Intro to Challenge 1

Data Sets

What you need to know

Challenge 1: Data Sets

Status: Challenge Complete

What to do next: All Done!

1. Get the Skills you Need

Study the practice questions in these levels. Mark them "Ready for Challenge" whne you are done.

Python skill level(s) you need

- ✓ **Level 1**: What's Programming?
- ✓ **Level 2**: In and out of Python

Data Science skill level(s) you need

- ✓ **Level 1**: What's Data Science?
- ✓ **Level 2**: Whats a data set?

Statistics skill level(s) you need

- ✓ **Level 1**: What's Statistics?
- ✓ **Level 2**: What's Data?

What you do

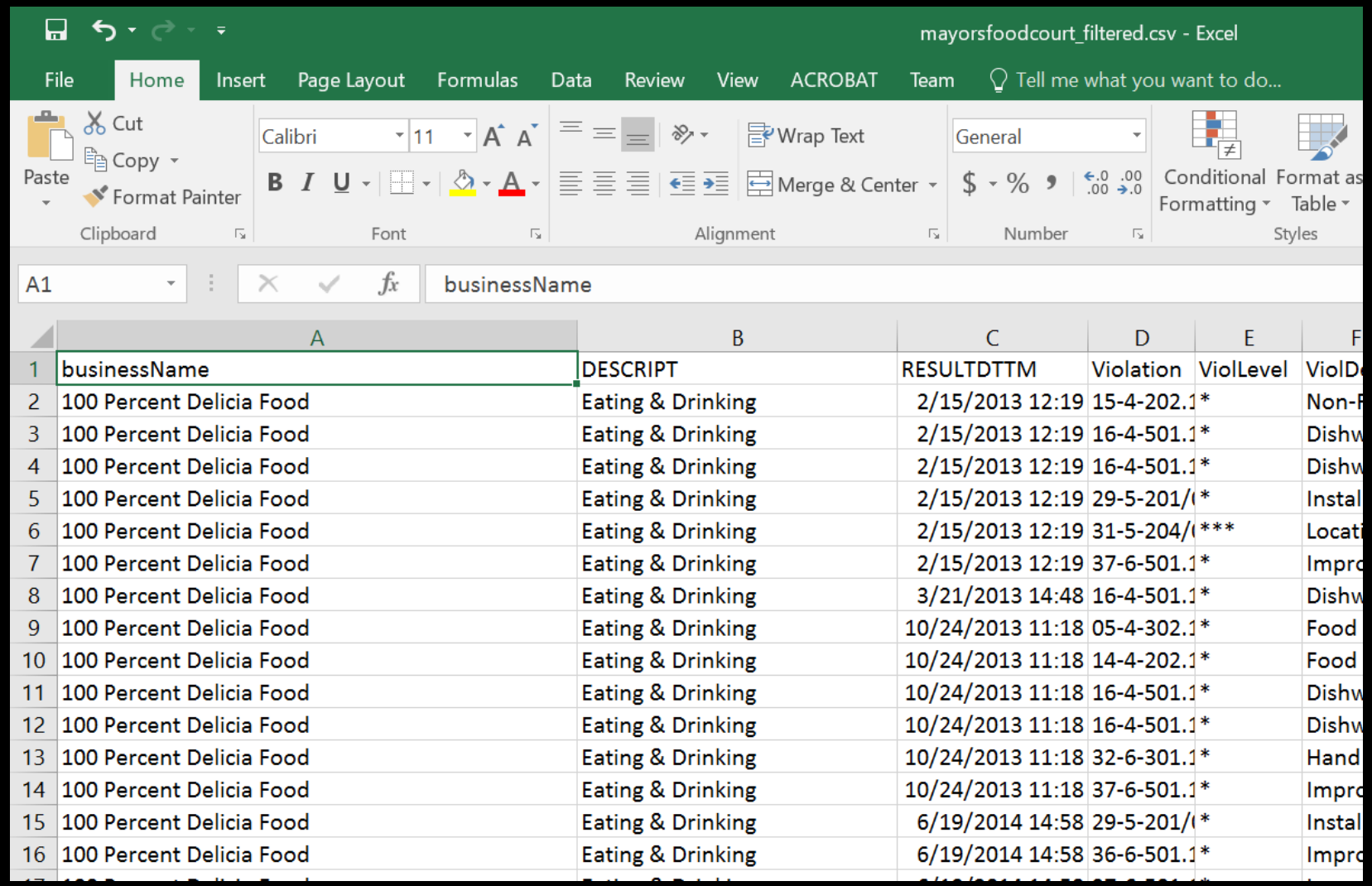
1. Download and open a data set
2. Talk about the social context of the data
3. Talk about the data itself
4. Do a data analysis to find answers

Key Data Science Points

- What is in data science that is not in stats and python?
- What is a data context?
- What is a data set?

