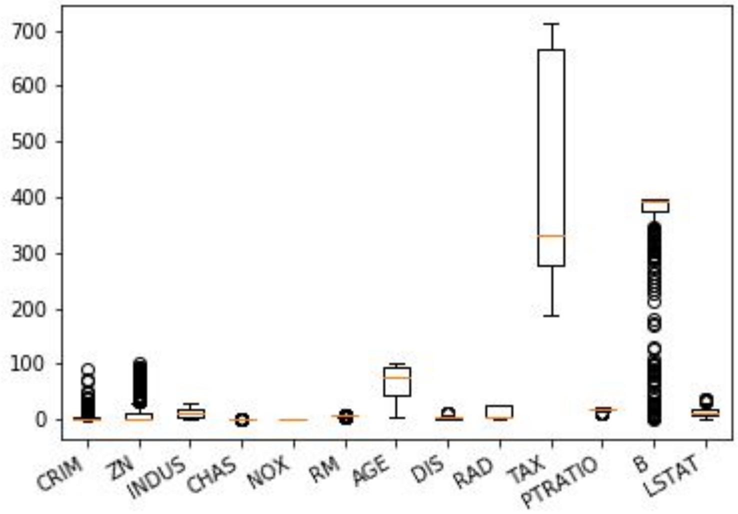
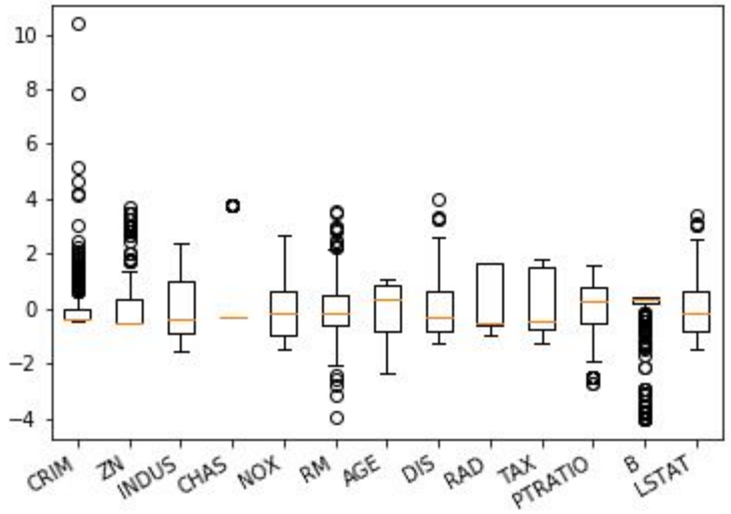
Aprendiz de Machine Learning

