

```

In [1]: ## Matthew Lambaot homework 3
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

##### questions using Kirb21 data

kirb21_df= pd.read_csv("https://raw.githubusercontent.com/smart-stats/ds4bio_book/main/book/assets/kirby21AllLevels.csv") ## reading csv file
print(kirb21_df.head(4)) ## checking csv file contents
kirby906a_ax_df= kirb21_df.loc[(kirb21_df['rawid']=="kirby906a_ax.img").copy() ##creating new dataframe with only values that where rawid= kirby906a_ax.img
print(kirby906a_ax_df.head(4)) # print new dataframe
telencephalon_L= kirby906a_ax_df.loc[(kirby906a_ax_df['type'] ==1) & (kirby906a_ax_df['level'] ==1) & (kirby906a_ax_df['roi'] =='Telencephalon_L')].copy() ##creating new dataframe that only consists of type1 level1 data for Telecephalon_L
print(telencephalon_L.head(4))
telencephalon_L_volume= telencephalon_L.loc[12540,'volume'] ##assigning new variable to be the value for volume of this dataframe
print(telencephalon_L_volume) #printing volume value

telencephalon_R= kirby906a_ax_df.loc[(kirby906a_ax_df['type'] ==1) & (kirby906a_ax_df['level'] ==1) & (kirby906a_ax_df['roi'] =='Telencephalon_R')].copy()##creating new dataframe that only consists of type1 level1 data for Telecephalon_R
print(telencephalon_R.head(4))
telencephalon_R_volume= telencephalon_R.loc[12541,'volume']##assigning new variable to be the value for volume of this dataframe
print(telencephalon_R_volume) # printing volume value

ICV_fraction= (telencephalon_L_volume+telencephalon_R_volume)/kirby906a_ax_df.loc[12540,'icv'] ##creating ICV fraction variable that is the fraction of telencephalon volume to ICV
TBV_fraction= (telencephalon_L_volume+telencephalon_R_volume)/kirby906a_ax_df.loc[12540,'tbv']## creation of TBV fraction variable that is the fraciton of telecephalon volume to TBV
print(ICV_fraction)
print(TBV_fraction)

all_regions_df= (kirby906a_ax_df.groupby(["type","level"], as_index=False)[["volume"]].sum())## create new dataframe that groups the data in kirby906 by type and level, then takes the sum of the volumes for those parameters
all_regions_df= all_regions_df.rename(columns={"volume":"total_volume"}) ##renames the column name to be total_volume instead of volume
print(all_regions_df)

t1l2= kirby906a_ax_df.loc[(kirby906a_ax_df['type'] ==1) & (kirby906a_ax_df['level'] ==2)].copy() ##creating new dataframe that only contains type 1 level 2 data
Volume_bar= sns.barplot(x='roi', y='volume', data=t1l2) #creates a bar graph depicting the volume from each part of the brain from t1l2
plt.xticks(rotation=90) #rotates tick
plt.show()
plt.clf() ##clear the current plot figure
plt.cla() ##clear the current play axes

t1l3= kirby906a_ax_df.loc[(kirby906a_ax_df['type'] ==1) & (kirby906a_ax_df['level'] ==3)] ##creates new dataframe for only data that is type 1 level 3 from kirby
labels= t1l3['roi'] ##creates labels from roi of t1l3
Volume_donut= plt.pie( t1l3['volume'],textprops={'fontSize':'smaller'}, rotatelabels=270)##create pie plot
center_circle= plt.Circle((0,0), 0.70, fc='white') ##creates a white circle
fig=plt.gcf() ## return reference to the pie chart
fig.gca().add_artist(center_circle) ##adds center circle to the pie plot to make it a donut
plt.legend(labels,loc="upper right", fontsize=6, bbox_to_anchor=(1.25,5))
plt.show()

##### data using class_interests_df
plt.clf()
plt.cla()
class_interests_df= pd.read_table("https://raw.githubusercontent.com/bcaffo/ds4ph-bme/refs/heads/master/data/classInterests.txt")
sns.countplot(x='Program', hue='Year', data=class_interests_df) ##plots the data based the proportion of students from each year in the class per program
plt.xticks(rotation=90) #rotates tick
plt.show()

#####data using gene expression dataset

gene_expression_df= pd.read_csv("https://raw.githubusercontent.com/jhu-advdatasci/2018/refs/heads/master/data/GSE5859_exprs.csv") ## reading csv file
print(gene_expression_df.head(4))
gene_ids= gene_expression_df.iloc[:,0] ## splitting dataframe to remove string values and numerical values
expression_val= gene_expression_df.iloc[:,1:] ## splitting dataframe to remove string values and numerical values
row_means=expression_val.mean(axis=1) ## taking the mean of each row
gene_expression_df_pt1= expression_val.sub(row_means, axis=0) ##subtracting the row mean from each row
col_means= gene_expression_df_pt1.mean(axis=0) ##takes the mean of each column in the previous matrix
gene_expression_df_pt2= gene_expression_df_pt1.sub(col_means, axis=1) ##subtracts the column mean from each column
col_dev= gene_expression_df_pt2.std(axis=0) # calculates the standard deviation of each column
gene_expression_df_pt3= gene_expression_df_pt2.div(col_dev, axis=1) #subtracts the standard deviation from each column

gene_expression_df_pt1.insert(0, "gene_id", gene_ids) ## adds back gene ids to each matrix
gene_expression_df_pt2.insert(0, "gene_id", gene_ids)## adds back gene ids to each matrix
gene_expression_df_pt3.insert(0, "gene_id", gene_ids)## adds back gene ids to each matrix

print(gene_expression_df_pt1.head(4))
print(gene_expression_df_pt2.head(4))

```

```

print(gene_expression_df_pt3.head(4)) ##shows first 4 rows of final data

##### data using healthcare_df
healthcare_df= pd.read_csv("https://raw.githubusercontent.com/jhu-advdatasci/2018/master/data/KFF/healthcare-spending.csv", skiprows=2) ## reading csv file
print(healthcare_df.head(10))
us_states = [
    "Alabama", "Alaska", "Arizona", "Arkansas", "California", "Colorado", "Connecticut", "Delaware",
    "Florida", "Georgia", "Hawaii", "Idaho", "Illinois", "Indiana", "Iowa", "Kansas", "Kentucky",
    "Louisiana", "Maine", "Maryland", "Massachusetts", "Michigan", "Minnesota", "Mississippi",
    "Missouri", "Montana", "Nebraska", "Nevada", "New Hampshire", "New Jersey", "New Mexico",
    "New York", "North Carolina", "North Dakota", "Ohio", "Oklahoma", "Oregon", "Pennsylvania",
    "Rhode Island", "South Carolina", "South Dakota", "Tennessee", "Texas", "Utah", "Vermont",
    "Virginia", "Washington", "West Virginia", "Wisconsin", "Wyoming"
]
healthcare_df= healthcare_df[healthcare_df["Location"].isin(us_states)] ##filters out all data that isn't linked to one of the states in us_states
healthcare_df_melt= healthcare_df.melt(id_vars=["Location"], var_name="Year", value_name="Spending") ##transforms the dataframe from wide format to Long which makes it easier to graph by grouping each Location to a year and spending
healthcare_df_melt["Year"]=healthcare_df_melt["Year"].str.extract(r"(\d{4})").astype(int) ##pulls the first 4 characters in the Year tab to get only the year from 1991_Total Health Spending
print(healthcare_df_melt.head())
plt.figure(figsize=(16,8))
for state, group in healthcare_df_melt.groupby("Location"): ## groups the data by Location to create a line plot from each state for spending vs year
    plt.plot(group["Year"], group["Spending"], label=state)
plt.xlabel("Year")
plt.ylabel("Spending")
plt.legend(bbox_to_anchor=(1,1), loc="upper left")
plt.tight_layout()
plt.show()

plt.figure()

avg_Spending=healthcare_df_melt.groupby("Location")["Spending"].mean().sort_values() ##finds the mean spending from each state
avg_Spending.plot(kind="bar")
plt.ylabel("Average Healthcare Spending")
plt.tight_layout()
plt.show()

```

```

Unnamed: 0      rawid      roi  volume  min  max  \
0      1 kirby127a_3_1_ax.img  Telencephalon_L  531111  0.0  374.0
1      2 kirby127a_3_1_ax.img  Telencephalon_R  543404  0.0  300.0
2      3 kirby127a_3_1_ax.img  Diencephalon_L   9683  15.0  295.0
3      4 kirby127a_3_1_ax.img  Diencephalon_R   9678  10.0  335.0

      mean      std  type  level  id      icv      tbv
0  128.3013  51.8593    1      1  127  1378295  1268519
1  135.0683  53.6471    1      1  127  1378295  1268519
2  193.5488  32.2733    1      1  127  1378295  1268519
3  193.7051  32.7869    1      1  127  1378295  1268519

Unnamed: 0      rawid      roi  volume  min  max  \
12540  12541 kirby906a_ax.img  Telencephalon_L  467063  2.0  350.0
12541  12542 kirby906a_ax.img  Telencephalon_R  470488  2.0  337.0
12542  12543 kirby906a_ax.img  Diencephalon_L   8801  60.0  327.0
12543  12544 kirby906a_ax.img  Diencephalon_R   9054  63.0  415.0

      mean      std  type  level  id      icv      tbv
12540  165.2599  57.1707    1      1  906  1195015  1123076
12541  171.8695  59.3001    1      1  906  1195015  1123076
12542  227.1878  31.2303    1      1  906  1195015  1123076
12543  231.6770  31.1780    1      1  906  1195015  1123076

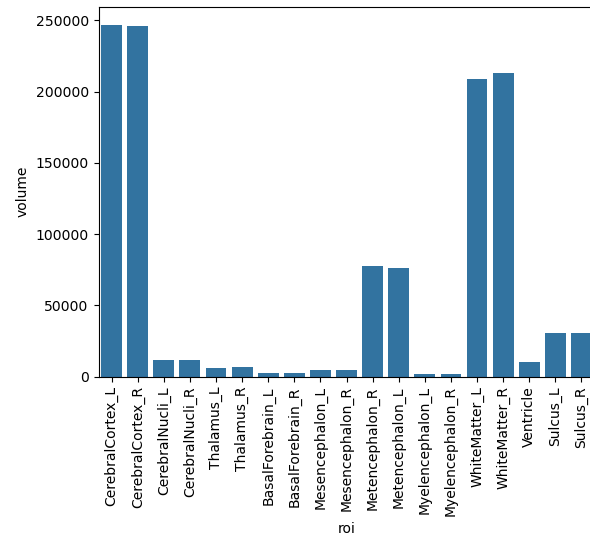
Unnamed: 0      rawid      roi  volume  min  max  \
12540  12541 kirby906a_ax.img  Telencephalon_L  467063  2.0  350.0

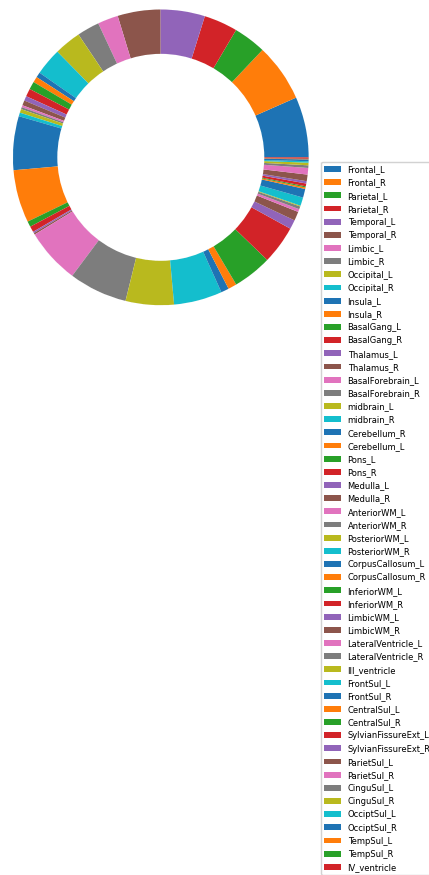
      mean      std  type  level  id      icv      tbv
12540  165.2599  57.1707    1      1  906  1195015  1123076
467063

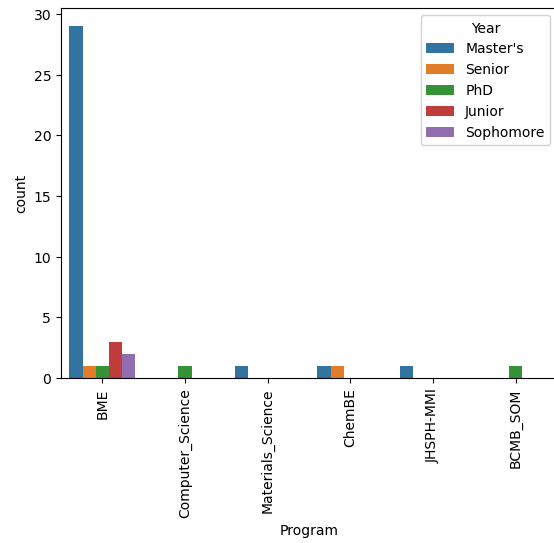
Unnamed: 0      rawid      roi  volume  min  max  \
12541  12542 kirby906a_ax.img  Telencephalon_R  470488  2.0  337.0

      mean      std  type  level  id      icv      tbv
12541  171.8695  59.3001    1      1  906  1195015  1123076
470488
0.7845516583473847
0.8348063710737297
type  level  total_volume
0      1      1      1195015
1      1      2      1195021
2      1      3      1195034
3      1      4      1195065
4      1      5      1195124
5      2      1      1195015
6      2      2      1195022
7      2      3      1195032
8      2      4      1195041
9      2      5      1195092

```







```

Unnamed: 0    GSM25581.CEL.gz    GSM25681.CEL.gz    GSM136524.CEL.gz    \
0  1007_s_at      6.333951      5.736190      6.156754
1  1053_at       7.255622      7.399993      7.491967
2  117_at        5.760106      4.825169      5.039387
3  121_at        6.935451      7.025210      7.543667

GSM136707.CEL.gz    GSM25553.CEL.gz    GSM136676.CEL.gz    GSM136711.CEL.gz    \
0  6.513010      6.061013      6.506403      6.863426
1  7.006123      7.250995      7.082581      6.842236
2  5.414160      5.205697      5.300078      5.099337
3  7.959781      7.223323      8.006816      8.102504

GSM136542.CEL.gz    GSM136535.CEL.gz    ...    GSM48650.CEL.gz    GSM25687.CEL.gz    \
0  6.369645      6.646321      ...      7.082636      6.315866
1  7.048487      7.042361      ...      6.599718      7.412731
2  5.156459      5.736109      ...      6.231457      5.240717
3  7.434617      7.532321      ...      8.215096      7.677593

GSM25685.CEL.gz    GSM136549.CEL.gz    GSM25427.CEL.gz    GSM25525.CEL.gz    \
0  7.010165      6.460107      6.122355      6.330314
1  7.274429      6.928642      7.168782      7.235648
2  5.275062      5.759827      5.409720      5.491938
3  7.923624      7.219401      7.432338      6.785174

GSM25349.CEL.gz    GSM136727.CEL.gz    GSM25626.CEL.gz    GSM136725.CEL.gz
0  6.627014      6.133068      6.419444      6.488579
1  6.939184      7.280781      6.351776      7.517410
2  5.113570      5.401876      5.537605      5.247190
3  7.833862      7.607461      7.302935      7.331864

[4 rows x 209 columns]
gene_id    GSM25581.CEL.gz    GSM25681.CEL.gz    GSM136524.CEL.gz    \
0  1007_s_at    -0.049313    -0.647073    -0.226509
1  1053_at     0.163992     0.308363     0.400337
2  117_at      0.283074    -0.651863    -0.437645
3  121_at     -0.648591    -0.558832    -0.040375

GSM136707.CEL.gz    GSM25553.CEL.gz    GSM136676.CEL.gz    GSM136711.CEL.gz    \
0  0.129747    -0.322250    0.123230    0.480163
1  -0.085507    0.159365    -0.009050    -0.249394
2  -0.062873    -0.271335    -0.176954    -0.377696
3  0.375740    -0.360718    0.422774    0.518462

GSM136542.CEL.gz    GSM136535.CEL.gz    ...    GSM48650.CEL.gz    GSM25687.CEL.gz    \
0  -0.013618     0.263058     ...      0.699373    -0.067397
1  -0.043143    -0.049269     ...     -0.491913     0.321100
2  -0.320573     0.259077     ...      0.754425    -0.236315
3  -0.149425    -0.051720     ...      0.631054     0.093552

GSM25685.CEL.gz    GSM136549.CEL.gz    GSM25427.CEL.gz    GSM25525.CEL.gz    \
0  0.626902     0.076844    -0.260908    -0.052949
1  0.182799    -0.162988     0.077152     0.144017
2  -0.201970     0.282794    -0.067313     0.014905
3  0.339583    -0.364641    -0.151704    -0.798867

GSM25349.CEL.gz    GSM136727.CEL.gz    GSM25626.CEL.gz    GSM136725.CEL.gz
0  0.243751    -0.250195     0.036100     0.105316
1  -0.152446     0.189150    -0.739855     0.425779
2  -0.363462    -0.075156     0.060573    -0.229842
3  0.249820     0.023419    -0.281107    -0.252178

[4 rows x 209 columns]
gene_id    GSM25581.CEL.gz    GSM25681.CEL.gz    GSM136524.CEL.gz    \
0  1007_s_at    -0.024490    -0.640032    -0.223988
1  1053_at      0.188815     0.315404     0.402858
2  117_at       0.307897    -0.644822    -0.435124
3  121_at     -0.623768    -0.551790    -0.037853

GSM136707.CEL.gz    GSM25553.CEL.gz    GSM136676.CEL.gz    GSM136711.CEL.gz    \
0  0.114936    -0.428929     0.109485     0.457383
1  -0.100318     0.052687    -0.022794    -0.272174
2  -0.077684    -0.378014    -0.190609    -0.400475
3  0.360928    -0.467397     0.409030     0.495682

GSM136542.CEL.gz    GSM136535.CEL.gz    ...    GSM48650.CEL.gz    GSM25687.CEL.gz    \
0  -0.017008     0.250137     ...      0.678049    -0.061456
1  -0.046533    -0.062190     ...     -0.513237     0.327042
2  -0.323963     0.246157     ...      0.733100    -0.230374
3  -0.152814    -0.064641     ...      0.609730     0.099493

GSM25685.CEL.gz    GSM136549.CEL.gz    GSM25427.CEL.gz    GSM25525.CEL.gz    \
0  0.665913     0.076758    -0.251423    -0.031388
1  0.221810    -0.163074     0.086637     0.165579
2  -0.162959     0.282708    -0.057828     0.036467

