Lecture 1

Introduction to Modeling

DSC 40A

Plan for Today

- Introductions.
- What is DSC 40A about?
- Logistics.
- Modeling.
- The constant model.

Introductions

...who the heck is this guy?

I'm Kyle

coffee enthusiast | entrepreneur | data soothsayer

At the end of the day my goal is to decipher order from the chaos. Join me.

kshannon@ucsd.edu Linkedin: kmshannon Github: kshannon

- From San Diego, I enjoy traveling, hiking, eating out, my dog 🐶 and family, tennis, etc.
- Undergrad + Masters @ UCSD 🪷 in Cognitive Neuroscience & DataSci/ML.
- Since 2019 teaching off and on at HDSI and Cognitive Science Dept.
    - DSC 40a, 100, and DSC 180ab Healthcare DataSci Section.
    - COGS 8, and 9

My Journey *(more or less)*

- Sprinkling of DS/ML Consulting gigs in the medical/healthcare DataSci field
- 2020 Director of Data Science - Volunteer Transportation Center, Inc. (VTC)
- 2019 Adjunct Professor UC San Diego
- 2018 Data Science Consultant - Healthgen.ai, Inc.
- 2016 CoFounder - Turnkey Trips, LLC `>>` merged w/ VTC in 2020
- 2013 Staff Data Scientist - Booz Allen Hamilton
- 2011 Research Associate & Educator - Backyard Brains, Inc.
- 2010 Research Assistant - UCSD

Education

- 2018 MAS in Data Science & Engineering - UCSD CSE
- 2012 BS in Cognitive Science & Neuroscience - UCSD CogSci

Course staff

We have several excellent tutors and TAs, all of whom are excited to help you in class, in discussion section and office hours! Read more about us at dsc40a.com/staff.

Questions 🤔

Any Thoughts?

Select the **FALSE** statement below.

A: I am currently starting center for the UC Health "dad" 🏀 b-ball league.

B: I have dual US/Argentinian 🇦🇷 citizenship.

C: In my first two years of college I worked full time as an apprentice 🪚 carpenter.

D: I play(ed) high(er) level tennis 🎾

E: I am less than 39.5 years old. Remember **age** is just a number! 👴

What is DSC 40A about?

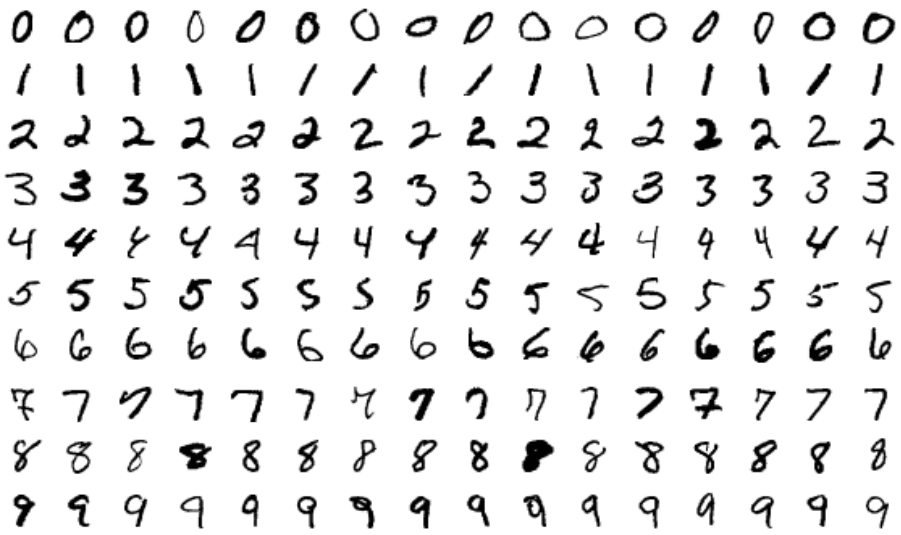
Theoretical Foundations of Data Science I

What have you *heard* about DSC 40A?

Here are some responses from the Welcome Survey this quarter.

- *I've heard the class seeks to uncover a lot of the key concepts of the math behind machine learning, while utilizing a lot of linear algebra. I've heard that the class can be difficult and proof-heavy.*
- *I heard it is conceptual, and therefore, a pretty hard class (to understand conceptually). I also heard it has a lot to do with linear algebra.*
- *That it's the most awful class in the DSC major, pretty much just pure math/all proofs.*
- *It's a pretty hard class but rewarding in the end.*
- *kyle is very chill!*

Why do we need to study theoretical foundations?



Machine learning is about **automatically learning patterns from data**.

Humans are good at understanding handwriting – but how do we get computers to understand handwriting?

Course overview

Part 1: Learning from Data (Weeks 1 - 6) || (🏖️ Weeks 1 - 3 🏖️)

- Summary statistics and loss functions; empirical risk minimization.
- Linear regression (including multiple variables); linear algebra.
- Clustering.

Part 2: Probability (Weeks 7 - 10) || (🏖️ Weeks 4 - 5 🏖️)

- Set theory and combinatorics; probability fundamentals.
- Conditional probability and independence.
- The Naïve Bayes classifier.

Learning objectives

After this quarter, you'll...

- understand the basic principles underlying almost every machine learning and data science method.
- be better prepared for the math in upper division: vector calculus, linear algebra, and probability.
- Understand the importance of theory and the connection to DataSci application.

What do DSC 80 students have to say about DSC 40A?

Here are some responses from the End-of-Quarter Survey last quarter in DSC 80.

- *study hardy, pay attention in DSC 40A and start work early :)*
- *40A and Math 18 is super important for this class. Don't wait till the last minute too!*
- *I think DSC40[A] was the most important prerequisite for this class.*

Logistics

Getting started

- The course website, **dsc40a.com**, contains all content. **Read the syllabus carefully!**
  - Click around; you'll find other helpful resources.
- Other sites you'll need to use:
  - **Gradescope** is where you'll submit all assignments. You'll be automatically added within 24 hours of enrolling.
  - **Ed** is where all announcements will be made. If you're not enrolled, there's a join link in the syllabus.
  - We aren't using Canvas.
- Make sure to fill out the **Welcome Survey** ASAP.

My Aproach to 40a based on a trial in the ☀️ Summer ☀️

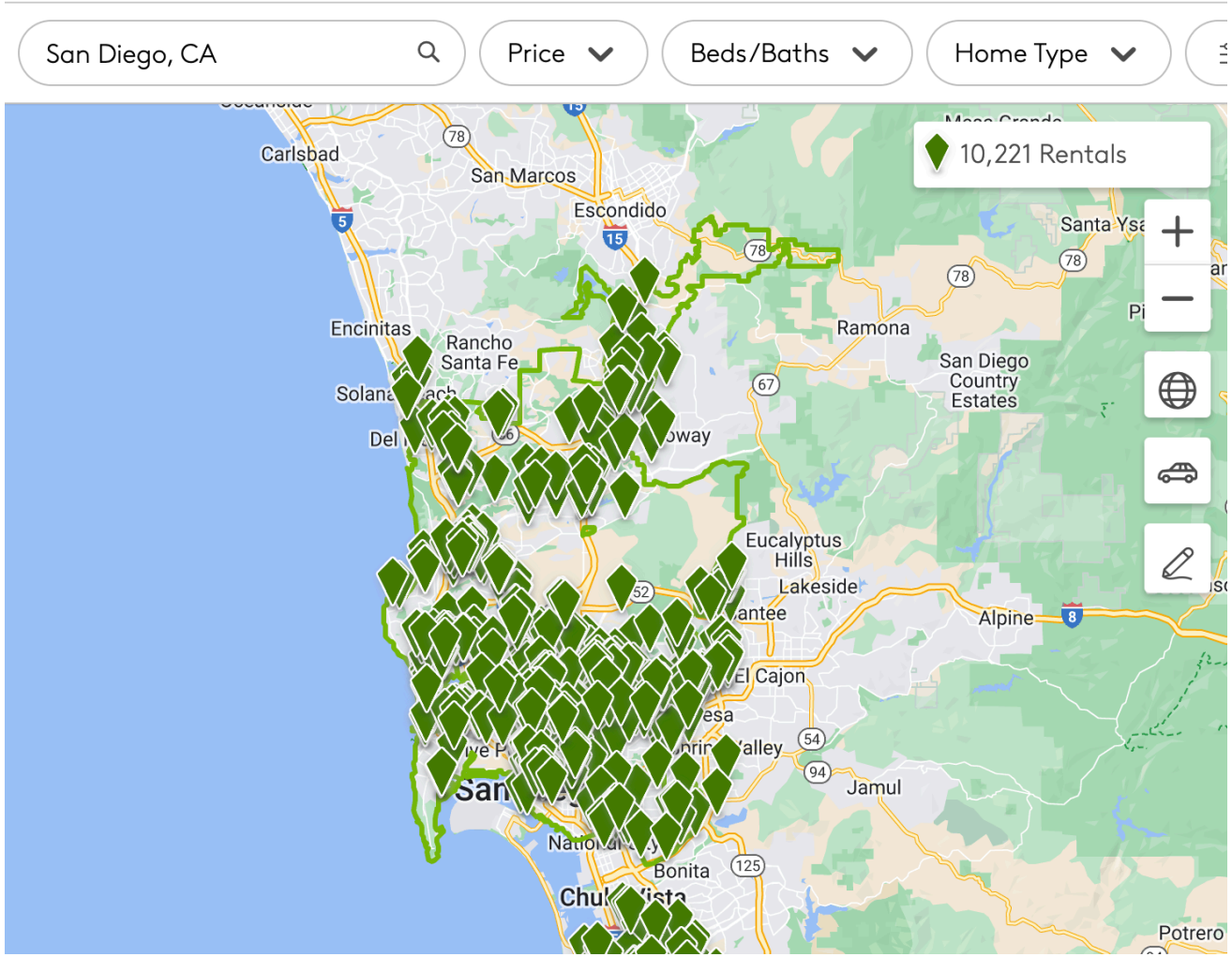
- I have designed this course to be a semi-flipped classroom!
  - You will watch pre-recorded lectures at home that have to do with theory.
  - In class time we will explore context driven problems that enforce course material.
  - You will have a chance to work on a problem alone, as if it was a quiz, to help you guage what you know so far and where you need to improve.
- Lecture slides will be posted on the course website along with annotated version from the recordings.
- All other material, e.g. in-class notebooks + notebook solutions will be made available on the course website.
- **The value of lecture is interaction and discussion, so even though attendance isn't required, it's highly, highly recommended.**

Support

We know this is a challenging class, and we're here to help:

- **Office hours**: Many OHs, most in HDSI, some on zoom. Plan to attend at least several a week.
- **Ed**: Use it! We're here to help you. Post conceptual questions publicly – just don't post answers to homework questions.
- We're developing practice.dsc40a.com to give you access to practice exam problems, categorized by topic.
- We're recording walkthrough videos to show you our thought process when answering questions.
- We're planning to spend more time reviewing linear algebra 📏, probability 🎲, and maybe some calculus 🧮.

Modeling



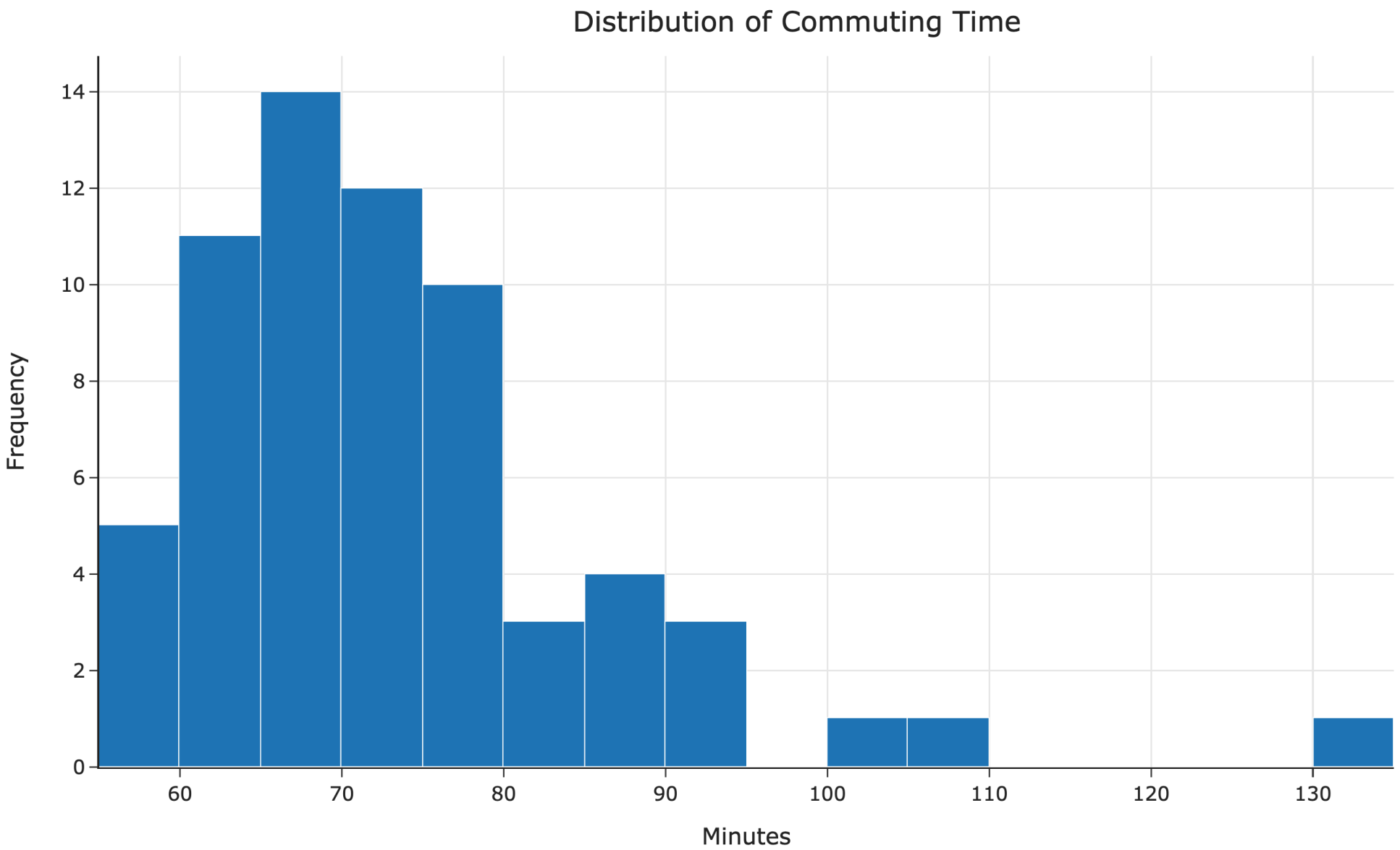
You might be starting to look for off-campus apartments, none of which are affordable.

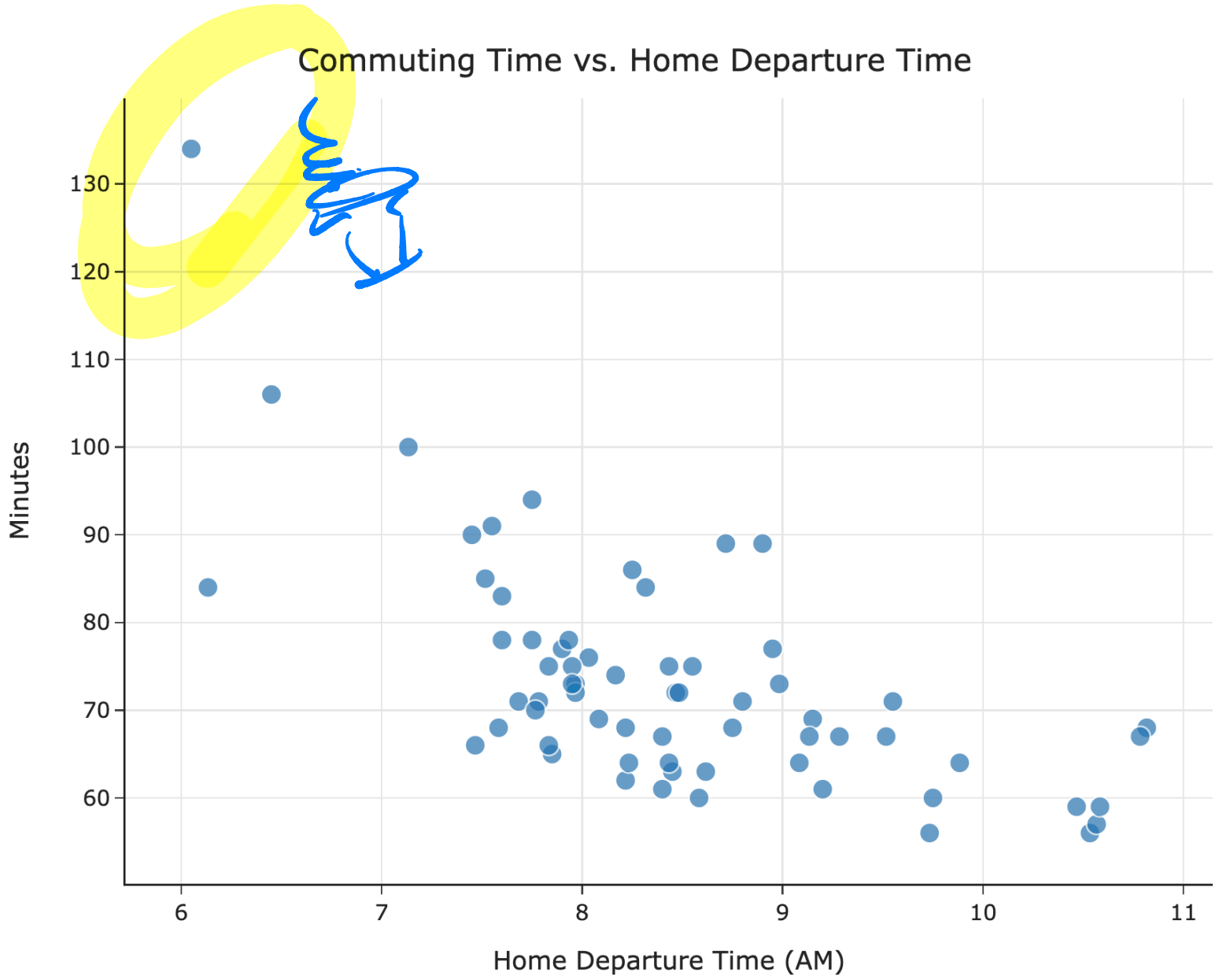
	date	day	departure_hour	minutes
0	5/22/2023	Mon	8.450000	63.0
1	9/18/2023	Mon	7.950000	75.0
2	10/17/2023	Tue	10.466667	59.0
3	11/28/2023	Tue	8.900000	89.0
4	2/15/2024	Thu	8.083333	69.0

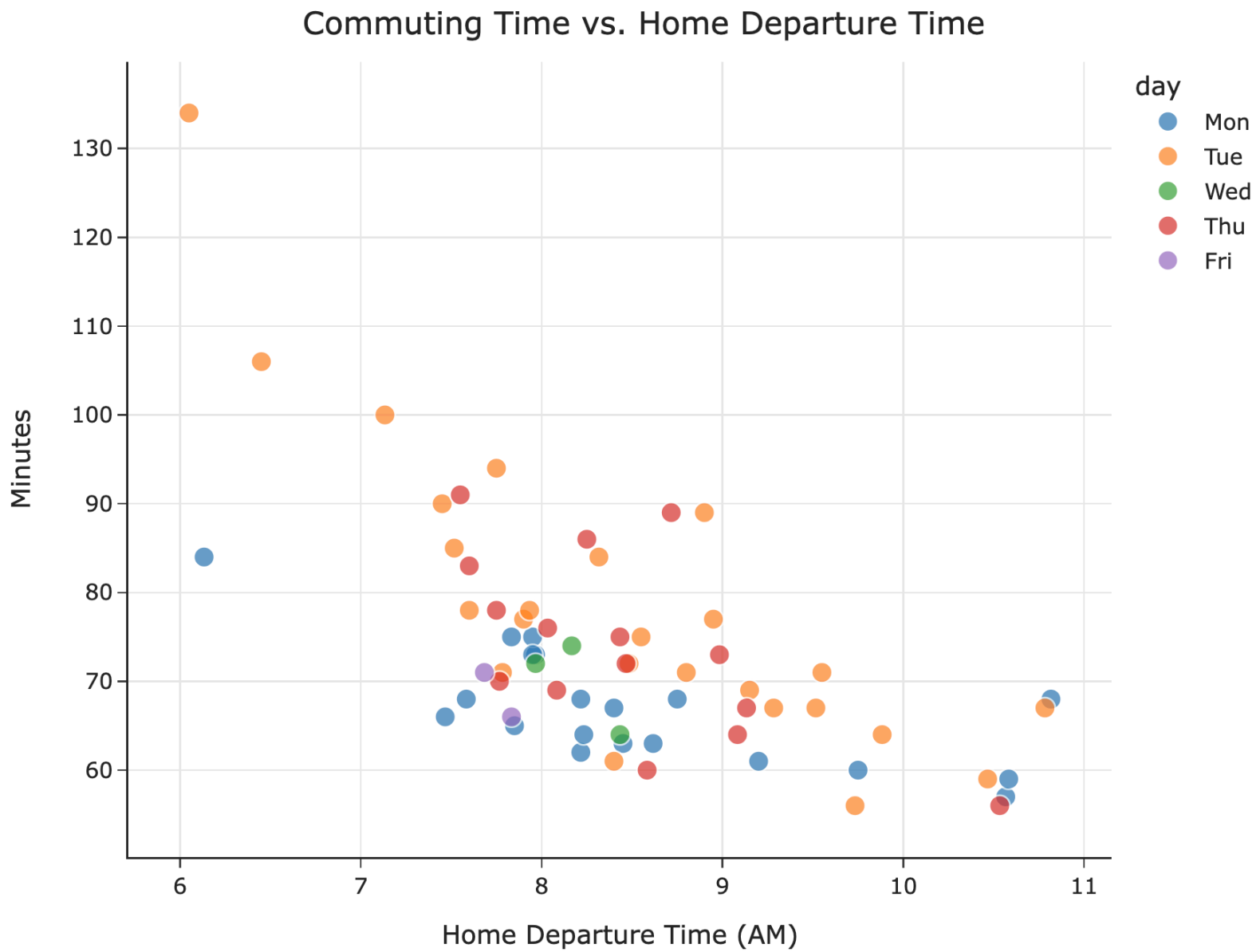
...

You decide to live with your parents in Orange County and commute.
You keep track of how long it takes you to get to school each day.

This is a real dataset, collected by Joseph Hearn! However, he lived in the Seattle area, not San Diego.

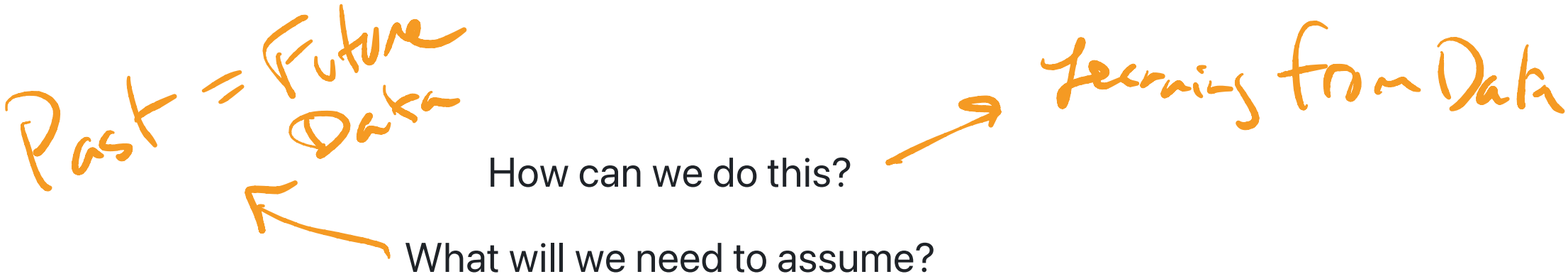

Distribution of Commuting Time
Frequency
0
2
4
6
8
10
12
14
60
70
80
90
100
110
120
130
Minutes


Commuting Time vs. Home Departure Time
Minutes
130
120
110
100
90
80
70
60
6
7
8
9
10
11
Home Departure Time (AM)


Commuting Time vs. Home Departure Time
Minutes
130
120
110
100
90
80
70
60
6
7
8
9
10
11
Home Departure Time (AM)
day
Mon
Tue
Wed
Thu
Fri

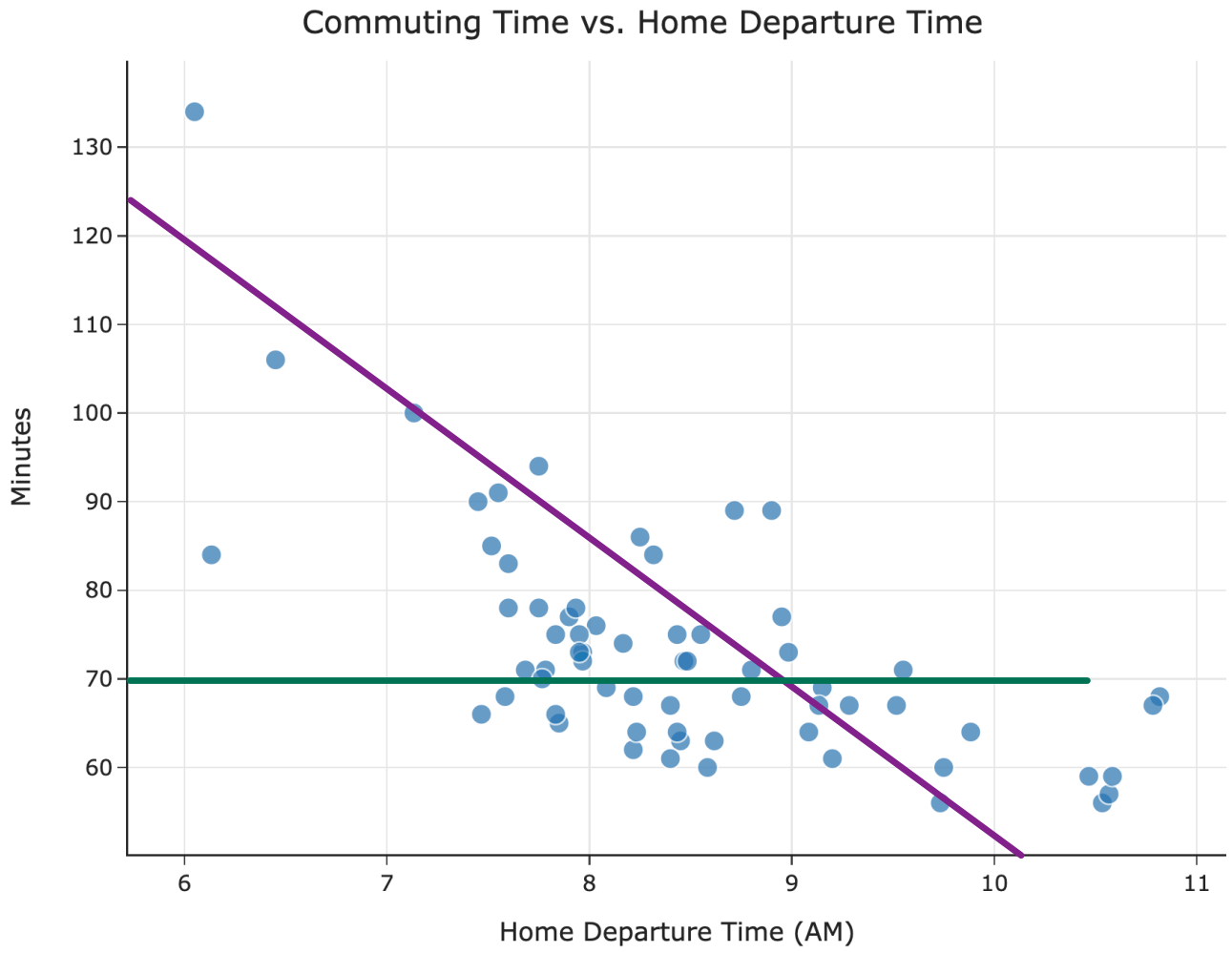
Goal: Predict your commute time.

That is, predict how long it'll take to get to school.



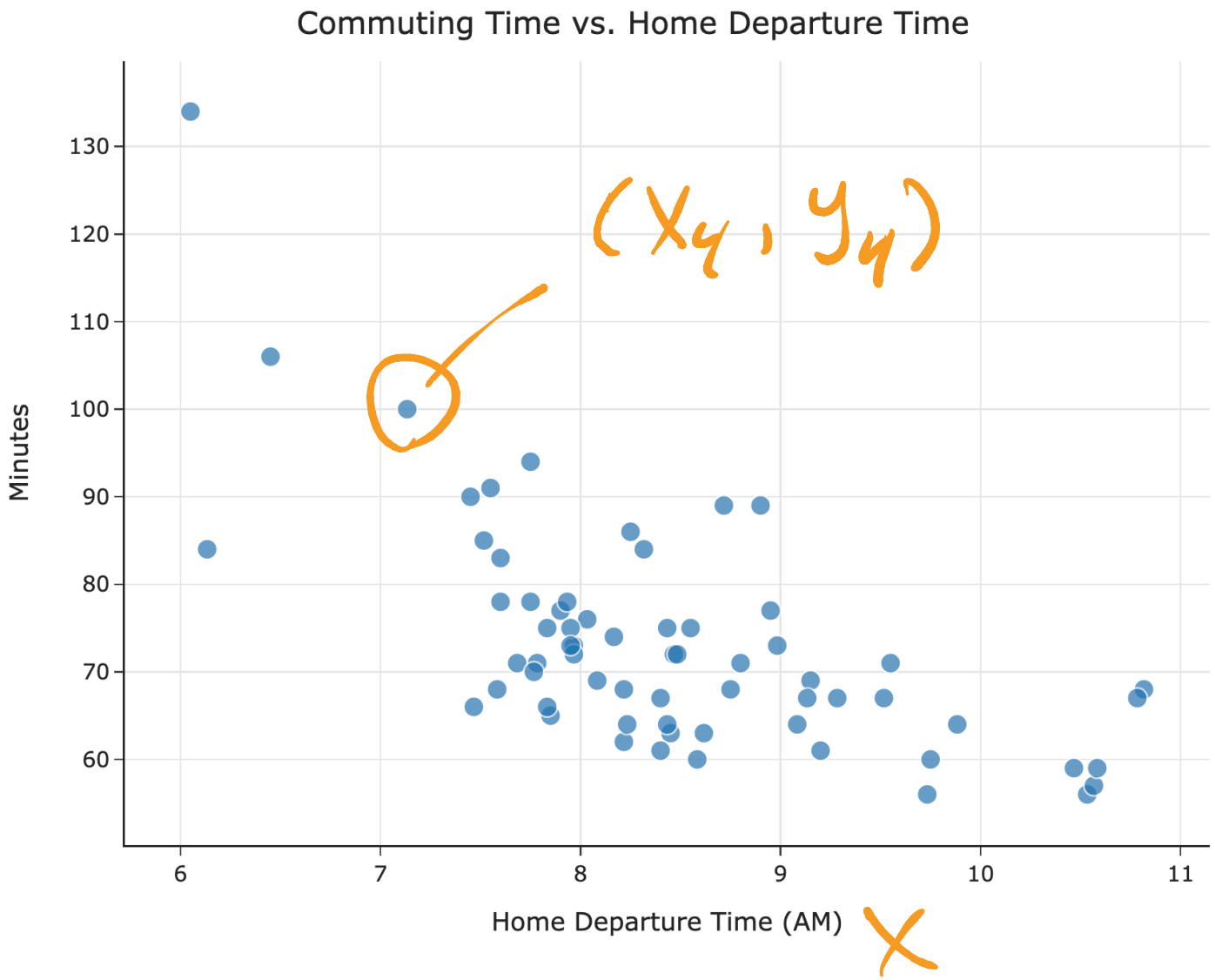
A **model** is a set of assumptions about how data were generated.

Possible models



- Linear Regression
- Constant model

Notation



X: is our input, feature, independent var

Y: Response var, target

use X to predict y

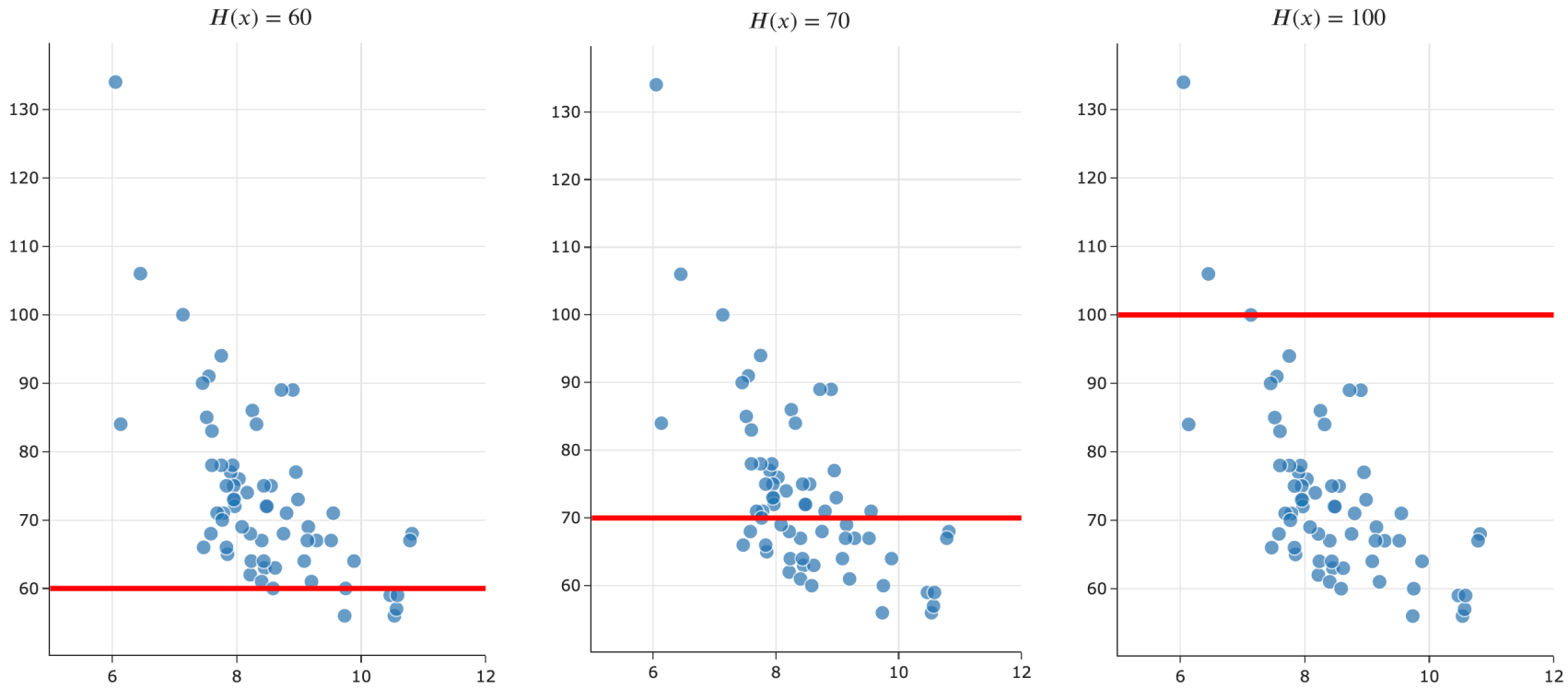
(x_i, y_i)

Hypothesis functions and parameters

A hypothesis function, H, takes in an x as input and returns a predicted y.

Parameters define the relationship between the input and output of a hypothesis function.

The constant model, $H(x) = h$, has one parameter: h.




```
def H(x):
    return 70
```

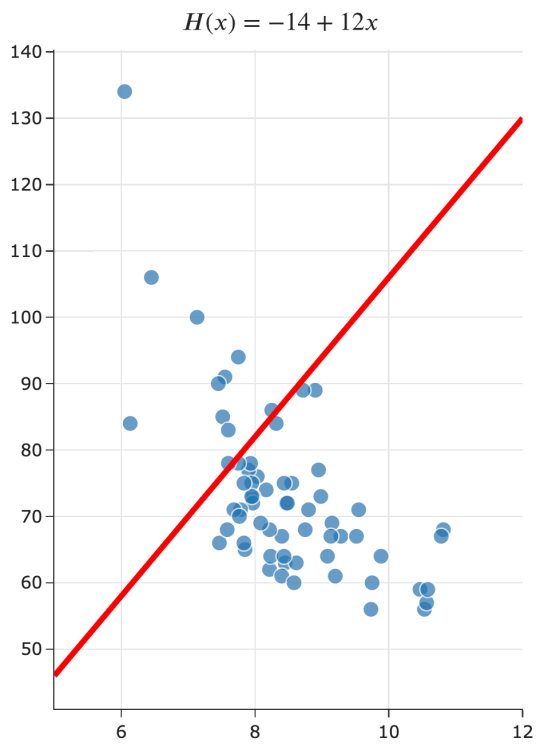
Hypothesis functions and parameters

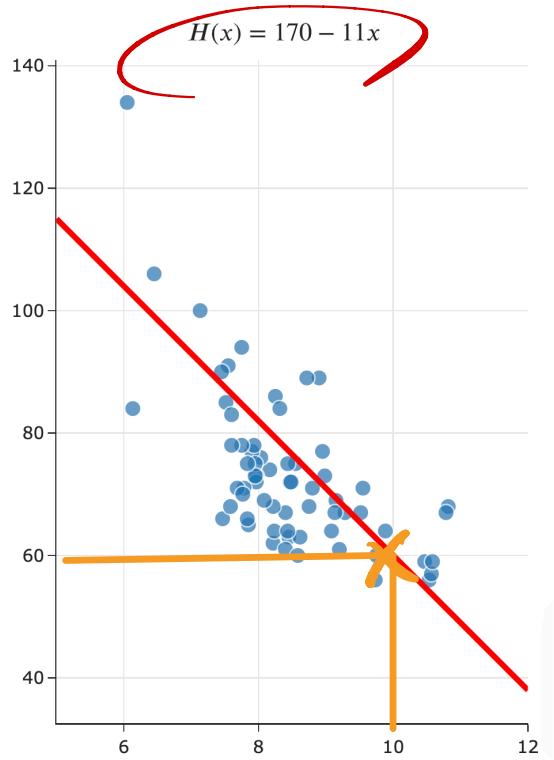
A hypothesis function, H, takes in an x as input and returns a predicted y.

Parameters define the relationship between the input and output of a hypothesis function.

The simple linear regression model, $H(x) = w_0 + w_1 x$, has two parameters: w_0 and w_1.

Int. slope





$y = mx + b$

$H(x) = 170 - 11(x)$

$H(10) = 170 - 11(10) = 60$ min.

y

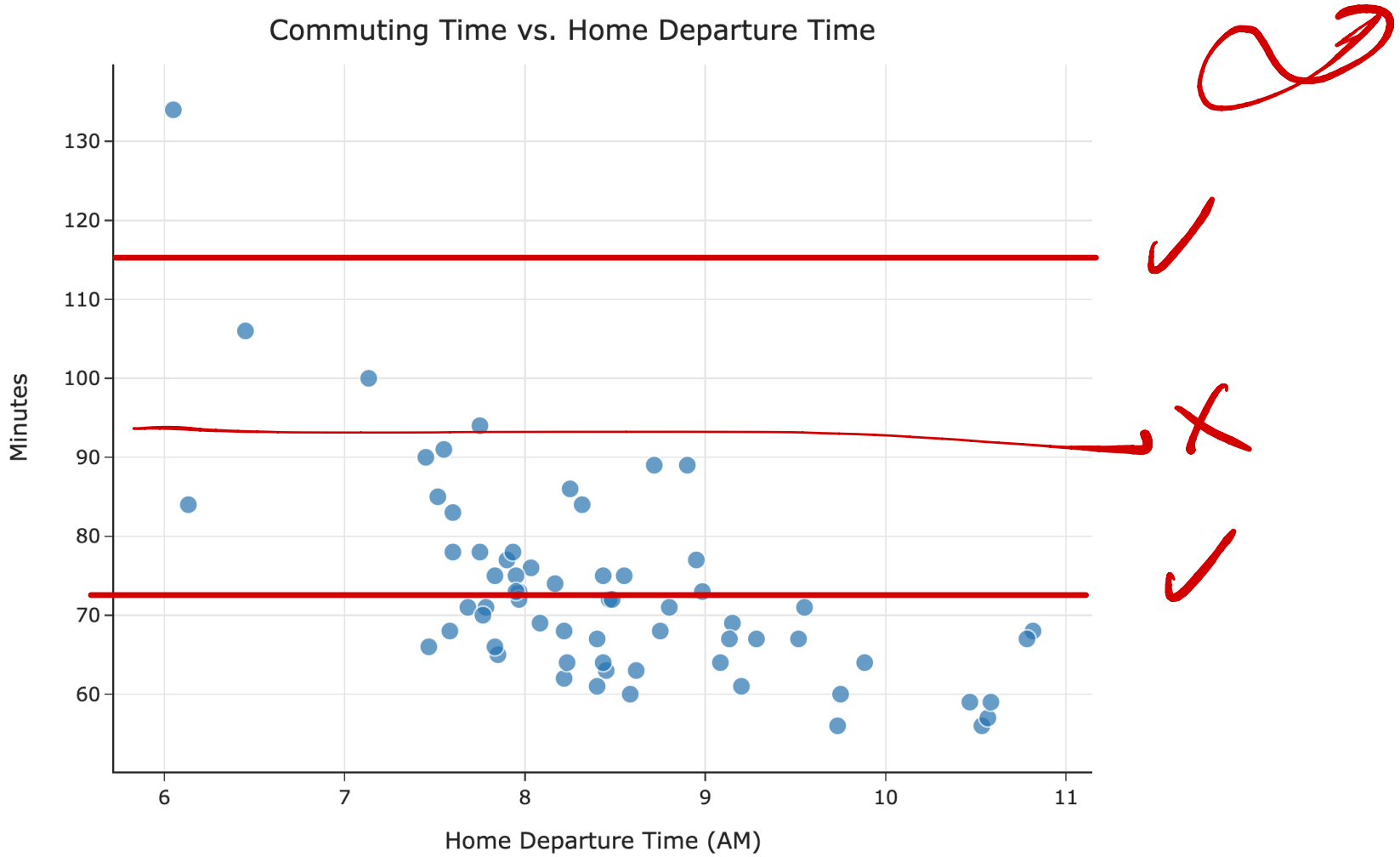
@ 10am

params.

