How to find your perfect home or an investment opportunity

CONTENTS OF THIS PRESENTATION

Why you should choose us for your real estate purchase:

OLS Regression and Sklearn Linear Regression Model

- Based on the 'King County House Price Dataset'
- Contains 21 different variables (e.g. price, number of bedrooms,...) for 21,597 houses
- The dataset was splitted into a train and a test set
- Aim was to find the best predictors for the house price

➔ **Found 6 predictors to work best from the original variables and added 2 more.**

OLS Regression Results

Dep. Variable:	log_price	**R-squared:**	0.745
Model:	OLS	**Adj. R-squared:**	0.745
Method:	Least Squares	**F-statistic:**	9001.
Date:	Tue, 09 Jun 2020	**Prob (F-statistic):**	0.00
Time:	20:27:50	**Log-Likelihood:**	-2044.4
No. Observations:	21596	**AIC:**	4105.
Df Residuals:	21588	**BIC:**	4169.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-29.0305	0.461	-62.971	0.000	-29.934	-28.127
bathrooms	0.0907	0.004	22.420	0.000	0.083	0.099
sqft_living	0.0001	4.11e-06	34.354	0.000	0.000	0.000
grade	0.1775	0.003	65.410	0.000	0.172	0.183
yr_built	-0.0041	7.98e-05	-51.404	0.000	-0.004	-0.004
lat	1.3176	0.014	97.279	0.000	1.291	1.344
sqft_living15	0.0001	4.29e-06	23.834	0.000	9.39e-05	0.000
reno_yes	-14.4783	0.230	-62.904	0.000	-14.929	-14.027
reno_no	-14.5522	0.231	-63.005	0.000	-15.005	-14.100

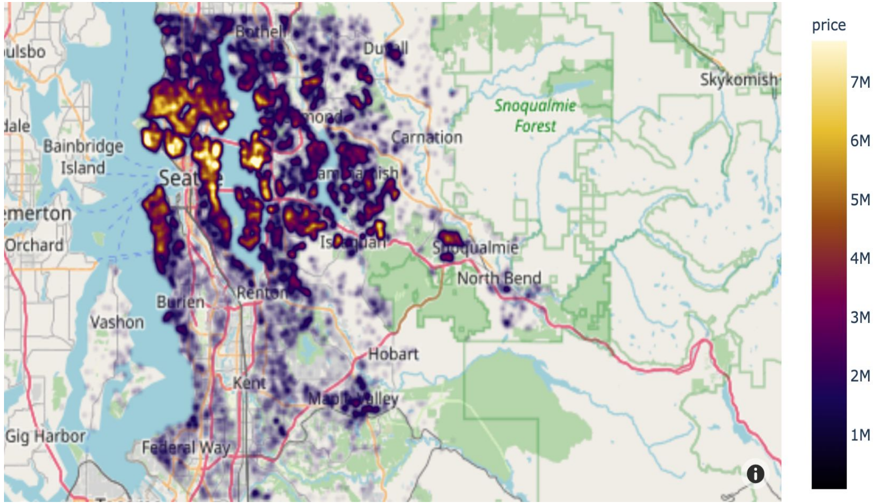
Omnibus:	409.895	**Durbin-Watson:**	1.976
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	844.156
Skew:	0.063	**Prob(JB):**	4.94e-184
Kurtosis:	3.960	**Cond. No.**	2.41e+19

The Price Situation

As you can see in the heatmap, the prices are **varying strongly**. The **highest prices** are located in the **urban areas** especially in Seattle and are declining towards the **outskirts** and **rural areas**.

- Owning a home equal to roughly 4x – 6x your household income puts you in the middle class.
- In Seattle the median income in 2015 was 80000 $ per year and household.

➔ Middle class houses range from 320000 to 480000 $.

