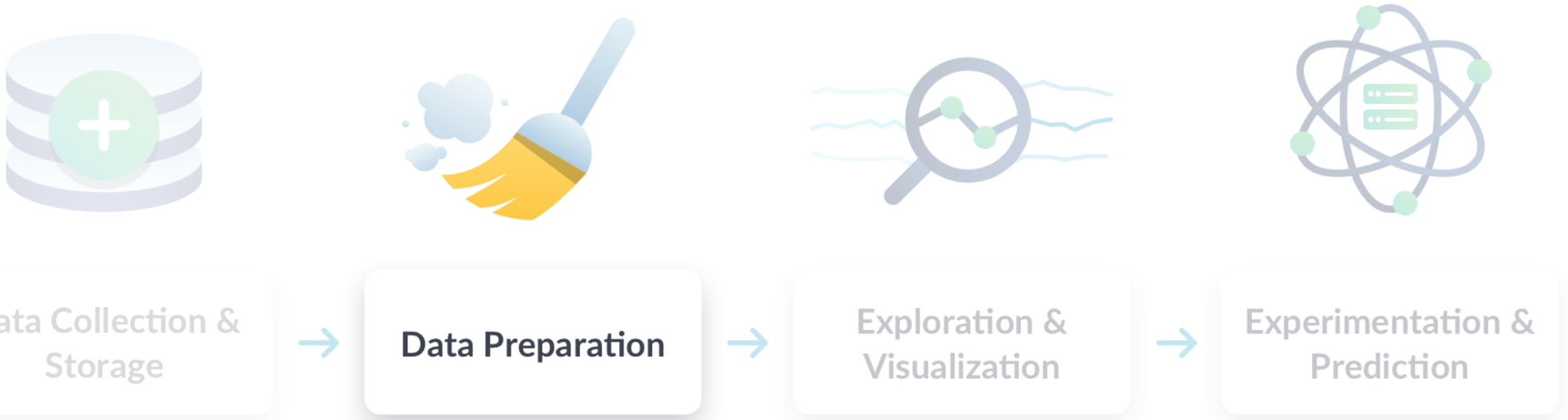
Data preparation

UNDERSTANDING DATA SCIENCE

Hadrien Lacroix
Content Developer, DataCamp

Data workflow



Why prepare data?

- Real-life data is messy
- Preparation is done to prevent:
  - errors
  - incorrect results
  - biasing algorithms

Let's start cleaning

	Sara	Lis	Hadrien	Lis
Age	"27"	"30"		"30"
Size	1.77	5.58	1.80	5.58
Country	"Belgium"	"USA"	"FR"	"USA"

Tidy data

Before

	Sara	Lis	Hadrien	Lis
Age	"27"	"30"		"30"
Size	1.77	5.58	1.80	5.58
Country	"Belgium"	"USA"	"FR"	"USA"

Tidy data output

Before

	Sara	Lis	Hadrien	Lis
Age	"27"	"30"		"30"
Size	1.77	5.58	1.80	5.58
Country	"Belgium"	"USA"	"FR"	"USA"

After

Name	Age	Size	Country
Sara	"26"	1.78	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.58	"USA"

Remove duplicates

Before

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.58	"USA"

Remove duplicates | output

Before

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.58	"USA"

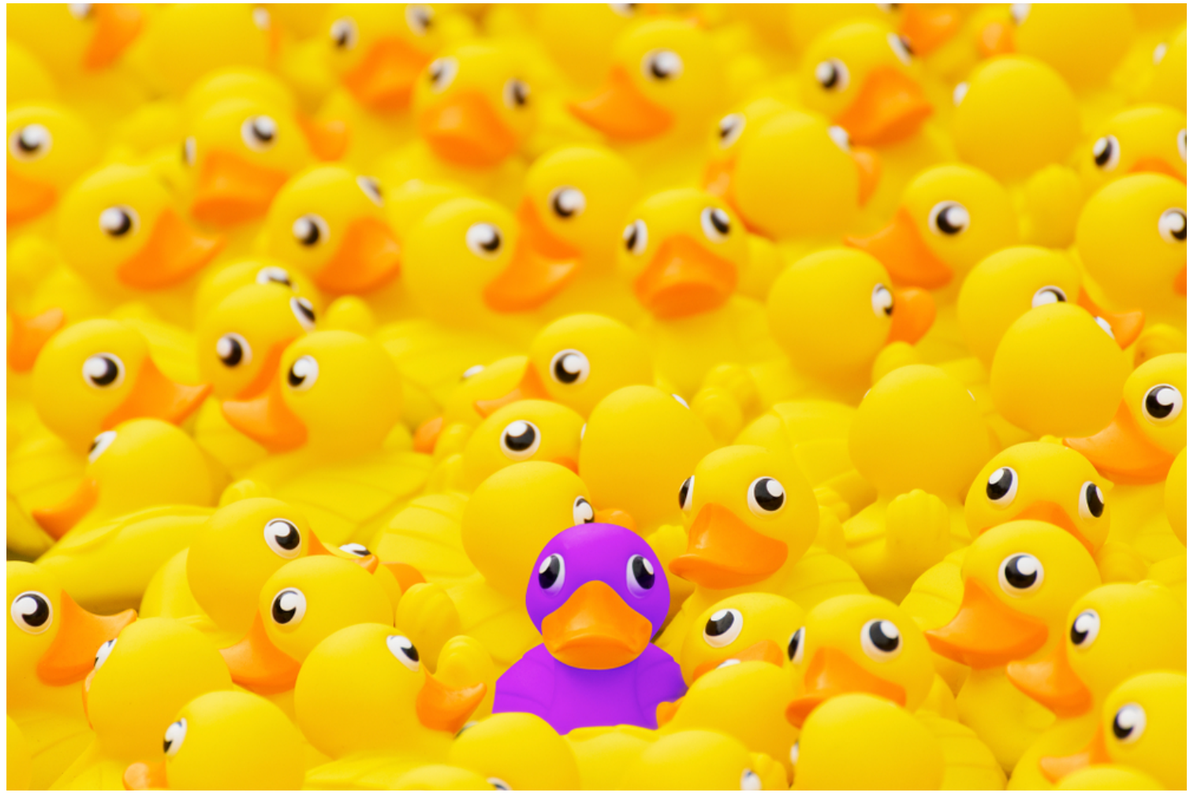
After

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"

Unique ID

Before

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"



Unique ID | output

Before

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"

After

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.58	"USA"
2	Hadrien		1.80	"FR"

Homogeneity

Before

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.58	"USA"
2	Hadrien		1.80	"FR"

Homogeneity | output

Before

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.58	"USA"
2	Hadrien		1.80	"FR"

After

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	1.70	"USA"
2	Hadrien		1.80	"FR"

Homogeneity, again

Before

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	1.70	"USA"
2	Hadrien		1.80	"FR"

Homogeneity, again | output

Before

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	1.70	"US"
2	Hadrien		1.80	"FR"

After

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"BE"
1	Lis	"30"	1.70	"US"
2	Hadrien		1.80	"FR"

Data types

Before

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"BE"
1	Lis	"30"	1.70	"US"
2	Hadrien		1.80	"FR"

Data types | output

Before

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"BE"
1	Lis	"30"	1.70	"US"
2	Hadrien		1.80	"FR"

After

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"US"
2	Hadrien		1.80	"FR"

Missing values

Before

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"US"
2	Hadrien		1.80	"FR"

Reasons:

- data entry
- error
- valid missing value

Solutions:

- impute
- drop
- keep

Missing values | output

Before

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"USA"
2	Hadrien		1.80	"FR"

After

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"US"
2	Hadrien	28	1.80	"FR"

Let's practice!

UNDERSTANDING DATA SCIENCE

Exploratory Data Analysis

UNDERSTANDING DATA SCIENCE

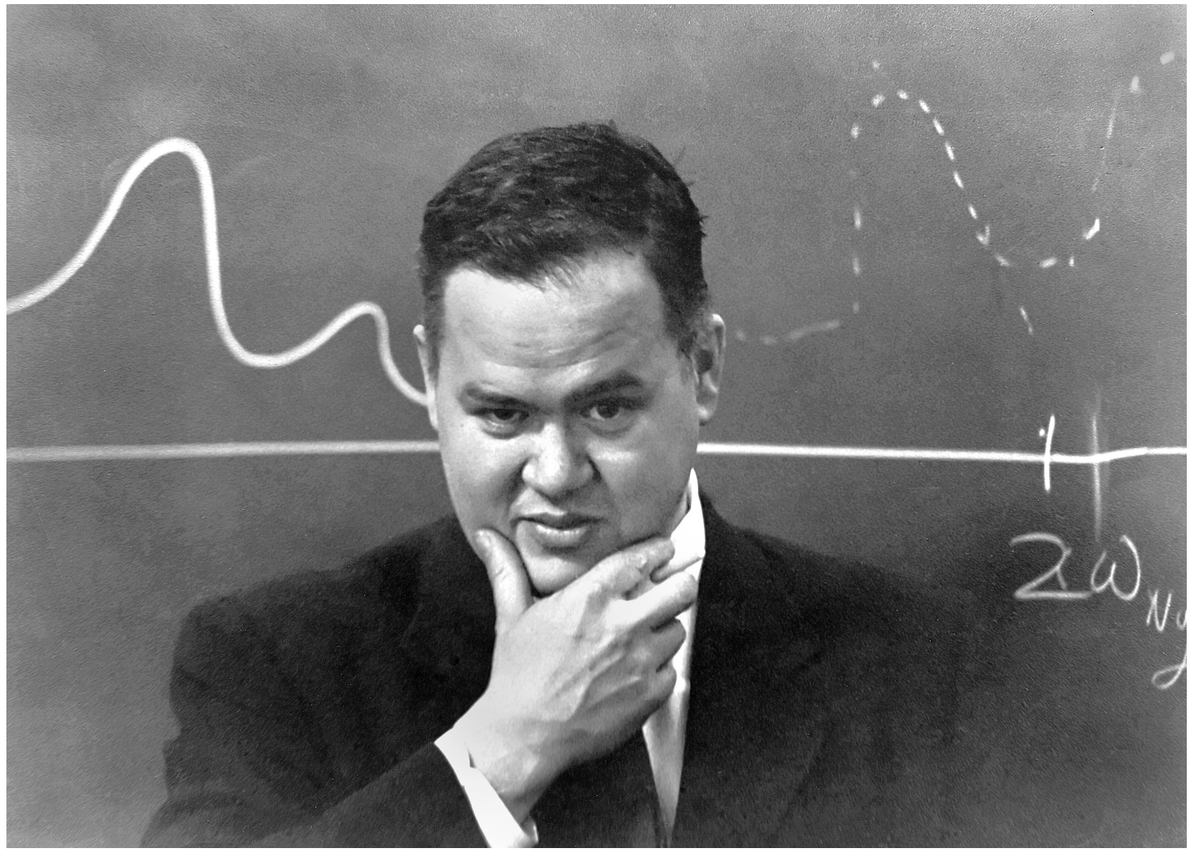
Hadrien Lacroix
Content Developer at DataCamp

datacamp

What is EDA?

