Pricing of Real Estate Apartments (Macro Factors)

Analytics Project

Submitted by: Alcala Osmar, Choudhury Saarthak

July 1st, 2020

Team

Saarthak Choudhury

403962

Osmar Alcalá

403966

Agenda

Introduction

Problem Statement and Hypothesis

Methodology

Analytical Framework

Results

Introduction

Munich is the city with the highest rents, followed by Frankfurt and Stuttgart

Most Germans live in multi-family houses with up to ten apartments. Roughly one quarter live in large housing blocks or high-rise buildings and one third in single-family homes.

Statistically, each household consists of two people

Depending on region, rental costs amount to between one quarter and one third of monthly income

54% of Germans live in rented accommodation – more than in any other country in Europe. Only roughly 46% own a house or apartment

Problem Statement and Hypothesis

Problem Statement

Based on Macro- Factors how are the Real Estates in Germany Priced?

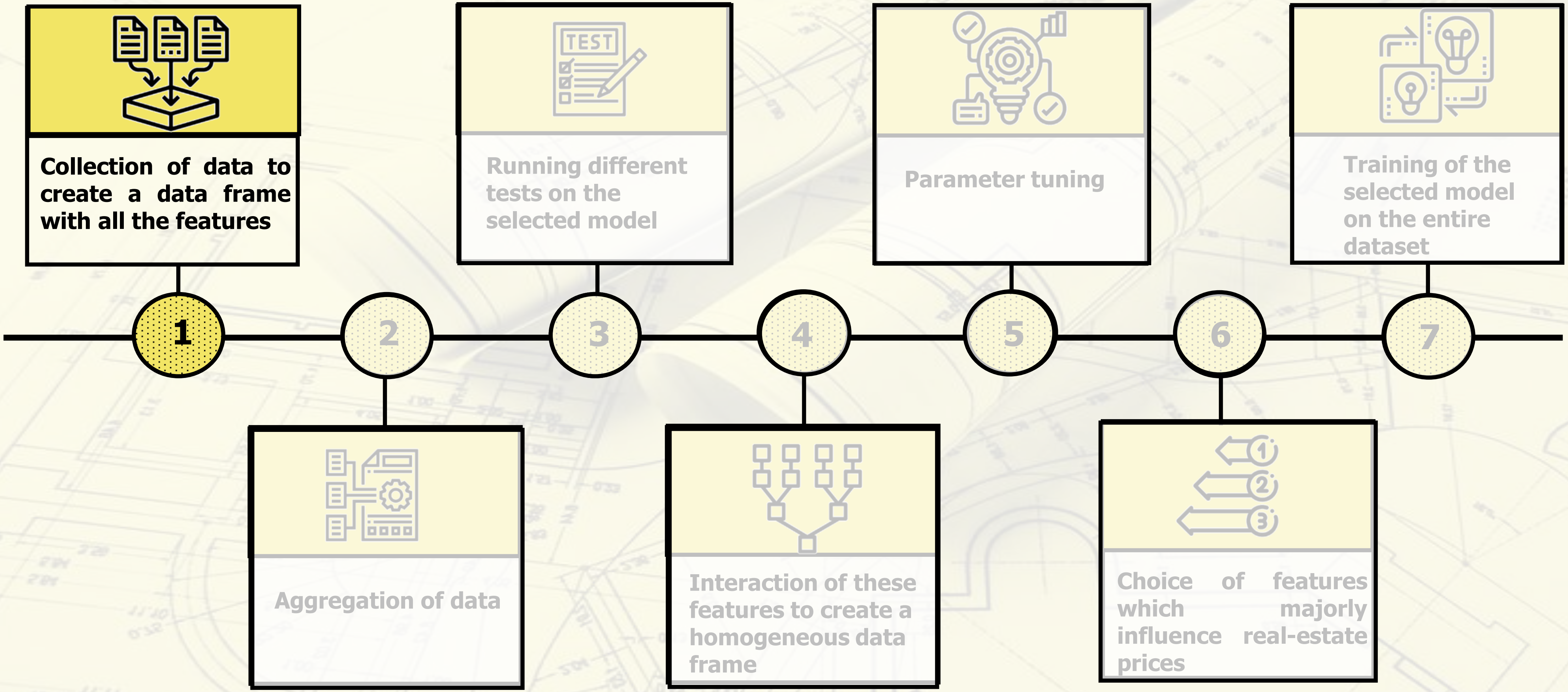
Hypothesis

There is a relationship between the macro factors proper to every zip code and the access to facilities to the price

Methodology

1. Collection of data to create a data frame with all the features
2. Aggregation of data
3. Running different tests on the selected model
4. Interaction of these features to create a homogeneous data frame
5. Parameter tuning
6. Choice of features which majorly influence real-estate prices
7. Training of the selected model on the entire dataset

Methodology



Main Data files

Master_data
File with the macroeconomic factors on district level

Zuordnung_plz_ort

File with postal code to city and bundesland mapping

Price

File with the pricing, obtained with web Scrapping

OSM
File with the facilities on a postal code level, obtained from Open Street Map

Plz_einwohner

Population assigned to each postal code

Plz-gebiete

Shapefile with Germany postal codes polygons.

Collection and Filtration of Data

Criteria for consideration of properties

- **Properties built after 2005**
- **Only properties with**
  - **- atmost six rooms**
  - **- price more than €10000**
- **Types of properties - Single Family House, Multi Family House, Semi-Detached House and Mid- Terrace House**

```
data["obj_yearConstructed"] = data["obj_yearConstructed"].astype(float)

x = data.copy()
x = x[x["obj_yearConstructed"] >= 2005]

x = x[x["obj_noRooms"] <= 6]
x = x[x["obj_purchasePrice"] >= 10000]
x = x[x["obj_purchasePrice"] < x["obj_purchasePrice"].quantile(0.99) ]


btype =["single_family_house","multi_family_house","semidetached_house","mid_terrace_house"]

x = x[x["obj_buildingType"].isin(btype)]
x['geo_plz'] = x['geo_plz'].astype(int)
x['geo_plz'] = x['geo_plz'].astype(str)
x['geo_plz'] = x['geo_plz'].apply(lambda x: x.zfill(5))
x['plz'] = x['geo_plz']
```

With further filtration we generated germany_df3 as our primary data frame which does not contain any NaN vaules or any duplicacies.

Methodology

1. Collection of data to create a data frame with all the features
2. **Aggregation of data**
3. Running different tests on the selected model
4. Interaction of these features to create a homogeneous data frame
5. Parameter tuning
6. Choice of features which majorly influence real-estate prices
7. Training of the selected model on the entire dataset

Aggregation of data

Since a district area was too big, we decided to segment our data in a smaller division which is Postal Code.

Aim - Differentiating whether a city belongs to a east or west using the postal codes

To prove our hypothesis we have created a new feature which contains the Euclidian distance of the centre of every postal code the top 10 most populated cities in Germany.

With help of Overpass API we were able to scrap the amenities data from OpenStreetMap database and sorted them by postal codes

	plz	amenity	count
0	01099	cafe	35
1	01099	doctors	13
2	01099	fast_food	50
3	01099	restaurant	83
4	01108	doctors	1

Amenities such as Restaurants, Cafés, Doctors, Hospitals, etc. are few key-contributors in price determination

	plz	university	train_station	bus
9929	99998	NaN	NaN	12.0
10005	99996	NaN	NaN	1.0
10244	99994	NaN	NaN	3.0
10002	99991	NaN	NaN	5.0
9337	99988	NaN	NaN	6.0
9575	99986	NaN	NaN	12.0
9305	99976	NaN	NaN	24.0
9715	99974	NaN	NaN	78.0

Proiximity to Bus & Train Station are also supposed to be key factors in determing the price of the apartments but our results did not project such elasticity.

