Read Files

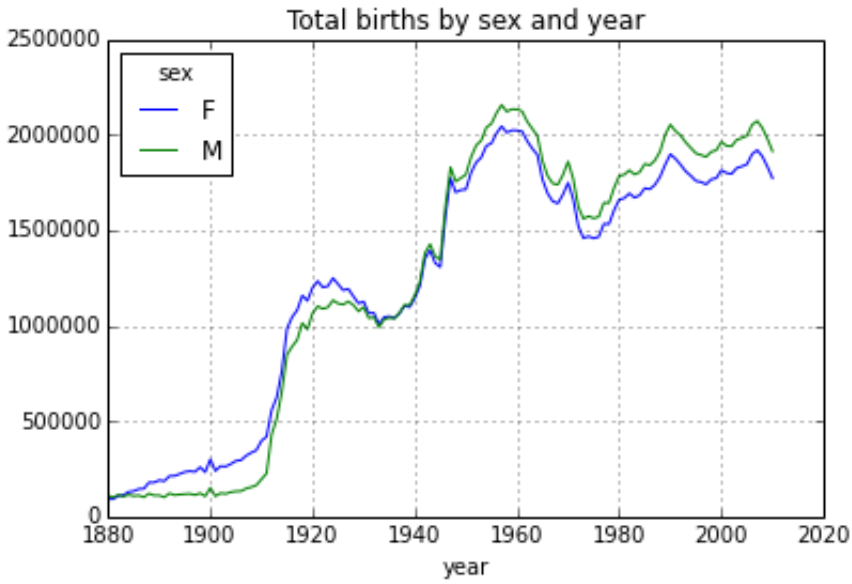
In [121]:
```
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
os.chdir('C:/Users/Giuseppe/Dropbox (Personal)/babynames')

years = range(1880, 2011)
pieces = []
columns = ['name', 'sex', 'births']
for year in years:
    path = 'names/yob%d.txt' % year
    #print(path)
    frame = pd.read_csv(path, names=columns)
    frame['year'] = year
    pieces.append(frame)
names = pd.concat(pieces, ignore_index=True)
```

Plots births by sex and year

In [122]:
```
total_births = names.pivot_table('births', rows='year', cols='sex', aggfunc=sum)
total_births.plot(title='Total births by sex and year')
```

Out[122]: <matplotlib.axes.AxesSubplot at 0x15439a58>



Fraction of total names, within sex and year

In [123]:
```
def add_prop(group):
    # Integer division floors
    births = group.births.astype(float)
    group['prop'] = births / births.sum()
    return group

names = names.groupby(['year', 'sex']).apply(add_prop)
names.head()
```

Out[123]:

	name	sex	births	year	prop
0	Mary	F	7065	1880	0.077643
1	Anna	F	2604	1880	0.028618
2	Emma	F	2003	1880	0.022013
3	Elizabeth	F	1939	1880	0.021309
4	Minnie	F	1746	1880	0.019188

5 rows × 5 columns

Top 1000 names by year, sex

In [124]:
```
def get_top1000(group):
    return group.sort_index(by='births', ascending=False)[:1000]
grouped = names.groupby(['year', 'sex'])
top1000 = grouped.apply(get_top1000)
top1000.index = np.arange(len(top1000))
top1000.head()
```

Out[124]:

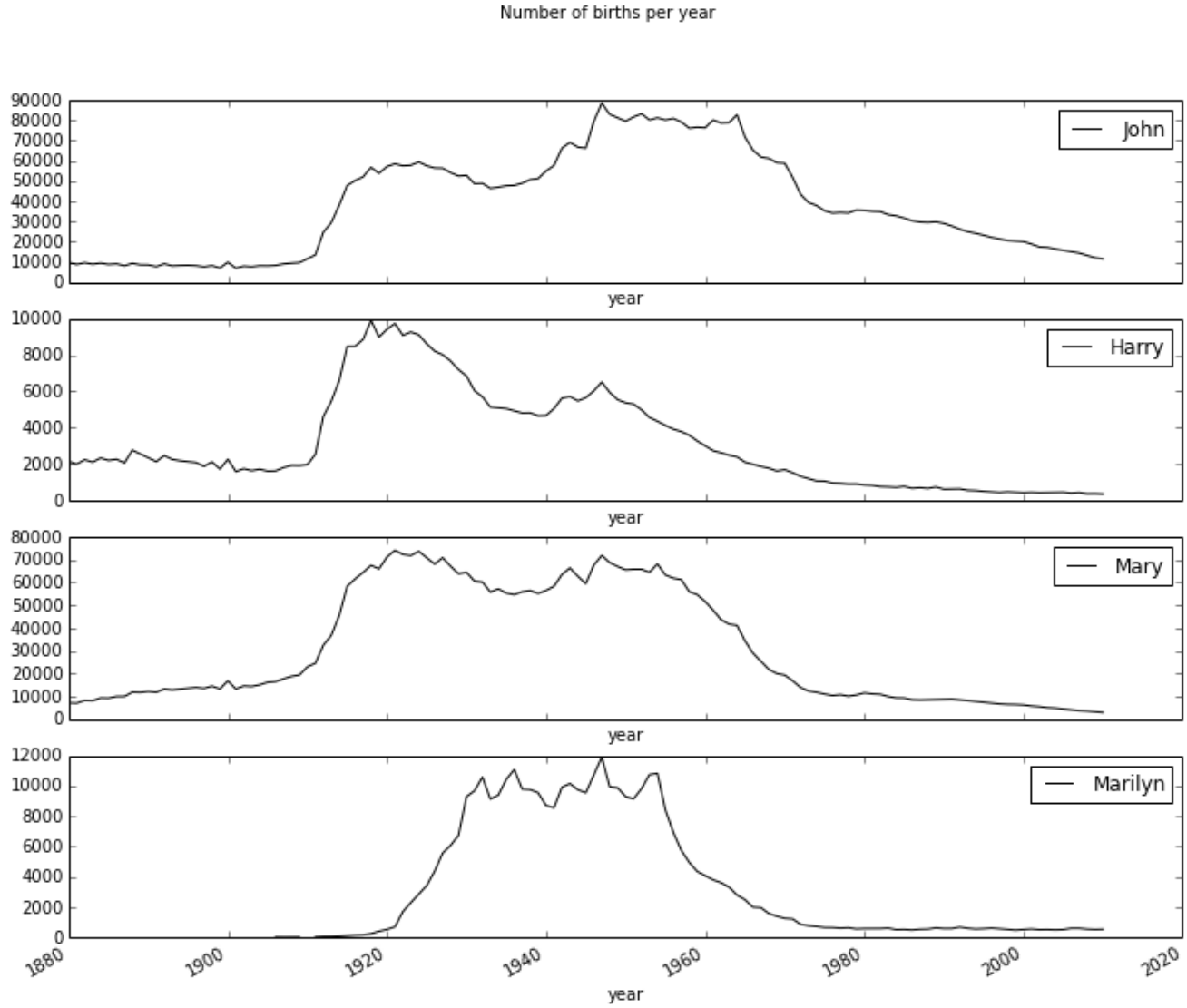
	name	sex	births	year	prop
0	Mary	F	7065	1880	0.077643
1	Anna	F	2604	1880	0.028618
2	Emma	F	2003	1880	0.022013
3	Elizabeth	F	1939	1880	0.021309
4	Minnie	F	1746	1880	0.019188

5 rows × 5 columns

Plots the number of babies named John, Harry, Mary, Marilyn over time

In [106]:
```
total_births = top1000.pivot_table('births', rows='year', cols='name', aggfunc=sum)
subset = total_births[['John', 'Harry', 'Mary', 'Marilyn']]
subset.plot(subplots=True, figsize=(12, 10), grid=False, title="Number of births per y
ear")
```

Out[106]:
```
array([<matplotlib.axes.AxesSubplot object at 0x000000005CAF46D8>,
       <matplotlib.axes.AxesSubplot object at 0x000000006DDF6240>,
       <matplotlib.axes.AxesSubplot object at 0x00000000641E8668>,
       <matplotlib.axes.AxesSubplot object at 0x000000006E17C0B8>], dtype=object)
```

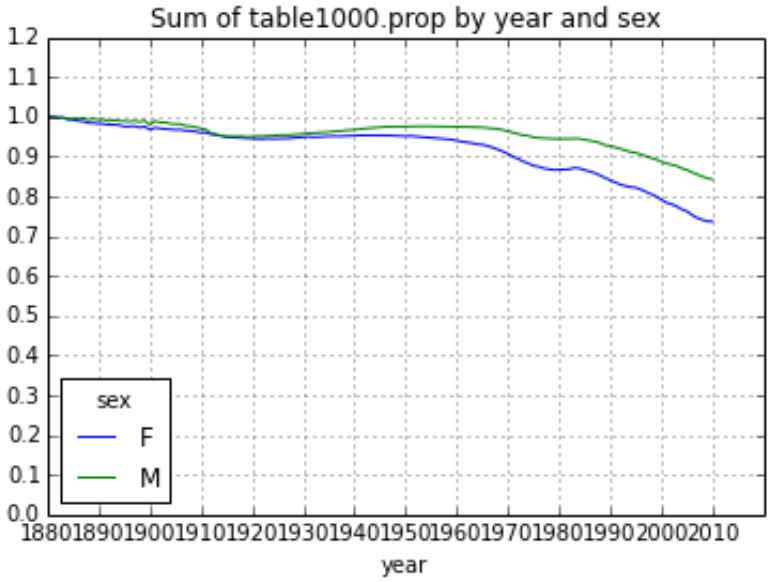


Plots the proportion of the top 1000 names as a percentage of total

In [125]:
```
table = top1000.pivot_table('prop', rows='year', cols='sex', aggfunc=sum)
```

```
table.plot(title='Sum of table1000.prop by year and sex', yticks=np.linspace(0, 1.2, 13), xticks=range(1880, 2020, 10))
```

Out[125]: <matplotlib.axes.AxesSubplot at 0x65230278>



How many boy names comprise 50% of the total in 2010?

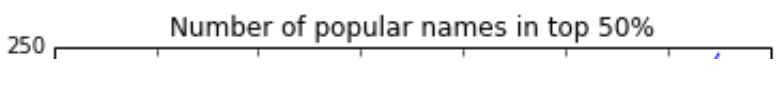
In [126]:
```
boys = top1000[top1000.sex == 'M']
df = boys[boys.year == 2010]
prop_cumsum = df.sort_index(by='prop', ascending=False).prop.cumsum()
prop_cumsum.values.searchsorted(0.5)
```

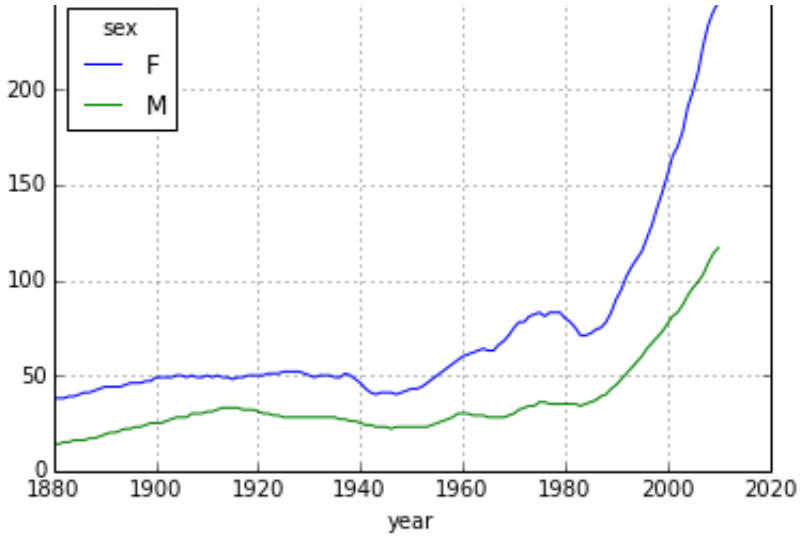
Out[126]: 116

Plots number of most popular names used by 50% of boys and girls over time

In [127]:
```
def get_quantile_count(group, q=0.5):
        group = group.sort_index(by='prop', ascending=False)
        return group.prop.cumsum().values.searchsorted(q) + 1
diversity = top1000.groupby(['year', 'sex']).apply(get_quantile_count)
diversity = diversity.unstack('sex')
diversity.plot(title="Number of popular names in top 50%")
```

Out[127]: <matplotlib.axes.AxesSubplot at 0x6caa4198>





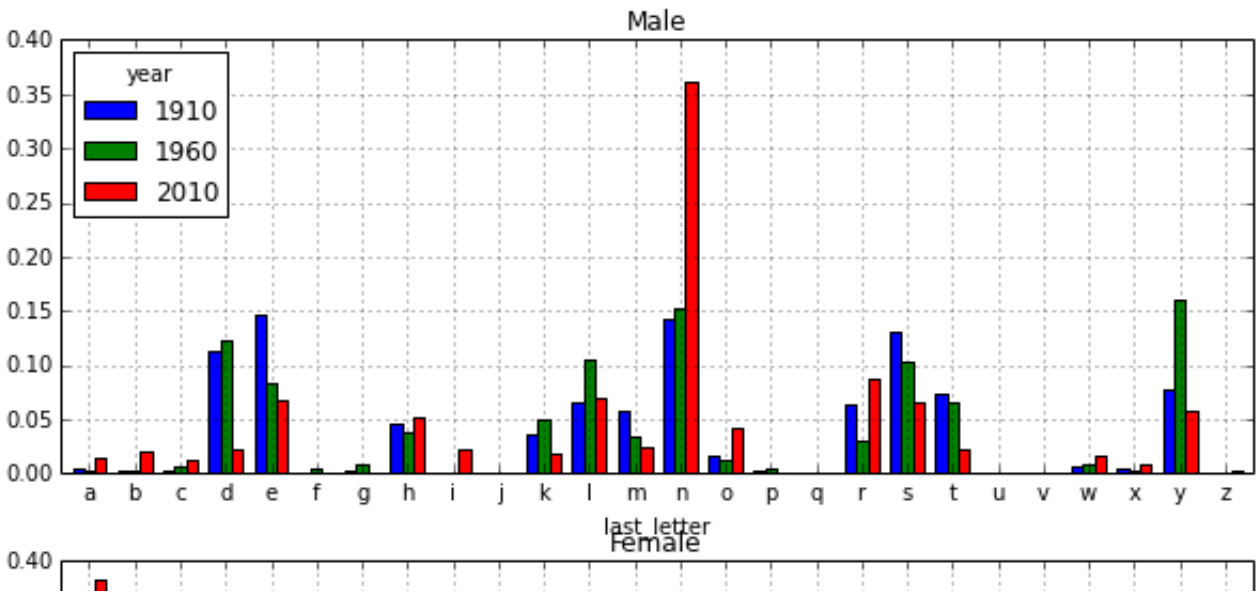
Plot the distribution of names by last letter for three time snapshots

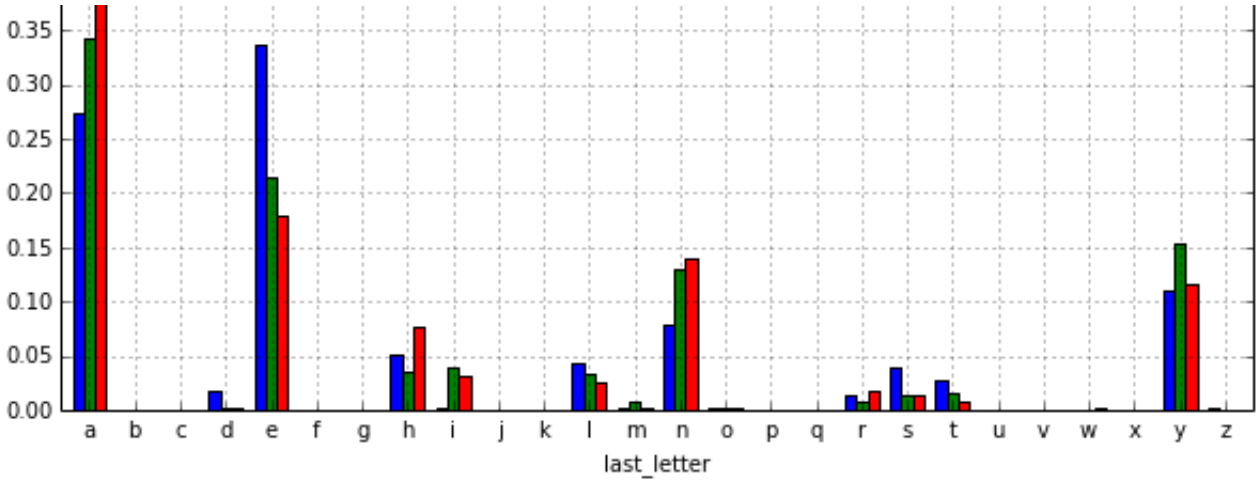
In [138]:
```
# extract last letter from name column
get_last_letter = lambda x: x[-1]
last_letters = names.name.map(get_last_letter)
last_letters.name = 'last_letter'
table = names.pivot_table('births', rows=last_letters, cols=['sex', 'year'], aggfunc=s
um)

subtable = table.reindex(columns=[1910, 1960, 2010], level='year')
letter_prop = subtable / subtable.sum().astype(float)
```

In [98]:
```
fig, axes = plt.subplots(2, 1, figsize=(10, 8))
letter_prop['M'].plot(kind='bar', rot=0, ax=axes[0], title='Male', legend=True)
letter_prop['F'].plot(kind='bar', rot=0, ax=axes[1], title='Female', legend=False)
```

Out[98]:
```
<matplotlib.axes.AxesSubplot at 0x5bd93a58>
```



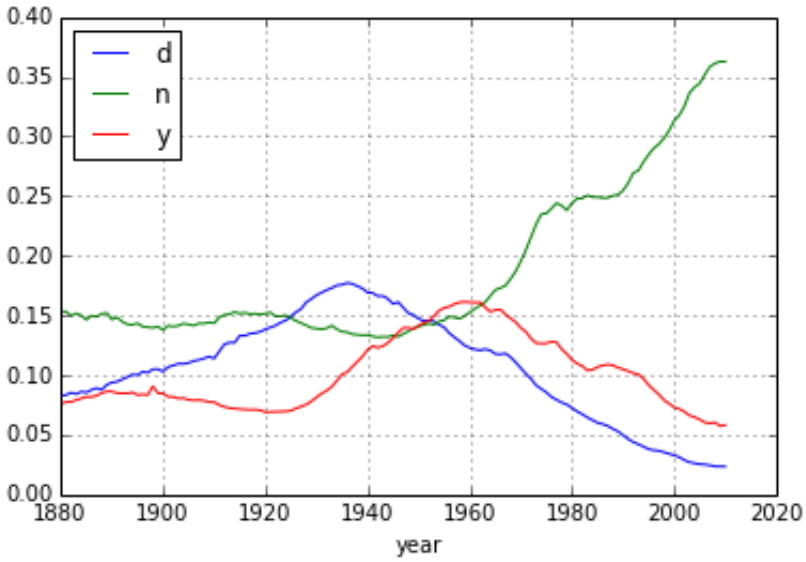


Plots the proportion of boy names ending in 'd', 'n', and 'y' over time

In [137]:
```
letter_prop = table / table.sum().astype(float)
dny_ts = letter_prop.ix[['d', 'n', 'y'], 'M'].T

dny_ts.plot()
# table.ix(last_letter=='d')
```

Out[137]: <matplotlib.axes.AxesSubplot at 0x5cb0af28>



In []: