DATA11001

INTRODUCTION TO DATA SCIENCE

EPISODE 3: EXPLORATORY DATA ANALYSIS
1. EXPLORATORY DATA ANALYSIS

2. VISUALIZATION
MEET MR DATA SCHEINTIST

TOOL #1: AN EYEBALL

TOOL #2: A MOUTH
EXPLORATORY ANALYSIS

• "[P]rocedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

  (John Tukey, The Future of Data Analysis, July 1961)

• Tukey found that too much emphasis was put on hypothesis testing (confirmatory analysis) and that a new kind of analysis was needed where hypotheses are suggested by the data

• The development of tools like S, S-Plus, and R was motivated by Tukey's ideas
EXPLORATORY ANALYSIS

• **LOOK AT THE DATA!**

• Big problems have been caused by not looking

• If you want to be sure, go look at the *actual thing* rather than the data (check cables connections, etc.)

• Second best would be to look at the raw data

• The further you get from the source, the higher the risk that you see artefacts of the pre-processing
NON-GRAphICAL EDA

• Again: **LOOK AT THE DATA**

• Raw data:

```
player_name, height, weight, p_id, id, player_fifa_api_id
"Abel Tamata", 182.88, 170, 240556, 1031, 202153, 240556,
"Abel", 177.8, 165, 40938, 1046, 17880, 40938, "2010-08-30 00
"Abella Perez Damia", 187.96, 174, 37422, 1051, 159580, 37
"Abiola Dauda", 180.34, 165, 114503, 1076, 187175, 114503,
"Abou Diaby", 193.04, 168, 27277, 1092, 163423, 27277, "201
"Aboubacar Tandia", 193.04, 185, 181344, 1119, 138675, 181
"Aboubakar Kamara", 177.8, 168, 581141, 1122, 225541, 5811
```

• Observations:
  – height in metric units [cm] but weight in imperial [lb]
  – name can be 1–3 words (at least)
NON-GRA P HICAL EDA

- Again: **LOOK AT THE DATA**

- Raw data:

```plaintext
70,right,,_0,60,50,60,74,,74,,53,62,73,74,70,,63,,63,right,,_0,49,64,65,54,58,58,46,44,27,64,69,54,66,500",63,72,left,medium,medium,42,52,39,64,40,72,67,4",56,62,right,high,medium,54,37,48,62,31,56,35,39,563,right,medium,medium,57,53,58,69,57,59,69,68,66,64,,right,,_0,56,34,70,70,,56,,26,71,60,69,68,,70,,72,,",61,65,right,,_0,42,60,54,52,,51,,42,35,60,74,77,,5,,left,medium,medium,42,30,61,59,37,47,41,32,67,50,67
```

- Observations:
  - height in metric units \([\text{cm}]\) but weight in imperial \([\text{lb}]\)
  - name can be 1–3 words (at least)
  - missing data
  - what are these '0'?
NON-GRAphical EDA

- For table-formatted data (csv), a spreadsheet can be useful (but mind the formatting)
**NON-GRAphical EDA**

- For table-formatted data (csv), a spreadsheet can be useful (but mind the formatting)

<table>
<thead>
<tr>
<th></th>
<th>Name</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>291</td>
<td>Alberto Fontana</td>
<td>185.42</td>
<td>161</td>
<td>39425</td>
<td>484</td>
</tr>
<tr>
<td>292</td>
<td>Alberto Frison,18</td>
<td>190.5</td>
<td>187</td>
<td>30143</td>
<td>484</td>
</tr>
<tr>
<td>293</td>
<td>Alberto Garcia</td>
<td>182.88</td>
<td>170</td>
<td>102394</td>
<td>487</td>
</tr>
<tr>
<td>294</td>
<td>Alberto Gilardino</td>
<td>182.88</td>
<td>174</td>
<td>30881</td>
<td>488</td>
</tr>
<tr>
<td>295</td>
<td>Alberto Giuliatto</td>
<td>180.34</td>
<td>170</td>
<td>42460</td>
<td>492</td>
</tr>
<tr>
<td>296</td>
<td>Alberto Grassi</td>
<td>182.88</td>
<td>165</td>
<td>575364</td>
<td>493</td>
</tr>
<tr>
<td>297</td>
<td>Alberto Guitian</td>
<td>182.88</td>
<td>163</td>
<td>543020</td>
<td>494</td>
</tr>
<tr>
<td>298</td>
<td>Alberto Lopo</td>
<td>185.42</td>
<td>179</td>
<td>37451</td>
<td>496</td>
</tr>
<tr>
<td>299</td>
<td><strong>Alberto Luque,21</strong></td>
<td>182.88</td>
<td>176</td>
<td>32763</td>
<td>498</td>
</tr>
<tr>
<td>300</td>
<td>Alberto Maria Fontana</td>
<td>187.96</td>
<td>183</td>
<td>39743</td>
<td>499</td>
</tr>
<tr>
<td>301</td>
<td>Alberto Moreno</td>
<td>170.18</td>
<td>143</td>
<td>314605</td>
<td>500</td>
</tr>
</tbody>
</table>
For table-formatted data (csv), a spreadsheet can be useful (but mind the formatting)

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>63</td>
<td>68</td>
<td>right</td>
<td>medium</td>
<td>high</td>
<td>36</td>
</tr>
<tr>
<td>62</td>
<td>66</td>
<td>right</td>
<td>le</td>
<td>ean</td>
<td>61</td>
</tr>
<tr>
<td>71</td>
<td>76</td>
<td>right</td>
<td>medium</td>
<td>medium</td>
<td>70</td>
</tr>
<tr>
<td>73</td>
<td>73</td>
<td>right</td>
<td>low</td>
<td>high</td>
<td>57</td>
</tr>
<tr>
<td>70</td>
<td>70</td>
<td>right</td>
<td>medium</td>
<td>medium</td>
<td>63</td>
</tr>
<tr>
<td>73</td>
<td>76</td>
<td>left</td>
<td>medium</td>
<td>low</td>
<td>47</td>
</tr>
<tr>
<td>69</td>
<td>71</td>
<td>right</td>
<td>medium</td>
<td>medium</td>
<td>68</td>
</tr>
<tr>
<td>61</td>
<td>70</td>
<td>right</td>
<td>medium</td>
<td>medium</td>
<td>58</td>
</tr>
<tr>
<td>68</td>
<td>71</td>
<td>right</td>
<td>high</td>
<td>medium</td>
<td>63</td>
</tr>
<tr>
<td>73</td>
<td>75</td>
<td>right</td>
<td>None</td>
<td>o</td>
<td>73</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>right</td>
<td>high</td>
<td>medium</td>
<td>82</td>
</tr>
</tbody>
</table>
EXPLORATORY DATA ANALYSIS (EDA)

• There are various classes of EDA:
  – non-graphical vs graphical
  – univariate vs bivariate vs multivariate
  – etc

• You should use them all

• The basic goals are:
  1. remove or correct erroneous data
  2. formulate initial hypotheses
  3. choose suitable analysis methods
WHY EDA

• Any suspicious data need to be investigated and corrected/removed

• This can be quite tricky: is it an outlier or a key observation?

• After EDA, it may be easier to formulate hypotheses, and to make an informed choice of the analysis approach

• E.g.:
  – choose between linear vs non-linear methods
  – decide whether clustering would be helpful
### SUMMARY STATISTICS

