

Data Structures & Visualization

BC2406 UNIT 4

Based on Chew C. H. (2019) textbook: Analytics, Data Science and AI. Vol 1., Chap 4.

Some Sources of Data

- Point-of-sale data
- Database Tables
- Sensor Readings
- Blood test results
- E-commerce server application logs
- Direct mail response records
- Call-centre records, including memos written.
- Survey response data

Data Structures

Records Data Format

Each column represents a variable (aka attribute).

custid	sex	is.employed	income	marital.stat	health.ins	housing.type
2068	F	NA	11300	Married	TRUE	Homeowner free and clear
2073	F	NA	0	Married	TRUE	Rented
2848	M	TRUE	4500	Never Married	FALSE	Rented
5641	M	TRUE	20000	Never Married	FALSE	Occupied with no rent
6369	F	TRUE	12000	Never Married	TRUE	Rented
8322	F	TRUE	180000	Never Married	TRUE	Homeowner with mortgage/lo
8521	M	TRUE	120000	Never Married	TRUE	Homeowner free and clear

Each row represents a record (aka observation).

Transactions Data Format

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

5

Variable (aka Data Column) Structure

- Continuous (aka real number)
- Categorical
 - Nominal
 - Ordinal
- Text

6

Continuous Variables

- Real number that can take on any value within a certain interval.
- In practice, some integer variables are treated as continuous for purpose of analysis e.g. price.
- In R, any numbers are typically treated as numeric (i.e. real number). Use the suffix L to treat as integer.

```
> x <- 5
> class(x)
[1] "numeric"
> str(x)
num 5
> y <- 5L
> class(y)
[1] "integer"
> str(y)
int 5
```

7

Nominal Variable

- Categorical variable without any order.
- Examples: Gender, Country,...
- R may/may not recognize a variable as nominal. Explicitly force R to treat a variable as nominal with the factor() function.

```
# month and day treated as integer. Convert to categorical.
flights.dt$month <- factor(flights.dt$month)
flights.dt$day <- factor(flights.dt$day)
summary(flights.dt$month)
summary(flights.dt$day)
```

8

Ordinal Variable

- Categorical variable with order.
- Examples: A/B/C/D/E; Gold/Silver/Bronze.
- R may/may not recognize a variable as ordinal. Explicitly force R to treat a variable as ordinal with the `ordered = T` option in `factor()` function.

10

Data Type determine Summary Table and Visualization

- `Table()` function will list the numerical summary in a table, depending on variable type
- Plots and other charts will use variable type to plot the data. Depending on variable type, some charts cannot be plotted (i.e. error) or chart can be plotted but does not make sense (e.g. scatterplot of categorical variable).

17

Visualization with ggplot2 package

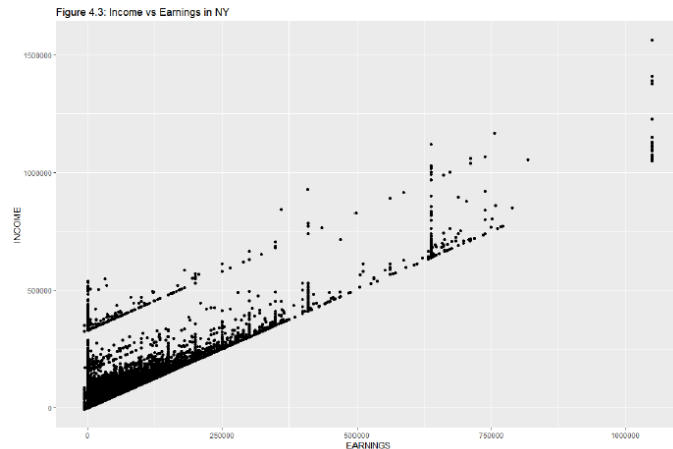
- ggplot2 builds chart by adding layers:
 - First layer defines the dataset and (optional) axis variables
 - Each subsequent layers adds one element to the chart.
- Refer to Chapter 3 of eTextbook R for Data Science: <http://r4ds.had.co.nz/data-visualisation.html>
 - Written by creator of ggplot2 package.
- See ggplot2 cheatsheet in NTUlearn main site.
 - Shows common charts for different variable types.

Visualization with ggplot2

18

19

```
library(ggplot2)
ggplot(data = ins.dt[STATE == 'NY'], aes(x = EARNINGS, y = INCOME)) + geom_point() +
  labs(title = "Figure 4.3: Income vs Earnings in NY")
```



Layer 1: ggplot(), Layer 2: geom_point(), Layer 3: labs()

20





Faceting with ggplot2 package

- Add panels to a chart

Faceting

Facets divide a plot into subplots based on the values of one or more discrete variables.

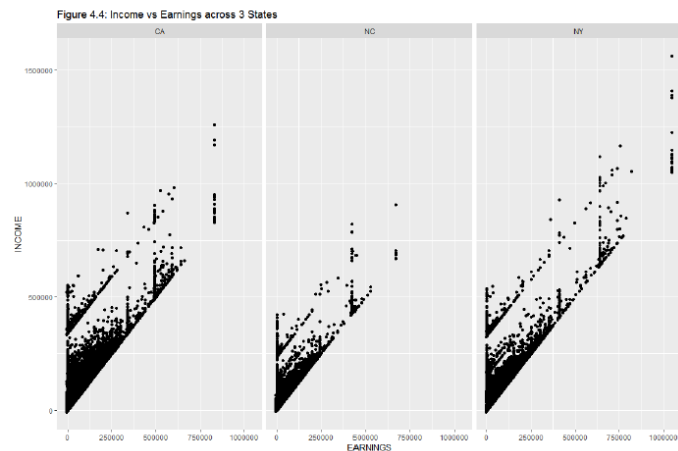
```
t <- ggplot(mpg, aes(cty, hwy)) + geom_point()
```

	t + facet_grid(. ~ fl) facet into columns based on fl
	t + facet_grid(year ~ .) facet into rows based on year
	t + facet_grid(year ~ fl) facet into both rows and columns
	t + facet_wrap(~ fl) wrap facets into a rectangular layout

Source: ggplot2 Cheatsheet

21

```
ggplot(data = ins.dt, aes(x = EARNINGS, y = INCOME)) + geom_point() +
  facet_grid(. ~ STATE) +
  labs(title = "Figure 4.4: Income vs Earnings across 3 States")
```



Layer 1: ggplot(), Layer 2: geom_point(), Layer 3: facet_grid(), Layer 4: labs()

22

eTextbook Chapter for Learning ggplot2

Chapter 3 of eTextbook R for Data Science:

<http://r4ds.had.co.nz/data-visualisation.html>

24

Summary

- Data Format as a record data.
- Variables data types impact analysis.
- Good Data Exploration, summary table and Visualization facilitated by correct data type.
- Visualization:
 - Data exploration for own self (focus on speed, probably discard after view)
 - Communication of findings to others (focus on the message and clarity, probably re-use elsewhere and on-record.)