1. **Data Set Review:**

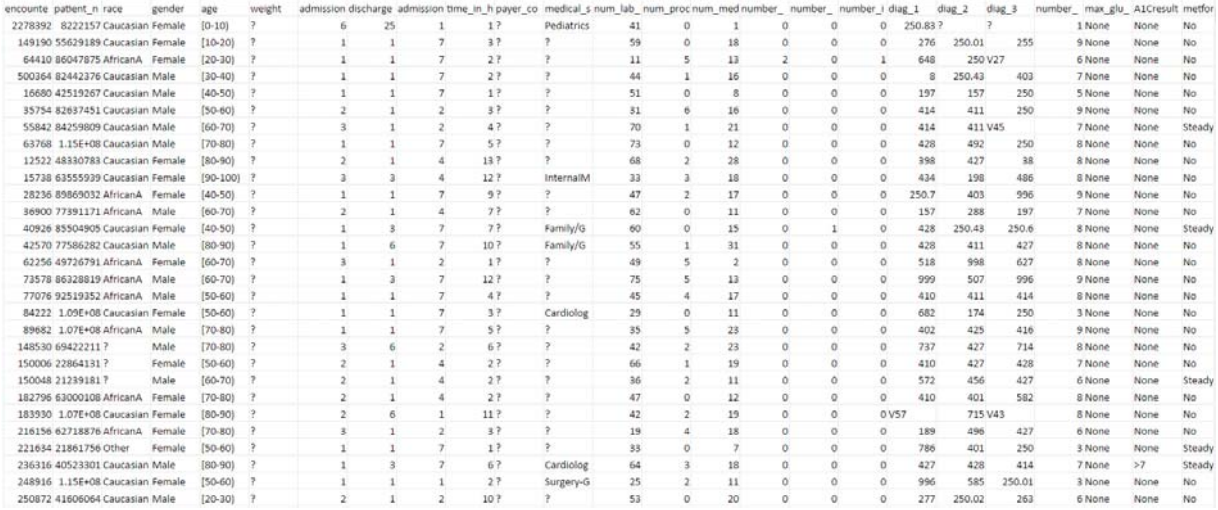
Present day in United States it is observed that nearly 30 million of Americans are suffering with diabetes. I had drawn co-relations between the factors that directly affect our cause. Also, I have done exploratory data analysis for various age categories and race of people as our first hand analysis. This analysis would be more of a kind of descriptive and less of a predictive analysis. The present analysis of a large clinical database was undertaken to examine historical patterns of diabetes care in patients with diabetes admitted to a US hospital and to inform future directions which might lead to improvements in patient safety.

2. **Data Set Sample:**

1. Diabetes data showing between years and states.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	States	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
2	AL	15.266	18.705	18.412	22.853	27.184	32.725	36.146	35.318	33.824	30.72	32.822	33.529	33.181	29.835	28.71
3	AK	1.34	1.693	2.109	2.016	1.912	1.612	1.896	2.146	2.649	3.131	3.673	3.727	3.291	2.549	2.287
4	AZ	14.842	15.199	17.254	17.396	19.531	20.79	21.129	22.16	25.23	30.866	39.837	46.17	48.033	46.386	45.579
5	AR	7.819	8.774	11.934	14.007	16.283	17.697	16.935	17.835	15.742	17.953	18.999	21.114	0	0	0
6	CA	130.516	160.836	158.254	171.509	192.167	201.714	220.641	209.276	209.33	189.559	186.795	197.197	197.792	210.345	208.537
7	CO	8.256	9.393	10.414	14.049	13.367	14.399	13.212	17.156	18.668	17.565	17.654	18.338	20.096	20.532	0
8	CT	0	0	13.345	12.81	13.854	15.761	17.546	15.159	15.22	15.029	17.684	17.379	15.636	14.462	13.383
9	DE	0	0	0	0	0	0	4.844	4.608	4.704	4.902	4.934	4.302	4.397	4.131	4.038
10	DC	1.578	1.578	2.446	2.583	2.651	2.885	0	0	0	0	3.473	3.678	2.928	2.622	2.477
11	FL	35.914	48.646	72.09	85.941	101.78	94.77	102.799	98.889	117.503	126.431	123.853	134.678	125.512	131.419	108.303
12	GA	27.963	38.417	33.004	30.587	39.751	52.556	58.348	51.244	54.285	54.52	61.54	59.82	62.161	62.561	61.563
13	HI	3.419	4.892	5.05	5.258	5.915	5.521	6.329	6.164	0	0	4.8	5.483	5.677	5.788	5.256
14	ID	2.379	3.241	4.005	5.15	5.909	6.433	7.14	7.077	7.197	7.297	8.33	8.625	9.197	7.61	0
15	IL	0	0	0	0	0	0	0	0	0	0	0	0	0	53.122	53.122
16	IN	22.153	23.053	0	0	0	27.659	37.158	35.37	39.513	37.861	39.888	33.931	33.247	33.286	36.164
17	IA	9.482	10.599	11.195	12.622	13.315	12.82	13.518	12.638	14.39	14.715	15.755	15.48	14.207	15.721	16.57
18	KS	8.931	6.728	6.884	10.096	12.403	13.79	12.912	13.841	0	0	16.596	16.765	16.851	16.86	16.739
19	KY	10.433	13.365	15.057	17.705	19.265	20.076	23.11	22.823	24.164	24.349	28.563	30.927	31.81	30.183	29.88
20	LA	0	0	17.79	20.322	20.322	27.47	29.427	29.371	25.705	26.915	26.963	28.837	31.679	32.648	32.648
21	ME	3.401	3.429	3.105	4.531	5.439	7.264	6.795	8.175	8.111	8.086	7.421	7.235	8.036	8.182	8.414
22	MD	0	0	0	0	0	26.902	27.282	25.212	26.462	22.903	28.976	28.976	39.645	41.821	41.821
23	MA	14.764	16.927	22.254	23.293	24.002	23.18	22.282	23.649	0	0	0	0	22.477	24.369	25.248
24	MI	38.551	45.537	45.33	46.167	50.604	53.072	0	0	0	0	69.87	67.3	66.623	67.457	70.105
25	MN	9.166	12.103	14.129	14.612	13.92	13.767	17.077	20.247	20.352	18.468	16.894	14.156	14.441	14.873	16.298
26	MS	10.448	10.93	13.054	13.368	0	15.621	0	18.61	0	21.623	0	20.946	0	20.957	0
27	MO	19.225	19.225	0	0	0	0	0	23.209	23.004	23.786	23.238	21.052	20.2	27.023	0

