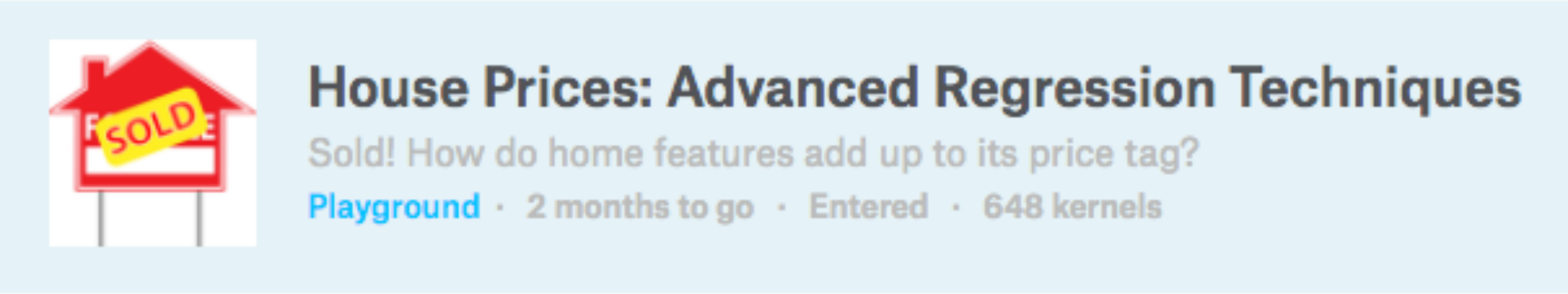
Feature engineering

- 1: discrete -> continuous
- 2: spatial engineering
- 3: semantics
- 4: time

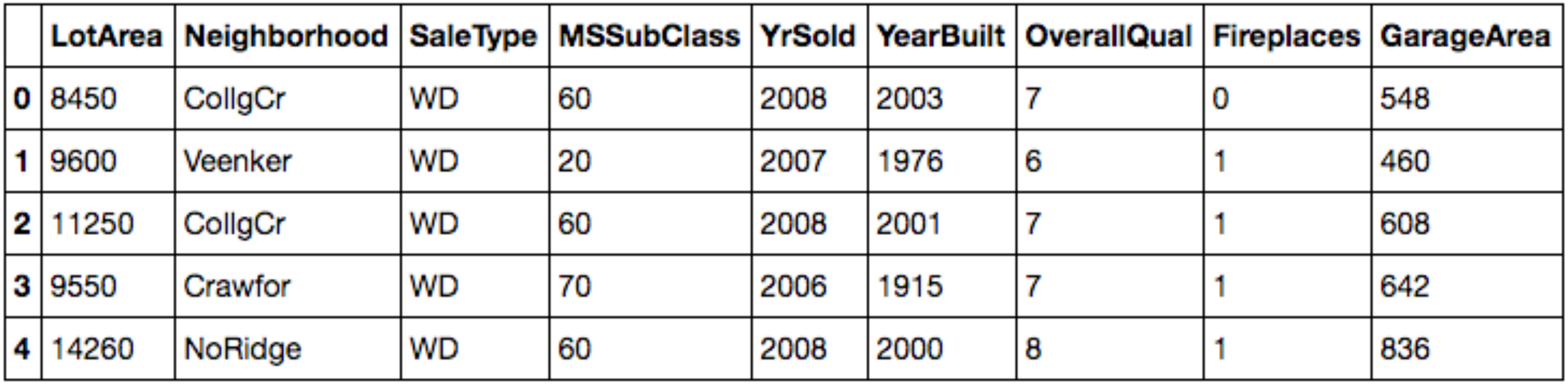
Feature Engineering 1: One Hot



House Prices: Advanced Regression Techniques

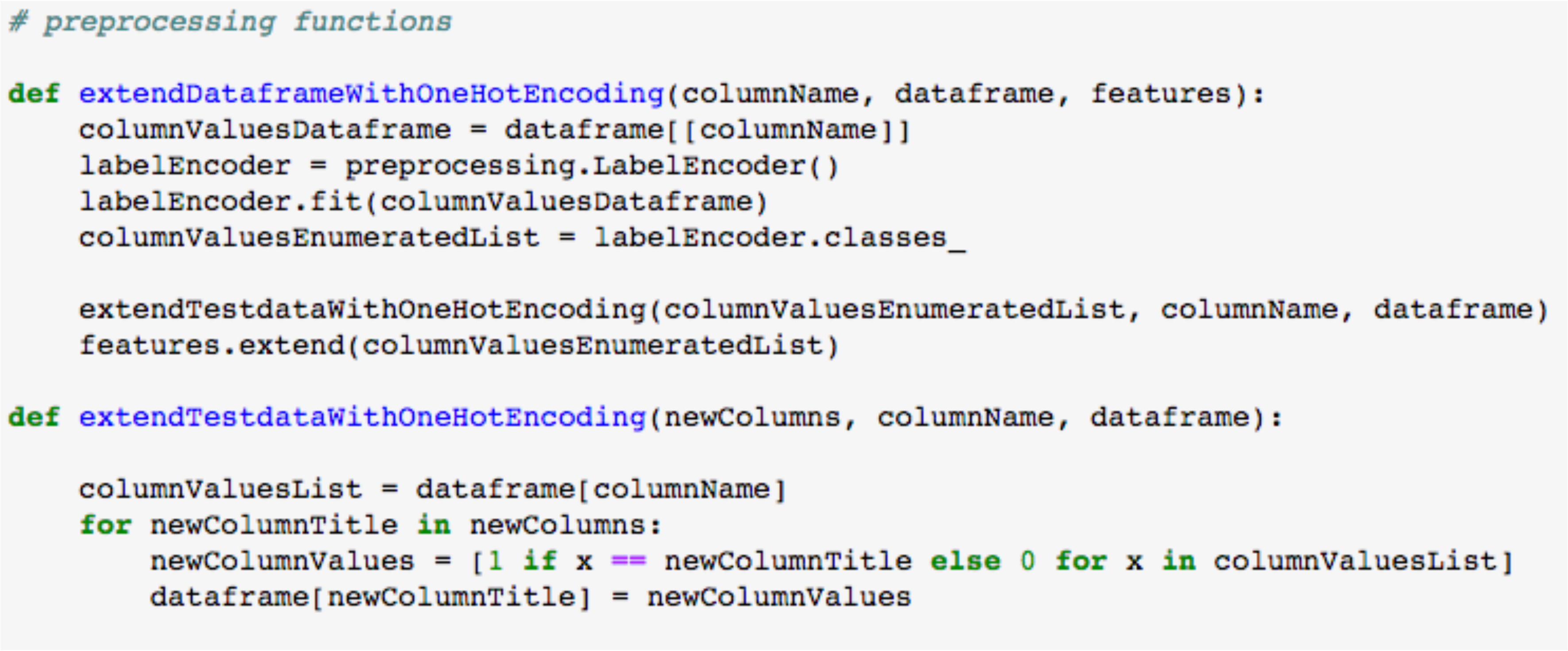
Sold! How do home features add up to its price tag?

