STAT515 - Applied Statistics & Visualization for Analytics

FINAL PROJECT REPORT ON

UNITED STATES STATISTICS

BY

SriUdaya Damuluri

G01111092

About Data Set:

The data set consists of varied statistics in general about the states in United States of America. The intention behind choosing this project is to observe different patterns pertaining to population with respect to each state. The data set was collected from Kaggle. It includes numeric values about population for different years, population increase in a decade, percent of total population residing in metro area, death rate due to heart disease and cancer, percent of population without insurance, crime rate, unemployment rate, etc.

Data Pre-processing:

The data set encapsulated huge information about all the states in US which needed processing to derive meaningful conclusions. The raw data had 100 variables which are not useful for the analysis and visualization. Since data cleaning is an essential part of data analysis, I had to trim the extrinsic data and include only the variables which were required for the final conclusion. I also had to transpose the rows and columns of the data to make it convenient for the analysis since variables are required to be arranged in columns. The data cleaning was done after clearly understanding the structure of the data thus maintaining the data integrity. The difference between raw data and processed data can be seen from the figures 1.1 and 1.2.

	Unit	U.S.	AL	AK	AZ	AR	CA	CO	CT	DE	DC	FL	GA
POPULATION													
Total persons: (July 1)													
1993	1000's	257908	4187	599	3936	2424	31211	3566	3277	700	578	13679	6917
2000	1000's	276242	4485	699	4437	2578	34888	4059	3271	759	537	15313	7637
Percent increase:													
1990 to 1993	Percent	3.7	3.6	8.9	7.4	3.1	4.9	8.2	-0.3	5.1	-4.7	5.7	6.8
1990 to 2000 (1)	Percent	11.1	11	27.1	21.1	9.7	17.2	23.2	-0.5	13.9	-11.5	18.4	17.9
65 yrs and over	Percent	12.7	13	4.4	13.4	15	10.6	10	14.1	12.4	13.3	18.6	10.1
Residing in a metro area	Percent	79.7	67.4	41.8	84.7	44.7	96.7	81.8	95.7	82.7	100	93	67.7
VITAL STATISTICS AND HEALTH													
Birth rate per 1,000 people	Rate	16.3	15.4	20.5	18.2	15	20.1	15.9	14.8	16.5	19.7	14.6	16.7
Infant deaths per 1,000 live births,	Rate	8.9	11.2	8.9	8.6	10.2	7.6	8.9	7.4	11.8	21	9	
Death rate per 100,000 people													
Heart disease	Rate	286	322	83	234	346	222	182	291	283	312	348	249
Cancer	Rate	204	216	88	191	236	165	154	214	224	259	260	175
Persons without health insurance	Percent	14.2	17.3	14.9	15.8	17.6	19	12.4	7.5	12.7	22	18.7	16.1
Community hospitals													
Occupancy rate (2)	Rate	65.6	62.1	53.7	60.2	59.1	62.5	61.6	75.6	70.2	74.4	61.2	65.3
Avg cost per patient per day (3)	Dollars	820	729	1116	1051	633	1134	904	1012	920	1124	886	721

Figure 1.1 Data Set Before Pre-Processing

Data Set:

	A	B	C	D	E	F	G	H	I	J	K	L
1	State	Pop_1993	Pop_2000	Percent_Increase	ResideInMetro	Death_HeartDisease	Death_Cancer	Crime	FederalAndStatePrisoners	UnemploymentRate	GrossStateIncome	BirthRate
2	Alabama	4187	4485	11	3023	322	216	47	42	7.5	70	15.4
3	Alaska	599	699	27.1	292	83	88	73	49	7.6	26	20.5
4	Arizona	3936	4437	21.1	3758	234	191	41	34	6.2	67	18.2
5	Arkansas	2424	2578	9.7	1152	346	236	52	35	6.2	39	15
6	California	31211	34888	17.2	33737	222	165	43	35	9.2	745	20.1
7	Colorado	3566	4059	23.2	3320	182	154	32	26	5.2	71	15.9
8	Connecticut	3277	3271	-0.5	3130	291	214	22	17	6.2	94	14.8
9	Delaware	700	759	13.9	628	283	224	50	41	5.3	20	16.5
10	Florida	13679	15313	18.4	14241	348	260	46	36	7	245	14.6
11	Georgia	6917	7637	17.9	5170	249	175	38	31	5.8	137	16.7
12	Hawaii	1172	1327	19.7	991	180	146	30	25	4.2	29	17.6
13	Idaho	1099	1290	28.1	387	225	165	22	20	6.1	19	16.2
14	Illinois	11697	12168	6.5	10221	309	212	38	27	7.4	272	16.8
15	Indiana	5713	6045	9	4328	299	214	39	25	5.3	112	15.3
16	Iowa	2814	2930	5.5	1283	346	228	29	16	4	56	13.9
17	Kansas	2531	2722	9.9	1486	301	206	39	24	5	51	15.2
18	Kentucky	3789	3989	8.2	1935	322	230	22	14	6.2	67	14.6
19	Louisiana	4295	4478	6.1	3359	293	209	54	35	7.4	91	17
20	Maine	1239	1240	1	443	300	238	26	12	7.9	23	13.6
21	Maryland	4965	5322	11.3	4939	242	201	46	37	6.2	109	16.3
22	Massachuset	6012	5950	-1.1	5724	285	230	39	17	6.9	154	14.7
23	Michigan	9478	9759	5	8071	295	206	42	37	7	188	16
24	Minnesota	4517	4824	10.3	3343	241	189	24	9	5.1	100	15.1
25	Mississippi	2643	2750	6.9	844	372	213	28	19	6.3	40	16.7
26	Missouri	5234	5437	6.3	3713	345	229	50	31	6.4	104	15.3
27	Montana	839	920	15.1	221	241	203	35	18	6	13	14.2
28	Nebraska	1607	1704	8	862	322	200	27	16	2.6	33	15.1
29	Nevada	1389	1691	40.7	1434	250	186	70	45	7.2	31	17.2
30	New Hampsh	1125	1165	5	692	246	203	20	16	6.6	24	14.8
31	New Jersey	7879	8135	5.2	8135	301	234	26	17	7.4	208	15.6
32	New Mexico	1616	1823	20.3	1021	200	152	66	32	7.5	27	18
33	New York	18197	18237	1.4	16723	353	213	38	32	7.7	467	16.2
34	North Carolir	6945	7617	14.9	5050	281	198	35	3	4.9	141	15.2
35	North Dakot	635	643	0.7	267	286	216	24	9	4.3	12	14
36	Ohio	11091	11453	5.6	9311	320	222	30	23	6.5	222	15.2
37	Oklahoma	3231	3382	7.5	2033	340	215	42	31	6	56	15.1
38	Oregon	3032	3404	19.8	2383	246	213	26	21	7.2	55	14.5
39	Pennsylvania	12048	12296	3.5	10427	363	251	32	21	7	245	14.1
40	Rhode Island	1000	998	-0.5	934	323	236	24	16	7.7	21	14.7

Figure 1.2 Data Set for Analysis

The processed data consists of the following variables:

- State
- Population in 1990’s
- Population in 2000’s
- Percent increase in population
- Population residing in metro area
- Death rate due to Heart disease
- Death rate due to Cancer
- Crime rate
- State prisoners
- Unemployment rate
- Gross State Income
- Birth rate
- Infant Deaths

VISUALIZATIONS:

Line Plot:

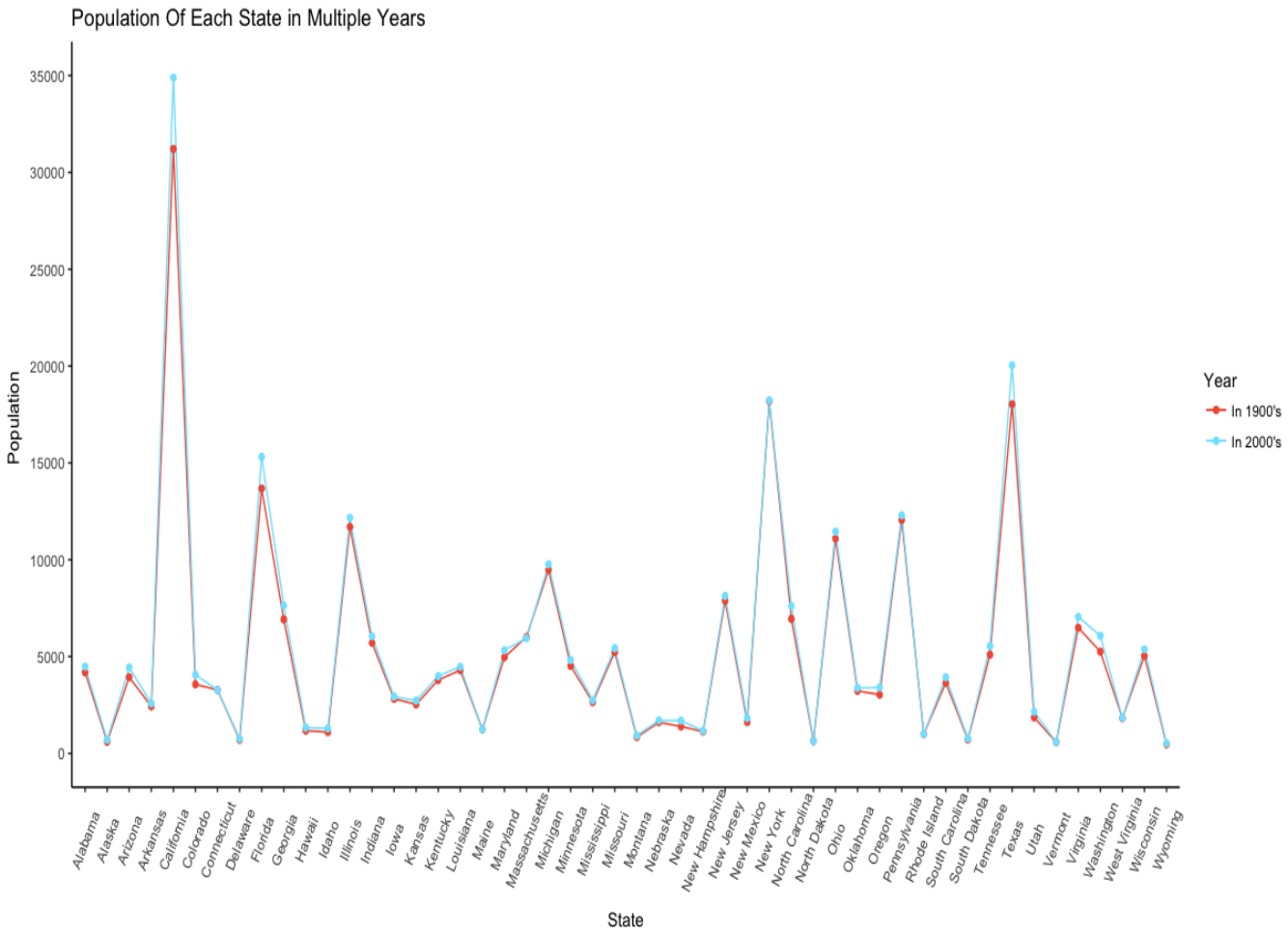

Figure 1.3 Line plot comparing population change in one decade (In Thousands)

The above line graph represents the total population of each state for a decade. The red line represents the population in the 1990's and the blue line represents the total population in 2000's. We can observe from the plot that California is the most populous state followed by Texas, New York, Florida. We can also observe that the population has increased over the years by observing the blue line in most of the states except for Alaska, Connecticut, Massachusetts, Rhode Island.

