In [25]:
```
from csv import reader
csv_reader = reader(open("popdata.csv"))
```

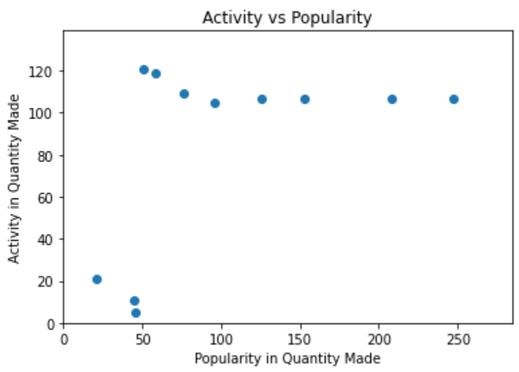
In [26]:
```
POP = {}
ACT = {}
next(csv_reader)
for row in csv_reader:
    if(row[0]) == 'TOTAL':
        continue
    POP[row[0]] = float(row[1])
    ACT[row[0]] = float(row[2])
print(POP)
print(ACT)
```

```
{'account5': 45.0, 'blah': 125.27, 'dataman': 46.0, 'jrc4615': 95.75, 'me': 247.5, 'newguy': 153.0, 'patrick': 208.25, 'test': 76.5, 'test3': 21.5, 'username1': 58.5, 'wrong': 51.0}
{'account5': 11.0, 'blah': 106.5, 'dataman': 5.0, 'jrc4615': 104.88, 'me': 106.5, 'newguy': 106.5, 'patrick': 106.5, 'test': 109.13, 'test3': 21.0, 'username1': 118.88, 'wrong': 120.88}
```

In [27]:
```
from matplotlib import pyplot as plt
```

In [28]:
```
plt.scatter(POP.values(), ACT.values())
plt.title("Activity vs Popularity")
plt.xlim(0, max(POP.values()) * 1.15)
plt.ylim(0, max(ACT.values()) * 1.15)
plt.xlabel("Popularity in Quantity Made")
plt.ylabel("Activity in Quantity Made")
```

Out[28]:
```
Text(0, 0.5, 'Activity in Quantity Made')
```



In [29]:
```
from sklearn import metrics
from sklearn import cluster
```

In [30]:
```
Kmean = cluster.KMeans(n_clusters = 2, init = "random", algorithm = "auto")
```

In [32]:
```
# convert data into [(pop1, act1), ...]
l1 = list(map(float, POP.values()))
l2 = list(map(float, ACT.values()))
X = list(map(list, zip(l1, l2)))
```

In [33]:
```
kmeans = Kmean.fit(X)
# Run a prediction on the dataset and encode 'red' for cluster 0
# encode 'blue' for cluster 1
colors = ['red' if x == 1 else 'blue' for x in kmeans.predict(X)]
blues = []
reds = []
for i, point in enumerate(X):
    if(colors[i] == 'blue'):
        blues.append(point)
    else:
        reds.append(point)
bluesx, bluesy = [p[0] for p in blues], [p[1] for p in blues]
redsx, redsy = [p[0] for p in reds], [p[1] for p in reds]
```

In [77]:
```
# regression
from sklearn import linear_model as lm
import numpy as np
x = np.array(bluesx + redsx).reshape((-1, 1))
y = np.array(bluesy + redsy).reshape((-1, 1))
reg = lm.LinearRegression().fit(x, y)
m = reg.coef_[0][0]
b = reg.intercept_[0]
r = reg.score(x, y)
```

In [86]:
```
fig = plt.figure(1, facecolor = "grey")
ax = fig.add_axes([0, 0, 1, 1])
ax.scatter(bluesx, bluesy, color = 'b')
ax.scatter(redsx, redsy, color = 'r')
ax.set_title("Clustered Activity vs Popularity")
ax.set_xlabel("Popularity in Quantity Made")
ax.set_ylabel("Activity in Quantity Made")
plt.plot(x, m*x + b)
text = f'\ny = {round(m,3)}x + {round(b, 3)}\nr={round(r, 4)}'
ax.text(0.65, 0.2, text, transform=ax.transAxes, fontsize=12,
        verticalalignment='top')
plt.show()
```