Exploratory data analysis

Overview

1. Exploratory Data Analysis (EDA): what and why?

Overview

1. Exploratory Data Analysis (EDA): what and why?
2. Things to explore

Overview

1. Exploratory Data Analysis (EDA): what and why?
2. Things to explore
3. Exploration and visualization tools

Overview

1. Exploratory Data Analysis (EDA): what and why?
2. Things to explore
3. Exploration and visualization tools
4. (A bit of) dataset cleaning

Overview

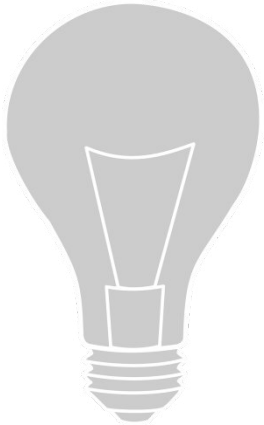
1. Exploratory Data Analysis (EDA): what and why?
2. Things to explore
3. Exploration and visualization tools
4. (A bit of) dataset cleaning
5. Kaggle competition EDA

Overview

1. Exploratory Data Analysis (EDA): what and why?
2. Things to explore
3. Exploration and visualization tools
4. (A bit of) dataset cleaning
5. Kaggle competition EDA

Exploratory Data Analysis (EDA)

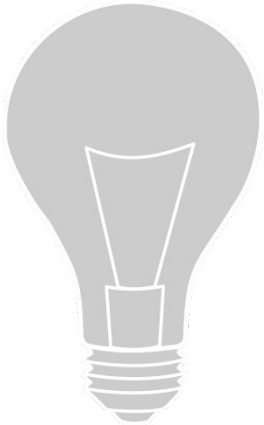
EDA allows to:



Exploratory Data Analysis (EDA)

EDA allows to:

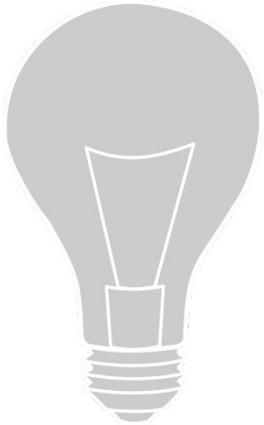
- Better understand the data



Exploratory Data Analysis (EDA)

EDA allows to:

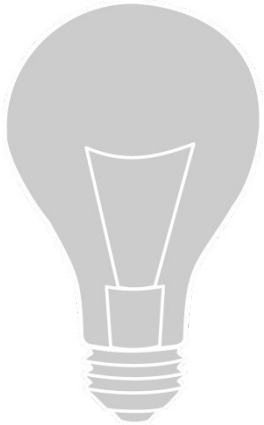
- Better understand the data
- Build an intuition about the data



Exploratory Data Analysis (EDA)

EDA allows to:

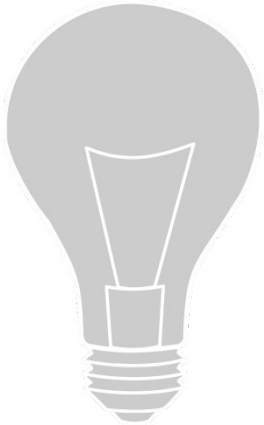
- Better understand the data
- Build an intuition about the data
- Generate hypothesizes



Exploratory Data Analysis (EDA)

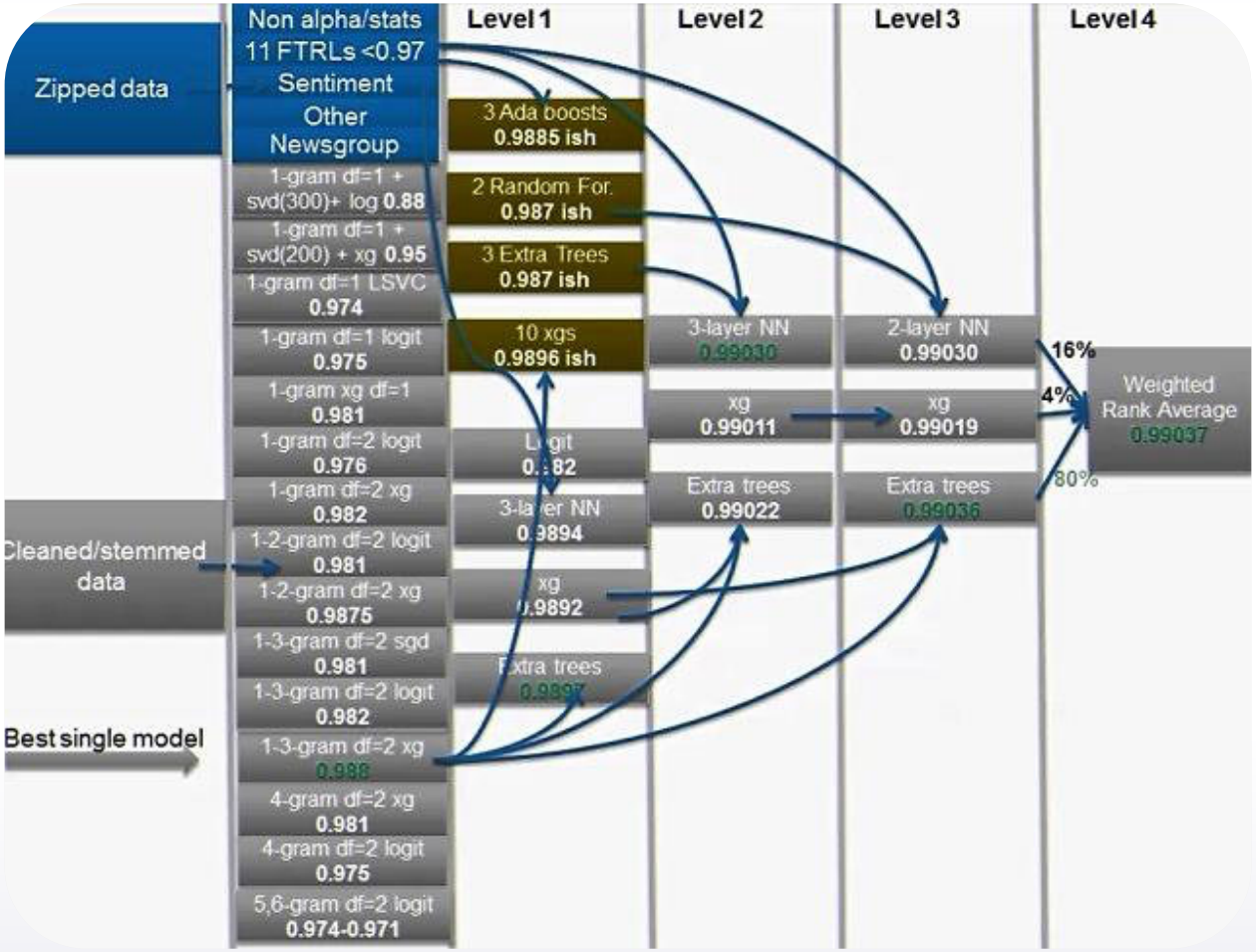
EDA allows to:

- Better understand the data
- Build an intuition about the data
- Generate hypothesizes
- Find insights



Exploratory Data Analysis (EDA)

- Please, do not start with stacking...



Kaggle Team, http://blog.kaggle.com/2015/12/03/dato-winners-interview-1st-place-mad-professors/

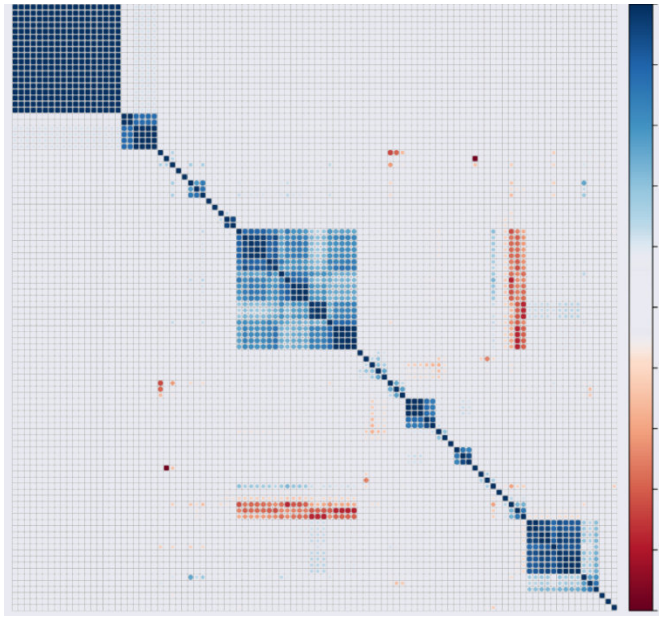
Visualizations

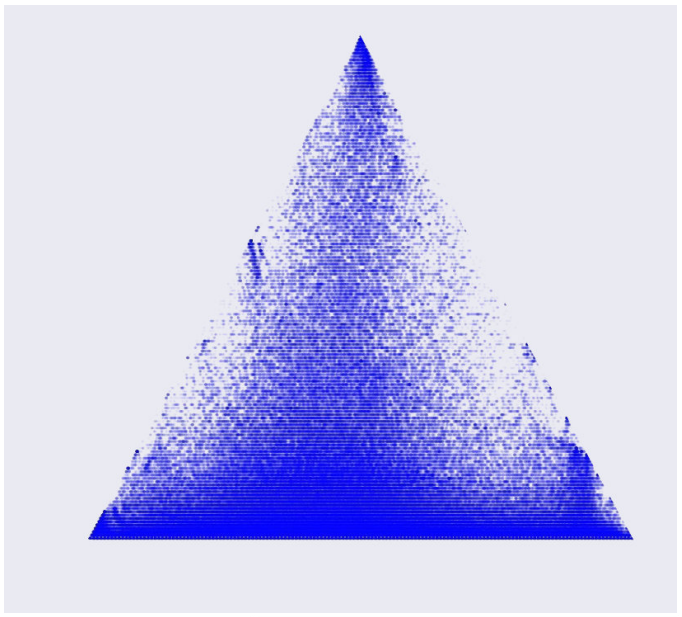
Visualization ⟶ Idea

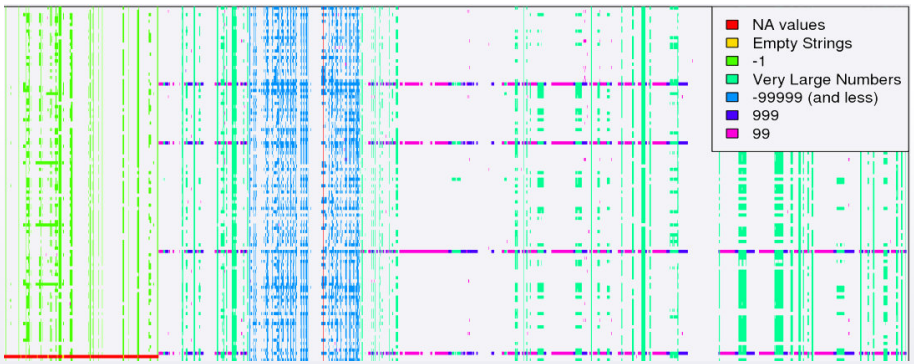
Patterns lead to questions

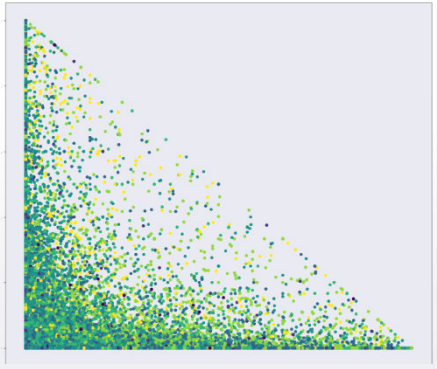
Idea ⟶ Visualization

Hypothesis testing









Motivating example

Alexander D'yakonov

Moscow, Russian Federation
Joined 7 years ago · last seen 21 days ago
http://alexanderdyakonov.narod.ru/english.htm

Followers 2

Competitions Grandmaster

Home | Competitions (36) | Kernels (1) | Discussion (104) | Followers (2) | Contact User | Follow User

Competitions Grandmaster

Current Rank	Highest Rank
199 of 60,591	1

9	14	4

Greek Media Monitoring M... · 3 years ago · Top 1%	1st	of 120
dunnhumby's Shopper Cha... · 6 years ago · Top 1%	1st	of 277
Large Scale Hierarchical Te... · 3 years ago · Top 2%	2nd	of 119

Kernels Contributor

Unranked

0	0	0

No kernel results

Discussion Contributor

Unranked

2	7	27

Code sharing · 3 years ago	21 votes
Thanks · 6 years ago	14 votes
congrats to the winners! · 2 years ago	10 votes

Motivating example

person id	person info	promo info	# promos sent	# promos used	*used this promo?*
14	...	...	13	4	**1**
3	...	...	43	35	**0**
0	...	..	6	0	**1**
32	...	...	15	13	**1**

Motivating example

id	...	# promos sent	# promos used	*diff*	*used this promo?*
13	...	0	0	1	**1**
13	...	1	1	0	**0**
13	...	2	1	1	**0**
13	...	4	2	1	**1**
13	...	5	3	1	**1**
13	...	6	3	NaN	**0**

1. For each person sort by '**# promos sent**'
2. Look at difference between consecutive rows in '**# promos used**' column ('***diff***' feature)

Conclusion

With EDA we can:

- get comfortable with the data
- find *magic features*

Do EDA first. Do not immediately dig into modelling.

