Data Strategies

+ Data Tactics

Thomas Stoeger
thomas.stoeger@northwestern.edu

advertising (postdoc and grad student symposium): August 22

https://www.nupostdocs.org/symposium.html

aim: prepared to approach data

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

break

Setup

- Computer
- Organizing data
- Coding

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

break

Setup

- Computer
- Organizing data
- Coding

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

break

Setup

- Computer
- Organizing data
- Coding

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

break

Setup

- Computer
- Organizing data
- Coding

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

break

Setup

- Computer
- Organizing data
- Coding

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

break

Setup

- Computer
- Organizing data
- Coding

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

break

Setup

- Computer
- Organizing data
- Coding

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

break

Setup

- Computer
- Organizing data
- Coding

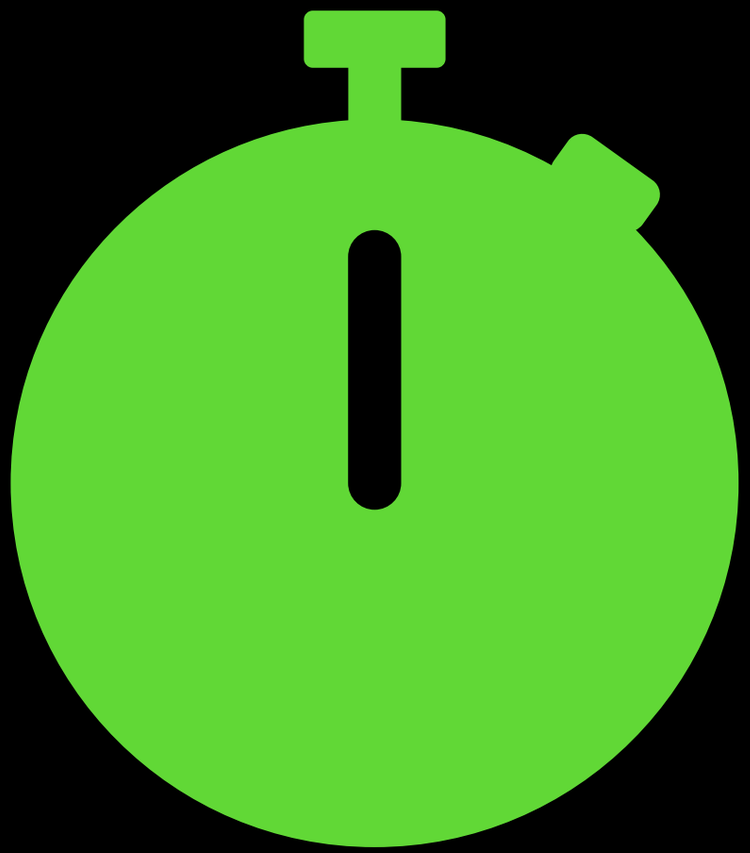
Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

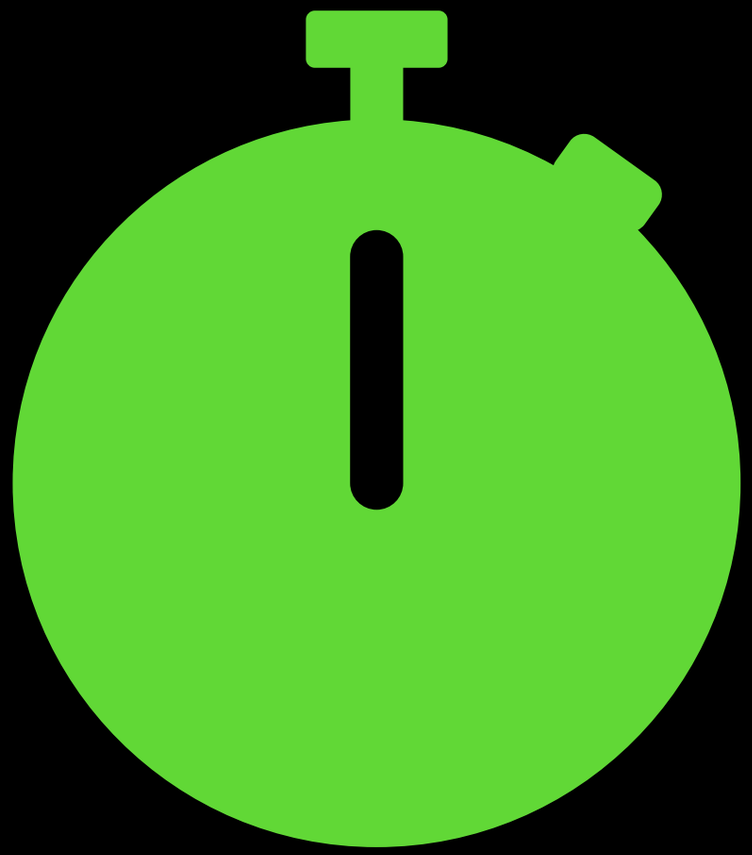
break

Setup

- Computer
- Organizing data
- Coding

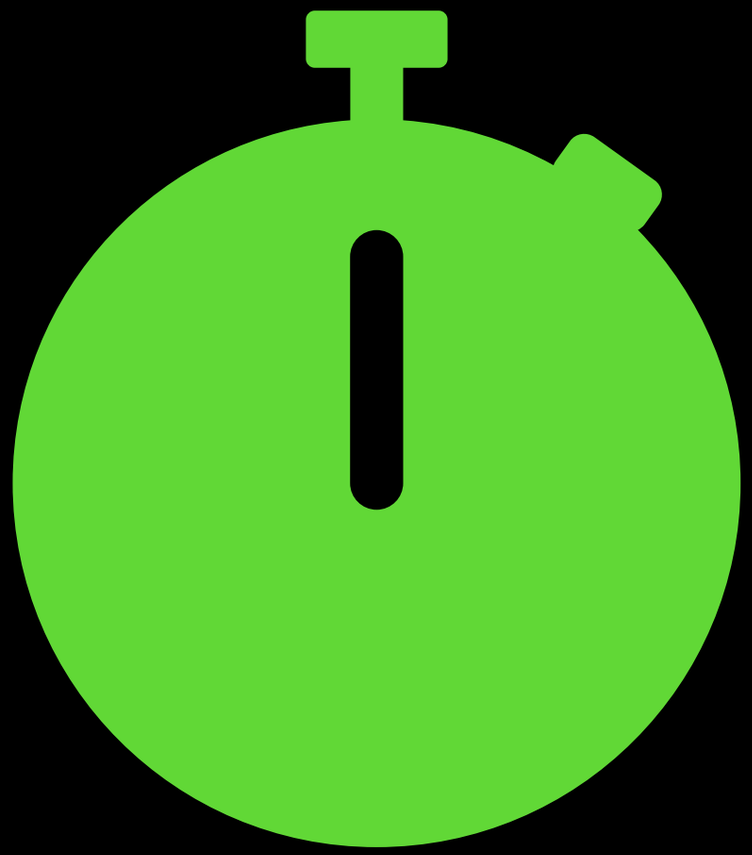


Two minute exercise.



Two minute exercise.

Find a neighbor, or neighbors, whom you do not know.



Two minute exercise.

Find a neighbor, or neighbors, whom you do not know.

Then talk about the worst data you have encountered.

How can we anticipate problems?

How can we anticipate problems?

“just”, “easy”, “simple”

How can we anticipate problems?

“just”, “easy”, “simple”

Can not share data or code.

Understand the past history of the data.

External:

↓

Read paper
or documentation

Understand the past history of the data.

External:

Read paper
or documentation

Internal:

Ask

"is generated by program"

Read documentation

Custom

Communication

Understand the past history of the data.

External:

Read paper
or documentation

Internal:

Ask

"is generated by program"

Read documentation

Custom

Communication

Write consistency checks!
(programmatically)

Understand the past history of the data.

External:

Read paper
or documentation

Internal:

Ask

"is generated by program"

Read documentation

Custom

Communication

Write consistency checks!
(programmatically)

Expect to spend
days or weeks

Understand the past history of the data.

External:

Read paper
or documentation

"is generated by program"

Read documentation

Internal:

Ask

Custom

Communication

Write consistency checks!
(programmatically)

Expect to spend
days or weeks

Tips for getting all information:

- be decent person
- remove fear that you will scoop

Workflow

- Anticipate problems
- **Get overview**
- Clean data
- Find connections

break

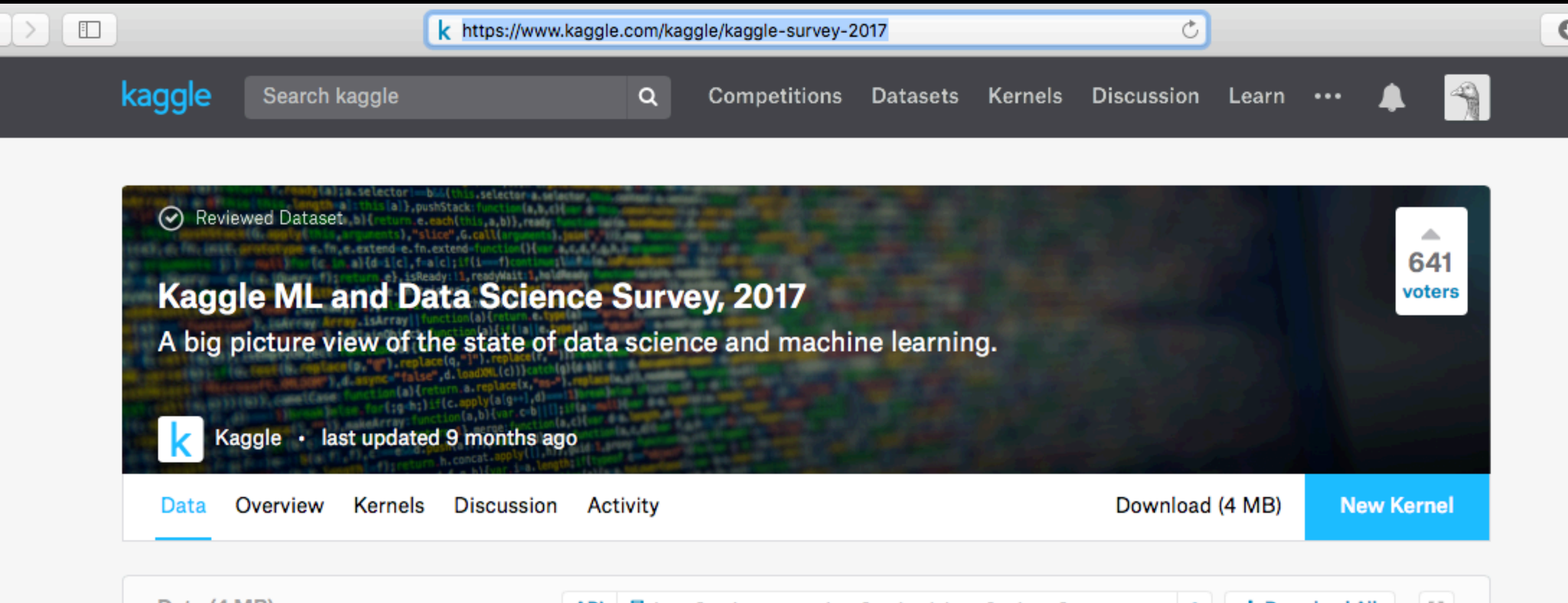
Setup

- Environment
- Organizing data
- Coding

Introducing

Kaggle ML and Data Science Survey, 2017

An example data set, which we will be using



Introducing

Kaggle ML and Data Science Survey, 2017

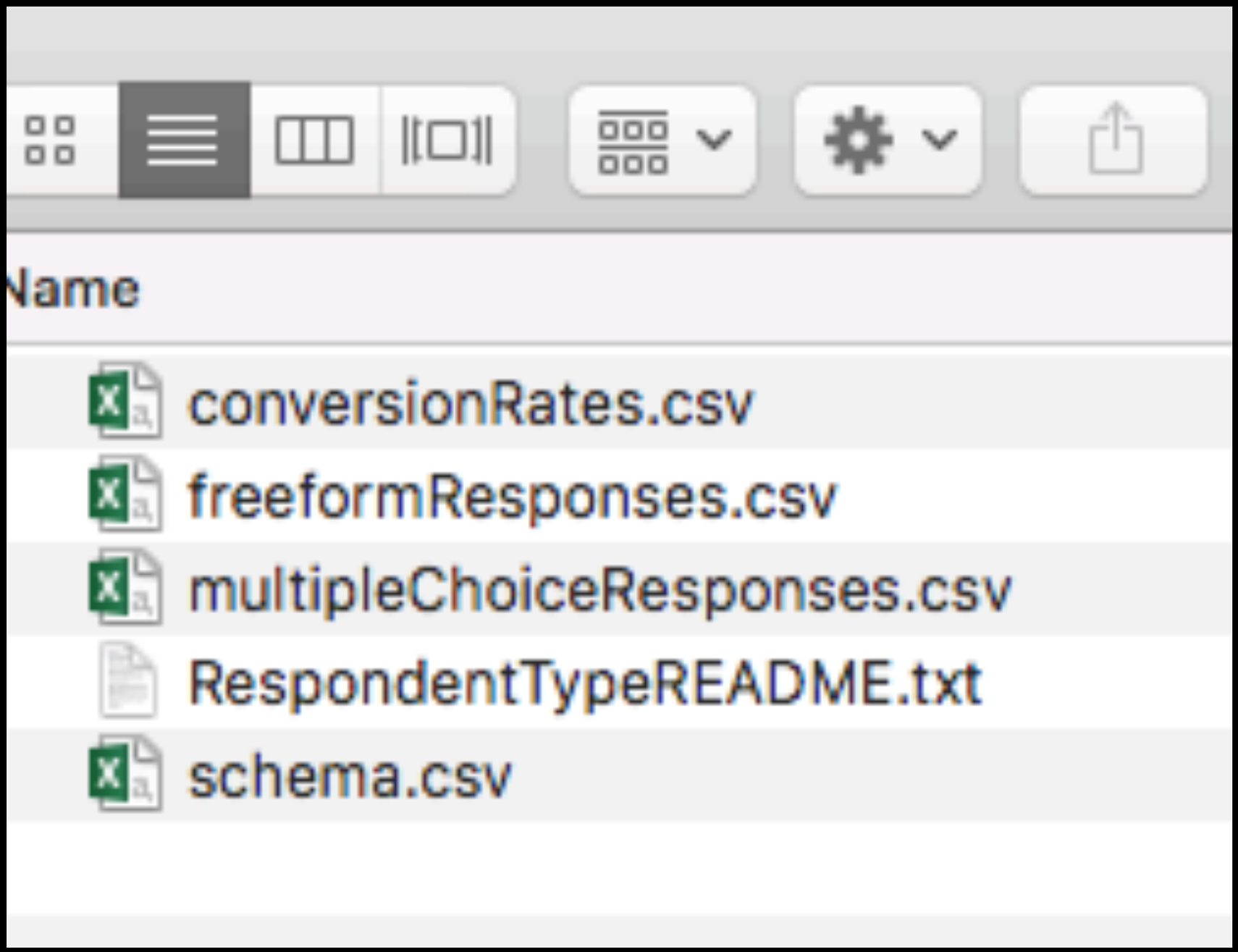
An example data set, which we will be using

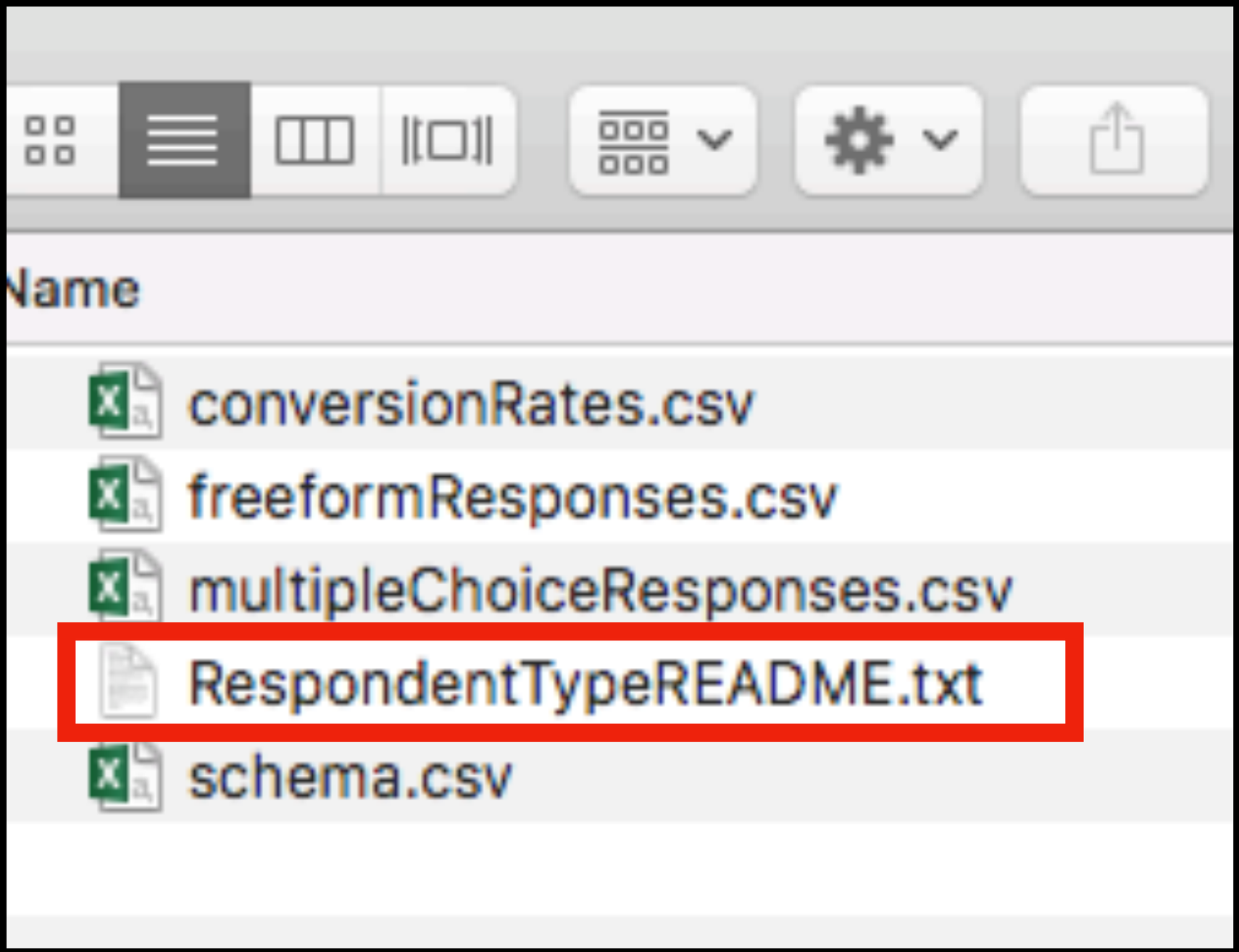
https://www.kaggle.com/kaggle/kaggle-survey-2017

kaggle Search kaggle Competitions Datasets Kernels Discussion Learn

Data Overview Kernels Discussion Activity

super nice example


Name
conversionRates.csv
freeformResponses.csv
multipleChoiceResponses.csv
RespondentTypeREADME.txt
schema.csv


Name
conversionRates.csv
freeformResponses.csv
multipleChoiceResponses.csv
RespondentTypeREADME.txt
schema.csv

RespondentTypeREADME.txt

All: Every respondent was shown this question

Non-worker: Respondents who indicated that they were "Not employed, and not looking for work" or "I prefer not to say"

Non-switcher: Respondents that indicated that they were not actively looking to switch careers to data science.

Worker: Respondents who indicted that they were "Employed full-time", "Employed part-time", "Independent contractor, freelancer, or self-employed", or "retired"

CodingWorker: Respondents who indicated that they were "Employed full-time", "Employed part-time", or an "Independent contractor, freelancer, or self-employed" AND that they write code to analyze data in their current job

CodingWorker-NC: Respondents who indicated that they were "Employed full-time" or "Employed part-time" AND that they write code to analyze data in their current job.

Learners: Respondents who indicated that they were either students, formally or informally learning data science skills, planning to transition into data science, or not employed but looking for work

RespondentTypeREADME.txt

```
All: Every respondent was shown this question

Non-worker: Respondents who indicated that they were "Not employed, and not looking for
work" or "I prefer not to say"

Non-switcher: Respondents that indicated that they were not actively looking to switch
careers to data science.

Worker: Respondents who indicted that they were "Employed full-time", "Employed part-
time", "Independent contractor, freelancer, or self-employed", or "retired"

CodingWorker: Respondents who indicated that they were "Employed full-time", "Employed
part-time", or an "Independent contractor, freelancer, or self-employed" AND that they
write code to analyze data in their current job

CodingWorker-NC: Respondents who indicated that they were "Employed full-time" or
"Employed part-time" AND that they write code to analyze data in their current job.

Learners: Respondents who indicated that they were either students, formally or informally
learning data science skills, planning to transition into data science, or not employed
but looking for work
```

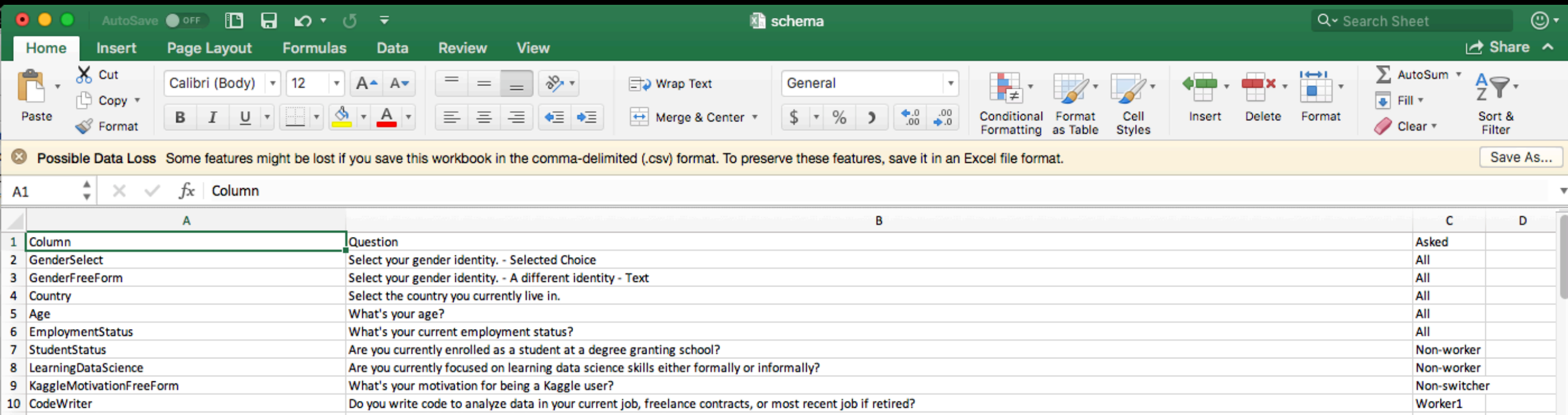
Understand terminology!

RespondentTypeREADME.txt

```
All: Every respondent was shown this question

Non-worker: Respondents who indicated that they were "Not employed, and not looking for
work" or "I prefer not to say"

Non-switcher: Respondents that                d that they were not actively looking to switch
careers to data science.

Worker: Respondents who indicte                          ed full-time", "Employed part-
time", "Independent contractor                           yed", or "retired"

CodingWorker: Respondents who                            ere "Employed full-time", "Employed
part-time", or an "Independent                           ncer, or self-employed" AND that they
write code to analyze data in

CodingWorker-NC: Respondents                             were "Employed full-time" or
"Employed part-time" AND that                            yze data in their current job.

Learners: Respondents who ind ated that t                er students, formally or informally
learning data science skills, planning to t              into data science, or not employed
but looking for work
```

Curators might follow specific assumptions.
e.g.: “I prefer not to say” at employment status is essentially “not employed”.

Understand terminology!



	A	B	C	D
1	Column	Question	Asked	
2	GenderSelect	Select your gender identity. - Selected Choice	All	
3	GenderFreeForm	Select your gender identity. - A different identity - Text	All	
4	Country	Select the country you currently live in.	All	
5	Age	What's your age?	All	
6	EmploymentStatus	What's your current employment status?	All	
7	StudentStatus	Are you currently enrolled as a student at a degree granting school?	Non-worker	
8	LearningDataScience	Are you currently focused on learning data science skills either formally or informally?	Non-worker	
9	KaggleMotivationFreeForm	What's your motivation for being a Kaggle user?	Non-switcher	
10	CodeWriter	Do you write code to analyze data in your current job, freelance contracts, or most recent job if retired?	Worker1	
11	CareerSwitcher	Are you actively looking to switch careers to data science?	Worker1	
12	CurrentJobTitleSelect	Select the option that's most similar to your current job/professional title (or most recent title if retired). - Selected Choice	Worker1	
13	CurrentJobTitleFreeForm	Select the option that's most similar to your current job/professional title (or most recent title if retired). - Other - Text	Worker1	
14	TitleFit	How adequately do you feel your title describes what you do (or what you did if retired)?	Worker1	
15	CurrentEmployerType	Which of the following describe your current employer (or most recent employer if retired)? (Select all that apply)	Worker1	
16	MLToolNextYearSelect	Which tool or technology are you most excited about learning in the next year? (Select one option) - Selected Choice	All	
17	MLToolNextYearFreeForm	Which tool or technology are you most excited about learning in the next year? (Select one option) - Other - Text	All	
18	MLMethodNextYearSelect	Which ML/DS method are you most excited about learning in the next year? (Select one option) - Selected Choice	All	
19	MLMethodNextYearFreeForm	Which ML/DS method are you most excited about learning in the next year? (Select one option) - Other - Text	All	
20	LanguageRecommendationSelect	What programming language would you recommend a new data scientist learn first? (Select one option) - Selected Choice	All	
21	LanguageRecommendationFreeForm	What programming language would you recommend a new data scientist learn first? (Select one option) - Other - Text	All	
22	PublicDatasetsSelect	Where do you find public datasets to practice data science skills? (Select all that apply) - Selected Choice	All	
23	PublicDatasetsFreeForm	Where do you find public datasets to practice data science skills? (Select all that apply) - Other - Text	All	
24	PersonalProjectsChallengeFreeForm	What is your biggest challenge with the public datasets you find for personal projects?	All	
25	LearningPlatformSelect	What platforms & resources have you used to continue learning data science skills? (Select all that apply) - Selected Choice	All	
26	LearningPlatformCommunityFreeForm	What platforms & resources have you used to continue learning data science skills? (Select all that apply) - Non-Kaggle online communities - Text	All	
27	LearningPlatformFreeForm1	What platforms & resources have you used to continue learning data science skills? (Select all that apply) - Other - Text1	All	
28	LearningPlatformFreeForm2	What platforms & resources have you used to continue learning data science skills? (Select all that apply) - Other - Text2	All	
29	LearningPlatformFreeForm3	What platforms & resources have you used to continue learning data science skills? (Select all that apply) - Other - Text3	All	
30	LearningPlatformUsefulnessArxiv	How useful did you find these platforms & resources for learning data science skills? - Arxiv	All	
31	LearningPlatformUsefulnessBlogs	How useful did you find these platforms & resources for learning data science skills? - Blogs	All	
32	LearningPlatformUsefulnessCollege	How useful did you find these platforms & resources for learning data science skills? - College/University	All	
33	LearningPlatformUsefulnessCompany	How useful did you find these platforms & resources for learning data science skills? - Company internal community	All	
34	LearningPlatformUsefulnessConferences	How useful did you find these platforms & resources for learning data science skills? - Conferences	All	
35	LearningPlatformUsefulnessFriends	How useful did you find these platforms & resources for learning data science skills? - Friends network	All	
36	LearningPlatformUsefulnessKaggle	How useful did you find these platforms & resources for learning data science skills? - Kaggle	All	
37	LearningPlatformUsefulnessNewsletters	How useful did you find these platforms & resources for learning data science skills? - Newsletters	All	
38	LearningPlatformUsefulnessCommunities	How useful did you find these platforms & resources for learning data science skills? - Non-Kaggle online communities	All	

schema

Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

A1 Column

	A	B	C	D
1	Column	Question	Asked	
2	GenderSelect	Select your gender identity. - Selected Choice	All	
3	GenderFreeForm	Select your gender identity. - A different identity - Text	All	
4	Country	Select the country you currently live in.	All	
5	Age	What's your age?	All	
6	EmploymentStatus	What's your current employment status?	All	
7	StudentStatus	Are you currently enrolled as a student at a degree granting school?	Non-worker	
8	LearningDataScience	Are you currently focused on learning data science skills either formally or informally?	Non-worker	
9	KaggleMotivationFreeForm	What's your motivation for being a Kaggle user?	Non-switcher	
10	CodeWriter	Do you write code to analyze data in your current job, freelance contracts, or most recent job if retired?	Worker1	
11	CareerSwitcher	Are you actively looking to switch careers to data science?	Worker1	
12	CurrentJobTitleSelect	Select the option that's most similar to your current job/professional title (or most recent title if retired). - Selected Choice	Worker1	
13	CurrentJobTitleFreeForm	Select the option that's most similar to your current job/professional title (or most recent title if retired). - Other - Text	Worker1	
14	TitleFit	How adequately do you feel your title describes what you do (or what you did if retired)?	Worker1	
15	CurrentEmployerType	Which of the following describe your current employer (or most recent employer if retired)? (Select all that apply)	Worker1	
16	MLToolNextYearSelect	Which tool or technology are you most excited about learning in the next year? (Select one option) - Selected Choice	All	
17	MLToolNextYearFreeForm	Which tool or technology are you most excited about learning in the next year? (Select one option) - Other - Text	All	
18	MLMethodNextYearSelect	Which ML/DS method are you most excited about learning in the next year? (Select one option) - Selected Choice	All	
19	MLMethodNextYearFreeForm	Which ML/DS method are you most excited about learning in the next year? (Select one option) - Other - Text	All	
20	LanguageRecommendationSelect	What programming language would you recommend a new data scientist learn first? (Select one option) - Selected Choice	All	
21	LanguageRecommendationFreeForm	What programming language would you recommend a new data scientist learn first? (Select one option) - Other - Text	All	
22	PublicDatasetsSelect	Where do you find public datasets to practice data science skills? (Select all that apply) - Selected Choice	All	
23	PublicDatasetsFreeForm	Where do you find public datasets to practice data science skills? (Select all that apply) - Other - Text	All	
24	PersonalProjectsChallengeFreeForm	What is your biggest challenge with the public datasets you find for personal projects?	All	
25	LearningPlatformSelect	What platforms & resources have you used to continue learning data science skills? (Select all that apply) - Selected Choice	All	
26	LearningPlatformCommunityFreeForm	What platforms & resources have you used to continue learning data science skills? (Select all that apply) - Non-Kaggle online communities - Text	All	
27	LearningPlatformFreeForm1	What platforms & resources have you used to continue learning data science skills? (Select all that apply) - Other - Text1	All	
28	LearningPlatformFreeForm2	What platforms & resources have you used to continue learning data science skills? (Select all that apply) - Other - Text2	All	
29	LearningPlatformFreeForm3	What platforms & resources have you used to continue learning data science skills? (Select all that apply) - Other - Text3	All	
30	LearningPlatformUsefulnessArxiv	How useful did you find these platforms & resources for learning data science skills? - Arxiv	All	
31	LearningPlatformUsefulnessBlogs	How useful did you find these platforms & resources for learning data science skills? - Blogs	All	
32	LearningPlatformUsefulnessCollege	How useful did you find these platforms & resources for learning data science skills? - College/University	All	
33	LearningPlatformUsefulnessCompany	How useful did you find these platforms & resources for learning data science skills? - Company internal community	All	
34	LearningPlatformUsefulnessConferences	How useful did you find these platforms & resources for learning data science skills? - Conferences	All	
35	LearningPlatformUsefulnessFriends	How useful did you find these platforms & resources for learning data science skills? - Friends network	All	
36	LearningPlatformUsefulnessKaggle	How useful did you find these platforms & resources for learning data science skills? - Kaggle	All	
37	LearningPlatformUsefulnessNewsletters	How useful did you find these platforms & resources for learning data science skills? - Newsletters	All	
38	LearningPlatformUsefulnessCommunities	How useful did you find these platforms & resources for learning data science skills? - Non-Kaggle online communities	All	

schema

Ready 100%

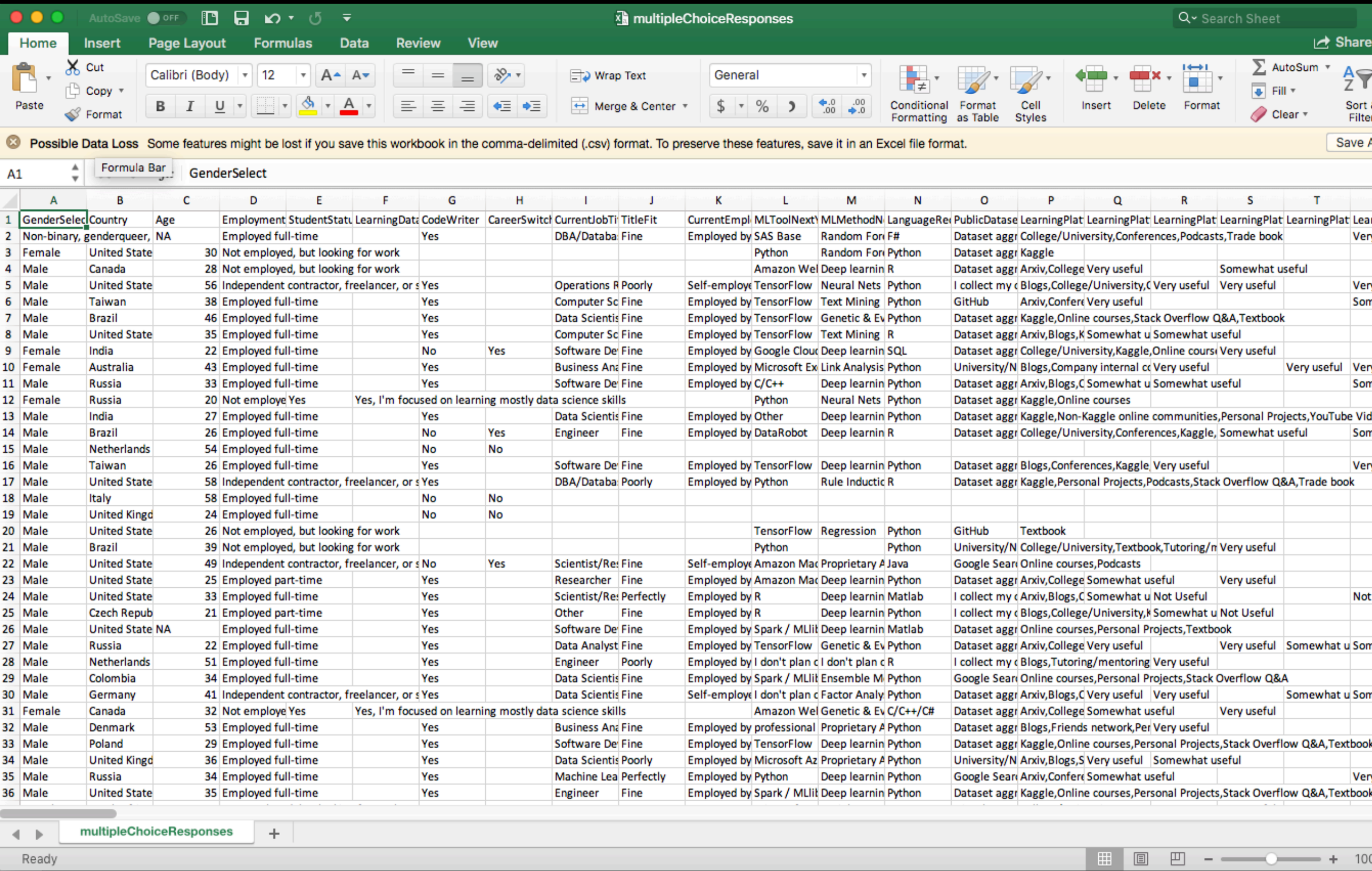
Understand possibilities!

Paste Format B I U A Merge & Center $

Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve th

A1 fx Column

	A	
1	Column	Question
2	GenderSelect	Select your gender identity. - Selected Choice
3	GenderFreeForm	Select your gender identity. - A different identity - Text
4	Country	Select the country you currently live in.
5	Age	What's your age?
6	EmploymentStatus	What's your current employment status?
7	StudentStatus	Are you currently enrolled as a student at a degree granting school?
8	LearningDataScience	Are you currently focused on learning data science skills either formally or informall
9	KaggleMotivationFreeForm	What's your motivation for being a Kaggle user?
10	CodeWriter	Do you write code to analyze data in your current job, freelance contracts, or most r
11	CareerSwitcher	Are you actively looking to switch careers to data science?
12	CurrentJobTitleSelect	Select the option that's most similar to your current job/professional title (or most
13	CurrentJobTitleFreeForm	Select the option that's most similar to your current job/professional title (or most
14	TitleFit	How adequately do you feel your title describes what you do (or what you did if reti
15	CurrentEmployerType	Which of the following describe your current employer (or most recent employer if
16	MLToolNextYearSelect	Which tool or technology are you most excited about learning in the next year? (Sel
17	MLToolNextYearFreeForm	Which tool or technology are you most excited about learning in the next year? (Sel
18	MLMethodNextYearSelect	Which ML/DS method are you most excited about learning in the next year? (Select
19	MLMethodNextYearFreeForm	Which ML/DS method are you most excited about learning in the next year? (Select
20	LanguageRecommendationSelect	What programming language would you recommend a new data scientist learn firs
21	LanguageRecommendationFreeForm	What programming language would you recommend a new data scientist learn firs
22	PublicDatasetsSelect	Where do you find public datasets to practice data science skills? (Select all that ap
23	PublicDatasetsFreeForm	Where do you find public datasets to practice data science skills? (Select all that ap
24	PersonalProjectsChallengeFreeForm	What is your biggest challenge with the public datasets you find for personal projec
25	LearningPlatformSelect	What platforms & resources have you used to continue learning data science skills?
26	LearningPlatformCommunityFreeForm	What platforms & resources have you used to continue learning data science skills?
27	LearningPlatformFreeForm1	What platforms & resources have you used to continue learning data science skills?



AutoSave OFF

multipleChoiceResponses

Search Sheet

Home Insert Page Layout Formulas Data Review View

Share

Paste Cut Copy Format

Calibri (Body) 12

Wrap Text Merge & Center

General

Conditional Formatting Format as Table Cell Styles Insert Delete Format

AutoSum Fill Clear

Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

Save A

A1

Formula Bar

GenderSelect

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	GenderSelec	Country	Age	Employment	StudentStatu	LearningData	CodeWriter	CareerSwitch	CurrentJobTi	TitleFit	CurrentEmpl	MLToolNextY	MLMethodN	LanguageRe	PublicDatase	LearningPlat	LearningPlat	LearningPlat	LearningPlat	LearningPlat
2	Non-binary, genderqueer,		NA	Employed full-time			Yes		DBA/Databa	Fine	Employed by	SAS Base	Random For	F#	Dataset aggr	College/University,Conferences,Podcasts,Trade book				
3	Female	United State	30	Not employed, but looking for work								Python	Random For	Python	Dataset aggr	Kaggle				
4	Male	Canada	28	Not employed, but looking for work								Amazon Wel	Deep learnin	R	Dataset aggr	Arxiv,College	Very useful		Somewhat useful	
5	Male	United State	56	Independent contractor, freelancer, or s			Yes		Operations R	Poorly	Self-employe	TensorFlow	Neural Nets	Python	I collect my c	Blogs,College/University,C		Very useful	Very useful	
6	Male	Taiwan	38	Employed full-time			Yes		Computer Sc	Fine	Employed by	TensorFlow	Text Mining	Python	GitHub	Arxiv,Confere	Very useful			
7	Male	Brazil	46	Employed full-time			Yes		Data Scientis	Fine	Employed by	TensorFlow	Genetic & Ev	Python	Dataset aggr	Kaggle,Online courses,Stack Overflow Q&A,Textbook				
8	Male	United State	35	Employed full-time			Yes		Computer Sc	Fine	Employed by	TensorFlow	Text Mining	R	Dataset aggr	Arxiv,Blogs,K	Somewhat u	Somewhat useful		
9	Female	India	22	Employed full-time			No	Yes	Software De	Fine	Employed by	Google Cloud	Deep learnin	SQL	Dataset aggr	College/University,Kaggle,Online cours	Very useful			
10	Female	Australia	43	Employed full-time			Yes		Business Ana	Fine	Employed by	Microsoft Ex	Link Analysis	Python	University/N	Blogs,Company internal c		Very useful		Very useful
11	Male	Russia	33	Employed full-time			Yes		Software De	Fine	Employed by	C/C++	Deep learnin	Python	Dataset aggr	Arxiv,Blogs,C	Somewhat u	Somewhat useful		
12	Female	Russia	20	Not employe	Yes	Yes, I'm focused on learning mostly data science skills						Python	Neural Nets	Python	Dataset aggr	Kaggle,Online courses				
13	Male	India	27	Employed full-time			Yes		Data Scientis	Fine	Employed by	Other	Deep learnin	Python	Dataset aggr	Kaggle,Non-Kaggle online communities,Personal Projects,YouTube				
14	Male	Brazil	26	Employed full-time			No	Yes	Engineer	Fine	Employed by	DataRobot	Deep learnin	R	Dataset aggr	College/University,Conferences,Kaggle,	Somewhat useful			
15	Male	Netherlands	54	Employed full-time			No	No												
16	Male	Taiwan	26	Employed full-time			Yes		Software De	Fine	Employed by	TensorFlow	Deep learnin	Python	Dataset aggr	Blogs,Conferences,Kaggle		Very useful		
17	Male	United State	58	Independent contractor, freelancer, or s			Yes		DBA/Databa	Poorly	Employed by	Python	Rule Inductic	R	Dataset aggr	Kaggle,Personal Projects,Podcasts,Stack Overflow Q&A,Trade book				
18	Male	Italy	58	Employed full-time			No	No												
19	Male	United Kingd	24	Employed full-time			No	No												
20	Male	United State	26	Not employed, but looking for work								TensorFlow	Regression	Python	GitHub	Textbook				
21	Male	Brazil	39	Not employed, but looking for work								Python		Python	University/N	College/University,Textbook,Tutoring/n			Very useful	
22	Male	United State	49	Independent contractor, freelancer, or s			No	Yes	Scientist/Res	Fine	Self-employe	Amazon Mac	Proprietary A	Java	Google Sear	Online courses,Podcasts				
23	Male	United State	25	Employed part-time			Yes		Researcher	Fine	Employed by	Amazon Mac	Deep learnin	Python	Dataset aggr	Arxiv,College	Somewhat useful		Very useful	
24	Male	United State	33	Employed full-time			Yes		Scientist/Res	Perfectly	Employed by	R	Deep learnin	Matlab	I collect my c	Arxiv,Blogs,C	Somewhat u	Not Useful		
25	Male	Czech Repub	21	Employed part-time			Yes		Other	Fine	Employed by	R	Deep learnin	Python	I collect my c	Blogs,College/University,K	Somewhat u	Not Useful		
26	Male	United State	NA	Employed full-time			Yes		Software De	Fine	Employed by	Spark / MLlib	Deep learnin	Matlab	Dataset aggr	Online courses,Personal Projects,Textbook				
27	Male	Russia	22	Employed full-time			Yes		Data Analyst	Fine	Employed by	TensorFlow	Genetic & Ev	Python	Dataset aggr	Arxiv,College	Very useful		Very useful	Somewhat u
28	Male	Netherlands	51	Employed full-time			Yes		Engineer	Poorly	Employed by	I don't plan c	I don't plan c	R	I collect my c	Blogs,Tutoring/mentoring		Very useful		
29	Male	Colombia	34	Employed full-time			Yes		Data Scientis	Fine	Employed by	Spark / MLlib	Ensemble M	Python	Google Sear	Online courses,Personal Projects,Stack Overflow Q&A				
30	Male	Germany	41	Independent contractor, freelancer, or s			Yes		Data Scientis	Fine	Self-employe	I don't plan c	Factor Analy	Python	Dataset aggr	Arxiv,Blogs,C	Very useful	Very useful		Somewhat u
31	Female	Canada	32	Not employe	Yes	Yes, I'm focused on learning mostly data science skills						Amazon Wel	Genetic & Ev	C/C++/C#	Dataset aggr	Arxiv,College	Somewhat useful		Very useful	
32	Male	Denmark	53	Employed full-time			Yes		Business Ana	Fine	Employed by	professional	Proprietary A	Python	Dataset aggr	Blogs,Friends network,Per		Very useful		
33	Male	Poland	29	Employed full-time			Yes		Software De	Fine	Employed by	TensorFlow	Deep learnin	Python	Dataset aggr	Kaggle,Online courses,Personal Projects,Stack Overflow Q&A,Textbook				
34	Male	United Kingd	36	Employed full-time			Yes		Data Scientis	Poorly	Employed by	Microsoft Az	Proprietary A	Python	University/N	Arxiv,Blogs,S	Very useful	Somewhat useful		
35	Male	Russia	34	Employed full-time			Yes		Machine Lea	Perfectly	Employed by	Python	Deep learnin	Python	Google Sear	Arxiv,Confere	Somewhat useful			
36	Male	United State	35	Employed full-time			Yes		Engineer	Fine	Employed by	Spark / MLlib	Deep learnin	Python	Dataset aggr	Kaggle,Online courses,Personal Projects,Stack Overflow Q&A,Textbook				

multipleChoiceResponses

Ready

Some answers are well defined (multiple choice).

AutoSave OFF

multipleC

Home Insert Page Layout Formulas Data Review View

Paste Cut Copy Format

Calibri (Body) 12

Wrap Text Merge & Center

Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To pres

A1 Formula Bar GenderSelect

	A	B	C	D	E	F	G	H	I	J
1	GenderSelec	Country	Age	Employment	StudentStatu	LearningData	CodeWriter	CareerSwitch	CurrentJobTi	TitleFit
2	Non-binary, genderqueer,		NA	Employed full-time			Yes		DBA/Databa	Fine
3	Female	United State	30	Not employed, but looking for work						
4	Male	Canada	28	Not employed, but looking for work						
5	Male	United State	56	Independent contractor, freelancer, or s			Yes		Operations R	Poorly
6	Male	Taiwan	38	Employed full-time			Yes		Computer Sc	Fine
7	Male	Brazil	46	Employed full-time			Yes		Data Scientis	Fine
8	Male	United State	35	Employed full-time			Yes		Computer Sc	Fine
9	Female	India	22	Employed full-time			No	Yes	Software De	Fine
10	Female	Australia	43	Employed full-time			Yes		Business Ana	Fine
11	Male	Russia	33	Employed full-time			Yes		Software De	Fine
12	Female	Russia	20	Not employe	Yes	Yes, I'm focused on learning mostly data science skills				
13	Male	India	27	Employed full-time			Yes		Data Scientis	Fine
14	Male	Brazil	26	Employed full-time			No	Yes	Engineer	Fine
15	Male	Netherlands	54	Employed full-time			No	No		
16	Male	Taiwan	26	Employed full-time			Yes		Software De	Fine
17	Male	United State	58	Independent contractor, freelancer, or s			Yes		DBA/Databa	Poorly
18	Male	Italy	58	Employed full-time			No	No		
19	Male	United Kingd	24	Employed full-time			No	No		

multipleChoiceResponses

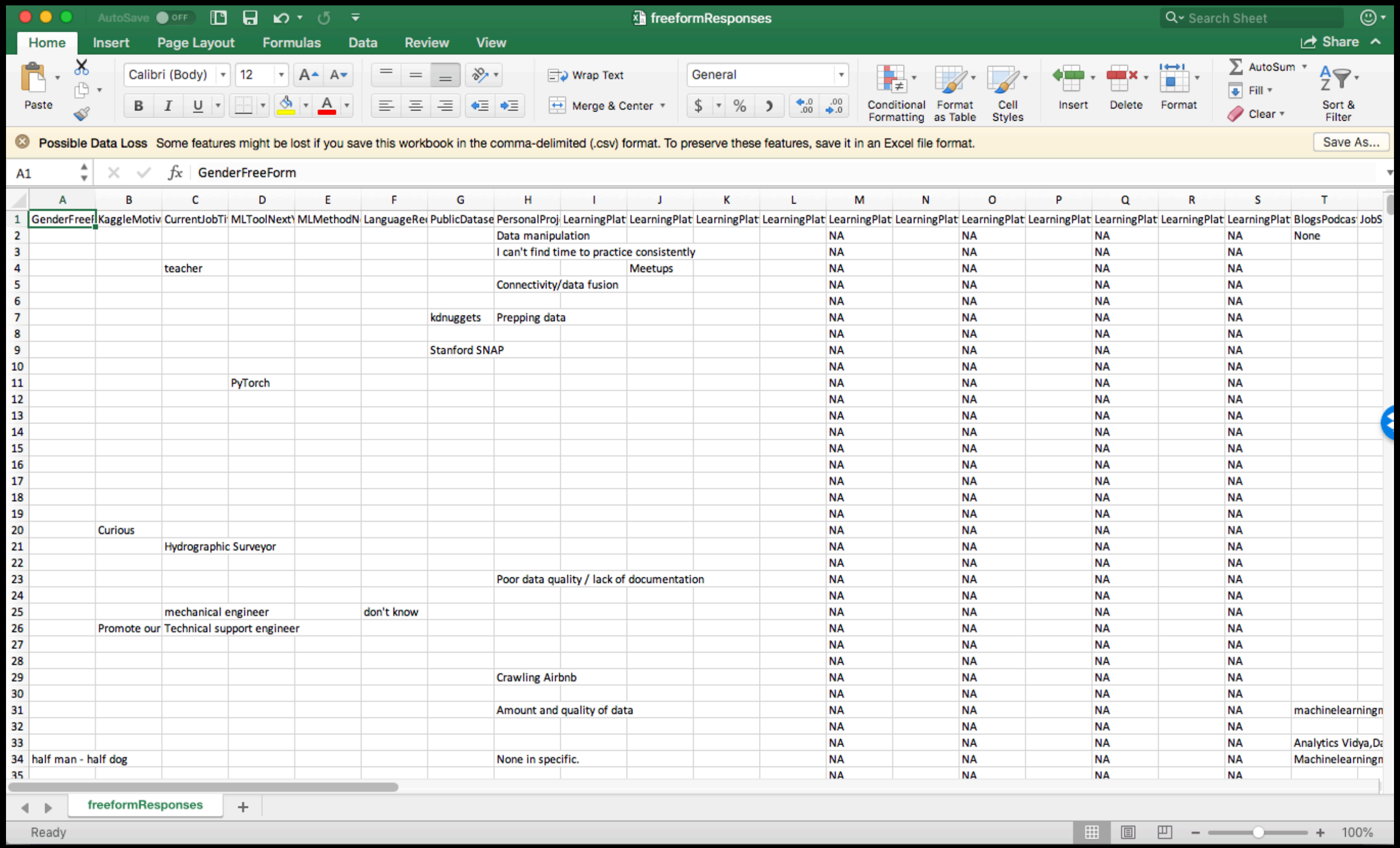
Wrap Text

General

Merge & Center

-delimited (.csv) format. To preserve these features, save it in an E

	I	J	K	L	M
witch	CurrentJobTi	TitleFit	CurrentEmpl	MLToolNextY	MLMethodN



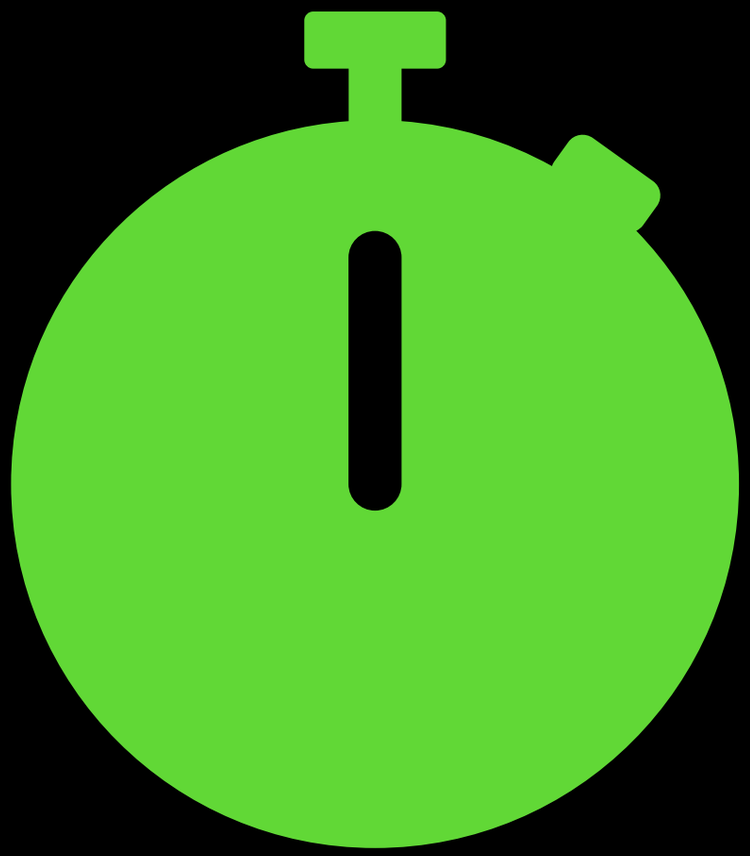
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	GenderFreef	KaggleMotiv	CurrentJobTi	MLToolNextY	MLMethodN	LanguageRe	PublicDatase	PersonalProj	LearningPlat	LearningPlat	LearningPlat	LearningPlat	LearningPlat	LearningPlat	LearningPlat	LearningPlat	LearningPlat	LearningPlat	LearningPlat	BlogsPodcas
2								Data manipulation					NA		NA		NA		NA	None
3								I can't find time to practice consistently					NA		NA		NA		NA	
4			teacher							Meetups			NA		NA		NA		NA	
5								Connectivity/data fusion					NA		NA		NA		NA	
6													NA		NA		NA		NA	
7							kdnuggets	Prepping data					NA		NA		NA		NA	
8													NA		NA		NA		NA	
9							Stanford SNAP						NA		NA		NA		NA	
10													NA		NA		NA		NA	
11				PyTorch									NA		NA		NA		NA	
12													NA		NA		NA		NA	
13													NA		NA		NA		NA	
14													NA		NA		NA		NA	
15													NA		NA		NA		NA	
16													NA		NA		NA		NA	
17													NA		NA		NA		NA	
18													NA		NA		NA		NA	
19													NA		NA		NA		NA	
20		Curious											NA		NA		NA		NA	
21			Hydrographic Surveyor										NA		NA		NA		NA	
22													NA		NA		NA		NA	
23								Poor data quality / lack of documentation					NA		NA		NA		NA	
24													NA		NA		NA		NA	
25			mechanical engineer			don't know							NA		NA		NA		NA	
26		Promote our	Technical support engineer										NA		NA		NA		NA	
27													NA		NA		NA		NA	
28													NA		NA		NA		NA	
29								Crawling Airbnb					NA		NA		NA		NA	
30													NA		NA		NA		NA	
31								Amount and quality of data					NA		NA		NA		NA	machinelearningn
32													NA		NA		NA		NA	
33													NA		NA		NA		NA	Analytics Vidya,Da
34	half man - half dog							None in specific.					NA		NA		NA		NA	Machinelearningn
35													NA		NA		NA		NA	

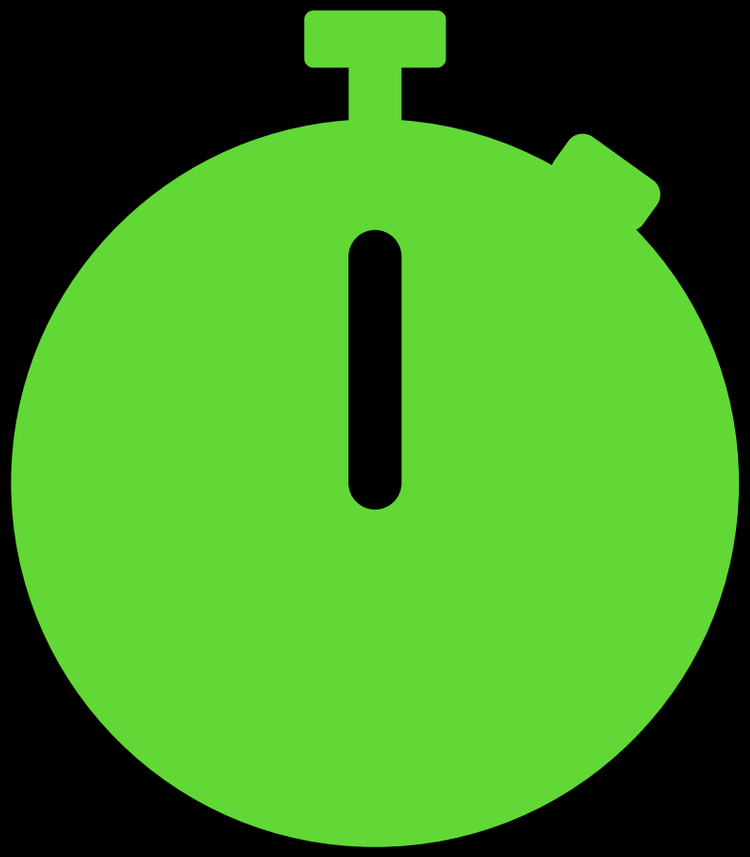
Some answers are not well defined (free form).

25			mechanical engineer	
26		Promote our	Technical support engineer	
27				
28				
29				
30				
31				
32	**Gender (free form).**			
33				
34	half man - half dog			
35				

freeformResponses

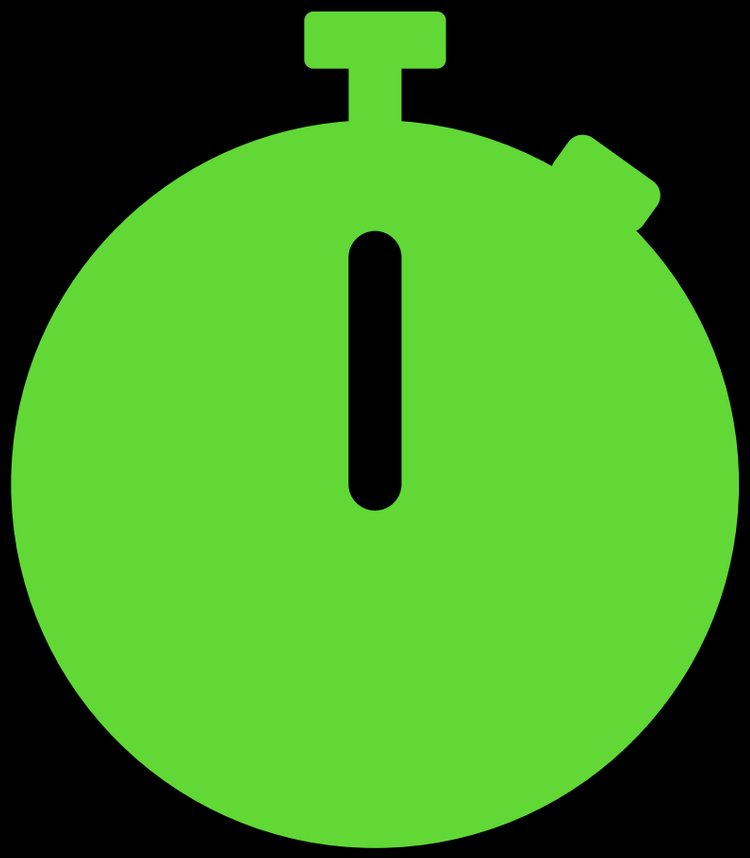
Ready





Find another neighbor.

How does the Kaggle dataset help users to understand the meaning of the data?



Find another neighbor.

How does the Kaggle dataset help users to understand the meaning of the data?

Do academic datasets usually come with good descriptions?

Let us pretend we don’t know much about the data.

In [1]:
```
%matplotlib inline
```

In [2]:
```
cd '/Users/tstoeger/Dropbox/Work/kaggle_survey'
```

```
/Users/tstoeger/Dropbox/Work/kaggle_survey
```

In [3]:
```
import pandas as pd
import seaborn as sns
```

In [1]:
```
%matplotlib inline
```

In [2]:
```
cd '/Users/tstoeger/Dropbox/Work/kaggle_survey'
```

```
/Users/tstoeger/Dropbox/Work/kaggle_survey
```

In [3]:
```
import pandas as pd
import seaborn as sns
```

In [4]:
```
df_multiple_choice = pd.read_csv(
    './multipleChoiceResponses.csv',
    encoding='latin-1', # files have some special characters
    low_memory=False)   # some columns have mixed type
```

In [1]:
```
%matplotlib inline
```

In [2]:
```
cd '/Users/tstoeger/Dropbox/Work/kaggle_survey'
```

```
/Users/tstoeger/Dropbox/Work/kaggle_survey
```

In [3]:
```
import pandas as pd
import seaborn as sns
```

In [4]:
```
df_multiple_choice = pd.read_csv(
    './multipleChoiceResponses.csv',
    encoding='latin-1', # files have some special characters
    low_memory=False)   # some columns have mixed type
```

In [5]:
```
df_multiple_choice.shape   # get idea of number of datasets
```

Out[5]:
```
(16716, 228)
```

In [1]:
```
%matplotlib inline
```

In [2]:
```
cd '/Users/tstoeger/Dropbox/Work/kaggle_survey'
```

```
/Users/tstoeger/Dropbox/Work/kaggle_survey
```

In [3]:
```
import pandas as pd
import seaborn as sns
```

In [4]:
```
df_multiple_choice = pd.read_csv(
    './multipleChoiceResponses.csv',
    encoding='latin-1', # files have some special characters
    low_memory=False)   # some columns have mixed type
```

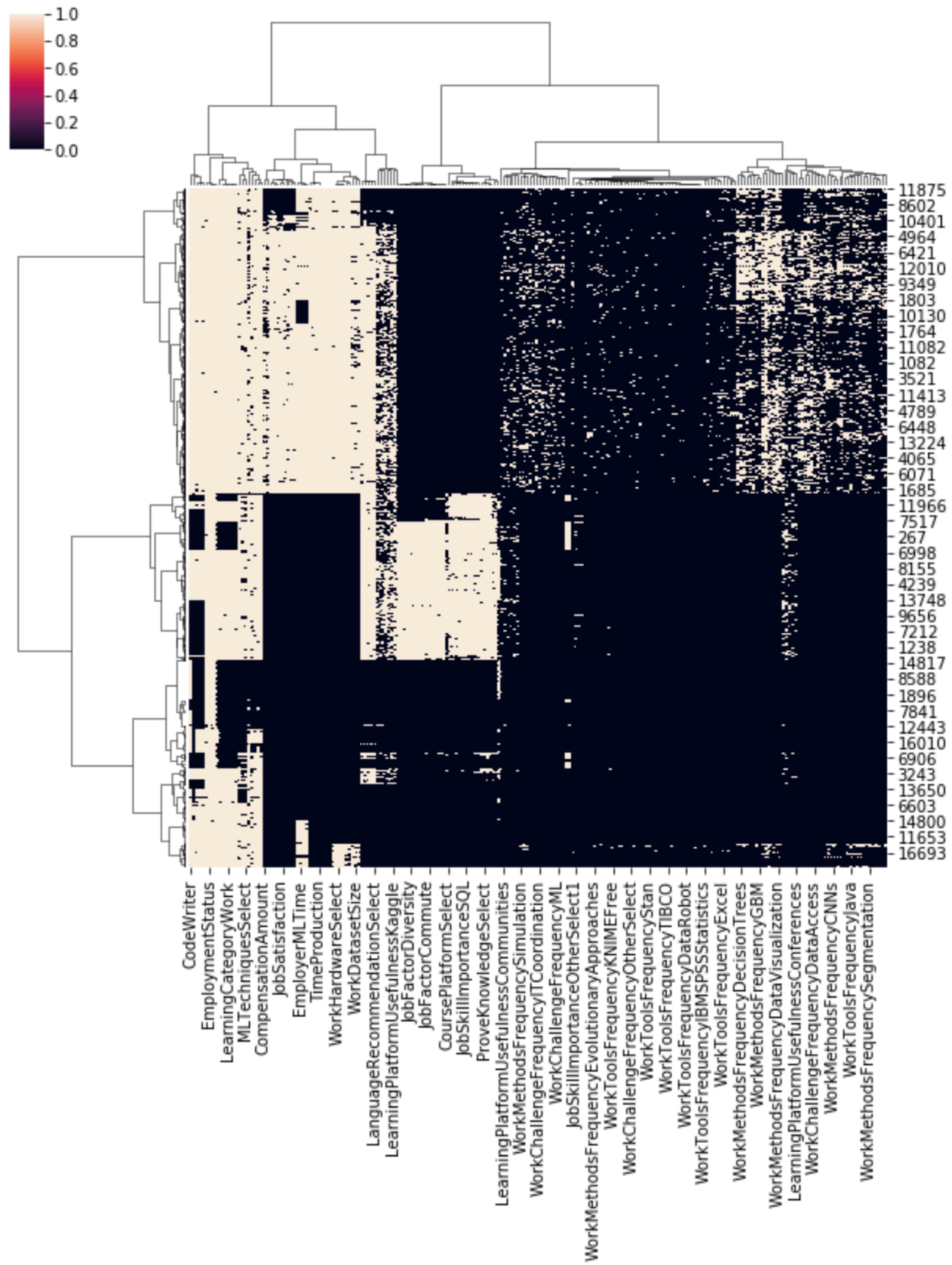
In [5]:
```
df_multiple_choice.shape   # get idea of number of datasets
```

Out[5]:
```
(16716, 228)
```

In [6]:
```
sns.clustermap(                    # visualize presence of data
    df_multiple_choice.notnull(),
    method='ward')
```

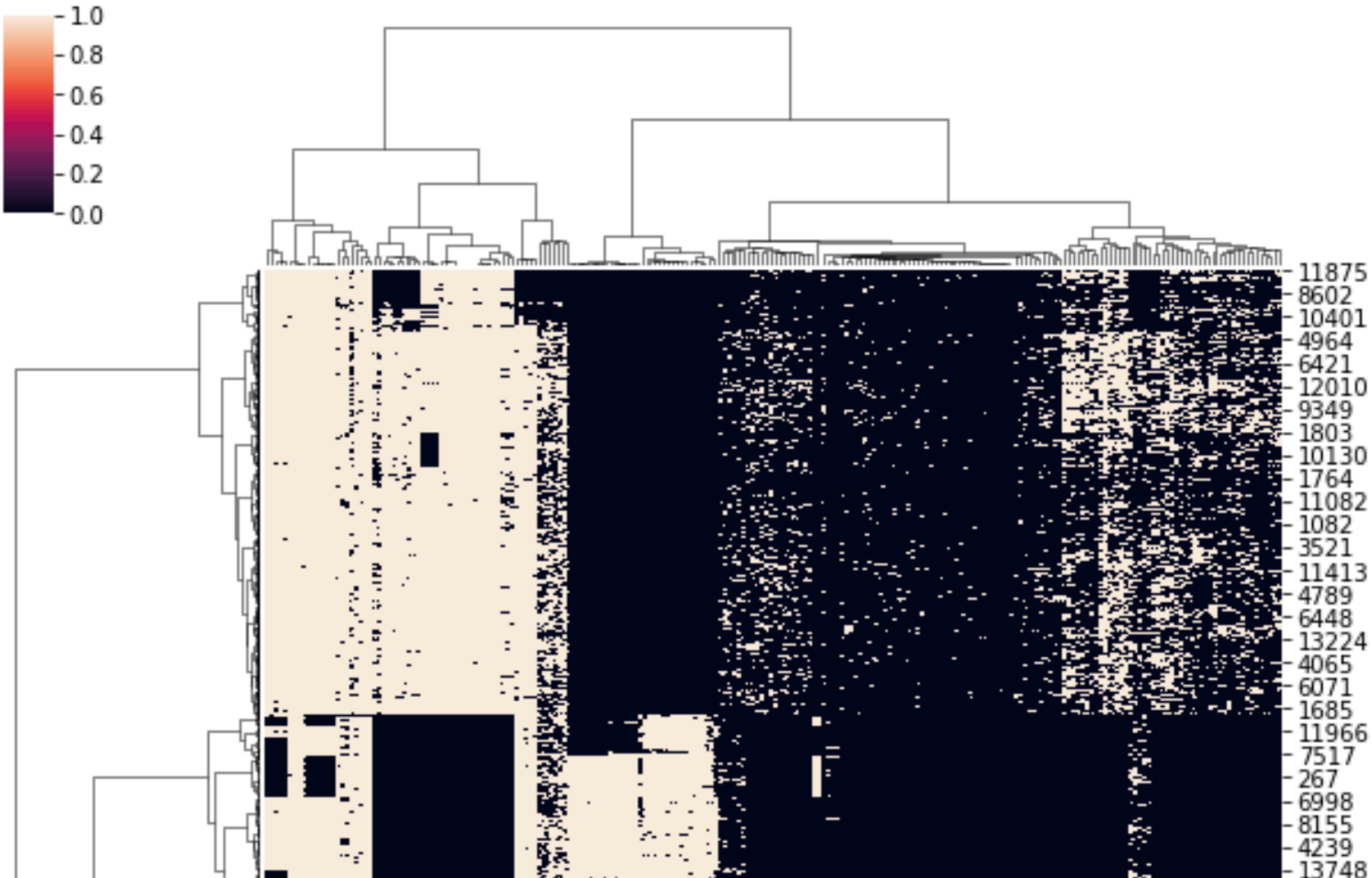
In [6]:
```
sns.clustermap(             # visualize presence of data
    df_multiple_choice.notnull(),
    method='ward')
```

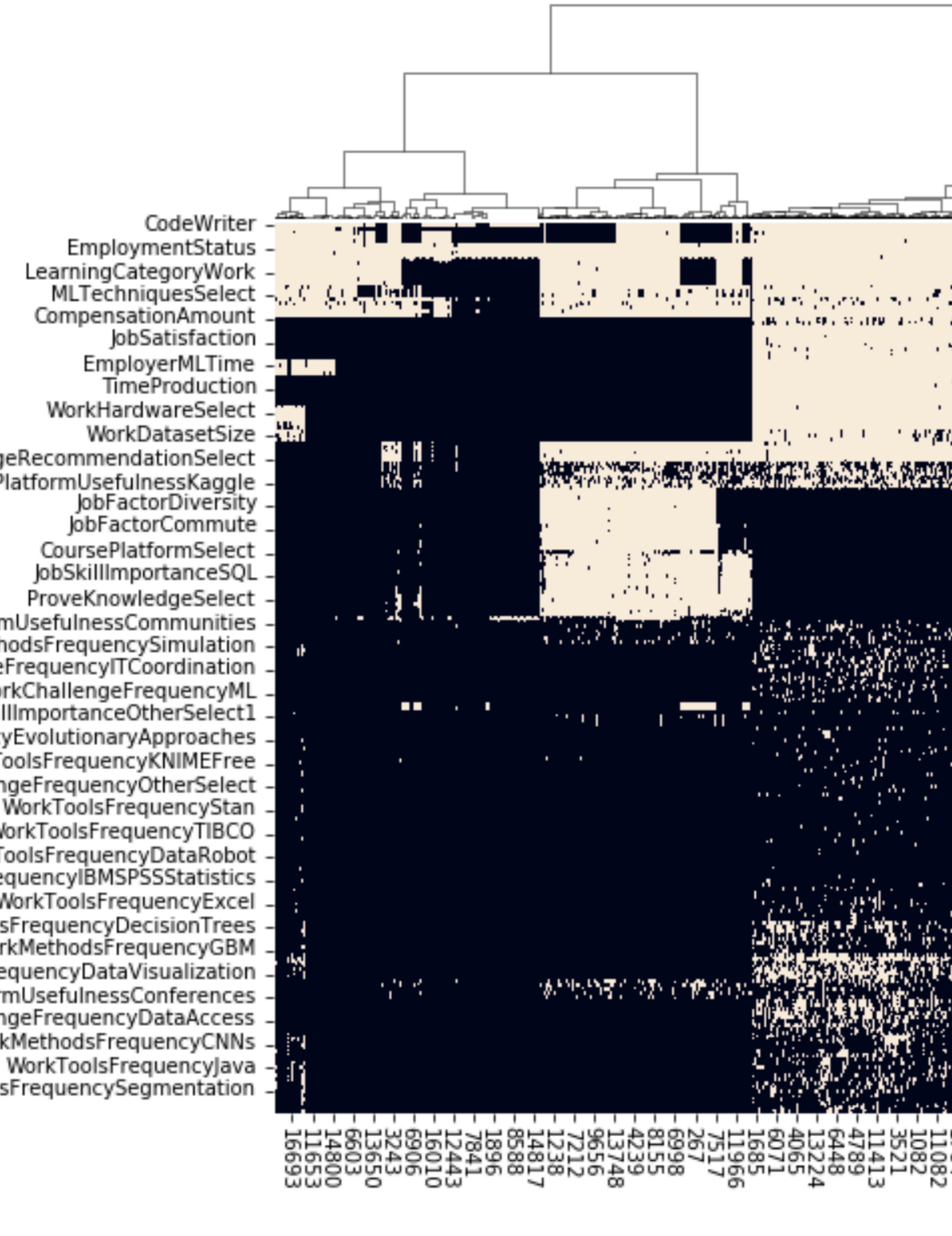
Out[6]: <seaborn.matrix.ClusterGrid at 0x10cd02a58>



```
sns.clustermap(                    # visualize presence of data
    df_multiple_choice.notnull(),
    method='ward')
```

```
<seaborn.matrix.ClusterGrid at 0x10cd02a58>
```




CodeWriter
EmploymentStatus
LearningCategoryWork
MLTechniquesSelect
CompensationAmount
JobSatisfaction
EmployerMLTime
TimeProduction
WorkHardwareSelect
WorkDatasetSize
geRecommendationSelect
PlatformUsefulnessKaggle
JobFactorDiversity
JobFactorCommute
CoursePlatformSelect
JobSkillImportanceSQL
ProveKnowledgeSelect
mUsefulnessCommunities
hodsFrequencySimulation
eFrequencyITCoordination
rkChallengeFrequencyML
IlImportanceOtherSelect1
yEvolutionaryApproaches
oolsFrequencyKNIMEFree
ngeFrequencyOtherSelect
WorkToolsFrequencyStan
orkToolsFrequencyTIBCO
ToolsFrequencyDataRobot
equencyIBMSPSSStatistics
WorkToolsFrequencyExcel
sFrequencyDecisionTrees
rkMethodsFrequencyGBM
equencyDataVisualization
mUsefulnessConferences
ngeFrequencyDataAccess
kMethodsFrequencyCNNs
WorkToolsFrequencyJava
sFrequencySegmentation
16693
11653
14800
6603
13650
3243
6906
16010
12443
7841
1896
8588
14817
1238
7212
9656
13748
4239
8155
6998
267
7517
11966
1685
6071
4065
13224
6448
4789
11413
3521
1082
11082

Always look at data!

```
In [7]: df_multiple_choice
```

Out[7]:

	GenderSelect	Country	Age	EmploymentStatus	StudentStatus	LearningDataScience
0	Non-binary, genderqueer, or gender non-conforming	NaN	NaN	Employed full-time	NaN	NaN
1	Female	United States	30.0	Not employed, but looking for work	NaN	NaN
2	Male	Canada	28.0	Not employed, but looking for work	NaN	NaN
3	Male	United States	56.0	Independent contractor, freelancer, or self-em...	NaN	NaN
4	Male	Taiwan	38.0	Employed full-time	NaN	NaN
5	Male	Brazil	46.0	Employed full-time	NaN	NaN
6	Male	United States	35.0	Employed full-time	NaN	NaN
7	Female	India	22.0	Employed full-time	NaN	NaN
8	Female	Australia	43.0	Employed full-time	NaN	NaN
9	Male	Russia	33.0	Employed full-time	NaN	NaN

Always look at data!

```
In [7]: df_multiple_choice
```

Out[7]:

	GenderSelect	Country	Age	EmploymentStatus	StudentStatus	LearningDataScience
0	Non-binary, genderqueer, or gender non-conforming	NaN	NaN	Empl…time	NaN	NaN
1	Female	United States	30.0		NaN	NaN
2	Male	Canada	…		…aN	NaN
3	Male	United States	56.0	free…	NaN	NaN
4	Male	Taiwan	38.0	Employed full-time	NaN	NaN
						NaN
						NaN
7	Female	India	22.0	Employed full-time	NaN	NaN
8	Female	Australia	43.0	Employed full-time	NaN	NaN
9	Male	Russia	33.0	Employed full-time	NaN	NaN

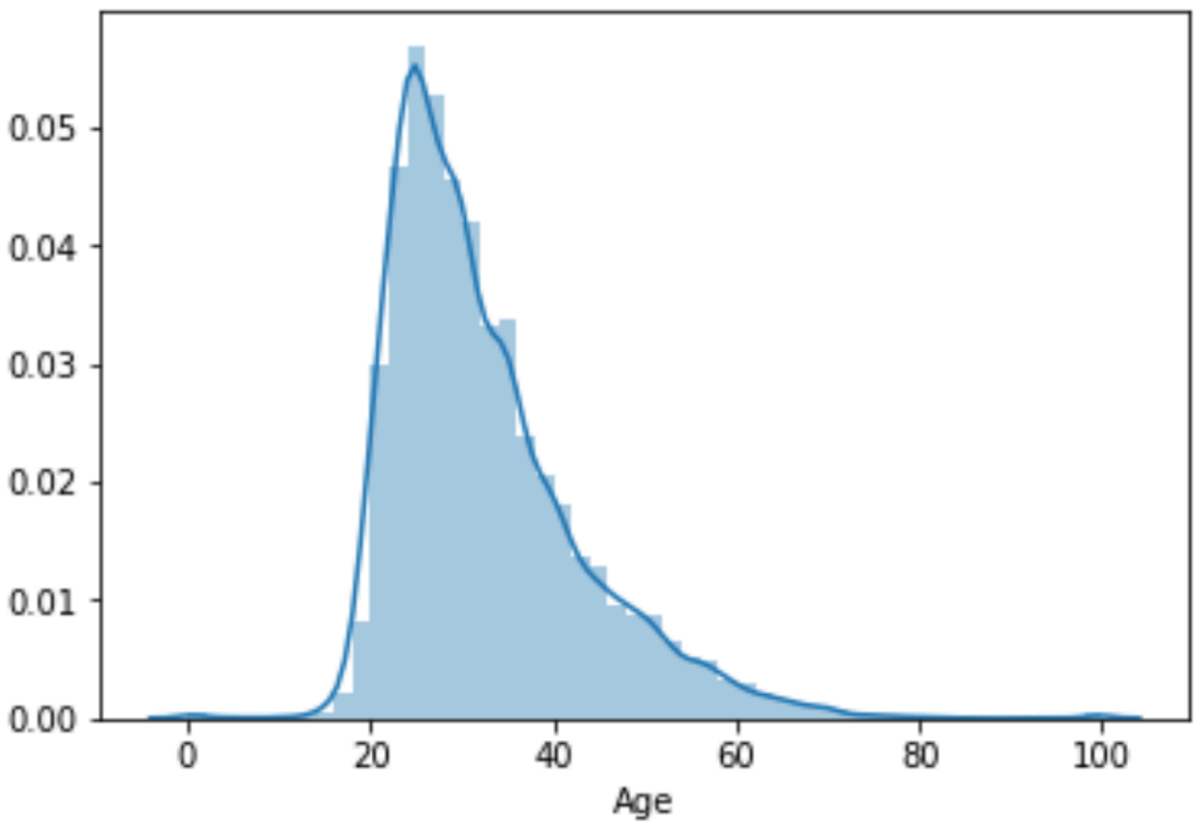
In a nicely organized data set, columns are organized in a meaningful way

```
In [8]: sns.distplot(
            df_multiple_choice['Age'].dropna(),
        )
```

Sanity-check numbers, and test for unusual values.

In [8]:
```
sns.distplot(
    df_multiple_choice['Age'].dropna(),
)
```

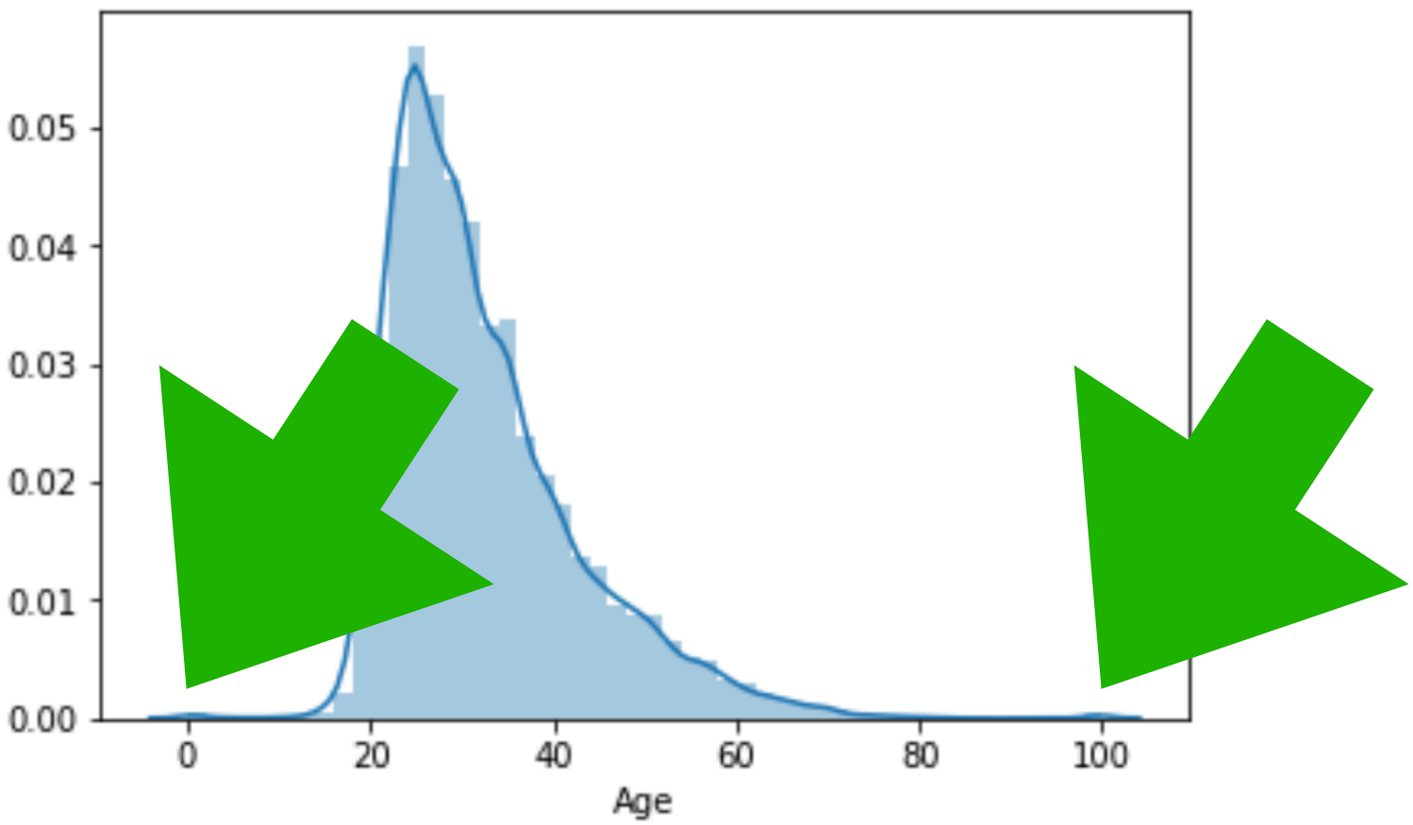
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x1a124307f0>



Sanity-check numbers.

```
In [8]: sns.distplot(
            df_multiple_choice['Age'].dropna(),
        )
```

```
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x1a124307f0>
```



Sanity-check numbers.

Workflow

- Anticipate problems
- **Get overview**
- Clean data
- Find connections

break

Setup

- Environment
- Organizing data
- Coding

Wor

What are two strategies?

- Anticipate problems
- **Get overview**
- Clean data
- Find connections

break

- Environment
- Organizing data
- Coding

Data science is: 80% *data cleaning* and 20% complaining about *data cleaning*.

freeformResponses

Review View

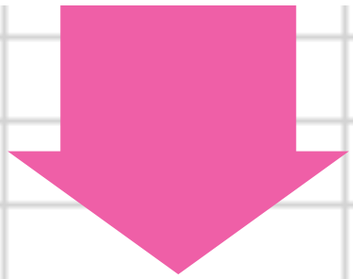
PersonalProjectsChallenge

this workbook in the comma-delimited (.csv) format. To preserve these feature

You are not alone.

F	G		I	J	K	L
nguageRe	PublicDatase	PersonalProj	LearningPlat	LearningPlat	LearningPlat	Learning
		Data manipulation				
		I can't find time to practice consistently				
				Meetups		
		Connectivity/data fusion				
	kdnuggets	Prepping data				
	Stanford SNAP					

PersonalProjectsChallenge



	Poor data quality / lack of documentation
n't know	
	Crawling Airbnb
	Amount and quality of data
	None in specific.

You are not alone.

Main tactics for data cleaning.

- Codify anticipation of unanticipated.
- Regular expressions.
- Entity matching.

Main tactics for data cleaning.

- Codify anticipation of unanticipated.
- Regular expressions.
- Entity matching.

Main tactics for data cleaning.

- Codify anticipation of unanticipated.
- Regular expressions.
- Entity matching.

Main tactics for data cleaning.

- Codify anticipation of unanticipated.
- Regular expressions.
- Entity matching.

Salaries of data scientists (part of Kaggle Survey)

In [13]:
```
# For the sake of this example, let us only consider
# salaries that have been converted to USD
f = df_multiple_choice['CompensationCurrency'] == 'USD'
df_multiple_choice = df_multiple_choice[f]
```

In [13]:
```
# For the sake of this example, let us only consider
# salaries that have been converted to USD
f = df_multiple_choice['CompensationCurrency'] == 'USD'
df_multiple_choice = df_multiple_choice[f]
```

In [14]:
```
# For the sake of this example, let us only consider
# records, where CompensationAmount is defined
# (we still keep records, if something else, e.g.
# StudentStatus, whould be not defined)
df_multiple_choice = df_multiple_choice.dropna(
    subset=['CompensationAmount']
)
```

In [13]:
```
# For the sake of this example, let us only consider
# salaries that have been converted to USD
f = df_multiple_choice['CompensationCurrency'] == 'USD'
df_multiple_choice = df_multiple_choice[f]
```

In [14]:
```
# For the sake of this example, let us only consider
# records, where CompensationAmount is defined
# (we still keep records, if something else, e.g.
# StudentStatus, whould be not defined)
df_multiple_choice = df_multiple_choice.dropna(
    subset=['CompensationAmount']
)
```

In [15]:
```
# Manually inspect the data in CompensationAmount
df_multiple_choice['CompensationAmount']
```

Out[15]:
```
3          250,000
21           20000
22          100000
34          133000
37           80000
61           15000
75          215000
86           83500
```

In [16]:
```
# Let us convert these values to a number
df_multiple_choice['CompensationAmount'] = df_multiple_choice[
    'CompensationAmount'].astype(float)
```

Error

```
# Let us convert these values to a number
df_multiple_choice['CompensationAmount'] = df_multiple_choice[
    'CompensationAmount'].astype(float)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-16-f163d2412d49> in <module>()
      1 df_multiple_choice['CompensationAmount'] = df_multiple_choice[
----> 2     'CompensationAmount'].astype(float)

~/anaconda3/lib/python3.6/site-packages/pandas/util/_decorators.py in wrapper(*args, **kwargs)
    176                 else:
    177                     kwargs[new_arg_name] = new_arg_value
--> 178             return func(*args, **kwargs)
    179         return wrapper
    180     return _deprecate_kwarg

~/anaconda3/lib/python3.6/site-packages/pandas/core/generic.py in astype(self, dtype, copy, errors, **kwargs)
   4995             # else, only a single dtype is given
   4996             new_data = self._data.astype(dtype=dtype, copy=copy, errors=errors,
-> 4997                                          **kwargs)
   4998             return self._constructor(new_data).__finalize__(self)
   4999

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in astype(self, dtype, **kwargs)
   3712
   3713     def astype(self, dtype, **kwargs):
-> 3714         return self.apply('astype', dtype=dtype, **kwargs)
   3715
   3716     def convert(self, **kwargs):

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in apply(self, f, axes, filter, do_integrity_check,
consolidate, **kwargs)
   3579
   3580             kwargs['mgr'] = self
-> 3581             applied = getattr(b, f)(**kwargs)
   3582             result_blocks = _extend_blocks(applied, result_blocks)
   3583

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in astype(self, dtype, copy, errors, values, **kwar
gs)
    573     def astype(self, dtype, copy=False, errors='raise', values=None, **kwargs):
    574         return self._astype(dtype, copy=copy, errors=errors, values=values,
--> 575                             **kwargs)
    576
    577     def _astype(self, dtype, copy=False, errors='raise', values=None,

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in _astype(self, dtype, copy, errors, values, klass
, mgr, **kwargs)
    662
    663                 # _astype_nansafe works fine with 1-d only
--> 664                 values = astype_nansafe(values.ravel(), dtype, copy=True)
    665                 values = values.reshape(self.shape)
    666

~/anaconda3/lib/python3.6/site-packages/pandas/core/dtypes/cast.py in astype_nansafe(arr, dtype, copy)
    728
    729     if copy:
--> 730         return arr.astype(dtype, copy=True)
    731     return arr.view(dtype)
    732

ValueError: could not convert string to float: '85,000'
```

Could not convert string to float: ’85,000’

Only fix what you want to fix. (85,000 and similar)

```
ValueError: could not convert string to float: '85,000'
```

```
In [17]: f = df_multiple_choice['CompensationAmount'].str.contains(
             '[0-9]*,[0-9]{3}$')   # create a highly specific regular expression
```

[0-9]*,[0-9]{3}$

Only fix what you want to fix. (85,000 and similar)

```
ValueError: could not convert string to float: '85,000'
```

```
In [17]: f = df_multiple_choice['CompensationAmount'].str.contains(
             '[0-9]*,[0-9]{3}$')   # create a highly specific regular expression
```

Can someone translate this?

[0-9]*,[0-9]{3}$

Only fix what you want to fix. (85,000 and similar)

```
ValueError: could not convert string to float: '85,000'
```

In [17]:
```
f = df_multiple_choice['CompensationAmount'].str.contains(
    '[0-9]*,[0-9]{3}$')   # create a highly specific regular expression
```

In [18]:
```
df_multiple_choice.loc[f, 'CompensationAmount'] = df_multiple_choice.loc[
    f, 'CompensationAmount'].str.replace(',', '')
```

Can someone translate this?

[0-9]*,[0-9]{3}$

Only fix what you want to fix. (85,000 and similar)

```
ValueError: could not convert string to float: '85,000'
```

In [17]:
```
f = df_multiple_choice['CompensationAmount'].str.contains(
    '[0-9]*,[0-9]{3}$')   # create a highly specific regular expression
```

In [18]:
```
df_multiple_choice.loc[f, 'CompensationAmount'] = df_multiple_choice.loc[
    f, 'CompensationAmount'].str.replace(',', '')
```

Can someone translate this?

[0-9]*,[0-9]{3}$

Trick question: Could this be shorter to solve our problem?

Cleaning data is an interactive process.

```
In [19]: df_multiple_choice['CompensationAmount'] = df_multiple_choice[
             'CompensationAmount'].astype(float)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-19-f163d2412d49> in <module>()
      1 df_multiple_choice['CompensationAmount'] = df_multiple_choice[
----> 2     'CompensationAmount'].astype(float)

~/anaconda3/lib/python3.6/site-packages/pandas/util/_decorators.py in wrapper(*args, **kwargs)
    176                 else:
    177                     kwargs[new_arg_name] = new_arg_value
--> 178             return func(*args, **kwargs)
    179         return wrapper
    180     return _deprecate_kwarg

~/anaconda3/lib/python3.6/site-packages/pandas/core/generic.py in astype(self, dtype, copy, errors, **kwargs)
   4995             # else, only a single dtype is given
   4996             new_data = self._data.astype(dtype=dtype, copy=copy, errors=errors,
-> 4997                                          **kwargs)
   4998             return self._constructor(new_data).__finalize__(self)
   4999

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in astype(self, dtype, **kwargs)
   3712
   3713     def astype(self, dtype, **kwargs):
-> 3714         return self.apply('astype', dtype=dtype, **kwargs)
   3715
   3716     def convert(self, **kwargs):

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in apply(self, f, axes, filter, do_integrity_check,
consolidate, **kwargs)
   3579
   3580             kwargs['mgr'] = self
-> 3581             applied = getattr(b, f)(**kwargs)
   3582             result_blocks = _extend_blocks(applied, result_blocks)
   3583

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in astype(self, dtype, copy, errors, values, **kwar
gs)
    573     def astype(self, dtype, copy=False, errors='raise', values=None, **kwargs):
    574         return self._astype(dtype, copy=copy, errors=errors, values=values,
--> 575                             **kwargs)
    576
    577     def _astype(self, dtype, copy=False, errors='raise', values=None,

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in _astype(self, dtype, copy, errors, values, klass
, mgr, **kwargs)
    662
    663                 # _astype_nansafe works fine with 1-d only
--> 664                 values = astype_nansafe(values.ravel(), dtype, copy=True)
    665                 values = values.reshape(self.shape)
    666

~/anaconda3/lib/python3.6/site-packages/pandas/core/dtypes/cast.py in astype_nansafe(arr, dtype, copy)
    728
    729     if copy:
--> 730         return arr.astype(dtype, copy=True)
    731     return arr.view(dtype)
    732

ValueError: could not convert string to float: '-'
```

Main tactics for data cleaning.

- Codify anticipation of unanticipated.
  (e.g.: assert, raise Error)
- Regular expressions.
- Entity matching.

e.g: match Thomas Stoeger to Thomas Stoger

Main tactics for data cleaning.

- Codify anticipation of unanticipated.
  (e.g.: assert, raise Error)
- Regular expressions.
- Entity matching.

e.g: match Thomas Stoeger to Thomas Stoger

Main tactics for data cleaning.

- Codify anticipation of unanticipated.
  (e.g.: assert, raise Error)
- Regular expressions.
- Entity matching.

e.g: match Thomas Stoeger to Thomas Stoger

Main tactics for data cleaning.

- Codify anticipation of unanticipated.
  (e.g.: assert, raise Error)
- Regular expressions.
- Entity matching.

e.g: match Thomas Stoeger to Thomas Stoger

Missing data can be tricky.

Advance Access publication August 22, 2016

Political Analysis (2016) 24:414–433
doi:10.1093/pan/mpw020

How Multiple Imputation Makes a Difference

Ranjit Lall

Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138
e-mail: ranjitlall@fas.harvard.edu (corresponding author)

Edited by R. Michael Alvarez

Political scientists increasingly recognize that multiple imputation represents a superior strategy for analyzing missing data to the widely used method of listwise deletion. However, there has been little systematic investigation of *how* multiple imputation affects existing empirical knowledge in the discipline. This article presents the first large-scale examination of the empirical effects of substituting multiple imputation for listwise deletion in political science. The examination focuses on research in the major subfield of comparative and international political economy (CIPE) as an illustrative example. Specifically, I use multiple imputation to reanalyze the results of almost every quantitative CIPE study published during a recent five-year period in *International Organization* and *World Politics*, two of the leading subfield journals in CIPE. The outcome is striking: in almost half of the studies, key results "disappear" (by conventional statistical standards) when reanalyzed.

Missing data can be tricky.

Advance Access publication August 22, 2016

Political Analysis (2016) 24:414–433
doi:10.1093/pan/mpw020

How Multiple Imputation Makes a Difference

Ra

Department of Government, Harvard Univers
e-mail: ranjitlall@fas.har

Edited by

Political scientists increasingly recognize that multi
missing data to the widely used method of listw
investigation of *how* multiple imputation affects e
presents the first large-scale examination of the
listwise deletion in political science. The examina
parative and international political economy (CIP
imputation to reanalyze the results of almost ever
year period in *International Organization* and *World*
outcome is striking: in almost half of the studies,
ards) when reanalyzed.

Sometimes, there are smart people who actively work on misleading others.

A rough guide to missing data:

Random

Non-random

A rough guide to missing data:

Random → Discard records

Non-random

A rough guide to missing data:

Random

- Discard records
- Impute based on similar records

Non-random

A rough guide to missing data:

Random

- **Discard records**
- **Impute based on similar records**

Non-random → **Use as information**

A rough guide to missing data:

Random

- Discard records
- Impute based on similar records

Non-random

- Use as information
- Consider as "below measurement limit"

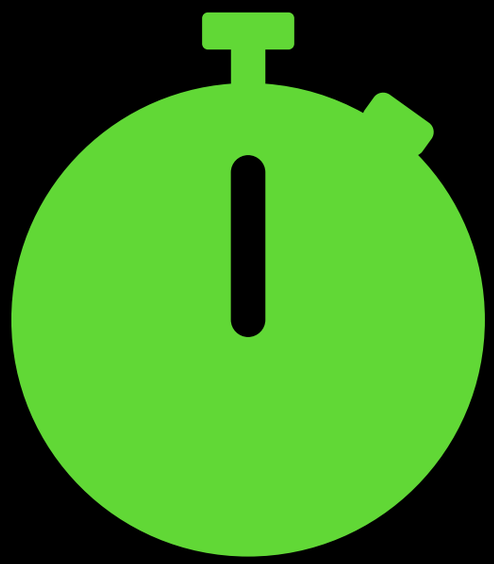
A rough guide to missing data:

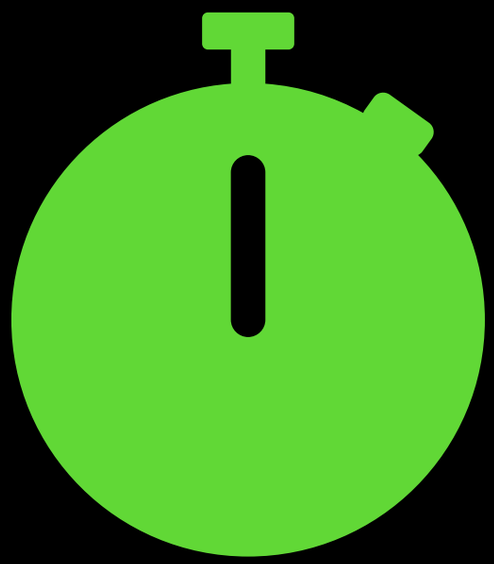
Random

- Discard records
- Impute based on similar records

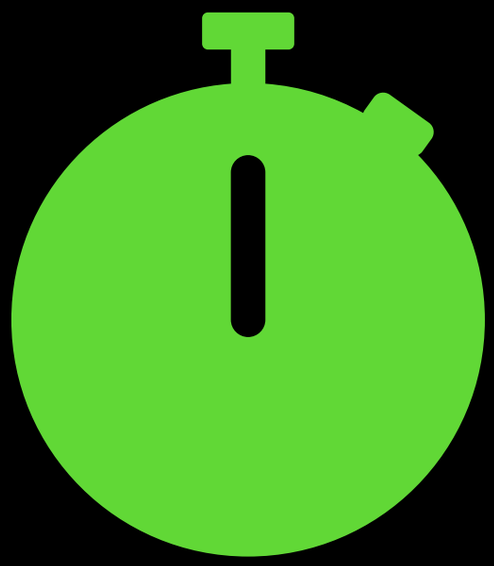
Non-random

- Use as information
- Consider as "below measurement limit"
- Advanced (often domain-specific) imputation





Remember your worst encounter with data.



Remember your worst encounter with data.

How should the data have been cleaned to simplify your work?

Workflow

- Anticipate problems
- Get overview
- Clean data
- **Find connections**

break

Setup

- Computer
- Organizing data
- Coding

Main tactics for finding connections.

- Exploratory visualization
- Correlations
- Machine learning

Main tactics for finding connections.

- Exploratory visualization
- Correlations
- Machine learning

Main tactics for finding connections.

- Exploratory visualization
- Correlations
- Machine learning

Main tactics for finding connections.

- Exploratory visualization
- Correlations
- Machine learning

Is there a gender bias in the salary of data scientists?

Is there a gender bias in the salary of data scientists?

In [108]:
```
f = (
    df_multiple_choice['GenderSelect'].isin(['Male', 'Female'])) & (
    df_multiple_choice['Country'].isin(['United States'])) & (
    df_multiple_choice['EmploymentStatus'].isin(['Employed full-time']) & (
    df_multiple_choice['Age'].isin(range(20, 35))) & (
    df_multiple_choice['CurrentJobTitleSelect'].isin(['Data Scientist']))
)
```

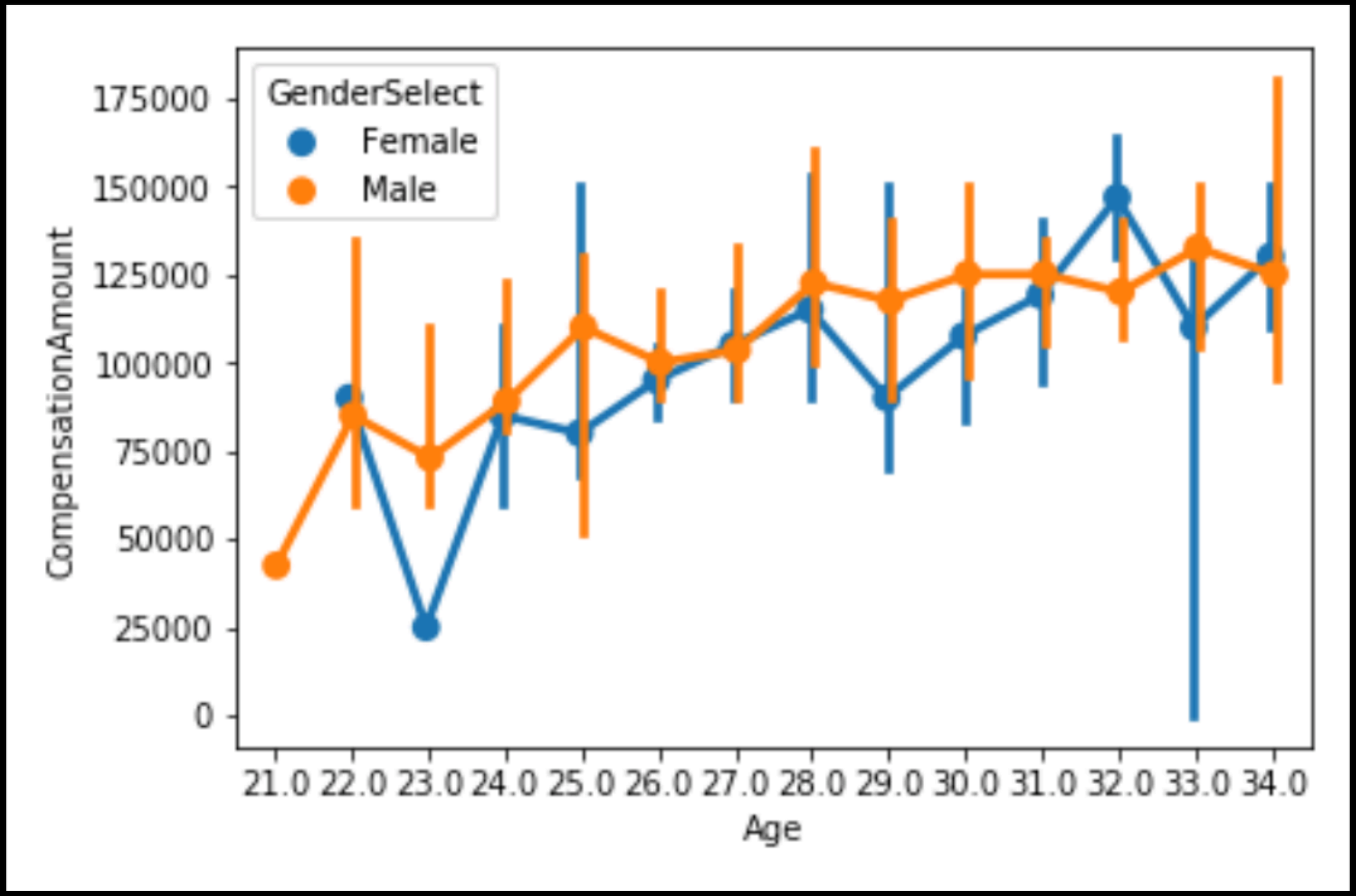
Is there a gender bias in the salary of data scientists?

In [108]:
```
f = (
    df_multiple_choice['GenderSelect'].isin(['Male', 'Female'])) & (
    df_multiple_choice['Country'].isin(['United States'])) & (
    df_multiple_choice['EmploymentStatus'].isin(['Employed full-time']) & (
    df_multiple_choice['Age'].isin(range(20, 35))) & (
    df_multiple_choice['CurrentJobTitleSelect'].isin(['Data Scientist']))
)
```

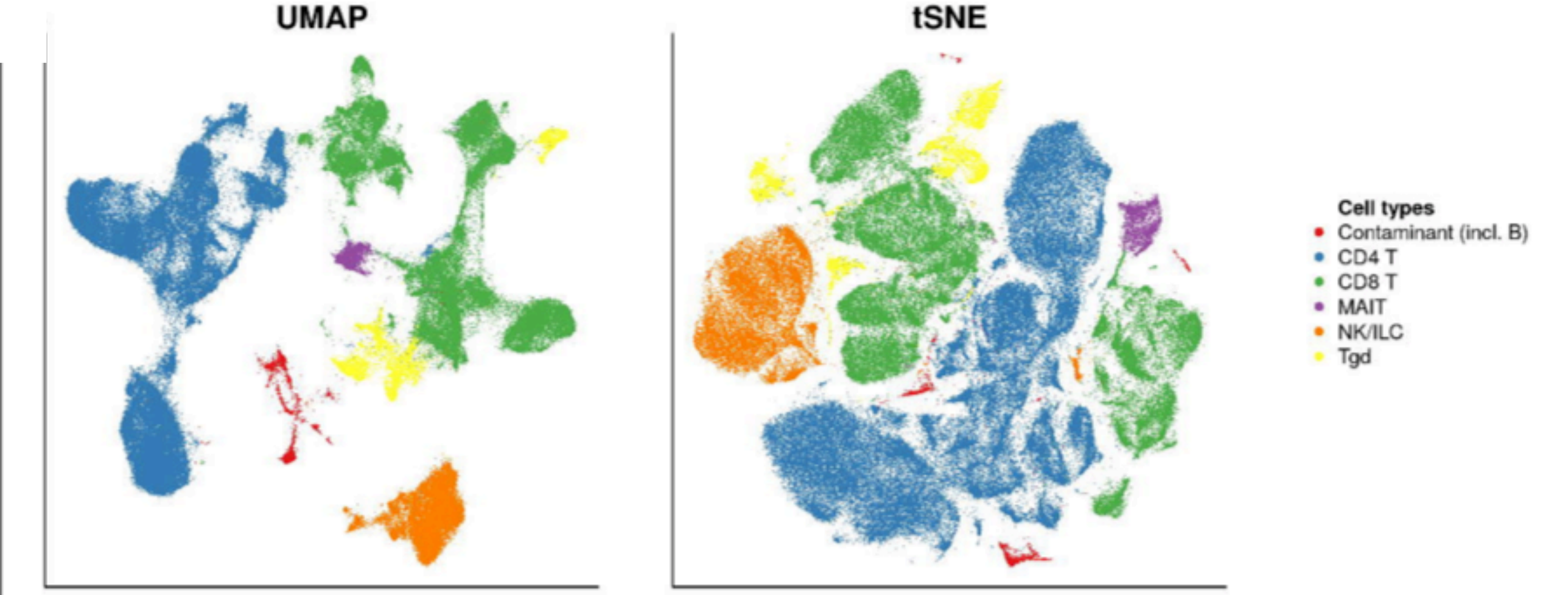
In [109]:
```
import numpy as np
```

In [110]:
```
sns.pointplot(
    x='Age',
    y='CompensationAmount',
    data=df_multiple_choice[f],
    estimator=np.median,
    hue='GenderSelect',
    dodge=True
)
```

Median salary of data scientists at university ages.



Further default tactic: visualize upon dimensionality reduction.



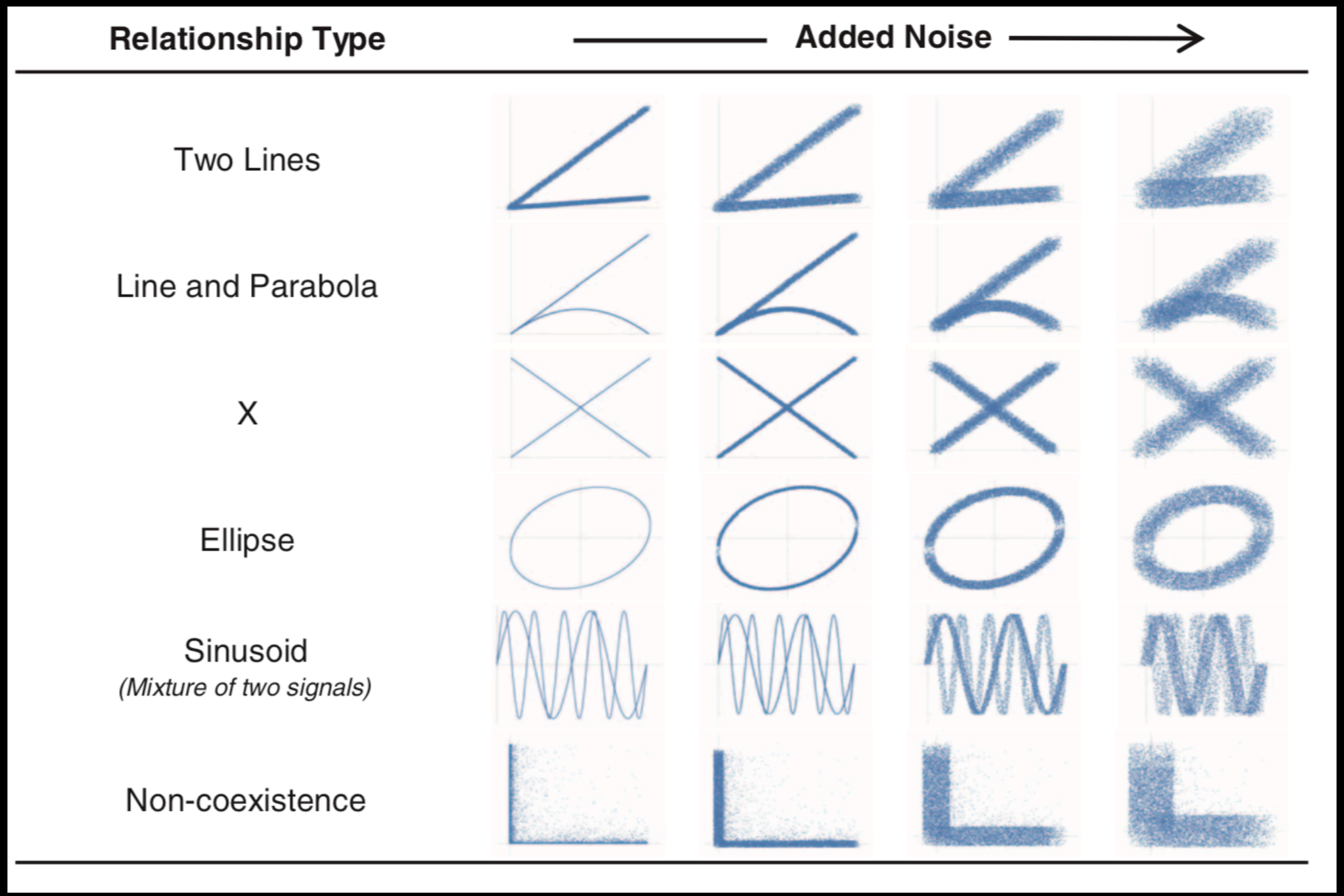
Dimensionality reduction
Fast

Only visualization
Slow

Becht et al., bioRxiv 2018

Distinct types of correlations can embody distinct assumptions on the data.

Distinct types of correlations can embody distinct assumptions on the data.



Reshef et al. 2011

Distinct types of correlations can embody distinct assumptions on the data.

Relationship Type	MIC	Pearson	Spearman	Mutual Information *(KDE)*	Mutual Information *(Kraskov)*	CorGC (Principal Curve-Based)	Maximal Correlation
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal *(Fourier frequency)*	1.00	-0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.31	0.69	0.45	0.49	0.91
Parabolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal *(non-Fourier frequency)*	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal *(varying frequency)*	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76

Reshef et al. 2011

Main tactics for finding connections.

- Exploratory visualization
- Correlations
- Machine learning

Main tactics for finding connections.

- Exploratory visualization
- Correlations
- Machine learning

In [113]:
```
df_multiple_choice['MLMethodNextYearSelect'].value_counts()
```

```
In [113]: df_multiple_choice['MLMethodNextYearSelect'].value_counts()
```

```
Out[113]: Deep learning                                    535
          Neural Nets                                      207
          Bayesian Methods                                 101
          Time Series Analysis                              90
          Genetic & Evolutionary Algorithms                 66
          Anomaly Detection                                 61
          Text Mining                                       52
          Social Network Analysis                           48
          Other                                             46
          Ensemble Methods (e.g. boosting, bagging)         42
          I don't plan on learning a new ML/DS method       33
          Cluster Analysis                                  31
          Monte Carlo Methods                               28
          Support Vector Machines (SVM)                     26
          Proprietary Algorithms                            21
          Random Forests                                    18
          Survival Analysis                                 18
          Regression                                        15
          Rule Induction                                     8
          Link Analysis                                      8
          Decision Trees                                     8
          Uplift Modeling                                    6
          MARS                                               5
          Factor Analysis                                    5
          Association Rules                                  3
          Name: MLMethodNextYearSelect, dtype: int64
```

```
In [113]: df_multiple_choice['MLMethodNextYearSelect'].value_counts()

Out[113]: Deep learning                                        535
          Neural Nets                                         207
          Bayesian Methods                                    101
          Time Series Analysis                                 90
          Genetic & Evolutionary Algorithms                    66
          Anomaly Detection                                    61
          Text Mining                                          52
          Social Network Analysis                              48
          Other                                                46
          Ensemble Methods (e.g. boosting, bagging)            42
          I don't plan on learning a new ML/DS method          33
          Cluster Analysis                                     31
          Monte Carlo Methods                                  28
          Support Vector Machines (SVM)                        26
          Proprietary Algorithms                               21
          Random Forests                                       18
          Survival Analysis                                    18
          Regression                                           15
          Rule Induction                                        8
          Link Analysis                                         8
          Decision Trees                                        8
          Uplift Modeling                                       6
          MARS                                                  5
          Factor Analysis                                       5
          Association Rules                                     3
          Name: MLMethodNextYearSelect, dtype: int64
```

Reminder.

Any questions?

Note: Some discipline-specific approaches have very interesting properties.

TIME SERIES ANALYSIS

Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality

Turn high dimensionality from a curse to a blessing.

Hao Ye and George Sugihara*

In ecological analysis, complexity has been regarded as an obstacle to overcome. Here we present a straightforward approach for addressing complexity in dynamic interconnected systems. We show that complexity, in the form of multiple interacting components, can actually be an asset for studying natural systems from temporal data. The central idea is that multidimensional time series enable system dynamics to be reconstructed from multiple viewpoints, and these viewpoints can be combined into a single model. We show how our approach, multiview embedding (MVE), can improve forecasts for simulated ecosystems and a mesocosm experiment. By leveraging complexity, MVE is particularly effective for overcoming the limitations of short and noisy time series and should be highly relevant for many areas of science.

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

break

Setup

- Computer
- Organizing data
- Coding

Workflow

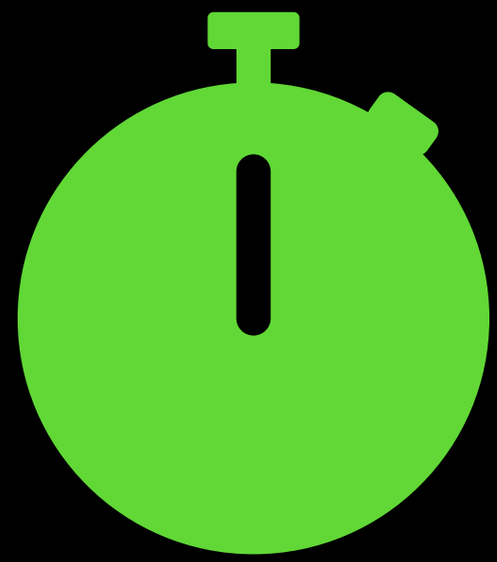
- Anticipate problems
- Get overview
- Clean data
- Find connections

break

Setup

- Computer
- Organizing data
- Coding

How does your setup (computer, code, libraries, data formats, …) reduce the efficiency of your work?



Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

break

Setup

- **Environment**
- Organizing data
- Coding

Workstation

- much RAM
- Fast-single core
- Don’t use workstation for computations!

Space

- Avoid isolated rooms.

Constant Learning

- Books, scientific literature; e.g.: note that NU has license for huge ebook collection on data science through Safari Online (https://www-safaribooksonline-com.turing.library.northwestern.edu/library/view/temporary-access/)
- Meetups
- www.data-science-nights.org

Workstation

- much RAM
- Fast-single core
- Don’t use workstation for computations!

Space

- Avoid isolated rooms.

Constant Learning

- Books, scientific literature; e.g.: note that NU has license for huge ebook collection on data science through Safari Online (https://www-safaribooksonline-com.turing.library.northwestern.edu/library/view/temporary-access/)
- Meetups
- www.data-science-nights.org

Workstation

- much RAM
- Fast-single core
- Don’t use workstation for computations!

Space

- Avoid isolated rooms.

Constant Learning

- Books, scientific literature; e.g.: note that NU has license for huge ebook collection on data science through Safari Online (https://www-safaribooksonline-com.turing.library.northwestern.edu/library/view/temporary-access/)
- Meetups
- www.data-science-nights.org

Workstation

- much RAM
- Fast-single core
- Don’t use workstation for computations!

Space

- Avoid isolated rooms.

Constant Learning

- Books, scientific literature; e.g.: note that NU has license for huge ebook collection on data science through Safari Online (https://www-safaribooksonline-com.turing.library.northwestern.edu/library/view/temporary-access/)
- Meetups
- www.data-science-nights.org

Workstation

- much RAM
- Fast-single core
- Don’t use workstation for computations!

Space

- Avoid isolated rooms.

Constant Learning

- Books, scientific literature; e.g.: note that NU has license for huge ebook collection on data science through Safari Online (https://www-safaribooksonline-com.turing.library.northwestern.edu/library/view/temporary-access/)
- Meetups
- www.data-science-nights.org

Workstation

- much RAM
- Fast-single core
- Don’t use workstation for computations!

Space

- Avoid isolated rooms.

Constant Learning

- Books, scientific literature; e.g.: note that NU has license for huge ebook collection on data science through Safari Online (https://www-safaribooksonline-com.turing.library.northwestern.edu/library/view/temporary-access/)
- Meetups
- www.data-science-nights.org

Workstation

- much RAM
- Fast-single core
- Don’t use workstation for computations!

Space

- Avoid isolated rooms.

Constant Learning

- Books, scientific literature; e.g.: note that NU has license for huge ebook collection on data science through Safari Online (https://www-safaribooksonline-com.turing.library.northwestern.edu/library/view/temporary-access/)
- Meetups
- www.data-science-nights.org

Workstation

- much RAM
- Fast-single core
- Don’t use workstation for computations!

Space

- Avoid isolated rooms.

Constant Learning

- Books, scientific literature; e.g.: note that NU has license for huge ebook collection on data science through Safari Online (https://www-safaribooksonline-com.turing.library.northwestern.edu/library/view/temporary-access/)
- Meetups
- www.data-science-nights.org

Workstation

- much RAM
- Fast-single core
- Don’t use workstation for computations!

Space

- Avoid isolated rooms.

Constant Learning

- Books, scientific literature; e.g.: note that NU has license for huge ebook collection on data science through Safari Online (https://www-safaribooksonline-com.turing.library.northwestern.edu/library/view/temporary-access/)
- Meetups
- www.data-science-nights.org

Please object. As some of this is scientific field specific.

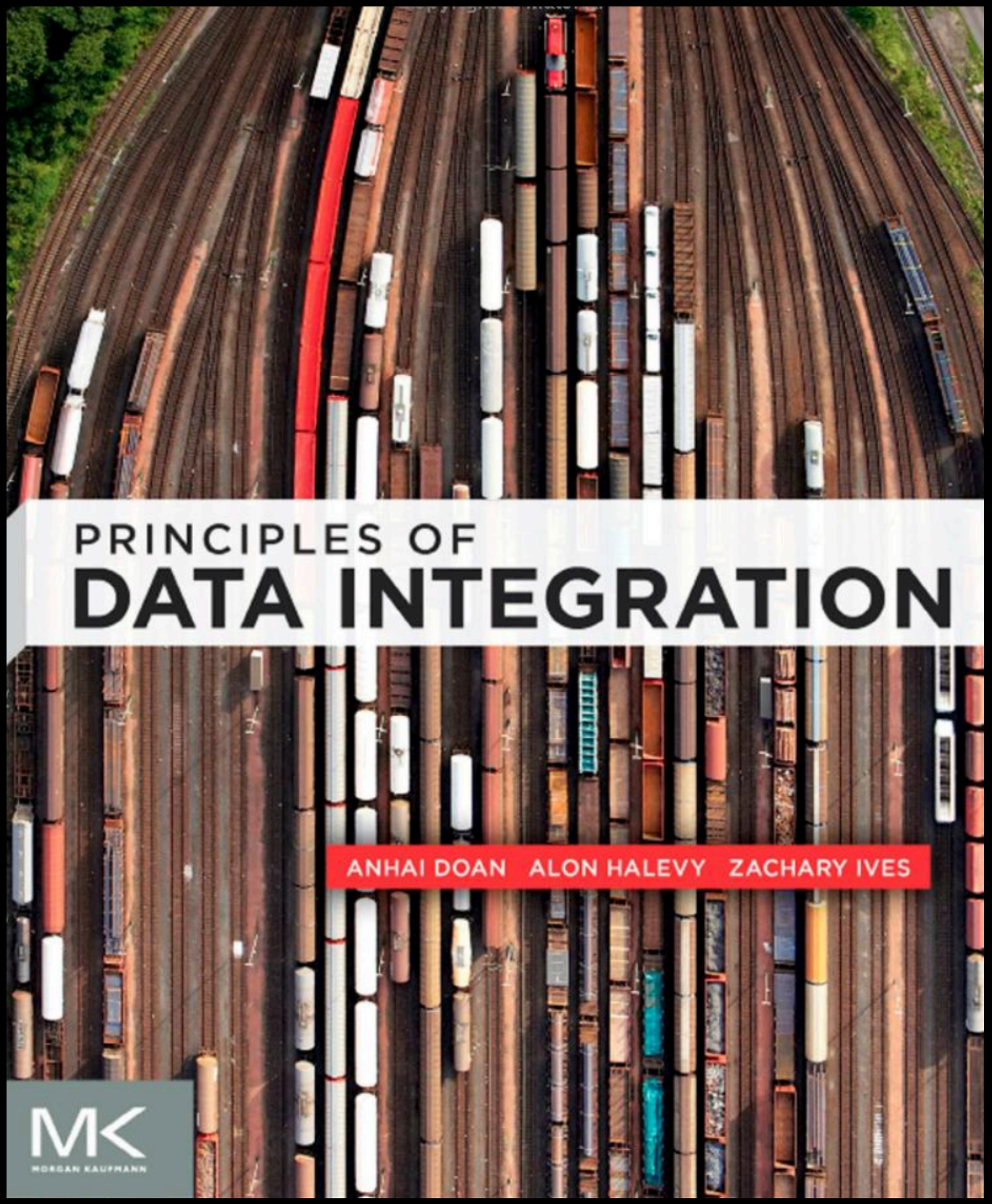
Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

break

Setup

- Computer
- **Organizing data**
- Coding



Data integration is an entire scientific field.

e.g.: dealing with ambiguity in matching of entries, or speeding processing up,…

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
https://doi.org/10.1080/00031305.2017.1375989

Taylor & Francis
Taylor & Francis Group

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
https://doi.org/10.1080/00031305.2017.1375989

Taylor & Francis
Taylor & Francis Group

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

Be consistent.

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
https://doi.org/10.1080/00031305.2017.1375989

Taylor & Francis
Taylor & Francis Group

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

Be consistent.

Just put one thing in a cell.

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
https://doi.org/10.1080/00031305.2017.1375989

Taylor & Francis
Taylor & Francis Group

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

Be consistent.

Just put one thing in a cell.

Create a data dictionary.

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
https://doi.org/10.1080/00031305.2017.1375989

Taylor & Francis
Taylor & Francis Group

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

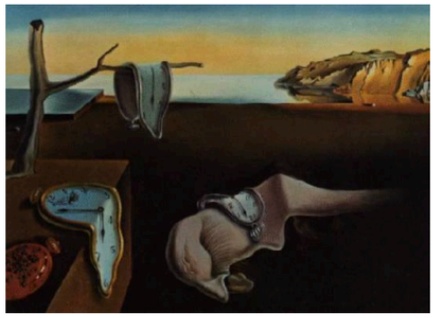
[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

Be consistent.

Just put one thing in a cell.

Create a data dictionary.

Do not use color as data.



Journal of Statistical Softwar

August 2014, Volume 59, Issue 10. *http://www.jstats*

popular in some academic fields

Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualize, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Portability and access.

Central location

Database

Folder with datasets

Domain-specific (e.g.: NIH's Synapse)

Local copies

Single reference!

Portability and access.

Central location

Database

Folder with datasets

Domain-specific (e.g.: NIH's Synapse)

Local copies

Single reference!

Portability and access.

Central location

Database

Folder with datasets

Domain-specific (e.g.: NIH's Synapse)

Local copies

Single reference!

Portability and access.

Central location

Database

Folder with datasets

Domain-specific (e.g.: NIH's Synapse)

Local copies

Single reference!

Portability and access.

Central location

Database

Folder with datasets

Domain-specific (e.g.: NIH's Synapse)

Local copies

Single reference!

many technical choices

```
In [ ]: df_multiple_choice.to_
        df_multiple_choice.to_clipboard
        df_multiple_choice.to_csv
        df_multiple_choice.to_dense
        df_multiple_choice.to_dict
        df_multiple_choice.to_excel
        df_multiple_choice.to_feather
        df_multiple_choice.to_gbq
        df_multiple_choice.to_hdf
        df_multiple_choice.to_html
        df_multiple_choice.to_json
```

```
In [ ]: df_multiple_choice.to_
        df_multiple_choice.to_latex
        df_multiple_choice.to_msgpack
        df_multiple_choice.to_panel
        df_multiple_choice.to_parquet
        df_multiple_choice.to_period
        df_multiple_choice.to_pickle
        df_multiple_choice.to_records
        df_multiple_choice.to_sparse
        df_multiple_choice.to_sql
        df_multiple_choice.to_stata
```

many technical choices

```
In [ ]: df_multiple_choice.to_
        df_multiple_choice.to_clipboard
        df_multiple_choice.to_csv
        df_multiple_choice.to_dense
        df_multiple_choice.to_dict
        df_multiple_choice.to_excel
        df_multiple_choice.to_feather
        df_multiple_choice.to_gbq
        df_multiple_choice.to_hdf
        df_multiple_choice.to_html
        df_multiple_choice.to_json
```

CSV
Spreadsheets

```
In [ ]: df_multiple_choice.to_
        df_multiple_choice.to_latex
        df_multiple_choice.to_msgpack
        df_multiple_choice.to_panel
        df_multiple_choice.to_parquet
        df_multiple_choice.to_period
        df_multiple_choice.to_pickle
        df_multiple_choice.to_records
        df_multiple_choice.to_sparse
        df_multiple_choice.to_sql
        df_multiple_choice.to_stata
```

many technical choices

```
In [ ]: df_multiple_choice.to_
        df_multiple_choice.to_clipboard
        df_multiple_choice.to_csv
        df_multiple_choice.to_dense
        df_multiple_choice.to_dict
        df_multiple_choice.to_excel
        df_multiple_choice.to_feather
        df_multiple_choice.to_gbq
        df_multiple_choice.to_hdf
        df_multiple_choice.to_html
        df_multiple_choice.to_json
```

CSV
Spreadsheets

HDF5
Selectively load Parts of data (avoid breaking cluster)

```
In [ ]: df_multiple_choice.to_
        df_multiple_choice.to_latex
        df_multiple_choice.to_msgpack
        df_multiple_choice.to_panel
        df_multiple_choice.to_parquet
        df_multiple_choice.to_period
        df_multiple_choice.to_pickle
        df_multiple_choice.to_records
        df_multiple_choice.to_sparse
        df_multiple_choice.to_sql
        df_multiple_choice.to_stata
```

many technical choices

```
In [ ]: df_multiple_choice.to_
        df_multiple_choice.to_clipboard
        df_multiple_choice.to_csv
        df_multiple_choice.to_dense
        df_multiple_choice.to_dict
        df_multiple_choice.to_excel
        df_multiple_choice.to_feather
        df_multiple_choice.to_gbq
        df_multiple_choice.to_hdf
        df_multiple_choice.to_html
        df_multiple_choice.to_json
```

CSV
Spreadsheets

HDF5
Selectively load
Parts of data
(avoid breaking
cluster)

```
In [ ]: df_multiple_choice.to_
        df_multiple_choice.to_latex
        df_multiple_choice.to_msgpack
        df_multiple_choice.to_panel
        df_multiple_choice.to_parquet
        df_multiple_choice.to_period
        df_multiple_choice.to_pickle
        df_multiple_choice.to_records
        df_multiple_choice.to_sparse
        df_multiple_choice.to_sql
        df_multiple_choice.to_stata
```

PARQUET
Very fast reading
and writing of
entire tables

many technical choices

```
In [ ]: df_multiple_choice.to_
        df_multiple_choice.to_clipboard
        df_multiple_choice.to_csv
        df_multiple_choice.to_dense
        df_multiple_choice.to_dict
        df_multiple_choice.to_excel
        df_multiple_choice.to_feather
        df_multiple_choice.to_gbq
        df_multiple_choice.to_hdf
        df_multiple_choice.to_html
        df_multiple_choice.to_json
```

CSV
Spreadsheets

HDF5
Selectively load Parts of data (avoid breaking cluster)

```
In [ ]: df_multiple_choice.to_
        df_multiple_choice.to_latex
        df_multiple_choice.to_msgpack
        df_multiple_choice.to_panel
        df_multiple_choice.to_parquet
        df_multiple_choice.to_period
        df_multiple_choice.to_pickle
        df_multiple_choice.to_records
        df_multiple_choice.to_sparse
        df_multiple_choice.to_sql
        df_multiple_choice.to_stata
```

PARQUET
Very fast reading and writing of entire tables

SQL
Elegant query language

In [172]:
```
example_data = df_multiple_choice[['Country', 'Age', 'GenderSelect']].dropna()
```

In [173]:
```
example_data.head()
```

Out[173]:

	Country	Age	GenderSelect
1	United States	30.0	Female
2	Canada	28.0	Male
3	United States	56.0	Male
4	Taiwan	38.0	Male
5	Brazil	46.0	Male

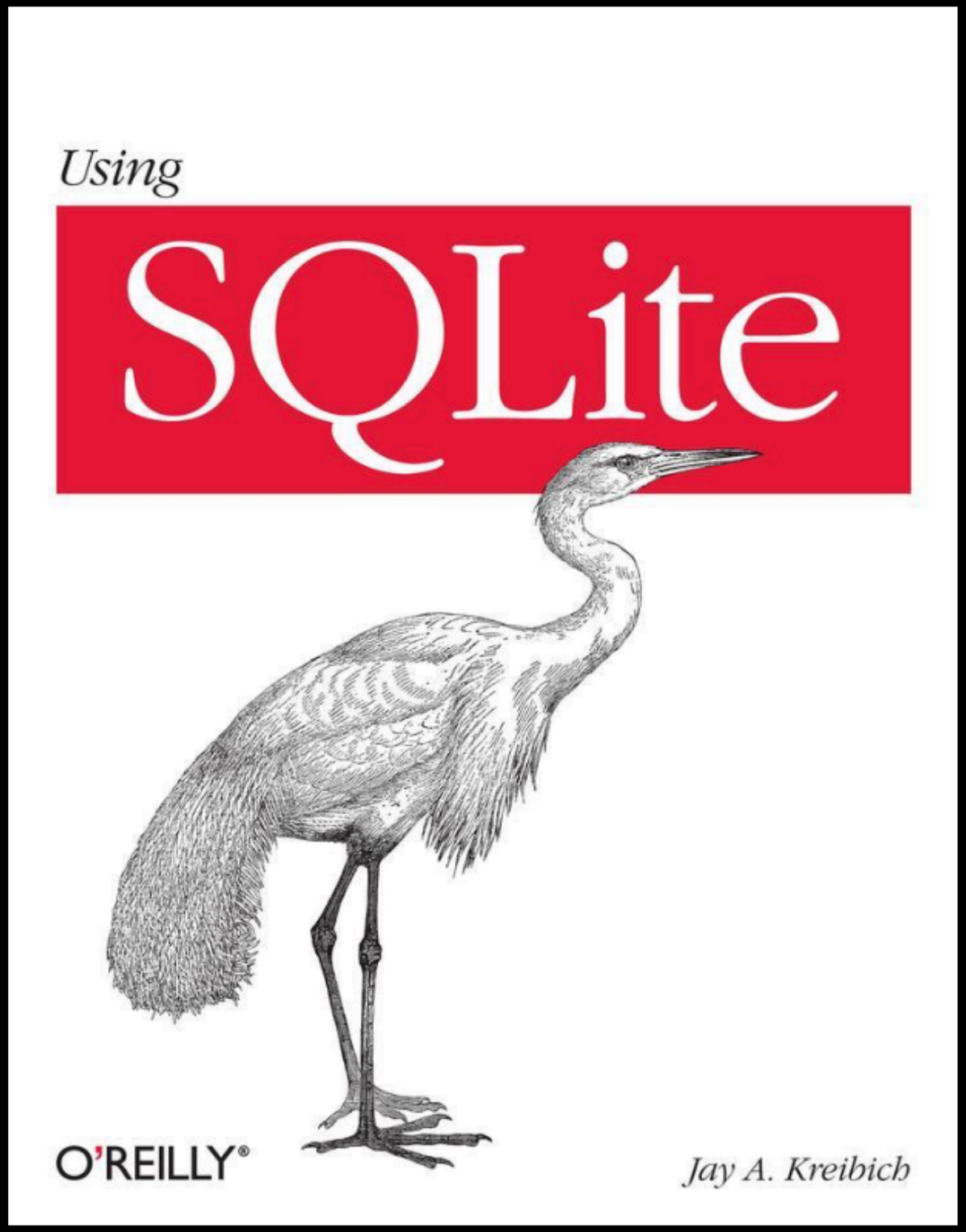
HDF5

In [174]:
```
example_data.to_hdf(
    './example.h5',
    format='table',
    key='searchable_data',
    mode='w',
    data_columns=True,
    )
```

