

Data Pre-Processing

William Steimel

Table of Contents

- ▶ Introduction to Data Pre-processing
 - ▶ Data Cleaning
 - ▶ Data Integration
 - ▶ Data Transformation
 - ▶ Data Reduction
 - ▶ Tidy Data

Sources

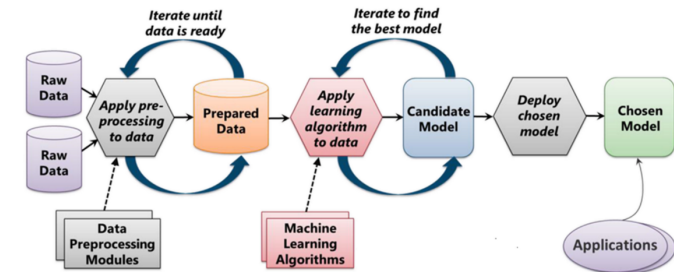
- ▶ Data Pre-processing
 - ▶ Secondary Analysis of Electronic Health Records Chapter 12,13,14
 - ▶ Medical Data seems to be very dirty which makes it a good place to learn about Data Pre-processing.
- ▶ Tidy Data - Hadley Wickham
 - ▶ Journal of Statistical Software August 2014, Volume 59, Issue 10.

Introduction

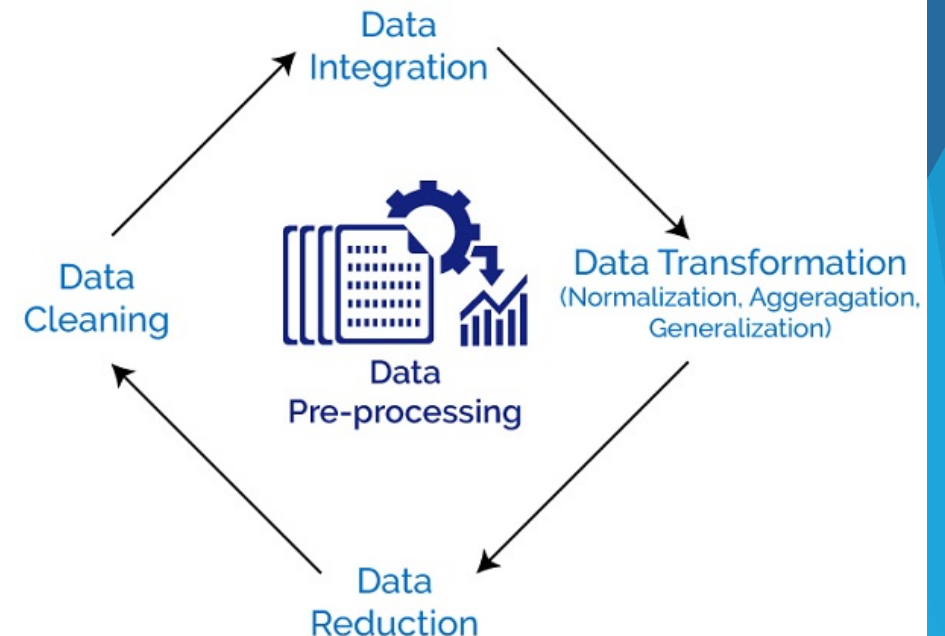
Data Preprocessing

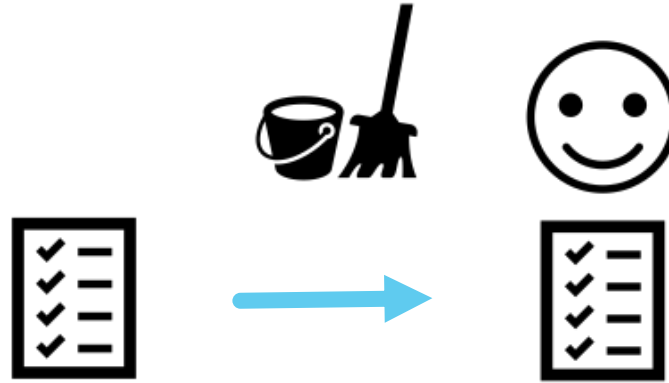
- ▶ Data Pre-processing consists of a series of steps to transform raw data to "clean" and "tidy" data prior to statistical analysis and Machine Learning modeling.
- ▶ *There are several steps in Data Pre-Processing including*
 - ▶ "Data Cleaning"
 - ▶ "Data Integration"
 - ▶ "Data Transformation"
 - ▶ "Data Reduction"
- ▶ *This presentation will discuss these steps along with methods for handling data.*

The Machine Learning Process



From "Introduction to Microsoft Azure" by David Chappell





Data Cleaning

What is Data Cleaning/Cleansing?

