# DATA VISUALIZATION

Abhijit Dasgupta

BIOF 439

# WHY VISUALIZE DATA?

To communicate a story

# WHY VISUALIZE DATA?

- Data-generating processes can be complex
- The data only mirrors that complexity
- We can use visualization techniques to represent that complexity in a cognitively accessible manner

# WHY VISUALIZE DATA?

- Take advantage of our human capability to understand visual patterns quickly and, often, intuitively
- Can be a gateway to interpreting what the data is saying

# “INSIGHTS”

- Exploration
- Contextual understanding
- Reporting

# "INSIGHTS"

- Exploration
- Contextual understanding
- Reporting

Data → Information → Evaluation / Action

# EARLY DAYS

- Florence Nightangle
- Message
  - Most military deaths caused by infection

# EARLY DAYS

- John Snow, 1854
- Mapping cholera outbreak in London
- Association with certain wells

# EARLY DAYS

Minard, 1869

NUMBERS DON'T TELL THE

WHOLE STORY

# TYPICAL SUMMARIES

- We compare
  - averages
  - variances
  - correlations

# SUMMARIES DON'T DIFFERENTIATE

- Anscombe (1973) created this toy examples
- Averages of x and y are the same
- Correlation between x and y are the same
- Relationships are VERY different

The Datasaurus
Matejka & Fitzmaurice, 2017

Same positive correlation overall

Negative correlation within groups

# PICTURES DON'T DIFFERENTIATE

Strip plot

Boxplot

Violin plot

Matejka & Fitzmaurice, 2017

# HOW TO THINK ABOUT DATAVIZ

# BEGIN WITH THE CONSUMER IN MIND

- You have a deep understanding of the data you're presenting
- The person seeing the visualization DOESN'T
- Develop simpler visualizations first which are easier to explain

40
20
0
-20
mean arrival delay (min.)
0
1000
2000
3000
distance (miles)
airline
American
Delta
other
# of flights
4000
8000
12000

American
Delta
US Airways
United
Endeavor
other
JetBlue
Southwest
Envoy
ExpressJet
0
4
8
12
mean arrival delay (min.)

United
JetBlue
ExpressJet
Delta
American
Envoy
US Airways
Endeavor
Southwest
other
0
10000
20000
30000
40000
50000
number of flights

# TELL A STORY

- Make sure the graphic is clear
- Make sure the main point you want to make "pops"

# TELL A STORY

Chicago
San Diego
Houston
Death Valley
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month
temperature (°F)
100
80
60
40

# TELL A STORY

# TELL A STORY

Cell type
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
Gene 7
Gene 8
Gene 9
Gene 10
Gene 11
Gene 12
Gene 13
Gene 14
Gene 15
Gene 16
Gene 17
Gene 18
Gene 19
Gene 20
Gene 21
Gene 22
Gene 23
Gene 24
Gene 25
Gene 26
Gene 27
Gene 28
Gene 29
Gene 30
Gene 31
Gene 32
Gene 33
Gene 34
Gene 35
Gene 36
Gene 37
Gene 38
Gene 39
Gene 40
Gene 41
Gene 42
Gene 43
Gene 44
Gene 45
Gene 46
Gene 47
Gene 48
Gene 49
Gene 50
Type 1 rep 1
Type 1 rep 4
Type 1 rep 2
Type 1 rep 3
Type 2 rep 3
Type 2 rep 4
Type 2 rep 1
Type 2 rep 2
18
16
14
12
10
8
6
4
Cell type
Type 1
Type 2

# WHAT MAKES A GOOD VISUALIZATION?

# WHAT MAKES A BAD VISUALIZATION?

Africa
Americas
Asia
Europe
Oceania
0
22.5
45
67.5
90
Years
CONTINENT
Africa
Americas
Asia
Europe
Oceania
Life Expectancy: 2007

MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions
$300
250
200
150
100
50
1972
'74
'76
'78
'80
'82 est.
VOTE

A
B
C
D
1
2
3
4
3D Column Charts in Excel are Awesome
4
3
2
1
0
A
B
C
D

135
130
125
120
115
110
stock price (USD)
Nov 1, 2016
Dec 1, 2016
Jan 1, 2017

150
100
50
0
stock price (USD)
Nov 1, 2016
Dec 1, 2016
Jan 1, 2017

# Comment lire ?

## EN FRANCE

★ Film succès de l'année

Part des entrées réalisées par des films français

Part des entrées réalisées par des films américains

Part des films américains dans le nombre total de films

Autres nationalités

Part des films français dans le nombre total de films

### LE CINÉMA FAIT LE PLEIN

**Le nombre d'entrées en salles a presque doublé en vingt ans.** Il est passé de 124 millions en 1994, soit 2 entrées en moyenne par personne, à 206 millions, l'équivalent de 3,1 entrées par Français.

La meilleure années des Etats-Unis

Le nombre d'entrées selon la nationalité des films varie beaucoup d'une année à l'autre, car **un grand succès peut bouleverser le classement.**

Le record de la France

Alors que les films américains ne représentent qu'un quart des films projetés, **ils concentrent près de la moitié des entrées en salle.**

### La réponse

**Non** Si les films américains accueillent toujours près de la moitié du public français, leur part a baissé de presque dix points en vingt ans.

Contrairement au petit écran, les salles de cinéma ne sont pas sujettes à des contraintes de diffusion. Cependant, les obligations faites aux chaînes de financer la production française influence l'offre cinématographique.

### SUR LE PETIT ÉCRAN...

en 2014

**42%** des films diffusés à la télévision sont français

**37%** des films diffusés à la télévision sont américains

Chiffre d'affaires du marché français de films en DVD et Blu-ray en 2015.

### DE MOINS EN MOINS DE FILMS AMÉRICAINS

Depuis vingt ans, **40 % des films sortis en salle sont français.** La part des films américains représentait environ un tiers des films dans les années 1990, c'est aujourd'hui plutôt aux alentours d'un quart.

**Méthodologie :** le nombre d'entrées correspond aux entrées payantes des salles commerciales (hors cinémathèques, festivals, ciné-clubs) pour les films projetés au moins une fois au cours de l'année. La nationalité est automatiquement française lorsqu'il s'agit d'un film ayant reçu l'agrément du CNC. Sont compris dans les films français les coproductions, y compris minoritaires. Pour les autres films, la nationalité est celle figurant sur la demande de visa d'exploitation. **Sources :** CNC, GfK, AlloCiné. **Infographie :** ASKmedia

The French are Watching More French Movies
1994
1999
2004
2009
2014
The Lion King
Titanic
Sixth Sense
French Movies
US Movies
Les Bronzes 3
Bienvenue chez les Ch'tis
Qu'est ce qu'on a fait au Bon Dieu?
Half the Market
Kaiser Fung | Junkcharts

# Domestic Seat Departures by Aircraft Type

DELTA

# Domestic Seat Departures by Aircraft Type

DELTA

# Domestic Seat Departures by Aircraft Type

DELTA

# Delta Continues to Upgauge Domestic Aircraft

Kaiser Fung / JunkCharts: 2018

## Age Distribution of Religious Groups, 2010

Figures may not add to 100% due to rounding.
Source: The Future of World Religions: Population Growth Projections, 2010-2050

PEW RESEARCH CENTER

Muslims and Hindus are younger, while Jews and Buddhists are older, relative to average
Under 15
Over 60
Muslims
Hindus
Christians
Everyone
Folk Religions
Other Religions
Jews
Buddhists
Unaffiliated
Jews
Buddhists
Other Religions
Christians
Unaffiliated
Folk Religions
Everyone
Hindus
Muslims
10
30%
20%
Proportion
Kaiser Fung | Junkcharts

# A MATTER OF PERCEPTION

$10,000
$3,000
$1,000
$300
price (USD)
0
1
2
3
carat
cut
ideal
premium
very good
good
fair

$10,000
$3,000
$1,000
$300
price (USD)
0.0
0.5
1.0
1.5
2.0
carat
cut
ideal
premium
very good
good
fair

fair
good
very good
premium
ideal
price (USD)
$10,000
$3,000
$1,000
$300
0.0
0.5
1.0
1.5
2.0
carat

# COLORS

bad
30%
20%
10%
0%
population growth
2000 to 2010
$10^6$
$3 \times 10^6$
$10^7$
$3 \times 10^7$
population size in 2000
state
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming

Nevada
Utah
Arizona
Texas
New Mexico
Florida
Alaska
Oregon
California
Montana
Tennessee
Pennsylvania
Wisconsin
Vermont
Michigan
Rhode Island
Louisiana
New York
30%
20%
10%
0%
population growth, 2000 to 2010
$10^6$
$3 \times 10^6$
$10^7$
$3 \times 10^7$
population size in 2000
region
West
South
Midwest
Northeast

# COLOR-VISION DEFICIENCY

# PRINCIPLES

# DATA-INK RATIO

# DATA-INK RATIO

- Make sure most of the “ink” reflects the data

# DATA-INK RATIO

- Make sure most of the "ink" reflects the data
  - **within reason**

# NO MENTAL GYMNASTICS

- Self-sufficiency: The graphic should be self-evident without captions
- Context should be clear (labels, axes, etc)

# NO MENTAL GYMNASTICS

- Several small graphs can be better than one complicated graph
  - small multiples
  - trellis graphs

# NO MENTAL GYMNASTICS

- Is a graph really appropriate
  - can a table be clearer
  - are you doing graphs for graphs' sake

# PRESENTING INFORMATION

- Visualizations need to
  - be clear & uncluttered
  - make the point
  - provide easy access to the underlying data "talking points" and metadata
  - allow the consumer to explore a bit and think
  - be geared towards what the customer wants

# CONSUMER FOCUS

# CONSUMER FOCUS

- What is important to the consumer?

# CONSUMER FOCUS

- What is important to the consumer?
  - Often <u>not</u> what is important to the producer

# CONSUMER FOCUS

- What is important to the consumer?
  - Often <u>not</u> what is important to the producer
- We may have a perfectly clear graphic …

# CONSUMER FOCUS

- What is important to the consumer?
  - Often <u>not</u> what is important to the producer
- We may have a perfectly clear graphic …
  - which is not what the consumer wants

# CONSUMER FOCUS

- What is important to the consumer?
  - Often <u>not</u> what is important to the producer
- We may have a perfectly clear graphic …
  - which is not what the consumer wants
  - and not what (s)he needs

# CONSUMER FOCUS

Harvard Business Review

DESIGN

## What Happens When Data Scientists and Designers Work Together

by Jon Wettersten and Dean Malmgren

MARCH 05, 2018

"Instead of a version of data science that is narrowly focused on researching new statistical models or building better data visualizations, a design-thinking approach recognizes data scientists as creative problem solvers."