Visualizations

Video overview

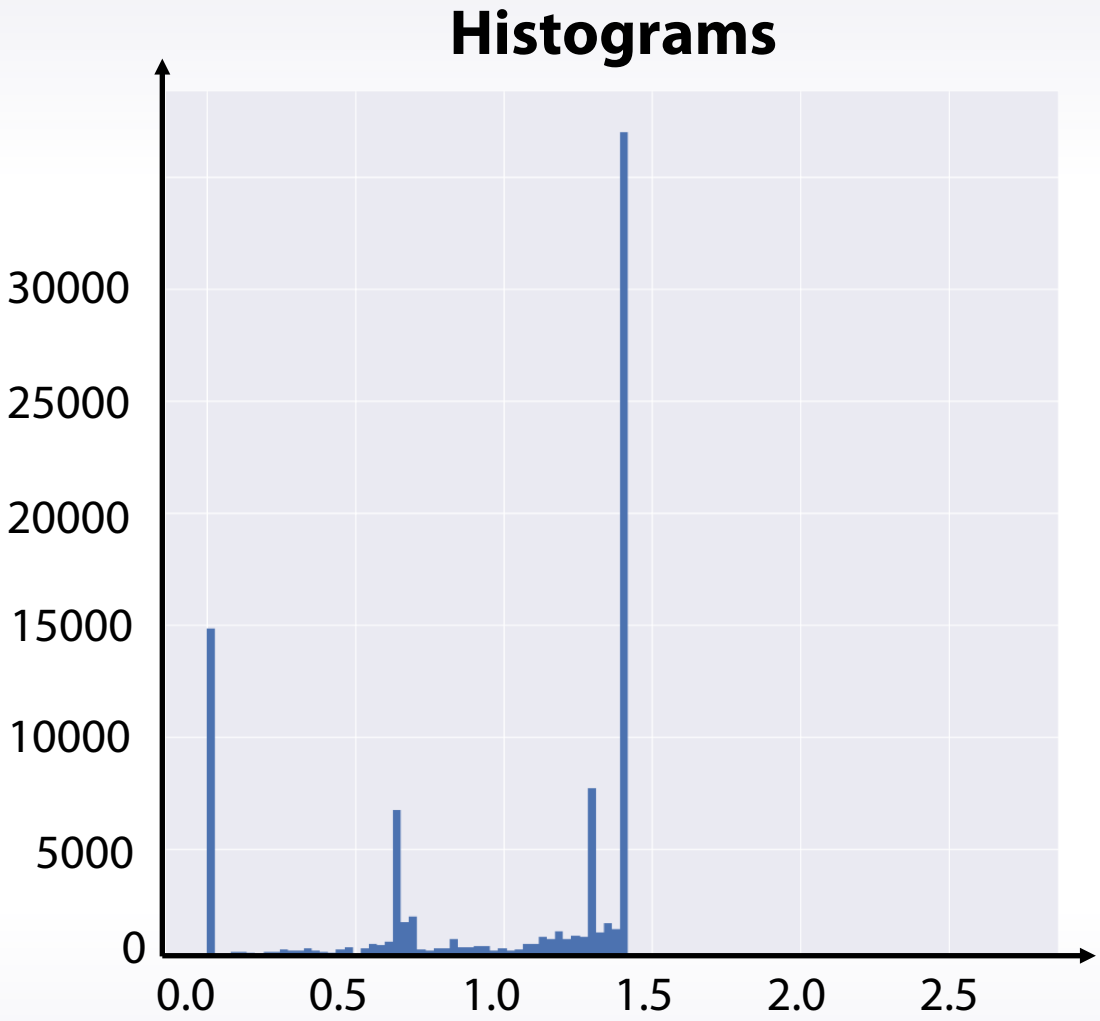
- Visualization tools to…
  - **Explore individual features**
    - Histograms
    - Plots
    - Statistics
  - **Explore feature relations**
    - Scatter plots
    - Correlation plots
    - Plot (index vs feature statistics)
    - And more

Visualizations

EDA is an art!

