Analysis of Cantonal Rents in Switzerland (OFS, 2023)

Tramont Maxim

Nov 14, 2025

Contents

Chapter 1

Analysis of Cantonal Rents in Switzerland (FSO, 2023)

1.1 Project Objective

The objective of this project is to analyze the structural determinants of rents in Switzerland at the cantonal level. Based on official data from the Swiss Federal Statistical Office (FSO), this work aims to explain variations in average rent per square meter across cantons using demographic, economic, and real estate variables.

More specifically, the project seeks to identify which factors are most closely associated with the differences in rent levels observed between cantons and to quantify their relative influence. The emphasis is placed on the economic interpretation of the results, in order to link statistical observations to well-known housing market mechanisms, such as demographic pressure, supply–demand tensions, and the structure of the residential market.

The project follows an applied data analysis approach, combining descriptive statistics, visualization, and simple econometric modeling. Interpretable linear regression models are used as the main analytical tools, while cross-validation methods and a non-linear model (Random Forest) are employed as complementary tools to assess robustness and generalization capacity of the identified relationships.

The goal is therefore not only to predict rents, but above all to understand the structural determinants that explain territorial disparities in the Swiss rental market.

1.2 Installation of Dependencies

The working environment is based on Python ($\geq$ 3.10). The dependencies required to run the project are listed in the `requirements.txt` file.

```
python -m venv venv
source ./venv/bin/activate
pip install -r requirements.txt
```

1.3 Data Preparation and Normalization

The data used in this project come from several statistical tables published by the Swiss Federal Statistical Office (FSO) and describe various socio-economic and real estate characteristics of Swiss cantons for the year 2023. Since these datasets are provided separately, a preparation step was necessary in order to build a single, coherent, and exploitable dataset for analysis and modeling.

The data come exclusively from the Swiss Federal Statistical Office (FSO) and concern the year 2023. They are aggregated at the cantonal level, with each observation corresponding to a Swiss canton.

1.4 Target Variable

1.4.1 Average Rent per Square Meter (loyer_moyen_m2)

The average rent per square meter corresponds to the average annual rent, expressed in Swiss francs, divided by the living area of dwellings occupied by tenants in each canton. This variable constitutes the target variable of the project. It allows for comparison of rental price levels across cantons independently of dwelling size and represents a central indicator of pressure on the housing market.

1.5 Explanatory Variables

1.5.1 Population Density (densite_population)

Population density represents the number of inhabitants per square kilometer in each canton. It is used as an indicator of demographic pressure and urbanization intensity. High density is generally associated with increased housing demand and limited land availability, which may contribute to higher rents.

1.5.2 Vacancy Rate (taux_logements_vacants)

The vacancy rate measures the proportion of unoccupied dwellings relative to the total housing stock in each canton. It is a direct indicator of tension in the housing market. A low vacancy rate reflects supply scarcity relative to demand, which may lead to rising rents, while a high rate indicates a more relaxed market.

1.5.3 Total Number of Dwellings (nombre_logements_total)

The total number of dwellings corresponds to the overall size of the housing stock in each canton. This variable helps capture the structural supply of housing and is mainly used as a control variable to account for differences in size and structure across cantons.

1.5.4 Unemployment Rate (taux_chomage)

The unemployment rate is defined as the ratio between the number of unemployed persons and the total number of economically active persons in each canton. It serves as an indicator of local economic conditions and household purchasing power. The unemployment rate may influence rental demand and households' ability to afford high rent levels.

1.5.5 Homeownership Rate (taux_logements_proprietaire)

The homeownership rate corresponds to the proportion of dwellings occupied by their owners, expressed as a percentage of total occupied dwellings. This indicator reflects the structure of the local residential market by distinguishing cantons dominated by rental markets from those where ownership is more widespread. A low share of homeowners is generally associated with greater pressure on the rental market.

Housing and construction data are drawn from FSO housing stock statistics: https://www.bfs.admin.ch/bfs/fr/home/statistiques/construction-logement/logements.gnpdetail.2025-0428.html

Demographic and labor market data come from FSO population and employment statistics: https://www.bfs.admin.ch/bfs/fr/home/statistiques/population.html

1.6 Data Preparation Steps

First, each file was imported individually using the `pandas` library. The selected datasets include average rent per square meter, population density, vacancy rate, total number of dwellings, as well as the number of unemployed persons and economically active persons by canton. All tables share a common key corresponding to the canton, allowing them to be merged.