In the following videos

1. Exploratory Data Analysis (EDA): what and why?
2. Things to explore
3. Exploration and visualization tools
4. (A bit of) dataset cleaning
5. Kaggle competition EDA

Building intuition about the data

Video overview

1. Getting domain knowledge
2. Checking if the data is intuitive
3. Understanding how the data was generated

Get domain knowledge

Passenger Screening Algorithm Challenge — $1,500,000
Improve the accuracy of the Department of Homeland Security's threat recognition algorithms — 96 teams
Featured · 5 months to go

Zillow Prize: Zillow's Home Value Prediction (Zestimate) — $1,200,000
Can you improve the algorithm that changed the world of real estate? — 1,489 teams
Featured · 6 months to go

Planet: Understanding the Amazon from Space — $60,000
Use satellite data to track the human footprint in the Amazon rainforest — 875 teams
Featured · 7 days to go

Instacart Market Basket Analysis — $25,000
Which products will an Instacart consumer purchase again? — 1,427 teams
Featured · a month to go

Get domain knowledge, example

Task: Predict advertiser's cost

Data:

AdGroupId	AdNetwork Type2	MaxCpc	Slot	Clicks	Impressions	…
78db044136	s	0.28	s_2	3	0	…
68a0110c33	s	1	s_2	1	13	…
2r39fw11w3	p	1.2	p_1	3	419	…

Check if the data is intuitive

...	*Age*	...
...	21	...
...	45	...
...	336	...
...	19	...
...	...	...

Check if the data is intuitive

…	*Age*	…
…	21	…
…	45	…
…	336	…
…	19	…
…	…	…

- Is *336* a typo?

Check if the data is intuitive

...	*Age*	...
...	21	...
...	45	...
...	336	...
...	19	...
...	...	...

- Is *336* a typo?
- Or we misinterpret the feature and age 336 is normal

Check if the data is intuitive

Task: Predict advertiser's cost

Data:

AdGroupId	AdNetwork Type2	MaxCpc	Slot	Clicks	Impressions	…
78db044136	s	0.28	s_2	*3*	*0*	…
68a0110c33	s	1	s_2	1	13	…
2r39fw11w3	p	1.2	p_1	3	419	…

Check if the data is intuitive

Task: Predict advertiser's cost

Data:

AdGroupId	AdNetwork Type2	MaxCpc	Slot	Clicks	Impressions	*is_incorrect*
78db044136	s	0.28	s_2	*3*	*0*	***True***
68a0110c33	s	1	s_2	1	13	***False***
2r39fw11w3	p	1.2	p_1	3	419	***False***

Understand how the data was generated

Understand how the data was generated

It is crucial to understand the generation process to set up a proper validation scheme

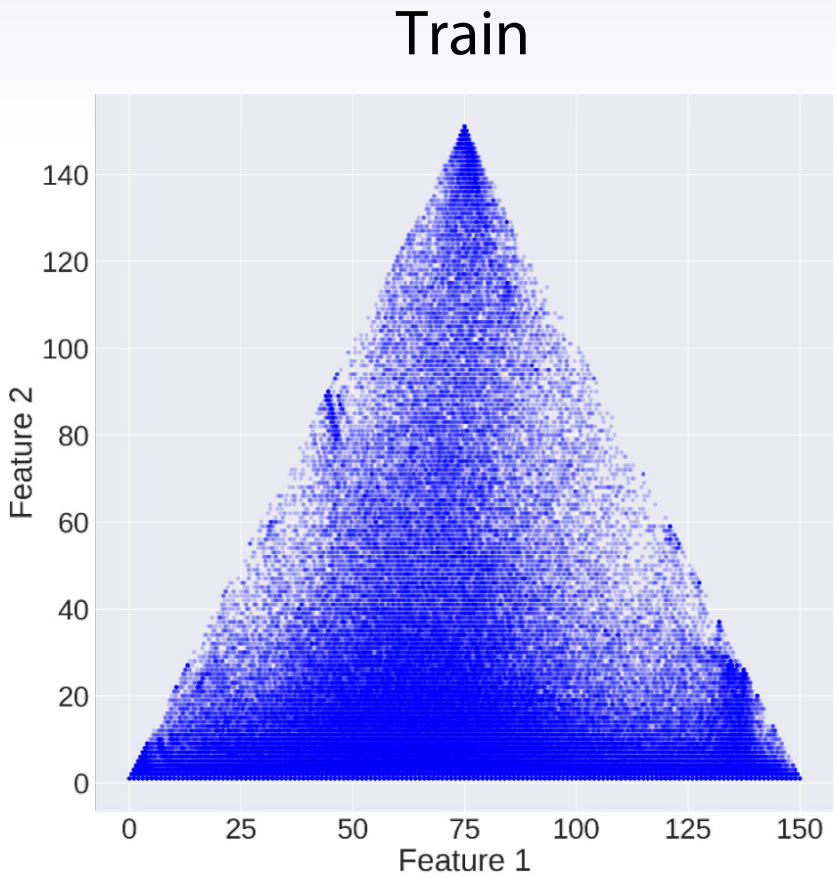
Check if the data is intuitive

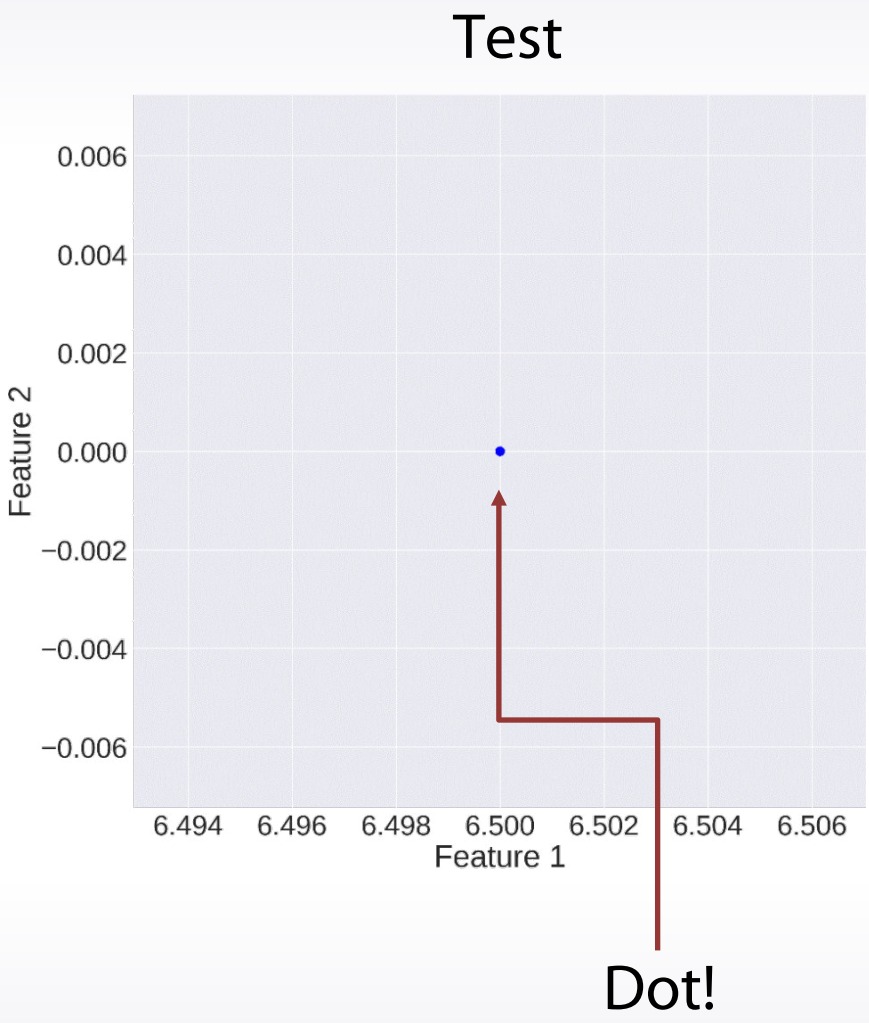
Task: Predict advertiser's cost

Data:

AdGroupId	AdNetwork Type2	MaxCpc	Slot	*Clicks*	*Impressions*	*is_incorrect*
78db044136	s	0.28	s_2	*3*	*0*	*True*
68a0110c33	s	1	s_2	1	13	*False*
2r39fw11w3	p	1.2	p_1	3	419	*False*

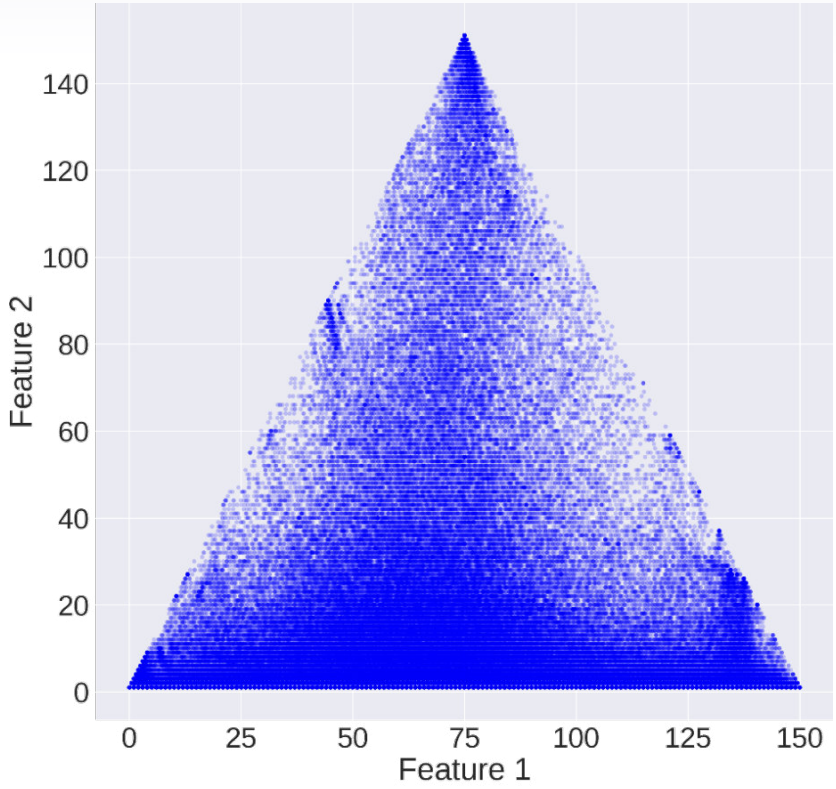
Understand how the data was generated



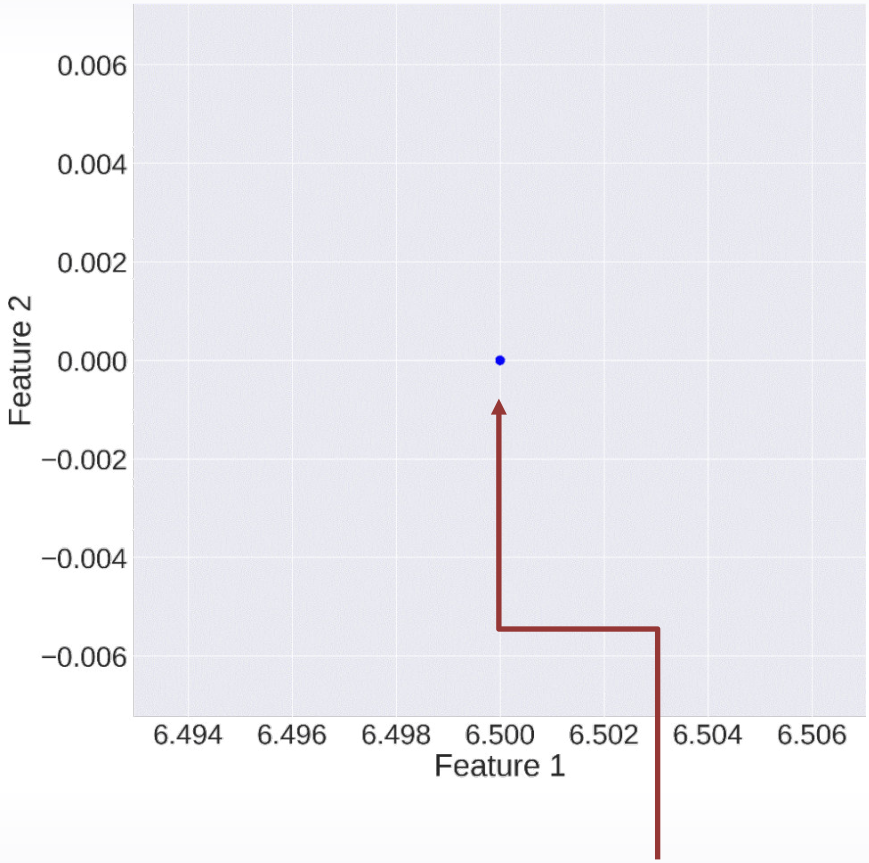


Understand how the data was generated

Train



Test



#days in *train* > #days in *test*

#rows in *train* < #rows in *test*

Conclusion

Conclusion

- **Get domain knowledge**
  - It helps to deeper understand the problem

Conclusion

- **Get domain knowledge**
  - It helps to deeper understand the problem
- **Check if the data is intuitive**
  - And agrees with domain knowledge

