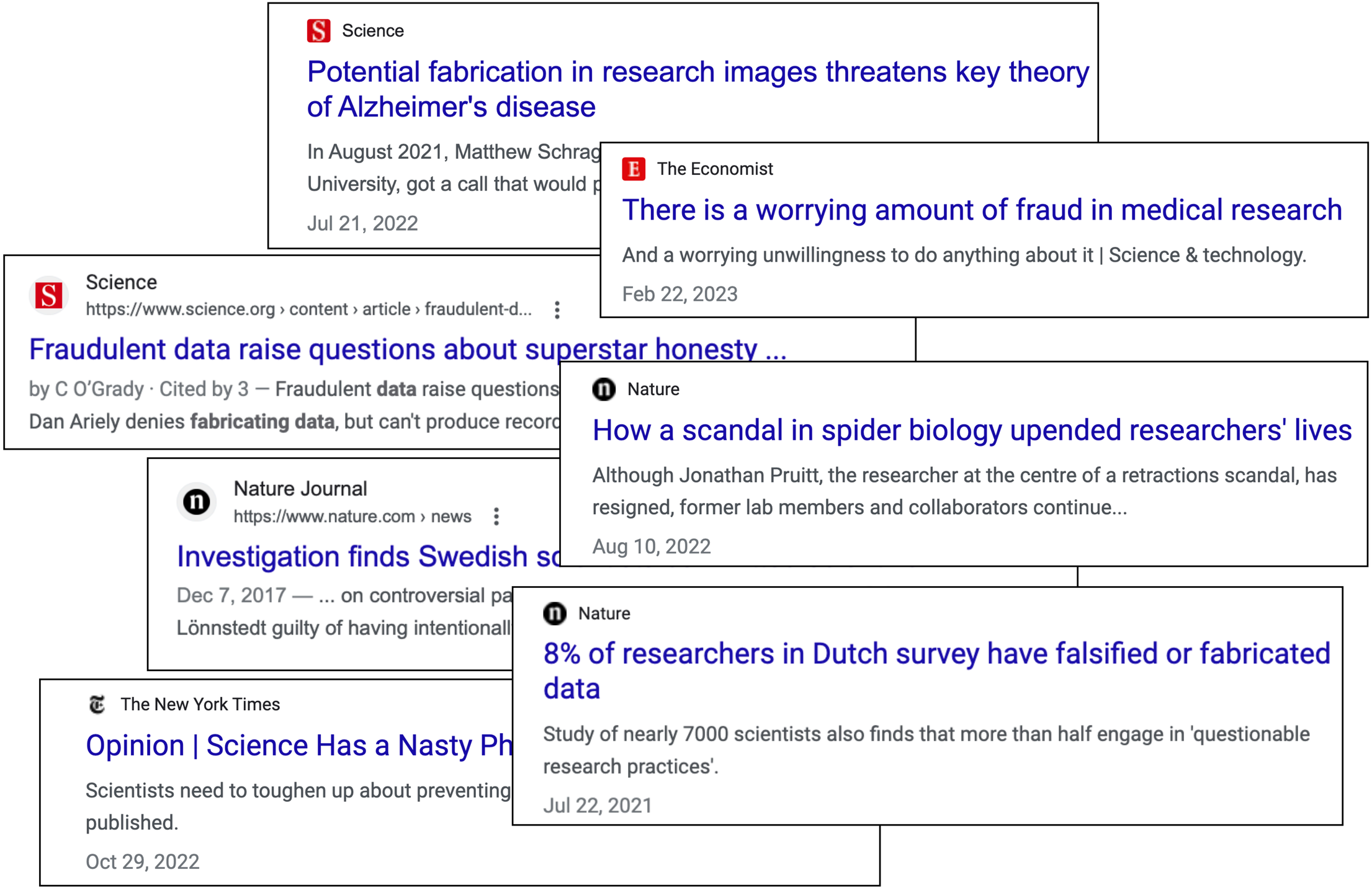
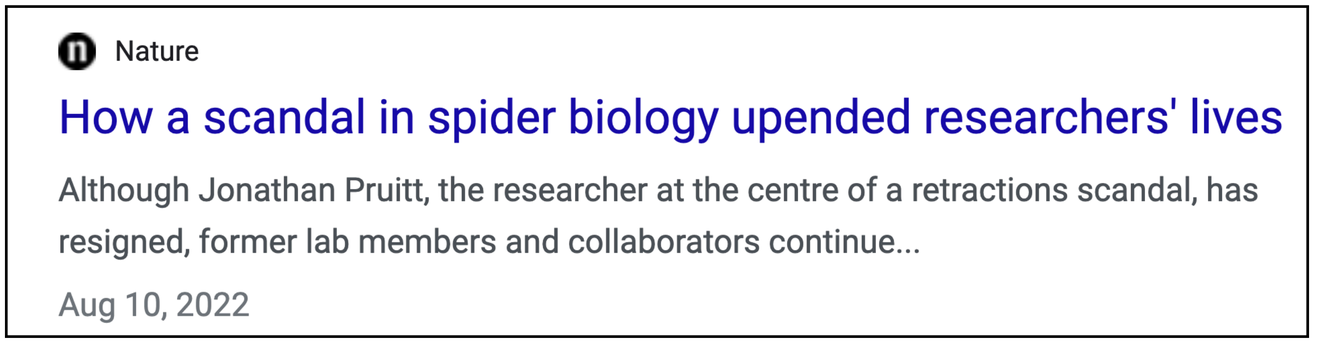
Ferret: Reviewing Tabular Datasets for Manipulation

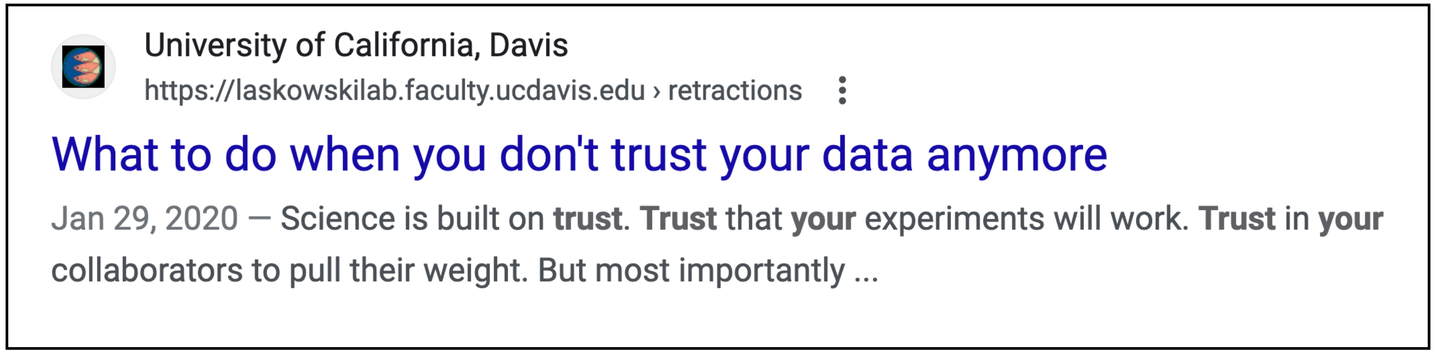
Devin Lange, Shaurya Sahai, Jeff M. Phillips, Alexander Lex

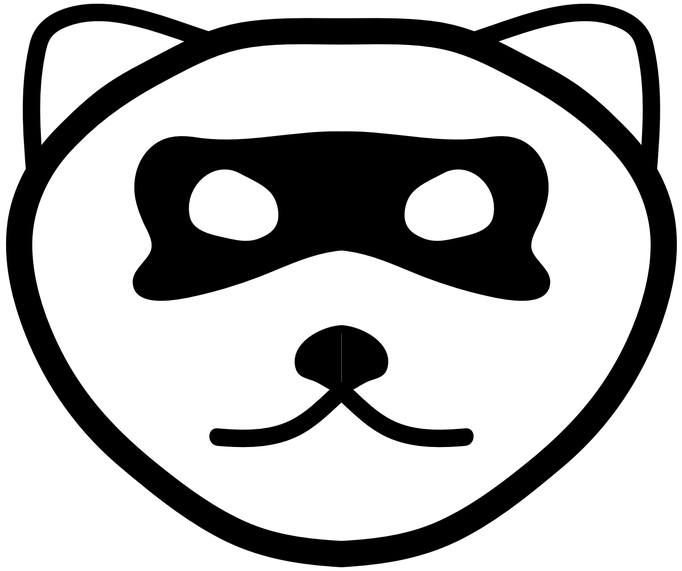
THE UNIVERSITY OF UTAH

visualization design lab


Science
Potential fabrication in research images threatens key theory of Alzheimer's disease
In August 2021, Matthew Schrag
University, got a call that would p
Jul 21, 2022
The Economist
There is a worrying amount of fraud in medical research
And a worrying unwillingness to do anything about it | Science & technology.
Feb 22, 2023
Science
https://www.science.org › content › article › fraudulent-d...
Fraudulent data raise questions about superstar honesty ...
by C O'Grady · Cited by 3 — Fraudulent data raise questions
Dan Ariely denies fabricating data, but can't produce record
Nature
How a scandal in spider biology upended researchers' lives
Although Jonathan Pruitt, the researcher at the centre of a retractions scandal, has resigned, former lab members and collaborators continue...
Aug 10, 2022
Nature Journal
https://www.nature.com › news
Investigation finds Swedish sc
Dec 7, 2017 — ... on controversial pa
Lönnstedt guilty of having intentionall
Nature
8% of researchers in Dutch survey have falsified or fabricated data
Study of nearly 7000 scientists also finds that more than half engage in 'questionable research practices'.
Jul 22, 2021
The New York Times
Opinion | Science Has a Nasty Ph
Scientists need to toughen up about preventing
published.
Oct 29, 2022


Nature
How a scandal in spider biology upended researchers' lives
Although Jonathan Pruitt, the researcher at the centre of a retractions scandal, has resigned, former lab members and collaborators continue...
Aug 10, 2022

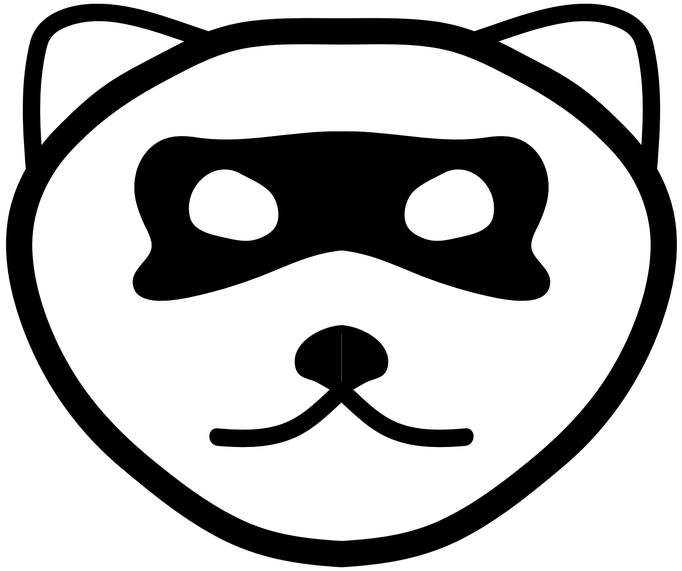

University of California, Davis
https://laskowskilab.faculty.ucdavis.edu › retractions
What to do when you don't trust your data anymore
Jan 29, 2020 — Science is built on trust. Trust that your experiments will work. Trust in your collaborators to pull their weight. But most importantly ...



Understand Fraud

Create Tools

Implement Policies



Understand Fraud

Methods: Datasets and Patterns of Manipulation

Literature Review → **3** datasets and initial set of patterns

Retraction Watch Database[1] → **1161** papers with "**Falsification/Fabrication of Data**" tag
Reviewed **103** → (only) **1** tabular dataset

Social Media → **4** datasets

Interviews → **2** datasets and additional patterns

[1] THE CENTER FOR SCIENTIFIC INTEGRITY: The Retraction Watch Database, 2018.

We identified 8 common patterns of manipulation

Unexpected Formatting

Real Data

Column A	Column B	Column C
45	79	10
16	93	92
3	96	98

Fabricated Data

Column A	Column B	Column C
45	79	10
16	93	92
3	96	98

Unexpected Formatting

Real/Fabricated Data

Column A	Column B	Column C
45	79	10
16	93	92
3	96	98
30	87	96
12	45	33
45	89	11

Fabricated Data

Column A	Column B	Column C
45	79	10
16	93	92
3	96	98

Duplicate Numbers and Digits

Column A	Column B	Column C

Duplicate Numbers and Digits

Column A	Column B	Column C
56	28	564
82	56	67
56	8	245

Duplicate Numbers and Digits

Column A	Column B	Column C
56	28	564
82	56	67
56	8	245

Unexpected Leading Digits

Unexpected Trailing Digits

Repeated Regions

Column A	Column B	Column C
45	79	10
16	**93**	92
3	**96**	98
30	87	96
12	45	33
45	89	11
70	56	14

Repeated Regions

Column A	Column B	Column C
45	79	10
16	**93**	92
3	**96**	98
30	87	96
12	45	33
45	**16**	**93**
70	**3**	**96**

Unexpected Variation of Precision

Column A
=random()

Unexpected Variation of Precision

Column A
0.9126470121876

Unexpected Variation of Precision

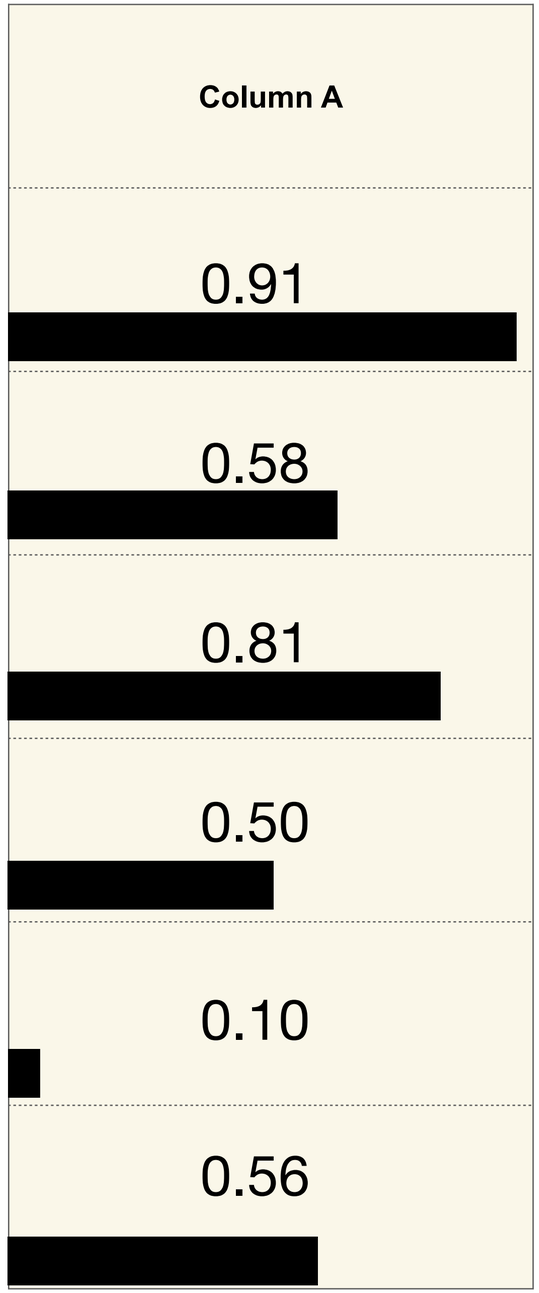
Column A
0.9126470121876
0.5870513854838
0.8126026019113
0.5059135817684
0.56808564109646
0.10941148406697

