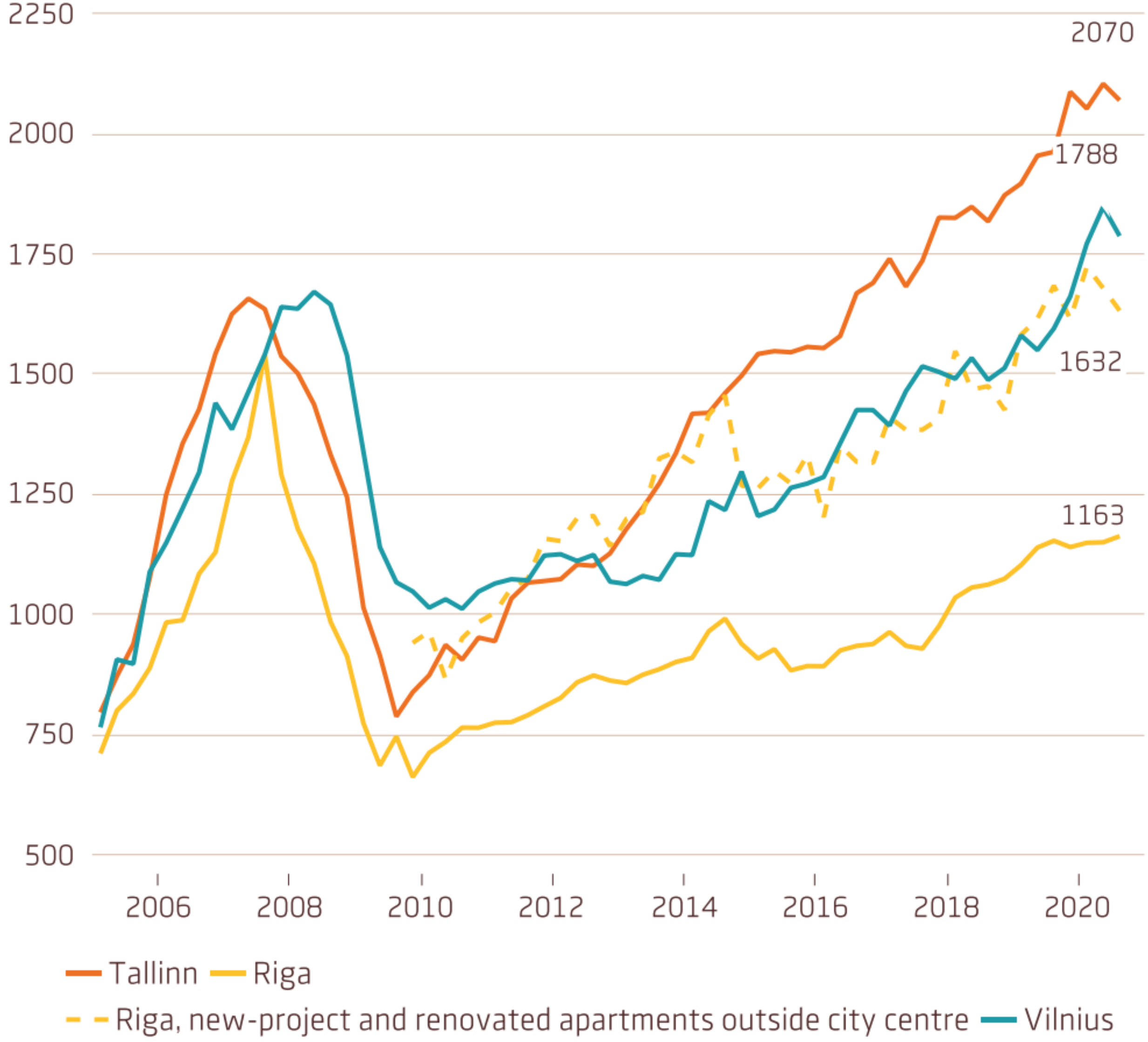
How can Data Analytics help in predicting house prices?

Cracking Riga real estate market

Average apartment transaction price, EUR/m2



Sources: Swedbank Research & Macrobond

Maskavas iela 46 - 3, Centrs — **120 000 EUR**

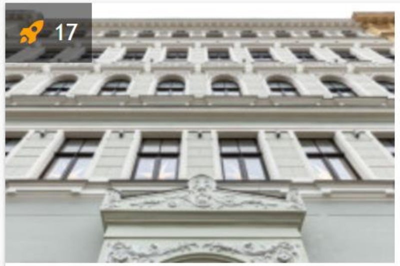
Rīga

1 809.95 EUR/m²

Monthly payment 403 EUR

66.3 m²
4 Rooms
3/4 On Floor

Apartments, For Sale — Rēdera Nami, Amela Al- Asbahi



Krišjāņa Barona iela 6 - 13, Centrs — **112 960 EUR**

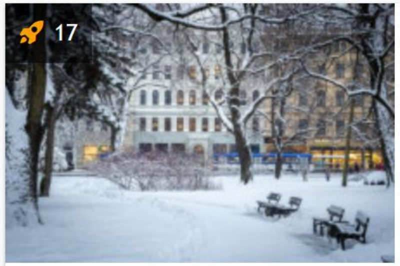
Rīga **Project of the month**

3 200 EUR/m²

Monthly payment 379 EUR

35.3 m²
2 Rooms
5/5 On Floor

Apartments, For Sale — Riverproperties, Laura Semjonova



Krišjāņa Barona iela 6 - 12, Centrs — **99 900 EUR**

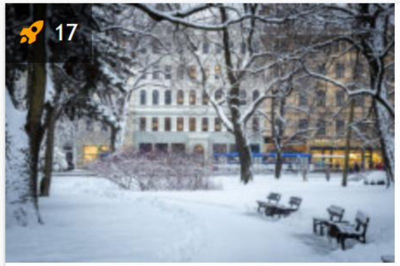
Rīga **Project of the month**

3 000 EUR/m²

Monthly payment 336 EUR

33.3 m²
2 Rooms
2/5 On Floor

Apartments, For Sale — Riverproperties, Laura Semjonova



Krišjāņa Barona iela 6 - 11, Centrs — **102 300 EUR**

Rīga **Project of the month**

3 000 EUR/m²

Monthly payment 344 EUR

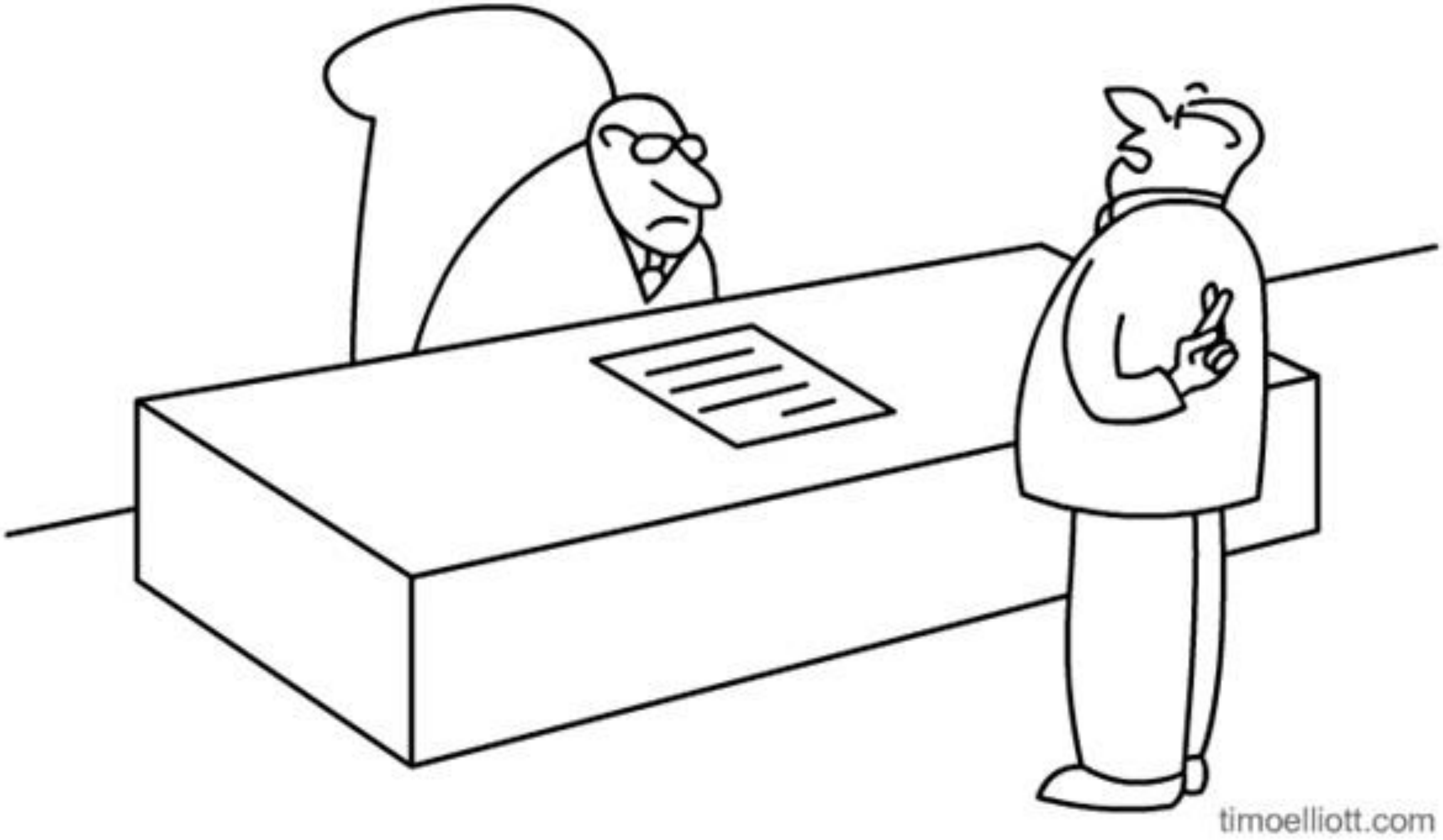
34.1 m²
2 Rooms
2/5 On Floor

Apartments, For Sale — Riverproperties, Laura Semjonova

PROPERTY VALUE
$
LOCATION
SELLER'S TIME FRAME
IMPROVEMENTS
CONDITION
MARKET
NEIGHBORHOOD


Valuations

Benchmarking of houses



"Yes sir, you can absolutely trust those numbers"

"When you two have finished arguing your opinions, I actually have data!"

