

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

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

feature

Duplicated and constant features

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

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

train test set F0 가
- > F0가

Duplicated and constant features

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

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

train

test

Duplicated and constant features

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

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

f2

f3

feature

Duplicated and constant features

f4 : A -> 1, B -> 2, C -> 3

f5 : C -> 1, A -> 2, B -> 3

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

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

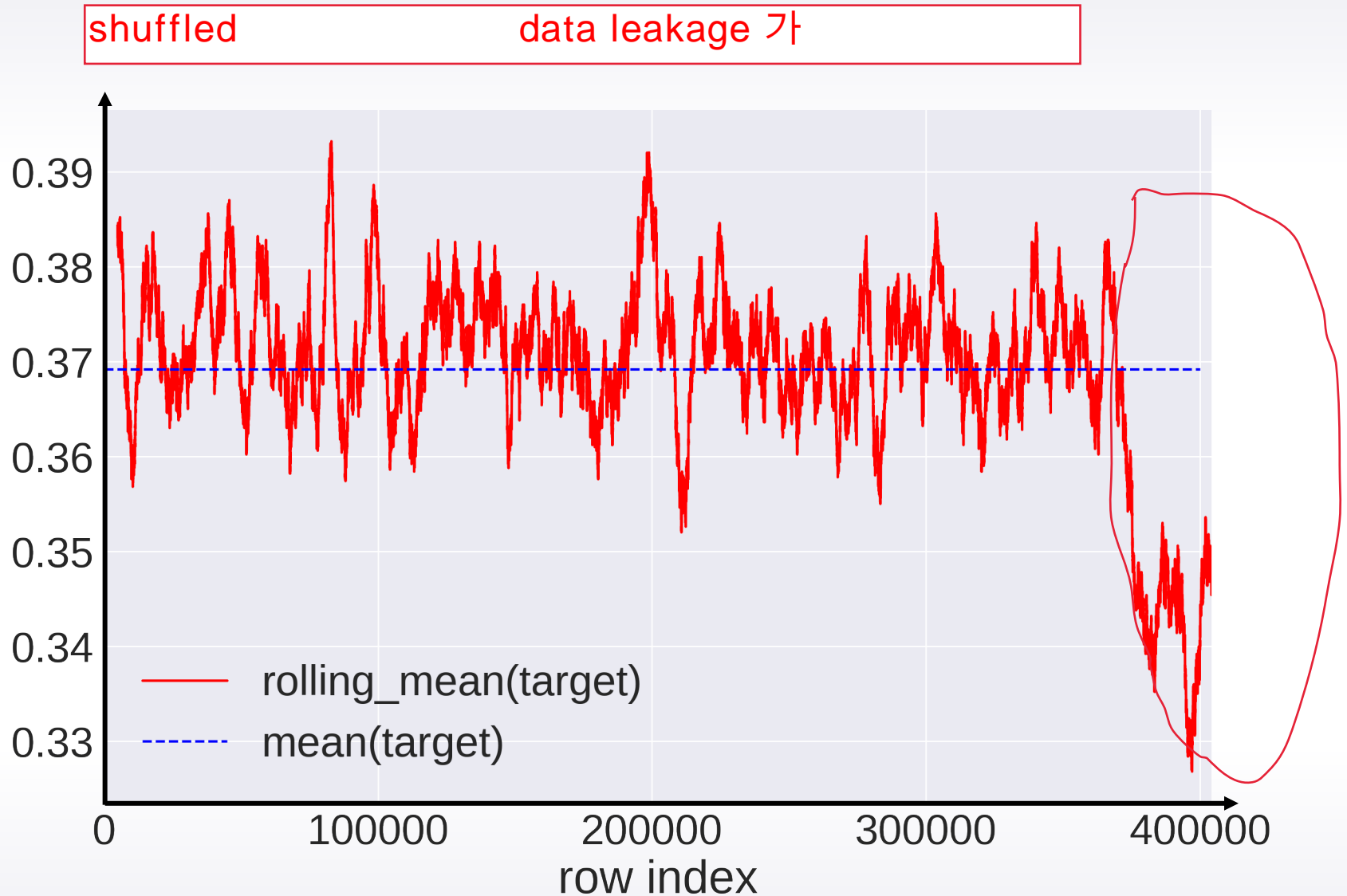
encoding f4 f5 -> feature

Duplicated rows

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

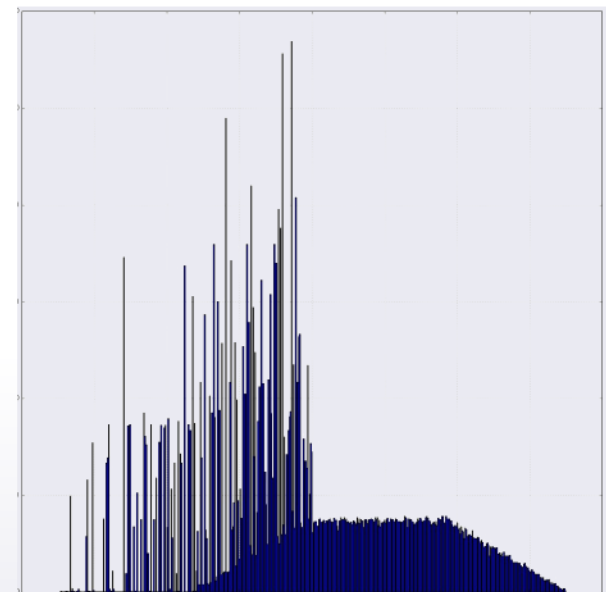
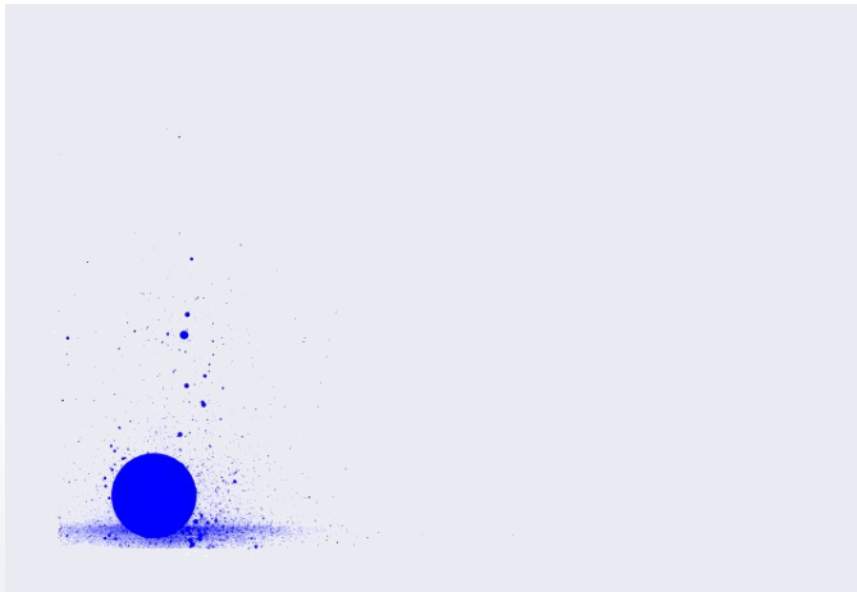
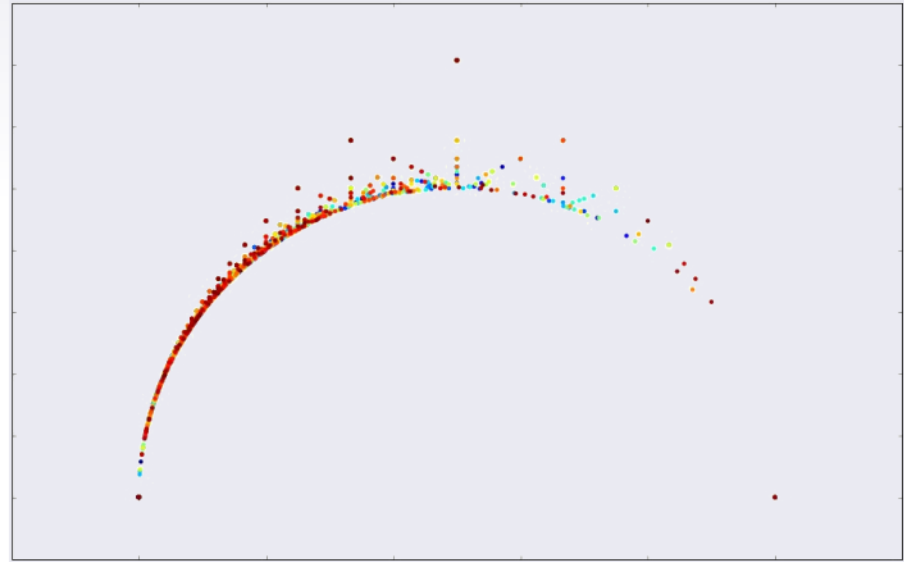
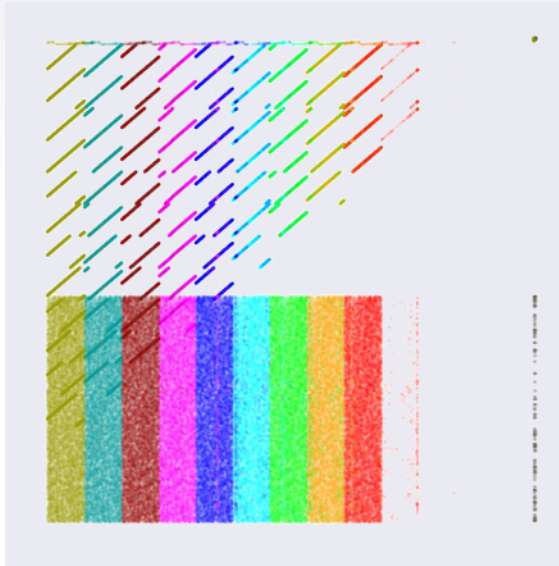
- Check if same rows have same label
- Find duplicated rows, understand why they are duplicated

Check if dataset is shuffled



Data leakage is the unauthorized transmission of data from within an organization to an external destination

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)