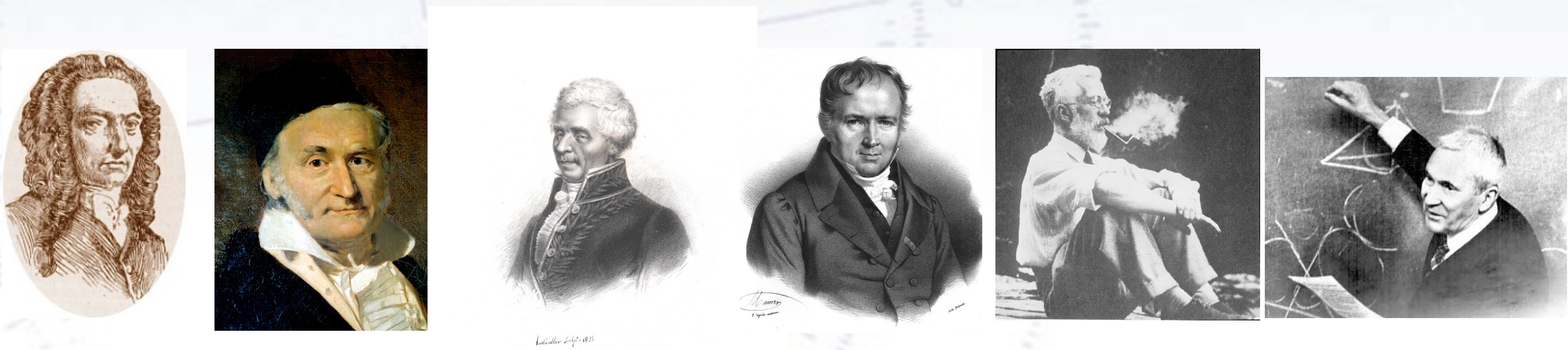
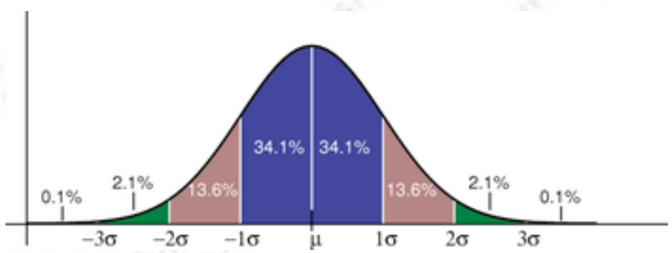
A simple ML example

Data set: Housing Prices



Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense - Machine Learning is a sharpening of it!"

Data, goal, and misc.

The data:
About **50.000 real estate sales**, including the final sales price along with several descriptive variables, many incomplete or missing.

The goal:
To determine the final sales price as accurately as possible.
NOTE: “As accurately” is not a well determined measure, and we will discuss this.

Miscellaneous:
While the dataset is on the border of “Big Data”, we have chosen it, as it fits all the ML methods well, and since its analysis can be **done in finite time**.

Dataset variables - 90 in total

0 MI_OBJ_OIS_PROPERTY_ID
1 MI_OBJ_OIS_PROPERTY_NUMBER
2 MI_OBJ_OIS_MOTHER_ID
3 MI_OBJ_OIS_MUNICIPALITY_NUMBER
4 MI_OBJ_OIS_POSTAL_CODE
5 MI_OBJ_OIS_RENTED_PLOT
6 MI_OBJ_OIS_OWNERSHIP_CODE_PROPERTY
7 MI_OBJ_OIS_OWNERSHIP_CODE_UNIT
8 MI_OBJ_OIS_PROPERTY_APPLICATION_CODE_UNIT
9 MI_OBJ_OIS_PROPERTY_APPLICATION_CODE_BUILDING
10 MI_OBJ_OIS_PROPERTY_USE_CODE
11 MI_OBJ_OIS_SALES_PRICE
12 MI_OBJ_OIS_DATE_OF_SALES_PRICE
13 MI_OBJ_OIS_PREVIOUS_SALES_PRICE_FIRST
14 MI_OBJ_OIS_DATE_OF_PREVIOUS_SALES_PRICE_FIRST
15 MI_OBJ_OIS_PREVIOUS_SALES_PRICE_SECOND
16 MI_OBJ_OIS_DATE_OF_PREVIOUS_SALES_PRICE_SECOND
17 MI_OBJ_OIS_PREVIOUS_SALES_PRICE_THIRD
18 MI_OBJ_OIS_DATE_OF_PREVIOUS_SALES_PRICE_THIRD
19 MI_OBJ_OIS_PREVIOUS_SALES_PRICE_FOURTH
20 MI_OBJ_OIS_DATE_OF_PREVIOUS_SALES_PRICE_FOURTH
21 MI_OBJ_OIS_TAXATION_VALUE
22 MI_OBJ_OIS_TAXATION_VALUE_PLOT
23 MI_OBJ_OIS_TAXATION_VALUE_FARMHOUSE
24 MI_OBJ_OIS_DATE_OF_TAXATION_VALUE
25 MI_OBJ_OIS_PROPERTY_ADDRESS
26 MI_OBJ_OIS_HOUSE_NUMBER
27 MI_OBJ_OIS_HOUSE_LETTER
28 MI_OBJ_OIS_DOOR_CODE
29 MI_OBJ_OIS_FLOOR_NUMBER
30 MI_OBJ_OIS_MAX_FLOOR_NUMBER_BUILDING
31 MI_OBJ_OIS_LAND_ZONE
32 MI_OBJ_OIS_SIZE_OF_HOUSE
33 MI_OBJ_OIS_SIZE_OF_BUSINESS_AREA
34 MI_OBJ_OIS_SIZE_OF_PLOT
35 MI_OBJ_OIS_SIZE_OF_INTEGRATED_CARPORT
36 MI_OBJ_OIS_SIZE_OF_NOT_INTEGRATED_CARPORT
37 MI_OBJ_OIS_SIZE_OF_OUTDOOR_LIVING_ROOM
38 MI_OBJ_OIS_SIZE_OF_INTEGRATED_OUTHOUSE
39 MI_OBJ_OIS_SIZE_OF_INTEGRATED_GARAGE
40 MI_OBJ_OIS_SIZE_OF_LEGAL_BASEMENT
41 MI_OBJ_OIS_SIZE_OF_BASEMENT
42 MI_OBJ_OIS_SIZE_OF_ATTIC
43 MI_OBJ_OIS_SIZE_OF_USED_ATTIC
44 MI_OBJ_OIS_SIZE_OF_HOUSE_EXCL_UTILIZED_ATTIC
45 MI_OBJ_OIS_SIZE_OF_BUSINESS_AREA_BUILDING
46 MI_OBJ_OIS_SIZE_OF_NOT_INTEGRATED_GARAGE
47 MI_OBJ_OIS_NUMBER_OF_FLOORS
48 MI_OBJ_OIS_CONSTRUCTION_YEAR
49 MI_OBJ_OIS_CONSTRUCTION_MATERIAL
50 MI_OBJ_OIS_REBUILD_YEAR
51 MI_OBJ_OIS_ROOF_MATERIAL
52 MI_KNN_PROPERTY_CONDITION
53 MI_KNN_TOP_FLOOR_INDICATOR
54 MI_KNN_GROUND_FLOOR_INDICATOR
55 MI_KNN_GROUP_VALID_REGRESSION_INPUT
56 MI_KNN_GRP_PERCENTILE_MIN_WEIGHTED_SIZE_OF_HOUSE
57 MI_KNN_GROUP_PERCENTILE_MIN_SIZE_OF_PLOT
58 MI_KNN_GROUP_PERCENTILE_MIN_CONSTRUCTION_YEAR
59 MI_KNN_GROUP_PERCENTILE_MIN_TAXATION_VALUE
60 MI_KNN_GROUP_PERCENTILE_MIN_TAXATION_VALUE_PLOT
61 MI_KNN_GRP_PERCENTILE_MAX_WEIGHTED_SIZE_OF_HOUSE
62 MI_KNN_GROUP_PERCENTILE_MAX_SIZE_OF_PLOT
63 MI_KNN_GROUP_PERCENTILE_MAX_TAXATION_VALUE
64 MI_KNN_GROUP_PERCENTILE_MAX_TAXATION_VALUE_PLOT
65 MI_KNN_M2_P_PREDIC
66 MI_KNN_STD_SALES_PRICE_NEIGHBORS
67 MI_KNN_AVG_GEO_DISTANCE_NEIGHBORS
68 MI_KNN_AVG_CONSTRUCTION_YEAR_NEIGHBORS
69 MI_KNN_AVG_WEIGHTED_SIZE_OF_HOUSE_NEIGHBORS
70 MI_KNN_AVG_SIZE_OF_PLOT_NEIGHBORS
71 MI_KNN_APARTMENTS_NEIGHBORS_INDICATOR
72 MI_KNN_MATERIAL_TYPE
73 MI_KNN_APARTMENTS_ACTUAL_NUM_OF_NEIGHBORS
74 MI_KNN_STATUS
75 MI_OBJ_NUMBER_OF_EXTERNAL_MATRS
76 MI_OBJ_OIS_SUM_OF_TAXATION_VALUES
77 MI_OBJ_OIS_N_COORDINATE
78 MI_OBJ_OIS_E_COORDINATE
79 C20_1MONTH%
80 C20_3MONTH%
81 C20_6MONTH%
82 C20_12MONTH%
83 SCHOOL_DISTANCE_1
84 SCHOOL_DISTANCE_2
85 SCHOOL_DISTANCE_3
86 SUPERMARKET_DISTANCE_1
87 SUPERMARKET_DISTANCE_2
88 SUPERMARKET_DISTANCE_3
89 KOEBESUM_BELOEB

