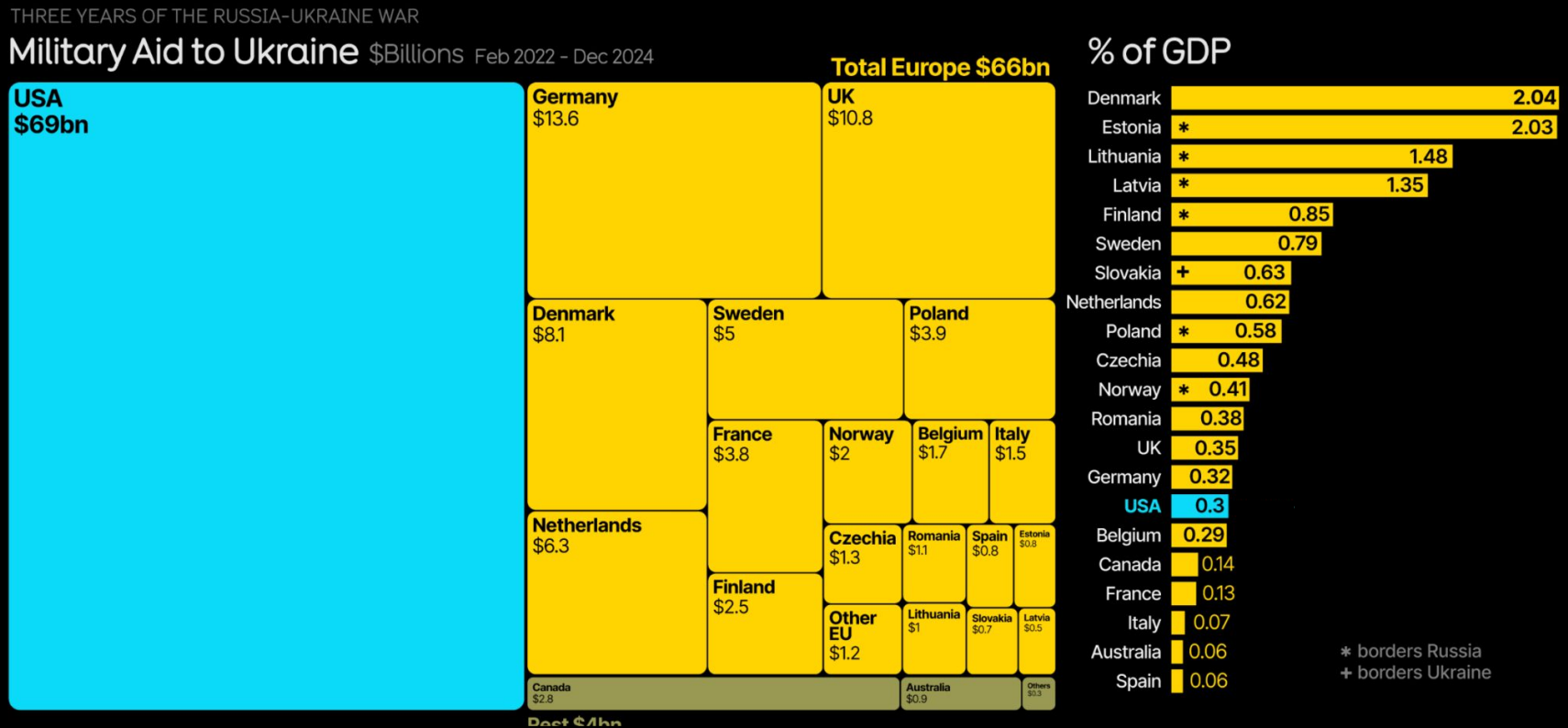
Understanding & Visualizing Data

Lecture 2

Emma Ning, M.A.

If You Were a Politician, How Would You Use the Left vs. Right Figure?



https://informationisbeautiful.net/visualizations/ukraine-russian-war-infographics-data-visuals/

TODAY'S PLAN

01

Intro to Dataframe

How does a dataframe look like?

02

Types of Data

What kind of data can we analyze?

03

Summarizing & Visualizing Data

Creating frequency tables to summarize our data & Plotting our data

04

Wrap Up

Review + Reminders

Learning objectives

- Differentiate and give examples of **nominal**, **ordinal**, **interval**, and **ratio** data
- Construct a **frequency table** based on a given set of data
- Construct univariate visualizations based on the four types of data, including basic forms of **pie chart, bar chart, histogram,** and **line graph**
- Interpret the univariate visualizations you create
- Optional (if time): Able to differentiate between univariate vs. bivariate visualizations

Intro to Dataframe

How do “data” look like, typically?

user	app_version	age	sleep_hours	subscription	stress_level
Hafsah	v3.1	23	8.5	Free	Med
Jackie	v2.9	30	6	Pro	Low
Aseem	v3.1	19	9.5	Free	High
Oscar	v3.3	21	5.9	Free	High

These are **columns**. Each column represents one **variable**.

	user	app_version	age	sleep_hours	subscription	stress_level
→	Hafsah	v3.1	23	8.5	Free	Med
→	Jackie	v2.9	30	6	Pro	Low
→	Aseem	v3.1	19	9.5	Free	High
•••	Oscar	v3.3	21	5.9	Free	High

These are **rows**. Each row represents one **observation** (e.g., person).

user	app_version	age	sleep_hours	subscription	stress_level
Hafsah	v3.1	23	8.5	Free	Med
Jackie	v2.9	30	6	Pro	Low
Aseem	v3.1	19	9.5	Free	High
Oscar	v3.3	21	5.9	Free	High

The variables in red seem to include some text.
The variables in blue seem to only have numbers.

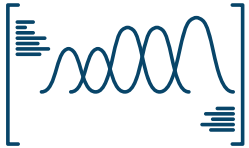
02

Types of Data

Two Broad Types of Data/Variable

Categorical

These are variables that are **categories**, such as first language, hometown, or favorite song. Think of categorical variables as **labels**.



Continuous

These are **numeric** variables that are coded with **meaningful numbers**, including height, heart rate, or stress level. Think of them as **quantities**.

Scales of Measurement

The 4 scales of measurement “zoom into” the smaller differences between categorical vs. continuous variables, and are what scientists use.

Categorical **Ordinal** **Interval** **Ratio**

Four Levels of Measurement

Categorical (nominal)	These are **categories** or **names**; they have no inherent order	→	name; gender identity; hometown
Ordinal	The are names or numbers that represent a **rank-order**; the distance between each rank is ***not*** equal.	→	order in a competition; income bracket
Interval	The numbers represent **equal distances**, but there is <u>not</u> a **true zero** (e.g., zero point is arbitrary, like calendar year)	→	Temperature in *Celsius/Fahrenheit*; credit score
Ratio	The numbers represent equal distances, but there is a **true & meaningful zero**.	→	Income; Age

	Nominal	Ordinal	Interval	Ratio
Categorizes and labels variables	✔	✔	✔	✔
Ranks categories in order		✔	✔	✔
Has known, **equal intervals**			✔	✔
Has a true, meaningful **zero**				✔

Back to the Dataset...

categorical,
app_version is ordinal

continuous

Categorical - but seem to have some order? - ordinal!

user	app_version	age	sleep_hours	subscription	stress_level
Hafsah	v3.1	23	8.5	Free	Med
Jackie	v2.9	30	6	Pro	Low
Aseem	v3.1	19	9.5	Free	High
Oscar	v3.3	21	5.9	Free	High

The variables in red seem to include some text.
The variables in blue seem to only have numbers.

