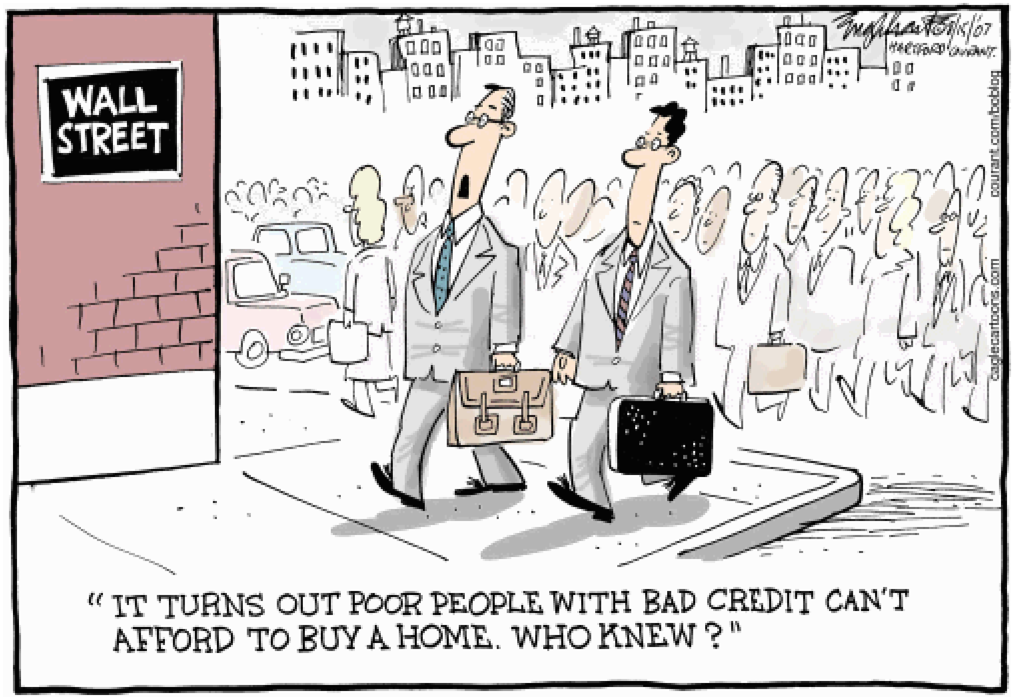
Exploring origination trends in single-family mortgage data

JEANA CURRO

DATA SCIENCE- FINAL PROJECT

Table of contents:

The mortgage lending we all know:



But has mortgage lending changed? Yes!

Problem Statement:

Using Freddie Mac loan level origination data from 2006 and 2014, we will explore different origination trends pre and post the housing crisis.

Hypothesis:

We hypothesize that loans originated prior to the crisis generally show weaker credit characteristics (e.g. lower credit scores and higher debt to income ratios) than loans originated post crisis.

Past research:

1. The Urban Institute, July 21 2015: *The Credit Shows Early Signs of Loosening*
2. Federal Reserve Bank of Kansas City, July 7 2014: *Tight Credit Conditions Continue to Constrain the Housing Recovery*
3. Federal Reserve Board Divisions of Research, Statistics and Monetary Affairs, November 2008: *The Rise in Mortgage Defaults*

Our goal: to update these findings and add some statistical significance

The dataset:

- Released November 2016
- Approximately 22.5 million single-family residential mortgage loans originated between January 1, 1999 – September 30, 2015 that are currently guaranteed by Freddie Mac
- We will use a smaller **sample dataset** which consists of one origination file/year (1999-2015)
- Each yearly file contains origination data on 50,000 loans randomly selected from each origination year
- The 2015 file is only populated through September 2015, so only has 37,500 loans.
- Links to the publicly available dataset:
  - Dataset download: http://www.freddiemac.com/news/finance/sf_loanlevel_dataset.html
  - Dataset user guide: http://www.freddiemac.com/news/finance/pdf/user_guide.pdf

The (26!) features given:

VARIABLE	TYPE	VARIABLE	TYPE
fico	integer	channel	categorical (1)
first_pay_date	date	prepay	categorical (1)
first_time_homebuyer	categorical (1)	product	categorical (5)
mature_date	date	state	categorical (2)
msa	categorical (5)	prop_type	categorical (2)
mi_pct	float	prop_zip	categorical (5)
units	integer	loan_id	categorical (12)
occupancy	categorical (1)	purpose	categorical (1)
cltv	float	orig_term	integer
dti	float	num_borrowers	integer
orig_bal	float	seller	categorical (20)
ltv	float	servicer	categorical (20)
init_rate	float	super_conforming_flag	categorical (1)

http://www.freddiemac.com/news/finance/pdf/user_guide.pdf

Analysis Step 1: data munging

1. We use 2006 and 2014 origination files as proxies for pre and post housing crisis:
   - 100,000 loans total, 50,000 for each origination year.
2. Merge the two datasets into one
3. Eliminate some rows: control for the most generic type of mortgage:
   - 30 year mortgage
   - fixed rate
   - single family unit
   - no prepayment penalty

