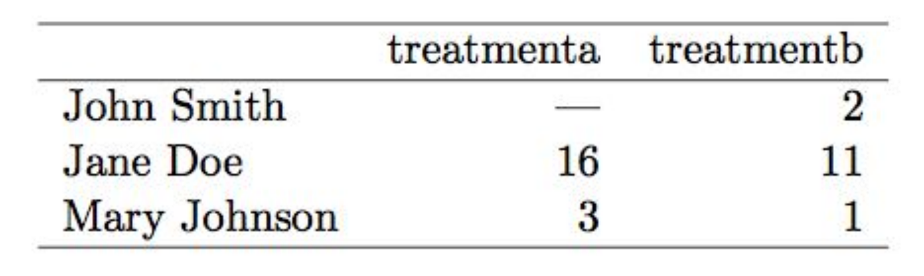
Data Visualization

Quiz

- Explain the importance of data types and why we may not want to use Excel in our data pipelines.
- What’s the definition of the tidy data?
- Is this data table tidy? Why? Why not? Can you make it tidy?



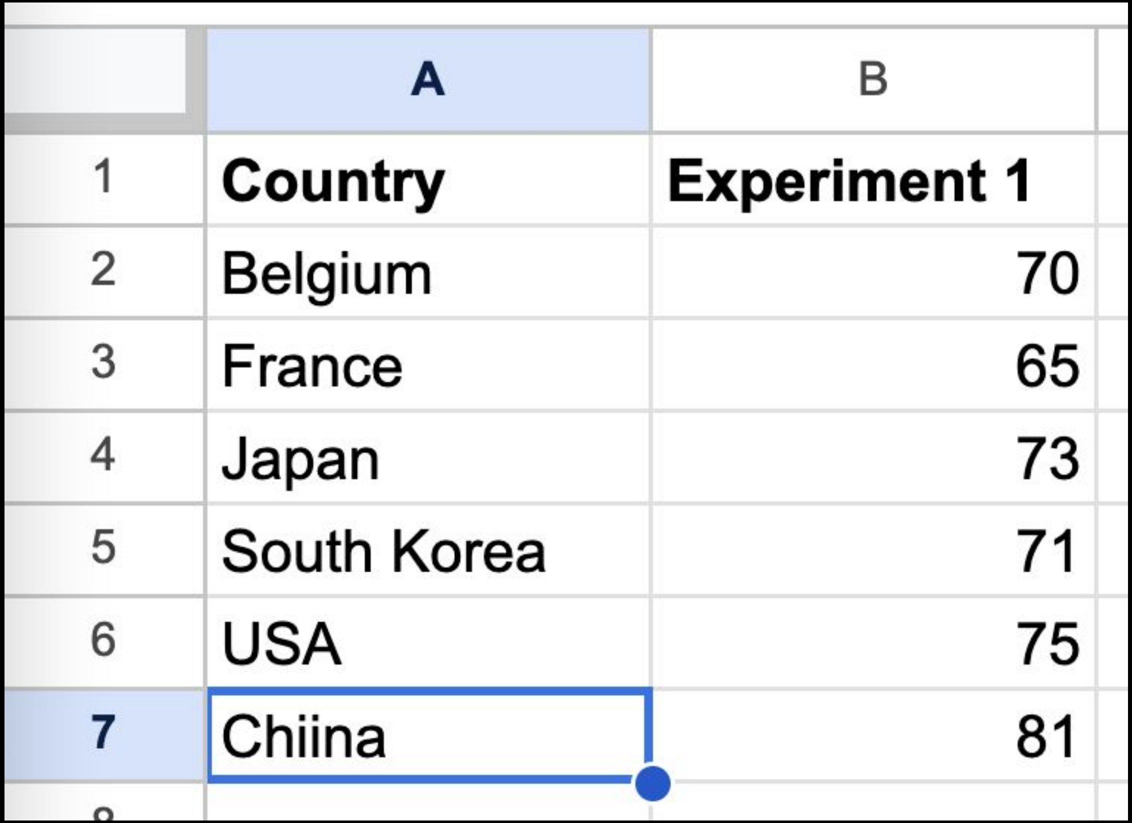
	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Part II

Data Science 💔 Excel

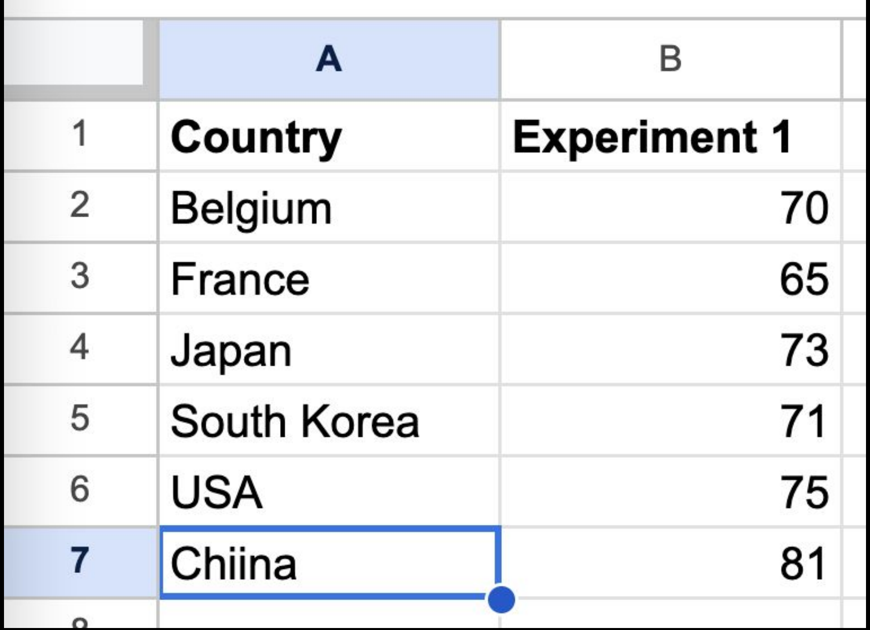
😭

You've just found some typos in country_data.csv you downloaded. What would you do?



	A	B
1	**Country**	**Experiment 1**
2	Belgium	70
3	France	65
4	Japan	73
5	South Korea	71
6	USA	75
7	Chiina	81
8		

You’ve just found some typos in a dataset. What would you do?



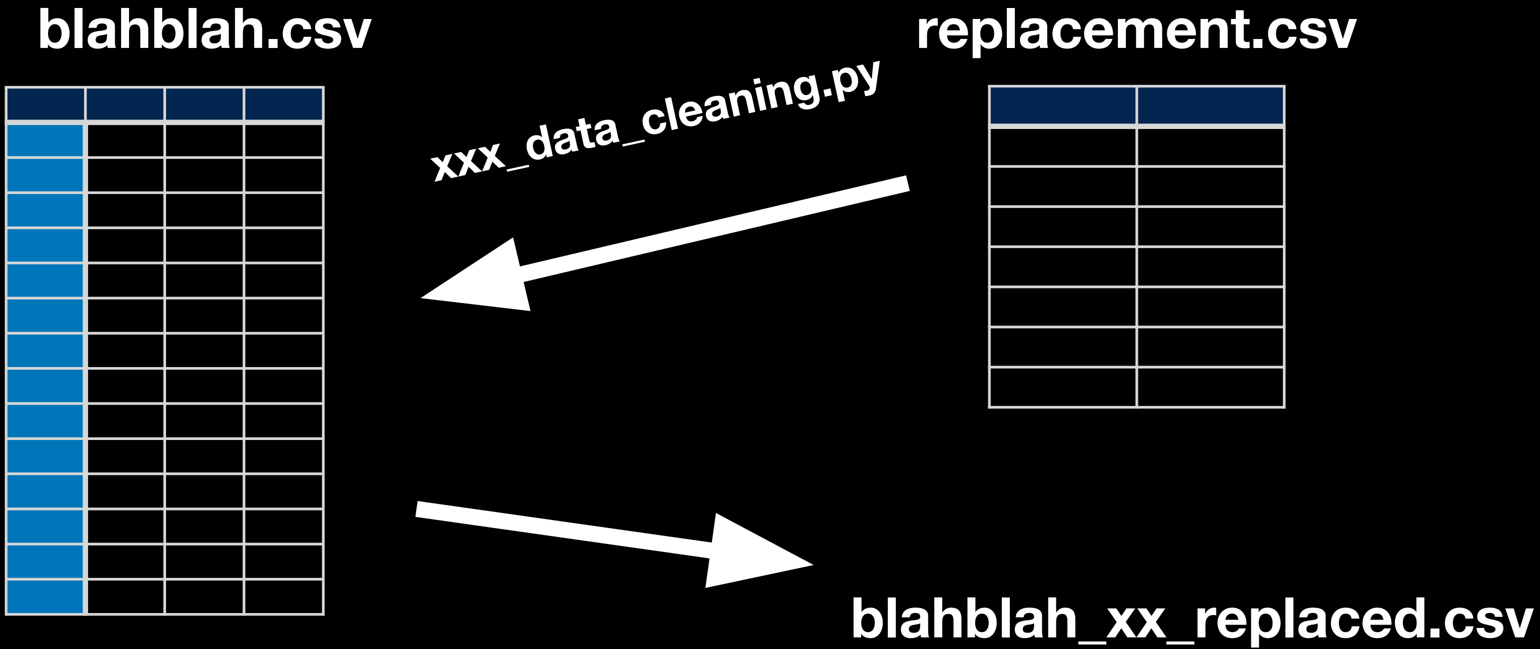
	A	B
1	**Country**	**Experiment 1**
2	Belgium	70
3	France	65
4	Japan	73
5	South Korea	71
6	USA	75
7	Chiina	81

Ok, let me quickly fix them by hand… 👨‍💻

What could be problems?

1. What if you introduce a different error?
2. What if you re-download the raw data?
3. What if you have to explain the process but you can’t remember?
4. What if it breaks the pipeline and you can’t remember exactly what you fixed?
5. What if someone else takes over your job?

Never handle your data manually &
Be explicit!



blahblah.csv

xxx_data_cleaning.py

replacement.csv

blahblah_xx_replaced.csv

Clean/process you data with auxiliary data files and scripts.

- The code & auxiliary data serves as concrete documentation of the process (provenance).
- Easier to spot errors.
- You can simply re-apply (or improve) the script/data when you have a new version of the raw dataset.
- Automation! 🤖
- Others can catch up quickly and test/check the process (more 👀 is good)

Data Provenance

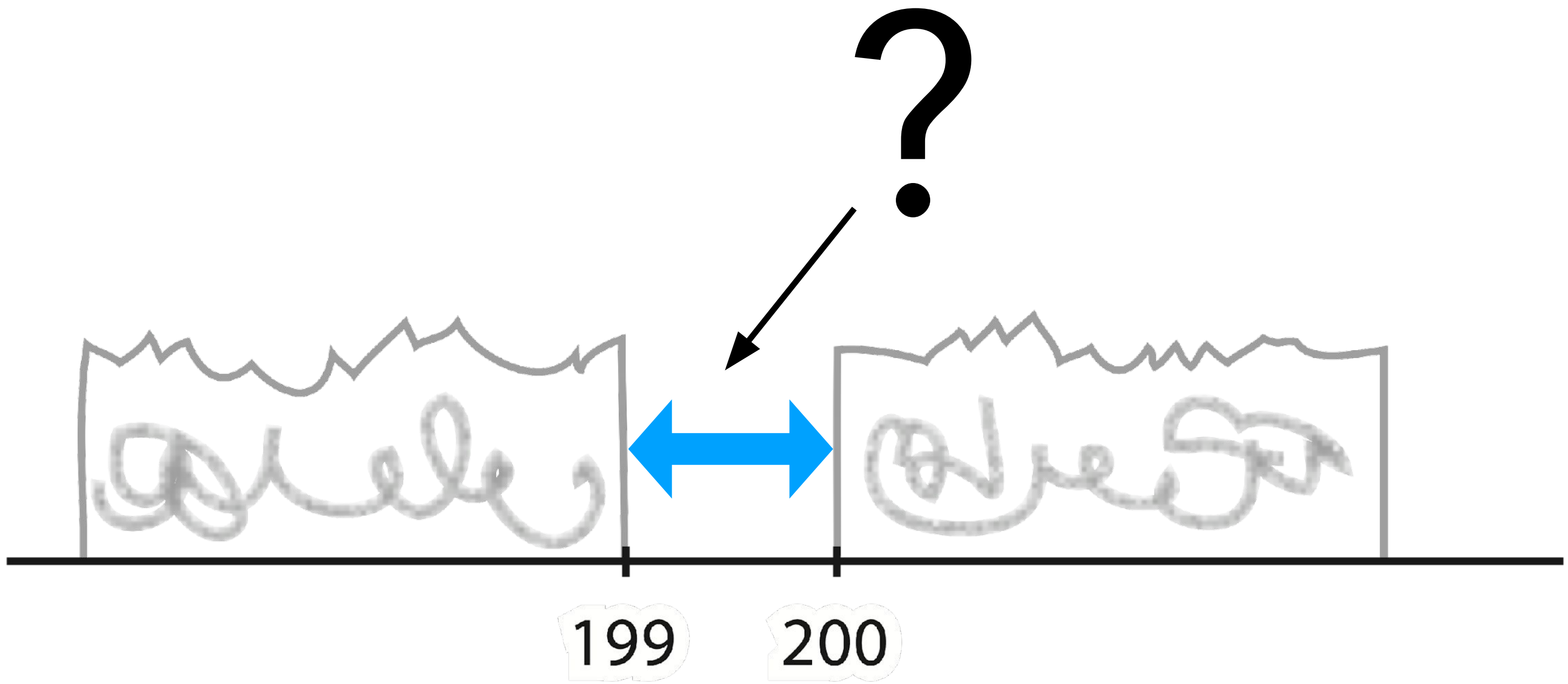
EXCEL
PYTHON
imgflip.com

Histogram

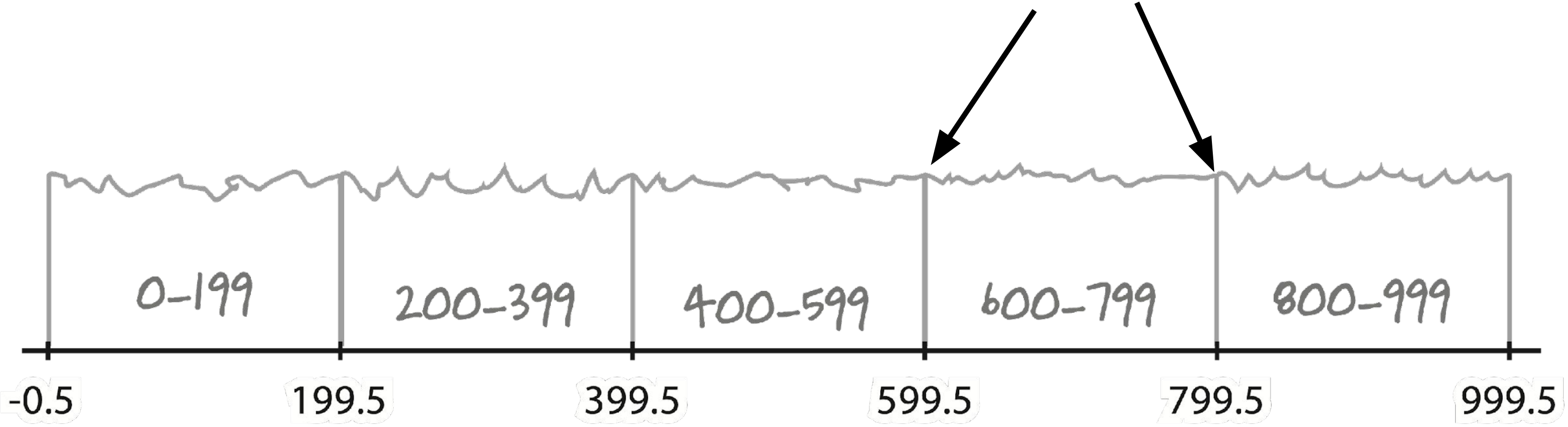
Draw a histogram

Score	Frequency
0-199	5
200-399	29
400-599	56
600-799	17
800-999	3

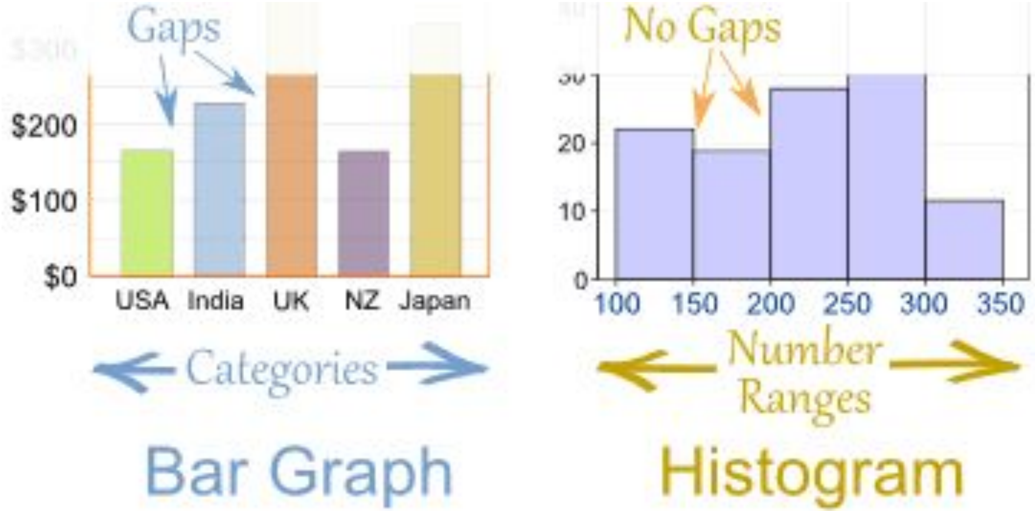
Head First Statistics


?
199
200

No gaps!