Exploratory Data Analysis:

- Exploring the data
- Formulating hypotheses
- Assessing characteristics
- Visualizing



Data workflow

Data Collection & Storage → Data Preparation → **Exploration & Visualization** → Experimentation & Prediction

Let's dive right in

Dataset 1		Dataset 2		Dataset 3		Dataset 4	
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

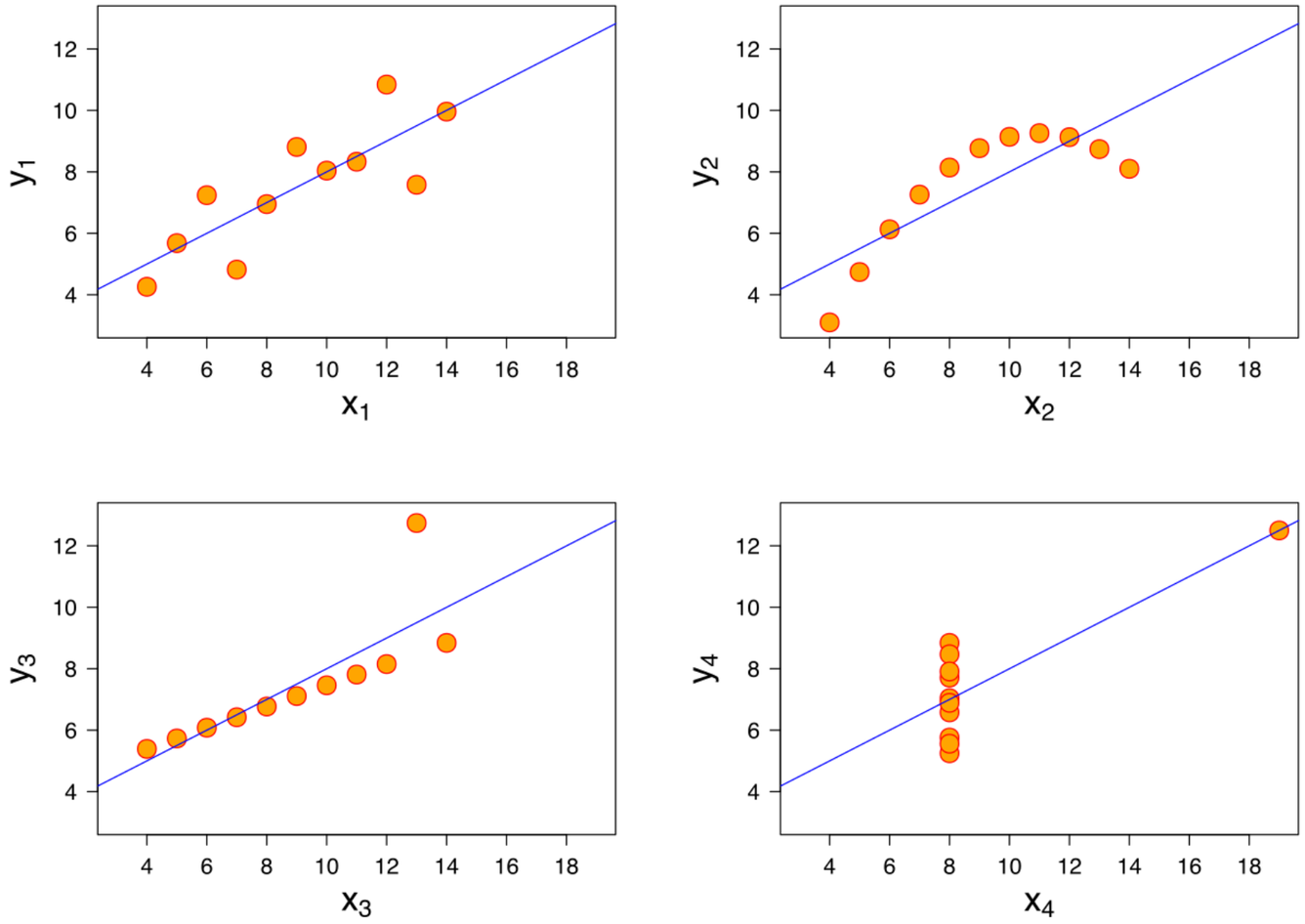
Surprise!

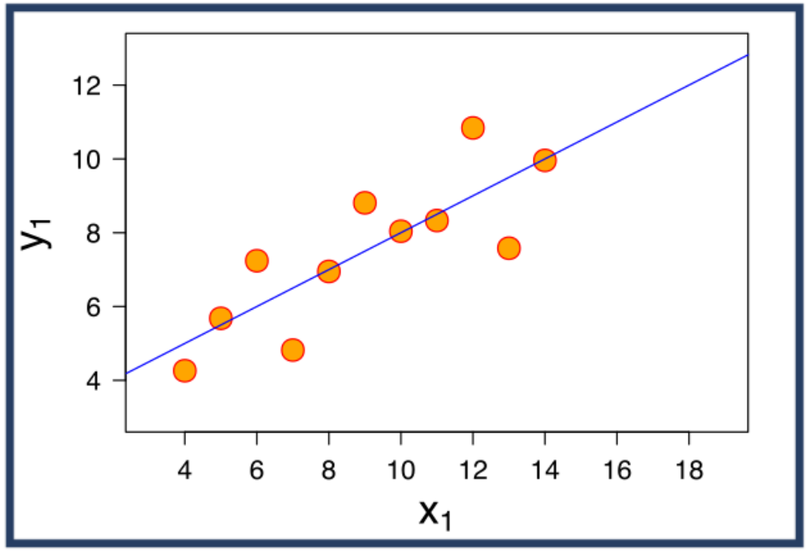
All four datasets display:

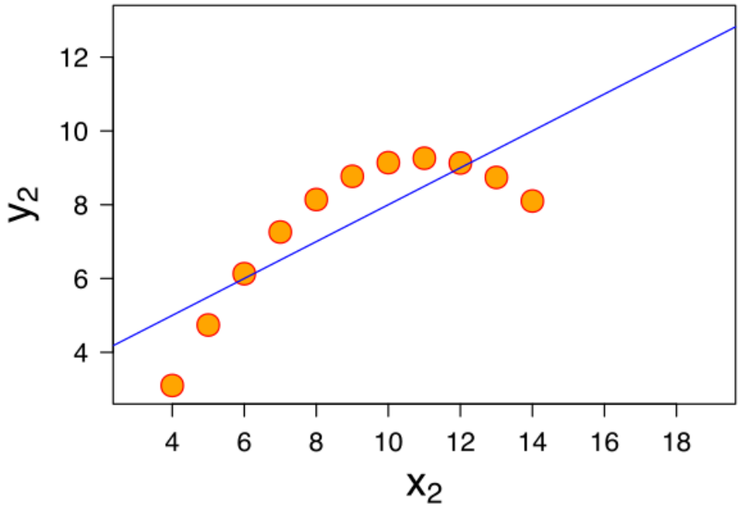
- identical mean and variance for x
- identical mean and variance for y
- identical correlation coefficient
- identical linear regression equation