Downscaling

Question **Why is downscaling important?**

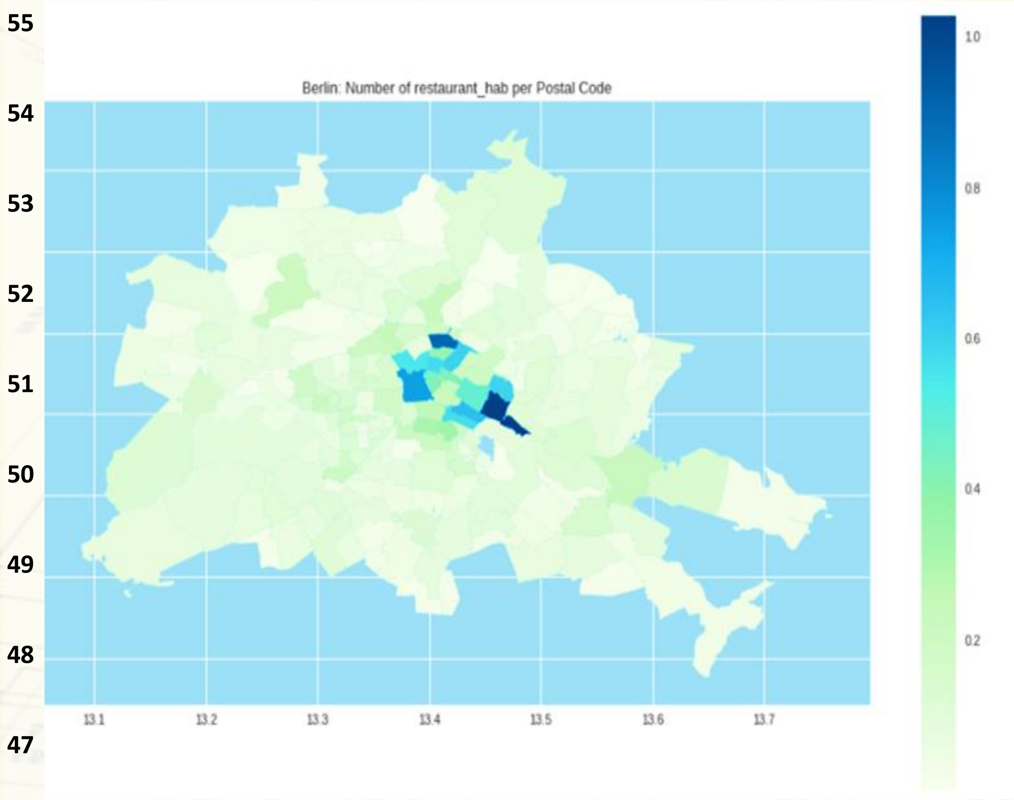
Since macro-economic factors are mainly found in on a district level, downscaling was necessary to project this features to a zip code level

$$\textit{Feature\_hab} = \sum_{i=1}^{n} \frac{habitant\_zipcode}{habitant\ district} \times \textit{feature}$$

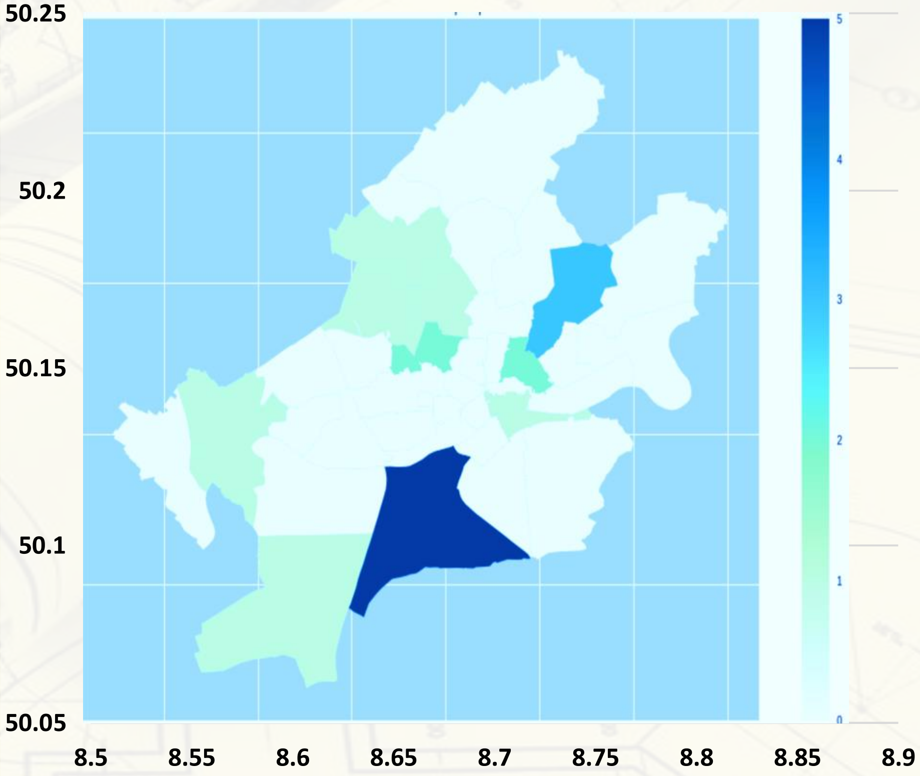
Where n = number of postal codes
habitant_zipcode = people living in a zip code
habitant district = people living in a district

Source "Germany property and metropolis market outlook 2019"

Projecting Amenities Country and City-Wise



Germany: Hospitals per Postal Code



Frankfurt Am Main: Hospital per Postal Code

Methodology

1. Collection of data to create a data frame with all the features
2. Aggregation of data
3. **Running different tests on the selected model**
4. Interaction of these features to create a homogeneous data frame
5. Parameter tuning
6. Choice of features which majorly influence real-estate prices
7. Training of the selected model on the entire dataset

Spatial Autocorrelation

The Spatial Autocorrelation (Global Moran's I) tool measures spatial autocorrelation based on both feature locations and feature values simultaneously. Given a set of features and an associated attribute, it evaluates whether the pattern expressed is clustered, dispersed, or random.