And visualizations are our art tools

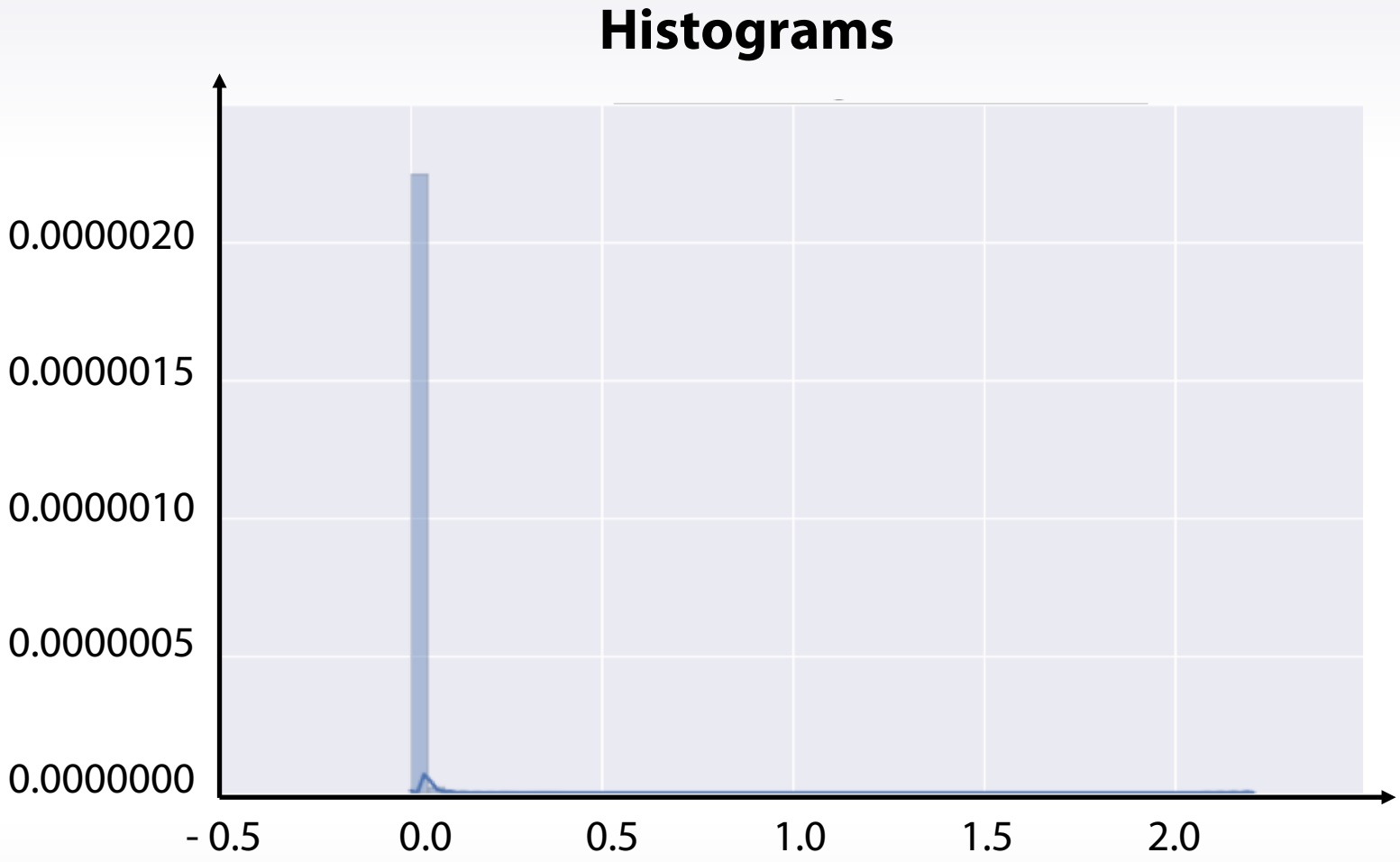
Art tools




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

단점1: aggregate이라서 unique 값인지 반복되는 값인지 알 수 없다

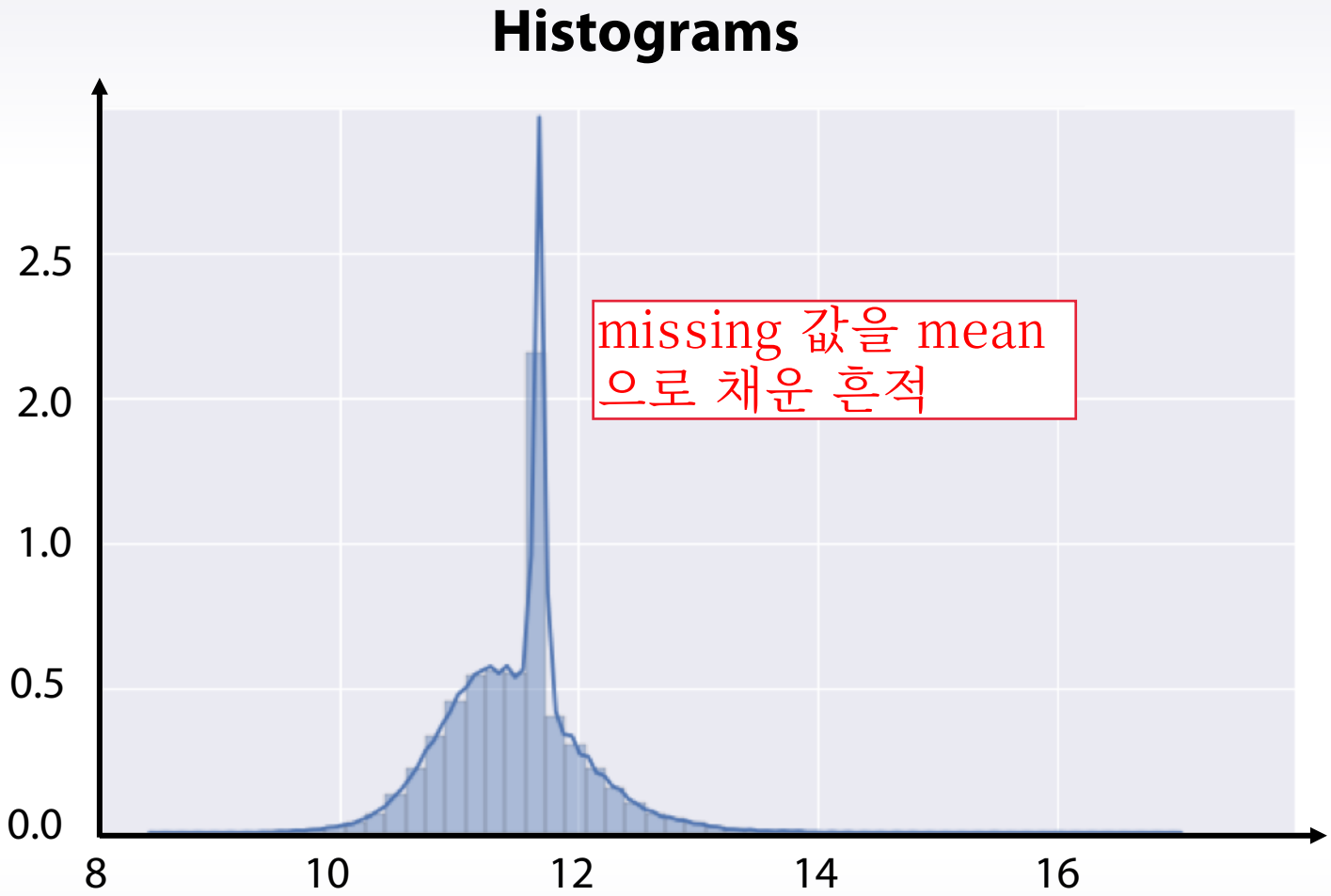
Art tools




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

0이 많고 degenerate 분포인 것 같이 보이지만
log 취해서 그려보면 - - >

Art tools

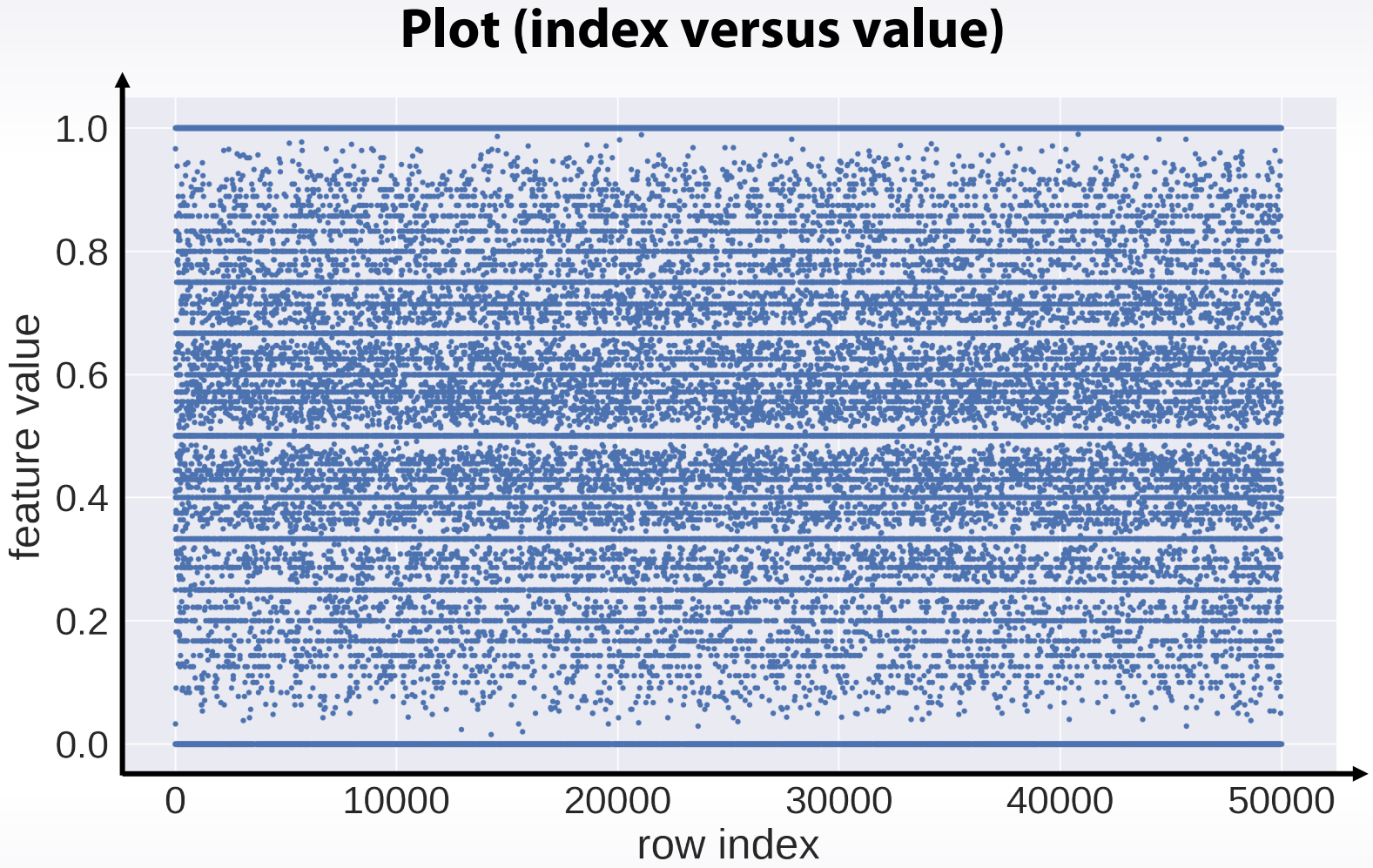


0이 아닌 값이 많음, non - degenerate 분포

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