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"


AC
ON/C
0
3

KEEP
CALM

AND

LET`S SEE WHAT WE
FOUND OUT

AGENDA

1 What we analyzed

Setting the context | Defining the problem statement

2 What we found

Conclusions and unexpected insights

3 How we analyzed

What goes in the background in making all of this…

1 What we analyzed

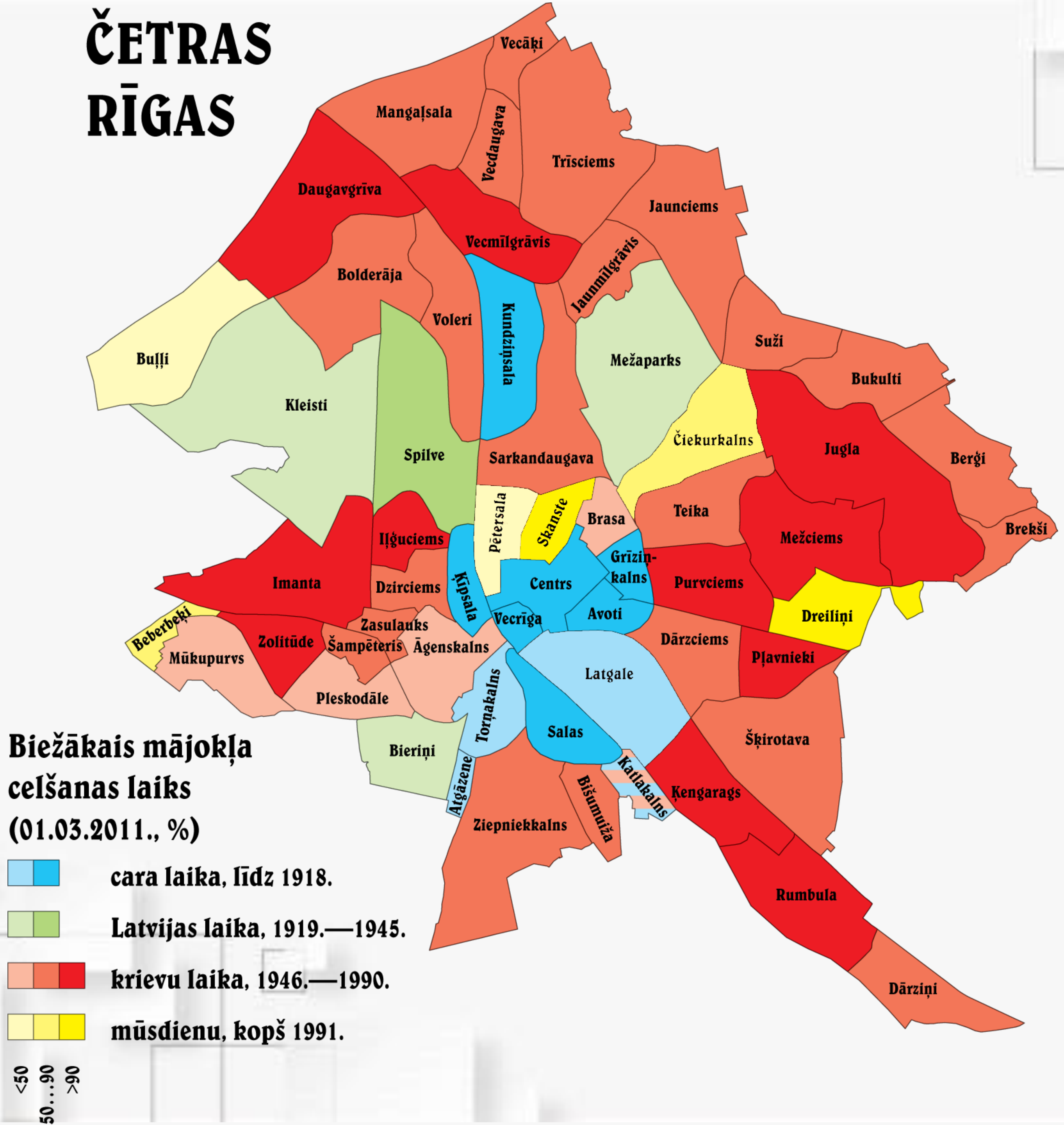
2 What we found

Cleaning of data | Regression analysis

3 How we analyzed

Conclusions and

Setting the context



Examples of house types

Classical

Stalin era houses

New houses

Neo Classical

Soviet mass housing

New houses exclusive

Most common areas

City center	Historic heighbourhoods	Soviet built areas

Other smaller regions

Private residences within city limits

Seaside region

Criminal, run down areas

New area

1 What we analyzed

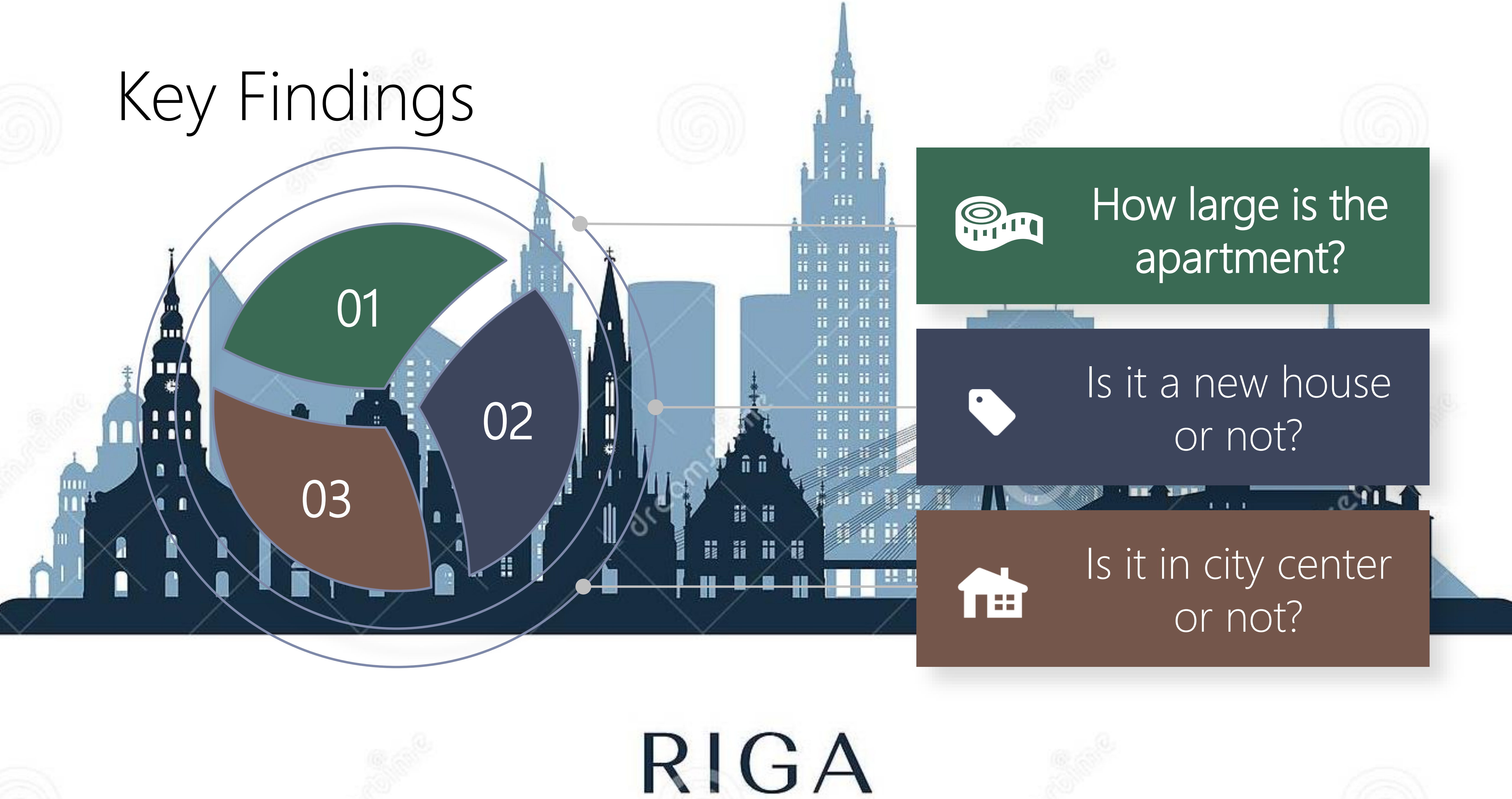
Setting the context | Defining the problem statement

2 What we found

3 How we analyzed

Conclusions and inferences

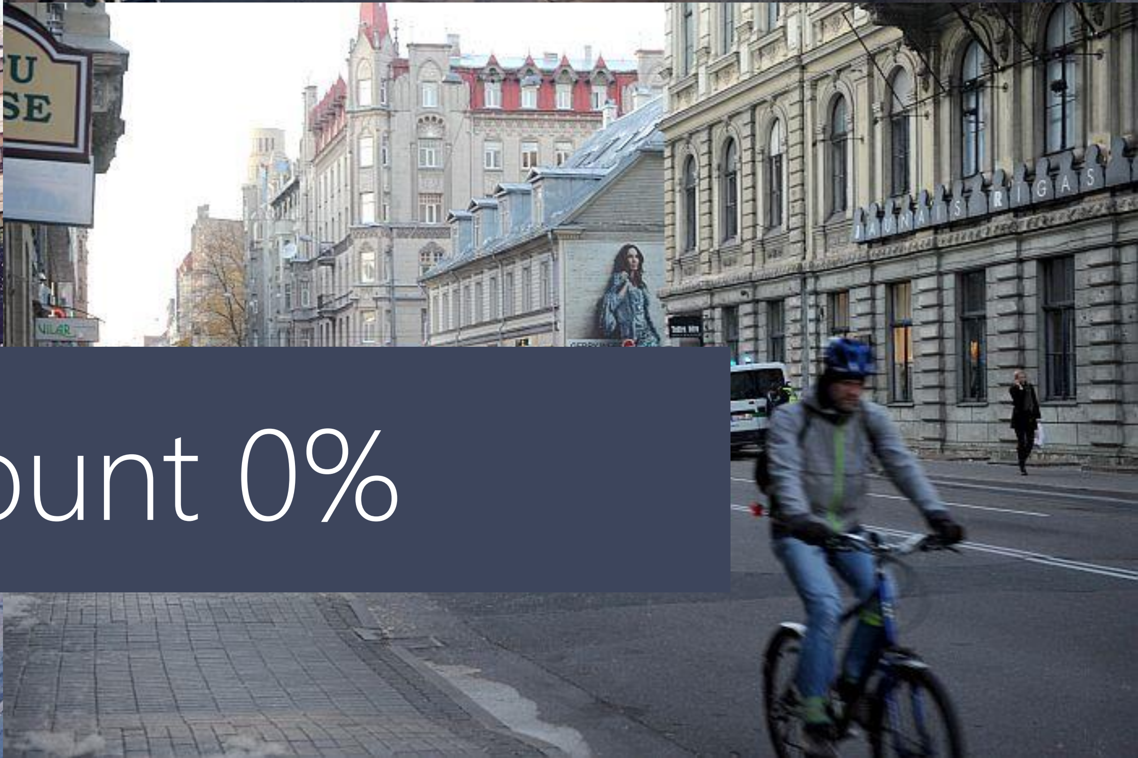
Key Findings



Size matters, every 1%↑ in house size increases price by 1.1%


Whilst similar, there is 40% price difference between both house types

Apartments in city center are more expensive than suburbs



Historic center discount 0%

Historic regions discount -20%


Soviet housing regions discount -40%

City limits discount -60%

Criminal & run-down areas discount -80%

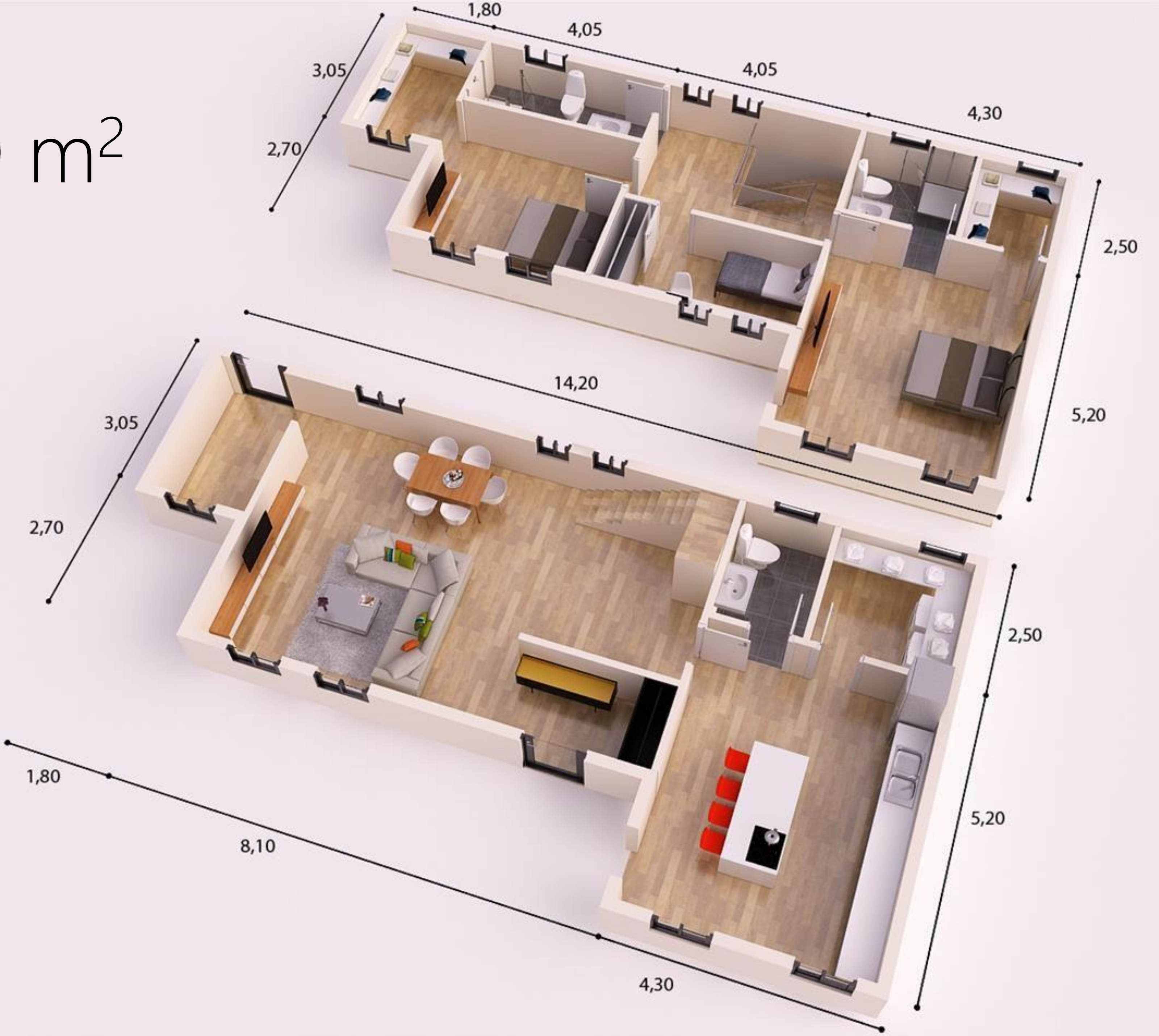

Private residences premium +10%



Historic houses in city center were
2x more valuable than other regions

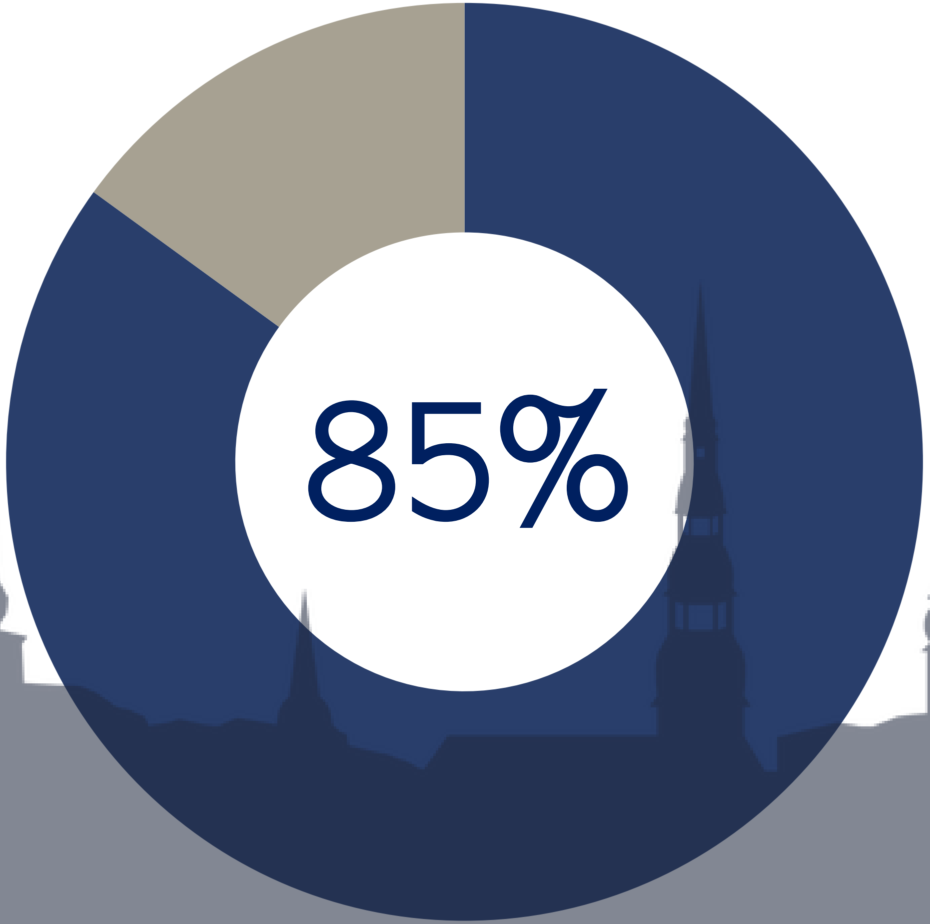
1ST FLOOR

+200 m^2




VESTABALT
NEKUSTAMIE ĪPAŠUMI

How much we can understand the price?



What to do with rest of it? Well, you make the call!

