Data Discovery using Tableau

Jugal.Patel

RIT

MGIS 650 - Introduction to Data Analytics

and Business Intelligence

Professor: Ali Tosyali

04/17/2022

Table of Contents:

I. Acknowledgement

I would like to take this opportunity to thank Prof. Ali Tosyali for guiding us throughout the whole duration of the course. His teaching style and commitment towards this course really helped me tremendously to understand the core fundamentals of business intelligence.

II. Introduction

In the present world, there is a huge demand for cleaner and greener sources of energies. Energy production forms a major part of an economy of a state. The geographical location of the states affects the energy production and usage. As non renewable forms of energies are depleting, there is a huge incentive for states to produce more green energies. Every state is looking to cut down its dependence on non renewable forms of energies and looking to produce more cleaner and renewable energies. As energy policies in the United States are decentralized to the State level, these states are responsible for producing more cleaner forms of energies.

III. Objectives

Our main goal here is to:

i) Perform data analysis on the energy data of the four states. California(CA), Texas(TX), New Mexico(NM), Arizona(AZ).

ii) From the data analysis, figure out the development needs of these states and suggest a plan to each state on how they could increase their production of cleaner and greener energies.

iii) To also create an energy profile for each state.

iv) To extrapolate from the analysis, which state has the best profile for cleaner, greener renewable forms of energies.

IV. Dataset Description :

The dataset “ProblemCDataWide.csv” consisted of 605 variables and 200 rows.

For each state the time frame was 1960 - 2009.

Each column in the dataset represented some information regarding energy production, consumption, demographic or economic information.

Another dataset description file “msscodes” contained information regarding the column names and their measuring unit.

V. Analysis:

i) **Exploratory Data Analysis(EDA):** To begin the process, we first remove irrelevant columns from the dataset. This can be achieved via following a two step procedure.

- Removal of columns that consist of NA values. Attributes either have ‘forty’ or ‘zero’ NA values. And NA values are those that have a ‘blank’ cell in the excel worksheet. So we have removed all columns that have ‘forty’ missing values. The function to check for NA values in excel is ‘=COUNTBLANK’. As missing values are empty cells, this function will count the number of empty cells, which is the number of NA values.

  Here is a screenshot depicting the columns that have no missing values in them.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
175	TX	1983	38.81881	77779.9732	762.28418	15.679	14.8558	15.68112	12.65946	8951.8677	344.29987	54.513	-19638.128	8.17014	1.79736	7.15826	8167900.52	2520.4698	0	0
176	TX	1984	4.65957	71200.3982	962.63304	15.95326	14.66297	15.95326	12.52324	7085.20316	272.50983	53.424	28018.3849	4.12288	5.30975	8.57686	8639364.47	3170.15056	0	0
177	TX	1985	207.94202	78354.8307	1317.12418	15.57497	14.80734	15.57653	12.64291	6777.61951	27.45067	5.394	60741.8514	7.6803	18.48613	9.22877	8662150.09	3424.52813	0	0
178	TX	1986	28.74489	90548.4586	1539.40854	15.90663	14.58271	15.90799	12.66701	4010.19781	15.29324	0	95559.5259	7.83558	17.05267	9.78918	8499342.75	0	0	0
179	TX	1987	2.33894	83620.0417	1150.16591	15.15325	14.48383	15.15529	12.66701	5124.95152	51.28411	0.356	109715.478	12.38018	22.77244	10.73926	8718342.61	0	0	0
180	TX	1988	-55.03351	95781.0095	1012.67885	14.06793	14.60802	14.0694	12.58779	3545.80423	20.16648	0	109864.97	9.71261	16.40931	10.8112	9257462.5	0	0	0
181	TX	1989	5.3761	64248.8443	819.82112	14.56476	14.61001	14.56476	12.53908	2754.85728	9.07132	1.33	-12651.739	7.1792	13.89483	10.22532	9585112.03	0	179.1	0
182	TX	1990	12.1767	92991.4225	837.89486	14.78802	14.57822	14.78967	12.67546	2224.52538	1.77751	1.121	27947.4915	6.52258	12.33019	8.5944	9617721.55	0	198.39969	0
183	TX	1991	-4.17681	62183.0244	655.43535	15.05199	14.45537	15.05322	12.5233	2222.88264	2.48277	0.014	14853.9474	4.75632	13.53133	10.08427	9589263.49	0	215.15803	0
184	TX	1992	7.82705	78306.851	783.4834	14.30566	14.46625	14.31012	12.44741	2339.00588	1.69868	0	-21377.435	4.74296	14.22844	9.00667	9706242.98	0	231.93558	0
185	TX	1993	7.42308	84504.2766	693.01064	15.18034	14.7574	15.18809	12.51612	1956.50557	2.7655	0.014	9522.81338	0.11486	2.48199	0	9915385.79	0	246.68333	0
186	TX	1994	313.34416	72642.514	773.07663	15.47725	14.76697	15.48368	12.60926	2214.30835	5.35818	0.07	-13362.817	0.27223	6.36453	0	10039254.9	0	227.457	0
187	TX	1995	272.05164	78262.5188	644.76546	14.9602	14.72568	14.96538	12.61315	2669.33325	6.22879	0	-25655.259	0.93283	22.45895	0	10160405.3	0	242.99145	0
188	TX	1996	357.6387	79380.4479	624.97312	15.3402	14.98921	15.3402	12.86182	2679.58936	0.06902	5.566	45138.6404	0.32646	8.07416	0	10951761.6	0	257.12929	0
189	TX	1997	468.18762	69738.8638	657.61525	15.54805	15.01066	15.55204	12.89968	2410.56577	0.18262	526.185	44362.5621	0.77356	20.12291	0	11279859.5	0	274.87422	0
190	TX	1998	205.21045	74329.5184	554.54126	14.2223	15.057	14.23099	12.87519	3072.08749	0.26128	738.369	39110.2132	1.09312	33.1719	0	11461644.7	0	284.769	0
191	TX	1999	329.74287	55993.4709	795.73055	14.2223	15.01573	14.22843	12.66183	2870.65831	2.11864	204.117	7530.71092	0.92371	14.03892	0	11194641.7	0	343.5	0
192	TX	2000	197.5877	52804.2621	608.6949	0	15.19314	16.28	12.77547	5657.24346	2.72023	2.388	18224.5305	1.04404	16.11391	0	11542905.2	0	343.5	0
193	TX	2001	319.76578	77777.1584	467.9077	0	15.33008	17.00044	12.84791	3627.03552	0.91923	3.604	63988.0353	1.08706	28.55661	0	11314207.9	0	383.5	0
194	TX	2002	391.02382	87617.128	532.9662	0	15.44303	17.70065	12.84434	2315.88074	3.87707	80.264	31467.6258	0.45741	12.84592	0	11582320.6	0	416.5	0
195	TX	2003	394.53016	92343.78	511.36974	0	15.2467	17.54537	12.51624	2625.84625	0.21804	79.58376	91441.3344	0.36864	10.91662	0	11416320.5	0	535.79572	0
196	TX	2004	572.25097	86174.8261	483.9473	0	15.27875	17.09972	12.48258	1796.48832	145.12374	78.99675	32066.2527	0.4293	14.52405	0	11478409	0	579.34486	0
197	TX	2005	425.9147	114067.504	510.84561	0	15.38507	17.16594	12.96527	2717.0287	4.76478	78.09796	80524.0294	0.25901	8.30827	0	10926534.6	0	660.43218	0
198	TX	2006	32.40434	104457.807	494.35035	0	15.44616	17.29	13.03012	2419.60344	0.31846	79.52296	117898.228	7.10287	231.37797	0	11249998.5	0	762.96185	0
199	TX	2007	92.40479	100464.913	492.04553	0	15.24276	21.64758	13.224	2441.43332	0.26083	159.85336	60715.2705	19.78302	243.72728	0	11332934.2	0	920.21051	0
200	TX	2008	5.14593	71106.8615	418.34594	0	15.38326	21.58698	13.212	2241.45205	0.26144	960.99372	66610.3567	23.03075	246.83303	4495.0813	10922957.7	0	1091.41766	0
201	TX	2009	-43.61959	61348.0105	347.27032	0	15.51691	20.48223	12.98	3477.55121	1.69734	447.30832	100588.547	20.73308	253.9078	3984.52016	10405746.3	0	1358.13442	0
202																				
203		NA values	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 1.1 Columns having no missing values

- Constructing the correlation matrix. The next step to remove excess columns is via correlation analysis. Here we create a correlation matrix for a group of columns having the same attributes. Then remove the columns that have a high

  correlation among them. Correlation is a statistical measure between two or more variables of their size and direction.

  In simple terms, if the correlation between two variables is positive then both of them will increase together and vice versa.

  So the strategy here is to remove columns that have the highest correlation i.e tending towards 'one'. The closer the correlation to 'one' the higher its chance of removal.

After following these two steps, the columns are reduced from 605 to 71.

Therefore, our working column number is 71.

A screenshot of a correlation matrix after the removal of highly correlated values.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		*ICP*	*NGACD*	*NGACV*	*NGEID*	*NGICV*	*NGISB*	*NGLPP*	*NGMPK*	*NGPZP*	*NGRCP*	*NGRCV*	*NGTCK*
2	ICP	1											
3	NGACD	-0.1436012	1										
4	NGACV	-0.059144	0.56436108	1									
5	NGEID	-0.0514229	0.70258059	0.51122972	1								
6	NGICV	0.11283487	0.47416693	0.49234485	0.46058089	1							
7	NGISB	0.19849886	0.10630752	0.10697898	-0.0078184	0.74026161	1						
8	NGLPP	0.2266092	-0.1108103	-0.0762682	-0.2029719	0.31992581	0.72984842	1					
9	NGMPK	0.03574853	-0.3117348	-0.0471158	-0.2041045	0.19719347	0.33674573	0.16822562	1				
10	NGPZP	0.30350668	-0.0644153	-0.1694064	-0.1794051	0.39139285	0.71320129	0.76111713	0.17842834	1			
11	NGRCP	0.04129738	0.02414853	0.25784137	0.06611375	0.38028044	0.3943505	0.12248029	0.17002317	-0.1077409	1		
12	NGRCV	-0.016954	0.43703088	0.65874678	0.49410564	0.66873103	0.35793871	-0.0183784	0.08162767	-0.1552699	0.73900342	1	
13	NGTCK	-0.0490505	-0.5667579	-0.1927549	-0.4238309	-0.2036241	-0.1165041	-0.1728998	0.1609756	-0.1997748	-0.0713336	-0.2390751	1

Figure 1.2 Correlation matrix after the deletion of highly related columns

In the correlation matrix above, the values are color coded to indicate the strength of the correlation.

ii) **Data Visualization**: In the step we export our excel data consisting of 71 columns to Tableau and create plots that depict a certain behavior and provide a better understanding of our data.

1) **Aviation Gasoline (AVTXP) :** Aviation gasoline total end use consumption.