Data Preprocessing

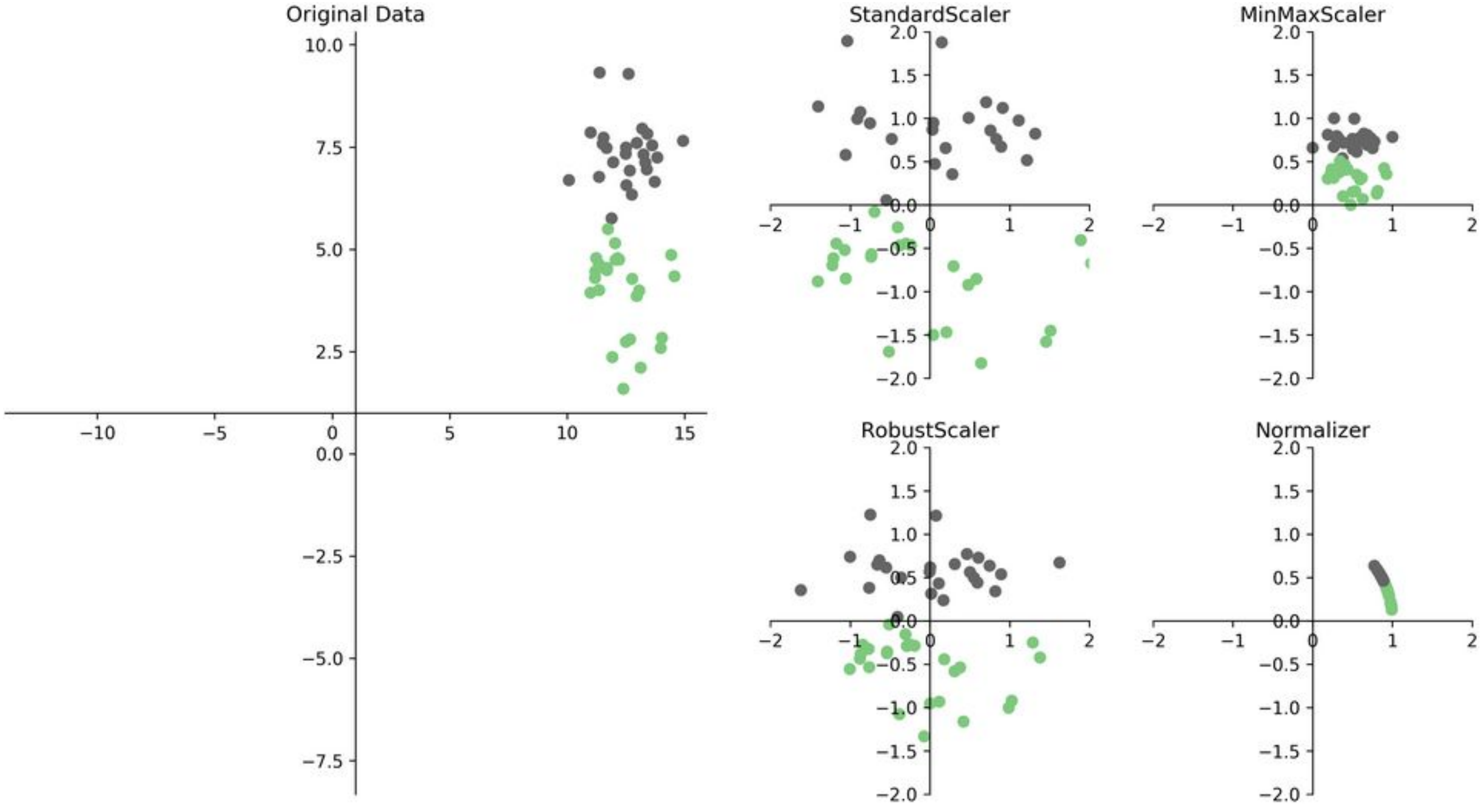
Continuous Variables

Standard Scaler





Scaling Alternatives

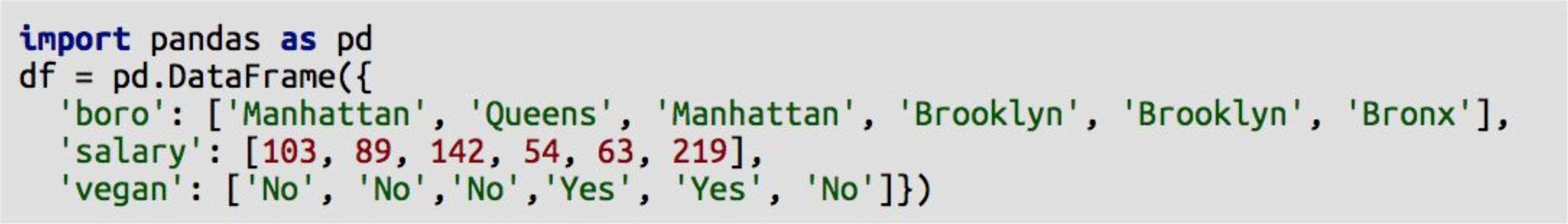


Sparse Data

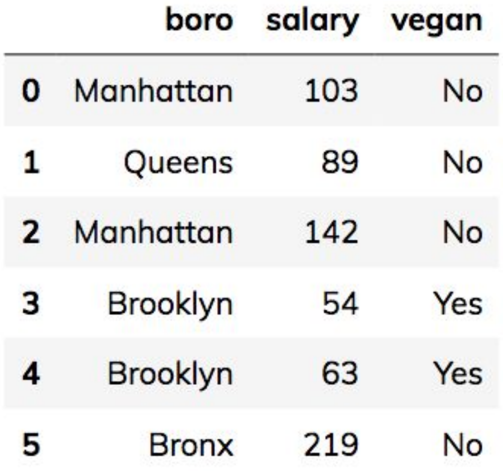
- Data with many zeros
- Don’t center: will destroy the sparnesess structure
- Use MaxAbsScaler:
  - don’t center
  - only scale between [-1, +1]

Categorical Variables

Categorical Variables

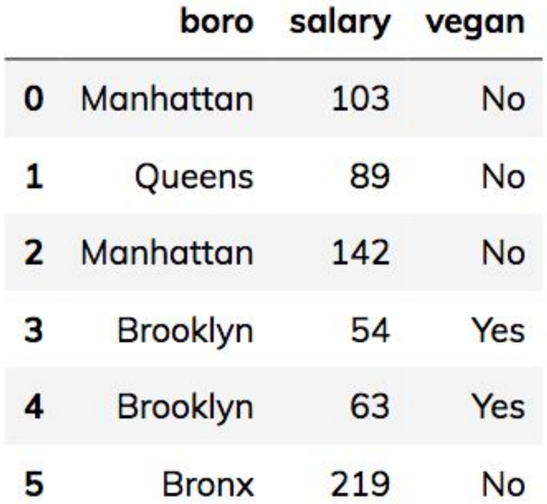


```
import pandas as pd
df = pd.DataFrame({
   'boro': ['Manhattan', 'Queens', 'Manhattan', 'Brooklyn', 'Brooklyn', 'Bronx'],
   'salary': [103, 89, 142, 54, 63, 219],
   'vegan': ['No', 'No','No','Yes', 'Yes', 'No']})
```

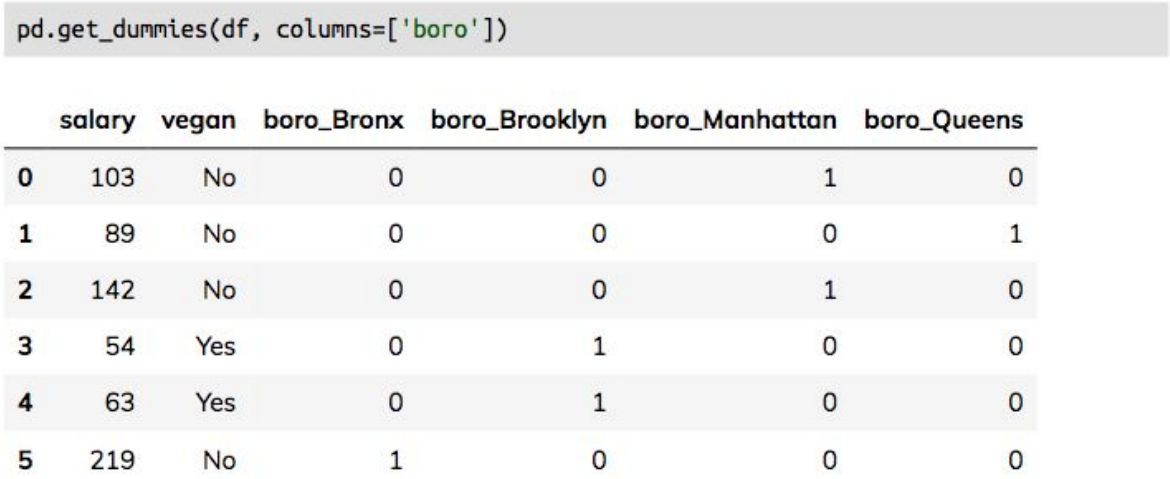


	boro	salary	vegan
0	Manhattan	103	No
1	Queens	89	No
2	Manhattan	142	No
3	Brooklyn	54	Yes
4	Brooklyn	63	Yes
5	Bronx	219	No

Dummy Encoding



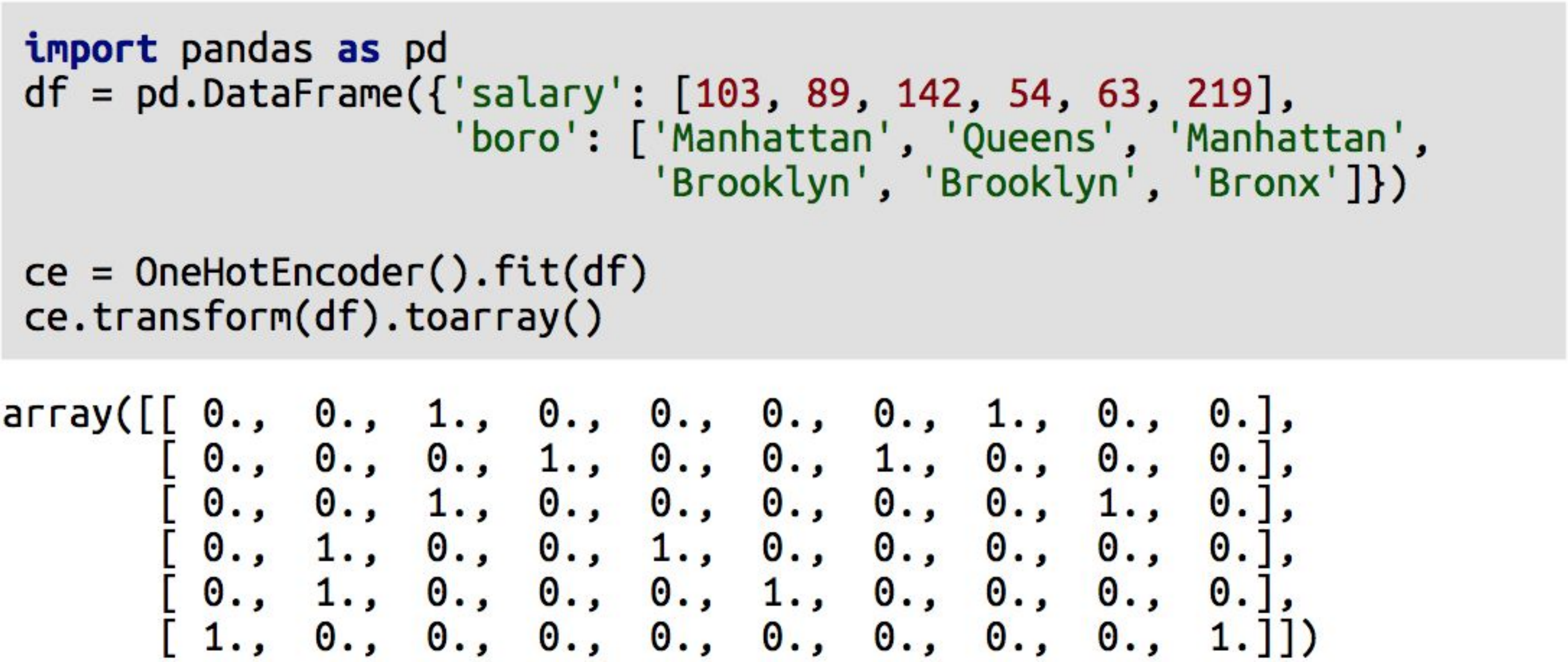
	boro	salary	vegan
0	Manhattan	103	No
1	Queens	89	No
2	Manhattan	142	No
3	Brooklyn	54	Yes
4	Brooklyn	63	Yes
5	Bronx	219	No



```
pd.get_dummies(df, columns=['boro'])
```

	salary	vegan	boro_Bronx	boro_Brooklyn	boro_Manhattan	boro_Queens
0	103	No	0	0	1	0
1	89	No	0	0	0	1
2	142	No	0	0	1	0
3	54	Yes	0	1	0	0
4	63	Yes	0	1	0	0
5	219	No	1	0	0	0

OneHotEncoder



```
import pandas as pd
df = pd.DataFrame({'salary': [103, 89, 142, 54, 63, 219],
                   'boro': ['Manhattan', 'Queens', 'Manhattan',
                            'Brooklyn', 'Brooklyn', 'Bronx']})

ce = OneHotEncoder().fit(df)
ce.transform(df).toarray()
```