1 What we analyzed

Setting the context | Defining the problem statement

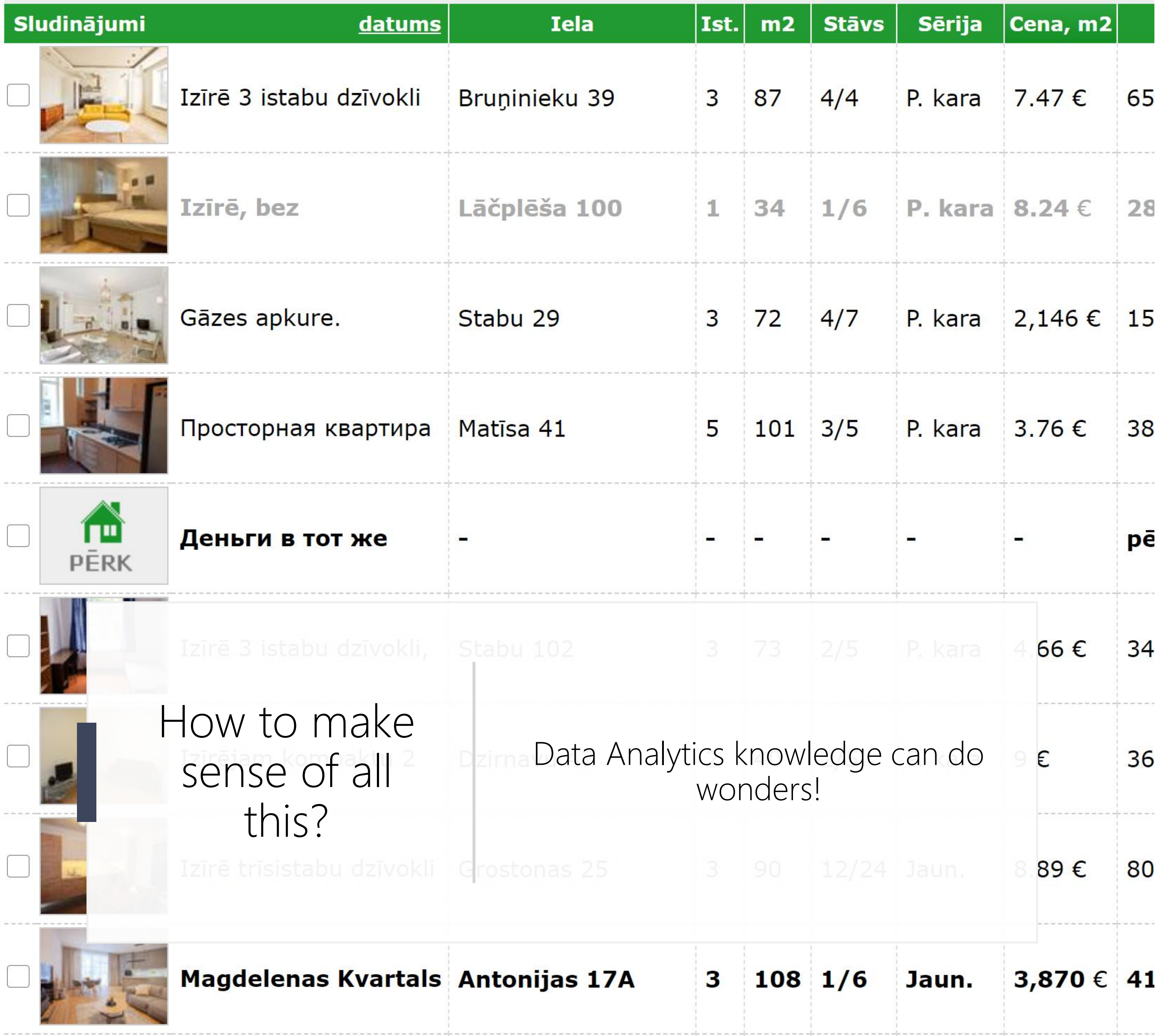
2 What we found

Cleaning of data | Regression analysis

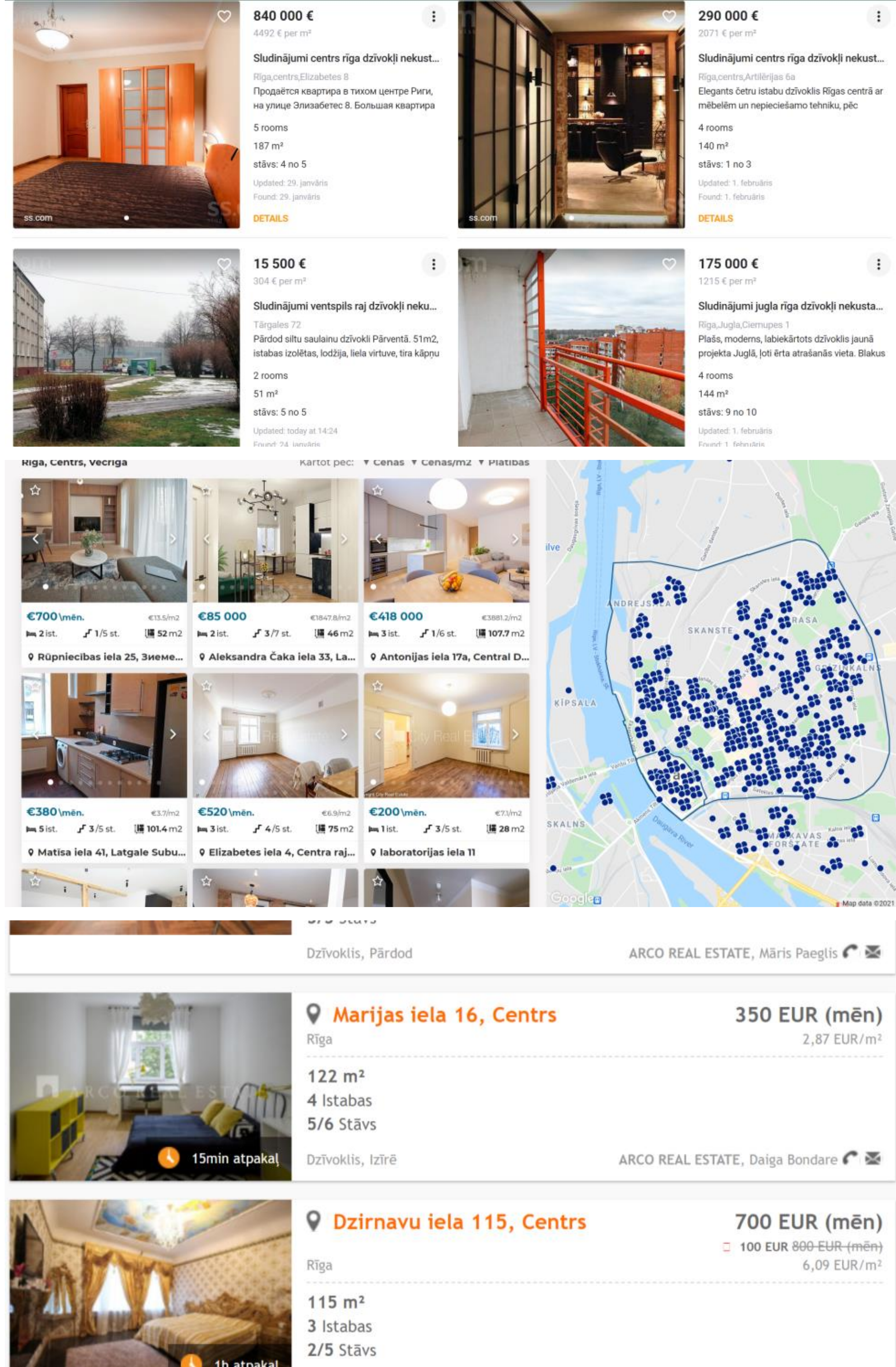
3 How we analyzed

How to make sense of all this?

Data Analytics knowledge can do wonders!

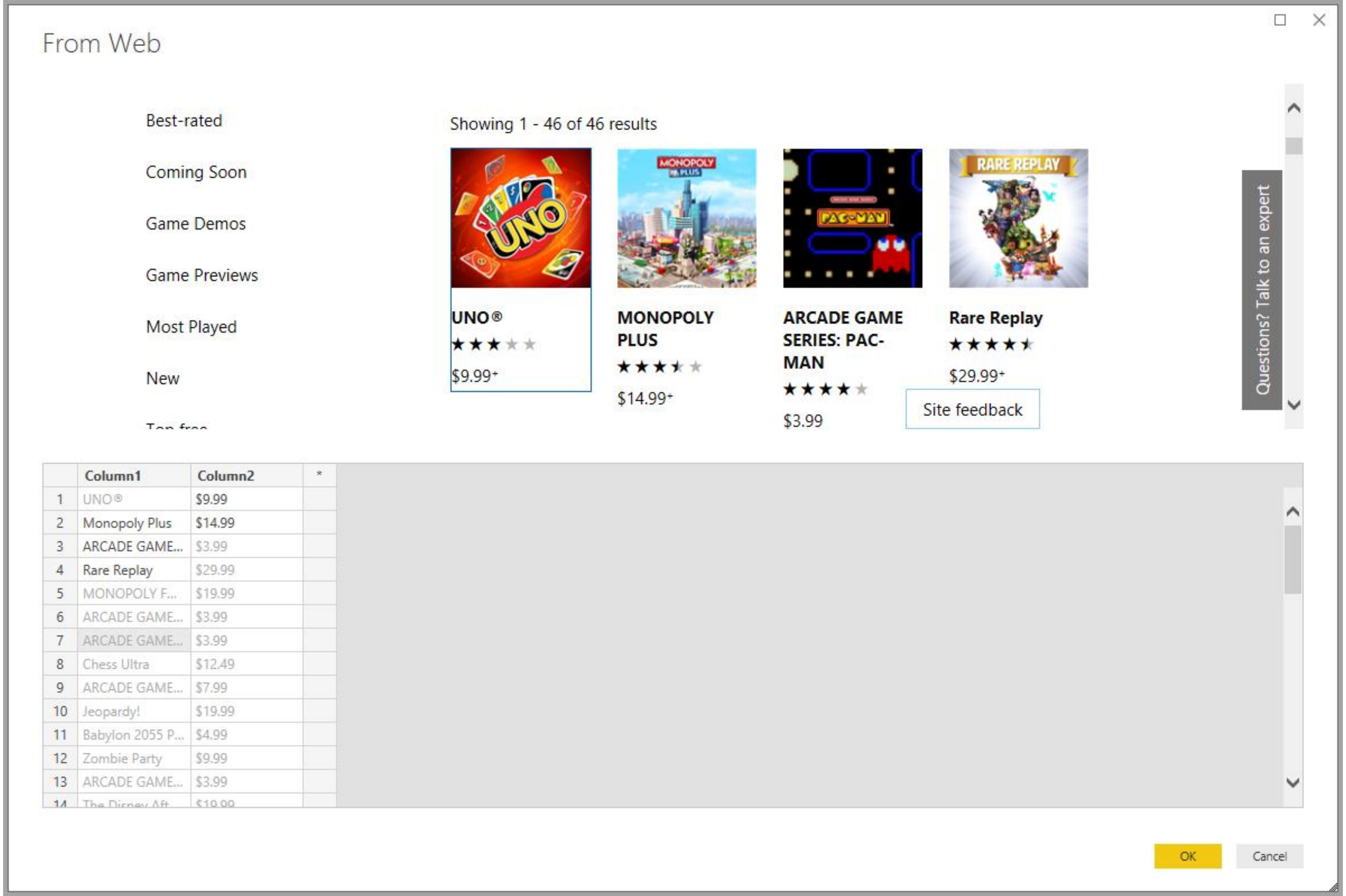


Sludinājumi datums	Iela	Ist.	m2	Stāvs	Sērija	Cena, m2	
Izīrē 3 istabu dzīvokli	Bruņinieku 39	3	87	4/4	P. kara	7.47 €	65
Izīrē, bez	Lāčplēša 100	1	34	1/6	P. kara	8.24 €	28
Gāzes apkure.	Stabu 29	3	72	4/7	P. kara	2,146 €	15
Просторная квартира	Matīsa 41	5	101	3/5	P. kara	3.76 €	38
PĒRK Деньги в тот же	-	-	-	-	-	-	pē
Izīrē 3 istabu dzīvokli,	Stabu 102	3	73	2/5	P. kara	4.66 €	34
Izīrējam kompakto 2	Dzirna[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]9 €	36
Izīrē trīsistabu dzīvokli	Grostonas 25	3	90	12/24	Jaun.	8.89 €	80
Magdelenas Kvartals	Antonijas 17A	3	108	1/6	Jaun.	3,870 €	41