   Unit of Measurement: Thousand Barrels

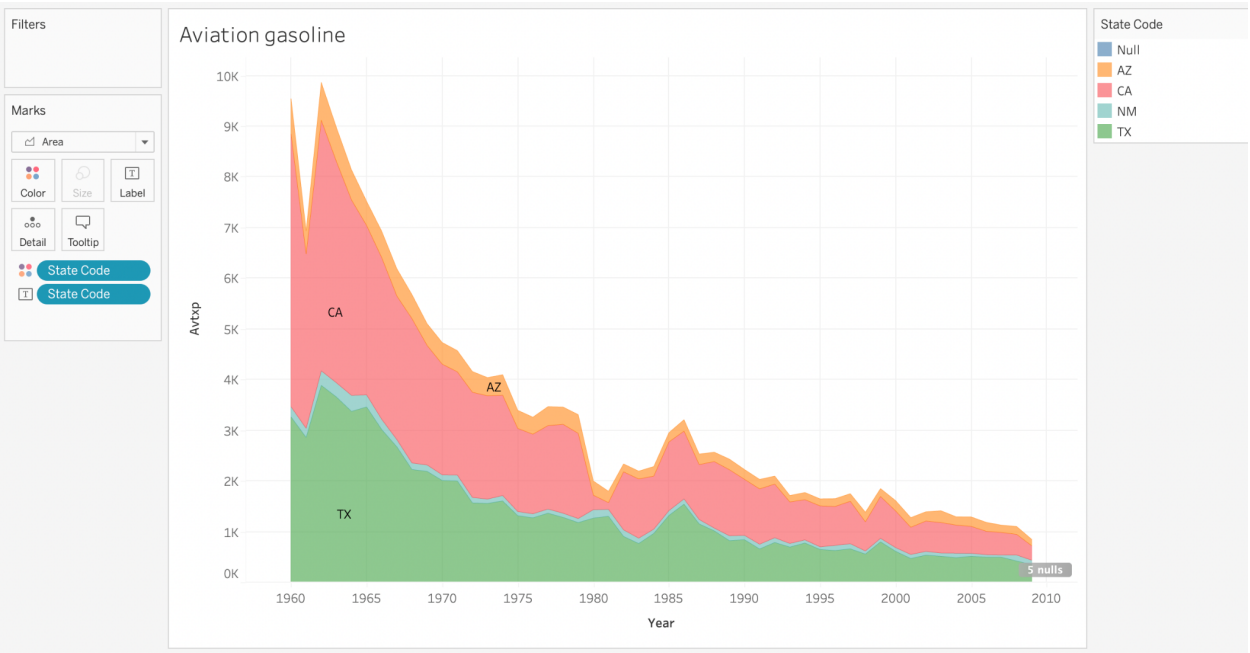


Figure 2.1

The area chart above in Figure 2.1 shows that in all four states aviation gasoline usage has decreased over time. The highest usage was in CA then TX followed by AZ and NM.

2) **Coal Production (CLPRK):** Factor for converting coal from physical units to Btu.

   Unit of Measurement: Million Btu per short ton

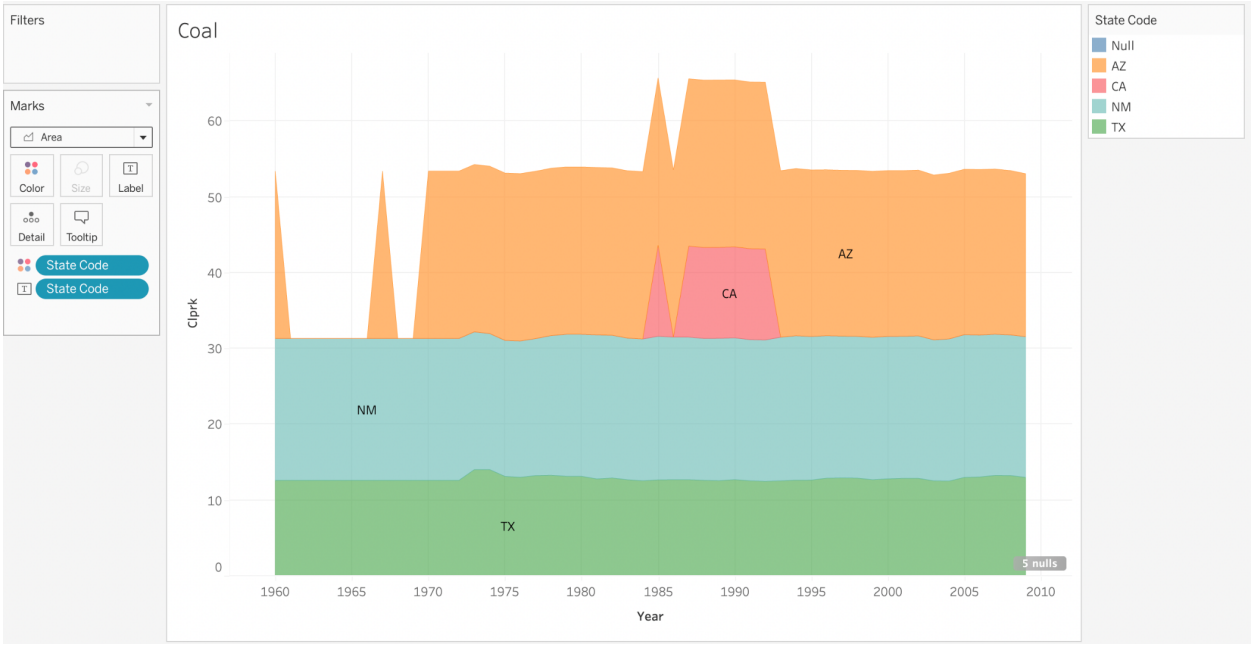


Figure 2.2

California(CA) has a lot of values filled with '0', therefore there is a small area covered by it.

