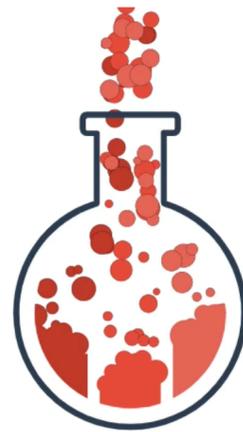




# Ferret: Reviewing Tabular Datasets for Manipulation

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



**visualization**  
**design lab**

 Science

## Potential fabrication in research images threatens key theory of Alzheimer's disease

In August 2021, Matthew Schrag at the University of Cambridge got a call that would change his life.  
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...](https://www.science.org/content/article/fraudulent-data-raise-questions-about-superstar-honesty)

## Fraudulent data raise questions about superstar honesty ...

by C O'Grady · Cited by 3 — Fraudulent **data** raise questions about honesty  
Dan Ariely denies **fabricating data**, but can't produce records

 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 to work on his research.  
Aug 10, 2022

 Nature Journal

<https://www.nature.com/news>

## Investigation finds Swedish scientist guilty of fabricating data

Dec 7, 2017 — ... on controversial paper by Lennart Lönnstedt  
Lönnstedt guilty of having intentionally fabricated data

 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 Problem

Scientists need to toughen up about preventing fraud from being 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

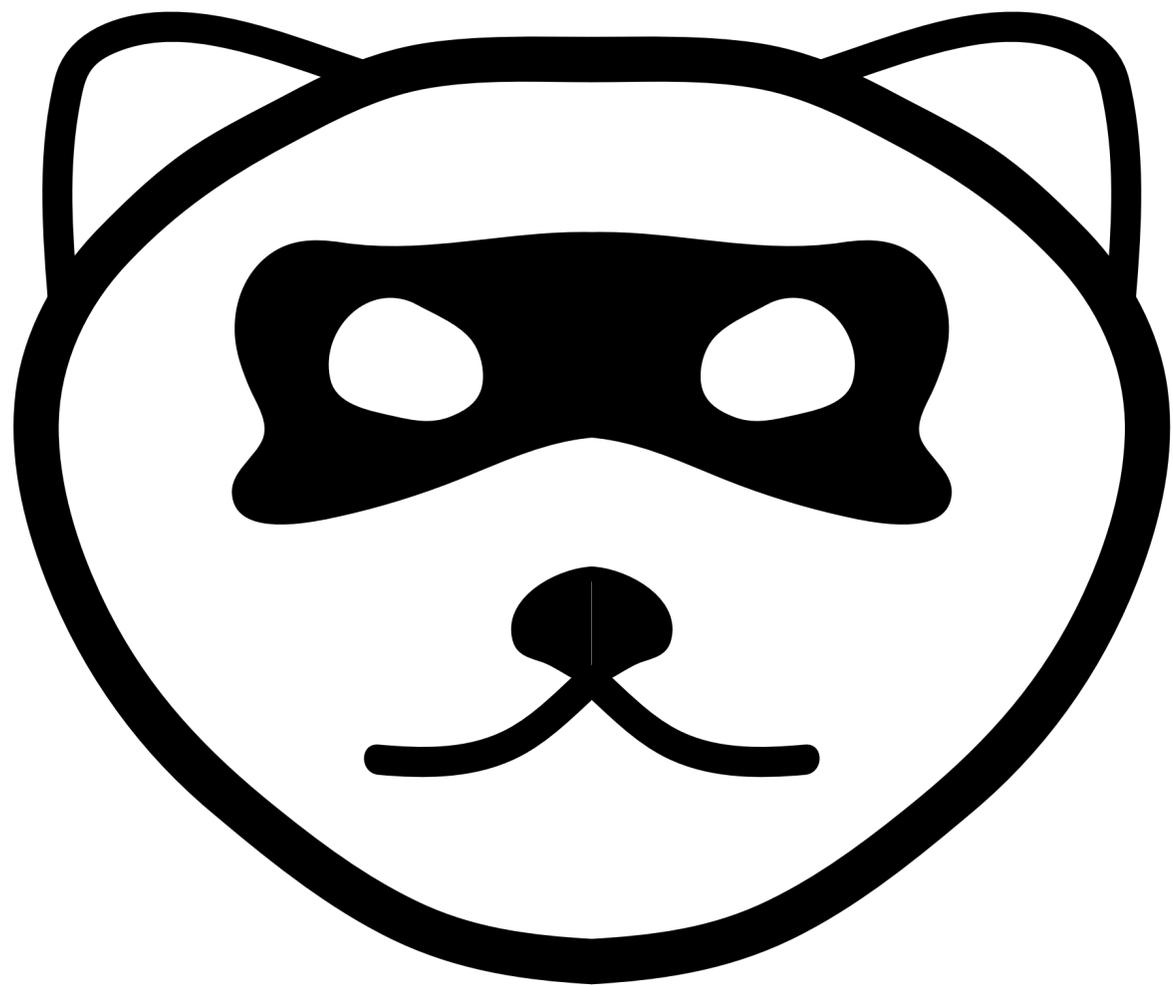


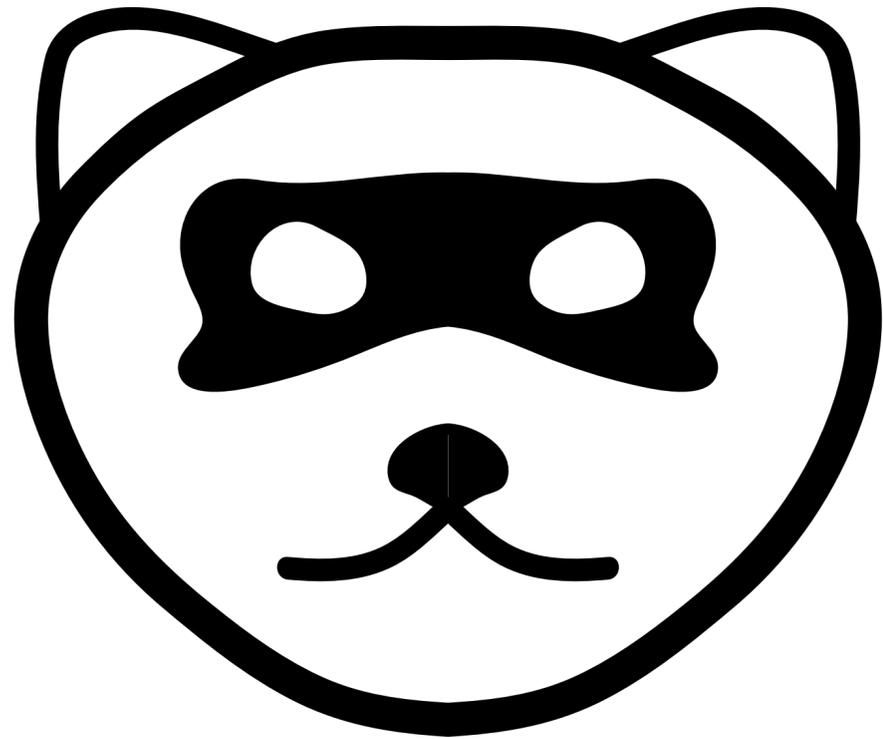
University of California, Davis

<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 ...



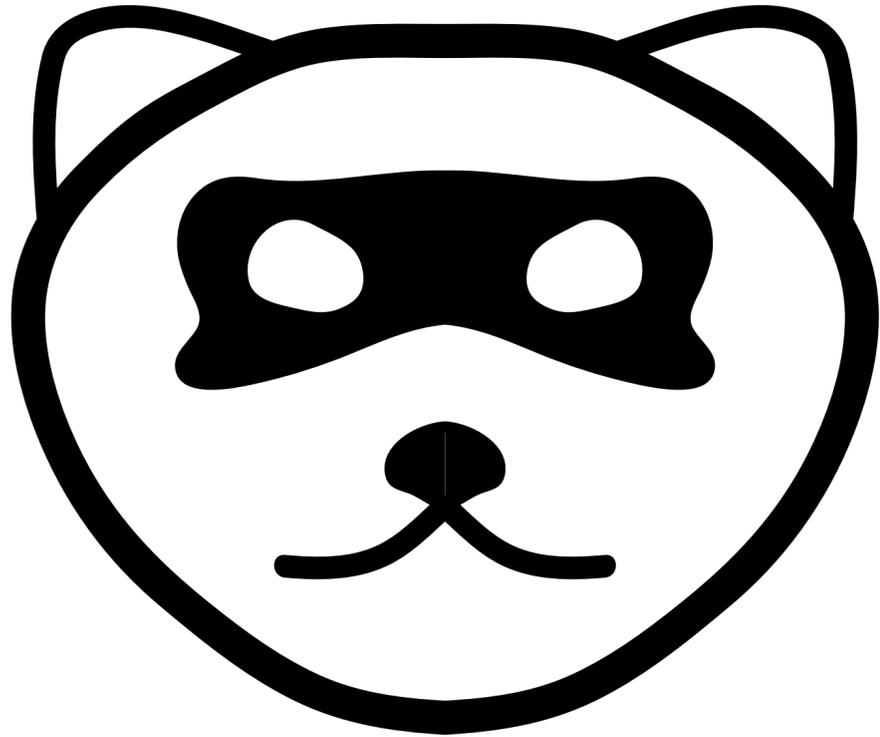


**Understand Fraud**

**Create Tools**

**Implement Policies**

# Understand Fraud



# Methods: Datasets and Patterns of Manipulation

**Literature Review**

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

**Retraction Watch Database<sup>1</sup>**

→ **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
<b>16</b>	<b>93</b>	92
<b>3</b>	<b>96</b>	98
30	87	96
12	45	33
45	89	11
70	56	14

# Repeated Regions

Column A	Column B	Column C
45	79	10
<b>16</b>	<b>93</b>	92
<b>3</b>	<b>96</b>	98
30	87	96
12	45	33
45	<b>16</b>	<b>93</b>
70	<b>3</b>	<b>96</b>

Column A

=random()