From labor market data, a synthetic indicator was computed: the unemployment rate. This was obtained by dividing the number of unemployed persons by the total number of economically active persons for each canton. Once constructed, intermediate variables (unemployed persons and active persons) were removed to avoid redundancy and limit collinearity among explanatory variables.

The different datasets were then merged by canton. This choice ensures that only complete observations present in all sources are retained in the final dataset. The result is a single table where each row corresponds to a canton and each column to a quantitative variable describing its socio-economic and real estate characteristics.

Finally, in order to make variables comparable and facilitate statistical analysis and machine learning model training, data normalization was applied. All quantitative variables were standardized using z-score normalization, which centers the data (zero mean) and scales them (unit variance). This transformation prevents variables with large scales (e.g., population density) from artificially dominating multivariate analyses.

In the final normalized file, a NaN value may appear, indicating that a value does not exist. This occurs because some cantons lack data for certain parameters. In such cases, missing values are imputed using the mean of the corresponding variable.

Two datasets were thus produced: a merged, non-normalized dataset retaining original units, and a normalized dataset used for exploratory analysis, correlation studies, and predictive modeling.

1.7 Exploratory Data Analysis

Exploratory analysis is based on the standardized (z-score) dataset, allowing direct comparison of variables. This step aims to describe the distribution of average rent per square meter and identify key relationships between the target variable and selected explanatory variables.

1.7.1 Distribution of Average Rent per m^2

The distribution of average rent per square meter shows significant dispersion across cantons. While most values cluster around the mean, some cantons exhibit notably higher or lower rents.

These differences reflect strong structural disparities, particularly between attractive urban areas and more rural regions. This heterogeneity justifies analyzing the determinants of rents.

1.7.2 Bivariate Analysis of Explanatory Variables

This section analyzes relationships between average rent per square meter and each explanatory variable using scatter plots and regression lines.

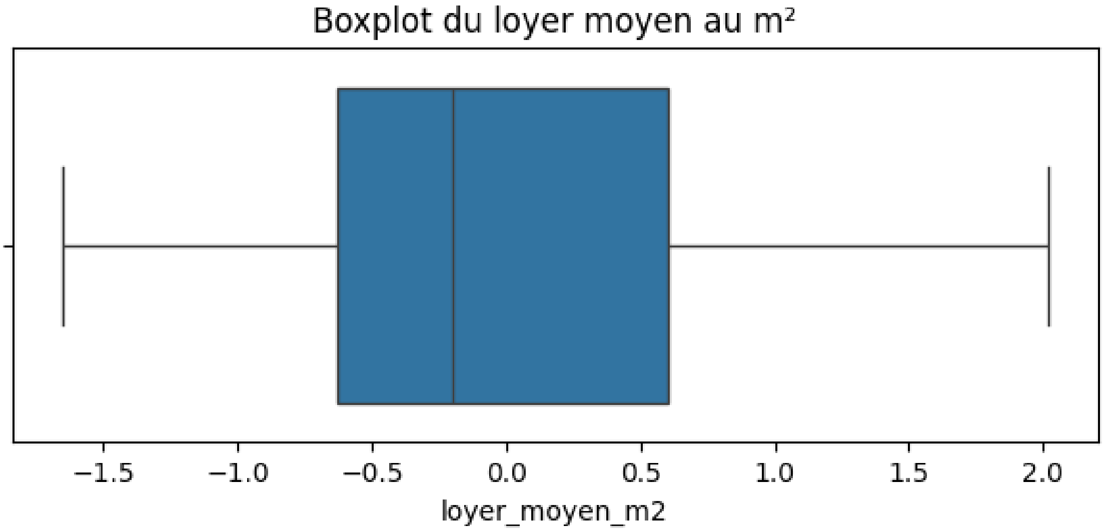


Figure 1.1: Distribution of average rent per m^2

1.7.2.1 Homeownership Rate

A strong negative relationship is observed between average rent and the homeownership rate (correlation of -0.72). Cantons dominated by rental markets generally exhibit higher rents, while those with higher proportions of homeowners display lower rents.

