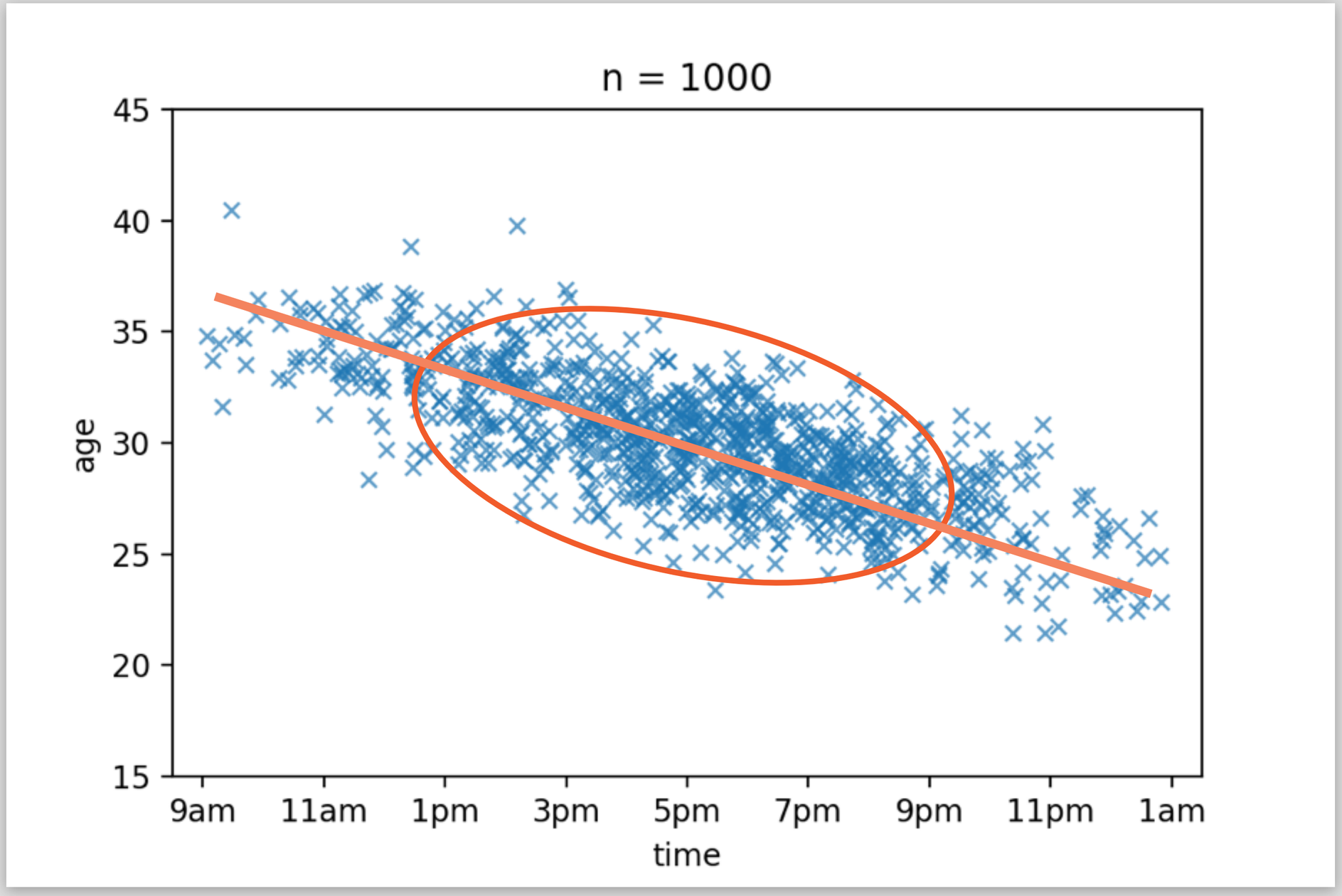