Something similar is the case with Arizona(AZ) but at the same time it has the largest area hence it is the highest producer of coal among the four states.

3) **Distillate fuel oil consumed by the commercial and residential sector (DFCCP) (DFRCP) :** Distillate fuel oil is one of the general classification for petroleum fractions produced in conventional distillation operations. It has two types: diesel fuels and fuel oils.

Diesel fuels are used in on highway diesel engines as well as off highway engines.

Fuel oils are primarily used for space heating and electric power generation.

Unit of Measurement: Thousand barrels

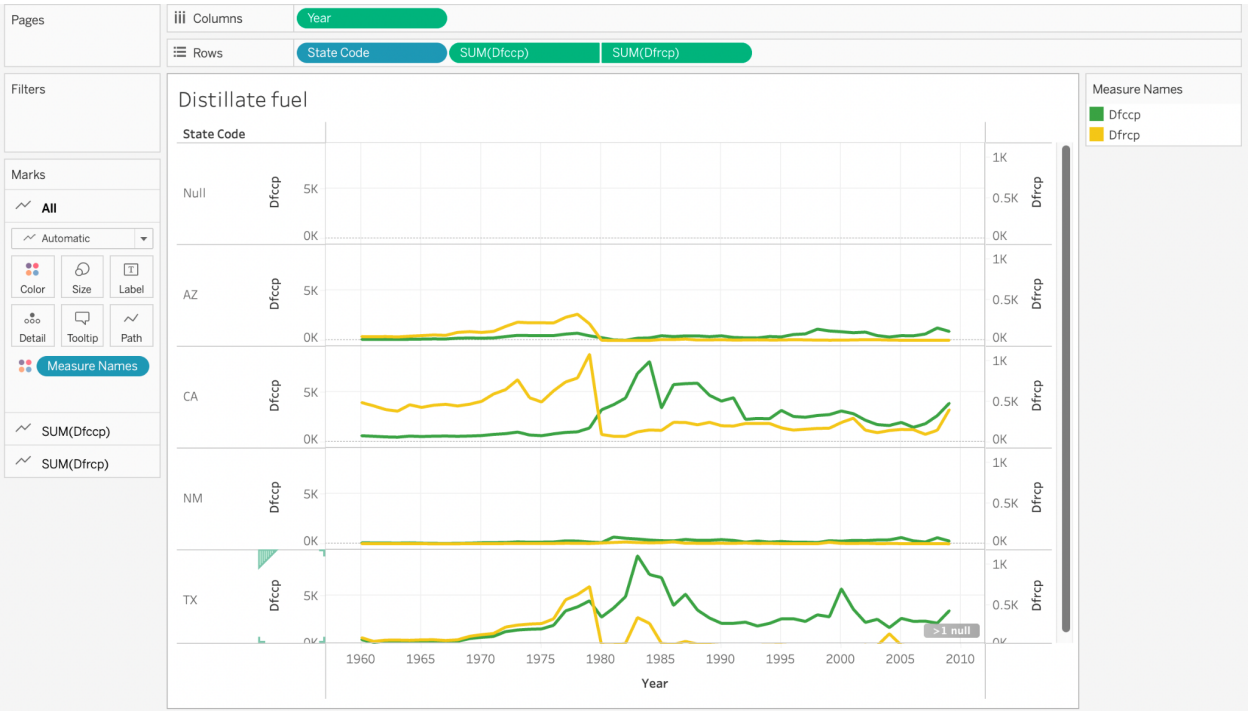


Figure 2.3

The figure above shows that distillate fuel for commercial purposes is the highest in Texas(TX) and for residential purposes the highest in the state of California(CA).

4) **Fossil Fuels(FFTCB) :** Consumption of fossil fuels by the four states.

Unit of Measurement: Billion Btu

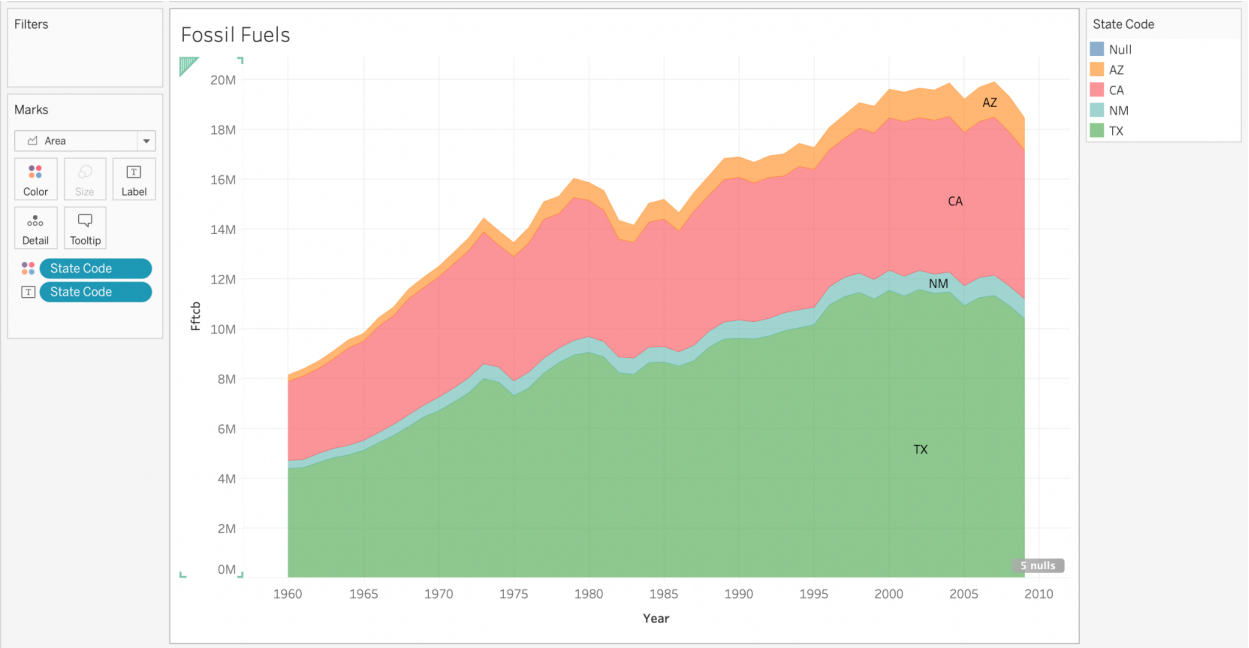


Figure 2.4

The consumption of fossil fuels is highest in the state of Texas(TX) followed by California(CA) then Arizona(AZ) and finally the lowest in New Mexico(NM).

Facts:

- The percent increase in the consumption of fossil fuels for the state of texas was a staggering 136.903% increase.
- For California the percent increase was about 87.38%.
- The state of Arizona witnessed a drastic increase of 400.35% in the consumption of fossil fuels.
- For the state with the least amount of consumption i.e. New Mexico saw an increase of 151.31%.

5) **Hydroelectricity total production(HYTCP) :** Generation of electricity by moving water.

   Unit of measurement: Million kilowatt hours

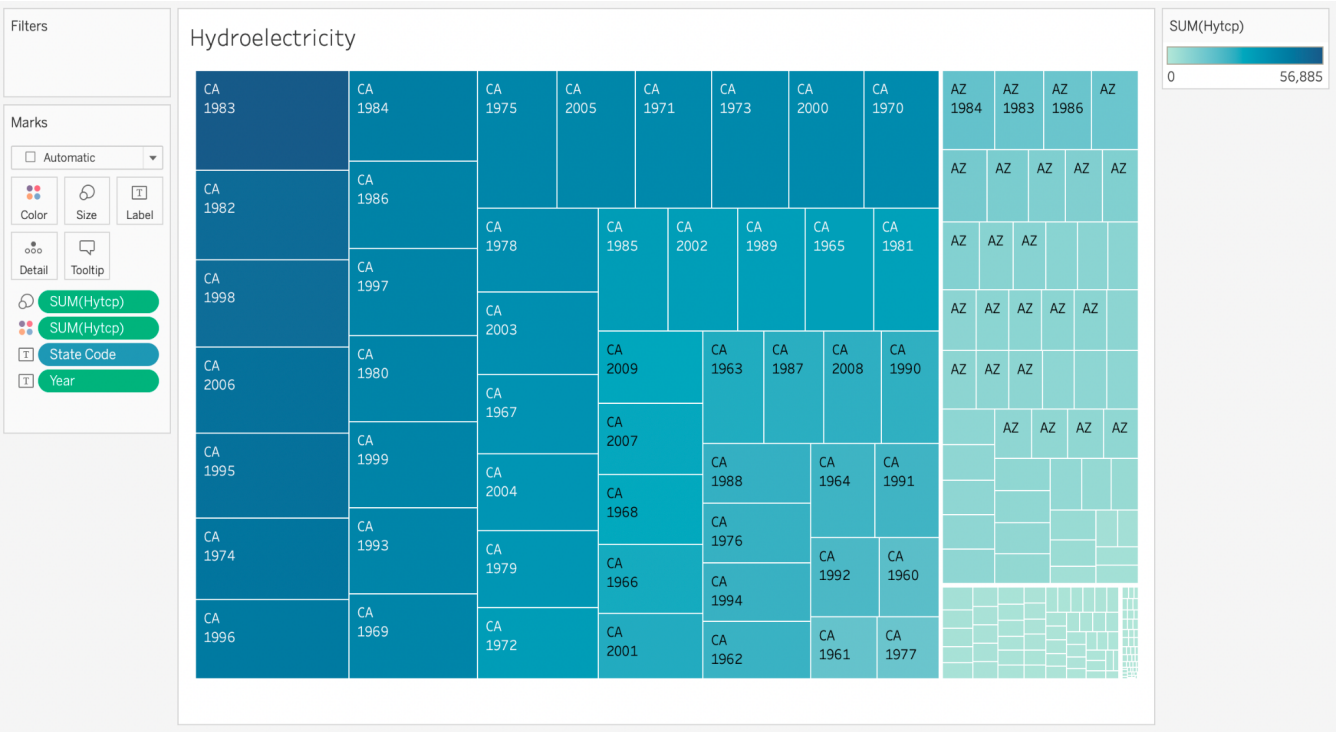


Figure 2.5

Facts:

- A total of 2108871.25 Million kilowatt hours of hydroelectricity was produced by the four states combined.

- Out of which, California produced 78.92% of the total hydroelectricity.
- New Mexico produced the least amount of hydroelectricity amounting to 0.32%.