Key Stats Points

- What is Data?
- What's a data type?
- What is a table
- Row
- Column



	A	B	C	D	E	F
1	businessName	DESCRIPT	RESULTDTTM	Violation	ViolLevel	ViolD
2	100 Percent Delicia Food	Eating & Drinking	2/15/2013 12:19	15-4-202.1	*	Non-F
3	100 Percent Delicia Food	Eating & Drinking	2/15/2013 12:19	16-4-501.1	*	Dishw
4	100 Percent Delicia Food	Eating & Drinking	2/15/2013 12:19	16-4-501.1	*	Dishw
5	100 Percent Delicia Food	Eating & Drinking	2/15/2013 12:19	29-5-201/(	*	Instal
6	100 Percent Delicia Food	Eating & Drinking	2/15/2013 12:19	31-5-204/(	***	Locati
7	100 Percent Delicia Food	Eating & Drinking	2/15/2013 12:19	37-6-501.1	*	Impro
8	100 Percent Delicia Food	Eating & Drinking	3/21/2013 14:48	16-4-501.1	*	Dishw
9	100 Percent Delicia Food	Eating & Drinking	10/24/2013 11:18	05-4-302.1	*	Food
10	100 Percent Delicia Food	Eating & Drinking	10/24/2013 11:18	14-4-202.1	*	Food
11	100 Percent Delicia Food	Eating & Drinking	10/24/2013 11:18	16-4-501.1	*	Dishw
12	100 Percent Delicia Food	Eating & Drinking	10/24/2013 11:18	16-4-501.1	*	Dishw
13	100 Percent Delicia Food	Eating & Drinking	10/24/2013 11:18	32-6-301.1	*	Hand
14	100 Percent Delicia Food	Eating & Drinking	10/24/2013 11:18	37-6-501.1	*	Impro
15	100 Percent Delicia Food	Eating & Drinking	6/19/2014 14:58	29-5-201/(	*	Instal
16	100 Percent Delicia Food	Eating & Drinking	6/19/2014 14:58	36-6-501.1	*	Impro

Key Python Points

- import pandas as pd
- dataFrame = pd.read_csv('', sep="")
- print()
- dataFrame.head()

Intro to Challenge 2

Describing Data

What you need to know

Challenge 2: Describing Data

Status: Challenge Complete

What to do next: All Done!

1. Get the Skills you Need

Study the practice questions in these levels. Mark them "Ready for Challenge" when you are done.

Python skill level(s) you need

- ✓ **Level 3**: Making Calculations in Python
- ✓ **Level 4**: Python Data Structures

Data Science skill level(s) you need

- ✓ **Level 3**: Describing Data: Google Trends

Statistics skill level(s) you need

- ✓ **Level 4**: Working with Graphs and Charts
- ✓ **Level 3**: Describing Data

What you do

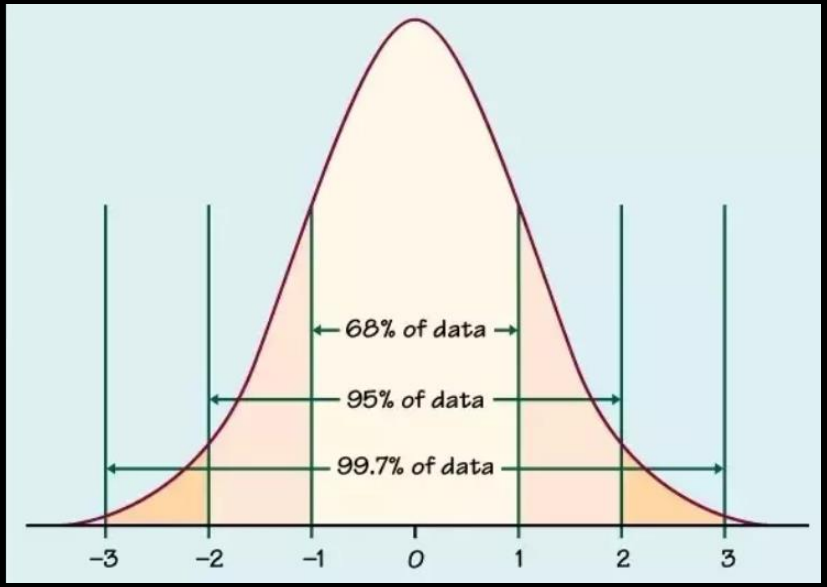
1. Download and open a data set
2. Talk about the social context of the data
3. Talk about the data itself
4. Do a data analysis to find answers

Key Data Science Points

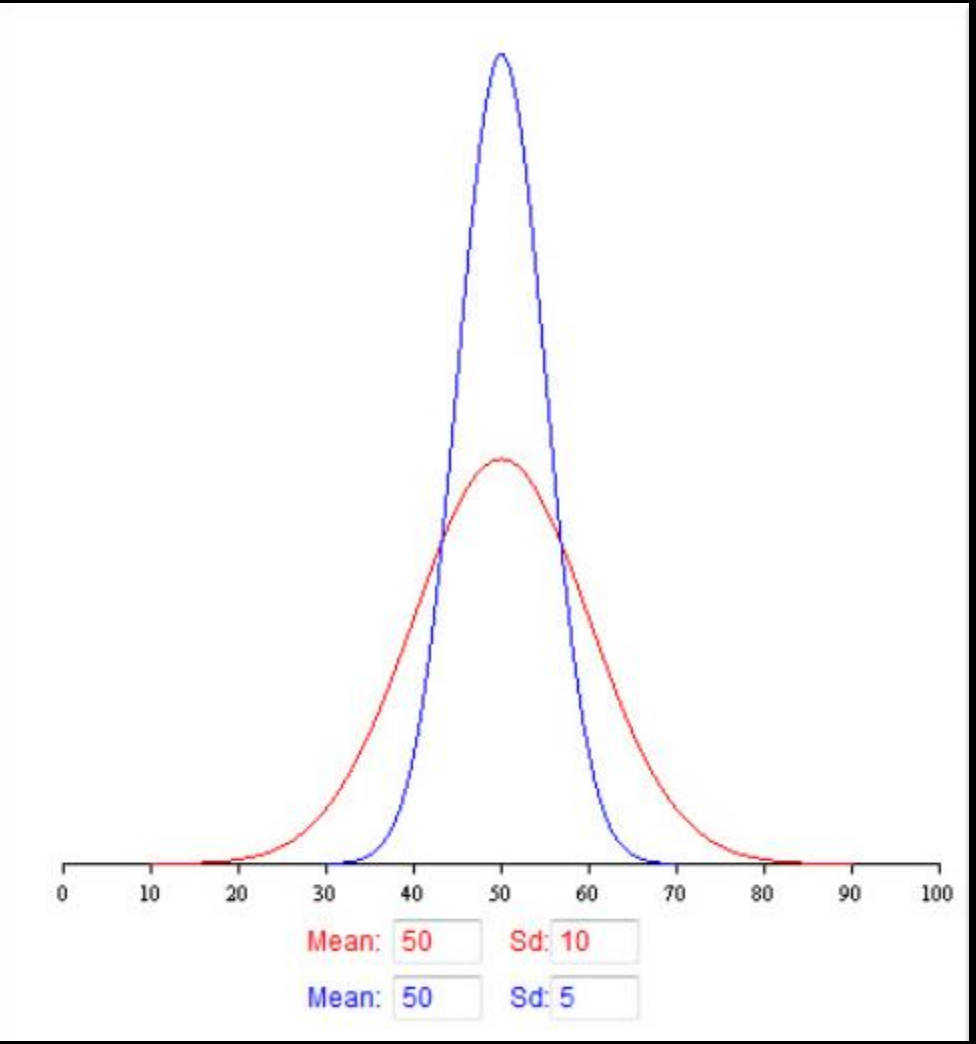
- Bias
- Normalized
- Anonymized
- Missing values
- Google Trends

Key Stats Points

- Central tendency
- Mean
- Standard deviation



Quora



OnlineStatBook.com

Key Python Points

- df = pd.read_csv('')
- meanAsNumber = df[''].mean()
- meanAsMoney = "${???}".format(meanAsNumber)

Intro to Challenge 3

Data Tables

What you need to know

Challenge 3: Data Tables

Status: Qualifying Exam Open

What to do next: Click the button below to take your Qualifying Exam

1. Get the Skills you Need

Study the practice questions in these levels. Mark them "Ready for Challenge" when you are done.

Python skill level(s) you need

- ✓ **Level 5**: Working with Tables
- ✓ **Level 6**: Functions in Python

Data Science skill level(s) you need

- ✓ **Level 4**: Data Tables: Auto Power

Statistics skill level(s) you need

- ✓ **Level 5**: Working with Categorical Data

What you do

1. Download and open a data set
2. Talk about the social context of the data
3. Talk about the data itself
4. Do a data analysis to find answers

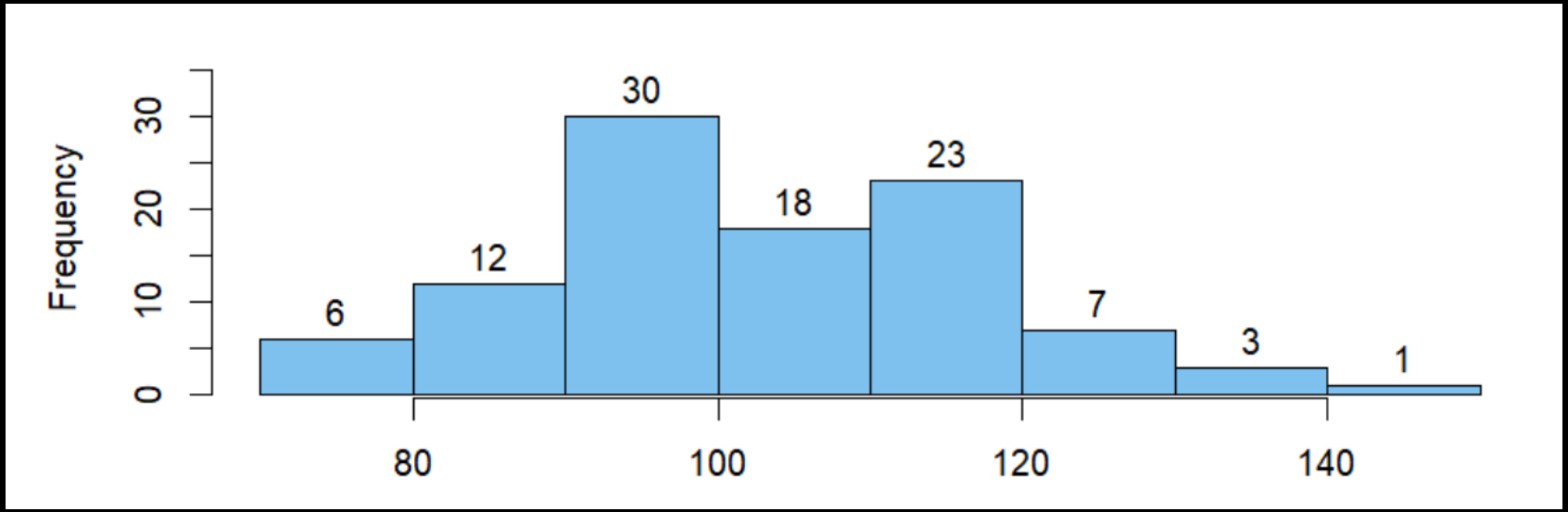
Key Data Science Points

- Data types
  - Categorical data vs. continuous data
  - Ints, floats and strings
- Bad Data

	A	B	C	D	E	F	G	H	I	J	K
1	mpg	cylinders	displaceme	horsepowe	weight	acceleratio	year	origin	name		
29	11	8	318	210	4382	13.5	70	1	dodge d200		
30	9	8	304	193	4732	18.5	70	1	hi 1200d		
31	27	4	97	88	2130	14.5	71	3	datsun pl510		
32	28	4	140	90	2264	15.5	71	1	chevrolet vega 2300		
33	25	4	113	95	2228	14	71	3	toyota corona		
34	25	4	98	?	2046	19	71	1	ford pinto		
35	19	6	232	100	2634	13	71	1	amc gremlin		
36	16	6	225	105	3439	15.5	71	1	plymouth satellite custom		
37	17	6	250	100	3329	15.5	71	1	chevrolet chevelle malibu		
38	19	6	250	88	3302	15.5	71	1	ford torino 500		
39	18	6	232	100	3288	15.5	71	1	amc matador		

Key Stats Points

- Counts vs Frequencies



https://math.stackexchange.com/questions/2666834/what-is-the-difference-between-frequency-and-density-in-a-histogram

- Two-way tables

	Sport Utility Vehicle (SUV)	Sports Car	Totals
male	21	39	60
female	135	45	180
Totals	156	84	240

MathBits.com

https://mathbitsnotebook.com/Algebra1/StatisticsReg/ST2TwoWayTable.html

Key Python Points: DataFrames(Tables)

```
df = pd.DataFrame({
        "strings" : ["1", "2", "3", "4", "5", "6"],
        "more strings" : ["1.1","1.2", "1.3", "1.4", "1.5", "1.6"]
})
```

```python
import numpy as np
import pandas as pd

trialNumberArray = np.arange(1,6,1)
initialSizeArray = np.arange(1,3.5,.5)
np.random.shuffle(initialSizeArray)
finalSizeArray = np.arange(2,4.5,.5)
np.random.shuffle(finalSizeArray)

dict = {
    'Trial Number': trialNumberArray,
    'Initial Size': initialSizeArray,
    'Final Size': finalSizeArray

}
experimentTable = pd.DataFrame(dict)
print("****Experiment table****")
print(experimentTable)
```

Create arrays (lists)

Add them to a dictionary

Turn the dictionary into a DataFrame

Key Python Points: Functions

Name of the function

Parameters

```
def calculateGrowth(initial, final):
    result = final - initial
    return result
```

Says I'm a function

What the function does

What the function give back

```
print(calculateGrowth(3, 10))
table["Growth"] = calculateGrowth(table["Initial Size"],table["Final Size"])
```

Intro to Challenge 4

Charts and graphs

What you need to know

Challenge 4: Data Charts and Graphs

Status: Not yet ready for challenge

What to do next: Study the skill levels and mark the as 'Ready for Challenge'

1. Get the Skills you Need

Study the practice questions in these levels. Mark them "Ready for Challenge" when you are done.

Python skill level(s) you need

- ✗ **Level 7**: Visualizing Bar Charts, Line Graphs and Scatter Plots
- ✗ **Level 8**: Grouping in Python

Data Science skill level(s) you need

- ✗ **Level 5**: Charts and Graphs: Journal Articles

Statistics skill level(s) you need

- ✓ **Level 5**: Working with Categorical Data

What you do

1. Download and open a data set
2. Talk about the social context of the data
3. Talk about the data itself
4. Prepare the data
5. Do a data analysis to find answers
6. Report your results

Key Data Science Points: Credibility

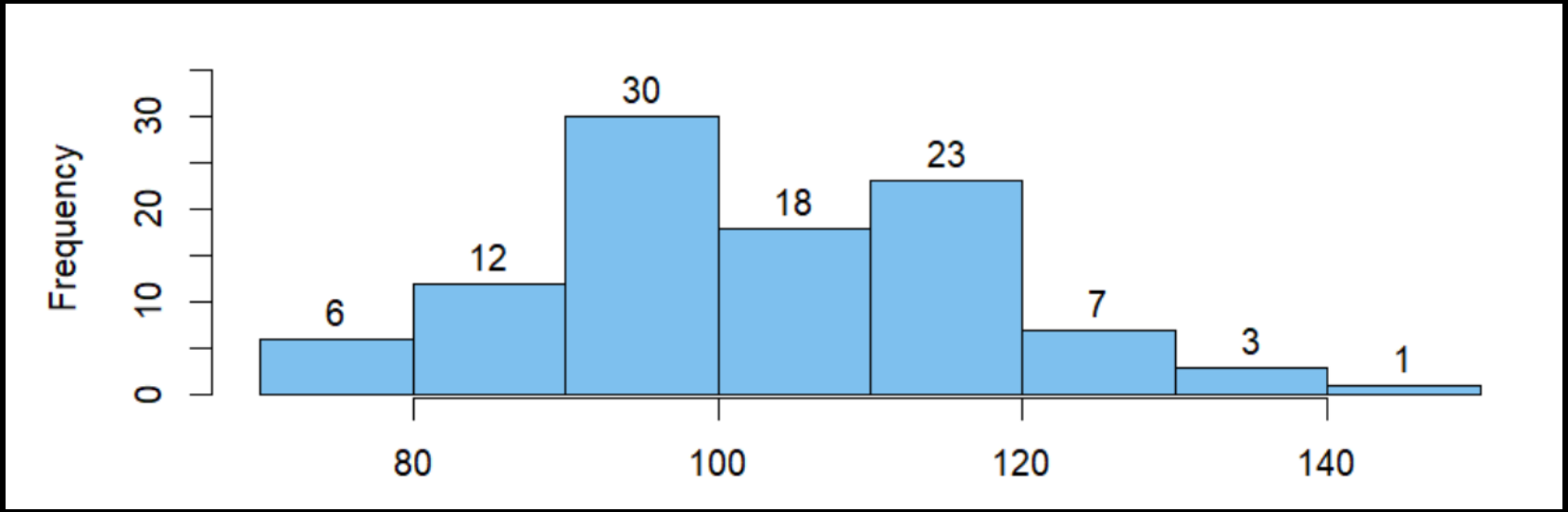
- Creators.
  - Credible data sources clearly specify who collected the data. The collectors are established and reputable. Judge the creators of this data set.
- Providers.
  - Credible data sources are provided by established and reputable organizations. Judge the data provider.
- Intentions.
  - The creators and providers of credible data sources have no motivation to deceive. Judge the motivations of the creators and providers.
- Expertise.
  - Credible data creators have the data collection skills needed to create unbiased and complete data. Judge the expertise of these creators.
- Methods.
  - Credible data has well documented industry standard collection methods. Judge the collection methods of these creators.

Key Data Science Points: Data prep

- Filter columns
- Filter rows
- Rename columns
- Turn column numbers into strings
- Find and kill bad data
  - Nulls
  - Garbage
  - Impossible values

Key Stats Points – No new levels

- Counts vs Frequencies



https://math.stackexchange.com/questions/2666834/what-is-the-difference-between-frequency-and-density-in-a-histogram

- Two-way tables

	Sport Utility Vehicle (SUV)	Sports Car	Totals
male	21	39	60
female	135	45	180
Totals	156	84	240

MathBits.com

https://mathbitsnotebook.com/Algebra1/StatisticsReg/ST2TwoWayTable.html

Key Python Points: Bar Charts

```
import matplotlib
%matplotlib inline
```

Where the real work happens

```
plot1 = femaleMale.plot(kind="barh")
plot1.set_xlabel("Number of responders")
plot1.set_title("Gender of responders in top 10")
```

Set up the plot

```
print("****Three Variables****")
voting["Candidate 2"] = candidate2
voting["Candidate 3"] = candidate3
print(voting)
plot2 = voting.plot(kind="bar")
plot2.set_ylabel("% Voting for Candidate")
plot2.set_xlabel("Age of Voter")
plot2.set_title("Three Variable Plot")
```

Multiple plots from multiple columns

Key Python Points: Line Charts

```
import matplotlib
%matplotlib inline
```

Where the real work happens

```
plot1 = airQuality.plot(kind="line", x="Time", y="Carbon Monoxide")
plot1.set_title("Carbon Monoxide Over Time")
plot1.set_xlabel("Time (24-hour)")
```

Set up the plot

Multiple plots from multiple rows

```
plot1 = airQuality.plot(kind="line", x="Time",  y=["Carbon Monoxide","Hydrocarbons","Nitrogen Dioxide"])
plot1.set_title("All Gasses Over Time")
plot1.set_xlabel("Time (24-hour)")
```

Data Science Level 3

Describing Data

What's normalization?

1. What's the average number of calories a person ***should*** eat?
2. (2600+2000+2200+2000+1600+1400)/6 = 1966?

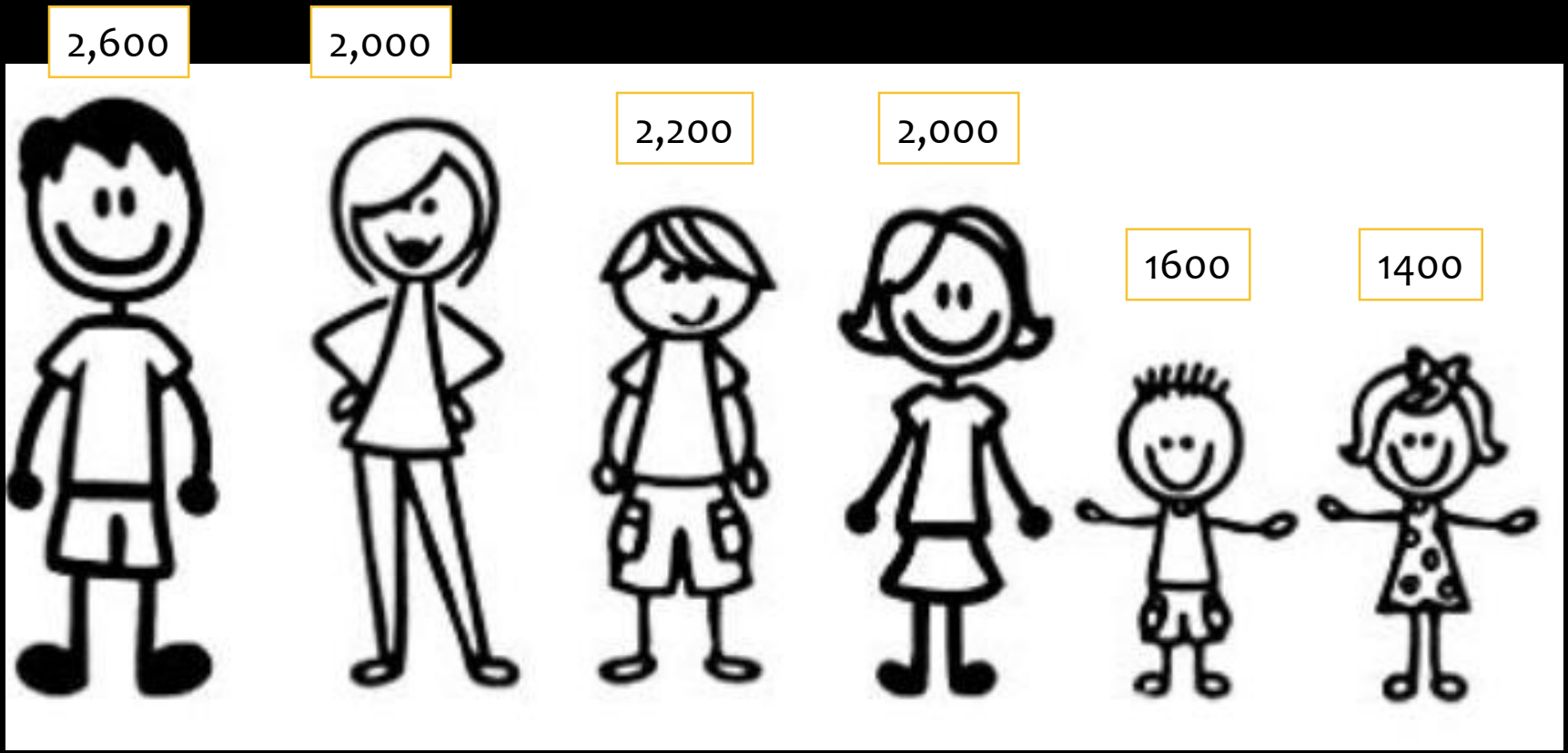
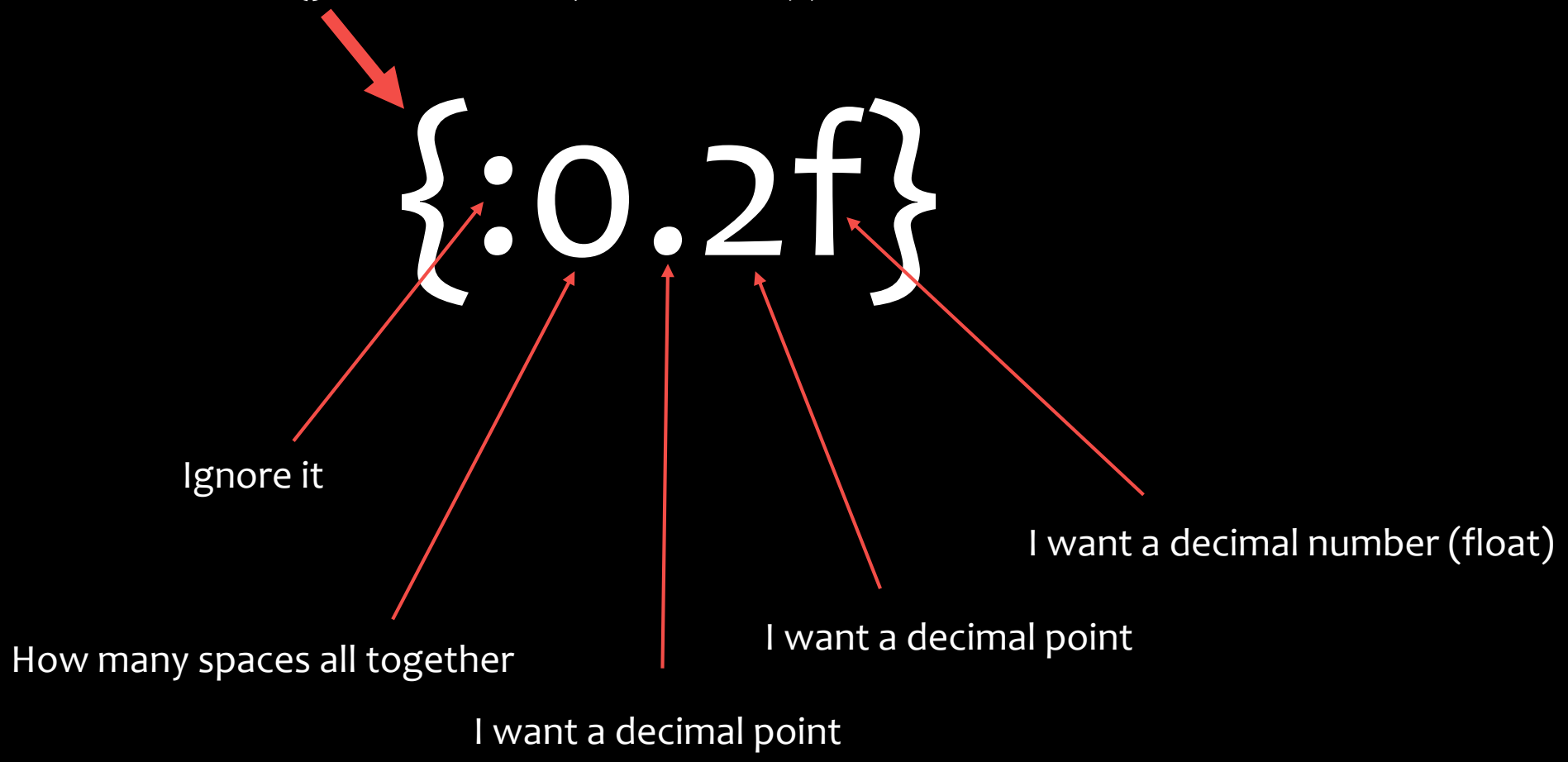


Table drill: Open a new Jupyter notebook

- Do it!
  - Name it practiceFormatString
  - Enter
    - number = 897.654
- Once you got it done, help anyone else at your table get it done
- The first table done gets a bump for all

Personal Drill: formatting

Print(“some text {}”.format(variable))



Personal drill: Write the script to make this happen

```
897.657
The number is: 897.657
The integer number is: 898
The number as money is: $897.66
The number with 10 spaces is: 897.657
```