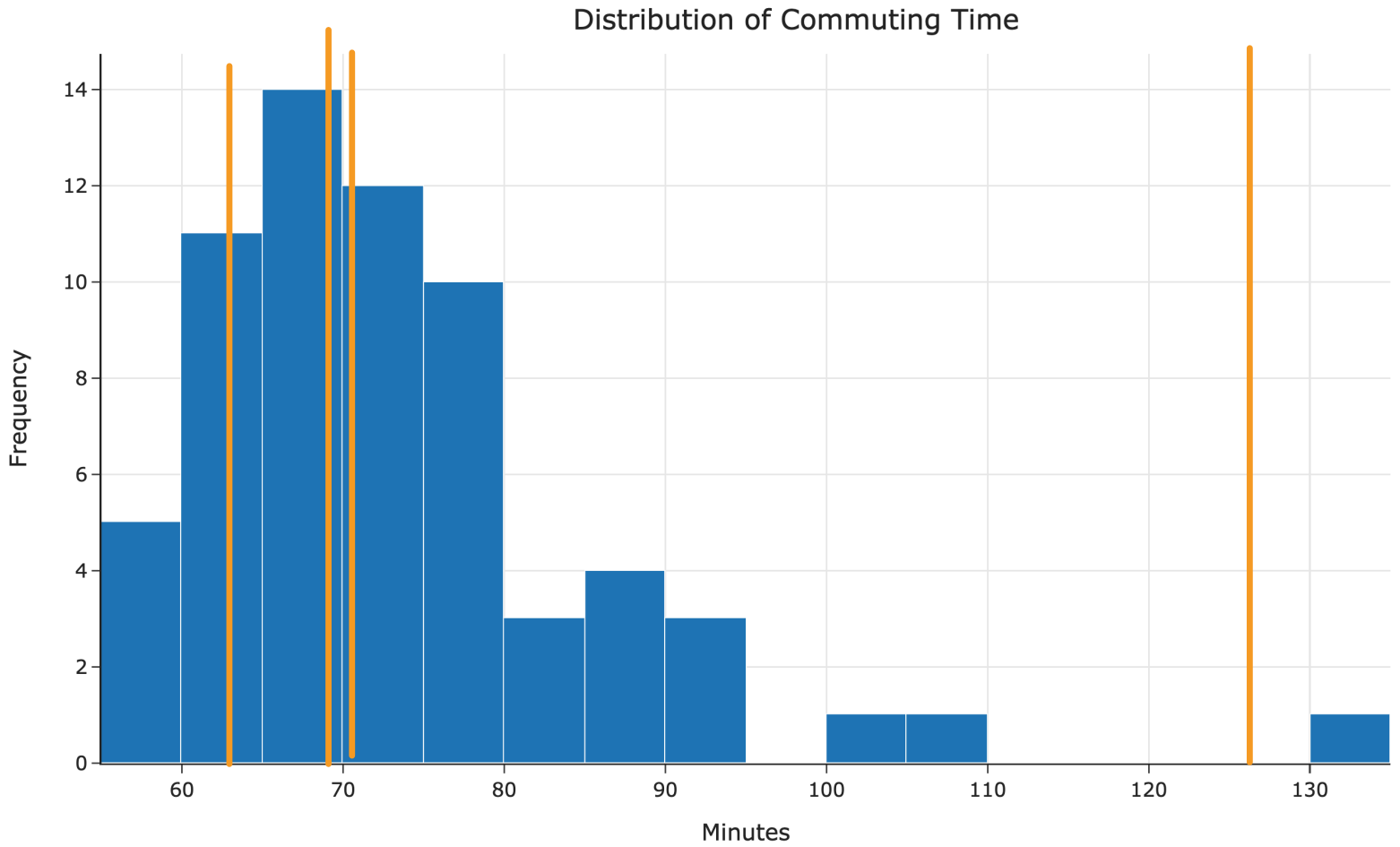
x

The constant model

The constant model



The constant model



A concrete example

Let's suppose we have just a smaller dataset of just five historical commute times in minutes.

$$
\begin{aligned}
y_1 &= 72 \\
y_2 &= 90 \\
y_3 &= 61 \\
y_4 &= 85 \\
y_5 &= 92
\end{aligned}
$$

Given this data, can you come up with a prediction for your future commute time? How?

mean: 80 midrange: 76.5 quartiles

median: 85 max min mode last y_5 or first y_1

Some common approaches

- The **mean**:

$$\frac{1}{5}(72 + 90 + 61 + 85 + 92) = \boxed{80}$$

- The **median**:

$$61 \quad 72 \quad \boxed{85} \quad 90 \quad 92$$

- Both of these are familiar **summary statistics** – they summarize a collection of numbers with a single number.
- But which one is better? Is there a "best" prediction we can make?

The cost of making predictions

A **loss function** quantifies how bad a prediction is for a single data point.

- If our prediction is **close** to the actual value, we should have **low** loss.
- If our prediction is **far** from the actual value, we should have **high** loss.

A good starting point is error, which is the difference between **actual** and **predicted** values.

$$e_i = y_i - H(x_i)$$

Actual value

Predicted value

Suppose my commute **actually** takes 80 minutes.

- If I predict 75 minutes: error = 80 - 75 = 5
- If I predict 72 minutes: error = 80 - 72 = 8
- If I predict 100 minutes: error = 80 - 100 = -20

Square errors

Absolute value

Squared loss

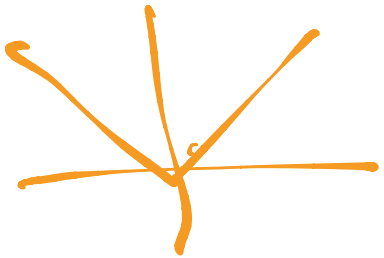
One loss function is squared loss, L_{sq}, which computes $(\text{actual} - \text{predicted})^2$.

$$L_{\text{sq}}(y_i, H(x_i)) = (y_i - H(x_i))^2$$

Note that for the constant model, $H(x_i) = h$, so we can simplify this to:

$$L_{\text{sq}}(y_i, h) = (y_i - h)^2$$

$(y_i - h)^2 = (h - y_i)^2$



Squared loss is not the only loss function that exists! Soon, we'll learn about absolute loss.

A concrete example, revisited

Consider again our smaller dataset of just five historical commute times in minutes. Suppose we predict the median, $h = 85$. What is the squared loss of 85 for each data point?

$y_1 = 72$ → $(72 - 85)^2 = 169$

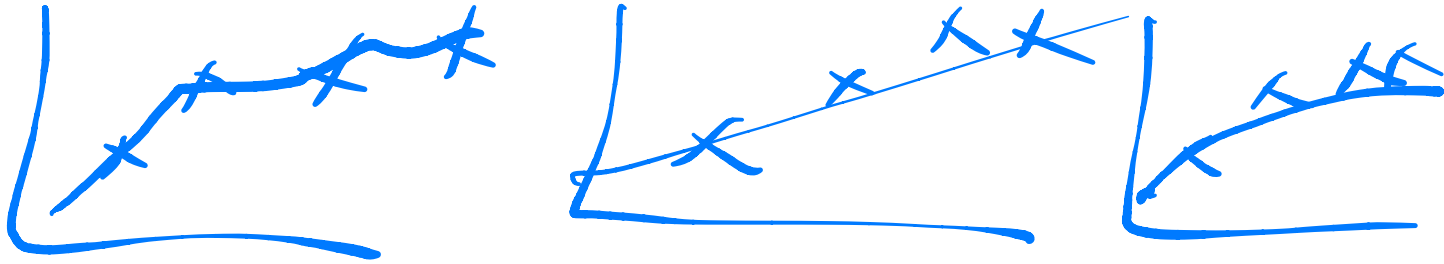
$y_2 = 90$ → $= 25$

$y_3 = 61$ $= 576$

$y_4 = 85$ $= 0$

$y_5 = 92$ $= 49$

Averaging squared losses



We'd like a single number that describes the quality of our predictions across our entire dataset. One way to compute this is as the **average of the squared losses**.

- For the median, $h = 85$:

$$\frac{1}{5}\left((72 - 85)^2 + (90 - 85)^2 + (61 - 85)^2 + (85 - 85)^2 + (92 - 85)^2\right) = \boxed{163.8}$$

- For the mean, $h = 80$:

$$\frac{1}{5}\left((72 - 80)^2 + (90 - 80)^2 + (61 - 80)^2 + (85 - 80)^2 + (92 - 80)^2\right) = \boxed{138.8}$$

Which prediction is better? Could there be an even better prediction?

Mean squared error

- Another term for **average squared loss** is **mean squared error (MSE)**.
- The mean squared error on our smaller dataset for any prediction h is of the form:

$$R_{\text{sq}}(h) = \frac{1}{5}\left((72-h)^2 + (90-h)^2 + (61-h)^2 + (85-h)^2 + (92-h)^2\right)$$

R stands for "risk", as in "empirical risk." We'll see this term again soon.

- For example, if we predict $h = 100$, then:

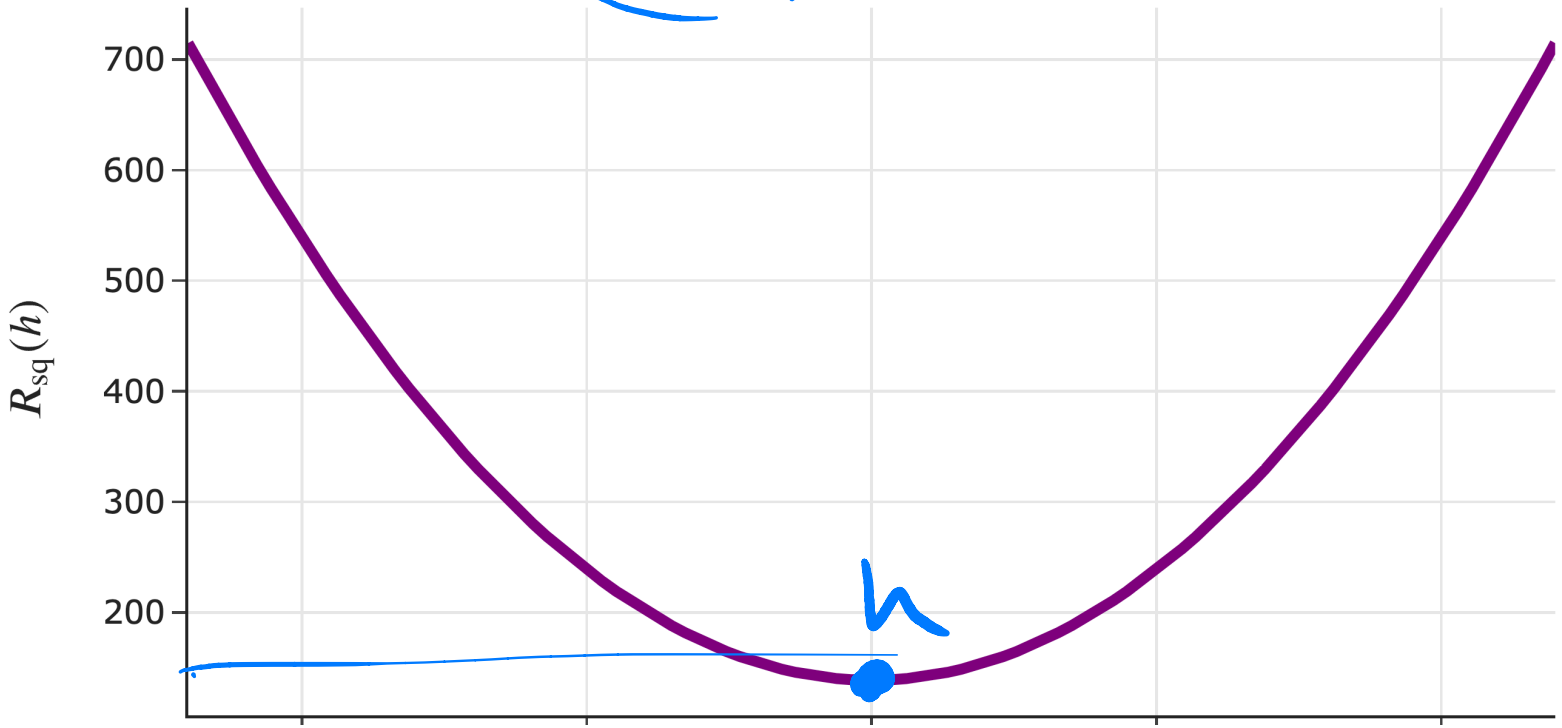
$$R_{\text{sq}}(100) = \frac{1}{5}\left((72-100)^2 + (90-100)^2 + (61-100)^2 + (85-100)^2 + (92-100)^2\right)$$

$$= \boxed{538.8}$$

- **We can pick any h as a prediction, but the smaller $R_{\text{sq}}(h)$ is, the better h is!**

Visualizing mean squared error

$$R_{\text{sq}}(h) = \frac{1}{5}\left((72-h)^2 + (90-h)^2 + (61-h)^2 + (85-h)^2 + (92-h)^2\right)$$



Which h corresponds to the vertex of $R_{\text{sq}}(h)$?

Mean squared error, in general

- Suppose we collect n commute times, $y_1, y_2, \ldots, y_n$.
- The mean squared error of the prediction h is:

$$R_{sq}(h) = \frac{1}{n}\left[(y_1 - h)^2 + (y_2 - h)^2 + \cdots + (y_i - h)^2\right]$$

- Or, using **summation notation**:

$$R_{sq}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2$$

```
total = 0
for i in range(1, n+1):
    total += (y[i] - h)**2
total = total/n
```

The best prediction

$$R_{\text{sq}}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2$$

- We want the **best** prediction, h^*.
- The smaller $R_{\text{sq}}(h)$ is, the better h is.
- **Goal**: Find the h that minimizes $R_{\text{sq}}(h)$.
  The resulting h will be called h^*.
- **How do we find h^*?** w/ calc. i.e. set Derivative = 0

Summary, next time

- We started with the abstract problem:

> Given historical commute times, predict your future commute time.

- We've turned it into a formal optimization problem:

> Find the prediction h^* that has the smallest mean squared error $R_{\text{sq}}(h)$ on the data.

- Implicitly, we introduced a three-step modeling process that we'll keep revisiting:
    - i. Choose a model.
    - ii. Choose a loss function.
    - iii. Minimize average loss, R.
- **Next time**: We'll solve this optimization problem by-hand.