Data Visualization

Workshop

David Sasson / Wei-Hung Weng

HST.953 Collaborative Data Science in Medicine
Sep 14, 2018

Massachusetts
Institute of
Technology

Agenda

- Quick introduction
- Tutorial 1: Data Visualization with R and ggplot2
  - Powerful visualization tool in R
- Tutorial 2: Visualization for clinical applications
  - Applied visualization with simple R commands

Why Visualization?

- A method of encoding quantitative, relational, or spatial information into images
- Taps into the visual system – an enormously powerful pattern-finding device – which can reveal structure in data in a compelling and accessible way

David Sasson

Goal of Visualization

- The greatest value of a picture is when it forces us to notice what we never expected to see.

  John Tukey (1977)

- The purpose of visualization is insight, not pictures.

  Ben Shneiderman (1999)

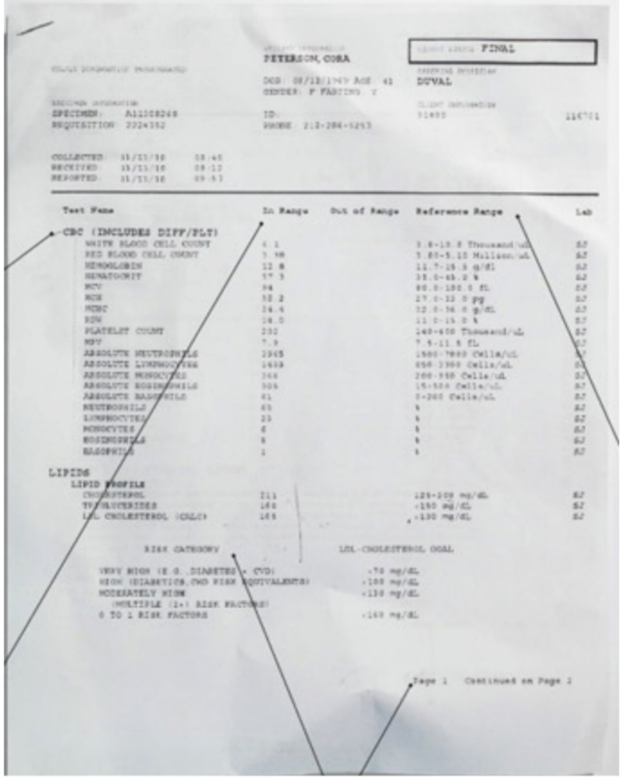
- Understanding and exploring trends and patterns inside data
- Summarizing statistics
  - Instead of reading thousands raw data points
- Telling a story!

Medical Data?

Why challenging (for analysis as well as visualization)?

- Volume
- Missing data
- Trusting source of data/resolving conflicting data
- Time series
- Change/acceleration vs. absolute (whether in spending or in disease progression)
- Bias

Do This



[illegible]

PETERSON, CORA

DOB: 08/12/1969 AGE: 41
GENDER: F FASTING: Y

ID:
PHONE: 212-284-5253

[illegible] FINAL

[illegible]
DUVAL

[illegible]
31485 116701

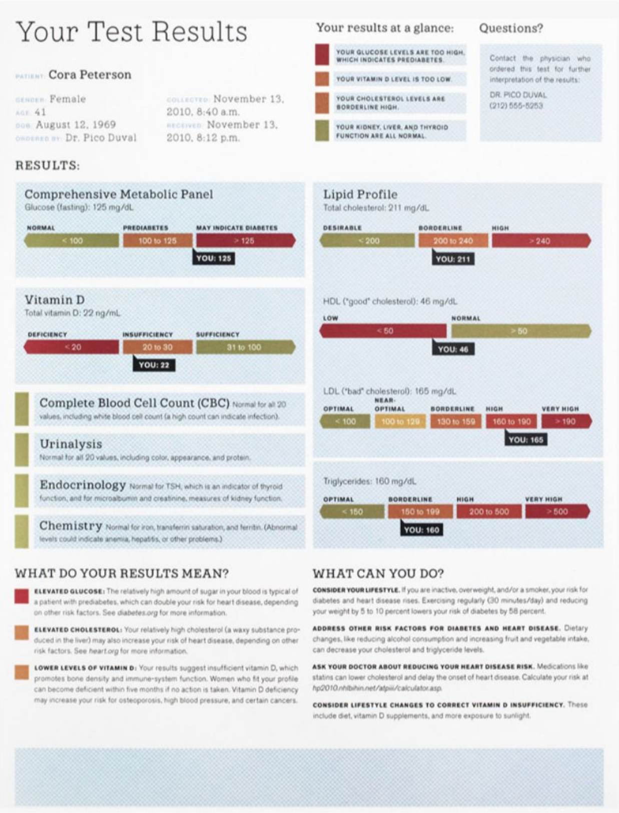
[illegible]
SPECIMEN: A11208248
REQUISITION: 2224352

