Lec 09 - Visualization with ggplot2

Statistical Programming

Sem 1, 2020

Dr. Colin Rundel

Why do we visualize?

Asncombe's Quartet

```r
datasets::anscombe %>% as_tibble()
```

```
## # A tibble: 11 x 8
##       x1    x2    x3    x4    y1    y2    y3    y4
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1    10    10    10     8  8.04  9.14  7.46  6.58
##  2     8     8     8     8  6.95  8.14  6.77  5.76
##  3    13    13    13     8  7.58  8.74 12.7   7.71
##  4     9     9     9     8  8.81  8.77  7.11  8.84
##  5    11    11    11     8  8.33  9.26  7.81  8.47
##  6    14    14    14     8  9.96  8.1   8.84  7.04
##  7     6     6     6     8  7.24  6.13  6.08  5.25
##  8     4     4     4    19  4.26  3.1   5.39 12.5 
##  9    12    12    12     8 10.8   9.13  8.15  5.56
## 10     7     7     7     8  4.82  7.26  6.42  7.91
## 11     5     5     5     8  5.68  4.74  5.73  6.89
```

Tidy anscombe

```
(tidy_anscombe = datasets::anscombe %>%
  pivot_longer(everything(), names_sep = 1, names_to = c("var", "group")) %>%
  pivot_wider(id_cols = group, names_from = var,
              values_from = value, values_fn = list(value = list)) %>%
  unnest(cols = c(x,y)))
```

```
## # A tibble: 44 x 3
##    group     x     y
##    <chr> <dbl> <dbl>
##  1 1        10  8.04
##  2 1         8  6.95
##  3 1        13  7.58
##  4 1         9  8.81
##  5 1        11  8.33
##  6 1        14  9.96
##  7 1         6  7.24
##  8 1         4  4.26
##  9 1        12 10.8
## 10 1         7  4.82
## # … with 34 more rows
```

```
tidy_anscombe %>%
  group_by(group) %>%
  summarize(mean_x = mean(x), mean_y = mean(y), sd_x = sd(x), sd_y = sd(y), cor = cor(x,y))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 4 x 6
```

```r
ggplot(tidy_anscombe, aes(x = x, y = y, color = as.factor(group))) +
  geom_point(size=2) +
  facet_wrap(vars(group)) +
  geom_smooth(method="lm", se=FALSE, fullrange=TRUE) +
  guides(color=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

DatasauRus

```
datasauRus::datasaurus_dozen
```

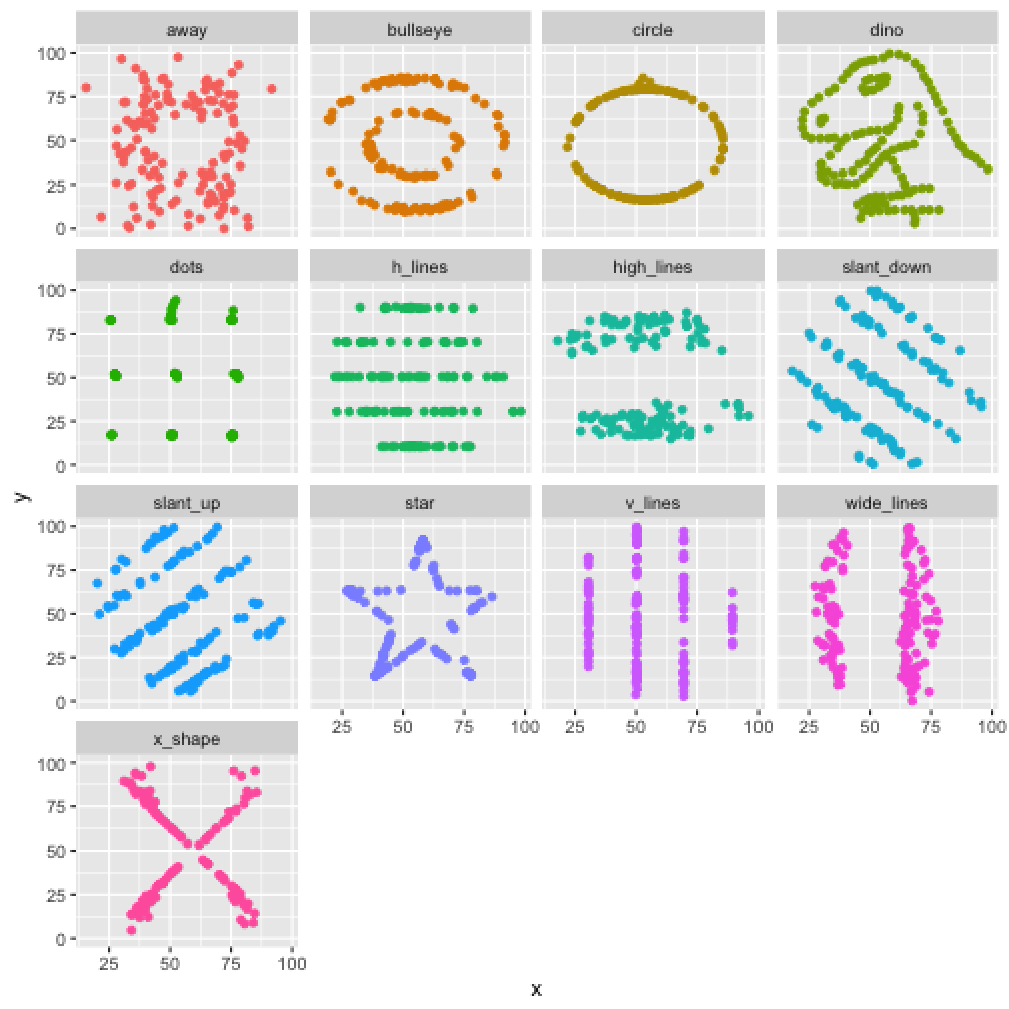
```
## # A tibble: 1,846 x 3
##    dataset     x     y
##    <chr>   <dbl> <dbl>
##  1 dino     55.4  97.2
##  2 dino     51.5  96.0
##  3 dino     46.2  94.5
##  4 dino     42.8  91.4
##  5 dino     40.8  88.3
##  6 dino     38.7  84.9
##  7 dino     35.6  79.9
##  8 dino     33.1  77.6
##  9 dino     29.0  74.5
## 10 dino     26.2  71.4
## # … with 1,836 more rows
```

```
datasauRus::datasaurus_dozen %>%
  group_by(dataset) %>%
  summarize(mean_x = mean(x), mean_y = mean(y),
            sd_x = sd(x), sd_y = sd(y),
            cor = cor(x,y))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

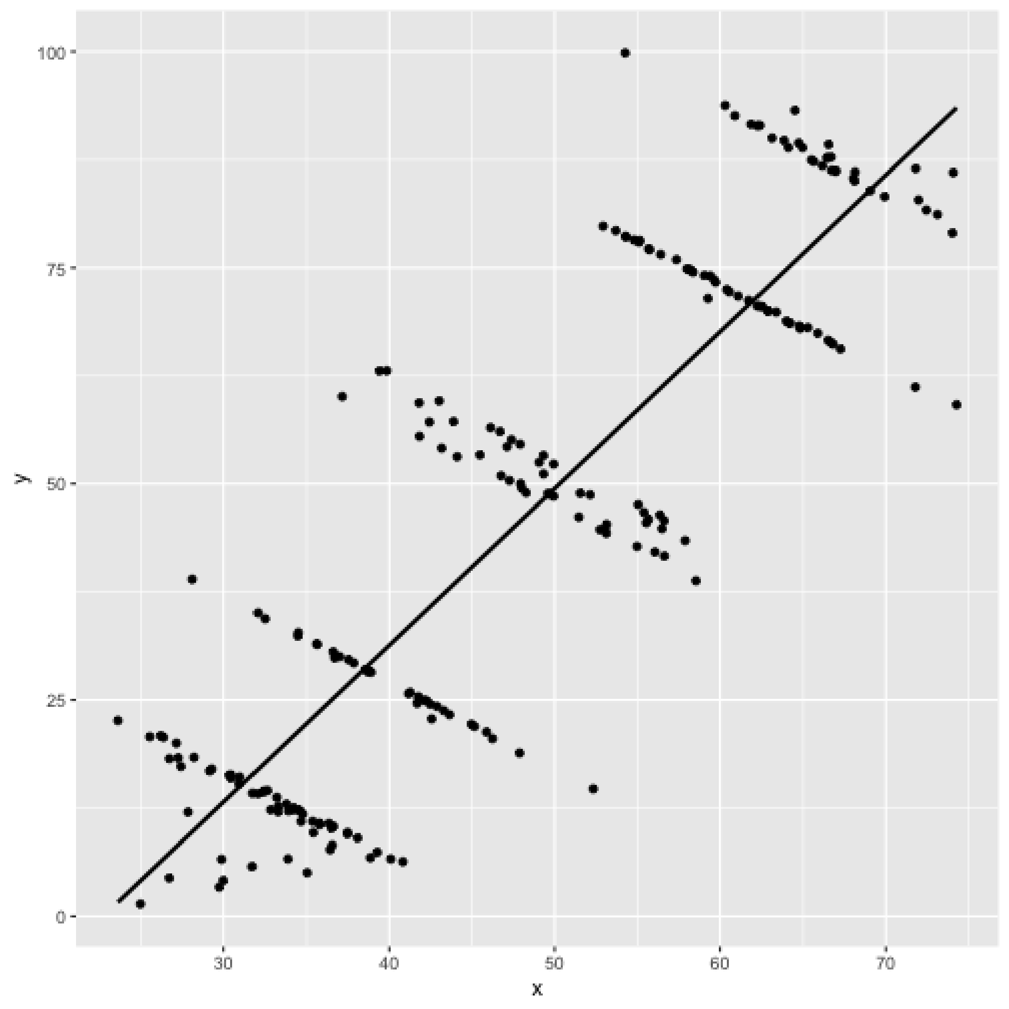
```
