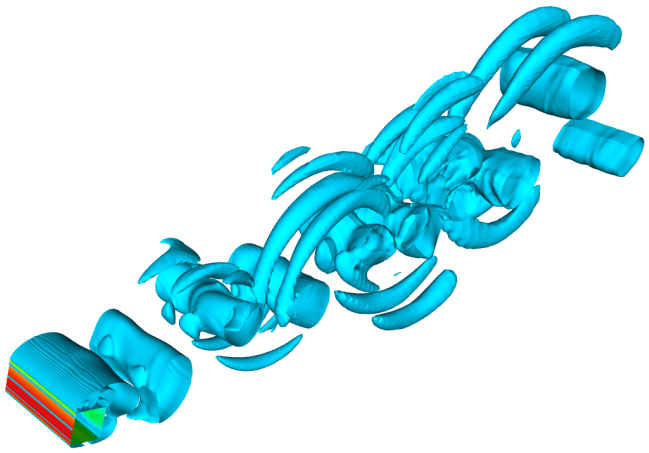
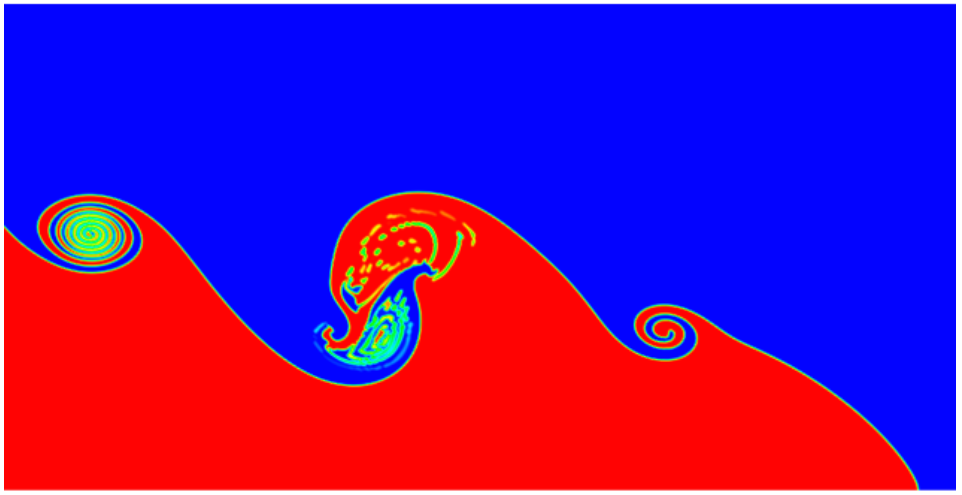
Visual storytelling and data visualization in numerical simulations

Joel Guerrero
University of Genova + Wolf Dynamics

Giovanni Bailardi & Haileyesus Kifle
DLTM La Spezia

This presentation is NOT about colorful fluid dynamics (CFD)

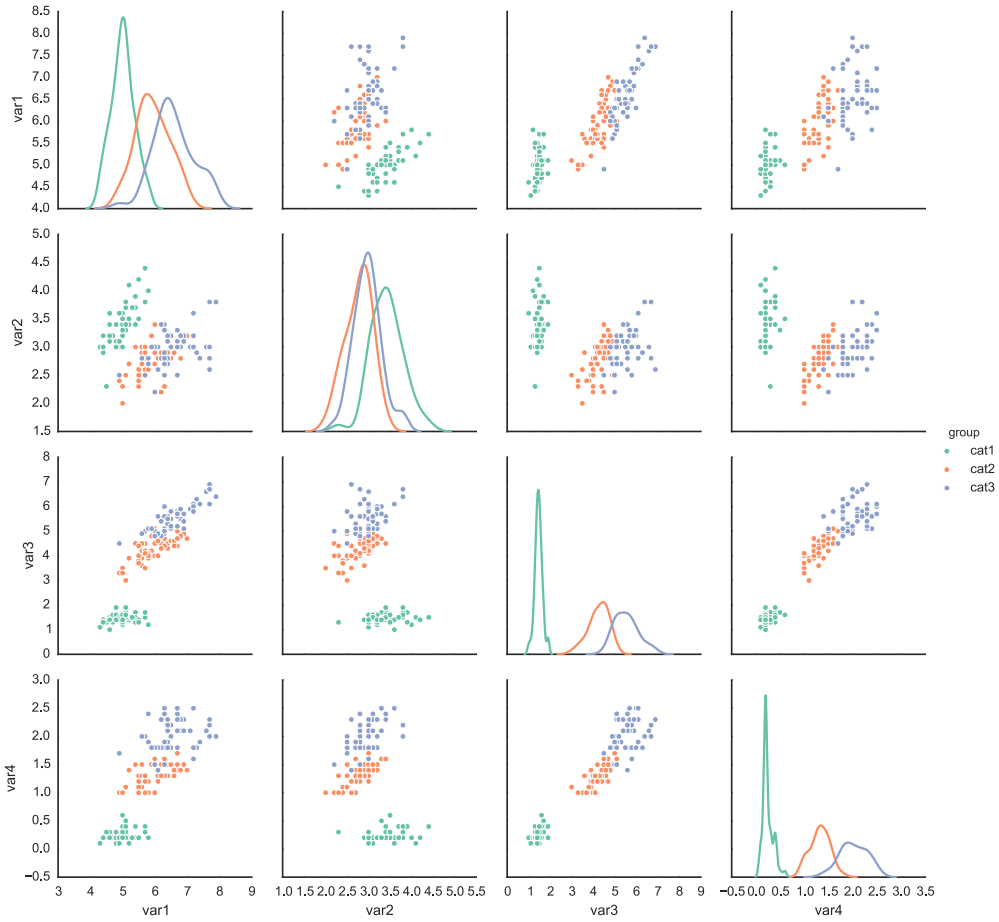


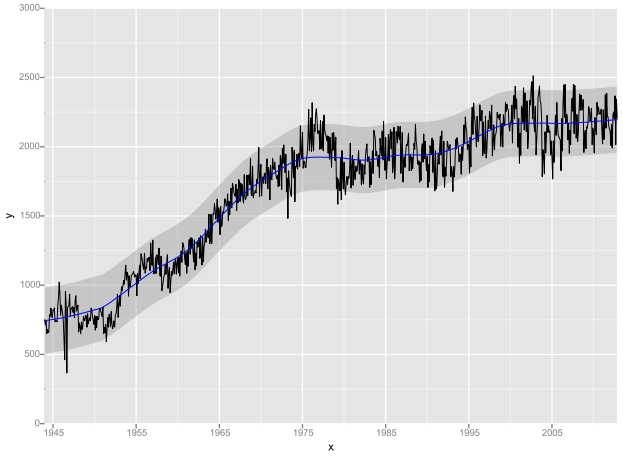


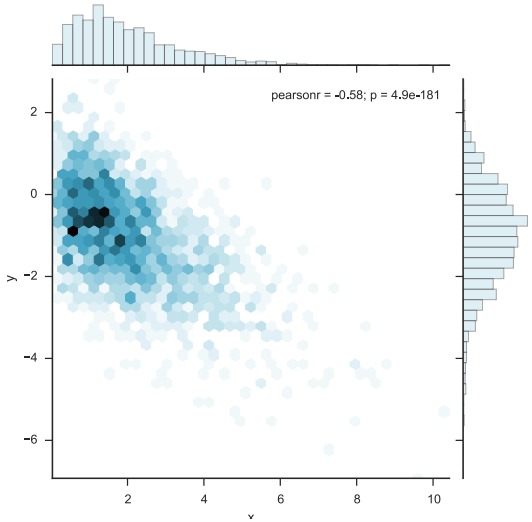
- However, to gather most of the data to be presented I ran many numerical simulations (DAKOTA-OpenFOAM®).
- In particular, design space exploration and design optimization studies.
- And thanks to data analytics (DA) and exploratory data analysis (EDA*), I was able to turn all the quantitative information into valuable insight.

* EDA is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

This presentation is about charts and plots

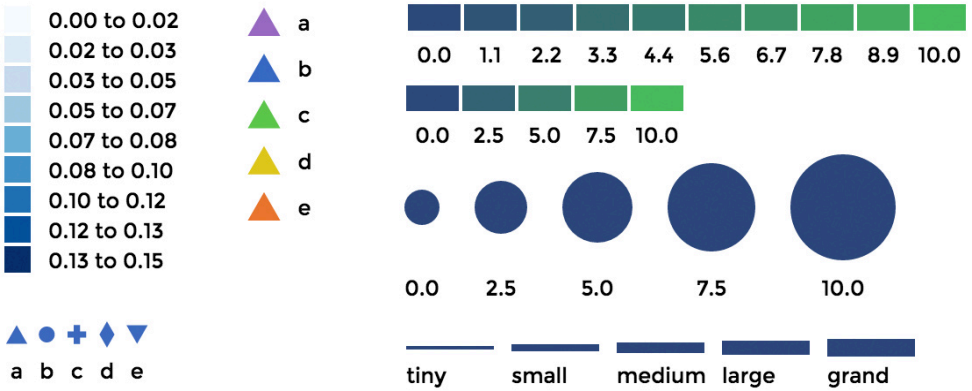






and also visual encoders

- The key to effectively exposing meaningful patterns in data comes down to thoughtful visual encoding.
- Incremental changes in aesthetics should reflect and be perceived as proportional and meaningful changes in data.



D3 Legend by Susie Lu
http://www.susielu.com/

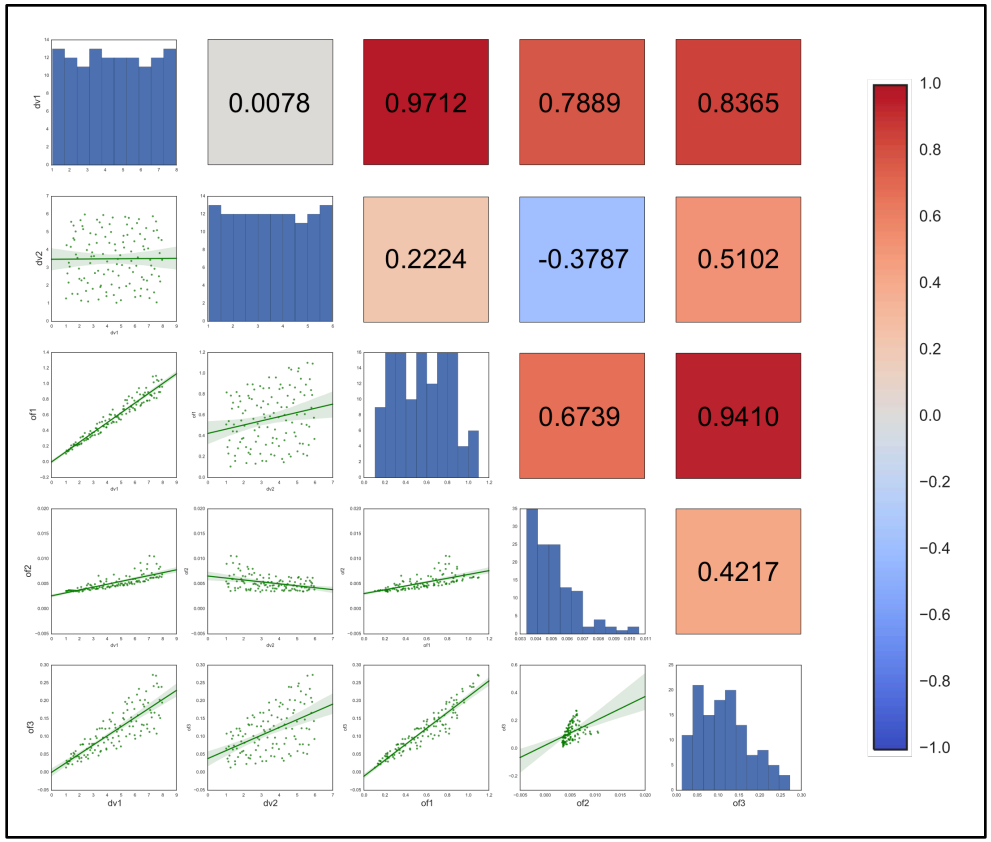
“The greatest value of a picture is when it forces us to notice what we never expected to see.”

— John Tukey

The what of data visualization and visual storytelling

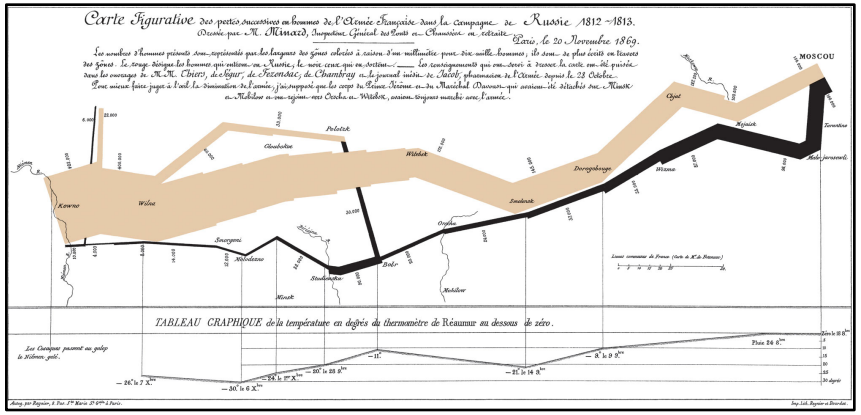
What is data visualization?

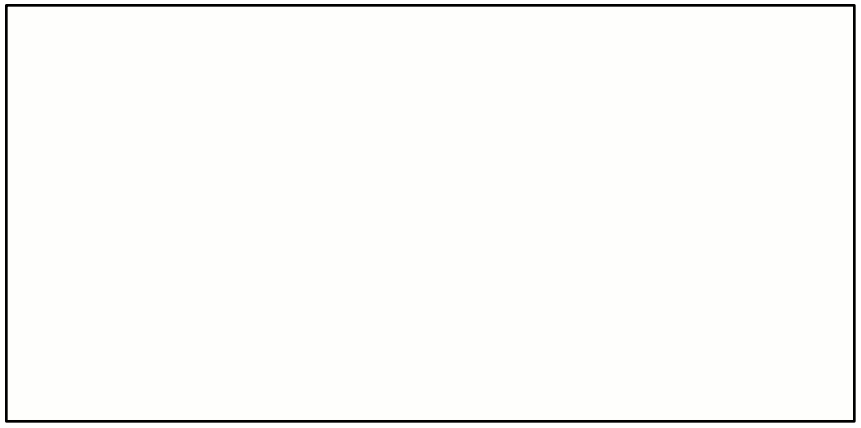
- Data visualization is the presentation of data in a pictorial or graphical format in order to amplify cognition.
- It is where visual storytelling meets math.



What is visual storytelling?

- Communication of a story or known information through visual components.
- It is about reaching a general audience.





The why of data visualization and visual storytelling

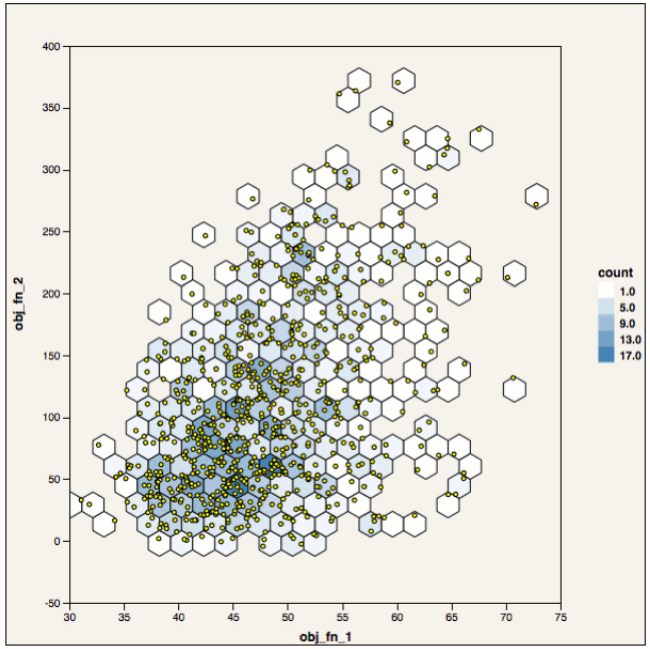
“Transform information from a format efficient for computation into a format efficient for human perception, cognition, and communication.”

— John C. Hart

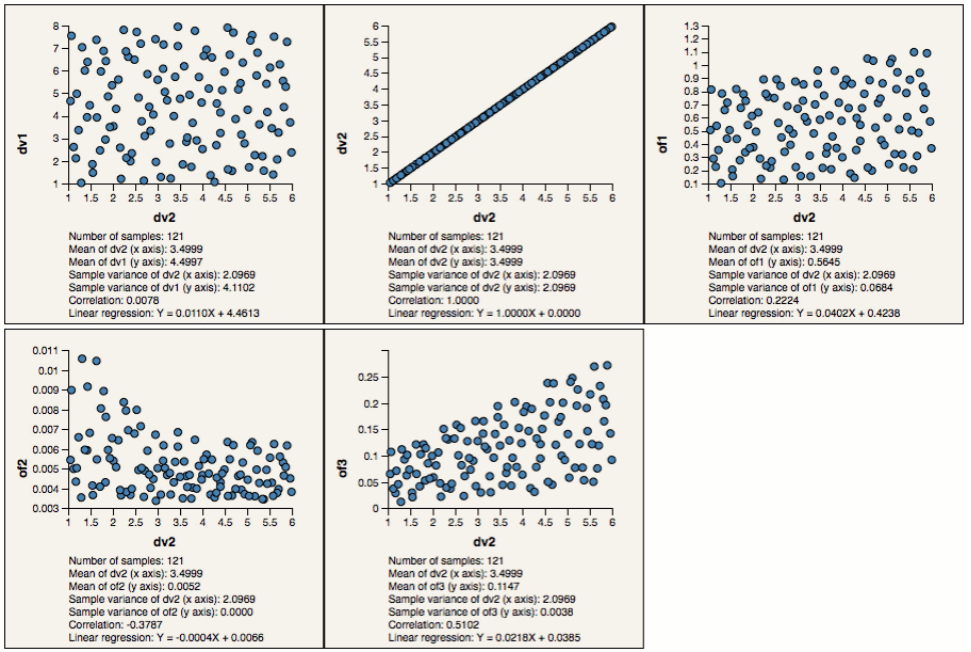
University of Illinois at Urbana-Champaign

Why data visualization and visual storytelling?

- Patterns, trends, correlations and anomalies that might go undetected in raw data can be exposed and recognized easily when visualizing it.
- Turn data into valuable insights and make informed decisions.



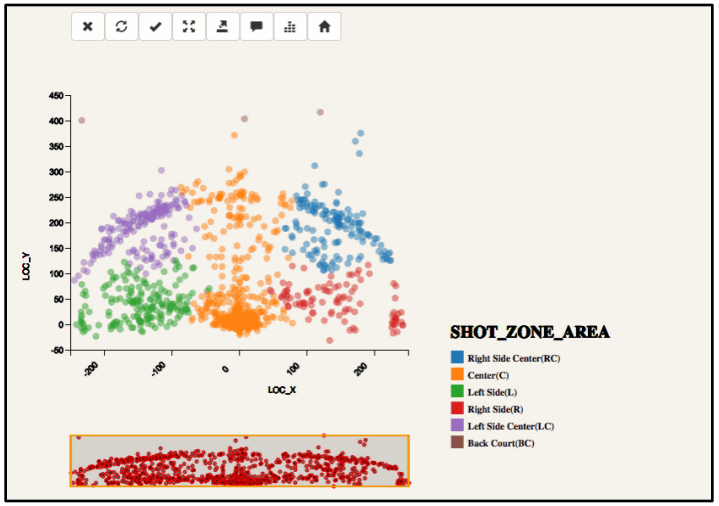
Hexbin plot



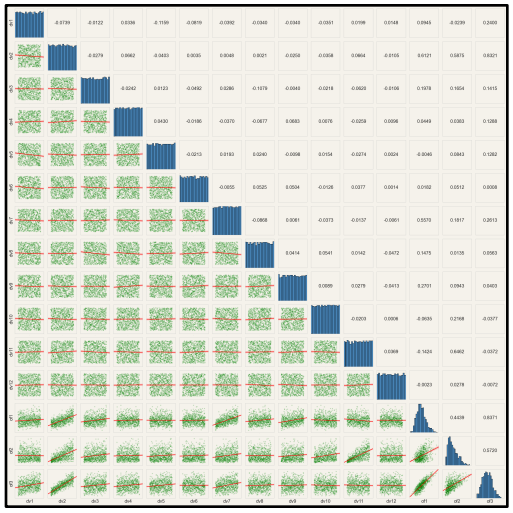
Regression plot – Paired plot

Why data visualization and visual storytelling?

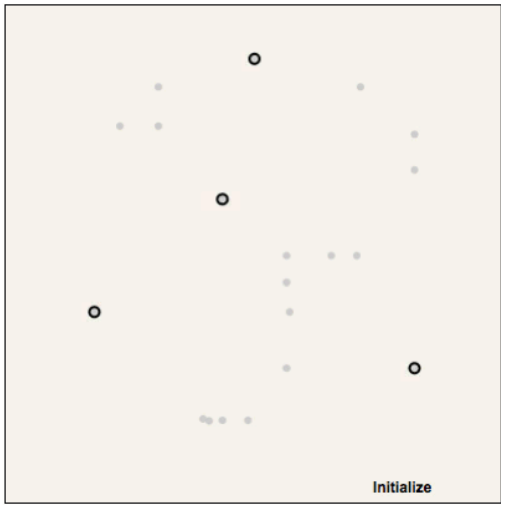
- Spur new questions and prompt skepticism.
- Data interaction (manipulation, cleaning and on-the-fly statistics)
- Explore combinations in the data.



Interactive scatter plot



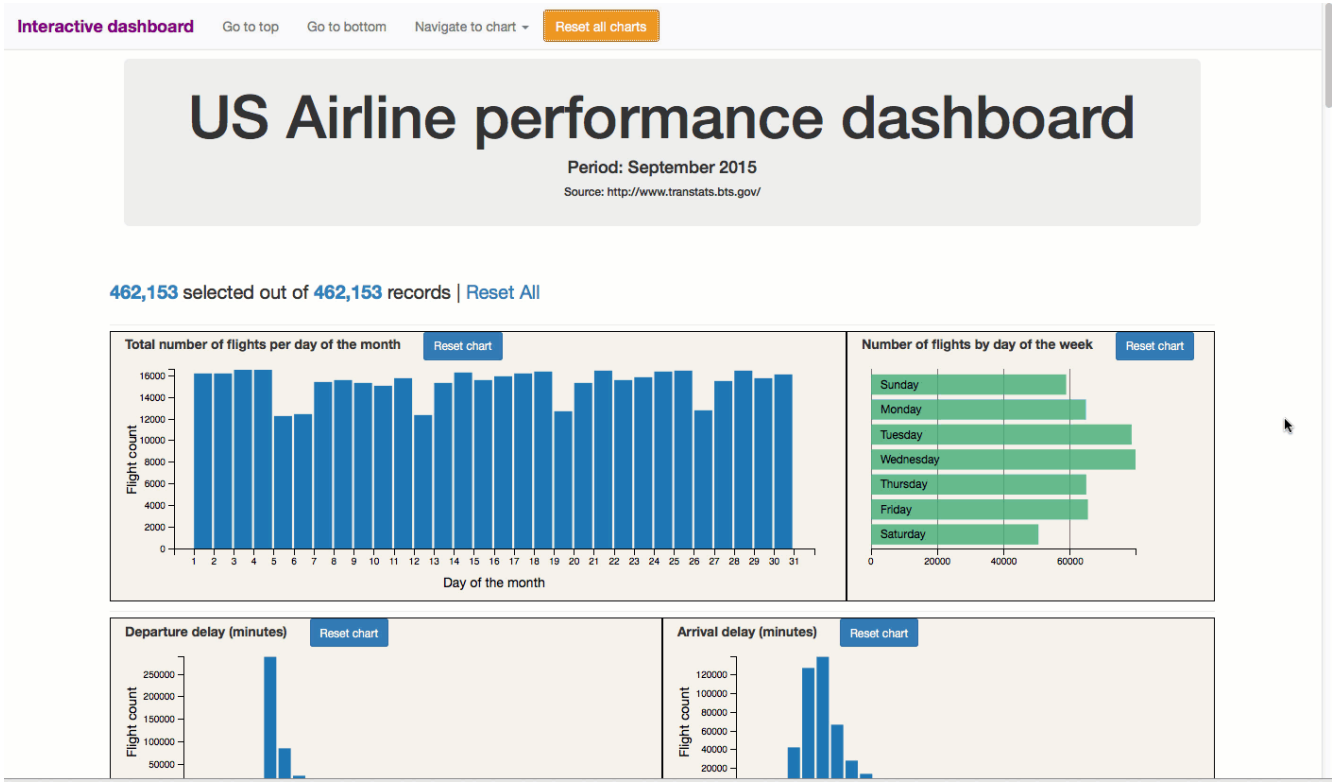
Scatter matrix –
Correlation matrix plot



Clustering (kmeans)

Why data visualization and visual storytelling?

- Cross-filtering of data.
- Create living-interactive documents.
- Communicating information in an effective way to a general audience.



Why data visualization and visual storytelling?

- And finally, because we have raw data and big tables are difficult to follow.

Sat Nov 14 21:39:05 CET 2015	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58	Left
Sat Nov 14 21:39:10 CET 2015	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76	Left
Sat Nov 14 21:39:15 CET 2015	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71	Right
Sat Nov 14 21:39:20 CET 2015	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84	Top
Sat Nov 14 21:39:25 CET 2015	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47	Top
Sat Nov 14 21:39:30 CET 2015	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04	Bottom
Sat Nov 14 21:39:35 CET 2015	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25	Bottom
Sat Nov 14 21:39:40 CET 2015	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50	Right
Sat Nov 14 21:39:45 CET 2015	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56	Up
Sat Nov 14 21:39:50 CET 2015	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91	Up
Sat Nov 14 21:39:55 CET 2015	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89	Top

Do we really need data visualization?

An enlightening example of data visualization

- Let us visually inspect the following table (raw data).
- Do you spot any correlation or peculiarity on this dataset?

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

An enlightening example of data visualization

- Even if the four datasets are different, they have nearly identical simple statistical properties.
- This dataset is known as anscombe's quartet.
- What will we see when the data is graphed?

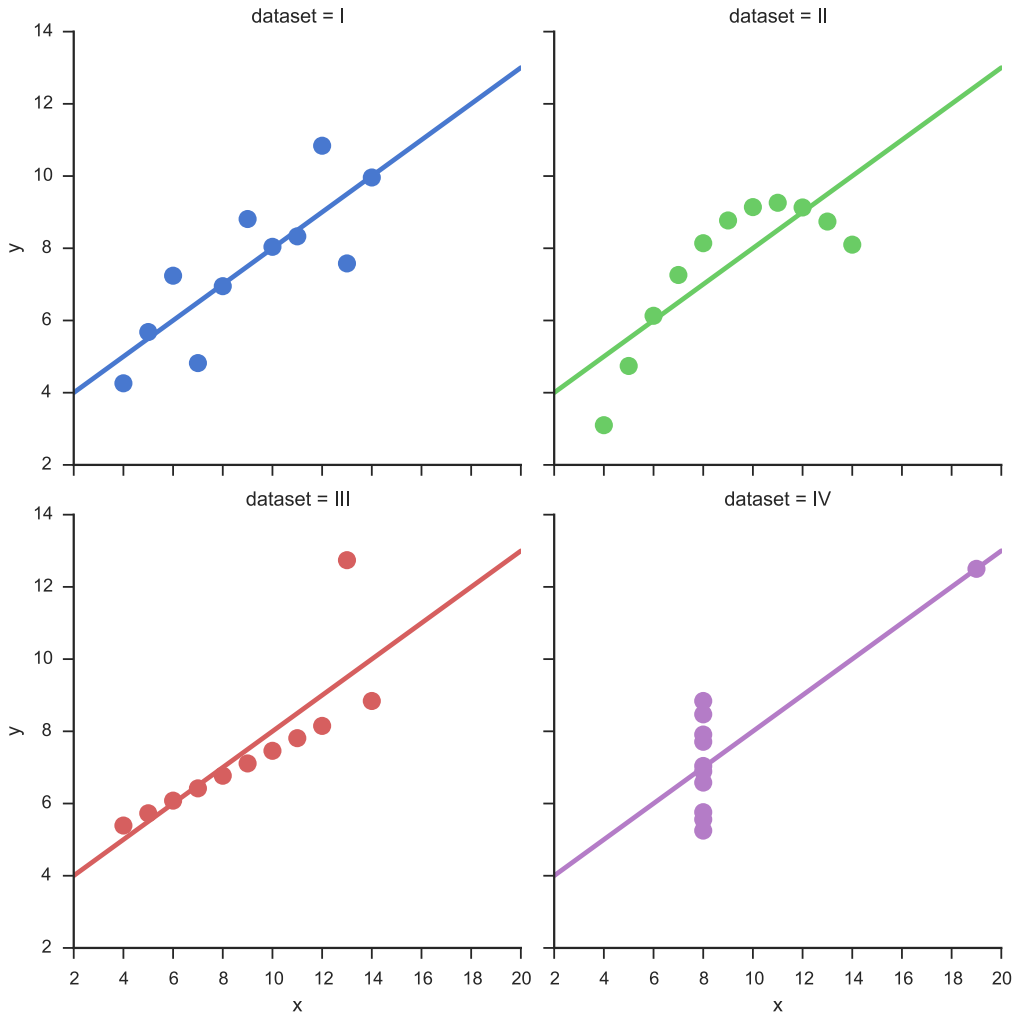
I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

For all datasets:

Statistical property	Value
Sample size	11
Mean (x)	9
Variance (x)	11
Mean (y)	7.50
Variance (y)	4.122
Correlation	0.816
Linear regression	Y = 3.00 + 0.5000X

An enlightening example of data visualization

- Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.



For all datasets:

Statistical property	Value
Sample size	11
Mean (x)	9
Variance (x)	11
Mean (y)	7.50
Variance (y)	4.122
Correlation	0.816
Linear regression	Y = 3.00 + 0.5000X

The how of data visualization and visual storytelling

Our approach

- **A web-based interactive data visualization and analysis toolkit.**
  - The toolkit uses:
    - On the client side javascript, jQuery, D3.js, webGL, bootstrap, and html5.
    - On the server side node.js, Python, and R.
  - We speak the language of the web.
  - We are able to control every pixel of the screen.
  - The server tools give us access to extensive, advanced and scalable data analytics capabilities.

Our approach

- **A web-based interactive data visualization and analysis toolkit.**

  - As the tools are implemented using the language of the web (javascript and html5), they can run from any device with a working web browser (PC, tablet, smart-phone, raspberry pi).
  - The learning curve is minimal as the user only needs to interact with the web browser interface.
  - The tool supports DSV*, JSON, XML and SQL data formats.

* Delimiter-separated value. CSV and TSV files are examples of DSV data files.

About the data that can be used

- The data can be obtained from any discipline (social sciences, econometrics, marketing, the social web, sports, health care, bioinformatics, engineering, etc.) or the user's daily activity (blood pressure measurement, time to arrive to your workplace, daily calories intake, etc.).
- However, during this presentation I will mainly address data obtained from numerical simulations and optimization studies.
- We are talking about

Visual storytelling for CFD

The toolkit enables the user to do

- **Data visualization and exploration.**
  - Plotting of multidimensional data.
  - Machine learning and predictive analytics.
  - Summary statistics (but do not just rely on this).
- **Interactive visualization.**
  - Manipulation and exploration of the data.
  - Cross-filtering of data.
  - Summaries with access to the details.
- **Reports and data communication.**
  - You can create living documents.
- **More exploration, more connections, more interaction, more insight.**

Show me some numbers and plots!

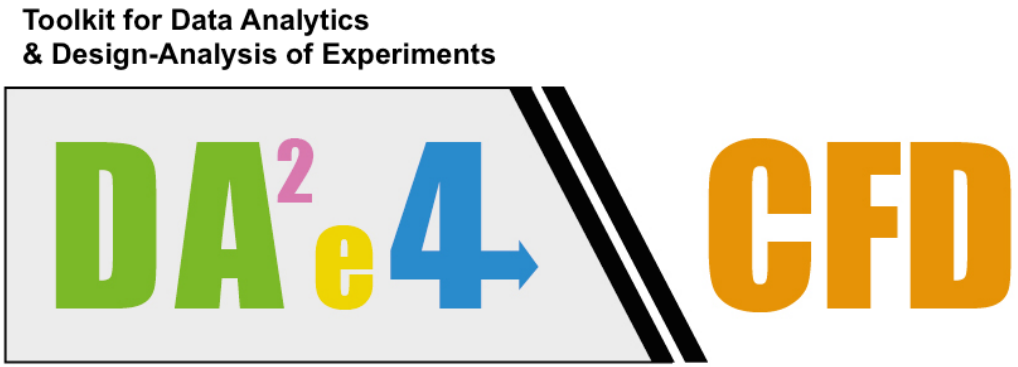
“Show data variation and not design variation.”

— Edward Tufte

On the datasets and scripts

The following datasets and scripts are available in the following link:

`https://github.com/joelguerrero/dae4cfd`
`http://joelguerrero.github.io/p1/`



Dataset 1

Column 1	Column 2	Column 3	Column 4	Column 5
3.95E+00	1.63E+00	0.44242	0.0055127	0.0666262
7.11E+00	3.44E+00	0.860229	0.00615626	0.174343
6.90E+00	1.78E+00	0.782869	0.0089705	0.122455
7.79E+00	3.82E+00	0.960738	0.00650422	0.203037
1.25E+00	3.28E+00	0.158067	0.00373114	0.031634
5.63E+00	2.11E+00	0.643855	0.00648422	0.107407
...				
N				

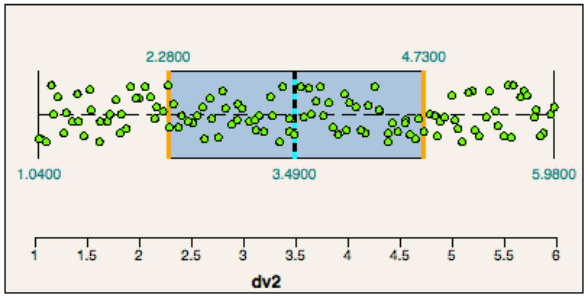
- Data obtained from a design space exploration study.
- Sample size: 5 X 121
- All the data is numerical.
- Tidy data.

How do I explore this dataset or any dataset?

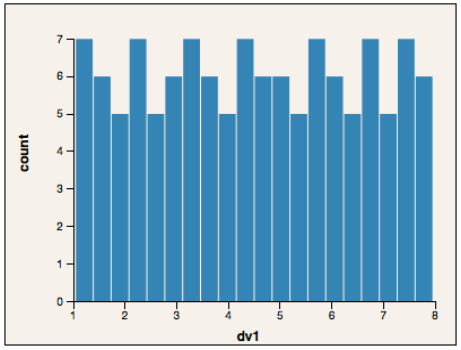
- Getting and prepping the data.
- Cleaning the data. If the data is tidy data, clear skies.
- Most of the times the data is not clean so you need to do some data wrangling, and this the most annoying and time consuming part.
- Visualizing the data.
- Drawing conclusions.

How do I explore this dataset or any dataset?