주의: 여러plot 그려서 특징 확인 필요

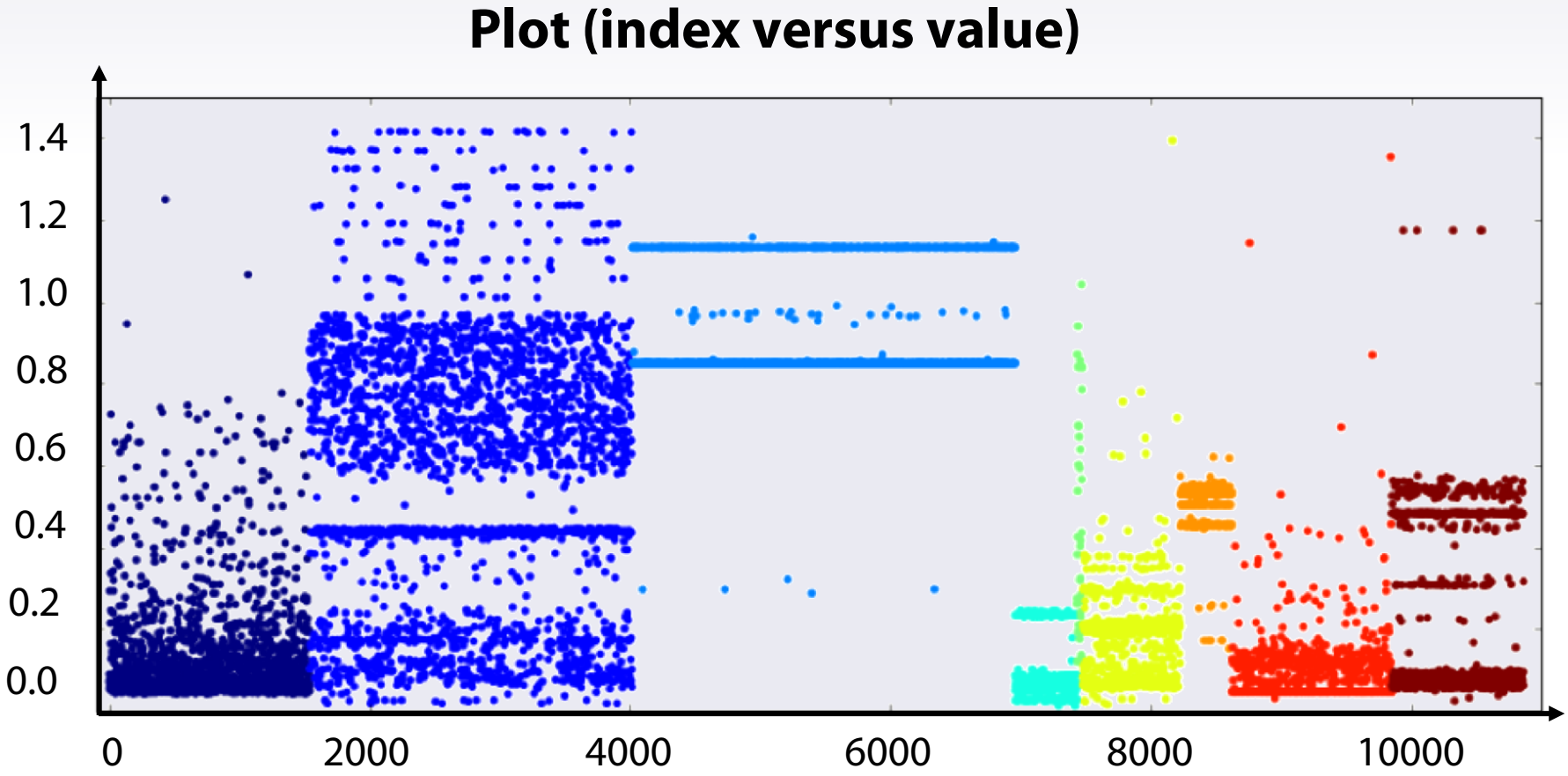
Art tools




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

수평line은 반복값을 의미

Art tools



label에 따라 색을 다르게 그림

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

Art tools

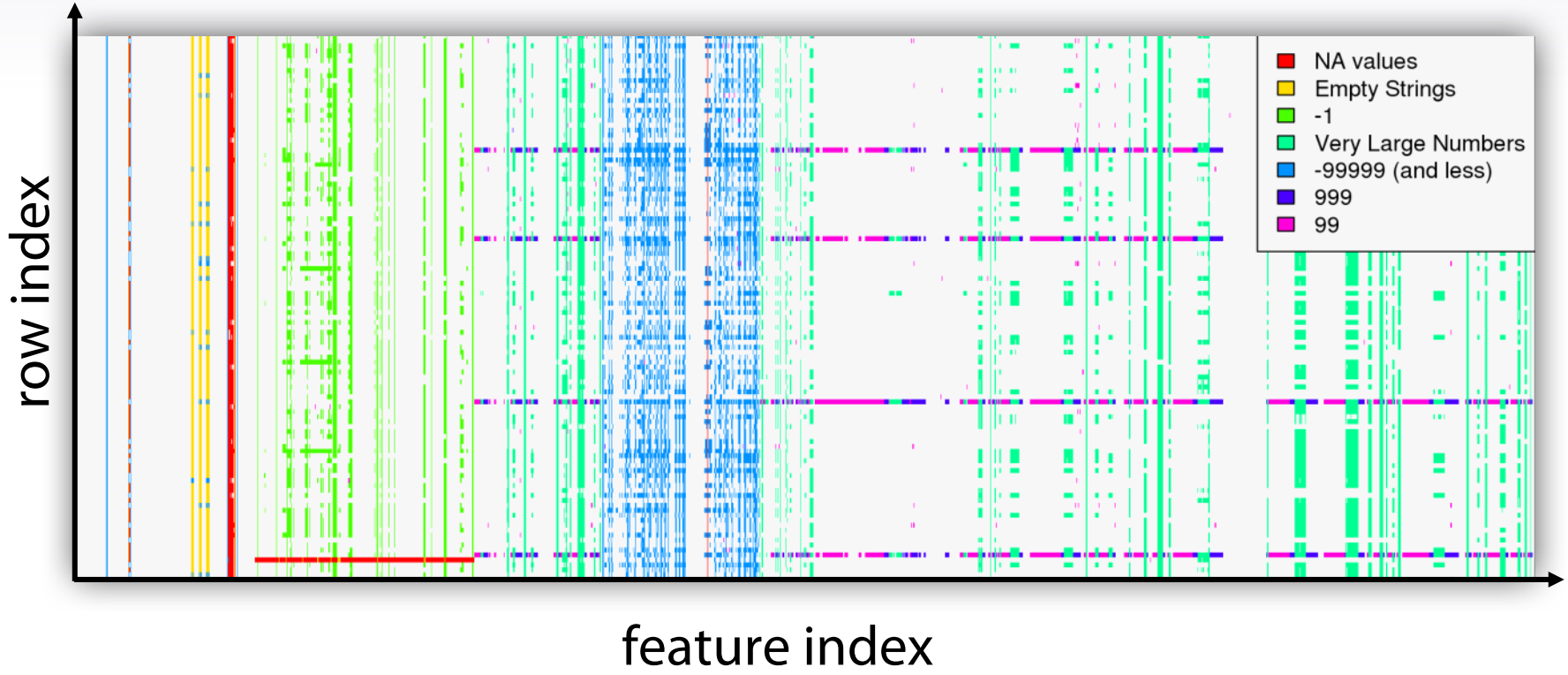
Feature statistics

	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



feature값의 발생빈도를 시각화

```
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()
```