6) Kerosene total end use consumption(KSTXP) :

Unit of Measurement: Thousand barrels

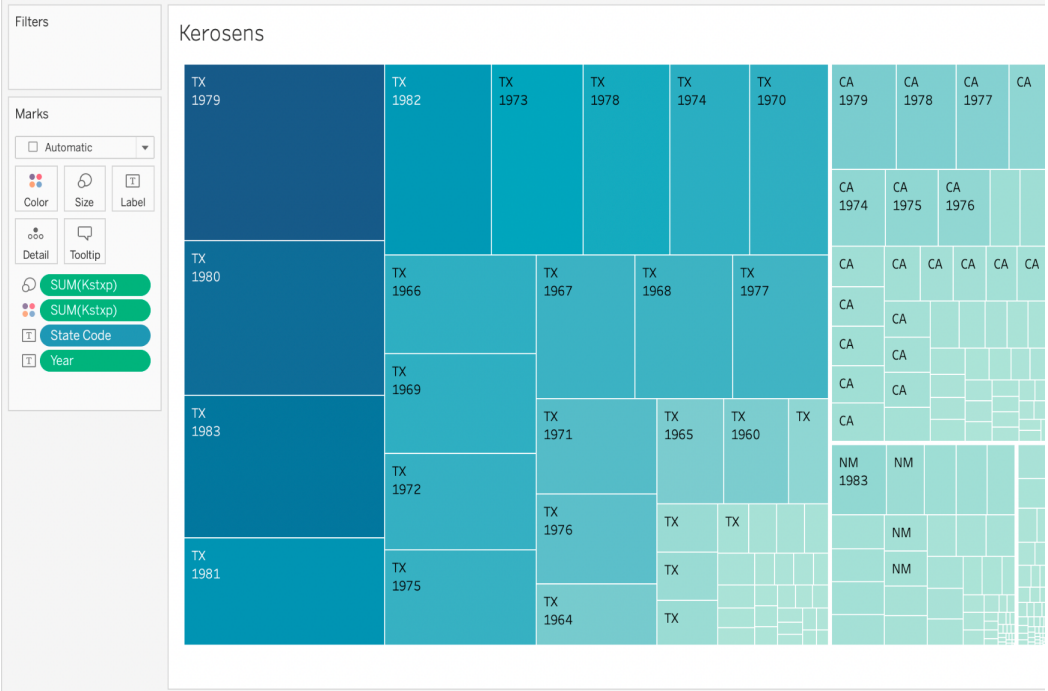


Figure 2.6

Facts:

- In 1979, Texas was responsible for the highest use of kerosene with a total of 17,670 thousand barrels. Also out of the four states for the time frame 1960-2009, Texas was the largest user of kerosene.
- In the year 1993, Arizona used the least amount of kerosene amounting to 3 thousand barrels.
- A total of 249202.163 thousand barrels was consumed by the four states.

7) **Crude oil production(PAPRP)** :

Unit of Measurement : Thousand barrels

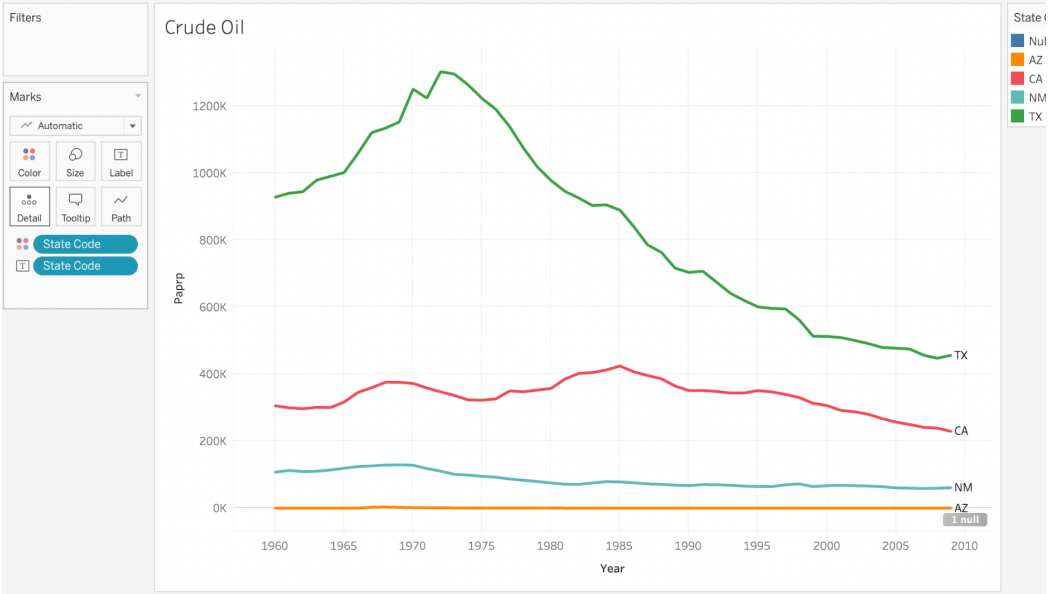


Figure 2.7

Facts:

- The state of Arizona is not depicted in the area plot as its crude oil production is very low compared to the other three states.
- Around the 1970's all the three states hit their peak production of crude oil. From there onwards the production falls and keeps on decreasing.
- A total of 62826296.5 thousand barrels were produced by the four states.
- Texas was responsible for about 66.64% of the total crude oil produced by the four states.
- Arizona was responsible for just 0.03% production of crude oil among the four states.

8) **Petroleum coke total consumption(PCTCP)** : Petroleum coke is a by-product of crude oil refining.

