908': [], 'year_1909': [], 'year_1910': [], 'year_1911': [], 'year_1912': [], 'year_1913': [], 'year_1914': [], 'year_1915': [], 'year_1916': [], 'year_1917': [], 'year_1918': [], 'year_1919': [], 'year_1920': [], 'year_1921': [], 'year_1922': [], 'year_1923': [], 'year_1924': [], 'year_1925': [], 'year_1926': [], 'year_1927': [], 'year_1928': [], 'year_1929': [], 'year_1930': [], 'year_1931': [], 'year_1932': [], 'year_1933': [], 'year_1934': [], 'year_1935': [], 'year_1936': [], 'year_1937': [], 'year_1938': [], 'year_1939': [], 'year_1940': [], 'year_1941': [], 'year_1942': [], 'year_1943': [], 'year_1944': [], 'year_1945': [], 'year_1946': [], 'year_1947': [], 'year_1948': [], 'year_1949': [], 'year_1950': [], 'year_1951': [], 'year_1952': [], 'year_1953': [], 'year_1954': [], 'year_1955': [], 'year_1956': [], 'year_1957': [], 'year_1958': [], 'year_1959': [], 'year_1960': [], 'year_1961': [], 'year_1962': [], 'year_1963': [], 'year_1964': [], 'year_1965': [], 'year_1966': [], 'year_1967': [], 'year_1968': [], 'year_1969': [], 'year_1970': [], 'year_1971': [], 'year_1972': [], 'year_1973': [], 'year_1974': [], 'year_1975': [], 'year_1976': [], 'year_1977': [], 'year_1978': [], 'year_1979': [], 'year_1980': [], 'year_1981': [], 'year_1982': [], 'year_1983': [], 'year_1984': [], 'year_1985': [], 'year_1986': [], 'year_1987': [], 'year_1988': [], 'year_1989': [], 'year_1990': [], 'year_1991': [], 'year_1992': [], 'year_1993': [], 'year_1994': [], 'year_1995': [], 'year_1996': [], 'year_1997': [], 'year_1998': [], 'year_1999': [], 'year_2000': [], 'year_2001': [], 'year_2002': [], 'year_2003': [], 'year_2004': [], 'year_2005': [], 'year_2006': [], 'year_2007': [], 'year_2008': [], 'year_2009': [], 'year_2010': [], 'year_2011': [], 'year_2012': [], 'year_2013': [], 'year_2014': [], 'year_2015': [], 'year_2016': [], 'year_2017': [], 'year_2018': [], 'year_2019': [], 'year_2020': []}
```

```
In [10]: movies = csv.reader(open('./movielens/movies.csv','r'))
```

```
dct = {}

for line in movies:
    line = ''.join(line) # converts lines to String

    #i = 1800

    for i in range(1800, 2020):
        year_parentheses = '(' + str(i) + ')'

        if year_parentheses in line:
            dct['year_%s' % i] = [line]
            i = i + 1

print(len(dct))
```

118

As a dictionary

```
In [11]: movies = csv.reader(open('./movielens/movies.csv','r'))

years = {}

for line in movies:
    line = ''.join(line)

    year = line[line.find("(") + 1 : line.find(")")]

    if year.isdigit():
        #print(year + "/" + line)      # List all the years from each line
        years[year] = [line]

print(years)

{'1995': ['131128Flodder 3 (1995)Comedy'], '1994': ['131142Voll Normaaal (1994)Comedy'], '1996
```

```
In [13]: from collections import Counter
import matplotlib.pyplot as plt
import numpy as np
%matplotlib inline

movies = csv.reader(open('./movielens/movies.csv','r'))

years = []

for line in movies:
    line = ''.join(line)

    year = line[line.find("(") + 1 : line.find(")")]

    if year.isdigit():
        #print(year + "/" + line)      # List all the years from each line
        years.append(int(year))

print(len(years)) # 22,059 is the number of lines with years included
```

```

C = Counter(years)
year_list_of_lists = [[k,]*v for k,v in C.items()]
print(len(year_list_of_lists)) # 119 unique years
#print(year_list_of_lists)

unique_years = [item[0] for item in year_list_of_lists]
#print(len(unique_years))

sorted_years = sorted(unique_years)
#print(sorted_years)

x_axis = []
y_axis = []
graph_pairs = []
r = 0
while r < 119:
    year_num = year_list_of_lists[0 + r][0]
    entries_num = len(year_list_of_lists[0 + r])

    x_axis.append(year_num)
    y_axis.append(entries_num)

    plt.bar(x_axis, y_axis, label='Movie Entries by Year') # Create Bar Graph
    ax = plt.gca()
    ax.set_xlim([1890, 2020])

    ### Change the size of the plot and save it to folder as PNG
    fig = plt.gcf()
    fig.set_size_inches(18.5, 10.5)
    fig.savefig('test2019.png', dpi=100)

    combo = str(year_num) + "|" + str(entries_num)
    #print(combo)

    graph_pairs.append(year_num | entries_num)

    r = r + 1

plt.xlabel('Year')
plt.ylabel('Number of Films')

```

```
plt.title('Movie Entries by Year')

#print(graph)

graph_pairs.sort()
print(graph_pairs[0:-3])
```

22059
119
[7, 69, 501, 1891, 1893, 1894, 1895, 1898, 1901, 1901, 1902, 1903, 1905, 1910, 1911, 1913, 191