2. Diabetes data between races, ages, no of cases, etc.



encounte	patient_n	race	gender	age	weight	admission	discharge	admission	time_in_h	payer_co	medical_s	num_lab_	num_proc	num_med	number_	number_	number_i	diag_1	diag_2	diag_3	number_	max_glu_	A1Cresult	metfor
2278392	8222157	Caucasian	Female	[0-10)	?	6	25	1	1	?	Pediatrics	41	0	1	0	0	0	250.83	?	?	1	None	None	No
149190	55629189	Caucasian	Female	[10-20)	?	1	1	7	3	?	?	59	0	18	0	0	0	276	250.01	255	9	None	None	No
64410	86047875	AfricanA	Female	[20-30)	?	1	1	7	2	?	?	11	5	13	2	0	1	648	250	V27	6	None	None	No
500364	82442376	Caucasian	Male	[30-40)	?	1	1	7	2	?	?	44	1	16	0	0	0	8	250.43	403	7	None	None	No
16680	42519267	Caucasian	Male	[40-50)	?	1	1	7	1	?	?	51	0	8	0	0	0	197	157	250	5	None	None	No
35754	82637451	Caucasian	Male	[50-60)	?	2	1	2	3	?	?	31	6	16	0	0	0	414	411	250	9	None	None	No
55842	84259809	Caucasian	Male	[60-70)	?	3	1	2	4	?	?	70	1	21	0	0	0	414	411	V45	7	None	None	Steady
63768	1.15E+08	Caucasian	Male	[70-80)	?	1	1	7	5	?	?	73	0	12	0	0	0	428	492	250	8	None	None	No
12522	48330783	Caucasian	Female	[80-90)	?	2	1	4	13	?	?	68	2	28	0	0	0	398	427	38	8	None	None	No
15738	63555939	Caucasian	Female	[90-100)	?	3	3	4	12	?	InternalM	33	3	18	0	0	0	434	198	486	8	None	None	No
28236	89869032	AfricanA	Female	[40-50)	?	1	1	7	9	?	?	47	2	17	0	0	0	250.7	403	996	9	None	None	No
36900	77391171	AfricanA	Male	[60-70)	?	2	1	4	7	?	?	62	0	11	0	0	0	157	288	197	7	None	None	No
40926	85504905	Caucasian	Female	[40-50)	?	1	3	7	7	?	Family/G	60	0	15	0	1	0	428	250.43	250.6	8	None	None	Steady
42570	77586282	Caucasian	Male	[80-90)	?	1	6	7	10	?	Family/G	55	1	31	0	0	0	428	411	427	8	None	None	No
62256	49726791	AfricanA	Female	[60-70)	?	3	1	2	1	?	?	49	5	2	0	0	0	518	998	627	8	None	None	No
73578	86328819	AfricanA	Male	[60-70)	?	1	3	7	12	?	?	75	5	13	0	0	0	999	507	996	9	None	None	No
77076	92519352	AfricanA	Male	[50-60)	?	1	1	7	4	?	?	45	4	17	0	0	0	410	411	414	8	None	None	No
84222	1.09E+08	Caucasian	Female	[50-60)	?	1	1	7	3	?	Cardiolog	29	0	11	0	0	0	682	174	250	3	None	None	No
89682	1.07E+08	AfricanA	Male	[70-80)	?	1	1	7	5	?	?	35	5	23	0	0	0	402	425	416	9	None	None	No
148530	69422211	?	Male	[70-80)	?	3	6	2	6	?	?	42	2	23	0	0	0	737	427	714	8	None	None	No
150006	22864131	?	Female	[50-60)	?	2	1	4	2	?	?	66	1	19	0	0	0	410	427	428	7	None	None	No
150048	21239181	?	Male	[60-70)	?	2	1	4	2	?	?	36	2	11	0	0	0	572	456	427	6	None	None	Steady
182796	63000108	AfricanA	Female	[70-80)	?	2	1	4	2	?	?	47	0	12	0	0	0	410	401	582	8	None	None	No
183930	1.07E+08	Caucasian	Female	[80-90)	?	2	6	1	11	?	?	42	2	19	0	0	0	V57	715	V43	8	None	None	No
216156	62718876	AfricanA	Female	[70-80)	?	3	1	2	3	?	?	19	4	18	0	0	0	189	496	427	6	None	None	No
221634	21861756	Other	Female	[50-60)	?	1	1	7	1	?	?	33	0	7	0	0	0	786	401	250	3	None	None	Steady
236316	40523301	Caucasian	Male	[80-90)	?	1	3	7	6	?	Cardiolog	64	3	18	0	0	0	427	428	414	7	None	>7	Steady
248916	1.15E+08	Caucasian	Female	[50-60)	?	1	1	1	2	?	Surgery-G	25	2	11	0	0	0	996	585	250.01	3	None	None	No
250872	41606064	Caucasian	Male	[20-30)	?	2	1	2	10	?	?	53	0	20	0	0	0	277	250.02	263	6	None	None	No

3. **Data Sources:**

This data is being collected from different sources in internet and has been further developed into one set. Among many helpful websites these few need a special mention.

