First-year courses and outcomes

In [3]:
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [4]:
```python
df1 = pd.read_excel("STEM Outcomes Individual Data 2002 - 2018.xlsx")
df2 = pd.read_excel("STEM Outcomes Grades 2002 - 2018.xlsx")
```

In [8]:
```python
#Merge the data on the Student ID key keeping only those records with IDs in both data sets.
df_merge = pd.merge(df1, df2, left_on = 'Student ID', right_on = 'Student ID', how = 'inner')
df_merge.head()

#Here I called the merge function onto the two datasets of grades and outcomes. I merged together the left data
```

Out[8]:

	Student ID	Start Term	End Date	Gender	Races	Ethnicity	First Generation	Major 1	Major 2	Major 3	Term	Academic Division	Subject	Course Number	Course Title
0	6599	2002FA	2005-05-22	F	WH	NHS	N	GE.AS	NaN	NaN	2002FA	4	GENS	145	Antiquity and Modernity
1	6599	2002FA	2005-05-22	F	WH	NHS	N	GE.AS	NaN	NaN	2002FA	3	GEOL	110	The Physical Earth
2	6599	2002FA	2005-05-22	F	WH	NHS	N	GE.AS	NaN	NaN	2002FA	3	GEOL	110L	The Physical Earth Lab
3	6599	2002FA	2005-05-22	F	WH	NHS	N	GE.AS	NaN	NaN	2002FA	3	ASTR	178	Sun and Stars
4	6599	2002FA	2005-05-22	F	WH	NHS	N	GE.AS	NaN	NaN	2002FA	3	ASTR	178L	Sun and Stars Laboratory

In [9]:
```python
common_degree = df_merge['Major 1'].value_counts()
common_degree.head()
#I decided to print out the top 5 awarded degrees by looking into the column of "Major 1" within the df_merge v
```

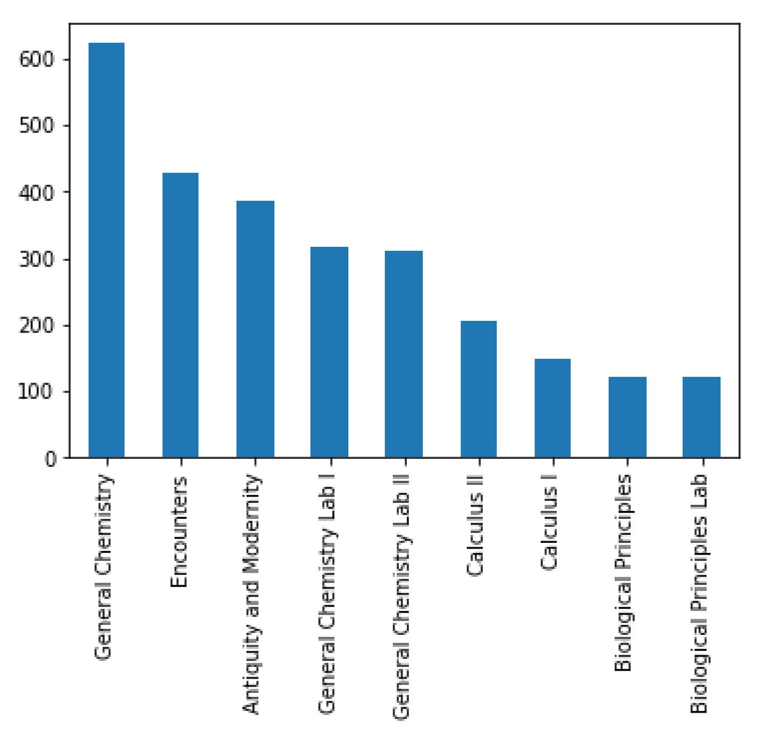
Out[9]:
```
BIOL    5100
PSYC    4151
POL     3215
BBMB    3006
ENGL    2978
Name: Major 1, dtype: int64
```

In [10]:
```python
common_classes = df_merge['Course Title'][df_merge['Major 1']== "BIOL"].value_counts()
#Here looked into the Biology major and compare it to the first year courses which Biology majors have taken.
```

Out[10]:
```
General Chemistry              623
Encounters                     429
Antiquity and Modernity        386
General Chemistry Lab I        317
General Chemistry Lab II       310
                              ...
ST:Inscribing the Archive        1
Intro to Lit and Humanities      1
CSR: Science and Religion        1
Intercoll Golf (women)           1
Intro to Asian Studies           1
Name: Course Title, Length: 486, dtype: int64
```

In [11]:
```python
common_classes[common_classes>100].plot(kind='bar')
#This figure shows a bar graph of first year classes that have been taken by Biology majors. I also made a limi
```

Out[11]: <AxesSubplot:>

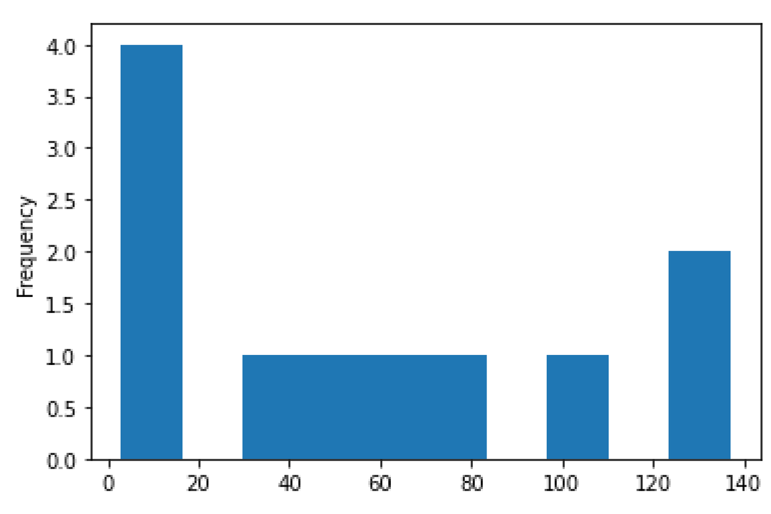


In [19]:
```python
gen_chem_bio = df_merge['Grade'][df_merge['Course Title']== "General Chemistry"][df_merge['Major 1']=="BIOL"].v
gen_chem_bio
#Here I wanted to compare the grades of Biology majors that had taken General Chemistry
```

Out[19]:
```
4.0    137
3.0    132
3.7    103
3.3     70
2.7     63
2.0     53
2.3     41
1.7      8
1.3      5
1.0      4
0.7      3
Name: Grade, dtype: int64
```

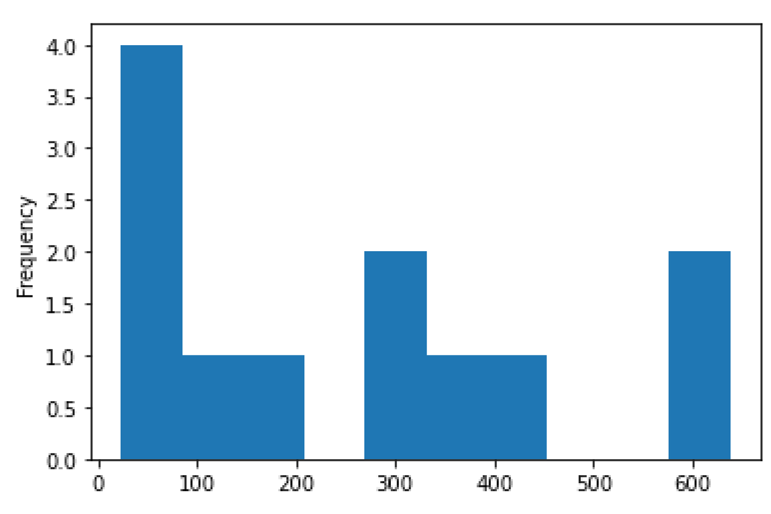
In [20]:
```python
gen_chem_bio.plot(kind="hist")
#Here is a historgram which shows the grades of Biology majors who had taken the first year course of General C
```

Out[20]: <AxesSubplot:ylabel='Frequency'>



In [23]:
```python
gen_chem = df_merge['Grade'][df_merge['Course Title']== "General Chemistry"][df_merge['Major 1']!="BIOL"].value
gen_chem.plot(kind="hist")
#Now I compared non-biology majors' grades of General Chemistry and plotted the outcome.
#It suprised me at first because I did not think that there would be this many non-Biology majors who had score
```

Out[23]: <AxesSubplot:ylabel='Frequency'>



In [24]:
```python
student = pd.merge(df1, df2, left_on = 'Student ID', right_on = 'Student ID', how = 'inner')
student
#For my next analysis, I wanted to focus on the relationship between race and majors.
```

Out[24]:

	Student ID	Start Term	End Date	Gender	Races	Ethnicity	First Generation	Major 1	Major 2	Major 3	Term	Academic Division	Subject	Course Number	Cours
0	6599	2002FA	2005-05-22	F	WH	NHS	N	GE.AS	NaN	NaN	2002FA	4	GENS	145	Antiqu Mo
1	6599	2002FA	2005-05-22	F	WH	NHS	N	GE.AS	NaN	NaN	2002FA	3	GEOL	110	The P
2	6599	2002FA	2005-05-22	F	WH	NHS	N	GE.AS	NaN	NaN	2002FA	3	GEOL	110L	The P Ea
3	6599	2002FA	2005-05-22	F	WH	NHS	N	GE.AS	NaN	NaN	2002FA	3	ASTR	178	Sun an
4	6599	2002FA	2005-05-22	F	WH	NHS	N	GE.AS	NaN	NaN	2002FA	3	ASTR	178L	Sun an Lab
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
74496	1067807	2018SP	NaT	M	WH	NHS	N	NaN	NaN	NaN	2018SP	1	POL	121	Intr Medie
74497	1067807	2018SP	NaT	M	WH	NHS	N	NaN	NaN	NaN	2018SP	1	POL	201	S Mi Ir
74498	1067807	2018SP	NaT	M	WH	NHS	N	NaN	NaN	NaN	2018SP	1	SSRA	257	Intercol B
74499	1067807	2018SP	NaT	M	WH	NHS	N	NaN	NaN	NaN	2018SP	4	GENS	146	Enco
74500	1075669	2018SP	NaT	M	NaN	NaN	N	NaN	NaN	NaN	2018SP	2	MUS	480	Comp

74501 rows × 17 columns

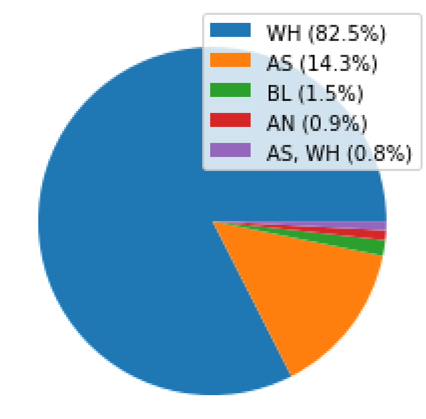
In [25]:
```python
races = student['Races'].value_counts()
races
#Got the value counts on what races students identify as.
```

Out[25]:
```
WH                  53560
AS                   7068
BL                   1622
AS, WH               1202
AN                    771
HP                    314
AN, WH                245
BL, WH                171
HP, WH                 78
AS, HP                 55
WH, AS                 54
AN, AS, WH             51
WH, AN                 47
AS, HP, WH             39
AN, AS                 34
HP, AS                 33
AN, BL, WH             28
AN, WH, AS             24
HP, AS, WH             21
AS, BL, WH             19
AN, AS, BL, WH         19
AS, BL                 15
BL, HP, WH             13
BL, AS, WH             12
WH, WH                 12
AS, BL, WH, AN         11
WH, HP                 10
AN, AS, HP, WH         10
WH, AN, AS             10
AS, AN                 10
Name: Races, dtype: int64
```

In [26]:
```python
major_race = student['Races'][student['Major 1']== "CHEM"].value_counts()
major_race
#I now compared the races with a certain major. Since I noticed that General Chemistry was one of the most popu
```

Out[26]:
```
WH        1188
AS         206
BL          21
AN          13
AS, WH      12
Name: Races, dtype: int64
```

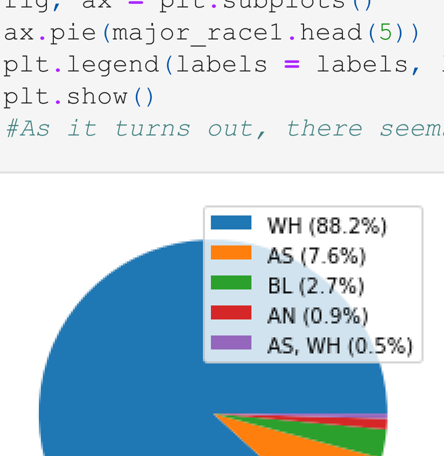
In [55]:
```python
labels = ['WH (82.5%)','AS (14.3%)','BL (1.5%)','AN (0.9%)','AS, WH (0.8%)']
fig, ax = plt.subplots()
ax.pie(major_race)
plt.legend(labels = labels, loc='best')
plt.show()
#Here I displayed a piechart that shows what percentage of Chemistry majors are associated with what race. As y
```



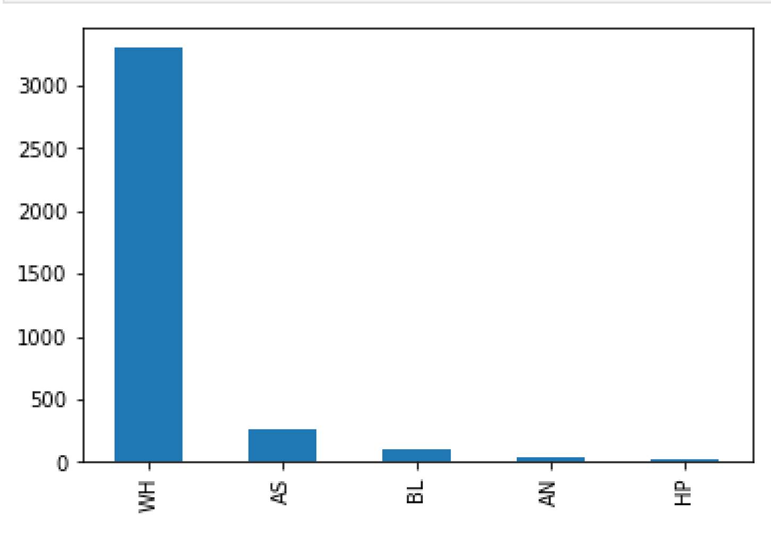
In [48]:
```python
major_race1 = student['Races'][student['Major 1']== "ENGL"].value_counts()
major_race1.head(5)
#I repeated the comparison from above with another major, English.
```

Out[48]:
```
WH        2312
AS         200
BL          70
AN          26
WH, AN      12
Name: Races, dtype: int64
```

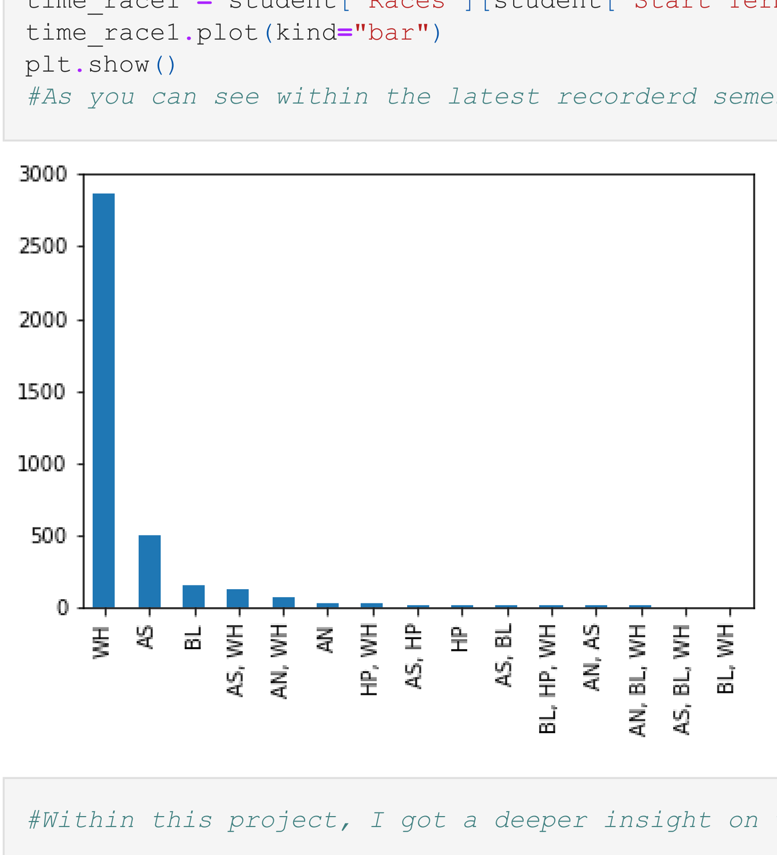
In [50]:
```python
labels = ['WH (88.2%)','AS (7.6%)','BL (2.7%)','AN (0.9%)','AS, WH (0.5%)']
fig, ax = plt.subplots()
ax.pie(major_race1.head(5))
plt.legend(labels = labels, loc='best')
plt.show()
#As it turns out, there seems to be slightly larger percentage of English majors that are white than Chemistry
```



In [53]:
```python
#The data above got me curious about the diversity of Whitman at the start(earliest date) of the dataset and at
time_race = student['Races'][student['Start Term']=="2002FA"].value_counts()
time_race.plot(kind="bar")
plt.show()
#This bar plot shows that at the earliest recorded semester within the dataset, there were only 5 different rac
```



In [54]:
```python
time_race1 = student['Races'][student['Start Term']=="2017FA"].value_counts()
time_race1.plot(kind="bar")
plt.show()
#As you can see within the latest recorderd semester at Whitman, there is a much larger range of diversity iden
```



In [ ]:
```python
#Within this project, I got a deeper insight on what type of education route students tend to go on depending o
```