0-199
200-399
400-599
600-799
800-999
-0.5
199.5
399.5
599.5
799.5
999.5

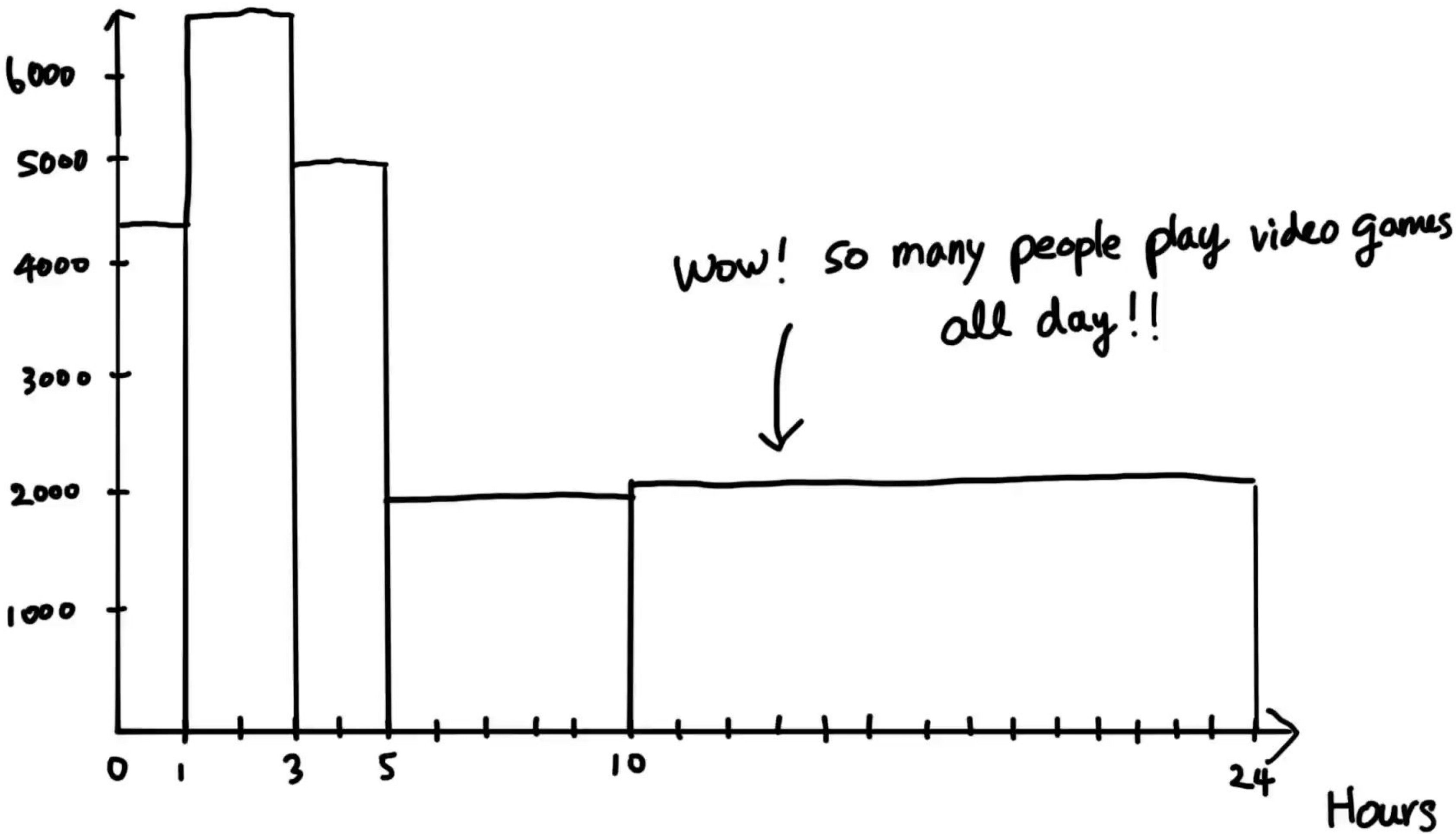
Bar graph vs. histogram

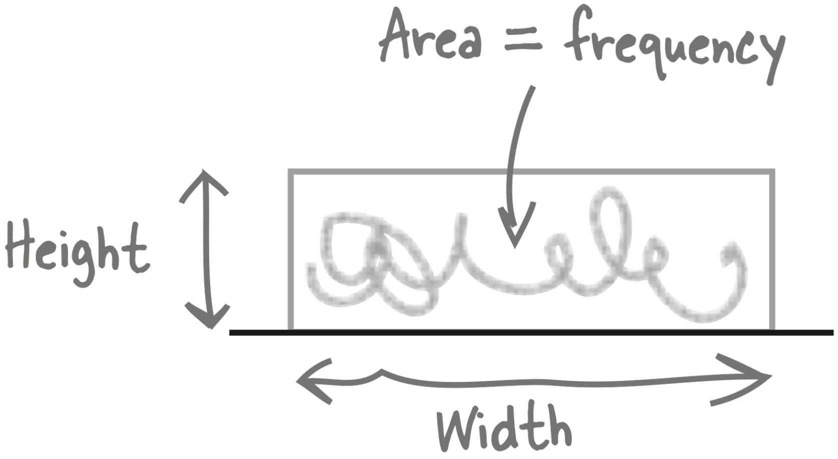


Draw a histogram

Hours	Frequency
0-1	4,300
1-3	6,900
3-5	4,900
5-10	2,000
10-24	2,100

Head First Statistics


6000
5000
4000
3000
2000
1000
0
1
3
5
10
24
Hours
Wow! so many people play video games all day!!



Hours	Frequency	Width	Height (Frequency Density)
0–1	4,300	1	4,300 ÷ 1 = 4,300
1–3	6,900	2	6,900 ÷ 2 = 3,450
3–5	4,900	2	4,900 ÷2 = 2,450
5–10	2,000	5	2,000 ÷ 5 = 400
10–24	2100	14	2,100 ÷ 14 = 150

Head First Statistics

In histogram, area, not the height, represents the frequency!