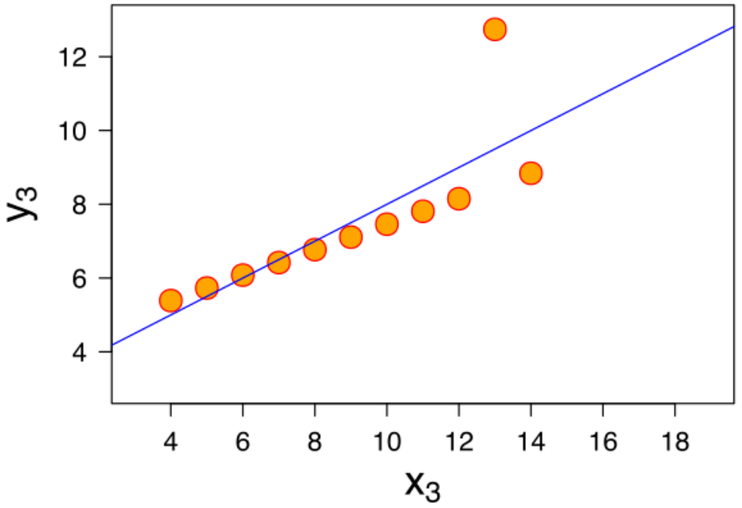
In short: **they look quite similar**

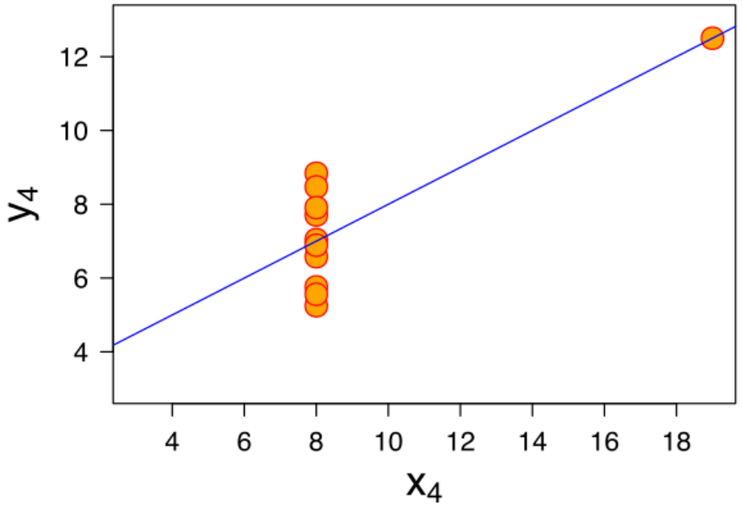
Anscombe's quartet

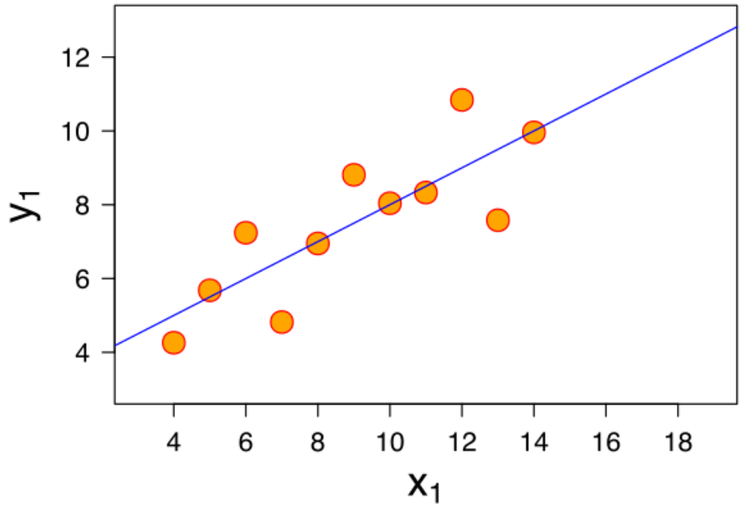


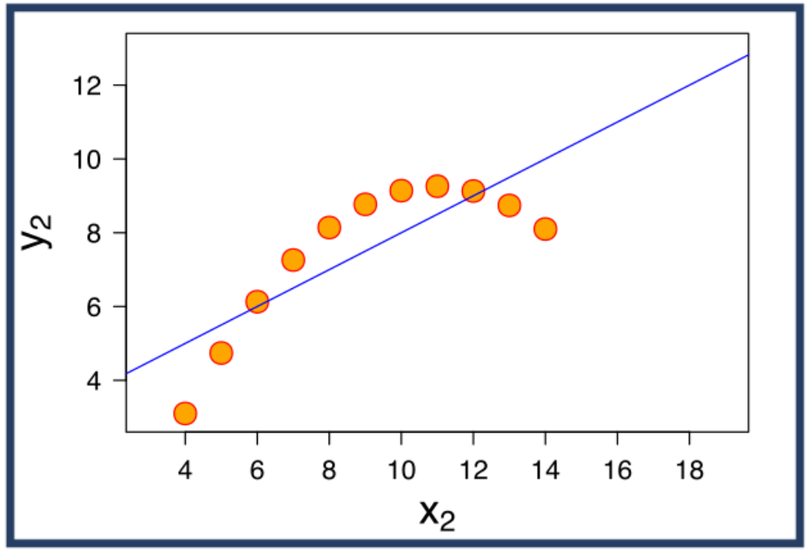

12
10
8
6
4
y_1
4 6 8 10 12 14 16 18
x_1

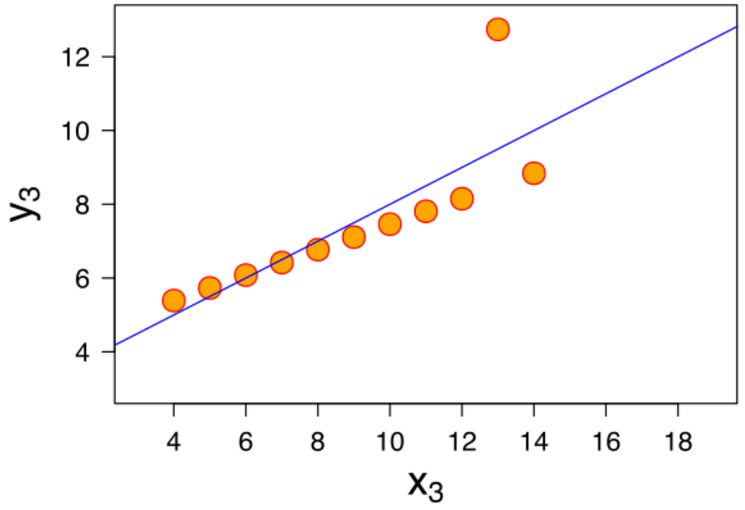

12
10
8
6
4
y_2
4 6 8 10 12 14 16 18
x_2

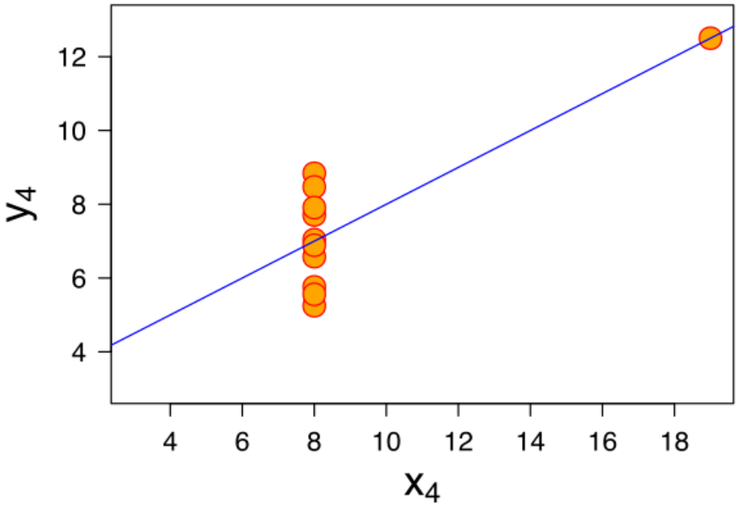

12
10
8
6
4
y_3
4 6 8 10 12 14 16 18
x_3

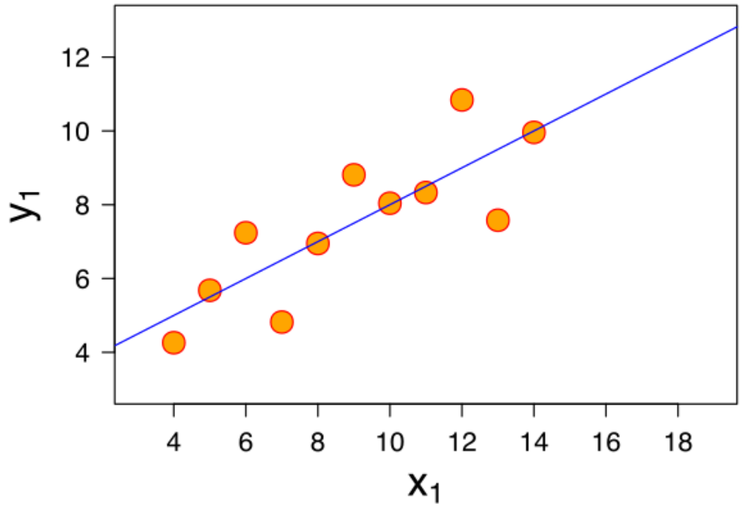

12
10
8
6
4
y_4
4 6 8 10 12 14 16 18
x_4

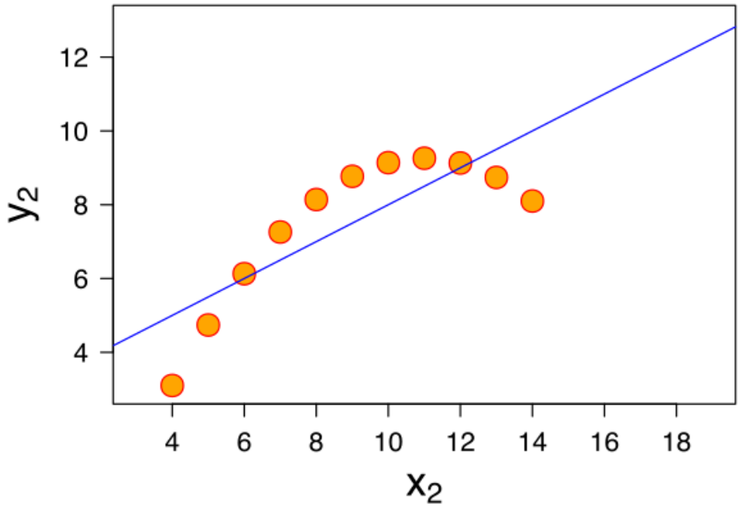

12
10
8
6
4
y_1
4
6
8
10
12
14
16
18
x_1

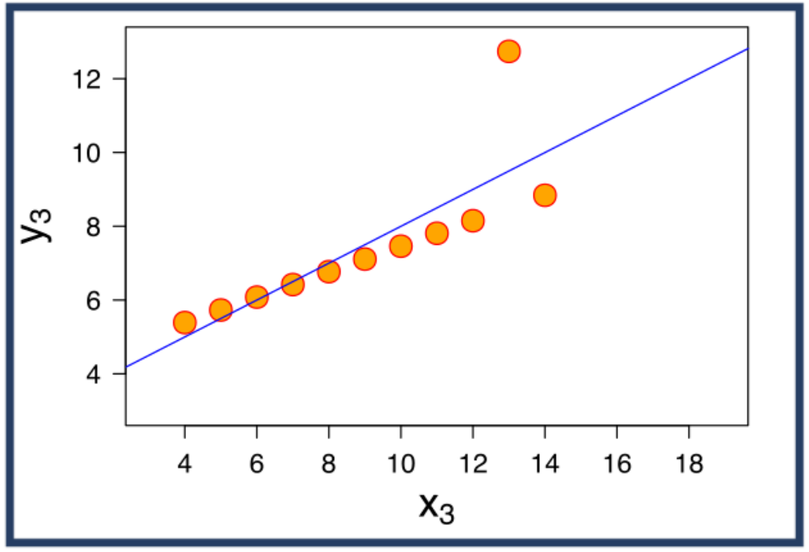

12
10
8
6
4
y_2
4
6
8
10
12
14
16
18
x_2

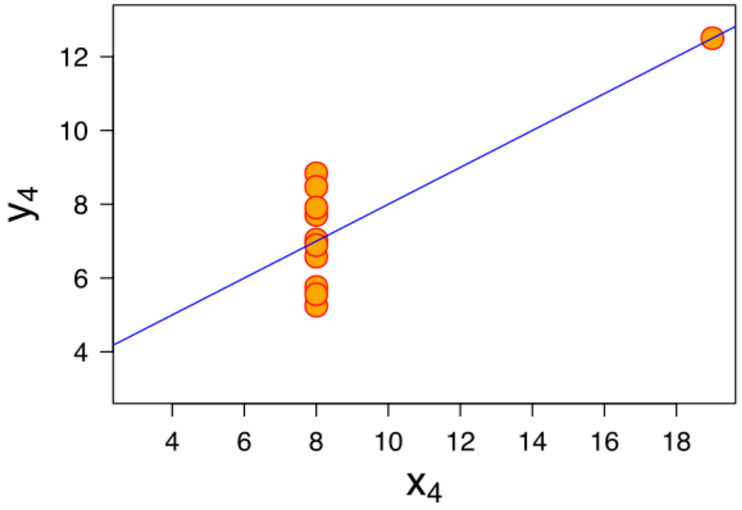

12
10
8
6
4
y_3
4
6
8
10
12
14
16
18
x_3

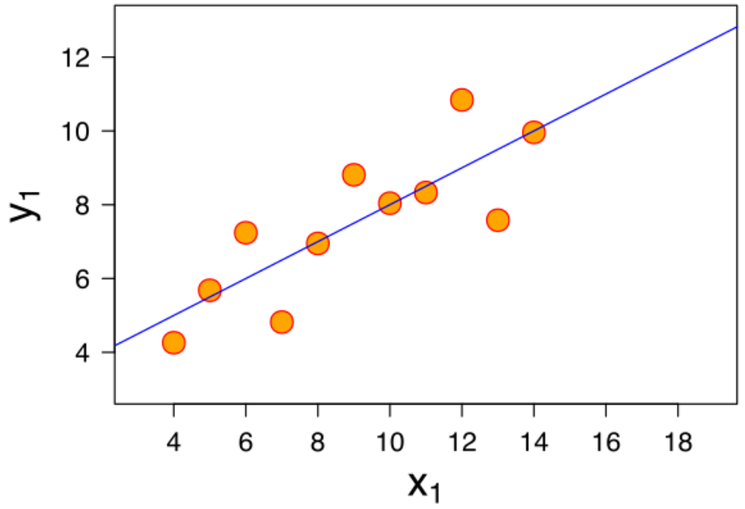

12
10
8
6
4
y_4
4
6
8
10
12
14
16
18
x_4

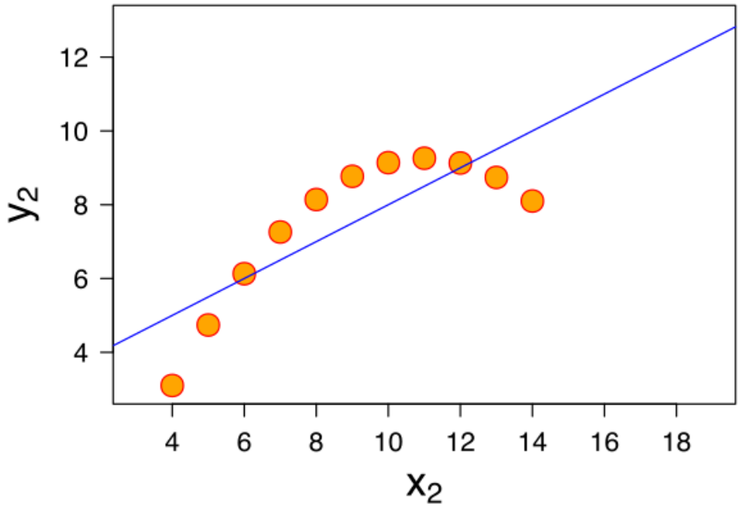

12
10
8
6
4
y_1
4
6
8
10
12
14
16
18
x_1

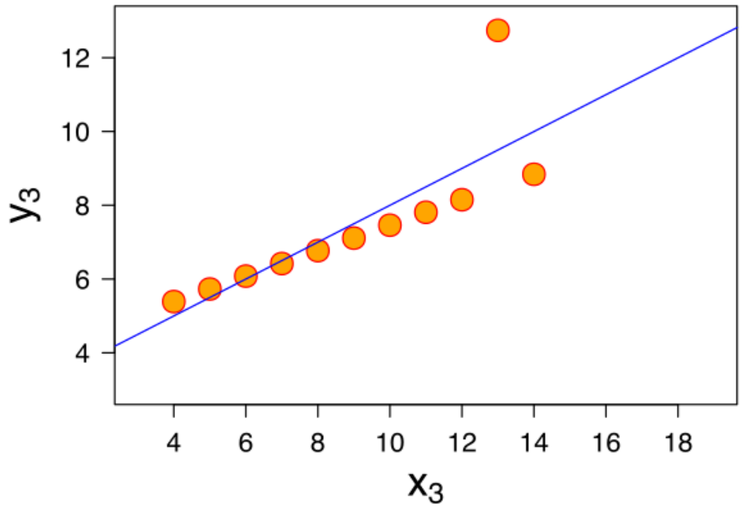

12
10
8
6
4
y_2
4
6
8
10
12
14
16
18
x_2

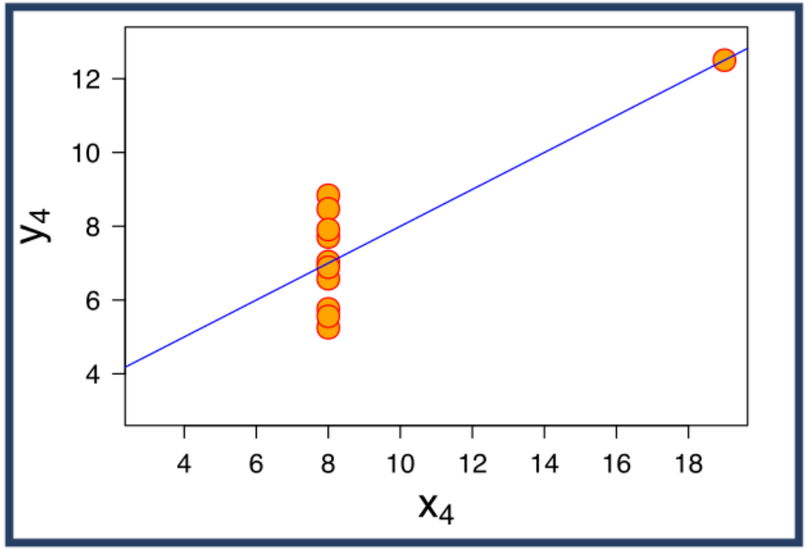

12
10
8
6
4
y_3
4
6
8
10
12
14
16
18
x_3


12
10
8
6
4
y_4
4
6
8
10
12
14
16
18
x_4


12
10
8
6
4
y_1
4 6 8 10 12 14 16 18
x_1


12
10
8
6
4
y_2
4 6 8 10 12 14 16 18
x_2


12
10
8
6
4
y_3
4 6 8 10 12 14 16 18
x_3


12
10
8
6
4
y_4
4 6 8 10 12 14 16 18
x_4

Knowing your data

- Flight Number (number)