Unit of Measurement: Thousand barrels

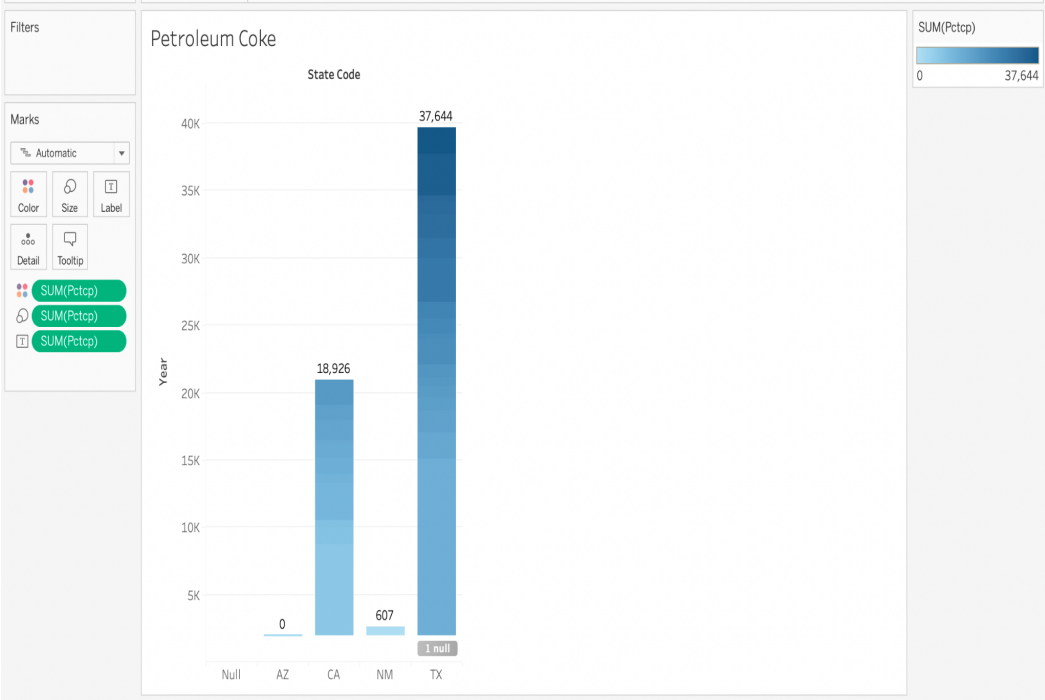


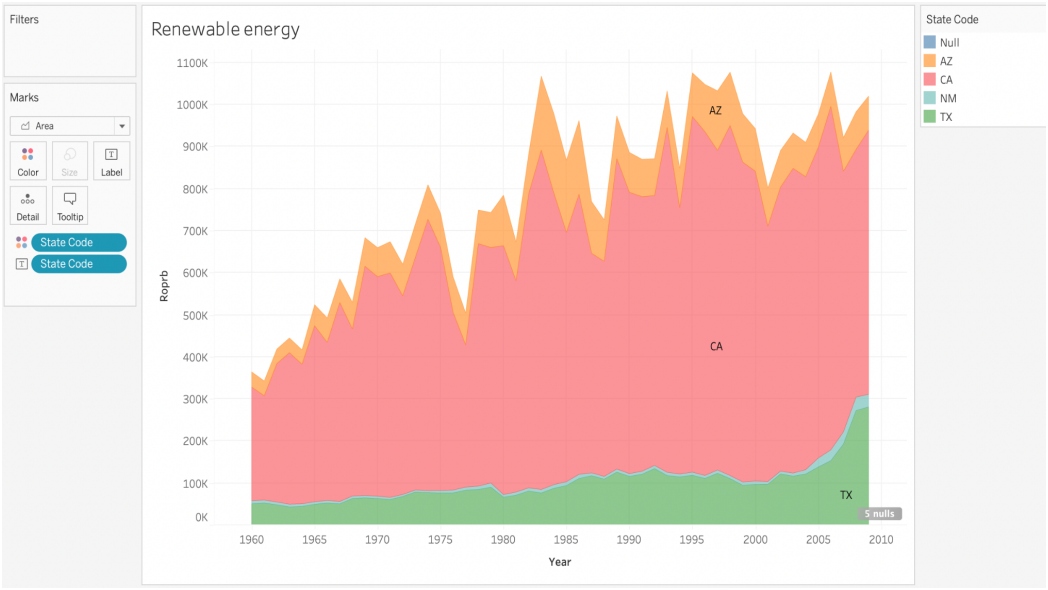
Figure 2.8

Facts:

- As it is clearly from the gantt chart, California and Texas are largely responsible for producing petroleum coke.
- The two states i.e. Texas and California were jointly responsible for a total of 98.98% production of petroleum coke.

9) **Renewable energy production(ROPRB)** : There are several types of renewable forms of energy. Here we exclude the production of fuel ethanol.

