K-Nearest Neighbor Predictive Models

Karsten Maurer

November 11th, 2019

Who am I?

Karsten Maurer

- Originally from Champlin, MN
- Went to Minnesota, Morris for BA in history and statistics
- Went to Iowa State for PhD in statistics
- Currently an Assistant Professor of Statistics at Miami University
    - Teaching: Data Visualization, Predictive Modeling, Statistical Programming
    - Research: Statistics Education, Machine Learning, Visualization, Collaborative Consulting

Predictive Modeling Background - Discussion

What types of predictive models have you encountered in your classes so far?

What are the two general types of responses that we might try to predict? (Supervised Learning)

A **K-Nearest Neighbor model** generates predictions based on the responses of the **K most similar** previously observed instances.

... before getting into any notation, let's build the intuition with an exercise.

Predicting Sleep

You have two tasks:

1. Think back to last night. How many hours of sleep did you get? **This is your response value**

2. Ask the 3 people sitting closest to you how much sleep they got. **These are your neighbor's responses**

Making a prediction about your sleep

Suppose that where you sit in the room (predictor) holds information about how much you sleep (response)

How could we use your neighbor's sleep values to predict for you?

KNN Regression

If $Y_i \in \mathbb{R}$ is your numeric response for instances $i = 1, 2, \ldots, n$

and $\mathbf{X_i} \in \mathbb{R}^{\mathbf{p}}$ is your p-dimensional numeric predictors

Suppose we have new instance with predictors $\mathbf{X_0}$

Define the set of K-Nearest Neighbors based on some distance function $d(\,.\,)$ (Typically Euclidean)

$\mathcal{N}_0 = \{i | d(\mathbf{X_i}, \mathbf{X_0}) \leq d(\mathbf{X_{[k]}}, \mathbf{X_0})\}$

Prediction is average of K-neighbors: $\hat{Y}_0 = \frac{1}{K} \sum_{i \in \mathcal{N}_0} Y_i$

KNN Sleep Regression

Y = sleep hours, X_1 =Position Left/Right, X_2 =Position Front/Back

$\mathbf{X_0}$ is your seat location

$\mathcal{N}_0$ are indices that identify your 3 closest neighbors

Our prediction for you is the average sleep from your 3 closest neighbors

KNN Regression Algorithm

To predict for any new observation $\mathbf{X_0}$

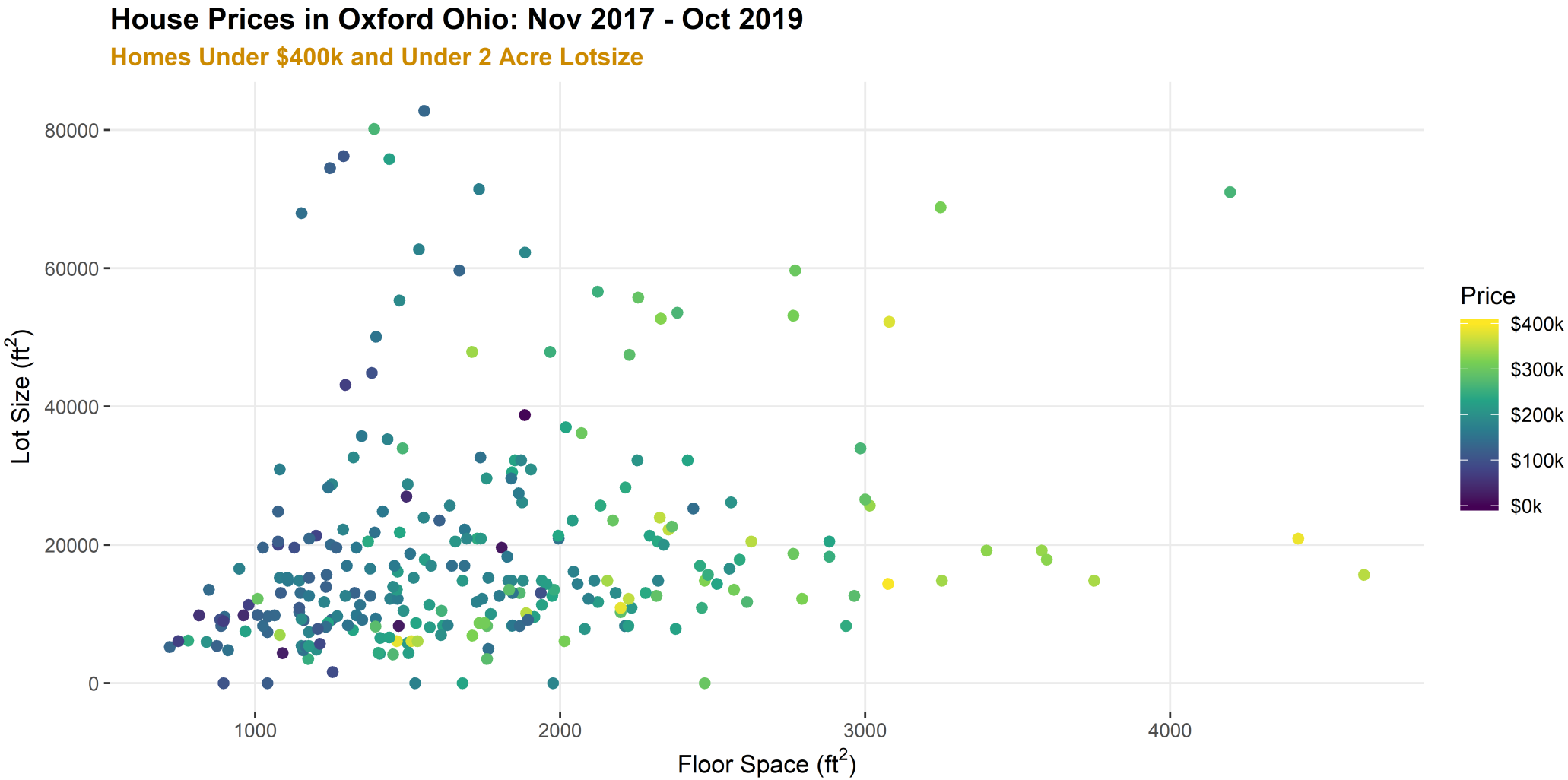
1. Compute the distances between $\mathbf{X_0}$ and all previous observations

2. Sort distances smallest to largest

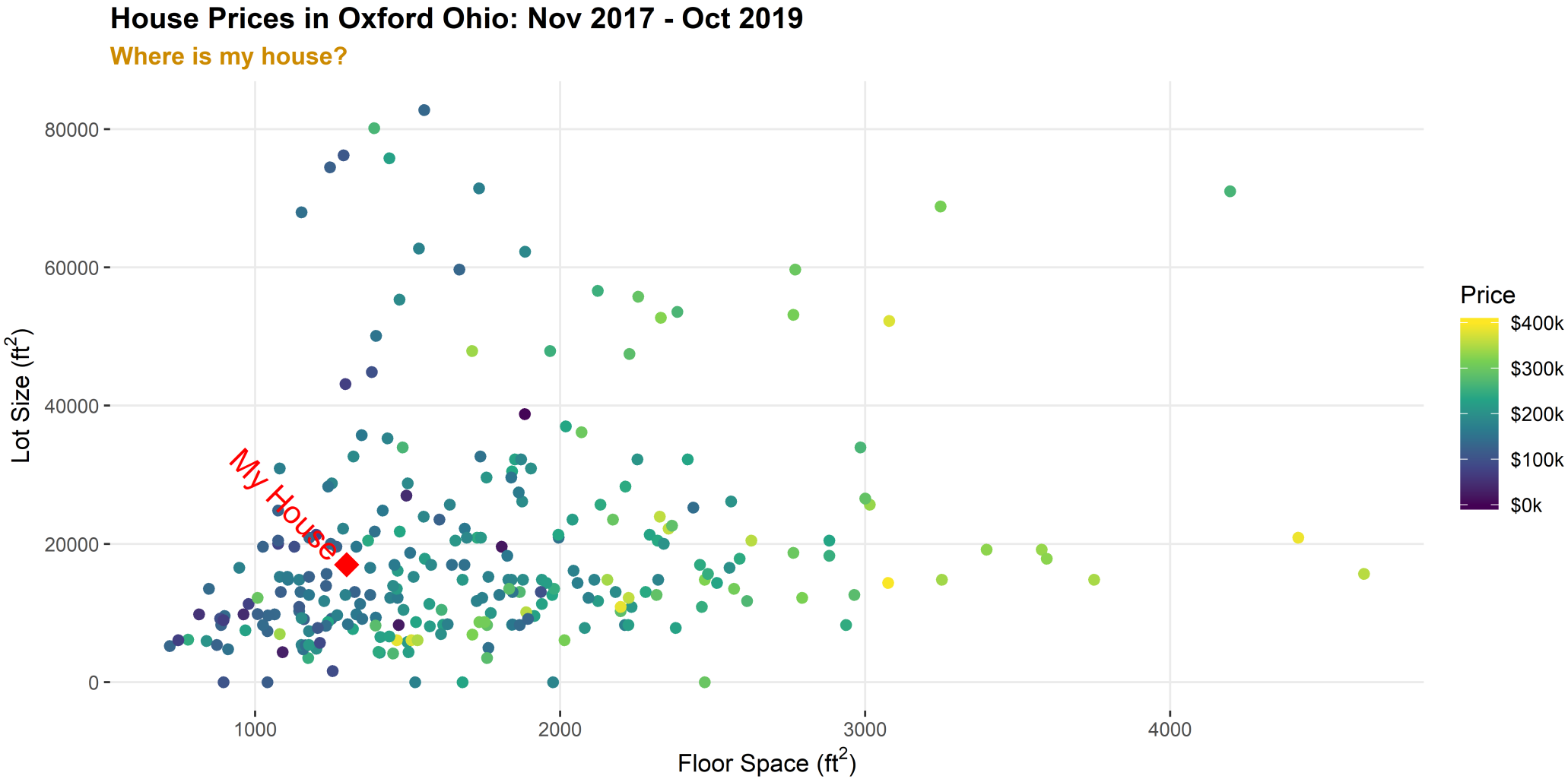
3. Select K smallest, record which observations these belong to

4. Compute the average response of these observations

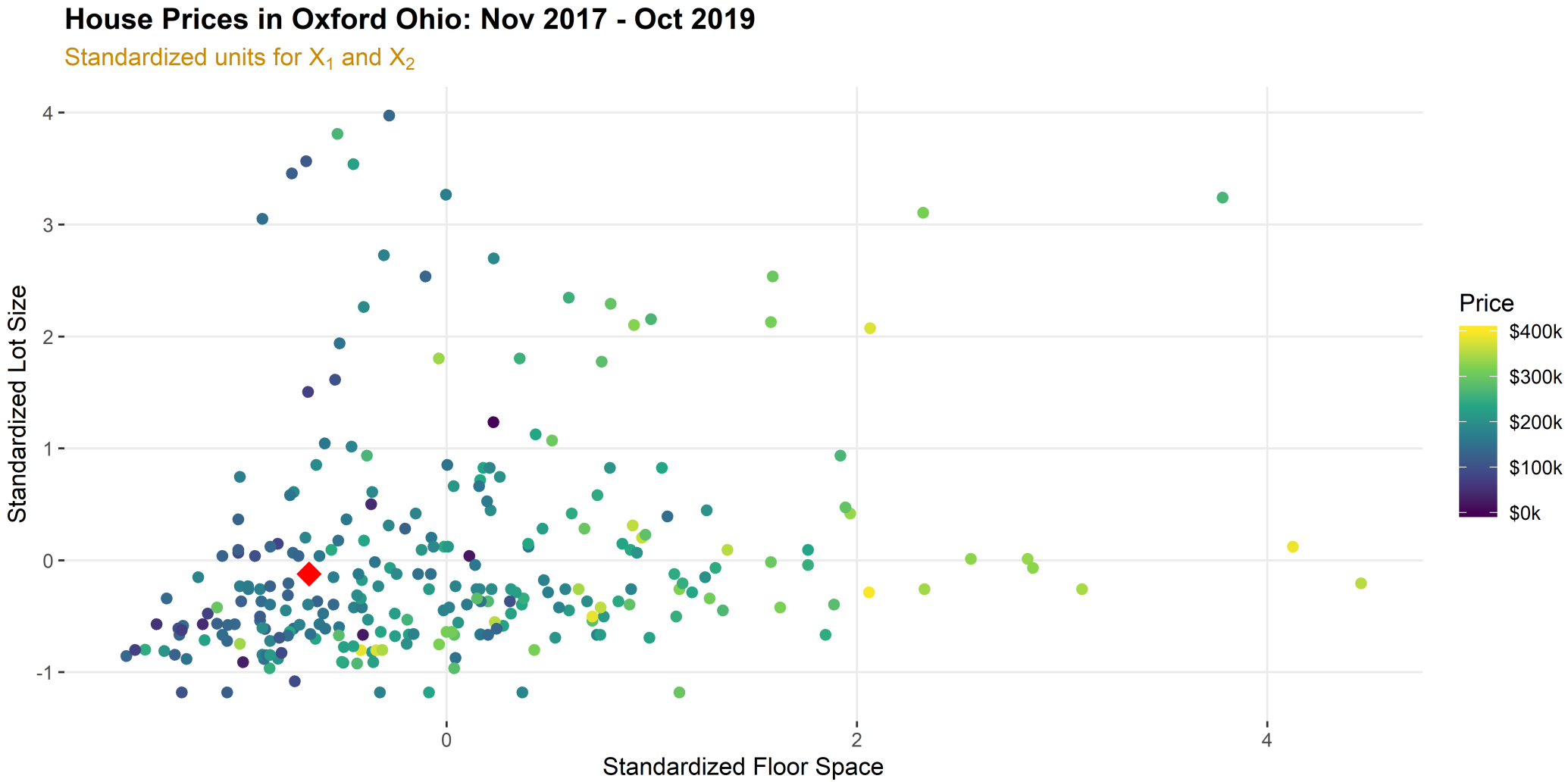
Case Study: Real Estate



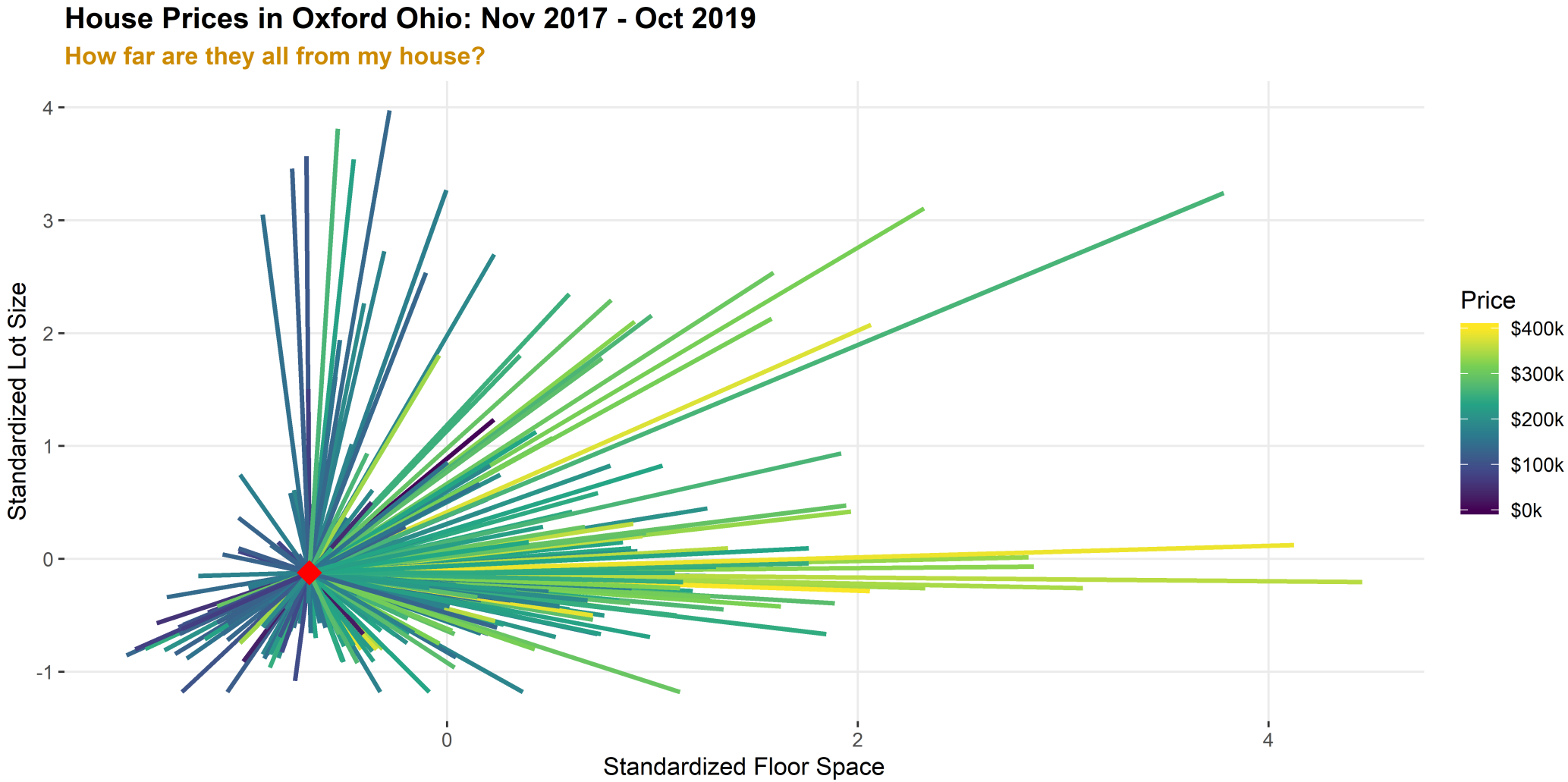
Case Study: Real Estate



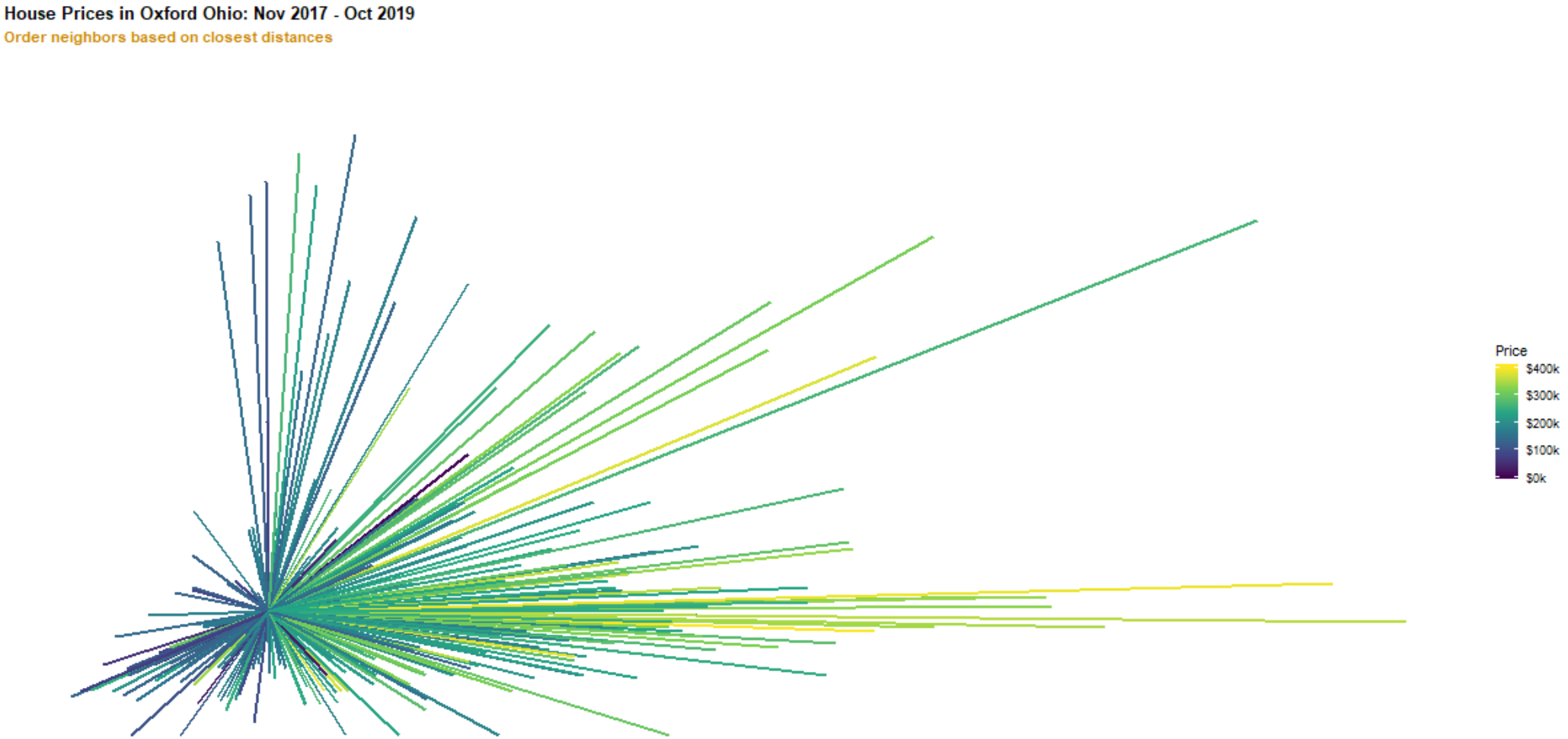
Case Study: Real Estate



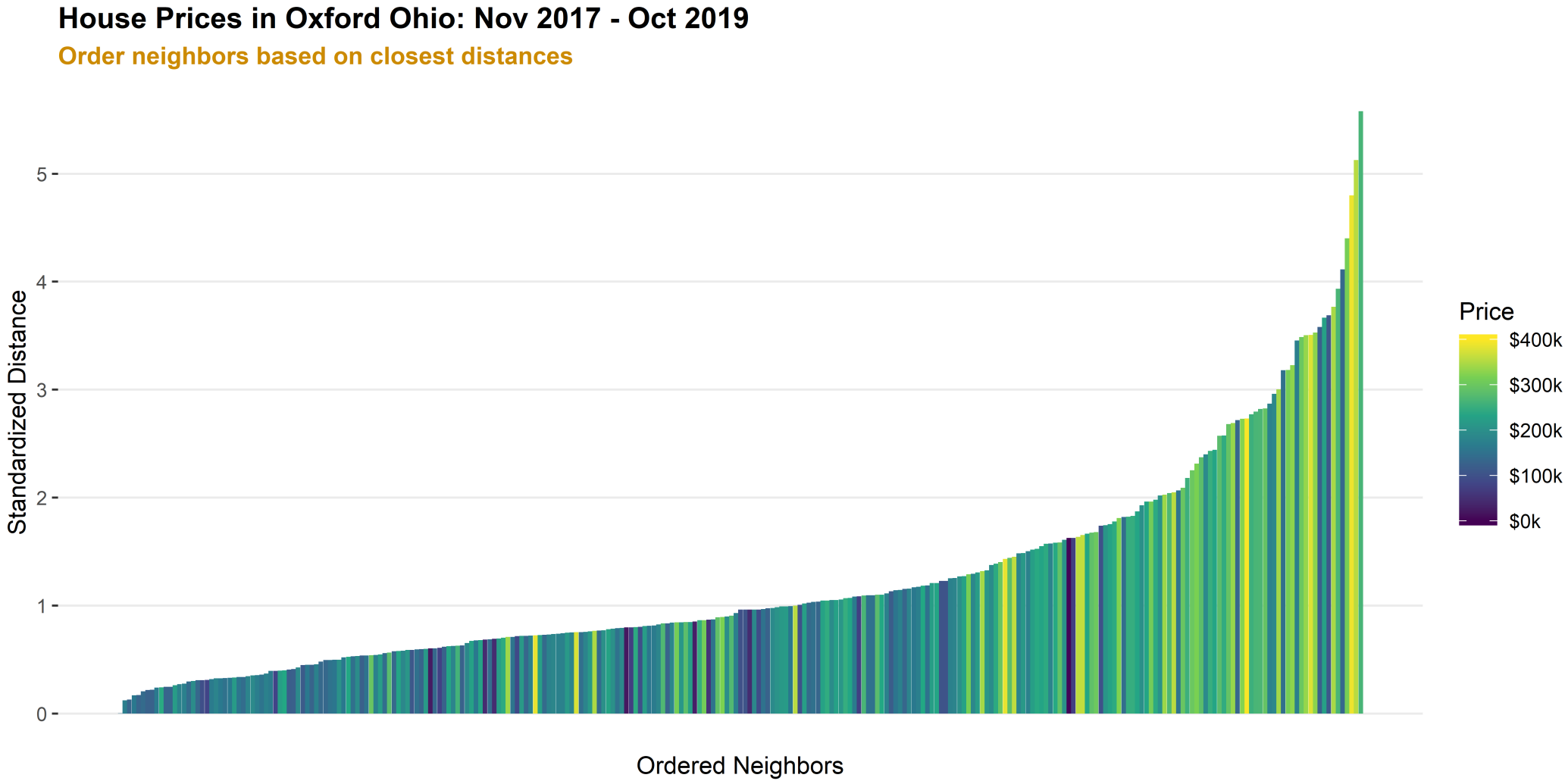
Case Study: Real Estate



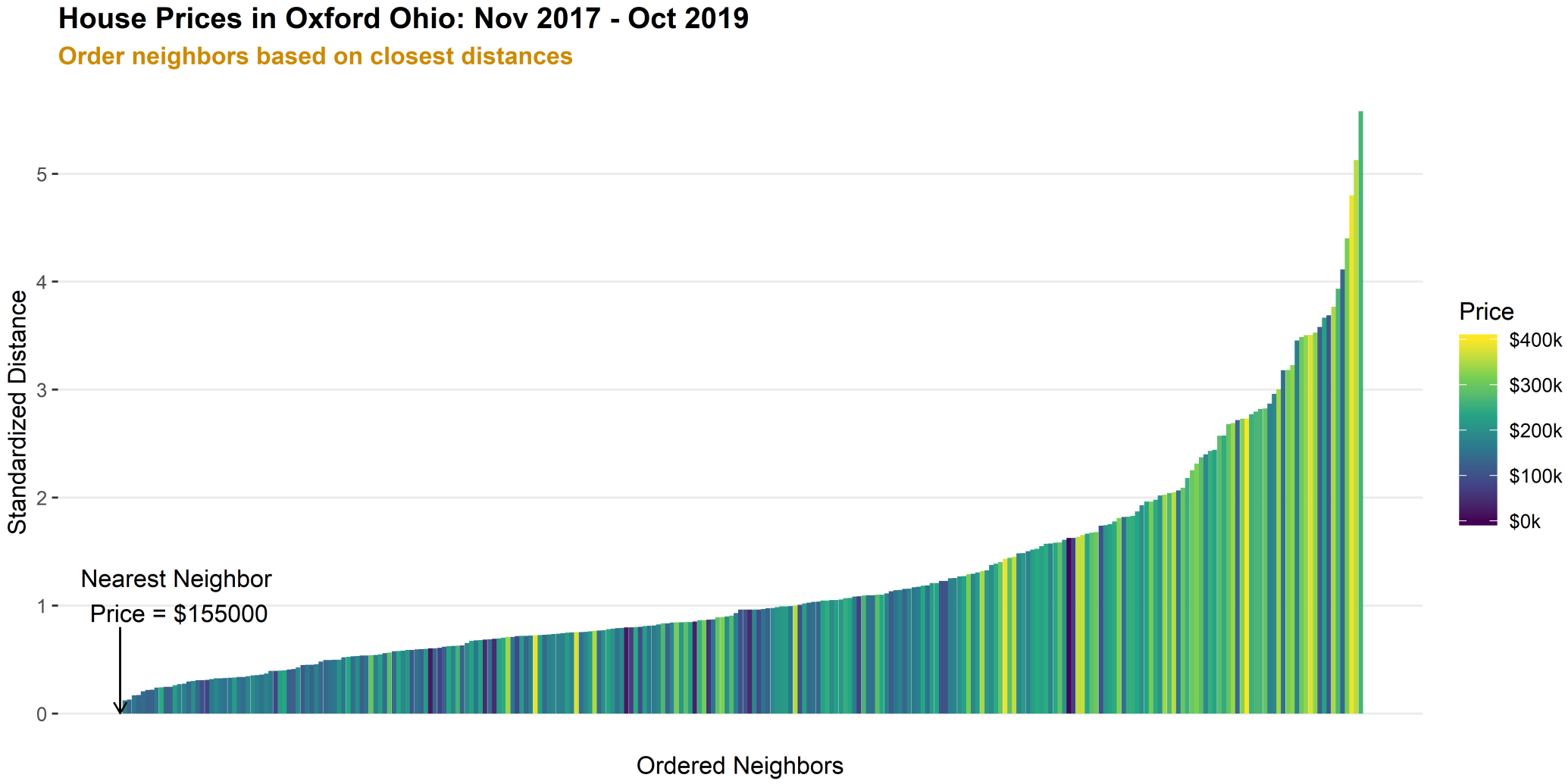
Case Study: Real Estate



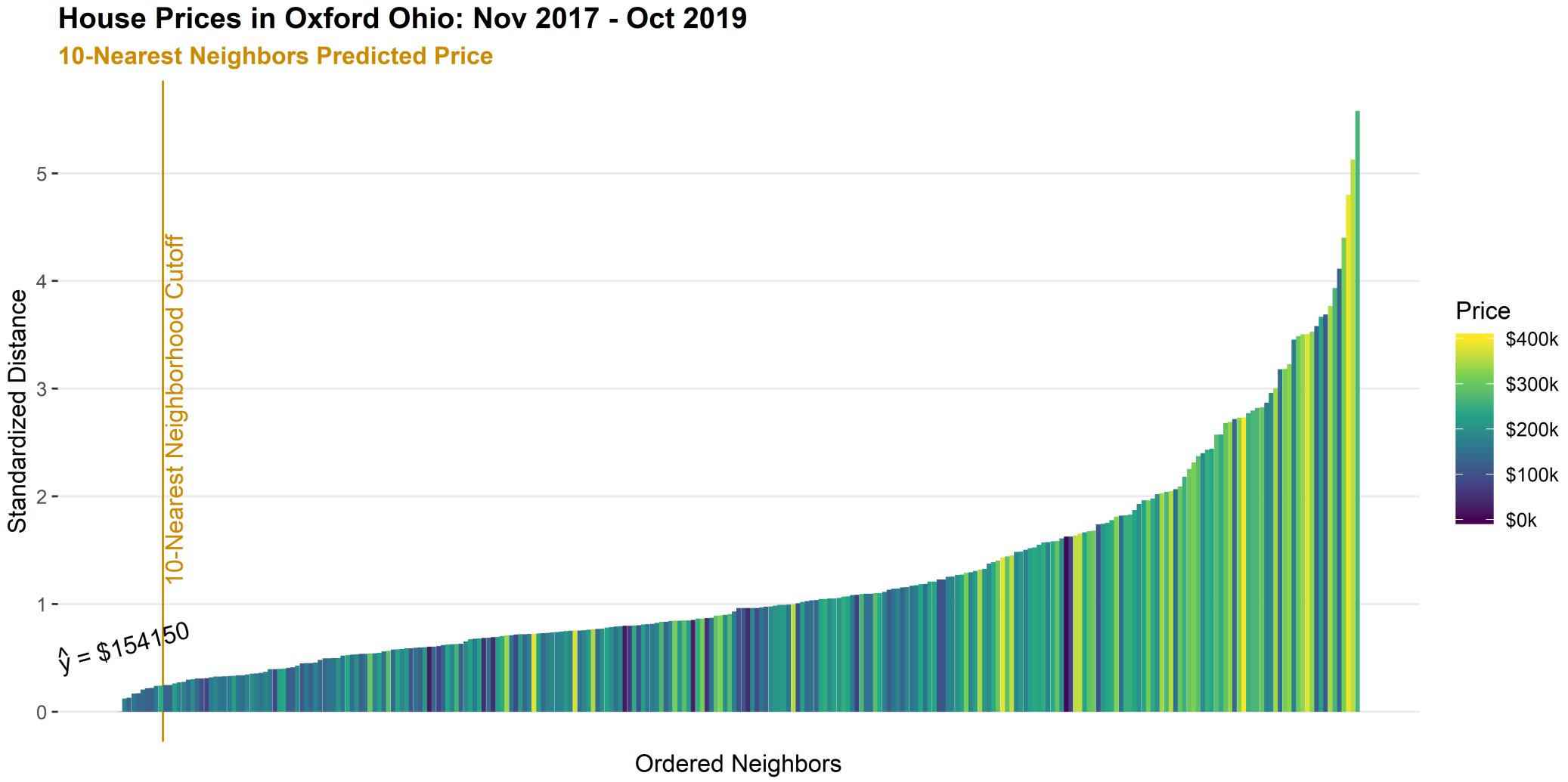
Case Study: Real Estate



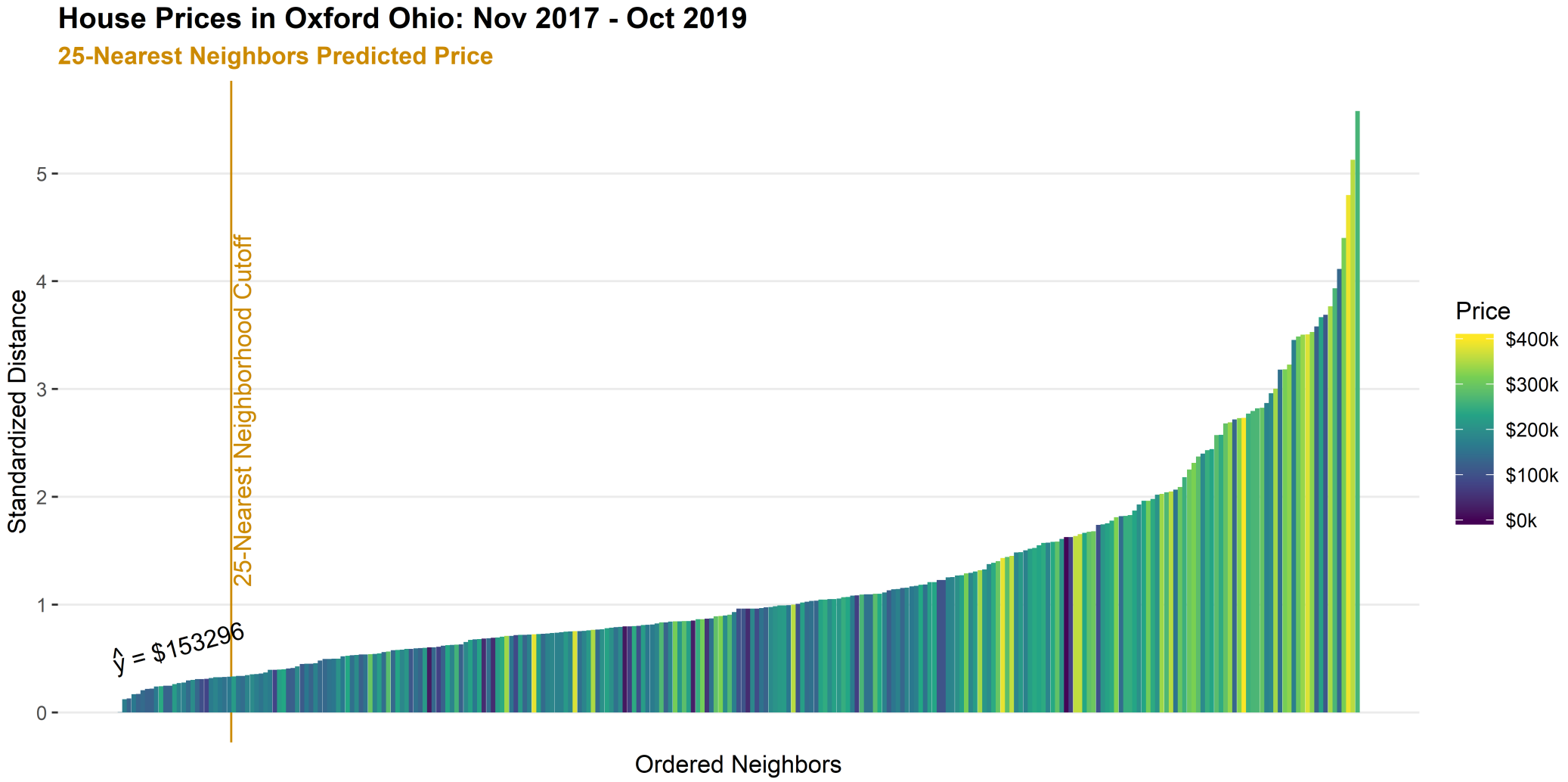
Case Study: Real Estate



Case Study: Real Estate



Case Study: Real Estate



A **K-Nearest Neighbor model** generates predictions based on the responses of the **K most similar** previously observed instances.

Now that we have some intuition for the algorithms,

how do we **use** KNN regression in an **applied** setting?

Let's **code** it up with the house prices data!

If you have your laptop with **R** and **RStudio** installed, please code along!

Starter Code Available at kmaurer.github.io

Benefits and Drawbacks of KNN

Predictions not based on linear combinations of input variables

- This could be *bad* if there is a linear relationship between inputs and the response
- This could be *good* if there is a non-linear, localized structure to predictive structure
- In the real estate context, how are prices set?

Computationally Demanding

- For each new instance, must calculate and sort n distances

Distance Metrics

- Euclidean distances depend on scaling in each dimension
- Other distance metrics possible (e.g. Manhattan distance)

Easily adapts for *classification* problems

- Make prediction based on *most common class* in neighborhood

Big Takeaways

A K-Nearest Neighbor model generates predictions based on the responses of the K most similar previously observed instances.

Algorithm: Compute the distances, Sort, Select K neighbors, Make prediction based on neighbors

There are user-friendly implementations of KNN available in open-source programming languages

Suggested Supplementary Readings

- James, Witten, Hastie and Tibshirani (2013) Introduction to Statistical Learning. URL *http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf*
- Grolemund and Wickham (2017) R For Data Science. URL *https://r4ds.had.co.nz/*

Software and Data

- Garrett, Nar, Fisher, Maurer (2018). ggvoronoi: Voronoi Diagrams and Heatmaps with ggplot2. J. Open Source Software, 3(32), 1096.
- Kuhn (2019). caret: Classification and Regression Training. R package version 6.0-84. *https://CRAN.R-project.org/package=caret*
- R Core Team (2019). R: A language and environment for statistical computing.
- R Foundation for Statistical Computing, Vienna, Austria. URL *https://www.R-project.org/*.
- RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL *http://www.rstudio.com/*.
- Temple Lang, CRAN Team (2019). XML: Tools for Parsing and Generating XML Within R and S-Plus. R package version 3.98-1.20. *https://CRAN.R-project.org/package=XML*
- Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. *https://CRAN.R-project.org/package=tidyverse*
- Wickham (2019). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.4. *https://CRAN.R-project.org/package=rvest*
- Xie (2019). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.22.
- Zillow home listings data from Oxford, Ohio. *https://www.zillow.com/*

Thanks!

Slides created via the R package **xaringan**.