# Unexpected Variation of Precision

Column A

0.9126470121876

Unexpected Variation of Precision

# 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

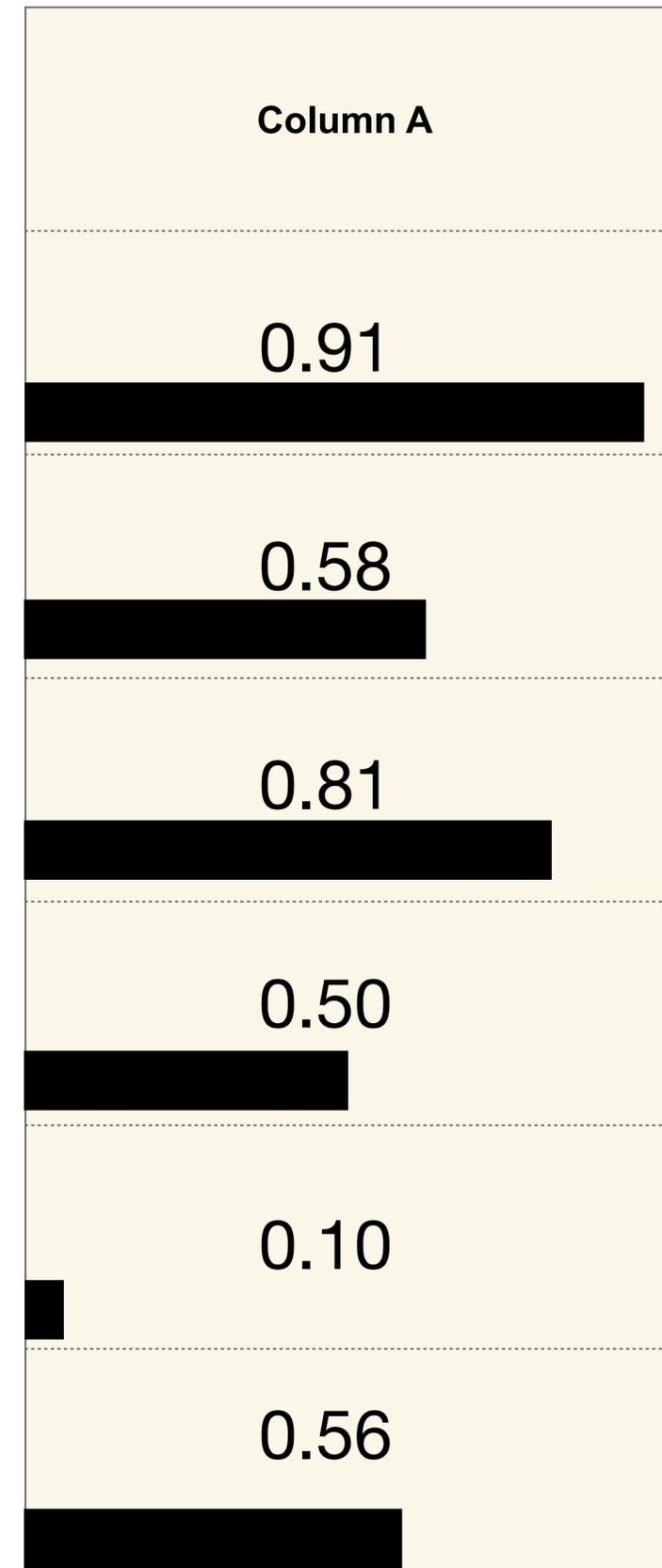
0.5059135817684

0.12

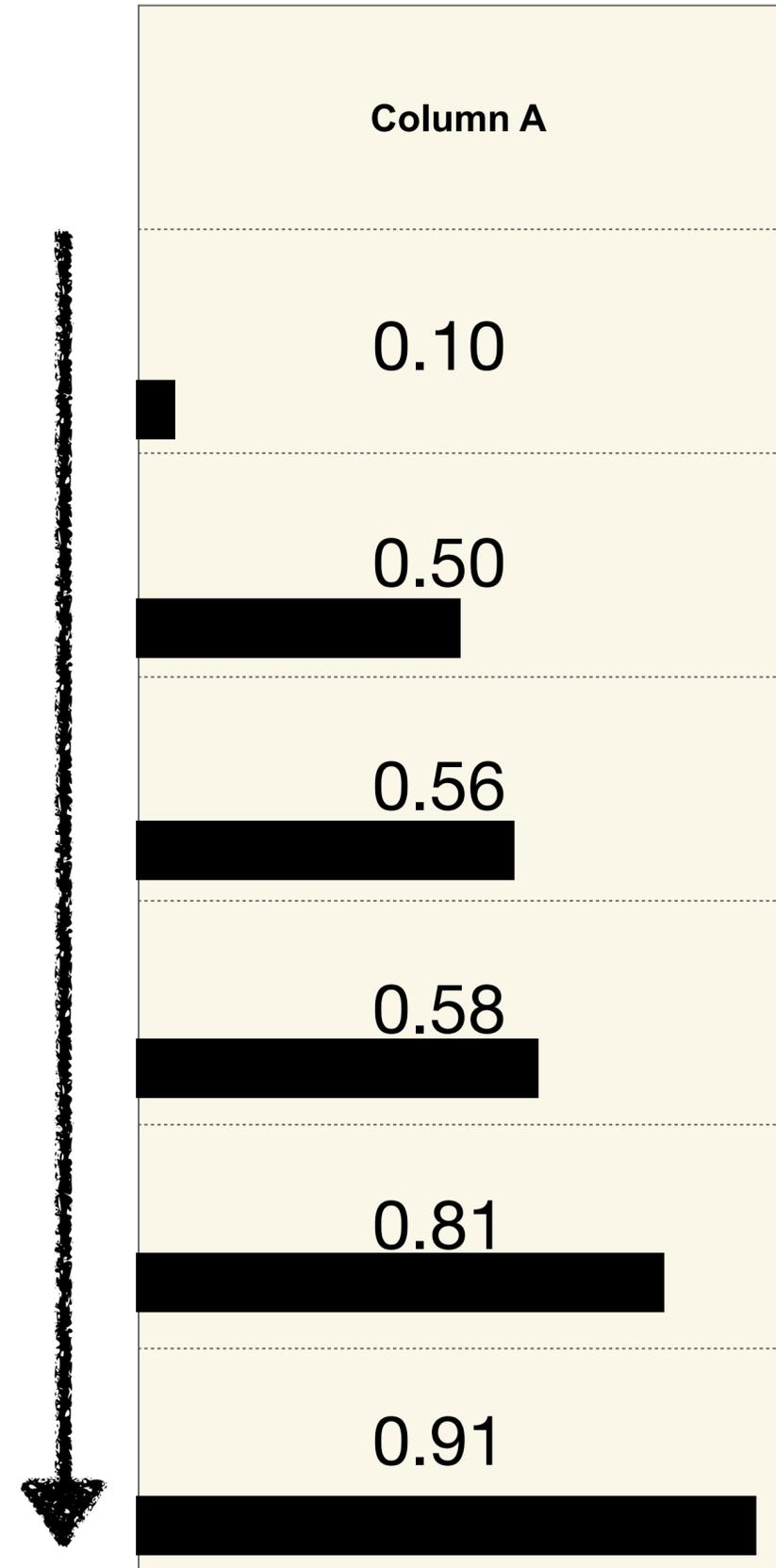
0.10941148406697

# Ordering Artifacts

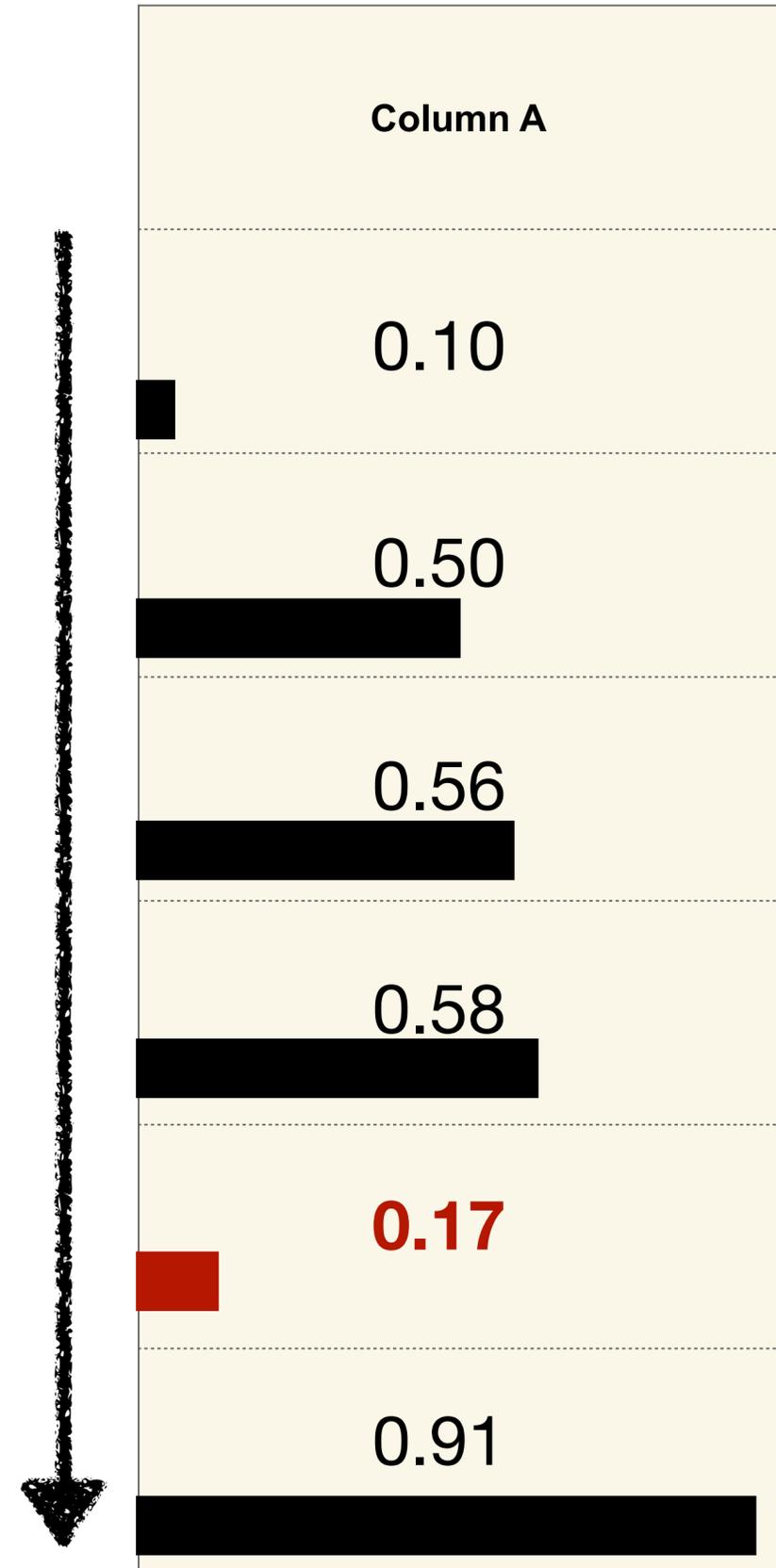
# Ordering Artifacts



# 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

## Artifacts of Manipulation

### Formatting

Unexpected Formatting

DS-PRIMING  
DS-DRIVING  
DS-COVID  
DS-GAMIGN  
DS-SPIDER-P  
DS-SPIDER-E  
DS-SPIDER-I  
DS-GLIOMA  
DS-FLY  
DS-FISH



Duplicate Numbers and Digits



### Numerical

Unexpected Leading Digits

Unexpected Trailing Digits



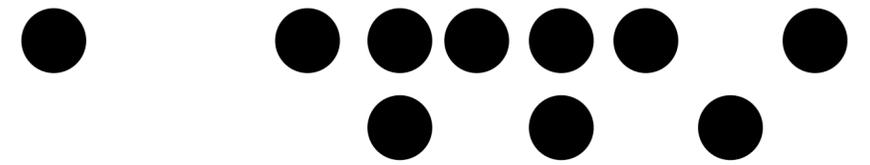
Unexpected Variation of Precision



### Structural

Repeated Regions

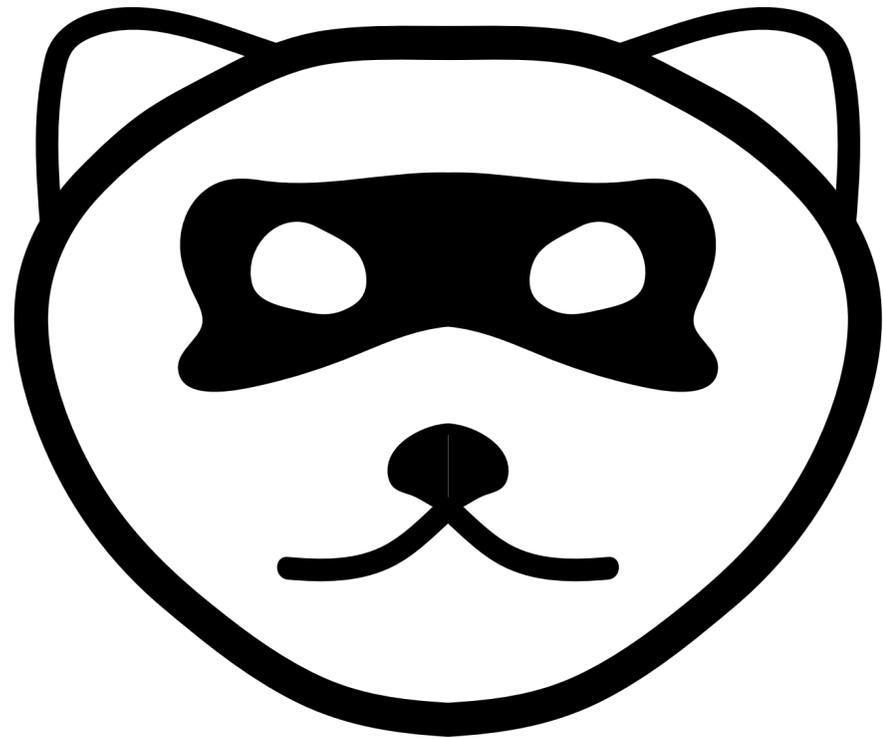
Ordering Artifacts



### Domain

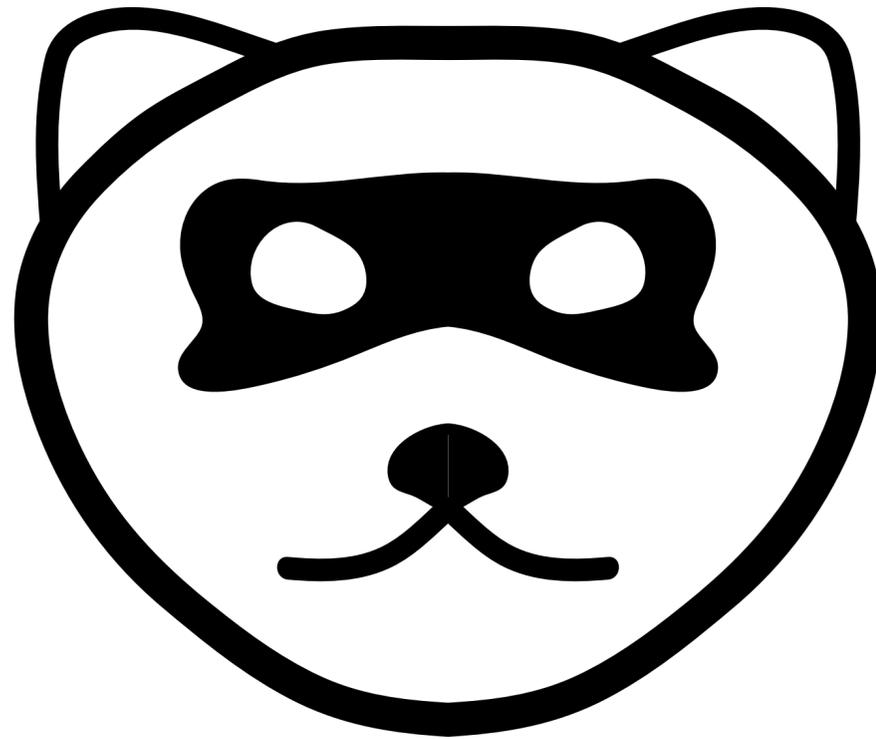
Deviation from Domain Expectations





**Create Tools**

# Create Tools



**iThenticate** Quotes Included  
Bibliography Included **43%**  
SIMILAR

treatment of hepatocellular malignancy (Schutte et al., 2009)

many surveillance programs that aimed at detecting early stage HCC have been widely recommended. These programs are based on the use of ultrasound and alpha-fetoprotein (AFP) (Wong et al., 2009) the reliability of imaging techniques has greatly improved in the last years but such diagnostic procedures are expensive and subject to misinterpretation. On the other hand, AFP is the only diagnostic serologic test currently used in surveillance programs either in asymptomatic cases or cirrhotic patients (Sterling et al., 2010). But its reliability remains satisfactory as its sensitivity (41% - 65%), therefore identification of a new biomarkers to establish the risk of cancer and / or detect its appearance at a preclinical stage is urgently needed (Zhang et al., 2013).

