Introduction to Modeling and Machine Learning

Code

Instructor - Soumya Mukherjee

Content Credit- Dr. Matthew Beckman and Olivia Beck

July 28, 2023

We have spent most of our time on two subjects in the context of EDA:

1. Data wrangling: getting from the data you are given to the “glyph-ready” data that you need to make a graphic or some other mode to guide interpretation of the data.
2. Data visualization

Visualization works well with 1-3 variables, and in some situations can work with more variables.

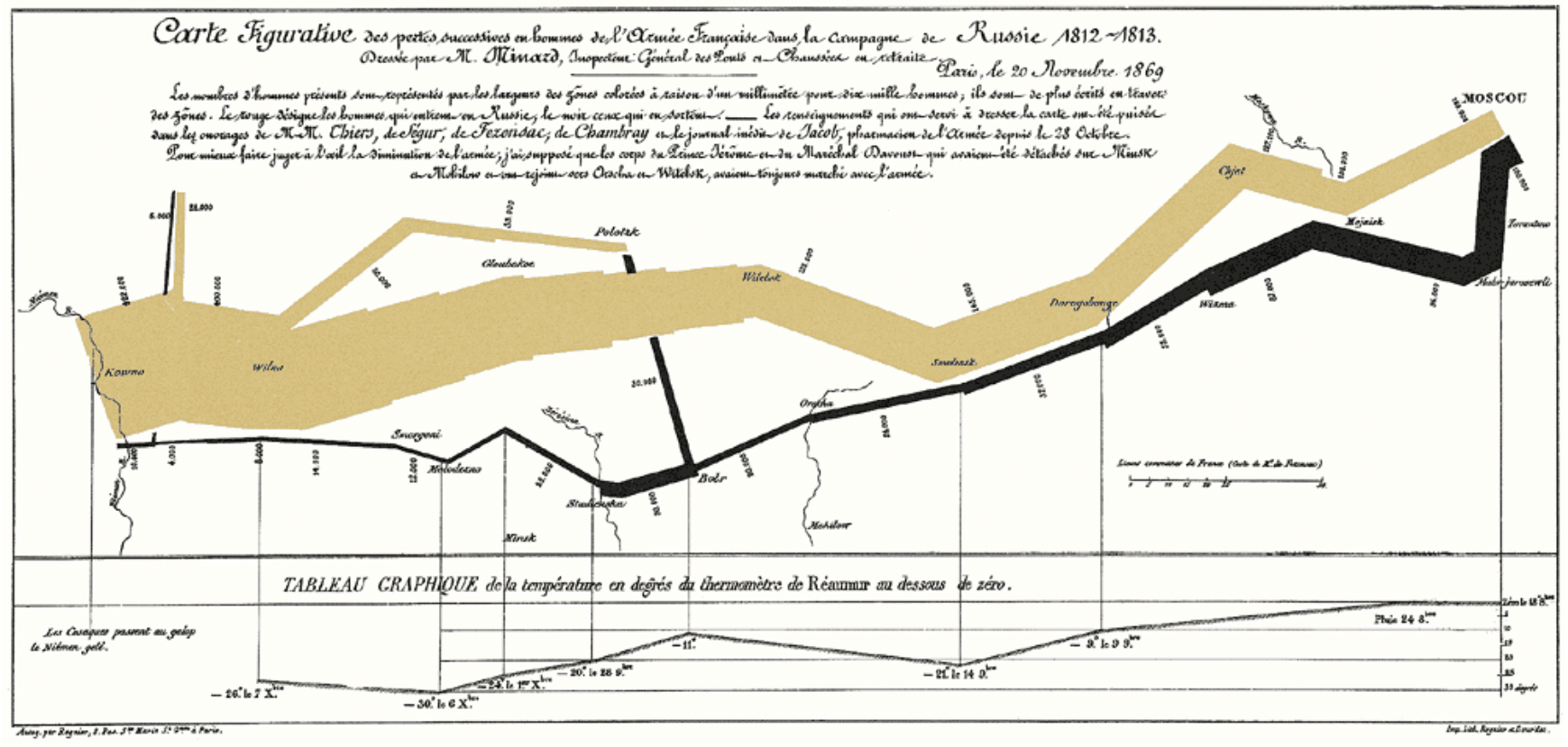
A multivariable graphic (& review)

In 1812, Napoleon Bonaparte attempted to invade Russia in what is known as the **French invasion of Russia** or the **Russian Campaign**. “Napoleon’s invasion of Russia is one of the best studied military campaigns in history and is listed among the most lethal military operations in world history. It is characterized by the massive toll on human life: in less than six months nearly a million soldiers and civilians died.” (source (https://en.wikipedia.org/wiki/French_invasion_of_Russia))

Minard’s Map of French casualties

Can you identify one or more of the following elements in the graphic?

- Glyphs
- Aesthetics
- Scale
- Guide



source (http://www.edwardtufte.com/tufte/graphics/minard_lg.gif)

Here is is in English:

https://en.wikipedia.org/wiki/French_invasion_of_Russia
https://en.wikipedia.org/wiki/French_invasion_of_Russia (https://en.wikipedia.org/wiki/French_invasion_of_Russia)
source (https://www.datavis.ca/papers/jebs.pdf) This paper also has alternative ways to map this data.

We can plot this in `ggplot`

Hide

```
library(gridExtra)
```

```
Warning: package ‘gridExtra’ was built under R version 4.2.3
Attaching package: ‘gridExtra’

The following object is masked from ‘package:dplyr’:

    combine
```

Hide

```
library(maps)
```

```
Warning: package ‘maps’ was built under R version 4.2.3
Attaching package: ‘maps’

The following object is masked from ‘package:purrr’:

    map
```

Hide

```
troops <- read.table("troops.txt", header = T)
cities <- read.table("cities.txt", header = T)
temps <- read.table("temps.txt", header = T)
temps$date <- as.Date(strptime(temps$date,"%d%b%Y"))

borders <- data.frame(map(.x = "world", .f = ~ . ,  xlim = c(10,50), ylim = c(40, 80), plot = F)[c("x", "y")])

xlim <- scale_x_continuous(limits = c(24, 39))

# Troop Survival
march <-
ggplot(cities, aes(x = long, y = lat)) +
geom_path(aes(size = survivors, colour = direction, group = group), data = troops) +
theme(legend.position = "none") +
geom_point() +
 geom_text(aes(label = city), hjust = 0, vjust = 1, size = 4) +
 # scale_size(to = c(1, 10)) +
 scale_colour_manual(values = c("tan","black")) +
ggtitle("Carte figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813", subti
tle = "[English:] Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812–1813. \n Dr
awn by M. Minard, Inspector General of Bridges and Roads (retired). Paris, November 20, 1869.")
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
Please use `linewidth` instead.
```

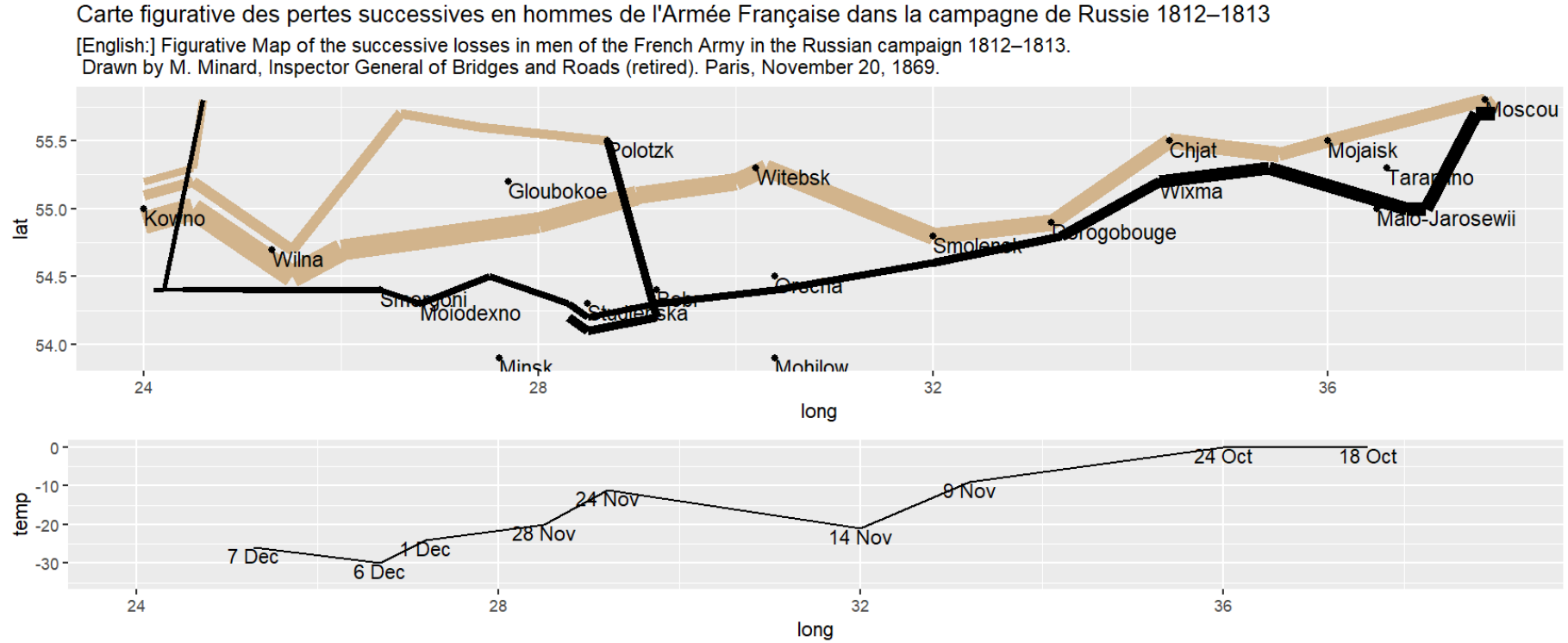
Hide

```
 # Temperature
 temp <-
   qplot(long, temp, data = temps, geom = "line") +
   geom_text(aes(label = paste(day, month)), vjust = 1) +
   xlim + ylim(c(-35, 0))
```

```
Warning: `qplot()` was deprecated in ggplot2 3.4.0.
```

Hide

```
grid.arrange(march, temp, nrow = 2, heights = c(3, 1.5))
```



I don't blame you if you prefer Charles Minard's version to the `ggplot2` attempt in this case

p.s. I don't take credit for all of the original code; I did the same thing I often recommend you do:

- start with working code that does something close (https://github.com/slygent/statlearningpres/blob/master/minard.r (https://github.com/slygent/statlearningpres/blob/master/minard.r))
- tweak it until it works for your specific purpose
- even fancier version: https://github.com/andrewheiss/fancy-minard (https://github.com/andrewheiss/fancy-minard)

Exploration with many variables?

- If we need to relate more variables, a visualization may not suffice
- Statistical/Mathematical representations are an important tool:
    - model formulas, e.g. `lm(Y ~ X1 + X2 + X3)`

- lots of different types of models

Side note on formulas

response variable ~ predictor1 + predictor 2 + ……

- Can handle both numeric (`as.numeric()`) and categorical (`as.factor()`) variables

More advanced:

- transformations of variables: log(predictor1) or I(predictor1^2)
- interactions: predictor1 * predictor 2
- conditional: response variable ~ predictor1 | predictor 2
- random effects: (1| predictor1)
- remove the intercept: y ~ -1 + predictor1

Side note about linear regression

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \epsilon_i$$

- y_i = response variable for data point i
- $x_{1,i}$ = 1st predictor variable for data point i
- $x_{2,i}$ = 2nd predictor variable for data point i
- … = we can add as many predictors as we would like
- $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

We are attempting to estimate $\beta_0, \beta_1, \beta_2, \ldots$.

Our statistical test is:

$H_0 : \beta_j = 0$ (the jth predictor is **NOT** related to the response variable)

$H_1 : \beta_j \neq 0$(predictor is related to the response variable)

Interpretation of β_j: for ever one unit increase in x_j, the response variable changes β_j amount HOLDING ALL OTHER VALUES CONSTANT.

Example: Child carseat sales

Source: James, Witten, Hastie, & Tibshirani (2021). *Introduction to Statistical Learning with applications in R*. New York: Springer.

Research Question

What is the relationship between store & community characteristics and carseat sales?

Britax Frontier Car Seat

First look at `Carseats` data

Hide

```
# data intake from ISLR2 package
data(Carseats, package = "ISLR2")

glimpse(Carseats)
```

```
Rows: 400
Columns: 11
$ Sales       <dbl> 9.50, 11.22, 10.06, 7.40, 4.15, 10.81…
$ CompPrice   <dbl> 138, 111, 113, 117, 141, 124, 115, 13…
$ Income      <dbl> 73, 48, 35, 100, 64, 113, 105, 81, 11…
$ Advertising <dbl> 11, 16, 10, 4, 3, 13, 0, 15, 0, 0, 9,…
$ Population  <dbl> 276, 260, 269, 466, 340, 501, 45, 425…
$ Price       <dbl> 120, 83, 80, 97, 128, 72, 108, 120, 1…
$ ShelveLoc   <fct> Bad, Good, Medium, Medium, Bad, Bad, …
$ Age         <dbl> 42, 65, 59, 55, 38, 78, 71, 67, 76, 7…
$ Education   <dbl> 17, 10, 12, 14, 13, 16, 15, 10, 10, 1…
$ Urban       <fct> Yes, Yes, Yes, Yes, Yes, No, Yes, Yes…
$ US          <fct> Yes, Yes, Yes, Yes, No, Yes, No, Yes,…
```

Hide

```
# shorten several variable names to investigate
Carseats %>%
  rename(CompP = CompPrice, Ads = Advertising,
         Pop = Population, Shelf = ShelveLoc, Edu = Education) %>%
  head()
```

	Sales <dbl>	CompP <dbl>	Income <dbl>	Ads <dbl>	Pop <dbl>	Price <dbl>	Shelf <fctr>	Age <dbl>	Edu <dbl>
1	9.50	138	73	11	276	120	Bad	42	17
2	11.22	111	48	16	260	83	Good	65	10
3	10.06	113	35	10	269	80	Medium	59	12
4	7.40	117	100	4	466	97	Medium	55	14
5	4.15	141	64	3	340	128	Bad	38	13
6	10.81	124	113	13	501	72	Bad	78	16

6 rows | 1-10 of 11 columns

Building intuition for research questions

What is the relationship between store & community characteristics and car seat sales?

Response Variable (Y): carseat sales ($ thousands) at each store

1. Price relative to competitor's price is probably relevant.
2. Expect larger population to be associated with higher sales
3. Does education level matter?
4. Advertising efficiency?

Hide

```
# add relative price comparison
Carseats <-
  Carseats %>%
  mutate(rel_price = Price / CompPrice)

# fit regression model: lm(Y ~ X, data = [Name of data source])
regressMod <- lm(Sales ~ rel_price + Population + Education + Advertising, data = Carseats )

# recall: extract coefficient estimates from `regressMod` object using `$`
regressMod$coefficients
```

```
  (Intercept)     rel_price    Population     Education
 17.497352098 -10.907759402  -0.000139262  -0.055051345
  Advertising
  0.136911760
```

Interpreting the model?

Hide

```
summary(regressMod)
```

```
Call:
lm(formula = Sales ~ rel_price + Population + Education + Advertising,
    data = Carseats)

Residuals:
   Min     1Q Median     3Q    Max
-6.005 -1.470 -0.118  1.310  5.280

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.750e+01  8.728e-01  20.048  < 2e-16 ***
rel_price   -1.091e+01  6.673e-01 -16.345  < 2e-16 ***
Population  -1.393e-04  7.470e-04  -0.186    0.852
Education   -5.505e-02  4.051e-02  -1.359    0.175
Advertising  1.369e-01  1.651e-02   8.294 1.74e-15 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.108 on 395 degrees of freedom
Multiple R-squared:  0.4483,    Adjusted R-squared:  0.4427
F-statistic: 80.25 on 4 and 395 DF,  p-value: < 2.2e-16
```

Purposes for modeling

Statistical modeling generally serves to formalize relationships inherent in the data using efficient mathematical approximations. Some typical purposes for statistical modeling might include:

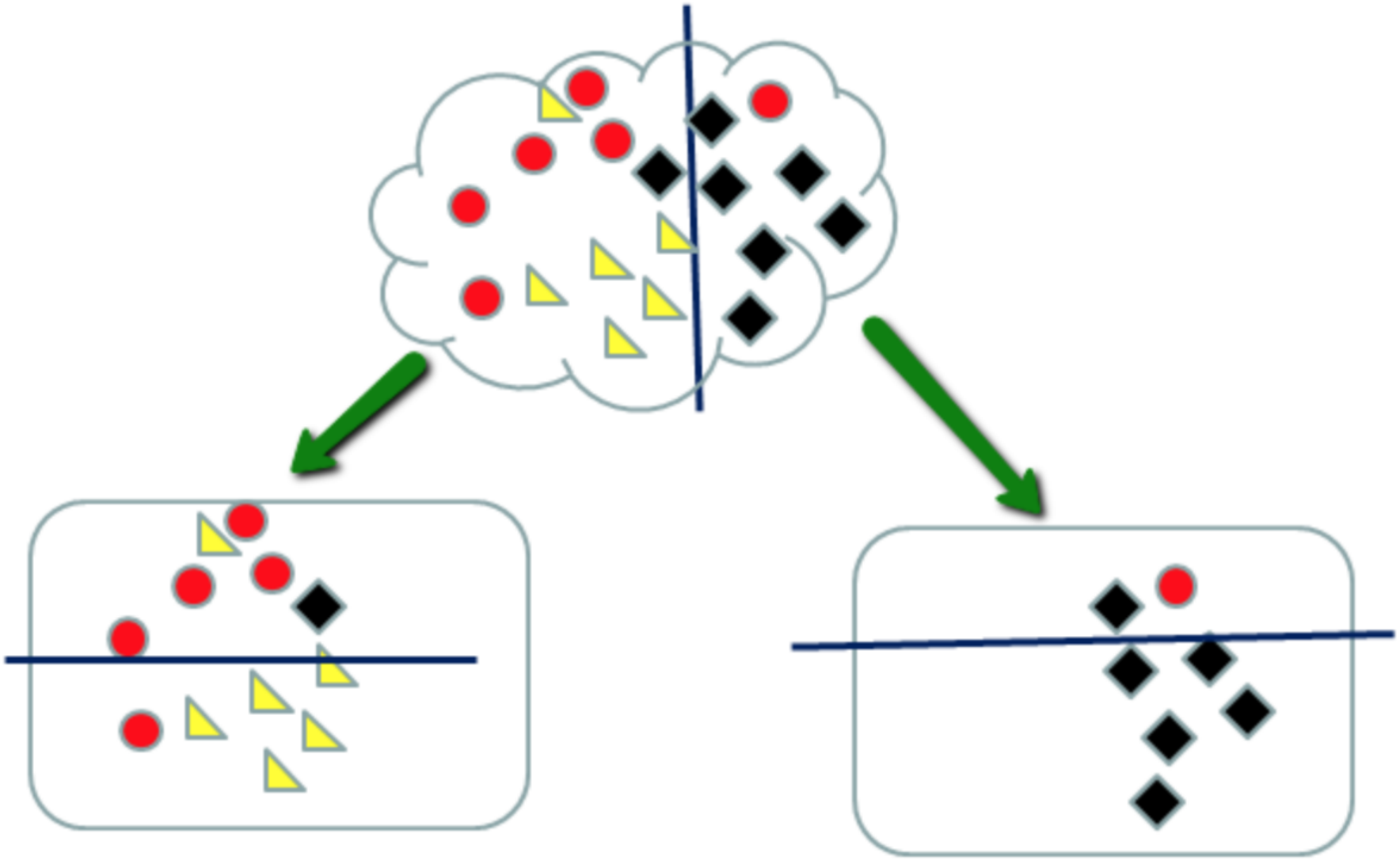
- descriptive / exploratory: *what can I learn about the structure of the data in front of me?*
- inferential: *what can these data tell me about a more general population they represent?*
- predictive: *what outcomes might we expect in the future based on what we have observed so far?*

These are not necessarily mutually exclusive, but it’s sometimes useful to consider them as distinct purposes.

How might these purposes be interpreted in the context of our regression model of carseat sales?

Another formalism: Regression trees

Idea of recursive partitioning



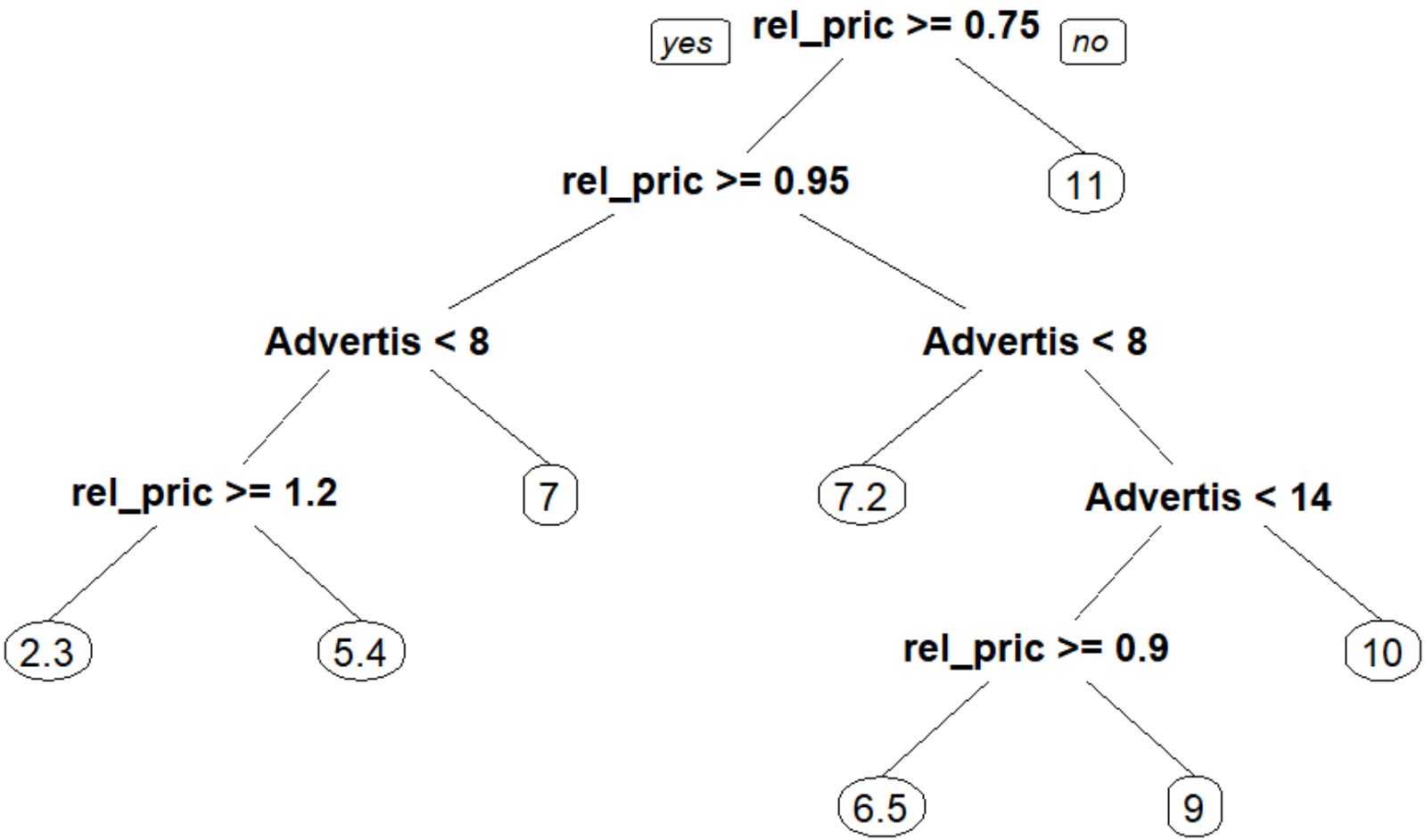
Source: Penn State World Campus STAT 555 (https://online.stat.psu.edu/stat555/node/100/ (https://online.stat.psu.edu/stat555/node/100/))

Another formalism: Regression trees

Hide

```
# use help to learn about `rpart` function... note the similarity of syntax
treeMod <- rpart(Sales ~ rel_price + Population + Education + Advertising, data = Carseats)

prp(treeMod)  # use help to learn about the `prp( )` function
```



Linear Regression vs Regression Trees

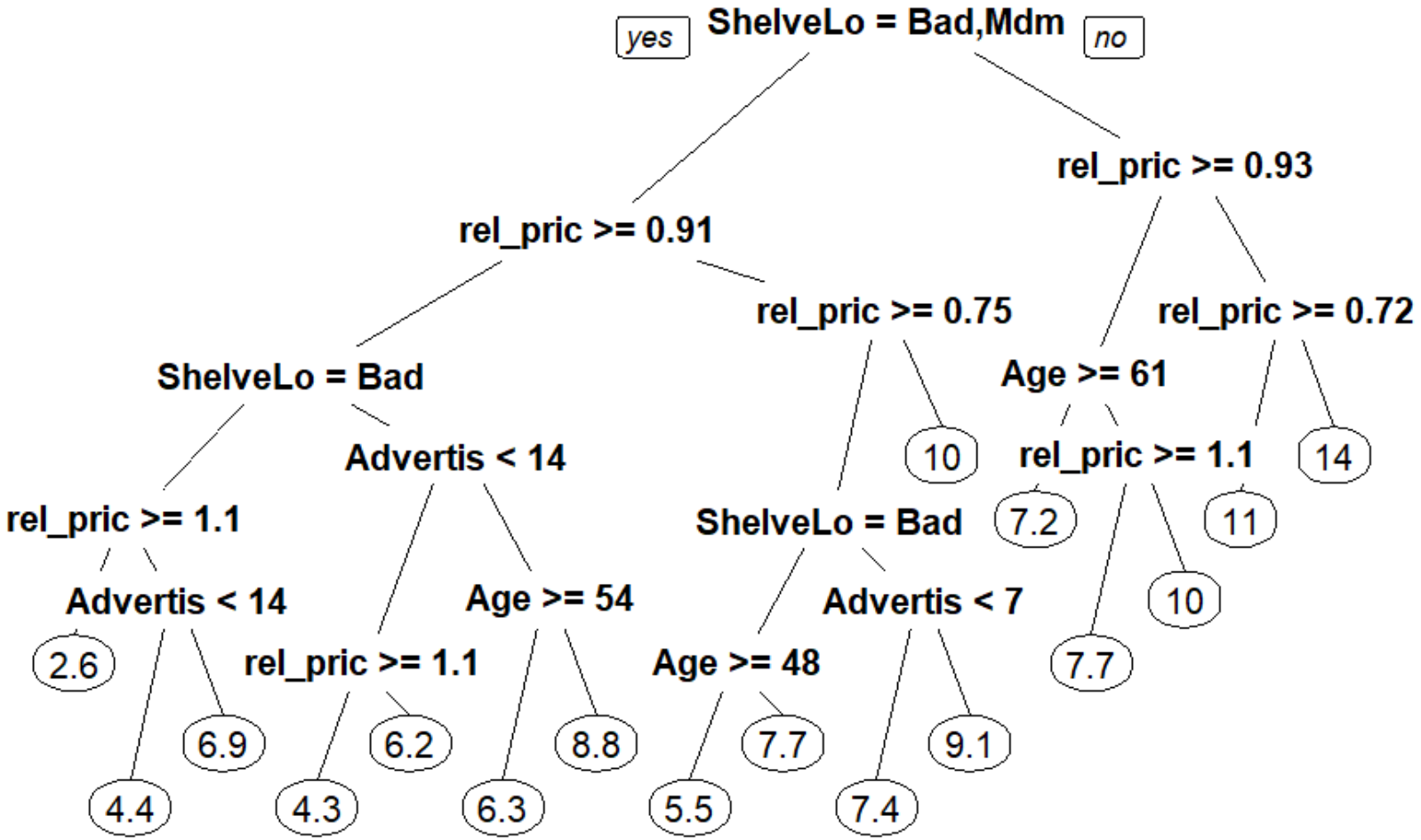
compare/contrast: syntax, output, interpretation, assumptions, conclusions, etc

Sales vs all other variables in the data

Hide

```
# what do you suppose the "dot" means here?
treeModFull <- rpart(Sales ~ . , data = Carseats)

prp(treeModFull)
```



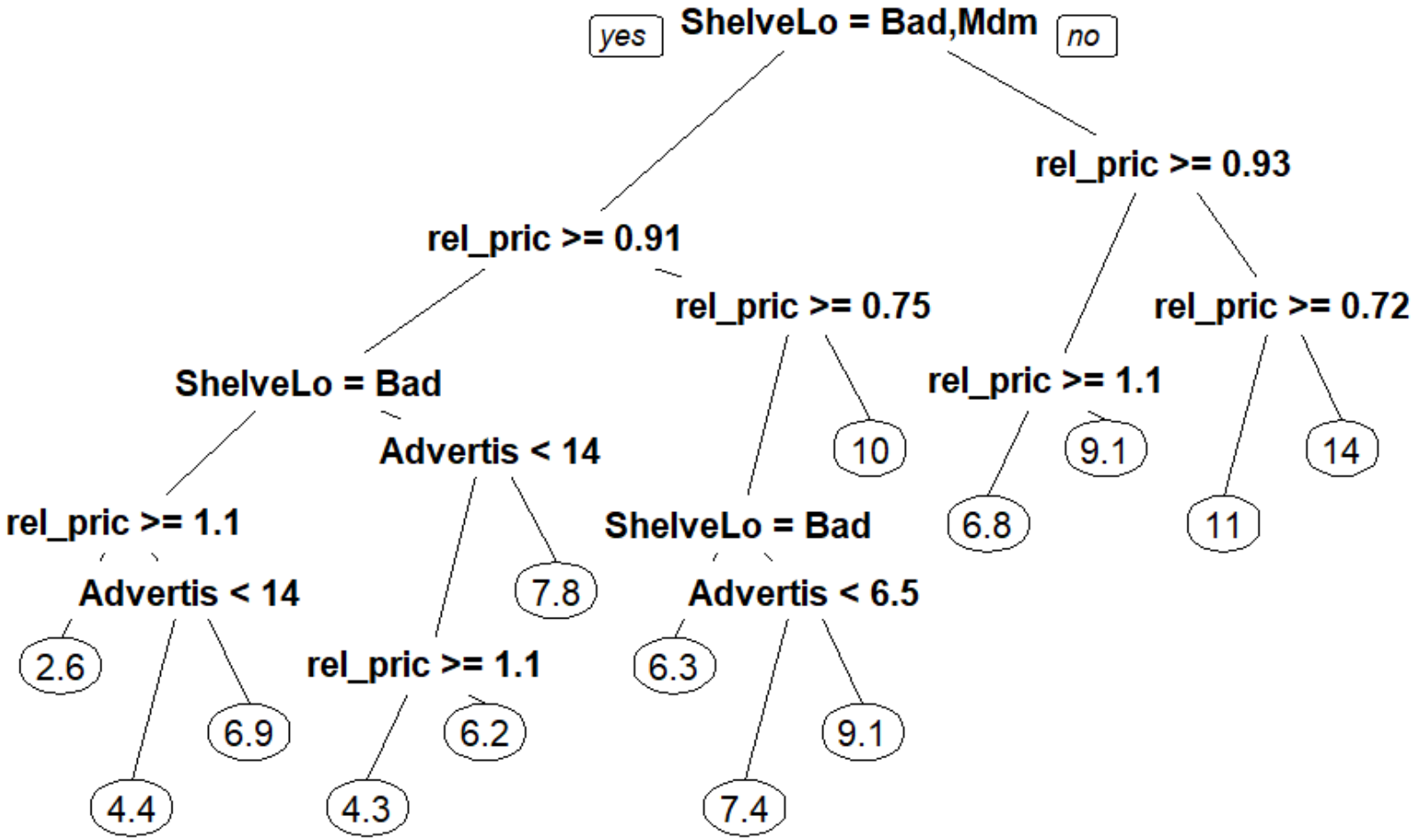
Shelf location?

Hide

```
# what do you suppose the "dot" means here?
treeModFinal <-
  Carseats %>%
  rpart(Sales ~ rel_price + Advertising + ShelveLoc + Population, data = . )

prp(treeModFinal)
```

```
Warning: Cannot retrieve the data used to build the model (so cannot determine roundint and is.binary for the variables).
To silence this warning:
    Call prp with roundint=FALSE,
    or rebuild the rpart model with model=TRUE.
```



Unsupervised learning intro

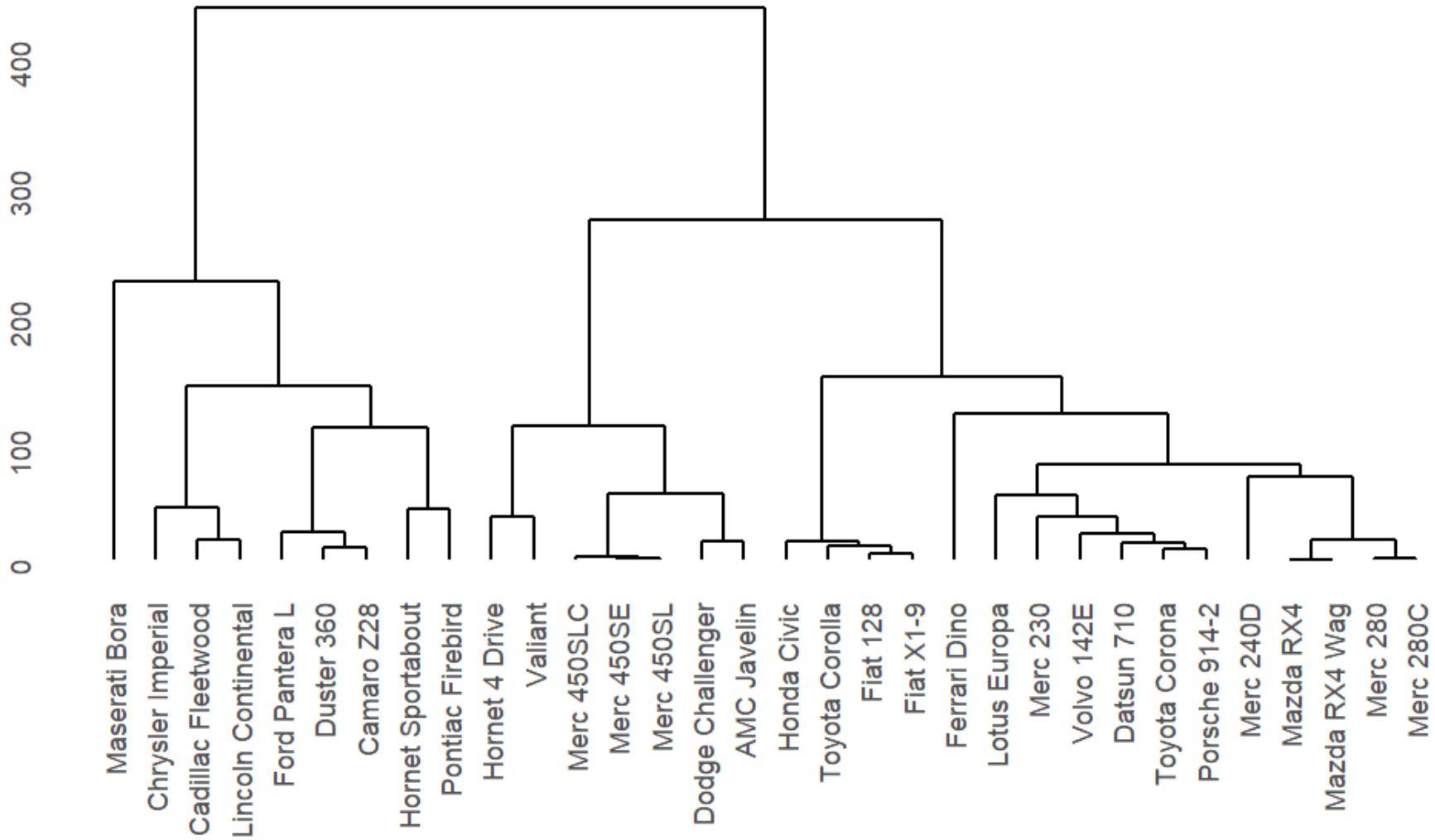
- If you do the opposite of our binary partitioning process, we call it hierarchical clustering
- Here we join elements based on their similarity to one another (according to available information)
- Why “Unsupervised”?

Hide

```
# compute distance matrix
Dists <- dist(mtcars)

# hierarchical cluster analysis
Dendrogram <- hclust(Dists)

# plot dendogram of cars
ggdendrogram(Dendrogram)
```



Some key concepts in Machine Learning (ML) / Statistical Learning

1. Supervised vs unsupervised learning
2. Model assumptions & nonparametric methods (e.g., Recursive partitioning)
3. Protect against “overfitting” the data
    - *cross-validation* tests out-of-sample performance to diagnose overfitting among other purposes

    - *bias-variance trade-off* balance model performance between overfitting and underfitting
    - *regularization* (e.g. Ridge Regression & LASSO) impose model constraints to prevent overfitting
4. Dimension reduction

Remarks

- **Goal**: characterize patterns or structure present in the data
- **Supervised** learning: values of both inputs (e.g. “X” variables) *and* outputs (e.g. “Y” variable) are used
    - frequently used in exploratory, inferential, & predictive applications
    - examples:
        - classification and regression trees (CART);
        - support vector machines;
        - simple linear regression
- **Unsupervised** learning: values of only the inputs (e.g. “X” variables) are used
    - commonly used for grouping/clustering similar observations
    - variable/dimension reduction (i.e. compressing the number of “inputs” in the data)
    - **Data mining**: often refers to unsupervised learning for exploratory purposes
- **Tools**: R has lots of useful packages for ML
    - `rpart` : used in the course notes today
    - `party` : demonstrated in the book
    - `e1071` , `nnet` , `randomForest` , `caret` , `gbm` , `SuperLearner` ,
    - other software have great tools too (e.g., python’s `scikit-learn` library)

Extra credit Activity (Due August 4, 9:59am)

- Activity: Street or Road (5% extra credit)