CS M148 –

Data Science Fundamentals

Lecture #3: kNN & Linear regression

Baharan Mirzasoleiman

UCLA Computer Science

Announcements

Data Science minor

- The applications can be accepted to the minor starting from Spring
- The information will be posted online this quarter

Waitlist

- We are asking the department for more reader hours, and will admit the waitlist (and a few PTEs) if we get more reader hours
- I will update you on this on Wednesday

Announcements

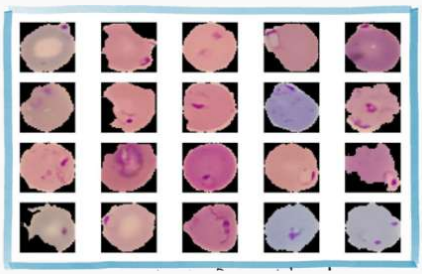
Exam

- To accommodate the current situation, we will not have the in-class midterm exam and will instead have 4 homework assignments
- Your top 3 assignments will be taken into consideration
- We will not accept extensions requests
- Please check the updated syllabus in Canvas

Let’s quickly review what we’ve seen so far

The Potential of Data Science

Disease Diagnosis

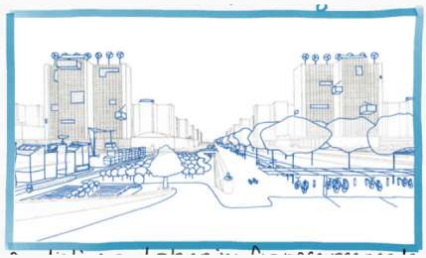


Detecting malaria from blood smears

Drug Discovery

Quickly discovering new drugs for COVID

Urban Planning



Predicting and planning for resource needs

Agriculture

Precision agriculture

The Potential of Data Science

Gender Bias

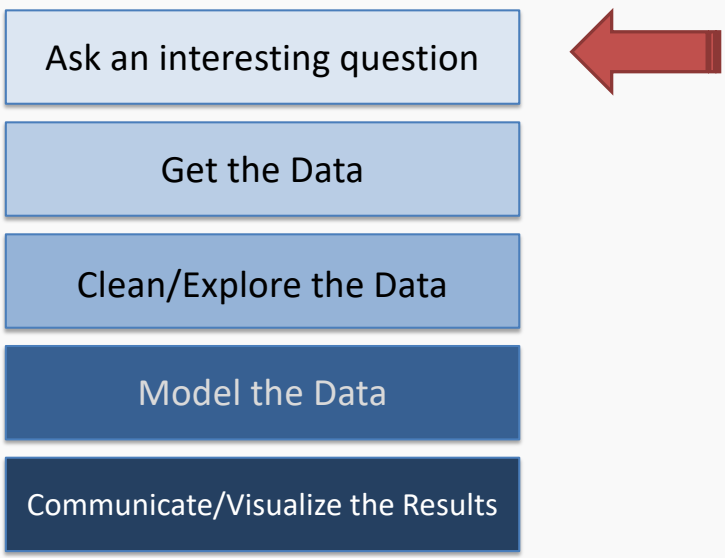
Some DS models for evaluating job applications show bias in favor of male candidate

Racial Bias

Risk models used in US courts have shown to be biased against non-white defendants

What is Data Science?

The Data Science Process

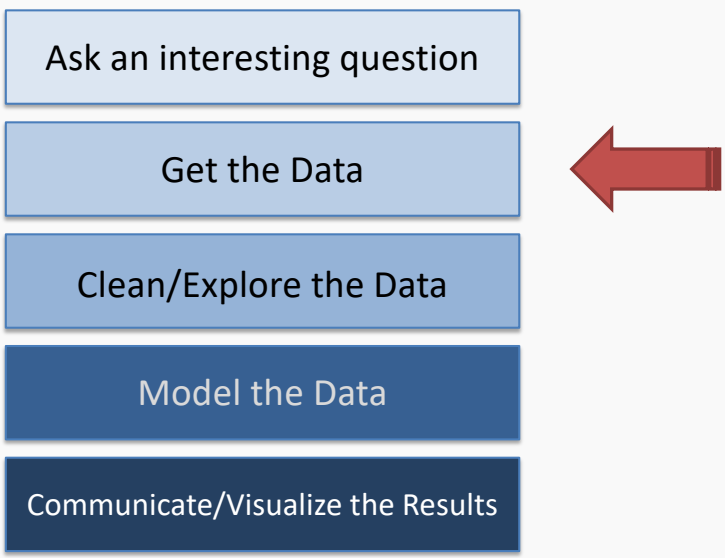


Problem Statement

Which club will win the EPL?

What?

The Data Science Process

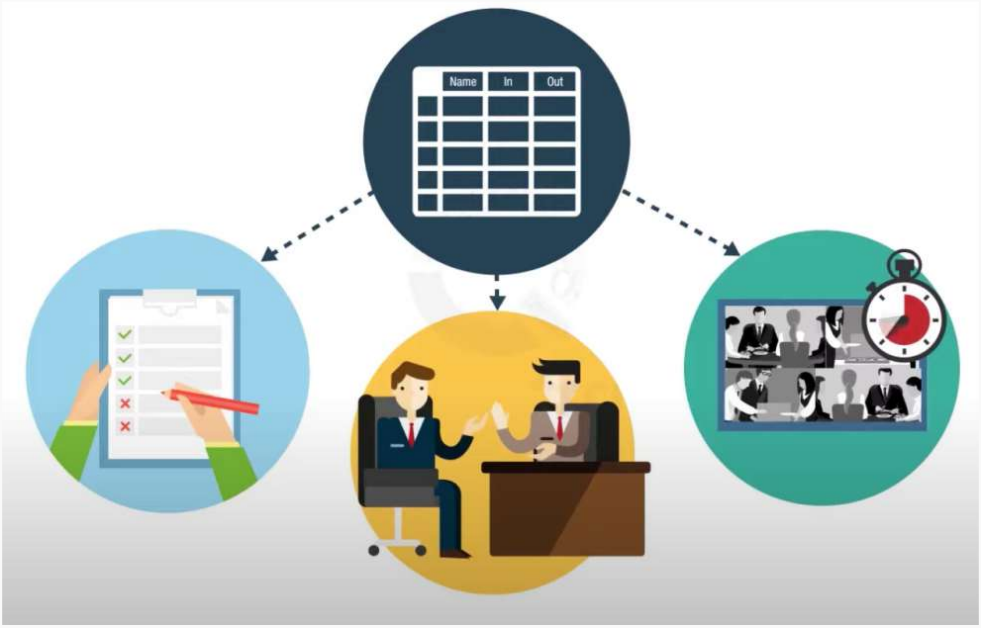


Data Collection

- Primary: When you have a unique problem and no related research is done on the subject.
- Secondary: use the data which is readily available or collected by someone else

Primary Data Collection

Surveys, interviews, observations, etc.



Secondary Data Collection

Can be found on open-source websites such as Kaggle, Gapminder, news articles, government census, magazines, etc.

Data Collection Procedure

Steps you will take to gather data that is consistent, accurate, and unbiased

Consider these questions:

- How will you define and measure your variables?
- How will you ensure your measurements are reliable and valid?
- How will you select and contact your sample?

How will you chose your participants?

Population: the entire group that you want to make conclusion about

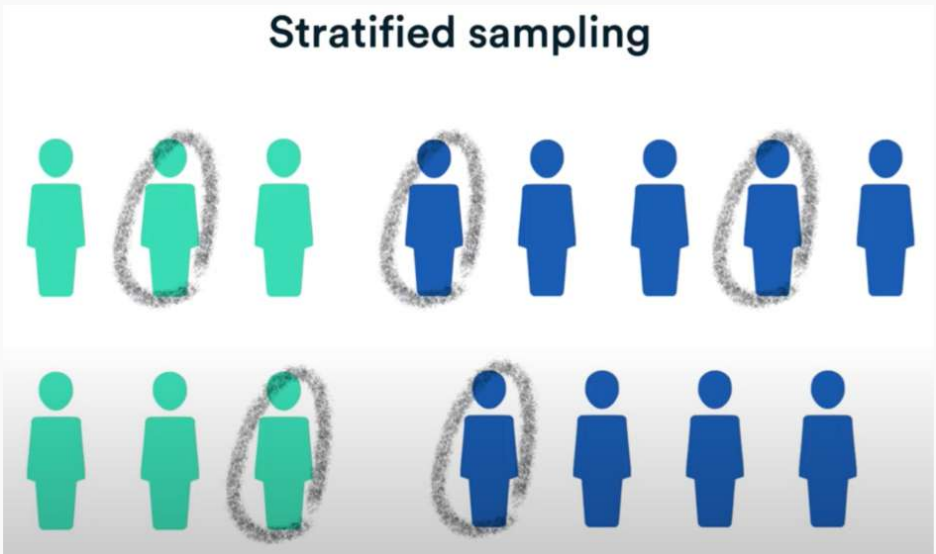
Sample: smaller group of individuals you'll collect the data from

Probability Sampling methods

Simple random sampling: Select a sample completely at random from the whole population

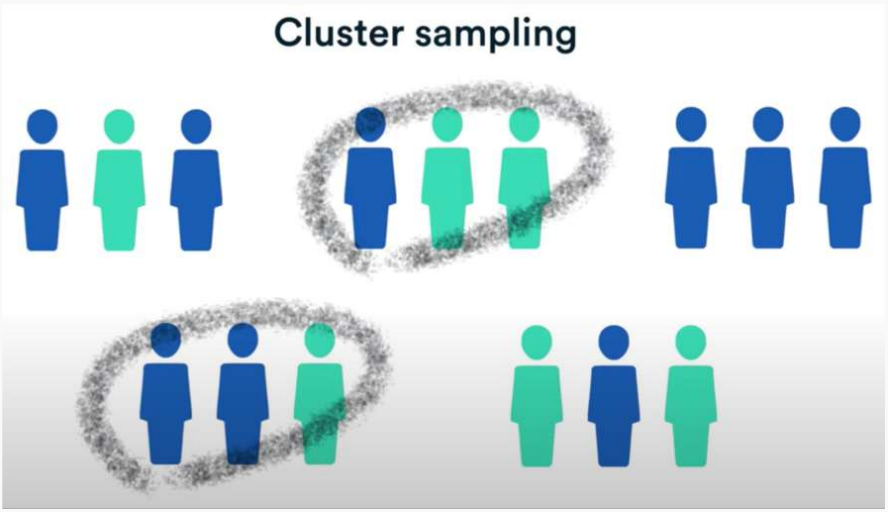
Probability Sampling methods

Stratified sampling: divide the population into subgroups, and draw a random sample from each subgroup



Probability Sampling methods

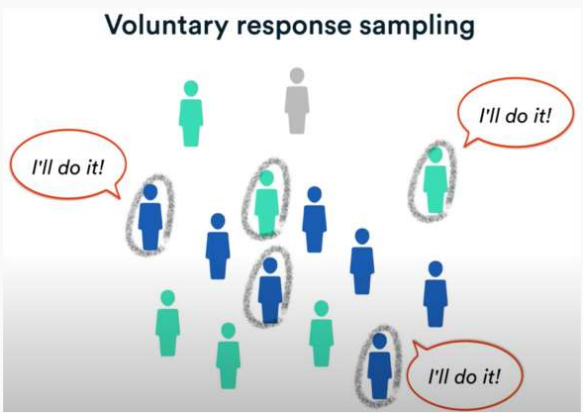
Cluster sampling: divide the population into clusters (e.g. geographical areas), and randomly select some of these cluster for your sample



Non-probability sampling

Non-probability samples are much easier to achieve, but they have more risk of bias

- If you chose a sample based on the most convenient and accessible member of the population, or
- If you rely on volunteers for your study



Data Collection Bias

For practical reasons, many studies rely on convenience samples

- It's important to be aware of the limitations and carefully consider potential biases!
- Always make an effort to gather a sample that's as representative as possible of the population

Example: Amazon decided to shut down its AI recruiting tool after discovering it discriminated against women.

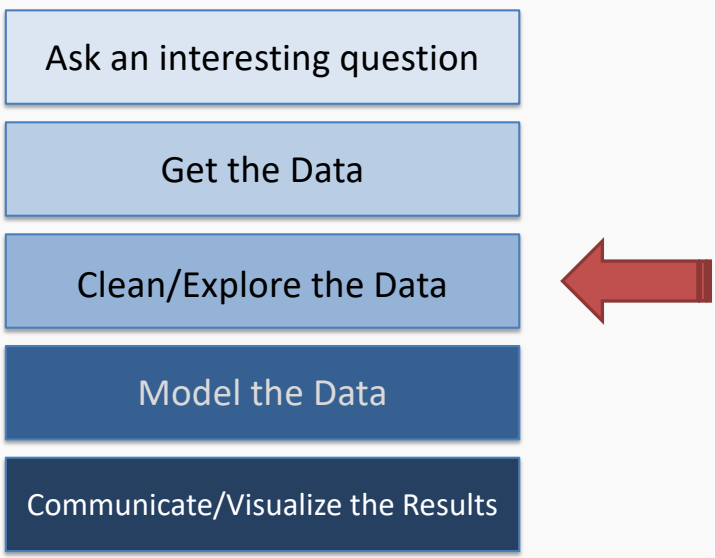
- What was the reason for the bias?
  - not as many women in the data

Selection Bias

- Voluntary bias
- Under-coverage bias
- Non-response bias
- Convenience bias
- Response bias
- Over-coverage bias

What?

The Data Science Process



Clean/explore the Data

Which club will win the EPL?

Always Sanity Check First

If you start the analysis without ensuring data quality then you might get unexpected results such as the Crystal Palace club will win the next EPL

Player Name	Age	Club	Height	Weight	Foot	Joined
Pierre-Emerick Aubameyang	29	Arsenal	6'2"	176lbs	Right	Jan 31, 2018
Alexandre Lacazette	27	Arsenal	5'9"	161lbs	Right	Jul 5, 2017
Bernd Leno		Arsenal	6'3"	183lbs	Right	Jul 1, 2018
Henrikh Mkhitaryan	29	Arsenal	5'10"	165lbs	Right	Jan 22, 2018
Granit Xhaka	25	Arsenal	6'1"	181lbs	Left	Jul 1, 2016
Shkodran Mustafi	26	Arsenal	6'0"	181lbs	Right	Aug 30, 2016
Jack Grealish	22	Aston Villa	5'9"	150lbs	Right	Mar 1, 2012
John McGinn	23	Ast	5'10"	150lbs	Left	Aug 8, 2018
Anwar El Ghazi	23		'2"	550lbs	Right	Jan 31, 2017
Conor Hourihane	27			137lbs	Left	Jan 26, 2017
James Chester	29			174lbs		Aug 12, 2016
James Chester	29			174lbs		Aug 12, 2016
James Chester	29			174lbs		Aug 12, 2016
James Chester	29			174lbs		Aug 12, 2016
Jonathan Kodjia	2	Aston Villa		170lbs	Right	Aug 30, 2016
Callum Wilson	26	Crystal Palace	5'11	146lbs	Right	Jul 4, 2014

Bad quality data

Factors Causing Data Quality Issue

- Improper data collection

Company	Employee Name	Age	Time Spent (hours)
Apple	John S.	23	100
Apple	Evan B.	27	8
Apple	Emily B.	31	12
Google	Ava W.		7
Google	Noah A.	34	9

Incorrect measurement

Incorrect time

Incomplete data

Factors Causing Data Quality Issue

- Improper data integration

Player Name	Team	Weight (lbs.)
P. Bardsley	Chelsea	150
D. McNeil	Chelsea	198
Adam Legzdins	Chelsea	170
Dan Agyei	Chelsea	168
David Luiz	Chelsea	192

Source: X (in lbs.)

Player Name	Team	Weight (kgs.)
Jamal Blackman	Chelsea	72
Ethan Ampadu	Chelsea	68
Billy Gilmour	Chelsea	73
Ike Ugbo	Chelsea	64.5
George McEachran	Chelsea	75

Source: Y (in kgs.)

Data Quality Issues

Some issues are difficult to spot. For example, can you spot what is wrong in this data set? If you follow EPL, then there is no club with the name of Real Madrid in EPL

Player Name	Age	Club	Height	Weight	Foot	Joined
Eden Hazard	27	Chelsea	5'6"	159lbs	Right	Jul 16, 2016
N'Golo Kanté	28	Chelsea	5'10"	168lbs	Right	Aug 24, 2012
César Azpilicueta	23	Chelsea	6'1"	187lbs	Right	Aug 8, 2018
Kepa Arrizabalaga	29	Chelsea	5'9"	172lbs	Right	Aug 28, 2013
Willian	31	Chelsea	6'2"	190lbs	Right	Aug 31, 2016
David Luiz	27	Chelsea	6'2"	192lbs	Left	Aug 31, 2016
Ferland Mendy	23	Real Madrid	5'9"	161lbs	Left	Jun 8, 1995

Requires domain knowledge

Domain Knowledge

As a data scientist, you should develop a good understanding of the domain, and the problem you are solving.

Data Quality Issues

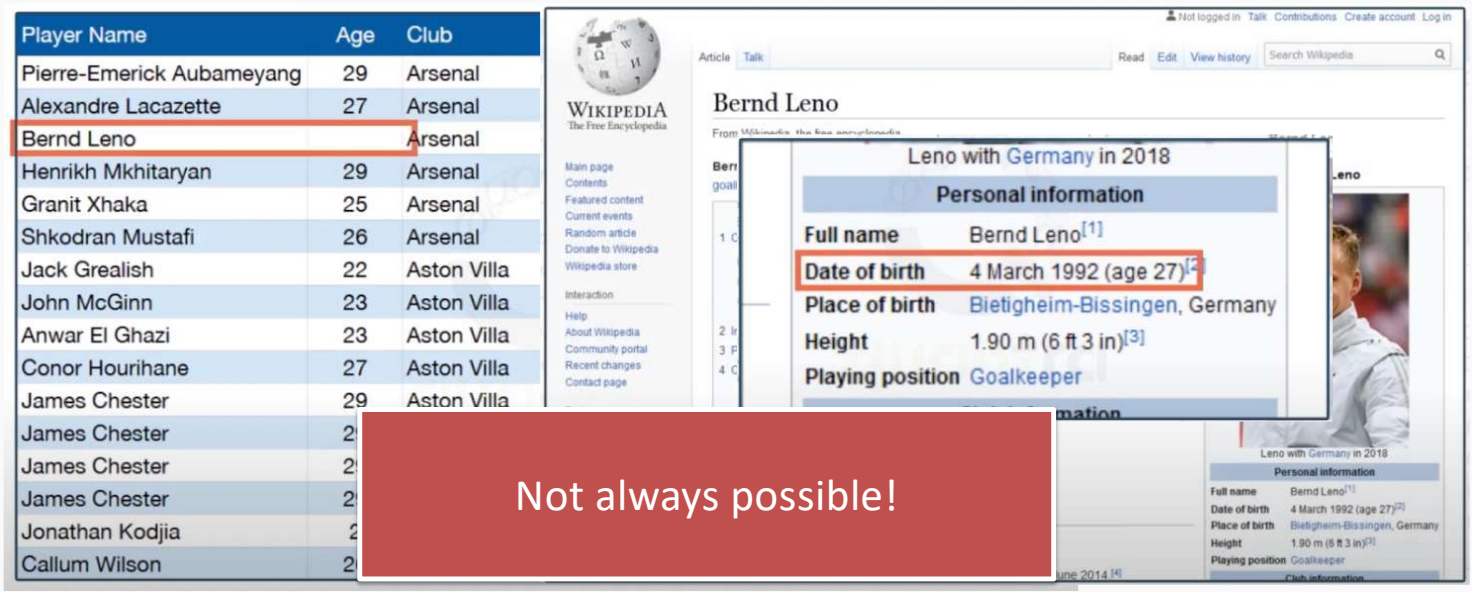
The common data quality issues that are easy to spot are missing values, duplicate values, and inconsistent data.

Player Name	Age	Club	Height	Weight	Foot	Joined
Pierre-Emerick Aubameyang	29	Arsenal	6'2"	176lbs	Right	Jan 31, 2018
Alexandre Lacazette	27	Arsenal	5'9"	161lbs	Right	Jul 5, 2017
Bernd Leno		Arsenal	6'3"	183lbs	Right	Jul 1, 2018
Henrikh Mkhitaryan	29	Arsenal	5'10"	165lbs	Right	Jan 22, 2018
Granit Xhaka	25	Arsenal	6'1"	181lbs	Left	Jul 1, 2016
Shkodran Mustafi	26	Arsenal	6'0"	181lbs	Right	Aug 30, 2016
Jack Grealish	22	Aston Villa	5'9"	150lbs	Right	Mar 1, 2012
John McGinn	23	Aston Villa	5'10"	150lbs	Left	Aug 8, 2018
Anwar El Ghazi	23	Aston Villa	6'2"	550lbs	Right	Jan 31, 2017
Conor Hourihane	27	Aston Villa	5'11"	137lbs	Left	Jan 26, 2017
James Chester	29	Aston Villa	5'11"	174lbs		Aug 12, 2016
James Chester	29	Aston Villa	5'11"	174lbs		Aug 12, 2016
James Chester	29	Aston Villa	5'11"	174lbs		Aug 12, 2016
James Chester	29	Aston Villa	5'11"	174lbs		Aug 12, 2016
Jonathan Kodjia	2	Aston Villa	6'2"	170lbs	Right	Aug 30, 2016
Callum Wilson	26		5'11"	146lbs	Right	Jul 4, 2014

Missing Duplicate Inconsistent

How to Fix Data Quality Issues?

Once you identify the inaccurate and missing data, you can use the alternate source of data, if available.



Data quality remediation

A simple approach is to remove the inaccurate data

- Can work well if you have a few inaccurate data points.
- But, if there are many records with data quality problems, then this approach can reduce the data size, resulting in a poor analysis.

Player Name	Age	Club	Height	Weight	Foot	Joined
Pierre-Emerick Aubameyang	29	Arsenal	6'2"	176lbs	Right	Jan 31, 2018
Alexandre Lacazette	27	Arsenal	5'9"	161lbs	Right	Jul 5, 2017
Bernd Leno		Arsenal	6'3"	183lbs	Right	Jul 1, 2018
Henrikh Mkhitaryan	29	Arsenal	5'10"	165lbs	Right	Jan 22, 2018
Granit Xhaka	25	Arsenal	6'1"	181lbs	Left	Jul 1, 2016
Shkodran Mustafi	26	Arsenal	6'0"	181lbs	Right	Aug 30, 2016
Jack Grealish	22	Aston Villa	5'9"	150lbs	Right	Mar 1, 2012
John McGinn	23	Aston Villa	5'10"	150lbs	Left	Aug 8, 2018
Anwar El Ghazi	23	Aston Villa	6'2"	550lbs	Right	Jan 31, 2017
Conor Hourihane	27	Aston Villa	5'11"	137lbs	Left	Jan 26, 2017
James Chester	29	Aston Villa	5'11"	174lbs		Aug 12, 2016
James Chester	29	Aston Villa	5'11"	174lbs		Aug 12, 2016
James Chester	29	Aston Villa	5'11"	174lbs		Aug 12, 2016
James Chester	29	Aston Villa	5'11"	174lbs		Aug 12, 2016
Jonathan Kodjia	2	Aston Villa	6'2"	170lbs	Right	Aug 30, 2016
Callum Wilson	26		5'11"	146lbs	Right	Jul 4, 2014

Data quality remediation

A better approach, would be to impute the incorrect or missing values.

- The mean, mode, and the median of attributes, can be used for this.

Player Name	Age	Club	Height	Weight
Joe Hart	30	Burnley	5'9"	185lbs

Mode

Player Name	Age	Club	Height	Weight
Joe Hart	30	Burnley	5'9"	178.3lbs

Mean

Player Name	Age	Club	Height	Weight
Joe Hart	30	Burnley	5'9"	178.5lbs
Steven Defour	26	Burnley	6'2"	203lbs
Chris Wood	28	Burnley	6'1"	185lbs
Ashley Barnes	29	Burnley	5'11"	172lbs
Matthew Lowton	30	Burnley	5'9"	171lbs
Robert Brady	24	Burnley	6'1"	154lbs
Charlie Taylor	26	Burnley	6'0"	185lbs

Median

Data quality remediation

Another approach, is to estimate the missing weight, based on the player whose height and age is similar to Joe Hart.

- Not all values can be estimated from the values of other attributes

Player Name	Age	Club	Height	Weight
Joe Hart	30	Burnley	5'9"	171lbs
Steven Defour	26	Burnley	6'2"	203lbs
Chris Wood	28	Burnley	6'1"	185lbs
Ashley Barnes	29	Burnley	5'11"	172lbs
Matthew Lowton	30	Burnley	5'9"	171lbs
Robert Brady	24	Burnley	6'1"	154lbs
Charlie Taylor	26	Burnley	6'0"	185lbs

- the remediation approach depends, on the type of data, and the domain understanding of the data.

EDA: Explore and Ensure Data Quality

- Ensure your data is as expected/valid/appropriate for the task
- Provides insights into a dataset
- Extract/determine important variables/attributes/features
- Detect outliers and anomalies
- Test underlying assumptions
- Make informed decisions in developing models

EDA: Explore the Data

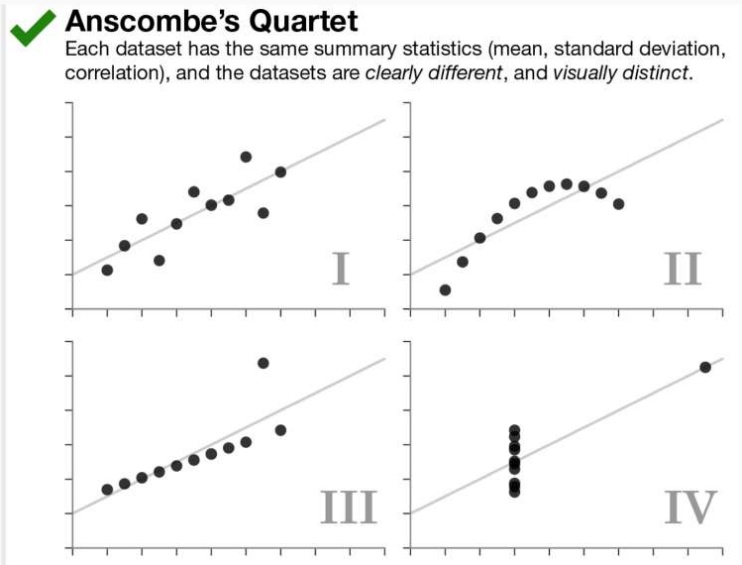
- Explore **global** properties: use histograms, scatter plots, and aggregation functions to summarize the data
- Explore **group** properties: group like-items together to compare subsets of the data (are the comparison results reasonable/expected?)
- This approach can be done at any time and any stage of the data science process

EDA: Explore the Data

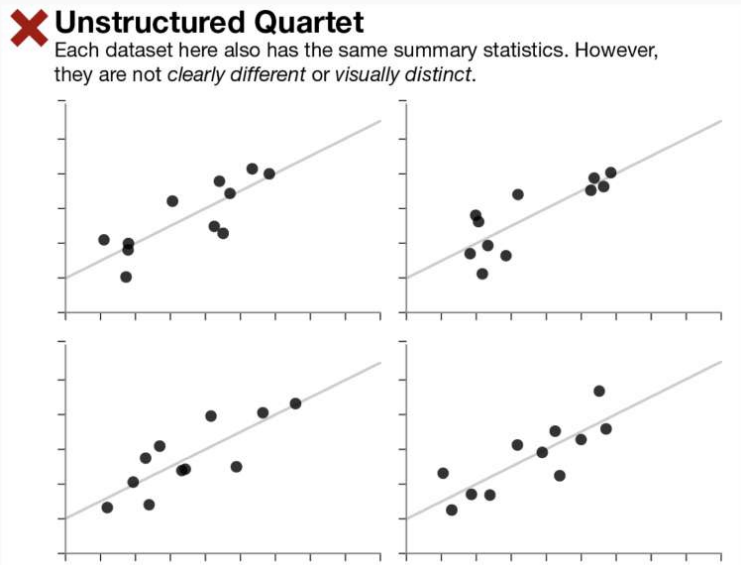
	age	page_views	fpl_value	fpl_points	market_value
count	461.000000	461.000000	461.000000	461.000000	461.000000
mean	26.804772	763.776573	5.447939	57.314534	11.012039
std	3.961892	931.805757	1.346695	53.113811	12.257403
min	17.000000	3.000000	4.000000	0.000000	0.050000
25%	24.000000	220.000000	4.500000	5.000000	3.000000
50%	27.000000	460.000000	5.000000	51.000000	7.000000
75%	30.000000	896.000000	5.500000	94.000000	15.000000
max	38.000000	7664.000000	12.500000	264.000000	75.000000

Summary statistics can only reveal so much

EDA: Visualization

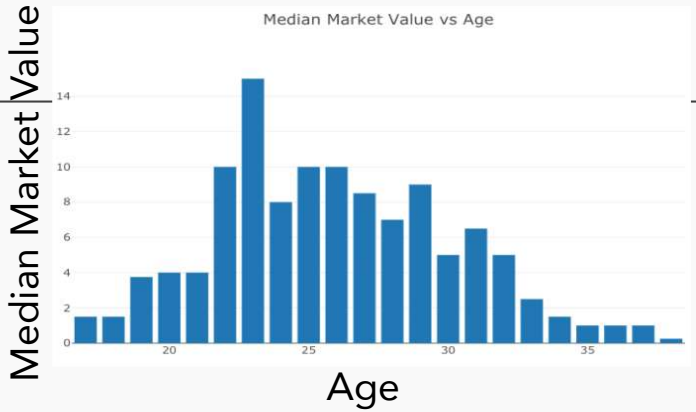


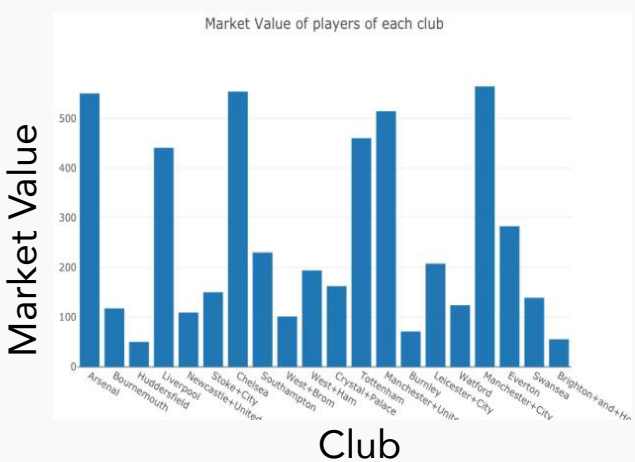
Same stats do not imply same graphs

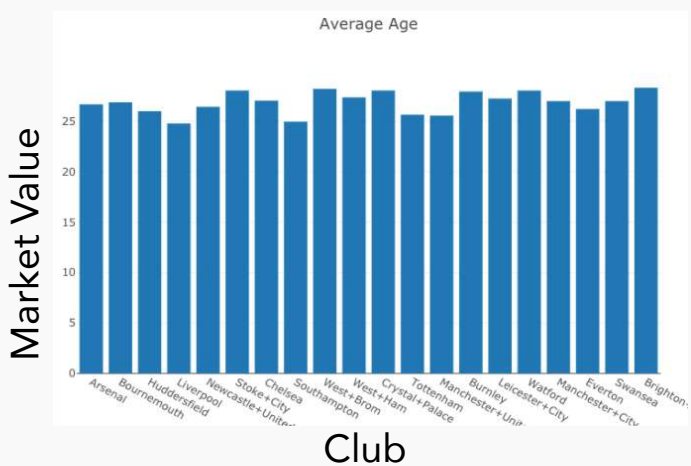


Same graphs do not imply same stats


Mean Market Value vs Age
Mean Market Value
Age

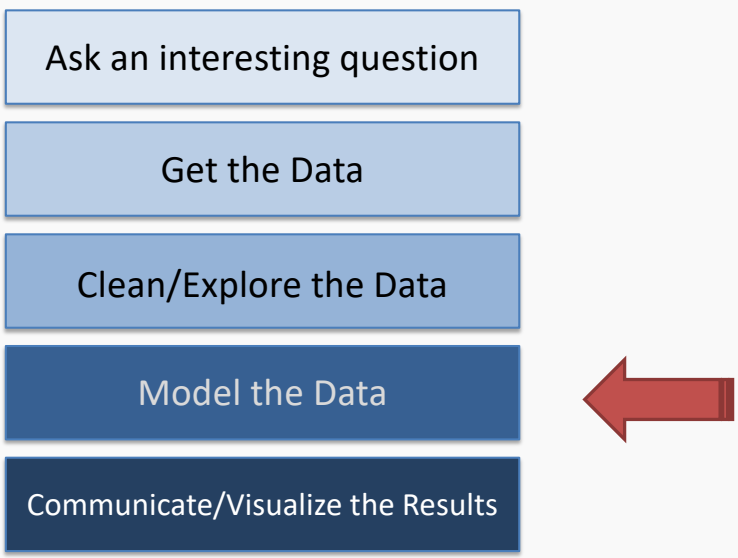

Median Market Value vs Age
Median Market Value
Age


Market Value of players of each club
Market Value
Club


Average Age
Market Value
Club

Ready to Model the Data!

The Data Science Process



Lecture Outline

Part A: Statistical Modeling

k-Nearest Neighbors (kNN)

Part B: Model Fitness

How does the model perform predicting?

Part B: Comparison of Two Models

How do we choose from two different models?

Part C: Linear Models

Lecture Outline

Part A: Statistical Modeling

k-Nearest Neighbors (kNN)

Part B: Model Fitness

How does the model perform predicting?

Part B: Comparison of Two Models

How do we choose from two different models?

Part C: Linear Models

Predicting a Variable

Let’s imagine a scenario where we'd like to predict one variable using another (or a set of other) variables.

Examples:

- Predicting the number of views a YouTube video will get next week based on video length, the date it was posted, the previous number of views, etc.
- Predicting which movies a Netflix user will rate highly based on their previous movie ratings, demographic data, etc.

Data

The **Advertising data set** consists of the sales of a particular product in 200 different markets, and advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper. Everything is given in units of $1000.

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

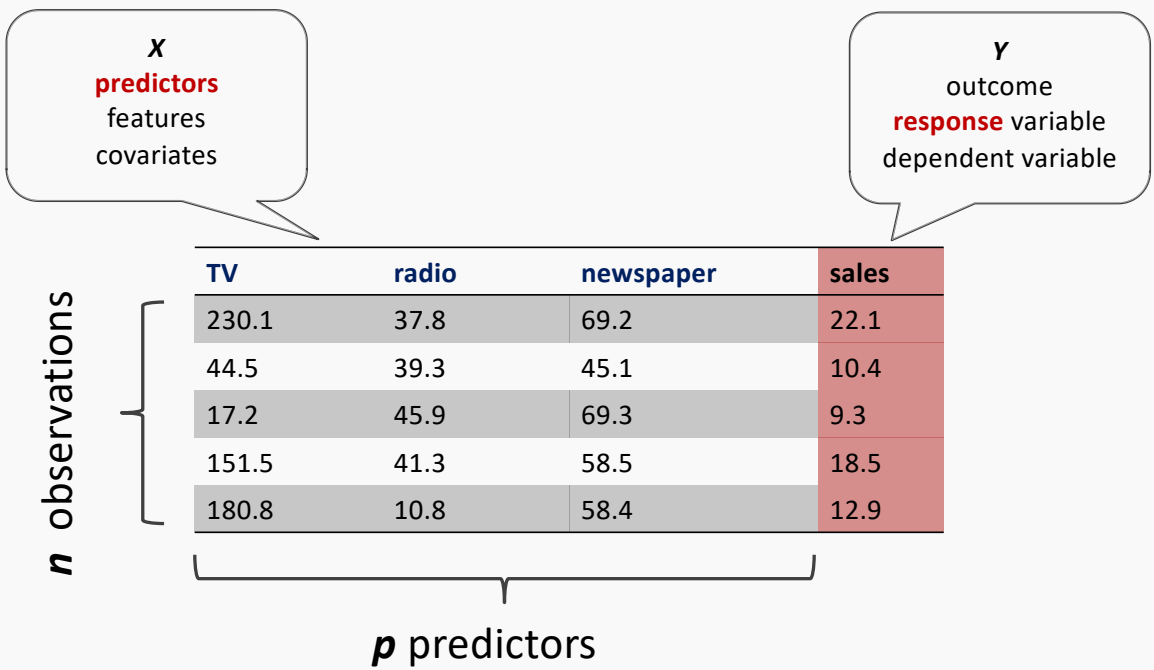
There is an asymmetry in many of these problems:

The variable we would like to predict may be more difficult to measure, is more important than the other(s), or maybe directly or indirectly influenced by the other variable(s).

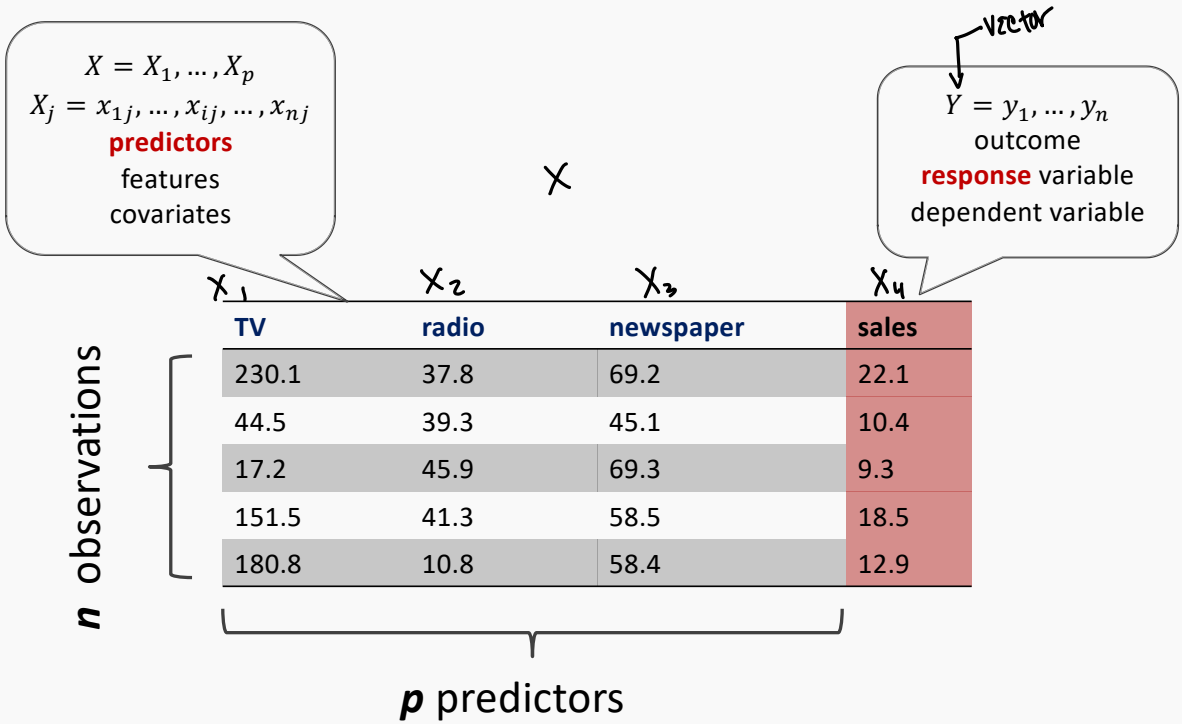
Thus, we'd like to define two categories of variables:

- Response: variables whose values we want to predict
- Predictors: variables whose values we use to make our prediction

Response vs. Predictor Variables



Response vs. Predictor Variables



Statistical Model

True vs. Statistical Model

We will assume that the response variable, Y, relates to the predictors, X, through some unknown function expressed generally as:

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

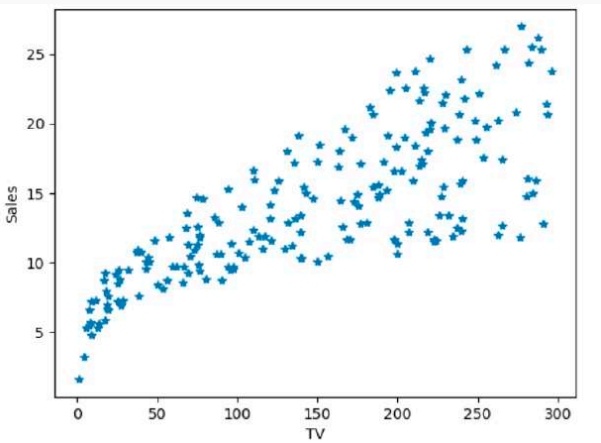
Here, f is the unknown function expressing an underlying rule for relating Y to X, ε is the random amount (unrelated to X) that Y differs from the rule $f(X)$.

A ***statistical model*** is any algorithm that estimates f. We denote the estimated function as $\hat{f}$.

Example: predicting sales

Motivation: Predict Sales

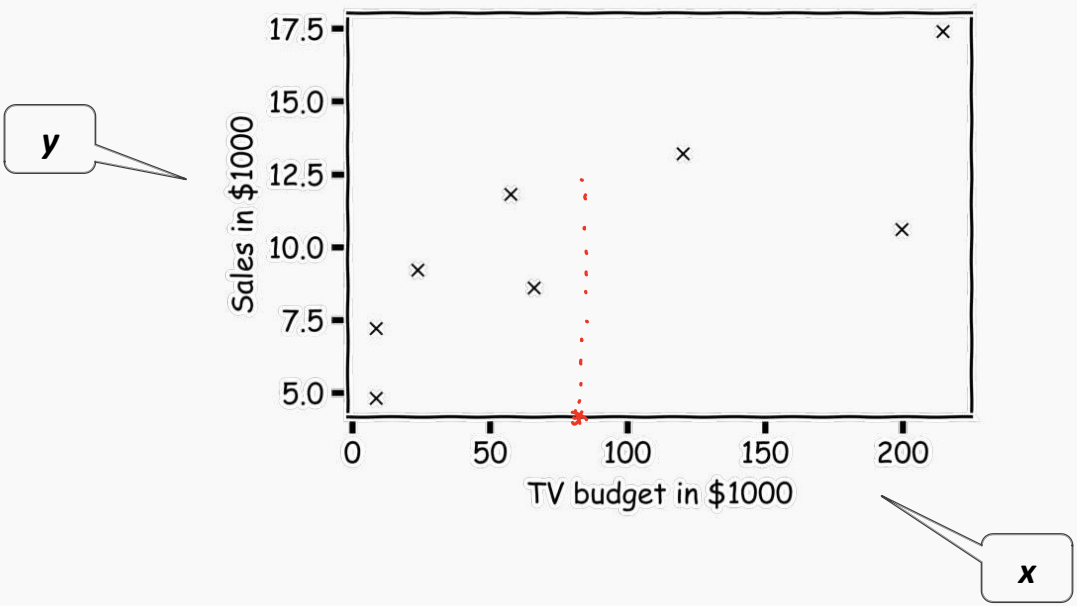
Build a model to **predict** sales based on TV budget



The response, y, is the sales

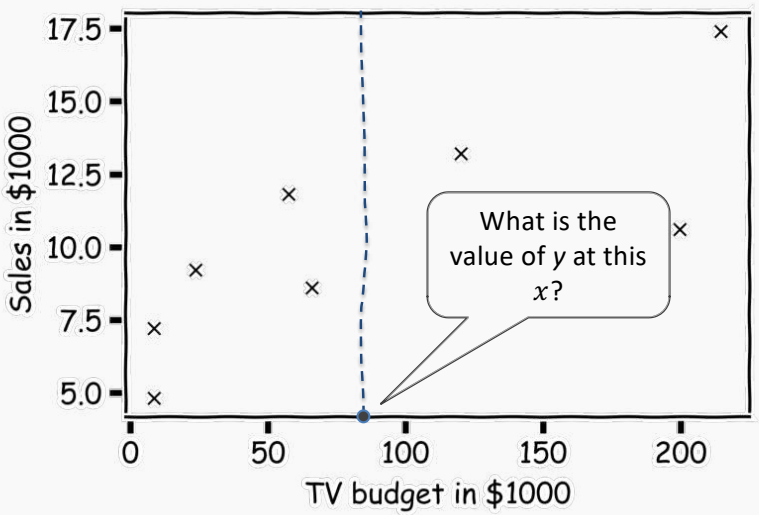
The predictor, x, is TV budget

Statistical Model



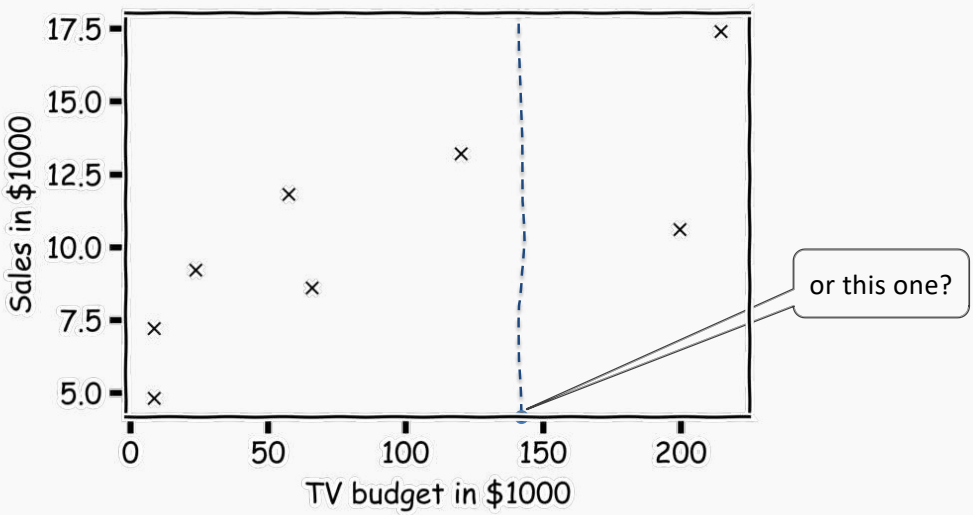
Statistical Model

How do we predict y for some x



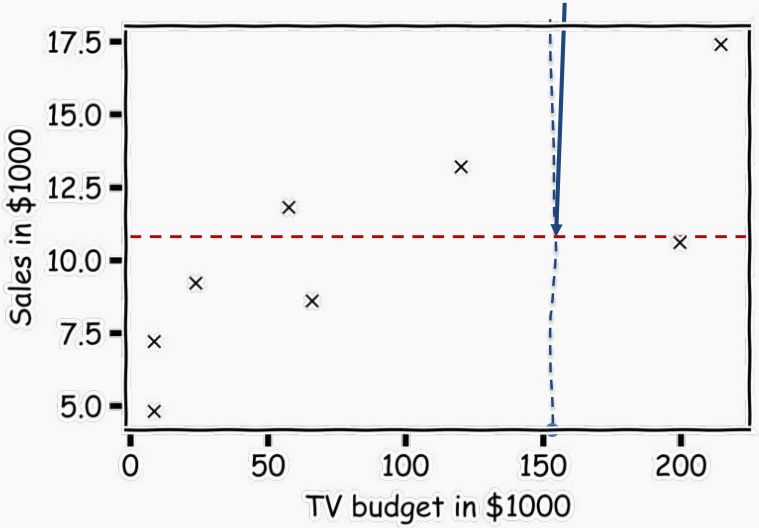
Statistical Model

How do we predict y for some x



Statistical Model

Simple idea is to take the mean of all *y's*, $\hat{f}(x) = \frac{1}{n}\sum_1^n y_i$



Prediction vs. Estimation

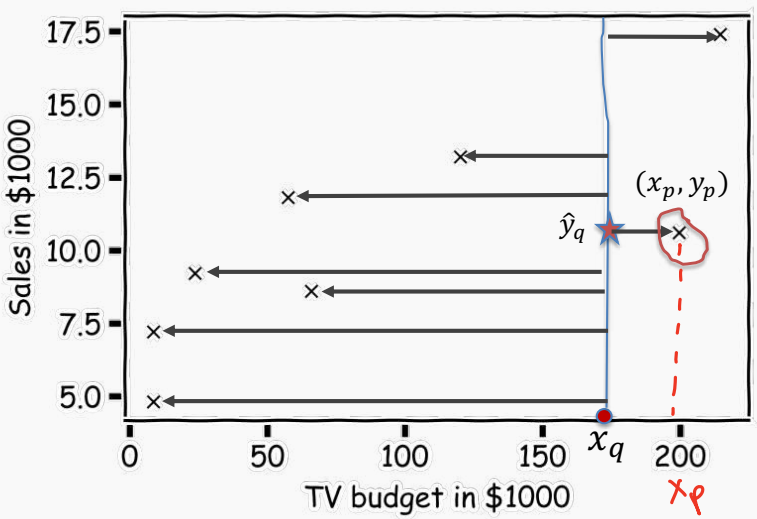
For some problems, what's important is obtaining $\hat{f}$, our estimate of f. These are called ***inference*** problems.

When we use a set of measurements, $(x_{i,1}, \dots, x_{i,p})$ to predict a value for the response variable, we denote the ***predicted*** value by:

$$\hat{y}_i = \hat{f}(x_{i,1}, \dots, x_{i,p}).$$

For some problems, we don't care about the specific form of $\hat{f}$, we just want to make our predictions $\hat{y}$'s as close to the observed values y's as possible. These are called ***prediction problems*.**

Simple Prediction Model



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

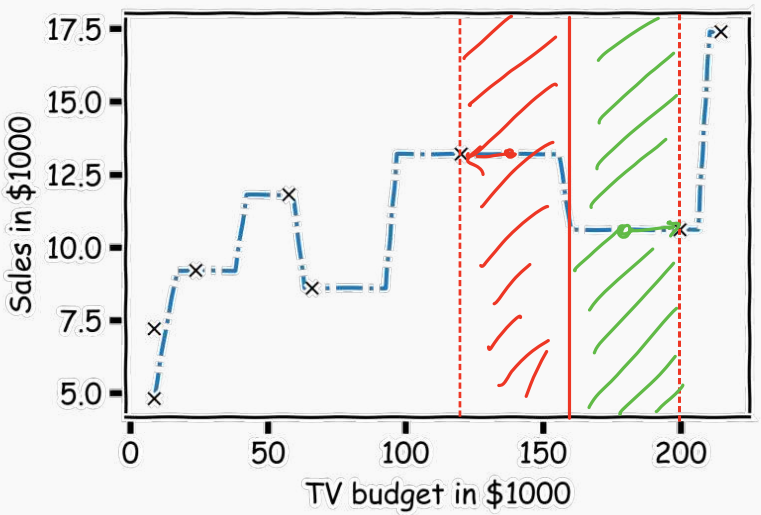
Find distances to all other points $D(x_q, x_i)$

Find the nearest neighbor, (x_p, y_p)

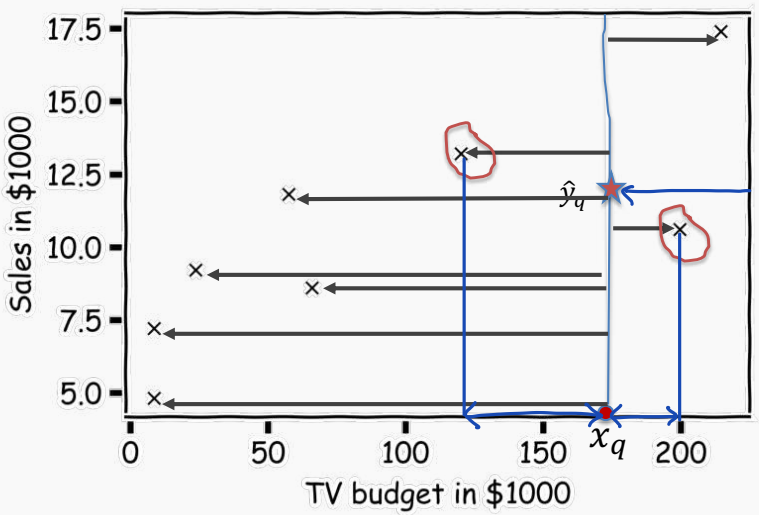
Predict $\hat{y}_q = y_p$

Simple Prediction Model

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



Extend the Prediction Model



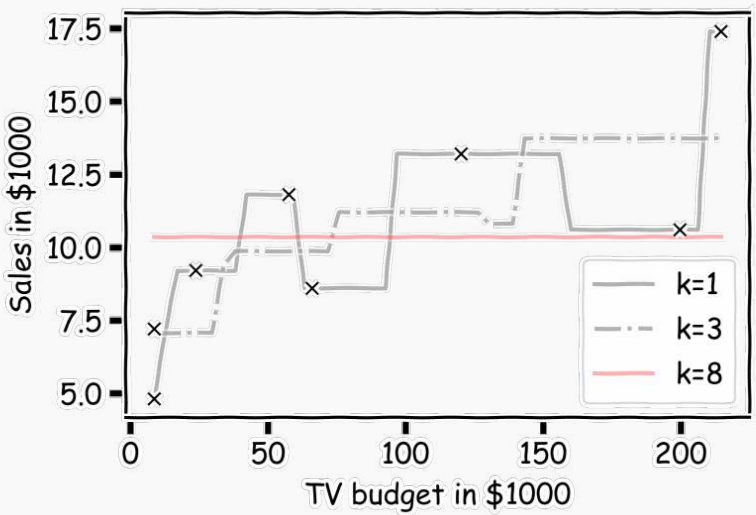
What is $\hat{y}_q$ at some x_q ?

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

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

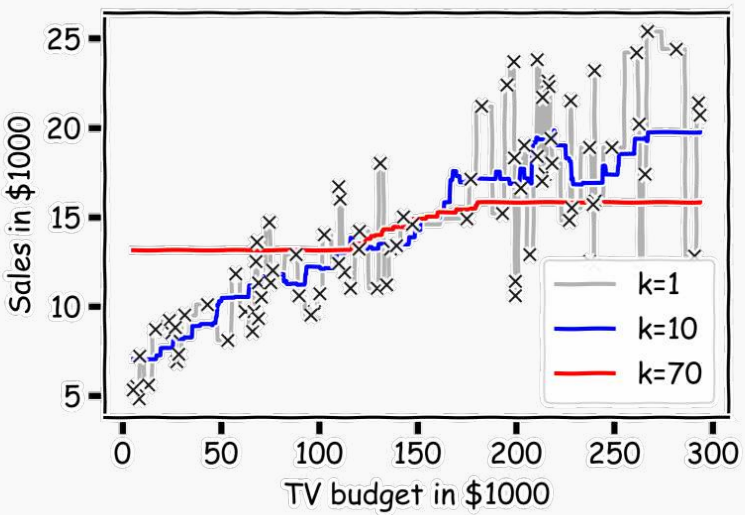
Predict $\hat{y}_q = \frac{1}{k}\sum_i^k y_{q_i}$

Simple Prediction Models



Simple Prediction Models

We can try different k-models on more data



k-Nearest Neighbors

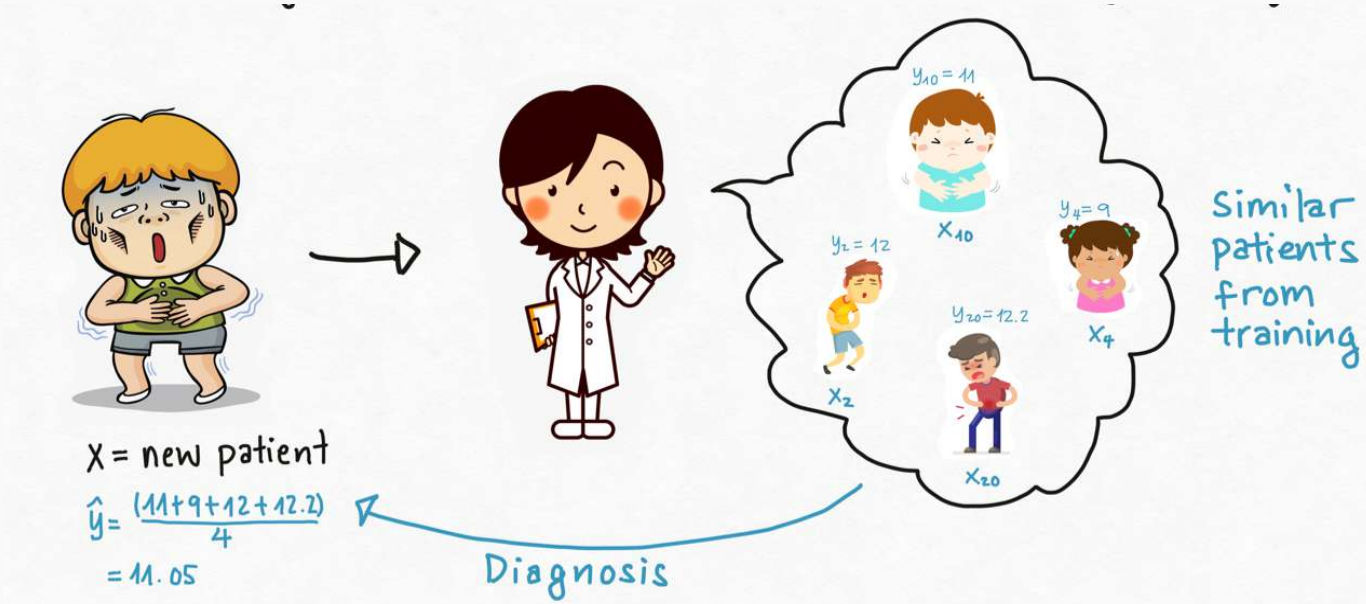
The ***k-Nearest Neighbor (kNN) model*** is an intuitive way to predict a quantitative response variable:

> *to predict a response for a set of observed predictor values, we use the responses of other observations most similar to it*

kNN is a non-parametric learning algorithm. When we say a technique is non parametric, it means that it does not make any assumptions on the underlying data distribution.

Note: this strategy can also be applied in classification to predict a categorical variable. We will encounter kNN again later in the course in the context of classification.

k-Nearest Neighbors – kNN



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

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

Lecture Outline

Part A: Statistical Modeling

k-Nearest Neighbors (kNN)

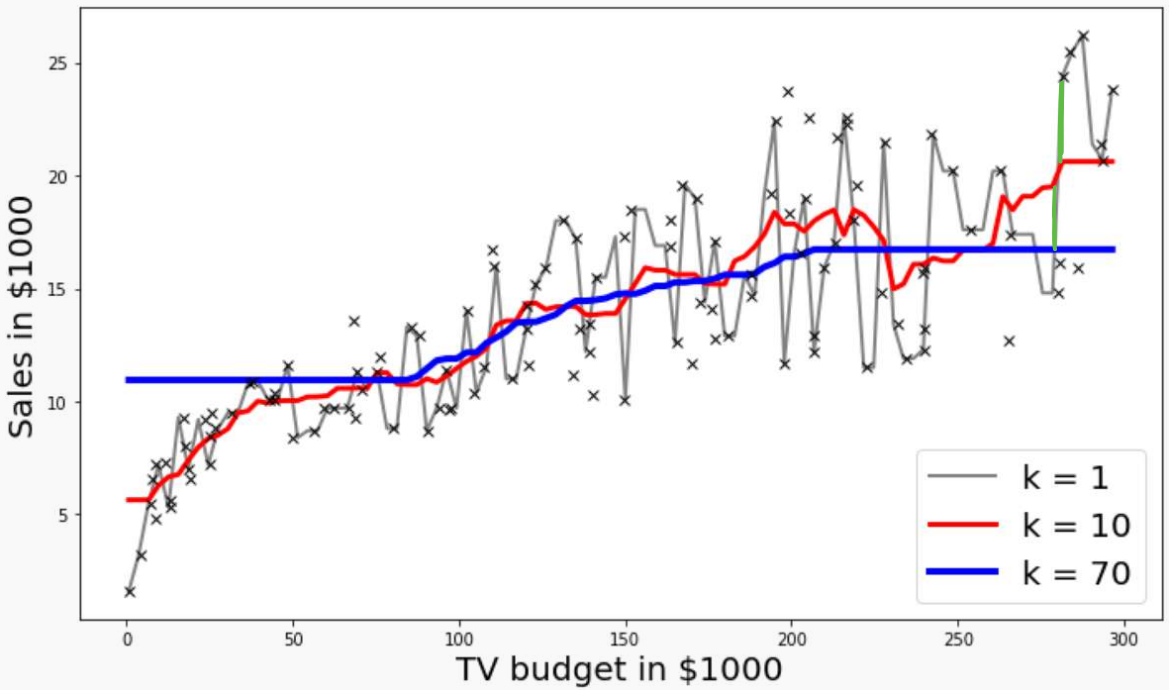
Part B: Model Fitness

How does the model perform predicting?

Part B: Comparison of Two Models

How do we choose from two different models?

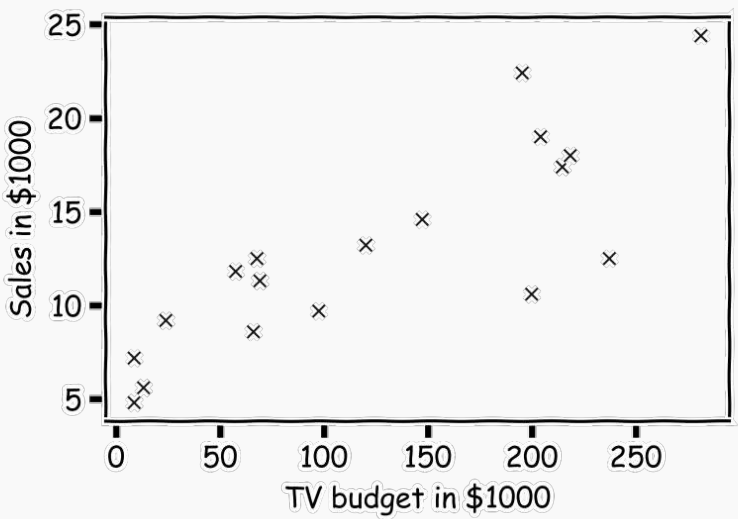
Part C: Linear Models


Sales in $1000
TV budget in $1000
k = 1
k = 10
k = 70
0
50
100
150
200
250
300
5
10
15
20
25

Error Evaluation

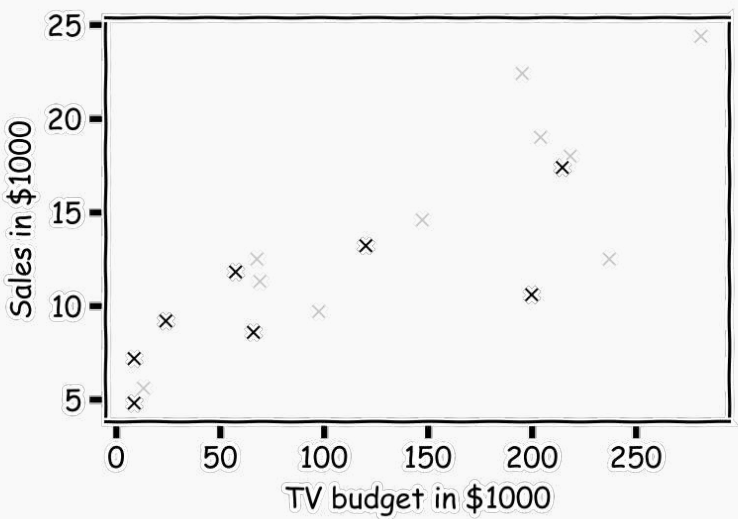
Error Evaluation

Start with some data.



Error Evaluation

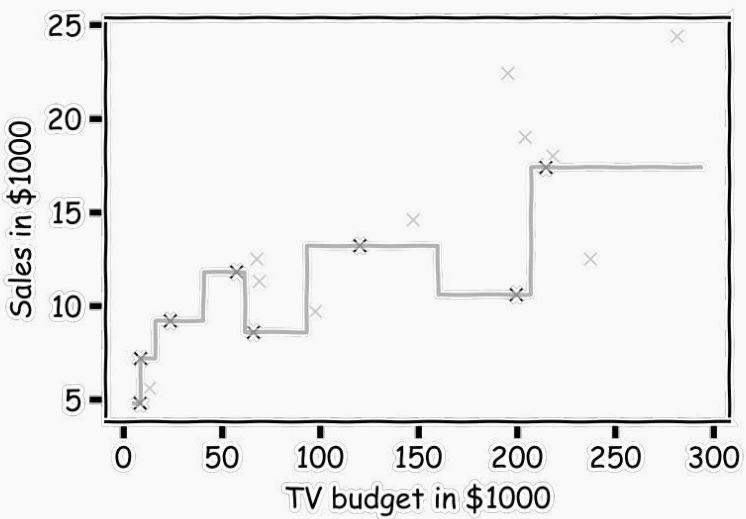
Hide some of the data from the model. This is called **train-test** split.



We use the train set to estimate $\hat{y}$, and the test set to evaluate the model.

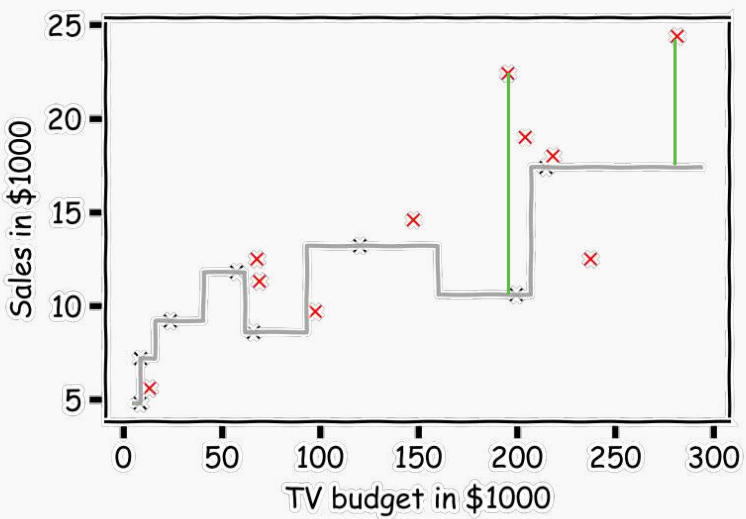
Error Evaluation

Estimate $\hat{y}$ for *k=1* .



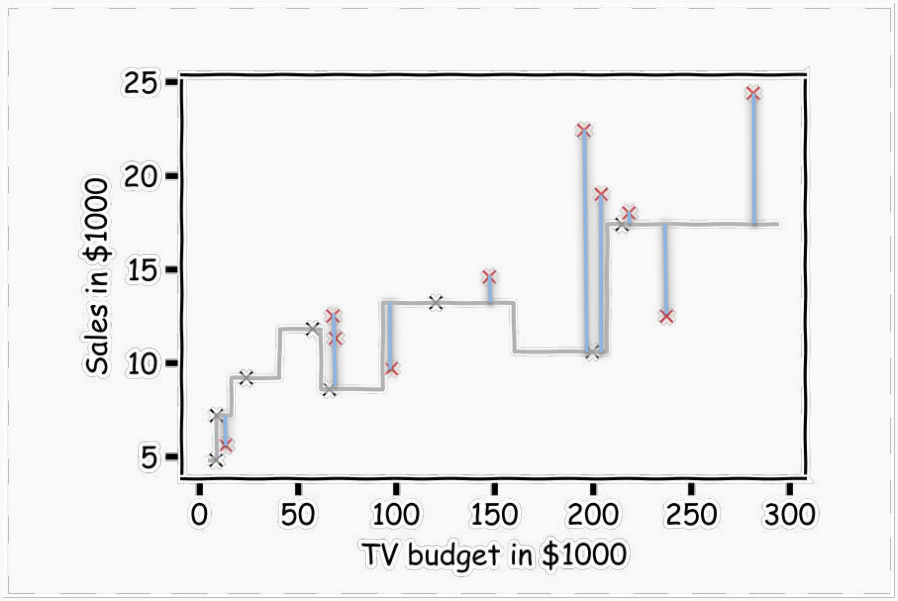
Error Evaluation

Now, we look at the data we have not used, the **test data** (red crosses).



Error Evaluation

Calculate the **residuals** $(y_i - \hat{y}_i)$.



Error Evaluation

In order to quantify how well a model performs, we aggregate the errors and we call that the ***loss*** or ***error*** *or* ***cost function.***

A common loss function for quantitative outcomes is the **Mean Squared Error (MSE):**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

Note: Loss and cost function refer to the same thing. Cost usually refers to the total loss where loss refers to a single training point.

Error Evaluation

Caution: The MSE is by no means the only valid (or the best) loss function!

1. Max Absolute Error
2. Mean Absolute Error
3. Mean Squared Error → more sensitive to outliers

We will motivate MSE when we introduce probabilistic modeling.

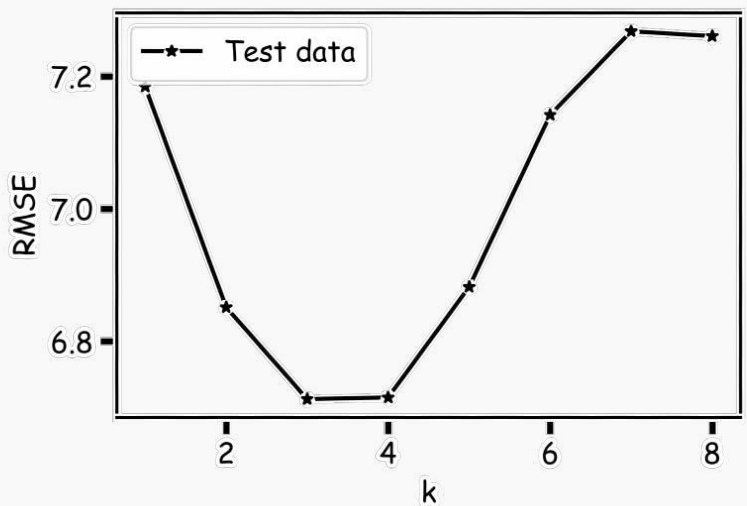
Note: The square **R**oot of the **M**ean of the **S**quared **E**rrors (RMSE) is also commonly used.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$$

Model Comparison

Model Comparison

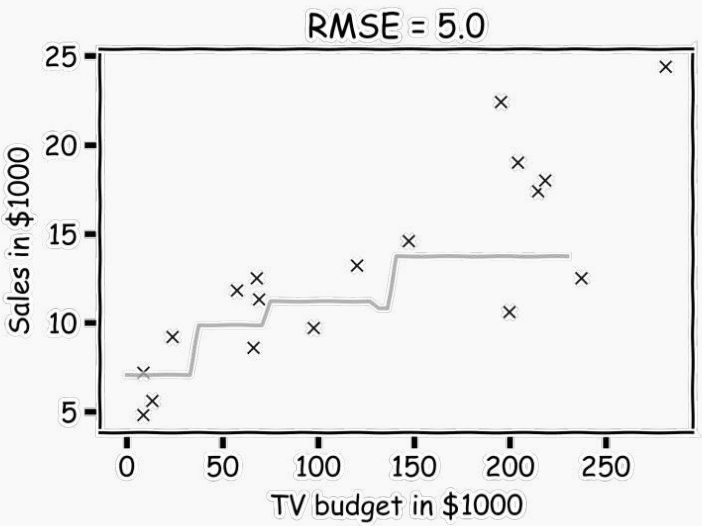
Do the same for all *k*'s and compare the RMSEs. *k=3* seems to be the **best model**.



Model Fitness

Model fitness

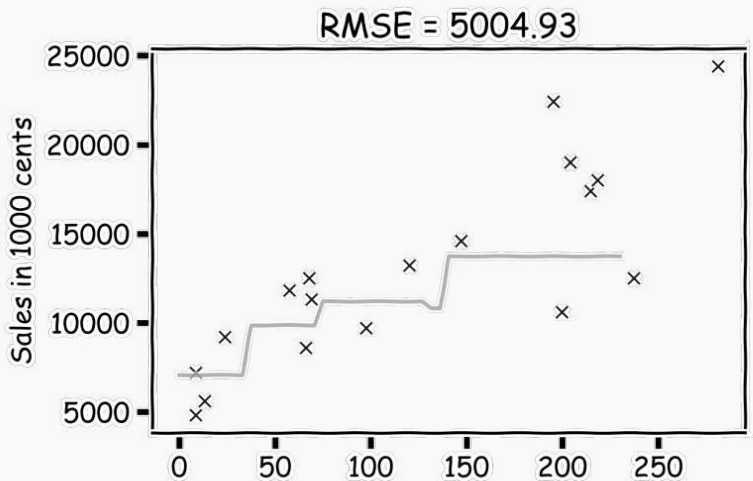
For a subset of the data, calculate the RMSE for *k=3.*



Is RMSE=5.0 good enough?

Model fitness

What if we measure the Sales in cents instead of dollars?

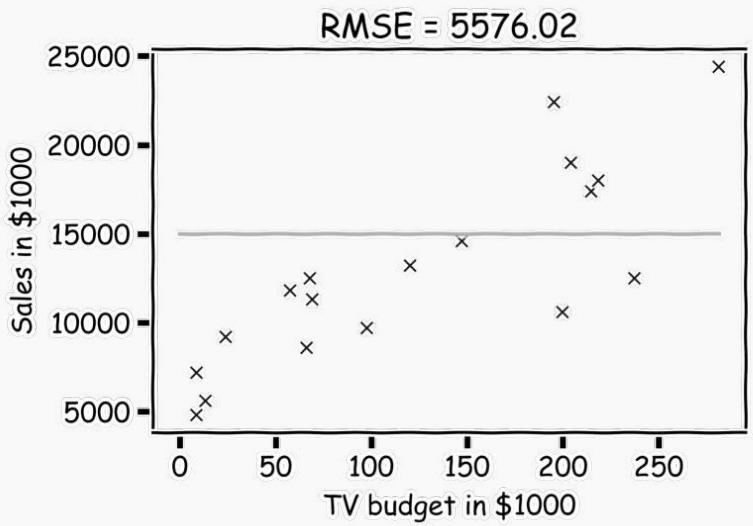


RMSE is now 5004.93.

Is that good?

Model fitness

It is better if we compare it to something.



We will use the simplest model:

$$\hat{y} = \bar{y} = \frac{1}{n}\sum_i y_i$$

as the worst possible model and

$$\hat{y_i} = y_i$$

as the best possible model.

R-squared

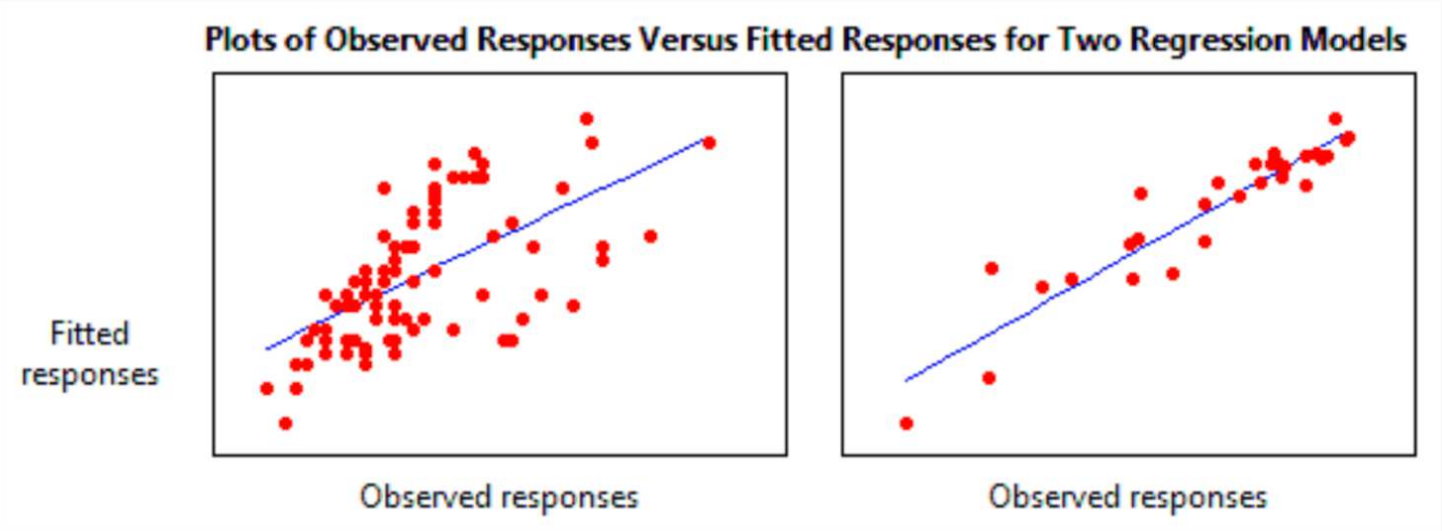
$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2}$$

→ our model

→ simplest model (predict the mean)

- If our model is as good as the mean value, $\bar{y}$, then $R^2 = 0$
- If our model is perfect then $R^2 = 1$
- R^2 can be negative if the model is worst than the average. This can happen when we evaluate the model on the test set.

R-squared



$R^2 = 0.38$ $R^2 = 0.87$

Lecture Outline

Part A: Statistical Modeling

k-Nearest Neighbors (kNN)

Part B: Model Fitness

How does the model perform predicting?

Part B: Comparison of Two Models

How do we choose from two different models?

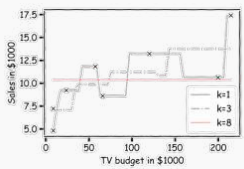
Part C: Linear Models

Lecture Outline

- Linear models
- Estimate of the regression coefficients
- Model evaluation
- Interpretation

Linear Models

Note that in building our kNN model for prediction, we did not compute a closed form for $\hat{f}$.



What if we ask the question:

"how much more sales do we expect if we double the TV advertising budget?"

Alternatively, we can build a model by first assuming a simple form of f:

$$f(x) = \beta_0 + \beta_1 X$$

Linear Regression

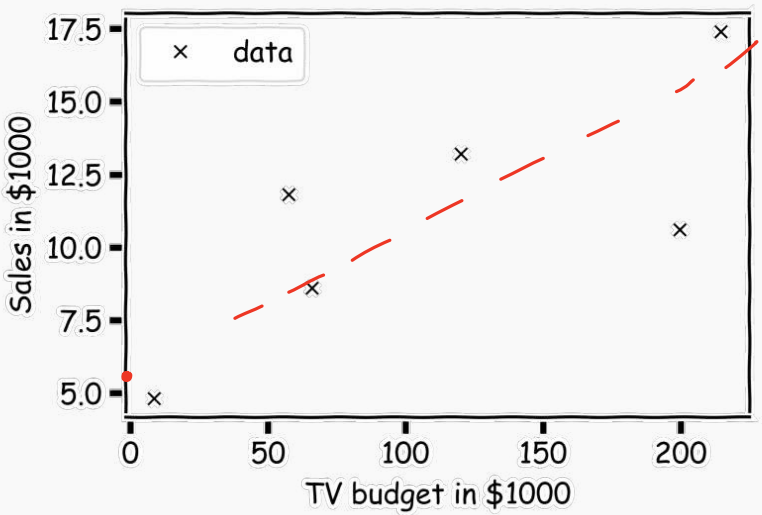
… then it follows that our estimate is:

$$\widehat{Y} = \widehat{f}(X) = \widehat{\beta_1} X + \widehat{\beta_0}$$

where $\hat{\beta}_1$ and $\hat{\beta}_0$ are **estimates** of β_1 and β_0 respectively, that we compute using observations.

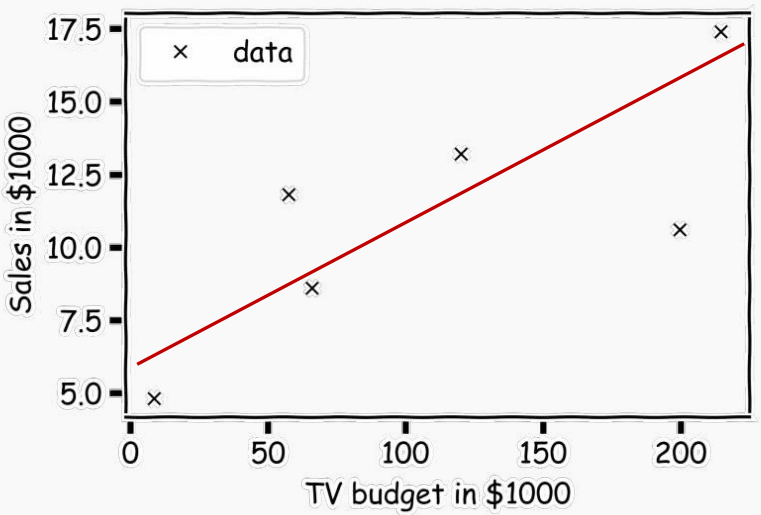
Estimate of the regression coefficients

For a given data set



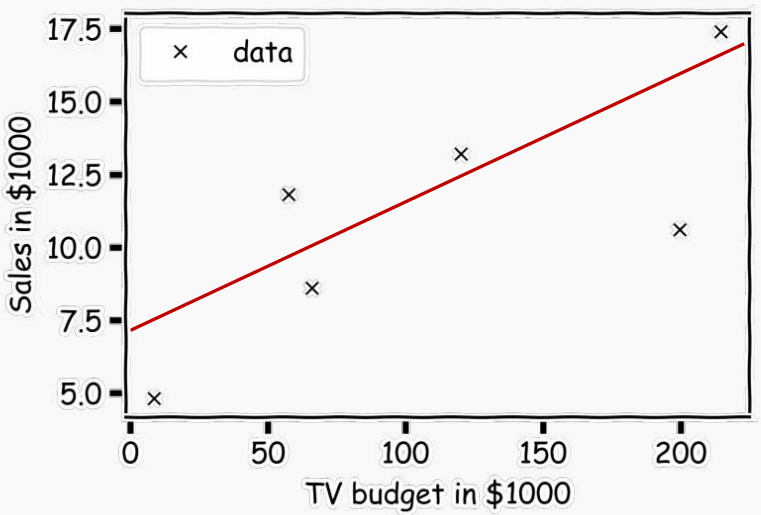
Estimate of the regression coefficients (cont)

Is this line good?



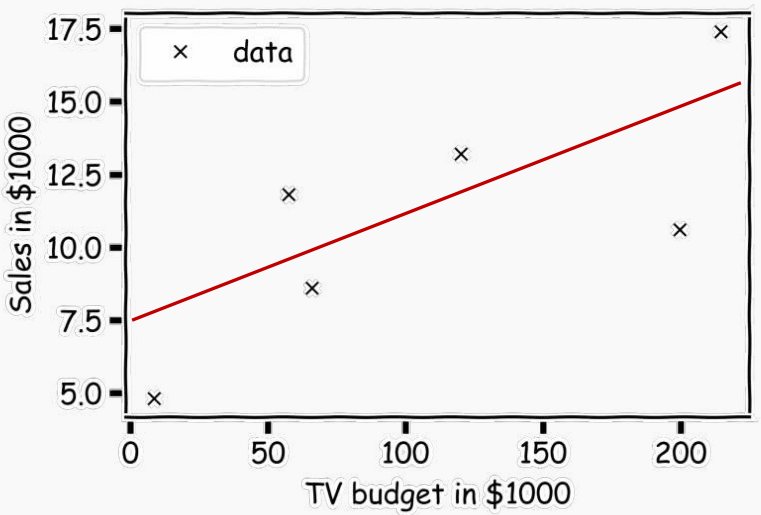
Estimate of the regression coefficients (cont)

Maybe this one?



Estimate of the regression coefficients (cont)

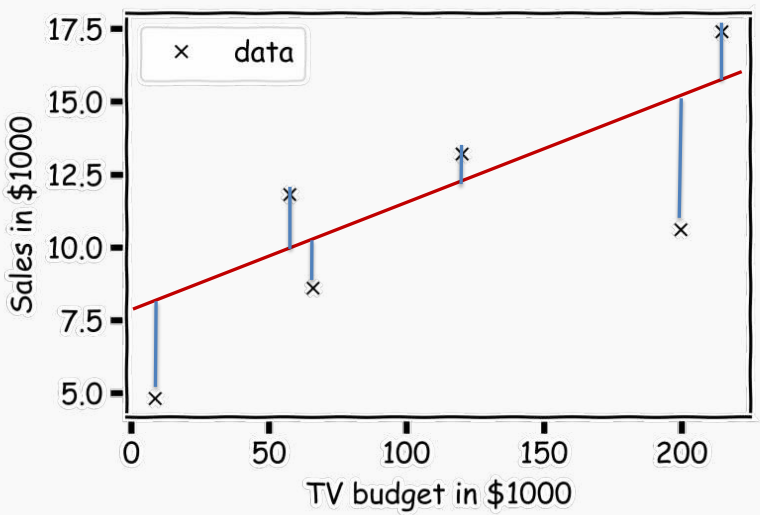
Or this one?



Estimate of the regression coefficients (cont)

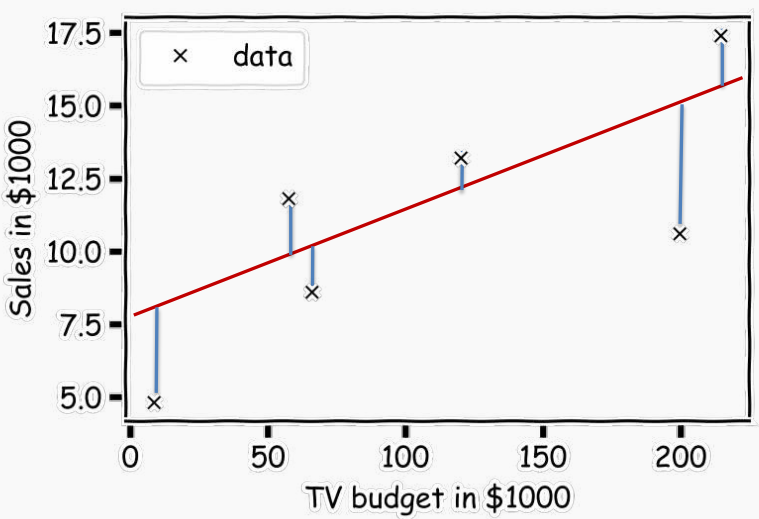
Question: Which line is the best?

For each observation (x_n, y_n), the absolute residual is $r_i = |y_i - \hat{y}_i|$.



Loss Function: Aggregate Residuals

How do we aggregate residuals across the entire dateset?



1. Max Absolute Error
2. Mean Absolute Error
3. Mean Squared Error

Estimate of the regression coefficients (cont)

Again we use MSE as our **loss function**,

$$L(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - (\beta_1 X + \beta_0)]^2.$$

We choose $\hat{\beta}_1$ and $\hat{\beta}_0$ in order to minimize the predictive errors made by our model, i.e. minimize our loss function.

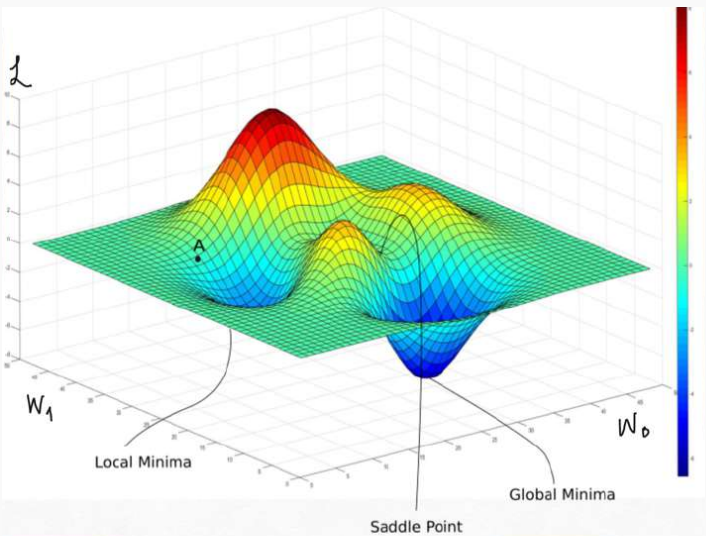
Then the optimal values for $\hat{\beta}_0$ and $\hat{\beta}_1$ should be:

$$\widehat{\beta}_0, \widehat{\beta}_1 = \underset{\beta_0, \beta_1}{\operatorname{argmin}} L(\beta_0, \beta_1).$$

WE CALL THIS **FITTING** OR **TRAINING** THE MODEL

Optimization

How does one minimize a loss function?



The global minima or maxima of $L(\beta_0, \beta_1)$ must occur at a point where the gradient (slope)

$$\nabla L = \left[\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}\right] = 0$$

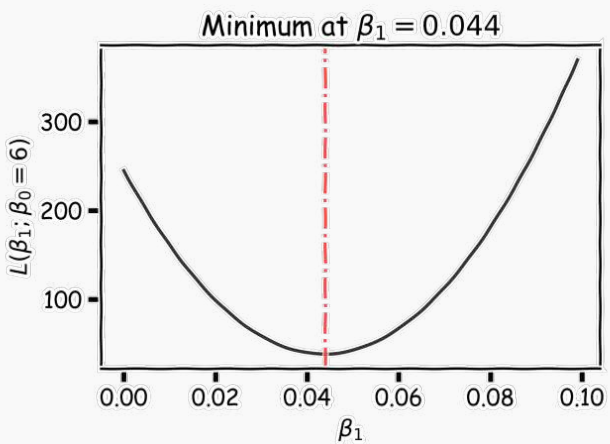
- Brute Force: Try every combination
- Exact: Solve the above equation
- Greedy Algorithm: Gradient Descent

Optimization: Estimate of the regression coefficients

Brute force

A way to estimate $\text{argmin}_{\beta_0,\beta_1} L$ is to calculate the loss function for every possible β_0 and β_1. Then select the β_0 and β_1 where the loss function is minimum.

E.g. the loss function for different β_1 when β_0 is fixed to be 6:



Very **computationally expensive** with many coefficients

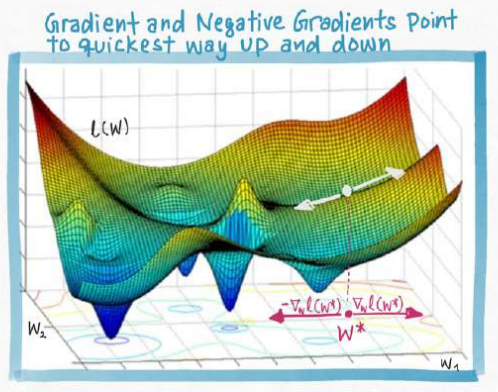
Gradient Descent

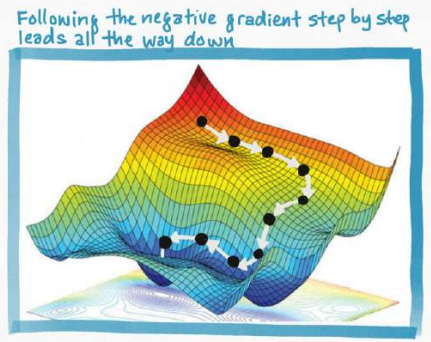
When we can't analytically solve for the stationary points of the gradient, we can still exploit the information in the gradient.

The gradient ∇L at any point is the direction of the steepest increase. The negative gradient is the direction of steepest decrease.

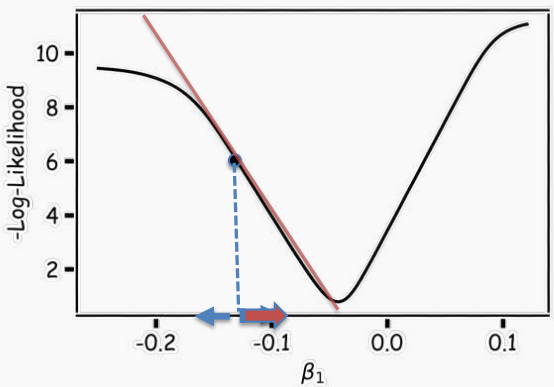
By following the negative gradient, we can eventually find the lowest point.

This method is called Gradient Descent





Gradient Descent



- Start from a random point
  1. Determine which direction to go to reduce the loss (left or right)
  2. Compute the slope of the function at this point and step to the right if slope is negative or step to the left if slope is positive
  3. Goto to #1

Estimate of the regression coefficients: analytical solution

Take the gradient of the loss function and find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ where the gradient is zero: $\nabla L = \left[\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}\right]$=0

This does not usually yield to a close form solution. However for linear regression this procedure gives us explicit formulae for $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{y}$ and $\bar{x}$ are sample means.

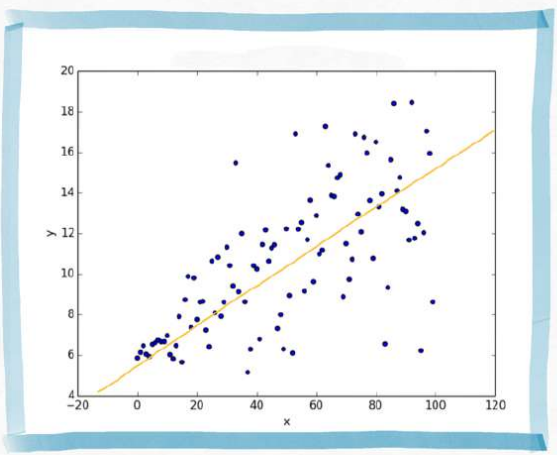
The line:

$$\widehat{Y} = \widehat{\beta}_1 X + \widehat{\beta}_0$$

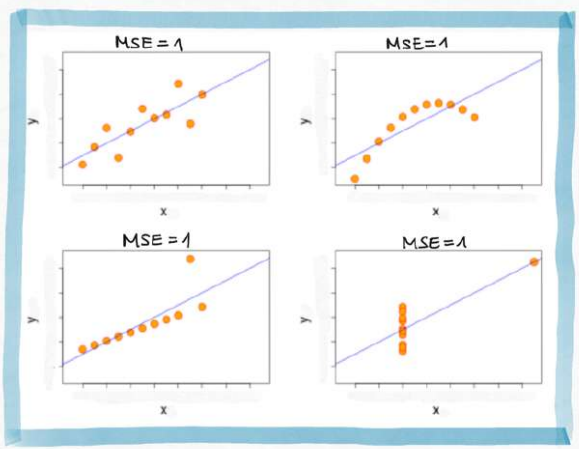
is called the **regression line**.

Evaluation: Training Error

Just because we found the model that minimizes the squared error it doesn't mean that it's a good model. We investigate the R2 but also:



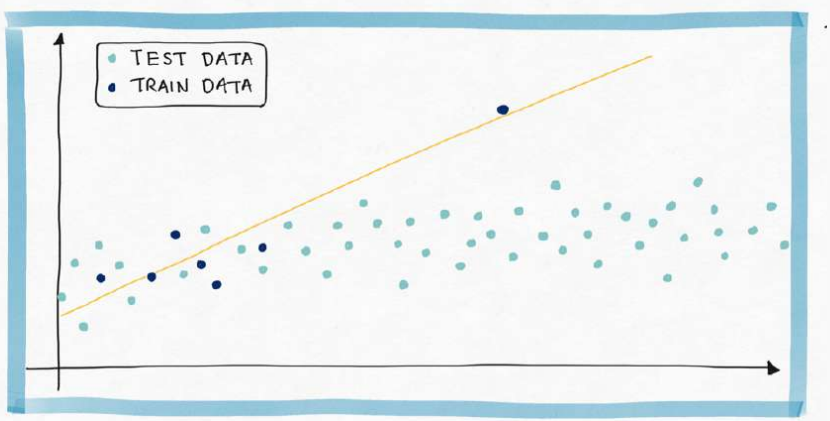
The MSE is high due to noise in the data.



The MSE is high in all four models but the models are not equal.

Evaluation: Test Error

We need to evaluate the fitted model on new data, data that the model did not train on, the test data.



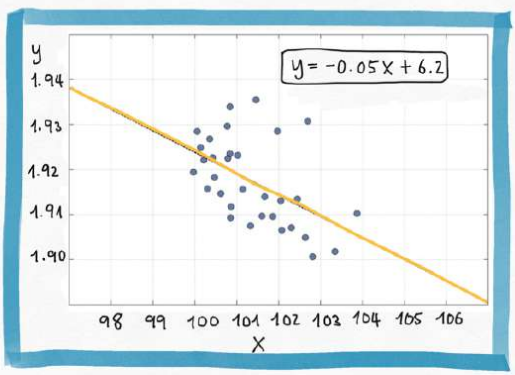
The training MSE here is 2.0 where the test MSE is 12.3.

The training data contains a strange point – an outlier – which confuses the model.

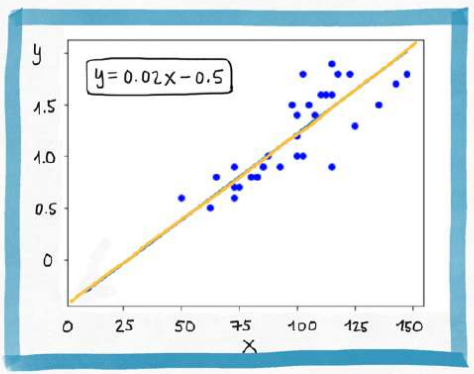
Fitting to meaningless patterns in the training is called **overfitting.**

Evaluation: Model Interpretation

For linear models it's important to interpret the parameters



The MSE of this model is very small. But the slope is -0.05. That means the larger the budget the less the sales.



The MSE is very small but the intercept is -0.5 which means that for very small budget we will have negative sales.