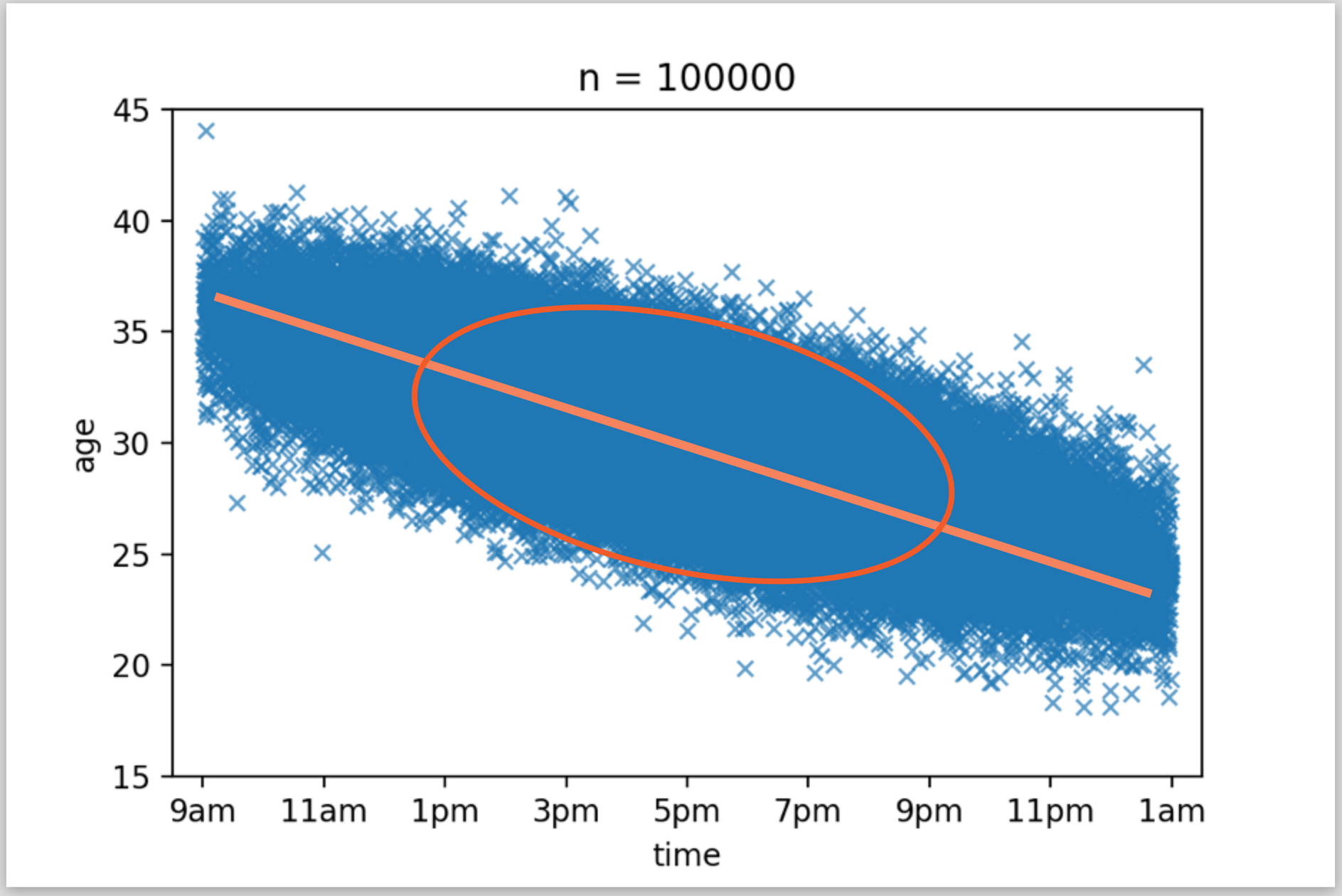
What to Do When Your Data Is Too Big

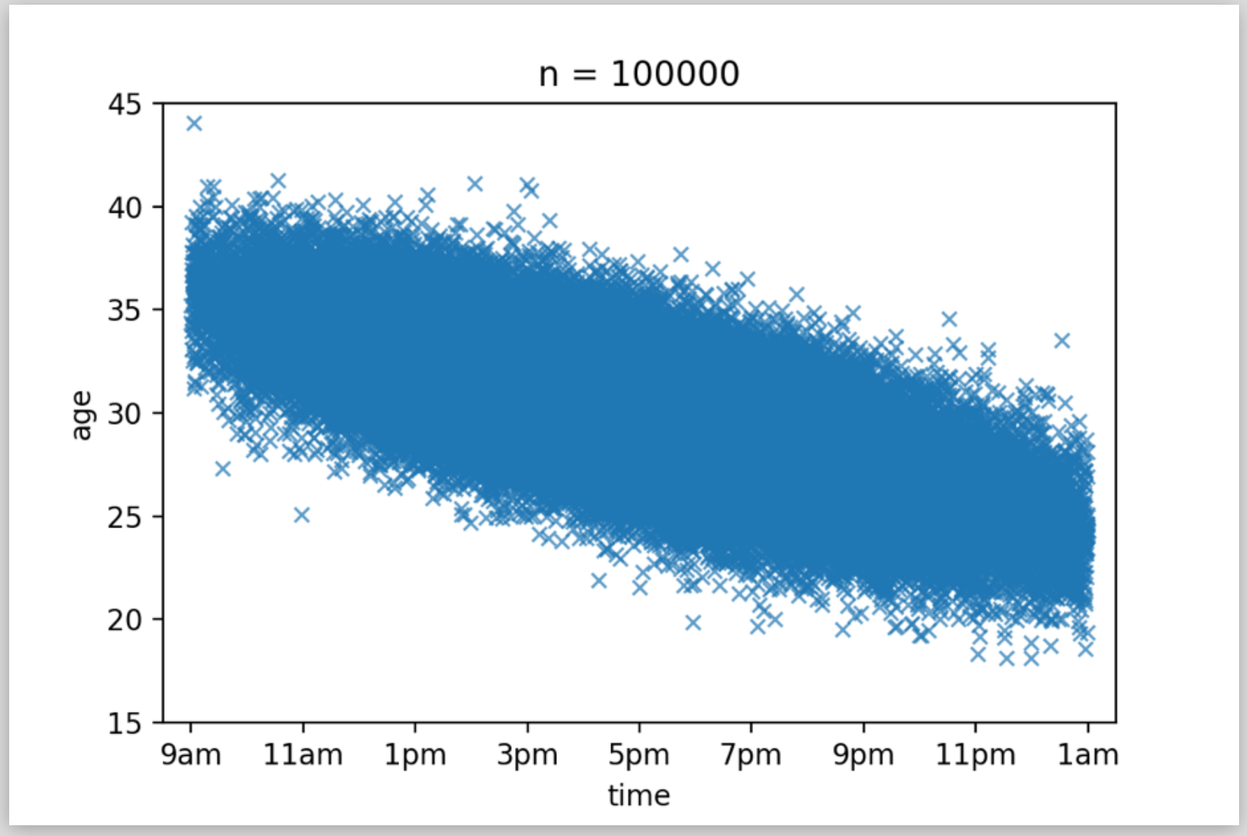
YK Sugishita
SOFTWARE DEVELOPER / DATA SCIENTIST
www.csdojo.io

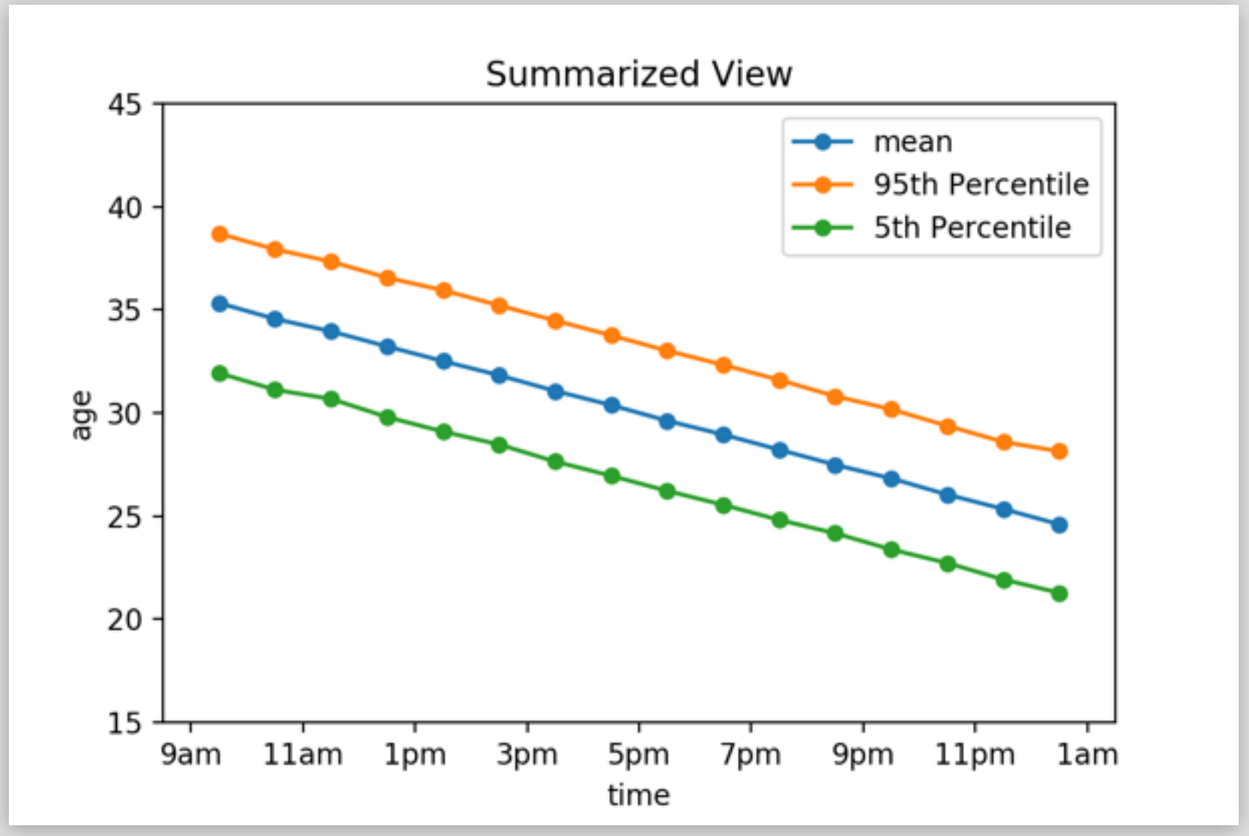

n = 1000
45
40
35
30
25
20
15
age
9am
11am
1pm
3pm
5pm
7pm
9pm
11pm
1am
time

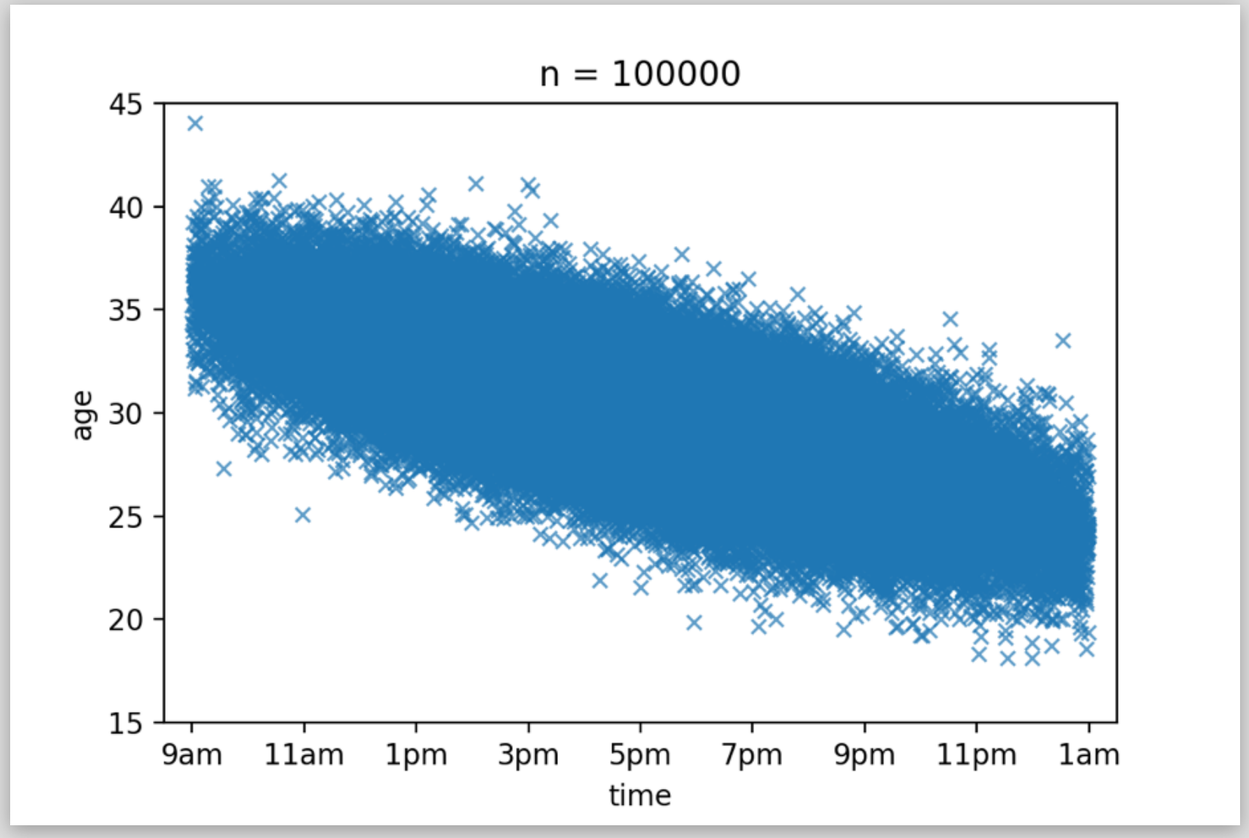

n = 100000
45
40
35
30
25
20
15
age
9am
11am
1pm
3pm
5pm
