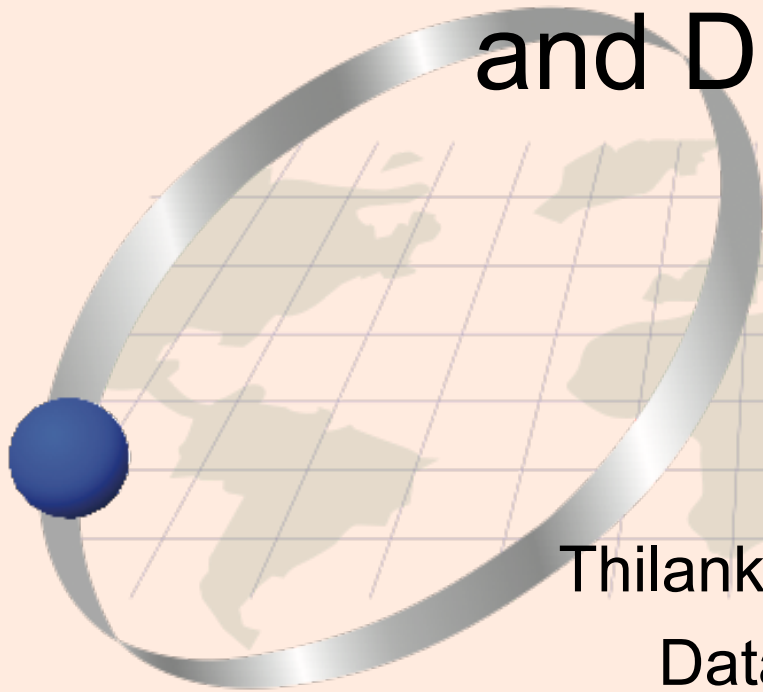


Data and Information Resources, Role of Hypothesis, Exploration and Distributions



Thilanka Munasinghe

Data Analytics

ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960 MGMT-
4962/6962 BCBP 4960

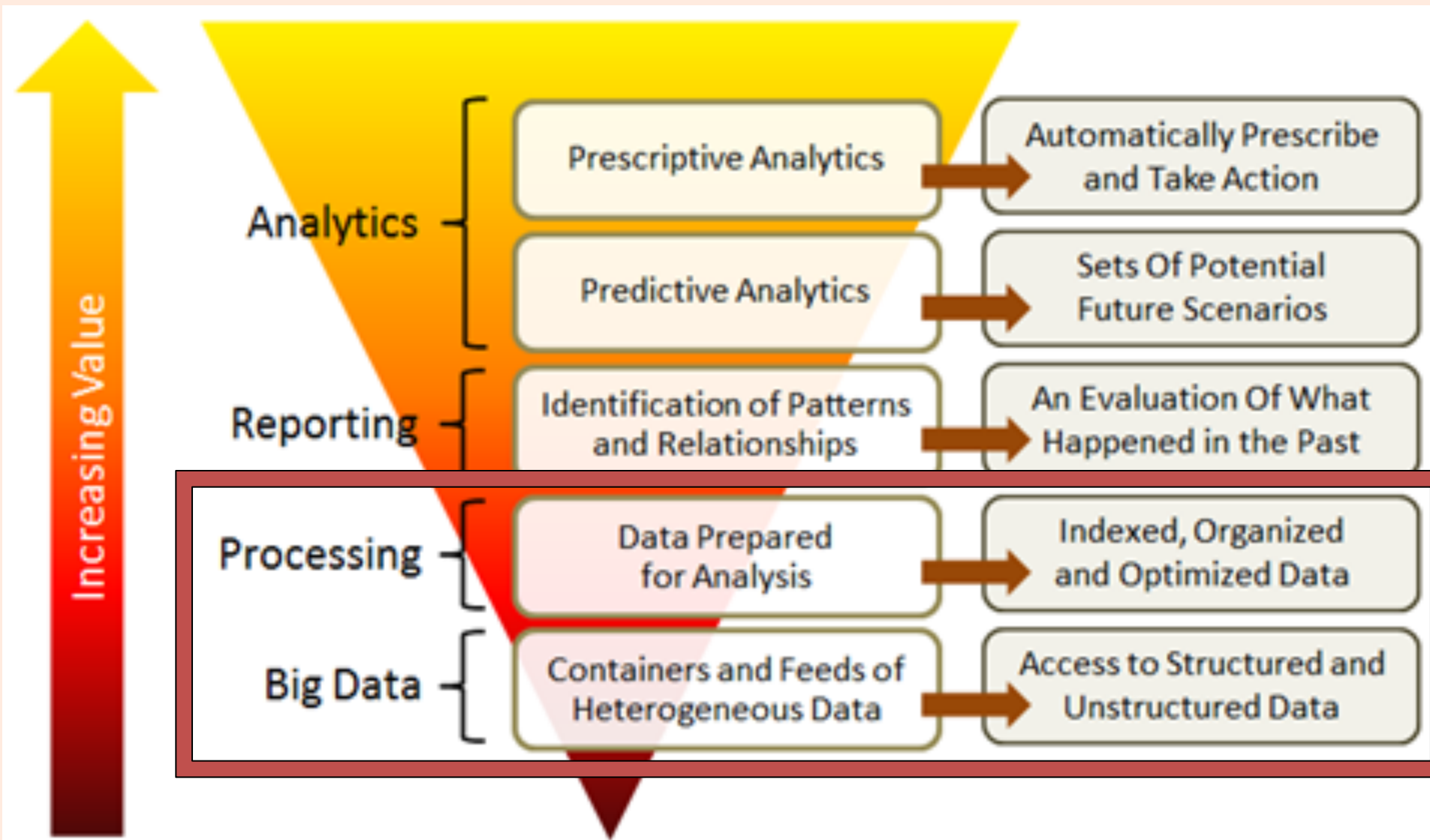
Group 1 Module 3(a), ~ Sept 14th, 2021

Contents

- Data sources
 - Cyber
 - Human
- “Munging”
- Exploring
 - Distributions...
 - Summaries
 - Visualization
- Testing and evaluating the results (beginning)



Lower layers in the Analytics Stack





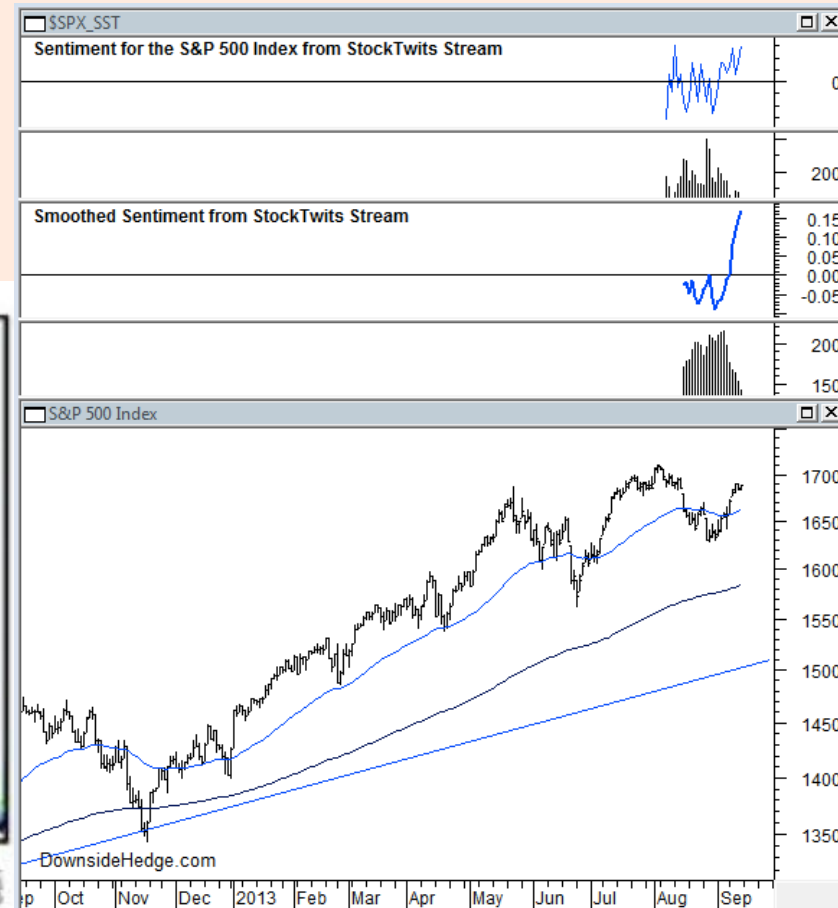
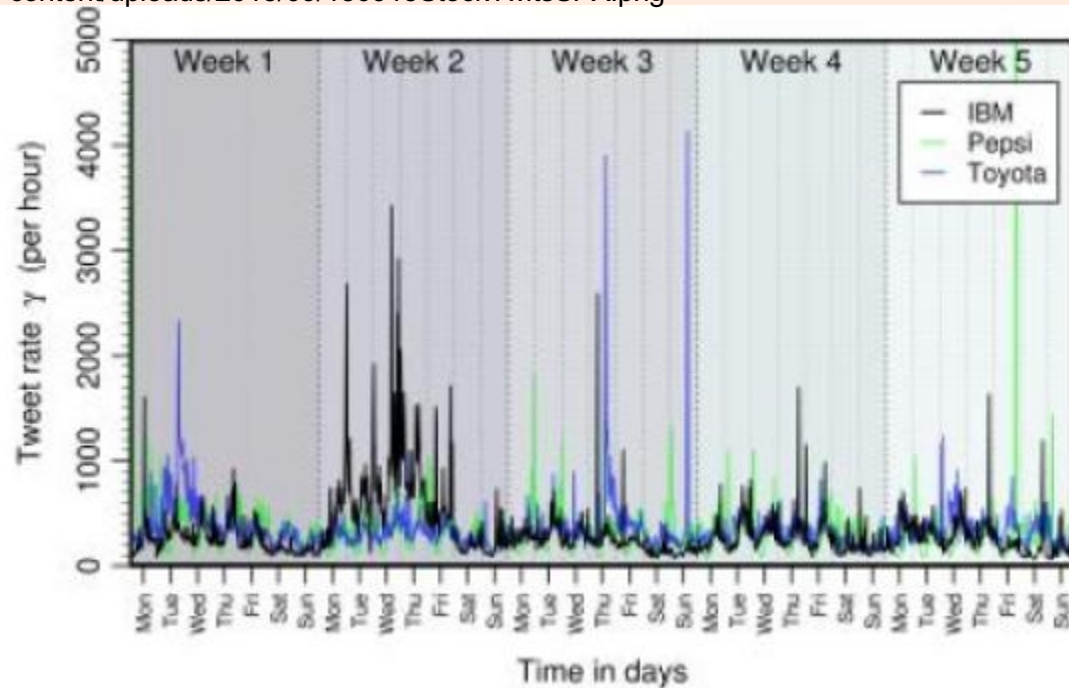
“Cyber Data”

...

iStock.
by Getty Images™

“Human Data” ...