## # A tibble: 13 x 6
##    dataset    mean_x mean_y   sd_x   sd_y      cor
##    <chr>       <dbl>  <dbl> <dbl> <dbl>    <dbl>
```

```r
ggplot(datasauRus::datasaurus_dozen, aes(x = x, y = y, color = dataset)) +
  geom_point() +
  facet_wrap(vars(dataset)) +
  guides(color=FALSE)
```

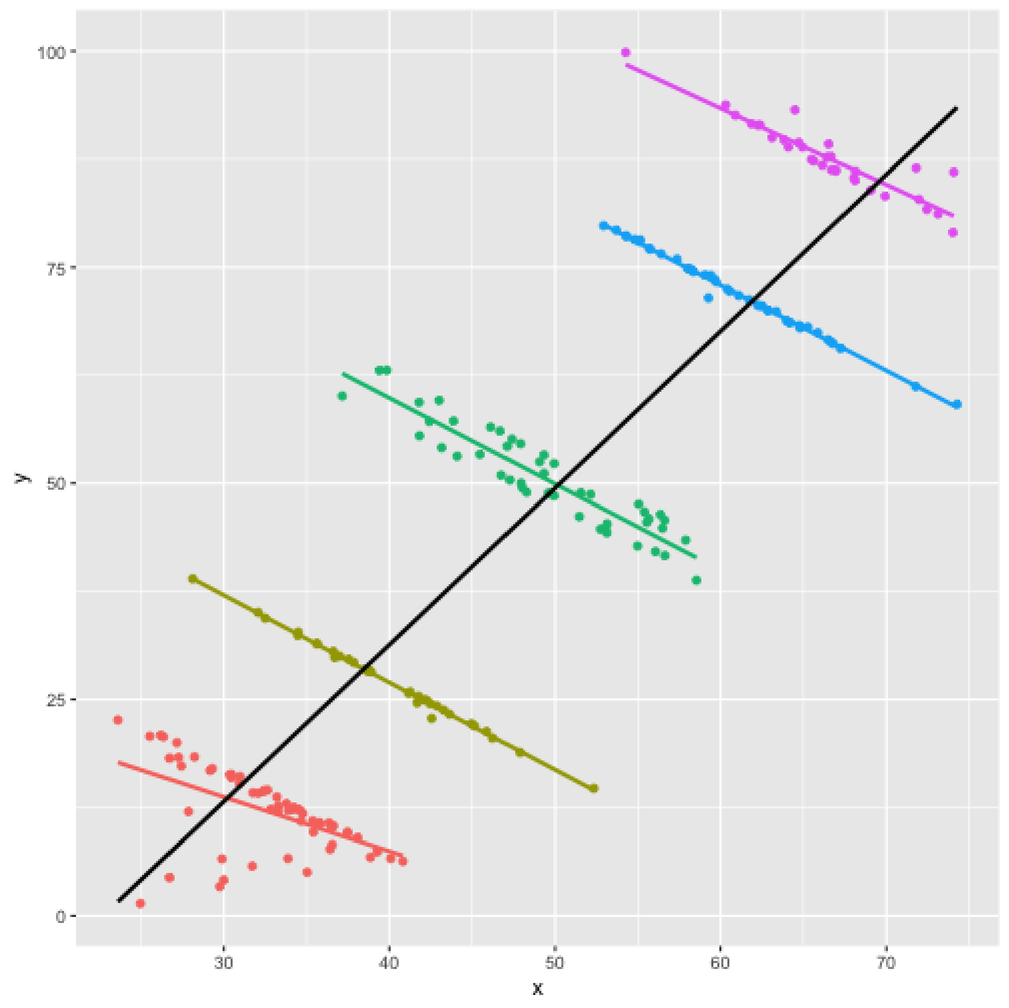


Simpson's Paradox

```
## `geom_smooth()` using formula 'y ~ x'
```

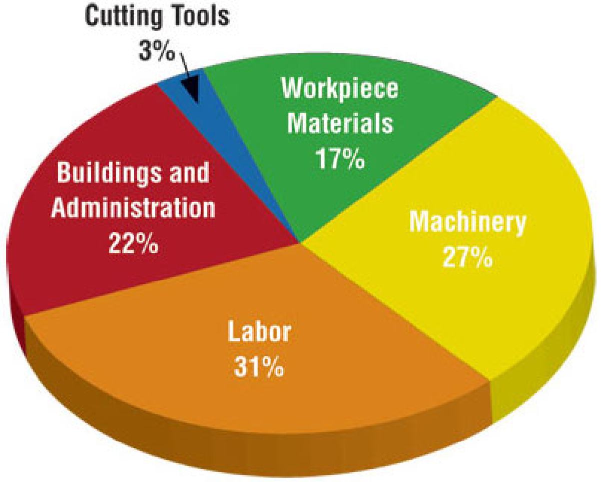


```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



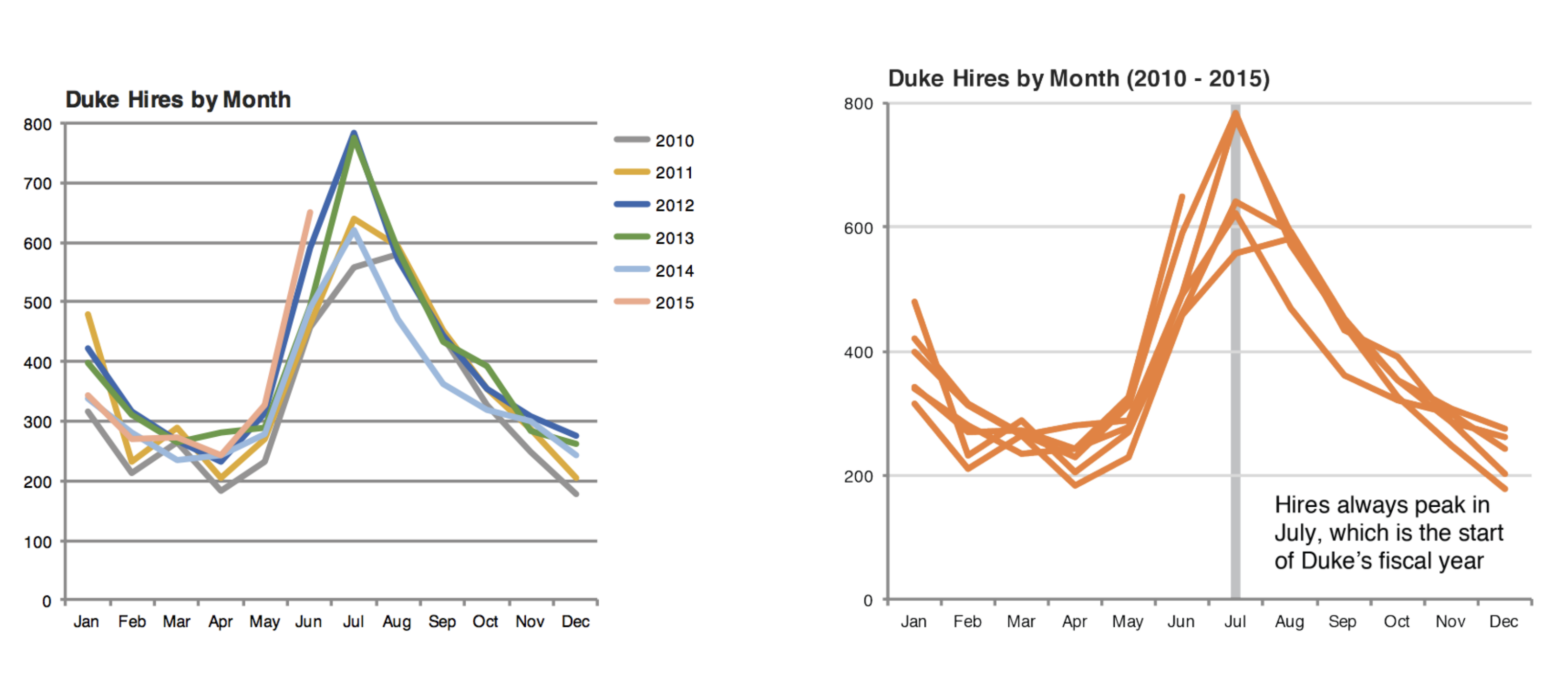
Designing effective visualizations

Keep it simple



Use color to draw attention

Tell a story



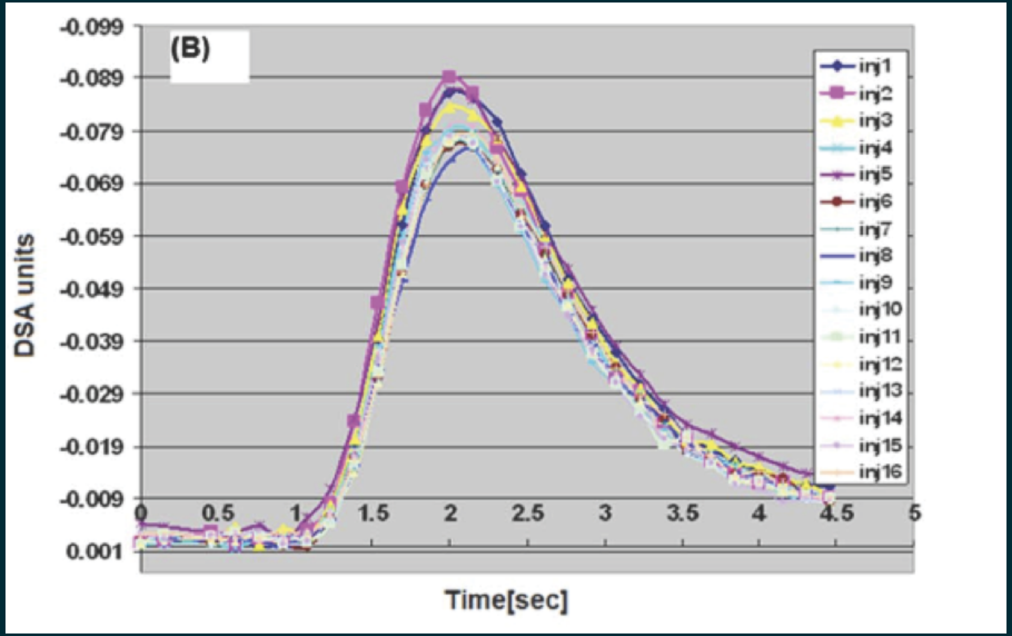
Duke Hires by Month

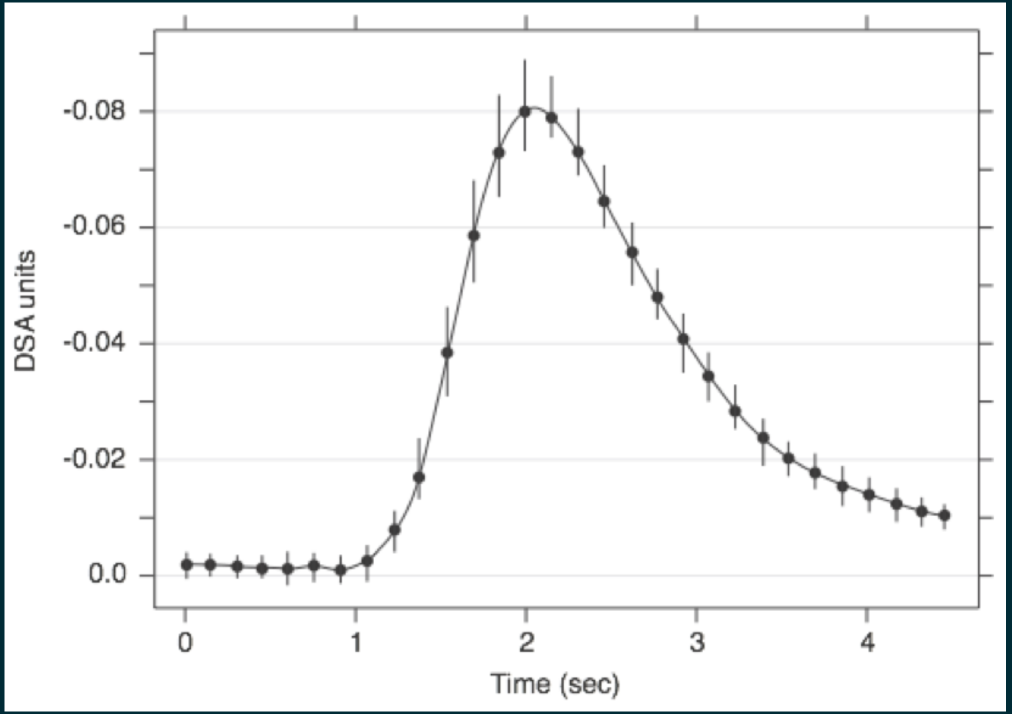