Unexpected Variation of Precision

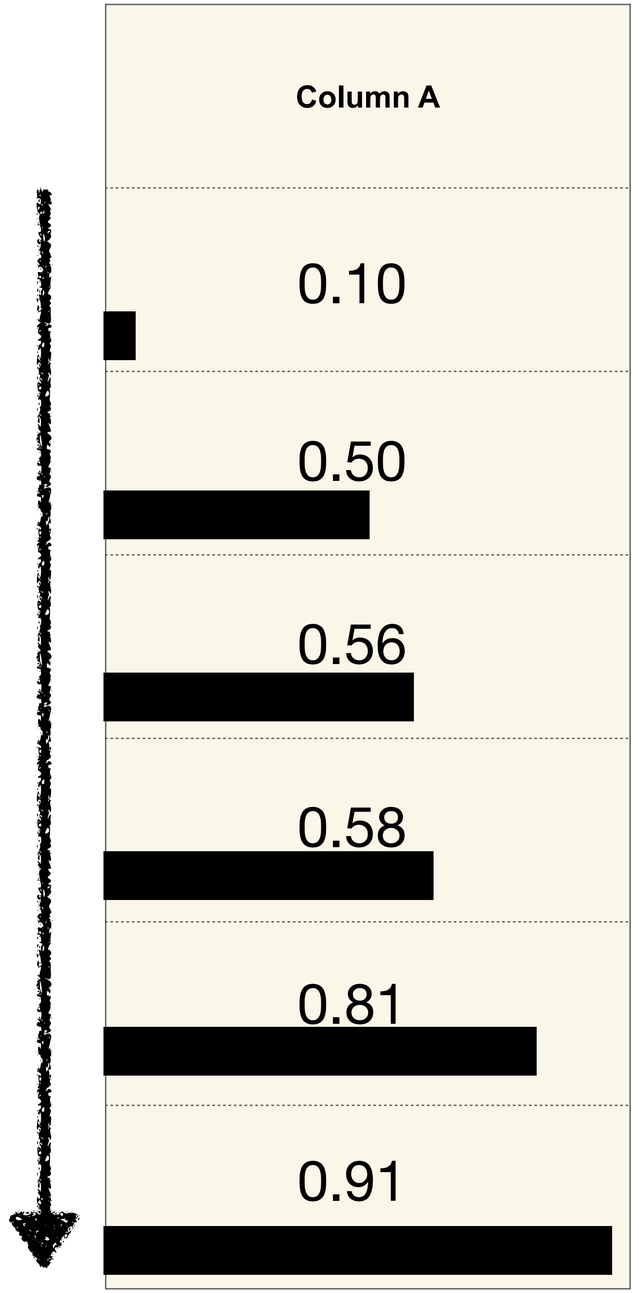
Column A
0.9126470121876
0.27
0.8126026019113
0.5059135817684
0.12
0.10941148406697

Ordering Artifacts

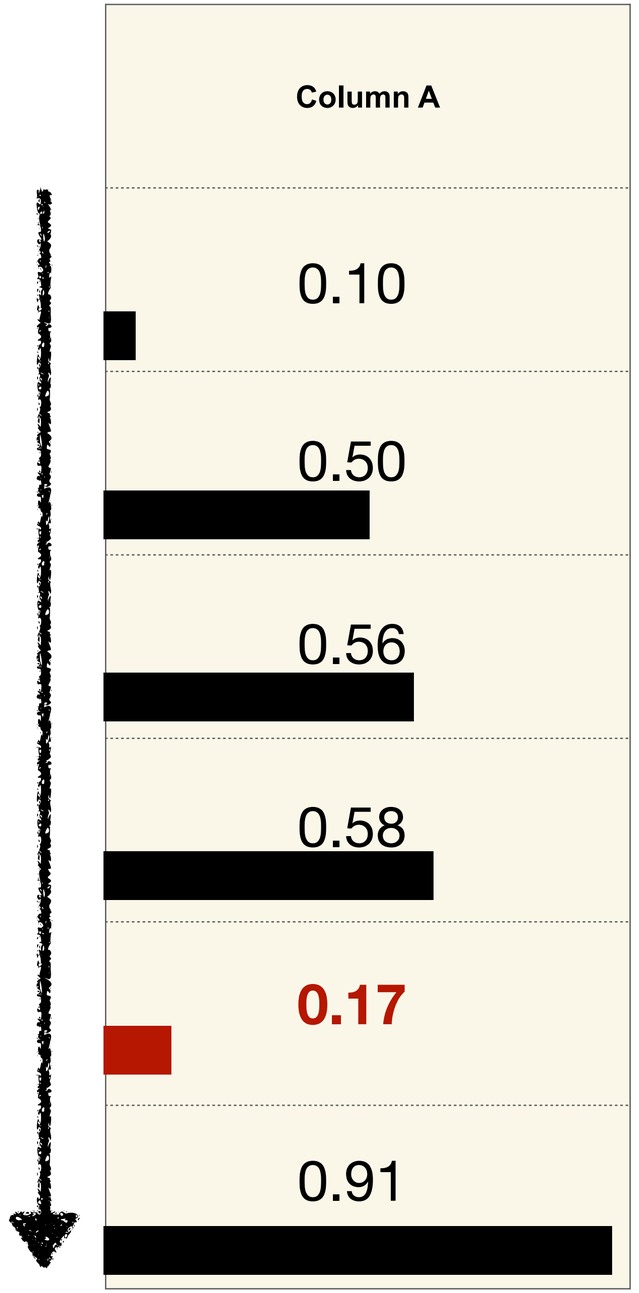
Ordering Artifacts



Ordering Artifacts



Ordering Artifacts

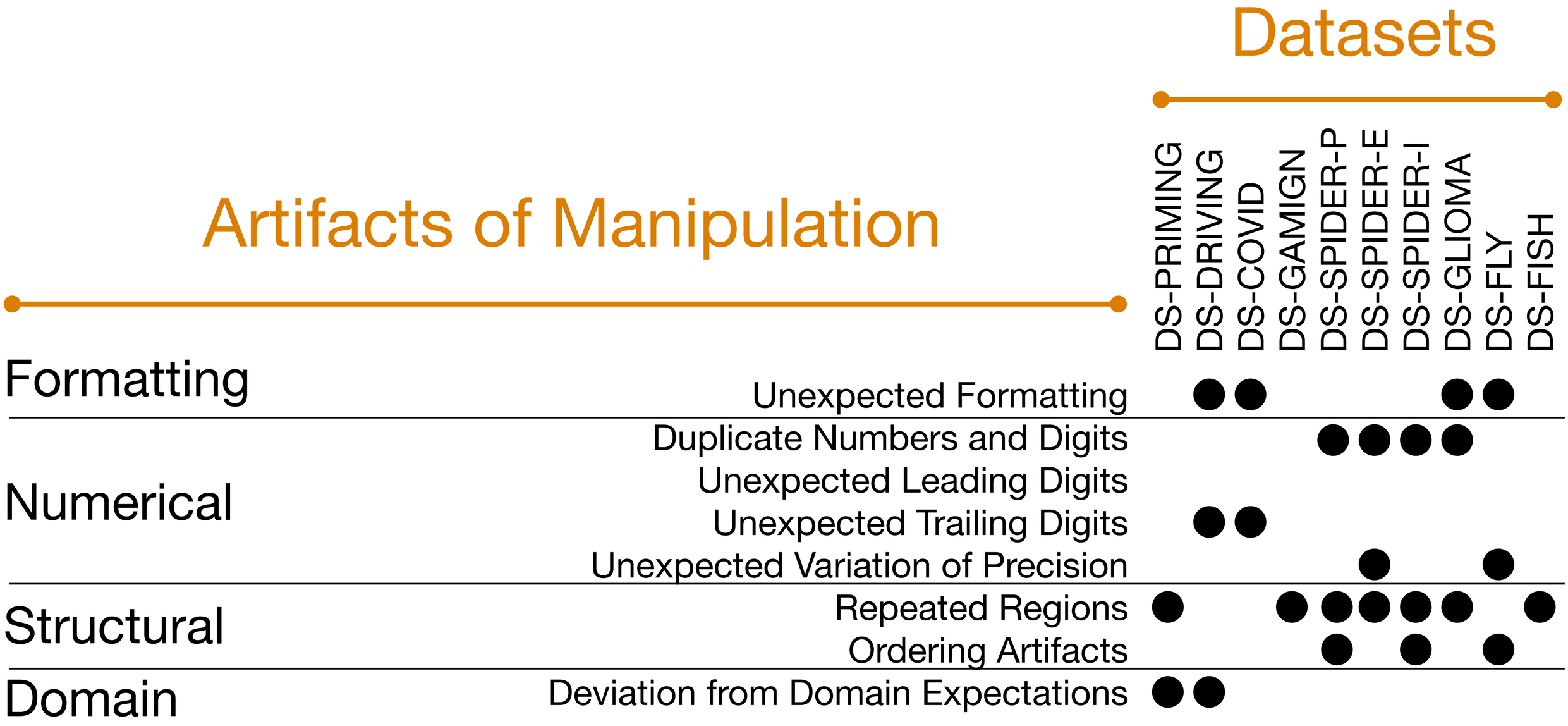


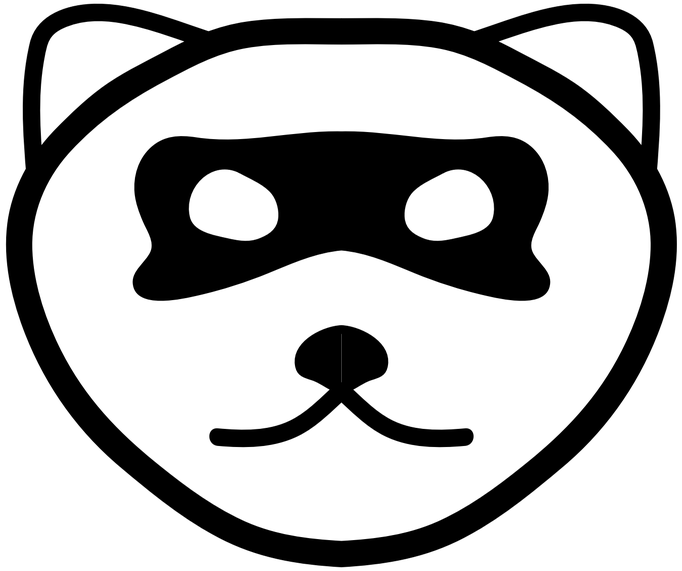
Deviation from Domain Expectations

Height (cm)
1510
155
182
212

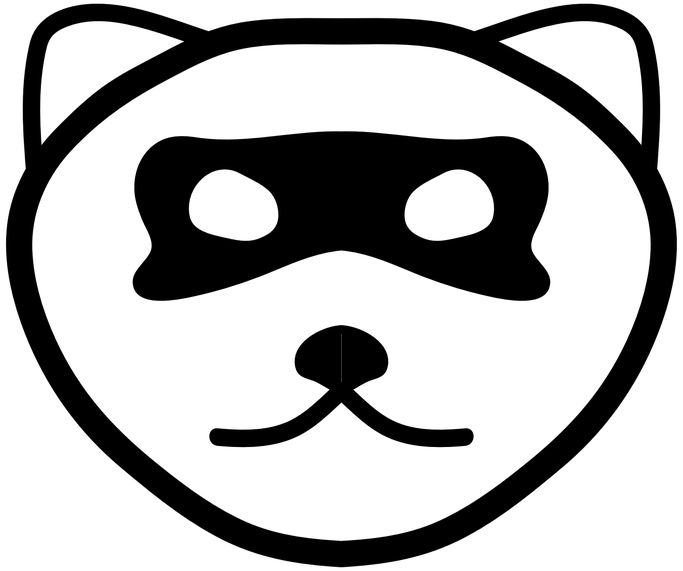
Deviation from Domain Expectations

Height (cm)
1510
155
182
212

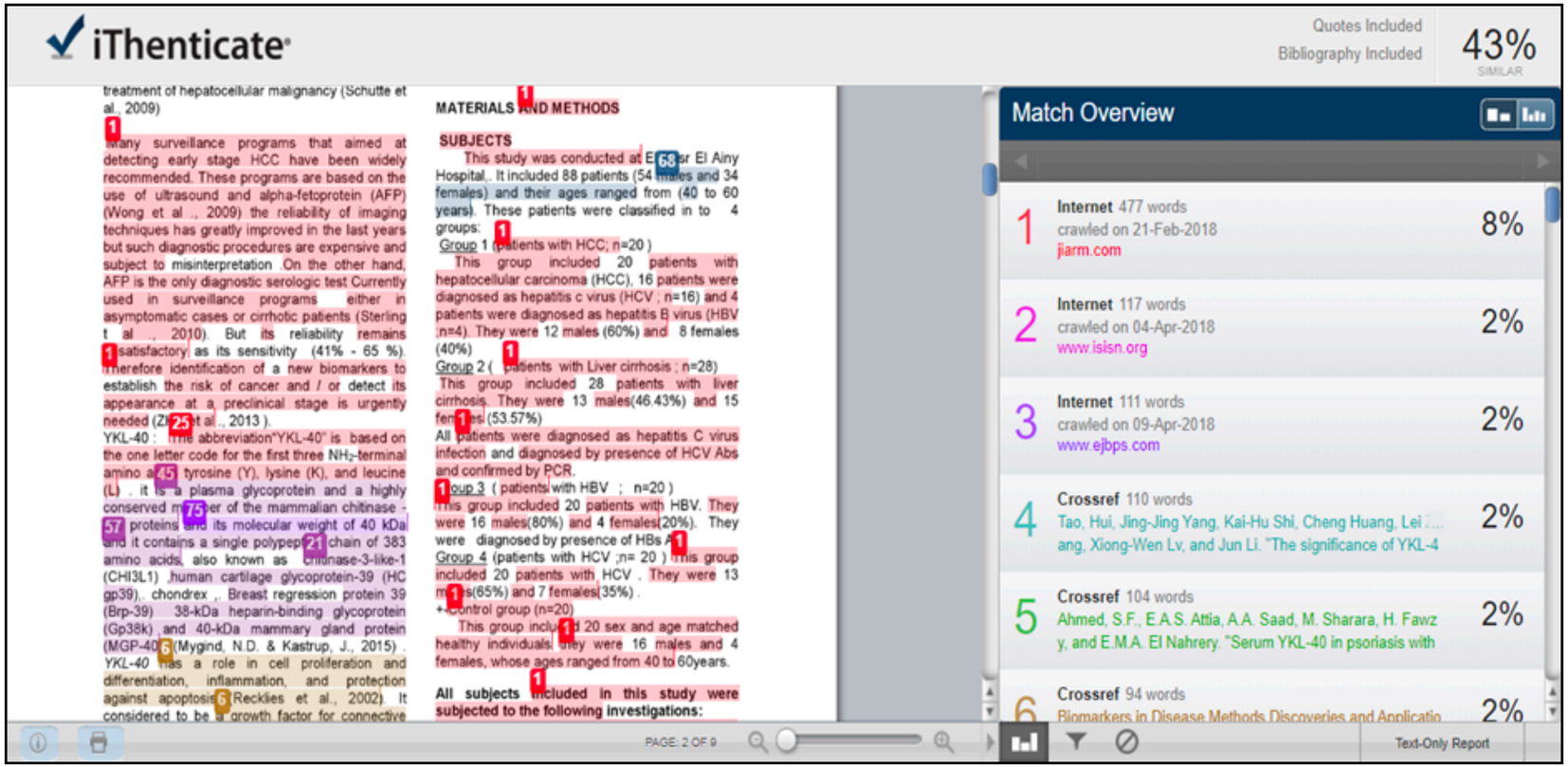

Datasets
DS-PRIMING
DS-DRIVING
DS-COVID
DS-GAMIGN
DS-SPIDER-P
DS-SPIDER-E
DS-SPIDER-I
DS-GLIOMA
DS-FLY
DS-FISH
Artifacts of Manipulation
Formatting
Unexpected Formatting
Numerical
Duplicate Numbers and Digits
Unexpected Leading Digits
Unexpected Trailing Digits
Unexpected Variation of Precision
Structural
Repeated Regions
Ordering Artifacts
Domain
Deviation from Domain Expectations



