CHAPTER 7 GETTING YOUR DATA IN SHAPE

In the previous chapter, we learned how to use integrity checks to uncover flaws in our data. In this chapter we will learn some tricks that we can use to scrub some of the most common data dirt. We will also learn some techniques for transforming our data, so they are better prepared for analysis and visualization.

We will use Excel to perform most of our data cleaning and transforming tasks, but we will also use OpenRefine and a PDF conversion service.

Excel and other spreadsheet programs have built-in features and functions that can help whip data into shape. For example, we can use Excel's Text to Columns tool to **parse**—or split—data that are stored in one column into multiple columns. This is useful when we have a column for full names that are stored like, "Doe, Jane". Using **Text to Columns**, we could carve this into one column for the last name, and another column for the first name. Excel also has a Remove Duplicates tool that allows us to remove rows that hold data that are repeated. In addition, we can write formulas in our Excel spreadsheets to carve out data and then assemble them. This is handy when we have dates stored as text, such as "20150101" for January 1, 2015.

OpenRefine (which runs on Windows, Mac and Linux computers) is an open-source program from Google that lets users quickly run data integrity checks, then clean flawed data. Its creators call it a power tool for working with messy data (http://openrefine.org).

Cometdocs and Zamzar are two online converters that allow us to extract Excel files from PDF tables.

Then there are even more-powerful and more-complex tools for data cleaning that are outside the scope of this book. More-advanced data users will use database managers, such as Microsoft Access or MySQL, to clean data using string functions. String functions are code that can be inserted into the **Structured Query Language** (SQL) that these database managers execute. For instance, someone who's adept with SQL can write queries to rearrange dates into a proper format and put them into a new column. Data users who are even more advanced write scripts using programming languages like **Python**, **Ruby**, PHP or **Perl** to clean and transform data sets. These programming languages allow users to process huge amounts of data more quickly than they could using traditional programs like database managers and spreadsheets. In addition, a number of companies offer data cleaning services and software.

For a deeper look at data cleaning, check out the *Bad Data Handbook* (McCallum, 2012) or *Best Practices in Data Cleaning* (Osborne, 2013), which is geared to working with primary data collected for the purposes of research.

96 DATA LITERACY

There's plenty that we can do, though, with Excel, PDF converters and OpenRefine, so we'll start to tackle some of the most common challenges now.

COLUMN CARVING

Government agencies often stuff too much data into one column, which makes it difficult for us to work with them. In Chapter 5, we took a look at political campaign contribution data from the Missouri Ethics Commission that stores all the identifying data about contributors into one column. This is an extreme example of dirty data. Someone with advanced data-cleaning skills might be able to parse these data elements into separate fields, but that would take a lot of work. As shown in that chapter, the commission now provides these data with the data broken down into multiple columns that users can work with better.

Most column parsing challenges are far simpler, such as the one in our spreadsheet of dangerous dogs that have been reported to animal control authorities in the city of Austin, Texas. The city's Animal Services Office allows people to fill out forms asserting that a dog is dangerous or has bitten another animal (AustinTexas.gov, n.d.). The original file has been modified for this exercise. The file is called Austin_Declared_Dangerous_Dogs.xlsx and has 40 rows, including one for the column headers. (You should have your data notebook open so you can record any activities.)

We see four columns, one each for the street address, zip code, description of the dog and the dog's location, which includes latitude and longitude points that someone



Source: Retrieved from Data.austintexas.gov.

Note: Austin declared dangerous dog data. Note the Description of Dog column has four pieces of data, separated by a comma.

could use to create a webmap. The Description of Dog column has a lot of different bits of data mashed together: the dog's name, sex, color and breed. Each piece of data is separated by a comma, a pattern that is important to note before using the Text to Columns tool.

Let's get ready by saving a copy of the spreadsheet with a name other than the original, so we preserve the original copy. We will work with the new file.

If we use Text to Columns, it will overwrite our original data in Column C. We want to preserve that for reference, so insert four columns to the right of Column C. We'll need four columns because we're going to parse out four data items. Right-click on Column D and pick Insert from the popup menu four times. You should have blank columns from D through G.

Practice safe computing by copying the contents of Column C into Column D: Rightclick on C and pick Copy from the popup menu, then right-click on D and pick Paste. Your spreadsheet should look like this with Column D highlighted:



Source: Retrieved from Data.austintexas.gov. Note: Description of Dog column, ready for parsing

Now we're ready to tell Excel to perform its magic. Click on the Data tab and then the Text to Columns button. Excel launches a Wizard that walks us through carving up the data in three steps. In the first step, tell Excel the correct data type, which is delimited.

The Text Wizard has determined that your data is Delimited.
If this is correct, choose Next, or choose the data type that best describes your data.
Original data type
Choose the file type that best describes your data:
Delimited - Characters such as commas or tabs separate each field.
Fixed width - Fields are aligned in columns with spaces between each field.
Preview of selected data:
1 Description of Dog 2 Rex, neutered male, tan and white, Pit Bull mix 3 Scooby, neutered male, red, Labrador Retriever mix 4 Salty, male, brown and white, Boxer 5 Junebug, spayed female, blue tick, Australian Cattle dog 4

Note: Step 1 of the Text to Columns Wizard for delimited data.

Click Next and it's time to set more options. We need to tell Excel that the delimiter is a comma, so make sure only that box is checked in the Delimiters section. We have no text qualifiers (single or double quotation marks to denote text), so change that to None. In the preview area, Excel draws lines where the column breaks will go, based on the position of the comma delimiters.

	This screen lets you set the in the preview below. Delimiters	delimiters your dat	a contains. You can s	ee how your text is affected
not	Tab Semicolon © Comma Space Other Data greview	reat consecutive du gualifien (nonce)	elimiters as one	
$\sqrt{2}$	Description of Dog Rex Scooby	neutered male	tan and white red	Pit Bull mix

Source: Microsoft Excel for Windows 2013.

Note: Step 2 of the Text to Columns Wizard for delimited data.

Click Next and we're in the last step of our wizard, where we set the data types for the four columns that we are creating. Change all of these from General to Text. Do that by selecting all of the headers, which say General, and changing the Column data format to Text. Click Finish and Excel carves up the data into four columns. If you get a warning

about overwriting existing data, click OK. That's just Excel's way of telling you that the data in Column D will be modified. (This is why we saved the original in Column C.)