Playground · 2 months to go · Entered · 648 kernels



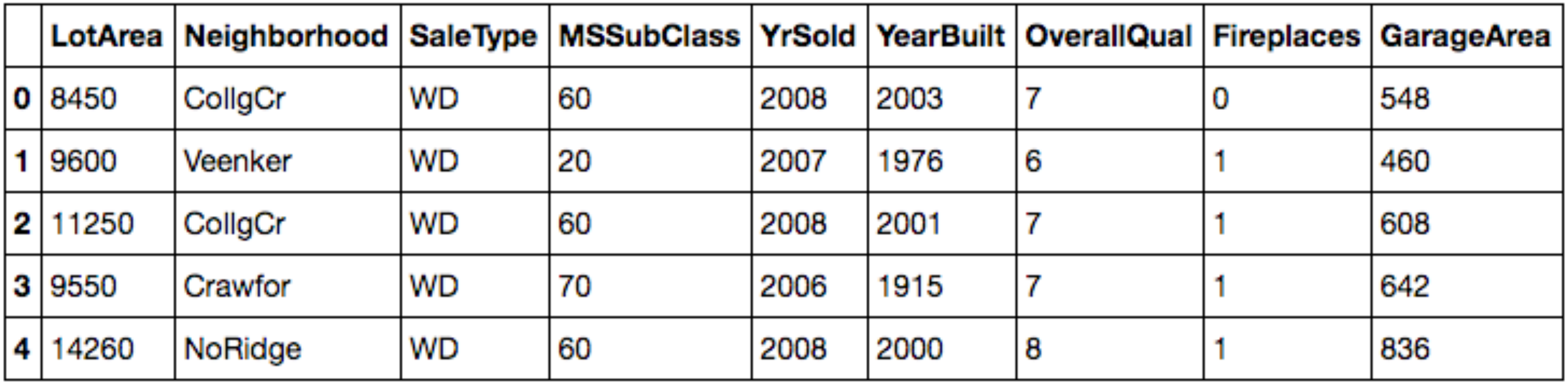
	LotArea	Neighborhood	SaleType	MSSubClass	YrSold	YearBuilt	OverallQual	Fireplaces	GarageArea
0	8450	CollgCr	WD	60	2008	2003	7	0	548
1	9600	Veenker	WD	20	2007	1976	6	1	460
2	11250	CollgCr	WD	60	2008	2001	7	1	608
3	9550	Crawfor	WD	70	2006	1915	7	1	642
4	14260	NoRidge	WD	60	2008	2000	8	1	836

Feature Engineering 1: One Hot

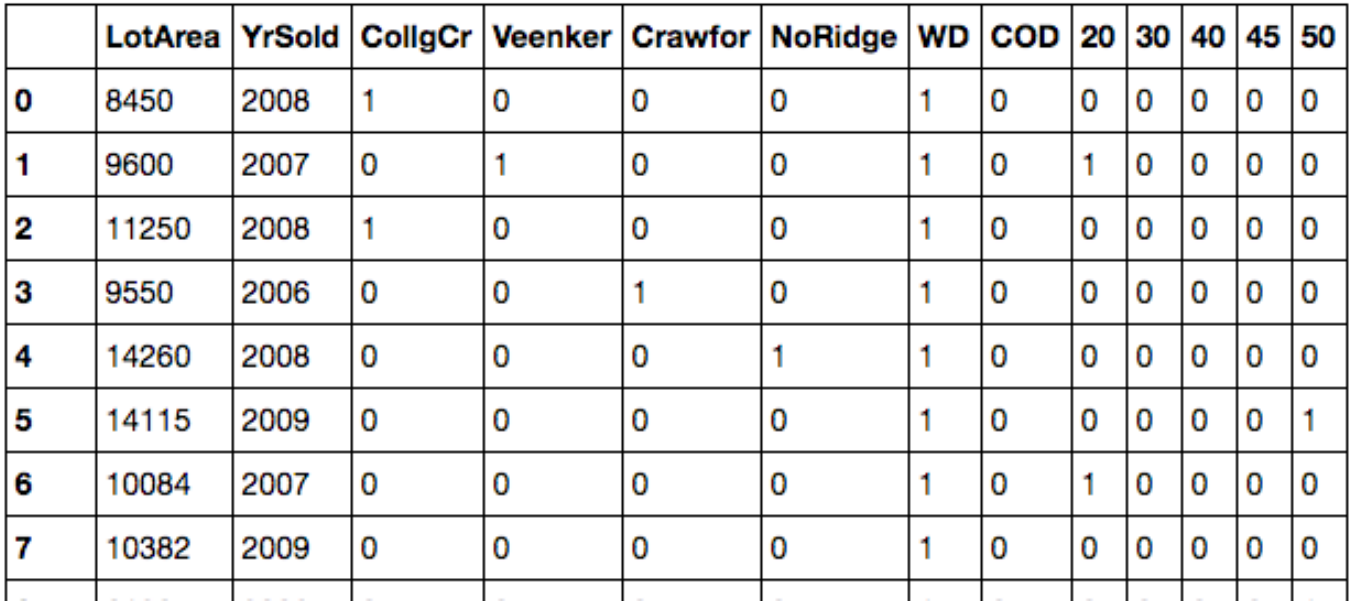


```python
# preprocessing functions

def extendDataframeWithOneHotEncoding(columnName, dataframe, features):
    columnValuesDataframe = dataframe[[columnName]]
    labelEncoder = preprocessing.LabelEncoder()
    labelEncoder.fit(columnValuesDataframe)
    columnValuesEnumeratedList = labelEncoder.classes_

    extendTestdataWithOneHotEncoding(columnValuesEnumeratedList, columnName, dataframe)
    features.extend(columnValuesEnumeratedList)

def extendTestdataWithOneHotEncoding(newColumns, columnName, dataframe):

    columnValuesList = dataframe[columnName]
    for newColumnTitle in newColumns:
        newColumnValues = [1 if x == newColumnTitle else 0 for x in columnValuesList]
        dataframe[newColumnTitle] = newColumnValues
```