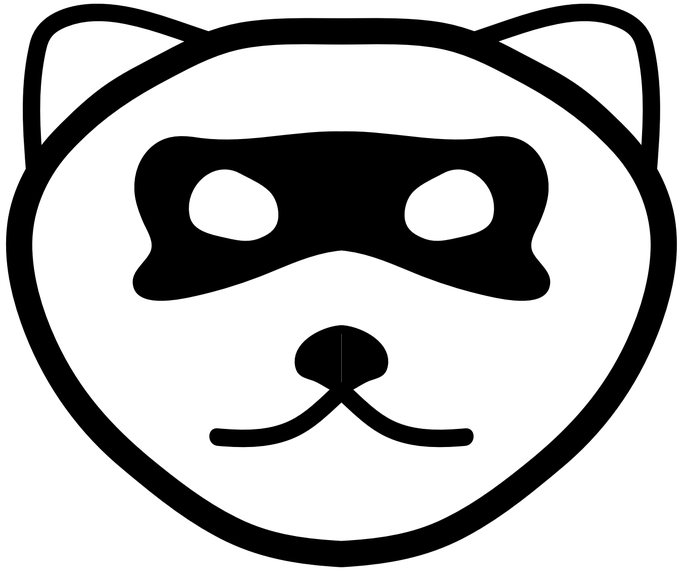
Create Tools

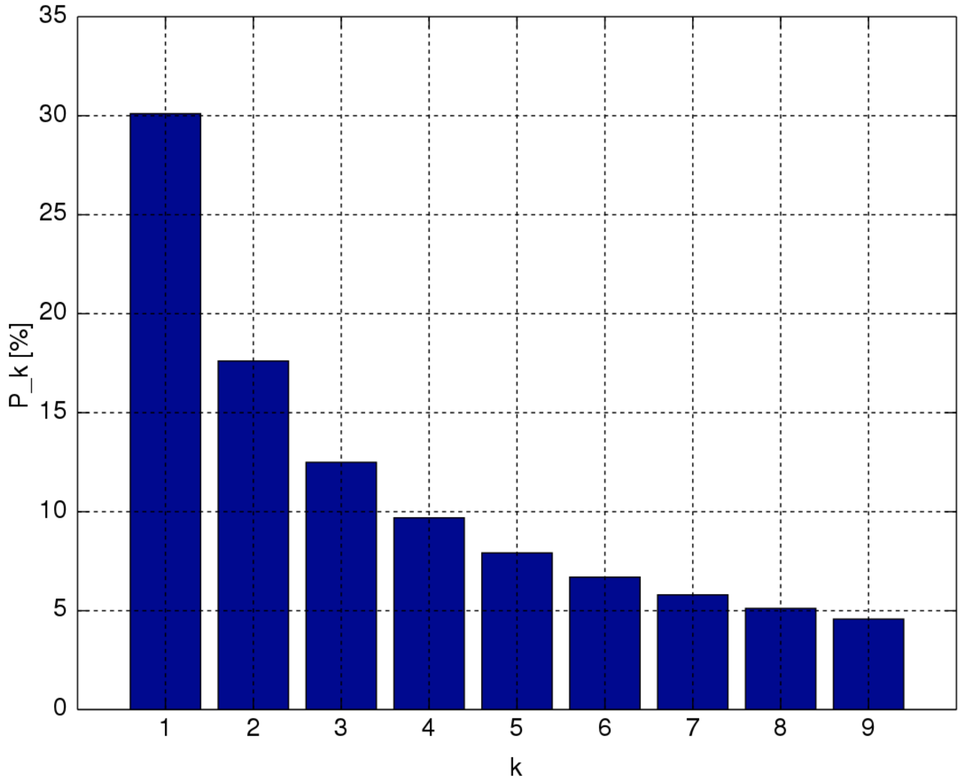


Create Tools



Create Tools





Ferret Overview

Home | Upload | Settings | Descriptions

Select Analysis

Previous | Next

- Formatting
- Value Distribution
- Duplicate Numbers
- Show all unique values
- Replicates
- Duplicate Digits
- Leading Digits
- Trailing Digits
- Precision
- Structural Overview
- Check Domain Expectations

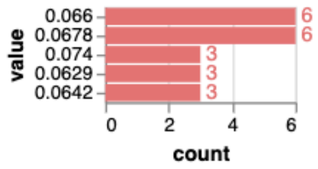
Ignore (1)

Highlight (1)

Summary Stats

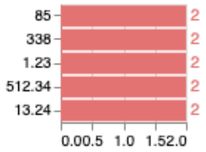
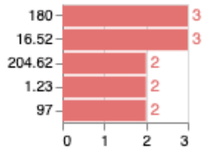
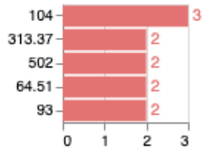
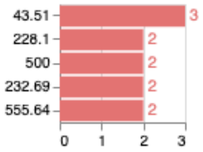
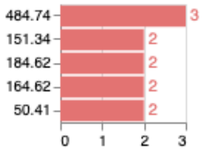
Select Columns

Duplicate Numbers



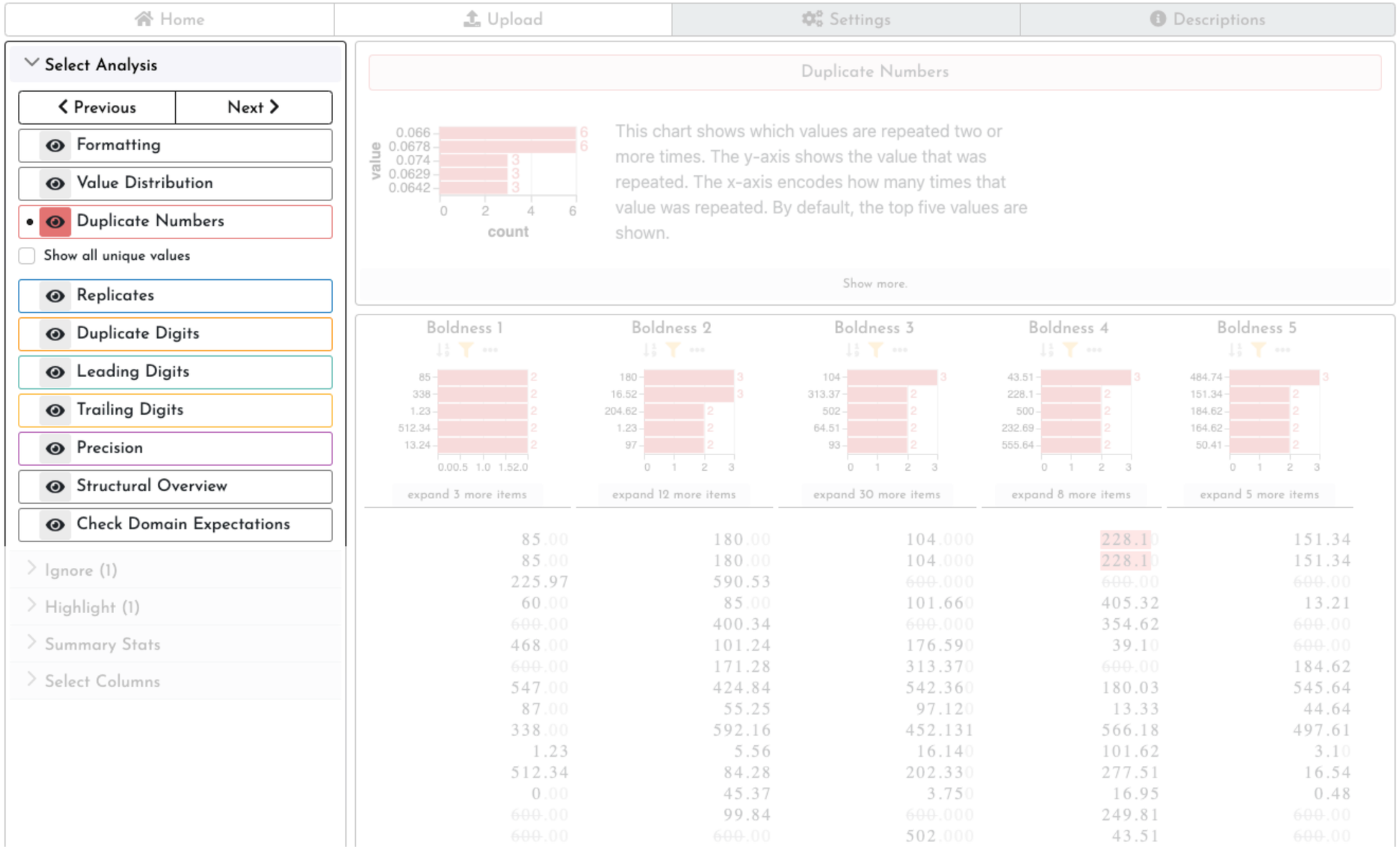
This chart shows which values are repeated two or more times. The y-axis shows the value that was repeated. The x-axis encodes how many times that value was repeated. By default, the top five values are shown.

Show more.

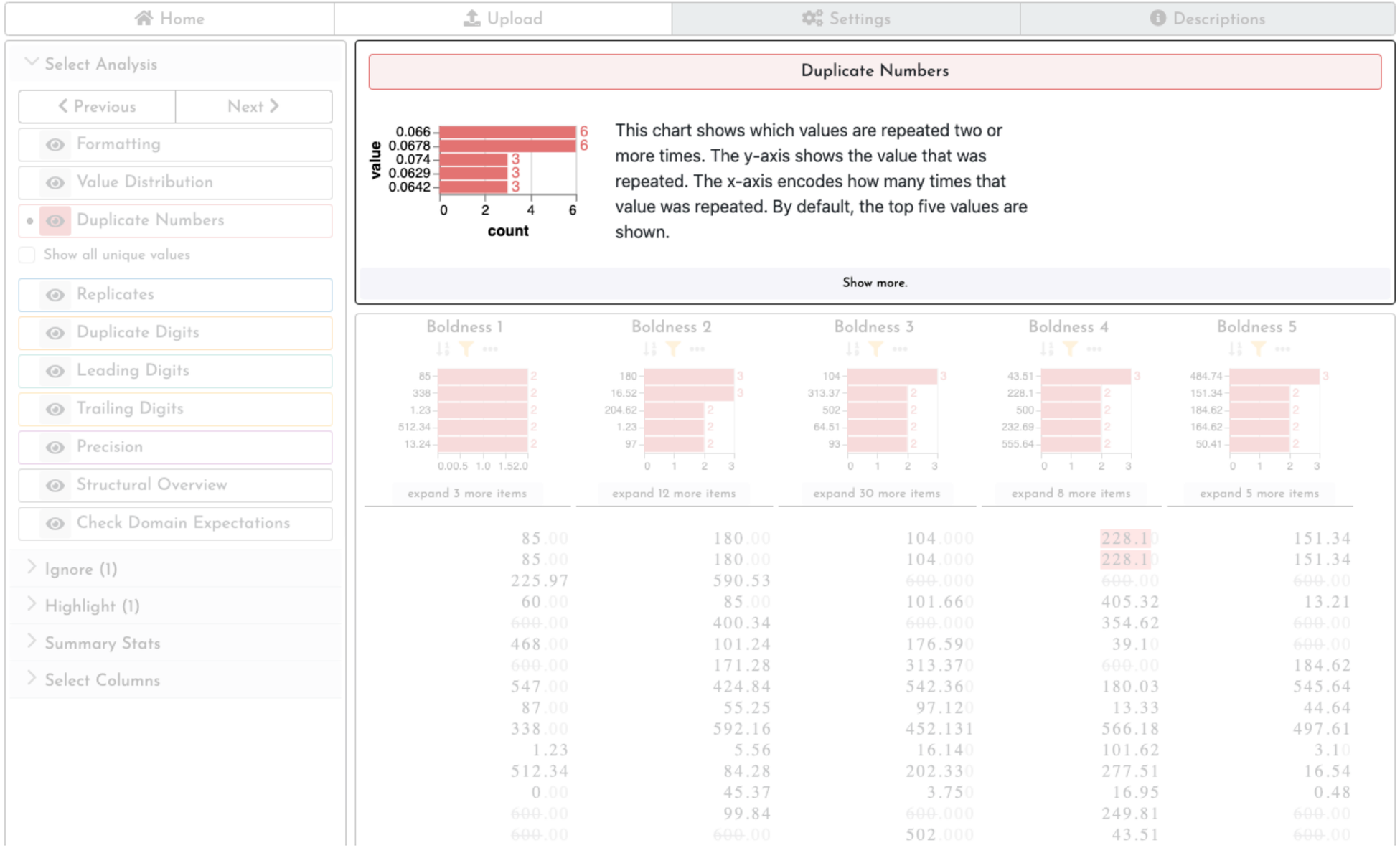
Boldness 1	Boldness 2	Boldness 3	Boldness 4	Boldness 5
				
expand 3 more items	expand 12 more items	expand 30 more items	expand 8 more items	expand 5 more items
85.00	180.00	104.000	228.10	151.34
85.00	180.00	104.000	228.10	151.34
225.97	590.53	600.000	600.00	600.00
60.00	85.00	101.660	405.32	13.21
600.00	400.34	600.000	354.62	600.00
468.00	101.24	176.590	39.10	600.00
600.00	171.28	313.370	600.00	184.62
547.00	424.84	542.360	180.03	545.64
87.00	55.25	97.120	13.33	44.64
338.00	592.16	452.131	566.18	497.61
1.23	5.56	16.140	101.62	3.10
512.34	84.28	202.330	277.51	16.54
0.00	45.37	3.750	16.95	0.48
600.00	99.84	600.000	249.81	600.00
600.00	600.00	502.000	43.51	600.00

Ferret Overview

Analysis Selection



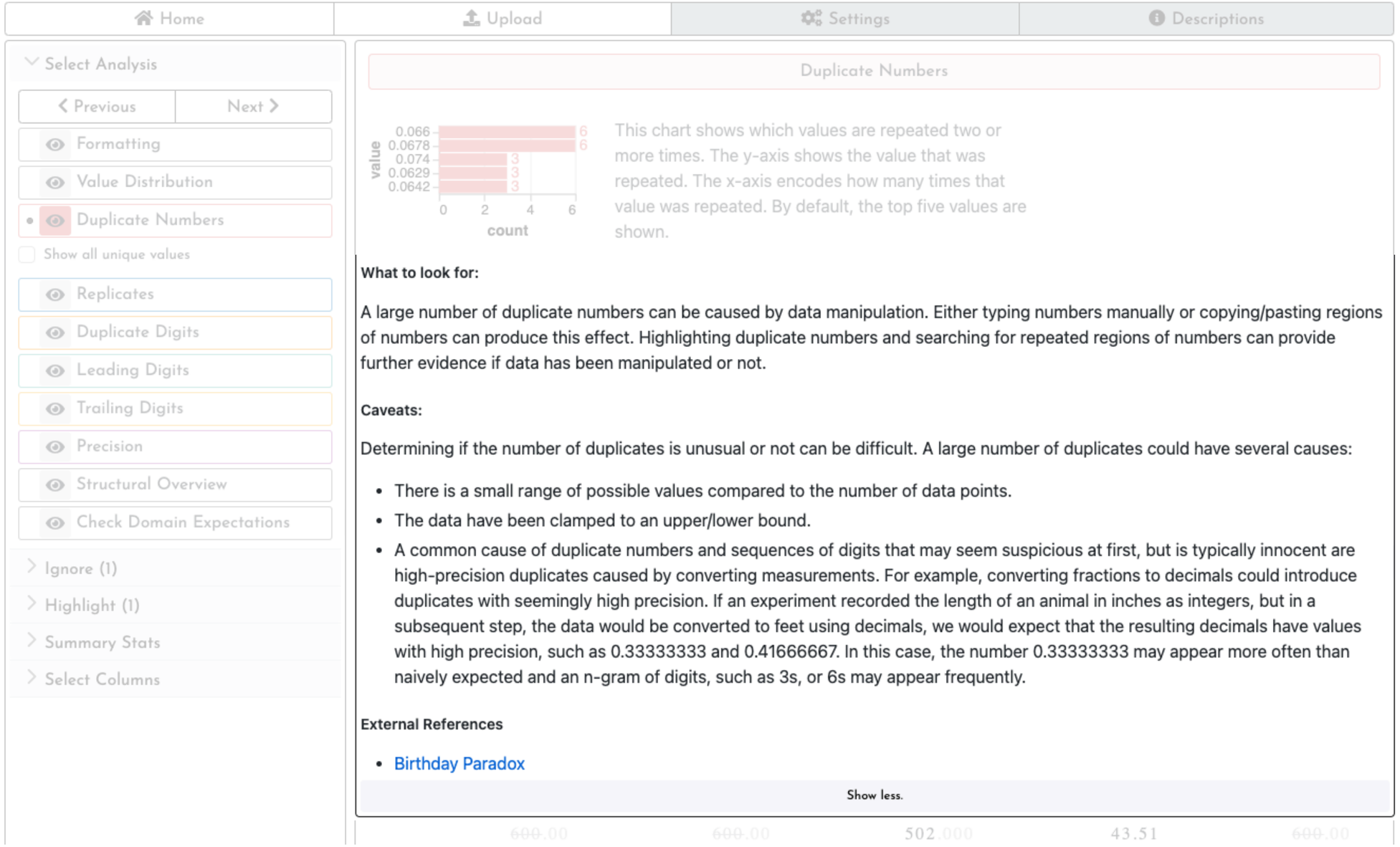
Ferret Overview



Analysis Selection

Analysis Explanation

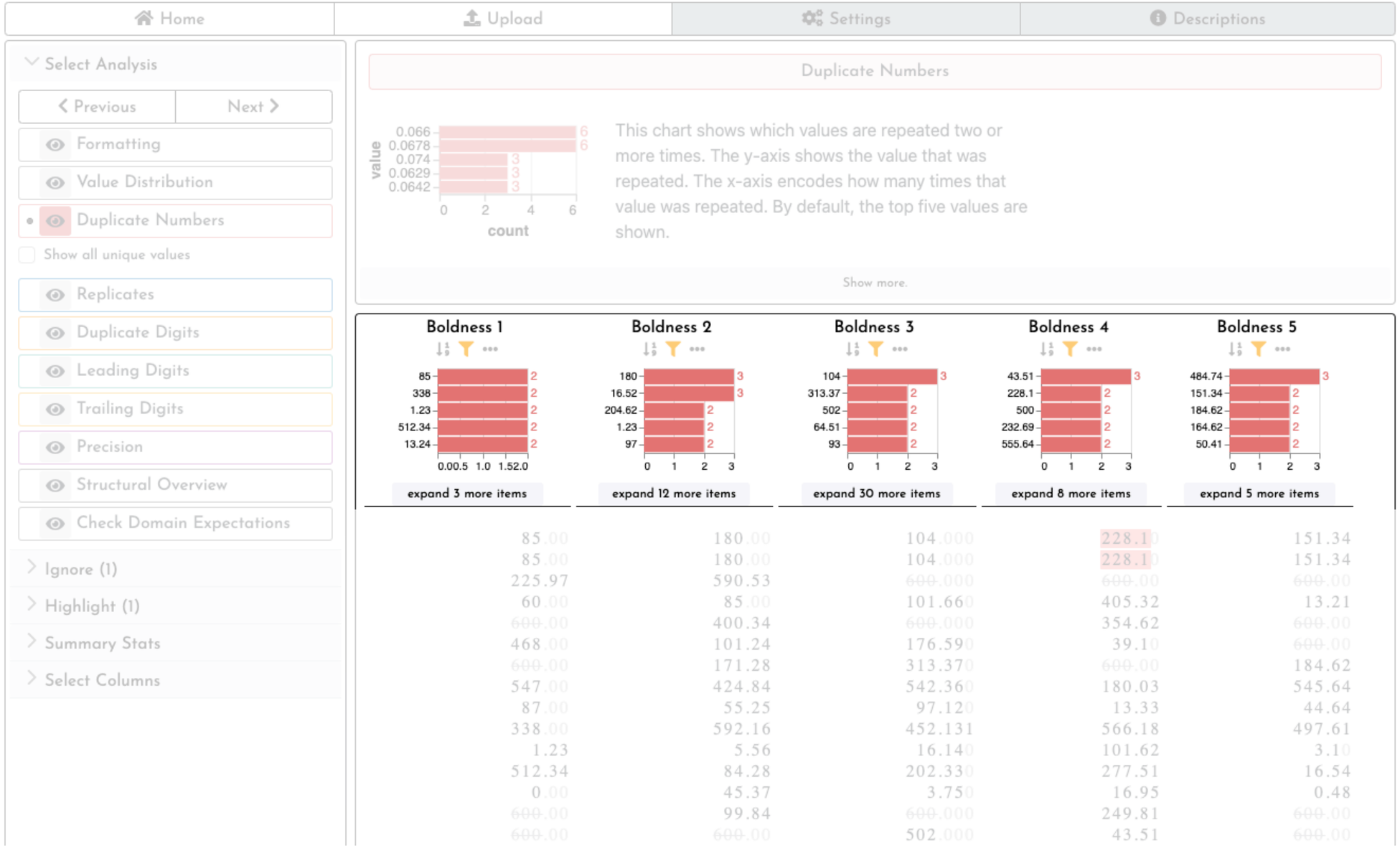
Ferret Overview



Analysis Selection

Analysis Explanation

Ferret Overview

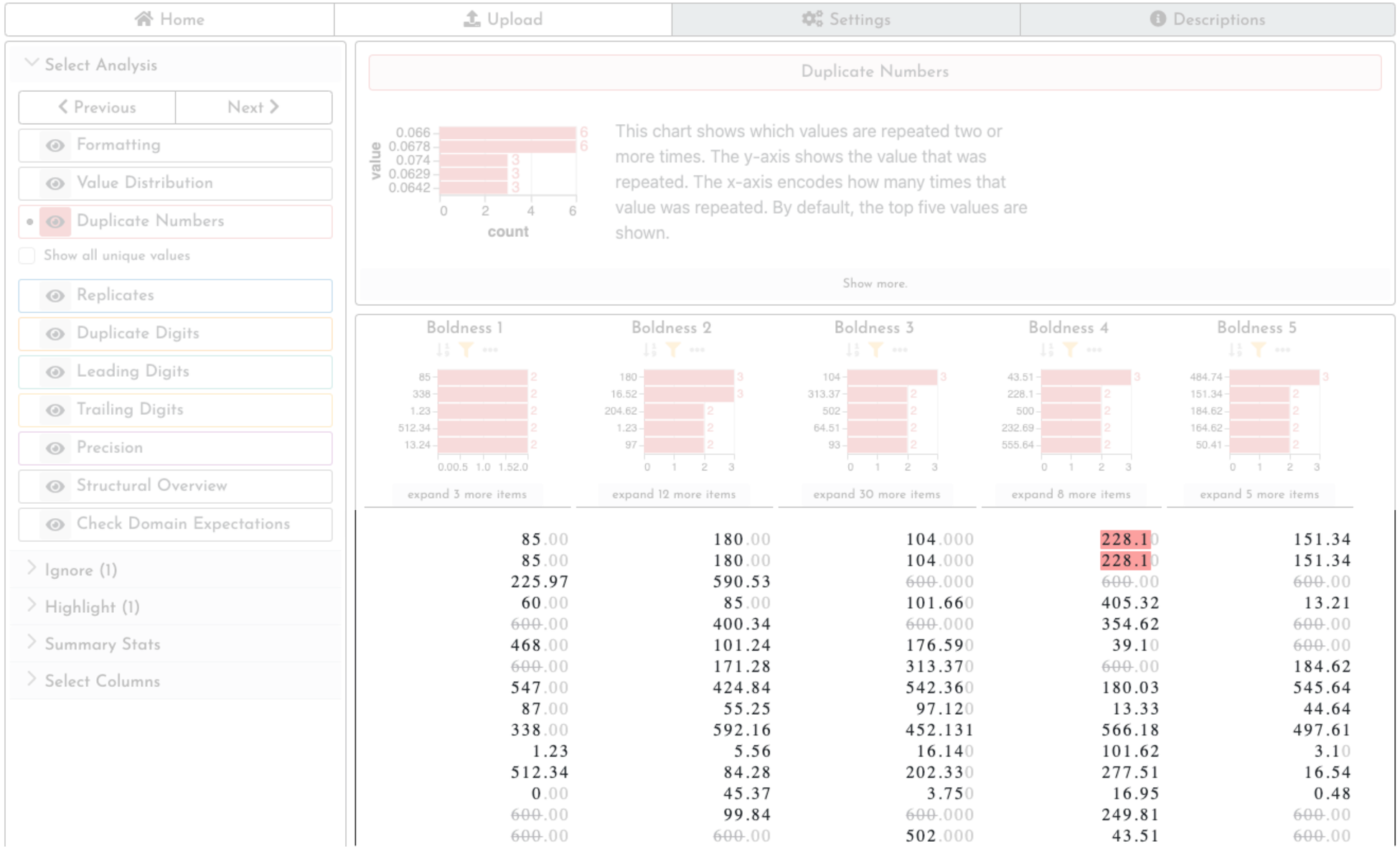


Analysis Selection

Analysis Explanation

Summary Charts

Ferret Overview



Analysis Selection

Analysis Explanation

Summary Charts

Tabular Visualization

Ferret Overview

Home | Upload | Settings | Descriptions

Analysis Selection

Select Analysis

Previous | Next

- Formatting
- Value Distribution
- Duplicate Numbers
- Show all unique values
- Replicates
- Duplicate Digits
- Leading Digits
- Trailing Digits
- Precision
- Structural Overview
- Check Domain Expectations

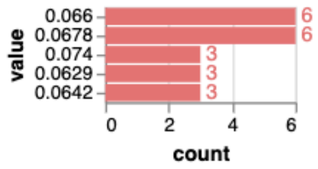
Ignore (1)

Highlight (1)

Summary Stats

Select Columns

Duplicate Numbers



This chart shows which values are repeated two or more times. The y-axis shows the value that was repeated. The x-axis encodes how many times that value was repeated. By default, the top five values are shown.

Show more.

Analysis Explanation

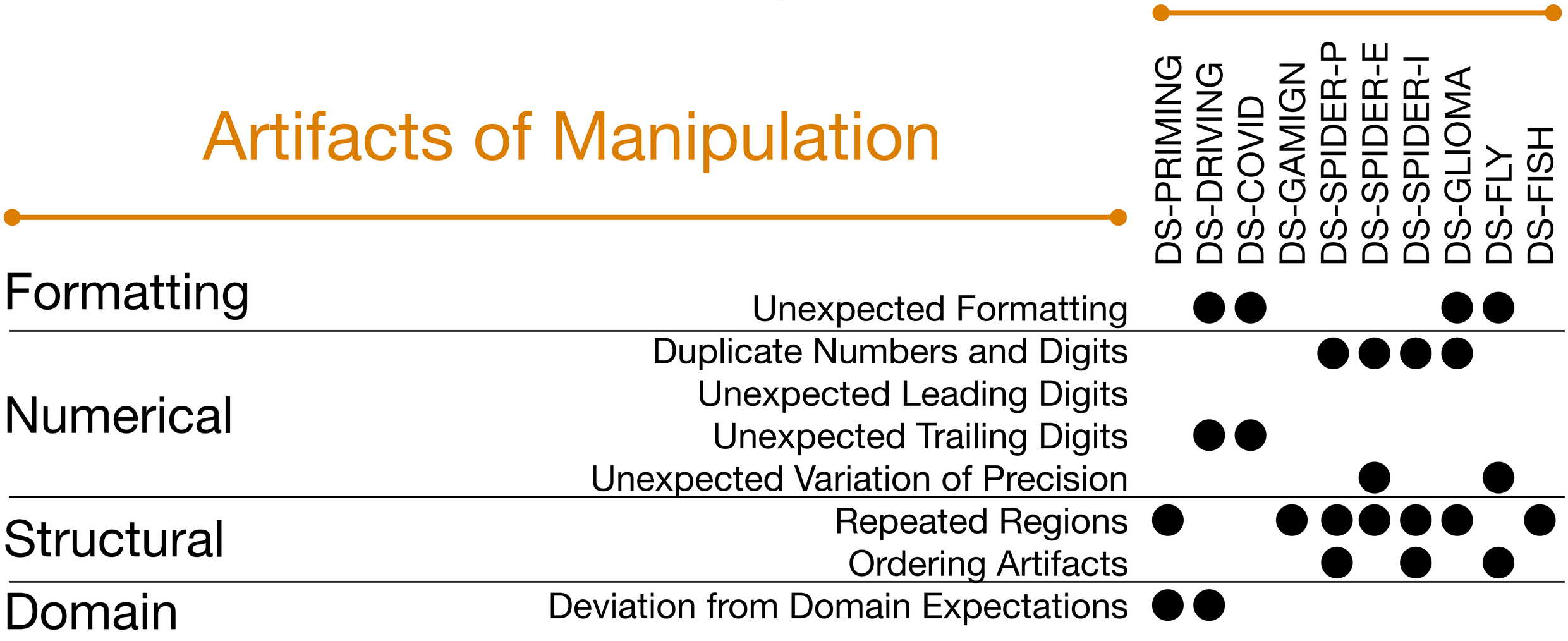
expand 3 more items | expand 12 more items | expand 30 more items | expand 8 more items | expand 5 more items

Summary Charts

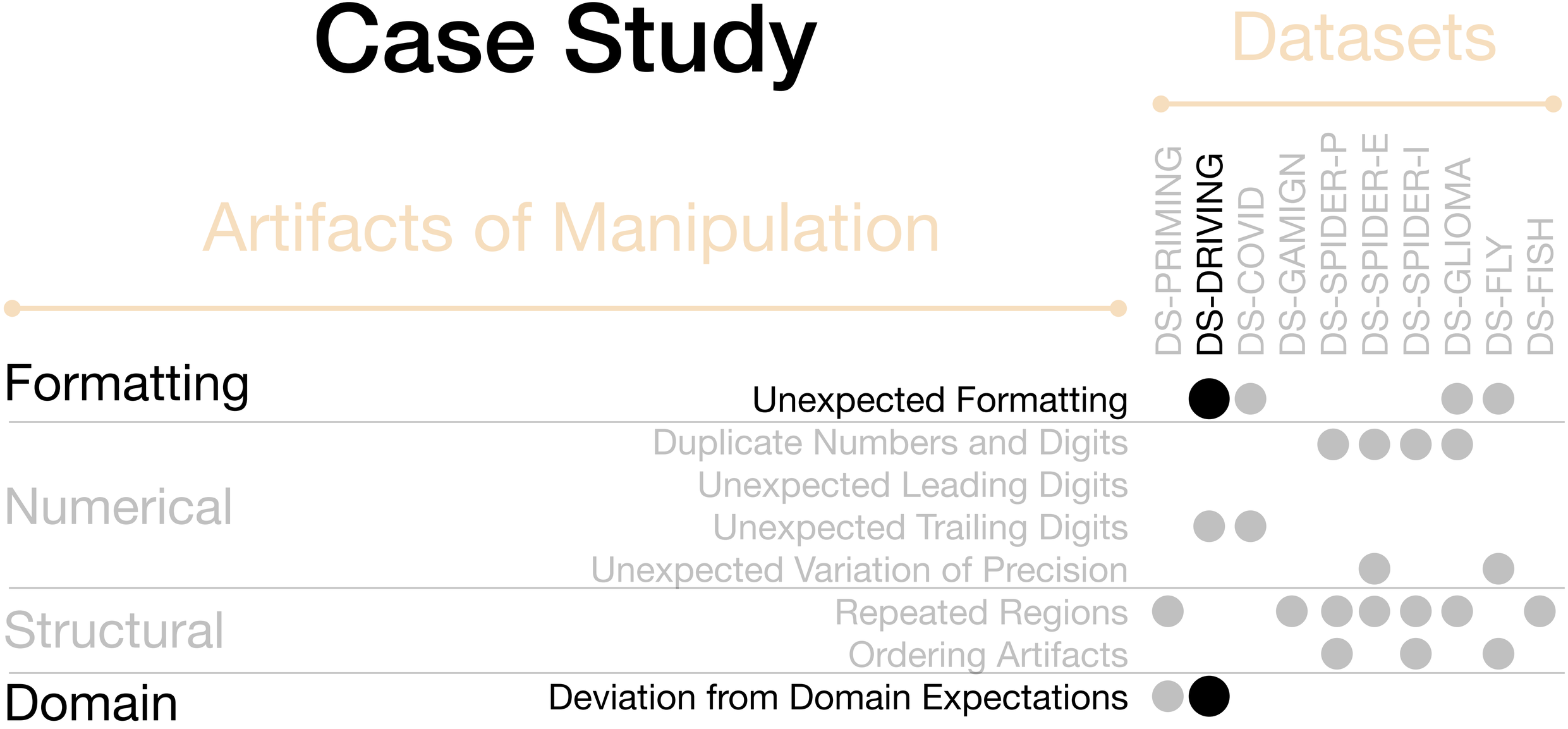
85.00	180.00	104.000	228.10	151.34
85.00	180.00	104.000	228.10	151.34
225.97	590.53	600.000	600.00	600.00
60.00	85.00	101.660	405.32	13.21
600.00	400.34	600.000	354.62	600.00
468.00	101.24	176.590	39.10	600.00
600.00	171.28	313.370	600.00	184.62
547.00	424.84	542.360	180.03	545.64
87.00	55.25	97.120	13.33	44.64
338.00	592.16	452.131	566.18	497.61
1.23	5.56	16.140	101.62	3.10
512.34	84.28	202.330	277.51	16.54
0.00	45.37	3.750	16.95	0.48
600.00	99.84	600.000	249.81	600.00
600.00	600.00	502.000	43.51	600.00

Tabular Visualization

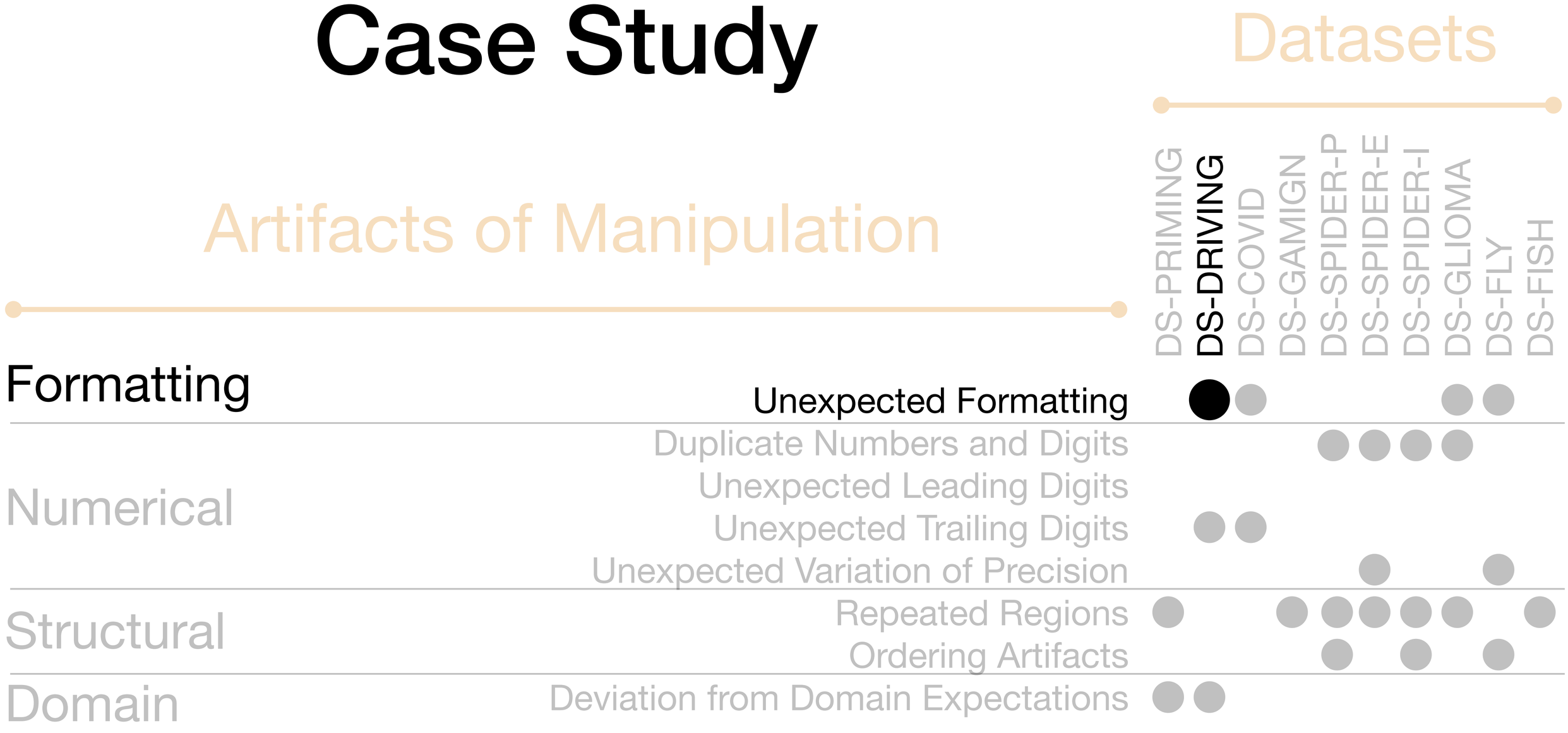
Case Study


Datasets
DS-PRIMING
DS-DRIVING
DS-COVID
DS-GAMIGN
DS-SPIDER-P
DS-SPIDER-E
DS-SPIDER-I
DS-GLIOMA
DS-FLY
DS-FISH
Artifacts of Manipulation
Formatting
Unexpected Formatting
Numerical
Duplicate Numbers and Digits
Unexpected Leading Digits
Unexpected Trailing Digits
Unexpected Variation of Precision
Structural
Repeated Regions
Ordering Artifacts
Domain
Deviation from Domain Expectations

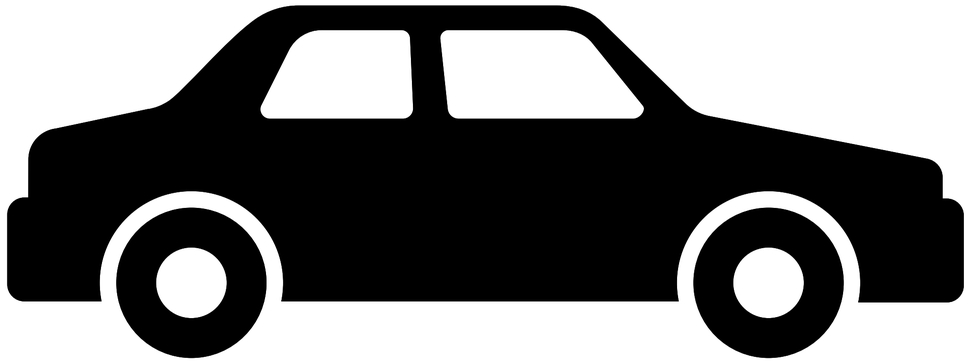
Case Study


Datasets
DS-PRIMING
DS-DRIVING
DS-COVID
DS-GAMIGN
DS-SPIDER-P
DS-SPIDER-E
DS-SPIDER-I
DS-GLIOMA
DS-FLY
DS-FISH
Artifacts of Manipulation
Formatting
Unexpected Formatting
Numerical
Duplicate Numbers and Digits
Unexpected Leading Digits
Unexpected Trailing Digits
Unexpected Variation of Precision
Structural
Repeated Regions
Ordering Artifacts
Domain
Deviation from Domain Expectations

Case Study


Datasets
DS-PRIMING
DS-DRIVING
DS-COVID
DS-GAMIGN
DS-SPIDER-P
DS-SPIDER-E
DS-SPIDER-I
DS-GLIOMA
DS-FLY
DS-FISH
Artifacts of Manipulation
Formatting
Unexpected Formatting
Numerical
Duplicate Numbers and Digits
Unexpected Leading Digits
Unexpected Trailing Digits
Unexpected Variation of Precision
Structural
Repeated Regions
Ordering Artifacts
Domain
Deviation from Domain Expectations