- Simple summary statistics:
  - **continuous**: average, median, min, max
  - **categorical**: set of values, mode

```r
D <- read.csv("player_stats.csv")
summary(D)
```

```
player_name        height          weight
Danilo  :    7   Min.   :157.5   Min.   :117.0   Min.   :  2625
Paulinho:    6   1st Qu.:177.8   1st Qu.:159.0   1st Qu.: 35558
Ricardo :    5   Median :182.9   Median :168.0   Median : 96622
Adriano :    4   Mean   :181.9   Mean   :168.4   Mean   :156573
Douglas :    4   3rd Qu.:185.4   3rd Qu.:179.0   3rd Qu.:212442
Felipe  :    4   Max.   :208.3   Max.   :243.0   Max.   :750584
(Other) :11034

id         player_fifa_api_id player_api_id
Min.   :     1   Min.   :     2     Min.   :  2625
1st Qu.: 46173   1st Qu.:151895     1st Qu.: 35558
Median : 92080   Median :184705     Median : 96622
Mean   : 92278   Mean   :165686     Mean   :156573
3rd Qu.:138874   3rd Qu.:203884     3rd Qu.:212442
Max.   :183969   Max.   :234141     Max.   :750584
```
**SUMMARY STATISTICS**

- Simple summary statistics:
  - continuous variables: *average, median, min, max*
  - categorical variables: *set of values, mode*

<table>
<thead>
<tr>
<th></th>
<th>attacking_work_rate</th>
<th>defensive_work_rate</th>
<th>crossing</th>
</tr>
</thead>
<tbody>
<tr>
<td>medium</td>
<td>6967</td>
<td>7311</td>
<td>Min. : 6.00</td>
</tr>
<tr>
<td>high</td>
<td>2375</td>
<td>1555</td>
<td>1st Qu. :44.00</td>
</tr>
<tr>
<td>low</td>
<td>565</td>
<td>1041</td>
<td>Median :58.00</td>
</tr>
<tr>
<td>_0</td>
<td>544</td>
<td>540</td>
<td>Mean : 53.98</td>
</tr>
<tr>
<td>(Other) :</td>
<td>544</td>
<td>540</td>
<td>3rd Qu. :67.00</td>
</tr>
<tr>
<td>norm</td>
<td>495</td>
<td>278</td>
<td>Max. : 92.00</td>
</tr>
<tr>
<td>(Other) :</td>
<td>52</td>
<td>273</td>
<td>NA's : 4</td>
</tr>
</tbody>
</table>
WHY TABLES ARE MUCH BETTER THAN GRAPHS (APRIL 1ST, 2009)

"And I recommend using Excel, which has some really nice defaults as well as options such as those 3-D colored bar charts."

(Andrew Gelman, "Why Tables are Really Much Better than Graphs", Journal of Computational and Graphical Statistics, Volume 20, Number 1, Pages 3–7)

<table>
<thead>
<tr>
<th>Method</th>
<th>Saturated fat (≤ 30 g/day; ≥ 30 g/day)</th>
<th>Total energy intake (1,000 kcal/day)</th>
<th>Alcohol (20 g/day; ≥ 20 g/day)</th>
<th>Age (10 years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC</td>
<td>.94 (.75–1.17)</td>
<td>.95 (.78–1.16)</td>
<td>1.35 (1.05–1.74)</td>
<td>1.60 (1.43–1.87)</td>
</tr>
<tr>
<td>RC</td>
<td>.78 (.33–1.80)</td>
<td>.80 (.27–2.38)</td>
<td>1.57 (1.09–2.28)</td>
<td>1.54 (1.42–1.67)</td>
</tr>
<tr>
<td>ML-WICI</td>
<td>.65 (.18–2.44)</td>
<td>.89 (.26–3.05)</td>
<td>1.55 (1.06–2.27)</td>
<td>1.53 (1.33–1.76)</td>
</tr>
<tr>
<td>ML-WGCI</td>
<td>.65 (.22–1.99)</td>
<td>.89 (.26–3.01)</td>
<td>1.55 (1.05–2.27)</td>
<td>1.53 (1.33–1.77)</td>
</tr>
<tr>
<td>ML-PLCI</td>
<td>.65 (0–2.13)</td>
<td>.89 (.19–3.61)</td>
<td>1.55 (1.07–2.60)</td>
<td>1.53 (1.25–1.75)</td>
</tr>
<tr>
<td>ML-RbCI</td>
<td>.65 (.13–3.22)</td>
<td>.89 (.24–3.30)</td>
<td>1.55 (1.05–2.28)</td>
<td>1.53 (1.34–1.75)</td>
</tr>
<tr>
<td>GEE a*-RbCI</td>
<td>.71 (.21–2.38)</td>
<td>.88 (.25–3.15)</td>
<td>1.53 (1.05–2.25)</td>
<td>1.53 (1.38–1.75)</td>
</tr>
<tr>
<td>GEE b-RbCI</td>
<td>.69 (.15–3.10)</td>
<td>.93 (.20–4.42)</td>
<td>1.52 (1.01–2.30)</td>
<td>1.55 (1.37–1.76)</td>
</tr>
<tr>
<td>GEE c-RbCI</td>
<td>.69 (.19–2.44)</td>
<td>.94 (.24–3.71)</td>
<td>1.52 (1.03–2.23)</td>
<td>1.55 (1.38–1.75)</td>
</tr>
</tbody>
</table>

WHY TABLES ARE MUCH BETTER THAN GRAPHS (APRIL 1ST, 2009)

- "And I recommend using Excel, which has some really nice defaults as options such as those 3-D colored bar charts."

(Andrew Gelman, "Why Tables are Really Much Better than Graphs," *Journal of Computational and Graphical Statistics*, Volume 20, Number 1, Pages 3–7)

<table>
<thead>
<tr>
<th>Method</th>
<th>Saturated fat ($\leq 30$ g/day; $\geq 30$ g/day)</th>
<th>Total energy intake ($1,000$ kcal/day)</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC</td>
<td>$0.94 (0.75–1.17)$</td>
<td>$0.95 (0.78–1.16)$</td>
<td>1.</td>
</tr>
<tr>
<td>RC</td>
<td>$0.78 (0.33–1.80)$</td>
<td>$0.80 (0.27–2.38)$</td>
<td>1.</td>
</tr>
<tr>
<td>ML-WICI</td>
<td>$0.65 (0.18–2.44)$</td>
<td>$0.89 (0.26–3.05)$</td>
<td>1.</td>
</tr>
<tr>
<td>ML-WGCI</td>
<td>$0.65 (0.22–1.99)$</td>
<td>$0.89 (0.26–3.01)$</td>
<td>1.</td>
</tr>
<tr>
<td>ML-PLCI</td>
<td>$0.65 (0.02–2.13)$</td>
<td>$0.89 (0.19–3.61)$</td>
<td>1.</td>
</tr>
<tr>
<td>ML-RbCI</td>
<td>$0.65 (0.13–3.22)$</td>
<td>$0.89 (0.24–3.30)$</td>
<td>1.</td>
</tr>
<tr>
<td>GEE a*-RBCI</td>
<td>$0.71 (0.21–2.38)$</td>
<td>$0.88 (0.25–3.15)$</td>
<td>1.</td>
</tr>
<tr>
<td>GEE b-RbCl</td>
<td>$0.69 (0.15–3.10)$</td>
<td>$0.93 (0.20–4.42)$</td>
<td>1.</td>
</tr>
<tr>
<td>GEE c-RbCl</td>
<td>$0.69 (0.19–2.44)$</td>
<td>$0.94 (0.24–3.71)$</td>
<td>1.</td>
</tr>
</tbody>
</table>

CHOOSING THE RIGHT KIND OF CHART

How to Choose the Right Chart for Your Data

- Variable Width Column Chart
- Table or Table with Embedded Charts
- Bar Chart
- Column Chart
- Circular Area Chart
- Line Chart
- Column Chart
- Line Chart

- Two Variables per Item
- Many Categories
- Many Items
- Few Items
- Cyclical Data
- Non-Cyclical Data
- Single or Few Categories
- Many Categories

- Few Categories
- Many Periods
- Few Periods

- Among Items
- Over Time

- Scatter Chart
- Bubble Chart

- Two Variables
- Three Variables

- Relationship

- Comparison

- What would you like to show?

