Data Mining: Homework 1

Maha Alkhairy

Question 1:

1.a:

number of distinct authors: 1478733

number of distinct venues: 255685

number of distinct publications: 1976815

1.b:

The number of publications is likely to be correct because they are indexed uniquely.

The number of distinct venues could be off because the venue could have more that one way of being referred to.

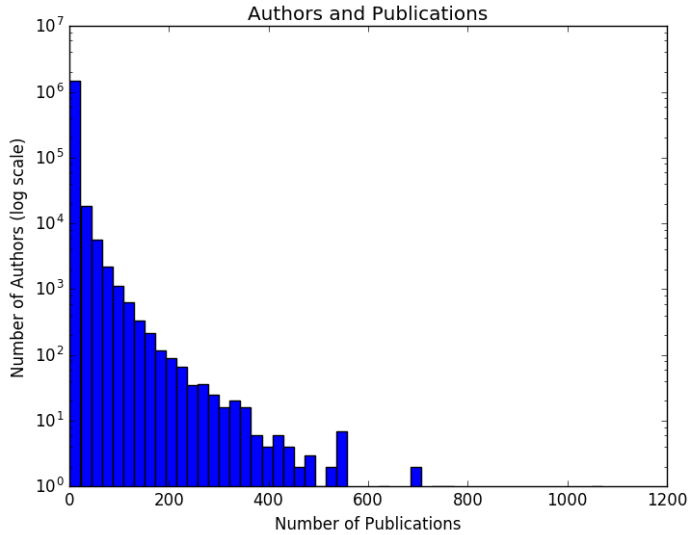
The number of distinct authors could be off because the author could have be referred to by more than one different way in the publications.

1.c:

The authors could have been referred to by more than one different way in the publications.

Question 2:

2.a:



2.b:

Number of publications per author:

- Mean: 3.29178560295
- Standard deviation: 8.87188969104
- Quantiles: [1. 1. 1. 2.]
- Median 1
- Q1: 1
- Q3: 1

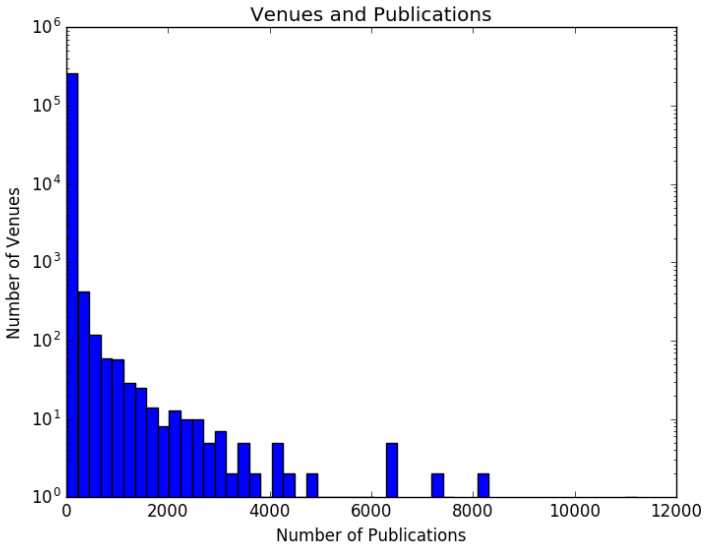
median = 1, mean = 3.2917856029, third quartile = 1 so 75% of the data has the same value. since the mean is higher than the median, that means that the data is skewed and that the difference between the max and min values is high. the standard deviation is 8 which means the data is spread out.

2.c:

Number of publications per venue:

- Mean: 7.73096583687
- Standard deviation: 83.3809075408
- Quantiles: [1. 1. 1. 1.]
- Median: 1
- Q1: 1
- Q3: 1

Venue with largest number of publications: IEEE Transactions on Information Theory

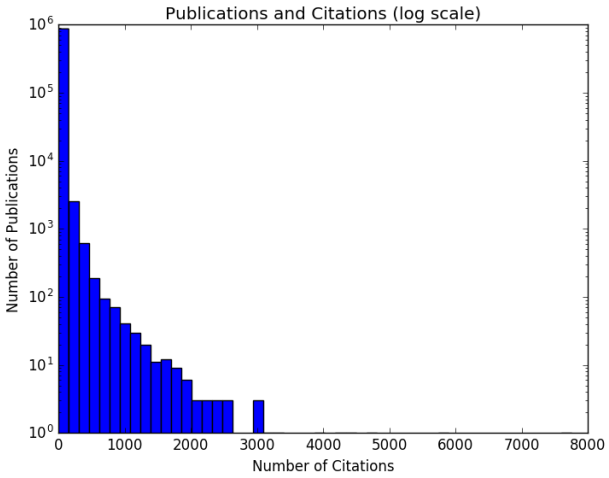


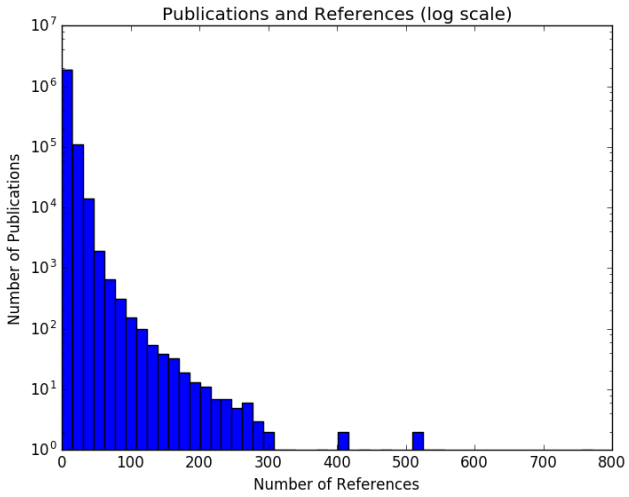
Question 3:

3.a:

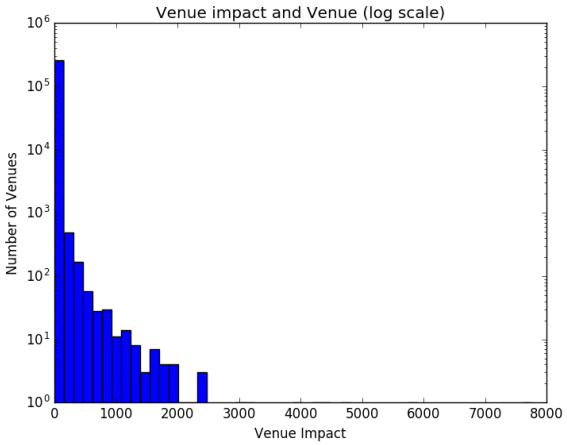
Publication with largest number of references (index, title): (719353, 'Cited References')

Publication with largest number of citations (index, title): (408396, 'Computers and Intractability: A Guide to the Theory of NP-Completeness')





3.b:



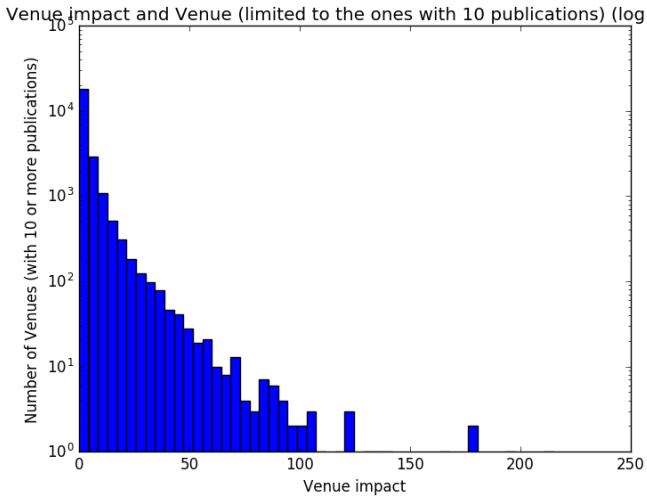
3.c:

Venue with largest apparent impact: Computers and Intractability: A Guide to the Theory of NP-Completeness

Largest impact factor: 7753.0

The maximum impact factor for a venue could be off because the venues might not be uniquely named as people refer to the same venue using different names.

3.d:



The histogram changes to be a subset of the histogram of the venue impact factor per venue. This decreases the range of the impact factors for venues with 10 or more publications.

Venue with largest apparent impact and with 10 or more publications: Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications

Largest impact factor for venues with 10 or more publications: 214.82608695652175

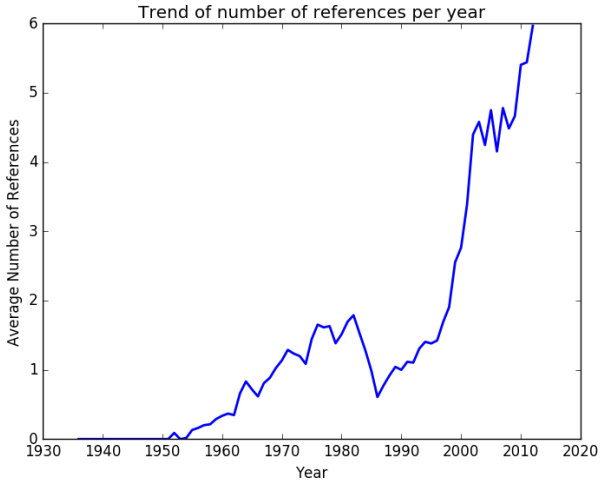
The citation counts of all the publications from the venue with the largest impact factor: [7753]

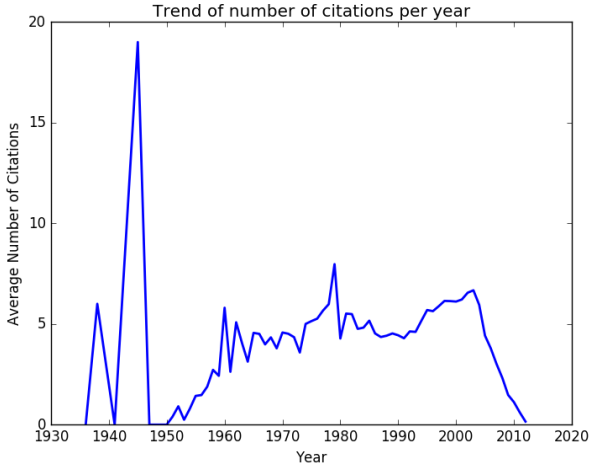
The citation counts of all the publications from the venue (with at least 10 publications) with the largest impact factor: [167, 115, 48, 11, 138, 50, 10, 11, 112, 103, 27, 2140, 1609, 87, 22, 67, 19, 28, 22, 21, 42, 46, 46]

mean number of citations for the venue with the largest impact and 10 or more publications: 214.826086957

median number of citations for the venue with the largest impact and 10 or more publications: 22.

The mean is much greater than the median which means that the data is skewed and that the maximum number of citation for the venue with the highest impact factor and 10 or more publications is a lot greater than the minimum.





3.e:

In the trend for the average number of references per year, it is low in the older years and peaks in year 2010.

In the trend for the average number of citations, it is higher in the older years and is decreasing in the later years and peaks in year 1945.

The average number pf references generally increase as a function of time (except for 1982 – 1990) but the average number of citations is not increasing or decreasing as a function time.