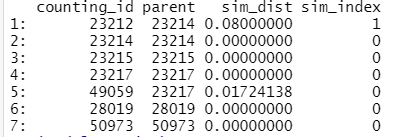
Die Classification ist nun fertig und zur Übersichtlichkeit zweiteilig:

* 1. Im ersten Schritt bestimmen wir, wie gehabt, die similiarity der individuellen Listings (bzw. deren unique-combinations) auf Ebene der exakten Koordinate. Dazu wird zunächst iterativ die Euklidische Distanz[[1]](#footnote-1) zwischen den Listings und allen anderen bestimmt. Das Ganze wird skaliert um das aktuell betrachtete Listing, sodass die Distanz als prozentuale Abweichung interpretiert werden kann (similarity distance).[[2]](#footnote-2)
  2. In einem parallelen Schritt wird dann der ursprünglich abgesprochene similarity index bestimmt, welcher angibt, ob die Werte der anderen Listings (child) innerhalb eines Intervalls um das betrachtete Listing (parent) liegen. Es gibt drei Ausprägungen:
     + 0 = exact match (<=5% Abweichung der Wohnflaeche)
     + 1 = resembling match ( <=10% Abweichung der Wohnflaeche)
     + NA = weder noch

Mithilfe dieser beiden Informationen gilt es nun zwei Probleme zu bewältigen. Zunächst muss bestimmt werden, welches child zu welchem parent gehört. Dieser Schritt ist relativ straightforward mit den Informationen von oben, da jedes child mit einem Index von 0 oder 1 zu dem entsprechenden parent gehört. Die Herausforderung an dieser Stelle sind überlappende claims, also wenn mehrere parents sagen, dass dasselbe child zu ihnen gehört oder eine parent-child Kombination in beide Richtungen funktionieren würde (Bspw. Wohnfläche 50 und 53 qm, Rest gleich, sind jeweils im Intervall voneinander). Um sich hier für einen Kombination zu entscheiden, benutzte ich die aktuell den minimalen Durchschnitt der similarity distance der parents. Das lässt sich ungefähr als ols-fit interpretieren, da die Summe der Euklidischen Distanz minimiert wird.[[3]](#footnote-3) Anschließend muss noch die Entscheidung getroffen werden, ob ein jeweiliges Listing eher child oder parent sein soll, insofern der oben beschriebene Prozess beides als Ergebnis hat. In diesem Fall wähle ich, ähnlich zu oben, den besten fit, also die minimale similarity distance. Das sorgt konkret dafür, dass insofern ein parent als solcher bestimmt wurde auch immer als solcher klassifiziert wird und nicht etwa als child eines anderen parents.[[4]](#footnote-4)

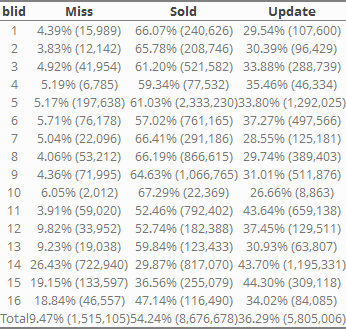
Als Ergebnis dieses Prozesses haben wir nun zu jedem Listing auf der Koordinate eine Zuordnung zu einem parent (ggf. das Listing selber, insofern es nicht zu einem parent zu geordnet werden konnte), samt similarity index und distance zu dem gewählten parent. Beispiel:



Daraufhin kann nun die Panel-Struktur generiert werden, da wir nun einheitliche IDs für die zusammengehörigen Listings haben. Dazu wird eine Art zweite Classification vorgenommen, da es nun gilt zu identifizieren, welche Listings lediglich Updates und welche tatsächlich realistischerweise verkauft worden sind. Dazu berechne ich den zeitlichen Abstand zum jeweils darauffolgenden Listing (sortiert nach Anfangsjahr/-monat). Für den Fall, dass es kein drauffolgendes Listing gibt, wird das Enddatum der Daten substituiert. Basierend auf diesem zeitlichen Abstand gibt es nun drei Ausprägungen (time\_offset ist aktuell testweise bei 3 Monaten für WM):