1.7.2.2 Vacancy Rate

The vacancy rate is also negatively correlated with average rent (correlation of -0.64). Cantons with low vacancy rates—indicating supply–demand tension—show higher rents, consistent with classical housing market mechanisms.

1.7.2.3 Population Density

Population density exhibits a moderate positive correlation with average rent (0.51). More densely populated, often more urbanized cantons tend to have higher rents, although this variable alone does not explain all observed differences.

1.7.2.4 Total Number of Dwellings

The total number of dwellings is positively correlated with average rent (0.43). This mainly reflects that cantons with large housing stocks are also economically and demographically attractive. This variable primarily plays a structural control role.

1.7.2.5 Unemployment Rate

The unemployment rate shows a moderate positive correlation with average rent (0.34), though the relationship is more diffuse. This variable appears to capture specific urban and socio-economic characteristics rather than a direct effect on rent levels.

1.7.3 Multivariate Analysis: Correlation Matrix

The correlation matrix summarizes linear relationships among all variables. It confirms:

- a strong negative correlation between average rent and vacancy rate,

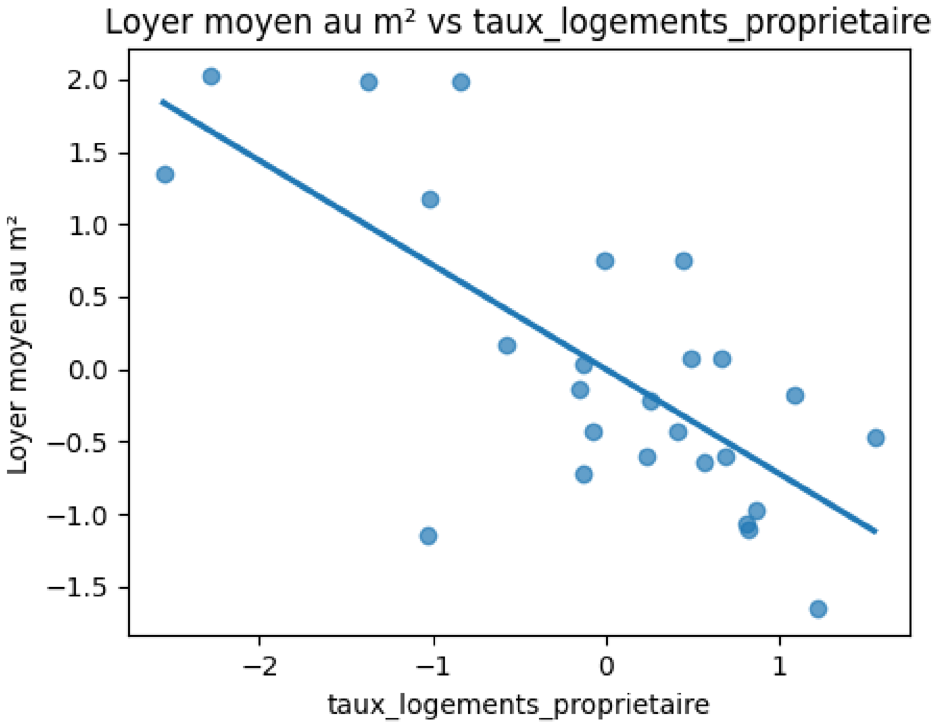


Figure 1.2: Relationship between average rent per m^2 and homeownership rate

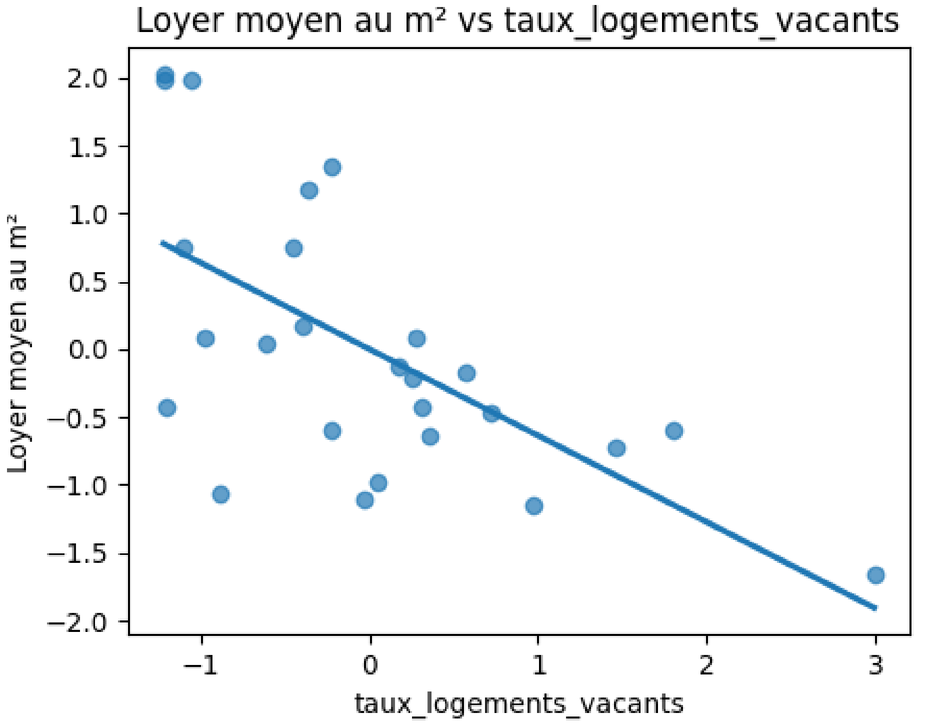


Figure 1.3: Relationship between average rent per m^2 and vacancy rate

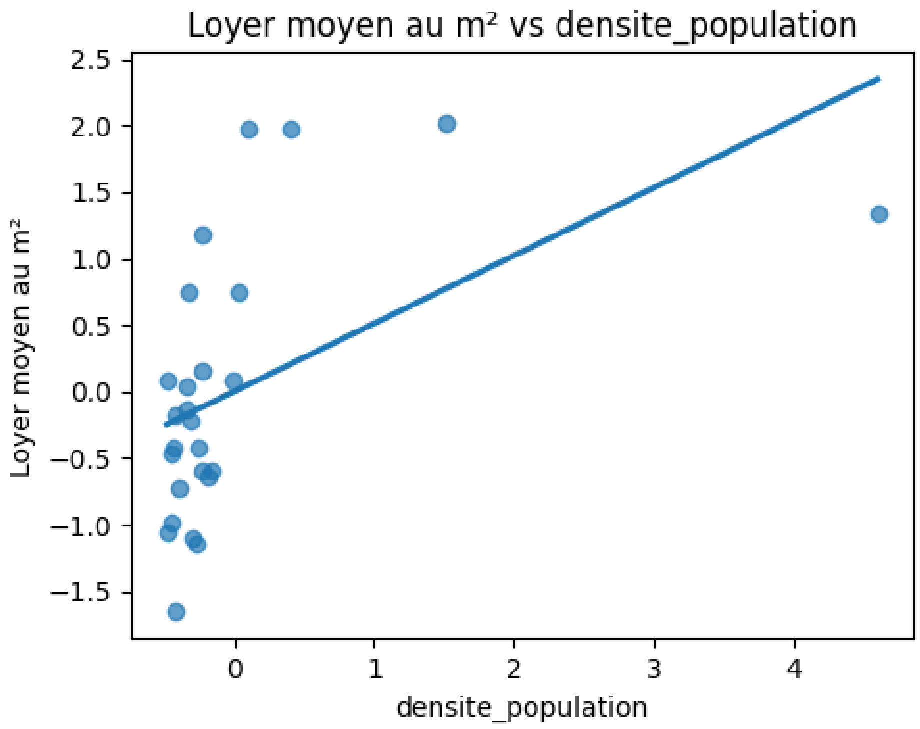


Figure 1.4: Relationship between average rent per m² and population density

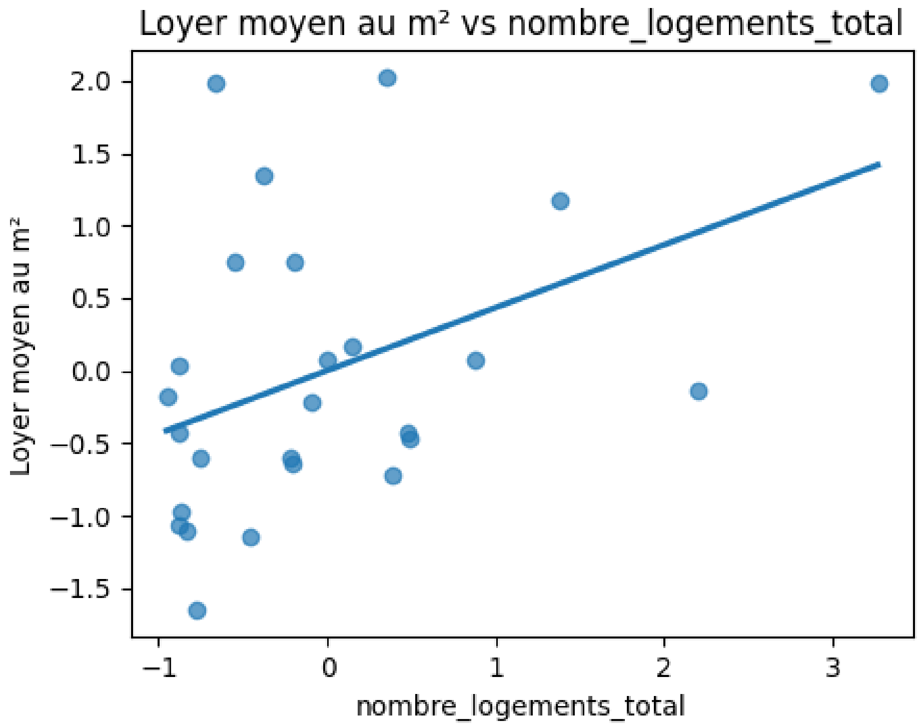


Figure 1.5: Relationship between average rent per m² and total number of dwellings

