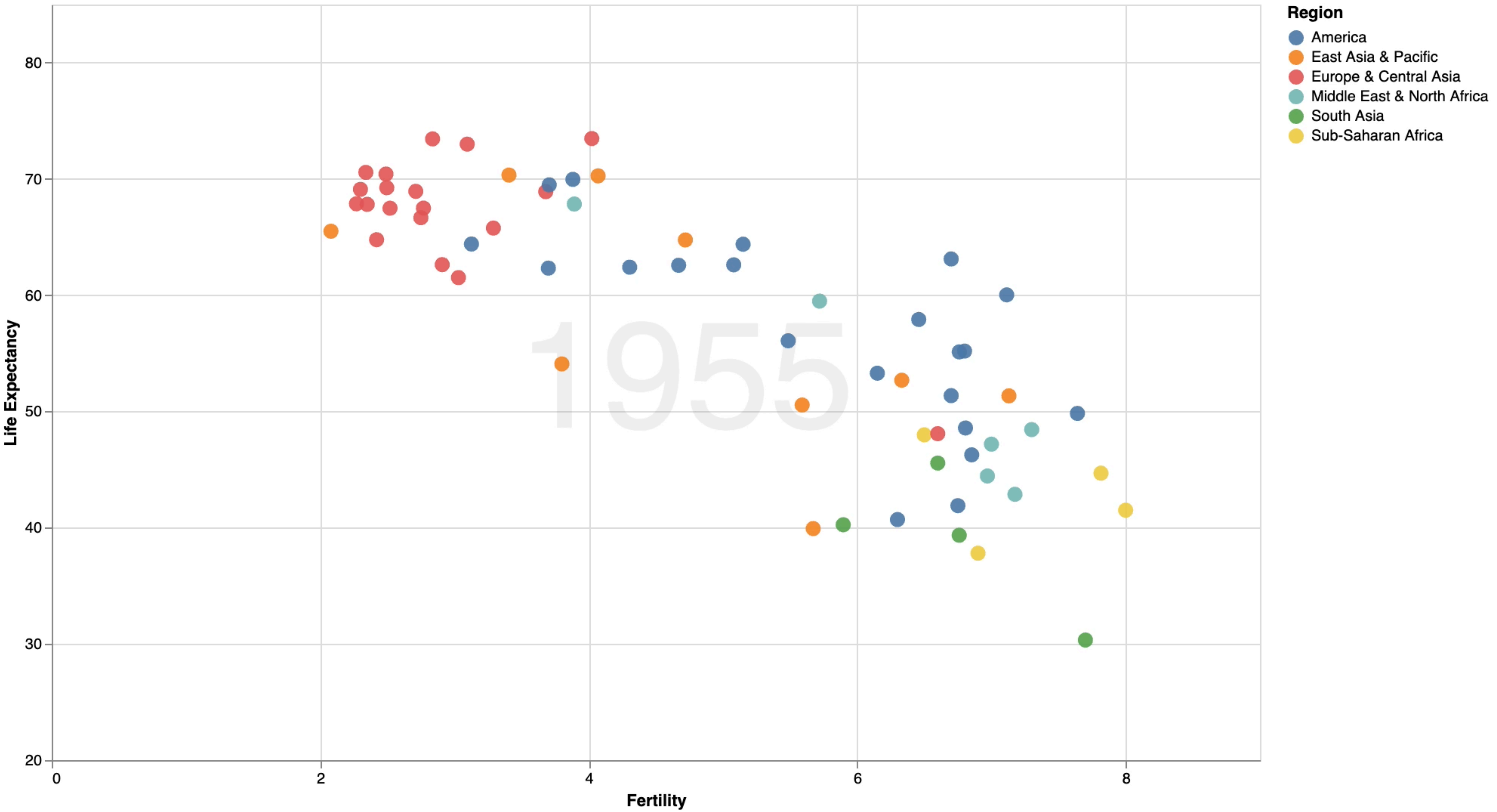
Enhancing Decision-Making through Interactive Data Visualization

Dominik Moritz @domoritz

Carnegie Mellon University

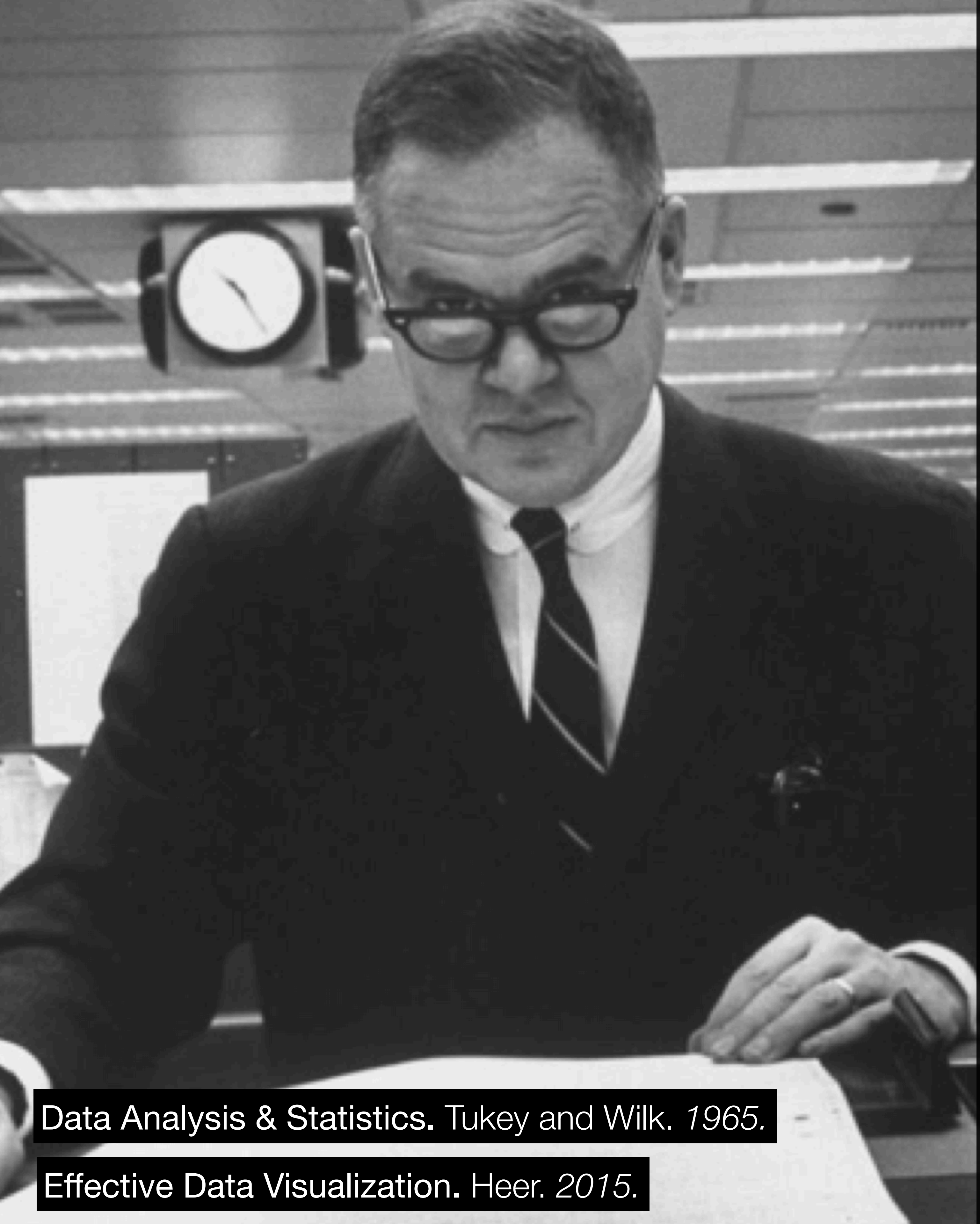
dig.cmu.edu

Based in part on materials by Jeffrey Heer


Region
● America
● East Asia & Pacific
● Europe & Central Asia
● Middle East & North Africa
● South Asia
● Sub-Saharan Africa
1955
Life Expectancy
20
30
40
50
60
70
80
Fertility
0
2
4
6
8

Year 1955

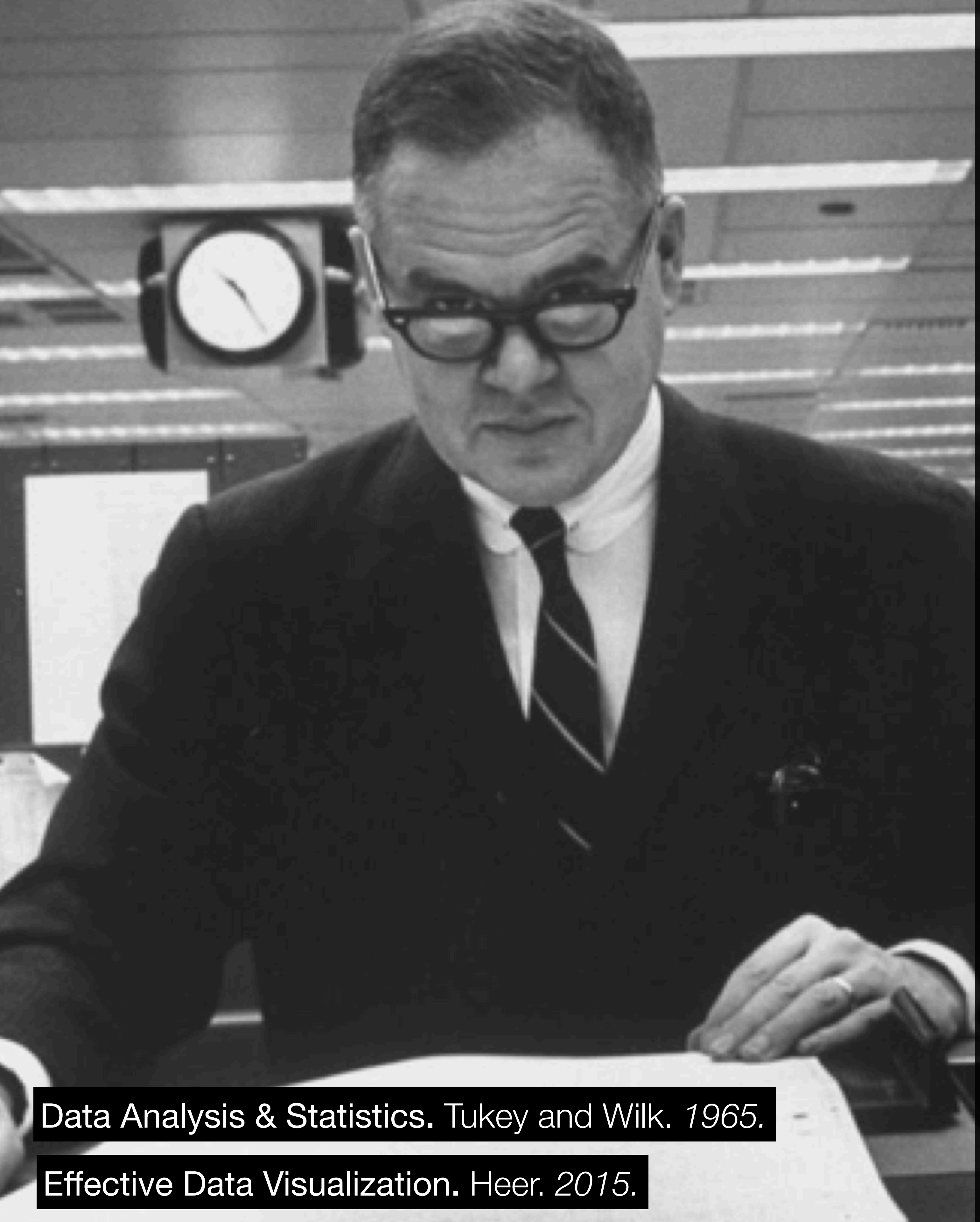
6 takeaways about data visualization



Data Analysis & Statistics. Tukey and Wilk. *1965.*

Effective Data Visualization. Heer. *2015.*

Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the **informality** and **flexibility** appropriate to the **exploratory character of exposure** can be fitted into any of the structures of formal statistics so far proposed.



Nothing - not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers - nothing can substitute here for the **flexibility of the informed human** mind.

Accordingly, both approaches and techniques need to be structured so as to **facilitate human involvement and intervention**.

Data Analysis & Statistics. Tukey and Wilk. *1965.*

Effective Data Visualization. Heer. *2015.*

Set 1

X	Y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

Set 2

X	Y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.11
7	7.26
5	4.74

Set 3

X	Y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

Set 4

X	Y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89

Summary Statistics

$u_X = 9.0$ $\sigma_X = 3.317$

$u_Y = 7.5$ $\sigma_Y = 2.03$

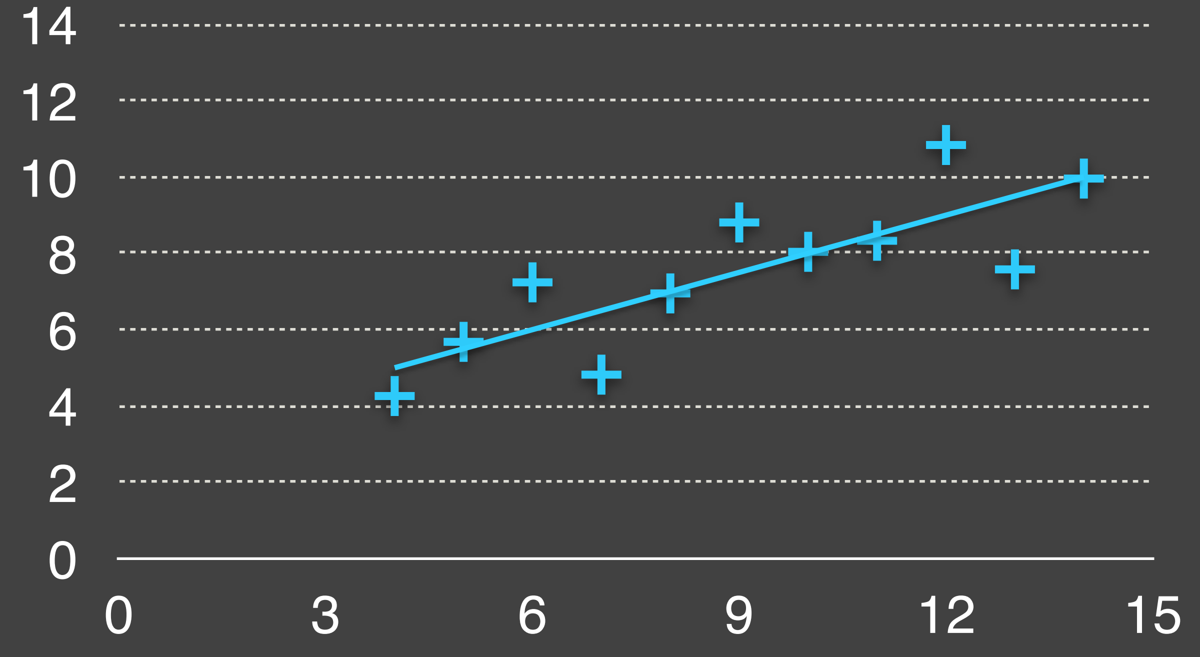
Linear Regression

$Y^2 = 3 + 0.5\ X$

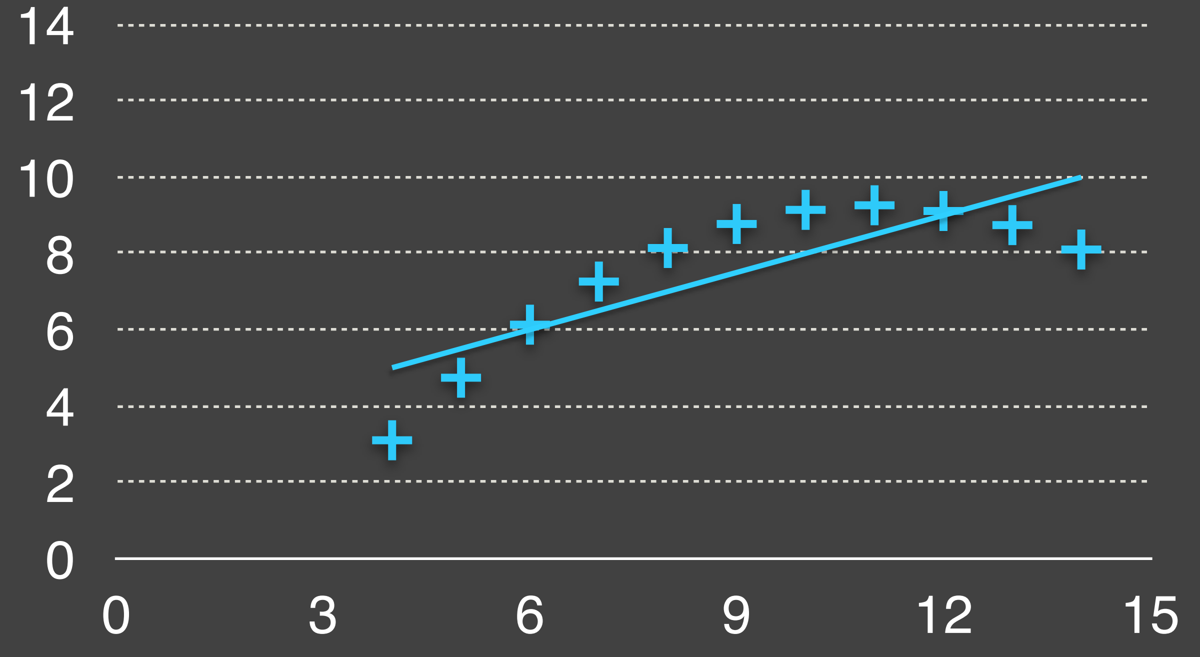
$R^2 = 0.67$

[Anscombe 1973]