Dataset variables - 90 in total

0 MI_OBJ_OIS_PROPERTY_ID
1 MI_OBJ_OIS_PROPERTY_NUMBER
2 MI_OBJ_OIS_MOTHER_ID
3 MI_OBJ_OIS_MUNICIPALITY_NUMBER
4 MI_OBJ_OIS_POSTAL_CODE
5 MI_OBJ_OIS_RENTED_PLOT
6 MI_OBJ_OIS_OWNERSHIP_CODE_PROPERTY
7 MI_OBJ_OIS_OWNERSHIP_CODE_UNIT
8 MI_OBJ_OIS_PROPERTY_APPLICATION_CODE_UNIT
9 MI_OBJ_OIS_PROPERTY_APPLICATION_
10 MI_OBJ_OIS_PROPERTY_USE_CODE
11 MI_OBJ_OIS_SALES_PRICE
12 MI_OBJ_OIS_DATE_OF_SALES_PRICE
13 MI_OBJ_OIS_PREVIOUS_SALES_PRICE
14 MI_OBJ_OIS_DATE_OF_PREVIOUS_SAL
15 MI_OBJ_OIS_PREVIOUS_SALES_PRICE_SECOND
16 MI_OBJ_OIS_DATE_OF_PREVIOUS_SALES_PRICE_SECOND
17 MI_OBJ_OIS_PREVIOUS_SALE
18 MI_OBJ_OIS_DATE_OF_PREVI
19 MI_OBJ_OIS_PREVIOUS_SALE
20 MI_OBJ_OIS_DATE_OF_PREVIOUS_SALES_PRICE_FOURTH
21 MI_OBJ_OIS_TAXATION_VALU
22 MI_OBJ_OIS_TAXATION_VALU
23 MI_OBJ_OIS_TAXATION_VALUE_FARMHOUSE
24 MI_OBJ_OIS_DATE_OF_TAXATION_VALUE
25 MI_OBJ_OIS_PROPERTY_ADDRESS
26 MI_OBJ_OIS_HOUSE_NUMBER
27 MI_OBJ_OIS_HOUSE_LETTER
28 MI_OBJ_OIS_DOOR_CODE
29 MI_OBJ_OIS_FLOOR_NUMBER
30 MI_OBJ_OIS_MAX_FLOOR_NUMBER
31 MI_OBJ_OIS_LAND_ZONE
32 MI_OBJ_OIS_SIZE_OF_HOUSE
33 MI_OBJ_OIS_SIZE_OF_BUSINE
34 MI_OBJ_OIS_SIZE_OF_PLOT
35 MI_OBJ_OIS_SIZE_OF_INTEGRATED_C
36 MI_OBJ_OIS_SIZE_OF_NOT_INTEGRAT
37 MI_OBJ_OIS_SIZE_OF_OUTDOOR_LIVI
38 MI_OBJ_OIS_SIZE_OF_INTEGRATED_OUTHOUSE
39 MI_OBJ_OIS_SIZE_OF_INTEGRATED_
40 MI_OBJ_OIS_SIZE_OF_LEGAL_BASE
41 MI_OBJ_OIS_SIZE_OF_BASEMENT
42 MI_OBJ_OIS_SIZE_OF_ATTIC
43 MI_OBJ_OIS_SIZE_OF_USED_ATTIC
44 MI_OBJ_OIS_SIZE_OF_HOUSE_EXCL_UTILIZED_ATTIC

Postal code
Sales price
Sales date
Taxation value
Taxation value plot
Address
Floor number
Floor number max
Size indoor
Size outdoor

45 MI_OBJ_OIS_SIZE_OF_BUSINESS_AREA_BUILDING
46 MI_OBJ_OIS_SIZE_OF_NOT_INTEGRATED_GARAGE
47 MI_OBJ_OIS_NUMBER_OF_FLOORS
48 MI_OBJ_OIS_CONSTRUCTION_YEAR
49 MI_OBJ_OIS_CONSTRUCTION_MATERI
50 MI_OBJ_OIS_REBUILD_YEAR
51 MI_OBJ_OIS_ROOF_MATERIAL
52 MI_KNN_PROPERTY_CONDITION
53 MI_KNN_TOP_FLOOR_INDICATOR
54 MI_KNN_GROUND_FLOOR_INDICATOR
55 MI_KNN_GROUP_VALID_REGRESSION_INPUT
56 MI_KNN_GRP_PERCENTILE_MIN_WEIGHTED_SIZE_OF_HOUSE
57 MI_KNN_GROUP_PERCENTILE_MIN_SIZE_OF_PLOT
58 MI_KNN_GROUP_PERCENTILE_MIN_CONSTRUCTION_YEAR
59 MI_KNN_GROUP_PERCENTILE_MIN_TAXATION_VALUE
60 MI_KNN_GROUP_PERCENTILE_MIN_TAXATION_VALUE_PLOT
61 MI_KNN_GRP_PERCENTILE_MAX_WEIGHTED_SIZE_OF_HOUSE
62 MI_KNN_GROUP_PERCENTILE_MAX_SIZE_OF_PLOT
63 MI_KNN_GROUP_PERCENTILE_MAX_TAXATION_VALUE
64 MI_KNN_GROUP_PERCENTILE_MAX_TAXATION_VALUE_PLOT
65 MI_KNN_M2_P_PREDIC
66 MI_KNN_STD_SALES_PRICE_NEIGHBORS
67 MI_KNN_AVG_GEO_DISTANCE_NEIGHBORS
68 MI_KNN_AVG_CONSTRUCTION_YEAR_NEIGHBORS
69 MI_KNN_AVG_WEIGHTED_SIZE_OF_HOUSE_NEIGHBORS
70 MI_KNN_AVG_SIZE_OF_PLOT_NEIGHBORS
71 MI_KNN_APARTMENTS_NEIGHBORS_INDICATOR
72 MI_KNN_MATERIAL_TYPE
73 MI_KNN_APARTMENTS_ACTUAL_NUM_OF_NEIGHBORS
74 MI_KNN_STATUS
75 MI_OBJ_NUMBER_OF_EXTERNAL_MATRS
76 MI_OBJ_OIS_SUM_OF_TAXATION_VALUES
77 MI_OBJ_OIS_N_COORDINATE
78 MI_OBJ_OIS_E_COORDINATE
79 C20_1MONTH%
80 C20_3MONTH%
81 C20_6MONTH%
82 C20_12MONTH%
83 SCHOOL_DISTANCE_1
84 SCHOOL_DISTANCE_2
85 SCHOOL_DISTANCE_3
86 SUPERMARKET_DISTANC
87 SUPERMARKET_DISTANCE_2
88 SUPERMARK
89 KOEBESUM_E

Construction year
Property condition
Distance to school
Distance to supermarket
PRICE PAID FOR PROPERTY

Information available

While there are in principle 90 pieces of information on each property sale, it is in practice not the case! As it turns out, most entries are empty!!!

In the figure we consider the most crucial variables (see page before), and check what fraction of entries have information available here.

The conclusions is, that if we wanted all entries filled, we would only have < 1% of data remaining… not a great way forward!

Figure 1(a): Is the info there?

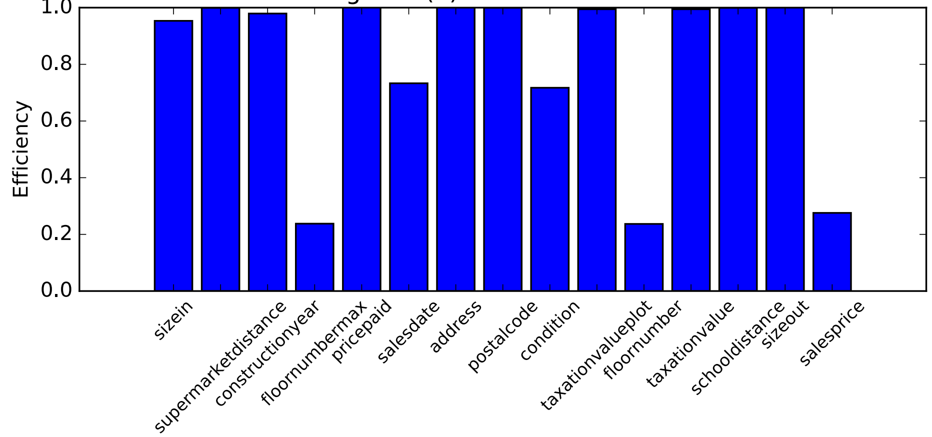


Figure 1(b): How much do we lose from our essentials?

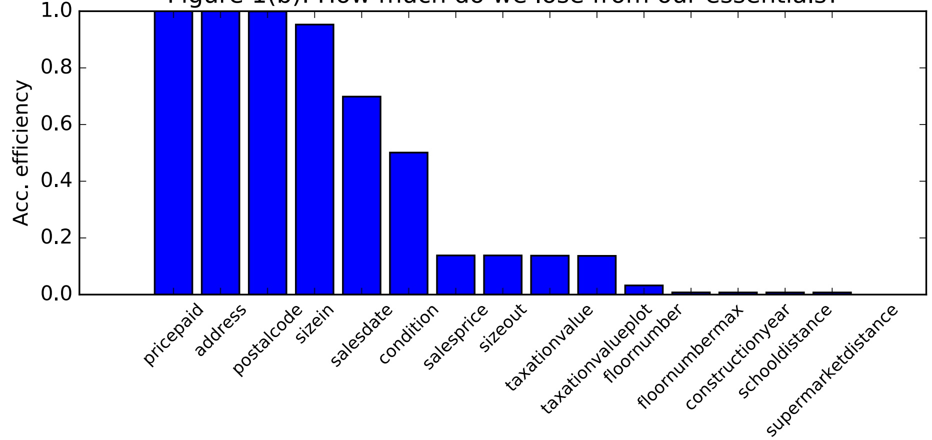
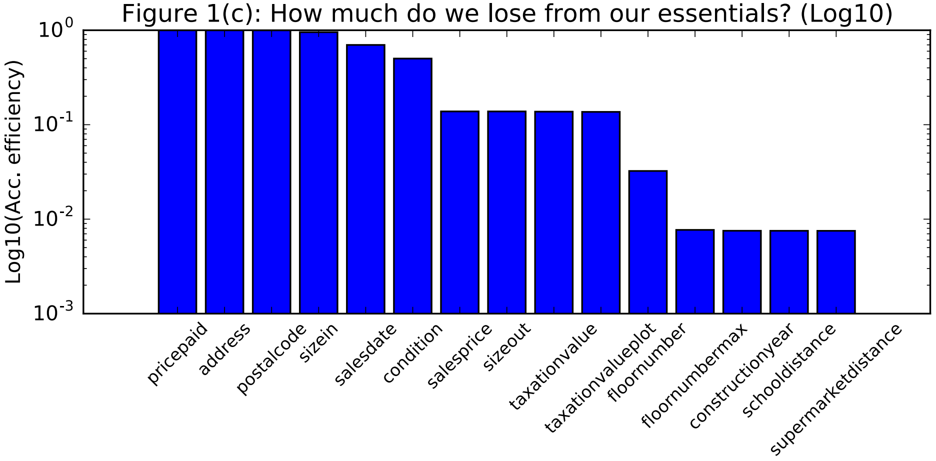


Figure 1(c): How much do we lose from our essentials? (Log10)



Information availal

While there are in princ
of information on each
it is in practice not the c
out, most entries are en

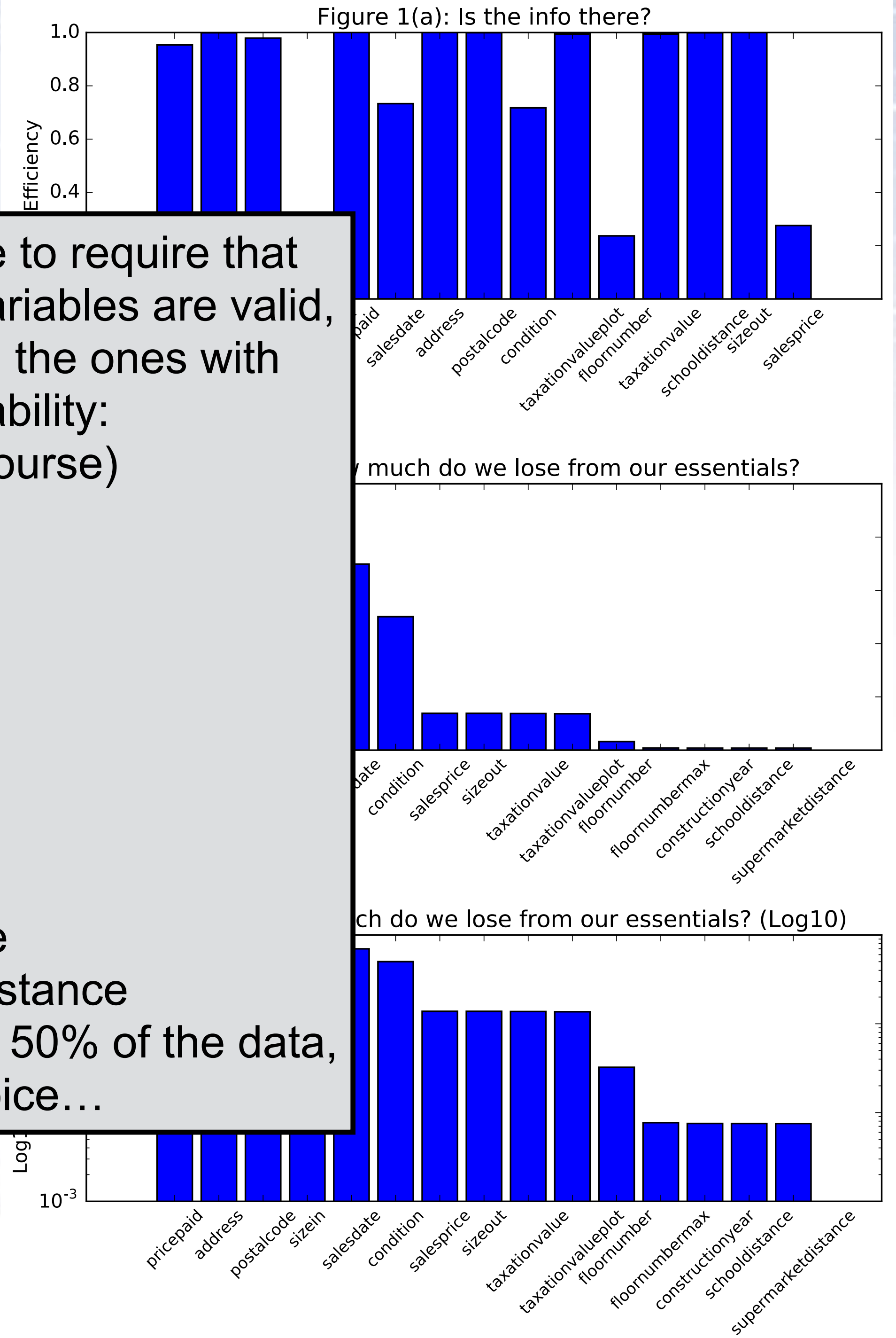
In the figure we conside
crucial variables (see pa
check what fraction of e
information available h

The conclusions is, that
all entries filled, we wo
< 1% of data remaining… not a great
way forward!

One could choose to require that e.g. the first six variables are valid, and then only add the ones with (almost) full availability:

- Price paid (of course)
- Address
- Postal Code
- Size inside
- Sales date
- Condition
- Size outside
- Taxation value
- Floor number
- School distance
- Supermarket distance

This leaves about 50% of the data, which is a fair choice…

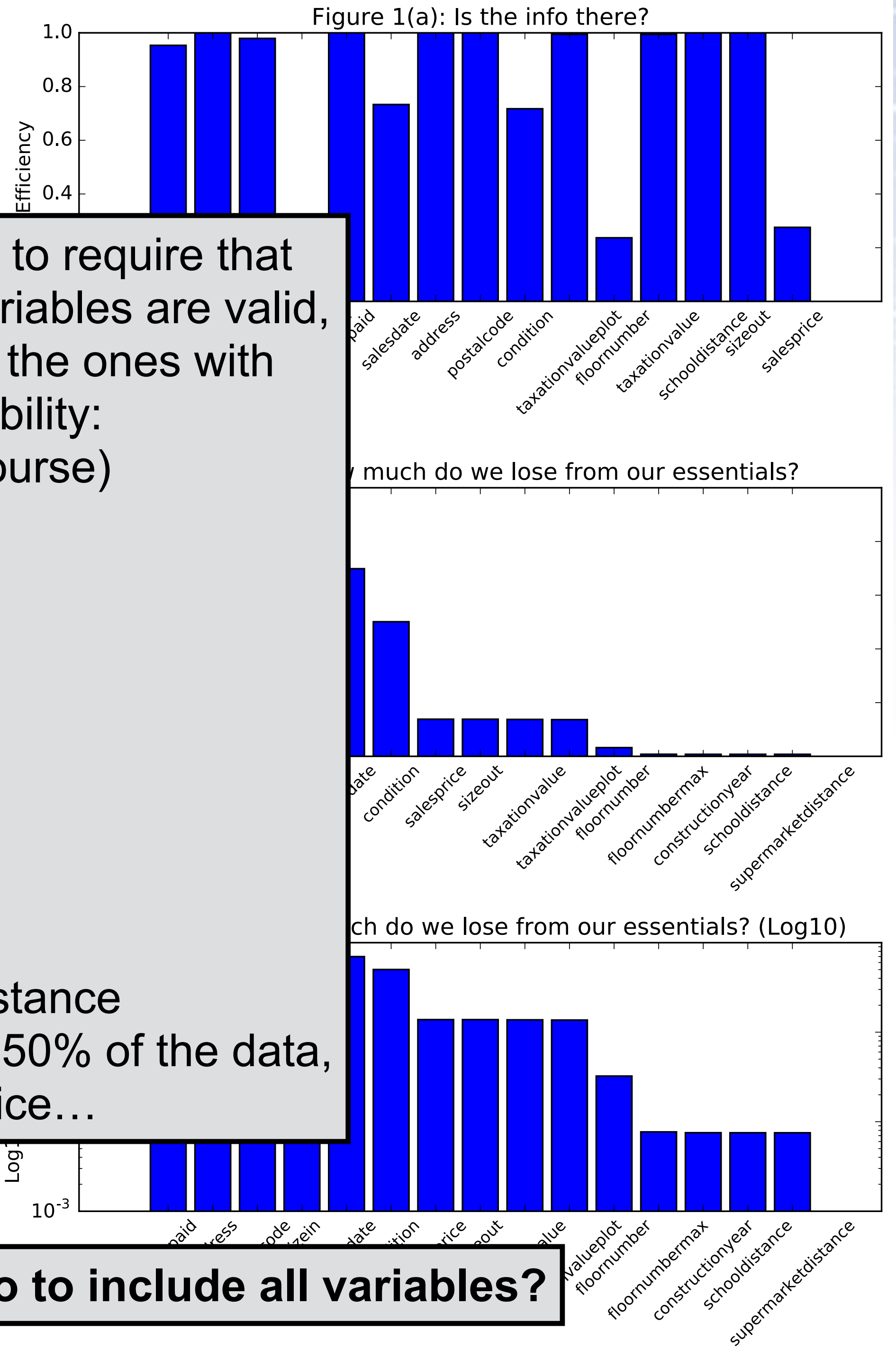


Information availa

While there are in princ
of information on each
it is in practice not the c
out, most entries are em

In the figure we conside
crucial variables (see pa
check what fraction of e
information available h

The conclusions is, that
all entries filled, we wo
< 1% of data remaining… not a great
way forward!



One could choose to require that e.g. the first six variables are valid, and then only add the ones with (almost) full availability:

- Price paid (of course)
- Address
- Postal Code
- Size inside
- Sales date
- Condition
- Size outside
- Taxation value
- Floor number
- School distance
- Supermarket distance

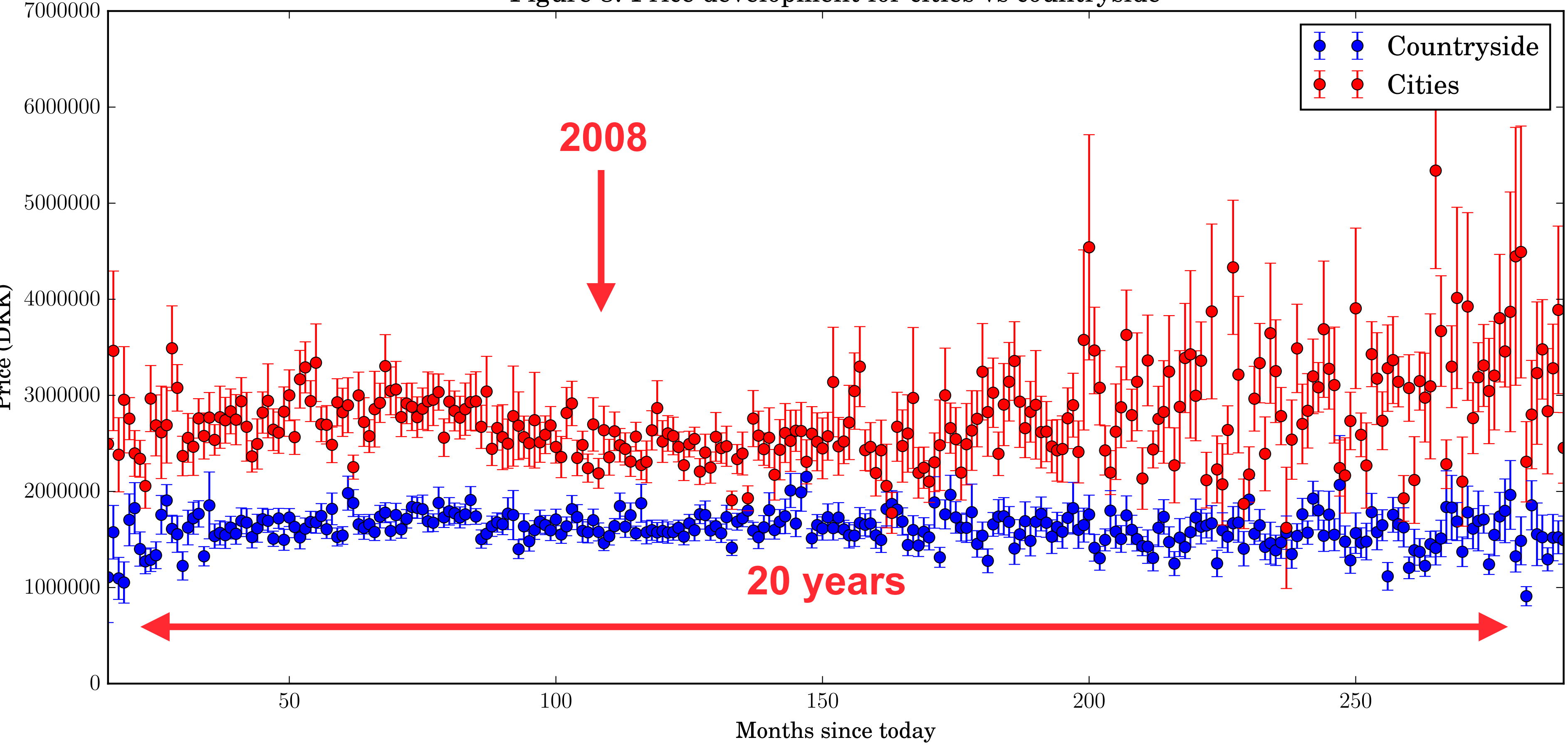
This leaves about 50% of the data, which is a fair choice…

Discuss shortly, what to do to include all variables?

Price vs. time

Just to gauge the data, we try to plot the average price over time:

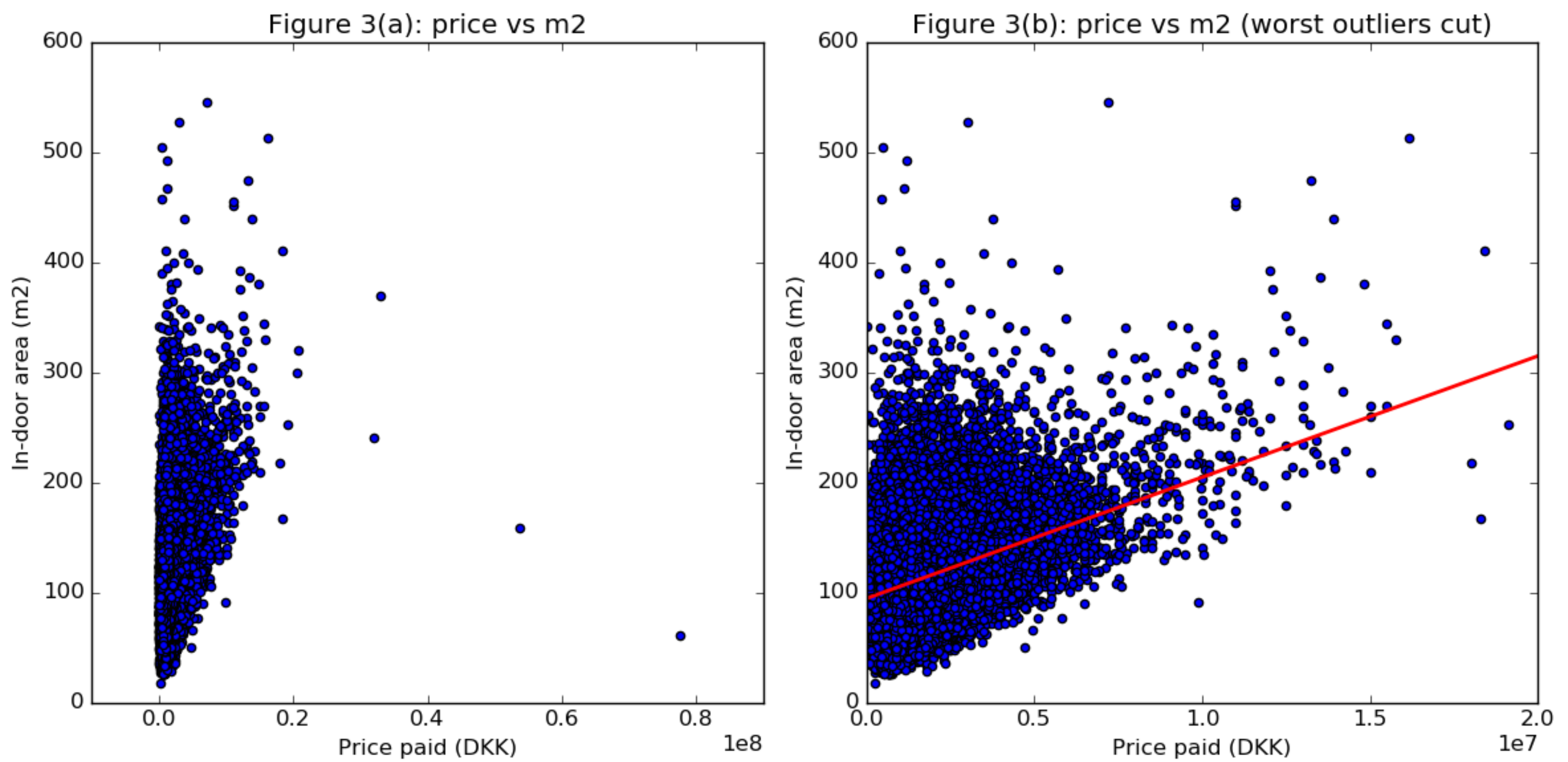
Figure 8: Price development for cities vs countryside



Clearly, the data is corrected for inflation, but not much else, since 2008 doesn't clearly show up.

Price per square meter

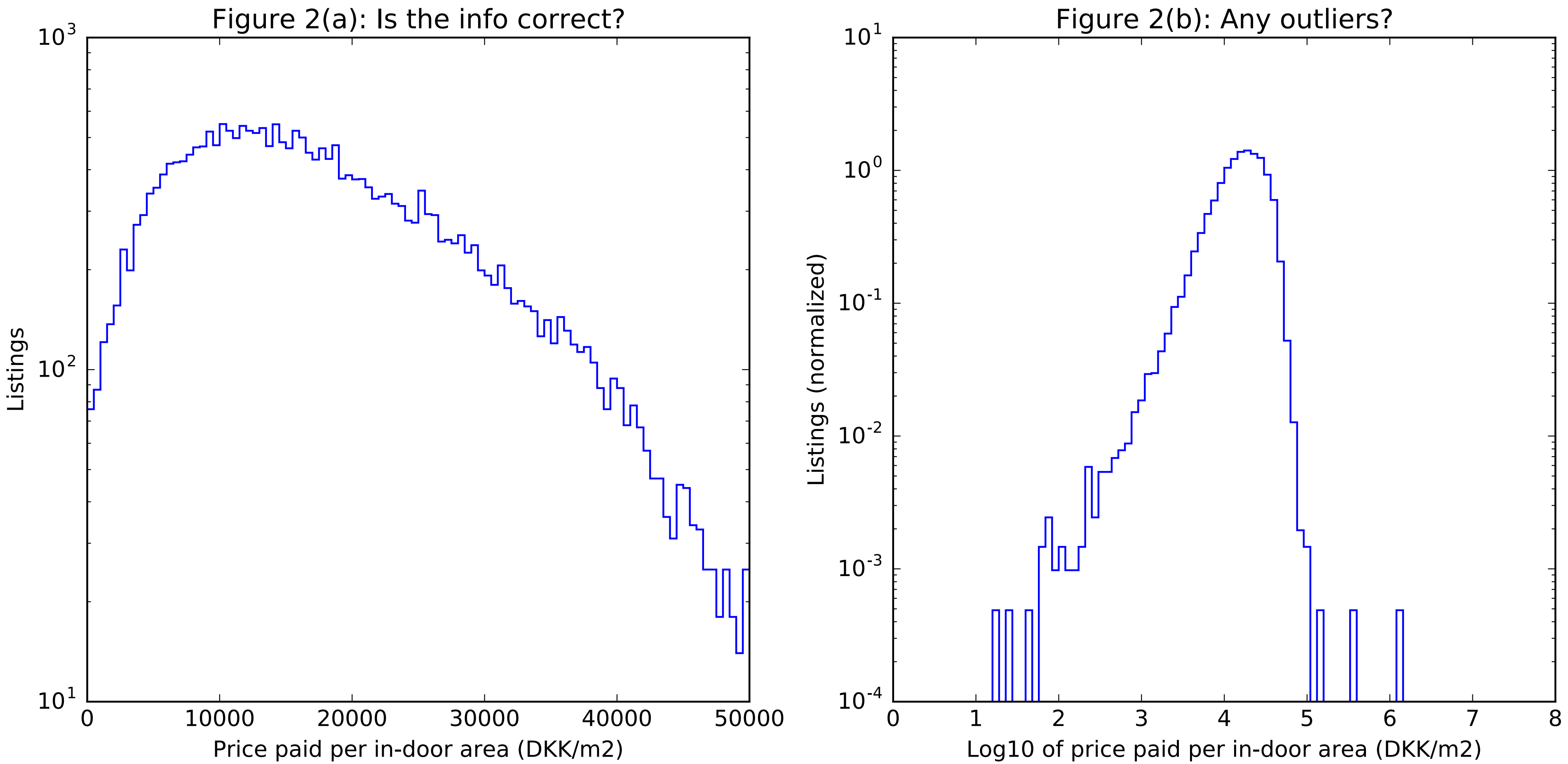
As a first step, one would estimate the price from the size, i.e. assume that the price per square meter was constant, and so we plot price vs. size:



As can be seen from the figure, this does not seem to be the case, and even after filtering away the worst outliers, we don't get any reasonable estimate!

Price per square meter

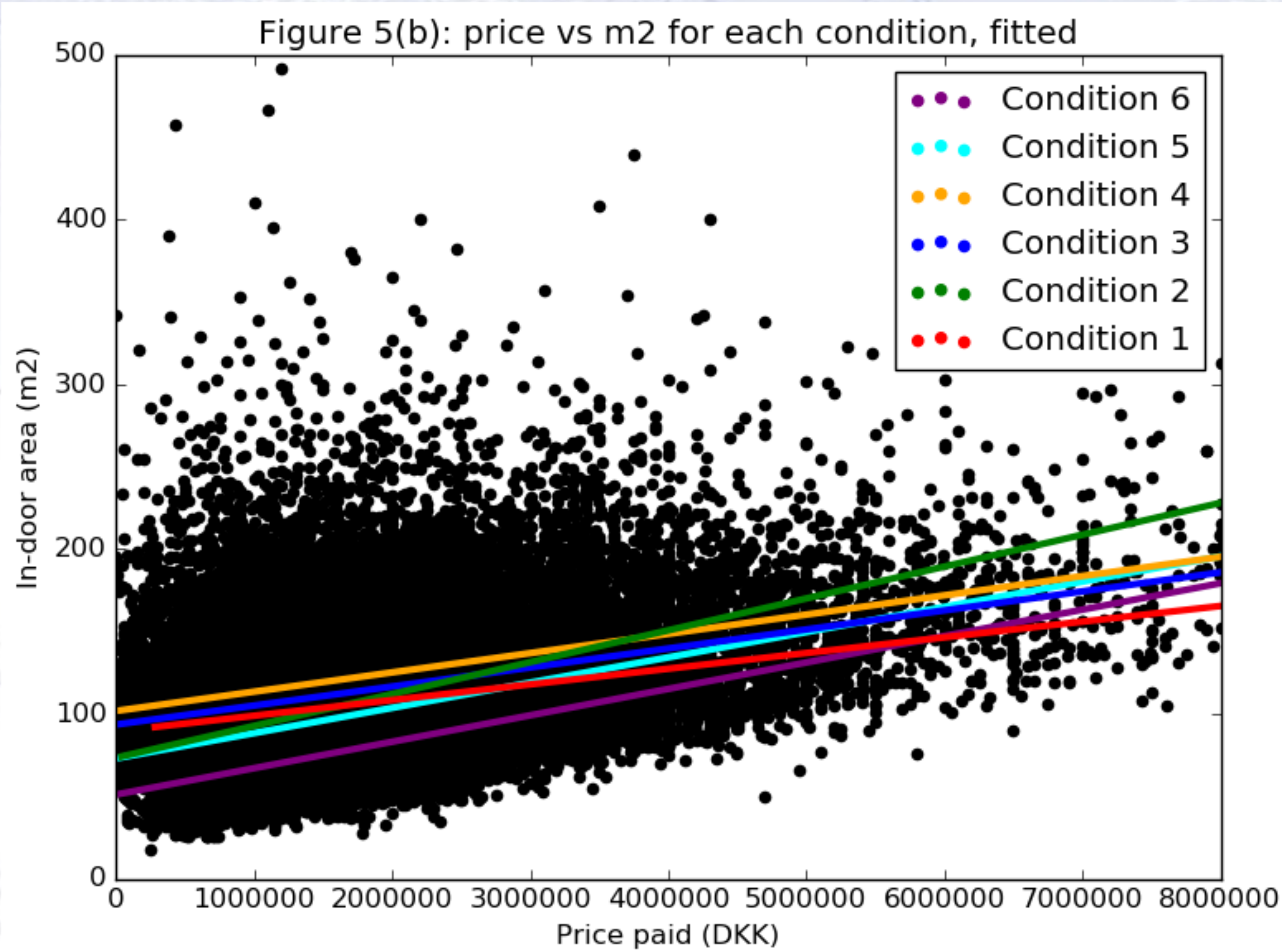
Looking at the price/m2, most values are reasonable, but there are exceptions:



I don't know who paid 1.000.000+ Kr./m2, but that is not a normal value!
Similarly, < 100 Kr./m2 seems odd, and also needs further investigation.

Price per square meter

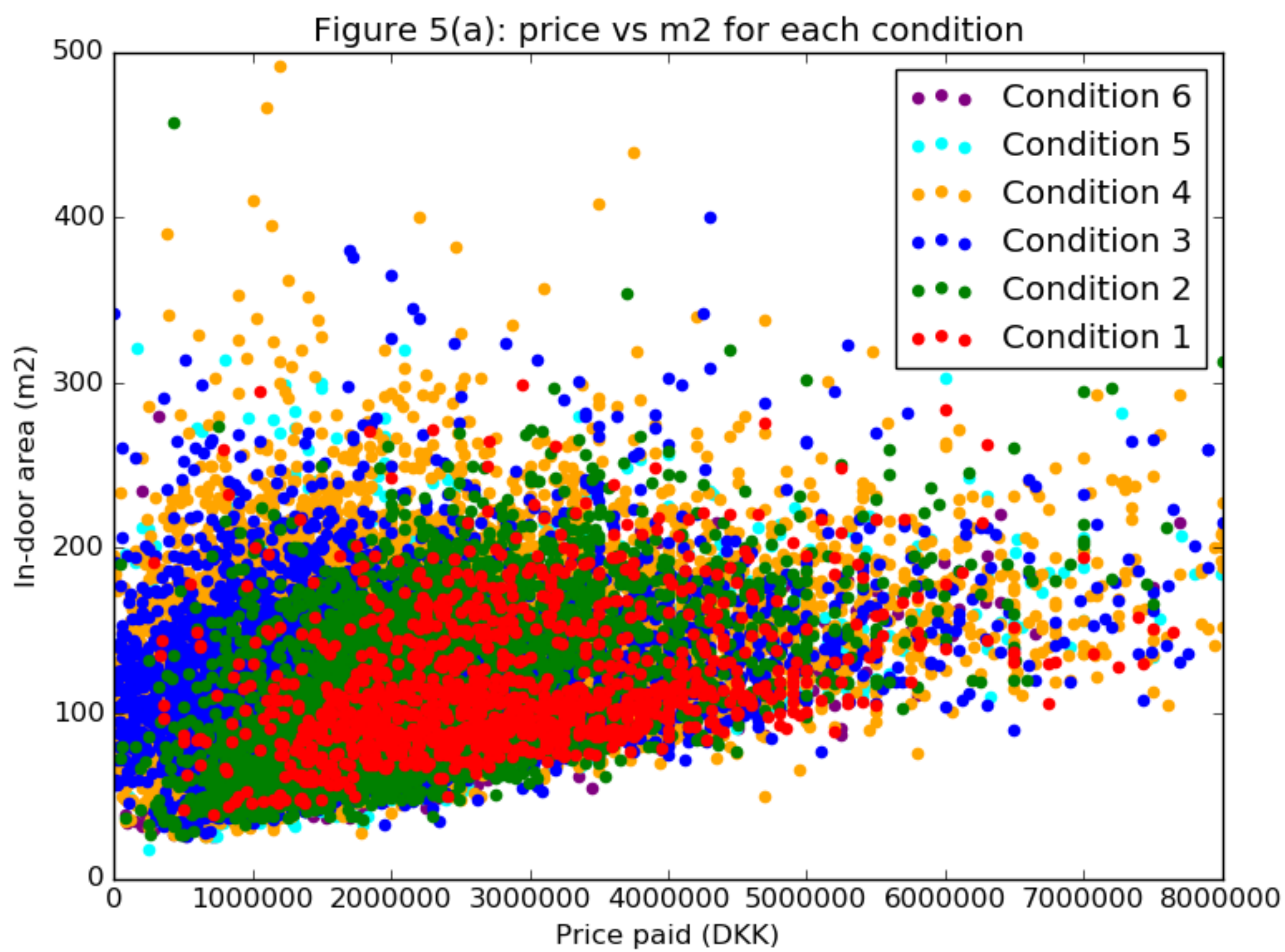
Dividing according to condition, one might expect a higher price/m2, but…



…the pattern is rather, that the basic price is higher!

Price per square meter

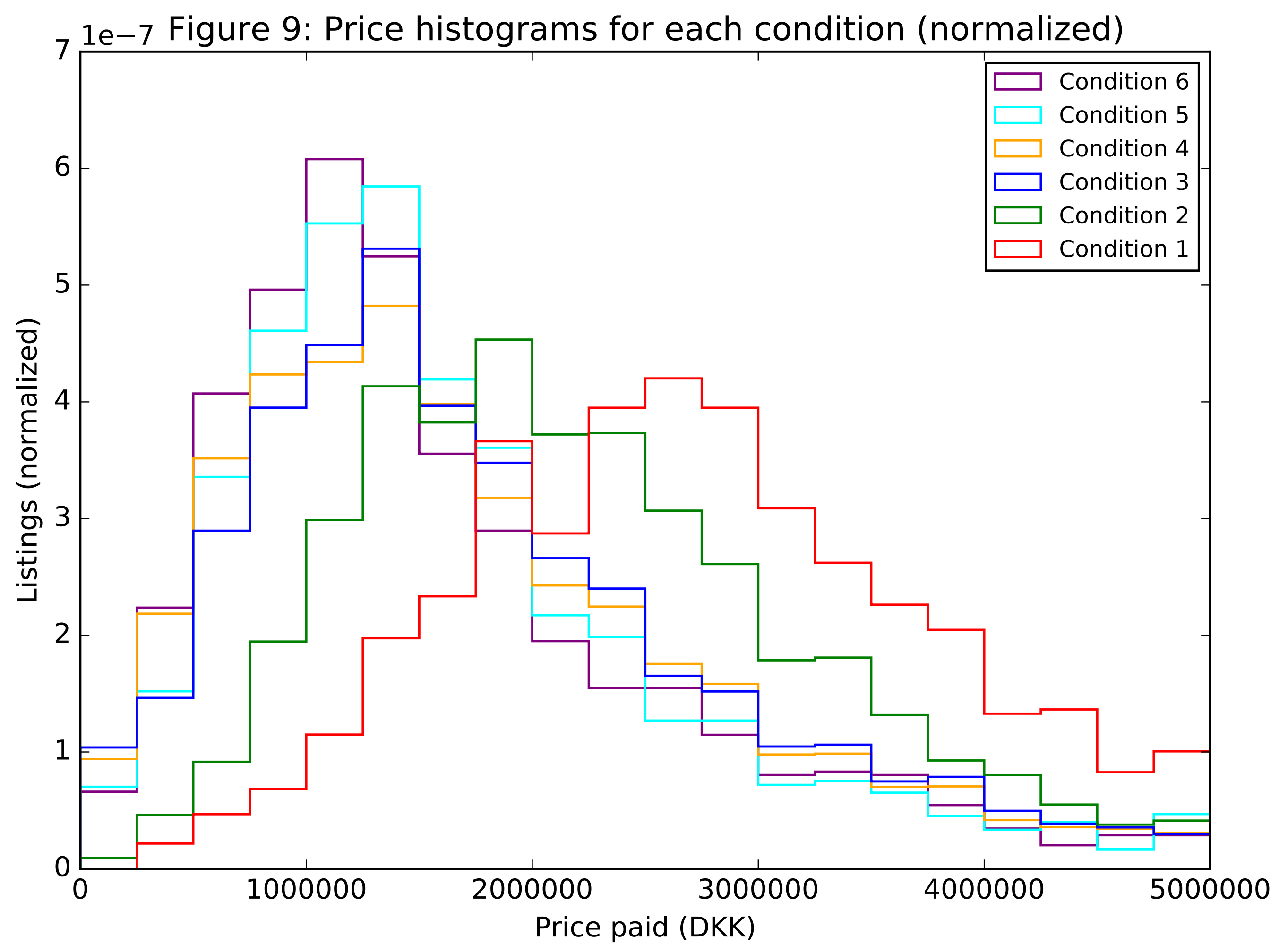
Dividing according to condition, one might expect a higher price/m2, but…



…the pattern is rather, that the basic price is higher! And condition 1 is best!!!

Price per square meter

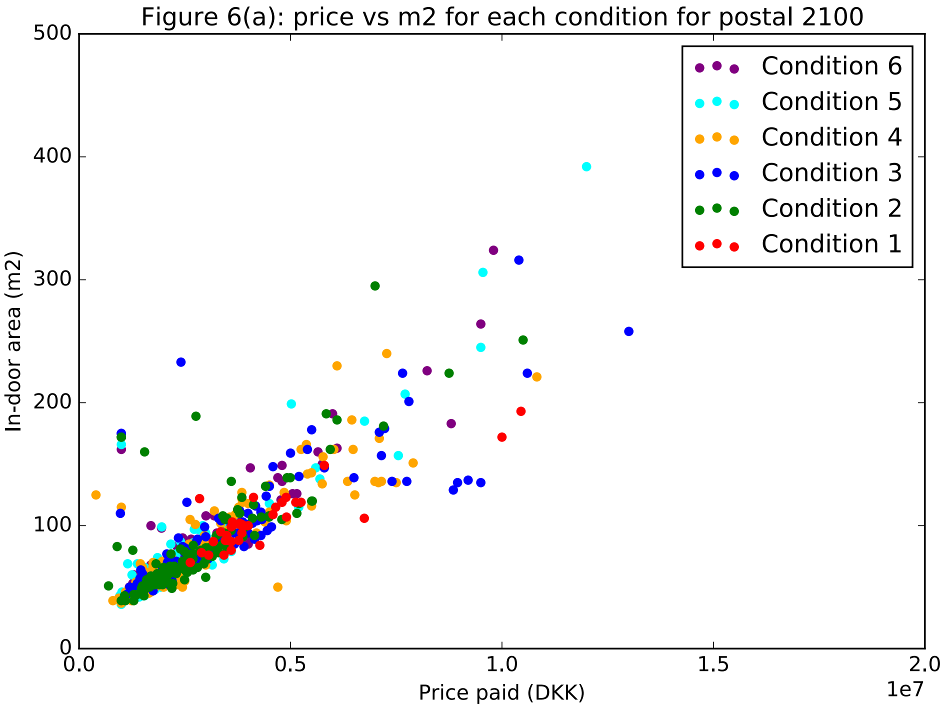
Dividing according to condition, one might expect a higher price/m2, but…

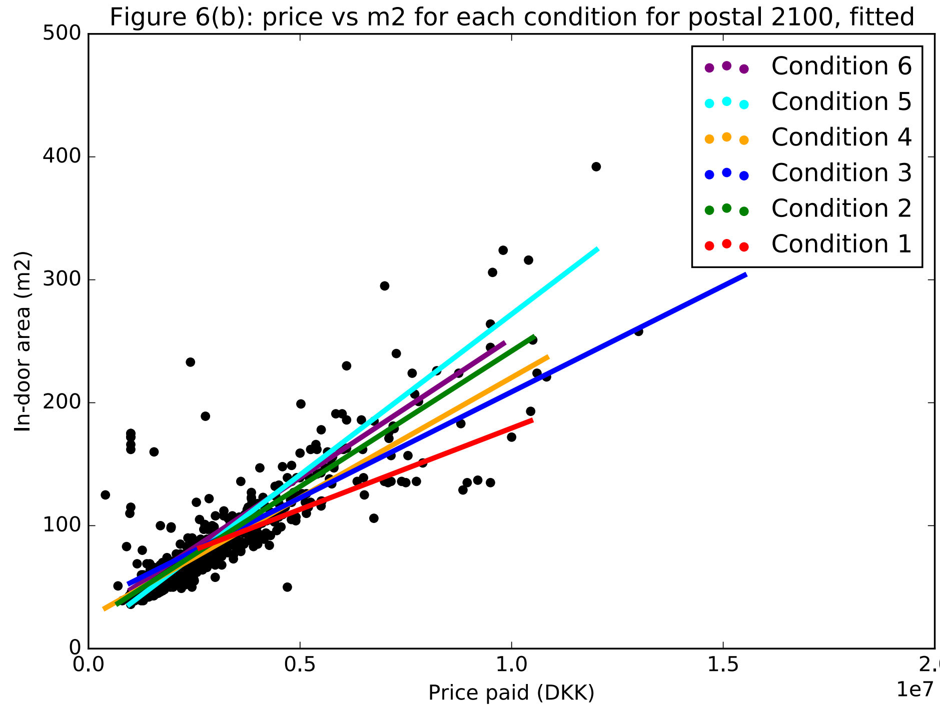


…the pattern is rather, that the basic price is higher! And condition 1 is best!!!

Considering Østerbro only

If we restrict ourselves to Østerbro, the pattern suddenly becomes more clear:



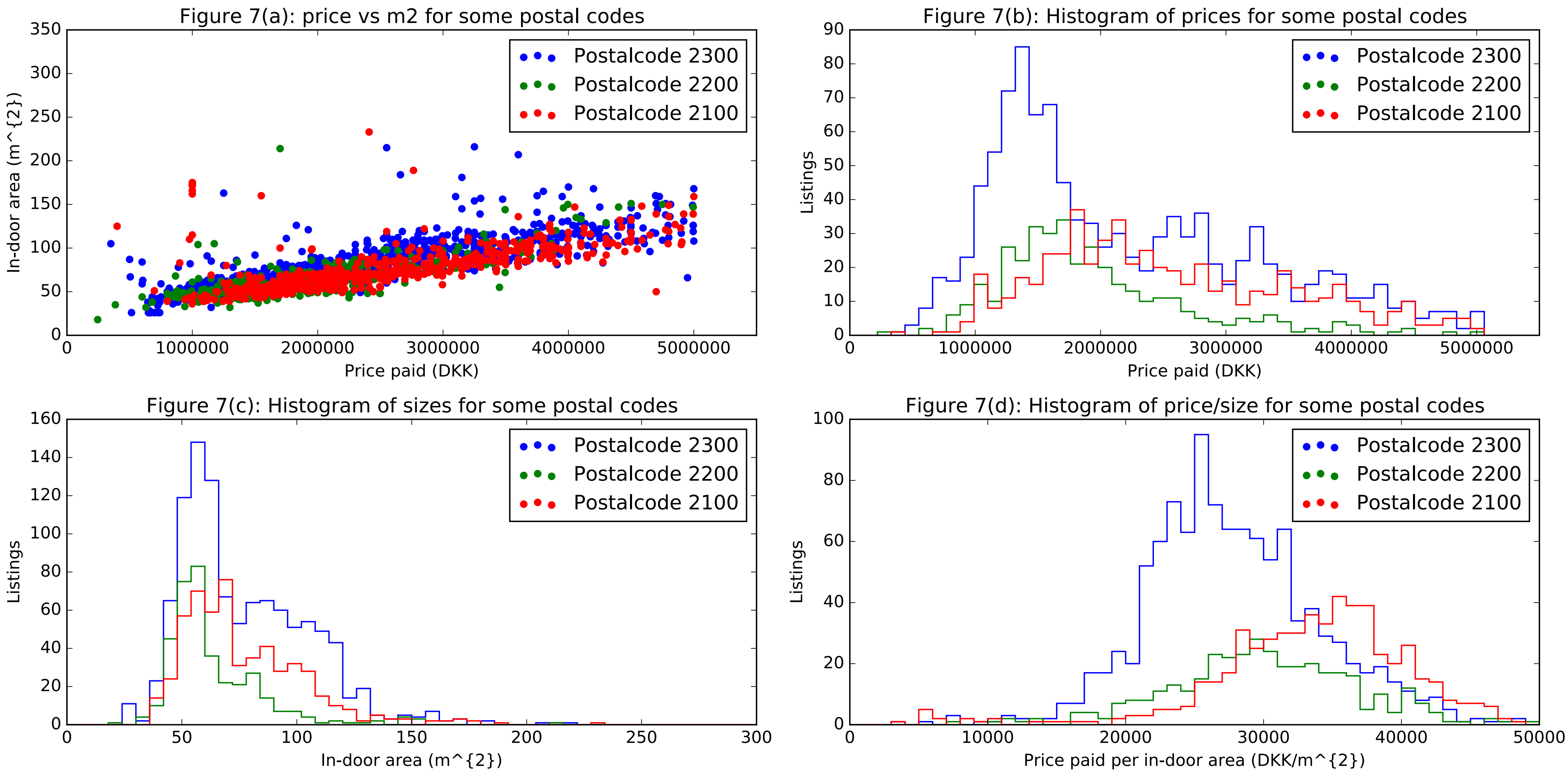


The number of square meters suddenly become a much better indicator, and a condition suddenly also becomes a better variable.

So clearly, district/postal code is also a factor, as should be no surprise.

Comparing districts

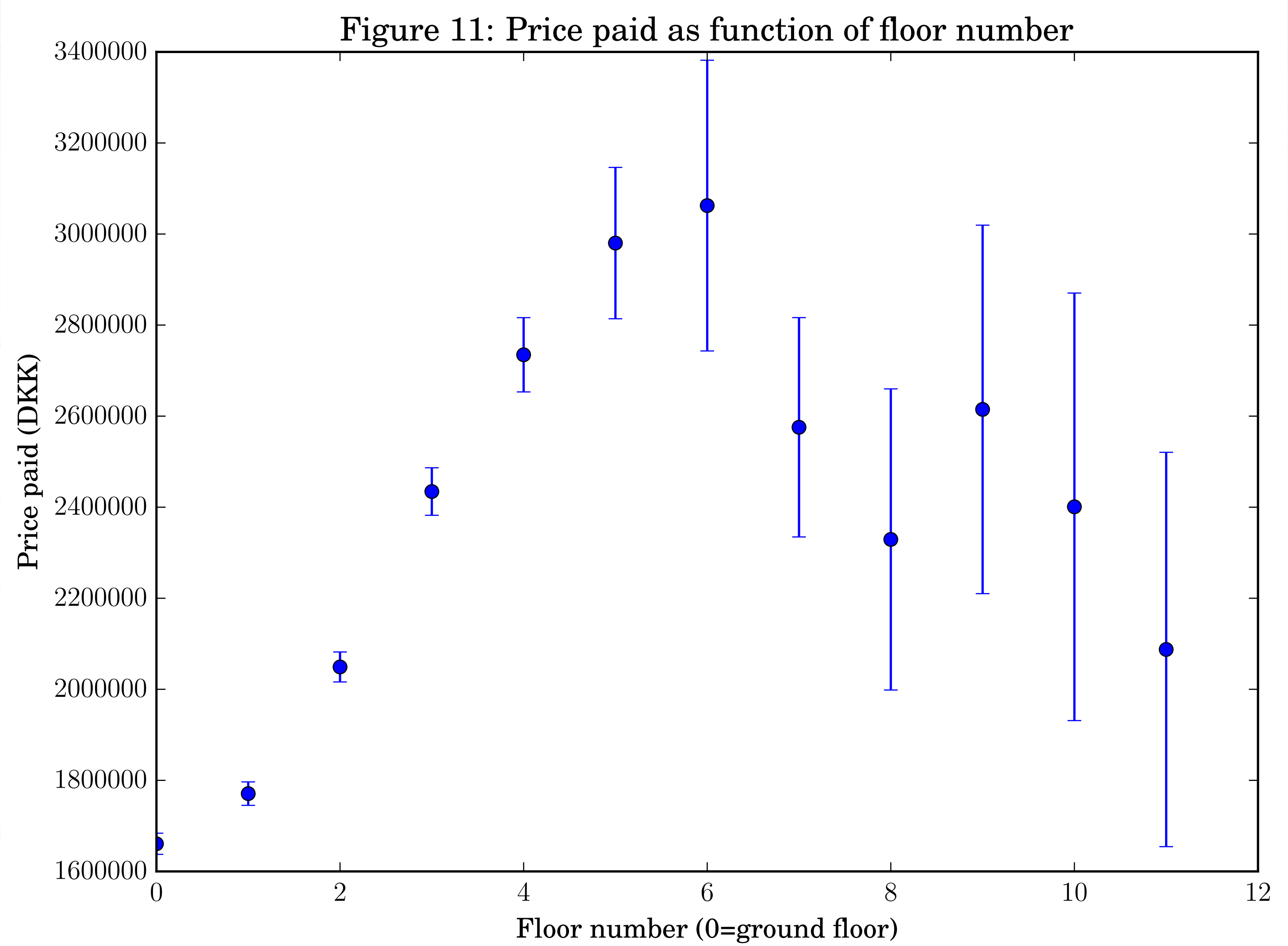
Now we consider the various postal codes (Østerbro, Nørrebro og Amager):



Amager has small apartments and lower price/m2, and the linear model (price = price/m2 * size) holds OK for each district.

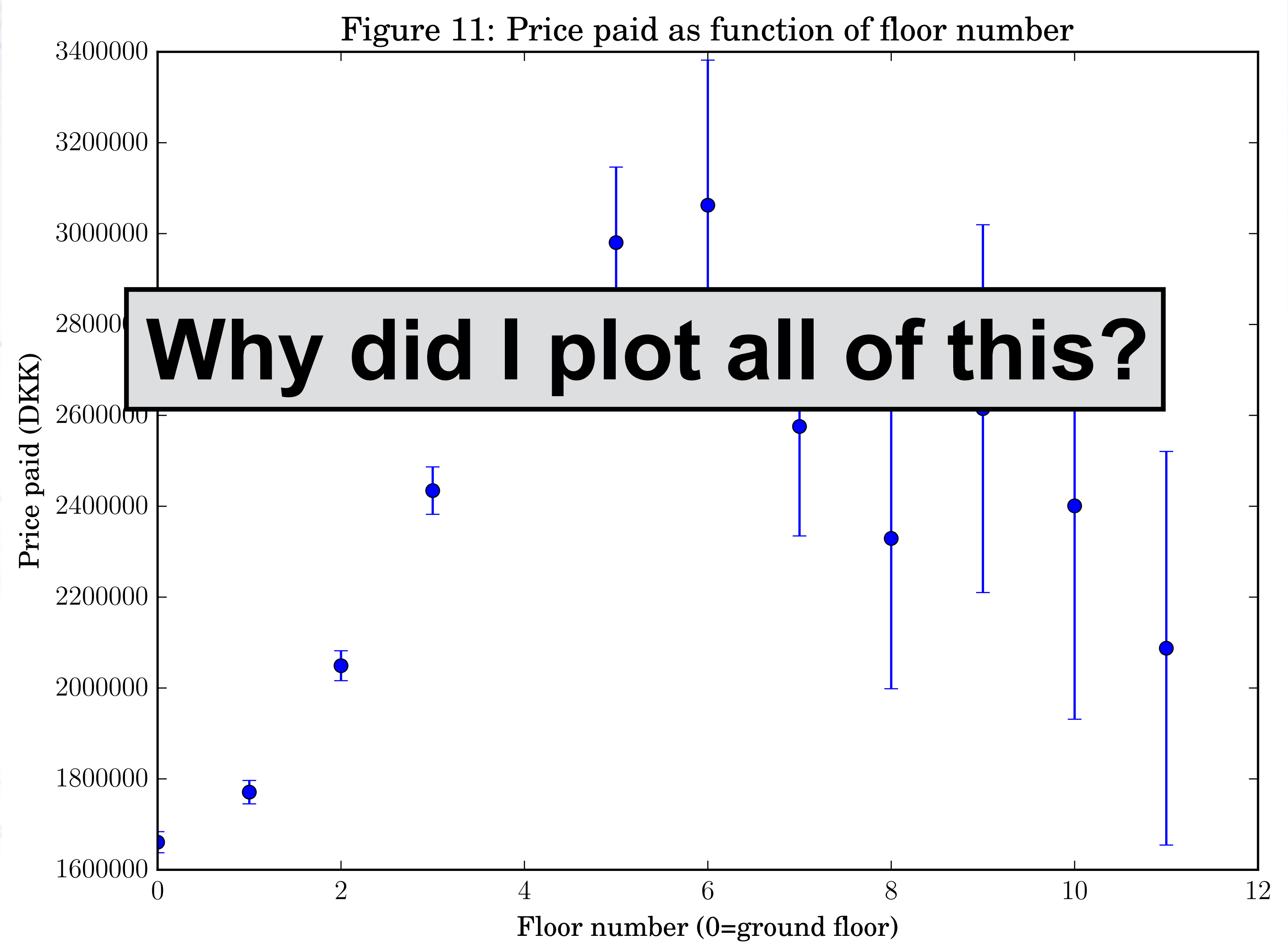
Floor vs. price

One can continue with all sorts of variables, such as e.g. floor:



Floor vs. price

One can continue with all sorts of variables, such as e.g. floor:

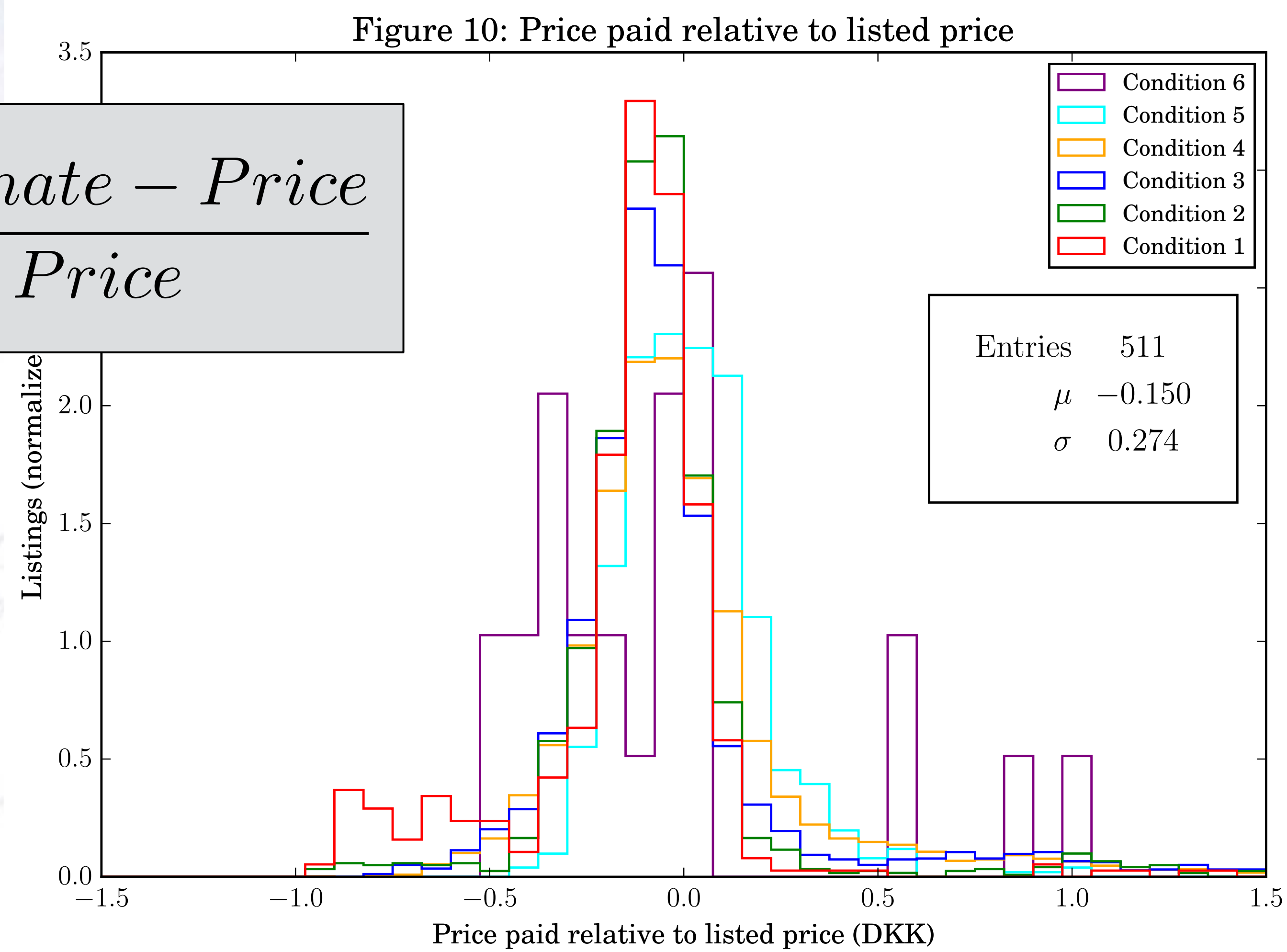


A "measure-of-goodness"

Q: How do we know, that we are improving our price estimates?

A: Well, consider how close the predictions are compared to actual price.

$$\frac{Estimate - Price}{Price}$$

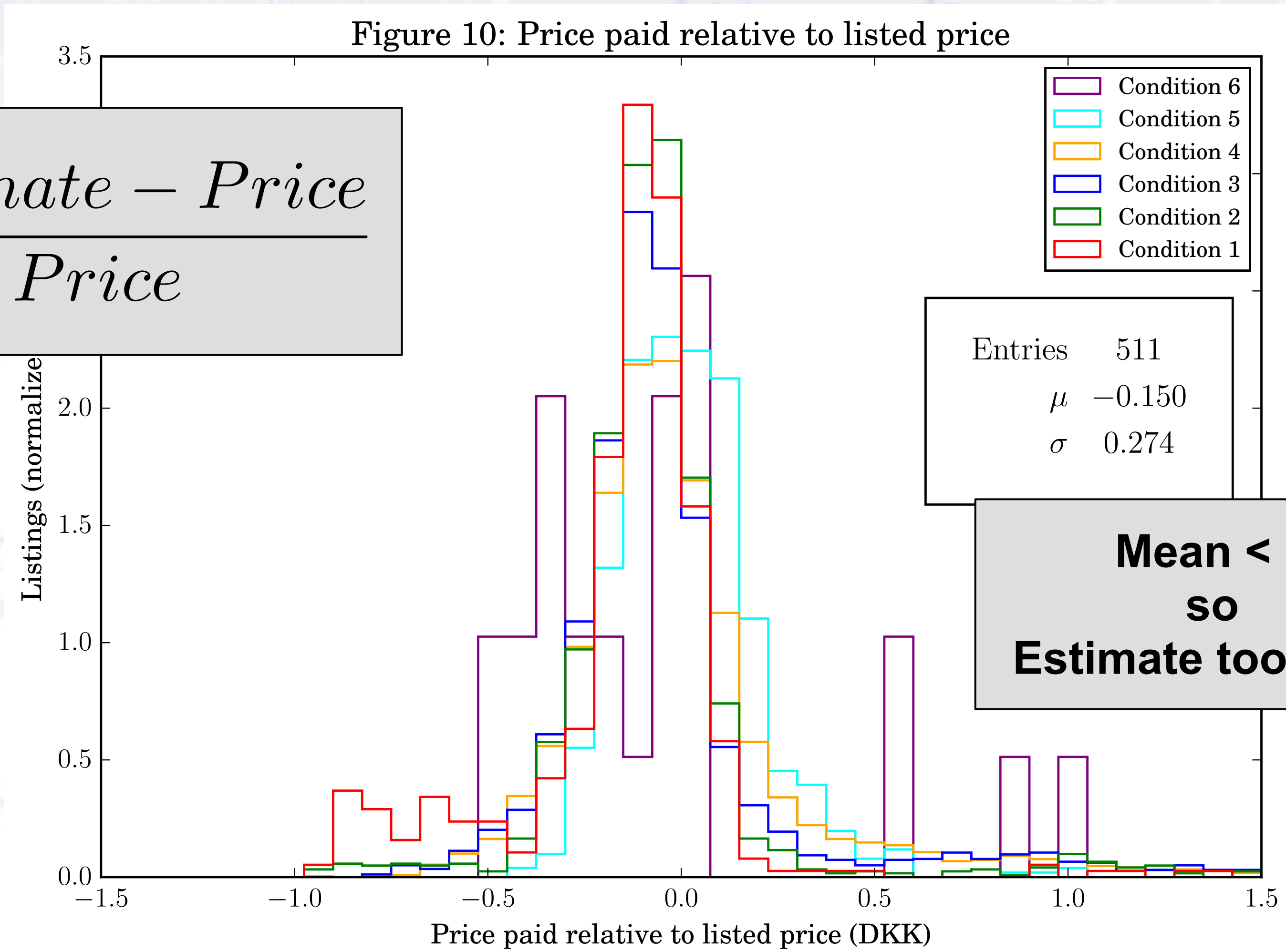


A "measure-of-goodness"

Q: How do we know, that we are improving our price estimates?

A: Well, consider how close the predictions are compared to actual price.

$$\frac{Estimate - Price}{Price}$$



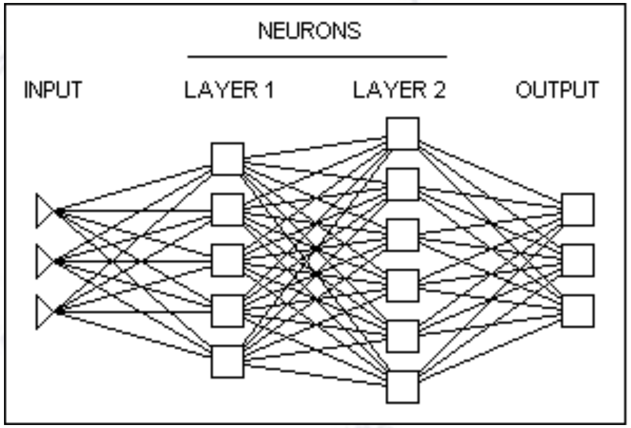
Mean < 0
so
Estimate too high!

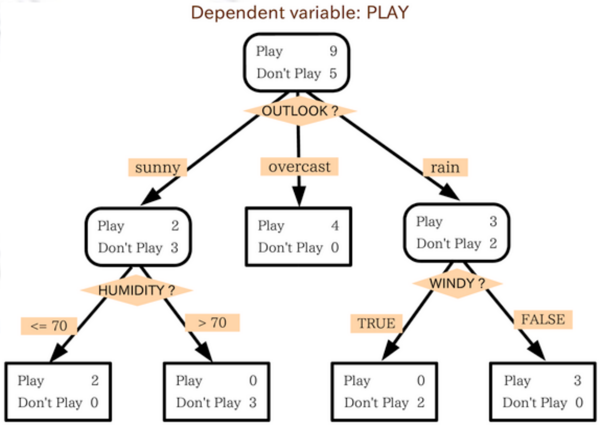
The path forward

Clearly, we could continue in this way, and produce a more and more refined model, which would give a rough estimate for most cases, but…

- The model gets more and more complicated to update or improve.
- There is no "system" by which the model can be improved.
- **The process is very manpower intensive.**

The solution is of course to use Machine Learning (ML) on large datasets (what in industry is often called Big Data analysis), which in an automated and often very powerful way can combine many variables into one "optimal" prediction (or separation, if categorising).





Discussion of path forward

Which considerations do you have in mind regarding doing an ML approach?

- Data size and splitting.
- Current and potential input variables.
- ML algorithms.
- Loss function.
- Output(s).

Discuss first with your collaborators (5 min), and then we'll do it in plenum.