Data munging (cont'd)

4. Transform categorical variables to Boolean ones: MI%, occupancy, channel, purpose, orig yr

In [16]:
```
orig_subset.sample(10)
```

Out[16]:

	fico	mi_pct	occupancy	dti	orig_bal	ltv	init_rate	channel	state	purpose	num_borrowers	seller	servicer	orig_yr
loan_id														
F106Q2074101	780	0	O	25	125000	80.0	6.625	T	NY	P	1.0	WELLSFARGOBANK,NA	WELLSFARGOBANK,NA	2006
F106Q4030733	729	0	O	26	110000	80.0	6.750	R	IN	C	1.0	CHASEHOMEFINANCELLC	JPMORGANCHASEBANK,NA	2006
F106Q3039726	681	0	I	58	48000	80.0	7.375	T	MN	P	2.0	USBANKNA	USBANKNA	2006

In [25]:
```
#we change the colu  hea ers for Occupancy, Channel, MI_ ct, and Ori  Yr to make them more intuitive:
orig_subset=orig_su set.re ame(columns = {'occupancy':'ow er_occ', 'c annel':'retail','mi_pct':'mi','purpose':'purchase'
orig_subset.sample(
```

Out[25]:

	fico	mi	owner_occ	dti	orig_bal	ltv	init_rate	retail	state	purchase	num_borrowers	seller	servicer	orig_precrisis
loan_id														
F114Q1038299	740.0	0	1	29.0	264000	61.0	4.625	0	VA	0	1.0	PENNYMACCORP	PENNYMACCORP	0
F114Q3114180	729.0	1	1	21.0	179000	95.0	4.250	1	MT	1	2.0	WELLSFARGOBANK,NA	WELLSFARGOBANK,NA	0
F106Q2029430	691.0	0	1	34.0	131000	53.0	6.750	1	IN	0	2.0	Other sellers	Other servicers	1

Data munging (cont'd)

5. Drop categorical features with too many dummies
   - state, seller, and servicer
   - We can always add back in later!

```
orig_subset=orig_subset.drop(['state','seller','servicer'],axis=1)
print orig_subset.dtypes
print orig_subset.shape
```

```
fico                float64
mi                    int64
owner_occ             int64
dti                 float64
orig_bal              int64
ltv                 float64
init_rate           float64
retail                int64
purchase              int64
num_borrowers       float64
orig_precrisis        int64
dtype: object
(54817, 11)
```

Now we have 11 features, all numeric.
Next step, EDA!

Analysis Step 2: EDA, check correlation

1. Check for correlated variables: drop MI (correlated with LTV), drop rate (correlated w orig yr)

In [27]:

```
orig_subset.corr()
#mi and ltv are highly correlated (59%) - makes sense since loans with >80% need to have MI
#orig year and rate are super highly correlated (-95%); we will drop rate
```

Out[27]:

	fico	mi	owner_occ	dti	orig_bal	ltv	init_rate	retail	purchase	num_borrowers	orig_precrisis
fico	1.000000	-0.026487	-0.093066	-0.163893	0.095348	-0.045062	-0.325834	0.061234	0.198110	-0.028981	-0.273143
mi	-0.026487	1.000000	0.108517	0.014701	-0.030235	0.599915	-0.163232	-0.000672	0.278975	-0.028610	-0.204144
owner_occ	-0.093066	0.108517	1.000000	0.024907	0.113040	0.053670	-0.025081	-0.020667	-0.065044	-0.012242	0.038253
dti	-0.163893	0.014701	0.024907	1.000000	0.098343	0.048135	0.142423	-0.071995	-0.041548	-0.063890	0.134969
orig_bal	0.095348	-0.030235	0.113040	0.098343	1.000000	0.050505	-0.235868	-0.074053	-0.014282	0.177651	-0.187765
ltv	-0.045062	0.599915	0.053670	0.048135	0.050505	1.000000	-0.120956	-0.031758	0.371818	-0.008417	-0.162118
init_rate	-0.325834	-0.163232	-0.025081	0.142423	-0.235868	-0.120956	1.000000	-0.134933	-0.185723	0.027835	0.946250
retail	0.061234	-0.000672	-0.020667	-0.071995	-0.074053	-0.031758	-0.134933	1.000000	0.006814	0.004405	-0.157173
purchase	0.198110	0.278975	-0.065044	-0.041548	-0.014282	0.371818	-0.185723	0.006814	1.000000	0.007372	-0.182110
num_borrowers	-0.028981	-0.028610	-0.012242	-0.063890	0.177651	-0.008417	0.027835	0.004405	0.007372	1.000000	0.047334
orig_precrisis	-0.273143	-0.204144	0.038253	0.134969	-0.187765	-0.162118	0.946250	-0.157173	-0.182110	0.047334	1.000000