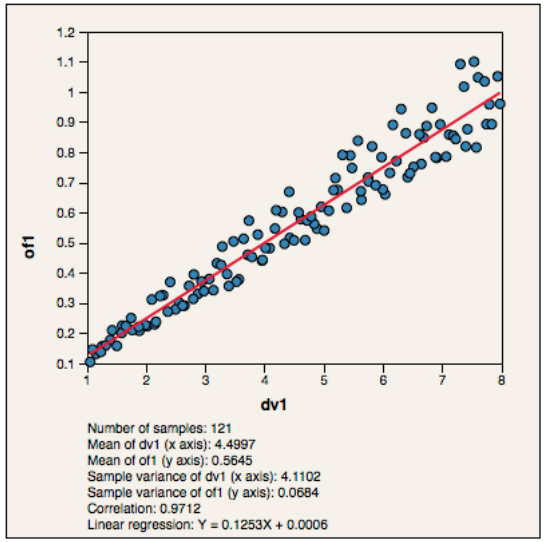
Boxplot



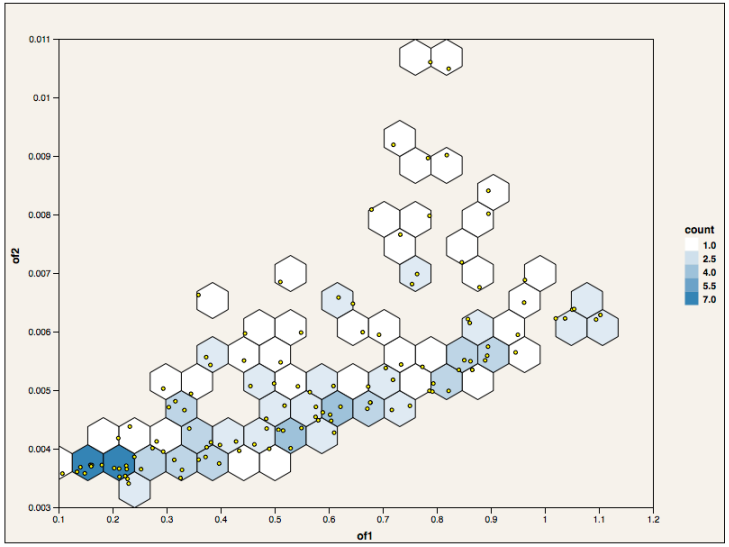
Histogram



Scatter plot + regression

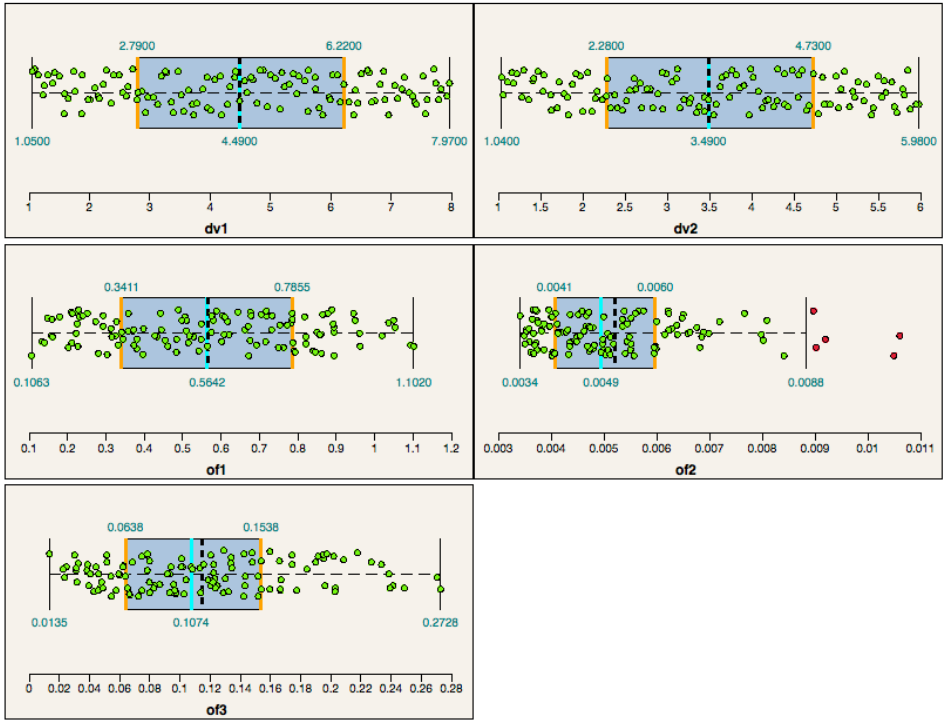


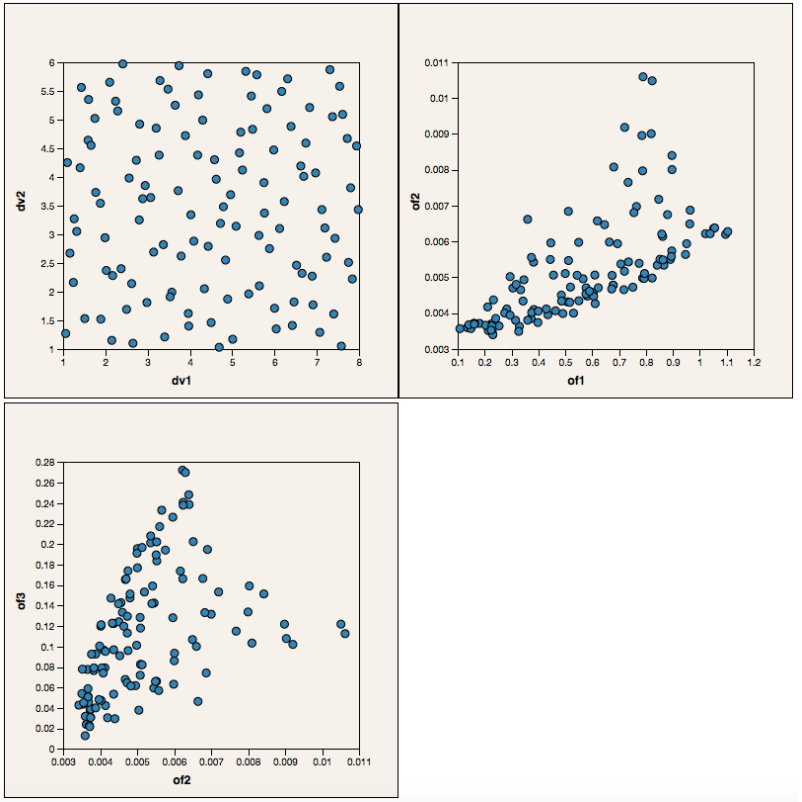
Hexbin



How do I explore this dataset or any dataset?

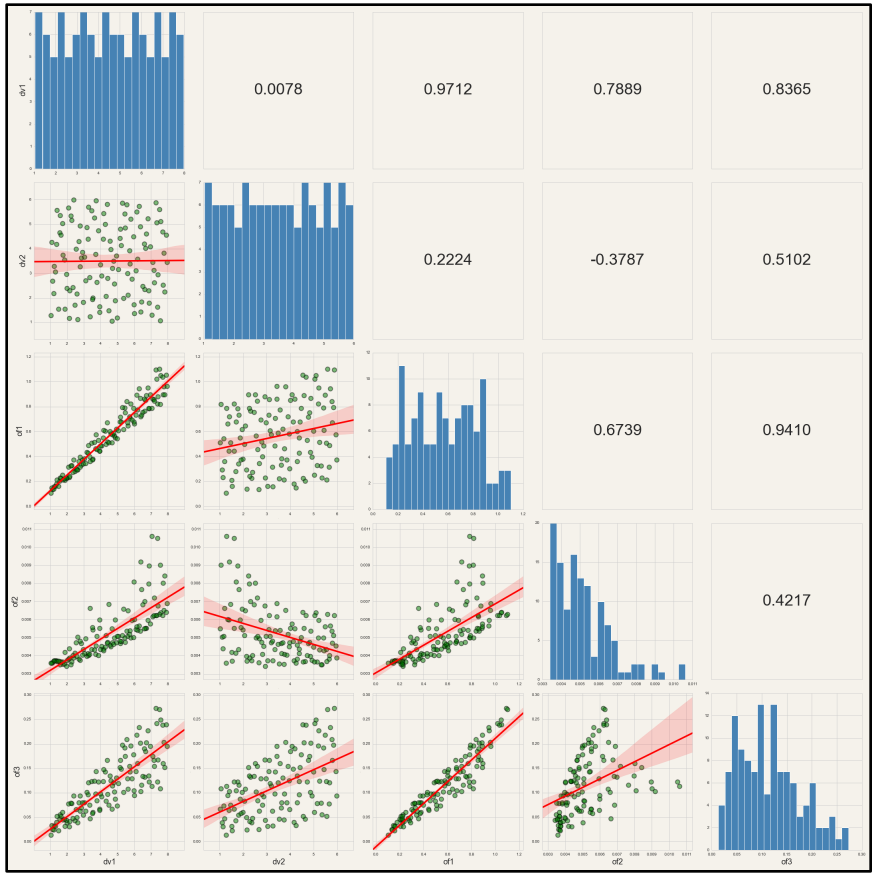
Small multiples (univariate data), paired-data and faceted plots (bivariate data)

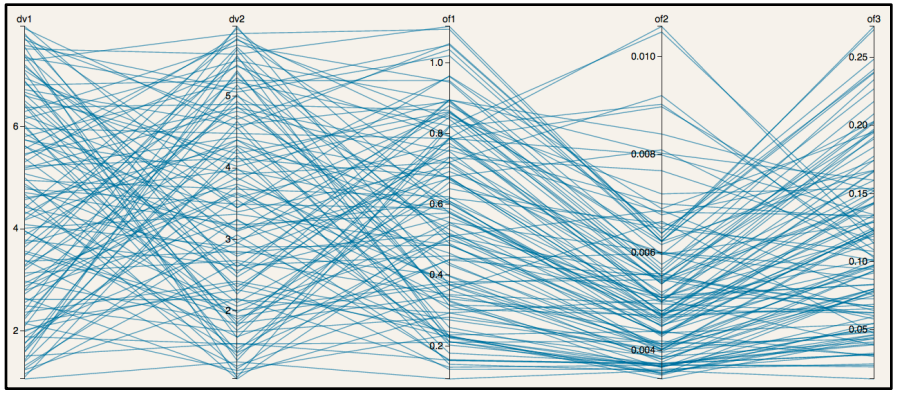




How do I explore this dataset or any dataset?

Scatter matrix plot and parallel coordinates for multivariate data





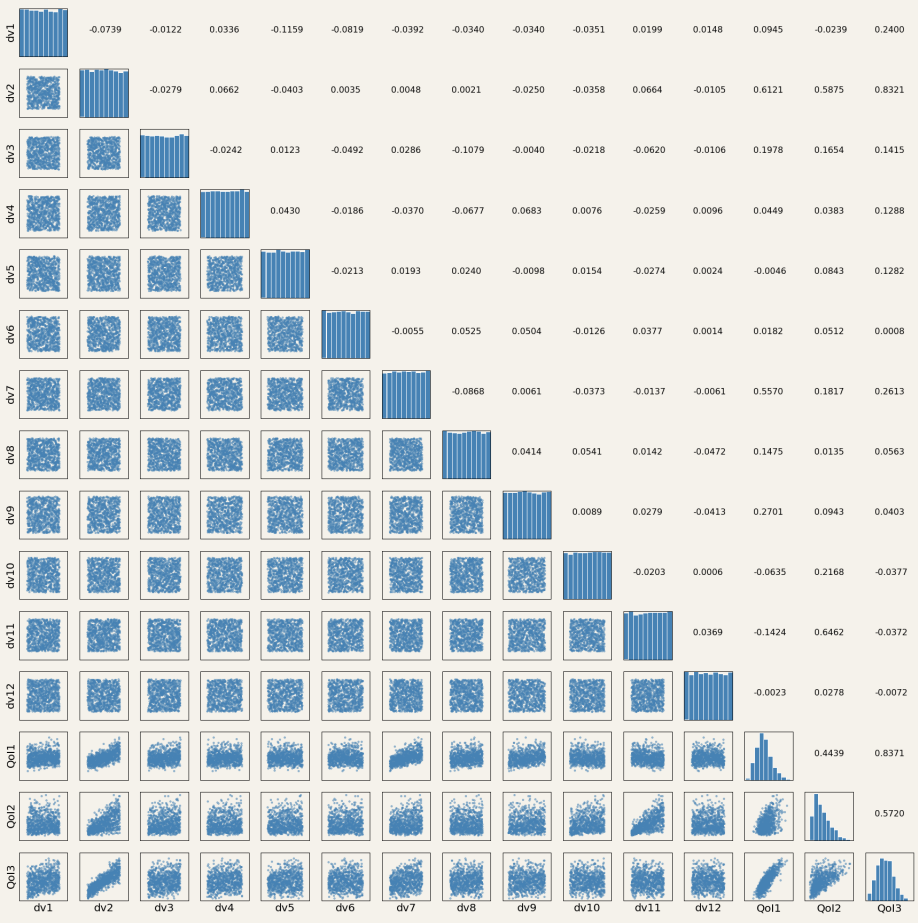
Dataset 2

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
4.87E-01	1.43E-01	5.90E-01	7.99E-01	8.64E-01	6.29E-01	2.08E-01	2.28E-01	2.58E-01	7.49E+00	3.88E+01	-1.34E+01	35.2251	122.277	339.623
2.85E-01	1.43E-01	6.49E-01	6.33E-02	8.88E-01	9.87E-01	3.09E-01	2.24E-01	1.66E-01	1.32E+00	9.31E-01	-4.96E-01	42.1183	4.13759	309.449
3.49E-01	5.40E-01	1.93E-01	4.72E-01	6.07E-02	3.80E-01	2.60E-01	2.02E-01	1.47E-01	6.05E+00	2.95E+00	-4.68E+00	41.7185	37.3891	467.299
3.07E-01	4.09E-01	5.77E-01	1.81E-02	1.97E-01	2.30E-01	3.44E-01	2.36E-01	2.02E-01	3.21E+00	3.34E+01	-1.17E+01	45.6606	155.98	484.209
6.53E-01	1.56E-02	5.53E-01	9.93E-02	6.88E-01	2.09E-01	3.05E-01	3.46E-01	1.50E-01	8.64E+00	1.12E-01	-6.07E+00	45.3468	30.2206	291.584
3.09E-01	7.93E-01	3.16E-01	4.10E-01	9.95E-01	2.74E-01	3.26E-01	2.82E-01	1.88E-01	3.95E+00	6.59E+00	-2.71E+00	55.0919	79.864	930.964
...														
N														

- Data obtained from a design space exploration study.
- Sample size: 15 X 777
- All the data is numerical.
- Not so tidy data.

