



SFU

DataPrep - Accelerate Data Preparation for AI

Jiannan Wang, Associate Professor
Director, Professional Master's Program
School of Computing Science
Simon Fraser University
<http://www.cs.sfu.ca/~jnwang>

Oct 07, 2021, NLP Zurich



~1000 Stars



200K+ Downloads



25+ Contributors

From Model-Centric to Data-Centric



“If 80 percent of our work is data preparation, then ensuring data quality is the important work of a machine learning team.”

Andrew Ng

Data Preparation Is **Still** the Bottleneck!!!

2014

The New York Times

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

Yet far too much handcrafted work — what data scientists call “data wrangling,” “data munging” and “data janitor work” — is still required. Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.

<https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html>

2020

 ANACONDA

The State of Data Science 2020 Moving from hype toward maturity

We were disappointed, if not surprised, to see that data wrangling still takes the lion's share of time in a typical data professional's day. Our respondents reported that almost half of their time is spent on the combined tasks of data loading and cleansing. Data

<https://www.anaconda.com/state-of-data-science-2020>

Three Questions

1. What makes data preparation hard?
2. Why has this problem not been solved?
3. How to solve it in the next 5-10 years?

Three Questions

1. What makes data preparation hard?
2. Why has this problem not been solved?
3. How to solve it in the next 5-10 years?

Why Is Data Preparation Hard?



Collection



Cleaning



Integration



Analysis

How much time is spent on preparation?

1. **Too many small problems** (e.g., standardize date, dedup address, etc)
2. Humans have **different levels of expertise** (in data science and programming)
3. **Domain specific** (finance, social science, healthcare, economics, etc.)

Three Questions

1. What makes data preparation hard?
- 2. Why has this problem not been solved?**
3. How to solve it in the next 5-10 years?

Human-in-the-loop Data Preparation

Three Directions:

- Spreadsheet GUI
- Workflow GUI
- Notebook GUI

Spreadsheet GUI

CUSTOMER ANALYSIS >
customer ▾
Random

Run Job

Preview

#	IMSI	CONTRACT_END	CONTRACT_START	#	SUBSCRIBER_AGE	RBC	STATUS
310T - 310.26T		Jan 2013 - Dec 2016	Jan 2000 - Dec 2014	0 - 15			2 Categories
310170226812721	6/4/16	7/29/09					ACTIVE
310160900766700	3/28/15	10/6/13		1			ACTIVE
310170546822541	9/23/16	1/9/07		7			ACTIVE
310005432849230	5/29/15	2/14/01		13			ACTIVE
310026939721905	9/11/15	9/18/10		4			ACTIVE
310026015466952	8/27/15	3/13/06		8			ACTIVE
310170484724861	1/16/16	5/11/04					ACTIVE
310170765640471	05-Jul-2011	9/11/06		4			INACTIVE
310260310245556	12/24/15	3/28/01		13			ACTIVE
310150834295817	3/6/15	7/26/00		14			ACTIVE
310160464252516	9/25/15	4/4/04		10			ACTIVE
310120438750772	4/30/16	9/8/04		10			ACTIVE
310260195729676	1/16/15	1/3/04		11			ACTIVE
310026261822880	8/13/13	11/23/08		4			INACTIVE
310005667082048	8/4/16	10/22/14					ACTIVE
310170836020164	1/22/15	10/19/14		0			ACTIVE
310160772267782	11/21/15	12/28/14					ACTIVE
310170116249240	27-Sep-2011	2/9/09					INACTIVE
310026110612337	5/29/15	3/29/05		9			ACTIVE
310260681676970	11/17/16	5/21/07		7			ACTIVE
310004436630316	9/15/16	7/24/11					ACTIVE
310120423699542	2/27/15	6/29/11		3			ACTIVE
310120773194729	4/28/16	6/15/04		10			ACTIVE
310030295859214	2/7/15	3/24/12		2			ACTIVE
310012150088547	13-Jan-2009	12/10/05		3			INACTIVE
310120387060694	10/1/16	10/25/11		3			ACTIVE

Pattern Details CONTRACT_END

Hide Example Values

12.65k

m / dd / yy

9/18/15
6/13/15
5/21/15
12/12/15
1/16/16

5.37k

m / d / yy

6/5/15
4/4/15
12/8/16
7/2/14
11/6/15

dd - month-abbrev - yyyy

14-Nov-2012
11-Jul-2007
20-Jul-2010

Trifacta

19 Columns 20,000 Rows 8 Data Types

Show only affected ☐ Rows

Workflow GUI

Input Data

- Customers CRM
- Transactions AWS

Data Preparation

- Crosstab
- Filter
- Formula / Calculate Fields
- Summarize / Pivot Table

Data Blend

- Join
- VLOOKUP
- Visualytics

Profile: ZIP

Frequency

Showing only top 20 unique values

Data Quality

0.0 % NOT OK 0.0 % NULL 0.0 % EMPTY 100.0 % OK

ZIP	
Data Type	V_String
Size	255
Non-Nulls	1735
Uniques	86
Nulls	0
Blanks	0
Values with Leading Whitespace	0
Values with Trailing Whitespace	0
Shortest (Non-Blank) Length	5
Average Length	5.0

Results - Browse (41) - Input

14 of 14 Fields | Cell Viewer | 1,735 records displayed, 121 KB

Record #	Customer ID	Address	City	Customer_Segment	First_Name	Last_Name	Responder	State	Store_Number	Suite	ZIP
1	5	5360 Zuni St	Denver	Home Office	LINDA	TREVINO	No	CO	100	[Null]	80221
2	6	1599 Williams St	Denver	Home Office	H	MACK	No	CO	106	[Null]	80218
3	7	12066 E Lake Cir	Greenwood Village	Home Office	MARISSA	LATTA	No	CO	105	[Null]	80111
4	8	7225 S Gaylord St	Centennial	Home Office	PHYLLIS	WALKER	No	CO	101	[Null]	80122
5	9	4497 Cornish Way	Denver	Home Office	VIVIAN	GAULDEN	No	CO	105	[Null]	80239

Notebook GUI

The screenshot shows a Jupyter Notebook interface. The top menu bar includes File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. On the right, it says 'Trusted' and 'Python 3'. Below the menu is a toolbar with icons for saving, adding, deleting, copying, pasting, undo, redo, running, and other actions. The main area contains a code cell with the following Python code:

```
In [1]: import qgrid
import numpy as np
import pandas as pd

df = pd.DataFrame(np.random.rand(3,5))
qg = qgrid.show_grid(df)
qg
```

Below the code cell, a table is rendered, showing the output of the code. The table has 6 columns and 4 rows (including the header).

	0	1	2	3	4
0	0.09835	0.65988	0.24213	0.1324	0.21811
1	0.11348	0.17392	0.18452	0.25759	0.45406
2	0.77328	0.2921	0.37934	0.87462	0.93311

What happened to the past?

SFU

Spreadsheet/Workflow GUIs

Targeted at non-programmers



Three Questions

1. What makes data preparation hard?
2. Why has this problem not been solved?
- 3. How to solve it in the next 5-10 years?**

Our Vision

SFU

Machine Learning Made Easy



Initial release: 2007; 14 years ago

Deep Learning Made Easy



Initial release: 2016; 5 years ago

Data Preparation Made Easy



Initial release: 2019; 2 years ago

Unique Selling Points

DataPrep vs Commercial Software

- Open Source
- Python Data Science Ecosystem
- Notebook GUI

DataPrep vs Existing Python Libraries

- All-in-one
- Fast
- Task-Centric API Design

DataPrep Components

May 2019 - Now	DataPrep.EDA	Simplify Exploratory Data Analysis
Nov 2019 - Now	DataPrep.Connector	Simplify Web Data Collection
Sept 2020 - Now	DataPrep.Clean	Simplify Data Cleaning
Planning	DataPrep.Feature	Simplify Feature Engineering
Planning	DataPrep.Integrate	Simplify Data Integration

DataPrep.EDA

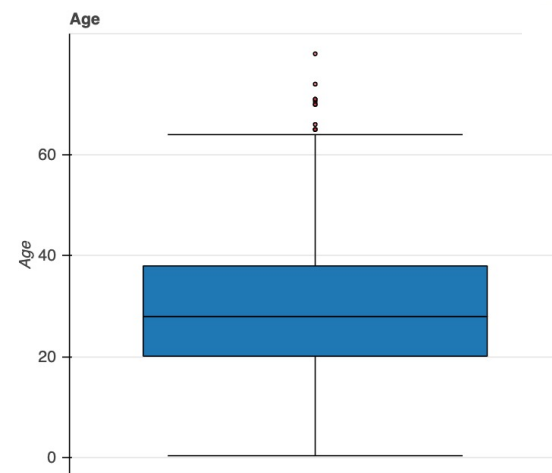
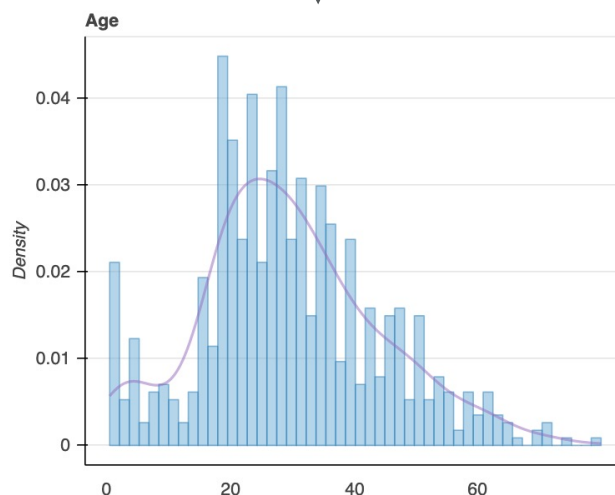
Task-Centric Exploratory Data Analysis

Exploratory Data Analysis (EDA)

Understand data and discover insights
via data visualization, data summarization, etc.

Understand “Age” column

Minimum	0.42
5-th Percentile	4
Q1	20.125
Median	28
Q3	38
95-th Percentile	56
Maximum	80
Range	79.58
IQR	17.875



Solution 1: Plot-Centric EDA

Pandas + Matplotlib

☹ Hard to Use

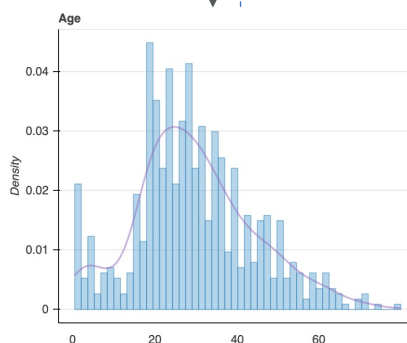
- Need to write lengthy and repetitive code

Understand “Age” column

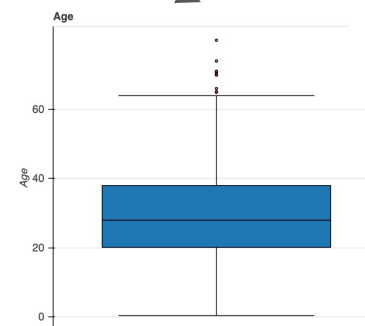
Write Code

Minimum	0.42
5-th Percentile	4
Q1	20.125
Median	28
Q3	38
95-th Percentile	56
Maximum	80
Range	79.58
IQR	17.875

Write Code



Write Code



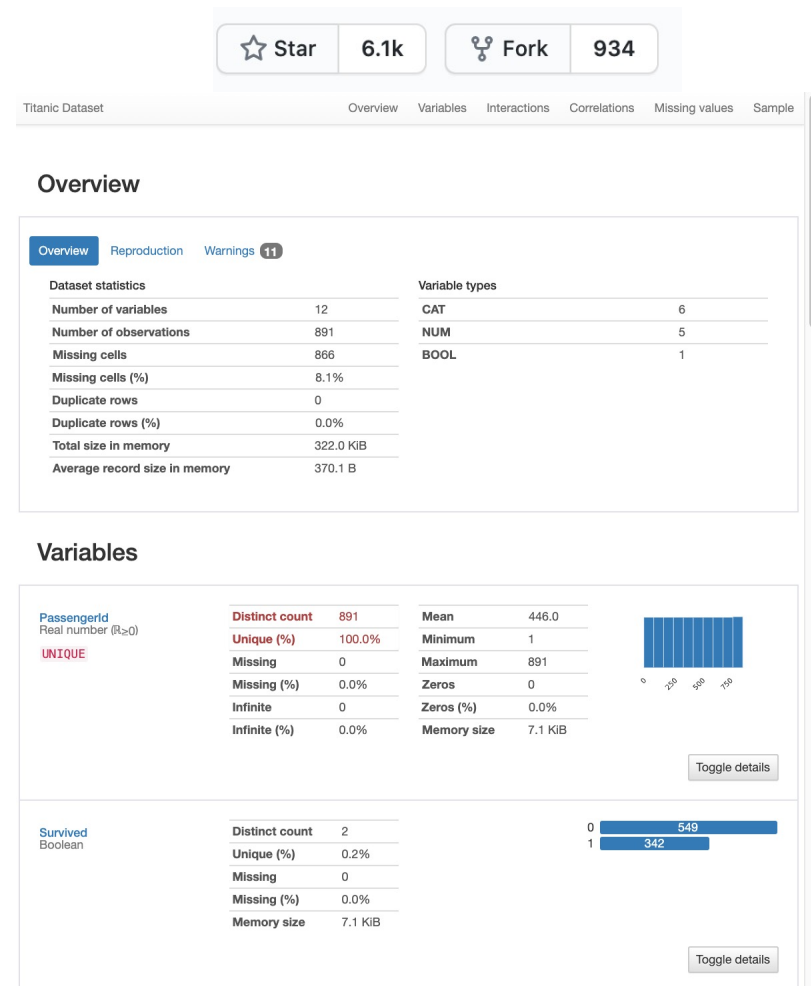
Solution 2: Profiling-Centric EDA

Pandas-profiling










🐢 Slow

🐢 Hard to Customize

```
profile = ProfileReport(df, title="Pandas Profiling Report")
```



Our Solution: Task-Centric EDA

EDA Solutions	Easy to Use	Interactive Speed	Easy to Customize
1. Pandas + Matplotlib			
2. Pandas-profiling			
3. DataPrep.EDA			

Key Ideas:

- Declarative
- Support both coarse-grained and fine-grained EDA tasks



Posted by u/jnwang 7 days ago

Understand your data with a few lines of code in seconds using DataPrep.eda



I Made This

samdof 4 points · 7 days ago

I'll look into it and get back to you. By the way what you guys are doing is amazing and have the potential to be a game-changer if it cut some time out of data prep.

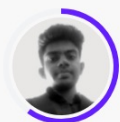
apivan191 3 points · 5 months ago

This will save me so much time even just exploring my data, not to mention coding all of it up. You've done good in the world

cym13 9 months ago

Time will tell if this is the right solution, but at least I think you're tackling the right problem. Thank's for sharing.

Why do users love DataPrep?



Kannan Ravinther

Topic Author

Best Way to do EDA (AutoEDA)

Posted in [General](#) 14 days ago

kaggle

▲
13

I recently got to know about Pandas Profiling which can generate EDA report in a few seconds. But now, I found a better option called **DataPrep**, it is faster than Pandas Profiling and also customizable.👍

7 Cool Python Packages Kagglers Are Using Without Telling You

Let me reveal the secrets...



Bex T. Aug 6 · 8 min read ★

towards
data science

- **DataPrep** — the most comprehensive auto EDA [[GitHub](#), [Documentation](#)]

<https://www.kaggle.com/general/273848#1521057>

<https://towardsdatascience.com/7-cool-python-packages-kagglers-are-using-without-telling-you-e83298781cf4>



GettingStarted Prediction Competition

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics



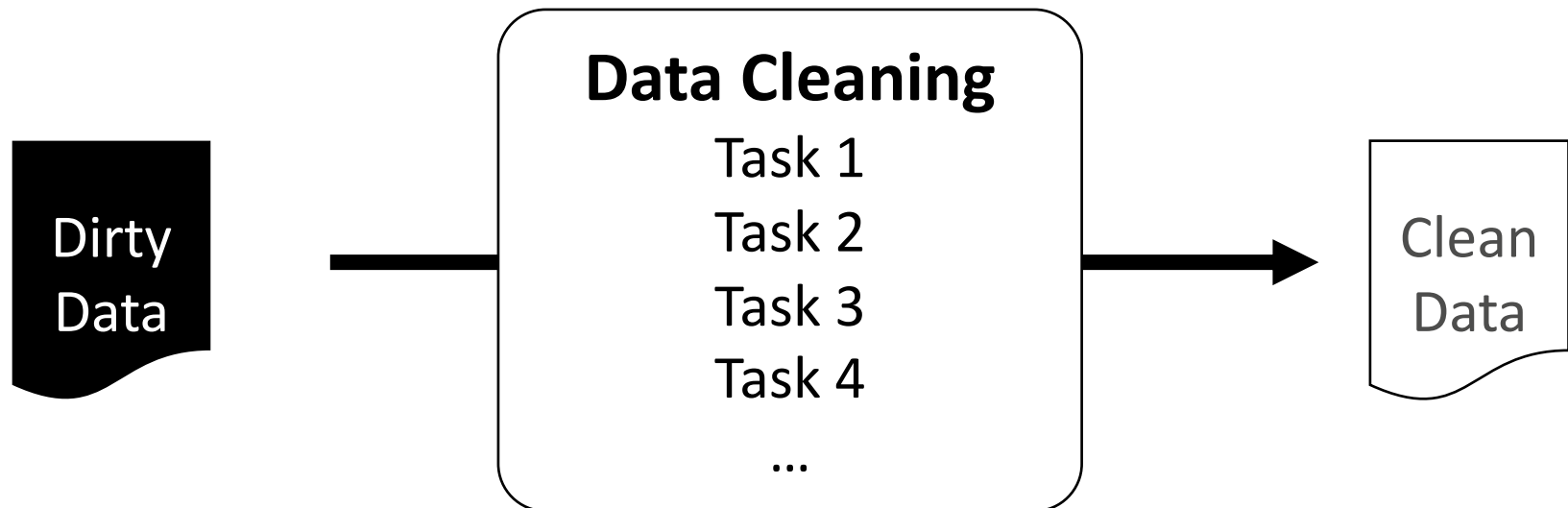
Kaggle · 19,840 teams · Ongoing

Demo Time

DataPrep.Clean

Task-Centric Data Cleaning

Data Cleaning



Example: DataPrep.Clean

	phone.....num	DATE★	CITY
0	(555) 234-5678	25-Sep-03	Quebec
1	555.234.5678	Sep-25-2003	Vancouver
2	555/234/5678	09-25-2003	Vancouver
3	800-299-JUNK	2003 09 25	Ottowa
4	15551234567	10-09-2003	vancouver
5	(1) 555-234-5678	2010-09-03	Vancouver
6	+1 (234) 567-8901 x. 1234	2003.Sep.25	Toronto
7	2345678901 extension 1234	2003-09-25	Toront
8	2345678	2003 Sep 25	Tronto
9	800-299-JUNK	2003 09 25	Ottowa
10	1-866-4ZIPCAR	Sep of 03 2003	otowa
11	555-234-5678	2003-Sep-25	Québec

Data Cleaning

1. clean_header
2. clean_phone
3. clean_date
4. clean_duplication

Demo Time

DataPrep.Connector

A Unified API Wrapper
to Simplify Web Data Collection

Data Collection Through Restful APIs



Social Data



Business Data



Event Data



Publication Data

 [public-apis](#) / [public-apis](#) 

A collective list of free APIs for use in software and web development.

 [ultimatecourses.com](#)

☆ 98.4k stars  12k forks

Index

- [Animals](#)
- [Anime](#)
- [Anti-Malware](#)
- [Art & Design](#)
- [Books](#)
- [Business](#)
- [Calendar](#)
- [Cloud Storage & File Sharing](#)
- [Continuous Integration](#)
- [Cryptocurrency](#)
- [Currency Exchange](#)
- [Data Validation](#)
- [Development](#)
- [Dictionaries](#)
- [Documents & Productivity](#)
- [Environment](#)
- [Events](#)
- [Finance](#)

The Unified API

1. Connect

```
conn = connect(website_name, _auth, _concurrency)
```

2. Understand (Optional)

```
conn.table_names  
conn.show_schema(table_name)
```

3. Query

```
conn.query(table_name, query_parameter_list, _count)
```

Why is it a good idea?

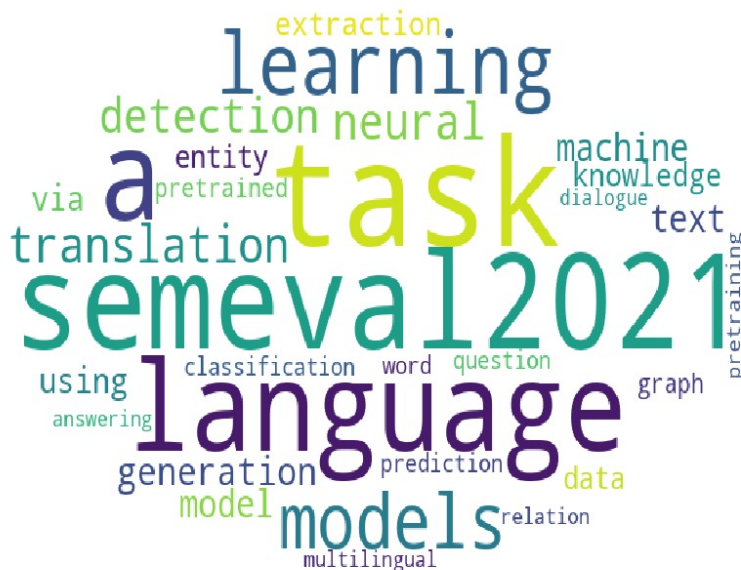
- Easy to maintain
- Easy to learn

Hot Topics in ACL 2021

```
from dataprep.connector import Connector
from dataprep.eda import plot
```

```
# Write your code
```

3 lines of code



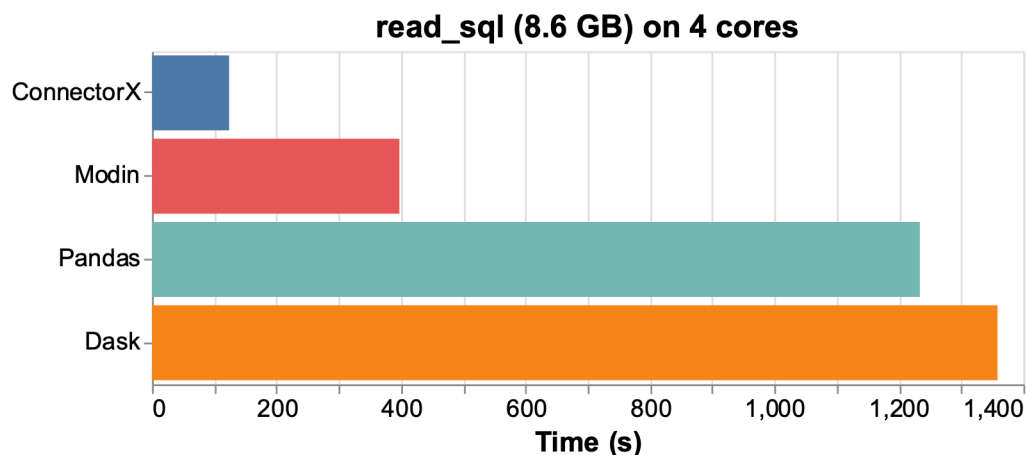
Roadmap 2021

DataPrep.EDA

- Make plots look **attractive**
- Understand **multiple** dataframes (plot_db, plot_lake...)

DataPrep.Connector <http://cx.dataprep.ai>

- Speed up **read_sql()** with arrow and parallel connection



Roadmap 2021

DataPrep.Clean

- Goal: Implement 100+ `clean_{type}(df, x)` functions
- Example: `clean_email`, `clean_date`, `clean_phone`, `clean_country`, etc.
- Application: Data Validation, Data Standardization, Semantic Type Detection

```
1 import pandas as pd
2 from dataprep.clean import clean_country
3 df = pd.DataFrame({"country": ["USA", "country: Canada", " France ", "233", " tr "]})
4 clean_country(df, "country")
```

	country	country_clean
0	USA	United States
1	country: Canada	Canada
2	France	France
3	233	Estonia
4	tr	Turkey

Join the DataPrep Community



```
pip install -U dataprep
```

Thank you!

<http://dataprep.ai>



~1000 Stars



200K+ Downloads



25+ Contributors