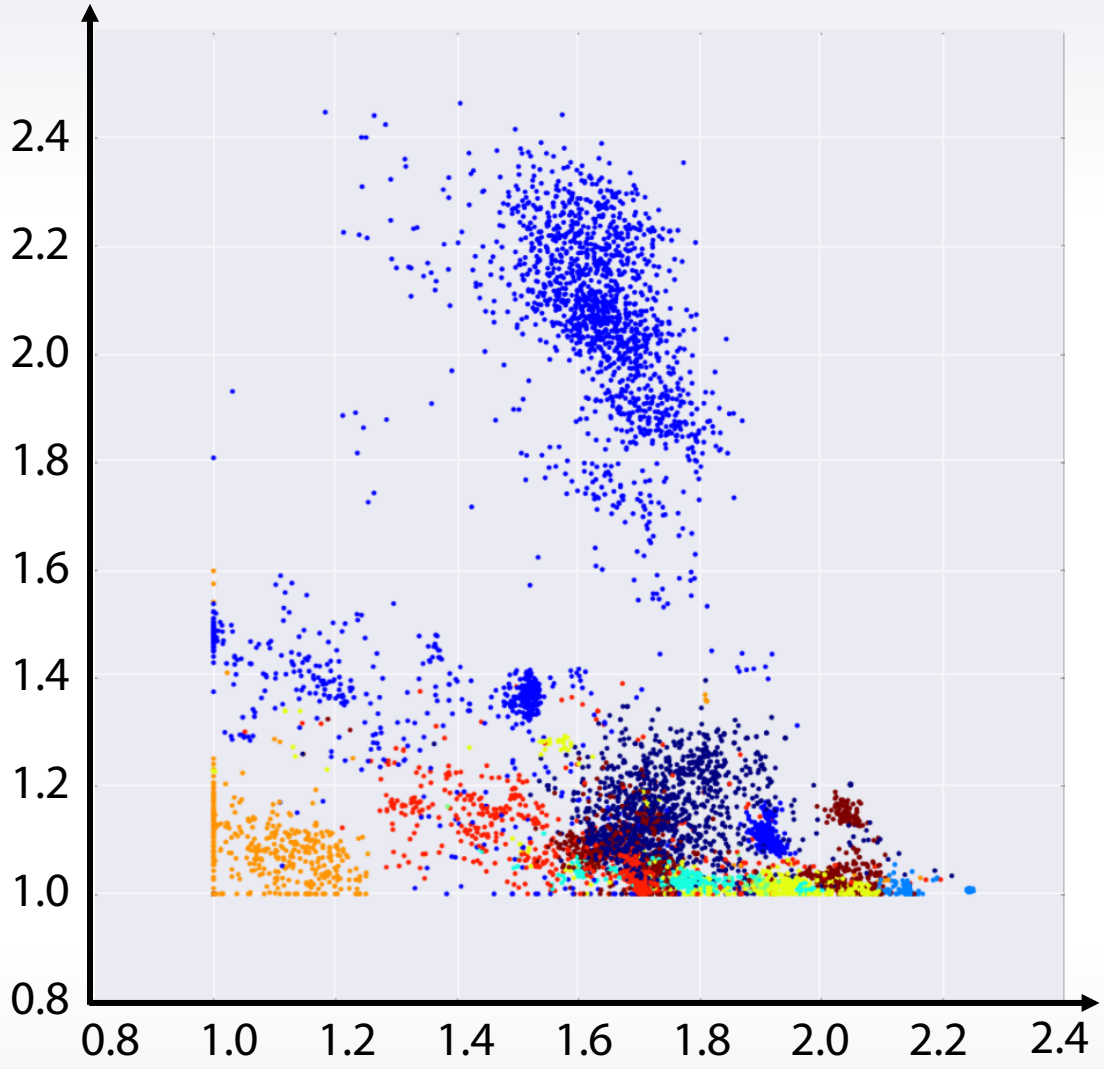
Other tools:

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

Next in this video

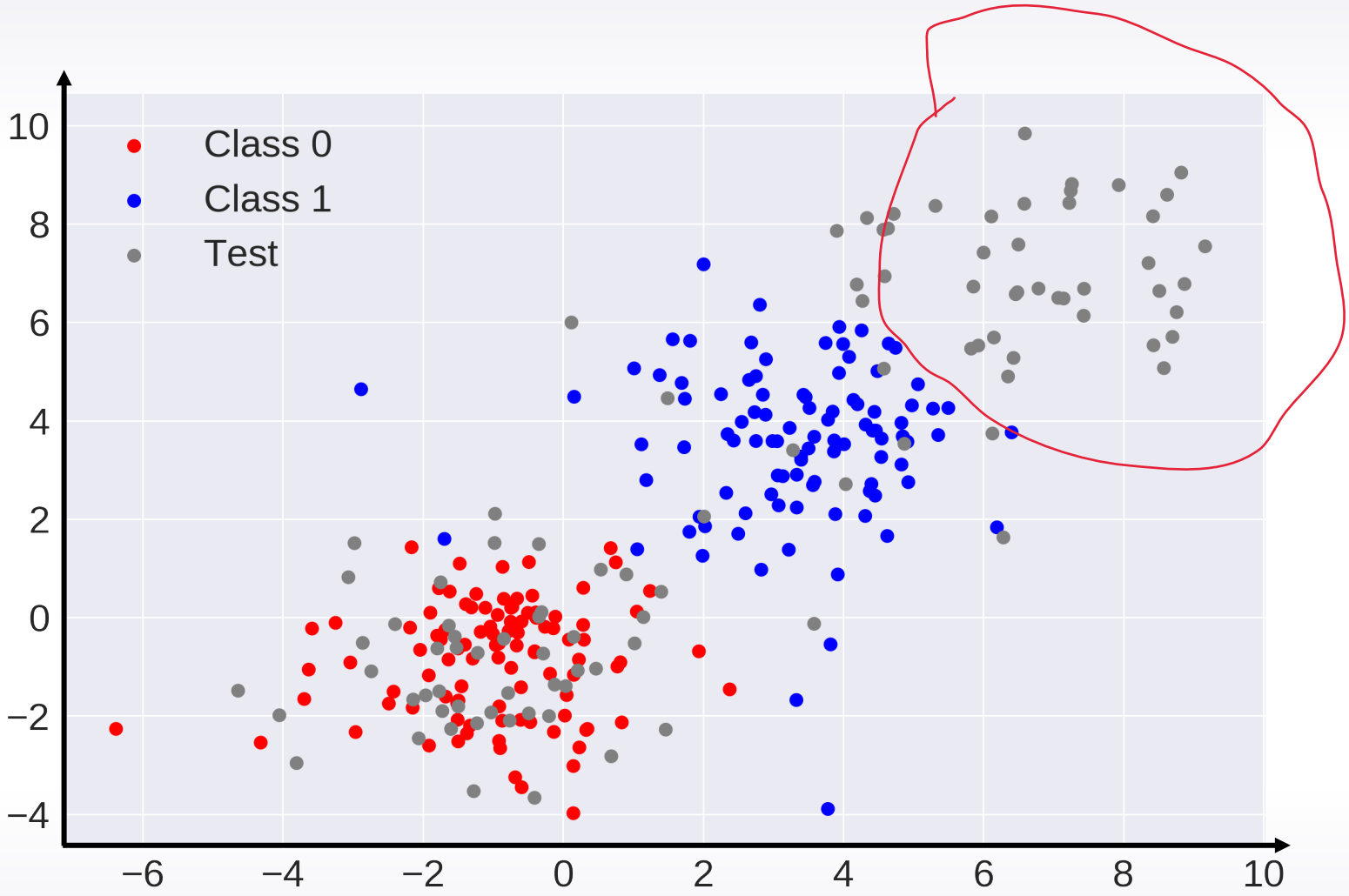
- Visualization tools to…
  - **Explore individual features**
    - Histograms
    - Plots
    - Statistics
  - **Explore feature relations**
    - Scatter plots
    - Correlation matrices
    - …

Exploring feature relations




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

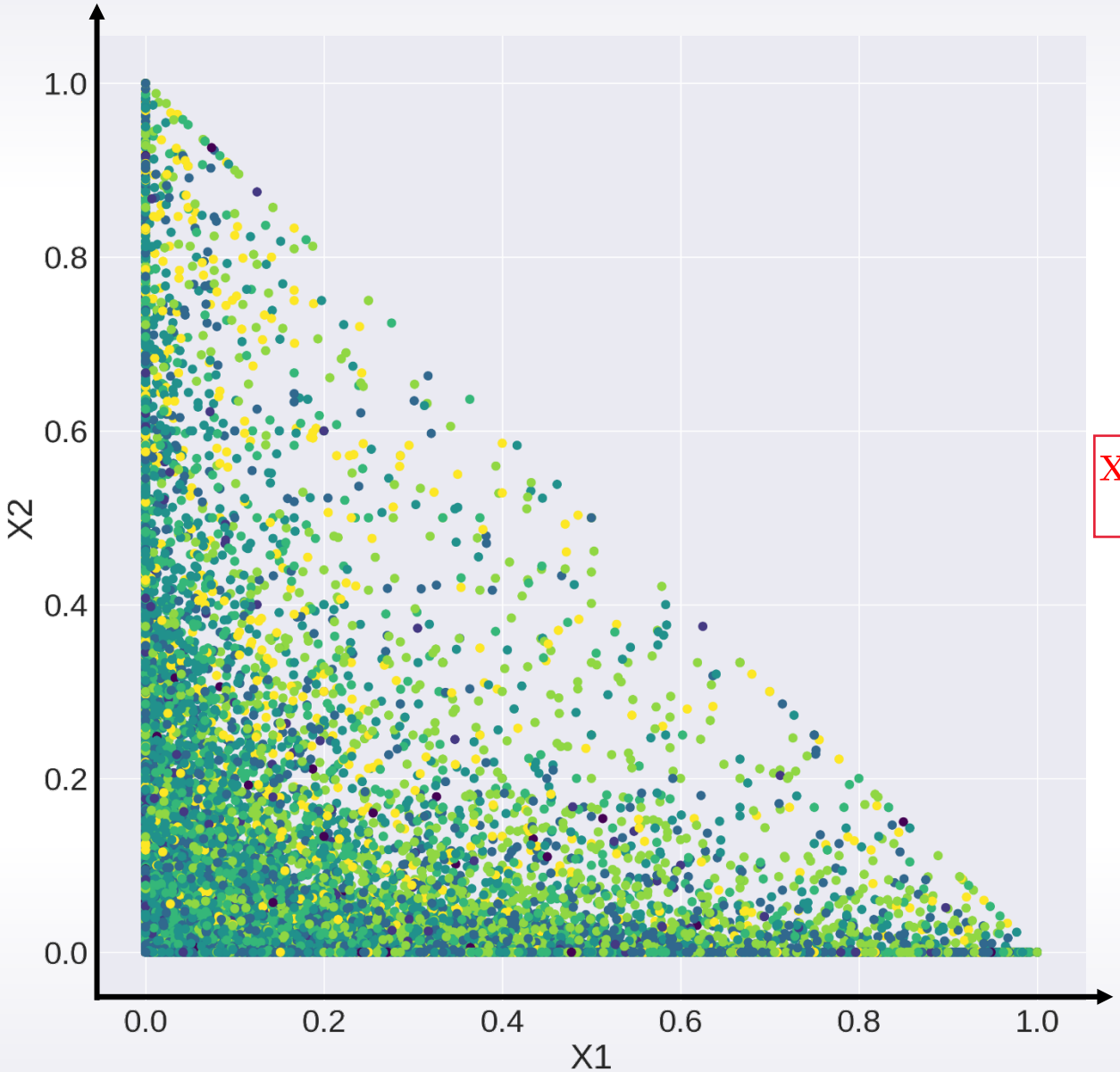
Exploring feature relations




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

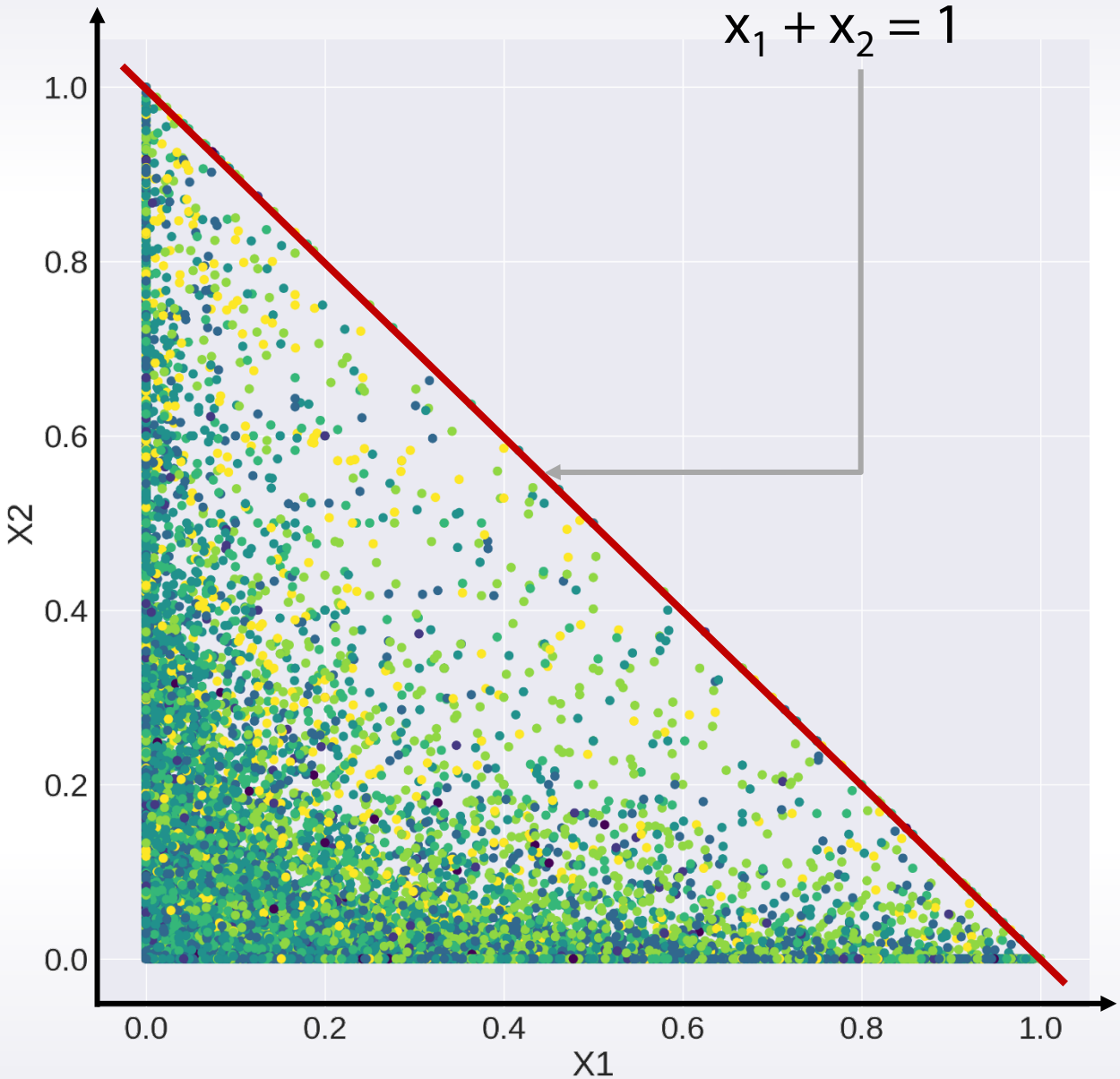
train, test 세트의 데이터 분포가 같은지 확인 - - > 회색만 있는 영역이 존재하면 세트간 분포가 다른 것임

