INTRODUCTION TO STATISTICS

LECTURE 1

ABOUT ME

- EVGENIYA Korneva  evgeniakorneva@gmail.com

ABOUT ME

- EVGENIYA Korneva  evgeniakorneva@gmail.com

- Education:
  - *2015* - Bachelor of Applied Mathematics *(Moscow, Russia)*
  - *2016* - Master of Artificial Intelligence *(Leuven, Belgium)*

NATIONAL RESEARCH UNIVERSITY

KU LEUVEN

ABOUT ME

- EVGENIYA Korneva  evgeniakorneva@gmail.com
- Education:
  - *2015* - Bachelor of Applied Mathematics *(Moscow, Russia)*
  - *2016* - Master of Artificial Intelligence *(Leuven, Belgium)*
- *2016* - … Doctoral researcher at KU Leuven

ABOUT ME

- EVGENIYA Korneva evgeniakorneva@gmail.com
- Education:
  - *2015* - Bachelor of Applied Mathematics *(Moscow, Russia)*
  - *2016* - Master of Artificial Intelligence *(Leuven, Belgium)*
- *2016* - … Doctoral researcher at KU Leuven

NATIONAL RESEARCH UNIVERSITY

KU LEUVEN

DTAI
Declarative Languages & Artificial Intelligence

Prague, Czech Republic

ABOUT THE COURSE

- Class: 09:00 – 12.20
  - *two 10-minute break*

- Office hours: 13.00 – 14.00
  - **Want to talk?**
    Send me an email first!

- Materials will be posted after class.

ABOUT THE COURSE

- Class: 09:00 – 12.20
  - *two 10-minute break*
- Office hours: 13.00 – 14.00
  - **Want to talk? Send me an email first!**
- Materials will be posted after class.
- Final grade:
  - 5 graded assignments *(20 pts each)*
  - mid-term exam *(? December 11 or 12)*
  - final exam *(? December 17 or 18)*
- First graded assignment out today. Deadline: this Thursday.
- **Please complete the ENTRY TEST by tomorrow evening!**

LET’S STRAT!

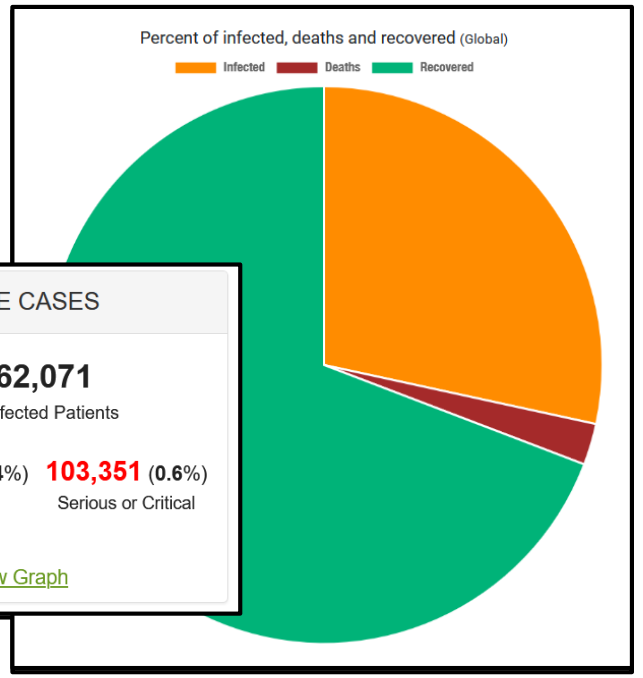
WHAT IS STATISTICS?

- How do *you* imagine Statistics?

WHAT IS STATISTICS?

- How do *you* imagine Statistics?
  - Can you think of any example of Statistics?

WE'RE ALL FOLLOWING THESE STATISTICS...



ACTIVE CASES

16,862,071
Currently Infected Patients

16,758,720 (99.4%) in Mild Condition

103,351 (0.6%) Serious or Critical

Show Graph

Coronavirus Statistics

Global

59229918 Infected

1397583 2.4% Deaths

40976963 69.2% Recovered

Last update: 23.11.20 17:03 GMT

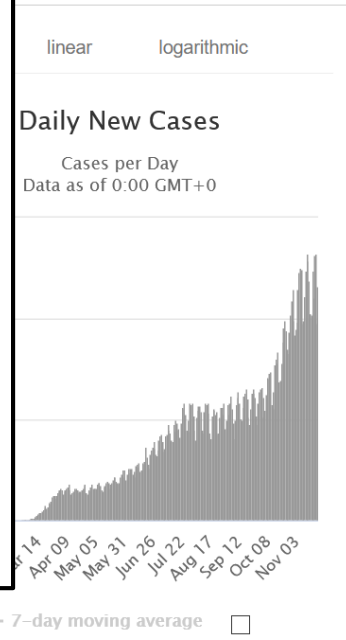
Show 10 countries

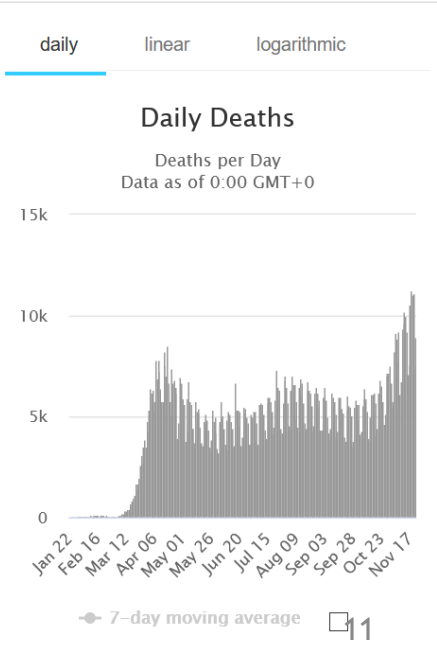
Search:

Country	Infected	Deaths	Recovered	Deaths Percent	Recovered Percent	Infected per million	Deaths per million
USA	12603019	262781	7453682	2.1%	59.1%	37987	792
India	9170825	134088	8592303	1.5%	93.7%	6620	97
Brazil	6073058	169213	5432505	2.8%	89.5%	28491	794
France	2140208	48732	149521	2.3%	7.0%	32760	746
Russia	2114502	36540	1611445	1.7%	76.2%	14487	250
Spain	1589219	42619	N/A	2.7%	0.0%	33985	911
UK	1527495	55230	N/A	3.6%	0.0%	22454	812
Italy	1431795	50453	584493	3.5%	40.8%	23695	835
Argentina	1370366	37002	1195492	2.7%	87.2%	30212	816
Colombia	1248417	35287	1150932	2.8%	92.2%	24433	691

Showing 1 to 10 of 209 countries

Previous 1 2 3 4 5 ... 21 Next





WHAT IS STATISTICS?

- How do *you* imagine Statistics?
  - Can you think of any example of Statistics?
  - What is it?

WHAT IS STATISTICS?

- How do *you* imagine Statistics?
  - Can you think of any example of Statistics?
  - What is it?
  - Why do we need it?