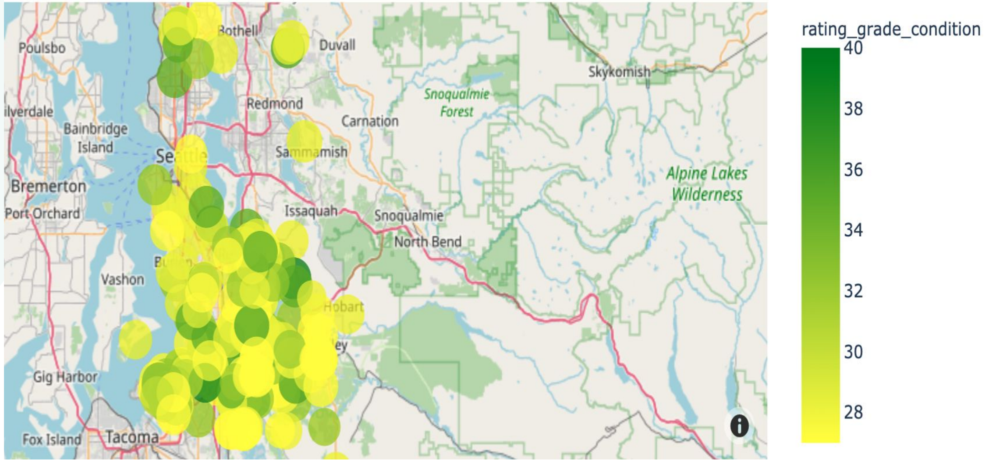


*https://www.financialsamurai.com/the-ideal-house-size-and-layout-to-raise-a-family/

First Case: A family of four

The ideal house for a **family of four** should have at least a square footage of **2200** or more*, **4 bedrooms,** optimally a **2nd bathroom**. It should have the **same size** of the surrounding houses and should be in a **good condition.**

The number of houses with matching features was: 258

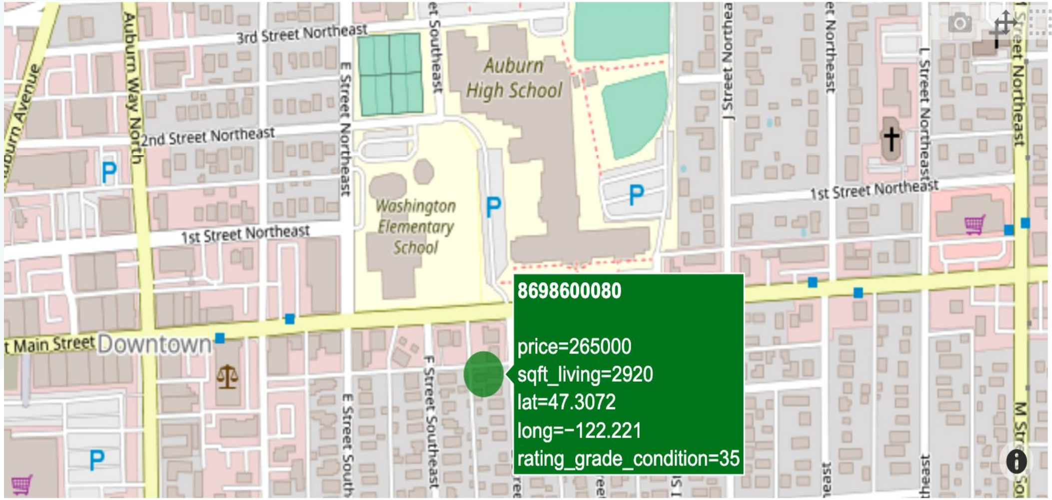


*https://saterdesign.com/blogs/news/living-area-explained

First Case: A family of four (Part1)

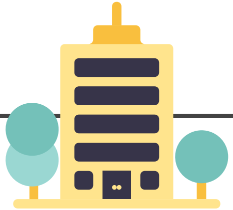
So the data was filtered again and sorted the results after the price.
We found a house for **265000 $** with **2920 sqft of living space** and additional **5250 sqft of land space**.

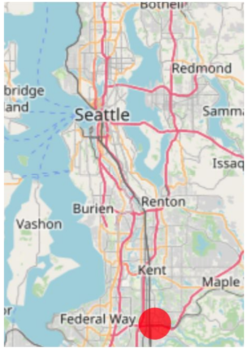
On top there is an Elementary and a High School within the neighborhood.

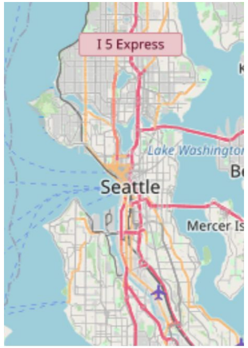


BUT! WHAT IF?

What if you don't want to live in the outskirts but in the city?



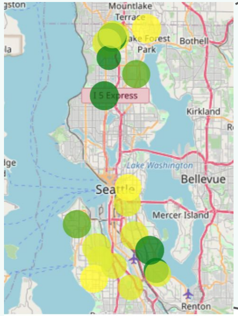


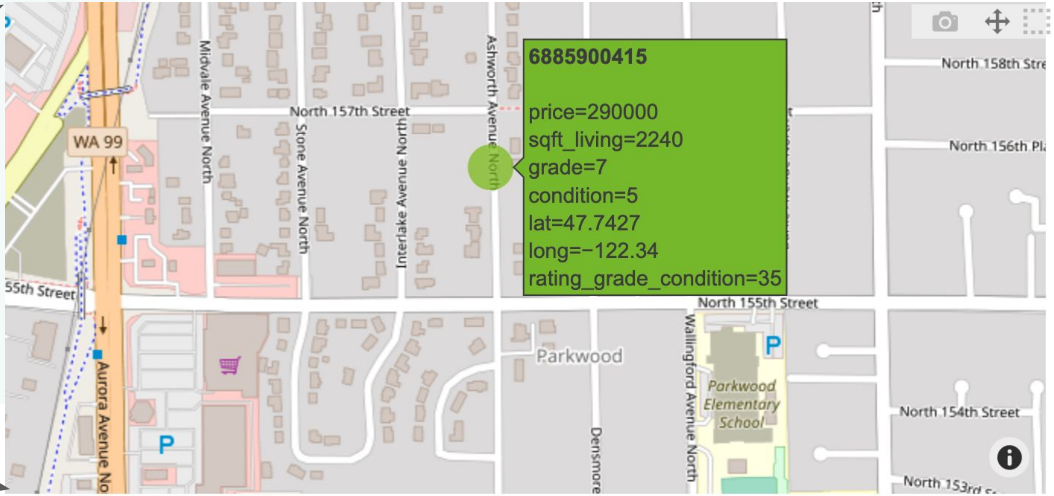


First Case: A family of four (Part2)

An additional filter was added to look exclusively at houses in Seattle. We found a house for **290000.00 $** with **2240 sqft of living space** and additional **8162 sqft of land space**.

There is an Elementary School within the neighborhood as well.



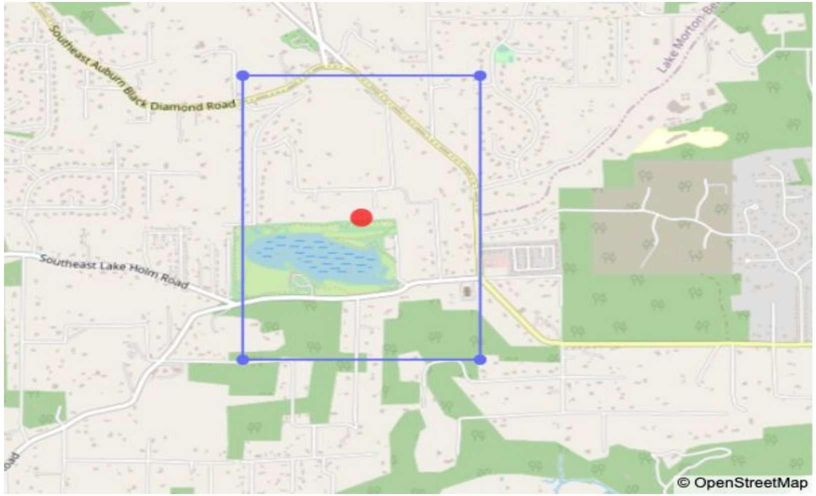


Second Case: Investing in real estate

This time we looked for **cheap houses** with **average feature values** and **bad conditions** that have **not been renovated**.

- Based on our prediction **6 houses** were identified as potential investments.
- Additionally we compared the **price of the house** with the **average house price** of a defined area.

