# Introduction to Regression
## Part A - kNN

Pavlos Protopapas, Natesh Pillai and Chris Gumb

# Lecture Outline

**Part A: Statistical Modeling**

k-Nearest Neighbors (kNN)

**Part B: Model Fitness**

How does the model perform predicting?

How do we choose from two different models?

# Response and Predictor Variables

# Predicting a Variable

Let's consider a scenario in which we aim to predict the value of one variable based on another variable or a set of other variables.

**Examples:**

Predicting the number of views that a **TikTok** video will receive next week, based on factors such as video length, posting date, and previous view count.

# Predicting a Variable

Let's consider a scenario in which we aim to predict the value of one variable based on another variable or a set of other variables.

**Examples:**

Forecasting which movies, a Netflix user is likely to rate highly, considering their previous movie ratings and demographic data.

# Working example

The Advertising dataset contains sales (in 1000 units) data for a specific product across 200 different markets. It also includes advertising budgets in $1000 allocated to three different media channels: *TV, radio, and newspaper*, for each of those markets.

| TV | radio | newspaper | sales |
|---|---|---|---|
| 230.1 | 37.8 | 69.2 | 22.1 |
| 44.5 | 39.3 | 45.1 | 10.4 |
| 17.2 | 45.9 | 69.3 | 9.3 |
| 151.5 | 41.3 | 58.5 | 18.5 |
| 180.8 | 10.8 | 58.4 | 12.9 |

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani "

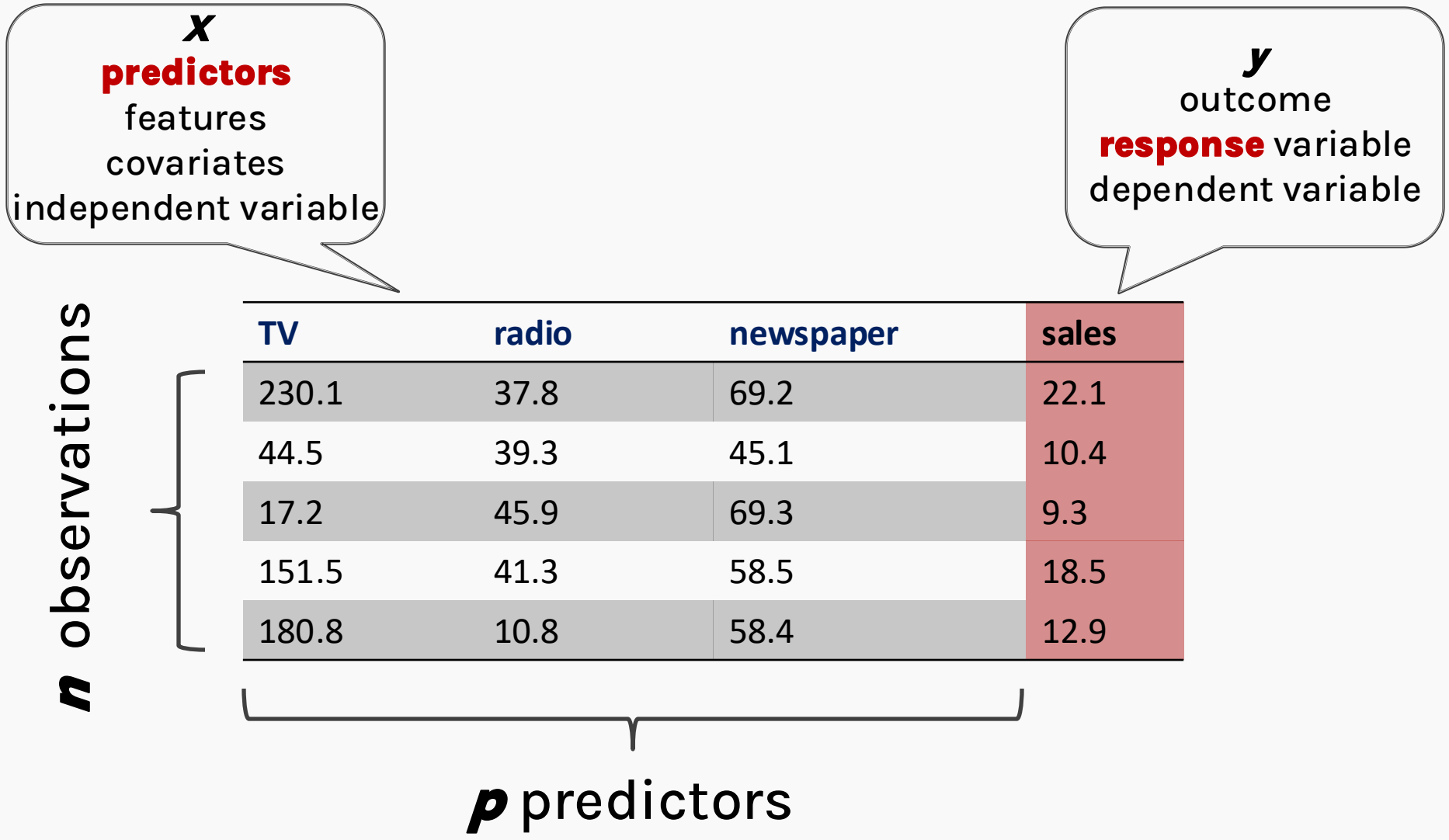
# Response vs. Predictor Variables

Many of these problems exhibit an asymmetry: the variable we aim to predict may be harder to measure, more significant, or directly or indirectly influenced by other variables.

Therefore, we can classify variables into two categories:

- Variables whose values we aim to predict
- Variables used as inputs to inform our prediction

# Response vs. Predictor Variables




| TV | radio | newspaper | sales |
|---|---|---|---|
| 230.1 | 37.8 | 69.2 | 22.1 |
| 44.5 | 39.3 | 45.1 | 10.4 |
| 17.2 | 45.9 | 69.3 | 9.3 |
| 151.5 | 41.3 | 58.5 | 18.5 |
| 180.8 | 10.8 | 58.4 | 12.9 |

# Response vs. Predictor Variables

$X = X_1, \dots, X_p$

$X_j = x_{1j}, \dots, x_{ij}, \dots, x_{nj}$

**predictors**

features

covariates

independent variable

$\boldsymbol{y} = y_1, \dots, y_i, \dots, y_n$

outcome

**response** variable

dependent variable

***n*** observations

| TV | radio | newspaper | sales |
|---|---|---|---|
| 230.1 | 37.8 | 69.2 | 22.1 |
| 44.5 | 39.3 | 45.1 | 10.4 |
| 17.2 | 45.9 | 69.3 | 9.3 |
| 151.5 | 41.3 | 58.5 | 18.5 |
| 180.8 | 10.8 | 58.4 | 12.9 |

***p*** predictors

# Response vs. Predictor Variables

This is called $X$: a.k.a.
***The Design Matrix***

$y$:
The response variable

***n*** observations

| TV | radio | newspaper |
|---|---|---|
| 230.1 | 37.8 | 69.2 |
| 44.5 | 39.3 | 45.1 |
| 17.2 | 45.9 | 69.3 |
| 151.5 | 41.3 | 58.5 |
| 180.8 | 10.8 | 58.4 |

| sales |
|---|
| 22.1 |
| 10.4 |
| 9.3 |
| 18.5 |
| 12.9 |

# Response vs. Predictor Variables

This is called $X$: a.k.a.
***The Design Matrix***

$y$:
The response variable

***n*** observations

| TV | radio | newspaper |
|---|---|---|
| 230.1 | 37.8 | 69.2 |
| 44.5 | 39.3 | 45.1 |
| 17.2 | 45.9 | 69.3 |
| 151.5 | 41.3 | 58.5 |
| 180.8 | 10.8 | 58.4 |

| sales |
|---|
| 22.1 |
| 10.4 |
| 9.3 |
| 18.5 |
| 12.9 |

Capital letters mean matrices,

# Response vs. Predictor Variables

This is called $X$: a.k.a.
***The Design Matrix***

$y$
The response variable

***n*** observations

| TV | radio | newspaper |
|---|---|---|
| 230.1 | 37.8 | 69.2 |
| 44.5 | 39.3 | 45.1 |
| 17.2 | 45.9 | 69.3 |
| 151.5 | 41.3 | 58.5 |
| 180.8 | 10.8 | 58.4 |

| sales |
|---|
| 22.1 |
| 10.4 |
| 9.3 |
| 18.5 |
| 12.9 |

Capital letters mean matrices, lower case letters mean vectors

# Sklearn expects certain dimensions

```
>>> X.shape
(n,p)
```

```
>>> y.shape
(n,) OR (n,1)
```

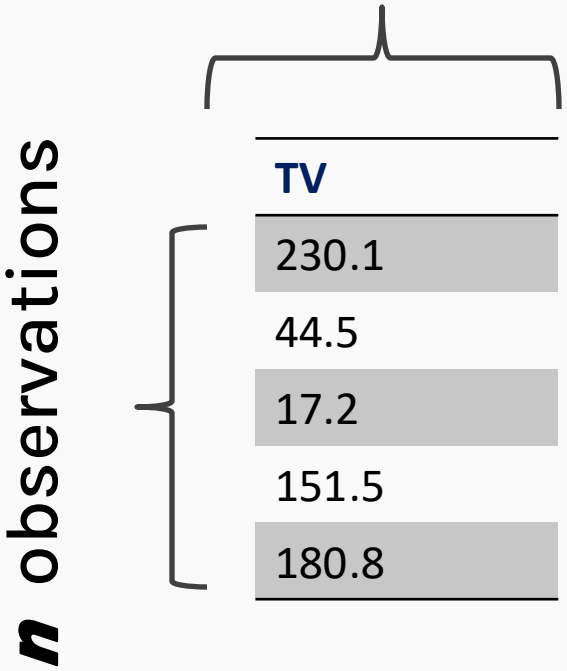
***n*** observations

| TV | radio | newspaper |
|---|---|---|
| 230.1 | 37.8 | 69.2 |
| 44.5 | 39.3 | 45.1 |
| 17.2 | 45.9 | 69.3 |
| 151.5 | 41.3 | 58.5 |
| 180.8 | 10.8 | 58.4 |

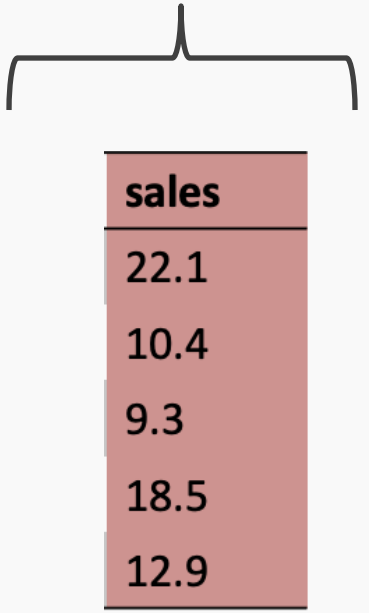
| sales |
|---|
| 22.1 |
| 10.4 |
| 9.3 |
| 18.5 |
| 12.9 |

# Sklearn expects certain dimensions

```
>>> X.shape
(n,) OR (n,1)
```




```
>>> y.shape
(n,) OR (n,1)
```




# CS109A
# GAME Time

# Game time: Pandas Review

df[['x']] vs df['x']

Which of the statements below is correct?

**Options:**

A. df[['x']] returns a pd.Series object whereas df['x'] returns a pd.DataFrame.

B. df[['x']] is invalid operation.

C. df[['x']] returns a pd.DataFrame whereas df['x'] returns a pd.Series object.

D. df['x'] is invalid operation.

# Statistical Model

# True vs. Statistical Model

- Imagine an ice cream cone so perfect that it captures every flavor, topping, and swirl of deliciousness. That is what a ***true model*** is.

# True vs. Statistical Model

- Imagine an ice cream cone so perfect that it captures every flavor, topping, and swirl of deliciousness. That is what a ***true model*** is.

- But reality is like an ice cream shop with infinite flavors and toppings. Trying to fit all of them into one cone is impossible.

# True vs. Statistical Model

- Imagine an ice cream cone so perfect that it captures every flavor, topping, and swirl of deliciousness. That is what a ***true model*** is.

- But reality is like an ice cream shop with infinite flavors and toppings. Trying to fit all of them into one cone is impossible.

- This is why we use ***statistical models***: instead of trying to scoop the impossible sundae, we craft a tasty treat from the flavors we have.

# True vs. Statistical Model

We assume that the response variable, $Y$, is related to the predictor variables, $X$, through an unknown function which can be generally expressed as:

$$Y = f(X) + \varepsilon$$

Here, $f$ represents the unknown function expressing an underlying rule for relating $Y$ to $X$. $\varepsilon$ represents the random amount (unrelated to $X$) that $Y$ differs from the rule $f(X)$.

A ***statistical model*** is any algorithm used to estimate $f$. We denote the estimated function as $\widehat{f}$.

# Prediction vs. Estimation

## Inference Problems:

- The primary focus is on obtaining $\hat{f}$, which is an estimate of the true function $f$
- Objective: Understand the form and characteristics of $\hat{f}$.

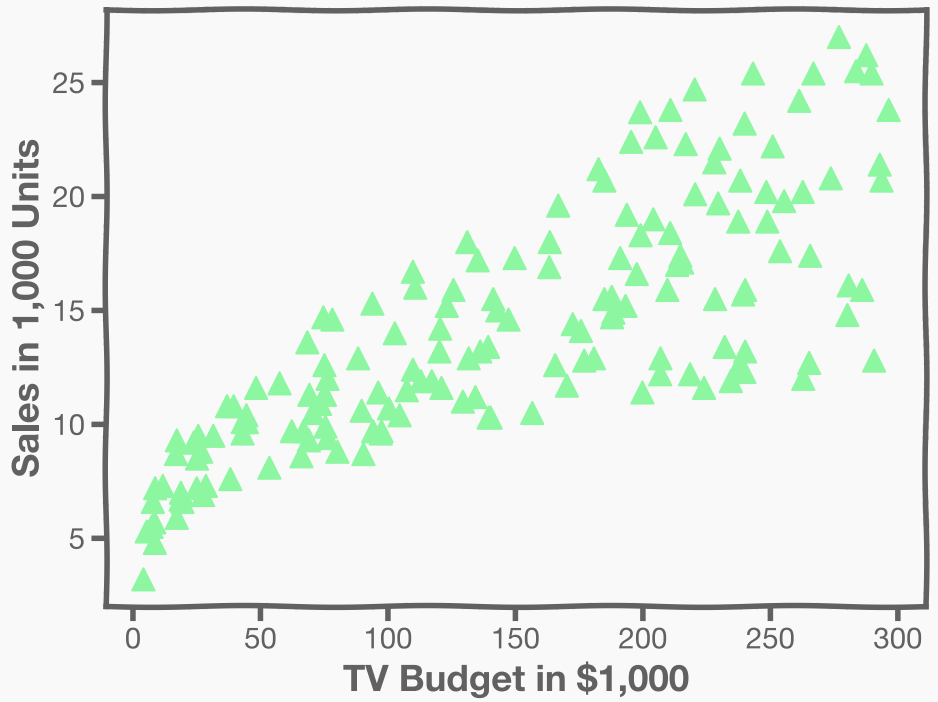
## Prediction Problems

- The specific form of $\hat{f}$ is less important than the accuracy of the predictions.
- Objective: Minimize the difference between predicted values $\hat{y}$ and observed values $y$.

# Example: predicting sales

**Motivation**: Predict Sales

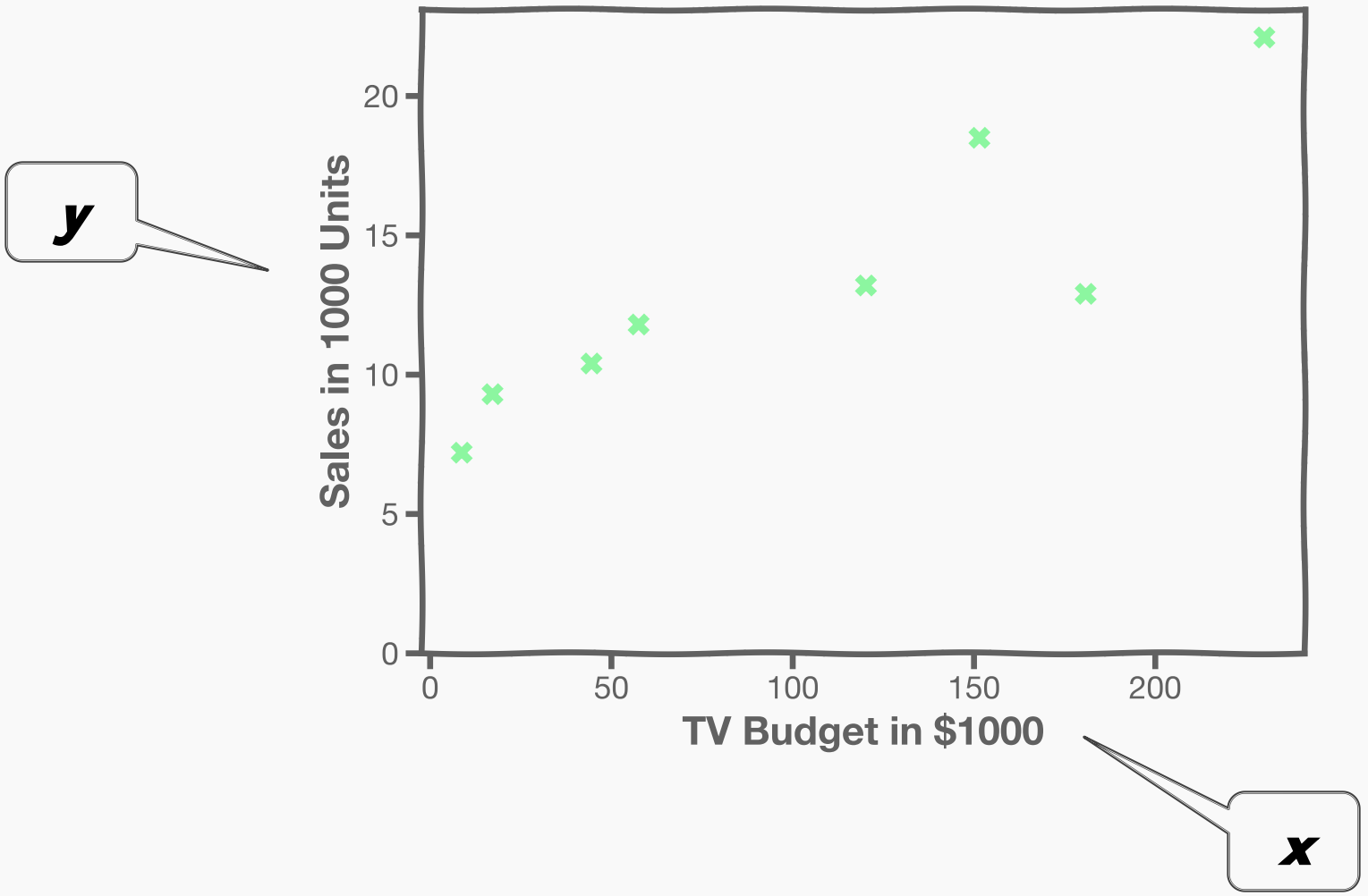
Build a model to **predict** sales based on TV budget




The response, ***y***, is the sales.

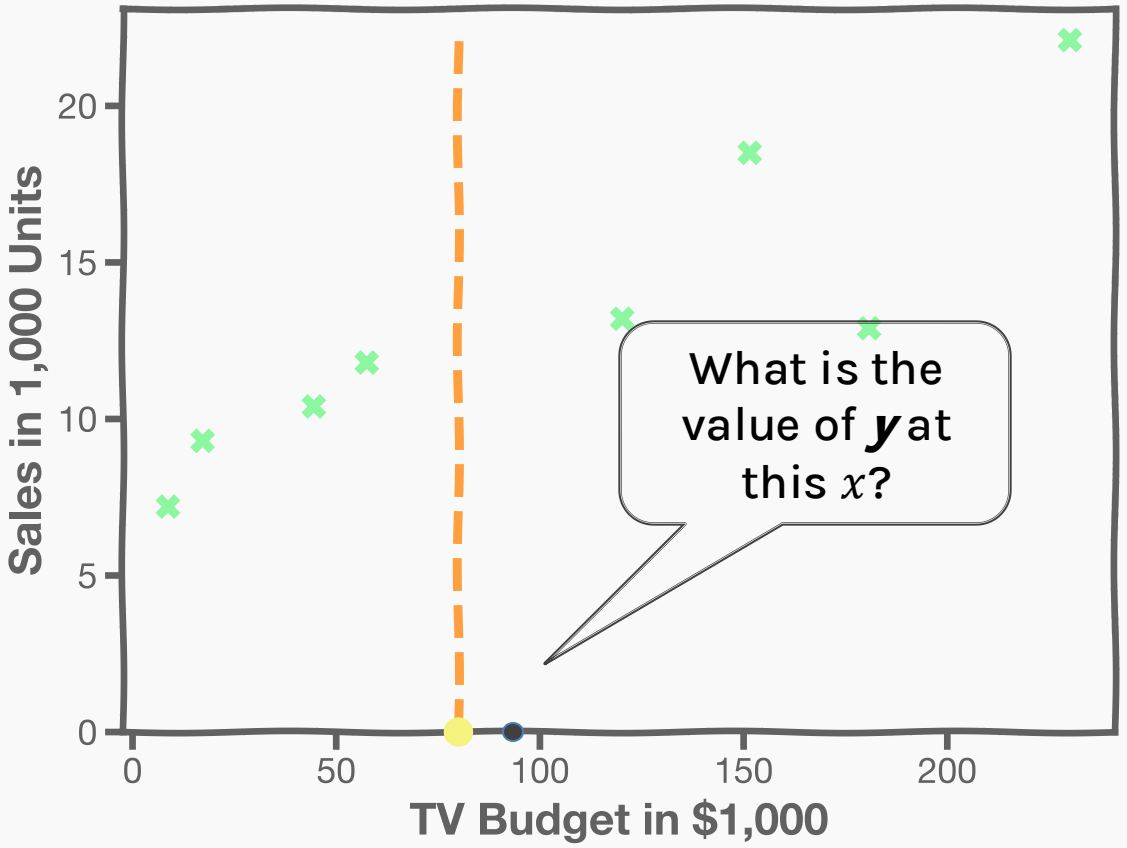
The predictor, ***x***, is TV budget.

# Statistical Model



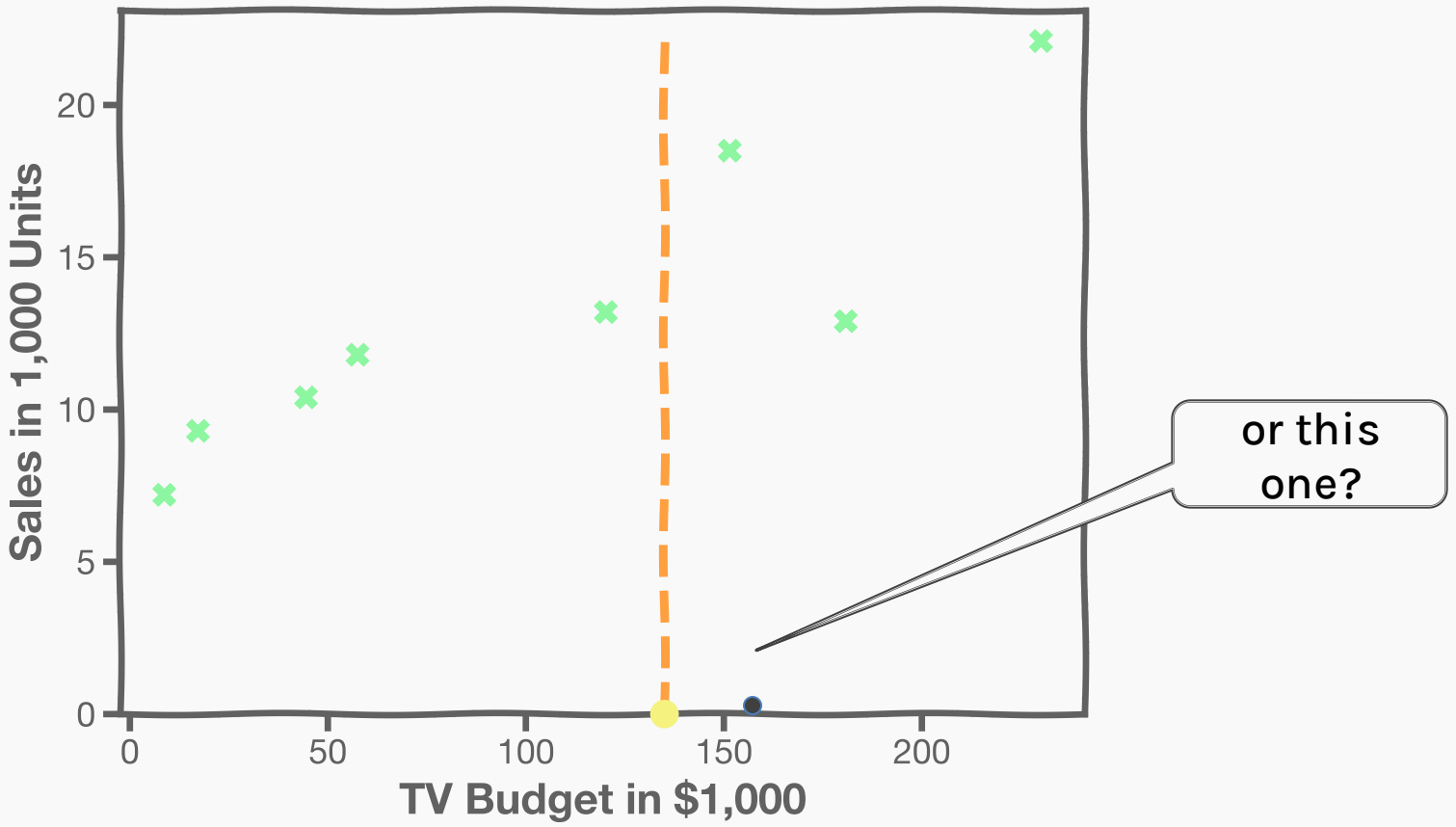
# Statistical Model

How do we predict $y$ for some $x$?



# Statistical Model

How do we predict $y$ for some $x$?



# Game time: Choices for model

Which of the following 5 methods could be used to predict the value of y given x?

**Options:**

A. Utilize a Convolutional Neural Network (CNN).

B. Use a Linear Regression Model with a slope of 3 and an intercept of 2.

C. Identify examples that closely resemble the input data point.

D. Consult a TF during office hours for the answer.

E. Calculate the average value of y from the available data points.

# Game time: Choices for model

Which of the following 5 methods could be used to predict the value of y given x?

**Options:**

A. Utilize a Convolutional Neural Network (CNN).

B. Use a Linear Regression Model with a slope of 3 and an intercept of 2.

C. Identify examples that closely resemble the input data point.

D. Consult a TF during office hours for the answer.

E. Calculate the average value of y from the available data points.

# Game time: Choices for model

Which of the following 5 methods could be used to predict the value of y given x?

**Options:**

A. Utilize a Convolutional Neural Network (CNN).

B. Use a Linear Regression Model with a slope of 3 and an intercept of 2.

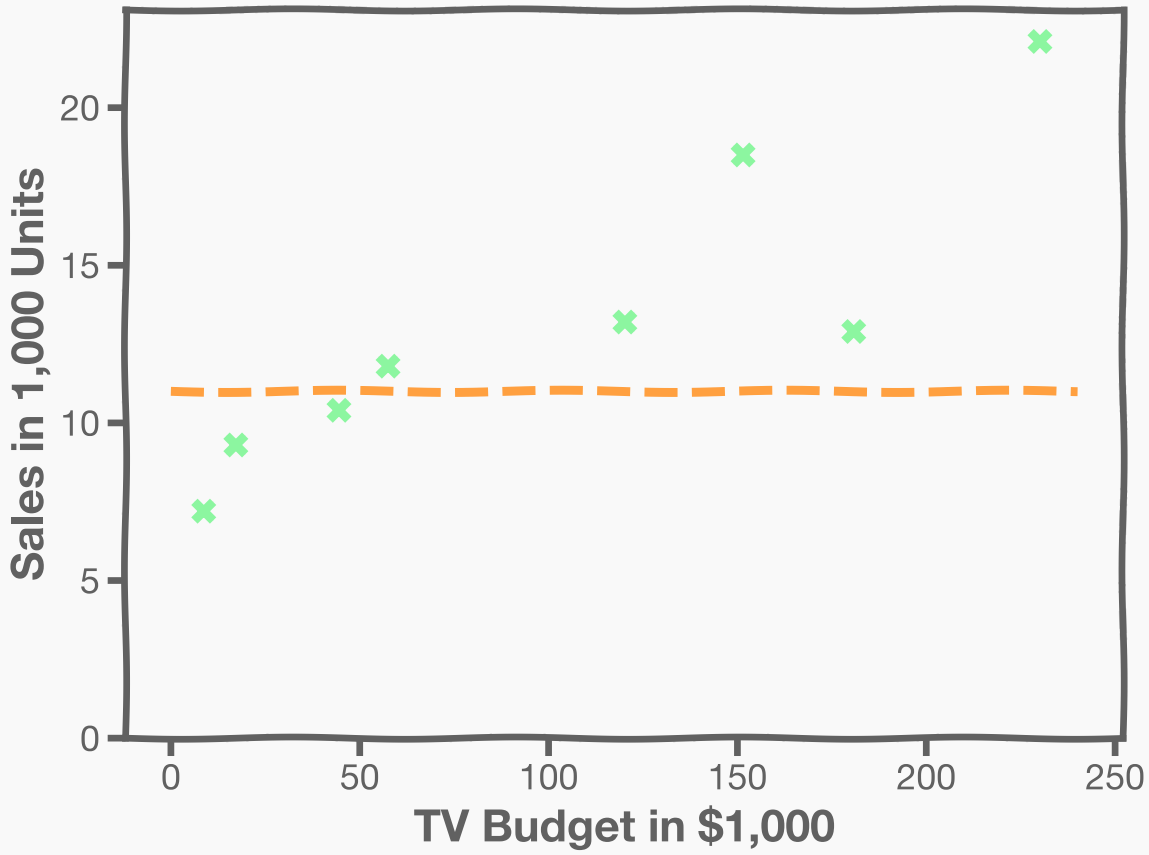
C. Identify examples that closely resemble the input data point.

D. Consult a TF during office hours for the answer.

E. Calculate the average value of y from the available data points.

# Statistical Model

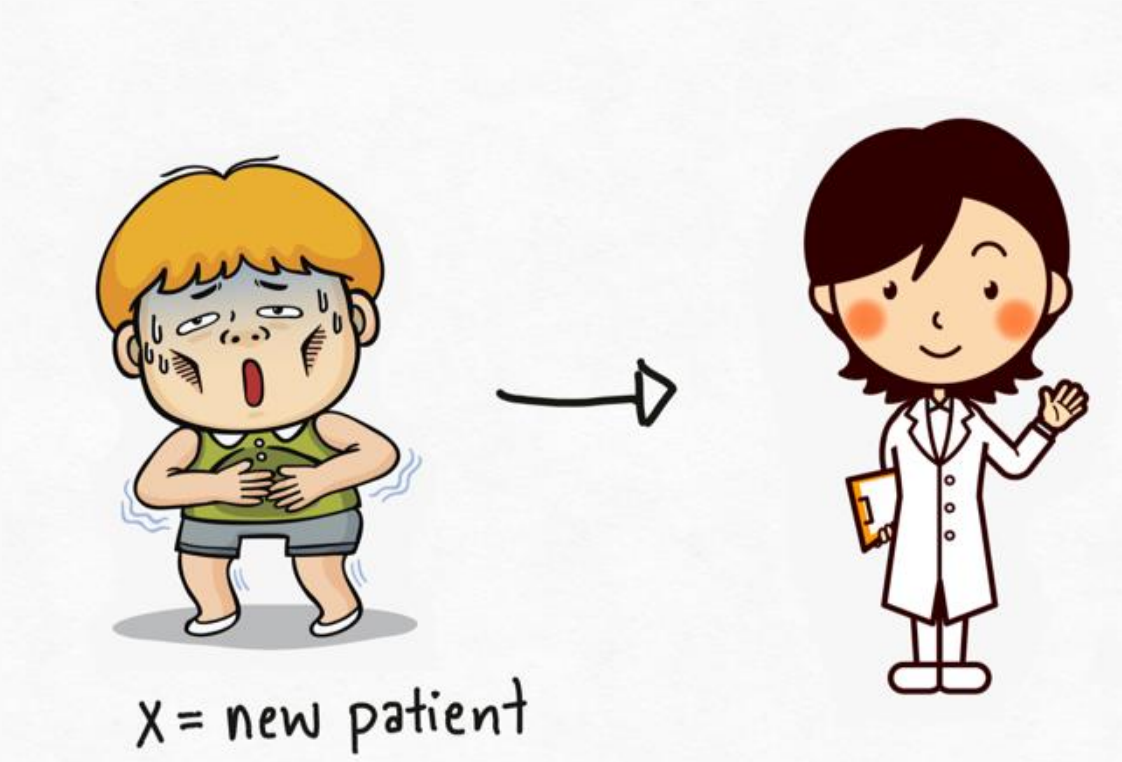
A simple idea is to take the mean of all $y$'s: $\frac{1}{n}\sum_{i=1}^{n} y_i$



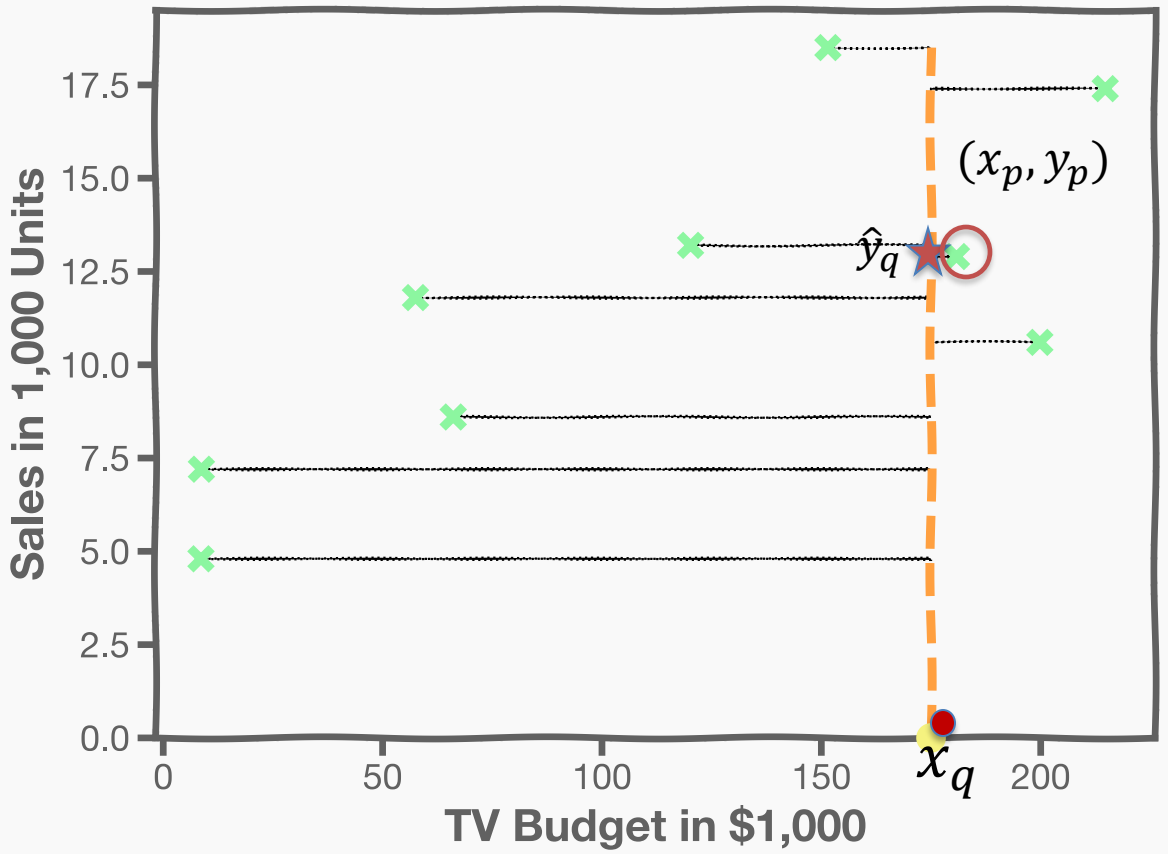

$$\frac{1}{n}\sum_{i=1}^{n} y_i$$

# K-Nearest Neighbors

# k-Nearest Neighbors – kNN



# k-Nearest Neighbors – kNN




What is $\hat{y}_q$ at some $x_q$ ?

Find distances to all other points $D(x_q, x_i)$

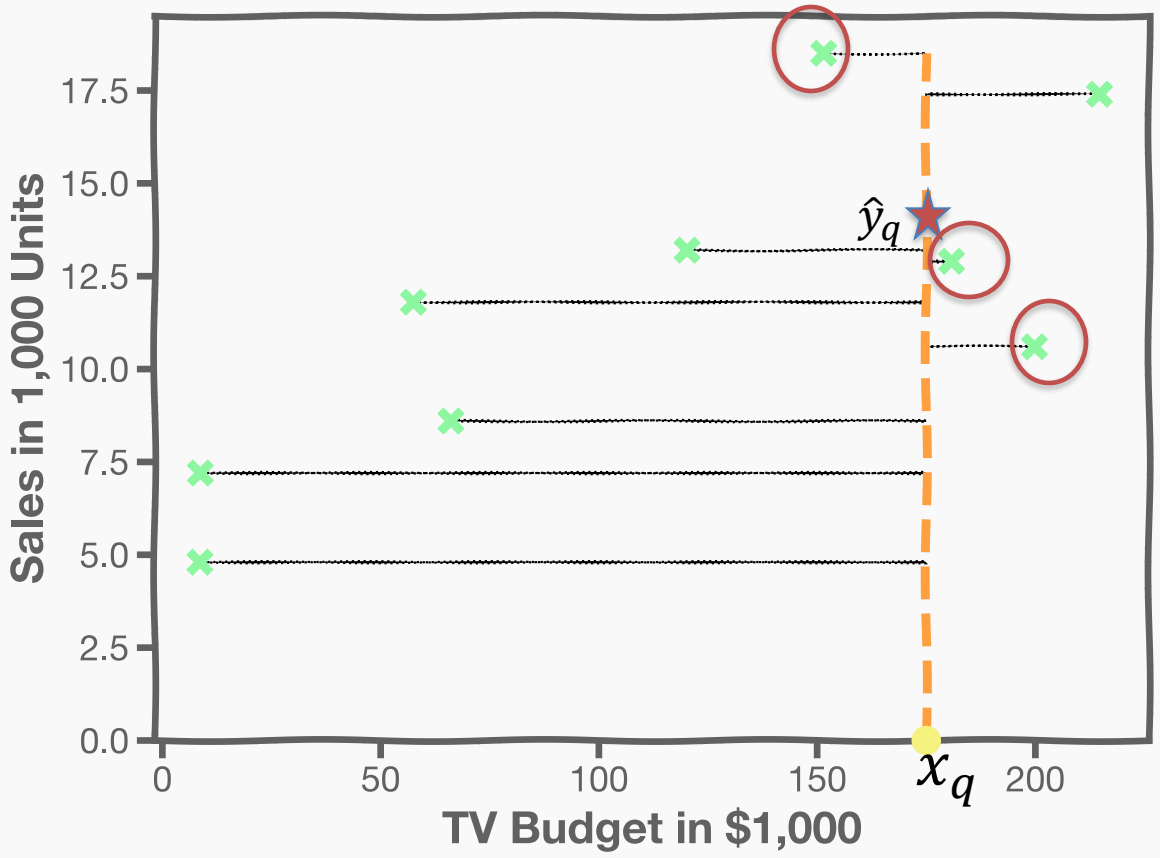
Find the nearest neighbor, $(x_p, y_p)$

Predict $\hat{y}_q = y_p$

# k-Nearest Neighbors – kNN

Do the same for "all" $x's$

# k-Nearest Neighbors – kNN



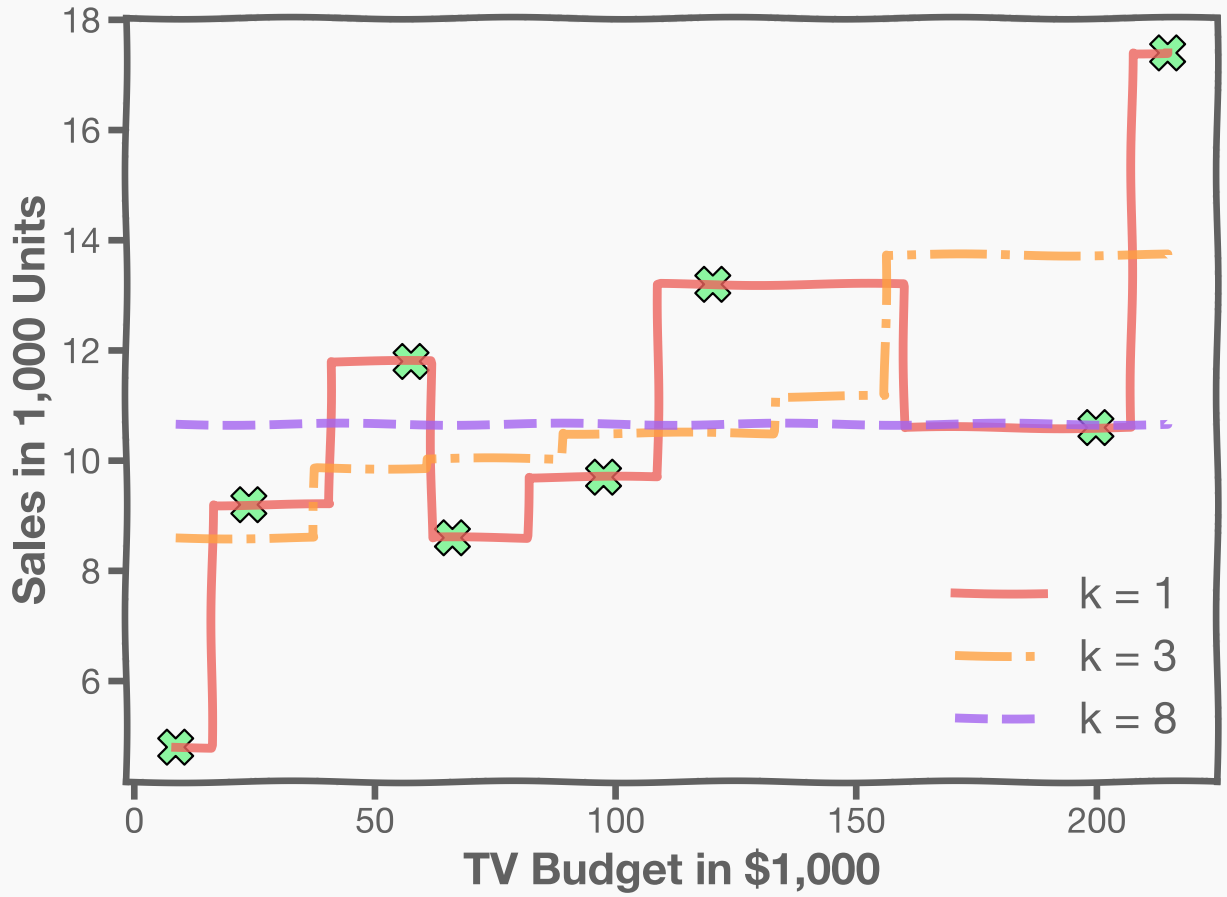

What is $\hat{y}_q$ at some $x_q$ ?

Find distances to all other points $D(x_q, x_i)$

Find the k-nearest neighbors, $x_{q_1}, \ldots, x_{q_k}$

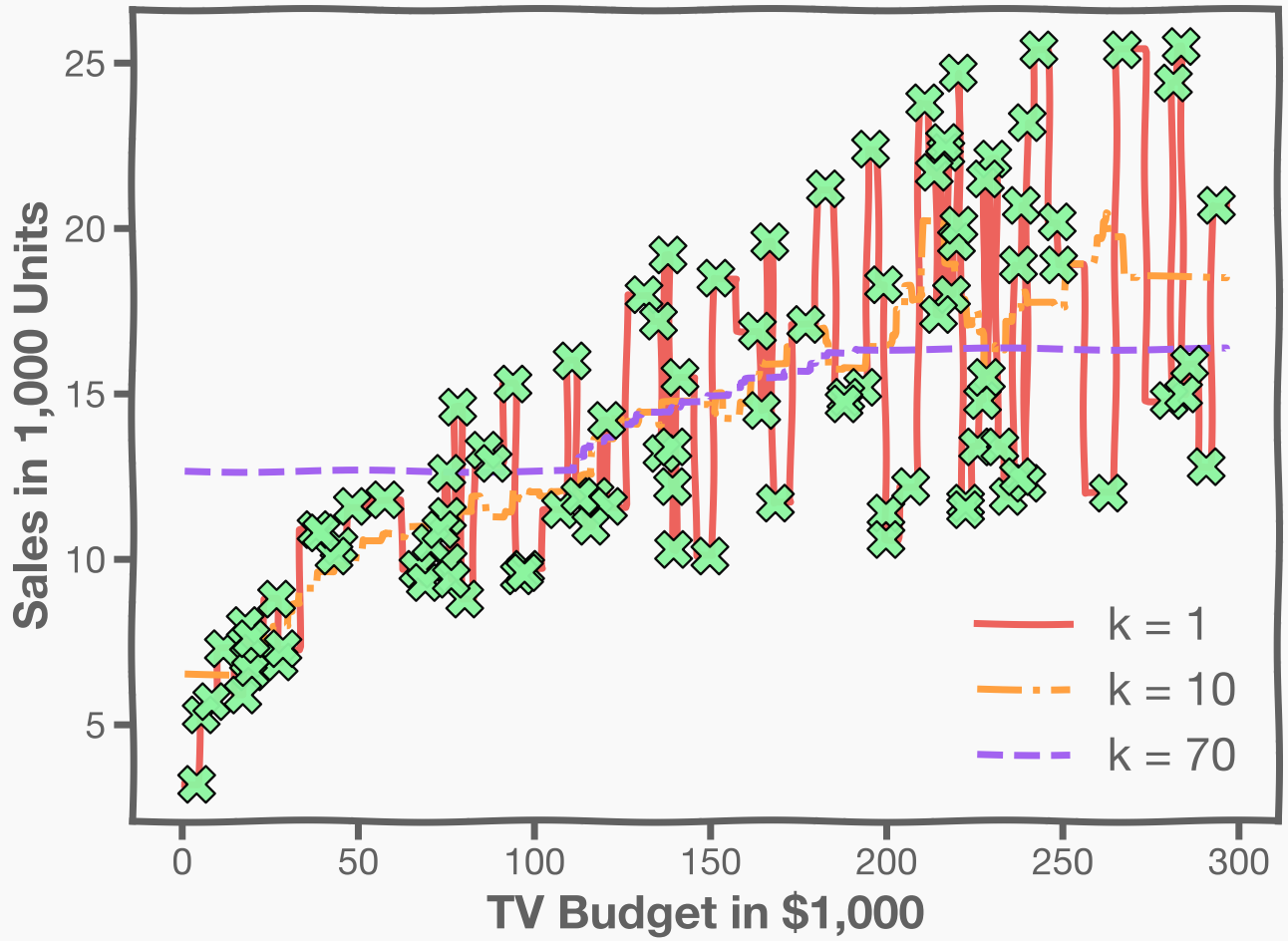
Predict $\hat{y}_q = \frac{1}{k}\sum_i^k y_{q_i}$

# k-Nearest Neighbors – kNN



# k-Nearest Neighbors – kNN

We can try different k-models on more data



# k-Nearest Neighbors – kNN

The very human way of decision making by similar examples. kNN is a **non-parametric** learning algorithm.

The k-Nearest Neighbor Algorithm:

Given a dataset $D = \{(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})\}$, for every new $X$:

1. Find the k-number of observations in $D$ most similar to $X$:

$$\{(x^{(n_1)}, y^{(n_1)}), \dots, (x^{(n_k)}, y^{(n_k)})\}$$

   These are called the k-nearest neighbors of $x$

2. Average the output of the k-nearest neighbors of $x$

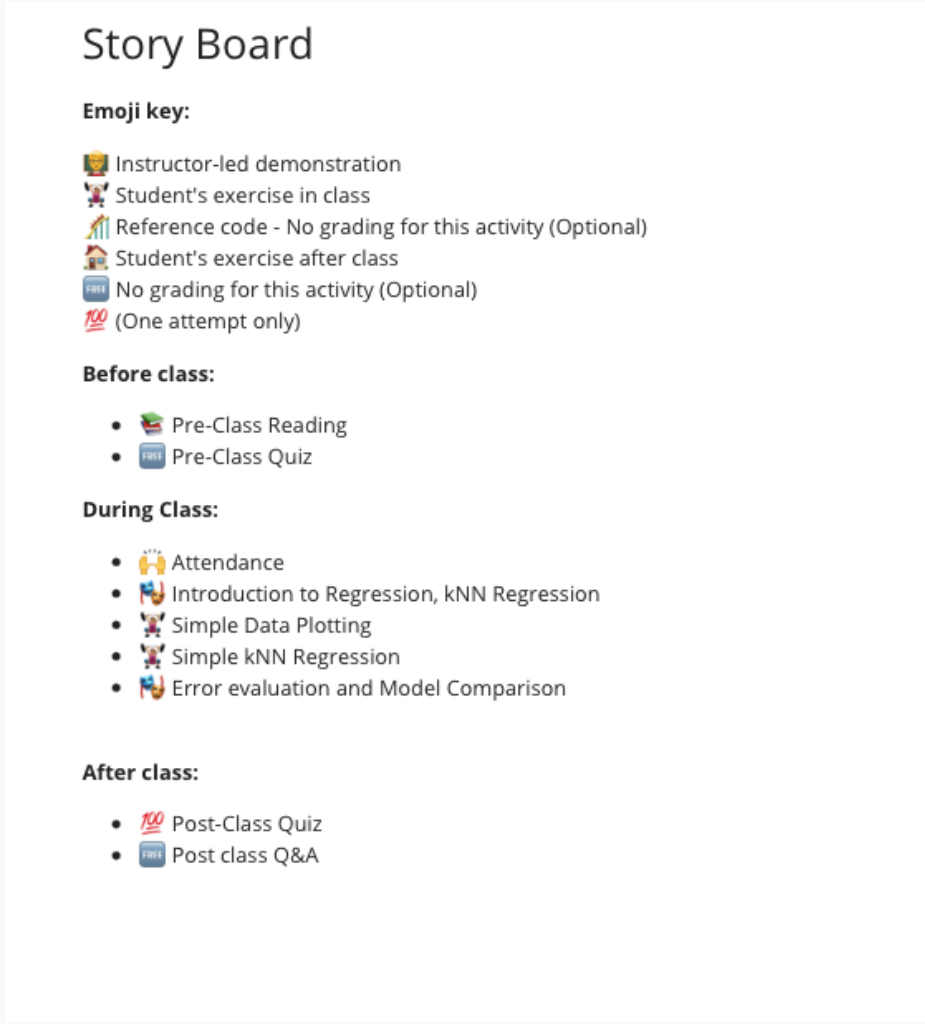
$$\hat{y} = \frac{1}{K}\sum_{k=1}^{K} y^{(n_k)}$$

# Exercises + Quizzes Review

- **Pre-class quiz**: Not graded (see 🆓 emoji).
- **Post-class Q&A:** Not graded (see 🆓 emoji).
- **Attendance**: For every **8 sessions attended**, you earn an additional late day.

***Note: This does not apply to Extension School students.***

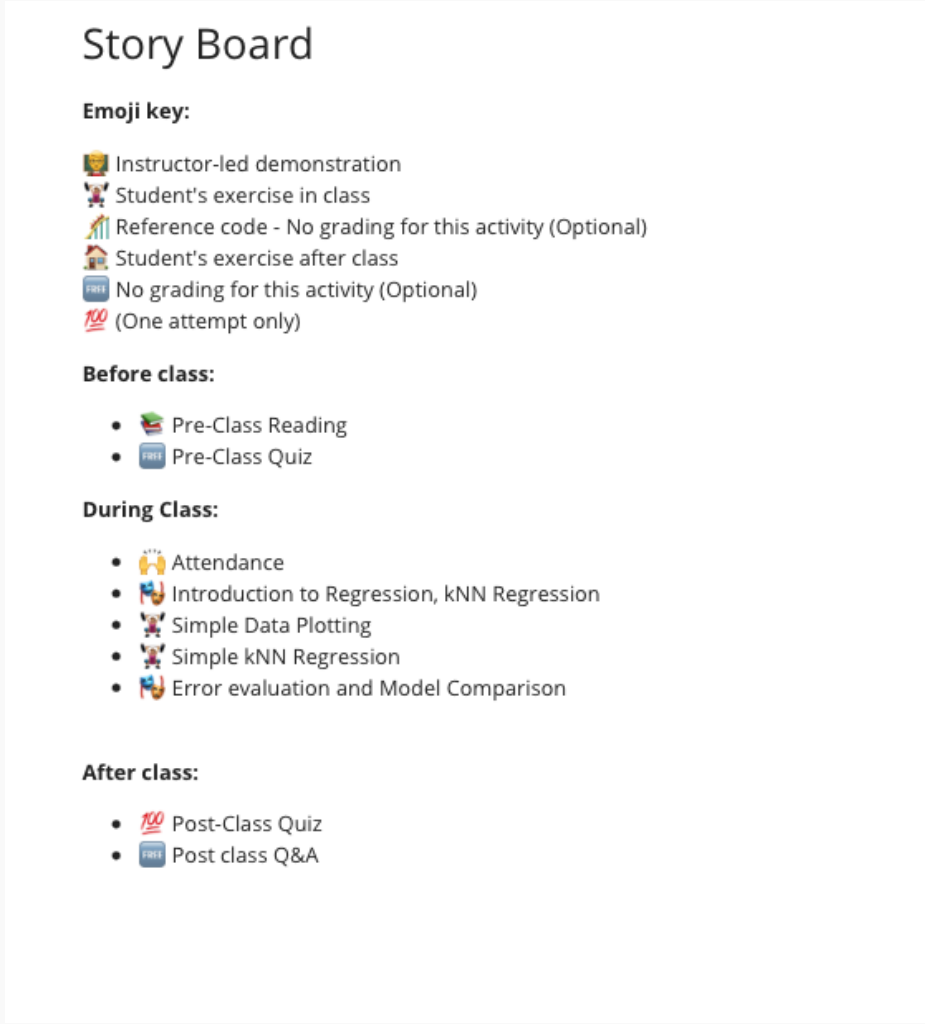
- **Quizzes**: Must be completed before the next lecture. The lowest 1/3 of quiz grades (including missed ones) will be dropped.



# Exercises + Quizzes Review

- **Exercises**: Sometimes led by the instructor 👩‍🏫, sometimes done in class 🏋️ or at home 🏠. Regardless, they are **due by the beginning of the next lecture**.

***Note: Exercises are only counted in the final grade if they improve your grade. Otherwise, their weight is shifted to quizzes (from 8% to 10%).***



Story Board

**Emoji key:**

👩‍🏫 Instructor-led demonstration
🏋️ Student's exercise in class
📈 Reference code - No grading for this activity (Optional)
🏠 Student's exercise after class
🆓 No grading for this activity (Optional)
💯 (One attempt only)

**Before class:**

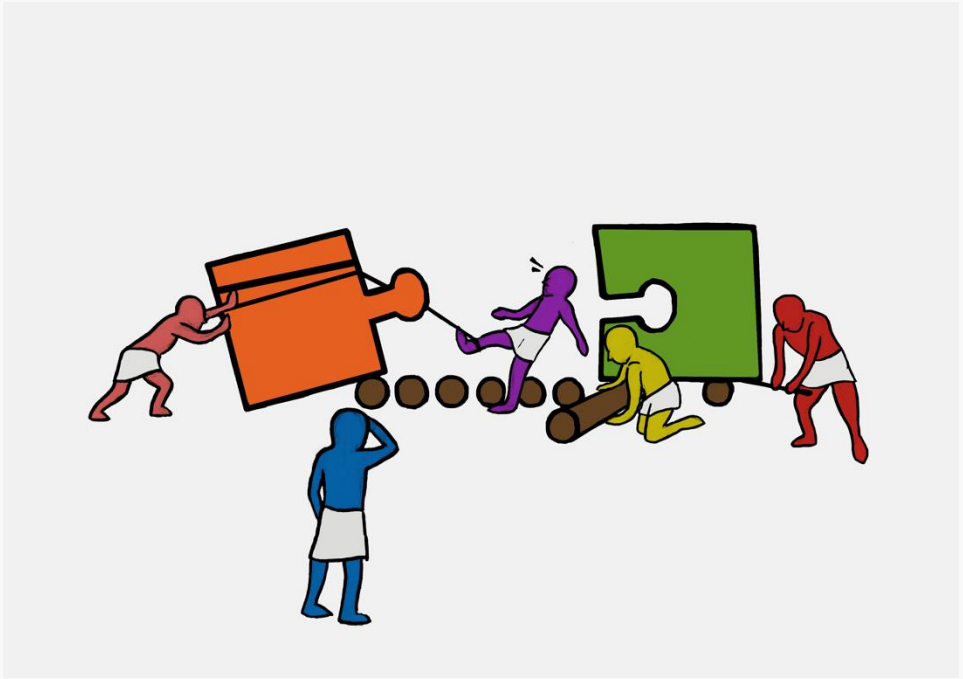
- 📚 Pre-Class Reading
- 🆓 Pre-Class Quiz

**During Class:**

- 🙌 Attendance
- 🤹 Introduction to Regression, kNN Regression
- 🏋️ Simple Data Plotting
- 🏋️ Simple kNN Regression
- 🤹 Error evaluation and Model Comparison

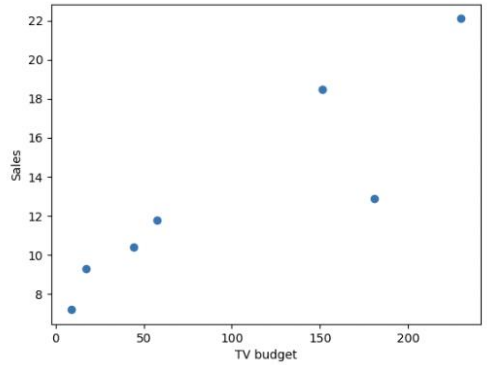
**After class:**

- 💯 Post-Class Quiz
- 🆓 Post class Q&A



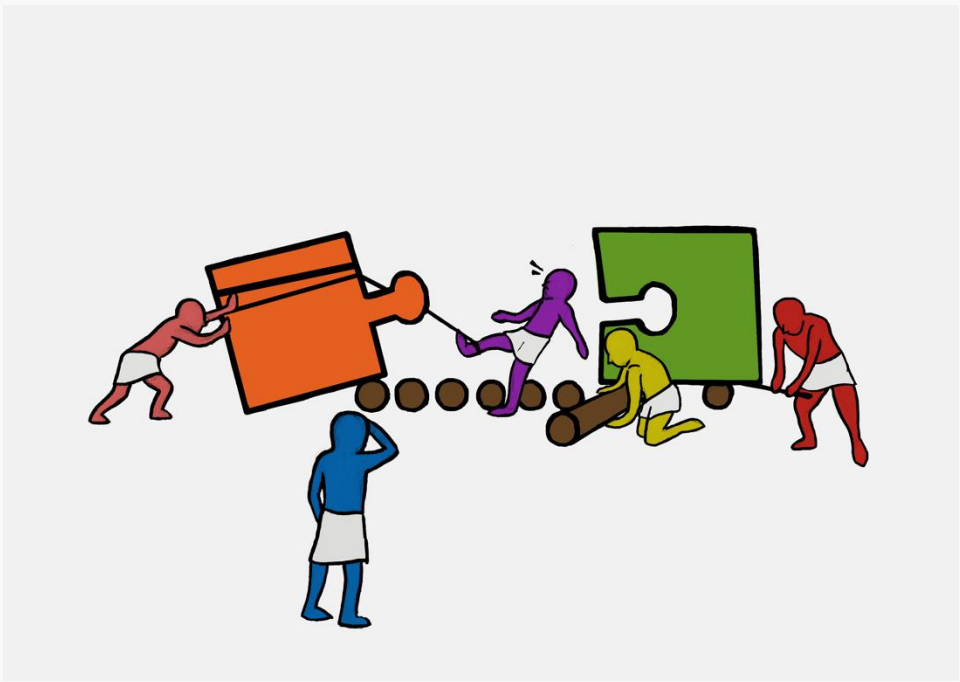
# 👩‍🏫 Exercise: A.1 - Simple Data Plotting

The aim of this exercise is to **plot** TV Ads vs Sales based on the Advertisement dataset which should look similar to the graph given below.




## Instructions:

- Read the Advertisement data and view the top rows of the dataframe to get an understanding of the data and the columns.
- Select the first 7 observations and the columns `TV` and `sales` to make a new data frame.



## 🏋️‍♀️ Exercise: A.2 - Simple kNN Regression

The goal of this exercise is to **re-create the plots** given below. You would have come across these graphs in the lecture as well.