Image Credit: <http://images.sciencedaily.com/2013/10/131007151731-large.jpg>
<http://www.downsidehedge.com/wp-content/uploads/2013/09/130915StockTwitsSPX.png>



Statistics Review – Probability

You have learned this in your Statistics class.

Probability ...

- Before dive into the Naïve Bayes lecture in upcoming classes, lets go over some definitions in probability.
- Probability is the measure of the likelihood that an event will occur.
- In other words, probability is a measurement of how likely an event occurs.
- *Probability of event **A**:*

$$P(A) = \frac{\textit{Number of ways for A}}{\textit{Total number of possible outcomes}}^7$$

Probability ...

- Before we do a deep dive, You should know/understand the two probability concepts:

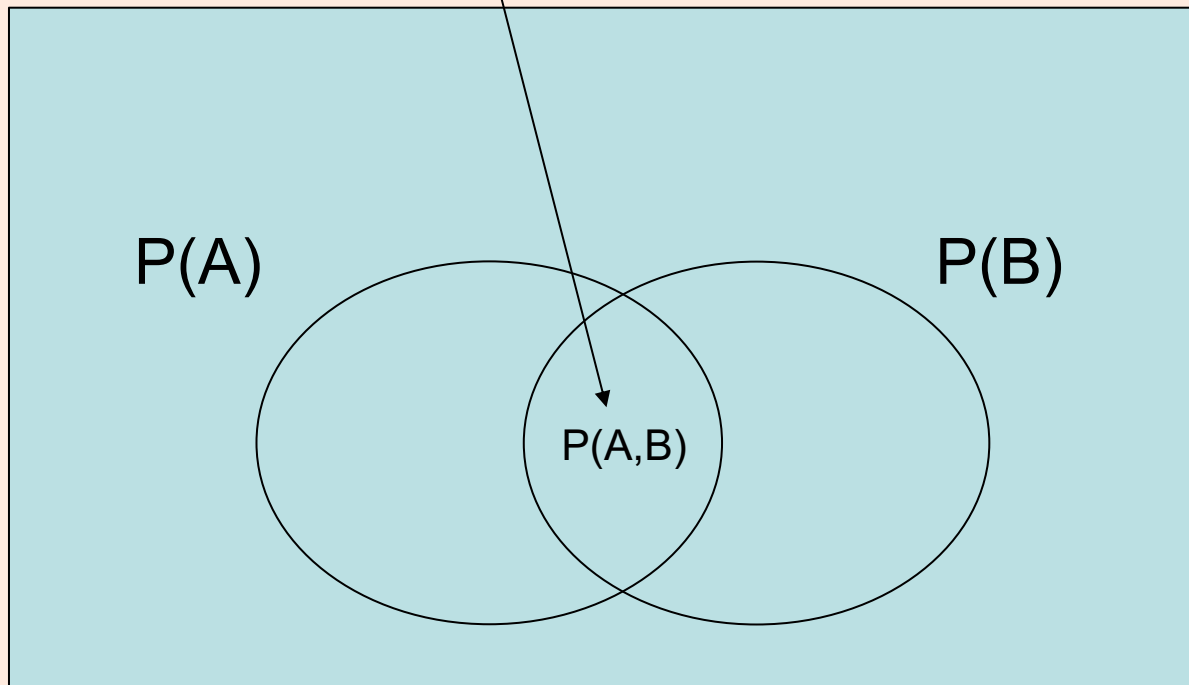
1) Joint Probability

2) Conditional Probability

Probability ...

Joint Probability: *specifies the probability of event A and event B occurring together.*

Joint Probability A and B



Probability ...

Joint Probability: specifies the probability of event *A* and event *B* occurring together.

If the two events are independent,

What is the probability of getting two 6's when you roll two dice?

The probability of rolling(getting) two 6's:

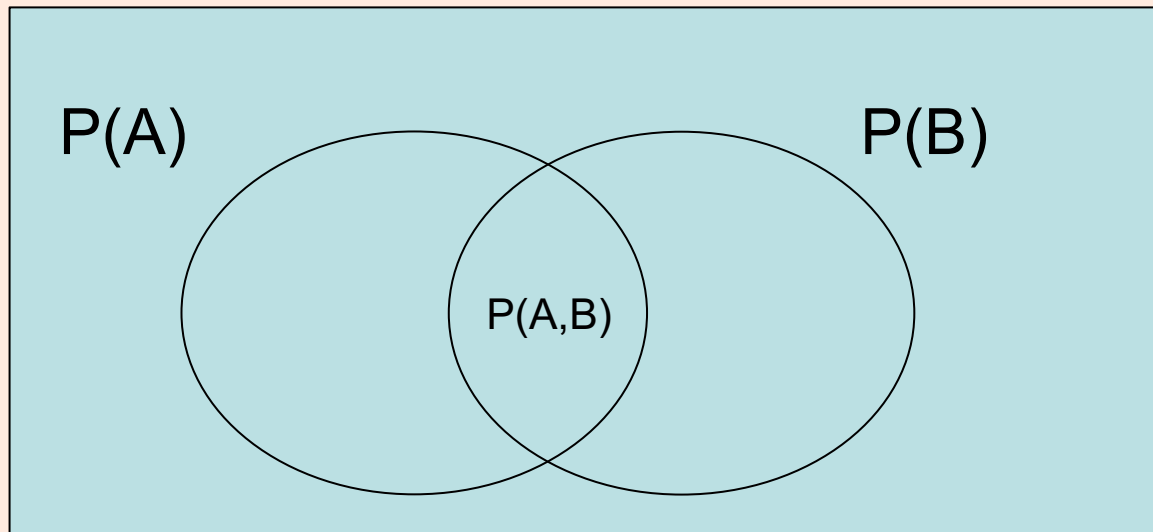
$$P(A, B) = P(A) * P(B) = \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$$



Probability ...

Conditional Probability: *probability of event A occurring, given that event B occurred.*

$$P(A|B) = \frac{P(A,B)}{P(B)} = \text{Probability of A, given B ; } P(B) > 0$$



Bayes Theorem...

- The relationship between conditional probabilities, $P(B|A)$ and $P(A|B)$ can be expressed using the Bayes Theorem.

$$P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$$

EPI data set from previous lecture...

Index of /html/DA/EPI			
	<u>Name</u>	<u>Last modified</u>	<u>Size</u> <u>Description</u>
	Parent Directory	-	
	2010EPI_data.csv	05-Feb-2016 00:28	10M
	2010EPI_data.xls	05-Feb-2016 00:35	11M
	2016 EPI Wastewater Data Appendix.xls	19-Jan-2018 16:01	907K
	2016EPI Backcasted Scores.xls	19-Jan-2018 16:01	1.3M
	2016EPI Full Report_opt.pdf	19-Jan-2018 16:02	15M
	2016EPI Raw Data.xls	19-Jan-2018 16:02	1.5M
	2016 epi framework indicator scores friendly.xls	19-Jan-2018 16:02	740K
	2016epi weightings 0.xls	19-Jan-2018 16:02	660K
	EPI_data.csv	05-Feb-2016 00:28	232K
	EPI_data.xls	05-Feb-2016 00:36	11M
	Fisheries Penalties.xls	19-Jan-2018 16:02	120K
	OnlyEPI_data.csv	05-Feb-2016 00:29	10M
	OnlyEPI_data.xls	05-Feb-2016 00:37	11M
	filters materiality for 2016epi.xls	19-Jan-2018 16:02	64K
Apache/2.2.14 (Ubuntu) Server at aquarius.tw.rpi.edu Port 443			

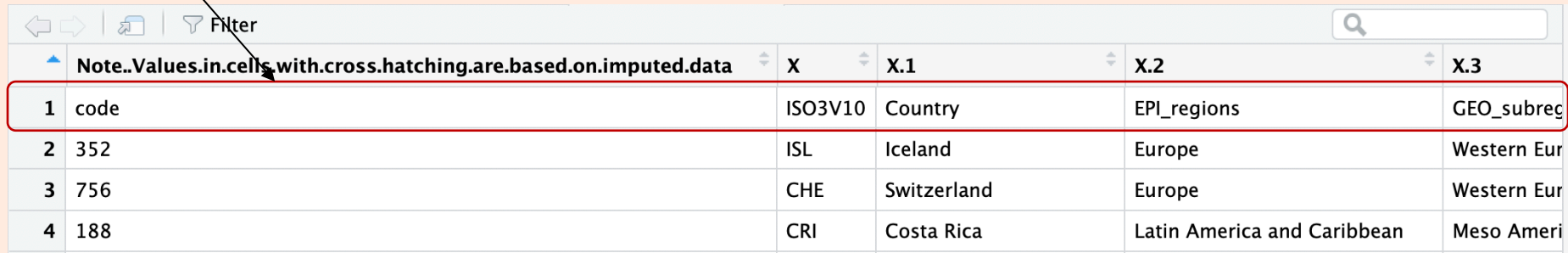
<https://aquarius.tw.rpi.edu/html/DA/EPI/>

2010EPI_data.xls

Home										Insert	Draw	Page Layout	Formulas	Data	Review	View	
<div><div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div></div><div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></div><div></div></div></div></div> <div><div><div></div><div></div></div><div><div><div></</div></div></div></div>																	

2010EPI dataset in R

Before:



The screenshot shows a data table interface. At the top, there is a toolbar with navigation arrows, a filter icon, and a search bar. Below the toolbar, a note reads: "Note..Values.in.cells.with.cross.hatching.are.based.on.imputed.data". The table has five columns: "code", "ISO3V10", "Country", "EPI_regions", and "GEO_subregion". The first row is highlighted with a red border. Below it are three more rows of data.

	code	ISO3V10	Country	EPI_regions	GEO_subregion
1	352	ISL	Iceland	Europe	Western Eur
2	756	CHE	Switzerland	Europe	Western Eur
3	188	CRI	Costa Rica	Latin America and Caribbean	Meso Ameri

We want this: Change the first row to be the header:



The screenshot shows a data table interface where the first row is now the header. The table has five columns: "code", "ISO3V10", "Country", "EPI_regions", and "GEO_subregion". The first row is highlighted with a light blue background. Below it are three more rows of data.

	code	ISO3V10	Country	EPI_regions	GEO_subregion
2	352	ISL	Iceland	Europe	Western Europe
3	756	CHE	Switzerland	Europe	Western Europe
4	188	CRI	Costa Rica	Latin America and Caribbean	Meso America

How to change the first row to be the header in 2010EPI ?

How to change the first row to be the header in R?

```
names(data_2010EPI) <- as.matrix(data_2010EPI[1, ])
```

```
data_2010EPI <- data_2010EPI[-1, ]
```

```
data_2010EPI[] <- lapply(data_2010EPI, function(x)
```

```
type.convert(as.character(x)))
```

```
data_2010EPI
```

```
View(data_2010EPI)
```

```
# How to change the first row to be the header in R?
```

```
names(data_2010EPI) <- as.matrix(data_2010EPI[1, ])
```

```
data_2010EPI <- data_2010EPI[-1, ]
```

```
data_2010EPI[] <- lapply(data_2010EPI, function(x) type.convert(as.character(x)))
```

```
data_2010EPI
```

```
View(data_2010EPI)
```

Data Prepared for Analysis = Munging

- Missing values, null values, etc.
- E.g. in the EPI_data – they use “--”
 - Most data applications provide built ins for these higher-order functions – in R “NA” is used and functions such as `is.na(var)`, etc. provide powerful filtering options (we’ll cover these on next class)
- Of course, different variables often are missing “different” values
- In R – higher-order functions such as: Reduce, Filter, Map, Find, Position and Negate will become your enemies and then your friends:

<http://www.johnmyleswhite.com/notebook/2010/09/23/higher-order-functions-in-r/>

Explore the “Missing values” -- NA

ISL	Iceland	59039821	NA	100
CHE	Switzerland		NA	100
CRI	Costa Rica	.61820969	75	100
SWE	Sweden	.16391001	NA	100
NOR	Norway	88187967	NA	100
MUS	Mauritius	45367567	83.333333	88.4616
FRA	France	.73853548	50	100
AUT	Austria		NA	100
CUB	Cuba	.76040364	47.058824	100
COL	Colombia	22333964	47.142857	NA
MLT	Malta	49471766	NA	100
FIN	Finland	87493735	NA	100
SVK	Slovakia		NA	100
GBR	United Kingdom	24868389	66.666667	100
NZL	New Zealand	33441573	78.571429	NA
CHL	Chile	61701405	28.571429	100
DEU	Germany	100	NA	100
ITA	Italy	.67003818	100	100
PRT	Portugal	.93377589	100	100
JPN	Japan	.38311329	45	100
LVA	Latvia	.72311585	NA	100
CZE	Czech Republic		NA	100
ALB	Albania	24993836	NA	100
PAN	Panama	.66377944	50	77.4416

Getting started – summarize data

- Summary statistic
 - Ranges, “hinges”
 - Tukey’s five numbers
- Look for a distribution match
- Tests...for...
 - Normality – shapiro-wilksand a p-value – what is the null hypothesis here?
> shapiro.test(EPI_data\$EPI)
Shapiro-Wilk normality test
data: EPI_data\$EPI
p-value = 0.1188

Accept or Reject?

- **Reject the null hypothesis if the p-value is less than the level of significance.**
- **You will fail to reject the null hypothesis if the p-value is greater than or equal to the level of significance.**
- **Typical significance 0.05 (!)**

Another variable in EPI

```
> shapiro.test(EPI_data$DALY)
```

Shapiro-Wilk normality test

data: EPI_data\$DALY

W = 0.9365, p-value = 1.891e-07

Read: [1] https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

[2] <https://www.dummies.com/programming/r/how-to-test-data-normality-in-a-formal-way-in-r/>

[3] <https://emilkirkegaard.dk/en/?p=4452>

Distribution tests

most distributions have tests

- Wilcoxon (Mann-Whitney)

- Comparing populations

Two data samples are independent if they come from distinct populations and the samples do not affect each other. Using the Mann-Whitney-Wilcoxon Test, we can decide whether the population distributions are identical without assuming them to follow the normal distribution. <http://www.r-tutor.com/elementary-statistics/non-parametric-methods/mann-whitney-wilcoxon-test>

- Kolmogorov-Smirnov (KS)

- It got out of control when people realized they can name the test after themselves, v. someone else...

Getting started – look at the data

- Visually
 - What is the improvement in the understanding of the data as compared to the situation without visualization?
 - Which visualization techniques are suitable for one's data?
 - Scatter plot diagrams
 - Box plots (min, 1st quartile, median, 3rd quartile, max)
 - Stem and leaf plots
 - Frequency plots
 - Group Frequency Distributions plot
 - Cumulative Frequency plots
 - Distribution plots

Why visualization?

- Reducing amount of data, quantization
- Patterns
- Features
- Events
- Trends
- Irregularities
- Leading to presentation of data, i.e. information products
- *Exit points for analysis*

Exploring the distribution

> summary(EPI) # stats

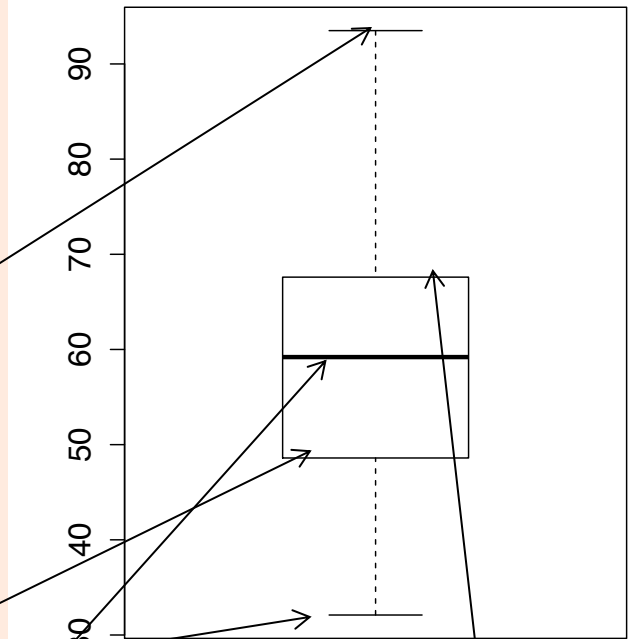
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
32.10	48.60	59.20	58.37	67.60	93.50	68

> boxplot(EPI)



> fivenum(EPI, na.rm=TRUE)

[1] 32.1 48.6 59.2 67.6 93.5



Tukey: min, lower hinge, median, upper hinge, max

Stem and leaf plot

> stem(EPI) # like-a histogram

The decimal point is 1 digit(s) to the right of the | - but the scale of the stem is 10... watch carefully..

3 | 234

3 | 66889

4 | 00011112222223344444

4 | 5555677788888999

5 | 0000111111111244444

5 | 55666677778888999999

6 | 000001111111222333344444

6 | 5555666666677778888889999999

7 | 000111233333334

7 | 5567888

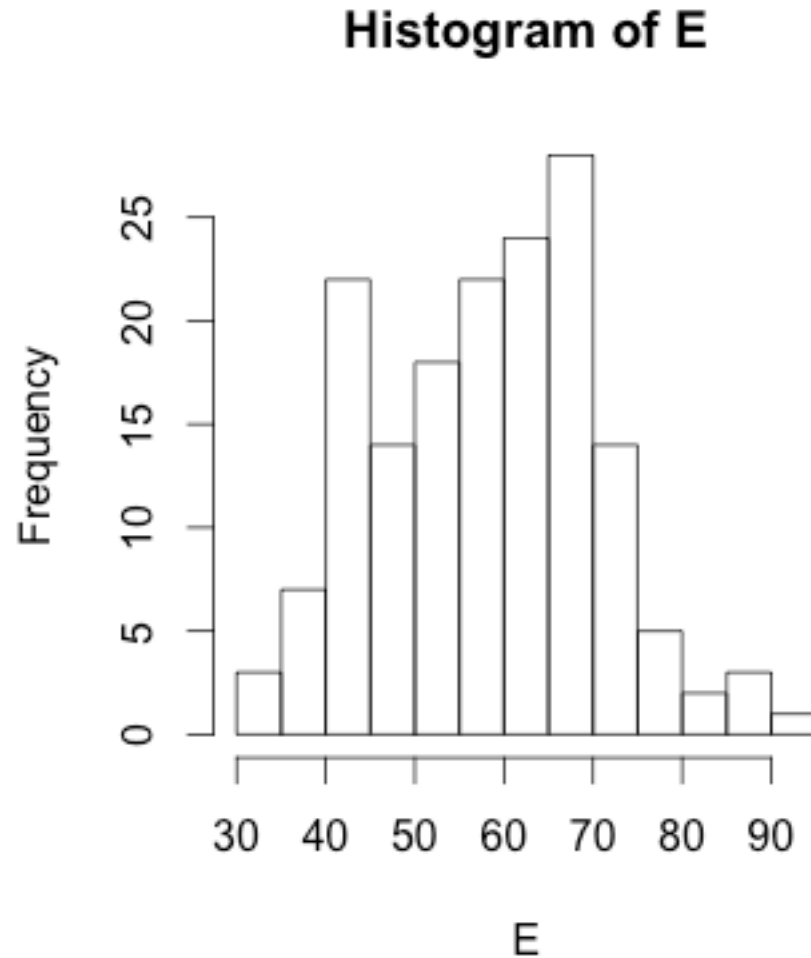
8 | 11

8 | 669

9 | 4

Grouped Frequency Distribution aka binning

```
> hist(EPI)      #defaults
```



Distributions

- Shape
- Character
- Parameter(s)
- Which one fits?

