In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import os
%matplotlib inline
```

In [2]:

```
# Exploring the folder
os.listdir()
```

Out[2]:

```
['.ipynb_checkpoints',
 'data_row',
 'Do more educated people go to cinema more.ipynb',
 'main_df.xls']
```

In [3]:

```
os.listdir("./data_row")
```

Out[3]:

```
['cine_audience.xls', 'city_region.xls', 'population.xlsx', 'tr_phd.xls']
```

In [4]:

```
# Reading the data
df = pd.read_excel("./data_row/city_region.xls")
df.head()
```

Out[4]:

	city	region
0	Türkiye / Turkey	Türkiye / Turkey
1	Adana	Akdeniz Bölgesi / The Mediterranean Region
2	Adıyaman	Güneydoğu Anadolu Bölgesi / The Southeastern A...
3	Afyonkarahisar	Ege Bölgesi / The Aegean Region
4	Ağrı	Doğu Anadolu Bölgesi / The Eastern Anatolia Re...

In [5]:

```
df.shape
```

Out[5]:

```
(82, 2)
```

In [6]:

```
# Expanding the df for adding data acording to years
df = pd.concat([df]*20, ignore_index=True)
df.head()
```

Out[6]:

	city	region
0	Türkiye / Turkey	Türkiye / Turkey
1	Adana	Akdeniz Bölgesi / The Mediterranean Region
2	Adıyaman	Güneydoğu Anadolu Bölgesi / The Southeastern A...
3	Afyonkarahisar	Ege Bölgesi / The Aegean Region
4	Ağrı	Doğu Anadolu Bölgesi / The Eastern Anatolia Re...

In [7]:

```
# Sorting
df = df.sort_values("city")
df.head()
```

Out[7]:

	city	region
821	Adana	Akdeniz Bölgesi / The Mediterranean Region
329	Adana	Akdeniz Bölgesi / The Mediterranean Region
1395	Adana	Akdeniz Bölgesi / The Mediterranean Region
903	Adana	Akdeniz Bölgesi / The Mediterranean Region
575	Adana	Akdeniz Bölgesi / The Mediterranean Region

In [8]:

```
# Resetting the index
df.reset_index(inplace=True)
df.head()
```

Out[8]:

	index	city	region
0	821	Adana	Akdeniz Bölgesi / The Mediterranean Region
1	329	Adana	Akdeniz Bölgesi / The Mediterranean Region
2	1395	Adana	Akdeniz Bölgesi / The Mediterranean Region
3	903	Adana	Akdeniz Bölgesi / The Mediterranean Region
4	575	Adana	Akdeniz Bölgesi / The Mediterranean Region

In [9]:

```
# Deleting old index column
df.drop("index", axis=1, inplace=True)
df.head()
```

Out[9]:

	city	region
0	Adana	Akdeniz Bölgesi / The Mediterranean Region
1	Adana	Akdeniz Bölgesi / The Mediterranean Region
2	Adana	Akdeniz Bölgesi / The Mediterranean Region
3	Adana	Akdeniz Bölgesi / The Mediterranean Region
4	Adana	Akdeniz Bölgesi / The Mediterranean Region

In [10]:

```
# Defining years series
years = pd.Series([i for i in range(2000, 2020)])
years
```

Out[10]:

```
0     2000
1     2001
2     2002
3     2003
4     2004
5     2005
6     2006
7     2007
8     2008
9     2009
10    2010
11    2011
12    2012
13    2013
14    2014
15    2015
16    2016
17    2017
18    2018
19    2019
dtype: int64
```

In [11]:

```
# Expanding years series for each city
years = pd.concat([years]*82, ignore_index=True)
years.head()
```

Out[11]:

```
0    2000
1    2001
2    2002
3    2003
4    2004
dtype: int64
```

In [12]:

```
# Adding year values
df["years"] = years
df.head()
```

Out[12]:

	city	region	years
0	Adana	Akdeniz Bölgesi / The Mediterranean Region	2000
1	Adana	Akdeniz Bölgesi / The Mediterranean Region	2001
2	Adana	Akdeniz Bölgesi / The Mediterranean Region	2002
3	Adana	Akdeniz Bölgesi / The Mediterranean Region	2003
4	Adana	Akdeniz Bölgesi / The Mediterranean Region	2004

In [13]:

```
# Bringing population data
population = pd.read_excel("./data_row/population.xlsx")
population
```

Out[13]:

	city	2000	2001	2002	2003	2004	2005	
0	Türkiye / Turkey	64729501	65603160.0	66401851.0	67187251.0	68010215.0	68860539.0	69729
1	Adana	1879695	1899324.0	1916637.0	1933428.0	1951142.0	1969512.0	1988
2	Adıyaman	568432	571180.0	573149.0	574886.0	576808.0	578852.0	580
3	Afyonkarahisar	696292	698029.0	698773.0	699193.0	699794.0	700502.0	701
4	Ağrı	519190	521514.0	523123.0	524514.0	526070.0	527732.0	529
5	Amasya	333927	333768.0	333110.0	332271.0	331491.0	330739.0	329
6	Ankara	3889199	3971642.0	4050309.0	4128889.0	4210596.0	4294678.0	4380
7	Antalya	1430539	1480282.0	1529110.0	1578367.0	1629338.0	1681656.0	1735
8	Artvin	167909	168184.0	168215.0	168164.0	168153.0	168164.0	168
9	Aydın	870460	881911.0	892345.0	902594.0	913340.0	924446.0	935
10	Balıkesir	1069260	1077362.0	1084072.0	1090411.0	1097187.0	1104261.0	1111
11	Bilecik	197625	198736.0	199580.0	200346.0	201182.0	202063.0	202
12	Bingöl	240337	242183.0	243717.0	245168.0	246718.0	248336.0	249
13	Bitlis	318886	320555.0	321791.0	322898.0	324114.0	325401.0	326
14	Bolu	255576	257926.0	259953.0	261902.0	263967.0	266114.0	268
15	Burdur	246060	247106.0	247811.0	248412.0	249090.0	249816.0	250
16	Bursa	2150571	2192169.0	2231582.0	2270852.0	2311735.0	2353834.0	2396
17	Çanakkale	449418	453632.0	457280.0	460792.0	464511.0	468375.0	472
18	Çankırı	169044	169955.0	170637.0	171252.0	171924.0	172635.0	173
19	Çorum	567609	566094.0	563698.0	560968.0	558300.0	555649.0	552
20	Denizli	845493	854958.0	863396.0	871614.0	880267.0	889229.0	898
21	Diyarbakır	1317750	1338378.0	1357550.0	1376518.0	1396333.0	1416775.0	1437
22	Edirne	392134	393292.0	393896.0	394320.0	394852.0	395449.0	396
23	Elazığ	517551	521467.0	524710.0	527774.0	531048.0	534467.0	537
24	Erzincan	206815	208015.0	208937.0	209779.0	210694.0	211658.0	212
25	Erzurum	801287	800311.0	798119.0	795482.0	792968.0	790505.0	787
26	Eskişehir	651672	662354.0	672328.0	682212.0	692529.0	703168.0	714
27	Gaziantep	1292817	1330205.0	1366581.0	1403165.0	1441079.0	1480026.0	1519
28	Giresun	410946	412428.0	413335.0	414062.0	414909.0	415830.0	416
29	Gümüşhane	116008	118147.0	120166.0	122175.0	124267.0	126423.0	128

	city	2000	2001	2002	2003	2004	2005	
...	...	...	...	...	...	...	...	
52	Ordu	705746	708079.0	709420.0	710444.0	711670.0	713018.0	714
53	Rize	307133	308800.0	310052.0	311181.0	312417.0	313722.0	315
54	Sakarya	750485	762848.0	774397.0	785845.0	797793.0	810112.0	822
55	Samsun	1191926	1198574.0	1203611.0	1208179.0	1213165.0	1218424.0	1223
56	Siirt	270832	273982.0	276806.0	279562.0	282461.0	285462.0	288
57	Sinop	194318	195151.0	195715.0	196196.0	196739.0	197319.0	197
58	Sivas	651825	650946.0	649078.0	646845.0	644709.0	642614.0	640
59	Tekirdağ	577812	598658.0	619152.0	639837.0	661237.0	683199.0	705
60	Tokat	641033	639371.0	636715.0	633682.0	630722.0	627781.0	624
61	Trabzon	720620	724340.0	727080.0	729529.0	732221.0	735072.0	737
62	Tunceli	82554	82871.0	83074.0	83241.0	83433.0	83640.0	83
63	Şanlıurfa	1257753	1294842.0	1330964.0	1367305.0	1404961.0	1443639.0	1483
64	Uşak	320535	322814.0	324673.0	326417.0	328287.0	330243.0	332
65	Van	895836	908296.0	919727.0	930984.0	942771.0	954945.0	967
66	Yozgat	544446	538313.0	531220.0	523696.0	516096.0	508398.0	500
67	Zonguldak	630323	629346.0	627407.0	625114.0	622912.0	620744.0	618
68	Aksaray	351474	353939.0	355942.0	357819.0	359834.0	361941.0	364
69	Bayburt	75221	75517.0	75709.0	75868.0	76050.0	76246.0	76
70	Karaman	214461	216318.0	217902.0	219417.0	221026.0	222700.0	224
71	Kırıkkale	287427	286900.0	285933.0	284803.0	283711.0	282633.0	281
72	Batman	408820	418186.0	427172.0	436165.0	445508.0	455118.0	464
73	Şırnak	362700	370314.0	377574.0	384824.0	392364.0	400123.0	408
74	Bartın	175982	177060.0	177903.0	178678.0	179519.0	180401.0	181
75	Ardahan	122409	121305.0	119993.0	118590.0	117178.0	115750.0	114
76	Iğdır	174285	175550.0	176588.0	177563.0	178609.0	179701.0	180
77	Yalova	144923	150027.0	155041.0	160099.0	165333.0	170705.0	176
78	Karabük	205172	207241.0	209056.0	210812.0	212667.0	214591.0	216
79	Kilis	109698	111024.0	112219.0	113387.0	114615.0	115886.0	117
80	Osmaniye	411163	417418.0	423214.0	428943.0	434930.0	441108.0	447
81	Düzce	296712	300686.0	304316.0	307884.0	311623.0	315487.0	319

82 rows × 20 columns

In [14]:

```
# Adding year 2019, it needs to match with df
population[2019]=0
```

In [15]:

```
population = population.sort_values("city")
population.head()
```

Out[15]:

	city	2000	2001	2002	2003	2004	2005	2006	
1	Adana	1879695	1899324.0	1916637.0	1933428.0	1951142.0	1969512.0	1988277.0	2
2	Adıyaman	568432	571180.0	573149.0	574886.0	576808.0	578852.0	580926.0	
3	Afyonkarahisar	696292	698029.0	698773.0	699193.0	699794.0	700502.0	701204.0	
68	Aksaray	351474	353939.0	355942.0	357819.0	359834.0	361941.0	364089.0	
5	Amasya	333927	333768.0	333110.0	332271.0	331491.0	330739.0	329956.0	

5 rows × 21 columns

In [16]:

```
# Resetting the index
population.reset_index(inplace=True)
population.head()
```

Out[16]:

	index	city	2000	2001	2002	2003	2004	2005	2
0	1	Adana	1879695	1899324.0	1916637.0	1933428.0	1951142.0	1969512.0	198827
1	2	Adıyaman	568432	571180.0	573149.0	574886.0	576808.0	578852.0	58092
2	3	Afyonkarahisar	696292	698029.0	698773.0	699193.0	699794.0	700502.0	7012
3	68	Aksaray	351474	353939.0	355942.0	357819.0	359834.0	361941.0	36408
4	5	Amasya	333927	333768.0	333110.0	332271.0	331491.0	330739.0	32995

5 rows × 22 columns

In [17]:

```
# Deleting old index column
population.drop("index", axis=1, inplace=True)
population.head()
```

Out[17]:

	city	2000	2001	2002	2003	2004	2005	2006	
0	Adana	1879695	1899324.0	1916637.0	1933428.0	1951142.0	1969512.0	1988277.0	20
1	Adıyaman	568432	571180.0	573149.0	574886.0	576808.0	578852.0	580926.0	5
2	Afyonkarahisar	696292	698029.0	698773.0	699193.0	699794.0	700502.0	701204.0	7
3	Aksaray	351474	353939.0	355942.0	357819.0	359834.0	361941.0	364089.0	3
4	Amasya	333927	333768.0	333110.0	332271.0	331491.0	330739.0	329956.0	3

5 rows × 21 columns

In [18]:

```
population.head()
```

Out[18]:

	city	2000	2001	2002	2003	2004	2005	2006	
0	Adana	1879695	1899324.0	1916637.0	1933428.0	1951142.0	1969512.0	1988277.0	20
1	Adıyaman	568432	571180.0	573149.0	574886.0	576808.0	578852.0	580926.0	5
2	Afyonkarahisar	696292	698029.0	698773.0	699193.0	699794.0	700502.0	701204.0	7
3	Aksaray	351474	353939.0	355942.0	357819.0	359834.0	361941.0	364089.0	3
4	Amasya	333927	333768.0	333110.0	332271.0	331491.0	330739.0	329956.0	3

5 rows × 21 columns

In [19]:

```
# Adding population column to the df
df["pop"] = 0
```

In [20]:

```
# Adding populations to the dataframe
m = 0
for a in range(len(population.index)):
    for b in range(len(population.columns)-1): #I need -1 becouse of city column
        df.iloc[m, 3] = population.iloc[a, b+1]
        m += 1
```

In [21]:

```
df.head()
```

Out[21]:

	city	region	years	pop
0	Adana	Akdeniz Bölgesi / The Mediterranean Region	2000	1879695.0
1	Adana	Akdeniz Bölgesi / The Mediterranean Region	2001	1899324.0
2	Adana	Akdeniz Bölgesi / The Mediterranean Region	2002	1916637.0
3	Adana	Akdeniz Bölgesi / The Mediterranean Region	2003	1933428.0
4	Adana	Akdeniz Bölgesi / The Mediterranean Region	2004	1951142.0

In [22]:

```
#Cheacking the results
df.groupby("city").pop.mean().values == population.mean(axis=1).values # Looks like the pro
```

Out[22]:

```
array([ True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True])
```

In [23]:

```
audiences = pd.read_excel("./data_row/cine_audience.xls")
audiences
```

Out[23]:

	year	Adana- 1	Adıyaman- 2	Afyonkarahisar- 3	Aksaray- 68	Amasya- 5	Ankara- 6	Antalya- 7	Ardahar 7
0	2000	157500	10200	38661	16800	26747	2944678	336753	
1	2001	289500	0	44597	2500	18082	3378571	536289	
2	2002	215000	3200	17445	13100	26451	2714496	548476	
3	2003	579673	0	38000	54898	0	2545293	271676	
4	2004	851200	0	52325	68406	17680	2890955	735858	
5	2005	749490	12620	58500	107924	21450	2308746	926263	
6	2006	890328	8600	60200	80578	43000	3647310	954949	
7	2007	341113	12500	54000	33900	17038	2845002	944761	
8	2008	880246	17300	131225	56000	38180	3094337	1042985	
9	2009	617750	2500	113431	46575	42000	3484478	962324	
10	2010	953800	14235	125671	53640	49740	4428579	1356812	380
11	2011	847310	17387	159470	23000	39500	4424255	975199	180
12	2012	1077880	131900	129382	34919	27382	3802358	1327489	
13	2013	1247917	14944	116078	64776	22933	4929341	1535748	
14	2014	1459378	15854	241278	78082	59010	5582224	1711212	
15	2015	1567552	17732	239045	104598	36016	5988088	2098651	
16	2016	1330989	52361	300257	88086	75701	5463044	1840716	
17	2017	1501452	36273	392657	92000	108640	6940217	2424173	

18 rows × 82 columns

In [24]:

```
# Matching new data set with our main data set.
# Creating Turkey column.
audiences["Türkiye / Turkey"] = audiences.sum(axis=1) - audiences["year"]
audiences.head()
```

Out[24]:

	year	Adana-1	Adıyaman-2	Afyonkarahisar-3	Aksaray-68	Amasya-5	Ankara-6	Antalya-7	Ardahan-75
0	2000	157500	10200	38661	16800	26747	2944678	336753	0
1	2001	289500	0	44597	2500	18082	3378571	536289	0
2	2002	215000	3200	17445	13100	26451	2714496	548476	0
3	2003	579673	0	38000	54898	0	2545293	271676	0
4	2004	851200	0	52325	68406	17680	2890955	735858	0

5 rows × 83 columns

In [25]:

```
#Adding two new rows for years 2018 and 2019
audiences = audiences.append(audiences.iloc[0:2], ignore_index=True)
audiences.tail()
```

Out[25]:

	year	Adana-1	Adıyaman-2	Afyonkarahisar-3	Aksaray-68	Amasya-5	Ankara-6	Antalya-7	Ardahar 7
15	2015	1567552	17732	239045	104598	36016	5988088	2098651	
16	2016	1330989	52361	300257	88086	75701	5463044	1840716	
17	2017	1501452	36273	392657	92000	108640	6940217	2424173	
18	2000	157500	10200	38661	16800	26747	2944678	336753	
19	2001	289500	0	44597	2500	18082	3378571	536289	

5 rows × 83 columns

In [26]:

```
# Setting new rows to zero since we have no data
audiences.iloc[18:20] = 0
audiences.tail()
```

Out[26]:

	year	Adana- 1	Adıyaman- 2	Afyonkarahisar- 3	Aksaray- 68	Amasya- 5	Ankara- 6	Antalya- 7	Ardahar 7
15	2015	1567552	17732	239045	104598	36016	5988088	2098651	
16	2016	1330989	52361	300257	88086	75701	5463044	1840716	
17	2017	1501452	36273	392657	92000	108640	6940217	2424173	
18	0	0	0	0	0	0	0	0	
19	0	0	0	0	0	0	0	0	

5 rows × 83 columns

In [27]:

```
# No need for year column anymore
audiences.drop("year", axis=1, inplace=True)
```

In [28]:

```
# We need to sort by columns to match it with main df
audiences.sort_index(axis=1, inplace=True)
```

In [29]:

```
# creating a series from audiences dataframes columns.
cine_audiences=pd.Series()
cine_audiences = cine_audiences.append([audiences.iloc[:, i] for i in range(len(audiences.c
cine_audiences.head()
```

Out[29]:

```
0    157500
1    289500
2    215000
3    579673
4    851200
dtype: int64
```

In [30]:

```
# Assigning it to main df
df["cinema_audiences"] = cine_audiences
```

In [31]:

```
# Changing data type to int
df = df.astype(int, errors="ignore")
```

In [32]:

```
df.dtypes
```

Out[32]:

```
city                  object
region                object
years                  int32
pop                    int32
cinema_audiences       int32
dtype: object
```

In [33]:

```
# Checking the values
df.groupby("city").cinema_audiences.mean().values == audiences.mean().values
```

Out[33]:

```
array([ True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True])
```

In [34]:

```
# Reading phd data
phd = pd.read_excel("./data_row/tr_phd.xls")
phd
```

Out[34]:

	years	Adana- 1	Adıyaman- 2	Afyonkarahisar- 3	Aksaray- 68	Amasya- 5	Ankara- 6	Antalya- 7	Ardahan 7
0	2008	1505	103	412	153	103	16239	1188	2
1	2009	2044	151	588	226	136	19446	1780	2
2	2010	2589	286	687	314	207	20926	2360	5
3	2011	2785	353	737	338	230	21333	2661	5
4	2012	2762	364	714	331	229	21038	2783	5
5	2013	2941	459	908	448	302	28853	3434	7
6	2014	3101	491	955	478	311	29329	3680	9
7	2015	3183	516	1016	526	336	30486	3929	11
8	2016	3398	536	1027	540	346	30744	4033	12
9	2017	3926	753	1294	703	531	33979	4857	18
10	2018	3992	769	1291	716	571	33831	5112	19

11 rows × 82 columns

In [35]:

```
phd.drop("years", axis=1, inplace=True)
phd.head()
```

Out[35]:

	Adana- 1	Adıyaman- 2	Afyonkarahisar- 3	Aksaray- 68	Amasya- 5	Ankara- 6	Antalya- 7	Ardahan- 75	Artvin
0	1505	103	412	153	103	16239	1188	21	4
1	2044	151	588	226	136	19446	1780	22	7
2	2589	286	687	314	207	20926	2360	52	11
3	2785	353	737	338	230	21333	2661	57	12
4	2762	364	714	331	229	21038	2783	53	12

5 rows × 81 columns

In [36]:

```python
# Adding turkey column
phd["Türkiye / Turkey"] = phd.sum(axis=1)
phd.head()
```

Out[36]:

	Adana-1	Adıyaman-2	Afyonkarahisar-3	Aksaray-68	Amasya-5	Ankara-6	Antalya-7	Ardahan-75	Artvin
0	1505	103	412	153	103	16239	1188	21	4
1	2044	151	588	226	136	19446	1780	22	7
2	2589	286	687	314	207	20926	2360	52	11
3	2785	353	737	338	230	21333	2661	57	12
4	2762	364	714	331	229	21038	2783	53	12

5 rows × 82 columns

In [37]:

```python
phd.sort_index(axis=1, inplace=True)
phd.head()
```

Out[37]:

	Adana-1	Adıyaman-2	Afyonkarahisar-3	Aksaray-68	Amasya-5	Ankara-6	Antalya-7	Ardahan-75	Artvin
0	1505	103	412	153	103	16239	1188	21	4
1	2044	151	588	226	136	19446	1780	22	7
2	2589	286	687	314	207	20926	2360	52	11
3	2785	353	737	338	230	21333	2661	57	12
4	2762	364	714	331	229	21038	2783	53	12

5 rows × 82 columns

In [38]:

```python
#defining new serie from all columns of has_phd
has_phd =pd.Series()
has_phd = has_phd.append([phd.iloc[:, i] for i in range(len(phd.columns))], ignore_index=Tr
```

In [39]:

```
has_phd
```

Out[39]:

```
0       1505
1       2044
2       2589
3       2785
4       2762
5       2941
6       3101
7       3183
8       3398
9       3926
10      3992
11       103
12       151
13       286
14       353
15       364
16       459
17       491
18       516
19       536
20       753
21       769
22       412
23       588
24       687
25       737
26       714
27       908
28       955
29      1016
       ...  
872     9705
873     9739
874    11529
875    12089
876    12843
877    13164
878    14627
879    15082
880      344
881      628
882      801
883     1092
884     1053
885     1094
886     1103
887     1119
888     1102
889     1388
890     1347
891       59
892       89
893      105
894      141
```

```
895      121
896      152
897      161
898      159
899      141
900      202
901      210
Length: 902, dtype: int64
```

In [40]:

```python
# Adding new column
df["has_phd"] = 0
```

In [41]:

```
# Filtering the df in has_phd range
df[((2007 < df["years"]) & (df["years"] < 2019))]
```

Out[41]:

	city	region	years	pop	cinema_audiences	has_phd
8	Adana	Akdeniz Bölgesi / The Mediterranean Region	2008	2026319	880246	0
9	Adana	Akdeniz Bölgesi / The Mediterranean Region	2009	2062226	617750	0
10	Adana	Akdeniz Bölgesi / The Mediterranean Region	2010	2085225	953800	0
11	Adana	Akdeniz Bölgesi / The Mediterranean Region	2011	2108805	847310	0
12	Adana	Akdeniz Bölgesi / The Mediterranean Region	2012	2125635	1077880	0
13	Adana	Akdeniz Bölgesi / The Mediterranean Region	2013	2149260	1247917	0
14	Adana	Akdeniz Bölgesi / The Mediterranean Region	2014	2165595	1459378	0
15	Adana	Akdeniz Bölgesi / The Mediterranean Region	2015	2183167	1567552	0
16	Adana	Akdeniz Bölgesi / The Mediterranean Region	2016	2201670	1330989	0
17	Adana	Akdeniz Bölgesi / The Mediterranean Region	2017	2216475	1501452	0
18	Adana	Akdeniz Bölgesi / The Mediterranean Region	2018	2220125	0	0
28	Adıyaman	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2008	585067	17300	0
29	Adıyaman	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2009	588475	2500	0
30	Adıyaman	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2010	590935	14235	0
31	Adıyaman	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2011	593931	17387	0
32	Adıyaman	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2012	595261	131900	0
33	Adıyaman	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2013	597184	14944	0
34	Adıyaman	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2014	597835	15854	0
35	Adıyaman	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2015	602774	17732	0

	city	region	years	pop	cinema_audiences	has_phd
36	Adıyaman	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2016	610484	52361	0
37	Adıyaman	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2017	615076	36273	0
38	Adıyaman	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2018	624513	0	0
48	Afyonkarahisar	Ege Bölgesi / The Aegean Region	2008	697365	131225	0
49	Afyonkarahisar	Ege Bölgesi / The Aegean Region	2009	701326	113431	0
50	Afyonkarahisar	Ege Bölgesi / The Aegean Region	2010	697559	125671	0
51	Afyonkarahisar	Ege Bölgesi / The Aegean Region	2011	698626	159470	0
52	Afyonkarahisar	Ege Bölgesi / The Aegean Region	2012	703948	129382	0
53	Afyonkarahisar	Ege Bölgesi / The Aegean Region	2013	707123	116078	0
54	Afyonkarahisar	Ege Bölgesi / The Aegean Region	2014	706371	241278	0
55	Afyonkarahisar	Ege Bölgesi / The Aegean Region	2015	709015	239045	0
...	...	...	...	...	...	...
1591	İzmir	Ege Bölgesi / The Aegean Region	2011	3965232	2719564	0
1592	İzmir	Ege Bölgesi / The Aegean Region	2012	4005459	2677782	0
1593	İzmir	Ege Bölgesi / The Aegean Region	2013	4061074	3319804	0
1594	İzmir	Ege Bölgesi / The Aegean Region	2014	4113072	3662618	0
1595	İzmir	Ege Bölgesi / The Aegean Region	2015	4168415	3448945	0
1596	İzmir	Ege Bölgesi / The Aegean Region	2016	4223545	3621719	0
1597	İzmir	Ege Bölgesi / The Aegean Region	2017	4279677	4479193	0
1598	İzmir	Ege Bölgesi / The Aegean Region	2018	4320519	0	0
1608	Şanlıurfa	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2008	1574224	69453	0
1609	Şanlıurfa	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2009	1613737	155100	0
1610	Şanlıurfa	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2010	1663371	129180	0

	city	region	years	pop	cinema_audiences	has_phd
1611	Şanlıurfa	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2011	1716254	135396	0
1612	Şanlıurfa	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2012	1762075	106000	0
1613	Şanlıurfa	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2013	1801980	220208	0
1614	Şanlıurfa	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2014	1845667	325435	0
1615	Şanlıurfa	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2015	1892320	277382	0
1616	Şanlıurfa	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2016	1940627	268630	0
1617	Şanlıurfa	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2017	1985753	356683	0
1618	Şanlıurfa	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2018	2035809	0	0
1628	Şırnak	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2008	429287	0	0
1629	Şırnak	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2009	430424	0	0
1630	Şırnak	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2010	430109	0	0
1631	Şırnak	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2011	457997	0	0
1632	Şırnak	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2012	466982	0	0
1633	Şırnak	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2013	475255	0	0
1634	Şırnak	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2014	488966	0	0
1635	Şırnak	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2015	490184	0	0
1636	Şırnak	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2016	483788	0	0
1637	Şırnak	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2017	503236	0	0
1638	Şırnak	Güneydoğu Anadolu Bölgesi / The Southeastern A...	2018	524190	0	0

902 rows × 6 columns

```
In [42]:
```

```
index_1 = df[((2007 < df["years"]) & (df["years"] < 2019))].index
```

In [43]:

```
has_phd.index=index_1
```

In [44]:

```
has_phd
```

Out[44]:

```
8        1505
9        2044
10       2589
11       2785
12       2762
13       2941
14       3101
15       3183
16       3398
17       3926
18       3992
28        103
29        151
30        286
31        353
32        364
33        459
34        491
35        516
36        536
37        753
38        769
48        412
49        588
50        687
51        737
52        714
53        908
54        955
55       1016
        ...  
1591     9705
1592     9739
1593    11529
1594    12089
1595    12843
1596    13164
1597    14627
1598    15082
1608      344
1609      628
1610      801
1611     1092
1612     1053
1613     1094
1614     1103
1615     1119
1616     1102
1617     1388
1618     1347
1628       59
1629       89
1630      105
1631      141
```

```
1632      121
1633      152
1634      161
1635      159
1636      141
1637      202
1638      210
Length: 902, dtype: int64
```

In [45]:

```
# Asigning the values
df[(2007 < df["years"]) & (df["years"] < 2019)]["has_phd"] = has_phd
```

```
C:\Users\mkogu\Anaconda3\lib\site-packages\ipykernel_launcher.py:2: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/s
table/indexing.html#indexing-view-versus-copy (http://pandas.pydata.org/pand
as-docs/stable/indexing.html#indexing-view-versus-copy)
```

In [46]:

```
# Use loc to get rid of SettingWithCopyWarning. Python get confused if you try to update th
df.loc[((2007 < df["years"]) & (df["years"] < 2019)), "has_phd"] = has_phd
```

In [47]:

```
df.head(10)
```

Out[47]:

	city	region	years	pop	cinema_audiences	has_phd
0	Adana	Akdeniz Bölgesi / The Mediterranean Region	2000	1879695	157500	0
1	Adana	Akdeniz Bölgesi / The Mediterranean Region	2001	1899324	289500	0
2	Adana	Akdeniz Bölgesi / The Mediterranean Region	2002	1916637	215000	0
3	Adana	Akdeniz Bölgesi / The Mediterranean Region	2003	1933428	579673	0
4	Adana	Akdeniz Bölgesi / The Mediterranean Region	2004	1951142	851200	0
5	Adana	Akdeniz Bölgesi / The Mediterranean Region	2005	1969512	749490	0
6	Adana	Akdeniz Bölgesi / The Mediterranean Region	2006	1988277	890328	0
7	Adana	Akdeniz Bölgesi / The Mediterranean Region	2007	2006650	341113	0
8	Adana	Akdeniz Bölgesi / The Mediterranean Region	2008	2026319	880246	1505
9	Adana	Akdeniz Bölgesi / The Mediterranean Region	2009	2062226	617750	2044

In [48]:

```
# Saving the data.
df.to_excel("main_df.xls")
```

In [49]:

```
df.dtypes
```

Out[49]:

```
city                   object
region                 object
years                   int32
pop                     int32
cinema_audiences        int32
has_phd                 int64
dtype: object
```

In [50]:

```
# Creating new data frame to plot
df_1 = df[(2007 < df["years"]) & (df["years"] < 2018)].copy()
```

In [51]:

```
# Some zero values cousing some math problem(divided by zero, infinity problems). I will re
# it will have no impact on data and no couse infinity problems.
df_1.pop == 0
```

Out[51]:

```
False
```

In [52]:

```
df_1.cinema_audiences == 0
```

Out[52]:

```
8       False
9       False
10      False
11      False
12      False
13      False
14      False
15      False
16      False
17      False
28      False
29      False
30      False
31      False
32      False
33      False
34      False
35      False
36      False
37      False
48      False
49      False
50      False
51      False
52      False
53      False
54      False
55      False
56      False
57      False
        ...
1588    False
1589    False
1590    False
1591    False
1592    False
1593    False
1594    False
1595    False
1596    False
1597    False
1608    False
1609    False
1610    False
1611    False
1612    False
1613    False
1614    False
1615    False
1616    False
1617    False
1628     True
1629     True
1630     True
```

```
1631     True
1632     True
1633     True
1634     True
1635     True
1636     True
1637     True
Name: cinema_audiences, Length: 820, dtype: bool
```

In [53]:

```
df_1[df_1.cinema_audiences == 0]["cinema_audiences"] = 0
```

```
C:\Users\mkogu\Anaconda3\lib\site-packages\ipykernel_launcher.py:1: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/s
table/indexing.html#indexing-view-versus-copy (http://pandas.pydata.org/pand
as-docs/stable/indexing.html#indexing-view-versus-copy)
  """Entry point for launching an IPython kernel.
```

In [54]:

```
df_1.loc[(df_1.cinema_audiences == 0), "cinema_audiences" ] = 1
```

In [55]:

```
df_1[df_1.has_phd == 0] # No zero value
```

Out[55]:

	city	region	years	pop	cinema_audiences	has_phd

In [56]:

```
# I will exclude also the Turkey data. It ruins the chart due to high values.
df_1[df_1.city.str.contains("Türkiye")]
```

Out[56]:

	city	region	years	pop	cinema_audiences	has_phd
1388	Türkiye / Turkey	Türkiye / Turkey	2008	71517100	31132231	73244
1389	Türkiye / Turkey	Türkiye / Turkey	2009	72561312	31334447	95500
1390	Türkiye / Turkey	Türkiye / Turkey	2010	73722988	35787380	113862
1391	Türkiye / Turkey	Türkiye / Turkey	2011	74724269	37439786	121923
1392	Türkiye / Turkey	Türkiye / Turkey	2012	75627384	39002190	122619
1393	Türkiye / Turkey	Türkiye / Turkey	2013	76667864	45077509	154180
1394	Türkiye / Turkey	Türkiye / Turkey	2014	77695904	55378716	160410
1395	Türkiye / Turkey	Türkiye / Turkey	2015	78741053	57148011	168211
1396	Türkiye / Turkey	Türkiye / Turkey	2016	79814871	55260600	171486
1397	Türkiye / Turkey	Türkiye / Turkey	2017	80810525	68482526	203811

In [57]:

