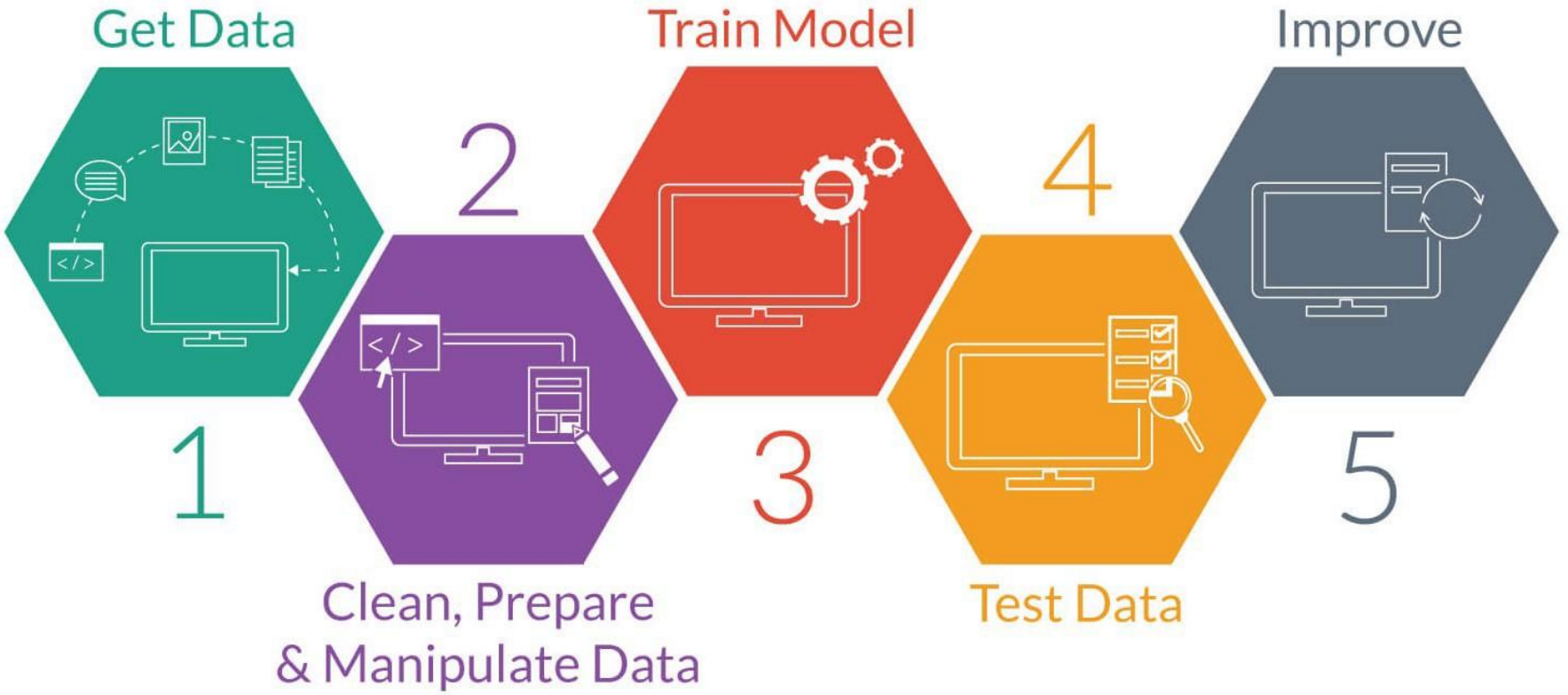
Feature selection

...or variables, or attributes..

Boom!


Get Data
Train Model
Improve
2
4
1
3
5
Clean, Prepare
& Manipulate Data
Test Data

Lesson 1

GIGO

Training data				
var1	var2	var3	var4	result
5	2	2	1	dead
5	1	2	2	allve
3	3	1	2	dead
5	3	2	1	alive
4	3	2	2	alive

Lesson 2

Advantages

1. Could improve the results
2. Faster training

Lesson 3

Univariante feature selection

Training data				
var1	var2	var3	var4	result
5	2	2	1	dead
5	1	2	2	allve
3	3	1	2	dead
5	3	2	1	alive
4	3	2	2	alive

Lesson 4

Remove no sense variables

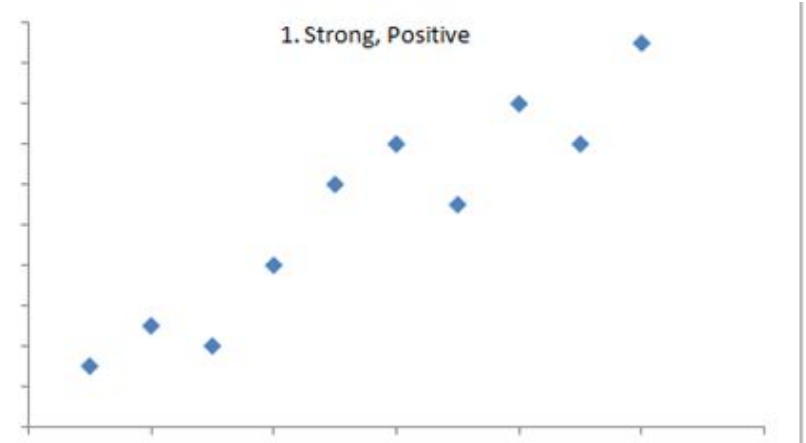
Lesson 5

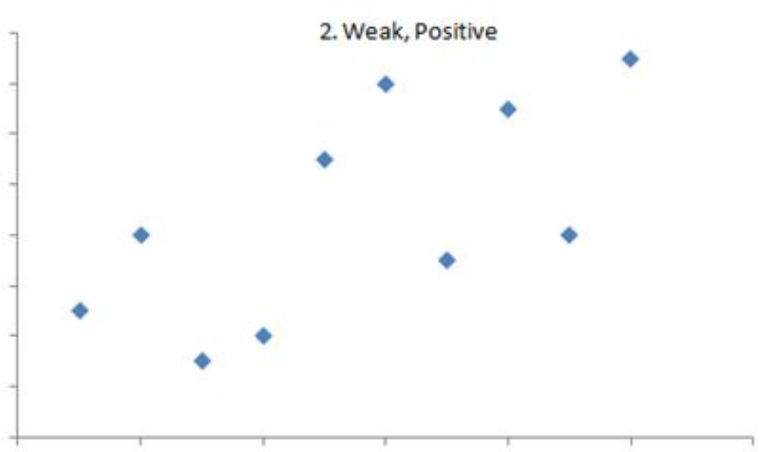
Low variance variables

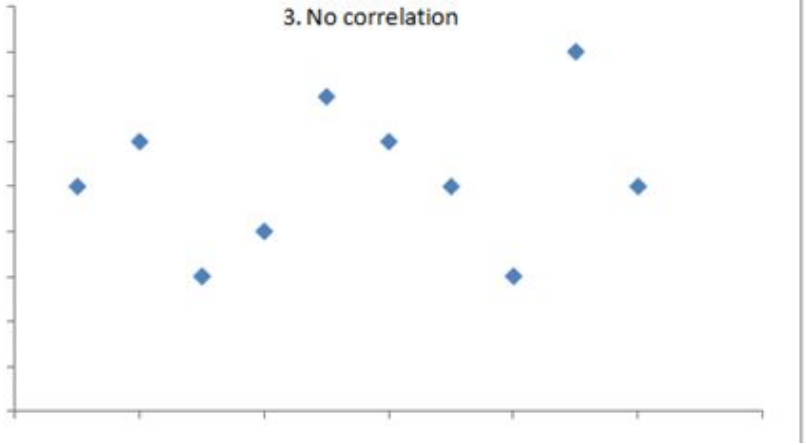
Lesson 6

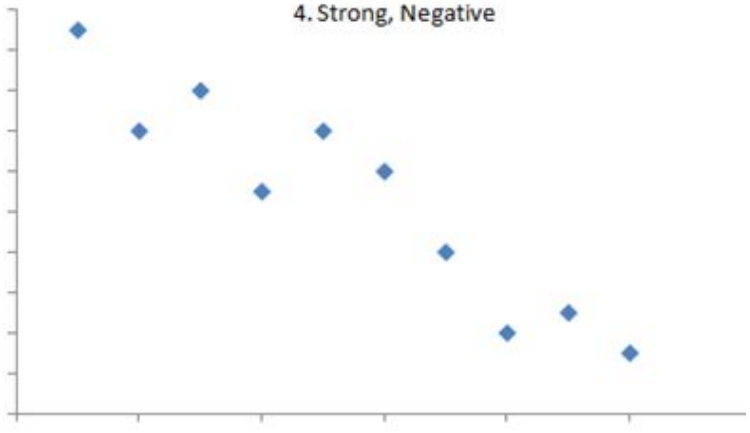
Pearson's Correlation

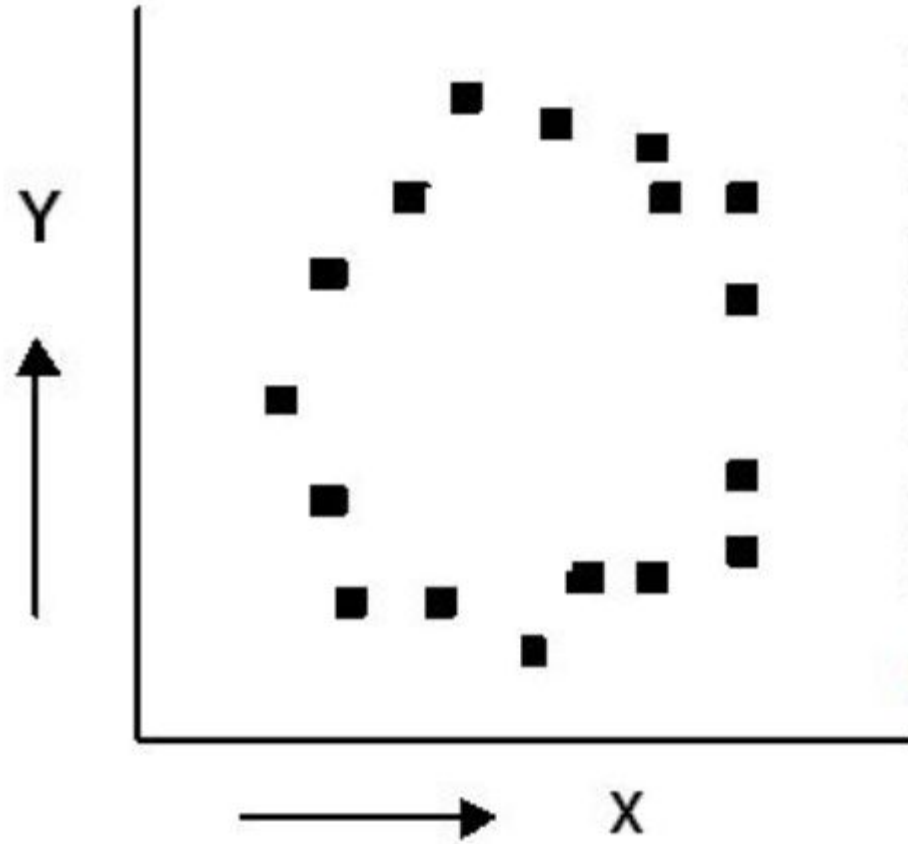
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$


1. Strong, Positive


2. Weak, Positive


3. No correlation


4. Strong, Negative


Y
X

Lesson 7

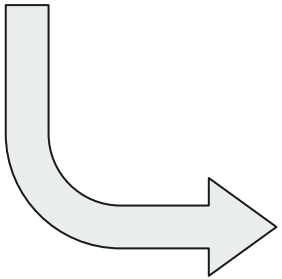
Deal with categorical variables

- Numerical Value
- One hot encoding

Get dummies!

	numeric_variable	categorical_variable
0	1	A
1	2	A
2	5	B
3	8	A

df = pd.get_dummies(df, columns=[])