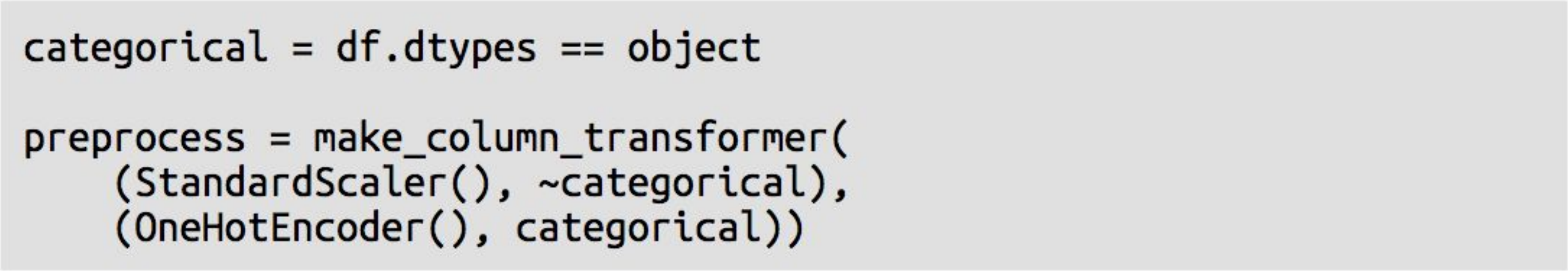
```
array([[ 0.,  0.,  1.,  0.,  0.,  0.,  0.,  1.,  0.,  0.],
       [ 0.,  0.,  0.,  1.,  0.,  0.,  1.,  0.,  0.,  0.],
       [ 0.,  0.,  1.,  0.,  0.,  0.,  0.,  0.,  1.,  0.],
       [ 0.,  1.,  0.,  0.,  1.,  0.,  0.,  0.,  0.,  0.],
       [ 0.,  1.,  0.,  0.,  0.,  1.,  0.,  0.,  0.,  0.],
       [ 1.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  1.]])
```

Collinearity Problem

- Dummy variables are redundant
- Last one is a linear combination: 1 - sum(others)
- Can drop one
- Keeping all can make the model more interpretable

All Variables

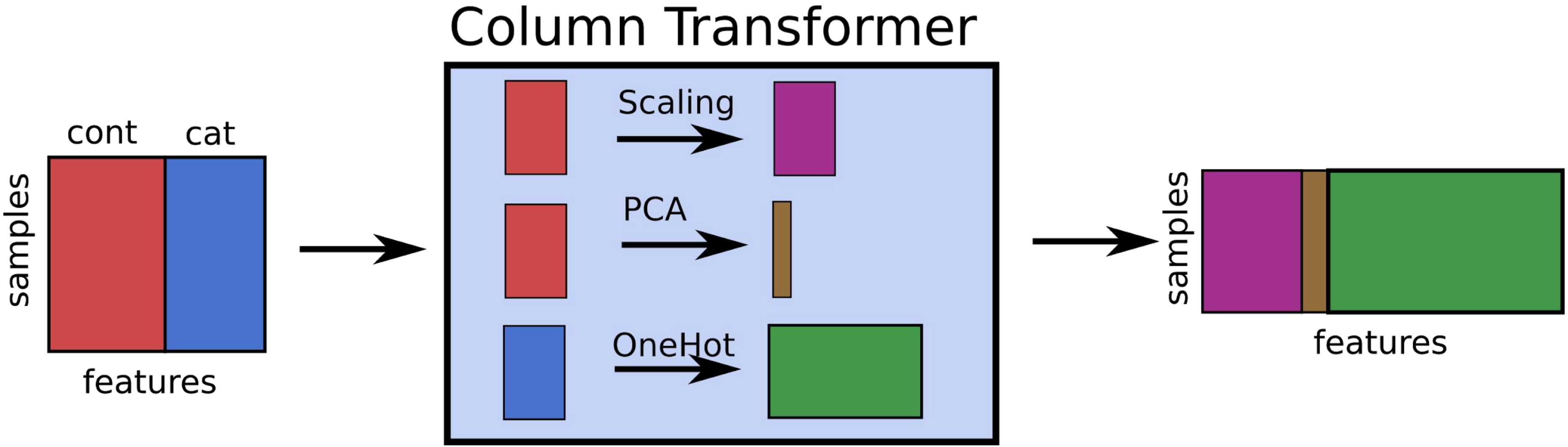
OneHotEncoder + ColumnTransformer



```
categorical = df.dtypes == object

preprocess = make_column_transformer(
    (StandardScaler(), ~categorical),
    (OneHotEncoder(), categorical))
```

ColumnTransformer



Thank You!