

Beuth Hochschule für Technik Berlin

Computer Science für Big Data Übg

A4-Data Preparation Exercise

February 2020

Ilia Ozhmegov

Berlin 2020

Exercise:

A. Do the 5 Day Data Cleaning Challenge by Rachel Tateman:

kaggle.com/getting-started. The sign up here link is broken but if you search the 5 courses in Kaggle you will find them easily.

Fork her repositories and send me one or five links to your Kaggle solutions.

B. Play around with Trifacta! Import your favourite dataset where you might have inserted some errors before. Upload one PDF where you show me screenshots, explain the problems with the dataset and show the recipes that solved the problems.

A. Data Cleaning Challenge

Accomplished challenges:

1. iliaozhmegov/data-cleaning-challenge-handling-missing-values
2. iliaozhmegov/data-cleaning-challenge-scale-and-normalize-data
3. iliaozhmegov/data-cleaning-challenge-parsing-dates
4. iliaozhmegov/data-cleaning-challenge
5. iliaozhmegov/data-cleaning-challenge-inconsistent-data-entry

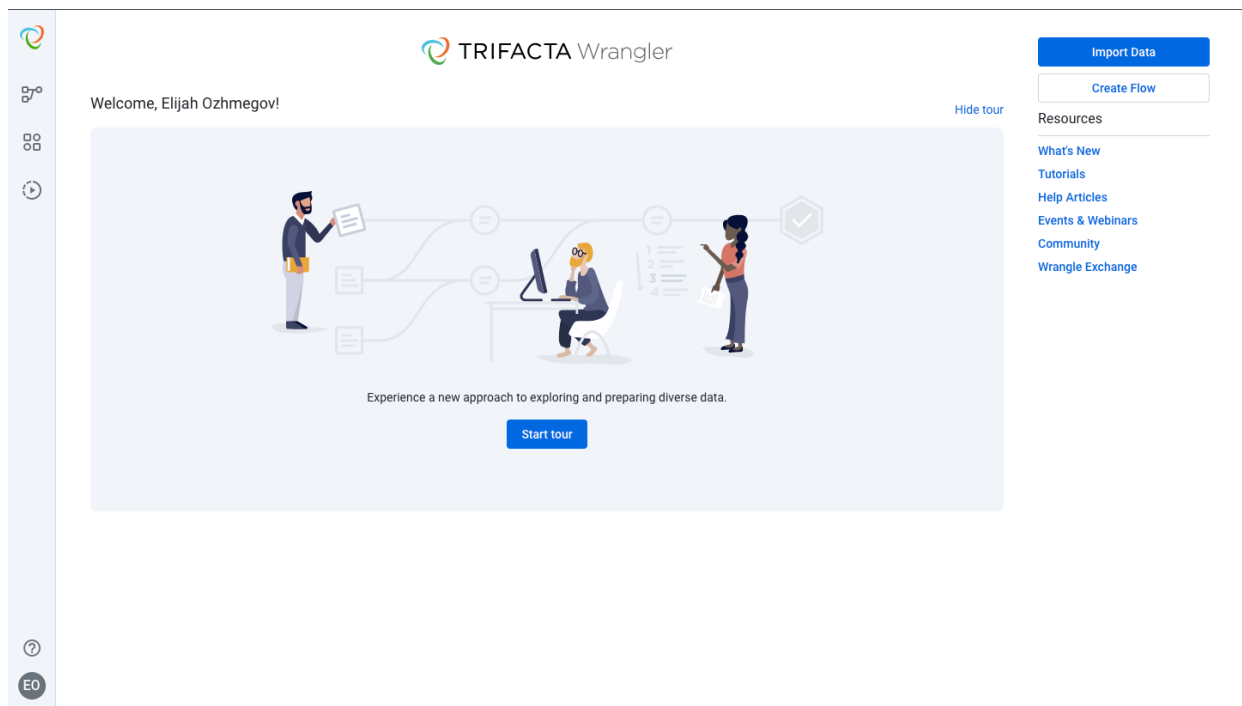
Useful libraries:

- chardet — to determine a text encoding
- fuzzywuzzy — to measure a distance between two different words and replace one of them

B. Playing around with Trifacta

1. What I saw after registration

My name here is Elijah instead of Ilia, I hope it is ok.



2. Then, I uploaded the dataset, that I have web-scraped for my BI project. It have information about commits of one company.

Library

All Data

Imported Datasets

References

Macros

All Data

Import Data

Find Datasets...

Name	In Flows	Source	Last Updated
github_cleaned.csv	0	Upload	Today at 2:18 AM

3. Then, I look at the data itself. It gave me immediately the year graph, which I had to draw with python package **matplotlib**. I think it is useful.

GITHUB_CLEANED FLOW

github_cleaned

Full Data

Details

#	id	rbc	nickname	datetime	date	year	#	mc
1 - 5.08k	5885	ArjaanBuijk		2020-01-16 21:56:24+00:00	2020-01-16	2020		
	5884	SachinKalsi		2020-01-16 17:29:15+00:00	2020-01-16	2020		
	5883	alwx		2020-01-16 12:51:26+00:00	2020-01-16	2020		
	5882	JEM-Mosig		2020-01-16 12:13:44+00:00	2020-01-16	2020		
	5881	wochinge		2020-01-16 12:01:33+00:00	2020-01-16	2020		
	5880	alwx		2020-01-16 10:03:38+00:00	2020-01-16	2020		
	5879	tabergma		2020-01-15 15:57:59+00:00	2020-01-15	2020		
	5878	melindaLoubser1		2020-01-15 14:54:19+00:00	2020-01-15	2020		
	5877	alwx		2020-01-15 14:03:39+00:00	2020-01-15	2020		
	5876	melindaLoubser1		2020-01-15 13:07:53+00:00	2020-01-15	2020		
	5875	nbeuchat		2020-01-15 11:21:33+00:00	2020-01-15	2020		
	5874	RaminParker		2020-01-15 10:30:20+00:00	2020-01-15	2020		
	5873	srinathgnath		2020-01-15 07:57:18+00:00	2020-01-15	2020		
	5872	tmbo		2020-01-14 22:57:47+00:00	2020-01-14	2020		
	5871	staticdev		2020-01-16 08:31:17+00:00	2020-01-16	2020		
	5870	alwx		2020-01-14 10:16:35+00:00	2020-01-14	2020		
	5869	nmohona		2020-01-14 10:53:01+00:00	2020-01-14	2020		
	5868	tdzienniak		2020-01-14 09:01:45+00:00	2020-01-14	2020		
	5867	ncoder		2020-01-16 10:09:46+00:00	2020-01-16	2020		
	5866	mtworld		2020-01-14 07:00:04+00:00	2020-01-14	2020		
	5865	tabergma		2020-01-13 14:08:33+00:00	2020-01-13	2020		
	5864	tabergma		2020-01-13 18:00:33+00:00	2020-01-13	2020		
	5863	alwx		2020-01-16 10:11:19+00:00	2020-01-16	2020		
	5862	koaning		2020-01-13 10:16:21+00:00	2020-01-13	2020		
	5861	WahabBhatti		2020-01-13 07:00:05+00:00	2020-01-13	2020		
	5860	muelerma		2020-01-11 13:16:16+00:00	2020-01-11	2020		
	5859	muelerma		2020-01-10 19:51:17+00:00	2020-01-10	2020		

6 Columns, 5,079 Rows, 5 Data Types

RBC nickname

Quality

Valid	Mismatched	Missing
5079	0	0
100%	0%	0%

Unique Values

tmbo	wochinge	ann41	erohmensing	tabergma
220	195	173	148	122

Show more values...

Patterns

{alpha}+{alpha-numeric}+	{lower}+	{lower}{3}-{lower}{4}	{lower}{6}-{lower}{6}	{lower}{6}-{lower}{3}
4,754	2,729	15	9	5

Show pattern details...