Exploring feature relations

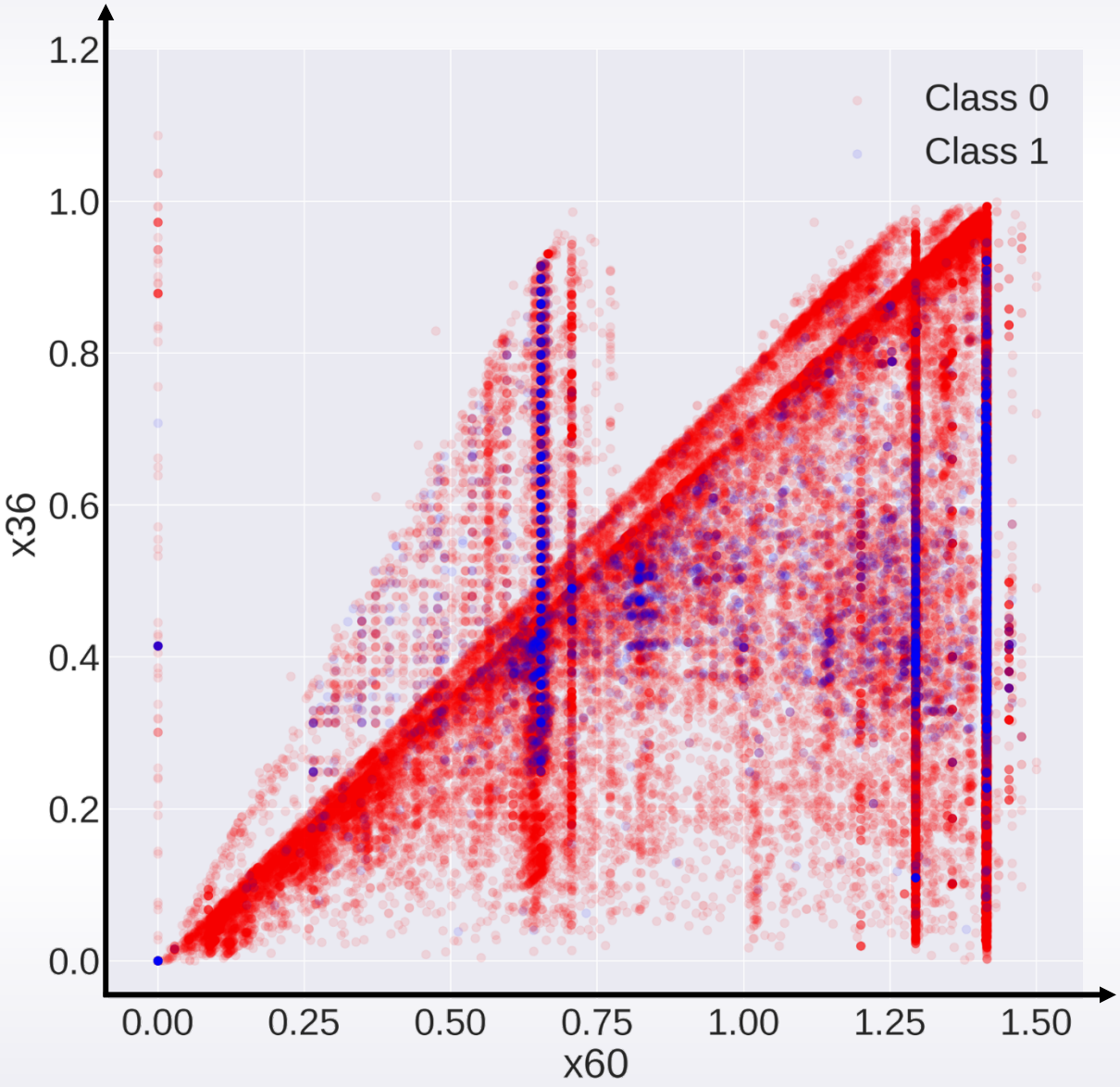


$x1 + x2 < 1$

Exploring feature relations

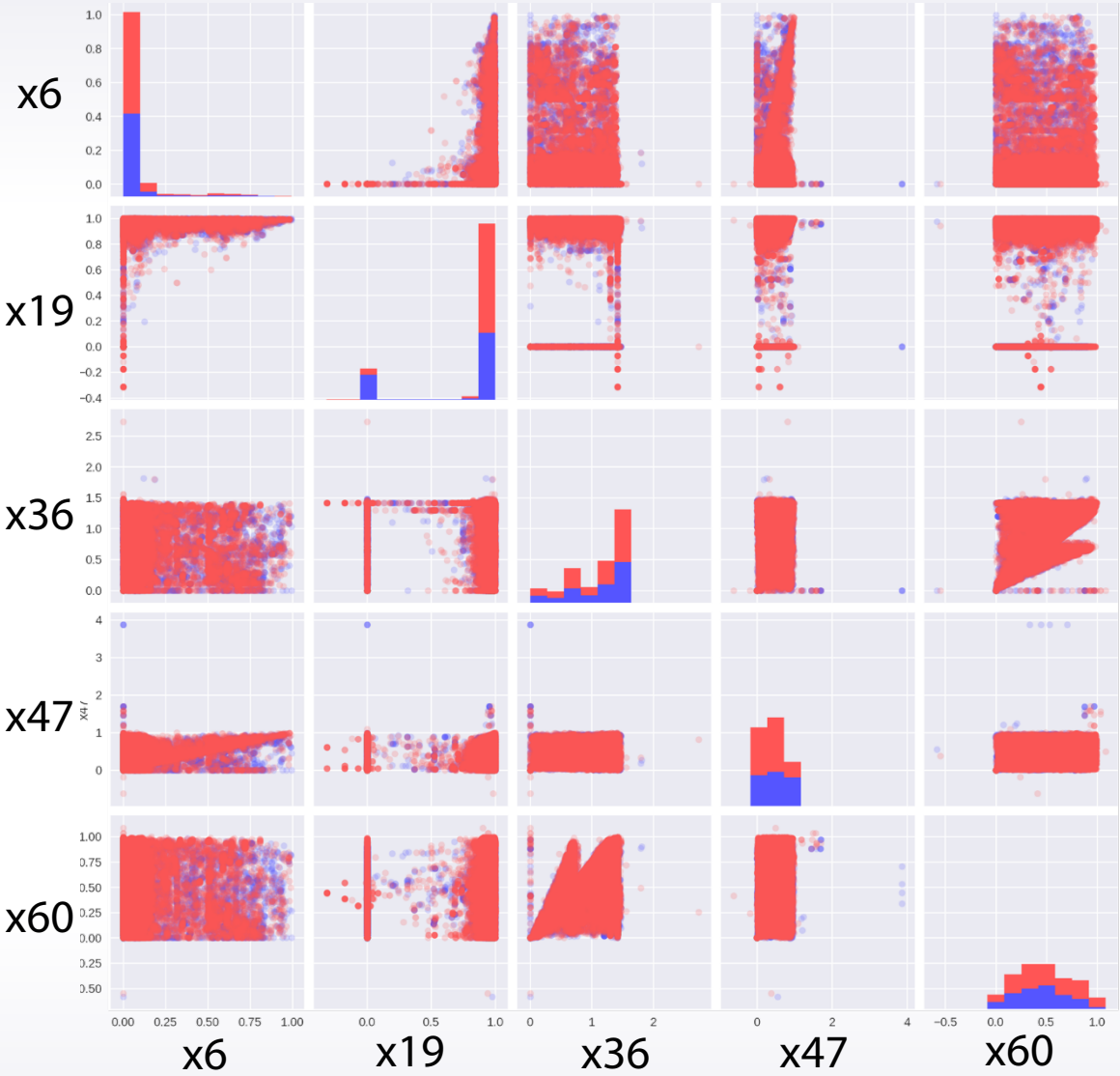


Exploring feature relations



new feature generation 의 단서 제공

Exploring individual features: pairs

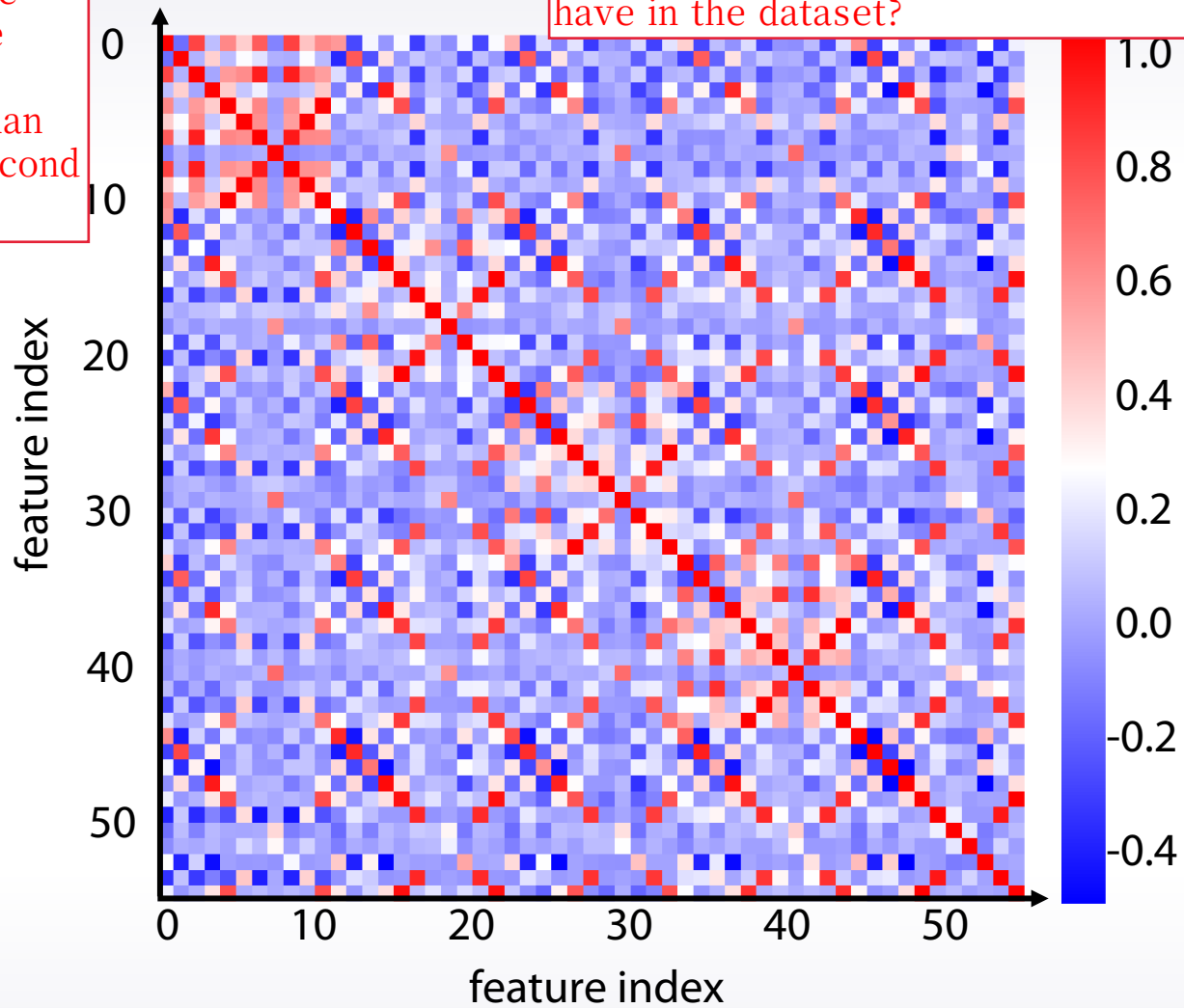



```
