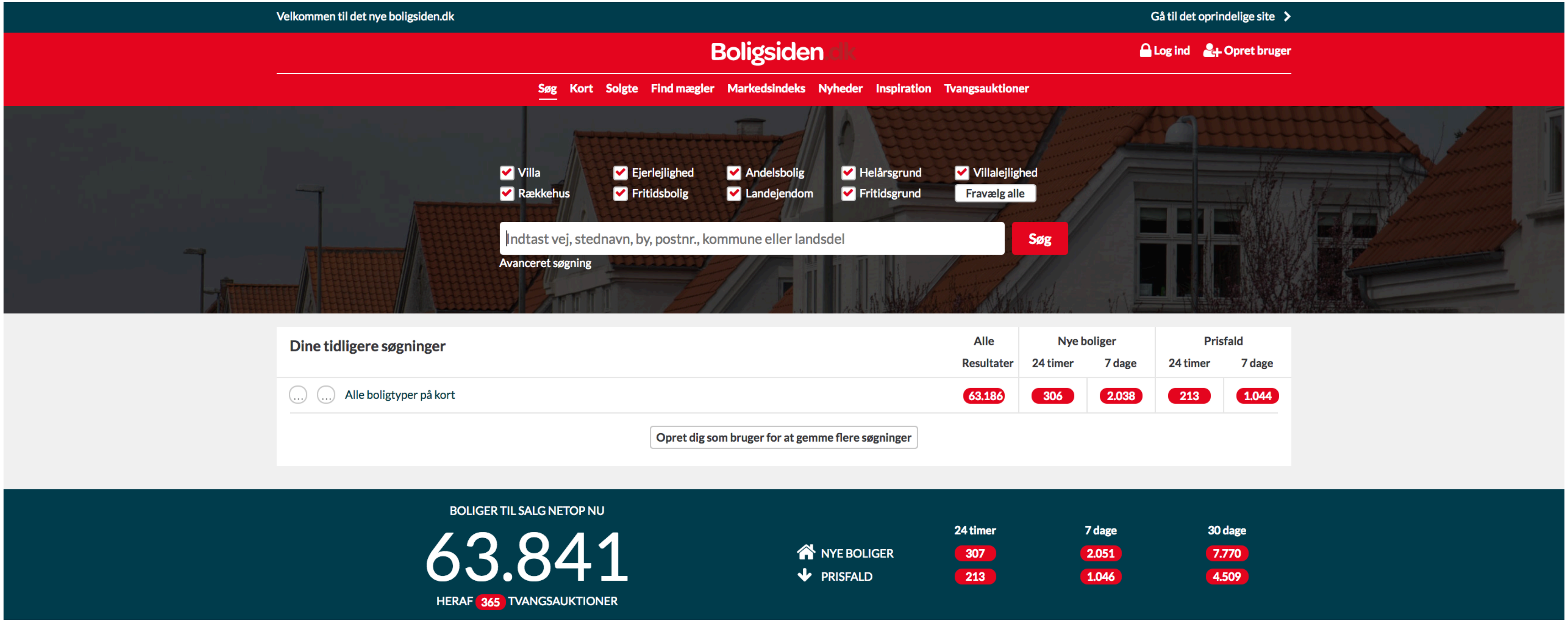
Boligsiden.dk

Forecasting House Prices

Agenda

- Background info
- Feature inspection and selecting based on EDA
- Multiple regression analysis
- OLS residual diagnostics
- Transformed model inference

Boligsiden.dk is a site owned by the real estate agencies in Denmark with the objective to make it easier for potential buyers to find their dream home



By scraping information of houses sold in the last 5 years in the city of Aarhus, we can start exploring what factors that can predict and explain house prices

