Rent Analysis and Prediction

FOR APARTMENTS IN GERMANY

FEMI ONAFALUJO

SPRINGBOARD DATA SCIENCE

SEPTEMBER 26, 2021

"The more you know about the past, the better prepared you are for the future."

- THEODORE ROOSEVELT (1858-1919)

Project Objectives

Problem

- Help landlords set the rent for their apartments

Solution

- Develop a predictive model

It is a Journey:
Solutions Areas & Scopes

Data Wrangling

- Collect and organize data
- Clean data

Exploratory Data Analysis

- Relationships between rent and numerical features
- Relationship between rent and categories

Machine Learning

- Training data development
- Metrics and testing
- Model development
- Model selection
- Model application

Data Wrangling

COLLECTING DATA | CLEANING DATA

Primary Dataset – Rental Information

Source: Immoscout24 – largest real estate website in Germany

Information: Record of apartments for rent

Size: 267,859 records, 49 features

Examples of features

- Heating cost
- Rent (€)
- Living space (square meters)
- Number of rooms
- Interior quality
- Location information (State, city / town, municipality, zip code, street address, house number)
- Facilities (Balcony, garden, cellar, etc.)

Secondary Dataset – State Information

Source: Wikipedia

Information: State macro-economic data

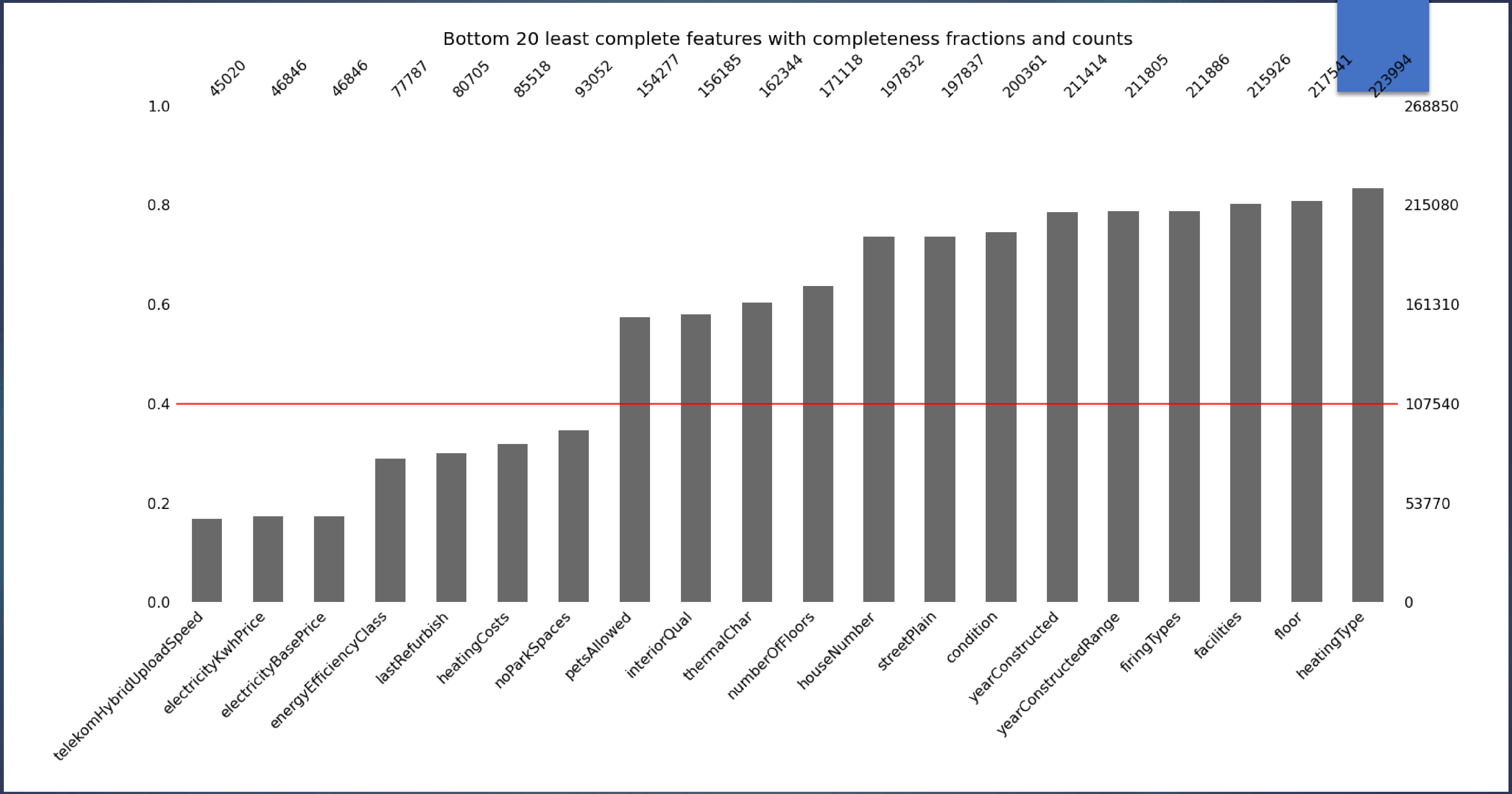
Size: 16 states x 5 features

Features

- State area
- State population
- State population per area
- GDP per capita
- Human development index (HDI)