DS-DRIVING



Car Mileage T_1

DS-DRIVING

Time

Car Mileage T_1 → Car Mileage T_2

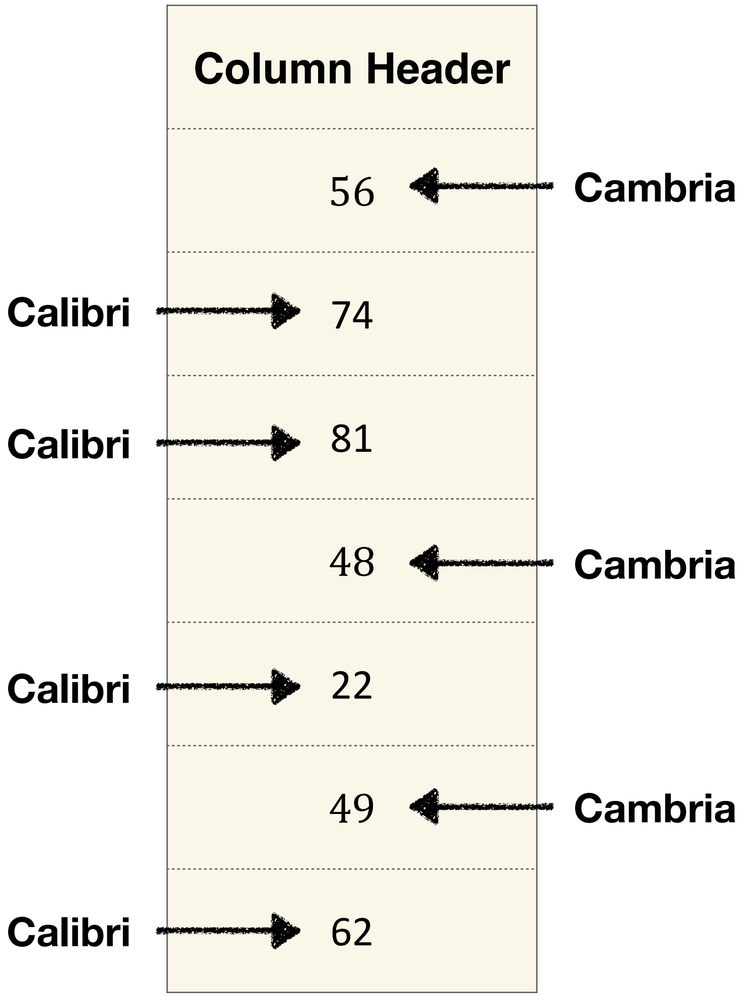
Unexpected Formatting

DS-DRIVING

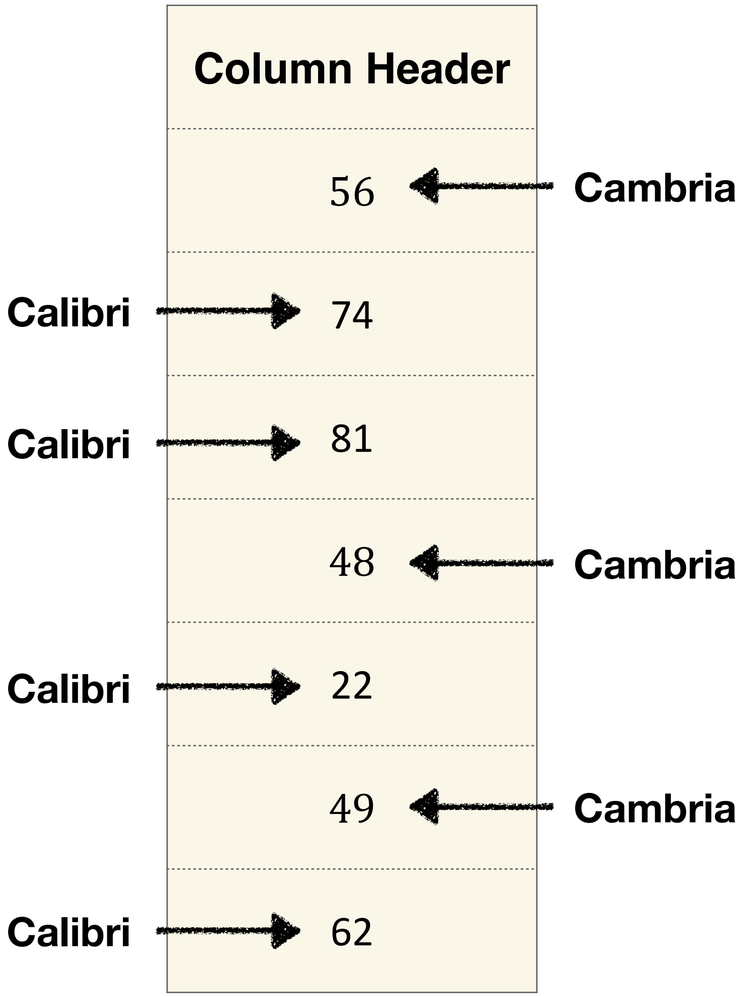
Column Header
56
74
81
48
22
49
62

Unexpected Formatting

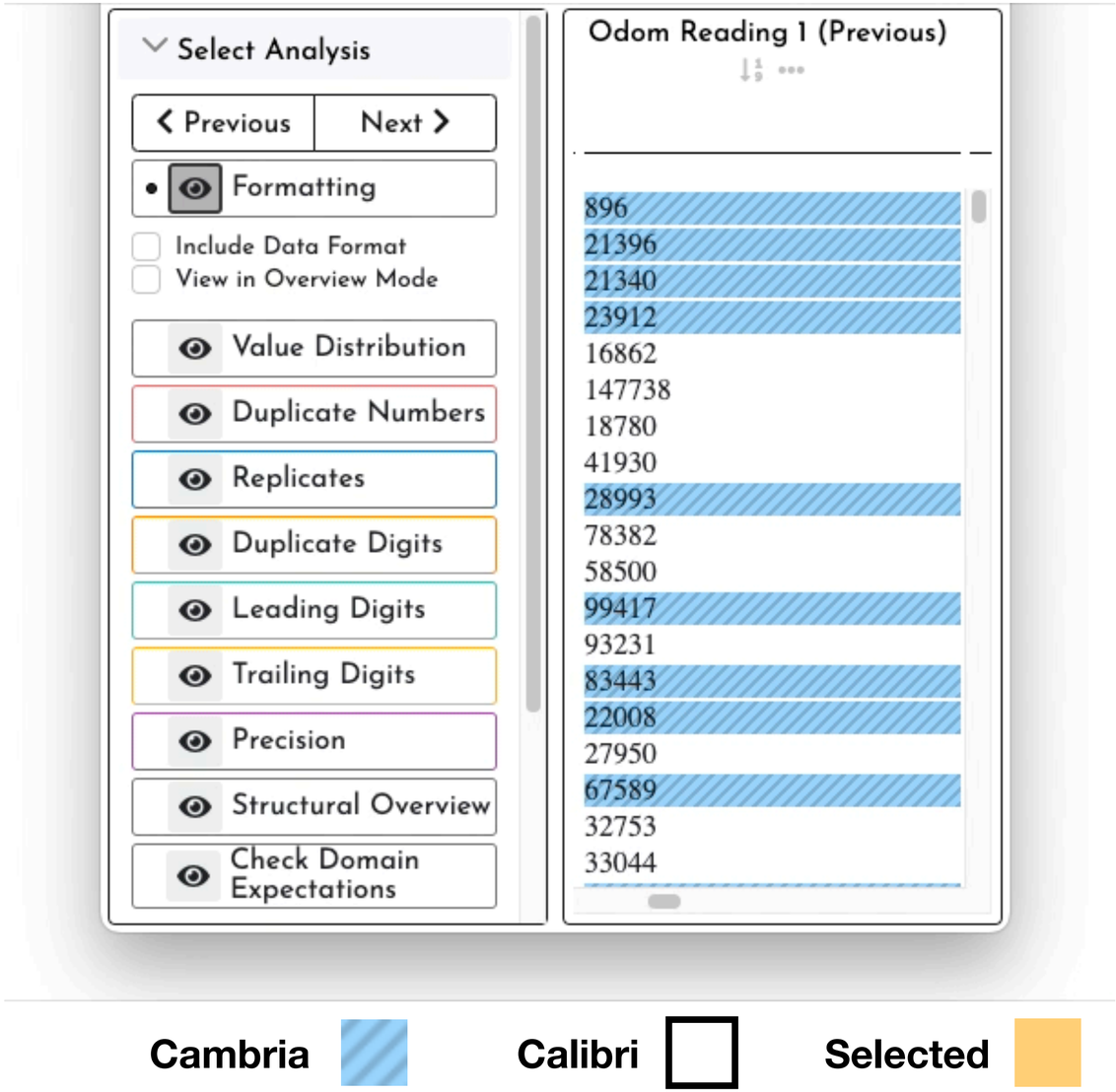
DS-DRIVING



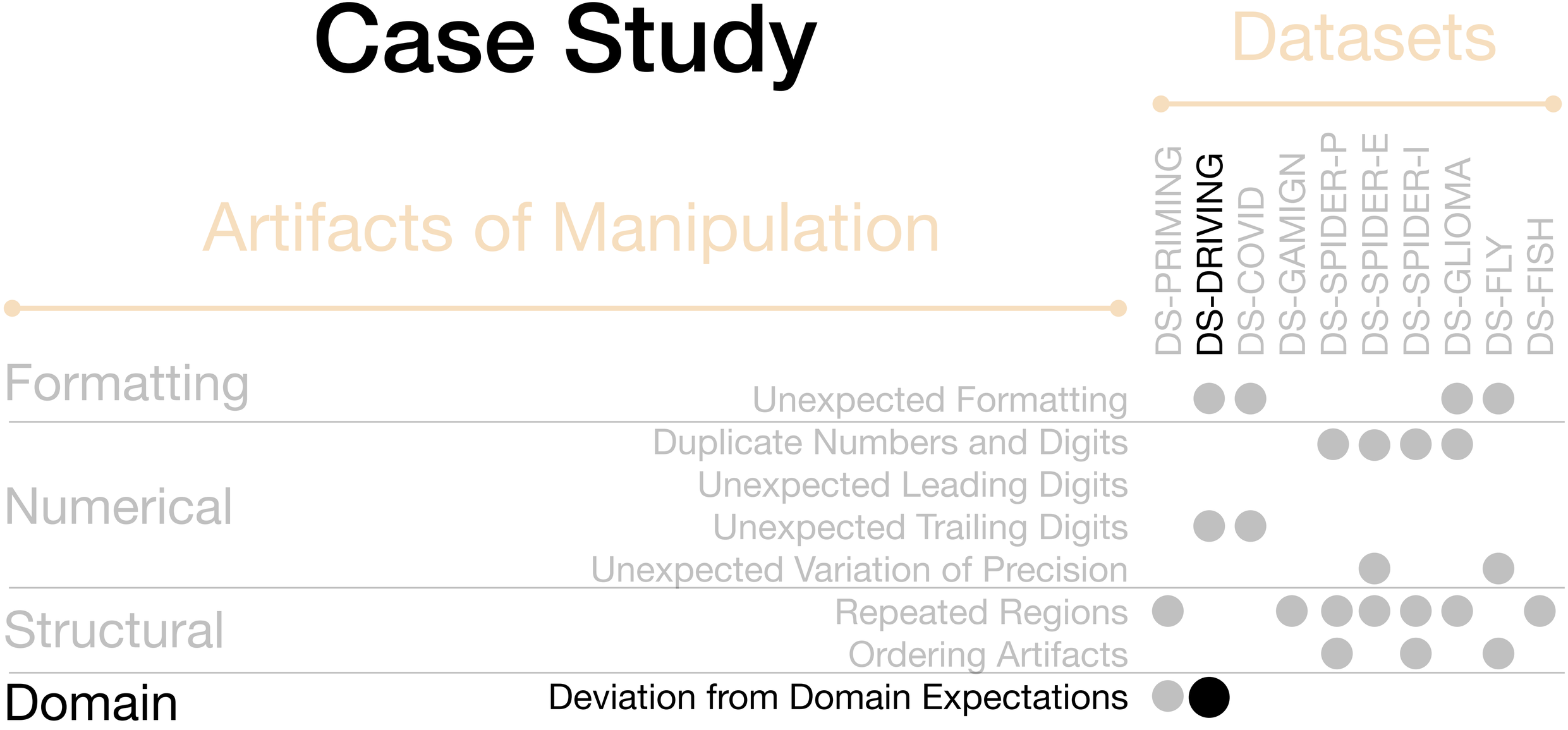
Unexpected Formatting



DS-DRIVING



Case Study


Datasets
DS-PRIMING
DS-DRIVING
DS-COVID
DS-GAMIGN
DS-SPIDER-P
DS-SPIDER-E
DS-SPIDER-I
DS-GLIOMA
DS-FLY
DS-FISH
Artifacts of Manipulation
Formatting
Unexpected Formatting
Numerical
Duplicate Numbers and Digits
Unexpected Leading Digits
Unexpected Trailing Digits
Unexpected Variation of Precision
Structural
Repeated Regions
Ordering Artifacts
Domain
Deviation from Domain Expectations

Deviation from Domain Expectations

DS-DRIVING

Deviation from Domain Expectations

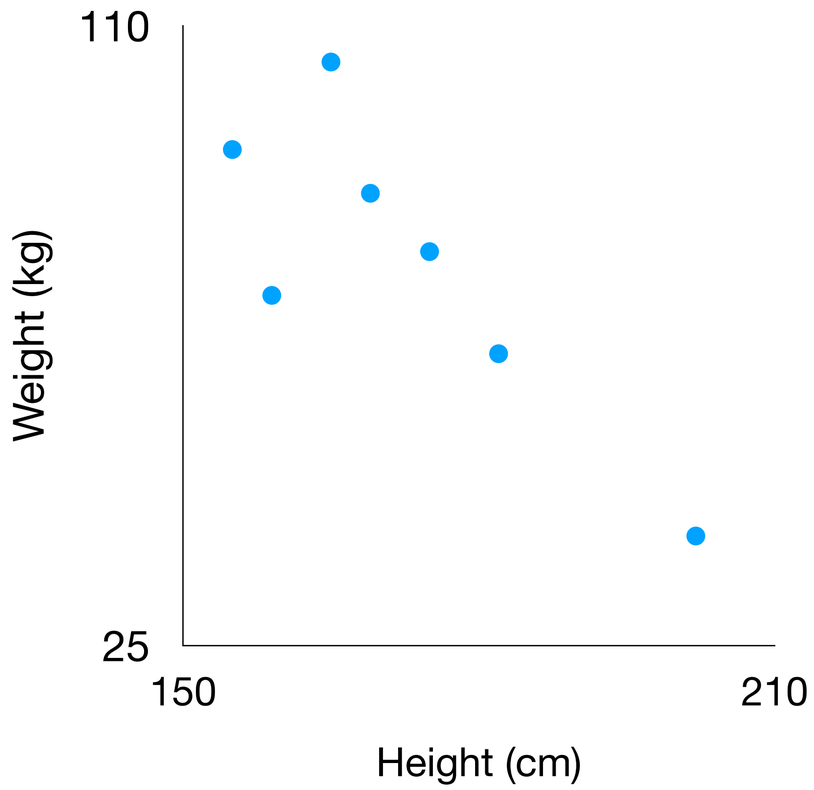
DS-DRIVING

Height (cm)	Weight (kg)
175	79
155	93
182	65
202	40
165	105
169	87
159	73

Deviation from Domain Expectations

DS-DRIVING

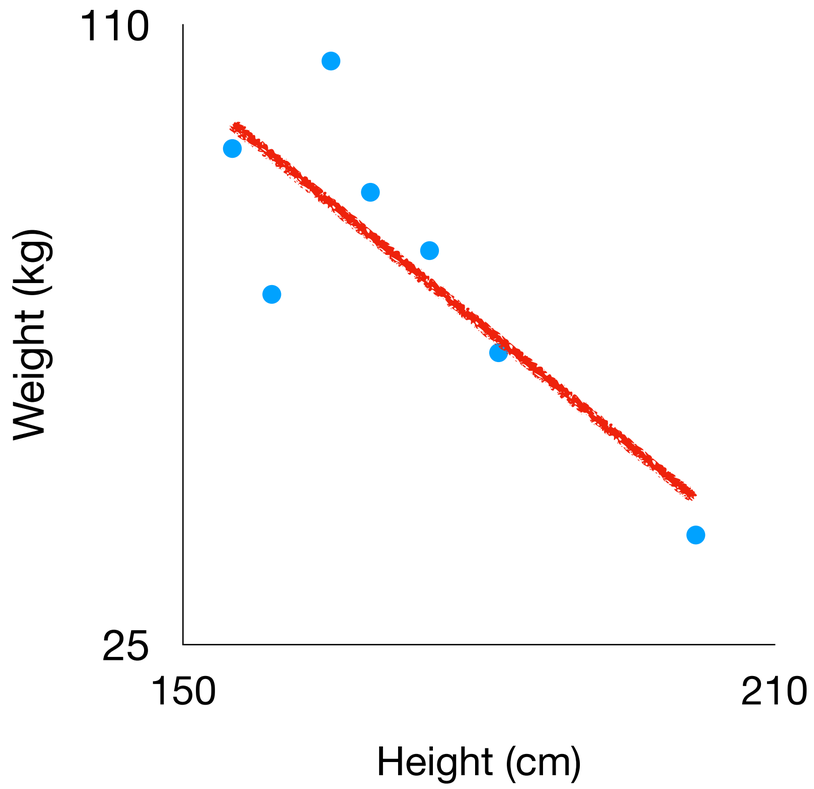
Height (cm)	Weight (kg)
175	79
155	93
182	65
202	40
165	105
169	87
159	73



Deviation from Domain Expectations

DS-DRIVING

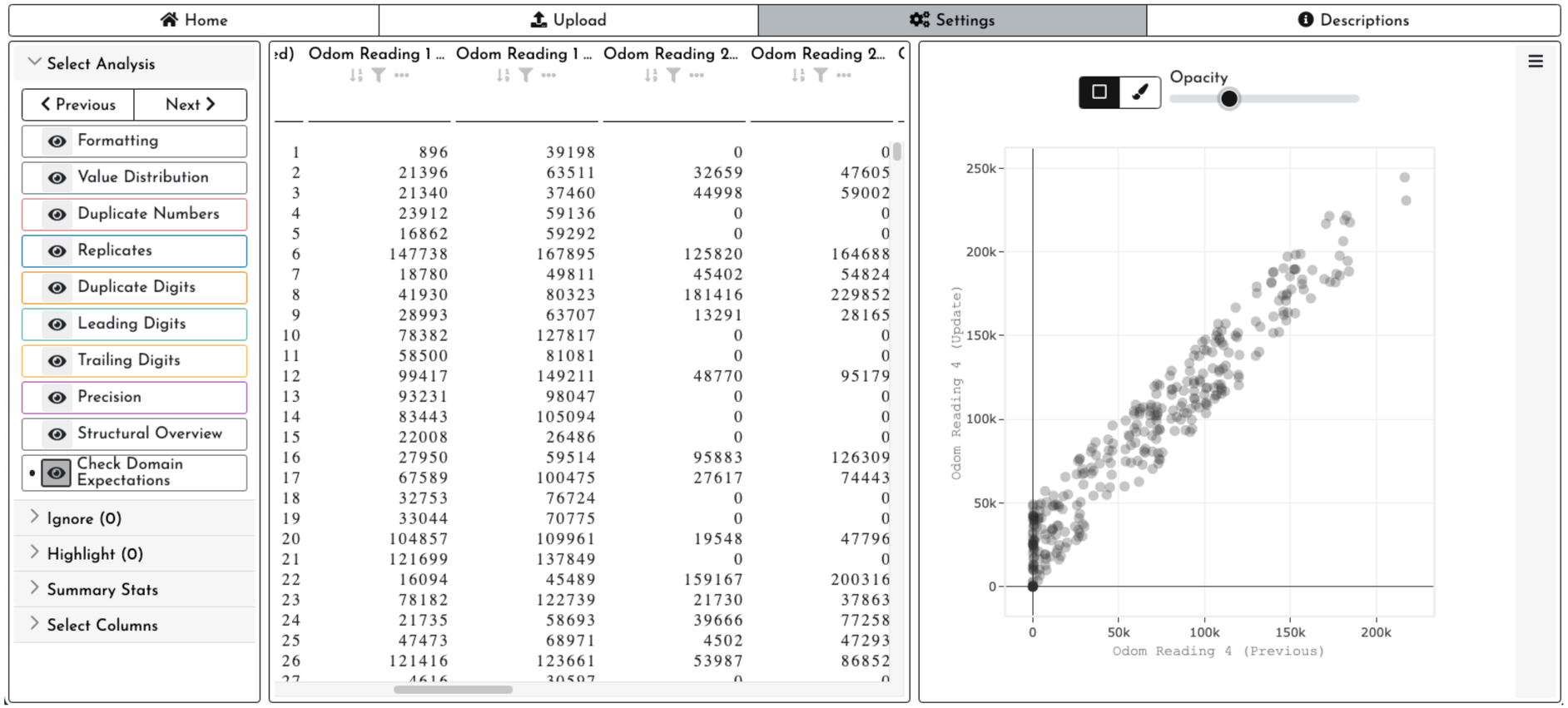
Height (cm)	Weight (kg)
175	79
155	93
182	65
202	40
165	105
169	87
159	73



Deviation from Domain Expectations

DS-DRIVING

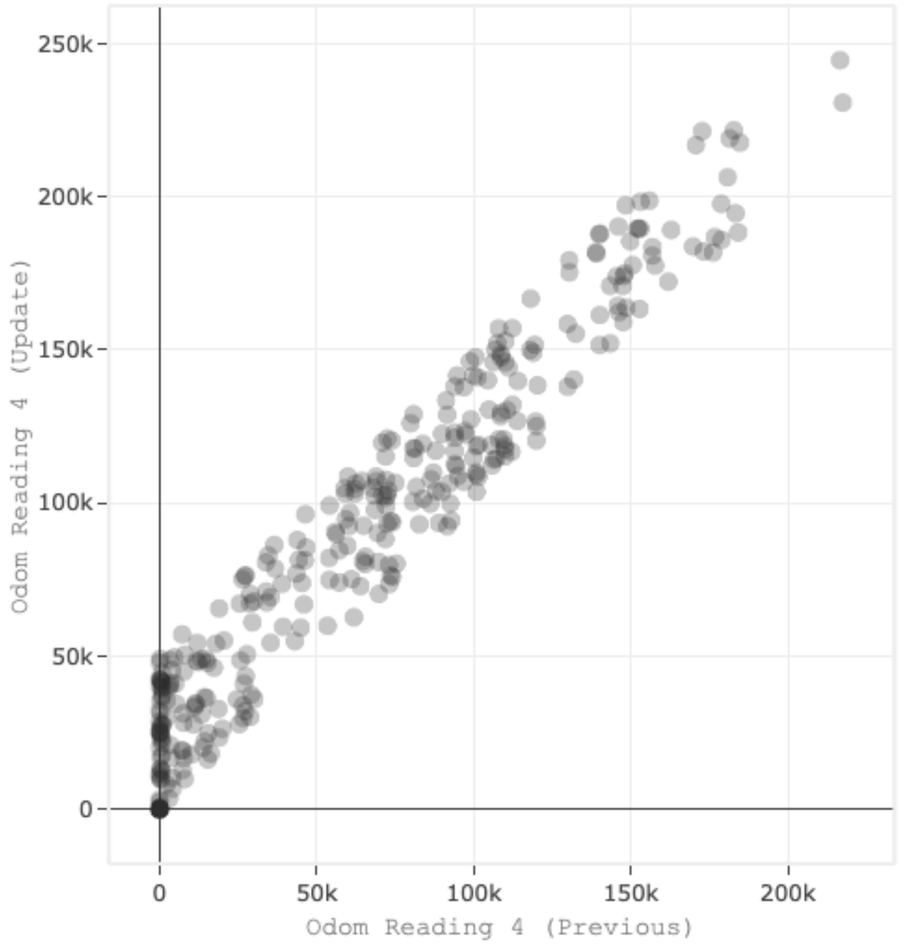
Height (cm)	Weight (kg)
175	79
155	93
182	65
202	40
165	105
169	87
159	73



Deviation from Domain Expectations

DS-DRIVING

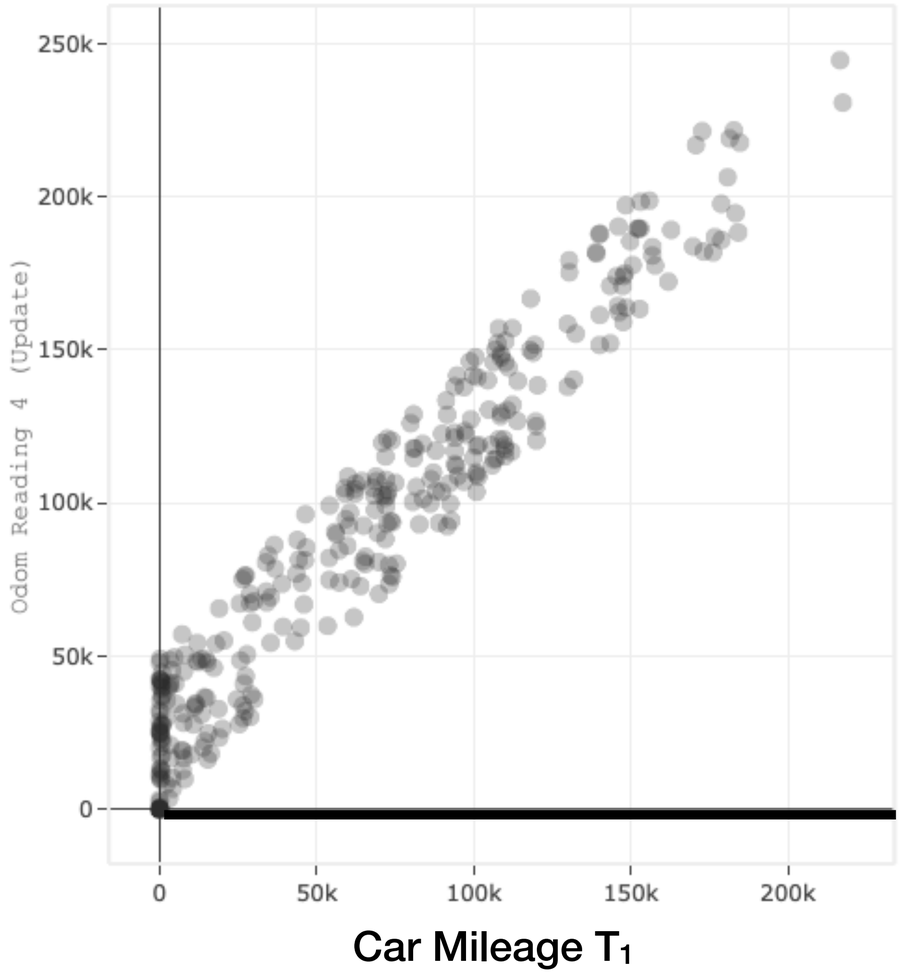
Height (cm)	Weight (kg)
175	79
155	93
182	65
202	40
165	105
169	87
159	73



Deviation from Domain Expectations

DS-DRIVING

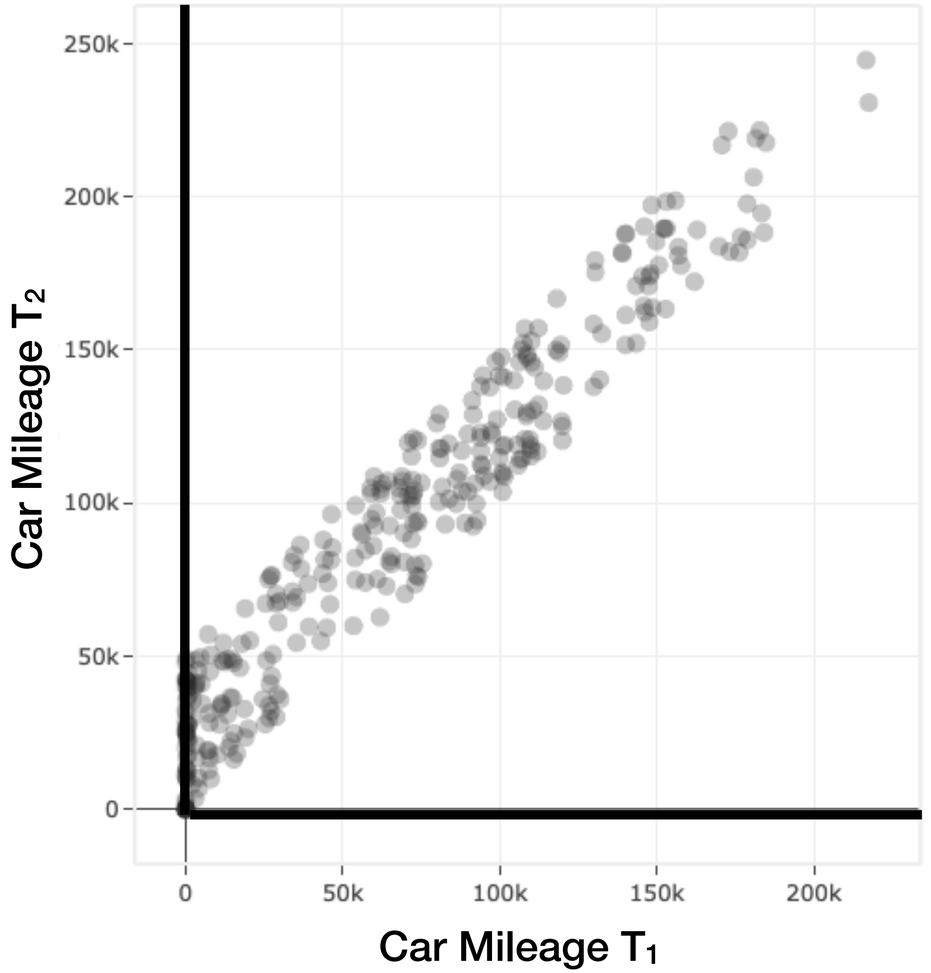
Height (cm)	Weight (kg)
175	79
155	93
182	65
202	40
165	105
169	87
159	73



Deviation from Domain Expectations

DS-DRIVING

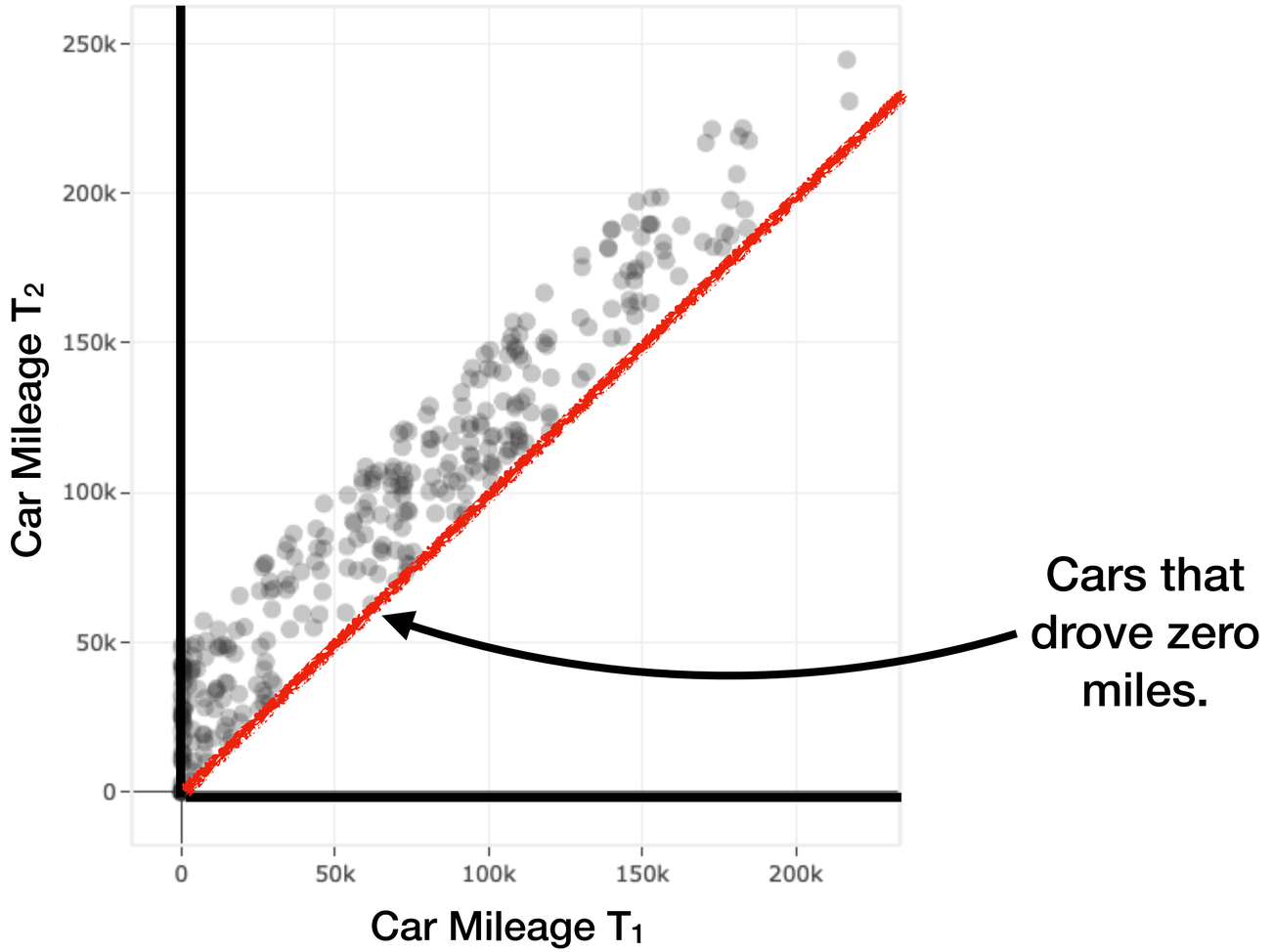
Height (cm)	Weight (kg)
175	79
155	93
182	65
202	40
165	105
169	87
159	73



Deviation from Domain Expectations

DS-DRIVING

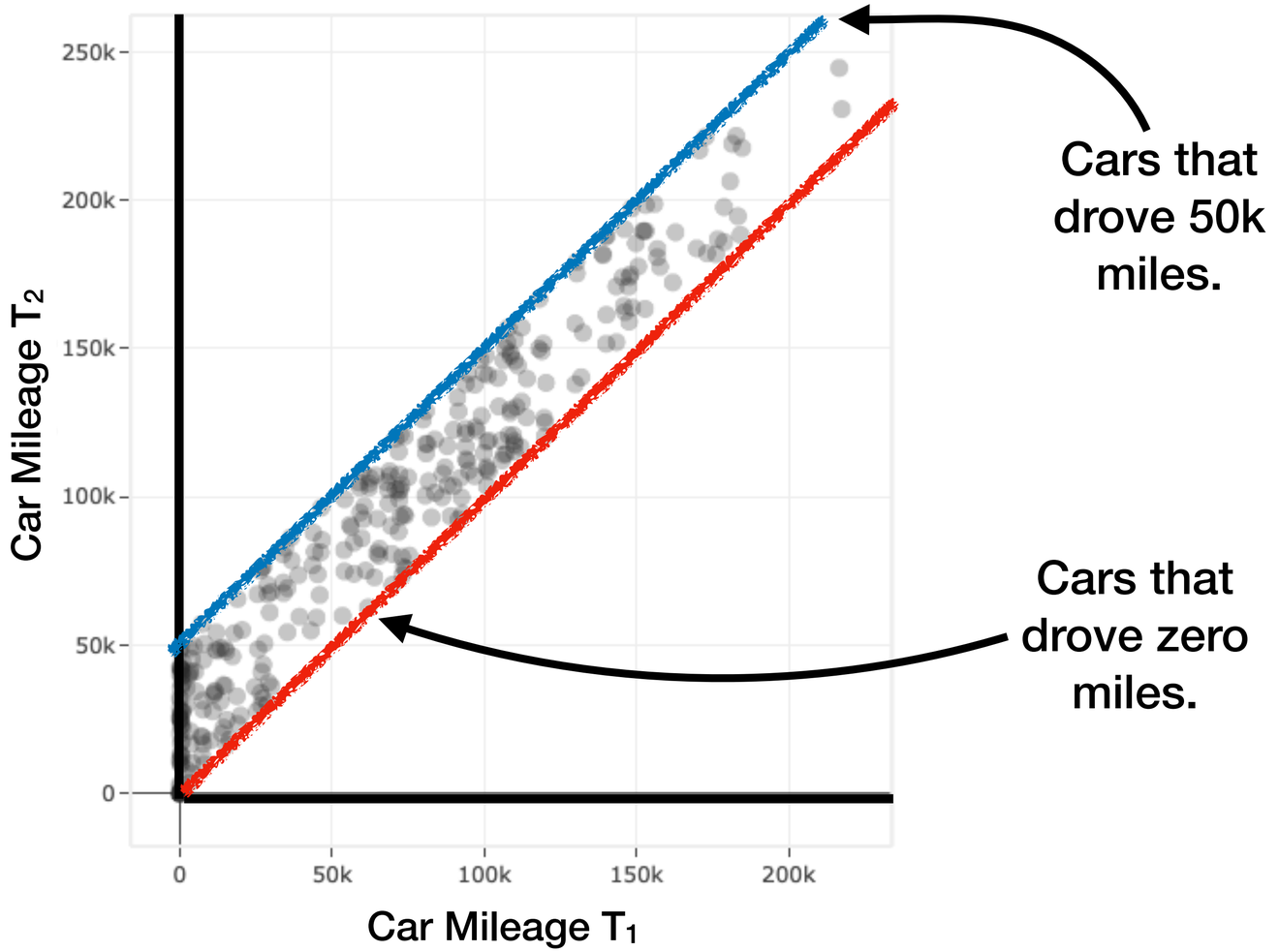
Height (cm)	Weight (kg)
175	79
155	93
182	65
202	40
165	105
169	87
159	73



Deviation from Domain Expectations

DS-DRIVING

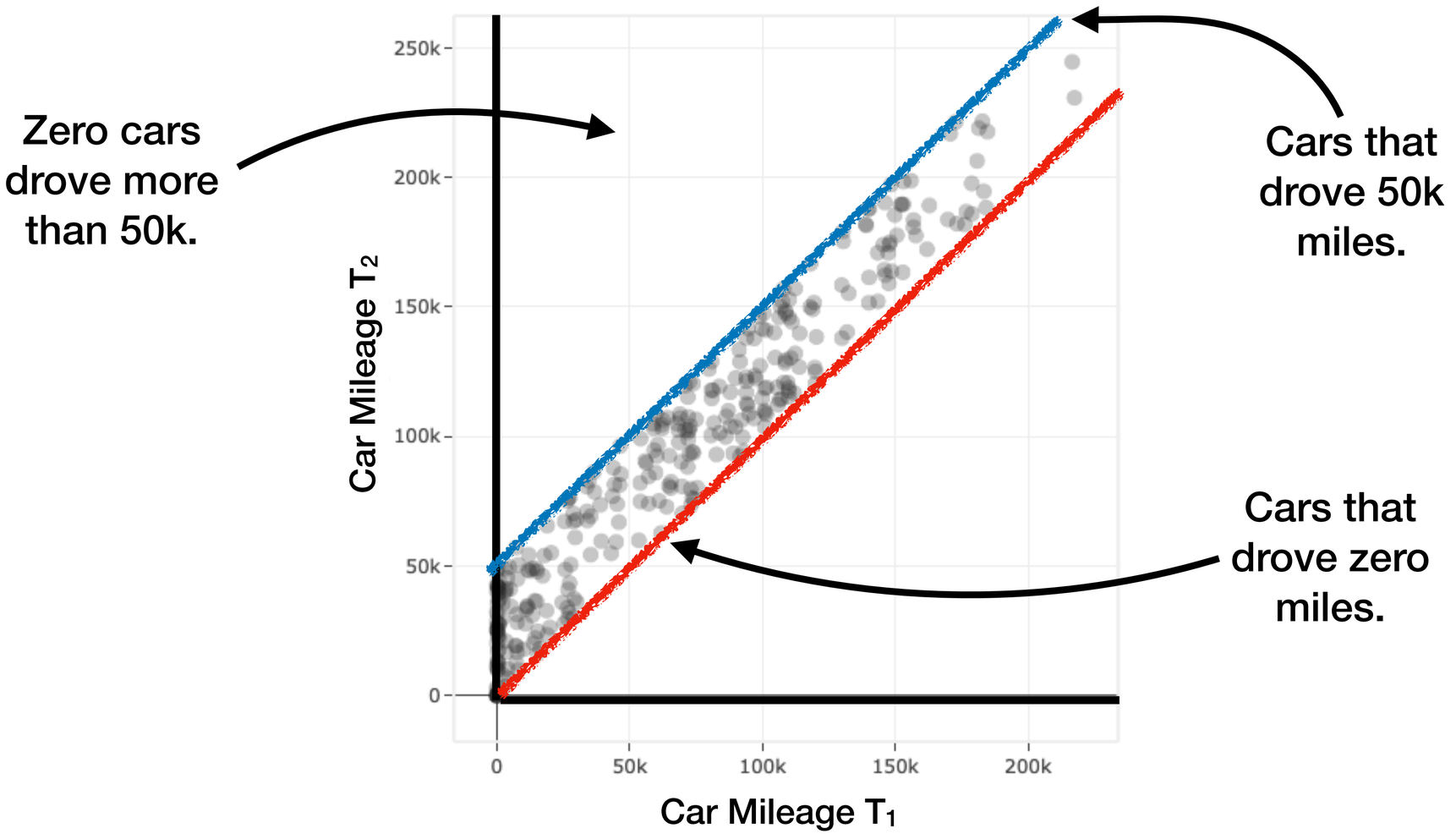
Height (cm)	Weight (kg)
175	79
155	93
182	65
202	40
165	105
169	87
159	73