Feature Engineering 1: One Hot

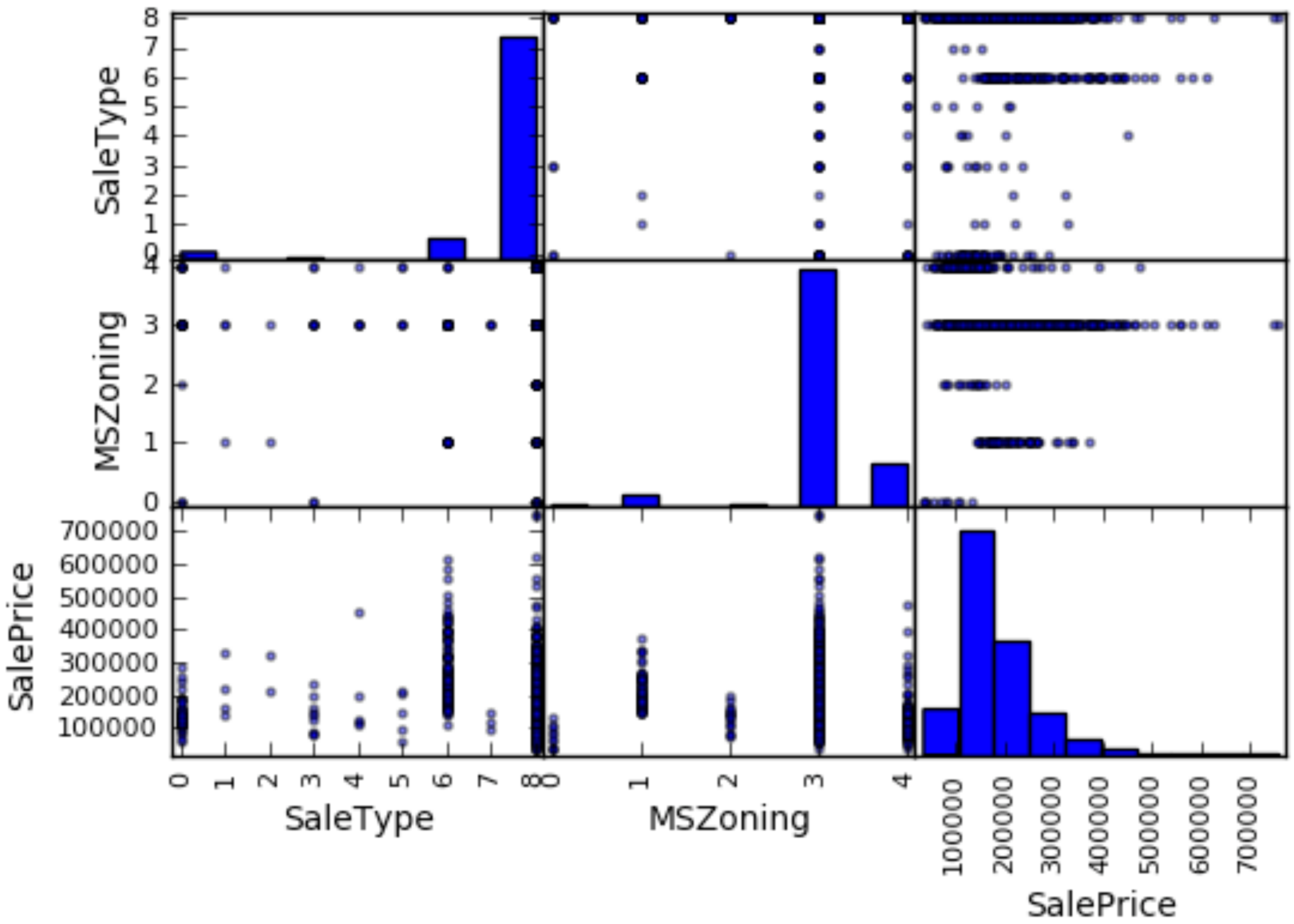


	LotArea	Neighborhood	SaleType	MSSubClass	YrSold	YearBuilt	OverallQual	Fireplaces	GarageArea
0	8450	CollgCr	WD	60	2008	2003	7	0	548
1	9600	Veenker	WD	20	2007	1976	6	1	460
2	11250	CollgCr	WD	60	2008	2001	7	1	608
3	9550	Crawfor	WD	70	2006	1915	7	1	642
4	14260	NoRidge	WD	60	2008	2000	8	1	836



	LotArea	YrSold	CollgCr	Veenker	Crawfor	NoRidge	WD	COD	20	30	40	45	50
0	8450	2008	1	0	0	0	1	0	0	0	0	0	0
1	9600	2007	0	1	0	0	1	0	1	0	0	0	0
2	11250	2008	1	0	0	0	1	0	0	0	0	0	0
3	9550	2006	0	0	1	0	1	0	0	0	0	0	0
4	14260	2008	0	0	0	1	1	0	0	0	0	0	0
5	14115	2009	0	0	0	0	1	0	0	0	0	0	1
6	10084	2007	0	0	0	0	1	0	1	0	0	0	0
7	10382	2009	0	0	0	0	1	0	0	0	0	0	0

MSZoning + SaleType vs SalePrice

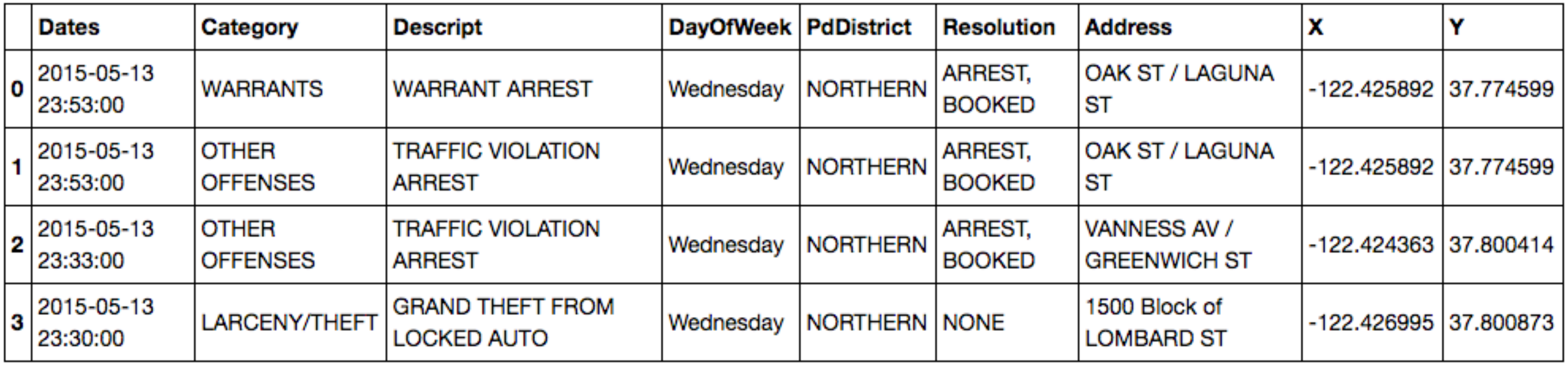


Feature Engineering 2, 3 & 4

San Francisco Crime Classification

Predict the category of crimes that occurred in the city by the bay

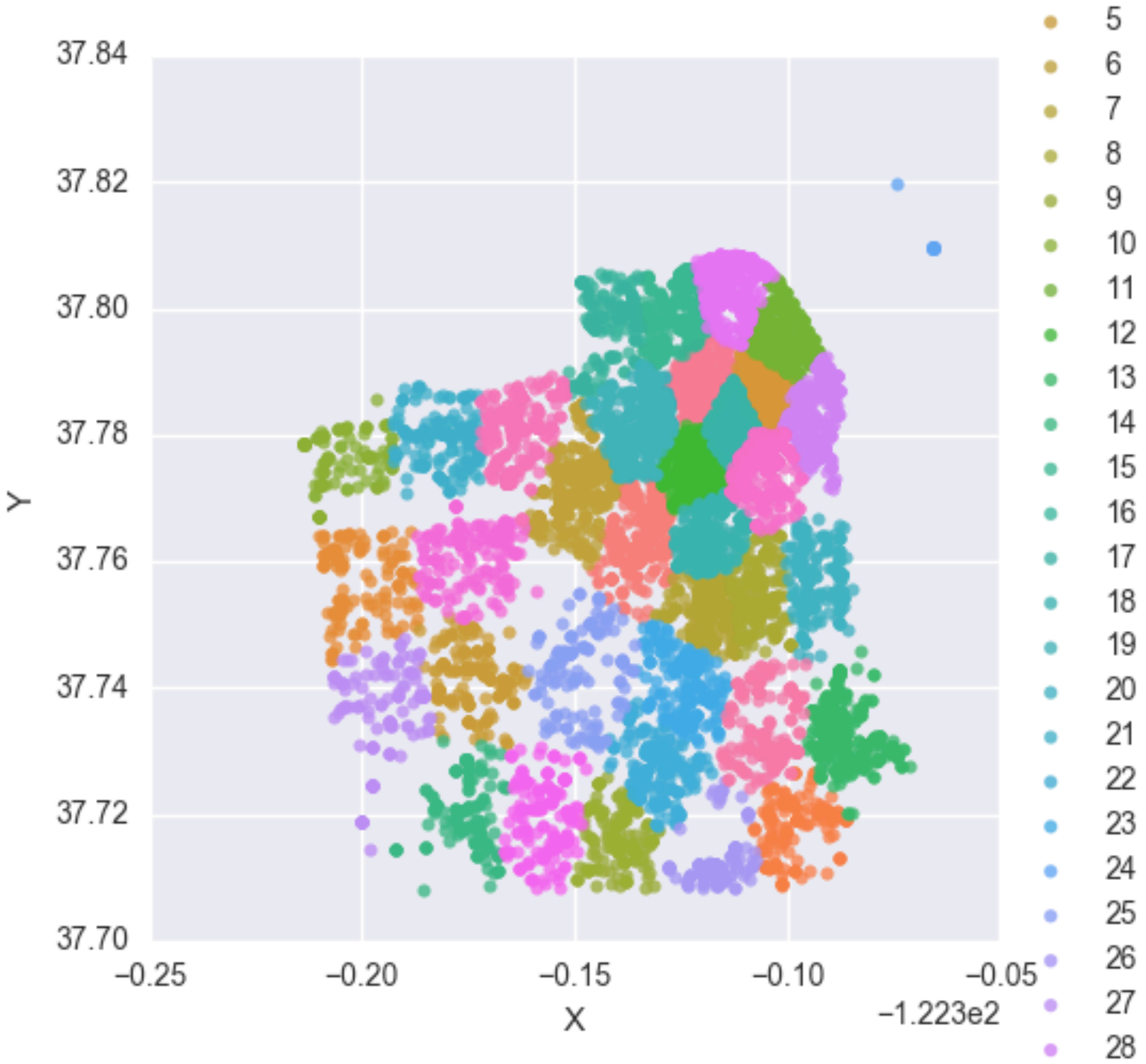
Playground · 7 months ago · 528 kernels



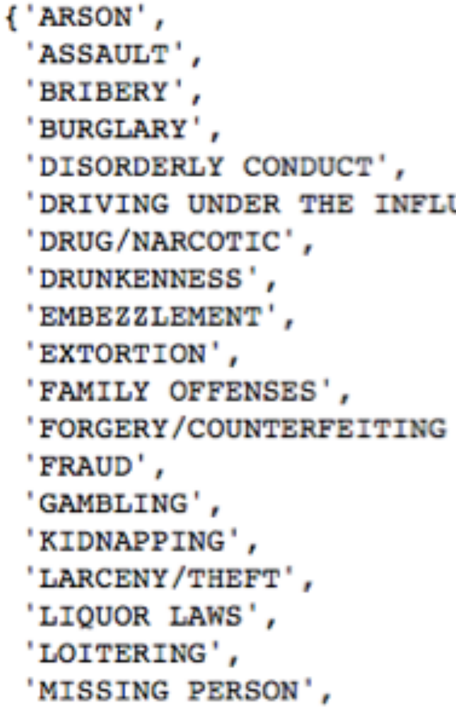
	Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y
0	2015-05-13 23:53:00	WARRANTS	WARRANT ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599
1	2015-05-13 23:53:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599
2	2015-05-13 23:33:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	VANNESS AV / GREENWICH ST	-122.424363	37.800414
3	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	NORTHERN	NONE	1500 Block of LOMBARD ST	-122.426995	37.800873

Feature Engineering 2: Spatial Data

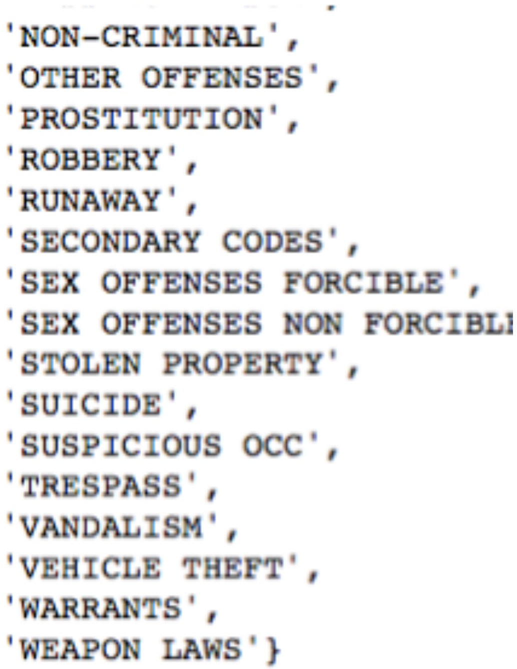
- K Means Clustering (k = 35)!
- ~1.5% improvement



