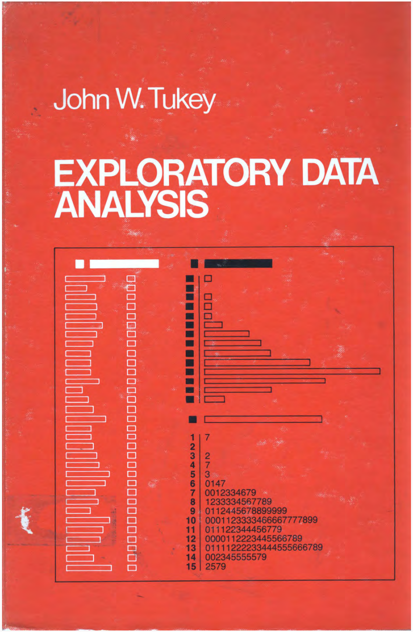
Stat 470/670 Lecture 1

What is Exploratory Data Analysis?


John W. Tukey
EXPLORATORY DATA ANALYSIS
1 | 7
2 |
3 | 2
4 | 7
5 | 3
6 | 0147
7 | 0012334679
8 | 1233334567789
9 | 0112445678899999
10 | 0001123333466667777899
11 | 011122344456779
12 | 00001122234455566789
13 | 0111122222334445555666789
14 | 002345555579
15 | 2579

We will be exploring numbers. We need to handle them easily and look at them effectively. Techniques for handling and looking — whether graphical, arithmetic, or intermediate — will be important.

Tukey, Exploratory Data Analysis (1977)

A first example: Heights of the highest points by state

```
## load required packages and data
library(tidyverse)

## -- Attaching packages ---------------------------------------
tidyverse 1.3.0 --

## v tibble  3.0.1     v dplyr   1.0.2
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## v purrr   0.3.4

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

options(tibble.print_min = 15)
heights = read_csv("highest-points-by-state.csv")

## Parsed with column specification:
## cols(
##   elevation = col_double(),
##   state = col_character()
```

A first try at looking at the data

```
heights

## # A tibble: 50 x 2
##    elevation state
##        <dbl> <chr>
##  1      733. Alabama
##  2     6168. Alaska
##  3     3851. Arizona
##  4      839. Arkansas
##  5     4418. California
##  6     4399. Colorado
##  7      725. Connecticut
##  8      137. Delaware
##  9      105. Florida
## 10     1458. Georgia
## 11     4205. Hawaii
## 12     3859. Idaho
## 13      376. Illinois
## 14      383. Indiana
## 15      509. Iowa
## # ... with 35 more rows
```

A second try at looking at the data

```r
arrange(heights, elevation)
```

```
## # A tibble: 50 x 2
##    elevation state
##        <dbl> <chr>
##  1      105. Florida
##  2      137. Delaware
##  3      163. Louisiana
##  4      246. Mississippi
##  5      247. Rhode Island
##  6      376. Illinois
##  7      383. Indiana
##  8      472. Ohio
##  9      509. Iowa
## 10      540. Missouri
## 11      550. New Jersey
## 12      595. Wisconsin
## 13      603. Michigan
## 14      701. Minnesota
## 15      725. Connecticut
## # ... with 35 more rows
```

```r
arrange(heights, desc(elevation))
```

```
## # A tibble: 50 x 2
##    elevation state
##        <dbl> <chr>
##  1     6168. Alaska
##  2     4418. California
##  3     4399. Colorado
##  4     4392. Washington
##  5     4207. Wyoming
##  6     4205. Hawaii
##  7     4123. Utah
##  8     4011. New Mexico
##  9     4005. Nevada
## 10     3901. Montana
## 11     3859. Idaho
## 12     3851. Arizona
## 13     3426. Oregon
## 14     2667. Texas
## 15     2207. South Dakota
## # ... with 35 more rows
```

Stem-and-leaf plots

Goals:

- Write down the set of numbers, keeping as much detail as possible
- Pack the numbers efficiently, so you can see all of them at once

Stem-and-leaf plots

Goals:

- Write down the set of numbers, keeping as much detail as possible
- Pack the numbers efficiently, so you can see all of them at once

These are in conflict!

