01- Reading Files

R Workshop

- Data Formatting and Reshaping -

Outline

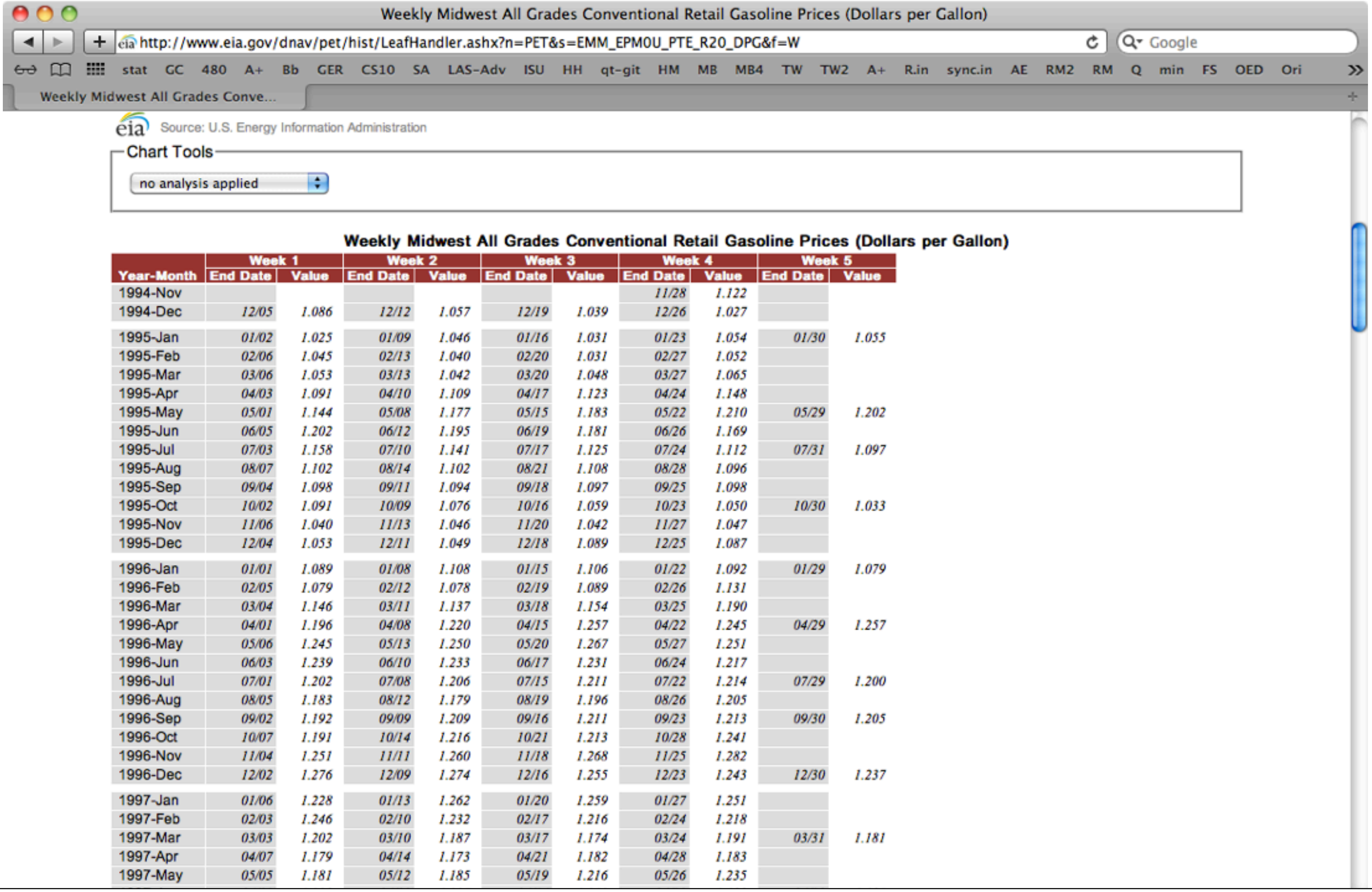
- Reading files: Excel and R
- packages gdata and foreign
- reading SAS xport files

Data in Excel

- Formats xls and csv - what's the difference?
- File extensions xls or xlsx are proprietary Excel formats, they are binary files
- csv is the extension for comma separated value files. They are text files - and directly readable.
- Example: gas prices in the Midwest since 1994 (from data.gov and EIA)

Gas Prices

xls file in Website:


Weekly Midwest All Grades Conventional Retail Gasoline Prices (Dollars per Gallon)
http://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=EMM_EPM0U_PTE_R20_DPG&f=W
Google
stat GC 480 A+ Bb GER CS10 SA LAS-Adv ISU HH qt-git HM MB MB4 TW TW2 A+ R.in sync.in AE RM2 RM Q min FS OED Ori
Weekly Midwest All Grades Conve...
eia

Source: U.S. Energy Information Administration

Chart Tools

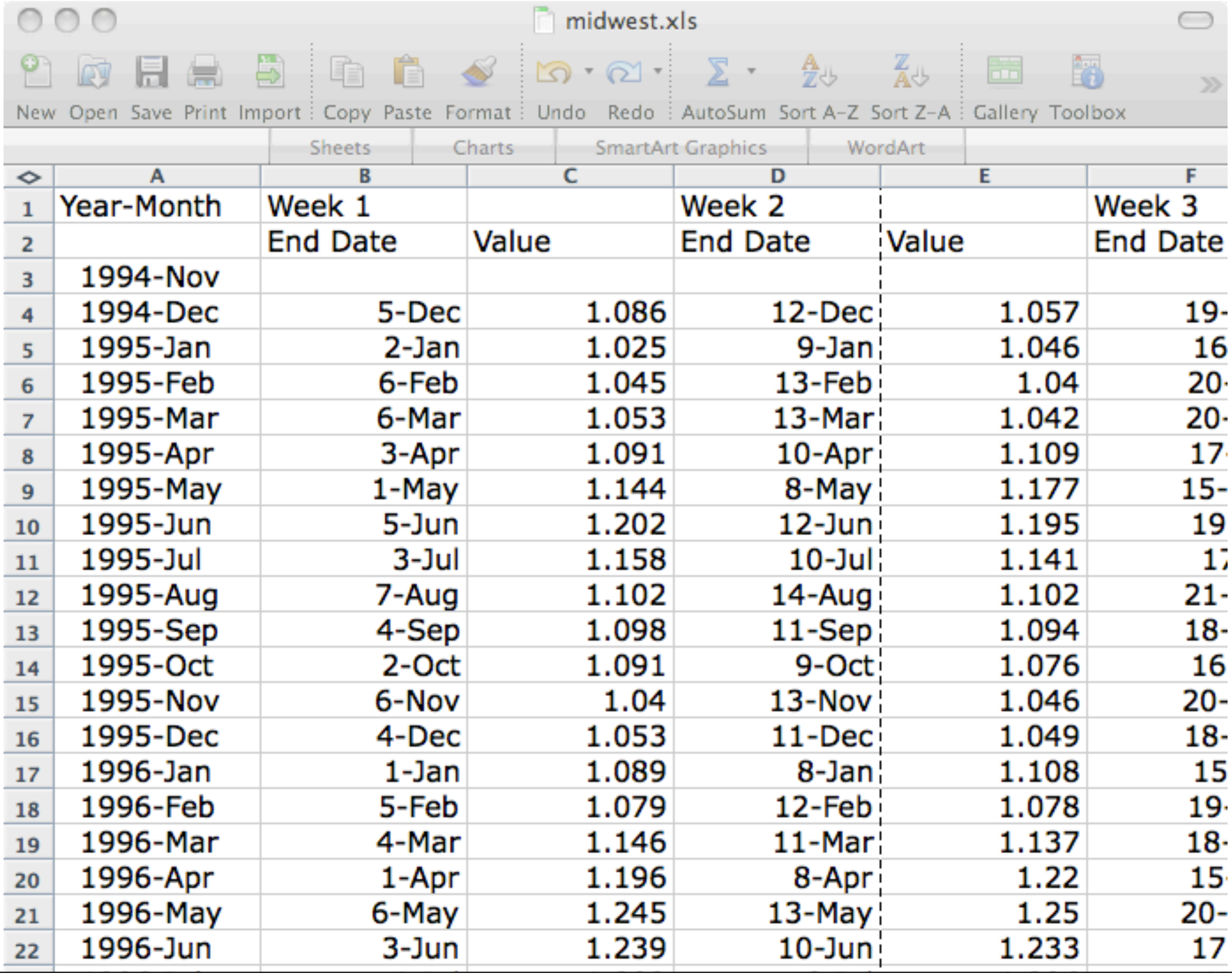
no analysis applied

Weekly Midwest All Grades Conventional Retail Gasoline Prices (Dollars per Gallon)

	Week 1		Week 2		Week 3		Week 4		Week 5	
Year-Month	End Date	Value	End Date	Value	End Date	Value	End Date	Value	End Date	Value
1994-Nov							11/28	1.122		
1994-Dec	12/05	1.086	12/12	1.057	12/19	1.039	12/26	1.027		
1995-Jan	01/02	1.025	01/09	1.046	01/16	1.031	01/23	1.054	01/30	1.055
1995-Feb	02/06	1.045	02/13	1.040	02/20	1.031	02/27	1.052		
1995-Mar	03/06	1.053	03/13	1.042	03/20	1.048	03/27	1.065		
1995-Apr	04/03	1.091	04/10	1.109	04/17	1.123	04/24	1.148		
1995-May	05/01	1.144	05/08	1.177	05/15	1.183	05/22	1.210	05/29	1.202
1995-Jun	06/05	1.202	06/12	1.195	06/19	1.181	06/26	1.169		
1995-Jul	07/03	1.158	07/10	1.141	07/17	1.125	07/24	1.112	07/31	1.097
1995-Aug	08/07	1.102	08/14	1.102	08/21	1.108	08/28	1.096		
1995-Sep	09/04	1.098	09/11	1.094	09/18	1.097	09/25	1.098		
1995-Oct	10/02	1.091	10/09	1.076	10/16	1.059	10/23	1.050	10/30	1.033
1995-Nov	11/06	1.040	11/13	1.046	11/20	1.042	11/27	1.047		
1995-Dec	12/04	1.053	12/11	1.049	12/18	1.089	12/25	1.087		
1996-Jan	01/01	1.089	01/08	1.108	01/15	1.106	01/22	1.092	01/29	1.079
1996-Feb	02/05	1.079	02/12	1.078	02/19	1.089	02/26	1.131		
1996-Mar	03/04	1.146	03/11	1.137	03/18	1.154	03/25	1.190		
1996-Apr	04/01	1.196	04/08	1.220	04/15	1.257	04/22	1.245	04/29	1.257
1996-May	05/06	1.245	05/13	1.250	05/20	1.267	05/27	1.251		
1996-Jun	06/03	1.239	06/10	1.233	06/17	1.231	06/24	1.217		
1996-Jul	07/01	1.202	07/08	1.206	07/15	1.211	07/22	1.214	07/29	1.200
1996-Aug	08/05	1.183	08/12	1.179	08/19	1.196	08/26	1.205		
1996-Sep	09/02	1.192	09/09	1.209	09/16	1.211	09/23	1.213	09/30	1.205
1996-Oct	10/07	1.191	10/14	1.216	10/21	1.213	10/28	1.241		
1996-Nov	11/04	1.251	11/11	1.260	11/18	1.268	11/25	1.282		
1996-Dec	12/02	1.276	12/09	1.274	12/16	1.255	12/23	1.243	12/30	1.237
1997-Jan	01/06	1.228	01/13	1.262	01/20	1.259	01/27	1.251		
1997-Feb	02/03	1.246	02/10	1.232	02/17	1.216	02/24	1.218		
1997-Mar	03/03	1.202	03/10	1.187	03/17	1.174	03/24	1.191	03/31	1.181
1997-Apr	04/07	1.179	04/14	1.173	04/21	1.182	04/28	1.183		
1997-May	05/05	1.181	05/12	1.185	05/19	1.216	05/26	1.235		

Gas Prices

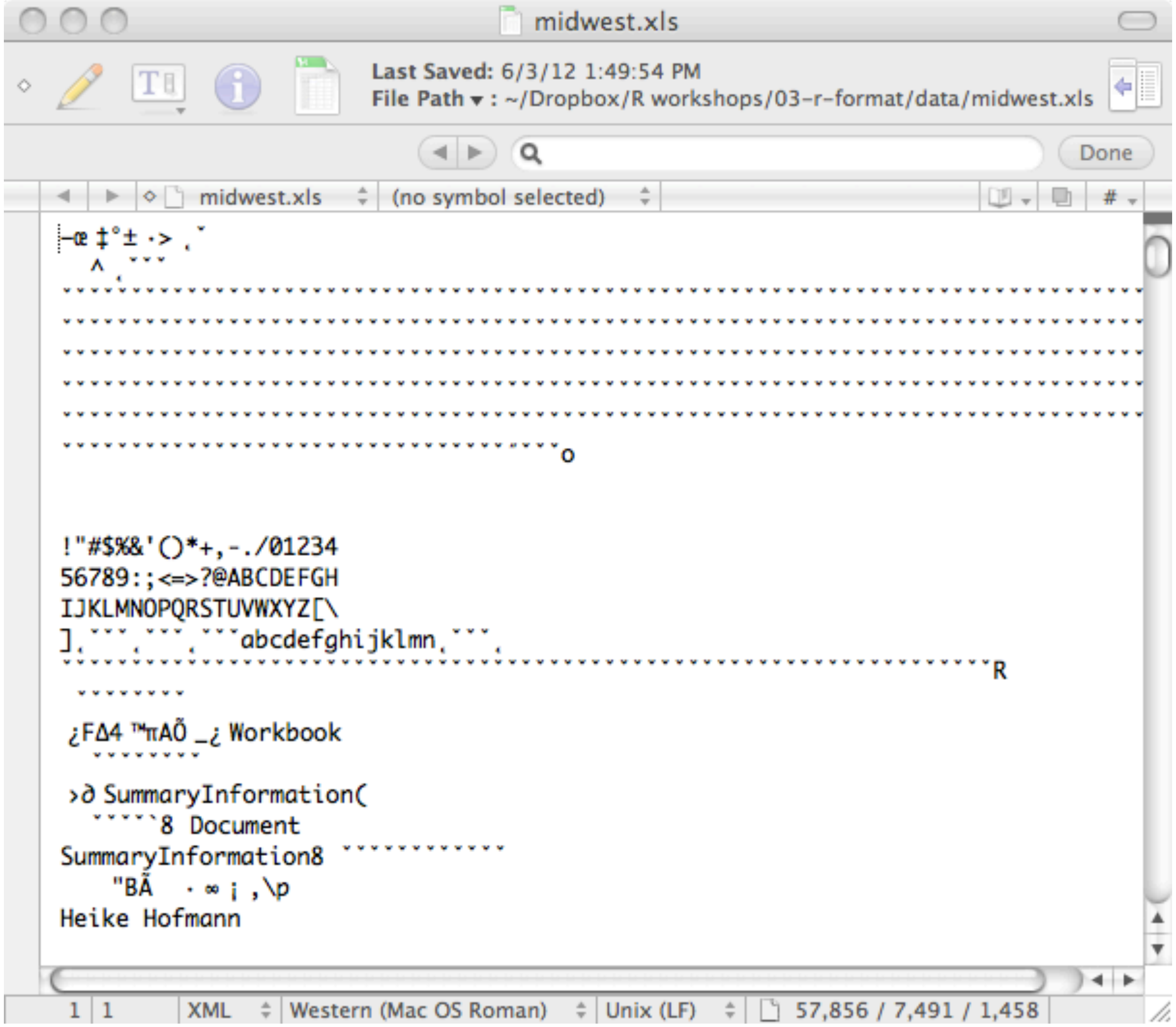
xls file in Excel:



	A	B	C	D	E	F
1	Year-Month	Week 1		Week 2		Week 3
2		End Date	Value	End Date	Value	End Date
3	1994-Nov					
4	1994-Dec	5-Dec	1.086	12-Dec	1.057	19-
5	1995-Jan	2-Jan	1.025	9-Jan	1.046	16
6	1995-Feb	6-Feb	1.045	13-Feb	1.04	20-
7	1995-Mar	6-Mar	1.053	13-Mar	1.042	20-
8	1995-Apr	3-Apr	1.091	10-Apr	1.109	17-
9	1995-May	1-May	1.144	8-May	1.177	15-
10	1995-Jun	5-Jun	1.202	12-Jun	1.195	19
11	1995-Jul	3-Jul	1.158	10-Jul	1.141	17
12	1995-Aug	7-Aug	1.102	14-Aug	1.102	21-
13	1995-Sep	4-Sep	1.098	11-Sep	1.094	18-
14	1995-Oct	2-Oct	1.091	9-Oct	1.076	16
15	1995-Nov	6-Nov	1.04	13-Nov	1.046	20-
16	1995-Dec	4-Dec	1.053	11-Dec	1.049	18-
17	1996-Jan	1-Jan	1.089	8-Jan	1.108	15
18	1996-Feb	5-Feb	1.079	12-Feb	1.078	19-
19	1996-Mar	4-Mar	1.146	11-Mar	1.137	18-
20	1996-Apr	1-Apr	1.196	8-Apr	1.22	15-
21	1996-May	6-May	1.245	13-May	1.25	20-
22	1996-Jun	3-Jun	1.239	10-Jun	1.233	17

Gas Prices

- xls file in text editor



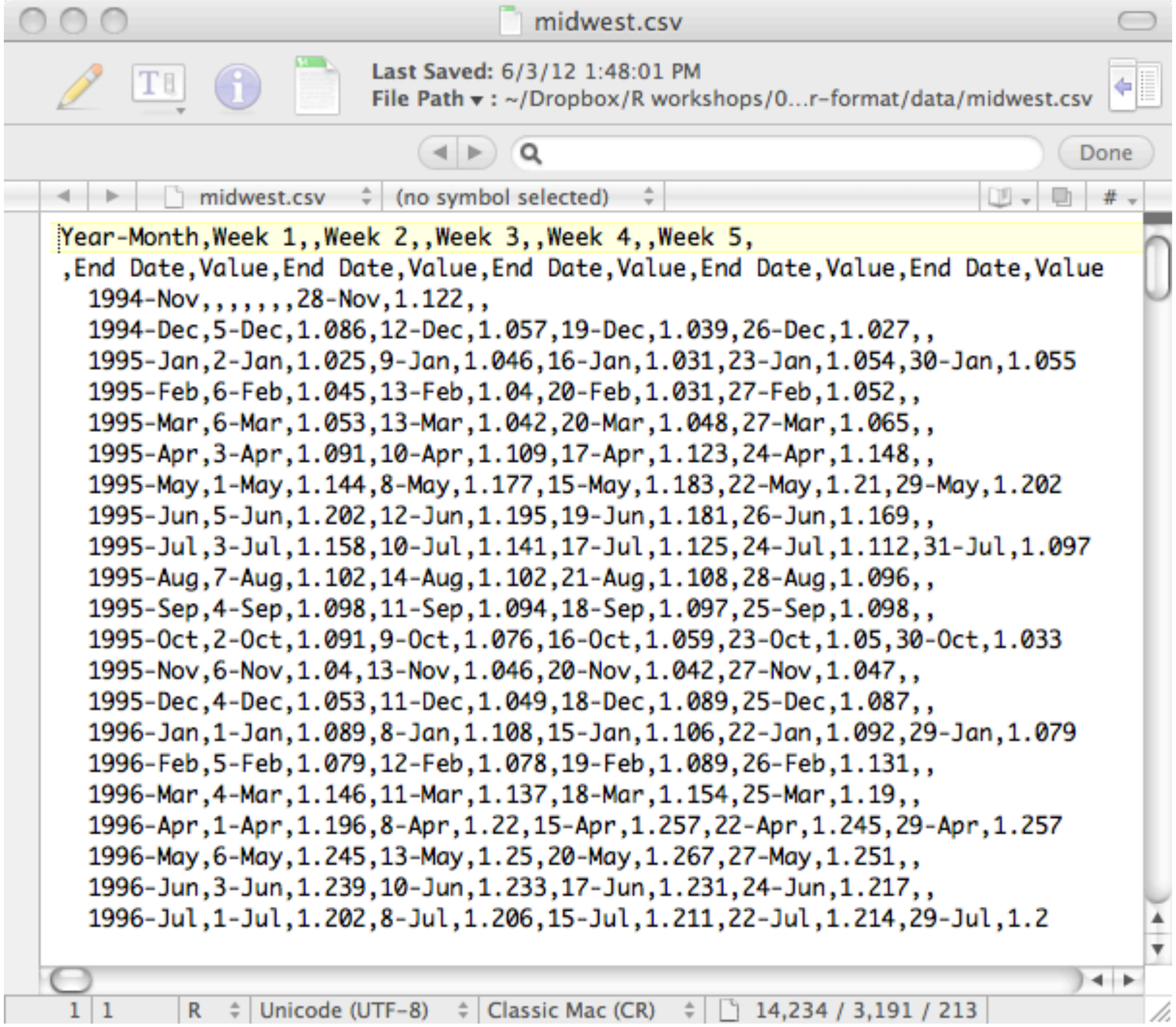
Last Saved: 6/3/12 1:49:54 PM
File Path ▾ : ~/Dropbox/R workshops/03-r-format/data/midwest.xls

```
├œ‡°±·> ˛ˇ
  ^ ˛ˇˇˇ
ˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇ
ˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇ
ˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇ
ˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇ
ˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇ
ˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇ˝o



!"#$%&'()*+,-./01234
56789:;<=>?@ABCDEFGH
IJKLMNOPQRSTUVWXYZ[\
],ˇˇˇ,ˇˇˇ,ˇˇˇabcdefghijklmn,ˇˇˇ,
ˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇˇR
 ˇˇˇˇˇˇˇˇ
 ¿F∆4 ™πAÕ _¿ Workbook
  ˇˇˇˇˇˇˇˇ

 ›∂ SummaryInformation(
  ˇˇˇˇ`8 Document
SummaryInformation8 ˇˇˇˇˇˇˇˇˇˇˇˇ
    "BÃ  · ∞ ¡ ,\p
Heike Hofmann
```

1 | 1   XML   Western (Mac OS Roman)   Unix (LF)   57,856 / 7,491 / 1,458

Gas Prices

- csv file in text editor

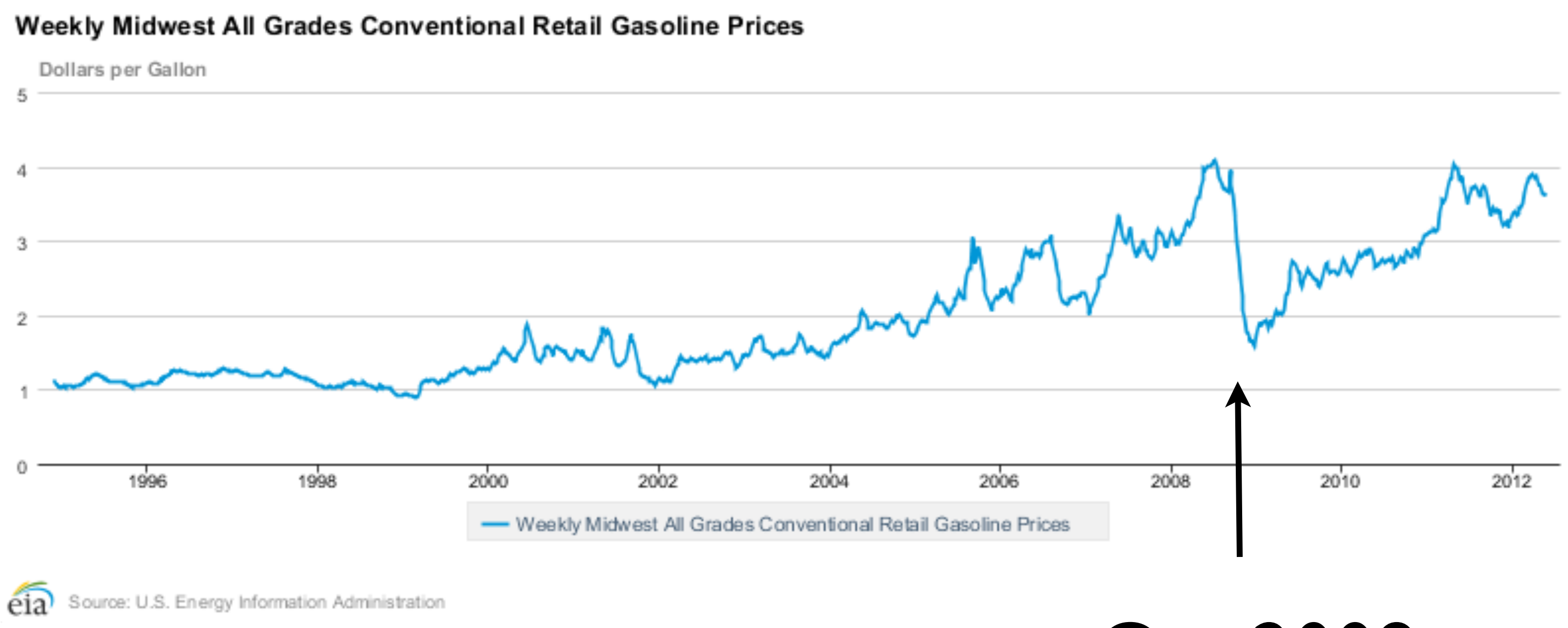


Last Saved: 6/3/12 1:48:01 PM
File Path ▾ : ~/Dropbox/R workshops/0...r-format/data/midwest.csv

```
Year-Month,Week 1,,Week 2,,Week 3,,Week 4,,Week 5,
,End Date,Value,End Date,Value,End Date,Value,End Date,Value,End Date,Value
  1994-Nov,,,,,,,28-Nov,1.122,,
  1994-Dec,5-Dec,1.086,12-Dec,1.057,19-Dec,1.039,26-Dec,1.027,,
  1995-Jan,2-Jan,1.025,9-Jan,1.046,16-Jan,1.031,23-Jan,1.054,30-Jan,1.055
  1995-Feb,6-Feb,1.045,13-Feb,1.04,20-Feb,1.031,27-Feb,1.052,,
  1995-Mar,6-Mar,1.053,13-Mar,1.042,20-Mar,1.048,27-Mar,1.065,,
  1995-Apr,3-Apr,1.091,10-Apr,1.109,17-Apr,1.123,24-Apr,1.148,,
  1995-May,1-May,1.144,8-May,1.177,15-May,1.183,22-May,1.21,29-May,1.202
  1995-Jun,5-Jun,1.202,12-Jun,1.195,19-Jun,1.181,26-Jun,1.169,,
  1995-Jul,3-Jul,1.158,10-Jul,1.141,17-Jul,1.125,24-Jul,1.112,31-Jul,1.097
  1995-Aug,7-Aug,1.102,14-Aug,1.102,21-Aug,1.108,28-Aug,1.096,,
  1995-Sep,4-Sep,1.098,11-Sep,1.094,18-Sep,1.097,25-Sep,1.098,,
  1995-Oct,2-Oct,1.091,9-Oct,1.076,16-Oct,1.059,23-Oct,1.05,30-Oct,1.033
  1995-Nov,6-Nov,1.04,13-Nov,1.046,20-Nov,1.042,27-Nov,1.047,,
  1995-Dec,4-Dec,1.053,11-Dec,1.049,18-Dec,1.089,25-Dec,1.087,,
  1996-Jan,1-Jan,1.089,8-Jan,1.108,15-Jan,1.106,22-Jan,1.092,29-Jan,1.079
  1996-Feb,5-Feb,1.079,12-Feb,1.078,19-Feb,1.089,26-Feb,1.131,,
  1996-Mar,4-Mar,1.146,11-Mar,1.137,18-Mar,1.154,25-Mar,1.19,,
  1996-Apr,1-Apr,1.196,8-Apr,1.22,15-Apr,1.257,22-Apr,1.245,29-Apr,1.257
  1996-May,6-May,1.245,13-May,1.25,20-May,1.267,27-May,1.251,,
  1996-Jun,3-Jun,1.239,10-Jun,1.233,17-Jun,1.231,24-Jun,1.217,,
  1996-Jul,1-Jul,1.202,8-Jul,1.206,15-Jul,1.211,22-Jul,1.214,29-Jul,1.2
```

1 | 1 | R | Unicode (UTF-8) | Classic Mac (CR) | 14,234 / 3,191 / 213

… what we'd like to do with the data …



Weekly Midwest All Grades Conventional Retail Gasoline Prices

Source: U.S. Energy Information Administration

Oct 2008

Reading Files in R

- Textfiles: usually comma separated (or tabular separated)

```
read.csv(file, header = TRUE, sep = ",", quote = "\\"",
         dec = ".",  fill = TRUE, comment.char = "", ...)

read.table (file, header = FALSE, sep = "", quote = "\\"'", dec = ".",
      row.names, col.names, as.is = !stringsAsFactors, na.strings = "NA",
      colClasses = NA, nrows = -1, skip = 0, check.names = TRUE,
      fill = !blank.lines.skip, strip.white = FALSE, blank.lines.skip = TRUE,
      comment.char = "#", allowEscapes = FALSE, flush = FALSE,
      stringsAsFactors = default.stringsAsFactors(), fileEncoding = "",
      encoding = "unknown", text)
```

Gas Prices in the Midwest

```
# read a csv file published as a webfile
gp <- read.csv("http://www.hofroe.net/R workshops/03-r-format/data/midwest.csv")

# read (and find) a local csv file
gp <- read.csv(file.choose())
```

```
# reveals awful format
head(gp)
```

```
  Year.Month   Week.1     X   Week.2   X.1   Week.3   X.2   Week.4   X.3   Week.5   X.4
1            End Date Value End Date Value End Date Value End Date Value End Date Value
2   1994-Nov                                                  28-Nov 1.122
3   1994-Dec    5-Dec 1.086   12-Dec 1.057   19-Dec 1.039   26-Dec 1.027
4   1995-Jan    2-Jan 1.025    9-Jan 1.046   16-Jan 1.031   23-Jan 1.054   30-Jan 1.055
5   1995-Feb    6-Feb 1.045   13-Feb  1.04   20-Feb 1.031   27-Feb 1.052
6   1995-Mar    6-Mar 1.053   13-Mar 1.042   20-Mar 1.048   27-Mar 1.065
```

