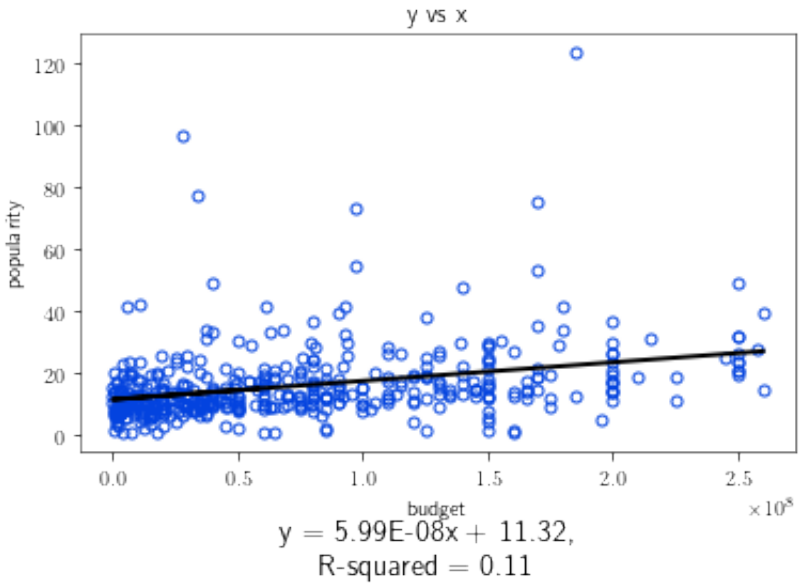
In [37]:
```
from imdb import IMDb
import pandas as pd
from scipy import stats
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
from plotting import master_plot
```

In [38]:
```
path = '/Users/austinbenny/Documents/python/movie budget ratings/the-m
```

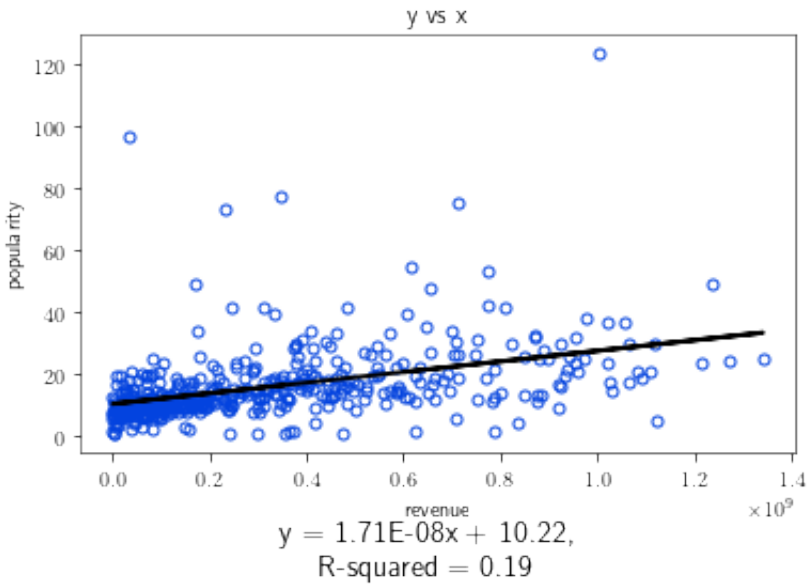
In [40]:
```
df=pd.read_csv(path)
df.replace([np.inf, -np.inf], np.nan)
df.dropna(inplace=True)
df = df[(df[['budget','revenue','popularity','vote_average']] != 0).al
z_scores=np.abs(stats.zscore(df[['budget','revenue','popularity','vote
new_df=df[(z_scores < 3.0).all(axis=1)]
df=new_df
df.columns.values
```

Out[40]:
```
array(['adult', 'belongs_to_collection', 'budget', 'genres', 'homepag
e',
       'id', 'imdb_id', 'original_language', 'original_title', 'overv
iew',
       'popularity', 'poster_path', 'production_companies',
       'production_countries', 'release_date', 'revenue', 'runtime',
       'spoken_languages', 'status', 'tagline', 'title', 'video',
       'vote_average', 'vote_count'], dtype=object)
```

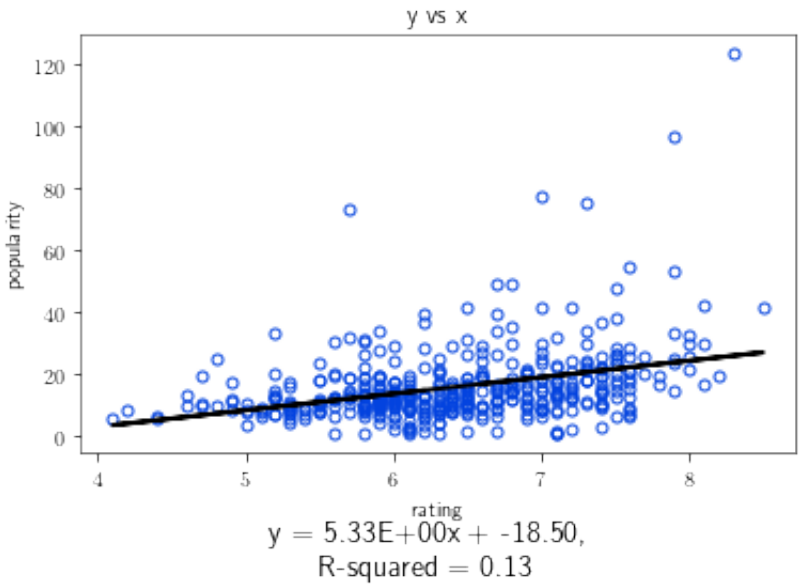
In [44]:
```
# plt.scatter(df.loc[:,'budget'],df.loc[:,'popularity'],alpha=0.5)
# plt.xlabel('budget');plt.ylabel('popularity')
master_plot(df.loc[:,'budget'],df.loc[:,'popularity'],linestyle='none'
            xlabel='budget',ylabel='popularity',legend=False)
```



In [45]:
```
master_plot(df.loc[:,'revenue'],df.loc[:,'popularity'],linestyle='none
            xlabel='revenue',ylabel='popularity',legend=False)
```



In [46]:
```
master_plot(df.loc[:,'vote_average'],df.loc[:,'popularity'],linestyle=
            xlabel='rating',ylabel='popularity')
```



What this means?

the tmdb popularity values for the cleared up df arent really functions of the rating, revenue, or budget

In [ ]:
```
```