➔ The most lucrative house would cost **335,000 $** and is **108,277.78 $ cheaper** than the average house price in the area (9 houses located in the area).



Outlook

Further Optimisation
$R^2 = 0.745$ could be better

Using other Models
Testing if other models lead to a better predictions

More business ideas
Coming up with more business ideas

THANKS

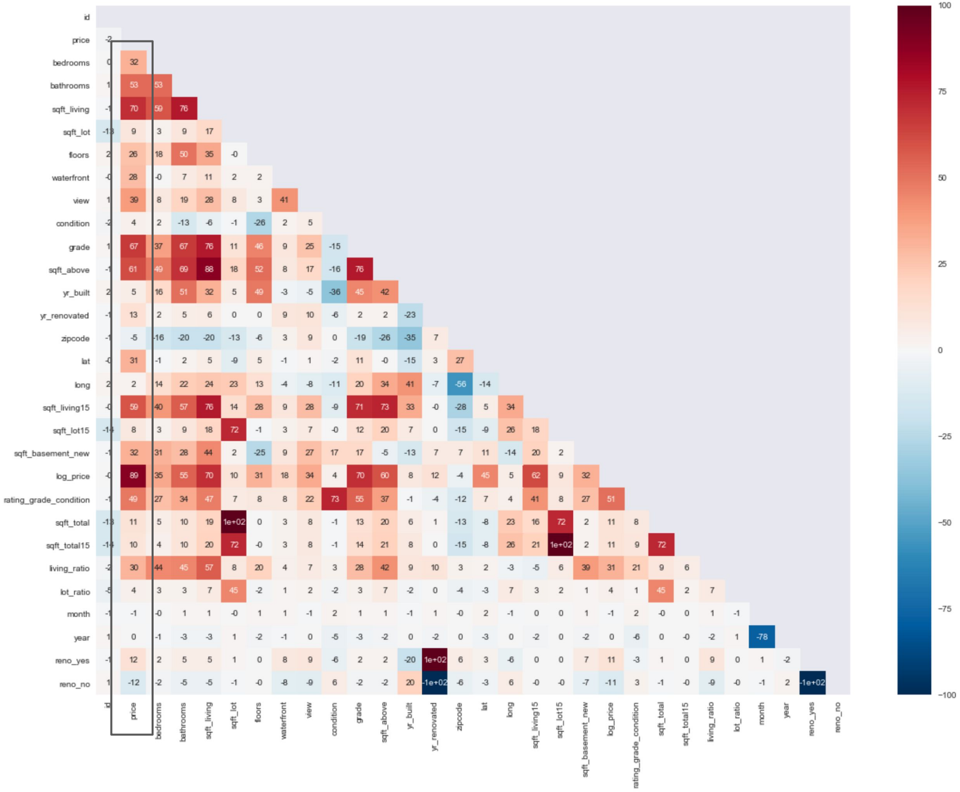
Does anyone have any questions?

eike.rogall@gmail.com
github.com/rogall-e
neuefische.de

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.

Please keep this slide for attribution.

Appendix



Correlation heatmap:

- the darker the color the stronger the correlation
- Red: positiv correlation
- Blue: negativ correlation

Appendix

OLS Regression Results

Dep. Variable:	log_price	**R-squared:**	0.771
Model:	OLS	**Adj. R-squared:**	0.771
Method:	Least Squares	**F-statistic:**	3032.
Date:	Wed, 10 Jun 2020	**Prob (F-statistic):**	0.00
Time:	10:22:38	**Log-Likelihood:**	-859.03
No. Observations:	21596	**AIC:**	1768.
Df Residuals:	21571	**BIC:**	1968.
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-87.2950	8.315	-10.498	0.000	-103.594	-70.996
bedrooms	-0.0219	0.003	-8.706	0.000	-0.027	-0.017
bathrooms	0.0680	0.004	16.584	0.000	0.060	0.076
sqft_living	3.173e-05	4.2e-06	7.549	0.000	2.35e-05	4e-05
sqft_lot	-1.572e-05	2.1e-06	-7.475	0.000	-1.98e-05	-1.16e-05
floors	0.0738	0.005	16.308	0.000	0.065	0.083
view	0.0774	0.002	31.114	0.000	0.073	0.082
grade	0.1233	0.009	13.984	0.000	0.106	0.141
sqft_above	1.125e-05	3.26e-06	3.452	0.001	4.86e-06	1.76e-05
yr_built	-0.0035	9.12e-05	-38.124	0.000	-0.004	-0.003
zipcode	-0.0007	4.15e-05	-16.414	0.000	-0.001	-0.001
lat	1.3967	0.013	103.776	0.000	1.370	1.423
long	-0.1636	0.016	-9.921	0.000	-0.196	-0.131
sqft_living15	0.0001	6.15e-06	20.594	0.000	0.000	0.000
sqft_lot15	-6.368e-05	3.08e-06	-20.674	0.000	-6.97e-05	-5.76e-05
sqft_basement_new	2.052e-05	3.64e-06	5.640	0.000	1.34e-05	2.77e-05
rating_grade_condition	0.0102	0.002	4.088	0.000	0.005	0.015
sqft_total	1.625e-05	2.1e-06	7.734	0.000	1.21e-05	2.04e-05
sqft_total15	6.331e-05	3.07e-06	20.598	0.000	5.73e-05	6.93e-05
living_ratio	0.1801	0.015	11.645	0.000	0.150	0.210
lot_ratio	-0.0024	0.002	-1.386	0.166	-0.006	0.001
month	0.0025	0.001	2.796	0.005	0.001	0.004
year	0.0641	0.006	10.851	0.000	0.052	0.076
reno_yes	-43.6074	4.157	-10.489	0.000	-51.756	-35.458
reno_no	-43.6876	4.158	-10.507	0.000	-51.837	-35.538
condition_2	0.0334	0.052	0.639	0.523	-0.069	0.136
condition_3	0.0959	0.055	1.747	0.081	-0.012	0.204
condition_4	0.0763	0.066	1.149	0.251	-0.054	0.207
condition_5	0.0723	0.080	0.900	0.368	-0.085	0.230

Omnibus:	333.316	**Durbin-Watson:**	1.987
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	651.740
Skew:	-0.014	**Prob(JB):**	3.00e-142
Kurtosis:	3.851	**Cond. No.**	1.03e+16

OLS Regression Results

Dep. Variable:	log_price	**R-squared:**	0.745
Model:	OLS	**Adj. R-squared:**	0.745
Method:	Least Squares	**F-statistic:**	9001.
Date:	Tue, 09 Jun 2020	**Prob (F-statistic):**	0.00
Time:	20:27:50	**Log-Likelihood:**	-2044.4
No. Observations:	21596	**AIC:**	4105.
Df Residuals:	21588	**BIC:**	4169.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-29.0305	0.461	-62.971	0.000	-29.934	-28.127
bathrooms	0.0907	0.004	22.420	0.000	0.083	0.099
sqft_living	0.0001	4.11e-06	34.354	0.000	0.000	0.000
grade	0.1775	0.003	65.410	0.000	0.172	0.183
yr_built	-0.0041	7.98e-05	-51.404	0.000	-0.004	-0.004
lat	1.3176	0.014	97.279	0.000	1.291	1.344
sqft_living15	0.0001	4.29e-06	23.834	0.000	9.39e-05	0.000
reno_yes	-14.4783	0.230	-62.904	0.000	-14.929	-14.027
reno_no	-14.5522	0.231	-63.005	0.000	-15.005	-14.100

Omnibus:	409.895	**Durbin-Watson:**	1.976
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	844.156
Skew:	0.063	**Prob(JB):**	4.94e-184
Kurtosis:	3.960	**Cond. No.**	2.41e+19

The Data of the two Models:

1. On the left the OLS Regression Model with all features:
   a. $R^2 = 0.771$
   b. Some features with high p-values
2. On the right the OLS Regression Model with the strongest features:
   a. $R^2 = 0.745$
3. On the bottom the Sklearn Model after splitting the data in Test and Train sets:
   a. $R^2 = 0.745$
   b. RSME = 0.27

```
Root Mean squared error (RMSE): 0.27
R**2: 0.75
```