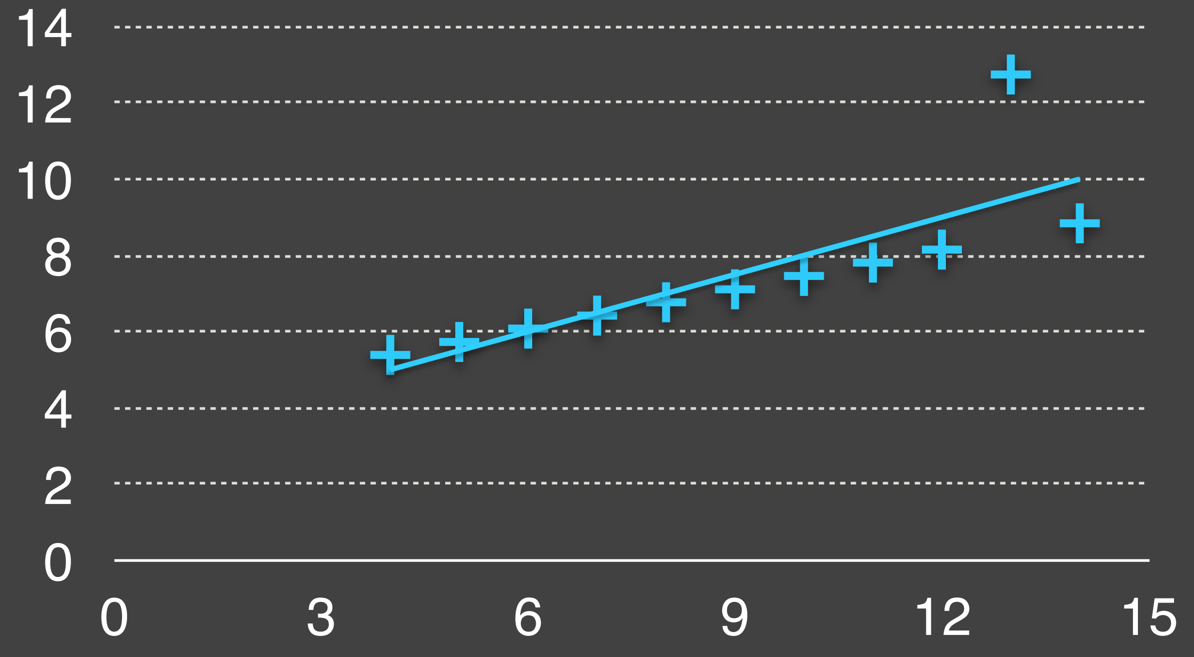
Set 1



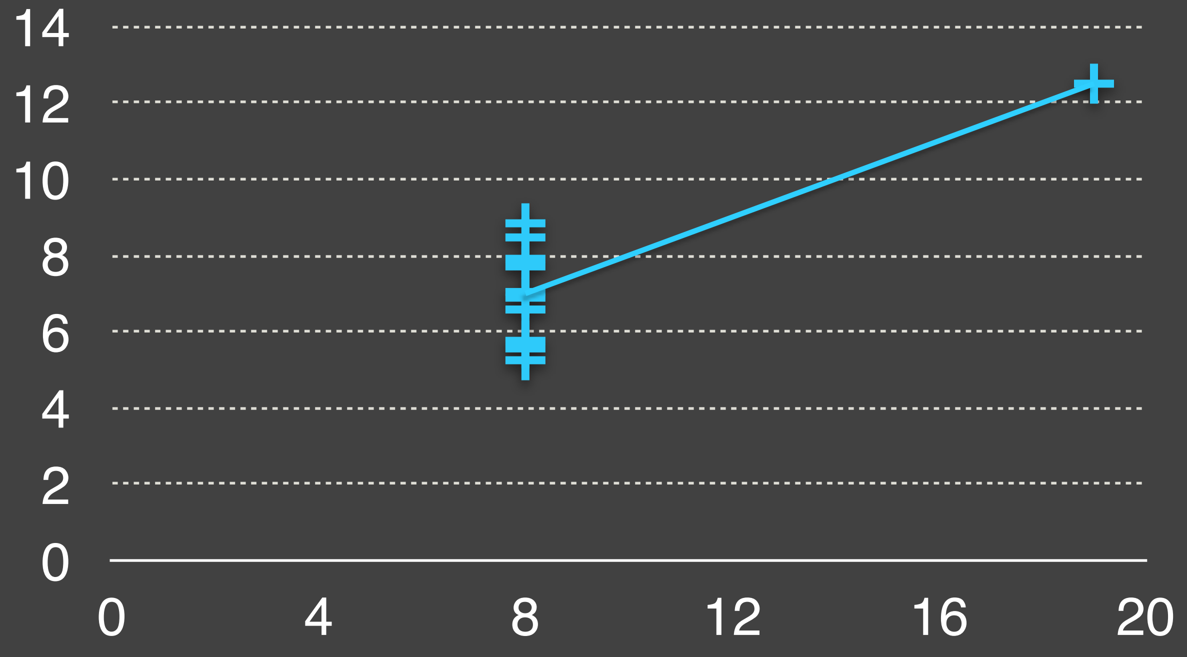
Set 2



Set 3

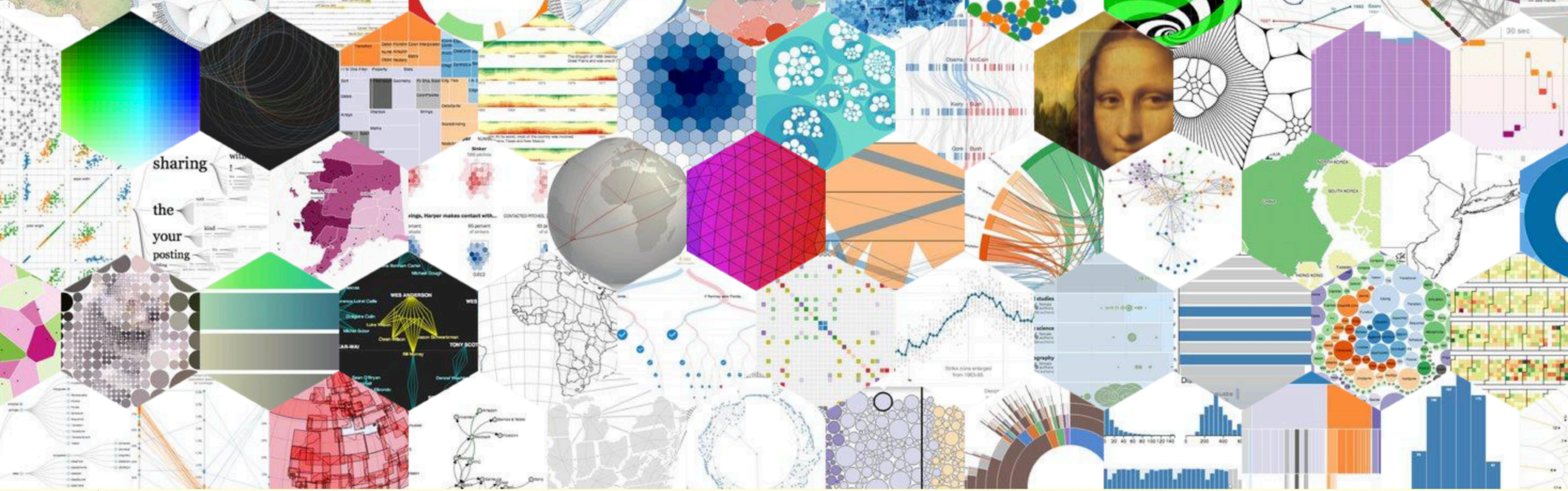


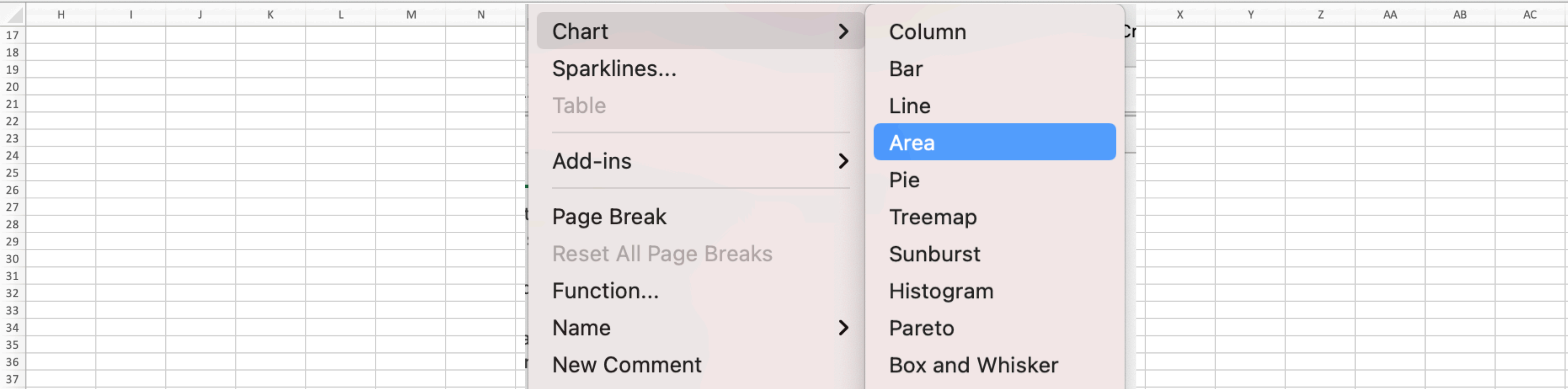
Set 4

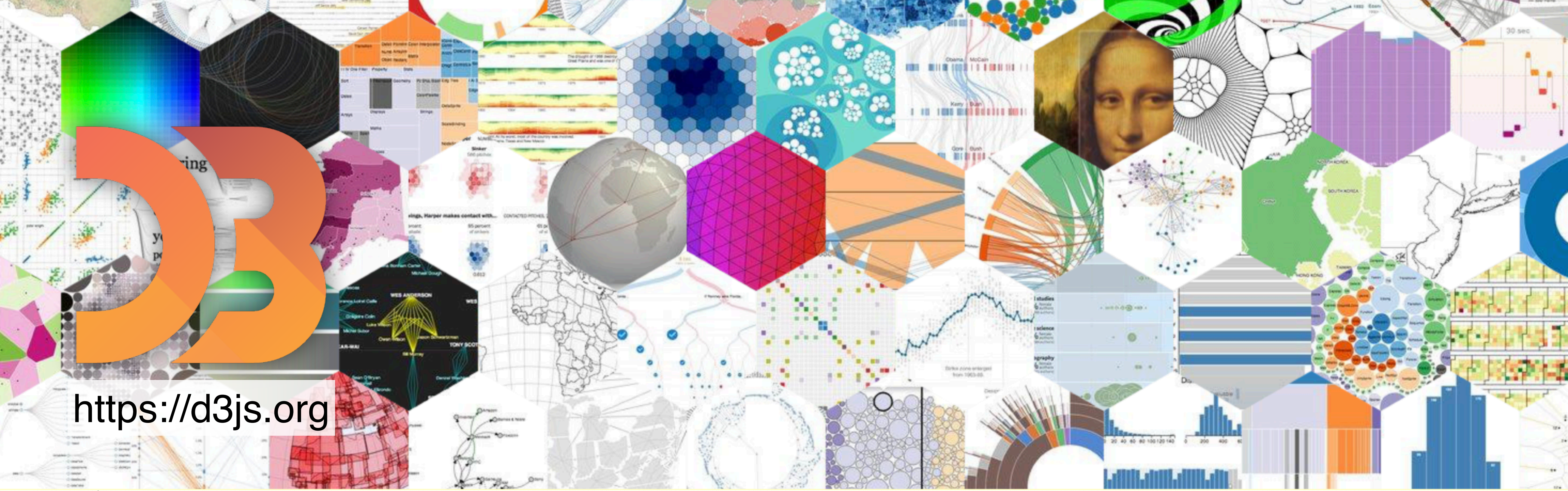


Takeaway:

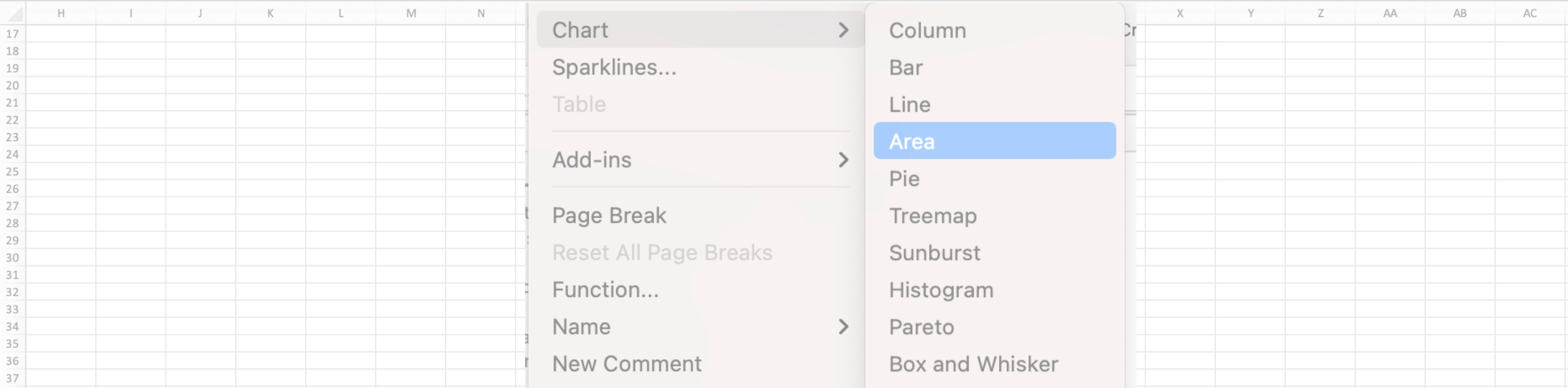
Machine Learning, AI, and Statistics are people problems. For them to be effective, we need to design for human involvement. 🤖 👩‍💻


sharing
the
your
posting
WES ANDERSON

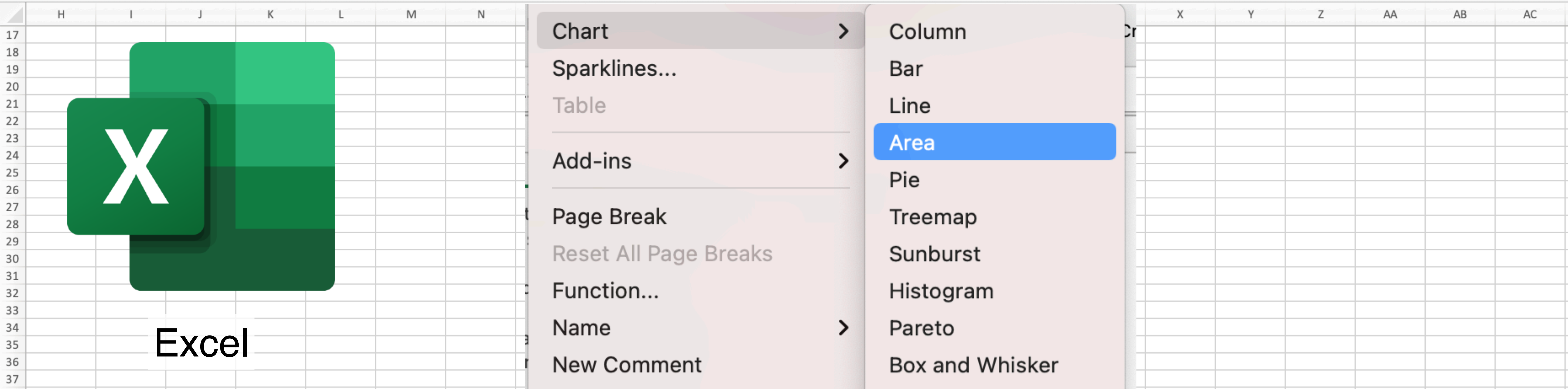

H
I
J
K
L
M
N
X
Y
Z
AA
AB
AC
Chart
Sparklines...
Table
Add-ins
Page Break
Reset All Page Breaks
Function...
Name
New Comment
Column
Bar
Line
Area
Pie
Treemap
Sunburst
Histogram
Pareto
Box and Whisker