H_0

Null Hypothesis For Moran’s I

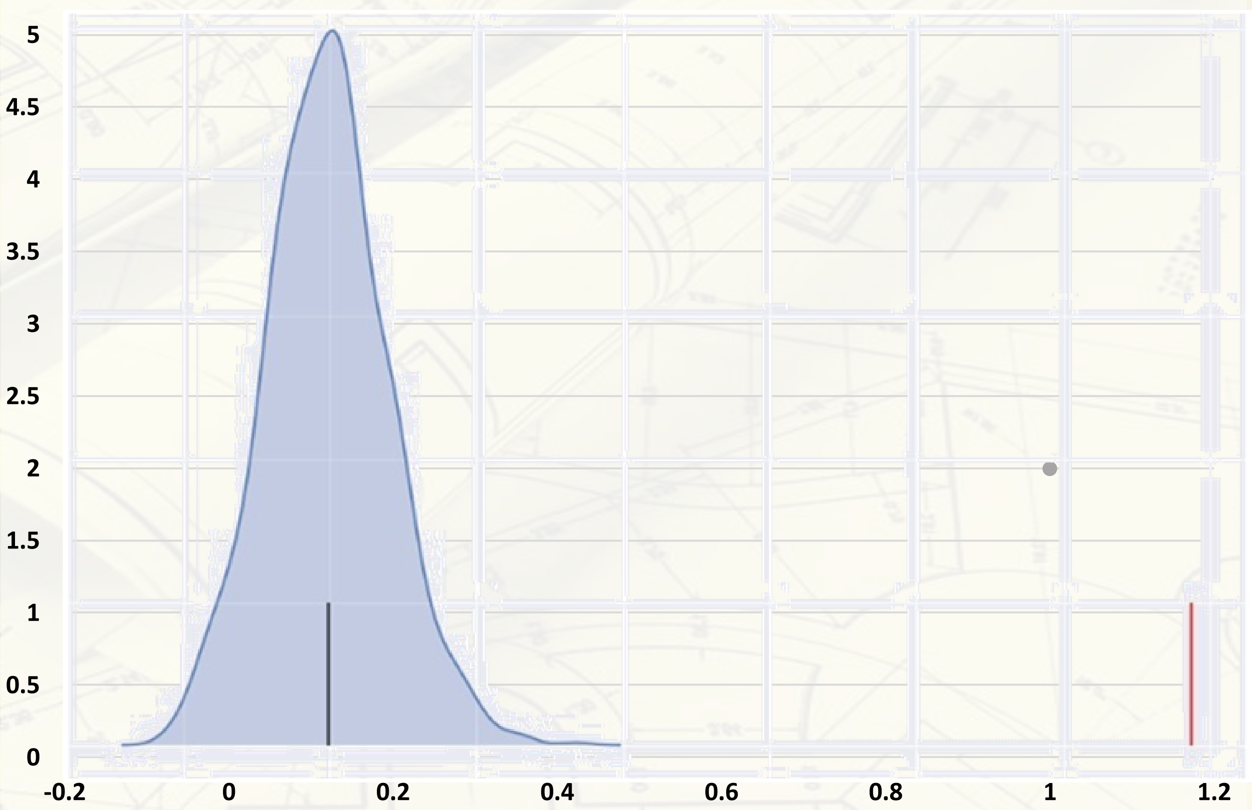
For the Global Moran's I statistic, the null hypothesis states that the attribute being analysed is randomly distributed among the features in your study area

Global Spatial Autocorrelation

The density portrays the distribution of the log price, with the black vertical line indicating the mean log price from the synthetic realizations and the red line the observed log price for our prices.

Clearly our observed value is extremely high.

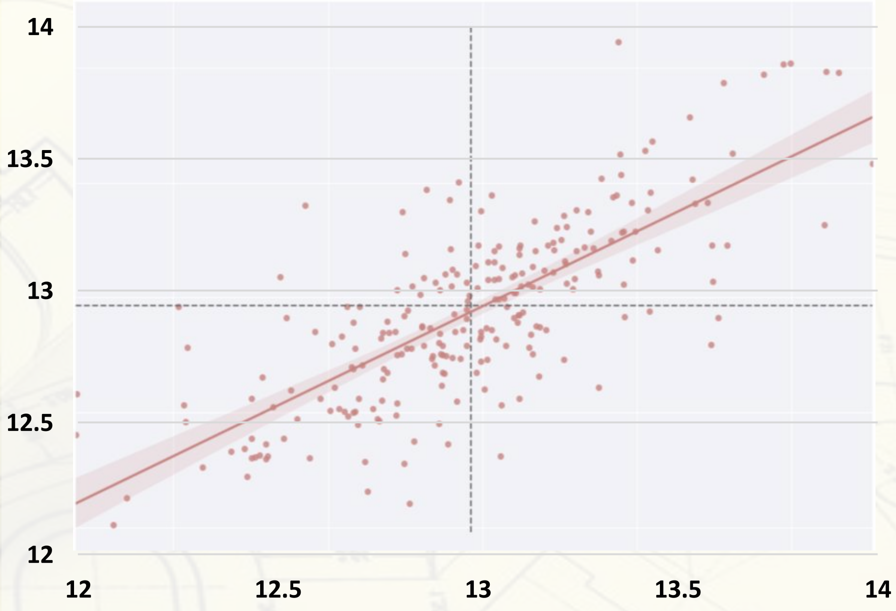
Since this is below conventional significance levels, we would reject the null of complete spatial randomness in favour of spatial autocorrelation in prices.



Log transformation of price

Moran Scatterplot

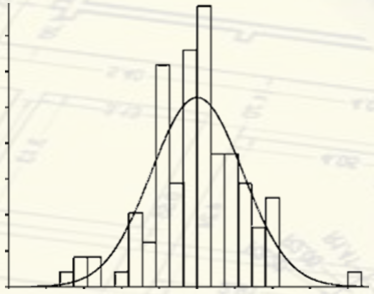
- After the removal of the outliers we that each observation falls into provides an indication of how the scatterplot works. All observations in the top right quadrant are apartments that are above the **mean log price in the data and whose local average log prices are large as well**.
- This means that observations that fall in this quadrant all tend to be more **expensive than the average listing** and are surrounded by pricier-than-average listings.
- Likewise, the **bottom left are cheap listings in cheap surroundings**.
- In the top left and lower right quadrants, the focal observation is different from its surroundings; Apartments listings in the lower right quadrant tend to **have larger-than-average (log) prices but are surrounded by cheaper-than-average apartments**.



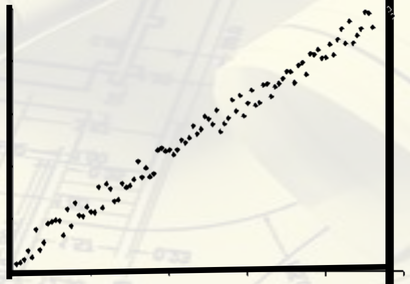
Log transformation of price

OLS Assumptions

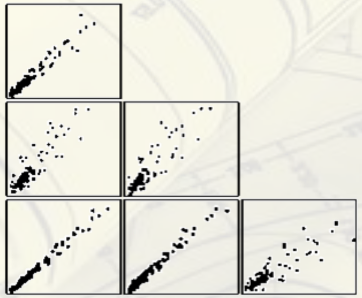
Normality of the Residuals



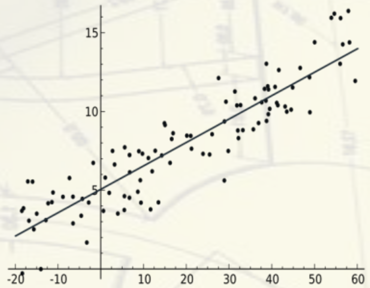
Homoscadasticity



Multicollinearity



Linearity of the Model



OLS Regression on all features

Ordinary Least Squares Regression:

Ordinary Least Squares is the simplest and most common estimator in which the independent are chosen to minimize the square of the distance between the predicted values and the actual values.

- Before implementing regression we divide our data into Test data and Train Data.
- After splitting the data we check our data for NaN values and replace them with the mean of the column using SimpleImputer
- Our Aim here is to reduce the MSE value.
- R-Squared Value: **0.481**

Dep. Variable:	log_price	R-squared:	0.481
Model:	OLS	Adj. R-squared:	0.477
Method:	Least Squares	F-statistic:	104.9
Date:	Tue, 30 Jun 2020	Prob (F-statistic):	0.00
Time:	23:20:15	Log-Likelihood:	-1869.1
No. Observations:	4792	AIC:	3824.
Df Residuals:	4749	BIC:	4103.
Df Model:	42		
Covariance Type:	nonrobust		

Omnibus:	286.365	Durbin-Watson:	2.009
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1215.541
Skew:	0.080	Prob(JB):	1.12e-264
Kurtosis:	5.462	Cond. No.	70.0

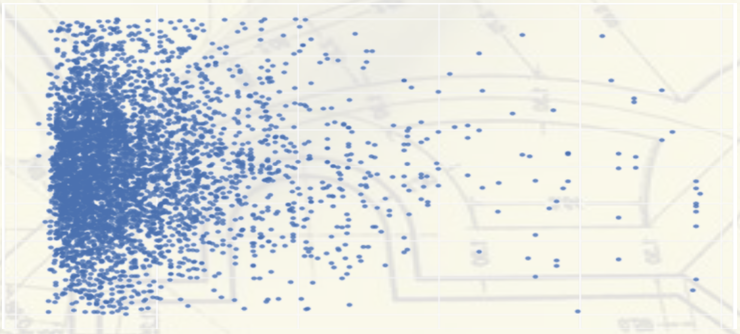
OLS Regression Results

Normality of the residuals

Studentized Residuals and Leverage

When trying to identify outliers, one problem that can arise is when there is a potential outlier that influences the regression model to such an extent that the estimated regression function is "pulled" towards the potential outlier, so that it isn't flagged as an outlier using the standardized residual criterion.