- Distribution

- Composition

- Changing over Time
- Static

- Few Periods
- Many Periods

- Only Relative Differences Matter
- Relative and Absolute Differences Matter
- Only Relative Differences Matter
- Relative and Absolute Differences Matter

- Simple Share of Total

- Accumulation or Subtraction to Total

- Components of Components

- Stacked 100% Column Chart
- Stacked Column Chart
- Stacked 100% Area Chart
- Stacked Area Chart
- Pie Chart
- Waterfall Chart
- Stacked 100% Column Chart with Subcomponents

© A. Abela
• Distribution of a single quantity: histogram, boxplot

```r
D <- read.csv("player_stats.csv")
hist(D$weight)
boxplot(D$weight)
```
ON HISTOGRAMS

- Choosing the bin width may change how things look

```r
> par(mfrow=c(3,1), mar=c(2.5,2.5,2.5,0.5))
> hist(D$height, breaks=10, col='orange')
> hist(D$height, breaks=20, col='orange')
```
GRAPHICAL EDA

• relationship between two variables: scatter plot

```r
> D <- read.csv("player_stats.csv")
> plot(D$weight, D$height)
```
• relationship between two variables: scatter plot

• more variables can be included using color and size

red = left footed
green = right footed
size = overall rating
Graphical EDA

- relationship between two variables: scatter plot
- more variables can be included using color and size

Additional remarks (thanks to students!)

On why red and green are a bad choice

1. Color blindness
2. Green vs red also carries the connotation good vs bad for most people
GRAPHICAL EDA

- A majority of graphical EDA is 1D or 2D

```python
import seaborn as sns
import matplotlib.pyplot as plt
sns.set(style="white")

df = sns.load_dataset("iris")

g = sns.PairGrid(df, diag_sharey=False)
g.map_lower(sns.kdeplot, cmap="Blues_d")
g.map_upper(plt.scatter)
g.map_diag(sns.kdeplot, lw=3)
```

examples on this page from: "seaborn: statistical data visualization", Michael Waskom
**GRAPHICAL EDA**

- A majority of graphical EDA is 1D or 2D

```python
import seaborn as sns
import matplotlib.pyplot as plt
sns.set(style="white")
df = sns.load_dataset("iris")
g = sns.PairGrid(df, diag_sharey=False)
g.map_lower(sns.kdeplot, cmap="Blues_d")
g.map_upper(plt.scatter)
g.map_diag(sns.kdeplot, lw=3)
```

**PYTHON MATPLOTLIB**
GRAPHICAL EDA

- Scatter plot: seeing structure

LOOKS LIKE CLUSTERS
ABOUT STATISTICAL GRAPHICS

• Visualization techniques and even the details can be crucial in data analysis and communication of the findings

• Darrell Huff, *How to Lie with Statistics*, 1954

TUFTE, 1983

• Principles of graphical **integrity:**
  1. Representation (length, area, ...) should be directly proportional to the number
  2. Clear labeling on the graphic itself – avoid legends
  3. Show data variation, not design variation
  4. Apply inflation-adjustment for money
  5. The number of information-carrying dimensions depicted should not exceed the number of dimensions in the data.

• Principles of graphical **excellence:**
  1. **Fundamental principle:** Above all else show the data.
  2. Maximize the data-ink ratio.
  3. Erase non-data ink.
  4. Erase redundant data ink.
  5. Revise and edit.
TUFTÉ, 1983

- data-ink

"data–ink ratio = ____________________________

total ink used to print the graphic

= proportion of a graphic's ink devoted to the non-redundant display of data information "

TUFTE, 1983
THE GOOD, THE BAD, AND THE UGLY

#1

- Apparently "cosmetic" things such as line weight and decorations (borders, labels, etc) can make all the difference.

examples on this page from: Introduction to R Graphics with ggplot2, Harvard University
THE GOOD, THE BAD, AND THE UGLY

#1

- Apparently "cosmetic" things such as line weight and decorations (borders, labels, etc) can make all the difference

examples on this page from: Introduction to R Graphics with ggplot2, Harvard University
THE GOOD, THE BAD, AND THE UGLY

#2 & #3

- Even though you can produce a really cool visualization, it's not necessarily the best way to show the data or convey the idea.

examples on this page from: English 419: Multimedia writing
4. Apply inflation-adjustment for money
5. "Representation (length, area, ...) should be directly proportional to the number" (Tufte, 1983)

- If you map a one-dimensional feature (such as money, weight) to 2D, the area (not the diameter) should be directly proportional to the feature.

from: (Darrell Huff, 1954)
THE GOOD, THE BAD, AND THE UGLY
#6

- 16 books is *much* better than five books, right?

UNDER PRESIDENT OBAMA, MORE STUDENTS ARE EARNING THEIR HIGH SCHOOL DIPLOMAS THAN EVER BEFORE

HIGH SCHOOL GRADUATION RATE

- 2007-08: 75%
- 2008-09: 75%
- 2009-10: 78%
- 2010-11: 79%
- 2011-12: 80%
- 2012-13: 81%
- 2013-14: 82%

SOURCE: U.S. DEPARTMENT OF EDUCATION, NATIONAL CENTER FOR EDUCATION STATISTICS

#LeadOnEducation
THE GOOD, THE BAD, AND THE UGLY

#6

• 16 books is much better than five books, right?

1. Representation (length, area, ...) should be directly proportional to the number

• Therefore, column (bar) charts should always start at zero!

From: "The most misleading charts of 2015, fixed", qz.com
THE GOOD, THE BAD, AND THE UGLY

#6

- 16 books is much better than five books, right?

High school graduation rates in the US, 1975 to 2012

- Carter: -2.2% points
- Reagan: +1.2
- Bush: +0.4
- Clinton: -2.2
- Bush: +3.8
- Obama: +5.3

Class of '75

'80 '85 '90 '95 '00 '05 '10

Data: National Center for Education Statistics

examples on this page from: "The most misleading charts of 2015, fixed", qz.com
THE GOOD, THE BAD, AND THE UGLY

THE GOOD, THE BAD, AND THE UGLY
#7

TUFTE, 1983: "THIS MAY WELL BE THE WORST GRAPHIC EVER TO FIND ITS WAY TO PRINT." 

- data = five numbers
- area not directly proportional to number
- arbitrary interpolation/smoothing
- a lot of ink!
- colors, 3D & mirroring completely redundant
THE GOOD, THE BAD, AND THE UGLY

#8

- It's not just about the pretty pictures
THE GOOD, THE BAD, AND THE UGLY
#8

- It's not just about the pretty pictures
THE GOOD, THE BAD, AND THE UGLY

#8

• It's not just about the pretty pictures

Los Angeles Times

Shuttle Explodes; All 7 Die
Teacher on Board as Challenger Blows Up on Liftoff

Reagan Postpones Future Flights Pending a Probe

By MICHAEL SEILER and PETER H. KING,
Times Staff Writers

KENNEDY SPACE CENTER, Fla.—The space shuttle Challenger exploded in a huge fireball less than two minutes after takeoff today, with all seven crew members—including New Hampshire teacher Sharon Christa McAuliffe—on board.

Airborne paramedics parachuted quickly into the calf waters off Cape Canaveral in a vain search for survivors. Though there was no immediate announcement on the fate of the crew, all were believed dead.

The disaster—the worst in the history of America’s manned space program—came shortly after the Challenger blasted off on a cold Florida morning on the 25th shuttle flight.

As the contrails formed to the west of Cape Canaveral, a parachute appeared in the clear blue sky, giving spectators a small glimpse...
THE GOOD, THE BAD, AND THE UGLY

#8

- It's not just about the pretty pictures
ONE MORE TIP

• Transparency or jitter helps with overlapping points

```R
> D <- read.csv("player_stats.csv")
> library(ggplot2)
> p = ggplot(D, aes(height, weight))
> p = p + geom_point(color='red', alpha=.03)
> p
```