WOMEN WHO

CODE

OLGA KUZMINA

DATA CLEANING

DATA ANALYSIS

datagovsg Follow
Official Medium account for https://data.gov.sg, Singapore's open data portal.
Nov 30, 2016 · 8 min read

How the Circle Line rogue train was caught with data

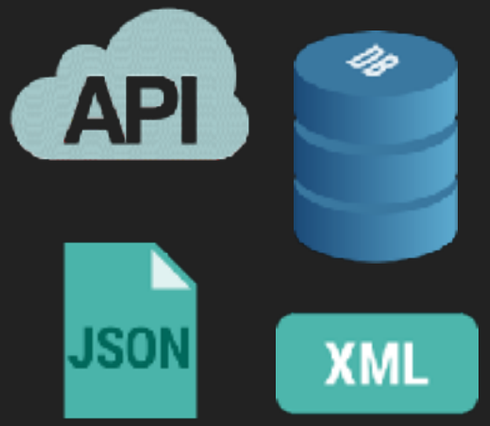
Text: Daniel Sim | Analysis: Lee Shangqian, Daniel Sim & Clarence Ng

Singapore's MRT Circle Line was hit by a spate of mysterious disruptions in recent months, causing much confusion and distress to thousands of commuters.

- https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

DATA ANALYSIS STEPS

DATA COLLECTION → RAW DATA → DATA CLEANING → TIDY DATA → DATA ANALYSIS



pandas

$$y_it - \beta' x_{it} + \mu_i + \epsilon_{it}$$

TIDY DATA

H17

	A	B	C	D	E
1	id	name	surname	phone	
2	1	Bob	Douglas	4567281	
3	2	Kate	Fox	3484932	
4	3	Alison	Lee	4329424	
5	4	James	Brown	2423432	
6	5	Mike	Hall	2323843	
7					
8					

- Each observation should be in a different row
- Each variable should be in one column

RAW DATA

```
1794595:-1,1,-1,-211,-1551,0.00;-1,2,-1,0,-2654,0.00;-1,3,-1,355,-2802,0.00;-1,4,-1,-606
,-1512,0.00;-1,5,-1,-666,-3099,0.00;-1,6,-1,-441,-2563,0.00;-1,7,-1,2812,-2790,0.00;-1,8
,-1,-451,-3190,0.00;-1,9,-1,264,-3218,0.00;-1,10,-1,294,-2321,0.00;-1,11,-1,-157,-3257,0
.00;-1,12,-1,803,-3644,0.00;4,13,-1,5550,4400,0.00;-1,14,-1,2169,-3114,0.00;4,15,-1,5550
,4400,0.00;4,16,-1,5550,4400,0.00;4,17,-1,5550,4400,0.00;4,18,-1,5550,4400,0.00;4,19,-1,
5550,4400,0.00;4,20,-1,5550,4400,0.00;4,21,-1,5550,4400,0.00;4,22,-1,5550,4400,0.00;4,23
,-1,5550,4400,0.00;4,24,-1,5550,4400,0.00;4,25,-1,5550,4400,0.00;4,26,-1,5550,4400,0.00;
4,27,-1,5550,4400,0.00;4,28,-1,5550,4400,0.00;4,29,-1,5550,4400,0.00;:-606,-1512,0,1.00,
A,Dead;:
1794596:-1,1,-1,-211,-1552,0.00;-1,2,-1,0,-2653,0.00;-1,3,-1,357,-2796,0.00;-1,4,-1,-606
,-1512,0.00;-1,5,-1,-666,-3099,0.00;-1,6,-1,-441,-2559,0.00;-1,7,-1,2815,-2787,0.00;-1,8
,-1,-452,-3180,0.00;-1,9,-1,265,-3212,0.00;-1,10,-1,294,-2315,0.00;-1,11,-1,-161,-3261,0
.00;-1,12,-1,806,-3640,0.00;4,13,-1,5550,4400,0.00;-1,14,-1,2170,-3117,0.00;4,15,-1,5550
,4400,0.00;4,16,-1,5550,4400,0.00;4,17,-1,5550,4400,0.00;4,18,-1,5550,4400,0.00;4,19,-1,
5550,4400,0.00;4,20,-1,5550,4400,0.00;4,21,-1,5550,4400,0.00;4,22,-1,5550,4400,0.00;4,23
,-1,5550,4400,0.00;4,24,-1,5550,4400,0.00;4,25,-1,5550,4400,0.00;4,26,-1,5550,4400,0.00;
4,27,-1,5550,4400,0.00;4,28,-1,5550,4400,0.00;4,29,-1,5550,4400,0.00;:-532,-1496,58,85.6
6,A,Dead;:
```

