DATA 101

Ides of February

We have to have an EXAM

Not Today Obviously

Next Week...72 hours: your choice

‘My Homework’

My Homework Walkthrough

Kicking it over to Jupyter

Correlation ideas from media - good and bad

A person who is currently renting and whose Facebook friends saw their homes appreciate 5 percent more than the market average in the past two years is 3.1 percentage points more likely to buy a home themselves in the next two years.

To put that in perspective, that increased likelihood is about half the size of the effect of having a child — one of the major life events that people typically tie to house ownership.They also buy a 1.7 percent larger house, pay 3.3 percent more for a given house and make a 7 percent larger downpayment, the researchers found.

So what's going on here? By controlling for different demographic characteristics, researchers said they discovered it's not just groups of people coming of age at the same time. Likewise, clustered occupations are controlled for, so positive shocks to a certain industry (e.g., tech) are accounted for.

"I think this is leading to that sense of envy and what the millennial generation calls 'fomo,' or 'fear of missing out,'" Lawrence Yun, chief economist with the National Association of Realtors... "If someone's friends are doing well, they fear they're missing out and they want to be more active in the home-buying process."

An Aside

For those of you interested:

https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

March Project ~ 7-10 Minute Presentation

WHAT IS EXPECTED

- Choose GREAT data
  - GREAT == You care about it
  - GREAT == Has {appropriate, interesting, workable} FEATURES
  - Data is robust in FEATURES or Volume
- Determine INTERESTING *Business Questions*
  - Set out to look for something you care about or find intriguing
  - ITERATE! Use initial findings to go down the rabbit hole
- Formulate INCREDIBLE conclusions and insights
  - Visual & Qualitative TAKEAWAYS - you actually think others should **‘care and share’**
- Spend time writing GOOD CODE (Ask Questions)
  - Don’t take shortcuts on quality non-code work, because you didn’t put the time in
  - Create incredible visuals with key insights, means setting up a good coding ‘test bed’

March Project - You are presenting to The Board

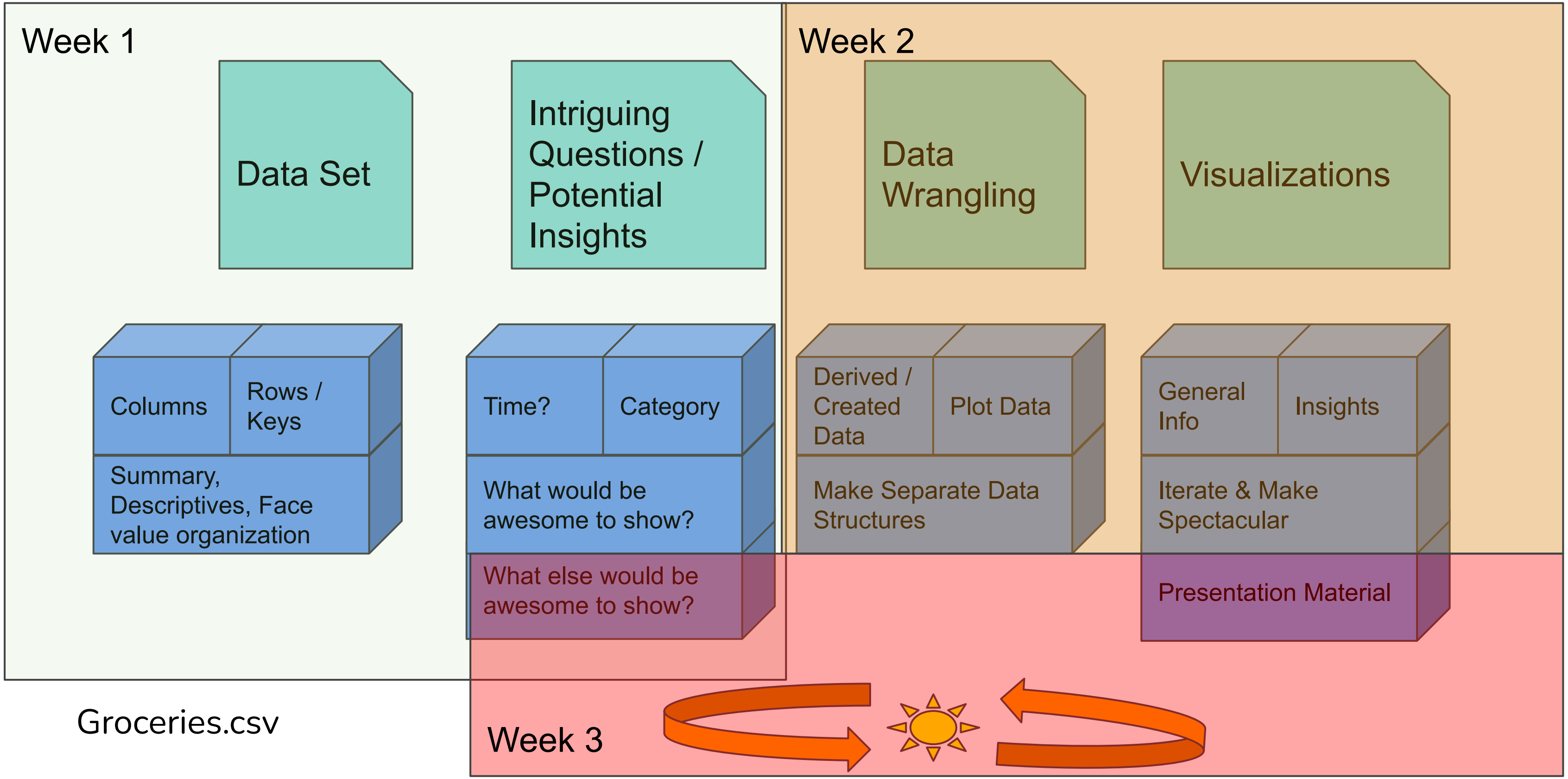
A SUGGESTED AGENDA FOR YOU TO FOLLOW

- Choose GREAT data, Determine INTERESTING *Business Questions*, & Sharpening HW
  - Week of Feb 15-21
- Data Munging, Wrangling, & Exam
  - Week of Feb 22 - Feb 28
  - ITERATE! Weeks of March 1 - 15
- Formulate INCREDIBLE conclusions and insights - Steve’s Presentation
  - Weeks of March 8 -15
- Write CODE - Let me reiterate everything...
  - Do your Exploratory Data Analysis in FEB
  - Set yourself up to do A LOT of interesting discovery, try different segmentations, insights, correlations, visual inspection, ITERATE March 1-7ish
  - Produce visualizations, stories, presentation talk track and presentation March 8-15

Subjectivity

PTS? TBD.	Low	Average	High	X Factor
Coding	Straightforward dataset, little to no derivative data, minimal ‘munging’ attempts	Some *derivative data created*, wrangling attempted and succeeded, data structures used	Several creative data slicing and filtering methods, m*uch of the used data was derived or created*	Did you have to? Comparison? Roadblocks conquered?
Math / Algorithms	Count, sum, max, min, etc.	Trends, rate of change, %, mean	Probability, covariance, standard deviation, Naive Bayes, Regression, Clustering	Did you use these ‘correctly’ and ‘interestingly’ or just use them to use them?
Analysis	Tell us about the above	Opinions and insights on why, call outs to external events or knowledge	Correlated nuggets of information coming from *derived data creation*, external callouts, attempting to answer why, how, potential cause and effects	Is this information you would not have guessed was the case anyway? Why is this ah ha? Shareable? Comparable?
X Factor	Basic plots of the above, presentation skills	The right plots, appropriate and informative information within, presentation skills	Incredibly insightful and visually appealing plots, incredible presentation techniques	**Comparison**. Take away story? Was there a surprise?

Project Planning - My Week 1+



Week 1

Data Set

Intriguing Questions / Potential Insights

Columns

Rows / Keys

Summary, Descriptives, Face value organization

Time?

Category

What would be awesome to show?

What else would be awesome to show?

Week 2

Data Wrangling

Visualizations

Derived / Created Data

Plot Data

Make Separate Data Structures

General Info

Insights

Iterate & Make Spectacular

Presentation Material

Week 3

Groceries.csv

March Project - Last Notes

Algorithms / Concepts Covered by March 8th

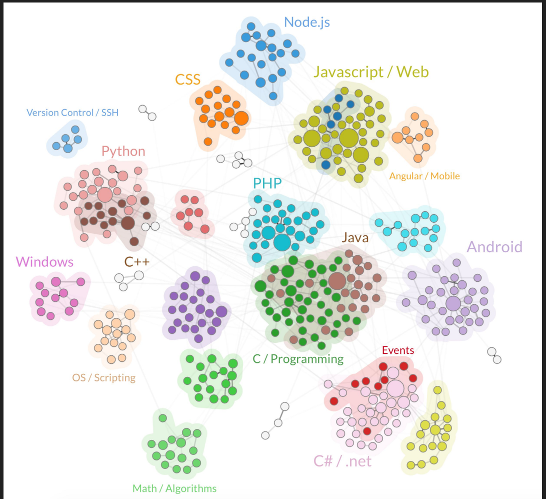
Clustering and Regression

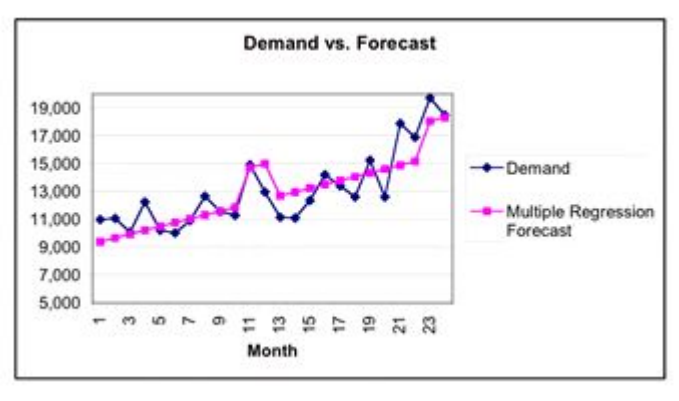
Presentations by np.random.choice(students)

- March 15 (submissions)...some March 22

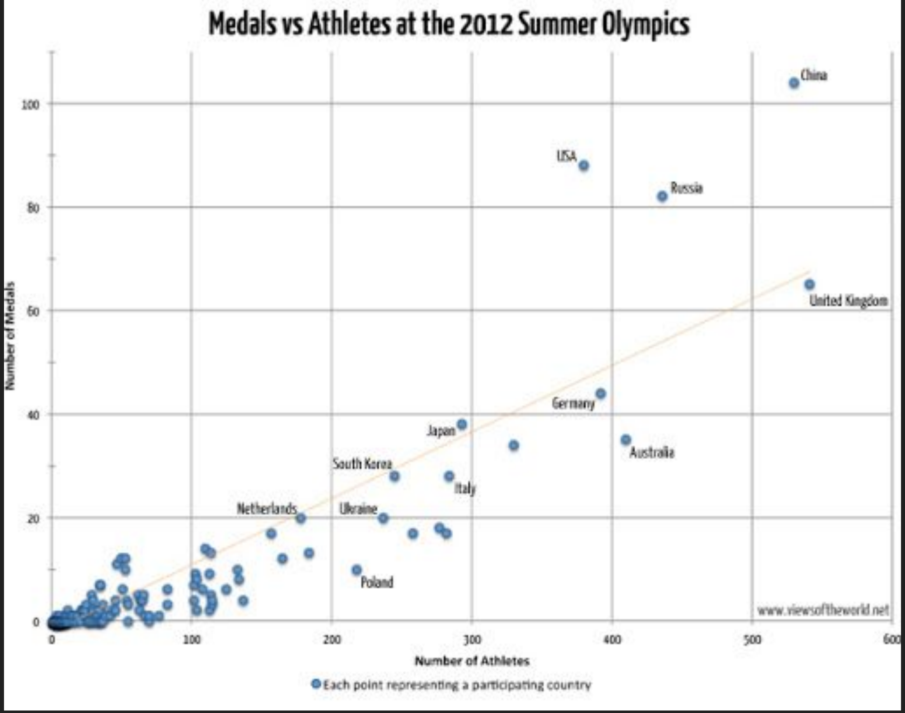
NOTES:

1. Really get into what the data/columns mean
2. Enumerate how you could roll up the data
3. Location, time, loop through each column…
4. Off the top of my head...12 Hours



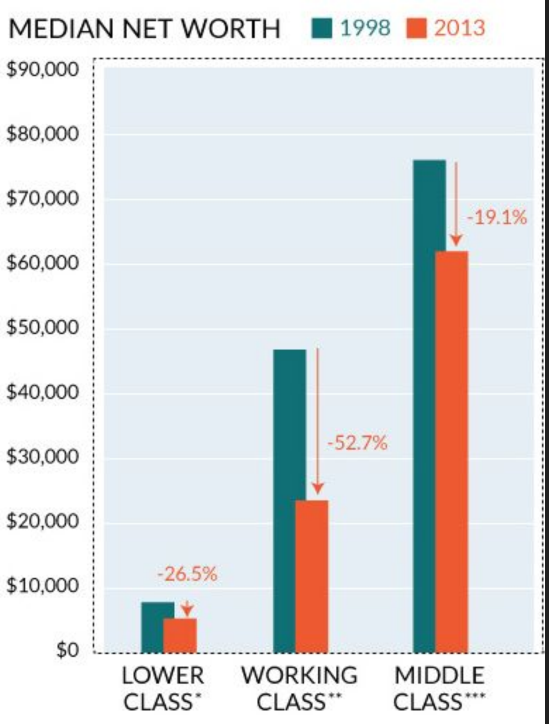


Simple scatter plot shows 2 variables plotted against each other often showing ‘subjective correlation’



- Athletes on X-axis
- Count of medals on Y-axis
- Title, labels, scale

Simple bar chart with (assuming) grouped by data by some definition by the author

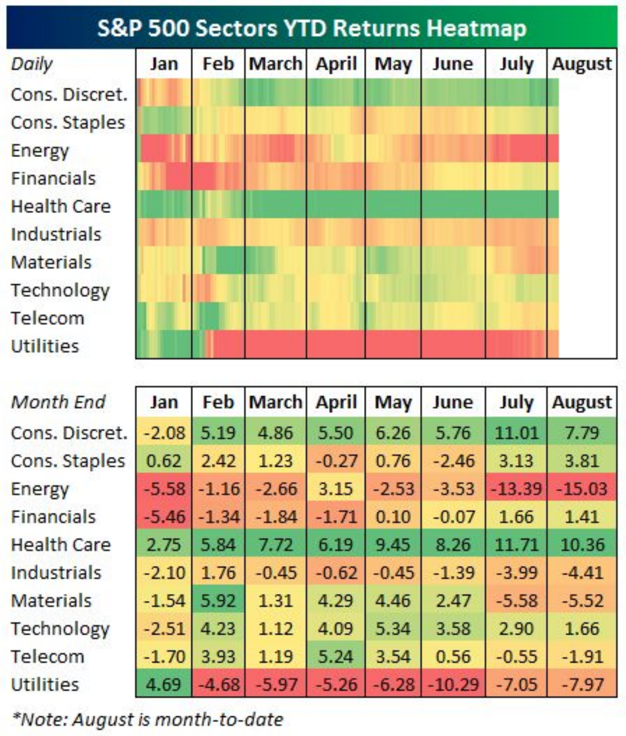


- Pivot on X-axis
- Median value on Y-axis
- Title, labels, scale

- Both axis are most likely derived data, meaning they are most likely ‘data.groupby’ operations in our terms

Many folks say the ‘rule’ of presenting data is the reader understand and take away your objective with only visuals

Heat maps generate ‘3D’ data, when you really want a third *aggregate* with two variables



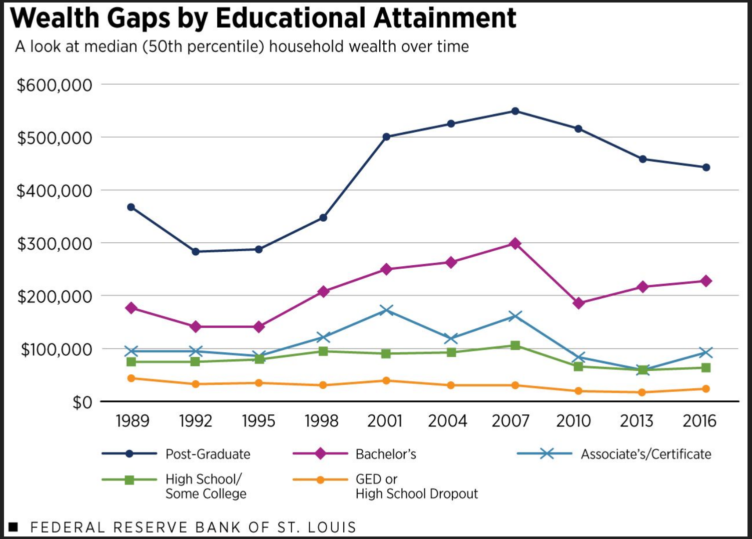
S&P 500 Sectors YTD Returns Heatmap

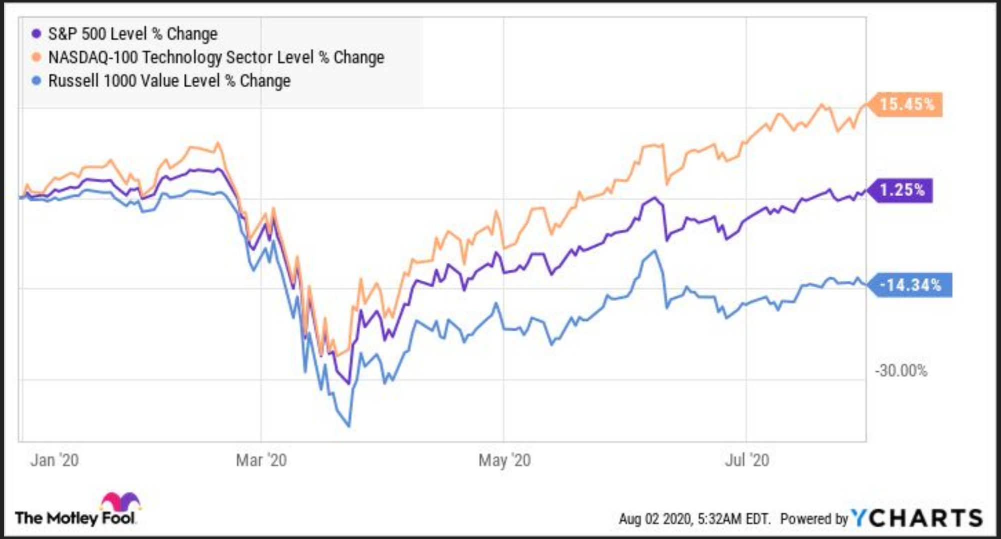
Month End	Jan	Feb	March	April	May	June	July	August
Cons. Discret.	-2.08	5.19	4.86	5.50	6.26	5.76	11.01	7.79
Cons. Staples	0.62	2.42	1.23	-0.27	0.76	-2.46	3.13	3.81
Energy	-5.58	-1.16	-2.66	3.15	-2.53	-3.53	-13.39	-15.03
Financials	-5.46	-1.34	-1.84	-1.71	0.10	-0.07	1.66	1.41
Health Care	2.75	5.84	7.72	6.19	9.45	8.26	11.71	10.36
Industrials	-2.10	1.76	-0.45	-0.62	-0.45	-1.39	-3.99	-4.41
Materials	-1.54	5.92	1.31	4.29	4.46	2.47	-5.58	-5.52
Technology	-2.51	4.23	1.12	4.09	5.34	3.58	2.90	1.66
Telecom	-1.70	3.93	1.19	5.24	3.54	0.56	-0.55	-1.91
Utilities	4.69	-4.68	-5.97	-5.26	-6.28	-10.29	-7.05	-7.97

**Note: August is month-to-date*

- We have seen group by operations take an ‘aggregate function’
  - sum(), mean(), count()
- These are great candidates for a heat map along with two other variables which you claim to be linked
- Natural roll ups are good candidates to think about
- For loops work to generate pairs of heat maps for various columns

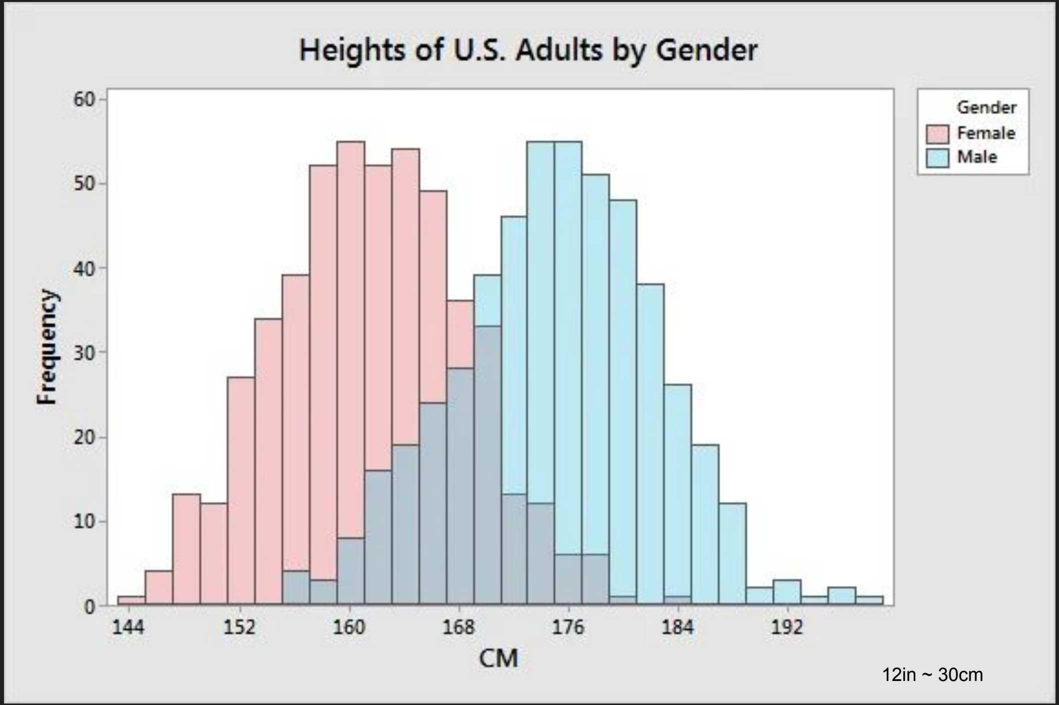
Time series analysis is a large field - this is a way to get *better* trend analysis





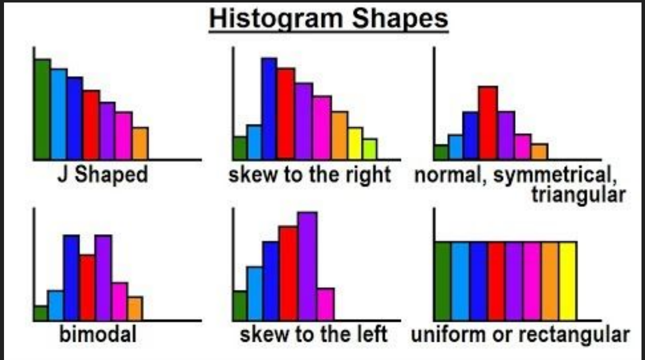
- Trend analysis is a more powerful way to look at data as it changes over time
- This is a very large field and we could not conquer it in a semester
- This is where scale comes into play, does your data have the same scale?
  - Usually it is better to normalize the data somehow, growth rate is a popular choice

Histograms represent counts, which again typically give us visual clues as to *a distribution*

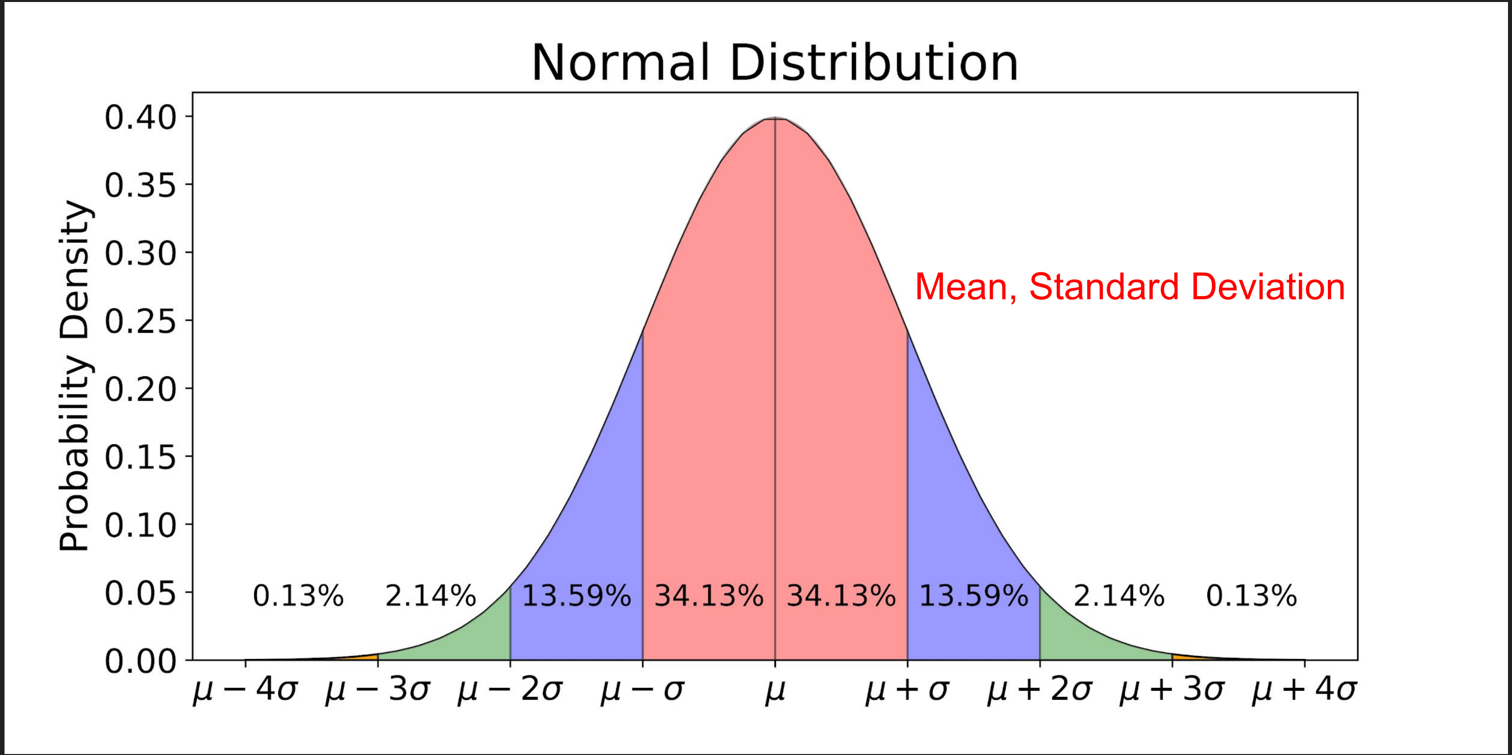


12in ~ 30cm

- Binning (how many bars)
- Frequency (count in each)
- Starting to think about distributions...



Let’s revisit the normal distribution to get a sense of where things could fall

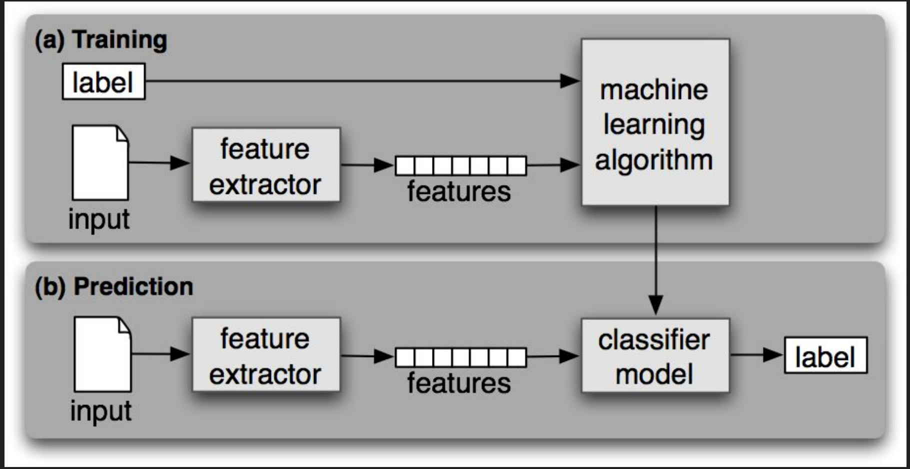


Mean, Standard Deviation

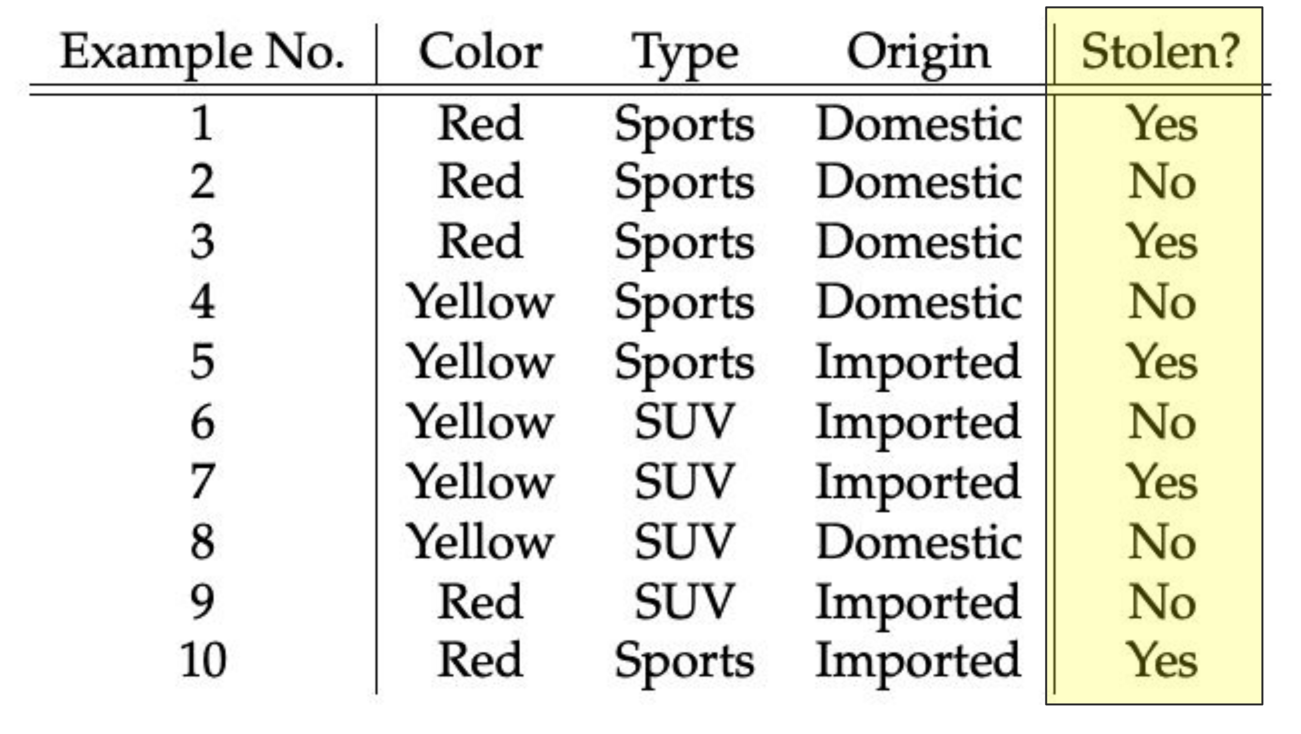
Someone give me an algorithm

Anyone…

Here begins our advanced path...



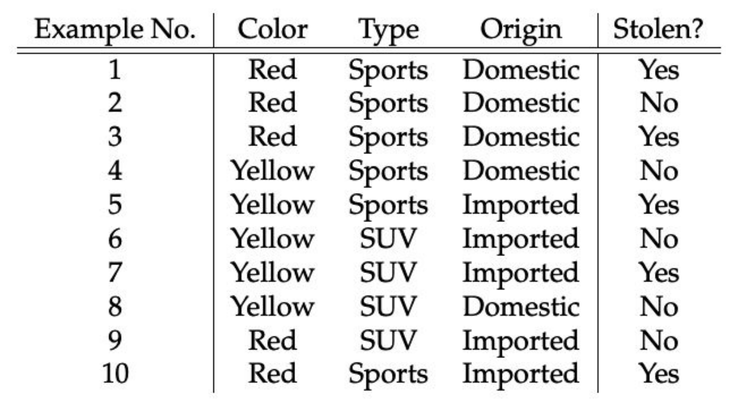
Back to Naive Bayes - our first algorithm...example data



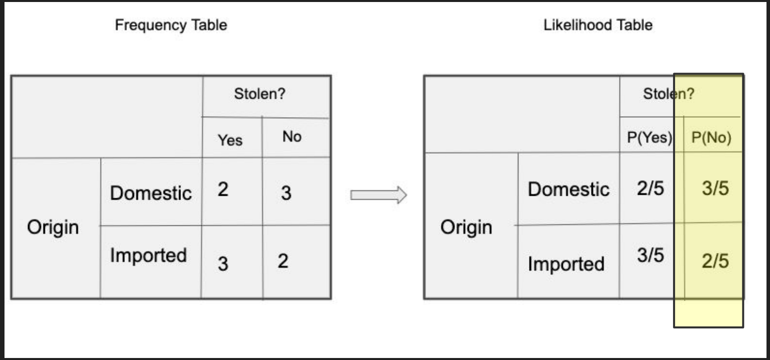
Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

We want to know if certain FEATURES produce information which correlate with a LABEL

Working through an example...for prediction



Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

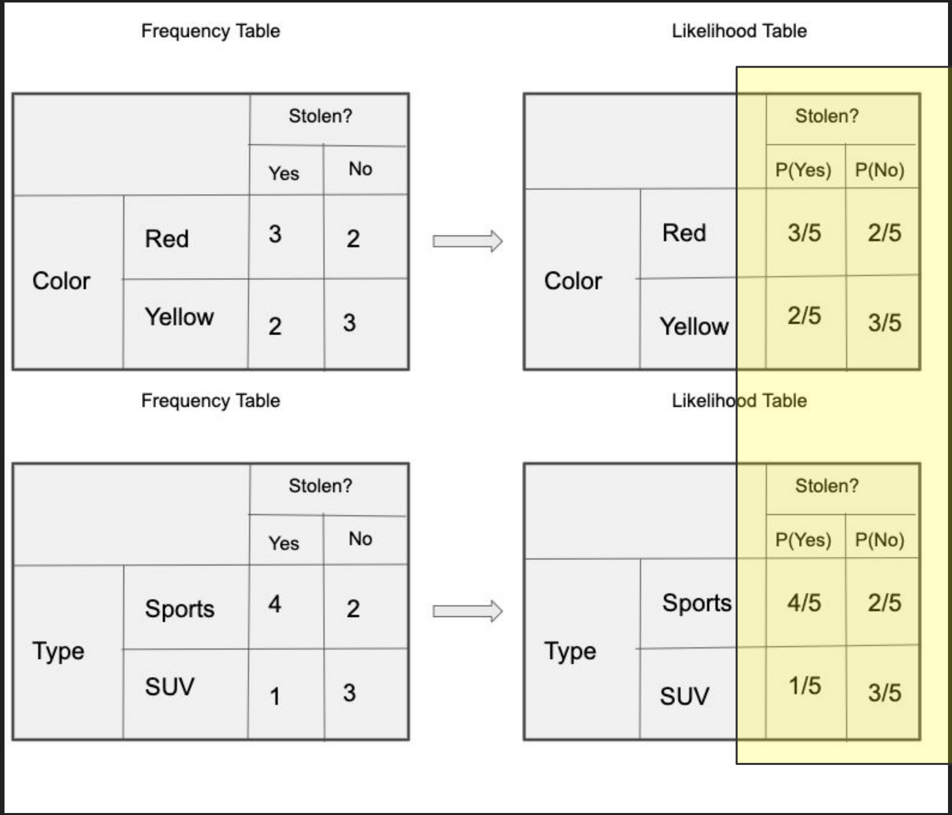


Frequency Table

		Stolen?	
		Yes	No
Origin	Domestic	2	3
	Imported	3	2

Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Origin	Domestic	2/5	3/5
	Imported	3/5	2/5



Frequency Table

		Stolen?	
		Yes	No
Color	Red	3	2
	Yellow	2	3

Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Color	Red	3/5	2/5
	Yellow	2/5	3/5

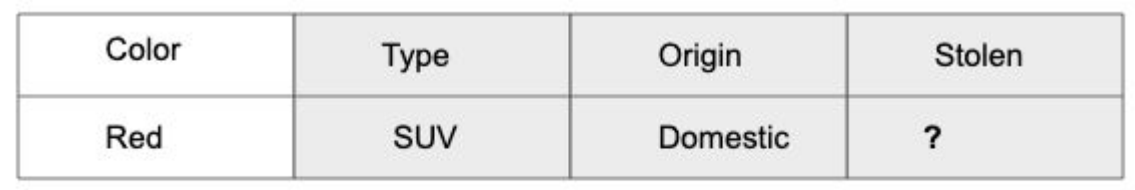
Frequency Table

		Stolen?	
		Yes	No
Type	Sports	4	2
	SUV	1	3

Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Type	Sports	4/5	2/5
	SUV	1/5	3/5

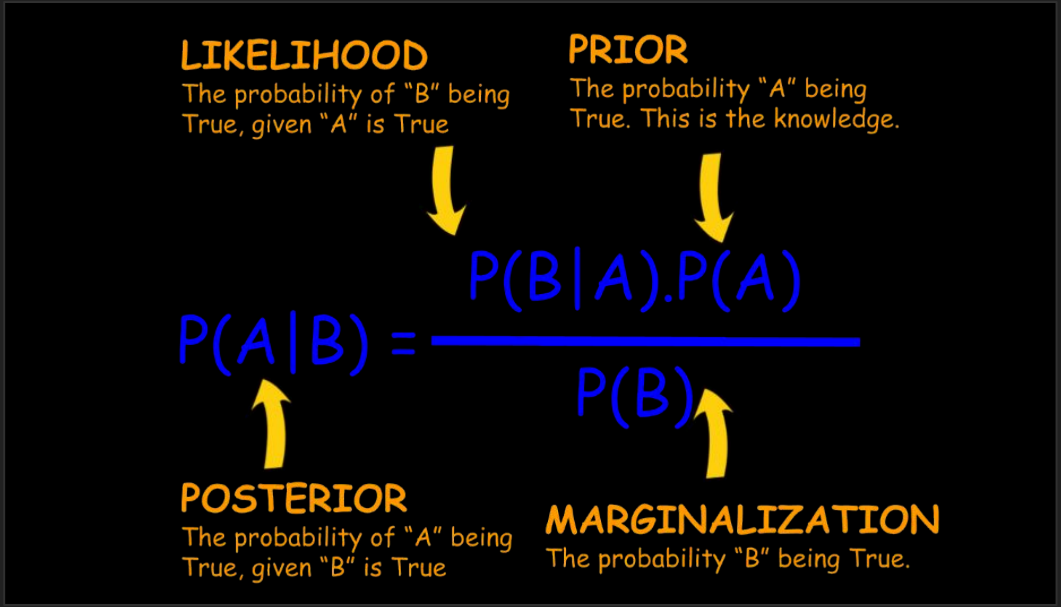
Let’s use Naive Bayes to predict if a car is stolen...



Color	Type	Origin	Stolen
Red	SUV	Domestic	?

Let’s work through this:
Vehicle = {...,...,...}
P(YES|B) =
P(B|A) = ⅕ * ⅖ * ⅗ = 0.048
P(A) = .5

P(NO|B) =
P(B|A) = ⅖ * ⅗ * ⅖ = 0.140
P(A) = .5
...NO is our prediction (higher value)



LIKELIHOOD
The probability of “B” being True, given “A” is True

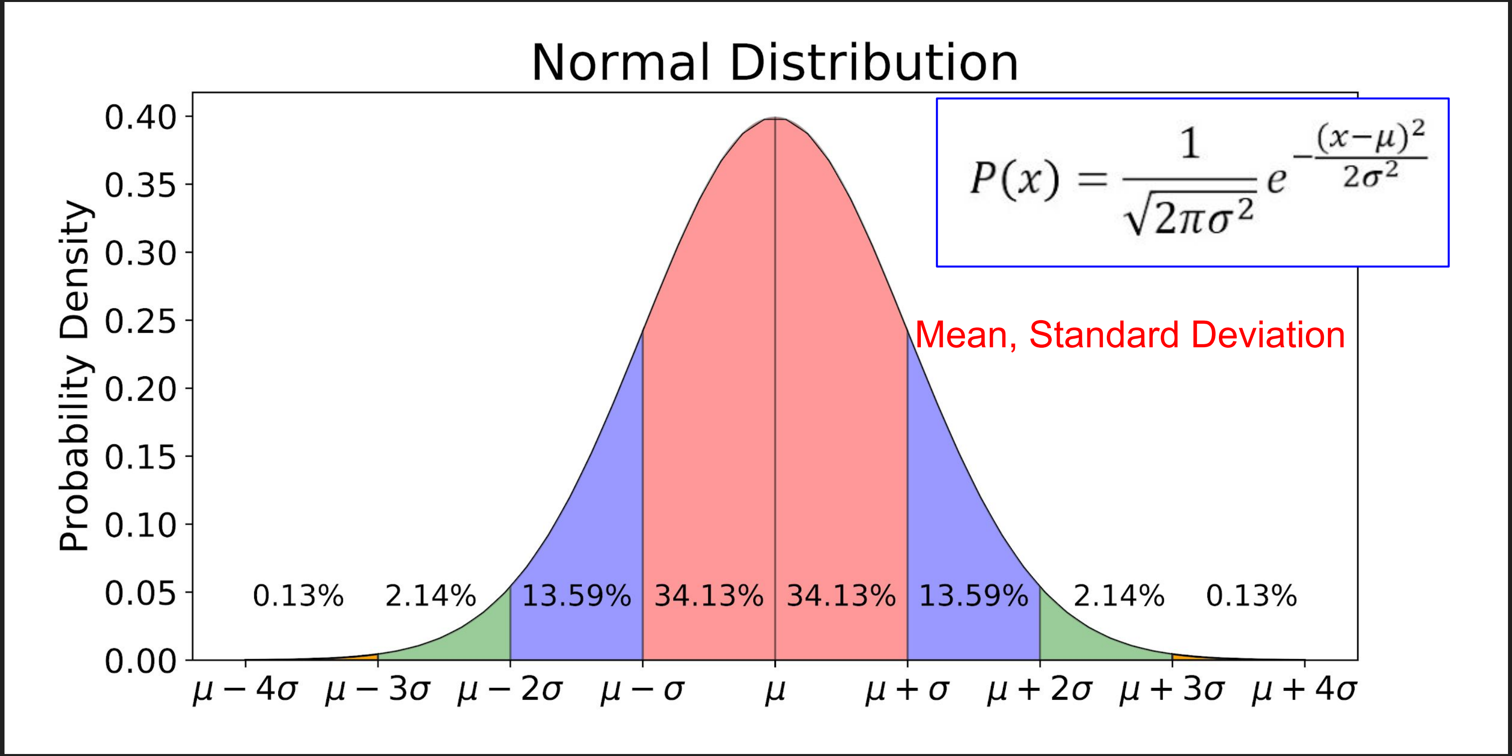
PRIOR
The probability “A” being True. This is the knowledge.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

POSTERIOR
The probability of “A” being True, given “B” is True

MARGINALIZATION
The probability “B” being True.

Recall this slide...if our data was not categorical then this applies to the probability of our feature instance



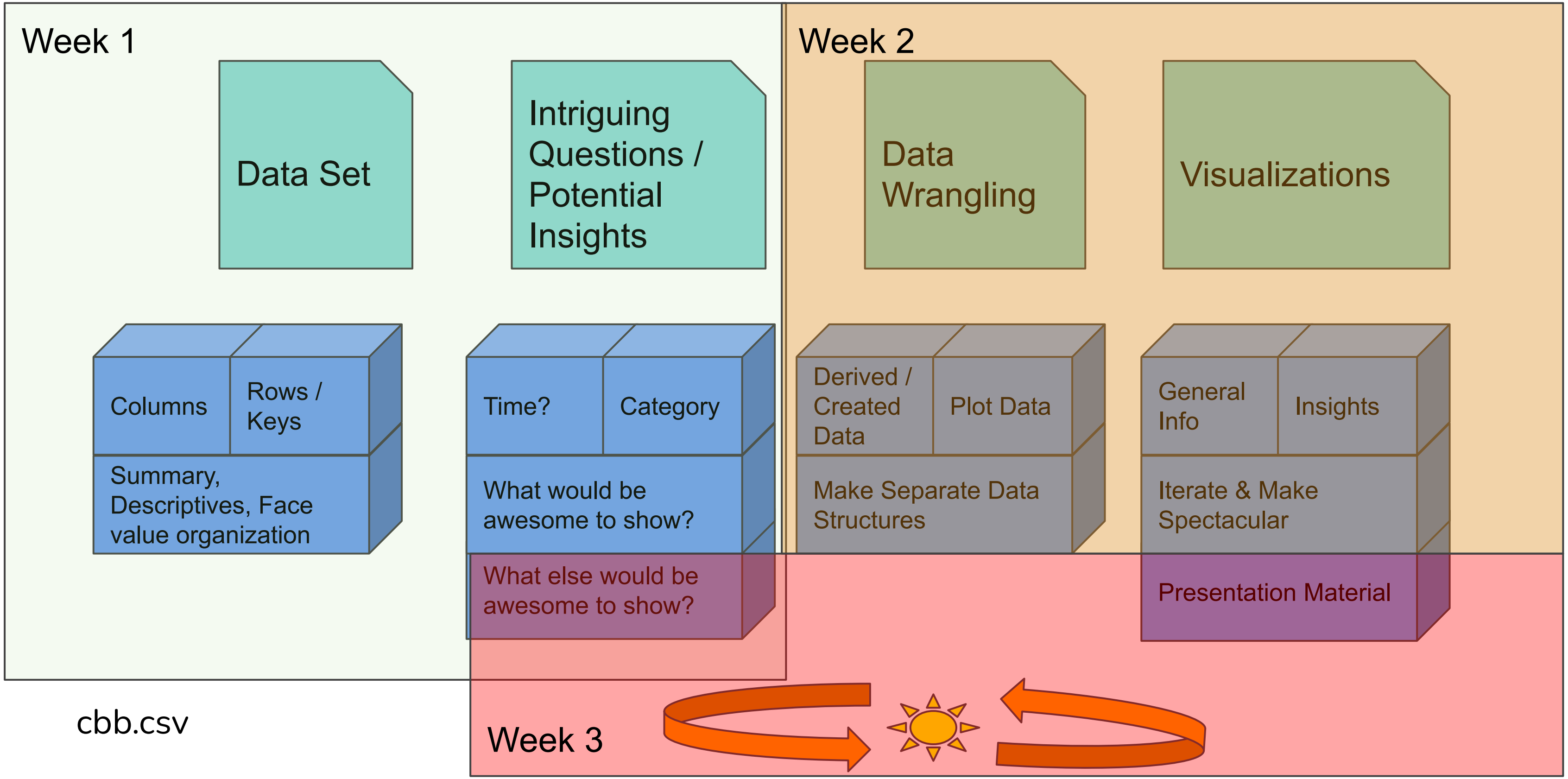
$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Mean, Standard Deviation

Naive Bayes Coding

Kicking it over to Jupyter

Project Planning - My Week 1+



Week 1
Data Set
Intriguing Questions / Potential Insights
Columns
Rows / Keys
Summary, Descriptives, Face value organization
Time?
Category
What would be awesome to show?
What else would be awesome to show?
Week 2
Data Wrangling
Visualizations
Derived / Created Data
Plot Data
Make Separate Data Structures
General Info
Insights
Iterate & Make Spectacular
Presentation Material
Week 3

cbb.csv