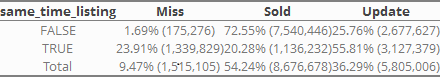
1. Abstand < 0 = „Miss“ [Fehlklassifizierung]
2. Abstand >=0 & <= time\_offset = „Update“
3. Abstand > time\_offset = „Sold“

Anhand von diesen Ausprägungen lassen sich dann die relevanten/non-update Listings identifizieren und verknüpfen, sodass die tatsächliche Listing-Duration bestimmt werden kann. Hier das Ergebnis:

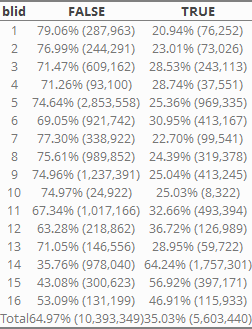


Kurioserweise entsteht ein sehr deutlicher Ost/West-Unterschied, wofür ich zwar eine Erklärung finden konnte, aber keinerlei Verständnis habe. Dazu habe ich mir die same\_time\_listings angeschaut, also Listings mit identischem Anfangsjahr/-monat. Diese sind meiner Einschätzung nach unser größtes verbleibendes Problem, da eine akkurate Klassifizierung dieser Listings mit der jetzigen Methodik nicht möglich ist. Ein same\_time\_listing tritt beispielsweise durch die zeitgleiche Fertigstellung (Bau oder Sanierung?) von Apartments innerhalb eines Hauses auf. Diese Apartments werden dann anschießend innerhalb desselben Monats auf Immobilienscout gelistet und einzeln vermietet. Daher haben wir ein eindeutiges Indiz, dass es sich um verschiedene Wohnung handeln muss, welches allerdings in unsere Methodik unberücksichtigt bleibt. Dies ist insbesondere problematisch, da eben diese same\_time\_listings über 85% unserer Fehlklassifizierung verursachen:

[Erste Spalte: 1.3/1.5 = 0.86]



Ferner liegt an diesem Punkt auch die Ursache für den oben erwähnten Ost/West-Unterschied:



Die Zahl der same\_time\_listings ist aus irgendeinem Grund (hier liegt mein Verständnisproblem) in Ost-Deutschland teilweise doppelt so hoch wie im Westen mit Werten von um die 50%. Da diese same\_time\_listings naturgemäß maximal als Updates, regulärer Weise aber als Miss klassifiziert werden erklärt sich hier durch auch die hohe Missquote im Osten. Den Umgang mit diesem Phänomen sollten wir defintiv nochmal besprechen.

Ein paar weitere Summary-Statistics findest du unter:

„N:\FDZ\Intern\HiWi-Praktikanten\Mitarbeiter\Thorben\repeated offerings\output\WM\v8\2023\_07\_12“

1. Bin hier noch am Lesen ob es andere/bessere Alternativen gibt. [↑](#footnote-ref-1)
2. Die Etage wird für diese Berechnung um jeweils eins erhöht, da eine Skalierung der Etage 0 nicht sinnvoll ist. Ist ein bisschen vergleichbar mit der Berechnung eines logs. [↑](#footnote-ref-2)
3. Hier gibt es Edgecases, wo ein child, was ursprünglich zu einem Kandidaten-clusters gehörte, nicht in dem Intervall des gewählten parents liegt. Das sind aktuell c.a. 0.03% des Datensatzes (siehe blid\_sim\_index.png). [↑](#footnote-ref-3)
4. Hier wäre eine komplizierte Lösung denkbar, wo die jeweiligen Kosten gegeneinander abgewogen werden. Das würde ich demnächst mal ausprobieren. [↑](#footnote-ref-4)