Conclusion

- **Get domain knowledge**
  - It helps to deeper understand the problem
- **Check if the data is intuitive**
  - And agrees with domain knowledge
- **Understand how the data was generated**
  - As it is crucial to set up a proper validation

Exploring anonymized data

Video overview

1. What is anonymized data?
2. What can we do with it?

Anonymized data

Anonymized data

Text	Encoded text
I want this table	7ugy 972h 98ww hj34
Table is what I want	hj34 4f08 rtte 7ugy 972h
This table is red	98ww hj34 4f08 4rj9
And this is me	jk8r 98ww 4f08 9jo4

Anonymized data

id	x1	x2	x3	x4	x5	x6
1	m268i97y	0	NO	105.4	14	
2	j0gheu6	1	YES	25.631	12	
3	26fmsp6u	1	NO	12.0	12	m268i97y
4	13e5dpzp	0	NO	140.12	14	m268i97y

Anonymized data

id	x1	x2	x3	x4	x5	x6
1	m268i97y	0	NO	105.4	14	
2	j0gheu6	1	YES	25.631	12	
3	26fmsp6u	1	NO	12.0	12	m268i97y
4	13e5dpzp	0	NO	140.12	14	m268i97y

- Explore individual features
  - Guess the meaning of the columns
  - Guess the types of the column
- Explore feature relations
  - Find relations between pairs
  - Find feature groups

Anonymized data

id	x1	x2	x3	x4	x5	x6
1	m268i97y	0	NO	105.4	14	
2	j0gheu6	1	YES	25.631	12	
3	26fmsp6u	1	NO	12.0	12	m268i97y
4	13e5dpzp	0	NO	140.12	14	m268i97y

- Explore individual features
  - Guess the meaning of the columns
  - Guess the types of the column
- Explore feature relations
  - Find relations between pairs
  - Find feature groups

Notebook

Exploring individual features: guessing types

id	x1	x2	x3	x4	x5	x6
1	m268i97y	0	NO	105.4	14	
2	j0gheu6	1	YES	25.631	12	
3	26fmsp6u	1	NO	12.0	12	m268i97y
4	13e5dpzp	0	NO	140.12	14	m268i97y

Exploring individual features: guessing types

id	x1	x2	x3	x4	x5	x6
1	m268i97y	0	NO	105.4	14	
2	j0gheu6	1	YES	25.631	12	
3	26fmsp6u	1	NO	12.0	12	m268i97y
4	13e5dpzp	0	NO	140.12	14	m268i97y

Helpful functions:

```
df.dtypes
df.info()
x.value_counts()
x.isnull()
```

Conclusion

- Two things to do with anonymized features:
  - **Try to decode the features**
    - Guess the true meaning of the feature
  - **Guess the feature types**
    - Each type needs its own preprocessing

Visualizations

Video overview

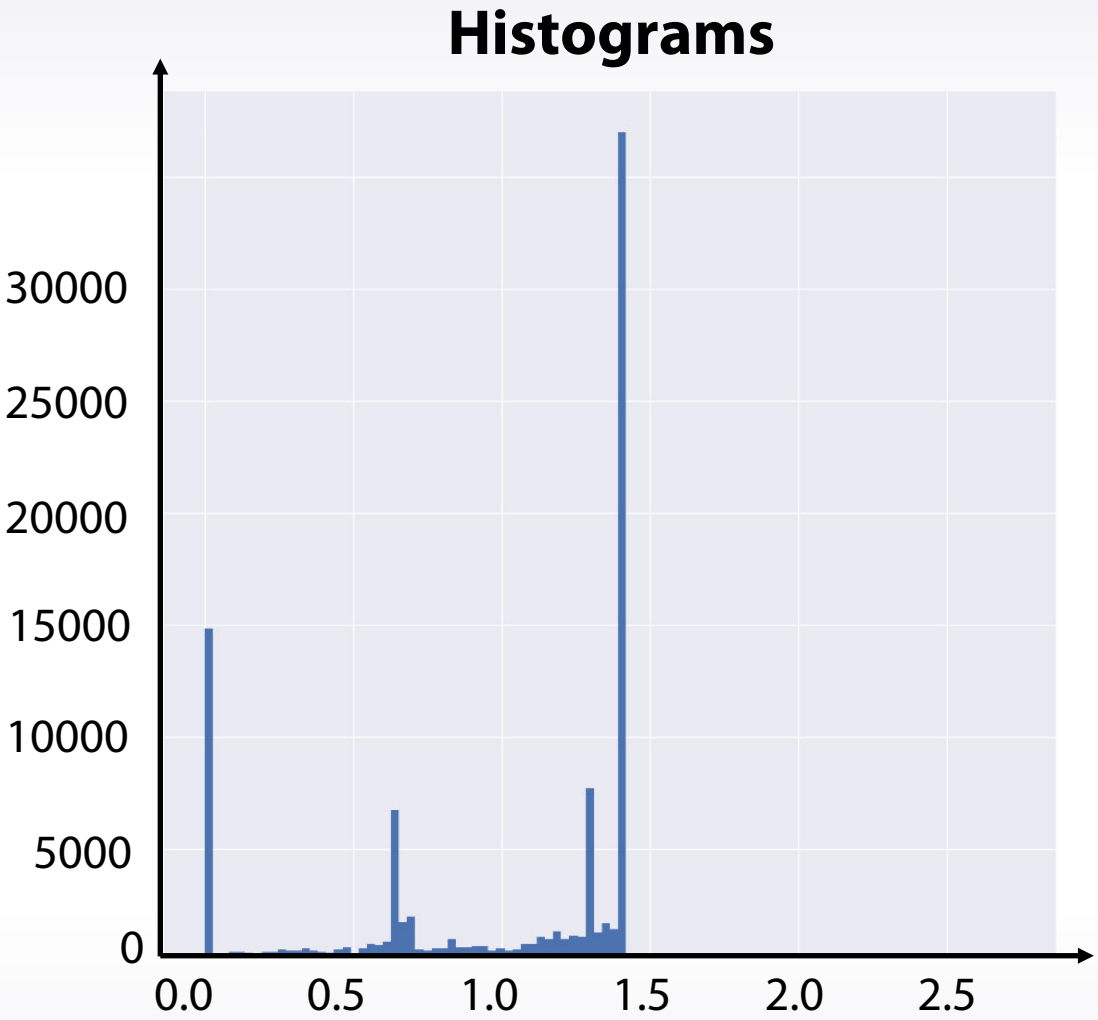
- Visualization tools to…
  - **Explore individual features**
    - Histograms
    - Plots
    - Statistics
  - **Explore feature relations**
    - Scatter plots
    - Correlation plots
    - Plot (index vs feature statistics)
    - And more

Visualizations

EDA is an art!

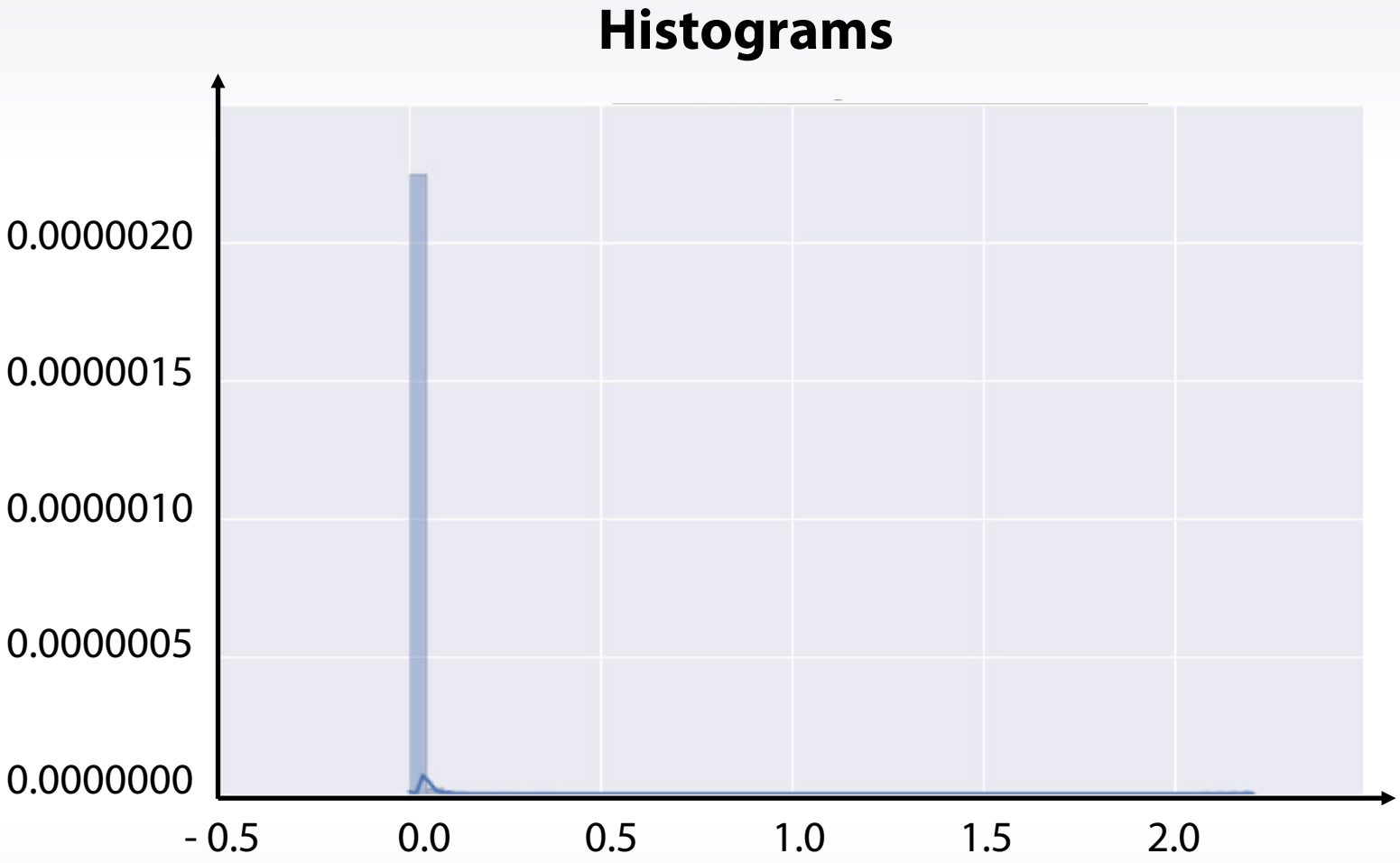
And visualizations are our art tools

Art tools



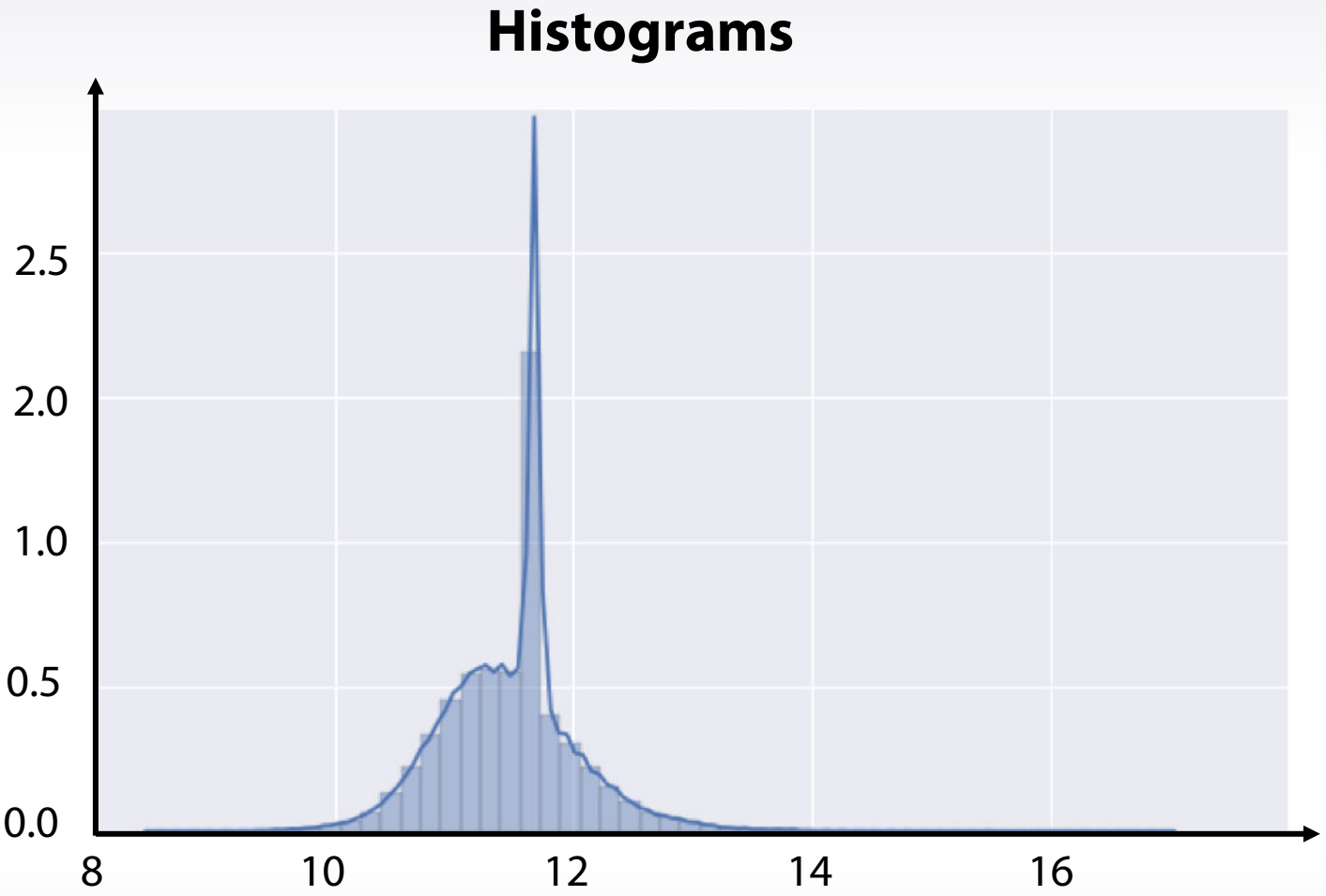

```
plt.hist(x)
```

Art tools



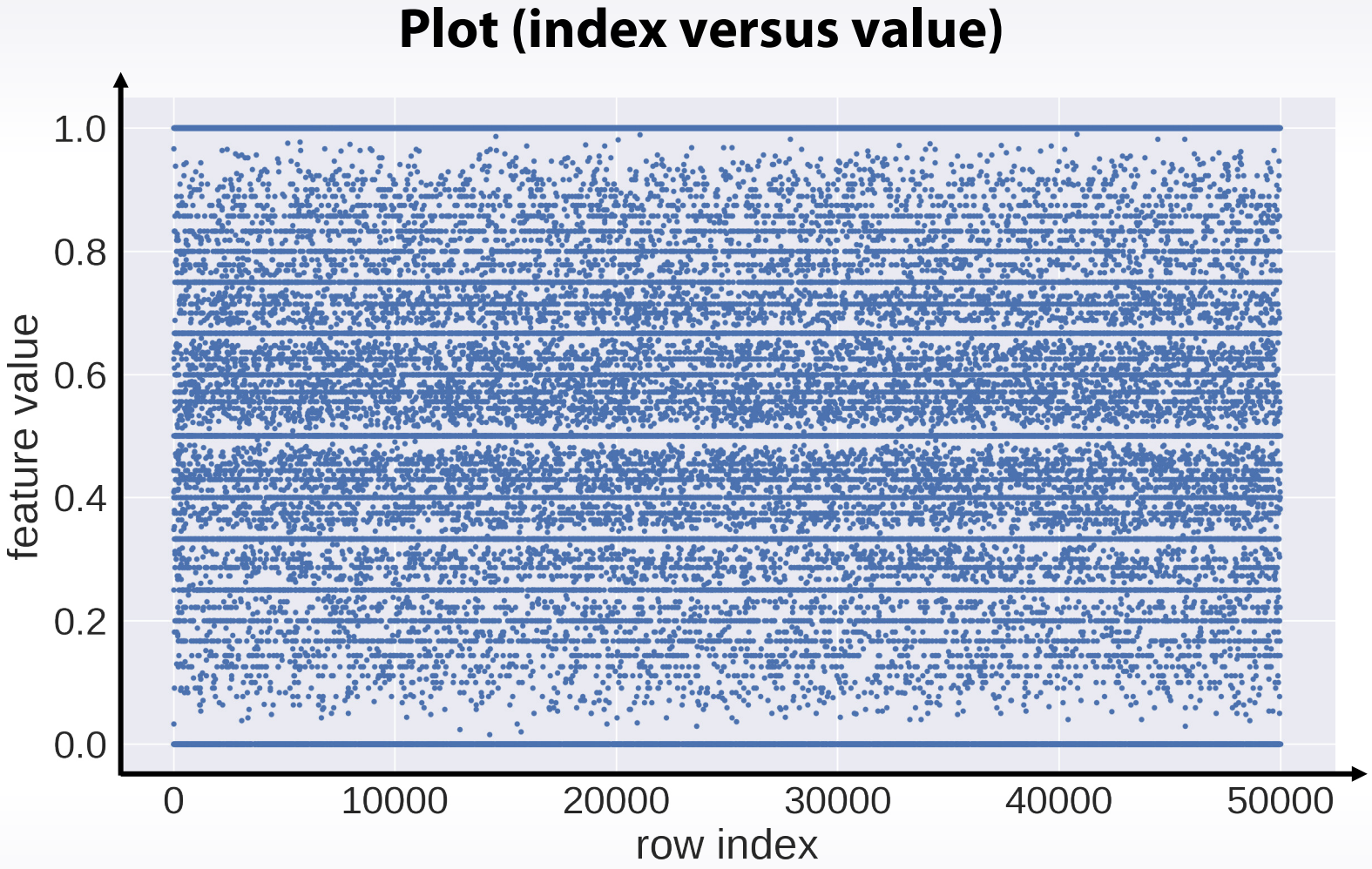

```
plt.hist(x)
```

Art tools



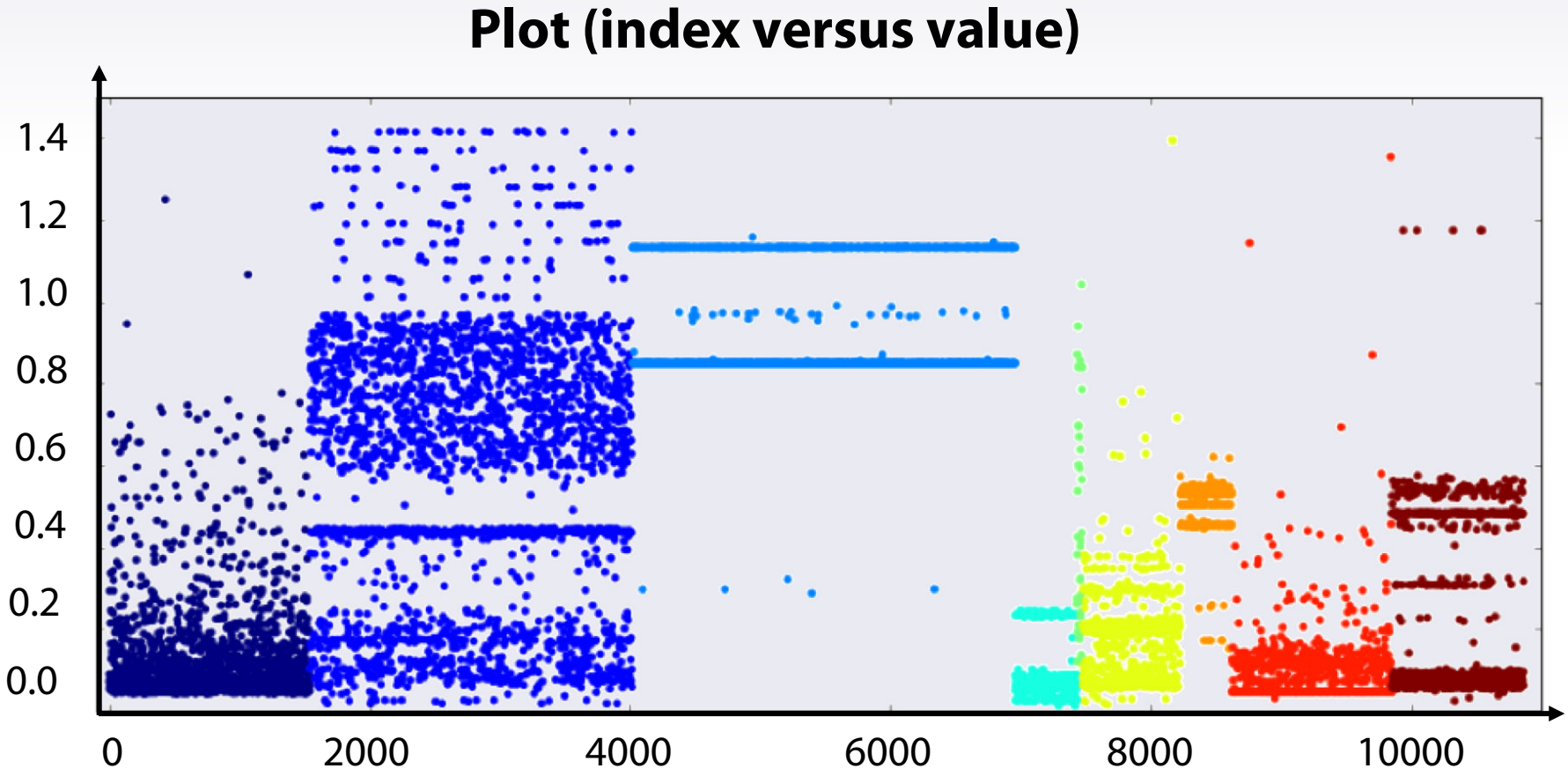

```
plt.hist(x)
```

Art tools




```
plt.plot(x,'.')
```

Art tools




```
plt.scatter(range(len(x)), x, c=y)
```

Art tools

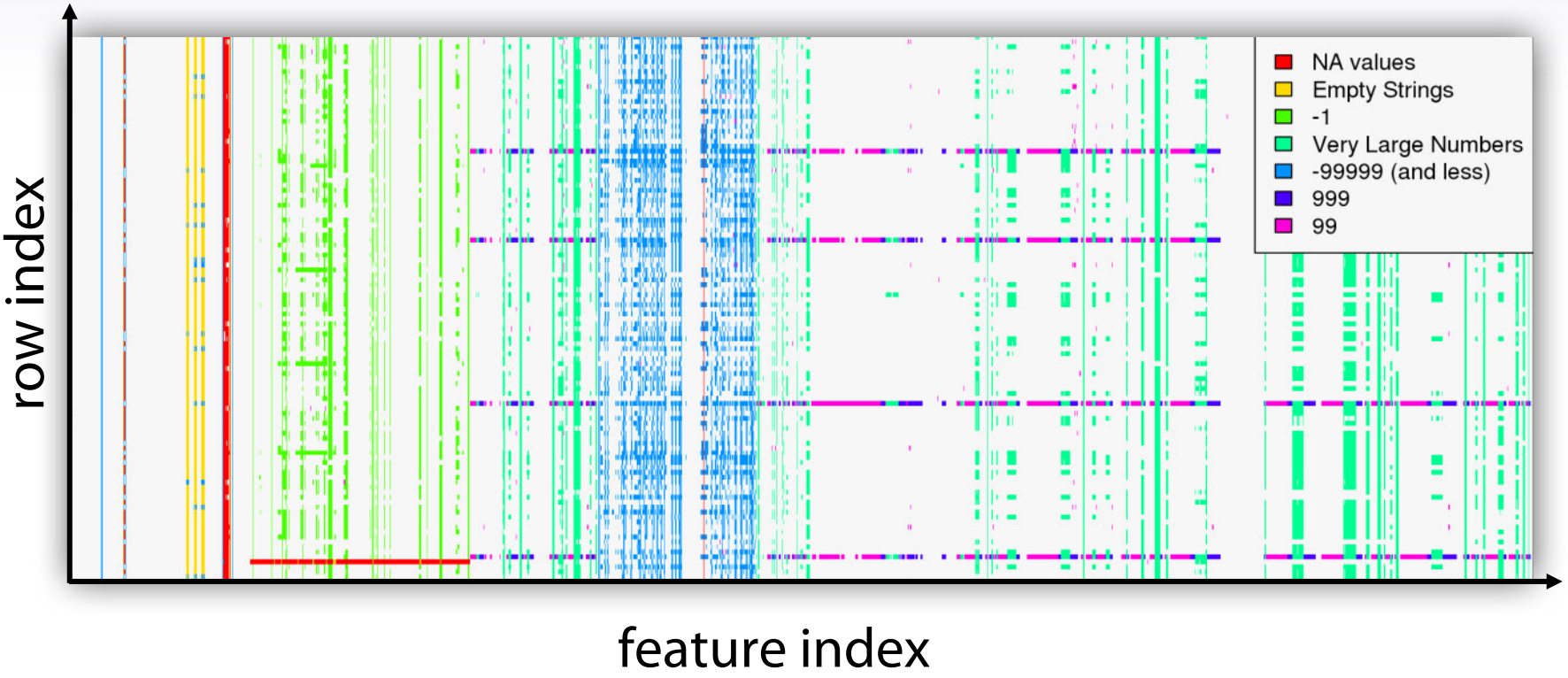
Feature statistics

	x6	x7	x8	x13
count	50000.00000	50000.000000	48793.000000	45512.000000
mean	0.99296	0.975860	-0.000252	4428.915253
std	0.08361	0.153485	1.023282	10943.884658
min	0.00000	0.000000	-85.252444	-99.000000
25%	1.00000	1.000000	-0.255490	0.000000
50%	1.00000	1.000000	0.221047	1817.000000
75%	1.00000	1.000000	0.567620	5582.000000
max	1.00000	1.000000	3.426844	776759.000000

```
df.describe()
x.mean()
x.var()
```

Art tools

Other tools




```
x.value_counts()
x.isnull()
```

Tools for individual features exploration

Histograms:

```
plt.hist(x)
```

Plot (index versus value):

```
plt.plot(x, '.')
```

Statistics:

```
df.describe()
x.mean()
x.var()
```

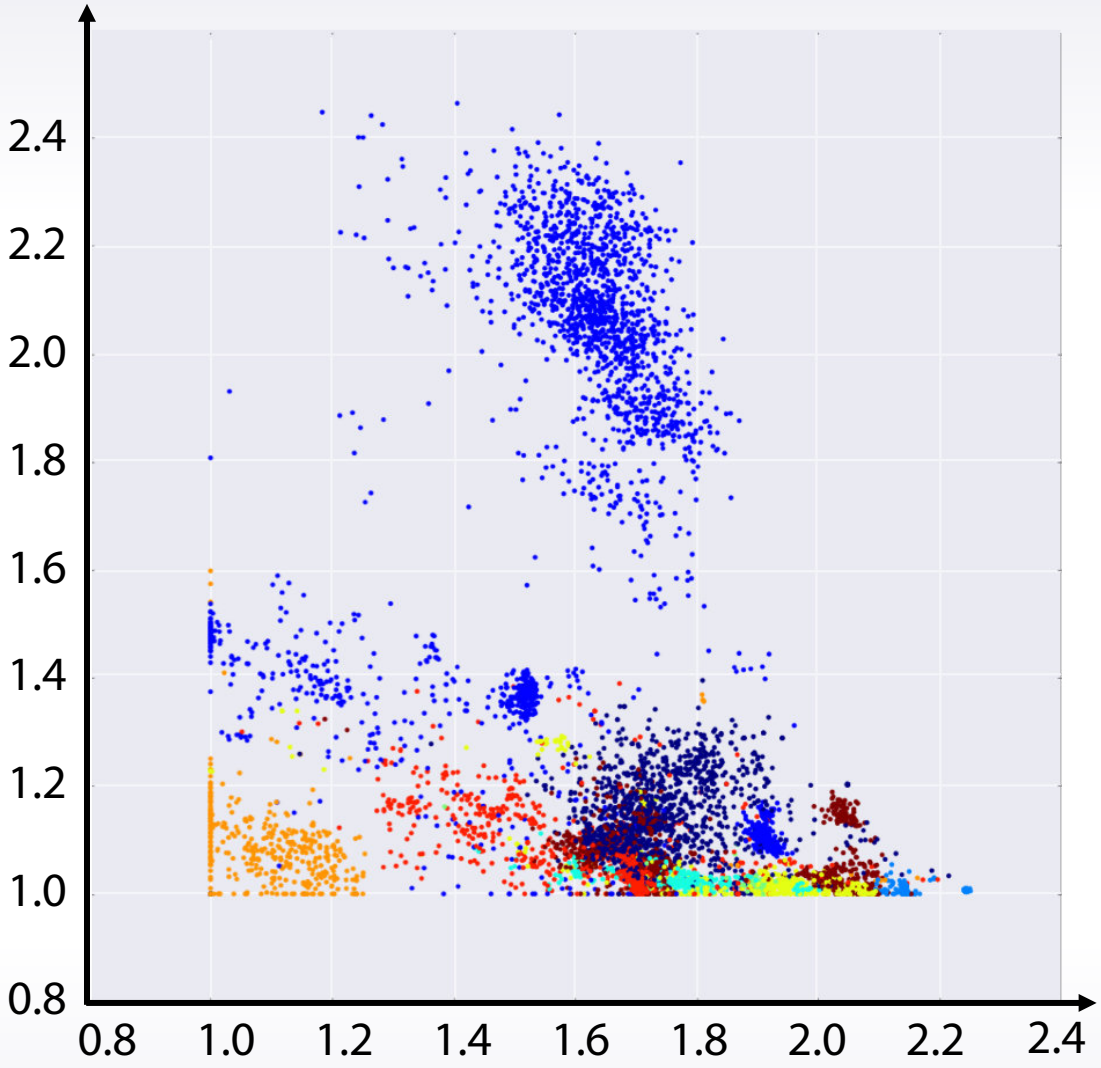
Other tools:

```
x.value_counts()
x.isnull()
```

Next in this video

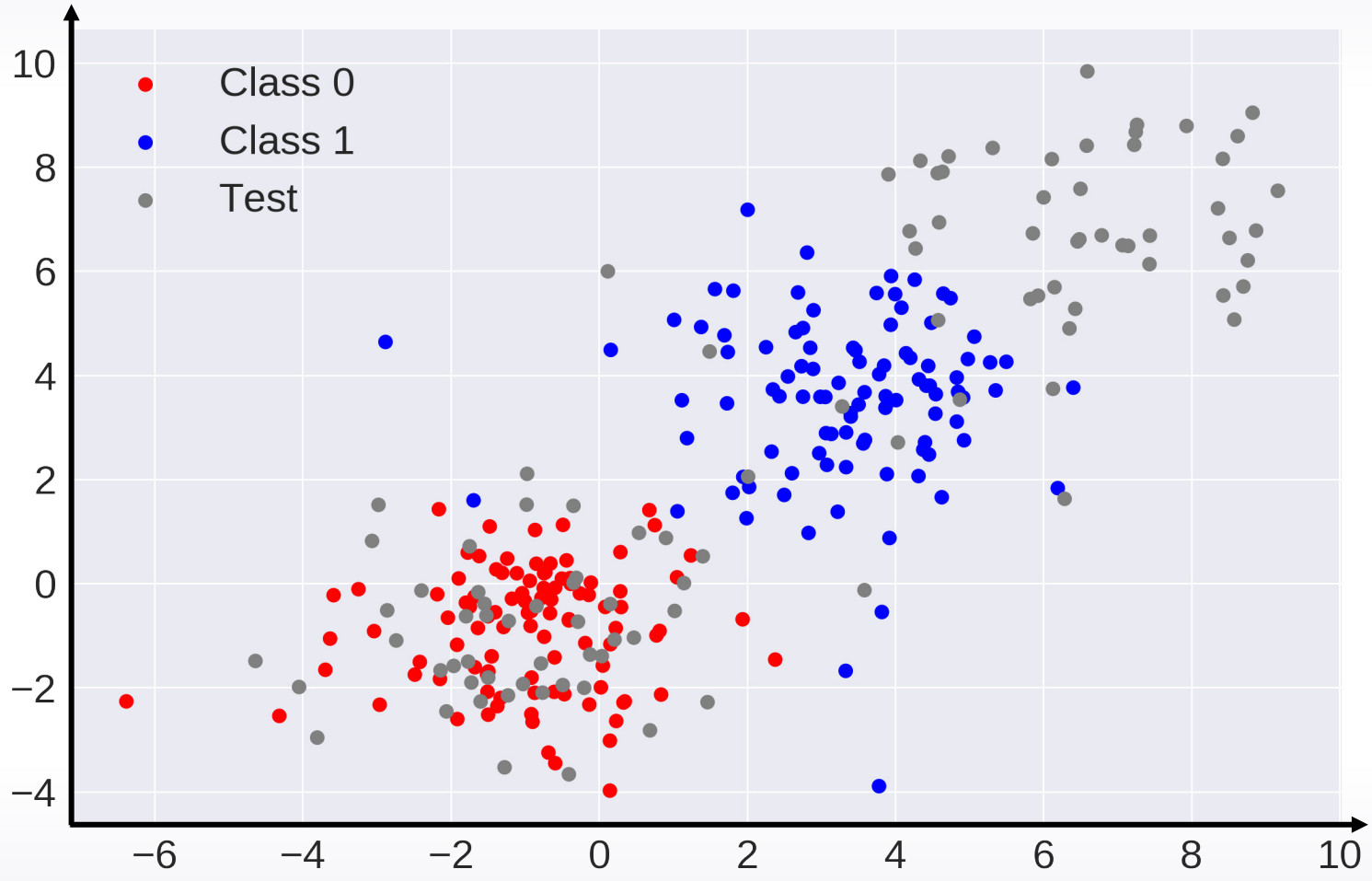
- Visualization tools to…
  - **Explore individual features**
    - Histograms
    - Plots
    - Statistics
  - **Explore feature relations**
    - Scatter plots
    - Correlation matrices
    - …

Exploring feature relations



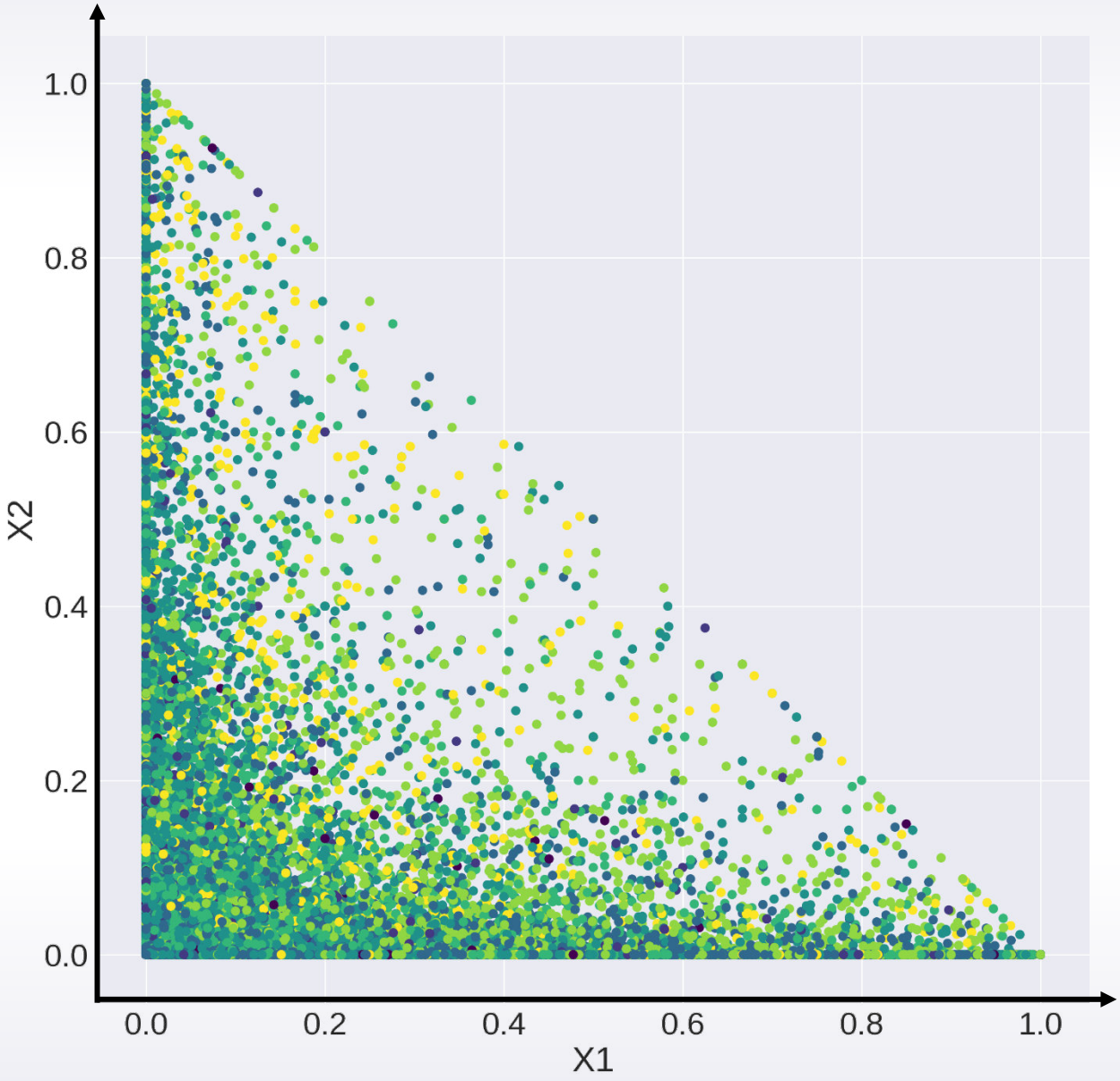

```
plt.scatter(x1, x2)
```

Exploring feature relations

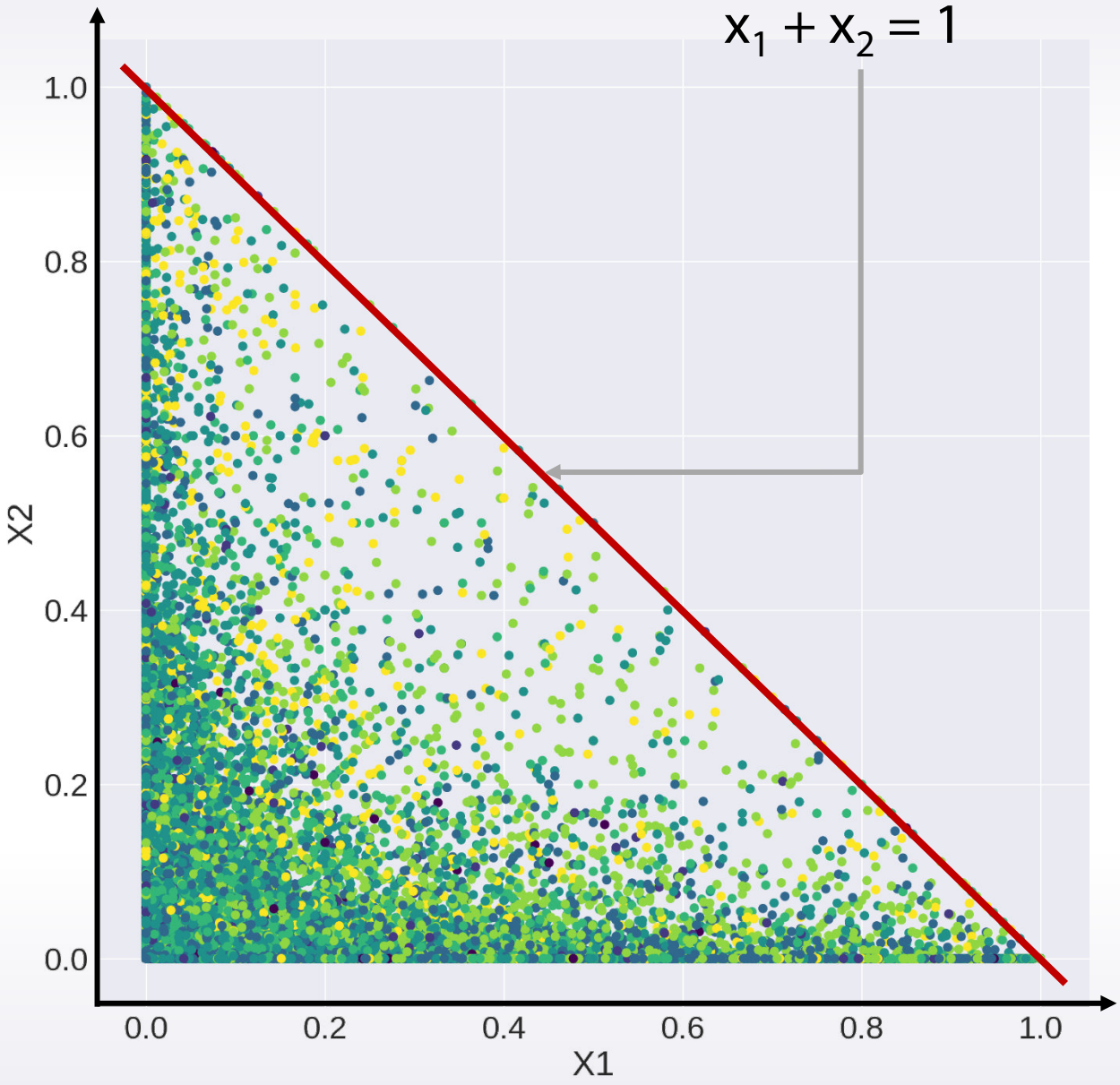



```
plt.scatter(x1, x2)
```

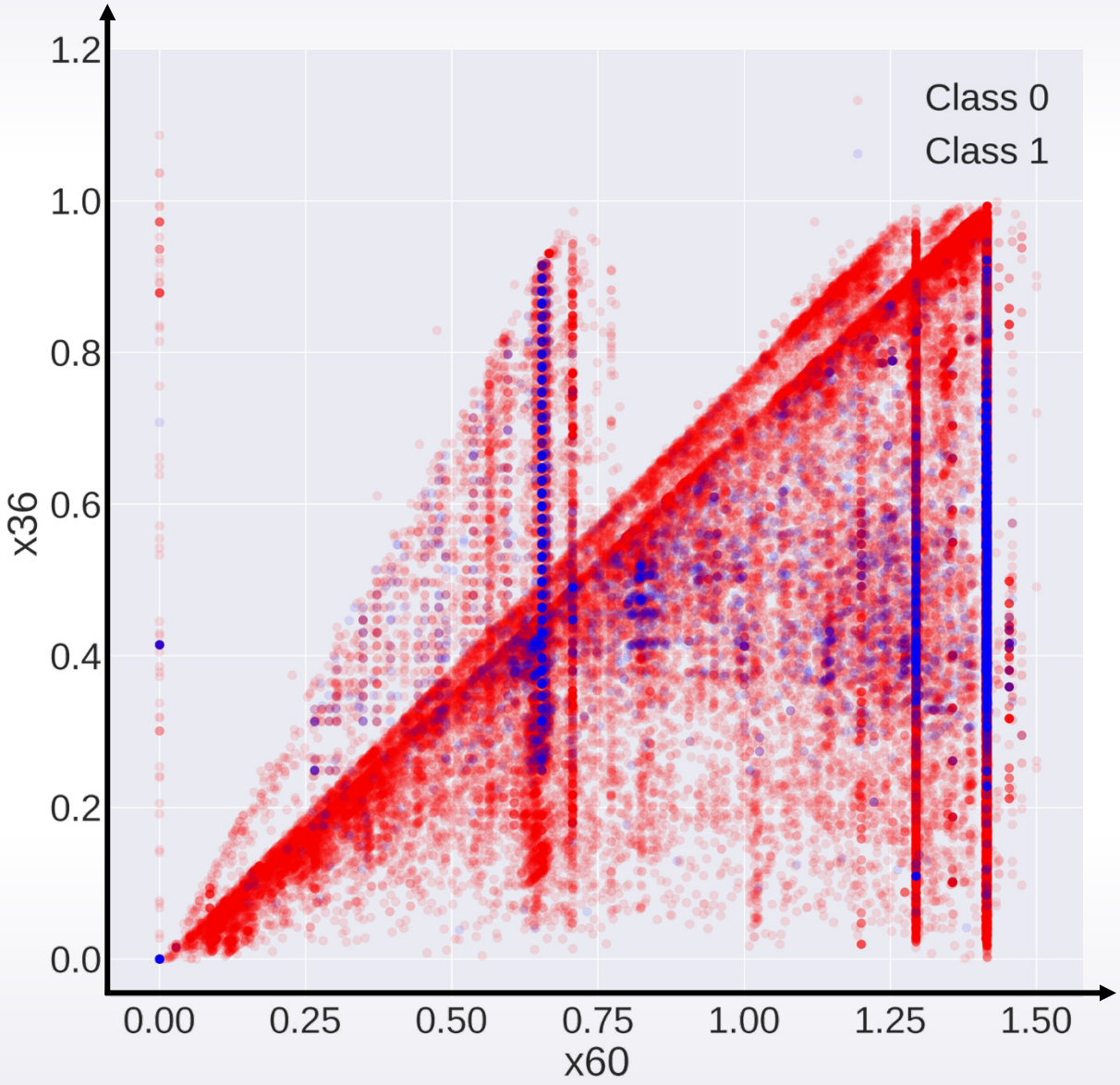
Exploring feature relations



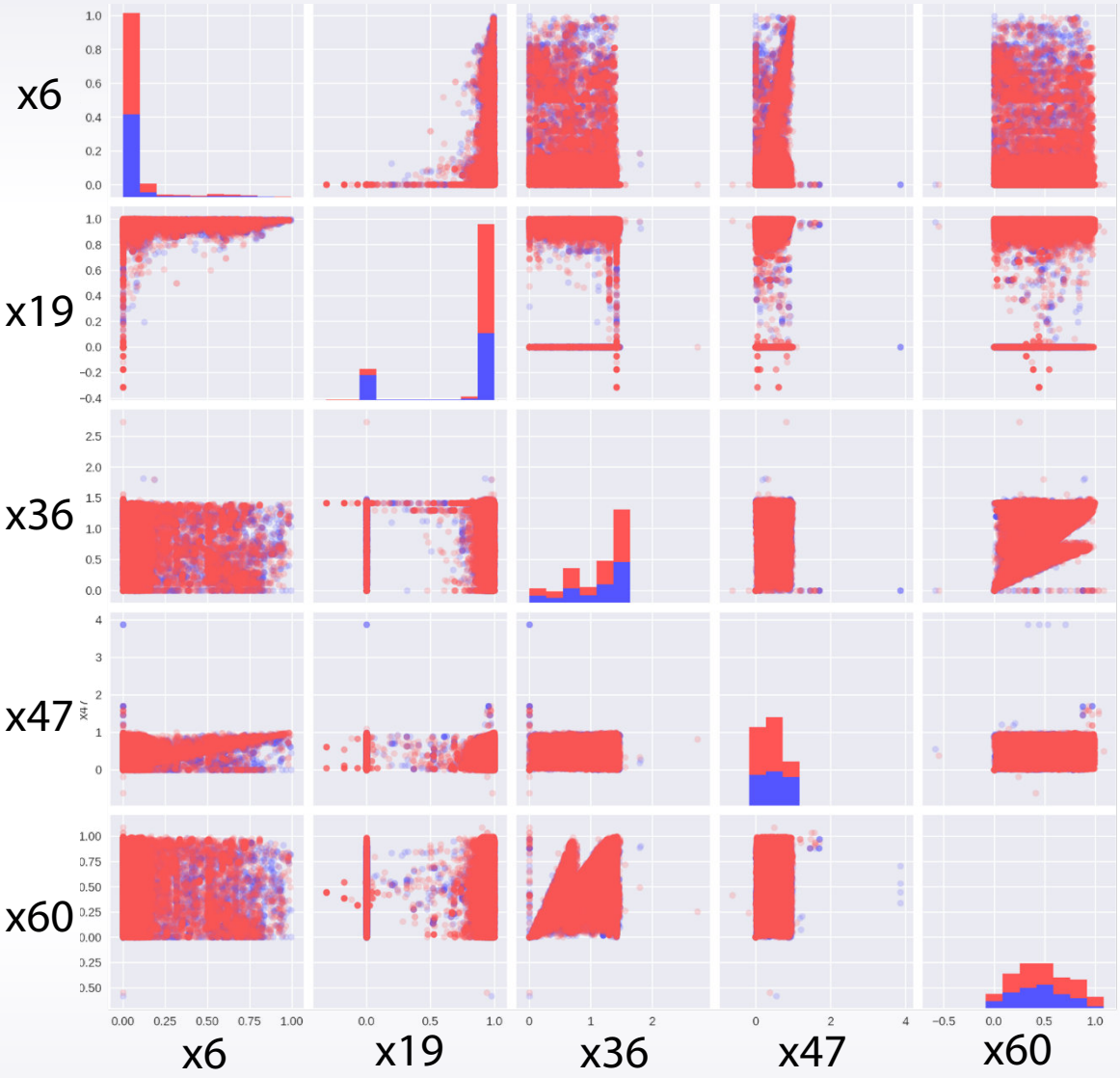
Exploring feature relations



Exploring feature relations

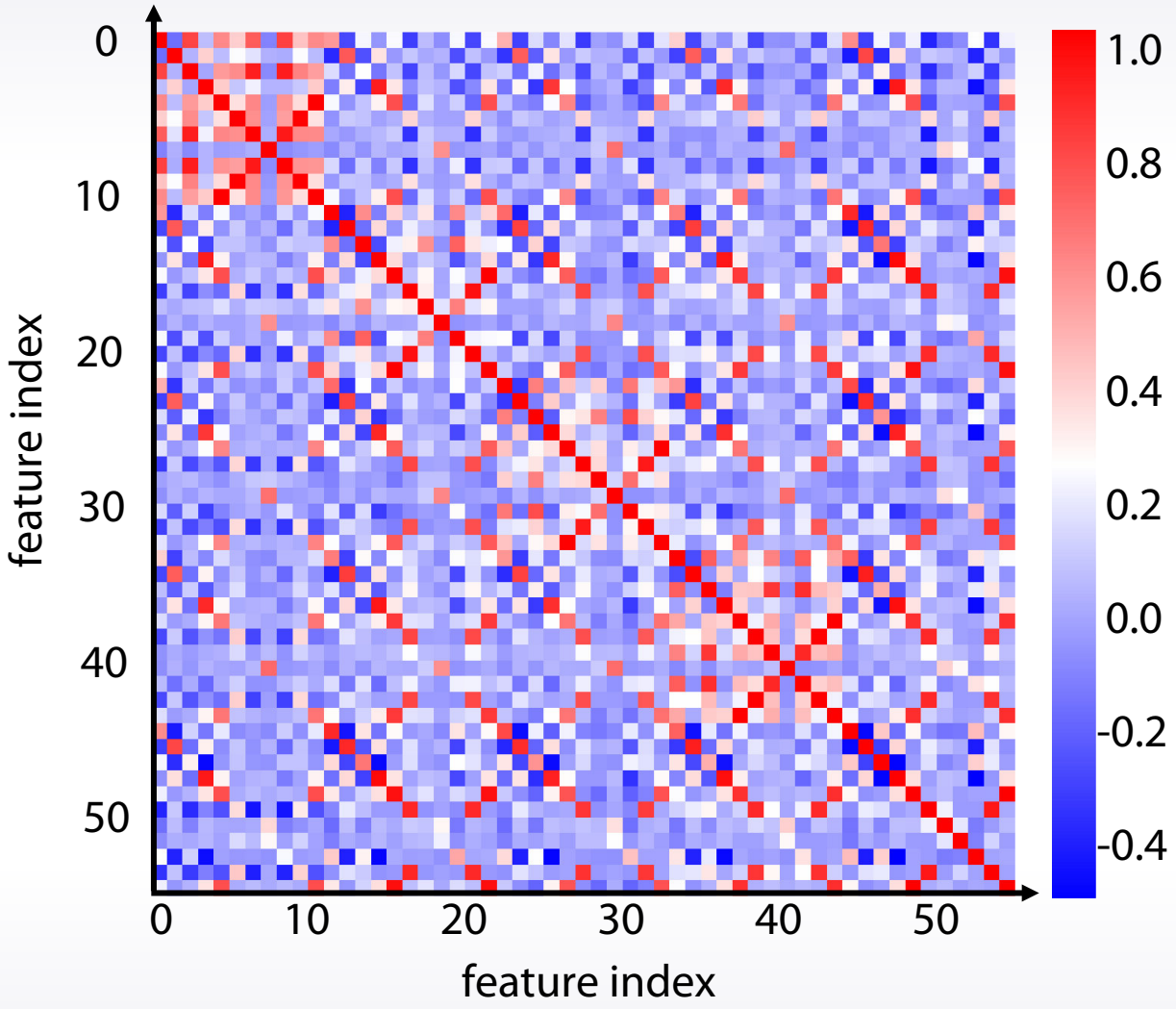


Exploring individual features: pairs



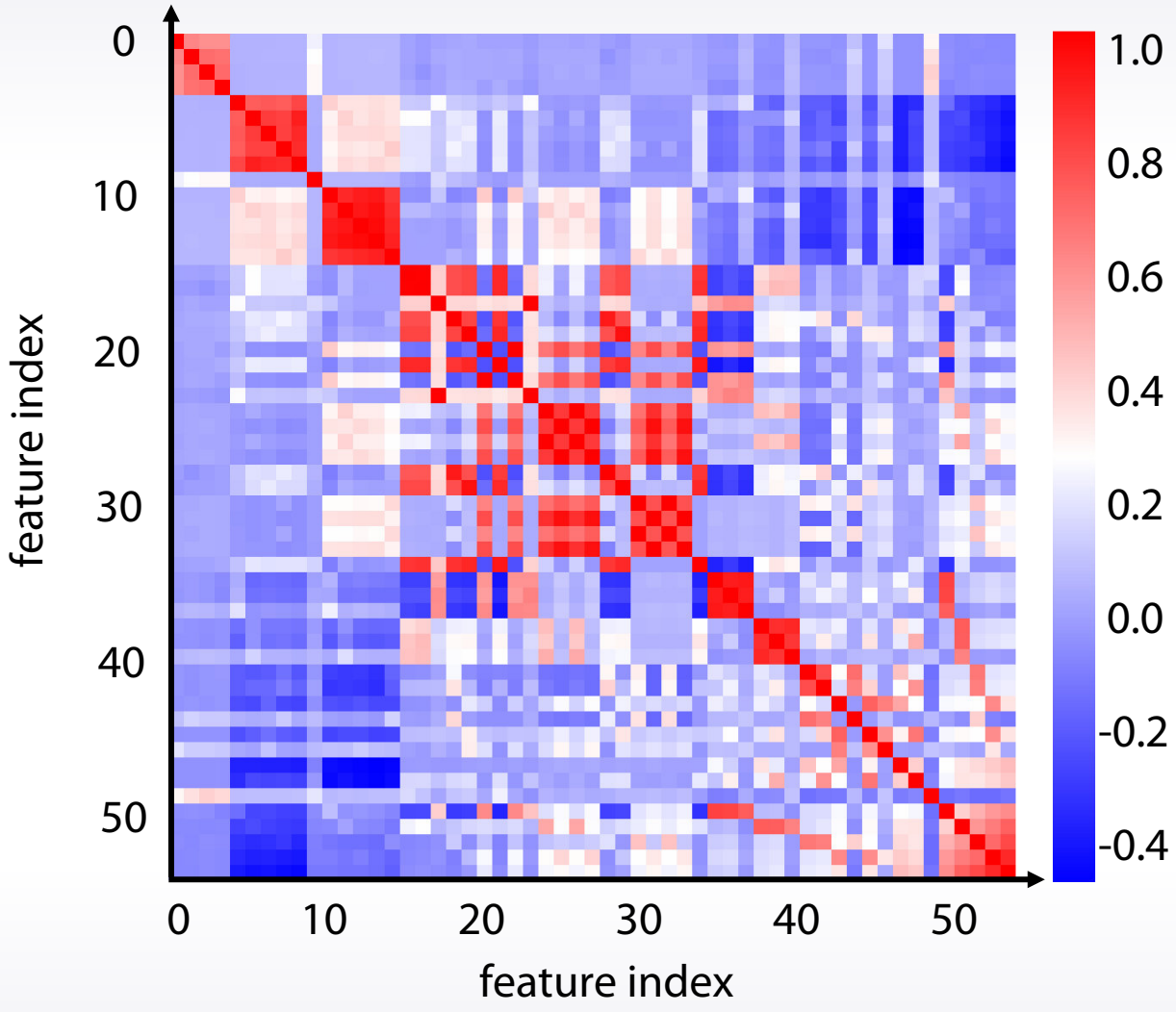

```
pd.scatter_matrix(df)
```

Exploring individual features: pairs




