data_engineering

November 25, 2019

In [1]:
```
import re
import pandas as pd
import pickle
import json
import numpy as np
import matplotlib.pyplot as plt
```

In [2]:
```
df = pd.read_stata(
    './ICPSR_20367-V1/ICPSR_20367/DS0001/20367-0001-Data.dta'
)
#'/home/ethan/Repos/data_project/data'
```

This section of the pdf is aggregating data by state. The organiztion of data here will be helpful in understanding the trends seen among the states

In [3]:
```
states = df.groupby('STATE', sort=True)
indices = states.indices
new_index = list(indices.keys())
factors = states.GOVTYPE.count().values
```

In [4]:
```
new_df = pd.DataFrame(index=new_index)
```

In [5]:
```
new_df['CONFPOP'] = states.CONFPOP.sum()
# new_df['TOTPOP'] = states.TOTPOP.sum()
new_df['MALE'] = states.ADMALE.sum()
new_df['MALE_PERC'] = new_df.MALE / new_df.CONFPOP
new_df['JUVMALE'] = states.JUVMALE.sum()
new_df['JUVMALE_PERC'] = new_df.JUVMALE / new_df.CONFPOP
new_df['FEM'] = states.ADFEML.sum()
new_df['FEM_PERC'] = new_df.FEM / new_df.CONFPOP
new_df['JUVFEM'] = states.JUVFEML.sum()
new_df['JUVFEM_PERC'] = new_df.JUVFEM / new_df.CONFPOP
new_df['WHITE'] = states.WHITE.sum()
new_df['WHITE_PERC'] = new_df.WHITE / new_df.CONFPOP
new_df['BLACK'] = states.BLACK.sum()
new_df['BLACK_PERC'] = new_df.BLACK / new_df.CONFPOP
new_df['HISP'] = states.HISP.sum()
new_df['HISP_PERC'] = new_df.HISP / new_df.CONFPOP
```

```
new_df['ASIAN'] = states.ASIAN.sum()
new_df['ASIAN_PERC'] = new_df.ASIAN / new_df.CONFPOP
new_df['ICE'] = states.ICE.sum()
new_df['ICE_PERC'] = new_df.ICE / new_df.CONFPOP
new_df['BIA'] = states.BIA.sum()
# new_df['Males'] = states.ADMALE.mean() * states.ADMALE.count()
# new_df['Males'] = states.ADMALE.mean() * states.ADMALE.count()
# new_df['Males'] = states.ADMALE.mean() * states.ADMALE.count()
# new_df['Males'] = states.ADMALE.mean() * states.ADMALE.count()
# new_df['Males'] = states.ADMALE.mean() * states.ADMALE.count()
```

This new data frame has the sum of each of the desired columns by state and I have also added different rate (of incarceration) columns. These will make the data more comparable between states. All of this data is in jails only.

```
In [6]: new_df
```

Out[6]:

	CONFPOP	MALE	MALE_PERC	JUVMALE	JUVMALE_PERC	FEM	FEM_PERC	JUVFEM
AL	15143	13062	0.862577	31	0.002047	2050	0.135376	0
AK	65	55	0.846154	0	0.000000	10	0.153846	0
AZ	15479	12973	0.838103	217	0.014019	2281	0.147361	8
AR	6125	4905	0.800816	58	0.009469	1147	0.187265	15
CA	84030	73159	0.870630	6	0.000071	10865	0.129299	0
CO	13638	11680	0.856431	39	0.002860	1918	0.140636	1
DC	3552	3115	0.876971	15	0.004223	422	0.118806	0
FL	65166	56032	0.859835	599	0.009192	8509	0.130574	26
GA	44965	38863	0.864294	344	0.007650	5710	0.126988	48
HI	605	529	0.874380	0	0.000000	76	0.125620	0
ID	3787	2961	0.781885	17	0.004489	809	0.213626	0
IL	20795	18204	0.875403	265	0.012743	2307	0.110940	19
IN	17567	15271	0.869300	107	0.006091	2166	0.123299	23
IA	3637	3123	0.858675	23	0.006324	491	0.135001	0
KS	6904	5934	0.859502	15	0.002173	955	0.138326	0
KY	16761	14131	0.843088	5	0.000298	2623	0.156494	2
LA	32579	28649	0.879370	448	0.013751	3396	0.104239	86
ME	1545	1377	0.891262	0	0.000000	168	0.108738	0
MD	12386	10795	0.871549	195	0.015744	1388	0.112062	8
MA	12619	11878	0.941279	73	0.005785	668	0.052936	0
MI	18118	15546	0.858042	314	0.017331	2235	0.123358	23
MN	7023	6152	0.875979	15	0.002136	855	0.121743	1
MS	11422	10386	0.909298	95	0.008317	936	0.081947	5
MO	10461	8969	0.857375	110	0.010515	1377	0.131632	5
MT	2265	1944	0.858278	11	0.004857	303	0.133775	7
NE	3098	2695	0.869916	1	0.000323	401	0.129438	1
NV	7110	5920	0.832630	26	0.003657	1164	0.163713	0
NH	1728	1471	0.851273	35	0.020255	220	0.127315	2
NJ	17621	15824	0.898019	25	0.001419	1771	0.100505	1
NM	8514	7482	0.878788	54	0.006342	962	0.112990	16

NY	33341	28991	0.869530	1316	0.039471	2900	0.086980	134
NC	17171	14700	0.856095	463	0.026964	1934	0.112632	74
ND	944	790	0.836864	25	0.026483	129	0.136653	0
OH	19853	16904	0.851458	47	0.002367	2902	0.146174	0
OK	9585	8195	0.854982	47	0.004903	1341	0.139906	2
OR	6549	5637	0.860742	23	0.003512	888	0.135593	1
PA	35573	31162	0.876001	214	0.006016	4191	0.117814	6
SC	12226	10756	0.879764	86	0.007034	1368	0.111893	16
SD	1432	1166	0.814246	2	0.001397	264	0.184358	0
TN	24233	20692	0.853877	67	0.002765	3471	0.143234	3
TX	67418	58124	0.862144	322	0.004776	8931	0.132472	41
UT	6739	5562	0.825345	20	0.002968	1157	0.171687	0
VA	26424	23155	0.876287	60	0.002271	3209	0.121443	0
WA	13611	11591	0.851591	32	0.002351	1988	0.146058	0
WV	4077	3692	0.905568	0	0.000000	385	0.094432	0
WI	14304	12317	0.861088	270	0.018876	1685	0.117799	32
WY	1551	1305	0.841393	14	0.009026	230	0.148291	2
PR	1056	974	0.922348	0	0.000000	82	0.077652	0

	JUVFEM_PERC	WHITE	WHITE_PERC	BLACK	BLACK_PERC	HISP	HISP_PERC	\
AL	0.000000	6747	0.445552	6847	0.452156	457	0.030179	
AK	0.000000	34	0.523077	0	0.000000	2	0.030769	
AZ	0.000517	6974	0.450546	1706	0.110214	5802	0.374830	
AR	0.002449	3234	0.528000	2179	0.355755	204	0.033306	
CA	0.000000	25818	0.307247	17562	0.208997	32169	0.382828	
CO	0.000073	7915	0.580364	1946	0.142690	3101	0.227379	
DC	0.000000	93	0.026182	3310	0.931869	125	0.035191	
FL	0.000399	33351	0.511785	27111	0.416030	4364	0.066967	
GA	0.001067	15597	0.346870	25193	0.560280	2315	0.051484	
HI	0.000000	208	0.343802	27	0.044628	0	0.000000	
ID	0.000000	1579	0.416953	32	0.008450	540	0.142593	
IL	0.000914	6274	0.301707	11204	0.538783	2219	0.106708	
IN	0.001309	10493	0.597313	4541	0.258496	706	0.040189	
IA	0.000000	2283	0.627715	786	0.216112	295	0.081111	
KS	0.000000	4048	0.586327	1773	0.256808	814	0.117903	
KY	0.000119	11698	0.697930	4327	0.258159	416	0.024820	
LA	0.002640	8518	0.261457	20069	0.616010	266	0.008165	
ME	0.000000	1426	0.922977	63	0.040777	29	0.018770	
MD	0.000646	4073	0.328839	7627	0.615776	555	0.044809	
MA	0.000000	5084	0.402885	2925	0.231793	3141	0.248910	
MI	0.001269	10282	0.567502	6577	0.363009	595	0.032840	
MN	0.000142	3515	0.500498	1030	0.146661	411	0.058522	
MS	0.000438	3364	0.294519	7504	0.656978	177	0.015496	
MO	0.000478	5121	0.489533	4043	0.386483	391	0.037377	
MT	0.003091	1514	0.668433	57	0.025166	100	0.044150	
NE	0.000323	1748	0.564235	672	0.216914	416	0.134280	
NV	0.000000	3131	0.440366	1837	0.258368	1526	0.214627	
NH	0.001157	1364	0.789352	163	0.094329	177	0.102431	