In [175]:
```
pd.read_hdf('./example.h5', 'searchable_data', where=['Age=20'])
```

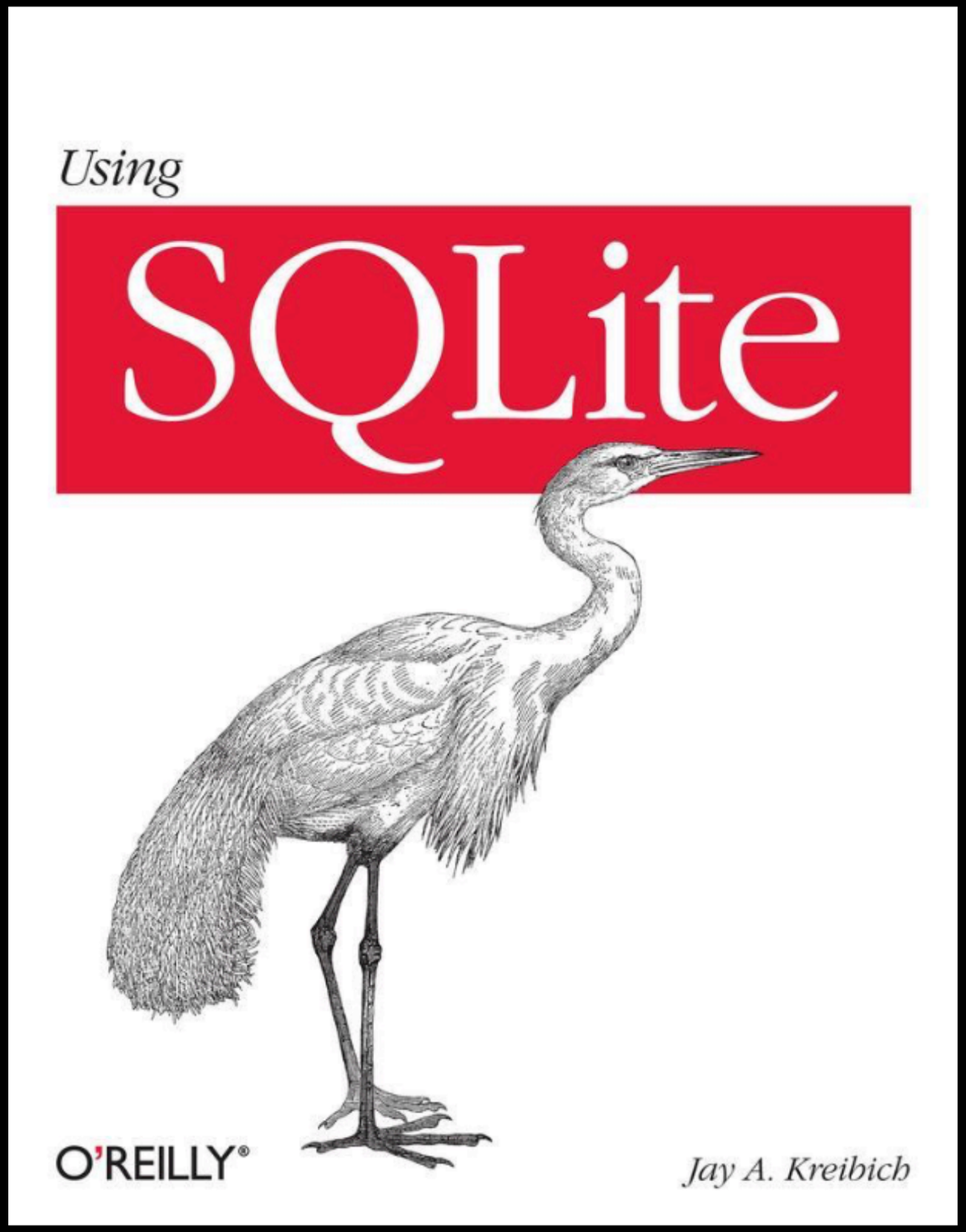
Out[175]:

	Country	Age	GenderSelect
10	Russia	20.0	Female
123	United States	20.0	Female
125	Turkey	20.0	Male
144	United States	20.0	Male
148	India	20.0	Male



Remember

You can have a server-less database.



Remember

You can have a server-less database.

NU provides tons of free ebooks.

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

break

Setup

- Computer
- Organizing data
- **Coding**

Save time!

Important to many.

Reduce time coding.

Specialist applications.

Reduce time running code.

Save time!

Important to many.

Reduce time coding.

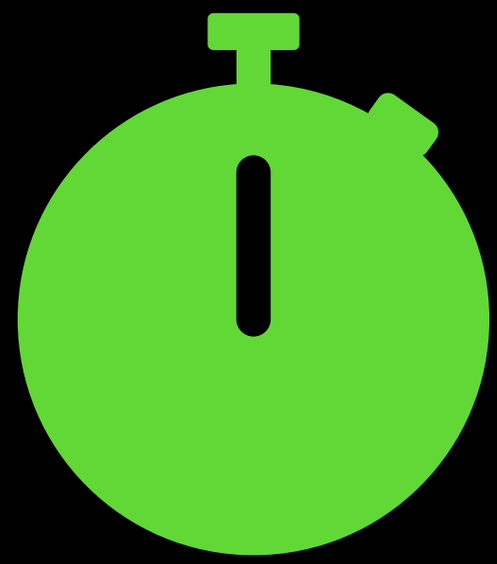
Important to even more

Reduce time others need to understand code.

Specialist applications.

Reduce time running code.

Which coding practices could safe time?



Watch your language!

PLOS ONE | PUBLISH ABOUT BROWSE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

A large-scale analysis of bioinformatics code on GitHub

Pamela H. Russell, Rachel L. Johnson, Shreyas Ananthan, Benjamin Harnke, Nichole E. Carlson

Published: October 31, 2018 • https://doi.org/10.1371/journal.pone.0205898

Article | Authors | Metrics | Comments | Media Coverage

Abstract

In recent years, the explosion of genomic data and bioinformatic tools has been accompanied by a growing conversation around reproducibility of results and usability of software. However, the actual state of the body of bioinformatics software remains largely unknown. The purpose of this paper is to investigate the state of source code in the bioinformatics community, specifically looking at relationships between code properties, development activity, developer communities, and software impact. To investigate these issues, we curated a list of 1,720 bioinformatics repositories on GitHub through their mention in peer-reviewed bioinformatics articles. Additionally, we included 23 high-profile repositories identified by their popularity in an online bioinformatics forum. We analyzed repository metadata, source code, development activity, and team dynamics using data made available publicly through the GitHub API, as well as article metadata. We found key relationships within our dataset, including: certain scientific topics are associated with more active code development and higher community interest in the repository; most of the code in the main dataset is written in dynamically typed languages, while most of

education.github.com

Back to GitHub.com GitHub Support Contact GitHub

GitHub Education

Students Teachers Schools Events **Get benefits**

Real-world tools, engaged students

GitHub Education helps students, teachers, and schools access the tools and events they need to shape the next generation of software development.

GitHub Student Developer Pack

The best developer tools, free for students

GitHub Campus Experts

Training to enrich the technology community at your school

GitHub Campus Program

GitHub for your whole school, with everything you need to make it great

GitHub Classroom

The GitHub workflow, scaled for the needs of students

GitHub Campus Advisors

Teacher training to master Git and GitHub

Track and share your code!

e.g.: GitHub, which offers free repositories for academics and research groups (education.github.com)

The art of naming variables

July 30th 2018

> There are only two hard things in Computer Science: cache invalidation and naming things.

https://hackernoon.com/the-art-of-naming-variables-52f44de00aad

The art of naming variables

July 30th 2018

> There are only two hard things in Computer Science: cache invalidation and naming things.

['apple', 'banana', 'cucumber']

https://hackernoon.com/the-art-of-naming-variables-52f44de00aad

The art of naming variables

July 30th 2018

> There are only two hard things in Computer Science: cache invalidation and naming things.

['apple', 'banana', 'cucumber']

fruits

https://hackernoon.com/the-art-of-naming-variables-52f44de00aad

The art of naming variables

July 30th 2018

> There are only two hard things in Computer Science: cache invalidation and naming things.

['apple', 'banana', 'cucumber'] **fruits**

function any(x.isin(fruits))

https://hackernoon.com/the-art-of-naming-variables-52f44de00aad

The art of naming variables

July 30th 2018

> There are only two hard things in Computer Science: cache invalidation and naming things.

['apple', 'banana', 'cucumber'] fruits

function any(x.isin(fruits)) is_fruit

https://hackernoon.com/the-art-of-naming-variables-52f44de00aad

Principles [edit]

Principles are listed as follows:

> Beautiful is better than ugly.
> Explicit is better than implicit.
> Simple is better than complex.
> Complex is better than complicated.
> Flat is better than nested.
> Sparse is better than dense.
> Readability counts.
> Special cases aren't special enough to break the rules.
> Although practicality beats purity.
> Errors should never pass silently.
> Unless explicitly silenced.
> In the face of ambiguity, refuse the temptation to guess.
> There should be one—and preferably only one—obvious way to do it.
> Although that way may not be obvious at first unless you're Dutch.
> Now is better than never.
> Although never is often better than *right* now.[n 1]
> If the implementation is hard to explain, it's a bad idea.
> If the implementation is easy to explain, it may be a good idea.
> Namespaces are one honking great idea—let's do more of those!

style conventions
e.g. Zen of Python (and subsequently PEP8)

Did we miss some useful?

the last lesson:

Plan for changing data.

Saves time!

allows expansion of project

allows to only spend time on improving access when needed

the last lesson:

Plan for changing data.

Saves time!

allows expansion of project

allows to only spend time on improving access when needed

the last lesson:

Plan for changing data.

Saves time!

allows expansion of project

allows to only spend time on improving access when needed

the last lesson:

Plan for changing data.

Saves time!

allows expansion of project

allows to only spend time on improving access when needed

Don’t access data directly!

Don’t access data directly!

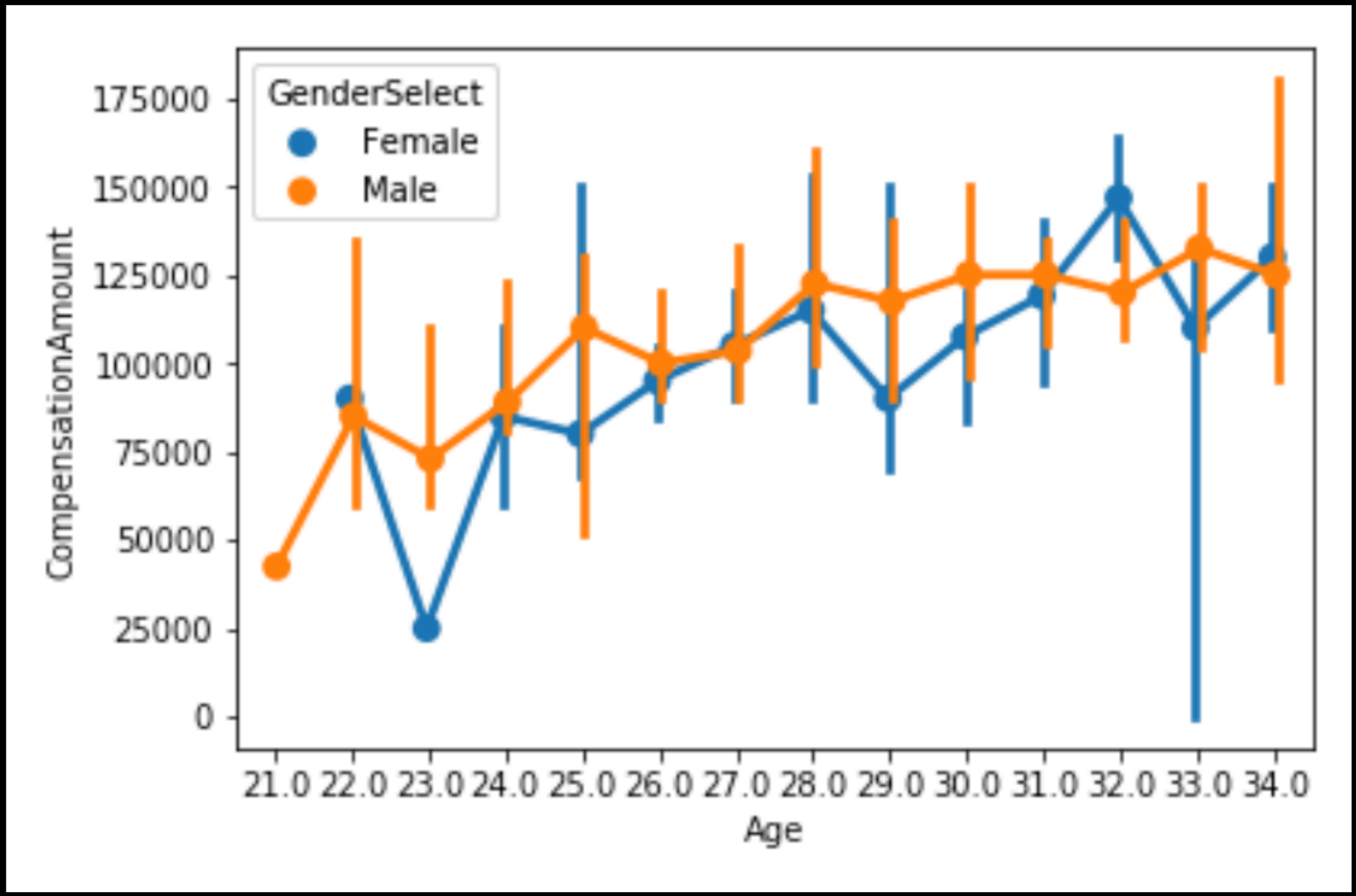
Notebook → **Adaptor Function** → **Data resource**

Don’t access data directly!

Notebook → **Adaptor** → **Data resource**

Function

Can change data resource
without worrying in ongoing research.


GenderSelect
Female
Male
175000
150000
125000
100000
75000
50000
25000
0
CompensationAmount
21.0 22.0 23.0 24.0 25.0 26.0 27.0 28.0 29.0 30.0 31.0 32.0 33.0 34.0
Age

Direct loading is acceptable for previewing new datasets.

In [1]:
```
%matplotlib inline
```

In [2]:
```
cd '/Users/tstoeger/Dropbox/Work/kaggle_survey'
```

```
/Users/tstoeger/Dropbox/Work/kaggle_survey
```

In [3]:
```
import pandas as pd
import seaborn as sns
```

In [4]:
```
df_multiple_choice = pd.read_csv(
    './multipleChoiceResponses.csv',
    encoding='latin-1', # files have some special characters
    low_memory=False)   # some columns have mixed type
```

In [5]:
```
df_multiple_choice.shape   # get idea of number of datasets
```

Out[5]:
```
(16716, 228)
```

… but move your loading into adaptor functions in the long run.

In [190]:

```
def load_country_gender_at_given_age(age_of_interest):
    """
    Loads the country and gender of survey takes, which
    are of a given age

    Input:
        age_of_interest  int

    Output:
        dataframe

    """

    df = pd.read_csv(
        './multipleChoiceResponses.csv',
        encoding='latin-1', # files have some special characters
        low_memory=False)   # some columns have mixed type
    df = df=df[['Country', 'Age', 'GenderSelect']].dropna()
    df = df[df['Age']==age_of_interest]

    return df
```

Define function
In human digestible
Manner

… but move your loading into adaptor functions in the long run.

In [190]:
```
def load_country_gender_at_given_age(age_of_interest):
    """
    Loads the country and gender of survey takes, which
    are of a given age

    Input:
        age_of_interest  int

    Output:
        dataframe

    """

    df = pd.read_csv(
        './multipleChoiceResponses.csv',
        encoding='latin-1', # files have some special characters
        low_memory=False)   # some columns have mixed type
    df = df=df[['Country', 'Age', 'GenderSelect']].dropna()
    df = df[df['Age']==age_of_interest]

    return df
```

In [191]:
```
d = load_country_gender_at_given_age(20)
```

In [192]:
```
d.head()
```

Out[192]:

	Country	Age	GenderSelect
10	Russia	20.0	Female
123	United States	20.0	Female

In [193]:

```
def load_country_gender_at_given_age(age_of_interest):
    """
    Loads the country and gender of survey takes, which
    are of a given age

    Input:
        age_of_interest  int

    Output:
        dataframe

    """

    df = pd.read_hdf('./example.h5', 'searchable_data', where=['Age={}'.format(
        int(age_of_interest))])

    return df
```

Changed to HDF5

In [193]:
```
def load_country_gender_at_given_age(age_of_interest):
    """
    Loads the country and gender of survey takes, which
    are of a given age

    Input:
        age_of_interest  int

    Output:
        dataframe

    """

    df = pd.read_hdf('./example.h5', 'searchable_data', where=['Age={}'.format(
        int(age_of_interest))])

    return df
```

In [194]:
```
d = load_country_gender_at_given_age(20)
```

In [195]:
```
d.head()
```

Out[195]:

	Country	Age	GenderSelect
10	Russia	20.0	Female
123	United States	20.0	Female
125	Turkey	20.0	Male
144	United States	20.0	Male
148	India	20.0	Male

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

Setup

- Computer
- Organizing data
- Coding

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

Setup

- Computer
- Organizing data
- Coding

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

Setup

- Computer
- Organizing data
- Coding

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

Setup

- Computer
- Organizing data
- Coding

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

Setup

- Computer
- Organizing data
- Coding

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

Setup

- Computer
- Organizing data
- Coding

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

Setup

- Computer
- Organizing data
- Coding

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

Setup

- Computer
- Organizing data
- Coding

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

Setup

- Computer
- Organizing data
- Coding

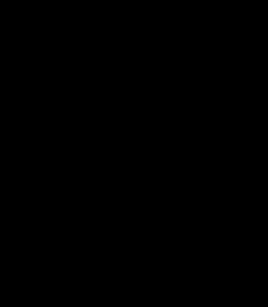
2 minutes:
Which encountered strategy could facilitate your work?

Workflow

- Anticipate problems
- Get overview
- Clean data
- Find connections

Setup

- Computer
- Organizing data
- Coding



2 minutes:
Which encountered strategy could facilitate your work?

Questions / Comments?

thank you