7pm
9pm
11pm
1am
time

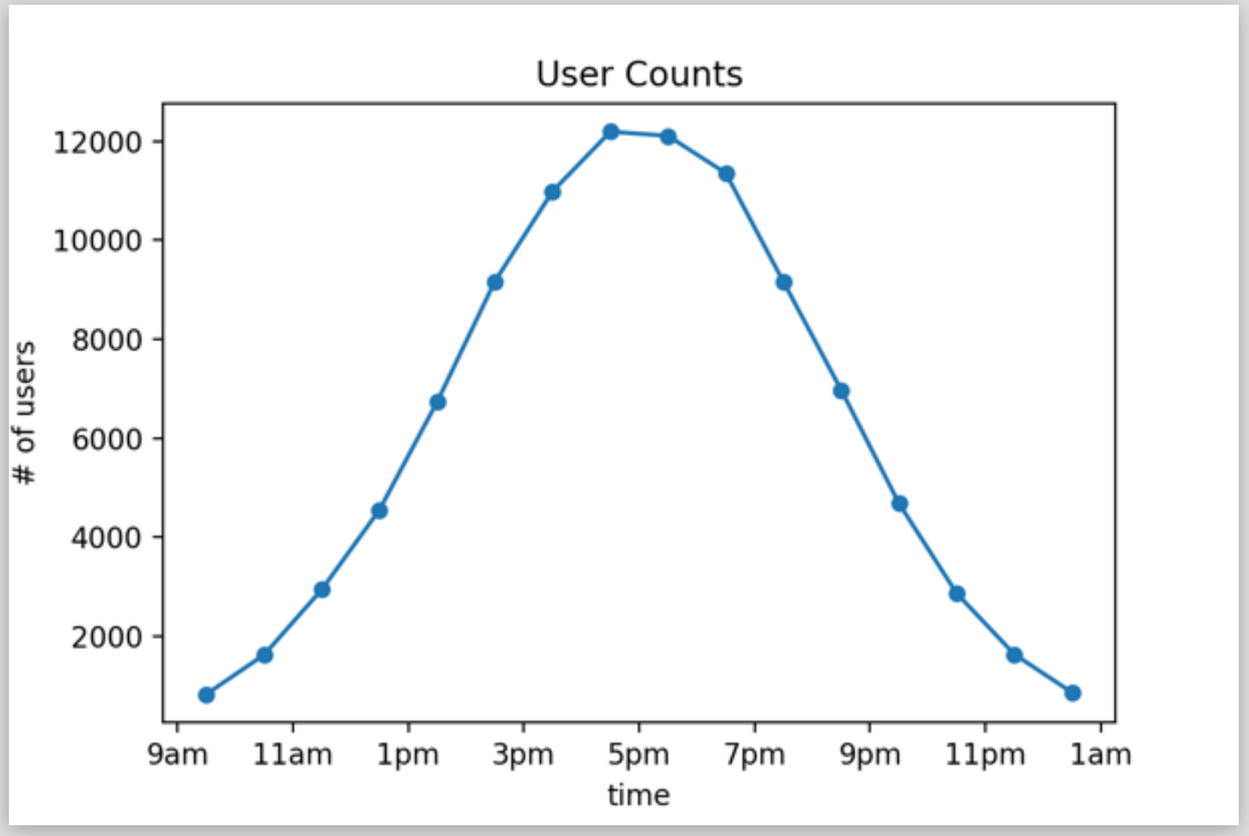
Introduction to Data Aggregation and Sampling

Data Aggregation

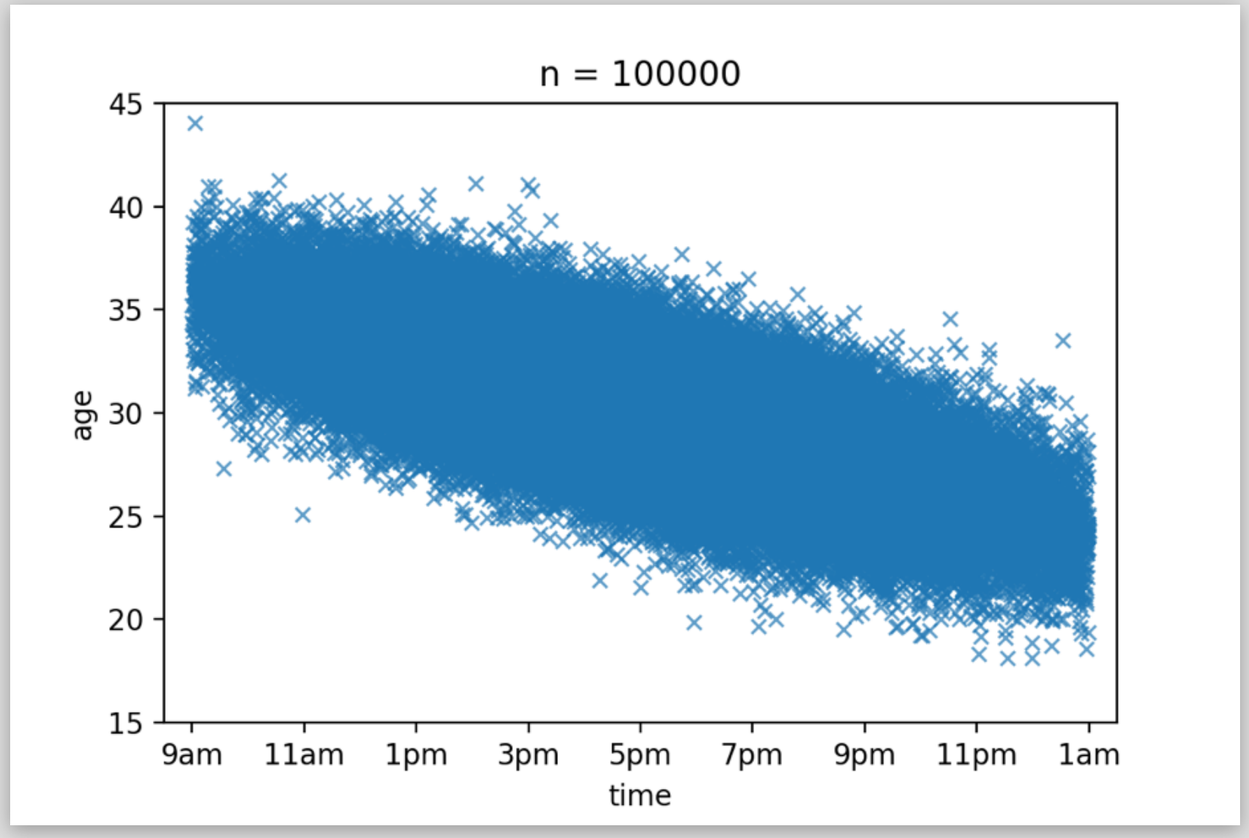

n = 100000
45
40
35
30
25
20
15
age
9am
11am
1pm
3pm
5pm
7pm
9pm
11pm
1am
time

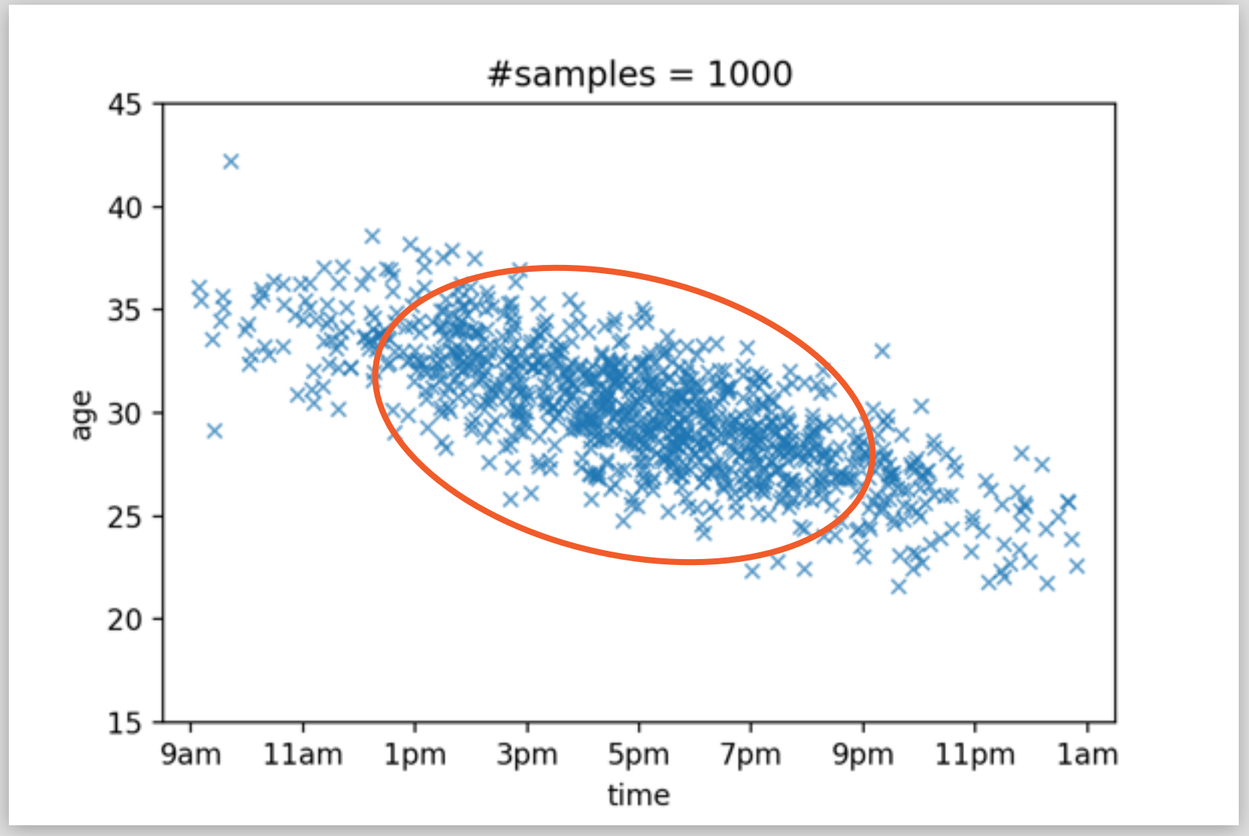

Summarized View
mean
95th Percentile
5th Percentile
45
40
35
30
25
20
15
age
9am
11am
1pm
3pm
5pm
7pm
9pm
11pm
1am
time


n = 100000
45
40
35
30
25
20
15
age
9am
11am
1pm
3pm
5pm
7pm
9pm
11pm
1am
time


User Counts
12000
10000
8000
6000
4000
2000
# of users
9am
11am
1pm
3pm
5pm
7pm
9pm
11pm
1am
time

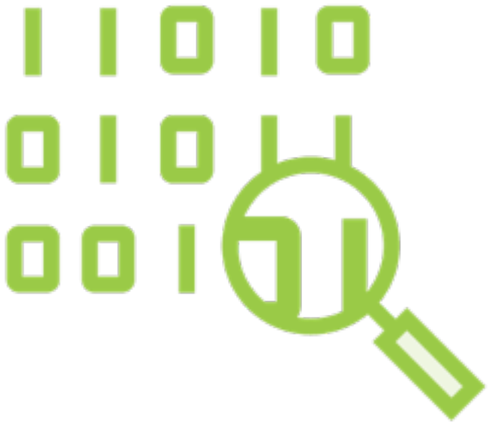
Sampling





Note: Remember to try sampling a few times.

What to Do When Your Data Is Too Big



Aggregate data (mean, percentiles, etc.)

Sample a random subset of data

- Sometimes you don’t need all data

A Data Aggregation Example with Python

obama.csv

year_month	survey_organization	approve_percent	disapprove_percent
2009-01	ABC/Post	80	15
2009-01	AP-GFK	74	15
2009-01	CNN	84	14
		.	
		.	
		.	
2017-01	SurveyMonkey	62	37.5
2017-01	YouGov/Economist	47	42
2017-01	Zogby	50	48

Demo

Data Aggregation Example

- A scatter plot of approval ratings
- Mean and median of approval ratings

A Random Sampling Example with Python

obama_too_big.csv

year_month	survey_organization	approve_percent	disapprove_percent
2009-01	Generated Data	67	31
2009-01	Harris (Phone)	60	29
2009-01	Generated Data	65.9	19.1

.
.
.

2015-12	Monmouth University	36	53
2016-01	Generated Data	46.8	49.2
2016-01	Generated Data	40.2	53.8

Demo

Random Sampling Example

- A scatter plot of approval ratings
- Randomly sample rows