THINK - PAIR - SHARE

What **scales of measurement** are the following variables on?

1

Type of Commute

Train, Car, Bus, Walk, Bike

2

T-shirt Size

small, medium, large, x-large

3

Hometown

Decatur, AL; Chicago, IL; Seoul, South Korea

4

Commute Time

5 mins, 23 mins, 105 mins

5

Temperature in Fahrenheit

70 ℉, 32℉, -10℉

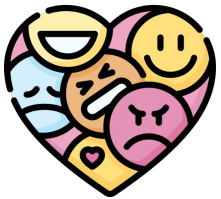
6

Stress rating (Likert Scale)

3 out of 7, 6 out of 7

Categorical | Ordinal | Interval | Ratio

Summarizing & Visualizing Data



Say we are interested in people's emotions:
We want to know how people are feeling

What emotion do you feel, right now?

Anxious Curious Stressed Calm Distracted

○ ○ ○ ○ ○

Our data

ID	Emotion
Hijab	Curious
Yocelyn	Curious
Arleth	Calm
Kaden	Distracted
Andres	Anxious
Ken	Curious

Hard to see patterns?
Let's **summarize** it:

→

Frequency Table

Emotion label	Count/ Absolute frequency	Proportion/ Relative Frequency
Anxious	1	⅙ ≈ 0.167 = 16.7%
Curious	3	3/6 = 0.5 = 50.0%
Stressed	0	0/6 = 0 = 0%
Calm	1	⅙ ≈ 0.167 = 16.7%
Distracted	1	⅙ ≈ 0.167 = 16.7%
N	6	100%

On the left is **absolute** value (spending). On the right is **relative/proportional** value (spending).

https://informationisbeautiful.net/visualizations/ukraine-russian-war-infographics-data-visuals/

What emotion do you feel, right now?

Anxious Curious Stressed Calm Distracted

Our data

ID	Emotion
Hijab	Curious
Yocelyn	Curious
Arleth	Calm
Kaden	Distracted
Andres	Anxious
Ken	Curious

How do we see patterns?

Let's **summarize** it:

Let's come back to our question. As opposed to 6 people listed here, let's assume we have a lot more people in our dataset. Specifically, a sample size of *N* = 53.

Frequency Table

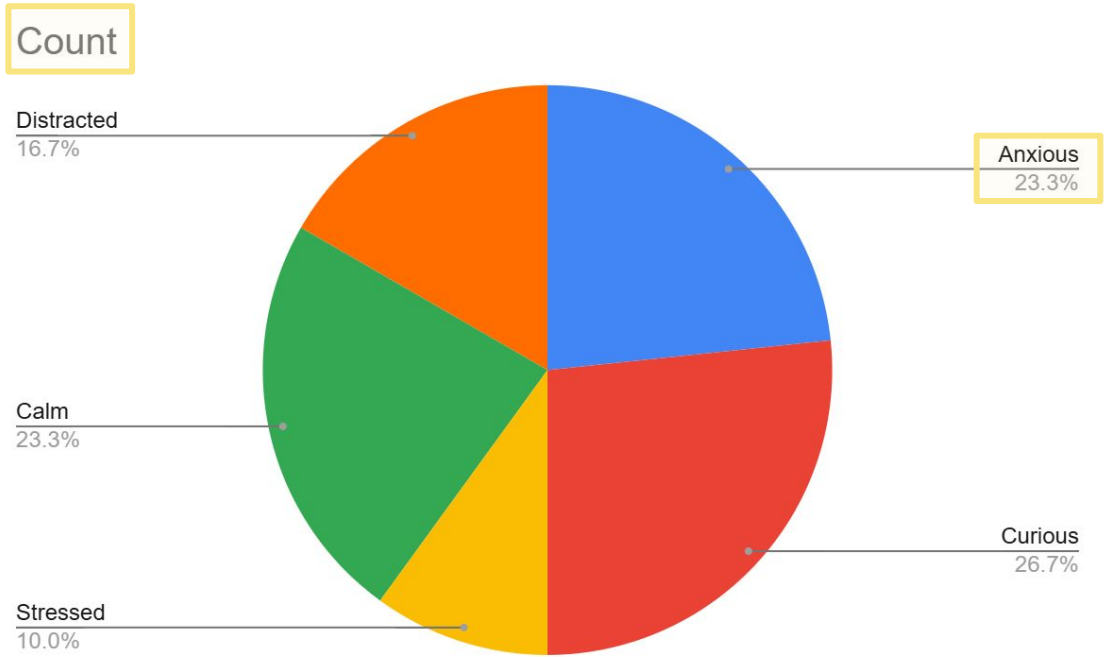
Emotion label	Count/ Absolute frequency	Proportion/ Relative Frequency
Anxious	1	1/6 ≈ 0.167 = 16.7%
Curious	3	3/6 = 0.5 = 50.0%
Stressed	0	0/6 = 0 = 0%
Calm	1	1/6 ≈ 0.167 = 16.7%
Distracted	1	1/6 ≈ 0.167 = 16.7%
N	6	100%

What emotion do you feel, right now?

Anxious Curious Stressed Calm Distracted

Since we have a lot more people, we can create our pie chart:

Notice the only “action” we can do to a nominal variable is to **count** it.



The % is our relative frequency/ proportion.

What emotion do you feel, right now?

Anxious Curious Stressed Calm Distracted

Since we have a lot more people, we can create our pie chart:

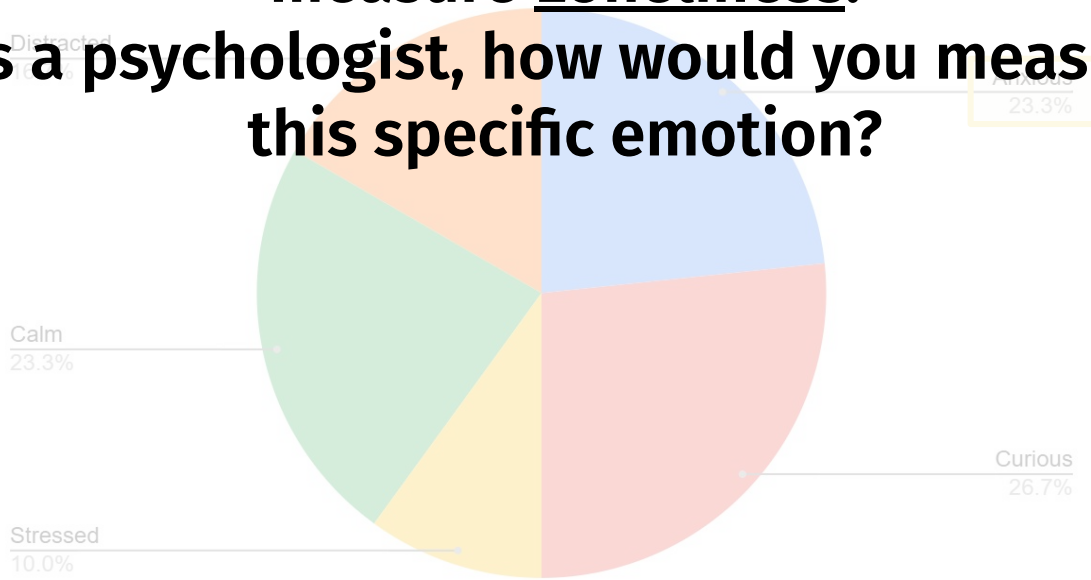
Let's pick a specific emotion: say we want to measure Loneliness.

As a psychologist, how would you measure this specific emotion?

Notice the only "action" we can do to a nominal variable is to **count** it.