Dataset 2

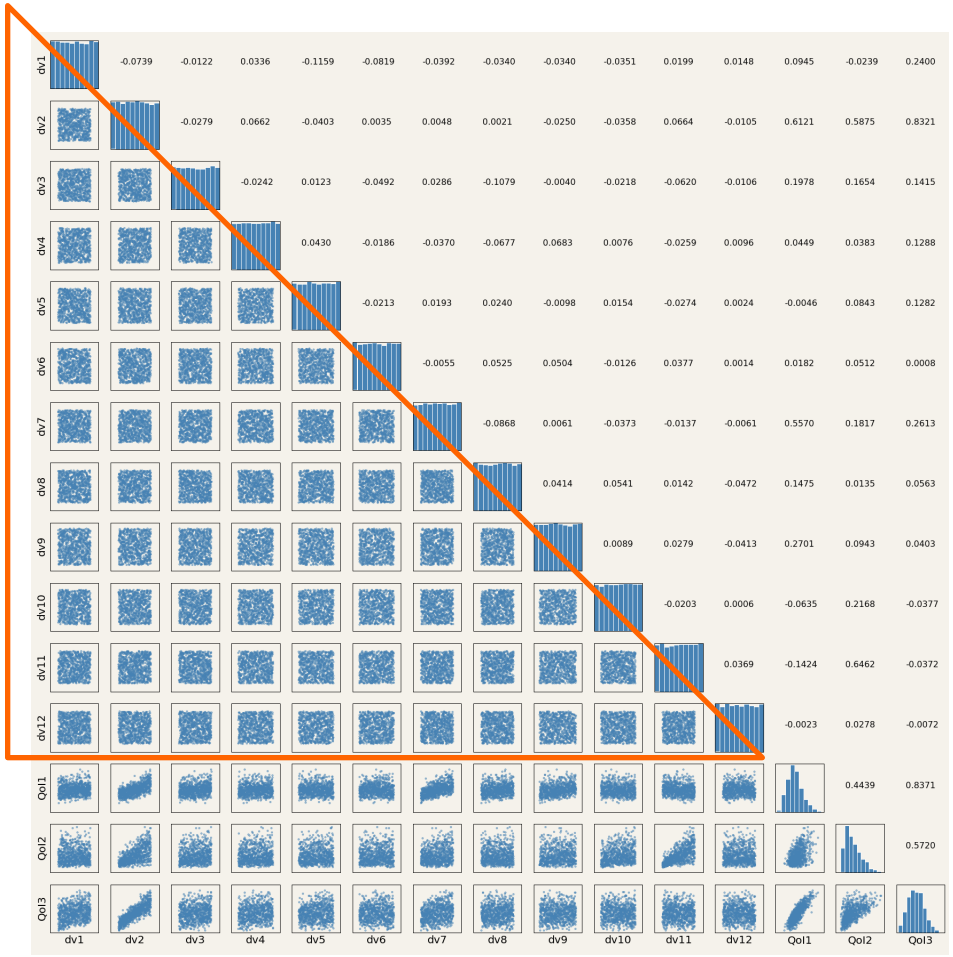
Scatter matrix plot of a design space exploration study
This plot shows correlation, skewness, kurtosis, tendency and distribution of the data



Dataset 2

Scatter matrix plot of a design space exploration study

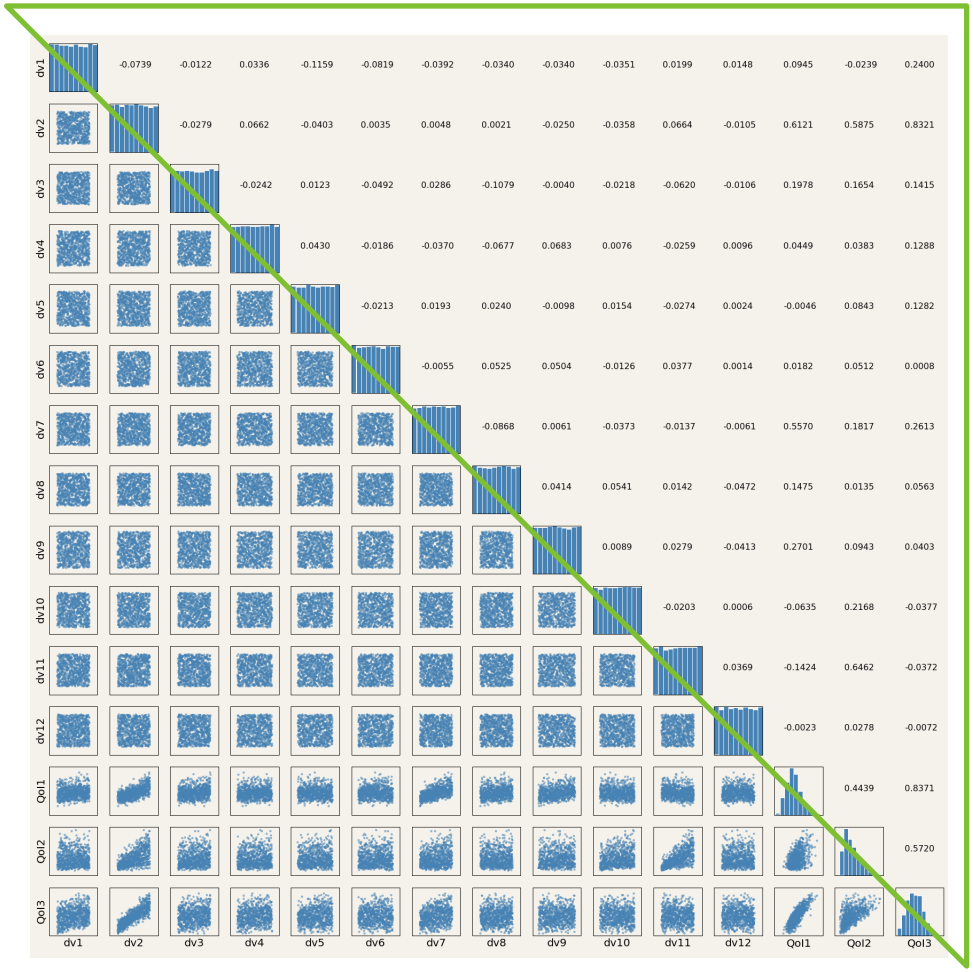
Scatter plot of design variables distribution (sampling distribution in design space)



Dataset 2

Scatter matrix plot of a design space exploration study

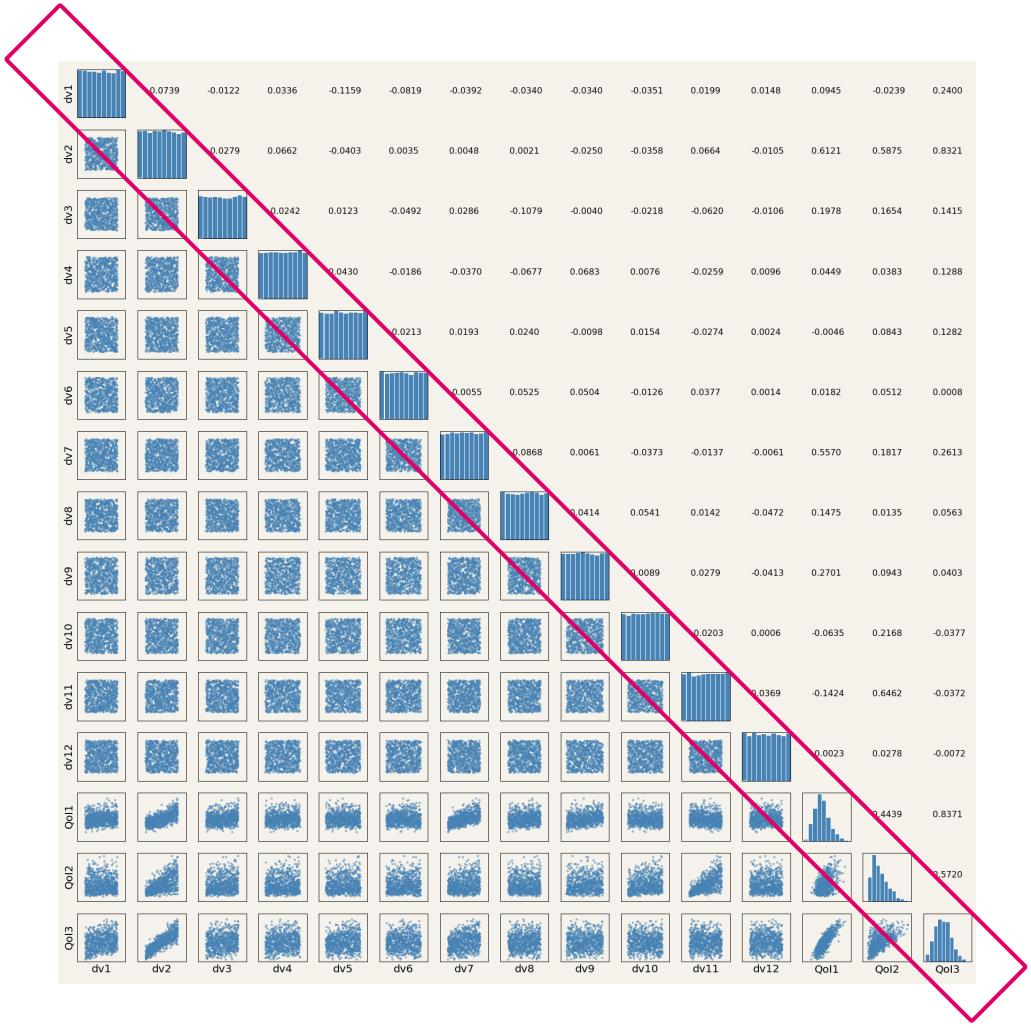
Correlation matrix of design space variables (design variables and objective functions)



Dataset 2

Scatter matrix plot of a design space exploration study

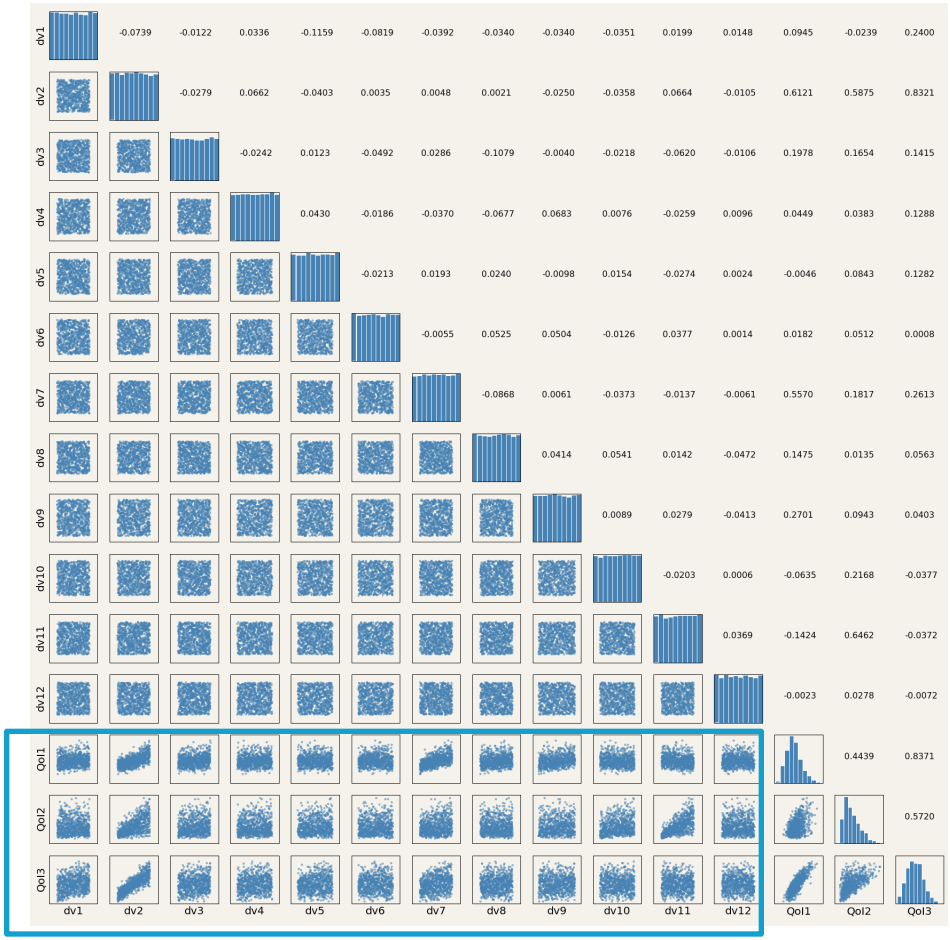
Histograms of design variables and objective functions



Dataset 2

Scatter matrix plot of a design space exploration study

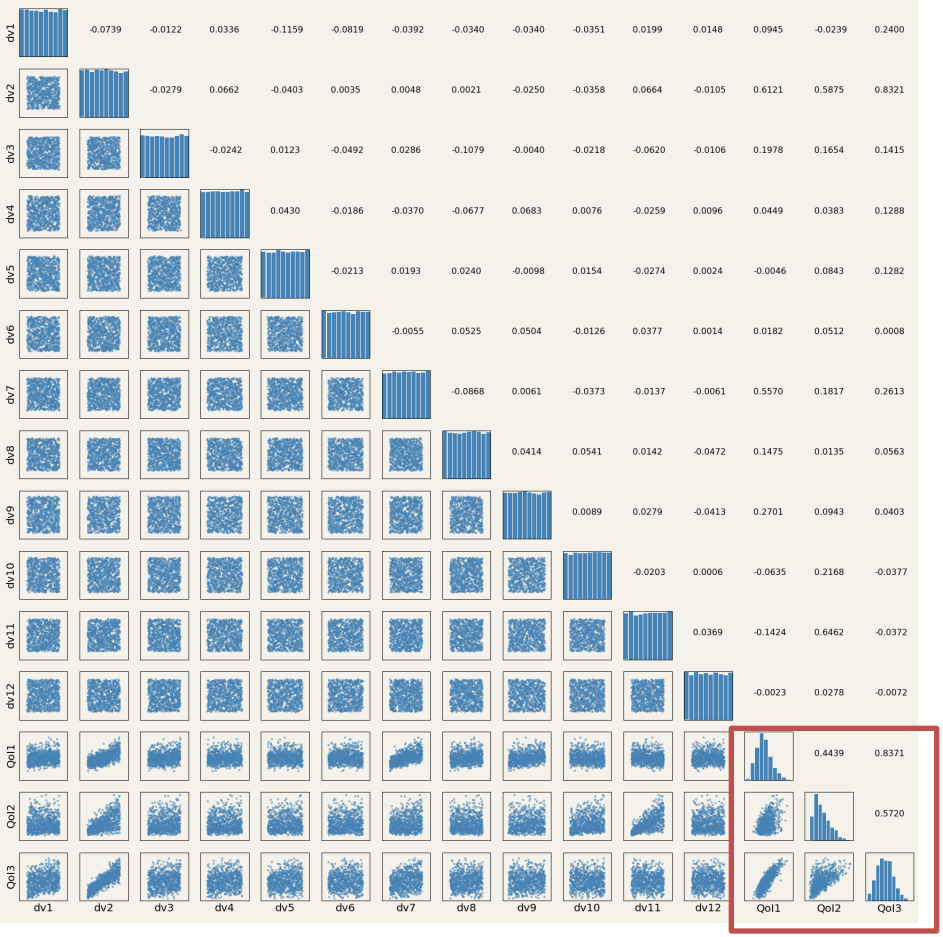
Response of design space (scatter plot of design variables vs. objective functions)



Dataset 2

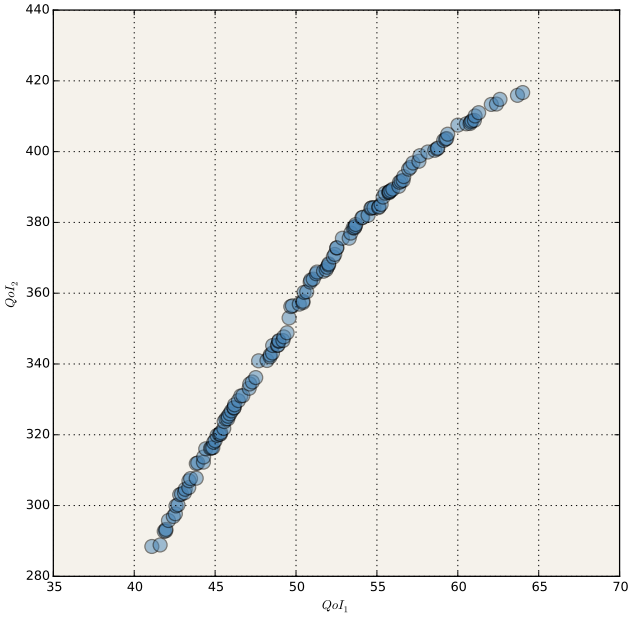
Scatter matrix plot of a design space exploration study

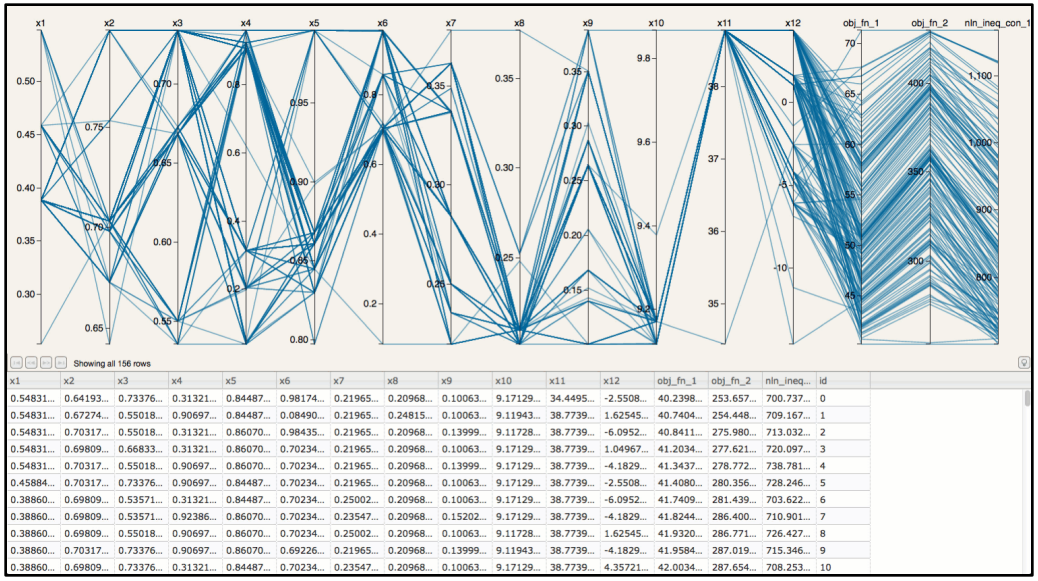
Response or trade-off of objective functions