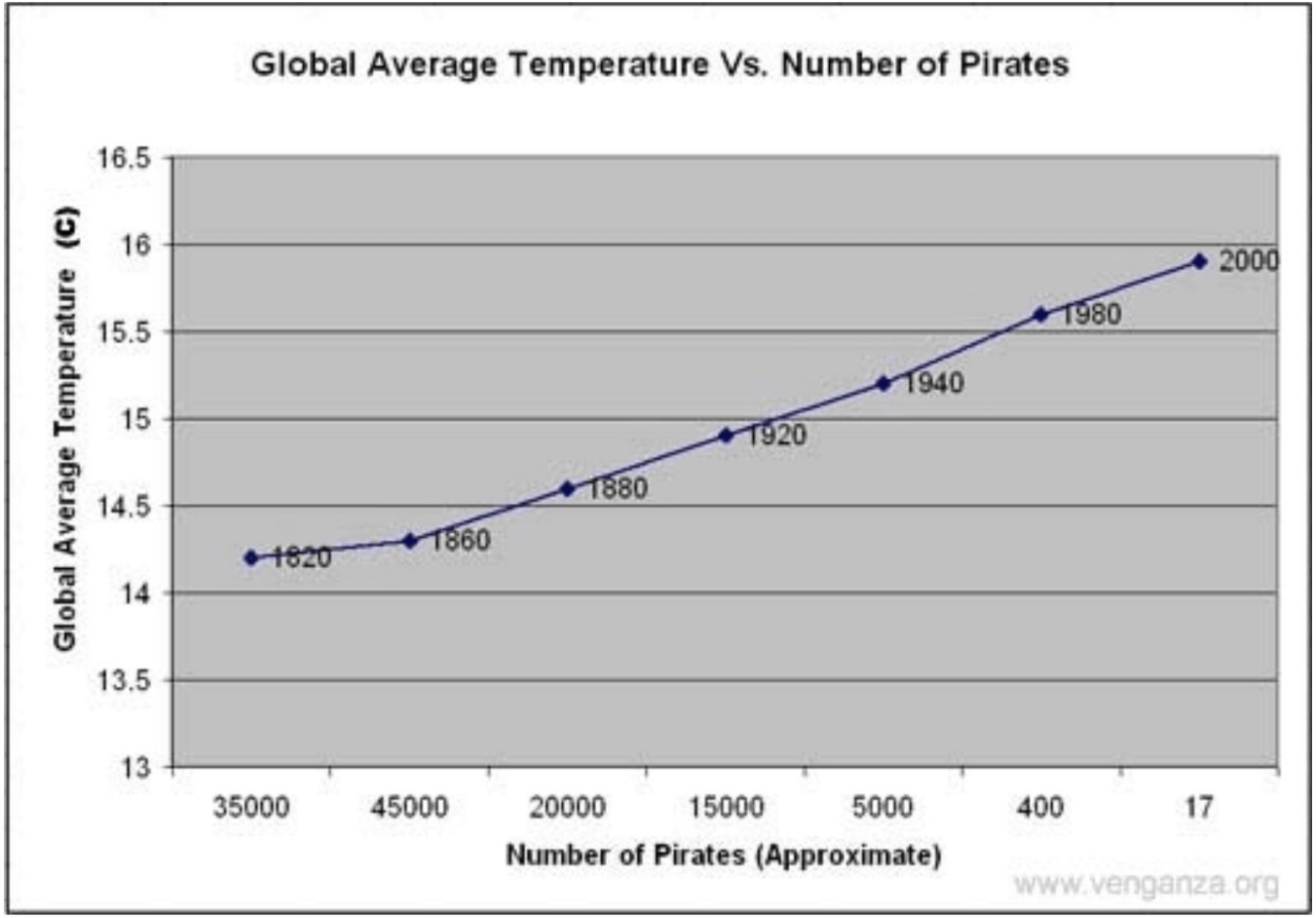


	numeric_variable	categorical_variable_A	categorical_variable_B
0	1	1	0
1	2	1	0
2	5	0	1
3	8	1	0

Practice time

Lesson 8

Correlation isn’t causation


Global Average Temperature Vs. Number of Pirates
Global Average Temperature (C)
16.5
16
15.5
15
14.5
14
13.5
13
1620
1860
1880
1920
1940
1980
2000
35000
45000
20000
15000
5000
400
17
Number of Pirates (Approximate)
www.venganza.org

Lesson 9

Chi-square test

- For categorical variables
- Statistical test
- Likelihood of correlation

Practice time

Lesson 10

Tree-based feature selection

Practice time

Lesson 11

Multivariante methods

- Slower than univariante
- More powerful

Lesson 12

Wrapper methods

- Forward Selection
- Recursive Feature elimination

Lesson 13

Embedded Methods

- LASSO
- Ridge Regression

Practice time

Thank you!