Data Exploration

Machine Learning for Behavioral Data

February 27, 2023

EPFL IC

Today's Topic

Week	Lecture/Lab
1	Introduction
2	**Data Exploration**
3	Regression
4	Classification
5	Model Evaluation
6	Time Series Prediction
7	Time Series Prediction
8	Time Series Prediction

Complete pipeline for one use case:
- Data exploration
- Prediction
- Model evaluation

Getting ready for today's lecture…

- **If not done yet**: clone the repository containing the Jupyter notebook and data for today's lecture into your Noto workspace..
- SpeakUp room for today's lecture:

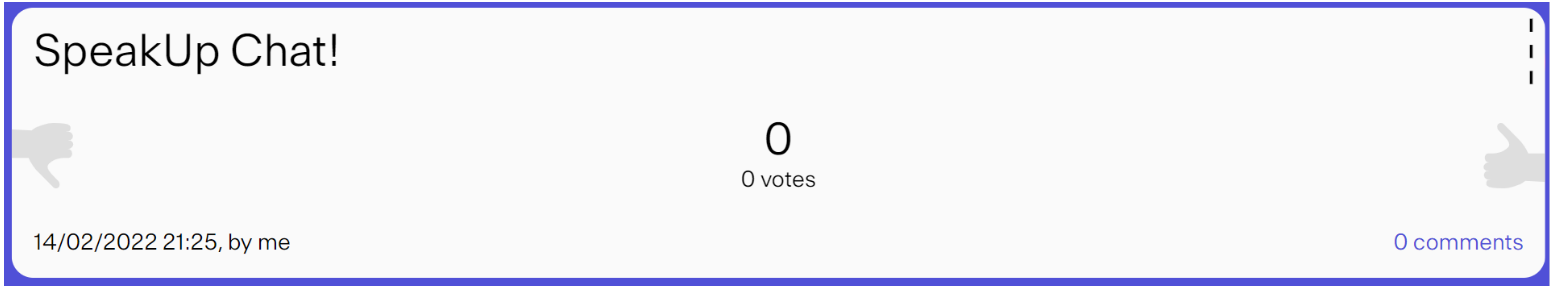
https://go.epfl.ch/speakup-mlbd

Noto: Student notebook

- Go to https://noto.epfl.ch/
- Login with your GASPAR
- Go to Git → Clone
- Clone the course repository: https://github.com/epfl-ml4ed/mlbd-2023

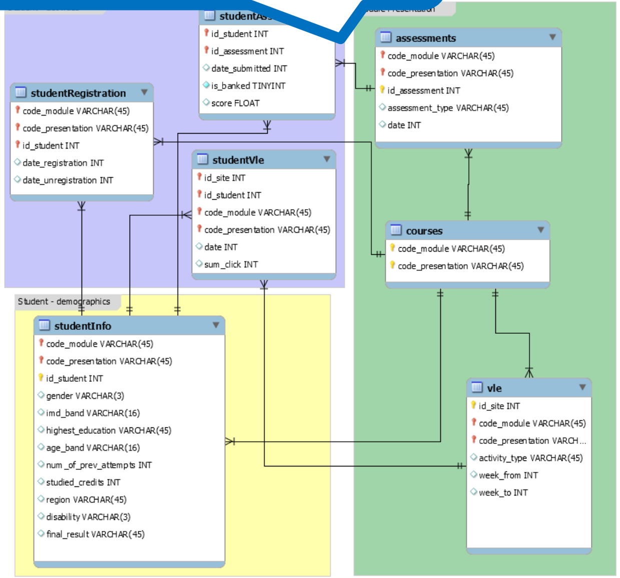
Why is data handling important?

- Why do we not just use the *raw data*?

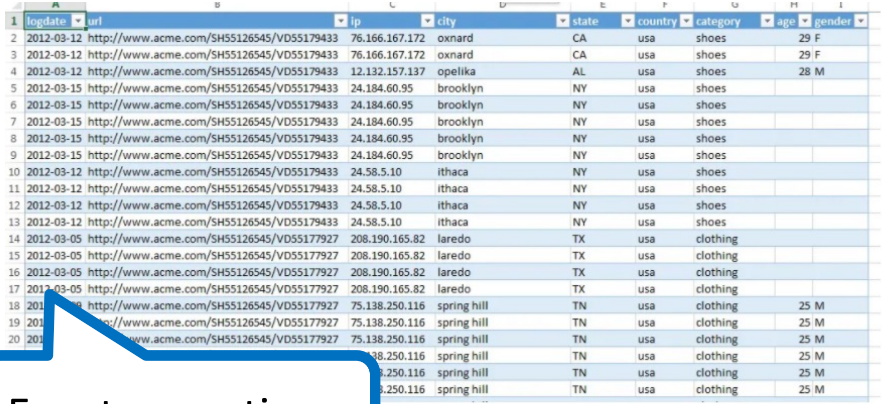


Different types of input data

Relational databases



Clickstream data



	logdate	url	ip	city	state	country	category	age	gender
2	2012-03-12	http://www.acme.com/SH55126545/VD55179433	76.166.167.172	oxnard	CA	usa	shoes	29	F
3	2012-03-12	http://www.acme.com/SH55126545/VD55179433	76.166.167.172	oxnard	CA	usa	shoes	29	F
4	2012-03-12	http://www.acme.com/SH55126545/VD55179433	12.132.157.137	opelika	AL	usa	shoes	28	M
5	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.184.60.95	brooklyn	NY	usa	shoes		
6	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.184.60.95	brooklyn	NY	usa	shoes		
7	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.184.60.95	brooklyn	NY	usa	shoes		
8	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.184.60.95	brooklyn	NY	usa	shoes		
9	2012-03-15	http://www.acme.com/SH55126545/VD55179433	24.184.60.95	brooklyn	NY	usa	shoes		
10	2012-03-12	http://www.acme.com/SH55126545/VD55179433	24.58.5.10	ithaca	NY	usa	shoes		
11	2012-03-12	http://www.acme.com/SH55126545/VD55179433	24.58.5.10	ithaca	NY	usa	shoes		
12	2012-03-12	http://www.acme.com/SH55126545/VD55179433	24.58.5.10	ithaca	NY	usa	shoes		
13	2012-03-12	http://www.acme.com/SH55126545/VD55179433	24.58.5.10	ithaca	NY	usa	shoes		
14	2012-03-05	http://www.acme.com/SH55126545/VD55177927	208.190.165.82	laredo	TX	usa	clothing		
15	2012-03-05	http://www.acme.com/SH55126545/VD55177927	208.190.165.82	laredo	TX	usa	clothing		
16	2012-03-05	http://www.acme.com/SH55126545/VD55177927	208.190.165.82	laredo	TX	usa	clothing		
17	2012-03-05	http://www.acme.com/SH55126545/VD55177927	208.190.165.82	laredo	TX	usa	clothing		
18	201	http://www.acme.com/SH55126545/VD55177927	75.138.250.116	spring hill	TN	usa	clothing	25	M
19	201	://www.acme.com/SH55126545/VD55177927	75.138.250.116	spring hill	TN	usa	clothing	25	M
20	201	ww.acme.com/SH55126545/VD55177927	75.138.250.116	spring hill	TN	usa	clothing	25	M
			38.250.116	spring hill	TN	usa	clothing	25	M
			.250.116	spring hill	TN	usa	clothing	25	M
			.250.116	spring hill	TN	usa	clothing	25	M

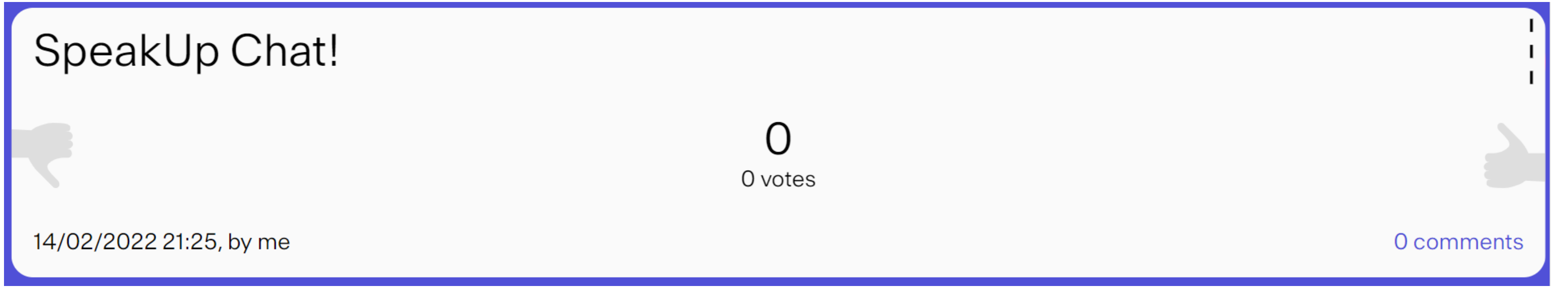
Events over time

Data Problems

- Incorrect data
- Duplicates
- Inconsistent data
- Missing data
- Outliers

Why is data handling important?

- What is the purpose of *data exploration*?



