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

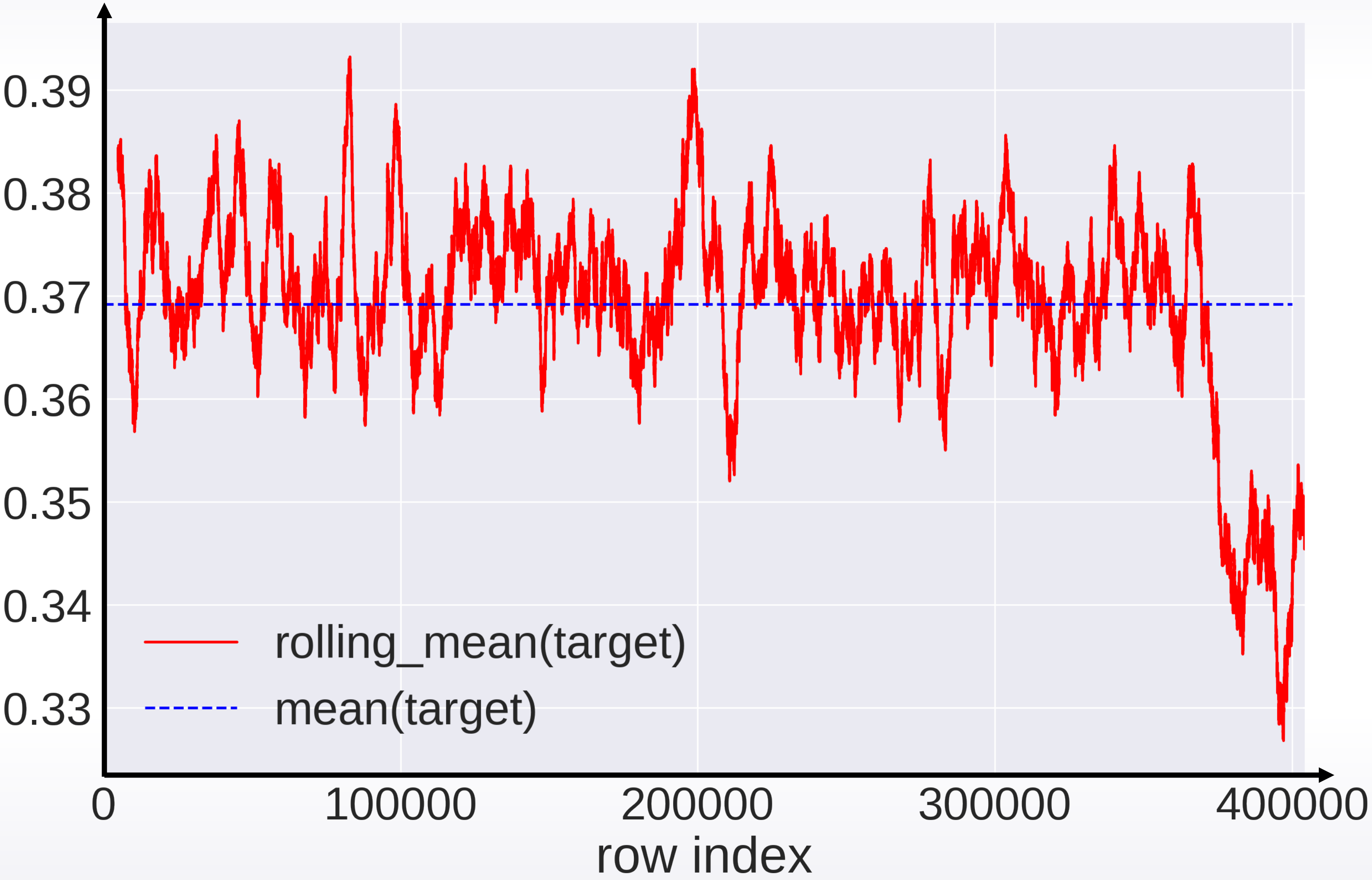
```
for f in categorical_feats:
   traintest[f] =raintest[f].factorize()

traintest.T.drop_duplicates()
```

Duplicated rows

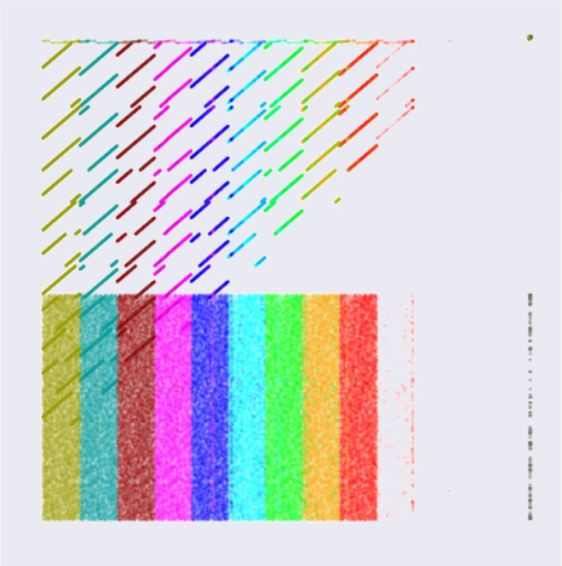
f1	f2	f3	y
13	34r9	A	**0**
13	34r9	A	**1**
13	34r9	A	**1**

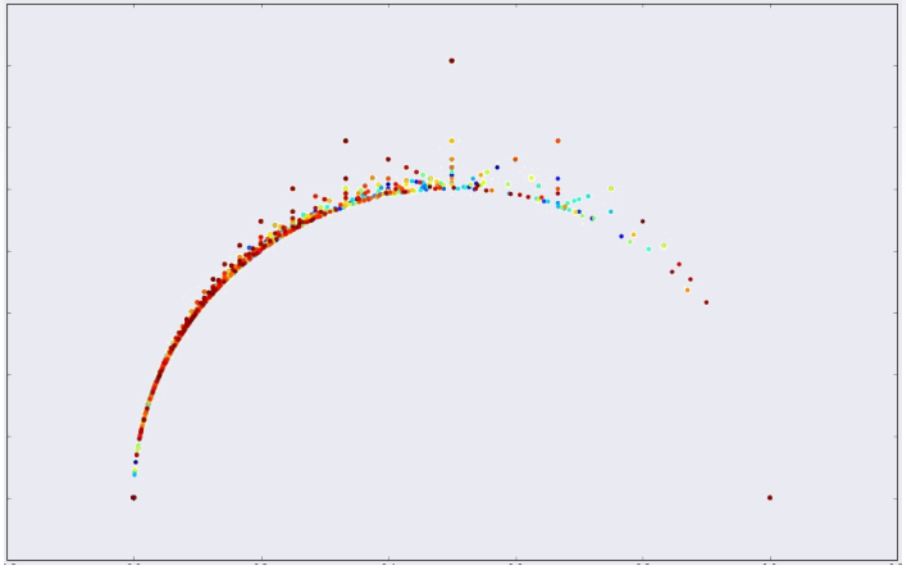
- Check if same rows have same label
- Find duplicated rows, understand why they are duplicated

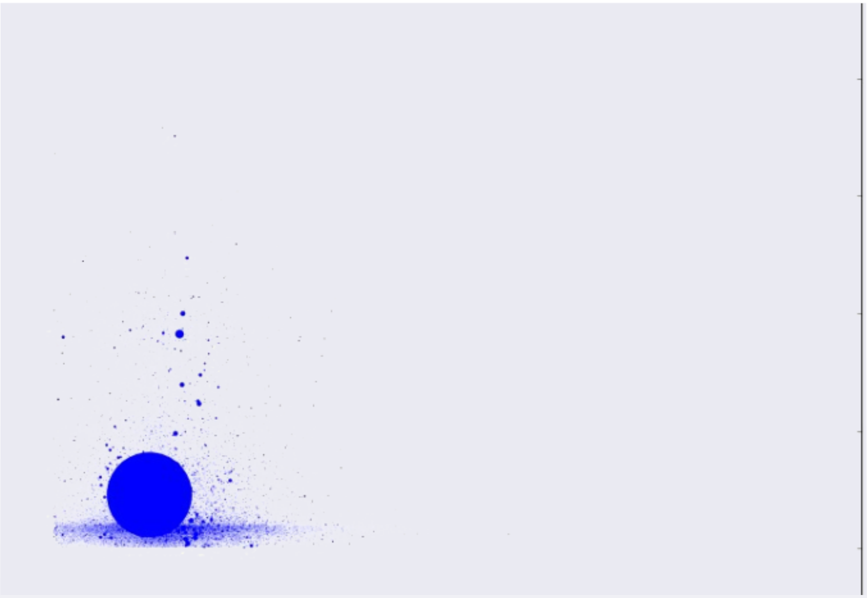
Check if dataset is shuffled

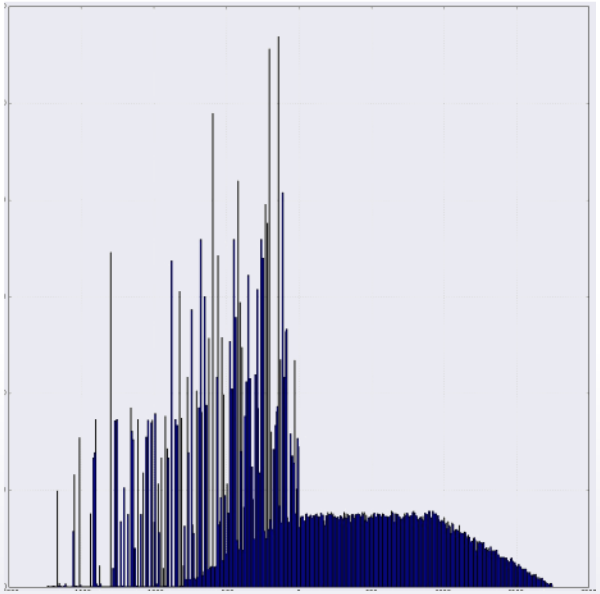


Cool visualizations









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)