MTAT.03.319

Business Data Analytics

Lecture 2
Descriptive analysis and visualization

The slides are available under creative common license.

The original owner of these slides is the University of Tartu.
Lecture 2: Descriptive analysis and visualization
“about 1/5 of business decision-makers don’t really understand what big data is or still believe that big data is a lot of hype”

Informative business decisions vs. intuition

many business decisions remain based on intuitive hunches, not facts

many business decisions remain based on intuitive hunches, not facts

analytics helps to reduce the gap between intuition and factual decision-making

Informative business decisions vs. intuition

- many business decisions remain based on intuitive hunches, not facts

- analytics helps to reduce the gap between intuition and factual decision-making

- sophisticated data usage brings competitive gains

image source: making of mass effect andromeda on behance
Informative business decisions vs. intuition

- Many business decisions remain based on intuitive hunches, not facts.
- Analytics helps to reduce the gap between intuition and factual decision-making.
- Sophisticated data usage brings competitive gains.
- Data does not speak for itself. It should be analyzed to take full advantage of its potential.

Source: Forrester’s Global Business Technographics Data And Analytics Survey, 2015
Data is not yet knowledge

Data → Information → Knowledge → Decision
Data → clean, organize, summarize, filter → Information

Knowledge → analyze → Decision

collect, action, decision making
Usable data is clean, consistent, comprehensive, and current.
Most common risks

“Our customer data is a mess. We know we could be doing more with it. Being smarter...”

“We already have some great insights, but we’re too busy (and to be perfectly honest, too disorganized) to actually use them...”

source: custora.com
Most common risks

- Organization will not have the expertise to use the tools
- Organization will not have the expertise of concepts and techniques
- Business people will not understand how to obtain business values out of BA
Course agenda

1. Introduction
2. Tools: descriptive analysis and visualization
3. ...
Examples are easy and clean. Real data is messy.
Analysis: principles
Steps of data analysis

source: R for Data Science
http://r4ds.had.co.nz/introduction.html
Steps of data analysis

source: R for Data Science
http://r4ds.had.co.nz/introduction.html
Exploratory phase

Descriptive statistics

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.72</td>
<td>35.13</td>
<td>51.54</td>
<td>48.60</td>
<td>63.33</td>
<td>77.95</td>
</tr>
</tbody>
</table>

Explore via Descriptive Plots
What is tidy data

**wide format**

<table>
<thead>
<tr>
<th></th>
<th>treatment _a</th>
<th>treatment _b</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Smith</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Jane Doe</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>Mary Johnson</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

**wide format**

<table>
<thead>
<tr>
<th></th>
<th>John Smith</th>
<th>Jane Doe</th>
<th>Mary Johnson</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment _a</td>
<td></td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>treatment _b</td>
<td>2</td>
<td>11</td>
<td>1</td>
</tr>
</tbody>
</table>

**long format**

<table>
<thead>
<tr>
<th>person</th>
<th>treatment</th>
<th>result</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Smith</td>
<td>a</td>
<td>—</td>
</tr>
<tr>
<td>Jane Doe</td>
<td>a</td>
<td>16</td>
</tr>
<tr>
<td>Mary Johnson</td>
<td>a</td>
<td>3</td>
</tr>
<tr>
<td>John Smith</td>
<td>b</td>
<td>2</td>
</tr>
<tr>
<td>Jane Doe</td>
<td>b</td>
<td>11</td>
</tr>
<tr>
<td>Mary Johnson</td>
<td>b</td>
<td>1</td>
</tr>
</tbody>
</table>
What is tidy data

3 principles:

- each variable forms a column
- each observation forms a row
- each type of observational unit forms a table

source: R for Data Science
http://r4ds.had.co.nz/introduction.html
What is tidy data

3 principles:
- each variable forms a column
- each observations forms a row
- each type of observational unit forms a table
Data types

- Categorical
  - Binary
  - Nominal
  - Ordinal

- Numerical
  - Discrete
  - Continuous
Central tendency measures: Mean, Mode, Median.

Variation measures: Variance, Standard deviation.

Relative measures: Percentiles
Central tendency measures. Mean, Mode, Median.
computed to provide a ‘center’ around which observations are distributed

Variation measures. Variance, Standard deviation.
describe ‘data spread’ or the distance from the center.

Relative measures. Percentiles
description of relative positions of observations
The Mode, the Median, and the Mean

\[
x \leftarrow c(4,5,2,5,0,0,4,0,9,3)
\]

mode: 0

\[
\text{sort}(x): 0\ 0\ 0\ 2\ 3\ 4\ 4\ 5\ 5\ 9
\]

median: 3.5

\[
\frac{\sum_{i=1}^{n} x_i}{n}
\]

mean: 3.2
Variance and standard deviation

\[
\text{variance} = \frac{\sum (\text{each observation} - \text{mean})^2}{\text{number of observations}}
\]

\[
\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}
\]

standard deviation = \sqrt{\text{variance}}
The Mode, the Median, and the Mean
Distributions

Quartiles, percentiles and IQR

The $n$th percentile is a value such that $n\%$ of the observations fall at or below it.
Quartiles, percentiles and IQR

The first quartile, $q_1$, is the value for which 25% of the observations are smaller and 75% are larger.

Only 25% of the observations are greater than $q_3$.

IQR = Q3 – Q1
Visualization
A visualization is a graphical representation designed to enable exploration, analysis, or communication.
The goal of the visualization

Data exploration
The goal of the visualization

Conveying the message
Cut through B2B information overload with effective storytelling.

Everyone is overloaded with information, so how do you cut through the noise and prioritize what you want your B2B audience to care about?

Everyone wants multimedia: 77% of respondents prefer multimedia to text-only content.

“Show, don’t tell”: It’s classic storytelling advice and more than ever in the age of information overload.

No one reads: Only 79% of people read an average of 100,000 words per day.

Make sure your message doesn’t get skinned: People scan the web rather than reading word by word.

Marketing is noisy: 5.3 trillion ads are served up on the Internet in 2015.

The problems of information overload in making your voice heard are many, but their solution is simple — your story, told by you.
The goal of the visualization

Topic exploration by the end-user
http://hint.fm/wind/index.html
https://www.bloomberg.com/billionaires/

16.5M  
troy ounces of gold

407M  
barrels of crude oil

... and is equivalent to ...

0.115%  
of the GDP of the United States

0.422%  
of the total wealth of the 500 richest people in the world

5.4%  
of the top 100 U.S. college endowments

338%  
of the top 200 U.S. executives' total awarded compensation
<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: 23.0769</td>
<td>70.3125</td>
</tr>
<tr>
<td>2: 24.3590</td>
<td>81.0817</td>
</tr>
<tr>
<td>3: 26.9231</td>
<td>90.3125</td>
</tr>
<tr>
<td>4: 29.7436</td>
<td>86.8510</td>
</tr>
<tr>
<td>5: 31.5385</td>
<td>82.2356</td>
</tr>
<tr>
<td>6: 34.3590</td>
<td>76.8510</td>
</tr>
<tr>
<td>7: 38.9744</td>
<td>77.6202</td>
</tr>
<tr>
<td>8: 42.8205</td>
<td>79.5433</td>
</tr>
<tr>
<td>9: 22.3077</td>
<td>63.3894</td>
</tr>
<tr>
<td>10: 22.0513</td>
<td>53.3894</td>
</tr>
<tr>
<td>11: 24.6154</td>
<td>47.2356</td>
</tr>
<tr>
<td>12: 28.7179</td>
<td>41.4663</td>
</tr>
</tbody>
</table>

<truncated>

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>68: 21.2821</td>
<td>46.4663</td>
</tr>
<tr>
<td>69: 27.1795</td>
<td>48.7740</td>
</tr>
<tr>
<td>70: 31.0256</td>
<td>49.1587</td>
</tr>
<tr>
<td>71: 35.1282</td>
<td>49.5433</td>
</tr>
<tr>
<td>72: 40.2564</td>
<td>51.4663</td>
</tr>
<tr>
<td>73: 45.8974</td>
<td>53.0048</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>1:</td>
<td>23.0769</td>
</tr>
<tr>
<td>2:</td>
<td>24.3590</td>
</tr>
<tr>
<td>3:</td>
<td>26.9231</td>
</tr>
<tr>
<td>4:</td>
<td>29.7436</td>
</tr>
<tr>
<td>5:</td>
<td>31.5385</td>
</tr>
<tr>
<td>6:</td>
<td>34.3590</td>
</tr>
<tr>
<td>7:</td>
<td>38.9744</td>
</tr>
<tr>
<td>8:</td>
<td>42.8205</td>
</tr>
<tr>
<td>9:</td>
<td>22.3077</td>
</tr>
<tr>
<td>10:</td>
<td>22.0513</td>
</tr>
<tr>
<td>11:</td>
<td>24.6154</td>
</tr>
<tr>
<td>12:</td>
<td>28.7179</td>
</tr>
</tbody>
</table>

<truncated>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td></td>
</tr>
<tr>
<td>Min. :18.72</td>
<td>Min. :33.77</td>
<td></td>
</tr>
<tr>
<td>1st Qu.:35.13</td>
<td>1st Qu.:49.16</td>
<td></td>
</tr>
<tr>
<td>Median :51.54</td>
<td>Median :55.70</td>
<td></td>
</tr>
<tr>
<td>Mean :48.60</td>
<td>Mean :59.43</td>
<td></td>
</tr>
<tr>
<td>3rd Qu.:63.33</td>
<td>3rd Qu.:69.16</td>
<td></td>
</tr>
<tr>
<td>Max. :77.95</td>
<td>Max. :90.31</td>
<td></td>
</tr>
</tbody>
</table>

> cor(dt$x, dt$y)

[1] -0.005949079
\[
\begin{array}{cc}
1: & 23.0769 \, 70.3125 \\
2: & 24.3590 \, 81.0817 \\
3: & 26.9231 \, 90.3125 \\
4: & 29.7436 \, 86.8510 \\
5: & 31.5385 \, 82.2356 \\
6: & 34.3590 \, 76.8510 \\
7: & 38.9744 \, 77.6202 \\
8: & 42.8205 \, 79.5433 \\
9: & 22.3077 \, 63.3894 \\
10: & 22.0513 \, 53.3894 \\
11: & 24.6154 \, 47.2356 \\
12: & 28.7179 \, 41.4663 \\
\end{array}
\]
Visualization ABCs

- bar chart
- multi-set bar chart
- histogram
- density plot

- boxplot
- line chart
- scatter plot
- network/graph

http://www.datavizcatalogue.com/
Dataset

A transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.
Dataset

```r
> head(initial_dt)

<table>
<thead>
<tr>
<th>InvoiceNo</th>
<th>StockCode</th>
<th>Description</th>
<th>Quantity</th>
<th>InvoiceDate</th>
<th>UnitPrice</th>
<th>CustomerID</th>
</tr>
</thead>
<tbody>
<tr>
<td>536365</td>
<td>85123A</td>
<td>WHITE HANGING HEART T-LIGHT HOLDER</td>
<td>6</td>
<td>01/12/10 08:26</td>
<td>2.55</td>
<td>17850</td>
</tr>
<tr>
<td>536365</td>
<td>71053</td>
<td>WHITE METAL LANTERN</td>
<td>6</td>
<td>01/12/10 08:26</td>
<td>3.39</td>
<td>17850</td>
</tr>
<tr>
<td>536365</td>
<td>84406B</td>
<td>CREAM CUPID HEARTS COAT HANGER</td>
<td>8</td>
<td>01/12/10 08:26</td>
<td>2.75</td>
<td>17850</td>
</tr>
<tr>
<td>536365</td>
<td>84029G</td>
<td>KNITTED UNION FLAG HOT WATER BOTTLE</td>
<td>6</td>
<td>01/12/10 08:26</td>
<td>3.39</td>
<td>17850</td>
</tr>
<tr>
<td>536365</td>
<td>84029E</td>
<td>RED WOOLLY HOTTIE WHITE HEART.</td>
<td>6</td>
<td>01/12/10 08:26</td>
<td>3.39</td>
<td>17850</td>
</tr>
<tr>
<td>536365</td>
<td>22752</td>
<td>SET 7 BABUSHKA NESTING BOXES</td>
<td>2</td>
<td>01/12/10 08:26</td>
<td>7.65</td>
<td>17850</td>
</tr>
</tbody>
</table>

Country
---
1: United Kingdom
2: United Kingdom
3: United Kingdom
4: United Kingdom
5: United Kingdom
6: United Kingdom
```
Description of one discrete feature that displays counts

Bar chart
2500 transactions were made from Spain.
Multi-set bar chart

Description of two discrete features that displays counts

Country

Returns
Regular
Return
Multi-set bar chart

Description of two discrete features that displays counts

USA returns are exceptionally high compared to all orders.
Description of one continuous feature. Displays general distribution
Histogram

Description of one continuous feature. Displays general distribution

Mostly the order size is 1
Effect of a bin size on histogram
Density

Description of one continuous feature. Displays smoothed general distribution.
Boxplot

The actual values in a distribution

How a histogram would display the values (rotated)

How a boxplot would display the values

Outliers

Whisker to farthest non-outlier point

75th percentile

50th percentile

25th percentile

1.5 x IQR

Inter-Quartile Range (IQR)

source: Alberto Cairo
Line chart

2 continuous features. Usually time series
Scatter plot

Relationship of 2 continuous features.
Network

https://martenveskimaeh.shinyapps.io/R-packages/
What is the best way to understand correctly the differences without reading the numbers?

Length or height

Position

Area

Hue and shade

Angle/area

Line weight

source: Alberto Cairo

Figures represented in all these graphics: 22%, 25%, 34%, 29%, 32%
Chart Suggestions—A Thought-Starter

Comparison

Relationship

What would you like to show?

Distribution

Composition

© 2006 A. Abela — a.v.abela@gmail.com
Which one is better?
Building blocks of a graph include:
1. data
2. aesthetic mapping (color, position, fill, shape, linetype, size)
3. geometric object (points, lines, boxplots)
4. statistical transformations
5. scales
6. coordinate system
7. position adjustments
8. faceting
ggplot2

ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))

ggplot(dataset) + geom_histogram(aes(x=Quantity))
Demo time!

https://courses.cs.ut.ee/2018/bda/fall/Main/Practice
References

• http://ggplot2.tidyverse.org/reference/
• http://www.datavizcatalogue.com/
• The Truthful Art by Alberto Cairo