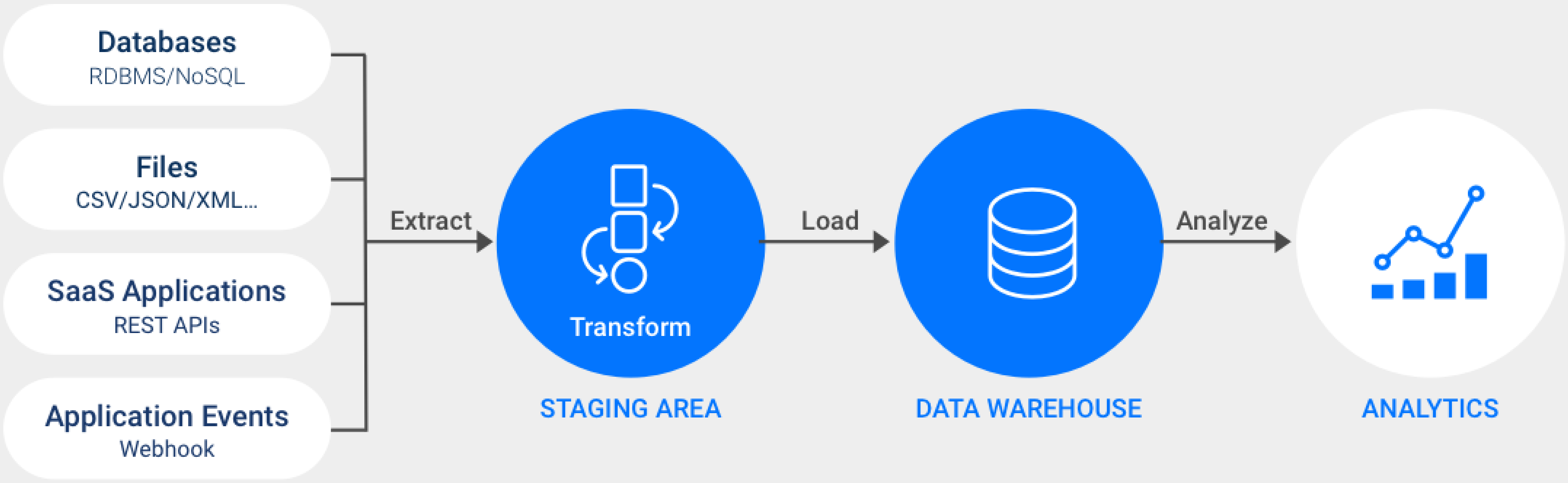
edureka!

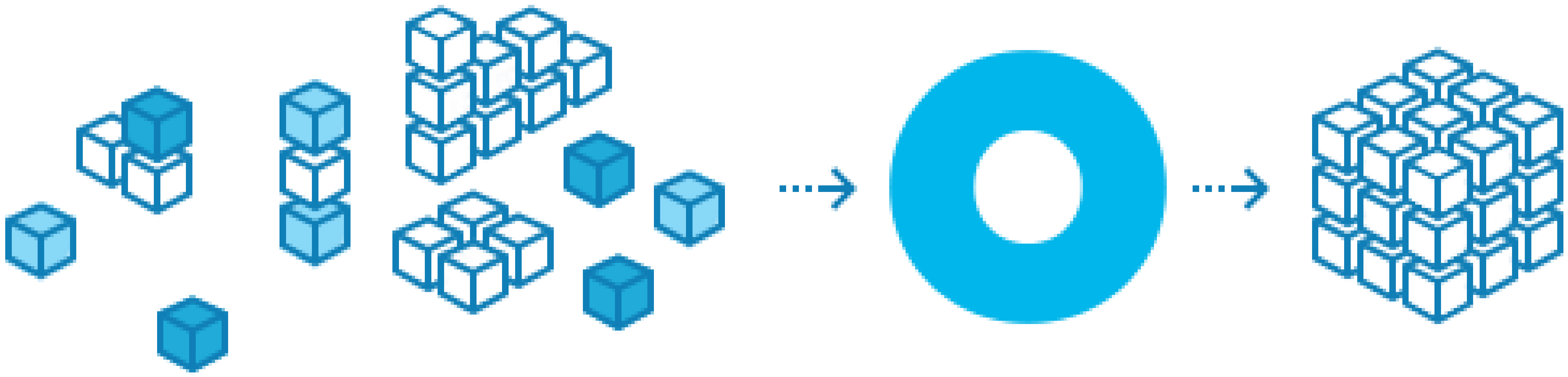
Web Scraping with Python



	Column1	Column2
1	UNO®	$9.99
2	Monopoly Plus	$14.99
3	ARCADE GAME...	$3.99
4	Rare Replay	$29.99
5	MONOPOLY F...	$19.99
6	ARCADE GAME...	$3.99
7	ARCADE GAME...	$3.99
8	Chess Ultra	$12.49
9	ARCADE GAME...	$7.99
10	Jeopardy!	$19.99
11	Babylon 2055 P...	$4.99
12	Zombie Party	$9.99
13	ARCADE GAME...	$3.99

ETL PROCESS


Databases
RDBMS/NoSQL
Files
CSV/JSON/XML...
SaaS Applications
REST APIs
Application Events
Webhook
Extract
Transform
STAGING AREA
Load
DATA WAREHOUSE
Analyze
ANALYTICS



Source data | Cleaning and identification of variables

We web extracted data from a largest listings website's real estate section [ss.com/lv/real-estate/flats/riga/].

For the purposes of this analysis, we simplified the data as follows:

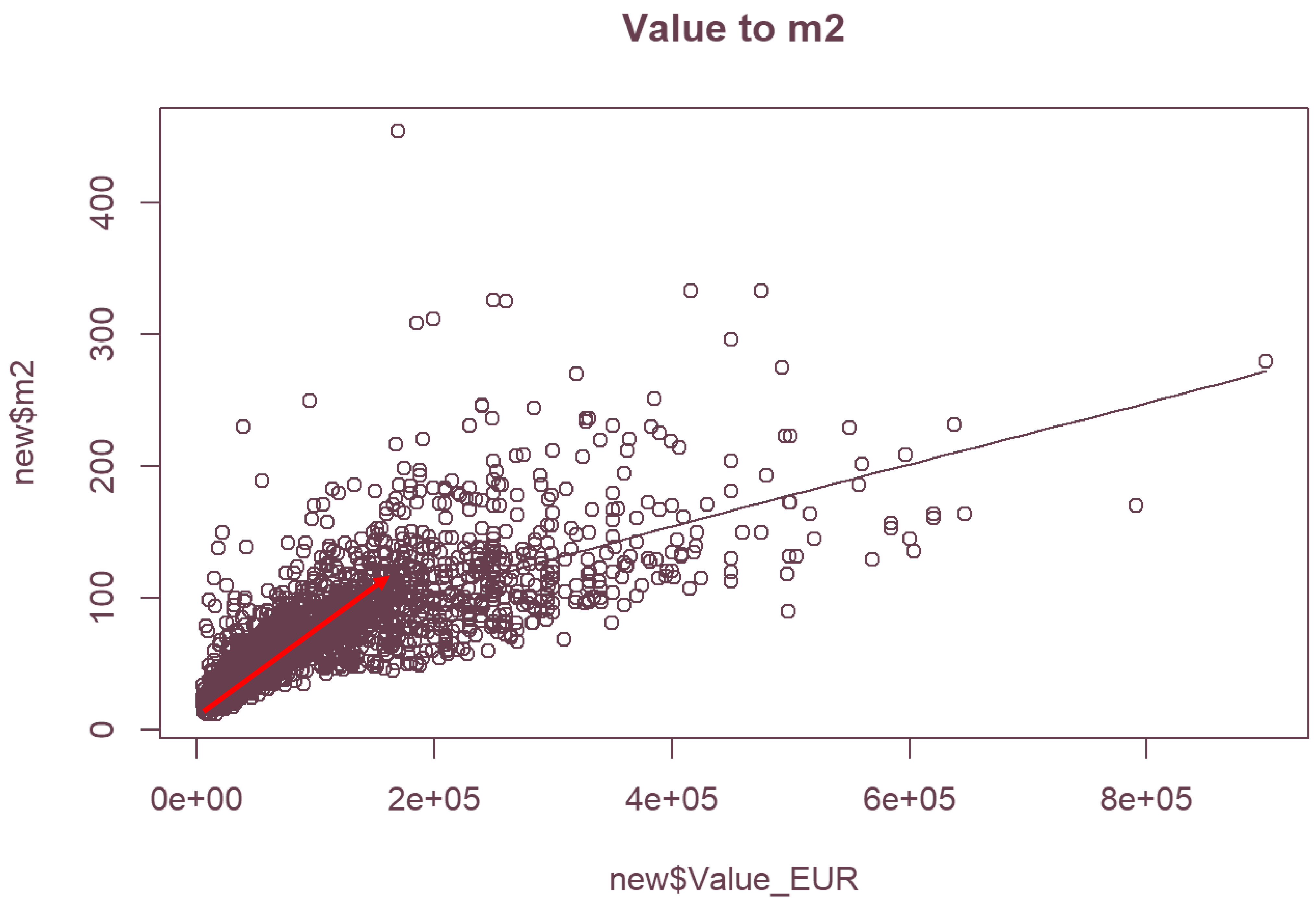
- Considered only those listings that are "for sale"
- Re-categorized 'Region' to limit the number of unique/distinct regions to 8.
- Re-categorized 'House_type' to limit the number of unique house types to 6.

Street	Region	Street_name	Rooms	m2	Value_EUR	Price_per_m2	Sale_type	House_type
Valnu 4	City center	Valnu	5	170	790500	4,650	For Sale	New
Vilandes 8	City center	Vilandes	3	164	647010	3,945	For Sale	Neo Classical
Ausekla 6	City center	Ausekla	6	232	638000	2,750	For Sale	Neo Classical
Krisjana Valdemara 23	City center	Krisjana	5	164	620000	3,780	For Sale	Classical

We also recognized that 'Street' and 'Street_name', would be drilling down even further into the region and hence, were not relevant. In addition, we did not expect that few such entries would yield statistically significant result.