```

```

3      0.378593      -0.364727      -0.142219      -0.777306

GSM25349.CEL.gz GSM136727.CEL.gz GSM25626.CEL.gz GSM136725.CEL.gz
0      0.235241      -0.261128      0.041275      0.090015
1      -0.160956      0.178217      -0.734760      0.410478
2      -0.371972      -0.086090      0.065668      -0.245143
3      0.241310      0.012486      -0.276012      -0.267479

[4 rows x 209 columns]
gene_id GSM25581.CEL.gz GSM25681.CEL.gz GSM136524.CEL.gz \
0 1007_s_at -0.062810 -1.694428 -0.797346
1 1053_at 0.484259 0.835005 1.434085
2 117_at 0.789673 -1.707110 -1.548943
3 121_at -1.599798 -1.460817 -0.134750

GSM136707.CEL.gz GSM25553.CEL.gz GSM136676.CEL.gz GSM136711.CEL.gz \
0 0.362538 -1.152413 0.366177 1.252187
1 -0.316432 0.141554 -0.076237 -0.745136
2 -0.245036 -1.015618 -0.637799 -1.096388
3 1.138467 -1.255766 1.368016 1.357039

GSM136542.CEL.gz GSM136535.CEL.gz ... GSM48650.CEL.gz GSM25687.CEL.gz \
0 -0.054154 1.008166 ... 1.446603 -0.185273
1 -0.148166 -0.250652 ... -1.094982 0.985940
2 -1.031533 0.992122 ... 1.564055 -0.694512
3 -0.486577 -0.260531 ... 1.300847 0.299944

GSM25685.CEL.gz GSM136549.CEL.gz GSM25427.CEL.gz GSM25525.CEL.gz \
0 1.615970 0.245537 -0.568262 -0.069236
1 0.538265 -0.521649 0.195814 0.365237
2 -0.395453 0.904341 -0.130702 0.080440
3 0.918732 -1.166704 -0.321440 -1.714599

GSM25349.CEL.gz GSM136727.CEL.gz GSM25626.CEL.gz GSM136725.CEL.gz
0 0.617954 -0.598607 0.088393 0.257410
1 -0.422815 0.408542 -1.573530 1.173822
2 -0.977132 -0.197351 0.140632 -0.701024
3 0.633897 0.028623 -0.591095 -0.764897

[4 rows x 209 columns]
Location 1991__Total Health Spending \
0 United States 675896.0
1 Alabama 10393.0
2 Alaska 1458.0
3 Arizona 9269.0
4 Arkansas 5632.0
5 California 81438.0
6 Colorado 8460.0
7 Connecticut 10950.0
8 Delaware 1938.0
9 District of Columbia 2800.0

1992__Total Health Spending 1993__Total Health Spending \
0 731455.0 778684.0
1 11284.0 12028.0
2 1558.0 1661.0
3 9815.0 10655.0
4 6022.0 6397.0
5 87949.0 91963.0
6 9215.0 9803.0
7 11635.0 12081.0
8 2111.0 2285.0
9 3098.0 3240.0

1994__Total Health Spending 1995__Total Health Spending \
0 820172.0 869578.0
1 12742.0 13590.0
2 1728.0 1879.0
3 11364.0 12042.0
4 6810.0 7343.0
5 94245.0 96870.0
6 10382.0 11153.0
7 12772.0 13649.0
8 2489.0 2655.0
9 3255.0 3285.0

1996__Total Health Spending 1997__Total Health Spending \
0 917540.0 969531.0
1 14450.0 15462.0
2 2076.0 2240.0
3 12850.0 13418.0
4 7817.0 8393.0
5 100215.0 103681.0
6 11863.0 12572.0

