Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com



QUANTIFYING THE FUTURE



When Data Isn't Enough

Kellie Scarbrough, ICEAA 2019, Tampa May 2019

Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Agenda

- Bad Data
- Government Data 2.0
- Pre-Processing and Exploring
- Text Mining in R
- Visualizing Data



What makes data "bad"?

BAD DATA



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Bad'' = Inaccurate

Inaccurate "big data" sets can still have predictive power



Analysts already have the skills to clean messy data:

- Normalization
- Removal of duplicates
- Correction of errors and blanks
- Correction of spelling
- Standardization
- Transformation



How government data has evolved

GOVERNMENT DATA 2.0



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com $Big \ Data \ 1.0 \ and \ 2.0$

1.0: The original adoption of internet technologies and electronic commerce

Businesses establish their online presence

2.0: The establishment of infrastructure and processes to collect large volumes of data

Businesses shift to data-analytic thinking

Foster, Provost and Tom Fawcett. Data Science for Business: What You Need to Know About Data Mining and Analytics Thinking. O'Reilly, 2013.



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Gov. Data 1.0 and 2.0

1.0: The adoption of systems with streamlined data collection and reporting

Government establishes access to data

2.0: Government adopts best practices centered around data analytics

Government uses data to improve the estimating, budgeting, and execution of its programs

Foster, Provost and Tom Fawcett. Data Science for Business: What You Need to Know About Data Mining and Analytics Thinking. O'Reilly, 2013.



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Government Big Data Initiatives

- California Natural Resources Agency's shared services data lake
- Joint Improvised Thread Defense Operations operations research and process improvement
- Department of Homeland Security integration of various departments
- GSA's Data Center Optimization Initiative (DCOI)
- DOE's Scalable Data Management, Analysis, and Visualization (SDAV) institute
- USGS Big Data for Earth System Science



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Going Rogue

Data science is a broad field that can be described as¹:

data science = { statistics \cap informatics \cap computing \cap communication \cap sociology \cap management | data \cap domain \cap thinking }

Analysts who do not adapt to big data and the evolution of data analytics will produce inferior estimates and models.

1. Longbing, Cao, "Data science: challenges and directions." In: *Communications of the ACM* 60.8 (Aug. 2017), pp. 59-68.





PRE-PROCESSING AND EXPLORING

What techniques can analysts use?

Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Visual Aids and R

- Visualization can be used during preprocessing to:
 - Identify inaccurate values
 - Find missing values
 - Find duplicate values
 - Establish candidates for bins
 - Identify places to consolidate data
 - Find outliers
 - Assess relationships



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com R and Resources

- R is a command line interface and was chosen because it is free, available to any analyst
- R is a powerful statistical computing environment
- Requires learning the R programming language
- Can be used with GUI like RStudio to facilitate use, reduce learning curve
- Code for the examples shown here can be found in the corresponding paper for this presentation and in many online forums – you can easily adapt it for your own needs



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Inspecting Data

When reading in files to R, be sure to set argument stringsAsFactors equal to FALSE:

orders.df <- read.csv("C:/Users/.../orders.csv", stringsAsFactors = False)

head(orders.df)
str(orders.df)
summary(orders.df)

head() returns the first six rows of data

str(orders.df)

data.frame': 275932 obs. of	68 variables:
<pre>\$ ïNumber.of.Records : int</pre>	1111111111
<pre>\$ Statusgroup. : chr</pre>	"active" "active" "active"
\$ Billing.Id : int	42 9090 15702 9090 8918 8 44 44 20 8
<pre>\$ braintree_customer_id : chr</pre>	
<pre>\$ business_name : chr</pre>	"ClearSight Studio" "ClearSight Studio" "ClearSight Studi
\$ Charge.Total : num	10.19 19.63 12.12 39.62 7.12
<pre>\$ contact_me : chr</pre>	"False" "False" "False"
\$ Cost : num	1.65 7.7 1.65 23.1 1.9 0 0 0 0 0
\$ Coupon.Id : int	9 na na na 15 na 9 9 na na
\$ Coupon.Total : num	3.16 0 0 0 5 0 3.14 3.16 0 0



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Exploring Data

There are several functions available to generate correlation matrices and visualize the results

												Coupon.To	Cost	group_id	Margin	Shipping.7	Tax	Discount.T	Charge.Tot	Line.tem.T
										С	oupon.Total									
											Cost	•								
											group_id					•				
e Edit Cod			m Build Debu								Margin									
le Edit Cod			io to file/function	g Profile Too						Sh	Margin ipping.Total		•	•		•				
e Edit Codi - ≪3 ≪* ≣BULcomplet× (⇔⇔ @	• 🔓 🔐 Orders,Co 9 Riter	arr Matrix.R*	io to file, function	Addima 🗄 - Addima	•	Line Hern Total	Maroin ²	Shipping Total	Q. Tar	Sh	ipping.Total			•	•	•				
e Edit Codi - %]@? EULcomplet × ©© @	• 🔓 🔐 Orders,Co 9 Riter	Coet	So to Fle,/function		ęroup_id ÷	Line. Item. Total		Shipping Total	Q, Taz ÷ 0.003165548	Sh			•	•	•	•				
e Edit Cod - Si Con EUI.complet × Cod Col Cod	 Gridert_Co Pilter Charge.Total 	Coet	So to Fle,/function	Discount.Total	group.id 0.004601855		-0.004205555	-0.001465625			ipping.Total Tax				•	•	•			
e Edit Cod - Silicomplet × - Cod Charge.Total Cost	Griders, Co Griders, Co Filter Charge. Total 1.0000000000	Cost	Coupen.Total 0 4 0.0004735722 0.0576122804	Discount.Total	group_id	0.9999880150	-0.004205056 -0.161508369	-0.001465625	0.003165548 0.042518460		ipping.Total				•		•			
Edit Codi Ulcampiet * Charge Total Coet Coupon Total		Coet 0.01544018 0.05701229	Coupen.Total [©] -0.0004753722 0.0576122864 1.400000000	Diecourt.Total	• group.jel 0.004601855 0.1228/73461 0.01188/943	0.9999880150	-0.004205856 -0.161508369 -0.012909967	-0.001465625 -0.195743489	0.005165548 0.042518460 0.098751824		ipping.Total Tax						•	•	•	
le Edit Codi - Si Complet X Charge Total Coet Coupon Total	Gridert, Co Gridert, Co Filter Acrossococcocc Coccocccocc Coccoccccocc	Cost Cost Cost Cost Cost Cost Cost Cost	Coupon.Total * Coupon.Total * -0.0004753722 0.0576122884 1.40000000 0.0006042544	Diecourt.Total	group.jd 0.004001055 0.1338/73461 0.01189(943 0.005006389	0.9999850150 0.0166925812 0.0007954746	-0.004205858 -0.161508389 -0.012909867 -0.005456771	-0.001405825 -0.195742489 -0.015623796 -0.001622481	0.005165548 0.042518460 0.098751824	Di	ipping.Total Tax scount.Total				•		•		•	
e Edit Cod FUIcomplet & Charge. Total Cost Coupon. Total Discount. Total group. Jd	Grider, Ci Filter Accordence, Ci Filter Accordence, Ci Filter Accordence, Ci Constance,	Cost Cost Cost Cost Cost Cost Cost Cost	Coupon.Total * Coupon.Total * -0.0004753722 0.0576122884 1.40000000 0.0006042544	Discount.Total	grsup.jd 0.004001855 0.122872461 0.011886943 0.005006899 1.00000000	0.9999850150 0.0169935812 0.0007954746 0.9991821046	-0.004205836 -0.161508369 -0.012909867 -0.005456771 -0.035061065	-0.001465625 -0.195743489 -0.015623796 -0.001622481 -0.005226297	0.003165548 0.042518480 0.098751624 0.098366198	Di	ipping.Total Tax				•		•	•	•	
e Edit Cod FUIcomplet & Charge. Total Cost Coupon. Total Discount. Total group. Jd	Grider, Ci G	Cost Cost Cost Cost Cost Cost Cost Cost	Coupen.Total Cuerchart	Diecount.Total 0.000019150114 0.01154.01909 0.0000042544 1.000000000 0.00000002544 1.0000000000 0.00000000000000 0.00000000	group.jd c.004001855 d.128673461 d.011880943 d.005006389 1.00000000 d.004821410 -0.855061063	0.999888130 0.0168835812 0.0007954746 0.9991821846 0.0048214182 1.0008330800 -0.0048255439	-0.004225856 -0.161508389 -0.012909967 -0.005456771 -0.033081065 -0.004225544 1.000000099	-0.001405625 -0.195743469 -0.015623796 -0.001622481 -0.005228287 -0.0023286297 -0.0023616679 -0.009038281	0.003165548 0.042518460 0.038751824 0.033368198 -0.020175005 0.03891521 -0.097769655	Di	ipping.Total Tax scount.Total Charge.Total						•	•••	•	
EULcomplet * EULcomplet * Charge.Total Cost Coupon.Total Discount.Total group.jd Line.Hem.Sotal	Grder.co Filter Goder.co Filter Lococococo October.co October.co	Cost Cost	Coupen.Total Coupen.Total Coupen.Total Coupen.Total Coupen.Total Coupen.Total Coupen.Total Coupen.Total Coupen.Total Couper.Total Coupe	Discount.Total	group.jd c.004001855 d.128673461 d.011880943 d.005006389 1.00000000 d.004821410 -0.855061063	0.9999850150 0.0169935812 0.0007954746 0.9991821046 0.0049214102 1.0000330000	-0.004225856 -0.161508389 -0.012909967 -0.005456771 -0.033081065 -0.004225544 1.000000099	-0.001405625 -0.195743469 -0.015623796 -0.001622481 -0.005228287 -0.0023286297 -0.0023616679 -0.009038281	0.003165548 0.042318480 0.098751824 0.098368198 -0.020179505 0.09894521	Di	ipping.Total Tax scount.Total							••••	••••	

(a) Correlation matrix from EUI.corrplot (b) EUI corrplot() result



eta

 \overline{a}

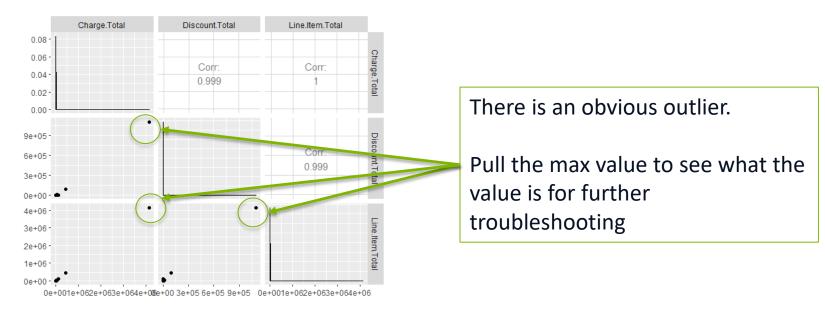
al da

- ta

0.8 0.6 0.4 0.2 -0.2 -0.4 -0.6

Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Explore with ggpairs()

#explore potential drivers of total charge ggpairs(orders.df[,c(6,20,38)])



#Find max value for troubleshooting
max(orders.df\$Charge.Total, na.rm = FALSE, dims = 1, n = NULL)

Result: \$4.1M – We know this is an error!



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Error Correction: Apply Filter

```
orders.df <-orders.df %>% filter(Status..Orders. == "placed" |
    Status..Orders. == "manually_placed" |
    Status..Orders. == "shipped" |
    Status..Orders. == "shipping")
```

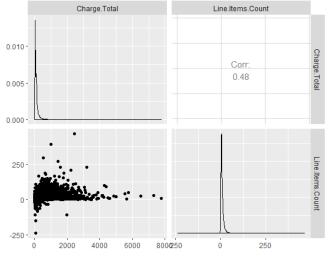
summary(orders.df\$Status..Orders.)

NSULTING

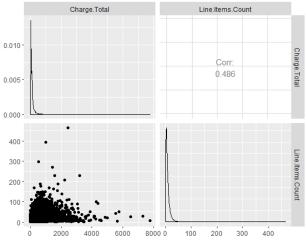


Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Another Example: Negative Values

#explore line item count and total charge
ggpairs(orders.df[,c(6,39)])



#remove negative line item counts and re-run
orders_licorr.df <- orders.df %>% filter(Line.Items.Count > 0)
ggpairs(orders_licorr.df[,c(6,39)])





How to extract value from text fields

TEXT MINING



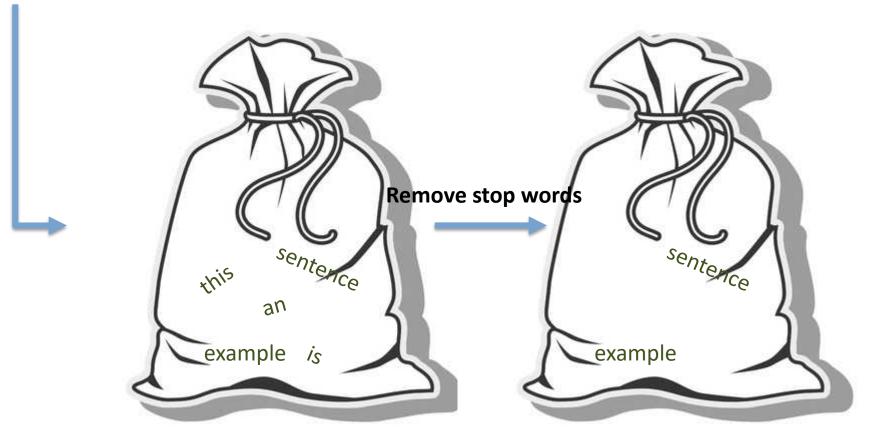
Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Text mining Overview

- Text fields are common in government data
- We will use "bag-of-words" text mining
 - Each word is treated as an individual item
 - Disregards sentence structure
 - Requires removal of stop words



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Bag-of-Words

This sentence is an example.





Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com The Code: Create a Corpus

The library Rweka contains a function which allows you to view words in groups

```
library(RWeka)
tokenizer <- function(x) {
NGramTokenizer(x, Weka_control(min = 1, max = 2))
}</pre>
```

The primary tools for text mining are vectors and corpora

```
#Create vector
parts_source <- VectorSource(parts_desc_clean)
#Create corpus
parts_corpus <- VCorpus(parts_source)</pre>
```



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Clean the Corpus

```
# Alter the function code to match the instructions
clean_corpus <- function(corpus) {</pre>
# Remove punctuation
corpus <- tm_map(corpus, removePunctuation)</pre>
# Transform to lower case
corpus <- tm_map(corpus, content_transformer(tolower))</pre>
# Remove stop words using common English stop words
corpus <- tm_map(corpus, removeWords, c(stopwords("en")))</pre>
# Strip whitespace
corpus <- tm_map(corpus, stripWhitespace)</pre>
return(corpus)
}
parts_clean <- clean_corpus(parts_corpus)</pre>
parts_tdm <- TermDocumentMatrix(parts_clean)</pre>
parts_m <- as.matrix(parts_tdm)</pre>
term_frequency <- rowSums(parts_m)</pre>
term_frequency<- sort(term_frequency, decreasing = TRUE)</pre>
```



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com The Code: View the Results

```
# View the top 10 most common words
term_frequency[1:10]
```

```
# Plot a barchart of the 10 most common words
barplot(term_frequency[1:10], col ="blue", las=2)
```

```
# Load wordcloud package
library(wordcloud)
```

```
terms_vec <- names(term_frequency)</pre>
```

wordcloud(terms_vec, term_frequency, max.words = 50, colors = "red")

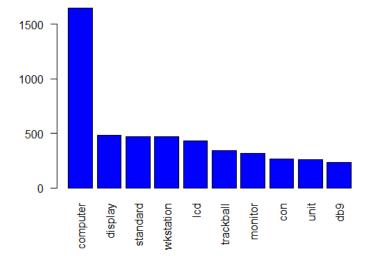
```
wordcloud(terms_vec, term_frequency, max.words =50,
colors = c("grey80","darkgoldenrod1","tomato"))
```



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Text Mining Results

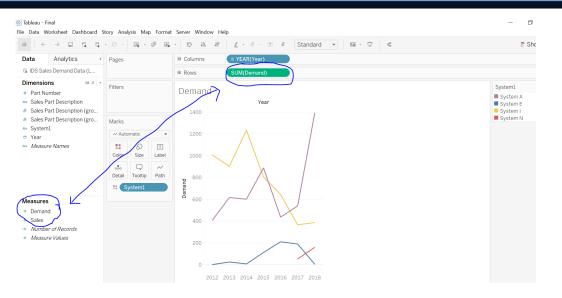
> # View t > term_fre			on words	-					
computer	display	standard	wkstation	lcd tr	ackball	monitor	con	unit	db9
1648	486	469	469	435	347	316	266	260	234
>									







Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Using Tableau for Exploration



In Tableau, we can quickly interchange variables and break out data by various dimensions

We can also quickly group data and filter to drill down and extract meaningful insights





How to Communicate Visually **VISUALIZING RESULTS**



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Visualizing the Results

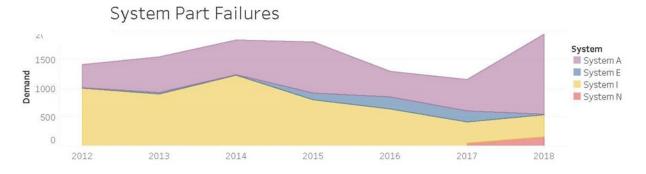
If you can't communicate and defend your analysis, your efforts are wasted.

Provide results that are:

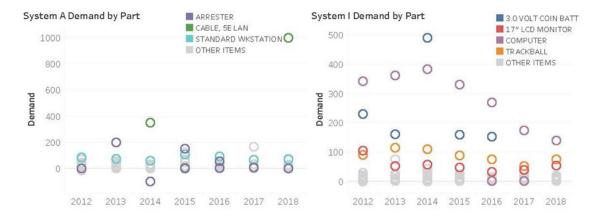
- Clear
- Meaningful
- Actionable



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Logistics Dashboard Example



Insurance ps2 In Most demand for parts come from Systems A and I. The highest demand for parts are computers for System I, indicating complete system failures are occurring. Failures spiked in 2014 and have been declining steadily since. Steady failures of 17" monitors, 3.0 volt coin batteries, and trackballs are also occurring. System A experiences a moderate but steady level of workstation replacement, and periodic spikes in SE LAN cable and arrester part demands.





Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com Summary

- Bad vs messy
- Role of data science in data analytics
- Explored some techniques for inspecting large data sets
- Explored text mining
- Visualizing the results



Presented at the 2019 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

QUESTIONS?