Scraped features

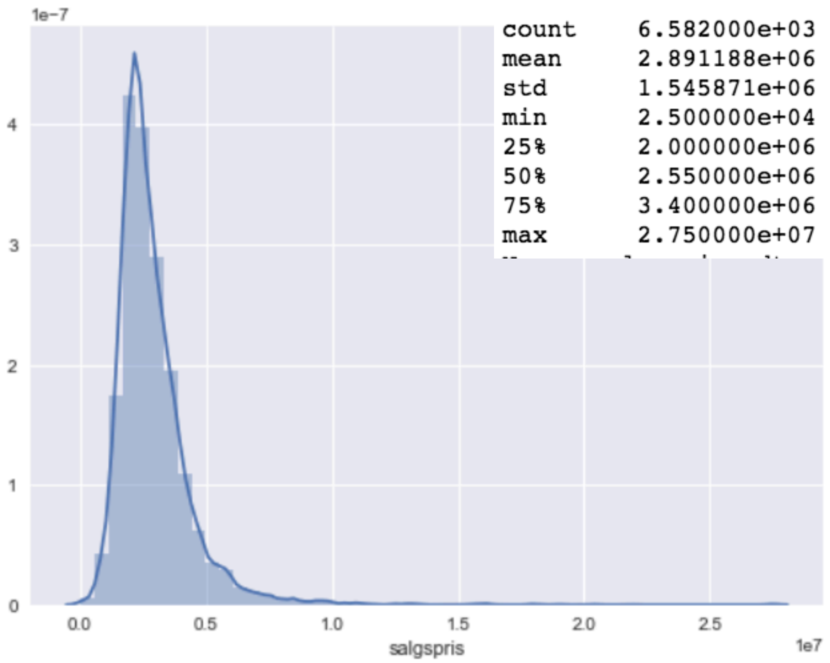
- Sales price
- Type of house
- Number of rooms
- Zip code
- Address
- Energy tag
- Property size
- Public property value
- Days on market
- Number of floors
- Year built
- Heating
- Exterior material

Plotting the distribution for sales price clearly indicate positive skewness and a overall non-symmetric uniform distribution

