Abstract

We address the questions about the connections between financial data (cost of tuition) and completion rate, and also consider the percentage of student loans. We process the columns by filling in null values with samples from sample distribution. We then use linear regression to explore possible correlation between each attributes. We mostly focus on approaching and discussing the problem from the student level as student loan and completion rate may be affected by each student's characteristics and factors compared to the instituion. There may be effect from the institutions, but we believe that self determination is a key factor that affect completion.

Addressing Questions

The question we choose is about connections between financial data (cost of tuition) and completion rates. We start by looking at several columns that are not part of the basic columns which are 'HIGHDEG', 'C150_4_POOLED_SUPP','C150_L4_POOLED_SUPP','NPT4_PRIV','NPT4_PUB', 'PCTFLOAN'. HIGHDEG indicates what is the highest level of award for that institution. C150_4_POOLED_SUPP and C150_L4_POOLED_SUPP indicates the completion rates for full-time, first-time students who complete within 150 percent of the expected time to completion for four years and less than four years respectively. NPT4_PRIV and NPT4_PUB describes the cost of attendance in private and public college respectively. PCTFLOAN indicates the percent of undergraduate students that receive loans. Lastly, we also use CONTROl from the basic column to differentiate the data between different control type.

Dealing with data

Before we start analyzing the data, we have to first look through the dataset and clean the data if necessary. We replace the college with `translation_dict(datadict)` from previous questions. We start by looking at NPT4_PRIV and NPT4_PUB columns from the college dataset. As there are null values in two columns and not all the rows have null values for both columns, we will create a new column NET_PRICE with the value from both columns. We will have to first fill the two columns NPT4_PRIV and NPT4_PUB that are null for one column with the value from another column if it is not also null. Afterwards for the entries that are null for both columns, we will fill Net_Price with values by drawing samples from the sample distribution of its control type. We show that the private colleges are all labeled private in NPT4_PRIV and public colleges are all labeled public in NPT4_PUB. After processing the data with null values, we also list out the deviation of mean and standard deviation for each college type to ensure it does not affect the data distribution significantly. We also use the same approach the process the columns C150_4_POOLED_SUPP and C150_L4_POOLED_SUPP.

Findings

We use linear regression to explore for correlation between the attributes. First, we try to answer the question we stated above directly by fitting a regression line to cost of tuition and completion rate of students. We get a r squared of around 0.58 and a weak postive correlation between two attributes. There is a trend based on the data where increasing completion rate of the college also indicates higher tuition fee. As using these two attributes are insufficient to tell us any interesting finding, we will also consider the student loan percentage attributes which is the second part of the question. First we will look at the relationship between the percentage of undergraduates receiving loans and average tuition fee. Interestingly, we see a stronger positive relation between the two attributes, where the r squared is around 0.78. There is a trend of college that charge higher tution fee has a higher percentage of student that also receive loan. Moreover, we also look at the the relation between percentage of students receiving loans and the completion_rate. Surprisingly, the postive correlation is very weak and only have r squared of around 0.19. We see a parabola shape for the data distribution, where more data are above the regression line (higher percentage of students receiving loan) when the completion is about 0.5. The percentage of students receiving loan on both end of completion rate are lower than the mid point of completion rate.

Short comings

Before talking about correlations and possible causation between the findings, we will talk about possible shortcomings of our analysis. A shortcoming that we thought of while analyzing the data is the lack of information of each school besides knowing it is a public or private college through `COLLEGE` field. We use `groupby` to process our data into three groups: `Private for-profit` , `Private nonprofit` and `Public` . Although this allow us to analyze the data easily with only three groups, we also lose certain information about the school in the process. School that are satellite campus rather than main campus, or in different geographical location and state may have different reasons that cause their data to behave this way.

Another shortcoming is when we group the data with the field `HIGHDEG` , we use the information to fill the null value in completion rate by drawing sample from the sample distribution. we check the field in the completion rate that are more than and less than four years. As it is only possible to graduate under 4 years or more than 4 years, we combine the two columns into one. However, as certain degree such that bachelor degree is more likely to graduate more than 4 years compare to a certificate degree. There are loss of information when we used `groupby` as we did not pay attention to whether below 4 years is higher than above 4 years when we merge two columns. Some school may mainly offers bachelor degree and some other school may only offer certain types of degree such as graduate degree.

Relationship between completion rate and average tuition fee

There are signs of correlation between certain attributes and it is helpful to think about the causation to help deciding whether the correlation really exists or the data may be biased. The first regression we show that there is a decent positive correlation between completion rate and the average tuition costs. We can see that there are more data points below the regression line when the completion rate is the lowest (lower quartile).

The reasoning behind this may be because of the lower tution fee of community college and public college which also has higher acceptance rate, which the students inside may not be as determined as others to finish their degree. As the completion rate increases, the tuiton fee also increases which may be a sign of better college such as research university or private university that has a much higher tuition fee. However, we acknowledge that sign of correlation between increasing tution fee and increasing completion fee does not mean causation as there are a lot of factors that determine whether a student can complete the college, and a college that has higher tuition fee does not guarantee better education. To support this, we can see from the graph that there are positive correlation in the beginning, but the data points slowly become a straight line after the midpoint of completion rate, which is actually shows the the increase of tuition fee from that point does not really increase the completion rate. Another possible confounders is that completion rate itself is not a good measure of anything such as the quality of the college, where including attributes such as average GPA or average happiness for the students in the college can be helpful. Graduating is definitely the main goal but there are other factors to consider that lead to the graduation. Lastly, as the field 'NPT4_PRIV','NPT4_PUB' only includes cost of attendance limited to undergraduates that pay in state tuition and does not include any financial support, there are confounders here as at some college many students receive financial aid and many students pay out of state tuition fee.

Relationship between percentage of undergraduates receiving loans and average cost of attendance

As mentioned above, the cost of attendance does not include any financial aid data for the public univertsity. Therefore we bring in the data about the percentage of student loan to analyze. We see a stronger positive correlation relationship between percentage of undergraduates receiving loans and average cost of attendance. We obtain a r squared of around 0.78 where the percentage of students receiving loan is higher at a more expensive college. We can see the correltaion between the two attributes and the causation, intuitively, makes sense that more students have to take loan at a more expensive college, as they are more likely to be able to afford at a cheaper college. A possible confounder here that was mentioned in its description of the attributes is that community college that has lower cost of tuition also has a lower loan borrowing rate. Although it can provide certain context to the loan borrowing rate, it may be difficult to compare the different borrowing behaviors across institutions that have different costs.

Relationship between percentage of undergraduates receiving loans and completion rate

Interestingly, the positive correlation between the percentage of undergraduates receiving loans and completion rate is very weak with only r squared of 0.19. Although it shows very weak correlation, increasing percentage of students loans in an instition does not provide information on its completion rate. Again, we see a parabola shape of distribution where higher and lower quartile of completion rate has lower rate of percentage of student loan. A possible reason for the lower quartile can be community that has lower cost of attendance also has a lower percentage of student loan, which may also has a lower completion rate as indiciated below. On the other hand of the higher quartile, the completion rate is high and lower percentage of student loan which may be a more prestigious institutions where students tend to study harder. There are also possibilities of scholarships from the more prestigious university either from the university itself or outside sources. The peak of the data are at the mid point of completion rate, which may be public university where people take student loan to attend the college that are not very expensive.

In [528]:

```
# Replace college by the dictionary
college = college.replace(proj.translation_dict(datadict))
# Take Control, NPT4_PRIV, NPT4_PUB from college
financial = college[['CONTROL','NPT4_PRIV','NPT4_PUB']]
# create another column name NPT as the combination of NPT4_PRIV and N
PT4_PUB
financial = financial.assign(Net_Price = financial['NPT4_PRIV'].fillna
(financial['NPT4_PUB']))
```

In [529]:

```
# We show that the private colleges are all labeled private in NPT4_PR
IV and
# public colleges are all labeled public in NPT4_PUB
financial.groupby('CONTROL').count()
```

Out[529]:

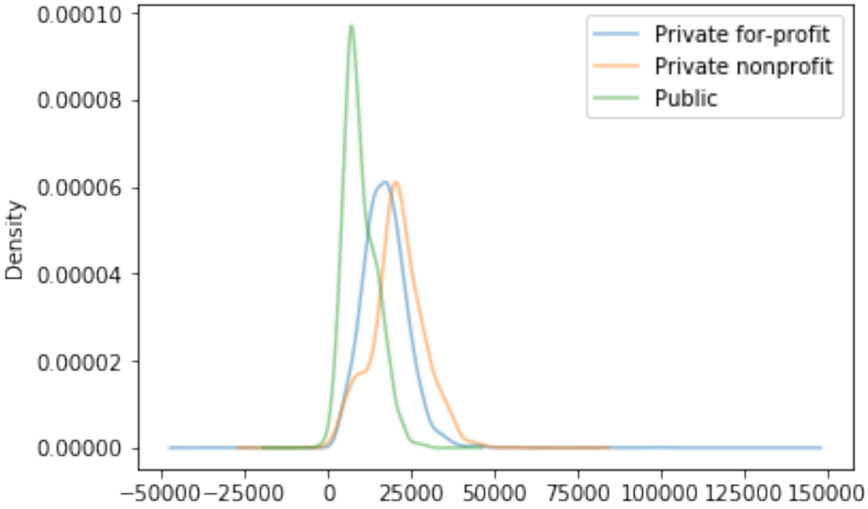
	NPT4_PRIV	NPT4_PUB	Net_Price
CONTROL			
Private for-profit	2598	0	2598
Private nonprofit	1476	0	1476
Public	0	1893	1893

In [530]:

```
# fill the na values in Net_Price based the sample distribution of its
control
financial_filled=financial.copy()
for i in financial_filled['CONTROL'].unique():
    current = financial_filled.loc[financial_filled['CONTROL']==i]
    num_null = current.Net_Price.isnull().sum()
    fill_val = current.Net_Price.dropna().sample(num_null,replace=True
)
    fill_val.index = current.loc[current.Net_Price.isnull()].index
    financial_filled = financial_filled.fillna({'Net_Price':fill_val.t
o_dict()})
```

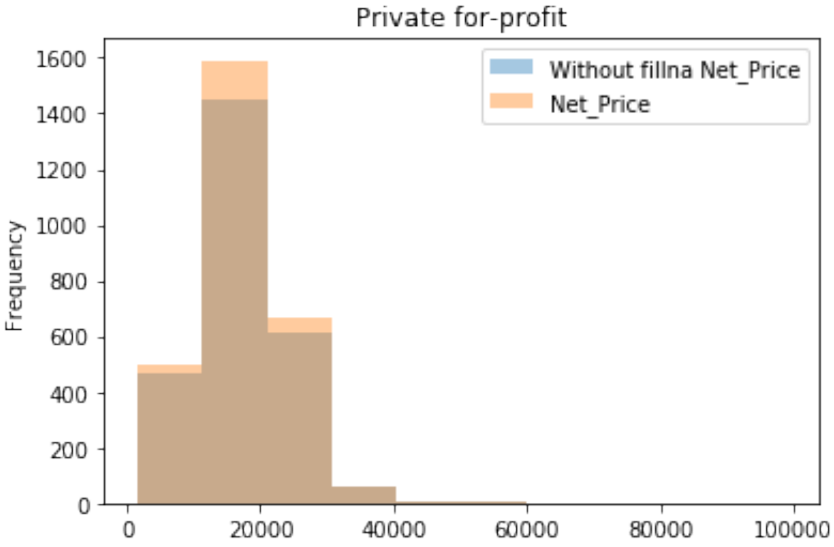
In [531]:
```
# thge private school's NPT distribution looks similar for nonprofit a
nd profit, but the public one looks different
financial_filled.groupby('CONTROL').Net_Price.plot(kind='kde',alpha=0.
5,legend=True)
```

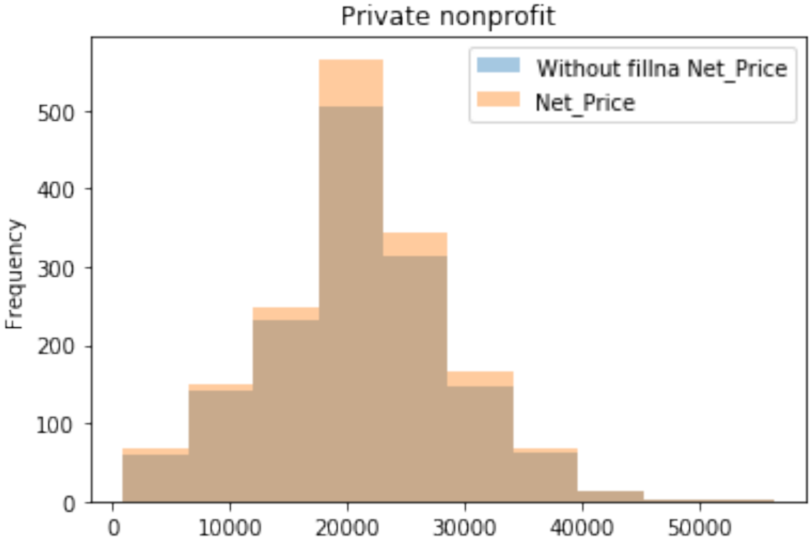
Out[531]:
```
CONTROL
Private for-profit    AxesSubplot(0.125,0.125;0.775x0.755)
Private nonprofit     AxesSubplot(0.125,0.125;0.775x0.755)
Public                AxesSubplot(0.125,0.125;0.775x0.755)
Name: Net_Price, dtype: object
```

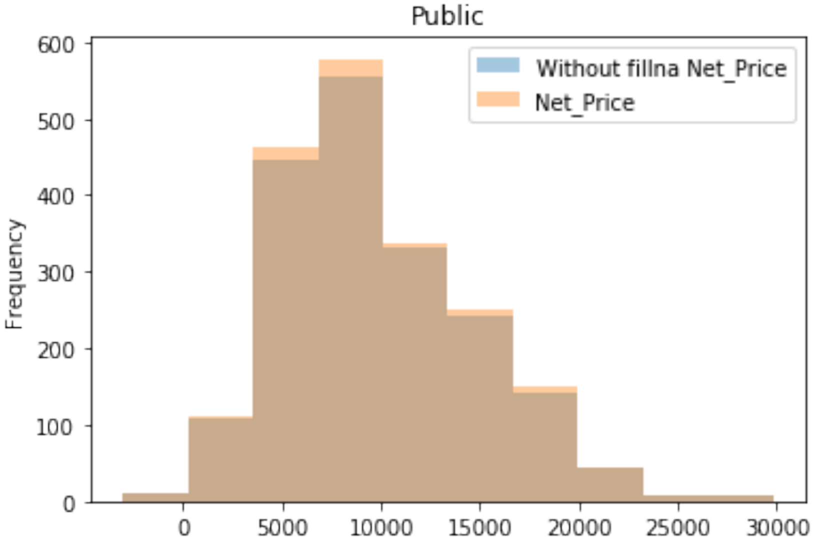


In [532]:
```
# combine the control, old NPT, and the filledna NPT
financial_graph = pd.concat([financial['CONTROL'],financial['Net_Price
'].rename(
    'Without fillna Net_Price'),financial_filled['Net_Price']],axis=1)
financial_graph.loc[financial_graph.CONTROL.isin(['Private for-profit'
])].plot(
    kind='hist',alpha=0.4,title='Private for-profit')
financial_graph.loc[financial_graph.CONTROL.isin(['Private nonprofit']
)].plot(
    kind='hist',alpha=0.4,title='Private nonprofit')
financial_graph.loc[financial_graph.CONTROL.isin(['Public'])].plot(kin
d='hist',alpha=0.4,title='Public')
```

Out[532]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a36cbc828>
```


Private for-profit
Without fillna Net_Price
Net_Price
Frequency
0
200
400
600
800
1000
1200
1400
1600
0
20000
40000
60000
80000
100000


Private nonprofit
Without fillna Net_Price
Net_Price
Frequency
0
100
200
300
400
500
0
10000
20000
30000
40000
50000


Public
Without fillna Net_Price
Net_Price
Frequency
0
100
200
300
400
500
600
0
5000
10000
15000
20000
25000
30000

In [533]:

```
# calculate the mean, std, and deviation for each control
for i in sorted(financial_filled['CONTROL'].unique()):
    i_financial = financial.loc[financial.CONTROL==i]
    i_financial_filled = financial_filled.loc[financial_filled.CONTROL
==i]

    errormean = np.abs(i_financial.Net_Price.mean()-i_financial_filled
.Net_Price.mean()
    )/i_financial.Net_Price.mean()*100
    errorstd = np.abs(i_financial.Net_Price.std()-i_financial_filled.N
et_Price.std()
    )/i_financial.Net_Price.std()*100
    print('CONTROL',i)
    print(
        'mean (original):  %f' % i_financial.Net_Price.mean(),
        'mean (filled): %f' % i_financial_filled.Net_Price.mean(),
        'mean percentage deviation: %f' % errormean,
        sep='\n'
    )
    print(
        'std (original):  %f' % i_financial.Net_Price.std(),
        'std (filled):   %f' % i_financial_filled.Net_Price.std(),
        'std percentage deviation: %f' % errorstd,
        sep='\n'
    )
    print()
```

```
CONTROL Private for-profit
mean (original):  17256.612394
mean (filled): 17278.098129
mean percentage deviation: 0.124507
std (original):  6913.645776
std (filled):   6857.700805
std percentage deviation: 0.809196

CONTROL Private nonprofit
mean (original):  20950.223577
mean (filled): 21031.460640
mean percentage deviation: 0.387762
std (original):  7941.576898
std (filled):   7898.359983
std percentage deviation: 0.544186

CONTROL Public
mean (original):  9851.179081
mean (filled): 9845.191327
mean percentage deviation: 0.060782
std (original):  4911.007691
std (filled):   4907.344353
std percentage deviation: 0.074594
```

In [534]:

```python
# Take HIGHDEG, C150_4_POOLED_SUPP, C150_L4_POOLED_SUPP from college
complete = college[['HIGHDEG','C150_4_POOLED_SUPP','C150_L4_POOLED_SUPP']]
# combine C150_4_POOLED_SUPP and C150_L4_POOLED_SUPP into one column, C150
complete = complete.assign(Completion_rate = complete['C150_L4_POOLED_SUPP'].fillna(complete['C150_4_POOLED_SUPP']))
complete['Completion_rate'] = complete['Completion_rate'].replace('PrivacySuppressed',np.NaN)
complete['Completion_rate'] = complete['Completion_rate'].astype('float')
display(complete.groupby('HIGHDEG').count())
```

	C150_4_POOLED_SUPP	C150_L4_POOLED_SUPP	Completion_rate
HIGHDEG			
Associate degree	58	1472	1497
Bachelor's degree	736	2	685
Certificate degree	3	2151	2022
Graduate degree	1611	0	1532
Non-degree-granting	1	11	7

In [535]:

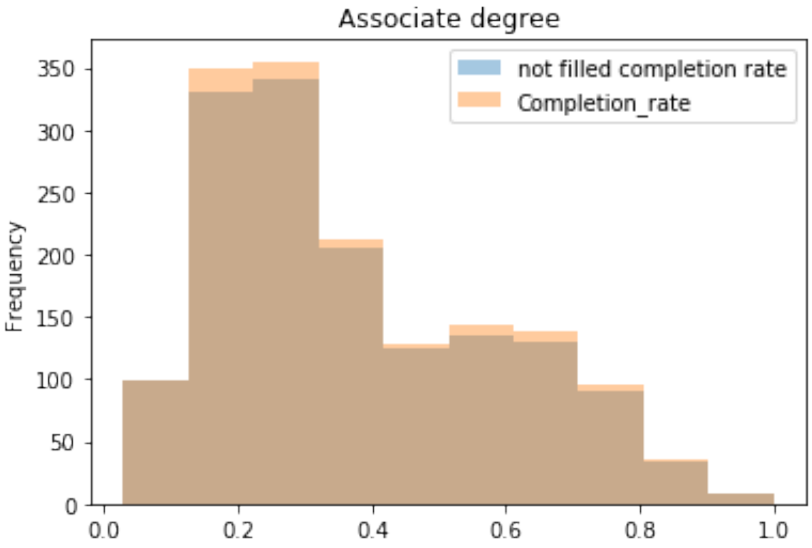
```python
# fill the na values in C150 based the sample distribution of its control
complete_filled=complete.copy()
for i in complete_filled['HIGHDEG'].unique():
    current = complete_filled.loc[complete_filled['HIGHDEG']==i]
    num_null = current.Completion_rate.isnull().sum()
    fill_val = current.Completion_rate.dropna().sample(num_null,replace=True)
    fill_val.index = current.loc[current.Completion_rate.isnull()].index
    complete_filled = complete_filled.fillna({'Completion_rate':fill_val.to_dict()})
```

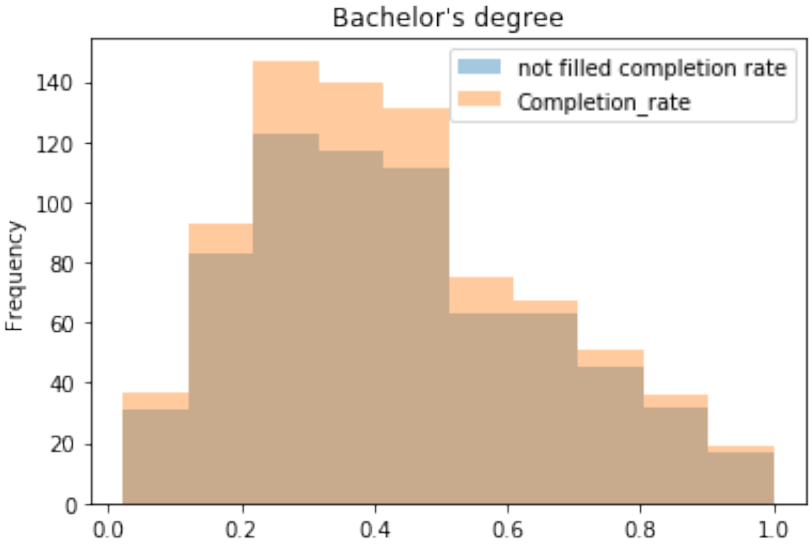
In [536]:

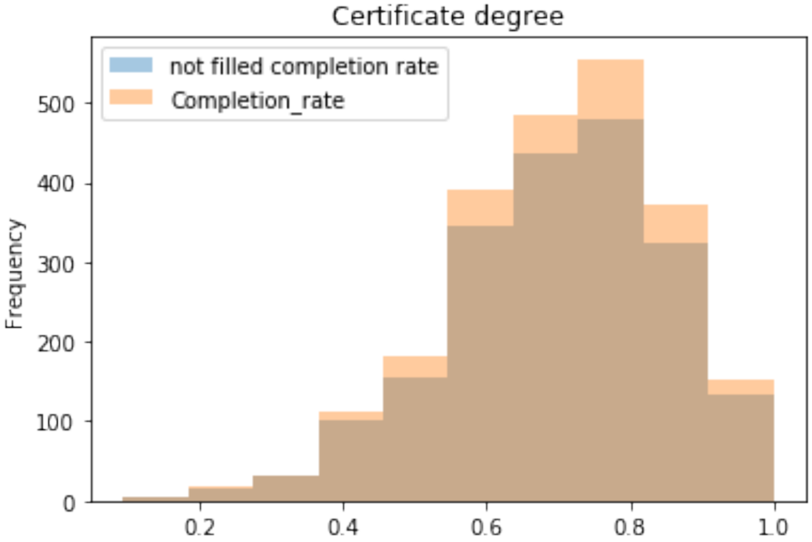
```
completion_graph = pd.concat([complete['HIGHDEG'],complete['Completion
_rate'].rename('not filled completion rate'),
        complete_filled['Completion_rate']],axis=1)
completion_graph.loc[completion_graph.HIGHDEG.isin(['Associate degree'
])].plot(kind='hist',alpha=0.4,

title='Associate degree')
completion_graph.loc[completion_graph.HIGHDEG.isin(['Bachelor\'s degre
e'])].plot(kind='hist',alpha=0.4,

title='Bachelor\'s degree')
completion_graph.loc[completion_graph.HIGHDEG.isin(['Certificate degre
e'])].plot(kind='hist',alpha=0.4,

title='Certificate degree')
completion_graph.loc[completion_graph.HIGHDEG.isin(['Graduate degree']
)].plot(kind='hist',alpha=0.4,

title='Graduate degree')
completion_graph.loc[completion_graph.HIGHDEG.isin(['Non-degree-granti
ng'])].plot(kind='hist',alpha=0.4,

title='Non-degree-granting')
```

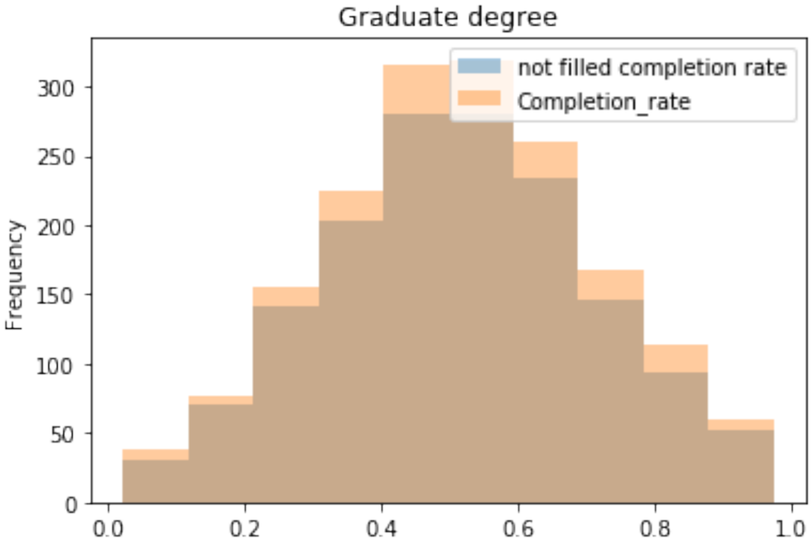
Out[536]:

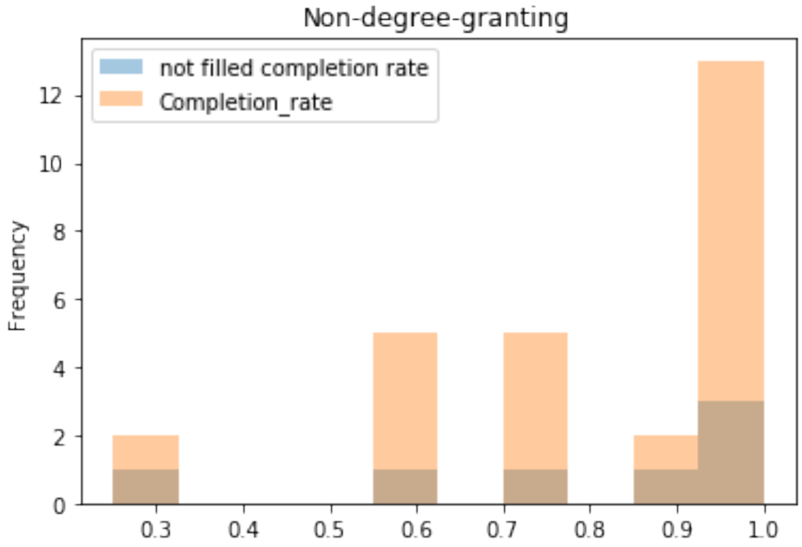
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a36c23fd0>
```




Bachelor's degree
not filled completion rate
Completion_rate
Frequency
0
20
40
60
80
100
120
140
0.0
0.2
0.4
0.6
0.8
1.0


Certificate degree
not filled completion rate
Completion_rate
Frequency
0
100
200
300
400
500
0.2
0.4
0.6
0.8
1.0


Graduate degree
not filled completion rate
Completion_rate
Frequency
0
50
100
150
200
250
300
0.0
0.2
0.4
0.6
0.8
1.0



In [537]:

```
# calculate the mean, std, and deviation for each HIGHDEG
for i in sorted(complete_filled['HIGHDEG'].unique()):
    i_complete = complete.loc[complete.HIGHDEG==i]
    i_complete_filled = complete_filled.loc[complete_filled.HIGHDEG==i
]
    errormean_C150 = np.abs(i_complete.Completion_rate.mean()-i_comple
te_filled.Completion_rate.mean()
    )/i_complete.Completion_rate.mean()*100
    errorstd_C150 = np.abs(i_complete.Completion_rate.std()-i_complete
_filled.Completion_rate.std()
    )/i_complete.Completion_rate.std()*100
    print('HIGHDEG',i)
    print(
        'mean (original):  %f' % i_complete.Completion_rate.mean(),
        'mean (filled): %f' % i_complete_filled.Completion_rate.mean()
,
        'mean percentage deviation: %f' % errormean_C150,
        sep='\n'
    )
    print(
        'std (original):  %f' % i_complete.Completion_rate.std(),
        'std (filled):   %f' % i_complete_filled.Completion_rate.std()
,
        'std percentage deviation: %f' % errorstd_C150,
        sep='\n'
    )
    print()
```

```
HIGHDEG Associate degree
mean (original):  0.372842
mean (filled): 0.373405
mean percentage deviation: 0.151015
std (original):  0.207987
std (filled):   0.207911
std percentage deviation: 0.036453

HIGHDEG Bachelor's degree
mean (original):  0.434364
mean (filled): 0.430506
mean percentage deviation: 0.888171
std (original):  0.217770
std (filled):   0.215509
std percentage deviation: 1.038449

HIGHDEG Certificate degree
mean (original):  0.698031
mean (filled): 0.699048
mean percentage deviation: 0.145693
std (original):  0.152298
std (filled):   0.152194
std percentage deviation: 0.068631

HIGHDEG Graduate degree
mean (original):  0.513049
mean (filled): 0.515796
mean percentage deviation: 0.535422
std (original):  0.195894
std (filled):   0.197104
std percentage deviation: 0.617502

HIGHDEG Non-degree-granting
mean (original):  0.776300
mean (filled): 0.809404
mean percentage deviation: 4.264293
std (original):  0.273458
std (filled):   0.220868
std percentage deviation: 19.231437
```

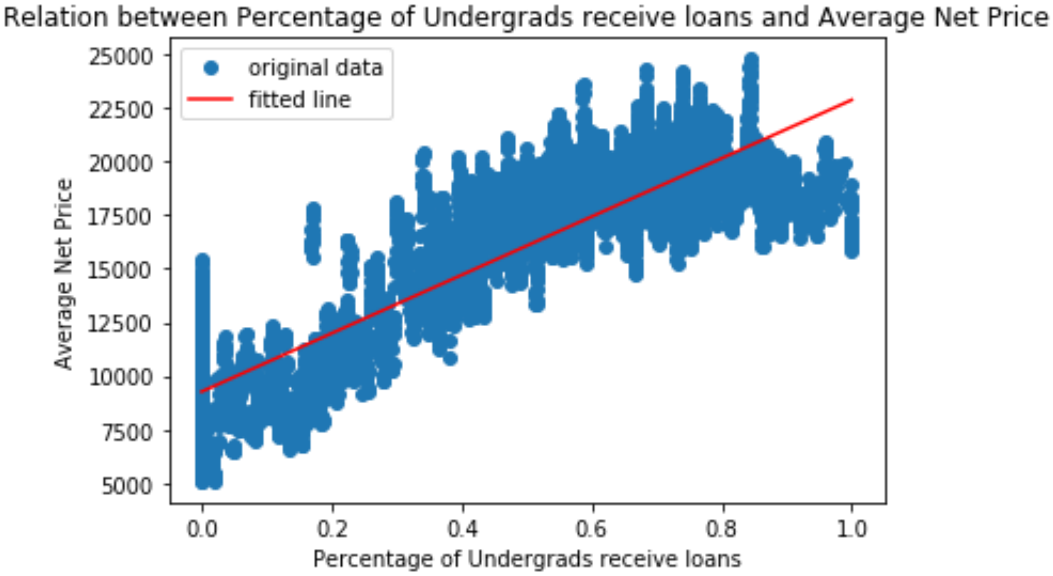
In [538]:

```
outputdf = pd.concat([college,complete_filled['Completion_rate'],financial_filled['Net_Price']],axis=1)
outputdf = outputdf[['INSTNM','PCTFLOAN','CONTROL','Net_Price','HIGHDEG','Completion_rate']]
```

In [539]:

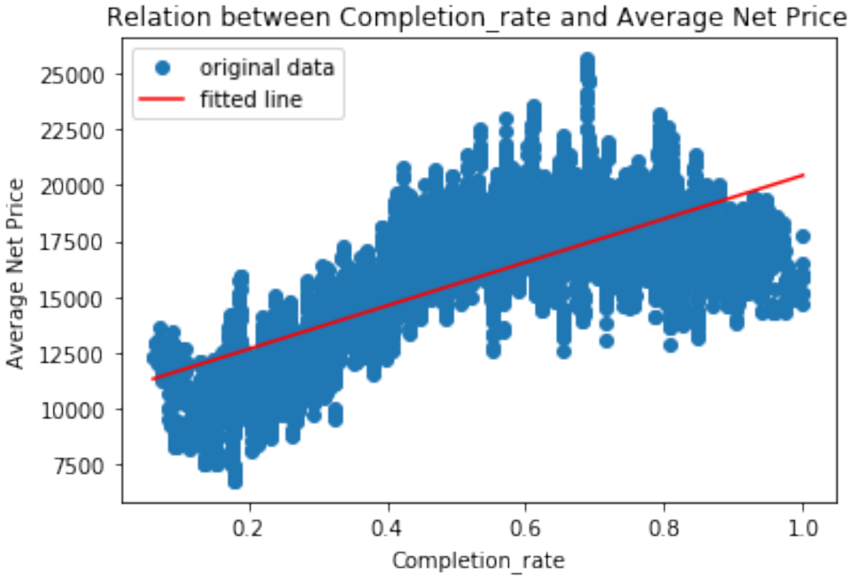
```
new = outputdf[['PCTFLOAN','Net_Price']].sort_values(by='PCTFLOAN').set_index('PCTFLOAN').rolling(window=20
).Net_Price.mean().to_frame().dropna().reset_index()
slope, intercept, r_value, p_value, std_err = stats.linregress(new['PCTFLOAN'], new['Net_Price'])
print("r-squared: %f" % r_value**2)
print("slope: %f" % slope)
x = new['PCTFLOAN']
y = new['Net_Price']
plt.plot(x, y, 'o', label='original data')
plt.plot(x, intercept + slope*x, 'r', label='fitted line')
plt.xlabel('Percentage of Undergrads receive loans')
plt.ylabel('Average Net Price')
plt.title('Relation between Percentage of Undergrads receive loans and Average Net Price')
plt.legend()
plt.show()
```

```
r-squared: 0.769147
slope: 13607.464854
```



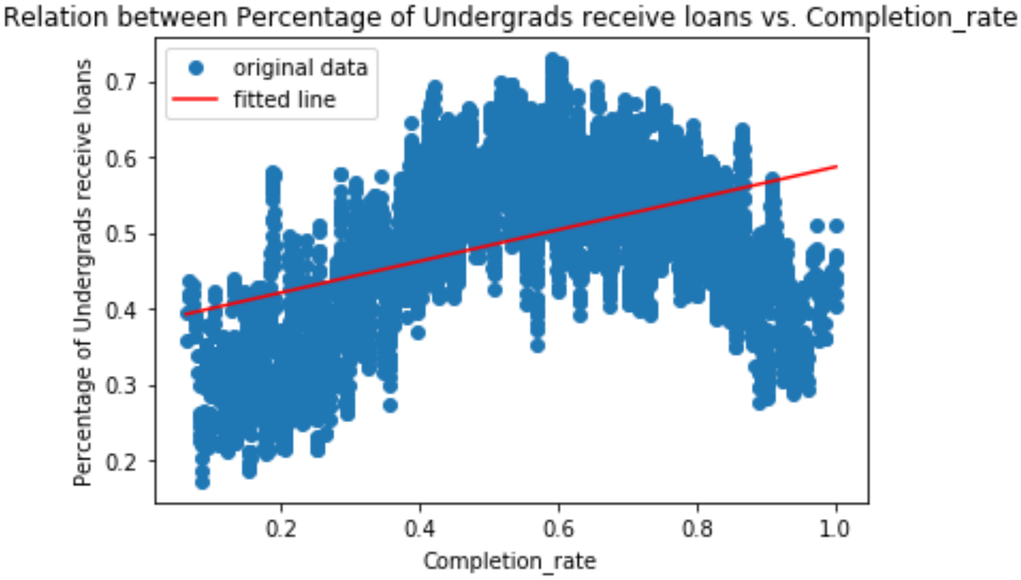
```
In [540]: new1 = outputdf[['Completion_rate','Net_Price']].sort_values(by='Completion_rate').set_index('Completion_rate'
              ).rolling(window=20).Net_Price.mean().to_frame().dropna().reset_index()
          slope, intercept, r_value, p_value, std_err = stats.linregress(new1['Completion_rate'], new1['Net_Price'])
          print("r-squared: %f" % r_value**2)
          print("slope: %f" % slope)
          x = new1['Completion_rate']
          y = new1['Net_Price']
          plt.plot(x, y, 'o', label='original data')
          plt.plot(x, intercept + slope*x, 'r', label='fitted line')
          plt.xlabel('Completion_rate')
          plt.ylabel('Average Net Price')
          plt.title('Relation between Completion_rate and Average Net Price')
          plt.legend()
          plt.show()
```

```
r-squared: 0.515485
slope: 9727.687038
```



```
In [542]: new2 = outputdf[['PCTFLOAN','Completion_rate']].sort_values(by='Completion_rate').set_index('Completion_rate').rolling(window=20).PCTFLOAN.mean().to_frame().dropna().reset_index()
slope, intercept, r_value, p_value, std_err = stats.linregress(new2['Completion_rate'], new2['PCTFLOAN'])
print("r-squared: %f" % r_value**2)
print("slope: %f" % slope)
x = new2['Completion_rate']
y = new2['PCTFLOAN']
plt.plot(x, y, 'o', label='original data')
plt.plot(x, intercept + slope*x, 'r', label='fitted line')
plt.xlabel('Completion_rate')
plt.ylabel('Percentage of Undergrads receive loans')
plt.title('Relation between Percentage of Undergrads receive loans vs. Completion_rate')
plt.legend()
plt.show()
```

```
r-squared: 0.198783
slope: 0.207411
```

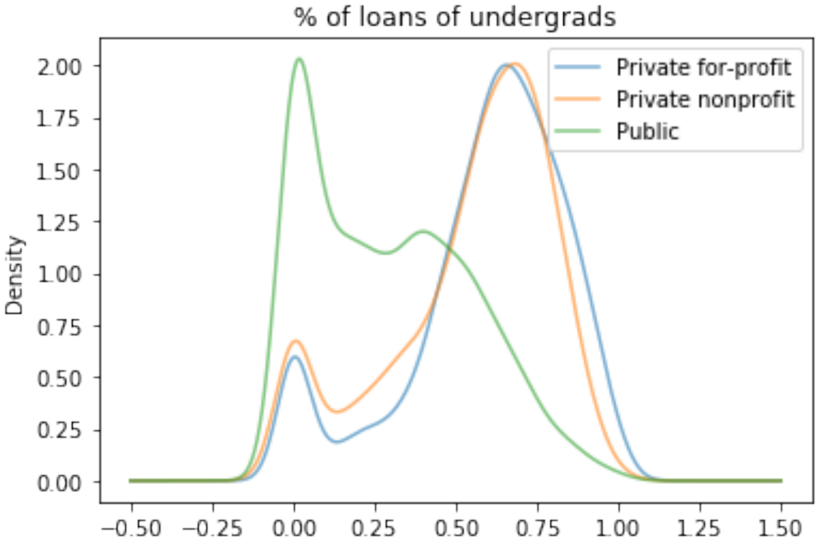


In [560]:

```
outputdf.groupby('CONTROL')['PCTFLOAN'].plot(kind='kde',alpha=0.6,legend=True,title='% of loans of undergrads')
```

Out[560]:

```
CONTROL
Private for-profit    AxesSubplot(0.125,0.125;0.775x0.755)
Private nonprofit     AxesSubplot(0.125,0.125;0.775x0.755)
Public                AxesSubplot(0.125,0.125;0.775x0.755)
Name: PCTFLOAN, dtype: object
```


% of loans of undergrads
Private for-profit
Private nonprofit
Public
Density