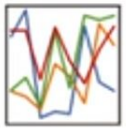
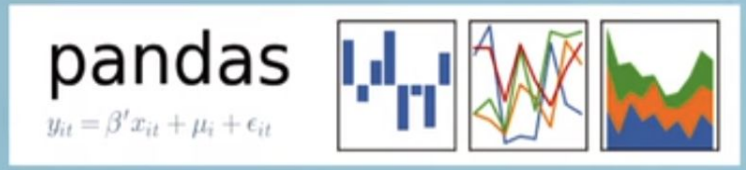
NumPy

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



matplotlib



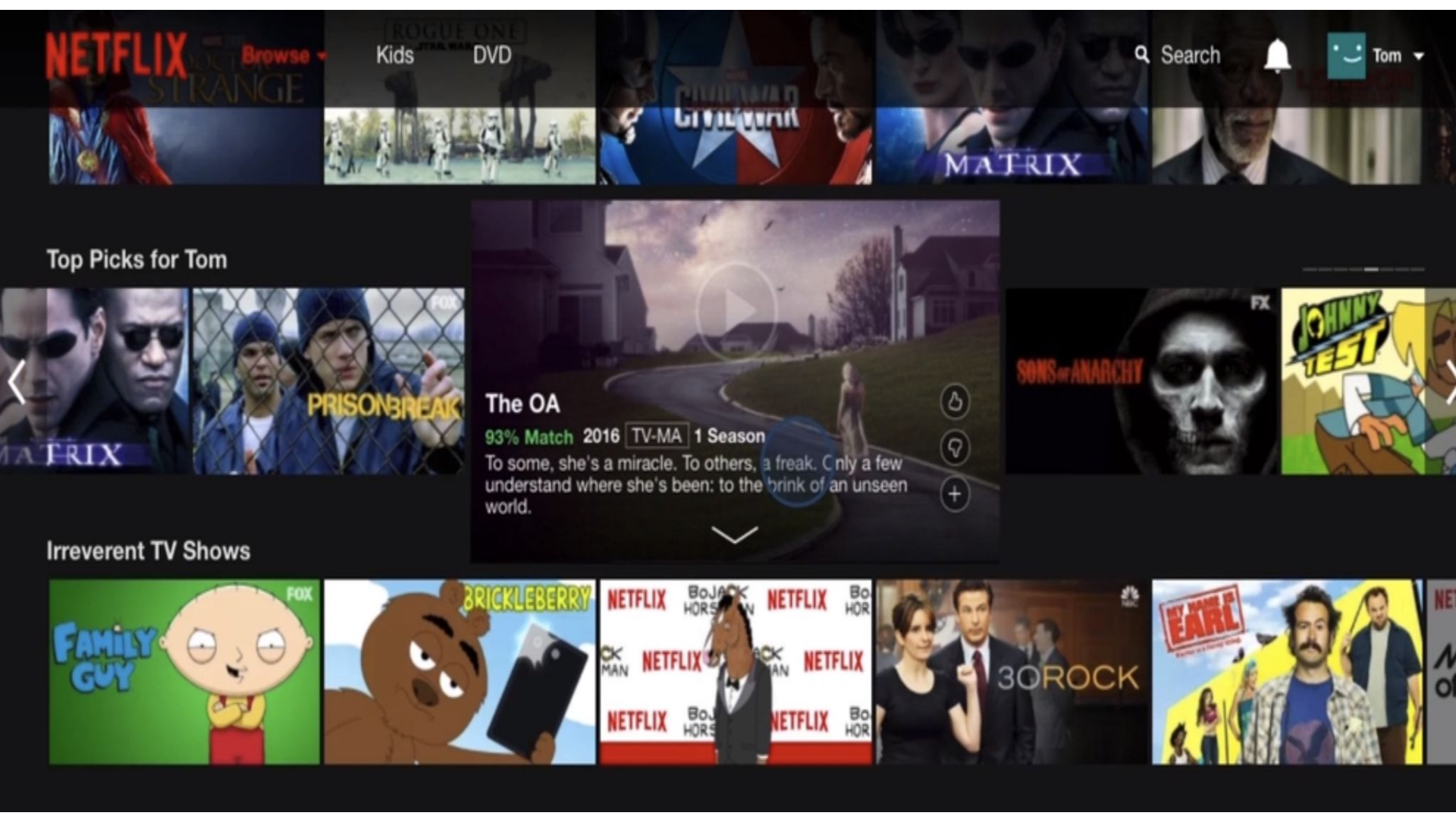
Data Frames

Data Visualization

SIMPLE PLOTS

SMALL AMOUNT OF CODE

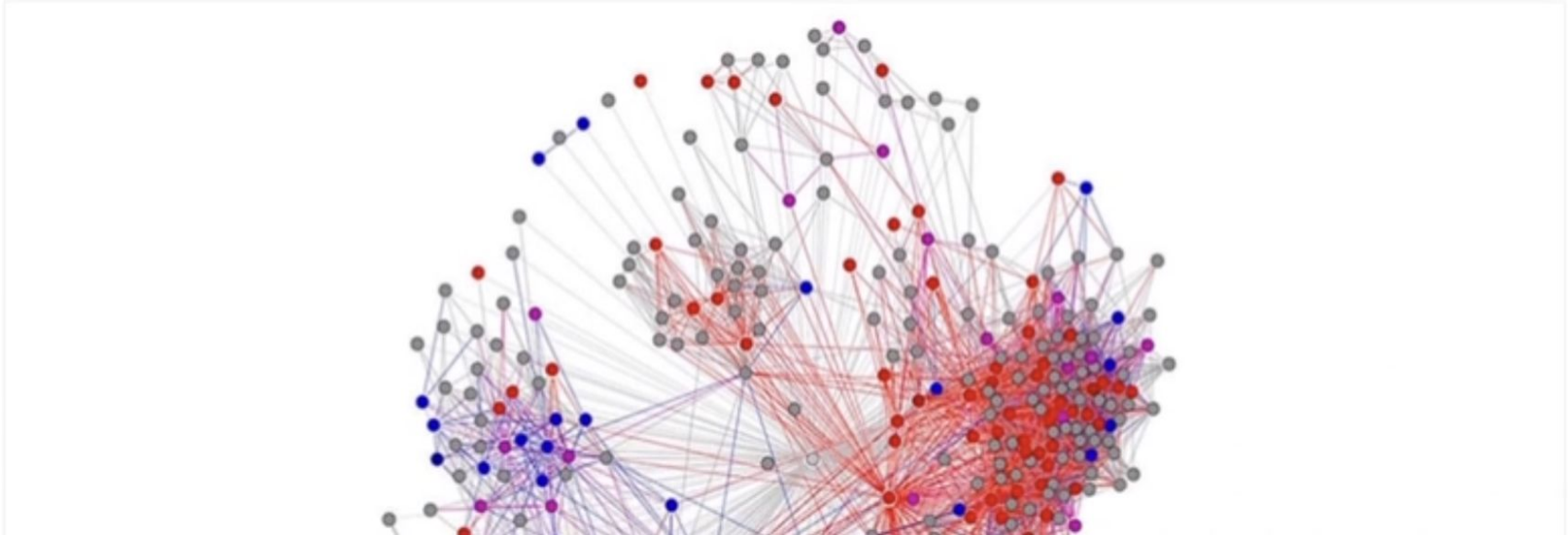
PROFESSIONAL LOOK


NETFLIX
Browse
Kids
DVD
Search
Tom
Top Picks for Tom
The OA
93% Match 2016 TV-MA 1 Season
To some, she's a miracle. To others, a freak. Only a few understand where she's been: to the brink of an unseen world.
Irreverent TV Shows
FAMILY GUY
BRICKLEBERRY
30ROCK
MY NAME IS EARL