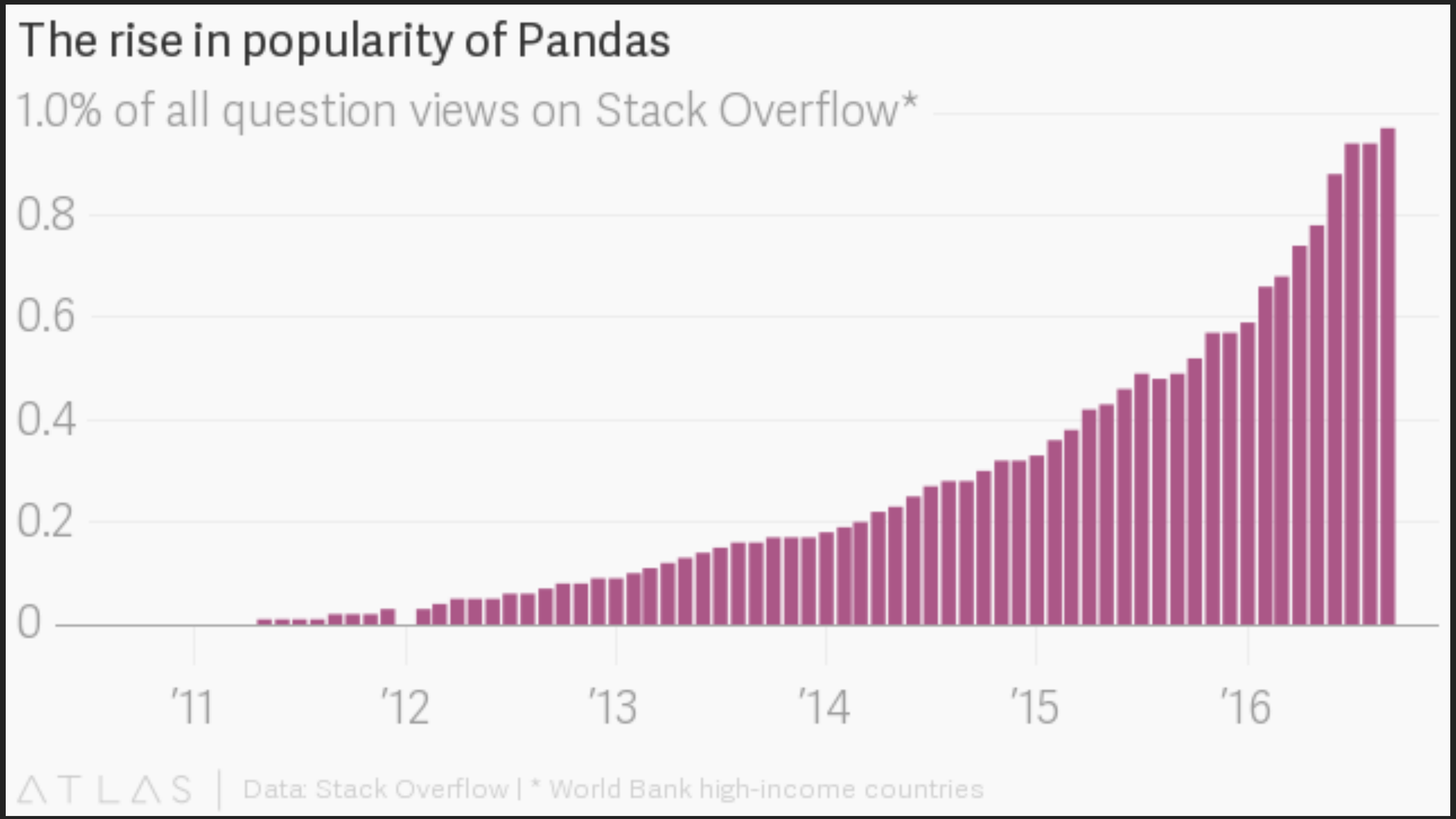
PANDAS

"...it enables people to analyze and work with data who are not expert computer scientists"

"You still have to write code, but it's making the code intuitive and accessible. It helps people move beyond just using Excel for data analysis"

Wes McKinney, developer of "Pandas",

PANDAS


The rise in popularity of Pandas
1.0% of all question views on Stack Overflow*
0.8
0.6
0.4
0.2
0
'11
'12
'13
'14
'15
'16
ATLAS
Data: Stack Overflow | * World Bank high-income countries

PANDAS.DATAFRAME

Columns

	Country	Capital	Population
0	Belgium	Brussels	11190846
1	India	New Delhi	1303171035
2	Brazil	Brasilia	207847528

Index

- pandas.DataFrame - Two-dimensional labeled data structure with columns of potentially different types.

PANDAS.DATAFRAME

Columns

	Country	Capital	Population
0	Belgium	Brussels	11190846
1	India	New Delhi	1303171035
2	Brazil	Brasilia	207847528

Index

H17

	A	B	C	D	E
1	id	name	surname	phone	
2	1	Bob	Douglas	4567281	
3	2	Kate	Fox	3484932	
4	3	Alison	Lee	4329424	
5	4	James	Brown	2423432	
6	5	Mike	Hall	2323843	
7					
8					

REFERENCES

1. https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

2. https://github.com/pandas-dev/pandas/blob/master/doc/cheatsheet/Pandas_Cheat_Sheet.pdf

3. https://s3.amazonaws.com/assets.datacamp.com/blog_assets/PandasPythonForDataScience.pdf

4. https://qz.com/1126615/the-story-of-the-most-important-tool-in-data-science/

5. https://www.coursera.org/learn/data-cleaning

TRY. EXPLORE. PRACTICE . . .

Thank you .

oskuzm@gmail.com

WOMEN WHO CODE