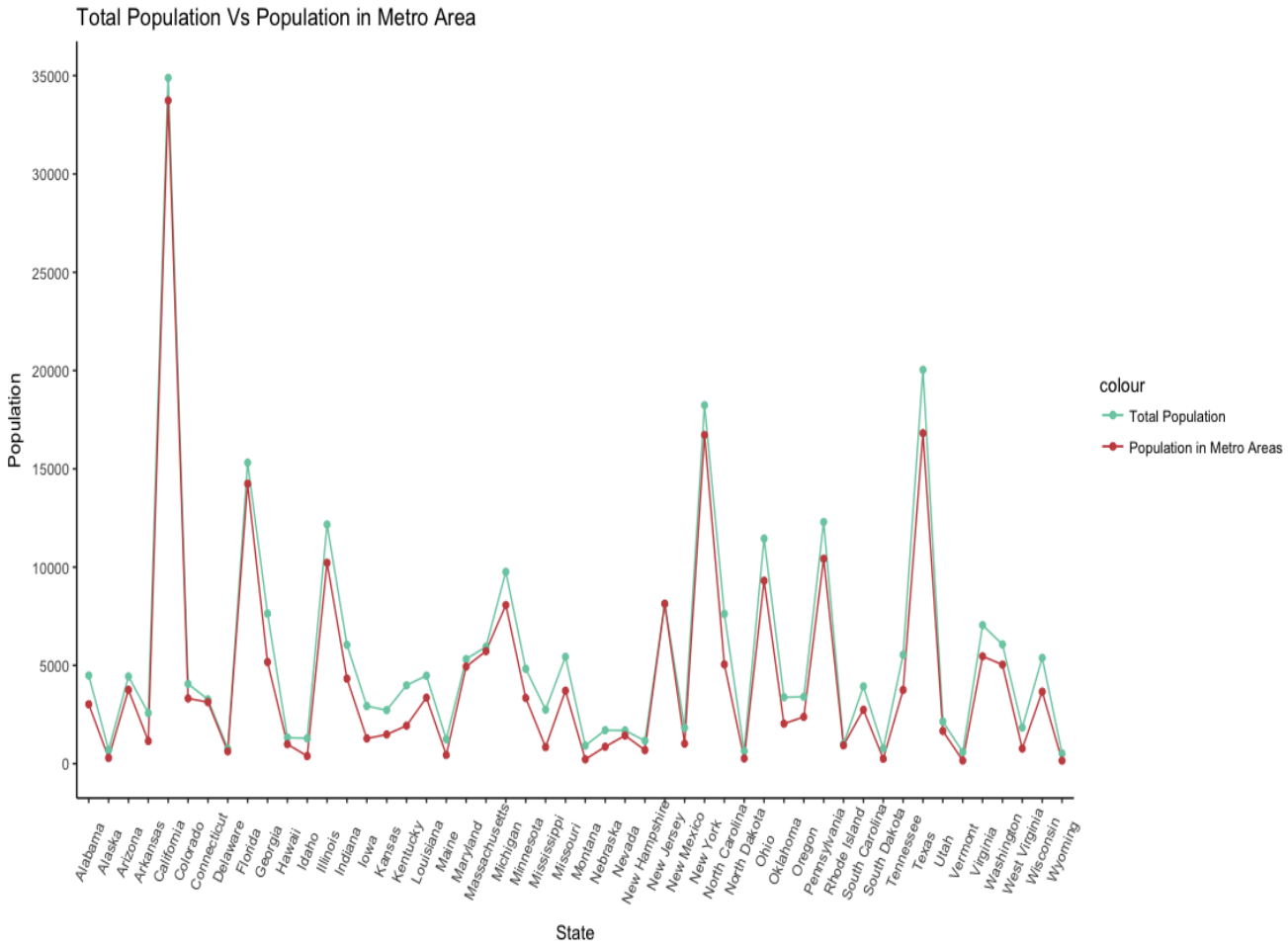

Figure 1.4 Line plot comparing the population residing in metro area (In Thousands)

The above plot depicts the number of people living in metro area and we can observe that more than 50 percent of the population in each state live in the metro areas.

I used a line plot to compare the above variables (Population in 1900, Population in 2000 and People residing in metro) since it is an easy to visualize data for all the states without any ambiguities.

Scatter Plot:

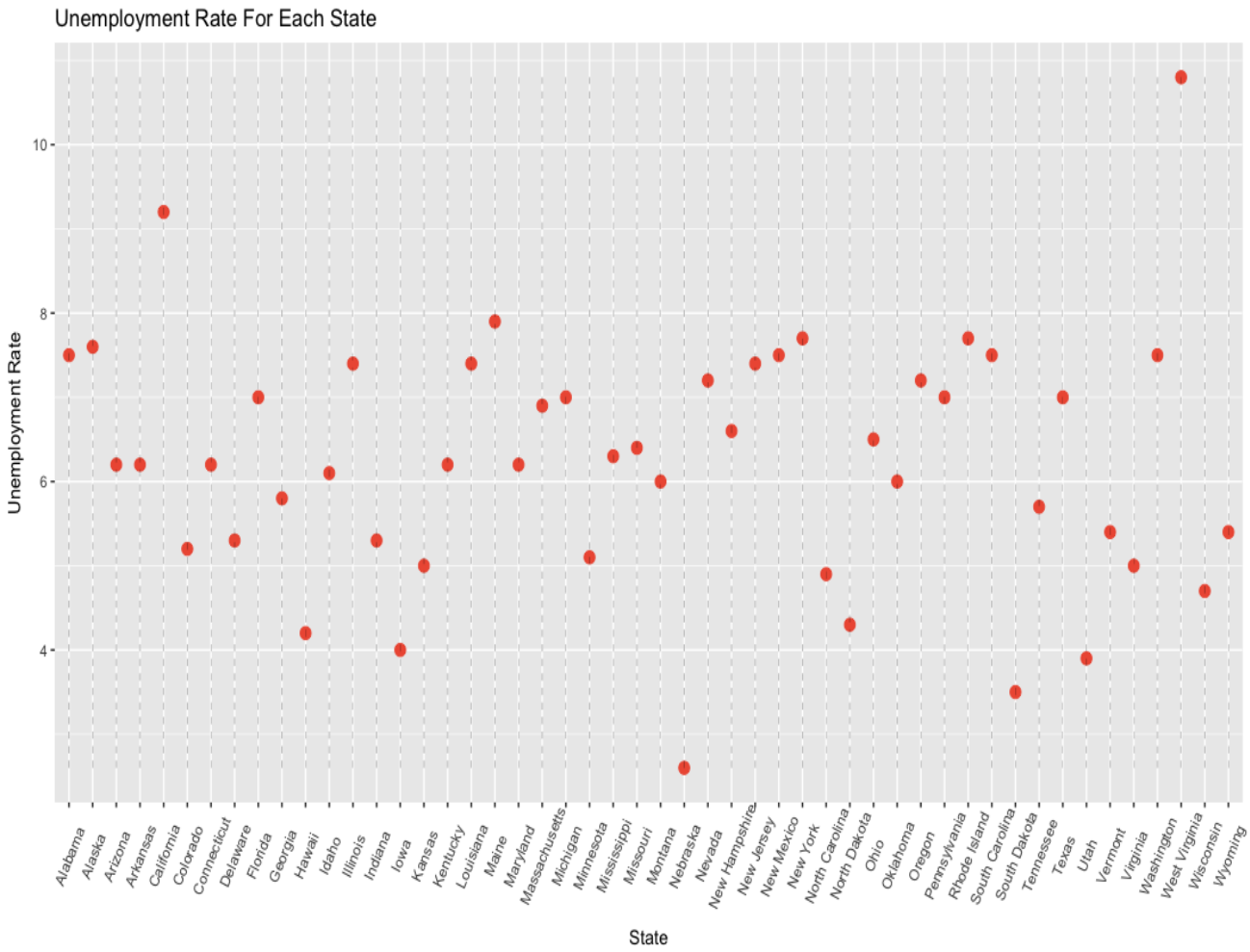

Figure 1.5 Unemployment rate for each state (In Percent)

The scatter plot above represents the unemployment rate for each state in percentage. We can notice that Nebraska has the lowest unemployment rate while West Virginia has the highest unemployment rate. There is no correlation between the X and Y axis since one of them is non-numeric but we can observe that the unemployment rate is quite similar in most of the states.

