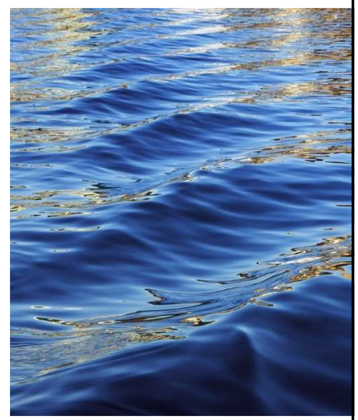
Converting Statistical Literacy Resources to Data Science Resources

Juana Sanchez
UCLA Dept of Statistics and Data Science



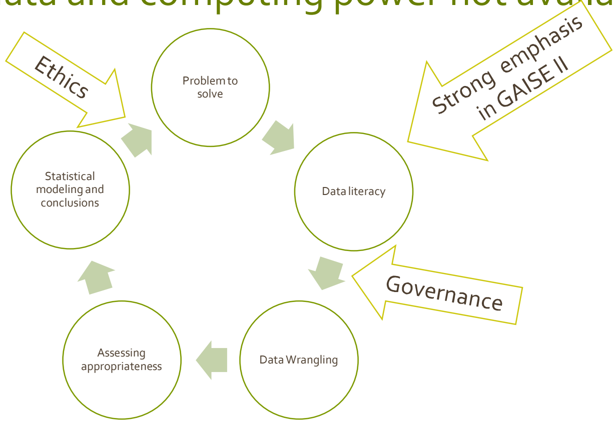
1

Thank you to the ISLP for inviting me to be here

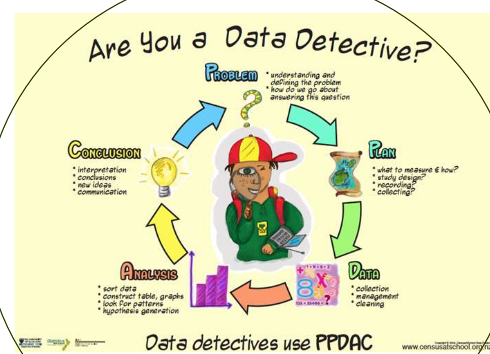
- I have been blessed to have worked for 25 years in an institution, UCLA Statistics and Data Science, where Statistics was always understood and introduced to undergraduates as the science of data. Labs with multivariate datasets, use of software, the PPDAC cycle, and the latest in stats education marked our approach to teaching (GAISE, the ISLP resources, Census@School, statistics education journals, ASA resources, all have played a role, ASA resources….)
- But in recent years, a new challenge emerged: students were hearing about machine learning, artificial intelligence, neural networks. Data Science majors were being created in other departments on campus. Words such as "data science," "data literacy," were popping up everywhere.
- So an existential question came up: what are they doing that we are not?
- This presentation is about some strategies and examples of how I help students realize that the classical curriculum is a crucial component of the emerging data science environment, that there is no data science without statistics.

2

I avoid telling students the obvious: data scientists do what we have always done, getting knowledge from data, but with larger VVV of data and computing power not available to everybody in the past.



PPDAC



Keller, S.A, et al. (2020): Doing Data Science: A Framework and Case Study. HDSR.

GAISE I, GAISE II, Census@School, ISLP, OECD, and many world venues and intro stats books for many years now.

3

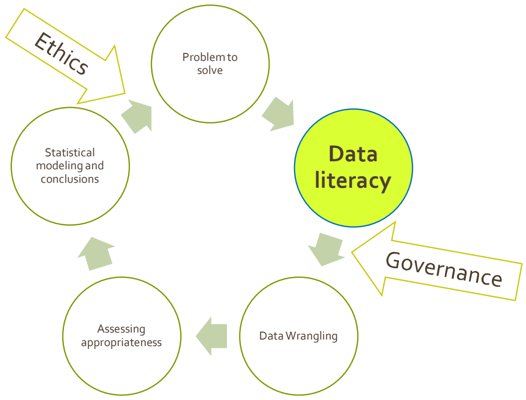
I tell students about language barriers constantly

With data science practitioners coming from different trainings (computer science, or statistics, or engineering field), the names we use in statistics have been renamed in different ways.

Action	Statistics	ML
Orders given to algorithm functions	Arguments of functions	Hyper-parameters
Given names for data collected	Variables	Features
Transformations or combinations of variables	Data wrangling or data management (cleaning, preparing, linking, exploring)	Features engineering
Finding the population model	Estimating the model	Learn the model
Data about the data (metadata, provenance)	Who, what, when, how, where/	Data literacy
Creating knowledge from data	Investigative process	Data pipeline
What lets us generate multivariate random numbers	Joint probability distribution	Generative model

4

The depth and breadth of the connection of our classical statistics curriculum to the widespread data science environment depends on the skill set of the students.



- **Minimum skill set:** "be able to understand information extracted from data and summarized into simple statistics, make further calculations using those statistics and use the statistics to make decisions." Bonikowska et al. (2019) –more than this done in College
- **Broader skill set:** "the ability to ask and answer a real world question from large and small data sets through an inquiry process, with consideration of ethical use of data." Wolff et al. (2016)- Sounds like the whole PPDAC. With different levels of computer skills in between.
- **Narrow definition:** ability to make a data inventory, be able to use all kinds of data available in as many forms as possible. Keller, S.A, et al. (2020)

5 8/9/2023 JSM 2023, Toronto, Canada. Sanchez, J.

