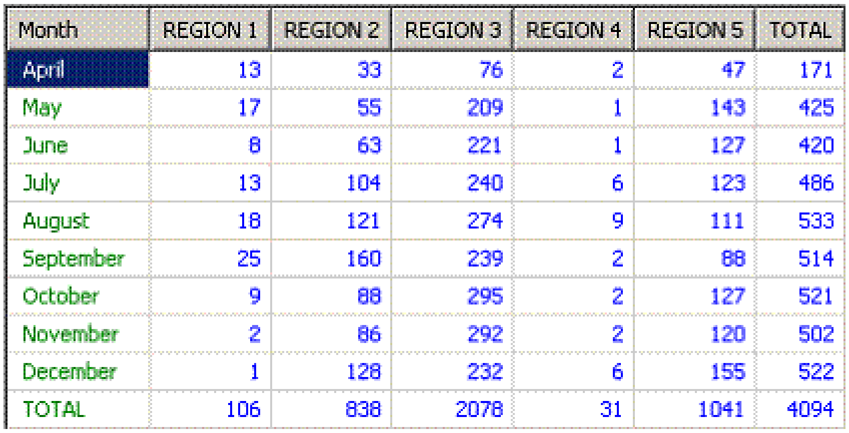
Collection

Users can Collect posts into a printable, sortable format. Collections are a good way to organize posts for quick reading. A Collection must be created to tag posts. More Help

Thread:	Crosstab Query	**Posted Date:**	September 25, 2016 9:38 PM
Post:	Crosstab Query	**Status:**	Published
Author:	Jose Zuniga	**Overall Rating:**	

1. Data

[Source: Tips for Simplifying Crosstab Query Statements, Rob Gravelle, Database Journal, 2010.]



Month	REGION 1	REGION 2	REGION 3	REGION 4	REGION 5	TOTAL
April	13	33	76	2	47	171
May	17	55	209	1	143	425
June	8	63	221	1	127	420
July	13	104	240	6	123	486
August	18	121	274	9	111	533
September	25	160	239	2	88	514
October	9	88	295	2	127	521
November	2	86	292	2	120	502
December	1	128	232	6	155	522
TOTAL	106	838	2078	31	1041	4094

2. Analysis

Compare monthly citizenship for the given regions.

Tags: None

(Post is Read)

Thread:	Crosstab Query	**Posted Date:**	October 2, 2016 10:25 PM
Post:	RE: Crosstab Query	**Status:**	Published
Author:	Aaron Grzasko	**Overall Rating:**	

Hi Jose:

This is an interesting and practical example of wide format data.

As your post mentions, the data is output from a crosstab query, which is a very common method (right up there with pivot tables) for summarizing data.

Aaron

Tags: None

(Post is Unread)

Thread: Crosstab Query
Post: RE: Crosstab Query
Author: Jose Zuniga

Posted Date: October 3, 2016 7:34 AM
Status: Published
Overall Rating:

Thank you Aaron. Glad you liked it.

Tags: None

(Post is Unread)

Thread: Tidy, same variable in several columns
Post: Tidy, same variable in several columns
Author: Marco Siqueira Campos

Posted Date: September 25, 2016 9:58 PM
Edited Date: September 26, 2016 12:06 PM
Status: Published
Overall Rating:

This is very simple and didactic example, as pointed out by Hadley Wicham, see text below. The standard to do a tidy data is:

1. Each variable forms a column.

2. Each observation form a row.

3. Each of observational unit forms a table.

In this case is about Income distribution by religious group, we have three variables: Religion, income and frequency. However, we have three columns to same variable, income, but with different values, to analyze is necessary melt or stack this columns in a single column called income.

Another very common trouble is to mix variable in rows and columns, sometimes we lose so much time to clean than data analyze.

data from: http://www.pewforum.org/religious-landscape-study/income-distribution/ (type table)

A must read text about tidy data from Hadley Wickam,

https://www.jstatsoft.org/article/view/v059i10/v59i10.pdf

Attachment: tidy.jpg (75.862 KB)

Tags: None

(Post is Read)

Thread: Tidy, same variable in several columns
Post: RE: Tidy, same variable in several columns
Author: **Marco Siqueira Campos**

Posted Date: September 25, 2016 10:04 PM
Status: Published
Overall Rating:

I was missing the main:

Analyze the Income by religion

Tags: None

(Post is Read)

Thread: Tidy, same variable in several columns
Post: RE: Tidy, same variable in several columns
Author: **Bruce Hao**

Posted Date: September 27, 2016 1:08 PM
Status: Published
Overall Rating:

Hi Marco,

This is a fascinating data set. Might I ask you how you found it?

Thanks for sharing!

Tags: None

(Post is Unread)

Thread: Tidy, same variable in several columns
Post: RE: Tidy, same variable in several columns
Author: **Kumudini Bhave**

Posted Date: September 27, 2016 10:11 PM
Status: Published
Overall Rating:

Hi Marco,

That is a pretty interesting data set and variables to compare for.

Also a good link to explore further.

Thanks!

Tags: None

(Post is Unread)

Thread:	Tidy, same variable in several columns	**Posted Date:**	September 27, 2016 10:46 PM
Post:	RE: Tidy, same variable in several columns	**Status:**	Published
Author:	**Sharon Morris**	**Overall Rating:**	

This data is very interested. I am looking forward to see the output from the assignment.

Tags: None

(Post is Unread)

Thread:	Untidy data and analysis	**Posted Date:**	September 26, 2016 9:22 AM
Post:	Untidy data and analysis	**Edited Date:**	September 26, 2016 9:27 AM
Author:	**Yifei Li**	**Status:**	Published
		Overall Rating:	

1. Data

[Source: Introduction to R. (2013). Retrieved from https://ramnathv.github.io

2. Analysis

The correlation between religious groups and income distribution.