Gas Prices in the Midwest

```
 str(gp)
```

```
'data.frame':  212 obs. of  11 variables:
 $ Year.Month: Factor w/ 212 levels "","  1994-Dec",..: 1 3 2 8 7 11 4 12 10 9 ...
 $ Week.1    : Factor w/ 86 levels "","1-Apr","1-Aug",..: 86 1 52 18 65 69 26 10 56 31 ...
 $ X         : Factor w/ 197 levels "","0.905","0.918",..: 197 1 19 7 12 13 21 29 42 31 ...
 $ Week.2    : Factor w/ 86 levels "","10-Apr","10-Aug",..: 86 1 28 78 41 45 2 70 32 7 ...
 $ X.1       : Factor w/ 206 levels "","0.919","0.921",..: 206 1 17 14 12 13 27 39 45 34 ...
 $ Week.3    : Factor w/ 86 levels "","15-Apr","15-Aug",..: 86 1 52 18 65 69 26 10 56 31 ...
 $ X.2       : Factor w/ 199 levels "","0.91","0.929",..: 199 1 11 9 9 15 28 40 38 29 ...
 $ Week.4    : Factor w/ 85 levels "22-Apr","22-Aug",..: 85 82 51 17 64 68 25 9 55 30 ...
 $ X.3       : Factor w/ 201 levels "0.883","0.921",..: 201 29 9 14 13 15 32 44 34 27 ...
 $ Week.5    : Factor w/ 31 levels "","29-Apr","29-Aug",..: 31 1 1 16 1 1 1 9 1 27 ...
 $ X.4       : Factor w/ 74 levels "","0.955","1.023",..: 74 1 1 5 1 1 1 18 1 11 ...
```

needs some more work before we can analyze
(or even visualize the data)

Gas Prices in the Midwest

Issues with the data:

- two lines of header information
- all variables are factor variables - but we know, that some are dates, some are numeric

Your Turn

- Have a look at the parameters of read.table in the help (Hint: try `?read.table` to view the help) to solve the following problems:
  - Read the first two lines of the file into an object called ‘gp_names’
  - Read everything but the first two lines into an object called ‘gp_data’

Reading Excel Data

- Need another package: `gdata`

```
read.xls(xls, sheet = 1, verbose = FALSE, pattern, ..., method = c("csv",
     "tsv", "tab"), perl = "perl")
```

```
library("gdata")
# get html page with an overview of the package functionality
help(package="gdata")
```

```
gp2 <- read.xls(file.choose(), sheet=1)
head(gp2)
```

```
  Year.Month   Week.1     X   Week.2   X.1   Week.3   X.2   Week.4   X.3   Week.5   X.4
1            End Date Value End Date Value End Date Value End Date Value End Date Value
2   1994-Nov                                                  28-Nov 1.122
3   1994-Dec    5-Dec 1.086   12-Dec 1.057   19-Dec 1.039   26-Dec 1.027
4   1995-Jan    2-Jan 1.025    9-Jan 1.046   16-Jan 1.031   23-Jan 1.054   30-Jan 1.055
5   1995-Feb    6-Feb 1.045   13-Feb  1.04   20-Feb 1.031   27-Feb 1.052
6   1995-Mar    6-Mar 1.053   13-Mar 1.042   20-Mar 1.048   27-Mar 1.065
```

Your Turn

- Read the file gasprices.xls into R and inspect it.
- What might be potential problems when analyzing the data?

Package foreign

- Other file formats can be read using functions from package `foreign`
- SPSS: `read.spss`
  SAS: `read.xport` (xport format)
  `read.ssd` (permanent SAS data)
  Minitab: `read.mtp`
  Systat: `read.systat`

Your Turn

- The NHANES (National Health and Nutrition Survey) publishes data in SAS export format: http://wwwn.cdc.gov/nchs/nhanes/search/nhanes09_10.aspx
- Download one of the files and load into R