Stem-and-leaf plots

Remedy:

- Notice that parts of the numbers (the beginnings) are repeated.
- The first digit of each number is printed at the beginning of the line, the remainder at the ends.
- The first digit is the “stem”, the remainder are the “leaves”.

Stem-and-leaf-plot example

Set of numbers:

```
16, 17, 17, 17, 17, 18
```

Stem-and-leaf display:

```
1 | 677778
```

Stem-and-leaf plot for the elevations in meters:

```
stem(heights$elevation)
```

```
##
##   The decimal point is 3 digit(s) to the right of the |
##
##   0 | 11222445555667778
##   1 | 0011123355566779
##   2 | 0027
##   3 | 4999
##   4 | 00122444
##   5 |
##   6 | 2
```

The stem-and-leaf plot shows that there are three groups of states:

- Alaska
- The western and Rocky Mountain states (California, Colorado, Washington, Wyoming, Hawaii, Utah, New Mexico, Nevada, Montana, Idaho, Arizona, Oregon)
- All the other states

Note 1

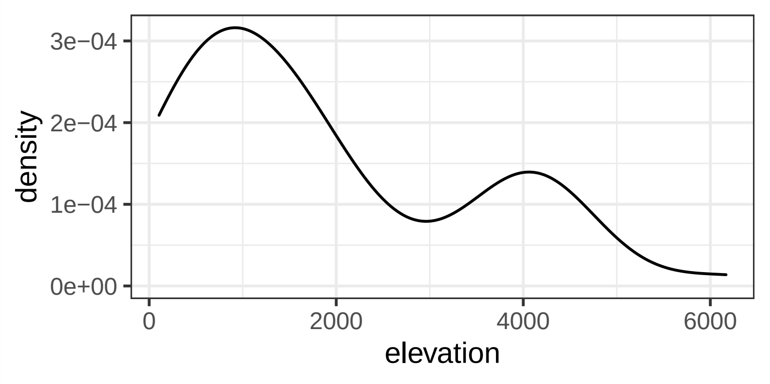
Hoosier Hill: Elevation 1257 feet

Source: google street view

Note 2

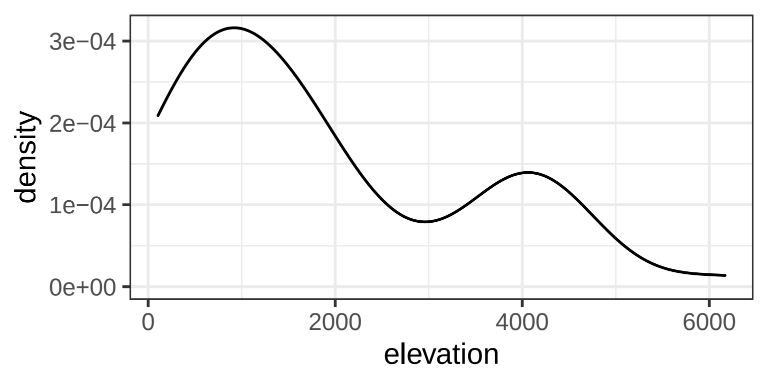
Compare the stem-and-leaf plot with a density estimate

```
ggplot(heights, aes(x = elevation)) + geom_density()
```



Compare the stem-and-leaf plot with a density estimate

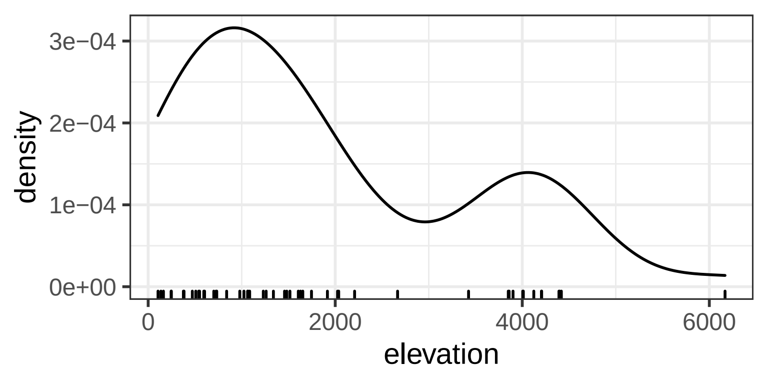
```
ggplot(heights, aes(x = elevation)) + geom_density()
```



Where is Alaska?

Compare the stem-and-leaf plot with a density estimate

```
ggplot(heights, aes(x = elevation)) + geom_density() + geom_rug()
```



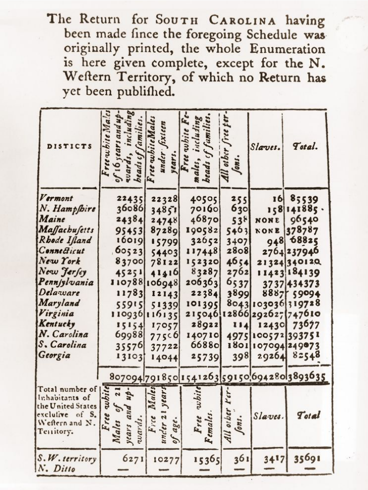
Where is Alaska?

We have made an advance in understanding this set of numbers!

We have made an advance in understanding this set of numbers!

What would traditional statistics have to say about these numbers?

What if we have a many more numbers, e.g. census data?



The Return for South Carolina having been made ſince the foregoing Schedule was originally printed, the whole Enumeration is here given complete, except for the N. Weſtern Territory, of which no Return has yet been publiſhed.

DISTICTS	Free white Males of 16 years and upwards, including heads of families.	Free white Males under ſixteen years.	Free white Females, including heads of families.	All other free perſons.	Slaves.	Total.
Vermont	22435	22328	40505	255	16	85539
N. Hampſhire	36086	34851	70160	630	158	141885
Maine	24384	24748	46870	538	NONE	96540
Maſſachuſetts	95453	87289	190582	5463	NONE	378787
Rhode Iſland	16019	15799	32652	3407	948	68825
Connecticut	60523	54403	117448	2808	2764	237946
New York	83700	78122	152320	4654	21324	340120
New Jerſey	45251	41416	83287	2762	11423	184139
Pennſylvania	110788	106948	206363	6537	3737	434373
Delaware	11783	12143	22384	3899	8887	59094
Maryland	55915	51339	101395	8043	103036	319728
Virginia	110936	116135	215046	12866	292627	747610
Kentucky	15154	17057	28922	114	12430	73677
N. Carolina	69988	77506	140710	4975	100572	393751
S. Carolina	35576	37722	66880	1801	107094	249073
Georgia	13103	14044	25739	398	29264	82548
	807094	791850	1541263	59150	694280	3893635

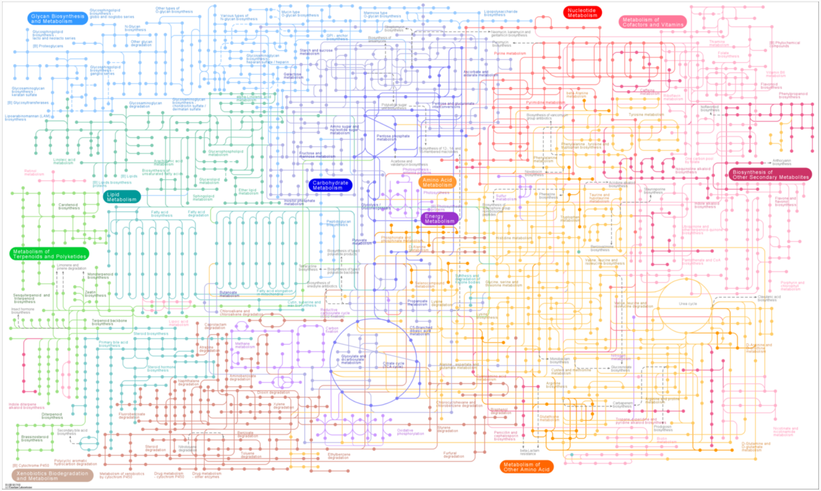
Total number of Inhabitants of the United States excluſive of S. Weſtern and N. Territory.	Free white Males of 21 years and upwards.	Free Males under 21 years of age.	Free white Females.	All other Perſons.	Slaves.	Total
S. W. territory	6271	10277	15365	361	3417	35691
N. Ditto	—	—	—	—	—	—

Source: US Census Bureau Public Information Office,
via the National Geographic Society

Or a large matrix?

Source: Still from "The Matrix"

Or graph data?



Source: KEGG PATHWAY Database

Exploratory vs. Confirmatory Analyses

Confirmatory analysis

- Probability model for the data specified before analysis takes place
- Given the probability model, test hypotheses or infer parameter values

Exploratory analysis: everything else! In particular:

- Check distributional assumptions
- Check for outliers
- Decide on variable transformations
- Decide on the form of the model: what variables to include

Exploratory analysis: everything else! In particular:

- Check distributional assumptions
- Check for outliers
- Decide on variable transformations
- Decide on the form of the model: what variables to include

BUT: Not limited to the work done before fitting a model! In the highest points example, we had an EDA-based advance that wasn't related to model fitting at all.

What does Tukey say?

Exploratory data analysis is detective work--numerical detective work--or counting detective work--or graphical detective work.

Exploratory data analysis is detective work--numerical detective work--or counting detective work--or graphical detective work.

As all detective stories remind us, many of the circumstances surrounding a crime are accidental or misleading. Equally, many of the indications to be discerned in bodies of data are accidental or misleading. To accept all appearances as conclusive would be destructively foolish, either in crime detection or in data analysis. **To fail to collect all appearances because some--or even most--are only accidents would, however, be gross misfeasance deserving (and often receiving) appropriate punishment.**

Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone--as the first step.

Tukey, *Exploratory Data Analysis* (1977) pp. 1-3

Exploratory: Collect everything that even seems to be true about the data, detective in character, “magical thinking”

Confirmatory: Given one pre-planned hypothesis, infer parameter values or test hypotheses, judicial in character, set a high bar for what we are willing to believe about the data.

The never ending data analysis cycle:

1. Get data.
2. Perform exploratory analysis to suggest a model.
3. Fit the model.
4. Perform exploratory analysis to critique the model and suggest a new model.
5. Return to step 3.

The never ending data analysis cycle:

1. Get data.
2. Perform exploratory analysis to suggest a model.
3. Fit the model.
4. Perform exploratory analysis to critique the model and suggest a new model.
5. Return to step 3.

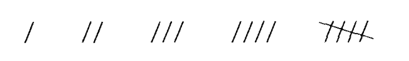
This workflow is dangerous!

- Using the data more than once
- Assiduous EDA means multiple comparison problems

Tukey's *EDA* also emphasizes tools and best practices for the practice of data analysis, all pen-and-paper based.

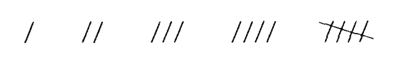
Example: Tallying

Standard method:

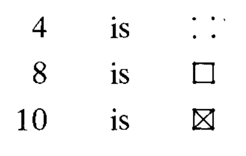


Example: Tallying

Standard method:



Tukey's proposal:



Pen-and-paper methods primarily of historical interest.

Pen-and-paper methods primarily of historical interest.

Philosophical descendants are the tidyverse packages in R.

What about this class?

What about this class?

Two categories of topics:
what to do and *how* to do it.

For *what* to do, organize by type of data:

- Univariate data
- Bivariate data
- Trivariate/Hypervariate data
- Categorical data
- Distance data
- Graph data
- Other topics according to interest

In addition:

- Dangers of EDA and how to avoid them

In the *how* to do it bin, we will learn to work with

- R
- ggplot2
- tidyverse packages

How is this class different from others?

- Machine learning: We put less emphasis on supervised learning.
- Data mining: More emphasis on visualization.
- Applied statistics: Less emphasis on p-values and inference, more flexibility in the methods used.

Texts:

- Cleveland, *Visualizing Data*
- Wickham, *ggplot2: Elemant Graphics for Data Analysis*
- Wickham and Grolemund, *R for Data Science*
- Other notes posted to the class website and canvas as necessary

Assessment:

- Homeworks (30%).
- Two mini projects (30%).
- Final project (40%).

How to succeed:

- Practice!
- Follow along with the code examples, actually type in the commands instead of copying and pasting.
- Start early on assignments and projects.
- Presentation matters – make your documents look nice enough thta you would be happy to show them to potential employers as examples of your work.

We will be exploring numbers. We need to handle them easily and look at them effectively. Techniques for handling and looking — whether graphical, arithmetic, or intermediate — will be important.

Tukey, Exploratory Data Analysis (1977)