```

7	14139.0	14948.0
8	2772.0	3026.0
9	3362.0	3374.0

	1998__Total Health Spending	1999__Total Health Spending	...	\
0	1026103.0	1086280.0	...	
1	15860.0	16451.0	...	
2	2386.0	2569.0	...	
3	14465.0	15550.0	...	
4	8814.0	9407.0	...	
5	111224.0	116036.0	...	
6	13790.0	14764.0	...	
7	15944.0	16785.0	...	
8	3207.0	3539.0	...	
9	3461.0	3578.0	...	

	2005__Total Health Spending	2006__Total Health Spending	\
0	1696222.0	1804672.0	
1	25338.0	26638.0	
2	4765.0	5048.0	
3	28190.0	30766.0	
4	14611.0	15431.0	
5	182958.0	194413.0	
6	22867.0	24849.0	
7	24538.0	25997.0	
8	5899.0	6285.0	
9	4971.0	5138.0	

	2007__Total Health Spending	2008__Total Health Spending	\
0	1918820.0	2010690.0	
1	27700.0	28765.0	
2	5426.0	5807.0	
3	33366.0	35547.0	
4	16426.0	17246.0	
5	209397.0	221013.0	
6	26525.0	27797.0	
7	27488.0	29141.0	
8	6735.0	7191.0	
9	5492.0	5779.0	

	2009__Total Health Spending	2010__Total Health Spending	\
0	2114221.0	2194625.0	
1	30095.0	30728.0	
2	6112.0	6519.0	
3	37258.0	38620.0	
4	18071.0	18735.0	
5	229541.0	241916.0	
6	29246.0	30187.0	
7	31132.0	31727.0	
8	7495.0	7938.0	
9	6182.0	6582.0	

	2011__Total Health Spending	2012__Total Health Spending	\
0	2272582.0	2365948.0	
1	31398.0	32848.0	
2	6928.0	7406.0	
3	39295.0	40495.0	
4	19356.0	20076.0	
5	253844.0	266767.0	
6	31372.0	32726.0	
7	32129.0	33421.0	
8	8365.0	8650.0	
9	7000.0	7130.0	

	2013__Total Health Spending	2014__Total Health Spending
0	2435624.0	2562824.0
1	33788.0	35263.0
2	7684.0	8151.0
3	41481.0	43356.0
4	20500.0	21980.0
5	278168.0	291989.0
6	34090.0	36398.0
7	34223.0	35413.0
8	9038.0	9587.0
9	7443.0	7871.0

[10 rows x 25 columns]

	Location	Year	Spending
0	Alabama	1991	10393.0
1	Alaska	1991	1458.0
2	Arizona	1991	9269.0
3	Arkansas	1991	5632.0
4	California	1991	81438.0