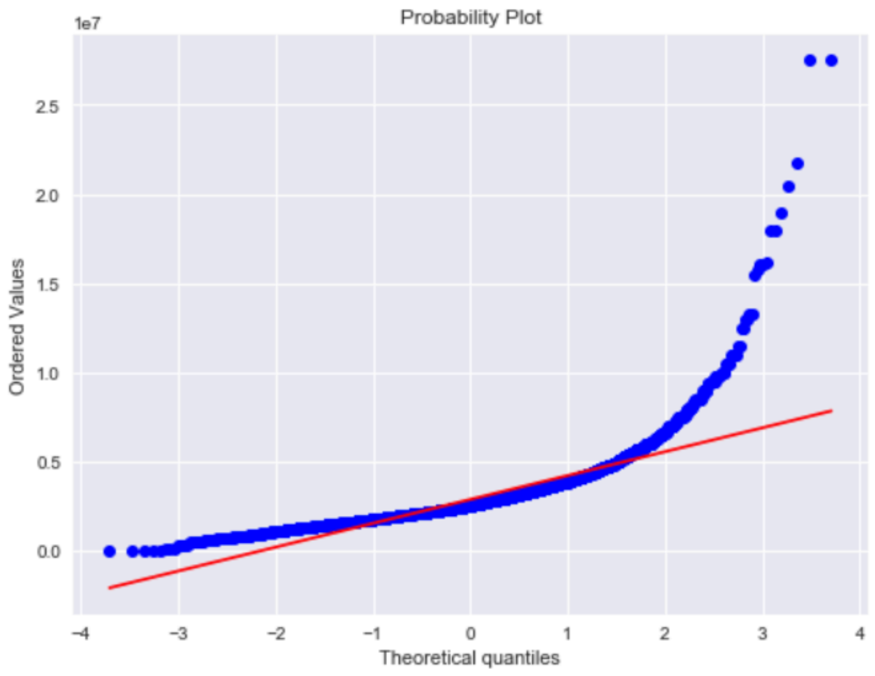
Analyzing target variable

Histogram

```
Skewness: 4.025724
Kurtosis: 37.236864
```



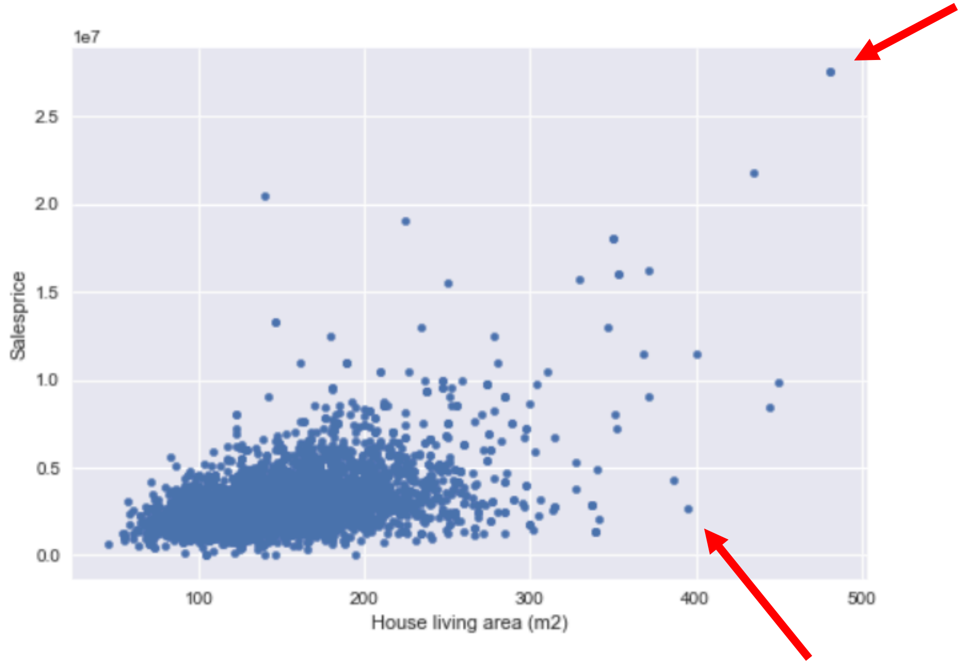
Probability plot



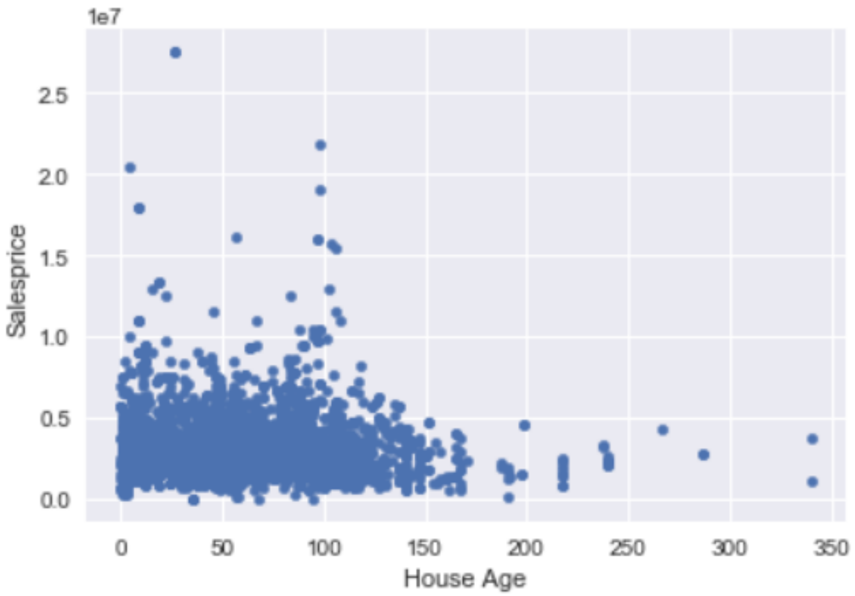
There seems to be a positive linear relationship between sales price and house living area.
However, the newer houses doesn’t visually seem to influence sales price more than we would think

