

## Chapter 2

# Data Visualization

Data and information are important resources to be managed in modern organizations. Business analytics refers to the skills, technologies, applications and practices for exploration and investigation of past business performance to gain insight and aid business planning. The focus is on developing new insights and understanding based on data and statistical analysis. The emphasis is on fact-based management to drive decision making.

Data visualization is an important aspect of decision maker and/or business analyst learning. There are many useful visualization tools offered by geographic information systems, which quickly plot data by map, usually by county. These are highly useful in politics, as well as in other forms of marketing, to include tracking where sales of different products might occur. They are also useful for law enforcement, seeking to identify hot areas of particular problems. This chapter will review the visualization tools offered by the open source data mining software R, and will demonstrate simple Excel models of time series data. These are meant as representative demonstrations—there clearly are many visualization tools offered by many different software products. All support the very important process of initial understanding of data relationships.

## Data Visualization

There are many excellent commercial data mining software products, although these tend to be expensive. These include SAS Enterprise Miner and IBM's Intelligent Miner, as well as many more recent variants and new products appearing regularly. One source of information is [www.kdnuggets.com](http://www.kdnuggets.com) under "software." Some of these are free. The most popular software by rdstats.com/articles/popularity (February 2016) by product are shown in Table 2.1.

**Table 2.1** Data mining software by popularity (rdstats.com)

Rank		
1	R	Open source
2	SAS	Commercial
3	SPSS	Commercial
4	WEKA	Open source
5	Statistica	Commercial
5	Rapid miner	Commercial

Rattle is a GUI system for R (also open source), and is also highly recommended. WEKA is a great system but we have found issues with reading test data making it a bit troublesome.

## R Software

Almost every data mining software provides some support in the form of visualization of data. We can use R as a case in point.

To install R, visit <https://cran.rstudio.com/>

Open a folder for R

Select Download R for windows.

To install Rattle:

Open the R Desktop icon (32 bit or 64 bit) and enter the following command at the R prompt. R will ask for a CRAN mirror. Choose a nearby location.

```
> install.packages("rattle")
```

Enter the following two commands at the R prompt. This loads the Rattle package into the library and then starts up Rattle.

```
> library(rattle)
> rattle()
```

If the RGtk2 package has yet to be installed, there will be an error popup indicating that libatk-1.0-0.dll is missing from your computer. Click on the OK and then you will be asked if you would like to install GTK+. Click OK to do so. This then downloads and installs the appropriate GTK+ libraries for your computer. After this has finished, do exit from R and restart it so that it can find the newly installed libraries.

When running Rattle a number of other packages will be downloaded and installed as needed, with Rattle asking for the user's permission before doing so. They only need to be downloaded once.

The installation has been tested to work on Microsoft Windows, 32bit and 64bit, XP, Vista and 7 with R 3.1.1, Rattle 3.1.0 and RGtk2 2.20.31. If you are missing something, you will get a message from R asking you to install a package. I read

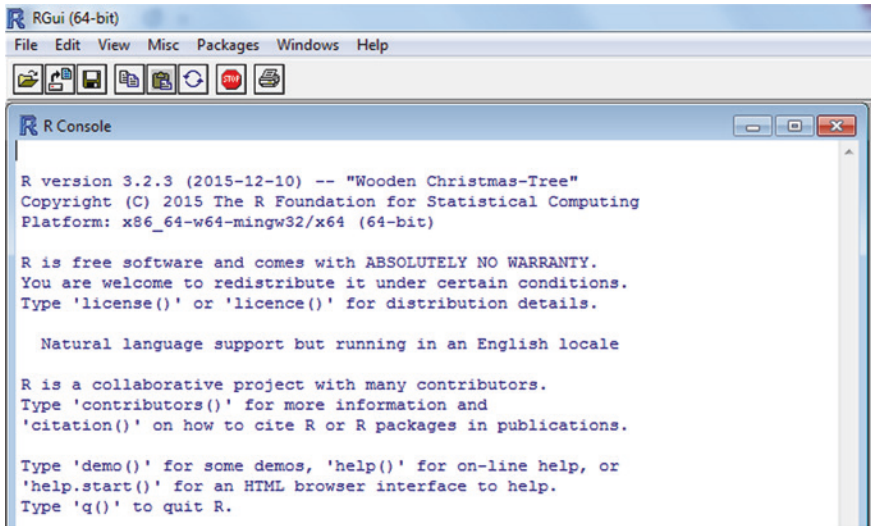


Fig. 2.1 R console

nominal data (string), and was prompted that I needed “stringr”. On the R console (see Fig. 2.1), click on the “Packages” word on the top line.

Give the command “Install packages” which will direct you to HTTPS CRAN mirror. Select one of the sites (like “USA(TX) [https]”) and find “stringr” and click on it. Then upload that package. You may have to restart R.

## Loan Data

This data set consists of information on applicants for appliance loans. The full data set, taken from a previous text (Olson and Shi 2007), involves 650 past observations. Applicant information on age, income, assets, debts, and credit rating (from a credit bureau, with red for bad credit, yellow for some credit problems, and green for clean credit record) is assumed available from loan applications. Variable Want is the amount requested in the appliance loan application. For past observations, variable On-Time is 1 if all payments were received on time, and 0 if not (Late or Default). The majority of past loans were paid on time. Asset, debt, and loan amount (variable Want) are used by rule to generate categorical variable risk. Risk was categorized as high if debts exceeded assets, as low if assets exceeded the sum of debts plus the borrowing amount requested, and average in between.

An extract of Loan Data is shown in Table 2.2.

In Fig. 2.2, 8 variables are identified from the file LoanRaw.csv. Rattle will hold out 30% of the data points for testing or other purposes by default. This

**Table 2.2** Extract of loan data

Age	Income	Assets	Debts	Want	Risk	Credit	Result
20	17,152	11,090	20,455	400	High	Green	On-time
23	25,862	24,756	30,083	2300	High	Green	On-time
28	26,169	47,355	49,341	3100	High	Yellow	Late
23	21,117	21,242	30,278	300	High	Red	Default
22	7127	23,903	17,231	900	Low	Yellow	On-time
26	42,083	35,726	41,421	300	High	Red	Late
24	55,557	27,040	48,191	1500	High	Green	On-time
27	34,843	0	21,031	2100	High	Red	On-time
29	74,295	88,827	100,599	100	High	Yellow	On-time
23	38,887	6260	33,635	9400	Low	Green	On-time
28	31,758	58,492	49,268	1000	Low	Green	On-time
25	80,180	31,696	69,529	1000	High	Green	Late
33	40,921	91,111	90,076	2900	Average	Yellow	Late
36	63,124	164,631	144,697	300	Low	Green	On-time
39	59,006	195,759	161,750	600	Low	Green	On-time
39	125,713	382,180	315,396	5200	Low	Yellow	On-time
55	80,149	511,937	21,923	1000	Low	Green	On-time
62	101,291	783,164	23,052	1800	Low	Green	On-time
71	81,723	776,344	20,277	900	Low	Green	On-time
63	99,522	783,491	24,643	200	Low	Green	On-time

leaves 448 observations. We could include all if we wished, but we proceed with this training set. Variables 3, 4 and 5 are used to calculate variable credit, so they are duplications. By clicking the “Ignore” radio button, these variables are deleted from analysis. The outcome variable is variable 8, “On-time”, so the use should make sure that the Target radio button is highlighted. When the use is satisfied with the variables for analysis, the **Execute** button on the top ribbon can be selected.

Next the **Explore** tab can be selected, which provides Fig. 2.3, where the user can select various visualization displays.

Figure 2.3 provides basic statistics for continuous variables, and the number of categories for categorical data. It has a number of visualization tools to examine each variable. In Fig. 2.3, we selected box plots for Age and Income. Selecting **Execute** yields Fig. 2.4.

Figure 2.4 gives an idea for the range and distribution of Age and Income. Each variable’s box plot is displayed using all data points (454 in this case), as well as by outcome variable result. We can see from Fig. 2.4 that borrowers that repaid on time have older observations, while problem borrowers have age distribution that is younger.

We can also explore categorical variables through bar plots in Fig. 2.5.

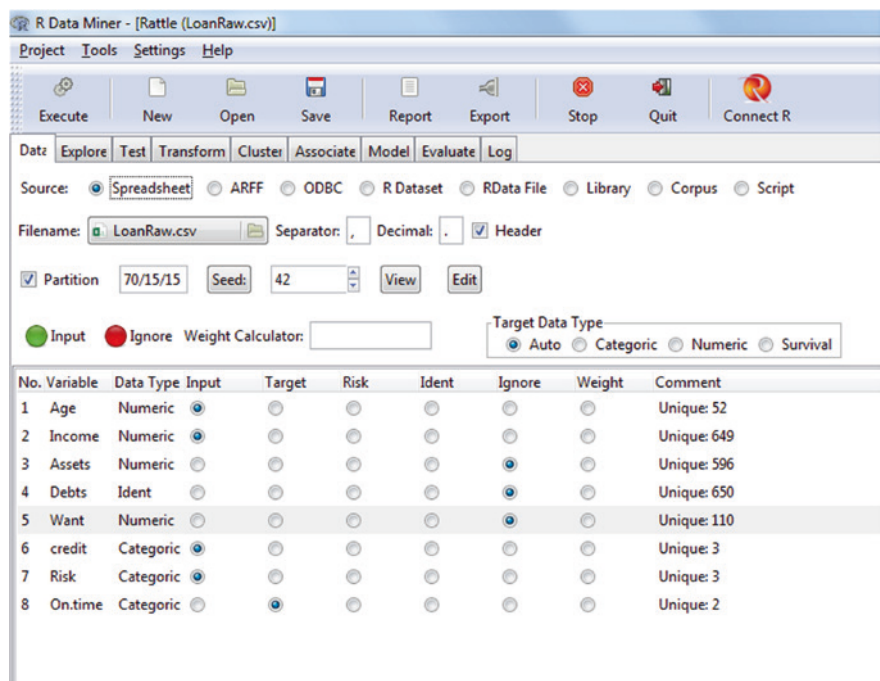


Fig. 2.2 Loading data file in R

Figure 2.5 shows the distributions of variables Credit, Risk, and On-time by categories. The On-time display is tautological, as the 409 OK outcomes by definition were OK. But the displays for variables Credit and Risk show the difference in outcome be each category. With respect to credit, there is a higher proportion of problem loans for category “red” than for either “amber” or “green. For variable Risk, the proportion of problems is clearly worse for “high” risk than for “medium” or “low” (as would be expected).

Another visualization option is **Mosaic**. Figure 2.6 shows the display using this tool for Credit and for Risk.

It says the same thing as Fig. 2.5, but in a different format.

We also can select **Pairs**, yielding Fig. 2.7.

This output provides the same information content as Figs. 2.5 and 2.6, but provides another way to show users data relationships.

Rattle also provides Principal Components display. This is called up as shown in Fig. 2.8.

The output obtained after selecting **Execute** is shown in Fig. 2.9.

The output is a mass of observations on the two principal component vectors obtained from the algorithm. Input variables Age and Income provide a bit of frame of reference, showing those outlying observations that are extreme. For

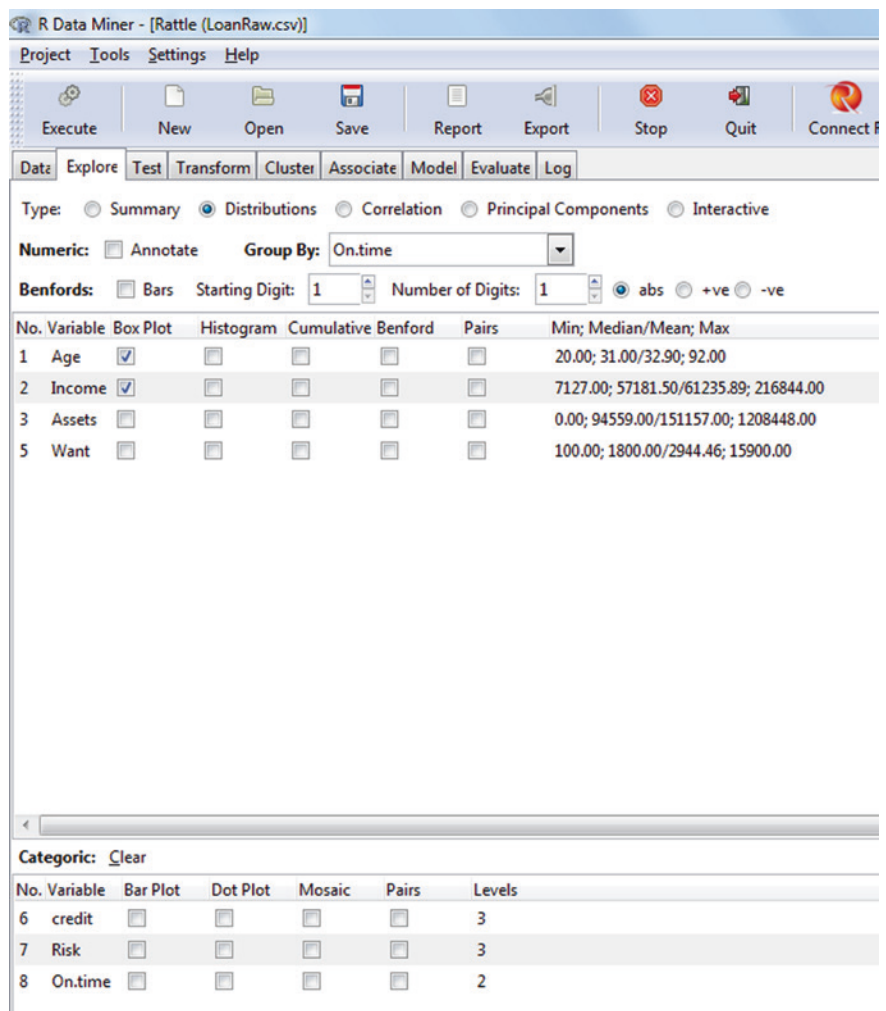


Fig. 2.3 Initial data visualization in R

instance, observation 104 has high income but average age, while observation 295 is the reverse. In the data set, observation 104 the age was 45, and income 215,033, while for observation 295 the age was 79 and income 92,010. From Fig. 2.3, we see that mean age was 32.90 and mean income 61,236. This provides some frame of reference for the Principal Components output.

There are other tools to aid data visualization. One of the most important is correlation, which will be covered in depth in Chap. 6, cluster analysis. Here we will redirect attention to basic data display obtained from simpler software tools. We will use Excel tools to obtain a visualization of energy data.

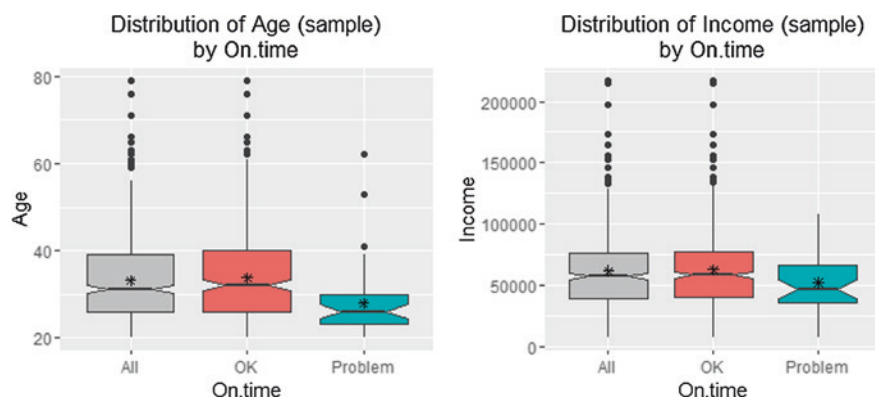


Fig. 2.4 Distribution visualization for continuous input variables

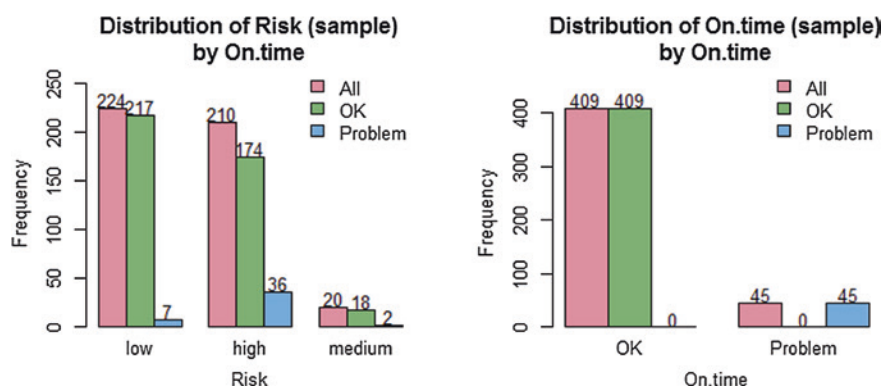
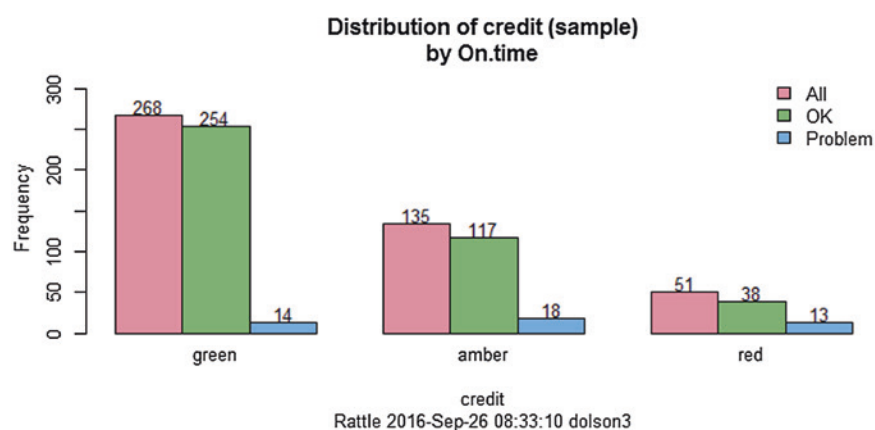


Fig. 2.5 Categorical data visualization in R

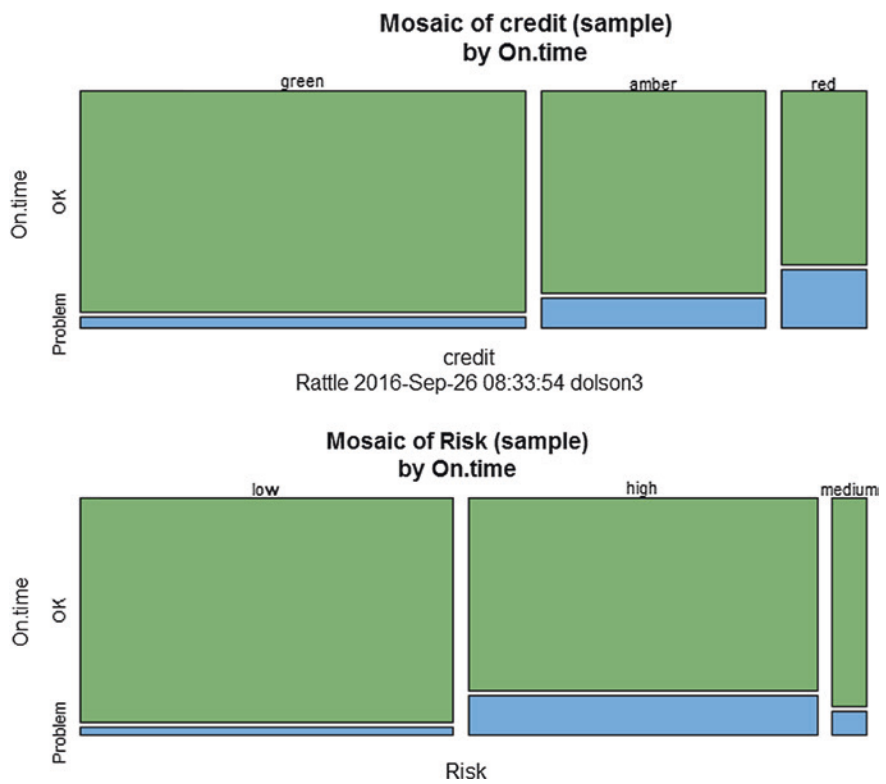


Fig. 2.6 Mosaic plot from R

## Energy Data

Energy is one of the major issues facing our society. The United States has prospered by utilizing a complex system of energy. Early US development relied on mills driven by hydro-power. Late in the 19th Century Edison and Tesla brought commercial grade electricity to cities, and ultimately to rural communities. John D. Rockefeller organized the oil industry, providing complex derivatives that have proven useful in many contexts. This includes a transportation system based on gasoline and diesel driven vehicles.

There is much disagreement about energy policy. First, there was a strong opinion that the world’s supply of crude oil was limited, and we were about to run out (Deffeyes 2001). But as oil prices soared and then in early summer 2008 plummeted, the pendulum returned with fracking, new prospects from Canada and Brazil, and the Bakken field in North Dakota coming on line. Furthermore, while some analysts argued that Saudis were running out of oil, Saudis kept quiet and appear to have lots of oil.



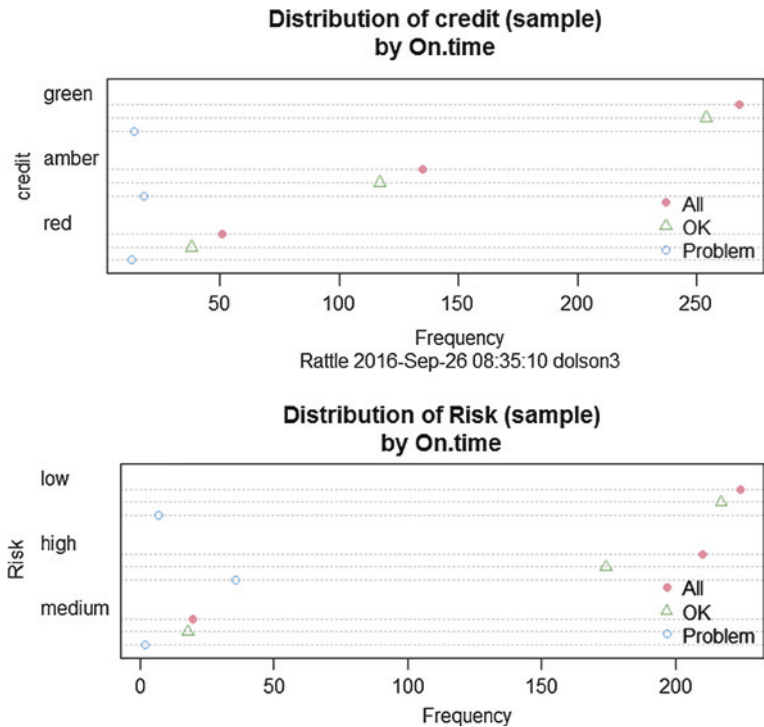
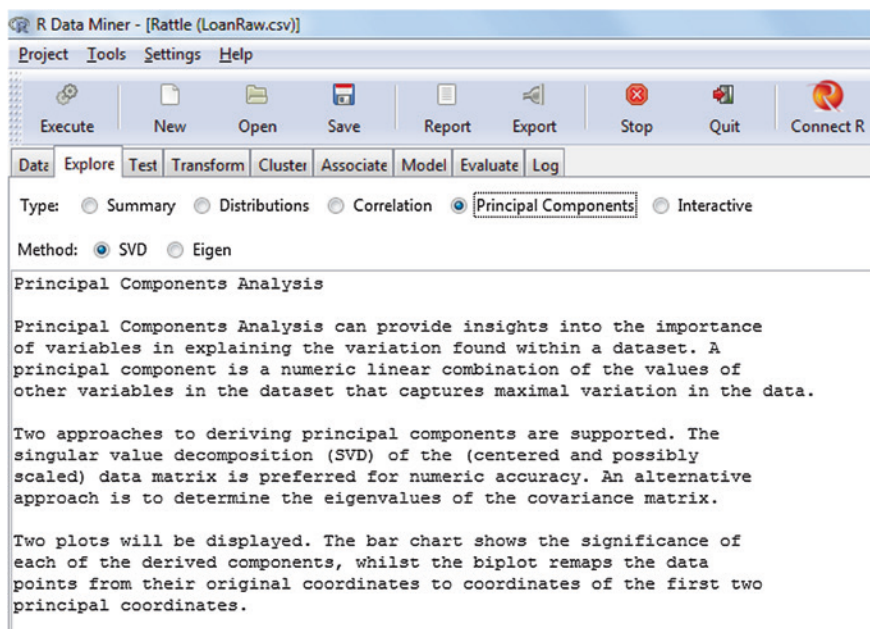


Fig. 2.7 Paired visualization in R

The second point relates to global warming. Clearly there is a rise in sea levels, as glaciers and small islands in the Pacific disappear, and we actually now can ship goods over the Northwest Passage of North America. This is caused in part by carbon emissions. One solution is to retain the current infrastructure where we can all drive our own automobiles with cleaner petroleum products, and run our coal electric plants with cleaner coal. Others feel that we need to scrap our current culture and eliminate carbon energy sources. This is a political issue that will not go away, and will drive election outcomes around the world into the foreseeable future. It is a fact of nature that people will disagree. Unfortunately, armaments often are involved more and more. Thus it is very important to understand the energy situation.

Those opposed to all carbon emissions propose wind and solar power. The US Government provided support to enterprises such as Solyndra to create viable solar power generation. This, however, has encountered problems, and Solyndra filed for bankruptcy in August 2011 (Meiners et al. 2011). There appear to be problems in translating what is physically possible with economic viability. Wind energy also involves some issues. If you fly into northern Europe, you can often see fields of 100 very large wind turbines just off-shore. They also exist off Brazil.



**Fig. 2.8** Principal components description from Rattle

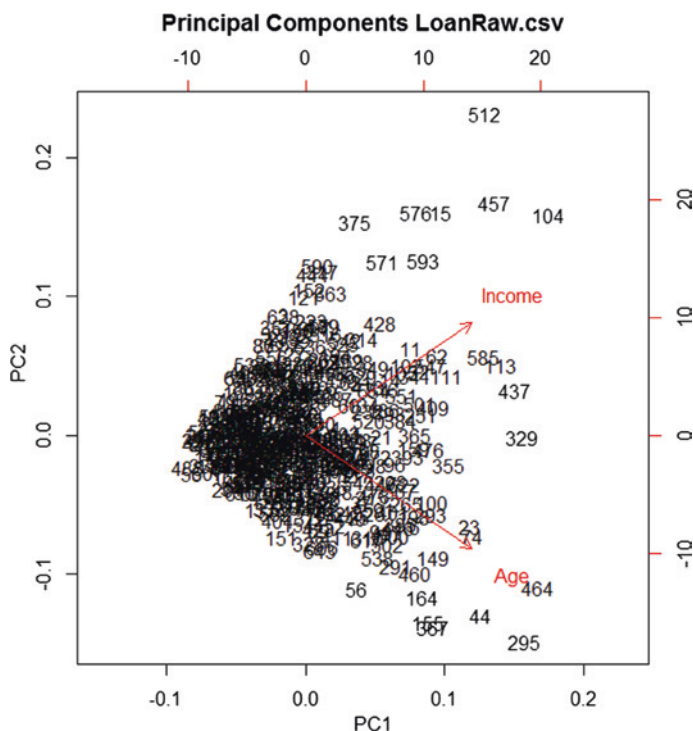
There are more and more wind farms in the US as well. But they have annoying features for nearby residents, kill birds, and like solar energy, are not capable of continuous generation of power.

Another non-carbon based energy source is nuclear. In the 1950s there was strong movement toward penny-cheap electricity generated by turning the sword of nuclear weapons into the plowshare of nuclear power plants. Nearly 100 such plants were built in the United States. But problems were encountered in 1979 at Three Mile Island in Pennsylvania (Levi 2013), and then the catastrophe at Chernobyl in the Ukraine in 1986. People don't want nuclear plants anymore, and what was a cheap source of power became expensive after the Federal Government insisted of retrofitting plants to provide very high levels of protection. There also is the issue of waste disposal that has become a major political issue, especially in Nevada.

Thus there are a number of important issues related to generation of energy in the United States, as well as in the world. The US Department of Energy provides monthly reports that provide an interesting picture that we might visualize with simple Excel tools. The source of this data over time is:

<http://www.eia.gov/totalenergy/data/monthly/>.

One of the obvious ways to visualize time series data is to plot it over time. Table 2.3 displays a recap of US energy production, displaying the emergence of geothermal energy in 1970, and solar and wind energy in 1990. Nuclear power didn't come on line until the late 1950s, and has slowly grown over the years.



**Fig. 2.9** Principal components plot for loan data in R

Natural gas plant liquids (NGPL) was small in 1950, but has exceeded hydro-power. Crude oil has been highly variable. The volumes for hydroelectricity has been fairly steady, and relatively small compared to natural gas, crude oil, and coal.

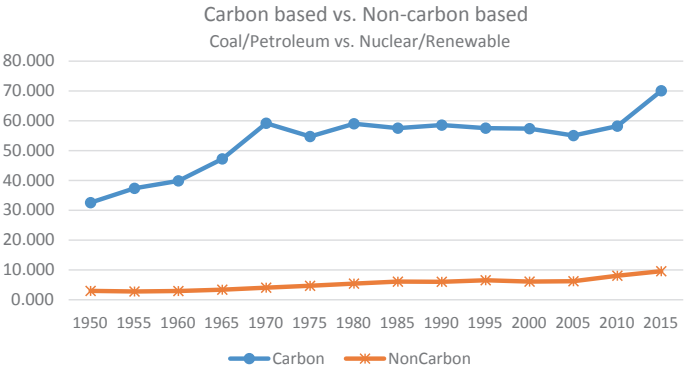
## Basic Visualization of Time Series

Figure 2.1 displays the growth in carbon versus non-carbon production. The data comes from Table 2.3, reorganized by adding columns Coal through NGPL for Carbon, columns Nuclear through Biomass for Non-Carbon. Using the Year column of Table 2.3 as the X-axis, Fig. 2.10 displays an Excel plot for the Carbon based total versus the Non-carbon based total by year.

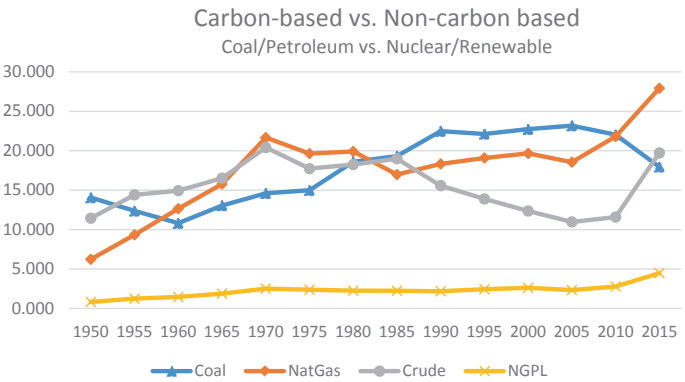
Figure 2.10 shows that there has been more variance in Carbon-based production of energy. There was a peak around 1965, with a drop in 1970 as the economy slowed. Growth resumed in 1975, but there was a decline due to massive increases in oil price due to OPEC and the Iranian crisis. Carbon-based production actually

Table 2.3 US energy production

Year	Coal	NatGas	Crude	NGPL	Nuclear	Hydro	Geotherm	Solar	Wind	Biomass	Total
1950	14.06	6.23	11.45	0.82	0.00	1.42	0.00	0.00	0.00	1.56	35.54
1955	12.37	9.34	14.41	1.24	0.00	1.36	0.00	0.00	0.00	1.42	40.15
1960	10.82	12.66	14.93	1.46	0.01	1.61	0.00	0.00	0.00	1.32	42.80
1965	13.06	15.78	16.52	1.88	0.04	2.06	0.00	0.00	0.00	1.33	50.67
1970	14.61	21.67	20.40	2.51	0.24	2.63	0.01	0.00	0.00	1.43	63.50
1975	14.99	19.64	17.73	2.37	1.90	3.15	0.03	0.00	0.00	1.50	61.32
1980	18.60	19.91	18.25	2.25	2.74	2.90	0.05	0.00	0.00	2.48	67.18
1985	19.33	16.98	18.99	2.24	4.08	2.97	0.10	0.00	0.00	3.02	67.70
1990	22.49	18.33	15.57	2.17	6.10	3.05	0.17	0.06	0.03	2.74	70.70
1995	22.13	19.08	13.89	2.44	7.08	3.21	0.15	0.07	0.03	3.10	71.17
2000	22.74	19.66	12.36	2.61	7.86	2.81	0.16	0.06	0.06	3.01	71.33
2005	23.19	18.56	10.97	2.33	8.16	2.70	0.18	0.06	0.18	3.10	69.43
2010	22.04	21.81	11.59	2.78	8.43	2.54	0.21	0.09	0.92	4.32	74.72
2015	17.95	27.93	19.72	4.47	8.34	2.39	0.22	0.43	1.82	4.72	87.99



**Fig. 2.10** Carbon versus non-carbon US production



**Fig. 2.11** US production of carbon-based energy

languished and even dropped around 2000. At that time there was a predominant thought that we had reached a “peak oil” point, where the amount of reserves available was finite and would soon run out (Simmons 2005). However, Saudi’s continued to contend that they had plenty of reserves, and in the US, a new major field in North Dakota was brought into production, while fracking increased output from old oil fields. The US then switched from a net oil importer to a net oil exporter. Figure 2.10 shows that oil production is increasing. Table 2.3 shows that coal production is dropping, but natural gas and crude production are growing significantly. This information is displayed graphically in Fig. 2.11.

As to alternative energy, Fig. 2.12 shows the small level of US energy production from wind and solar energy, the primary alternative energy sources.

Table 2.4 gives US Annual consumption in trillion BTUs by sector.

Figure 2.13 displays US energy use by sector graphically.

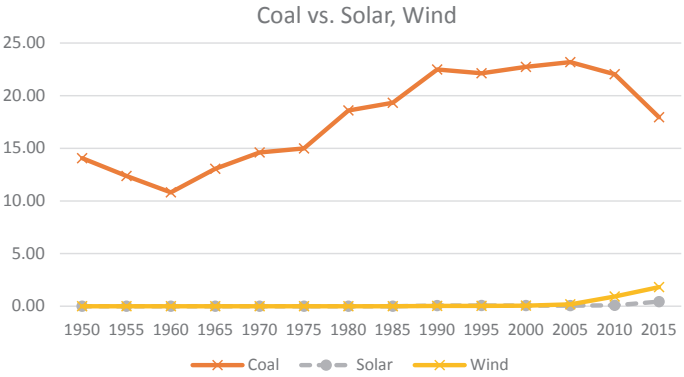


Fig. 2.12 Comparison of coal, solar, and wind energy—US

Table 2.4 US trillion BTUs/  
Year

Year	Residential	Commercial	Industrial	Transportation
1950	5989	3893	16,241	8492
1955	7278	3895	19,485	9550
1960	9039	4609	20,842	10,596
1965	10,639	5845	25,098	12,432
1970	13,766	8346	29,628	16,098
1975	14,813	9492	29,413	18,245
1980	15,753	10,578	32,039	19,697
1985	16,041	11,451	28,816	20,088
1990	16,944	13,320	31,810	22,420
1995	18,517	14,690	33,970	23,851
2000	20,421	17,175	34,662	26,555
2005	21,612	17,853	32,441	28,280
2010	21,793	18,057	30,525	27,059
2015	20,651	17,993	31,011	27,706

Looking at Fig. 2.13 we can see that all sectors have grown, with a bit of a dip around the 2008 economic crisis. The most volatile is the largest sector, industrial, which suffered downturns around the 1973 OPEC emergence (which raised the price of oil substantially, bringing on intensive inflation for the rest of the 1970s), and around 1985 when the Iranian crisis led to another major increase in the price of crude oil. There has been a noticeable slowing in industrial consumption since 2000. Residential consumption also has dipped a bit between 2010 and 2015. Transportation increased every period except around the 2008 global financial crisis.

Figure 2.14 shows another type of data display. In this case, the Department of Energy monitors detailed flows of production sources to consumption sectors.

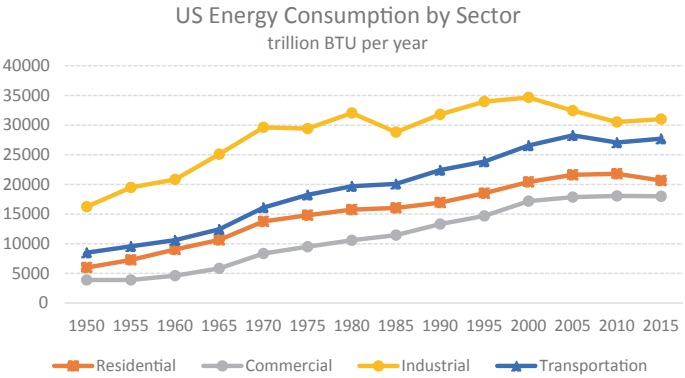


Fig. 2.13 Plot of US energy consumption by sector

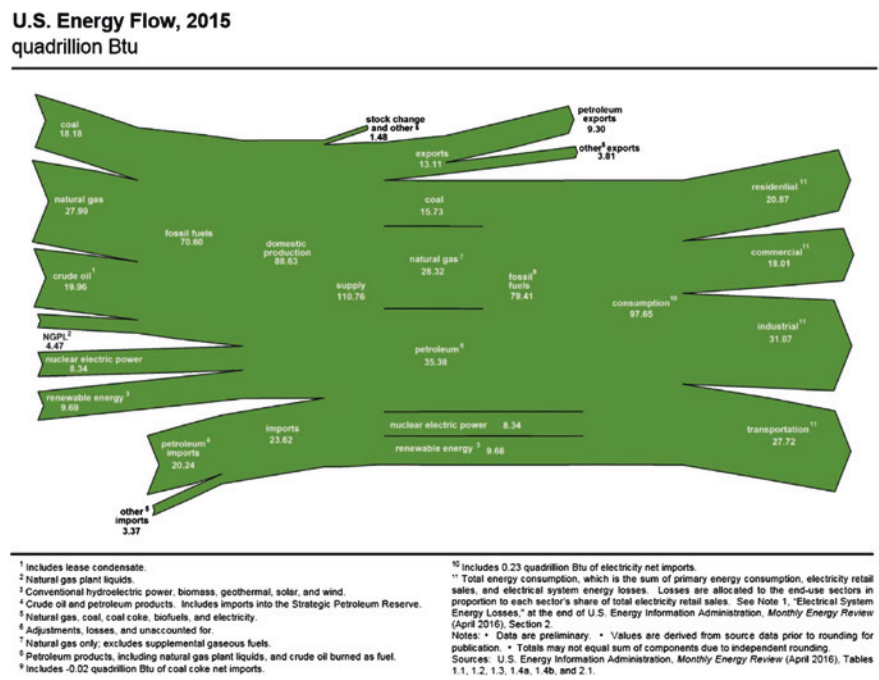


Fig. 2.14 US DOE display of energy flows in 2015

This excellent graphic gives a picture of the complex flows in aggregate terms. We have taken the individual annual graphs (available since 2001) and extracted values for major inputs and outputs in Table 2.5.

Table 2.5 US energy by source and sector

	Coal	NG	Crude	NGPL	Nuclear	Renew	Import	Resident	Comm	Indust	Trans	Export
2001	23.44	19.84	12.39	2.54	8.03	5.52	29.65	20.16	17.44	32.6	26.75	3.92
2002	22.55	19.56	12.31	2.56	8.15	5.9	29.04	20.94	17.4	32.49	26.52	3.65
2003	22.31	19.64	12.15	2.34	7.97	6.15	31.02	21.23	17.55	32.52	26.86	4.05
2004	22.69	19.34	11.53	2.47	8.23	6.12	33	21.18	17.52	33.25	27.79	4.43
2005	23.05	18.76	10.84	2.32	8.13	6.06	34.26	21.87	17.97	31.98	28.06	4.64
2006	23.79	19.02	10.87	2.35	8.21	6.79	34.49	21.05	18	32.43	28.4	4.93
2007	23.48	19.82	10.8	2.4	8.41	6.8	34.6	21.75	18.43	32.32	29.1	5.36
2008	23.86	21.15	10.52	2.41	8.46	7.32	32.84	21.64	18.54	31.21	27.92	7.06
2009	21.58	21.5	11.24	2.54	8.35	7.76	29.78	21.21	18.15	28.2	27.03	6.93
2010	22.08	22.1	11.67	2.69	8.44	8.06	29.79	22.15	18.21	30.14	27.51	8.17
2011	22.18	23.51	11.99	2.93	8.26	9.24	28.59	21.62	18.02	30.59	27.08	10.35
2012	20.68	24.63	13.76	3.25	8.06	6.84	27.07	20.08	17.41	30.77	26.75	11.36
2013	19.99	24.89	15.77	3.47	8.27	9.3	24.54	21.13	17.93	31.46	27.01	11.8
2014	20.28	26.43	18.32	4.03	8.33	9.68	23.31	21.53	18.34	31.33	27.12	12.22
2015	18.18	27.99	19.96	4.47	8.34	9.69	23.62	20.87	18.01	31.07	27.72	13.11



This data for the source side of the equation (the left side of Fig. 2.14) is graphed in Excel to provide Fig. 2.15.

It can be seen from this graph that imports were the primary source of US energy, and were bothersome to the public, which feared that oil reserves had peaked. However, around 2011 fracking increased crude production, as did North Dakota oil. Imports correspondingly declined, and the United States now finds itself the leading oil producer in the world. This same information can be displayed for any given year by a pie chart, as in Fig. 2.16.

Figure 2.17 plots the right side of Fig. 2.14.

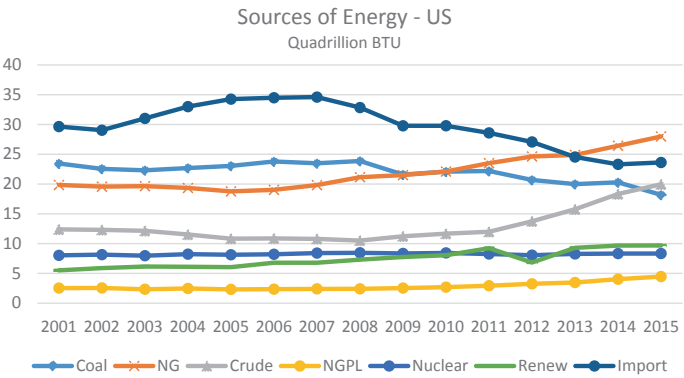
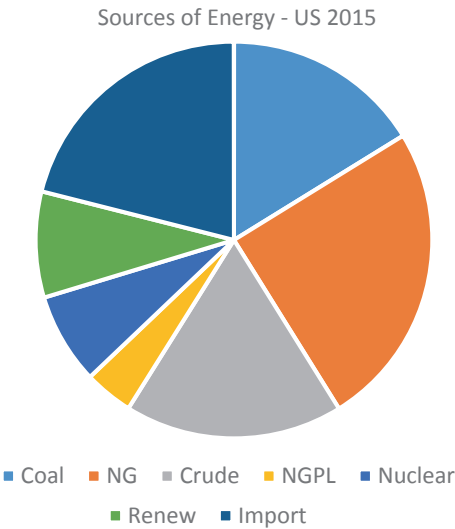


Fig. 2.15 Plot of US energy sources

Fig. 2.16 Pie chart of US energy sources in 2015



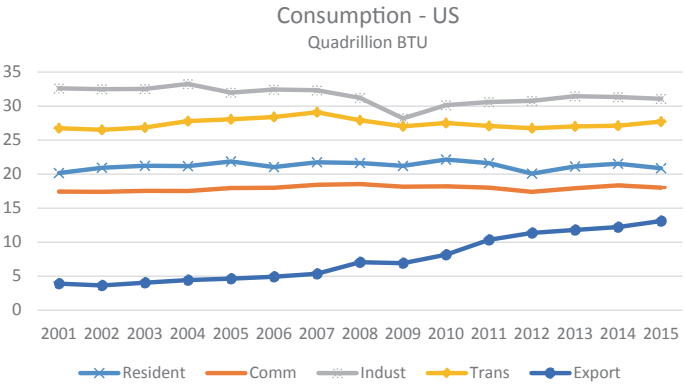
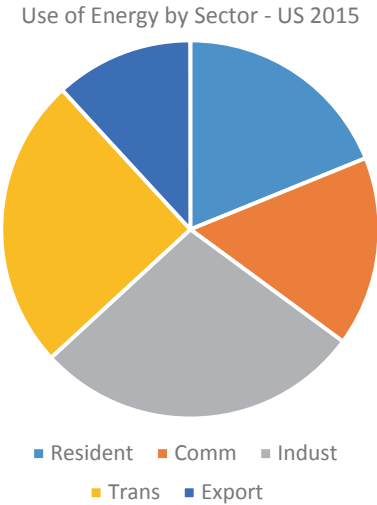


Fig. 2.17 Plot of US energy consumption

Fig. 2.18 US energy consumption by sector in 2015



This same information for the year 2015 in pie chart form is shown in Fig. 2.18. Consumption can be seen to be quite stable, with a drop in industrial consumption in 2009 (reflecting response to the 2008 financial crisis). Exports were quite low until 2007, after which time they have had steady increase.

Figure 2.19 displays another Department of Energy graphic, in this case providing details between each major source and each major consumption sector.

This data was extracted in tabular form in Table 2.6 by multiplying the given percentages by given quantities. Rounding leads to some imprecision:

Electrical power comes mostly from coal (about 37% in 2015). This number is expected to drop due to Federal policy, seeking to shift power generation away

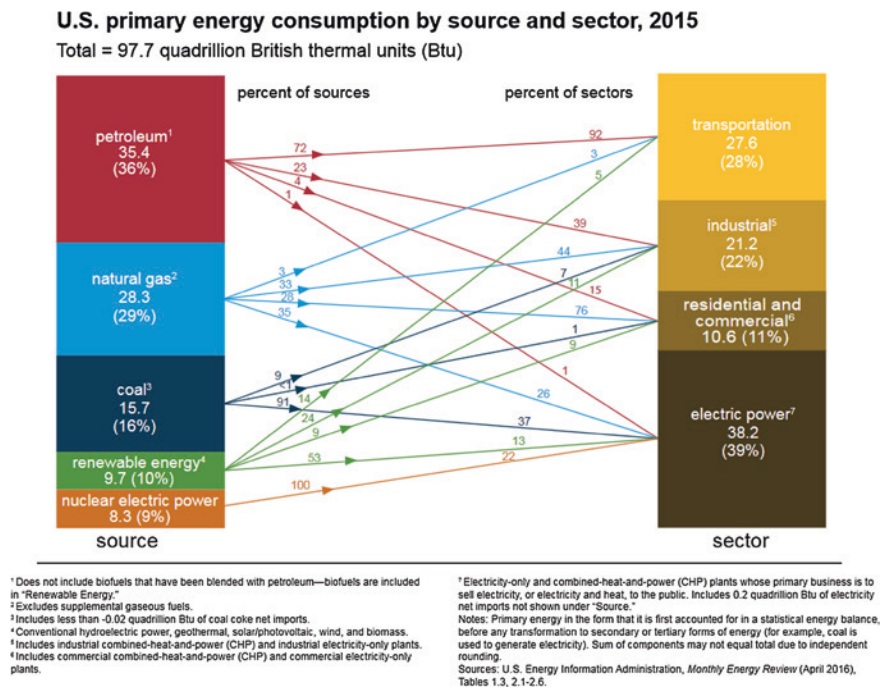


Fig. 2.19 US DOE display of US energy flow in 2015

Table 2.6 US energy flow—2015

2015	Qbillbtu	To TRANS	To IND	To RES/Com	To Elect
Petro	35.4	25.488	8.142	1.416	0.354
NG	28.3	0.849	9.339	7.924	10.188
Coal	15.7		1.413	0	14.287
Renew	9.7	1.261	2.328	1.067	5.044
Nuc	8.3				8.3
	97.4	27.598	21.222	10.407	38.173
		27.6	21.2	10.6	38.2

from coal. That policy seeks to have growth occur in the renewable sector, but at the moment, that source is only 13% of electrical power generation. Transportation energy comes predominately from petroleum. Electric vehicles are being emphasized by current government policy, but we are far away from a sustainable system of electrically powered vehicles. Furthermore, they would add to the need for electricity. Natural gas was discouraged by government policy in the 1970s, but has come back strong, and become a growing source of power that is flexible enough to deliver to a number of energy users.

## Conclusion

This chapter has brushed the tip of a very big ice berg relative to data visualization. Data visualization is important as it provides humans an initial understanding of data, which in our contemporary culture is overwhelming. We have demonstrated two types of data visualization. Data mining software provides tools as were shown from R. Excel also provides many tools that are within the reach of the millions who use Microsoft products. Neither are monopolists, and there are other products that can provide more and better visualization support. These sources have the benefit of being affordable.

We spent a great deal of attention to a specific field of data, obtained by US Government publication. Most of the work of data mining involves getting data, and then identifying which specific data is needed for the particular issue at hand. We have explored visualization of time series that can be informative relative to a number of important energy issue questions.

## References

- Deffeyes KS (2001) Hubbert's peak: the impending world oil shortage. Princeton University Press, Princeton
- Levi M (2013) The power surge: energy, opportunity, and the battle for America's future. Oxford University Press, Oxford
- Meiners RE, Morriss A, Bogart WT, Dorchak A (2011) The false promise of green energy. Cato Institute, Washington DC
- Olson DL, Shi Y (2007) Introduction to business data mining. McGraw-Hill/Irwin, New York
- Simmons MR (2005) Twilight in the desert: the coming Saudi oil shock and the world economy. Wiley, Hoboken

Descriptive Data Mining

Olson, D.L.

2017, XI, 116 p. 63 illus., 60 illus. in color., Hardcover

ISBN: 978-981-10-3339-1