To address this issue, studentized residuals offer an alternative criterion for identifying outliers. The basic idea is to delete the observations one at a time, each time refitting the regression model on the remaining n–1 observations.

With the presence of outliers	Without the presence of outliers
	

Studentized Residual vs Leverage

Normality of the residuals

Jarque-Bera Test

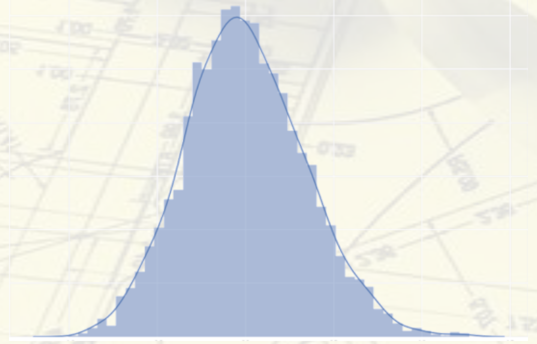
The Jarque–Bera test is a goodness-of-fit test of whether sample data have the **skewness and kurtosis matching a normal distribution**.

The **null hypothesis** is a joint hypothesis of the skewness being zero and the excess kurtosis being zero.

As we can se our data is almost normally distributed thus we **reject the null hypothesis**

Our data is normally distributed

Normal distribution curve



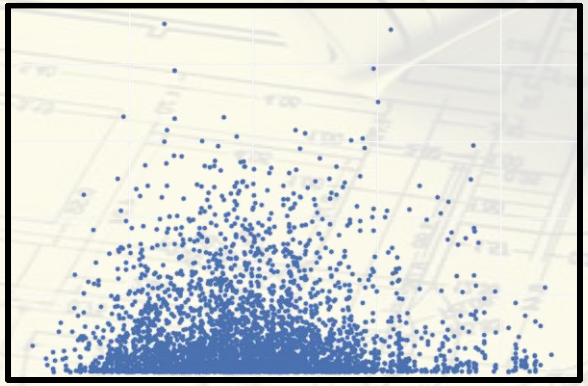
Normal probability plot

Heteroskedasticity

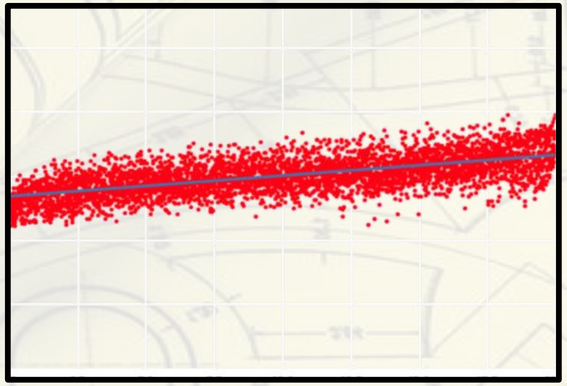
The Breusch–Pagan test, developed in 1979 is used to test for heteroskedasticity in a linear regression model.

If the test statistic has a p-value below an appropriate threshold (e.g. $p < 0.05$) then the null hypothesis of homoskedasticity is rejected and heteroskedasticity assumed.

We obtained a **p-value score of 0.00098.**
Hence we accepted the null hypothesis of homoskedasticity



No trend was seen when residuals were plotted



Homoskedasticity

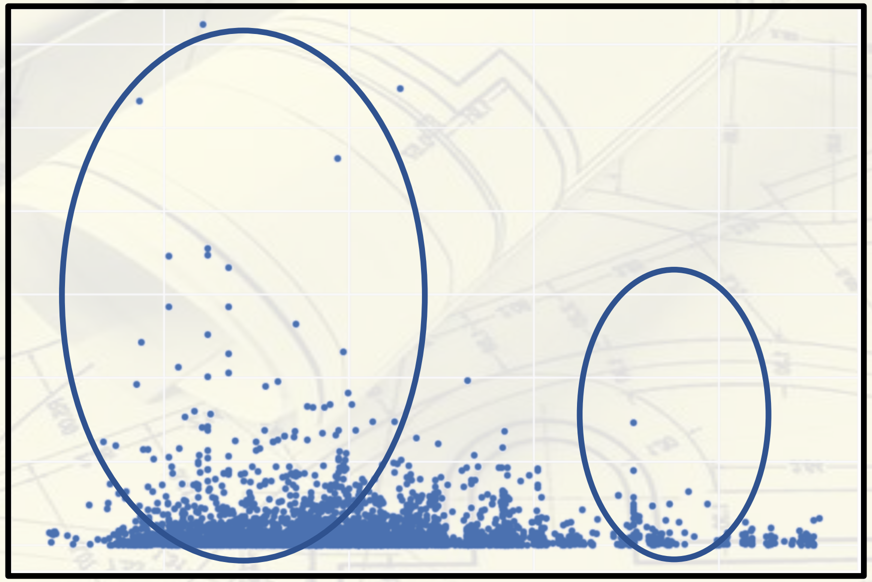
Heteroskedasticity

If features were not log transformed?

There would have been skewness in the data.

This skewness would have resulted in high variation in smaller price and smaller variation on higher price

Which would not have been good for our model



Squared residuals V/s squared values

Multicollinearity

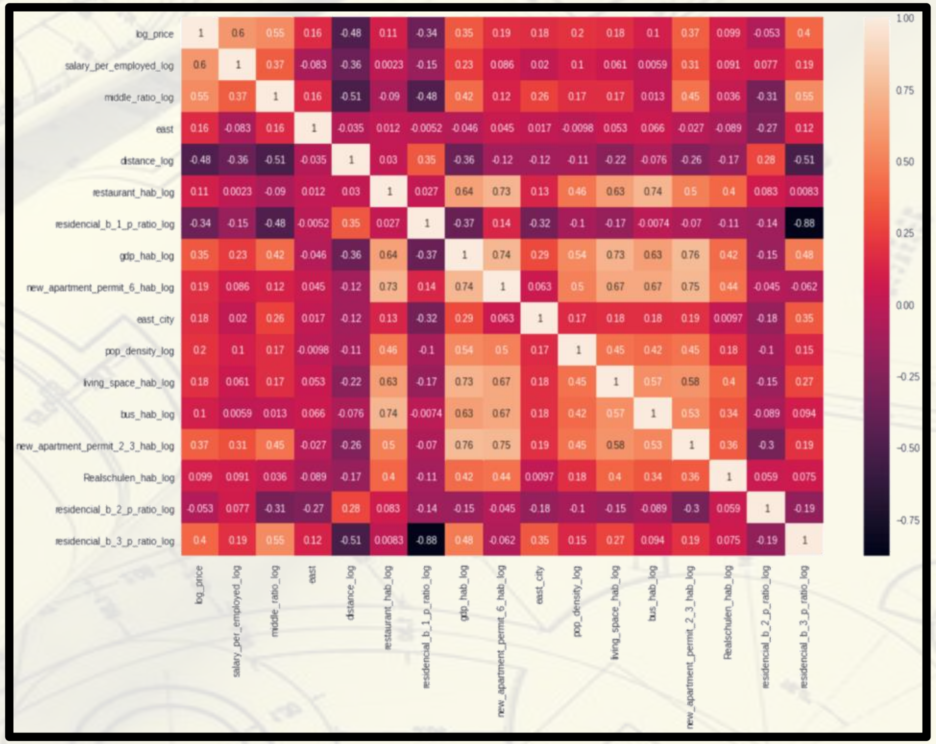
Is a measure of the relation between so-called independent variables within a regression

This phenomenon occurs when two or more predictor variables in a regression analysis are strongly associated or correlated with one another

Condition Number test

If the condition number is above 30, the regression may have severe multicollinearity.