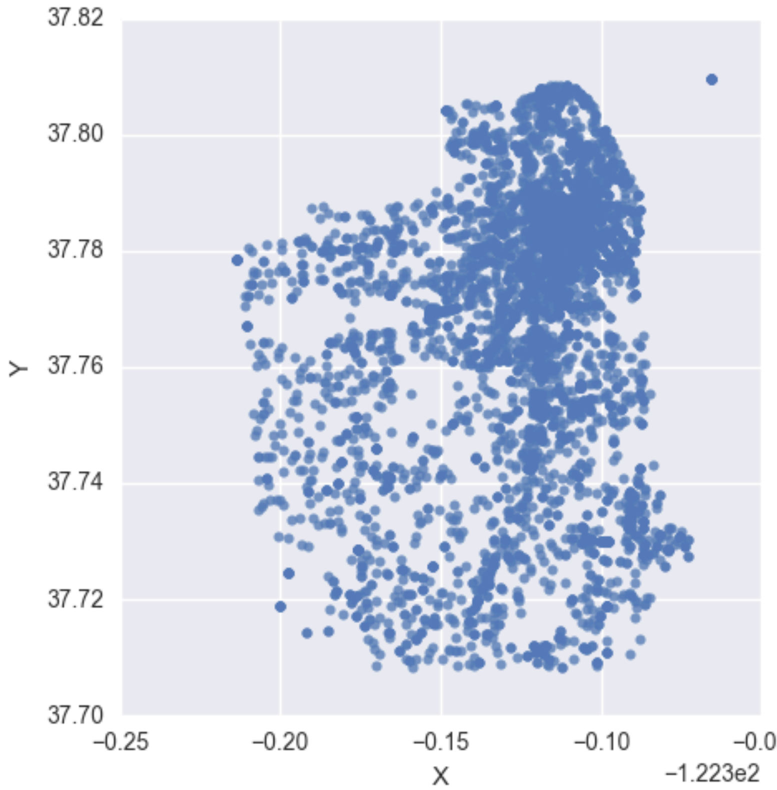
Feature Engineering 3: Semantics



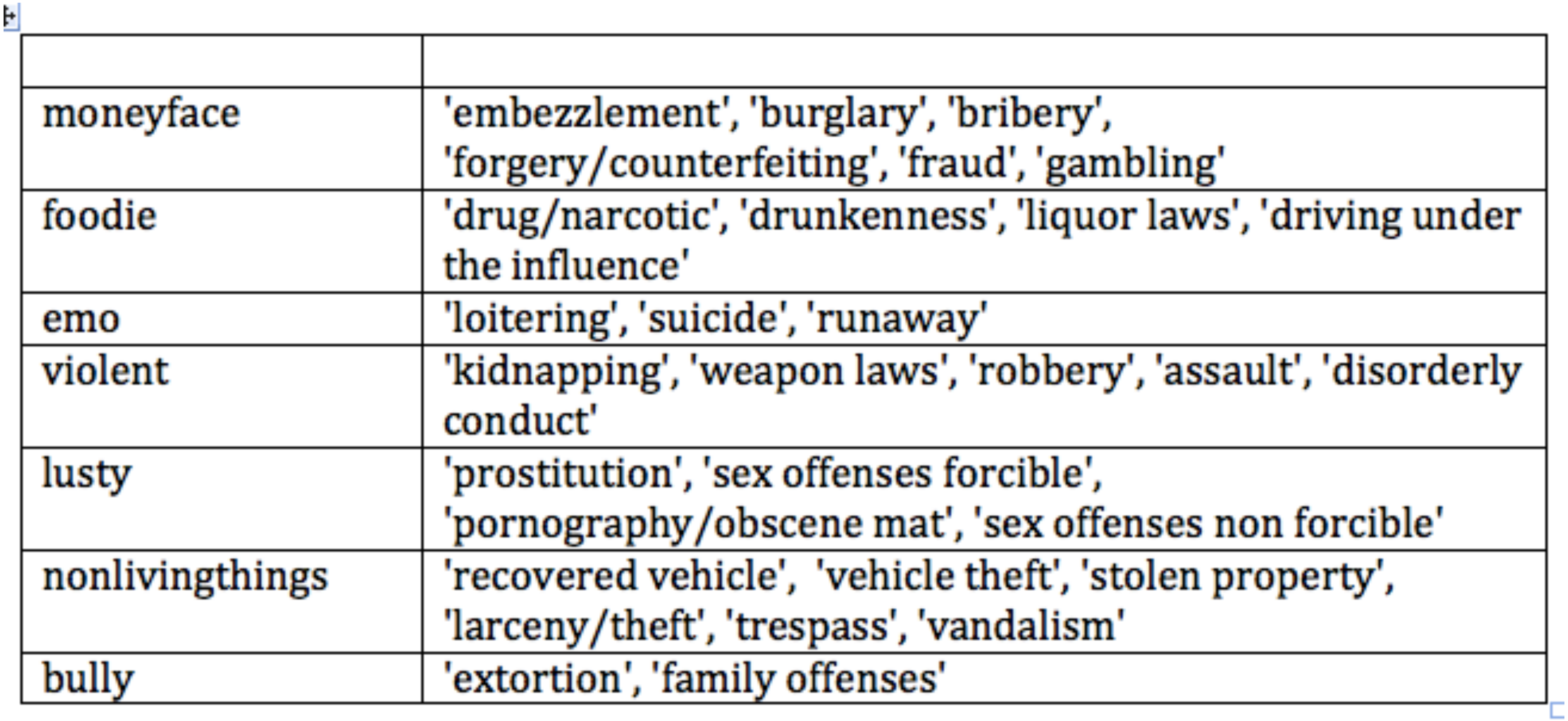
```
{'ARSON',
 'ASSAULT',
 'BRIBERY',
 'BURGLARY',
 'DISORDERLY CONDUCT',
 'DRIVING UNDER THE INFLUENCE',
 'DRUG/NARCOTIC',
 'DRUNKENNESS',
 'EMBEZZLEMENT',
 'EXTORTION',
 'FAMILY OFFENSES',
 'FORGERY/COUNTERFEITING
 'FRAUD',
 'GAMBLING',
 'KIDNAPPING',
 'LARCENY/THEFT',
 'LIQUOR LAWS',
 'LOITERING',
 'MISSING PERSON',
```



```
 'NON-CRIMINAL',
 'OTHER OFFENSES',
 'PROSTITUTION',
 'ROBBERY',
 'RUNAWAY',
 'SECONDARY CODES',
 'SEX OFFENSES FORCIBLE',
 'SEX OFFENSES NON FORCIBLI
 'STOLEN PROPERTY',
 'SUICIDE',
 'SUSPICIOUS OCC',
 'TRESPASS',
 'VANDALISM',
 'VEHICLE THEFT',
 'WARRANTS',
 'WEAPON LAWS'}
```



