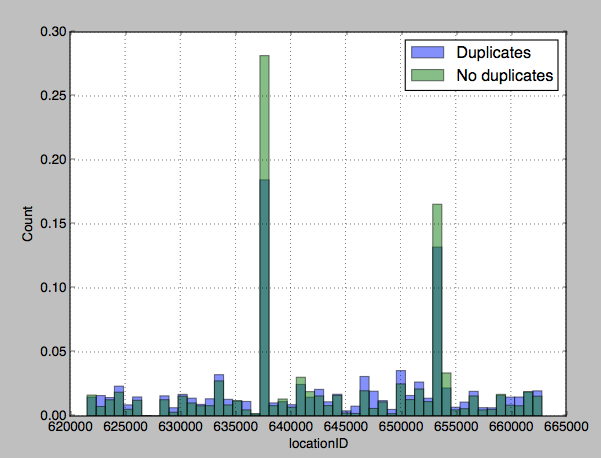
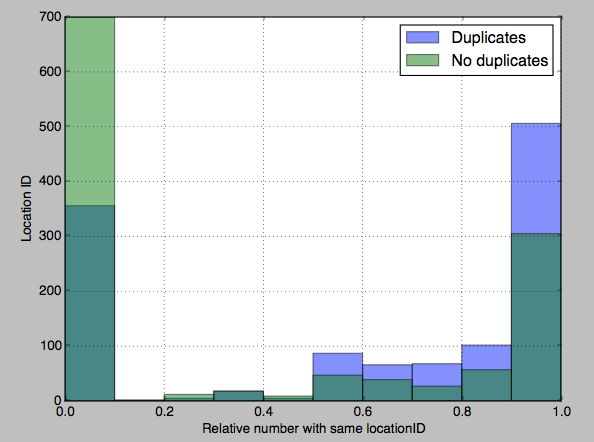
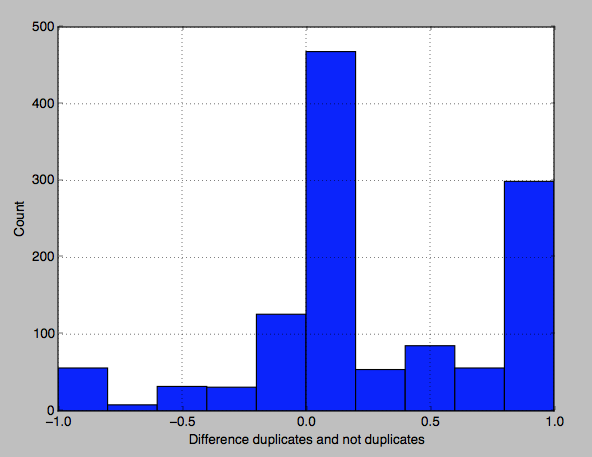
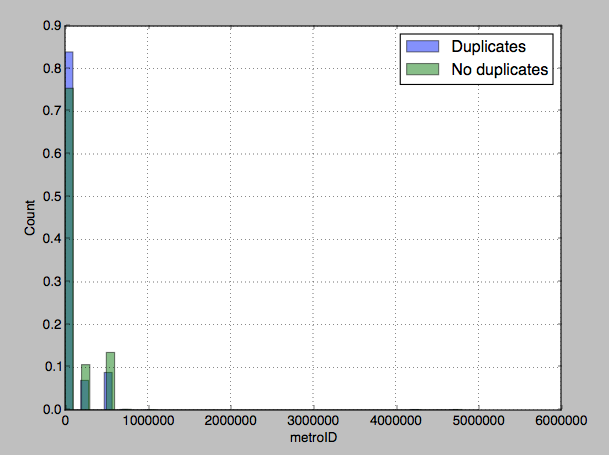
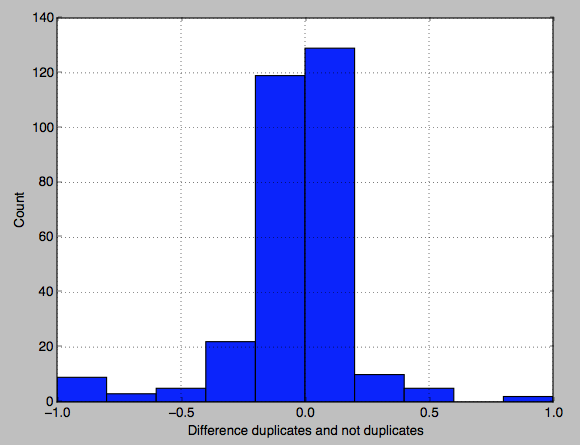
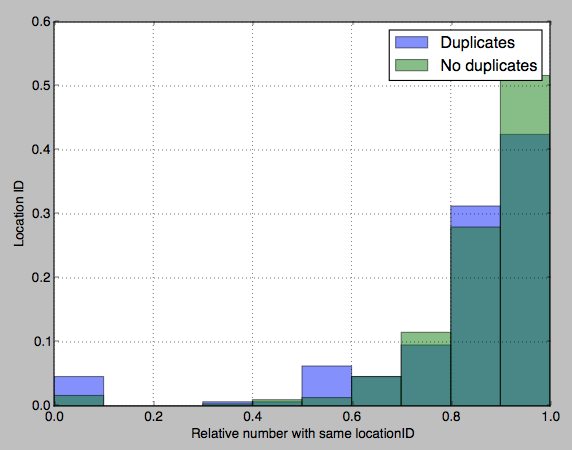
Price:

* Higher price more duplicates?

Location ID

* 91% of duplicates have same locationID
* 94% of non-duplicates have same locationID
* 
* Some locationID have very number of duplicates: smaller, so we can potentially identify locationID which are hot, but they are small
* 
* 
* Histogram of relative number with same locationID in duplicates - relative number with same locationID in non-duplicates
* Value 0: region where the relative number is the same
* Negative values: non-duplicates have more with same locationID
* Positve values: duplicates have more
* There are some regions with exclusively duplicates, but it represents a small fraction
* TO DO:
  + Is this true for alle dataset

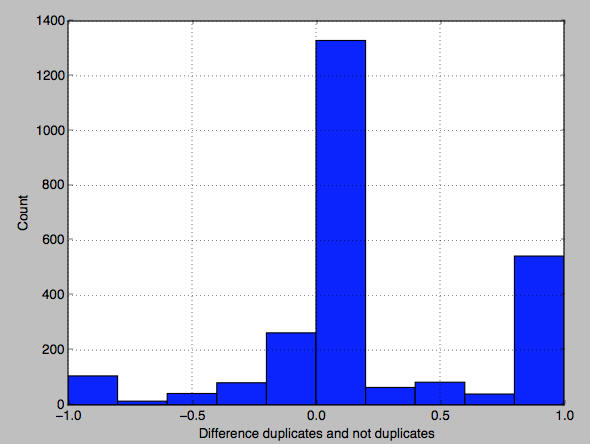
MetroID

* Equal for 96% duplicates vs 95% non-duplicates
* I do not see an opportunity to improve

CategoryID:

* Both 100%, only duplicates from within categories

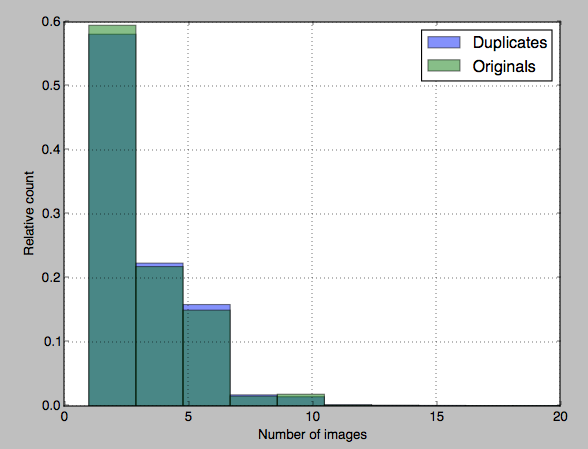
Lon:

* 77% duplicates vs 85% originals
* There are some coordinates with high amounts of duplicates, but it represents a very small fraction

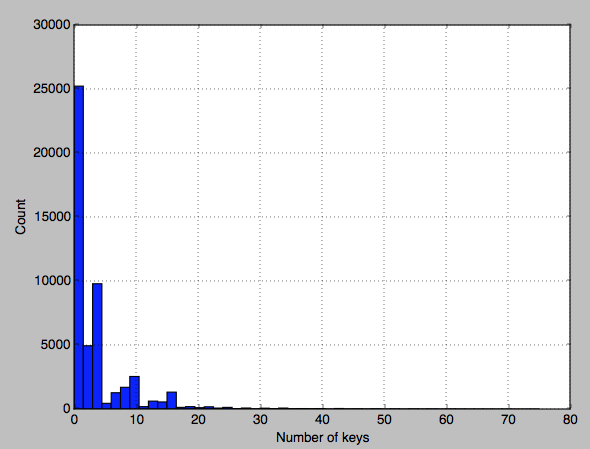
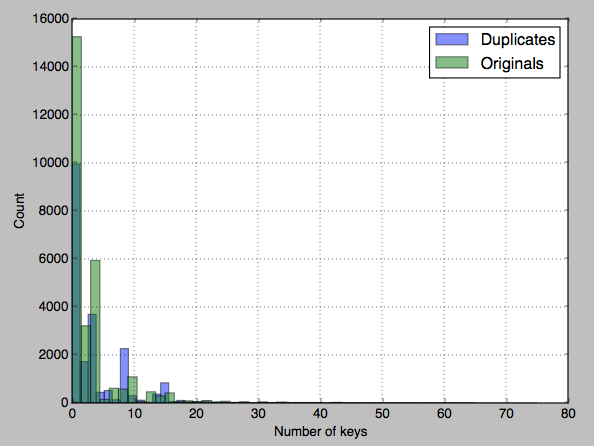
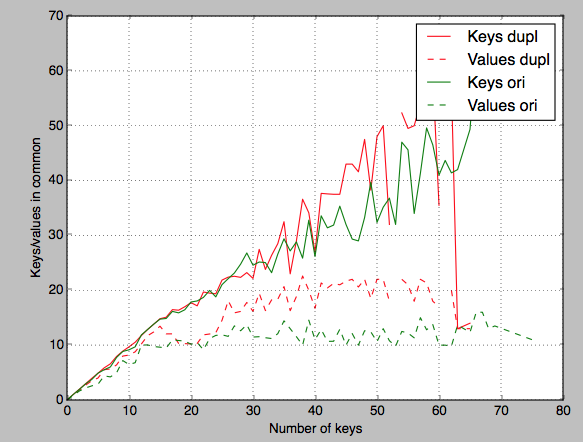
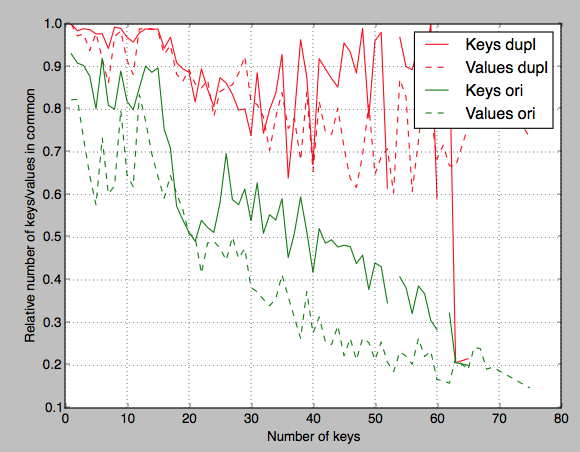
Lat:

* Same as lon

Images array

* 13% in duplicates same, 3% in non-duplicates have exactly the same images
* 66% in duplicates same number, 54% in non-duplicates same number
* 
* When looking at image length:
  + Only duplicates found for one image
    - 30% of duplicates same image
    - 3% or originals same image
  + Ads with more images always have completely different images
    - With
* Conclusion:
  + Include number of images
  + Include same image if images = 1
* TO DO:
  + Compare images itself

attrsJSON

* 73% of duplicates same, vs 65% of originals
* 176 different keys
* Most have a very small number of keys
* 50% has one key
* 5% has no keys
* 
* 
* Duplicates have on average 4 keys and 3.4 values in common
* Originals have on average 3.2 keys and 2.3 values in common
* More keys and large difference is good signal for duplicates. However, 10% has more than 10 keys and 5% has more than 15 keys
  + 
  + 
* To do: specific keys for keywords