EDA (cont'd): descriptive statistics

2. Check descriptive statistics:

```
In [35]: orig_subset.groupby(['orig_precrisis']).describe().round()
         # most post-crisis originations have smaller variance; bc most pre-crisis loans allowed for lower minimums
```

Out[35]:

		dti	fico	ltv	num_borrowers	orig_bal	owner_occ	purchase	retail
orig_precrisis									
0 (2014)	**count**	24073.0	24073.0	24073.0	24073.0	24073.0	24073.0	24073.0	24073.0
	mean	34.0	749.0	78.0	2.0	223751.0	1.0	1.0	1.0
	std	9.0	44.0	15.0	0.0	120336.0	0.0	0.0	0.0
	min	1.0	600.0	7.0	1.0	11000.0	0.0	0.0	0.0
	25%	28.0	718.0	72.0	1.0	130000.0	1.0	0.0	0.0
	50%	35.0	756.0	80.0	2.0	200000.0	1.0	1.0	1.0
	75%	42.0	785.0	90.0	2.0	299000.0	1.0	1.0	1.0
	max	50.0	832.0	95.0	2.0	626000.0	1.0	1.0	1.0
1 (2006)	**count**	30034.0	30034.0	30034.0	30034.0	30034.0	30034.0	30034.0	30034.0
	mean	37.0	719.0	72.0	2.0	182802.0	1.0	0.0	0.0
	std	12.0	58.0	16.0	0.0	92265.0	0.0	0.0	0.0
	min	2.0	300.0	7.0	1.0	14000.0	0.0	0.0	0.0
	25%	29.0	676.0	65.0	1.0	112000.0	1.0	0.0	0.0
	50%	37.0	722.0	79.0	2.0	164000.0	1.0	0.0	0.0
	75%	45.0	767.0	80.0	2.0	238000.0	1.0	1.0	1.0
	max	65.0	850.0	100.0	2.0	626000.0	1.0	1.0	1.0

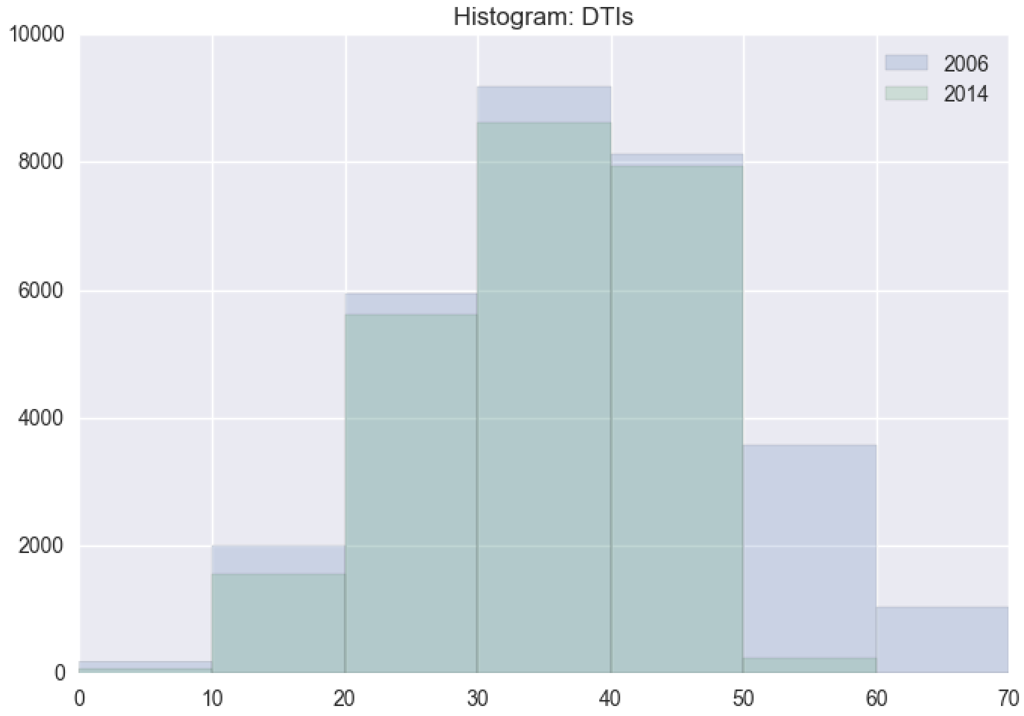
2006 has larger variance for most stats than **2014**.