Today: Data Exploration

- **Univariate Analysis**
- Multivariate Analysis
- Time Series

Today's Use Case: Flipped Classroom Course

- Participants: 157 EPFL students of a course taught in *flipped classroom* mode with a duration of 10 weeks
- Structure:
  - Preparation: watch videos (and solve simple quizzes) on **new content** at home as a preparation for the lecture
  - Lecture: discuss open questions and solve more complex tasks
  - Lab session: solve paper-and-pen assignments
- Data: clickstream data (all interactions of the student with the system)

Today’s Use Case: The Data

Video_Info | *Video_Events*

TimeStamp	DataPackageID	UniqueRowID	TableName	VideoID	EventType	SessionUserID
1436539064	hwts-002	0000000773b50de2958e6128ca6a01dc	Video_Events	75	Video.Download	9e6622aa3440f144edb91a7d6397
1348761147	progfun-2012-001	00000013631cd1107b9781b40c37ac07	Video_Events	37	Video.Play	a7e07c5f41369e0acdf08ec72794b
1362266322	dsp-001	0000002363c3bd0f73b783e3adc44fb3	Video_Events	29	Video.Pause	bf85620e711cc570f95763d9768c
1430601717	reactive-002	00000059c6fb3e38eb5639e1b9e6c863	Video_Events	133	Video.Seek	ec35ab9103eb35ffcafc74f12c7e97
1372391638	progfun-002	00000078c0f0685cc50a25a8d5734a88	Video_Events	33	Video.Play	ef64fb7b096008f7eaf8441684afdf
1348627928	progfun-2012-001	000000d6a01b089ecee6aea3ddb4589c	Video_Events	33	Video.Seek	f12fbe6298a9e46122ed11cfabc43
1366535543	progfun-002	0000013af9c71ddea9e67332e9f2220f	Video_Events	39	Video.Load	8d7c72c0dfe78d0dbeb187c6c464
1361863559	dsp-001	00000146053bbf1daf5e74539b695ae6	Video_Events	43	Video.Play	c0b7417192e8b38e8f6cb641fc7bd
1350842274	progfun-2012-001	0000016e472deac18413b2a7ccdc2e07	Video_Events	97	Video.Seek	0c8efe11945ef0f1d0017707ba930
1400493317	progfun-004	0000017c871f54fda701333bd0acf7ba	Video_Events	77	Video.Play	2487d6899365bd5f704979f91995
1426880606	villesafricaines-003	0000017ea64ccec0f405090cfd220b51	Video_Events	47	Video.Load	b27704ef3090a0f666907807c1d8
1417881517	intropoojava-001	0000019fa8f938d69cc019e7805edcba	Video_Events	67	Video.Pause	8ae201009a69aa6ee8c0ae790927
1395399921	java-fr-2013-001	000001cb3ef0ccf281d3b9f1c00e7d60	Video_Events	13	Video.Stalled	817fc9f1ede5e69d36641c8b2d937
1400786471	microcontroleurs-003	000001d606e9a4bea4544c1827275b89	Video_Events	19	Video.Pause	6c06a76c20df00c17f1d83e7c1832

Characteristics of a Variable/Feature

ID	Grade	Gender	Category	# Sessions	Time in videos	Time in problems	# clicks on weekdays	# clicks on weekends	Content alignment	Mean pause duration	Mean playback speed	# problem sub-missions	# correct sub-missions
1	4.5	M	Suisse. Autres	57	9227	1698	179	4	0.75	50	1.1	9	5.9
2	5.25	M	Suisse. Autres	41	10801	2340	129	95	0.35	231	0.8	6.1	3
3	4.5	F	Suisse. PAM	33	8185	2737	46	14	0.37	92	0.5	4.6	3.2
4	4.75	F	France	47	7040	3787		58	0.03	62	0.85	0.3	0.1

- Center of the data?
- Spread of the data?
- Shape/distribution of the data?

Descriptive Statistics

	Mean	Median	Mode	Variance	Std	Minimum	25%	75%	Maximum
grade	4.05	4.25	5.0	1.49e+00	1.22	1.00	3.25	5.00	6.00
sessions	33.89	34.00	36.0	2.38e+02	15.42	6.00	22.00	43.00	97.00
time_in_problem	28022.04	24209.50	0.0	4.83e+08	21980.95	0.00	10029.00	41756.75	111238.00
time_in_video	82851.62	81735.50	26699.0	2.20e+09	46942.02	0.00	48823.25	111431.25	274917.00
lecture_delay	820.27	0.00	0.0	1.85e+09	43010.20	-159250.48	-22921.90	24249.25	144964.21
content_anticipation	0.11	0.09	0.0	1.02e-02	0.10	0.00	0.01	0.20	0.31
mean_playback_speed	0.94	0.92	0.9	9.37e-02	0.31	0.00	0.80	1.11	1.76
relative_video_pause	0.22	0.23	0.0	1.05e-02	0.10	0.00	0.14	0.30	0.43
submissions	46.05	35.50	0.0	1.77e+03	42.12	0.00	9.75	77.00	171.00
submissions_correct	25.01	18.00	0.0	5.24e+02	22.90	0.00	4.75	41.00	89.00
clicks_weekend	679.80	465.00	0.0	4.93e+05	702.04	0.00	160.50	1012.75	4546.00
clicks_weekday	1130.64	930.50	108.0	8.13e+05	901.44	0.00	495.00	1534.00	6223.00

Center of the data

Spread of the data

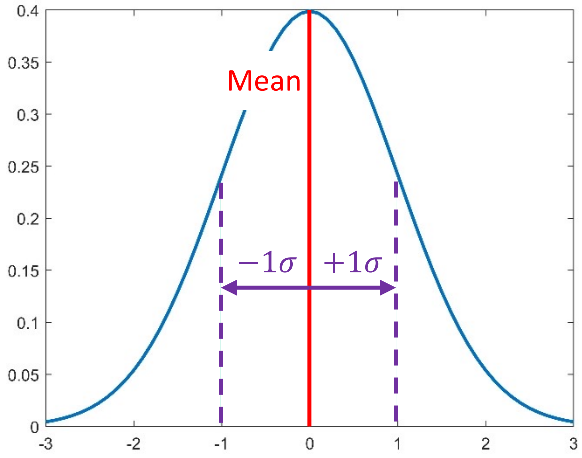
Example: Normal Distribution

- Sample mean: $\mu_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$

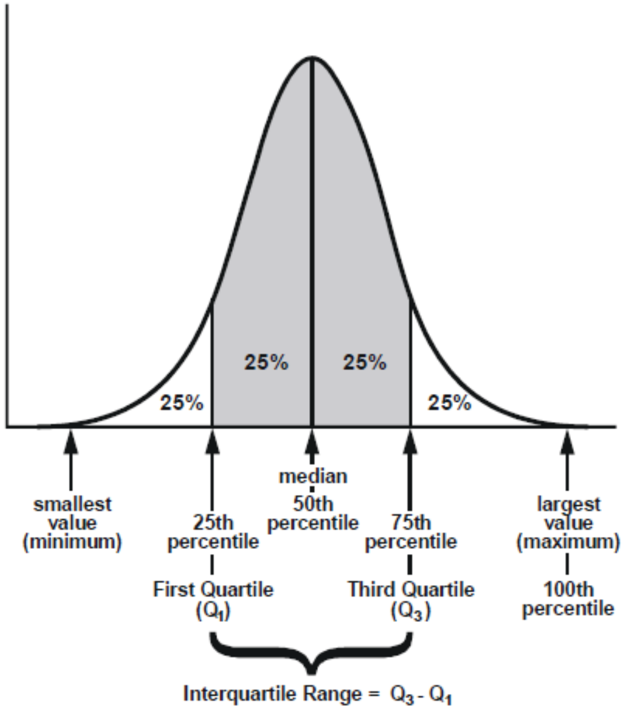
- Sample variance: $\sigma_{\bar{x}}{}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_{\bar{x}})^2$

- Mode: most frequent value in data set

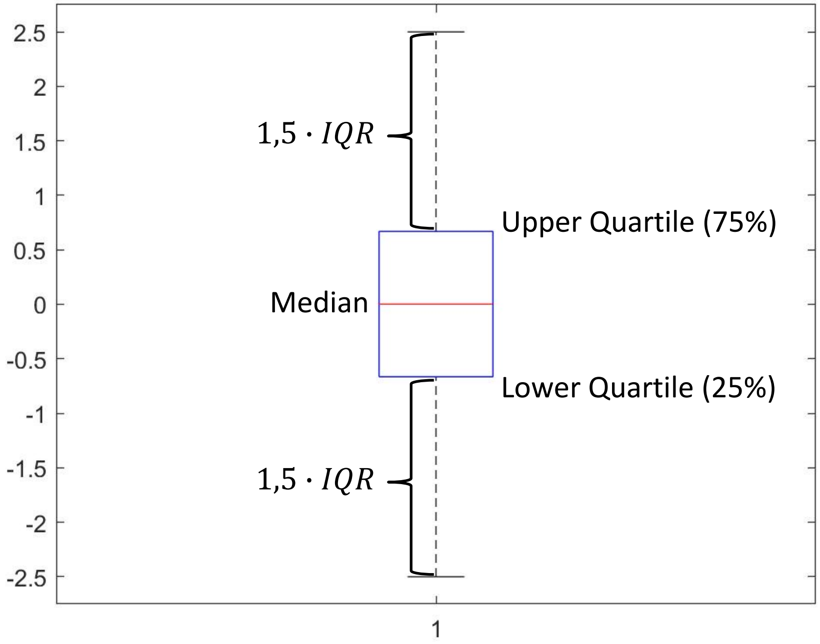
- Median: separates the lower and upper half of the data (1, 2, 2, **3**, 4, 7, 9)



Example: Normal Distribution



Boxplot



2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
1
$1{,}5 \cdot IQR$
Upper Quartile (75%)
Median
Lower Quartile (25%)
$1{,}5 \cdot IQR$

Descriptive Statistics

	Mean	Median	Mode	Variance	Std	Minimum	25%	75%	Maximum
grade	4.05	4.25	5.0	1.49e+00	1.22	1.00	3.25	5.00	6.00
sessions	33.89	34.00	36.0	2.38e+02	15.42	6.00	22.00	43.00	97.00
time_in_problem	28022.04	24209.50	0.0	4.83e+08	21980.95	0.00	10029.00	41756.75	111238.00
time_in_video	82851.62	81735.50	26699.0	2.20e+09	46942.02	0.00	48823.25	111431.25	274917.00
lecture_delay	820.27	0.00	0.0	1.85e+09	43010.20	-159250.48	-22921.90	24249.25	144964.21
content_anticipation	0.11	0.09	0.0	1.02e-02	0.10	0.00	0.01	0.20	0.31
mean_playback_speed	0.94	0.92	0.9	9.37e-02	0.31	0.00	0.80	1.11	1.76
relative_video_pause	0.22	0.23	0.0	1.05e-02	0.10	0.00	0.14	0.30	0.43
submissions	46.05	35.50	0.0	1.77e+03	42.12	0.00	9.75	77.00	171.00
submissions_correct	25.01	18.00	0.0	5.24e+02	22.90	0.00	4.75	41.00	89.00
clicks_weekend	679.80	465.00	0.0	4.93e+05	702.04	0.00	160.50	1012.75	4546.00
clicks_weekday	1130.64	930.50	108.0	8.13e+05	901.44	0.00	495.00	1534.00	6223.00

Variable Types

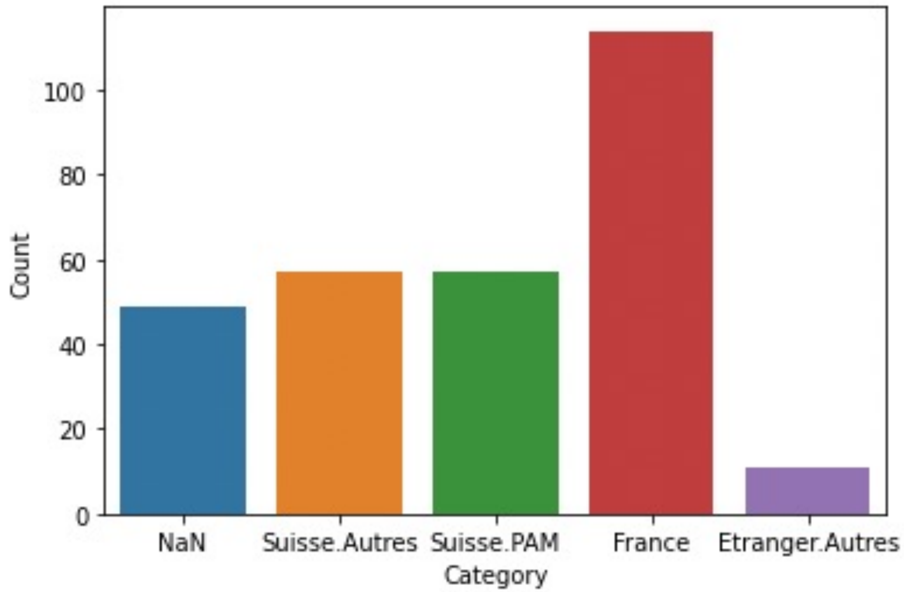
- Categorical
- Ordinal
- Numerical

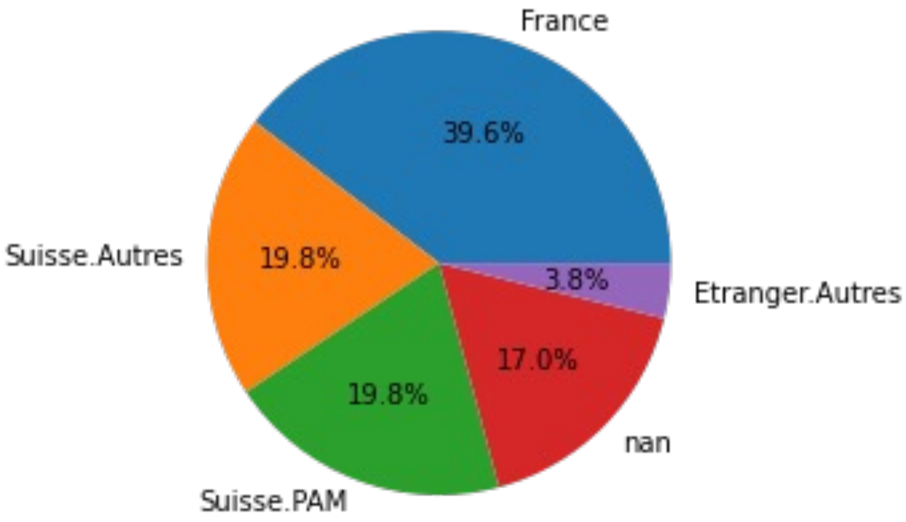
Categorical Variables

Category	Count	Count %
France	114	0.40
Suisse.Autres	57	0.20
Suisse.PAM	57	0.20
NaN	49	0.17
Etranger.Autres	11	0.04

Gender	Count	Count %
M	156	0.54
F	83	0.29
NaN	49	0.17

Number of students per category





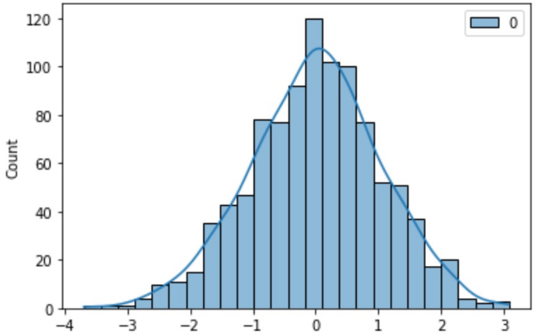
https://seaborn.pydata.org/tutorial/categorical.html

Characteristics of a Variable/Feature

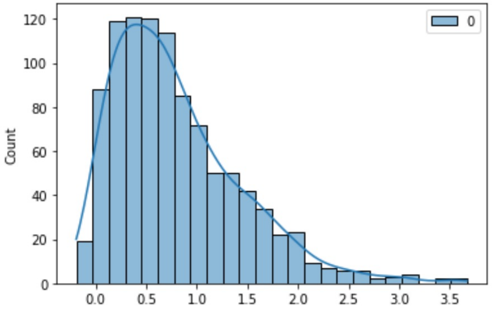
ID	Grade	Gender	Category	# Sessions	Time in videos	Time in problems	# clicks on weekdays	# clicks on weekends	Content alignment	Mean pause duration	Mean playback speed	# problem sub-missions	# correct sub-missions
1	4.5	M	Suisse. Autres	57	9227	1698	179	4	0.75	50	1.1	9	5.9
2	5.25	M	Suisse. Autres	41	10801	2340	129	95	0.35	231	0.8	6.1	3
3	4.5	F	Suisse. PAM	33	8185	2737	46	14	0.37	92	0.5	4.6	3.2
4	4.75	F	France	47	7040	3787		58	0.03	62	0.85	0.3	0.1

- Center of the data?
- Spread of the data?
- Shape/distribution of the data?

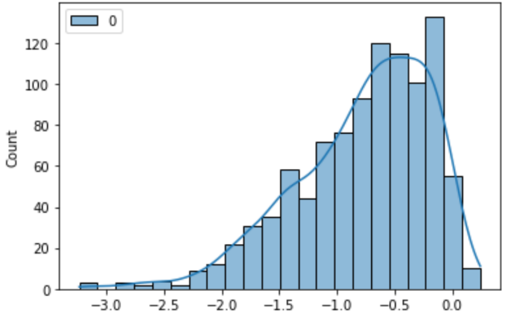
Does my data follow a normal distribution?



symmetric



right skewed



left skewed

Normal test p = 0.39

Normal test p = 8.7e-43

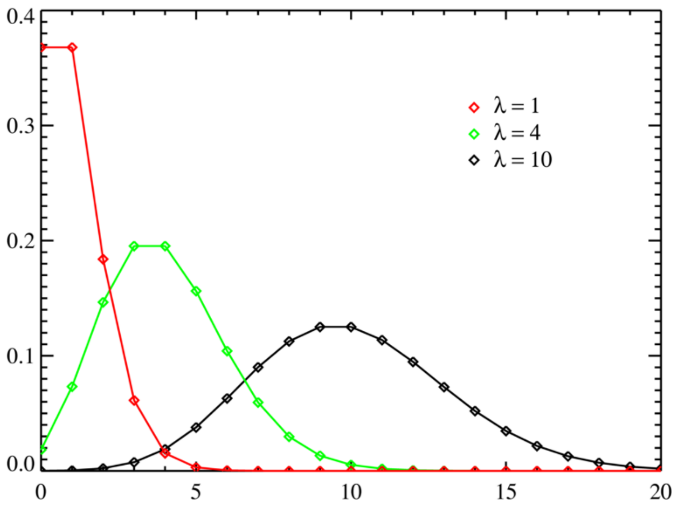
Normal test p = 6.0e-26

Important Distributions

- **Normal distribution :** *(continuous)* see previous slides
- **Poisson distribution:** *(discrete)* expresses the probability of a given number of events occurring in a fixed interval of time or space
- **Exponential distribution** *(continuous)* distribution of times between events in a Poisson process
- **Binomial distribution**: *(discrete)* models the number of successes in a sequence of independent experiments
- **Bernoulli distribution**: *(discrete)* special case of binomial distribution (n=1)

Important Distributions | Poisson

Models the number of events occurring within a given time interval.



Properties:

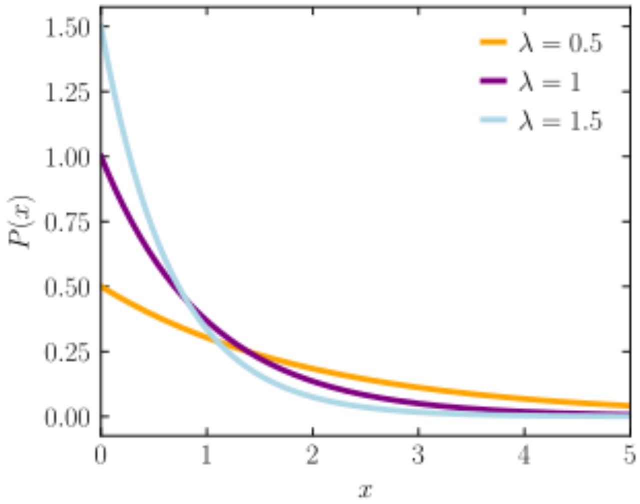
- Discrete (not continuous)
- Greater or equal to zero.

Examples:

- Number of calls a call center receives per minute
- Number of students that join the zoom meeting per minute during the first 15 minutes of the class

Important Distributions | Exponential

Probability distribution of time between events of a **Poisson** process.



Properties:

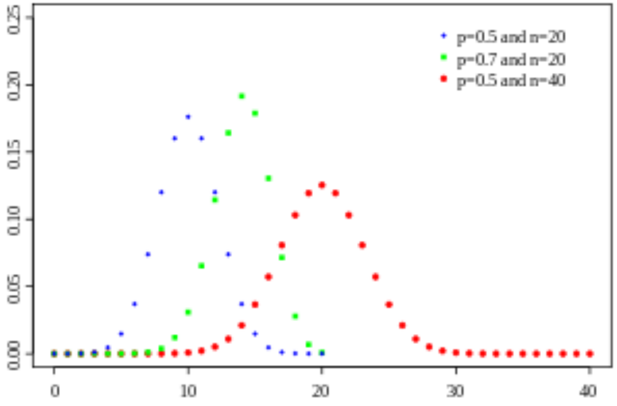
- Continuous
- Greater or equal to zero.

Examples:

- The time before the next telephone call in a call center.
- The time before the next student joins the zoom call.

Important Distributions | Binomial

Models the number of successes in a sequence of independent experiments.



Properties:

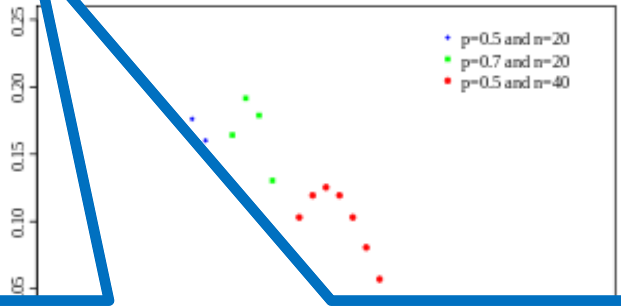
- Discrete (not continuous)
- Greater or equal to zero.

Examples:

- Number of passed tests in a course with 20 tests.
- Number of customers that redeemed a coupon.

Important Distributions | Binomial

Models the number of successes in a sequence of independent experiments.



Bernoulli is a special case of the Binomial distribution with one experiment: n = 1

Properties:

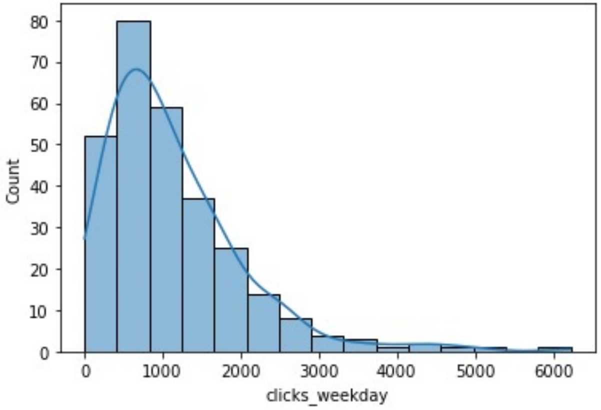
- Discrete (not continuous)
- Greater or equal to zero.

Examples:

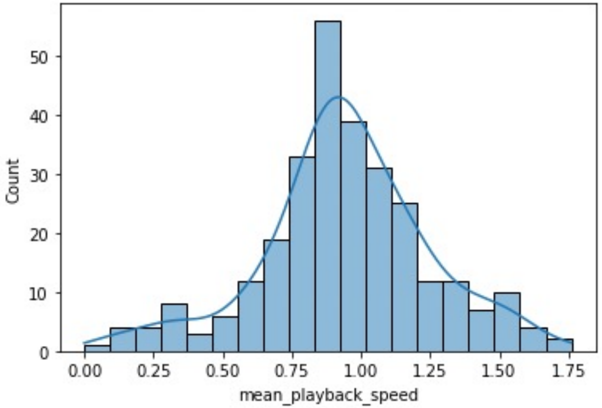
- Number of passed tests in a course with 20 tests.
- Number of customers that redeemed a coupon.

Visual Inspection

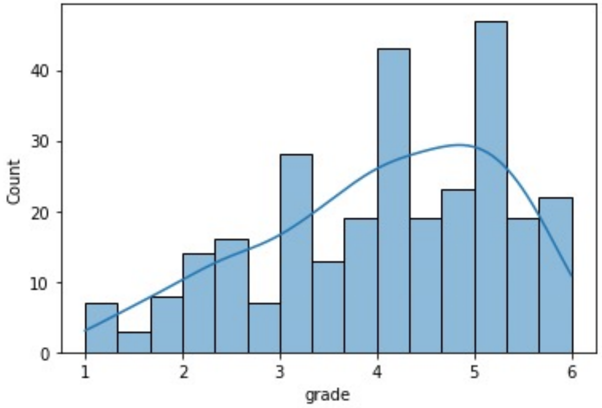
```
p = 6.20579e-29
The null hypothesis can be rejected
```




```
p = 0.0216998
The null hypothesis cannot be rejected
```




```
p = 5.78191e-05
The null hypothesis can be rejected
```

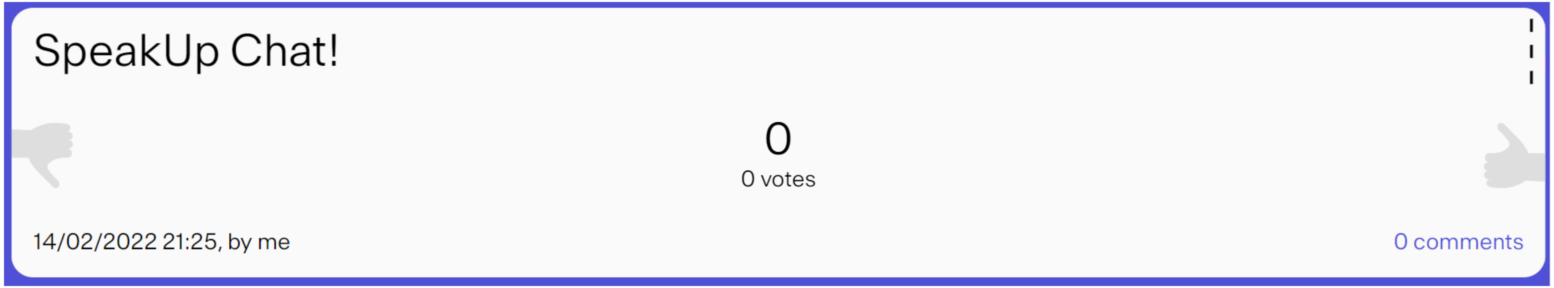


Data Exploration

- Univariate Analysis
- **Multivariate Analysis**
- Time Series

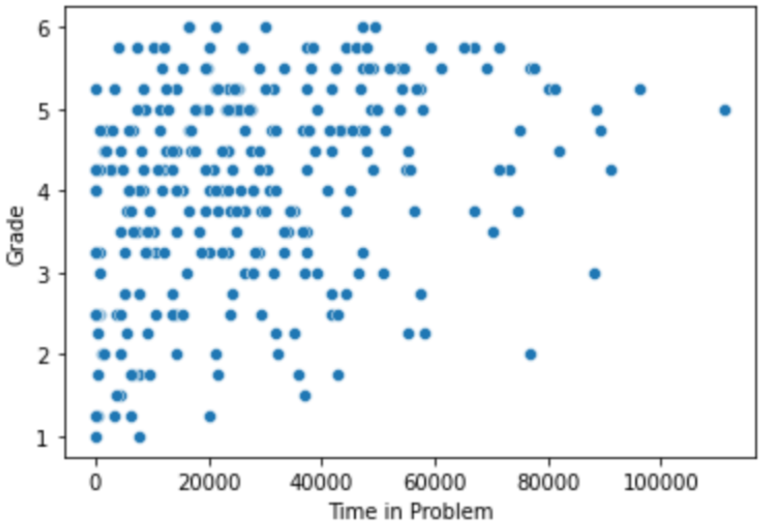
Multivariate Analysis

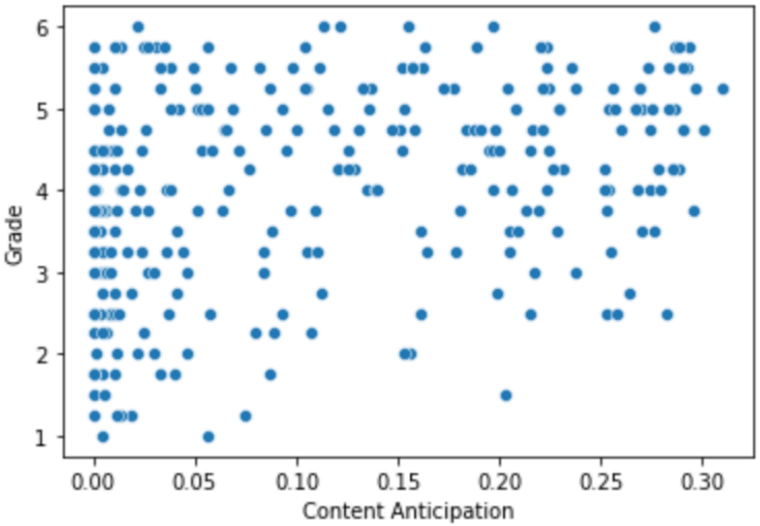
How can we explore the relationship between two variables?



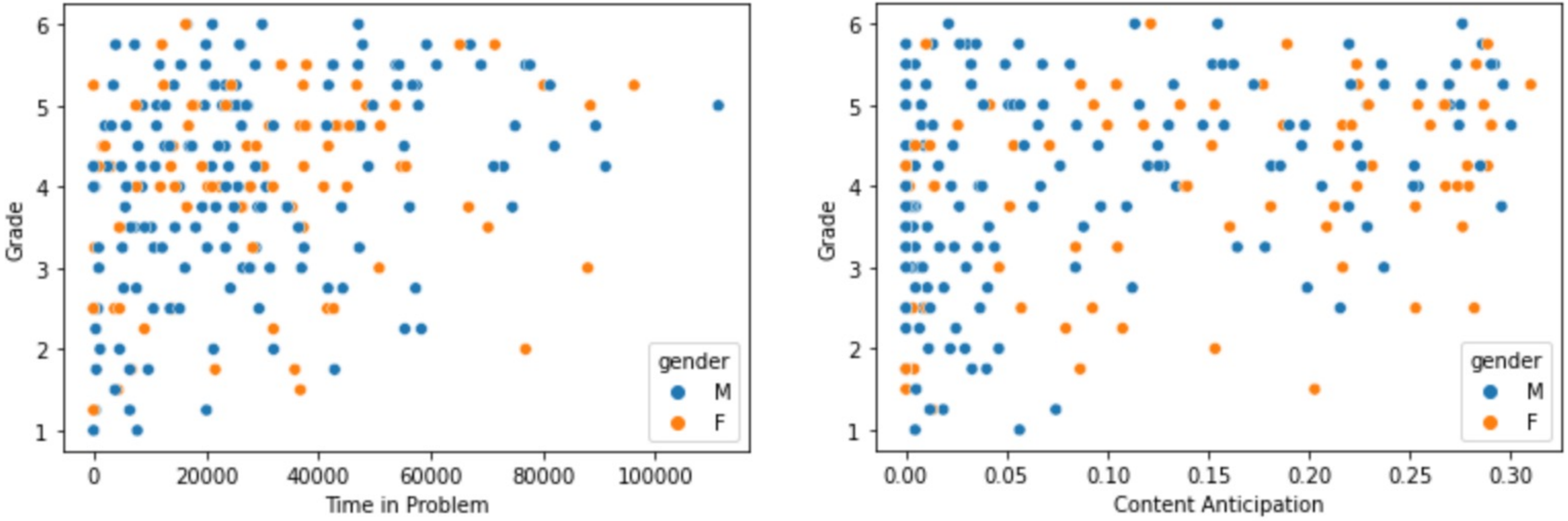
https://go.epfl.ch/speakup-mlbd

Relation between numerical variables

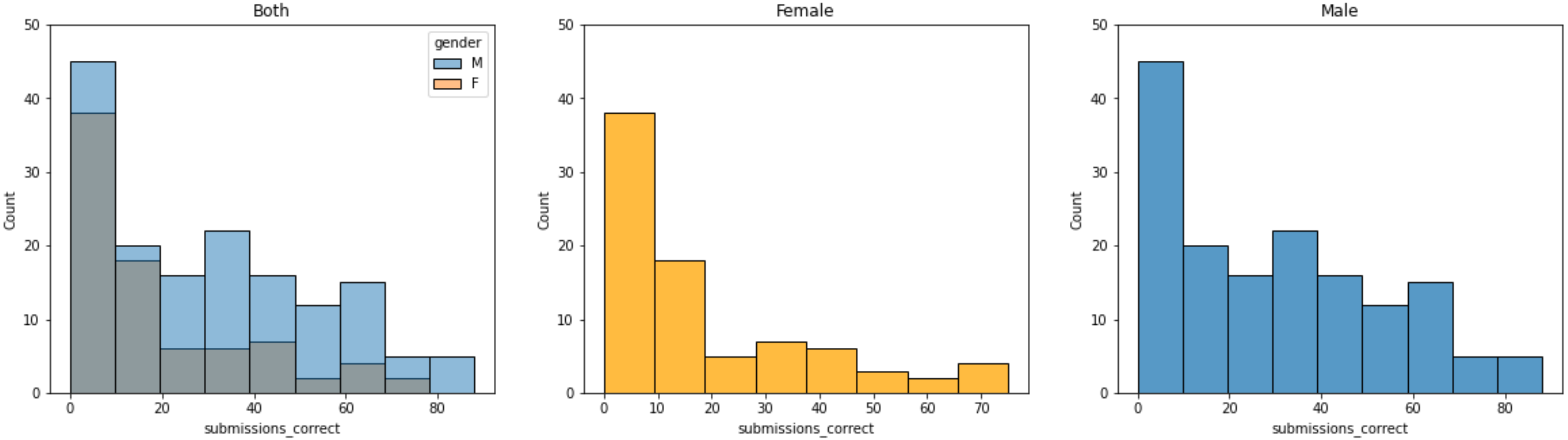




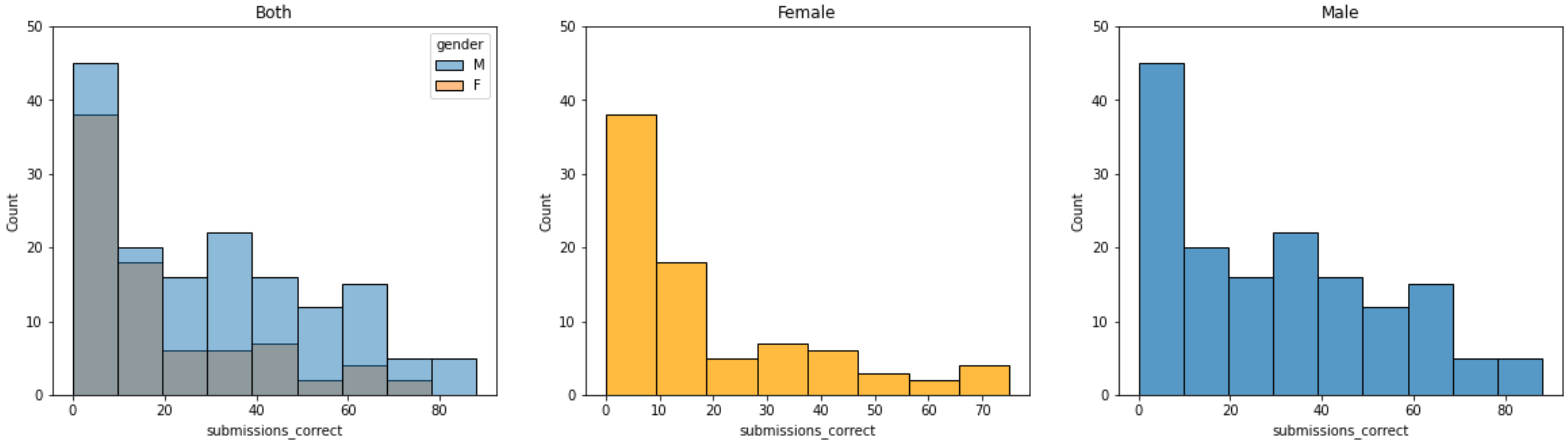
Relation between numerical & categorical variables



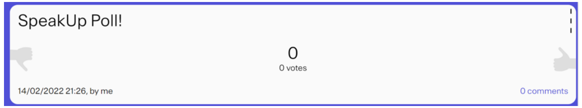
Submissions Correct by Gender



Who is more likely to have correct submissions?



a) Students identifying as male are more likely to have a correct submission.
b) Students identifying as female are more likely to have a correct submission.
c) I cannot answer based on the visualization.



https://go.epfl.ch/speakup-mlbd

Pearson's Correlation

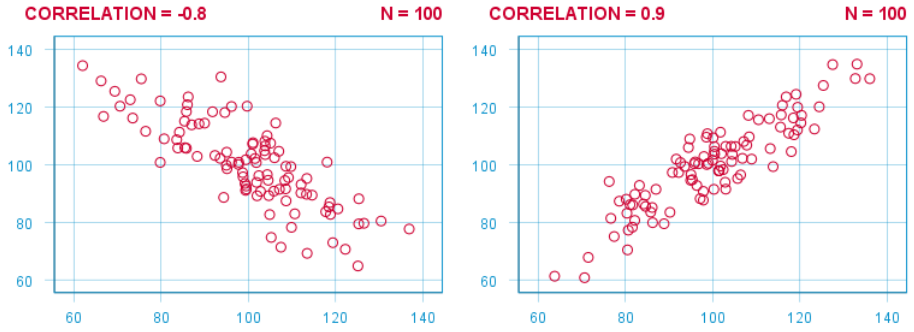
Linear correlation between two sets of data.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Where $cov(X,Y)$ is the covariance

σ_X is the standard deviation on X

σ_Y is the standard deviation on Y



Pearson's Correlation

Linear correlation between two sets of data.

