Welcome to Week #2!

Bureaucracy, Redux

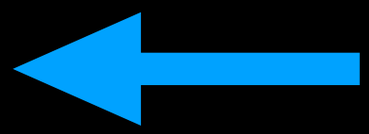
How to get help

Jill P. Naiman - jnaiman@illinois.edu

- **Office Hours: Tuesday 7-8pm, Wednesdays 8-9pm after class; other times by request**

Assignments

Weekly homework	55%
Midterm exam	20%
Final exam	25%



See Syllabus for late HW policies, but in general:

1. **we will drop your lowest HW score.**
2. **You are allowed 3 late HWs (due date Sundays)**
3. **HW #3 and beyond will be "self graded"**

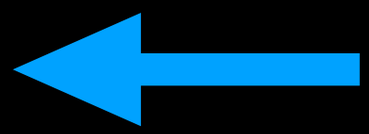
How to get help

Jill P. Naiman - jnaiman@illinois.edu

- **Office Hours: Thursdays 3-4pm, Slack; other times by request**

Assignments

Weekly homework	55%
Midterm exam	20%
Final exam	25%



Midterm/Final — see Moodle for more details, but generally:

1. **Pick a data set and do statistical "stuff" with it**
2. **Build predictive models with this data**

General Survey Results

Most of programming experience at about 1-3 years

Exercises will have wide range of options for different levels.

Wide range of data interests

Will use “real” datasets, wide variety of topics (feel free to suggest some!)

People are excited about learning more about programming

We will do a lot of collaborative coding in class, different options for different levels.

Week	Topic	Reading
1	• Data, Models, and Information • Elementary statistics: Definitions • Overview of R	OIS 1 (ISL 1)
2	• Elementary statistics: Applications & Plots	OIS 1 (ISL 1)
3	• Introduction to data analysis with R • Review of tabular and graphical displays of data	ITR 1, 2, 5, 6, 7, 12
4	• Random variables: expectation and variance • Joint and conditional probability • Bayes rule	OIS 2
5	• Random variables: distributions (normal, binomial, poisson)	OIS 3

Definitions, basic concepts, R practice

HW and Exam Formats

File name structure: lastname-first-module.ext
(e.g, naiman-jill-assignment1.pdf).

The submission must include:

1) A narrative document as a PDF file (to be read by a human). To preserve the natural flow of the narrative, figures (e.g., screenshots, code snippets) and tables should be embedded into the document near their first mention. Any supplementary files containing R programs or data should be referenced in the text and separately uploaded.

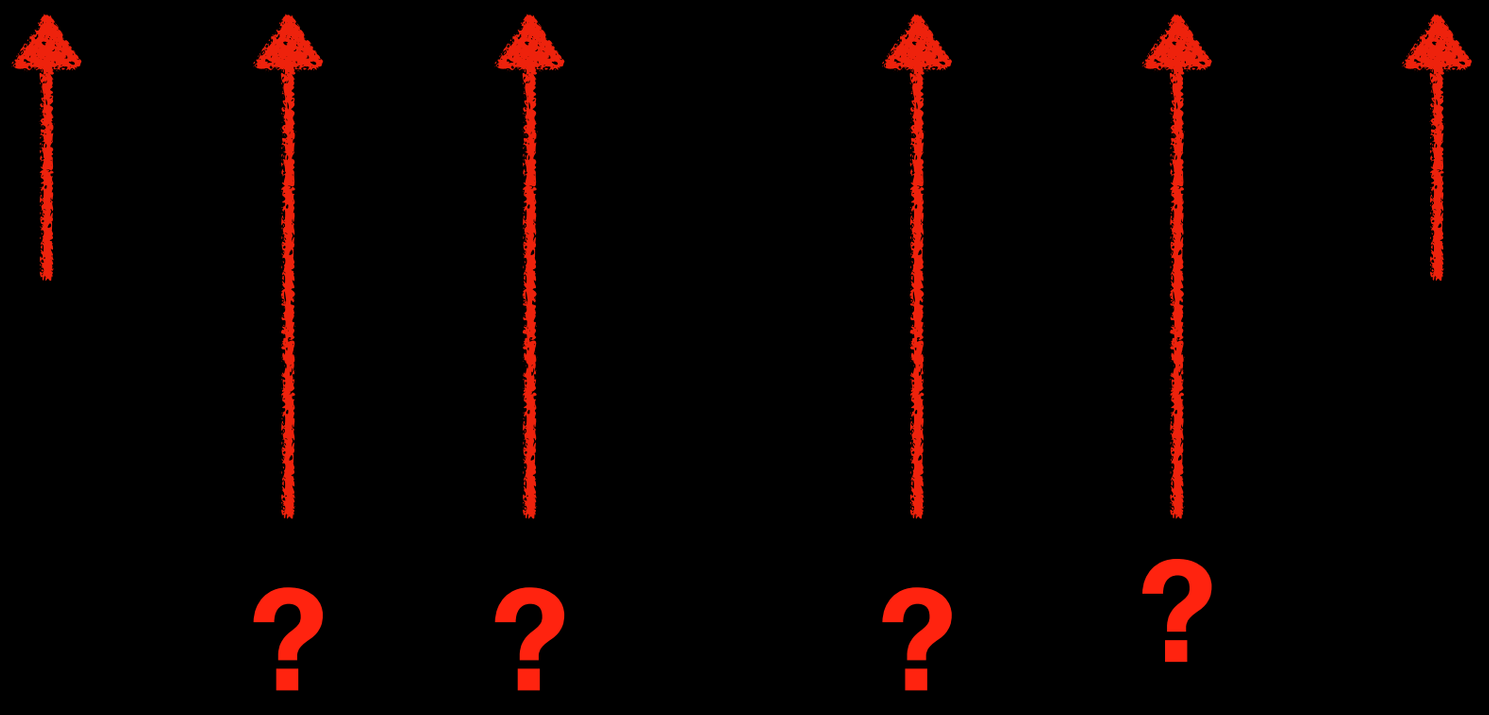
AND

2) All R code as separate files with an .R extension (to be read by a computer).

ALSO: make sure you include any files needed to run your R-script (data files for example)

Last time…

```
summary(after)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  39.00   41.25   44.50   44.21   46.75   50.00
```



Last time…

Summary Statistic Definitions!

Mean (Sample) = sum of all data values divided by number of data points

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

$$\text{Symbolically, } \bar{x} = \frac{\sum x}{n}$$

where $\bar{x}$ (read as 'x bar') is the mean of the set of x values,
$\sum x$ is the sum of all the x values, and
n is the number of x values.

(note - only works with "numerical" data types… more about data types later)

Median = if we order the data from smallest to largest, this is the observation in the middle (splits the data in 2 halves)

First/Third Quartiles = where 25% of the data falls below/above

Standard Deviation = this is the square root of the variance, where the variance is roughly the average distance of data values from the mean

$$\text{Standard Deviation (sample)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \text{mean})^2}{n-1}}$$

Last time…

Summary Statistic Definitions!

Mean (Sample) = sum of all data values divided by number of data points

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

$$\text{Symbolically, } \bar{x} = \frac{\sum x}{n}$$

where $\bar{x}$ (read as 'x bar') is the mean of the set of x values,
$\sum x$ is the sum of all the x values, and
n is the number of x values.

(note - only works with "numerical" data types… more about data types later)

Median = if we order the data from smallest to largest, this is the observation in the middle (splits the data in 2 halves)

First/Third Quartiles = where 25% of the data falls below/above

Standard Deviation = this is the square root of the variance, where the variance is roughly the average distance of data values from the mean

$$\text{Standard Deviation (sample)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \text{mean})^2}{n-1}}$$

Let’s play with *your* survey data! (Anonymized of course!)

IS 507 AO FA20: Data, Stat, Info (Naiman, J)

Dashboard / My courses / IS 507 AO FA20 / August 24 - August 30 [Week 1] / Intake Survey / View All Responses / Summary / View Default order

Advanced settings | Questions | Feedback | Preview | View All Responses | Non-respondents

Summary | List of responses

To R!

View Default order | Ascending order | Descending order | Delete ALL Responses | Download as CSV

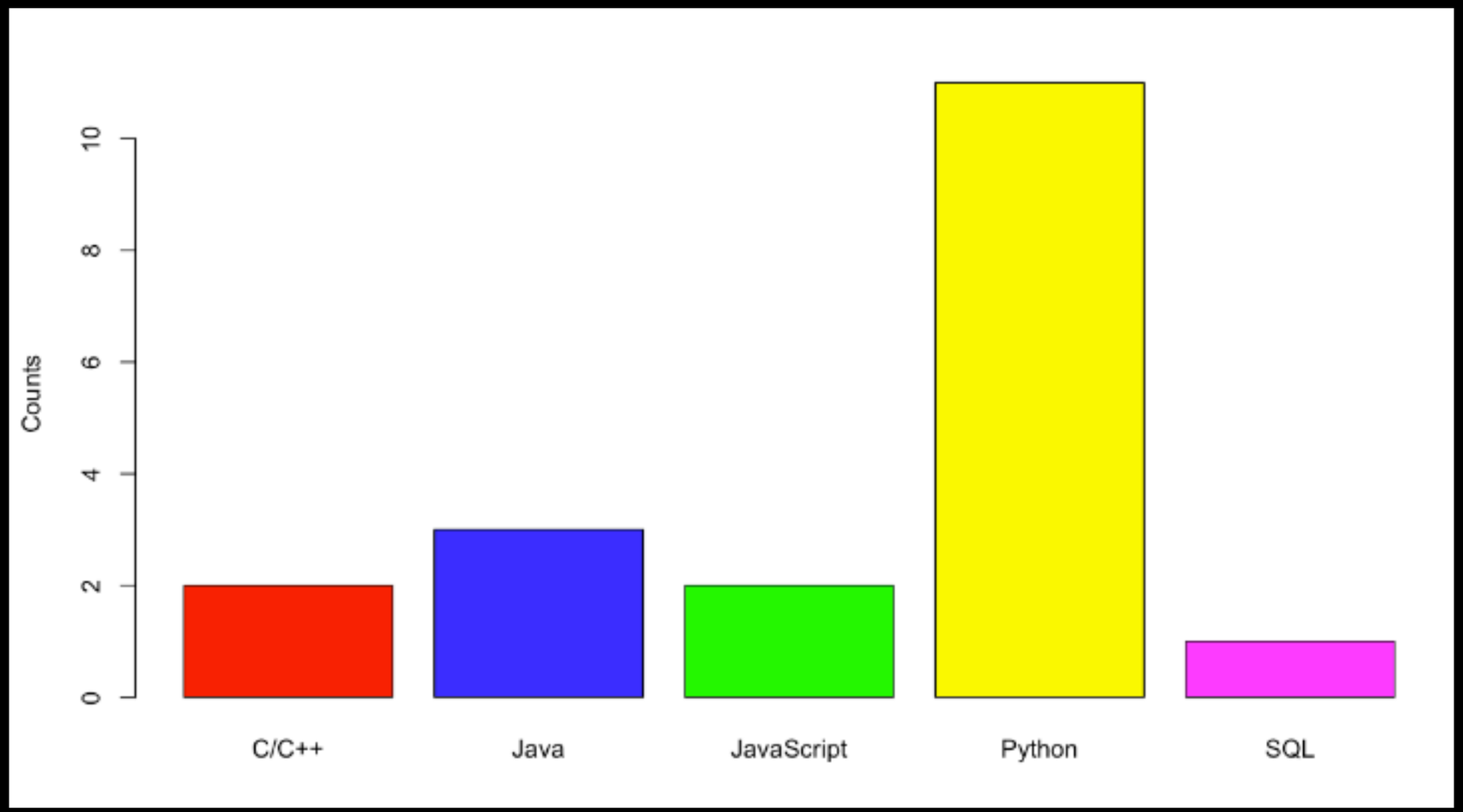
View All Responses. **All participants**. View Default order

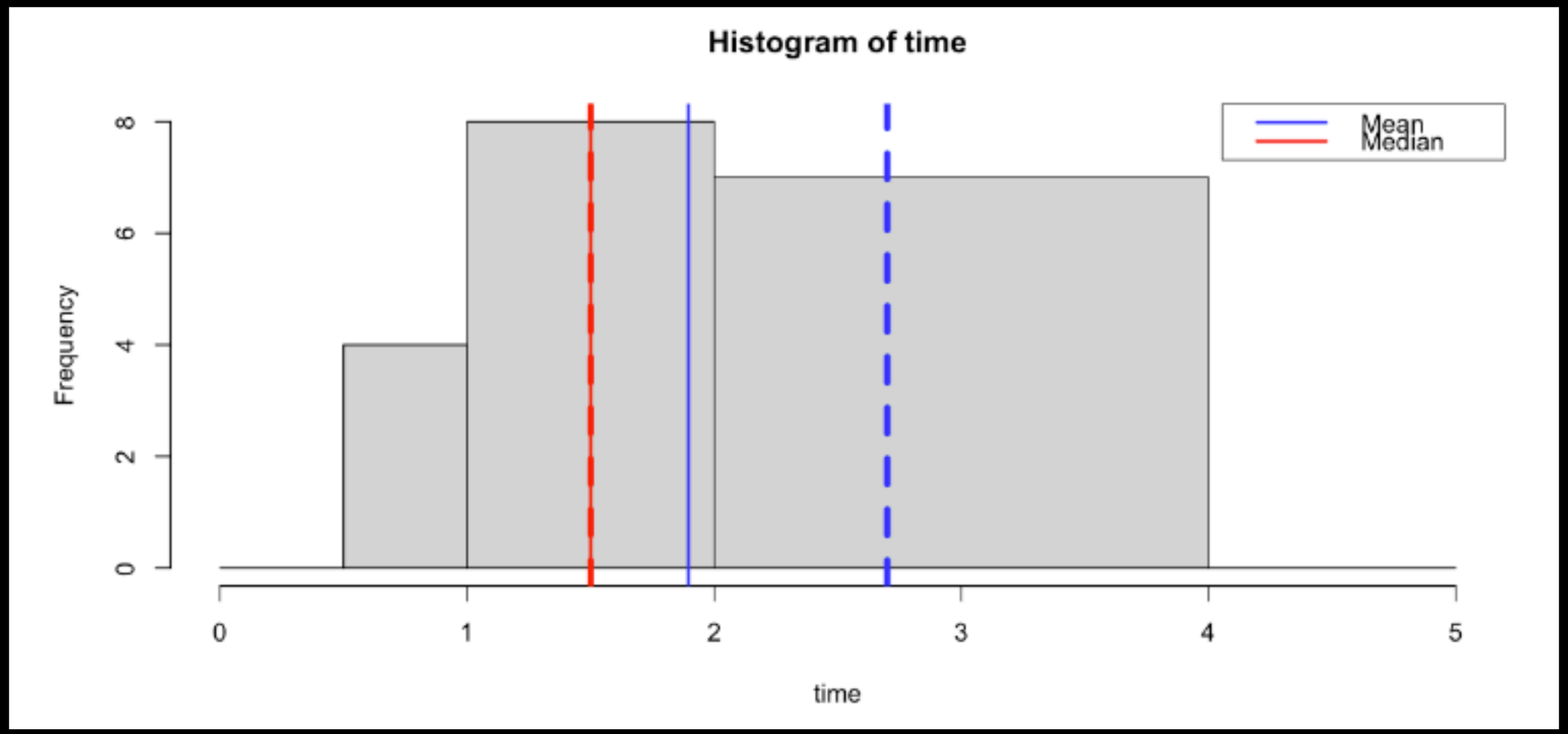
Intake Survey

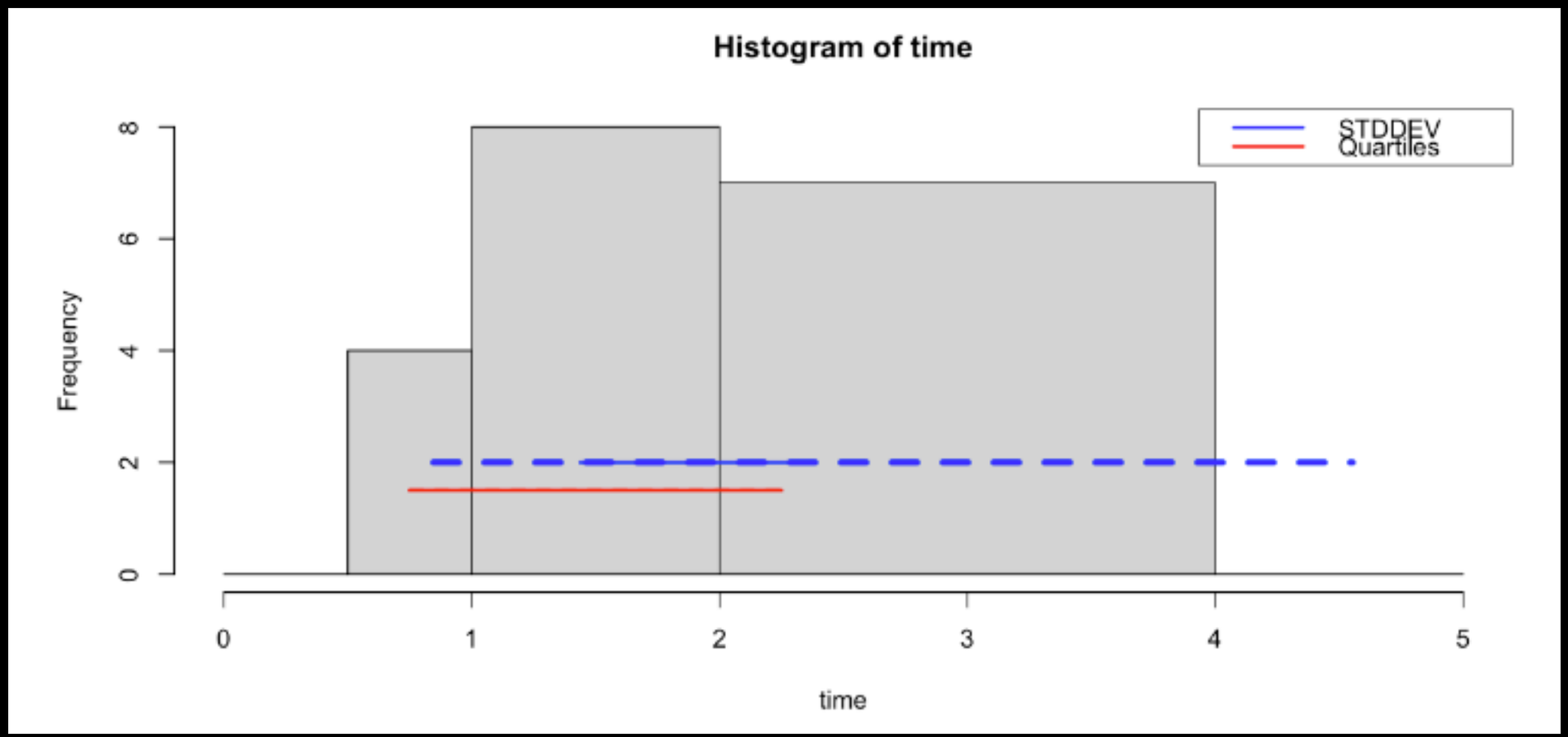
1 How would you gauge your familiarity with Statistics?

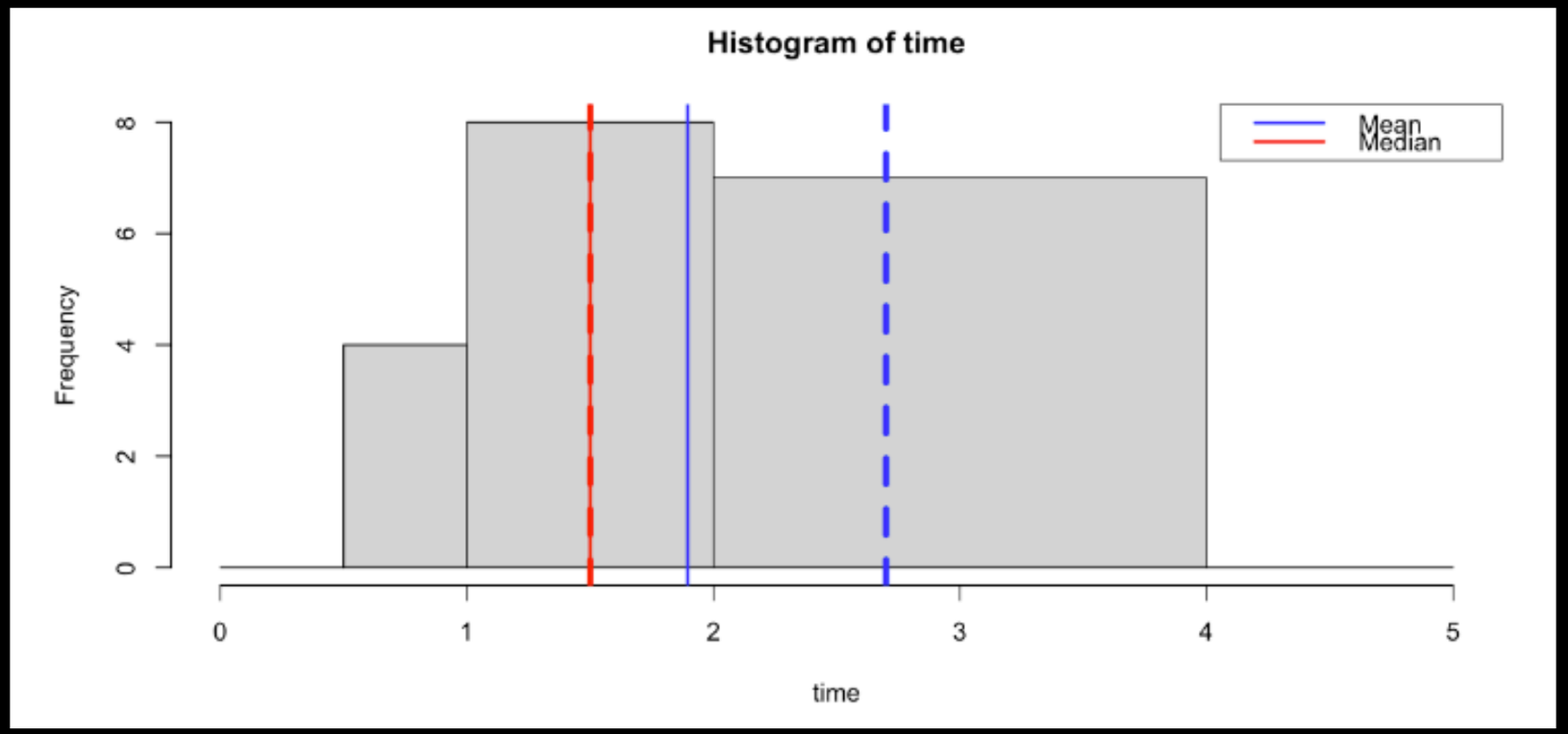
Average rank (and average values)

None	A bit	Medium	Use stats a lot	Expert

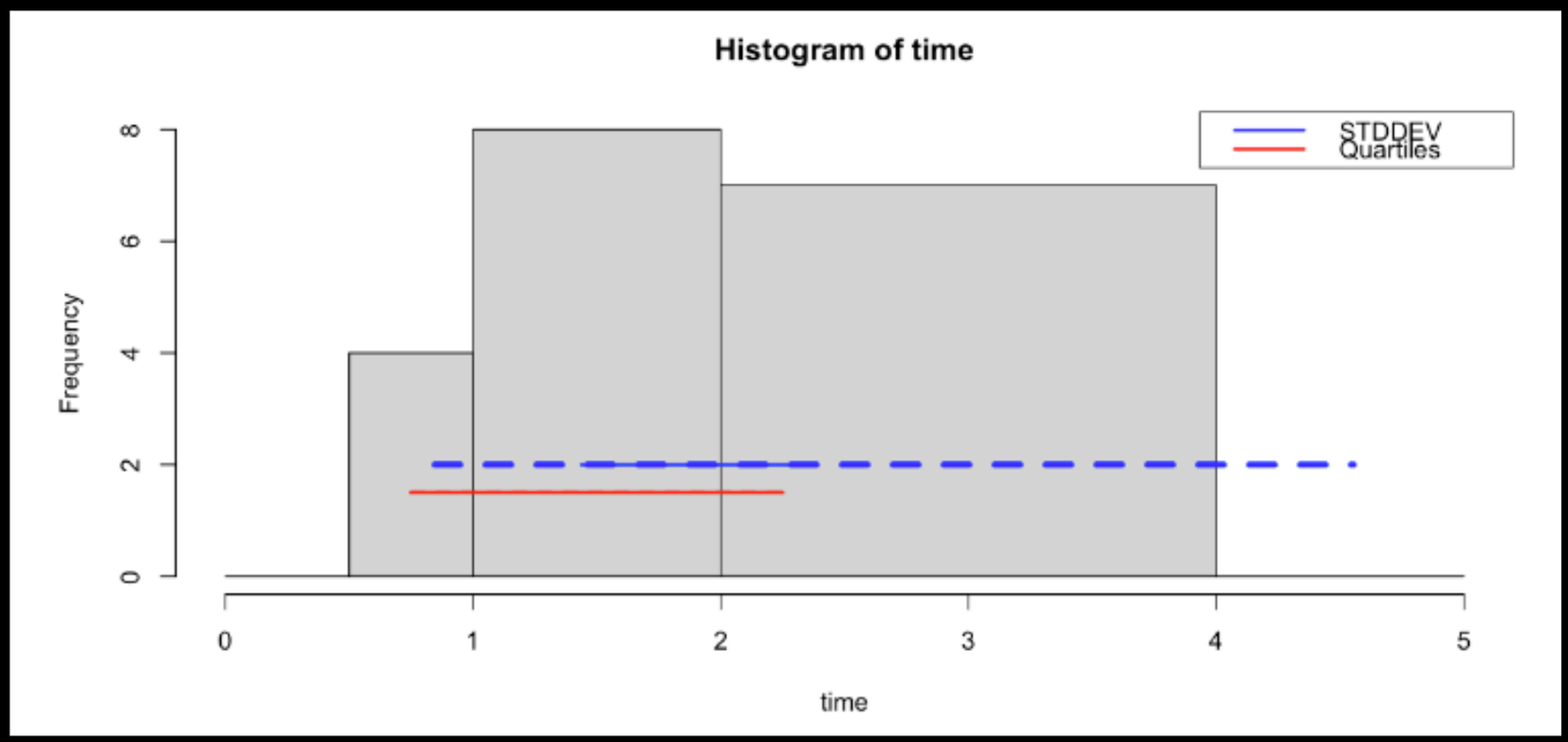

Counts
0
2
4
6
8
10
C/C++
Java
JavaScript
Python
SQL

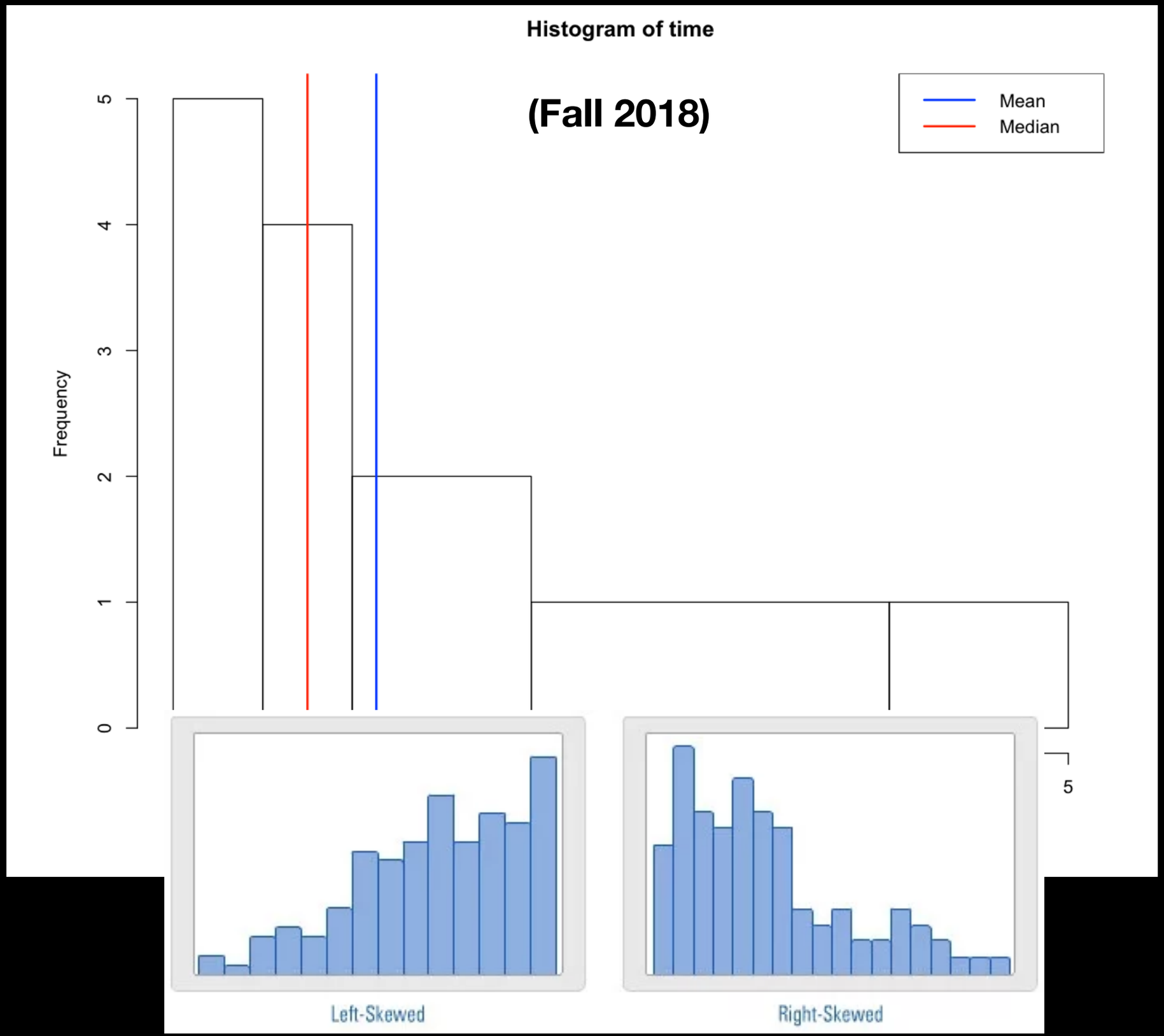

Histogram of time
Mean
Median
Frequency
0
2
4
6
8
0
1
2
3
4
5
time

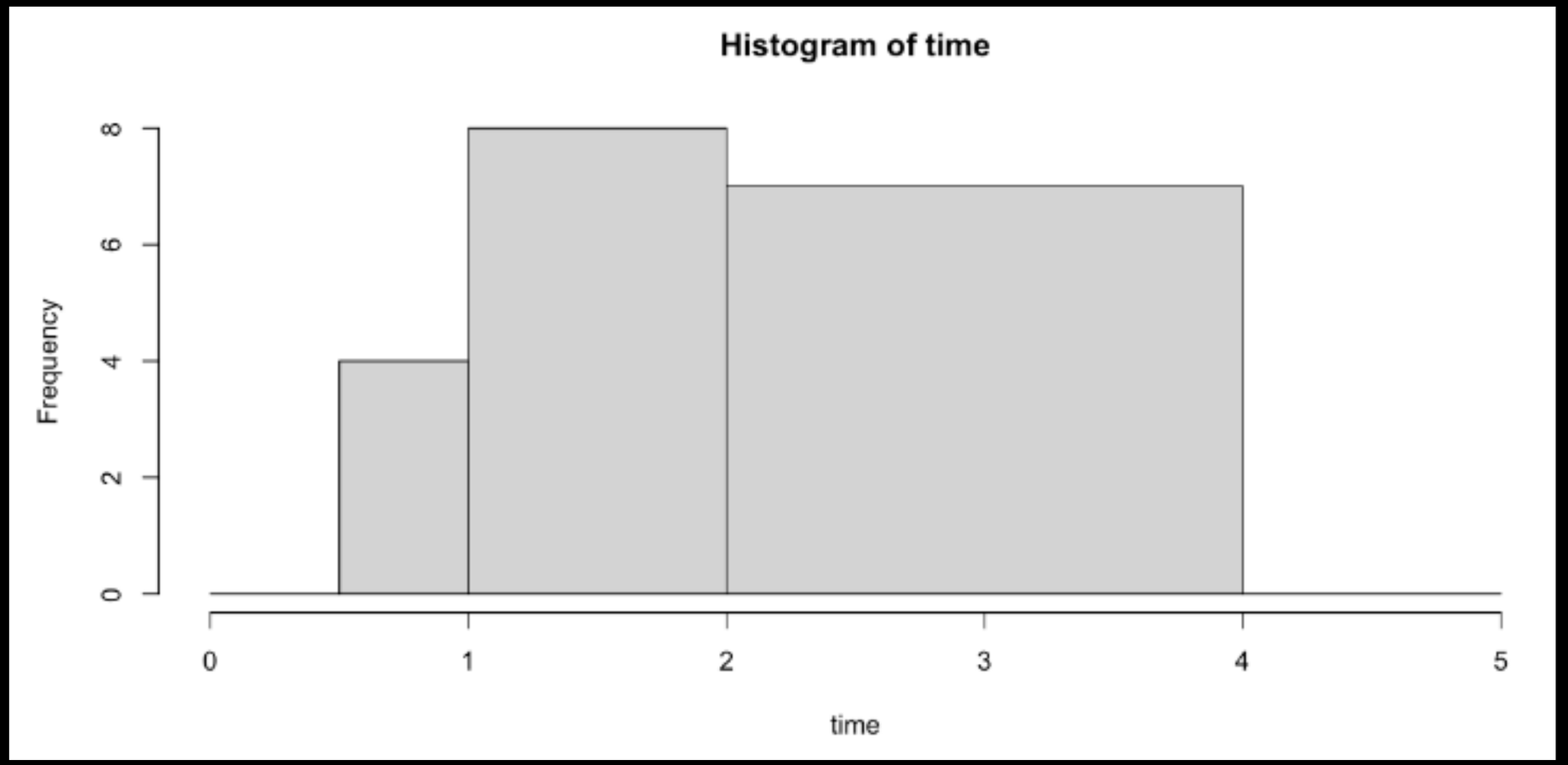

Histogram of time
STDDEV
Quartiles
Frequency
0
2
4
6
8
0
1
2
3
4
5
time

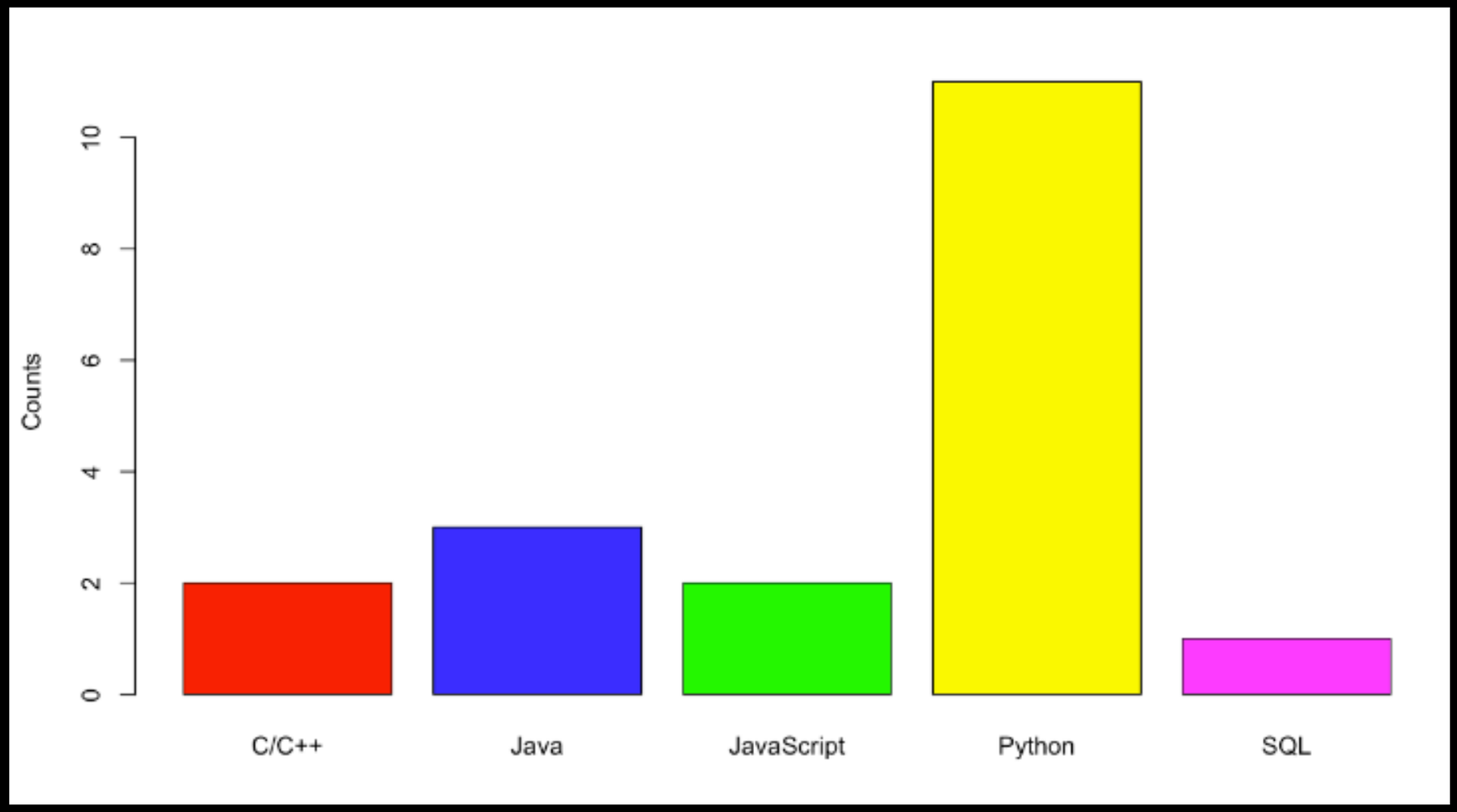


Median and quartile ranges are “robust” statistics - i.e. not as sensitive to variability




Histogram of time
(Fall 2018)
Mean
Median
Frequency
0
1
2
3
4
5
5
Left-Skewed
Right-Skewed


Histogram of time
Frequency
0
2
4
6
8
0
1
2
3
4
5
time


Counts
0
2
4
6
8
10
C/C++
Java
JavaScript
Python
SQL

Variables

Data Matrix

Case

	familiarity.with.stats	programming.language	time.programming
1	3	Python	2–4 years
2	2	JavaScript	2–4 years
3	2	Java	2–4 years
4	3	Python	1–2 years
5	2	Python	1–2 years
6	2	Java	2–4 years
7	3	Python	1–2 years
8	3	Java	1–2 years
9	3	Python	Between 6 months to 1 year
10	3	Python	1–2 years

Variable Types

	familiarity.with.stats	programming.language	time.programming
1	3	Python	2-4 years
2	2	JavaScript	2-4 years
3	2	Java	2-4 years
4	3	Python	1-2 years
5	2	Python	1-2 years
6	2	Java	2-4 years
7	3	Python	1-2 years
8	3	Java	1-2 years
9	3	Python	Between 6 months to 1 year
10	3	Python	1-2 years

Variable Types

NOT Numerical

Numerical (+/-, means, etc)

	familiarity.with.stats	programming.language	time.programming
1	3	Python	2-4 years
2	2	JavaScript	2-4 years
3	2	Java	2-4 years
4	3	Python	1-2 years
5	2	Python	1-2 years
6	2	Java	2-4 years
7	3	Python	1-2 years
8	3	Java	1-2 years
9	3	Python	Between 6 months to 1 year
10	3	Python	1-2 years

discrete numerical

vs.

continuous numerical

Variable Types

Numerical (+/-, means, etc)

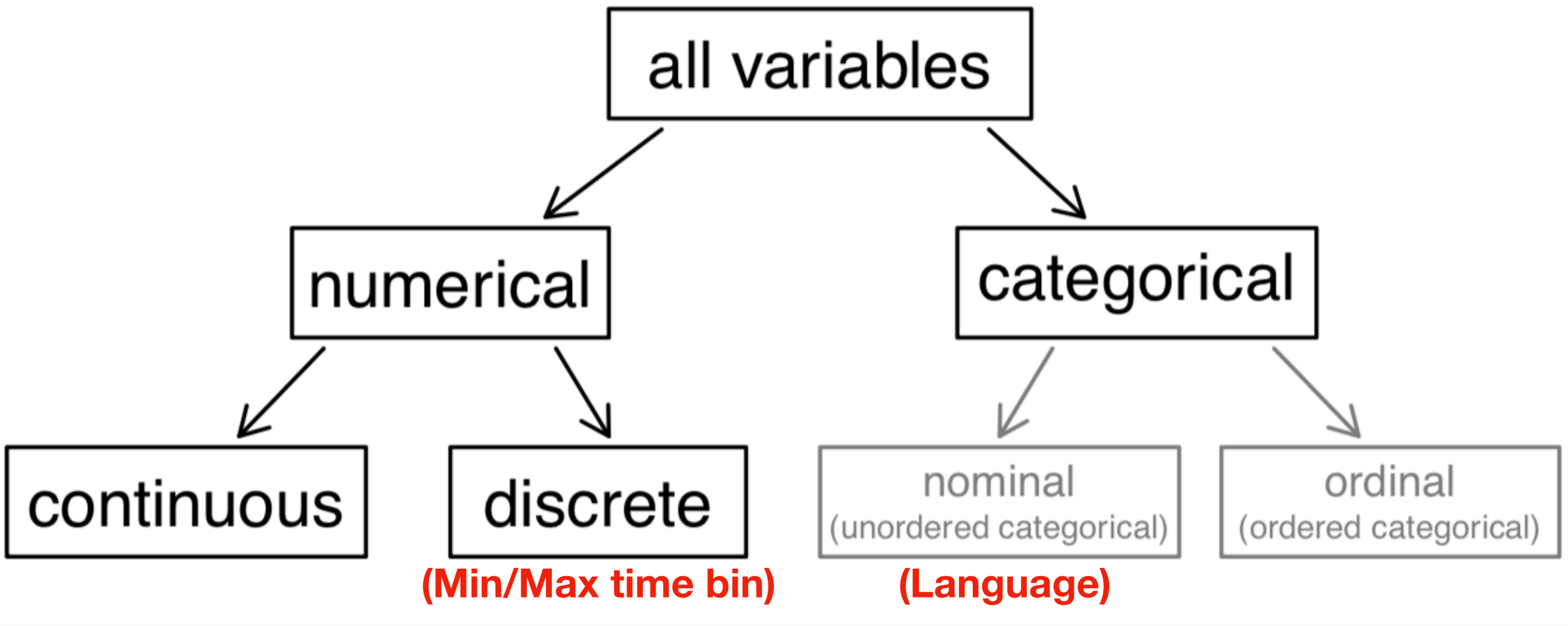
Categorical (levels)

	familiarity.with.stats	programming.language	time.programming
1	3	Python	2–4 years
2	2	JavaScript	2–4 years
3	2	Java	2–4 years
4	3	Python	1–2 years
5	2	Python	1–2 years
6	2	Java	2–4 years
7	3	Python	1–2 years
8	3	Java	1–2 years
9	3	Python	Between 6 months to 1 year
10	3	Python	1–2 years

nominal categorical

vs.