Potentially due to lower minimums (ficos) or higher maximums (DTI, LTV)

EDA (cont'd): check data distribution

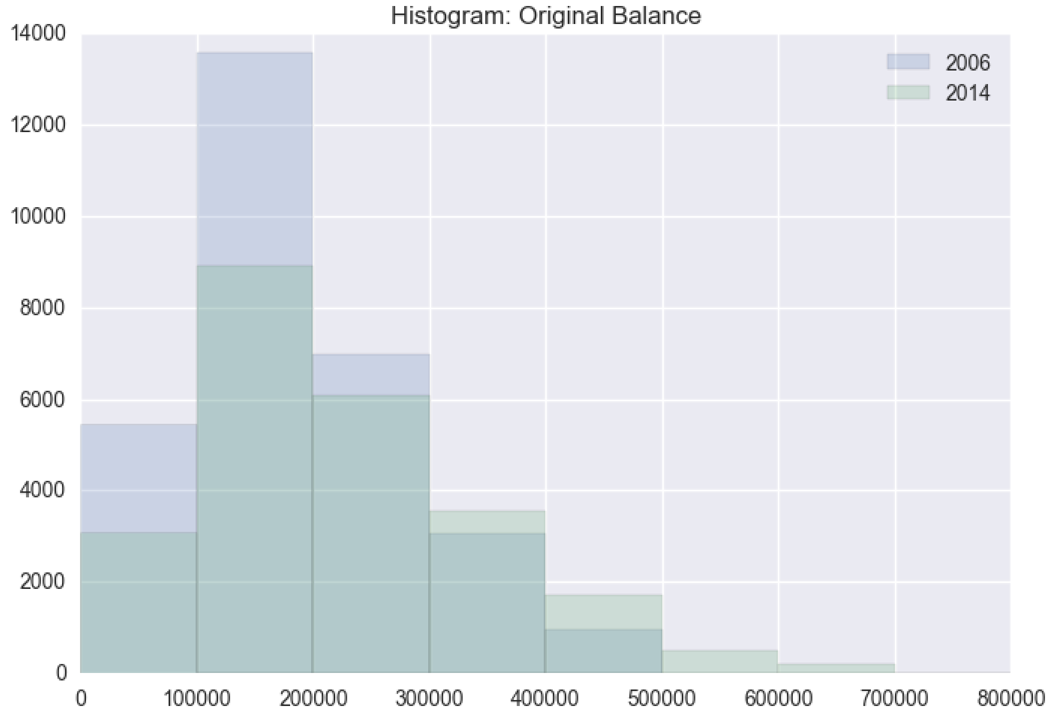
3. Check data distribution:

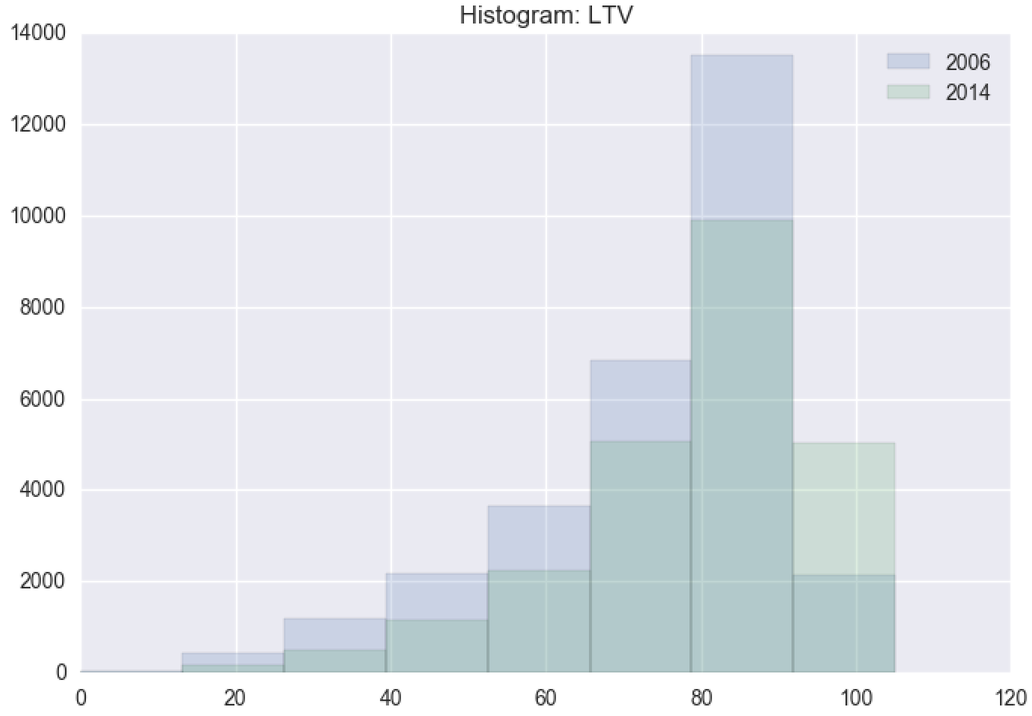
   - DTI appears normally distributed
   - 2006 shows generally higher DTIs – *as expected*