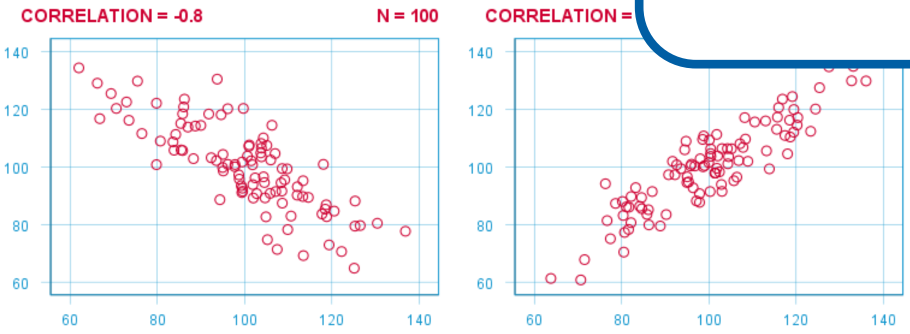
$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Where cov

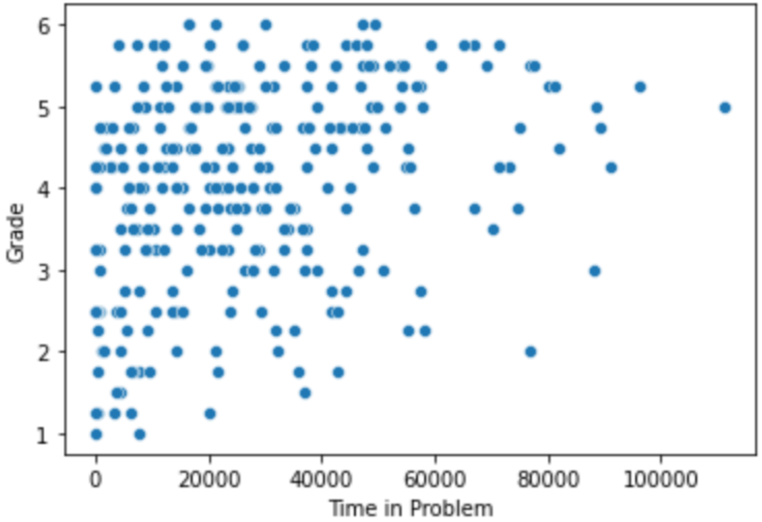
σ_X is the st

σ_Y is the st

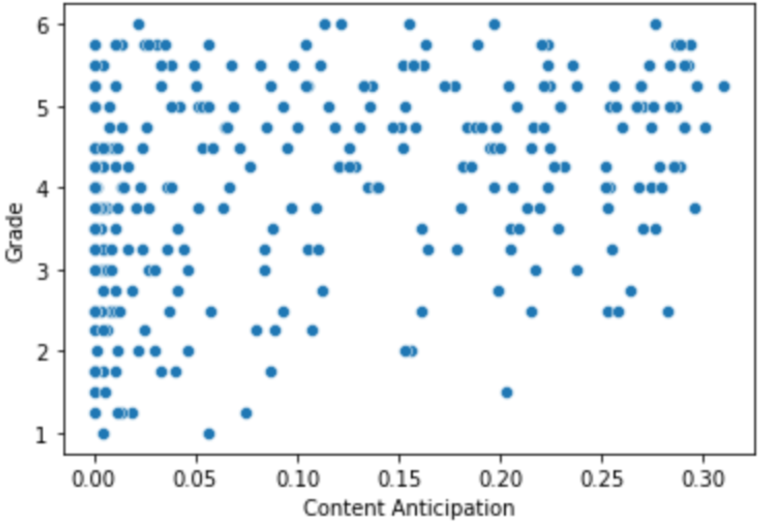
X and Y need to be numerical or at least ordinal variables



Correlation between variables



$$\rho = 0.31 \ (p = 6.8e - 8)$$



$$\rho = 0.32 \ (p = 1.5e - 08)$$

Pearson's Correlation

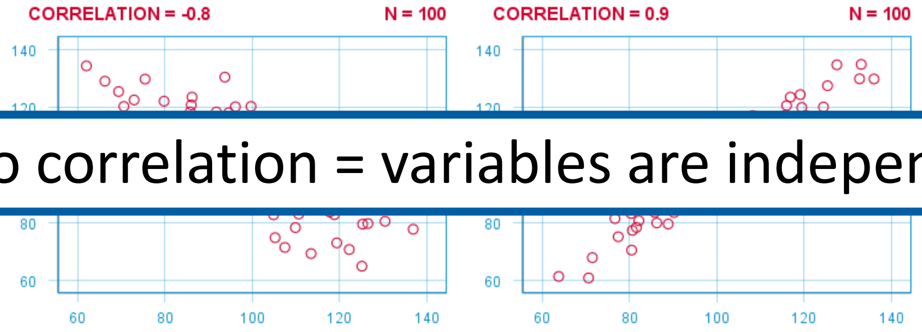
Linear correlation between two sets of data.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Where $cov(X,Y)$ is the covariance

σ_X is the standard deviation on X

σ_Y is the standard deviation on Y



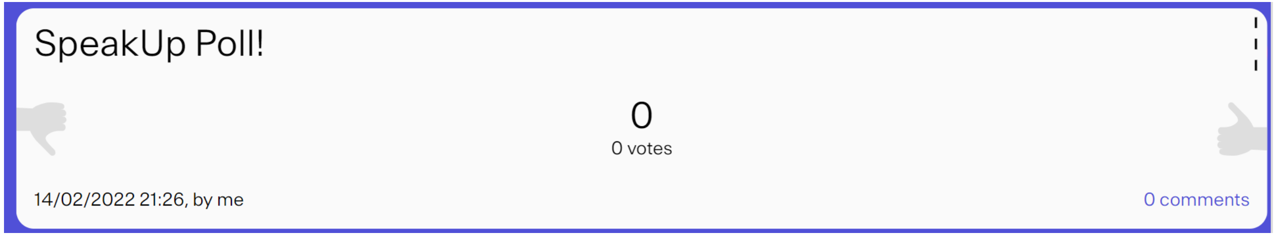
No correlation = variables are independent?

Pearson's Correlation

Linear correlation between two sets of data.

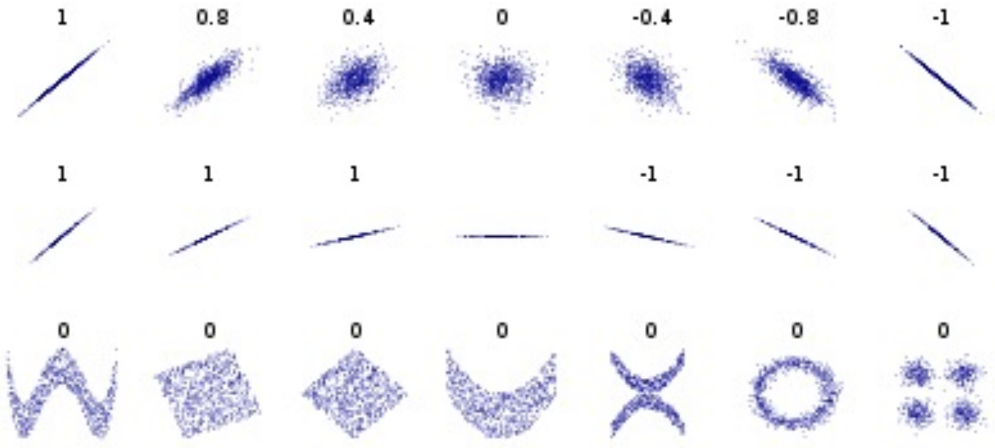
No correlation = variables are independent?

a) Yes
b) No



https://go.epfl.ch/speakup-mlbd

Pearson's Correlation



$$X, Y \text{ independent } \rightarrow \rho_{X,Y} = 0$$
$$\rho_{X,Y} = 0 \nrightarrow X, Y \text{ independent}$$

Image Source: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Mutual Information

- Dependence between two random variables: “Amount of information” obtained about one random variable through observing the other random variable

$$I(X;Y) = D_{KL}(P_{(X,Y)} || P_X \otimes P_Y)$$

where X and Y are random variables, $P_{(X,Y)}$ is their joint distribution, P_X and P_Y are the marginal distributions, and D_{KL} is the Kullback-Leibler divergence.

Mutual Information

- Dependence between two random variables: “Amount of information” obtained about one random variable through observing the other random variable

$$I(X;Y) = \ D_{KL}(P_{(X,Y)}||P_X \otimes P_Y)$$

where X and Y are random variables, $P_{(X,Y)}$ is their joint distribution, P_X and P_Y are the marginal distributions, and D_{KL} is the Kullback-Leibler divergence.

- For discrete distributions

$$I(X;Y) = \sum_{x \in \mathrm{X}} \sum_{y \in \mathrm{Y}} p(x,y) \cdot \log(\frac{p(x,y)}{p(x) \cdot p(y)})$$

Mutual Information - Motivation

- For discrete distributions

$$I(X;Y) = \sum_{x \in \mathrm{X}} \sum_{y \in \Upsilon} p(x,y) \cdot \log\left(\frac{p(x,y)}{p(x) \cdot p(y)}\right)$$

- If X and Y are *independent,* then $p(x,y) = p(x) \cdot p(y)$ and therefore:

$$\log\left(\frac{p(x,y)}{p(x) \cdot p(y)}\right) = \log(1) = 0$$

Pearson Correlation vs Mutual Information

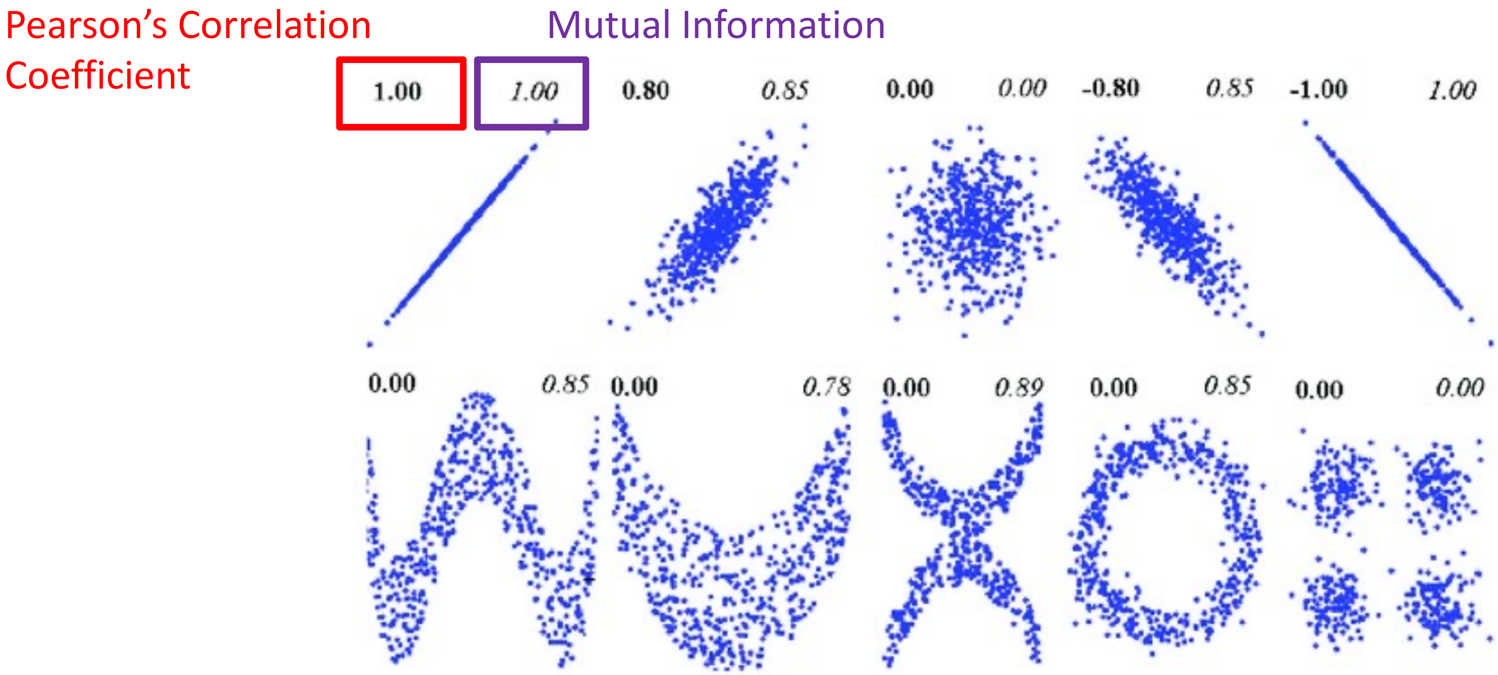
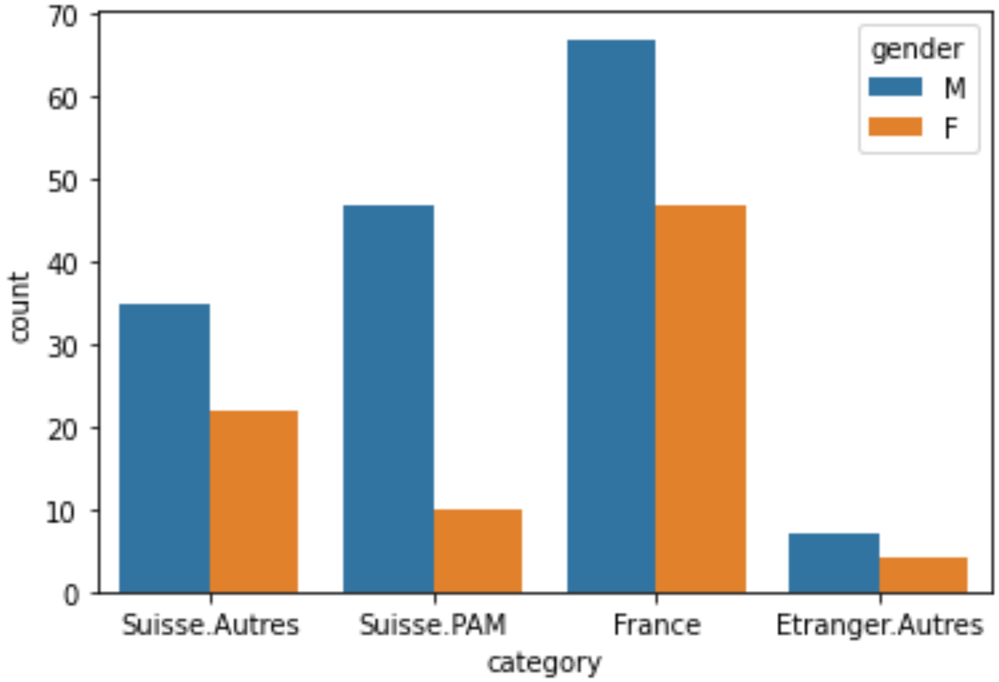


Image Source: https://www.researchgate.net/publication/322349267_Information_Entropy_Suggests_Stronger_Nonlinear_Associations_between_Hydro-Meteorological_Variables_and_ENSO

Mutual Information – Discrete



Mutual Information - Discrete

$P(X,Y)$

Y: Category

X: Gender

	France	Suisse.PAM	Suisse. Autres	Etranger.Autres
Male	0.28	0.20	0.15	0.02
Female	0.20	0.04	0.09	0.02

Mutual Information - Discrete

$P(X,Y)$

Y: Category

X: Gender

	France	Suisse.PAM	Suisse. Autres	Etranger.Autres
Male	0.28	0.20	0.15	0.02
Female	0.20	0.04	0.09	0.02

$P(Y)$

France	Suisse.PAM	Suisse. Autres	Etranger.Autres
0.48	0.24	0.24	0.04

$P(X)$

Female	Male
0.35	0.65

Mutual Information - Discrete

$P(X,Y)$

Y: Category

X: Gender

	France	Suisse.PAM	Suisse. Autres	Etranger.Autres
Male	0.28	0.20	0.15	0.02
Female	0.20	0.04	0.09	0.02

$P(Y)$

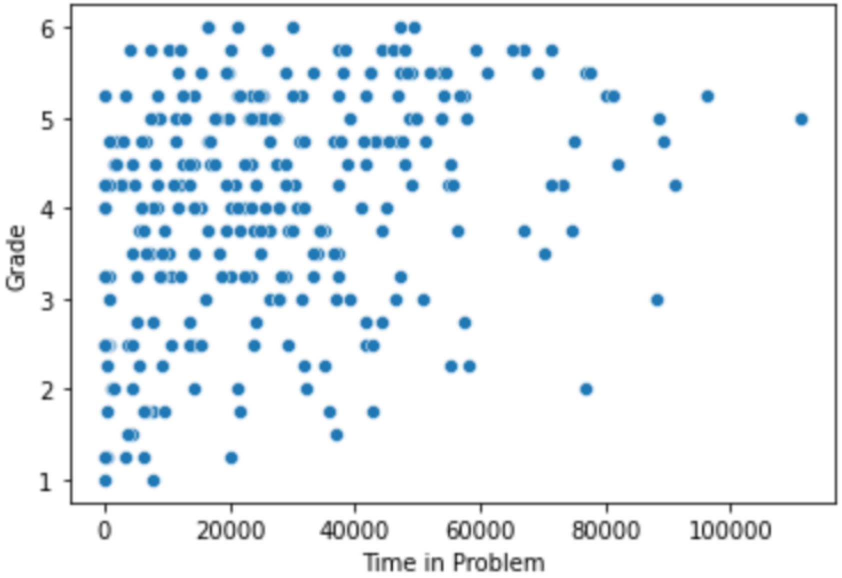
France	Suisse.PAM	Suisse. Autres	Etranger.Autres
0.48	0.24	0.24	0.04

$P(X)$

Female	Male
0.35	0.65

$$\boxed{I(X;Y) = 0.02}$$

Mutual Information - Continuos



$$I(X;Y) = 0.12$$

$$\rho = 0.31\ (p = 6.8e - 8)$$

Data Exploration

- Univariate Analysis
- Multivariate Analysis
- **Time Series**

Time Series Data

Records, which are measured sequentially over time:

- **Business**: sales figures, production numbers, customer frequencies, ...
- **Economics**: stock prices, exchange rates, interest rates, ...
- **Official Statistics**: census data, personal expenditures, road casualties, ...
- **Natural Sciences**: population sizes, sunspot activity, chemical process data, ...
- **Environmetrics**: precipitation, temperature or pollution recordings, ...