- Date (datetime)
- Time (UTC) (datetime)
- Booster Version (text)
- Launch Site (text)
- Payload (text)
- Payload Mass (kg) (number)
- Orbit (text)
- Customer (text)
- Mission Outcome (text)
- Landing Outcome (text)

Previewing your data

Flight	Date	Time (UTC)	Booster Version	Launch Site	Payload
1	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit
2	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats...
3	2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2+
4	2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1
5	2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2

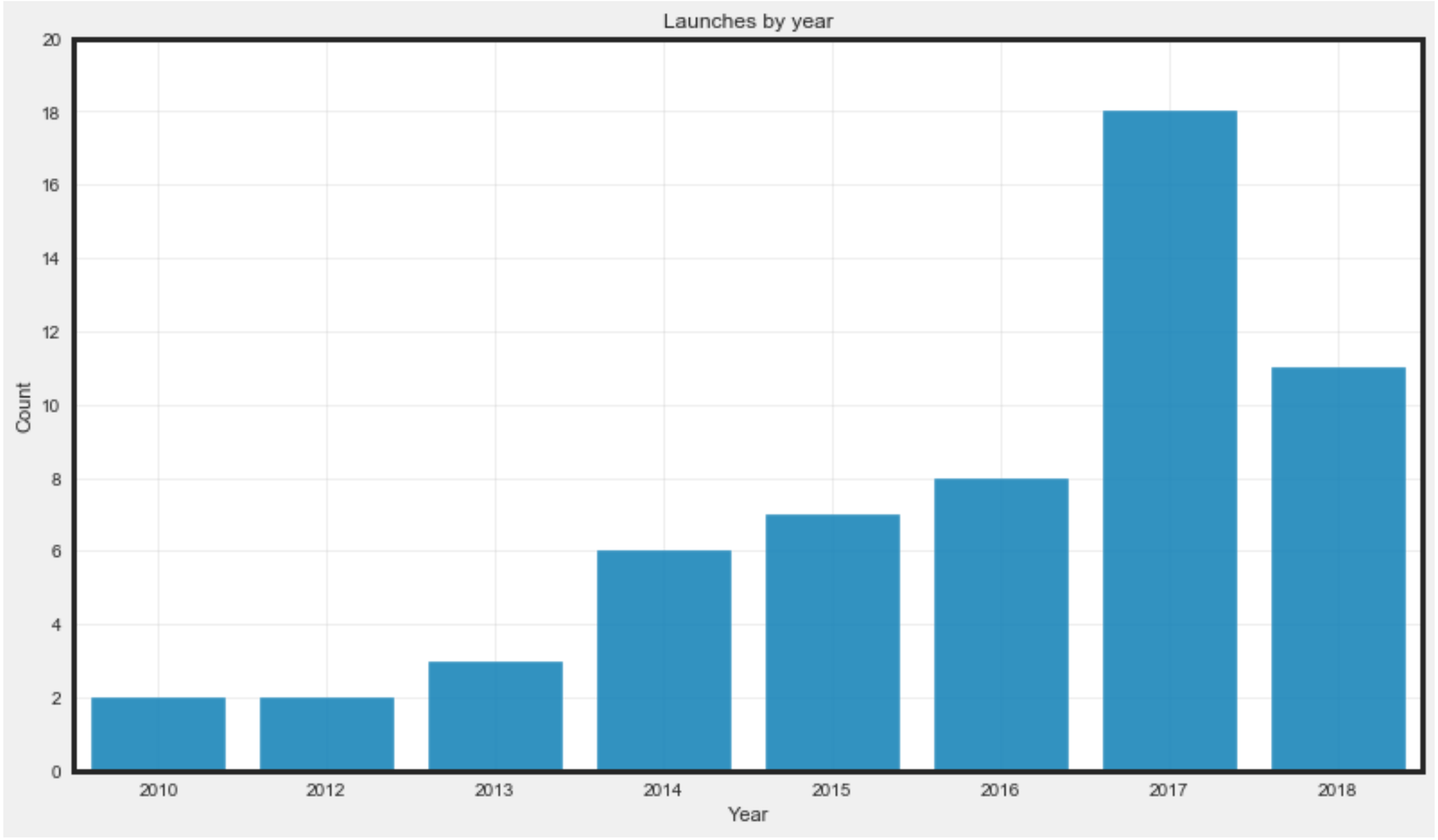
Payload Mass (kg)	Orbit	Customer	Mission Outcome	Landing Outcome
NaN	LEO	SpaceX	Success	Failure (parachute)
NaN	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
525	LEO (ISS)	NASA (COTS)	Success	No attempt
500	LEO (ISS)	NASA (CRS)	Success	No attempt
677	LEO (ISS)	NASA (CRS)	Success	No attempt

Descriptive statistics

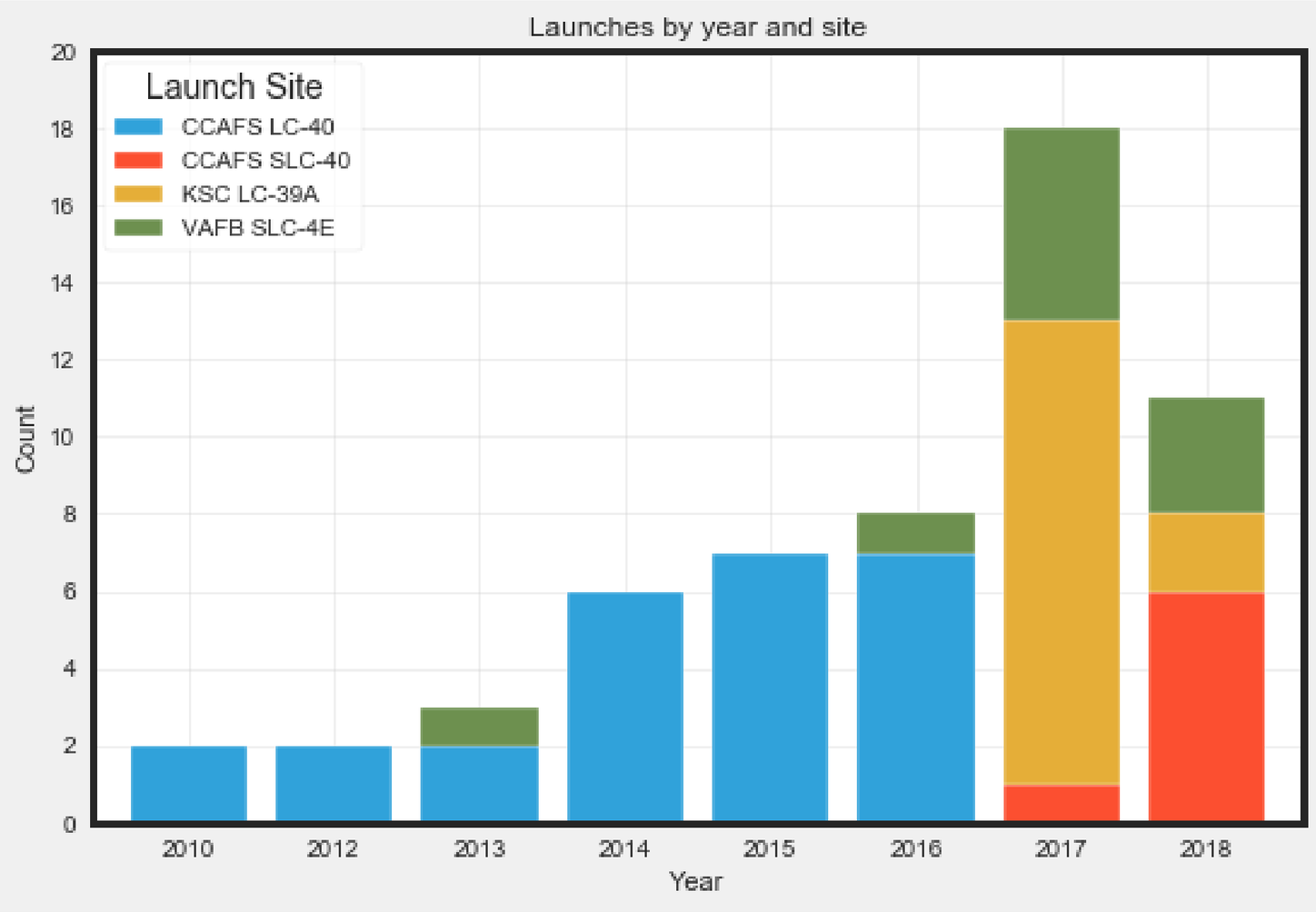
	Flight	Date	Time (UTC)	Booster Version	Launch Site	Payload
count	55	55	55	55	55	55
unique	55	55	53	51	4	55
top	6	2018-03-30	4:45:00	F9 v1.1	CCAFS LC-40	SES-9
freq	1	1	2	5	26	1

	Payload Mass (kg)	Orbit	Customer	Mission Outcome	Landing Outcome
count	53	55	55	55	55
unique	47	8	28	2	12
top	9,600	GTO	NASA (CRS)	Success	No attempt
freq	5	22	14	54	18

