Beuth Hochschule für Technik Berlin

# Computer Science für Big Data Übg

A4-Data Preparation Exercise

Februrary 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

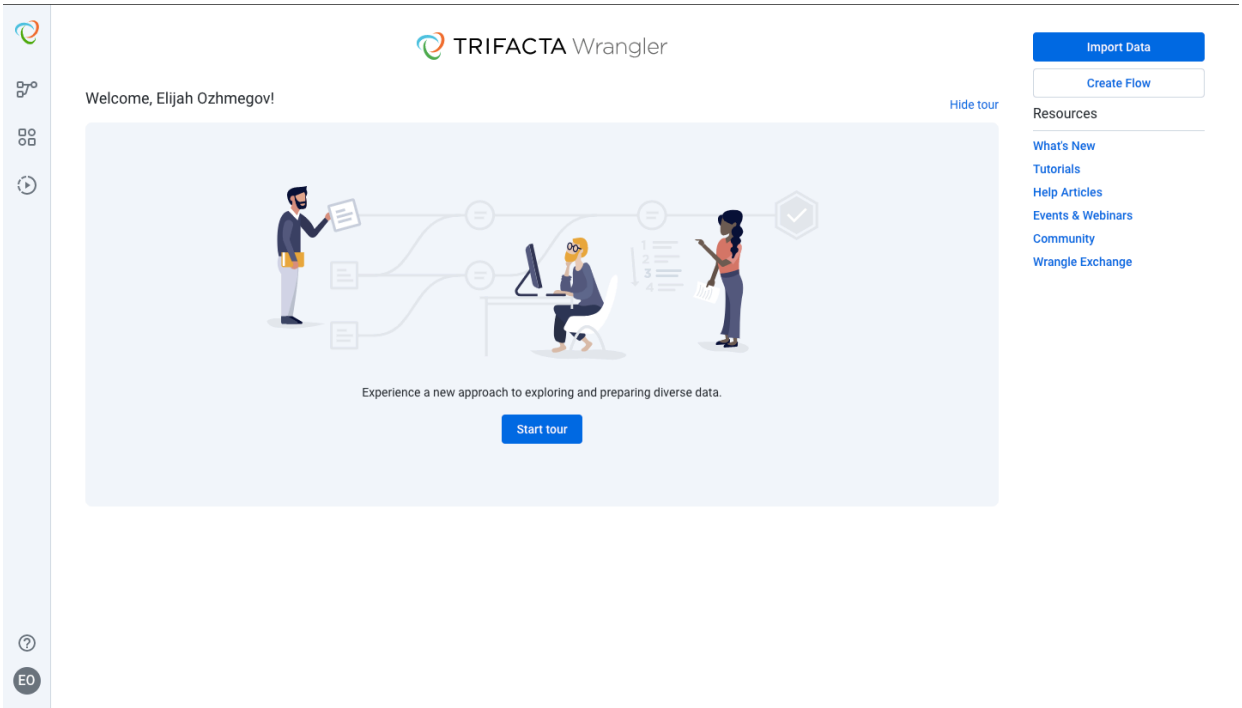
Useful libraries:

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

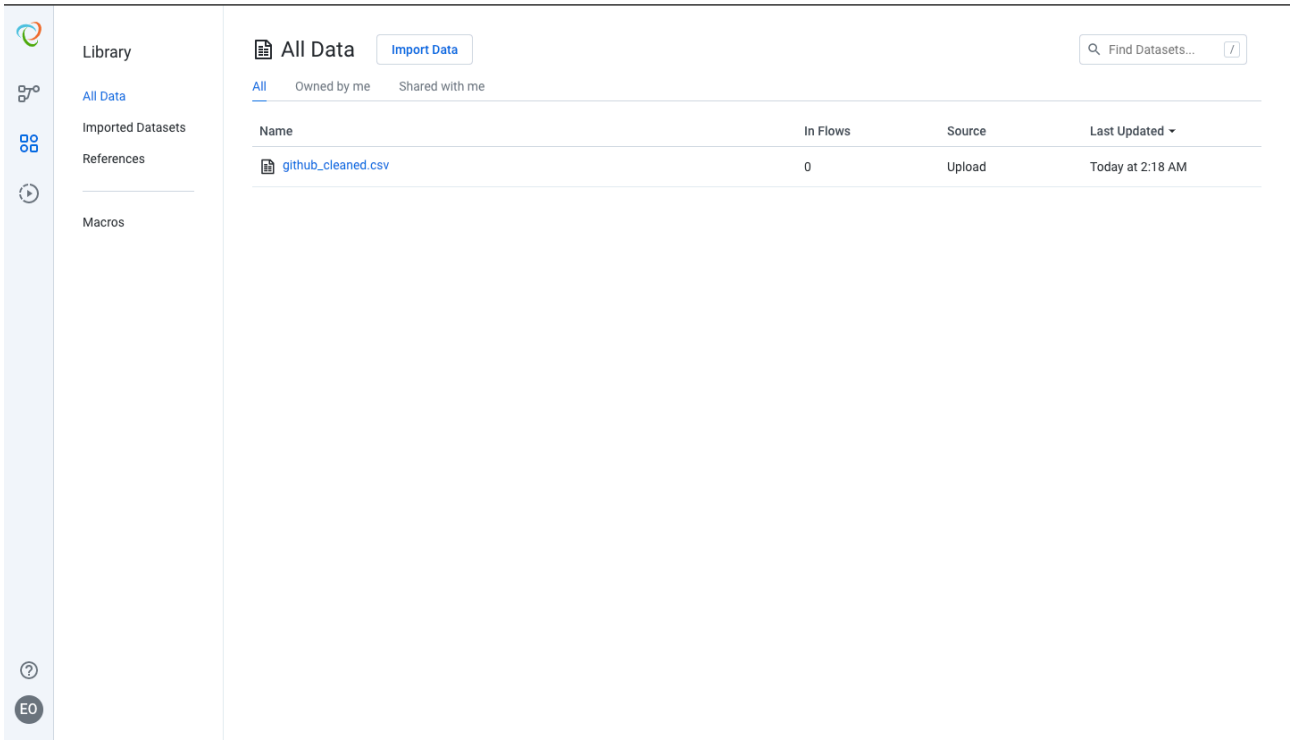
## B. Playing around with Trifacta

**1.** What I saw after registration

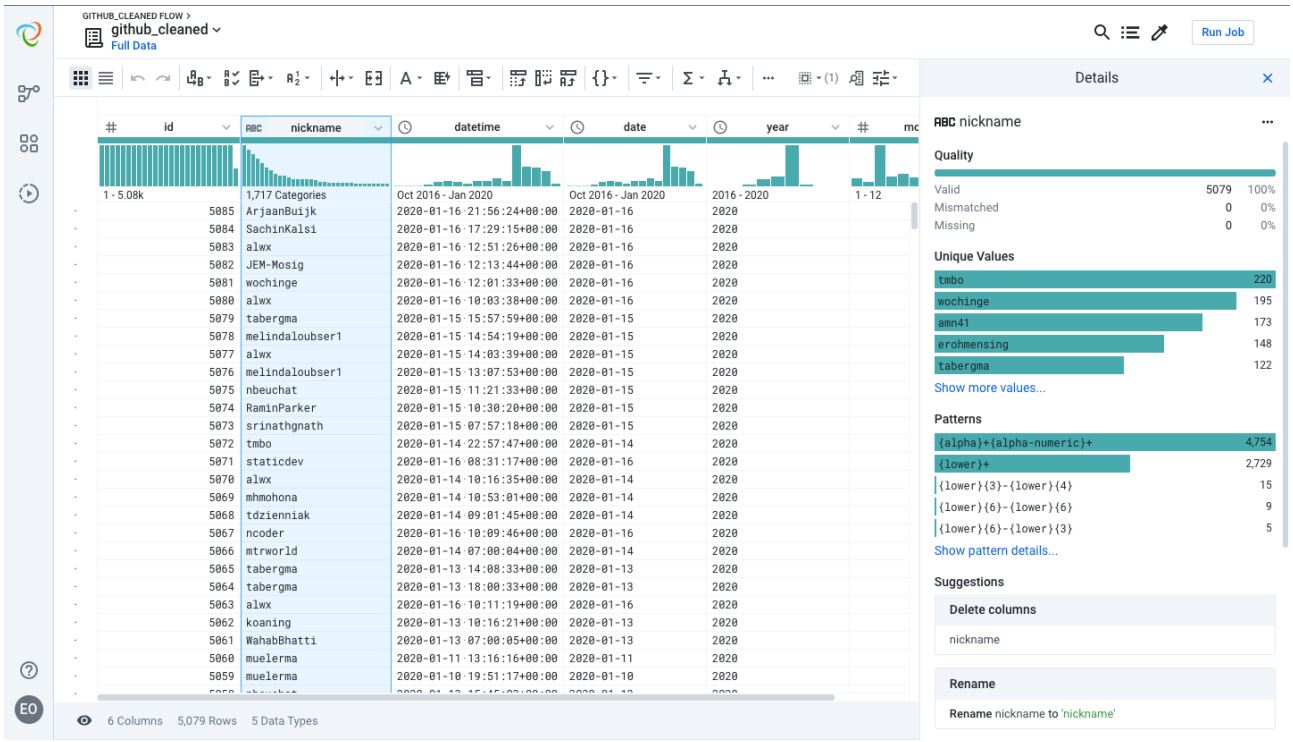
My name here is Elijah instead of Ilia, I hope it is ok.



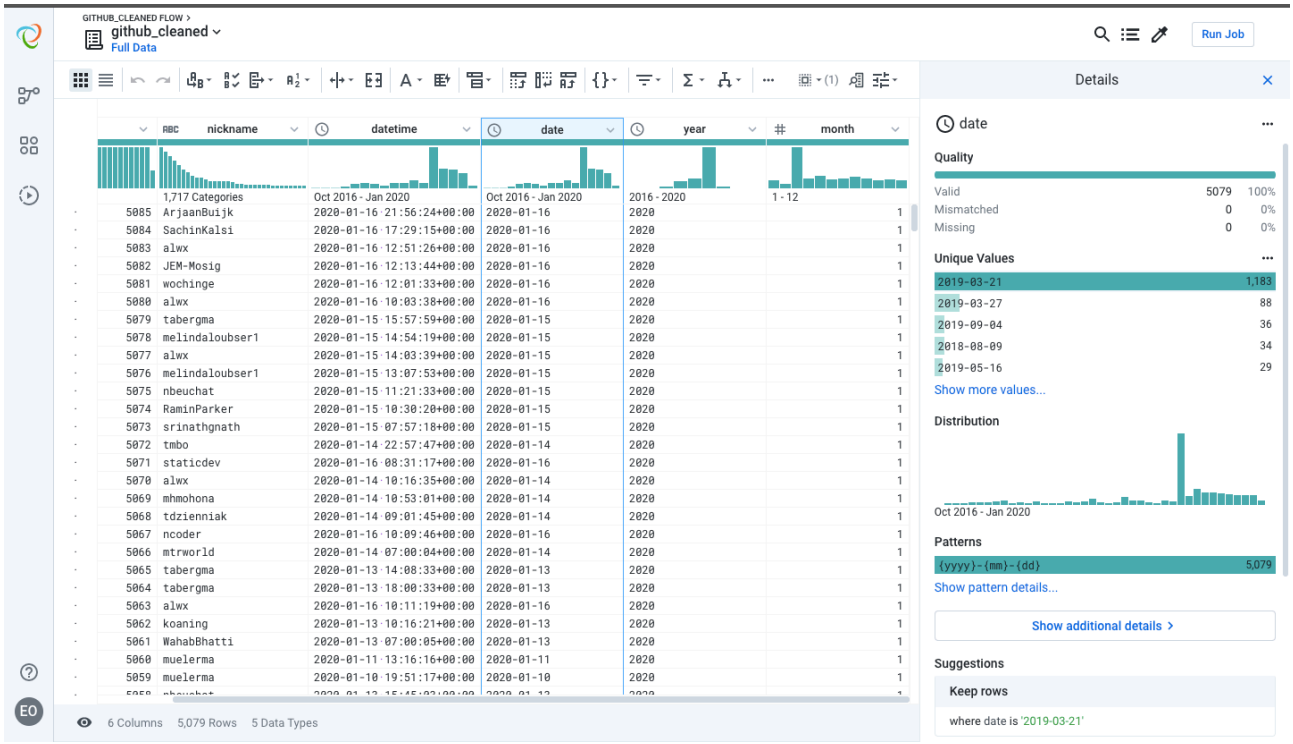
**2.** Then, I uploaded the dataset, that I have web-scrapped for my BI project. It have information about commits of one company.