COLLECTED:	11/13/10	08:40
RECEIVED:	11/13/10	08:12
REPORTED:	11/13/10	09:53

Test Name	In Range	Out of Range	Reference Range	Lab
CBC (INCLUDES DIFF/PLT)				
WHITE BLOOD CELL COUNT	6.1		3.8-10.8 Thousand/uL	EZ
RED BLOOD CELL COUNT	3.98		3.80-5.10 Million/uL	EZ
HEMOGLOBIN	12.8		11.7-15.5 g/dL	EZ
HEMATOCRIT	37.3		35.0-45.0 %	EZ
MCV	94		80.0-100.0 fL	EZ
MCH	32.2		27.0-33.0 pg	EZ
MCHC	34.4		32.0-36.0 g/dL	EZ
RDW	14.0		11.0-15.0 %	EZ
PLATELET COUNT	232		140-400 Thousand/uL	EZ
MPV	7.9		7.5-11.5 fL	EZ
ABSOLUTE NEUTROPHILS	3965		1500-7800 Cells/uL	EZ
ABSOLUTE LYMPHOCYTES	1403		850-3900 Cells/uL	EZ
ABSOLUTE MONOCYTES	366		200-950 Cells/uL	EZ
ABSOLUTE EOSINOPHILS	305		15-500 Cells/uL	EZ
ABSOLUTE BASOPHILS	61		0-200 Cells/uL	EZ
NEUTROPHILS	65		%	EZ
LYMPHOCYTES	23		%	EZ
MONOCYTES	6		%	EZ
EOSINOPHILS	5		%	EZ
BASOPHILS	1		%	EZ
LIPIDS				
LIPID PROFILE				
CHOLESTEROL	211		125-200 mg/dL	EZ
TRIGLYCERIDES	160		<150 mg/dL	EZ
LDL CHOLESTEROL (CALC)	165		<130 mg/dL	EZ

RISK CATEGORY	LDL-CHOLESTEROL GOAL
VERY HIGH (E.G., DIABETES + CVD)	<70 mg/dL
HIGH (DIABETICS,CHD RISK EQUIVALENTS)	<100 mg/dL
MODERATELY HIGH (MULTIPLE (2+) RISK FACTORS)	<130 mg/dL
0 TO 1 RISK FACTORS	<160 mg/dL

Page 1 Continued on Page 2



Your Test Results

PATIENT: Cora Peterson

GENDER: Female
AGE: 41
DOB: August 12, 1969
ORDERED BY: Dr. Pico Duval

COLLECTED: November 13, 2010, 8:40 a.m.
RECEIVED: November 13, 2010, 8:12 p.m.

Your results at a glance:

- 🟥 YOUR GLUCOSE LEVELS ARE TOO HIGH, WHICH INDICATES PREDIABETES.
- 🟧 YOUR VITAMIN D LEVEL IS TOO LOW.
- 🟧 YOUR CHOLESTEROL LEVELS ARE BORDERLINE HIGH.
- 🟩 YOUR KIDNEY, LIVER, AND THYROID FUNCTION ARE ALL NORMAL.

Questions?