NJ	0.000057	4820	0.273537	9221	0.523296	3161	0.179388
NM	0.001879	1712	0.201081	522	0.061311	5185	0.608997
NY	0.004019	9924	0.297652	15926	0.477670	6117	0.183468
NC	0.004310	5626	0.327645	9416	0.548366	1377	0.080193
ND	0.000000	620	0.656780	46	0.048729	37	0.039195
OH	0.000000	10902	0.549136	7993	0.402609	509	0.025638
OK	0.000209	5173	0.539697	2321	0.242149	726	0.075743
OR	0.000153	4783	0.730341	494	0.075431	732	0.111773
PA	0.000169	16681	0.468923	14518	0.408119	3580	0.100638
SC	0.001309	4029	0.329544	7599	0.621544	379	0.031000
SD	0.000000	816	0.569832	89	0.062151	68	0.047486
TN	0.000124	12271	0.506376	10845	0.447530	659	0.027194
TX	0.000608	24910	0.369486	19619	0.291005	17908	0.265626
UT	0.000000	4678	0.694168	307	0.045556	1316	0.195281
VA	0.000000	9022	0.341432	14111	0.534022	1552	0.058734
WA	0.000000	9084	0.667401	2093	0.153773	1168	0.085813
WV	0.000000	3070	0.753005	638	0.156488	20	0.004906
WI	0.002237	8306	0.580677	4012	0.280481	765	0.053482
WY	0.001289	1166	0.751773	58	0.037395	135	0.087041
PR	0.000000	462	0.437500	579	0.548295	0	0.000000

	ASIAN	ASIAN_PERC	ICE	ICE_PERC	BIA
AL	4	0.000264	350	0.023113	0
AK	2	0.030769	0	0.000000	0
AZ	66	0.004264	156	0.010078	18
AR	12	0.001959	27	0.004408	0
CA	1705	0.020290	2270	0.027014	0
CO	89	0.006526	279	0.020458	2
DC	10	0.002815	56	0.015766	0
FL	170	0.002609	1083	0.016619	0
GA	107	0.002380	274	0.006094	0
HI	365	0.603306	0	0.000000	0
ID	24	0.006337	49	0.012939	5
IL	28	0.001346	136	0.006540	0
IN	16	0.000911	8	0.000455	0
IA	23	0.006324	151	0.041518	0
KS	40	0.005794	79	0.011443	3
KY	11	0.000656	14	0.000835	0
LA	76	0.002333	95	0.002916	0
ME	7	0.004531	2	0.001294	0
MD	70	0.005652	227	0.018327	0
MA	125	0.009906	556	0.044061	0
MI	86	0.004747	257	0.014185	0
MN	98	0.013954	135	0.019223	3
MS	30	0.002627	11	0.000963	0
MO	30	0.002868	199	0.019023	0
MT	7	0.003091	2	0.000883	0
NE	18	0.005810	54	0.017431	2

```
NV    121    0.017018   178  0.025035    2
NH     15    0.008681     3  0.001736    0
NJ    228    0.012939  1055  0.059872    0
NM     14    0.001644    53  0.006225   13
NY    275    0.008248   789  0.023665    1
NC     30    0.001747   167  0.009726    9
ND      7    0.007415     3  0.003178    9
OH     40    0.002015    96  0.004836    0
OK     57    0.005947    69  0.007199    6
OR     53    0.008093    92  0.014048    4
PA    434    0.012200   734  0.020634    0
SC     10    0.000818     9  0.000736    0
SD      5    0.003492    20  0.013966    6
TN     16    0.000660   187  0.007717    0
TX    147    0.002180   904  0.013409    3
UT     28    0.004155    88  0.013058    1
VA    136    0.005147   691  0.026150    0
WA    394    0.028947    68  0.004996   33
WV      3    0.000736     2  0.000491    0
WI     97    0.006781   231  0.016149    0
WY      2    0.001289    10  0.006447    0
PR     15    0.014205     0  0.000000    0
```

```
In [7]: new_df.to_csv('state_jail_data.csv')
```

This next data set that I'm going to load in is prison and jail data. It also has some other state wide information that will be useful

```
In [8]: df = pd.read_csv('incarceration_by_race.csv')
```

Now we insert the data

```
In [9]: pop = df[['TotalPop','White_pop','Black_pop','Asian_pop']]
```

```
In [10]: new_df[['TotalPop','White_pop','Black_pop','Asian_pop']] = pop
```

```
In [ ]:
```

```
In [ ]:
```