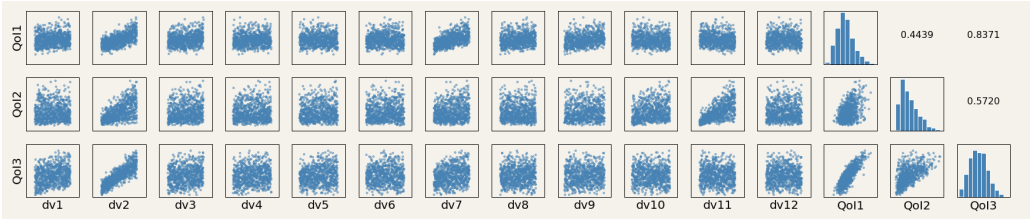
Dataset 2

Pareto front and overall response of the design space.





x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	obj_fn_1	obj_fn_2	nln_ineq...	id
0.54831...	0.64193...	0.73376...	0.31321...	0.84487...	0.98174...	0.21965...	0.20968...	0.10063...	9.17129...	34.4495...	-2.5508...	40.2398...	253.657...	700.737...	0
0.54831...	0.67274...	0.55018...	0.90697...	0.84487...	0.08490...	0.21965...	0.24815...	0.10063...	9.11943...	38.7739...	1.62545...	40.7404...	254.448...	709.167...	1
0.54831...	0.70317...	0.55018...	0.31321...	0.86070...	0.98435...	0.21965...	0.20968...	0.13999...	9.11728...	38.7739...	-6.0952...	40.8411...	275.980...	713.032...	2
0.54831...	0.69809...	0.66833...	0.31321...	0.86070...	0.70234...	0.21965...	0.20968...	0.10063...	9.17129...	38.7739...	1.04967...	41.2034...	277.621...	720.097...	3
0.54831...	0.70317...	0.55018...	0.90697...	0.86070...	0.70234...	0.21965...	0.20968...	0.13999...	9.17129...	38.7739...	-4.1829...	41.3437...	278.772...	738.781...	4
0.45884...	0.70317...	0.73376...	0.90697...	0.84487...	0.70234...	0.21965...	0.20968...	0.10063...	9.17129...	38.7739...	-2.5508...	41.4080...	280.356...	728.246...	5
0.38860...	0.69809...	0.53571...	0.31321...	0.84487...	0.70234...	0.25002...	0.20968...	0.10063...	9.17129...	38.7739...	-6.0952...	41.7409...	281.439...	703.622...	6
0.38860...	0.69809...	0.53571...	0.92386...	0.86070...	0.70234...	0.23547...	0.20968...	0.15202...	9.17129...	38.7739...	-4.1829...	41.8244...	286.400...	710.901...	7
0.38860...	0.69809...	0.55018...	0.90697...	0.86070...	0.70234...	0.25002...	0.20968...	0.10063...	9.11728...	38.7739...	1.62545...	41.9320...	286.771...	726.427...	8
0.38860...	0.70317...	0.73376...	0.90697...	0.86070...	0.69226...	0.21965...	0.20968...	0.13999...	9.11943...	38.7739...	-4.1829...	41.9584...	287.019...	715.346...	9
0.38860...	0.69809...	0.73376...	0.31321...	0.84487...	0.70234...	0.23547...	0.20968...	0.10063...	9.17129...	38.7739...	4.35721...	42.0034...	287.654...	708.253...	10



- The Pareto front was constructed using surrogate based optimization.
- How do we relate the trade-offs in the Pareto front with the design variables?

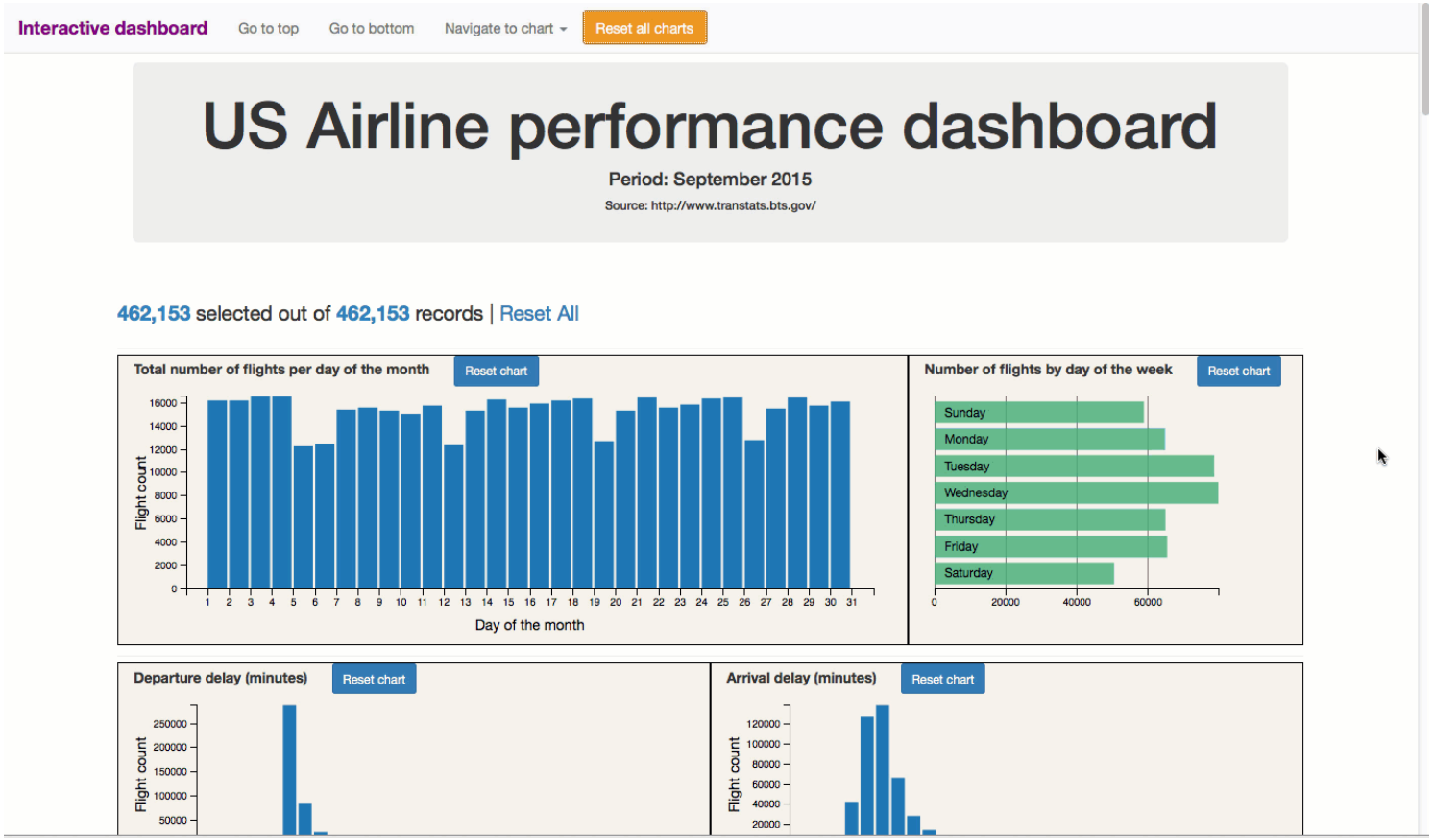
Dataset 3

YEAR	MONTH	DAY OF MONTH	DAY OF WEEK	ORIGIN	DESTINATION	DEPARTURE TIME	DEPARTURE DELAY	ARRIVAL DELAY	DISTANCE
2015	9	1	2	JFK	LAX	0853	-7	-28	2475
2015	9	2	3	JFK	LAX	0854	-6	14	2475
2015	9	3	4	JFK	LAX	0857	-3	-14	2475
2015	9	4	5	JFK	LAX	0852	-8	-32	2475
2015	9	5	6	JFK	LAX	0846	-14	-26	2475
2015	9	6	7	JFK	LAX	0855	-5	-58	2475
...									
N									

- Data obtained from the web (US airline performance data).
- Sample size: 10 X 464947
- Numerical data, categorical data and timestamps.
- Tidy data with missing values.

Dataset 3

- This example is about cross filtering data.
- X-filtering is about finding common dimensions, grouping data, using aggregators, filtering data and building interactive dashboards.



Key takeaways

- Data is being used by everybody all the time, and numerical simulations do not escape to this reality. The way to analyze and visualize the data is same.
- CFD is not anymore about submitting a few simulations and waiting long times. We should be ready to deal with large amount of quantitate data.
- At the end of the day, exploratory data analysis and data visualization is about presenting the data in a form which is comprehensible, insightful and actionable to anybody.
- Data visualization and visual storytelling is not just about a pretty picture but a structured, accurate visual presentation of evidence.

Good intentions

- UI, HCI and UX.
- Advanced ML and SL.
- UQ.
- Google bigQuery and webGL (big data visualization).
- Strong integration with a GUI framework (DICE http://dicehub.net/).

“Principles for the Development of a Complete Mind: Study the science of art. Study the art of science. Develop your senses - especially learn how to see. Realize that everything connects to everything else.”

— Leonardo da Vinci