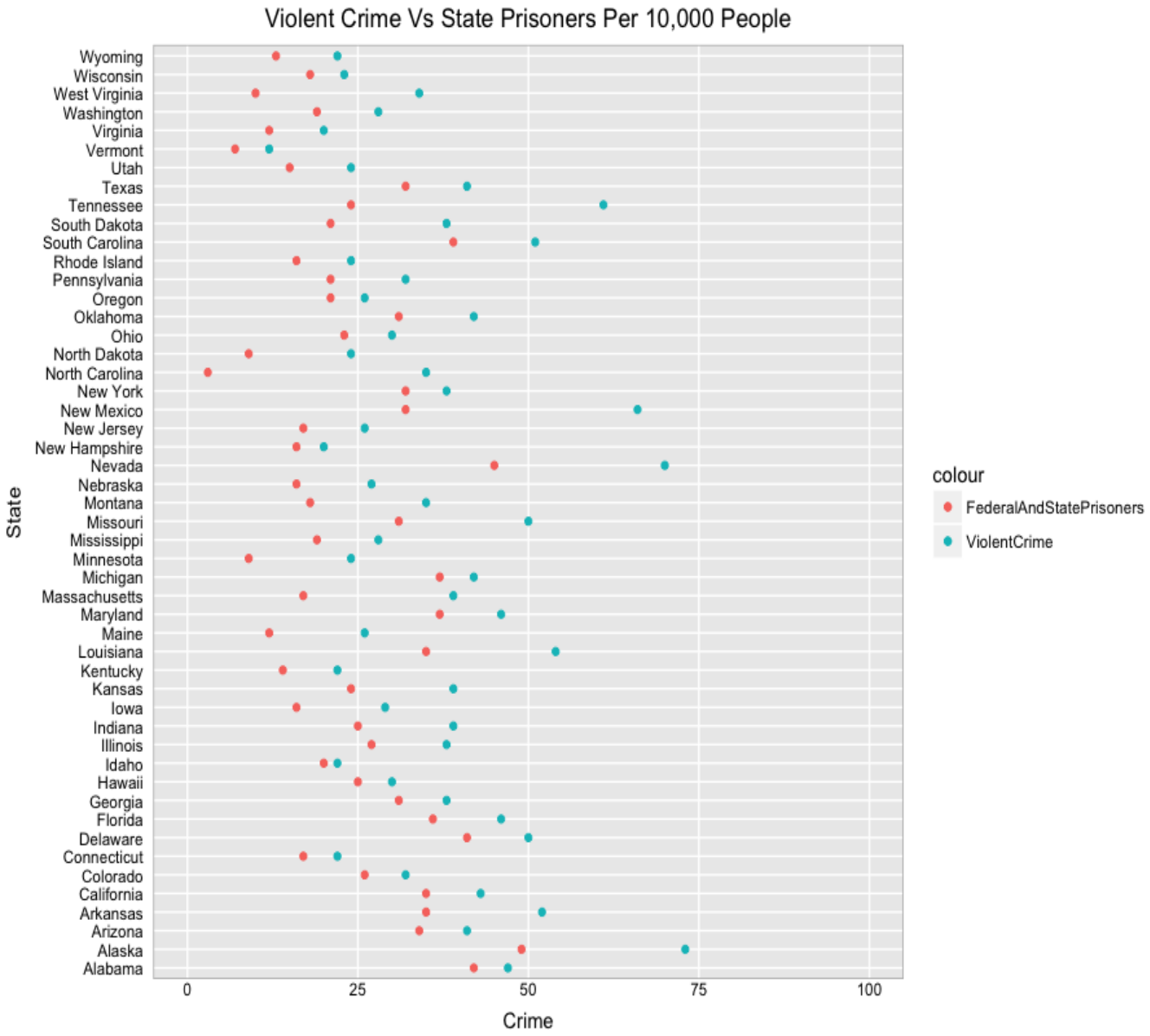

Figure 1.6 Comparing violent crime and state prisoners

The above graph represents the violent crime and state prisoners in each state. I choose scatter plot for this data because of the readability and the ease to understand data while comparing. We can see that Alaska has the highest crime rate while Vermont has the lowest crime rate. We can also observe the relation between crime and the state prisoners. Idaho has the lowest difference between crime and state prisoners which concludes that most of the crimes committed are being noticed and action is being taken against the crime.

Correlation Matrix:

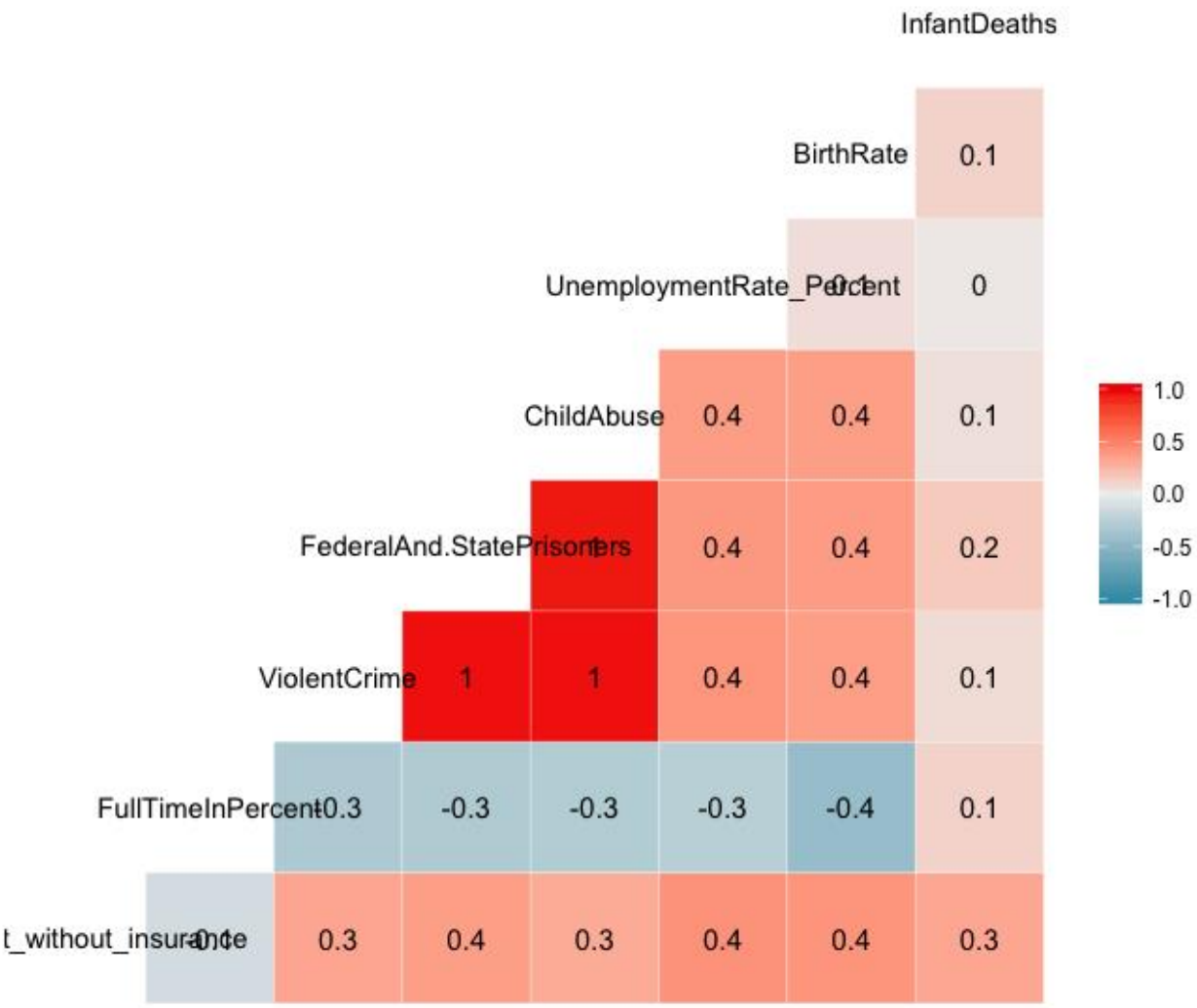

Figure 1.7 Correlation Matrix

The correlation matrix is used to determine the dependency between multiple variables at the same time. This can only be done with numeric data in the dataset. The correlation coefficient is indicated by the colour of the cell where red indicates that the variables are highly correlated. For this particular dataset, Violent crime and State prisoners have strong correlation.

Choropleth Map:

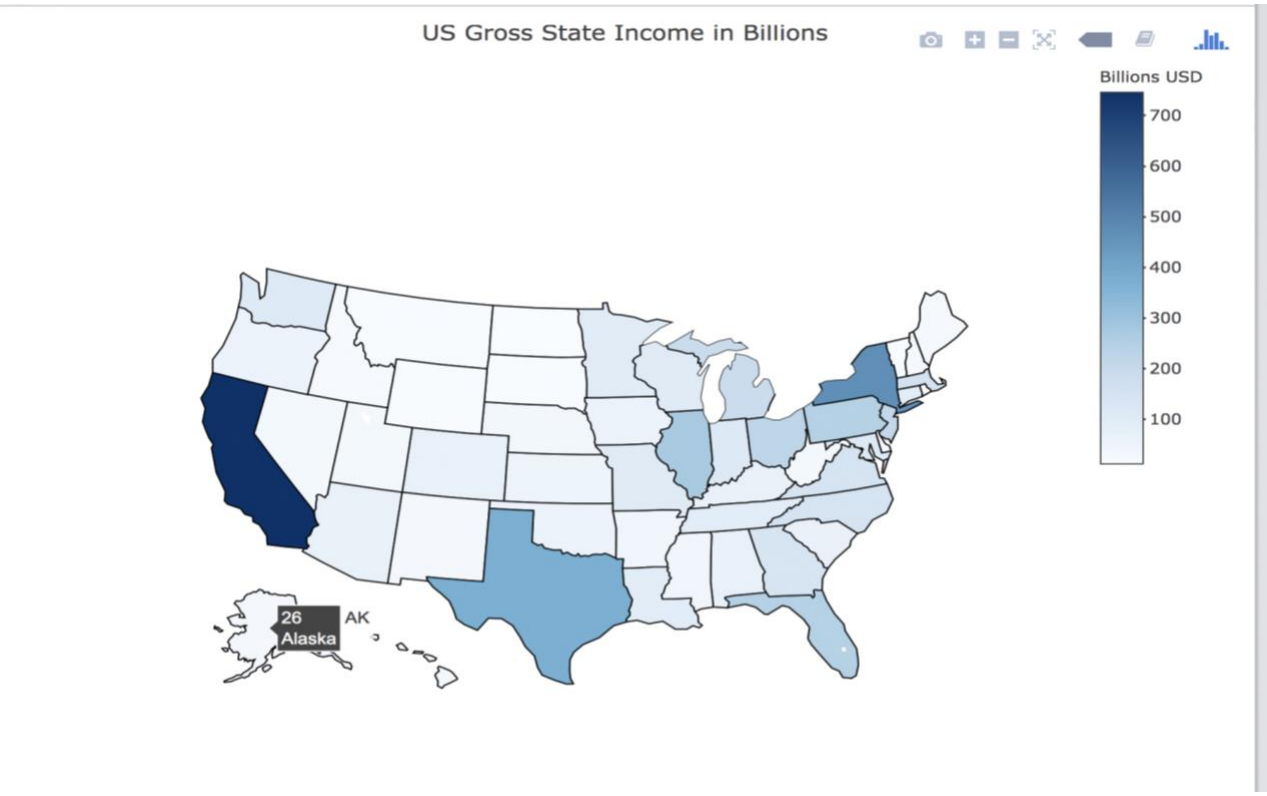

Figure 1.8 Gross income of each state (In Billions)

The above choropleth map shows the gross state income with colour variations representing the income from highest to the lowest, dark colour representing the highest income and light colour representing the least state income. We can conclude that California has the highest gross state income.

