Extracting Data in PDFs using R

Nathaniel Bechhofer

Some Background on R

- R is a programming language designed for statistical computing
- Free, open source, large community
  - RStudio, a powerful IDE
  - CRAN (The Comprehensive R Archive Network)

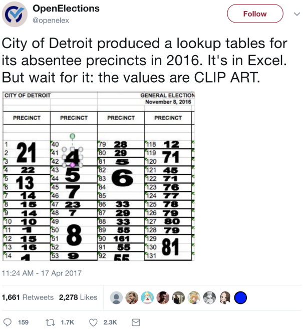
Examples of R packages

- `dplyr`, for a readable grammar of data manipulation
- `foreign`, for reading almost any kind of data file (including Stata files)
- `stringr`, for operating on strings
- `lubridate`, for parsing and processing dates
- `forecast`, for time series analysis
- `glmnet`, for regularized GLMs

along with over 10,000 other packages on CRAN.

Using the tabulizer package to extract data from PDFs

- Many institutions (especially government agencies) publish data that may be useful for research
- Generally, ease of processing for end-users is not a priority



OpenElections
@openelex

Follow

City of Detroit produced a lookup tables for its absentee precincts in 2016. It's in Excel. But wait for it: the values are CLIP ART.

CITY OF DETROIT

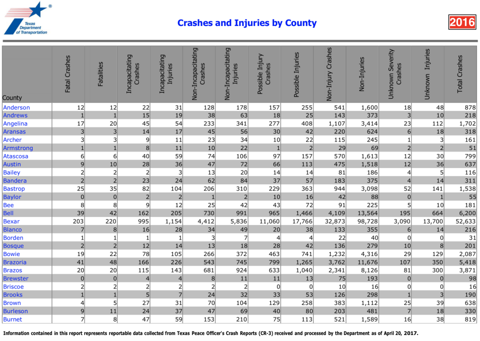
GENERAL ELECTION
November 8, 2016

PRECINCT		PRECINCT		PRECINCT		PRECINCT	
1	21	40		79	28	118	12
2		41	4	80	29	119	71
3		42		81	5	120	
4	22	43	5	82	6	121	45
5	13	44		83		122	71
6		45	7	84		123	76
7	14	46		85		124	77
8	15	47	23	86	33	125	78
9	14	48	7	87	29	126	79
10	10	49	8	88	33	127	80
11	1	50		89	55	128	79
12	15	51		90	161	129	81
13	16	52		91	55	130	
14	1	53	9	92	55	131	

11:24 AM - 17 Apr 2017

1,661 Retweets 2,278 Likes

159 1.7K 2.3K



Texas Department of Transportation

Crashes and Injuries by County

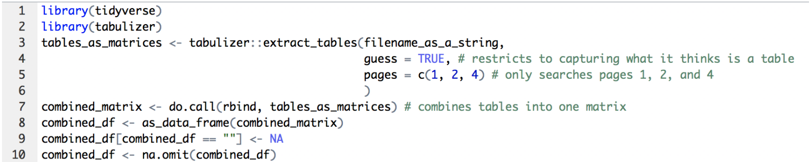
2016

County	Fatal Crashes	Fatalities	Incapacitating Crashes	Incapacitating Injuries	Non-Incapacitating Crashes	Non-Incapacitating Injuries	Possible Injury Crashes	Possible Injuries	Non-Injury Crashes	Non-Injuries	Unknown Severity Crashes	Unknown Injuries	Total Crashes
Anderson	12	12	22	31	128	178	157	255	541	1,600	18	48	878
Andrews	1	1	15	19	38	63	18	25	143	373	3	10	218
Angelina	17	20	45	54	233	341	277	408	1,107	3,414	23	112	1,702
Aransas	3	3	14	17	45	56	30	42	220	624	6	18	318
Archer	3	3	9	11	23	34	10	22	115	245	1	3	161
Armstrong	1	1	8	11	10	22	1	2	29	69	2	2	51
Atascosa	6	6	40	59	74	106	97	157	570	1,613	12	30	799
Austin	9	10	28	36	47	72	66	113	475	1,518	12	36	637
Bailey	2	2	2	3	13	20	14	14	81	186	4	5	116
Bandera	2	2	23	24	62	84	37	57	183	375	4	14	311
Bastrop	25	35	82	104	206	310	229	363	944	3,098	52	141	1,538
Baylor	0	0	2	2	1	2	10	16	42	88	0	1	55
Bee	8	8	9	12	25	42	43	72	91	225	5	10	181
Bell	39	42	162	205	730	991	965	1,466	4,109	13,564	195	664	6,200
Bexar	203	220	995	1,154	4,412	5,836	11,060	17,766	32,873	98,728	3,090	13,700	52,633
Blanco	7	8	16	28	34	49	20	38	133	355	6	14	216
Borden	1	1	1	1	3	7	4	4	22	40	0	0	31
Bosque	2	2	12	14	13	18	28	42	136	279	10	8	201
Bowie	19	22	78	105	266	372	463	741	1,232	4,316	29	129	2,087
Brazoria	41	48	166	226	543	745	799	1,265	3,762	11,676	107	350	5,418
Brazos	20	20	115	143	681	924	633	1,040	2,341	8,126	81	300	3,871
Brewster	0	0	4	4	8	11	11	13	75	193	0	0	98
Briscoe	2	2	2	2	2	2	0	0	10	16	0	0	16
Brooks	1	1	5	7	24	32	33	53	126	298	1	3	190
Brown	4	5	27	31	70	104	129	258	383	1,112	25	39	638
Burleson	9	11	24	37	47	69	40	80	203	481	7	18	330
Burnet	7	8	47	59	153	210	75	113	521	1,589	16	38	819

Information contained in this report represents reportable data collected from Texas Peace Officer's Crash Reports (CR-3) received and processed by the Department as of April 20, 2017.

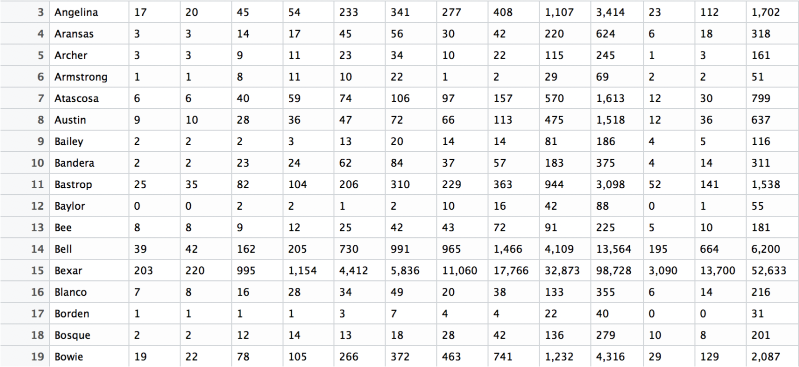
Using the tabulizer package to extract data from PDFs

The tabulizer package for R allows you to solve some of these problems. (The Natural Resource Governance Institute has an online version with much of the functionality.)
Use case: you have data stuck in PDF tables



```
library(tidyverse)
library(tabulizer)
tables_as_matrices <- tabulizer::extract_tables(filename_as_a_string,
                                                guess = TRUE, # restricts to capturing what it thinks is a table
                                                pages = c(1, 2, 4) # only searches pages 1, 2, and 4
                                                )
combined_matrix <- do.call(rbind, tables_as_matrices) # combines tables into one matrix
combined_df <- as_data_frame(combined_matrix)
combined_df[combined_df == ""] <- NA
combined_df <- na.omit(combined_df)
```

The result



3	Angelina	17	20	45	54	233	341	277	408	1,107	3,414	23	112	1,702
4	Aransas	3	3	14	17	45	56	30	42	220	624	6	18	318
5	Archer	3	3	9	11	23	34	10	22	115	245	1	3	161
6	Armstrong	1	1	8	11	10	22	1	2	29	69	2	2	51
7	Atascosa	6	6	40	59	74	106	97	157	570	1,613	12	30	799
8	Austin	9	10	28	36	47	72	66	113	475	1,518	12	36	637
9	Bailey	2	2	2	3	13	20	14	14	81	186	4	5	116
10	Bandera	2	2	23	24	62	84	37	57	183	375	4	14	311
11	Bastrop	25	35	82	104	206	310	229	363	944	3,098	52	141	1,538
12	Baylor	0	0	2	2	1	2	10	16	42	88	0	1	55
13	Bee	8	8	9	12	25	42	43	72	91	225	5	10	181
14	Bell	39	42	162	205	730	991	965	1,466	4,109	13,564	195	664	6,200
15	Bexar	203	220	995	1,154	4,412	5,836	11,060	17,766	32,873	98,728	3,090	13,700	52,633
16	Blanco	7	8	16	28	34	49	20	38	133	355	6	14	216
17	Borden	1	1	1	1	3	7	4	4	22	40	0	0	31
18	Bosque	2	2	12	14	13	18	28	42	136	279	10	8	201
19	Bowie	19	22	78	105	266	372	463	741	1,232	4,316	29	129	2,087