In our case there is a nominal presence of multicollinearity as we got **10.7** as our condition number.



Corelation Matrix

Multicollinearity

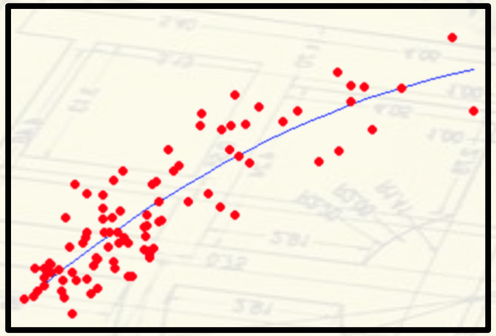
The Variance Inflation Factor (VIF) is a measure of collinearity among predictor variables within a multiple regression. It is calculated by taking the ratio of the variance of all a given model's betas divide by the variance of a single beta if it were fit alone.

	Features	VIF Factors
15	residencial_b_3_p_ratio_log	11.191559
5	residencial_b_1_p_ratio_log	10.117659
6	gdp_hab_log	9.185236
7	new_apartment_permit_6_hab_log	7.950205
12	new_apartment_permit_2_3_hab_log	4.815463
4	restaurant_hab_log	3.755383
14	residencial_b_2_p_ratio_log	3.010264
1	middle_ratio_log	2.785548
11	bus_hab_log	2.620985
10	living_space_hab_log	2.567191
3	distance_log	1.956155
0	salary_per_employed_log	1.604094
9	pop_density_log	1.528547
13	Realschulen_hab_log	1.446868
2	east	1.436891
8	east_city	1.266580

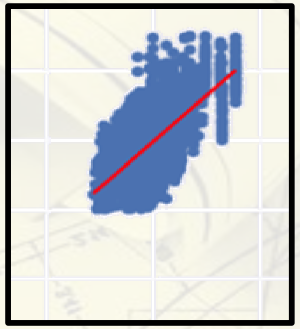
Tests on Nonlinearity

- Checking the linearity of the independent variables
- Linearity - the relationships between the predictors and the outcome variable should be linear Homogeneity of variance (homoscedasticity) - the error variance should be constant
- A good model should show linearity not non-linearity.

