

Exploratory Analysis Subgroup

Arizona BioSense Workgroup

March 28, 2019



ARIZONA DEPARTMENT
OF HEALTH SERVICES

Health and Wellness for all Arizonans

Objectives

- Welcome and Attendance
- Updates
- Syndromic Surveillance Exploratory Analysis Toolkit: What is Exploratory Analysis?
- Discussion
- Upcoming Meetings

ESSENCE Updates

- No new syndrome definitions this month
- **Are there any syndrome definitions you are working on?**
- ESSENCE Q&A v4:
<https://vimeo.com/325621569>
 - New myESSENCE features coming soon
 - Talk about Reports Manager

Syndromic Exploratory Analysis Toolkit:

What is Exploratory Analysis?

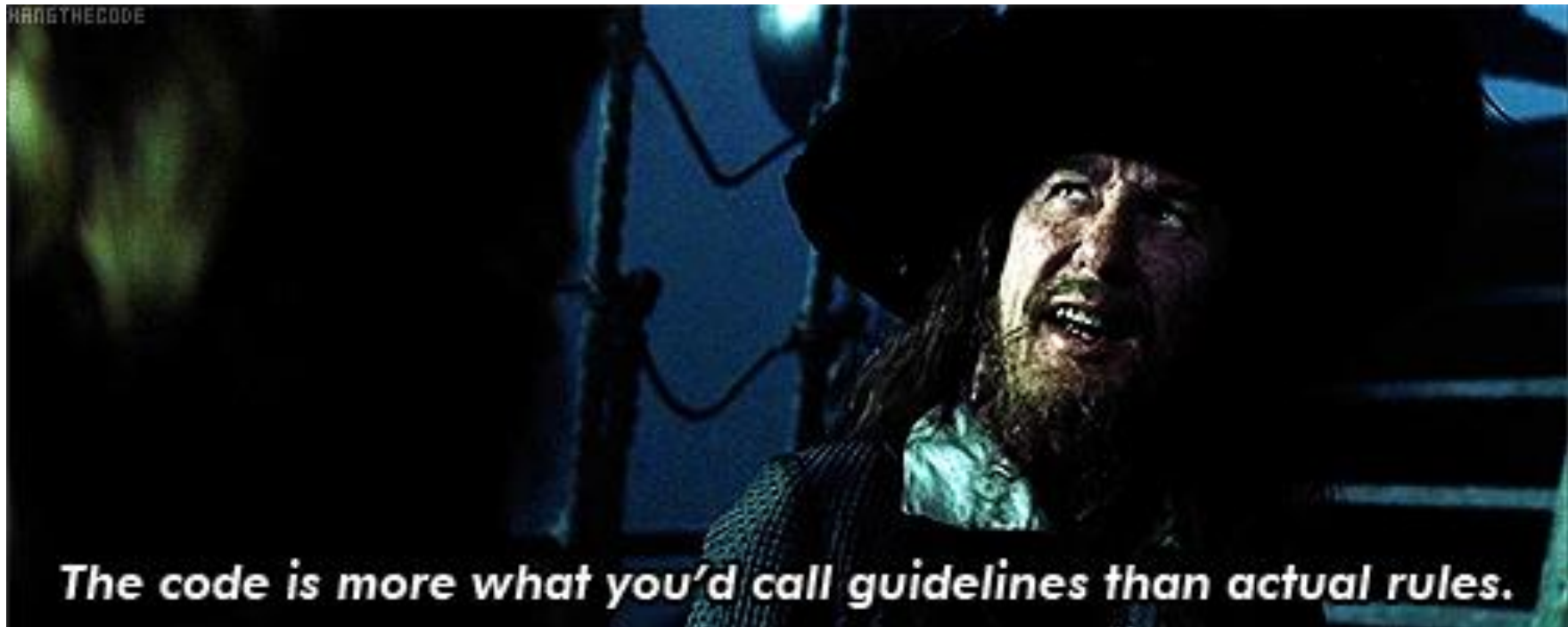
What is Exploratory Analysis?