- ▶ Real world data is usually very messy and can be incomplete (missing data), noisy (random error or outlier values that deviate from expectation), and can be inconsistent. (A name written in the phone number column)
- ▶ Data Cleaning - Process of dealing with missing data, noise, outliers, and duplicate incorrect records while minimizing introduction of bias into the data.
- ▶ Many reasons exist for messy data including user error and database related issues but it is essential to treat these data values before performing statistical analysis and modeling as a practitioner.

” **80 percent** of a data scientist's valuable time is spent simply finding, cleansing, and organizing data, leaving only 20 percent to actually perform analysis...

IBM Data Analytics

Questions for all of you

- ▶ When you download a dataset from the internet from a place like Kaggle, UCI Machine Learning Library, or other sources is the data clean?
 - ▶ Usually datasets from online sources are clean but this is not always the case.
 - ▶ From my working experience I was often given dirty datasets that were an extreme headache.

Garbage in - Garbage out

- ▶ Garbage in, garbage out is a popular saying in English that refers to incorrect or poor input will lead to faulty output.
 - ▶ If you perform computational experiments with dirty data it will likely lead to a faulty experiment
 - ▶ If your data isn't clean or preprocessed it will negatively impact your data analysis or machine learning model.
 - ▶ We will all perform experiments soon for our thesis and high quality output will be impacted by data quality.

“Garbage in, garbage out”



Your analysis is as good as your data.

Missing Data

- ▶ There are three general steps that should be followed for handling missing data.
 - ▶ Identify Patterns and reasons for missing data (Why is the data missing?)
 - ▶ Analyze the proportion of missing data (How much is missing per feature and observation?)
 - ▶ Choose the best imputation method (Choose the best method based on what you know)

Missing Data


- ▶ Step 1- Determine why the data is missing!
 - ▶ Types of Missingness
 - ▶ Missing Completely at Random (MCAR) - No relationship between missingness of data and any values
 - ▶ Missing at Random (MAR) (Conditional)- systematic relationship between rate of missing values and the observed data, but not the missing data.
 - ▶ If men are more likely to tell you their weight than women, weight is MAR.
 - ▶ Missing Not at Random (MNAR) - clear relationship between missingness and its values
 - ▶ On a survey people with low IQ have missing observations.
 - ▶ Not Ignorable

Missing Data

► Step 2 - Analyze Proportion of Missing Data

- The proportion of missing data should be calculated at the observation level and feature level.
- Observations and features with high ratios of missing values should be candidates for removal.

Table 13.1
Examples of missing data in EHR



	Gender	Glucose	AST	Age
Patient 1	?	120	?	?
Patient 2	M	105	?	68
Patient 3	F	203	45	63
Patient 4	M	145	?	42
Patient 5	M	89	?	80

Missing Data

▶ Step 3- Choose the Best Imputation Method

- ▶ The best method depends on situation and contributes to higher simplicity and little bias in the dataset.
 - ▶ Deletion Methods - The simplest way is to discard or delete the missing observations.
 - ▶ Single-Value Imputation - Missing Values are filled by some type of “predicted” values.
 - ▶ Mean/Median(Numeric) - Mode (Categorical)
 - ▶ Linear Interpolation (Time Series)
 - ▶ Hot Deck and Cold Deck - missing attribute replaced with a value from an estimated distribution of current data.
 - ▶ Last Observation Carried Forward - imputes the missing value with the last available observation of the class.
 - ▶ Model-Based Imputation - Predictive Model is created to estimate values to substitute missing data.
 - ▶ Linear Regression
 - ▶ Stochastic Regression
 - ▶ Multiple-Value Imputation
 - ▶ K-Nearest Neighbors

Missing Data

- ▶ There is no one size fits all approach to handling missing data and it is all situational.
- ▶ Use your best judgement and domain knowledge when making judgements regarding missing data.

Noisy Data/Outliers

- ▶ An outlier is a data point that differs from the remaining data.
 - ▶ Abnormalities, discordants, deviants, anomalies
- ▶ 1,3,2,4,7,3,2,98, 1
 - ▶ Can you spot the outlier?
- ▶ Negative Effect of Outliers:
 - ▶ Increase in error variance and reduction in statistical power
 - ▶ Decrease in normality for the cases where outliers are non-randomly distributed
 - ▶ Model bias by corrupting the true relationship between variables and outcome
- ▶ As you can see on the right- an outlier can have significant effect on the mean and standard deviation of the data.

	1,3,2,4,7,3,2,98, 1	
	Without Outlier	With Outlier
Mean	2.88	13.44
Median	2.50	3.00
Mode	1.00	1.00
Standard Deviation	1.83	29.94

Noisy Data/Outliers

- ▶ The simplest form of outlier detection looks at extreme value analysis of unidimensional data.
 - ▶ Boxplot and Histogram are useful for this.
- ▶ Discovering outliers relies on determining the statistical tails of the underlying distribution and assuming values either too large or too small are outliers.
- ▶ Statistical Methods - Data is assumed to follow a distribution model and a data point is considered an outlier if it deviates significantly from the model.
 - ▶ Tukey's Method
 - ▶ Z-Score
 - ▶ Modified Z-Score
 - ▶ Interquartile Range with Log-Normal Distribution
 - ▶ Ordinary and Studentized Residuals
 - ▶ Cook's Distance
 - ▶ Mahalanobis Distance
- ▶ Proximity Based Models - Using clustering to find outliers
 - ▶ K-Means
 - ▶ K-Medoids

Noisy Data and Outliers

▶ Handling Outliers

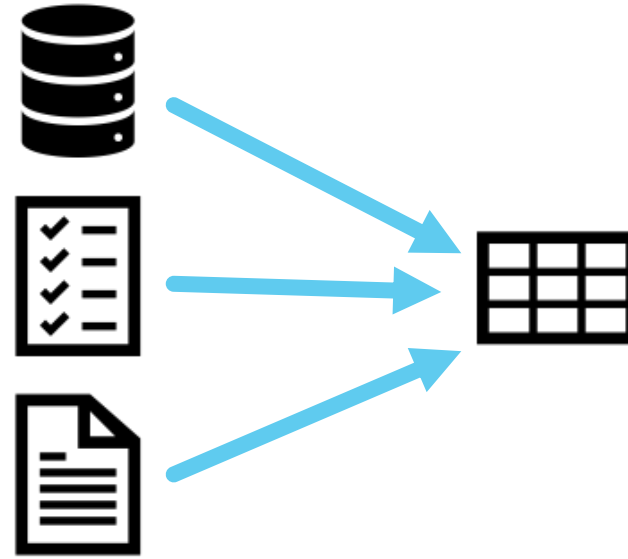
- ▶ Deleting Observations - removing entries due to error or at the tail of distribution
- ▶ Transforming and Binning Values - Smooth a sorted data value by considering the 'neighborhood' or values around it
- ▶ Imputing - Like with empty values we can impute mean, median
- ▶ Treat Separately - Treat separately or give it a new feature (Feature Engineering)

Noisy Data and Outliers

- ▶ In certain cases Outliers may represent valuable information that must not be thrown away.
- ▶ An alternative strategy is to also use models that are robust to outliers.

Inconsistent Data

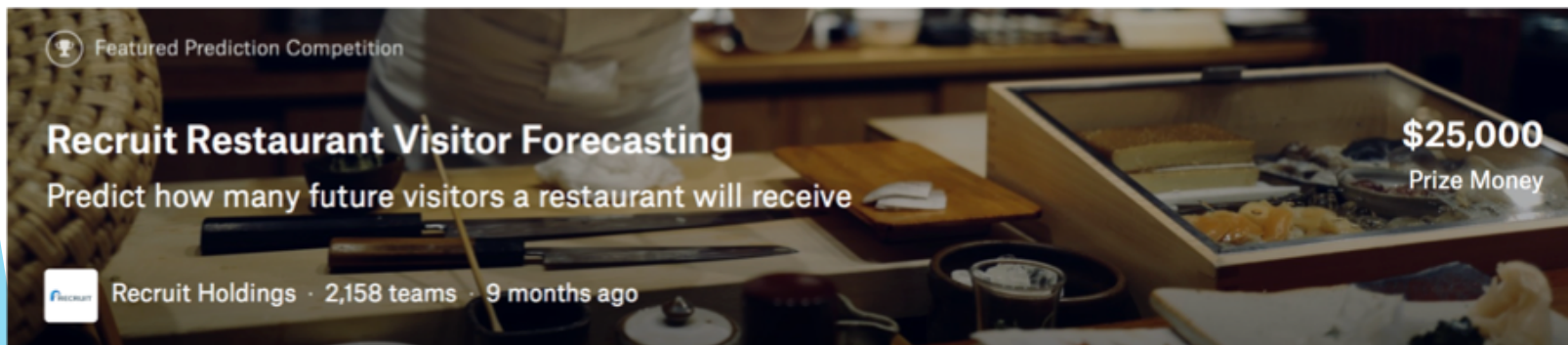
- ▶ There are sometimes inconsistencies and duplications in the data.
- ▶ These may be corrected using external references and deletion.
- ▶ Inconsistencies - Things that do not make sense from a data perspective based on domain knowledge.
 - ▶ Phone Number field contains a name (This could be a type)
 - ▶ Inconsistent formatting conventions for phone numbers (973-243-4244 vs 9732434244)
 - ▶ A Money column with Dollars for one observation and Yen for another observation
 - ▶ Typos/inconsistent capitalization/mislabeled classes
- ▶ Duplication - Repeated record in the dataset
- ▶ One must always examine the data and ask whether there is anything that doesn't seem right about the data based on previous experience and expertise.



Data Integration

Data Integration

- ▶ Data Integration is the process of combining data from various data sources into a consistent dataset.
- ▶ Data can come from various sources including csv, txt, xls among other files and are usually joined by some sort of key or ID's linking them.
- ▶ The below example from the Recruit Restaurant Visitor Forecasting Competition contains around 7 datasets that need to be integrated before modeling can occur.



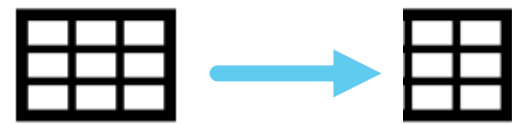
Data Sources

air_reserve.csv	92.4k x 4
air_store_info.csv	829 x 5
air_visit_data.csv	252k x 3
date_info.csv	517 x 3
hpg_reserve.csv	2.00m x 4
hpg_store_info.csv	4691 x 5
sample_submission...	32.0k x 2
store_id_relation.csv	150 x 2

Data Transformation

Data Transformation

- ▶ The goal of Data Transformation is to transform the data values into a format more suitable for statistical analysis.
- ▶ What is called Feature Engineering in Machine Learning is often done in this stage.
 - ▶ I have already discussed feature engineering and will not go into too much detail today.
- ▶ Normalization
- ▶ Aggregation
- ▶ Generalization



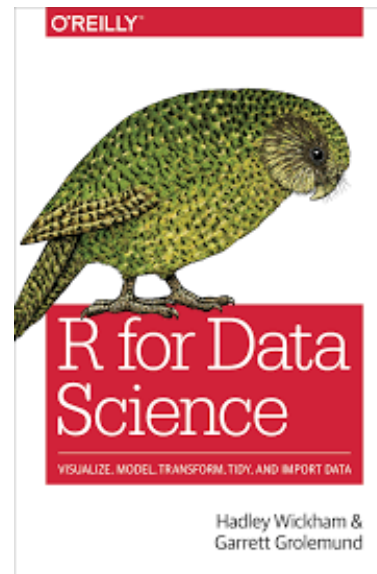
Data Reduction

Data Reduction

- ▶ Data Reduction is the process of reducing the input data without compromising the the integrity of the original data which will contribute to a more effective analysis.
- ▶ One example of applying this is the principles of Tidy Data.

About the Author

- ▶ Developer of various R data analysis packages called “Tidyverse”
- ▶ Author of R for Data Science Book
- ▶ Huge Contributor to Data Science and R Ecosystem.



Introduction

- ▶ “It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data” (Dasu and Johnson 2003)
- ▶ Data preparation is often done many times over the course of data analysis
- ▶ Despite this not much research has been done on how to clean data well.
- ▶ Data Cleaning is extremely broad which poses some challenges but this paper focuses on ”data tidying” or structuring datasets to facilitate analysis.
- ▶ This paper provides a framework and comprehensive philosophy of data called tidy data

Defining Tidy Data

- ▶ Tidy datasets provide a standardized way to link the structure of a dataset (physical layout) with its semantics (meaning).
- ▶ A dataset is messy or tidy based on how rows, columns, and tables are matched with observations, variables, and types
- ▶ Tidy Data -
 - ▶ Each variable forms a column
 - ▶ Each observation forms a row
 - ▶ Each type of observational unit forms a table
- ▶ Messy Data - Any other arrangement of the data

Illustrated with Mercari Dataset

- ▶ The below dataset comes from a Kaggle Competition titled Mercari- Price Suggestion Challenge which details prices of items sold on the website.
- ▶ Definitions
 - ▶ Variable: A measurement of an attribute. (item_condition_id, category_name, brand_name)
 - ▶ Value: The actual measurement or attribute (Value recorded in each column)
 - ▶ Observation: All values measure on the same unit (Each mercari sale)

train_id		name	item_condition_id	category_name	brand_name	price	shipping	item_description
0	0	MLB Cincinnati Reds T Shirt Size XL	3	Men/Tops/T-shirts	NaN	10.0	1	No description yet
1	1	Razer BlackWidow Chroma Keyboard	3	Electronics/Computers & Tablets/Components & P...	Razer	52.0	0	This keyboard is in great condition and works ...
2	2	AVA-VIV Blouse	1	Women/Tops & Blouses/Blouse	Target	10.0	1	Adorable top with a hint of lace and a key hol...
3	3	Leather Horse Statues	1	Home/Home Décor/Home Décor Accents	NaN	35.0	1	New with tags. Leather horses. Retail for [rm]...
4	4	24K GOLD plated rose	1	Women/Jewelry/Necklaces	NaN	44.0	0	Complete with certificate of authenticity

Defining Tidy Data

► Non-Tidy Data vs Tidy Data

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Table 1: Typical presentation dataset.

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Table 3: The same data as in Table 1 but with variables in columns and observations in rows.

Tidying Messy Datasets

- ▶ Real Datasets often violate the principles of Tidy Data
 - ▶ Occasionally you will get a dataset that you can start analyzing immediately but in the real world this is not very common
- ▶ Five Most Common Problems with messy datasets
 - ▶ Column headers are values, not variable names
 - ▶ Multiple variables are stored in one column
 - ▶ Variables are stored in both rows and columns
 - ▶ Multiple Types of observational units are stored in the same table
 - ▶ A single observational unit is stored in multiple tables
- ▶ This section reviews these common problems with real datasets.

Column headers are values, not variable names

- ▶ A common messy dataset is tabular data designed for presentations where variables form both rows and columns and column headers are values, not variable names
- ▶ The below dataset explores the relationship between income and religion in the US

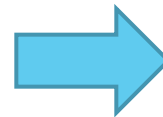
religion	<\$10k	\$10-20k	\$20-30k	\$30-40k	\$40-50k	\$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, \$75-100k, \$100-150k and >150k, have been omitted.

Column headers are values, not variable names

- ▶ There are three variables - religion, income, frequency
- ▶ To solve this we need to melt or stack the data - turn the columns into rows
- ▶ This is considered tidy because each column represents a variable and each row represents an observation

religion	<\$10k	\$10-20k	\$20-30k	\$30-40k	\$40-50k	\$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95



religion	income	freq
Agnostic	<\$10k	27
Agnostic	\$10-20k	34
Agnostic	\$20-30k	60
Agnostic	\$30-40k	81
Agnostic	\$40-50k	76
Agnostic	\$50-75k	137
Agnostic	\$75-100k	122
Agnostic	\$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, \$75-100k, \$100-150k and >150k, have been omitted.

Table 6: The first ten rows of the tidied Pew survey dataset on income and religion. The column has been renamed to `income`, and `value` to `freq`.

Multiple Variables Stored in One Column

- ▶ Another major problem with untidy data occurs when multiple variables are stored in one column
- ▶ The below dataset is from the World Health Organization and details counts of tuberculosis cases by country, year, and demographic group.
- ▶ The original dataset stores demographic data in the format gender(m,f) and age (0-14, 15-25, 25-34, 35-44, 45-54, 55-64, unknown)
 - ▶ Data in this format is often separated by a character (.,-,_,:.) and string processing will be needed

country	year	column	cases	country	year	sex	age	cases
AD	2000	m014	0	AD	2000	m	0-14	0
AD	2000	m1524	0	AD	2000	m	15-24	0
AD	2000	m2534	1	AD	2000	m	25-34	1
AD	2000	m3544	0	AD	2000	m	35-44	0
AD	2000	m4554	0	AD	2000	m	45-54	0
AD	2000	m5564	0	AD	2000	m	55-64	0
AD	2000	m65	0	AD	2000	m	65+	0
AE	2000	m014	2	AE	2000	m	0-14	2
AE	2000	m1524	4	AE	2000	m	15-24	4
AE	2000	m2534	4	AE	2000	m	25-34	4
AE	2000	m3544	6	AE	2000	m	35-44	6
AE	2000	m4554	5	AE	2000	m	45-54	5
AE	2000	m5564	12	AE	2000	m	55-64	12
AE	2000	m65	10	AE	2000	m	65+	10
AE	2000	f014	3	AE	2000	f	0-14	3

(a) Molten data

(b) Tidy data

Table 10: Tidying the TB dataset requires first melting, and then splitting the `column` column into two variables: `sex` and `age`.

Variables Are Stored in Both Rows and Columns

- ▶ The below dataset shows daily weather data from one weather station (MX17004) in Mexico for five months in 2010.
 - ▶ The problem is that its original format provides variables stored in both rows and columns as can be seen below
 - ▶ Column variable - d1,d2,d3,d4,d5,d6,d7,d8 (day in the month)
 - ▶ Row Variable- Tmax/tmin (Max and minimum temperature)

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—
MX17004	2010	4	tmax	—	—	—	—	—	—	—	—
MX17004	2010	4	tmin	—	—	—	—	—	—	—	—
MX17004	2010	5	tmax	—	—	—	—	—	—	—	—
MX17004	2010	5	tmin	—	—	—	—	—	—	—	—

Table 11: Original weather dataset. There is a column for each possible day in the month. Columns d9 to d31 have been omitted to conserve space.

Variables Are Stored in Both Rows and Columns

- ▶ To fix this issue we first melt with the colvars id, year, month and the column that contains the variable names, element.
- ▶ The tmin and tmax variables are variables stored in columns but can be fixed with the cast or unstack operation

id	date	element	value	id	date	tmax	tmin
MX17004	2010-01-30	tmax	27.8	MX17004	2010-01-30	27.8	14.5
MX17004	2010-01-30	tmin	14.5	MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-02	tmax	27.3	MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-02	tmin	14.4	MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-03	tmax	24.1	MX17004	2010-02-23	29.9	10.7
MX17004	2010-02-03	tmin	14.4	MX17004	2010-03-05	32.1	14.2
MX17004	2010-02-11	tmax	29.7	MX17004	2010-03-10	34.5	16.8
MX17004	2010-02-11	tmin	13.4	MX17004	2010-03-16	31.1	17.6
MX17004	2010-02-23	tmax	29.9	MX17004	2010-04-27	36.3	16.7
MX17004	2010-02-23	tmin	10.7	MX17004	2010-05-27	33.2	18.2

(a) Molten data

(b) Tidy data

Table 12: (a) Molten weather dataset. This is almost tidy, but instead of values, the `element` column contains names of variables. Missing values are dropped to conserve space. (b) Tidy weather dataset. Each row represents the meteorological measurements for a single day. There are two measured variables, minimum (`tmin`) and maximum (`tmax`) temperature; all other variables are fixed.

Multiple Types in one Table

- ▶ Datasets often involve values collected at multiple levels, on different types of observational units.
 - ▶ The below example uses the Billboard Dataset which contains different observational units including week and rank for each song
- ▶ During tidying each type of observational unit should be stored in its own table.
 - ▶ Database normalization
- ▶ However, Data analysis usually requires merging the datasets back into one table

year	artist	time	track	date	week	rank
2000	2 Pac	4:22	Baby Don't Cry	2000-02-26	1	87
2000	2 Pac	4:22	Baby Don't Cry	2000-03-04	2	82
2000	2 Pac	4:22	Baby Don't Cry	2000-03-11	3	72
2000	2 Pac	4:22	Baby Don't Cry	2000-03-18	4	77
2000	2 Pac	4:22	Baby Don't Cry	2000-03-25	5	87
2000	2 Pac	4:22	Baby Don't Cry	2000-04-01	6	94
2000	2 Pac	4:22	Baby Don't Cry	2000-04-08	7	99
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-02	1	91
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-09	2	87
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-16	3	92
2000	3 Doors Down	3:53	Kryptonite	2000-04-08	1	81
2000	3 Doors Down	3:53	Kryptonite	2000-04-15	2	70
2000	3 Doors Down	3:53	Kryptonite	2000-04-22	3	68
2000	3 Doors Down	3:53	Kryptonite	2000-04-29	4	67
2000	3 Doors Down	3:53	Kryptonite	2000-05-06	5	66



id	artist	track	time	id	date	rank
1	2 Pac	Baby Don't Cry	4:22	1	2000-02-26	87
2	2Ge+her	The Hardest Part Of ...	3:15	1	2000-03-04	82
3	3 Doors Down	Kryptonite	3:53	1	2000-03-11	72
4	3 Doors Down	Loser	4:24	1	2000-03-18	77
5	504 Boyz	Wobble Wobble	3:35	1	2000-03-25	87
6	98°0	Give Me Just One Nig...	3:24	1	2000-04-01	94
7	A*Teens	Dancing Queen	3:44	1	2000-04-08	99
8	Aaliyah	I Don't Wanna	4:15	2	2000-09-02	91
9	Aaliyah	Try Again	4:03	2	2000-09-09	87
10	Adams, Yolanda	Open My Heart	5:30	2	2000-09-16	92
11	Adkins, Trace	More	3:05	3	2000-04-08	81
12	Aguilera, Christina	Come On Over Baby	3:38	3	2000-04-15	70
13	Aguilera, Christina	I Turn To You	4:00	3	2000-04-22	68
14	Aguilera, Christina	What A Girl Wants	3:18	3	2000-04-29	67
15	Alice DeeJay	Better Off Alone	6:50	3	2000-05-06	66

Table 13: Normalized Billboard dataset split up into song dataset (left) and rank dataset (right). First 15 rows of each dataset shown; **genre** omitted from song dataset, **week** omitted from rank dataset.

One type in multiple tables

- ▶ It is also common to find data values about a single type of observational unit spread out over multiple tables or files
- ▶ The tables and files are often split by another variable so each represents a single year, person or location.
- ▶ Ways to solve this:
 - ▶ Read the files into a list of tables
 - ▶ For each table, add a new column that records the original file name (because the file name is often the value of an important variable).
 - ▶ Combine all of the tables into a single table

Other Sections

- ▶ The other sections discuss R tools which I have omitted from this presentation.

Discussion

- ▶ Data Cleaning is an important problem in data analysis but is not often researched.
- ▶ This paper creates a new framework for a small subset of data cleaning called tidying data-structuring datasets to facilitate manipulation, visualization, and modeling.
- ▶ Apart from tidying there are many other tasks involved in cleaning data:
 - ▶ Parsing dates and numbers
 - ▶ Identifying missing values
 - ▶ Correcting character encodings
 - ▶ Matching similar but not identical values (due to typos)
 - ▶ Verifying experimental design
 - ▶ Filling in structural missing values
 - ▶ Model-based data cleaning that identifies suspicious values
- ▶ Hadley Wickham - Can we develop other frameworks to make these tasks easier?

Conclusion

- ▶ Data Pre-processing and specifically Data Cleaning is extremely important for successful and accurate analysis and modeling.
- ▶ Data Pre-processing is often considered one of the tasks that take up the most of a data scientists time.
- ▶ The study of these techniques will be useful as I continue to encounter more challenging and dirty datasets in my daily studies.