In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:
```
data =pd.read_csv('./cleanData.csv')
```

In [3]:
```
data.head()
```

Out[3]:

	new_id	batch	ssctotal	sscyear	hsctotal	hscyear	dif_ssc_hsc	dif_hsc_uni	drop_out	semest
0	6	20160	4.63	2013.0	3.83	2015.0	0	1	regular	
1	7	20162	5.00	2009.0	5.00	2011.0	0	5	regular	
2	10	20161	5.00	2012.0	4.10	2014.0	0	2	dropOut	
3	12	20170	4.75	2013.0	4.00	2015.0	0	2	regular	
4	13	20120	3.94	2009.0	3.90	2011.0	0	1	dropOut	

5 rows × 25 columns

In [4]:

```
dropout=data.groupby(["batch"])["batch"].count()/len(data)*100
dropout.sort_values(ascending=False, inplace=True)
print(dropout)
```

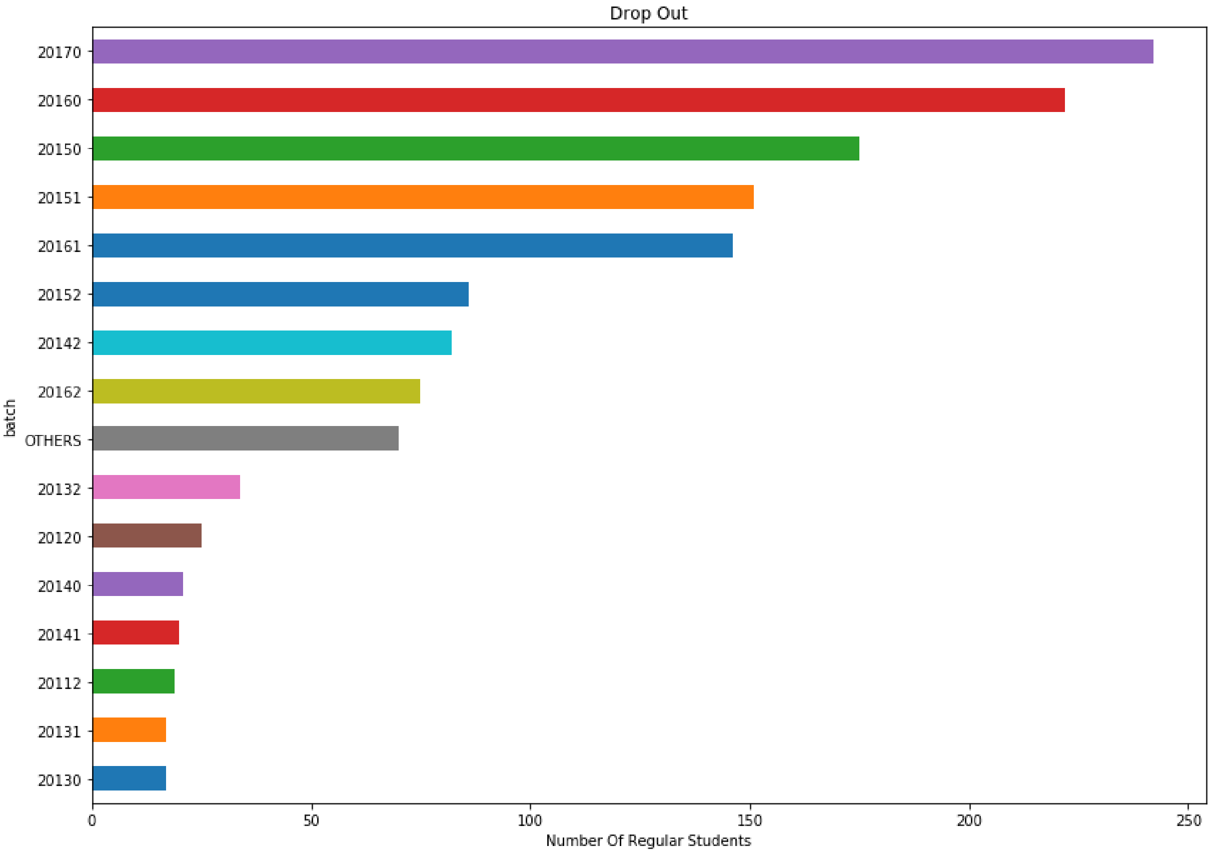
```
batch
20170    17.261056
20160    15.834522
20150    12.482168
20151    10.770328
20161    10.413695
20152     6.134094
20142     5.848787
20162     5.349501
20132     2.425107
20120     1.783167
20140     1.497860
20141     1.426534
20112     1.355207
20131     1.212553
20130     1.212553
20100     0.855920
20110     0.855920
20090     0.713267
20122     0.427960
20102     0.427960
20092     0.427960
20121     0.285307
20101     0.213980
20111     0.213980
20080     0.213980
20091     0.142653
20082     0.142653
20081     0.071327
Name: batch, dtype: float64
```

In [5]:
```
all_classes = data.groupby(['batch'])['dif_hsc_uni'].size().reset_index()
all_classes['Count'] = all_classes['dif_hsc_uni']
all_classes = all_classes.drop(['dif_hsc_uni'], axis=1)
all_classes = all_classes.sort_values(['Count'], ascending=[False])

unwanted_classes = all_classes.tail(13)
unwanted_classes
```

Out[5]:

	batch	Count
6	20100	12
9	20110	12
3	20090	10
14	20122	6
8	20102	6
5	20092	6
13	20121	4
10	20111	3
7	20101	3
0	20080	3
4	20091	2
2	20082	2
1	20081	1

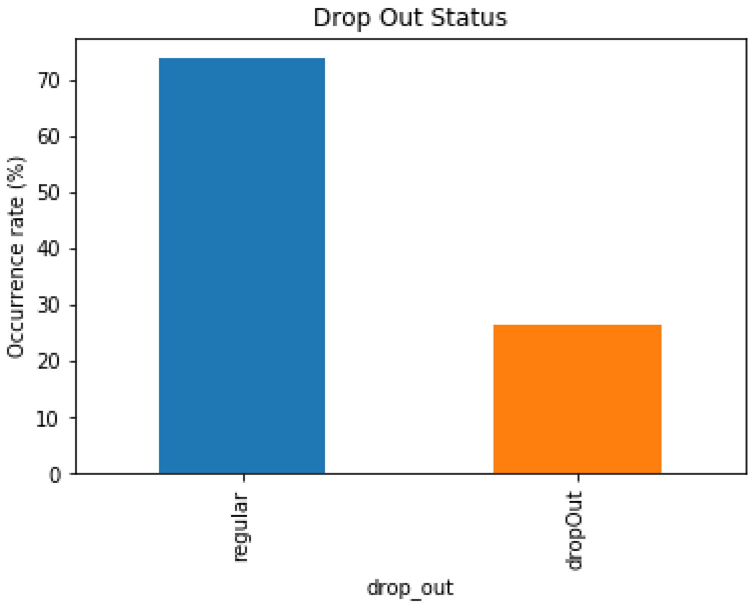
In [6]:
```
data.loc[data['batch'].isin(unwanted_classes['batch']), 'batch'] = 'OTHERS'

plt.figure(figsize=(14,10))
plt.title('Drop Out')
plt.ylabel('Batch')
plt.xlabel('Number Of Regular Students')

data.groupby([data['batch']]).size().sort_values(ascending=True).plot(kind='barh'

plt.show()
```



In [7]:
```
# Occurrence rates of student dropout
type=data.groupby(["drop_out"])["drop_out"].count()/len(data)*100
type.sort_values(ascending=False, inplace=True)
print(type)

# show graph for drop out and its occurance rate
type.plot(kind='bar',title="Drop Out Status")
plt.ylabel('Occurrence rate (%)')
```