- Based on this awesome resource:
<https://r4ds.had.co.nz/exploratory-data-analysis.html>
- Shows everything in R, but the same things can be done in Excel or SAS

Exploration of your data
in a systematic way:

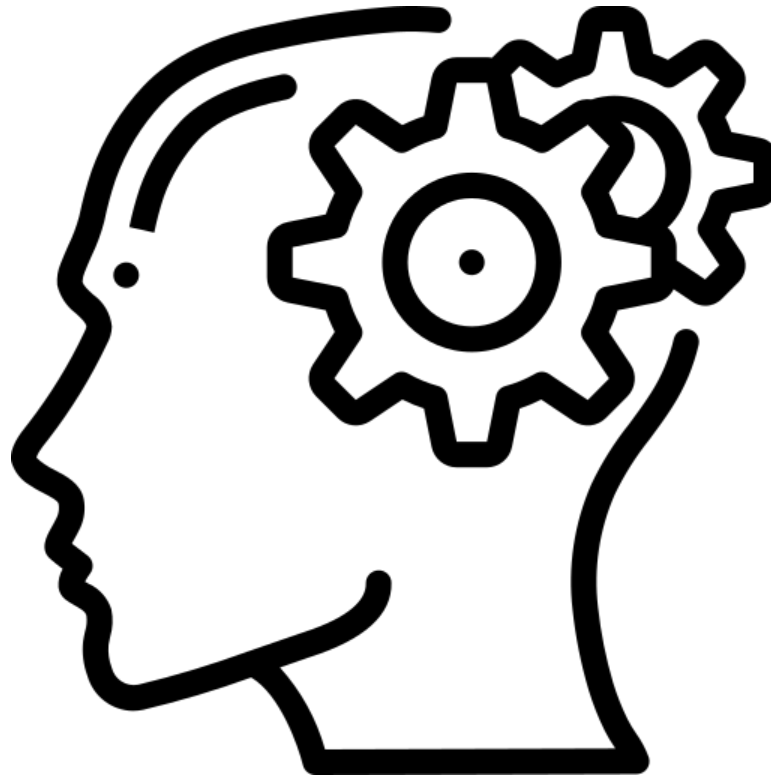
No strict set of rules

HANDTHECODE



The code is more what you'd call guidelines than actual rules.

EDA is a state of mind



Why is EDA important?

Investigate the quality of your data:

Does your data meet your expectations?

1. Generate questions
2. Search for answers
 - Visualize
 - Transform
 - Model
3. Refine or generate new questions

Far better an **approximate answer to the right question**, which is often vague, than an exact answer to the wrong question, which can always be made precise.

-- John Tukey

EDA is all about asking questions

- “The key to asking *quality* questions is to generate a large *quantity* of questions.”
- “Follow up each question with a new question based on what you find.”

Two general types of questions to ask:

- What type of variation occurs within my variables?
- What type of covariation occurs between my variables?

- **Variation:** phenomenon that creates uncertainty
- **Covariation:** phenomenon that reduces uncertainty

USE VISUALIZATIONS!!

- **Categorical:** bar chart
- **Continuous:** histogram
 - Multiple histograms: line graph

When looking at your data:

- What values are the most common?
Why?
- Which values are the least common?
Why?
- What patterns can you see? What might explain them?

Unusual data

- Outliers
- Missing data

Outliers: what do you do with them?

- Why are they outliers?
 - Data entry error?
 - True outlier?
- What would happen if you removed those observations?
- What would happen if you set them to missing?

Missing data: what do you do with it?

- Why is it missing?
- What would happen if you removed those observations?

