Repeated sales/rent walkthrough

Introduction outlining the idea and process..

Details of reading/reprocessing the RED data can be found in R/read__/, specifically read_RED.R and prepare_RED.R. These should be self explanatory. Note that the classification algorithm is designed to be run in parallel.[1] This is achieved by grouping the RED data on "blid", i.e. one group for each federal state. These groups are than classified in parallel for significant speedups. This classification digs down to coordinate-level, which is where i will start explaining the actual procedure.

Also dropped filtered and dropped balkon— The example data is taken from a coordinate with 25 observations of federal state Bremen. Since the data is already subset to the required dissolution for the next steps, the variables "blid" and "latlon_utm" were dropped beforehand (See R/misc/make_example_markdown_data.R for details).

```
## load example data
tar_load(example_markdown_data)
```

#General makeup of the data:

```
# show head
head(example_markdown_data)
```

```
##    wohnflaeche zimmeranzahl etage counting_id amonths emonths price_var
## 1:          71            3     2     1026832   24086   24088    135200
## 2:          56            2     1     1026833   24086   24088    102300
## 3:          76            3     6     1026834   24086   24089    139900
## 4:          71            3     2     1027667   24089   24100    142900
## 5:          56            2     1     1027668   24089   24100    107800
## 6:          82            3     2     1029808   24095   24097    164900
```

```
# show summary
summary(example_markdown_data)
```

```
##   wohnflaeche     zimmeranzahl      etage       counting_id
##  Min.   : 56.0   Min.   :2.00   Min.   :0.00   Min.   :1026832
##  1st Qu.: 71.0   1st Qu.:3.00   1st Qu.:1.00   1st Qu.:1030454
##  Median : 80.0   Median :3.00   Median :2.00   Median :1032333
##  Mean   : 80.0   Mean   :2.96   Mean   :2.24   Mean   :1033371
##  3rd Qu.: 81.0   3rd Qu.:3.00   3rd Qu.:3.00   3rd Qu.:1034542
##  Max.   :149.9   Max.   :5.00   Max.   :6.00   Max.   :1045696
##     amonths         emonths        price_var
##  Min.   :24086   Min.   :24088   Min.   :102300
##  1st Qu.:24097   1st Qu.:24100   1st Qu.:138000
##  Median :24102   Median :24104   Median :139900
##  Mean   :24104   Mean   :24108   Mean   :155220
##  3rd Qu.:24107   3rd Qu.:24117   3rd Qu.:155000
##  Max.   :24134   Max.   :24135   Max.   :335700
```

[1] I recommend using 'tar_make_future(workers = n)', where n is the number of cores. $n = 4$ is a decent value, when no one else is using the server the number can be set higher. More than 8 is overkill, since Berlin alone typically bottlenecks. If speed becomes a big issue with growing data consider using "foreach()" on the coordinate level.

There are two types of similarity which will be considered: resembling and exact. The former allows for slightly larger deviations, the latter is quite restrictive. "r_o" refers to resembeling offset and "e_o" to exact offset.

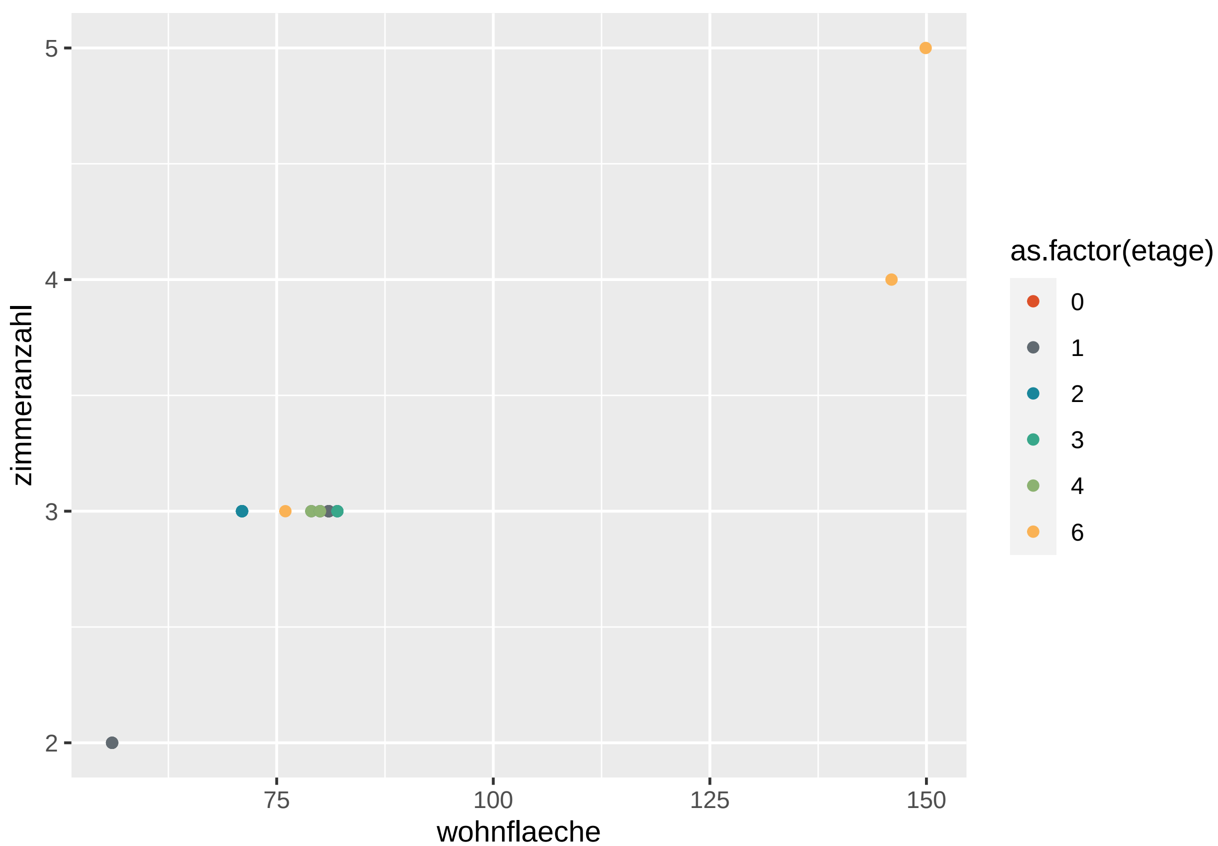
parameter used for classification:

```
# these are globally defined in _targets.R
print(exportJSON)
```

```
##    RED_type RED_version   categories wohnflaeche_r_o etage_r_o zimmeranzahl_r_o
## 1:       WK          v9  wohnflaeche             0.1         1              0.5
## 2:       WK          v9        etage             0.1         1              0.5
## 3:       WK          v9 zimmeranzahl             0.1         1              0.5
##    wohnflaeche_e_o zimmeranzahl_e_o time_offset
## 1:            0.05              0.5           6
## 2:            0.05              0.5           6
## 3:            0.05              0.5           6
```

```
# make color blind friendly palette
etage_colors = MetBrewer::met.brewer("Egypt", n = uniqueN(example_markdown_data$etage))

ggplot(example_markdown_data, aes(x  = wohnflaeche, y = zimmeranzahl, color = as.factor(etage)))+
  geom_point() +
  scale_color_manual(values = etage_colors)
```



Note that direct overlaps naturally arent visible. We can see that most of the combinations are fairly distinct with one notable exception: Etage 4 (green) has one near perfect match (etage and zimmeranzahl are the same) with slight deviations in wohnflaeche. The challenge now is to make a decision: are the lisings for two different apartments or are they the same apartment? To make this decision for any number of characteristic combinations, I use a modified version of k-nearest neighbors clustering. Before we can get to this stage however, I it is necessary to define and approach the issue formally. Visually classifying each combination at each coordinate in Germany would take quite some time (and would have to be re-done for every new wave, more on that later).

We will proceed in two overarching dimensions, each involving quite a few steps: the characteristics dimension (classifying similarity) and the subsequent time dimension ('classifying' non list reason):

characteristics dimension

```
# run tar_load_globals() if R cant find this function
out = similarity_classification(example_markdown_data)
head(out)
```

```
##    wohnflaeche zimmeranzahl etage counting_id amonths emonths price_var  parent
## 1:          71            3     2     1026832   24086   24088    135200 1026832
## 2:          71            3     2     1027667   24089   24100    142900 1026832
## 3:          71            3     2     1031914   24101   24106    139900 1026832
## 4:          71            3     2     1034541   24107   24115    139900 1026832
## 5:          71            3     2     1038183   24116   24127    136700 1026832
## 6:          56            2     1     1026833   24086   24088    102300 1026833
##    sim_dist sim_index
## 1:        0         0
## 2:        0         0
## 3:        0         0
## 4:        0         0
## 5:        0         0
## 6:        0         0
```