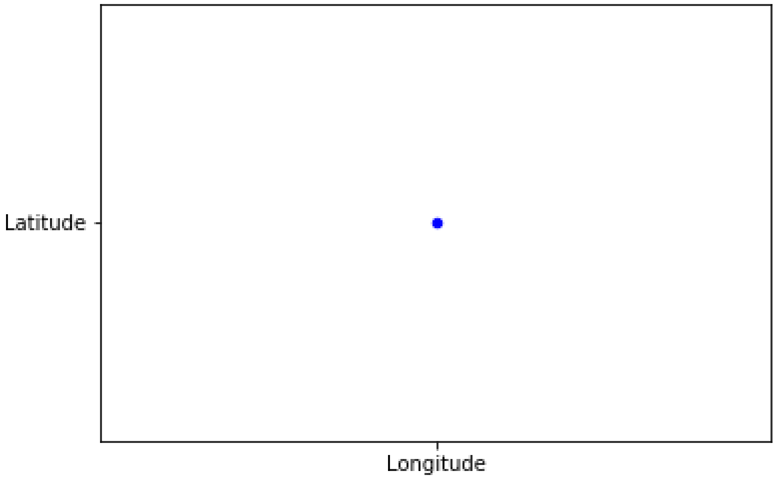
```
drop_out
regular    73.751783
dropOut    26.248217
Name: drop_out, dtype: float64
```

Out[7]:
```
Text(0,0.5,'Occurrence rate (%)')
```



Check if there are Outliers

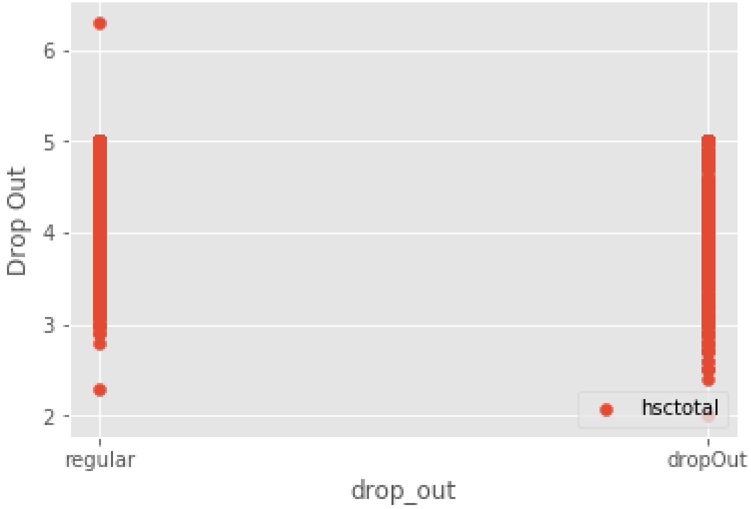
In [8]:
```
#Visualization of the Longitude and Latitude.
plt.scatter('Longitude', 'Latitude', c='blue', data=data, s=20)

plt.show();
```



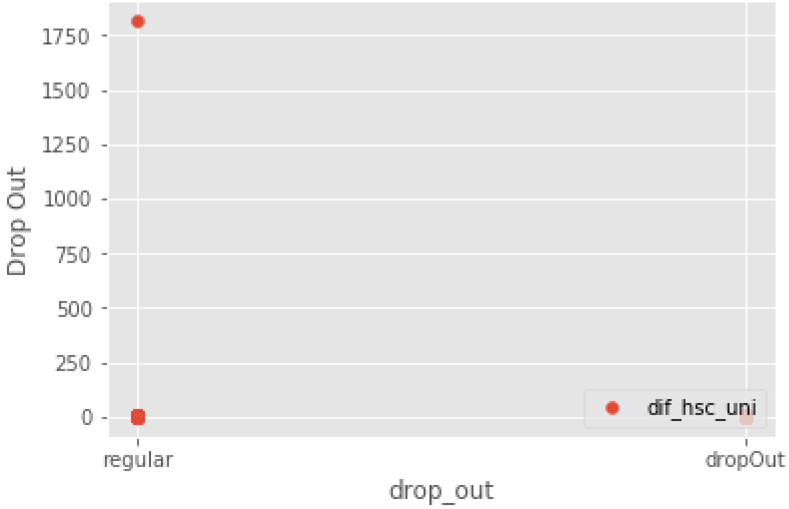
Data Plotting

In [9]:
```
from matplotlib import style
```

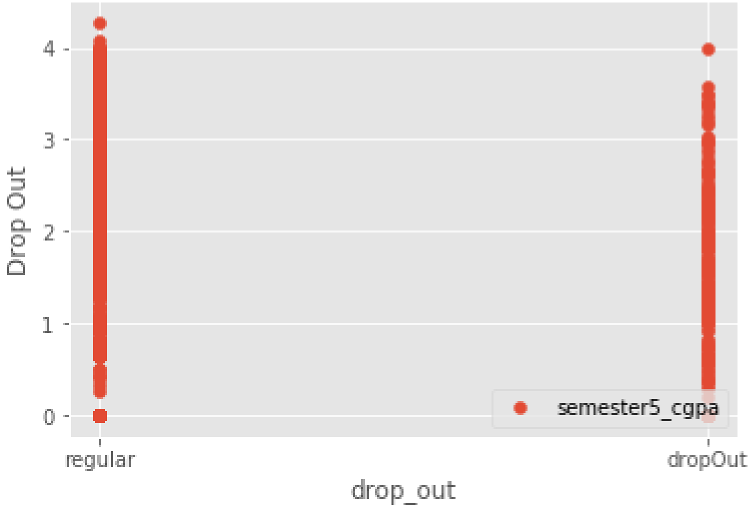
In [10]:
```
style.use("ggplot")


# Drawing and plotting model
plot = "drop_out"
plt.scatter(data[plot], data["hsctotal"])
plt.legend(loc=4)
plt.xlabel(plot)
plt.ylabel("Drop Out")
plt.show()
```



In [11]:
```
style.use("ggplot")


# Drawing and plotting model
plot = "drop_out"
plt.scatter(data[plot], data["dif_hsc_uni"])
plt.legend(loc=4)
plt.xlabel(plot)
plt.ylabel("Drop Out")
plt.show()
```

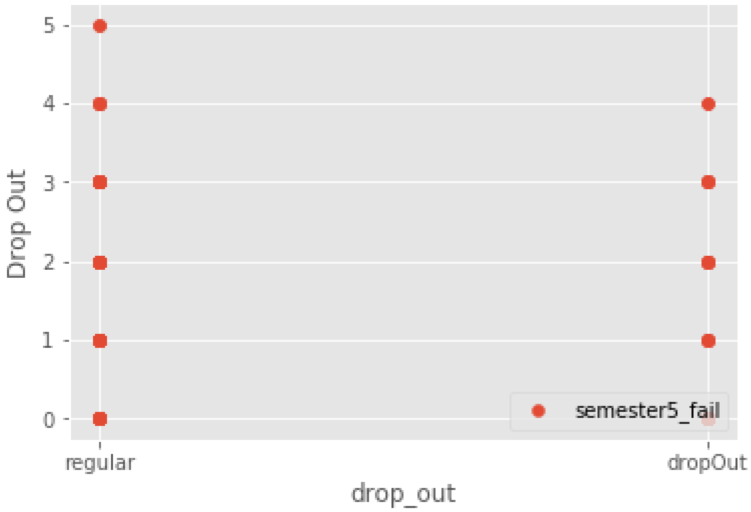


In [12]:
```
style.use("ggplot")


# Drawing and plotting model
plot = "drop_out"
plt.scatter(data[plot], data["semester5_cgpa"])
plt.legend(loc=4)
plt.xlabel(plot)
plt.ylabel("Drop Out")
plt.show()
```



In [13]:

```
style.use("ggplot")


# Drawing and plotting model
plot = "drop_out"
plt.scatter(data[plot], data["semester5_fail"])
plt.legend(loc=4)
plt.xlabel(plot)
plt.ylabel("Drop Out")
plt.show()
```



In [ ]: