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

The organizers could give us a fraction of objects they have or a fraction of features. And that is why we can have some issues with the data.

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
```

For example, we can encounter a feature which takes the same value for every object in both train and test set. This could be due to the sampling procedure.

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()
```

We can also have duplicated categorical features. The problem is that the features can be identical but their levels have different names. That is it can be possible to rename levels of one of the features and two columns will become identical. For example features f4 and f5. If we rename levels of the feature f5, C to A, A to B, and B to C. The result will look exactly as feature f4.

Duplicated and constant features

We need to label and code all the categorical features first, and then compare them as if they were numbers. The most important part here is label encoding. We need to do it right. We need to encode the features from top to bottom so that the first unique value we see gets label 1, the second gets 2 and so on. For example for feature f4, we will encode A with 1, B with 2 and C with 3. Now feature f5 will encode it differently C will be 1, A will be 2 and B will be 3.

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()
```

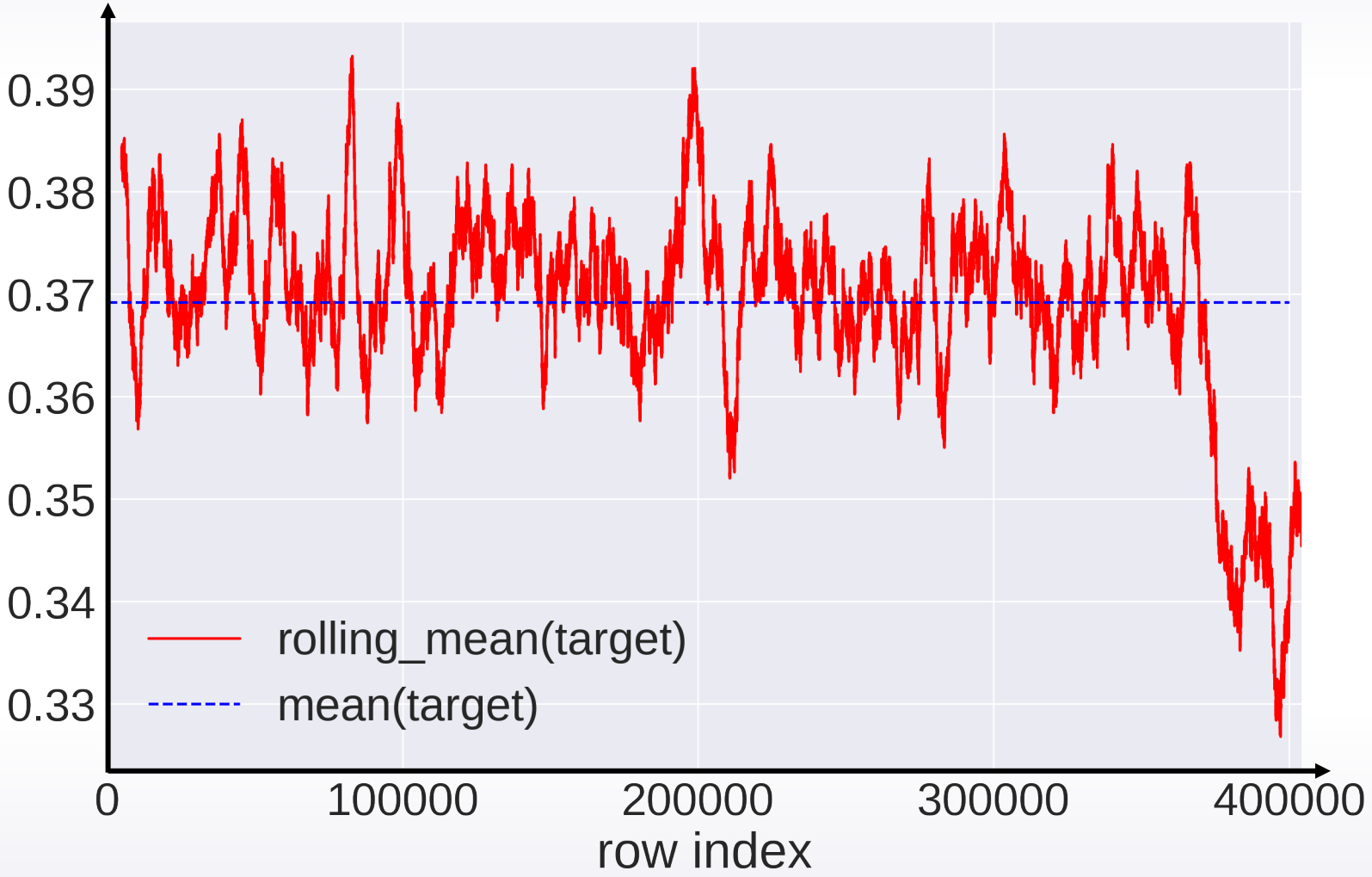
Duplicated rows

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

Finally, it is very useful to check that the data set is shuffled, because if it is not then, there is a high chance to find data leakage.



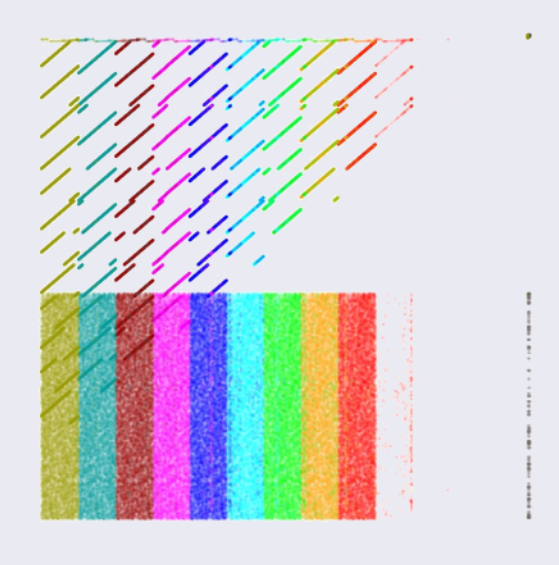
What we can do is we can plot a feature or target vector versus row index. We can optionally smooth the values using running average.

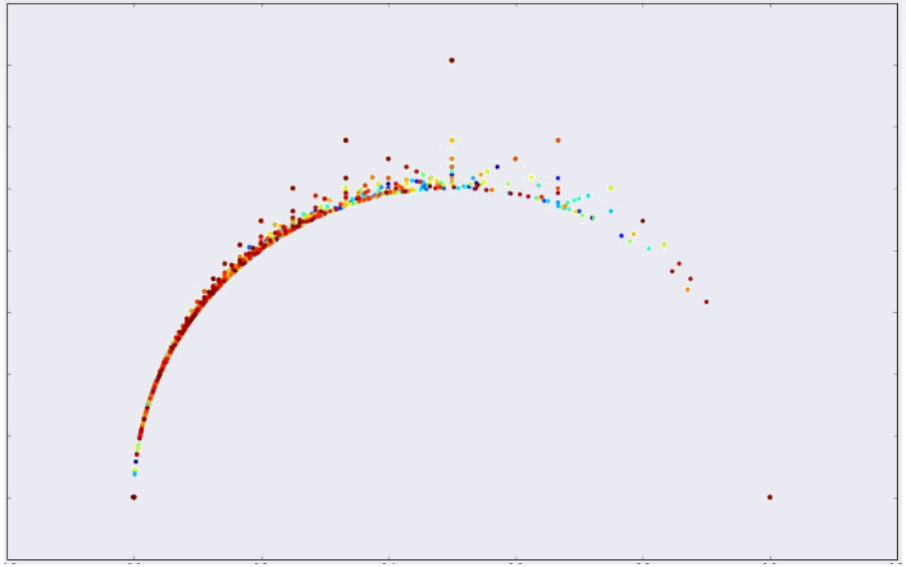
If the data was shuffled properly we would expect some kind of oscillation of the target values around the mean target value.

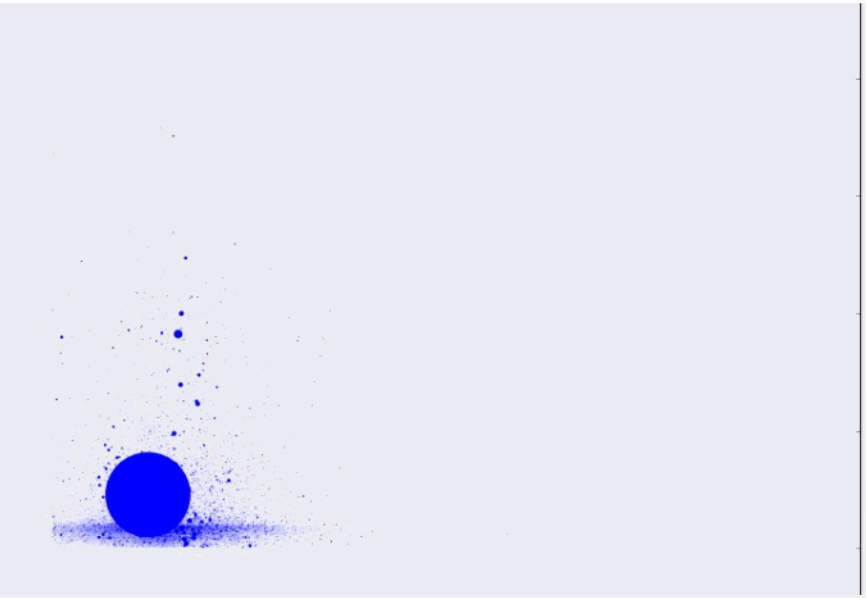
But in this case, it looks like the end of the train set is much different to the start, and we have some patterns.

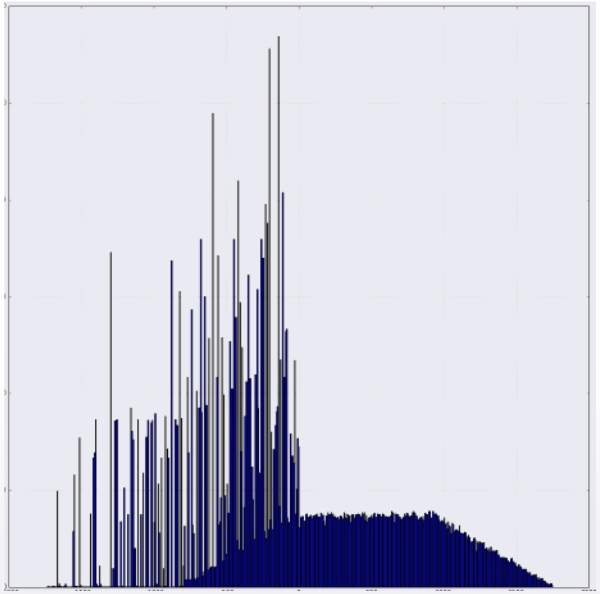
Maybe the information from this particular plot will not advance our model. But once again, we should find an explanation for all extraordinary things we observe.

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)