Attachment: Screen Shot 2016-09-26 at 9.14.32 PM.png (32.969 KB)

Tags: None

(Post is Read)

Thread:	Untidy data and analysis	**Posted Date:**	October 2, 2016 11:56 PM
Post:	RE: Untidy data and analysis	**Edited Date:**	October 2, 2016 11:57 PM
Author:	**Bin Lin**	**Status:**	Published
		Overall Rating:	

Hi Yifei,

I think your dataset can be combined with some other datasets, so that we might be able to find more interesting correlations. Maybe we can look for information about distribution of education levels or job types for each religion groups. So that there will be more variables from which we can do more analysis.

Tags: None

(Post is Unread)

Thread: Gaming, Jobs and Broadband data
Post: Gaming, Jobs and Broadband data
Author: **Bruce Hao**

Posted Date: September 27, 2016 1:25 PM
Status: Published
Overall Rating:

http://www.pewinternet.org/datasets/june-10-july-12-2015-gaming-jobs-and-broadband/

Following Marco's lead, I found a data set on a topic I found interesting. The data can be downloaded at the link above. The description of the data is as follows:

This dataset contains questions about video games and gaming; job seeking and the internet; workforce automation; online dating; and home broadband, cable and smartphone use among Americans.

Tags: None

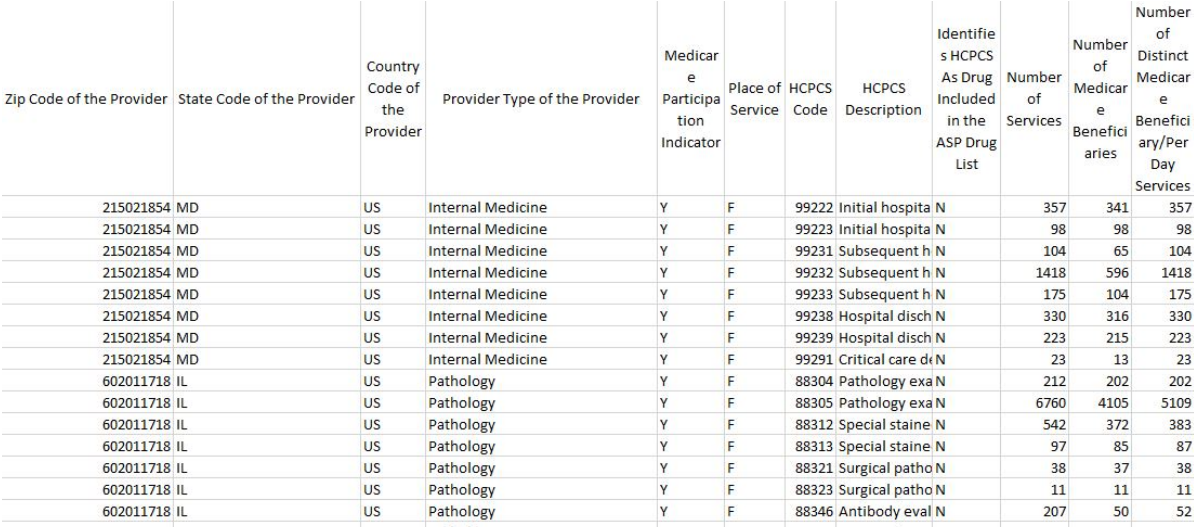
(Post is Unread)

Thread: Physician utilization
Post: Physician utilization
Author: **Christopher Estevez**

Posted Date: September 27, 2016 9:37 PM
Status: Published
Overall Rating:

1. Data

[Source: https://data.cms.gov/Public-Use-Files/Medicare-Provider-Utilization-and-Payment-Data-Phy/ee7f-sh97]



Zip Code of the Provider	State Code of the Provider	Country Code of the Provider	Provider Type of the Provider	Medicare Participation Indicator	Place of Service	HCPCS Code	HCPCS Description	Identifies HCPCS As Drug Included in the ASP Drug List	Number of Services	Number of Medicare Beneficiaries	Number of Distinct Medicare Beneficiary/Per Day Services
215021854	MD	US	Internal Medicine	Y	F	99222	Initial hospita	N	357	341	357
215021854	MD	US	Internal Medicine	Y	F	99223	Initial hospita	N	98	98	98
215021854	MD	US	Internal Medicine	Y	F	99231	Subsequent h	N	104	65	104
215021854	MD	US	Internal Medicine	Y	F	99232	Subsequent h	N	1418	596	1418
215021854	MD	US	Internal Medicine	Y	F	99233	Subsequent h	N	175	104	175
215021854	MD	US	Internal Medicine	Y	F	99238	Hospital disch	N	330	316	330
215021854	MD	US	Internal Medicine	Y	F	99239	Hospital disch	N	223	215	223
215021854	MD	US	Internal Medicine	Y	F	99291	Critical care de	N	23	13	23
602011718	IL	US	Pathology	Y	F	88304	Pathology exa	N	212	202	202
602011718	IL	US	Pathology	Y	F	88305	Pathology exa	N	6760	4105	5109
602011718	IL	US	Pathology	Y	F	88312	Special staine	N	542	372	383
602011718	IL	US	Pathology	Y	F	88313	Special staine	N	97	85	87
602011718	IL	US	Pathology	Y	F	88321	Surgical patho	N	38	37	38
602011718	IL	US	Pathology	Y	F	88323	Surgical patho	N	11	11	11
602011718	IL	US	Pathology	Y	F	88346	Antibody eval	N	207	50	52

The data consist of physician CPT utilization nationwide.

2. Analysis

Compare utilization distributions for provider within the New York area.

Tags: None

(Post is Unread)

Thread:	Physician utilization	**Posted Date:**	September 27, 2016 10:37 PM
Post:	RE: Physician utilization	**Status:**	Published
Author:	**Andrew Carson**	**Overall Rating:**	

What does CPT stand for? What made you think of this dataset? It's cool to see the difference in amount submitted vs. the amount paid. Very interesting dataset and I imagine very useful for sorting out what the "cost" of a procedure "should" be, at least to some degree.

Tags: None

(Post is Unread)

Thread:	Physician utilization	**Posted Date:**	October 2, 2016 9:33 AM
Post:	RE: Physician utilization	**Status:**	Published
Author:	**Christopher Estevez**	**Overall Rating:**	

CPT stands for Current Procedural Terminology. A CPT describes what type of visit was perform on a patient. For example, 99213 is Level 3 Established Office Visit .Differences can be because visits get rejected due to improper causes such as wrong diagnosis of patients etc. I use this data to compare my physicians to other providers in the same category nationally,locally and identify utilization for a particular provider.

Tags: None

(Post is Unread)

Thread: Untidy Data: Wide format to long format

Post: Untidy Data: Wide format to long format

Author: Kumudini Bhave

Posted Date: September 27, 2016 10:03 PM

Status: Published

Overall Rating:

Data :

Attached is the wide format data from demographic and health survey.

This is shortened to few observations with two characteristics of each birth (b2 and b4) for 3 possible births .

v012 is the mother's age, all the b2 variables are year of births, and the b4 variables are the sex of the child.

We can convert this into one observation per child born with mother caseid and age, and the year (b2) and gender of the child (b4) for each observation

Analysis :

This could be analyzed for relation between mother's age and gender of child.

Attachment: widebirthdata.png (14.693 KB)

Tags: None

(Post is Unread)

Thread: Untidy Data: Wide format to long format

Post: RE: Untidy Data: Wide format to long format

Author: Daniel Thonn

Posted Date: September 28, 2016 2:07 AM

Status: Published

Overall Rating:

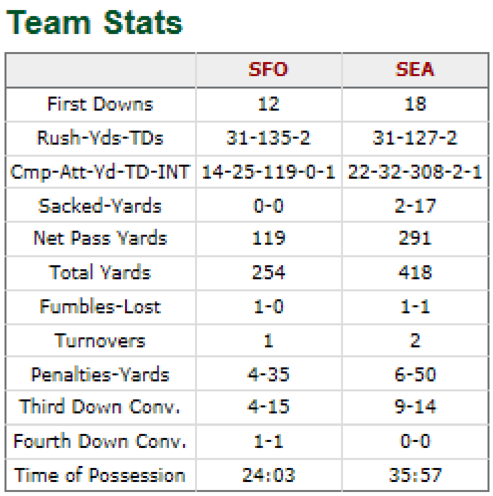
This is a good example of narrowing multiple columns, and cleaning up column names, as well as the values.

Tags: None

(Post is Unread)

Thread:	Football Data	**Posted Date:**	September 27, 2016 10:33 PM
Post:	Football Data	**Status:**	Published
Author:	Andrew Carson	**Overall Rating:**	

1. Data



Team Stats

	SFO	SEA
First Downs	12	18
Rush-Yds-TDs	31-135-2	31-127-2
Cmp-Att-Yd-TD-INT	14-25-119-0-1	22-32-308-2-1
Sacked-Yards	0-0	2-17
Net Pass Yards	119	291
Total Yards	254	418
Fumbles-Lost	1-0	1-1
Turnovers	1	2
Penalties-Yards	4-35	6-50
Third Down Conv.	4-15	9-14
Fourth Down Conv.	1-1	0-0
Time of Possession	24:03	35:57

Boxscore statistics from the Seahawks vs. 49ers game this past week.

Source: http://www.pro-football-reference.com/boxscores/201609250sea.htm

2. Analysis

Compare the yards per touchdown (both rushing and passing) for both teams. What does this mean?

Tags: None

(Post is Unread)

Thread:	NYC Citibike data	**Posted Date:**	September 27, 2016 10:43 PM
Post:	NYC Citibike data	**Edited Date:**	September 29, 2016 7:56 PM
Author:	Sharon Morris	**Status:**	Published
		Overall Rating:	

This dataset has 7 variables: when a Citibike is taken, when the bike is checked in, the location the bike is picked up, the type of user, the year of birth and gender of the user.

The data has some missing values that will have to be addressed. The data can be analyzed to create a profile of they typical Citibike users/subscriber. The year can be converted to an age and converted to a categorical variable. Start and stop times can be used to calculate the average time each bike is used.

Here are the variable names

Trip Duration (seconds)
Start Time and Date
Stop Time and Date
Start Station Name
End Station Name
Station ID
Station Lat/Long
Bike ID
User Type (Customer = 24-hour pass or 7-day pass user; Subscriber = Annual Member)
Gender (Zero=unknown; 1=male; 2=female)
Year of Birth

-

Attachment: 201606-citibike-tripdata.csv (10.231 MB)

Tags: None

(Post is Unread)

Thread:	NYC Citibike data	**Posted Date:**	September 28, 2016 10:44 PM
Post:	RE: NYC Citibike data	**Status:**	Published
Author:	**Talha Muhammad**	**Overall Rating:**	

was actually going to propose to use this dataset myself as well! very interesting

Tags: None

(Post is Unread)

Thread:	NYC Citibike data	**Posted Date:**	October 2, 2016 9:38 AM
Post:	RE: NYC Citibike data	**Status:**	Published
Author:	**Christopher Estevez**	**Overall Rating:**	

Great data set. I always wanted to analyze who are the bike riders for Citibike.

Tags: None

(Post is Unread)

Thread:	NYC Citibike data	**Posted Date:**	October 2, 2016 12:23 PM
Post:	RE: NYC Citibike data	**Status:**	Published
Author:	**Joseph Elikishvili**	**Overall Rating:**	

I really like this data set! I think There is a lot you can do with it.

Tags: None

(Post is Unread)

Thread:	NYC Citibike data	**Posted Date:**	October 2, 2016 10:09 PM
Post:	RE: NYC Citibike data	**Status:**	Published
Author:	**Ahsanul Choudhury**	**Overall Rating:**	

Very interesting dataset, would be fun .to find out the age group that uses the bikes the most

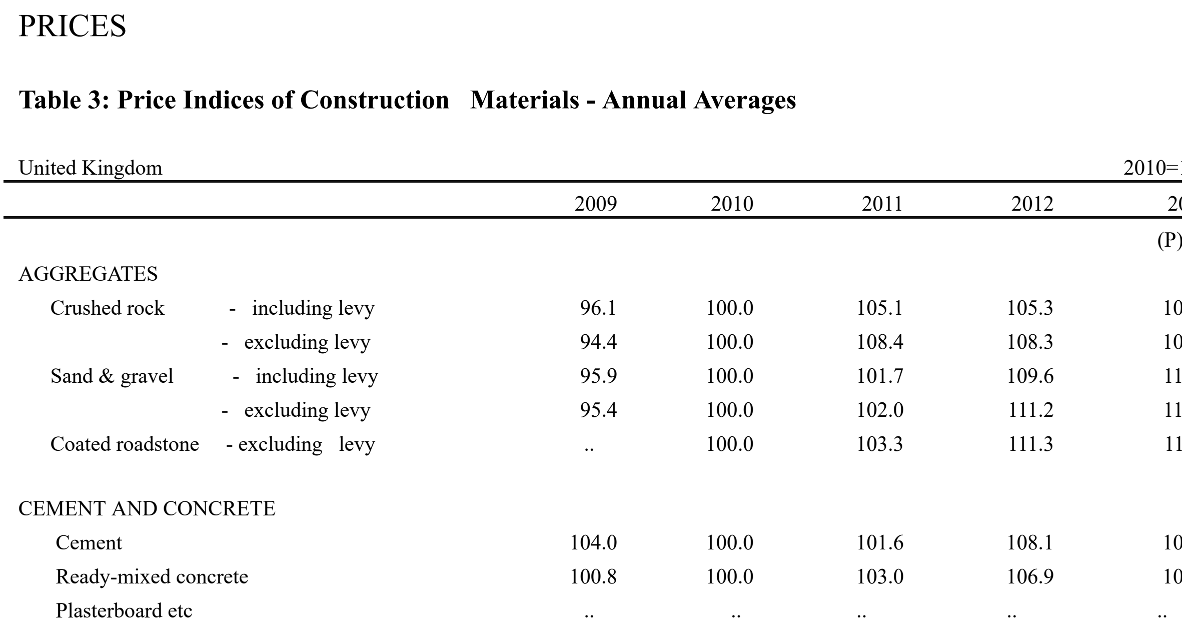
Tags: None

(Post is Unread)

Thread:	UK Construction Materials data	**Posted Date:**	September 28, 2016 1:57 AM
Post:	UK Construction Materials data	**Status:**	Published
Author:	**Daniel Thonn**	**Overall Rating:**	

1. Data: Construction_Price_Indices

Source: https://www.gov.uk/.../**14-313b-construction-building-materials-table**



PRICES

Table 3: Price Indices of Construction Materials - Annual Averages

United Kingdom 2010=

		2009	2010	2011	2012	2(
						(P)
AGGREGATES						
Crushed rock	- including levy	96.1	100.0	105.1	105.3	10
	- excluding levy	94.4	100.0	108.4	108.3	10
Sand & gravel	- including levy	95.9	100.0	101.7	109.6	11
	- excluding levy	95.4	100.0	102.0	111.2	11
Coated roadstone	- excluding levy	..	100.0	103.3	111.3	11
CEMENT AND CONCRETE						
Cement		104.0	100.0	101.6	108.1	10
Ready-mixed concrete		100.8	100.0	103.0	106.9	10
Plasterboard etc		..	..	..	..	..

Pre-cast concrete products	101.5	100.0	102.7	104.0	10
of which: Blocks, bricks, tiles & flagstones	100.2	100.0	103.3	105.6	10
Concrete re-inforcing bars	87.1	100.0	116.6	113.2	10
Fibre Cement Products	..	..	..	..	..
CLAY PRODUCTS					
All Bricks	98.9	100.0	100.2	104.7	10
Ceramic tiles	97.6	100.0	100.8	..	..
Ceramic sanitaryware	97.2	100.0	108.9	..	..
TIMBER AND JOINERY					
Imported sawn or planed wood	84.3	100.0	101.1	99.2	10
Imported plywood	83.5	100.0	104.2	110.1	11
Sawn Wood	90.1	100.0	104.7	105.7	10
Particle Board	95.3	100.0	113.5	120.2	12
Builders woodwork	96.3	100.0	103.4	107.4	10
of which					
Doors & windows	98.5	100.0	104.8	111.3	11
METAL PRODUCTS					
Fabricated structural steel	89.9	100.0	112.3	109.1	10
Doors & windows	98.8	100.0	104.1	110.8	11
Screws etc	..	..	..	..	10
Other builders' ironmongery	..	..	..	105.2	10
Central heating boilers	96.8	100.0	103.6	104.7	..
Central heating pumps	..	..	..	..	..
Taps & valves (domestic)	96.4	100.0	106.8	..	..
Metal sanitaryware	..	..	..	111.2	11
Copper pipes	..	..	..	..	..
PLASTIC PRODUCTS					
Pipes and fittings (rigid)	100.2	100.0	104.2	108.1	11
Pipes and fittings (flexible)	99.8	100.0	102.8	105.2	..
Sanitaryware	99.1	100.0	101.2	103.0	10
Doors & windows	100.0	100.0	102.5	104.3	10
Floor covering	..	..	..	..	..
OTHER BUILDING MATERIALS					
Asphalt products	93.5	100.0	..	..	..
Insulating materials (thermal or acoustic)	102.4	100.0	106.6	114.0	11
Paint (aqueous)	97.8	..	..	110.5	11
Paint (non-aqueous)	99.2	100.0	107.0	114.5	11
Lighting equipment for buildings	100.8	100.0	..	..	..
Lighting equipment for roads	102.4	100.0	100.0	103.1	10
Electric heating apparatus	..	..	..	..	..
Electric water heaters	97.2	100.0	104.0	107.4	10
Kitchen furniture	98.3	100.0	102.8	106.8	10

2. Analysis

Convert into long table with 3 columns:

1). Materials

2). Year

3). Price Indices

Tags: None

(Post is Unread)

Thread:	Pokemon	**Posted Date:**	September 28, 2016 10:19 PM
Post:	Pokemon	**Status:**	Published
Author:	**Ka Man Chan**	**Overall Rating:**	

Data:
Source: https://public.tableau.com/profile/ashlyn.opgrande#!/vizhome/PokemonStats_3/Pokemon
Analysis:
Compare Pokemons with Attack Level, Defense Level, Horse Power and Speed

Attachment: Pokemon.csv (110.623 KB)

Tags: None

(Post is Unread)

Thread: Interesting Transportation Datasets		**Posted Date:**	September 28, 2016 10:39 PM
Post: Interesting Transportation Datasets		**Status:**	Published
Author:	**Talha Muhammad**	**Overall Rating:**	

Interesting dataset available from the MTA on bus travel times across New York City. The data contains route and bus ID.

(web.mta.info/developers/MTA-Bus-Time-historical-data.html).

Another interesting dataset for taxis is:
www.nyc.gov/tlc/html/about/trip_record_data.shtml

Analysis that can be done on this data: Calculate average travel times by bus route (by time of day) for different locations and bus routes (including bus stoppage and excluding bus stoppage). Calculate the travel time reliability (95 percent travel time) and planning time (time you need to ensure you are on-time 95% of the time:

calculated as the difference between 95 percentile time and average travel time). How do travel times vary when school is in-season and out of season.

MTA Bus Time® Historical Data.pdf

mta.pdf

NYC Taxi & Limousine Commission - Trip Record Data.pdf

Tags: None

(Post is Unread)

Thread: Interesting Transportation Datasets
Post: RE: Interesting Transportation Datasets
Author: Mark Halpin

Posted Date: September 29, 2016 9:39 PM
Status: Published
Overall Rating:

Hi Talha,

Do you think that this data would be good for predictive analytics? I would be curious to see if one could use this data to predict when a bus would be late, or predict if a bus would need to take an alternative route based on traffic conditions or the previous bus's times on that route. If it could, they could funnel that information to a user app and tell the end user that they should plan an alternative route.

Tags: None

(Post is Unread)

Thread: Interesting Transportation Datasets
Post: RE: Interesting Transportation Datasets
Author: Talha Muhammad

Posted Date: October 2, 2016 2:23 PM
Status: Published
Overall Rating:

you can certainly use this data to develop predictive analytics - essentially a recommendation engine. However, the problem is not simple - since you need to consider all routes via bus between the locations they want to get to. I guess, google maps does that right now based existing or current travel times.

Tags: None

(Post is Unread)

Thread:	Chicago Crimes Data	**Posted Date:**	September 28, 2016 11:45 PM
Post:	Chicago Crimes Data	**Status:**	Published
Author:	**Mark Halpin**	**Overall Rating:**	

DATA
With Chicago being in the news very often, I thought it would be interesting to see what kind of data exists on crimes committed. The data set can be found here, with a screen shot here.

Analysis
From the data containing information on 6 million + crimes, I would analyze primary type of crime, and aggregate the information on arrests and compare arrest rate for theft vs assault.

Tags: None

(Post is Unread)

Thread:	Chicago Crimes Data	**Posted Date:**	September 29, 2016 10:54 AM
Post:	RE: Chicago Crimes Data	**Status:**	Published
Author:	**Upal Chowdhury**	**Overall Rating:**	

Mark, Nice post!!

Also i believe they are using this type of data to predict where next crime/assault would take place.

Tags: None

(Post is Unread)

Thread:	Chicago Crimes Data	**Posted Date:**	October 2, 2016 2:17 PM
Post:	RE: Chicago Crimes Data	**Status:**	Published
Author:	**Talha Muhammad**	**Overall Rating:**	

very cool and interesting!

Tags: None

(Post is Unread)

Thread:	Chicago Crimes Data	**Posted Date:**	October 2, 2016 9:15 PM
Post:	RE: Chicago Crimes Data	**Status:**	Published

Author: Scott Ogden **Overall Rating:**

Very interesting dataset! I recently attended a healthcare data analytics conference and listened to a presentation by Anne Milgram, the former attorney general for New Jersey - https://www.healthcatalyst.com/HAS-16-Anne-Milgram-Data-and-Criminal-Justice.

She spoke about integrating population health care statistics with crime and came to the conclusion that likelihood to re-offend criminally was strongly correlated with adverse health outcomes, specifically likelihood to be re-admitted for the same disease. By integrating two seeming separate fields from wildly different data sources we may just get a path forward for dealing with crime - a novel approach: treating their health. Law enforcement data is often biased and incomplete, integrations can help save lives..

Tags: None

(Post is Unread)

Thread:	Gender income inequality	**Posted Date:**	September 28, 2016 11:47 PM
Post:	Gender income inequality	**Status:**	Published
Author:	Ahsanul Choudhury	**Overall Rating:**	

1. Data

[Source: http://www.census.gov/population/age/data/files/2012/2012gender_table17.csv]

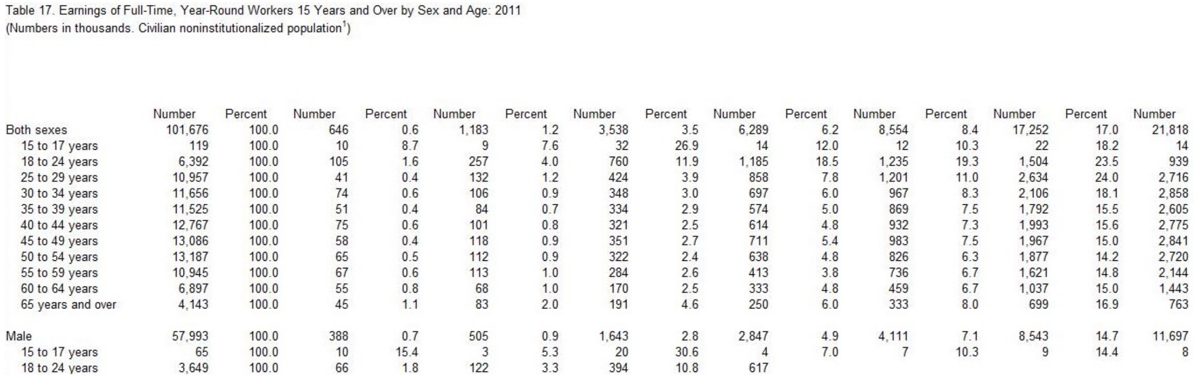


Table 17. Earnings of Full-Time, Year-Round Workers 15 Years and Over by Sex and Age: 2011
(Numbers in thousands. Civilian noninstitutionalized population[1])

	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number
Both sexes	101,676	100.0	646	0.6	1,183	1.2	3,538	3.5	6,289	6.2	8,554	8.4	17,252	17.0	21,818
15 to 17 years	119	100.0	10	8.7	9	7.6	32	26.9	14	12.0	12	10.3	22	18.2	14
18 to 24 years	6,392	100.0	105	1.6	257	4.0	760	11.9	1,185	18.5	1,235	19.3	1,504	23.5	939
25 to 29 years	10,957	100.0	41	0.4	132	1.2	424	3.9	858	7.8	1,201	11.0	2,634	24.0	2,716
30 to 34 years	11,656	100.0	74	0.6	106	0.9	348	3.0	697	6.0	967	8.3	2,106	18.1	2,858
35 to 39 years	11,525	100.0	51	0.4	84	0.7	334	2.9	574	5.0	869	7.5	1,792	15.5	2,605
40 to 44 years	12,767	100.0	75	0.6	101	0.8	321	2.5	614	4.8	932	7.3	1,993	15.6	2,775
45 to 49 years	13,086	100.0	58	0.4	118	0.9	351	2.7	711	5.4	983	7.5	1,967	15.0	2,841
50 to 54 years	13,187	100.0	65	0.5	112	0.9	322	2.4	638	4.8	826	6.3	1,877	14.2	2,720
55 to 59 years	10,945	100.0	67	0.6	113	1.0	284	2.6	413	3.8	736	6.7	1,621	14.8	2,144
60 to 64 years	6,897	100.0	55	0.8	68	1.0	170	2.5	333	4.8	459	6.7	1,037	15.0	1,443
65 years and over	4,143	100.0	45	1.1	83	2.0	191	4.6	250	6.0	333	8.0	699	16.9	763
Male	57,993	100.0	388	0.7	505	0.9	1,643	2.8	2,847	4.9	4,111	7.1	8,543	14.7	11,697
15 to 17 years	65	100.0	10	15.4	3	5.3	20	30.6	4	7.0	7	10.3	9	14.4	8
18 to 24 years	3,649	100.0	66	1.8	122	3.3	394	10.8	617						

2. Analysis

Compare income between same age group male and female.

Tags: None

(Post is Unread)

Thread:	infant and neonatal mortality and Sentiment Analysis	**Posted Date:**	September 29, 2016 10:51 AM
		Edited Date:	October 2, 2016 3:18 PM
Post:	infant and neonatal mortality and Sentiment Analysis	**Status:**	Published
		Overall Rating:	
Author:	Upal Chowdhury		

1. http://stanford.edu/~ejdemyr/r-tutorials/wide-and-long/

Above link will take to a good wide data example that can be downloaded from , www.childmortality.org ,

2. Common Crawl collects and stores billions of web page data and it can be access through AWS for free.

http://commoncrawl.org/

This data can be used for analyzing customer sentiment about certain product. Of course data needs to transform massively to make it analysis ready. Access to such data for free is amazing.

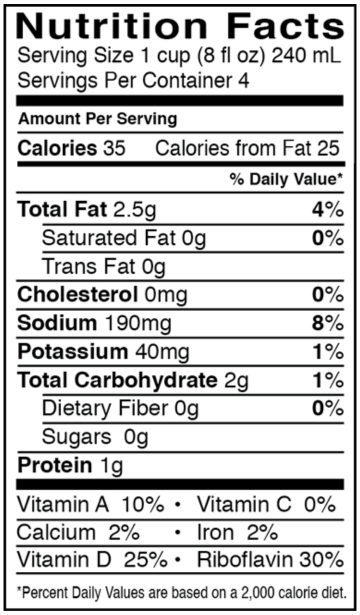
Tags: None

(Post is Unread)

Thread:	Food Nutritional Content	**Posted Date:**	September 29, 2016 1:32 PM
Post:	Food Nutritional Content	**Status:**	Published
Author:	Aaron Grzasko	**Overall Rating:**	

Data:

Food labels on common grocery items tend to follow a standard format. Below is an example for soy milk.



Nutrition Facts

Serving Size 1 cup (8 fl oz) 240 mL
Servings Per Container 4

Amount Per Serving

Calories 35 Calories from Fat 25

	% Daily Value*
Total Fat 2.5g	**4%**
Saturated Fat 0g	**0%**
Trans Fat 0g	
Cholesterol 0mg	**0%**
Sodium 190mg	**8%**
Potassium 40mg	**1%**
Total Carbohydrate 2g	**1%**
Dietary Fiber 0g	**0%**
Sugars 0g	
Protein 1g	

Vitamin A 10% •	Vitamin C 0%
Calcium 2% •	Iron 2%
Vitamin D 25% •	Riboflavin 30%

*Percent Daily Values are based on a 2,000 calorie diet.

It is hardly surprising, then, that web pages also display nutritional information in a similar fashion. Here is a link to another soy milk example from the USDA website.

The table on the USDA web page represents data for one observation--in this case, Silk brand, Vanilla soy milk.

The rows along the column *Nutrient* display relevant nutrition variables (e.g. calories, fat, protein, carbohydrates, etc.)

Analysis:

- Search the USDA website for five different types of milk (examples could include almond, cow, rice, goat, hemp), and combine into one master table.
- Manipulate the table format so that each row represents an observation for a specific type of milk. Nutritional variables should be displayed as separate fields
- Calculate the ratio of protein to total calories for each type of milk
- Calculate the average calorie content per cup across all milk types.
- Sort the observations in ascending order by total calories per cup
- Calculate net carbohydrates for each milk type (i.e. gross carbohydrates - dietary fiber)

Tags: None

(Post is Unread)

Thread:	Prescription Drug Expenditures	**Posted Date:**	September 29, 2016 4:07 PM
Post:	Prescription Drug Expenditures	**Status:**	Published
Author:	**Judd Anderman**	**Overall Rating:**	

Data: US National Health Expenditures by Service Type and Funding Source, 1960-2014

Source: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/NHE2014.zip

Analysis: Using the population and spending data in the NHE table, analyze the change in total nominal spending per capita on prescription drugs in the US from 1960 to 2014, and describe and plot the changes in the relative proportions of prescription drug expenditures across funding sources over the same interval.

Tags: None

(Post is Unread)

Thread:	Patient Experience	**Posted Date:**	September 29, 2016 4:49 PM
Post:	Patient Experience	**Status:**	Published
Author:	**Scott Ogden**	**Overall Rating:**	

https://data.medicare.gov/Hospital-Compare/Hospital-Value-Based-Purchasing-HVBP-Patient-Exper/avtz-f2ge

In Health care one of the biggest indicators of patient experience are measured by the HCAHPs satisfaction survey. CMS also uses these data to adjust Medicare reimbursements to over and under performing hospitals. The attached, wide, dataset shows hospital-specific points received for each question (in the columns). Tidying the data to long form would be an un-pivoting of sorts; shifting the the measure name to a column and the outcome in the column next to it.

For analysis you could determine which questions are most strongly associated with the aggregate question "How likely are you to recommend the hospital?" These sorts of analyses give operations leaders an ability to focus patient experience projects in areas that will help the most.

Additionally, on a higher level, you could find out which states' hospitals underperform in hopes of designing state based solutions to patient experience problems.

Tags: None

(Post is Unread)

Thread:	Top 10 US Auto Seller Leaders	**Posted Date:**	September 29, 2016 5:34 PM
Post:	Top 10 US Auto Seller Leaders	**Status:**	Published
Author:	**Joseph Elikishvili**	**Overall Rating:**	

Source: http://www.cleanmpg.com/community/index.php?threads/53096/

The table shows the rankings of top 10 selling models for March 16, Feb 16 and Jan 15 and sales numbers as of March 15 and March 16. We can take a look at most popular manufacturer for the specific time also we can take a look at the most volatile models in regards to the ranking, also find the average ranking across all dates.

Attachment: Top 10 Mar 2016 US Sales.jpg (59.543 KB)

Tags: None

(Post is Unread)

Thread:	Lending Club Data	**Posted Date:**	September 29, 2016 10:56 PM
Post:	Lending Club Data	**Status:**	Published
Author:	**Bin Lin**	**Overall Rating:**	

https://www.lendingclub.com/info/download-data.action

Just follow the link, we can download the data for the loans that were rejected and loans that is issued (Most current is 2016Q2).

We can analyses the relationship among amount requested loan amount, debt to income ratio and employment length. Or we can compare income with interest rates and the rating of loans etc.

Tags: None

(Post is Unread)

Thread:	Lending Club Data	**Posted Date:**	September 29, 2016 11:51 PM
Post:	RE: Lending Club Data	**Status:**	Published
Author:	**Dhananjay Kumar**	**Overall Rating:**	

Hi Bin Lin,

I have a demographic dataset of NYC based on zip codes and I see that your dataset has zip code as well. If we combine these two datasets, we can retrieve interesting insights out of it like which zip code in NYC has the most probability of loan rejection and then we can look at the demography of that zip code.

Tags: None

(Post is Unread)

Thread:	NYC Demographics	**Posted Date:**	September 29, 2016 11:44 PM
Post:	NYC Demographics	**Status:**	Published
Author:	**Dhananjay Kumar**	**Overall Rating:**	

I found a nice dataset of Demographics of NYC by Zip code. One can use this dataset for targeted Marketing Campaign based on Demography. For example, a manufacturer of Sporting Goods or a Sport Store can choose a county for their campaign based on residents age. Another example would be to do targeted campaign on the basis of second/native language.

Attachment: Demographic_Statistics_By_Zip_Code.csv (26.709 KB)

Tags: None

(Post is Unread)

Thread:	San Francisco salaries analysis	**Posted Date:**	September 30, 2016 11:26 AM
Post:	San Francisco salaries analysis	**Status:**	Published
Author:		**Overall Rating:**	

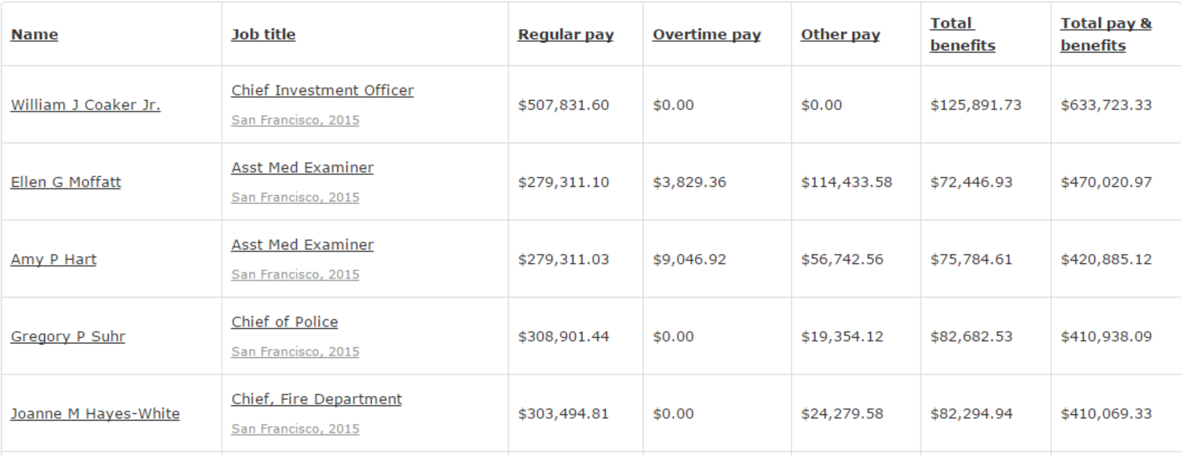
Shyam Balagurumurthy-Viswanathan

Below mentioned website has a good dataset about salaries/benefits of different job profiles.

Link for 2015: http://transparentcalifornia.com/salaries/2015/san-francisco/

We can filter out the necessary data and change for different years. Or we can fetch all the the data and perform some transformations and analysis. Below are some of it.

1. Categorize the data via job title and profile. Split job profile and designation in Job Title. We can also copy HTML data into excel and cleanup the data.
2. Salary changes over time between different groups
3. Base pay, overtime pay, and benefits allocated between different groups.



Name	Job title	Regular pay	Overtime pay	Other pay	Total benefits	Total pay & benefits
William J Coaker Jr.	Chief Investment Officer San Francisco, 2015	$507,831.60	$0.00	$0.00	$125,891.73	$633,723.33
Ellen G Moffatt	Asst Med Examiner San Francisco, 2015	$279,311.10	$3,829.36	$114,433.58	$72,446.93	$470,020.97
Amy P Hart	Asst Med Examiner San Francisco, 2015	$279,311.03	$9,046.92	$56,742.56	$75,784.61	$420,885.12
Gregory P Suhr	Chief of Police San Francisco, 2015	$308,901.44	$0.00	$19,354.12	$82,682.53	$410,938.09
Joanne M Hayes-White	Chief, Fire Department San Francisco, 2015	$303,494.81	$0.00	$24,279.58	$82,294.94	$410,069.33

Tags: None

(Post is Unread)

Thread:	San Francisco salaries analysis	**Posted Date:**	October 3, 2016 12:13 PM
Post:	RE: World food dataset	**Edited Date:**	October 3, 2016 12:14 PM
Author:		**Status:**	Published
	Shyam Balagurumurthy-Viswanathan	**Overall Rating:**	

I am adding one challenging/interesting data set from Kaggle.

Link: https://www.kaggle.com/openfoodfacts/world-food-facts

The columns in FoodFacts are as follows:

- **code** (text)
- **url** (text)
- **creator** (text)
- **created_t** (text)
- **created_datetime** (text)
- **last_modified_t** (text)
- **last_modified_datetime** (text)
- **product_name** (text)
- **generic_name** (text)
- **quantity** (text)
- **packaging** (text)
- **packaging_tags** (text)
- **brands** (text)
- **brands_tags** (text)
- **categories** (text)
- **categories_tags** (text)
- **categories_en** (text)
- **origins** (text)
- **origins_tags** (text)
- **manufacturing_places** (text)
- **manufacturing_places_tags** (text)
- **labels** (text)
- **labels_tags** (text)
- **labels_en** (text)
- **emb_codes** (text)
- **emb_codes_tags** (text)
- **first_packaging_code_geo** (text)
- **cities** (text)
- **cities_tags** (text)
- **purchase_places** (text)
- **stores** (text)

Tags: None

(Post is Unread)

← OK