This is an ipynb for exploratory analysis

It is very much a work in progress, so please forgive anything from typos to absence of graph labels.

Because there was too much noise in the full respondent data (many of them only had location services turned on for a few of their total semesters on campus, for example), I've decided to focus my analysis on the semesters themselves. This way, if a respondent spent two semesters working and two semesters not working, we have that exact relationship between each unit of time and work status (as opposed to a proportion which left a fuzzier picture).

The code I used to munge the data is still in the process of being consolidated into ipython notebooks, but I will use them almost exclusively for analysis moving forward. The exception will be for maps, for which I will likely use GRASS GIS or QGIS to generate and embed here.

Note on the semester data: I semi-arbitrarily removed all semesters with fewer than 20,000 points from the sample. I based this decision off of the fact that 20,000 points over the course of a semester is around 6 readings per hour, which for now seems like an acceptable lower limit.

In [1]:
```
# import libraries
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from matplotlib import rcParams
from scipy import stats
import json

# set the style
s = json.load( open("./style/538.json") )
rcParams.update(s)

#display inline
%matplotlib inline
```

In [2]:
```
# read data into a DataFrame
path = '/Users/fabio/code/thesis/stats/clean_semesters.csv'
data = pd.read_csv(path, index_col=0)
```

Descriptive Statistics

Basic descriptive statistics for the semester data I am working with.

In [3]:
```
data.describe().transpose()
```

Out[3]:

	count	mean	std	min	2
parent_edu_code	74	4.540541	0.725073	2.000000	4
parent_income_code	74	5.864865	2.411867	1.000000	4
grad_year	74	2016.851351	0.946261	2016.000000	2
worked	74	0.540541	0.501756	0.000000	0
assisted	74	0.783784	0.414473	0.000000	1
mhd_avg	74	7478.736038	2357.834886	1652.245771	6
mhd_sum	74	31495911.250568	31804491.492513	203906.500200	1
mhd_count	74	4998.405405	6752.440549	26.000000	1
total_points	74	71264.500000	35017.911189	21627.000000	4
percent_off_campus	74	0.077163	0.122775	0.001001	0
block_count	74	346.378378	171.461702	23.000000	1

Worked, Assisted

On the survey, for each semester, respondents were asked if they:

1. Worked for a wage during that semester
2. Received financial assistance from their family during that semester

Instead of relying on parent income, which does not necessarily translate directly into support or mean the same thing from region to region, these measures provide a basic yes-or-no measure of support from family and self-support via paid employment.

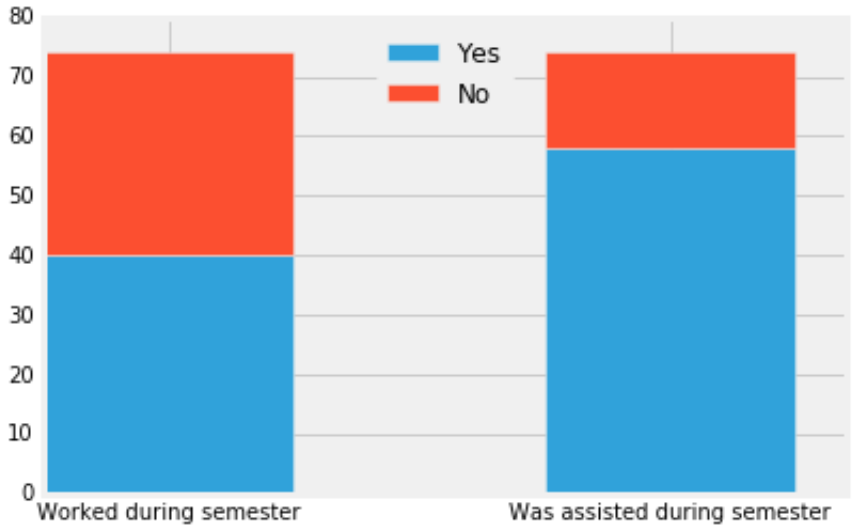
In [4]:
```
# stacked bar of whether or not respondent worked/was assisted duri
ng semester

# whether or not worked
worked = 0
not_worked = 0
for item in data['worked']:
 if item == 1:
    worked+=1
 else:
    not_worked+=1

# whether or not was assisted
assisted = 0
not_assisted = 0
for item in data['assisted']:
    if item == 1:
        assisted+=1
    else:
        not_assisted+=1

# make the arrays to plot stacked bars
yes = [worked, assisted]
no = [not_worked, not_assisted]
X = np.arange(2)
width = 0.5

plt.figure()
plt.bar(X, yes, width, color='#30a2da', label='Yes')
plt.bar(X, no, width, bottom = yes, color='#fc4f30',label='No')
plt.legend(loc='upper center')
plt.xticks(X + width/2., ('Worked during semester', 'Was assisted d
uring semester'))
plt.show()
```



This graph shows us two things:

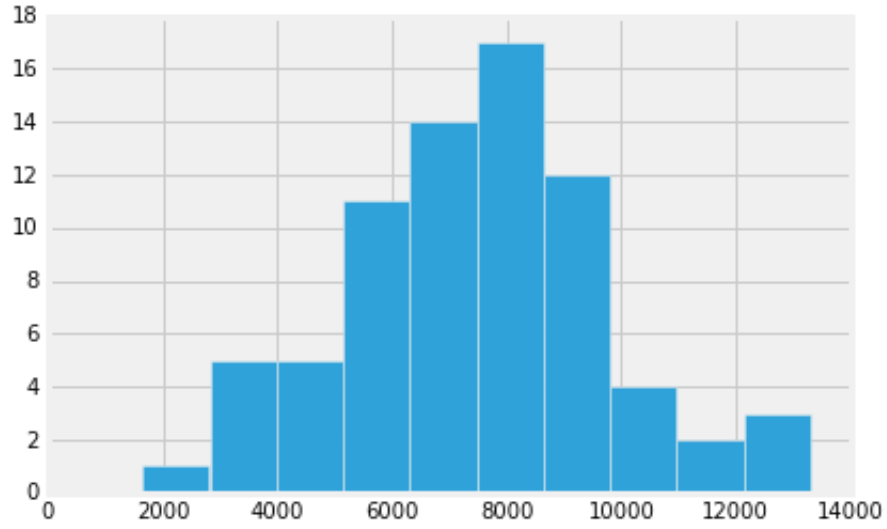
1. Respondents worked during a slight majority of semesters
2. Respondents were assisted during a large majority of semesters.

Average Manhattan Distance

This metric, which takes the average manhattan distance (https://xlinux.nist.gov/dads//HTML/manhattanDistance.html) of points collected off campus, gives us a rough approximation of how far off-campus (on average) a respondent went during a given semester. While flawed, this lets us look for any obvious differences.

```
In [5]: # histogram of average manhattan distance when off campus
        plt.hist(data['mhd_avg'])
```

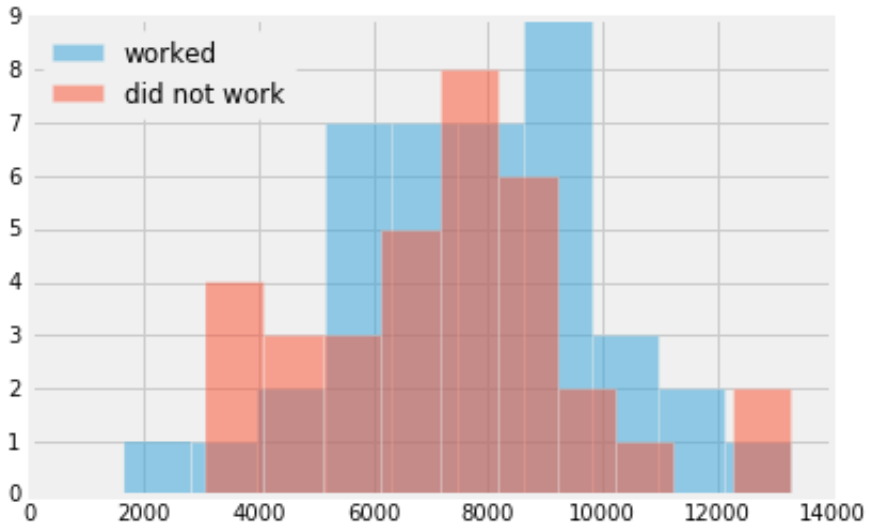
```
Out[5]: (array([  1.,   5.,   5.,  11.,  14.,  17.,  12.,   4.,   2.,
        3.]),
         array([  1652.245771 ,   2816.5010959,   3980.7564208,   5145.011
        7457,
                  6309.2670706,   7473.5223955,   8637.7777204,   9802.033
        0453,
                 10966.2883702,  12130.5436951,  13294.79902  ]),
         <a list of 10 Patch objects>)
```



In the above figure, a histogram of average manhattan distance off campus (in meters) for all semesters, we can see that there is a roughly normal distribution.

We can split them up by whether or not the respondent worked:

In [6]:
```
# overlaid histograms of average manhattan distance when off campu
s, divided by whether or not the respondent worked
worked = data.loc[data['worked']==1]
not_worked = data.loc[data['worked']==0]
worked_list = [worked['mhd_avg'], not_worked['mhd_avg']]
plt.figure()
plt.hist(worked_list[0], alpha=0.5, label = 'worked')
plt.hist(worked_list[1], alpha=0.5, label = 'did not work')
plt.legend(loc='upper left')
plt.show()
```



In the above figure, we can see that the histogram of average manhattan distance for semesters where the respondents worked seems to suggest that working for a wage could lead to being further away from campus than not working. However, a t-test (below) shows us that the means are not significantly different.

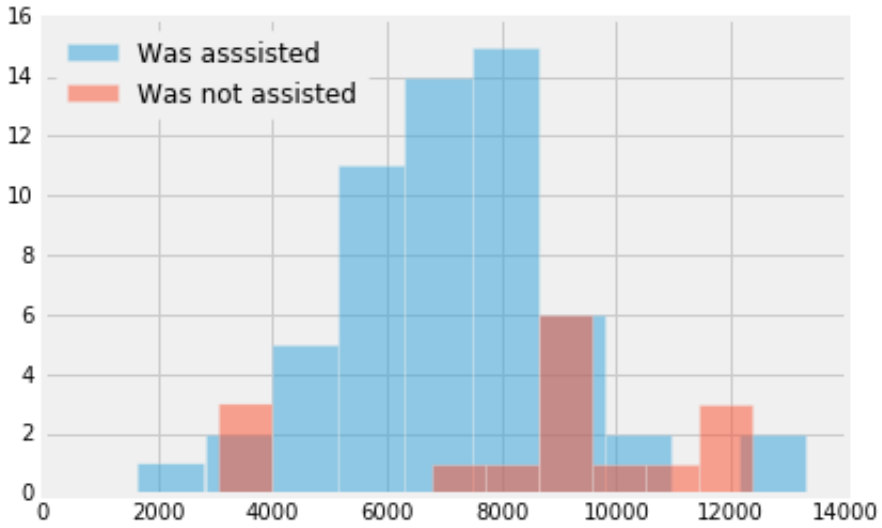
In [7]:
```
# t-test of worked/not-worked and percent off campus
print stats.ttest_ind(worked['mhd_avg'],not_worked['mhd_avg'], equa
l_var=False)

print 'Mean mhd_avg for semesters where respondent worked: ', worke
d['mhd_avg'].mean()
print 'Mean mhd_avg for semesters where respondent did not work: ',
not_worked['mhd_avg'].mean()
```

```
Ttest_indResult(statistic=0.97553219049736528, pvalue=0.3326876445
5806055)
Mean mhd_avg for semesters where respondent worked:  7725.96026192
Mean mhd_avg for semesters where respondent did not work:  7187.88
400994
```

We can also examine the difference in average manhattan distance between semesters where students were financially assisted or not.

In [8]:
```
# overlaid histograms of average manhattan distance when off campu
s, divided by whether or not the respondent was assisted
assisted = data.loc[data['assisted']==1]
not_assisted = data.loc[data['assisted']==0]
assisted_list = [assisted['mhd_avg'], not_assisted['mhd_avg']]
plt.figure()
plt.hist(assisted_list[0], alpha=0.5, label = 'Was asssisted')
plt.hist(assisted_list[1], alpha=0.5, label = 'Was not assisted')
plt.legend(loc='upper left')
plt.show()
```



While the difference in means is larger between assisted and not assisted semesters, it is still not significant for this sample (although a larger sample might be worth pursuing).

In [9]:
```
# t-test of assisted/not-assisted and percent off campus
print stats.ttest_ind(assisted['mhd_avg'],not_assisted['mhd_avg'],
equal_var=False)

print 'Mean mhd_avg for semesters where respondent was assisted: ',
assisted['mhd_avg'].mean()
print 'Mean mhd_avg for semesters where respondent was not assiste
d: ', not_assisted['mhd_avg'].mean()
```

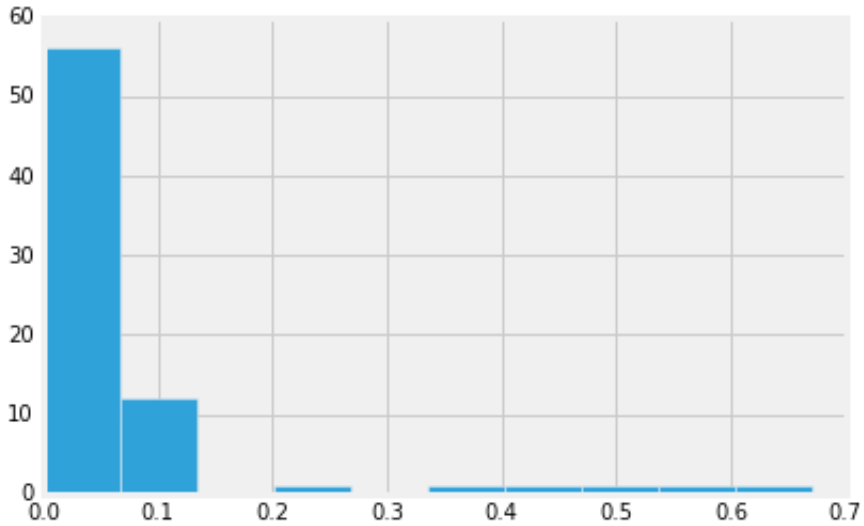
```
Ttest_indResult(statistic=-1.6210387360807037, pvalue=0.1211020387
9451864)
Mean mhd_avg for semesters where respondent was assisted:  7202.66
113771
Mean mhd_avg for semesters where respondent was not assisted:  847
9.50755175
```

Percentage off campus

This measure takes all of the points for a respondent's semester, and calculates the proportion of those points that were collected outside of the campus (110th Street - 123rd Street, Riverside Drive - Morningside Drive).

In [10]:
```
# histogram of percentage off campus
plt.hist(data['percent_off_campus'])
```

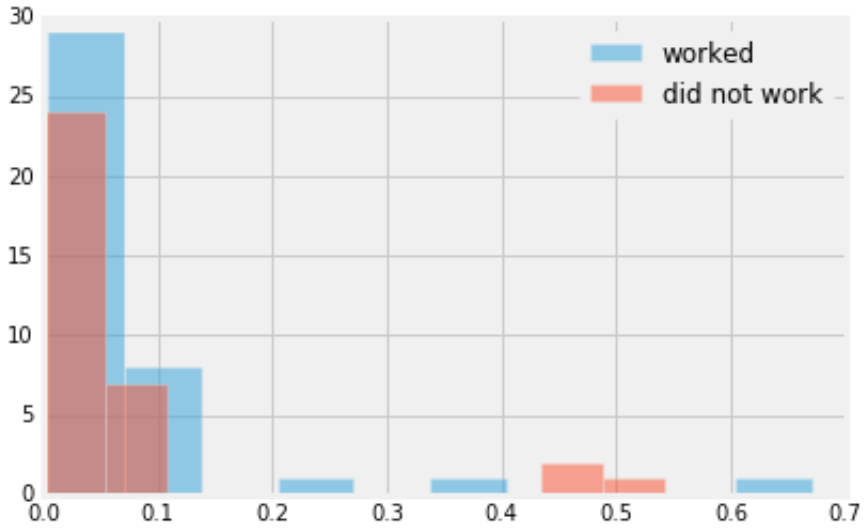
Out[10]:
```
(array([ 56.,  12.,   0.,   1.,   0.,   1.,   1.,   1.,   1.,
1.]),
 array([ 0.00100104,  0.06795589,  0.13491073,  0.20186558,  0.268
82043,
         0.33577527,  0.40273012,  0.46968497,  0.53663981,  0.603
59466,
         0.67054951]),
 <a list of 10 Patch objects>)
```



In the above figure, we can see that the vast majority of students spend less than 10 percent of their time off campus. The few outliers are worth looking into.

We can split them up by whether or not the respondent worked:

In [11]:
```
# overlaid histograms of percentage off campus, divided by whether
or not the respondent worked
worked = data.loc[data['worked']==1]
not_worked = data.loc[data['worked']==0]
worked_list = [worked['percent_off_campus'], not_worked['percent_of
f_campus']]
plt.figure()
plt.hist(worked_list[0], alpha=0.5, label = 'worked')
plt.hist(worked_list[1], alpha=0.5, label = 'did not work')
plt.legend(loc='upper right')
plt.show()
```



We can see a very similar distribution here, with both groups averaging only about 7.7% of their time spent off campus.

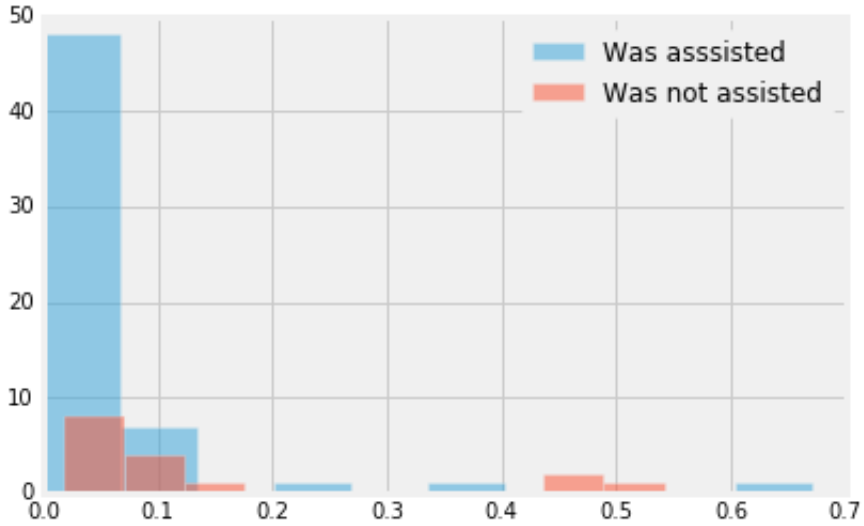
In [12]:
```
# t-test of worked/not-worked and percent off campus
print stats.ttest_ind(worked['percent_off_campus'],not_worked['perc
ent_off_campus'], equal_var=False)

print 'Mean % off campus for semesters where respondent worked: ',
worked['percent_off_campus'].mean()
print 'Mean % off campus for semesters where respondent did not wor
k: ', not_worked['percent_off_campus'].mean()
```

```
Ttest_indResult(statistic=0.020962108115270495, pvalue=0.983338753
84591207)
Mean % off campus for semesters where respondent worked:  0.077443
658825
Mean % off campus for semesters where respondent did not work:  0.
0768328146471
```

We can also examine the difference in the percentage of time spent off campus between semesters where students were financially assisted or not.

In [13]:
```
# overlaid histograms of percent off campus, divided by whether or
not the respondent was assisted
assisted = data.loc[data['assisted']==1]
not_assisted = data.loc[data['assisted']==0]
assisted_list = [assisted['percent_off_campus'], not_assisted['perc
ent_off_campus']]
plt.figure()
plt.hist(assisted_list[0], alpha=0.5, label = 'Was asssisted')
plt.hist(assisted_list[1], alpha=0.5, label = 'Was not assisted')
plt.legend(loc='upper right')
plt.show()
```



While the distributions here seem to remain similar, we see that the means are different at a p=0.1 significance level. The direction of the difference in this sample suggests that students spend more time off campus during semesters where they work for a wage. While the sample of unassisted semesters is too small at this point to begin confirming this, we can imagine that students with no financial assistance from their families are more likely to spend more time working; if their job was off campus, then it makes sense that they would spend less time on campus than their assisted peers.

In [19]:
```
# t-test of assisted/not-assisted and percent off campus
print stats.ttest_ind(assisted['percent_off_campus'],not_assisted
['percent_off_campus'], equal_var=False)

print 'Mean % off campus for semesters where respondent was assiste
d: ', assisted['percent_off_campus'].mean()
print 'Mean % off campus for semesters where respondent was not ass
isted: ', not_assisted['percent_off_campus'].mean()
```

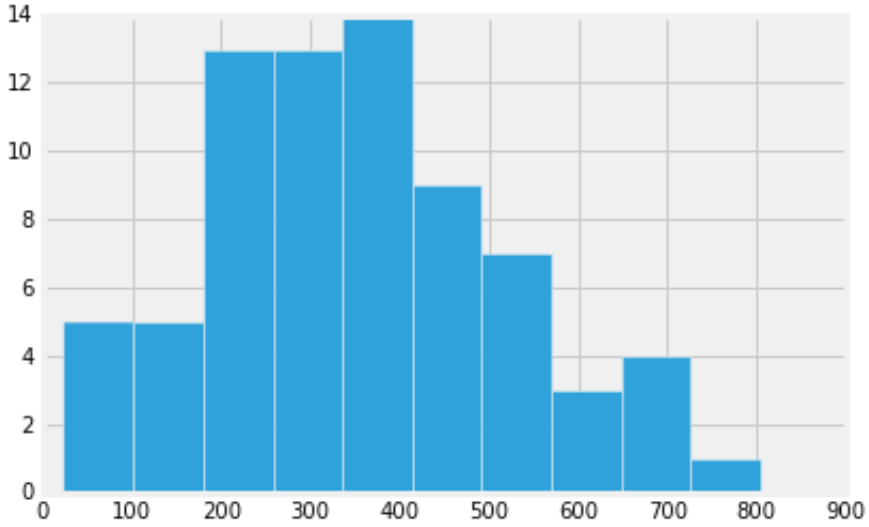
```
Ttest_indResult(statistic=-1.7256178756689975, pvalue=0.1018371884
0829355)
Mean % off campus for semesters where respondent was assisted:  0.
0600519388966
Mean % off campus for semesters where respondent was not assisted:
0.139190599687
```

Census Tracts visited

This measure is simple: how many unique census tracts did a respondent visit during a semester? While it's also a naive measure, it does provide a sense of diversity of place that manhattan distance and percentage off campus does not reveal. While it seems that students spend more of their time off-campus during semesters when they do not receive assistance, it is possible that they spend most of that time at the workplace, such that students who do not work might visit a broader range of places.

In [16]:
```
# histogram of count of unique census tracts visited in a semester
plt.hist(data['block_count'])
```

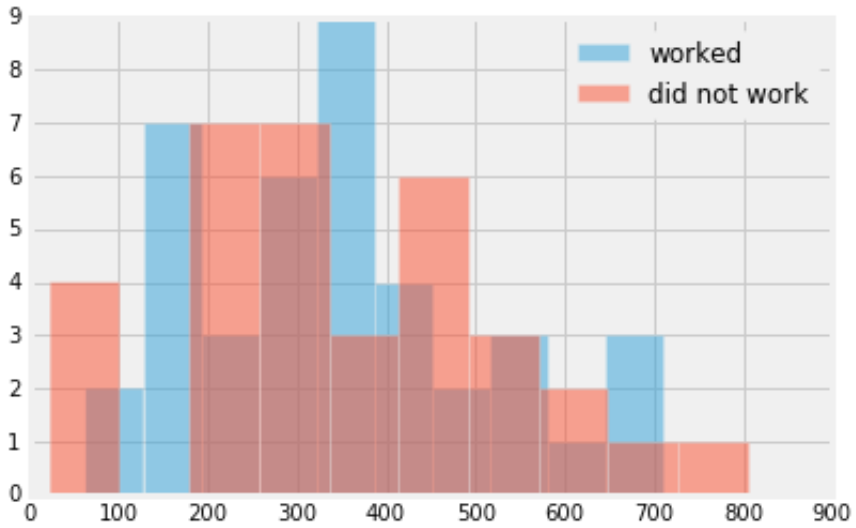
Out[16]:
```
(array([  5.,   5.,  13.,  13.,  14.,   9.,   7.,   3.,   4.,
1.]),
 array([  23. ,  101.2,  179.4,  257.6,  335.8,  414. ,  492.2,  5
70.4,
         648.6,  726.8,  805. ]),
 <a list of 10 Patch objects>)
```



A somewhat right-skewed distribution.

We can split them up by whether or not the respondent worked:

In [26]:
```
# overlaid histograms of count of unique census tracts visited in a
semester, divided by whether or not the respondent worked
worked = data.loc[data['worked']==1]
not_worked = data.loc[data['worked']==0]
worked_list = [worked['block_count'], not_worked['block_count']]
plt.figure()
plt.hist(worked_list[0], alpha=0.5, label = 'worked')
plt.hist(worked_list[1], alpha=0.5, label = 'did not work')
plt.legend(loc='upper right')
plt.show()
```



This doesn't reveal much, and the t-test shows us that there is no meaningful difference between semesters where respondents worked and semesters where they did not.

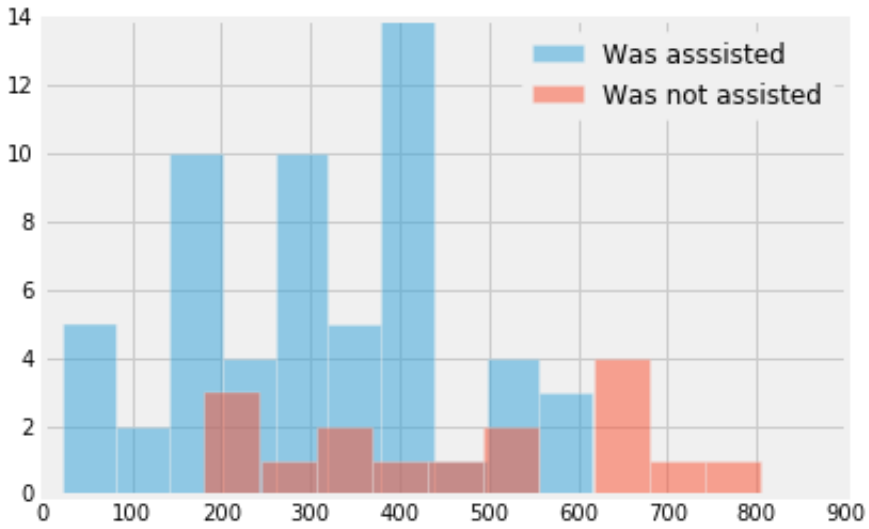
In [17]:
```
# t-test of work/no-work and census tracts count
print stats.ttest_ind(worked['block_count'],not_worked['block_coun
t'], equal_var=False)

print 'Mean # census tracts visited for semesters where respondent
worked: ', worked['block_count'].mean()
print 'Mean # census tracts for semesters where respondent did not
work: ', not_worked['block_count'].mean()
```

```
Ttest_indResult(statistic=0.11877007459409954, pvalue=0.9058172282
6142319)
Mean # census tracts visited for semesters where respondent worke
d:  348.6
Mean # census tracts for semesters where respondent did not work:
343.764705882
```

We can also examine the difference in the number of census tracts visited between semesters where students were financially assisted or not:

In [30]:
```
# overlaid histograms of count of unique census blocks visited, div
ided by whether or not the respondent was assisted
assisted = data.loc[data['assisted']==1]
not_assisted = data.loc[data['assisted']==0]
assisted_list = [assisted['block_count'], not_assisted['block_coun
t']]
plt.figure()
plt.hist(assisted_list[0], alpha=0.5, label = 'Was asssisted')
plt.hist(assisted_list[1], alpha=0.5, label = 'Was not assisted')
plt.legend(loc='upper right')
plt.show()
```



As was the case with average manhattan distance, and contrary to my hypothesis, students actually visit *more* census tracts during semesters in which they are not assisted financially by their family, at a p=.005 significance level. When I set out to examine this data, I expected students without financial assistance from their families to have less spatial access than their supported peers. However, the opposite seems to be true: According to this (admittedly flawed) sample, students who are not financially assisted by their families spend more time off-campus, and do so in a broader range of locations. If we assume that this is beneficial to students in terms of experienced cultural capital, then having less economic capital might be beneficial in some ways. As my grandfather would say (in Portuguese): Necessity makes makes the frog jump!

In [20]:
```
# t-test of assisted/not-assisted and census tracts count
print stats.ttest_ind(assisted['block_count'],not_assisted['block_count'], equal_var=False)

print 'Mean # of census tracts visited for semesters where respondent was assisted: ', assisted['block_count'].mean()
print 'Mean # of census tracts visited for semesters where respondent was not assisted: ', not_assisted['block_count'].mean()
```

```
Ttest_indResult(statistic=-3.1389481396088037, pvalue=0.0053377351141343163)
Mean # of census tracts visited for semesters where respondent was assisted:  309.603448276
Mean # of census tracts visited for semesters where respondent was not assisted:  479.6875
```

Of course, this sample is tiny, and there are a host of possible errors. Further work is needed to establish what the relationship between work and assistance is in order to confirm or reject my hypothesis of low assistance leading to more time spent working and therefore more spatial use. I can do this in part by eliminating semesters where a respondent reported no assistance and no work for a wage, in order to see if the relationship changes.

I can also check to see how family income relates to the number of semesters without assistance, and determine if the observed relationship remains true across income levels. Beyond that, I can also begin to dig in to the spatial data to look for differences in *where* students go, not just how much they go somewhere other than their campus.

In [ ]:
```
```