

datapilot

utilizing quality and usage information
for subset selection during visual data preparation



arpit **narechania**



fan **du**



atanu **sinha**



ryan **rossi**



jane **hoffswell**



shunan **guo**



eunyee **koh**



shamkant **navathe**

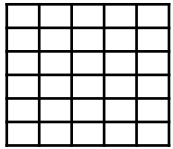


alex **endert**

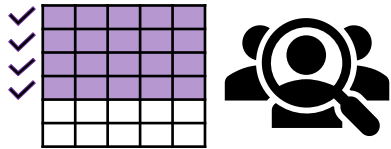


subset selection

what is it, who does it, and why?

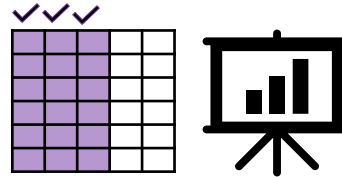


Consider a large tabular dataset about people, ...



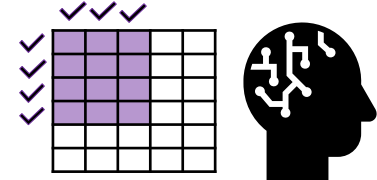
digital marketers

select specific people for marketing campaigns.



business analysts

select key attributes (KPIs) to show in dashboards, reports.

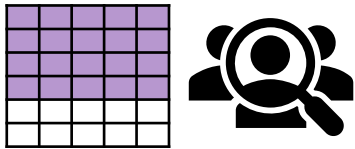


data scientists

select certain features and records to train ML models.

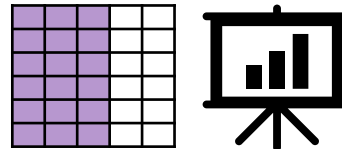
challenges with subset selection

why should we be concerned about this?



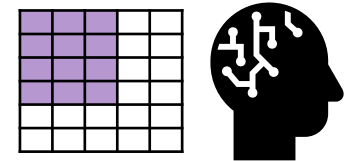
digital marketers

select specific people for marketing campaigns.



business analysts

select key attributes (KPIs) to show in dashboards, reports.



data scientists

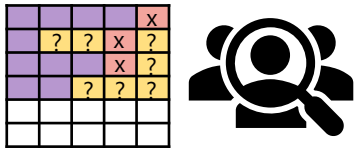
select certain features and records to train ML models.

challenges with subset selection

why should we be concerned about this?

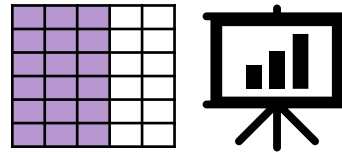
X incorrect?

? missing?



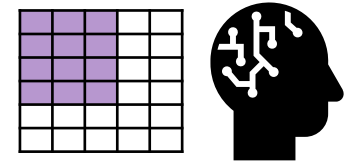
digital marketers

select specific people for marketing campaigns.



business analysts

select key attributes (KPIs) to show in dashboards, reports.



data scientists

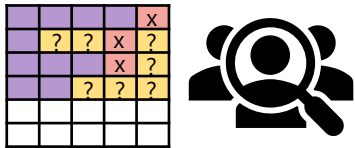
select certain features and records to train ML models.

challenges with subset selection

why should we be concerned about this?

 **incorrect?**

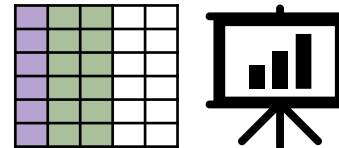
 **missing?**



digital marketers

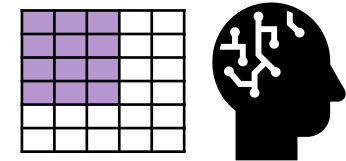
select specific people for marketing campaigns.

 **key, important attributes?**



business analysts

select key attributes (KPIs) to show in dashboards, reports.



data scientists

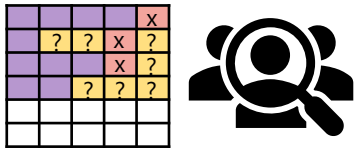
select certain features and records to train ML models.

challenges with subset selection

why should we be concerned about this?

X incorrect?

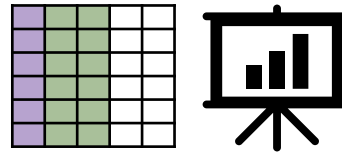
? missing?



digital marketers

select specific people for marketing campaigns.

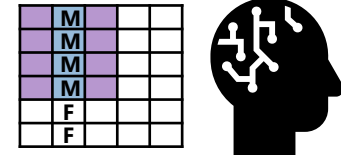
■ key, important attributes?



business analysts

select key attributes (KPIs) to show in dashboards, reports.

Male, **F**emale **skewed?**



data scientists

select certain features and records to train ML models.

design study with 14 data workers

data engineers, data analysts, data scientists, digital marketers (and a user interface engineer)

*“what **challenges** do you face during data preparation (subset selection)?”*

data quality is important

design study outcome #1

“is the data complete? correct? unbiased? ... showing such insights may instil feelings of curiosity and care in the user.”

“provision a visual report-card but for datasets can enable [organizational] decision-makers to enforce data policies.”

data quality

which metrics do we support?

missing (?)

				?
	?	?	?	?
			?	?
			?	?

completeness

frequency of non-missing
values in the data

incorrect (X)

	X			
		X		
			X	
	X	X	X	
		X		

correctness

frequency of correct
values in the data

skewed (M_{ale}, F_{emale})

	M			
	M			
	M			
	M			
	F	F		

objectivity

extent that values conform to
a target distribution

correctness, objectivity criteria are configurable by the user/org.

modeled for all attributes as well as records*.

scored from 0 – 100.

overall score = average (completeness, correctness, objectivity).

*objectivity is not computed for records

data usage can also be important

design study outcome #2

***"collaboration is tedious,** interactive self-service tools are needed."*

*"there are several low-quality datasets just lying around [in the data lake]. **knowing when, where, and by whom the datasets were last used** can help [with housekeeping]."*

data usage

which metrics do we support?

for a visual data exploration and dashboard building application,



in-subsets

% of users that select the attribute, record in their subset



in-filters

% of users that applied a filter on the attribute



in-visualizations

% of users that visualized the attribute

modeled for all attributes as well as records*.

scored from 0 – 100.

overall score = maximum (in-subsets, in-filters, in-visualizations) ^.

*by definition, record-level usage score is the same for all metrics.

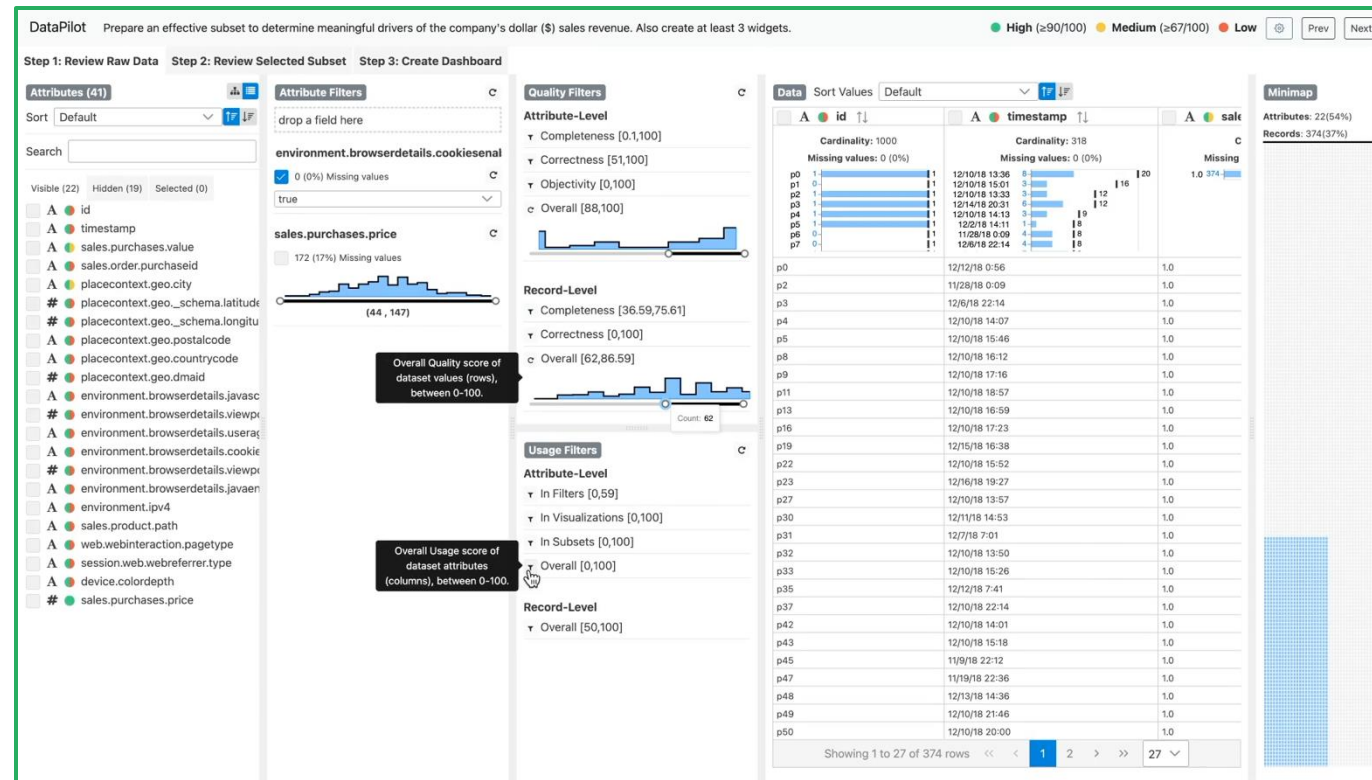
^ because the same attribute may not be used in-subsets, in-filters, and in-visualizations

datapilot

utilizing quality and usage information for subset selection during visual data preparation

data quality
missing values? errors?

data usage
which applications?
by whom?



navigate
large, unfamiliar datasets
select
relevant subsets
create
a visualization dashboard

Step 1: Review Raw Data Step 2: Review Selected Subset Step 3: Create Dashboard

Attributes (41)

Search

Visible (41) Hidden (0) Selected (0)

