Concepts of Visual Inference

Heike Hofmann
Statistics at Iowa State University

Outline

- Barcharts and Pies
- Visual Inference
- Framework for Comparing Designs

Other sources of data and charts

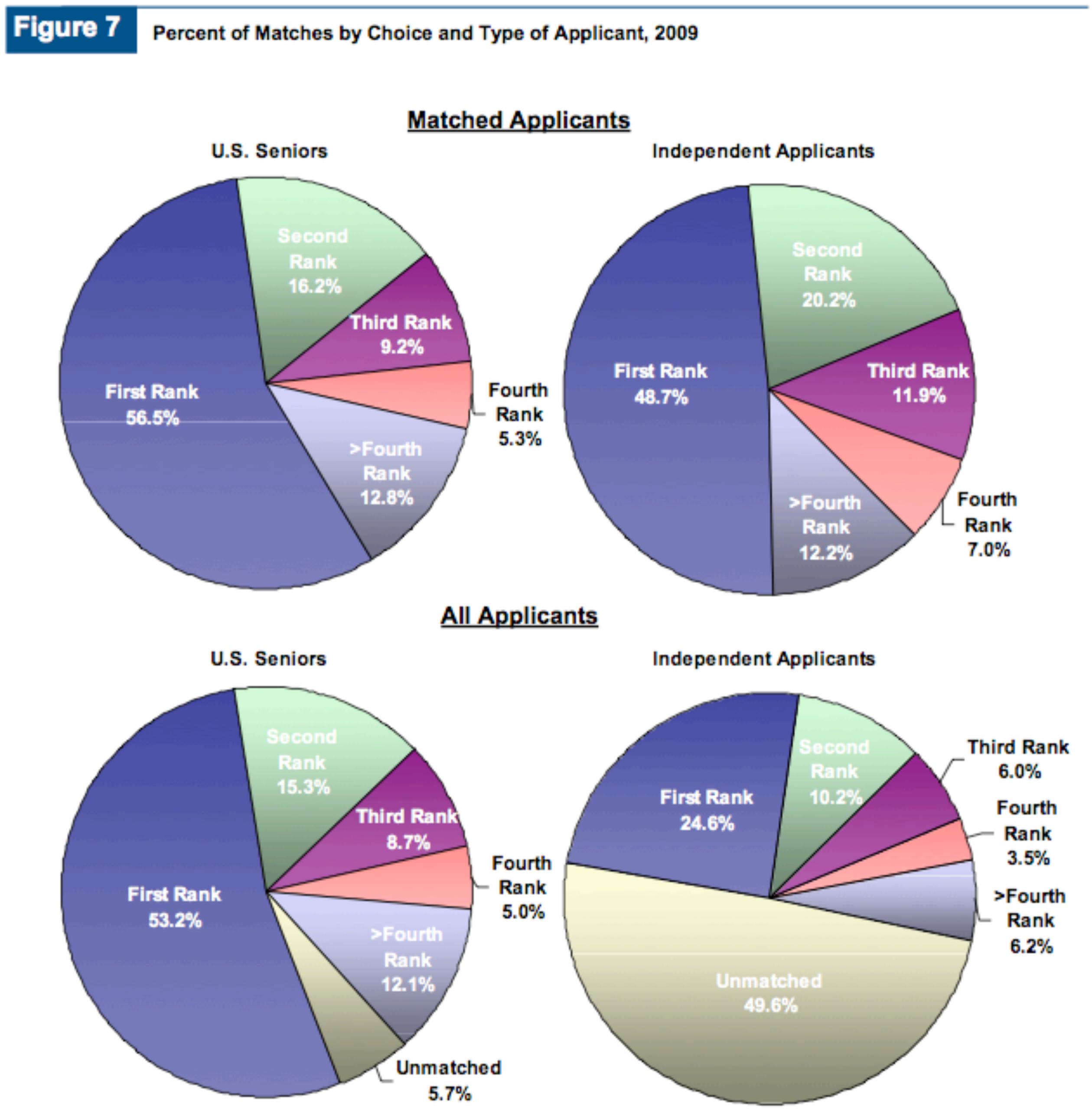
- Anesthesia Quality Institute: Anesthesia in the United States, 2009

  Excel graphics

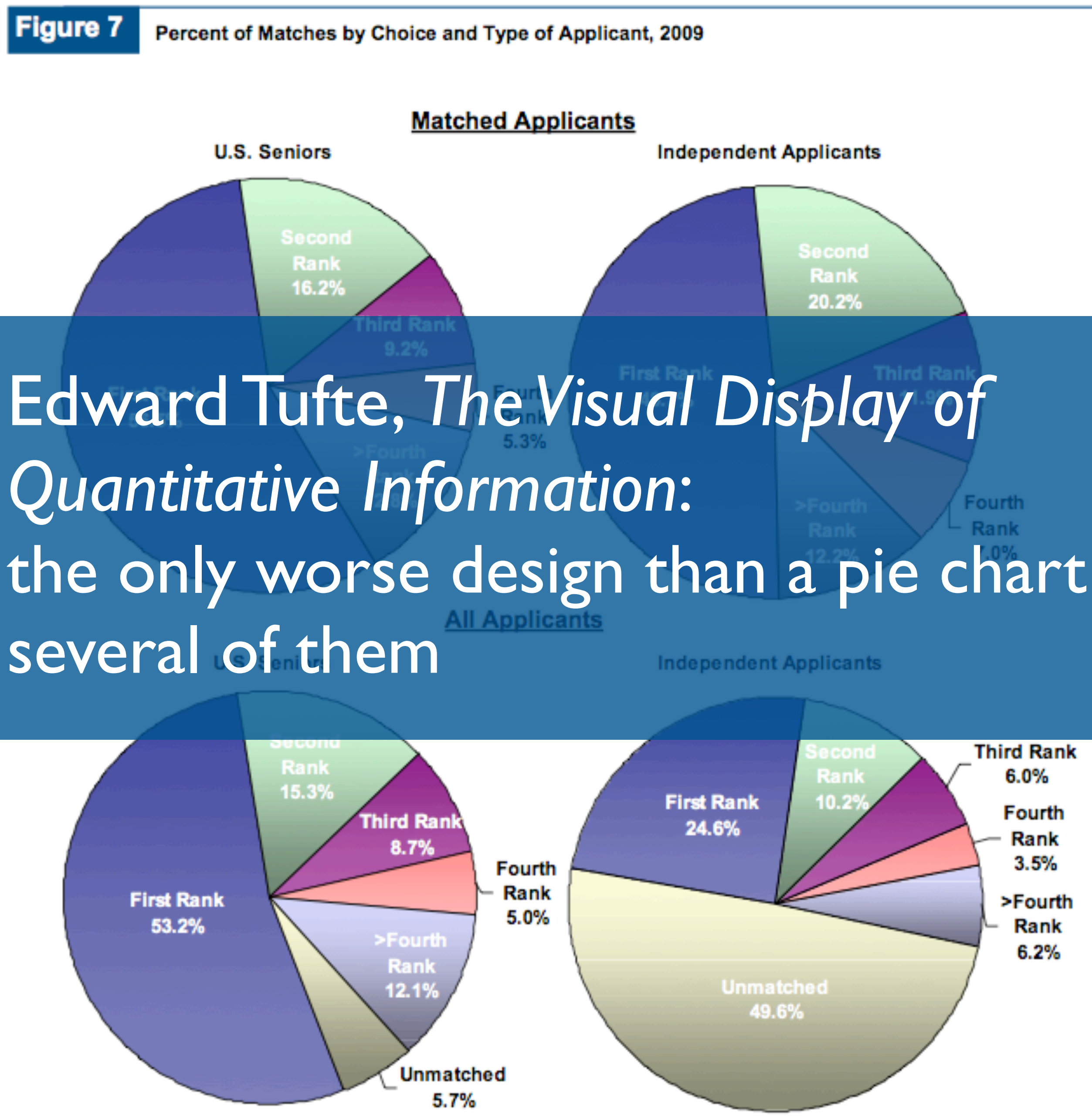
- National Resident Matching Program, Data and Report 2009

  Graphics are not in Excel

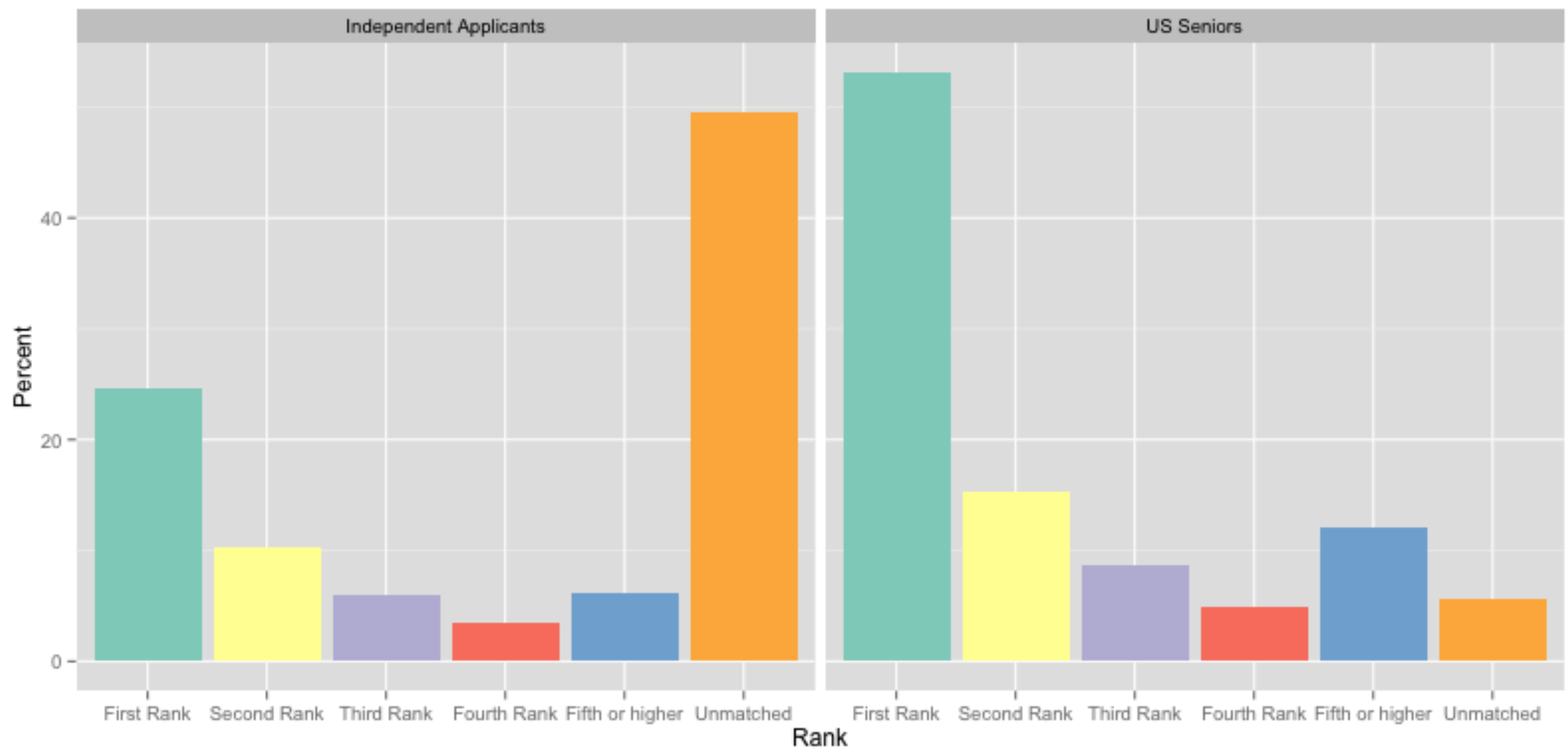
National Resident Matching Program



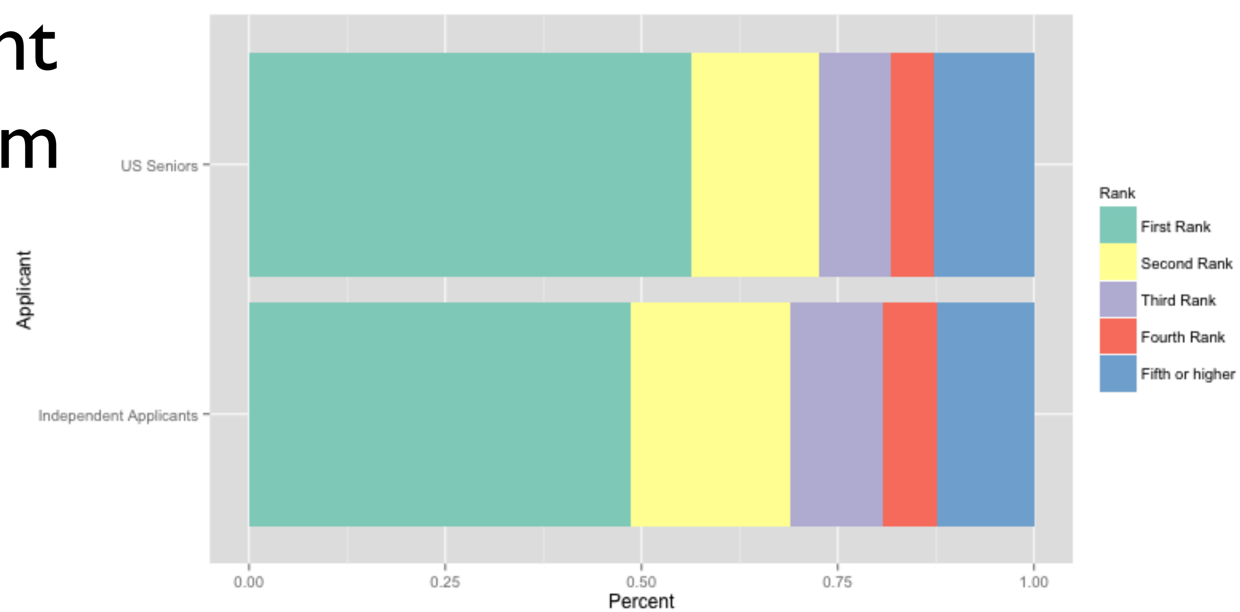
National Resident Matching Program



Edward Tufte, *The Visual Display of Quantitative Information*:
the only worse design than a pie chart is several of them



National Resident Matching Program -redone-



Evaluating Competing Designs

Evaluate perceptual strengths and weaknesses

- usually we are not interested in exact quantities
- ... But ... use accuracy as measure

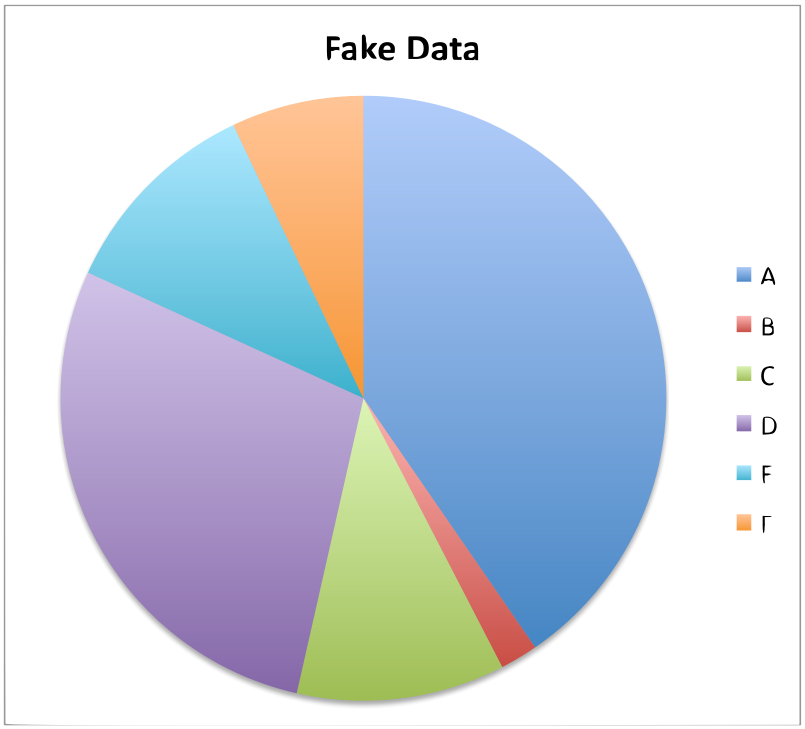
Cleveland & McGill (Science, 1985):

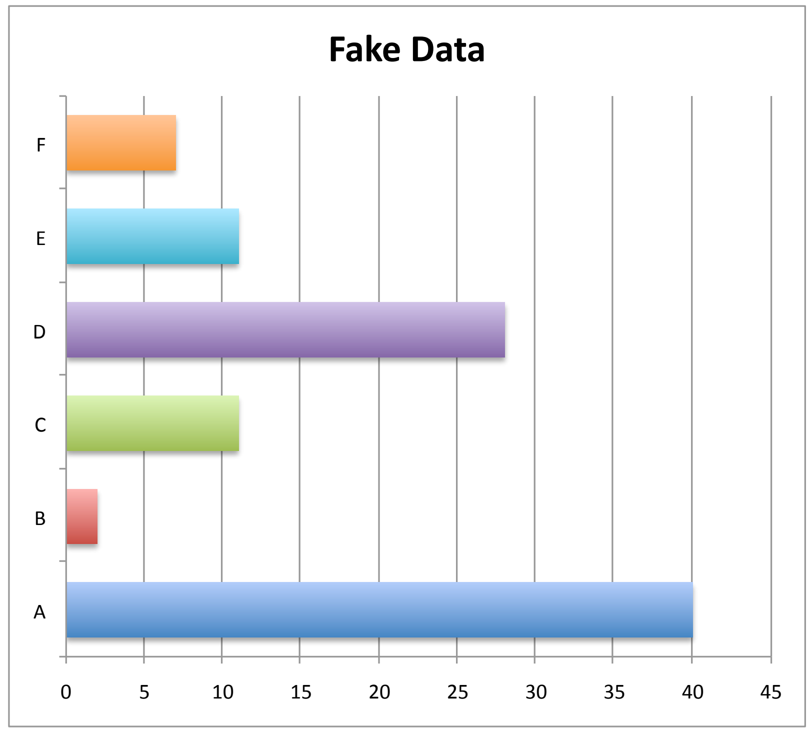
A graphical form that involves elementary perceptual tasks that lead to more accurate judgments than another graphical form (with the same quantitative information) will result in a better organization and increase the chances of correct perception of patterns and behavior.

Example: Bar vs Pie

What tasks are involved in comparisons?

Area is proportional to value





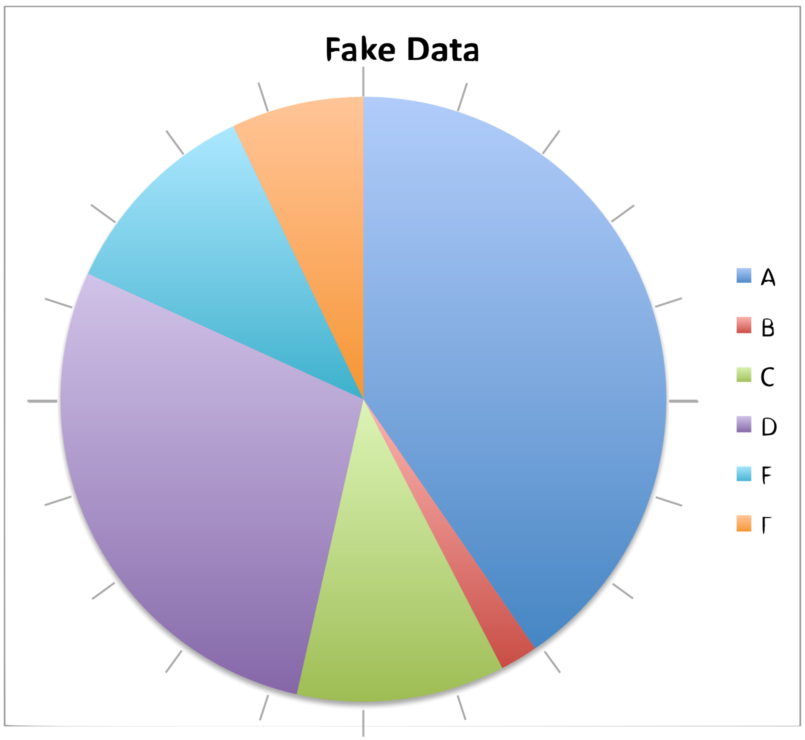
comparison of angles,
curve length

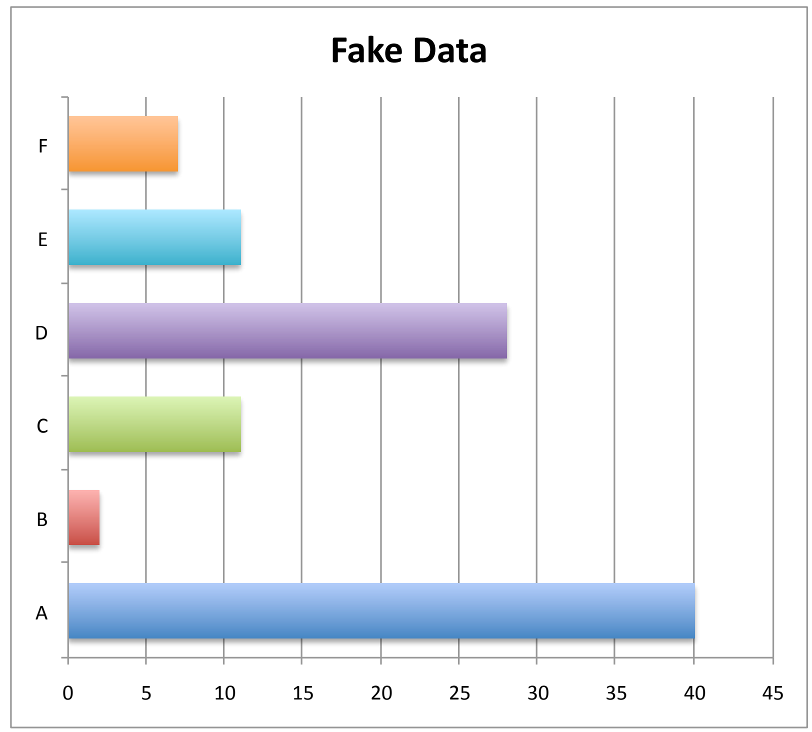
comparison of widths,
positions along a common scale

Example: Bar vs Pie

What tasks are involved in comparisons?

Area is proportional to value





comparison of angles, curve length

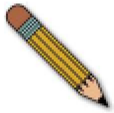
comparison of widths, positions along a common scale

Pies or Bars?

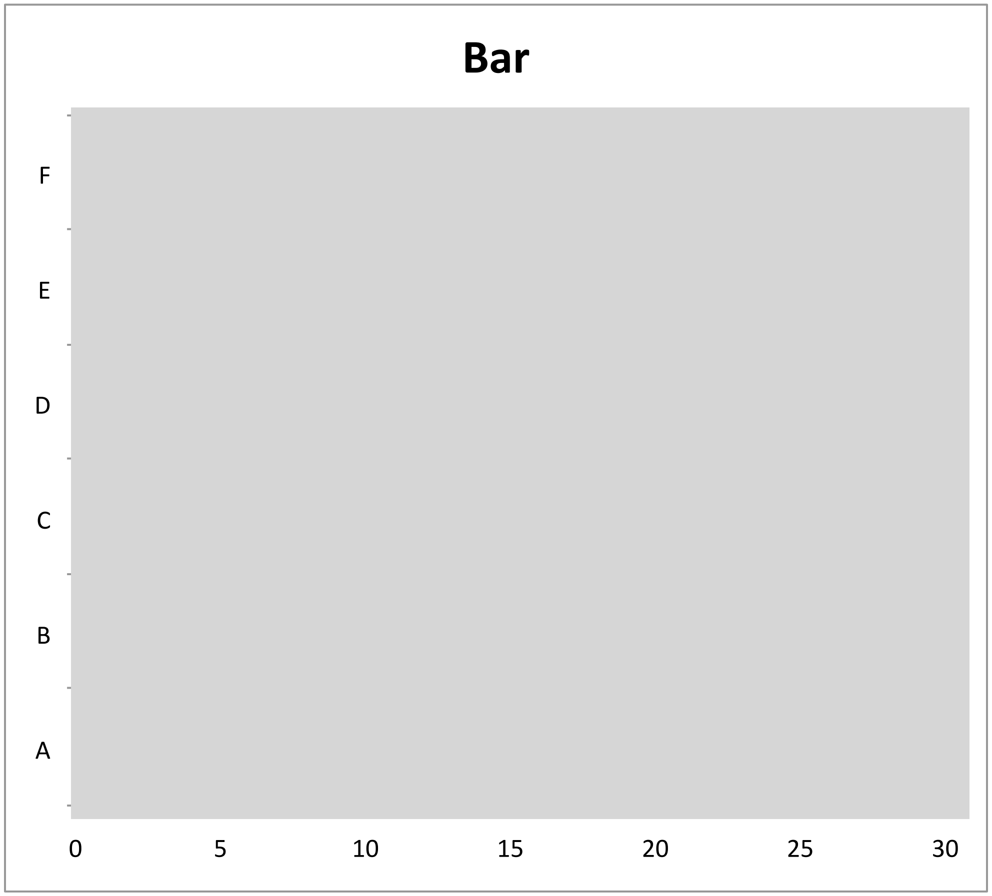
small

user studies

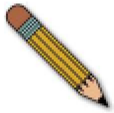
Positions along a common scale



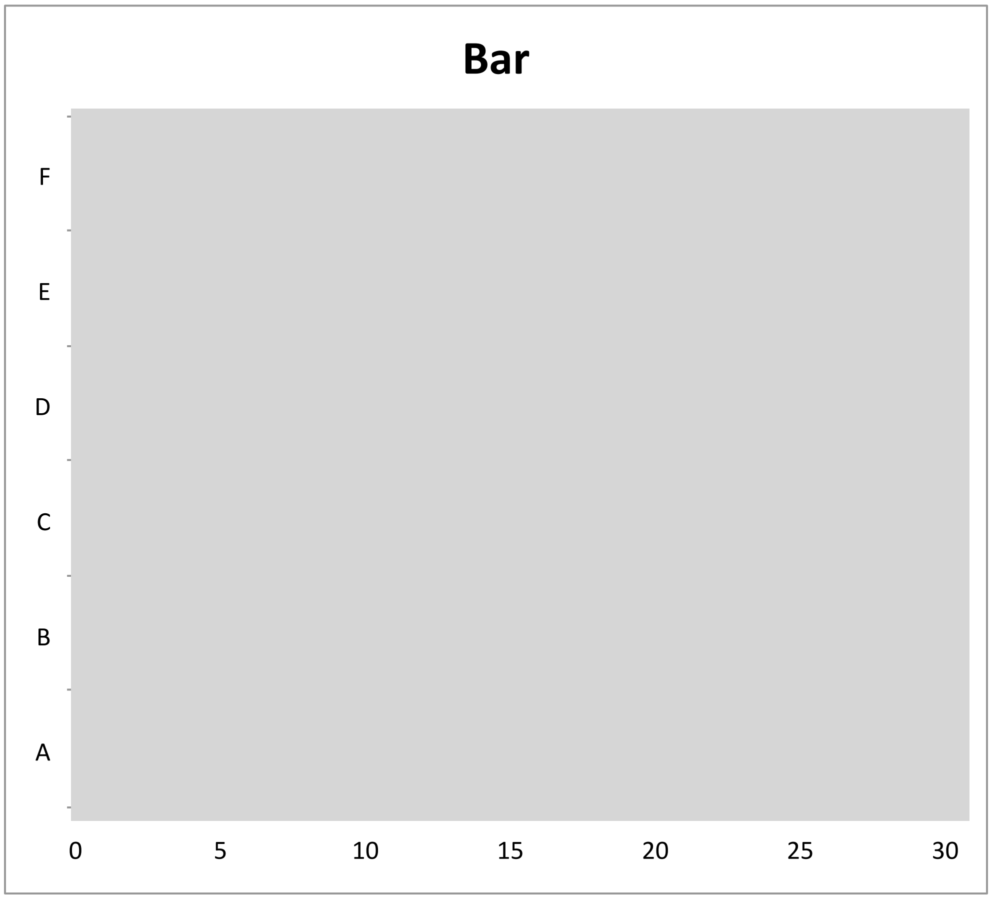
Determine the width for bins A to F as accurately as possible



Positions along a common scale

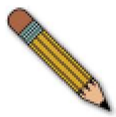


Determine the width for bins A to F as accurately as possible

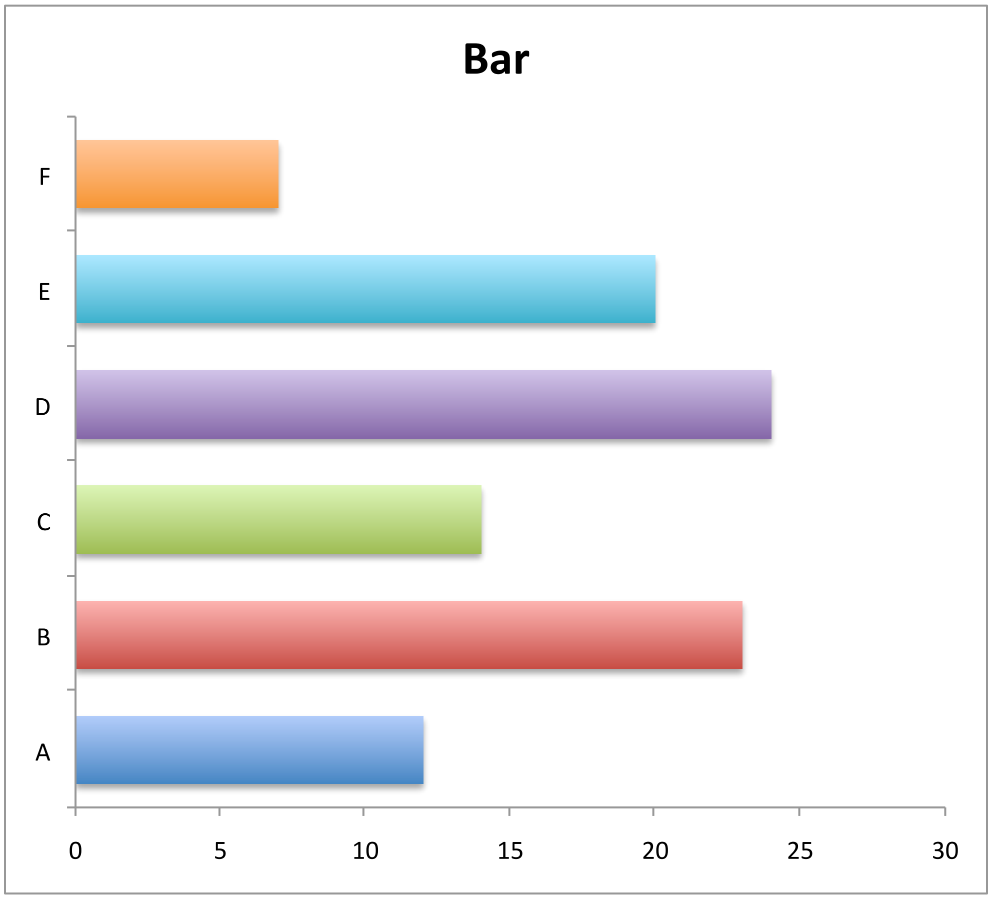


Positions along a common scale

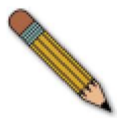
0:20



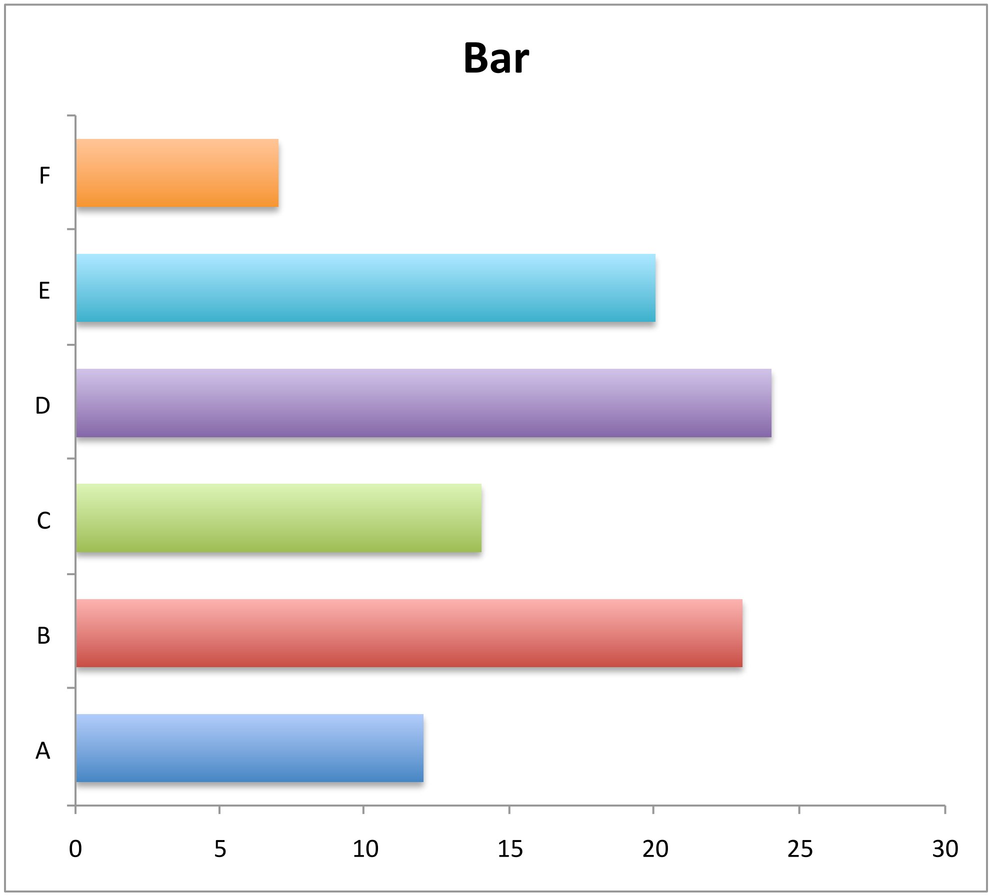
Determine the width for bins A to F as accurately as possible



Positions along a common scale



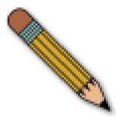
Determine the width for bins A to F as accurately as possible



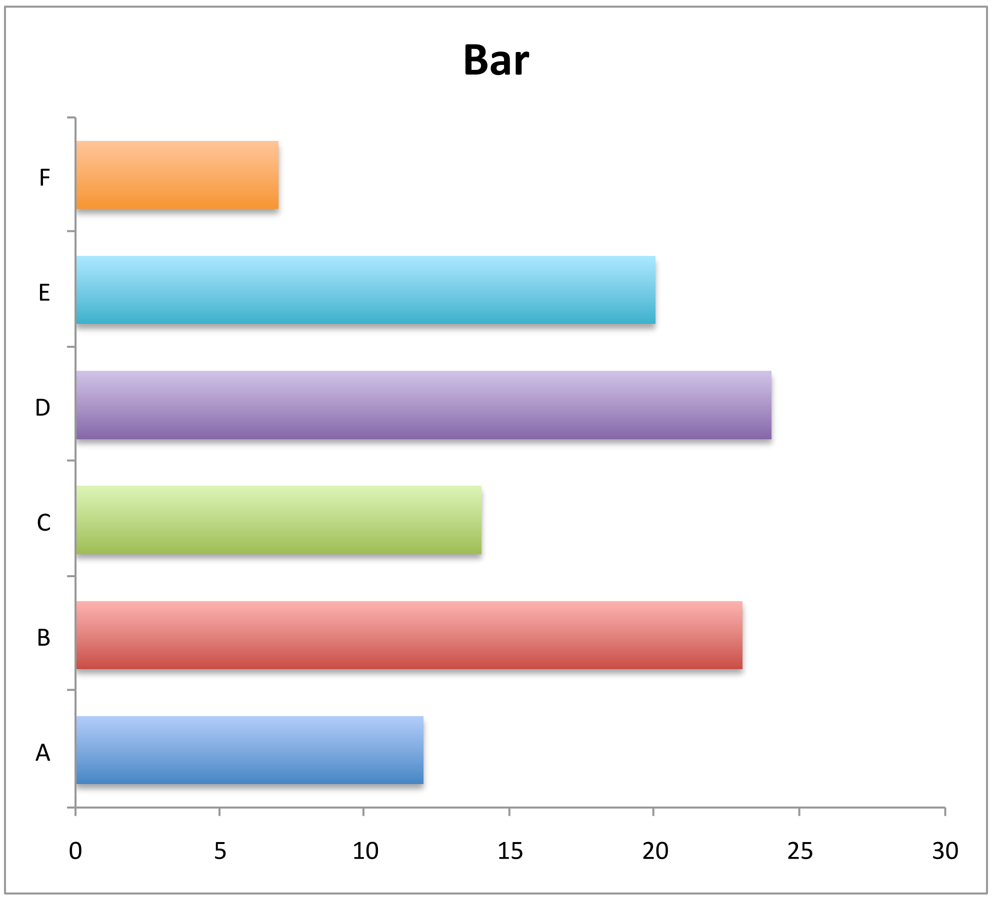
Bin	Value
A	12
B	23
C	14
D	24
E	20
F	7

Positions along a common scale

0:20



Determine the width for bins A to F as accurately as possible



Bin	Value
A	12
B	23
C	14
D	24
E	20
F	7

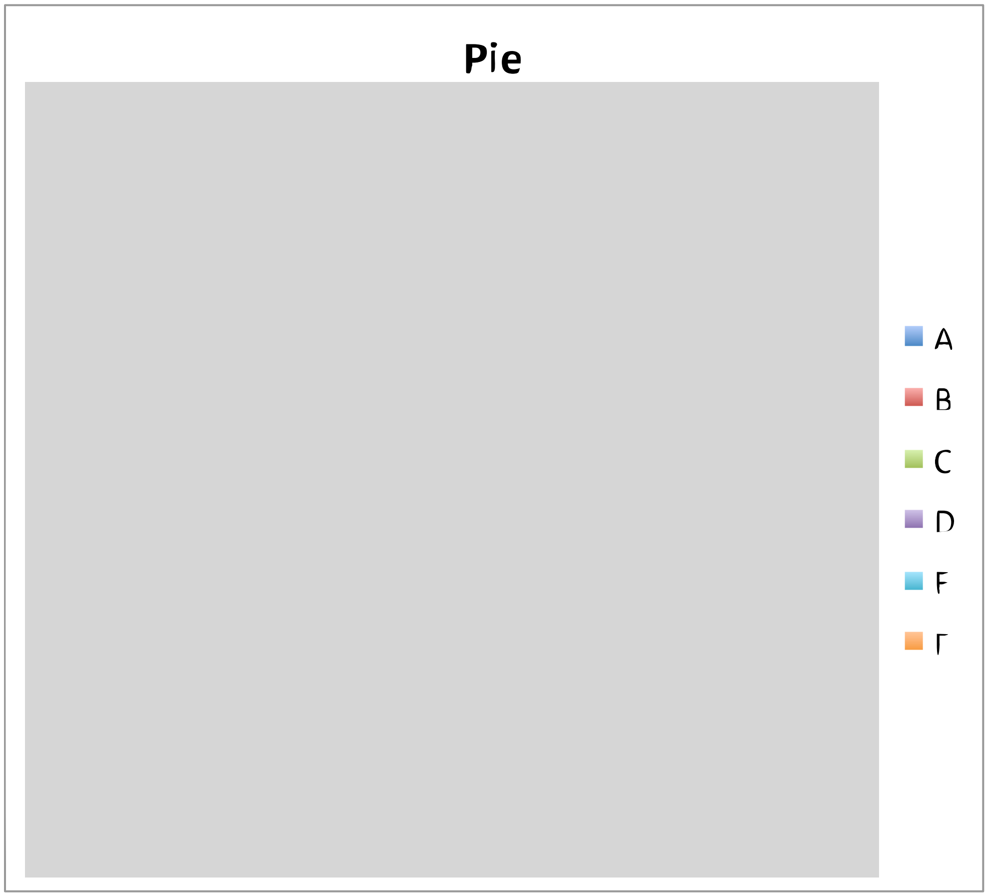
write down (absolute) differences between true values and your estimates

Show of hands: Sum of Errors

- 5 or less?
- 3 or less?
- Accurate?

Angle comparisons

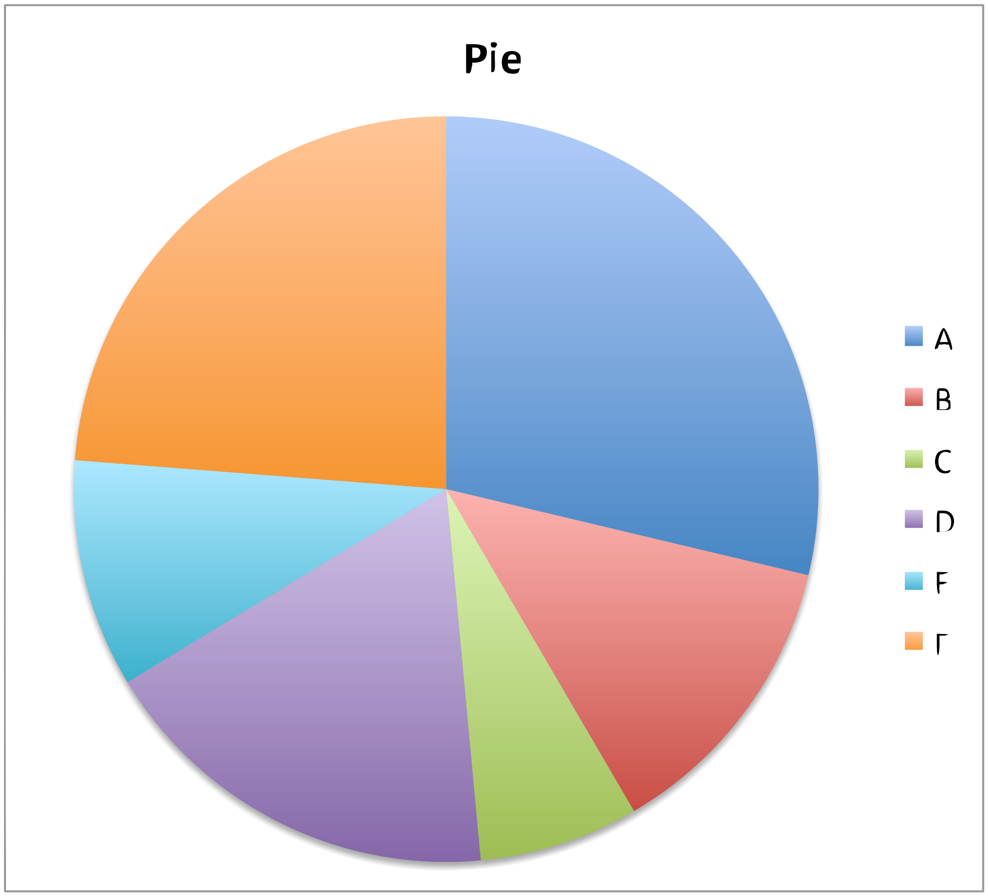
Determine the percentage for slices A to F as accurately as possible



Angle comparisons

0:25

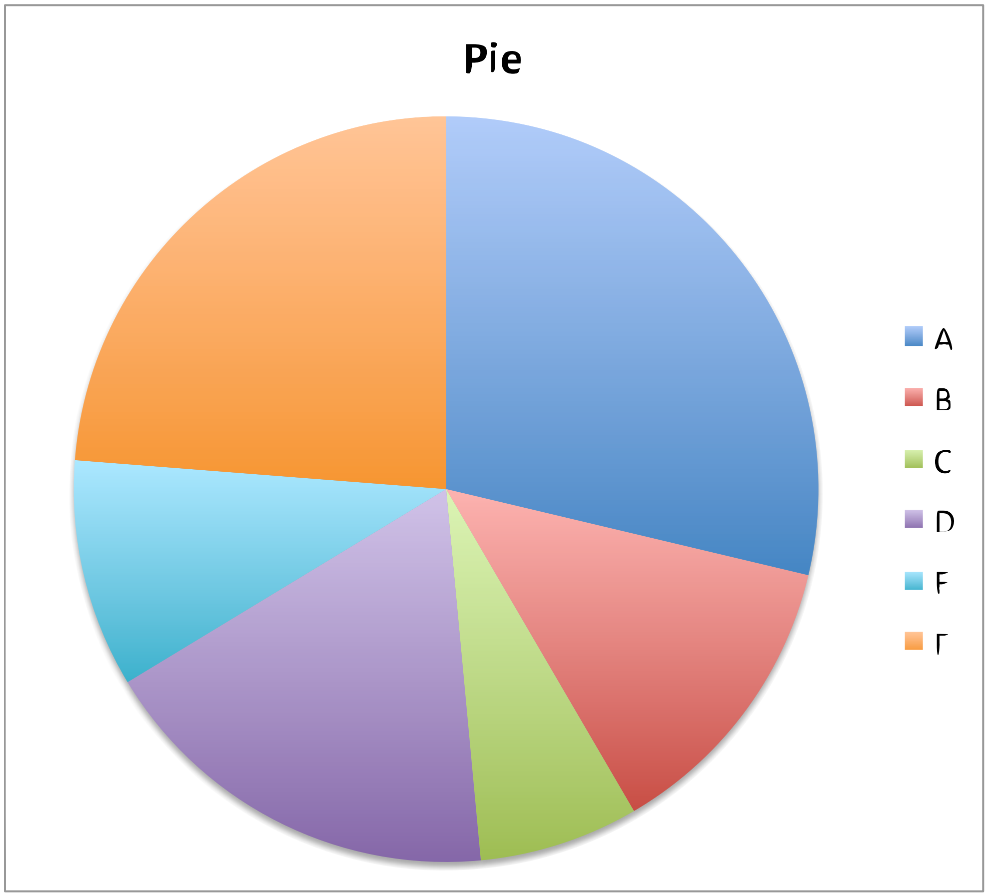
Determine the percentage for slices A to F as accurately as possible



Angle comparisons

0:25

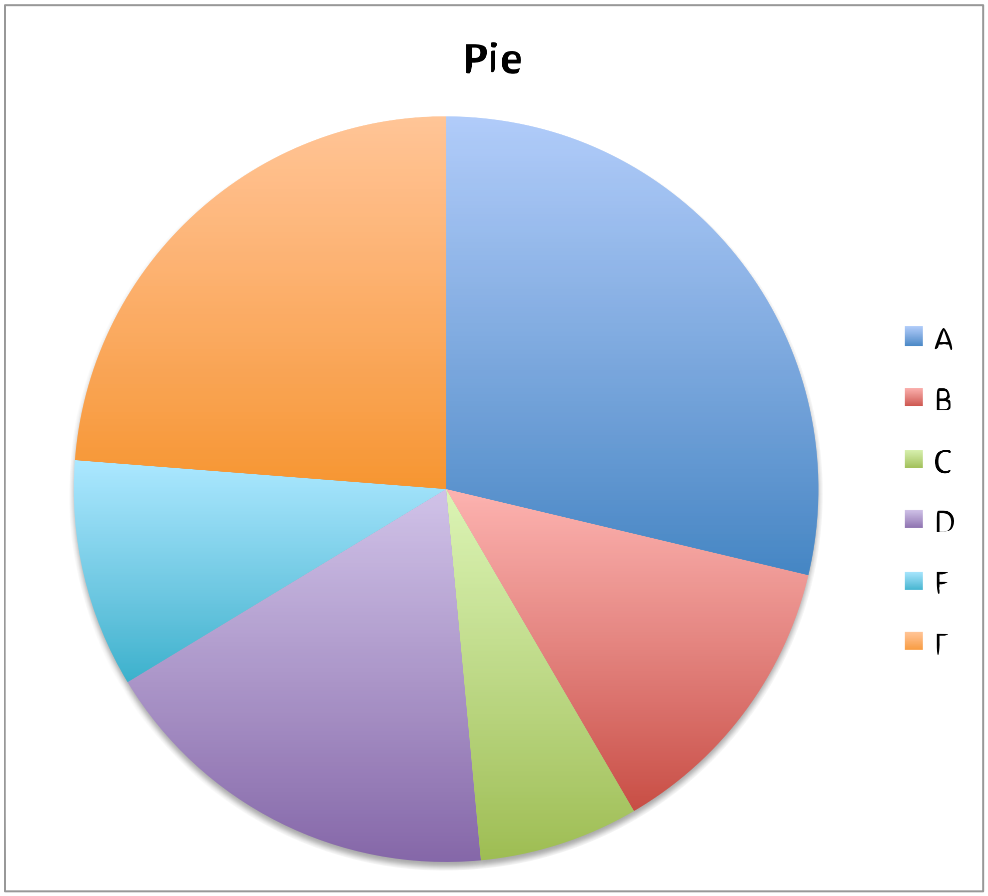
Determine the percentage for slices A to F as accurately as possible



Angle comparisons

0:25

Determine the percentage for slices A to F as accurately as possible

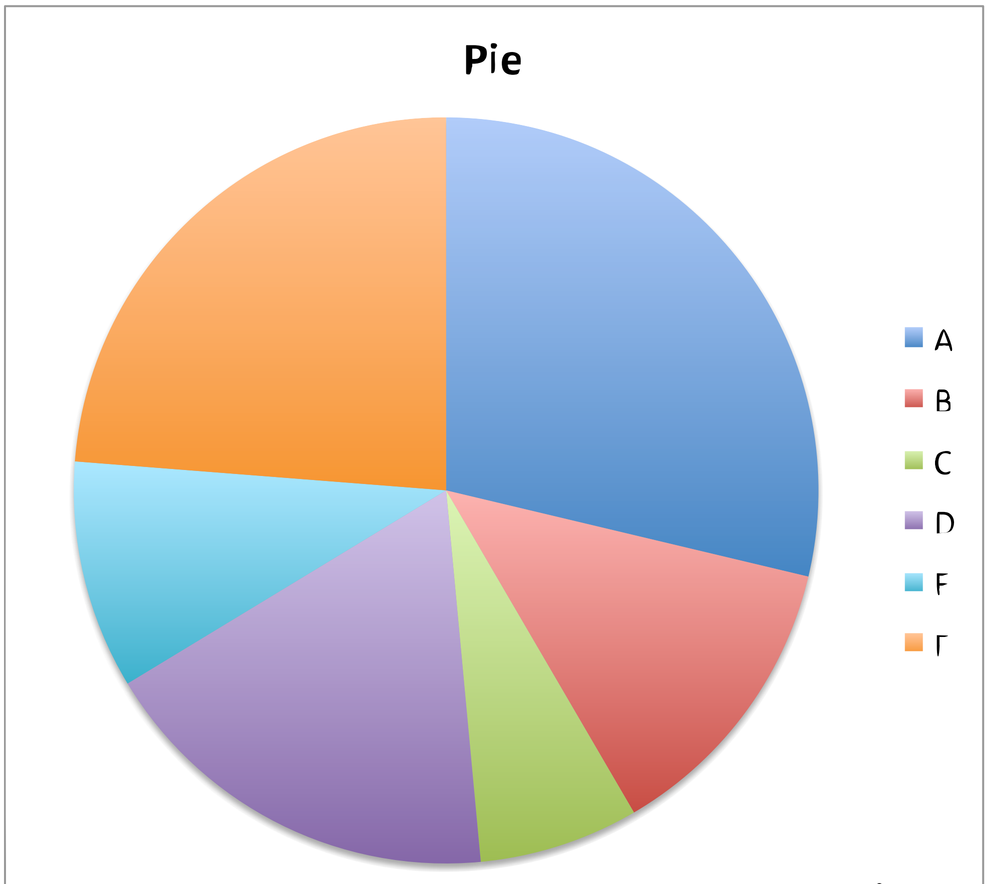


Slice	Value
A	29
B	13
C	7
D	18
E	10
F	24

Angle comparisons

0:25

Determine the percentage for slices A to F as accurately as possible



Slice	Value
A	29
B	13
C	7
D	18
E	10
F	24

write down differences between true values and your estimates

Show of hands: Sum of Errors

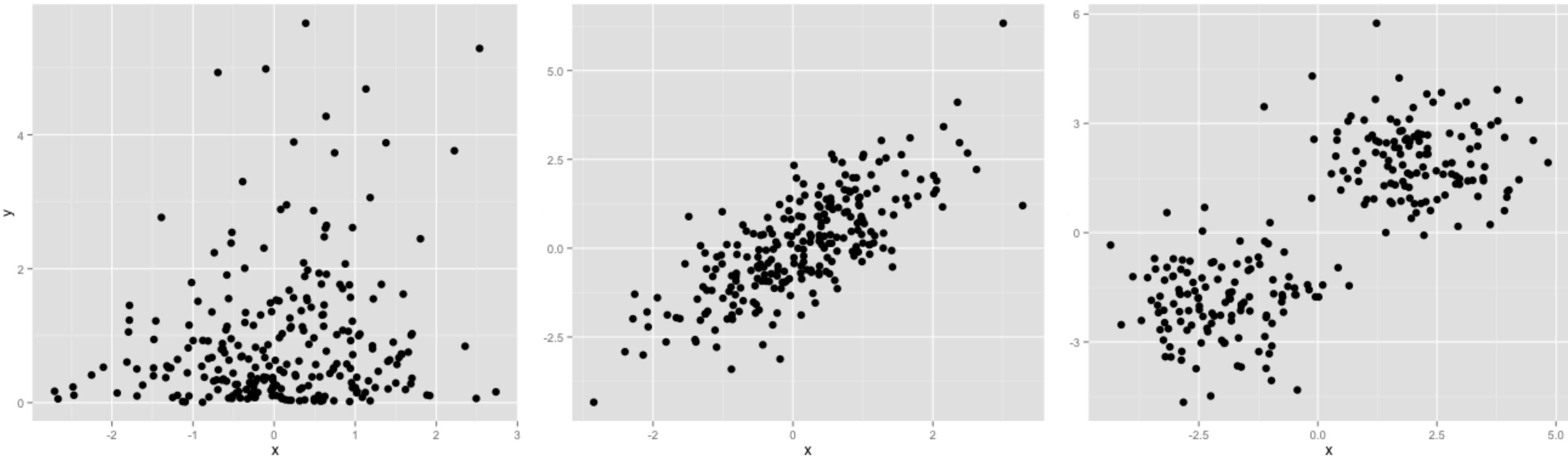
- Ran out of time?
- 5 or less?
- 3 or less?
- Accurate?

Show of hands: Sum of Errors

- Ran out of time?
- 5 or less?
- 3 or less?
- Accurate?

Barcharts give us more accurate results, faster ...

Fact or Artifact?



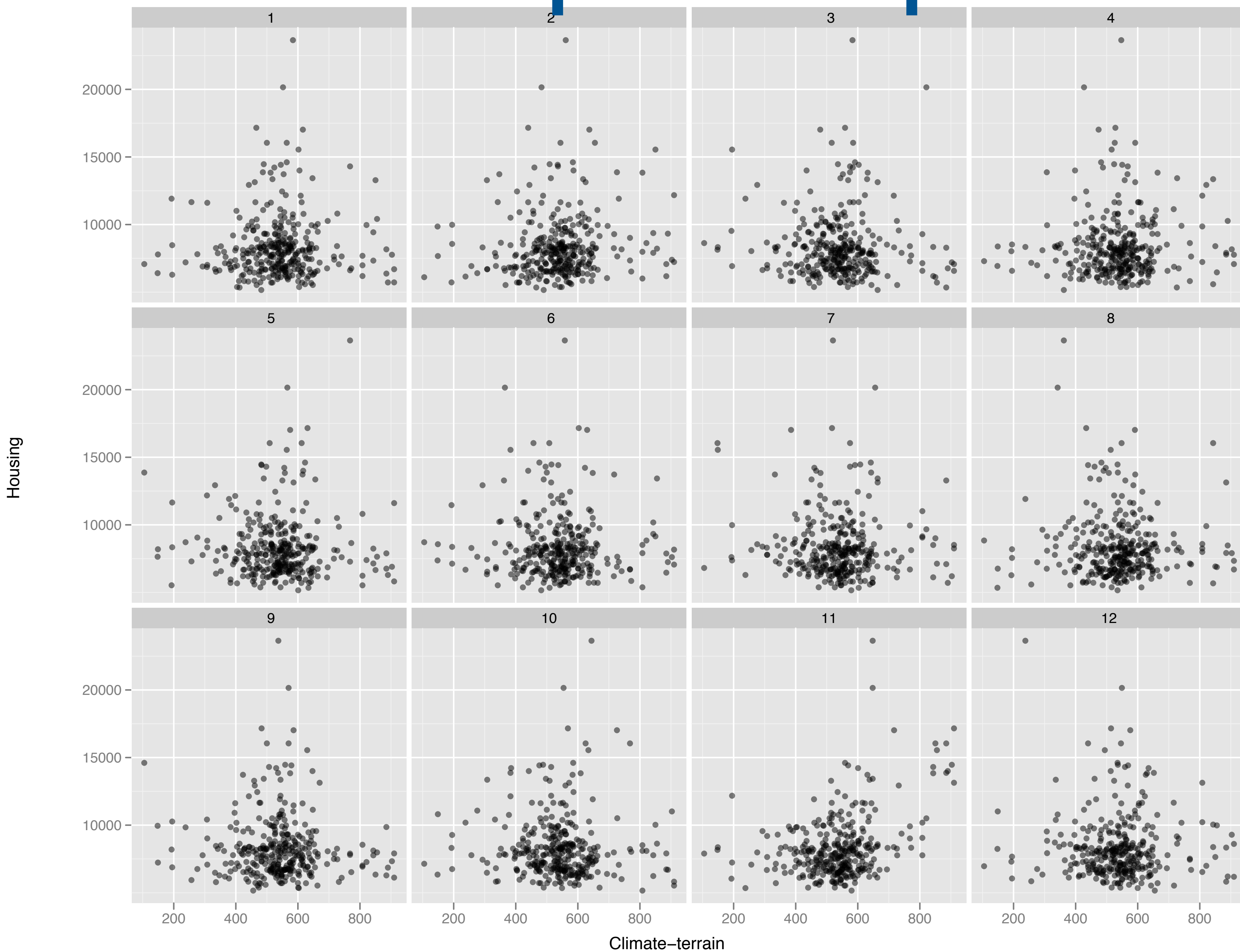
- Is what we see actually there? (or is it just random fluctuation in the data)
- Lineup protocol allows us to quantify significance of visual findings

Lineup Example

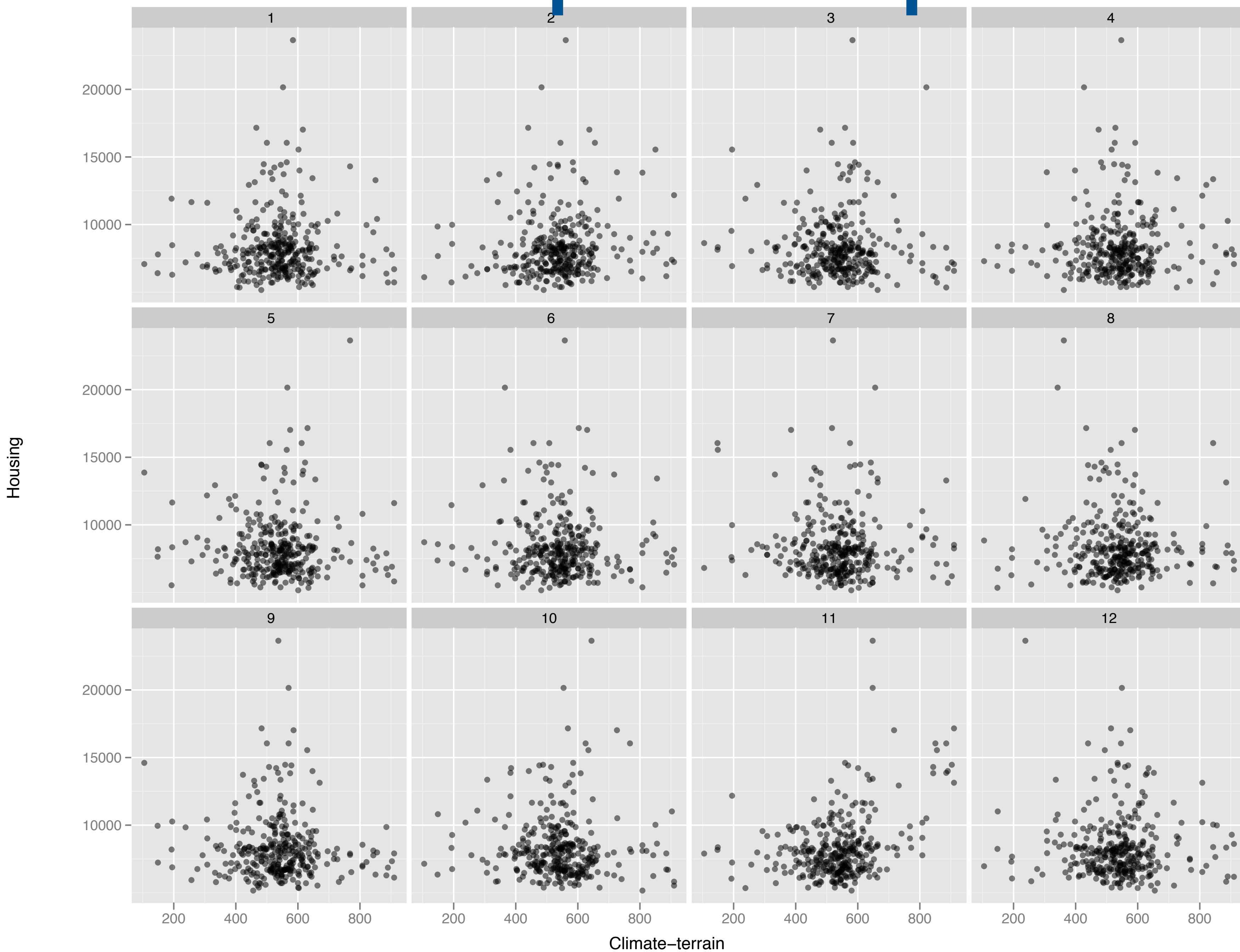
Which plot is the most different?

Lineup Example

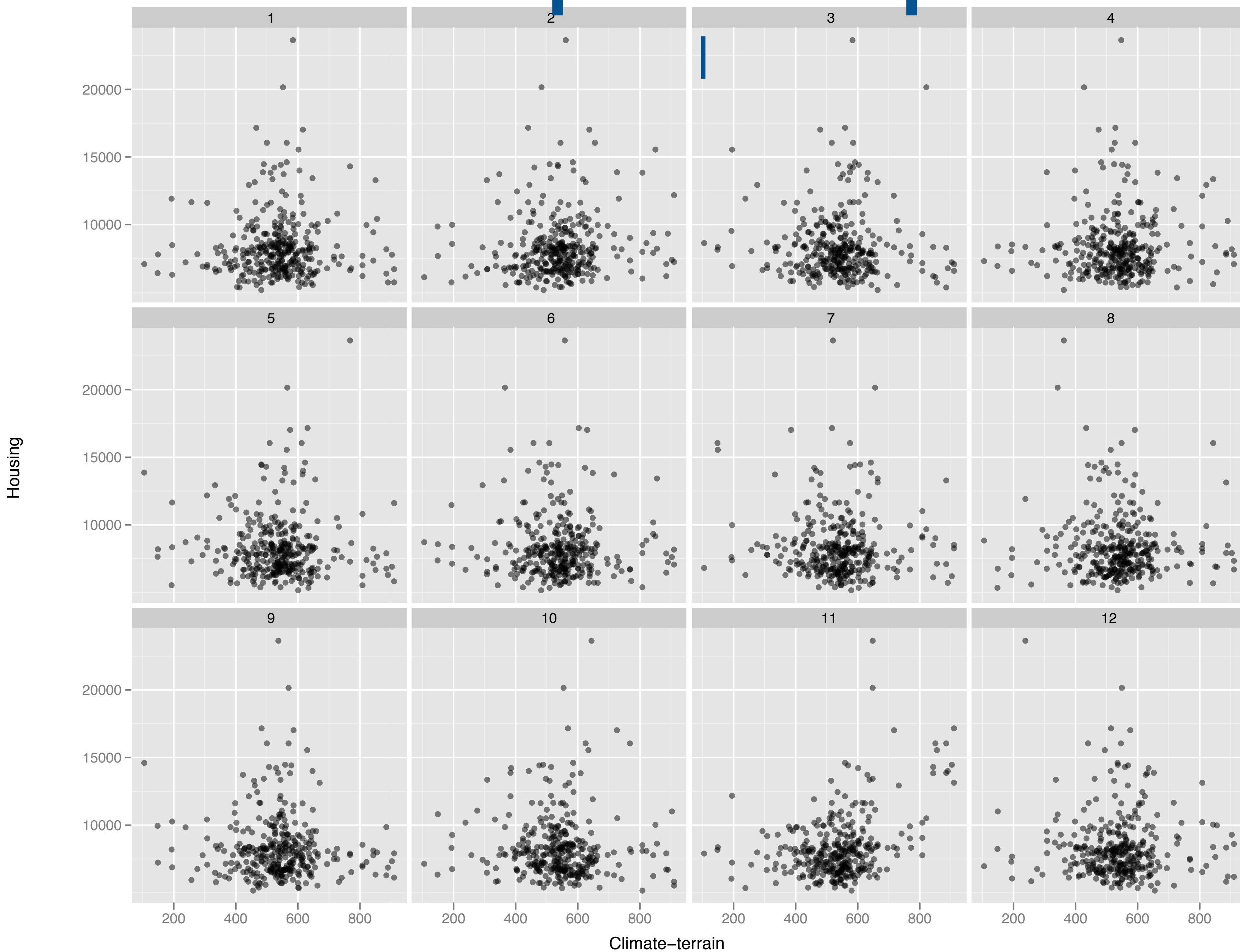
Lineup Example



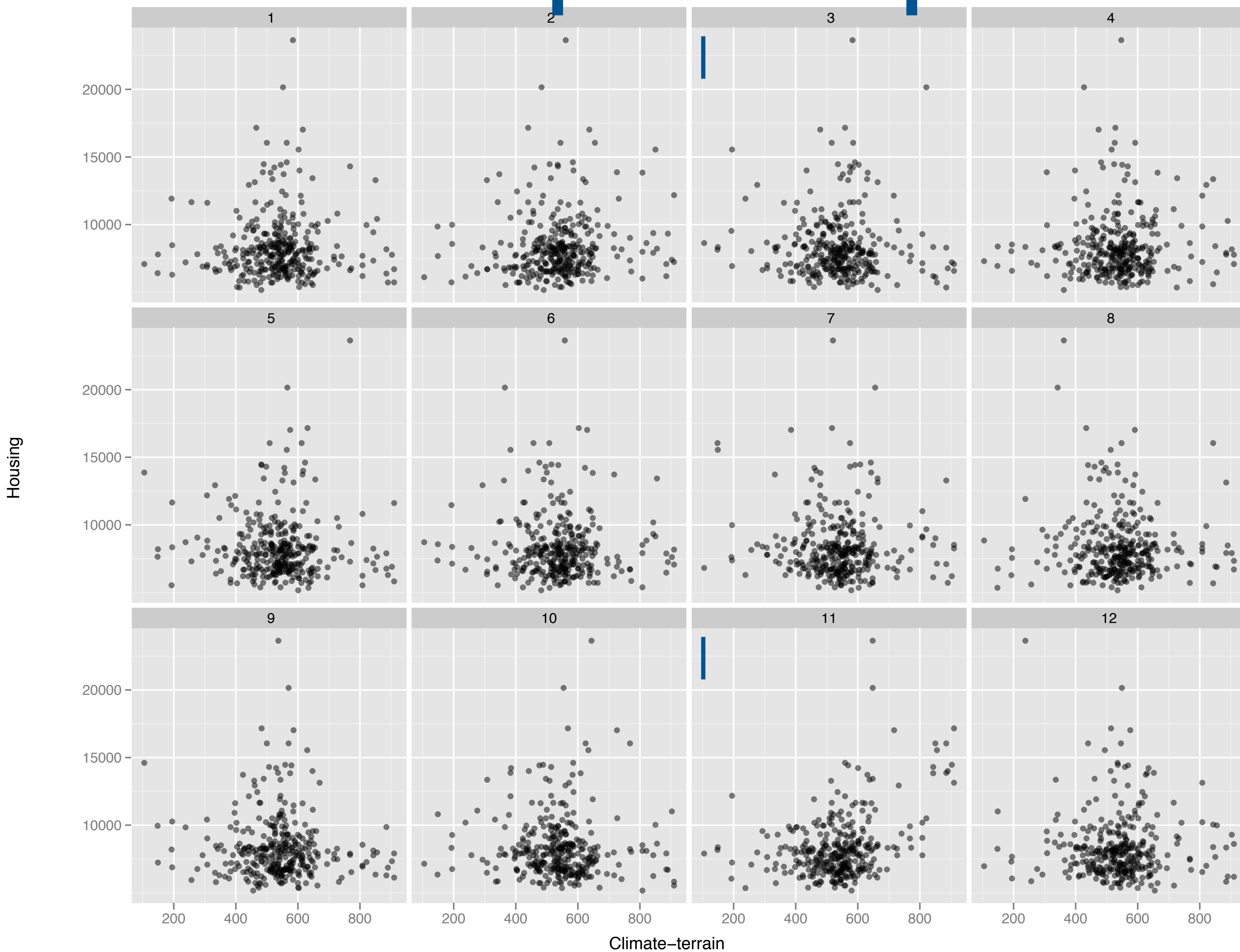
Lineup Example



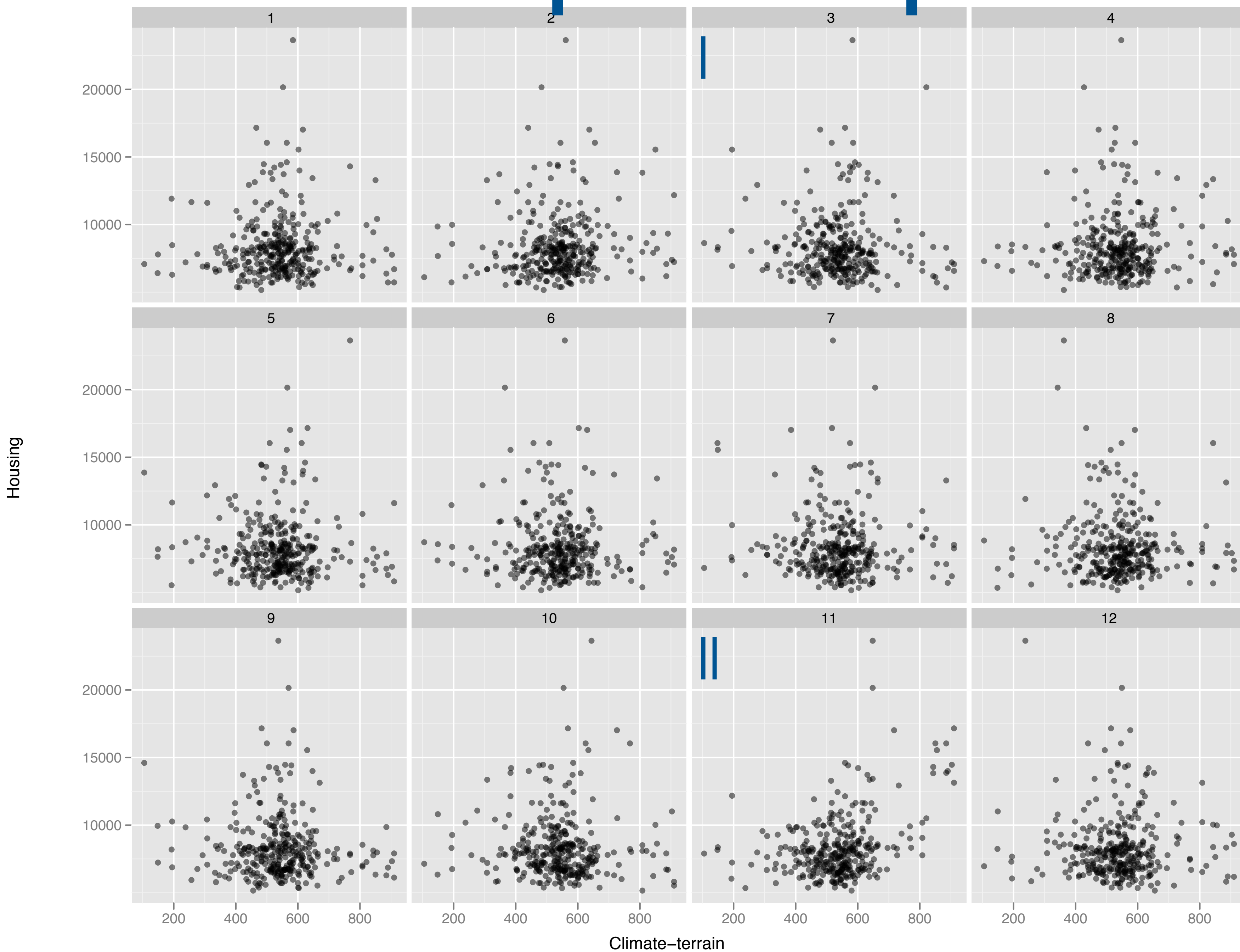
Lineup Example



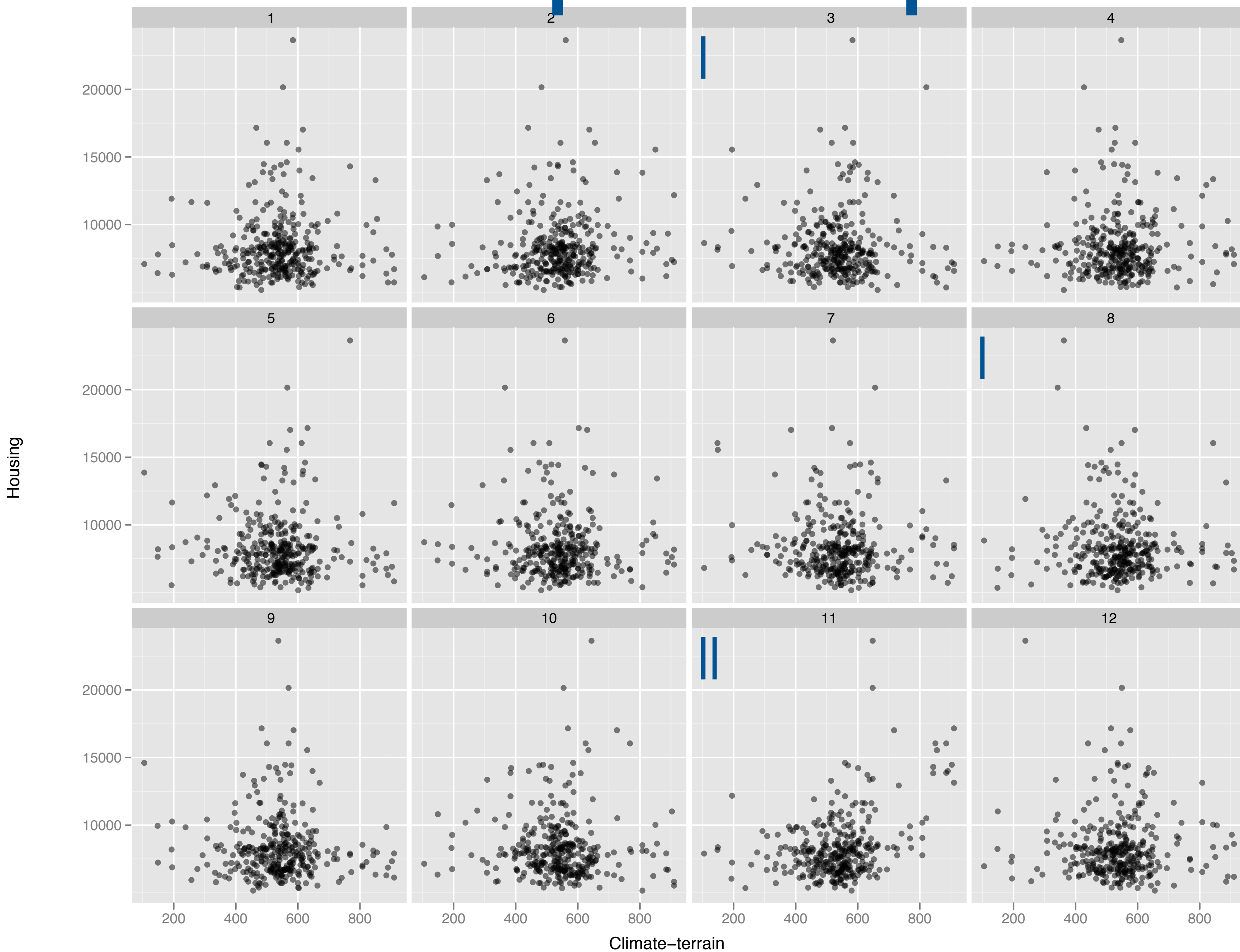
Lineup Example



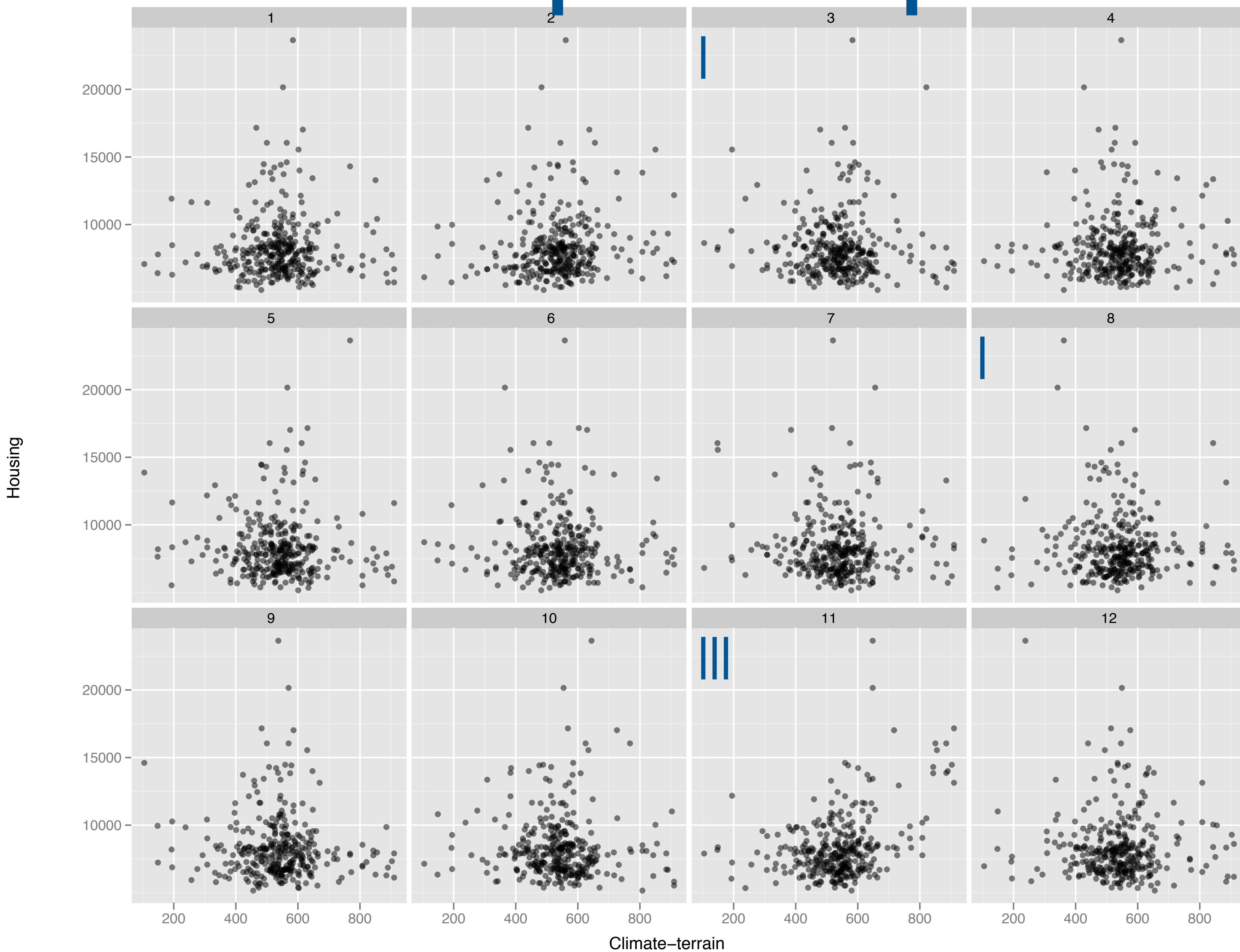
Lineup Example



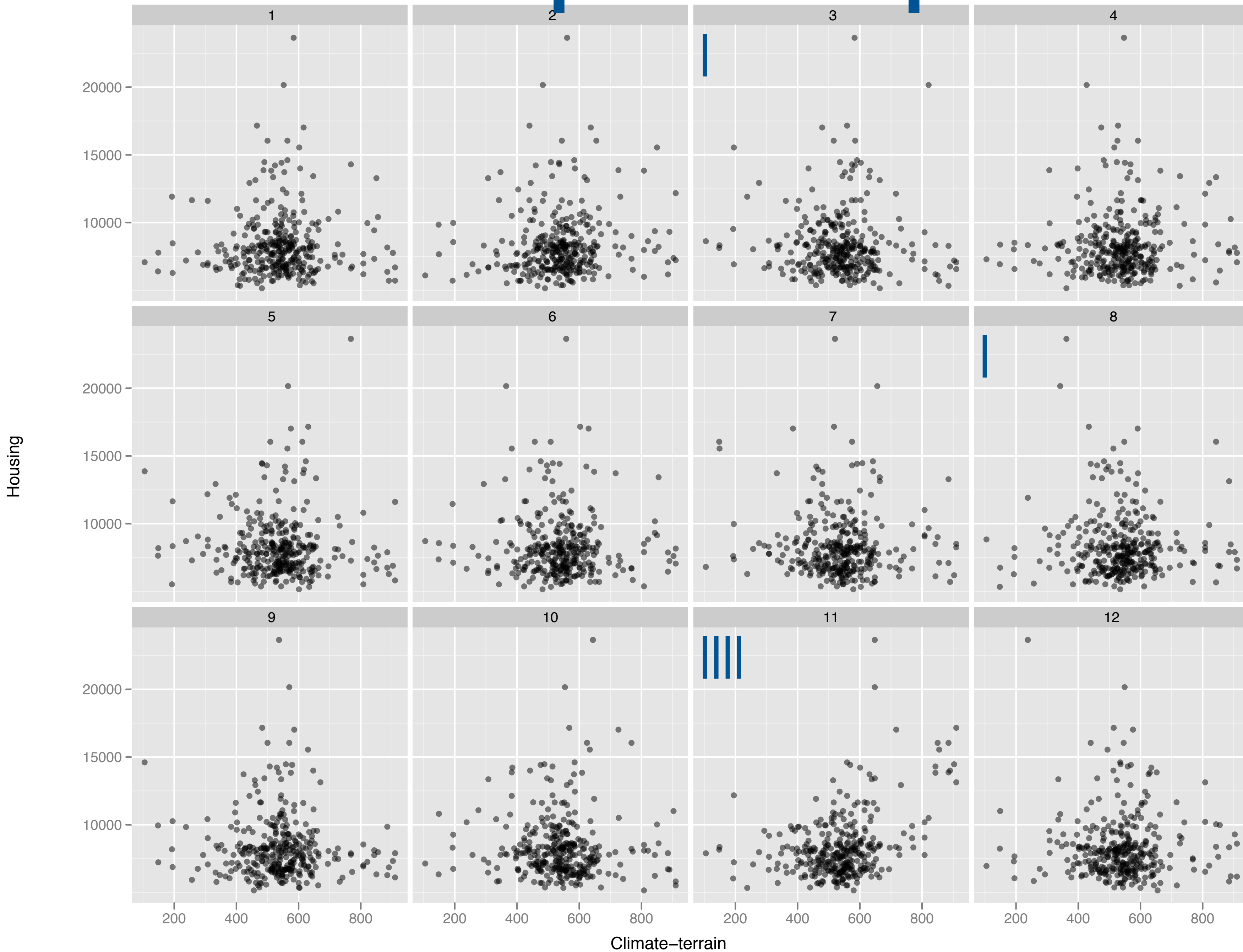
Lineup Example



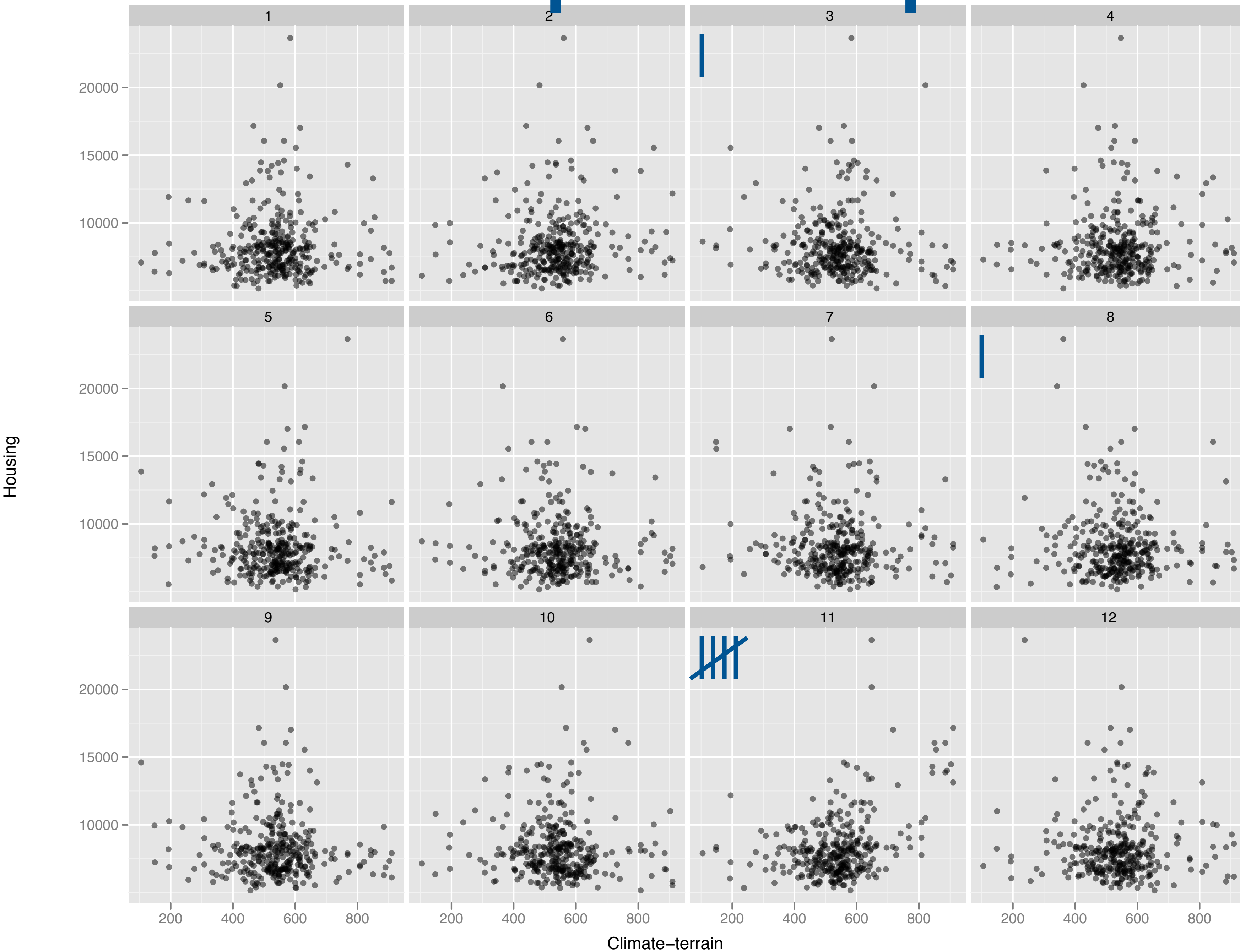
Lineup Example



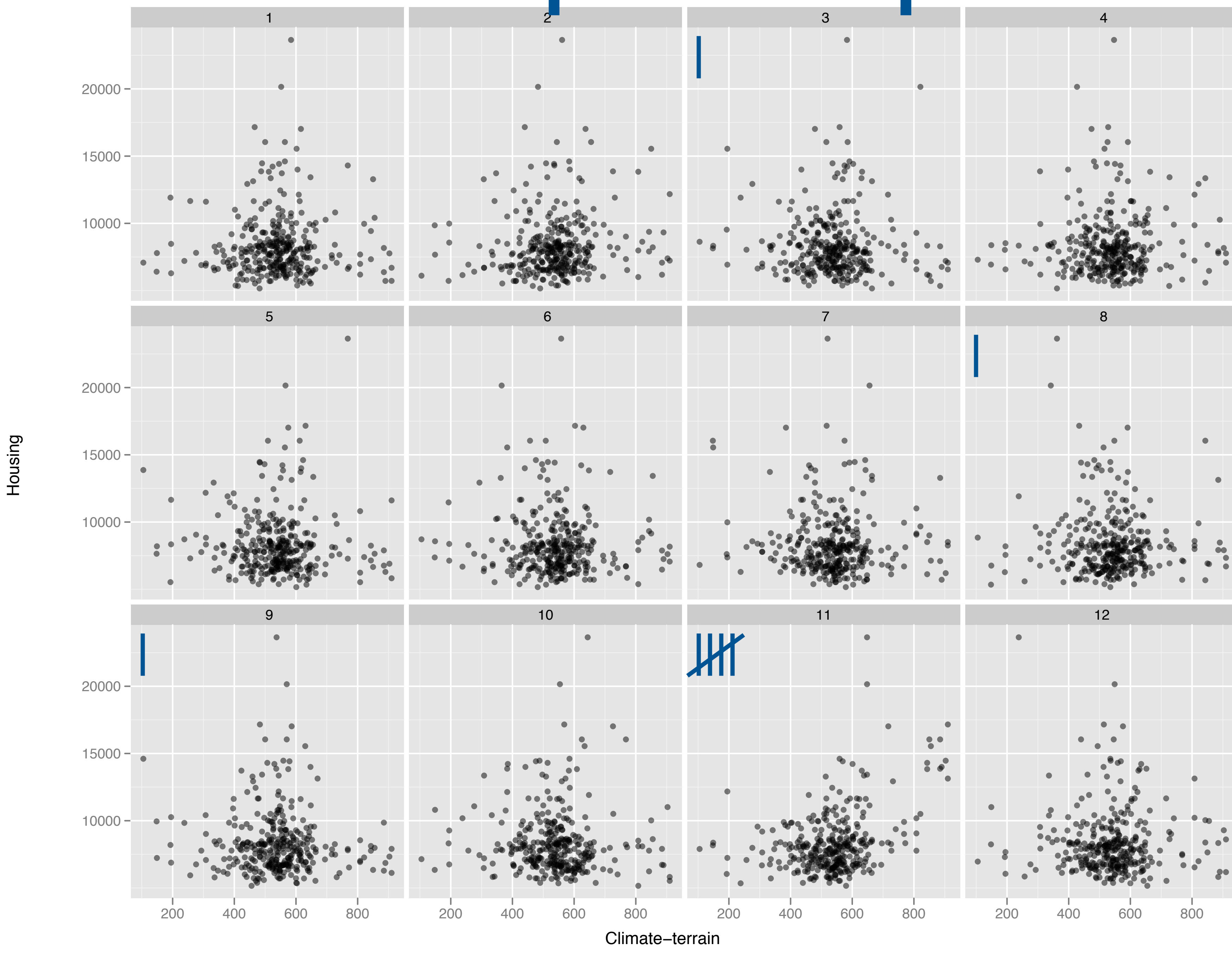
Lineup Example



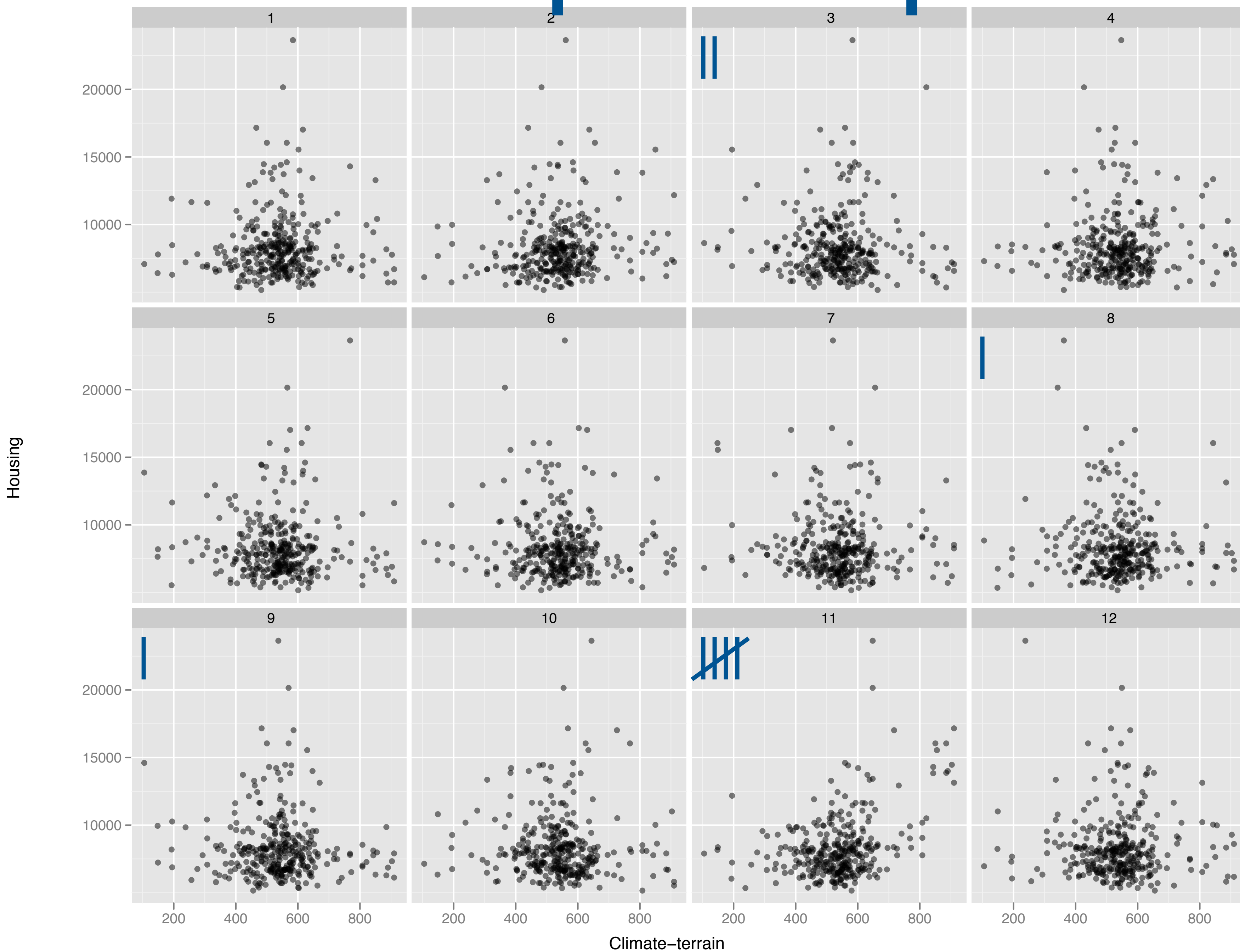
Lineup Example



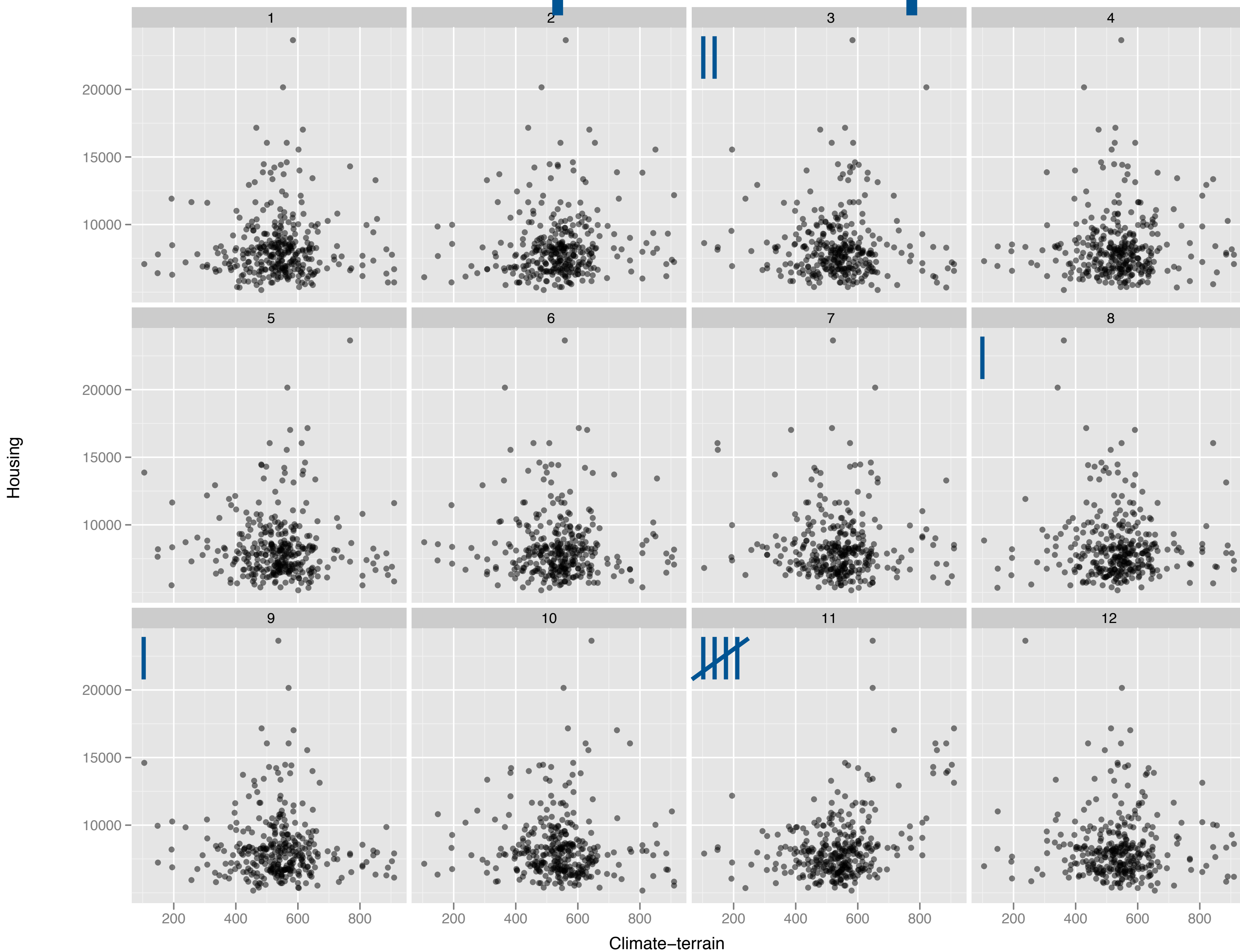
Lineup Example



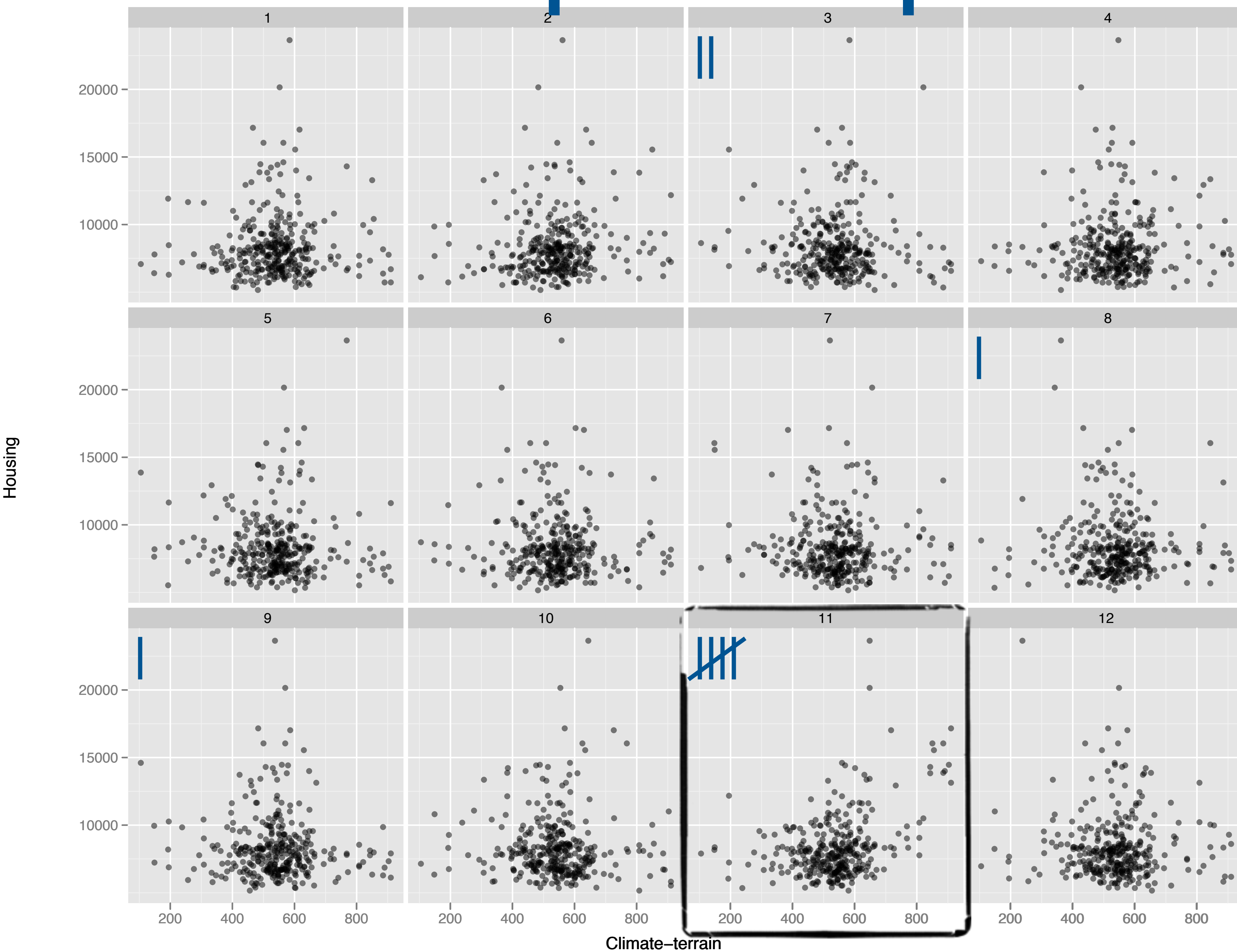
Lineup Example



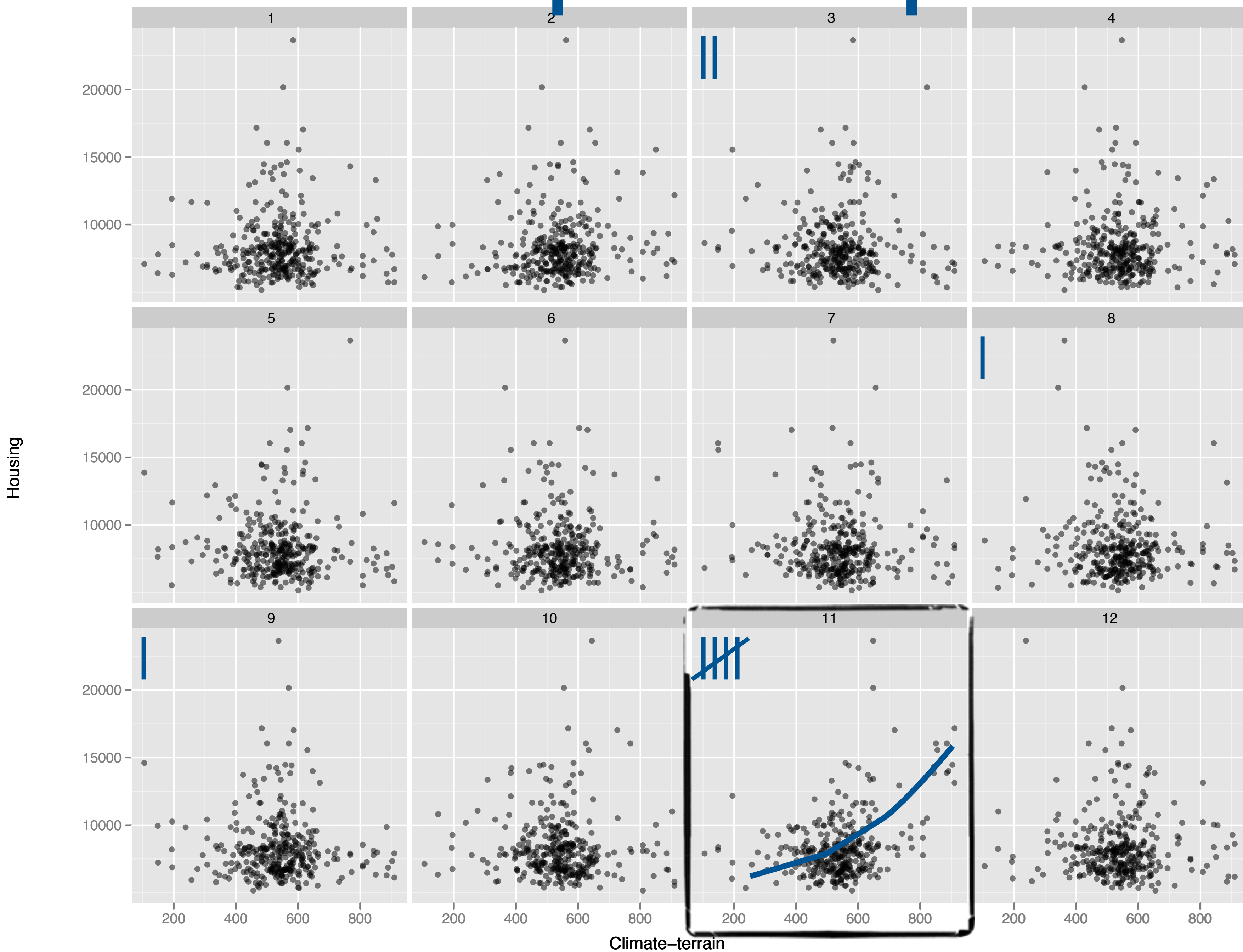
Lineup Example



Lineup Example



Lineup Example



Lineups

- data plot is placed randomly among decoys;
  "police lineup"

- are we able to still identify the data?
  … yes? - that's evidence that the data is different from the decoy plots

- Probability to identify data 'accidentally': 1 in m

- quantify difference as visual p-value:
  Pr(at least x out of n observers identified the data)

$$P(X \geq k) = \sum_{i=k}^{N} \binom{N}{i} \left(\frac{1}{m}\right)^i \left(1 - \frac{1}{m}\right)^{N-i}$$

Lineups

- data plot is placed randomly among decoys; *"police lineup"*
- are we able to still identify the data? … yes? - that's evidence that the data is different from the decoy plots
- Probability to identify data 'accidentally': 1 in m
- quantify difference as visual p-value: Pr(at least x out of n observers identified the data)

1st example: 5 out of 9 responses picked data

$$P(X \geq k) = \sum_{i=k}^{N} \binom{N}{i} \left(\frac{1}{m}\right)^{i} \left(1 - \frac{1}{m}\right)^{N-i}$$

Lineups

- data plot is placed randomly among decoys;
  "police lineup"

- are we able to still identify the data?
  … yes? - that's evidence that the data is different from the decoy plots

- Probability to identify data 'accidentally': 1 in m

- quantify difference as visual p-value:
  Pr(at least x out of n observers identified the data)

$P(X \geq 5) \leq 10^{-4}$

$$P(X \geq k) = \sum_{i=k}^{N} \binom{N}{i} \left(\frac{1}{m}\right)^{i} \left(1 - \frac{1}{m}\right)^{N-i}$$

Power of a design

- Premise: *given a choice of plot designs, that design is better that makes it the easiest for an observer to identify the signal*

- Power: Pr(pick data plot from lineup)

Power of a design

- Premise: *given a choice of plot designs, that design is better that makes it the easiest for an observer to identify the signal*

- Power: Pr(pick data plot from lineup)

5 out of 9 people picked first example:
Power is 5/9

Compare Designs

Simplest Scenario

- One data set, two designs:
  n_1 observers evaluate design 1, x_1 identify data
  n_2 observers evaluate design 2, x_2 identify data
- power $\hat{\pi}_1 = x_1/n_1$ and $\hat{\pi}_2 = x_2/n_2$
- t-test for differences in power:

$$\widehat{\pi}_1 - \widehat{\pi}_2 \pm t_{1-\alpha/2,n-1}\sqrt{\widehat{\pi}_1(1-\widehat{\pi}_1)/n_1 + \widehat{\pi}_2(1-\widehat{\pi}_2)/n_2},$$

More interesting: What affects Power?

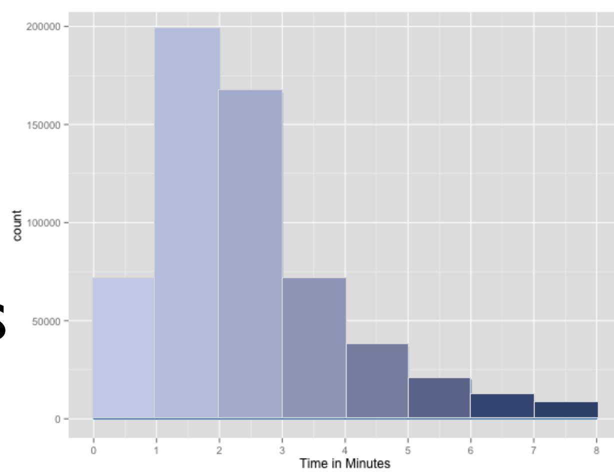
Add in covariates and assess power of

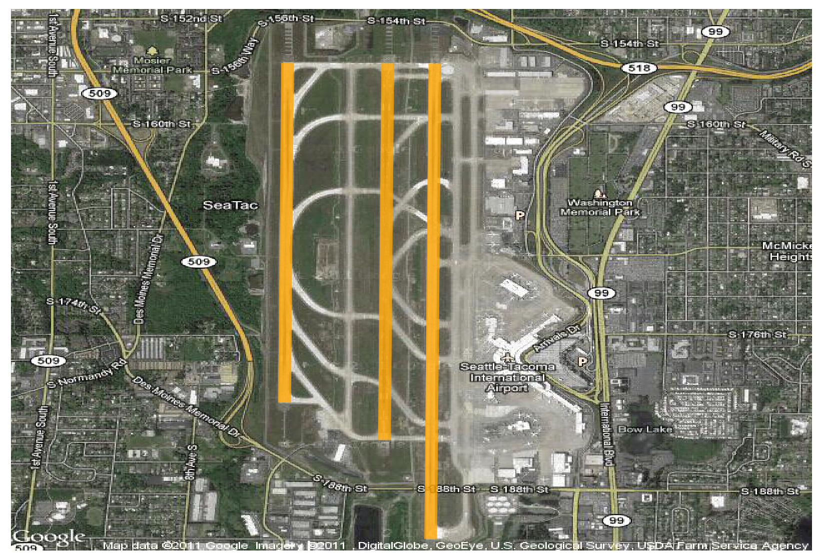
- signal strength
- individuals' visual abilities
- other problem specific properties

Statistical Method:
logistic regression with random effect for individuals

Airport Efficiency and Wind Direction

- Data: Wheel-on and -off events for three years (FAA), combined with weather (wind condition) for each event (restricted to normal operating hours between 6 am and 10 pm)
- results in approx. 500k events
- efficiency: time in mins between wheel events

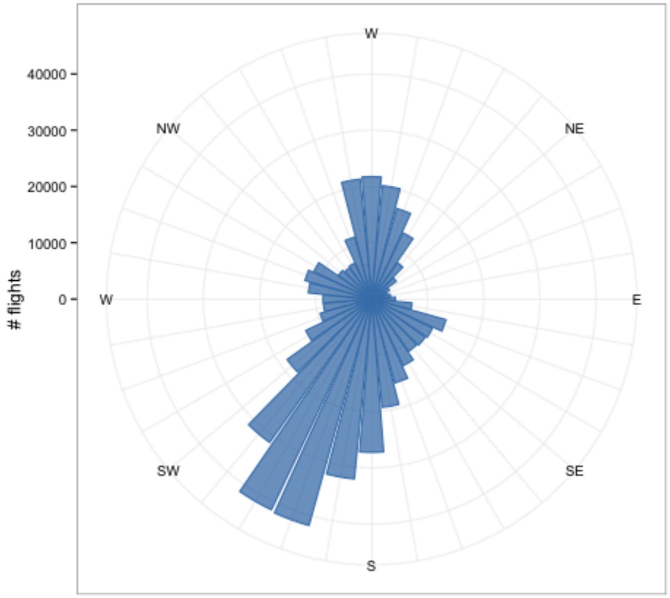




SEA airport

Displaying wind-efficiency relationship

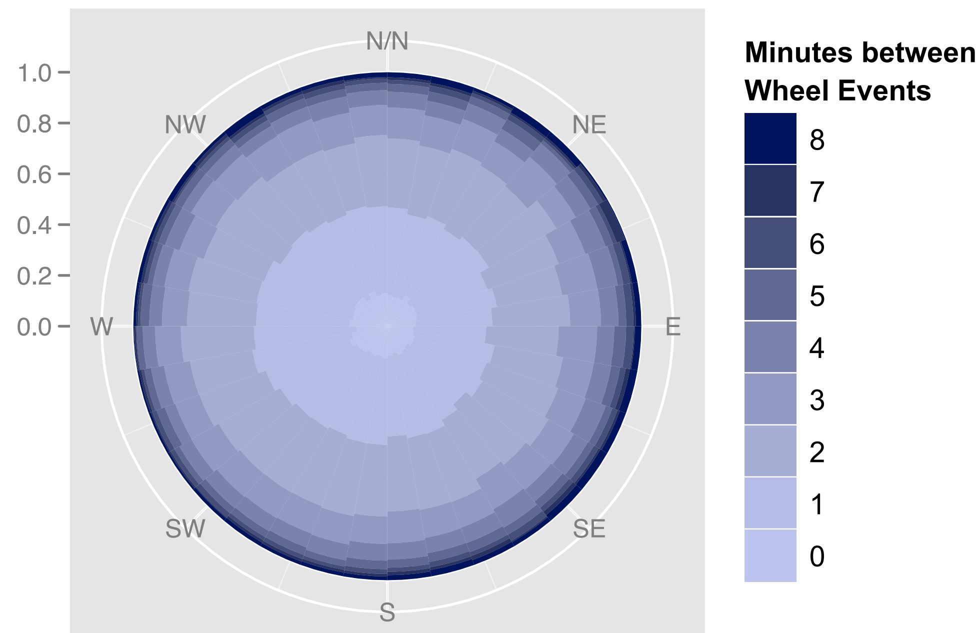
- Wind direction is measured in angles (discrete, in 10 degree intervals)



Wind direction in SEA

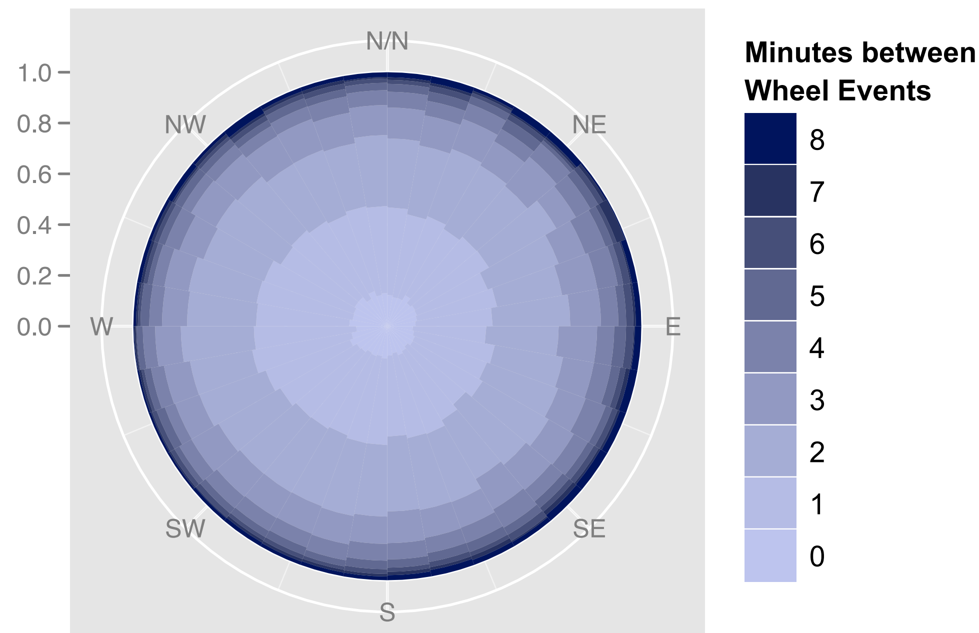
Displaying wind-efficiency relationship

- Wind direction is measured in angles (discrete, in 10 degree intervals)
- Fill color indicates time between wheel events



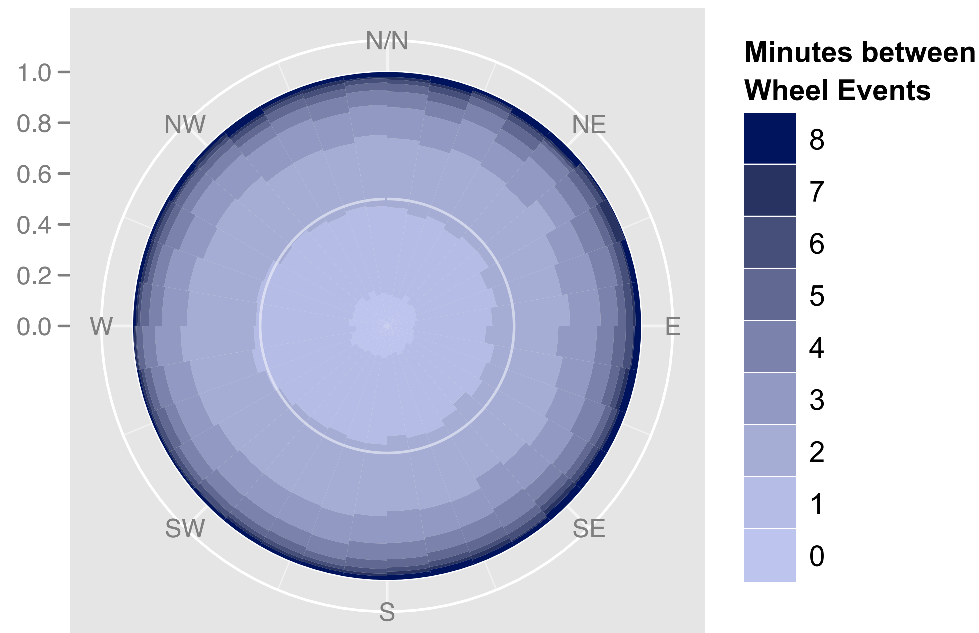
Displaying wind-efficiency relationship

- Wind direction is measured in angles (discrete, in 10 degree intervals)
- Fill color indicates time between wheel events
- Additional white helper line



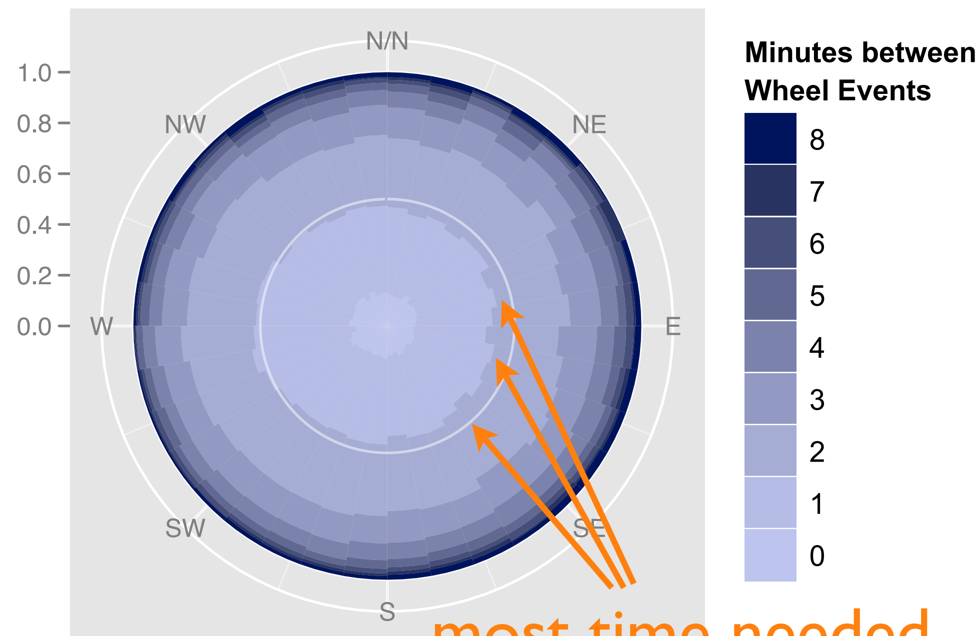
Displaying wind-efficiency relationship

- Wind direction is measured in angles (discrete, in 10 degree intervals)
- Fill color indicates time between wheel events
- Additional white helper line



Displaying wind-efficiency relationship

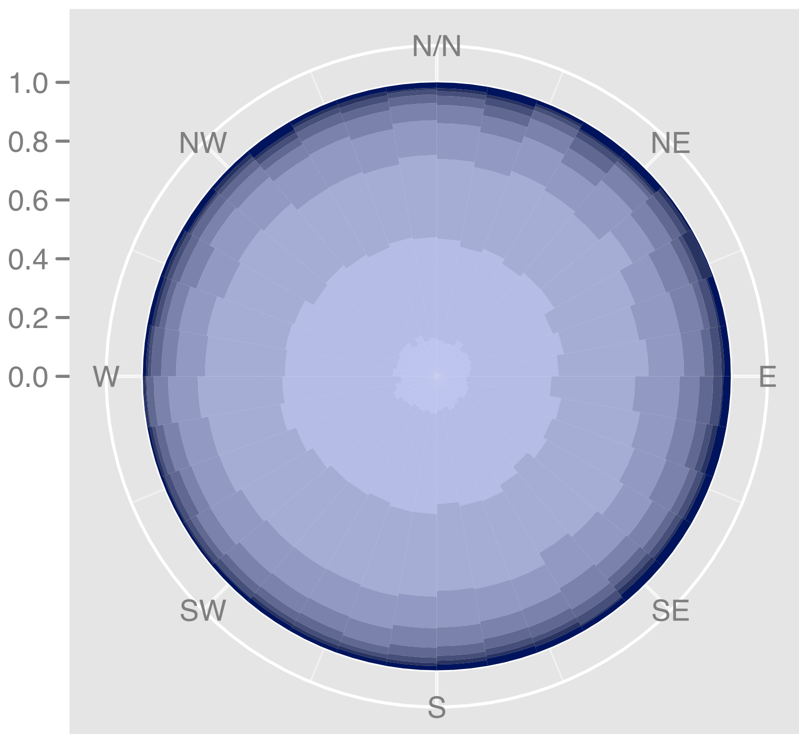
- Wind direction is measured in angles (discrete, in 10 degree intervals)
- Fill color indicates time between wheel events
- Additional white helper line

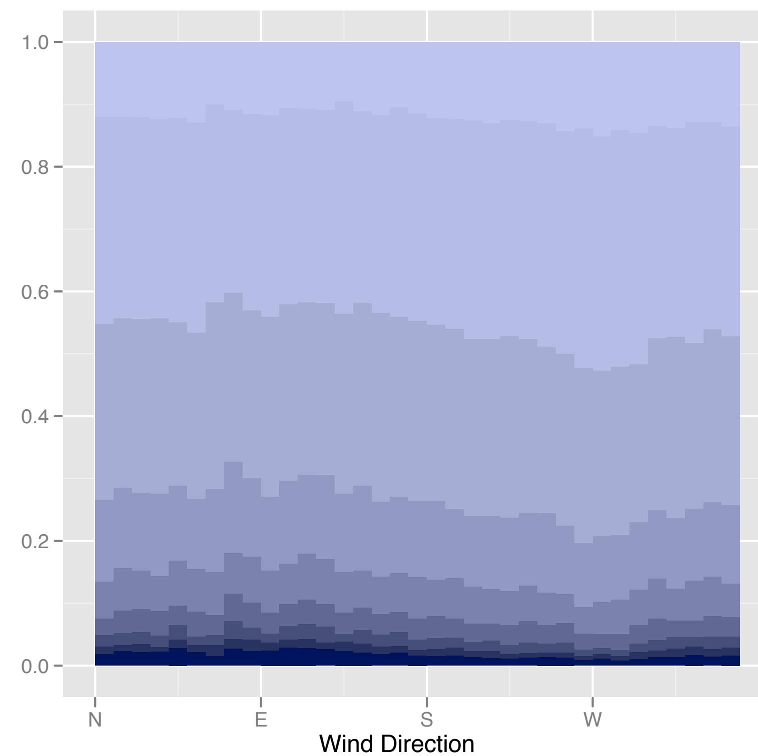


most time needed for these directions

Displaying wind-efficiency relationship

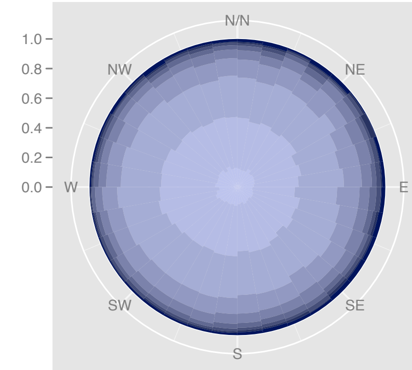
- Orthogonal instead of polar layout:

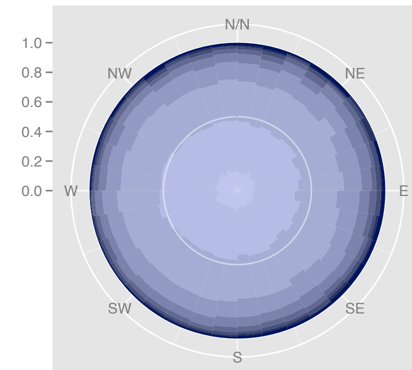


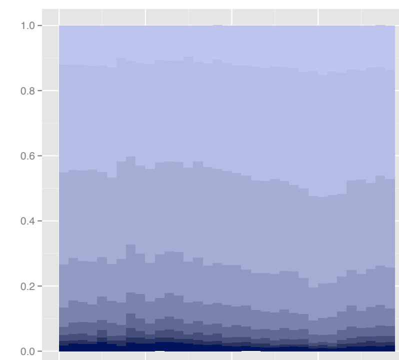


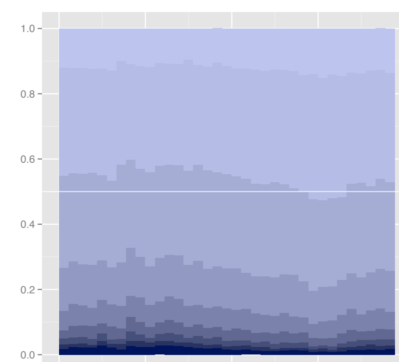
Designs & Experimental Setup

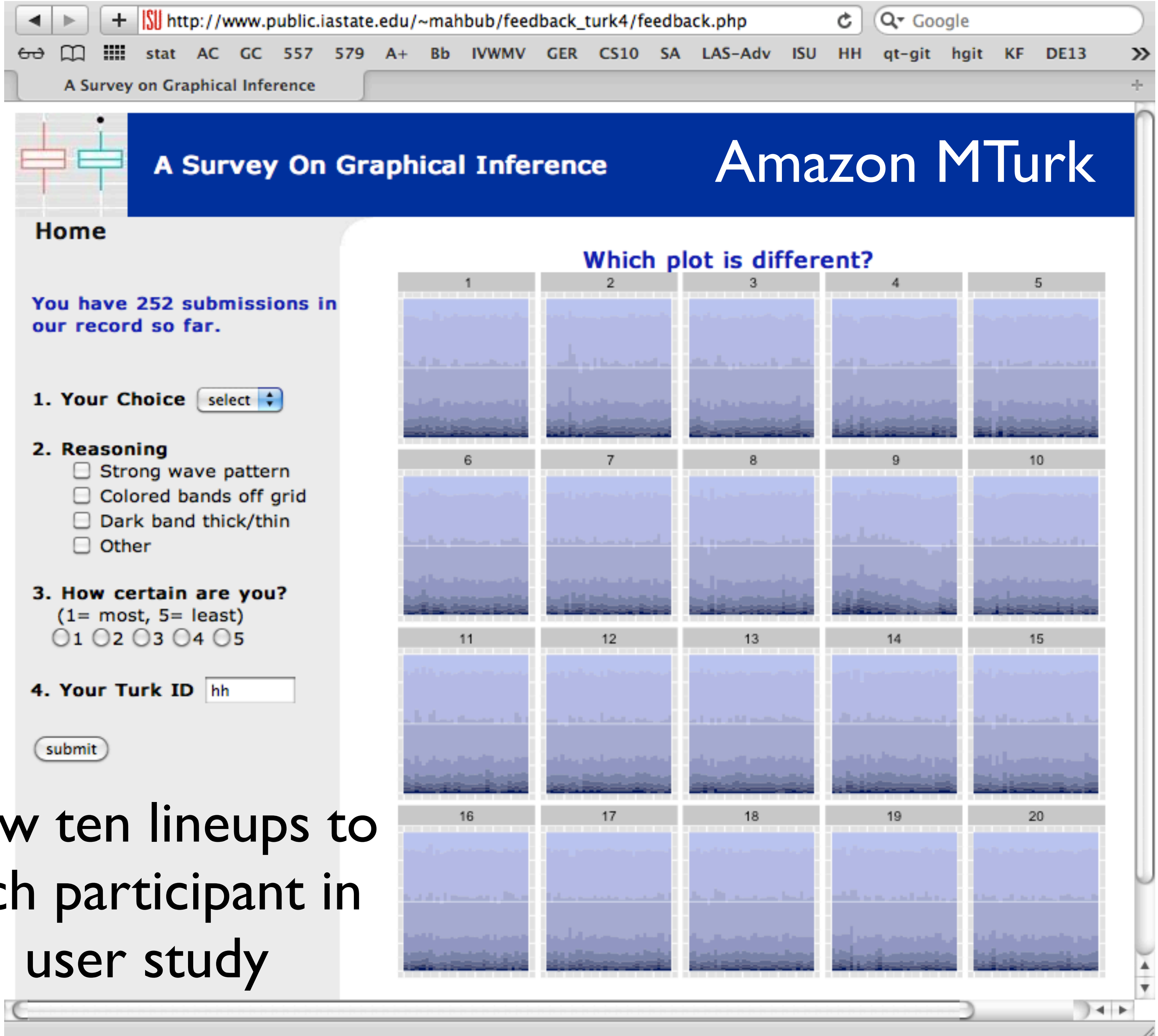
- design: polar versus orthogonal
  with and without grid lines
- sample size (in %): 2, 4, 6, 8, 10, 24
- shifts in direction (in °): 0, 90, 180, 270
- two replicates each
- results in 192 different plots, included in as many lineups











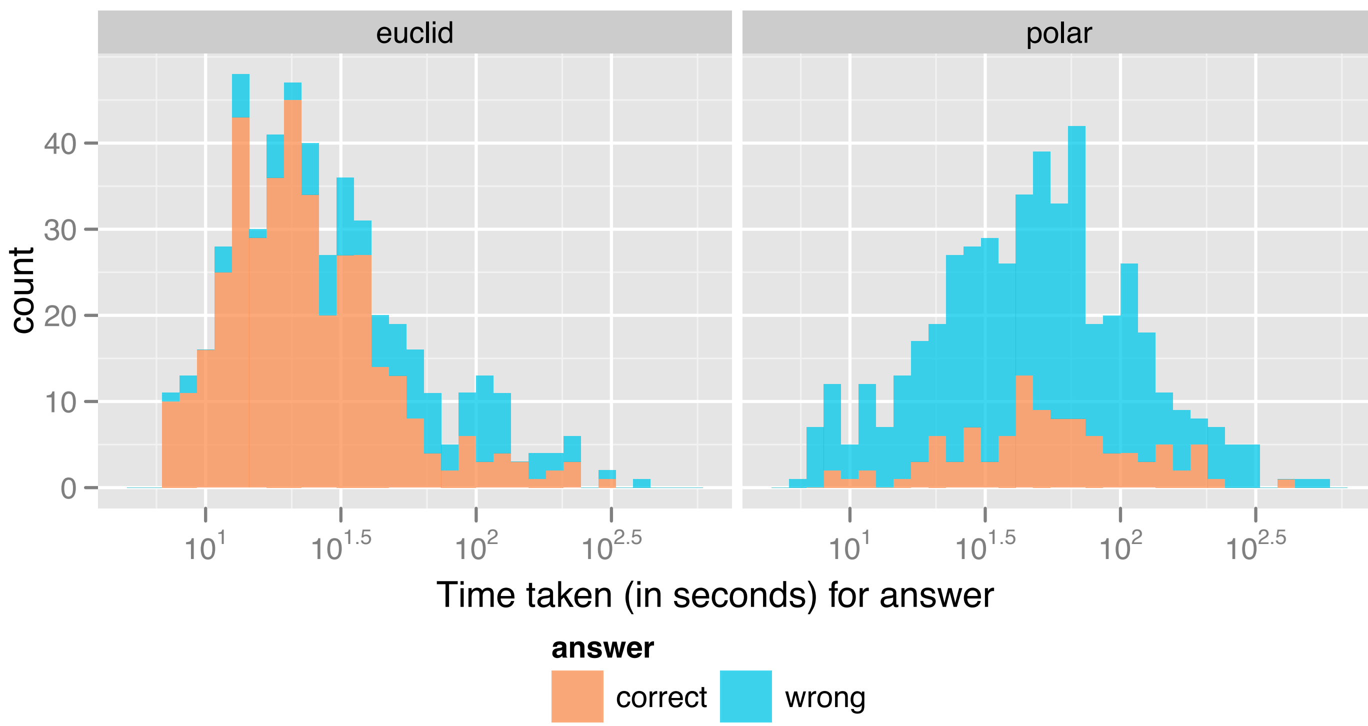
show ten lineups to each participant in user study

Evaluation

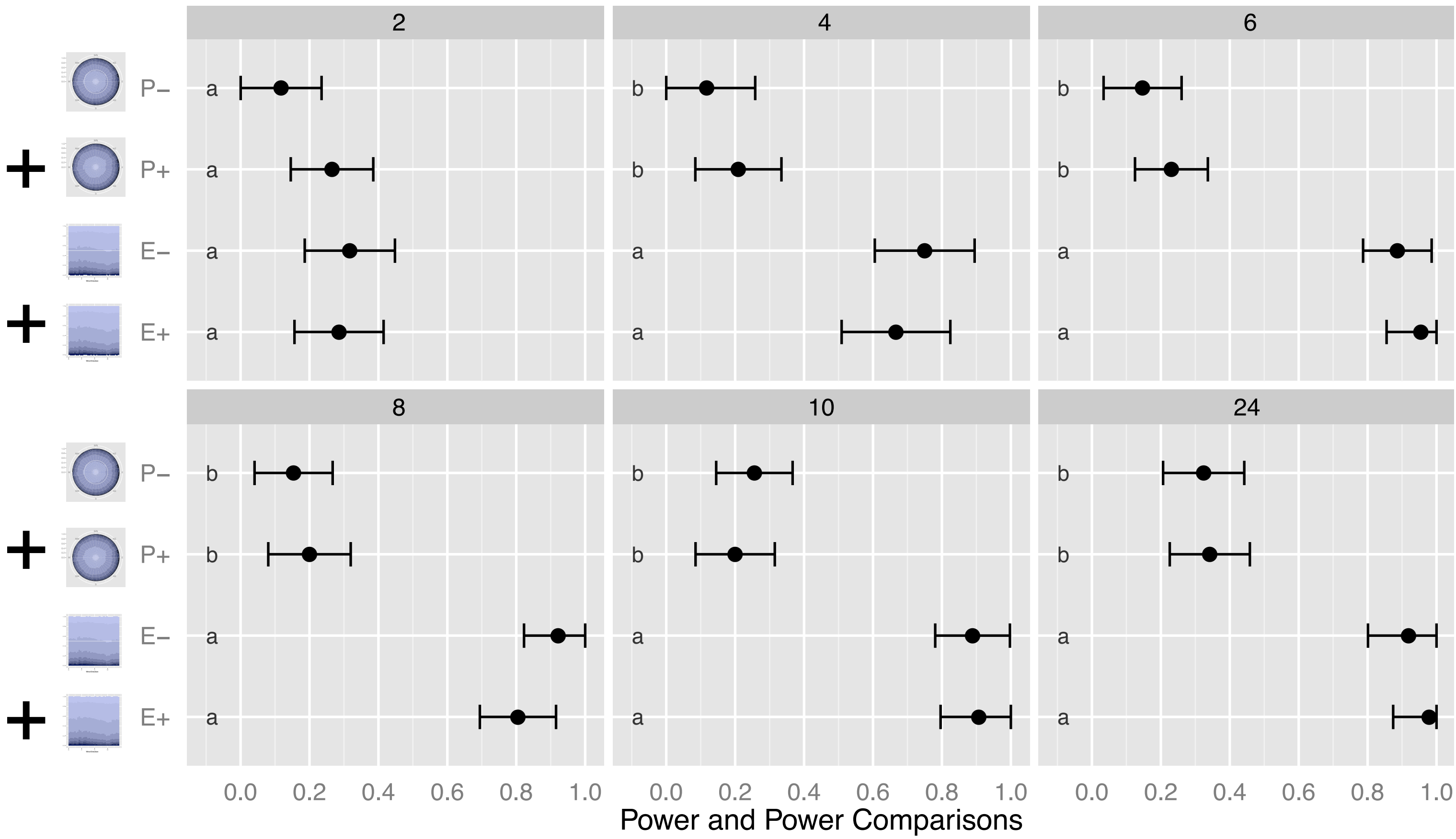
- 958 evaluations by 100 participants
- use one of ten lineups as reference - if people don't get a very easy one correct, we will exclude their data from the study

Evaluation

- 958 evaluations by 100 participants
- use one of ten lineups as reference - if people don't get a very easy one correct, we will exclude their data from the study

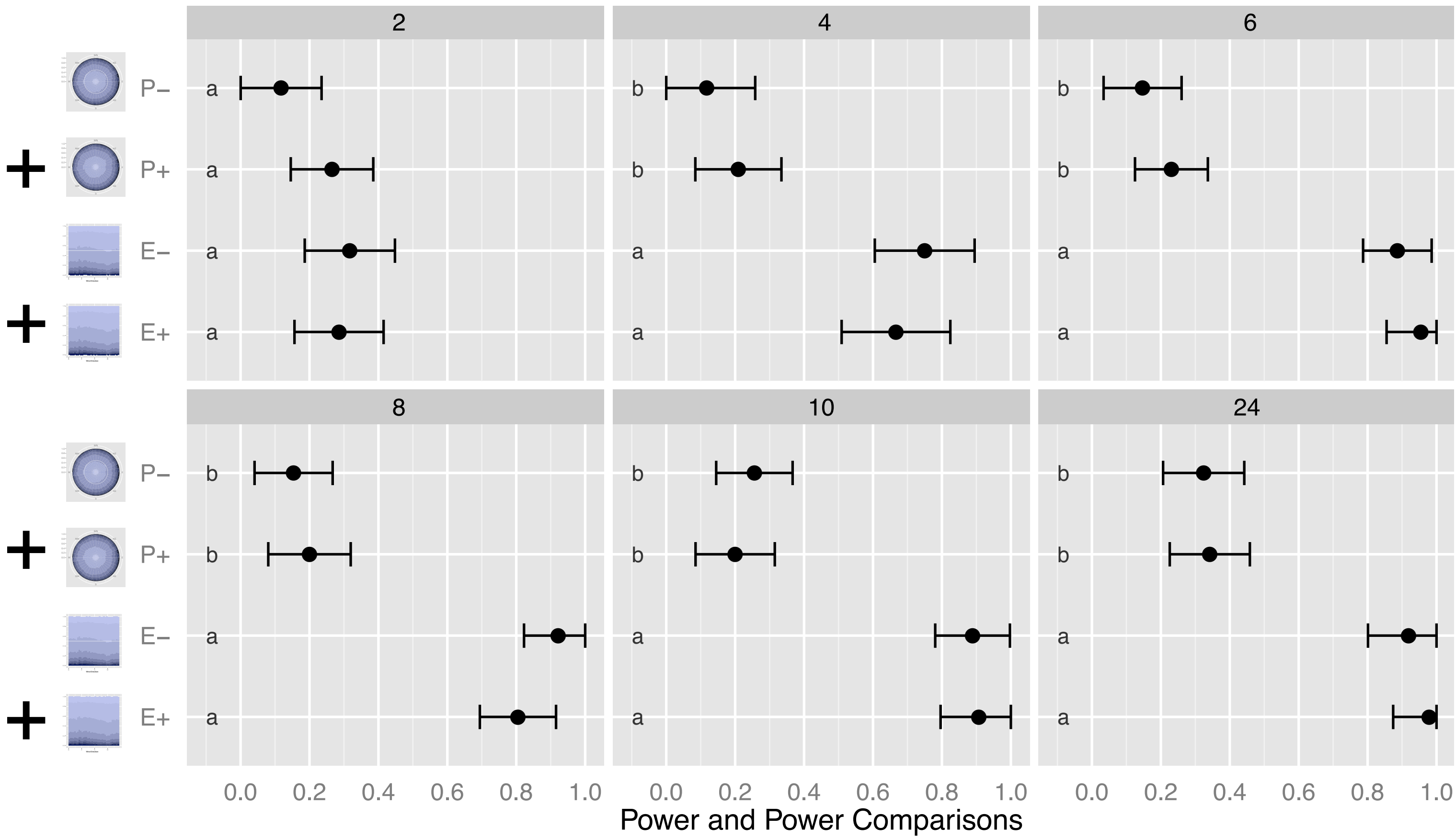


Comparison of Designs



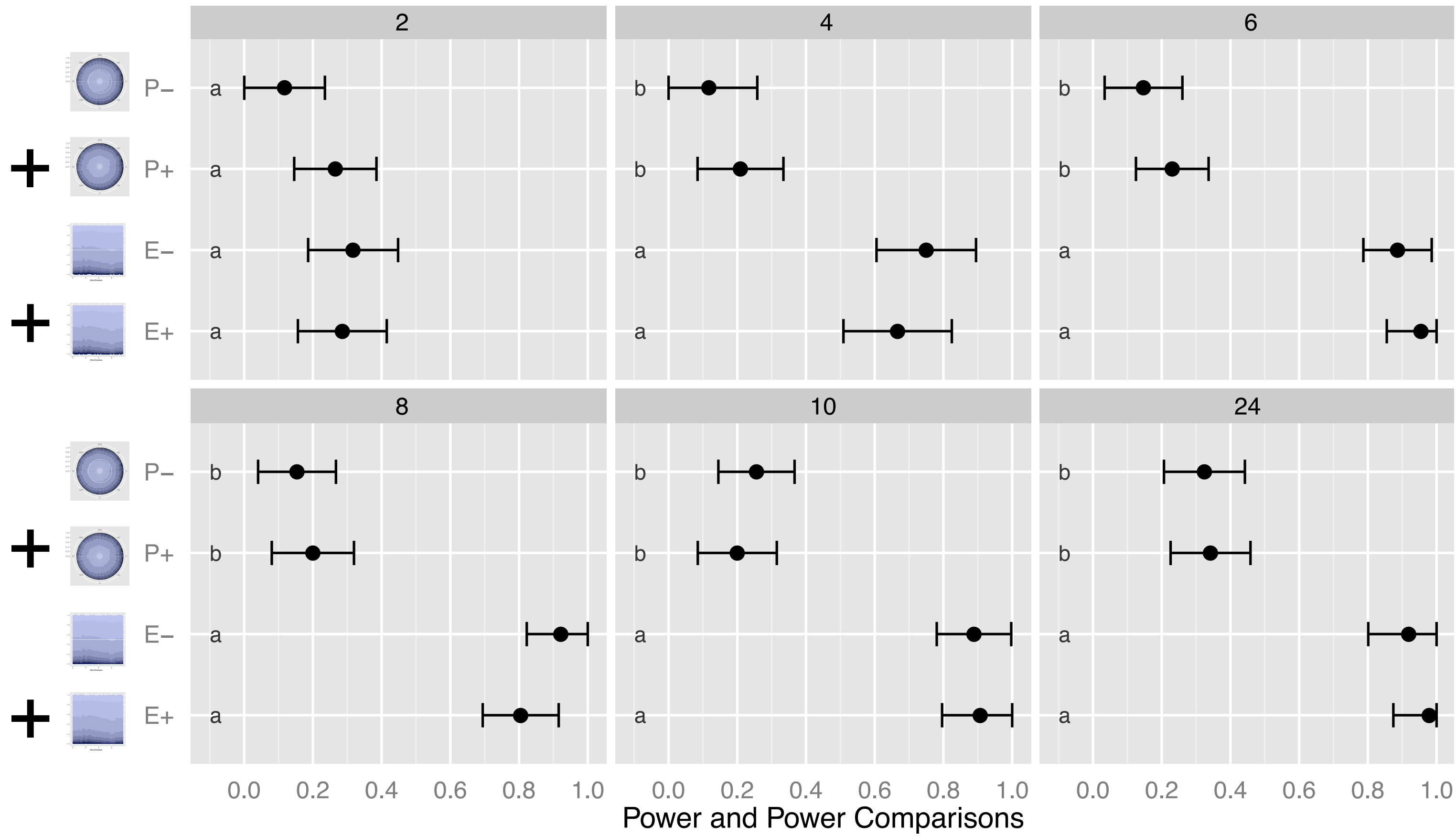
Proportion correct

Comparison of Designs



Polar charts perform significantly worse

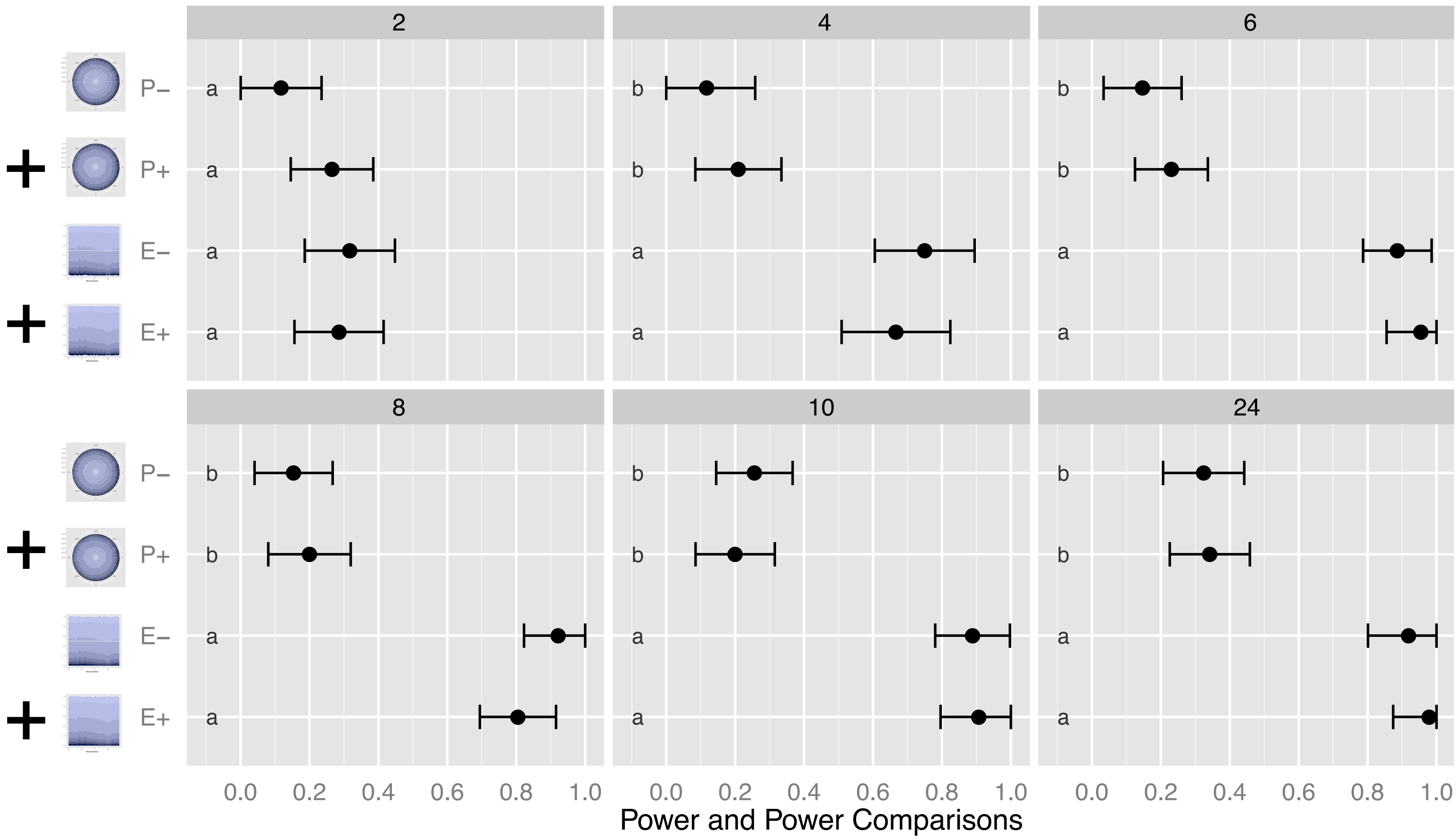
Comparison of Designs



Polar charts perform significantly worse

No significant benefit from helper lines (except in people's confidence)

Comparison of Designs

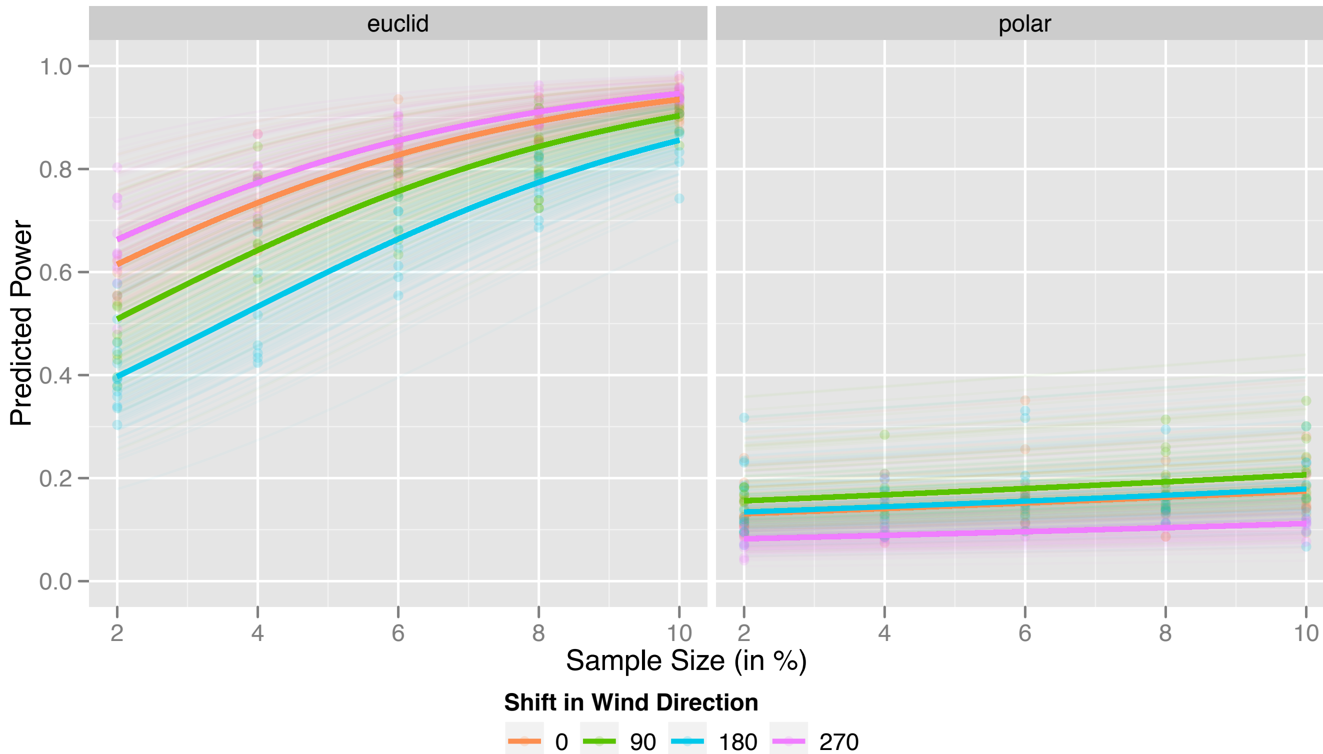


Polar charts perform significantly worse

No significant benefit from helper lines (except in people's confidence)

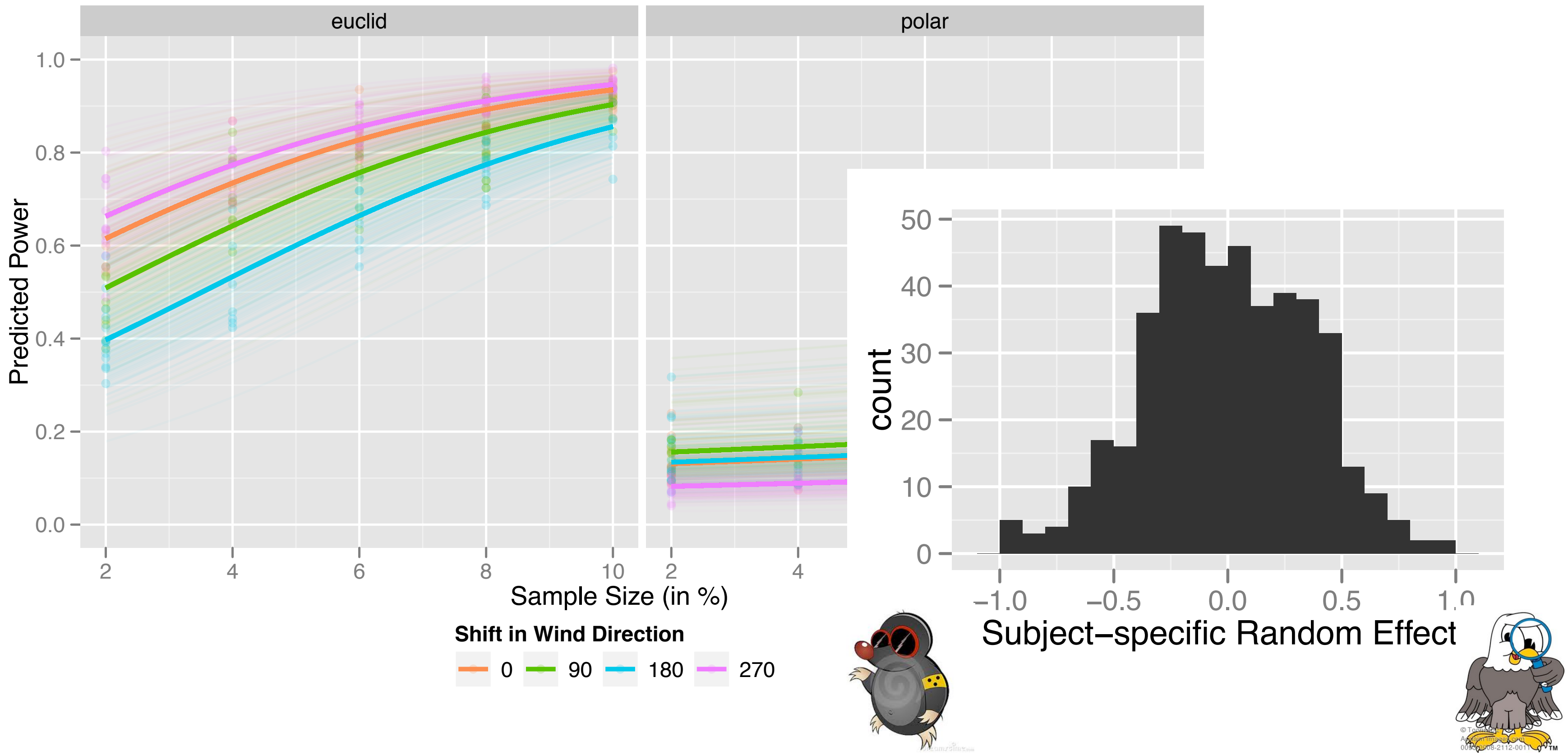
Shift in wind direction does not have an impact on performance ...

Effect of shifts



- average power drawn by thick solid lines
- subject-specific power shown with thin lines
- subject specific effects quite large - how do we get power observers?

Effect of shifts



- average power drawn by thick solid lines
- subject-specific power shown with thin lines
- subject specific effects quite large - how do we get power observers?

Conclusions

for Seattle

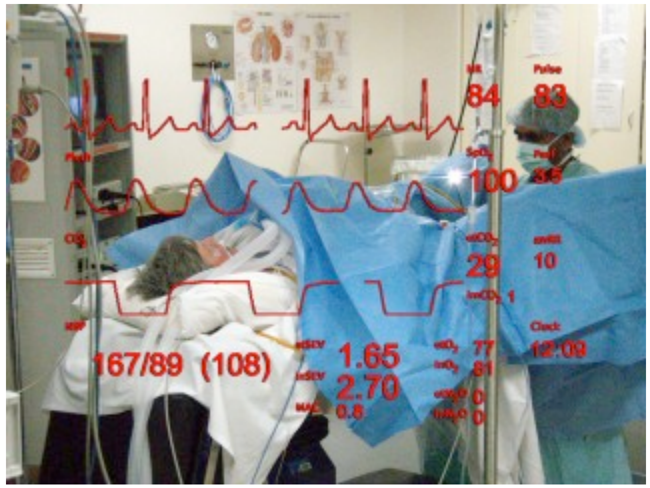
- overwhelming evidence that winds from SE lead to least efficient traffic flow
- BUT: winds from NW lead to most efficient traffic flow
- naive conclusion: use runways in other direction for days with SE winds?

Conclusions

- Use lineup scenario to get valid p-values for visual findings
- useful in situations where conventional methods break down (large or non-traditional data)
- define power (function) for lineups to evaluate
  - competing designs
  - measure impact of other co-variates on display
- Airport study: euclidean charts better at detecting patterns than polar charts

funded by NSF

Headsets for monitoring data



- http://www.newswise.com/articles/anesthesiologists-test-headsets-for-monitoring-data-during-surgery
- Anesthesia & Analgesia (Apr-2010)

graphs need to be highly efficient and preferably small