```
df.corr(), plt.matshow( … )
```

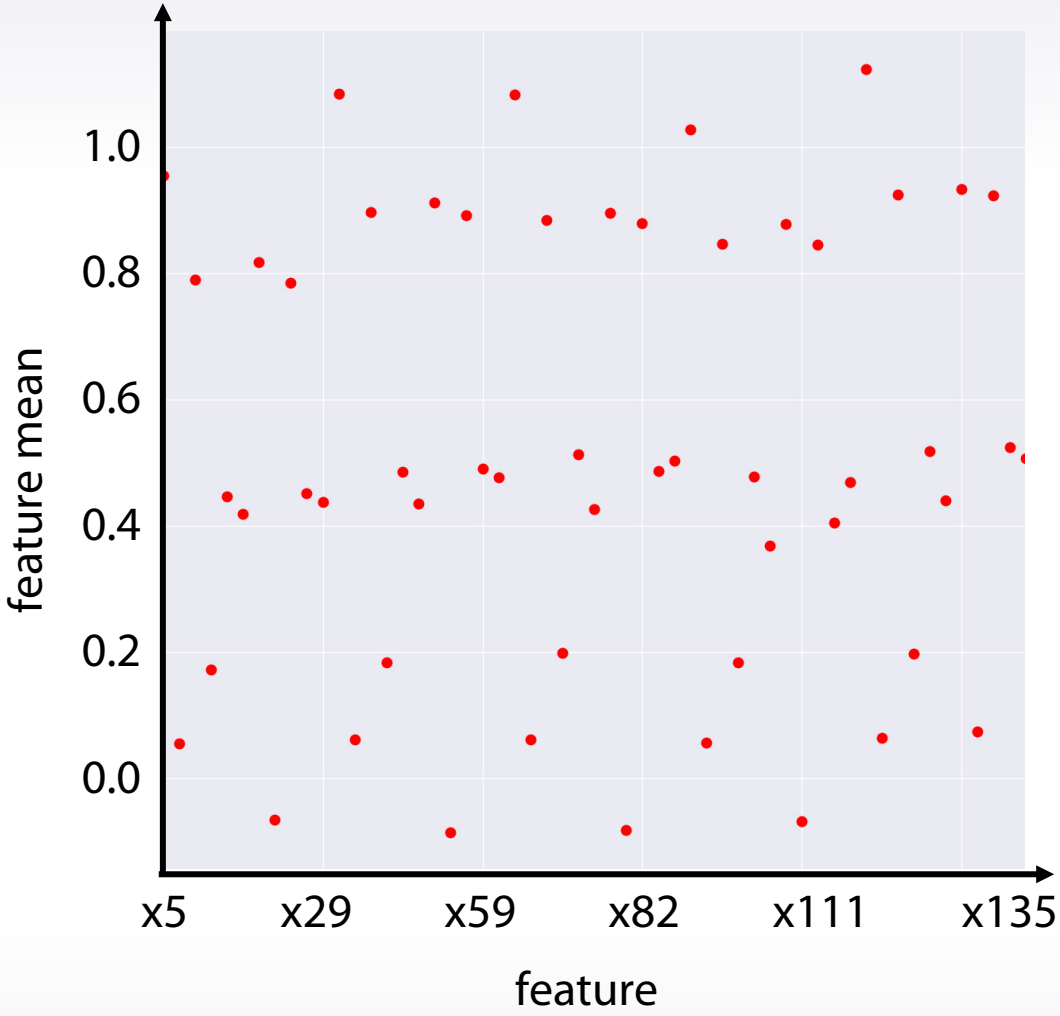
Exploring individual features: pairs/groups



Tools:

`df.corr(), plt.matshow( … )`

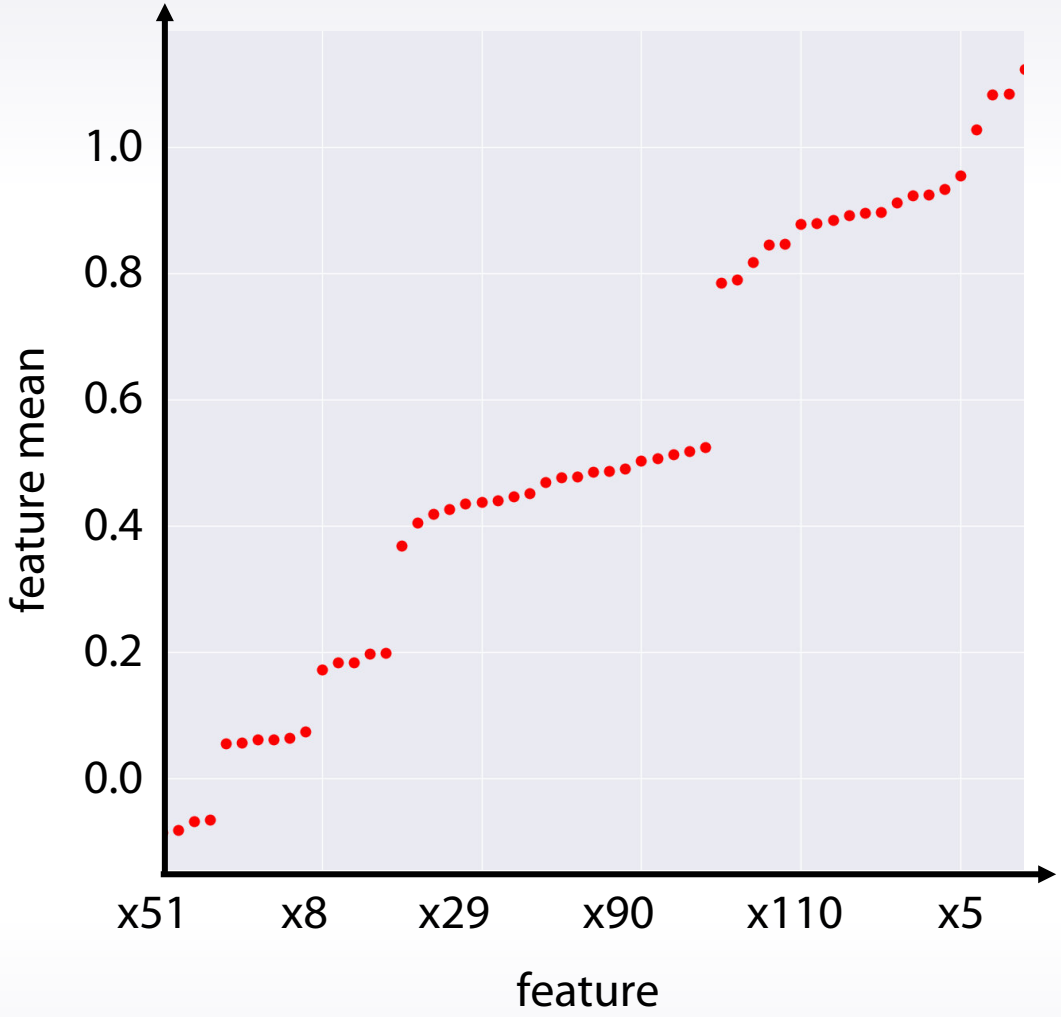
Exploring individual features: groups



Tools:

```
df.mean().plot(style='.')
```

Exploring individual features: groups



Tools:

```
df.mean().sort_values().plot(style='.')
```

Exploring individual features

Tools:

```
plt.scatter(x1, x2)
pd.scatter_matrix(df)
df.corr(), plt.matshow( … )
df.mean().sort_values().plot(style='.')
```

Conclusion

- Explore individual features
  - Histogram
  - Plot (index vs value)
  - Statistics

- Explore feature relations
  - **Pairs**
    - Scatter plot, scatter matrix
    - Corrplot
  - **Groups**
    - Corrplot + clustering
    - Plot (index vs feature statistics)

Dataset cleaning and other things to check

In this video

- Dataset cleaning
  - Constant features
  - Duplicated features
- Other things to check
  - Duplicated rows
  - Check if dataset is shuffled

Duplicated and constant features

is_train	f0	f1	f2	f3	f4	f5
True	13	H	1.2	1.2	A	C
True	13	H	36.6	36.6	B	A
False	13	H	0	0	A	C
False	13	G	-14	-14	C	B

Duplicated and constant features

is_train	*f0*	f1	f2	f3	f4	f5
True	*13*	H	1.2	1.2	A	C
True	*13*	H	36.6	36.6	B	A
False	*13*	H	0	0	A	C
False	*13*	G	-14	-14	C	B

```
traintest.nunique(axis=1) == 1
```

Duplicated and constant features

is_train	f0	*f1*	f2	f3	f4	f5
True	13	H	1.2	1.2	A	C
True	13	H	36.6	36.6	B	A
False	13	H	0	0	A	C
False	13	G	-14	-14	C	B

```
train.nunique(axis=1) == 1
```

Duplicated and constant features

is_train	f0	f1	*f2*	*f3*	f4	f5
True	13	H	1.2	1.2	A	C
True	13	H	36.6	36.6	B	A
False	13	H	0	0	A	C
False	13	G	-14	-14	C	B

```
traintest.T.drop_duplicates()
```

Duplicated and constant features

is_train	f0	f1	f2	f3	*f4*	*f5*
True	13	H	1.2	1.2	A	C
True	13	H	36.6	36.6	B	A
False	13	H	0	0	A	C
False	13	G	-14	-14	C	B

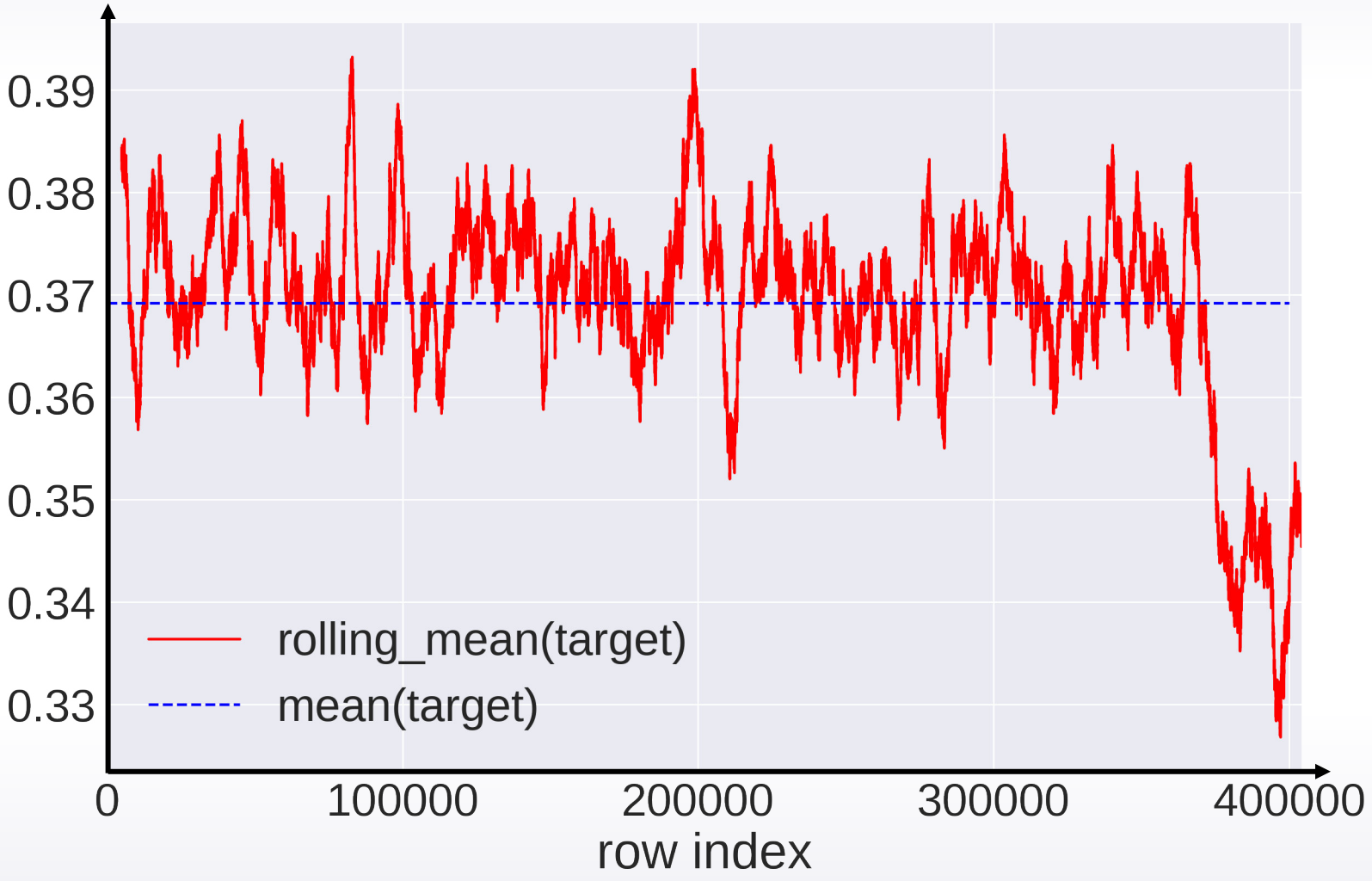
```
for f in categorical_feats:
  traintest[f] =raintest[f].factorize()

traintest.T.drop_duplicates()
```

Duplicated rows

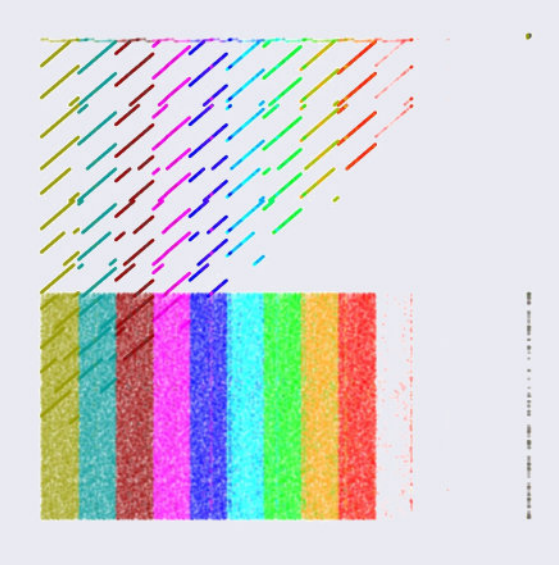
f1	f2	f3	y
13	34r9	A	**0**
13	34r9	A	**1**
13	34r9	A	**1**

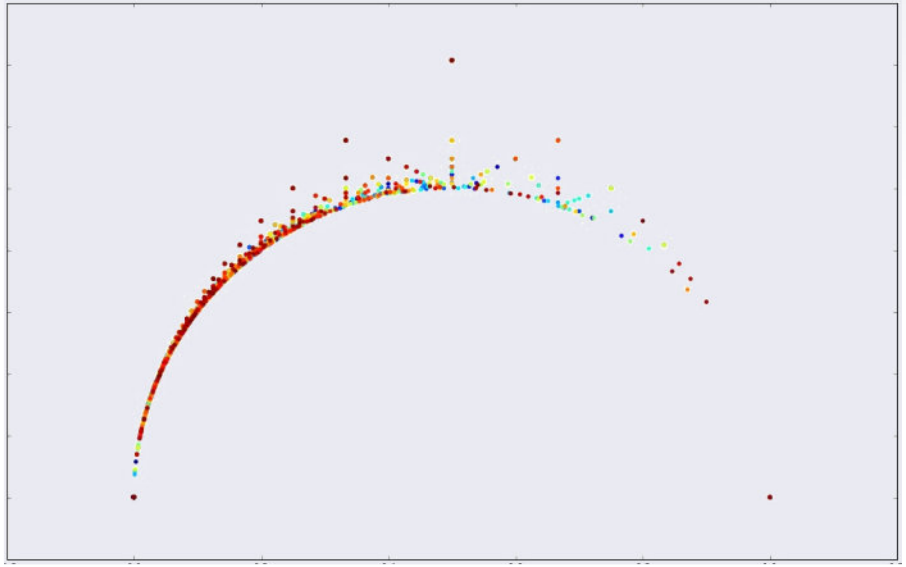
- Check if same rows have same label
- Find duplicated rows, understand why they are duplicated

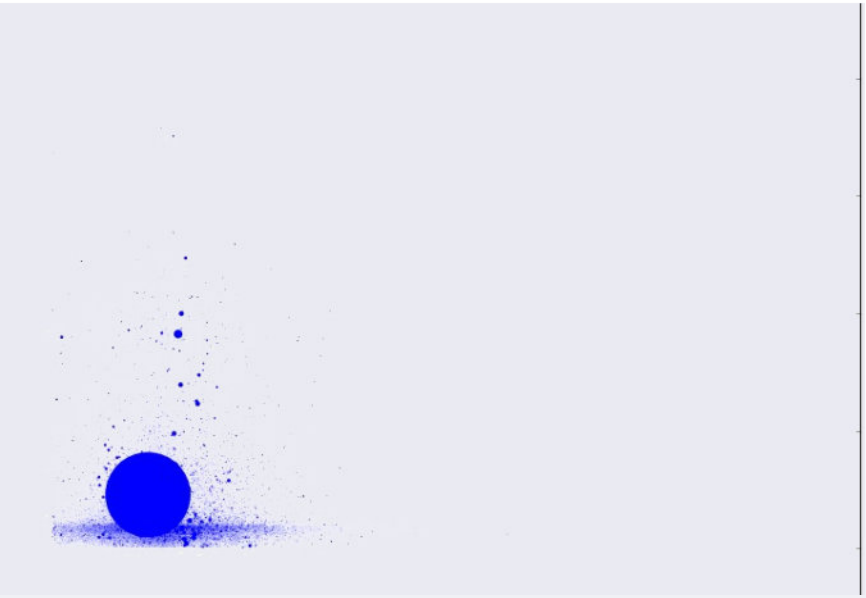
Check if dataset is shuffled

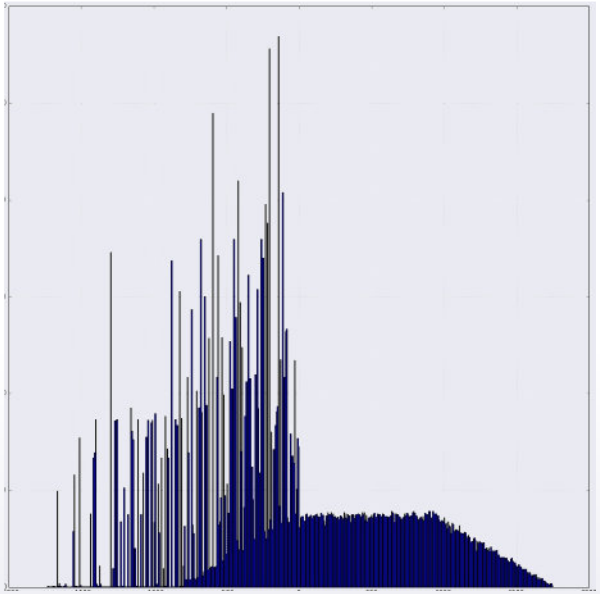


Cool visualizations









EDA check list

- Get domain knowledge
- Check if the data is intuitive
- Understand how the data was generated

- Explore individual features
- Explore pairs and groups

- Clean features up

- Check for leaks! (later in this course)