


This is a graded discussion: 10 points possible

due -

12 26

Tidy data practice

1. Find a tabular dataset (e.g. check out <https://github.com/yy/dviz-course/wiki/Resources#datasets>  [_.\(https://github.com/yy/dviz-course/wiki/Resources#datasets\)_](https://github.com/yy/dviz-course/wiki/Resources#datasets))
2. Determine whether it's tidy or not.
3. If it's tidy, explain what are the columns and rows and then explain why it's tidy.
4. If not, explain why not and how you can make it tidy.

Check out others' answers and discuss!



← **Reply.**

○




<https://iu.instructure.com/courses/2165942/users/6701715>

Dustin Cole (<https://iu.instructure.com/courses/2165942/users/6701715>)

Monday

⋮

1. I chose the kaggle school shooting dataset from Joakim Arvidsson. **School Shootings | Kaggle**  [_.\(https://www.kaggle.com/datasets/joebeachcapital/school-shootings\)_](https://www.kaggle.com/datasets/joebeachcapital/school-shootings)
2. This dataset is already tidy.
- 3-4.
 - None of the column headers are values. The headers are school id, school name, district id, district name, date, etc. If the columns were something like "shootings in 1990s, shootings in 2000s" it would break this rule.
 - Multiple variables are not stored in one column. If a column was "district_date" then it would break this rule.
 - Multiple types are not in one table since it is only providing a line per shooting with the variables as columns. (The data is not normalized)

school-shootings-data.csv (110.48 kB)

Detail Compact Column

About this file

School Shootings

uid	nces_school_id	school_name	nces_district_id	district_name	date	school_year	year	time	day_of_week
Unique ID	NCES School ID	School Name	NCES District ID	District Name	Date	School Year	Year	Time	Day of Week
1	391	2%	370 unique values	100k	5.52m	19Apr99	4Jun23	1999	2023
1	062271003034	1%	08848080787	Jefferson County R-1	4/28/1999	1998-1999	1999	11:19 AM	Tuesday
2	228054080422	98%	2280540	East Baton Rouge Parish School Board	4/22/1999	1998-1999	1999	12:38 PM	Thursday
3	138441801591	Heritage High School	1384418	Rockdale County	5/28/1999	1998-1999	1999	8:83 AM	Thursday
4	421899803847	John Bartram High School	4218998	Philadelphia City SD	18/4/1999	1999-2000	1999	18:08 AM	Monday
5	25827988225	Dorchester High School	2582798	Boston	11/3/1999	1999-2000	1999	7:48 AM	Wednesday
6	35886988248	Deming Middle School	3588698	Deming Public Schools	11/19/1999	1999-2000	1999	12:45 PM	Friday

[Reply](#) (1 like)

[https://](https://iu.instructure.com/courses/2165942/users/6679606)

[Sneha Satish \(https://iu.instructure.com/courses/2165942/users/6679606\)](https://iu.instructure.com/courses/2165942/users/6679606)

Monday

The dataset I am choosing is the 'Car Prices Jordan 2023' from kaggle . Here is the link to it.

<https://www.kaggle.com/datasets/farahalarbeed/car-prices-jordan>

(<https://www.kaggle.com/datasets/farahalarbeed/car-prices-jordan>)

This is what it looks like:

ID	Model	Property	Power	Price
0	Byd F0 2018	Manual	1000 CC	6,900
1	Suzuki Alto 2023	manual	800 CC	8,250
2	Suzuki Celerio 2019	Automatic	1000 CC	10,499
3	Changan E Star 2023	Automatic	0 CC	10,990
4	Hyundai Grand i10 2020	Automatic	1250 CC	11,500
5	Baic X35 2023	manual / MT	1500 CC	11,900
6	Nissan Micra 2020	Automatic	1500 CC	11,950
7	Kia Picanto 2021	Automatic	1200 CC	12,210
8	Hyundai Atos 2021	Manual	1100 CC	13,000
9	Kia Pegas 2021	Manual	1400 CC	13,400
10	Mitsubishi Mirage 2021	Automatic / GLX	1200 CC	13,950
11	Changan V7 2021	Automatic / Comfort	1500 CC	14,000
12	Changan V3 2023	Automatic	1500 CC	14,000
13	Kia Picanto 2023	Automatic	1200 CC	14,100

This dataset is tidy for the following reasons:

- Each variable forms a column: ID, Model, Property, Power, Price are distinct columns representing variables.
- Each observation forms a row: Each row corresponds to a single car.

- Each type of observational unit forms a table: This dataset contains information about cars, their type, power and price and there is no need of any additional tables.
- Column headers are variable names not values.
- There is no duplicate data.

← Reply 👍 (1 like)



Prem Amal (<https://iu.instructure.com/courses/2165942/users/6684842>)

Monday

Please find the referred dataset below:

Weather Data - Boston (Jul 2012 - Aug 2015) | Kaggle 📄

(<https://www.kaggle.com/datasets/naveenpandianv/weather-data-boston-jul-2012-aug-2015>)

In this study, data was collected from Kaggle, specifically from the "Weather Data - Boston Jul 2012-Aug 2015" dataset provided by Naveen Pandian. The dataset contains daily weather measurements, including temperature, humidity, dew point, wind, precipitation, and date, from July 2012 to August 2015. The data was collected from a weather station located in Boston, Massachusetts, USA.

Each Variable in a Column: It appears that each variable (precipitation, day, month, year, temperature, dewpoint, humidity, wind) is represented in separate columns, which aligns with the tidy data principle.

Each Observation in a Row: Each row represents a single observation or data point, which is also in line with the tidy data principle.

Each type of Observational Unit in a Table: From the provided glimpse, it seems that all the data pertains to the same type of observational unit (e.g., weather measurements for Boston over time). This also aligns with the tidy data principle.

There is no missing data here, we have clear column names and consistent data types. The given dataset follows the tidy principles, hence the dataset is tidy.

← Reply 👍



(https://

Thomas Jablenski (<https://iu.instructure.com/courses/2165942/users/6701599>)

⋮

Monday

The dataset that I chose comes from

<https://github.com/fivethirtyeight/data/blob/master/bob-ross/elements-by-episode.csv>

(<https://github.com/fivethirtyeight/data/blob/master/bob-ross/elements-by-episode.csv>) which contains occurrences of themes on Bob Ross shows for all episodes. This is not a tidy set of data. It is a wide dataset with many columns but can be summarized relatively easily. The columns include the episode, title of the episode, and then several boolean based columns depicting if the theme was shown within the episode. The one issue that this dataset has is the episode column is a combination of 2 different variables. It combines season and episode. If this column were to be broken into 2 different columns called season and episode this would be a tidy dataset.

← [Reply](#)



(http

Erik Gonzalez (<https://iu.instructure.com/courses/2165942/users/6352173>)

⋮

Tuesday

I selected the same dataset but missed the opportunity to split out season and episode, good catch!

← [Reply](#)



(https://

Gary Croke (<https://iu.instructure.com/courses/2165942/users/6706306>)

⋮


Tuesday

I selected the airline dataset from FiveThirtyEight. Here are the first few rows from the original table:

airline	avail_seat_km_per_week	incidents_85_99	fatal_accidents_85_99	fatalities_85_99	incidents_00_14	fatal_accidents_00_14	fatalities_00_14
Aer Lingus	320906734	2	0	0	0	0	0
Aeroflot*	1197672318	76	14	128	6	1	88
Aerolineas Argentinas	385803648	6	0	0	1	0	0
Aeromexico*	596871813	3	1	64	5	0	0
Air Canada	1865253802	2	0	0	2	0	0
Air France	3004002661	14	4	79	6	2	337

The data is not tidy. There are some values embedded in the selected variables (column names). Three fundamental attributes, incidents, fatal incidents, and fatalities are combined with date ranges to form six columns. Date range should instead form its own variable, and the table melted to produce something like this:

airline	avail_seat_km_per_week	year_range	incidents	fatal_accidents	fatalities
Aer Lingus	320906734	1985 - 1999	2	0	0
Aer Lingus	320906734	2000 - 2014	0	0	0
Aeroflot*	1197672318	1985 - 1999	76	14	128
Aeroflot*	1197672318	2000 - 2014	6	1	88
Aerolineas Argentinas	385803648	1985 - 1999	6	0	0
Aerolineas Argentinas	385803648	2000 - 2014	1	0	0
Aeromexico*	596871813	1985 - 1999	3	1	64
Aeromexico*	596871813	2000 - 2014	5	0	0
Air Canada	1865253802	1985 - 1999	2	0	0
Air Canada	1865253802	2000 - 2014	2	0	0
Air France	3004002661	1985 - 1999	14	4	79
Air France	3004002661	2000 - 2014	6	2	337

← [Reply](#) 



Onur Tekiner (<https://iu.instructure.com/courses/2165942/users/6758180>)

Thursday

This is perfect example!

Thanks for sharing this!

← [Reply](#) 



Erik Gonzalez (<https://iu.instructure.com/courses/2165942/users/6352173>)

Tuesday

1. Within the FiveThirtyEight file in public datasets link, I found a tabular dataset in the bob-ross folder entitled "elements by episode"

(<https://github.com/fivethirtyeight/data/blob/master/bob-ross/elements-by-episode.csv>)

↳ (<https://github.com/fivethirtyeight/data/blob/master/bob-ross/elements-by-episode.csv>.)

2. This dataset is not tidy

3. N/A

4. Some of the columns in this variable are tidy - specifically "Episode" and "Title", as each value represents one observation. The dataset is not tidy because third variable "elements" (which represents a Boolean value displaying if a variable appeared in the episode) is displayed in a way where each element forms it's own column. To normalize this dataset, I would split it into two tables. The first table would contain metadata specific to that episode ("Episode" and "Title"), while the second table would contain three columns ("Episode", to serve as a key and denote where an element occurs, and "Element", to keep track of which elements appear in a specific episode, and a Boolean value of 1 for the elements which

appeared). The second table arguably only needs to keep records for elements which appeared, and does not need to maintain records with values of false due to not appearing.

← [Reply](#) 👍



Ao Zhang (<https://iu.instructure.com/courses/2165942/users/6703098>)

Wednesday

I select NBA Draft 2015 dataset from FiveThirtyEight datasets. From my perspective, it is a tidy data since the information in the dataset are all clear. Every row stands for information of one player. Meanwhile, there are several columns which are helpful to predict performances of different players. The columns include basic information like name, year, ID, and draft year. Moreover, some predictable variables like probability of becoming a superstar, probability of becoming a starter, probability of becoming a role-player, and even probability of becoming a bust. All these information could be useful for GMs in different teams to determine which player is the best one for their team.

← [Reply](#) 👍



Onur Tekiner (<https://iu.instructure.com/courses/2165942/users/6758180>)

Thursday

I found this dataset.

<https://www.kaggle.com/datasets/iamsouravbanerjee/animal-information-dataset> ➡
(<https://www.kaggle.com/datasets/iamsouravbanerjee/animal-information-dataset>)


it is about animal information. It is a very tidy dataset.

Many factors make the dataset tidy:

Firstly, there are no null values. Also, Every column's information is clear. Information is given in each column in a very proper way. There is no need for melting techniques, as I see.

Animal	Height (cm)	Weight (kg)	Color	Lifespan (years)	Diet	Habitat	Predators	Average Speed (km/h)	Countries Found	Conservation Status	Family	Gestation Period (days)	Top Speed (km/h)
Aardvark	105-130	40-65	Grey	20-30	Insectivore	Savannas, Grasslands	Lions, Hyenas	40	Africa	Least Concern	Orycteropodidae	210-240	40
Aardwolf	40-50	8-14	Yellow-brown	10-12	Insectivore	Grasslands, Savannas	Lions, Leopards	24-30	Eastern and Southern Africa	Least Concern	Hyaenidae	90	40
African Elephant	270-310	2700-6000	Grey	60-70	Herbivore	Savannah, Forest	Lions, Hyenas	25	Africa	Vulnerable	Elephantidae	640-660	40
African Lion	80-110	120-250	Tan	10-14	Carnivore	Grasslands, Savannas	Hyenas, Crocodiles	58	Africa	Vulnerable	Felidae	98-105	80
African Wild Dog	75-80	18-36	Multicolored	10-12	Carnivore	Savannahs	Lions, Hyenas	56	Sub-Saharan Africa	Endangered	Canidae	70	56

Edited by [Onur Tekiner \(https://iu.instructure.com/courses/2165942/users/6758180\)](https://iu.instructure.com/courses/2165942/users/6758180) on Sep 21 at 12:50am


← [Reply](#) 



<https://iu.instructure.com/courses/2165942/users/6703376>

Thursday

I checked the top 100 Korean drama list among Kaggle's data.

<https://www.kaggle.com/datasets/chanoncharuchinda/top-100-korean-drama-mydramalists>  <https://www.kaggle.com/datasets/chanoncharuchinda/top-100-korean-drama-mydramalists>

First of all, all basic conditions are met. Each variable makes up a column, and each observation makes up a row. Of course, each observation unit constitutes a table. So I checked to see if this data doesn't belong to a case of dirty data.

Are the column names not variable names but values: All variables are listed as names.

Multiple variables in one column: There are no mixed variables, but there is too much text data in the synopsis and tags columns.

Have different units of observation in the same table: There is a lot of duplicate data in each column.


Are the variables contained in both rows and columns: There are no data included on either side.

Is one observation unit divided into multiple tables: This data consists of one table.

From the above, this data was not perfect. In particular, it would be a good idea to delete the synopsis and tag columns, and if someone wants to include them, it would be better to

manage them as a separate table. Also, because there is a lot of duplicate data in each column, it would be a good idea to divide the table by specifying a primary key.

Edited by [Sangzun Park \(https://iu.instructure.com/courses/2165942/users/6703376\)](https://iu.instructure.com/courses/2165942/users/6703376) on Sep 21 at 8:57am

[← Reply](#) 

 [Akash Patil \(https://iu.instructure.com/courses/2165942/users/6699404\)](https://iu.instructure.com/courses/2165942/users/6699404)

Thursday

I have used the Global YouTube Statistics dataset from Kaggle

(<https://www.kaggle.com/datasets/nelgiriyeewithana/global-youtube-statistics-2023>) (<https://www.kaggle.com/datasets/nelgiriyeewithana/global-youtube-statistics-2023>) as my example.

Global YouTube Statistics.csv (200.28 kB)

Detail Compact Column 10 of 28 columns

# rank	# Youtuber	# subscribers	# video views	A category	A Title	# uploads	Country	A Abbreviat...	A channel_t...
1	T-Series	245000000	2.28E+11	Music	T-Series	20082	India	IN	Music
2	YouTube Movies	170000000	0	Film & Animation	youtubemovies	1	United States	US	Games
3	MrBeast	166000000	28368841870	Entertainment	MrBeast	741	United States	US	Entertainment
4	Cocomelon - Nursery Rhymes	162000000	1.64E+11	Education	Cocomelon - Nursery Rhymes	966	United States	US	Education
5	SET India	159000000	1.48E+11	Shows	SET India	116536	India	IN	Entertainment
6	Music	119000000	0	nan	Music	0	nan	nan	Music
7	👶 Kids Diana Show	112000000	93247040539	People & Blogs	👶 Kids Diana Show	1111	United States	US	Entertainment
8	PewDiePie	111000000	29058044447	Gaming	PewDiePie	4716	Japan	JP	Entertainment
9	Like Nastya	106000000	90479060027	People & Blogs	Like Nastya Vlog	493	Russia	RU	People
10	Vlad and Niki	98900000	77180169894	Entertainment	Vlad and Niki	574	United States	US	Entertainment
11	Zee Music Company	96700000	57856289381	Music	Zee Music Company	8548	India	IN	Music
12	WWE	96000000	77428473662	Sports	WWE	70127	United States	US	Sports
13	Gaming	93600000	0	nan	Gaming	0	nan	nan	Games
14	BLACKPINK	89800000	32144597566	People & Blogs	BLACKPINK	543	South Korea	KR	Music

The data seems to be tidy, as each observation forms a row, each variable forms a column and there are no joint columns

Every row is one YouTube channel and statistics are provided for that channel along with its information. Some observations contain NA values that can be dropped.

Columns are Rank of the YouTube channel, name of Youtuber, no. of subscribers, no. of video views, category of the channel, title of the channel, no. of video uploads on the channel, the country of origin of the channel, the abbreviation of the country name and the channel type.

[← Reply](#) 



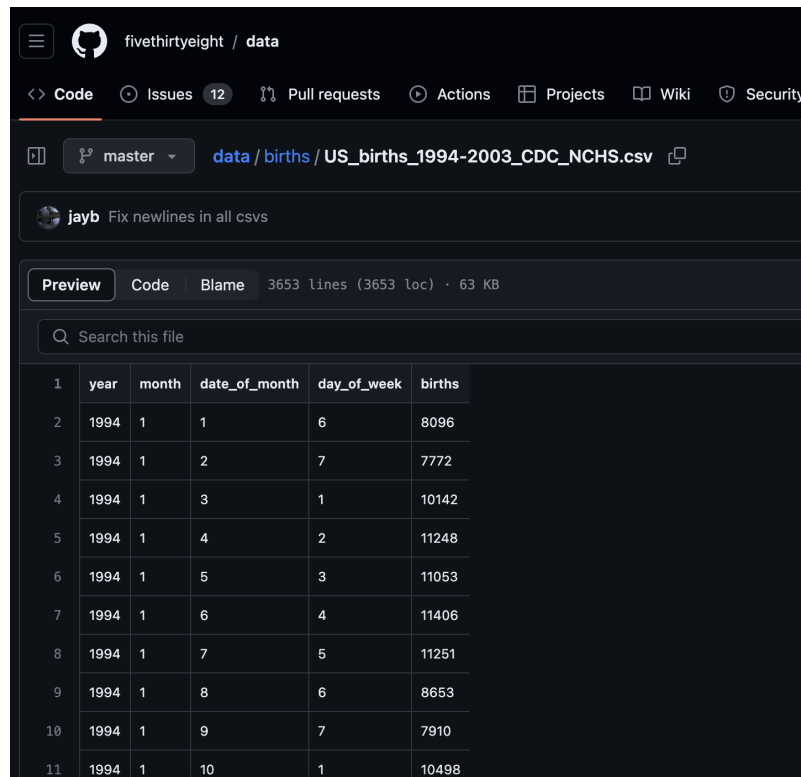
<https://iu.instructure.com/courses/2165942/users/6688148>

Thursday

1. The tabular data set is given below:

https://github.com/fivethirtyeight/data/blob/master/births/US_births_1994-2003_CDC_NCHS.csv

https://github.com/fivethirtyeight/data/blob/master/births/US_births_1994-2003_CDC_NCHS.csv



The screenshot shows the GitHub repository page for 'fivethirtyeight / data'. The file 'US_births_1994-2003_CDC_NCHS.csv' is selected, showing a preview of the first 11 rows. The table has 6 columns: year, month, date_of_month, day_of_week, and births. The data represents daily birth counts in the US from 1994 to 2003.

	year	month	date_of_month	day_of_week	births
1	1994	1	1	6	8096
2	1994	1	2	7	7772
3	1994	1	3	1	10142
4	1994	1	4	2	11248
5	1994	1	5	3	11053
6	1994	1	6	4	11406
7	1994	1	7	5	11251
8	1994	1	8	6	8653
9	1994	1	9	7	7910
10	1994	1	10	1	10498
11	1994	1	10	1	10498

2. This dataset is a tidy dataset. It is taken from FiveThirtyEight and contains U.S. birth data for the years 1994 to 2003, as provided by the Centers for Disease Control and Prevention's National Center for Health Statistics. It is tidy because each variable forms a column and each observation forms a column. There are no two variables in each column, and there is no date variable instead, it is broken down into year, month, day of month, day of week. Every row is one observation that represents the number of births in the US on each date and day in every month for almost 10 years(1994-2003). There are no melting techniques necessary as the rows and columns are well-defined. Each type of observational unit forms a table -here is the number of births in the US on every day of the month over 10 years.

3. The rows are observations of births on every day of every month from 1994-2003 in the US. The columns represent each variable-year, month, Date of month, Day of the week, that defines the birth variable. There are no duplicates in this data and no missing values so this is tidy data.

[Reply](#)



<https://iu.instructure.com/courses/2165942/users/6825193>

Yesterday

Student Grades

StudentID	Subject	Score
1	Economics	85
1	Science	78
2	Economics	92
2	Science	88
3	Economics	76
3	Science	80

This dataset is tidy because:

1. Each Column Represents a Variable: The columns represent different variables.
2. Each Row Represents an Observation: Each row represents a unique observation, which in this case is a student's score in a specific subject.
3. Each table contains data about one type of observation.
4. The data values for each variable are stored in their respective columns.
5. There are no duplicate rows

[← Reply](#)



<https://iu.instructure.com/courses/2165942/users/6443321>

Yesterday

I observed on "Top Streamers on Twitch" for this activity. Looking at the data table, I believe it is tidy. The data contains necessary information on top 1000 streamers from past one year who were streaming on Twitch platform. Each columns represent different aspects on each streamers such as watch time, stream time, peak viewers, average viewers, follower, followers gained, views gained, partnered (boolean), mature (boolean), and language. It has total of 11 columns and each variable forms a column. Also each observation for each streamers forms a row (1000 rows in total). Also, each observational unit can form a table.

[← Reply](#)



<https://iu.instructure.com/courses/2165942/users/6678592>

Yesterday

[congress-terms.csv \(https://iu.instructure.com/users/6678592/files/162620308?wrap=1&verifier=paNCTrB6OLAf2Ta8xjrlMnGvS43SLIGpQlotEnra\)](https://iu.instructure.com/users/6678592/files/162620308?wrap=1&verifier=paNCTrB6OLAf2Ta8xjrlMnGvS43SLIGpQlotEnra) ↓
(https://iu.instructure.com/users/6678592/files/162620308/download?verifier=paNCTrB6OLAf2Ta8xjrlMnGvS43SLIGpQlotEnra&download_frd=1)

The dataset I have used in this case is tidy. It adheres to the principles of tidy data:

1. Each variable forms a column: Each column in the dataset represents a variable, such as "congress", "chamber", "bioguide", "firstname", "middlename", "lastname", "suffix", "birthday", "state", "party", "incumbent", "termstart", and "age".
2. Each observation forms a row: Each row in the dataset represents an observation, i.e., a member of Congress with their associated information.
3. Each type of observational unit forms a table: The table represents one type of observational unit, which is a term of a member of Congress.

Each of the columns contains atomic pieces of data, and there are no nested tables or arrays within cells, which further supports the notion that this is a tidy dataset.

← Reply



<https://iu.instructure.com/courses/2165942/users/6762945>

Yesterday

I found a dataset with NFL play by play information.

(<https://www.dolthub.com/repositories/Liquidata/nfl-play-by-play/doc/master/README.md> ↗ (<https://www.dolthub.com/repositories/Liquidata/nfl-play-by-play/doc/master/README.md>))

There are three tables part of the dataset but I will look at the players one specifically.

This table is tidy in my opinion.

one example of why is that the middlename column handles nulls properly.

Another is that each row is for each player a certain year, an observation. All the data types are correctly assigned, with no exceptions it seems inside each column where there is a different data type.

It was easy to query the data how I'd like due to its tidyness.

← Reply 👍



[Yumeng Liang \(https://iu.instructure.com/courses/2165942/users/6587577\)](https://iu.instructure.com/courses/2165942/users/6587577)

Yesterday

I looked at the awesome public data set. It is a tidy dataset since it follows the tidy data principle. For this dataset, each variable forms a column, each observation forms a row, and each type of observational unit forms a table. Also, no data is duplicated.

← Reply 👍



[Alan Varkey \(https://iu.instructure.com/courses/2165942/users/6681532\)](https://iu.instructure.com/courses/2165942/users/6681532)

Yesterday

I chose this dataset: [fight-songs.csv](#) 📄

[\(https://github.com/fivethirtyeight/data/blob/master/fight-songs/fight-songs.csv\)](https://github.com/fivethirtyeight/data/blob/master/fight-songs/fight-songs.csv)

Each Column Represents a Variable: Each column in the dataset corresponds to a variable. For example, columns like "school," "conference," "song_name," "writers," "year," "student_writer," "official_song," and so on, represent different variables or attributes of the data.

Each Row Represents an Observation: Each row in the dataset represents a unique observation or entry for a particular school and its associated song details. For instance, the first row represents Notre Dame, its conference, song name, writers, year, and so on.

Each Cell Contains a Single Value: Each cell in the dataset contains a single value, and there are no instances where multiple values are combined within a single cell.

The Dataset Has a Clear and Meaningful Structure: The dataset is structured in a tabular format with clear headers for each column. It is organized in a way that makes it easy to understand and work with, and there is no ambiguity in how the data is organized.

Edited by [Alan Varkey \(https://iu.instructure.com/courses/2165942/users/6681532\)](https://iu.instructure.com/courses/2165942/users/6681532) on Sep 22 at 8:49pm

← Reply 👍



Aditya Sanjay Mhaske (<https://iu.instructure.com/courses/2165942/users/6692144>)

Yesterday

Iris Dataset: The Iris dataset consists of 150 samples of iris flowers, each from one of three species: Setosa, Versicolor, or Virginica. For each sample, four features are measured: the sepal length and width, and the petal length and width, all in centimeters.

Tidy Data Criteria:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

This dataset is tidy because it sticks to the criteria:

- Each variable forms a separate column.
- Each observation forms a single row.
- The dataset contains a single type of observational unit
- All the variables are related to these observations.

Why it's Tidy:

- Each row contains the measurements for a single iris flower, with its species specified.
- There's no redundancy or mixing of variables in the same column.

← Reply 👍



Simon Driver (<https://iu.instructure.com/courses/2165942/users/6818242>)

10:56am

I chose the dataset from 538 called "covid-geography". I would say that this data is tidy. Each variable forms a column (eg total % at risk is its own column). Each observation has its own row (each row is a different hospital). Finally, each type of observation unit does form its own table (with the hospital name, and then the associated colvar columns. This dataset matches the definitions and parameters for tidy data, and it looks like it could be operated upon already with R or some other program to start performing statistical analyses or summarizations (or visualizations!).

Link to dataset: <https://github.com/fivethirtyeight/data/tree/master/covid-geography>

← Reply



Renu Jaiswal (<https://iu.instructure.com/courses/2165942/users/6704404>)

11:27am

I am using covid geography data from fivethirtyeight. The data looks like this:

```
covid_df = pd.read_csv('https://raw.githubusercontent.com/fivethirtyeight/data/master/covid-geography/mmsa-icu-beds.csv')
covid_df
```

1 to 25 of 136 entries							
index	MMSA	total_percent_at_risk	high_risk_per_ICU_bed	high_risk_per_hospital	icu_beds	hospitals	total_at_risk
0	San Juan-Carolina-Caguas, PR	52.88	NaN	NaN	NaN	NaN	923725.203
1	Manhattan, KS	47.29	4489.84875	8979.6975	8.0	4.0	35918.79
2	Hilton Head Island-Bluffton-Beaufort, SC	62.72	3904.163571	36438.86	28.0	3.0	109316.58
3	Kahului-Wailuku-Lahaina, HI	59.13	3860.557	19302.785	20.0	4.0	77211.14
4	Spartanburg, SC	66.12	3786.115556	85187.6	45.0	2.0	170375.2
5	Baton Rouge, LA	66.6	3459.7325	39000.62091	124.0	11.0	429006.83
6	Rockingham County-Strafford County, NH, Metropolitan Division	57.72	3365.052	40380.624	60.0	5.0	201903.12
7	Salisbury, MD-DE	68.32	3292.271176	37312.40667	68.0	6.0	223874.44
8	Wichita Falls, TX	67.11	3279.425	19676.55	24.0	4.0	78706.2
9	Colorado Springs, CO	55.96	3251.603053	77225.5725	95.0	4.0	308902.29
10	Cambridge-Newton-Framingham, MA, Metropolitan Division	52.17	3161.025223	62035.12	314.0	16.0	992561.92
11	Albuquerque, NM	60.33	3091.331014	71100.61333	138.0	6.0	426603.68

The columns "MMSA," "total_percent_at_risk," "high_risk_per_ICU_bed," "high_risk_per_hospital," "icu_beds," "hospitals," and "total_at_risk" appear to represent different variables or measurements.

This table seems untidy to me. The main problem I see is it is not clear what the specific observational unit is, also there are some cells containing missing values. We can solve this by defining the observational unit such as a specific region or location. I tried to convert it into a tidy as follows:

```
[48] df = pd.DataFrame(covid_df)
# Melt the DataFrame
covid_tidy_df = pd.melt(covid_df, id_vars=['MMSA'], var_name='variable', value_name='value')
covid_tidy_df
```

1 to 25 of 816 entries			
index	MMSA	variable	value
0	San Juan-Carolina-Caguas, PR	total_percent_at_risk	52.88
1	Manhattan, KS	total_percent_at_risk	47.29
2	Hilton Head Island-Bluffton-Beaufort, SC	total_percent_at_risk	62.72
3	Kahului-Wailuku-Lahaina, HI	total_percent_at_risk	59.13
4	Spartanburg, SC	total_percent_at_risk	66.12
5	Baton Rouge, LA	total_percent_at_risk	66.6
6	Rockingham County-Strafford County, NH, Metropolitan Division	total_percent_at_risk	57.72
7	Salisbury, MD-DE	total_percent_at_risk	68.32
8	Wichita Falls, TX	total_percent_at_risk	67.11
9	Colorado Springs, CO	total percent at risk	55.96

← Reply





<https://iu.instructure.com/courses/2165942/users/6760559>

1:01pm



From Kaggle Data sets considered USA Country wise Covid data set. yes it is a not a tidy data, ON USA Country Wise file, Admin2 is null, lat and Log is Zero hence i am considering that this is not a tidy data. to make it tidy, i would prefer to filter out the Admin2 is null, lat and Log Zero data from the USA country wise and perform the required manipulations. Please find the attached file for reference.

[usa county wise.csv \(https://iu.instructure.com/files/162643595/download?download_frd=1&verifier=uNP3HBelP444ck14LimB2PBVA1aDEENGaSaoE7MR\)](https://iu.instructure.com/files/162643595/download?download_frd=1&verifier=uNP3HBelP444ck14LimB2PBVA1aDEENGaSaoE7MR)

[Reply](#)



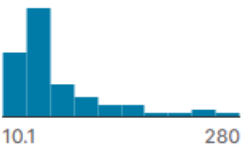
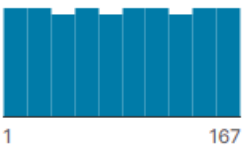
<https://iu.instructure.com/courses/2165942/users/6813278>

1:48pm



this dataset is :

based on year of 2017

▲ Countries	# Housing and utilit...	# Global rank	▲ Available data
The name of the country for which housing and utilities price data is provided	The combined cost of housing and utilities for the year 2017 in the local currency of each country	The global ranking of each country based on its housing and utilities prices in 2017, with lower numbers indicating higher	Indicates whether data for housing and utilities prices in the specified country is available or not.
167 unique values			1 unique value
Bermuda	279.748	1	2017 - 2017
Switzerland	266.926	2	2017 - 2017
Australia	257.057	3	2017 - 2017
New Zealand	236.631	4	2017 - 2017
UK	234.09	5	2017 - 2017
Israel	232.165	6	2017 - 2017
Luxembourg	230.278	7	2017 - 2017
Denmark	228.068	8	2017 - 2017
Ireland	222.32	9	2017 - 2017

1.) This data set is tidy since it has 4 different columns and each row has a unique value to it except available data column which shows the year.

2.) The columns for this dataset are Countries , Housing and Utilities, Global rank and Available data.

3-4) . The data is tidy because each column has a different name and no values are there in the column header

. There is no duplication of values except the available data which is used to show the year

. each row has a unique value

. The datatype for each column is unique and it remains constant for the entire dataset

Edited by [Shreedeeep Sadasivan Nair \(https://iu.instructure.com/courses/2165942/users/6813278\)](https://iu.instructure.com/courses/2165942/users/6813278) on Sep 23 at 2pm

← [Reply](#) 



<https://iu.instructure.com/courses/2165942/users/6684840>

[Jash Shah \(https://iu.instructure.com/courses/2165942/users/6684840\)](https://iu.instructure.com/courses/2165942/users/6684840)

3:32pm

⋮

I have selected the Boston House Pricing Dataset. In a tidy dataset, each row represents an observation, and each column represents a variable.

Here's an explanation of the columns:



1. CRIM: per capita crime rate by town
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX: nitric oxides concentration (parts per 10 million)
6. RM: average number of rooms per dwelling
7. AGE: proportion of owner-occupied units built prior to 1940
8. DIS: weighted distances to five Boston employment centers
9. RAD: index of accessibility to radial highways
10. TAX: full-value property-tax rate per \$10,000
11. PTRATIO: pupil-teacher ratio by town
12. B: $1000(B_k - 0.63)^2$ where B_k is the proportion of black people by town
13. LSTAT: % lower status of the population

This dataset is tidy because each row represents information about a specific town, and each column is a different attribute or characteristic of that town. The dataset follows the principles of tidy data:

1. Each variable (column) is in a separate column.
2. Each observation (row) represents a unique town.
3. There are no duplicate rows.

However, one aspect could be made clearer. The last two columns, "B" and "LSTAT," are not entirely self-explanatory. It would be helpful to have more detailed column names or include a data dictionary to explain these columns better. This would make the dataset even more user-friendly and interpretable.

Edited by [Jash Shah \(https://iu.instructure.com/courses/2165942/users/6684840\)](https://iu.instructure.com/courses/2165942/users/6684840) on Sep 23 at 3:32pm

 [Reply](#) 



[Robert Perez \(he/him/his\) \(https://iu.instructure.com/courses/2165942/users/6701521\)](https://iu.instructure.com/courses/2165942/users/6701521)

3:33pm

I chose the airline-safety dataset from FiveThirtyEight:

<https://github.com/fivethirtyeight/data/tree/master/airline-safety> 
(<https://github.com/fivethirtyeight/data/tree/master/airline-safety>)

This dataset is indeed tidy, perhaps the textbook definition of a tidy dataset. Each row represents a unique airline. There are no duplicates, and each row is easy to identify quickly. The columns are all unique variables that relate to the labeled observation. There are no nested columns (such as those that use JSON). Each variable is a number that represents a count of either miles or incidents. Plus, not that this makes the dataset tidy, but there aren't an overwhelming number of columns, making the dataset very easy to understand and ingest.

← [Reply](#) 