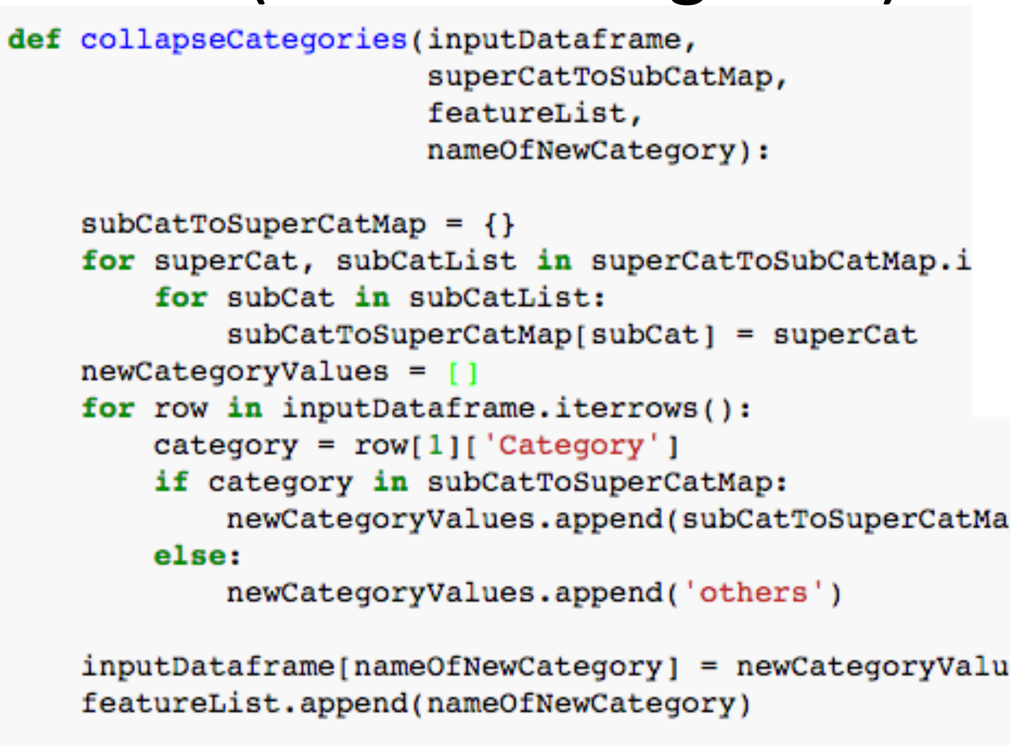
Feature Engineering 3: Semantics



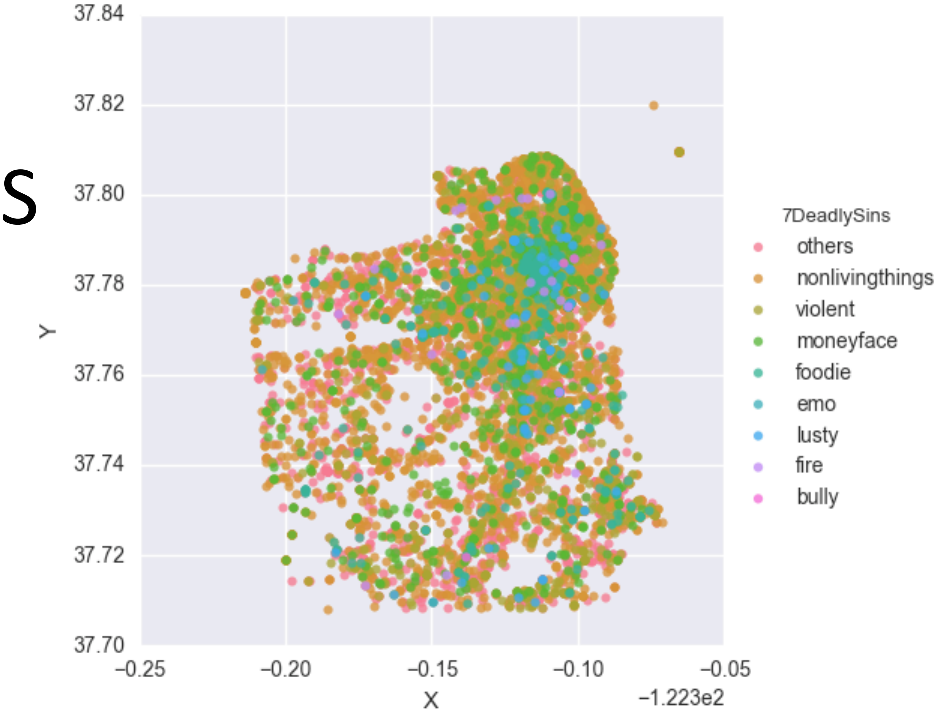
moneyface	'embezzlement', 'burglary', 'bribery', 'forgery/counterfeiting', 'fraud', 'gambling'
foodie	'drug/narcotic', 'drunkenness', 'liquor laws', 'driving under the influence'
emo	'loitering', 'suicide', 'runaway'
violent	'kidnapping', 'weapon laws', 'robbery', 'assault', 'disorderly conduct'
lusty	'prostitution', 'sex offenses forcible', 'pornography/obscene mat', 'sex offenses non forcible'
nonlivingthings	'recovered vehicle', 'vehicle theft', 'stolen property', 'larceny/theft', 'trespass', 'vandalism'
bully	'extortion', 'family offenses'

Feature Engineering 3: Semantics

- Reverse map is good! O(rows + categories) VS O(rows * categories)

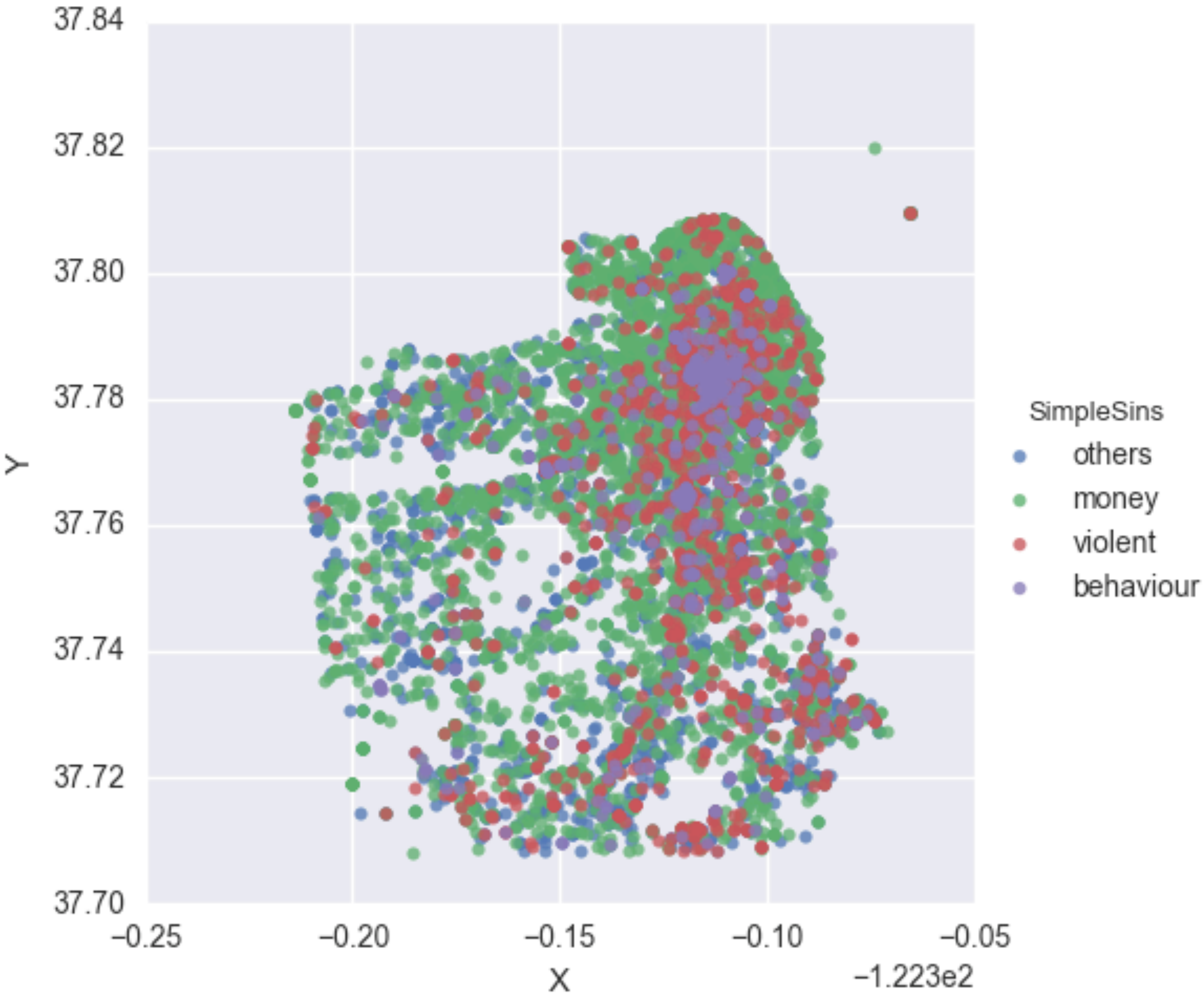


```
def collapseCategories(inputDataframe,
                       superCatToSubCatMap,
                       featureList,
                       nameOfNewCategory):

    subCatToSuperCatMap = {}
    for superCat, subCatList in superCatToSubCatMap.i
        for subCat in subCatList:
            subCatToSuperCatMap[subCat] = superCat
    newCategoryValues = []
    for row in inputDataframe.iterrows():
        category = row[1]['Category']
        if category in subCatToSuperCatMap:
            newCategoryValues.append(subCatToSuperCatMap[category])
        else:
            newCategoryValues.append('others')

    inputDataframe[nameOfNewCategory] = newCategoryValues
    featureList.append(nameOfNewCategory)
```

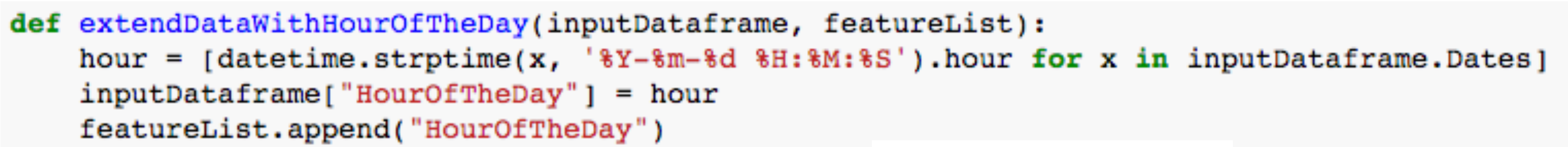


Feature Engineering 3: Semantics

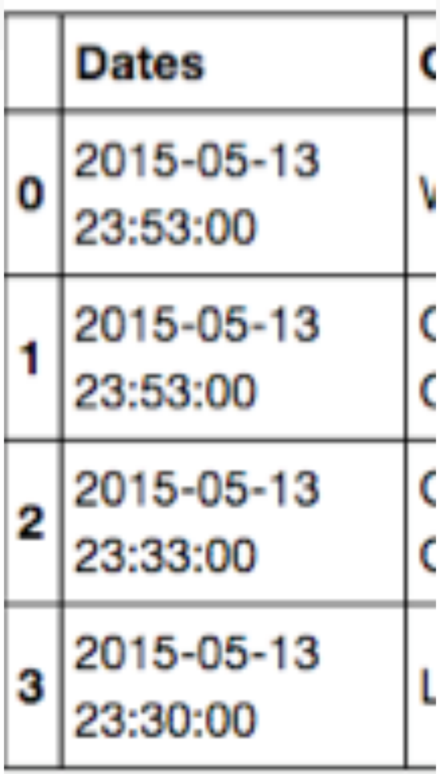
- Epicenters of crime!



Feature Engineering 4: Time

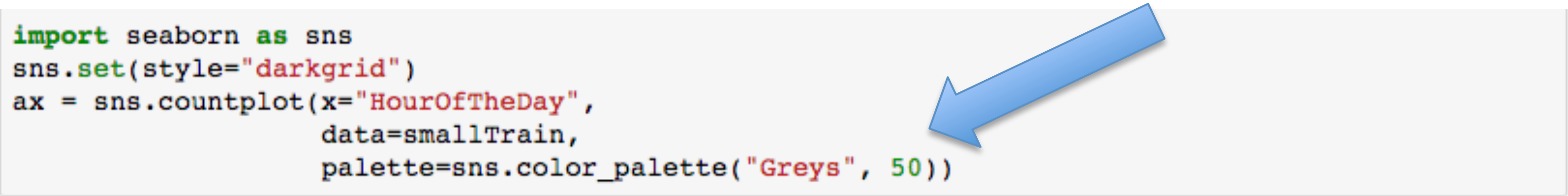


```
def extendDataWithHourOfTheDay(inputDataframe, featureList):
    hour = [datetime.strptime(x, '%Y-%m-%d %H:%M:%S').hour for x in inputDataframe.Dates]
    inputDataframe["HourOfTheDay"] = hour
    featureList.append("HourOfTheDay")
    print set(hour)
```

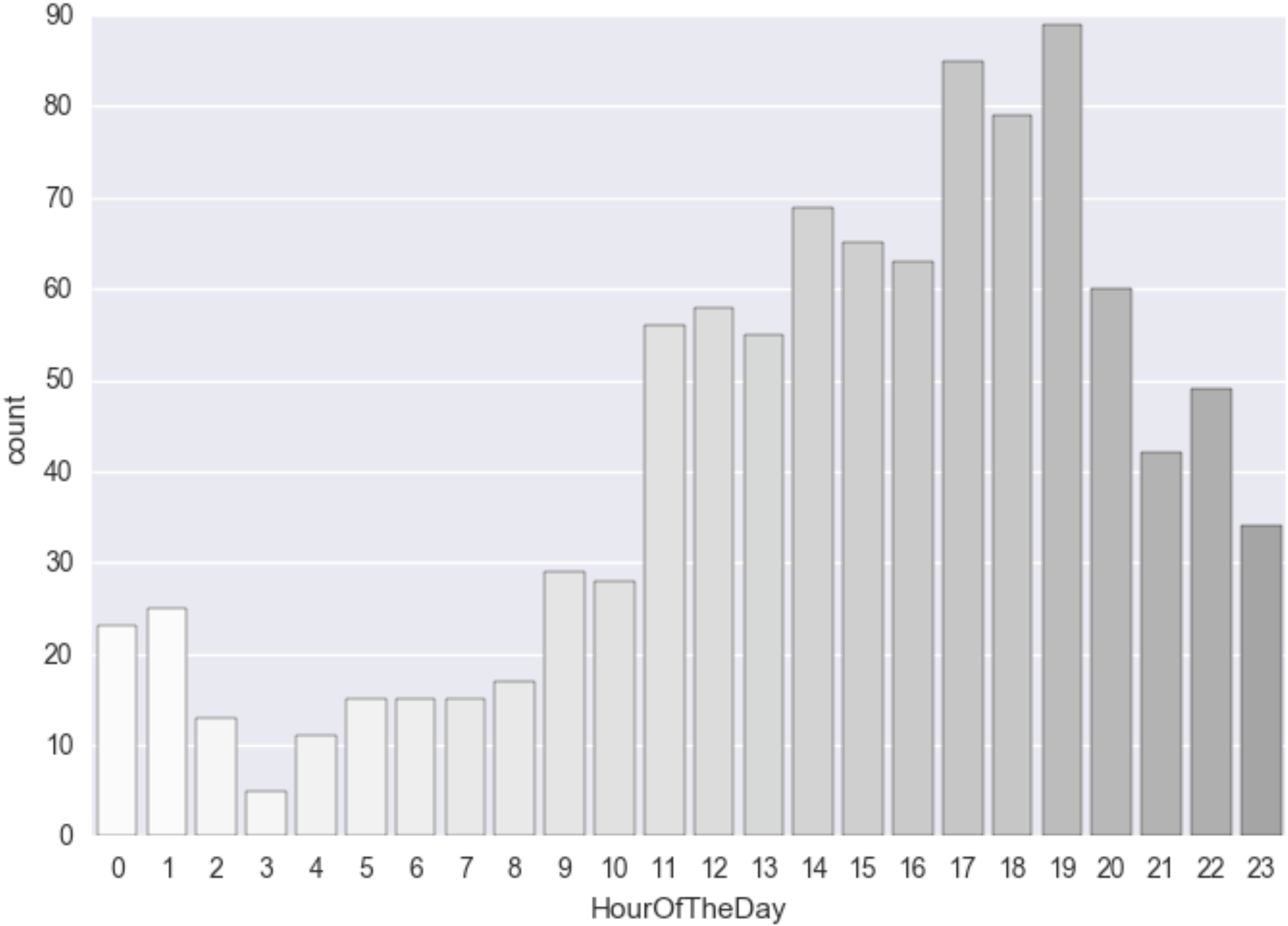


	Dates	C
0	2015-05-13 23:53:00	W
1	2015-05-13 23:53:00	C C
2	2015-05-13 23:33:00	C C
3	2015-05-13 23:30:00	L

Feature Engineering 4: Time

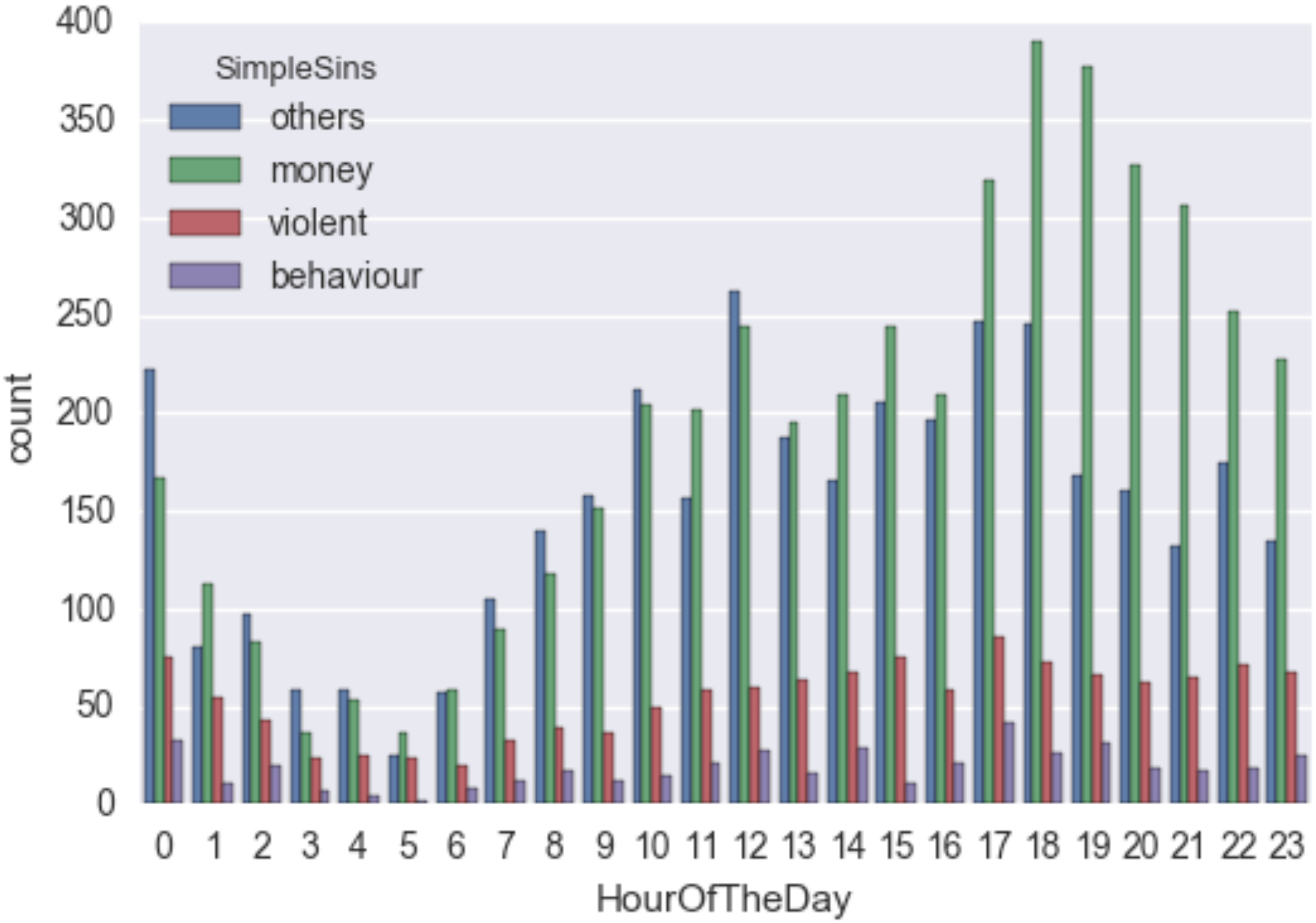


```
import seaborn as sns
sns.set(style="darkgrid")
ax = sns.countplot(x="HourOfTheDay",
                   data=smallTrain,
                   palette=sns.color_palette("Greys", 50))
```



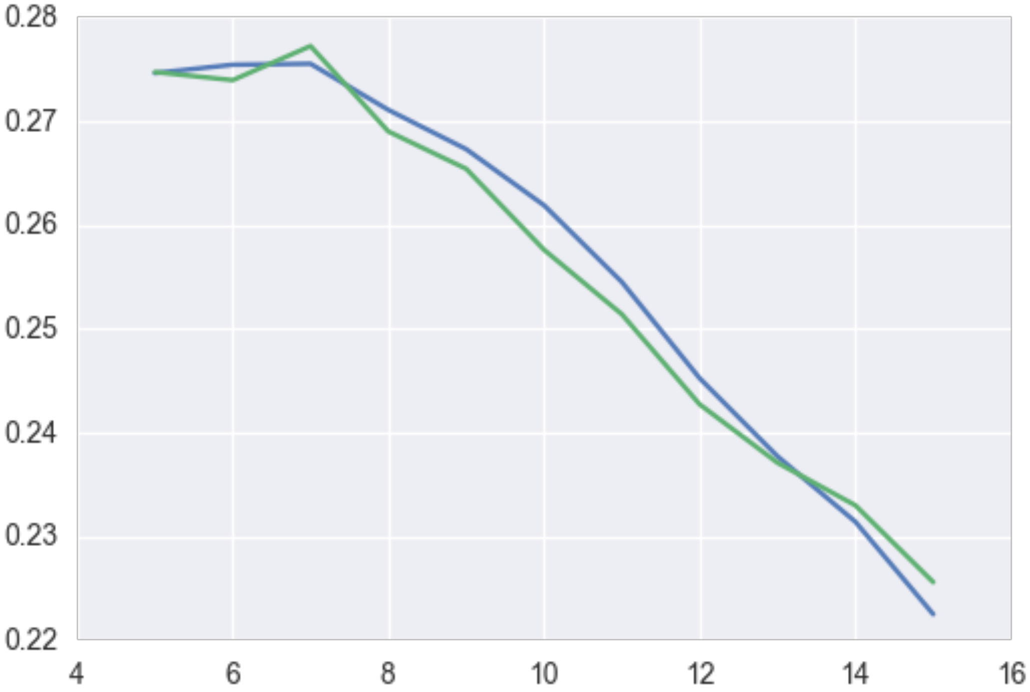
Feature Engineering 4: Time + semantics

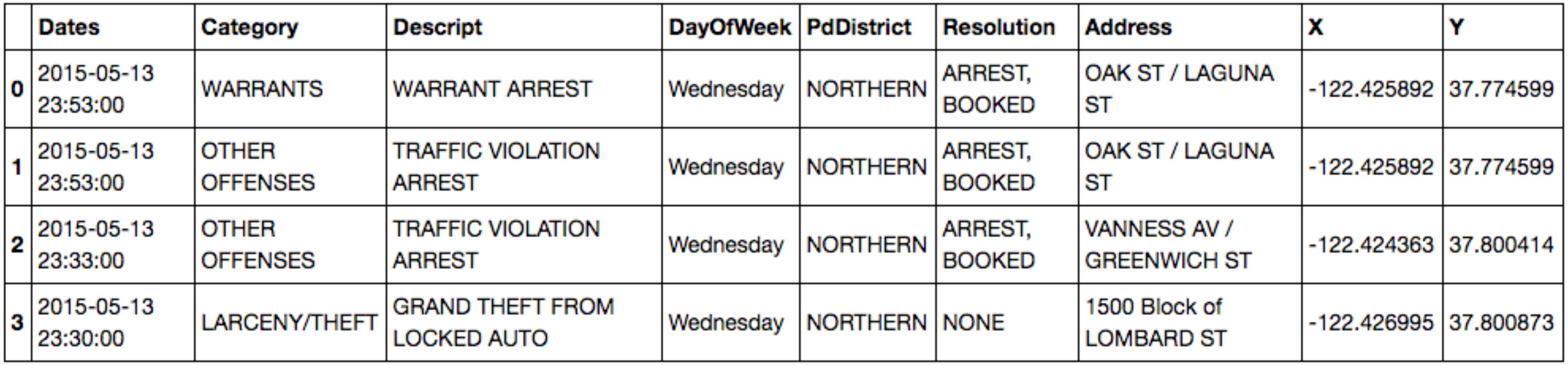
- Crime profile by time of day



Running the algorithms

- Elbow plot
- Description is a deal breaker





	Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y
0	2015-05-13 23:53:00	WARRANTS	WARRANT ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599
1	2015-05-13 23:53:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599
2	2015-05-13 23:33:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	VANNESS AV / GREENWICH ST	-122.424363	37.800414
3	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	NORTHERN	NONE	1500 Block of LOMBARD ST	-122.426995	37.800873