https://d3js.org

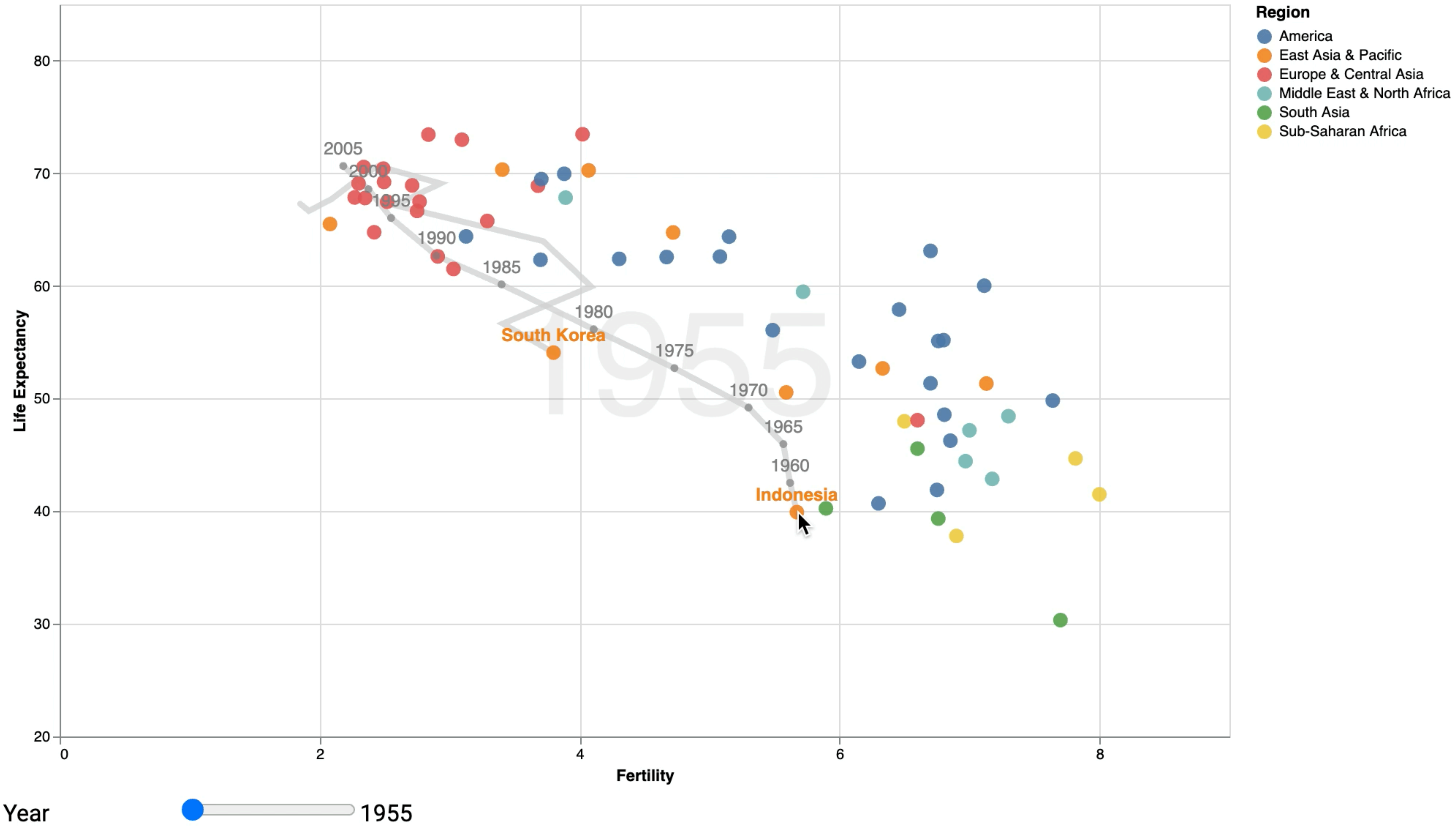


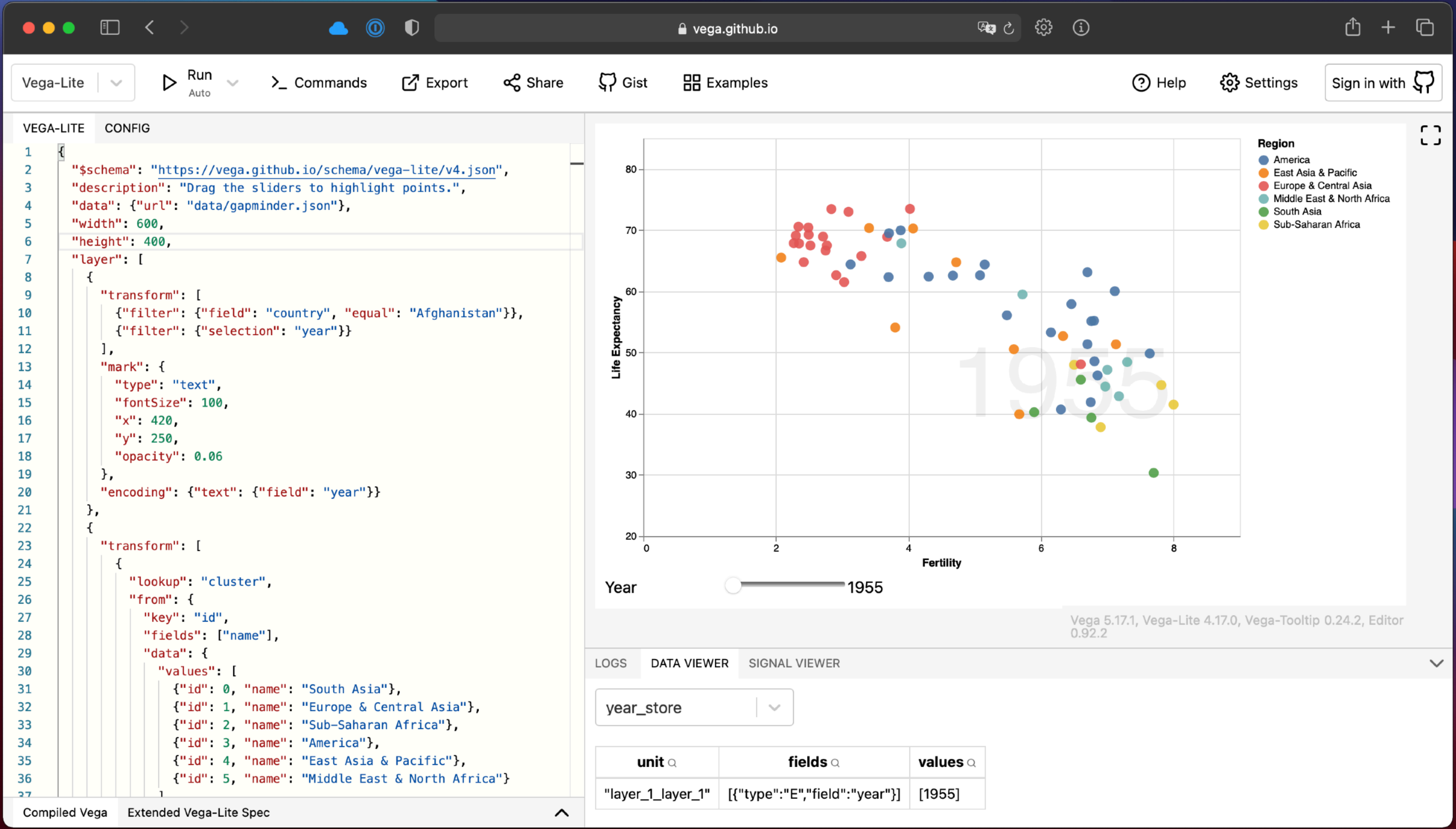
https://d3js.org



Excel

How do we make visualizations in the midst of an analysis?


Region
● America
● East Asia & Pacific
● Europe & Central Asia
● Middle East & North Africa
● South Asia
● Sub-Saharan Africa
Life Expectancy
80
70
60
50
40
30
20
Fertility
0
2
4
6
8
1955
2005
2000
1995
1990
1985
1980
1975
1970
1965
1960
South Korea
Indonesia
Year
1955



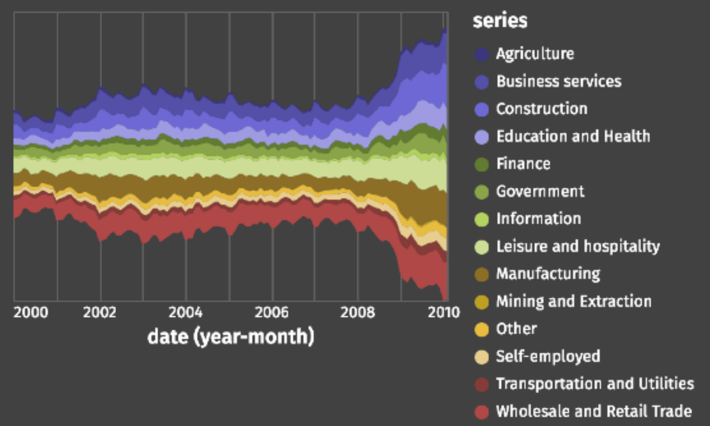
```
{
  "$schema": "https://vega.github.io/schema/vega-lite/v4.json",
  "description": "Drag the sliders to highlight points.",
  "data": {"url": "data/gapminder.json"},
  "width": 600,
  "height": 400,
  "layer": [
    {
      "transform": [
        {"filter": {"field": "country", "equal": "Afghanistan"}},
        {"filter": {"selection": "year"}}
      ],
      "mark": {
        "type": "text",
        "fontSize": 100,
        "x": 420,
        "y": 250,
        "opacity": 0.06
      },
      "encoding": {"text": {"field": "year"}}
    },
    {
      "transform": [
        {
          "lookup": "cluster",
          "from": {
            "key": "id",
            "fields": ["name"],
            "data": {
              "values": [
                {"id": 0, "name": "South Asia"},
                {"id": 1, "name": "Europe & Central Asia"},
                {"id": 2, "name": "Sub-Saharan Africa"},
                {"id": 3, "name": "America"},
                {"id": 4, "name": "East Asia & Pacific"},
                {"id": 5, "name": "Middle East & North Africa"}
              ]
```

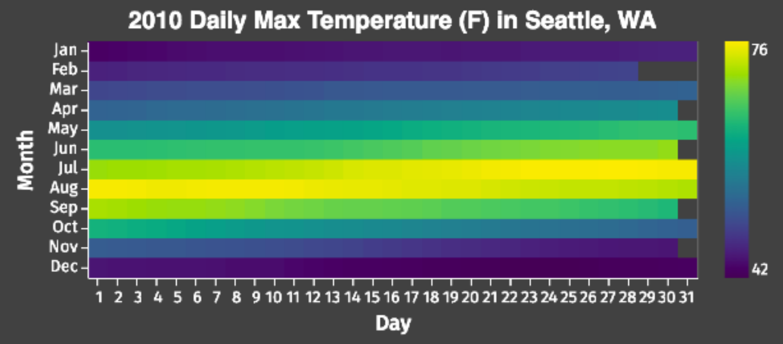
Vega 5.17.1, Vega-Lite 4.17.0, Vega-Tooltip 0.24.2, Editor 0.92.2

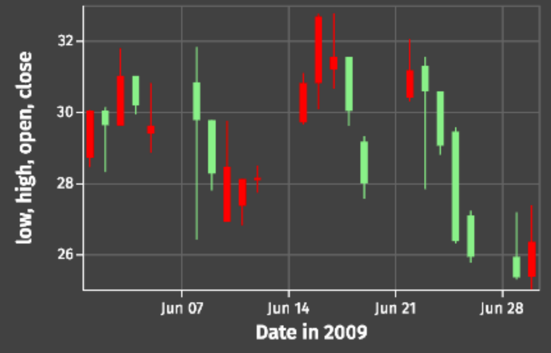
unit	fields	values
"layer_1_layer_1"	[{"type":"E","field":"year"}]	[1955]

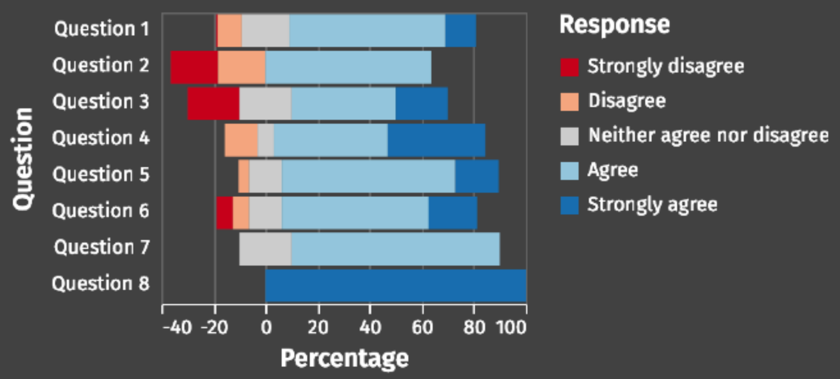
Vega-Lite

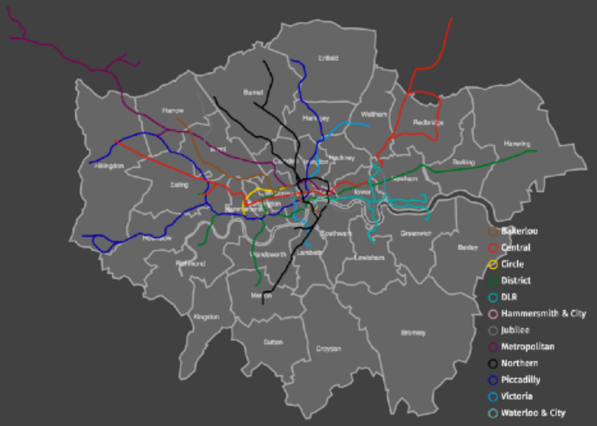
A high-level declarative grammar for interactive multi-view charts.