Suggestions

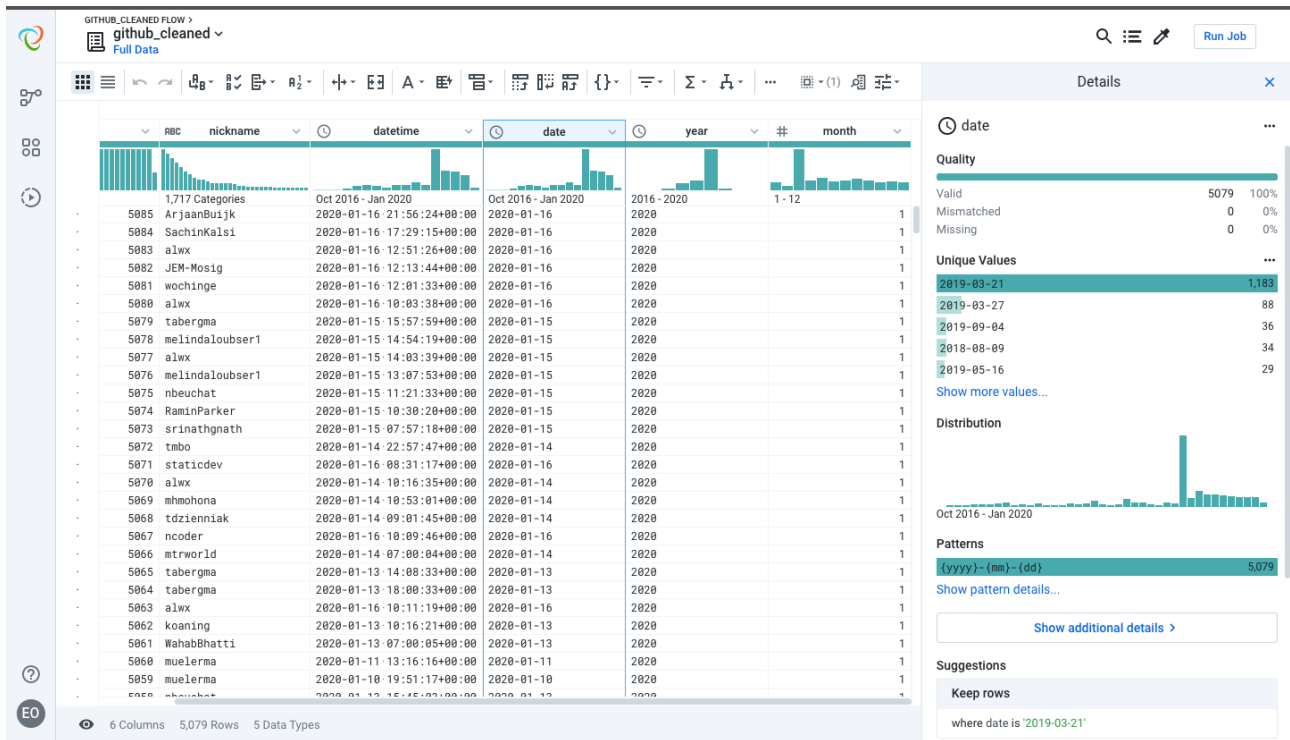
Delete columns

nickname

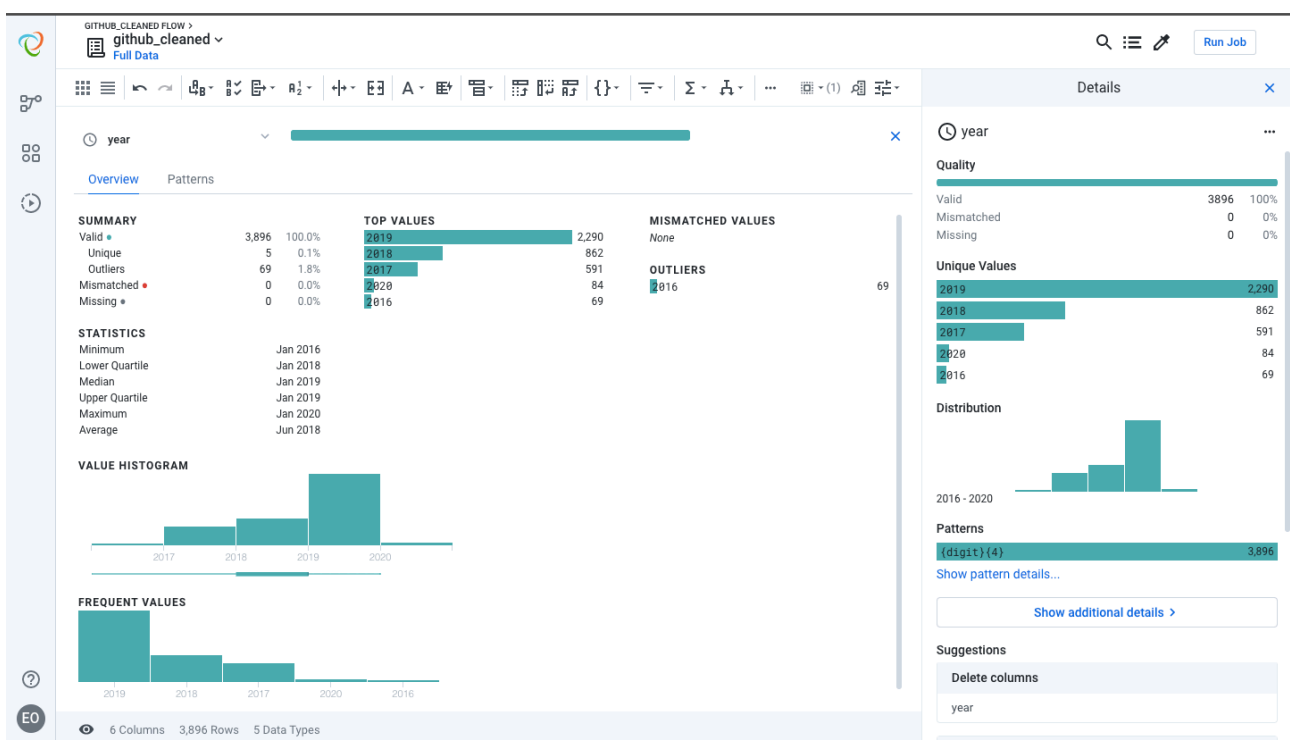
Rename

Rename nickname to 'nickname'

- Then, I decided to challenge the platform and remove all lines for March 21 of 2019, because it was a cheating day in the company, I think the company wanted to get some more money and artificially increased the number of commits. You can easily see it in the right panel.



- And now you can see more realistic and less impressive values per year.



Conclusion:

I think **trifacta** provide a good opportunity to look at massive datasets and make fast conclusions, but it still needs to be paid and it is not flexible as python packages or R set of tools. So I do not think I will use it in the future.