Data visualization for real-world machine learning

Julia Silge

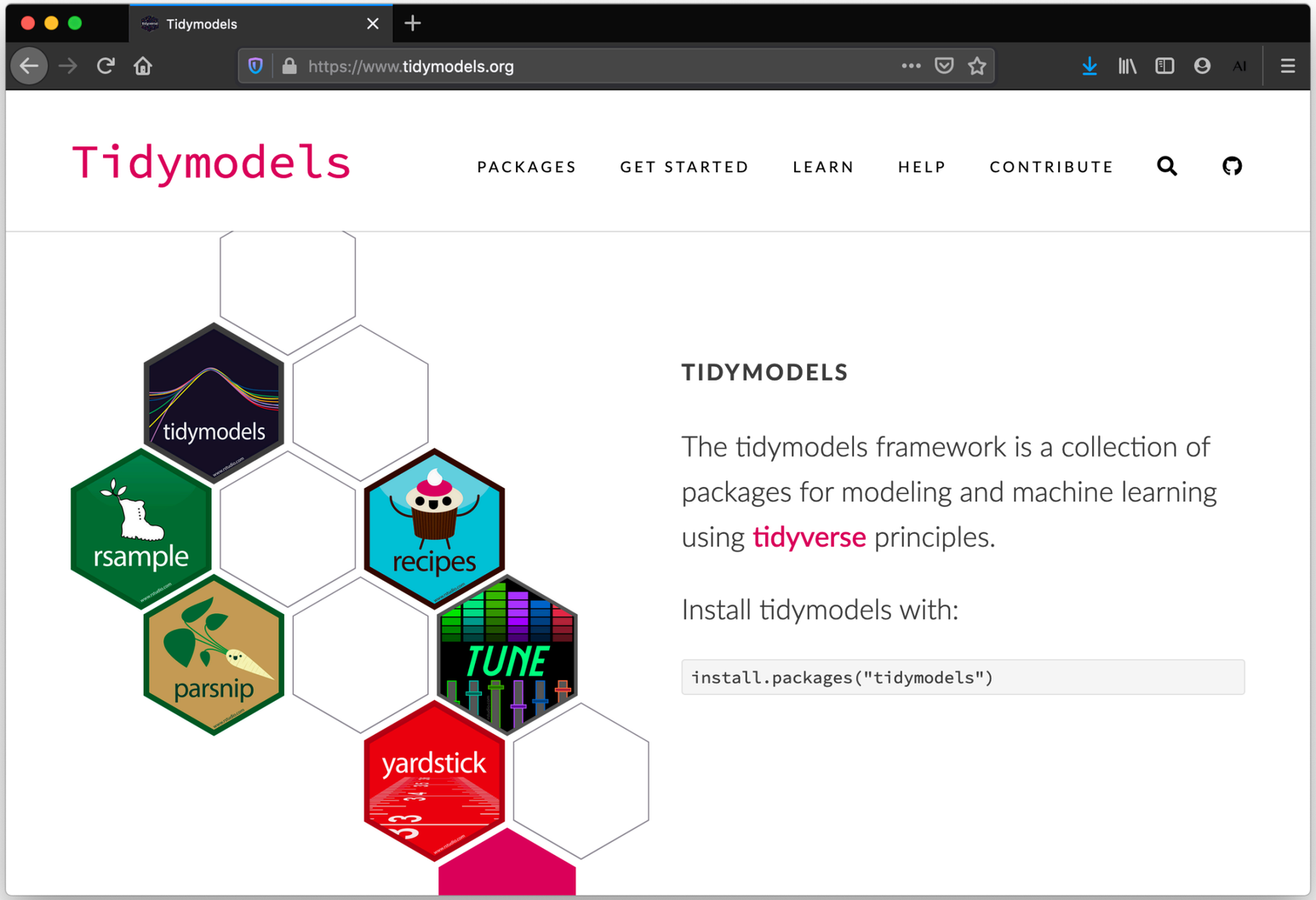
@juliasilge

data visualization informs how we think, understand, decide

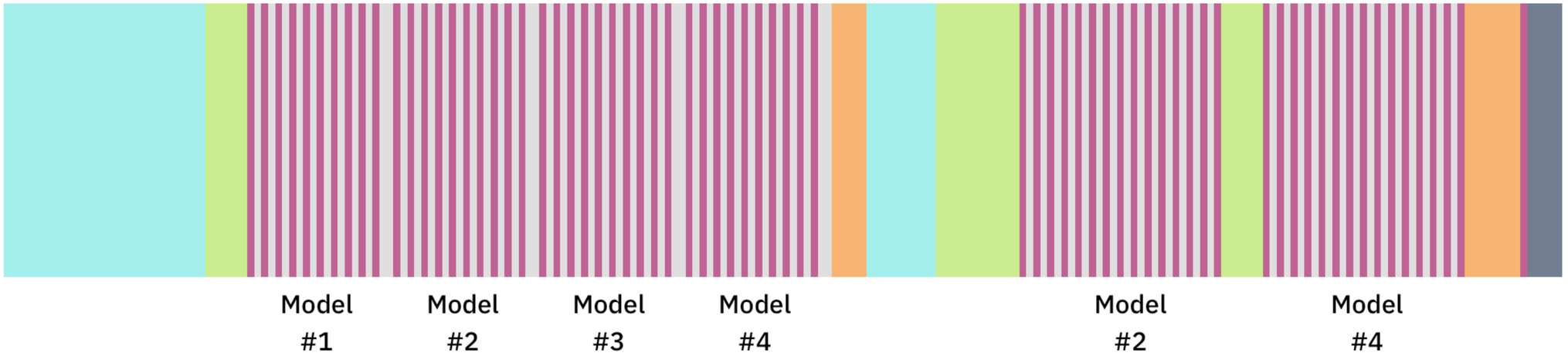
@juliasilge

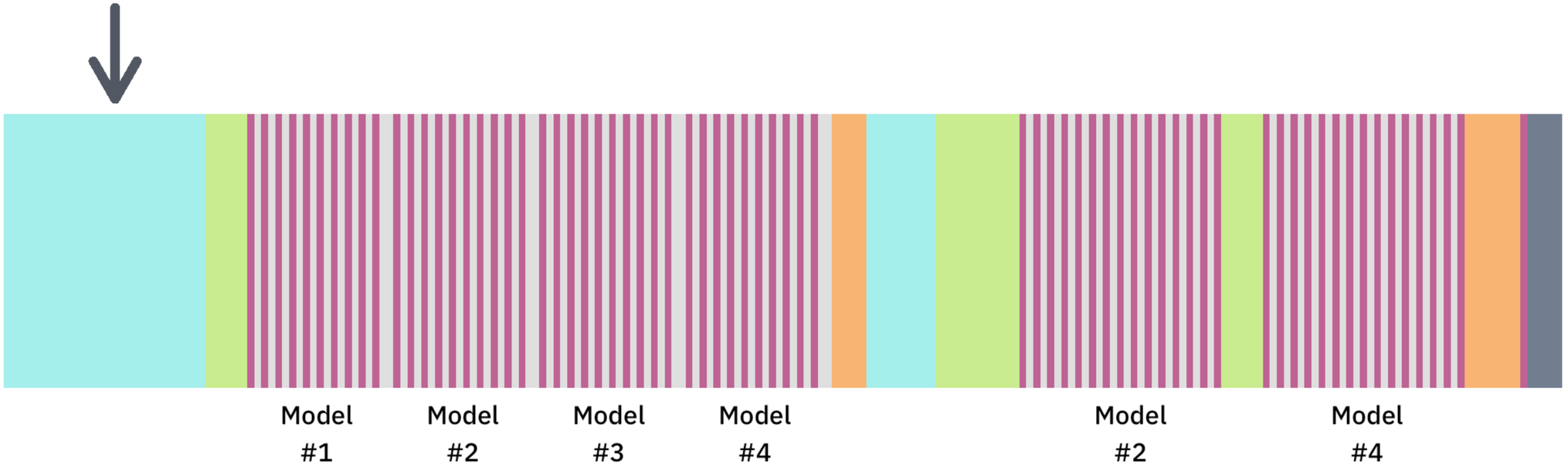
Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

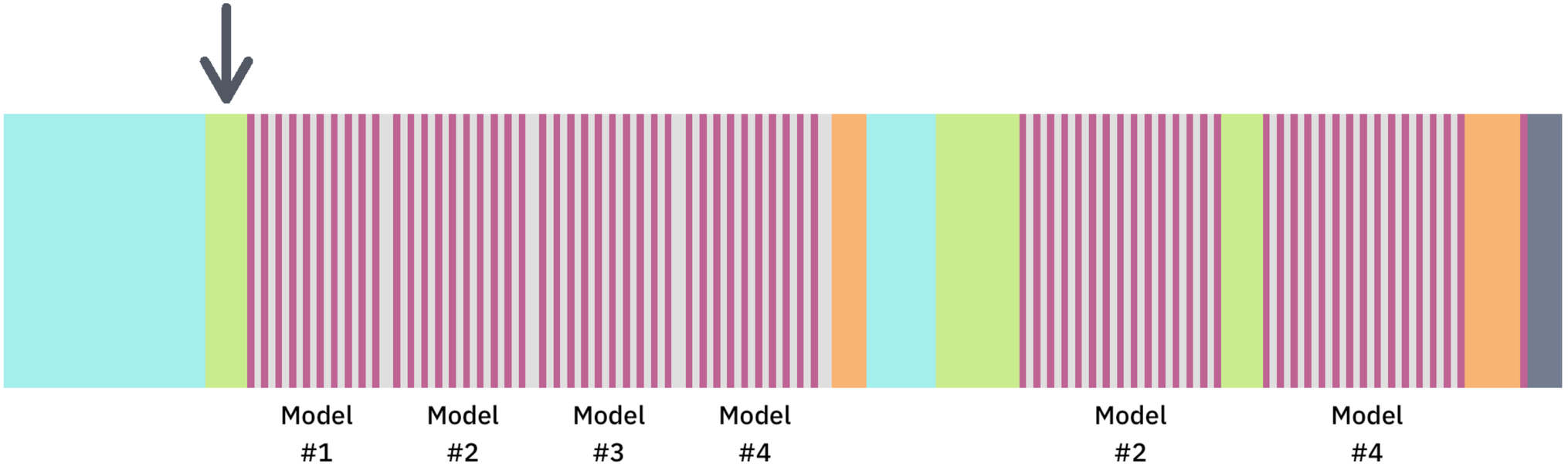
— Tamara Munzner in *Visualization Analysis & Design*

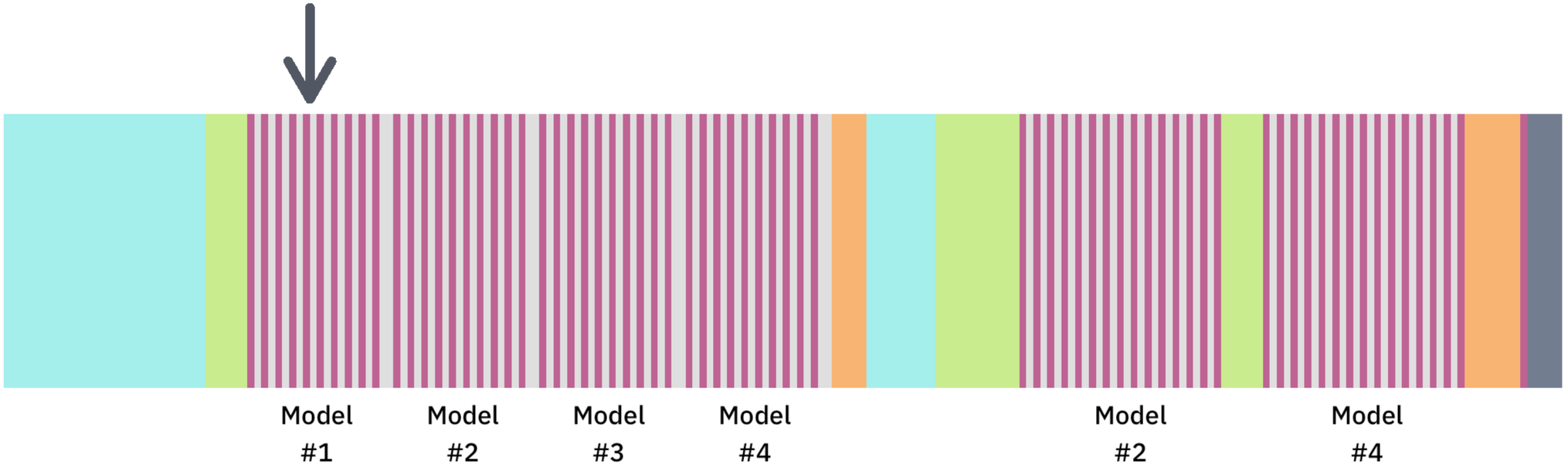

Tidymodels
https://www.tidymodels.org
Tidymodels
PACKAGES
GET STARTED
LEARN
HELP
CONTRIBUTE
tidymodels
rsample
recipes
parsnip
TUNE
yardstick
TIDYMODELS
The tidymodels framework is a collection of packages for modeling and machine learning using tidyverse principles.
Install tidymodels with:
install.packages("tidymodels")

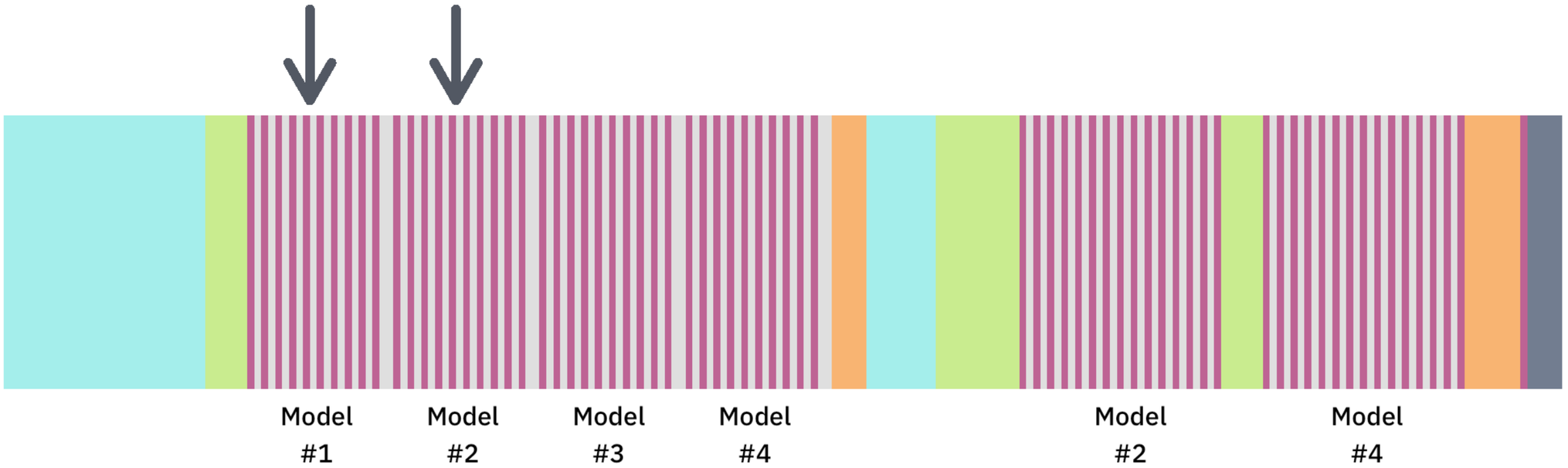
When do practitioners build data visualizations?

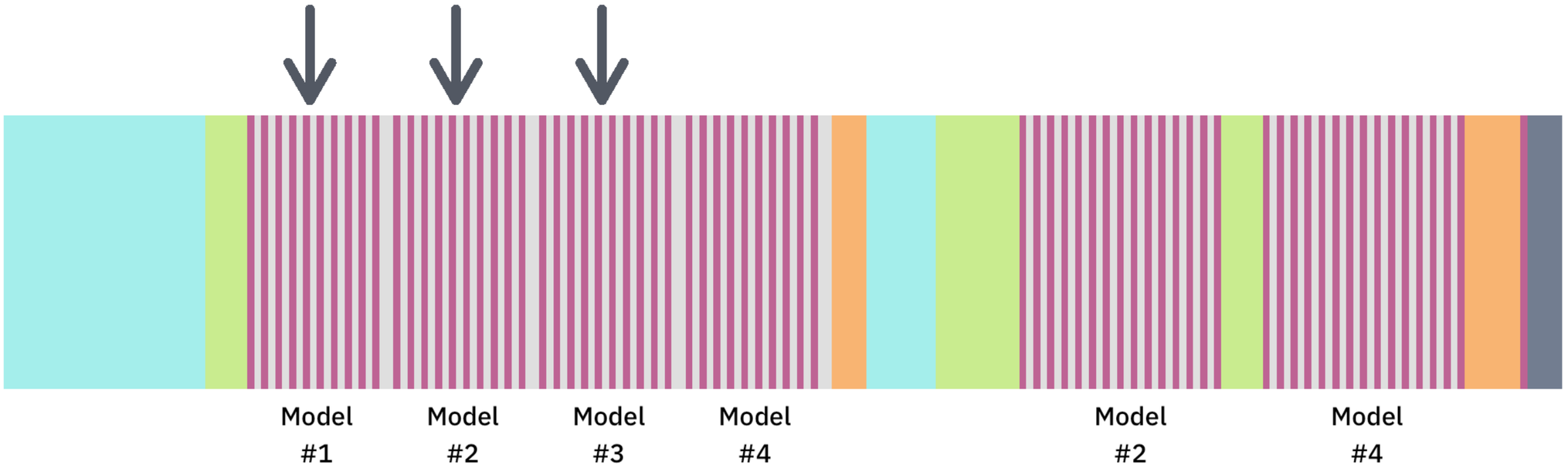

Model #1
Model #2
Model #3
Model #4
Model #2
Model #4
EDA
Feature Engineering
Model Fit
Model Tuning
Model Evaluation
Communication, deployment, etc.

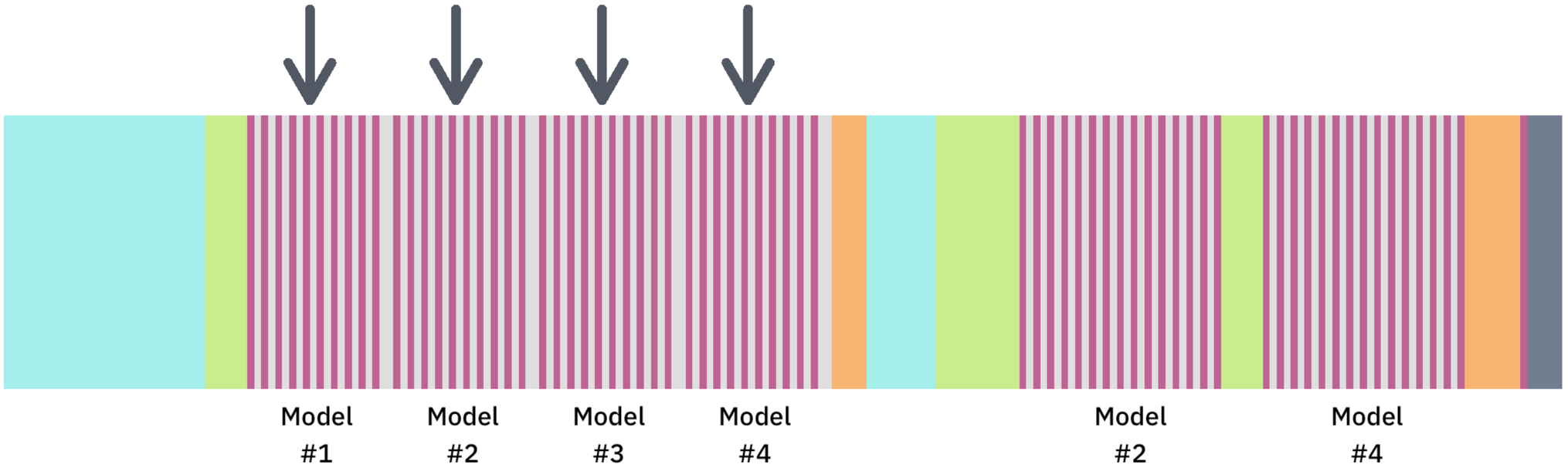

Model #1
Model #2
Model #3
Model #4
Model #2
Model #4
EDA
Feature Engineering
Model Fit
Model Tuning
Model Evaluation
Communication, deployment, etc.

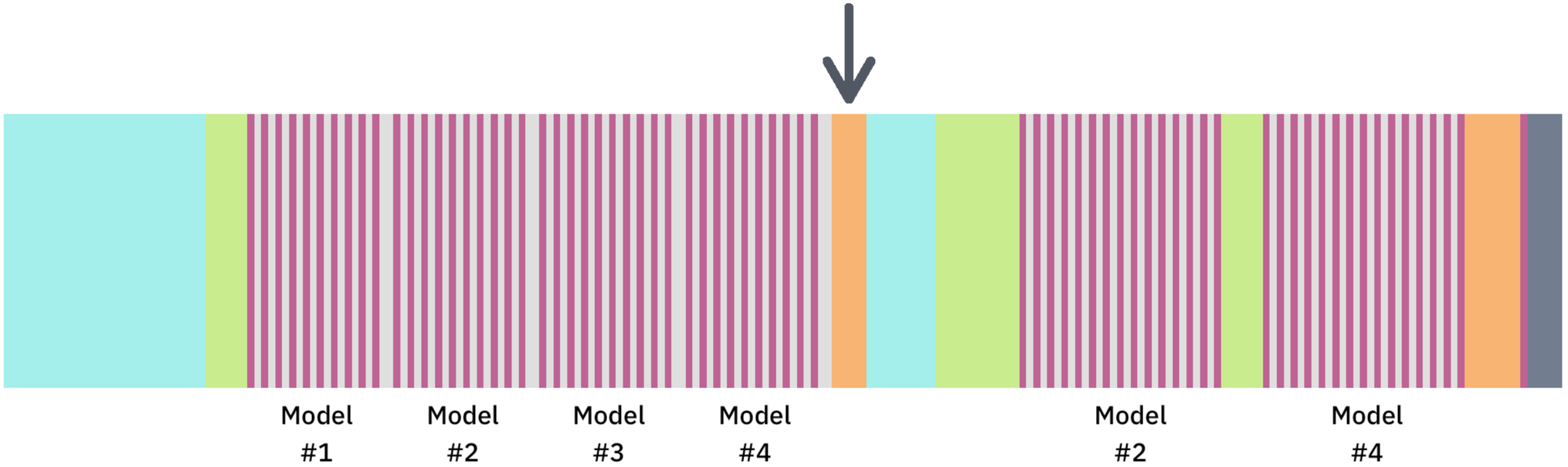

Model #1
Model #2
Model #3
Model #4
Model #2
Model #4
EDA
Feature Engineering
Model Fit
Model Tuning
Model Evaluation
Communication, deployment, etc.

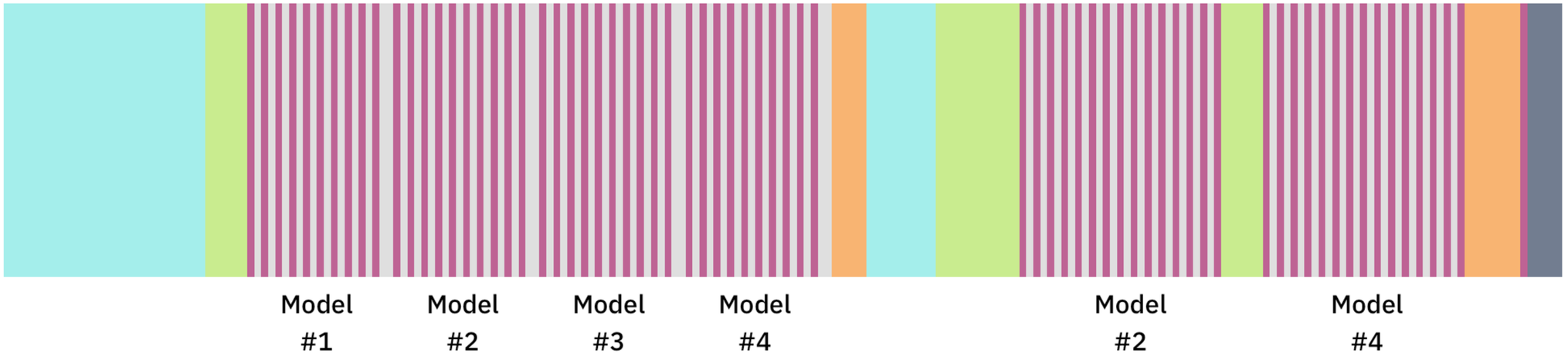

Model #1
Model #2
Model #3
Model #4
Model #2
Model #4
EDA
Feature Engineering
Model Fit
Model Tuning
Model Evaluation
Communication, deployment, etc.


Model #1
Model #2
Model #3
Model #4
Model #2
Model #4
EDA
Feature Engineering
Model Fit
Model Tuning
Model Evaluation
Communication, deployment, etc.


Model #1
Model #2
Model #3
Model #4
Model #2
Model #4
EDA
Feature Engineering
Model Fit
Model Tuning
Model Evaluation
Communication, deployment, etc.


Model #1
Model #2
Model #3
Model #4
Model #2
Model #4
EDA
Feature Engineering
Model Fit
Model Tuning
Model Evaluation
Communication, deployment, etc.

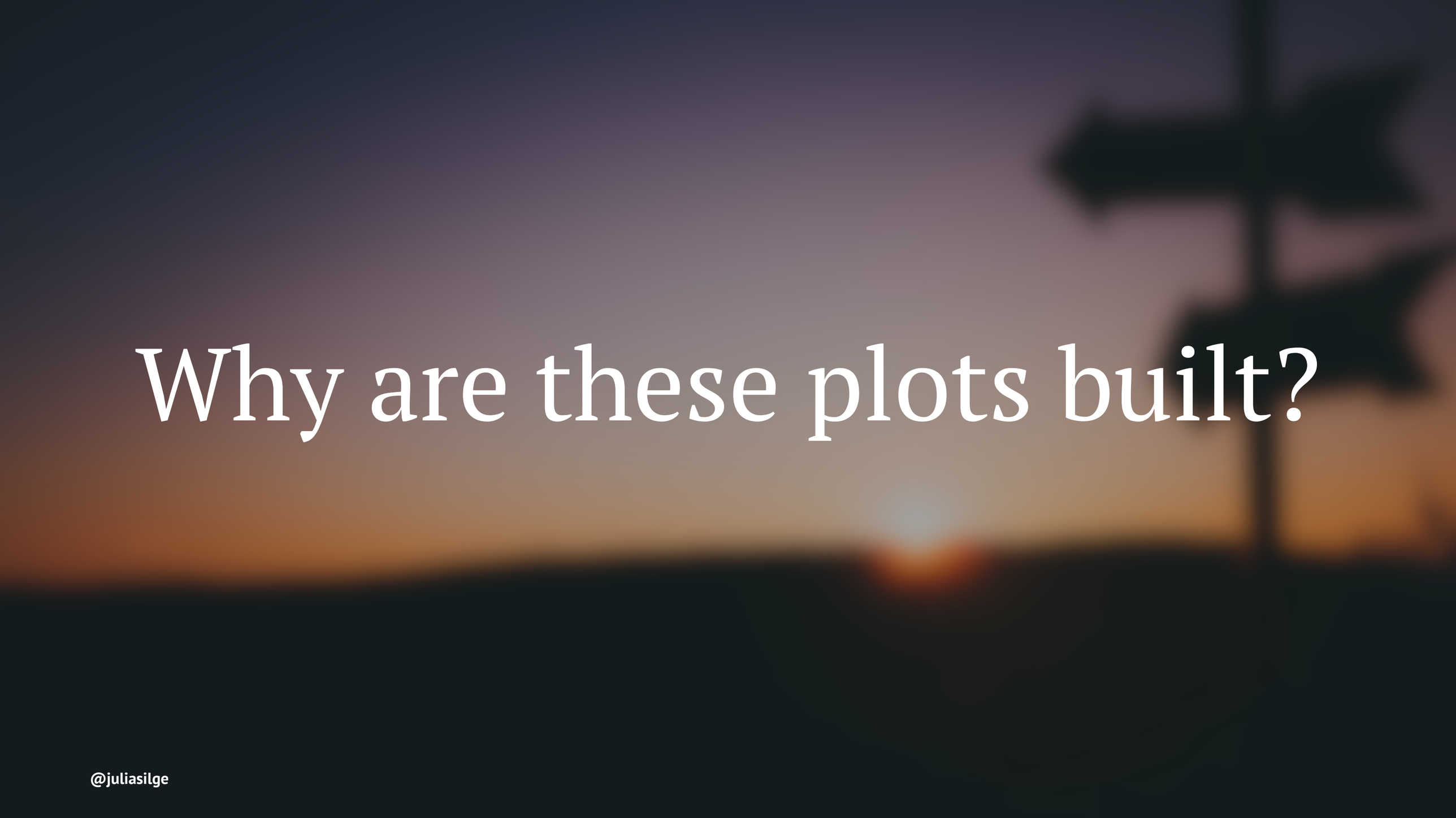

Model #1
Model #2
Model #3
Model #4
Model #2
Model #4
EDA
Feature Engineering
Model Fit
Model Tuning
Model Evaluation
Communication, deployment, etc.

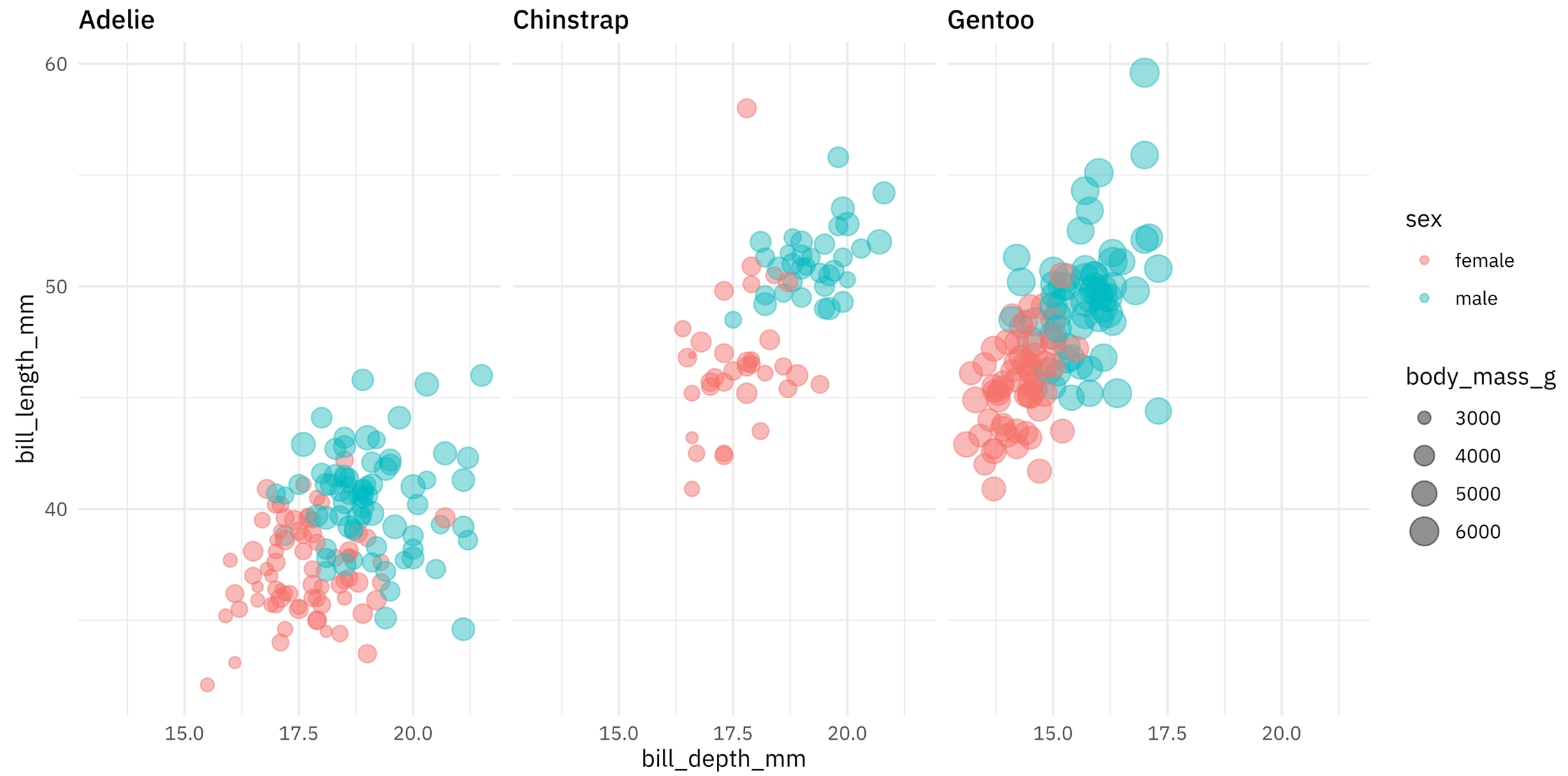

Model #1
Model #2
Model #3
Model #4
Model #2
Model #4
EDA
Feature Engineering
Model Fit
Model Tuning
Model Evaluation
Communication, deployment, etc.

exploratory data analysis

model evaluation

@juliasilge


Why are these plots built?
@juliasilge



Palmer penguins

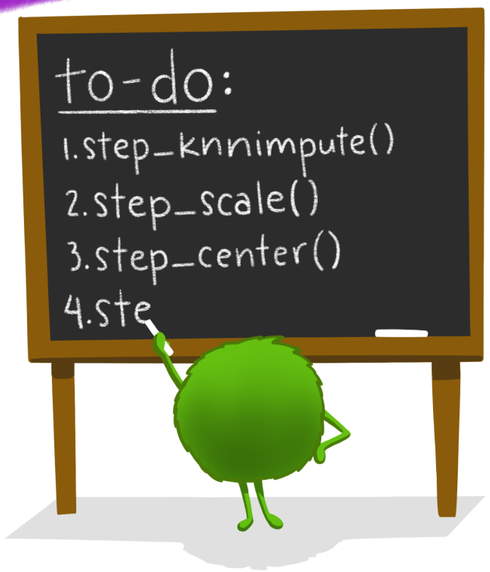
Trees in San Francisco

prioritize efficient iteration for EDA

recipes:

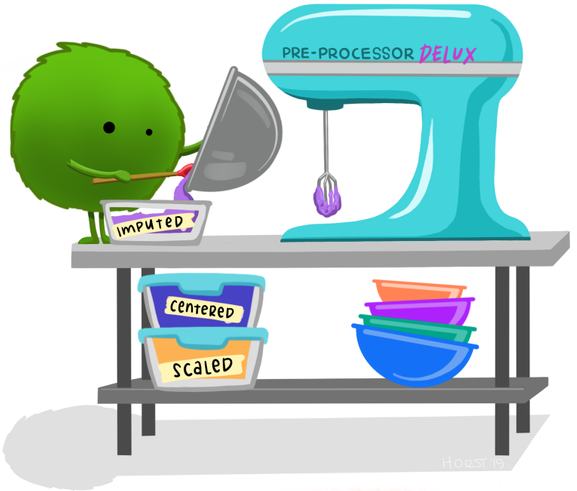
STREAMLINED DATA **PRE-PROCESSING** FOR STATISTICAL + MACHINE LEARNING MODELS

1. SPECIFY VARIABLES
recipe(y ~ a + b + ..., data = pantry)



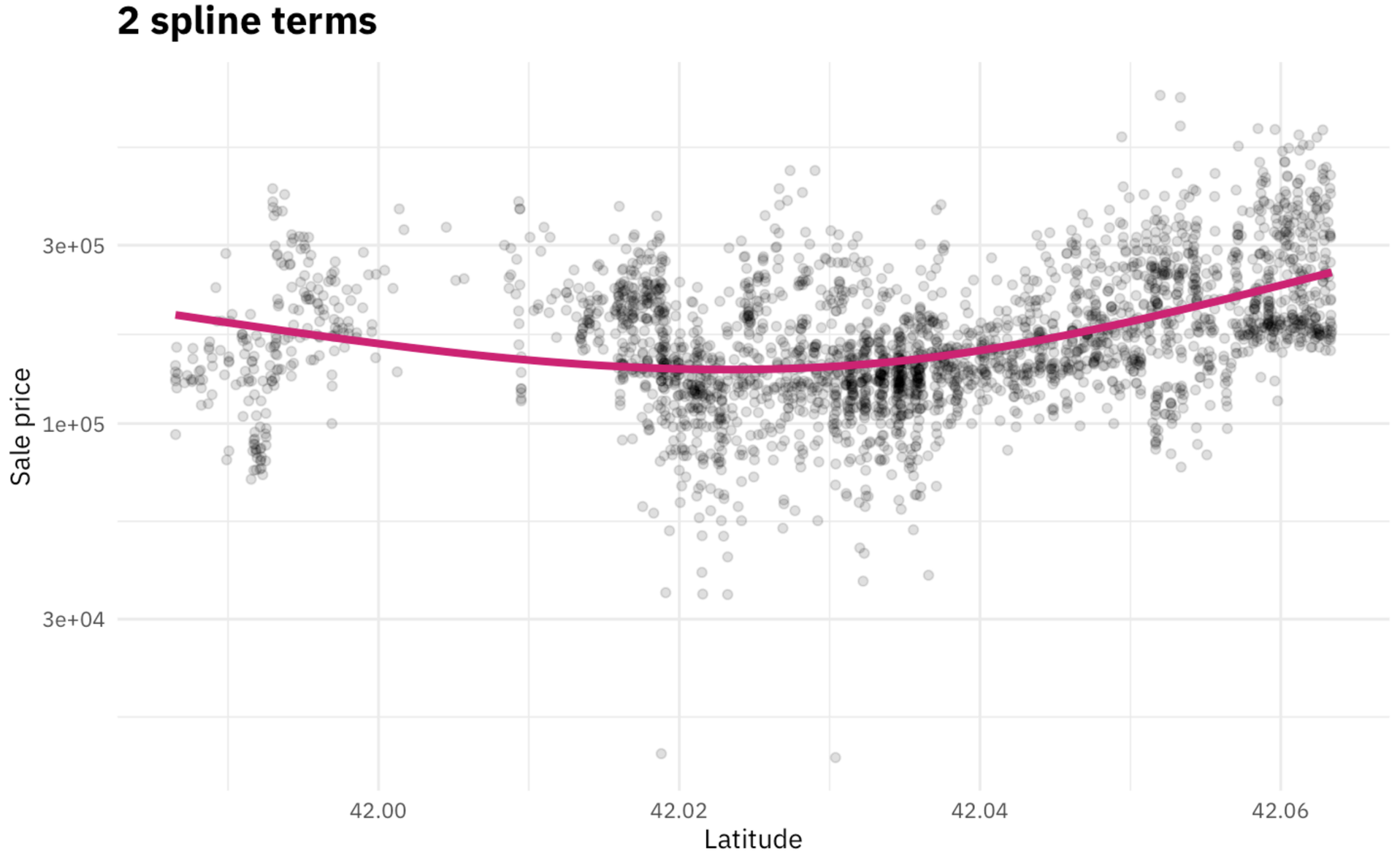
2. DEFINE PRE-PROCESSING STEPS (step_*)

3. PROVIDE DATASET(S) FOR RECIPE STEPS
prep()

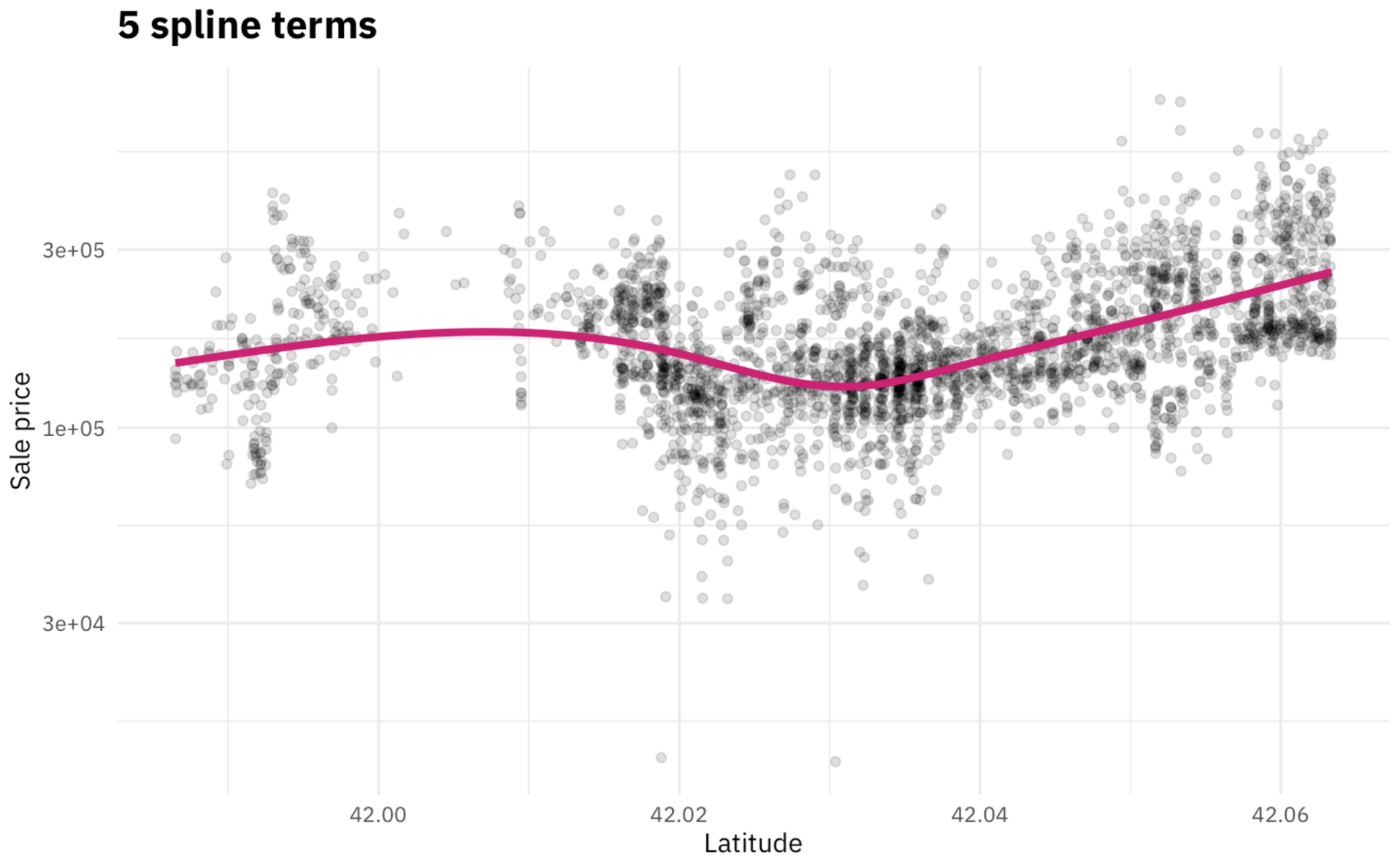


4. APPLY PRE-PROCESSING!
bake()

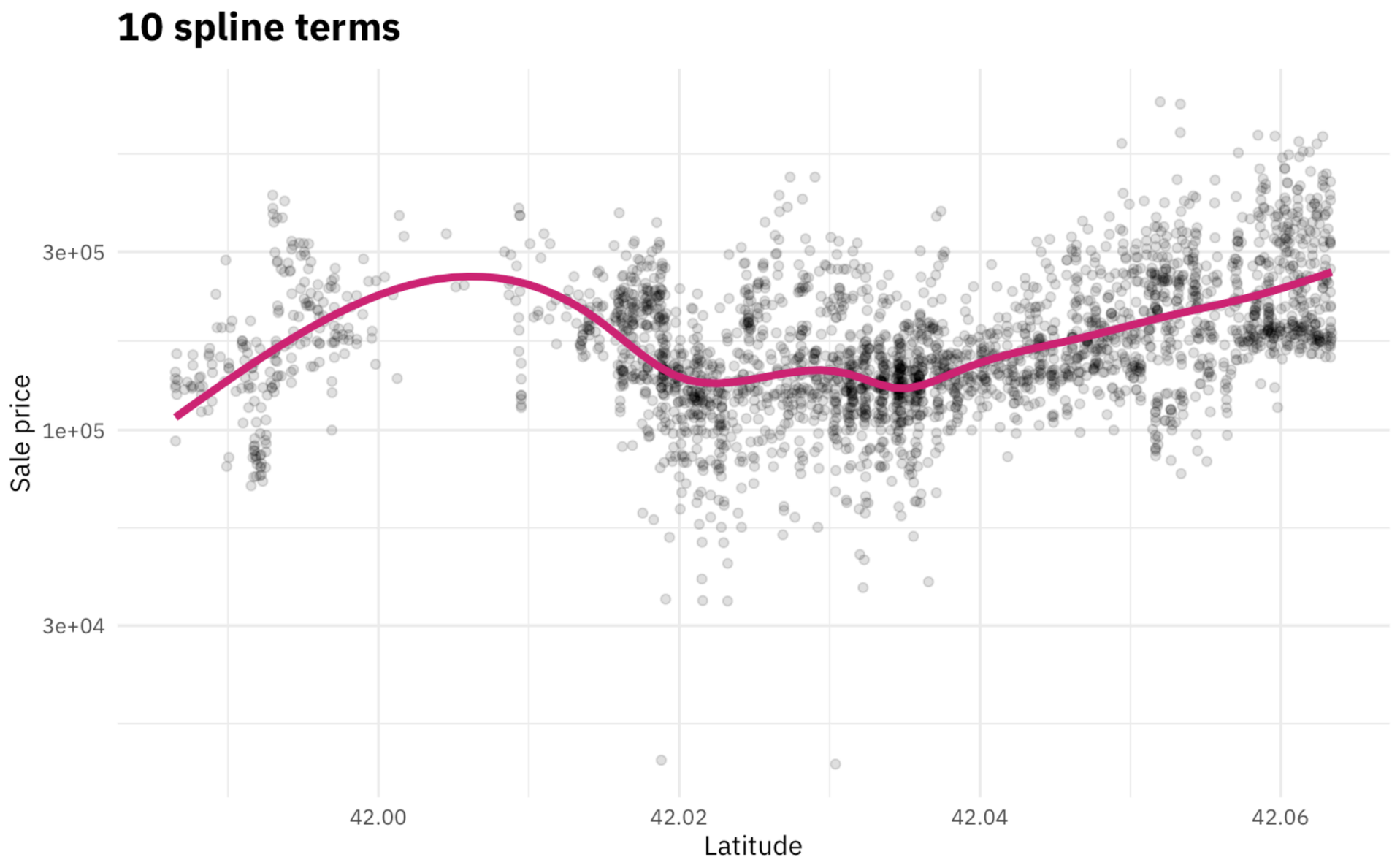
Illustration by Allison Horst



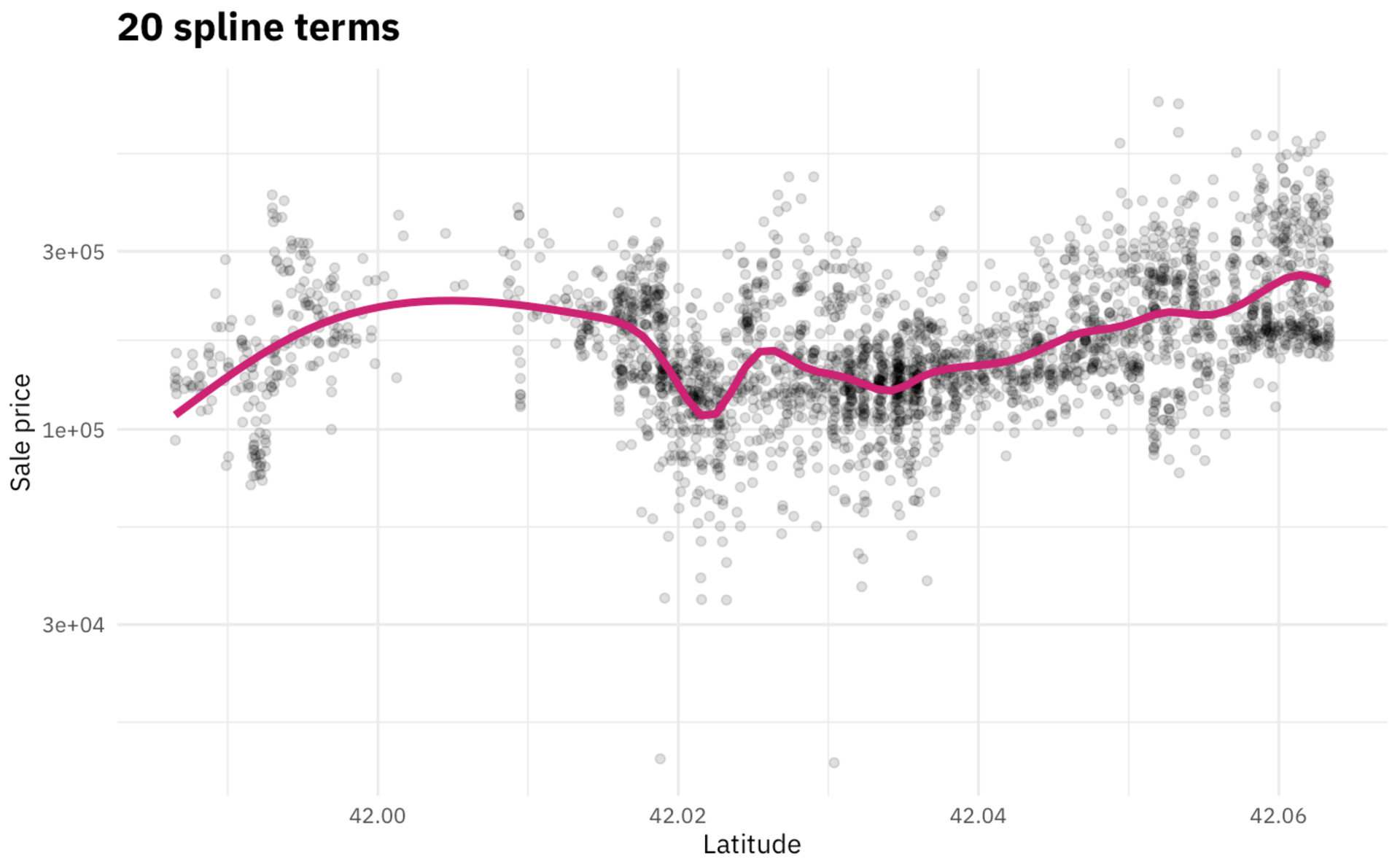
Houses in Ames, IA



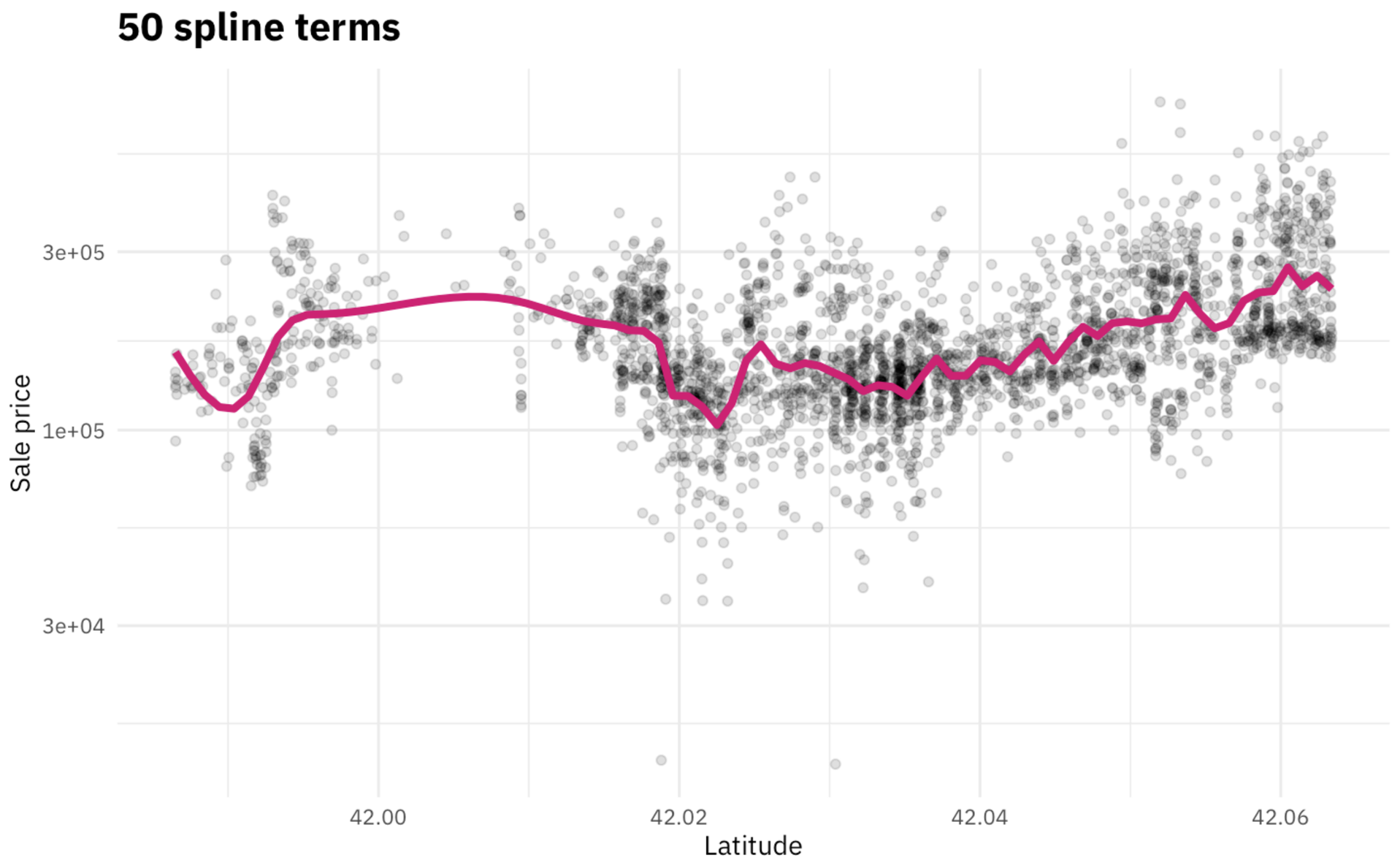
Houses in Ames, IA



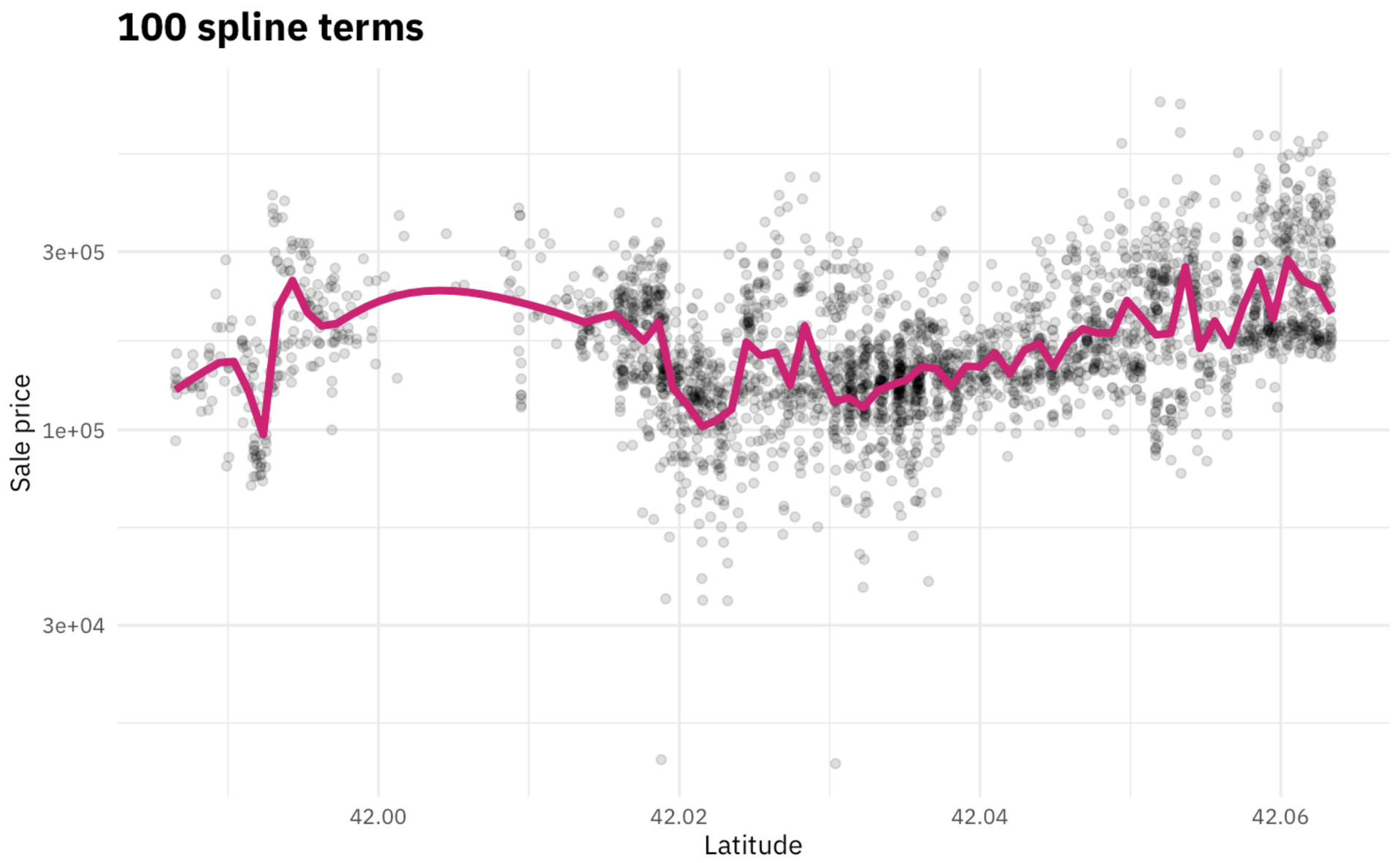
Houses in Ames, IA



Houses in Ames, IA



Houses in Ames, IA



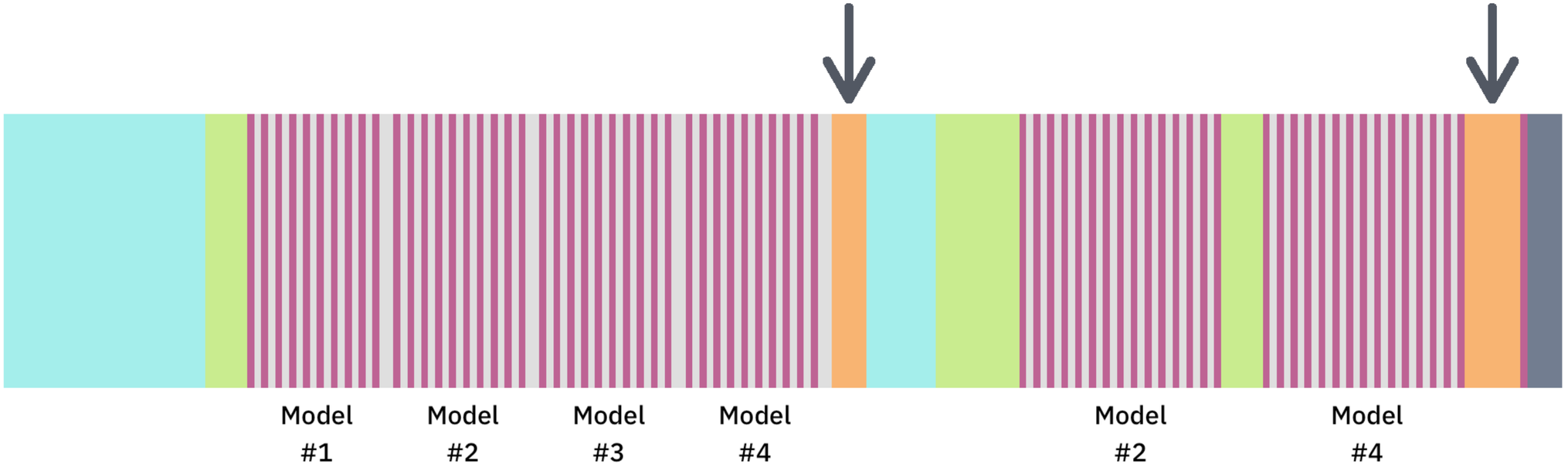
Houses in Ames, IA

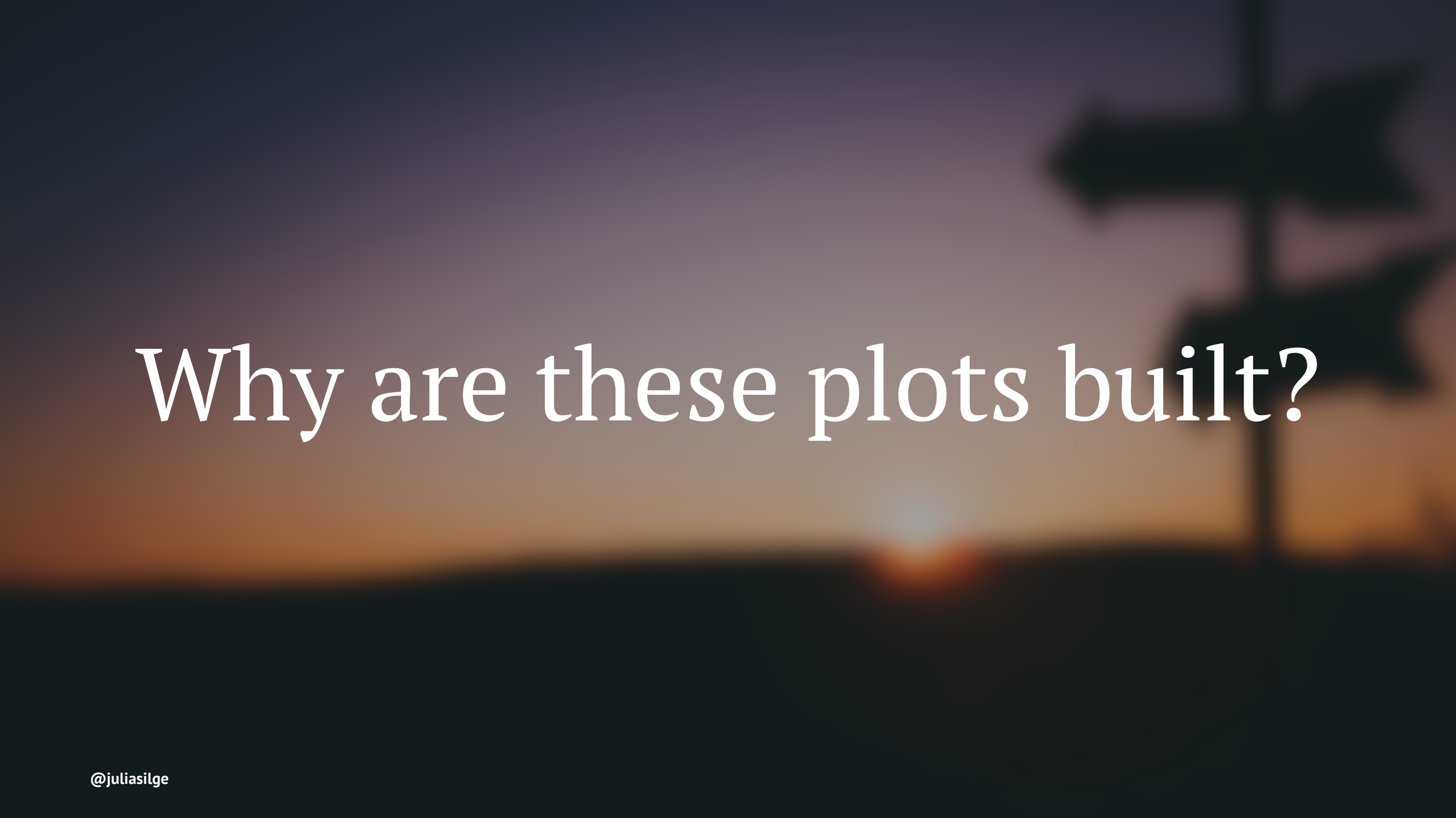
```
recipe(Sale_Price ~ Neighborhood + Gr_Liv_Area + Year_Built +
         Bldg_Type + Latitude + Longitude,
       data = ames_train) %>%
  step_log(Gr_Liv_Area, base = 10) %>%
  step_other(Neighborhood, threshold = 0.01) %>%
  step_dummy(all_nominal()) %>%
  step_ns(Latitude, Longitude, deg_free = 10)
```

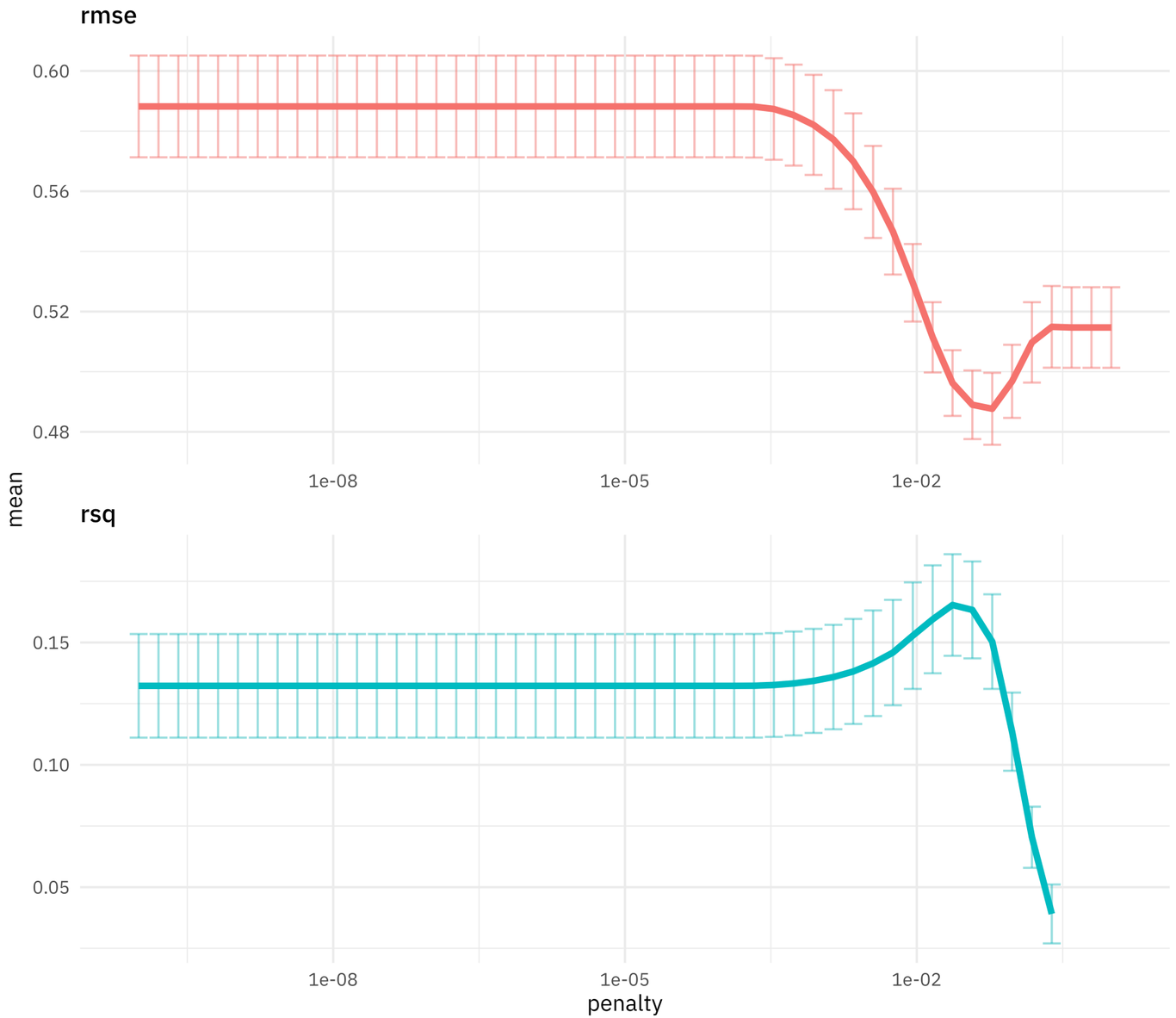
```
## Data Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          6
##
## Operations:
##
## Log transformation on Gr_Liv_Area
## Collapsing factor levels for Neighborhood
## Dummy variables from all_nominal()
## Natural Splines on Latitude, Longitude
```

Practitioners use data visualization during EDA to inform modeling choices

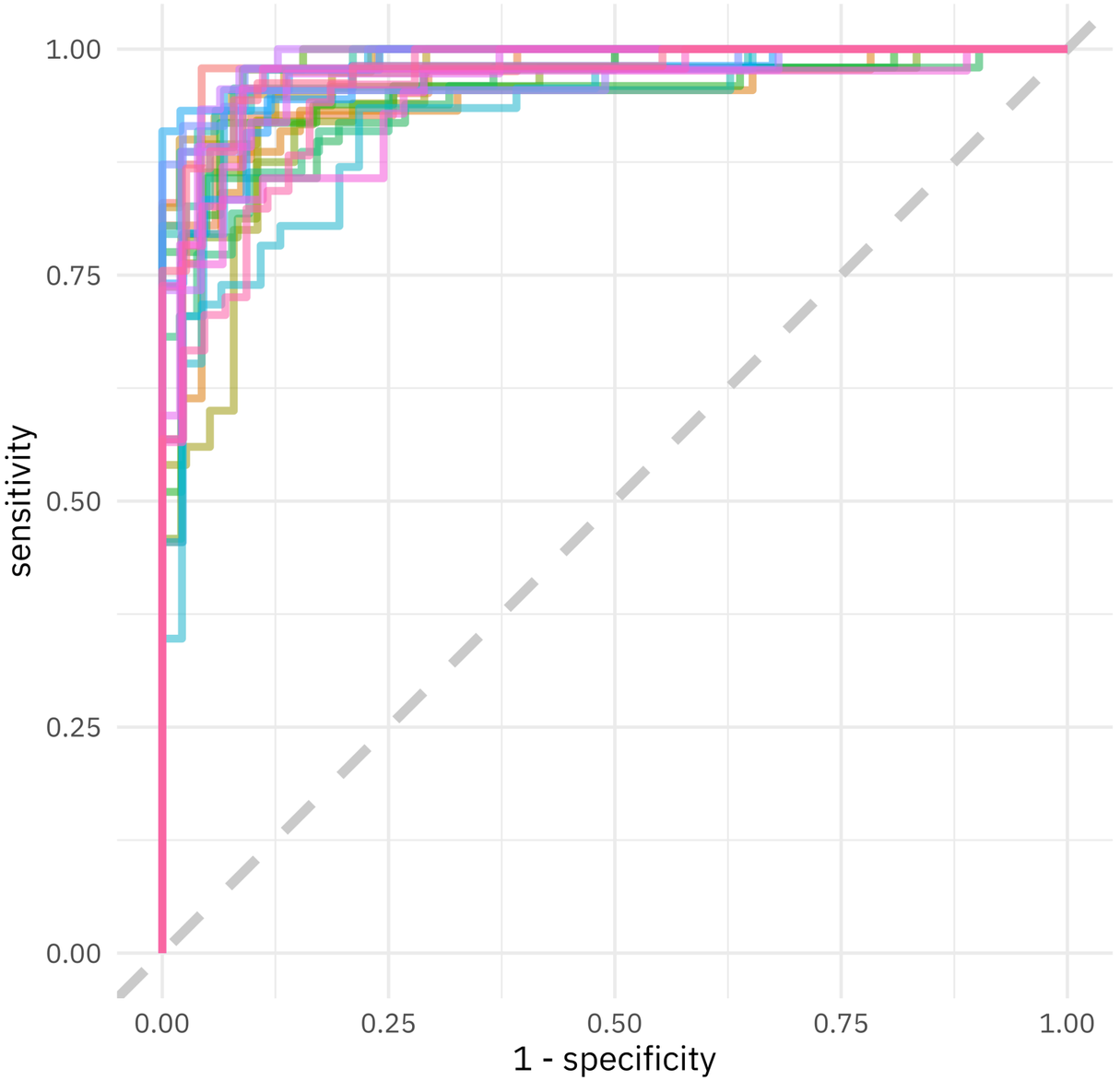

Pinboard
@Pinboard
Machine learning is like a deep-fat fryer. First time you try it you think "Amazing, I bet this will work on anything!" And it kind of does
10:16 PM · Jul 6, 2016 · YoruFukurou
321 Retweets and comments
501 Likes


Model #1
Model #2
Model #3
Model #4
Model #2
Model #4
EDA
Feature Engineering
Model Fit
Model Tuning
Model Evaluation
Communication, deployment, etc.


Why are these plots built?
@juliasilge



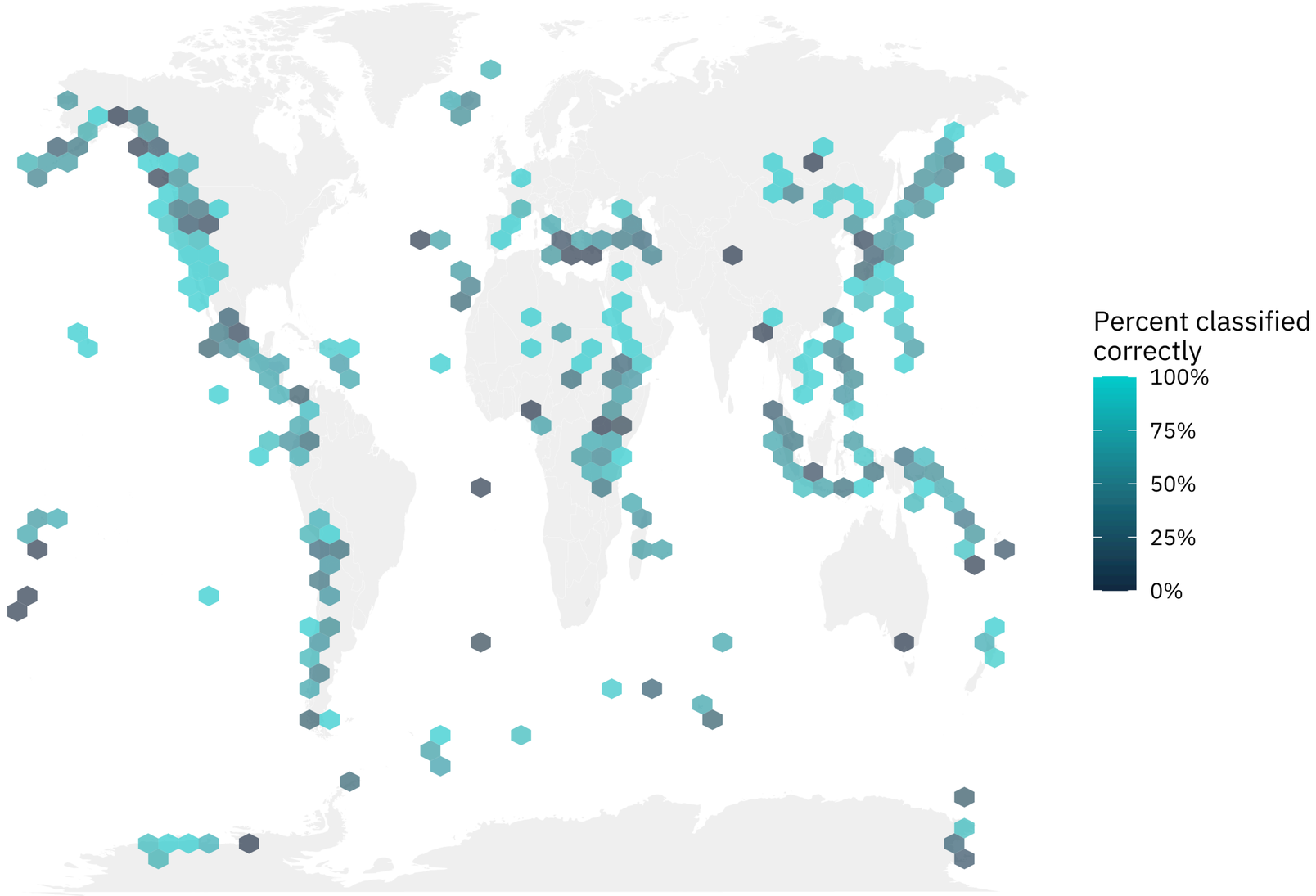
Episodes of *The Office*



Palmer penguins

automatic plotting methods

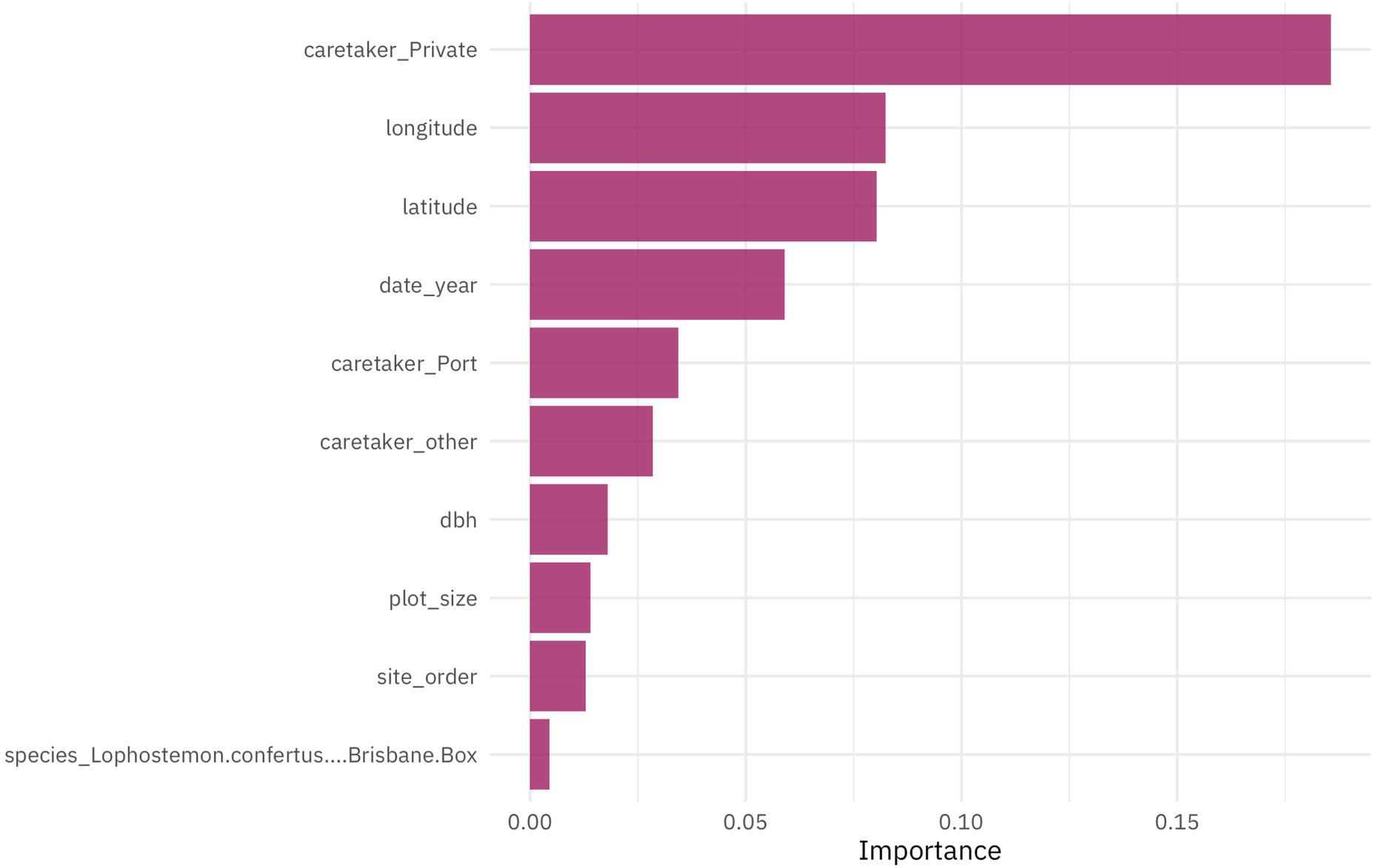
@juliasilge



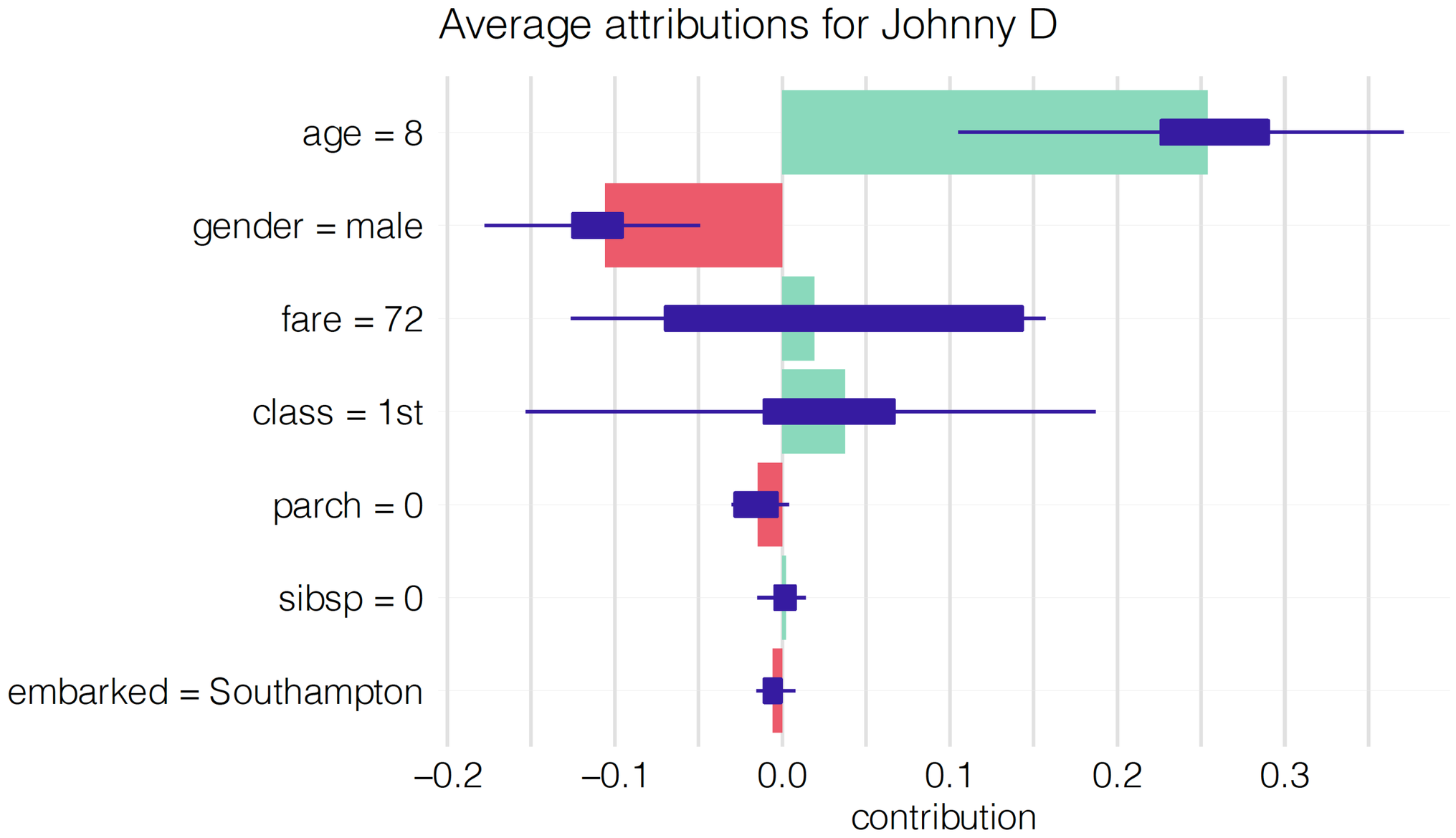
Volcano classification

explaining models

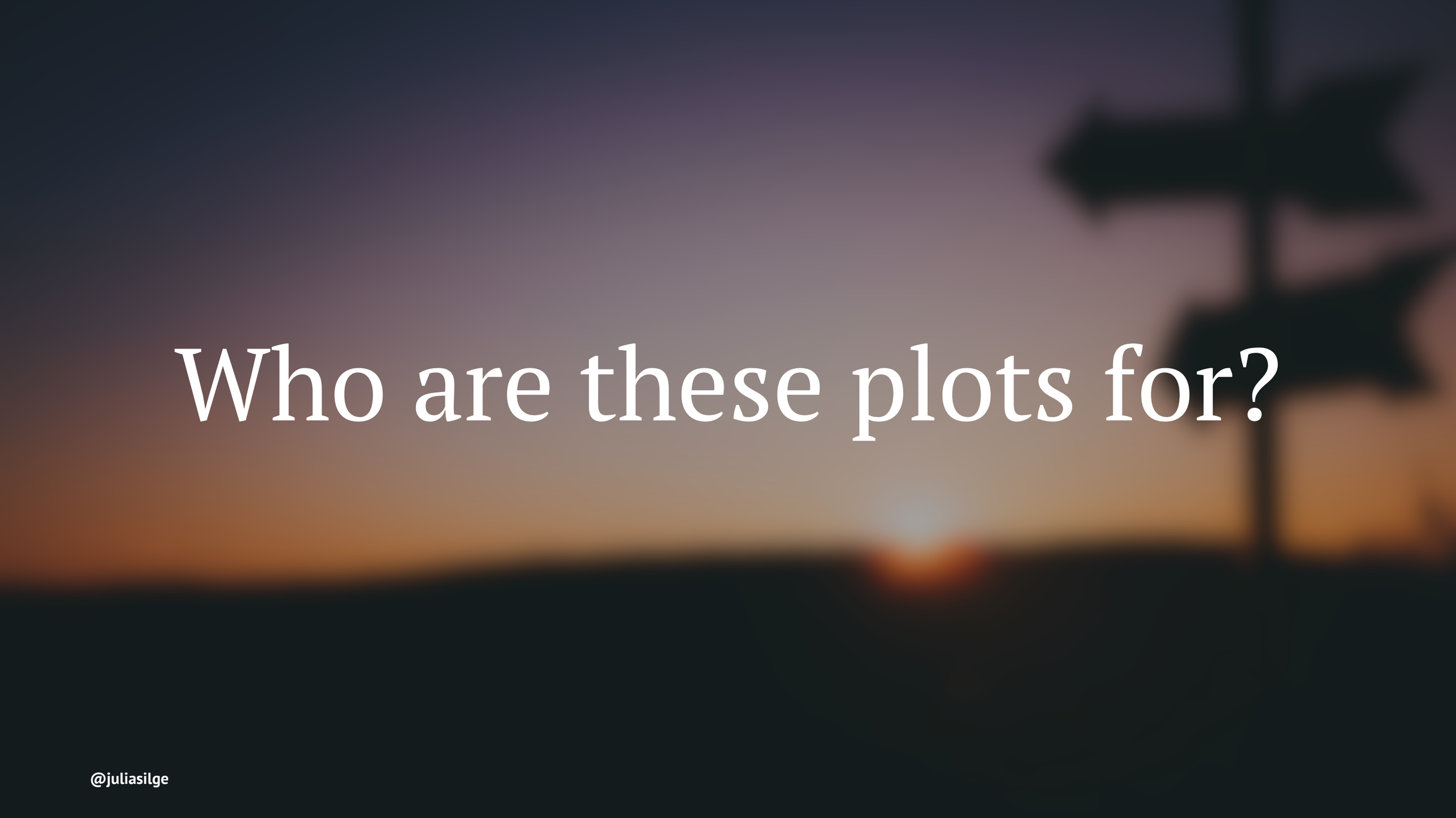
@juliasilge



Trees in San Francisco



from *Explanatory Model Analysis* by Przemyslaw Biecek and Tomasz Burzykowski


Who are these plots for?
@juliasilge

understand

data

@juliasilge

understand models

Thank you!

Julia Silge

Photo by Raul Petri on Unsplash