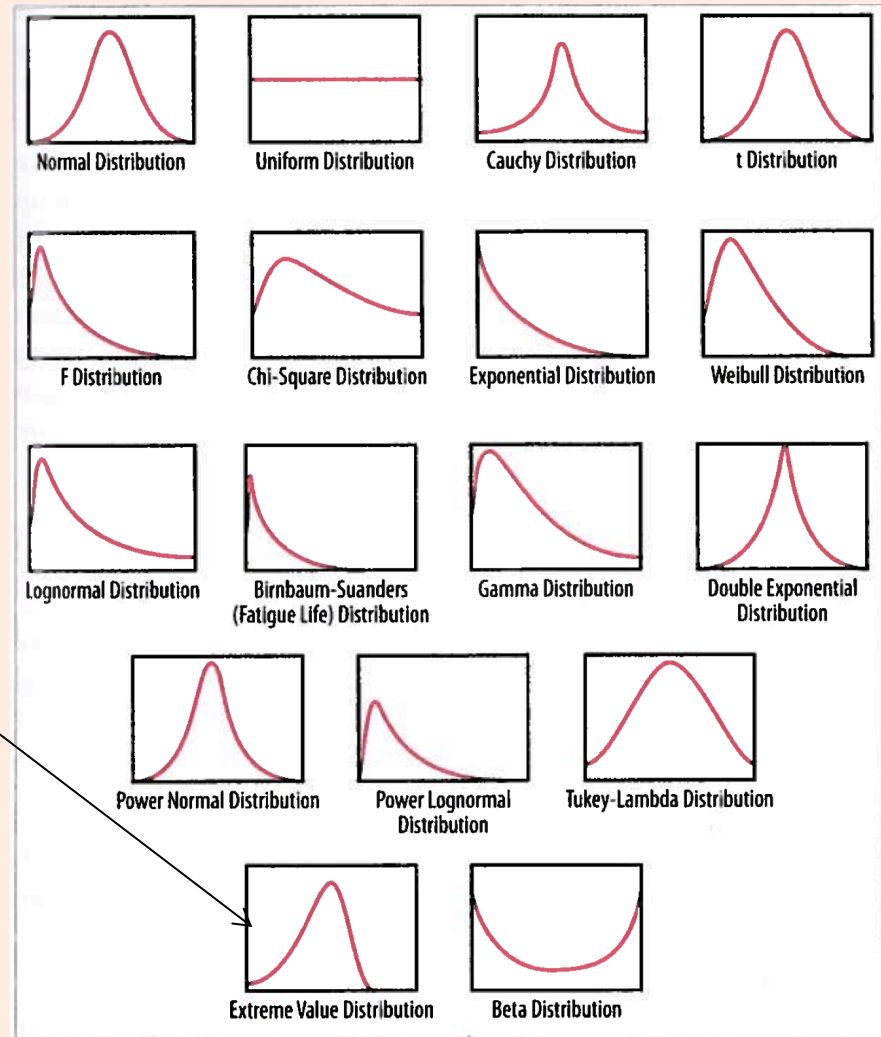
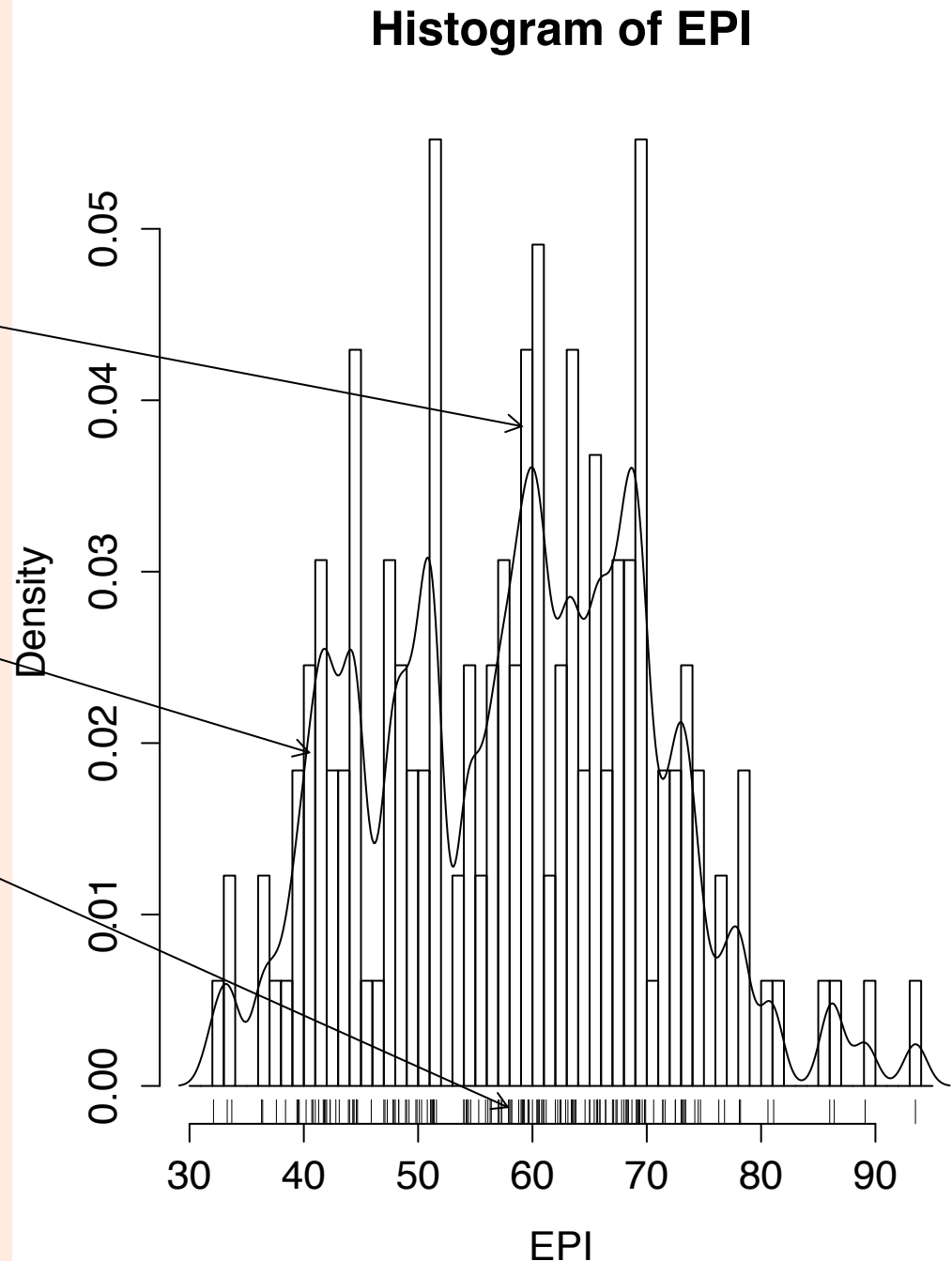


Figure 2-1. A bunch of continuous density functions (aka probability distributions)

```
> hist(EPI, seq(30.,  
95., 1.0), prob=TRUE)
```

```
> lines  
(density(EPI,na.rm=TR  
UE,bw=1.))
```

```
> rug(EPI)
```



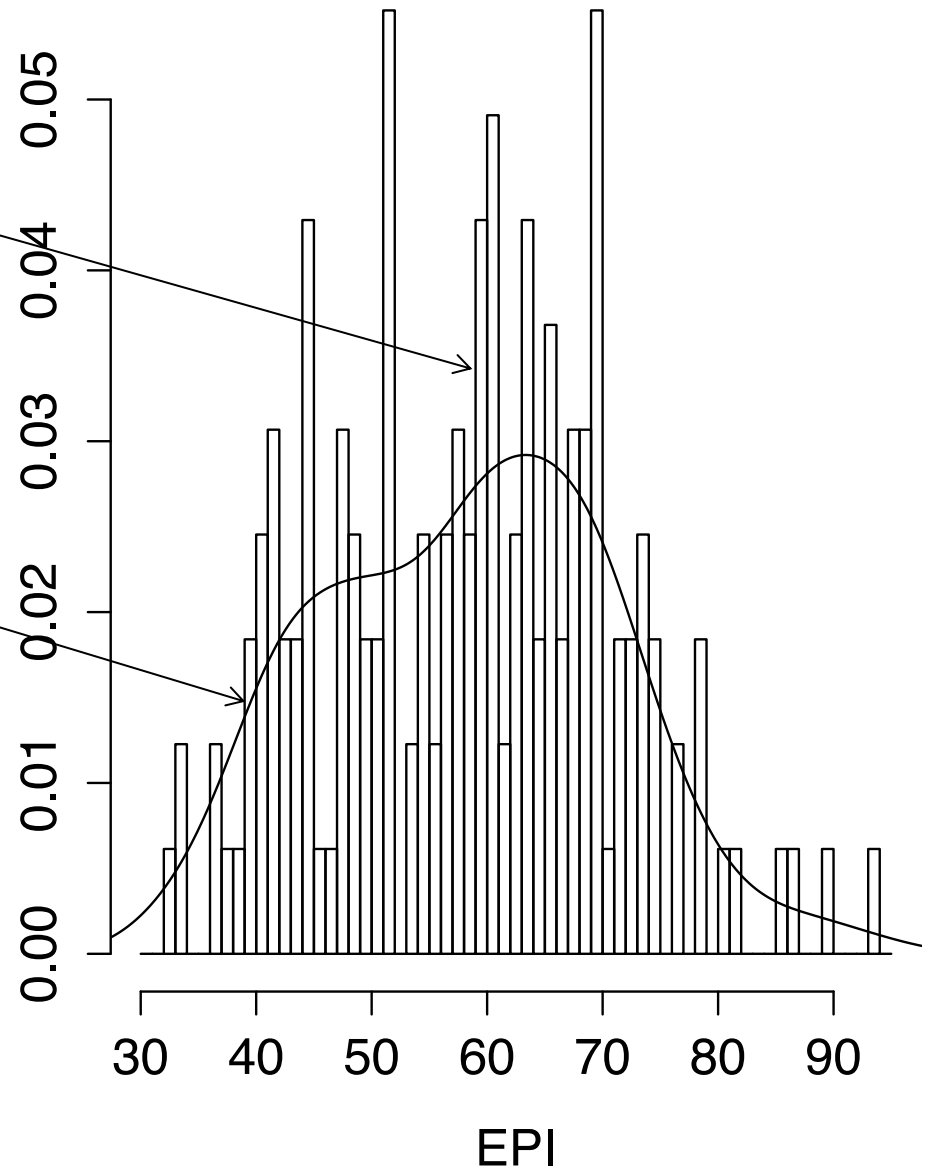
What is a rug plot ?

https://en.wikipedia.org/wiki/Rug_plot

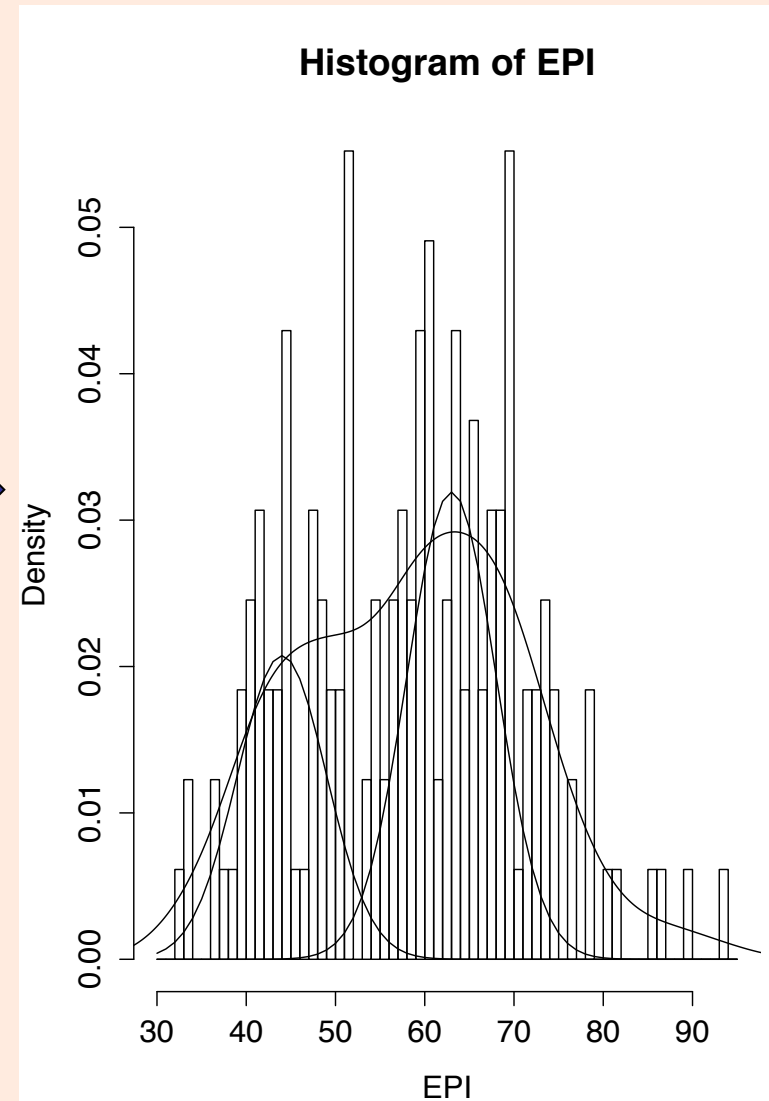
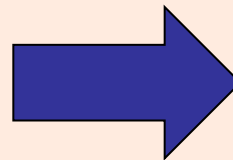
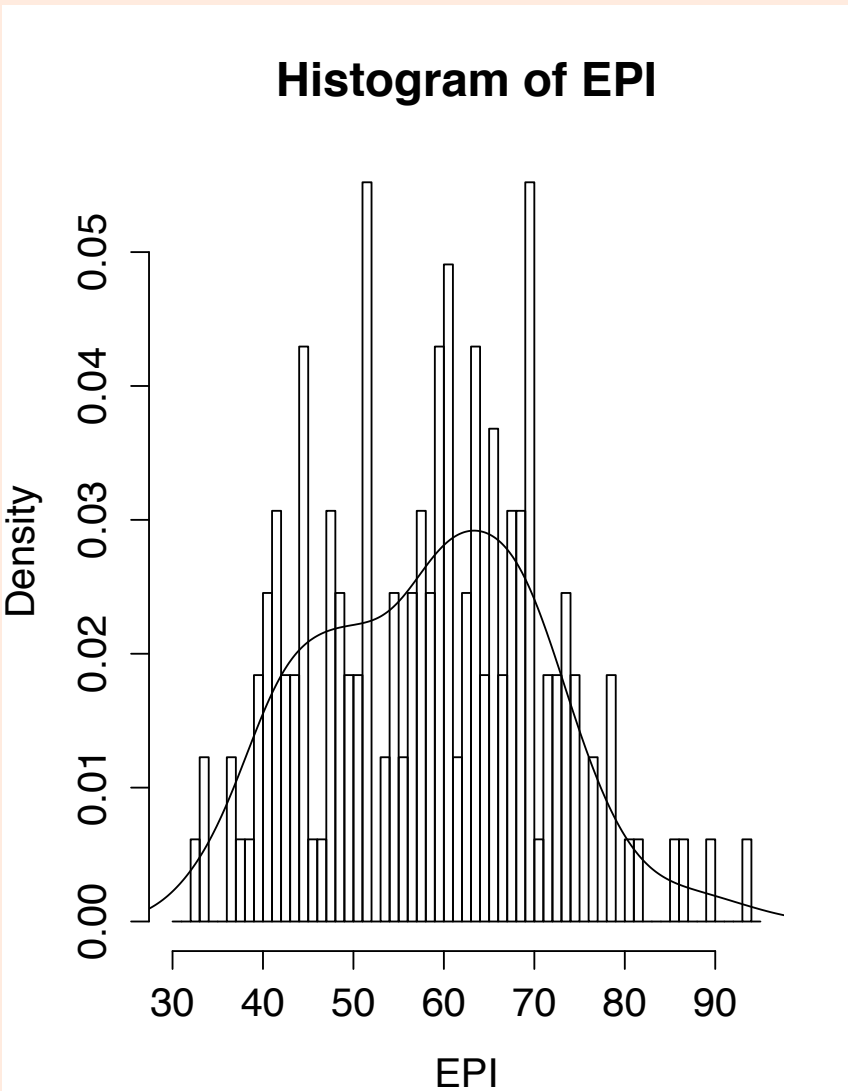
```
> hist(EPI, seq(30.,  
95., 1.0), prob=TRUE)
```

```
> lines  
(density(EPI,na.rm=TR  
UE,bw="SJ"))
```

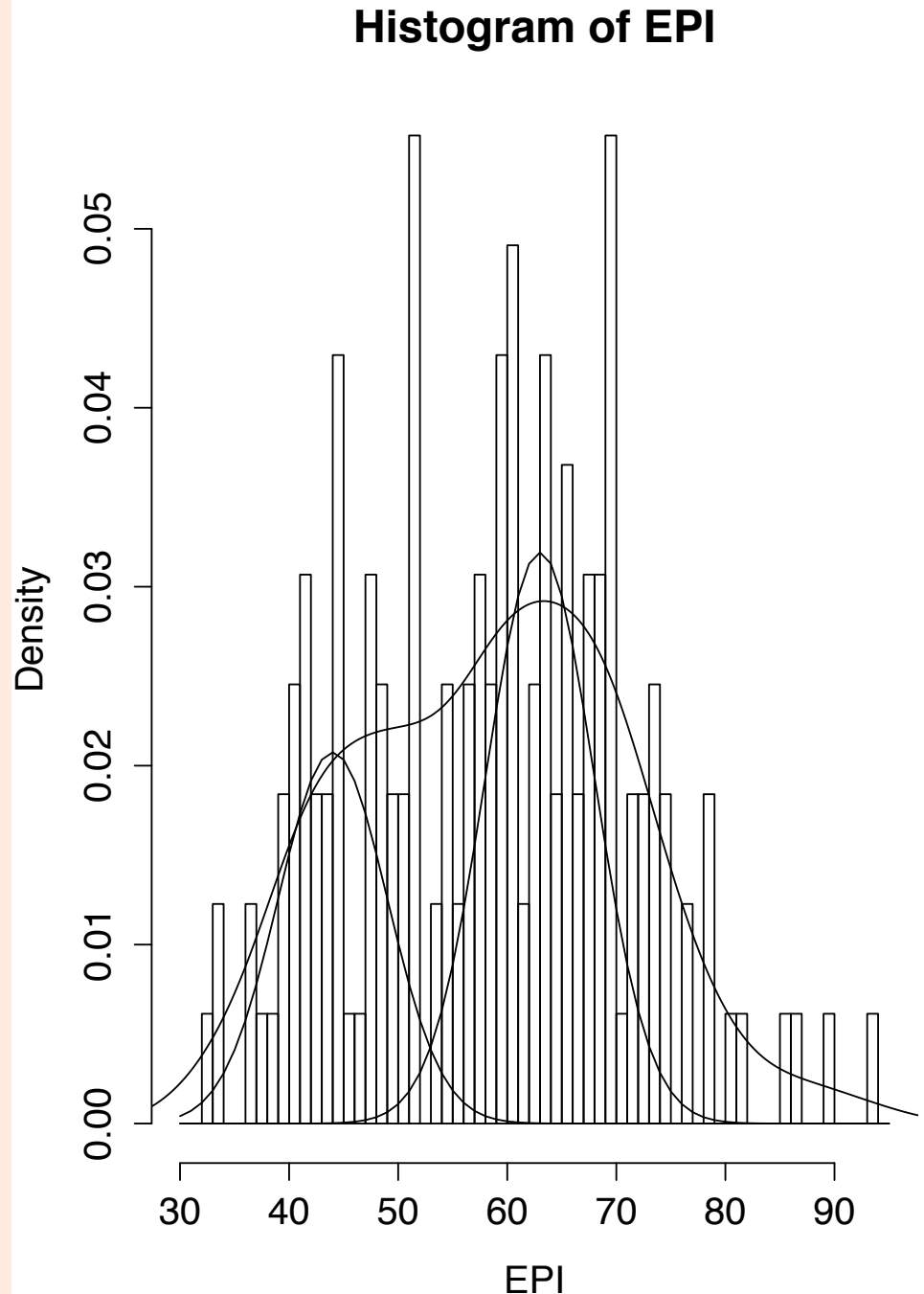
Density



Why are histograms so unsatisfying?



```
> xn<-seq(30,95,1)
> qn<-
dnorm(xn,mean=63,
sd=5,log=FALSE)
> lines(xn,qn)
> lines(xn,.4*qn)
> ln<-dnorm(xn,mean=44,
sd=5,log=FALSE)
> lines(xn,.26*ln)
```



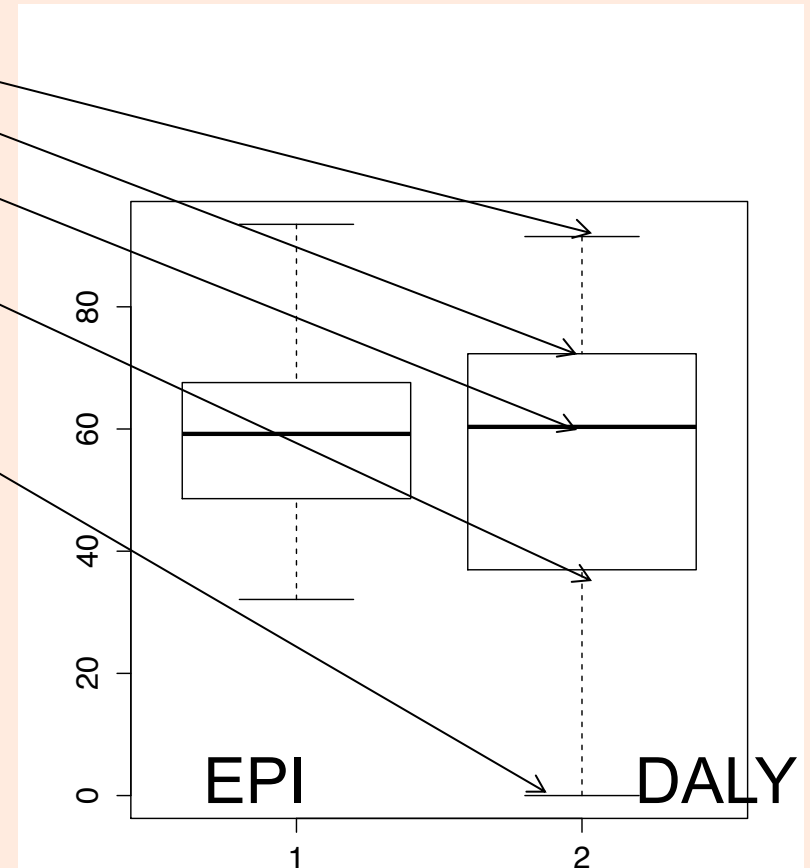
Exploring the distribution

> summary(DALY) # stats

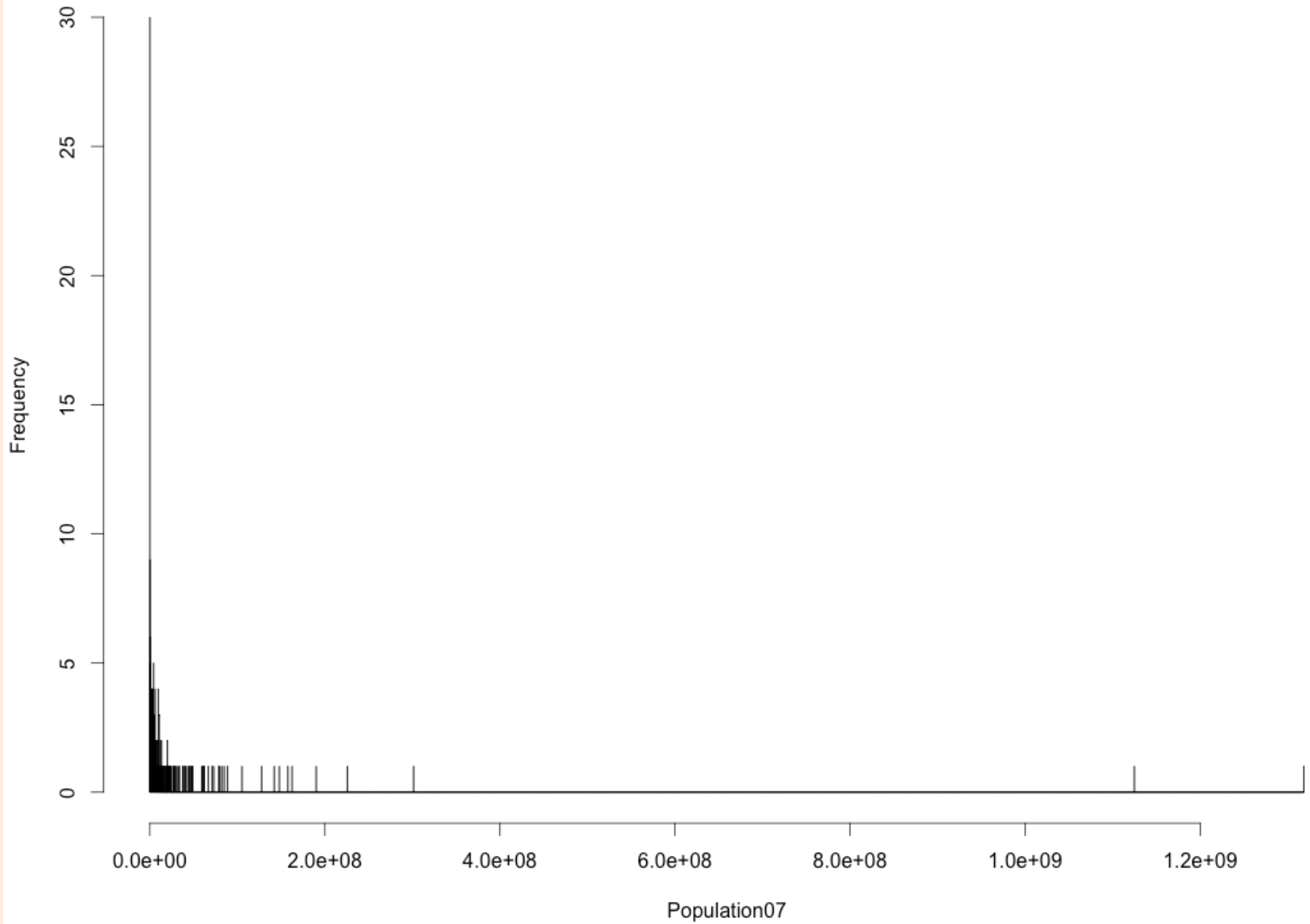
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.00	37.19	60.35	53.94	71.97	91.50	39

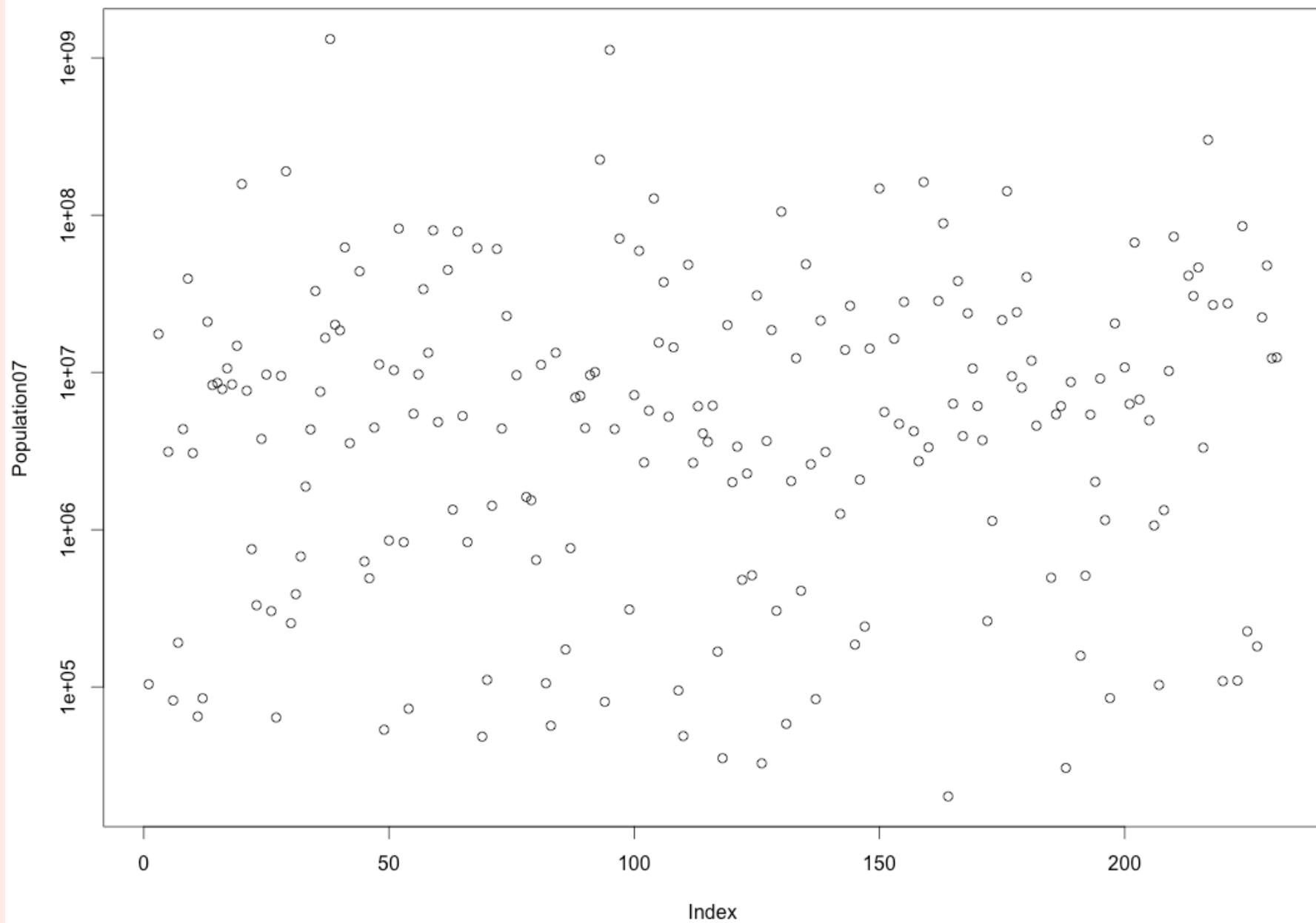
> fivenum(DALY,na.rm=TRUE)