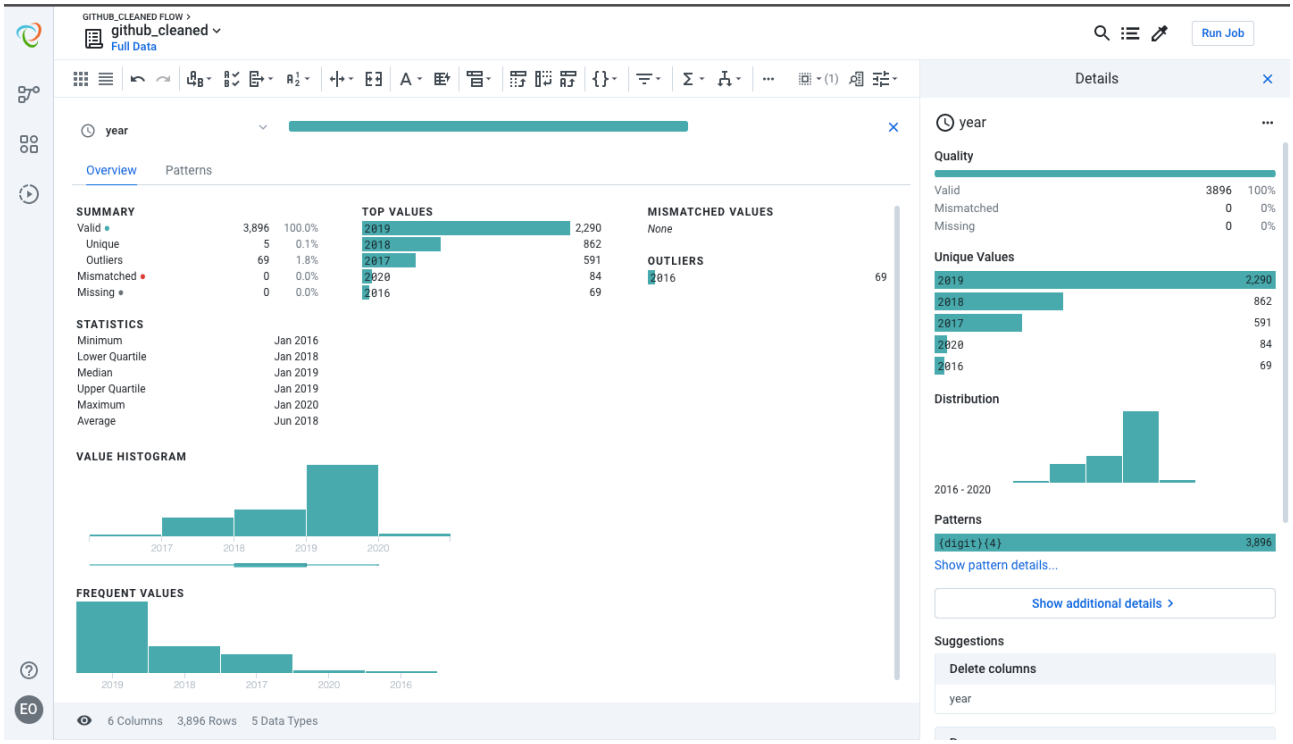

**3.** Then, I look at the data itself. It gave me immediately the year graph, which I had to draw with python package **matplotlib**. I think it is useful.



4. 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.




5. 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.