EDA (cont'd): distribution can be skewed

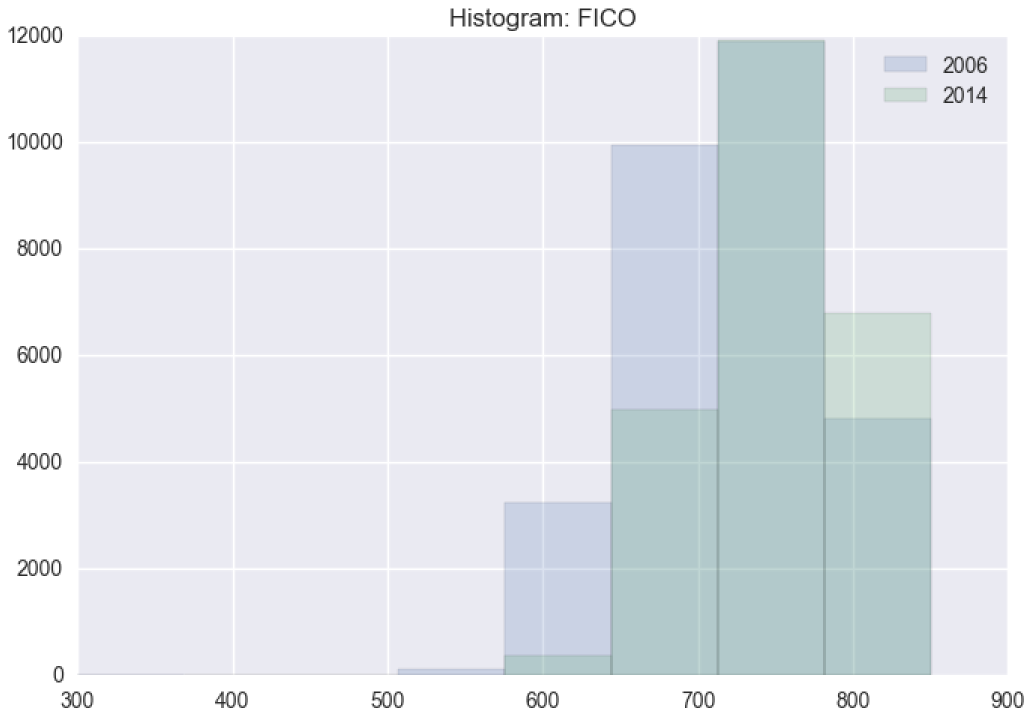
- Original Balance, LTV both slightly skewed