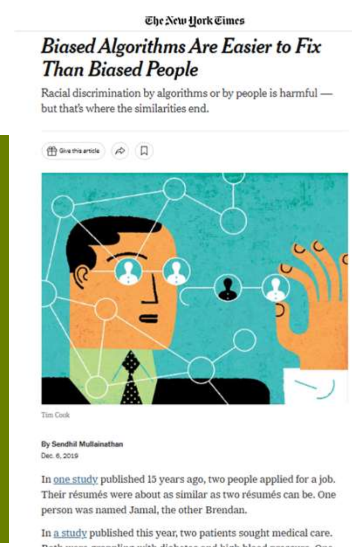
5

Example 1 in Intro Probability

Are artificial intelligent algorithms fair?

Data science context: algorithms used to extract knowledge from data. They are black boxes, some, or too complex, but we can measure their fairness with data about their outcomes and a simple intro stats/intro probability concept.

Intro Probability context: conditional probability, joint probabilities, marginal probabilities, construction of contingency tables from data.



The New York Times

Biased Algorithms Are Easier to Fix Than Biased People

Racial discrimination by algorithms or by people is harmful — but that's where the similarities end.

Give this article

Tim Cook

By Sendhil Mullainathan

Dec. 6, 2019

In one study published 15 years ago, two people applied for a job. Their résumés were about as similar as two résumés can be. One person was named Jamal, the other Brendan.

In a study published this year, two patients sought medical care. Both were grappling with diabetes and high blood pressure. One

https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html

6

LoanID	G	T	D
201	1	1	1
210	0	1	0
214	1	0	1
290	1	1	0
310	1	1	1
340	1	1	1
...	...	...	...
...	...	...	...

Algorithmic fairness

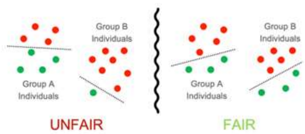
$D=0$

	$G=0$	$G=1$
$T=0$	0.21	0.32
$T=1$	0.07	0.28

$D=1$

	$G=0$	$G=1$
$T=0$	0.01	0.01
$T=1$	0.02	0.08

Tables could be tallied as counts



An artificial intelligence algorithm is going to be used to make a binary prediction for whether a person will repay a loan. The question has come up: is the algorithm "fair" with respect to a binary protected demographic? Notation: G=1 (predict person will pay loan); D =demographic group; T=1 (person pays the loan)

https://chrispiech.github.io/probabilityForComputerScientists/en/examples/fairness/

7

$D=0$

	$G=0$	$G=1$
$T=0$	0.21	0.32
$T=1$	0.07	0.28

$D=1$

	$G=0$	$G=1$
$T=0$	0.01	0.01
$T=1$	0.02	0.08

$$\begin{aligned} P(G=1|D=1) &= \frac{P(G=1,D=1)}{P(D=1)} \\ &= \frac{P(G=1,D=1,T=0)+P(G=1,D=1,T=1)}{P(D=1)} \\ &= \frac{0.01+0.08}{0.12} = 0.75 \end{aligned}$$

$$\begin{aligned} P(G=1|D=0) &= \frac{P(G=1,D=0)}{P(D=0)} \\ &= \frac{P(G=1,D=0,T=0)+P(G=1,D=0,T=1)}{P(D=0)} \\ &= \frac{0.32+0.28}{0.88} \approx 0.68 \end{aligned}$$

Algorithmic fairness concept 1 :demographic parity

https://chrispiech.github.io/probabilityForComputerScientists/en/examples/fairness/

8

	$D=0$			$D=1$	
	$G=0$	$G=1$		$G=0$	$G=1$
$T=0$	0.21	0.32	$T=0$	0.01	0.01
$T=1$	0.07	0.28	$T=1$	0.02	0.08

$$\begin{aligned} P(G=T|D=0) &= P(G=1,T=1|D=0)+P(G=0,T=0|D=0) \\ &= \frac{0.28+0.21}{0.88} \approx 0.56 \\ P(G=T|D=1) &= P(G=1,T=1|D=1)+P(G=0,T=0|D=1) \\ &= \frac{0.08+0.01}{0.12} = 0.75 \end{aligned}$$

Algorithmic fairness concept 2: calibration

https://chrispiech.github.io/probabilityForComputerScientists/en/examples/fairness/

9

	$D=0$			$D=1$	
	$G=0$	$G=1$		$G=0$	$G=1$
$T=0$	0.21	0.32	$T=0$	0.01	0.01
$T=1$	0.07	0.28	$T=1$	0.02	0.08

$$\begin{aligned} P(G=1|D=1,T=1) &= \frac{P(G=1,D=1,T=1)}{P(D=1,T=1)} \\ &= \frac{0.08}{0.08+0.02} = 0.8 \\ P(G=1|D=0,T=1) &= \frac{P(G=1,D=0,T=1)}{P(D=0,T=1)} \\ &= \frac{0.28}{0.28+0.07} = 0.8 \end{aligned}$$

Algorithmic fairness concept 3: equality of odds

https://chrispiech.github.io/probabilityForComputerScientists/en/examples/fairness/

10

For formative assessment, students do a survey of UCLA students and construct similar tables and demonstrate Bayes theorem.

For further discussion, talk about how generative AI models use joint probabilities to create new (synthetic) data and how discriminative AI models use existing data to classify it

https://chrispiech.github.io/probabilityForComputerScientists/en/examples/fairness/

11

Example 2 – In Intro Time Series

Features engineering

Data science context: Forecasting hourly electricity demand supplied by Southern Edison

https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html

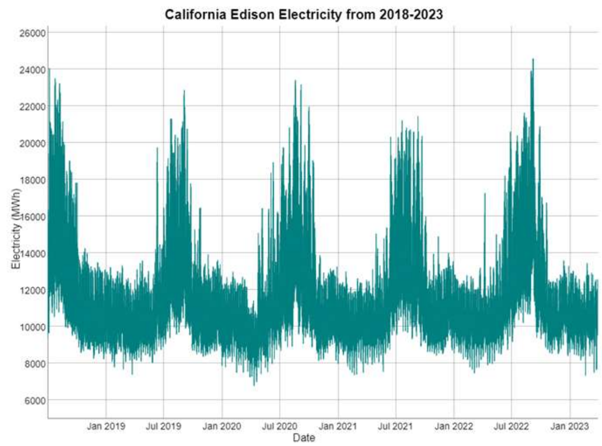
12

Most data collected nowadays is timestamped

- Hourly demand for electricity supplied by Southern California Edison 2018/07/01 8:00 AM- 2023/3/31 12:00AM.EIA, www.eia.gov

Sanchez, J. (2023) , case study for Chapter 10, found in timeseriestime.org

	date	value
	<dttm>	<dbl>
1	2018-07-01 08:00:00	10681
2	2018-07-01 09:00:00	10197
3	2018-07-01 10:00:00	9776
4	2018-07-01 11:00:00	9508
5	2018-07-01 12:00:00	9431
6	2018-07-01 13:00:00	9472
7	2018-07-01 14:00:00	9353
8	2018-07-01 15:00:00	9517
9	2018-07-01 16:00:00	9785
10	2018-07-01 17:00:00	10137
11	2018-07-01 18:00:00	10600
12	2018-07-01 19:00:00	11099
13	2018-07-01 20:00:00	11671
14	2018-07-01 21:00:00	12315
15	2018-07-01 22:00:00	12940
16	2018-07-01 23:00:00	13611
17	2018-07-02 00:00:00	14176
18	2018-07-02 01:00:00	14577
19	2018-07-02 02:00:00	14699
20	2018-07-02 03:00:00	14266
21	2018-07-02 04:00:00	14059
22	2018-07-02 05:00:00	13609
23	2018-07-02 06:00:00	12591
24	2018-07-02 07:00:00	11611



Prepare data for ML (RF, GB, NN) and regular multiple regression (and intro stats)

- Hourly demand for electricity supplied by Southern California Edison 2018/07/01 8:00 AM- 2023/3/31 12:00AM

	date	value
	<dttm>	<dbl>
1	2018-07-01 08:00:00	10681
2	2018-07-01 09:00:00	10197
3	2018-07-01 10:00:00	9776
4	2018-07-01 11:00:00	9508
5	2018-07-01 12:00:00	9431
6	2018-07-01 13:00:00	9472
7	2018-07-01 14:00:00	9353
8	2018-07-01 15:00:00	9517
9	2018-07-01 16:00:00	9785
10	2018-07-01 17:00:00	10137
11	2018-07-01 18:00:00	10600
12	2018-07-01 19:00:00	11099
13	2018-07-01 20:00:00	11671
14	2018-07-01 21:00:00	12315
15	2018-07-01 22:00:00	12940
16	2018-07-01 23:00:00	13611
17	2018-07-02 00:00:00	14176
18	2018-07-02 01:00:00	14577
19	2018-07-02 02:00:00	14699
20	2018-07-02 03:00:00	14266
21	2018-07-02 04:00:00	14059
22	2018-07-02 05:00:00	13609
23	2018-07-02 06:00:00	12591
24	2018-07-02 07:00:00	11611

Features engineering

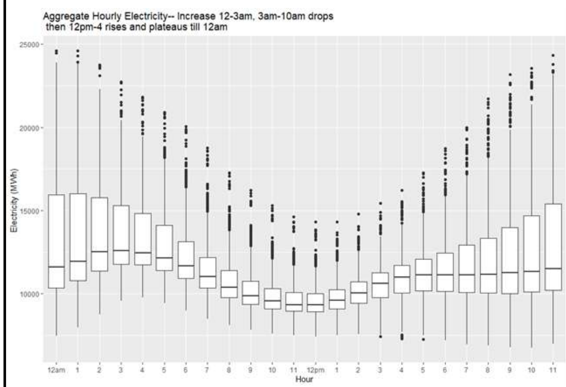
A multivariate data set format familiar to intro stats students for training ML models, such as NN, RF, GB

A tibble: 32,801 x 22

	date	y	hour	day_of_week	month	year	covid	lag_hour	lag_two	lag_three	lag_four
	<date>	<dbl>	<int>	<ord>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2019-07-02	9869	11	Tue	7	2019	0	10149	10646	11244	12161
2	2019-07-02	9982	12	Tue	7	2019	0	9869	10149	10646	11244
3	2019-07-02	10412	13	Tue	7	2019	0	9982	9869	10149	10646
4	2019-07-02	10864	14	Tue	7	2019	0	10412	9982	9869	10149
5	2019-07-02	11351	15	Tue	7	2019	0	10864	10412	9982	9869
6	2019-07-02	11745	16	Tue	7	2019	0	11351	10864	10412	9982
7	2019-07-02	12207	17	Tue	7	2019	0	11745	11351	10864	10412
8	2019-07-02	12643	18	Tue	7	2019	0	12207	11745	11351	10864
9	2019-07-02	13189	19	Tue	7	2019	0	12643	12207	11745	11351
10	2019-07-02	13716	20	Tue	7	2019	0	13189	12643	12207	11745
11	2019-07-02	14398	21	Tue	7	2019	0	13716	13189	12643	12207
12	2019-07-02	15073	22	Tue	7	2019	0	14398	13716	13189	12643
13	2019-07-02	15594	23	Tue	7	2019	0	15073	14398	13716	13189
14	2019-07-03	15931	0	Wed	7	2019	0	15594	15073	14398	13716
15	2019-07-03	16037	1	Wed	7	2019	0	15931	15594	15073	14398
16	2019-07-03	15878	2	Wed	7	2019	0	16037	15931	15594	15073
17	2019-07-03	15363	3	Wed	7	2019	0	15878	16037	15931	15594
18	2019-07-03	15010	4	Wed	7	2019	0	15363	15878	16037	15931
19	2019-07-03	14466	5	Wed	7	2019	0	15010	15363	15878	16037

Surprisingly the ML-ready data allows us to complete the PPDAC cycle. Many possible questions to start with.

- Hourly demand for electricity supplied by Southern California Edison 2018/07/01 8:00 AM- 2023/3/31 12:00AM



Features

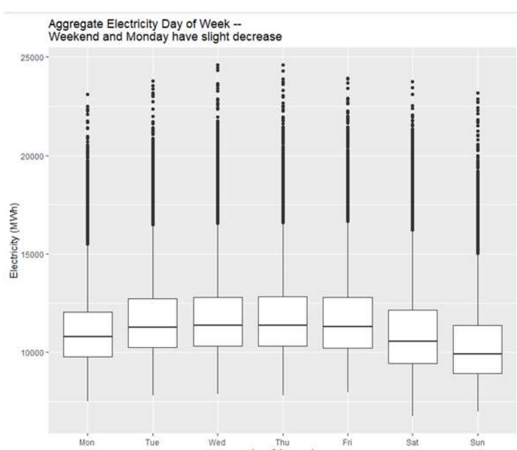
A tibble: 32,801 × 22

	date	y	hour	day_of_week	month	year	covid	lag_hour	lag_two	lag_three	lag_four
	<date>	<dbl>	<int>	<ord>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2019-07-02	9869	11	Tue	7	2019	0	10149	10646	11244	12161
2	2019-07-02	9982	12	Tue	7	2019	0	9869	10149	10646	11244
3	2019-07-02	10412	13	Tue	7	2019	0	9982	9869	10149	10646
4	2019-07-02	10864	14	Tue	7	2019	0	10412	9982	9869	10149
5	2019-07-02	11351	15	Tue	7	2019	0	10864	10412	9982	9869
6	2019-07-02	11745	16	Tue	7	2019	0	11351	10864	10412	9982
7	2019-07-02	12207	17	Tue	7	2019	0	11745	11351	10864	10412
8	2019-07-02	12643	18	Tue	7	2019	0	12207	11745	11351	10864
9	2019-07-02	13189	19	Tue	7	2019	0	12643	12207	11745	11351
10	2019-07-02	13716	20	Tue	7	2019	0	13189	12643	12207	11745
11	2019-07-02	14398	21	Tue	7	2019	0	13716	13189	12643	12207
12	2019-07-02	15073	22	Tue	7	2019	0	14398	13716	13189	12643
13	2019-07-02	15594	23	Tue	7	2019	0	15073	14398	13716	13189
14	2019-07-03	15931	0	Wed	7	2019	0	15594	15073	14398	13716
15	2019-07-03	16037	1	Wed	7	2019	0	15931	15594	15073	14398
16	2019-07-03	15878	2	Wed	7	2019	0	16037	15931	15594	15073
17	2019-07-03	15363	3	Wed	7	2019	0	15878	16037	15931	15594
18	2019-07-03	15010	4	Wed	7	2019	0	15363	15878	16037	15931
19	2019-07-03	14466	5	Wed	7	2019	0	15010	15363	15878	16037

15

Questioning throughout the analysis. Is the day of the week important?

- Hourly demand for electricity supplied by Southern California Edison 2018/07/01 8:00 AM- 2023/3/31 12:00AM



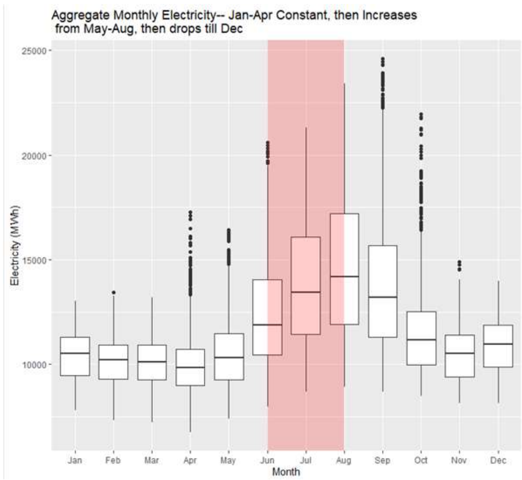
Features

A tibble: 32,801 × 22

	date	y	hour	day_of_week	month	year	covid	lag_hour	lag_two	lag_three	lag_four
	<date>	<dbl>	<int>	<ord>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2019-07-02	9869	11	Tue	7	2019	0	10149	10646	11244	12161
2	2019-07-02	9982	12	Tue	7	2019	0	9869	10149	10646	11244
3	2019-07-02	10412	13	Tue	7	2019	0	9982	9869	10149	10646
4	2019-07-02	10864	14	Tue	7	2019	0	10412	9982	9869	10149
5	2019-07-02	11351	15	Tue	7	2019	0	10864	10412	9982	9869
6	2019-07-02	11745	16	Tue	7	2019	0	11351	10864	10412	9982
7	2019-07-02	12207	17	Tue	7	2019	0	11745	11351	10864	10412
8	2019-07-02	12643	18	Tue	7	2019	0	12207	11745	11351	10864
9	2019-07-02	13189	19	Tue	7	2019	0	12643	12207	11745	11351
10	2019-07-02	13716	20	Tue	7	2019	0	13189	12643	12207	11745
11	2019-07-02	14398	21	Tue	7	2019	0	13716	13189	12643	12207
12	2019-07-02	15073	22	Tue	7	2019	0	14398	13716	13189	12643
13	2019-07-02	15594	23	Tue	7	2019	0	15073	14398	13716	13189
14	2019-07-03	15931	0	Wed	7	2019	0	15594	15073	14398	13716
15	2019-07-03	16037	1	Wed	7	2019	0	15931	15594	15073	14398
16	2019-07-03	15878	2	Wed	7	2019	0	16037	15931	15594	15073
17	2019-07-03	15363	3	Wed	7	2019	0	15878	16037	15931	15594
18	2019-07-03	15010	4	Wed	7	2019	0	15363	15878	16037	15931
19	2019-07-03	14466	5	Wed	7	2019	0	15010	15363	15878	16037

16

Do some months have more demand than others?



- Hourly demand for electricity supplied by Southern California Edison 2018/07/01 8:00 AM- 2023/3/31 12:00AM

Features

A tibble: 32,801 × 22

	date	y	hour	day_of_week	month	year	covid	lag_hour	lag_two	lag_three	lag_four
	<date>	<dbl>	<int>	<ord>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2019-07-02	9869	11	Tue	7	2019	0	10149	10646	11244	12161
2	2019-07-02	9982	12	Tue	7	2019	0	9869	10149	10646	11244
3	2019-07-02	10412	13	Tue	7	2019	0	9982	9869	10149	10646
4	2019-07-02	10864	14	Tue	7	2019	0	10412	9982	9869	10149
5	2019-07-02	11351	15	Tue	7	2019	0	10864	10412	9982	9869
6	2019-07-02	11745	16	Tue	7	2019	0	11351	10864	10412	9982
7	2019-07-02	12207	17	Tue	7	2019	0	11745	11351	10864	10412
8	2019-07-02	12643	18	Tue	7	2019	0	12207	11745	11351	10864
9	2019-07-02	13189	19	Tue	7	2019	0	12643	12207	11745	11351
10	2019-07-02	13716	20	Tue	7	2019	0	13189	12643	12207	11745
11	2019-07-02	14398	21	Tue	7	2019	0	13716	13189	12643	12207
12	2019-07-02	15073	22	Tue	7	2019	0	14398	13716	13189	12643
13	2019-07-02	15594	23	Tue	7	2019	0	15073	14398	13716	13189
14	2019-07-03	15931	0	Wed	7	2019	0	15594	15073	14398	13716
15	2019-07-03	16037	1	Wed	7	2019	0	15931	15594	15073	14398
16	2019-07-03	15878	2	Wed	7	2019	0	16037	15931	15594	15073
17	2019-07-03	15363	3	Wed	7	2019	0	15878	16037	15931	15594
18	2019-07-03	15010	4	Wed	7	2019	0	15363	15878	16037	15931
19	2019-07-03	14466	5	Wed	7	2019	0	15010	15363	15878	16037

17

Other questions: is demand the hour before important? Etc.

If we did a regression, which variable would be most important?

Difficult with a multiple regression, easier with a regression tree.

- Hourly demand for electricity supplied by Southern California Edison 2018/07/01 8:00 AM- 2023/3/31 12:00AM

Features

A tibble: 32,801 × 22

	date	y	hour	day_of_week	month	year	covid	lag_hour	lag_two	lag_three	lag_four
	<date>	<dbl>	<int>	<ord>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2019-07-02	9869	11	Tue	7	2019	0	10149	10646	11244	12161
2	2019-07-02	9982	12	Tue	7	2019	0	9869	10149	10646	11244
3	2019-07-02	10412	13	Tue	7	2019	0	9982	9869	10149	10646
4	2019-07-02	10864	14	Tue	7	2019	0	10412	9982	9869	10149
5	2019-07-02	11351	15	Tue	7	2019	0	10864	10412	9982	9869
6	2019-07-02	11745	16	Tue	7	2019	0	11351	10864	10412	9982
7	2019-07-02	12207	17	Tue	7	2019	0	11745	11351	10864	10412
8	2019-07-02	12643	18	Tue	7	2019	0	12207	11745	11351	10864
9	2019-07-02	13189	19	Tue	7	2019	0	12643	12207	11745	11351
10	2019-07-02	13716	20	Tue	7	2019	0	13189	12643	12207	11745
11	2019-07-02	14398	21	Tue	7	2019	0	13716	13189	12643	12207
12	2019-07-02	15073	22	Tue	7	2019	0	14398	13716	13189	12643
13	2019-07-02	15594	23	Tue	7	2019	0	15073	14398	13716	13189
14	2019-07-03	15931	0	Wed	7	2019	0	15594	15073	14398	13716
15	2019-07-03	16037	1	Wed	7	2019	0	15931	15594	15073	14398
16	2019-07-03	15878	2	Wed	7	2019	0	16037	15931	15594	15073
17	2019-07-03	15363	3	Wed	7	2019	0	15878	16037	15931	15594
18	2019-07-03	15010	4	Wed	7	2019	0	15363	15878	16037	15931
19	2019-07-03	14466	5	Wed	7	2019	0	15010	15363	15878	16037