- A id
- A application.crashes.value
- A application.launches.value
- A timestamp
- A sales.productlistviews.value
- A sales.checkouts.value
- A sales.purchases.value
- A sales.productlistadds.value
- A sales.order.purchaseid
- A sales.productlistopens.value
- A sales.productviews.value
- A web.webinteraction.linkclicks.value
- A placecontext.geo.city
- # placecontext.geo._schema.latitude
- # placecontext.geo._schema.longitude
- A placecontext.geo.postalcode
- A placecontext.geo.stateprovince
- A placecontext.geo.countrycode
- # placecontext.geo.dmaid
- A marketing.trackingcode
- A environment.browserdetails.javascript
- # environment.browserdetails.viewportwidth
- A environment.browserdetails.useragent
- A environment.browserdetails.cookiesenabled
- # environment.browserdetails.viewportwidth
- A environment.browserdetails.javaenabled
- A environment.operatingsystem
- A environment.operatingsystemversion
- A environment.ipv4
- A sales.product.color
- A sales.product.size

Attribute Filters

drop a field here

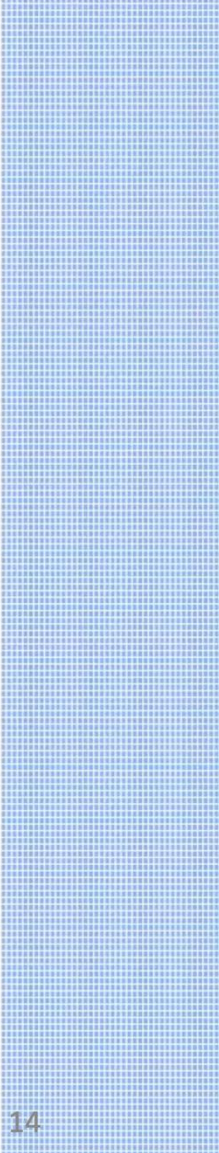
Data

A id ↑↓	A application.crashes.value ↑↓	A application.launches.value ↑↓	A timestamp
Cardinality: 1000	Cardinality: 1	Cardinality: 1	Cardinality: 1
p0 1	1.0 4	1.0 8	12/10/18 13:36 20
p1 1			12/10/18 15:01 16
p2 1			12/10/18 13:33 12
p3 1			12/14/18 20:31 12
p4 1			12/10/18 14:13 9
p5 1			12/2/18 14:11 8
p6 1			11/28/18 0:09 8
p7 1			12/6/18 22:14 8
p0			12/12/18 0:56
p1			12/2/18 14:11
p2			11/28/18 0:09
p3			12/6/18 22:14
p4			12/10/18 14:07
p5			12/10/18 15:46
p6			12/2/18 14:26
p7			12/16/18 18:25
p8			12/10/18 16:12
p9			12/10/18 17:16
p10			12/10/18 13:33
p11			12/10/18 18:57
p12			12/16/18 18:25
p13			12/10/18 16:59
p14			12/10/18 17:34
p15			12/16/18 19:32
p16			12/10/18 17:23
p17			11/24/18 0:48
p18			12/5/18 17:31
p19			12/15/18 16:38
p20			12/10/18 13:33
p21			12/10/18 14:43
p22			12/10/18 15:52
p23			12/16/18 19:27
p24			12/16/18 19:05
p25			12/14/18 20:31
p26			12/12/18 4:25

Minimap

Attributes: 41(100%)

Records: 1000(100%)



E-Commerce Transactions

“Select an effective subset that determines meaningful drivers of the company’s sales revenue.”

Step 1: Review Raw Data

Step 2: Review Selected Subset

Step 3: Create Dashboard

Attributes (41)

Search

Visible (41)

Hidden (0)

Selected (0)

- ☐ A id
- ☐ A application.crashes.value
- ☐ A application.launches.value
- ☐ A timestamp
- ☐ A sales.productlistviews.value
- ☐ A sales.checkouts.value
- ☐ A sales.purchases.value
- ☐ A sales.productlistadds.value
- ☐ A sales.order.purchaseid
- ☐ A sales.productlistopens.value
- ☐ A sales.productviews.value
- ☐ A web.webinteraction.linkclicks.value
- ☐ A placecontext.geo.city
- ☐ # placecontext.geo._schema.latitude
- ☐ # placecontext.geo._schema.longitude
- ☐ A placecontext.geo.postalcode
- ☐ A placecontext.geo.stateprovince
- ☐ A placecontext.geo.countrycode
- ☐ # placecontext.geo.dmaid
- ☐ A marketing.trackingcode
- ☐ A environment.browserdetails.javascript
- ☐ # environment.browserdetails.viewportv
- ☐ A environment.browserdetails.useragen
- ☐ A environment.browserdetails.cookieser
- ☐ # environment.browserdetails.viewportf
- ☐ A environment.browserdetails.javaenabl
- ☐ A environment.operatingsystem
- ☐ A environment.operatingsystemversion
- ☐ A environment.ipv4
- ☐ A sales.product.color
- ☐ A sales.product.size

Attribute Filters

C

drop a field here

Data

☐ A id ↑↓	☐ A application.crashes.value ↑↓	☐ A application.launches.value ↑↓	☐ A timestamp
Cardinality: 1000 p011 p11 p21 p31 p41 p51 p61 p71	Cardinality: 1 1.04 4	Cardinality: 1 1.08 8	Cardine 12/10/18 13:3620 12/10/18 15:0116 12/10/18 13:3312 12/14/18 20:3112 12/10/18 14:139 12/2/18 14:118 11/28/18 0:098 12/6/18 22:148
p0			12/12/18 0:56
p1			12/2/18 14:11
p2			11/28/18 0:09
p3			12/6/18 22:14
p4			12/10/18 14:07
p5			12/10/18 15:46
p6			12/2/18 14:26
p7			12/16/18 18:25
p8			12/10/18 16:12
p9			12/10/18 17:16
p10			12/10/18 13:33
p11			12/10/18 18:57
p12			12/16/18 18:25
p13			12/10/18 16:59
p14			12/10/18 17:34
p15			12/16/18 19:32
p16			12/10/18 17:23
p17			11/24/18 0:48
p18			12/5/18 17:31
p19			12/15/18 16:38
p20			12/10/18 13:33
p21			12/10/18 14:43
p22			12/10/18 15:52
p23			12/16/18 19:27
p24			12/16/18 19:05
p25			12/14/18 20:31
p26			12/12/18 4:25

Showing 1 to 27 of 1000 rows

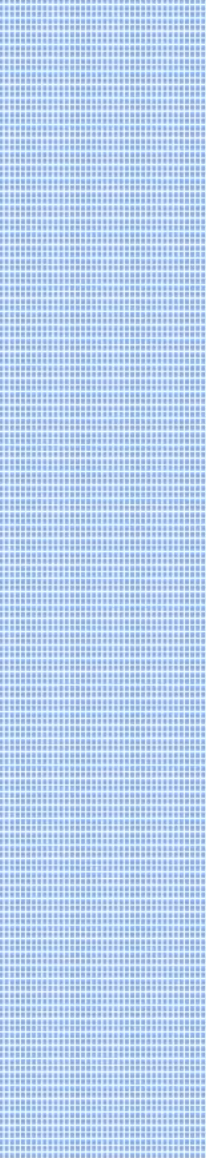
<<<<12>>>>

27

Minimap

Attributes: 41(100%)

Records: 1000(100%)



- Attributes (5)
- ☒

 A timestamp
- ☒

 A placecontext.geo.countrycode
- ☒

 A sales.product.name
- ☒

 A session.search.searchengine
- ☒

 # sales.purchases.price

☒ A timestamp ↑↓	☒ A placecontext.geo.countrycode ↑↓	☒ A sales.product.name ↑↓	☒ A session.search.searchengine ↑↓
Cardinality: 318 12/10/18 13:3620 12/10/18 15:0116 12/10/18 13:3312 12/14/18 20:3112 12/10/18 14:139 12/2/18 14:118 11/28/18 0:098 12/6/18 22:148 -	Cardinality: 21 US624 AA137 IN51 IE44 DE21 SE19 KR15 ES13 -	Cardinality: 98 harmony lumaflex™ strength b...24 mach street sweatshirt - (blue / s)16 sprite foam yoga brick16 montana wind jacket (black / s)12 sprite yoga straps12 ajax full-zip sweatshirt (blue / s)8 cronus yoga pant - (black / 32)...8 pursuit lumaflex™ tone band8 -	Cardinality: 1 Microsoft Bing1 -
12/12/18 0:56	US	ajax full-zip sweatshirt (blue / s)	
12/2/18 14:11	IN		
11/28/18 0:09	US		
12/6/18 22:14	IE		
12/10/18 14:07	US	cronus yoga pant - (black / 32) livingston all-purpose tight - (black / 32) affirm water bottle cruise dual analog watch	
12/10/18 15:46	US	montana wind jacket (black / s)	
12/2/18 14:26	IN		
12/16/18 18:25	US		
12/10/18 16:12	US	portia capri - (green / 28) arcadio gym short - (red / 32) compete track tote	
12/10/18 17:16	CA	daria bikram pant (white / 28) juno jacket - (green / xs) jade yoga jacket (green / s)	
12/10/18 13:33	US	sprite yoga straps sprite foam yoga brick sprite foam roller affirm water bottle	
12/10/18 18:57	US	orestes yoga pant - (black / 32)	
12/16/18 18:25	US		
12/10/18 16:59	DE	montana wind jacket (black / l)	
12/10/18 17:34	US		
12/16/18 19:32	US		
12/10/18 17:23	US	sprite yoga companion kit daphne full-zip hoodie - (purple / xs) juno jacket - (purple / xs)	
11/24/18 0:48	US		
12/5/18 17:31	US		
12/15/18 16:38	US	helena hooded fleece - (gray / s) mars heattech™ pullover - (red / s)	
12/10/18 13:33	IE	pursuit lumaflex™ tone band	
12/10/18 14:43	US		
12/10/18 15:52	AA	deirdre relaxed-fit capri (gray / 28) sprite foam yoga brick	
12/16/18 19:27	US	mach street sweatshirt - (blue / s)	
12/16/18 19:05	US		
12/14/18 20:31	AA		
12/12/18 4:25	AA	summit watch	

Showing 1 to 27 of 1000 rows

1

2

>

>>

27

Attributes (4)

☒ A ● timestamp

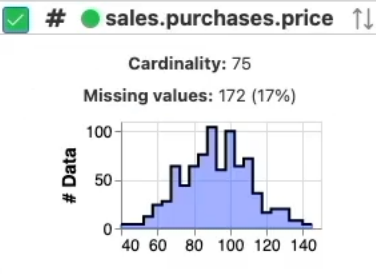
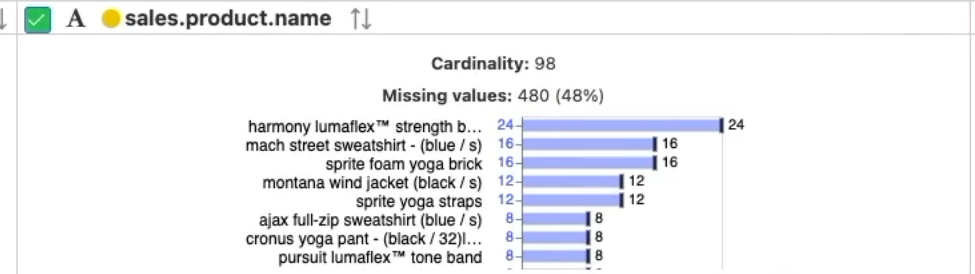
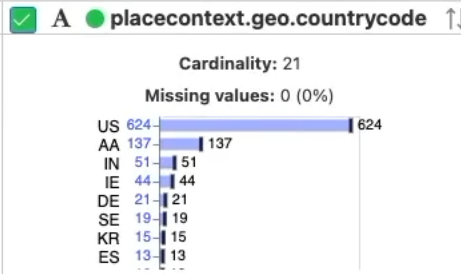
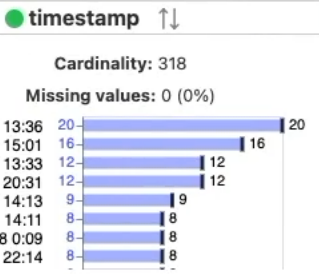
☒ A ● placecontext.geo.countrycode

☒ A ● sales.product.name

☒ # ● sales.purchases.price



Data



0:56	US	ajax full-zip sweatshirt (blue / s)	84
4:11	IN		99
0:09	US		70
:2:14	IE		64
14:07	US	cronus yoga pant - (black / 32) livingston all-purpose tight - (black / 32) affirm water bottle cruise dual analog watch	91
15:46	US	montana wind jacket (black / s)	74
4:26	IN		93
18:25	US		127
16:12	US	portia capri - (green / 28) arcadio gym short - (red / 32) compete track tote	93
17:16	CA	daria bikram pant (white / 28) juno jacket - (green / xs) jade yoga jacket (green / s)	83
13:33	US	sprite yoga straps sprite foam yoga brick sprite foam roller affirm water bottle	97
18:57	US	orestes yoga pant - (black / 32)	106
18:25	US		116
16:59	DE	montana wind jacket (black / l)	133
17:34	US		126
19:32	US		58
17:23	US	sprite yoga companion kit daphne full-zip hoodie - (purple / xs) juno jacket - (purple / xs)	78
0:48	US		77
7:31	US		92
16:38	US	helena hooded fleece - (gray / s) mars heattech™ pullover - (red / s)	110
13:33	IE	pursuit lumaflex™ tone band	99
14:43	US		112
15:52	AA	deirdre relaxed-fit capri (gray / 28) sprite foam yoga brick	93
19:27	US	mach street sweatshirt - (blue / s)	104
19:05	US		107
20:31	AA		114
4:25	AA	summit watch	100

Step 1: Review Raw Data Step 2: Review Selected Subset Step 3: Create Dashboard

Attributes (41)

Sort Default

Search

Visible (22) Hidden (19) Selected (4)

☐ A ● id

☒ A ● timestamp

☐ A ● sales.purchases.value

☐ A ● sales.order.purchaseid

☐ A ● placecontext.geo.city

☐ # ● placecontext.geo._schema.latitude

☐ # ● placecontext.geo._schema.longitude

☐ A ● placecontext.geo.postalcode

☒ A ● placecontext.geo.countrycode

☐ # ● placecontext.geo.dmaid

☐ A ● environment.browserdetails.javasc

☐ # ● environment.browserdetails.viewpo

☐ A ● environment.browserdetails.userag

☐ A ● environment.browserdetails.cookie

☐ # ● environment.browserdetails.viewpo

☐ A ● environment.browserdetails.javaen

☐ A ● environment.ipv4

☐ A ● sales.product.path

☐ A ● web.webinteraction.pagetype

☐ A ● session.web.webreferrer.type

☐ A ● device.colordepth

☒ # ● sales.purchases.price

Attribute Filters

drop a field here

Overall Quality score of dataset attributes (columns), between 0-100.

Quality Filters

Attribute-Level

Completeness [0.1,100]

Correctness [51,100]

Objectivity [0,100]

Overall [90,100]

Record-Level

Completeness [36.59,75.61]

Correctness [100,100]

Overall [19.51,86.59]

Data Sort Values Default

timestamp

Cardinality: 318

Missing values: 0 (0%)

sales.purchases.value

Cardinality: 1

Missing values: 172 (17.2%)

11/28/18 0:09	1.0
12/6/18 22:14	1.0
12/10/18 14:07	1.0
12/10/18 15:46	1.0
12/10/18 16:12	1.0
12/10/18 17:16	1.0
12/10/18 17:23	1.0
12/15/18 16:38	1.0
12/11/18 14:53	1.0
12/10/18 15:26	1.0
12/12/18 7:41	1.0
11/9/18 22:12	1.0
11/19/18 22:36	1.0
12/13/18 14:36	1.0
12/10/18 21:46	1.0
12/10/18 20:00	1.0
12/10/18 17:32	1.0
12/14/18 22:22	1.0
12/10/18 14:57	1.0
12/13/18 0:28	1.0
12/14/18 9:30	1.0
12/10/18 22:04	1.0
12/10/18 14:07	1.0
12/10/18 14:19	1.0
12/11/18 14:53	1.0
12/10/18 13:39	1.0
12/10/18 17:32	1.0

Showing 1 to 27 of 181 rows 1 27

Minimap

Attributes: 22(54%)

Records: 181(18%)

Step 1: Review Raw Data

Step 2: Review Selected Subset

Step 3: Create Dashboard

Attributes (6)

- ☒ A  timestamp
- ☒ A  sales.purchases.value
- ☒ A  placecontext.geo.city
- ☒ A  placecontext.geo.countrycode
- ☒ A  sales.product.name
- ☒ #  sales.purchases.price



Encodings

Swap XY ✕

Chart

X Axis

✕

drop a field here

Y Axis

✕

drop a field here

Visualization

Enter Name



Saved Visualizations

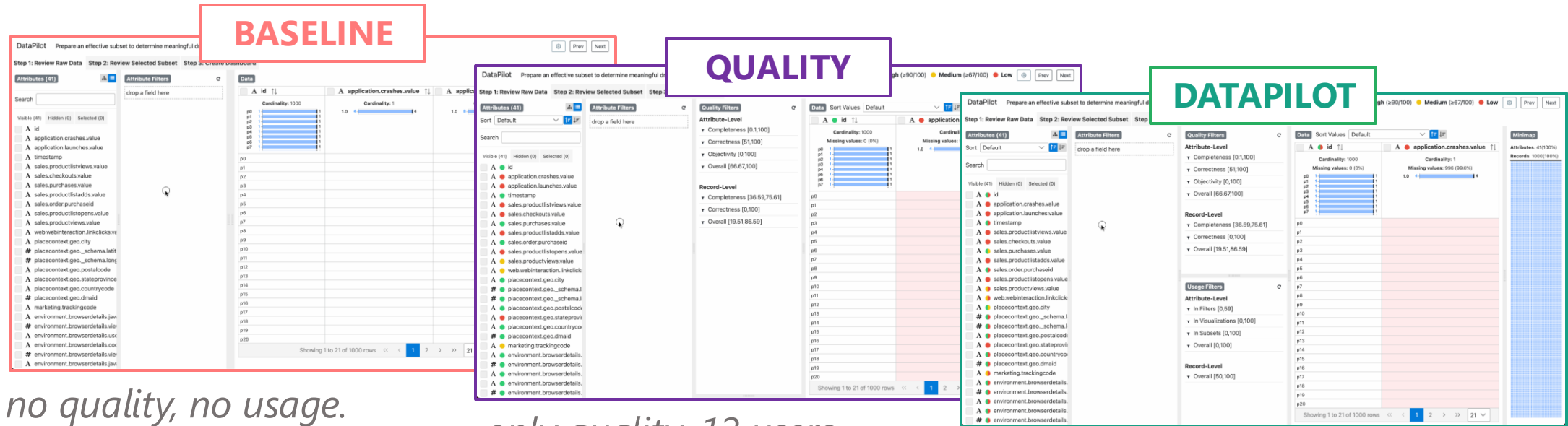


evaluation

between-subjects study with 36 users.

task

prepare an effective subset to determine meaningful drivers of the company's sales revenue. create three visualizations in support.



*no quality, no usage.
12 users.*

only quality. 12 users.

both quality and usage. 12 users.

hypotheses

evaluation

*no quality, no usage.
12 users.*

*only quality.
12 users.*

*both quality and usage.
12 users.*

subset size
smaller is desirable

BASELINE

>

QUALITY

>

DATAPILOT

subset effectiveness*
greater is desirable

BASELINE

<

QUALITY

<

DATAPILOT

success & confidence
greater is desirable

BASELINE

<

QUALITY

<

DATAPILOT



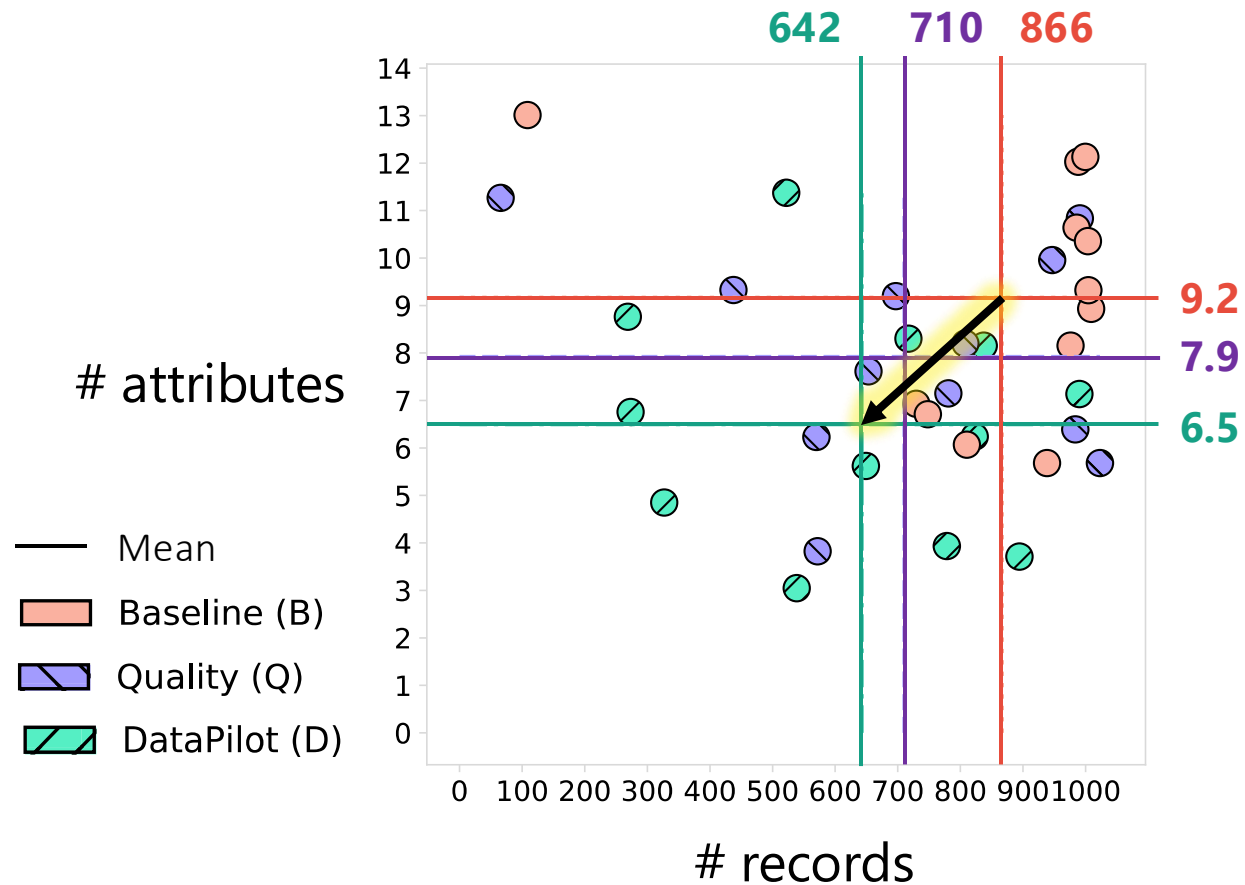
our study analyses validated these.

* [higher, lower] % of attributes and records with [high, low] overall quality & usage scores

> is greater than; < is less than

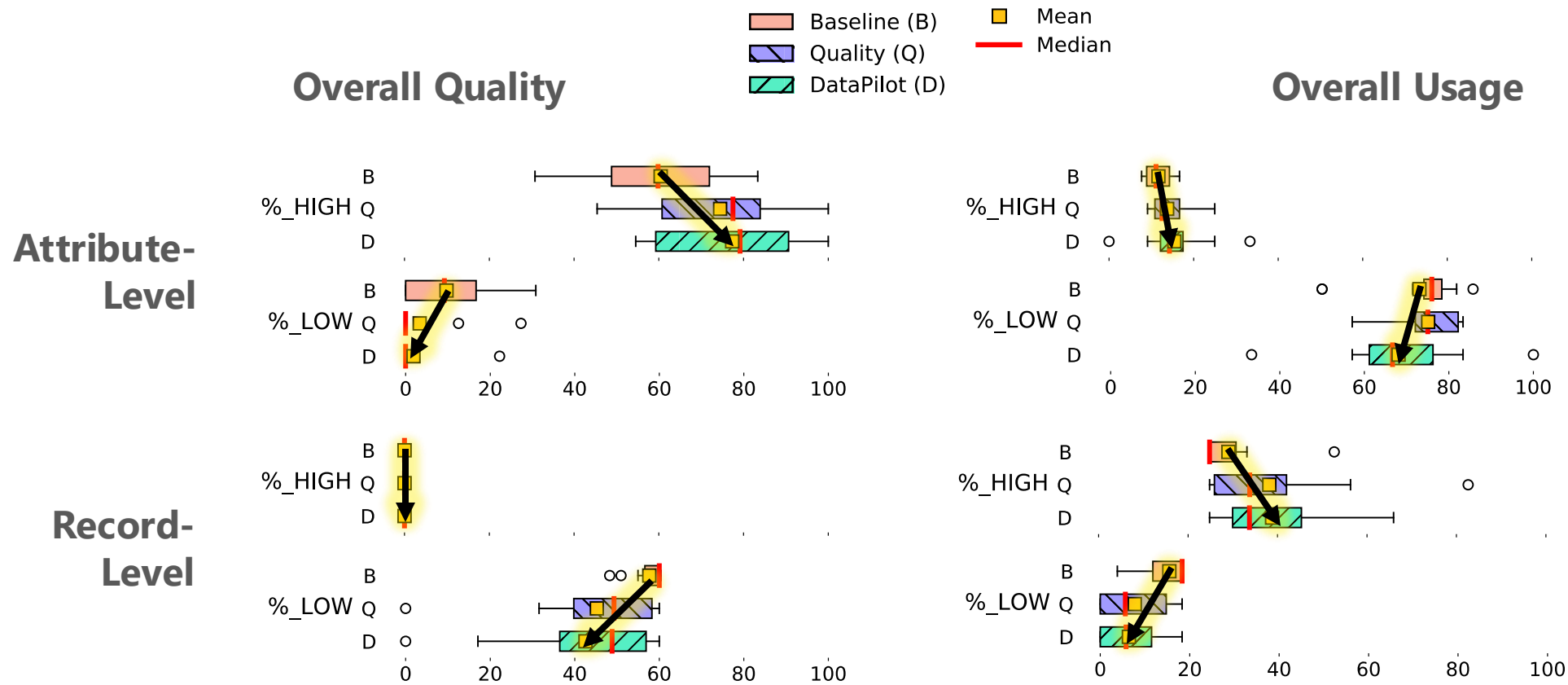
user study findings

DATAPILOT selected smaller subsets



user study findings

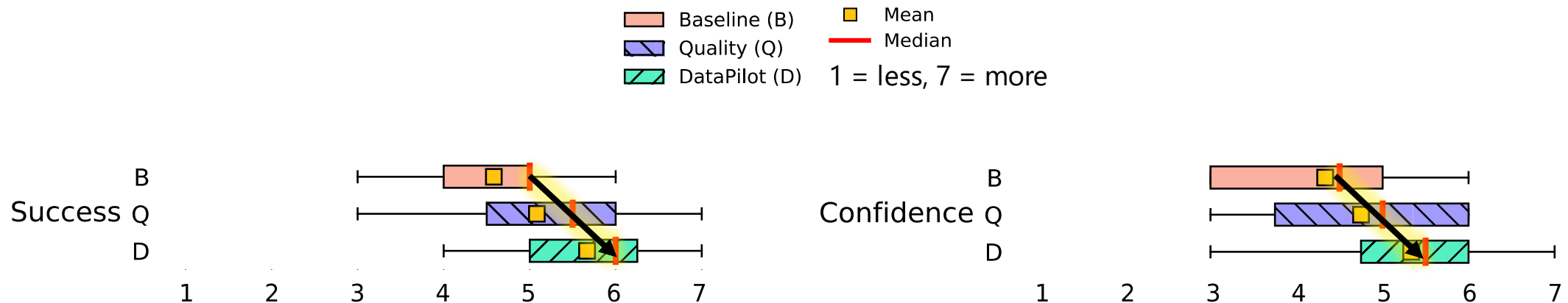
DATAPILOT selected more effective* subsets



* higher % of attributes and records with high overall quality & usage scores

user study findings

DATAPILOT reported greater success and confidence.



... but required some more effort.



general feedback

positive

System Usability Scale: Baseline: 80.21, Quality: 74.17, DataPilot: 71.67

"[this tool] is very powerful to get users to start working on their tasks and to also explore unintended consequences [of the data, their actions]."

"[this information] is missing in current tools like Tableau, PowerBI"

feedback on quality

positive

"users like me who don't write code [will not] ignore data quality aspects and can rely on the green attributes and just get started with analysis."

feedback on usage

mixed

"perform fast and efficient analysis"

"seek validation by checking others' opinions"

"identify new things where other people are not looking"

"avoid repeating past mistakes"

"I can't trust usage data if I don't know the folks and their task(s)"

"it will bias the perspective of people who are doing data exploration"

"it will hamper creativity and originality within users, organizations"

"is it ethical to extract usage data and share it?"

user strategies to select subsets

what data characteristics did they (not) use?

users* strategy

17/36 **data profile:** *attribute name, description, distribution*



7/12 **both quality & usage**

12/24 **only quality**



0/12 **only usage** **quality is still most important**



9/36 **others:** *intuitions, preconceptions, trial & error, common sense*

*per study design, not all users experienced all information.

subset, review, visualize workflow

positive

"it made me think about what [attribute] is important, whereas in Tableau, ... one immediately goes to the chart making step, dragging and dropping attributes hoping to find something interesting."

"I could focus on individual aspects of data and prevent junk data from reaching the chart creation step."

takeaways for current tools

focus on data preparation,
not just visual data analysis.

present quality information, always,
as it nudges users to pause, reflect, and act on the state of their data.

present usage information, but with caution,
as it can reduce exploration of valid data and increase exploitation of popular data.

datapilot

utilizing quality and usage information
for subset selection during visual data preparation



arpit narechania
anarechania3
@gatech.edu



fan du
dufan2013
@gmail.com



atanu sinha
atr
@adobe.com



ryan rossi
ryrossi
@adobe.com



jane hoffswell
jhoffs
@adobe.com



shunan guo
sguo
@adobe.com



eunye koh
eunye
@adobe.com



shamkant navathe
shamkant.navathe
@cc.gatech.edu



alex endert
endert
@gatech.edu

bit.ly/chi23datapilot