Contact the physician who ordered this test for further interpretation of the results:

DR. PICO DUVAL
(212) 555-5253

RESULTS:

Comprehensive Metabolic Panel
Glucose (fasting): 125 mg/dL

NORMAL	PREDIABETES	MAY INDICATE DIABETES
< 100	100 to 125	> 125

YOU: 125

Vitamin D
Total vitamin D: 22 ng/mL

DEFICIENCY	INSUFFICIENCY	SUFFICIENCY
< 20	20 to 30	31 to 100

YOU: 22

Complete Blood Cell Count (CBC) Normal for all 20 values, including white blood cell count (a high count can indicate infection).

Urinalysis Normal for all 20 values, including color, appearance, and protein.

Endocrinology Normal for TSH, which is an indicator of thyroid function, and for microalbumin and creatinine, measures of kidney function.

Chemistry Normal for iron, transferrin saturation, and ferritin. (Abnormal levels could indicate anemia, hepatitis, or other problems.)

Lipid Profile
Total cholesterol: 211 mg/dL

DESIRABLE	BORDERLINE	HIGH
< 200	200 to 240	> 240

YOU: 211

HDL ("good" cholesterol): 46 mg/dL

LOW	NORMAL
< 50	> 50

YOU: 46

LDL ("bad" cholesterol): 165 mg/dL

OPTIMAL	NEAR-OPTIMAL	BORDERLINE	HIGH	VERY HIGH
< 100	100 to 129	130 to 159	160 to 190	> 190

YOU: 165

Triglycerides: 160 mg/dL

OPTIMAL	BORDERLINE	HIGH	VERY HIGH
< 150	150 to 199	200 to 500	> 500

YOU: 160

WHAT DO YOUR RESULTS MEAN?

🟥 **ELEVATED GLUCOSE:** The relatively high amount of sugar in your blood is typical of a patient with prediabetes, which can double your risk for heart disease, depending on other risk factors. See diabetes.org for more information.

🟧 **ELEVATED CHOLESTEROL:** Your relatively high cholesterol (a waxy substance produced in the liver) may also increase your risk of heart disease, depending on other risk factors. See *heart.org* for more information.

🟧 **LOWER LEVELS OF VITAMIN D:** Your results suggest insufficient vitamin D, which promotes bone density and immune-system function. Women who fit your profile can become deficient within five months if no action is taken. Vitamin D deficiency may increase your risk for osteoporosis, high blood pressure, and certain cancers.

WHAT CAN YOU DO?

CONSIDER YOUR LIFESTYLE. If you are inactive, overweight, and/or a smoker, your risk for diabetes and heart disease rises. Exercising regularly (30 minutes/day) and reducing your weight by 5 to 10 percent lowers your risk of diabetes by 58 percent.

ADDRESS OTHER RISK FACTORS FOR DIABETES AND HEART DISEASE. Dietary changes, like reducing alcohol consumption and increasing fruit and vegetable intake, can decrease your cholesterol and triglyceride levels.

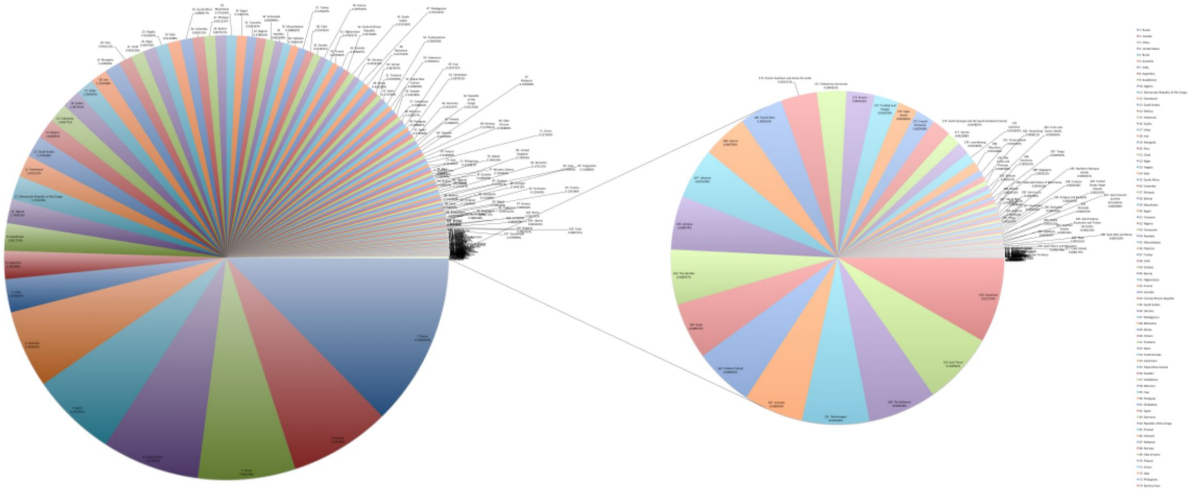
ASK YOUR DOCTOR ABOUT REDUCING YOUR HEART DISEASE RISK. Medications like statins can lower cholesterol and delay the onset of heart disease. Calculate your risk at *hp2010.nhlbihin.net/atpiii/calculator.asp.*

CONSIDER LIFESTYLE CHANGES TO CORRECT VITAMIN D INSUFFICIENCY. These include diet, vitamin D supplements, and more exposure to sunlight.

Don’t Do This!

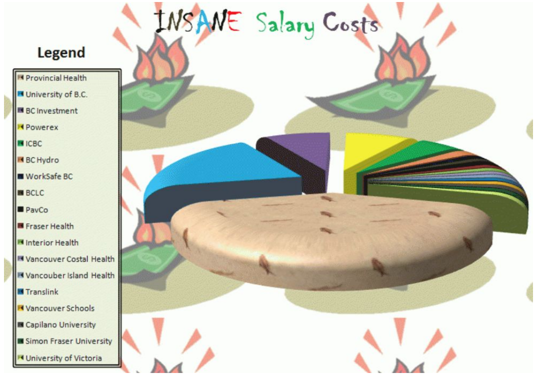
http://viz.wtf

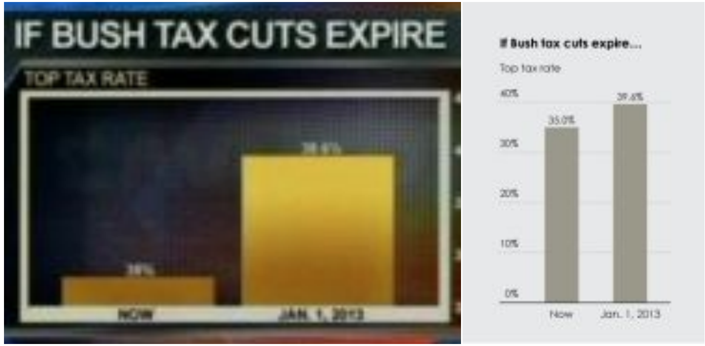
http://eagerpies.com/close-the-bars-down/

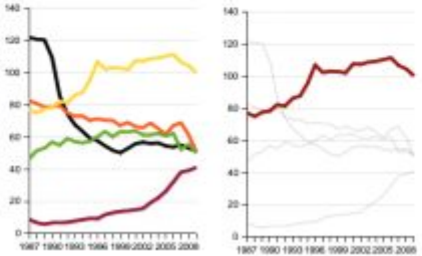


Some Principles

- Lets the data speak for itself
- The addition of extra fluff (shadows, 3D, extravagant colors) eclipses what the graph is actually showing
- Faithful to the data, and doesn't misrepresent it by modifying axes or colors the wrong way
- Data visualization is as much of an art as it is a science
- Minimalism







10 Commandments of Data Visualization

DO	DON'T
Use the full axis	Use 3D effects
Simplify less important information	Use more than six colors
Be creative with legends and labels	Change visual style
Utilize a hierarchy	Make people do visual math
Ask others for opinions	Overload the chart

David Sasson

Some Tips...

- Show the data
- Induce the viewer to think about the substance, rather than about methodology, graphic design, [or] the technology of graphic productions...
- Avoid distorting what the data have to say
- Present many numbers in a small space
- Make large data sets coherent
- Encourage the eye to compare different pieces of data
- Reveal the data at several levels of detail
- Serve a reasonably clear purpose
- Be closely integrated with the statistical and verbal descriptions

Edward Tufte, The Display of Quantitative Information

More Tips...

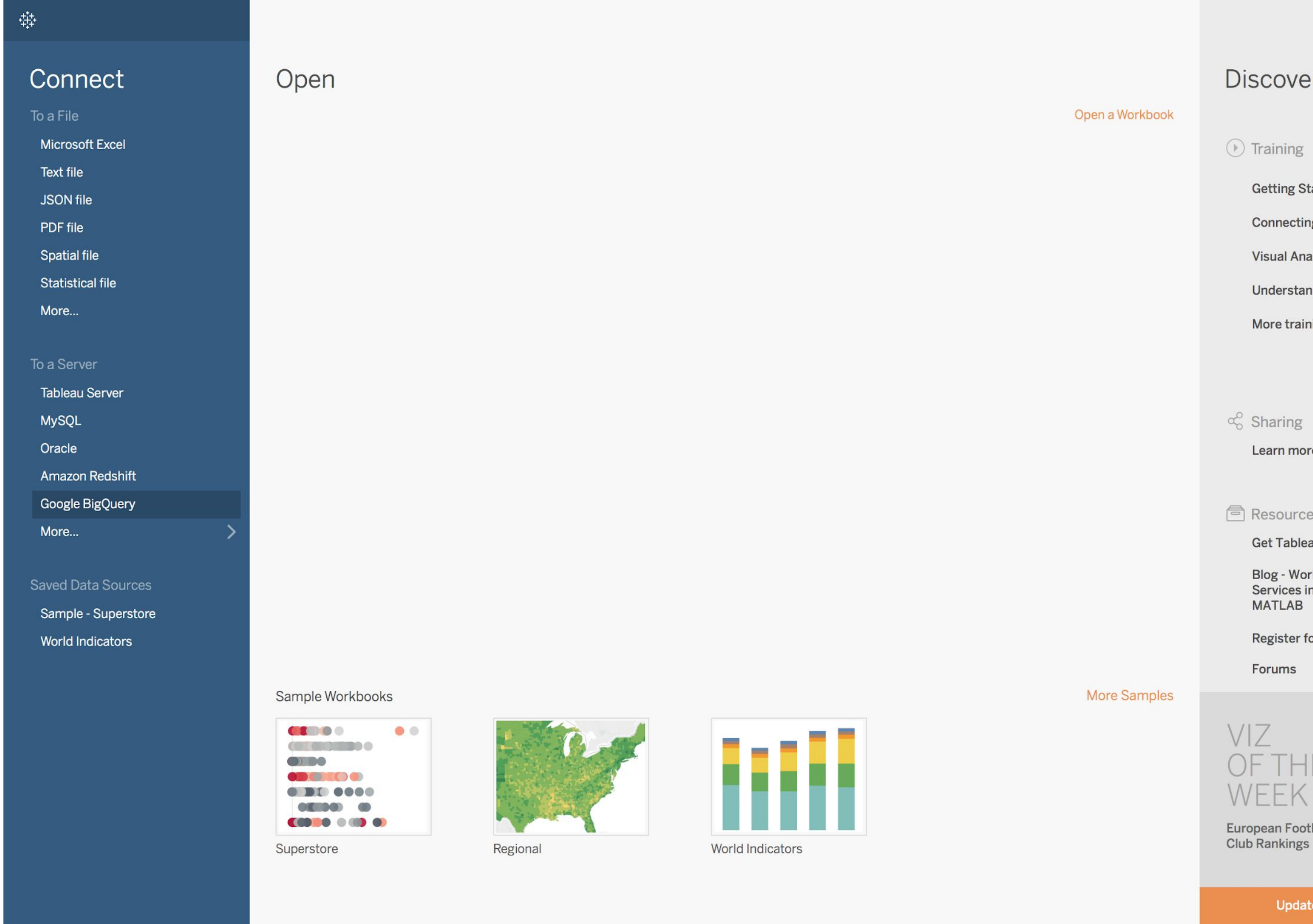
- Written things proceed from left to right (in English)
- Things proceed from top to bottom
- Center things are more important than periphery things
- Foreground things are more important than background things
- Thick things are more important than thin things
- Areas of activity contain the most important information
- Things with the same shape, size, color, or location are related
- Things stand out if they contrast with surroundings in terms of line thickness, type face, or color

T. Huckin and L. Olsen, English for Science and Technology

Further Readings

- Harvard CS171 - Visualization
  - http://www.cs171.org
- GaTech CS 7450 - Information Visualization
  - https://www.cc.gatech.edu/~stasko/7450/
- Edward Tufte
  - https://www.edwardtufte.com/tufte/
- David McCandless
  - https://informationisbeautiful.net/
- Toolkits
  - D3JS
  - R Shiny
  - Tableau (especially if you use Google BigQuery)
  - http://selection.datavisualization.ch/

Tableau



Connect

To a File

- Microsoft Excel
- Text file
- JSON file
- PDF file
- Spatial file
- Statistical file
- More...

To a Server

- Tableau Server
- MySQL
- Oracle
- Amazon Redshift
- Google BigQuery
- More...

Saved Data Sources

- Sample - Superstore
- World Indicators

Open

Open a Workbook

Sample Workbooks

More Samples

Superstore

Regional

World Indicators

Discove

Training

- Getting Sta
- Connectin
- Visual Ana
- Understan
- More train

Sharing

- Learn mor

Resource

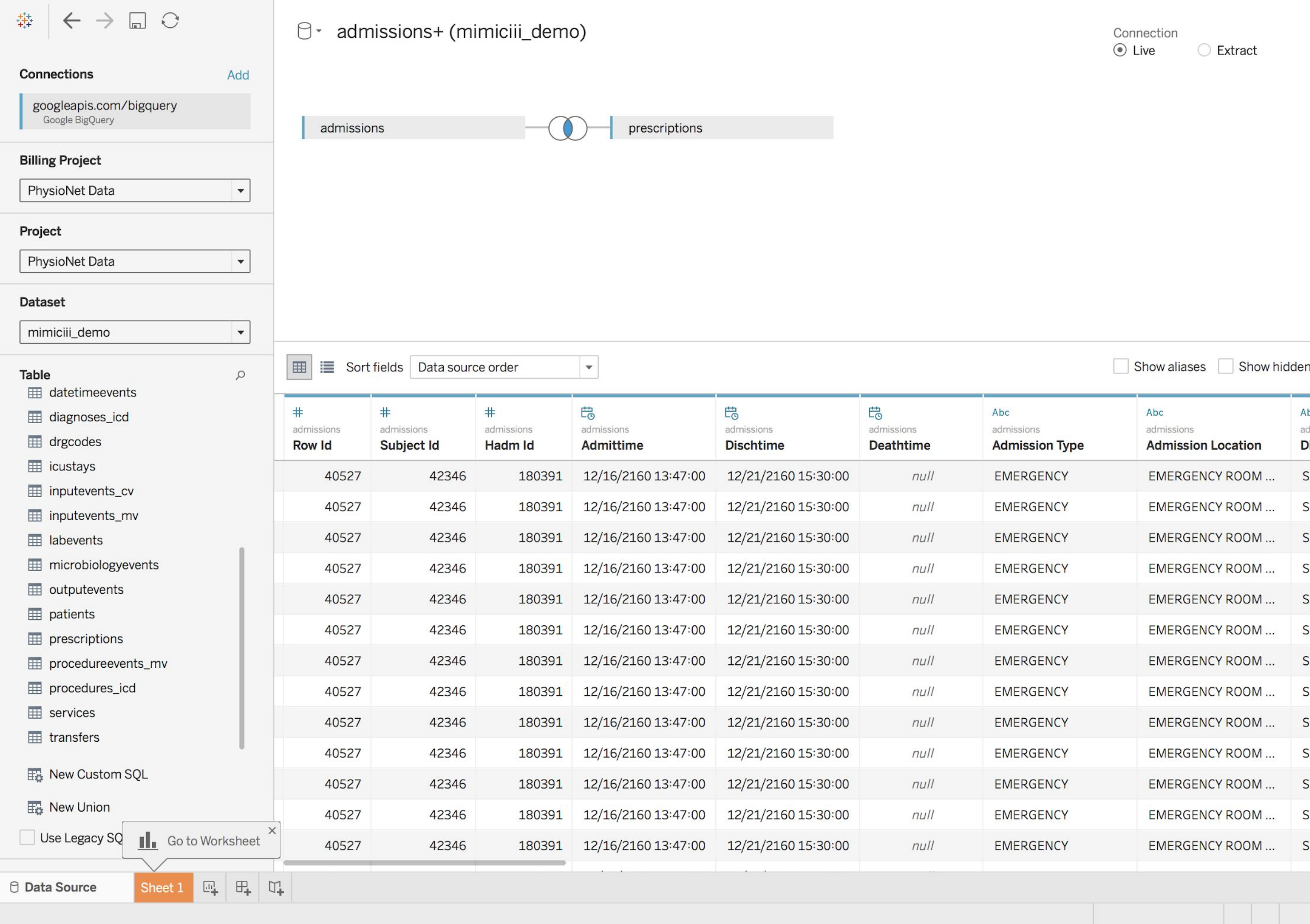
- Get Tablea
- Blog - Wor Services in MATLAB
- Register fo
- Forums

VIZ OF TH WEEK

European Foot Club Rankings

Updat

Tableau



Connections Add

googleapis.com/bigquery
Google BigQuery

Billing Project: PhysioNet Data

Project: PhysioNet Data

Dataset: mimiciii_demo

Table
- datetimeevents
- diagnoses_icd
- drgcodes
- icustays
- inputevents_cv
- inputevents_mv
- labevents
- microbiologyevents
- outputevents
- patients
- prescriptions
- procedureevents_mv
- procedures_icd
- services
- transfers

New Custom SQL

New Union

□ Use Legacy SQ

Go to Worksheet

admissions+ (mimiciii_demo)

Connection ● Live ○ Extract

admissions — prescriptions

Sort fields: Data source order

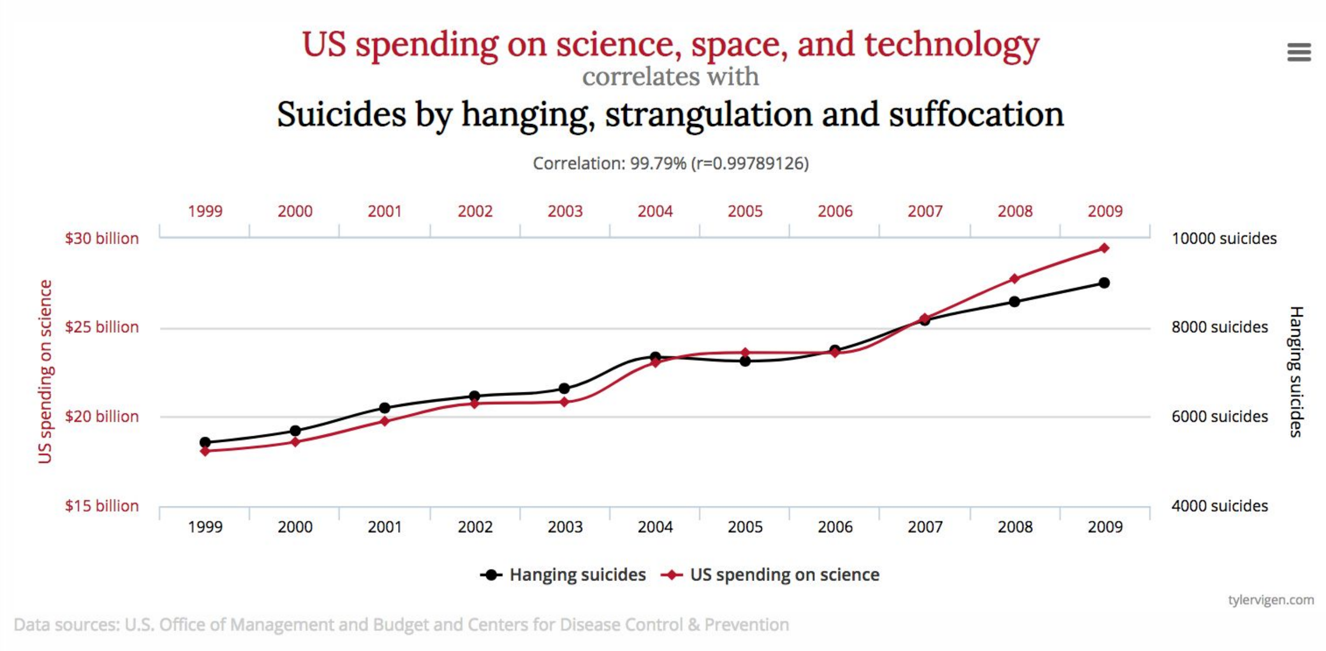
□ Show aliases □ Show hidden

admissions	admissions	admissions	admissions	admissions	admissions	admissions	admissions	ad
Row Id	Subject Id	Hadm Id	Admittime	Dischtime	Deathtime	Admission Type	Admission Location	D
40527	42346	180391	12/16/2160 13:47:00	12/21/2160 15:30:00	null	EMERGENCY	EMERGENCY ROOM ...	S
40527	42346	180391	12/16/2160 13:47:00	12/21/2160 15:30:00	null	EMERGENCY	EMERGENCY ROOM ...	S
40527	42346	180391	12/16/2160 13:47:00	12/21/2160 15:30:00	null	EMERGENCY	EMERGENCY ROOM ...	S
40527	42346	180391	12/16/2160 13:47:00	12/21/2160 15:30:00	null	EMERGENCY	EMERGENCY ROOM ...	S
40527	42346	180391	12/16/2160 13:47:00	12/21/2160 15:30:00	null	EMERGENCY	EMERGENCY ROOM ...	S
40527	42346	180391	12/16/2160 13:47:00	12/21/2160 15:30:00	null	EMERGENCY	EMERGENCY ROOM ...	S
40527	42346	180391	12/16/2160 13:47:00	12/21/2160 15:30:00	null	EMERGENCY	EMERGENCY ROOM ...	S
40527	42346	180391	12/16/2160 13:47:00	12/21/2160 15:30:00	null	EMERGENCY	EMERGENCY ROOM ...	S
40527	42346	180391	12/16/2160 13:47:00	12/21/2160 15:30:00	null	EMERGENCY	EMERGENCY ROOM ...	S
40527	42346	180391	12/16/2160 13:47:00	12/21/2160 15:30:00	null	EMERGENCY	EMERGENCY ROOM ...	S
40527	42346	180391	12/16/2160 13:47:00	12/21/2160 15:30:00	null	EMERGENCY	EMERGENCY ROOM ...	S
40527	42346	180391	12/16/2160 13:47:00	12/21/2160 15:30:00	null	EMERGENCY	EMERGENCY ROOM ...	S
40527	42346	180391	12/16/2160 13:47:00	12/21/2160 15:30:00	null	EMERGENCY	EMERGENCY ROOM ...	S

Data Source | Sheet 1

Next Step

- http://tylervigen.com/spurious-correlations



Tutorial 1 - Introduction to ggplot2

- You need R and RStudio locally OR using RStudio server
  - `http://35.231.235.240:8787`
- `https://github.com/ckbjimmy/hst953_viz`
  - **Clone or download**
  - Upload Rmd to RStudio server
- `https://github.com/dsasson48/dataviz`

Tutorial 2 - Visualization for Clinical Applications

- You need R and RStudio locally OR using RStudio server
  - `http://35.231.235.240:8787`
- `https://github.com/ckbjimmy/hst953_viz`
  - **Clone or download**
  - Upload Rmd to RStudio server
- What's inside the tutorial? Plotting Function in R
  - Histogram, Density estimation
  - Scatter plot, Boxplot
  - Interaction plot
  - Supervised visualization
    - Model validation
    - Summarization
  - Other issues...