<pre>International format General General</pre>		This screen lets you	select each column	and set the Data Format		
Concertain General: converts numeric values to numbers, date values to date values to date values to date values. Data greview General: converts numeric values to numbers, date values to date values. Destination: SDS1 Data greview Bescription of Doo motimport column (skip) Destination: SDS1 Data greview Bescription of Doo motimport column (skip) Bescription of Doo motimport column (skip) Data greview Bescription of Doo motimport columns wizard for delimited data: Data greview Prove Resett Excel for Windows 2013. Bescription of Doo motimport columns wizard for delimited data: Prove Resett Reset Latory Column Resett Reset Docolumns wizard for delimited data: New Mark Resett Reset Latory Column Resett Reset Docolumn Resett Reset Docolumn Resett Reset Docolumn Reset Resett Reset Re		This screen iers you	Select Each Column	and set the bata i ormat.		
General General Converts numeric values to numbers, date values i Date MDV General Converts numeric values to numbers, date values i Data greview Do not import column (slop) Destination: 1051 Data greview Destination: 1		Column data form	hat			
rce: Microsoft Excel for Windows 2013. E: Step 3 of the Text to Columns wizerd for delimited data.		<u>G</u> eneral	10	anaral' consists numaric values to num	harr data valuar	
Detrementer in de la		9 Jed	to	dates, and all remaining values to text	uers, uate values	
Do not import column (skip) Destination: 1051 Data greniew Text:		Date: MDY	-	Advanced		
Do not import column (skp) Destination: DSI Data greview Description of Dog Rescription: dog Dog Rescript		- Part mor	100	Advanced		~~
Description of Dor Description of Dor Parties and shife PE Balance There There is an and shife PE Balance There is an an an an and shife P		Do not import	t column (skip)			
Data greview Data greview Data greview Deta greview		Destanting land			(82)	
Data greview Image: Image		Ogstination: SDS	1			
Data greeview Text Text Text increase increase <th></th> <th></th> <th></th> <th></th> <th></th> <th></th>						
Image: Inter- Image: Inter- Image: Inter- Image: Inter- Image: Inter- Image: Inter- Image: Inter- Image: Inter- Image: Inter- Image: Inter- Image: Inter- Image: Inter- Image: Inter- Image: Inter- Image: Inter- Image: Inter- </th <th></th> <th>Data newiew</th> <th></th> <th></th> <th></th> <th></th>		Data newiew				
Text:		Dara Bicarea				
Text Text Text Text Bestription of Dor Dor Deviced male brain wind white 215 Buill must Scoopy Beltered male brain Exact Boxer State Deviced male brain Boxer Boxer Scoopy Back Exact Ended Boxer Scoopy State State Exact Ended Exact Scoopy State State Exact Exact Exact Exact Scoopy State State Exact		1	The second			
Peerst prior of Dog resultered male red Scoby Story S		Text	Text	Text Text		
Person and white provide the provided of the p		Description of	of Dog	The state of the second states and the second se	-	
Booksy heatsered male prod Booksy payed formed prove months and white booksy payed formed prove and and the booksy payed formed prove and and the booksy payed forme		Rex	neutered	male ban and white Pit Bul	1 mint	
Concel		Scooby	neutered	male red Labrado	r Retriet	
Cancel C		Saler	Tale.	brown and white Rover		
The second seco		Salty	nare di	brown and white Boxer	Lan Charles	
Connection C		Junepug	spayed fe	male blue tick Austral	Lan Catol -	
Concel		1				
Cancel Ensist Intervention of Dog Note Centred Damgenou, Dogunhor - Excel 7 0 Provide Centred Damgenou, Dogunhor - Excel 7 0 0 Provide Centred Damgenou, Dogunhor - Excel 7 0						
Cancel Back Emith rce: Microsoft Excel for Windows 2013. Excel for Windows 2013. Excel for Windows 2013. ex: Step 3 of the Text to Columns wizard for delimited data. Annual Reference Dergenous Doputer - Excel 7 00 - 01 HMME MCERT For the Columns wizard for delimited data. Here Dergenous Doputer - Excel 7 00 - 01 HMME MCERT For the Columns wizard for delimited data. Here Dergenous Doputer - Excel 7 00 - 01 HMME MCERT For the Columns 21 and the Columns For the Columns For the Columns Description of Dog Reverse For the Columns Extended and the Columns Extended and the Columns Extended and the Columns Socoty Annual and the Boore Extended and the Columns Extended and the Columns Extended and the Extended Annual An						
rce: Microsoft Excel for Windows 2013. e: Step 3 of the Text to Columns wizard for delimited data:			Can	Kel K Back	Finish	
rce: Microsoft Excel for Windows 2013. E: Step 3 of the Text to Columns wizard for delimited data: HOME HEART PAGE LAYOUT FORMULAS OUT A REVEW VEW VEW HEARD OF THE THE TO CONSIDER HOME HEART PAGE LAYOUT FORMULAS OUT A REVEW VEW VEW HEARD OF THE THE TO CONSIDER HOME HEART PAGE LAYOUT FORMULAS OUT A REVEW VEW VEW THE A Ausyah - Constituted For the Constituted of the Very VEW VEW THE A Ausyah - Constituted of the Very VEW VEW HEARD OF THE THE TO CONSTITUTE OF THE THE TO CONSTITUTE OF THE CONSTITUTE OF THE TO CONSTITUTE OF THE CONSTITUTE OF T						
rce: Microsoft Excel for Windows 2013. e: Step 3 of the Text to Columns wizard for delimited data: Home Poster Poster Poster Poster Press Poster						
rce: Microsoft Excel for Windows 2013. e: Step 3 of the Text to Columns wizard for delimited data. Howe PSERF RAGE LAYOUT FORMULAS OUT A SUPER VEW		-				
Image: Source	e: Step 3 of th	ne Text to Columns w	izard for delimited	l data.		
Bit - S - C - Annual United Difference Description - totel P	e: Step 3 of th	ne Text to Columns w	izard for delimited	l data.		
ET HOME PAGE LAYOUT FOMULAS DATA ENTRY VEW Herzeg Dard Image Decision Image D	e: Step 3 of th	ne Text to Columns w	izard for delimited	data.		
Literenti Reference Connections 21 III III III IIII IIII IIII IIII	e: Step 3 of th	e Text to Columns w	rizard for delimited	data.	~ (2) 1	0 - D
Catend Retern Comment 24 (Al2) Fort model	E Step 3 of th	PREFIX TO COlumns w	rizard for delimited	tared_Dangerous_Dogs.shor-Excel	- (2) 7 (Her	0 - D
Lateral Art Art of the Anticipa Art of the Antice Art	e: Step 3 of th	PISERT PAGE LAYOUT 5	rizard for delimited	Lata. tered Dergenson, Rogiselon - Escel VRW VRW	Green - Yi	0 – 11 ag David 5
Data Difference Description Description Difference Description 10 1 Brain France Description Description 10 1 Brain Brain Description Description 10 1 Brain Brain Description Description 10 1 Brain Brain Brain Description Description 10 1 Brain Brain Brain Brain Description Description 11 Catalian Brain Brain Brain Brain Description Description 11 Catalian Brain Brain Brain Description Description 11 Catalian Brain Brain Brain Description Description 11 Catalian Brain Brain Brain Description	e: Step 3 of th	PAGE LAYOUT	Aqual Processor	Lata. tared Damperoux Doguster - Excel NEW VEW Train Fill Fran Fill Fran Fill Fran Consolidate	Toropp - 11	0 – D ag David 5
Connection Solt Writer Data Tools Outline 10 D F G Description of Dog F G Rex mild register mail Ian and white Pt Bull mix 1305 Webbenite PG Scooby mild register mail Ian and white Pt Bull mix 1305 Webbenite PG Scooby mild register mail ref G 1305 Webbenite PG Scooby mild register mail ref G 1305 Webbenite PG Status mild register mail torum and white Boxer 3304 Carely Creat/7 Suby mail torum and white Boxer 3310 Catalina Ormal Bumpy mailed twinte and Back Amincare Bull Termer 11511 Catalonia Drev Station mailed ref and Tan Gliden Retrinver/Chow mox 1501 Dateview Drev Mathier ref and Tan Gliden Retrinver/Chow mox 1501 Dateview Drev 1201 Cec/1774403 Status maile white and tan Galar Retrinver/Chow mox 1201 Cec/1774403 Nebbias	e: Step 3 of th	PIGERT PAGE LAYOUT F Connections 21 12 5ort F	Autor Day Autor Day Autor Day Autor Day Autor Day Autor Day Autor Day	Lata: Uterd Dampenou, Bogusho-Escél NRW VRW Falan Fal Marcos Duplicates W Wat & Analysis	Group - *1 Ungeoup - *1	0 – D ag David 5
10 b b b b b b b b b b b b b b b b b b b	Estemal Refresh Caternal Refresh Caternal	Posent Page Layout P Connections & Lip Growthing & Series Connections	Autor Der Autor	Lata. taret Dargenou, Dopulor-Escel NEW VEW VEW Flat File End End Consolidate File framous Duplicates IV With & Analysis - E file A Valdation - E file A	7 (Herr Ungroup - 1 Subtest	0 – Dag David 5
B F C Description of Dog Rex neutrapid male tan and white Pit Bull mix 1305 Webbenile Ro Scooby neutrapid male tred Labrador Ratineer mox 1305 Webbenile Ro Scooby maile brown and white Boxer 1304 Webbenile Ro Suby payoed female brown and white Boxer 1304 Catalian Orne7 Bumpy payoed female brown and white Boxer 1304 Catalian Orne7 Bumpy payoed female brontle and black Aminican Bull Termer 11511 Catalionia Drev Bumpy payoed female brontle and white Labrador Ratineer max 705 Texas 07/270/20 Bumpy payoed female mail and Galden Ratineer/Chow mox 1501 Datexies Drev Nabla neularid mailwhite Boxer mix 2401 Cec/17/24/03 2401 Cec/17/24/03 Smakry maile brown forndle Pit Bull mix 2401 Cec/17/24/03 2100 Cec/17/24/03 Nbblas Remaile veha and tan Galden Ratineer/Chow mox 1306 Bishon Dharv	e: Step 3 of th	PAGE LAYOUT S Committions 21 212 Committions	Administration	Lata. tered Damperoca, Dogusko-Escel VMW VMW File Famove Doploates III With & Auson - El Data Valdation - El Data Tools	Crebe - 1 Ungioup Subtatil Cuther s	0 – D ag David 5
D F C Description of Dog Ian and white Pit Bull mix 1305 Webbenite Ro Scooby netEried male red Labrador Retriever mox 1305 Webbenite Ro Scooby mail trid mile Labrador Retriever mox 1305 Webbenite Ro Sonoby mail trid mile Boxe 1305 Webbenite Ro Junebug appyed female blow not wheb Boxer 1301 Catalina Drive Burny nesderid male white and btack Aminican Bull Timer 11511 Catalonia Drive Little G nesderid male white and btack Aminican Bull Timer 1511 Catalonia Drive Viettie Isemale rel and tan Golden Retriever/Chow mox 1601 Dateword Drive Stebb nuklwite Labador Retriever Mox 2401 Cect/8744030. Stebba Kernale white Pot Bull mox 2401 Cect/8744030. Nebbas Kernale white Pit Bull mox 2401 Cect/8744030. Nebbas Kerale Pit Bull mox 2401 Cect/8744030. 1306 Billiem Ditor-1	e: Step 3 of th	Poser Page Layour P Connections 24 23 P Commentions 24 24 24 P Commentions 24 24 24 P Commentions 24 24 24 24 24 24 24 24 24 24 24 24 24	REMAULAS DATA ST COMMULAS DATA ST Comment Comment Comment Comment	taret Dimperios Dogustor-Escel Inter Dimperios Dogustor-Escel Inter VEW VEW Enter State Tar Consolidate Effetenses Diplotes El Wate & Auspus - C Est Data Validation - Estate Auspus - C Data Taolis	7 E Herz Group - 1 Subtes Subtes curne - 5	0 – D ag David 5
Description of Dog Pit Bull mix 1305 Webbenite Ro. Rex neutrept male tan and white Pit Bull mix 1305 Webbenite Ro. Scooby neitered male red Labrador Ratinieer mix 1305 Webbenite Ro. Saiby maje traven white Boxer 1304 Catalian Orne7 Subpy paged female brown and white Boxer 1304 Catalian Orne7 Bumpy paged female brown and white Boxer 1304 Catalian Orne7 Bumpy paged female white and black Aminican Bull Termer 11911 Catalionia Dim Bumpy regard female moltered male moltered Boxer 1201 Catalionia Dim Malarie neulard male moltered Catalina Chronic Datalion 1201 Catalionia Dim Malarie neulared male moltered Boxer mix 2010 Cacel/1724420 Smakry male brown filmfall Pit Bull mix 2401 Cacel/1724420 Smakry male red and tan Galden Ratineer/Dow may 1306 Billien Diver Nbblas Remale	e: Step 3 of th Home External Refresh Later Later	NEET PAGE LAYOUT &	Adders Darie Residuces Contraction Contrac	Lata. tered Damperous, Bogscher-Excel NetW VOW Thath F2P F=Consolidate R16 femores Duplicates 20 Wite & Autyon - 40 Ext Teach - 20 Femore Argen - 40 Cata Teach	Prop - 1 Krez Unginup - 1 Subtest Culture S	0 – 10 ag David 5
Description of both Iter and white Pit Bull mix 1305 Webbenite Ro Scooby nittleferd male red Labrador Retriever mix 1305 Webbenite Ro Scooby nittleferd male red Labrador Retriever mix 1305 Webbenite Ro Scooby nittleferd male red Labrador Retriever mix 1305 Webbenite Ro Suby applyed female brown and white Boxer 3310 Catalina DrmP Bumpy neudered male white and back Aminican Bull Timme 11511 Catalonia Drm Litta Go neudered male white and back Aminican Bull Timme 11511 Catalonia Drm Applied female real and tan Golden Retriever/Chow mix 8501 Dateiwe DrmW Smbky mid brown female Pit Bull mix 2401 Gec/1774403 Smbks female refaile Pit Bull mix 2401 Gec/1774403 Smbks female refaile Pit Bull mix 5505 Shoabood Aw Shebas female refaile Pit Bull mix 5505 Shoabood Aw Keely apayed	e: Step 3 of th	PISERT PAGE LAYOUT F Connections 21 212 (1) Lower 21 501 Fill (1) Fill (1) 501 Fill (1) 501 Fill (1) 501 Fill (1) 501 Fill (1) 501 Fill (1) 501 Fill (1) 501 Fill (1) 501 Fill (1) 501 Fill (1) 501 Fill (1) 501 Fill (1) 501 Fill (1) 501 Fill (1) 50	Addition of the second se	Lata. United Dimperiora Dogushor-Escel VNW VDW First Fall First Consolidate First Consolidate Cata Taoni	7 (Her Group - 1 Subtate Cusine 10	0 – II ag David 5
next next 120 Million Soldby next and an	e: Step 3 of th HOME HOME Data- Lib HOME Com	PAGE LAYOUT	Anner Der Collhaulus Collhaulus Collhaulus Collhaneed Collhaneed	tarel Dangeroux Doguston-Escel NEW VEW Maren Daglacies IV Was / Ankron - I Data Valaion Cata Team	Page - 1 Group - 1 Ungioup - 1 Subtest outres - 5	0 - 13 ag David S
Society Instante Feat Status	e: Step 3 of th	PICERT PAGE LAYOUT S Connections 21 TTP Connections	Removal and the second	Lata: Steel Damperson, Dogushor-Escel VEW VOW Fitterson Dogusters (White & Anatyon - C Eff Deta Validation - C End Deta Validation - C Data Team	7 C Herr Ungroup - 1 Subtati Coultine - 5 1996 Mo.	0 - 13 ag David 5
Jamebu Droten mol wmite Doken 3594 Gamby Clearly Janebug payled fematie biue nck Australian Cattle dog 3310 Catalina Onen Bumpy payled fematie white and Stack American Bull Temie 11511 Catalonia Dime Bumpy payled fematie bennike and Stack American Bull Temie 11511 Catalonia Dime Julia payled fematie bennike and take Golden Retinever/Chow mox 6501 Dateiwey Drow Julia payled maile molwhite Labador Retinever mox 705 Tenas S1787052 Politik fematie entide Boxer mix 2401 Cacel/B744(30 Smokky maile white Politike 2401 Cacel/B744(30 Nebbus Remaile white Politike 2306 Elsen/Branck Nebbus Remaile white	e: Step 3 of th HOME Deternal, Refresh Deternal,	Poser Page Layour P Connections & A A A A A A A A A A A A A A A A A A	Izard for delimited	Lata. Stred. Dargenou. Dogusko- Ecel New VEW EM Flat File EM Consolidate Flat move Dopicates EM Weat & Anayon – Cata Tools Cata Tools Pt Bull max Pt Bull max	7 0 Herr Group - 1 Subset Cutane S 1305 We	0 – D ag David y obenville Ro
Jonneoug appyet termale Diot Ick Australian Cattle dog 2310 Catalina Dirrer Bumpy neudered male winite and black. Aminican Bull Termie 11511 Catalonia Dirr Utgle Ga appyet female brendle and white Aminican Bull Termie 11511 Catalonia Dirr Bumpy appyet female brendle and white Aminican Bull Termie 11511 Catalonia Dirr Bush nudered male real and tan Golden Retinwer/Chow mx 8501 Datewer Dire Status nudered male real and tan Golden Retinwer/Chow mx 705 Texas St/7050; Status nudered male write Boxer mix 2401 Cec/1774403; Status Amine write Boxer mix 2401 Cec/1774403; Strebba Aerrale write and Golden Retinwer/Chow mx 2306 Bilem Diver/D Nopy Ismale red and tan Golden Retinwer/Chow mx 2306 Bilem Diver/D Nopy Ismale Istatus Golden Retinwer/Chow mx 5306 Shaalwood Ave Nopy Ismale Istatus Stephered mix 2520 East	e: Step 3 of th	Proceedings of the second seco	Removal for delimited Amer.phy Accession Acces	Labador Retriever mox	Toreup - 1 Statest Statest Cutane S 1305 We 1305 We	0 - D ag David 5 obenville Ro obenville Ro
pumpy multime matter white and Stack. Anniscan Bull Timus 1151 Catalona Dim Ito Latalona Dim Mediane Litta Ge spayed female benife and white Anniscan Bull Timus 1151 Catalona Dim Mediane Jebbie malared male mellowine Galden Retinuer/Chow mal 850 Dialeskee Dinal Stabie multiwrite Babar Tonas 372 7020 F 705 Dialeskee Dinal Smakry male briven Ismitle Boar mix 260 Cec/IT374403 Smakry male briven Ismitle Ph Bull mix 2401 Cec/IT374403 Smakry male mala and tain Galden Retinuer Chow mot 113206 Billem Diver Nobbas fermale whita PH Bull mix 2401 Cec/IT374403 Shebba fermale whita PH Bull mix 2401 Cec/IT374403 Nobby fermale whita PH Bull mix 2505 Bills Diver Nobpy fermale malark and tain Stapped fermale 550 Shabuod Mer Nobpy fermale Red Labrador Retinever mix 2502 East Jol Shaee <	e: Step 3 of th HOME HOME External Refresh Cover 10 Description c Rex Scoopy Salty	Posent Page Layout P Connections 21 23 Connections 21 23 Connectio	Annu De Collours Out A Collours Out A Collours Out A Collours Collours Collours Collours Collours and while transfer and while bottom and while	tarel Dimperios Dogustor-Escel VRW VRW Fitas Fit Fit Proceedades Efference Dogustor III wat A range - III Data Teals Efference Dogustor III wat A range - III Cata Teals Pt Bull mix Librador Refilement mox Boxes Data Teals	7 0 Herr Uningsoup ~ 1 Subseta Coutine ~ 1305 We 1305 We 1305 We	0 – D ag David s obenutie Ro obenutie Ro obenutie Ro
Little Ger spaywil female benifie and white Annixican Bull Tamer 1151 Cataliona Den Michael Michael female mil and tan Gidine Retriever/Crear max 850 Daleves Donay Stack nundered male mil/white Labrador Retriever/Crear max 705 Texas S78705() Pank Ismale white Labrador Retriever/Crear max 705 Texas S78705() Stack male brinne tamolie Pil bull max 2401 Cecl/87844(3) Stackbax male brinne tamolie Pil bull max 2401 Cecl/87844(3) Stackbax fermale will an Gidine Retriever/Crev moz 2506 Ebitem Diom? Noppy fermale Mad tan Shepherd max 5565 Shoalwood Ave Noppy fermale Red Labrador Retriever max 2520 East 3d Street Noppy native chever and ave Aminican Bull max 2520 East 3d Street Noppy native chever and aver Aminican Bull max 252 Shep Street 777 Nopp male chever and aver Patibul Terrise 255 Shep Street 774	e: Step 3 of th HOME HOME Description of Saty Junebug	A Pace Lavour & Pace Lavour	Aner De Aner De Aner De Column Tex to Data Tex	L data: Utred Dimperson, Bioglobio - Escel VRW V VRW Fisch F2 Factor F2 Cata Visidation - Cata Station Autors Cata Tools Pit Bull max Labrador Retriever mnx Boxer Australian Cattle dog	7 (Herz Unginup - 1 Statetal outme - 5 1305 We 1305 We 1305 We 1305 We	0 - D ag David 5 oberville Ro berville Ro berville Ro alma Onve?
Identified Instantian Goldern Retiniver/Chow mix 16501 Dateway Drawt Night nucleared male mil/wife Labrador Retiniver mix 705 Texas S718705() Porbin Iamrala jentre Bose mix 2401 Cocc/17374420 Smakky male brown bundle Pit Bull mix 2401 Cocc/17374420 Smakky male brown bundle Pit Bull mix 2401 Cocc/17374420 Smakky male whas Pit Bull mix 2401 Cocc/17374420 Nebbas Remaile red and tan Goldern Retinver/Chow mox 13206 Billiem Direct/ 137440Billiem Direct/ 13206 Billiem Direct/ 13206 Billiem Direct/ 13206 Billiem Direct/ 13206 Billiem Direct/ 13206 Billiem Direct/ 1320 Bill	e: Step 3 of th HOME HOME Common HOME Common Home Common Home Common Home Home Home Home Home Home Home Home	Pose Text to Columns w Pose Page Layout S Connections 21 12 Connections 21 12 Connect	Annu De Annu D	terel Dangerous Dogustor-Escel VRIW VRIW Ertenso Dogustor-Escel Paterova Dogustor Er Yorke A Auson - El Deta Valdator - Cata Taos: Pt Bull mix Labador Retriever mox Boxer Australian Cattle dog Aminican Dull Temar	7 0 Herz Urogevap ~ 1 Subtest Coutine ~ 1305 We 1305 We 1305 We 1305 We 1305 We 1305 We 1305 We	0 – D ag David 5 oberstle Ro oberstle Ro berstle Ro sky Creek78 alana Driver talonia Driver
Jigk nuklered male multiWrite Labrador Retriever mot 705 Tonces 0777070; Perfor Ismails jentale Boxer mix 2401 Cacil/3744(30. Smokiny male brown fumdile Pit Bull mox 2401 Cacil/3744(30. Shebba kernale whas Pit Bull mox 2401 Cacil/3744(30. Shebba kernale whas Pit Bull mox 2401 Cacil/3744(30. Shebba kernale whas Pit Bull mox 2401 Cacil/3744(30. Shebba kernale whas and an Gallen Retriever/Chow mox 13206 Billem Direvit Nappy kernale Mad and an Shephard max 5506 Shoalwood Mwx Krefy spayed finalle Rad Labrador Ratriever mox 2520 East 3rd Street Marky naticare dimaile white and brown Aminican Diulog gricx 1128 Richadine Awx Ohico male chocylate and white Patter mix 525 East 3rd Street Chicol male chown and white Patter mix 525 Street/3rd 205 Toremail Pass Call	e: Step 3 of th HOME	A Text to Columns w A A A A A A A A A A A A A A A A A A A	Aner De Aner De Consults Data Test to Data Test Test Test Test Test Test Test Test Test Test Test Test Test Test	Labador Astrinova Burganhar - Escel Vistor Vistor V Vistor - Escel Fisch File Marcova Dopicates 12 Vistor J Autorn - 12 Marcova Dopicates 12 Vistor J Autorn - 12 Cata Visidation - 12 Fischer Marcova Cata Tools Pit Bull mix Labrador Retriever mox Booxe Australian Cattle dog Anteincan Bull Timer Amenican Bull Timer	7 (Hers Unginup Statetal culline - culline - 3390 Cat 1511 Ca 1511 Ca 1511 Ca	0 – D ag Divid 5 obendle Ro obendle Ro ey Creek78 alma Drive7 alma Drive7 alma Drive
PinkToc. Iamala Jeniste Boxe mix 2401 Coc/II774403 Smakty mia brum familia Pit Bull mox 2401 Coc/II774403 Shabba Armale wiaie Pit Bull mox 2401 Coc/II774403 Nabbas Armale red and tan Ocidian RetineverChow mox 13206 Billium Divert/ 13206 Billium Divert/ School School Schol School Schol School School Schol Schol School School Schol Sc	e: Step 3 of th HOME HOME Control Control Description of Rex Scooby Salty Junebug Bumpy Little Control	Pose Layout F Pose Layout F Connections 21 [1] Connections	Annu De Annu D	Lata. Uterd Dimperois Dopscher-Estel VRW VRW Field Fill Pill Pill Pill Pill Pill Pill Pill	7 0 Herr Greup - 1 Subtotal Cutane - 1305 We 1305 We	0 - D ag David 5 obervile Ro obervile Ro obervile Ro alma Onve? alma Drive? alma Drive? alma Drive?
Smakky maie brum temile Pit Bull mix 2411 Cect/T774403 Shebba Marale white Pit Bull mix 2401 Cect/T774403 Shebba Marale white Pit Bull mix 2401 Cect/T774403 Neppy Marale wind and tan Golden Retinever/Chow mox 13206 Billiem Dive/T Neppy Marale black and tan Shephard mix 5506 Shoahvod Aw Keely spayed Binale Red Labador Retinever mix 5506 Shoahvod Aw Markey neutred male white and brown Aminican Buildog mix 1128 Richardine Aw Ohico male chocabae and white Pointer mix 5205 Ebit ShoaPhood Aw Locoln male bawr and white Pointer mix 5205 Thommid Pass Cali female black Labrador Retriever mix 1202 Richcreek Roa	e: Step 3 of th HOME	NSERT PAGE LAYOUT S NSERT PAGE LAYOUT S Connections & A Ale S Con	Amer pro- Amer pro- control and american and and the real and tan mativities	Labrador Retriever mos Borten Durgenson, Dogusko-Escel VIEW VIEW Proceedidate Plate Teat Fill Data Validation - Consolidate Esta Tatos Data Validation - Consolidate Data Validation - Consolidate Cata Tatos Data Validation - Consolidate Data Validation - Consolidate Cata Tatos Plit Bull mile Labrador Retriever mos Boote Austratian Catific dog Aminican Bull Termer Aminican Bull Termer	7 0 Herr Ungroup - 1 Subtest outree - 1305 We 1305 We 1305 We 1305 We 1305 We 1305 Ue 1311 Ca 1511 Ca 1511 Ca 1511 Ca	0 - D ag David 5 bbendle Ro objendle Ro ey Creek78 alma Drive7 talonia Drive ristw Driva 5 st78705()
Shebba Kernale while Pt Bull mix 2401 Cect/0724020 Nöböss Karnale red and tan Goldan Retineve/Dow mox 13206 Billiem Drev? Nöböss Karnale Isak and Lan Shepherd mix 5566 Shaalwood Ave Kreely spayed female Red Labrack Reteinvert mix 2520 East 3rd Street Markey neutraced male white and brown Aminican Bulldog mix 1128 Richardine Ave Ohico male chocate and white Pointer mix 2555 Billion Shep Street/074 Lincolm male bourt and white Pointer mix 2555 Thormaid Pass Cal female black Labrador Retriever mix 1202 Pic/nevel Ras	e: Step 3 of th HOME HOME Com Description c Rex Scooby Salty Junebug Fump Address Addr	Proceedings of the second seco	Annu De Annu De Colliculas Column Annu De Column Annu De Column An	Latar. Latar. United Damperous Dopusher-Escel VNW VDW Fisch Fill Fisch Fill Methods Dapicates De Wited A Ausyon Ed Data Validation – Standonskyn Catar Econ Pit Bull mick Labrador Retriever max Boxer American Bull Temer American Bull Temer American Bull Temer Gatar Retriever Chow mot Labrador Retriever max Boxer mix	7 0 Herr Greup - 1 Subtotal Cuttine & 1305 We 1305 We 1305 Cat 1305 Cat 1305 Cat 1305 Cat 1315 Cat 1511 Ca 1515 Ca 1515 Ca 1515 Ca 1515 Ca 1515 Ca 1515 Ca	0 - 13 ag David 5 beendle Roo ey Creek7a alonia Drive talonia Drive s St78705() af St78705()
Nebbles Kemale red and tan Golden Retriever/Crow mot 13206 Billiem Drive? Neppy kemale Mach and tan Shepherd mix. 5506 Shoalwood Ave. Keely spyword firmale Red Labador Retriever mot 2520 East 30 Street Markey neutrared male white and brown Amsincan Buildog mix. 1128 Richardine Ave. Onico male choolae and white Pointer mot. 2525 East 305 Street/3721. Uncoln male bawr and white Pointer mot. 2305 Thormidi Pass Cali female black Labrador Retriever mot. 1202 Richcreek Roa	e: Step 3 of th HOME HOME Description of Rex Scooby Saby Jonebug Bumpy Litle General South Jonebug South Jonebug South Jonebug South Jonebug South Jonebug Jon	NEText to Columns w NEText to Columns w NETEXT Pace Layour S Connections AL Art Al Art Connections AL Art Art Al Art Al Art Art Art Art Art Art Art Art Art Art Art Art Art	Izard for delimited	Lata. Darred Dargeroux, Dogschor-Escel NEW VEW VEW Flat File File File Consolidate Filemoux Dogschor: El With & Anayou - El Ena Vaidation - Ena Tools: Pit Bull mix Labrador Retriever mox Boxie Australian Cattle dog Aminican Bull Times Golder Retriever Chow mox Labrador Retriever mox Boxie mix Boxie mix Boxie mix	7 0 Herr Ungroup - 1 Subtest Outene - 1305 We 1305 WE	0 - D ag David 5 oberville Ro berville Ro by CreeK78 aliana Driver talonia Driver siste Drive s St78705(2) s178744(30) s178744(30)
Neppy Kernale black and Ian Shephard mail 5506 Shoakwood Mw. Keely spayed finalle Rad Labrador Ratineer mix 2520 East 3rd Street Marky neutrated maile white and brown Aminican Dullog mix 1128 Richardine Awr Ohico maile choclate and white Pointer mix 525 East 5rd Street Uncoln maile bown and white Pointer mix 525 Shoakwood Mw. Cali female bown and white Pointer mix 225 Shoemwidi Pass Cali female black Labrador Ratinewr 1202 Richcreek Roa	e: Step 3 of th HOME HOME Care Correction Care Correction Co	A Proce Layour s Connections 21 22 2 Connections	ECHAULAS DATA EL CALANCERS DATA EL CALANCERS DATA EL CALANCERS DATA EL CALANCERS DATA EL CALANCERS DATA CALANCERS DATA CALANCE	Lata: Jared Damperson, Dopushar-Escel VRW VRW Fither Table Depictes (2) With & Autyon (2) Effermore Dopicates (2) With & Autyon (2) Effermore Dopicates (2) With & Autyon (2) Cata Teas Cata Te	7 C Herz Greige – *1 Subrotal Coutine © 1306 We 1306 We 1306 We 1306 Ve 1306 Ca 15151 Ca 1515	0 – II ag David 5 beenville Ro béprille Ro béprille Ro béprille Ro sé Creek 7 alona Drive rister Drive 7 alona Drive stratoria Drive stratoria Drive alonia Drive stratoria 178744(30 al78744(30
Keely spayed firmule Red Labrador Rathever mix 2520 East 3rd Street Markey neutreed male while and brown Aminician Buildog mix 1128 Richardina Awo Ohico male chocolate and white Pointer mix 525 Shep Street 734* Lincoln male bawn and white Pit Buil Terrier 2305 Thormaid Pass Cali lemale black Labrador Rathever mix 1202 Richcreek Roa	e: Step 3 of th HOME HOME HOME HOME HOME HOME HOME HOME	A Text to Columns w A HEAT Pace LAYOUT Connections A A A A A Connections A A A A A A A A A A A A A A A A A A	Annu DA Annu D	Labara Street, Dangeroux, Dogustor-Escel VNEW VEW Fittenson Dogustor-Escel Pattenson Dogustor-Escel Pattenson Dogustor Event Pattenson Dogustor Cata Taols Pt Bull mix Labardor Reference max Boxer Australian Cattle dog Aminican Bull Times Cata Taols Pt Bull mix Boxer max Boxer max	7 0 Herz Ungesup ~ 1 Subtest Cultime © 1305 We 1305 We 3304 Car 3310 Cat 1511 Ca 15151 Ca 15151 Ca 2401 Cac 2401 Cac 2401 Cac 2401 Cac	0 – D ag David - bohnile Ro ey Cresh78 ey Cresh78 statona Dma Statona Dma Stat
Marley neutoxed mails while and brown Aminican Buildog mix 1128 Richardine Awinican Buildog mix Ohico maile choclate and white Pointer mix 625 Shep Street/1721 Locoln maile bawn and white Pic Buil Terrifer 2305 Thormail Pass Cali female black Labrador Retriever mix 1202 Richcreek Roa	e: Step 3 of th HOME HOME Carrier C	A Text to Columns w A Text to Columns w A Start Pace Layour s Connections 21 A Text Connections 21 A Text Connections 21 A Text A Sort Sort s Connections 21 A Text A Sort S Connectio	Annu Der Reinhulles Re	Labrado Reference Doguster-Encel VERW VERW France Doguster-Encel Presson Dogusters (2014) Antyon - (2014) Effermous Dogusters (2014) Antyon - (2014) Effermous Dogusters (2014) Antyon - (2014) Data Validation - (2014) Antyon - (2014) Data Validation - (2014) Antyon - (2014) Cata Team Data Team Data Team Data Team Antyon - (2014) Antyon - (2014) Data Validation - (2014) Antyon - (2014) Data Validation - (2014) Antyon - (2014) Data Team Antyon - (2014) Antyon - (2014) Data Team Antyon - (2014) Antyon - (2014) Data Team Antyon - (2014) Antyon - (2014) Antyon - (2014) Anty	7 C Hez Greug - 1 Subrotal Coutine © 1305 We 3300 Cat 1305 We 3300 Cat 1511 Ca 1511 Ca	2 – 12 ag David 5 bobenile Ro bobenile Ro bobenile Ro bobenile Ro do bobenile Ro
Ohice male chocobae and white Pointer moc 525 Shep Streat773. Lincolm male favor and white Pix Buil Terrise 2305 Thormald Pass Cali farmale black Labodor Ratewort moc 1202 Pic/Lineek Road	e: Step 3 of th HOME	A Text to Columns w A HOERT Pace Layour A Connections & A A A A Connections & A A A A Connections & A A A A Connections & A A A A Connections & A A A A A A A A Connections & A A A A A A A A A A A A A A A A A A	Annu De Annu D	Labado References Dorgunhor-Escel VNW VEW Fitale Fital Fital Fital Consolidate Fitalences Diplocates EV What Manager (Ell Data Validation - Ell Statussical Consolidate Ell Data Validation - Ell Statussical Consolidate Cata Taolos Pit Bull mix Labador References Tonos mix Labador References Tonos mix Labador References mix Pit Bull mix Ell Data Validation Tames Pit Bull mix Labador References mix Pit Bull mix Ell Data References mix Staphard mix Labador References mix	7 0 Herz Greuge − 11 Subtest Coutine ∞ 1305 We 1305 W	0 – II ag David 5 bohnvite Road 5 bohnvite Roa
Lincoln male burn and white Pr. Bull Terrier 2305 Thornwild Pass Cali female black Labrador Ratnews mix 1202 Richcreek Roar	e: Step 3 of th HOME	A Text to Columns were text to	Anner Der Anner	Labrador Retriever max Boxer mix Pt Bull max Cata Total Pt Bull max Boxer mix Boxer mix Ammican Bull Times Ammican Bull Times Ammican Bull Times Ammican Bull Times Boxer mix Boxer mix Boxer mix Boxer mix Boxer mix Boxer mix Ammican Bull Times Ammican Bull Times Boxer mix Boxer	7 C Herr Jorginup - 1 Subtotal Coutine	0 - ang David 's ang David 's bebenutie Roa bebenutie Roa
Cali female black Labrador Retriever mic 1202 Richcreek Road	e: Step 3 of th HOME	Pace Layout Pace Layo	Annu De Annu D	Labada Ratineer Choe mo Cata Color Pitabili mic Labada Ratineer mic Boxe Manina Catlle dog Aminican Bull Times Autorian Catlle dog Aminican Bull Times Autorian Catlle dog Aminican Bull Times Autorian Bull Times Boxe mix Pit Bull mix Labada Ratineer Times Autorian Guildog mix Aminican Bull Times	7 0 Herr Greup - 1 Subtotal cutme - 1305 We 1305 WE 13	0 – David 5 obenetie Ro. obenetie Ro. Saltziakow Processor Saltziakow Processor Jatoria Drwe Jatoria Drwe Jat
The second secon	e: Step 3 of th HOME	A Text to Columns were text to	Izard for delimited	Labrador Refineer max Boxer Manine Bull max Pt Bull max Pt Bull max Data Tatas	7 C Herr Greige - 1 Subtotal Cuttine C 1305 We 3904 Car 3130 Cat 15151 Ca 15151 Ca 15151 Ca 15151 Ca 2401 Cac 2401 Cac 2401 Cac 2401 Cac 2401 Cac 2401 Cac 2401 Cac 2405 Cac 1505 Step 2555 Fac	0 - ang David 's ang David 's behavide Roo behavide Roo behavide Roo behavide Roo behavide Roo behavide Roo Balancia Draw Balancia Marka Straff 2014 178744(0) 17874(0) 17874(0) 17874(0)
	e: Step 3 of th HOME	Pack Layout Pack Layo	Annual An	Labrado Retineer Those mo Cater Jongenous Dogushor- Escel VNW VDV Fish Fill Fish Fill Fish Fill Fill Construction Cater Labrador Retineer Those Autorian Catille dog American Bull Theme Cater Theme Autorian Catille dog American Bull Theme Cater Theme Autorian Catille dog American Bull Theme Cater Theme Cater Theme Autorian Catille dog American Bull Theme Cater	7 0 Herr Greup = 1 Subtotal Cutane © 1305 We 3300 Car 3310 Cat 3350 We 3390 Car 3310 Cat 3310 Cat 3300	0 – II ag David 5 ag David 5 behavid Fo 5 behavid Fo 5 behavid Fo 5 behavid Fo 5 behavid Fo 5 stratog 5 st

Source: Retrieved from Data.austintexas.gov.

Note: Parsed Description of Dog column. Note that each of the four pieces of data is stored in its own column.

So Column D holds the name of the dog, Column E the sex and whether it's been neutered or spayed, F the color and G the breed. Now that we've placed the breed of the dog in its own column, we could create an Excel pivot table that tells us which breed is the most common on the dangerous dogs list. Let's practice safe computing

100 DATA LITERACY

by adding some labels that make sense: Change the contents of cell D1 to "name", E1 to "sex", F1 to "color" and G1 to "breed". Now save the file and close it because we are done with it for now.

Carving up data using Text to Columns usually is more complicated than this because of the way government agencies store data. Get the Excel file fuel_economy2013.xlsx from the book website and open it. Make a working copy using File | Save As. This file from the U.S Department of Energy and the Environmental Protection Agency lists fuel economy data for vehicles that are sold in the United States. The data are in 1,166 rows, including one for the headers. Let's scroll over to column AC, which provides descriptive data about each vehicle's drive system. The first line tells us that the drive system for one particular variant of the Aston Martin V8 Vantage is "2-Wheel Drive, Rear". That's actually two pieces of data: the drive system is two-wheel and those two wheels are the rear ones. We could use the comma and space that follows it to carve this into two distinct columns.

Using Insert, create two new columns, one for the number of wheels and the other to designate which wheels propel the vehicle. Copy the contents of AC and paste them into AD. Make sure AD is highlighted and start the Text to Columns Function.

Choose delimited in the first screen of the wizard and pick Next.



Source: Microsoft Excel for Windows 2013.

Note: Step 1 in the Text to Columns wizard for delimited data.

Set the delimiter to Comma only and make sure text qualifier is None. Excel gives a preview of how it's going to parse the data. Look closely and you will see that Excel is including the spaces on either side of the comma as text. That's OK for now. We'll fix that later.

This screen lets you s in the preview below	et the delimiters your data contains. You can see how your text is affected v.	
Delimiters Tab Semicolon Gomma Space Qther: Data greview	Treat consecutive delimiters as one Text gualifier:	Ň
Drive Desc	Rear	
2-Wheel Drive 2-Wheel Drive 2-Wheel Drive 2-Wheel Drive	Rear Rear	

Source: Microsoft Excel for Windows 2013.

Note: Step 2 in the Text to Columns wizard for delimited data

Click Next to move on to the third and final step. Change both of these columns to Text and then Finish.

	This screen lets you select each co Column data format	olumn and set the Data Format.	
	General Text Date: MDY Do not import column (kto	"General" converts numeric value to dates, and all remaining valu Advance	es to numbers, date values es to text. d
~	Destination: SADS1		
20	Text Drive Desc 2-Wheel Drive Rear		-
•	2-Wheel Drive Rear 2-Wheel Drive Rear		

Source: Microsoft Excel for Windows 2013.

Note: Step 3 in the Text to Columns wizard for delimited data.

This is great—Excel has carved the data up in Columns AD and AE. Label AD "NumWheels" (for number of wheels) and AE "DriveWheels".

AC	AD	AE	
Drive Desc	Drive Desc		
2-Wheel Drive, Rear	2-Wheel Drive	Rear	
2-Wheel Drive, Rear	2-Wheel Drive	Rear	
2-Wheel Drive, Rear	2-Wheel Drive	Rear	
2-Wheel Drive, Rear	2-Wheel Drive	Rear	
All Wheel Drive	All Wheel Drive		
2-Wheel Drive, Rear	2-Wheel Drive	Rear	
2-Wheel Drive, Rear	2-Wheel Drive	Rear	~
2-Wheel Drive, Rear	2-Wheel Drive	Rear	
2-Wheel Drive, Rear	2-Wheel Drive	Rear	
2-Wheel Drive, Rear	2-Wheel Drive	Rear	
All Wheel Drive	All Wheel Drive		

Source: Department of Energy.

Note: Data parsed, but with extraneous spaces.

Our next step is to remove those extraneous spaces using Excel. We want to remove these because the spaces technically are data (one of the basic ASCII characters) and could throw off any analysis. Column AE has leading spaces in the cells. We have "Rear" or "Front" for the values, when we really want no leading spaces in them. Excel's TRIM function can help here. TRIM removes any leading and trailing spaces, as well as duplicated spaces inside a text entry.

Insert a new column to the right of AE and make sure it is formatted as General, or else your formula will appear, instead of "Rear" or "Front". Format the column by highlighting it, then clicking the Home tab. Select General from the options in the drop-down list.

	NOME INS	ERT PAGE LAYOUT	FORMULAS DA	TA	fuel_economy REVIEW	2013.xlsx - E VIEW	Exc
A X	Arial	- 10 - A A	===*.	-	General	+]
Paste *	B I U	• · · · · · ·	三三王 化化	•	\$ - %	•00 ±00	
Clipboard 15		Font ra	Alignment	Ģ	Numb	er r	ā,
AF1	112	V fe					
41	AC	AD	AE			ÀF.	
1 Drive De	SC	NumWheels	DriveWheels		1		
2 2-Wheel	Drive, Rear	2-Wheel Drive	Rear				
3 2-Wheel	Drive, Rear	2-Wheel Drive	Rear				
4 2-Wheel	Drive, Rear	2-Wheel Drive	Rear				
5 2-Wheel I	Drive, Rear	2-Wheel Drive	Rear				
6 All Whee	Drive	All Wheel Drive					
7 2-Wheel	Drive, Rear	2-Wheel Drive	Rear		distant second		
8 2-Wheel	Drive, Rear	2-Wheel Drive	Rear				
9 2-Wheel	Drive, Rear	2-Wheel Drive	Rear				
10 2-Wheel	Drive, Rear	2-Wheel Drive	Rear		10000		
11 2 M/hool 1	Drive Done	2. Mhool Drive	Deat				

Source: Department of Energy.

Note: New column in fuel efficiency data.

In cell AF2, enter the trim formula: "=TRIM(AE2)". This says, Trim the contents of cell AE2. Hit Enter and you should see the new, trimmed version. Copy this for all of your cells in this column by double-clicking on the square at the bottom right of your cursor box.



These are just two examples of how you can carve data in Excel using Text to Columns. Sometimes we do the opposite and combine the contents of multiple columns into one. For that, we turn to concatenation.

CONCATENATE TO PASTE

We're going to use **concatenation** to put election date data into a proper form. Open the voterturnout.xlsx Excel file, which holds data about voter turnout by precinct in Boone County, Missouri, during the November 2, 2010, general election. The spreadsheet has columns for the precinct or voting district, number of people who voted, number registered and date. All of the dates are recorded as "20101102". Our goal is to turn all of those to 11/2/2010.

We'll use the Text to Columns tool to carve the date and then use concatenation to put the pieces back together. Practice safe computing by copying the contents of Column D into Column E, which we will parse. Highlight E and pick the Text to Columns button from the data menu. That launches a familiar wizard to help. Our data are not delimited this time, so punch the Fixed width button. We can tell that the dates are fixed width because the characters for year, month and day are in consistent locations.

e Text Wizard has d	etermined that your dat	ta is Delimited.					
this is correct, choos	e Next, or choose the c	data type that be	st describes your	data.			
Driginal data type							
Choose the file type	that best describes yo	ur data:					
© Delimited	- Characters such as	commas or tabs	separate each fiel	d.			
Fixed width	- Fields are aligned in	n columns with sp	aces between eac	h field.			
Preview of selected	data:					10	S
Preview of selected	data:					3	
Preview of selected 1 Date 2 20101102 3 20101102	data:						
Preview of selected 1 Date 2 20101102 3 20101102 4 20101102 4 20101102	data:				~		
Preview of selected 1 Date 2 20101102 3 20101102 4 20101102 5 20101102 6 20101102	data;				Š		
Preview of selected 1 Date 2 20101102 3 20101102 4 20101102 5 20101102 6 20101102 1	dətə:	_			Ś		
Preview of selected 2 20101102 3 20101102 4 20101102 5 20101102 6 20101102 4	dətə:			ż	Ś		

Source: Microsoft Excel for Windows 2013.

Note: Step 1 in Text to Columns for fixed-width text.

Click Next and we're at Step 2. In this screen, we create the column dividers by clicking where we'd like them to go inside the Data preview area. Don't worry if you put a break in the wrong position—Excel allows you to move or remove it. Create column breaks at positions 4 and 6, just like this.

s you set fiel ows signify a a break line, a break line, break line,	ld widths (co a column bro , click at the , double clic click and dro	olumn brea eak. desired po k on the lin ag it.	ks). Isition. Ie.		
a break line, a break line, i break line,	, click at the double clic click and dr	desired po k on the lin ag it.	isition. ie.		
10		30	40	50	
					-
	<u>10</u>	<u>10 20</u>	<u>10 20 30</u>	<u>10 20 30 40 .</u> .	<u>10</u>

Source: Microsoft Excel for Windows 2013.

Note: Step 2 in Text to Columns for fixed-width text.

	ts you select each col	lumn and set the Data Format.	-		
Column dat General Text Date: Do not	a format MDY 💽 import column (skip)	"General" converts numeric to dates, and all remaining <u>Adv</u>	values to numbers, o values to text. ranced	date values	5
Destination:	SES1				
Data grevie	w xt				
20101102 20101102 20101102					
	6	Cancel < Back	Next>	Einish	

Click Next and go to the final step. Make sure all of the new columns are set to text and click Finish.

Source: M

Note: Step 3 in

Success! Excel has put the year in Column E, month in F and date in G. Label those three columns accordingly and save a working copy of the file with File | Save as.

1	Precinct	Voting	Registered	Date	Date	ī	- 8.
2	1A	360	1030	20101102	2010	11	02
3,	1B	234	420	20101102	2010	11	02
4	10	708	1159	20101102	2010	11	02
5	1D	474	1136	20101102	2010	11	02
6	1E	75	2704	20101102	2010	11	02
7	1F	162	1011	20101102	2010	11	02
8	1G	373	759	20101102	2010	11	102
9	11	163	1026	20101102	2010	11	02
10	2A	629	957	20101102	2010	11	02
11	2B	503	1171	20101102	2010	11	02
1.2	20	E25	1212	20101102	5010	711	502

Source: Boone County, Missouri, Clerk.

Note: Fixed-width text parsed.

Label Column H as "Date2", which is where we'll put the restructured date that we'll build with concatenation, which uses the ampersand (&) character.

In cell H2 enter "=F2&"/"". This tells Excel to take the contents of cell F2 (or 11 for the month of November) and tack on a slash. We need to put the slash in double quotation marks, so Excel treats it as a text character. We get "11/" for a result, which is a good start.

Next we will tack on the date, by building on the formula "=F2&"/"&G2". Now we see "11/02".

As the last step, will add another slash and the year: "=F2&"/"&G2&"/"&E2". Now we have "11/02/2010" in our cell. Copy this for all of the cells in Column H and format it as a short date. Save the file and close it.

DATE TRICKS

Excel offers other functions that transform data stored in dates. These are handy for generating a month or a year. Open the disaster_declarations.xlsx Excel file and we'll see how this works. This file, originally downloaded from the federal government's Data.gov portal, holds data about more than 4,100 major disasters declared since 1953. Under federal law, state governors may request a declaration of a major disaster and receive federal assistance (Federal Emergency Management Agency, n.d.). In Column G, we have a declaration date, but not a month or year. So it would be difficult if we wanted to analyze the data from either of those dimensions.

Excel's YEAR and MONTH function can extract these data. We'll use another function to generate the day of week, by name, for the declaration dates.

Insert three columns between columns G and H, so we have a place to put the year, month and day of the week. Label these "Year", "Month" and "Day". Format these columns as General. In cell H4, enter "=Year(G4)" and "1953" appears. Copy this formula for all of the cells. In cell I4, enter "=Month(G4)" and "5" appears for May. Copy this formula for all of the cells. In cell J4, enter "=TEXT(A4, "dddd")" and Sunday appears. Copy this formula for all of the cells. In cell J4, enter "=TEXT(A4, "dddd")" and Sunday appears. Copy this formula for all of the cells. The TEXT function, in this case, returns a text value from the data stored in A4. dddd tells Excel to generate the day using its full name. Use ddd to generate an abbreviation. Use File I Save As to save a working copy of your file and close it.

	RIE I	HOME INS	ERT PAGE	LAYGUT	HORMULA	desaut	ar_declarations]. (TA REVIEW	VEW SEW	Biaty Model - two	d	2.13	Y II Herz	og David -
Pi Ci	abooid in	Arial B 7 U	·	к . . А -	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	₽ 60 40. 1411	General General S - 1 rs - 1	iumbei	Conditional Formitting -	Formult as Cell Table * Styles * Styles	Er Invert - Delete - Format - Calls	∑ - A Z Soit & For C - Filter - Sele Editing	180 ct -
34	£		1 10	=TENT	A4, "dddd")							
	A		c	ø	E	F	G	н	1.0	3	K	L	
- 1 N.	Neimber	Deciared	Deciared	Decrahed	Decrarea	24946	0.000	rear	Month	Dey	Type	Incident Type	
	100000-000	- I ORCINFED	Decision	Decisioned.	1 IMPCIDIFED	- Section -		TOUL	Month	Day	Type	Incident Type	
	1	No	Yes	Yes	Yes	GA	5/2/1953	1953	5	Sunday	DR	Toinado	TORNADO
$-\mathbf{Q}$	1 2	No No	Yes Yes	Yès Yès	Yes	GA TX	5/2/1953 5/15/1953	1953 1953	5	Sunday Monday	DR DR	Toinado Toinado Toinado	TORNAD
	1 2 3	No No No	Yes Yes Yes	Yès Yès Yès Yes	Yes Yes Yes	GA TX LA	5/2/1953 5/15/1953 5/29/1953	1953 1953 1953	5	Bay Sunday Monday Tuesday	DR DR DR	Tomado Tomado Tomado Filood	TORNADI TORNADI FLOOD
	1 2 3 4	No No No No	Yes Yes Yes Yes	Yès Yès Yès Yès	Yes Yes Yes Yes	GA TX LA MI	5/2/1953 5/15/1953 5/29/1953 6/2/1953	1953 1953 1953 1953	5 5	Sunday Sunday Monday Tuesday Wednesday	DR DR DR DR DR	Tomado Tomado Tomado Ficod Tomado	TORNADI TORNADI FLOOD TORNADI
	1 2 3 4 5	No No No No No	Yes Yes Yes Yes Yes	Yes Yes Yes Yes Yes	Yes Yes Yes Yes Yes	GA TX LA MI	5/2/1953 5/15/1053 5/29/1953 6/2/1953 6/8/1953	1953 1953 1953 1953 1953	5 5 6 6	Sunday Sunday Monday Tuesday Wednesday Thuraday	DR DR DR DR DR DR	Ternado Ternado Ficod Tornado Ficod	TORNADO TORNADO FLOOD TORNADO FLOODS
	1 2 3 4 5 8	No No No No No	Yes Yes Yes Yes Yes Yes	Yes Yes Yes Yes Yes Yes	Yes Yes Yes Yes Yes Yes	GA TX LA MI MI	5/2/1953 5/15/1053 5/29/1953 6/2/1953 6/8/1953 6/0/1953	1953 1953 1953 1953 1953 1953	5 5 6 6	Sunday Sunday Manday Tuesday Wednesday Thursday Friday	DR DR DR DR DR DR DR DR	Toinado Ternadu Ficod Toinado Ficod Tornado	TORNADO TORNADO FLOOD TORNADO FLOODS TORNADO
	1 2 3 4 5 6 7	No No No No No No	Yes Yes Yes Yes Yes Yes Yes	Yes Yes Yes Yes Yes Yes Yes	Yes Yes Yes Yes Yes Yes Yes	GA TX LA MI MI MA	5/2/1953 5/15/1053 5/29/1953 6/2/1953 6/6/1953 6/11/1953 6/11/1953	1953 1953 1953 1953 1953 1953 1953 1953	5 5 6 6 6	Sanday Sunday Monday Tuesday Wednesday Phuraday Fhiday Saturday	DR DR DR DR DR DR DR DR DR DR	Toinado Tornado Ficod Tornado Ficod Tornado Tornado	TORNADO TORNADO FLOOD TORNADO FLOODS TORNADO TORNADO
	1 2 3 4 5 6 7 8	No No No No No No No	Yes Yes Yes Yes Yes Yes Yes Yes Yes	Yes Yes Yes Yes Yes Yes Yes Yes	Yes Yes Yes Yes Yes Yes Yes Yes	GA TX LA MI MI MA IA TX	5/2/1953 5/15/1953 5/29/1953 6/2/1953 6/0/1953 6/11/1953 6/11/1953 6/11/1953	1953 1953 1953 1953 1953 1953 1953 1953	8 5 6 6 6 6 6	Sanday Sunday Monday Tuesday Wednesday Wednesday Thursday Fhiday Saturday Saturday Sunday	DR DR DR DR DR DR DR DR DR DR DR	Teinado Ternado Ficod Ternado Ficod Ficod Ternado Ficod Ficod Ficod	TORNADO TORNADO FLOOD TORNADO FLOODS TORNADO FLOOD FLOOD
	1 2 3 4 5 6 7 8 9	No No No No No No No	Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes	Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes	Yes Yes Yes Yes Yes Yes Yes Yes Yes	GA TX LA MI MI MI MA TX LA	5/2/1963 5/15/1963 5/29/1953 6/2/1953 6/8/1953 6/8/1953 6/11/1953 6/11/1953 6/11/1953 5/19/1953	1953 1953 1953 1953 1953 1953 1953 1953	5 5 6 6 6 6 6 7	Sunday Menday Tuesday Wednesday Phiday Friday Saturday Sunday Menday	DR DR DR DR DR DR DR DR DR DR DR DR DR	Tomado Tomado Tomado Tomado Flood Flood Flood Flood Flood Flood	TORNADO TORNADO FLOOD TORNADO FLOODS TORNADO FLOOD FLOOD FLOOD
6 7 a a 10 111 112	1 2 3 4 5 5 7 8 7 8 7 8 7 11	No No No No No No No No	Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes	Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes	Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes	GA TX LA MI MI MA IA TX NH B	5/2/1963 5/15/1963 5/29/1953 6/2/1953 6/2/1953 6/8/1953 6/19/1953 6/11/1953 6/19/1953 7/2/1953 1/022/1953	1953 1953 1953 1953 1953 1953 1953 1953	5 5 5 6 6 6 6 6 6 7 10	Sunday Monday Tuesday Wednesday Pricay Saturday Sunday Wednesday Sunday Wednesday Descedae	DR DR DR DR DR DR DR DR DR DR DR DR	Toinado Toinado Tarnado Filood Filood Tornado Filood Filood Filood Filood Filood	TORNADO TORNADO FLOOD TORNADO FLOODS TORNADO TORNADO FLOOD FLOOD FOREST I

Source: Federal Emergency Management Agency.

Note: Year, month and day extracted from date in Federal Emergency Management Agency disaster declaration data set.

Copyright ©2016 by SAGE Publications, Inc.

This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

POWER SCRUBBING WITH OPENREFINE

OpenRefine, a free and open-source program, can help for many data cleaning challenges. (Find the program for Windows, Mac and Linux operating systems here: https://github.com/OpenRefine/OpenRefine/wiki/Installation-Instructions.) We'll use OpenRefine to perform one of the most common, frustrating and time-consuming tasks: standardizing data entries. In Chapter 5, we learned how poor data-entry controls can lead to messy data. Such is the case in this Excel file of campaign contributions released by the Missouri Ethics Commission named mo12contribs.xlsx. Open the file and note that we have 19,867 rows, including one for the headers. Each row represents a campaign contribution made by a person, company or political committee. We can take an educated guess that Column J, which has the city of the contributor, might have a lot of typos.

Start OpenRefine. In Windows, the program launches a command-line window, which you might be familiar with if you have ever done any programming. Then it opens the program inside your default browser—here Google Chrome. Note that the address that Chrome is pointed to is http://127.0.0.1:3333, which refers to your machine. In other words, the browser is working locally, not over the Internet.

0-31-16 600 F	Contraction of the state from the	weat housed the 2199 0 0 1 1999
(Pms)	retime_server1 starting Ser	ver bound to 127.0.0.1:333
9:31:16.701 [refine_server1 Initializing	context: '/' From 'C: Prog
am Files (x86)\google-	refine-2.5-r2407.webapp' (11ms	a idstausth to detect week
ata path: resorting to	environment variables (1296ms) Juacapach co decece dael-
9:31:17.999 L	refine_server] Failed to us	e jdatapath to detect user
ata path: resorting to	environment variables (2ms)	wie Befine 2 5 Te24021 Y
4ms >	For the starting dou	gie her me and traderise. A
9:31:22.649 L	refinel FOST /comman	d/core/get-importing-config
ration (4636ms) 9-21-29 664 1	refinel CET /connadd	Annua (not -all-uno last spakad
ta (15ms)	Fer mer ser reommand	scoresgee all project metau
9:31:22.682 I	refinel GET /command	/core/get-version (18ms)

Source: OpenRefine.

Note: OpenRefine command-line window. OpenRefine opens this window upon startup on Windows machines.

Google refine	- Asser but to some part must be	a.						
Crease Project Open Project	Create a project by importing data. What kinds of data files can 1 impon? TSV, CSV, 'SV, Excit, Visa ad xisa, 3504, WL, RDF as XML, and Google Data documents we will supported. Support for other formats is addet and coople fielding actions:							
Import Project	Get data hom	Locate one or more files on your complifier to uplead.						
	This Computer	Chouse Files No Ne chosen						
	Web Addressee (URLe)	Next -						
	Clapbonet:							
100								
Among K.F.Fagin.		· · · · ·						
нир Арод								

Let's create a project in Refine, so we can clean the data. Click Create Project | Get data from | This Computer and find the mo12contribs.xlsx file. Click Next and Refine takes us to the next step for loading the file.

Change the project name to "Campaign contributions" and make sure that the option to parse next line as column header at the bottom is checked. This tells Refine to use the first row from the Excel file as column headers. Note that the top of the screen previews our data.

	C 127	00	133	333				_			_	_		_		_	\$		
	Google refine	A 90	ower h	cer for a	107.0	y sitt is	nig	inth.											
	Create Project	5	- Sta	n Ove		Contingor	é Pa	rsing Optioni				Pr		nami (ampaign cont	itutional	Cri	eate Project	i,
1	Open Project	17	- 0	211.A	6 1	NECIE	104	eranittee Nar		Committee	Company		10	of Name	LastBarre	Address 1	Address 2	Cay	ſ
	Import Project	1	1	12448	1	011119	4 10	ENNETH W W	N.SQN				5	anim	Diar	3 16 N Water		Liberty.	
		2	1	12451		C11110	+ 0.	TIZENS FOR ENVIETH W W	1.5724				Ke	/ 917	Dat	107 Weight Valley Rd		Stations	
		Ĵ,	1	12452		C17179	4 0 2	TIZENS FOR	R.S.CH				0	nglana	Cum	407 Shanron Aut			
		4	1	12453		C11119	4 0.8	TIZENS FOR ENNETH W W	LICH				0	artes	Hitchoom	548.5 Commercial		2-strake	
		14	-	15.05	-	-1444	1.0		- 0					-	10.0	10.52		5	4
		Ī	Pa	rse da	ata	as			Wee	instants to	E later			2.94	en black mus		Upd	ate Prinew	ï
			Vid. Net Door Document Format streadsheets			owl men		Import	frst	line(s) at	at	12 Ste	ve blarit mile	as male					
		12				12	R mo12smail 19867		bign of the	gnie	T Sto	re lite scorce							
		12	1.00	a)						ister.	(R Parse	1		in a	names URLS	1			
		18	RDF	FROM	Mir.						next	line(s) colum	20						
	100	B	JSC	Nillion	1						4.16	heade	10						
	ymme 14 spiny	12	Line	Dase	o têx	a met					EI Discard initial	0	int.						
		18	6.51	1 121	100	enerel o	1-049	191 2195				data	-						
	Plaip Atoxit	18	DO.	Avia h	1.18	in mild i					C Load at	0							
		100	- 1		-							data.	1 m						

Source: OpenRefine.

Note: OpenRefine project preview screen.

Click Create project and Refine loads the data. By default, it shows only 10 rows at a time. You can change that if you like. Working with data in Refine typically takes two steps: detecting data problems, then correcting them. One of the most common ways to detect the flaws is by creating **Facets**, or views of your data that are very much like spreadsheet pivot tables.

C [] 127.0.0.1:3333/project?p	roject=1	\$77	/03860956	-					12 L	18
Coogle refine Campaign contributi	ONS Ferm	-							Open. Ex	port - Hel;
acet / Filter Undo / Redo a	1986	6 rc	ws						Extensions	Freebase
~	Show a	s 1	prevenue 1 - 10 next last							
Using facets and filters	▼ All		* C01_A ID	* MECID	· Committee Nam	Committee	· Company	· First Note	* Last Name	* Address
Use facets and filters to select subsets	0.0	1	112448	C111194	CITIZENIS FOR KENINETH W WESON			Shawn	Ber	118 N Wwter
filter methods from the menus at the top of each data column	sala	2	112451	C111104	CITIZENIS FOR KENINETH W WILSON			Karte	Darr	107 Wright Valley Rd
Not save how to get started? Watch these screencasts		3.	112452	C11154	CITIZENS FOR KENNETH W WESON		.+.(Ciristana.	Carry	A07 Shannon Ave
	2.47	4	112453	C131184	CITIZENS FOR KENNETH W VIR.SON		X	Cratae	HECONOM	646.5 Commercial
	210		112456	C111104	CITIZENS FOR KENINETH W WESON		C	NER	Greete	P O Box 914
	0.9	8.	112457	C111194	CTIZENS FOR KENNETH W WILSON			William	Becker	3611 NE 1326 St
	0.5	7	112458	CITTIN	CITIZENS FOR KENNETH W WILSON		Potter. Gahagan & Bologna, LLC			1 18 N Water S
	975	8	112459	E)11154	CITIZENS FOR KENNETH W WISSOR		Diyde'a Paradise Ganeral			16406 Eim St
	0.00	8	112468	CHITH	CITIZENS FOR KENNETHIN WESSA		Pence Heating and Dooling			225 KK Hwy
	16/1	10	112401	C111194	CIPZENS FOR KENNETH W			Derothy	Perkes	404 4th St

Source: Missouri Ethics Commission.

Note: OpenRefine project screen, data successfully loaded.

Create a Facet for the City column by clicking on its drop-down arrow, then selecting Facet I Text facet from the menu. Scroll down the results and we can see, for example, that the city of St. Louis (proper spelling) has been misspelled many different ways. We could click on the edit line for each misspelling and manually change the city names to our proper spelling, but that would take a whole lot of time. Fortunately, Refine has some powerful tools that will allow us to clean up these misspelled city names en masse.

That's where Clustering can help. **Clustering** uses algorithms designed to detect text values that might be the

× City	change
1605 choices Sort by: name count	Cluster
St Loui 2	
st louis 17	
ST Louis 1	
ST LOUIS 118	
St LOuis 2	
St Louis 1337	
St LOUIS 1	
St louis 10	
st Louis 2	(=
St Louos 1	-
St Lous 1	
a. i i	

Source: Missouri Ethics Commission.

Note: Text facet for City column. Note that the results are very similar to those created by an Excel pivot table.

× City	course much	const.	DI A_I	· MECID	Committee Nan	* Committee	Company	
1605 choices Sort by nam	e count Clu	ster		C000913	ST LOUIS COUNTY	Dooley for St.		
St Loui 2					CENTRAL	cours see of		
st louis 17		- 11			COMMITTEE			
ST Louis 1								
ST LOUIS III								
St LOuis 2								
St Louis 1307								
St LOUIS								
St louis 10		- 11						
st Louis 2		121						
St Louis 1	-exclu	09 (ex	Tau	st all			1	
St Lous 1		20	, 200					
~			_					
		Ap	pty	Cancel				
		Er	n faise	5kg				

Source: Missouri Ethics Commission.

Note: Manually editing a misspelling.

same. Go ahead and click the Cluster button on the Facet box for City. Refine employs two methods for clustering—key collision and nearest neighbor. Each of those methods has multiple Keying Functions. By default, Refine clusters by key collision/fingerprint. For details about how these algorithms work, see the documentation at https://github.com/ OpenRefine/OpenRefine/wiki/Clustering-In-Depth.

Using key collision/fingerprint, Refine shows us how it has clustered city names, including many St. Louis variants at the top of the list. Let's look closely: Refine says the Cluster Size is 13, or that it found 13 different values for clustering here. The Row Count is 2,457. Under Values in Cluster, we see the 13 different spellings. Merge? asks whether we want to transform these values into something else. We do, so check this box. In the New Cell Value box, enter our proper spelling of St. Louis.

Method key c	offision 🔹	Keying Fi	unction fingerprint		289 cluste
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value	+ # Choices in Cluster
	2457	St Louis 1137 over St Louis (MT ST LOUIS ST LOUIS		St Louis	2 - 13 # Rows in Cluster
7-	102	OFallon (47 mm) OFallon (25 mm) OFallon (10 mm) Ofallon (10 mm) Ofallon (2 mm) Ofallon (2 mm) Ofallon (2 mm) OFALLON (1 mm)		OFallon	4 = 17 Length Variance of Choices
6	2075	Kansas City	E	Kansas City	0 = 0.0110000000000

Source: Missouri Ethics Commission.

Note: Clustering in OpenRefine.

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value	
13	2457	St Louis (1337 cours) St Louis (952 rows) ST LOUIS (118 rows) st Jouis (17 rows) St Louis (17 rows) St Louis (2 rows) st Louis (2 rows) ST LOUIS (11 rows) ST Louis (1 rows)	N	St Louis	č,
e: Missouri Etl	nics Commissi	on.			2

Click Merge Selected & Re-Cluster and Refine standardizes all of those St. Louis misspellings. We see that we have a whole bunch more for O'Fallon, Kansas City, Lee's Summit and so on. If we were really using this column for analysis or visualization, we'd want to go through each of these clusters, then apply all of the other methods and keying functions. Close the clustering box.

Let's export our cleaned data as an Excel file by picking Export | Excel at the top right. Refine creates a file in the older (.xls) format that has the same name as our project.

This is just a quick look at how Refine can detect and correct errors. Refine can also split and merge values in columns. It even has its own scripting language called GREL, which allows you to write code for cleaning data.

Close Refine by closing your browser window, then close the window with the command line terminal. If you forget to do the latter, Refine will continue to run in the background.

EXTRACTING DATA FROM PDFS

Let's tackle one last challenge that you might face while gathering data: converting data tables inside PDF files into Excel files. Many government agencies publish data tables trapped inside these PDFs. In order to work with the tables, we must first liberate them. Open the PDF called fy2012_US_gov_net_cost.pdf. This one-page table, which documents the net operating costs of federal agencies, was extracted from the larger Financial Report of the U.S. Government for Fiscal Year 2012 that was released by the Department of the Treasury. We could try copying and pasting the data into Excel, but that would not preserve the column layout. So we'll instead need to use an online converter that will turn the PDF file into an Excel table. Cometdocs (at www.cometdocs.com) and Zamzar (at www.zamzar.com) are two proven services. Sometimes one does better than the other, depending on how the PDF is structured. So try both in real-life situations. (Note that these conversions will not work with PDFs that were created by scanning other documents. You would need to use OCR-or Optical Character Recognition-software to convert those.)

112 DATA LITERACY

Go to Cometdocs and upload the PDF. Then drag the icon for the uploaded PDF onto the Convert button and pick Excel (xls) from the options list. Enter your email address and click Convert. Now check your email for a message from Cometdocs with the download link.

Open the Excel file and you'll see Cometdocs did a good job converting the PDF. The only glitch is that some agencies, like Health and Human Services, take up more than one line. We could clean that manually in the file if we needed to.

9	(Gain)/Loss					
10	from					
11		Gross	Earned		Changes in	Net
12	(In billions of dollars)	Cost	Revenue	Subtotal	Assumptions	Cost
13	Department of Health and Human					
4	Services	924.0	67.8	856.2	0.3	856.
15	Social Security Administration	825.4	0.3	825.1	-	825
16	Department of Defense	784.7	56.0	728.7	70.4	799.
7	Department of Veterans Affairs	213.6	4.1	209.5	149.3	358
18	Interest on Treasury Securities Held by					
19	the Public	245.4	+	245.4	÷	245.
20	Department of Agriculture	161.0	12.0	149.0		149.
21	Office of Personnel Management	48.2	19.1	29.1	98.9	128.
22	Department of Labor	107.3	-	107.3	-	107.
23	Department of Transportation	. 79.0	0.8	78.2	-	78.
24	Department of Housing and Urban					
25	Development	74.5	1.5	73.0		73.
26	Department of Energy	60.8	4.3	56.5	÷	56.
27	Department of Homeland Security	58.2	9.9	48.3	0.4	48.
28	Department of Education	62.7	20.0	42.7	(+	42
29	Department of Justice	38.9	1.3	37.6		37.

Source: Department of the Treasury.

Note: Data extracted from a PDF into an Excel file by Cometdocs.

For students and professionals working with data, cleaning is an essential part of the process. This chapter showed how you can get started to solve some of the most vexing problems with Excel, OpenRefine and online PDF converters.

Now that we know how to check and clean our data, we can move on to analyzing it to discover meaning.

ON YOUR OWN

Use OpenRefine to continue cleaning the campaign contributions data file. Clean the spellings of at least 10 cities. Use all of the algorithms available under clustering. Which one worked the best? Why?