```
df_1.drop(df_1[df_1.city.str.contains("Türkiye")].index, inplace=True)
```

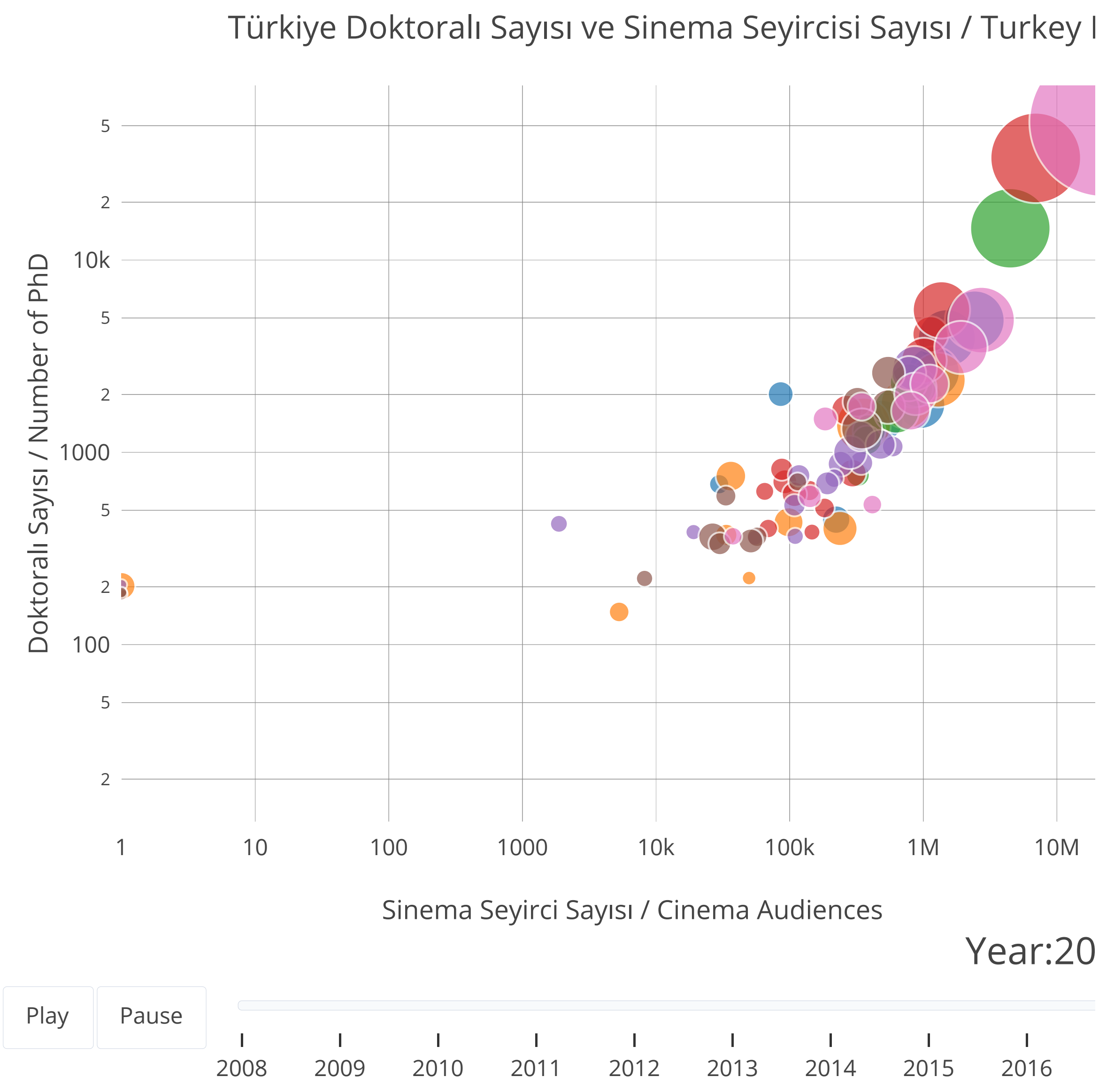
In [58]:

```
from __future__ import division
from plotly.offline import init_notebook_mode, iplot
init_notebook_mode()
from bubbly.bubbly import bubbleplot
```

In [59]:

```
figure = bubbleplot(dataset=df_1, x_column='cinema_audiences', y_column='has_phd',
    bubble_column='city', time_column='years', size_column='pop', color_column='region',
    x_title="Sinema Seyirci Sayısı / Cinema Audiences", y_title="Doktoralı Sayısı / Number
    title='Türkiye Doktoralı Sayısı ve Sinema Seyircisi Sayısı / Turkey Phd and Cinema Audi
    x_logscale=True, y_logscale=True, scale_bubble=1, width=1050, height=600)

iplot(figure)
```



In [60]:

```
# More populated cities are on the rigth top corner. Lets try to compare our values proport
# add also Turkey data.
# Creating new data frame to plot
df_2 = df[(2007 < df["years"]) & (df["years"] < 2018)].copy()
```

In [61]:

```
# Setting up to zero values to one
df_2.loc[(df_2.cinema_audiences == 0), "cinema_audiences" ] = 1
```

In [62]:

```
df_2.head()
```

Out[62]:

	city	region	years	pop	cinema_audiences	has_phd
8	Adana	Akdeniz Bölgesi / The Mediterranean Region	2008	2026319	880246	1505
9	Adana	Akdeniz Bölgesi / The Mediterranean Region	2009	2062226	617750	2044
10	Adana	Akdeniz Bölgesi / The Mediterranean Region	2010	2085225	953800	2589
11	Adana	Akdeniz Bölgesi / The Mediterranean Region	2011	2108805	847310	2785
12	Adana	Akdeniz Bölgesi / The Mediterranean Region	2012	2125635	1077880	2762

In [63]:

```

df_2["cine_aud_pop"] = df_2["cinema_audiences"] / df_2["pop"]
```

In [64]:

```
df_2["has_phd_pop"] = df_2["has_phd"] / df_2["pop"]
```

In [65]:

```
df_2.head()
```

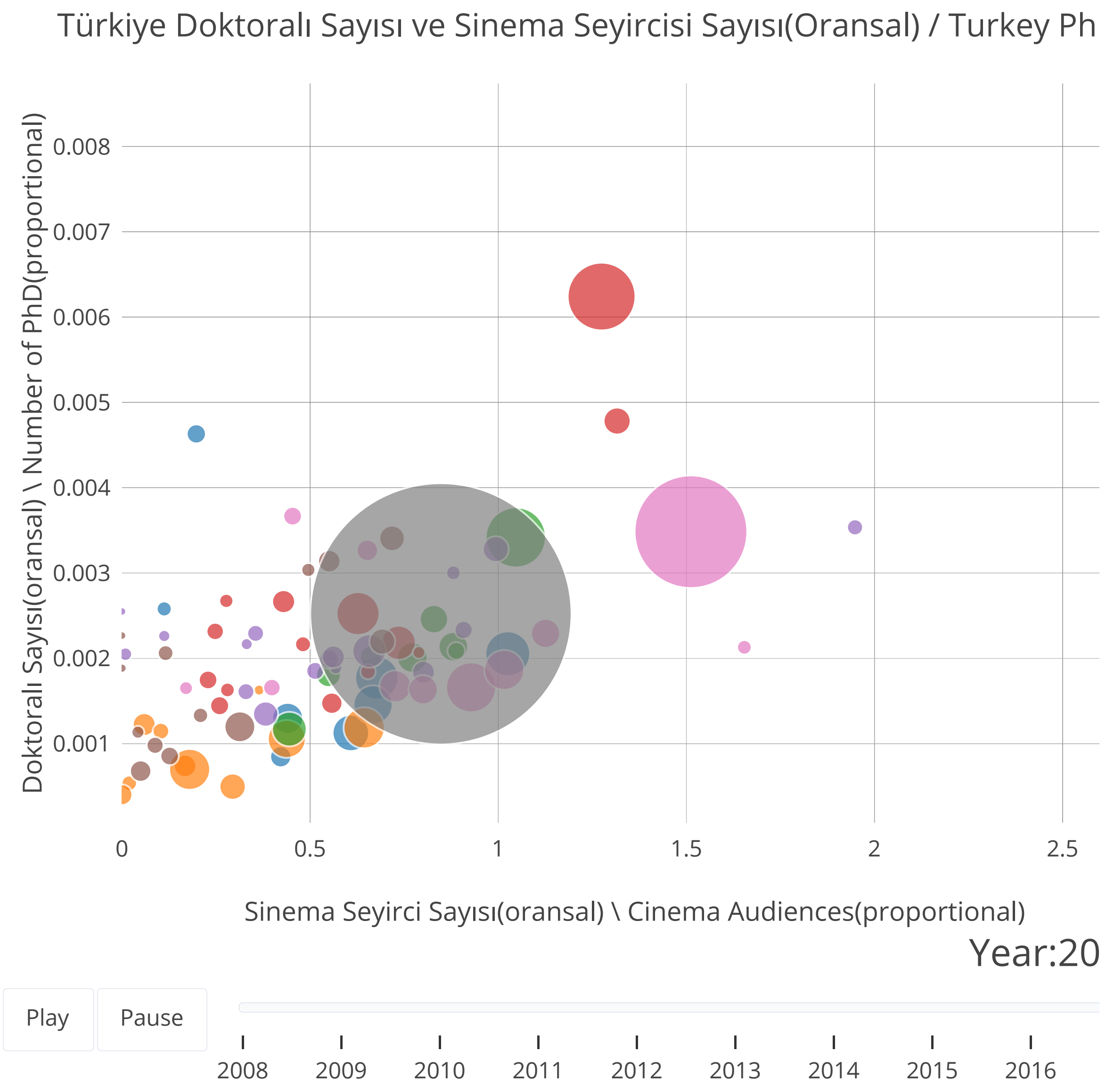
Out[65]:

	city	region	years	pop	cinema_audiences	has_phd	cine_aud_pop	has_phd_
8	Adana	Akdeniz Bölgesi / The Mediterranean Region	2008	2026319	880246	1505	0.434406	0.000
9	Adana	Akdeniz Bölgesi / The Mediterranean Region	2009	2062226	617750	2044	0.299555	0.000
10	Adana	Akdeniz Bölgesi / The Mediterranean Region	2010	2085225	953800	2589	0.457409	0.001
11	Adana	Akdeniz Bölgesi / The Mediterranean Region	2011	2108805	847310	2785	0.401796	0.001
12	Adana	Akdeniz Bölgesi / The Mediterranean Region	2012	2125635	1077880	2762	0.507086	0.001

In [66]:

```
figure_1 = bubbleplot(dataset=df_2, x_column="cine_aud_pop", y_column="has_phd_pop",
    bubble_column='city', time_column='years', size_column='pop', color_column='region',
    x_title="Sinema Seyirci Sayısı(oransal) \ Cinema Audiences(proportional)",
    y_title="Doktoralı Sayısı(oransal) \ Number of PhD(proportional)",
    title='Türkiye Doktoralı Sayısı ve Sinema Seyircisi Sayısı(Oransal) / Turkey Phd and Ci
    x_logscale=False, y_logscale=False, scale_bubble=3, width=1050, height=600)

iplot(figure_1)
```



In [ ]: