

CS M148 –

Data Science Fundamentals

Lecture #2: Data Collection & Bias

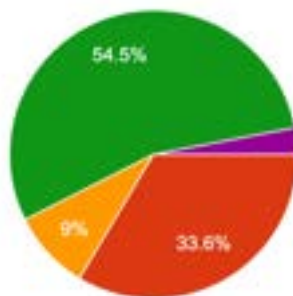
Baharan Mirzasoleiman

UCLA Computer Science

Survey Results

Which year you're at?

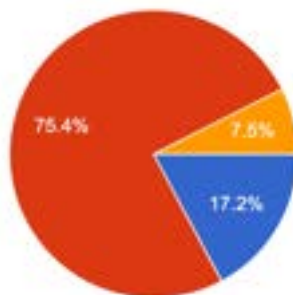
134 responses



- Freshman
- Junior
- Sophomore
- Senior
- Graduate student

Why are you taking this course?

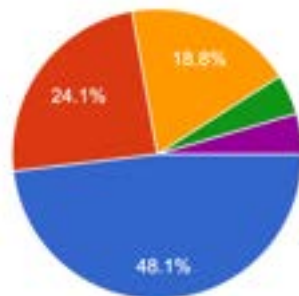
134 responses



- I'd like to become a data scientist
- I'm just interested in the materials
- I'm just curious, I may drop the course

How much math are you comfortable with?

133 responses



- Very comfortable
- I know some statistics
- I know some linear algebra
- I know some probability
- I don't have a good background in math

Have you taken CS146 or CS145 before?

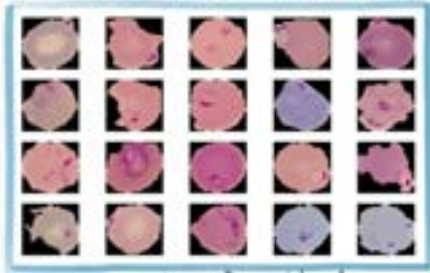
134 responses



- Yes
- No
- Yes, but I'm not comfortable with the materials
- I'm taking CS146 or CS145 this quarter

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

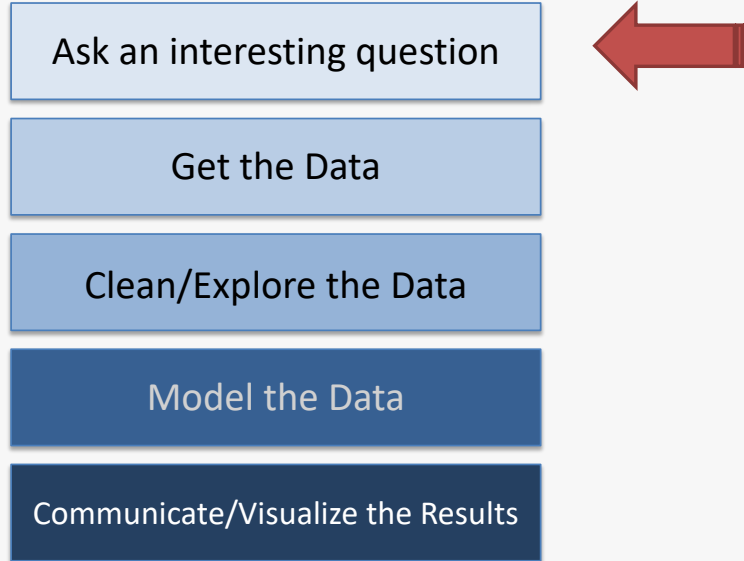
Racial Bias



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

What?

The Data Science Process



Problem Statement

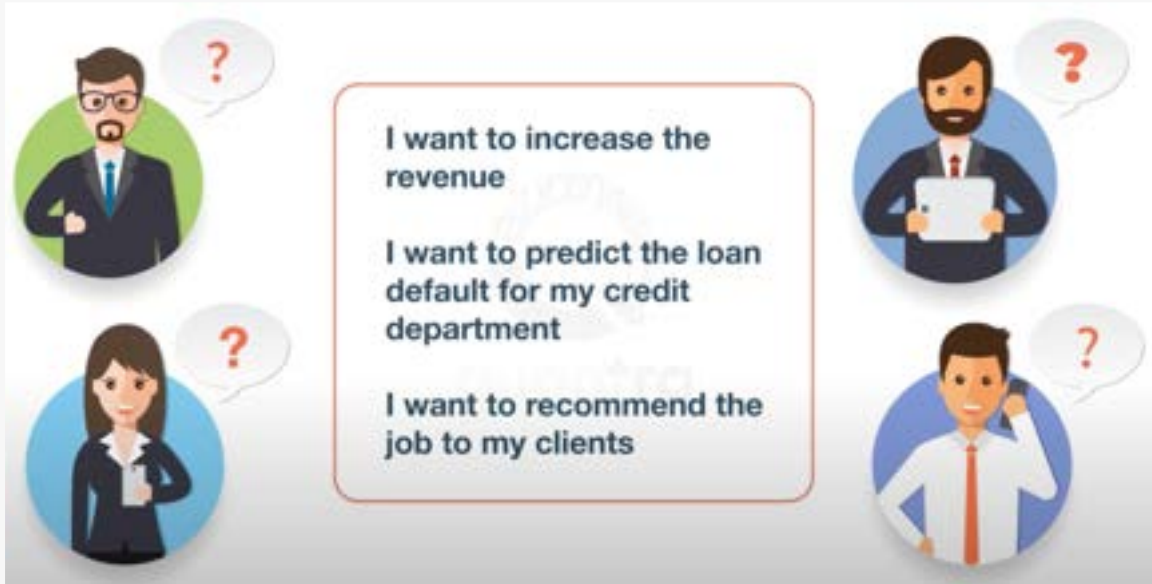
Creating a well-defined problem statement is the first and critical step in data science.

- A brief description of the problem that you are going to solve



Problem Statement

Most of the times, these initial set of problems shared with you is vague and ambiguous.



Problem Statement

You have to make the problem statement clear, goal-oriented and measurable, by asking the right set of questions.

- Are you satisfied with marketing strategies?
- What are the marketing strategies used by you?

Problem statement: What makes email marketing successful compared to other techniques?



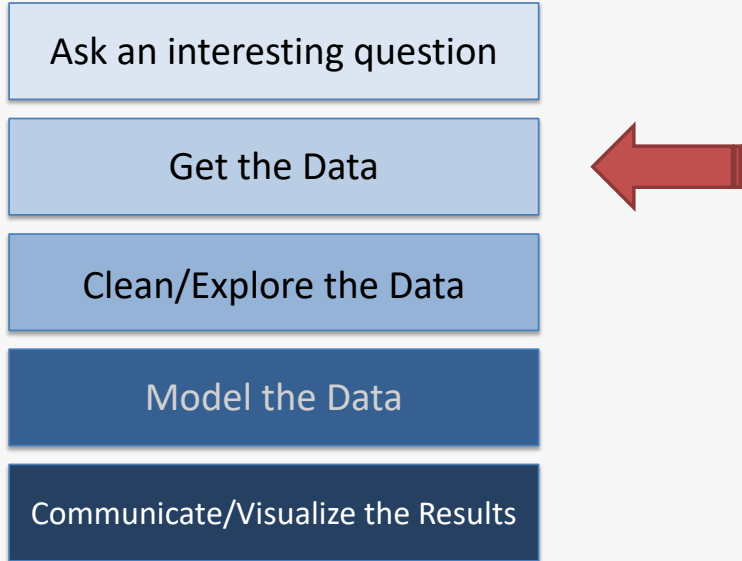
Problem Statement

Which club will win the EPL?



What?

The Data Science Process



Data Collection

Way of directly measuring variables and gathering information, allow you to gain first-hand knowledge and original insights into your research problem



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.



Primary Data Collection

Surveys: collect data by asking people directly

- Ask people to fill out questionnaire themselves
 - More common in quantitative research
 - Include closed questions with multiple-choice answers or rating scales
 - Collect consistent data and analyze the responses statistically
- Conduct interview where you ask questions and record the answers
 - More common in qualitative research
 - Allow participants to answer in their own words
 - You can ask follow up questions and explore ideas in more depth
 - However, it's more time-consuming and usually involves a smaller group of participants

Primary Data Collection

Observations: collect data unobtrusively

- Quantitative observations: systematically measuring or counting specific events, behaviors, etc.
 - You need to define the categories and criteria of your observation in advanced
- Qualitative observations: taking detailed notes and writing rich description of what is observed
 - You don't need to decide in advanced how to categorize your observations

In theory, observations allow you to collect data on how people really behave (and not what they just say they do)

- But being observed may make people behave differently!

Primary Data Collection

Data collection methods in other fields:

- Media and communication: a sample of text to be analyzed (e.g. speech, article, social media post)
- Psychology: technologies to measure things like attention or reaction time
- Education: tests or assignments to collect data about knowledge and skills
- Physical science: scientific instruments to measure e.g. weight or blood pressure

Secondary Data Collection



Secondary Data Collection

Instead of collecting your own data, you can use secondary data that is already collected

- Datasets from government surveys or previous studies
- Can be found on open-source websites such as Kaggle, Gapminder, news articles, government census, magazines, etc.
- Gives you access to much larger data
- However, you don't have any control over which variables to measurement or how to measure them. So, the conclusions you can draw might be limited

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?

Data Collection Procedure

Step 1: precisely define your variables and decide exactly how you'll measure them

- Some variables like height or age are easy to measure
- But often you'll deal with more abstract concepts like satisfaction, anxiety, or competence

Data Collection Procedure

Step 2 (operationalization): turning these fuzzy ideas into measurable indicators

- If you're using observations, which events or actions will you count?
- If you're using surveys, which questions will you ask and what range of responses you offer?



Data Collection Procedure

You should also consider the validity and reliability of your measurements

Reliability: consistently reproducible results

Validity: actually measuring the concept you're interested in

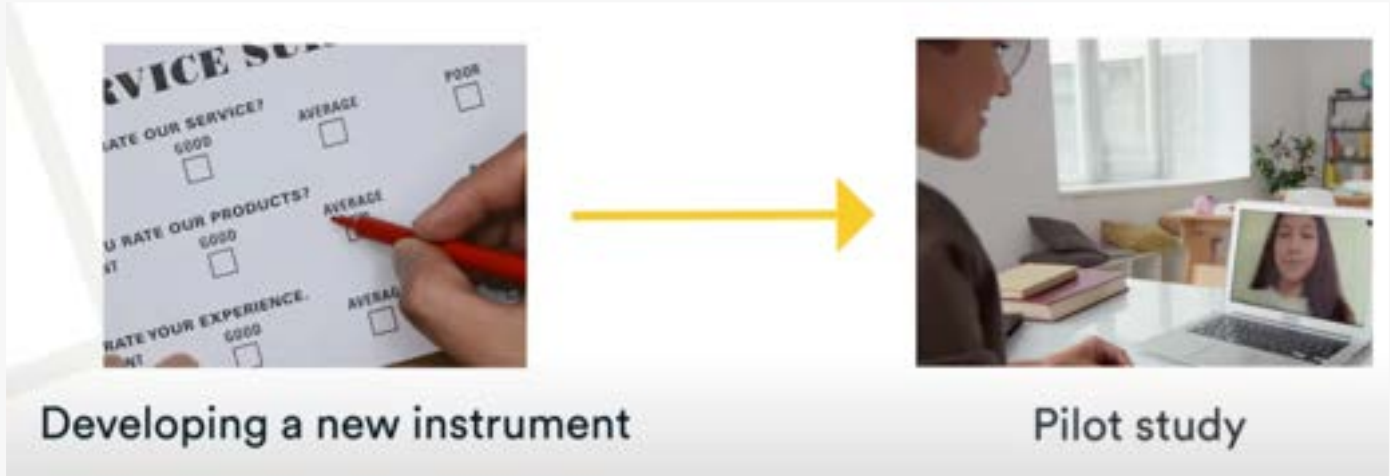
The diagram features the words "Reliability" and "Validity" in bold black font, separated by a large pink plus sign. Below "Reliability", there are two bullet points: "• Measurement materials should be thoroughly researched and carefully designed" and "• All steps should be carried out in the same way for each participant". The entire diagram is set against a light gray background with a subtle gradient.

Reliability + **Validity**

- Measurement materials should be thoroughly researched and carefully designed
- All steps should be carried out in the same way for each participant

Data Collection Procedure

If you're developing a new instrument to measure a specific concept, run a pilot study to check its validity and reliability in advanced



How will you chose your participants?

Step 3: Choosing a sample

- How many participants do you need for an adequate sample size?
- What criteria will you use to identify eligible participants?
- How will you 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



Population

Example:

- studying the effectiveness of online teaching in the US
 - Very difficult to get a representative sample!
- 9-th grade students in low-income areas on NY
 - Narrower population, more manageable!



9th grade students



Low-income areas of NY

Sample

Two main approaches to select a sample:

Probability sampling	Non-probability sampling
<ul style="list-style-type: none">• Sample is selected using random methods• Mainly used in quantitative research	<ul style="list-style-type: none">• Sample is selected in a non-random way• Used in qualitative and quantitative research

The sampling method affects how confidently you can generalize your results to the population

Probability Sampling Methods

Probability sampling helps ensure that your sample is representative and unbiased

- You can use statistics to draw strong conclusions about the whole



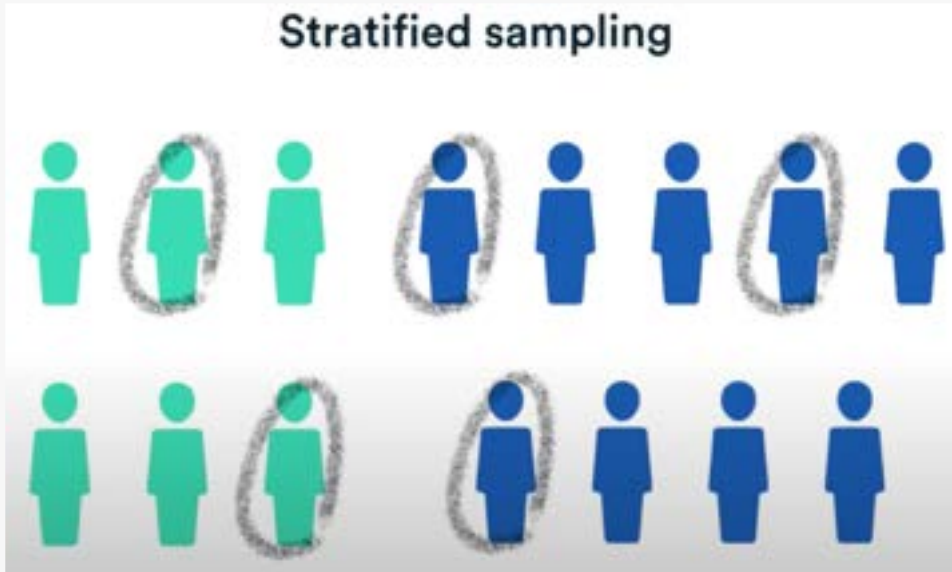
Sampling methods

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



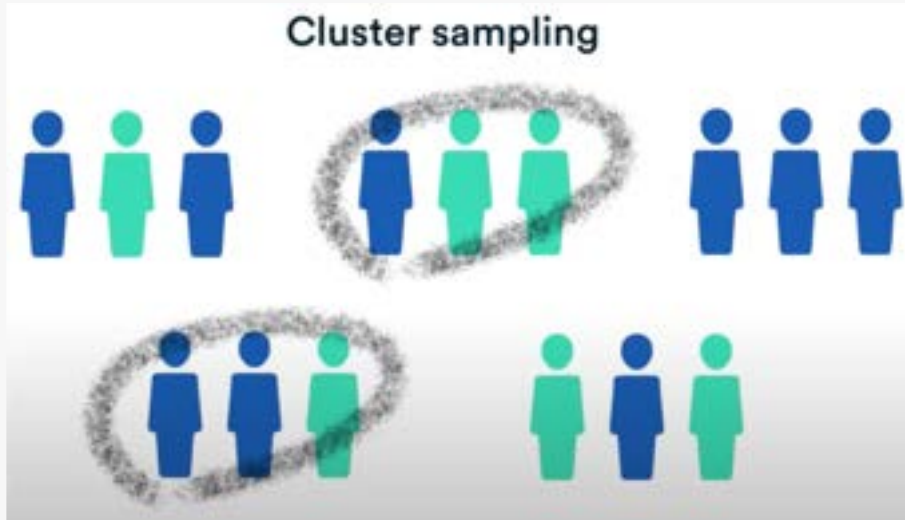
Sampling methods

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



Sampling methods

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



Probability sampling

Probability sampling requires that you have a list of all potential subjects or clusters in the population

- Difficult to achieve in practice, unless you're dealing with a very small and accessible population

Example: 9-th grade students in low-income areas of NY

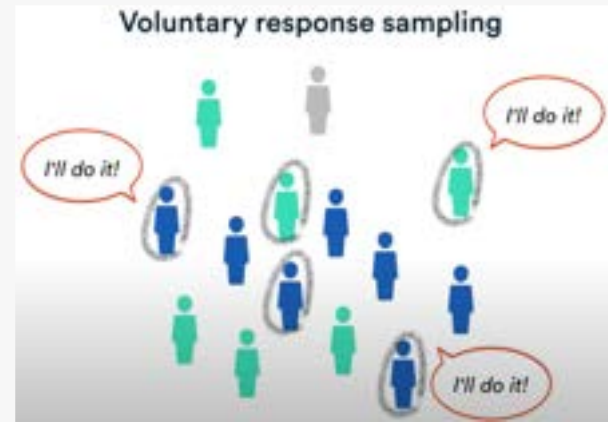
- Cluster sample: compile a list of all schools in low-income areas of NY and use a random number generator to select a sample of schools to collect data from



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



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

Your sample might differ in systematic ways from the population as a whole

Example: high-academic achievers might be more likely to volunteer to take part in an online teaching study than general students

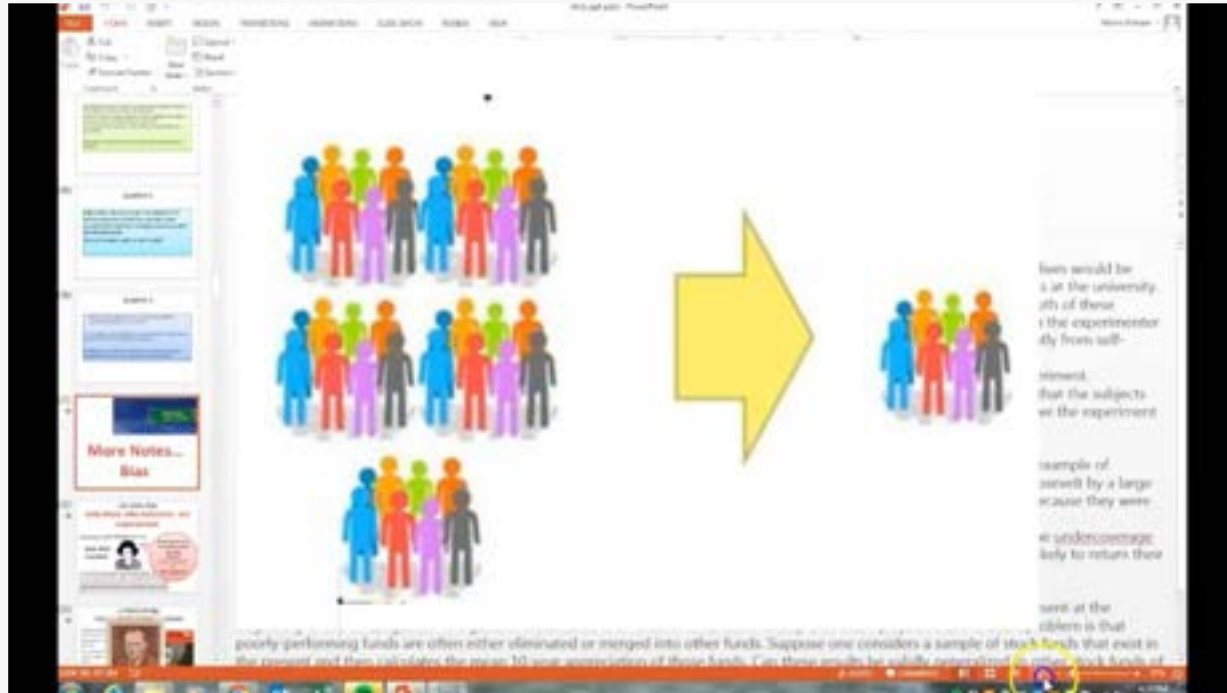
- Results will be biased towards students that have higher grades

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

Data Collection Bias



<https://www.youtube.com/watch?v=NJJdObWszAA>

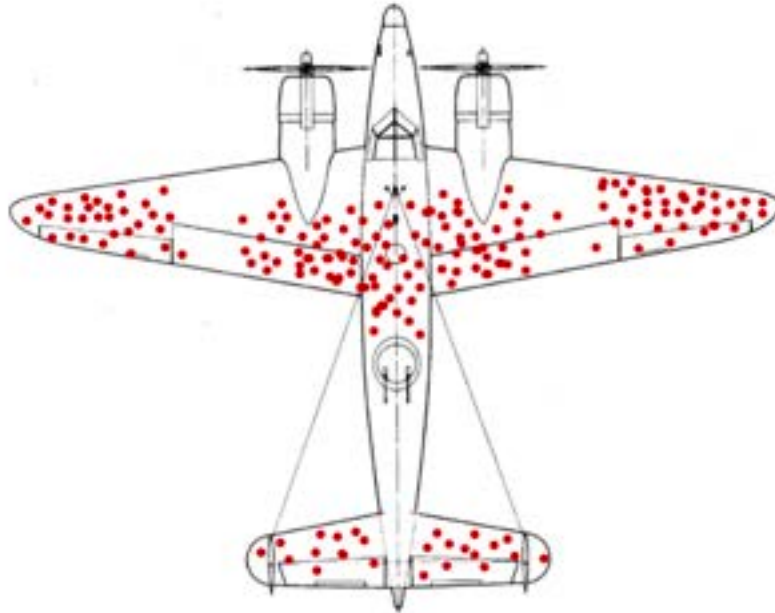
Selection Bias

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

Missing bullet holes (WWII)

During WWII, the Navy tried to determine where they needed to armor their aircraft to ensure they came back home. They ran an analysis of where planes had been shot up, and came up with this.

- Any issue?



Longevity Study from Lombard (1825)

Profession	Average Longevity

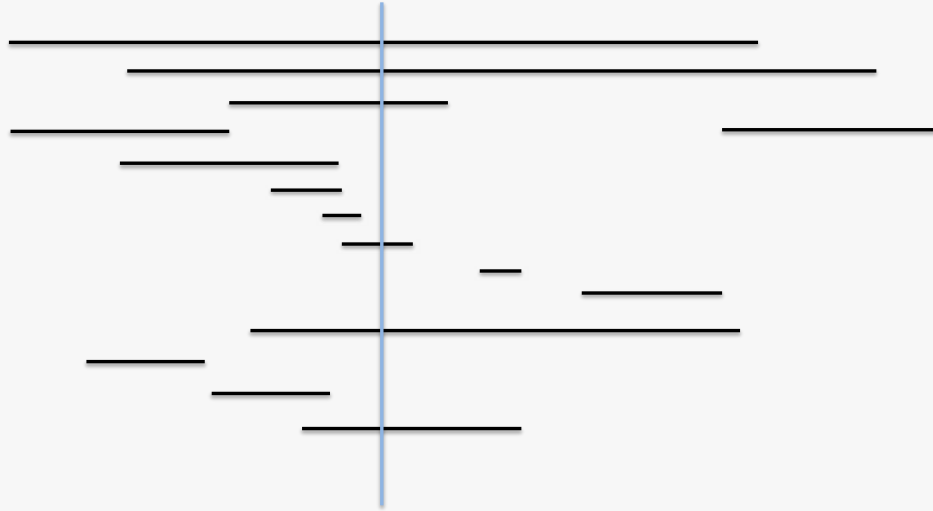
Sources: Lombard (1835), Wainer (1999), Stigler (2002)

“About 10 percent of the 1.6 million inmates in America’s prisons are serving life sentences; another 11 percent are serving over 20 years.”

source: [http://www.nytimes.com/2012/02/26/health/dealing-with-dementia-among-aging-criminals.html? pagewanted=all](http://www.nytimes.com/2012/02/26/health/dealing-with-dementia-among-aging-criminals.html?pagewanted=all)

Length-biasing Paradox

How would you measure the average prison sentence?



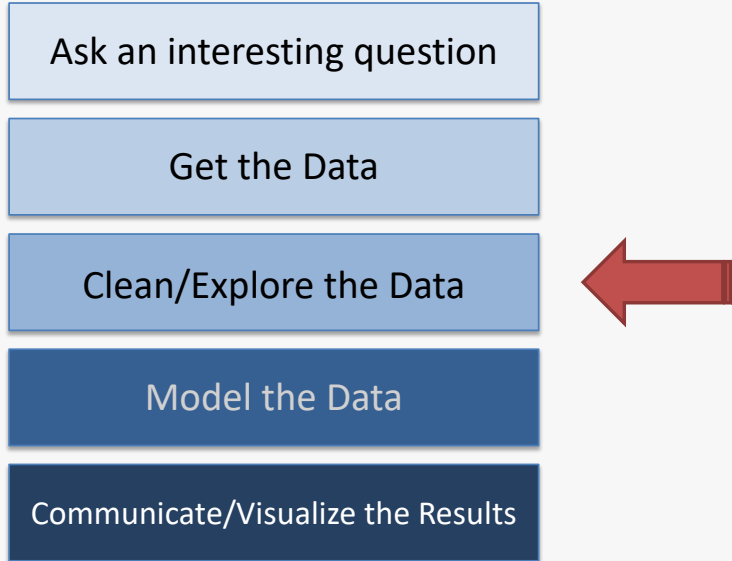
Bias in Data & AI



https://www.youtube.com/watch?v=gV0_raKR2UQ

What?

The Data Science Process



Clean/explore the Data

Which club will win the EPL?



Sample 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

 Quantitative Data

 Qualitative Data

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

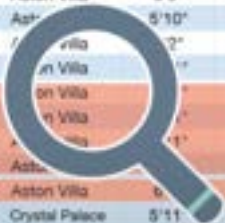
Player Name	Age	Club	Height	Weight	Foot	Joined
Pierre-Emerick Aubameyang	29	Arsenal	6'2"	176lbs	Right	Jan 31, 2018
Alexandre Lacazette	27	Arsenal	5'9"	161lbs	Right	Jul 5, 2017
Bernd Leno		Arsenal	6'3"	183lbs	Right	Jul 1, 2018
Henrikh Mkhitaryan	29	Arsenal	5'10"	165lbs	Right	Jan 22, 2018
Granit Xhaka	25	Arsenal	6'1"	161lbs	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	5'2"	150lbs	Right	Jan 31, 2017
Conor Hourihane	27	Aston Villa	5'11"	137lbs	Left	Jan 26, 2017
James Chester	29	Aston Villa	5'7"	174lbs	Right	Aug 12, 2016
James Chester	29	Aston Villa	5'7"	174lbs	Right	Aug 12, 2016
James Chester	29	Aston Villa	5'7"	174lbs	Right	Aug 12, 2016
James Chester	29	Aston Villa	5'7"	174lbs	Right	Aug 12, 2016
Jonathan Kodjia	2	Aston Villa	5'6"	170lbs	Right	Aug 30, 2018
Callum Wilson	26	Crystal Palace	5'11"	146lbs	Right	Jul 4, 2014

Bad quality data



Bad quality data can give misleading information

However, your domain knowledge on EPL says that the result looks inaccurate as Crystal Palace has never even finished in the top 4.



Player Name	Age	Club	Height	Weight	Foot	Joined
Pierre-Emerick Aubameyang	29	Arsenal	6'2"	175lbs	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	5'7"	155lbs	Right	Jan 31, 2017
Conor Hourihane	27	Aston Villa	5'10"	137lbs	Left	Jan 26, 2017
James Chester	29	Aston Villa	5'10"	174lbs	Right	Aug 12, 2016
James Chester	29	Aston Villa	5'10"	174lbs	Right	Aug 12, 2016
James Chester	29	Aston Villa	5'10"	174lbs	Right	Aug 12, 2016
James Chester	29	Aston Villa	5'10"	174lbs	Right	Aug 12, 2016
Jonathan Kodjia	22	Aston Villa	5'10"	170lbs	Right	Aug 30, 2016
Callum Wilson	26	Crystal Palace	5'11"	148lbs	Right	Jul 4, 2014

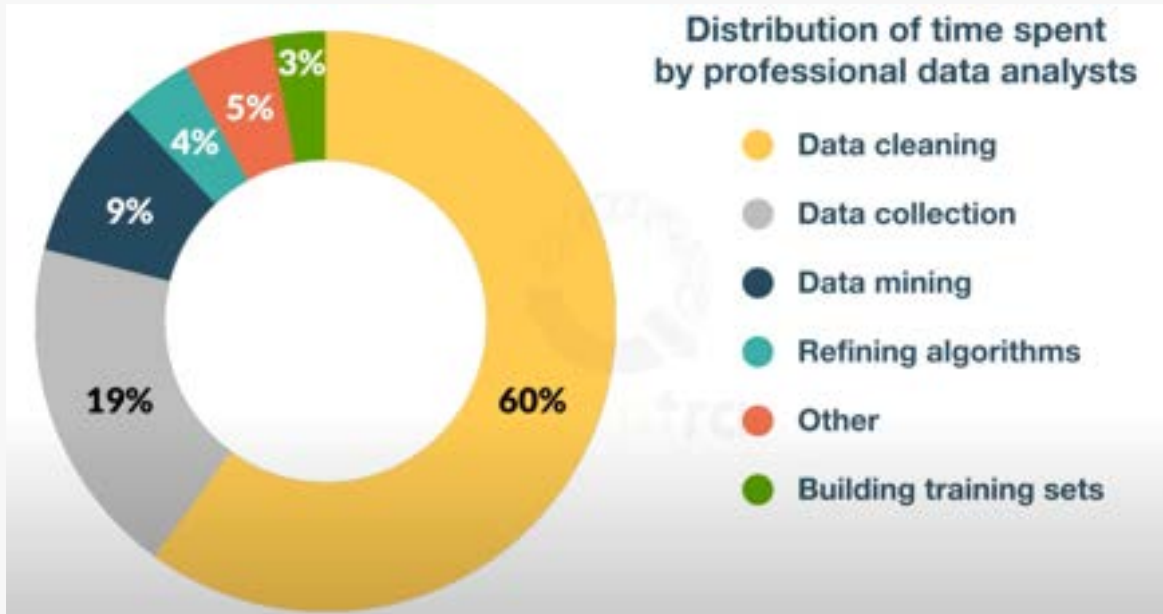
Bad quality data



Misleading information

Importance of Good Quality Data

a professional data scientist spends approximately 60% of his time ensuring that data is of high quality



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

Data Quality Issues (example from last lecture)

Question

Does age affect one's market value?

name	club	age	position	market value
Alexis Sanchez	Arsenal	28	LW	65
Mesut Ozil	Arsenal	28	AM	50
			GK	7
			RW	20
			CB	22

- Credible/Trustworthy?
- Possibly subjective market values?
- Sampled data

from www.transfermarkt.us

Data Quality Issues (example from last lecture)

Question

	age	page_views	fpl_value	fpl_points	market_value
count	461.000000				
mean	26.8047				
std	3.961892	931.803737	1.346695	53.113811	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

This seems abnormally low. Is it correct? Who is this?

Are the values reasonable? `DataFrame.describe()` ...

Data Quality Issues (example from last lecture)

	age	page_views	fpl_value	fpl_points	market_value
count					461.000000
mean					11.012039
std					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

This also seems suspicious. Is it correct? Who is this?

Are the values reasonable? `DataFrame.describe()` ...

Inspecting suspicious data

This accounts for both extreme values that we noticed. But, is this data **truly accurate**? It's worth validating online, elsewhere.

```
import pandas as pd
df = pd.read_csv("epl.csv")
df.iloc[df['market_value'].idxmin()]
```

name	Eduardo Carvalho
club	Chelsea
age	34
position	LW
position_cat	1
market_value	0.05
page_views	467
fpl_value	5
fpl_sel	0.10%
fpl_points	0
region	2
nationality	Portugal
new_foreign	0
age_cat	6
club_id	5
big_club	1
new_signing	1

Name: 109, dtype: object



Explore the Data

	age	page_views	fpl_value	fpl_points	market_value
count					
mean					
std					
min					
25%	24.000000	220.000000	4.500000	5.000000	0.000000
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

What is going on here?! Is someone actually paid that much more than others? And someone scores that much more?

from www.transfermarkt.us

Domain Knowledge

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



Domain Knowledge

Workblog

Census: Everybody's moving into their parents' basements

A [lock] [comment]

By Brad Plumer June 20, 2013 Follow Brad Plumer

Ever since the financial crisis hit, Americans have found it harder and harder to live on their own. According to a [new report \(pdf\)](#) from the Census Bureau, the number of "shared



Most Read [Business](#)

- 1 Here is everything we know about whether gentrification pushes poor people out 
- 2 Honey isn't as healthy as we think 

“The CPS counts students living in dormitories as living in their parents’ home.”

– Census Bureau, <http://www.census.gov/prod/2013pubs/p20-570.pdf>

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

Data Cleaning and exploration

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	28	Arsenal	5'9"	165lbs	Right	Jan 22, 2018
Granit Xhaka	25	Arsenal	6'1"	181lbs	Left	Jul 1, 2016
Shkodran Mustafi	26	Arsenal	6'4"	181lbs	Right	Aug 30, 2016
Jack Grealish	22	Aston Villa	5'10"	168lbs	Right	Jul 1, 2018
John McGinn	23	Aston Villa	5'9"	161lbs	Right	Jul 1, 2018
Anwar El Ghazi	23	Aston Villa	5'10"	168lbs	Right	Jul 1, 2018
Donor Hourihane	27	Aston Villa	5'10"	168lbs	Right	Jul 1, 2018
James Chester	29	Aston Villa	5'10"	168lbs	Right	Jul 1, 2018
James Chester	29	Aston Villa	5'10"	168lbs	Right	Jul 1, 2018
James Chester	29	Aston Villa	5'10"	168lbs	Right	Jul 1, 2018
James Chester	29	Aston Villa	5'10"	168lbs	Right	Jul 1, 2018
Jonathan Kodj	2	Aston Villa	5'10"	168lbs	Right	Jul 1, 2018
Dallum Wilson	26	Aston Villa	5'10"	168lbs	Right	Jul 1, 2018

Cleaning



Exploration

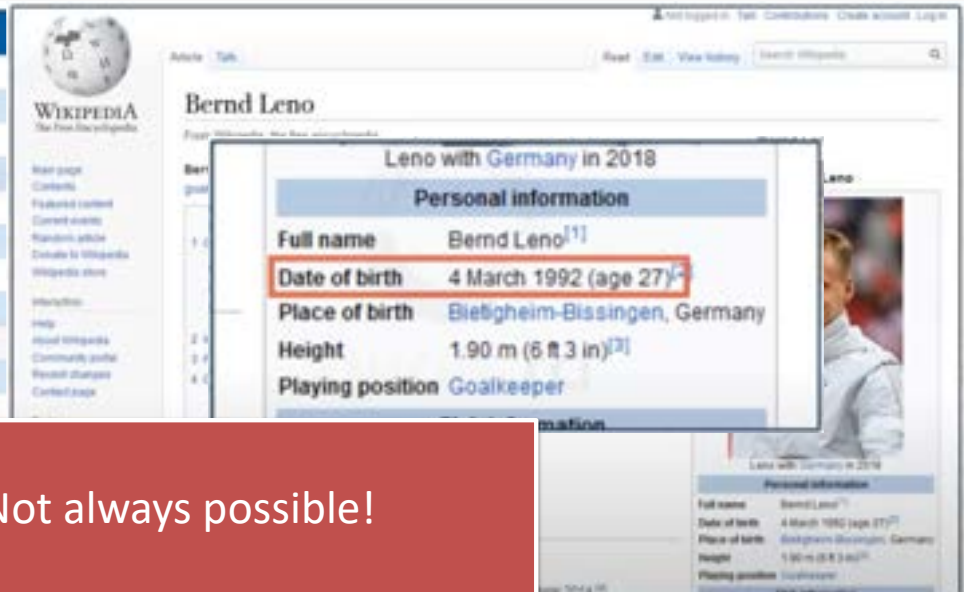
Explore and Ensure Data Quality

- 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

How to Fix Data Quality Issues?

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

Player Name	Age	Club
Pierre-Emerick Aubameyang	29	Arsenal
Alexandre Lacazette	27	Arsenal
Bernd Leno		Arsenal
Henrikh Mkhitaryan	29	Arsenal
Granit Xhaka	25	Arsenal
Shkodran Mustafi	26	Arsenal
Jack Grealish	22	Aston Villa
John McGinn	23	Aston Villa
Anwar El Ghazi	23	Aston Villa
Conor Hourihane	27	Aston Villa
James Chester	29	Aston Villa
James Chester	21	
James Chester	21	
James Chester	21	
Jonathan Kodjia	2	
Callum Wilson	21	



The screenshot shows the Wikipedia page for Bernd Leno. A red box highlights the 'Date of birth' field, which is '4 March 1992 (age 27)'. Another red box highlights the 'Full name' field, which is 'Bernd Leno^[1]'. The page also shows a table of 'Leno with Germany in 2018' and a 'Personal information' section.

Leno with Germany in 2018	
Personal information	
Full name	Bernd Leno ^[1]
Date of birth	4 March 1992 (age 27)
Place of birth	Bietigheim-Bissingen, Germany
Height	1.90 m (6 ft 3 in) ^[2]
Playing position	Goalkeeper

Not always possible!

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, 2018
Calum 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.

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

Explore the Data

name	club	age	position	market value
Alexis Sanchez	Arsenal	28	LW	65
Mesut Ozil	Arsenal	28	AM	50
Petr Cech	Arsenal	35	GK	7
Theo Walcott	Arsenal	28	RW	20
Laurent Koscielny	Arsenal	31	CB	22

Are the values reasonable? `DataFrame.describe()` ...

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

Are the values reasonable? `DataFrame.describe()` ...



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

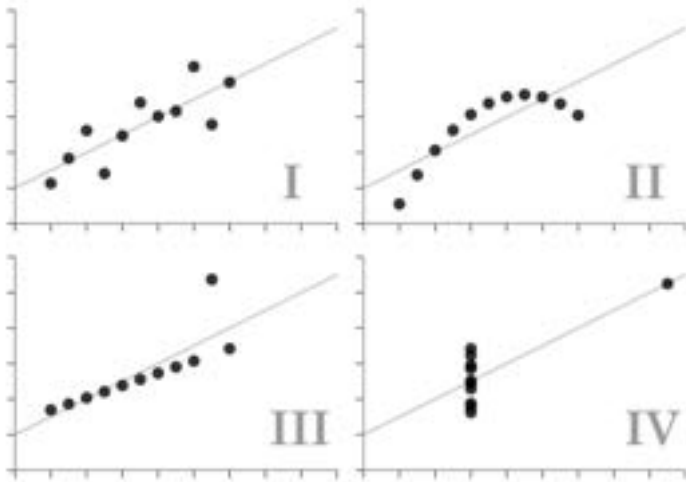
Summary statistics can only reveal so much

Visualization



Anscombe's Quartet

Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are clearly different, and visually distinct.