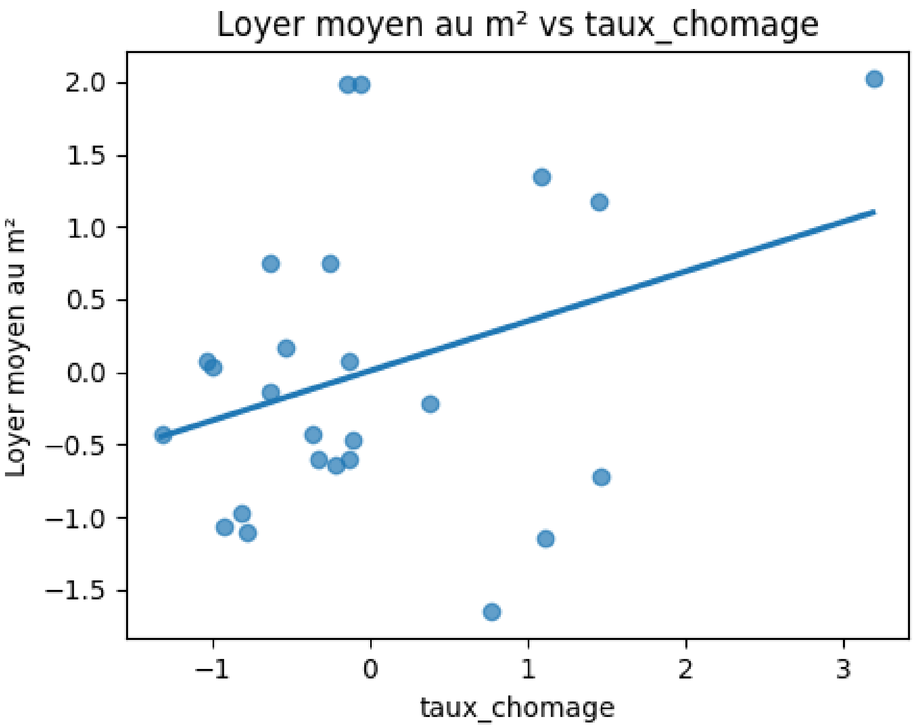
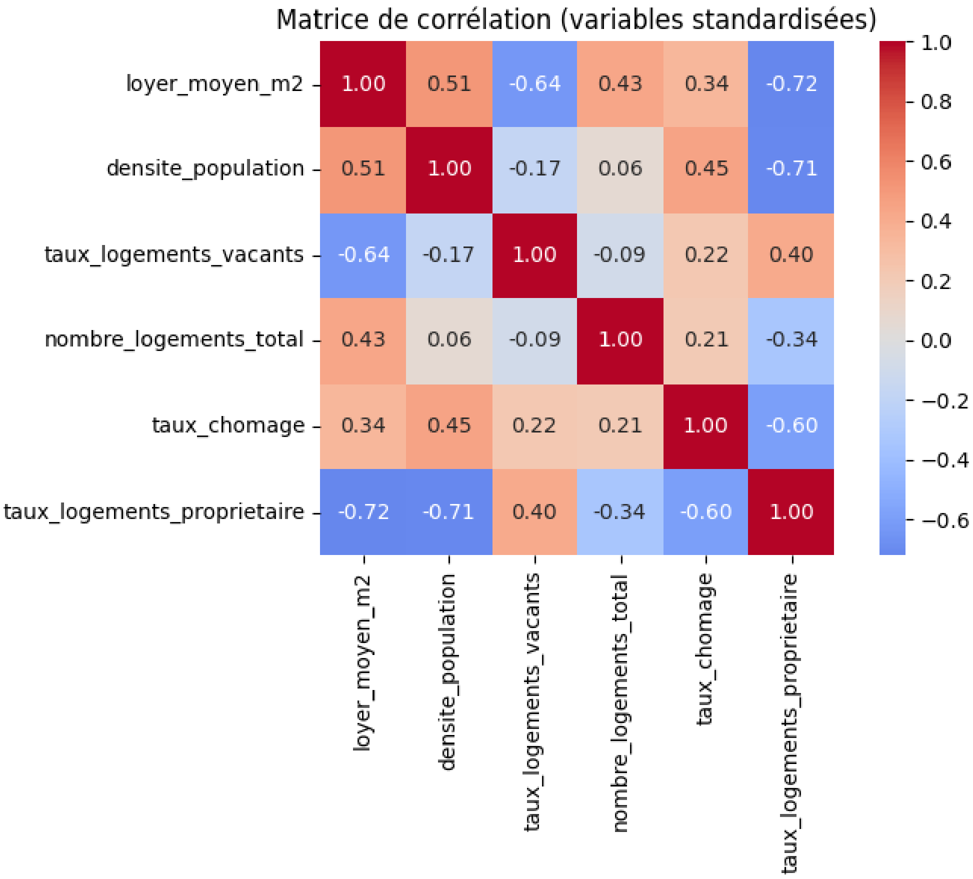


Figure 1.6: Relationship between average rent per m^2 and unemployment rate



	loyer_moyen_m2	densite_population	taux_logements_vacants	nombre_logements_total	taux_chomage	taux_logements_proprietaire
loyer_moyen_m2	1.00	0.51	-0.64	0.43	0.34	-0.72
densite_population	0.51	1.00	-0.17	0.06	0.45	-0.71
taux_logements_vacants	-0.64	-0.17	1.00	-0.09	0.22	0.40
nombre_logements_total	0.43	0.06	-0.09	1.00	0.21	-0.34
taux_chomage	0.34	0.45	0.22	0.21	1.00	-0.60
taux_logements_proprietaire	-0.72	-0.71	0.40	-0.34	-0.60	1.00

Figure 1.7: Correlation matrix of standardized explanatory variables

- a positive correlation with population density,
- notable correlations among explanatory variables, particularly between population density and homeownership rate (-0.71).

These relationships suggest partial multicollinearity, which must be addressed during modeling.

1.7.4 Summary of Exploratory Analysis

Exploratory analysis highlights that high rents are mainly associated with cantons characterized by strong urbanization, low vacancy rates, and a predominance of rental markets. However, no single variable fully explains observed disparities, justifying a multivariate approach.

1.8 Models Used: Linear Regression (OLS) and Ridge Regression

1.8.1 Model Objectives

The main objective is to analyze and predict average rent per square meter across Swiss cantons based on explanatory variables (population density, vacancy rate, unemployment rate, etc.). Two linear regression approaches are used: Ordinary Least Squares (OLS) and Ridge regression. These models quantify the impact of socio-economic factors on rents while ensuring robust and generalizable predictions.