18

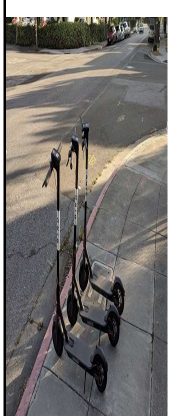
year	month	day	hour	osm_way_id	osm_start_node_id	osm_end_node_id	speed_mph_mean	speed_mph_stddev
2020	1	1	1	40722998	62385707	4927951349	26.636	4.483
2020	1	31	21	40722998	62385707	4927951349	25.513	4.276
2020	1	1	0	40722998	62385707	4927951349	27.521	5.105
2020	1	1	0	40722998	5780849015	4927951349	26.05	3.803
2020	1	1	1	40722998	5780849015	4927951349	25.459	3.585
2020	1	30	8	417094233	4714793573	1014244233	27.761	3.679
2020	1	7	15	416137931	239464357	4318478540	25.721	1.649
2020	1	30	18	416137931	239464357	4318478540	25.222	7.128
2020	1	4	11	416137931	239464357	4318478540	23.629	3.669
2020	1	17	17	416137931	239464357	4318478540	22.642	3.554
2020	1	22	17	416137931	239464357	4318478540	23.842	4.381
2020	1	9	17	416137931	239464357	4318478540	29.338	14.674
2020	1	29	10	416137931	239464357	4318478540	23.056	3.197
2020	1	17	15	416137931	239464357	4318478540	27.031	5.015
2020	1	5	18	416137931	239464357	4318478540	23.461	3.422
2020	1	30	19	416137931	239464357	4318478540	23.45	1.53
2020	1	25	14	416137931	239464357	4318478540	26.481	2.493
2020	1	27	14	416137931	239464357	4318478540	26.054	3.478
2020	1	27	17	416137931	239464357	4318478540	32.316	18.225

For further formative assessment, use Uber movement anonymized data to help urban planning

Uber already publishes its data in contemporary data science format ready to be used in ML models.

For further discussion, how would a regression tree be formed if we used just regression.

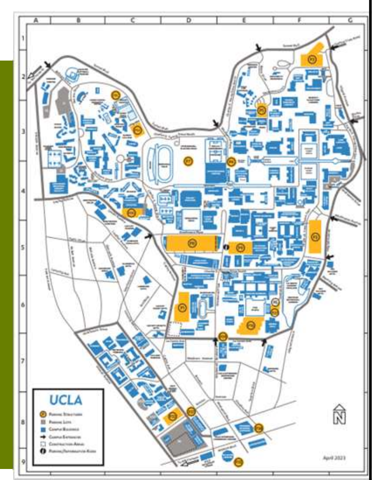
19



Example 3 – In Intro Probability

Micromobility at a small scale

Data science context: Does the distribution of scooters across campus follow a Poisson process?



https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html

20

Training students first

```
1951 2 3402 1191
--------------------------------4 births in the 20th hour
2010 1 3500 1210
2037 2 3736 1237
2051 2 3370 1251
------------------------------3 births in the 21st hour
2104 2 2121 1264
2123 2 3150 1283
------------------------------ 2 births in the 22nd hour
2217 1 3866 1337
--------------------------------1 birth in the 23rd hour
2327 1 3542 1407
2355 1 3278 1435
------------------------------ 2 births in the 24th hour
```

Number of Births per hour	Tally (in how many of the hours did we observe the number of births in column 1) (Observed)	Empirical Probability (this is the observed relative frequency)	Theoretical Probability (with Poisson model with lambda=44/24=1.83 births per hour)
0	3	3/24 = 0.125	$\frac{1.83^0 e^{-1.83}}{0!} = 0.160$
1	8	8/24 = 0.333	$\frac{1.83^1 e^{-1.83}}{1!} = 0.293$
2	6	0.250	0.269
3	4	0.167	0.164
4	3	0.125	0.075
5+	0	0.000	0.039
Total	24 hours	1	1

Number of Births per hour	Tally (in how many of the hours did we observe the number of births in column 1) (Observed)	Empirical Probability (this is the observed relative frequency)	Theoretical Probability (with Poisson model with lambda=44/24=1.83 births per hour) (Expected in red color)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	3	3/24 = 0.125	$\frac{1.83^0 e^{-1.83}}{0!} = 0.160$ (0.160*24=3.84)	$(3-3.84)^2 = 0.7056$	0.18375
1	8	8/24 = 0.333	$\frac{1.83^1 e^{-1.83}}{1!} = 0.293$ 0.293*24=7.032	(8-7.032)= 0.937024	0.13325142
2	6	0.250	0.269 0.269*24=6.456	6-6.456= 0.207936	0.03220818
3	4	0.167	0.164 0.164*24=3.936	4-3.936= 0.004096	0.00104065
4	3	0.125	0.075 0.075*24=1.8	3-1.8= 1.44	0.8
5+	0	0.000	0.039 0.039*24=0.936	0-0.936= 0.876096	0.9360
Total	24 hours	1	1		

$$Sum\ of\ \frac{(O-E)^2}{E} = 0.18375 + \cdots \ldots + 0.9360 = 2.08625$$

The Chi-square statistic equals 2.08625.

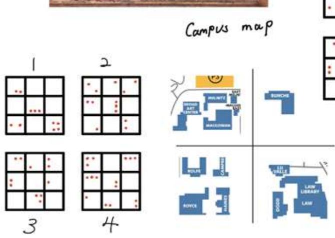
Looking at the app,

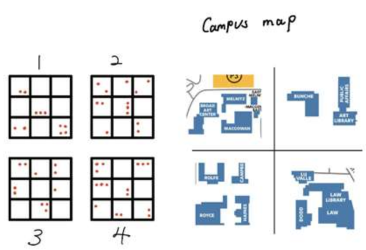
P("Chi-square with 5 degrees of freedom" > 2.08625) = 0.83709

Because the P-square statistic is larger than 0.05, a statistician would conclude that the Poisson Model with parameter lambda equal to 1.83 is a good fit to the birth data.