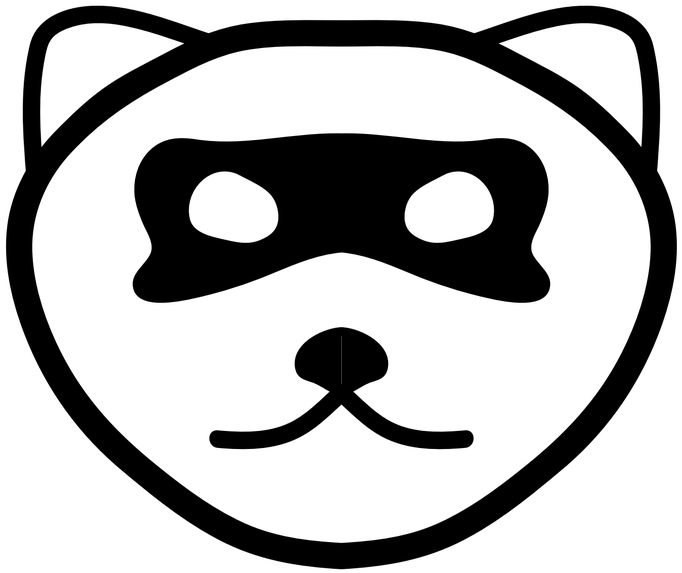


Deviation from Domain Expectations

DS-DRIVING

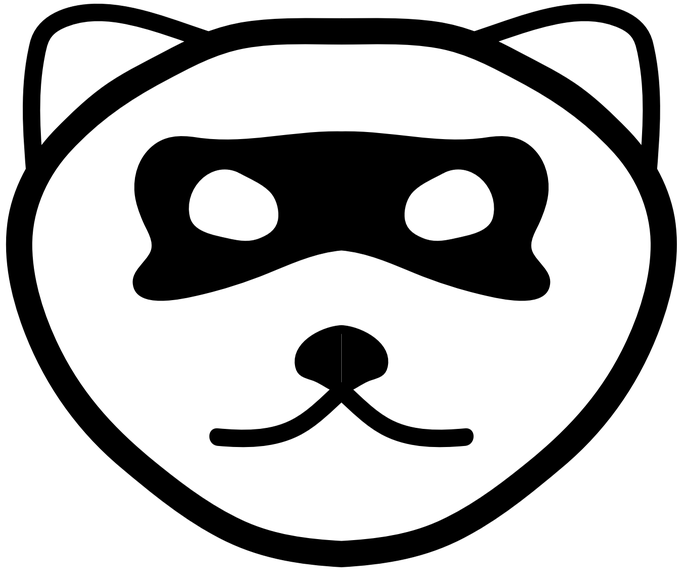
Height (cm)	Weight (kg)
175	79
155	93
182	65
202	40
165	105
169	87
159	73



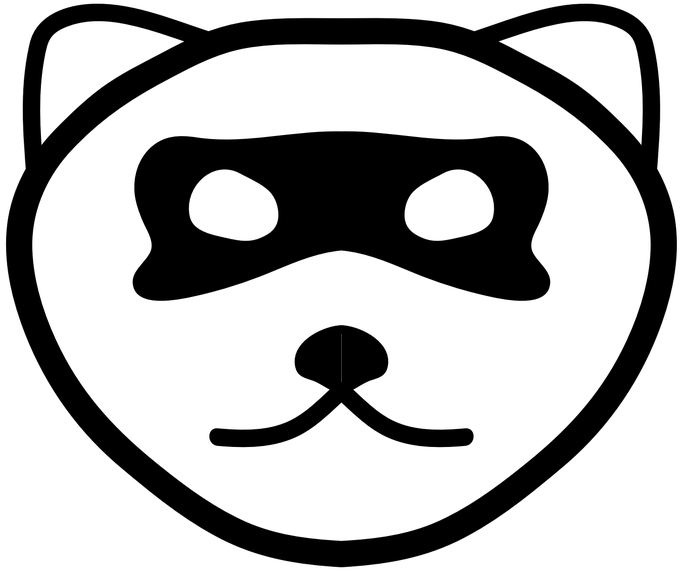


Implement Policies

Ethical Considerations

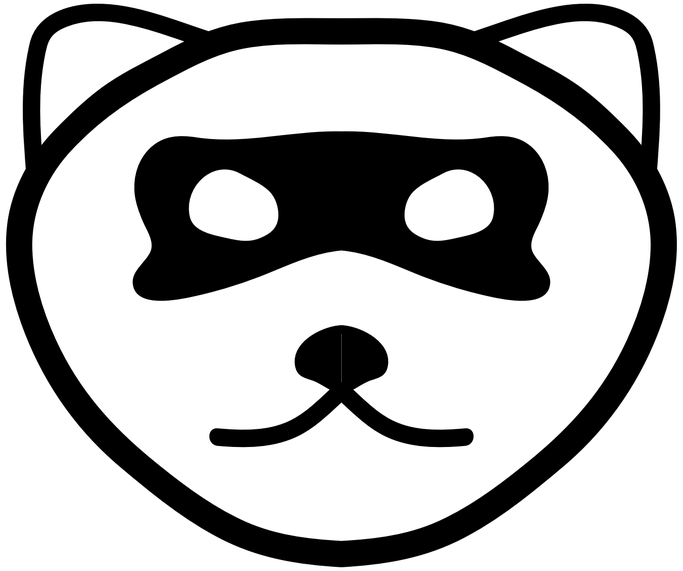


Ethical Considerations



False positives?

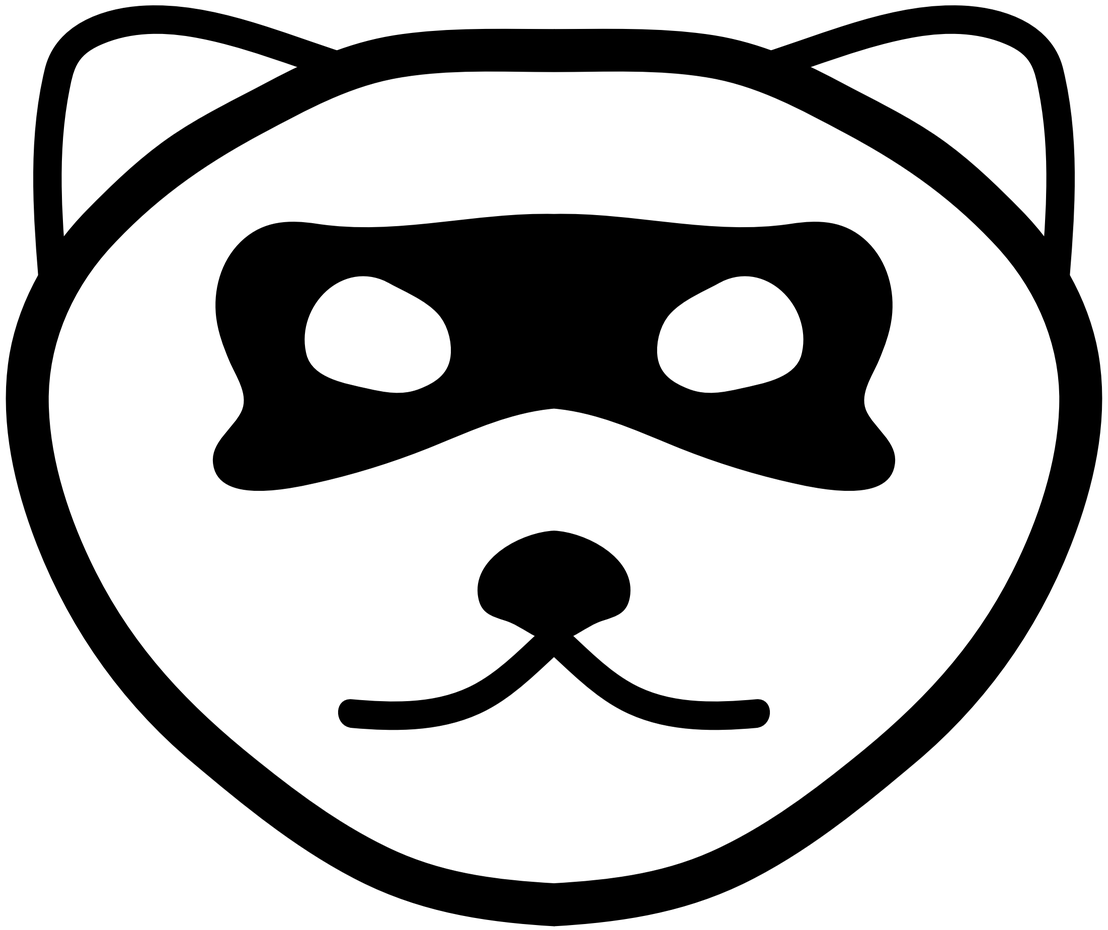
Ethical Considerations



False positives?

Use by bad actors?

Conclusion



Ferret: Reviewing Tabular Datasets for Manipulation

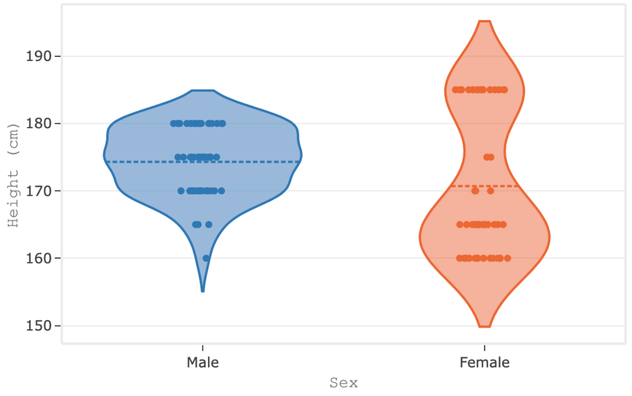
Devin Lange, Shaurya Sahai, Jeff M. Phillips, Alexander Lex

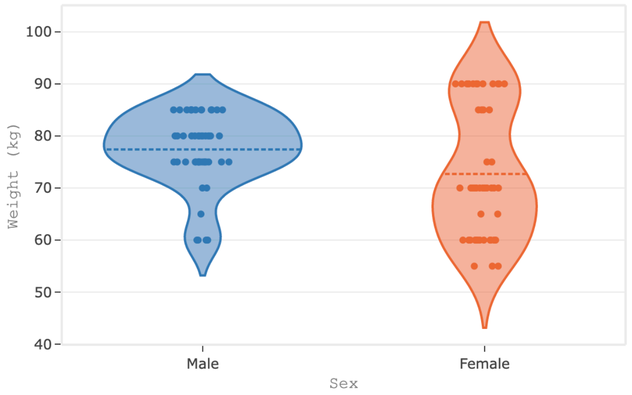
THE UNIVERSITY OF UTAH

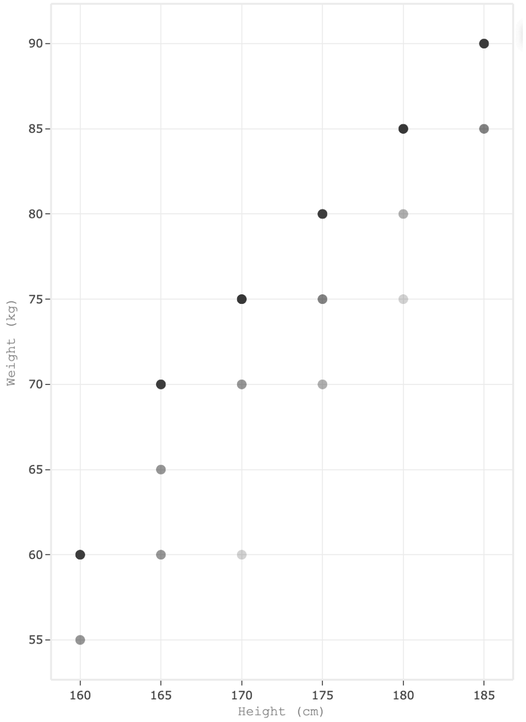
visualization design lab

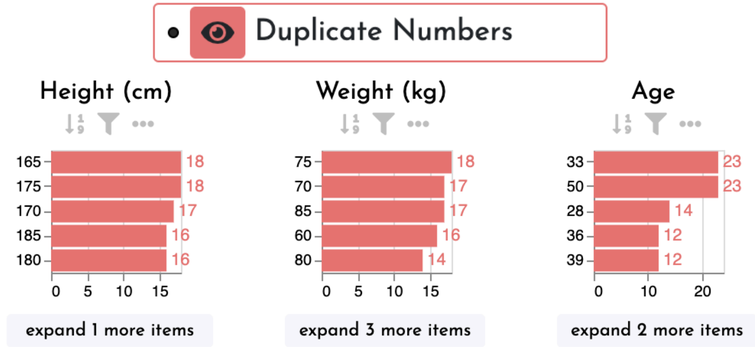
Appendix

Prompt: Please generate me a table with 100 rows of height, weight, age, sex, and BMI of people









Is this data fabricated?

Time (seconds)
2.57257
5.82583
7.82566
3.69847
2.44766
6.42546

Is this data fabricated?

Time (seconds)
2.57257
5.82583
7.82566
3.69847
2.44766
6.42546

What if…

Is this data fabricated?

Time (seconds)
2.57257
5.82583
7.82566
3.69847
2.44766
6.42546

What if…

data was recorded with a stopwatch?

Is this data fabricated?

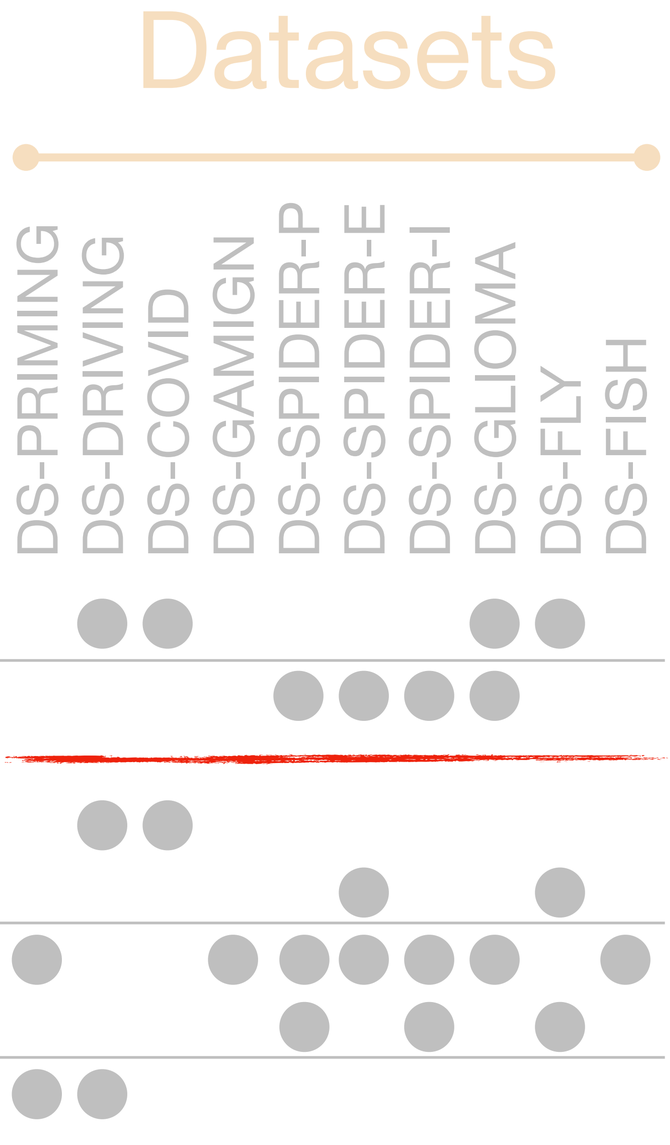
Time (seconds)
2.57257
5.82583
7.82566
3.69847
2.44766
6.42546

What if…

data was recorded with a stopwatch?

data was recorded automatically with high precision digital device?

Artifacts of Manipulation



Living Document Online

Artifacts of Manipulation

This document contains a living document of known artifacts of manipulation. If you would like to contribute, please submit an issue, or submit a pull request to suggest direct modifications to these documents. Check out the research project that helped create this document and the visualization tool that helps find these artifacts efficiently.

Expand All | Collapse All

Formatting

Formatting refers to metadata assigned to cells by spreadsheet tools like Excel or Google Sheets. This includes font, font size, methods of text emphasis such as bold, italics, and underlining, as well as background colors. In addition, a user can change the data format, e.g. changing a cell to a date format will alter how it is displayed.

Unexpected Formatting

What to look for:

A variation in formatting, especially within a single column, can be an indication there are problems with the data. These problems could be the result of clerical errors, or data processing done in another spreadsheet with different formatting, then copied back to the main spreadsheet.

Caveats:

In some cases, a variation of formatting is expected. Different types of data formats (numerical, string, etc) in columns are common. Also, sometimes certain data values are highlighted to make them easier to notice in the source dataset. Unusual formatting could be the result of valid processing or unintentional errors.

https://ferret.sci.utah.edu/artifacts.html