WHAT IS STATISTICS?

- Statistics is a collection of methods which help us to describe, summarize, interpret and analyse data.

WHAT IS STATISTICS?

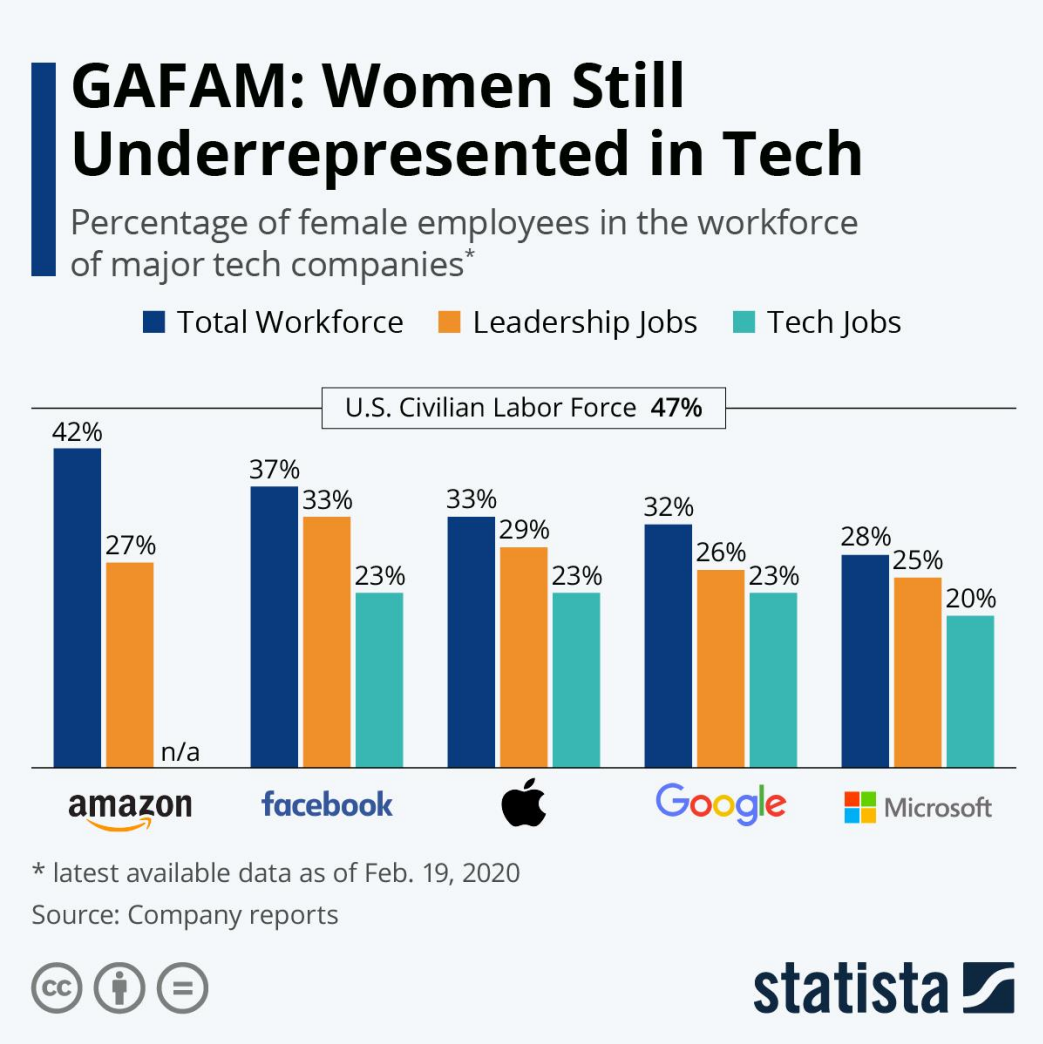
- Statistics is a collection of methods which help us to describe, summarize, interpret and analyse data.

- Vital in research, politics, management, business…

WHAT IS STATISTICS?

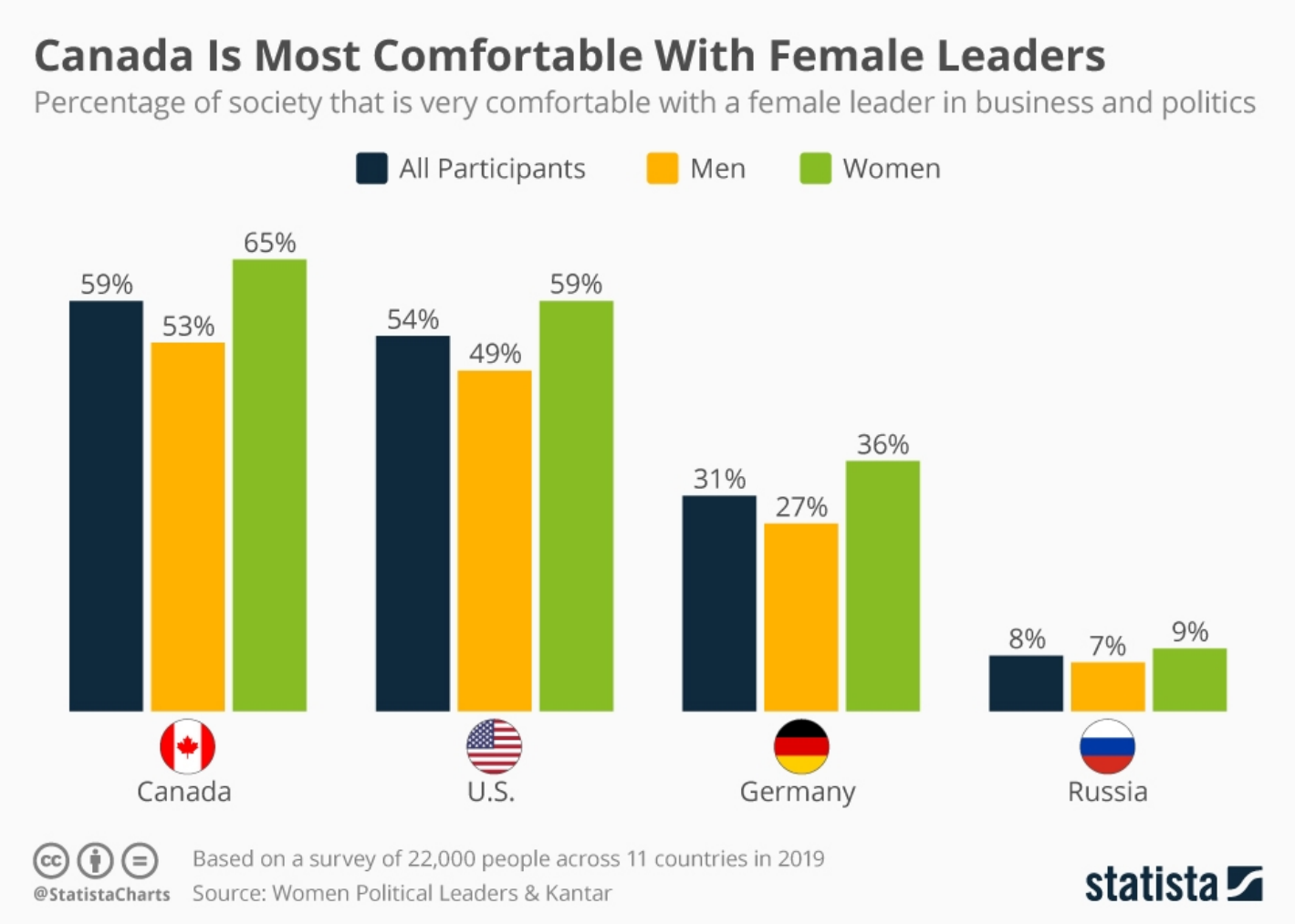
- Statistics is a collection of methods which help us to describe, summarize, interpret and analyse data.
- Vital in research, politics, management, business…
- There are different kinds of Statistics…

STATISTICS: EXAMPLE 1



Source: https://www.statista.com/chart/4467/female-employees-at-tech-companies/

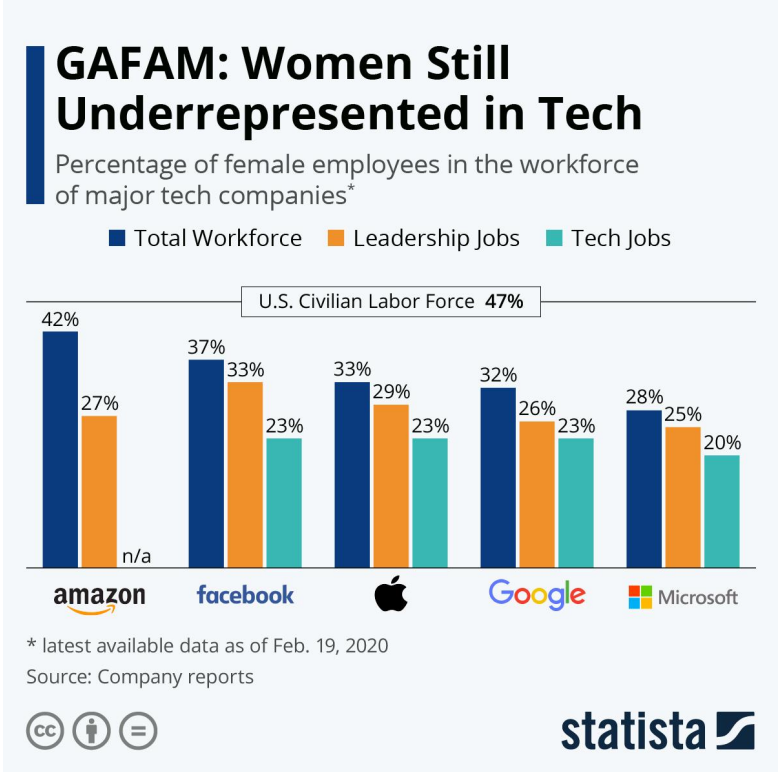
STATISTICS: EXAMPLE 2



Source: https://www.statista.com/chart/20018/canada-most-comfortable-with-female-leaders/

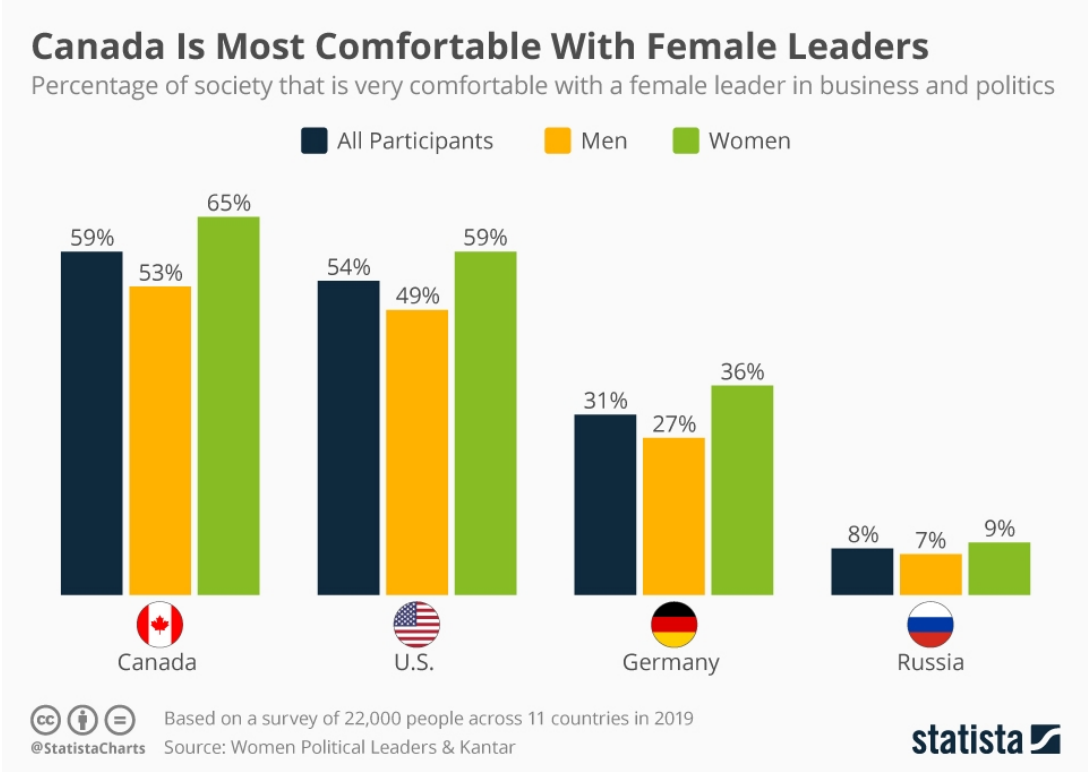
WHAT'S THE DIFFERENCE BETWEEN THE TWO?

EXAMPLE 1



Source: https://www.statista.com/chart/4467/female-employees-at-tech-companies/

EXAMPLE 2



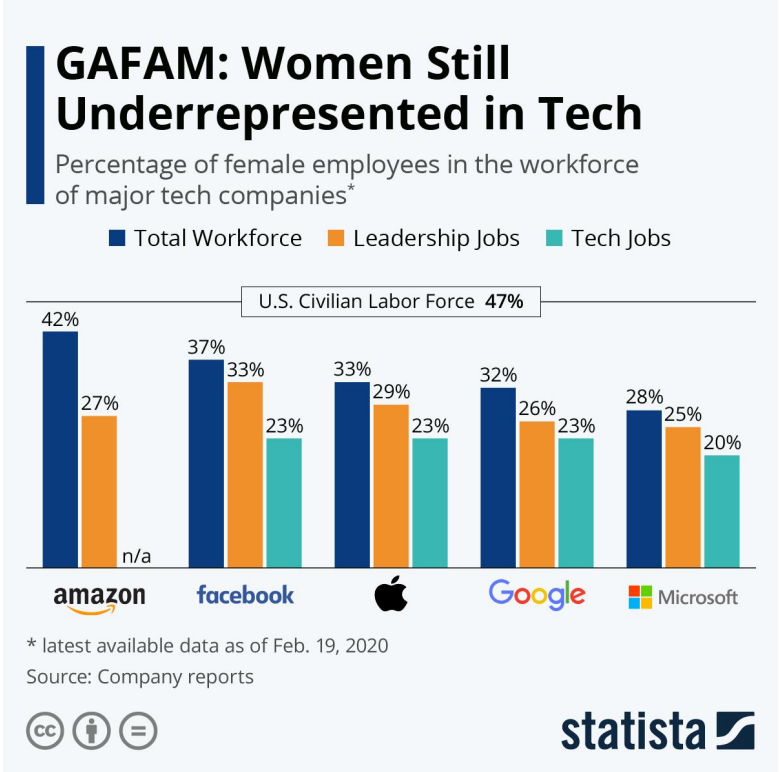
Source: https://www.statista.com/chart/20018/canada-most-comfortable-with-female-leaders/

WATCH THE VIDEO:

https://bit.ly/3fy8nzd

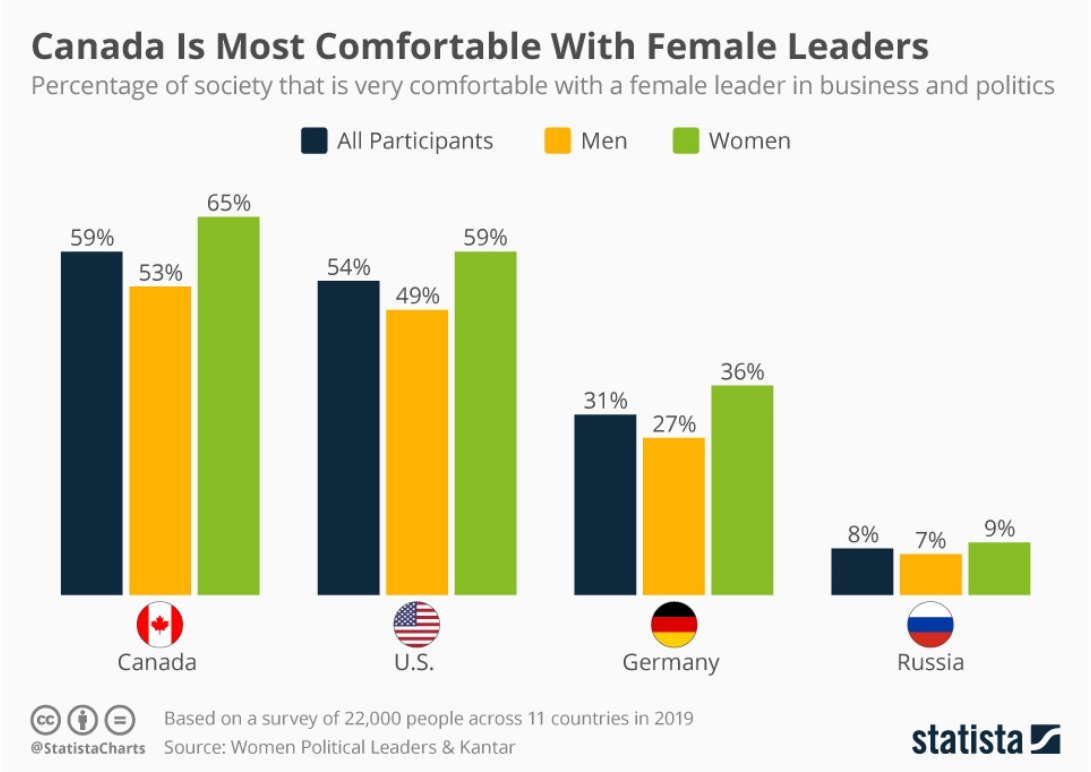
WHAT'S THE DIFFERENCE BETWEEN THE TWO?

EXAMPLE 1



Source: https://www.statista.com/chart/4467/female-employees-at-tech-companies/

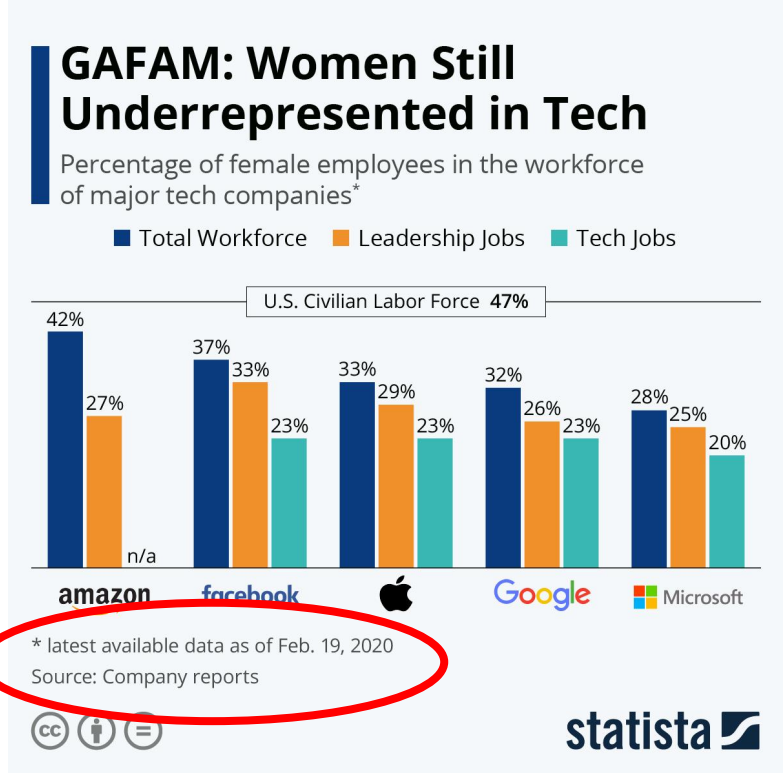
EXAMPLE 2



Source: https://www.statista.com/chart/20018/canada-most-comfortable-with-female-leaders/

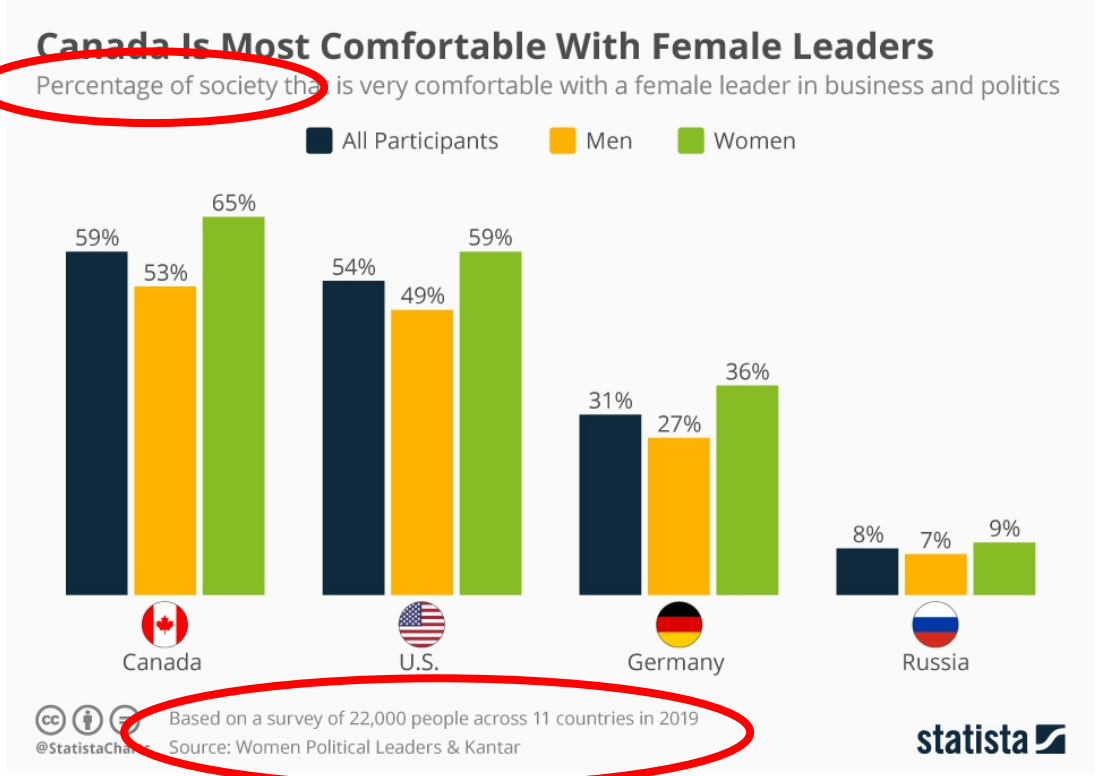
WHAT'S THE DIFFERENCE BETWEEN THE TWO?

DESCRIPTIVE STATISTICS



Source: https://www.statista.com/chart/4467/female-employees-at-tech-companies/

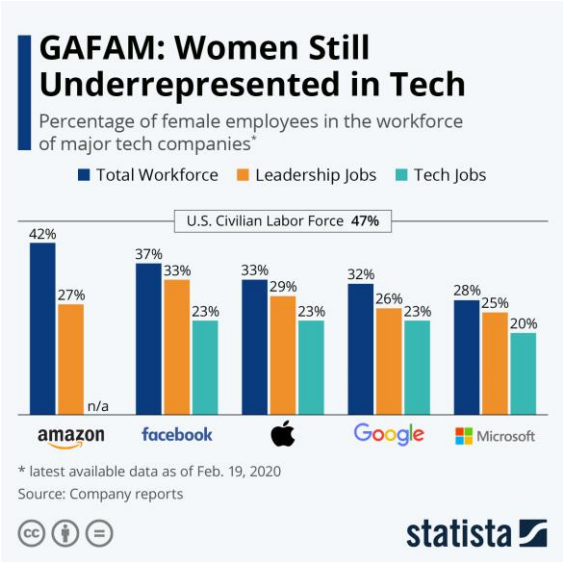
INFERENTIAL STATISTICS



Source: https://www.statista.com/chart/20018/canada-most-comfortable-with-female-leaders/

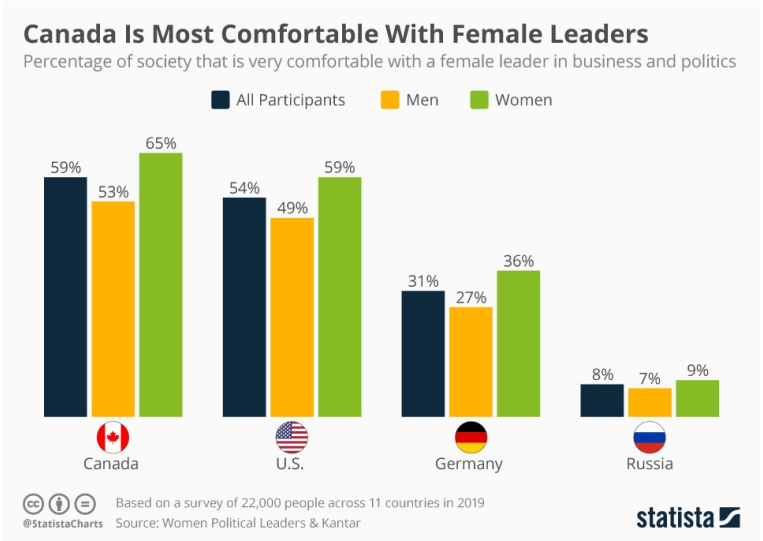
DESCRIPTIVE VS INFERENTIAL STATISTICS

DESCRIPTIVE STATISTICS



- *Describe* the data at hand.

INFERENTIAL STATISTICS



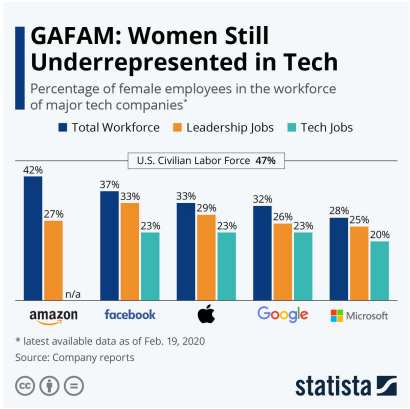
- From the data at hand, *make conclusions about a larger group.*

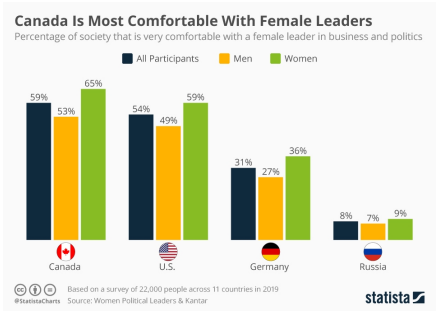
DESCRIPTIVE STATISTICS

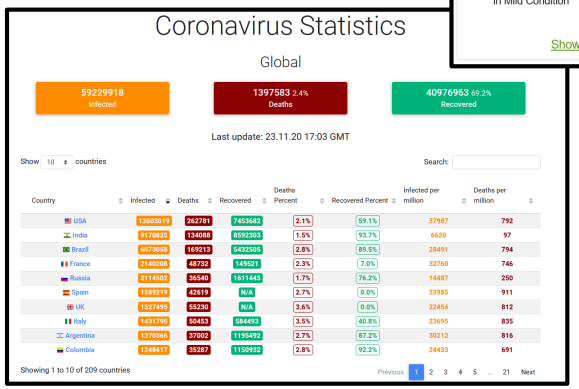
Describing the data at hand

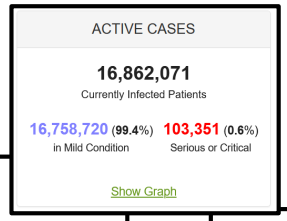
DATA

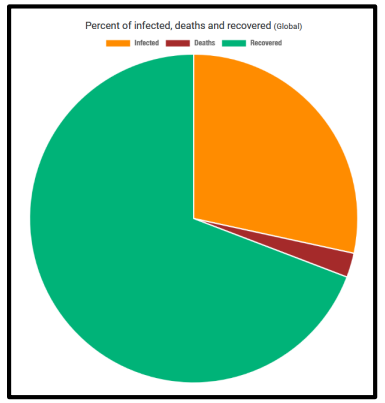
- Statistic is all about **data**.

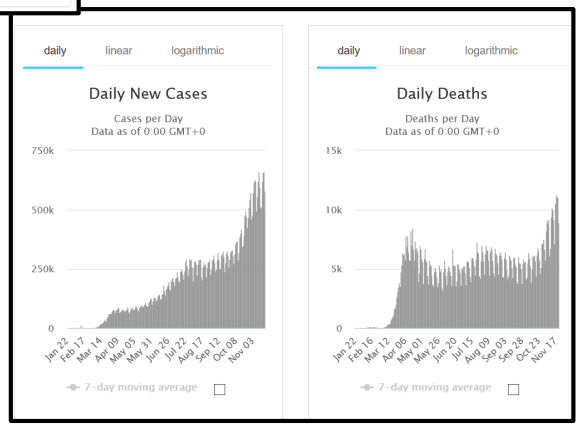






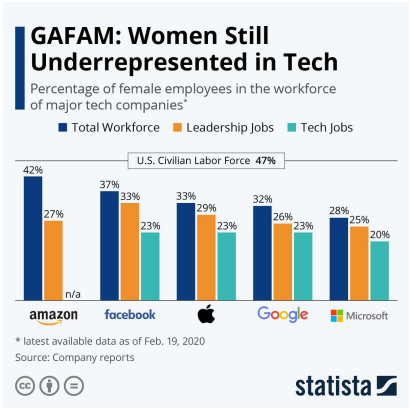


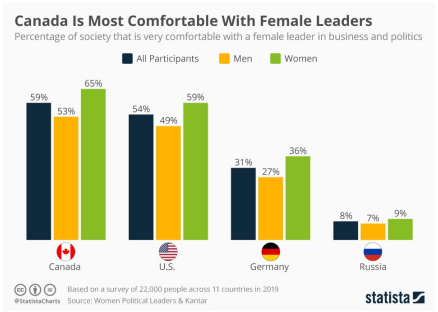


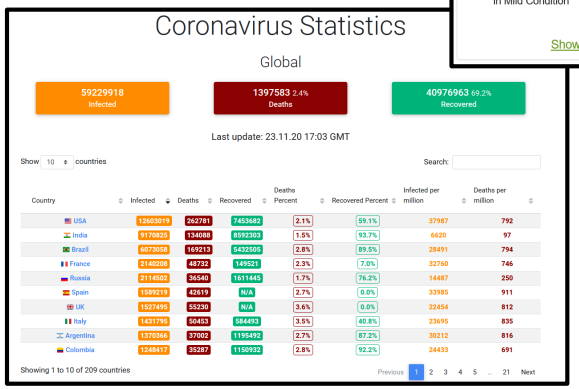


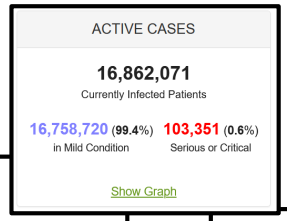
DATA

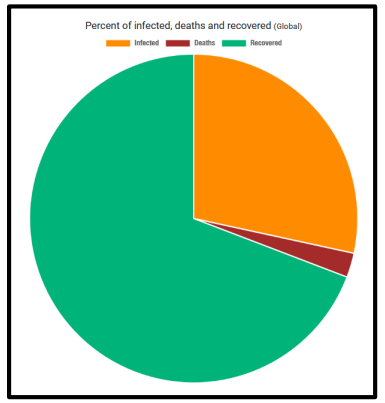
- Statistic is all about **data**.

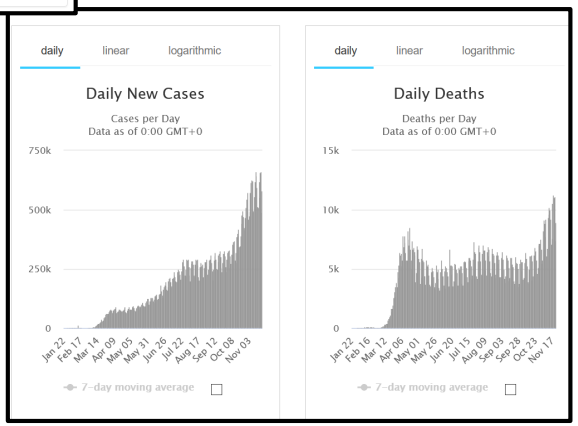












- But what ***kinds of data*** are there?
  - Important to know: different methods are applicable to different kinds of data.

Quantitative vs Qualitative Variables

QUANTITATIVE measurable quantities	**QUALITATIVE**

Quantitative vs Qualitative Variables

QUANTITATIVE measurable quantities	**QUALITATIVE**

- Air temperature
  (-1, 15, 21.5, …)
- Shoe size
  (38, 39, 40, …)
- Score for the final exam
  (100, 0, 85, …)

Quantitative vs Qualitative Variables

QUANTITATIVE
measurable quantities

- Air temperature
  (-1, 15, 21.5, …)
- Shoe size
  (38, 39, 40, …)
- Score for the final exam
  (100, 0, 85, …)

QUALITATIVE

- Eye colour
  (blue, brown, grey, …)
- Political party you support
  (Republicans, Democrats, …)
- Type of transport you use to commute
  (metro, bus, bike, …)

Quantitative vs Qualitative Variables

- Consider variable *SEX* that takes two values: *'male'* or *'female'*.

NAME	SEX
Ann	Female
Bob	Male
Kate	Female
Nick	Male

Quantitative vs Qualitative Variables

- Consider variable *SEX* that takes two values: *'male'* or *'female'*.

- In a dataset
  - *'male'* is represented as *0*
  - *'female'* is represented as *1*.

NAME	SEX
Ann	Female
Bob	Male
Kate	Female
Nick	Male

NAME	SEX
Ann	1
Bob	0
Kate	1
Nick	0

Quantitative vs Qualitative Variables

- Consider variable *SEX* that takes two values: *'male'* or *'female'*.

- In a dataset
  - *'male'* is represented as *0*
  - *'female'* is represented as *1*.

NAME	SEX
Ann	Female
Bob	Male
Kate	Female
Nick	Male

NAME	SEX
Ann	1
Bob	0
Kate	1
Nick	0

Is SEX a qualitative or a quantitative variable now?

Quantitative vs Qualitative Variables

QUANTITATIVE
measurable quantities
can be ordered

- Air temperature
  (-1, 15, 21.5, …)
- Shoe size
  (38, 39, 40, …)
- Score for the final exam
  (100, 0, 85, …)

QUALITATIVE
can't be ordered

- Eye colour
  (blue, brown, grey, …)
- Political party you support
  (Republicans, Democrats, …)
- Type of transport you use to commute
  (metro, bus, bike, …)

Discrete vs Continuous Variables

CONTINIOUS	DISCRETE

Discrete vs Continuous Variables

CONTINIOUS
take infinite number of values

DISCRETE
take finite number of values

Discrete vs Continuous Variables

CONTINIOUS	**DISCRETE**
take infinite number of values	take finite number of values

- *Some* quantitative variables:
  - height;
  - time to travel to work;
  - etc.

Discrete vs Continuous Variables

CONTINIOUS
take infinite number of values

- *Some* quantitative variables:
  - height;
  - time to travel to work;
  - etc.

DISCRETE
take finite number of values

- *All* qualitative variables.
- *Some* quantitative variables:
  - shoe size;
  - etc.

NOW, LET'S ANALYZE SOME DATA!

OUR DATASET

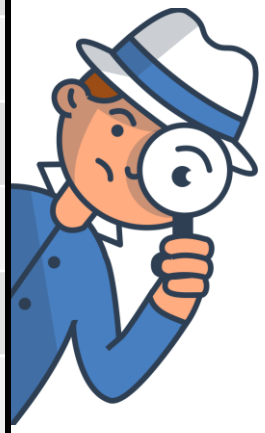
	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- You are considering joining a company.

OUR DATASET

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- You are considering joining a company.
- **What could you learn from the data available?**



How old are the employees there?
How much do they make?

Measures of center: mean and median

MEAN

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

MEAN

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- What is the mean age?

$$\frac{30 + 30 + 20 + 45 + 25 + 30 + 25 + 40 + 35 + 20}{10} = 30$$

MEAN

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- What is the mean age?

$$\frac{30 + 30 + 20 + 45 + 25 + 30 + 25 + 40 + 35 + 20}{10} = 30$$

- What is the mean salary?

MEAN

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- What is the mean age?

$$\frac{30 + 30 + 20 + 45 + 25 + 30 + 25 + 40 + 35 + 20}{10} = 30$$

- What is the mean salary?

$$\frac{3 + 4 + 1 + 50 + 2 + 3 + 4 + 1 + 2 + 1}{10} = 7.2$$

MEAN

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- What is the mean age?

$$\frac{30 + 30 + 20 + 45 + 25 + 30 + 25 + 40 + 35 + 20}{10} = 30$$

- What is the mean salary?

$$\frac{3 + 4 + 1 + 50 + 2 + 3 + 4 + 1 + 2 + 1}{10} = 7.2$$

MEAN

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO **outlier**	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- What is the mean age?

$$\frac{30 + 30 + 20 + 45 + 25 + 30 + 25 + 40 + 35 + 20}{10} = 30$$

- What is the mean salary?

$$\frac{3 + 4 + 1 + 50 + 2 + 3 + 4 + 1 + 2 + 1}{10} = 7.2$$

MEAN

SENSITIVE TO OUTLIERS ☹

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO **outlier**	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- What is the mean age?

$$\frac{30 + 30 + 20 + 45 + 25 + 30 + 25 + 40 + 35 + 20}{10} = 30$$

- What is the mean salary?

$$\frac{3 + 4 + 1 + 50 + 2 + 3 + 4 + 1 + 2 + 1}{10} = 7.1$$

MEDIAN

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- A value such that 50% of the data is lower than it, and 50% is higher.
- What is the median salary?

MEDIAN

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- A value such that 50% of the data is lower than it, and 50% is higher.
- What is the median salary?

MEDIAN

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- A value such that 50% of the data is lower than it, and 50% is higher.

sort from min to max

- What is the median salary?

1	1	1	2	2	3	3	4	4	50

MEDIAN

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- A value such that 50% of the data is lower than it, and 50% is higher.

sort from min to max

- What is the median salary?

take the value(s) in the middle

MEDIAN

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- A value such that 50% of the data is lower than it, and 50% is higher.

sort from min to max

- What is the median salary?

The median is
(2 + 3)/2 = 2.5

MEDIAN

LESS SENSITIVE TO OUTLIERS ☺

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- A value such that 50% of the data is lower than it, and 50% is higher.

sort from min to max

- What is the median salary?

1	1	1	2	2	3	3	4	4	50

The median is
(2 + 3)/2 = 2.5

Mean salary: 7.1
Median salary: 2.5

MEDIAN

LESS SENSITIVE TO OUTLIERS ☺

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- A value such that 50% of the data is lower than it, and 50% is higher.

sort from min to max

- What is the median salary?

1	1	1	2	2	3	3	4	4	50

The median is
(2 + 3)/2 = 2.5

- What is the median age?

MEDIAN

LESS SENSITIVE TO OUTLIERS ☺

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- A value such that 50% of the data is lower than it, and 50% is higher.

- What is the median salary?

1	1	1	2	2	3	3	4	4	50

The median is
(2 + 3)/2 = 2.5

sort from min to max

- What is the median age?

20	20	25	25	30	30	30	35	40	45

MEDIAN

LESS SENSITIVE TO OUTLIERS ☺

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- A value such that 50% of the data is lower than it, and 50% is higher.

- What is the median salary?

1	1	1	2	2	3	3	4	4	50

The median is
(2 + 3)/2 = 2.5

sort from min to max

- What is the median age?

20	20	25	25	30	30	30	35	40	45

30

How *different* are the salaries?
How *different* are the ages?

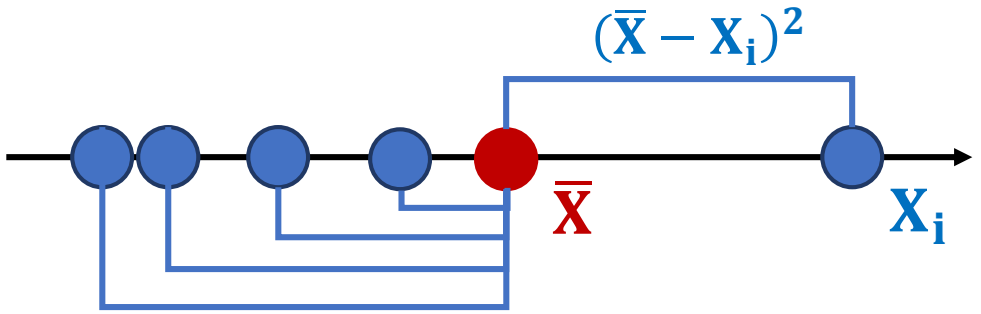
Measures of spread: variance and standard deviation, percentiles

SAMPLE VARIANCE

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Sample variance: average squared distance from all the points to the mean:

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$



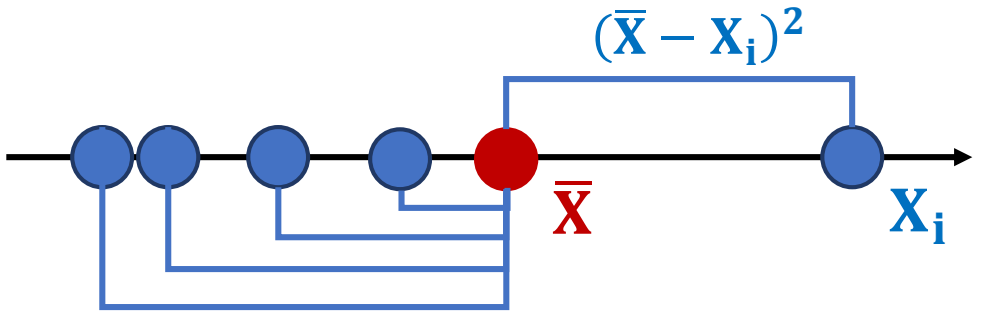
SAMPLE VARIANCE

*Sometimes, **(n-1)** instead of **n**.*
We'll discuss the reasons for that later

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Sample variance: average squared distance from all the points to the mean:

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$



SAMPLE VARIANCE

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- What is the variance of age?

$[(30-30)^2 +$
$(30-30)^2 +$
$(20-30)^2 +$
$(45-30)^2 +$
$(25-30)^2 +$
$(30-30)^2 +$
$(25-30)^2 +$
$(40-30)^2 +$
$(35-30)^2 +$
$(20-30)^2]/10 = ?$

SAMPLE VARIANCE

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- What is the variance of age?

$[(30-30)^2 +$
$(30-30)^2 +$
$(20-30)^2 +$
$(45-30)^2 +$
$(25-30)^2 +$
$(30-30)^2 +$
$(25-30)^2 +$
$(40-30)^2 +$
$(35-30)^2 +$
$(20-30)^2]/10$ = 60

SAMPLE VARIANCE

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- What is the variance of age?

$[(30-30)^2 +$
$(30-30)^2 +$
$(20-30)^2 +$
$(45-30)^2 +$
$(25-30)^2 +$
$(30-30)^2 +$
$(25-30)^2 +$
$(40-30)^2 +$
$(35-30)^2 +$
$(20-30)^2]/10 = 60$ YEARS^2

SAMPLE VARIANCE

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- What is the variance of age?

[(30-30)^2 +
(30-30)^2 +
(20-30)^2 +
(45-30)^2 +
(25-30)^2 +
(30-30)^2 +
(25-30)^2 +
(40-30)^2 +
(35-30)^2 +
(20-30)^2]/10 = 60 **YEARS^2**

HARD TO INTERPRET ☹

SAMPLE VARIANCE

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Standard deviation: square root of variance:

$$\mathbf{std}(\mathbf{X}) = \sqrt{\mathbf{Var}(\mathbf{X})}$$

SAMPLE VARIANCE

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Standard deviation: square root of variance:

$$\mathbf{std}(\mathbf{X}) = \sqrt{\mathbf{Var}(\mathbf{X})}$$

- What is the standard deviation of age?

SAMPLE VARIANCE

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Standard deviation: square root of variance:

$$\mathbf{std}(\mathbf{X}) = \sqrt{\mathbf{Var}(\mathbf{X})}$$

- What is the standard deviation of age?

std(AGE) = $\sqrt{\mathbf{60}}$ ~ 7.8

SAMPLE VARIANCE

EASY TO INTERPRET ☺

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Standard deviation: square root of variance:

$$\mathbf{std(X)} = \sqrt{\mathbf{Var(X)}}$$

- What is the standard deviation of age?

std(AGE) = $\sqrt{\mathbf{60}}$ ~ 7.8 **YEARS**

What positions are the most common there?

Measures of center: mode

MODE

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Mode: the most common value of a variable.

MODE

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Mode: the most common value of a variable.
- What's the mode of POSITION?

'Manager'

MODE

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Mode: the most common value of a variable.
- What's the mode of POSITION?

  'Manager'
- What's the mode of SEX?

MODE

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

- Mode: the most common value of a variable.
- What's the mode of POSITION?

  'Manager'
- What's the mode of SEX?

  'M'

Are men and women equally represented at each position?

Contingency tables

CONTIGENCY TABLE

CATEGORICAL x CATEGORICAL VARIABLE

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

Position / Sex	M	F	TOTAL
CEO	1	0	1
Manager	2	2	4
HR			
PR			
Secretary			
TOTAL			

CONTIGENCY TABLE

CATEGORICAL x CATEGORICAL VARIABLE

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

Position / Sex	M	F	TOTAL
CEO	1	0	1
Manager	2	2	4
HR	1	1	2
PR			
Secretary			
TOTAL			

CONTIGENCY TABLE

CATEGORICAL x CATEGORICAL VARIABLE

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

Position / Sex	M	F	TOTAL
CEO	1	0	1
Manager	2	2	4
HR	1	1	2
PR	2	0	2
Secretary			
TOTAL			

CONTIGENCY TABLE

CATEGORICAL x CATEGORICAL VARIABLE

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

Position / Sex	M	F	TOTAL
CEO	1	0	1
Manager	2	2	4
HR	1	1	2
PR	2	0	2
Secretary	0	1	1
TOTAL			

CONTIGENCY TABLE

CATEGORICAL x CATEGORICAL VARIABLE

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

Position / Sex	M	F	TOTAL
CEO	1	0	*1*
Manager	2	2	*4*
HR	1	1	*2*
PR	2	0	*2*
Secretary	0	1	*1*
TOTAL	*6*	*4*	*10*

Are men and women paid equally?

Pivot tables

PIVOT TABLE

CATEGORICAL x NUMERIC VARIABLE

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

SEX x SALARY

Sex	Mean salary (in $1000)
F	(3 + 1 + 3 + 2)/4 = 2.25
M	

PIVOT TABLE

CATEGORICAL x NUMERIC VARIABLE

	NAME	SEX	AGE	POSITION	SALARY (in $1000)
1	Ann	F	30	Manager	3
2	Bob	M	30	Manager	4
3	Kate	F	20	Secretary	1
4	Nick	M	45	CEO	50
5	John	M	25	HR	2
6	Alice	F	30	Manager	3
7	Joe	M	25	Manager	4
8	Dan	M	40	PR	1
9	Laura	F	35	HR	2
10	Jack	M	20	PR	1

SEX x SALARY

Sex	Mean salary (in $1000)
F	(3 + 1 + 3 + 2)/4 = 2.25
M	(4+50+2+4+1+1)/6= 10.3

NOW, LET'S PRACTICE WITH SOME REAL DATA!

Summary statistics + basic plots in Python

TO SUM UP…

- Two branches of Statistics.
- Data types
  - *numerical vs categorical;*
  - *continuous vs discrete.*
- Descriptive Statistics
  - *summary statistics;*
  - *tables;*
  - *plots.*

TO SUM UP…

- Two branches of Statistics.
- Data types
  - *numerical vs categorical;*
  - *continuous vs discrete.*
- Descriptive Statistics
  - *summary statistics;*
  - *tables;*
  - *plots.*

- **Please complete the ENTRY TEST!**
- Assignment 1 is online
  - 20 points
  - Deadline: Thursday 23:59.