Students go to the field, collect and describe

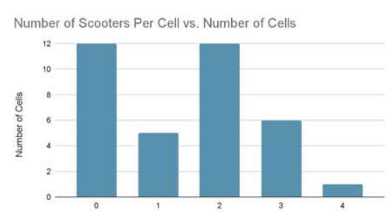
Group plans and collects data





Group tallies and summarizes (data wrangling)

Number of Scooters Per Cell	Number of Cells With That Number of Scooters
0	12
1	5
2	12
3	6
4	1



Students fit estimated probability model

Calculate what is needed

Estimate of $\lambda = \frac{0\times12 + 1\times5 + 2\times12 + 3\times6 + 4\times1}{36} = 1.42$

$P(X=x) = \frac{1.42^x e^{-1.42}}{x!}$

Theoretical Probabilities:

$\frac{1.42^0 e^{-1.42}}{0!} = 0.24$ $\frac{1.42^2 e^{-1.42}}{2!} = 0.24$ $\frac{1.42^4 e^{-1.42}}{4!} = 0.04$

$\frac{1.42^1 e^{-1.42}}{1!} = 0.34$ $\frac{1.42^3 e^{-1.42}}{3!} = 0.115$

Predicted # Per Cell:

$0.24 \times 36 = 8.64$ $0.24 \times 36 = 8.64$ $0.04 \times 36 = 1.44$

$0.34 \times 36 = 12.24$ $0.115 \times 36 = 4.14$

Realize that probability is also used to draw inferences

# of scooters per cell (X)	Observed (O)	$P(X=x)$	# of scooters Predicted (E)	$\frac{(E-O)^2}{E}$
0	12	0.24	8.64	1.3
1	5	0.34	12.24	4.28
2	12	0.24	8.64	1.3
3	6	0.115	4.14	0.83
4	1	0.04	1.44	0.13
		≈ 1	≈ 36	7.84

χ^2 = chi square statistic = 7.84

$5-1 = 4$ degrees of freedom

$P(\chi^2_4 > 7.84) = 0.097$

Poisson Distribution Formula

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where
x = 0, 1, 2, 3, ...
λ = mean number of occurrences in the interval
e = Euler's constant ≈ 2.71828

Because the P-square statistic is larger than 0.05, we can conclude the Poisson model with $\lambda = 1.42$ is a good fit to the data.

Source: students' paper.

23

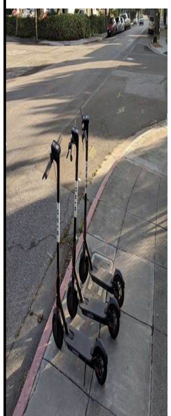
Students criticize the approach and suggest

More variables would help predict better

The data collection was not done the same day or hour

More data and better coverage of areas of campus in the sampling needed.

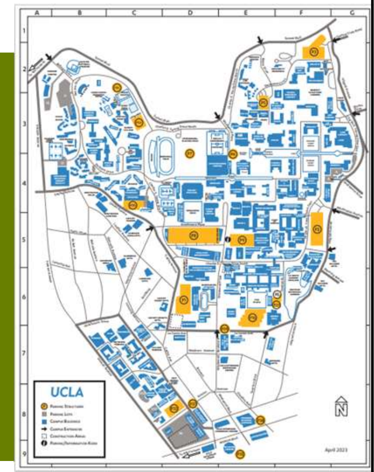
24



Example 4 – InTime Series

Clustering as in customer segmentation, but with rivers time series

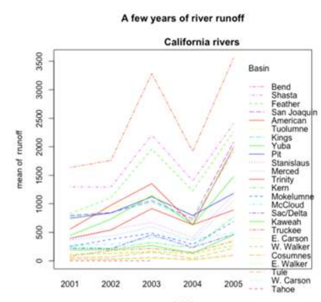
Data science context: Are rivers in California very different?

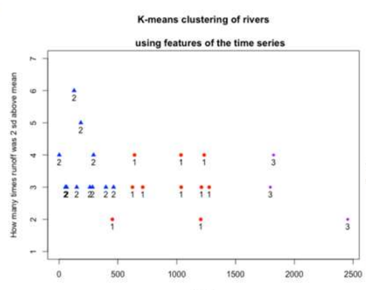


https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html

25

Time series data converted to summarized features data – simple features, for unsupervised machine learning



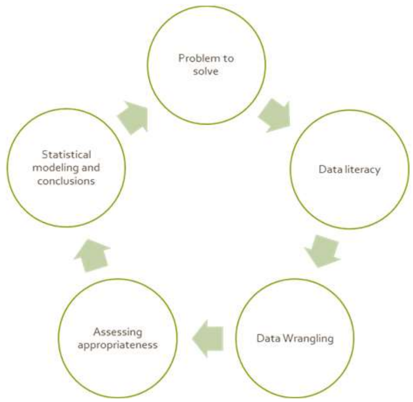


Basin	ID	1930	1931	1932	1933	1934	1935	1936	1937	1938	1939	1940	1941
Trinity	TNL	319.6	201.72	425.5	578.9	264.22	559.31	535.51	800.01	1087.4	256.21	620.28	1400.3
Sac/Delta	SDT	170.1	69.2	170.9	245.1	123.4	308.7	219.2	398.5	499.3	101.7	272.4	636.1
McCloud	MSS	284.1	193.9	285.4	342.8	242.4	459.98	318.68	470.82	744.07	279.12	481.56	748.03
Pit	PSH	703.1	480.1	805.1	759.7	535.3	1374.3	766.4	969	1615.6	592.4	968.2	1202.6
Shasta	SIS	1143	764	1312	1374	926	2275	1343	1969	3060	996	1849	2791
Bend	SBB	1644	943	1689	1770	1186	3336	1854	2600	4062	1267	2660	4235
Feather	FTO	1426	502.1	1742	1142	594.2	2892	1648	1940	4321	748.5	1833	2569
Yuba	YRS	752.8	279.8	1226.1	775.1	310.5	1547.2	1240.6	1220.2	2075.2	450.21	1056.32	1434.55
American	AMF	829.2	363.9	1579.8	977.2	361.7	1915	1663.8	1476.8	2475.1	572.7	1378.5	1531
Cosumnes	CSN	61.97	12.3	114.23	79.33	17.85	257.16	149.88	177.51	276.75	35.25	130.68	156.39

Basin	mean	sd.	min	. max	$+2sd$	$-1sd$
Trinity	641.12	294.28.	116.62.	1593.35	4	10
Sac/Delta	292.66	141.48	63.42	711.20	4	8
McCloud	397.87	132.07	184.67	748.03	3	9
Pit	1037.95	332.54	480.10	2097.72	3	8
Shasta	1795.64	660.68	764.00	3525.31	3	9
Bend	2453.62	989.65	943.00	5075.46	2	7
Feather	1822.91	962.31	391.85	4676.00	4	11
Yuba	1036.31	501.31	199.88	2424.09	4	12
American	1275.89	658.01	228.96	2912.26	3	13
Cosumnes	127.56	93.13	7.96	362.84	6	10
Mokelumne	463.32	219.65	101.59	1038.00	3	13
Stanislaus	710.88	351.49	115.51	1636.18	3	12
Tuolumne	1210.10	559.48	301.02	2645.28	3	13
Merced	623.45	332.30	123.29	1587.46	3	11
San Joaquin	1233.34	641.00	261.91	2898.00	4	10
Kings	1203.78	633.85	274.49	3112.61	2	13
Kaweah	283.91	170.07	61.72	799.70	3	10
Tule	63.26	56.19	2.36	259.14	3	3
Kern	452.69	328.28	84.39	1657.07	2	6
Truckee	261.81	147.88	52.42	712.73	3	12
Tahoe	1.39	0.82	0.17	3.57	4	15
W. Carson	54.12	26.40	12.06	135.21	3	12
E. Carson	184.80	88.67	42.57	406.72	5	12
W. Walker	149.87	63.97	34.79	303.33	3	13
E. Walker	62.13	45.11	6.66	209.04	3	7

26 8/9/2023 JSM 2023, Toronto, Canada. Sanchez, J.

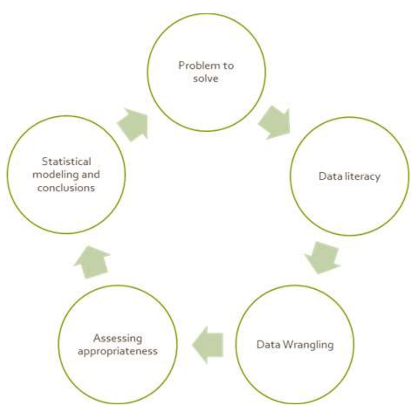
26



Conclusions

- In all the examples mentioned, everything involved one or more steps in the data science cycle, at the level appropriate for the moment and skill set of students, has been used.
- The examples involve a variety of data sets, and some very large data sets. In some we present the same data in very different ways, depending on our goals.
- But all the activities involve introductory statistics concepts in our classical curriculum for introductory stats, probability or time series.

27



Conclusions

- In all the examples mentioned, everything involved one or more steps in the data science cycle, at the level appropriate for the moment and skill set of students, has been used.
- The examples involve a variety of data sets, and some very large data sets. In some we present the same data in very different ways, depending on our goals.
- All the activities involve introductory statistics concepts in our classical curriculum for introductory stats, probability or time series, and yet are needed to do ML, AI, and

Thank you

28

My two favorite data literacy quotes involving the average. Certainly a conversation with students about what they mean for the use that is made of their social media data is important and only understood after gaining experience with the data science cycle.

"Let me assume that I am told that some cows ruminate. I can not infer logically from this that any particular cow does so, though I should feel some way removed from absolute disbelief, or even indifferent to assent, upon the subject; but if I saw a heard of cows I should feel more sure that some of them were ruminant than I did of the single cow, and my assurance would increase with the numbers of the herd about which I had to form an opinion. Here then we have a class of things as to the individuals of which we feel quite in uncertainty, whilst as we embrace larger numbers in our assertions we attach greater weight to our inferences. It is with such class of things and such inferences that the science of Probability is concerned." (Venn, 1888)

"Behavior modification, especially the modern kind implemented with gadgets like smartphones, is a statistical effect, meaning it's real but not comprehensively reliable; over a population, the effect is more or less predictable, but for each individual it's impossible to say." (Lanier 2018)

Bibliography

1. Arnold, P., Bargagliotti, A. Franklin, C. and Gould, R. (2022). *Bringing Complex Data into the Classroom.* HDSR. Issue 4.3. Summer 2022. https://doi.org/10.1162/99608f92.4ec90534
2. Peter K. Dunn (1999) *A Simple Dataset for Demonstrating Common Distributions, Journal of Statistics Education, 7:3*, DOI: 10.1080/10691898.1999.12131281
3. Fields, E. (2020). *Poisson Processes and Linear Programs for Allocating Shared Vehicles*. Notices of the American Statistical Association. Vol 67, No. 11, page 1804-1805.
4. Keller, S.A., Shipp, S.S., Schroeder, A.D. and Korkmaz, G. (2020). *Doing Data Science: A Framework and Case Study.* HDSR, Issue 2.1. Winter 2020. https://doi.org/10.1162/99608f92.2d83f7f5
5. Lanier, J. (2018). *Ten Arguments for Deleting your Social Media Accounts Right Now.* New York: Henry Holt and Company.
6. Piech, C. *Course Reader for CS 109.* Stanford University. Ongoing. https://chrispiech.github.io/probabilityForComputerScientists/en/examples/fairness/
7. Sanchez, J. (2023) *Time Series for Data Scientists*. Cambridge University Press.
8. Sanchez, J. (2023) timeseriestime.org A companion to Sanchez, J. Time Series Time for Data Scientists. https://timeseriestime.org/
9. Venn, J. (1888) The Logic of Chance. London. Macmillan and Co.