Time Series – Behavioral Data

Records of user behavior, which are measured sequentially over time:

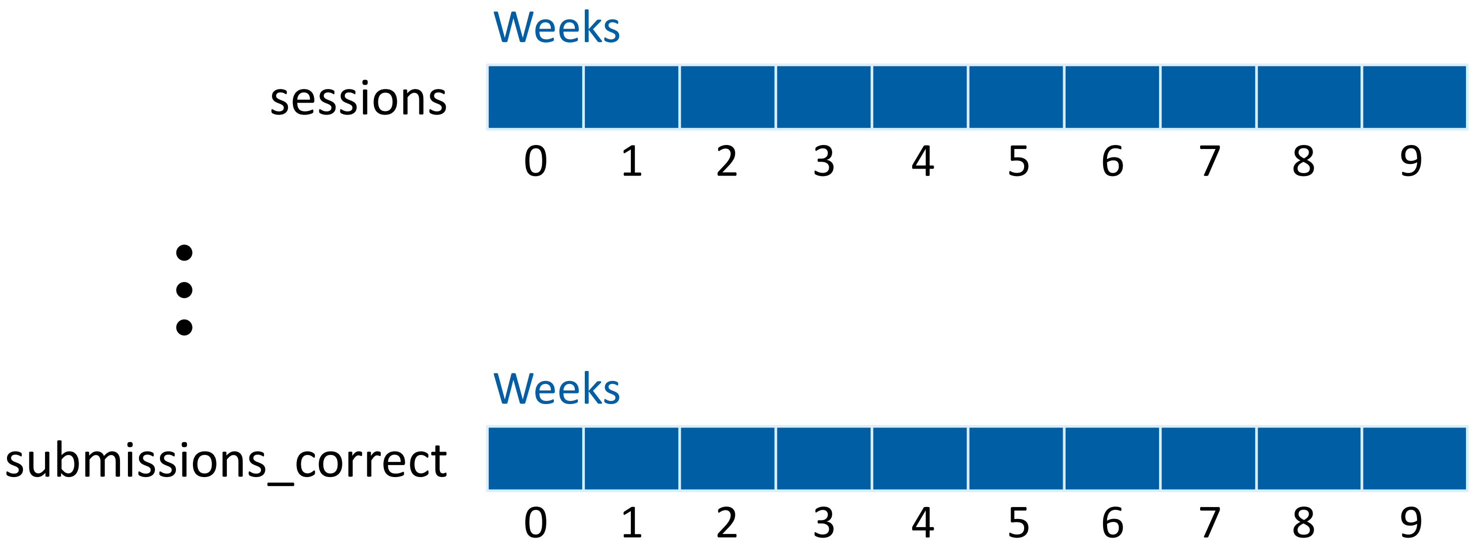
- we usually deal with multiple time series (i.e. one time series per user u)
- a record $r_{u,t}$ of a user u at time t can consists of multiple variables

We might be interested in representing, analyzing, and predicting behavior of single users or of group of users:

- Visualization and exploration of time series data (this lecture)
- Modeling time series data (later…)

Time Series – Our flipped classroom case

Student n

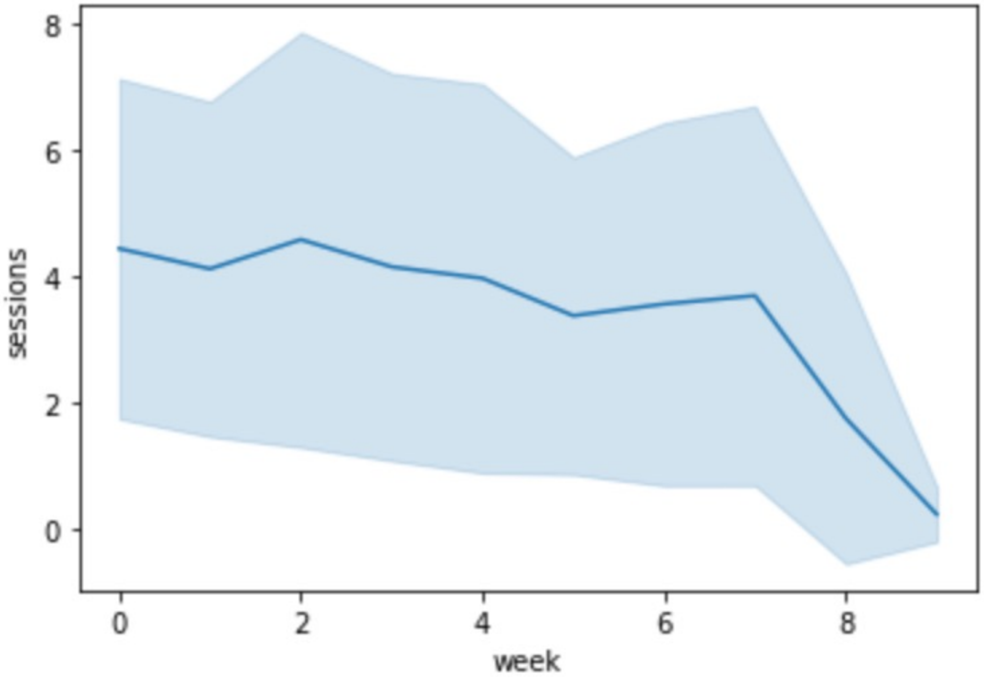


Hypothesis 1

The number of sessions will decrease over the course of the semester.

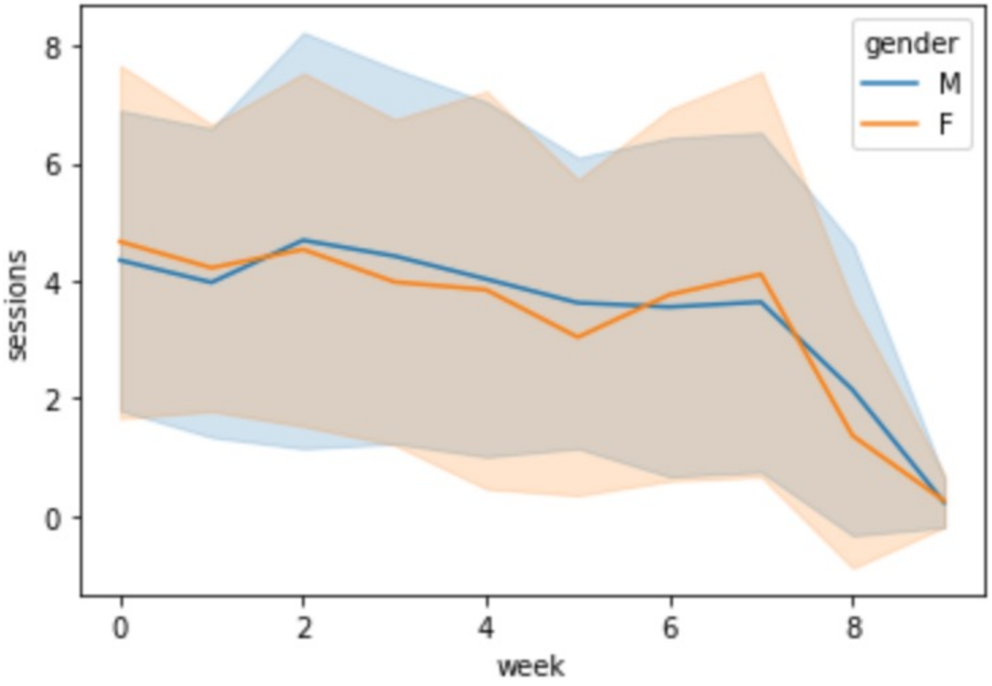
Hypothesis 1

The number of sessions will decrease over the course of the semester.



Hypothesis 2

There is no difference between males and females in terms of the number of sessions.



Your turn!

- Come up with a hypothesis on your own
- Produce a visualization
- Describe: what do you observe? Can your hypothesis be confirmed?

Your turn!

- Come up with a hypothesis on your own
- Produce a visualization
- Describe: what do you observe? Can your hypothesis be confirmed?

Do you want feedback or have questions?
(Optional) Upload your Jupyter Notebook here:
https://go.epfl.ch/notebooks-mlbd

Summary

- Compute descriptive statistics
- Visualize, visualize, visualize,…
  → Different types of visualizations or representations help to identify different types of problems
  → Different types of visualizations help to identify different patterns/properties in the data
- Try to gain as much knowledge as possible about the domain and the data collection