YKL-40: The abbreviation "YKL-40" is based on the one letter code for the first three NH<sub>2</sub>-terminal amino acids: tyrosine (Y), lysine (K), and leucine (L). It is a plasma glycoprotein and a highly conserved member of the mammalian chitinase-like proteins and its molecular weight of 40 kDa and it contains a single polypeptide chain of 383 amino acids, also known as chitinase-3-like-1 (CHI3L1), human cartilage glycoprotein-39 (HC gp39), chondrex, Breast regression protein 39 (Brp-39), 38-kDa heparin-binding glycoprotein (Gp38k), and 40-kDa mammary gland protein (MGP-40) (Mygind, N.D. & Kastrop, J., 2015). YKL-40 has a role in cell proliferation and differentiation, inflammation, and protection against apoptosis (Recklies et al., 2002). It considered to be a growth factor for connective

**MATERIALS AND METHODS**

**SUBJECTS**

This study was conducted at El Nasr El Ainy Hospital. It included 88 patients (54 males and 34 females) and their ages ranged from (40 to 60 years). These patients were classified in to 4 groups:

**Group 1 (patients with HCC; n=20)**

This group included 20 patients with hepatocellular carcinoma (HCC), 16 patients were diagnosed as hepatitis c virus (HCV; n=16) and 4 patients were diagnosed as hepatitis B virus (HBV; n=4). They were 12 males (60%) and 8 females (40%).

**Group 2 (patients with Liver cirrhosis; n=28)**

This group included 28 patients with liver cirrhosis. They were 13 males (46.43%) and 15 females (53.57%).

All patients were diagnosed as hepatitis C virus infection and diagnosed by presence of HCV Abs and confirmed by PCR.

**Group 3 (patients with HBV; n=20)**

This group included 20 patients with HBV. They were 16 males (80%) and 4 females (20%). They were diagnosed by presence of HBs Ag.

**Group 4 (patients with HCV; n= 20)**

This group included 20 patients with HCV. They were 13 males (65%) and 7 females (35%).

**+Control group (n=20)**

This group included 20 sex and age matched healthy individuals. They were 16 males and 4 females, whose ages ranged from 40 to 60 years.

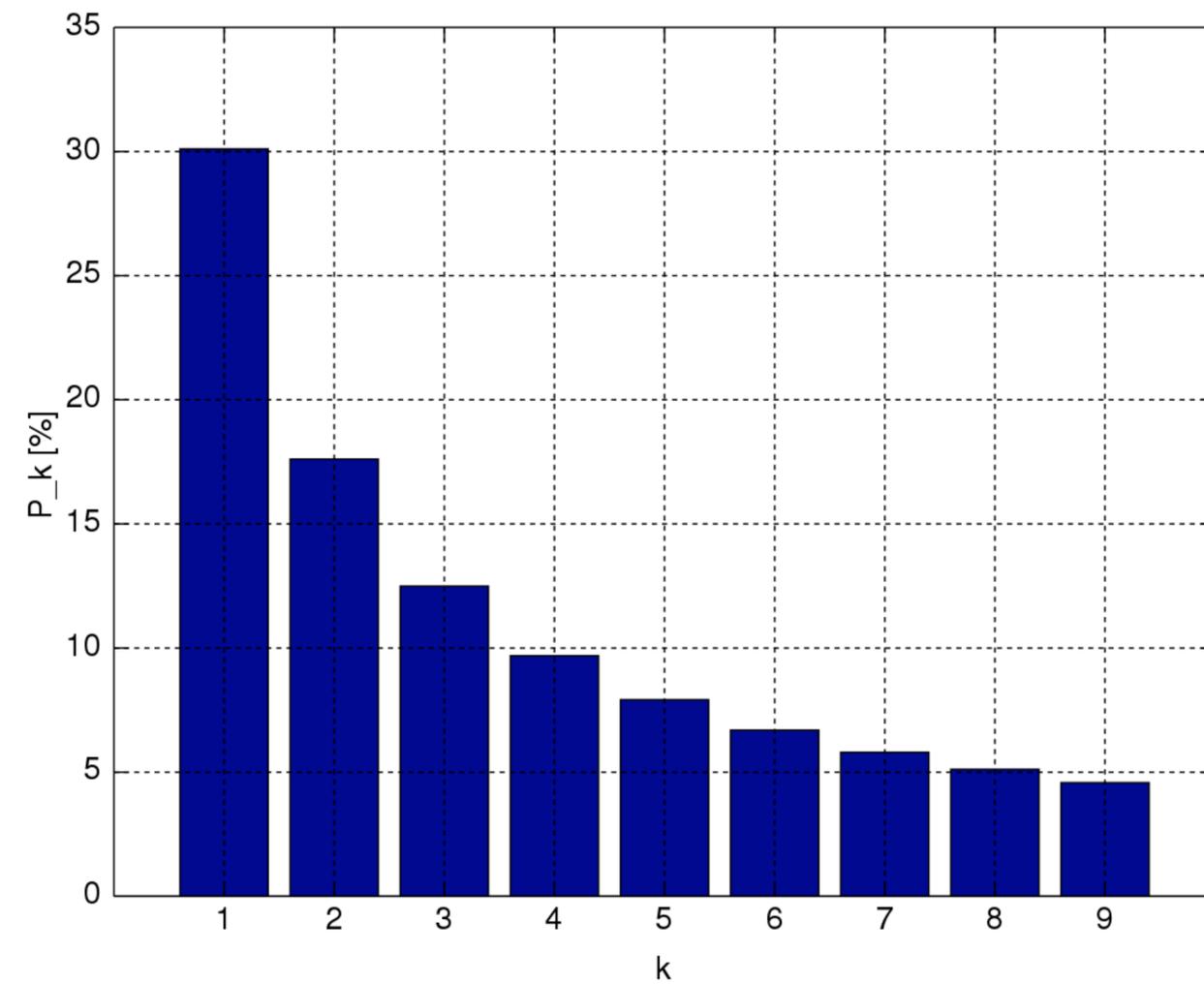
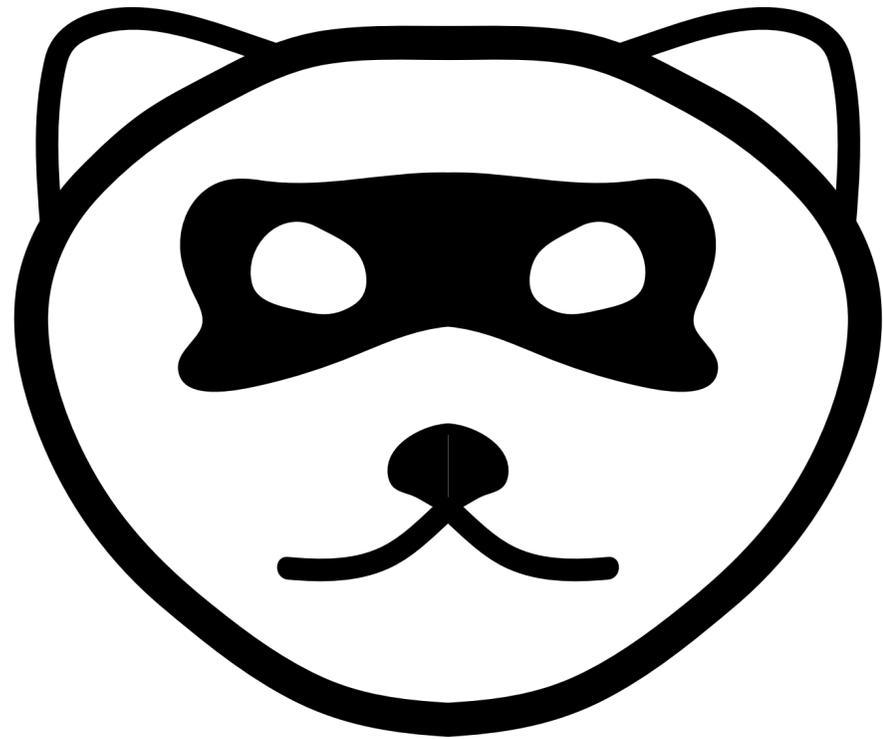
All subjects included in this study were subjected to the following investigations:

**Match Overview**

Match	Source	Words	Percentage
1	Internet	477 words crawled on 21-Feb-2018 jarm.com	8%
2	Internet	117 words crawled on 04-Apr-2018 www.isisn.org	2%
3	Internet	111 words crawled on 09-Apr-2018 www.ejbps.com	2%
4	Crossref	110 words Tao, Hui, Jing-Jing Yang, Kai-Hu Shi, Cheng Huang, Lei ... ang, Xiong-Wen Lv, and Jun Li. "The significance of YKL-4	2%
5	Crossref	104 words Ahmed, S.F., E.A.S. Attia, A.A. Saad, M. Sharara, H. Fawz y, and E.M.A. El Nahrery. "Serum YKL-40 in psoriasis with	2%
6	Crossref	94 words Biomarkers in Disease Methods Discoveries and Applicatio	2%

PAGE: 2 OF 9 Text-Only Report

# Create Tools





# Ferret Overview

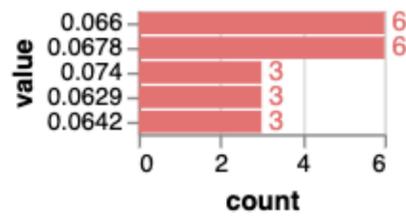
▼ Select Analysis

[← Previous](#)
[Next →](#)

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

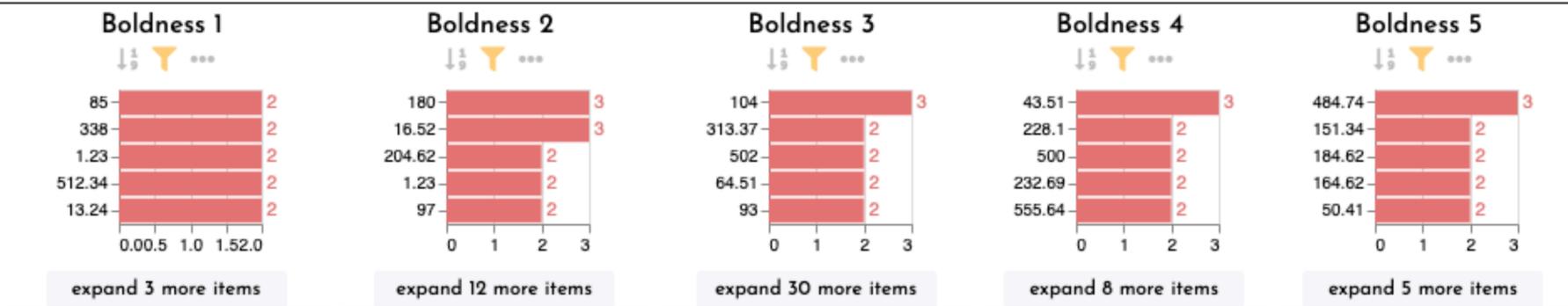
> Ignore (1)  
 > Highlight (1)  
 > Summary Stats  
 > Select Columns

## 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.



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

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

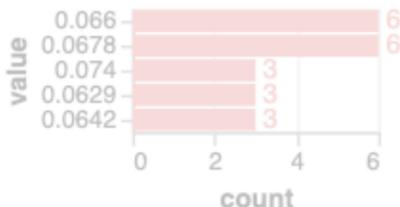
> Ignore (1)

> Highlight (1)

> Summary Stats

> Select Columns

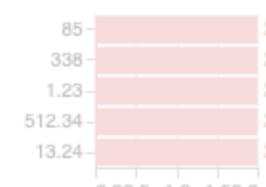
Duplicate Numbers



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.

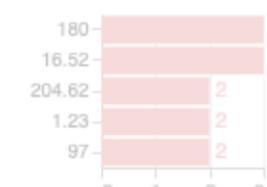
Show more.

**Boldness 1**



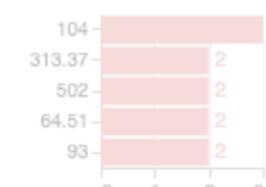
expand 3 more items

**Boldness 2**



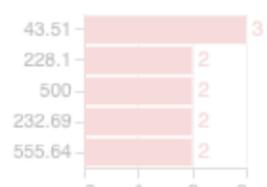
expand 12 more items

**Boldness 3**



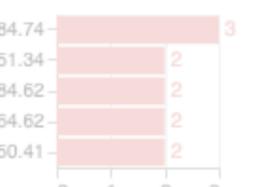
expand 30 more items

**Boldness 4**



expand 8 more items

**Boldness 5**



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

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

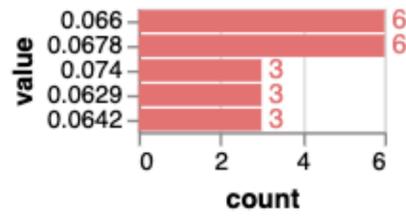
> Ignore (1)

> Highlight (1)

> Summary Stats

> Select Columns

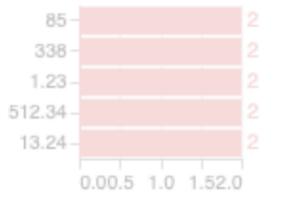
Duplicate Numbers



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.

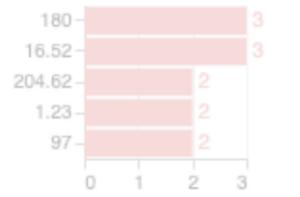
Show more.

**Boldness 1**



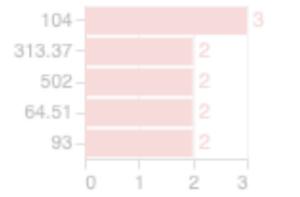
expand 3 more items

**Boldness 2**



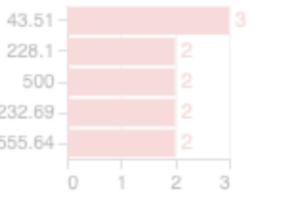
expand 12 more items

**Boldness 3**



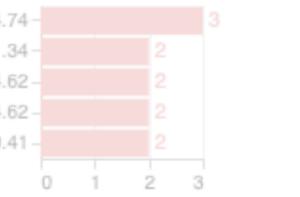
expand 30 more items

**Boldness 4**



expand 8 more items

**Boldness 5**



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

Analysis Selection

Analysis Explanation



# 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

value	count
0.066	6
0.0678	6
0.074	3
0.0629	3
0.0642	3

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.

**What to look for:**

A large number of duplicate numbers can be caused by data manipulation. Either typing numbers manually or copying/pasting regions of numbers can produce this effect. Highlighting duplicate numbers and searching for repeated regions of numbers can provide further evidence if data has been manipulated or not.

**Caveats:**

Determining if the number of duplicates is unusual or not can be difficult. A large number of duplicates could have several causes:

- There is a small range of possible values compared to the number of data points.
- The data have been clamped to an upper/lower bound.
- A common cause of duplicate numbers and sequences of digits that may seem suspicious at first, but is typically innocent are high-precision duplicates caused by converting measurements. For example, converting fractions to decimals could introduce duplicates with seemingly high precision. If an experiment recorded the length of an animal in inches as integers, but in a subsequent step, the data would be converted to feet using decimals, we would expect that the resulting decimals have values with high precision, such as 0.33333333 and 0.41666667. In this case, the number 0.33333333 may appear more often than naively expected and an n-gram of digits, such as 3s, or 6s may appear frequently.

**External References**

- [Birthday Paradox](#)

Show less.

600.00 600.00 502.000 43.51 600.00

Analysis Selection

Analysis Explanation



# Ferret Overview

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)

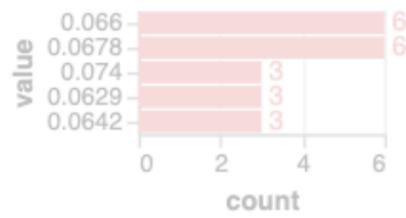
> Highlight (1)

> Summary Stats

> Select Columns

Analysis Selection

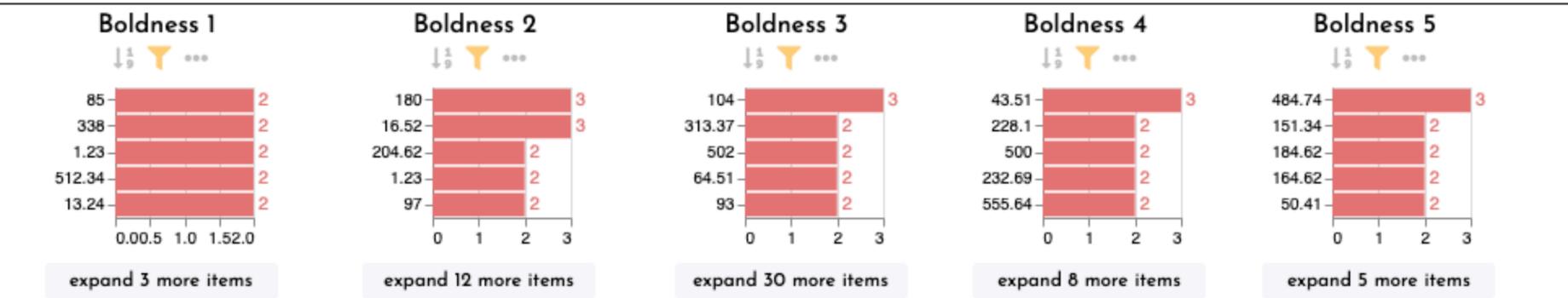
## Duplicate Numbers



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



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



# Ferret Overview

Home
Upload
Settings
Descriptions

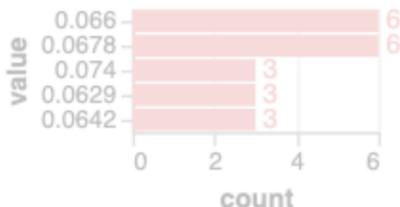
▼ Select Analysis

< Previous
Next >

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

Show all unique values

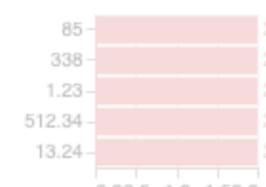
Duplicate Numbers



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.

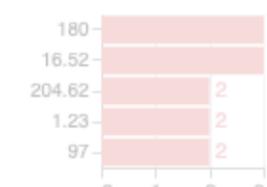
Show more.

Boldness 1



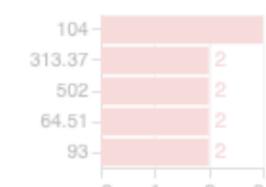
expand 3 more items

Boldness 2



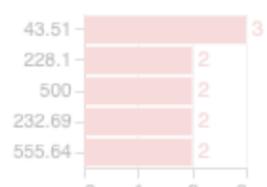
expand 12 more items

Boldness 3



expand 30 more items

Boldness 4



expand 8 more items

Boldness 5



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

Analysis Selection

Analysis Explanation

Summary Charts

Tabular Visualization



# 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

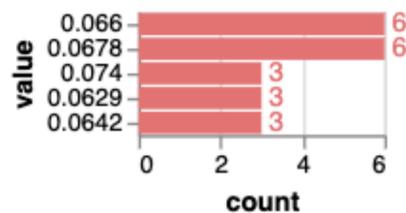
Ignore (1)

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.

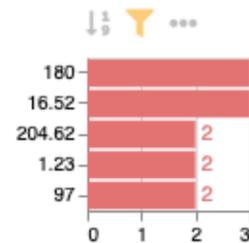
Show more.

### Boldness 1



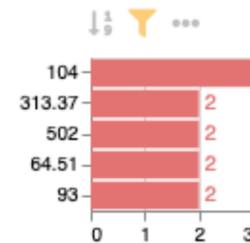
expand 3 more items

### Boldness 2



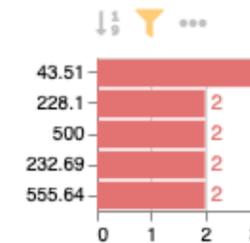
expand 12 more items

### Boldness 3



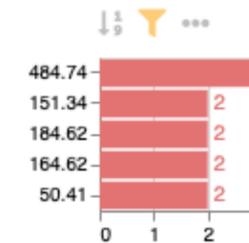
expand 30 more items

### Boldness 4



expand 8 more items

### Boldness 5



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

Analysis Selection

Analysis Explanation

Summary Charts

Tabular Visualization

# Case Study

## Datasets

### Artifacts of Manipulation

#### Formatting

Unexpected Formatting

DS-PRIMING  
DS-DRIVING  
DS-COVID  
DS-GAMIGN  
DS-SPIDER-P  
DS-SPIDER-E  
DS-SPIDER-I  
DS-GLIOMA  
DS-FLY  
DS-FISH

● ● ● ●

Duplicate Numbers and Digits

● ● ● ●

#### Numerical

Unexpected Leading Digits

Unexpected Trailing Digits

● ●

Unexpected Variation of Precision

● ●

#### Structural

Repeated Regions

Ordering Artifacts

● ● ● ● ● ● ● ● ● ●

#### Domain

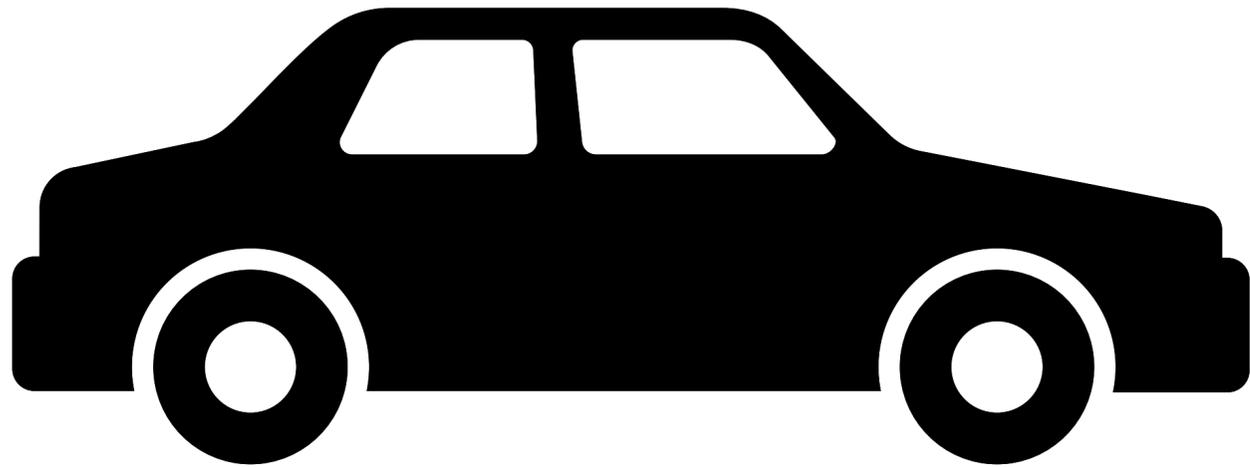
Deviation from Domain Expectations

● ●



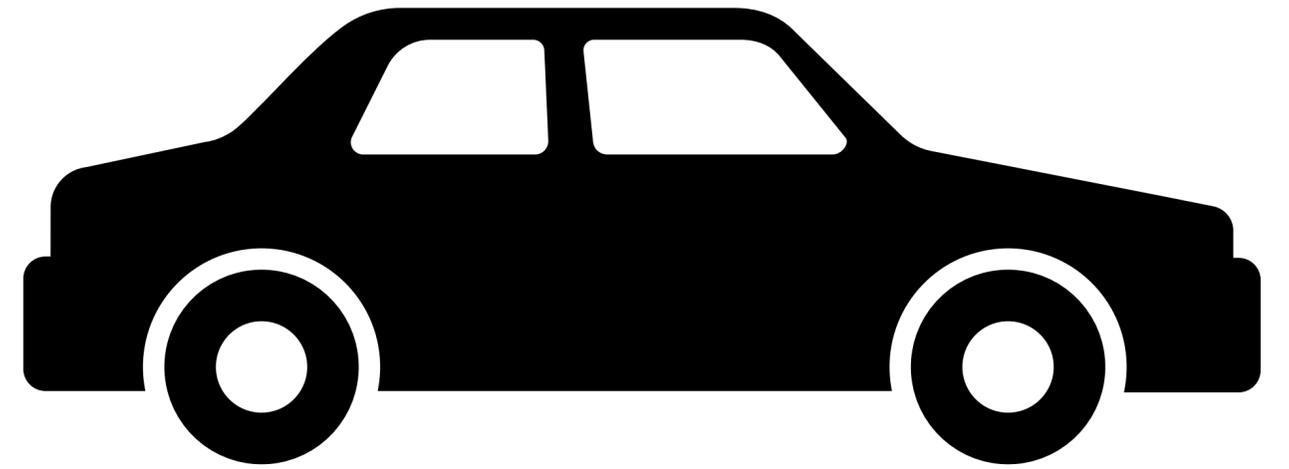
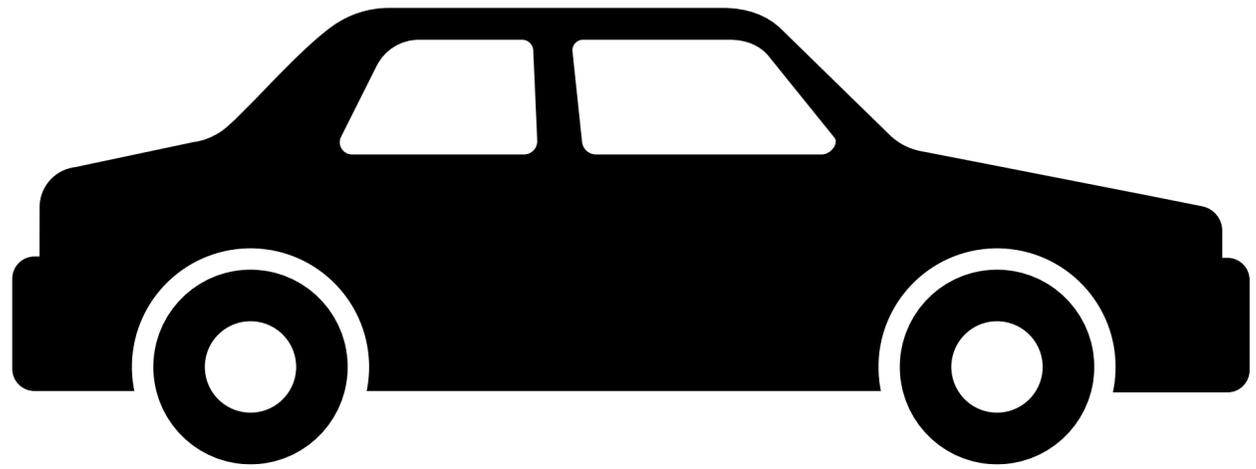


# DS-DRIVING



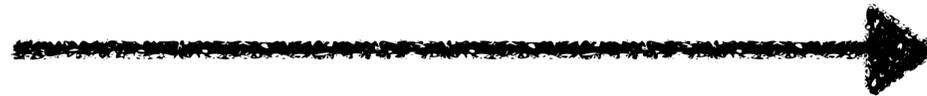
**Car Mileage  $T_1$**

# DS-DRIVING



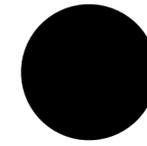
Time

Car Mileage  $T_1$



Car Mileage  $T_2$

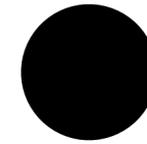
# Unexpected Formatting



DS-DRIVING

Column Header
56
74
81
48
22
49
62

# Unexpected Formatting



# DS-DRIVING

Column Header	
	56 ← Cambria
Calibri →	74
Calibri →	81
	48 ← Cambria
Calibri →	22
	49 ← Cambria
Calibri →	62

# Unexpected Formatting

Column Header	
	56 ← Cambria
Calibri →	74
Calibri →	81
	48 ← Cambria
Calibri →	22
	49 ← Cambria
Calibri →	62



# DS-DRIVING

Select Analysis

< Previous    Next >

- Formatting
- Include Data Format
- View in Overview Mode
- Value Distribution
- Duplicate Numbers
- Replicates
- Duplicate Digits
- Leading Digits
- Trailing Digits
- Precision
- Structural Overview
- Check Domain Expectations

Odom Reading 1 (Previous)

↓ ⋮

896
21396
21340
23912
16862
147738
18780
41930
28993
78382
58500
99417
93231
83443
22008
27950
67589
32753
33044

Cambria     Calibri     Selected 

# Case Study

## Datasets

### 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-PRIMING

DS-DRIVING

DS-COVID

DS-GAMIGN

DS-SPIDER-P

DS-SPIDER-E

DS-SPIDER-I

DS-GLIOMA

DS-FLY

DS-FISH

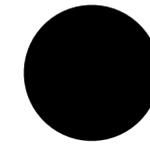


Deviation from Domain Expectations



DS-DRIVING

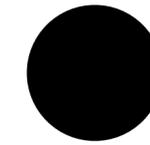
# Deviation from Domain Expectations



DS-DRIVING

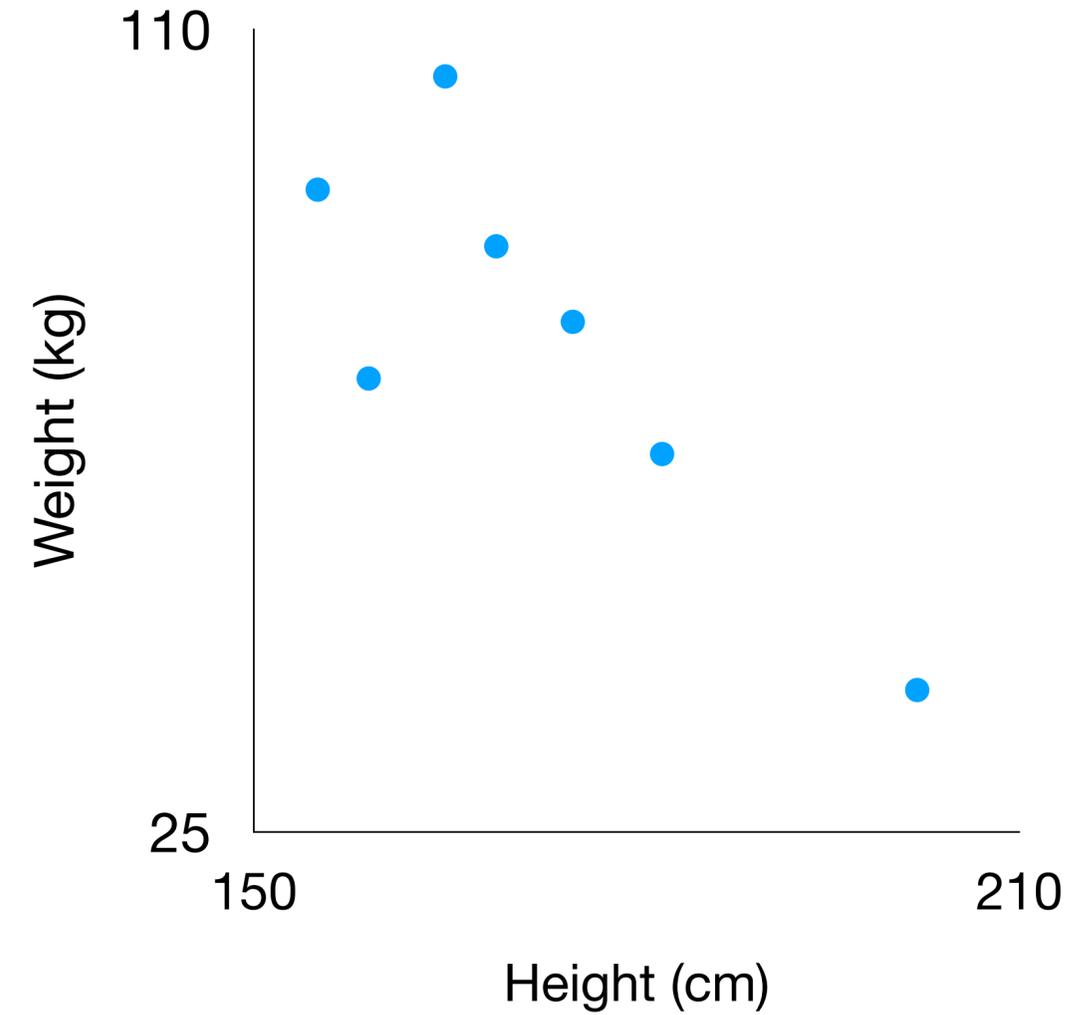
Height (cm)	Weight (kg)
175	79
155	93
182	65
202	40
165	105
169	87
159	73

# Deviation from Domain Expectations

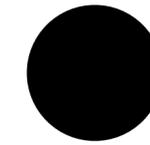


DS-DRIVING

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

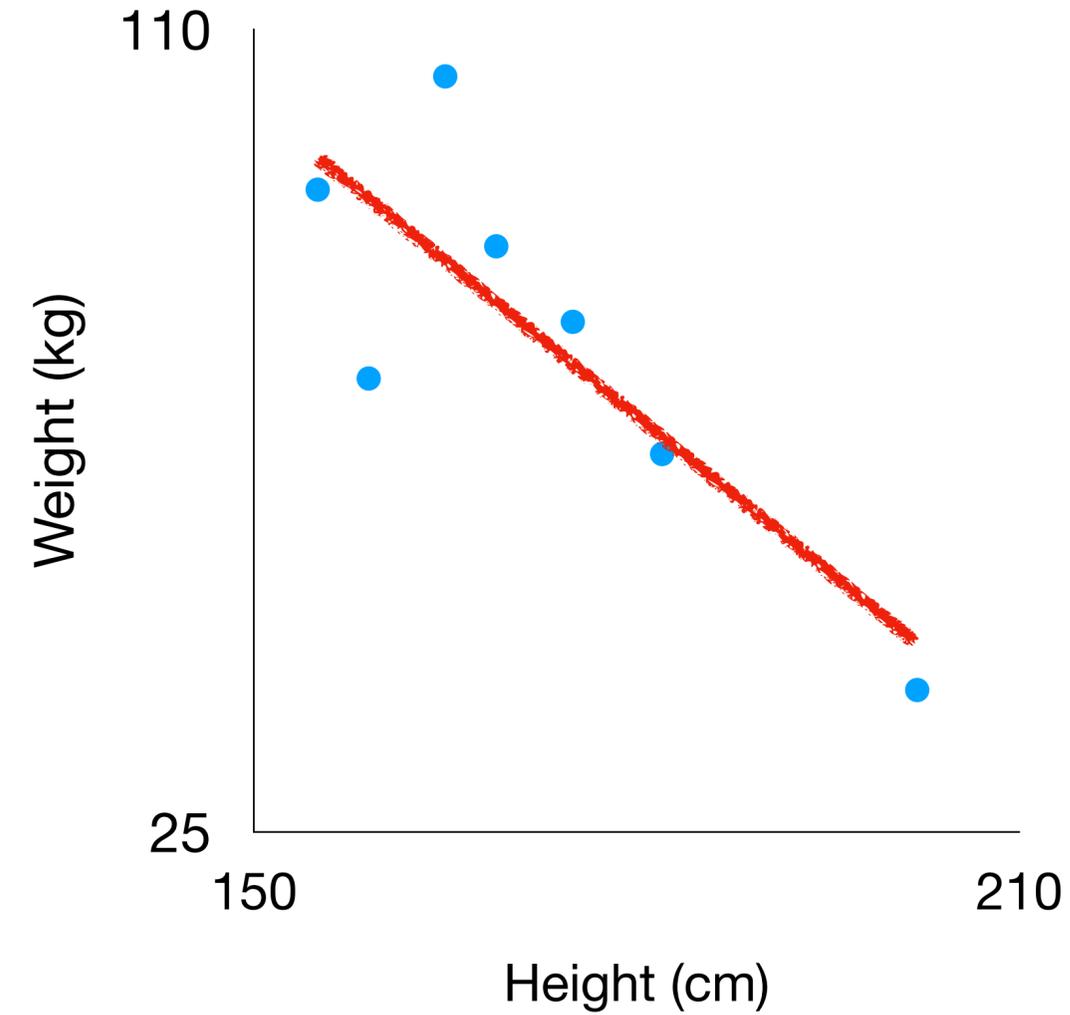


# Deviation from Domain Expectations

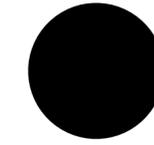


DS-DRIVING

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

Home
Upload
Settings
Descriptions

Select Analysis

< Previous    Next >

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

> Ignore (0)

> Highlight (0)

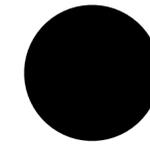
> Summary Stats

> Select Columns

	Odom Reading 1 ...	Odom Reading 1 ...	Odom Reading 2...	Odom Reading 2...
1	896	39198	0	0
2	21396	63511	32659	47605
3	21340	37460	44998	59002
4	23912	59136	0	0
5	16862	59292	0	0
6	147738	167895	125820	164688
7	18780	49811	45402	54824
8	41930	80323	181416	229852
9	28993	63707	13291	28165
10	78382	127817	0	0
11	58500	81081	0	0
12	99417	149211	48770	95179
13	93231	98047	0	0
14	83443	105094	0	0
15	22008	26486	0	0
16	27950	59514	95883	126309
17	67589	100475	27617	74443
18	32753	76724	0	0
19	33044	70775	0	0
20	104857	109961	19548	47796
21	121699	137849	0	0
22	16094	45489	159167	200316
23	78182	122739	21730	37863
24	21735	58693	39666	77258
25	47473	68971	4502	47293
26	121416	123661	53987	86852
27	4616	30507	0	0

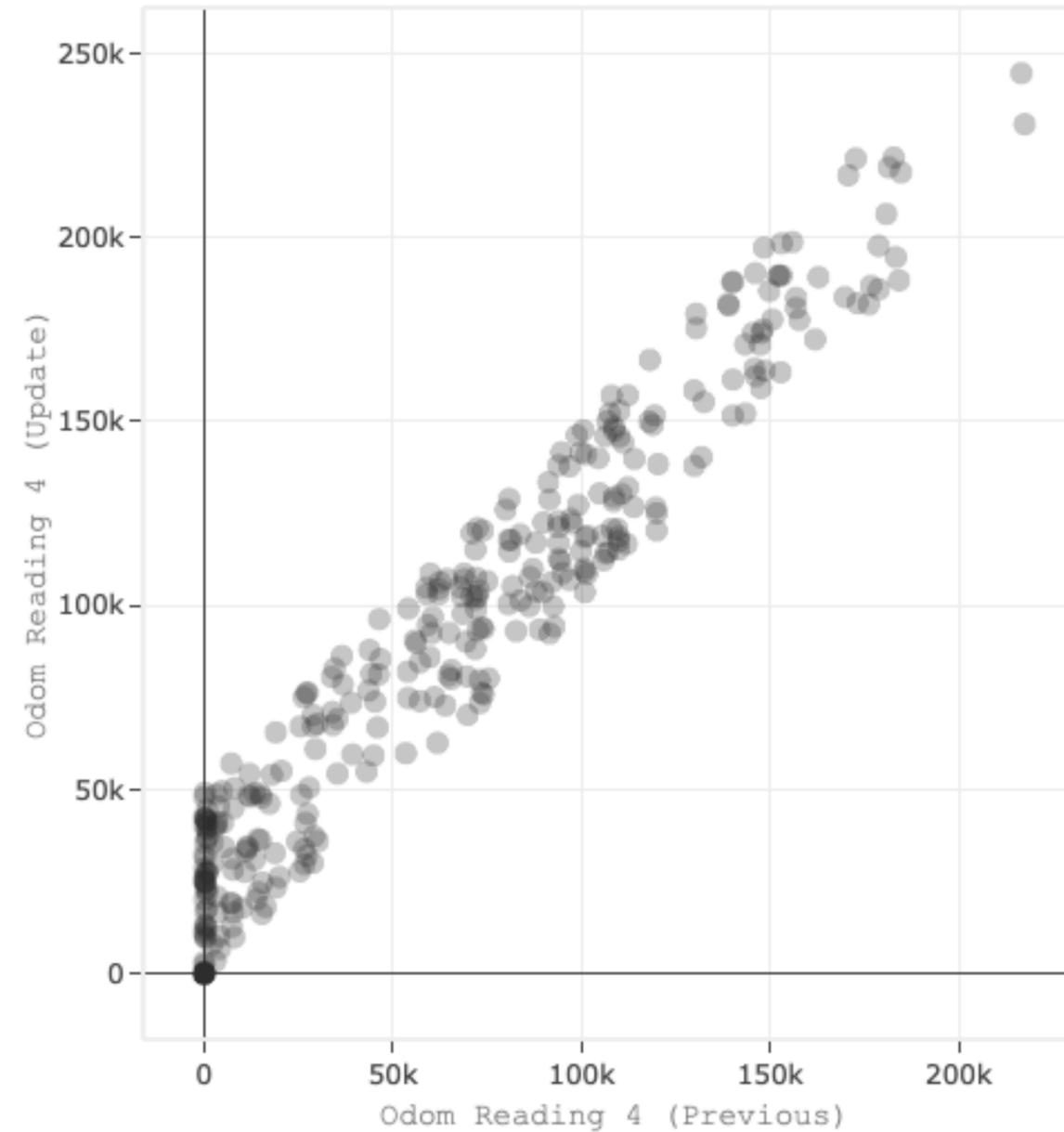
Opacity

# Deviation from Domain Expectations

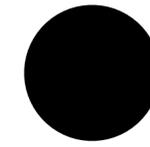


DS-DRIVING

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

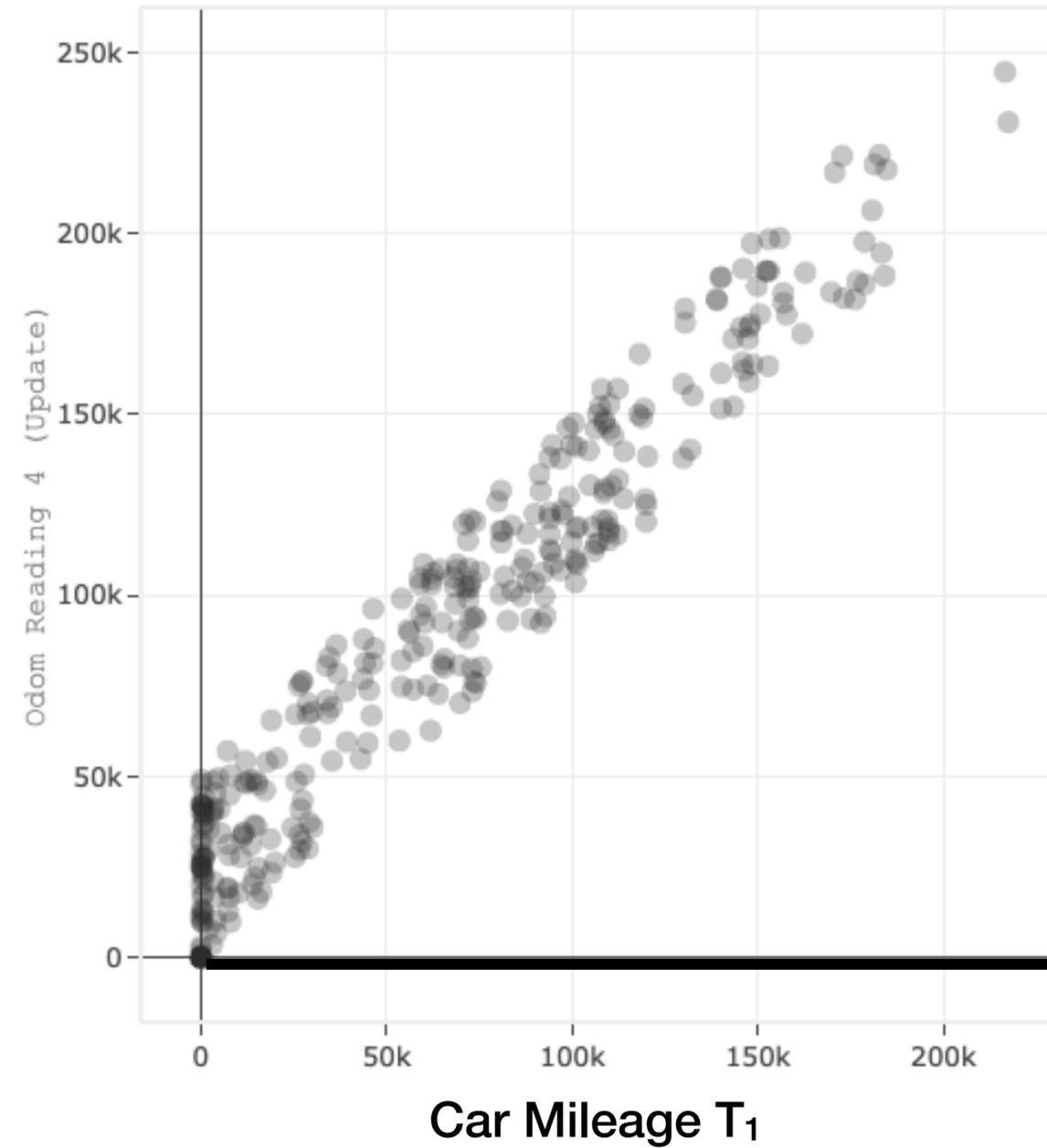


# Deviation from Domain Expectations

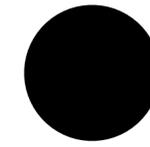


# DS-DRIVING

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

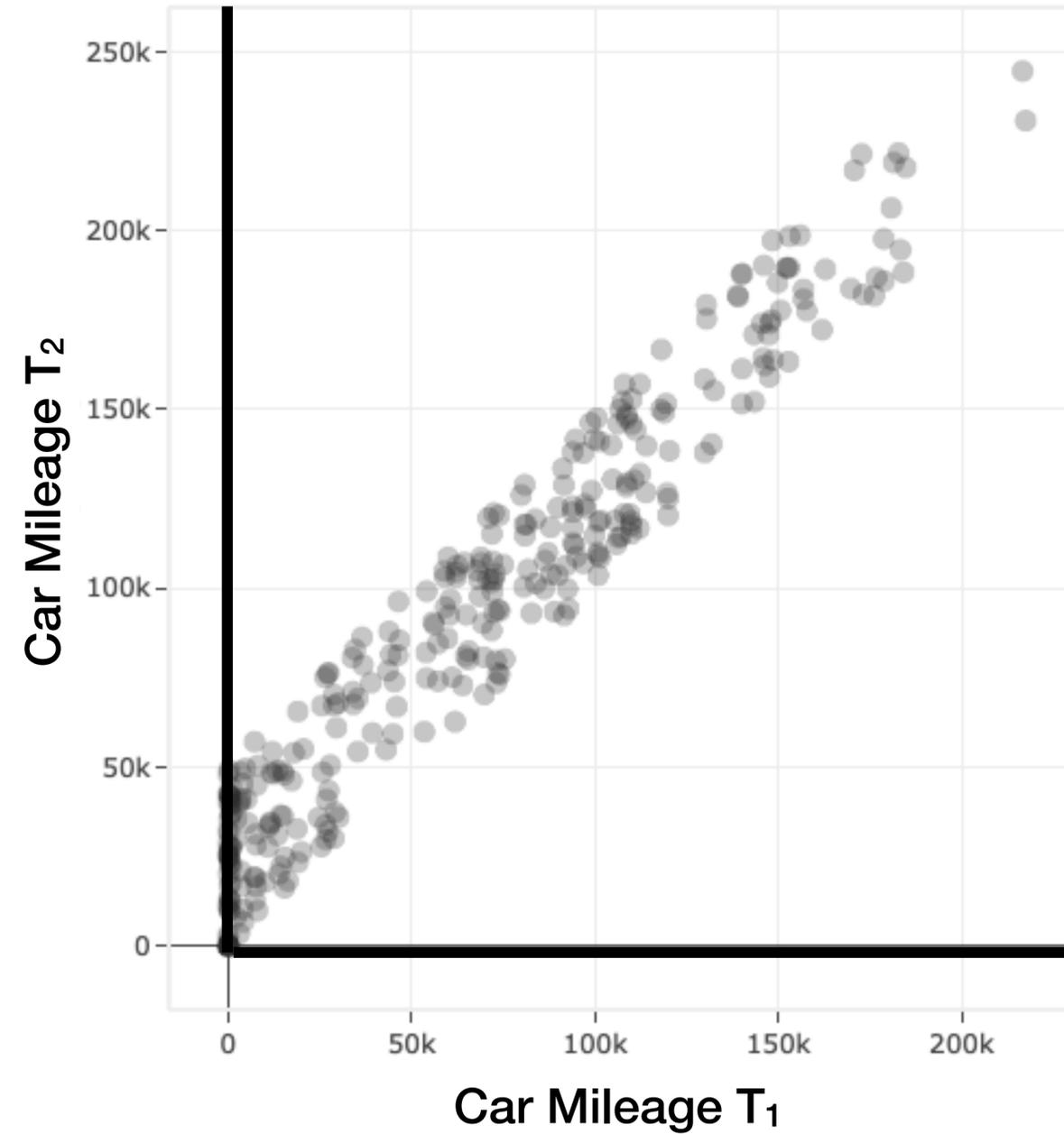


# Deviation from Domain Expectations

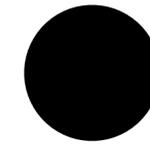


DS-DRIVING

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

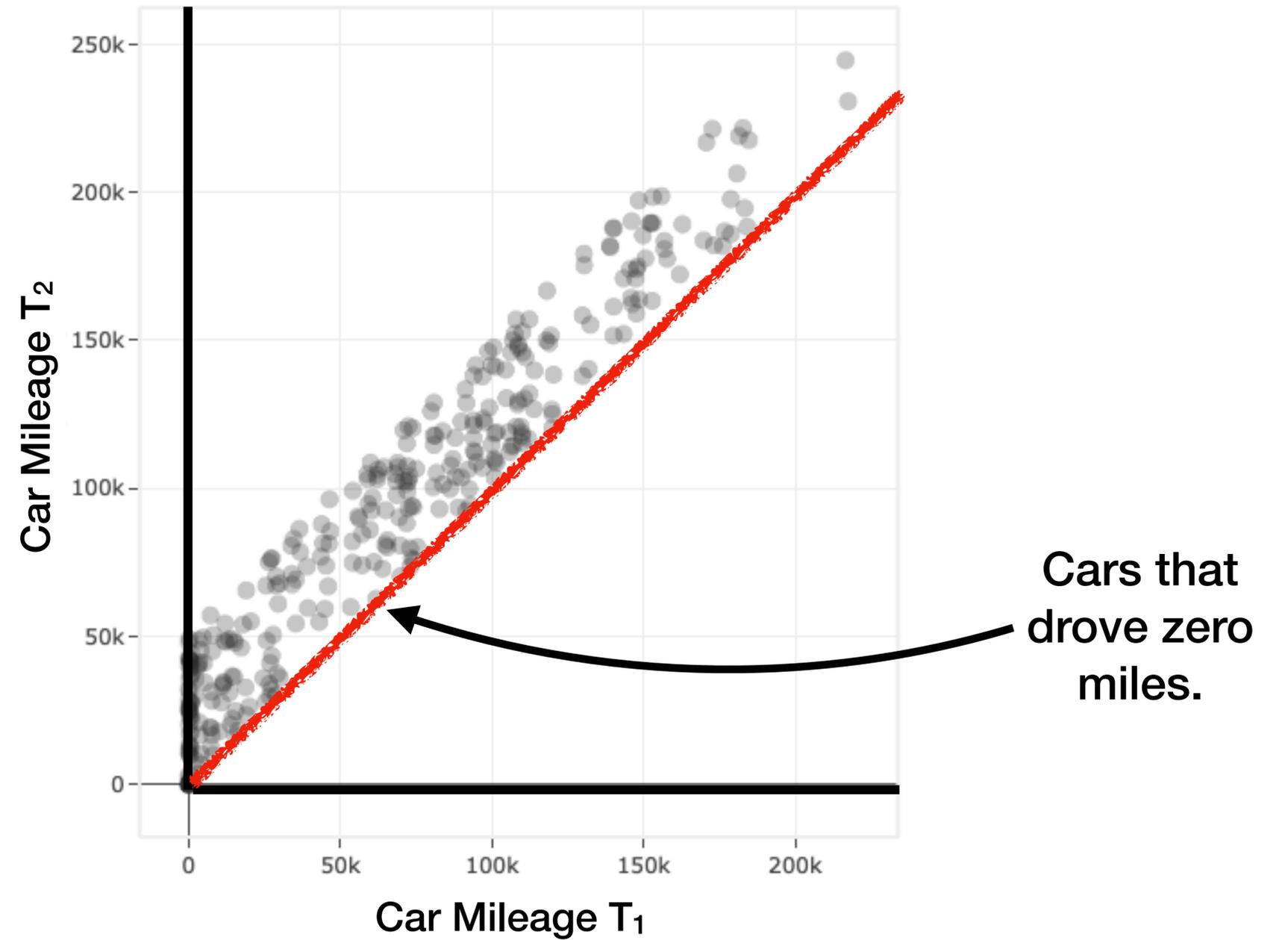


# Deviation from Domain Expectations

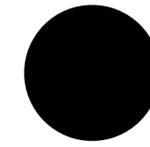


# DS-DRIVING

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

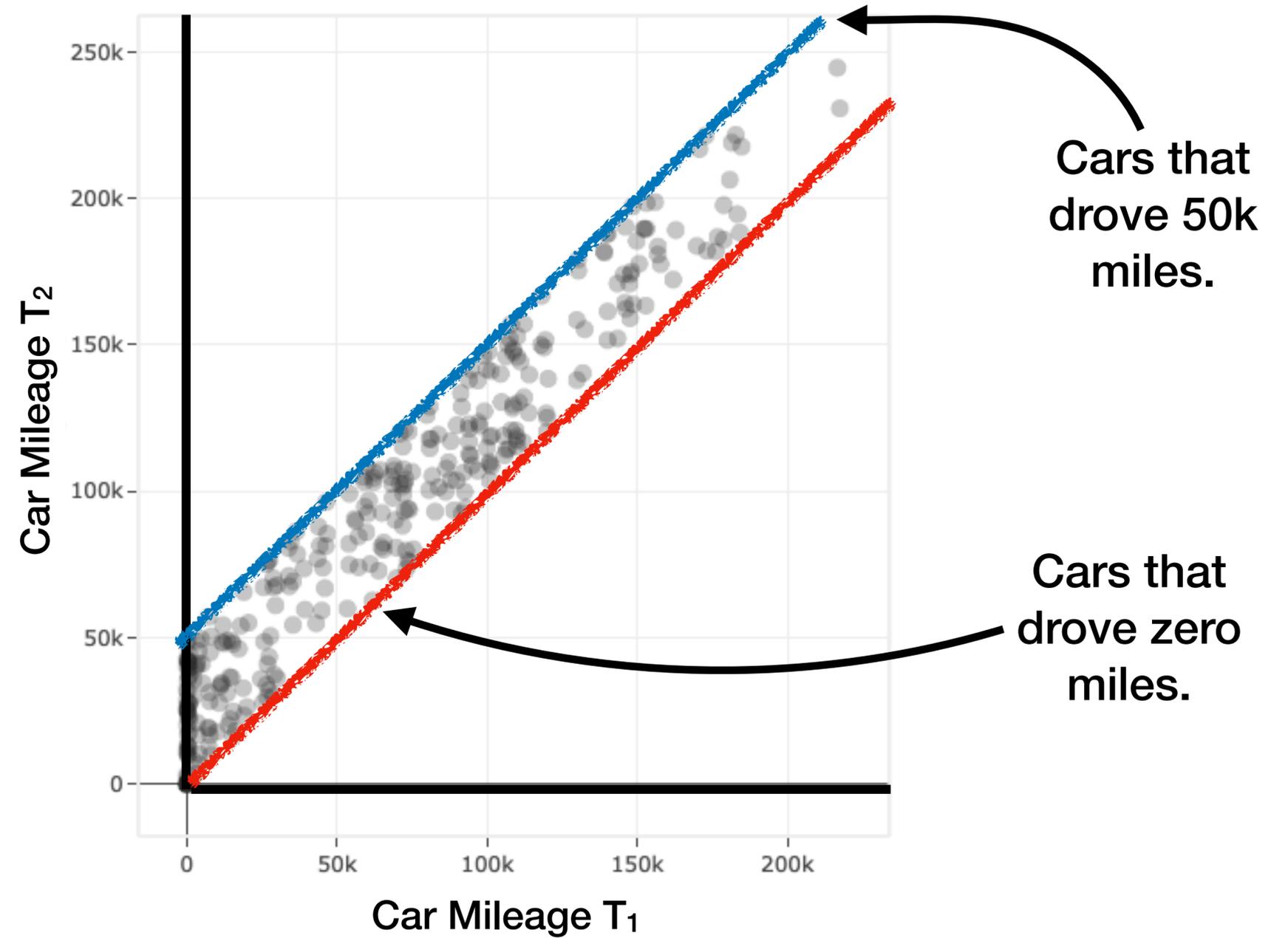


# Deviation from Domain Expectations

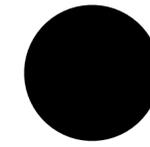


# DS-DRIVING

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

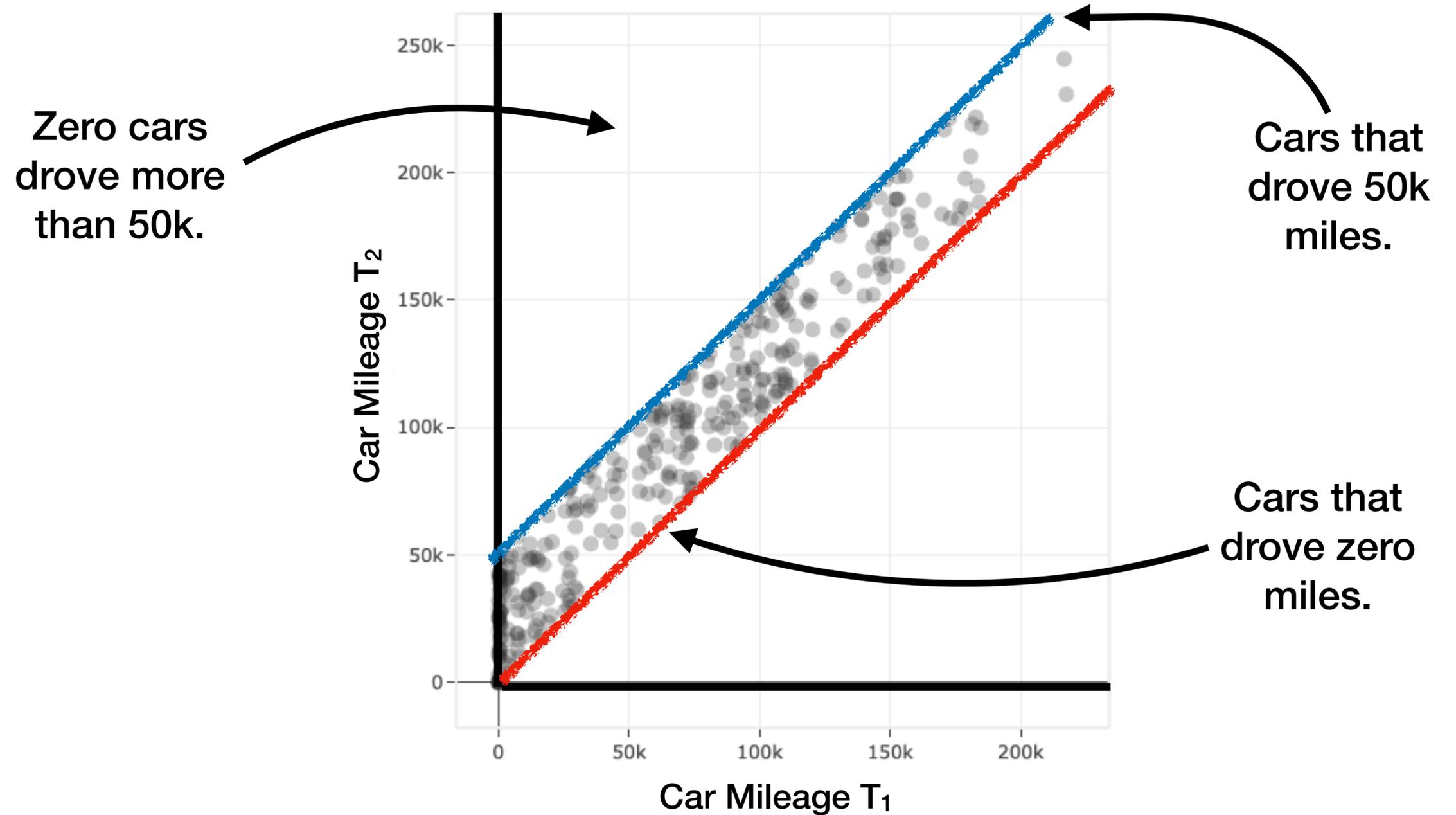


# Deviation from Domain Expectations



# DS-DRIVING

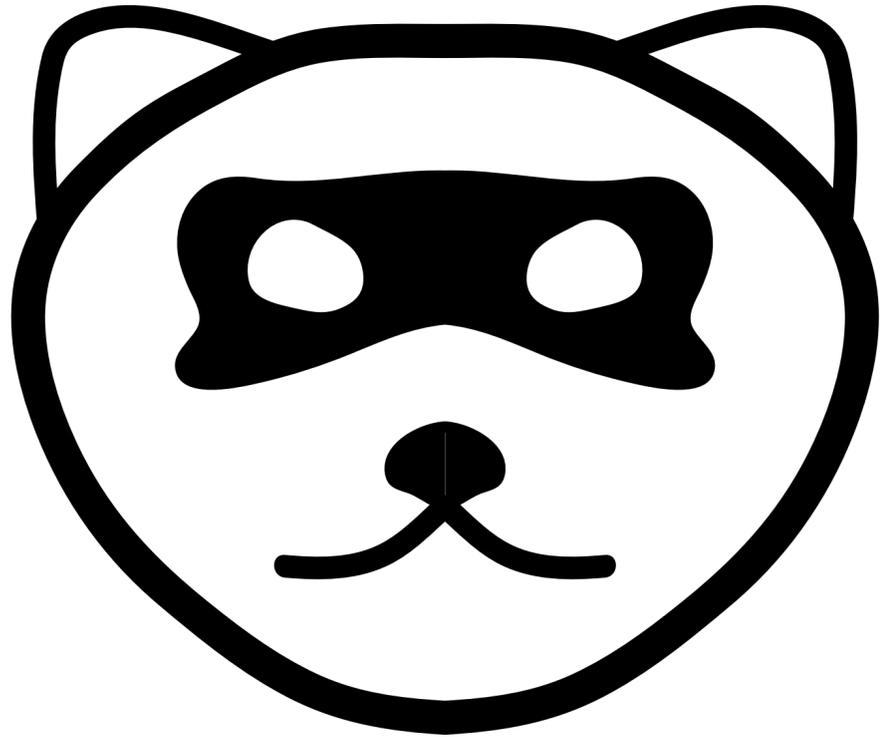
Height (cm)	Weight (kg)
175	79
155	93
182	65
202	40
165	105
169	87
159	73





**Implement Policies**

# Ethical Considerations

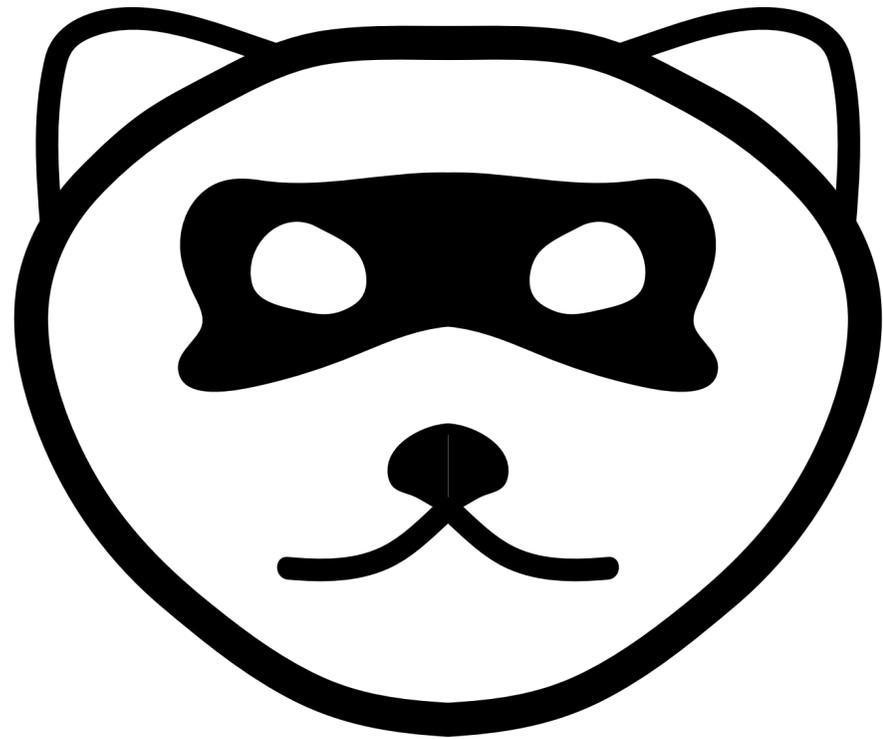


# Ethical Considerations



**False positives?**

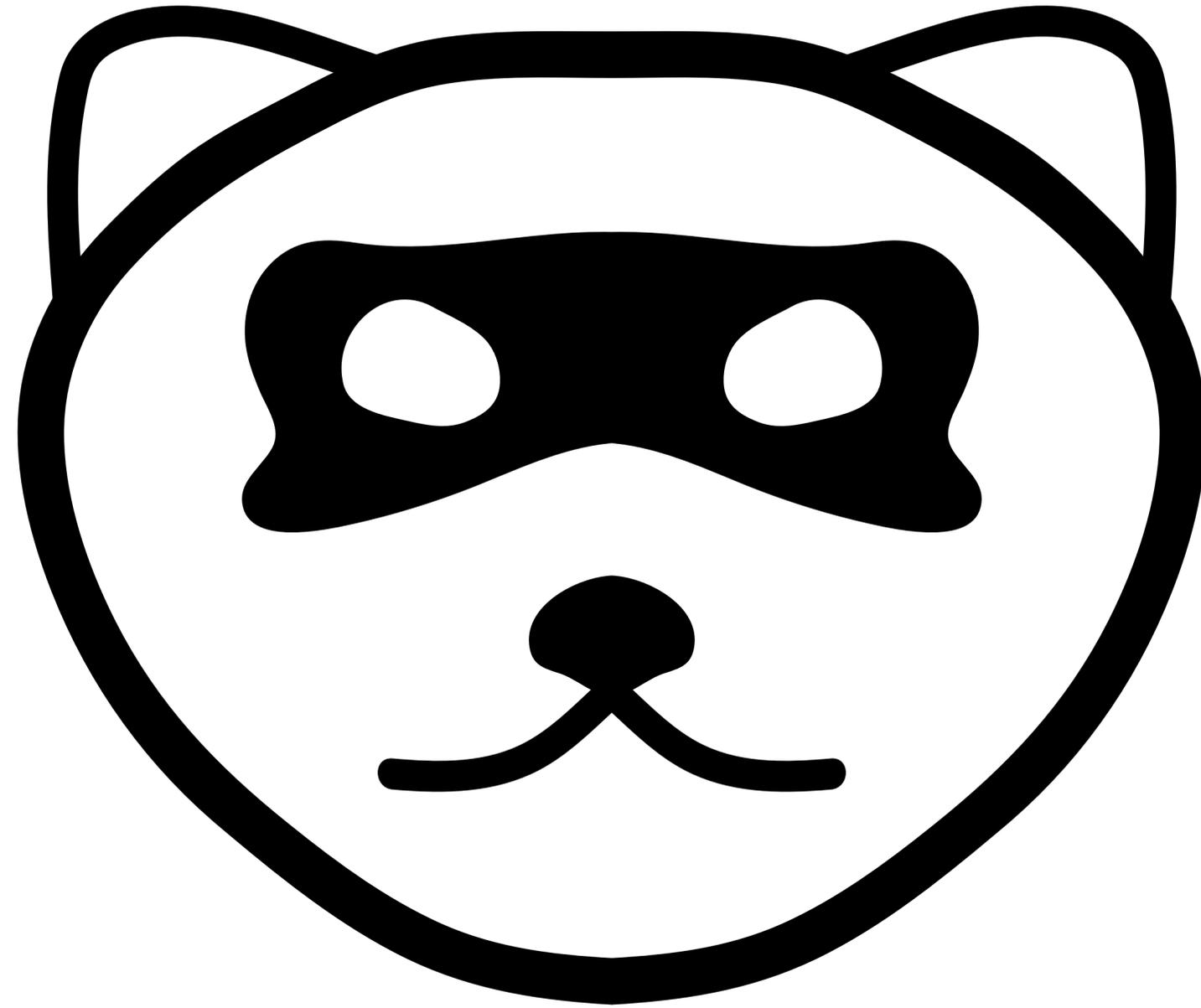
# Ethical Considerations



**False positives?**

**Use by bad actors?**

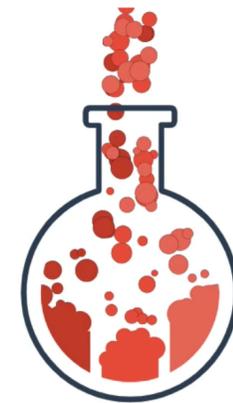
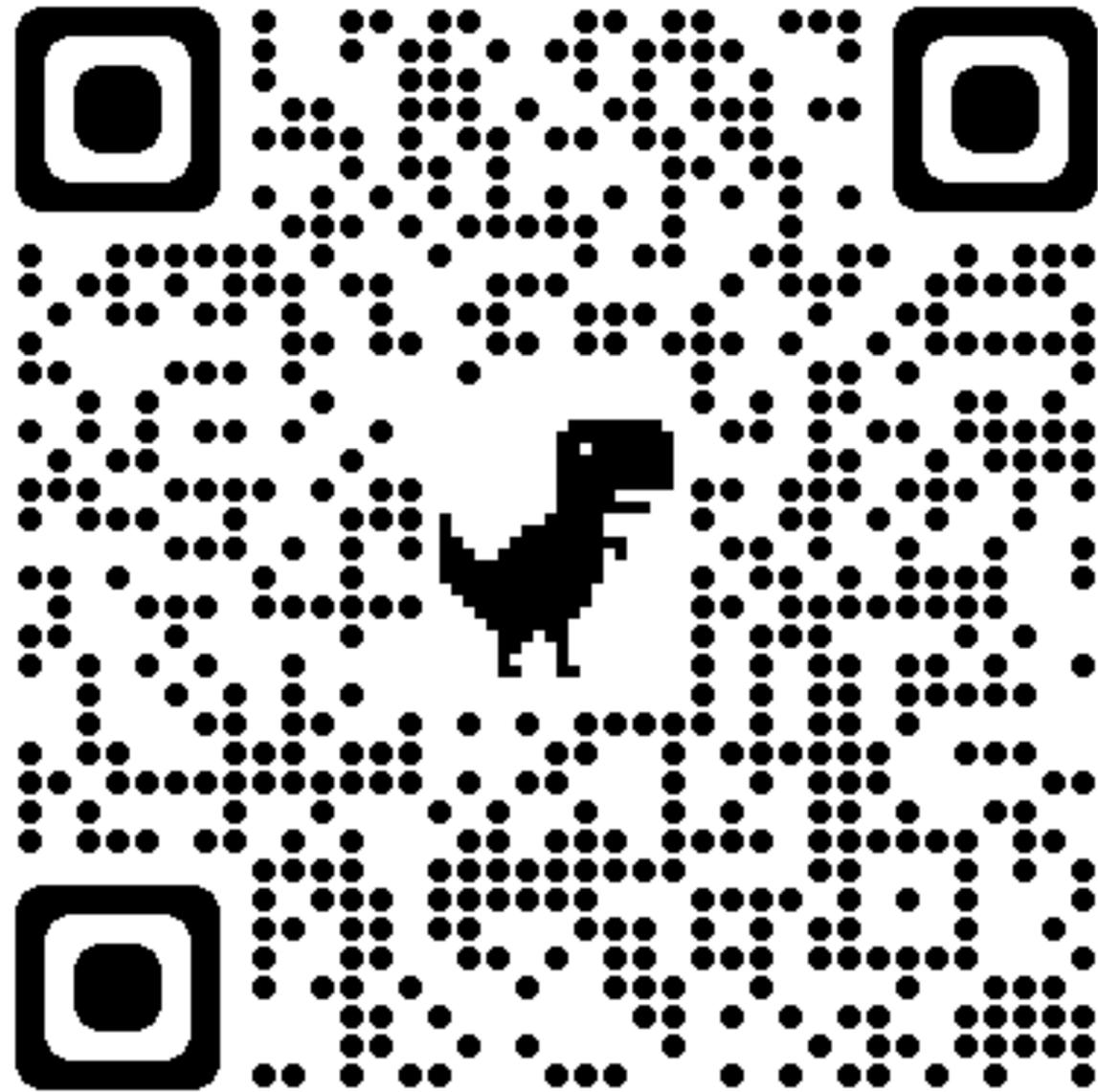
# Conclusion





# Ferret: Reviewing Tabular Datasets for Manipulation

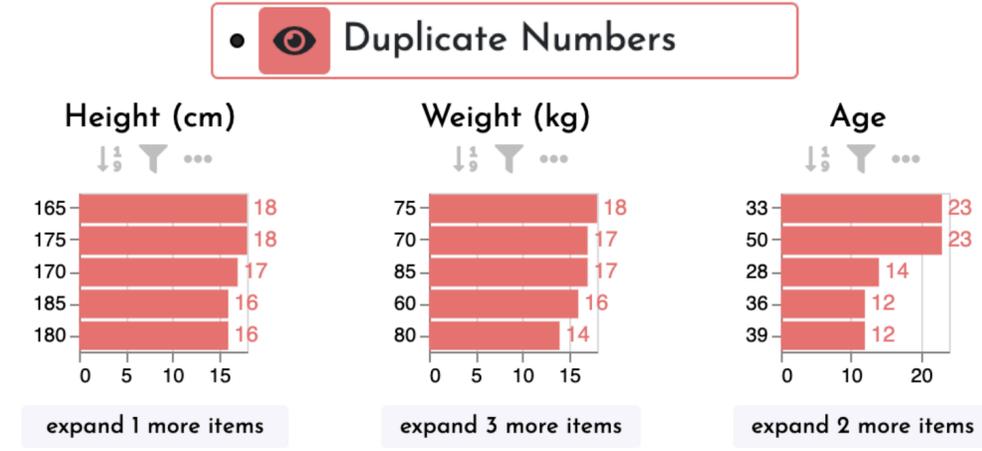
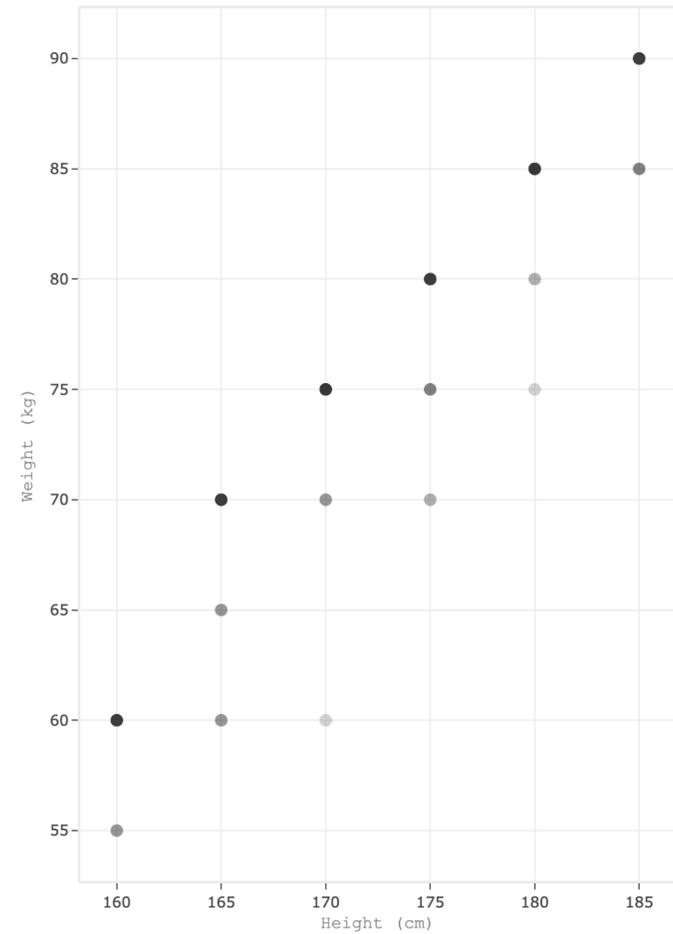
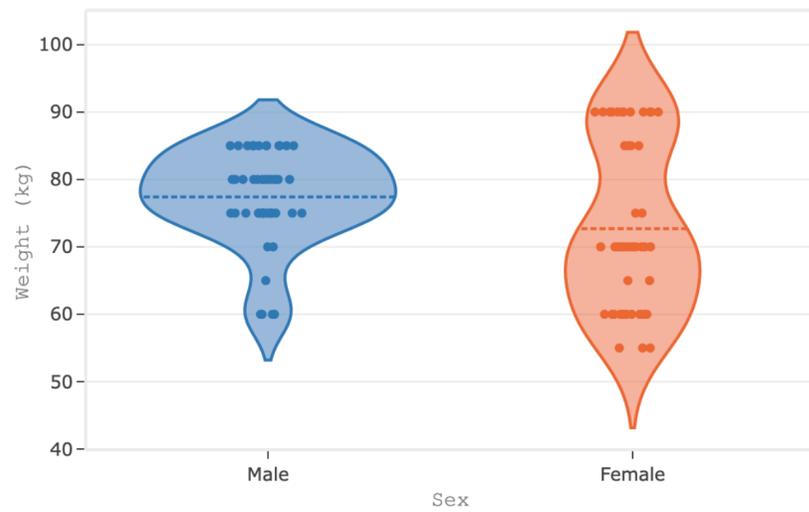
Devin Lange, Shaurya Sahai, Jeff M. Phillips, Alexander Lex



**visualization**  
**design lab**

# Appendix

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



Height (cm)	Weight (kg)	Age
175	70	33
180	75	36
165	60	39
170	70	33
185	80	36
160	65	39
175	75	33
180	80	36
165	70	39
170	75	33
185	85	36
160	55	39
165	65	33
170	70	36
175	75	39
180	80	33
185	85	36
160	60	39
165	70	33
170	75	36
175	80	39
180	85	33
185	90	36

# 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?**

# Datasets

## Artifacts of Manipulation

Formatting

UNEXPECTED FORMATTING

DS-PRIMING

DS-DRIVING

DS-COVID

DS-GAMIGN

DS-SPIDER-P

DS-SPIDER-E

DS-SPIDER-I

DS-GLIOMA

DS-FLY

DS-FISH



DUPLICATE NUMBERS AND DIGITS



Numerical

**UNEXPECTED LEADING DIGITS**

UNEXPECTED TRAILING DIGITS



UNEXPECTED VARIATION OF PRECISION



Structural

REPEATED REGIONS



ORDERING ARTIFACTS



Domain

DEVIATION FROM DOMAIN EXPECTATIONS



# 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>