# Data Visualization
## Data-based Storytelling

Daniel Winkler    Stephan Fally

Institute for Retailing & Data Science

Department Marketing

Always Visualize!

# Processing numbers is hard

| Dino: x | Dino: y | Star: x | Star: y |
| --- | --- | --- | --- |
| 55.38 | 97.18 | 58.21 | 91.88 |
| 51.54 | 96.03 | 58.20 | 92.21 |
| 46.15 | 94.49 | 58.72 | 90.31 |
| 42.82 | 91.41 | 57.28 | 89.91 |
| 40.77 | 88.33 | 58.08 | 92.01 |
| 38.72 | 84.87 | 57.49 | 88.09 |
| 35.64 | 79.87 | 28.09 | 63.51 |
| 33.08 | 77.56 | 28.09 | 63.59 |
| 28.97 | 74.49 | 28.09 | 63.12 |
| 26.15 | 71.41 | 27.58 | 62.82 |

## Summary statistics are limiting

| dataset | Mean of x | Std. Dev. of x | Mean of y | Std. Dev. of y | Correlation |
|---------|-----------|----------------|-----------|----------------|-------------|
| Dino    | 54.26     | 16.77          | 47.83     | 26.94          | -0.06       |
| Star    | 54.27     | 16.77          | 47.84     | 26.93          | -0.06       |



Dino

Star

y
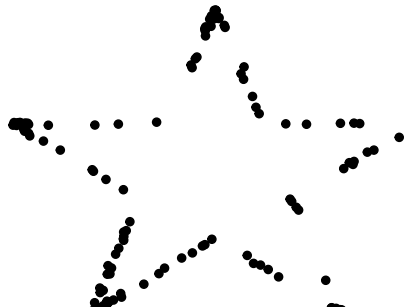
## Exercise

> **💡 See beyond summary statistics**
>
> ▶ Select the datasets x_shape & bullseye from the data.frame
>    datasaurus_dozen
> ▶ Create a table showing the following statistics for the two datasets:
>    ▶ mean of x and y,
>    ▶ standard deviation of x and y, and
>    ▶ covariance between x and y
> ▶ Create a plot showing the two datasets

```
tibble [284 x 3] (S3: tbl_df/tbl/data.frame)
 $ dataset: chr [1:284] "x_shape" "x_shape" "x_shape" "x_shape" ...
 $ x      : num [1:284] 38.3 35.8 32.8 33.7 37.2 ...
 $ y      : num [1:284] 92.5 94.1 88.5 88.6 83.7 ...
```
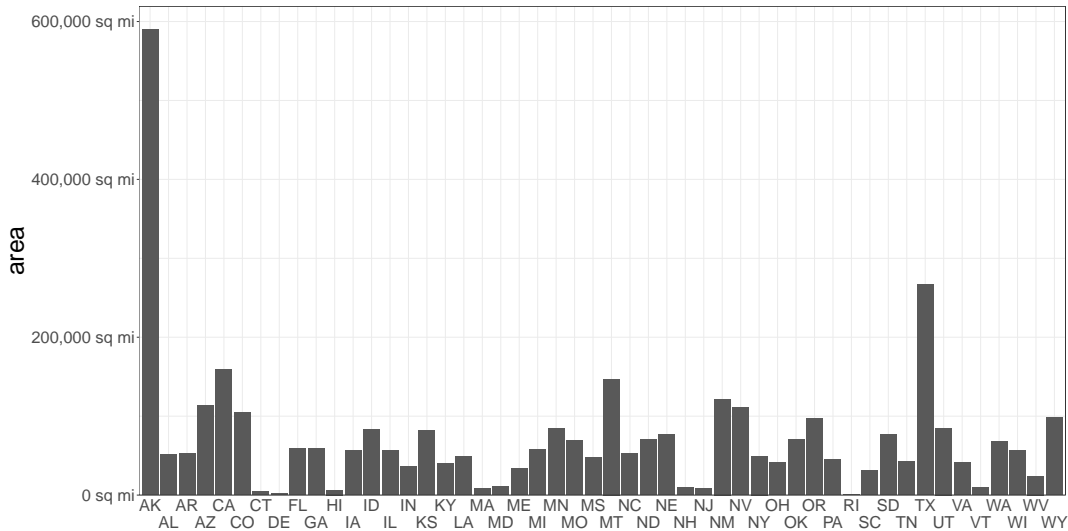
# Visual Channels

# Position is interpreted most accurately

- ▶ Common y-axis alignment allows for accurate interpretation
- ▶ Categories in the x-axis sorted by y-axis value
- ▶ Position should encode the answer to the questions the audience asks
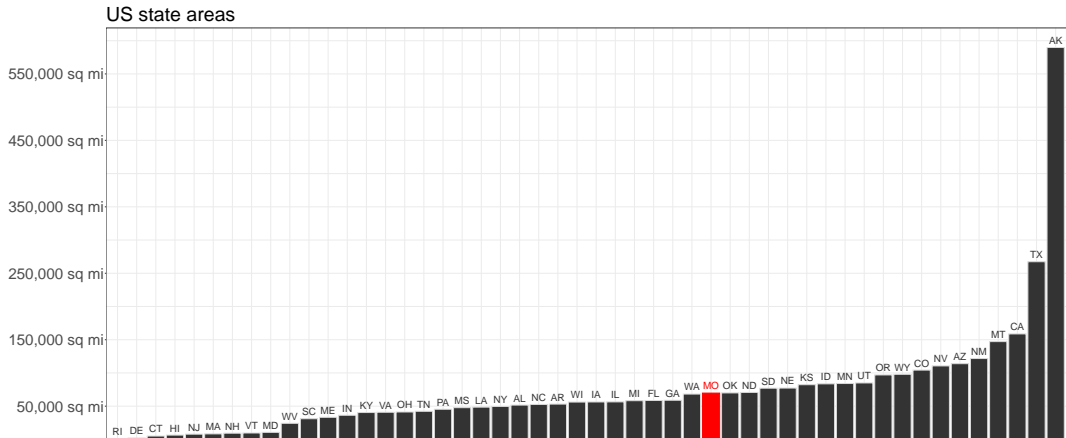
# Visualizing many categories (badly!)

- ▶ What is the 5th smallest state?
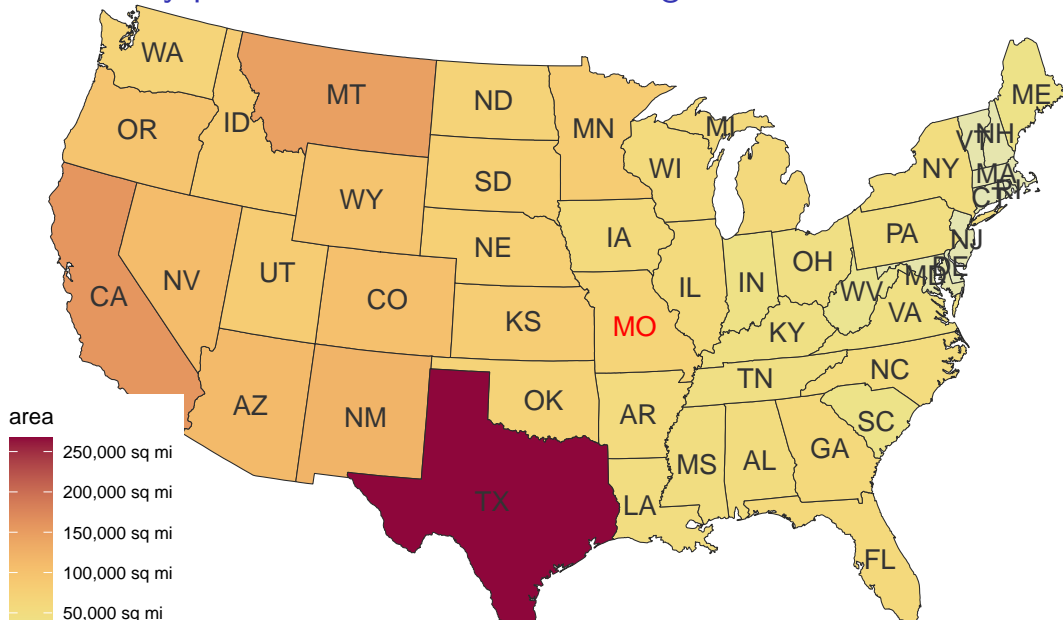- ▶ Where does MO rank in terms of size?

# Visualizing many categories

▶ What is the 5th smallest state?
▶ Where does MO rank in terms of size?



US state areas

# Color intensity provides less accurate encoding

# Colors are useful to mark targets for preattentive processing

▶ Large scale spatial data needs to be "projected" to 2D
▶ Most well known projection is 4326 measured in degrees latitude and longitude
▶ However, there are more accurate projections for specific regions e.g, 3035 for EU+candidates
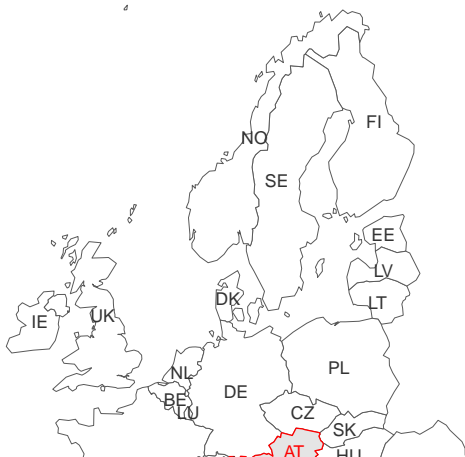
Mapping visual ratios

# Mapping visual ratios

If including $0$ in the y-axis would make the graph unreadable include additional annotation Or visualize the difference.