ordinal categorical


all variables
numerical
categorical
continuous
discrete
nominal
(unordered categorical)
ordinal
(ordered categorical)
(Min/Max time bin)
(Language)

Population of interest

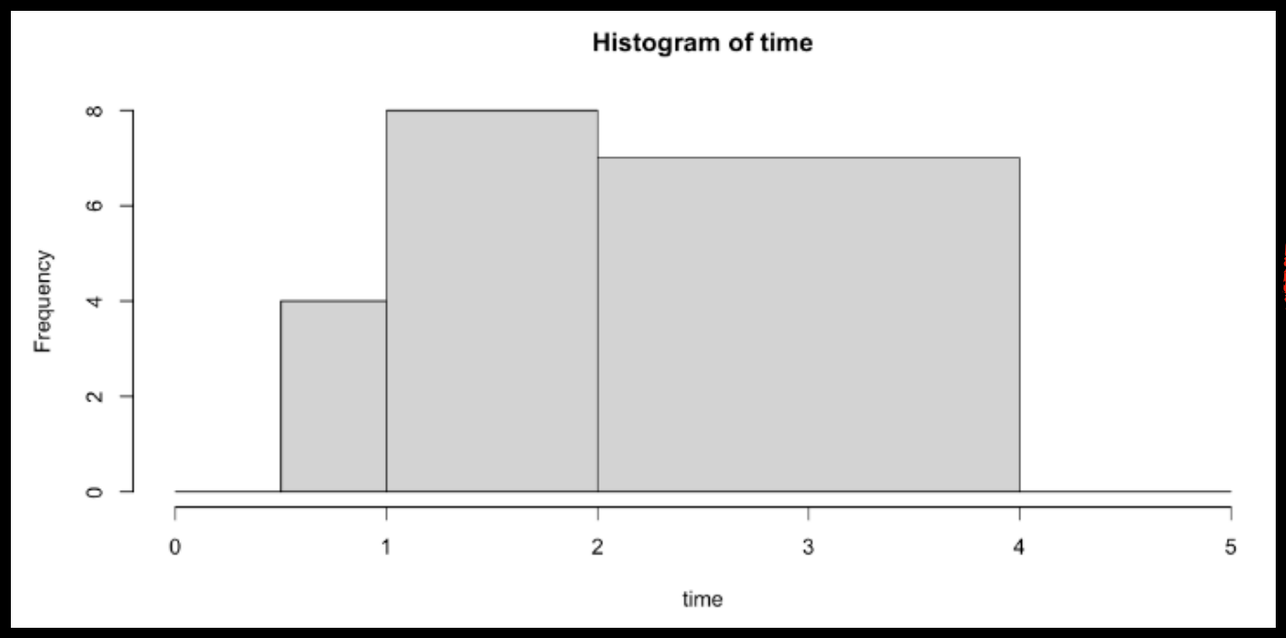
Can we use this data to ask the following questions about the typical iSchool student?

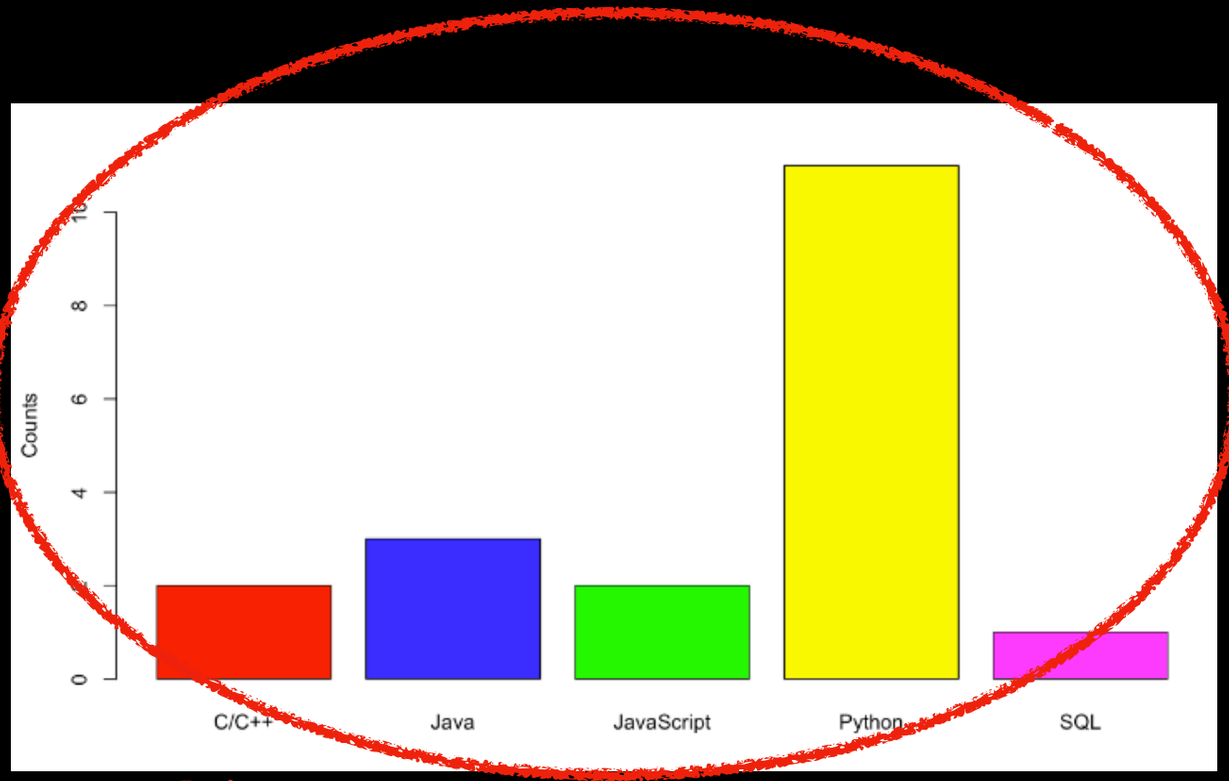
1. **Can we say typical iSchool student codes in python?**
2. **Can we say typical iSchool student has been programming for <1 year? >2 years?**

Why or why not?

Share: biases of sample?

Our survey was just a **sample**





In fact a **convenience sample** (not a **random sample**)

Response	
Medical studies (e.g. drug trials, treatment trials)	
Scientific studies (e.g. statistical astronomy, physics, genomics)	53%
Psychological studies (e.g. behavior studies, group behavior studies)	58%
Financial (e.g. stock market)	74%
Language studies (e.g. content of emails, books)	32%
I'm not sure yet	21%
Other	32%

Response	Average
Mac OSX	21%
Linux	5%
Windows 10	74%

Real Data! (ooooo!)

Dataset #1:

http://data.un.org/

to

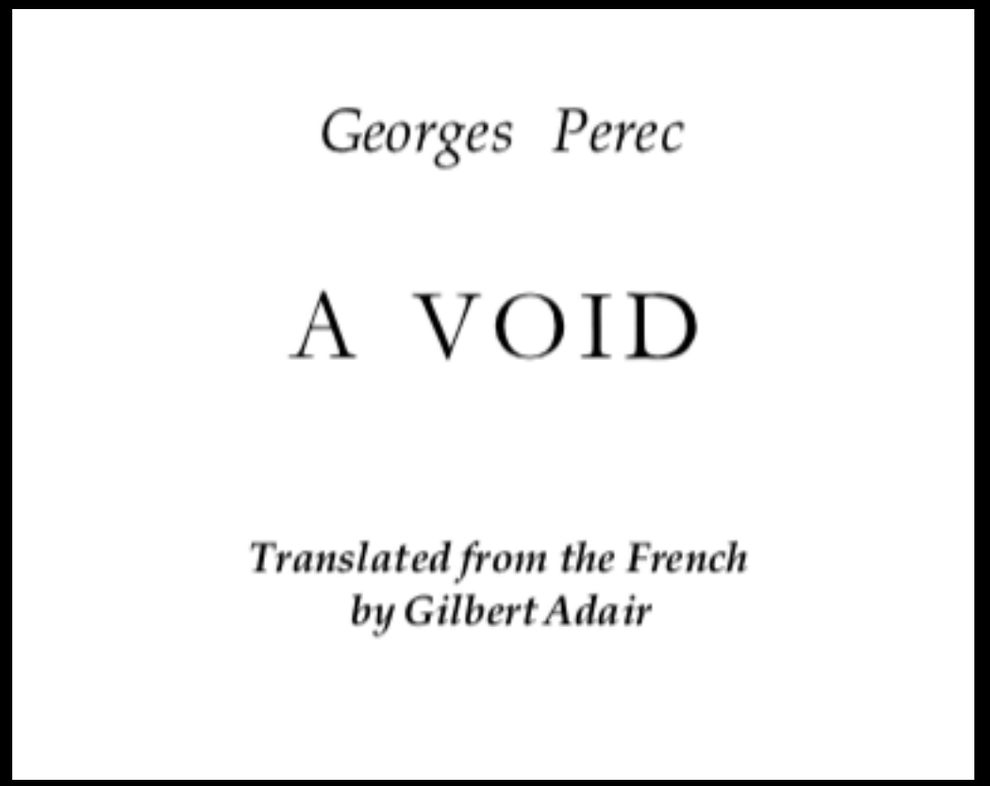
http://data.un.org/Explorer.aspx

to

Commodity trade stats =>
Fish, crustaceans, mollusks, etc

Dataset #2:

Excerpt of the
Introduction of “A Void”
by Georges Perec



Dataset #1

http://data.un.org/Data.aspx?d=ComTrade&f=_l1Code%3a4

or

https://tinyurl.com/y8vo7v8u

Download | Explore | Select columns | Select sort order | Link to this page

336777 records | Page 1 of 6736 |

Country or Area	Year	Commodity	Flow	Trade (USD)	Weight (kg)	Quantity Name	Quantity
Afghanistan	2010	Trout, fresh or chilled, whole	Import	8,600	9,000	Weight in kilograms	9,000
Albania	2016	Fish live, except trout, eel or carp	Import	2,202,944	39,896	Weight in kilograms	39,896
Albania	2016	Trout, fresh or chilled, whole	Export	1,973,381	266,283	Weight in kilograms	266,283
Albania	2016	Salmon fresh or chilled, whole	Import	387,938	96,083	Weight in kilograms	96,083
Albania	2016	Salmon fresh or chilled, whole	Export	39,162	5,974	Weight in kilograms	5,974
Albania	2016	Salmonidae, not trout or salmon,fresh or chilled whol	Import	70,993	11,065	Weight in kilograms	11,065
Albania	2016	Salmonidae, not trout or salmon,fresh or chilled whol	Export	49,287	10,808	Weight in kilograms	10,808
Albania	2016	Sardines,brisling,sprats, fresh or chilled, whole	Import	305,172	457,234	Weight in kilograms	457,234
Albania	2016	Fish nes, fresh or chilled, whole	Import	2,654,089	1,322,338	Weight in kilograms	1,322,338
Albania	2016	Fish nes, fresh or chilled, whole	Export	1,353,539	241,436	Weight in kilograms	241,436
Albania	2016	Fish livers and roes, fresh or chilled	Import	3,525	51	Weight in kilograms	51
Albania	2016	Salmon Atlantic or Danube, frozen, whole	Export	22,526	2,445	Weight in kilograms	2,445
Albania	2016	Fish nes, frozen, whole	Import	1,114,294	698,868	Weight in kilograms	698,868
Albania	2016	Fish nes, frozen, whole	Export	631,255	119,205	Weight in kilograms	119,205
Albania	2016	Fish fillets, frozen	Export	147,044	22,523	Weight in kilograms	22,523
Albania	2016	Fish meat & mince, except liver, roe & fillets, froze	Export	111,045	20,212	Weight in kilograms	20,212
Albania	2016	Fish fillets, dried, salted or in brine, not smoked	Export	7,529,293	558,797	Weight in kilograms	558,797
Albania	2016	Salmon, smoked, including fillets	Import	14,497	779	Weight in kilograms	779
Albania	2016	Anchovies, salted or in brine, not dried or smoked	Import	18,934,237	5,799,550	Weight in kilograms	5,799,550

Discuss:
What is the population?
What is a case?
What are each type of variable (numerical, categorical, etc)

Note: could be multiple answers!

Dataset #1

Download | Explore | Select columns | Select sort order | Link to this page

336777 records | Page 1 of 6736 |

Country or Area	Year	Commodity	Flow	Trade (USD)	Weight (kg)	Quantity Name	Quantity
Afghanistan	2010	Trout, fresh or chilled, whole	Import	8,600	9,000	Weight in kilograms	9,000
Albania	2016	Fish live, except trout, eel or carp	Import	2,202,944	39,896	Weight in kilograms	39,896
Albania	2016	Trout, fresh or chilled, whole	Export	1,973,381	266,283	Weight in kilograms	266,283
Albania	2016	Salmon fresh or chilled, whole	Import	387,938	96,083	Weight in kilograms	96,083
Albania	2016	Salmon fresh or chilled, whole	Export	39,162	5,974	Weight in kilograms	5,974
Albania	2016	Salmonidae, not trout or salmon,fresh or chilled whol	Import	70,993	11,065	Weight in kilograms	11,065
Albania	2016	Salmonidae, not trout or salmon,fresh or chilled whol	Export	49,287	10,808	Weight in kilograms	10,808
Albania	2016	Sardines,brisling,sprats, fresh or chilled, whole	Import	305,172	457,234	Weight in kilograms	457,234
Albania	2016	Fish nes, fresh or chilled, whole	Import	2,654,089	1,322,338	Weight in kilograms	1,322,338
Albania	2016	Fish nes, fresh or chilled, whole	Export	1,353,539	241,436	Weight in kilograms	241,436
Albania	2016	Fish livers and roes, fresh or chilled	Import	3,525	51	Weight in kilograms	51
Albania	2016	Salmon Atlantic or Danube, frozen, whole	Export	22,526	2,445	Weight in kilograms	2,445
Albania	2016	Fish nes, frozen, whole	Import	1,114,294	698,868	Weight in kilograms	698,868
Albania	2016	Fish nes, frozen, whole	Export	631,255	119,205	Weight in kilograms	119,205
Albania	2016	Fish fillets, frozen	Export	147,044	22,523	Weight in kilograms	22,523
Albania	2016	Fish meat & mince, except liver, roe & fillets, froze	Export	111,045	20,212	Weight in kilograms	20,212
Albania	2016	Fish fillets, dried, salted or in brine, not smoked	Export	7,529,293	558,797	Weight in kilograms	558,797
Albania	2016	Salmon, smoked, including fillets	Import	14,497	779	Weight in kilograms	779
Albania	2016	Anchovies, salted or in brine, not dried or smoked	Import	18,934,237	5,799,550	Weight in kilograms	5,799,550

Poll:
Type of variable “Country or Area”

1. Continuous Numerical
2. Discrete Numerical
3. Nominal (unordered) Categorical
4. Ordinal (ordered) Categorical

Dataset #1

Download | Explore | Select columns | Select sort order | Link to this page

336777 records | Page 1 of 6736 |

Country or Area	Year	Commodity	Flow	Trade (USD)	Weight (kg)	Quantity Name	Quantity
Afghanistan	2010	Trout, fresh or chilled, whole	Import	8,600	9,000	Weight in kilograms	9,000
Albania	2016	Fish live, except trout, eel or carp	Import	2,202,944	39,896	Weight in kilograms	39,896
Albania	2016	Trout, fresh or chilled, whole	Export	1,973,381	266,283	Weight in kilograms	266,283
Albania	2016	Salmon fresh or chilled, whole	Import	387,938	96,083	Weight in kilograms	96,083
Albania	2016	Salmon fresh or chilled, whole	Export	39,162	5,974	Weight in kilograms	5,974
Albania	2016	Salmonidae, not trout or salmon,fresh or chilled whol	Import	70,993	11,065	Weight in kilograms	11,065
Albania	2016	Salmonidae, not trout or salmon,fresh or chilled whol	Export	49,287	10,808	Weight in kilograms	10,808
Albania	2016	Sardines,brisling,sprats, fresh or chilled, whole	Import	305,172	457,234	Weight in kilograms	457,234
Albania	2016	Fish nes, fresh or chilled, whole	Import	2,654,089	1,322,338	Weight in kilograms	1,322,338
Albania	2016	Fish nes, fresh or chilled, whole	Export	1,353,539	241,436	Weight in kilograms	241,436
Albania	2016	Fish livers and roes, fresh or chilled	Import	3,525	51	Weight in kilograms	51
Albania	2016	Salmon Atlantic or Danube, frozen, whole	Export	22,526	2,445	Weight in kilograms	2,445
Albania	2016	Fish nes, frozen, whole	Import	1,114,294	698,868	Weight in kilograms	698,868
Albania	2016	Fish nes, frozen, whole	Export	631,255	119,205	Weight in kilograms	119,205
Albania	2016	Fish fillets, frozen	Export	147,044	22,523	Weight in kilograms	22,523
Albania	2016	Fish meat & mince, except liver, roe & fillets, froze	Export	111,045	20,212	Weight in kilograms	20,212
Albania	2016	Fish fillets, dried, salted or in brine, not smoked	Export	7,529,293	558,797	Weight in kilograms	558,797
Albania	2016	Salmon, smoked, including fillets	Import	14,497	779	Weight in kilograms	779
Albania	2016	Anchovies, salted or in brine, not dried or smoked	Import	18,934,237	5,799,550	Weight in kilograms	5,799,550

Poll:
Type of variable “Trade USD”

1. Continuous Numerical
2. Discrete Numerical
3. Nominal (unordered) Categorical
4. Ordinal (ordered) Categorical

Dataset #1

Download | Explore | Select columns | Select sort order | Link to this page

336777 records | Page 1 of 6736 |

Country or Area	Year	Commodity	Flow	Trade (USD)	Weight (kg)	Quantity Name	Quantity
Afghanistan	2010	Trout, fresh or chilled, whole	Import	8,600	9,000	Weight in kilograms	9,000
Albania	2016	Fish live, except trout, eel or carp	Import	2,202,944	39,896	Weight in kilograms	39,896
Albania	2016	Trout, fresh or chilled, whole	Export	1,973,381	266,283	Weight in kilograms	266,283
Albania	2016	Salmon fresh or chilled, whole	Import	387,938	96,083	Weight in kilograms	96,083
Albania	2016	Salmon fresh or chilled, whole	Export	39,162	5,974	Weight in kilograms	5,974
Albania	2016	Salmonidae, not trout or salmon,fresh or chilled whol	Import	70,993	11,065	Weight in kilograms	11,065
Albania	2016	Salmonidae, not trout or salmon,fresh or chilled whol	Export	49,287	10,808	Weight in kilograms	10,808
Albania	2016	Sardines,brisling,sprats, fresh or chilled, whole	Import	305,172	457,234	Weight in kilograms	457,234
Albania	2016	Fish nes, fresh or chilled, whole	Import	2,654,089	1,322,338	Weight in kilograms	1,322,338
Albania	2016	Fish nes, fresh or chilled, whole	Export	1,353,539	241,436	Weight in kilograms	241,436
Albania	2016	Fish livers and roes, fresh or chilled	Import	3,525	51	Weight in kilograms	51
Albania	2016	Salmon Atlantic or Danube, frozen, whole	Export	22,526	2,445	Weight in kilograms	2,445
Albania	2016	Fish nes, frozen, whole	Import	1,114,294	698,868	Weight in kilograms	698,868
Albania	2016	Fish nes, frozen, whole	Export	631,255	119,205	Weight in kilograms	119,205
Albania	2016	Fish fillets, frozen	Export	147,044	22,523	Weight in kilograms	22,523
Albania	2016	Fish meat & mince, except liver, roe & fillets, froze	Export	111,045	20,212	Weight in kilograms	20,212
Albania	2016	Fish fillets, dried, salted or in brine, not smoked	Export	7,529,293	558,797	Weight in kilograms	558,797
Albania	2016	Salmon, smoked, including fillets	Import	14,497	779	Weight in kilograms	779
Albania	2016	Anchovies, salted or in brine, not dried or smoked	Import	18,934,237	5,799,550	Weight in kilograms	5,799,550

Poll:
Type of variable “Year”

1. Continuous Numerical
2. Discrete Numerical
3. Nominal (unordered) Categorical
4. Ordinal (ordered) Categorical

Dataset #1

Download | Explore | Select columns | Select sort order | Link to this page

336777 records | Page 1 of 6736 |

Country or Area	Year	Commodity	Flow	Trade (USD)	Weight (kg)	Quantity Name	Quantity
Afghanistan	2010	Trout, fresh or chilled, whole	Import	8,600	9,000	Weight in kilograms	9,000
Albania	2016	Fish live, except trout, eel or carp	Import	2,202,944	39,896	Weight in kilograms	39,896
Albania	2016	Trout, fresh or chilled, whole	Export	1,973,381	266,283	Weight in kilograms	266,283
Albania	2016	Salmon fresh or chilled, whole	Import	387,938	96,083	Weight in kilograms	96,083
Albania	2016	Salmon fresh or chilled, whole	Export	39,162	5,974	Weight in kilograms	5,974
Albania	2016	Salmonidae, not trout or salmon,fresh or chilled whol	Import	70,993	11,065	Weight in kilograms	11,065
Albania	2016	Salmonidae, not trout or salmon,fresh or chilled whol	Export	49,287	10,808	Weight in kilograms	10,808
Albania	2016	Sardines,brisling,sprats, fresh or chilled, whole	Import	305,172	457,234	Weight in kilograms	457,234
Albania	2016	Fish nes, fresh or chilled, whole	Import	2,654,089	1,322,338	Weight in kilograms	1,322,338
Albania	2016	Fish nes, fresh or chilled, whole	Export	1,353,539	241,436	Weight in kilograms	241,436
Albania	2016	Fish livers and roes, fresh or chilled	Import	3,525	51	Weight in kilograms	51
Albania	2016	Salmon Atlantic or Danube, frozen, whole	Export	22,526	2,445	Weight in kilograms	2,445
Albania	2016	Fish nes, frozen, whole	Import	1,114,294	698,868	Weight in kilograms	698,868
Albania	2016	Fish nes, frozen, whole	Export	631,255	119,205	Weight in kilograms	119,205
Albania	2016	Fish fillets, frozen	Export	147,044	22,523	Weight in kilograms	22,523
Albania	2016	Fish meat & mince, except liver, roe & fillets, froze	Export	111,045	20,212	Weight in kilograms	20,212
Albania	2016	Fish fillets, dried, salted or in brine, not smoked	Export	7,529,293	558,797	Weight in kilograms	558,797
Albania	2016	Salmon, smoked, including fillets	Import	14,497	779	Weight in kilograms	779
Albania	2016	Anchovies, salted or in brine, not dried or smoked	Import	18,934,237	5,799,550	Weight in kilograms	5,799,550

What are some questions you might want to use your new R-skills (plotting, calculations) and summary statistics knowledge (mean, median, standard deviation, quartiles) to probe with this dataset?

Real Data! (ooooo!)

We'll go through this one

Dataset #1:

http://data.un.org/

to

http://data.un.org/Explorer.aspx

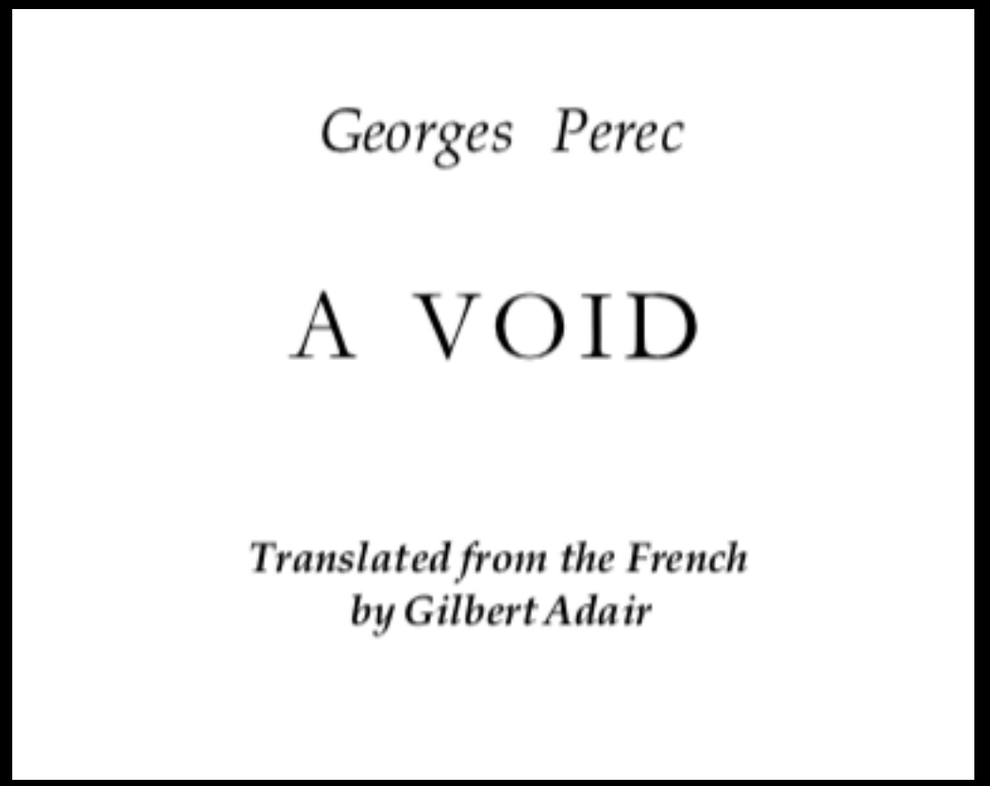
to

Commodity trade stats =>
Fish, crustaceans, mollusks, etc

We may or may not get to this one

Dataset #2:

Excerpt of the
Introduction of "A Void"
by Georges Perec



Real Data! (ooooo!)

To R!

Steps:

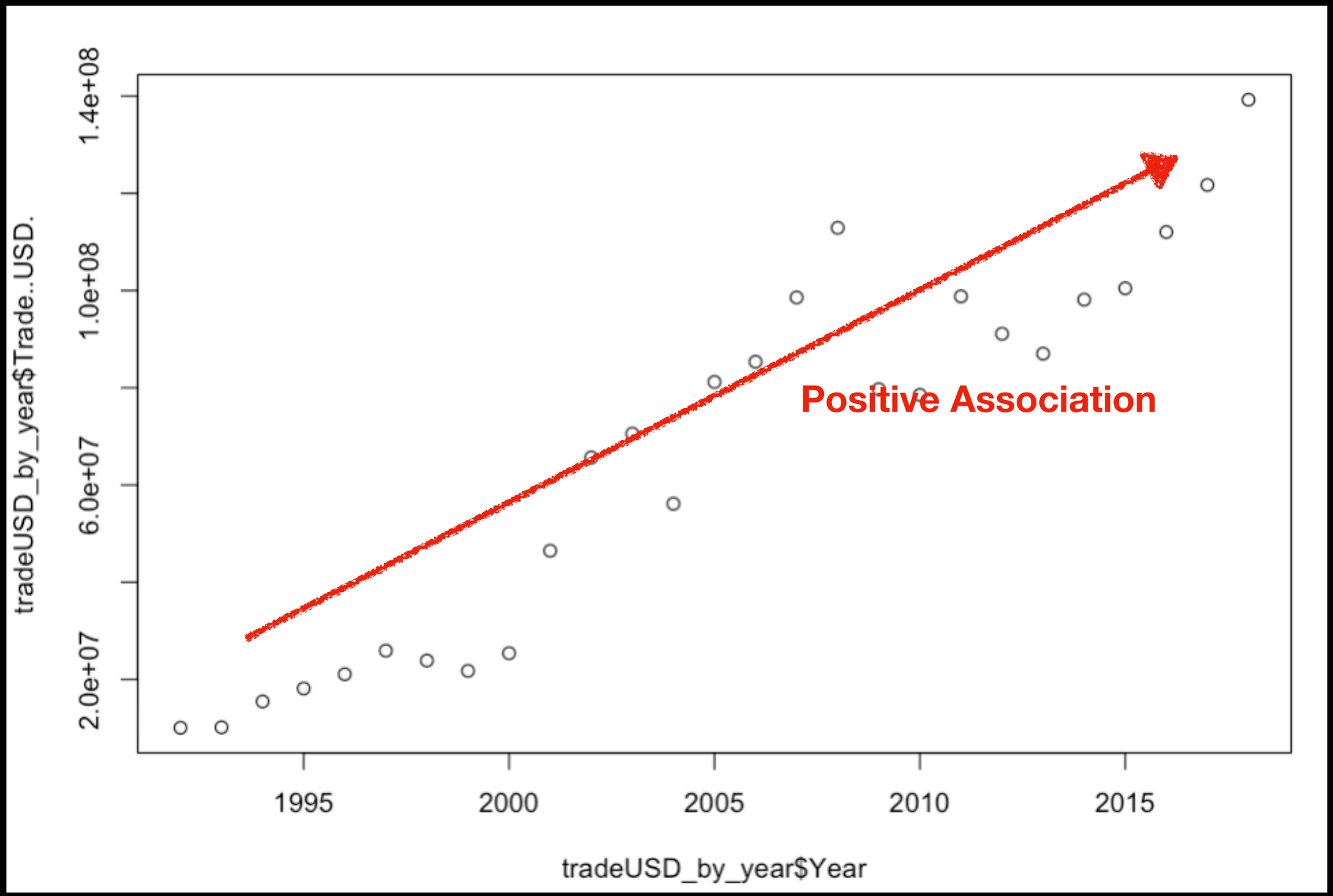
1. **Make sure everybody can download and open the data.**

2. **Make sure everybody can run the example code and look at the plots.**

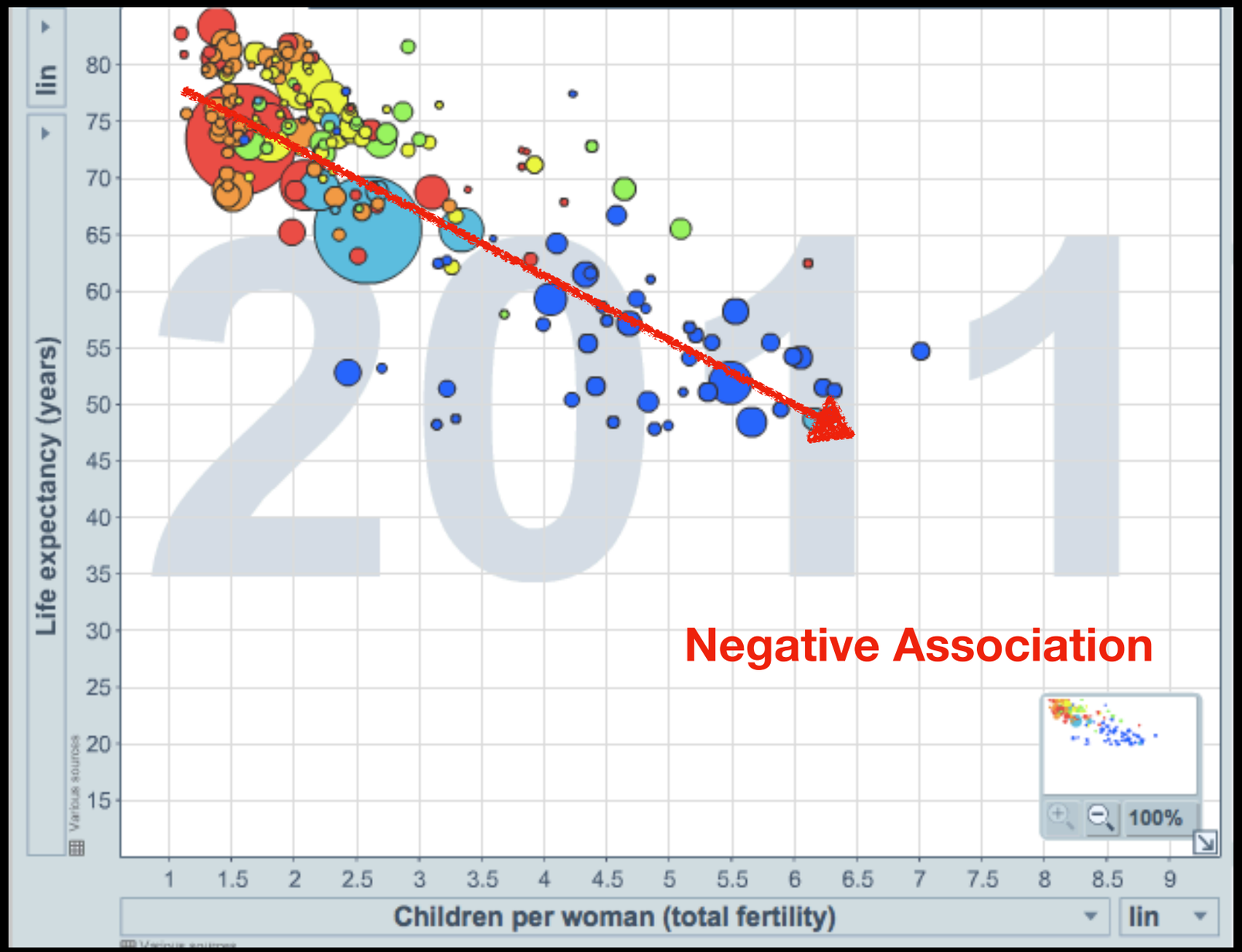
3. **Then we'll go through a R-things together**

Confusion at this stage is TOTALLY normal!

Fish Dataset: Croatian Oyster Trade in USD vs. Time



Fish Dataset: Life expectancy vs. Children Birthed per Woman



http://www.gapminder.org/world

Fish Dataset: Life expectancy vs. Children Birthed per Woman

Associated or independent, not both

A pair of variables are either related in some way (associated) or not (independent).
No pair of variables is both associated and independent.

Designing Studies: Observational & Experimental

There are more airplane deaths now than there were 50 years ago. Does that mean airplane travel is becoming more dangerous?

No, there are a lot more people flying now.
Compare rates, not absolute numbers.

The death rate in the Navy during the Spanish-American war was 9/1000. For civilians in the same time period in New York City it was 16/1000. Does that mean it is safer to be in the Navy?

No, civilians include the old and sick while the Navy was comprised of healthy, young men.
Make sure to compare like to like.

Designing Studies: Observational & Experimental

1975 study of 1286 British women.

- ★ 23% of smokers died by 20-year follow-up
- ★ 29% of nonsmokers died by 20-year follow-up

Explanatory ➡ Response

- So do smokers tend to live longer?
- What is a potential confounding factor?

TIP: Explanatory and response variables

To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other and plan an appropriate analysis.

explanatory variable —might affect→ response variable

Designing Studies: Table Proportions

	Died in 20 yrs	Alive	Total Smoker or Non
Non-smokers	205	499	704/1286 = 55%
Smokers	138	444	582/1286 = 45%
Total Alive or Dead	343	943	1286

What percentage of the study participants are smokers? Non-smokers?

Designing Studies: Table Proportions

	Died in 20 yrs	Alive	Total Smoker or Non
Non-smokers	205	499	704
Smokers	138	444	582
Total Alive or Dead	343/1286 = 27%	943/1286 = 73%	1286

What percentage of the study participants are alive? Dead?

Designing Studies: Table Proportions

	Died in 20 yrs	Alive	Total Smoker or Non
Non-smokers	205	499/704 = 71%	704
Smokers	138	444/582 = 76%	582
Total Alive or Dead	343	943	1286

What percentage of non-smokers (and smokers) are alive?

More alive smokers?

Designing Studies: Table Proportions

Age is a confounding factor

	Died in 20 yrs	Alive	Total in Row
Non-smokers Age 18-64	65	474/539 = 88%	539
Non-smokers Age 65+	140	25	165
Smokers Age 18-64	96	437/533 = 82%	533
Smokers Age 65+	42	7	49
Total In Column	343	943	1286

What percentage of non-smokers (and smokers) are alive from the <65 age group?

Now looks like non-smokers fair better than smokers.

Confounding factors

- A confounding factor is a variable associated with the both the explanatory and the response variable.

- Because of this, the response could be due to the supposed explanatory variable or to the confounding factor - the two are confounded.

Confounding factors

- In the previous example the supposed explanatory variable is:
  - ⋆ Smoking

- The response variable is:
  - ⋆ Dying within 20 years

- The confounding factor is:
  - ⋆ Age - there were more older people in the non-smoking group and older people are more likely to pass away within 20 years.

A case study

Eating more fruit, particularly blueberries, apples and grapes, is linked to a reduced risk of developing type-2 diabetes, suggests a study in the British Medical Journal.

http://www.bbc.com/news/health-23880701

“Blueberries cut the risk (of type-2 diabetes) by 26%...”

“The research looked at the diets of more than 187,000 people in the US...”

“The studies used food frequency questionnaires to follow up the participants every four years, asking how often, on average, they ate a standard portion of each fruit…”

Observation study vs. Experiment

- In an observational study researches watch/record information without imposing any treatment.

- In an experiment, researchers impose a treatment to try to draw a causal conclusion about the effect of the treatment.

- Why would we carry out one or the other?

A case study

- You have a hypothesis:
  Prolonged smoking leads to dental problems.
- How will you collect data to test this hypothesis?
- Should we do an experiment?

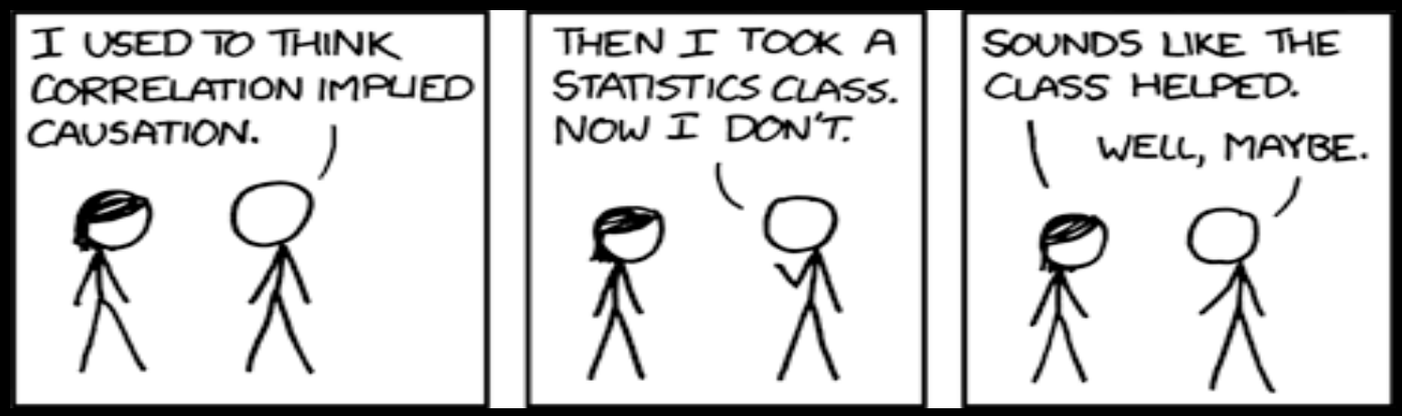
Blocking Variables

Blocking variables are characteristics that the experimental units come with, that we would like to control for.

Blocking is an effort to minimize confounding factors.

Correlation is not causation

- Correlation is not causation!



http:// xkcd.com/ 552/

- Observational studies alone cannot prove causation; only well designed experiments can prove causation.

More Experimental Design Terminology...

Placebo: fake treatment, often used as the control group for medical studies

Placebo effect: experimental units showing improvement simply because they believe they are receiving a special treatment

Blinding: when experimental units do not know whether they are in the control or treatment group

Double-blind: when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group

Practice #1

A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are equally represented in each group. Which of the below is correct?

(1) There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)

(2) There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)

(3) There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)

(4) There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

Practice #2

Protein May Hold the Key to Who Gets Alzheimer's
By PAM BELLUCK, MARCH 19, 2014, *New York Times*

It is one of the big scientific mysteries of Alzheimer's disease: Why do some people whose brains accumulate the plaques and tangles so strongly associated with Alzheimer's not develop the disease?

Now, a series of studies by Harvard scientists suggests a possible answer, one that could lead to new treatments if confirmed by other research….

The research, published on Wednesday in the journal Nature, focuses on a protein previously thought to act mostly in the brains of developing fetuses. The scientists found that the protein also appears to protect neurons in healthy older people from aging-related stresses. But in people with Alzheimer's and other dementias, the protein is sharply depleted in key brain regions.

Experts said if other scientists could replicate and expand upon the findings, the role of the protein, called REST, could spur development of new drugs for dementia, which has so far been virtually impossible to treat. But they cautioned that much more needed to be determined…