Data Dwarf

University of California: Santa Barbara

2015 Senior Computer Science Capstone Project

Team Members:

Sean Spearman
Cody Brown
Ray Smets
Aimee Galang
Tim Shen
Data is Valuable
The Problem

1. Data’s format is hard to understand.

2. Finding relationships is tedious and not guaranteed.
The Problem

1. Data’s format is hard to understand.
2. Finding relationships is tedious and not guaranteed.
Datasets sourced from World Bank Cross Country Data and ILOSTAT Database through Quandl
Datasets sourced from World Bank Cross Country Data and ILOSTAT Database through Quandl
The Problem

1. Data’s format is hard to understand.

2. Finding relationships is tedious and not guaranteed.
Correlation Indicates Potentially Meaningful Relationships
DataDwarf
<table>
<thead>
<tr>
<th>Technology</th>
<th>Front End</th>
<th>Back End</th>
<th>Data Source</th>
<th>Teamwork</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Elephant</td>
<td>Q</td>
<td>GitHub</td>
</tr>
<tr>
<td></td>
<td>Sass</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>jQuery</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Correlation Value Usage

- **Input:** 2 datasets
- **Output:** one number
- **Want to know:** how similar they are
Correlation Range

1: Directly correlated
0: Not correlated
-1: Inversely correlated
Pearson vs (Spearman or Kendall)

Linear and Monotonic

use Pearson

Only Monotonic

use Spearman or Kendall

Pearson

Example

- Dataset 2
- Dataset 1

Linear and Monotonic

Y values of Dataset 2

Y values of Dataset 1

Time

Y
Pearson
Pearson

Linear, Monotonic

always decreasing by $x$

Example

$Y$

Time

Dataset 2

Dataset 1
Spearman and Kendall

Example

- Red circle: Dataset 2
- Black circle: Dataset 1

Monotonic
Spearman and Kendall

Example

- Red dots: Dataset 2
- Black dots: Dataset 1

Y values of Dataset 2
Y values of Dataset 1

Only Monotonic

Time
Spearman and Kendall

Not Linear, Monotonic

Example

- Dataset 2
- Dataset 1

always increasing

Time

Y
Pearson vs (Spearman and Kendall)

Linear, Monotonic

always decreasing by x

Not Linear, Monotonic

always increasing
Pearson vs Spearman

Pearson = 0.37
Spearman = 0.35

Pearson = 0.67
Spearman = 0.84

Pearson = 0.88
Spearman = 1

http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient
Spurious Correlation Example

Total revenue generated by arcades correlates with Computer science doctorates awarded in the US
Demo
Bring your own data (BYOD) to DataDwarf.io