- http://www.cdc.gov
- http://www.diabetes.org/
- http://www.idf.org/
- http://stateofobesity.org/
- https://en.wikipedia.org/

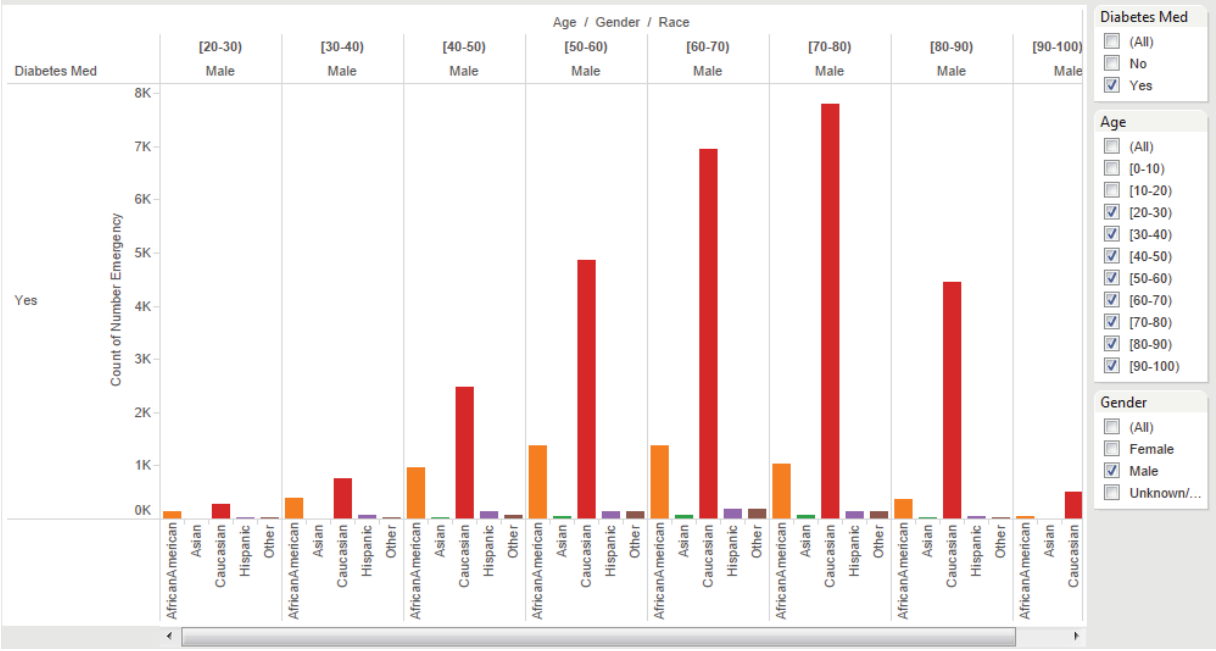
4. **Data Cleansing**

- There wasn't much need for cleaning except for formatting the data.
- Few columns were created in order to calculate the percentage change over a period of time.
- Excel has been used for most of the cleaning.
- Some of the data pre-processing tasks such as breaking down the dataset for a particular analysis has been done using R.

5. Visual Representation:

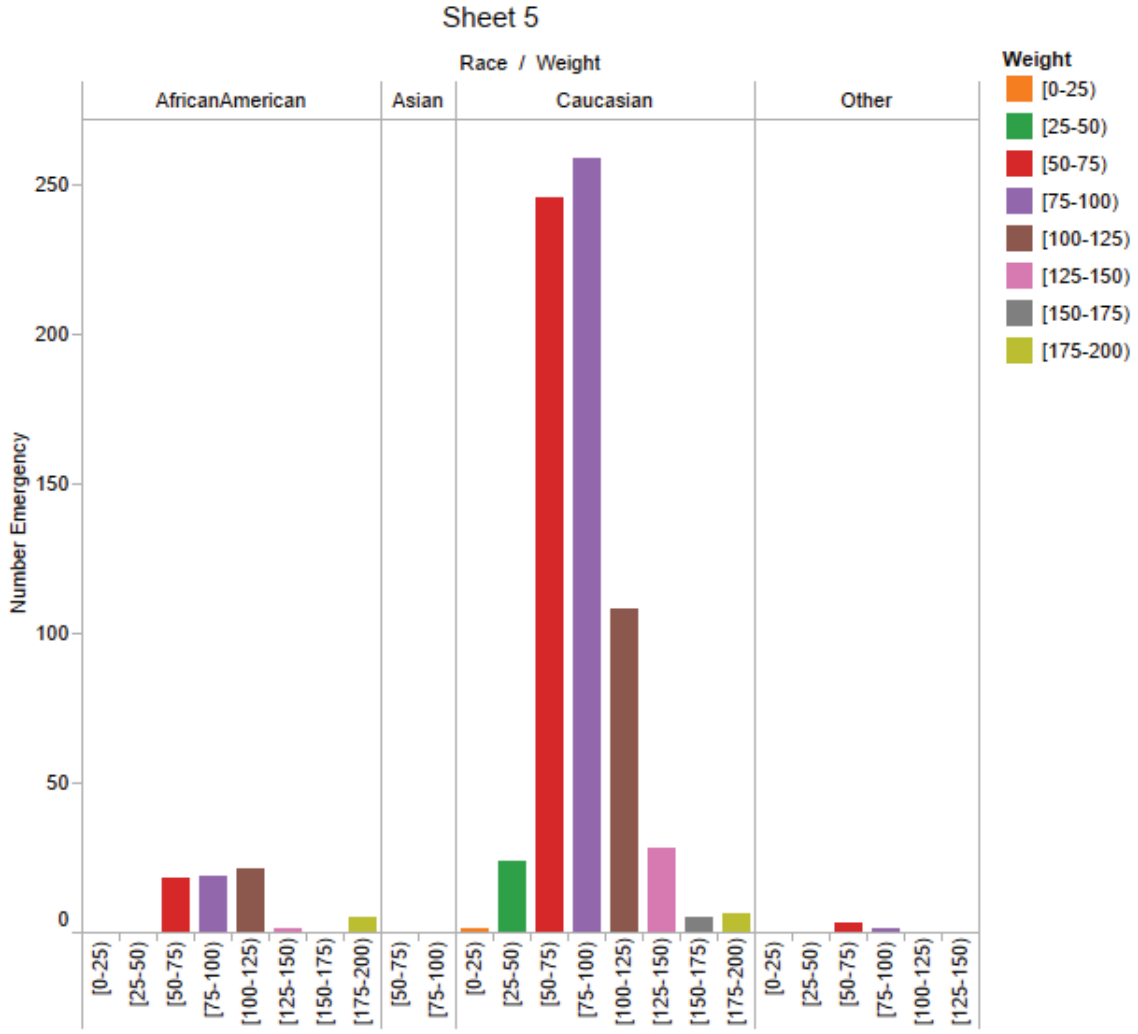
Tableau:

G1: Diabetes in males for different races in different age groups



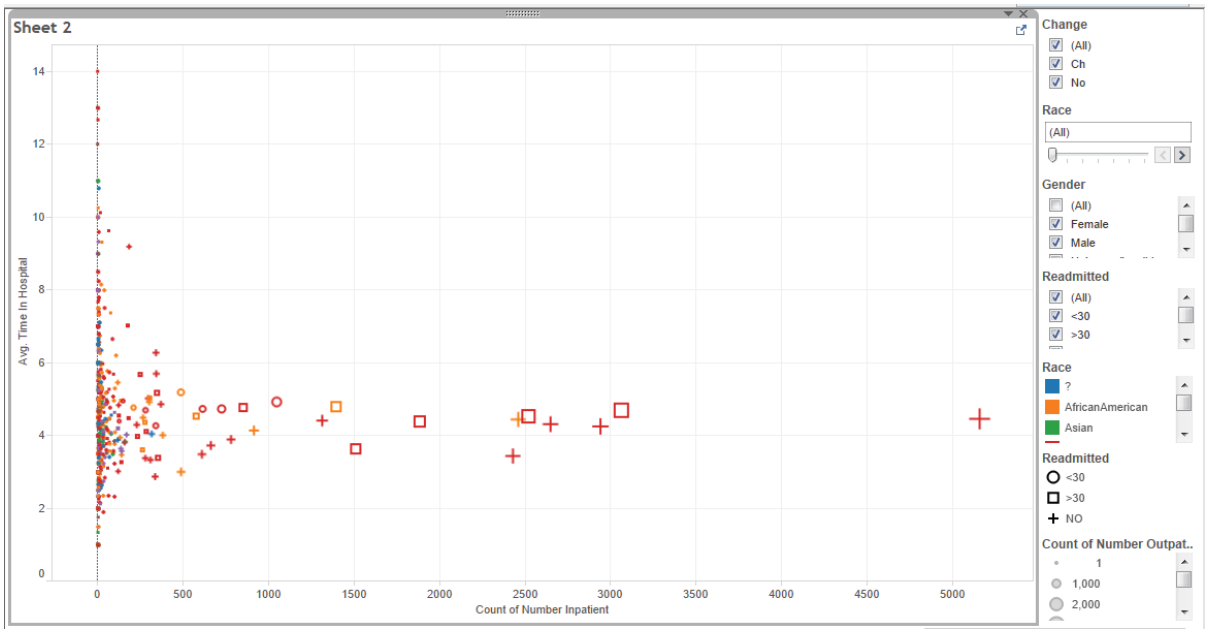
G2: Diabetes is females for different races in different age groups.

G3: Diabetes and Overweight for different races in different age groups:



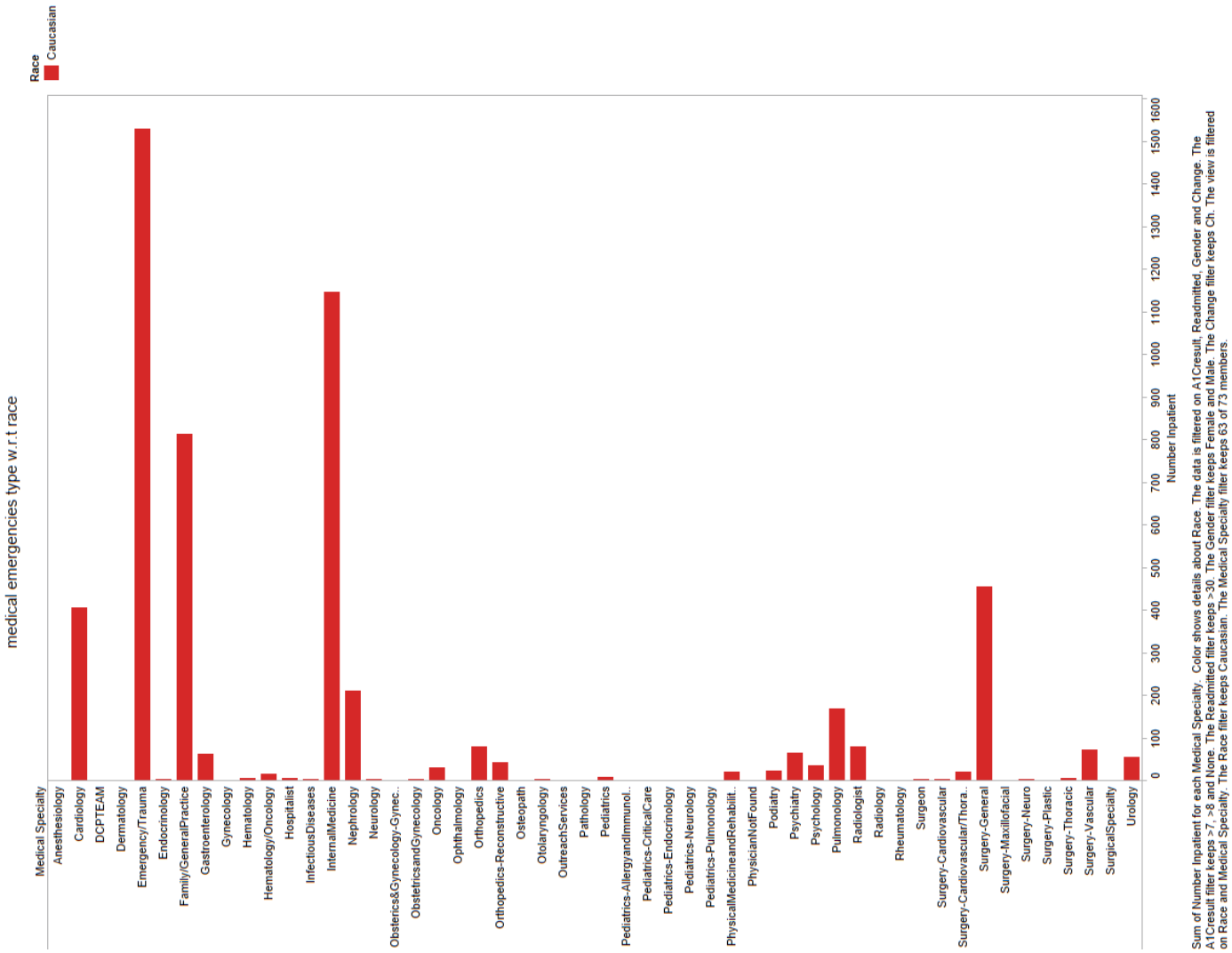
Sum of Number Emergency for each Weight broken down by Race. Color shows details about Weight. The view is filtered on Weight and Race. The Weight filter excludes ? and >200. The Race filter keeps AfricanAmerican, Asian, Caucasian, Hispanic and Other.