pd.scatter_matrix(df)
```

Exploring individual features: pairs

how many rows are there such that the value of the first feature is larger than the value of the second one?

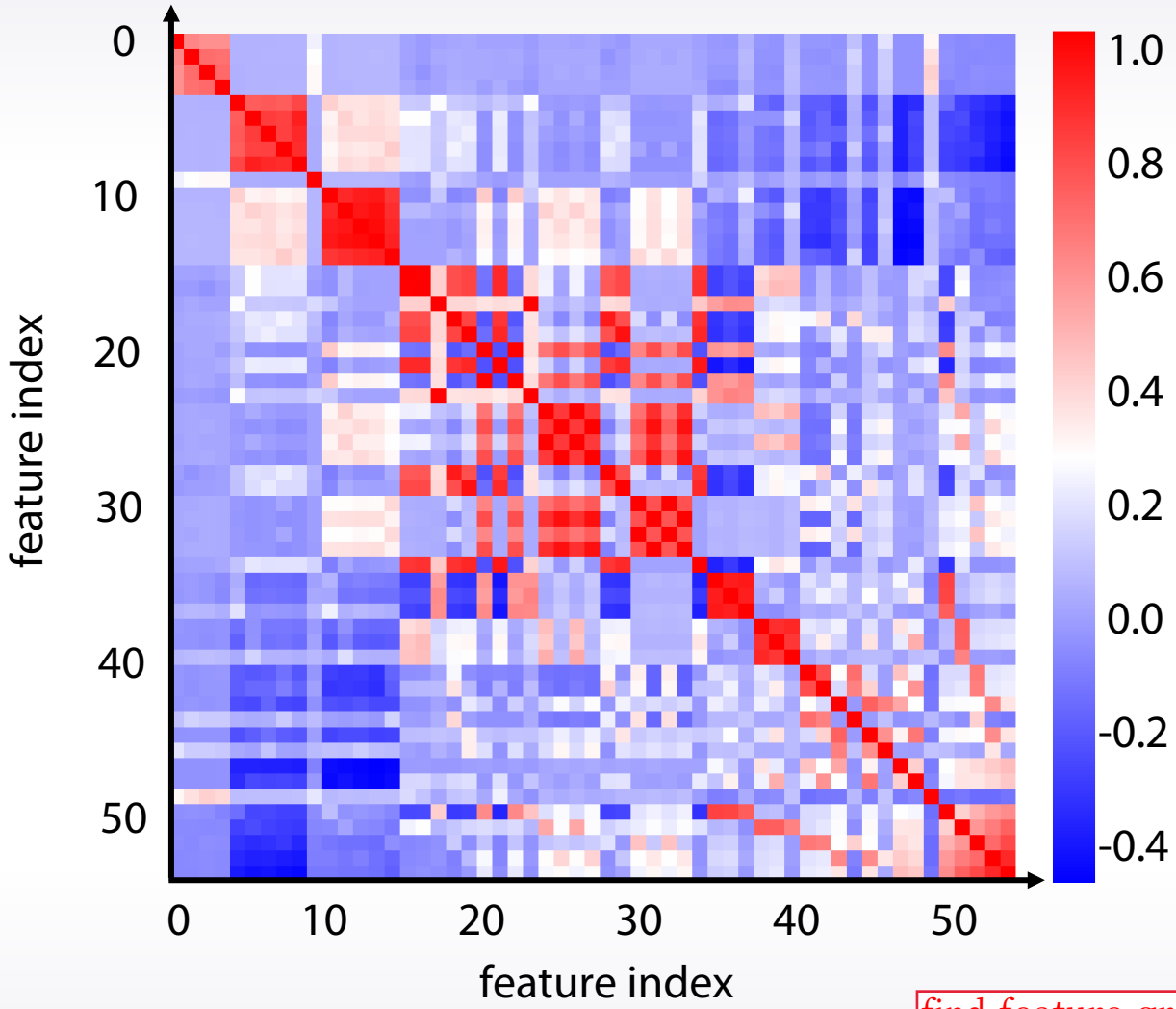
how many distinct combinations the features have in the dataset?




```
df.corr(), plt.matshow( … )
```