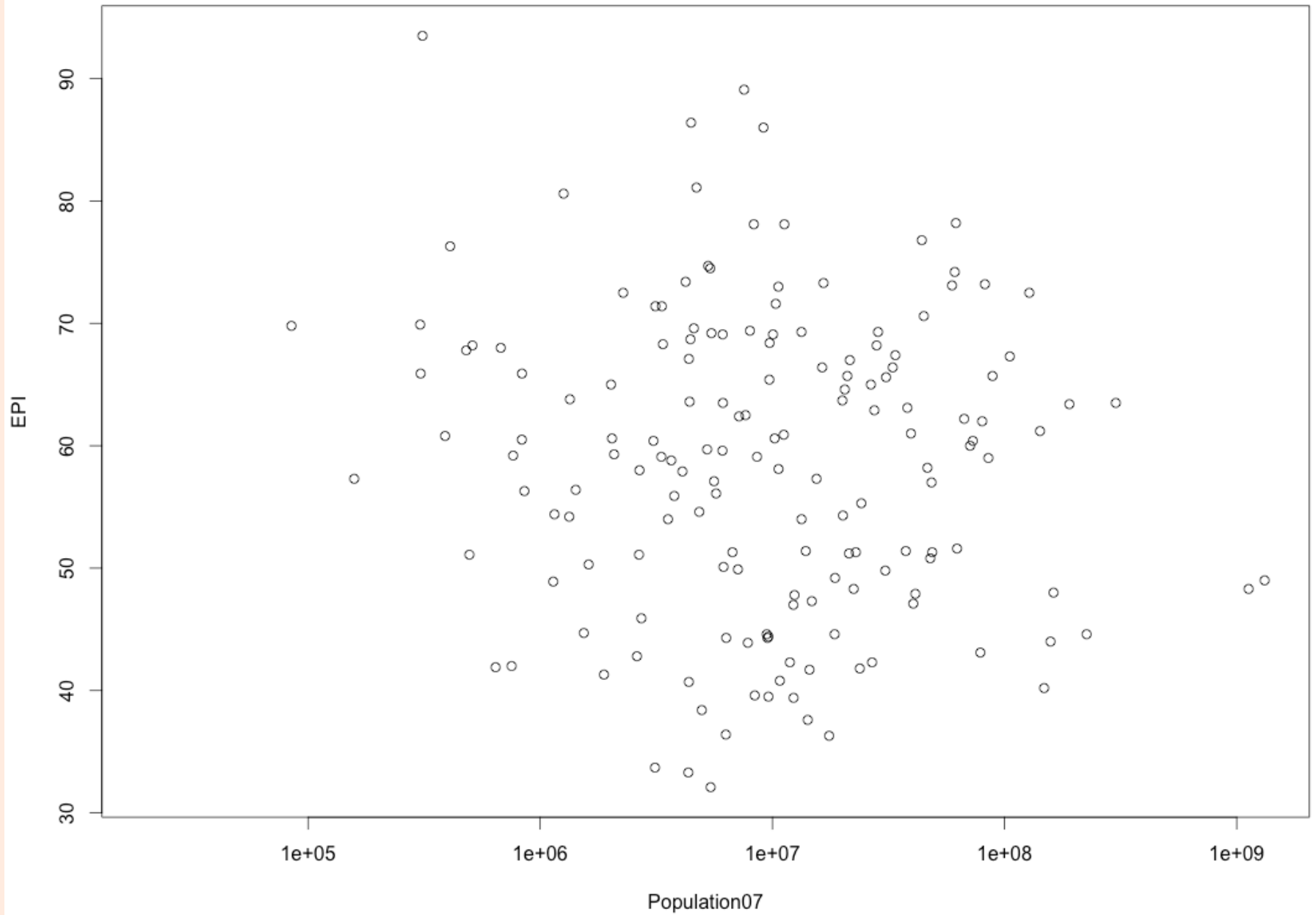
[1] 0.000 36.955 60.350 72.320 91.500

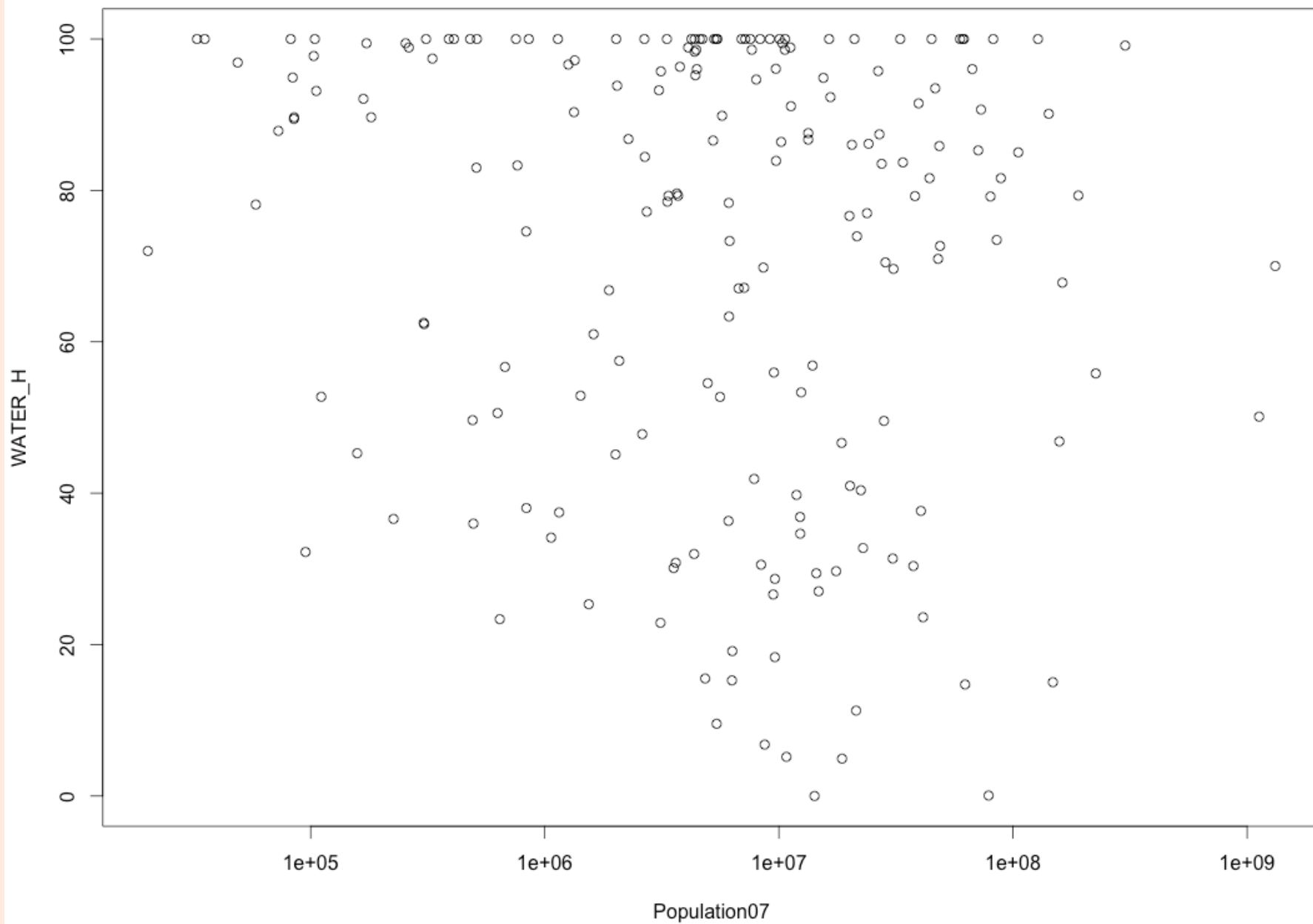


Histogram of Population07



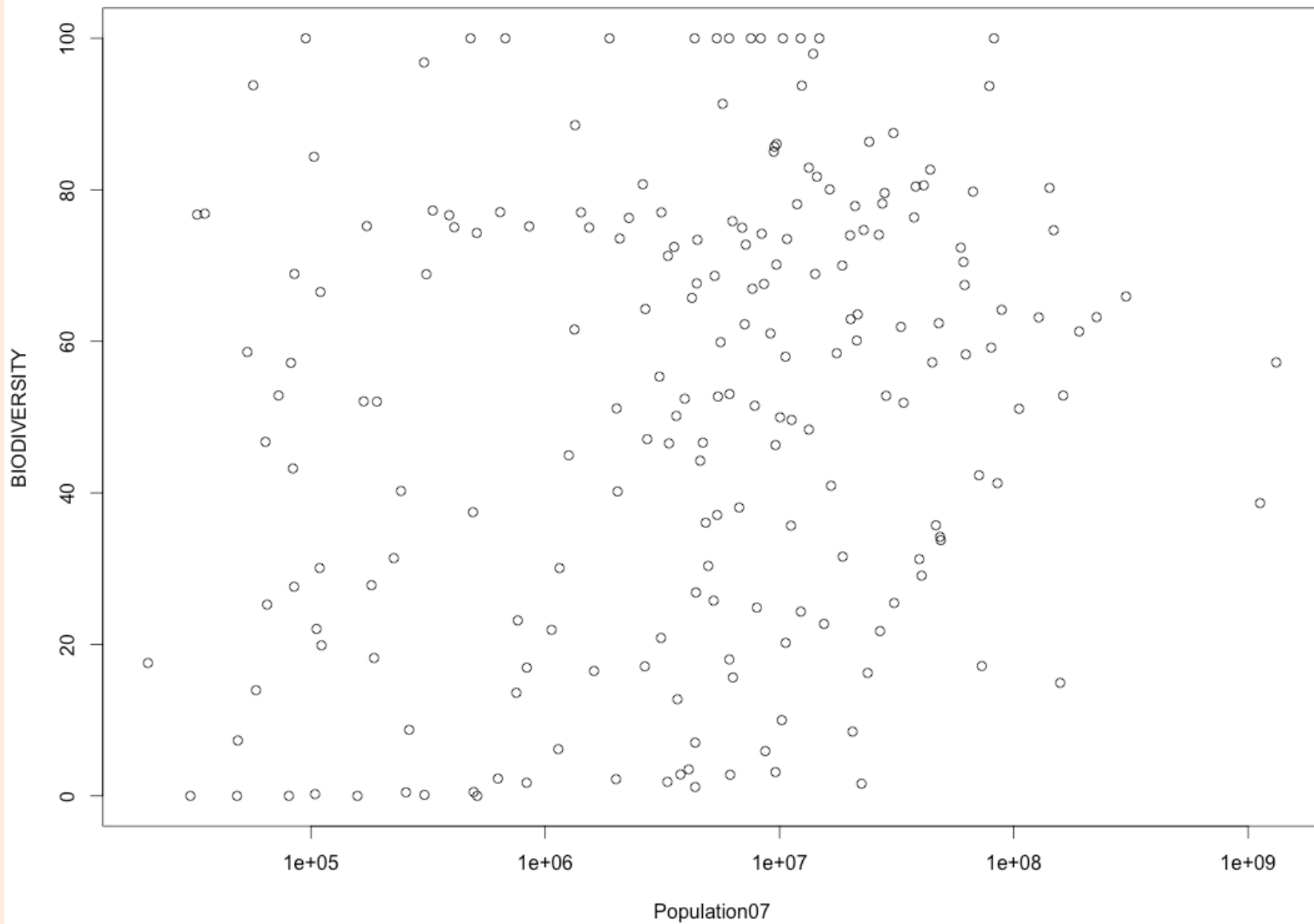


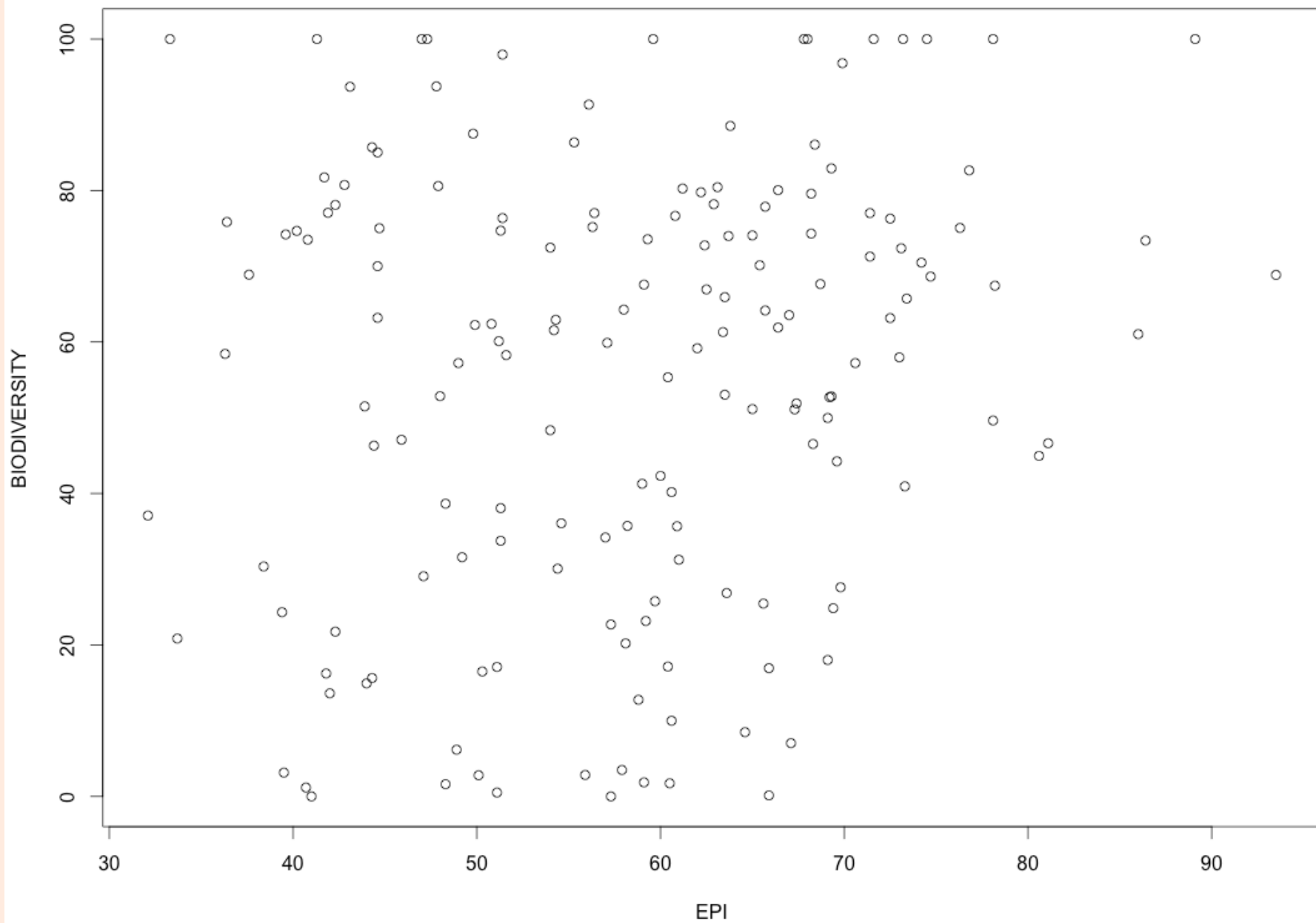




More munging

- Bad values, outliers, corrupted entries, thresholds ...
- Noise reduction – low-pass filtering, binning
- REMEMBER: when you munge you MUST record what you did (and why) and save copies of pre- and post- operations...






Populations within populations


- In the EPI example:
 - Geographic regions (GEO_subregion)
 - EPI_regions
 - Eco-regions
 - Primary industry(ies)
 - Climate region
- What would you do to start exploring?



Environmental Performance Index (EPI)


<https://sedac.ciesin.columbia.edu/data/collection/eipi>

Environmental Performance Index (EPI)







 Other DAACs ▾

**SOCIOECONOMIC DATA AND APPLICATIONS CENTER (SEDAC)**
A Data Center in NASA's Earth Observing System Data and Information System (EOSDIS) — Hosted by CIESIN at Columbia University

 Search SEDAC... **Data** 🔍 

 **DATA** ▾ **MAPS** ▾ **THEMES** ▾ **RESOURCES** ▾ **SOCIAL MEDIA** ▾ **ABOUT** ▾ **HELP**

Environmental Performance Index (EPI)

Follow Us:     | Share:  

Collection Overview

- Data Sets** (8)
- Map Gallery** (89)
- Map Services** (64)
- Citations**
- FAQs**
- Project Website**

Introduction

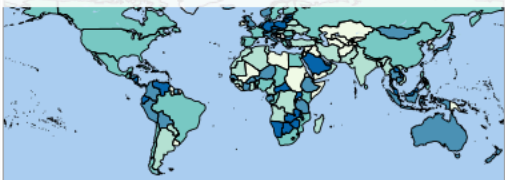
This collection includes an archive of 2006, 2008, 2010, 2012, 2014, 2016, 2018 and 2020 Environmental Performance Index (EPI) reports and data sets.

The EPI utilizes a proximity-to-target methodology focused on a core set of environmental outcomes linked to policy goals. Trend data for each indicator, where available, are provided for the 2012 and 2014 versions of the EPI, along with trend EPI scores.


Communications about the EPI may be directed to the addresses below.

The EPI is produced by a consortium, including the following groups:

EPI 2014: Ecosystem Vitality Objective - Biodiversity and Habitat



8 of 64



<https://sedac.ciesin.columbia.edu/data/collection/eipi>

EPI	ENVHEALTH	ECOSYSTEM	DALY	AIR_H	WATER_H	AIR_E	WATER_E	BIODIVERSITY	FORESTRY	FISHERIES	AGRICULTURE	CLIMATE	DALY
NA	NA	NA	NA	NA	100.00	33.13	NA	0.23	100.00	92.86	40.00	NA	NA
NA	11.55	NA	0.00	35.49	10.72	72.03	57.43	3.11	22.63	NA	39.59	NA	0.000
36.3	18.29	54.40	0.00	43.47	29.70	40.13	64.76	58.43	94.79	86.74	54.55	53.85	0.000
NA	NA	NA	NA	NA	NA	86.54	NA	0.26	100.00	NA	40.00	NA	NA
71.4	69.93	72.92	65.50	52.97	95.73	49.16	91.24	77.02	100.00	62.54	54.55	68.97	65.50
NA	90.21	NA	84.77	91.28	100.00	52.41	NA	57.16	100.00	NA	40.00	NA	84.77
NA	NA	NA	NA	79.04	NA	18.19	NA	52.05	100.00	68.27	40.00	NA	NA
40.7	81.29	0.06	89.10	48.63	98.32	34.00	5.27	1.17	100.00	50.00	39.09	20.58	89.09
61.0	74.49	47.60	71.63	63.21	91.50	48.24	72.91	31.25	82.81	48.08	95.45	49.58	71.63
60.4	71.63	49.26	62.31	68.66	93.23	61.96	50.52	55.34	63.02	NA	92.84	49.85	62.31
NA	NA	NA	NA	NA	NA	46.47	NA	25.25	95.02	94.36	40.00	NA	NA
69.8	83.21	56.38	73.01	97.37	89.44	36.03	70.87	27.61	100.00	75.54	54.55	61.88	73.01
65.7	91.73	39.58	84.77	97.37	100.00	29.46	57.95	77.86	96.89	96.53	97.19	27.64	84.77
78.1	89.47	66.80	86.86	84.15	100.00	39.83	97.55	100.00	100.00	NA	84.54	50.07	86.86
59.1	62.72	55.43	57.61	65.86	69.81	54.33	51.40	67.57	100.00	NA	44.13	58.09	57.60
43.9	22.56	65.25	4.09	40.17	41.89	51.96	69.90	51.50	0.00	NA	100.00	78.02	4.089
58.1	89.05	27.09	80.96	94.30	100.00	21.44	56.53	20.20	100.00	48.74	70.00	36.65	80.95
39.6	22.94	56.24	16.40	28.43	30.55	50.30	68.62	74.20	22.24	91.50	99.80	54.71	16.39
47.3	12.24	82.33	4.94	12.04	27.04	58.68	59.90	100.00	79.12	NA	81.00	76.31	4.939
44.0	32.33	55.57	39.85	2.78	46.84	42.98	79.96	14.92	87.64	26.02	54.55	70.72	39.85
62.5	73.21	51.85	65.50	63.26	98.58	41.33	68.68	66.94	100.00	93.20	95.44	39.92	65.50

Or, a twist – n=1 but many attributes?

EPI	ENVHEALTH	ECOSYSTEM	DALY	AIR_H	WATER_H	AIR_E	WATER_E	BIODIVERSITY	FORESTRY	FISHERIES	AGRICULTURE	CLIMATE	DALY
NA	NA	NA	NA	NA	100.00	33.13	NA	0.23	100.00	92.86	40.00	NA	NA
NA	11.55	NA	0.00	35.49	10.72	72.03	57.43	3.11	22.63	NA	39.59	NA	0.000
36.3	18.29	54.40	0.00	43.47	29.70	40.13	64.76	58.43	94.79	86.74	54.55	53.85	0.000
NA	NA	NA	NA	NA	NA	86.54	NA	0.26	100.00	NA	40.00	NA	NA
71.4	69.93	72.92	65.50	52.97	95.73	49.16	91.24	77.02	100.00	62.54	54.55	68.97	65.50
NA	90.21	NA	84.77	91.28	100.00	52.41	NA	57.16	100.00	NA	40.00	NA	84.77
NA	NA	NA	NA	79.04	NA	18.19	NA	52.05	100.00	68.27	40.00	NA	NA
40.7	81.29	0.06	89.10	48.63	98.32	34.00	5.27	1.17	100.00	50.00	39.09	20.58	89.09
61.0	74.49	47.60	71.63	63.21	91.50	48.24	72.91	31.25	82.81	48.08	95.45	49.58	71.63
60.4	71.63	49.26	62.31	68.66	93.23	61.96	50.52	55.34	63.02	NA	92.84	49.85	62.31
NA	NA	NA	NA	NA	NA	46.47	NA	25.25	95.02	94.36	40.00	NA	NA
69.8	83.21	56.38	73.01	97.37	89.44	36.03	70.87	27.61	100.00	75.54	54.55	61.88	73.01
65.7	91.73	39.58	84.77	97.37	100.00	29.46	57.95	77.86	96.89	96.53	97.19	27.64	84.77
78.1	89.47	66.80	86.86	84.15	100.00	39.83	97.55	100.00	100.00	NA	84.54	50.07	86.86
59.1	62.72	55.43	57.61	65.86	69.81	54.33	51.40	67.57	100.00	NA	44.13	58.09	57.60
43.9	22.56	65.25	4.09	40.17	41.89	51.96	69.90	51.50	0.00	NA	100.00	78.02	4.089
58.1	89.05	27.09	80.96	94.30	100.00	21.44	56.53	20.20	100.00	48.74	70.00	36.65	80.95
39.6	22.94	56.24	16.40	28.43	30.55	50.30	68.62	74.20	22.24	91.50	99.80	54.71	16.39
47.3	12.24	82.33	4.94	12.04	27.04	58.68	59.90	100.00	79.12	NA	81.00	76.31	4.939
44.0	32.33	55.57	39.85	2.78	46.84	42.98	79.96	14.92	87.64	26.02	54.55	70.72	39.85
62.5	73.21	51.85	65.50	63.26	98.58	41.33	68.68	66.94	100.00	93.20	95.44	39.92	65.50

The item of interest in relation to its attributes

Summary: exploration

- Going from preliminary to initial analysis...
- Determining if there is one or more common distributions involved – i.e. parametric statistics (assumes or asserts a probability distribution)
- Fitting that distribution -> provides a model!
- Or NOT
 - A hybrid model or
 - Non-parametric (statistics) approaches are needed – more on this to come

Summary

- Cyber and Human data; quality, uncertainty and bias – you will often spend a lot of time with the data
- Distributions – the common and not-so common ones and how cyber and human data can have distinct distributions
- How simple statistical distributions can mislead us
- Populations and samples and how inferential statistics will lead us to model choices (no we have not actually done that yet in detail)
- Munging toward exploratory analysis
- Toward models!

Reminder: finish Lab 0 and 1

- Reminder to finish the last week intro to Lab (Lab 0 and Lab 1)
- R! (how is your learning/coding in R going?) keep learning/coding...
- Create the Github repository for this class if you have not created so far and share the Data Analytics Class Github repo URL with the TA (please email TA your Github repo URL)