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:

<table>
<thead>
<tr>
<th>player_name</th>
<th>height</th>
<th>weight</th>
<th>p_id</th>
<th>id</th>
<th>player_fifa_api_id</th>
<th>date</th>
<th>size</th>
<th>shape</th>
<th>location</th>
<th>gender</th>
<th>hair</th>
<th>hair2</th>
<th>hair3</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Abel Tamata&quot;</td>
<td>182.88</td>
<td>170</td>
<td>240556</td>
<td>1031</td>
<td>202153</td>
<td>2010-08-30</td>
<td>63</td>
<td>62</td>
<td>right</td>
<td>medium</td>
<td>medium</td>
<td>medium</td>
<td></td>
</tr>
<tr>
<td>&quot;Abella Perez Damia&quot;</td>
<td>187.96</td>
<td>174</td>
<td>37422</td>
<td>1051</td>
<td>159580</td>
<td>2011-08-30</td>
<td>56</td>
<td>52</td>
<td>right</td>
<td>high</td>
<td>medium</td>
<td>medium</td>
<td></td>
</tr>
<tr>
<td>&quot;Abiola Dauda&quot;</td>
<td>180.34</td>
<td>165</td>
<td>114503</td>
<td>1076</td>
<td>187175</td>
<td>2012-08-30</td>
<td>42</td>
<td>52</td>
<td>medium</td>
<td>medium</td>
<td>medium</td>
<td>medium</td>
<td></td>
</tr>
<tr>
<td>&quot;Abou Diaby&quot;</td>
<td>193.04</td>
<td>168</td>
<td>27277</td>
<td>1092</td>
<td>163423</td>
<td>2013-08-30</td>
<td>40</td>
<td>72</td>
<td>right</td>
<td>medium</td>
<td>medium</td>
<td>medium</td>
<td></td>
</tr>
<tr>
<td>&quot;Aboubacar Tandia&quot;</td>
<td>193.04</td>
<td>185</td>
<td>181344</td>
<td>1119</td>
<td>138675</td>
<td>2014-08-30</td>
<td>50</td>
<td>66</td>
<td>right</td>
<td>medium</td>
<td>medium</td>
<td>medium</td>
<td></td>
</tr>
<tr>
<td>&quot;Aboubakar Kamara&quot;</td>
<td>177.8</td>
<td>168</td>
<td>581141</td>
<td>1122</td>
<td>225541</td>
<td>2015-08-30</td>
<td>46</td>
<td>58</td>
<td>right</td>
<td>medium</td>
<td>medium</td>
<td>medium</td>
<td></td>
</tr>
</tbody>
</table>

• Observations:
  - height in metric units [cm] but weight in imperial [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)

<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>Alberto Luque,21</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>
NON-GRAFICAL EDA

- 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.</td>
</tr>
<tr>
<td>high</td>
<td>2375</td>
<td>1555</td>
<td>1st Qu.</td>
</tr>
<tr>
<td>low</td>
<td>565</td>
<td>1041</td>
<td>Median</td>
</tr>
<tr>
<td>norm</td>
<td>544</td>
<td>0</td>
<td>Mean</td>
</tr>
<tr>
<td>(Other)</td>
<td>52</td>
<td>_0</td>
<td>3rd Qu.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>278</td>
<td>Max.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>normal</td>
<td>NA's</td>
</tr>
<tr>
<td></td>
<td></td>
<td>66</td>
<td>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 options such as those 3-D colored bar charts."

(Andrew Gelman, "Why Tables are Really Much Better than Graphs" 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>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC</td>
<td>.94 (.75–1.17)</td>
<td>.95 (.78–1.16)</td>
<td>1.</td>
</tr>
<tr>
<td>RC</td>
<td>.78 (.63–1.80)</td>
<td>.80 (.27–2.38)</td>
<td>1.</td>
</tr>
<tr>
<td>ML-WICI</td>
<td>.65 (.18–2.44)</td>
<td>.89 (.26–3.05)</td>
<td>1.</td>
</tr>
<tr>
<td>ML-WGCI</td>
<td>.65 (.22–1.99)</td>
<td>.89 (.26–3.01)</td>
<td>1.</td>
</tr>
<tr>
<td>ML-PLCI</td>
<td>.65 (0–2.13)</td>
<td>.89 (.19–3.61)</td>
<td>1.</td>
</tr>
<tr>
<td>ML-RbCI</td>
<td>.65 (.13–3.22)</td>
<td>.89 (.24–3.30)</td>
<td>1.</td>
</tr>
<tr>
<td>GEE a*-RBCl</td>
<td>.71 (.21–2.38)</td>
<td>.88 (.25–3.15)</td>
<td>1.</td>
</tr>
<tr>
<td>GEE b-RbCl</td>
<td>.69 (.15–3.10)</td>
<td>.93 (.20–4.42)</td>
<td>1.</td>
</tr>
<tr>
<td>GEE c-RbCl</td>
<td>.69 (.19–2.44)</td>
<td>.94 (.24–3.71)</td>
<td>1.</td>
</tr>
</tbody>
</table>

Choosing the Right Kind of Chart

- 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 vs. Many Categories
- Many Items vs. FEW Items
- Cyclical Data vs. Non-Cyclical Data
- Single or FEW Categories vs. Many Categories

Among Items vs. Over Time
- FEW Categories vs. Many Periods
- Many Periods vs. FEW Periods

What Would You Like to Show?

Comparison
- FEW Data Points

Relationship
- Scatter Chart
- Bubble Chart

Distribution
- Single Variable
- Many Data Points

Composition
- Changing Over Time
- Static

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

Components of Components
- Simple Share of Total
- Accumulation or Subtraction to Total
- Components of Components

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

```r
D <- read.csv("player_stats.csv")
hist(D$weight)
boxplot(D$weight)
```
Choosing the bin width may change how things look

> 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)
```
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
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
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)

GRAPHICAL EDA

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

examples on this page from: "seaborn: statistics"
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.
"data–ink ratio = \frac{\text{data-ink}}{\text{total ink used to print the graphic}}

= \text{proportion of a graphic's ink devoted to the non-redundant display of data information} \)
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.
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
THE GOOD, THE BAD, AND THE UGLY

#4

- 4. Apply inflation-adjustment for money

from: (Tufte, 1983)
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.

There it is, with dollars-per-week indicated up the left side. It is a clear and honest picture. Twice as much money is twice as big on the chart and looks it.

The chart lacks that eye-appeal though, doesn’t it? I can easily supply that by using something that looks more like money than a bar does: moneybags. One moneybag for the unfortunate Rotundian’s pittance, two for the American's wage. Or three for the Rotundian, six for the American's wage now dwarfs the foreigner’s.

Now that gives the impression I'm after. The American’s wage now dwarfs the foreigner’s.

from: (Darrell Huff, 1954)
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

SCHOOL DIPLOMAS THAN EVER BEFORE

- Therefore, column (bar) charts should always start at zero!
THE GOOD, THE BAD, AND THE UGLY #6

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

Example from: "The most misleading charts of 2015, fixed", qz.com
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
It's not just about the pretty pictures.
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
```