Duke Hires by Month (2010 - 2015)

Hires always peak in July, which is the start of Duke's fiscal year

Credit: Angela Zoss and Eric Monson, Duke DVS

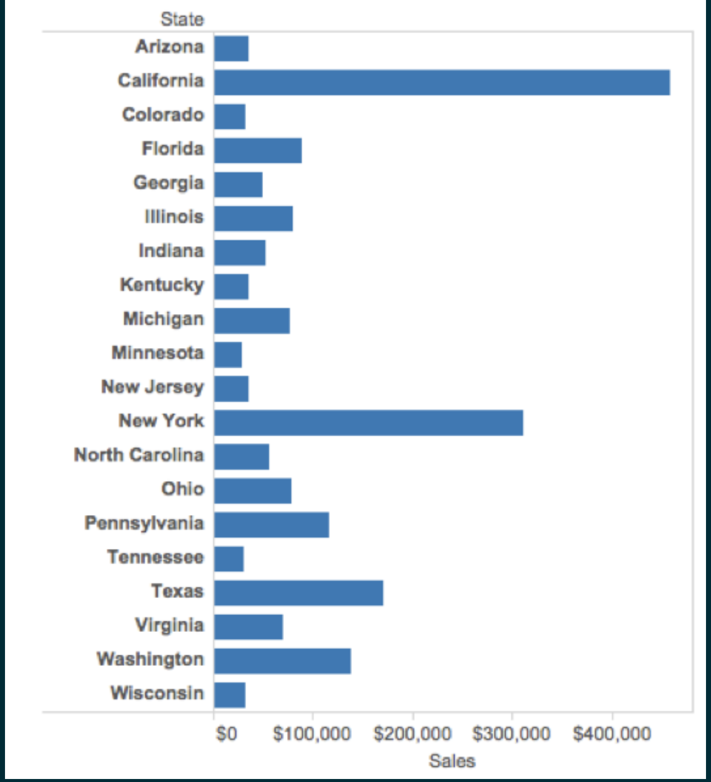
Leave out non-story details

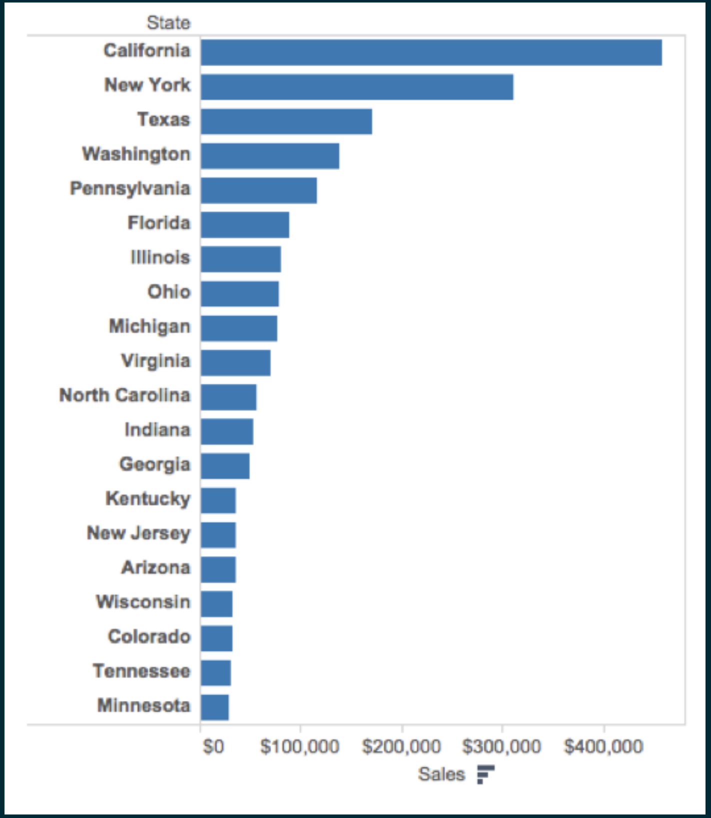




Credit: Angela Zoss and Eric Monson, Duke DVS

Order / usage matters





Credit: Angela Zoss and Eric Monson, Duke DVS

Be clear about missing data

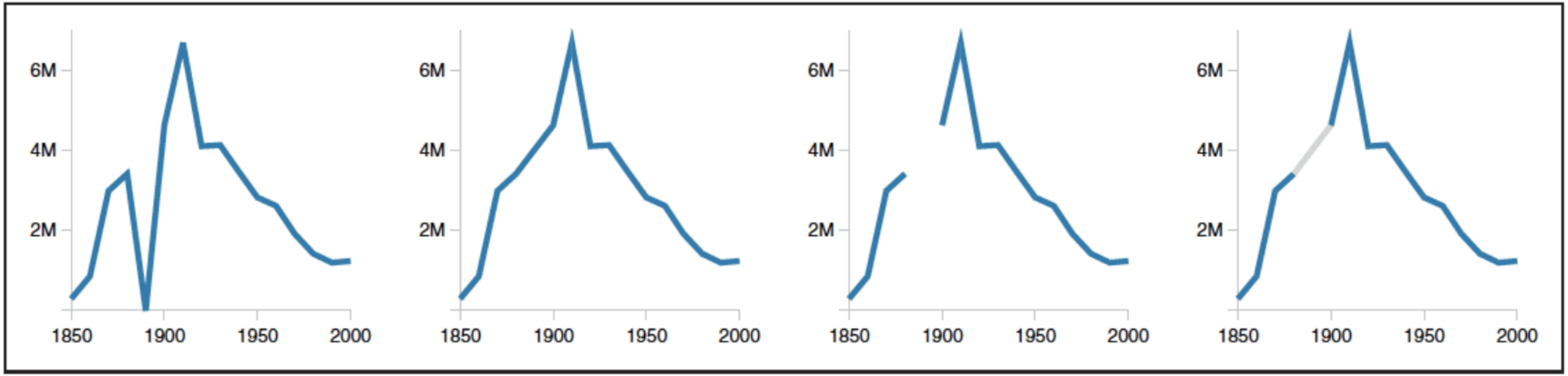
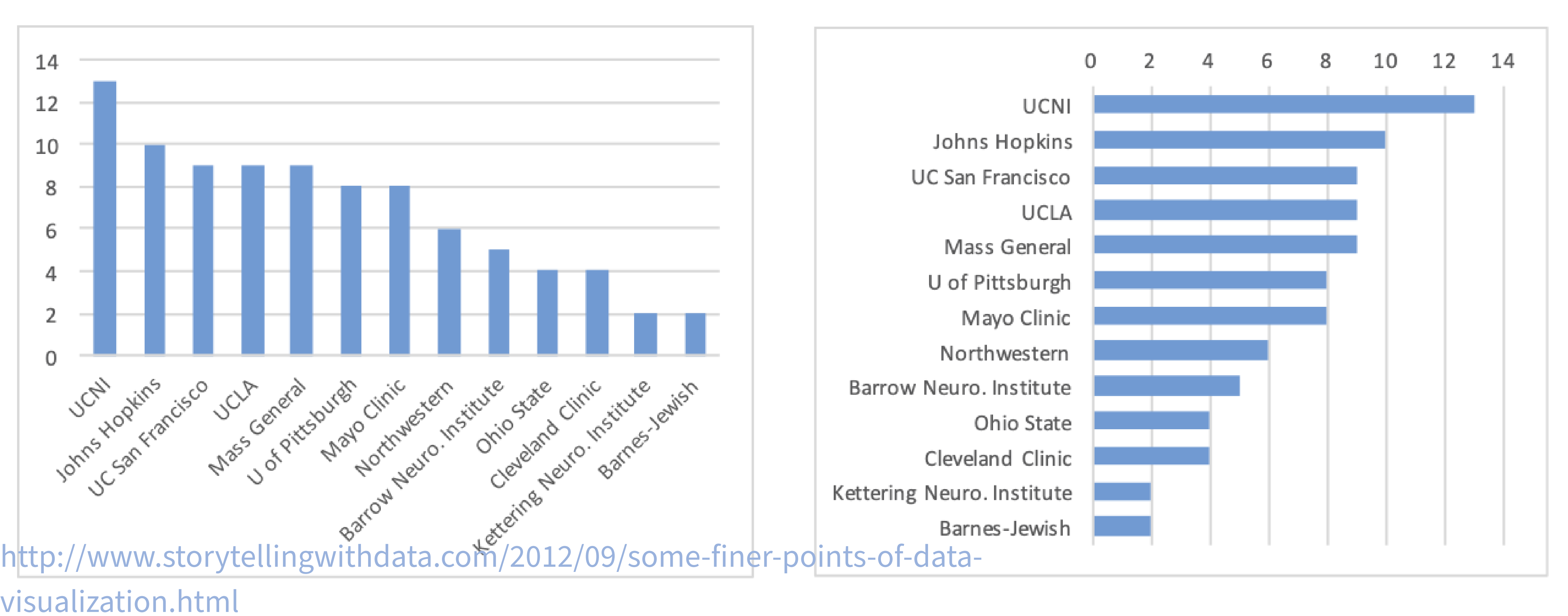


Figure 4. Alternative representations of missing data in a line chart. The data are U.S. census counts of people working as ‘Farm Laborers’; values from 1890 are missing due to records being burned in a fire. (a) Missing data is treated as a zero value. (b) Missing data is ignored, resulting in a line segment that interpolates the missing value. (c) Missing data is omitted from the chart. (d) Missing data is explicitly interpolated and rendered in gray.

http://ivi.sagepub.com/content/10/4/271

Angela Zoss and Eric Monson, Duke DVS

Reduce cognitive burden



http://www.storytellingwithdata.com/2012/09/some-finer-points-of-data-visualization.html

Angela Zoss and Eric Monson, Duke DVS

Use descriptive titles

Credit: Angela Zoss and Eric Monson, Duke DVS

Annotate figures directly

https://bl.ocks.org/susielu/23dc3082669ee026c552b85081d90976

All of the data doesn't tell a story

http://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html

All of the data doesn't tell a story

http://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html

All of the data doesn't tell a story

http://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html

Chart Remakes / Makeovers

The Why Axis - BLS

http://thewhyaxis.info/defaults/

The Why Axis - Gender Gap

http://thewhyaxis.info/gap-remake/

Acknowledgments

Acknowledgments

Above materials are derived in part from the following sources:

- Hadley Wickham - R for Data Science & Elegant Graphics for Data Analysis
- ggplot2 website
- Visualization training materials developed by Angela Zoss and Eric Monson, Duke DVS