1.8.2 Why Use These Models?

1.8.2.1 Linear Regression (OLS)

Simplicity and interpretability: OLS is a classical and simple method to model linear relationships between explanatory variables and average rent.

Coefficient interpretation: Each coefficient quantifies the effect of an explanatory variable on rent, which is crucial for economic interpretation.

1.8.2.2 Ridge Regression

Ridge regression introduces a penalty on coefficients to reduce the impact of highly correlated variables and prevent overfitting. It is particularly useful when multicollinearity exists (e.g., between population density and total dwellings).

By penalizing coefficients, Ridge improves generalization and reduces sensitivity to training data variations.

1.8.3 Random Forest

Random Forest is used to overcome limitations of linear models. This non-linear model captures complex interactions between variables without imposing linearity assumptions. It is especially relevant where economic mechanisms differ across cantons (urban vs rural, tight vs relaxed markets).

1.8.4 Cross-Validation

To validate model performance, cross-validation is applied. The dataset is split into multiple subsets to train and test models, reducing overfitting risk and enabling comparison of OLS and Ridge performance.

This process allows comparison in terms of predictive accuracy (MSE and R^2) while maintaining interpretability.

1.9 Formal Description of Models

Three supervised regression models are used: two interpretable linear models and one non-linear machine learning model focused on predictive performance.

1.9.1 Régression linéaire ordinaire (OLS)

OLS models the relationship between target variable y (average rent per m^2) and explanatory variables y $\mathbf{X} = (x_1, \dots, x_p)$ as :

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon$$

where:

- y is the average rent per square meter,
- x_j denotes the j-th explanatory variable,
- β_0 is the intercept,
- β_j is the coefficient associated with variable x_j,
- ε is the random error term.

The coefficients are estimated by minimizing the sum of squared residuals:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

This model serves as the statistical reference in the project, as it allows a direct interpretation of the effect of each variable on rent levels.

However, OLS regression can be sensitive to multicollinearity among explanatory variables and to overfitting, particularly in a small-sample context.

1.9.2 Ridge Regression

Ridge regression is a regularized extension of linear regression. It is based on minimizing the following cost function:

$$\min_{\beta} \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

where:

- $\lambda \geq 0$ is a regularization hyperparameter,
- the second term corresponds to an L_2-norm penalty on the coefficients.

This penalty:

- reduces the magnitude of the coefficients,
- limits the impact of multicollinearity,
- improves the model's generalization ability.

In this project, Ridge regression is used as a compromise between interpretability and predictive robustness, given the correlations observed among some explanatory variables during the exploratory analysis.

1.10 Random Forest Regression

The Random Forest model is an ensemble algorithm based on the aggregation of multiple decision trees, built on random subsamples of the data and of the explanatory variables.

Each tree produces a prediction $\hat{y}^{(t)}(x)$ of the average rent, and the final model prediction is obtained by averaging the individual predictions:

$$\hat{y}(x) = \frac{1}{T} \sum_{t=1}^{T} \hat{y}^{(t)}(x)$$

where T is the number of trees in the forest.

This model allows:

- capturing non-linear relationships,
- modeling complex interactions between variables,
- improving predictive performance relative to linear models.

In return, Random Forest is less interpretable, but it represents a relevant tool for comparing predictive performance with linear models in an applied data science framework.

1.11 Role of the Models in the Project

The three models play complementary roles:

- **OLS**: reference model and tool for economic interpretation,
- **Ridge**: regularized linear model favoring generalization,
- **Random Forest**: non-linear model focused on predictive performance.

This combination allows both the analysis of rent determinants and the comparison of different modeling approaches.

1.12 Results

1.12.1 Linear Regression with Cross-Validation

The following results were obtained for Ordinary Least Squares (OLS) regression using 5-fold cross-validation.

Cross-validation MSE scores for each fold:

```
[-1.42620849, -0.19716249, -0.44732334, -0.2113129, -0.93447193]
```

Model performance varies across data subsets. In particular, the first score (-1.426) is relatively large in absolute value, suggesting a higher prediction error on that subset. The other scores are closer to -0.5, which is more consistent.

This yields an average mean squared error (MSE) of **0.643**.