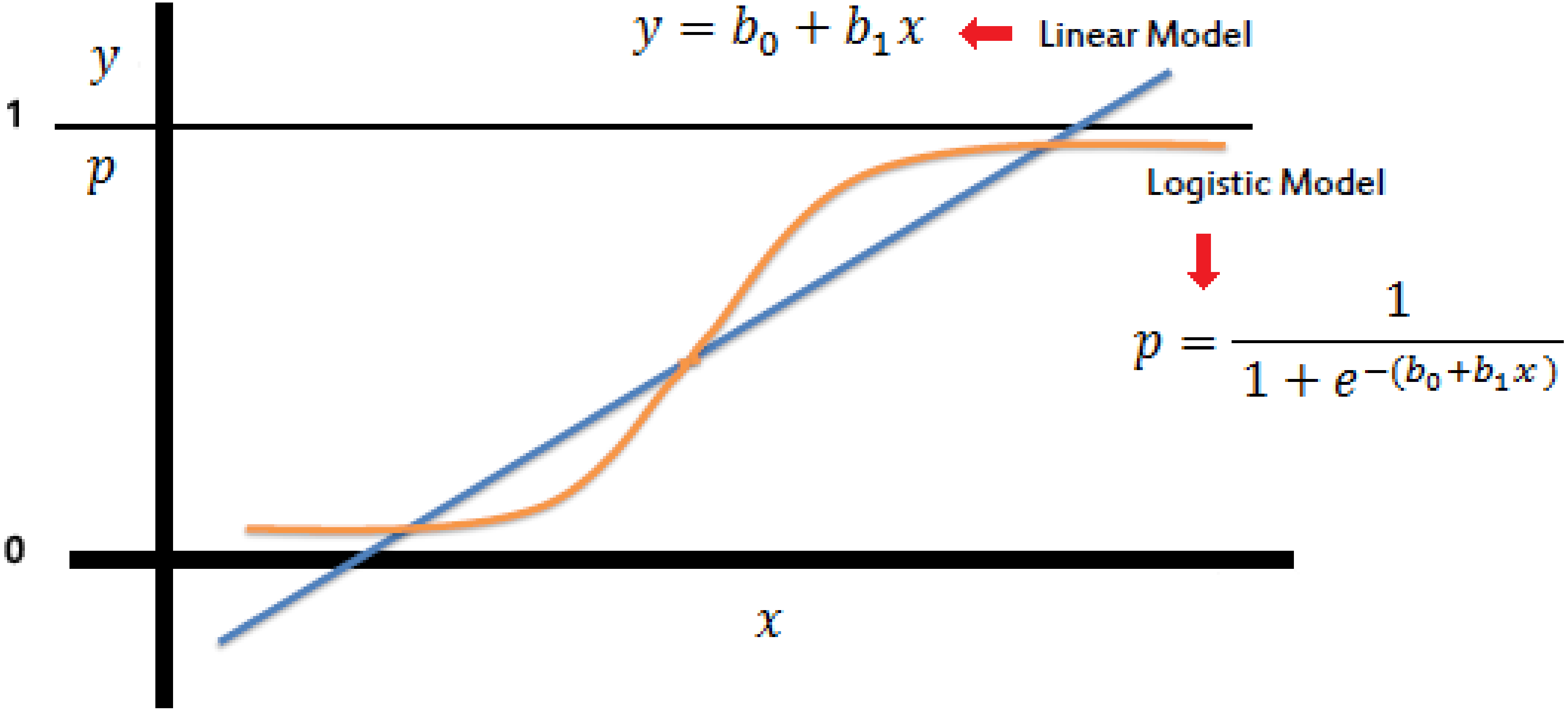
Additionally, 'Price_per_m2' would be highly correlated to Property price (i.e. 'Value_EUR') and hence was removed from analysis.

Hence, for the final analysis we retained the dependent variable 'Value_EUR' and the independent variables - 'Region', 'Rooms', 'm2', 'House_type'.


Value to m2
new$m2
0
100
200
300
400
new$Value_EUR
0e+00
2e+05
4e+05
6e+05
8e+05


CHECKLIST

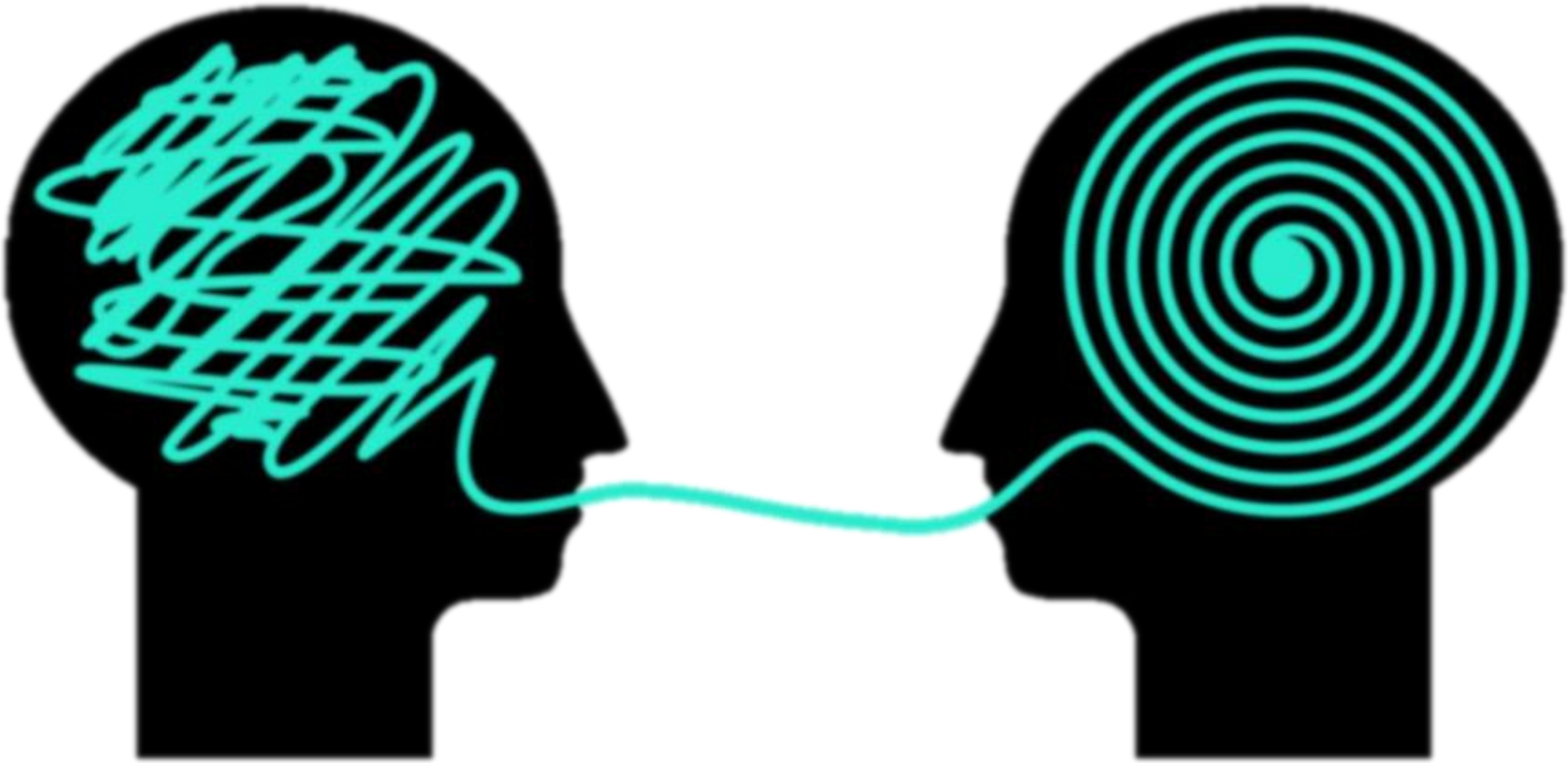
$$Ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_kX_k$$