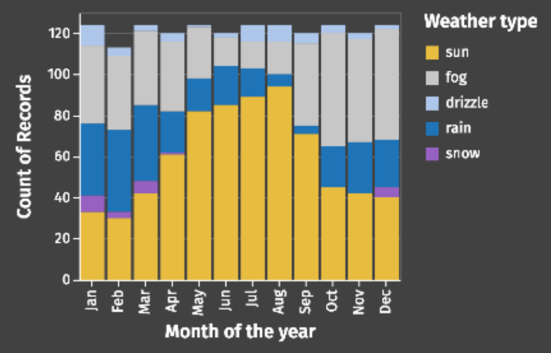


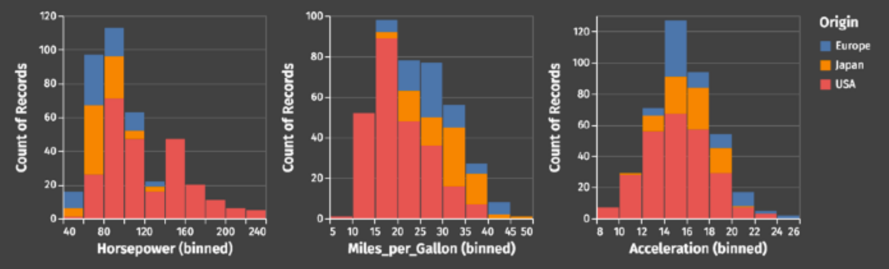






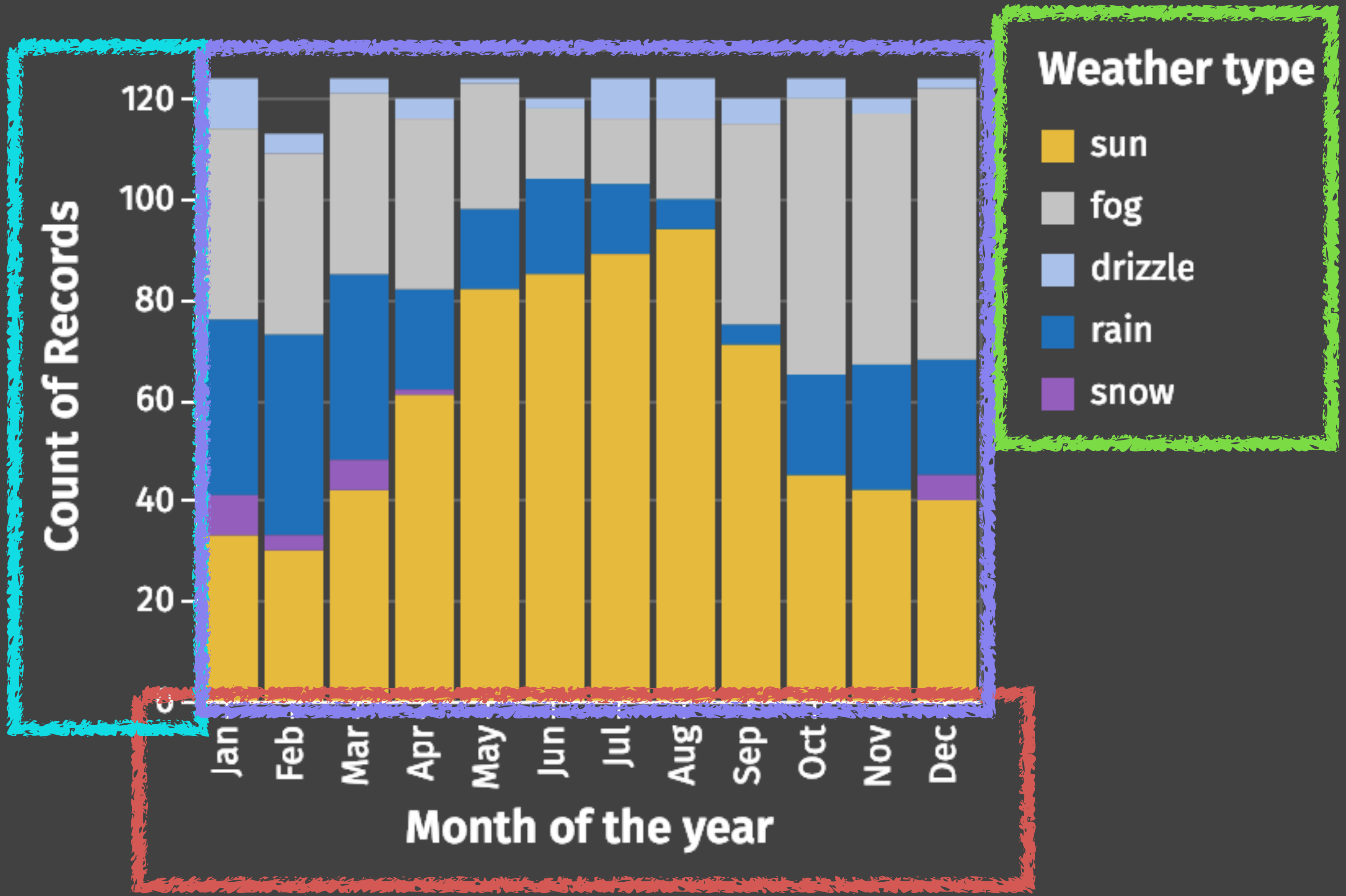






Vega-Lite

```
{
 data: {
    url: "seattle-weather.csv"
  },
  mark: "bar",
  encoding: {
    x: {
      timeUnit: "month",
      field: "date"
    },
    y: {
     aggregate: "count"
    },
    color: {
      field: "weather"
    }
  }
}
```

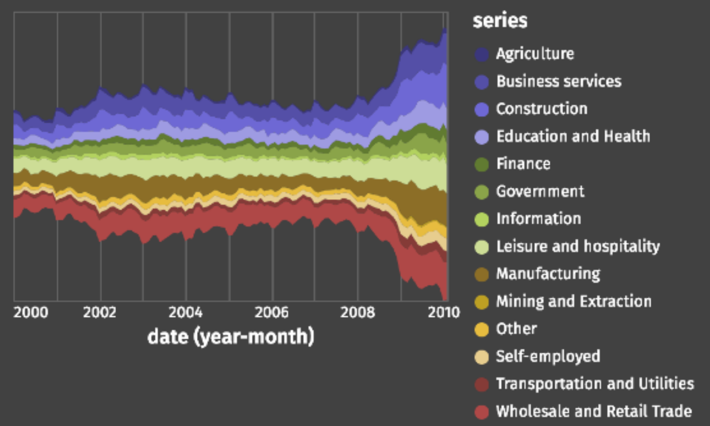


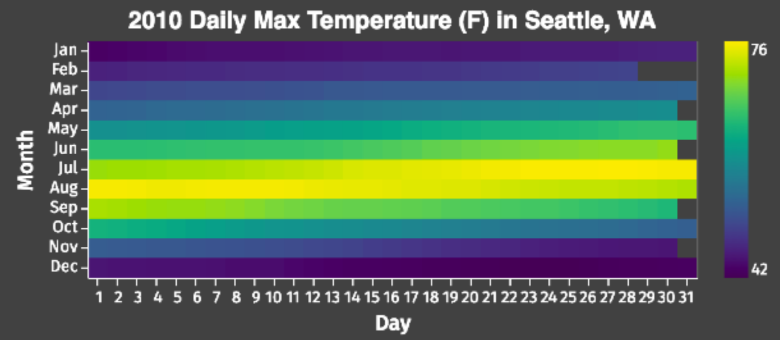
Vega-Lite

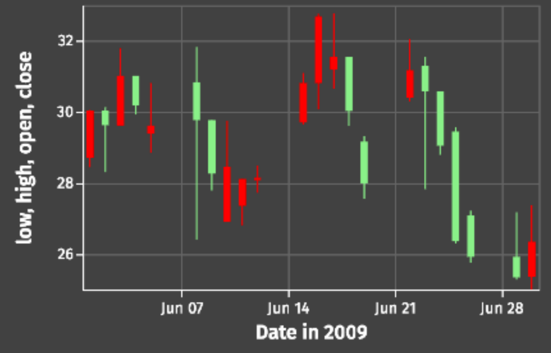
A high-level grammar for creating interactive multi-view charts.

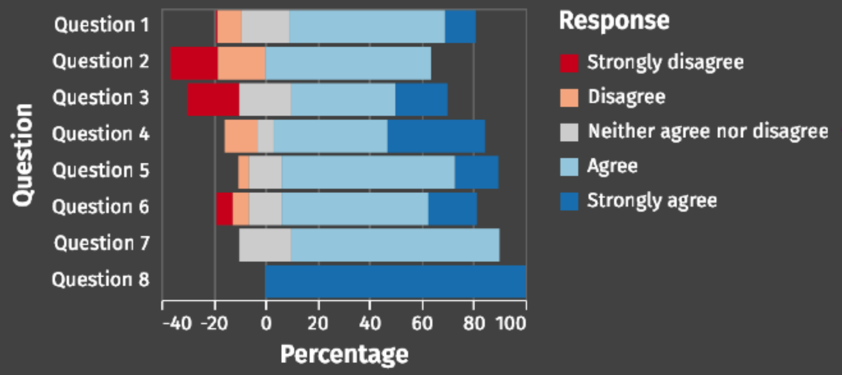
2.1M monthly downloads from CDN.
Used at , Microsoft, Google, Netflix, etc.

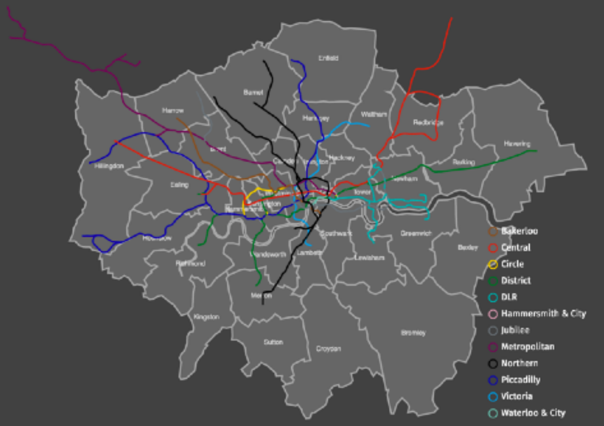
vega.github.io/vega-lite

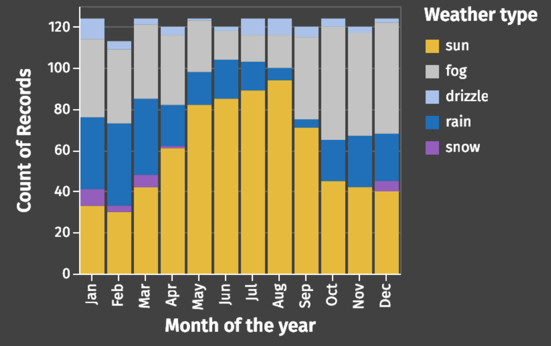


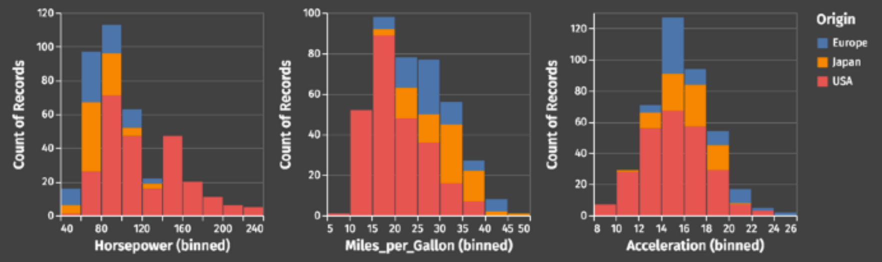






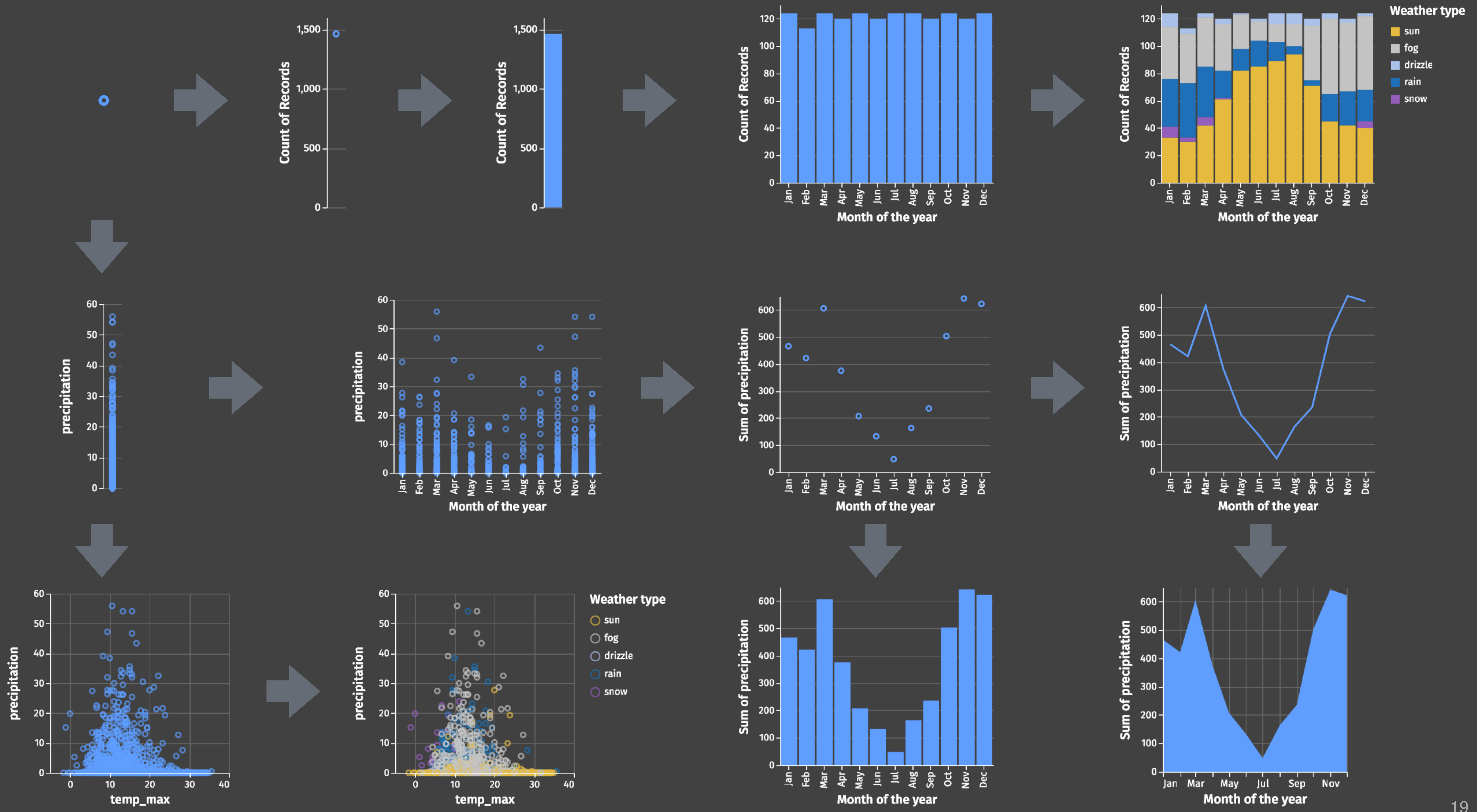






Takeaway:

Grammar-based visualization tools (such as Vega-Lite) support flexible interactive visualization and exploration. 🗣️ 📊

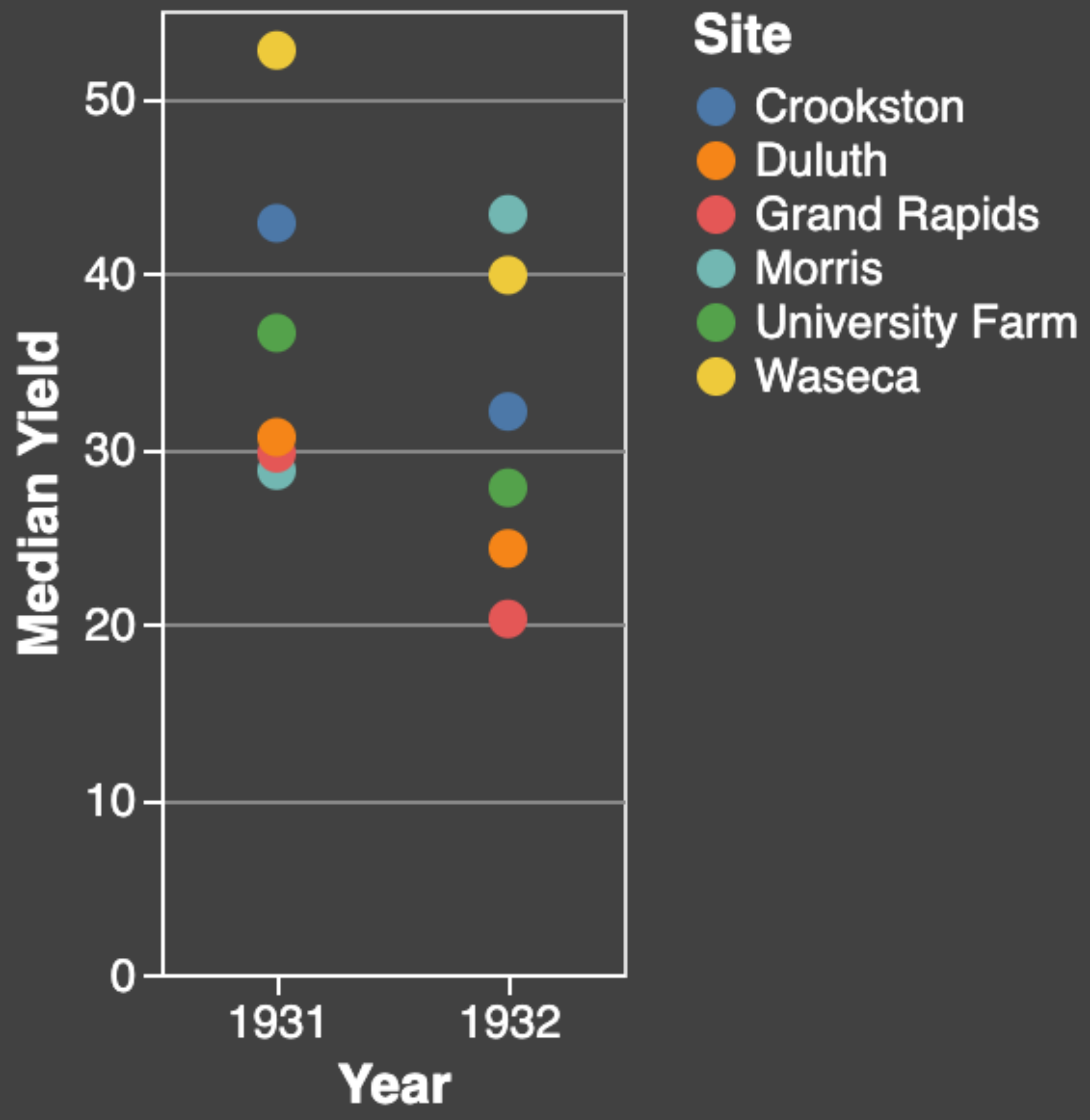

Count of Records
1,500
1,000
500
0
Count of Records
1,500
1,000
500
0
Count of Records
120
100
80
60
40
20
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month of the year
Weather type
sun
fog
drizzle
rain
snow
Count of Records
Month of the year
precipitation
60
50
40
30
20
10
0
precipitation
Month of the year
Sum of precipitation
600
500
400
300
200
100
0
Month of the year
Sum of precipitation
Month of the year
precipitation
temp_max
0
10
20
30
40
precipitation
temp_max
Weather type
sun
fog
drizzle
rain
snow
Sum of precipitation
Month of the year
Sum of precipitation
Jan Mar May Jul Sep Nov
Month of the year

How has Barley Yield Changed at Different Sites?

Year	Site	Median Yield
1931	"University Farm"	36.58
1931	"Waseca"	52.71
1931	"Morris"	28.73
1931	"Crookston"	42.85
1931	"Grand Rapids"	29.71
1931	"Duluth"	30.63
1932	"University Farm"	27.75
1932	"Waseca"	39.88
1932	"Morris"	43.36
1932	"Crookston"	32.09
1932	"Grand Rapids"	20.26
1932	"Duluth"	24.28

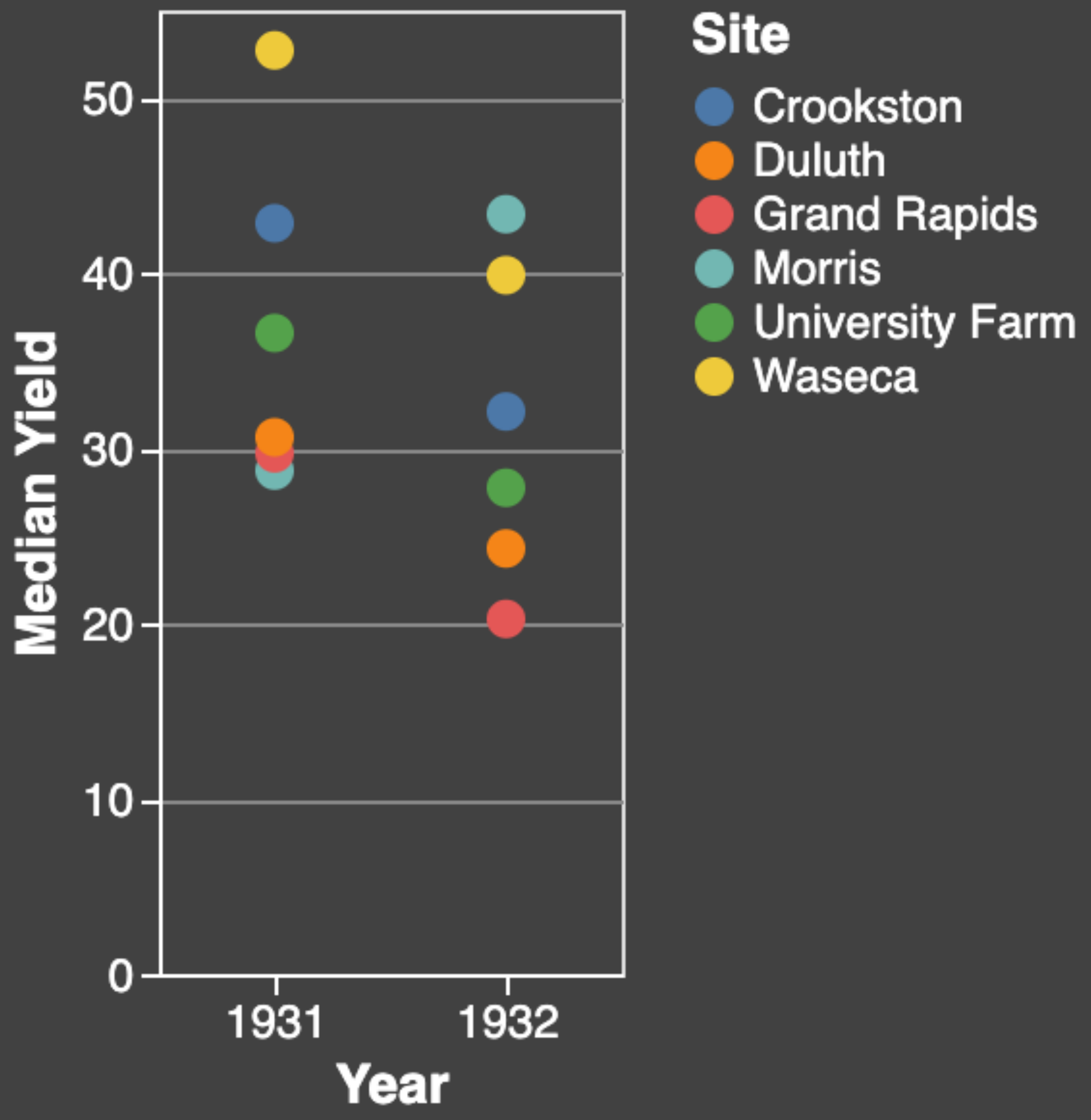
How has Barley Yield Changed at Different Sites?

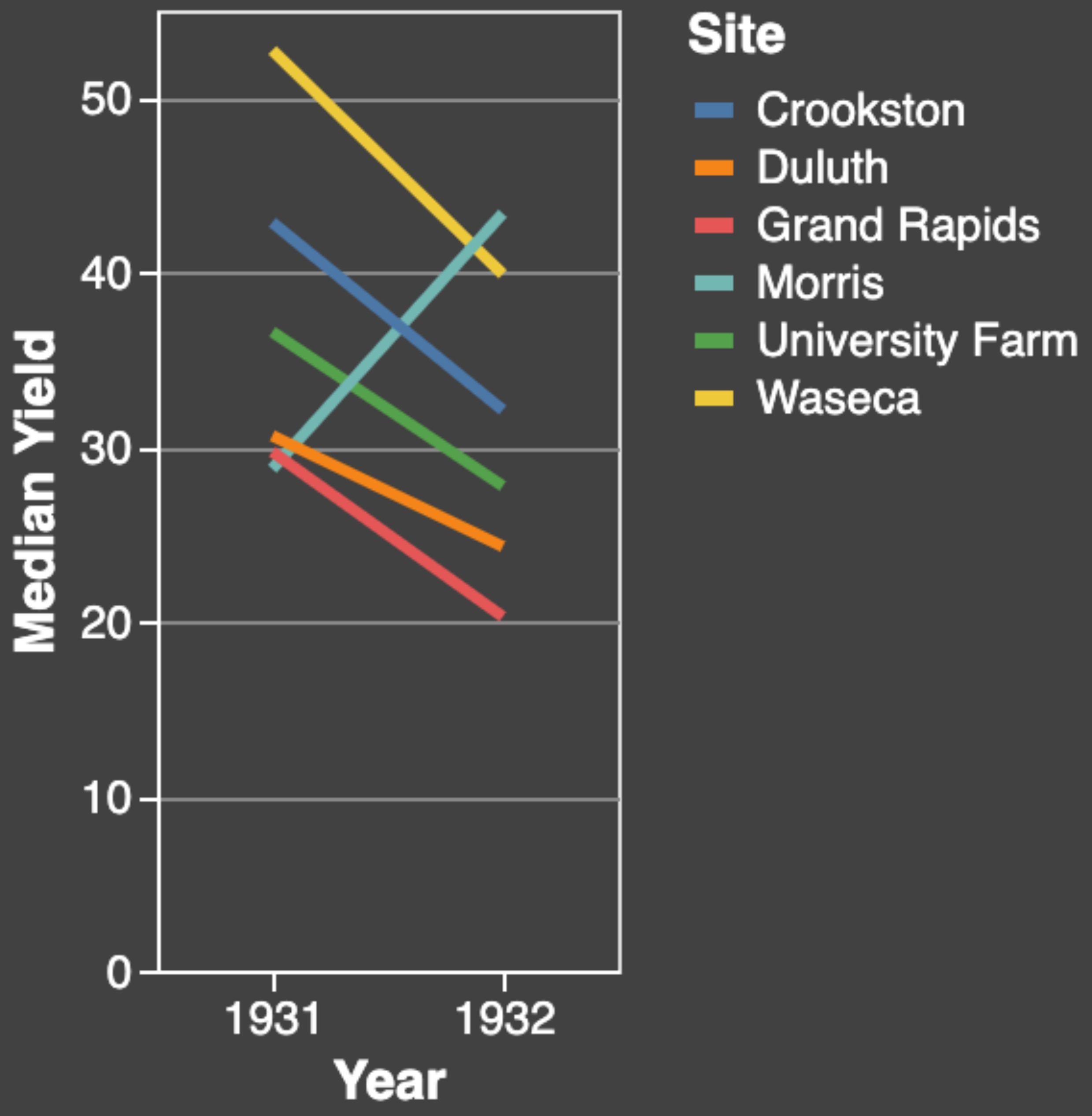
Year	Site	Median Yield
1931	"University Farm"	36.58
1931	"Waseca"	52.71
1931	"Morris"	28.73
1931	"Crookston"	42.85
1931	"Grand Rapids"	29.71
1931	"Duluth"	30.63
1932	"University Farm"	27.75
1932	"Waseca"	39.88
1932	"Morris"	43.36
1932	"Crookston"	32.09
1932	"Grand Rapids"	20.26
1932	"Duluth"	24.28



How has Barley Yield Changed at Different Sites?

n
.58
.71
.73
.85
.71
.63
.75
.88
.36
.09
.26
.28

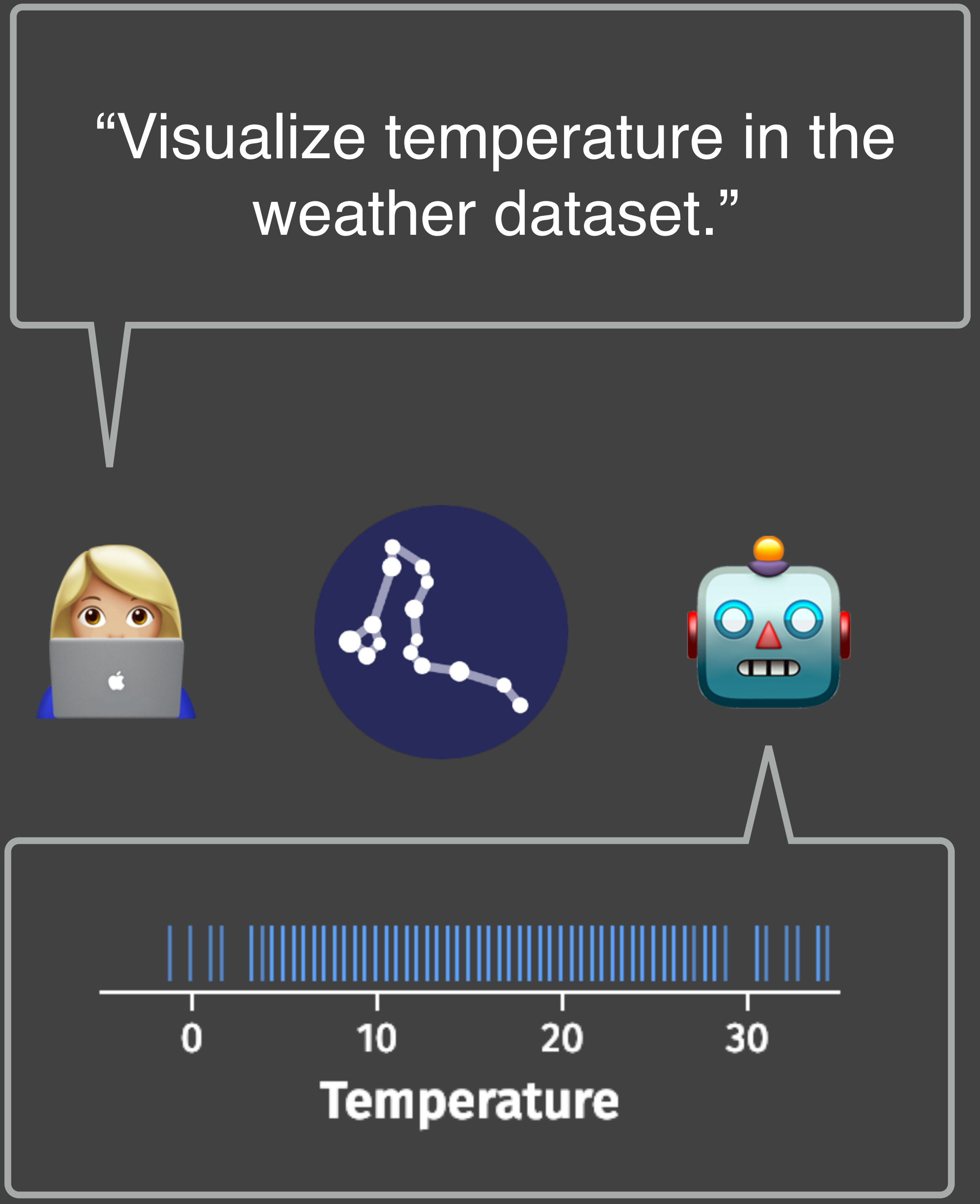




Draco

A formal model of visualization design with learned constraints. Can be used to automatically create “good” visualizations.

New version is in progress at github.com/cmudig/draco2



“Visualize temperature in the weather dataset.”

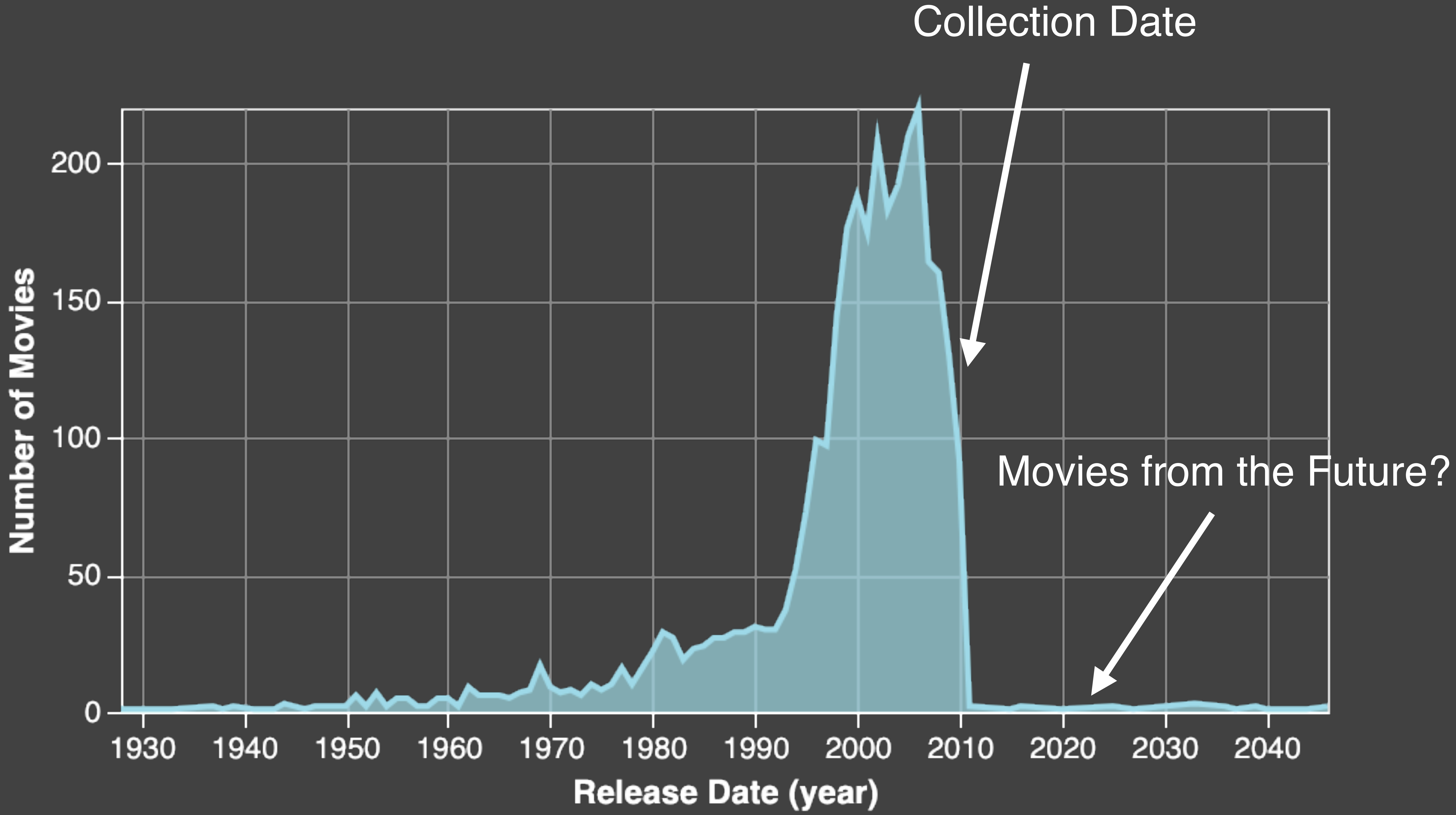
Takeaway:

When we design for perception, otherwise hidden patterns emerge. 👁️ 🧠

Design Visualizations → Explore Data
Make Decisions

Movie Data

Title	String (N)
IMDB Rating	Number (Q)
Rotten Tomatoes Rating	Number (Q)
MPAA Rating	String (O)
Release Date	Date (T)



Collection Date

Movies from the Future?

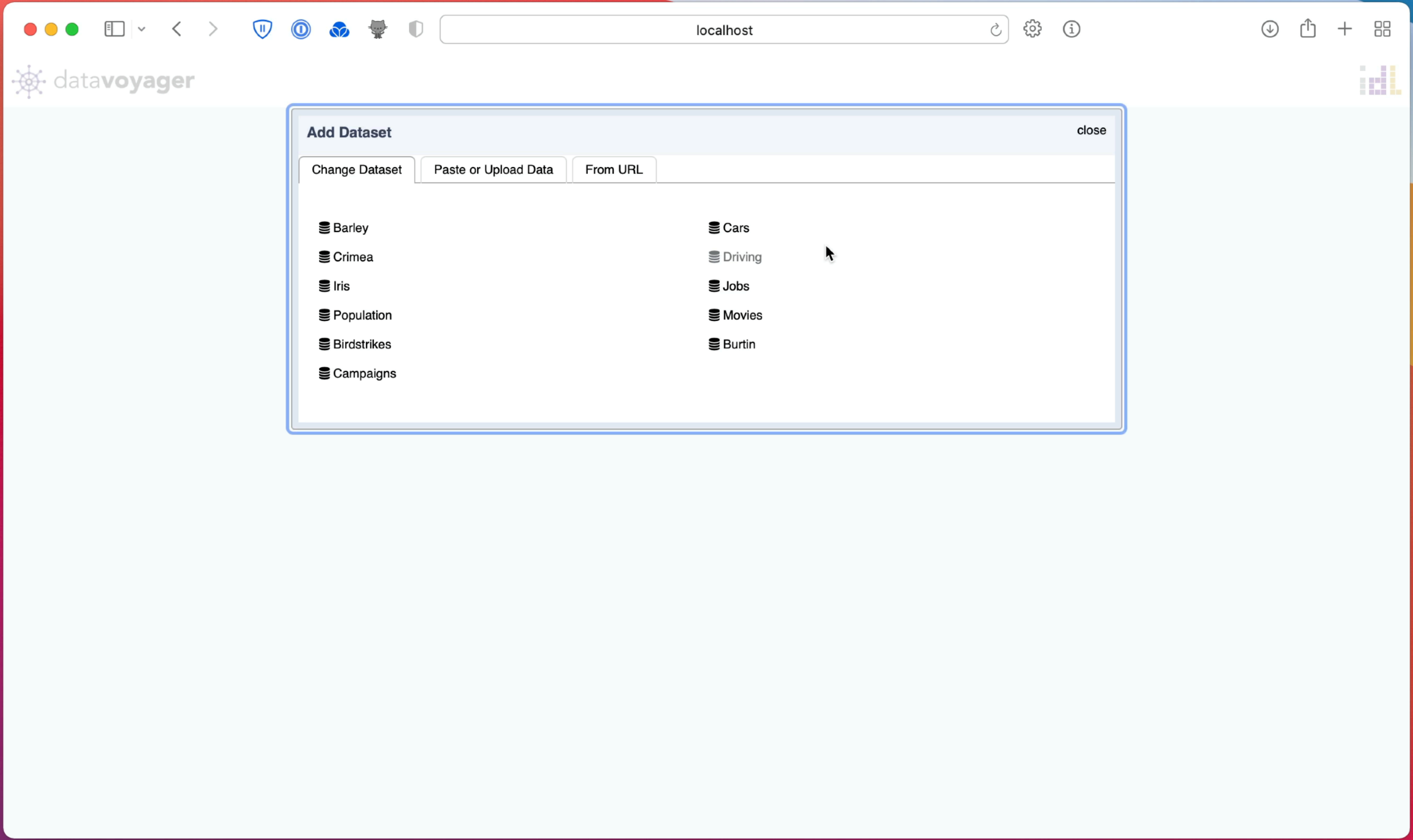
Common Analysis Pitfalls:

Overlook data quality issues

Fixate on specific relationships

Other cognitive biases

github.com/vega/voyager



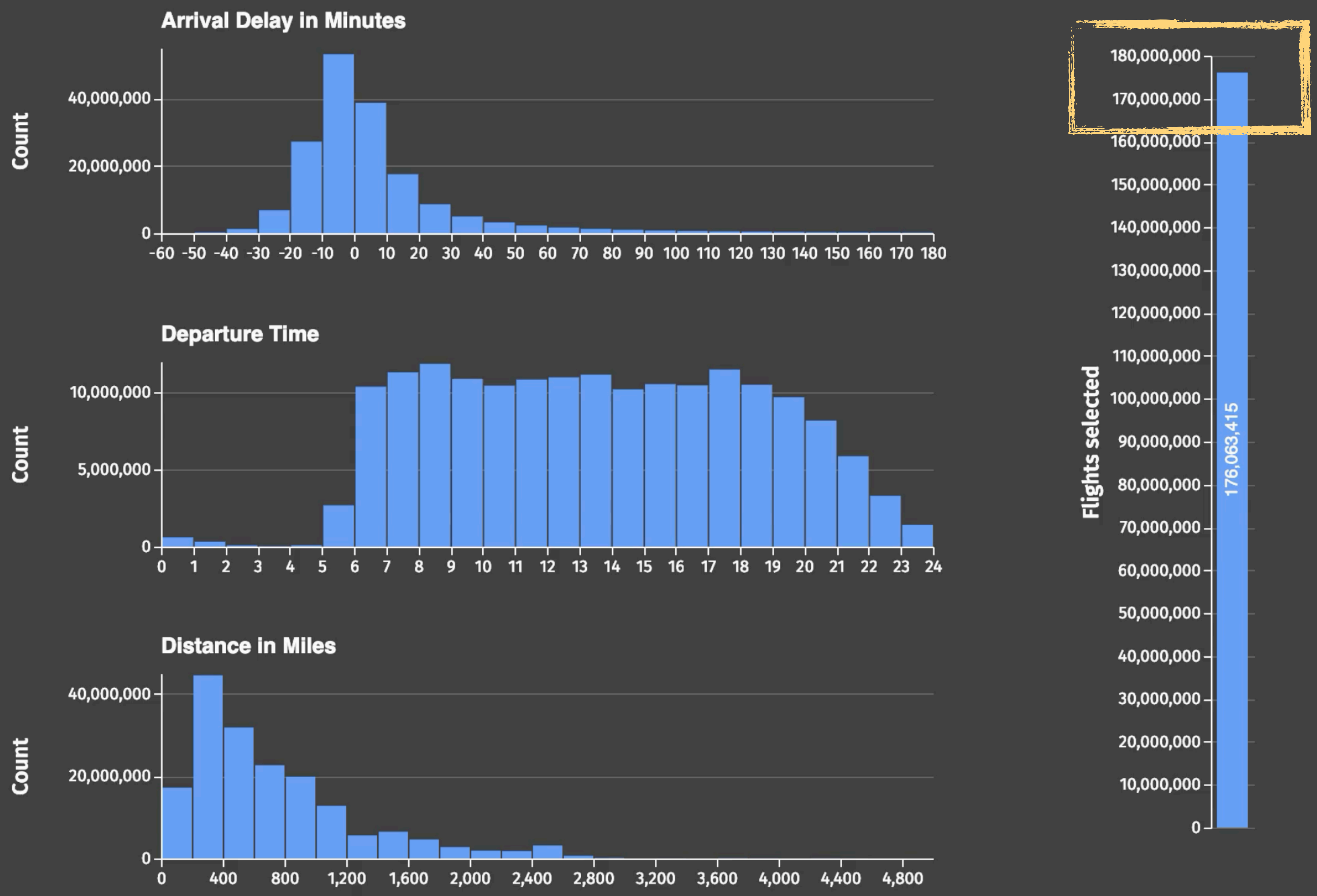
Add Dataset

- Barley
- Cars
- Crimea
- Driving
- Iris
- Jobs
- Population
- Movies
- Birdstrikes
- Burtin
- Campaigns

Takeaway:

UI tools can encourage best-practices. 🖥️ ✅

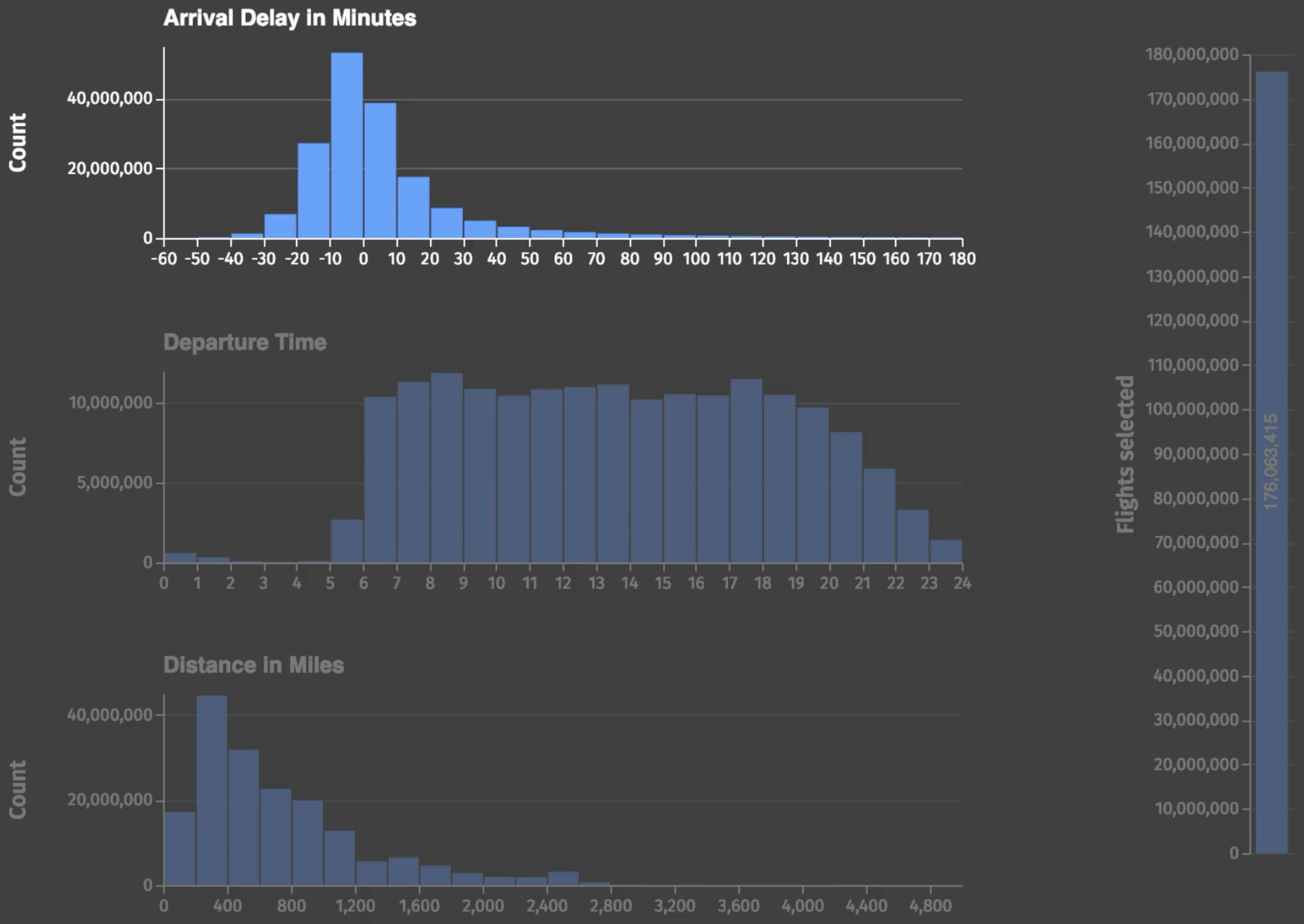
BRITISH AI
G-VIIA
IA

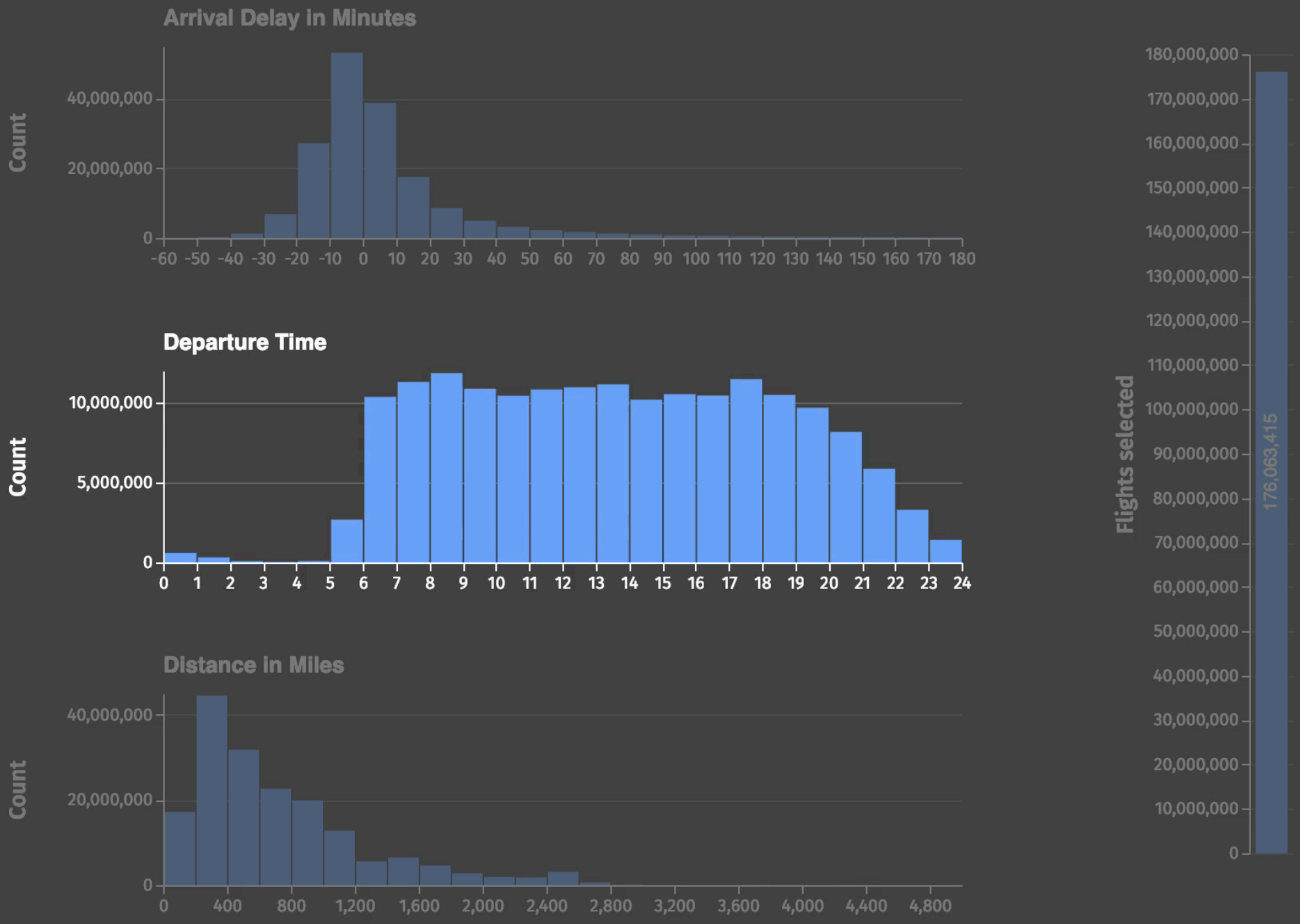


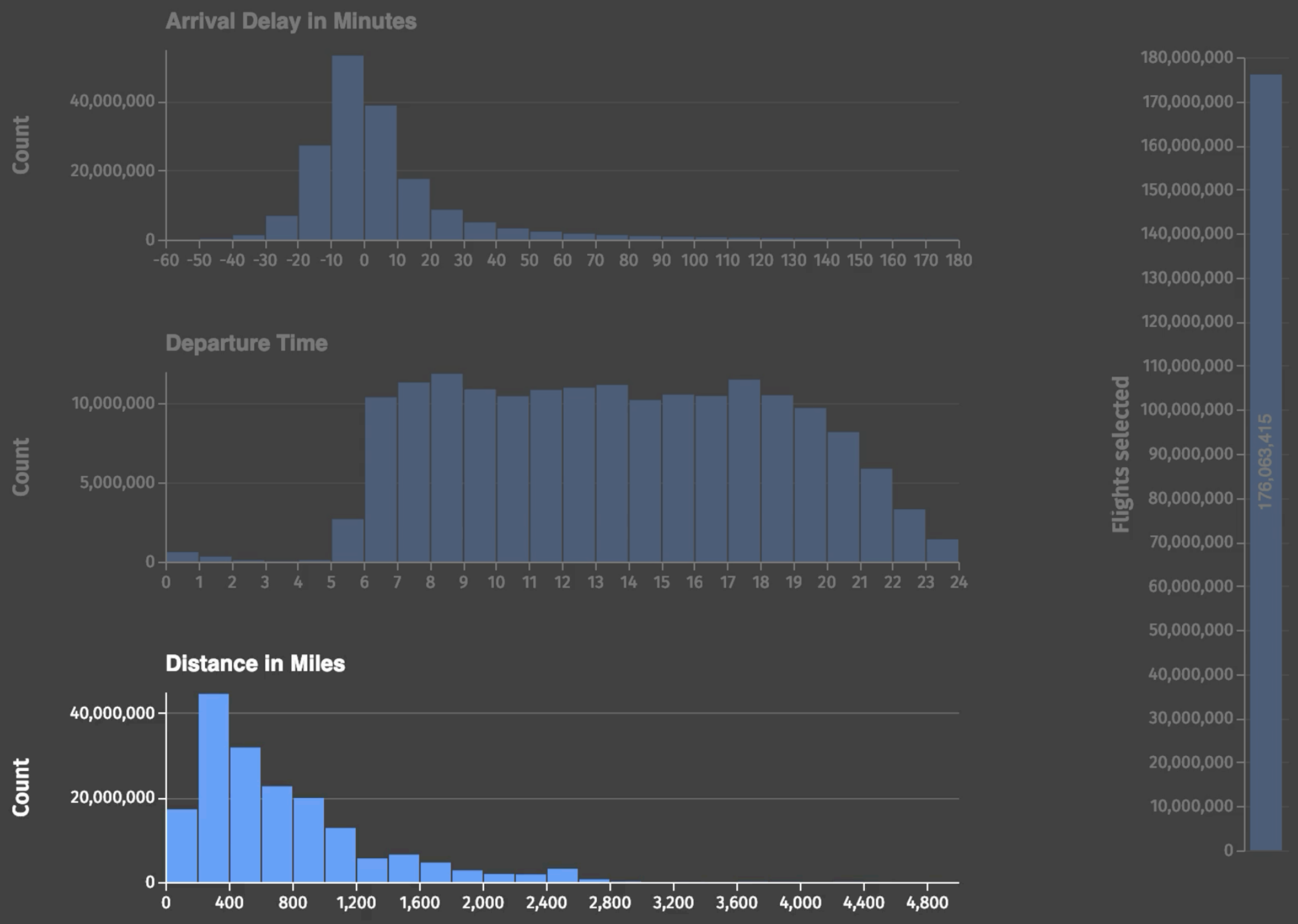
Arrival Delay in Minutes

Departure Time

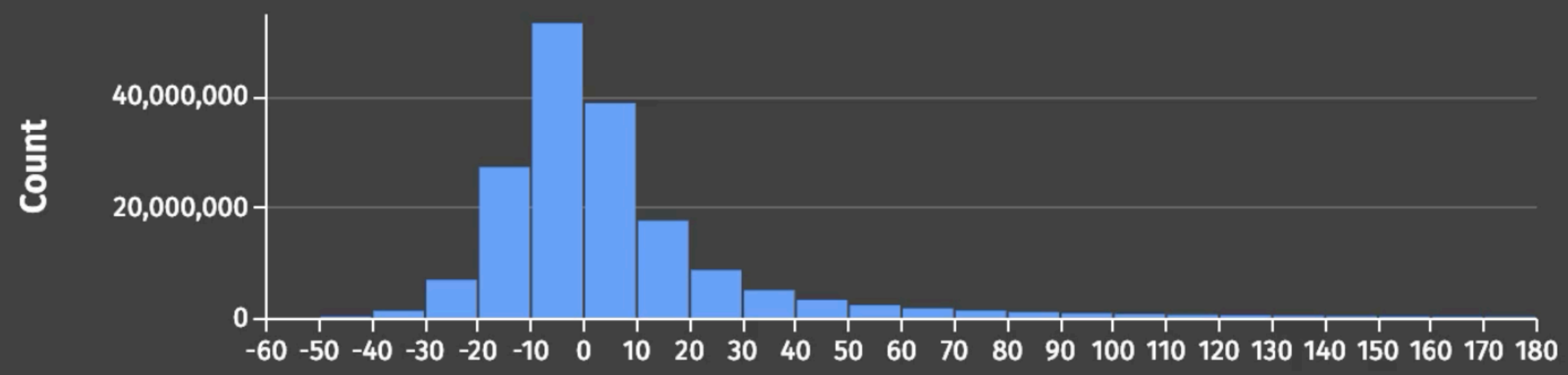
Distance in Miles


Arrival Delay in Minutes
Count
40,000,000
20,000,000
0
-60 -50 -40 -30 -20 -10 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180
Departure Time
Count
10,000,000
5,000,000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Distance in Miles
Count
40,000,000
20,000,000
0
0 400 800 1,200 1,600 2,000 2,400 2,800 3,200 3,600 4,000 4,400 4,800
Flights selected
180,000,000
170,000,000
160,000,000
150,000,000
140,000,000
130,000,000
120,000,000
110,000,000
100,000,000
90,000,000
80,000,000
70,000,000
60,000,000
50,000,000
40,000,000
30,000,000
20,000,000
10,000,000
0
176,063,415

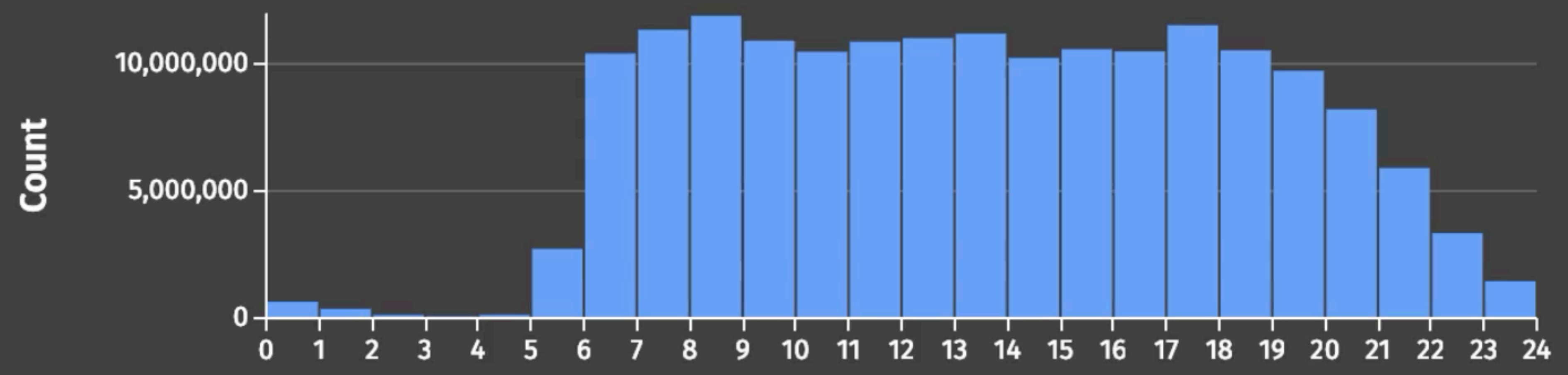

Arrival Delay in Minutes
Count
0
20,000,000
40,000,000
-60 -50 -40 -30 -20 -10 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180
Departure Time
Count
0
5,000,000
10,000,000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Distance in Miles
Count
0
20,000,000
40,000,000
0 400 800 1,200 1,600 2,000 2,400 2,800 3,200 3,600 4,000 4,400 4,800
Flights selected
0
10,000,000
20,000,000
30,000,000
40,000,000
50,000,000
60,000,000
70,000,000
80,000,000
90,000,000
100,000,000
110,000,000
120,000,000
130,000,000
140,000,000
150,000,000
160,000,000
170,000,000
180,000,000
176,063,415


Arrival Delay in Minutes
Count
40,000,000
20,000,000
0
-60 -50 -40 -30 -20 -10 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180
Departure Time
Count
10,000,000
5,000,000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Distance in Miles
Count
40,000,000
20,000,000
0
0 400 800 1,200 1,600 2,000 2,400 2,800 3,200 3,600 4,000 4,400 4,800
Flights selected
180,000,000
170,000,000
160,000,000
150,000,000
140,000,000
130,000,000
120,000,000
110,000,000
100,000,000
90,000,000
80,000,000
70,000,000
60,000,000
50,000,000
40,000,000
30,000,000
20,000,000
10,000,000
0
176,063,415

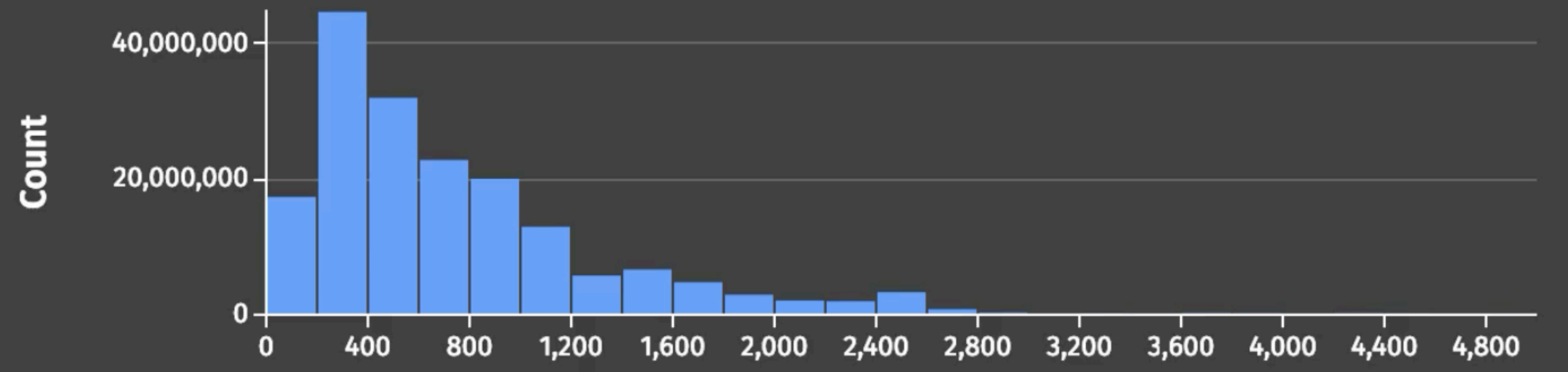
Arrival Delay in Minutes

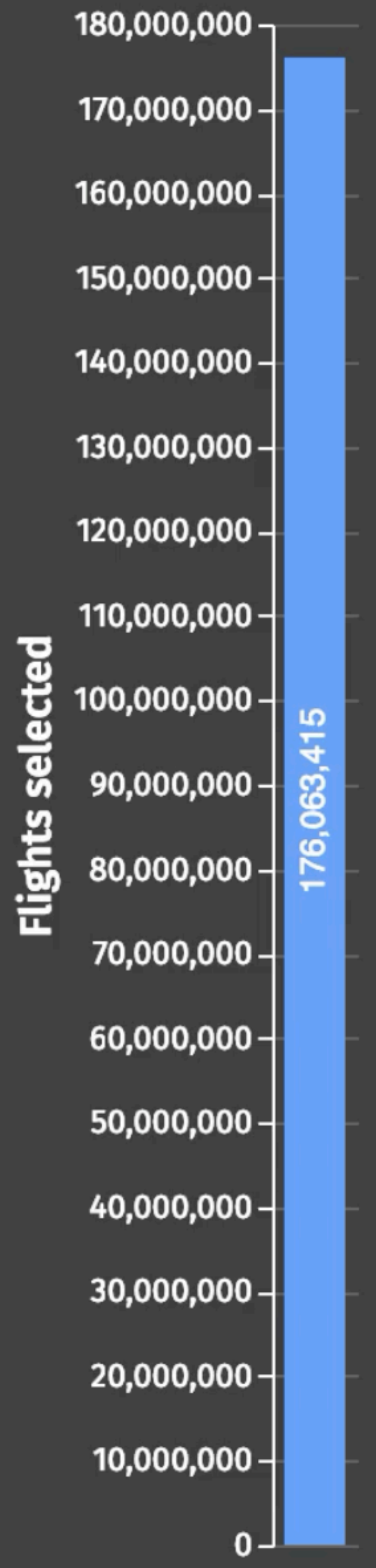


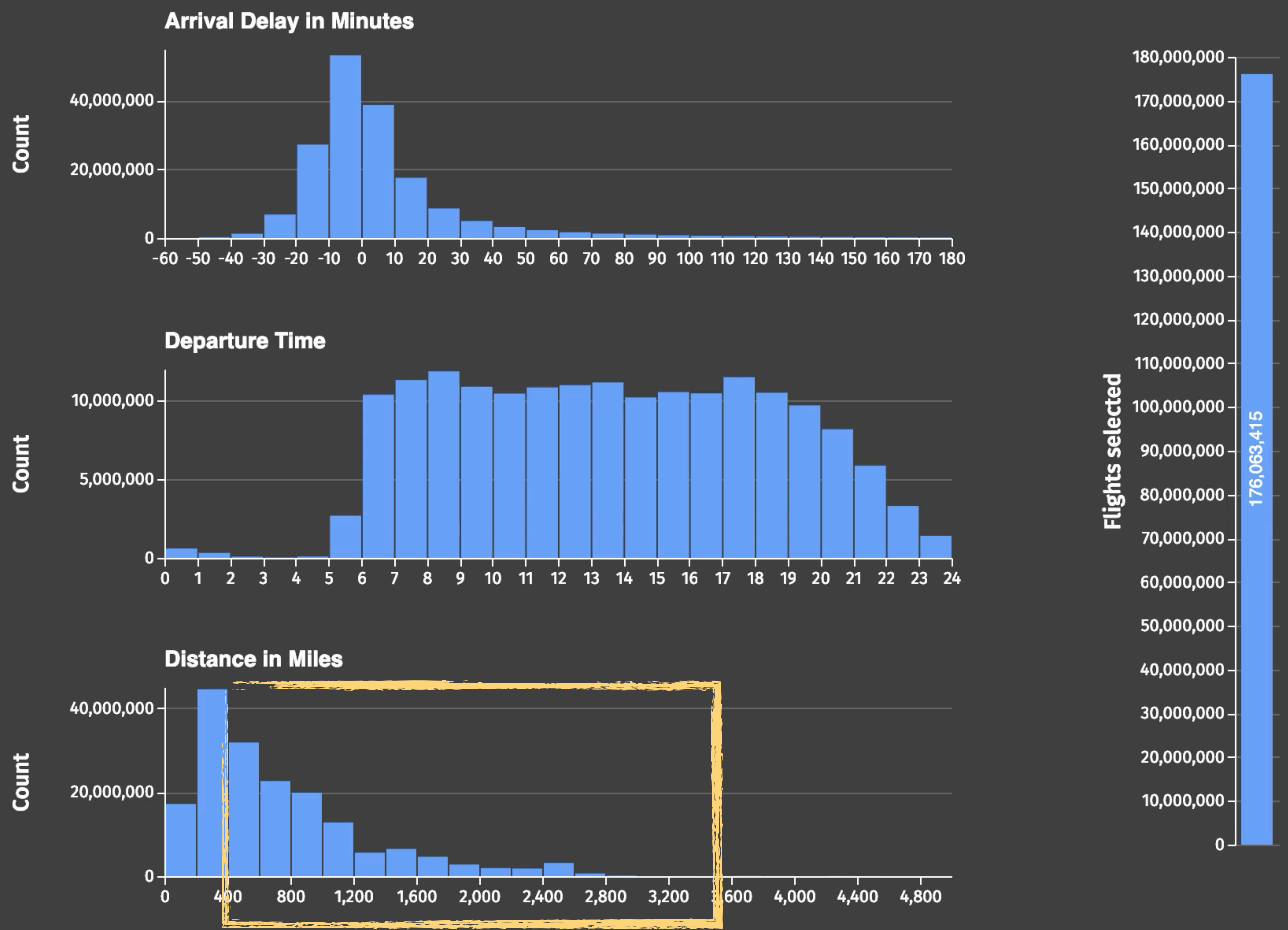
Departure Time



Distance in Miles



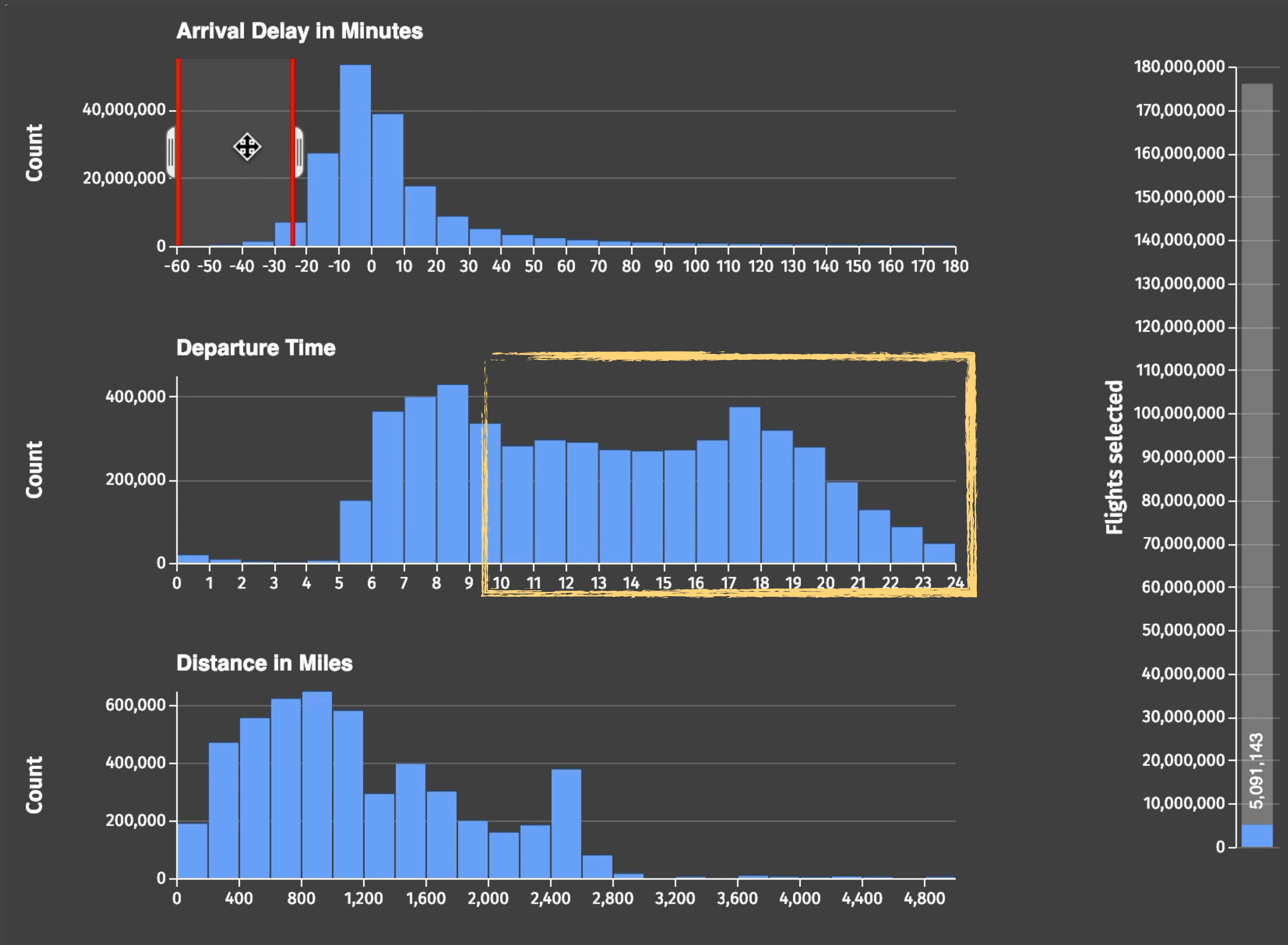


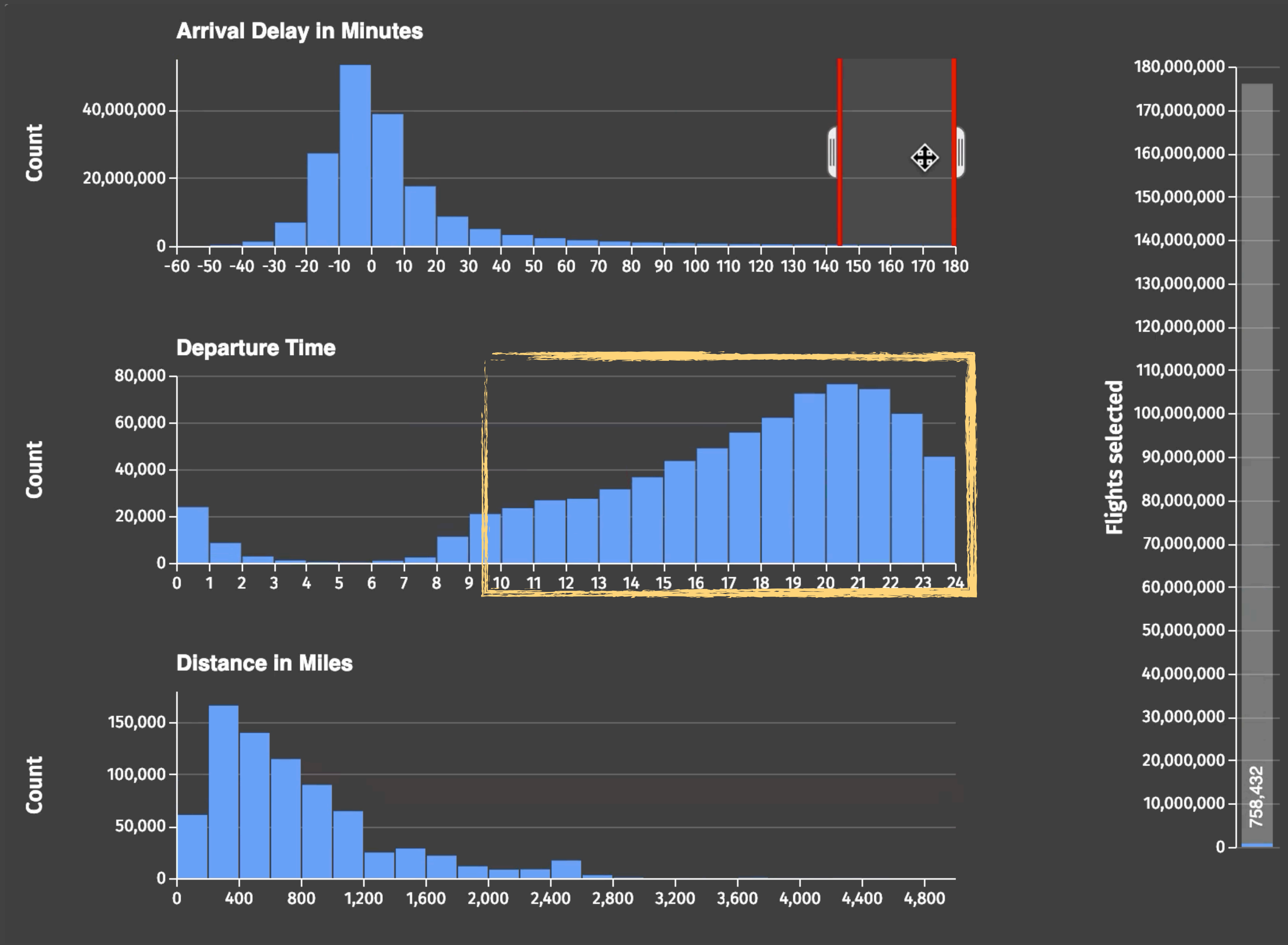


Arrival Delay in Minutes

Departure Time

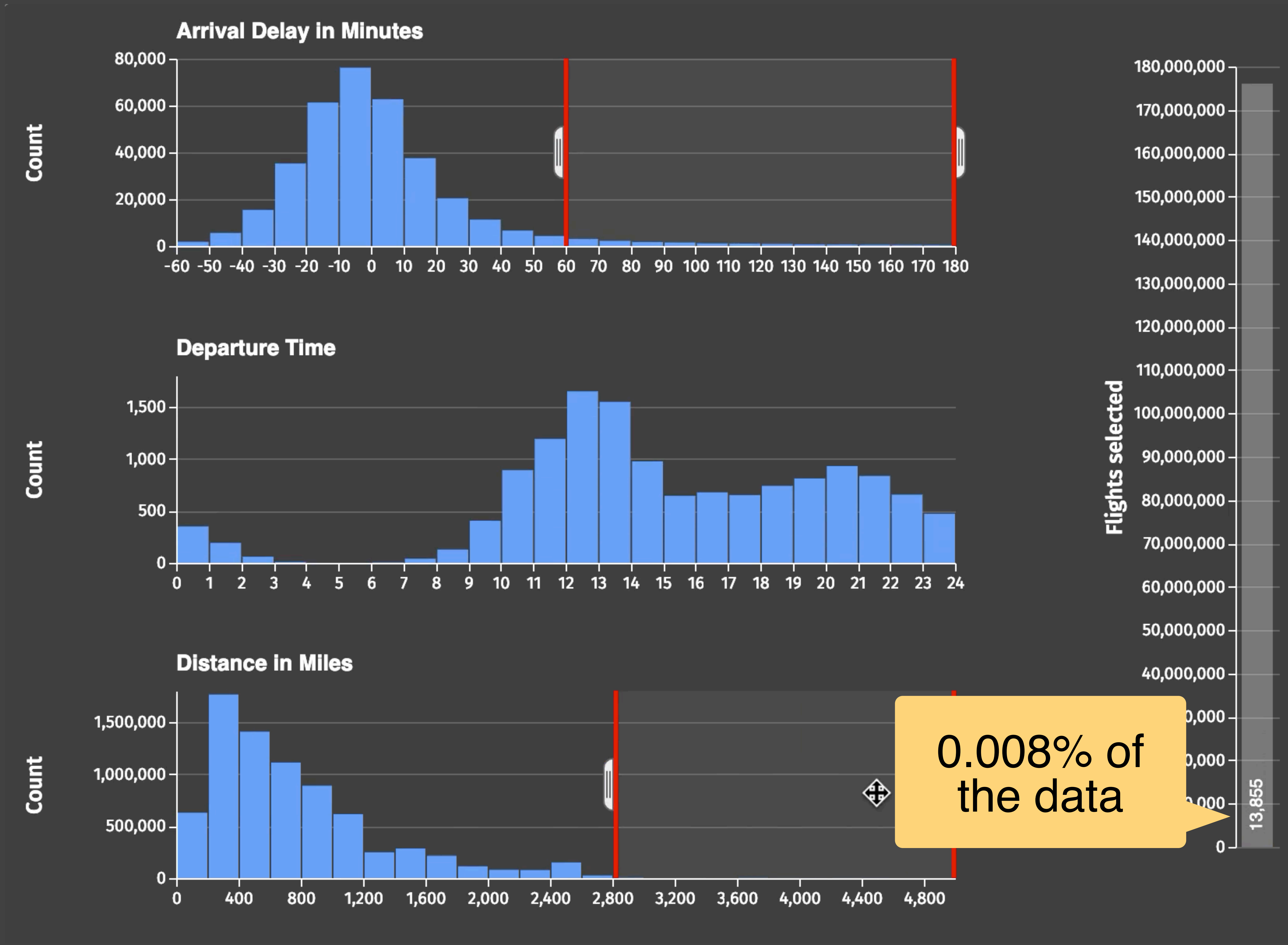
Distance in Miles


Arrival Delay in Minutes
Count
40,000,000
20,000,000
0
-60 -50 -40 -30 -20 -10 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180
Departure Time
Count
400,000
200,000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Distance in Miles
Count
600,000
400,000
200,000
0
0 400 800 1,200 1,600 2,000 2,400 2,800 3,200 3,600 4,000 4,400 4,800
Flights selected
180,000,000
170,000,000
160,000,000
150,000,000
140,000,000
130,000,000
120,000,000
110,000,000
100,000,000
90,000,000
80,000,000
70,000,000
60,000,000
50,000,000
40,000,000
30,000,000
20,000,000
10,000,000
0
5,091,143


Arrival Delay in Minutes
Count
40,000,000
20,000,000
0
-60 -50 -40 -30 -20 -10 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180
Departure Time
Count
80,000
60,000
40,000
20,000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Distance in Miles
Count
150,000
100,000
50,000
0
0 400 800 1,200 1,600 2,000 2,400 2,800 3,200 3,600 4,000 4,400 4,800
Flights selected
180,000,000
170,000,000
160,000,000
150,000,000
140,000,000
130,000,000
120,000,000
110,000,000
100,000,000
90,000,000
80,000,000
70,000,000
60,000,000
50,000,000
40,000,000
30,000,000
20,000,000
10,000,000
0
758,432

Takeaway:

Interactivity enables us to find patterns that exist across multiple dimensions. 👆 🔢



0.008% of the data

How do we interact with billion+record datasets in real-time?

How do we interact with billion+record datasets in real-time?

Delays reduce engagement and lead to fewer observations.

The Effect of Interactive Latency. Liu, Heer. *IEEE Infovis 2014.*

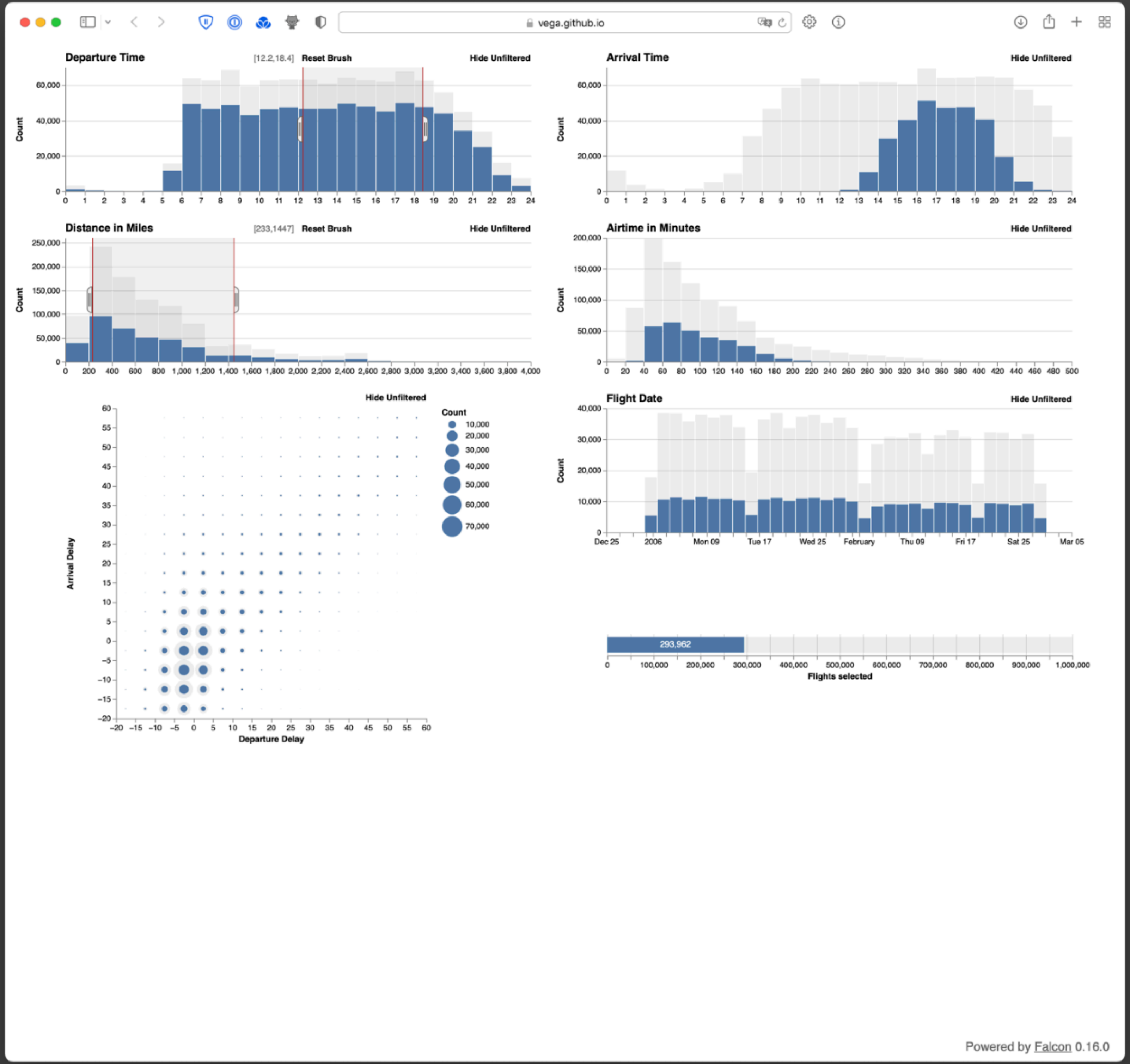
Delays may bias analysts towards convenient data.

Falcon

Zero-latency crossfiltering for massive datasets.

Leverages smart prefetching and precomputation designed around human perception.

Available as open source at github.com/vega/falcon



Takeaway:

Interactivity should be real-time regardless of the scale of the data. ⏱️ 🌎

Machine Learning, AI, and Statistics are people problems. For them to be effective, we need to **design for human involvement**. 🤖 👩‍💻

Grammar-based visualization tools (such as Vega-Lite) support flexible interactive visualization and exploration. 🗣️ 📊

When we **design for perception**, otherwise hidden patterns emerge. 👁️ 🧠

UI tools can encourage best-practices. 🖥️ ✅

Interactivity enables us to find patterns that exist **across multiple dimensions**. 👆 🔢

Interactivity should be **real-time** regardless of the scale of the data. ⏱️ 🌎

Research Mission

Empower everyone to effectively analyze and communicate data, by *designing interactive systems that richly integrate the strengths of both people and machines*.

dig.cmu.edu