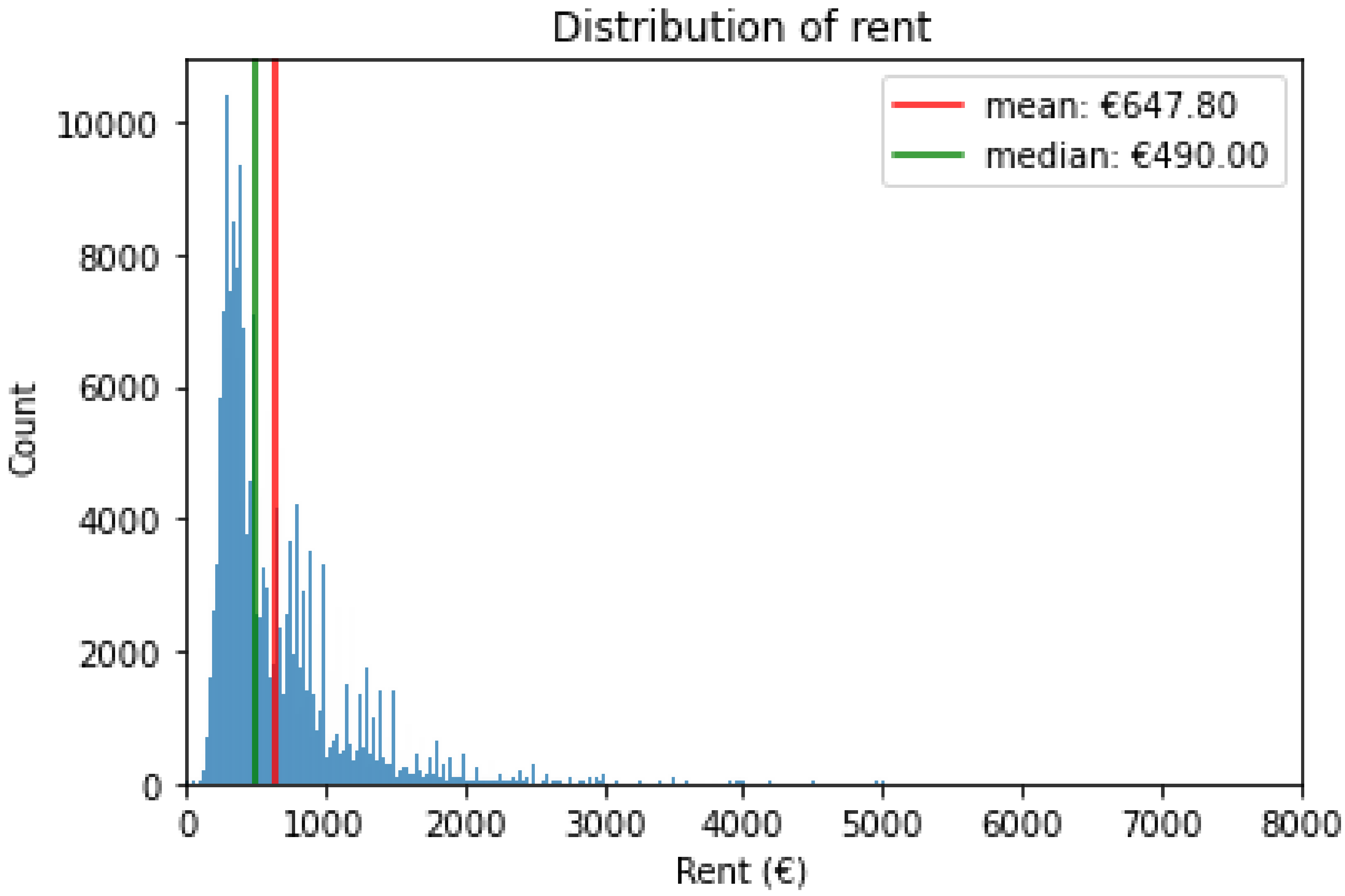
Issues: Cleaning the data

- Missing values
- High cardinality
- Duplicate entries
- Inconsistent category names


Bottom 20 least complete features with completeness fractions and counts
45020
46846
46846
77787
80705
85518
93052
154277
156185
162344
171118
197832
197837
200361
211414
211805
211886
215926
217541
223994
1.0
0.8
0.6
0.4
0.2
0.0
268850
215080
161310
107540
53770
0
telekomHybridUploadSpeed
electricityKwhPrice
electricityBasePrice
energyEfficiencyClass
lastRefurbish
heatingCosts
noParkSpaces
petsAllowed
interiorQual
thermalChar
numberOfFloors
houseNumber
streetPlain
condition
yearConstructed
yearConstructedRange
firingTypes
facilities
floor
heatingType

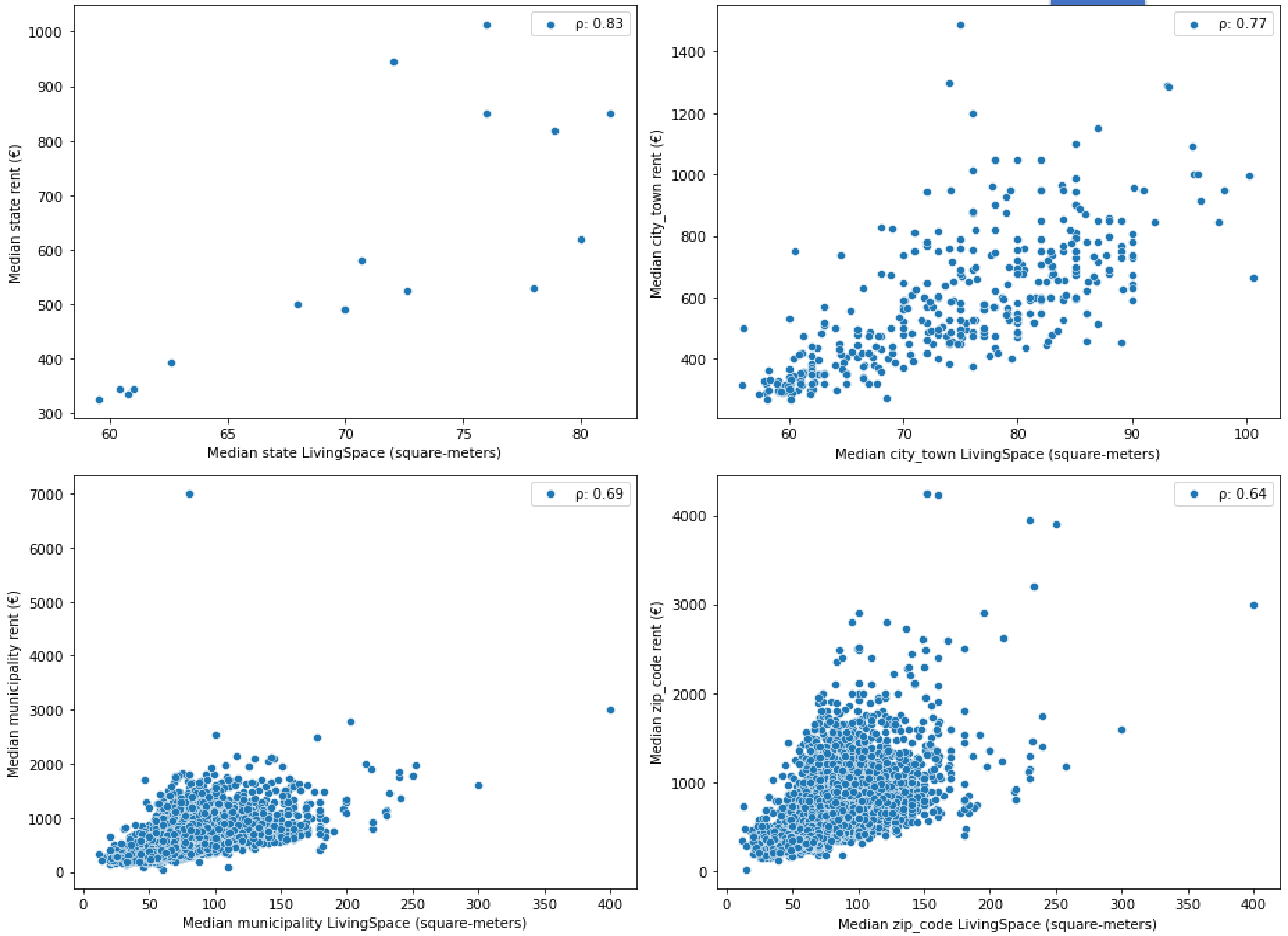
Exploratory Data Analysis

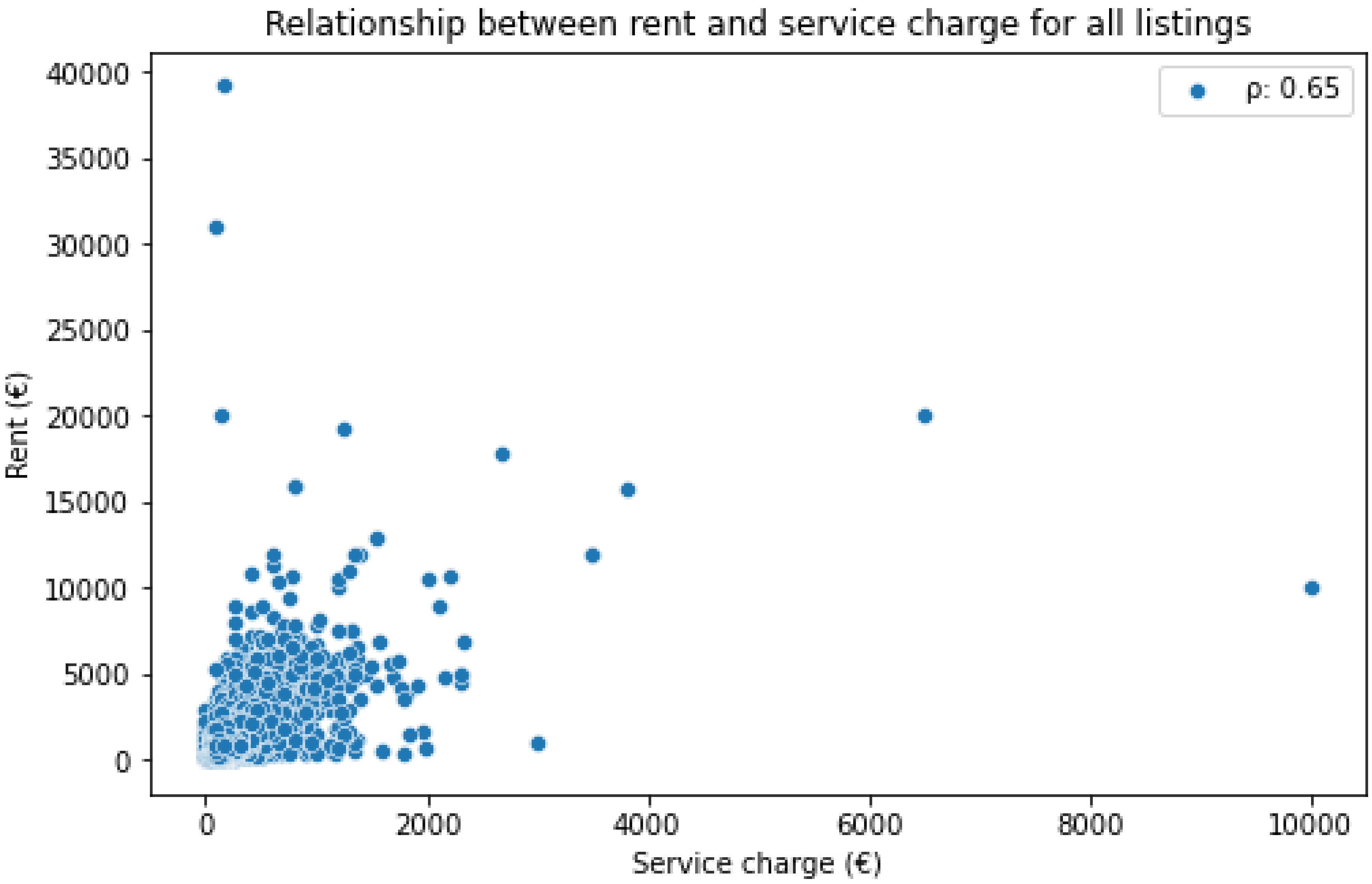
RENT VS. NUMERICAL AND CATEGORICAL FEATURES


Distribution of rent
mean: €647.80
median: €490.00
Count
10000
8000
6000
4000
2000
0
0
1000
2000
3000
4000
5000
6000
7000
8000
Rent (€)

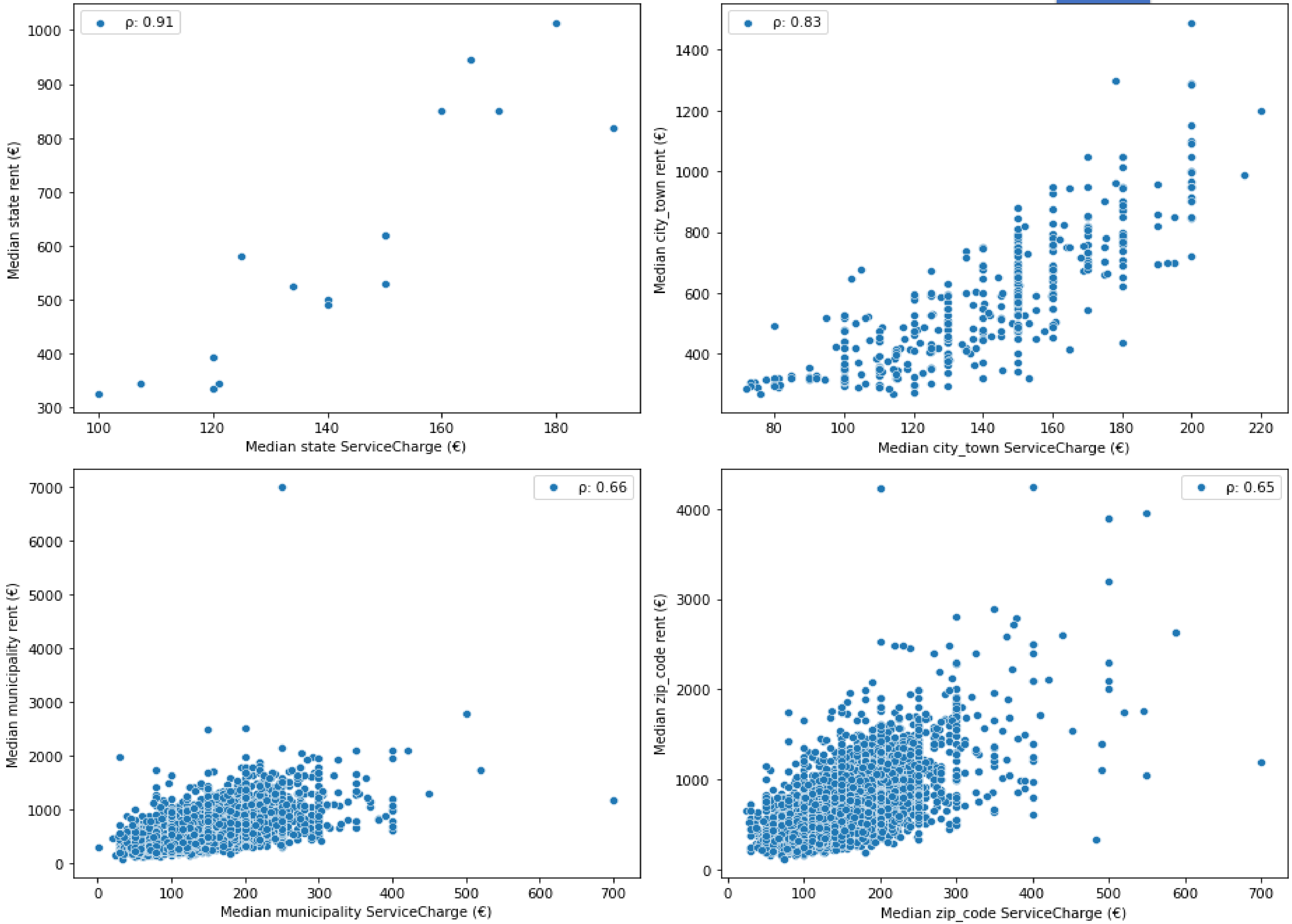

Relationship between rent and living space for all listings
ρ: 0.72
Rent (€)
0
5000
10000
15000
20000
25000
30000
35000
40000
0
200
400
600
800
1000
Living space (square-meters)

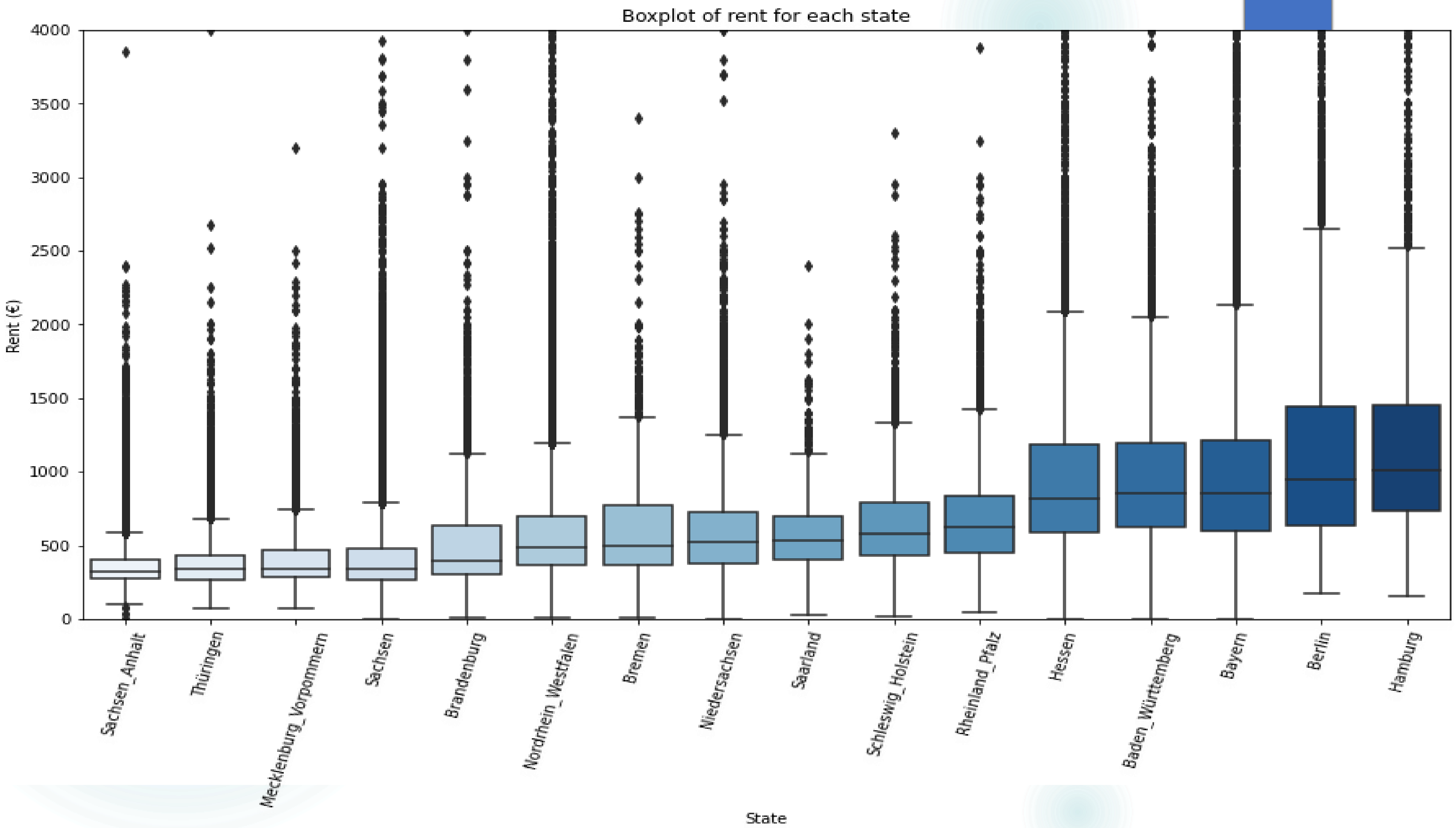
Plot of median rent versus median living space at the state, city / town, municipality, and zip code level



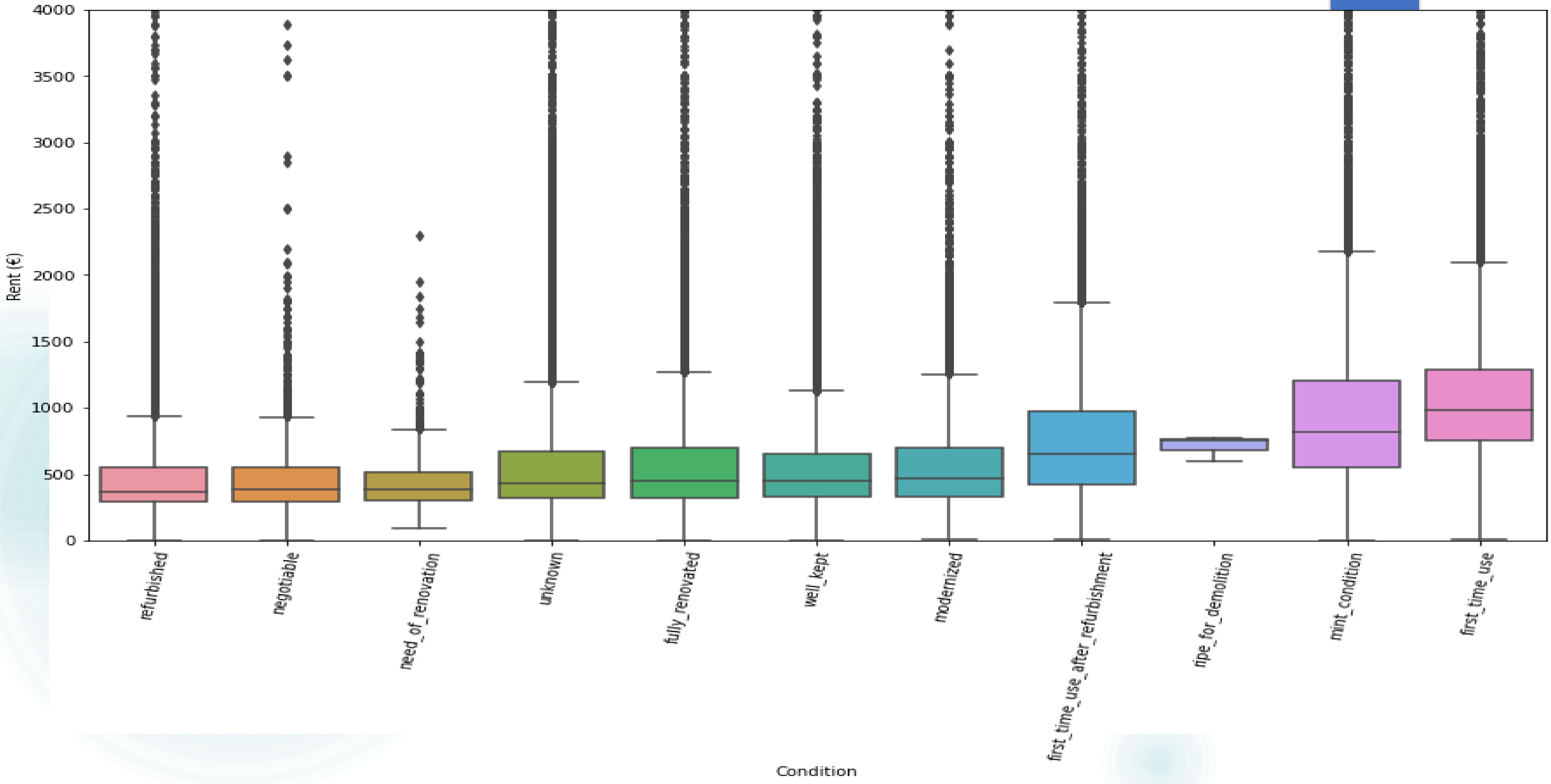

Relationship between rent and service charge for all listings
ρ: 0.65
40000
35000
30000
25000
20000
15000
10000
5000
0
Rent (€)
0
2000
4000
6000
8000
10000
Service charge (€)

Plot of median rent versus median service charge at the state, city / town, municipality, and zip code level

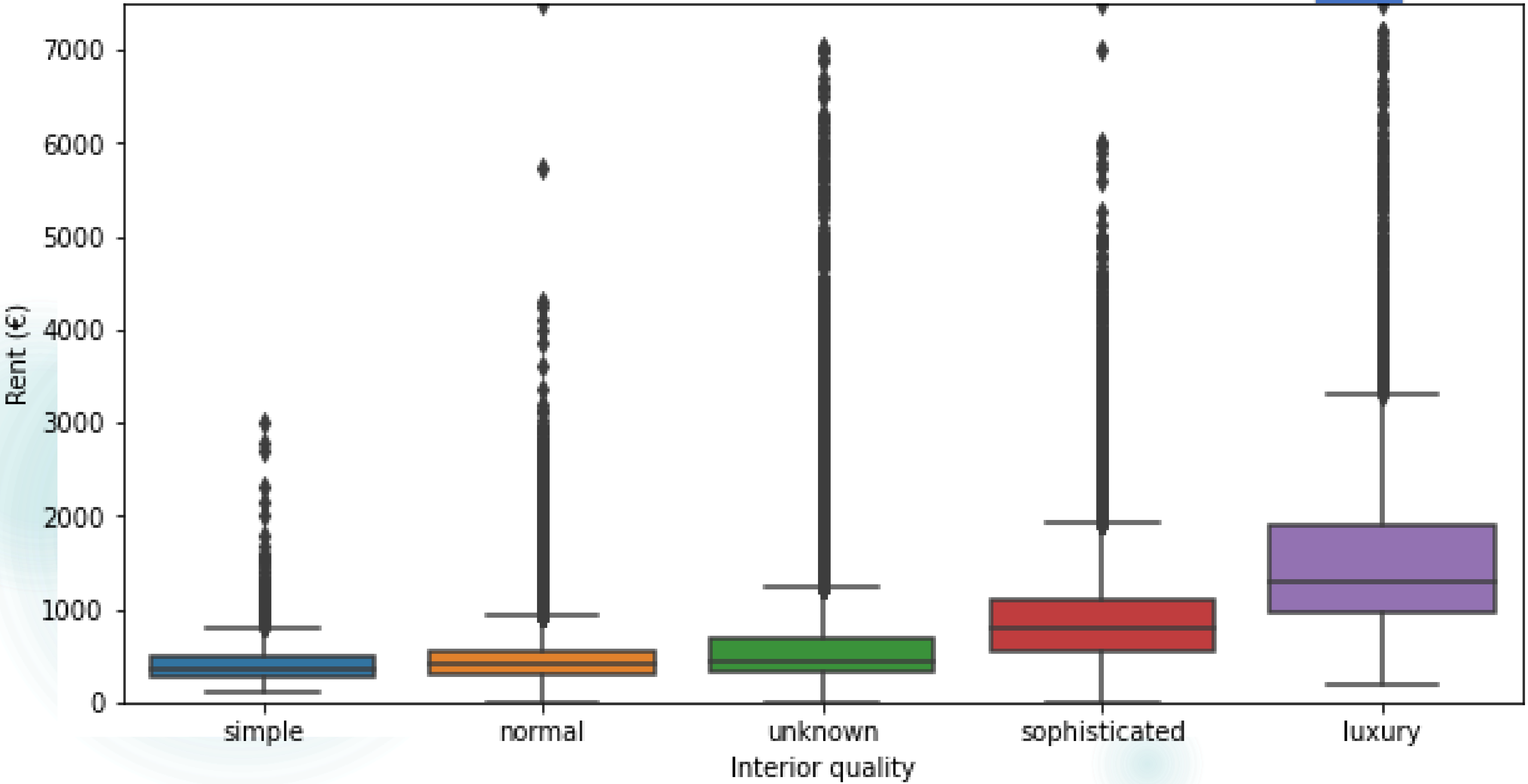



Boxplot of rent for each state
Rent (€)
4000
3500
3000
2500
2000
1500
1000
500
0
Sachsen_Anhalt
Thüringen
Mecklenburg_Vorpommern
Sachsen
Brandenburg
Nordrhein_Westfalen
Bremen
Niedersachsen
Saarland
Schleswig_Holstein
Rheinland_Pfalz
Hessen
Baden_Württemberg
Bayern
Berlin
Hamburg
State

Relationship between rent and condition



Relationship between rent and quality



Machine Learning

TRAINING DATA DEVELOPMENT | METRICS | TESTING | MODEL DEVELOPMENT | SELECTION | APPLICATION

Training data development

Application-set:
- 5 random samples

Training-set:
- 70% of remaining observations

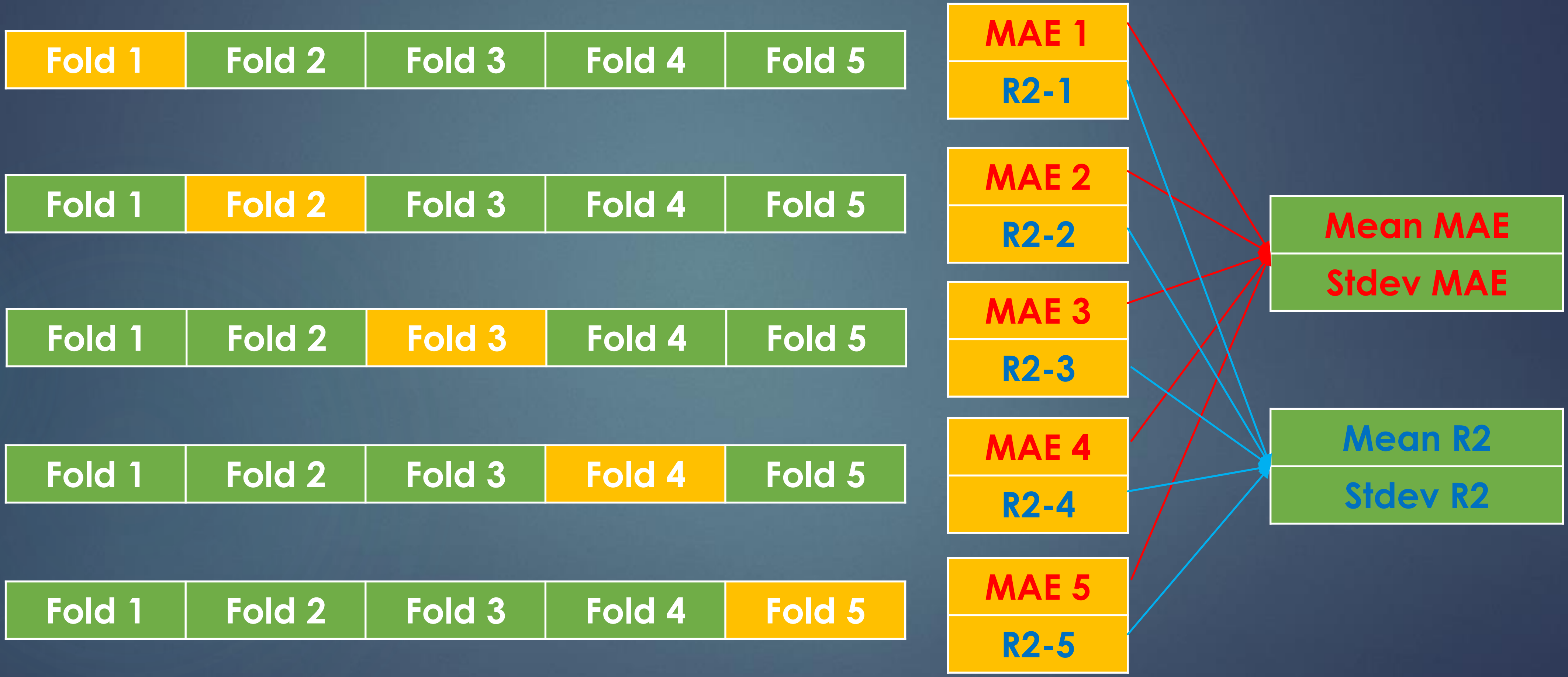
Test-set:
- 30% or remaining observations

Metrics

R-squared score (R2)	$1 - \frac{Residual\ \text{sum of square errors}}{Total\ \text{sum of square errors}}$
Mean Absolute Error (MAE)	$\sum \frac{\lvert \mathrm{y} - \ \hat{y})\rvert}{N}$

Testing strategy on training-set (70% of data)

- Cross-validation on 70% training set (5 fold)



Testing strategy (on test-set (30% of data))

Test set performance

MAE performance

R2-performance

Baseline model – linear regression

	R2 score	Mae (€)
Train set	0.73	146.36
Test set	0.75	147.29

Model development - Algorithms

Linear

- Lasso regression
- Ridge regression

Tree induction

- Random forest
- XGBoost

Model optimization for tree-induction

Feature selection

Used feature importance for tree-induction model

Gini importance threshold value of 0.001

Hyperparameter tuning

Random forest

- Number of estimators
- Maximum depth of tree

XGBoost

- Subsample
- Maximum depth of tree
- Column sample

Model Performance: r-squared

Model	Mean cv r2 score	Stdev of cv r2 scores	R2 test score
XGBoost	0.837	0.045	0.866
XGBoost w/ feat sel.	0.838	0.045	0.863
Random forest w/ hyper	0.819	0.047	0.861
XGBoost w/ hyper	0.827	0.047	0.849
Random forest w/ feat sel.	0.809	0.045	0.848
Random forest	0.807	0.049	0.841
Ridge	0.730	0.048	0.748
Lasso	0.730	0.048	0.747

Model performance: mean absolute error

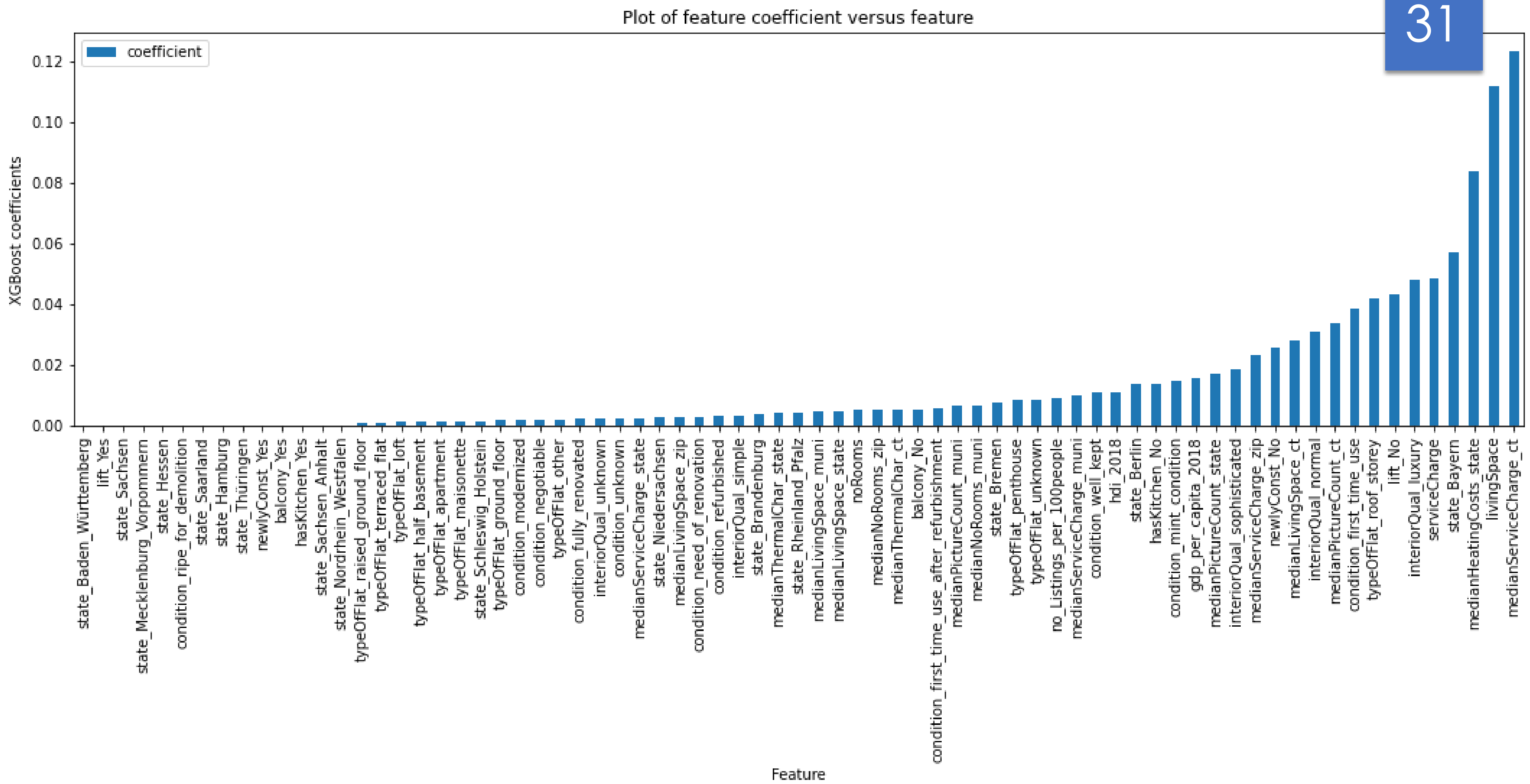
Model	Mean cv mae score (€)	Stdev of cv mae scores (€)	mae test score (€)
Random forest w/ hyper	91.34	0.67	90.34
XGBoost	91.90	0.31	92.04
XGBoost w/ feat sel.	91.79	0.52	92.14
Random forest w/ feat sel.	97.44	0.66	97.19
Random forest	97.30	0.80	97.20
XGBoost w/ hyper	103.52	0.35	103.49
Lasso	145.63	0.81	146.49
Ridge	146.35	0.82	147.20

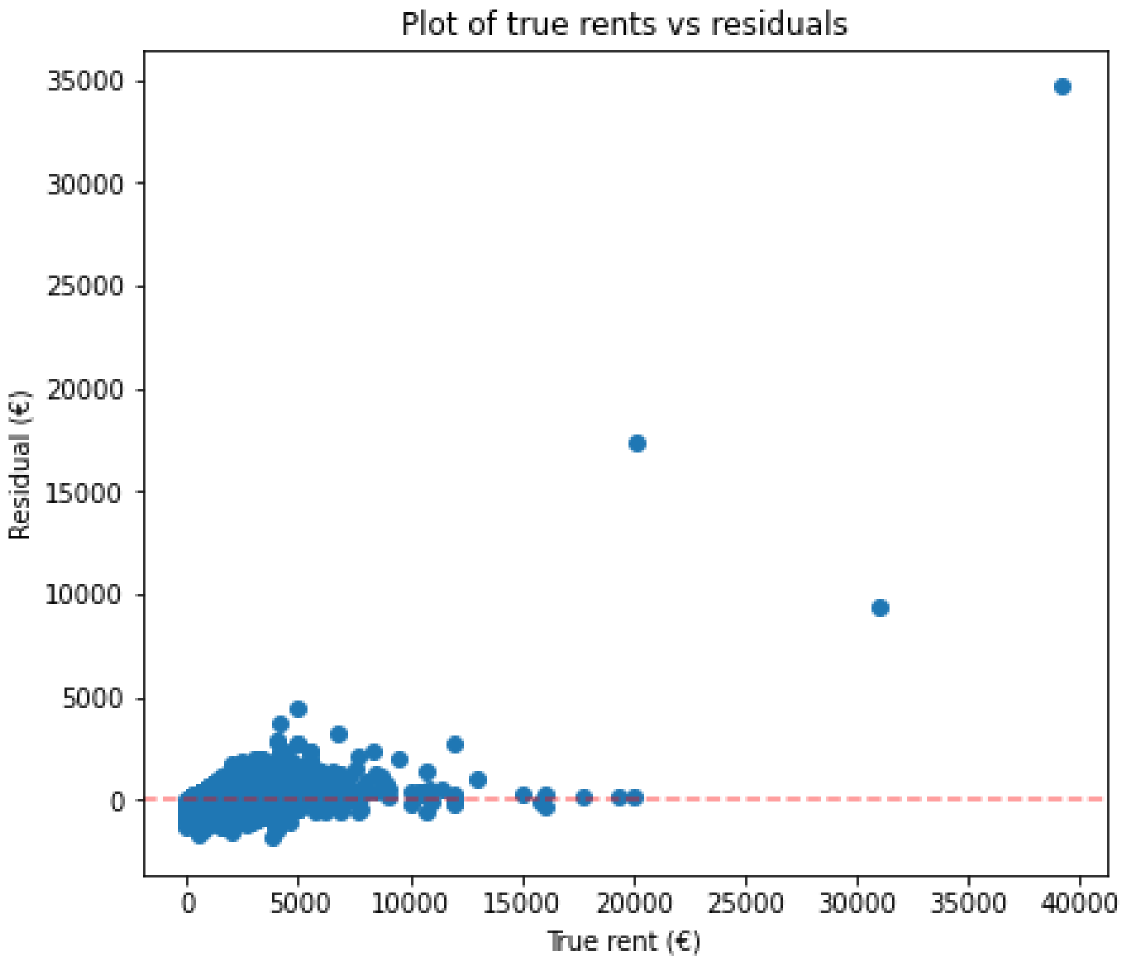
Common important features

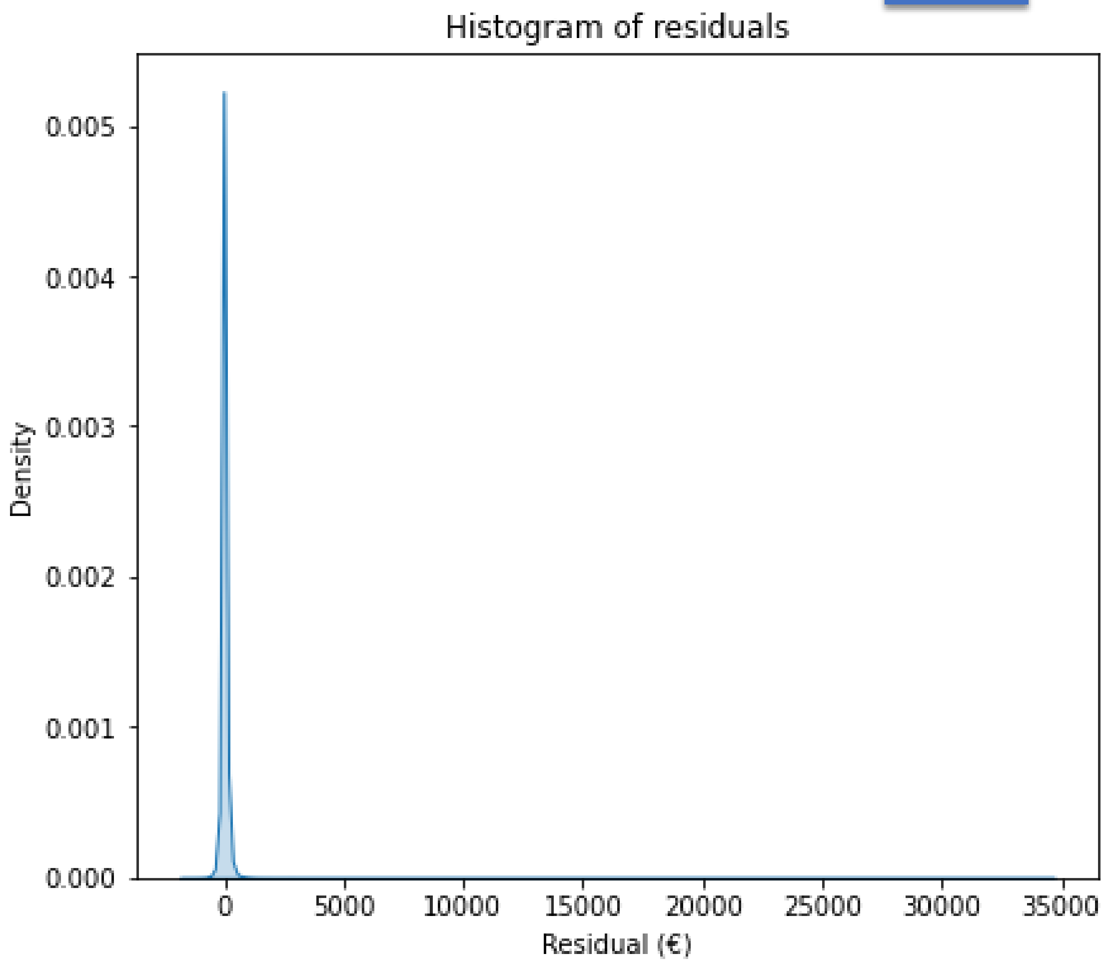
- Service charge
- Median city picture count
- Median zip code service charge
- Luxurious interior quality
- Living space
- Median city service charge
- Median city living space

Model selection: XGBoost

Data	Mean cv r2 score	Stdev of cv r2 scores	Mean cv mae score (€)	Stdev of cv mae scores (€)
All	0.774	0.073	114.73	29.66
Training	0.837	0.045	91.90	0.31


Plot of feature coefficient versus feature
coefficient
XGBoost coefficients
0.12
0.10
0.08
0.06
0.04
0.02
0.00
state_Baden_Württemberg
lift_Yes
state_Sachsen
state_Mecklenburg_Vorpommern
state_Hessen
condition_ripe_for_demolition
state_Saarland
state_Hamburg
state_Thüringen
newlyConst_Yes
balcony_Yes
hasKitchen_Yes
state_Sachsen_Anhalt
state_Nordrhein_Westfalen
typeOfFlat_raised_ground_floor
typeOfFlat_terraced_flat
typeOfFlat_loft
typeOfFlat_half_basement
typeOfFlat_apartment
typeOfFlat_maisonette
state_Schleswig_Holstein
typeOfFlat_ground_floor
condition_modernized
condition_negotiable
typeOfFlat_other
condition_fully_renovated
interiorQual_unknown
condition_unknown
medianServiceCharge_state
state_Niedersachsen
medianLivingSpace_zip
condition_need_of_renovation
condition_refurbished
interiorQual_simple
state_Brandenburg
medianThermalChar_state
state_Rheinland_Pfalz
medianLivingSpace_muni
medianLivingSpace_state
noRooms
medianNoRooms_zip
medianThermalChar_ct
balcony_No
condition_first_time_use_after_refurbishment
medianPictureCount_muni
medianNoRooms_muni
state_Bremen
typeOfFlat_penthouse
typeOfFlat_unknown
no_Listings_per_100people
medianServiceCharge_muni
condition_well_kept
hdi_2018
state_Berlin
hasKitchen_No
condition_mint_condition
gdp_per_capita_2018
medianPictureCount_state
interiorQual_sophisticated
medianServiceCharge_zip
newlyConst_No
medianLivingSpace_ct
interiorQual_normal
medianPictureCount_ct
condition_first_time_use
typeOfFlat_roof_storey
lift_No
interiorQual_luxury
serviceCharge
state_Bayern
medianHeatingCosts_state
livingSpace
medianServiceCharge_ct
Feature





Mean residual: €0 | Median residual: - €5 | Stdev: €157.5 | Skew: 46.5

Model application results

#	State	City/Town	Condition	rooms	Area (sqm)	Predicted rent (€)	Lower limit (€)	Set rent (€)	Higher limit (€)
1	Brandenburg	Uckermark	First time use	3.0	50.51	578.25	493.18	300.00	722.65
2	Schleswig Holstein	Dithmarschen	First time use after refurbishment	3.0	112.24	910.22	825.14	589.26	1054.61
3	Sachsen	Chemnitz	refurbished	2.0	51.67	641.11	556.03	518.00	785.50
4	Sachsen Anhalt	Halle Saale	negotiable	1.0	28.75	328.69	243.61	285.00	473.08
5	Hessen	Main Kinzig	Well kept	2.5	50.00	429.64	344.56	320.00	574.03

Conclusion / Recommendation

Determined most important factors for rent

- Living space
- Service charge
- Interior quality – luxurious

Predicted rent with XGBoost model

- R-squared: 0.77
- Mean absolute error: €114.73

Assumptions/Limitations/Opportunities

- Analysis of text features (i.e. description and facilities)
- Including time features in model
- Uncertainty around sampling methodology
- Higher compute capabilities for hyperparameter tuning

Questions