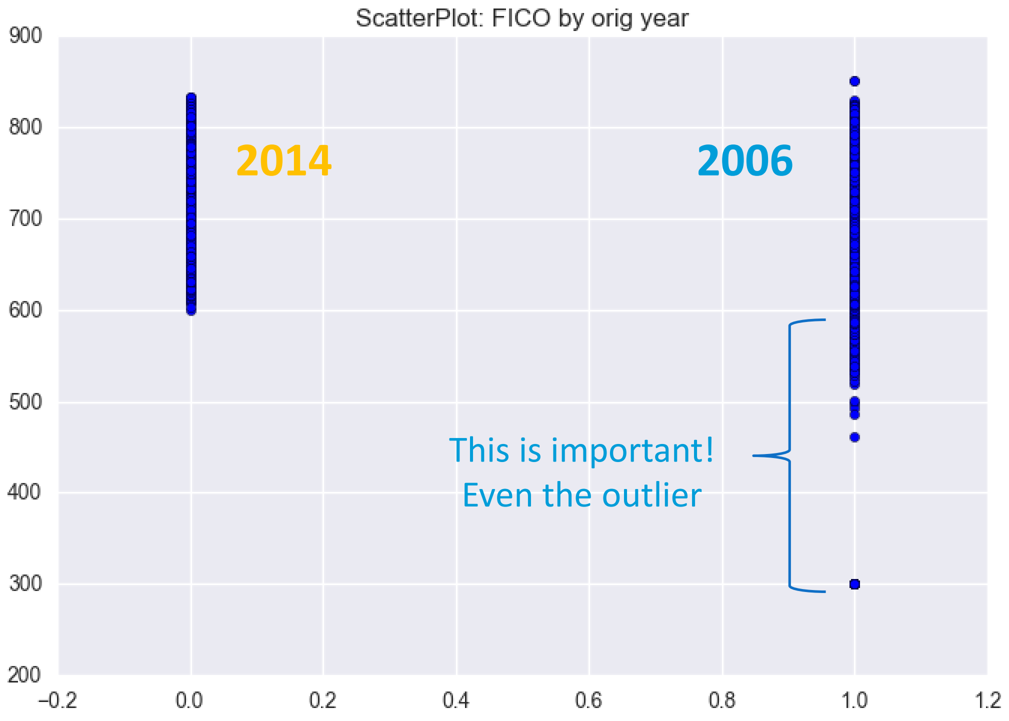




EDA (cont'd): some features heavily skewed

- FICO heavily skewed with outliers





Analysis Step 3: k-means clustering

Our EDA generally agrees with our hypothesis: pre-crisis loans reflect “weaker credit”

Will a 2-cluster analysis produce similar results?

In [33]:
```
orig_subset.groupby(['orig_precrisis']).median().round()
# from this we see 2006 had lower FICOs, loan balances, retail %, and purchase loans vs 2014 on average.
# 2006 also had higher interest rates and higher DTI ratios.
```

Out[33]:

	fico	owner_occ	dti	orig_bal	ltv	retail	purchase	num_borrowers
orig_precrisis								
0	756.0	1	35.0	200000	80.0	1	1	2.0
1	722.0	1	37.0	164000	79.0	0	0	2.0

In [34]:
```
# we sanity check the medians.
orig_subset.groupby(['orig_precrisis']).mean().round()
```

Out[34]:

	fico	owner_occ	dti	orig_bal	ltv	retail	purchase	num_borrowers
orig_precrisis								
0	749.0	1.0	34.0	223751.0	78.0	1.0	1.0	2.0
1	719.0	1.0	37.0	182802.0	72.0	0.0	0.0	2.0

K-means: two clusters not sufficient

Unfortunately no.

In [45]:
```
# inertia score is poor above and we see that its not clustering by orig year correctly.
# instead clustering by purchase vs refi
km = KMeans(n_clusters=2, n_init=20, random_state=1)
km.fit(X_scale)
columns = {str(x): scale.inverse_transform(km.cluster_centers_[x]) for x in range(0,len(km.cluster_centers_))}
pd.DataFrame(columns, index=example.columns)
```

Out[45]:

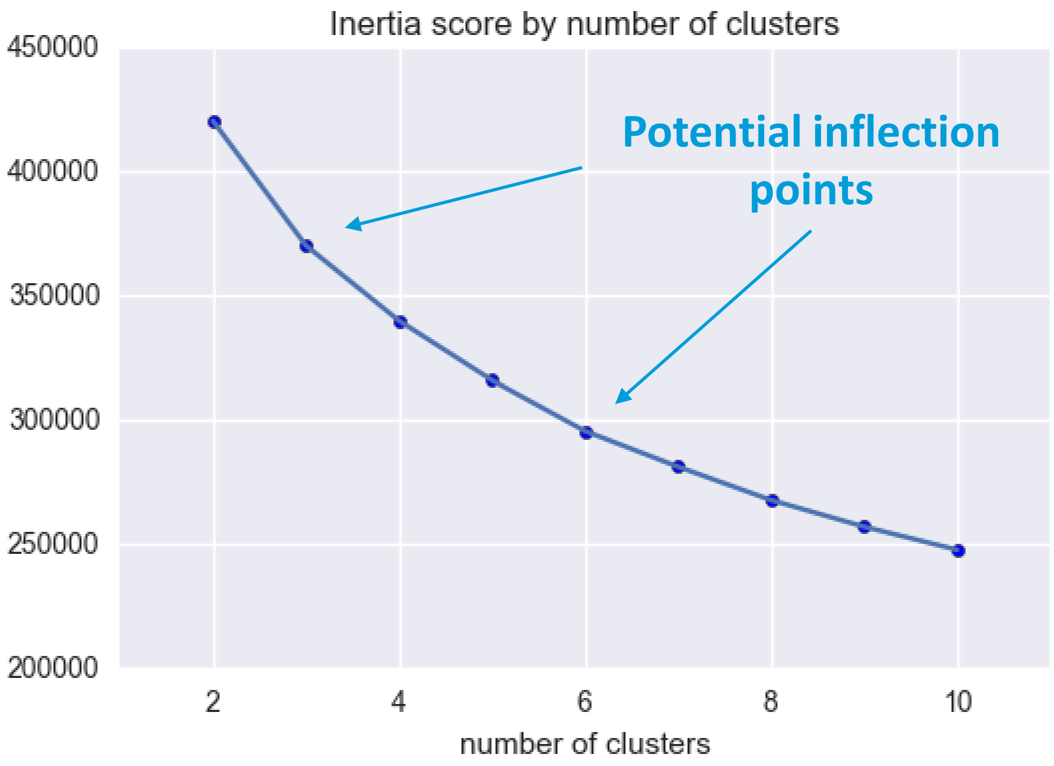
	0	1
fico	720.508967	742.864387
owner_occ	0.931615	0.891610
dti	36.276777	35.377029
orig_bal	200752.770057	201274.170378
ltv	67.512698	81.490052
retail	0.445913	0.454389
purchase	0.027110	0.998743
num_borrowers	1.545787	1.549526
orig_precrisis	0.659597	0.456508

Purchase (0,1) seems to be most discerning feature in a two-cluster analysis.

Both origination years are almost evenly represented in both clusters

K-means (cont'd): less clusters, better performance

Inertia score tells us performance is improved with more clusters anyway



Results using six clusters: much better!

	2014	2006	2006	2014	Both	Both
	0	**1**	**2**	**3**	**4**	**5**
fico	749.62	725.67	689.48	750.66	748.35	755.22
owner_occ	1.00	1.00	1.00	1.00	0.00	1.00
dti	33.26	38.30	39.65	35.80	34.94	30.20
orig_bal	171240.73	180191.55	188935.86	369468.69	159470.53	151946.77
ltv	85.17	80.89	74.16	75.93	71.99	48.37
retail	0.57	0.37	0.31	0.43	0.49	0.62
purchase	0.83	1.00	0.00	0.47	0.63	0.18
num_borrowers	1.41	1.56	1.55	1.74	1.57	1.54
orig_precrisis	0.00	1.00	0.94	0.10	0.49	0.68

Interpreting our six clusters: pre-crisis

	0	1	2	3	4	5
	2014	2006	2006	2014	Both	Both
fico	749.62	725.67	689.48	750.66	748.35	755.22
owner_occ	1.00	1.00	1.00	1.00	0.00	1.00
dti	33.26	38.30	39.65	35.80	34.94	30.20
orig_bal	171240.73	180191.55	188935.86	369468.69	159470.53	151946.77
ltv	85.17	80.89	74.16	75.93	71.99	48.37
retail	0.57	0.37	0.31	0.43	0.49	0.62
purchase	0.83	1.00	0.00	0.47	0.63	0.18
num_borrowers	1.41	1.56	1.55	1.74	1.57	1.54
orig_precrisis	0.00	1.00	0.94	0.10	0.49	0.68

2006 clusters (1, 2):

1 - leveraged buyers: high DTI ratios, exclusively purchase, put down the minimum acceptable down payment (20%)

2 - the struggling house-as-ATMers : worst credit scores and highest debt ratios, and 100% refinance loans; these borrowers likely refinanced to free up some cash to pay off their debts

Interpreting our six clusters: post-crisis

	2014	2006	2006	2014	Both	Both
	0	**1**	**2**	**3**	**4**	**5**
fico	749.62	725.67	689.48	750.66	748.35	755.22
owner_occ	1.00	1.00	1.00	1.00	0.00	1.00
dti	33.26	38.30	39.65	35.80	34.94	30.20
orig_bal	171240.73	180191.55	188935.86	369468.69	159470.53	151946.77
ltv	85.17	80.89	74.16	75.93	71.99	48.37
retail	0.57	0.37	0.31	0.43	0.49	0.62
purchase	0.83	1.00	0.00	0.47	0.63	0.18
num_borrowers	1.41	1.56	1.55	1.74	1.57	1.54
orig_precrisis	0.00	1.00	0.94	0.10	0.49	0.68

2014 clusters (0, 3):

0 - top notch starter homebuyers: very high credit score, carrying low level of debt, but loan size and down payment small

3 - the elite "mcmansions" : highest credit, largest loan sizes, put down 25% down payment

Interpreting our six clusters: found in both

	2014	2006	2006	2014	Both	Both
	0	**1**	**2**	**3**	**4**	**5**
fico	749.62	725.67	689.48	750.66	748.35	755.22
owner_occ	1.00	1.00	1.00	1.00	0.00	1.00
dti	33.26	38.30	39.65	35.80	34.94	30.20
orig_bal	171240.73	180191.55	188935.86	369468.69	159470.53	151946.77
ltv	85.17	80.89	74.16	75.93	71.99	48.37
retail	0.57	0.37	0.31	0.43	0.49	0.62
purchase	0.83	1.00	0.00	0.47	0.63	0.18
num_borrowers	1.41	1.56	1.55	1.74	1.57	1.54
orig_precrisis	0.00	1.00	0.94	0.10	0.49	0.68

clusters apparent pre and post crisis (4, 5):

4 - the sophisticated investor : exclusively small investment properties, borrower has good credit

5 - the responsible refinancer : over 50% equity in their home! high credit score, lowest DTI ratio, high refi %

Pros and cons of our six cluster analysis:

What we like about this analysis:

1. outputs come in generally as expected matching mean, median and intuition
2. almost all features are “contributing their fair share”; meaning for any one feature it does not have the same results across all columns

What we dislike:

1. Six clusters is a lot to explain

Results using three clusters: still very good!

2014 2006 Both

	0	1	2
fico	748.99	716.08	748.38
owner_occ	1.00	1.00	0.00
dti	33.98	37.38	34.95
orig_bal	231588.67	184156.90	162141.05
ltv	78.89	71.95	71.95
retail	0.54	0.38	0.48
purchase	0.64	0.42	0.63
num_borrowers	1.52	1.57	1.57
orig_precrisis	0.01	0.99	0.49

2014:

0 - top notch homeowners: very high credit score, carrying low level of debt, mostly buying homes but some refinancing

2006:

1 – debt-heavy, slightly blemished homeowner: materially worse credit scores, high debt to income ratios, the majority of which have been solicited perhaps to cash out (low retail %, lower purchase %)

Both:

2 - the sophisticated investor: exclusively small investment properties, borrower has very strong credit

Last step: eliminate insignificant feature

	0	1	2
fico	748.99	716.08	748.38
owner_occ	1.00	1.00	0.00
dti	33.98	37.38	34.95
orig_bal	231588.67	184156.90	162141.05
ltv	78.89	71.95	71.95
retail	0.54	0.38	0.48
purchase	0.64	0.42	0.63
~~**num_borrowers**~~	~~1.52~~	~~1.57~~	~~1.57~~
orig_precrisis	0.01	0.99	0.49

8 features: better performance, same clusters

Both 2006 2014

	0	1	2
fico	748.38	749.18	715.94
owner_occ	0.00	1.00	1.00
dti	34.95	33.96	37.40
orig_bal	162141.05	232070.85	183789.97
ltv	71.95	78.95	71.91
retail	0.48	0.54	0.37
purchase	0.63	0.64	0.42
orig_precrisis	0.49	0.01	0.99

Business implications

Investment opportunity:

1. Underwriting better now → Investors who were scarred by mortgage losses in 2008 should consider revisiting!

Borrower solicitation:

1. The recent borrowers (2014) were strong credit homebuyers; anyone with access to a borrowers credit history should cross market mortgage loans to their existing borrowers.
   - banks who also issue credit cards (e.g. chase, citi)
   - student loan companies may also want to issue mortgages (e.g. SoFi)
2. Keep marketing to investors in lower median home price areas

Caveats to our analysis and next steps:

1. We eliminated some features that would have resulted in a LOT of dummy variables (e.g. state, servicer).
   - Next steps: think about reintroducing these. Geo in particular could be useful for marketing.
2. We used single years 2006 and 2014 as proxies for pre- and post crisis origination
   - Next steps: test other proxies for pre and post to see if we get the same results
3. We use 2014 to represent current dynamics since it was the last data set presented to us that was fully populated (50,000 loans).
   - Next steps: use 2015 data when loans reach 50,000. Keep monitoring for more updates.