Count

The % is our relative frequency/ proportion.



Loneliness

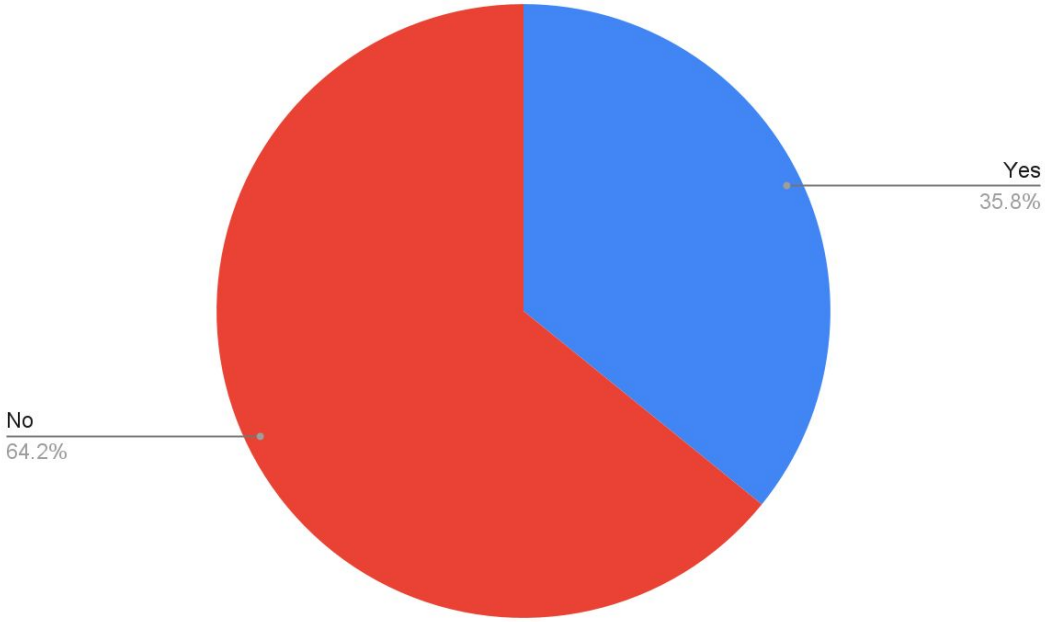
Are you lonely?

yes ◯ no ◯

Variable type: Nominal
Action: Count

Loneliness (Yes or no)	Count/ Absolute frequency	Proporton/ Relative frequency
No	34	34/53 = 0.642
Yes	19	19/53 = 0.358
Grand Total	53	1 (or 100%)

Pie Chart



Loneliness

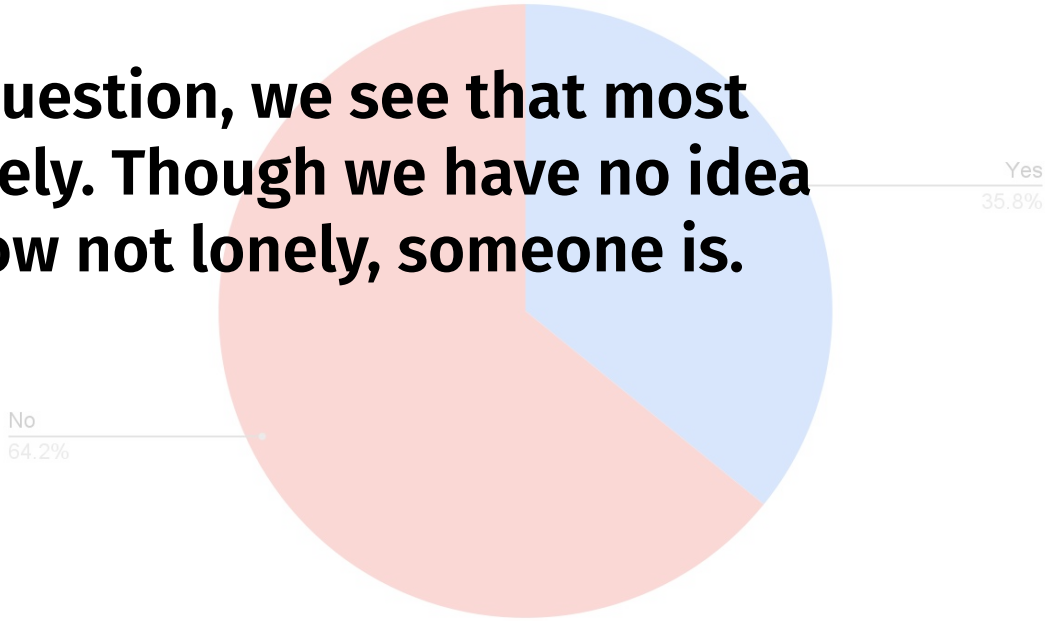
Are you lonely?

yes ○ no ○

Variable type: Nominal
Action: Count

Loneliness (Yes or no)	Count / Absolute frequency	Proportion / Relative frequency
No	34	34/53 = 0.642
Yes	19	19/53 = 0.358
Grand Total	53	1 (or 100%)

Okay, from this question, we see that most people are not lonely. Though we have no idea how lonely, or how not lonely, someone is.



Rate your level of loneliness.

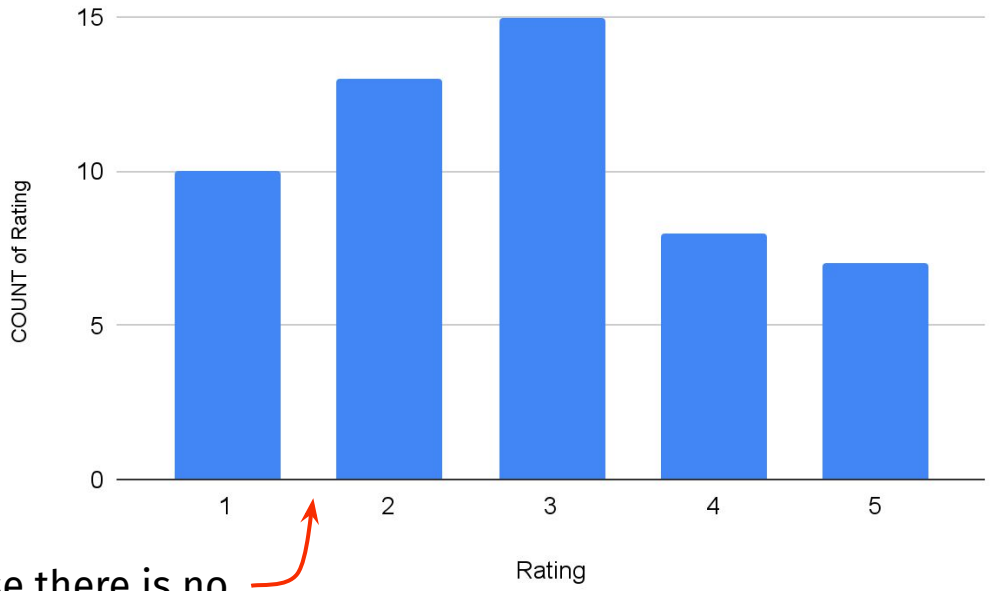
1	2	3	4	5
○	○	○	○	○
not lonely at all				extremely lonely

Variable type: Ordinal
Action: Count & Order

Frequency Table

Rating	Count/Absolute Frequency of Rating
1	10
2	13
3	15
4	8
5	7
Grand Total	53

Bar Plot



The bars are separated because there is no rating of 1.5 or anything in-between

Rate your level of loneliness.

1 2 3 4 5

not lonely at all — extremely lonely

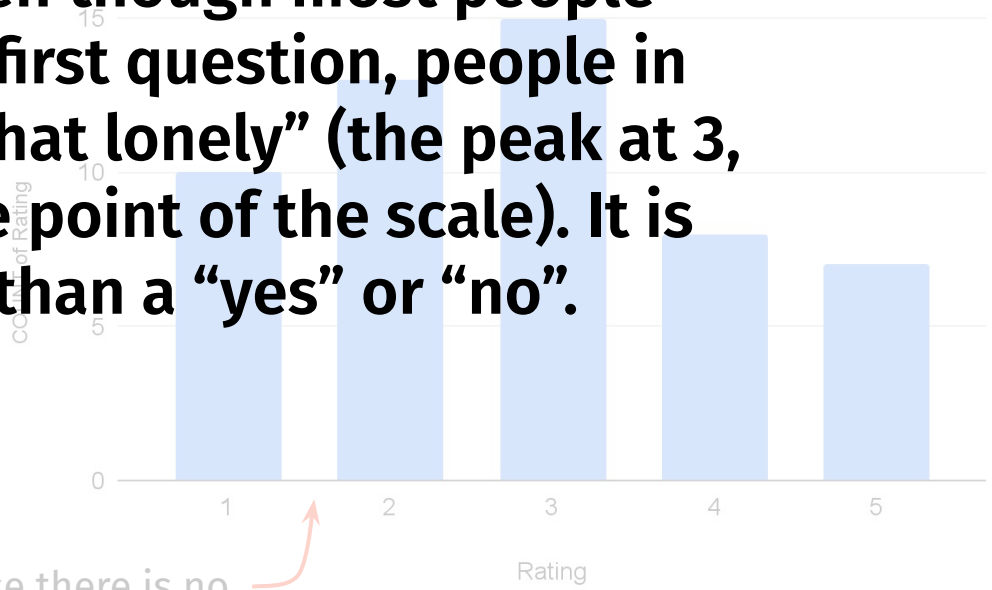
Variable type: Ordinal
Action: Count & Order

Now we know, even though most people report "No" to the first question, people in general feel "somewhat lonely" (the peak at 3, which is the middle point of the scale). It is more nuanced than a "yes" or "no".

Frequency Table

Rating	Count/Absolute Frequency of Rating
3	15
4	8
5	7
Grand Total	53

Bar Plot



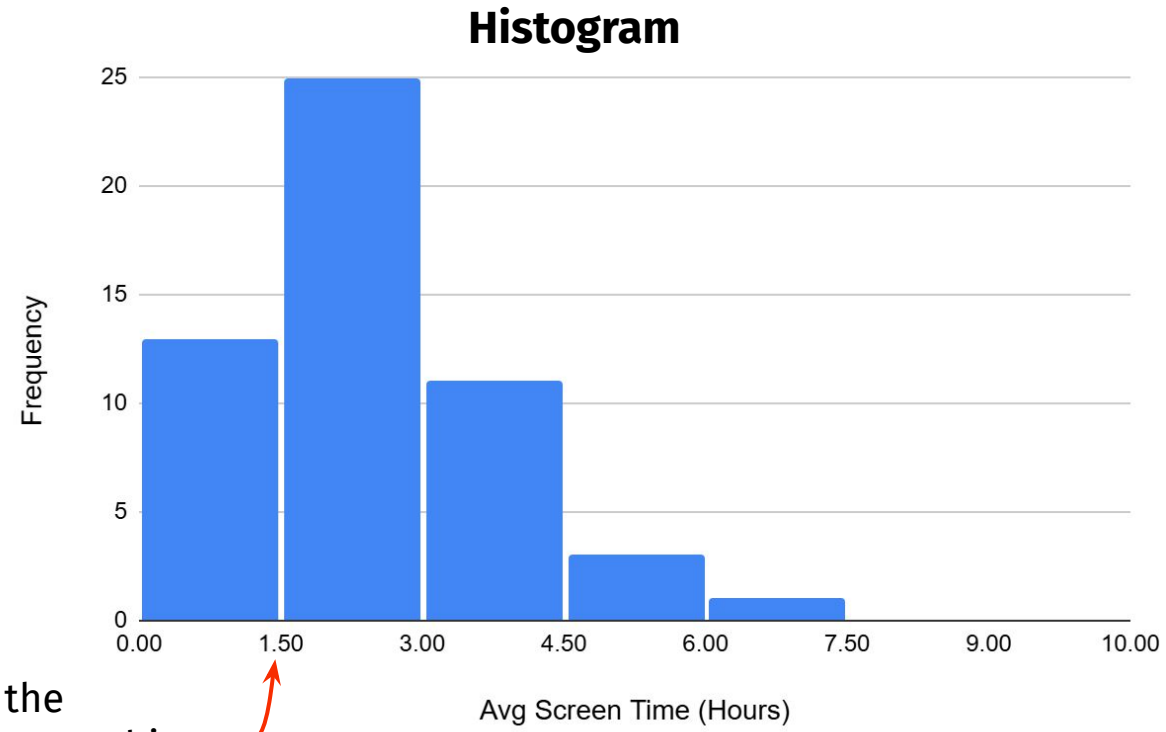
The bars are separated because there is no rating of 1.5 or anything in-between

Out of the last 7 days, **how long** is your average screen time (daily)?

Type in a number

Variable type: Ratio
Action: Bin & Aggregate

Avg Screen Time (Hrs)
1.2
0.6
2.4
1.9
2.8
2.7
1.5
3.2
3.2



The bars are not separated because the numbers are continuous: 1.51 means something

Out of the last 7 days, **how long** is your average screen time (hrs)?

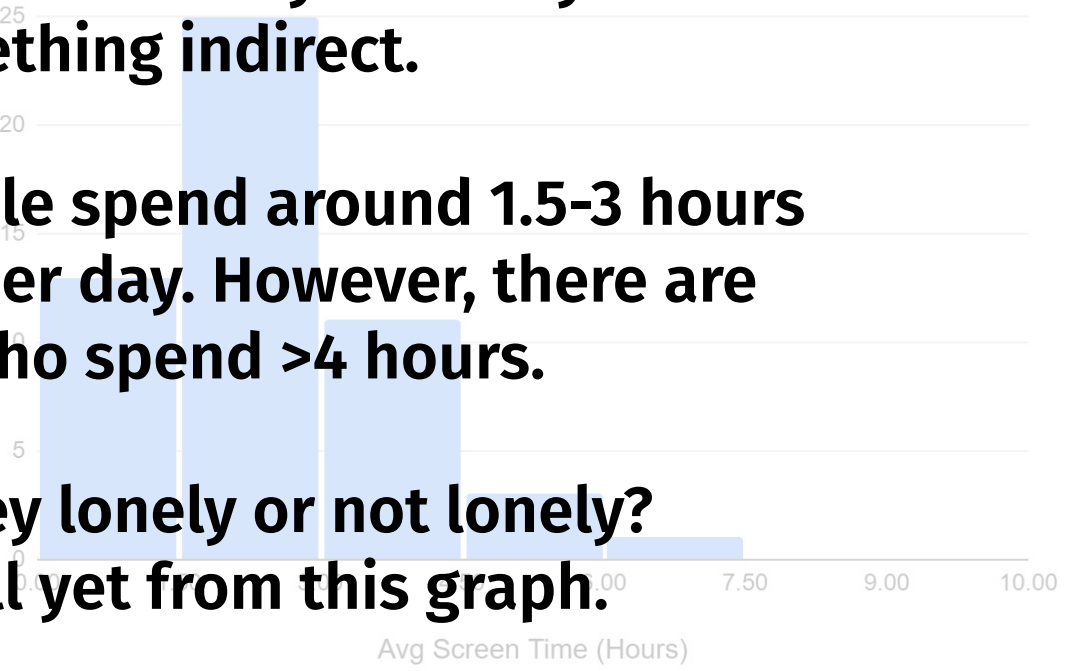
Type in a number

Avg Screen Time (Hrs)

1.2
0.6
2.4
1.9
2.7
1.5
3.2
3.2

Variable type: Ratio
Action: Bin & Aggregate

Histogram



This question operationalizes loneliness. People might lie to you for social desirability. However, they are less likely to lie if you ask something indirect.

We see most people spend around 1.5-3 hours on their phone per day. However, there are people who spend >4 hours.

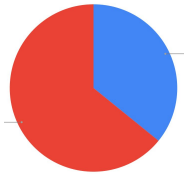
Hmm, are they lonely or not lonely? We can't tell yet from this graph.

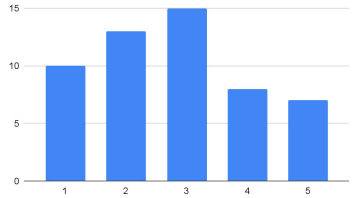
Recap: What do we have so far?

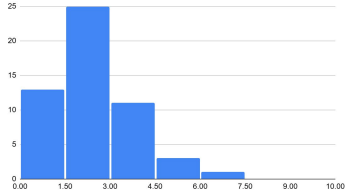
We investigated people's loneliness by asking them 3 questions. Now we can have a proper dataframe, like this:

Name	**Lonely - yes or no?**	**Self-rated loneliness (on a scale of 1-5)**	**Avg screen time per day (hours)**
Hafsah	No	2	3.1
Jackie	No	1	2.5
Aseem	Yes	5	3.9
Oscar	Yes	4	1.7
Claudia	No	3	2.1
Zion	No	2	1.3

Name	Lonely - yes or no?	Self-rated loneliness (on a scale of 1-5)	Avg screen time per day (hours)
Hafsah	No	2	3.1
Jackie	No	1	2.5
Aseem	Yes	5	3.9
Oscar	Yes	4	1.7
Claudia	No	3	2.1
Zion	No	2	1.3

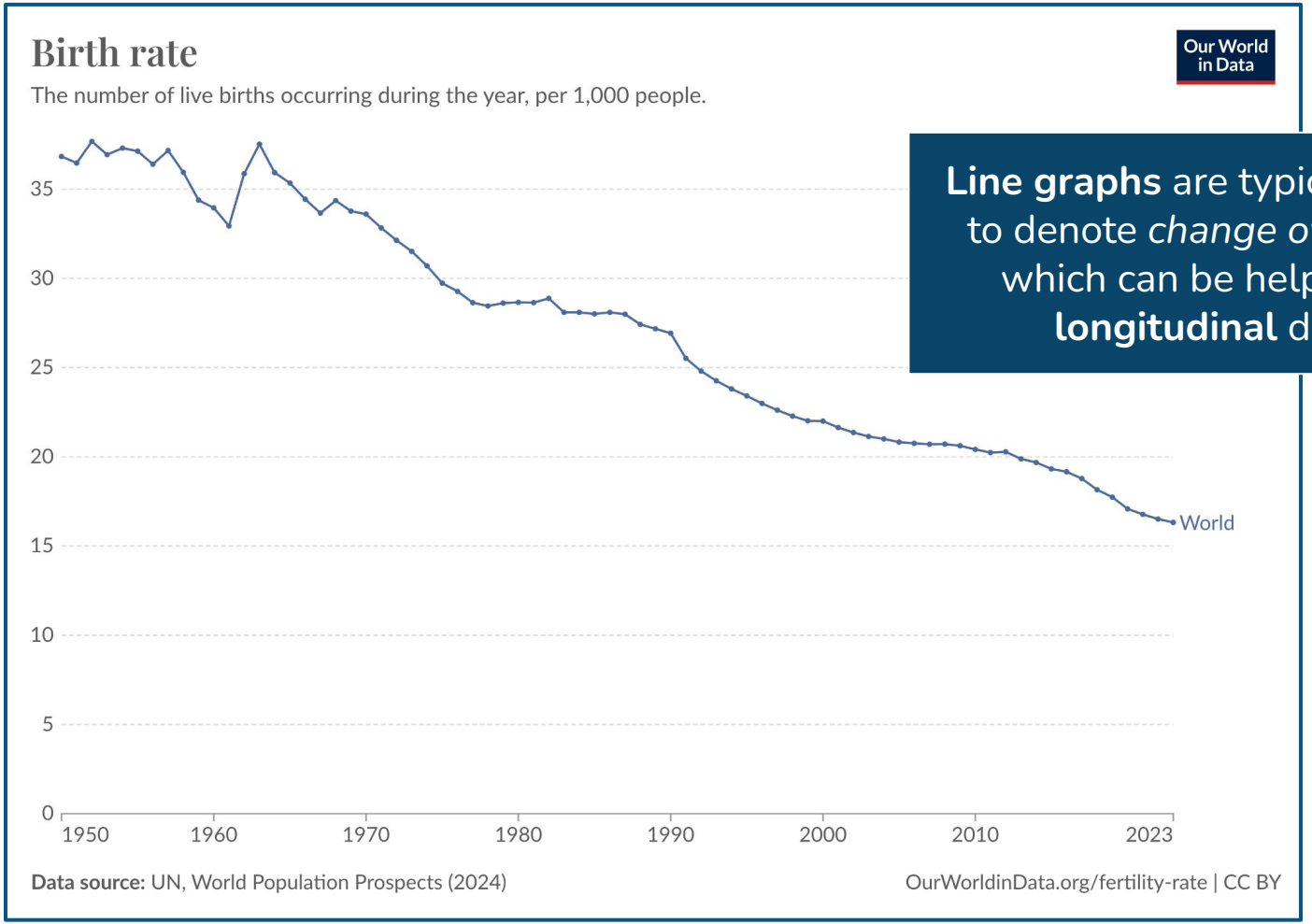






We took each column, visualized it separately. This is called **univariate** visualization. Data visualization like this summarizes data, and let us spot patterns.

There are also other types of univariate visualizations.

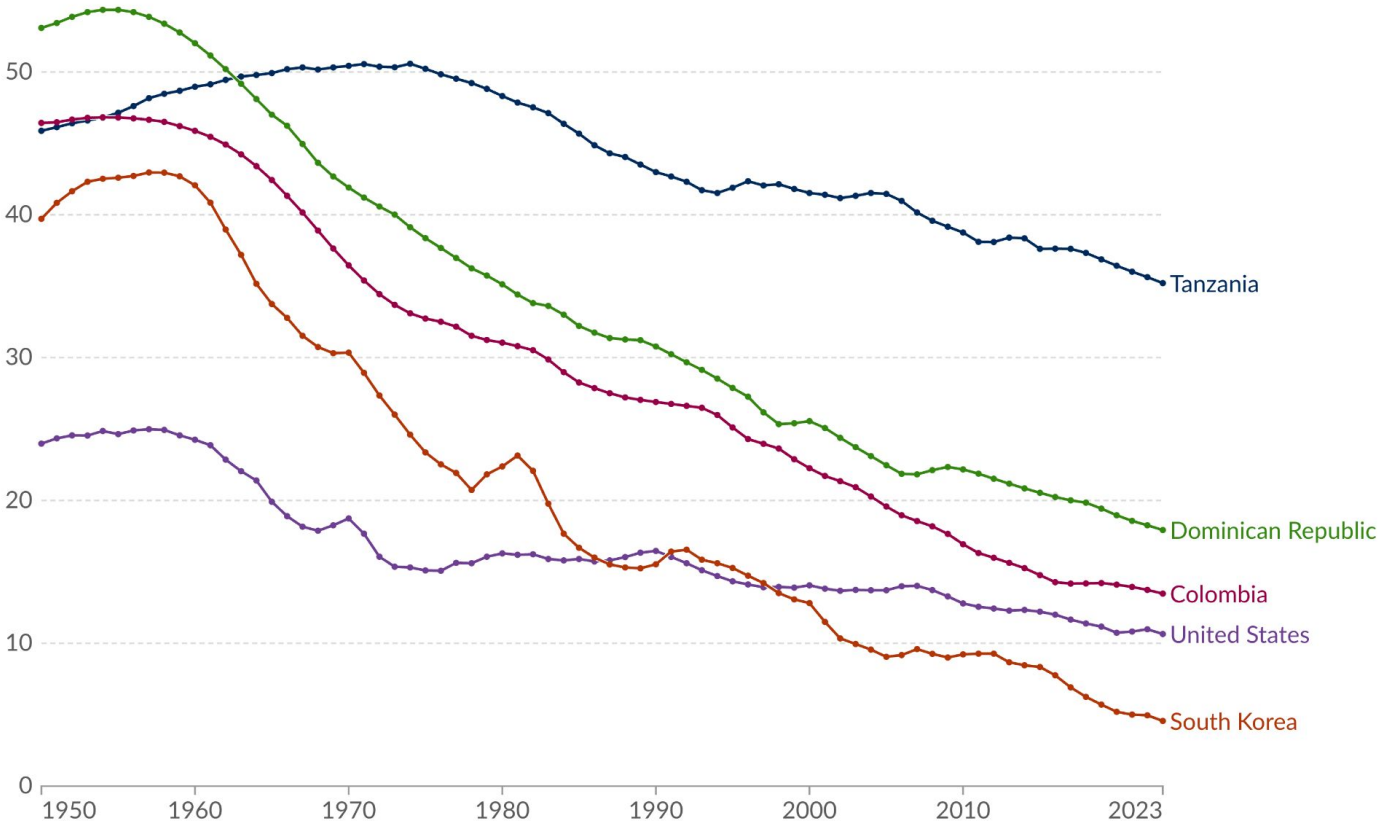


Line graphs are typically used to denote *change over time*, which can be helpful for **longitudinal** data.

Birth rate

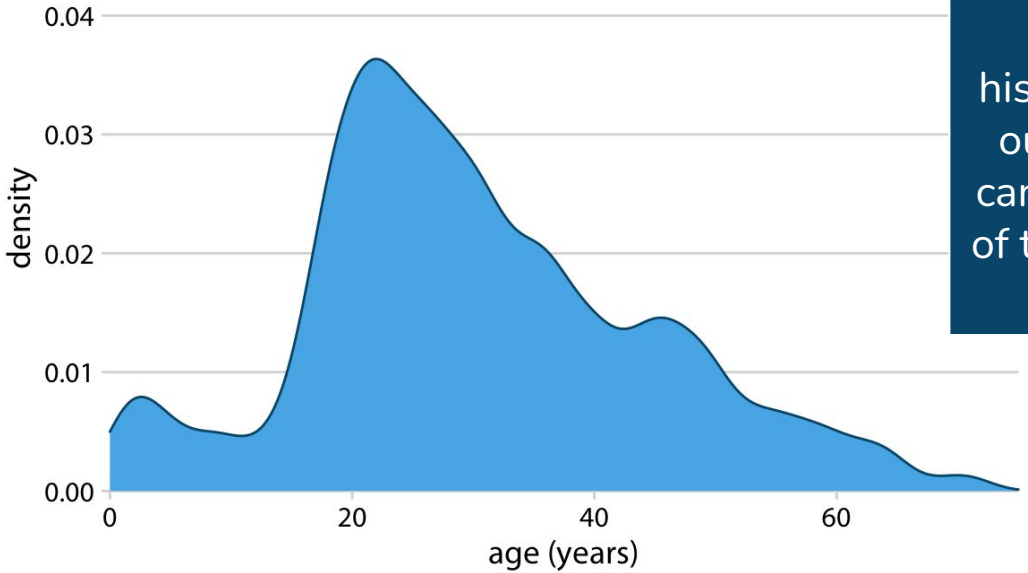
The number of live births occurring during the year, per 1,000 people.

Our World in Data

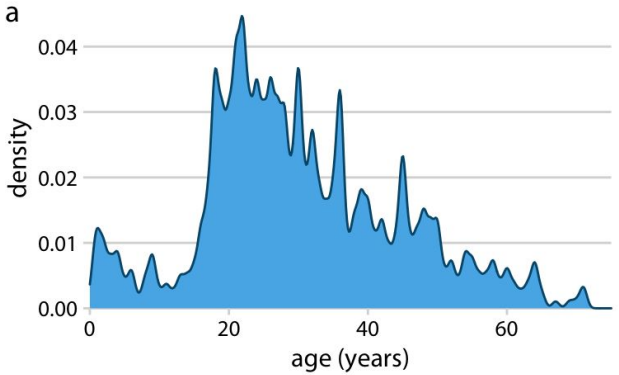


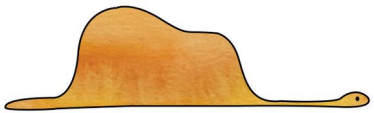
Data source: UN, World Population Prospects (2024)

OurWorldinData.org/fertility-rate | CC BY

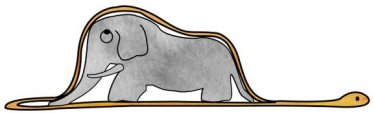


This is called a **density plot.** It is basically the same as a histogram, except you trace the outline of the histogram, so it can smooth out the ruggedness of the histogram (the one below is not ideal).

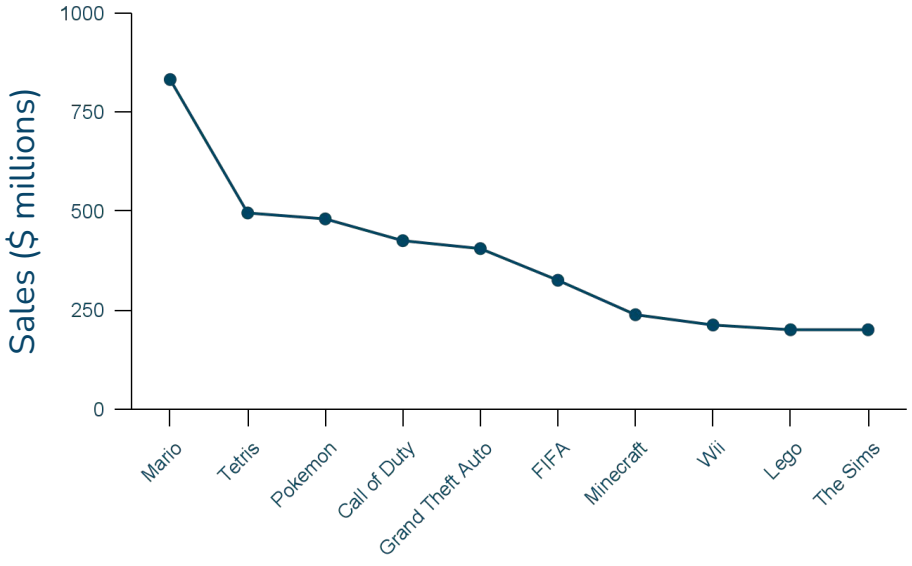


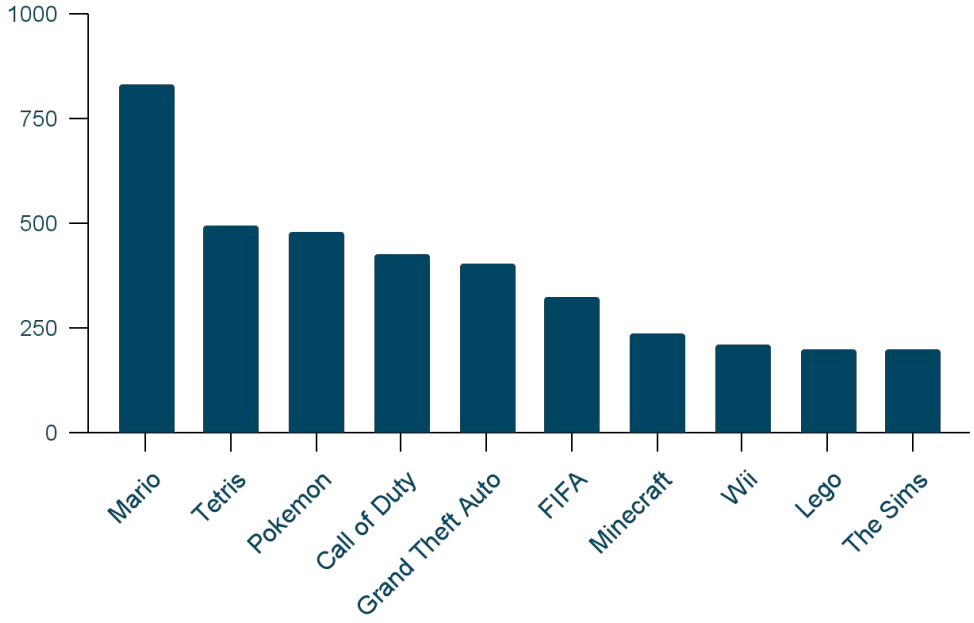


"My drawing was not a picture of a hat.
It was a picture of a boa constrictor digesting an elephant."

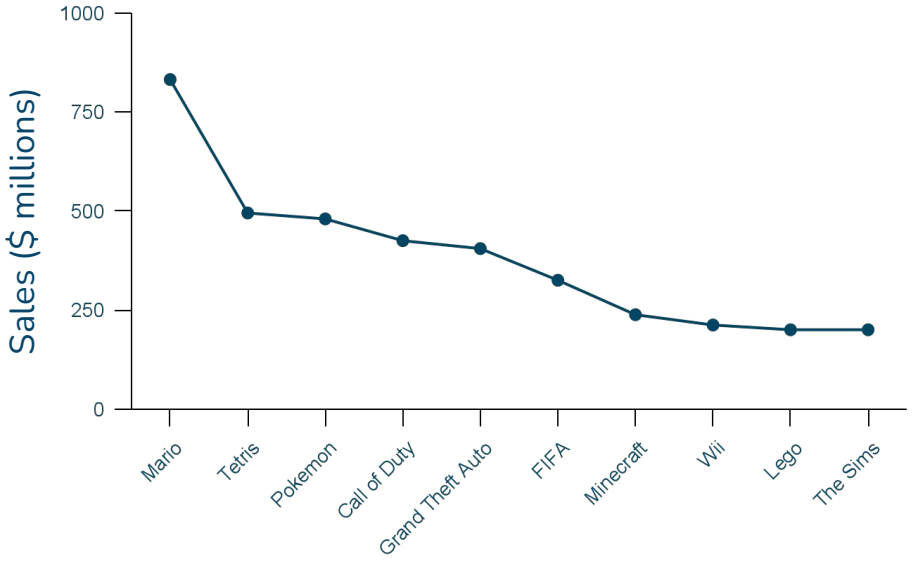


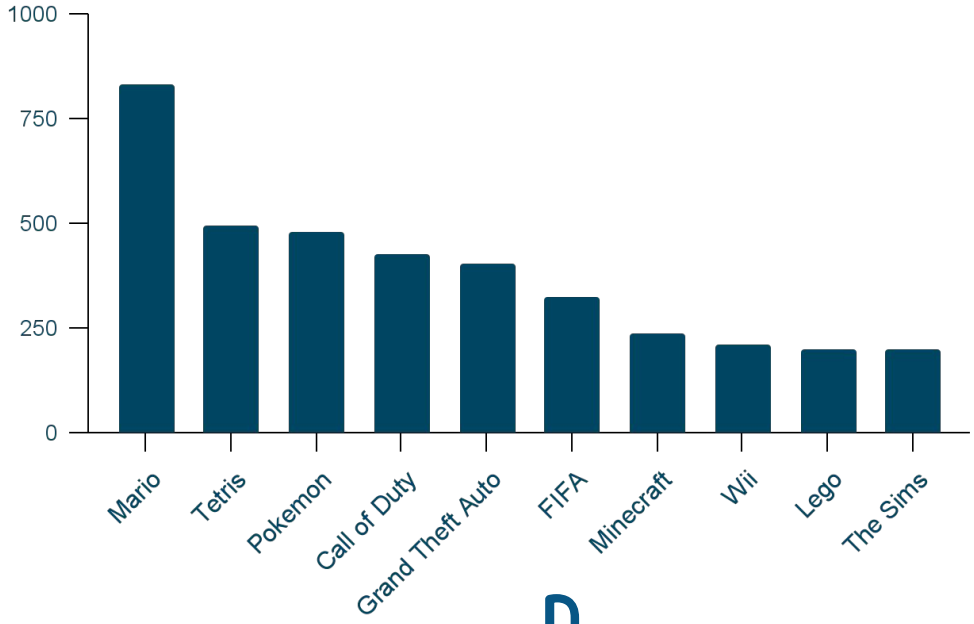
Which graph best represents the data?





Which graph best represents the data?





Bernie Sanders @BernieSand... · 9m

It is insane that in the richest country in the world, millions cannot afford a home and hundreds of thousands are homeless every night.

We need major investments in affordable housing, not tax breaks for billionaires.

56 61 314 16K

Let's think about how to help Bernie make this a better & more objective visualization!

Fill out ICA2 on Blackboard.

Wrap Up

Measurement scale - visualization choice summary

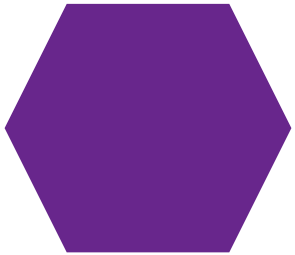
	Nominal	Ordinal	Interval	Ratio
Example	Pet type (e.g., dog, cat, fish)	Likert scale; Education level	Temperature	Weight; Commute time
Action Verbs	Count	Count & Order	Bin & Aggregate	Bin & Aggregate
Potential Visualization Choices	Pie chart; Bar chart	Pie chart; Bar chart (ordered)	Histogram; Line plot; Density plot	Histogram; Line plot; Density plot

A Note on Qualitative Data

Although not the focus of this class, many researchers work with **qualitative data**. This type of data often comes from interviews, diaries, focus groups, and open-ended survey answers. This type of data is often rich and detailed, and it may be more useful.

Survey Question: What are you looking forward to in this class?

"I am most excited about learning the various applications that statistics has in the real world and how we can analyze real data sets in our lab classes."

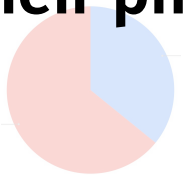


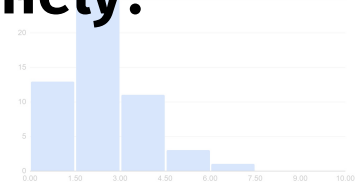
Optional Contents

Name	Lonely - yes or no?	Self-rated loneliness (on a scale of 1-5)	Avg screen time per day (hours)
Hafsah	No	2	3.1
Jackie	No	1	2.5
Aseel	No	1	3.5
Oscar	Yes	4	1.7
Claudia	Yes	5	4.2
Ziad	No	2	1.5

Univariate visualization is only useful to a certain extent.

For example, it cannot answer our previous question “are people who spent >4 hours on their phone lonely or not lonely?”

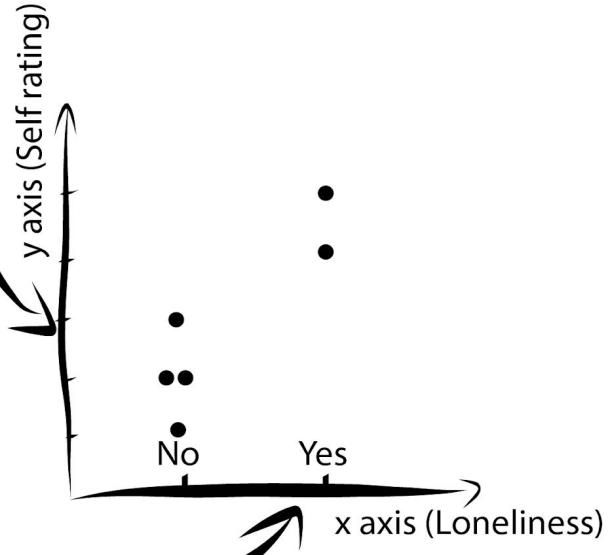




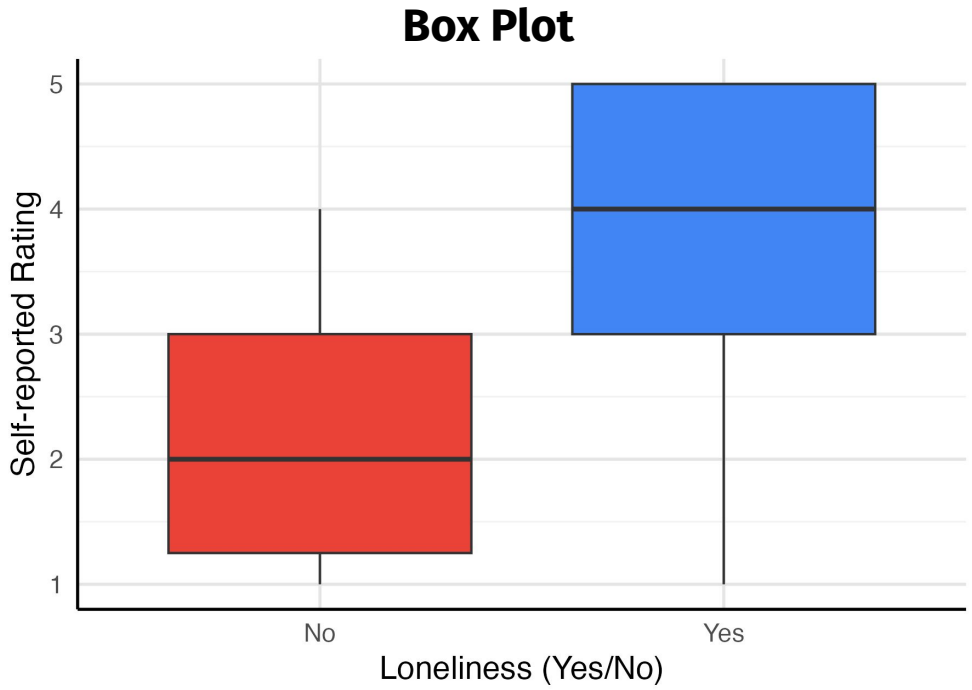
We took each column, visualized it separately. This is called **univariate** visualization.

Bivariate data visualization can help!

Name	Lonely - yes or no?	Self-rated loneliness (on a scale of 1-5)
Hafsah	No	2
Jackie	No	1
Aseem	Yes	5
Oscar	Yes	4
Claudia	No	3
Zion	No	2

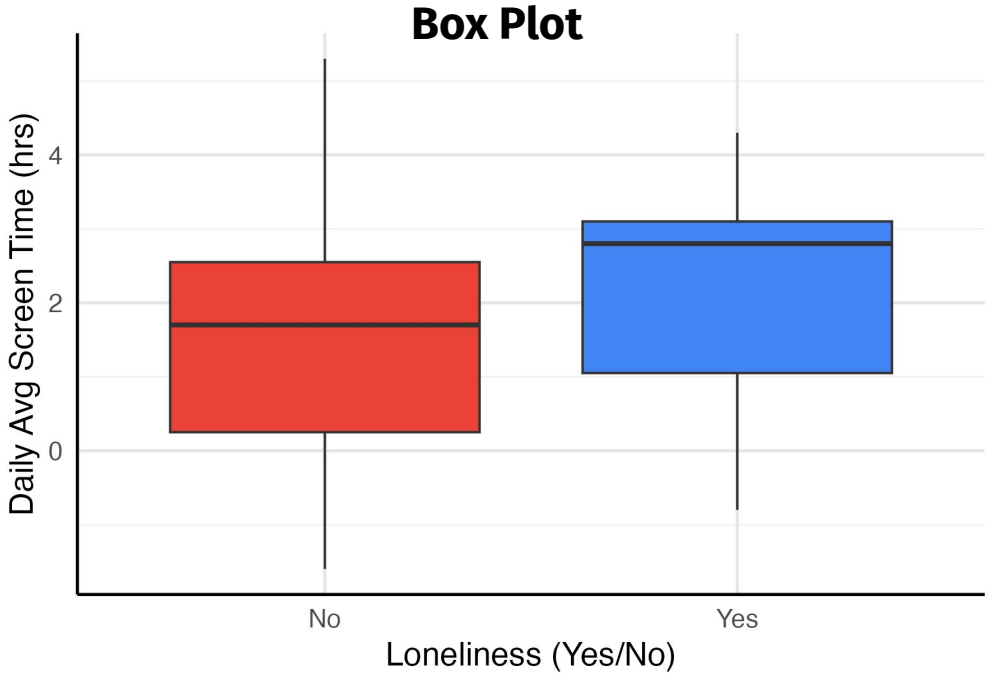


Name	Lonely - yes or no?	Self-rated loneliness (on a scale of 1-5)
Hafsah	No	2
Jackie	No	1
Aseem	Yes	5
Oscar	Yes	4
Claudia	No	3
Zion	No	2



Box plots are useful to compare between groups.
The center line in each box represents median (we will cover later this semester).
Overall, this plot tells us that people who answered “Yes” to being lonely, also self-reported more loneliness.

Name	Lonely - yes or no?	Avg screen time per day (hours)
Hafsah	No	3.1
Jackie	No	2.5
Aseem	Yes	3.9
Oscar	Yes	1.7
Claudia	No	2.1
Zion	No	1.3



We can interpret this box plot similar to the last one.
This plot tells us that people who answered “Yes” to being lonely, also had longer screen time.
Note: you do not have to completely understand this graph, but you will be required to differentiate between univariate vs. bivariate visualizations.