MAY 7, 2015

Exposure to Diverse Information on Facebook

By: Eytan Bakshy, Lada Adamic, Solomon Messing



Secure | https://www.walmart.com

Walmart | All | Search | Hello. Sign In | My Account

Sponsored products you might like

$295.80

Acer Sparkly Silver 14" CB3-431-CSFM…

★★★★★ 5

2-Day Shipping

Rollback

$149.00 list price $229.00

save $80.00

14" Full HD IPS screen Laptop, Ultra-Slim…

★★★★★ 86

2-Day Shipping

$49.89

Keurig® Classic Series Chili Red Keurig® K15…

★★★★★ 262

2-Day Shipping

$6.98

Starbucks Espresso Roast Dark Ground…

★★★★★ 53

2-Day Shipping

Rollback

$329.00 list price $399.00

save $70.00

Acer Aspire ES 15 Series ES1-572-33BP 15.6"…

★★★★★ 2

2-Day Shipping

$279.00

HP 14-ak040wm 14" Chromebook,…

★★★★★ 2

2-Day Shipping

Rollback

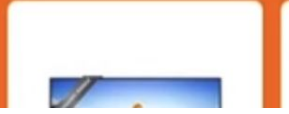
$249.00

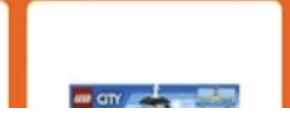
Dell Inspiron 11 3168 11.6" Laptop,…

★★★★★ 3116

2-Day Shipping

Top-selling Pickup Discount items





learning has shown promise in multiple fields but requires considerable computing power. Fortunately for Google, this is one area in which it is well equipped.

The search giant applied its large-scale neural network training system to the work. Google built the system to train networks of tens of thousands of CPU cores to perform a task. In the drug discovery project, the training entailed equipping the network to comb through 37.8 million data points covering more than 200 different biological processes. After running the system for more than 50 million CPU hours, Google has concluded the inclusion of data from multiple sources allows it to make more accurate predictions of the efficacy of a drug across different diseases.

Even greater scales are in Google's sights. At the time of writing a paper on the project, Google had scaled the system up to 239 tasks and the upward efficiency trend was yet to plateau. Similarly, the addition of more data was found to increase efficiency, too. The researchers have cast lustful looks at the "vast private stores of experimental measurements" locked away at Big Pharma companies as they try to figure out the next steps for the model. More data and more tasks are the near-term goals for the project.

Whether the efficiency gains touted by Google will have an effect on drug discovery remains to be seen. Even the paper's authors accept that the complexity of drug discovery could limit the impact of the approach, but overall are as optimistic as one would expect Google staffers to be about the potential for data and algorithms deliver improvements.

- read the paper (PDF)
- check out Google's blog post
- and *VentureBeat*'s take

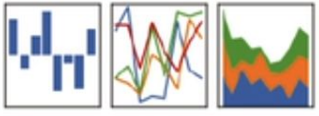
NumPy

SciPy

IP[y]:
IPython

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



matplotlib

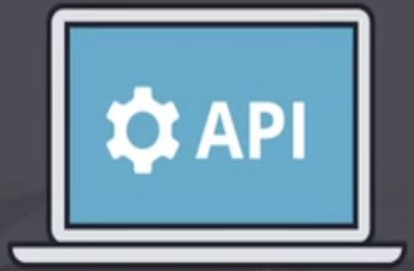
WRANGLING

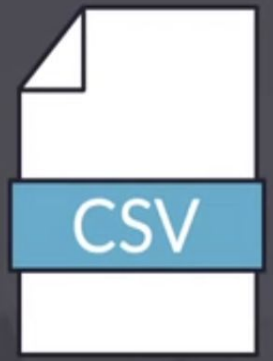
- Gather
- Assess
- Clean

EDA

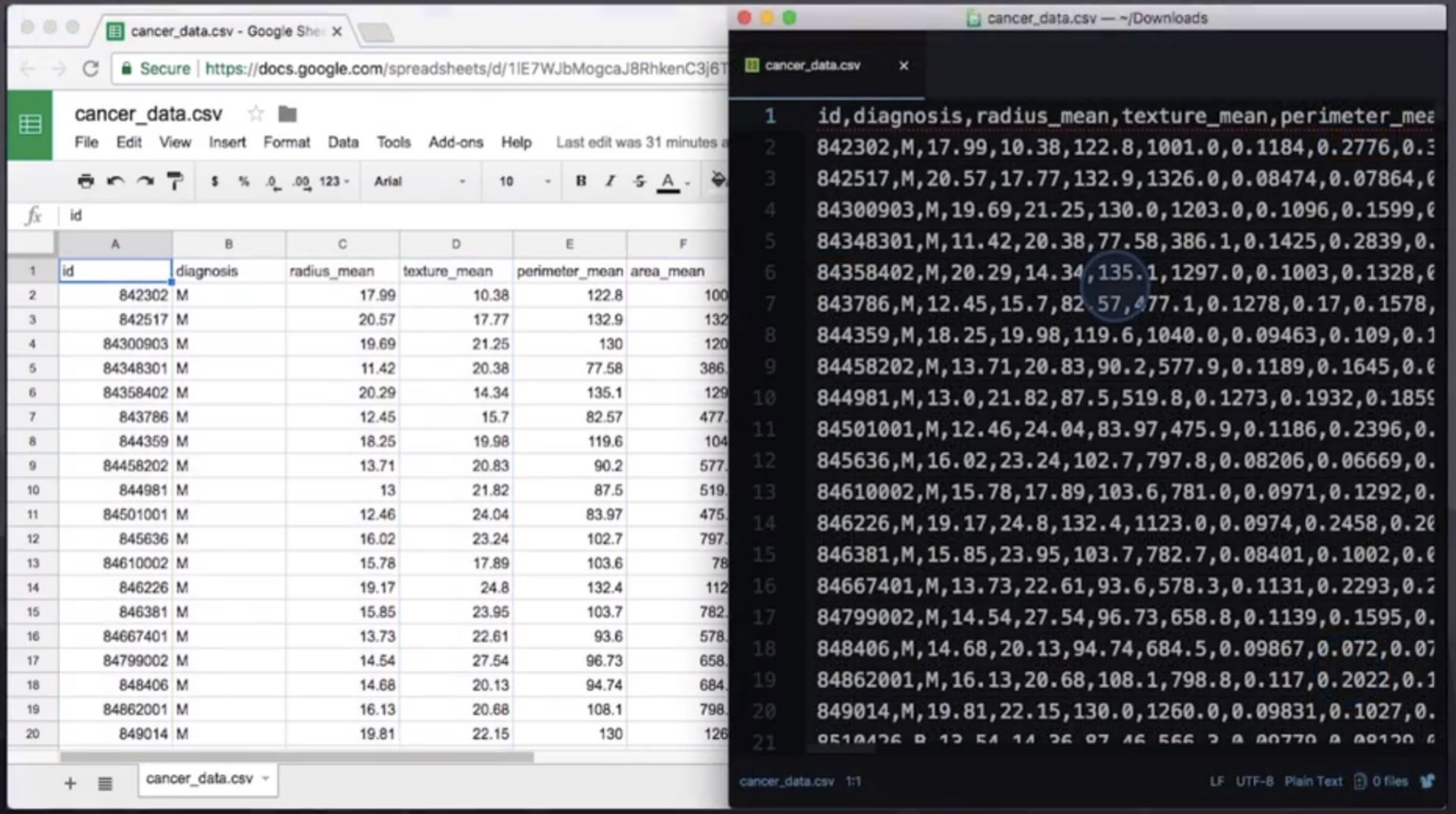
- Explore
- Augment

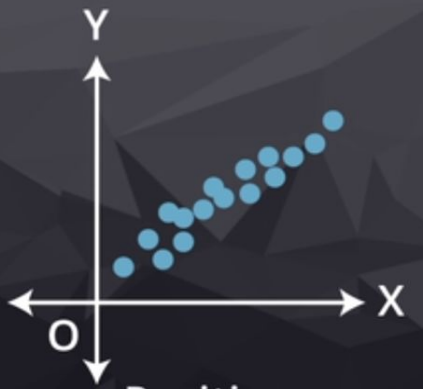
GATHERING



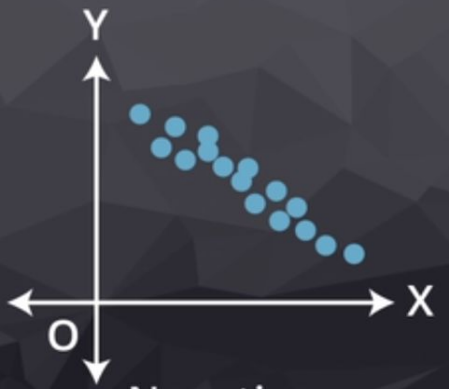


COMMA SEPARATED VALUES

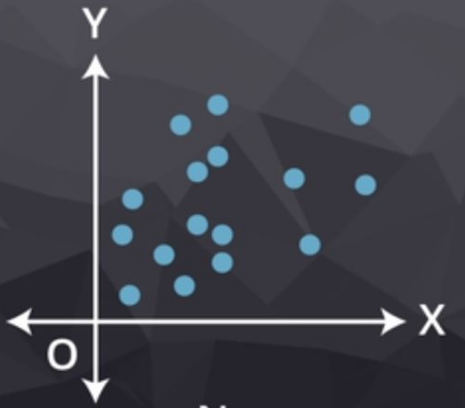

cancer_data.csv - Google Sheets
Secure | https://docs.google.com/spreadsheets/d/1lE7WJbMogcaJ8RhkenC3j6T
cancer_data.csv
File Edit View Insert Format Data Tools Add-ons Help Last edit was 31 minutes
fx id
id,diagnosis,radius_mean,texture_mean,perimeter_mean,area_mean
842302,M,17.99
842517,M,20.57
84300903,M,19.69
84348301,M,11.42
84358402,M,20.29
843786,M,12.45
844359,M,18.25
84458202,M,13.71
844981,M,13.0
84501001,M,12.46
845636,M,16.02
84610002,M,15.78
846226,M,19.17
846381,M,15.85
84667401,M,13.73
84799002,M,14.54
848406,M,14.68
84862001,M,16.13
849014,M,19.81
cancer_data.csv
cancer_data.csv — ~/Downloads
cancer_data.csv 1:1
LF UTF-8 Plain Text 0 files