G4: In Patient- Out Patient Details for different races in different age groups

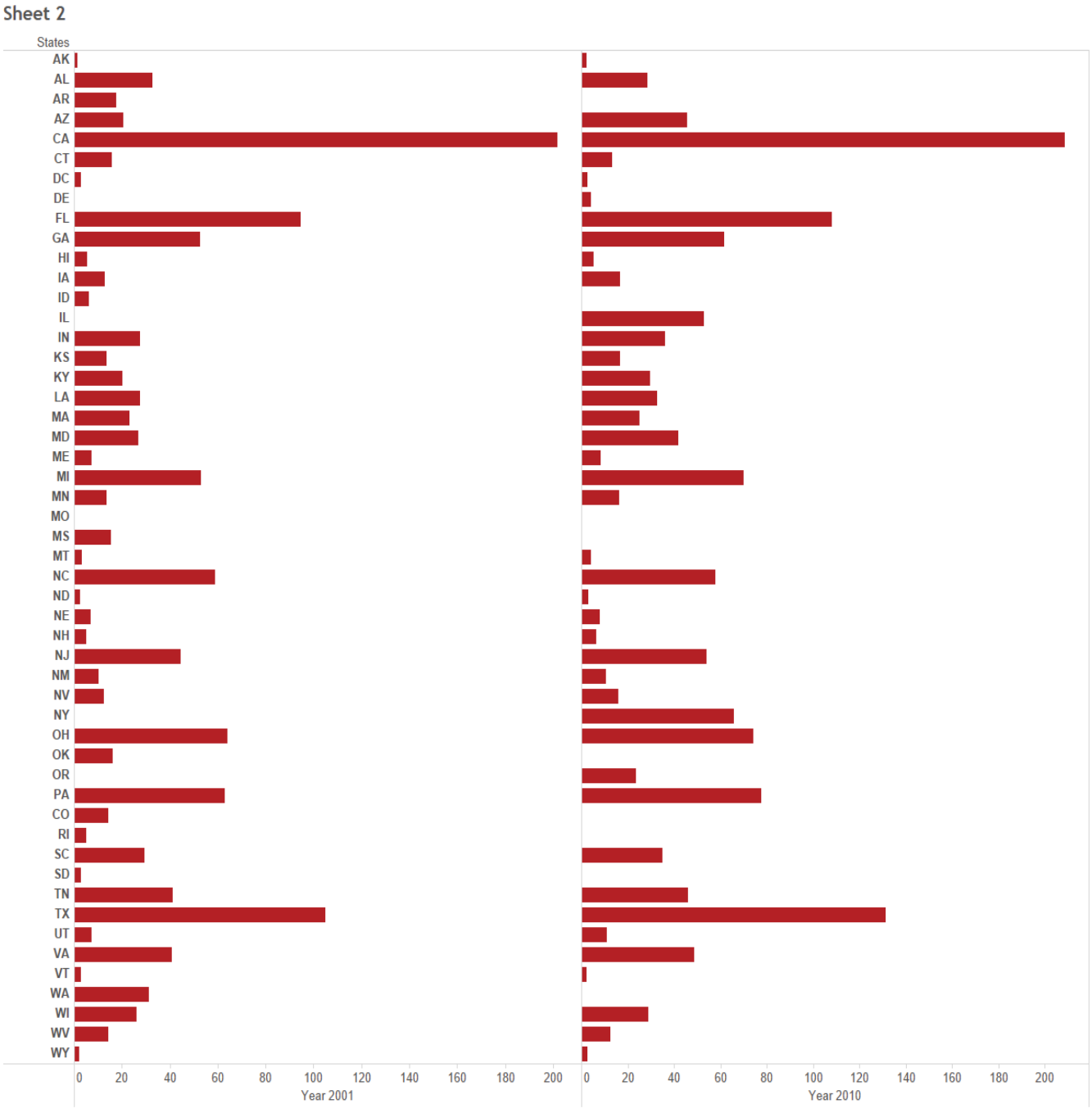


G5: Emergency Patients over male and female.

G6: Different health issues in Caucasians.



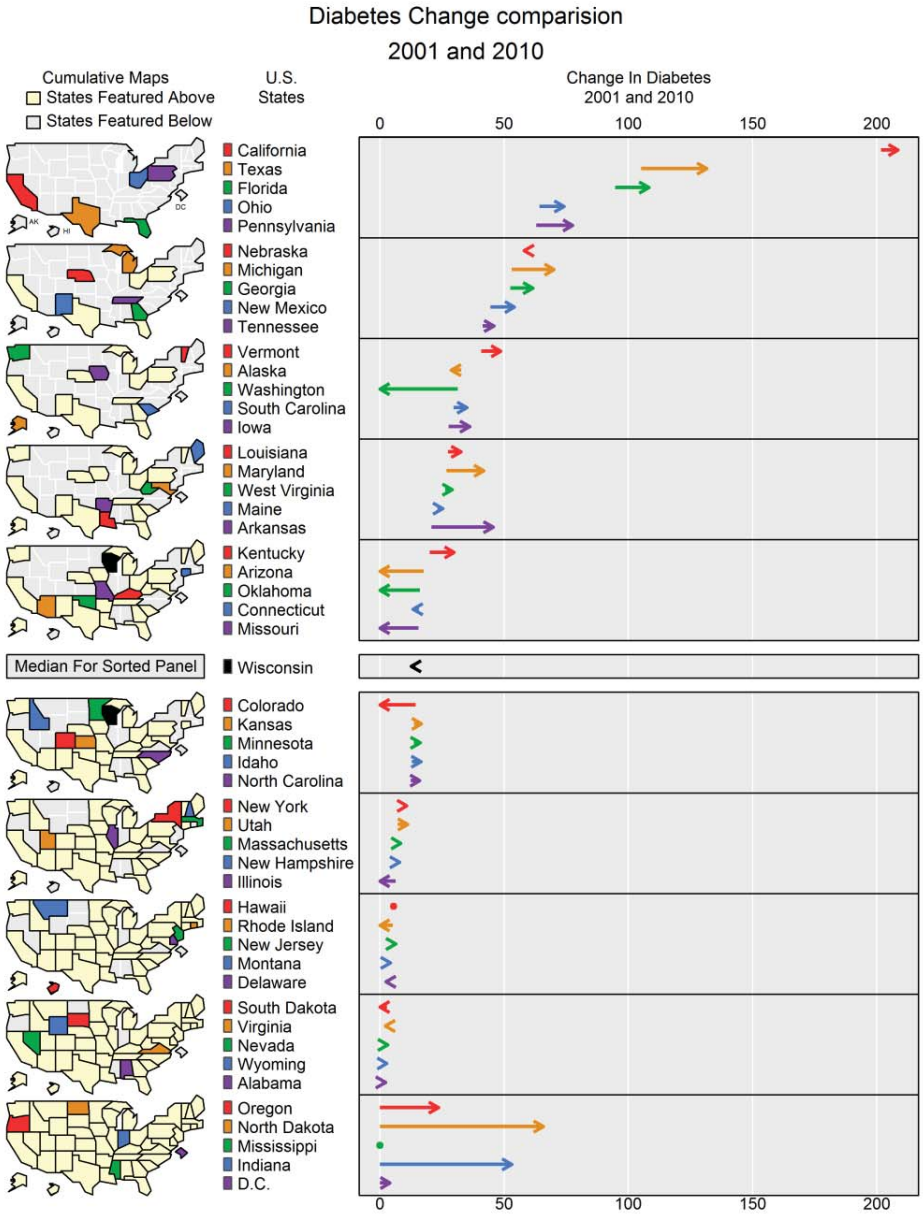
G7: Diabetes 2001 VS 2010 over all states



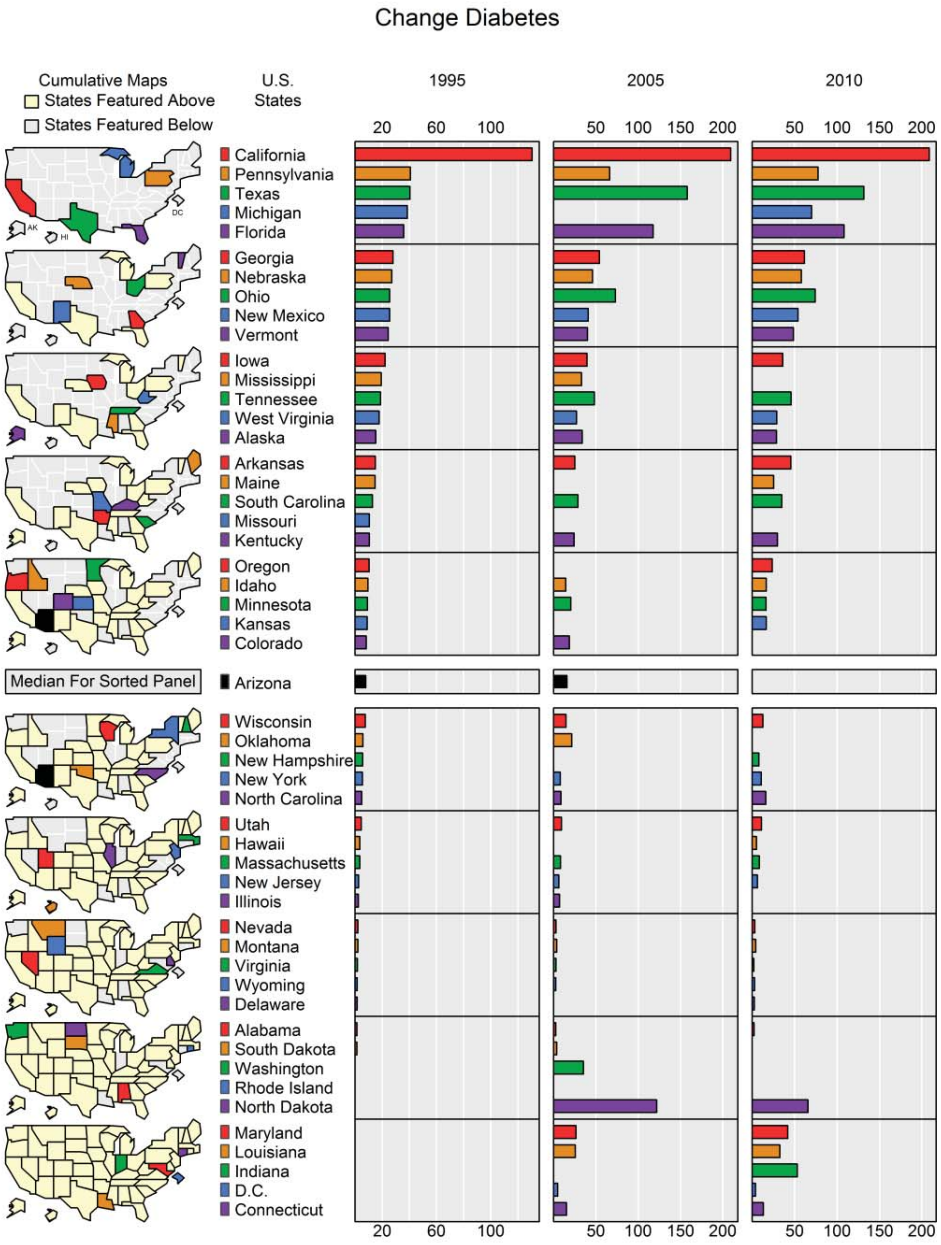
Year 2001 and Year 2010 for each States. The data is filtered on Year 2003, which ranges from 0 to 209.276. The view is filtered on Year 2001 and States. The Year 2001 filter ranges from 0.0 to 201.7. The States filter keeps 51 of 51 members.

Micro Maps:

G8: Diabetes 2001 VS 2010 over all states.

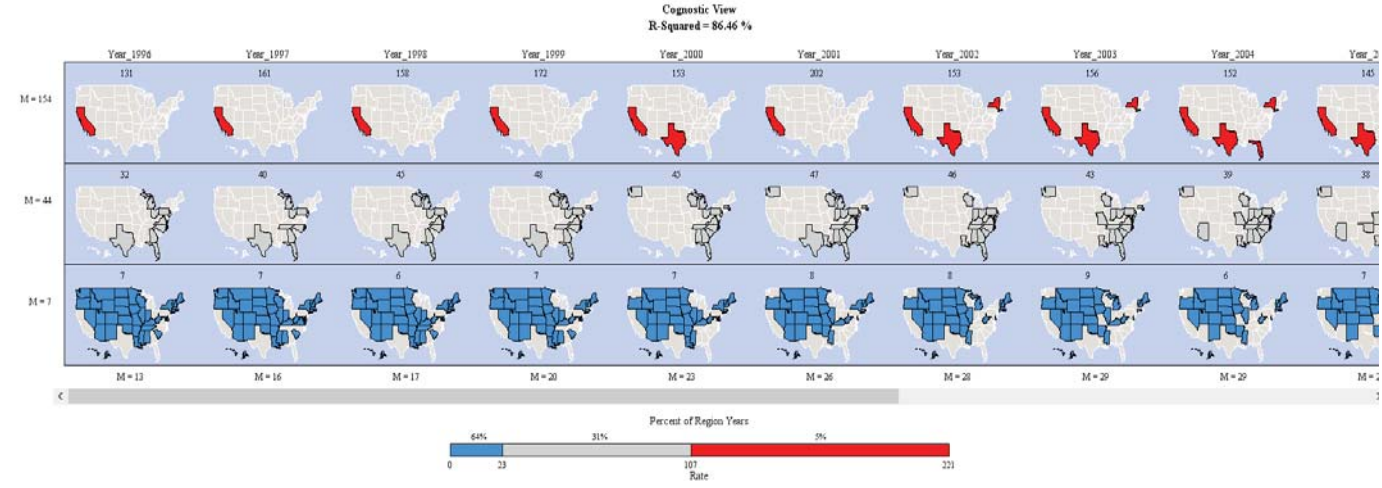


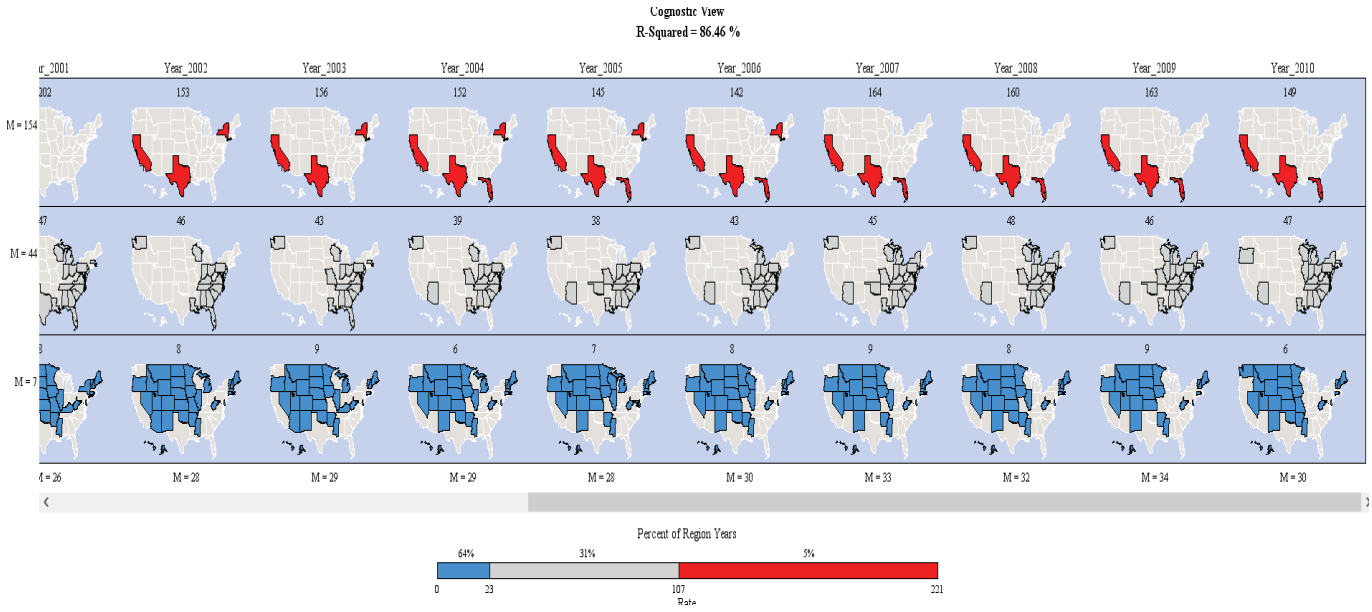
G9: Diabetes 1995 – 2005 – 2010 over all states.



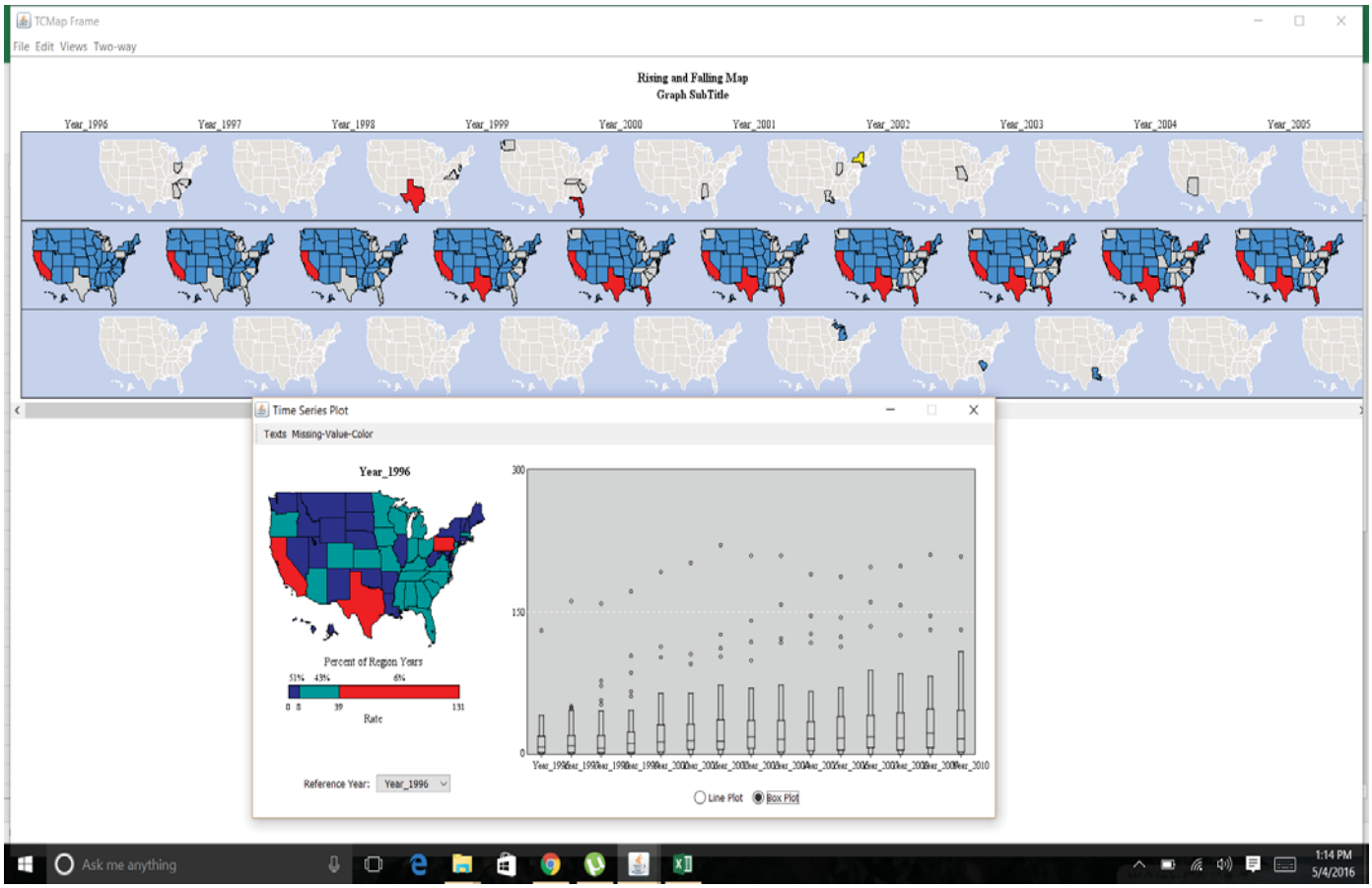
TC Maps:

G10: 1996 – 2006 yearly change over all states.





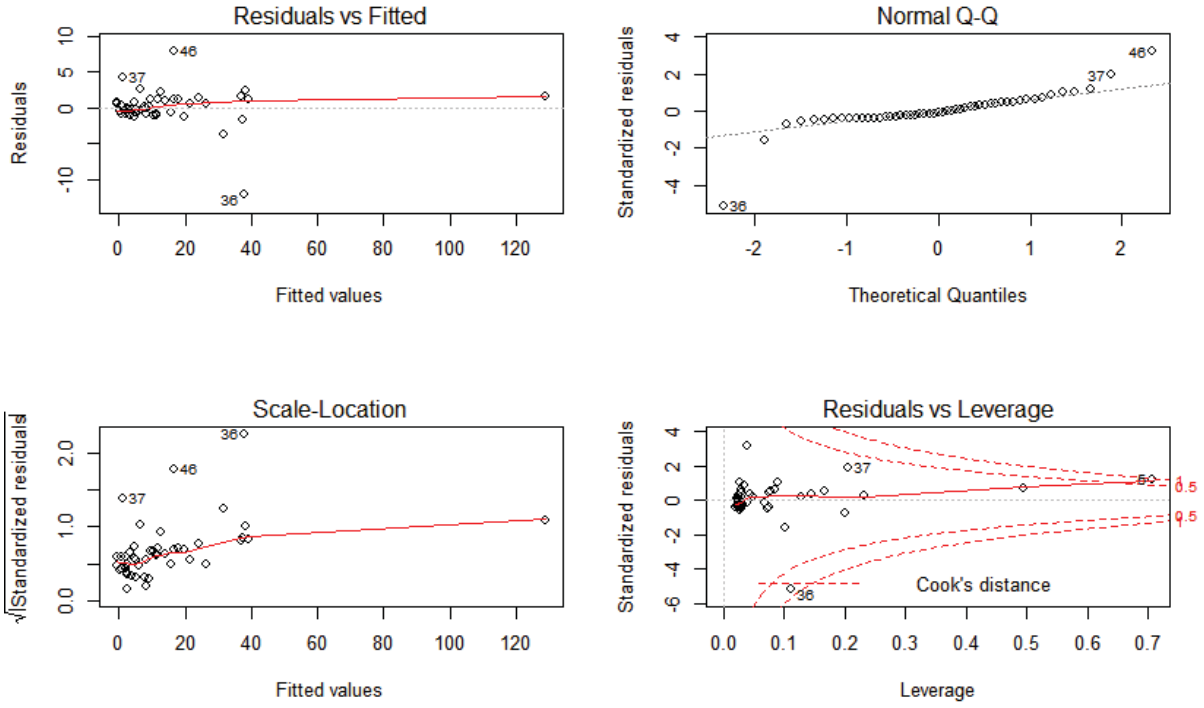
G11: Diabetes TC maps time series box plot.



Linear Regression

G12: Linear Regression of Patients to several factors – Year 1996-2000

Residual standard error: 2.516 on 47 degrees of freedom
Multiple R-squared: 0.9856, Adjusted R-squared: 0.9847
F-statistic: 1072 on 3 and 47 DF, p-value: < 2.2e-16



Conclusion:

1. Diabetes in USA is increasing at an alarming rate.
2. Type 2 diabetes has more effect then type 1.
3. It can be mostly seen in Caucasian people and people living around California.

References:

- Strack, B., DeShazo, J., Gennings, C., Olmo, J., Ventura, S., Cios, K. and Clore, J. (2014). Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. BioMed Research International, 2014, pp.1-11
- Archive.ics.uci.edu. (2016). UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999-2008 Data Set. [online] Available at: http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008 [Accessed 2 May 2016].
- https://public.tableau.com/profile/ashfaq93/•https://weka.wikispaces.com/What+do+those+numbers+mean+in+a+J48+tree%3F•https://www.rstudio.com/
- http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#
- http://www.healthline.com/health/type-2-diabetes/statistics
- http://www.cdc.gov/diabetes/data/statistics/2014statisticsreport.html
- http://www.idf.org/membership/nac/united-states
- http://www.diabetes.org/