This means that, on average, the error between observed and predicted values is 0.643 CHF/m^2. This error can be considered relatively low, especially given that average rents across cantons range between approximately 10 CHF/m^2 and 25 CHF/m^2.

R^2 scores for each fold:

- Fold 1: -1.4045
- Fold 2: 0.8724
- Fold 3: -1.1113
- Fold 4: 0.3255
- Fold 5: 0.4145

The R^2 values show substantial variation across folds. In particular, negative R^2 values (such as Fold 1 and Fold 3) indicate that the model performs poorly on these subsets and is even less accurate than a model that simply predicts the mean rent. This suggests that the linear relationship between the explanatory variables and the average rent is imperfect.

1.12.2 Ridge Regression with Cross-Validation

For Ridge regression, the following cross-validation results were obtained.

Cross-validation MSE scores:

```
[0.63289251, 0.08279395, 0.42805234, 0.19576885, 0.88375662]
```

The MSE scores for Ridge are more dispersed than those of linear regression, but the average MSE is lower, indicating better generalization.

The average MSE is **0.4447**.

This means that, on average, the prediction error of the Ridge regression model is 0.4447 CHF/m^2. This lower MSE suggests that Ridge generalizes better than OLS.

Cross-validation R^2 scores:

- Fold 1: -0.0670
- Fold 2: 0.9464
- Fold 3: -1.0204
- Fold 4: 0.3751
- Fold 5: 0.4463

The average R^2 for Ridge is **0.1361**, which remains low but positive. This indicates that Ridge explains a small share of the variance in rents. The R^2 is more stable than for OLS, but still insufficient to explain a large portion of rent variation.

1.12.3 Random Forest with Cross-Validation

A Random Forest model was then estimated to capture potentially non-linear relationships between explanatory variables and the average rent.

Cross-validation MSE scores:

```
[0.229, 0.230, 0.970, 0.723, 0.376]
```

The average MSE is **0.505**.

This value is slightly higher than that of Ridge but remains lower than that of OLS.

Cross-validation R^2 scores:

- Fold 1: 0.736

- Fold 2: 0.648
- Fold 3: −1.020
- Fold 4: 0.333
- Fold 5: 0.742

The average R^2 is **0.288**, which is the best performance among the three tested models. Despite substantial variability across folds (including one negative R^2), Random Forest overall explains a larger share of the variance in rents.

1.12.4 Model Comparison

The three tested models (OLS, Ridge, and Random Forest) exhibit different behaviors and performance levels, allowing for a clearer understanding of the trade-offs between interpretability and predictive capacity.

1.12.4.1 Linear Regression (OLS)

OLS provides a solid starting point for analysis. It is easy to interpret and helps identify linear relationships between explanatory variables and the average rent. However, cross-validation results reveal strong performance variability across folds, with several negative R^2 values. This indicates poor generalization, and the model is sometimes less accurate than a simple mean-based predictor. OLS therefore appears insufficient to capture the complexity of the relationships in the data.

1.12.4.2 Ridge Regression

Ridge regression improves performance relative to OLS by reducing the average MSE. Regularization limits the impact of multicollinearity and reduces overfitting. The model is more stable under cross-validation, although the R^2 remains relatively low. Ridge thus represents an appealing compromise, preserving the interpretability of a linear model while offering better generalization.

1.12.4.3 Random Forest

The Random Forest model achieves the highest average R^2 among the three approaches, suggesting that it captures rent variations more effectively. Its ability to model non-linear relationships and complex interactions between explanatory variables is a major advantage. However, performance remains variable across folds, likely due to the small sample size. Moreover, Random Forest is less interpretable than linear models, limiting its usefulness for detailed economic analysis of rent determinants.

1.13 Overall Assessment

Model	Average MSE	Average R^2	Interpretability
OLS	0.643	-0.181	High
Ridge	0.445	0.136	High
Random Forest	0.505	0.288	Low

OLS is useful for interpretation and exploratory analysis but exhibits weak generalization capacity.

Ridge improves robustness and reduces prediction error while remaining interpretable.

Random Forest delivers the strongest overall predictive performance, at the cost of interpretability.

The choice of model therefore depends on the objective: economic understanding of rent determinants or predictive performance. In this project, linear models are primarily used for interpretation, while Random Forest is employed to assess the maximum predictive potential of the available variables.