STAT 1291: Data Science

Lecture 8 - Data Wrangling

Sungkyu Jung

Where are we?

- *the grammar of graphics*
- a grammar of data manipulation

Introduced in Chapter 4 of MDSR is a grammar of data manipulation for statistical tables-not unlike SQL for databases. The verbs for this data manipulation are presented for single tables and two tables. Later, we will discuss the normalization of tables is called data tidying.

Motivating example

- Vater et al., “Trends in Cancer-Center Spending on Advertising in the United States, 2005 to 2014” published in 2016 in JAMA Internal Medicine,
- Hospital Ad. Media
- Hospital Ad. Article

Data?

- The raw data for the analysis was purchased from Kantar Media.
- I can only show you a scrambled, fake data set

Reproducing the analysis

Think about reproducing the figure and the table in the article.

How will you summarize your data to plot the time series?

Figure. Trends in Cancer Center Advertising Spending by Media Channel Between 2005 and 2014

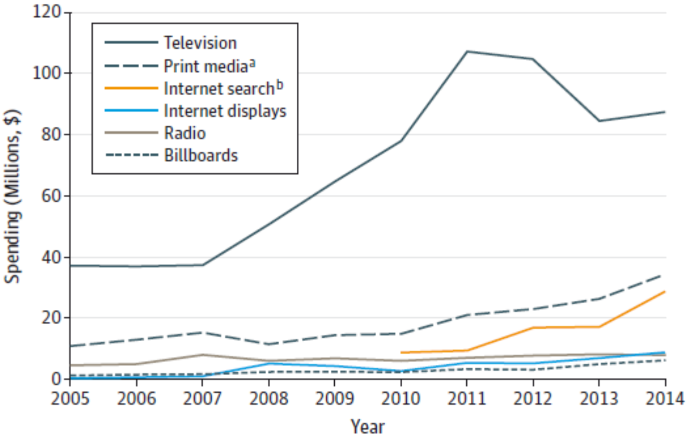


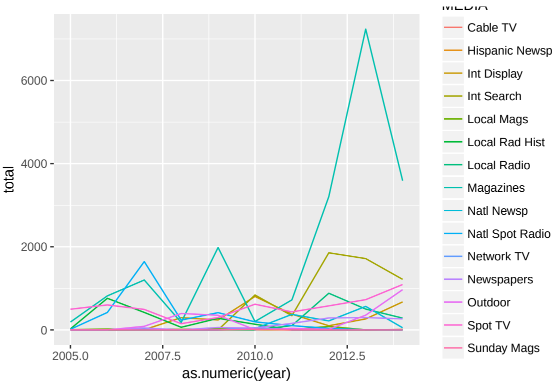
Figure 1:

The data set you need is something like this:

```
## # A tibble: 150 x 3
## # Groups:   MEDIA [?]
##       MEDIA  year total
##       <chr> <chr> <dbl>
##  1 Cable TV  2005  1.48
##  2 Cable TV  2006  0.00
##  3 Cable TV  2007  0.00
##  4 Cable TV  2008  0.00
##  5 Cable TV  2009  0.00
##  6 Cable TV  2010 21.12
##  7 Cable TV  2011  0.00
##  8 Cable TV  2012  0.00
##  9 Cable TV  2013  0.00
## 10 Cable TV  2014 11.53
## # ... with 140 more rows
```

Using ggplot2, creating a time series is now easy.

```
# 1 MEDIA-wise total
Transformed_data %>%
  ggplot(aes(x = as.numeric(year), y = total)) +
  geom_line(aes(color = MEDIA))
```



Problem?

Problem

The categorical values of `MEDIA` are not exatly what we want.

We would need to create another categorical variable by combining many values of `MEDIA` into one value. E.g. `Local Mags` and `Magazines` into `Print Media`.

Two types of data sets

- The raw data set you see is called **wide**. Note that `year` does not appear as a variable in the data set.
- Strange? In `Transformed_data`, however, `year` is a variable.
- For our purpose of creating the time series plot, a form of **thin** data set is more useful. Within the thin data set, we can treat `year` and `Expenditure` as variables.
- Both forms are useful for different occasions.

Re-creating the data table.

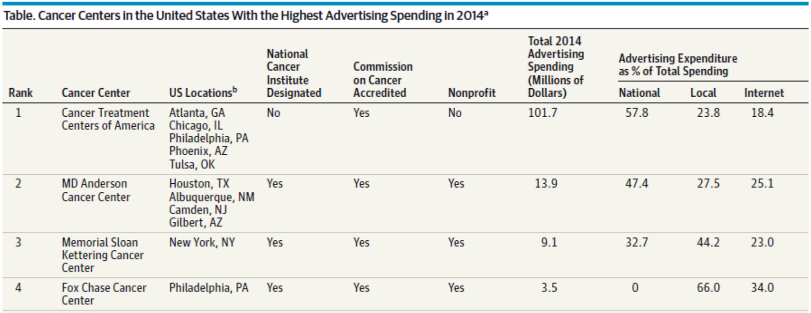


Table. Cancer Centers in the United States With the Highest Advertising Spending in 2014[a]

							Advertising Expenditure as % of Total Spending		
Rank	Cancer Center	US Locations[b]	National Cancer Institute Designated	Commission on Cancer Accredited	Nonprofit	Total 2014 Advertising Spending (Millions of Dollars)	National	Local	Internet
1	Cancer Treatment Centers of America	Atlanta, GA Chicago, IL Philadelphia, PA Phoenix, AZ Tulsa, OK	No	Yes	No	101.7	57.8	23.8	18.4
2	MD Anderson Cancer Center	Houston, TX Albuquerque, NM Camden, NJ Gilbert, AZ	Yes	Yes	Yes	13.9	47.4	27.5	25.1
3	Memorial Sloan Kettering Cancer Center	New York, NY	Yes	Yes	Yes	9.1	32.7	44.2	23.0
4	Fox Chase Cancer Center	Philadelphia, PA	Yes	Yes	Yes	3.5	0	66.0	34.0

Figure 2:

You need

1. *metadata* (or codebook): a set of data that describes and gives information about other data
2. to *combine* two datasets into one
3. to *filter* cases whose year is 2014 (in the thin form), or *select* the variable with name `DOI (000) 2014` (in the wide form)
4. to create a categorical variable that to tell whether a given MEDIA is National, Local or Internet (Perhaps by creating another metadata and combining)
5. to summarize all expenditure, for each of Cancer Center. (*summarize* expenditure *grouped by* Center)
6. *arrange* from the largest to the smallest