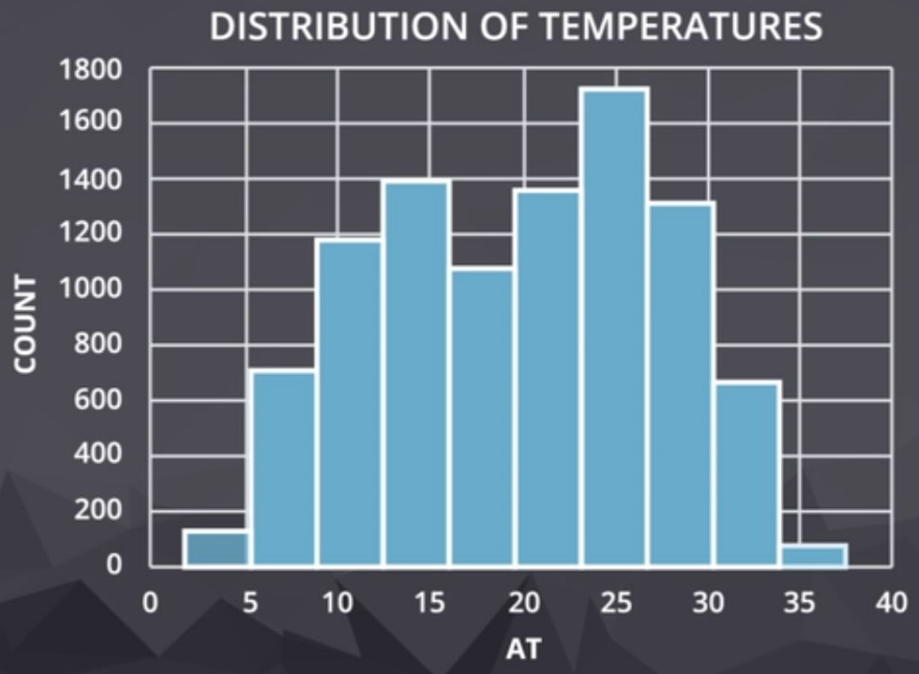
Positive Correlation

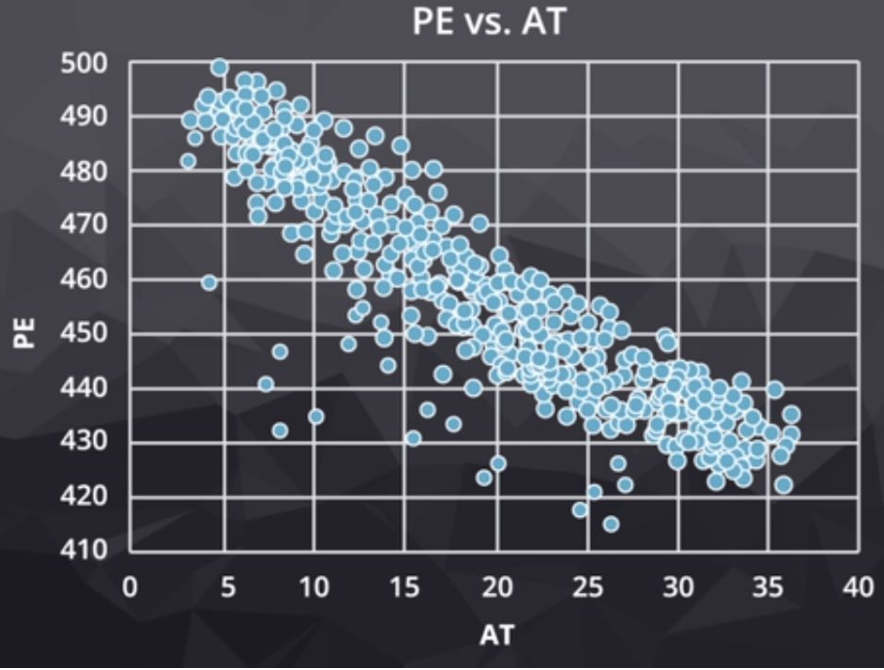


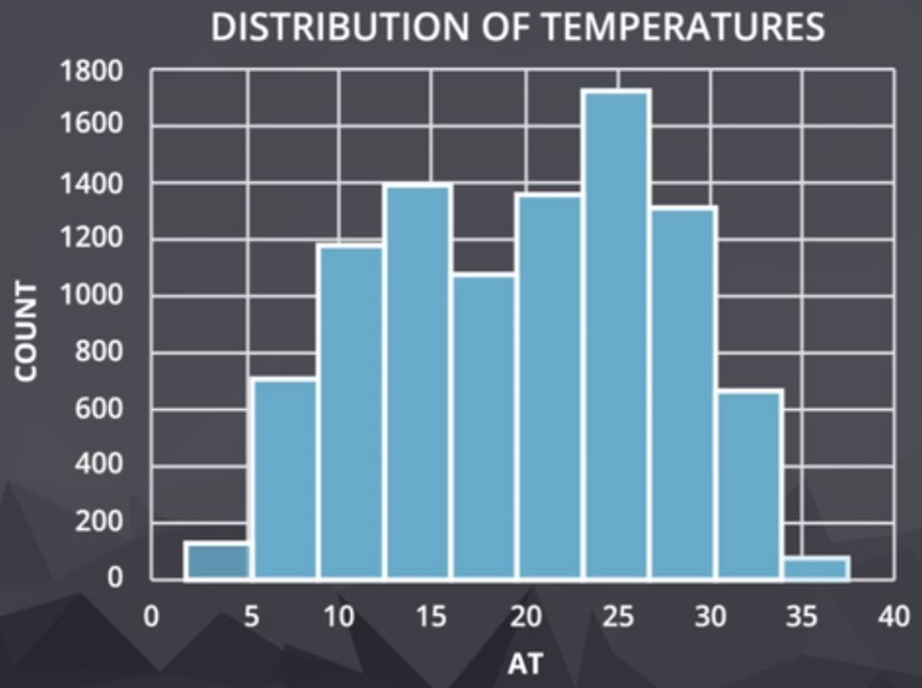
Negative Correlation

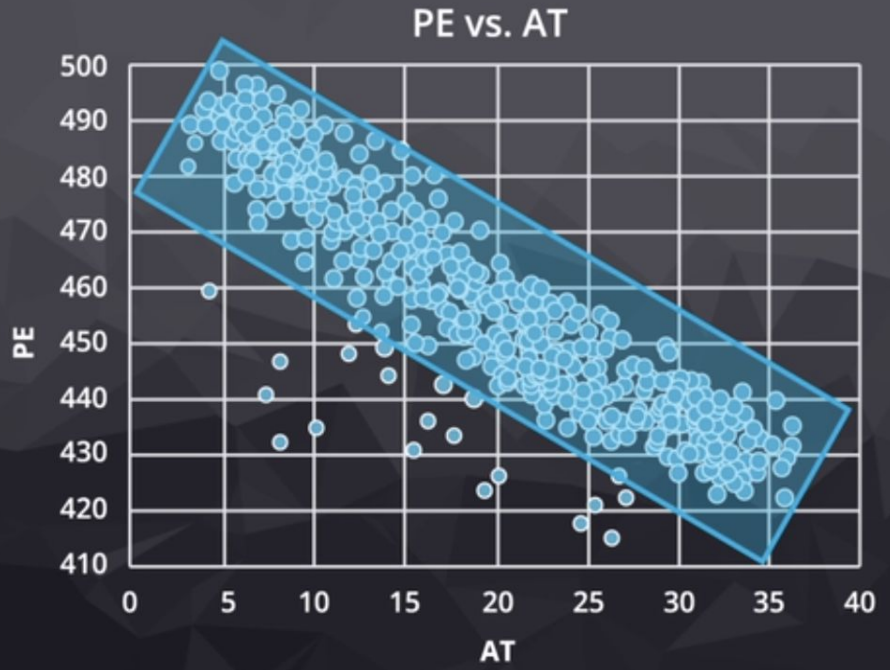


No Correlation


DISTRIBUTION OF TEMPERATURES
COUNT
1800
1600
1400
1200
1000
800
600
400
200
0
0
5
10
15
20
25
30
35
40
AT


PE vs. AT
PE
500
490
480
470
460
450
440
430
420
410
0
5
10
15
20
25
30
35
40
AT


DISTRIBUTION OF TEMPERATURES
COUNT
1800
1600
1400
1200
1000
800
600
400
200
0
0
5
10
15
20
25
30
35
40
AT


PE vs. AT
PE
500
490
480
470
460
450
440
430
420
410
0
5
10
15
20
25
30
35
40
AT