Relationship with numerical features

Sales price vs. house living area



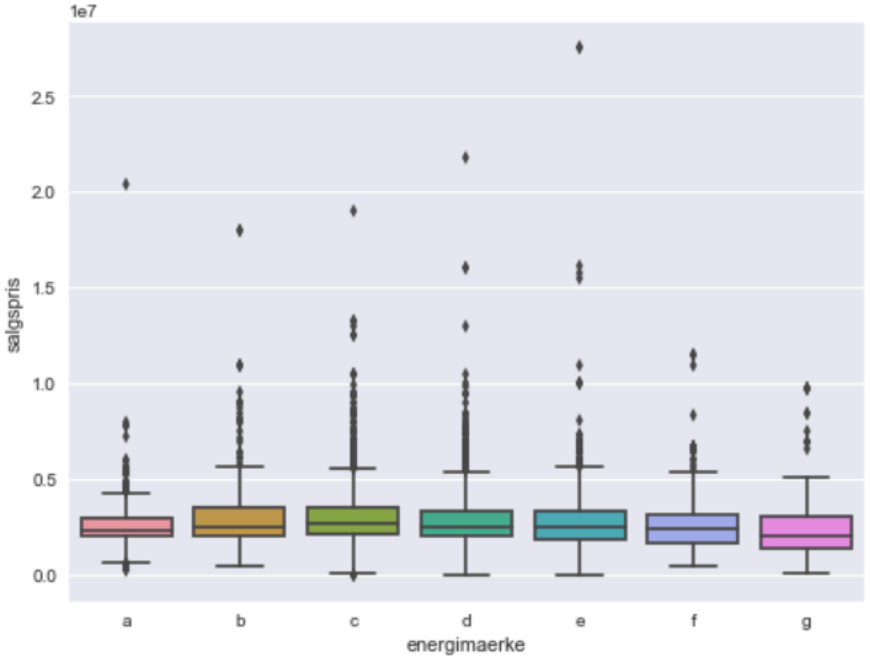
Sales price vs. Year Built



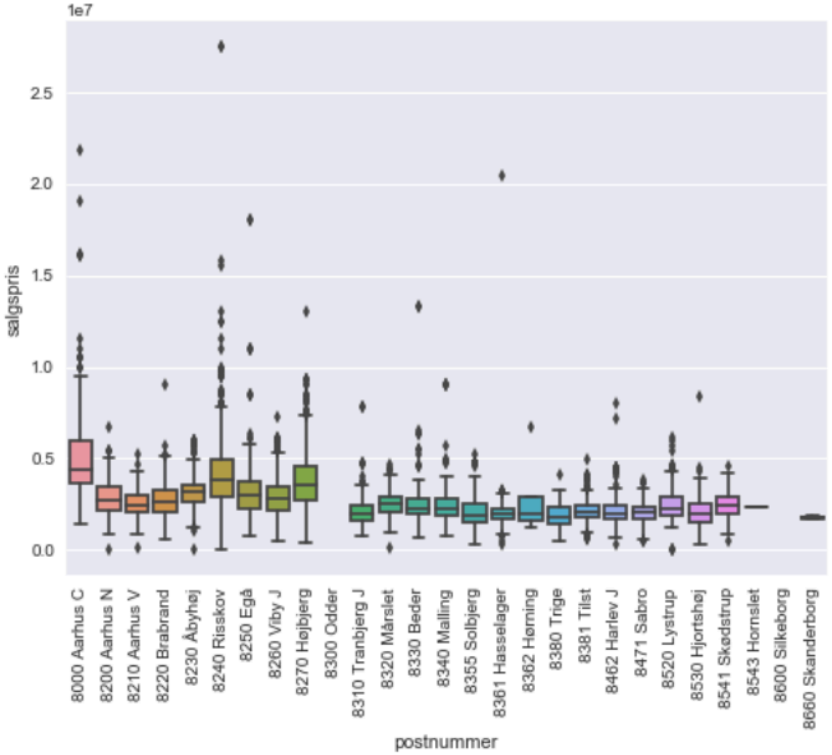
The initial though that the better the house is isolated the higher the price doesn’t seem to hold. However, sales prices seem to vary across zip codes

Relationship with categorical features

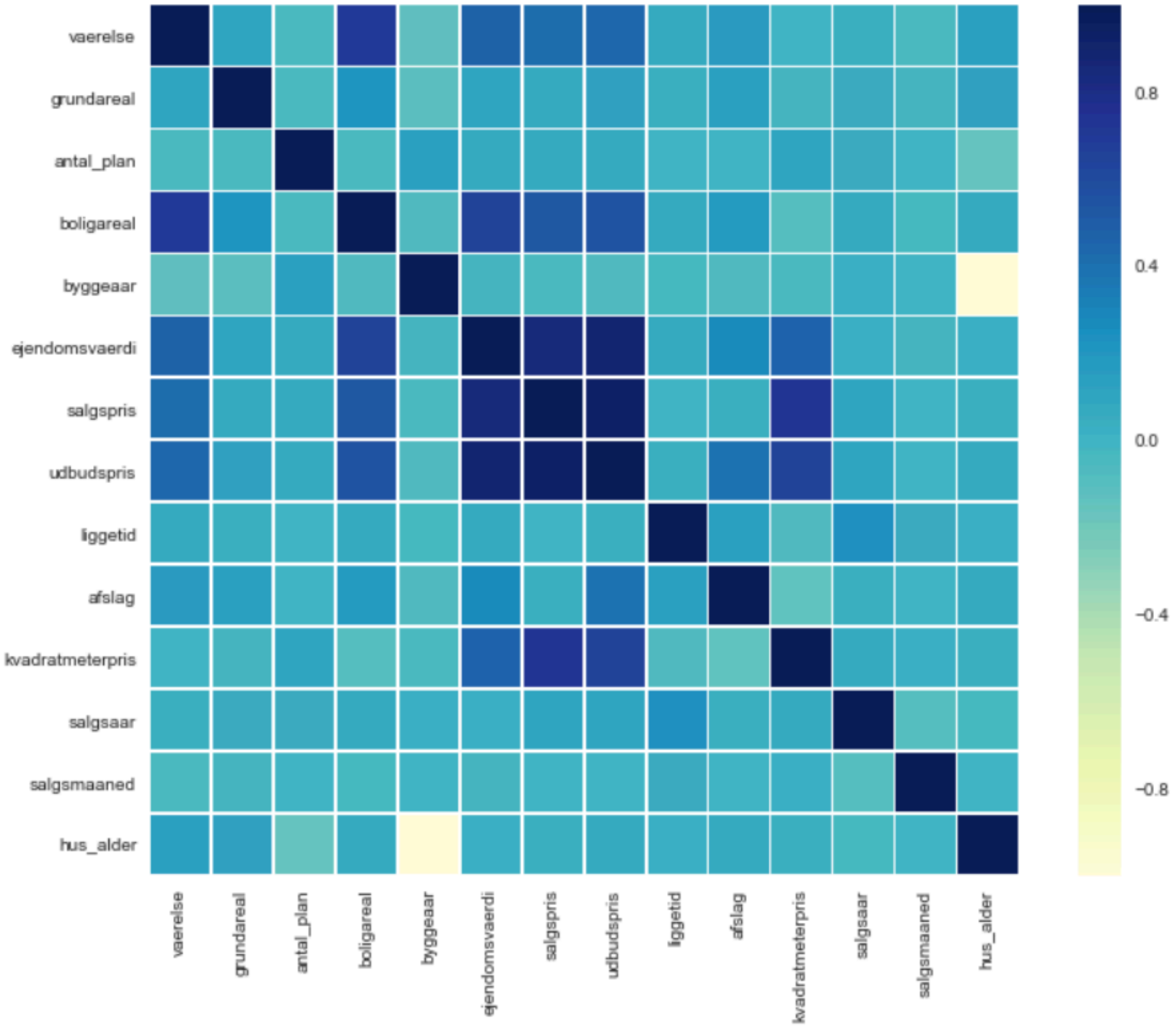
Sales price vs. energy tag



Sales price vs. House Age



The heat map correlation matrix gives us an estimate of the relationship between continuous variables and thus useful for feature selection



Features to test

- House living area
- Number of rooms
- Property size
- Zip code

Although house living area and number of rooms are significant, they don’t contribute that much to explain the variability in sales price

Multiple regression with 2 predictors

Dep. Variable:	salgspris	**R-squared:**	0.276
Model:	OLS	**Adj. R-squared:**	0.276
Method:	Least Squares	**F-statistic:**	1253.
Date:	Sun, 22 Oct 2017	**Prob (F-statistic):**	0.00
Time:	17:04:01	**Log-Likelihood:**	-1.0208e+05
No. Observations:	6582	**AIC:**	2.042e+05
Df Residuals:	6579	**BIC:**	2.042e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-5.697e+04	6.39e+04	-0.891	0.373	-1.82e+05	6.83e+04
boligareal	1.532e+04	519.381	29.488	0.000	1.43e+04	1.63e+04
vaerelse	1.444e+05	1.66e+04	8.711	0.000	1.12e+05	1.77e+05

Omnibus:	4812.833	**Durbin-Watson:**	1.856
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	233540.662
Skew:	2.989	**Prob(JB):**	0.00
Kurtosis:	31.562	**Cond. No.**	606.