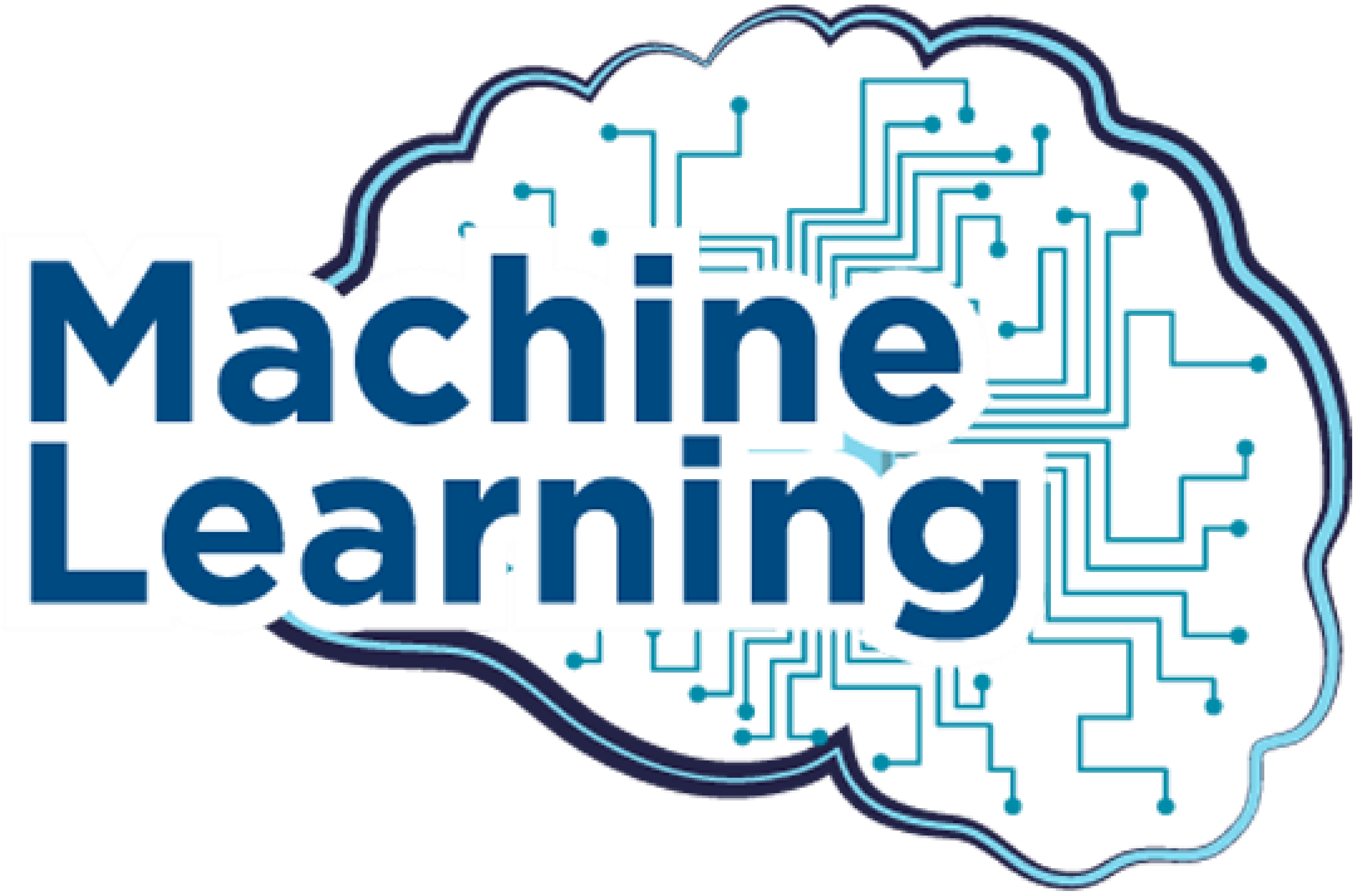


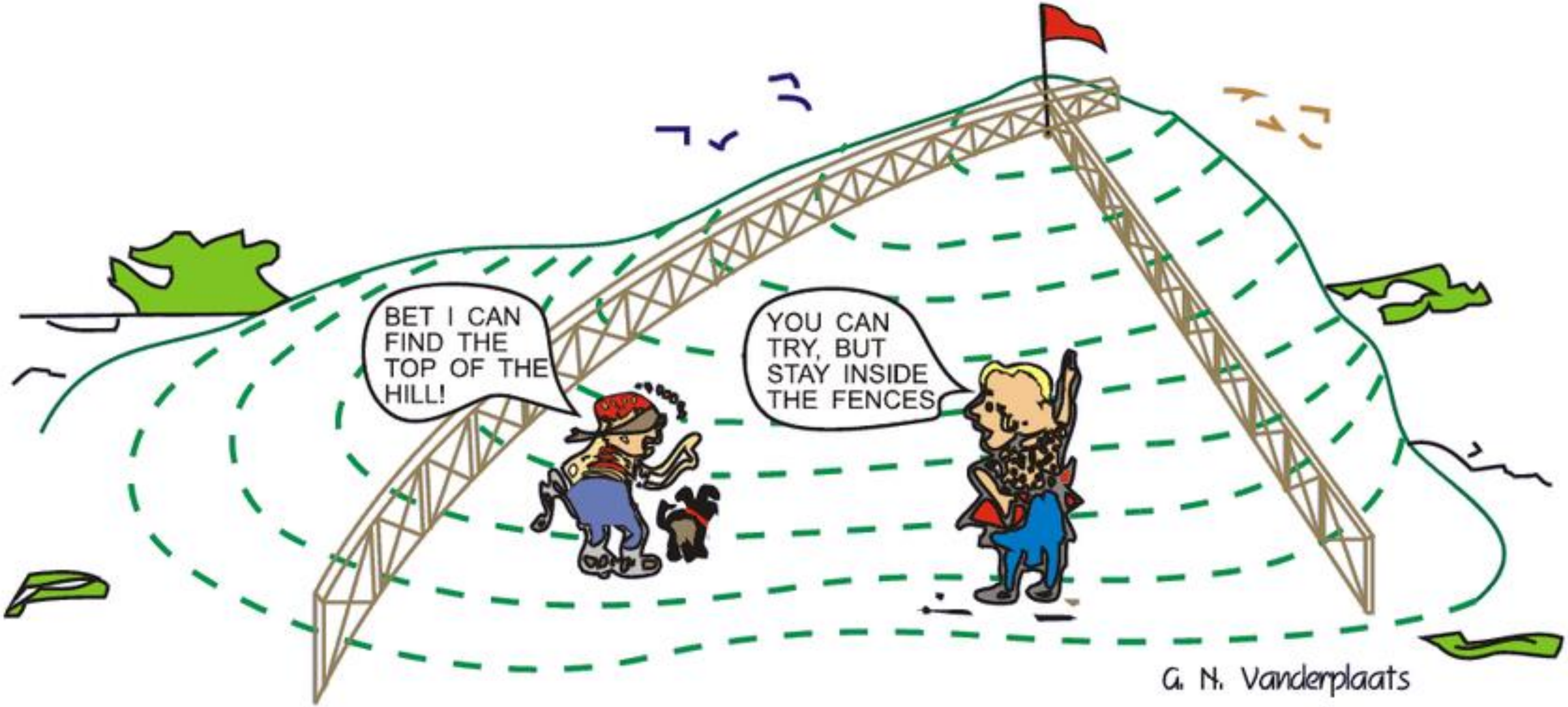
Your head has most probably exploded by this point

```
## Call:
## lm(formula = Value_EUR ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -433052  -20269   -1388   14468  498864
##
## Coefficients: (2 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -44727.33   894      -5.002 6.03e-07
## Rooms                         3828.3    1410.26   2.7
## m2                            1290        39.
## Floor                         245         43
## House_type_Classical         -924         48
## `House_type_Czech type`      -3438.63
## `House_type_Lithuania type`   2121.13
## `House_type_Neo Classical`   53750.99
## House_type_New               34237.96
## House_type_Other
## `House_type_Pre Soviet era`  -154        6342.50  -2.441 0.01
## `House_type_Private house`   -896       11793.71  -0.760 0.
## `House_type_Sieries 103`      4137.
## `House_type_Sieries 104`     -5155.48
## `House_type_Sieries 119`     -4038.
## `House_type_Sieries 467`     -6696
## `House_type_Sieries 602`     -1606
## `House_type_Soviet era`
## Reg_agenskalns               18710.
## Reg_Aplokciems                1116.3
## Reg_Bergi                    -8418.22
## Reg_Biekensala              -46056.49
## Reg_Bierini                  -8492.74
## Reg_Bolderaja                -3991.70
## Reg_Bukulti                 -42244.39    538
## Reg_ciekurkalns               9854.93   12826.
## Reg_Darzciems                 2530.89   12604.98
## Reg_Daugavgriva              -4808.51   15173.76  -0.
## Reg_Dreilini                -20982.21   53877.08  -0.389 0.
## Reg_Dzeguzkalns               5993.45   13481.07   0.445 0.656656
## Reg_Ilguciems                 5180.10    9726.11   0.533 0.594355
## Reg_Imanta                    8087.55    9080.31   0.891 0.373184
## Reg_Jaunciems                  337.44   17984.21   0.019 0.985032
```

```
## Reg_Katlakalns      -2879.07   27819.48  -0.103 0.917581
##                      5575.46    9149.23   0.609 0.542315
##                    208475.96   16782.35  12.422  < 2e-16 ***
##                     81260.83   16656.65   4.879 1.13e-06 ***
##                      8794.66   14342.76   0.613 0.539809
##                      1589.62   18890.62   0.084 0.932944
##                    -15813.43   20268.81  -0.780 0.435348
##          ta`          229.51   10999.84   1.112 0.266325
##                       953.14   11324.63   3.616 0.000304 ***
##     lezciems          366.47   10367.35   0.228 0.819460
##     her                   NA         NA      NA       NA
##                      2595.11    9161.34   0.283 0.776993
##                      4036.73    8302.61   0.486 0.626864
##                     58680.62    8164.06   7.188 8.44e-13 ***
##                     15845.60   13045.40   1.215 0.224603
##                      2543.90   10022.45   0.254 0.799653
##                     16727.60   38452.29  -0.435 0.663580
##                     17032.06    9791.50   1.739 0.082062 .
##                      8550.30   12140.24   0.704 0.481309
##                     -2618.14   15044.00  -0.174 0.861853
##                      3394.01   27880.56   0.122 0.903119
##                       999.78    9991.84   0.100 0.920305
##                     -1074.48   38567.68  -0.028 0.977776
##                     23032.94   53986.16   0.427 0.669671
##                      1623.89    8990.33   0.181 0.856674
##
##    odes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53280 on 2771 degrees of freedom
## Multiple R-squared:  0.7035, Adjusted R-squared:  0.6974
## F-statistic: 115.4 on 57 and 2771 DF,  p-value: < 2.2e-16
```