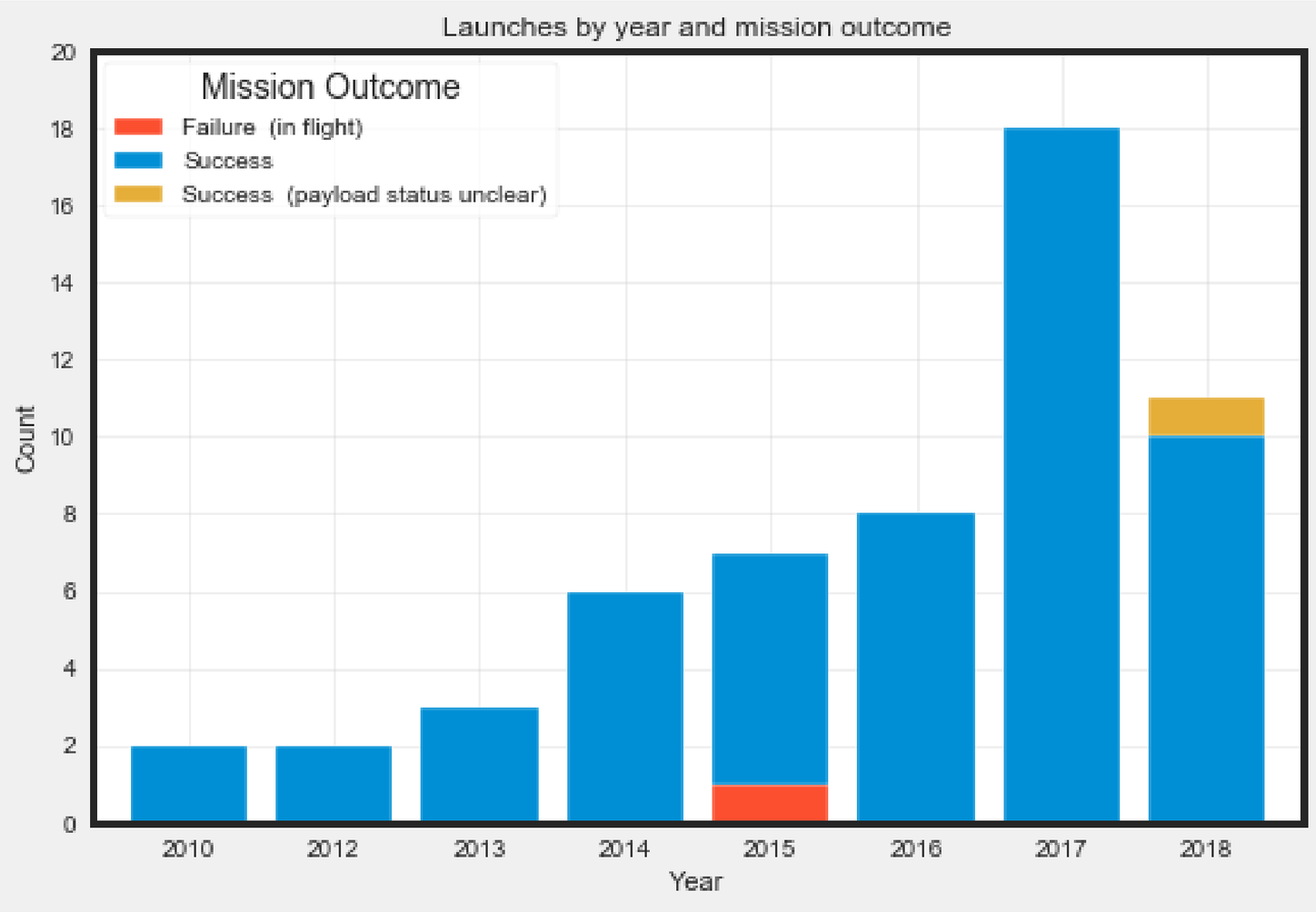
Visualize!


Launches by year
Count
Year
0
2
4
6
8
10
12
14
16
18
20
2010
2012
2013
2014
2015
2016
2017
2018

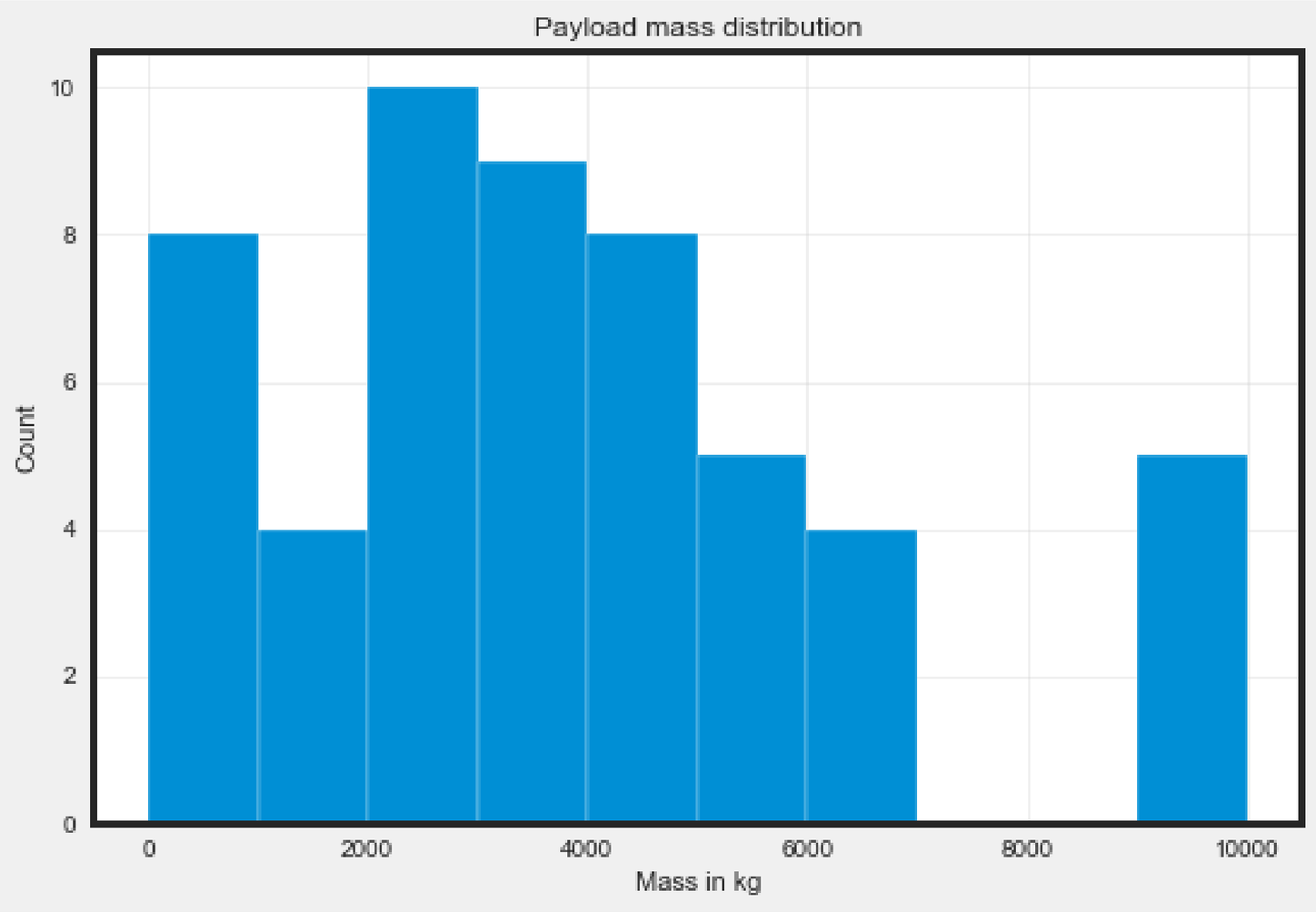
Ask more questions!



Ask more questions!



Outliers


Payload mass distribution
Count
0
2
4
6
8
10
0
2000
4000
6000
8000
10000
Mass in kg

Let's practice!

UNDERSTANDING DATA SCIENCE

Interactive dashboards

UNDERSTANDING DATA SCIENCE

Hadrien Lacroix
Content Developer at DataCamp

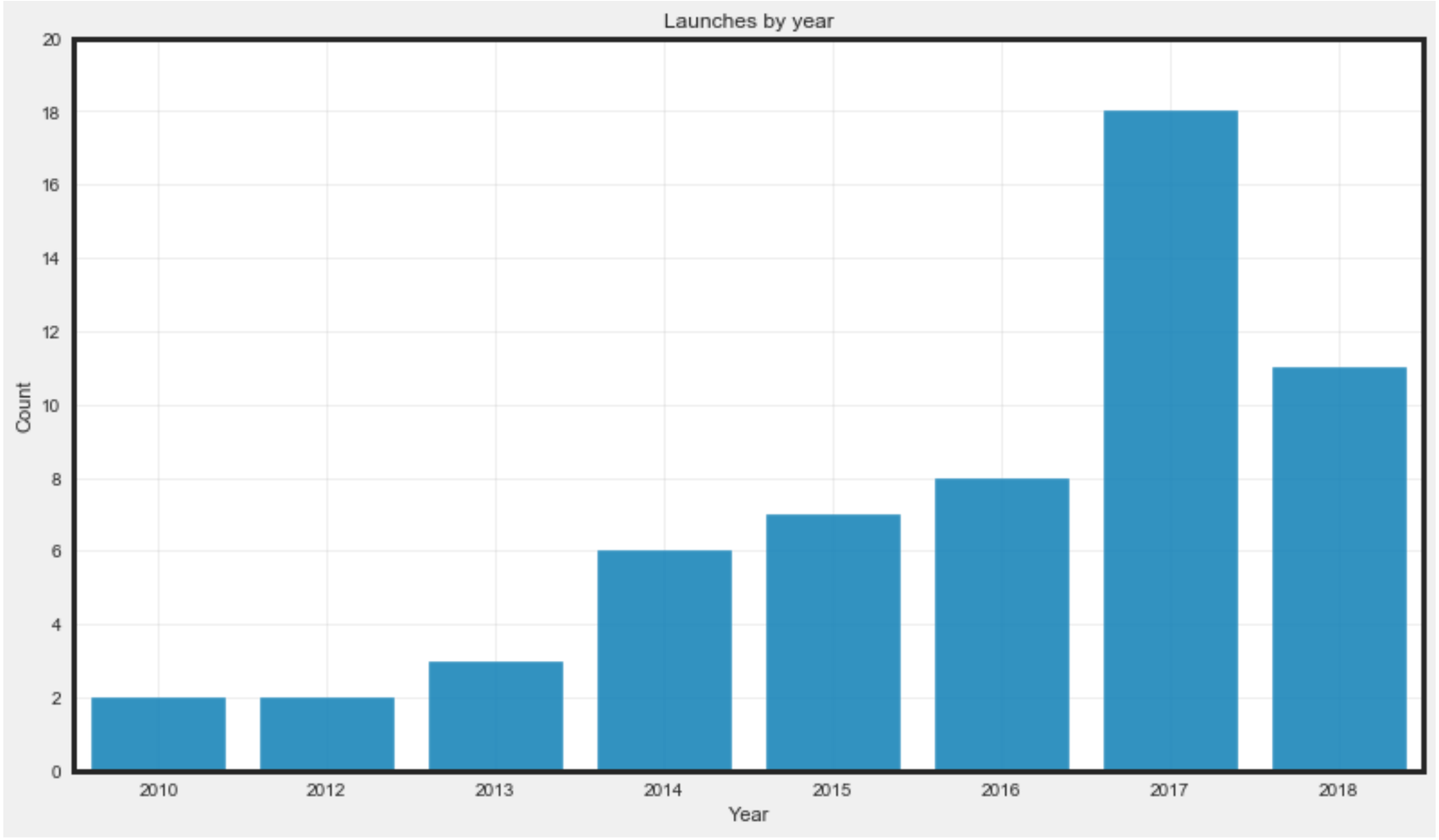
datacamp

One picture...