From the coefficient estimates it seems like ”number of rooms” has a higher impact on sales price than “house living area”, but can we be certain about that?

By standardizing the variables we are able to compare variables based on the same scale. House living area actually has a higher impact on sales price than number of rooms

Standardized coefficients

Dep. Variable:	y	**R-squared:**	0.276
Model:	OLS	**Adj. R-squared:**	0.276
Method:	Least Squares	**F-statistic:**	1253.
Date:	Sun, 22 Oct 2017	**Prob (F-statistic):**	0.00
Time:	19:11:42	**Log-Likelihood:**	-8277.2
No. Observations:	6582	**AIC:**	1.656e+04
Df Residuals:	6580	**BIC:**	1.657e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.4293	0.015	29.490	0.000	0.401	0.458
x2	0.1268	0.015	8.711	0.000	0.098	0.155

Omnibus:	4812.833	**Durbin-Watson:**	1.856
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	233540.662
Skew:	2.989	**Prob(JB):**	0.00
Kurtosis:	31.562	**Cond. No.**	2.35

Standardized values will have mean 0 and var 1

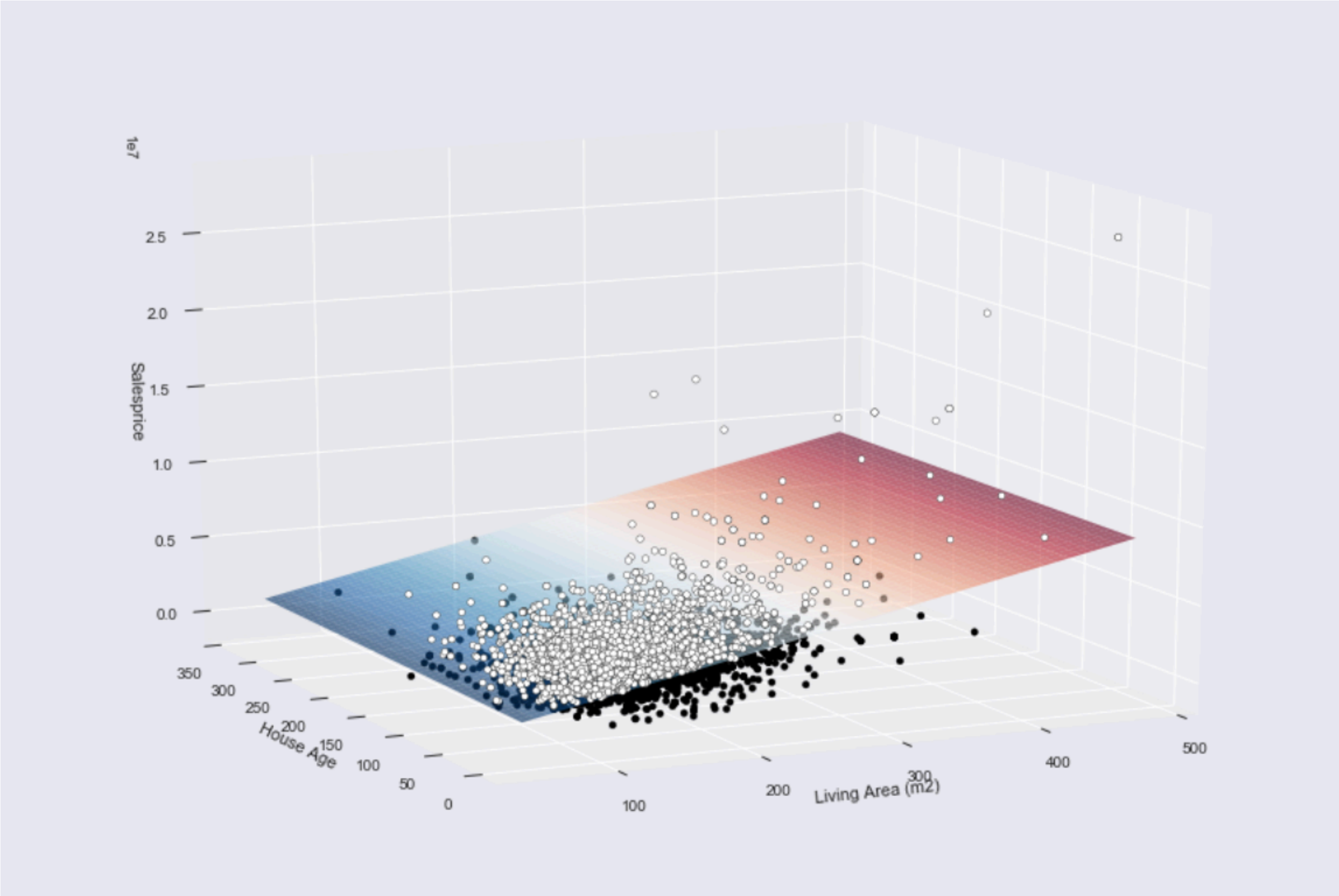
Alternatively manual adding predictor variables and inspect the change in the explanatory power could also be a solution

Constant equals zero and is therefore left out of the equation

The 3D plot too inspect the feature space between house living area and year built on sales price does not seem to provide that much insight

- The huge mass of observations are houses between 50 to 200 years old and have a square-meter size between 100-300

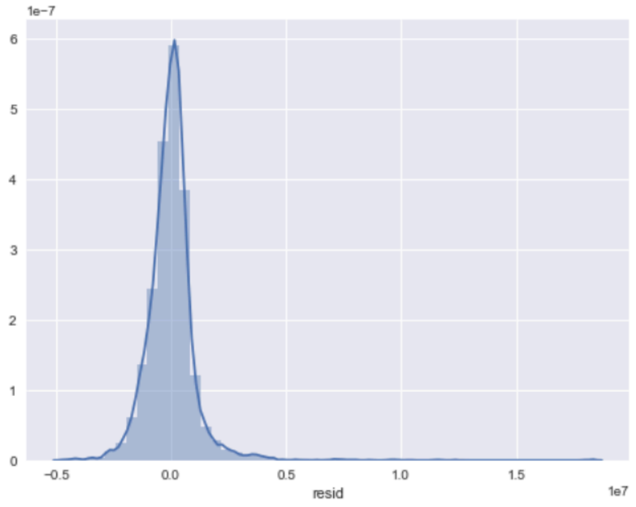
- The observations starts to spread as increases the three dimensions increase, thus indicating more fluctuation and potentially outliers



Residual diagnostics

OLS assumptions of residuals

- Normal distribution
- Homoscedasticity
- Independence
- Linearity



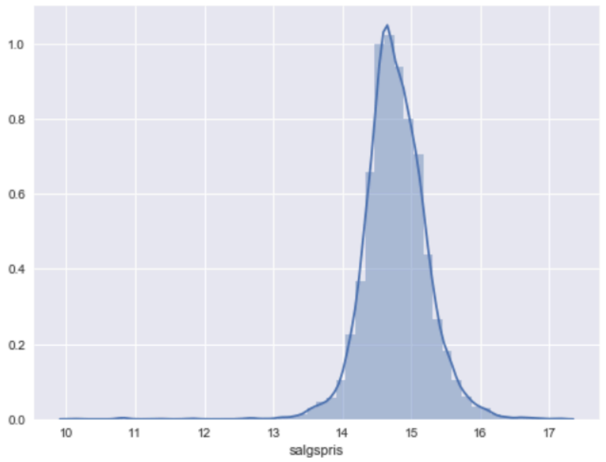
Clearly the normality assumptions is violated and the distribution of the residuals shares same characteristics as the target variable

Also heteroskedasticity seems to appear, since the residuals are more spread when the fitted values increase on the x-axis.

Statistical output after log transformation

Log transformation

```
Skewness: 4.025724
Kurtosis: 37.236864
```



Why are the dummy coefficients all negative?

How do we interpret the coefficients with the log transformation?

Estimation of an artificial situation:
- Property size = 500
- House living area = 200
- Number of rooms = 4
- Zip code = 8200 Aarhus N

14.60 + 1.272e-06 * 500 + 0.0038 * 200 + 0.033 * 4 - 0.4535 * 1 = exp(15) = **3.269.00 DKK**

Multiple regression with transformed target and dummy variables

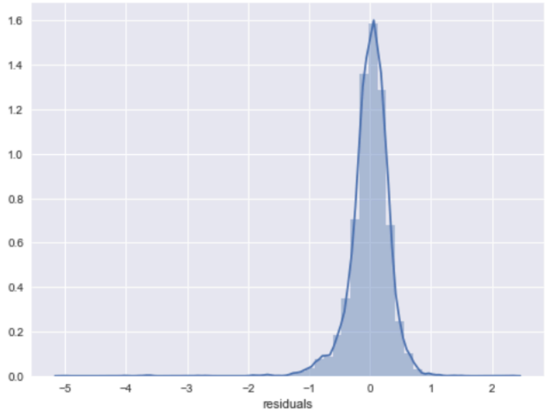
Dep. Variable:	salgspris	**R-squared:**	0.463
Model:	OLS	**Adj. R-squared:**	0.461
Method:	Least Squares	**F-statistic:**	209.5
Date:	Mon, 23 Oct 2017	**Prob (F-statistic):**	0.00
Time:	20:41:47	**Log-Likelihood:**	-2117.9
No. Observations:	6582	**AIC:**	4292.
Df Residuals:	6554	**BIC:**	4482.
Df Model:	27		
Covariance Type:	nonrobust		

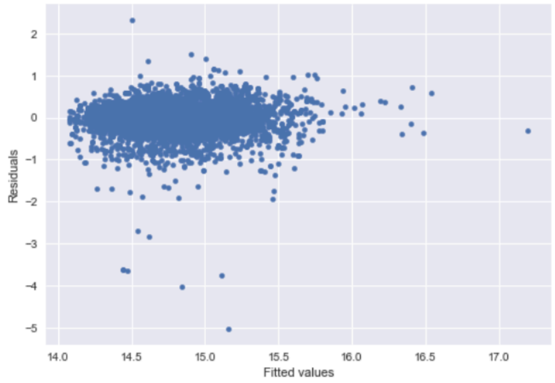
	coef	std err	t	P>|t|	[0.025	0.975]
const	14.5991	0.034	435.698	0.000	14.533	14.665
grundareal	1.272e-06	4.57e-07	2.783	0.005	3.76e-07	2.17e-06
boligareal	0.0038	0.000	27.521	0.000	0.003	0.004
vaerelse	0.0330	0.004	7.724	0.000	0.025	0.041
postnummer_8200 Aarhus N	-0.4535	0.035	-12.876	0.000	-0.523	-0.384
postnummer_8210 Aarhus V	-0.5412	0.033	-16.233	0.000	-0.607	-0.476
postnummer_8220 Brabrand	-0.5376	0.033	-16.363	0.000	-0.602	-0.473

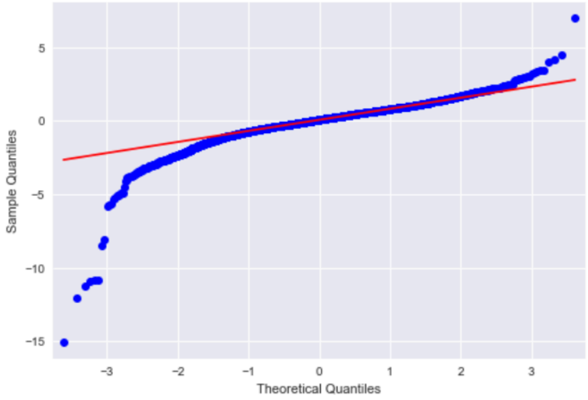
Residual diagnostics part 2

OLS assumptions of residuals

- Normal distribution
- Homoscedasticity
- Independence
- Linearity







- The log transformation seems to move the distribution closer to a normal distribution, however with a left-tail
- Visually, it seems like we removed the heteroskedasticity, but a formal test could be conducted to make a final conclusion. Also, no systematic pattern appears, so no autocorrelation.
- S shape of the qq-plot indicates skewness in the distribution and also that it has a tail, which confirmed the histogram

What are the next steps?

- Residual diagnostics based on statistical tests
- Implementation of non-linear models to capture more flexible patterns in the data
- Collection of more observations
- Gathering of more variables to implement in the model
- Prediction and accuracy measure on training and test data