k - means clustering on the rows and columns and reorder - - - >

Exploring individual features: pairs/groups

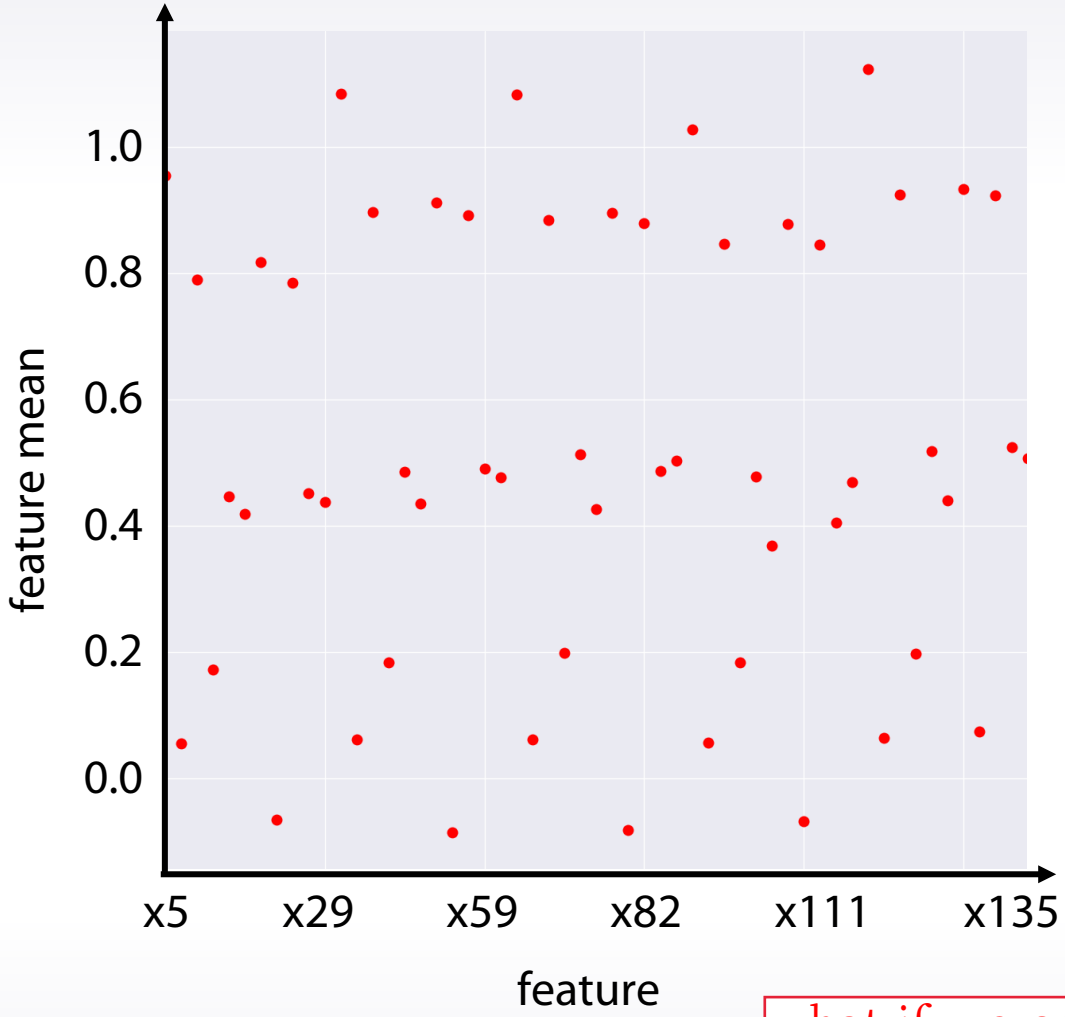


Tools:

```
df.corr(), plt.matshow( … )
```

find feature groups

Exploring individual features: groups

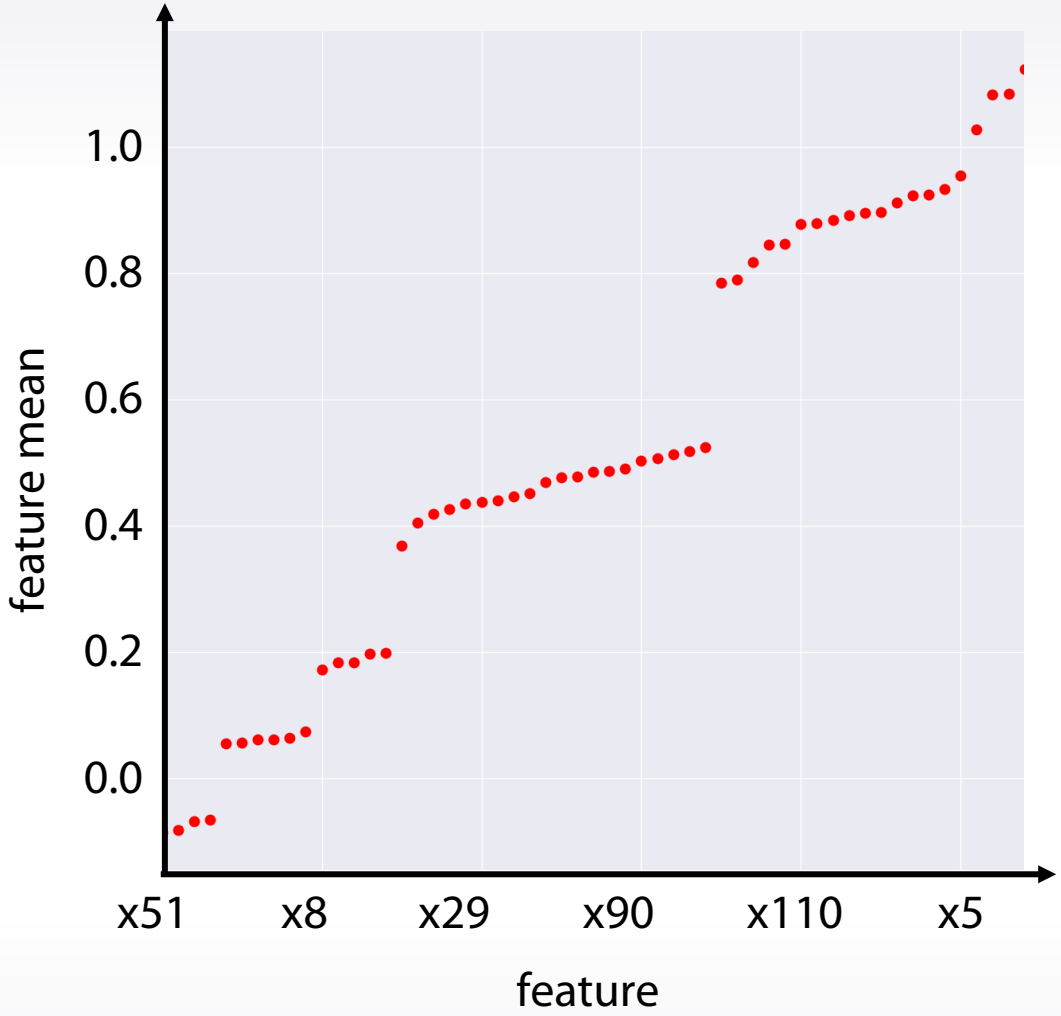


Tools:

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

what if we sorted the columns based on this statistic?
Feature and mean?

Exploring individual features: groups



find feature groups

use them to generate new features

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)