Lecture 5:
Line Charts and R Skills

February 26, 2018

Overview

Course Administration

Good, Bad and Ugly

Few, Chapter 10

Bar Charts in R

Course Administration

1. Rosa has graded problem sets – thank you
2. Still haven't set up grading format – apologies
3. Please be sure to
   - start early on your policy brief
   - book me early
4. Missing anything else from me?

Next Week's Good Bad and Ugly

Monday by 9 am. Earlier is ok.

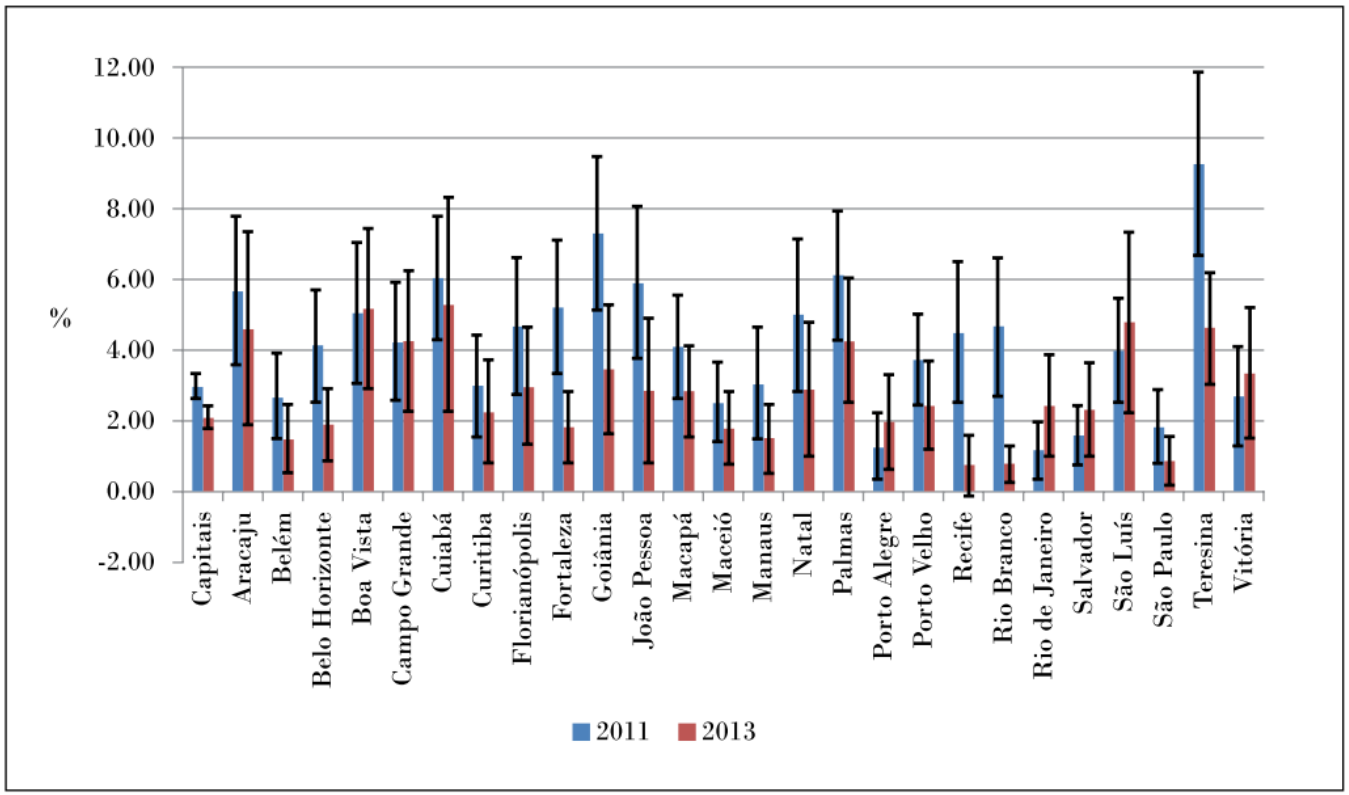
- Francisca Alba
- Julia Robertson
- Adam Brooks

This Week's Good Bad and Ugly

- Bruno Oliveira
- Gulfishan Khadim

Bruno's Example

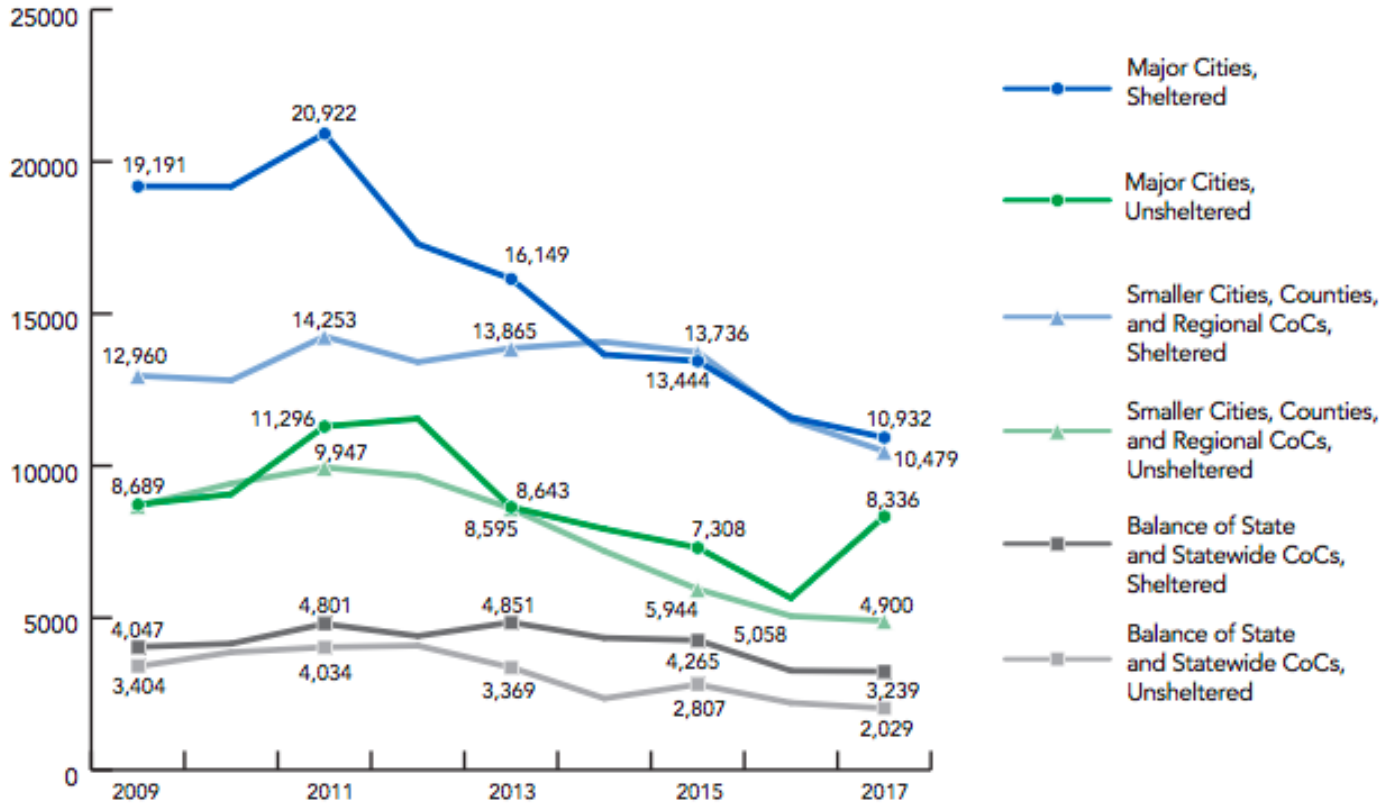
Graph 4 – Prevalence (%) among adults (≥18 years) who reported driving after abusive alcoholic consumption, by state capitals and the Federal District – Brazil, 2011-2013



Source: Surveillance System of Risk and Protective Factors for Chronic Diseases by Telephone Survey (*Vigitel*).
Note: Prevalence weighted and adjusted for the population existing in the year the survey was performed.

Gulfishan's Example

EXHIBIT 5.11: Veterans Experiencing Homelessness
By CoC Category and Sheltered Status, 2009–2017



Few:
Component Level Graph Design

Today

1. Primary Data Components
2. Secondary Data Components
3. Next Week: Non-Data Components

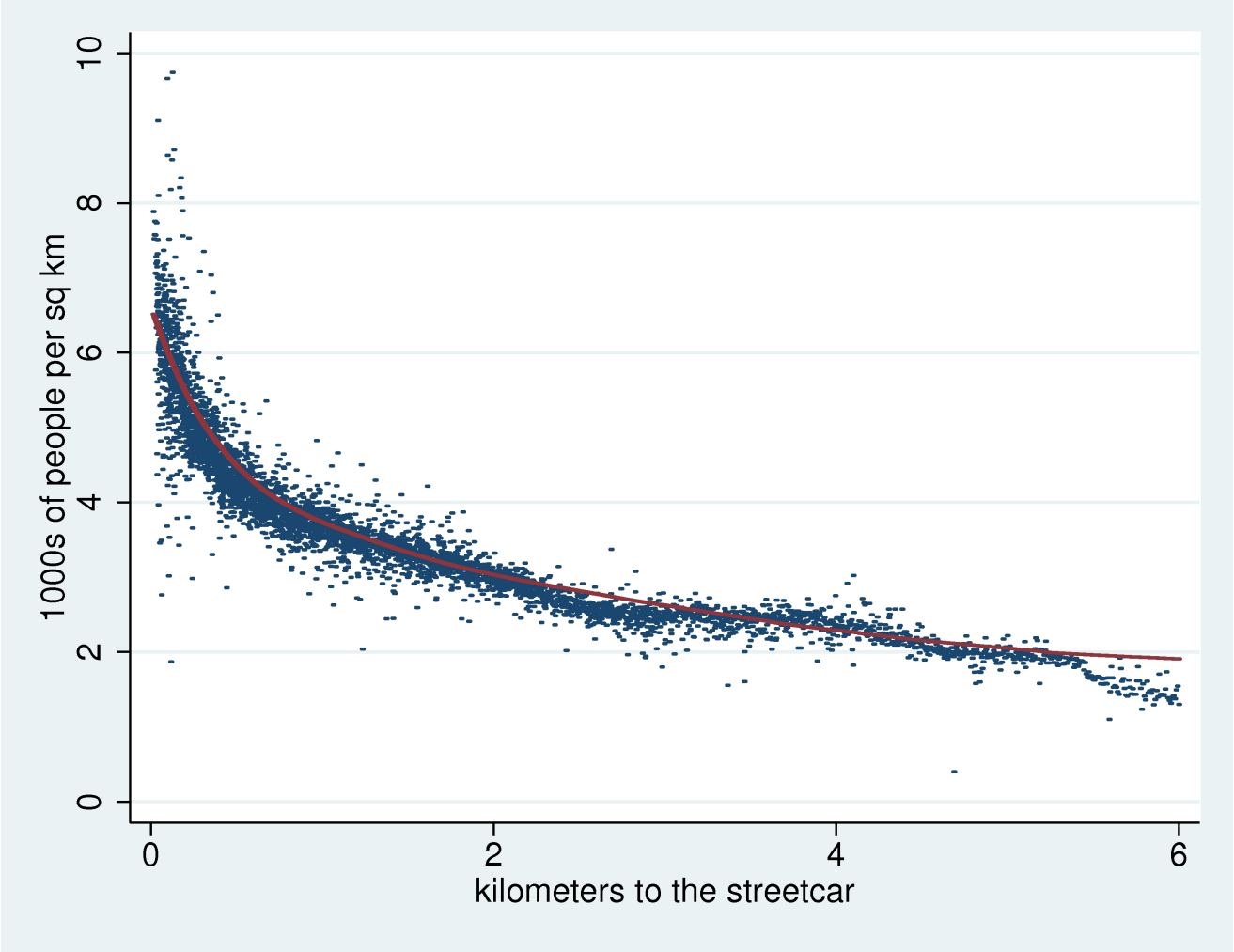
1. Primary Data Components

- Points
- Bars
- Lines
- Boxes
- And combinations thereof

Points

- Use colors to distinguish between points whenever possible
- Not in Few: Use summary statistics to make too many points legible
- Make open points as needed

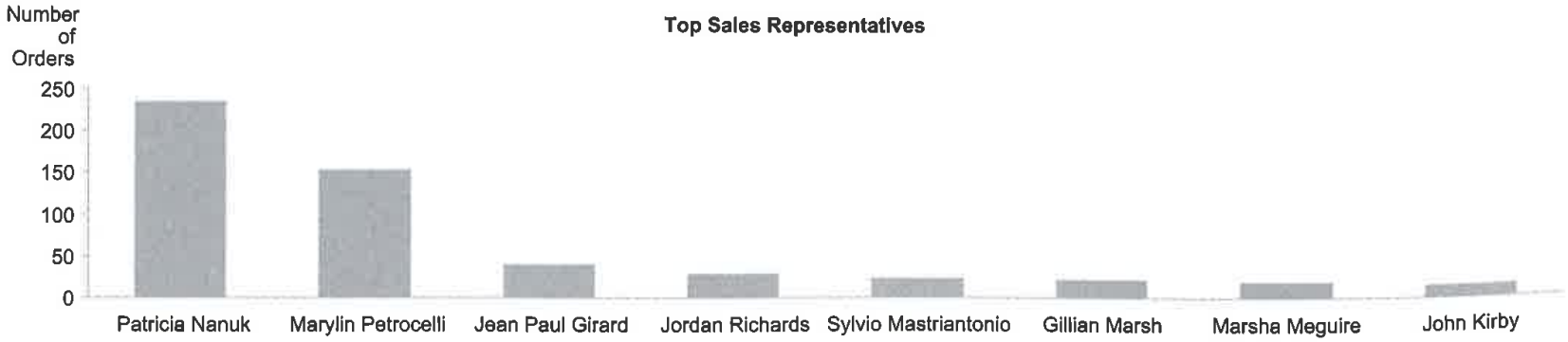
Using Summary Statistics: Streetcars in Los Angeles



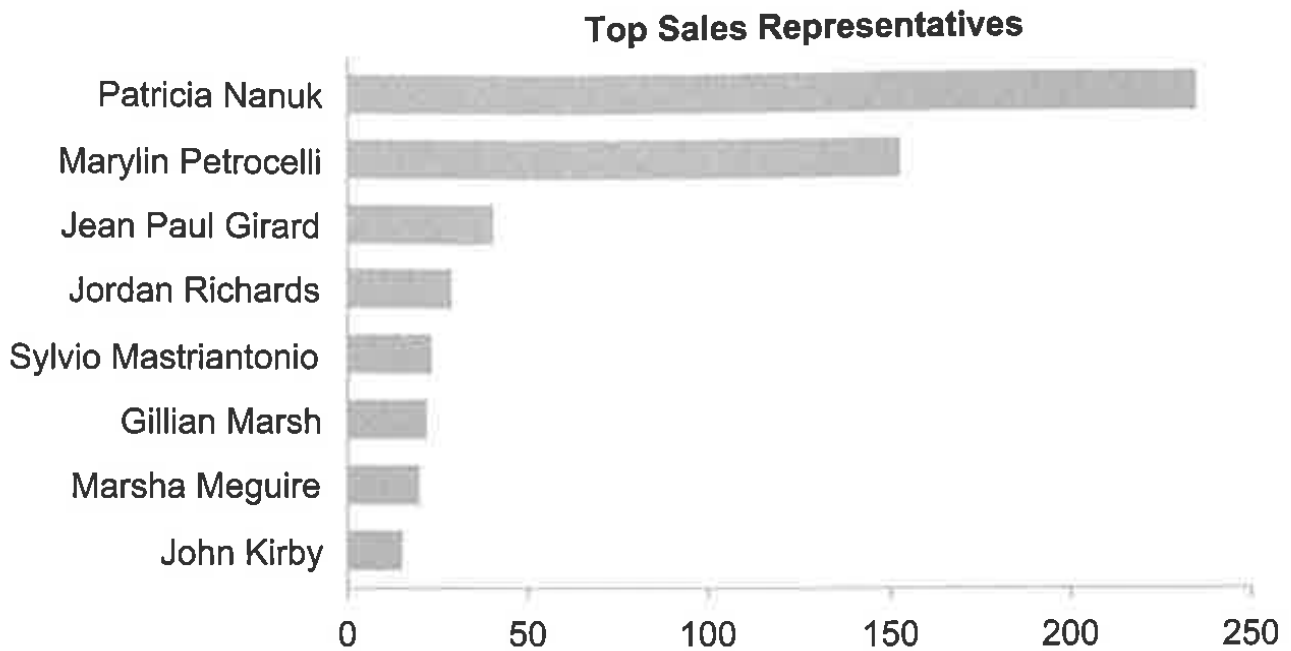
Bars

- Always start bars at zero
- If you don't want to start at zero, use a point or a dash
- If you have long category or label names, use horizontal bars
- With grouped bars, put groups together, but don't overlap
- Fill with the same hue unless you want to draw attention for some reason
- Use only one color per set of related values
- Don't put borders around bars without a very good reason

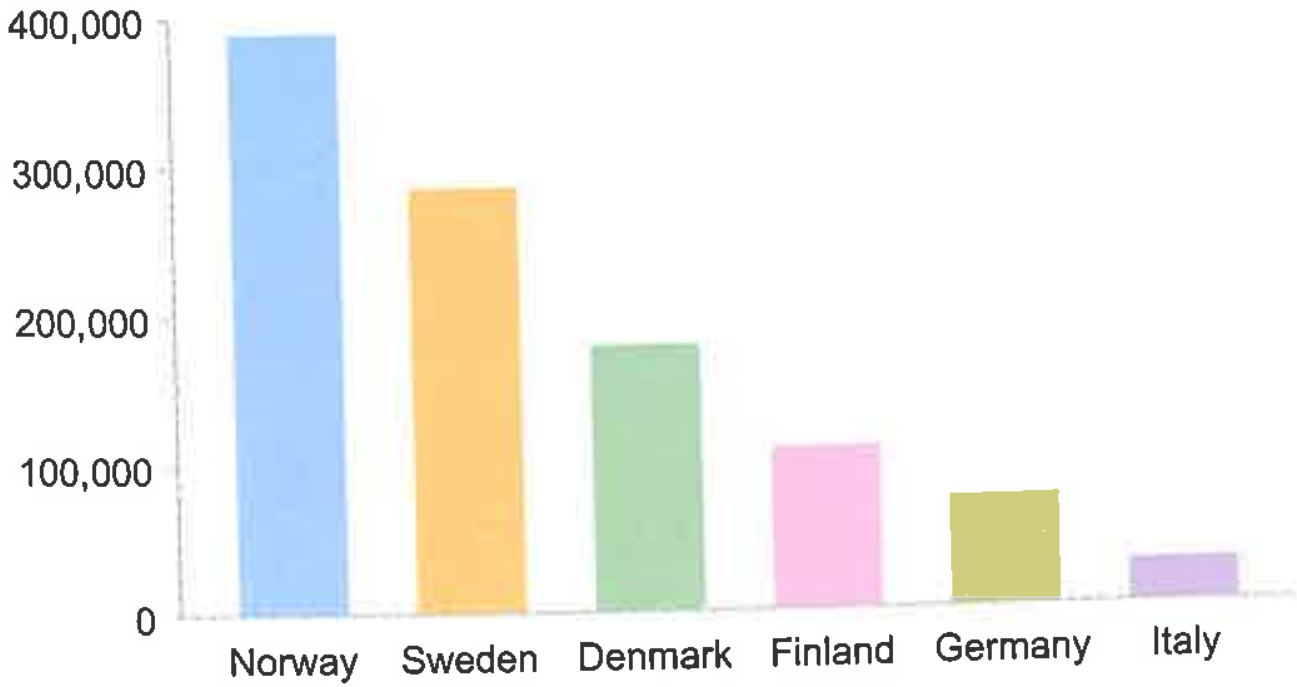
Why You May Prefer Horizontal Bars, 1 of 2



Why You May Prefer Horizontal Bars, 2 of 2



A Pretty But Bad Use of Color



Lines

- Make your lines distinguishable
- Marks along lines are hard to read
- Use color and thickness to distinguish lines when possible
- Indicate line height either with points or lines from axes

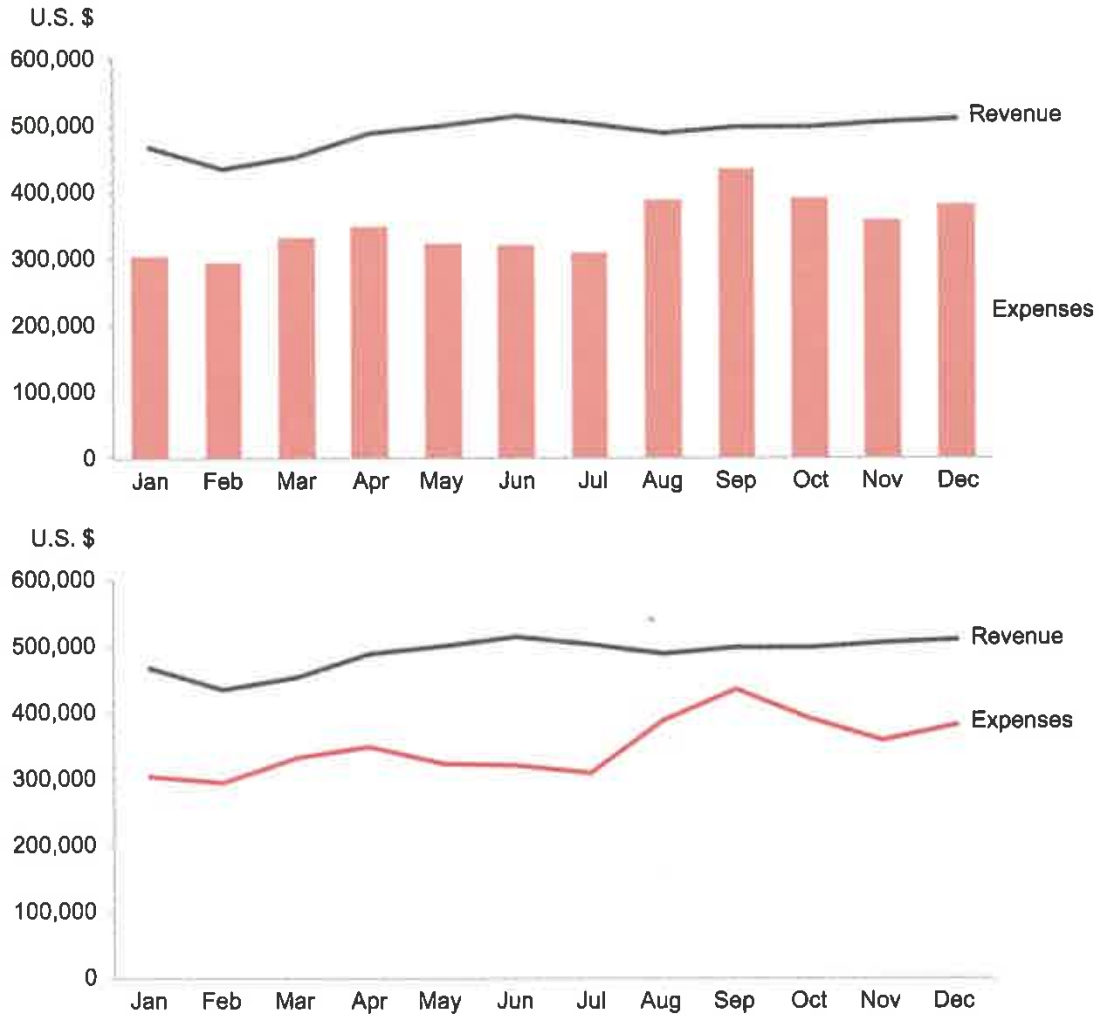
Boxes

- I just don't care for them
- Consider lines or shaded ares to show distributional points

Combinations of Plot Types

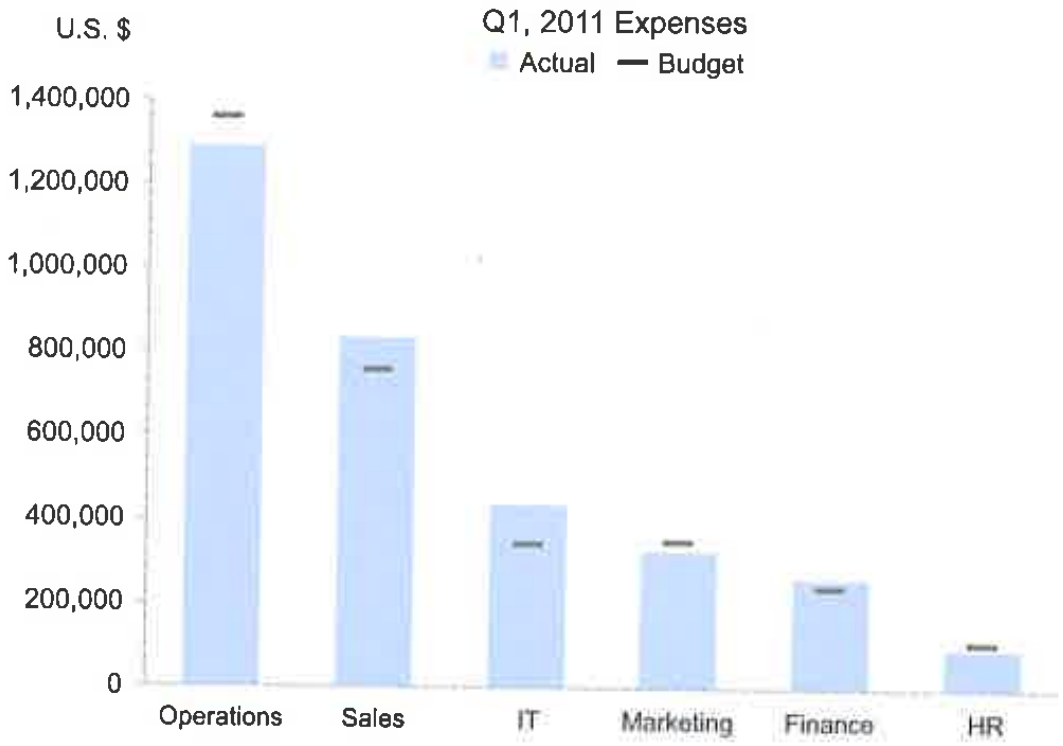
- Usually bar + line is not a good combination
- Consider instead bar + points
- For regression coefficients, bars + error markers or regions can be very useful

Bar with Line, Not so Fine

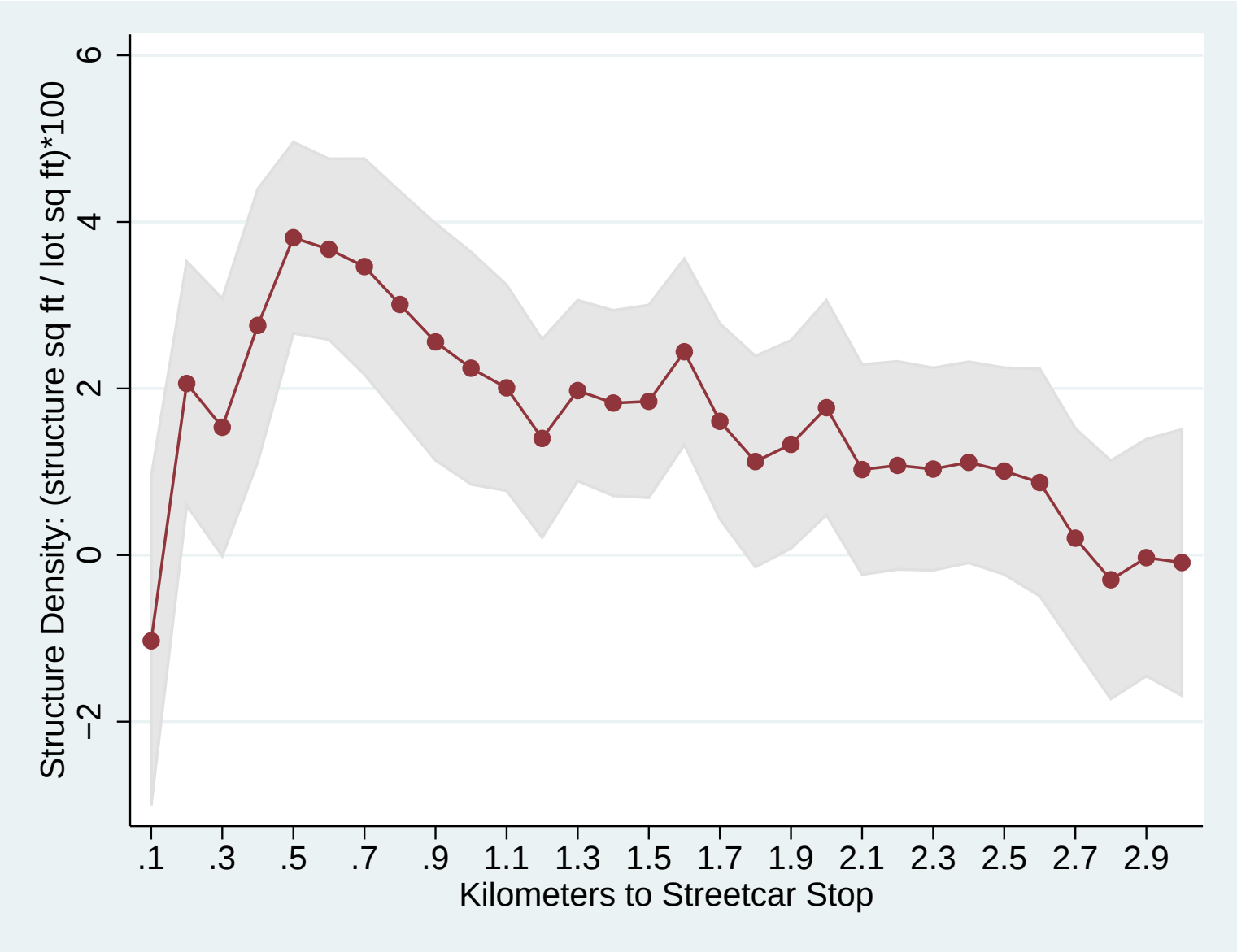


Bar with Point, An Improvement

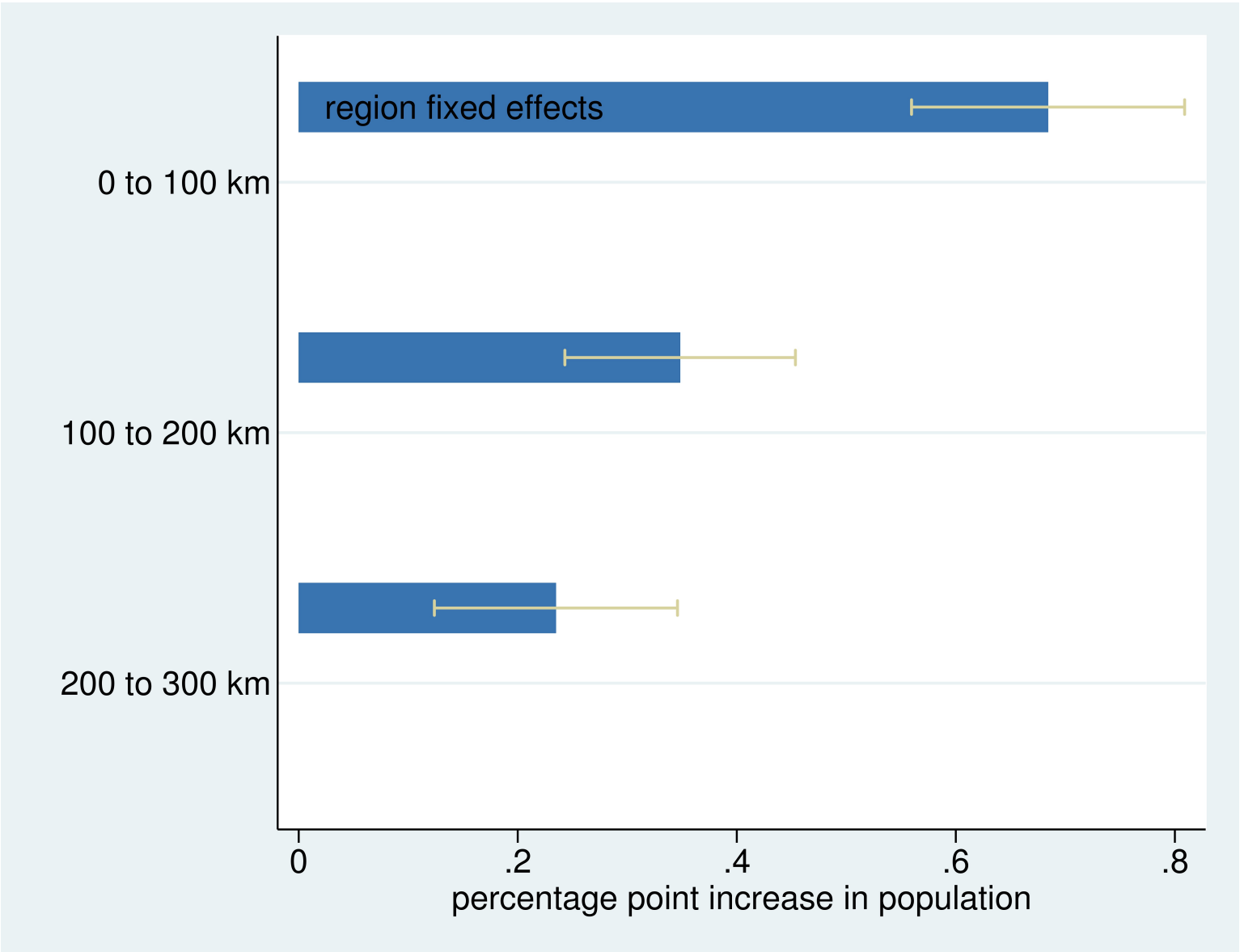
Highlights Actual Over Budgeted



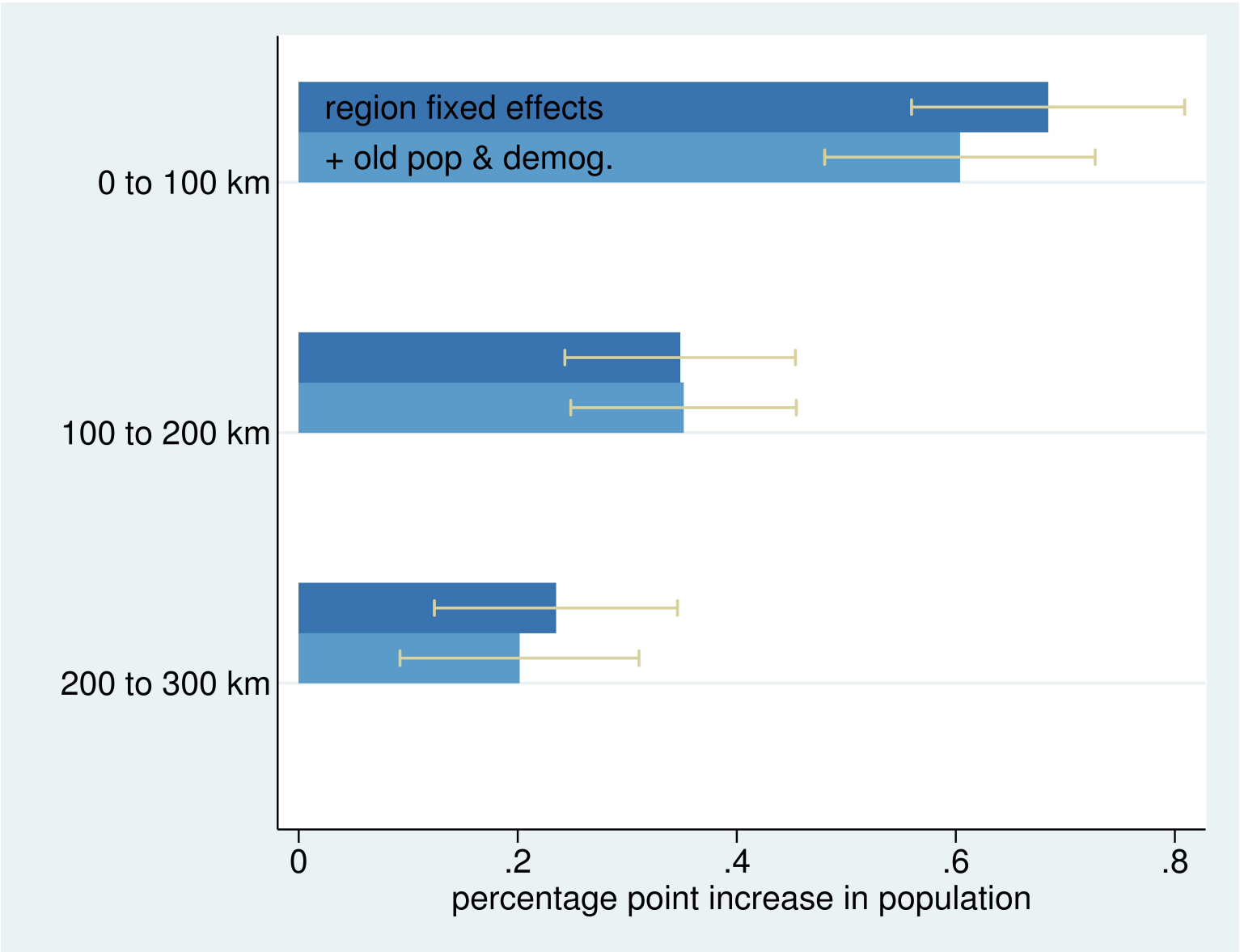
Points with Error Ranges



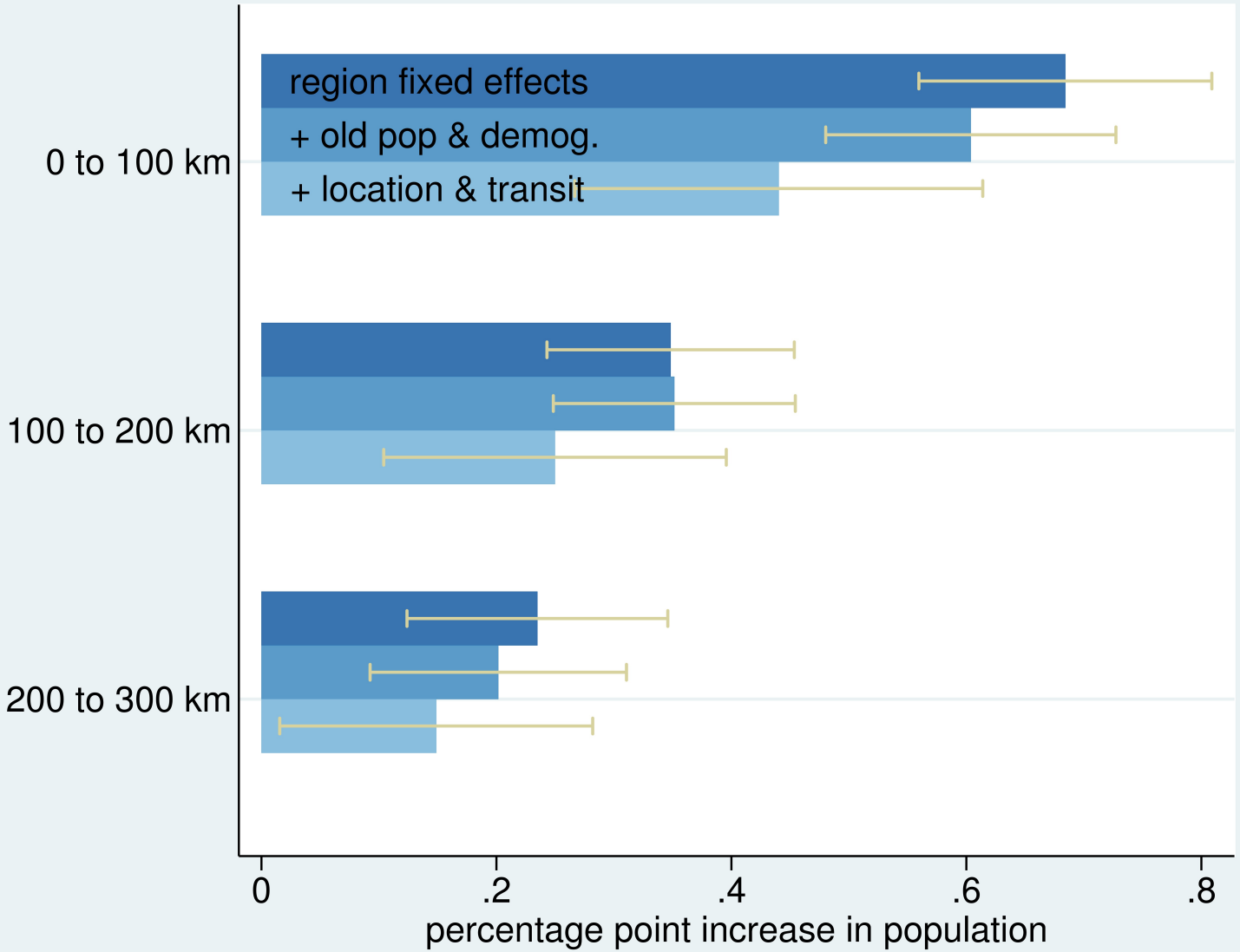
Bars with Error Bars, Building



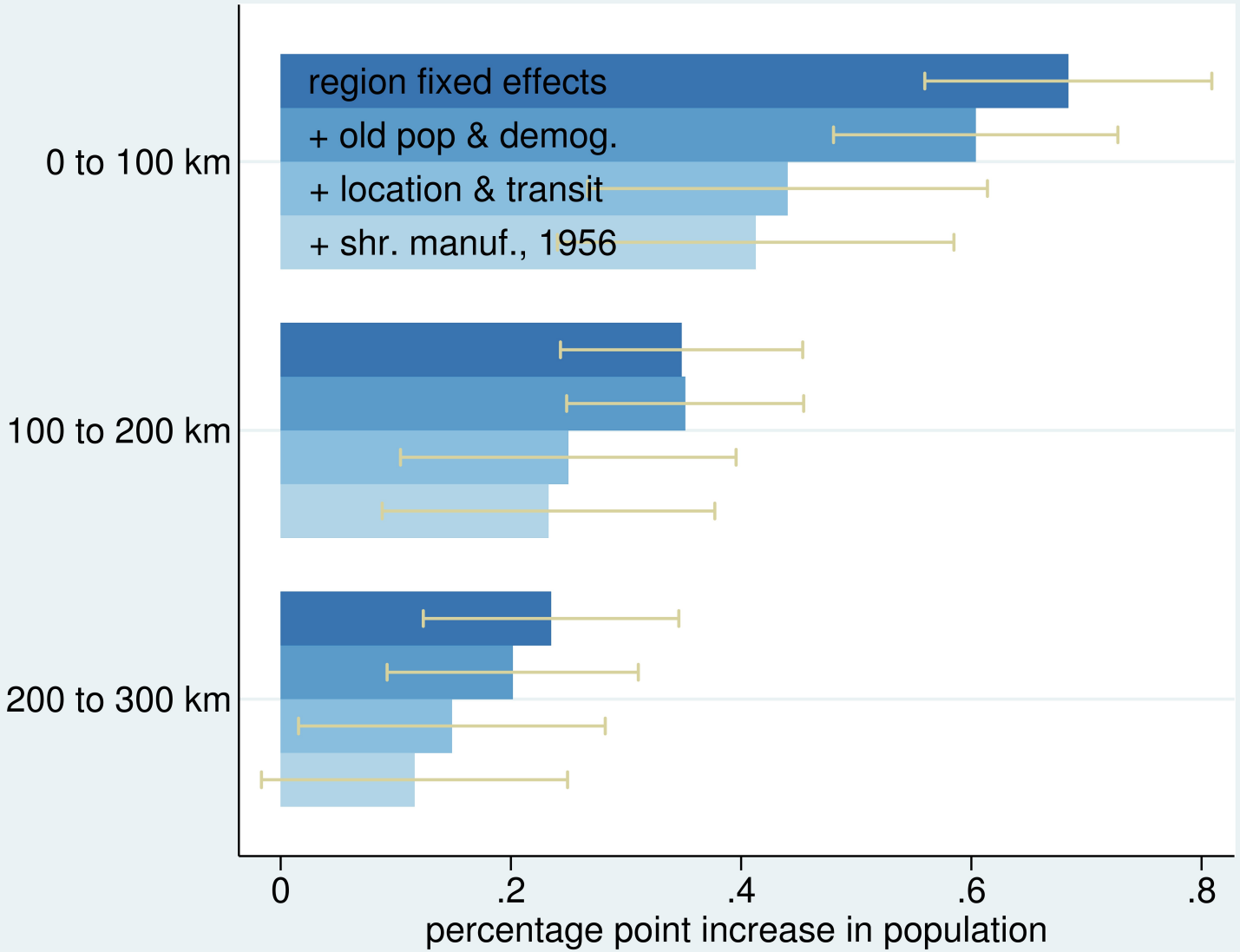
Bars with Error Bars, Building



Bars with Error Bars, Building



Bars with Error Bars, Building



2. Secondary Data Components

- Trend lines
- Reference lines
- Annotation
- Scales
- Tick marks
- Legends

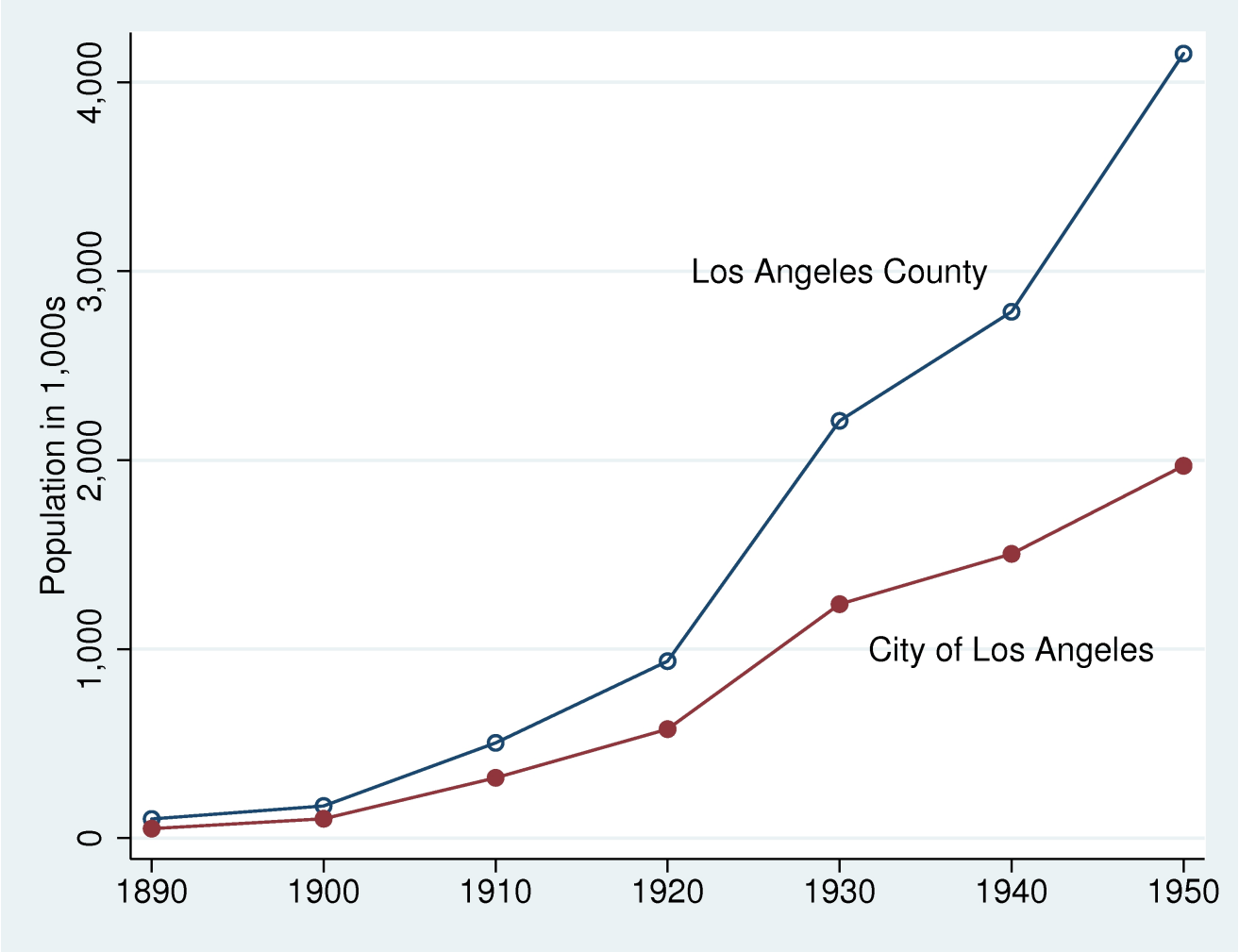
Trend Lines and Reference Lines

- In both cases, ask yourself why you want to include this line
- Why story is this line helping you tell?
- If the answer is nothing, omit it
- Include these types of lines if your goal is to highlight the relationship to trend or mean

Annotations

- Defined as on-graph writing
- Frequently key to successful charts

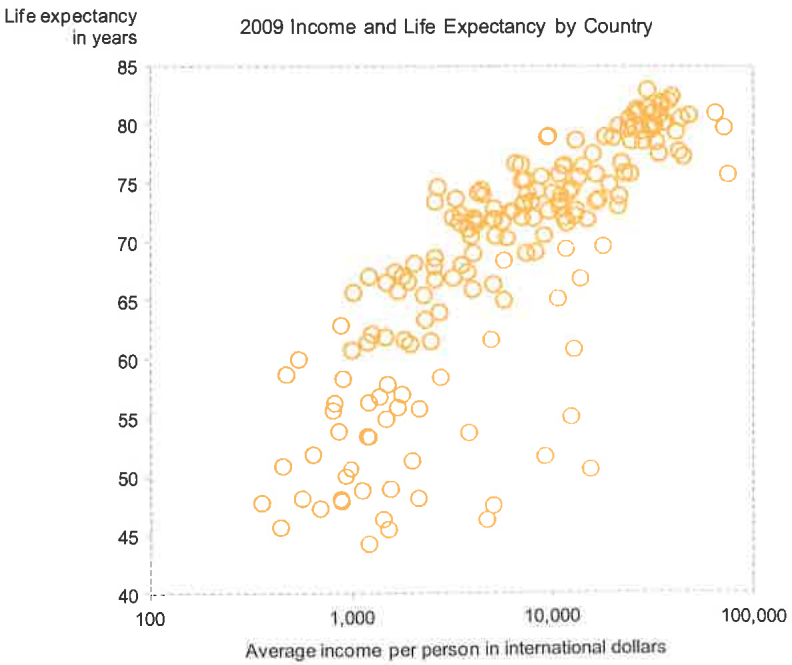
Annotations Simplify

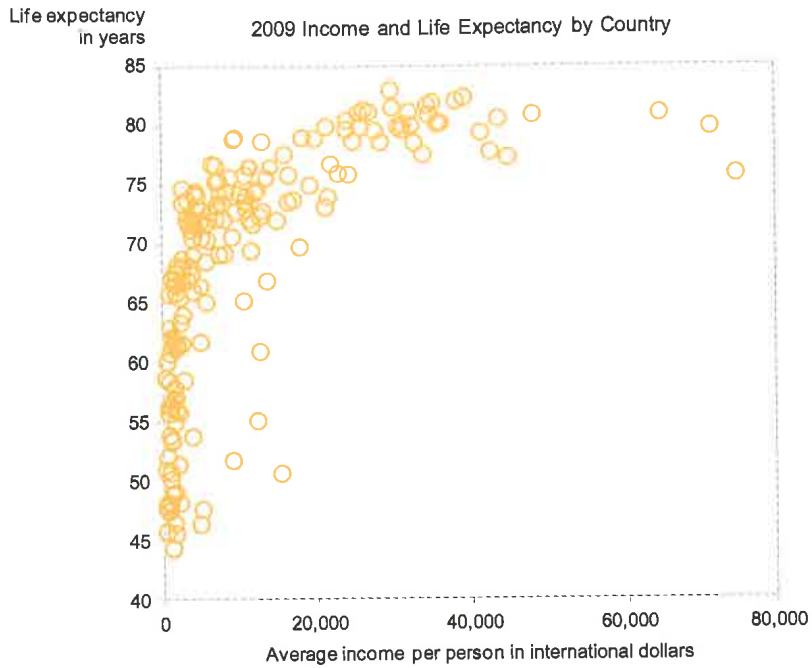


Scales

- You are usually choosing between levels and logs
- Use levels unless you have a good reason to use logs
- Good reasons to use logs are
  - The distribution is very skewed, and taking the log makes the visual easier
  - You care primarily about comparison of change
  - The difference in logs is change
- If using logs, label levels, not log values

Log vs Linear

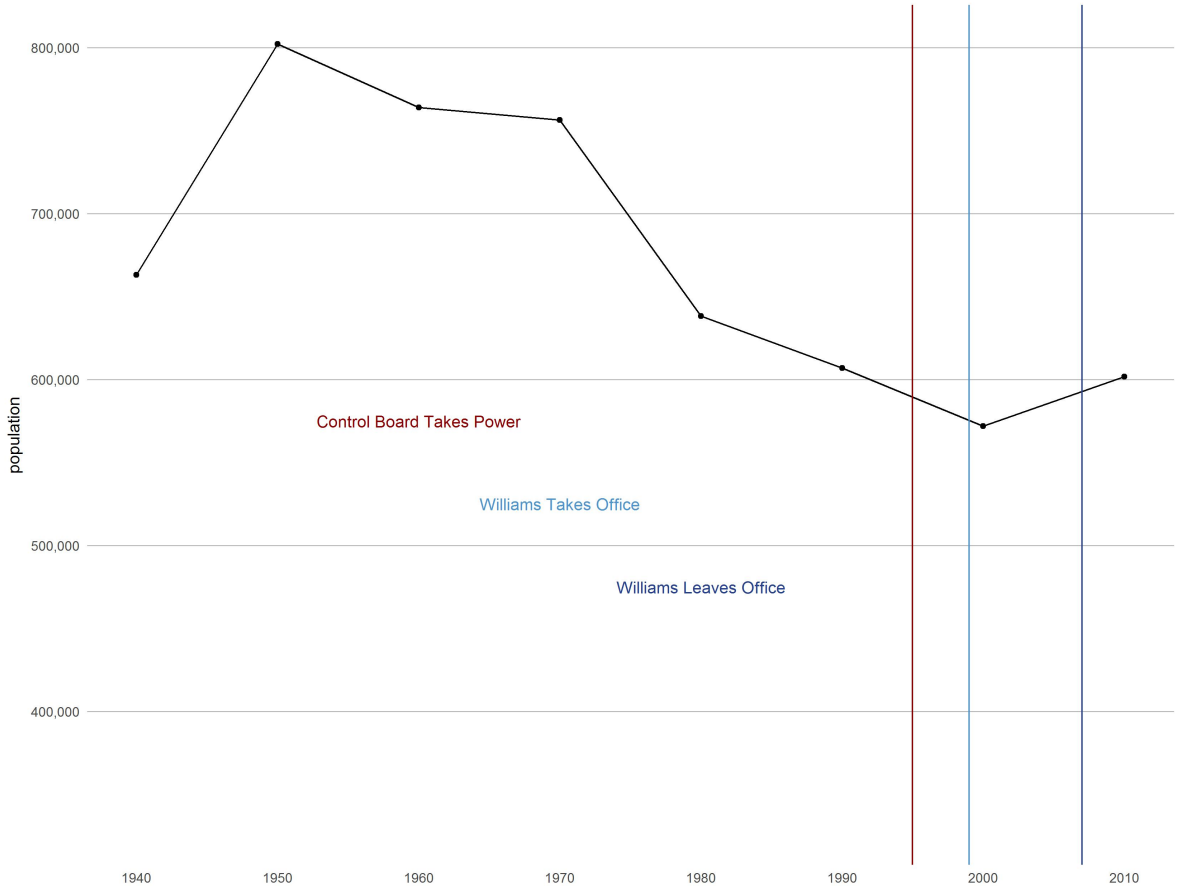




Tick Marks

- Get rid of them if you can
- Use to denote discrete numbers
- Not needed for categorical values
- Can remove if you have lines
- Lines serve as tick marks to compare values

From Today's Tutorial

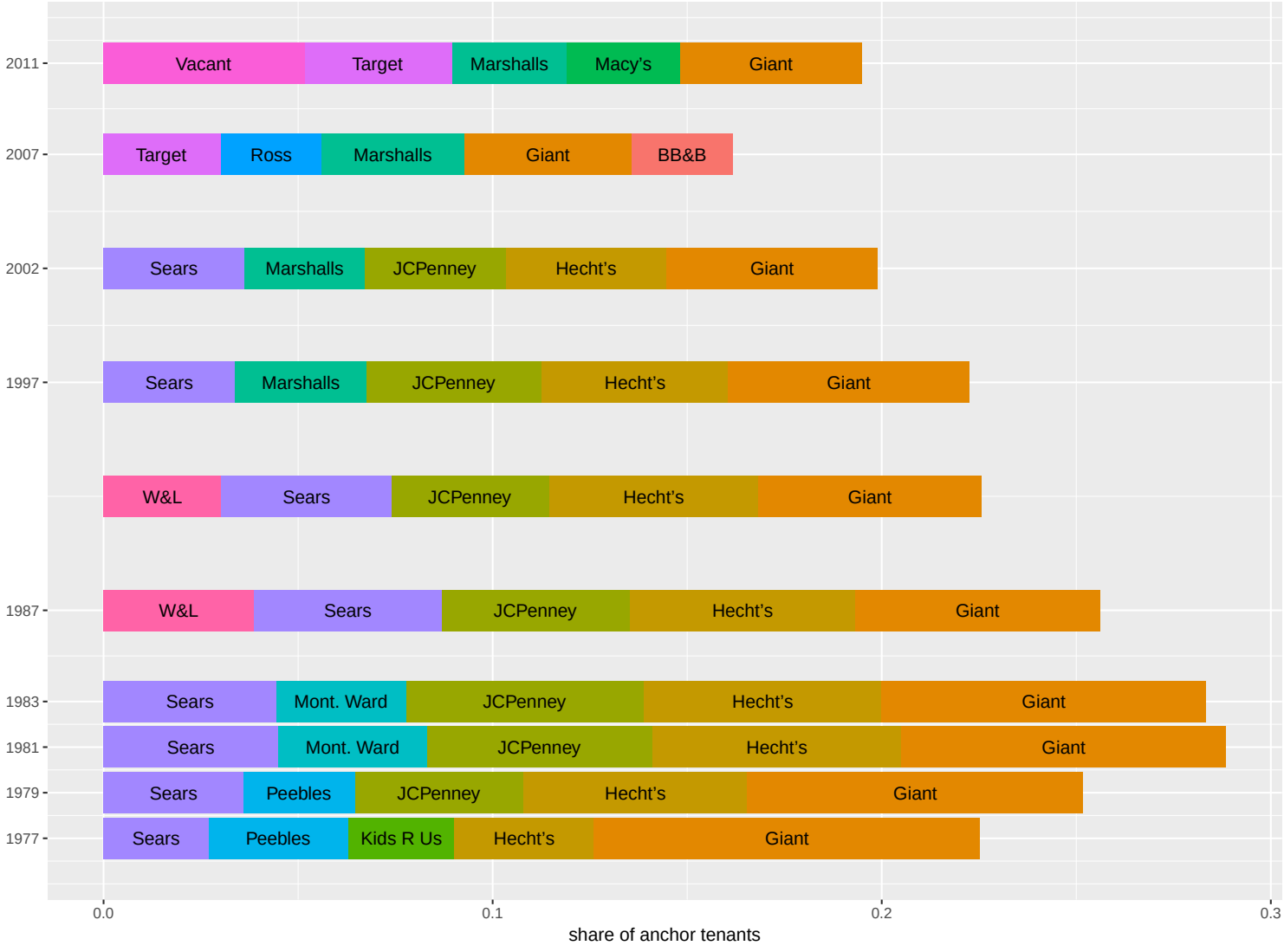


Legends

- Get rid of the box on the side if at all possible
- Don't use too many items
- Instead, fix the data or presentation

Anchor Stores Over Time

Top 5 Anchor Stores Per Year



Line Charts in R

Today's Goals

- De-bugging
- Data Manipulation
  - `reshape`
- Line graphs, via `ggplot`
  - `geom_line()`
- Annotating charts
  - `geom_v_line()`
  - `annotate()`

De-Bugging

- First question: Can R do what I'm asking it?
- Are my data appropriate to the command?
- Check the data
  - `names()`
  - `dim()`
  - `table(df$var.name)`
  - print a bit: `df[1:10,]`

R Matrix

```
mymat = matrix(1:12,4,3)
mymat
```

```
##      [,1] [,2] [,3]
## [1,]    1    5    9
## [2,]    2    6   10
## [3,]    3    7   11
## [4,]    4    8   12
```

- Rows are listed first
- Columns are listed second

Showing a little of the data: Rows 1 and 2

```
mymat = matrix(1:12,4,3)
mymat
```

```
##      [,1] [,2] [,3]
## [1,]    1    5    9
## [2,]    2    6   10
## [3,]    3    7   11
## [4,]    4    8   12
```

```
mymat[1:2,]
```

```
##      [,1] [,2] [,3]
## [1,]    1    5    9
## [2,]    2    6   10
```

Showing a little of the data: Column 3

```
mymat = matrix(1:12,4,3)
mymat
```

```
##      [,1] [,2] [,3]
## [1,]    1    5    9
## [2,]    2    6   10
## [3,]    3    7   11
## [4,]    4    8   12
```

```
mymat[,3]
```

```
## [1]  9 10 11 12
```

Data Manipulation

```
new.df <- reshape(data,
varying = NULL,
timevar = ``newname'',
idvar = ``org.id'',
direction = ``long'',
sep = ``[sep object]'')
```

- You have "wide" data and need "long" data or vice versa
- Plots frequently require "long" data
- A very standard thing to do in all kinds of languages that manipulate data
- Many alternatives in R

I relied on this page.

Wide data

```
wide <- data.frame(state = c("6","36","48"),
                   female_pop = c("10","12","14"),
                   male_pop = c("11","13","12"))
wide
```

```
##   state female_pop male_pop
## 1     6         10       11
## 2    36         12       13
## 3    48         14       12
```

Same data, long format

```
long <- data.frame(state = c("6","36","48","6","36","48"),
                   pop = c("10","12","14","11","13","12"),
                   sex = c("female","female","female",
                           "male","male","male"))
long
```

```
##   state pop    sex
## 1     6  10 female
## 2    36  12 female
## 3    48  14 female
## 4     6  11   male
## 5    36  13   male
## 6    48  12   male
```

Example in terms of `reshape()` for wide → long

- `varying` is female_pop, male_pop
- `timevar` is sex
- `idvar` is state
- `direction` is long
- `sep`: first rename variables

Rename the population variable to be something that ends in consecutive numbers so R can understand: `pop1, pop2`

And on to ggplot, geom_line()

Not entirely new to us. You can modify

- line width
- line color
- have multiple lines
- color under the lines

Annotating Charts

To just touch the tip of the iceberg

- `geom_vline(xintercept = x.value, color = ‘‘royalblue4’’)`
- `annotate(‘‘text’’, x=x.value, y=y.value, label=‘‘what it should say’’, color = ‘‘red4’’)`

I think I should have used

- `geom_segment(aes(x = x1, y = y1, xend = x2, yend = y2, color = ‘‘color’’)`
- and put the text above or below the line

Try Today's Tutorial

- Pay attention to the output of each bit
- Ask questions if the command doesn't make sense
- Go forth!

Next Lecture

- Turn in PS 5
- Read Few Chapter 13
- R Graphics Cookbook, Chapter 5: Scatter Plots
- Next policy brief deadline: April 2 for draft