Micro Map:

The micro map gives insights about the rate of people died due to heart disease, the rate of people died due to cancer for each state. The states with highest death rate due to heart disease are sorted from top to bottom with West Virginia being the first and Alaska being the least.

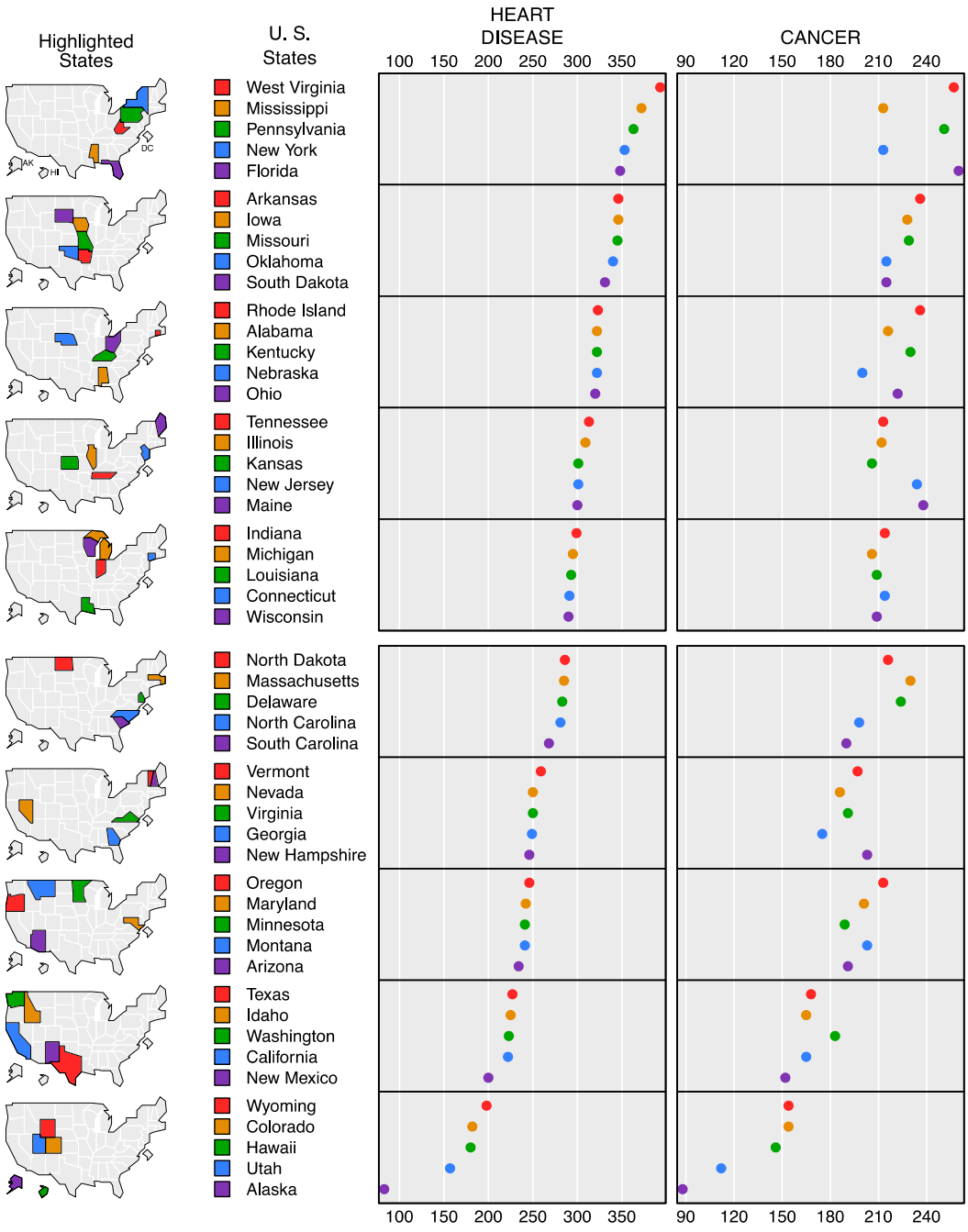

Figure 1.9 Comparing Death rates for each state (Per 10,000 people)

Conclusion:

From the data set, we were able to visualize various conclusions. Even though California is the most populated state, it is not ranked the highest with respect to crime, death rate or unemployment. California ranks first in population and gross state income while West Virginia ranks first in unemployment and death rate due to heart disease. These are the comprehensive statistics pertaining to each state in United States of America.