Machine
Learning


BET I CAN FIND THE TOP OF THE HILL!
YOU CAN TRY, BUT STAY INSIDE THE FENCES
G. N. Vanderplaats

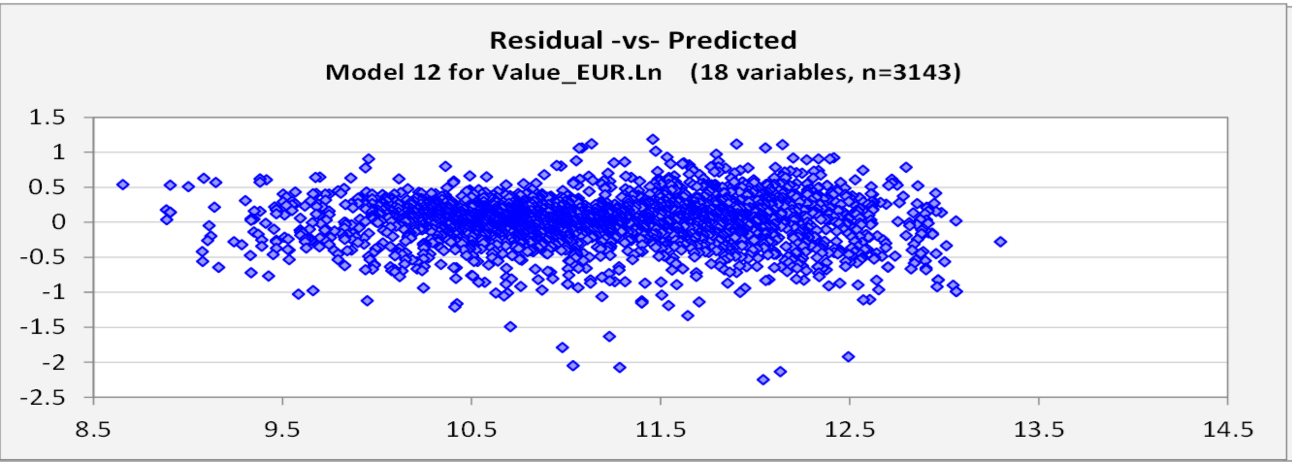
Final regression model

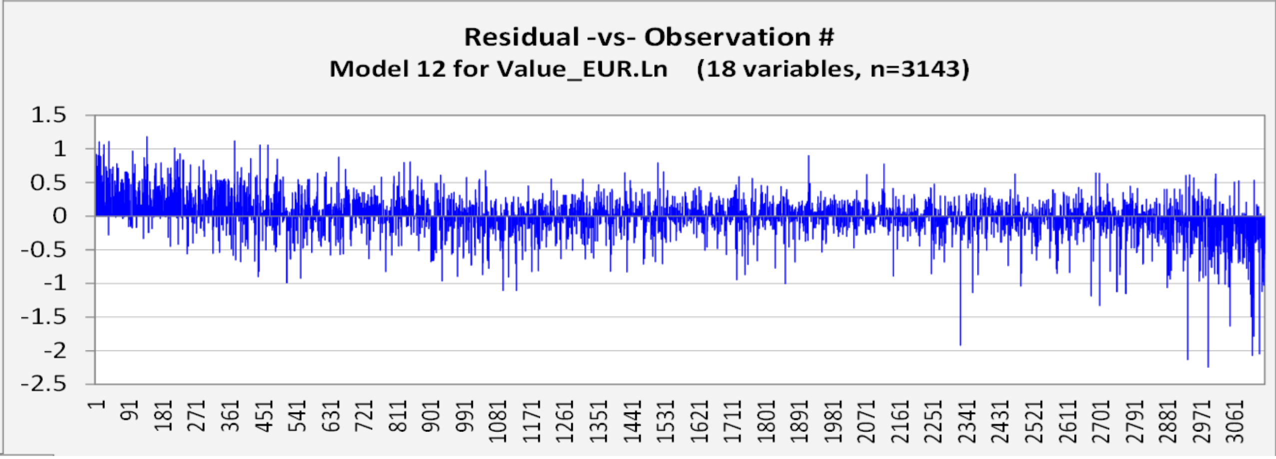
Dependent Variable:	R-Squared	Adj.R-Sqr.	Std.Err.Reg.	Std.Dep.Var.	# Fitted	# Missing	Critical t	Confidence
Value_EUR.Ln	0.852	0.852	0.328	0.852	3143	1	1.961	95.0%

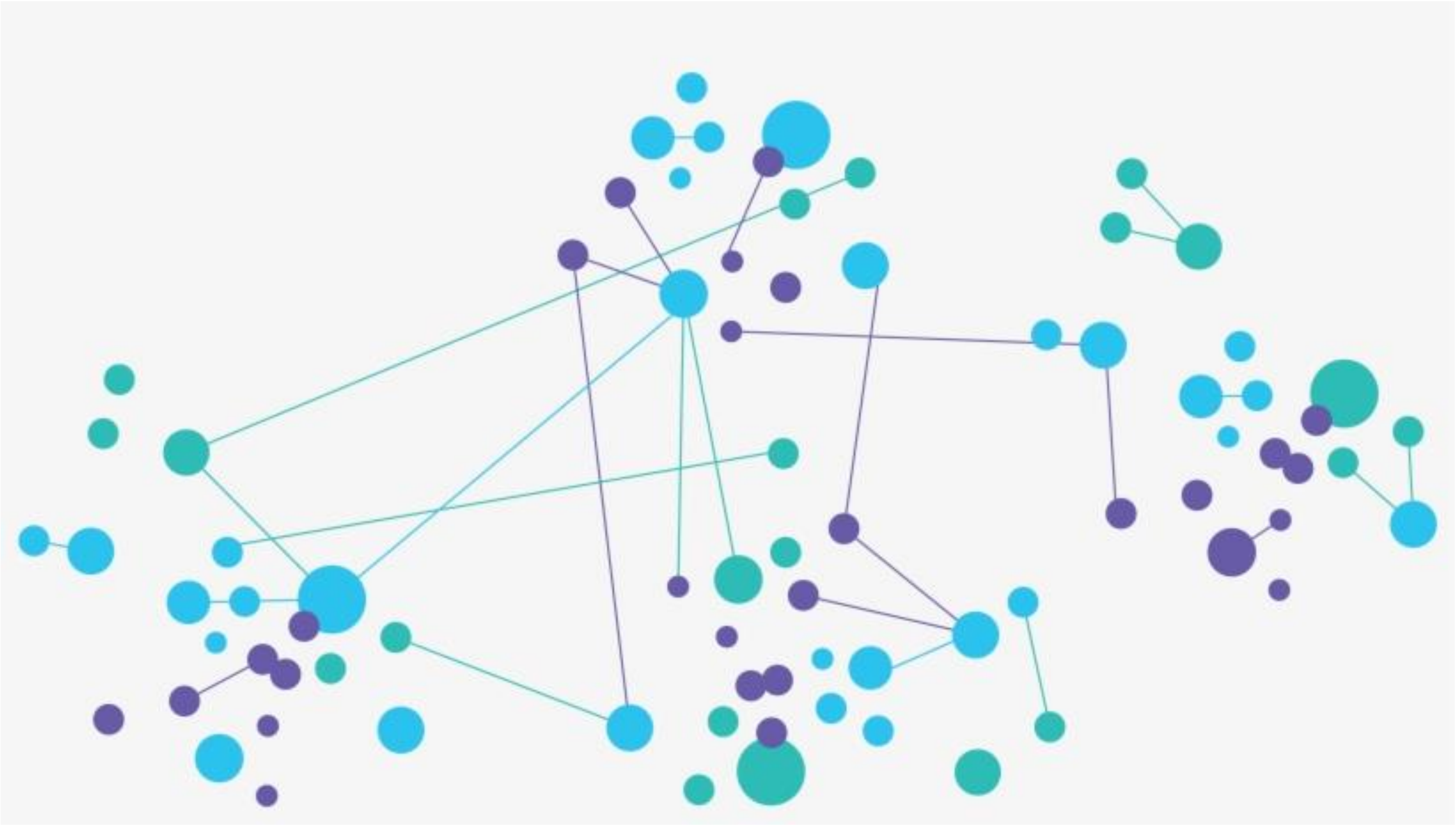
Variable	Coefficient	Std.Err.	t-Statistic	P-value	Lower95%	Upper95%	VIF	Std. Coeff.	Explanation
Constant	6.763	0.060	112.381	0.000	6.645	6.881	0.000	0.000	
Floor	0.007925	0.002825	2.805	0.005	0.002385	0.013	1.449	0.023	
Floor_1	-0.161	0.017	-9.313	0.000	-0.194	-0.127	1.389	-0.075	First floor discounted, nobody likes that people from outside can see through your windows
House_type.Eq.Classical	-0.224	0.042	-5.363	0.000	-0.306	-0.142	9.684	-0.115	Classical houses are cheaper overall because they are about 100 years old and not renovated
House_type.Eq.Neo_Classical	0.259	0.047	5.467	0.000	0.166	0.353	5.211	0.086	Renovated classical houses are in high regard as they usually require heavy investments
House_type.Eq.New	0.388	0.018	21.412	0.000	0.352	0.423	1.573	0.185	New houses command much higher price
House_type.Eq.Other	-0.064	0.020	-3.173	0.002	-0.103	-0.024	1.245	-0.024	
m2_.Ln	1.097	0.014	79.943	0.000	1.070	1.123	1.530	0.680	m2 has the highest predictive power
m2__200	-0.294	0.050	-5.904	0.000	-0.392	-0.196	1.133	-0.043	large m2 apartments are usually entire floors or rooftops that are in lower value than "actual" apar
Reg.Eq.City_center.Times.House_type.Eq.Class	0.253	0.047	5.371	0.000	0.161	0.345	10.324	0.119	Classical houses in the city center however are much more valued as they are made from bricks
Reg.Eq.City_center.Times.House_type.Eq.Neo_(	0.221	0.055	4.007	0.000	0.113	0.329	5.779	0.066	This could be greyed out as has similar finding as in the Neo_Classical case (most of renovated h
Reg.Eq.City_outskirts	-0.583	0.087	-6.703	0.000	-0.754	-0.413	1.049	-0.047	Further driving time to city center reduced the value
Reg.Eq.Criminal__run_down_area	-0.839	0.057	-14.691	0.000	-0.951	-0.727	1.194	-0.110	Criminal areas are most potent in negative pricing development
Reg.Eq.historic_neighbourhood	-0.214	0.028	-7.555	0.000	-0.269	-0.158	1.992	-0.073	Historic neighbourhood has the lowest price discount compared to city center among other region
Reg.Eq.New_area	-0.425	0.064	-6.627	0.000	-0.551	-0.300	1.099	-0.048	This is interesting and hard to explain, maybe because those new areas are located far from the c
Reg.Eq.Private_residences_within_city_limits	0.105	0.043	2.438	0.015	0.021	0.189	1.225	0.019	Private houses are in higher value, small number of them make low impact than it could potentially
Reg.Eq.Seaside_region	-0.521	0.047	-11.163	0.000	-0.612	-0.429	1.188	-0.084	This is interesting finding, most likely due to considered "cheap" seaside area (industrial area) cor
Reg.Eq.Soviet_built_region	-0.366	0.021	-17.207	0.000	-0.407	-0.324	3.277	-0.214	Soviet built region typically has social houses that are in low regarded value
Reg.Eq.Soviet_built_region.Times.House_type.E	-0.126	0.054	-2.319	0.020	-0.232	-0.019	2.375	-0.025	

Source	Deg. Freedom	Sum Squares	Mean Square	F-Statistic	P-value
Regression	18	1 944	107.994	1002.028	0.000
Residual	3124	336.692	0.108		
Total	3142	2 281			

	Mean Error	RMSE	MAE	Minimum	Maximum	MAPE	J-B stat
Fitted (n=3143)	0.000	0.327	0.236	-2.249	1.183	2.1%	:371.97 (P=0.000)







STOCKHOLM SCHOOL OF ECONOMICS

SSE RIGA