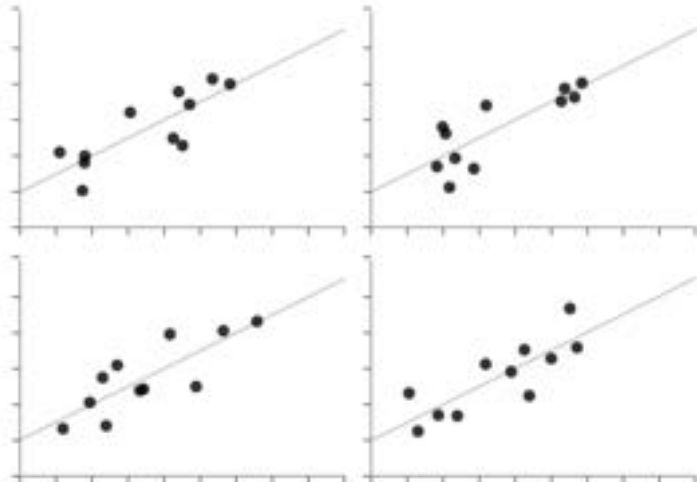


Same stats do not imply same graphs



Unstructured Quartet

Each dataset here also has the same summary statistics. However, they are not clearly different or visually distinct.



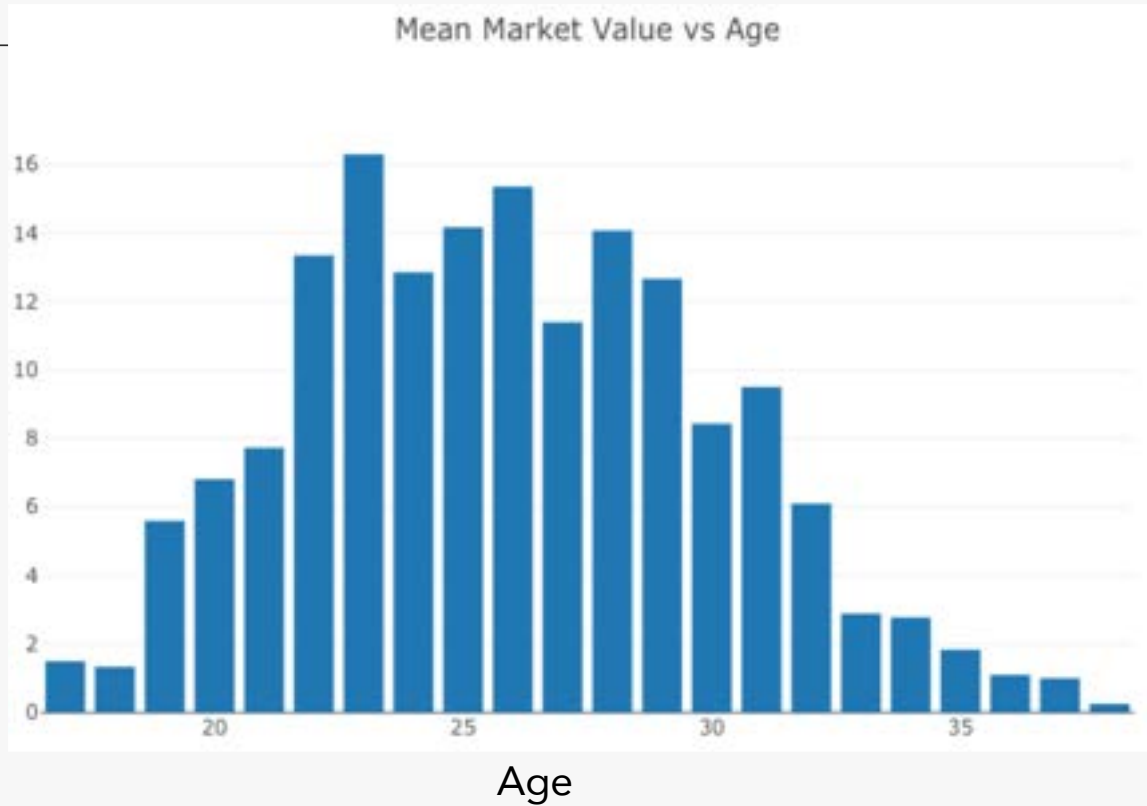
Same graphs do not imply same stats

Visualization

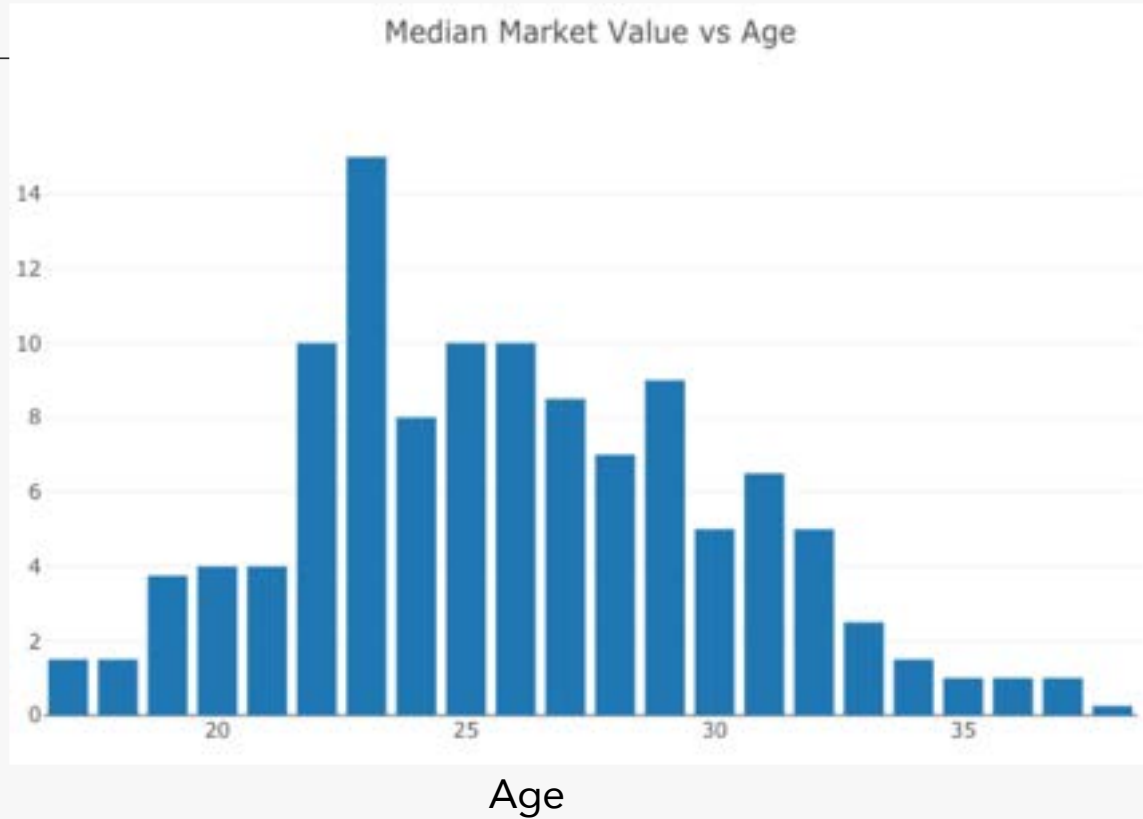
Visualization is incredibly important,
both for EDA and for communicating
your results to others.

Visualization packages will be used
throughout the semester.

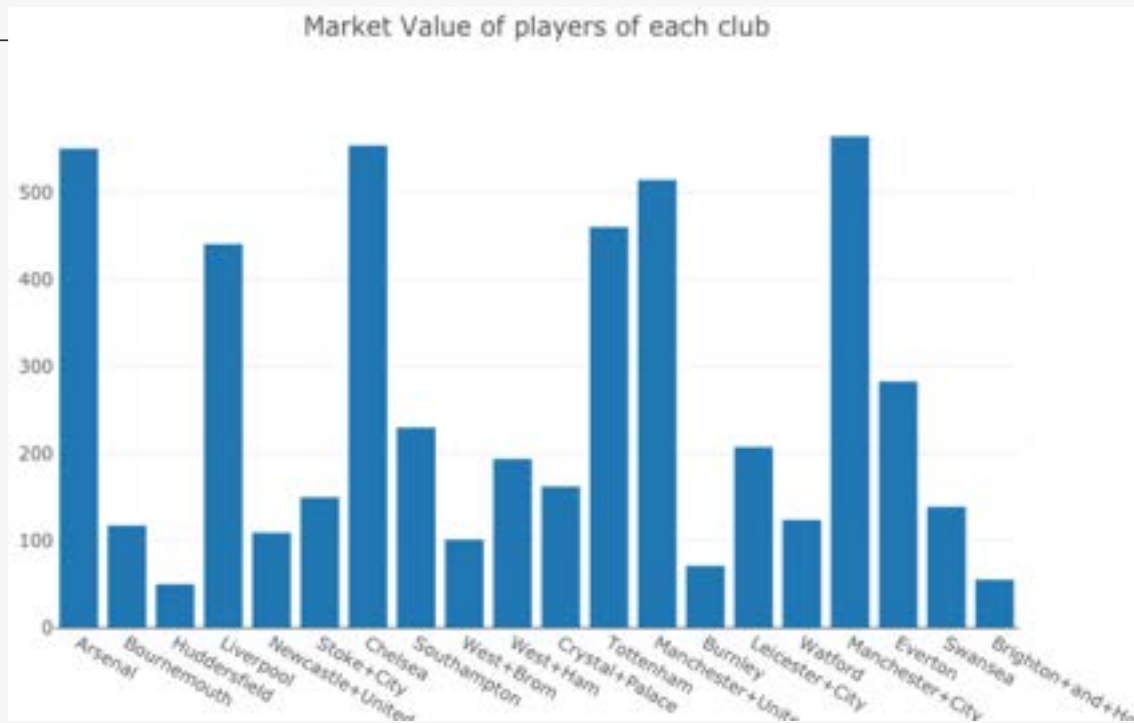
Mean Market Value



Median Market Value

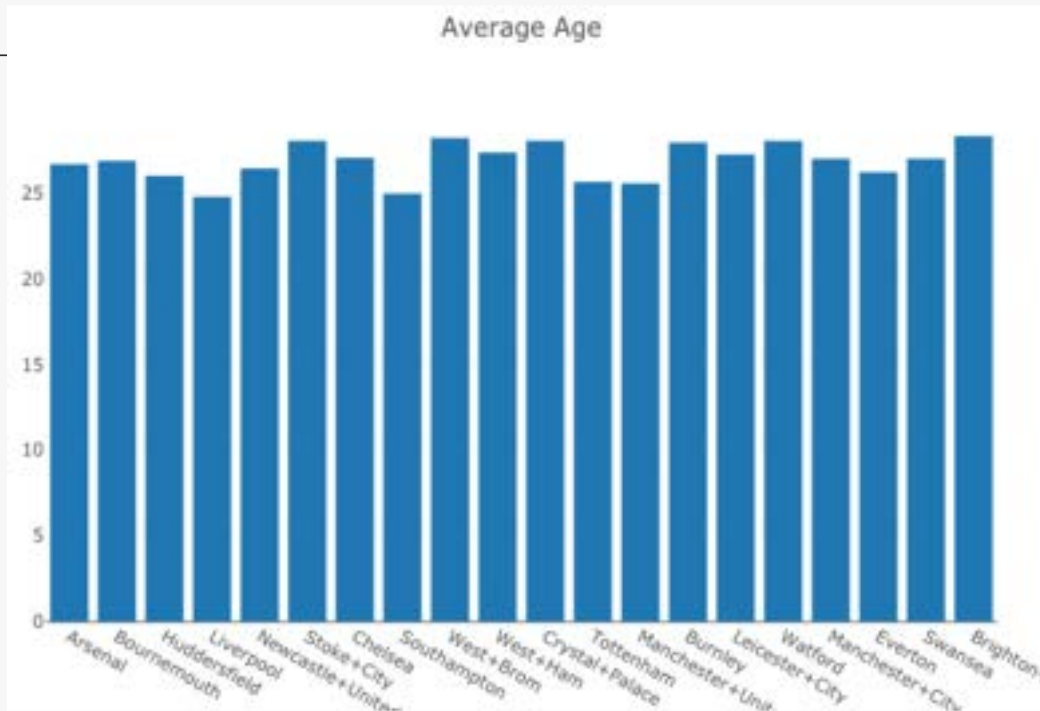


Market Value



Club

Market Value



Club

Ready to Model the Data!

The Data Science Process

Ask an interesting question

Get the Data

Clean/Explore the Data

Model the Data

Communicate/Visualize the Results