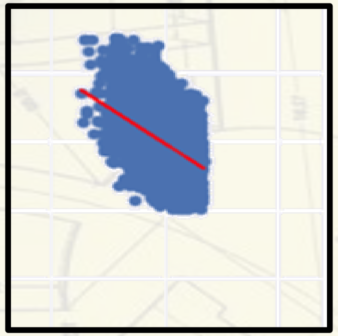
Our model shows Linearity



Example of Non-Lienarity

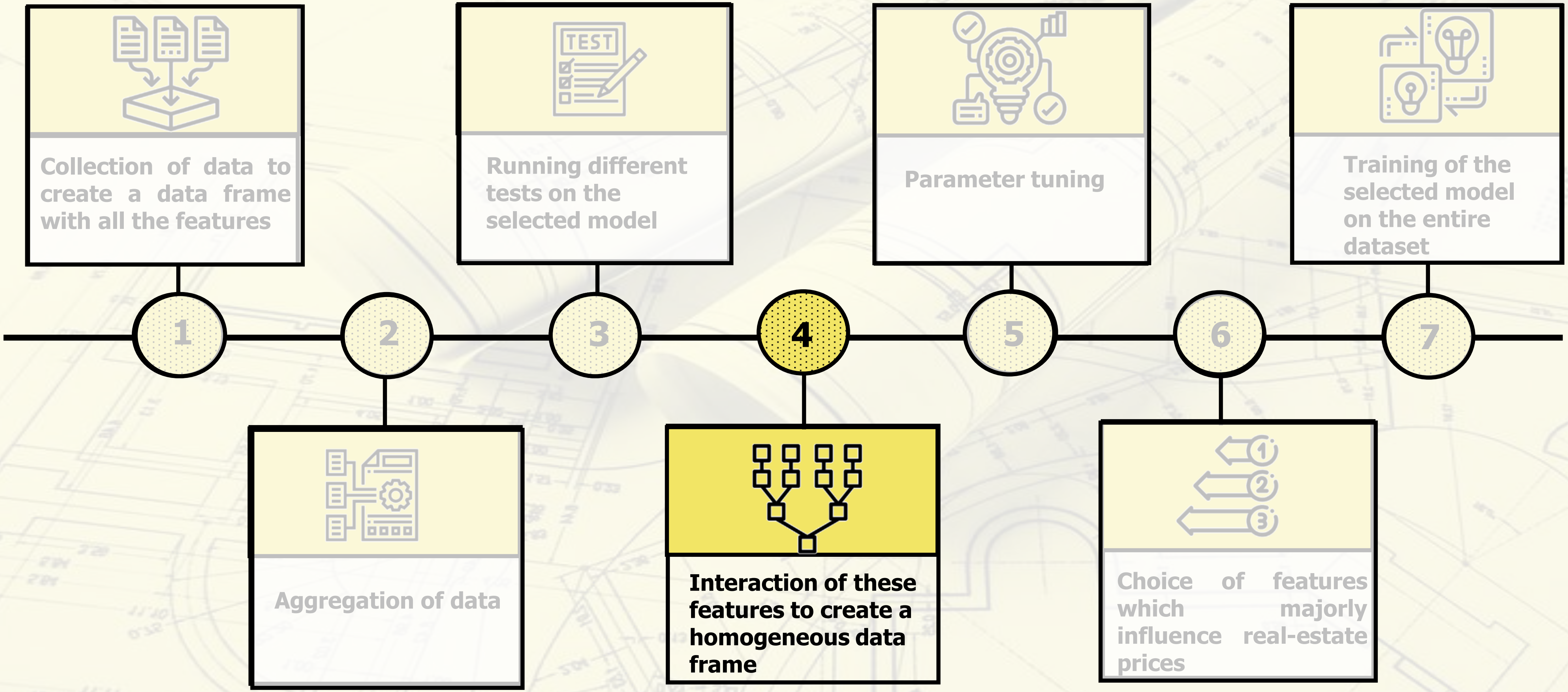


Log price vs Salary Per Employed



Log price vs distance_log

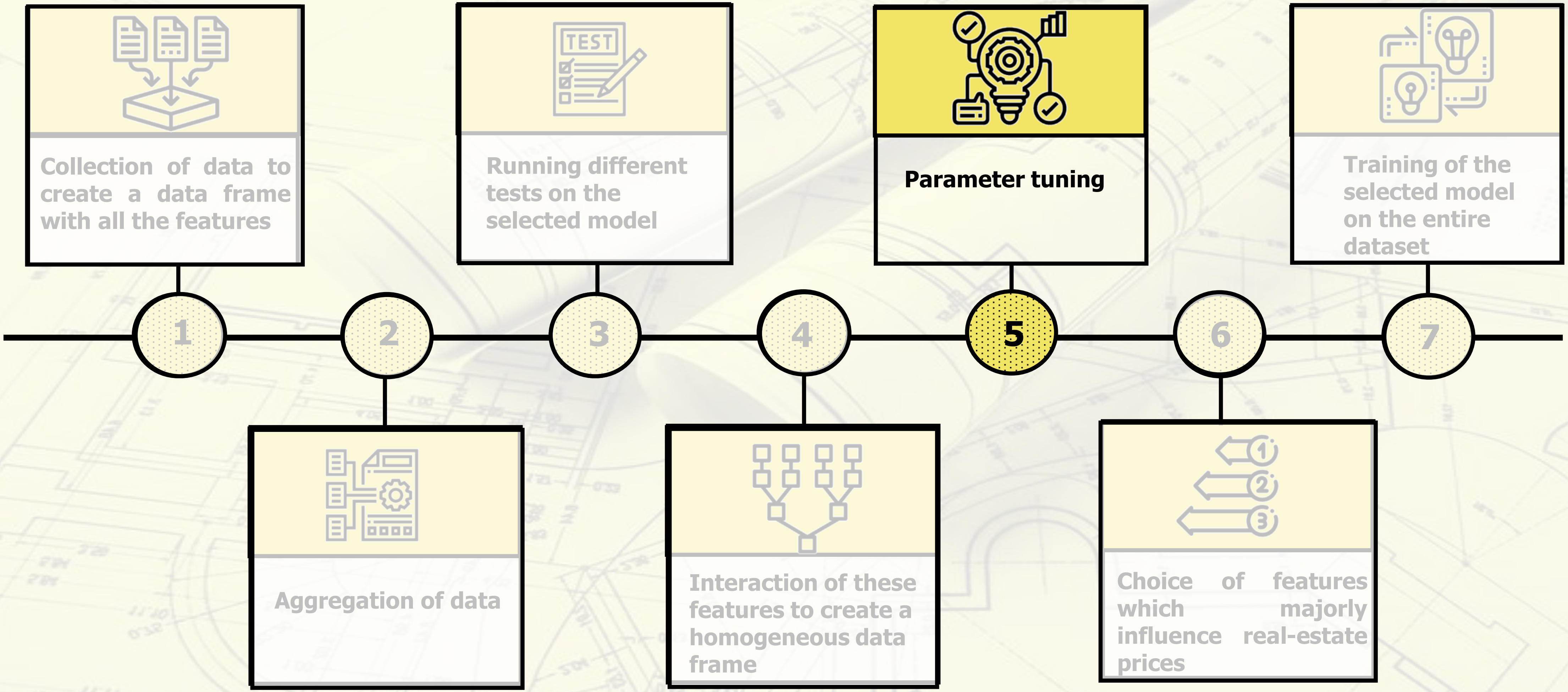
Methodology



Feature Interaction

Interacted Features	Description
Young_age_ratio	Age from 3 years to 25 years/ sum of every age
Middle_ratio	Age from 25 years to 50 years/ sum of every age
Old_age_ratio	Age from 50 years to 75 years/sum of every age
Floor_area_per_veg	Floor area per vegetation/ (floor area settlement+floor area traffic+floor area vegetation)
Salary_per_employed	Income total/employed
Residential_b_1_p_ratio	Residential_b_1_p_hab/(residential_b_1_p_hab+residential_b_2_p_hab+residential_b_3_p_hab)
Residential_b_2_p_ratio	Residential_b_2_p_hab/(residential_b_1_p_hab+residential_b_2_p_hab+residential_b_3_p_hab)
Residential_b_3_p_ratio	Residential_b_3_p_hab/(residential_b_1_p_hab+residential_b_2_p_hab+residential_b_3_p_hab)
east	Binary(1 if east of the country/state)
north_city	Binary(1 if north of the city)
east_city	Binary(1 if east of the city)
Pop_density	Einwohner/ (squaredkilometer/zipcode)

Methodology



Methodology

1. Collection of data to create a data frame with all the features
2. Aggregation of data
3. Running different tests on the selected model
4. Interaction of these features to create a homogeneous data frame
5. Parameter tuning
6. Choice of features which majorly influence real-estate prices
7. Training of the selected model on the entire dataset

Feature Selection

We are implementing **k-Cross Fold Validation** to evaluate our models, in our case we are keeping **k=5**

Our aim is to choose the best features possible which help in improving our model by reducing MSE value.

Strategy being used: Forward Stepwise Selection
- We start by creating an empty list and append only the relevant features.

We add a feature and check the MSE if the the MSE value improves then we add the Feature to list otherwise remove it and move to the next feature
In our case we are repeated this procedure untill we got 16 best features

	coef	std err	t	P>|t|	[0.025	0.975]
const	12.9806	0.005	2578.685	0.000	12.971	12.990
restaurant_hab_log	0.0834	0.011	7.824	0.000	0.062	0.104
living_space_hab_log	0.0090	0.008	1.075	0.283	-0.007	0.025
dormitories_hab_log	-0.0665	0.013	-5.211	0.000	-0.091	-0.041
cafe_hab_log	0.0082	0.008	0.995	0.320	-0.008	0.024
residencial_b_3_p_ratio_log	0.2094	0.017	12.167	0.000	0.176	0.243
qkm_hab_log	-0.1691	0.013	-13.204	0.000	-0.194	-0.144
floor_use_leisure_qkm_plz	-0.0106	0.006	-1.823	0.068	-0.022	0.001
new_apartment_permit_4_5_hab_log	0.1383	0.012	11.268	0.000	0.114	0.162
east	0.0967	0.006	15.179	0.000	0.084	0.109
old_age_ratio_log	-0.1400	0.008	-17.348	0.000	-0.156	-0.124
hospital_hab_log	-0.0101	0.006	-1.785	0.074	-0.021	0.001
Realschulen_hab_log	0.0169	0.006	2.797	0.005	0.005	0.029
new_apartment_permit_1_hab_log	0.0253	0.011	2.410	0.016	0.005	0.046
Gymnasien_hab_log	0.0058	0.015	0.392	0.695	-0.023	0.035
residencial_b_1_p_ratio_log	0.1684	0.015	11.211	0.000	0.139	0.198
residencial_b_2_p_ratio_log	0.1247	0.008	14.778	0.000	0.108	0.141

Dep. Variable	log_price	R-squared	0.420
Model	OLS	Adj. R-squared	0.418

Regression Summary Table

Coefficients of the parameters

Methodology

1. Collection of data to create a data frame with all the features
2. Aggregation of data
3. Running different tests on the selected model
4. Interaction of these features to create a homogeneous data frame
5. Parameter tuning
6. Choice of features which majorly influence real-estate prices
7. **Training of the selected model on the entire dataset**

Coefficients sorted by importance

Features	Description	coefficients	standard error	t-value	[0.025]
new_apartment_permit_6_hab_log	Permit of building apartment with more than 4 apartments	0.1926	0.021	9.133	0.151
salary_per_employed_log	Salary per employed people	0.1843	0.006	33.381	0.174
new_apartment_permit_2_3_hab_log	Permit of building apartment with upto 3 apartments	0.0933	0.015	6.421	0.065
middle_ratio_log	Age ranging from 25 to 50 years old	0.0919	0.007	13.404	0.078
residencial_b_3_p_ratio_log	Three people living in an apartment	0.0823	0.014	5.973	0.055
restaurant_hab_log	Restaurants per zip code	0.0717	0.008	9.292	0.057
cafe_hab_log	Cafe per zip code	0.0423	0.007	5.839	0.028
east	East side of country (1 if east of country)	0.0368	0.005	6.778	0.026
residencial_b_2_p_ratio_log	Two people sharing	0.036	0.008	4.763	0.021
east_city	East side of city	-0.0223	0.005	4.632	0.013
residencial_b_1_p_ratio_log	Einwohner apartment	-0.0188	0.013	-1.427	-0.045
Realschulen_hab_log	Schools per zipcode	-0.0224	0.007	-3.008	-0.037
distance_log	Euclidian distance	-0.0837	0.006	-13.598	-0.096

Log- Log Model

$$\boldsymbol{ln(y) = \beta_0 + \beta_1 ln(x)}$$

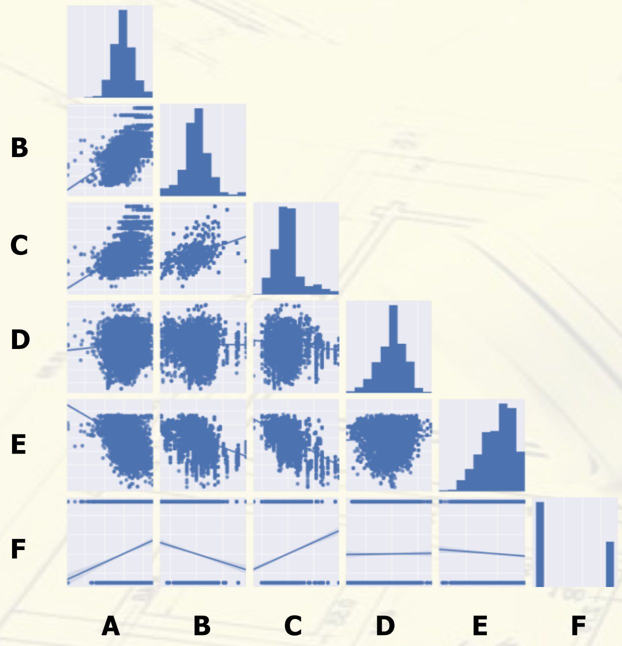
$$\boldsymbol{\frac{\partial y}{Y} = \beta_1 \frac{\partial x}{x}}$$

When both dependent/response variable and independent/predictor variable(s) are log-transformed, we interpret the coefficient as the percent increase in the dependent variable for every 1% increase in the independent variable.

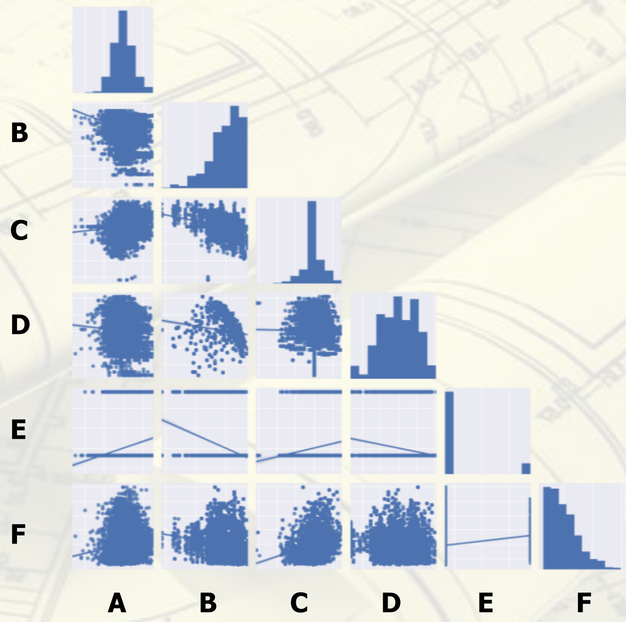
For example- 0.1926 * new_apartment_permit_6_hab_log

Means that for every 1% increase in the apartment buildings per habitant, the price of the apartment will increase 0.19 % .

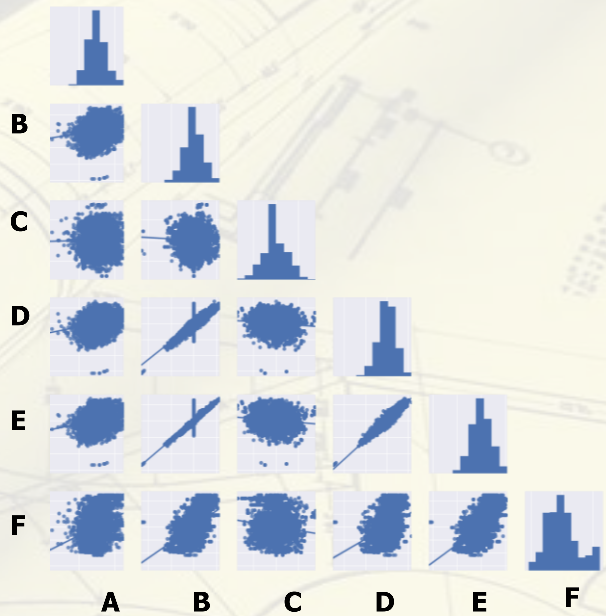
Distribution of selected features



Legend	
log_price	A
salary_per_employed_log	B
middle_ratio_log	C
restarurant_hab_log	D
distance_log	E
east	F

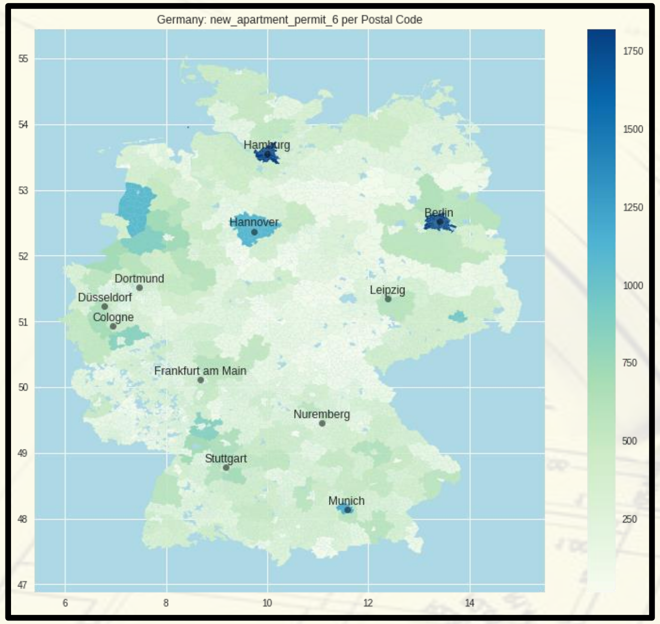


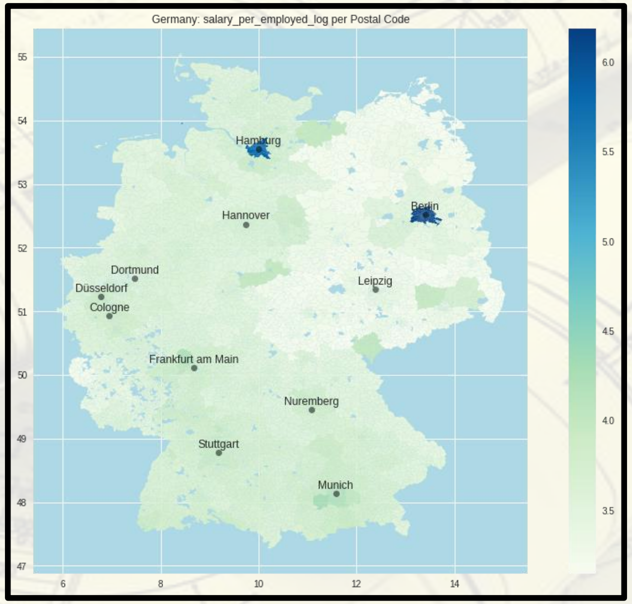
Legend	
log_price	A
residencial_b_1_p_ratio_log	B
realschulen_hab_log	C
residencial_b_2_p_ratio_log	D
east_city	E
einwohner	F

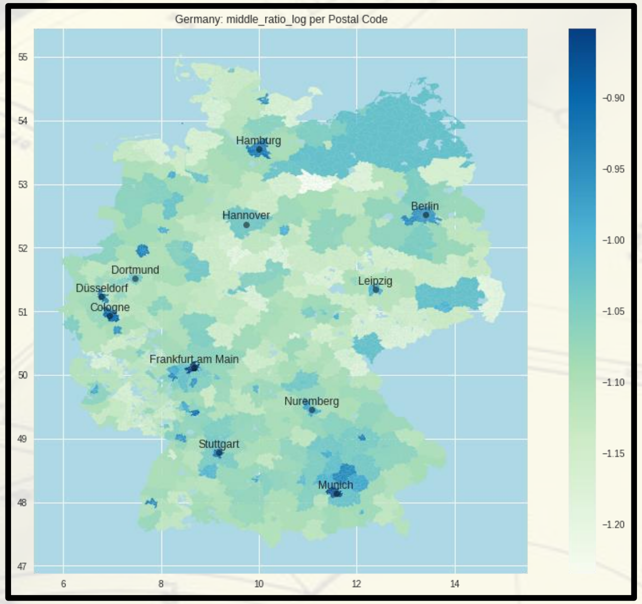


Legend	
log_price	A
employed_hab_log	B
café_hab_log	C
new_apartment_permit_hab_log	D
grad_school_hab_log	E
residencial_b_3_p_ratio_log	F

Geographical Distribution of selected district features





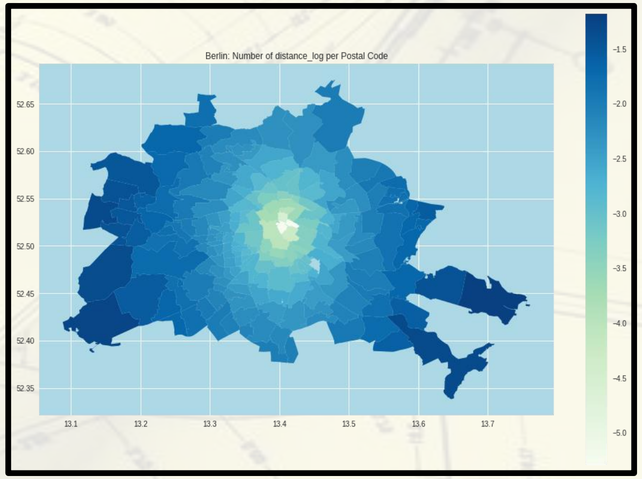


Germany: New apartment permit type 6

Germany: Salary per employed person

Germany: Middle age ratio

Geographical Distribution of selected features



Berlin: Restaurant per habitant

Berlin: cafe per habitant

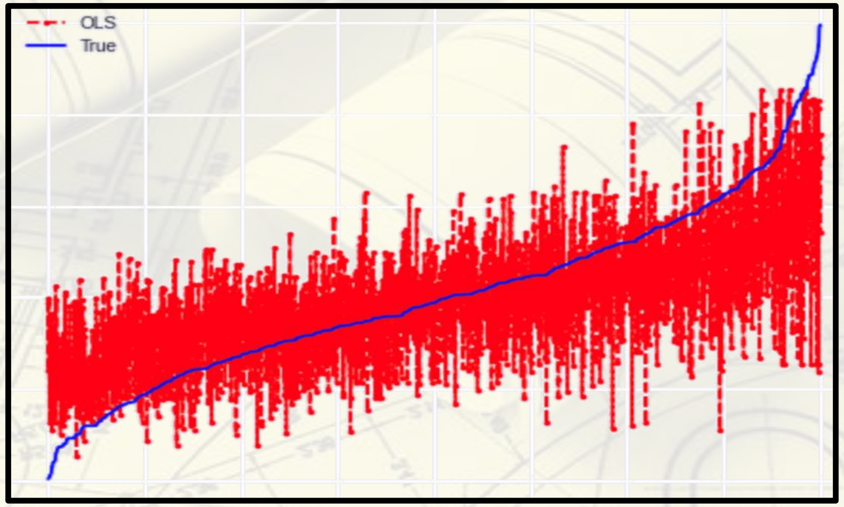
Berlin: Distance to closest main city

How does the model behave on unseen data?

The blue line depicts the actual price of the apartments

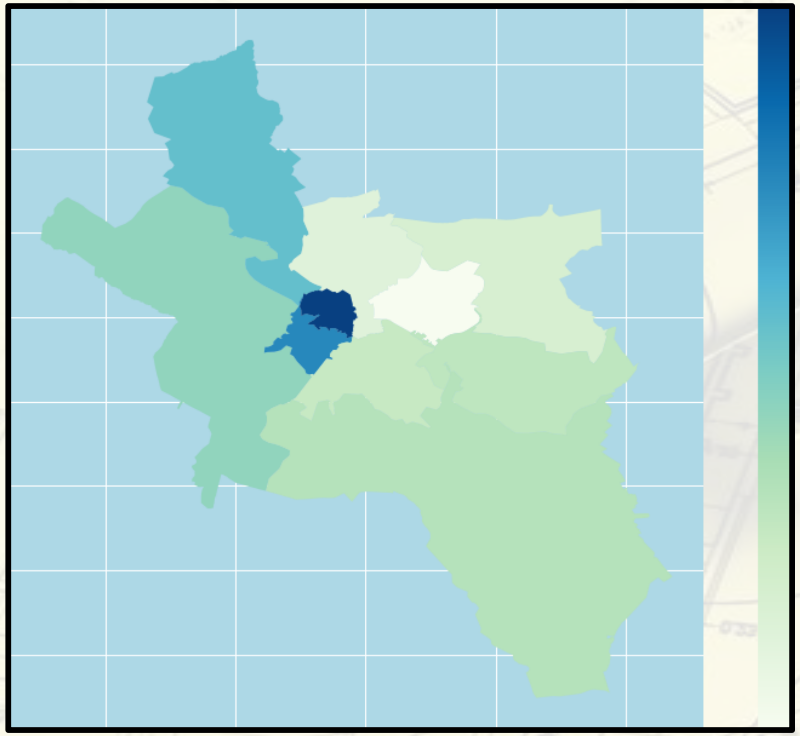
The red region is how our model predicted the price of the apartments with a MSE value of 0.119

We were almost able to achieve the perfect model



True value vs Predicted Value

Results



Aachen: Predicted Pricing of Apartments

Germany: Predicted Pricing of Apartments

Conclusion

- **Apartments in big cities are costlier than in towns or villages based on:**
    1. Number of the amenities closer to apartment
    2. Region with more working class
    3. As older people tend to live in the outskirts

- **Areas with mayor presence of Building with single apartment are cheaper than building with more apartments**

- **Apartments on the east side of Germany are costlier than west side of Germany**

- **Apartments on the east side of the city are costlier**

Future Work

IMMO SCOUT 24 <

Use more data sources than immobilienscout24

Redefine the transport, university, and bus variables as we think they should have had a higher correlation to the price but they did not

Try different aggregation levels like district, neighborhood pair of postal codes as there many postal code with no information

References

- Kajuth, Florian and Knetsch, Thomas and Pinkwart, Nicolas, Assessing House Prices in Germany: Evidence from an Estimated Stock-Flow Model Using Regional Data (2013). Bundesbank Discussion Paper No. 46/2013.
- Daisuke Murakami & Yoshiki Yamagata, 2019. "Estimation of Gridded Population and GDP Scenarios with Spatially Explicit Statistical Downscaling," Sustainability, MDPI, Open Access Journal, vol. 11(7), pages 1-18, April.
- "Germany property and metropolis market outlook 2019" Authors:Jochen Möbert Stable URL: https://www.dbresearch.com/PROD/RPS_EN-PROD/PROD0000000000488315/German_property_and_metropolis_market_outlook_2019.pdf
- Citation: Tomal M., 2019, The Impact of Macro Factors on Apartment Prices in Polish Counties: a Two-Stage
- Quantile Spatial Regression Approach, Real Estate Management and Valuation, vol. 27, no. 4, pp. 01-14.
- Anselin, Luc. *Spatial econometrics: methods and models*. Vol. 4. Springer Science & Business Media, 2013.

Thank you