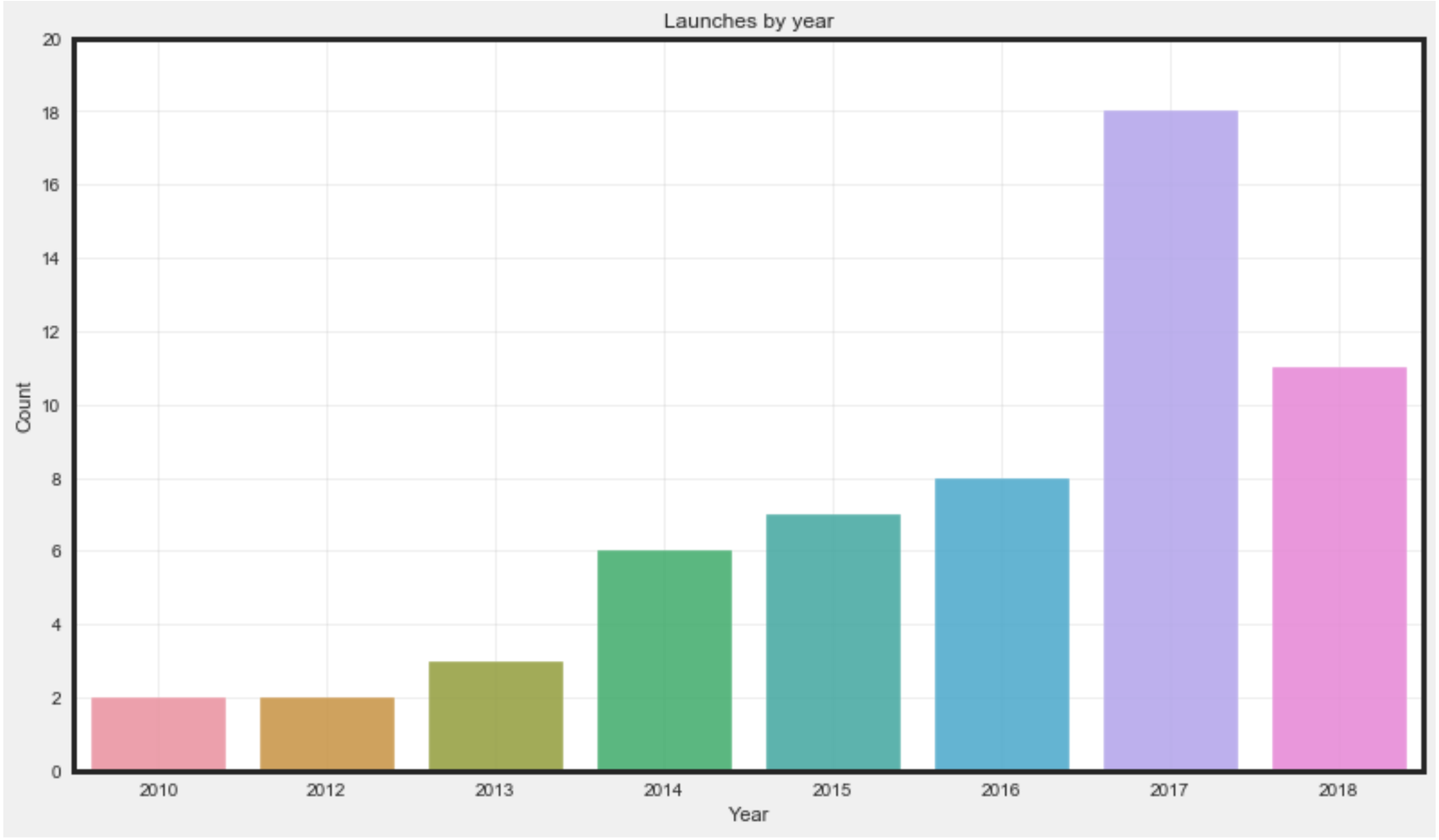
> One picture is worth a thousand words...

> ...if the picture makes sense.

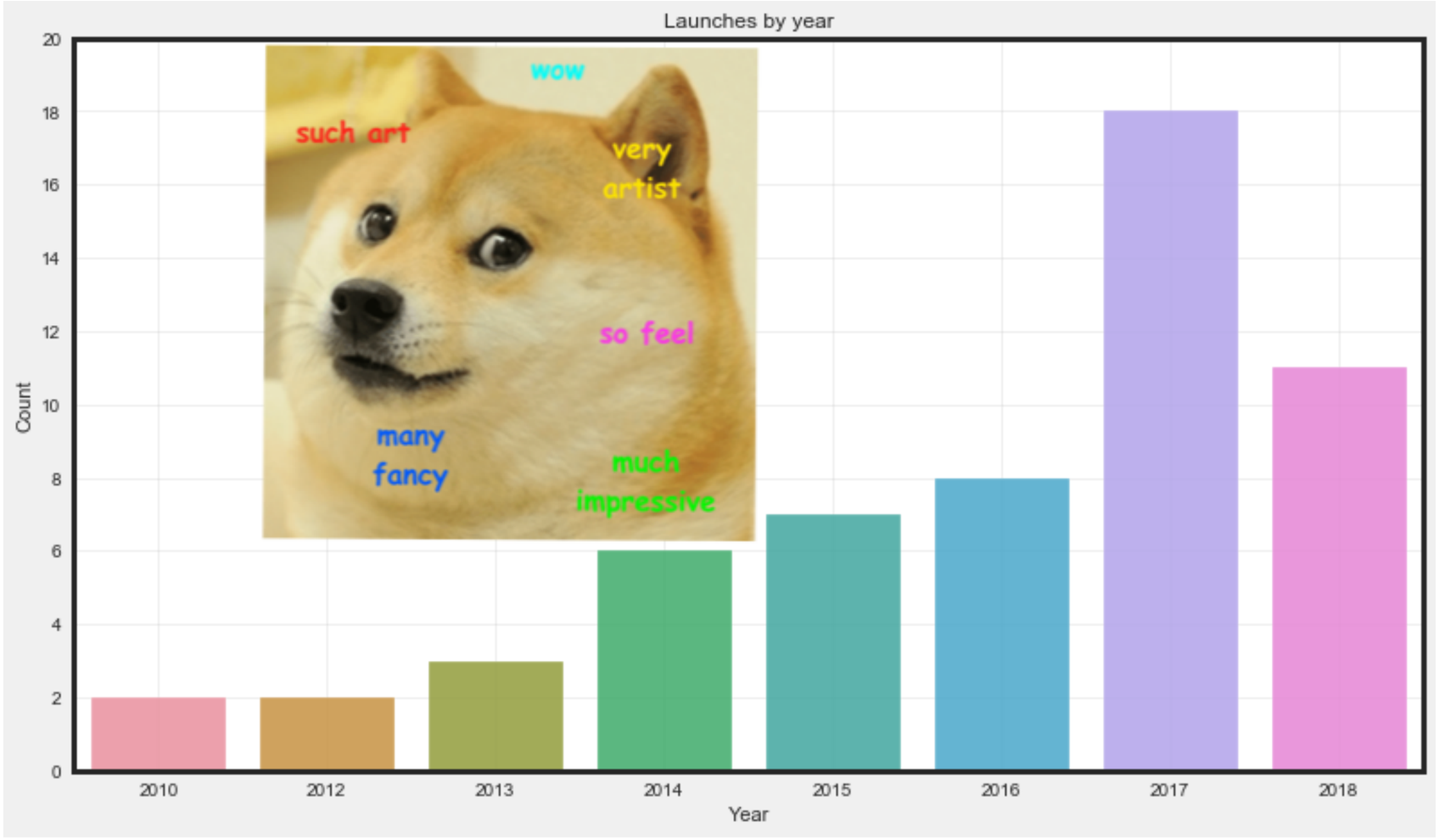
Use color purposefully



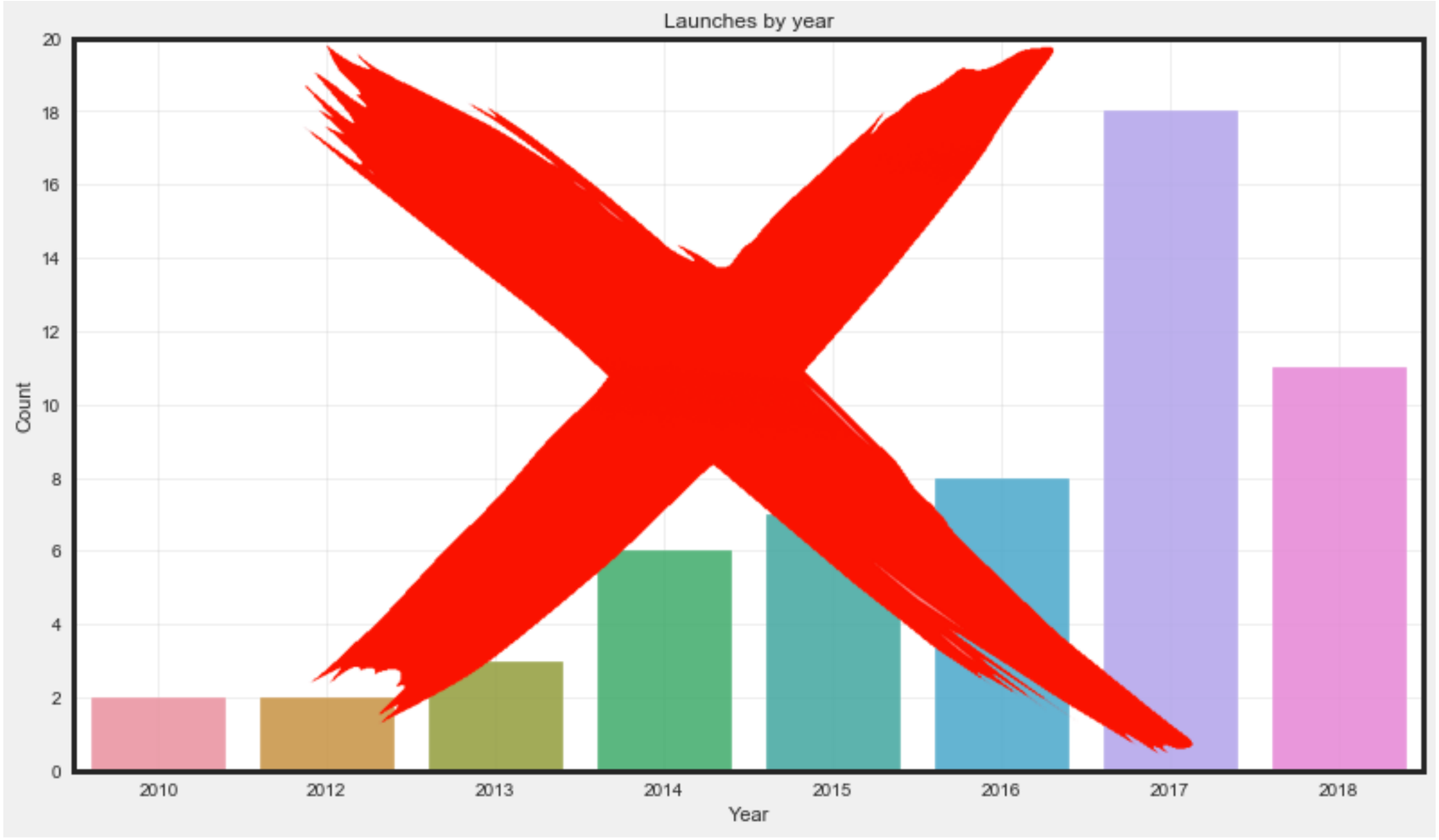
Use color purposefully



Use color purposefully



Use color purposefully



Colorblindness

- Red and green is the most common (but not the only one)
- Information and simulators online
- Existing color palettes accessible to everyone

Readable fonts

- sans-serif

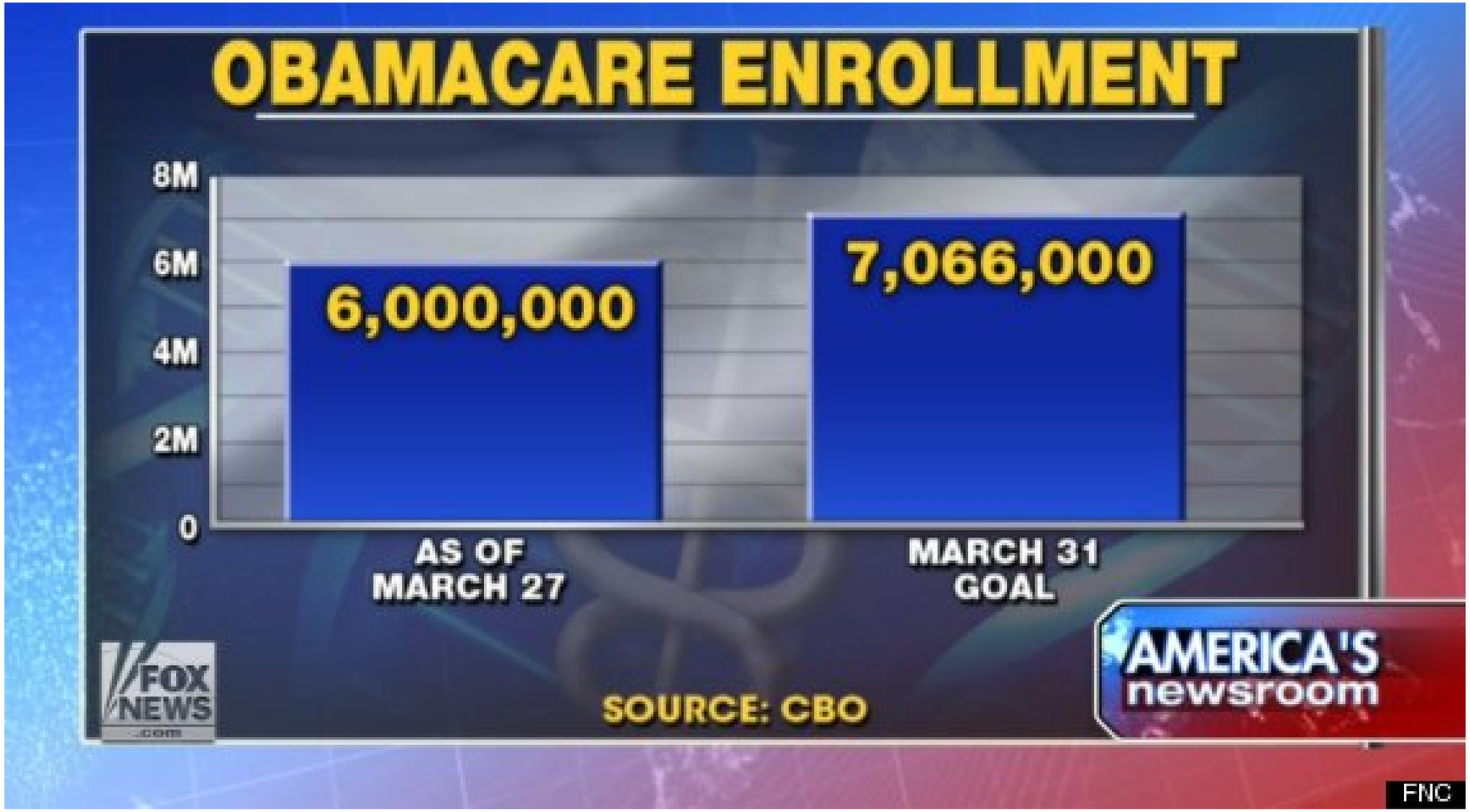
Label, label, label

- title
- x axis label
- y axis label
- legend

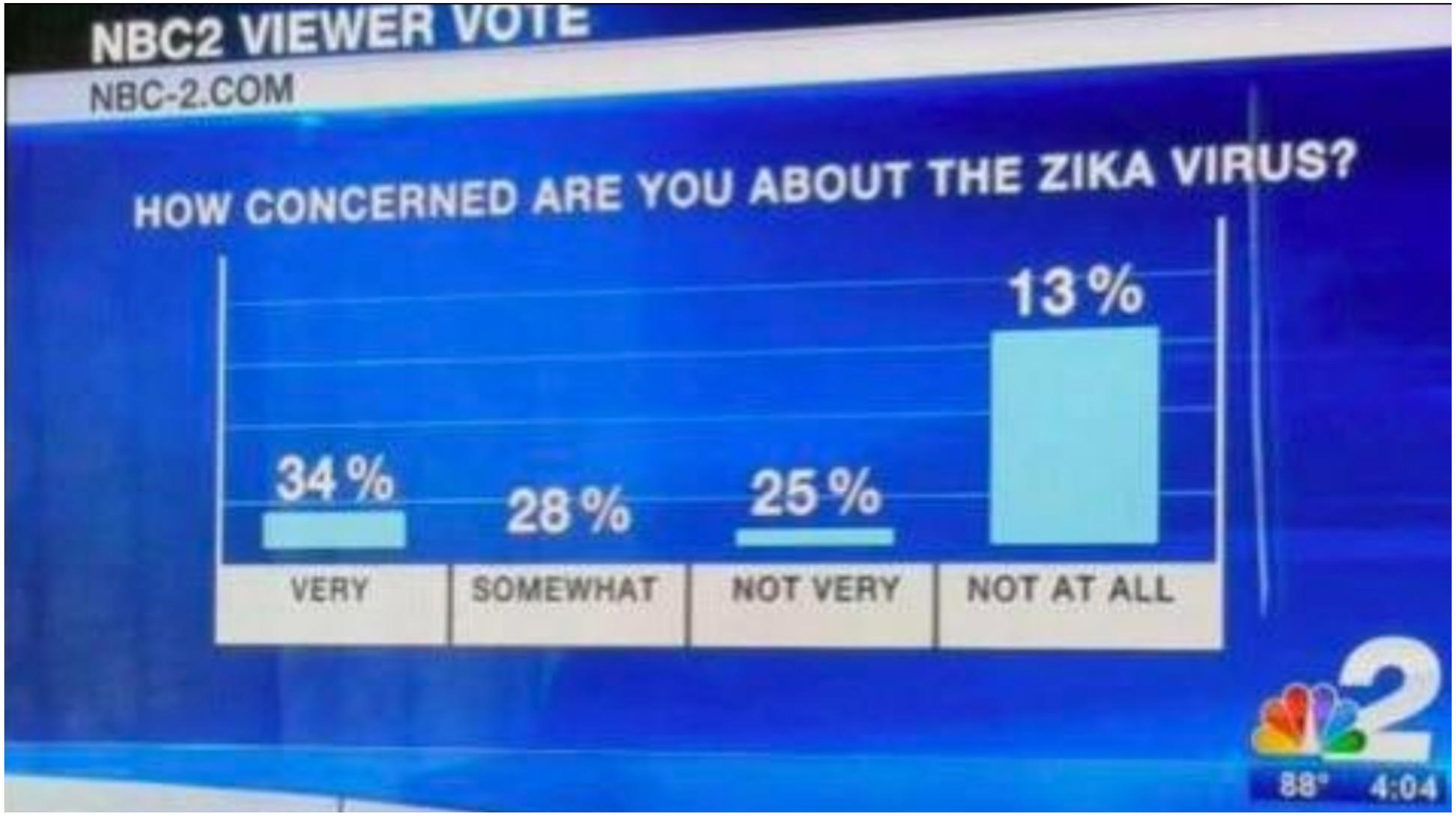
Axes


OBAMACARE ENROLLMENT
7,066,000
6,000,000
AS OF MARCH 27
MARCH 31 GOAL
SOURCE: HHS
FOX NEWS channel
mediamatters.org
E IN SECOND PLACE WITH $26.5 MIL, WHILE "MUPP
DOW FUT 16,325.00
Fox News

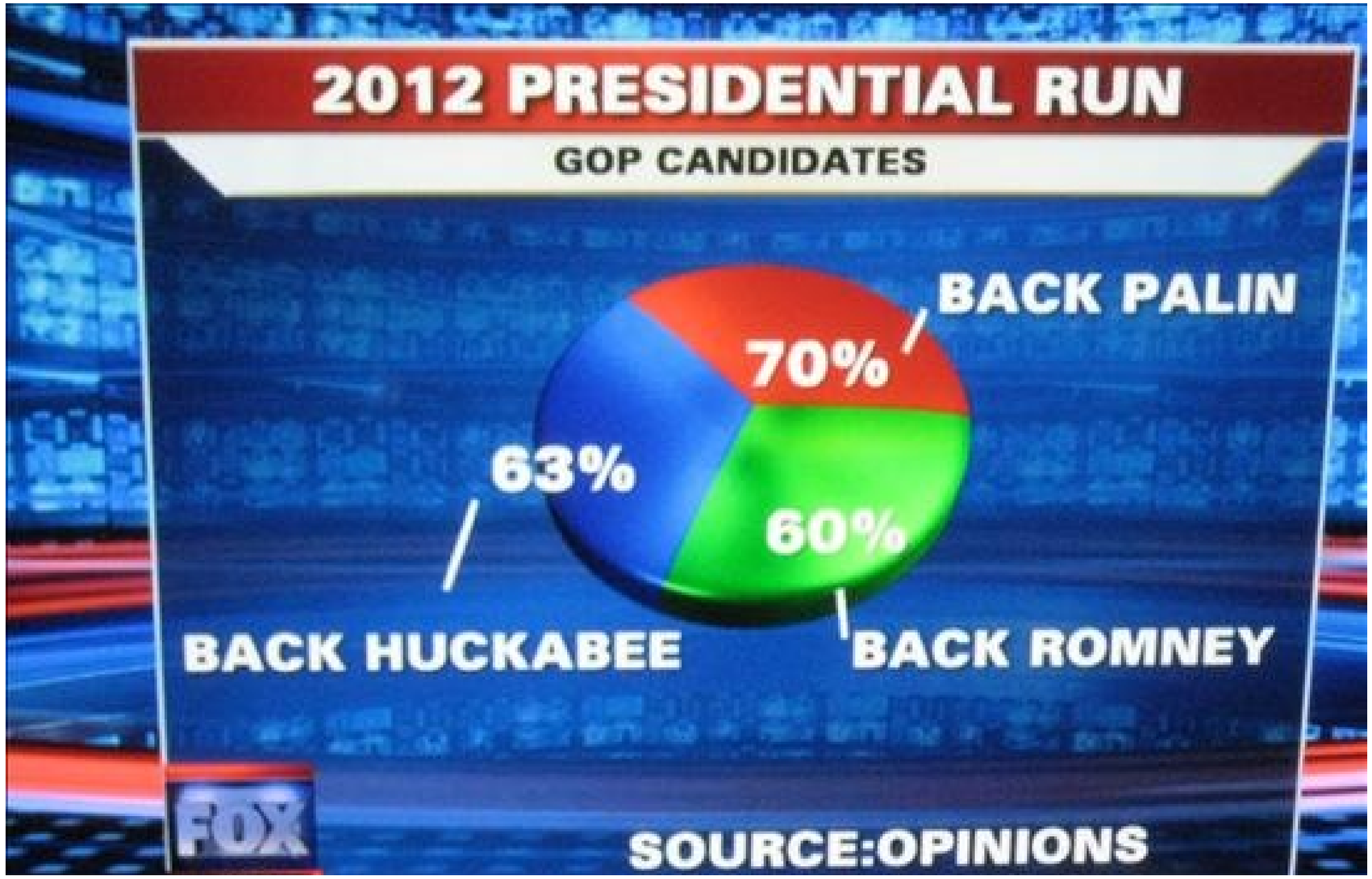
Axes


OBAMACARE ENROLLMENT
8M
6M
4M
2M
0
6,000,000
7,066,000
AS OF MARCH 27
MARCH 31 GOAL
FOX NEWS .com
SOURCE: CBO
AMERICA'S newsroom
FNC

And the award goes to...



Honorable mention



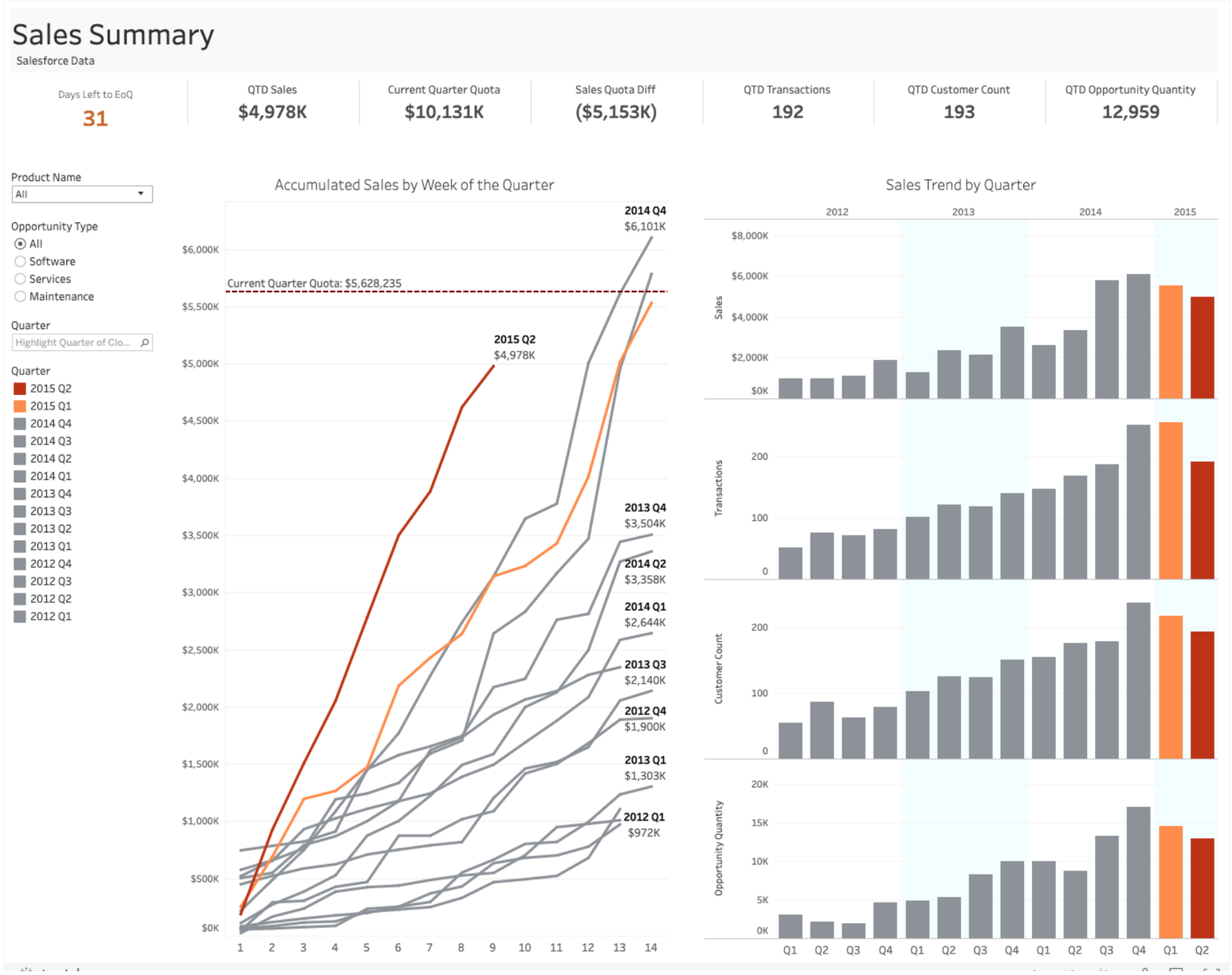
Question

$$1\ picture = 1000\ words$$

$$1000\ pictures = ?$$

A dashboard!

[1] Photo by Marek Szturc on Unsplash


Sales Summary
Salesforce Data
Days Left to EoQ
31
QTD Sales
$4,978K
Current Quarter Quota
$10,131K
Sales Quota Diff
($5,153K)
QTD Transactions
192
QTD Customer Count
193
QTD Opportunity Quantity
12,959
Product Name
All
Opportunity Type
All
Software
Services
Maintenance
Quarter
Highlight Quarter of Clo...
Quarter
2015 Q2
2015 Q1
2014 Q4
2014 Q3
2014 Q2
2014 Q1
2013 Q4
2013 Q3
2013 Q2
2013 Q1
2012 Q4
2012 Q3
2012 Q2
2012 Q1
Accumulated Sales by Week of the Quarter
Current Quarter Quota: $5,628,235
2014 Q4
$6,101K
2015 Q2
$4,978K
2013 Q4
$3,504K
2014 Q2
$3,358K
2014 Q1
$2,644K
2013 Q3
$2,140K
2012 Q4
$1,900K
2013 Q1
$1,303K
2012 Q1
$972K
$6,000K
$5,500K
$5,000K
$4,500K
$4,000K
$3,500K
$3,000K
$2,500K
$2,000K
$1,500K
$1,000K
$500K
$0K
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Sales Trend by Quarter
2012
2013
2014
2015
Sales
$8,000K
$6,000K
$4,000K
$2,000K
$0K
Transactions
200
100
0
Customer Count
200
100
0
Opportunity Quantity
20K
15K
10K
5K
0K
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2

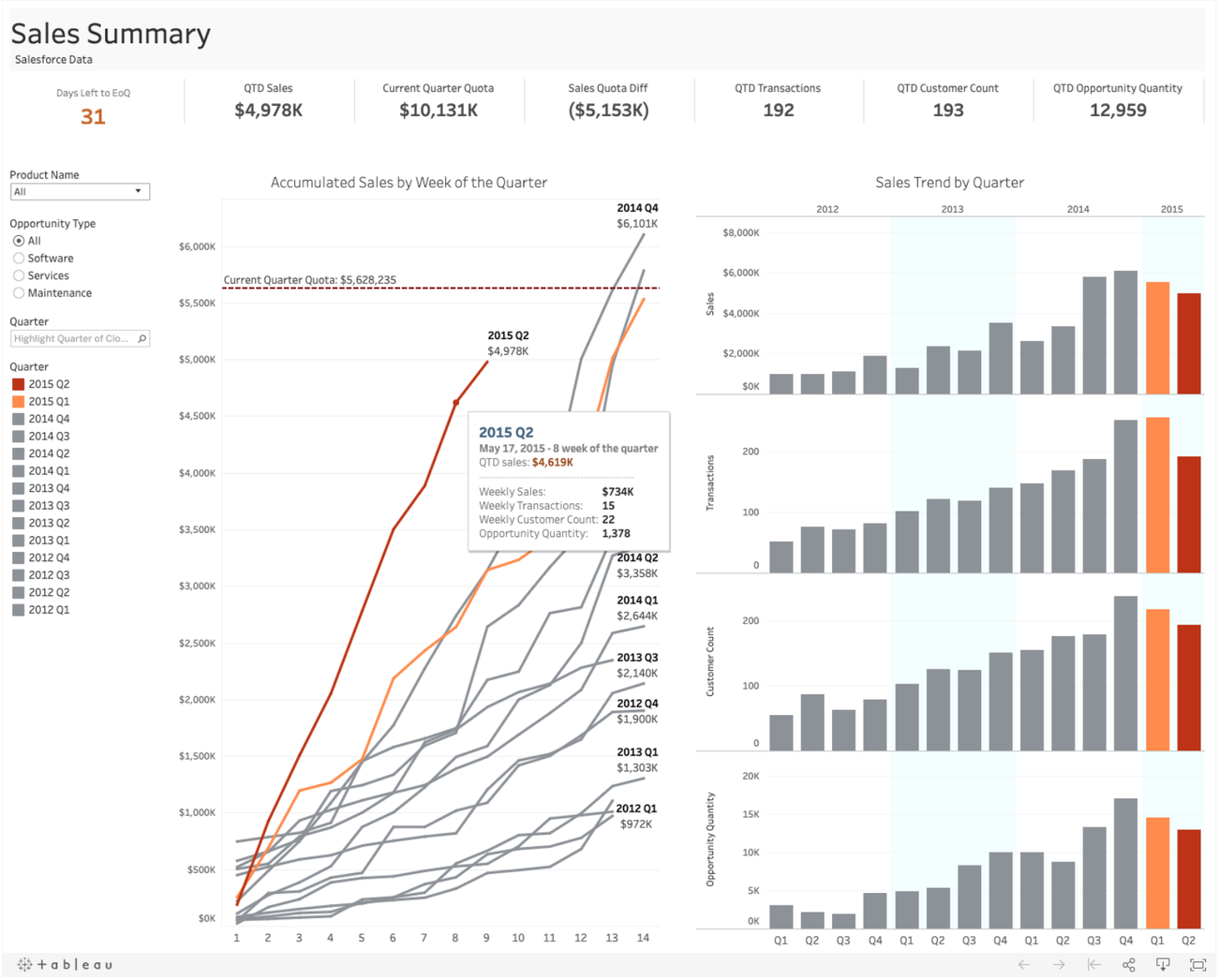
BI tools

tableau

looker

Power BI

Next level


Sales Summary
Salesforce Data
Days Left to EoQ
31
QTD Sales
$4,978K
Current Quarter Quota
$10,131K
Sales Quota Diff
($5,153K)
QTD Transactions
192
QTD Customer Count
193
QTD Opportunity Quantity
12,959
Product Name
All
Opportunity Type
All
Software
Services
Maintenance
Quarter
Highlight Quarter of Clo...
Quarter
2015 Q2
2015 Q1
2014 Q4
2014 Q3
2014 Q2
2014 Q1
2013 Q4
2013 Q3
2013 Q2
2013 Q1
2012 Q4
2012 Q3
2012 Q2
2012 Q1
Accumulated Sales by Week of the Quarter
Current Quarter Quota: $5,628,235
2014 Q4
$6,101K
2015 Q2
$4,978K
2015 Q2
May 17, 2015 - 8 week of the quarter
QTD sales: $4,619K
Weekly Sales: $734K
Weekly Transactions: 15
Weekly Customer Count: 22
Opportunity Quantity: 1,378
2014 Q2
$3,358K
2014 Q1
$2,644K
2013 Q3
$2,140K
2012 Q4
$1,900K
2013 Q1
$1,303K
2012 Q1
$972K
Sales Trend by Quarter
2012
2013
2014
2015
Sales
Transactions
Customer Count
Opportunity Quantity
tableau

Let's practice!

UNDERSTANDING DATA SCIENCE