Covariation:

- Behavior between variables
- Tendency for the values of two or more variables to vary together in a related way

- **Categorical and continuous:** histogram, line graph, density plot, boxplot
- **2 categorical:** count plot (geom_count), tile plot
- **2 continuous:** scatterplot, boxplot

Patterns and Models

- Is the pattern caused by random chance?
- How would you describe the relationship you see in the pattern?
- How strong is the relationship?
- What other variables impact the pattern?
- What happens when you look at subgroups?

Patterns and Models

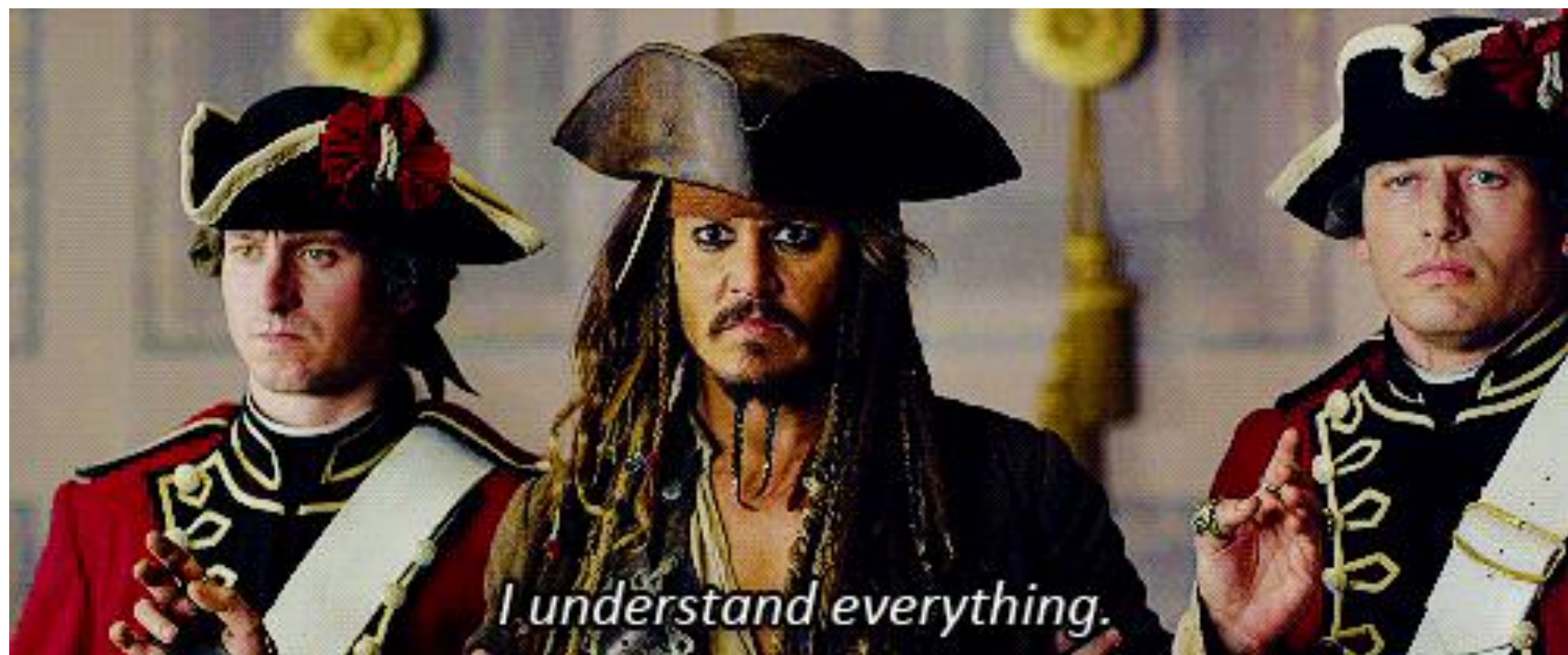
“If two variables covary, you can use the values of one variable to make better predictions about the values of the second. If the covariation is due to a **causal** relationship (a special case), then you can **use the value of one variable to control the value of the second.**”

Patterns and Models

Think of models as tools for extracting patterns out of data

**What does this mean for
syndromic surveillance and for
public health surveillance?**

- **Biggest barrier: TIME**
- Document what you do whenever possible
- Try developing a set of “standard EDA questions” that you can start with
- Set your own guidelines



Useful resources (not just for R users!)

- Book these slides are based on:
<https://r4ds.had.co.nz/exploratory-data-analysis.html>
- R Data viz cheat sheet:
<https://github.com/rstudio/cheatsheets/raw/master/data-visualization-2.1.pdf>

Data Visualization with ggplot2 : : CHEAT SHEET

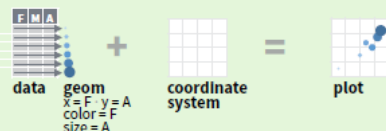


Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same components: a **data set**, a **coordinate system**, and **geoms**—visual marks that represent data points.



To display values, map variables in the data to visual properties of the geom (**aesthetics**) like **size**, **color**, and **x** and **y** locations.



Complete the template below to build a graph.

```
ggplot(data = <DATA>) +  
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>),  
  stat = <STAT>, position = <POSITION>) +  
  <COORDINATE_FUNCTION> +  
  <FACET_FUNCTION> +  
  <SCALE_FUNCTION> +  
  <THEME_FUNCTION>
```

required

Not required, sensible defaults supplied

ggplot(data = mpg, aes(x = cty, y = hwy)) Begins a plot that you finish by adding layers to. Add one geom function per layer.

qplot(x = cty, y = hwy, data = mpg, geom = "point") Creates a complete plot with given data, geom, and mappings. Supplies many useful defaults.

last_plot() Returns the last plot

ggsave("plot.png", width = 5, height = 5) Saves last plot as 5" x 5" file named "plot.png" in working directory. Matches file type to file extension.

Geoms

Use a geom function to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.

GRAPHICAL PRIMITIVES

```
a <- ggplot(economics, aes(date, unemployment))  
b <- ggplot(seals, aes(x = long, y = lat))
```

a + geom_blank()
(Useful for expanding limits)

b + geom_curve(aes(yend = lat + 1, xend = long + 1, curvature = z)) - x, xend, y, yend, alpha, angle, color, curvature, linetype, size

a + geom_path(lineend = "butt", linejoin = "round", linemitre = 1)
x, y, alpha, color, group, linetype, size

a + geom_polygon(aes(group = group))
x, y, alpha, color, fill, group, linetype, size

b + geom_rect(aes(xmin = long, ymin = lat, xmax = long + 1, ymax = lat + 1)) - xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size

a + geom_ribbon(aes(ymin = unemployment - 900, ymax = unemployment + 900)) - x, ymax, ymin, alpha, color, fill, group, linetype, size

LINE SEGMENTS

common aesthetics: x, y, alpha, color, linetype, size

```
b + geom_abline(aes(intercept = 0, slope = 1))  
b + geom_hline(aes(yintercept = lat))  
b + geom_vline(aes(xintercept = long))
```

b + geom_segment(aes(yend = lat + 1, xend = long + 1))
b + geom_spoke(aes(angle = 1:1155, radius = 1))

ONE VARIABLE continuous

```
c <- ggplot(mpg, aes(hwy)); c2 <- ggplot(mpg)
```

c + geom_area(stat = "bin")
x, y, alpha, color, fill, linetype, size

c + geom_density(kernel = "gaussian")
x, y, alpha, color, fill, group, linetype, size, weight

c + geom_dotplot()
x, y, alpha, color, fill

c + geom_freqpoly() x, y, alpha, color, group, linetype, size

c + geom_histogram(binwidth = 5) x, y, alpha, color, fill, linetype, size, weight

c2 + geom_qq(aes(sample = hwy)) x, y, alpha, color, fill, linetype, size, weight

discrete

```
d <- ggplot(mpg, aes(fl))
```

d + geom_bar()
x, alpha, color, fill, linetype, size, weight

TWO VARIABLES

continuous x, continuous y

```
e <- ggplot(mpg, aes(cty, hwy))
```

e + geom_label(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE) x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

e + geom_jitter(height = 2, width = 2)
x, y, alpha, color, fill, shape, size

e + geom_point() x, y, alpha, color, fill, shape, size, stroke

e + geom_quantile() x, y, alpha, color, group, linetype, size, weight

e + geom_rug(sides = "bl") x, y, alpha, color, linetype, size

e + geom_smooth(method = lm) x, y, alpha, color, fill, group, linetype, size, weight

e + geom_text(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE) x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

discrete x, continuous y

```
f <- ggplot(mpg, aes(class, hwy))
```

f + geom_col() x, y, alpha, color, fill, group, linetype, size

f + geom_boxplot() x, y, lower, middle, upper, ymax, ymin, alpha, color, fill, group, linetype, shape, size, weight

f + geom_dotplot(binaxis = "y", stackdir = "center") x, y, alpha, color, fill, group

f + geom_violin(scale = "area") x, y, alpha, color, fill, group, linetype, size, weight

discrete x, discrete y

```
g <- ggplot(diamonds, aes(cut, color))
```

g + geom_count() x, y, alpha, color, fill, shape, size, stroke

THREE VARIABLES

```
seals$z <- with(seals, sqrt(delta_long^2 + delta_lat^2)) l <- ggplot(seals, aes(long, lat))
```

l + geom_contour(aes(z = z))
x, y, z, alpha, colour, group, linetype, size, weight

continuous bivariate distribution

```
h <- ggplot(diamonds, aes(carat, price))
```

h + geom_bin2d(binwidth = c(0.25, 500))
x, y, alpha, color, fill, linetype, size, weight

h + geom_density2d()
x, y, alpha, colour, group, linetype, size

h + geom_hex()
x, y, alpha, colour, fill, size

continuous function

```
i <- ggplot(economics, aes(date, unemployment))
```

i + geom_area()
x, y, alpha, color, fill, linetype, size

i + geom_line()
x, y, alpha, color, group, linetype, size

i + geom_step(direction = "hv")
x, y, alpha, color, group, linetype, size

visualizing error

```
df <- data.frame(grp = c("A", "B"), fit = 4:5, se = 1:2)  
j <- ggplot(df, aes(grp, fit, ymin = fit-se, ymax = fit+se))
```

j + geom_crossbar(fatten = 2)
x, y, ymax, ymin, alpha, color, fill, group, linetype, size

j + geom_errorbar() x, ymax, ymin, alpha, color, group, linetype, size, width (also **geom_errorbarh**())

j + geom_linerange()
x, ymin, ymax, alpha, color, group, linetype, size

j + geom_pointrange()
x, y, ymin, ymax, alpha, color, fill, group, linetype, shape, size

maps

```
data <- data.frame(murder = USArrests$Murder, state = tolower(rownames(USArrests)))  
map <- map_data("state")  
k <- ggplot(data, aes(fill = murder))
```

k + geom_map(aes(map_id = state), map = map) + **expand_limits**(x = map\$long, y = map\$lat), map_id, alpha, color, fill, linetype, size

Discussion

- What questions do you have about current projects?
- What ideas do you have for future projects?
- What ESSENCE questions or issues have you come across?
- **What other resources can we work on putting together?**

Upcoming AZ Syndromic Meetings

AZ BioSense
Workgroup

Tuesday,
April 2 from
1:00 – 2:00

Varicella
Workgroup

Monday,
April 15 from
1:00 – 2:00

AZ BioSense
Opioid
Workgroup

Thursday,
April 18 from
2:00 – 3:00

THANK YOU

syndromicsurveillance@azdhs.gov | 602-542-6002

azhealth.gov

 @azdhs

 facebook.com/azdhs



ARIZONA DEPARTMENT
OF HEALTH SERVICES

Health and Wellness for all Arizonans