Unit of measurement: Billion Btu

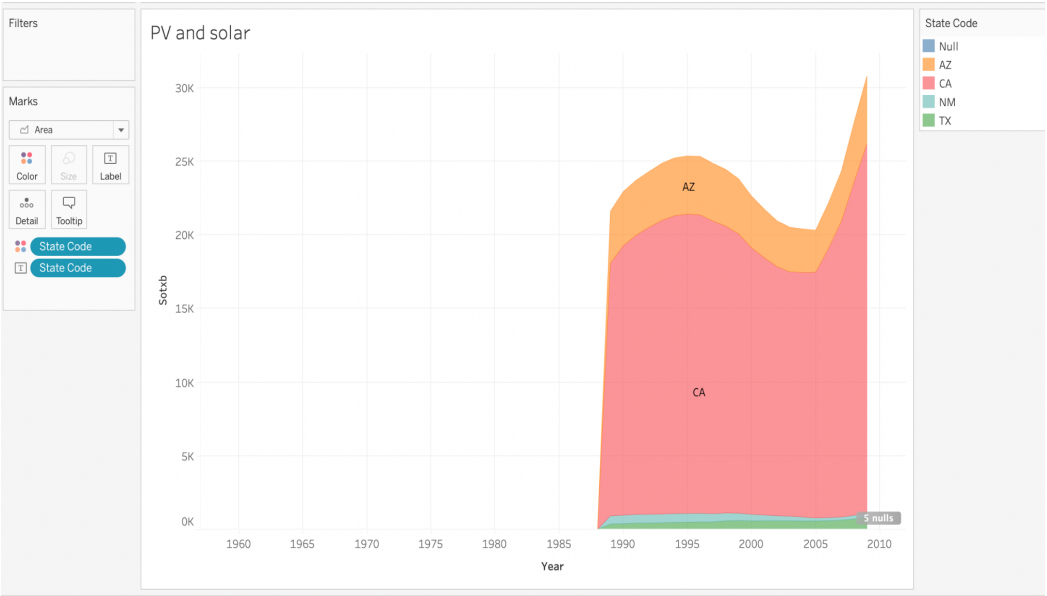


Facts:

- Here it is evident that California dominates the production of renewable forms of energy.
- Approximately 74.89% of renewable energy production is amassed by the state of California.
- The state following California was Texas. Texas accounted for a total of 12.53% of the renewable energies produced.

10) **Photovoltaic and solar thermal energy(SOTXB)** : This is a form of renewable source of energy. Solar panels are used to store the photovoltaic cells that store the energy produced from solar rays.

Unit of measurement: Billion Btu



Facts:

- California dominates this space of solar energy. It is clearly depicted by the area plot as well.
- New Mexico and Texas are not responsible for much of thermal energy.
- Arizona sits in the second spot.
- Nearly 80.62% of the solar energy produced is down to California.

11) **Total primary energy and electricity consumption(TNTXB)** : Primary sources of energy include fossil fuels, renewable energies and nuclear energy.

Unit of measurement : Billion Btu

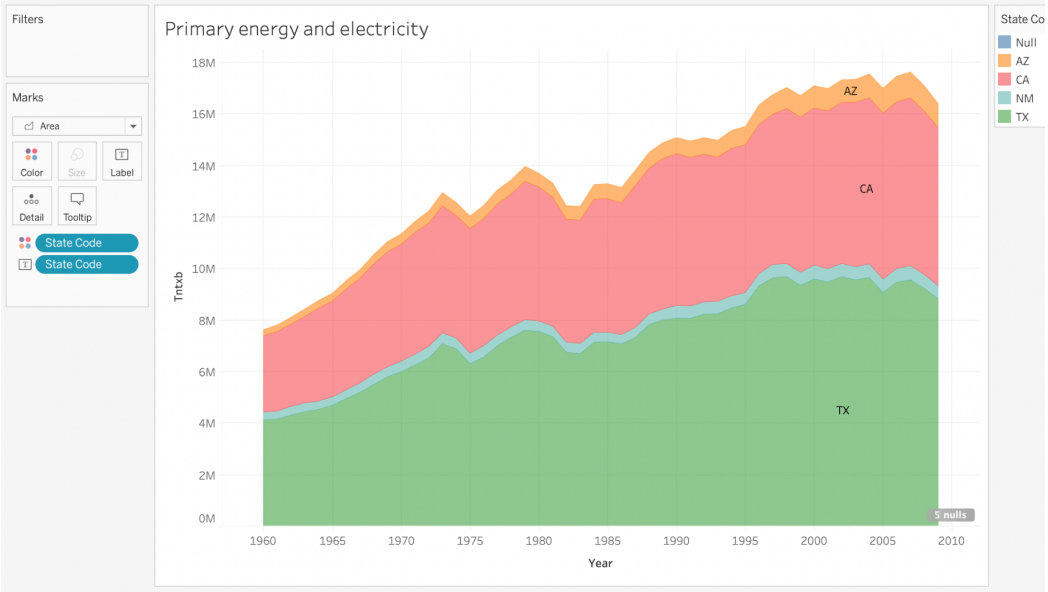


Figure 2.11

Facts:

- Texas takes the primary spot in the consumption of primary energies. Followed by California, then Arizona and finally New Mexico.
- Texas saw a rise of 113.48% of consumption from 1960-2009.
- Whereas California saw an increase of 106.9% for the time frame 1960-2009.

VI. Conclusions

With all the visual plots, it is evident that in most cases either California(CA) or Texas(TX) is leading the way. Be it production or consumption of energies, these two states are heavily involved and either state tops the ranking. As for the other two states Arizona(AZ) and New Mexico(NM) these two fall at the bottom of the ladder in regards to production or consumption of energies.

VII. Recommendations

The state of Arizona and New Mexico should invest in the production of cleaner and renewable sources of energies. They are much behind Texas and California in the production of renewable forms of energies.

Texas and California should look to decrease their production and consumption of non-renewable energies. They produce tremendous amounts of renewable energies also. The production of crude oil for all the states has decreased which is a good sign.
With all the evidence presented in the report I believe ‘California’ has the best energy profile among the four states. As its production for renewable energies is the highest.

VIII. References

i) eia.gov
ii) sciencedirect.com
iii) support.microsoft.com