Explorative Datenanalyse – Bericht

Yannic Lais, Marvin von Rappard, Luca Mazzotta

January 16, 2023

Abstract

In diesem Bericht wird eine explorative Datenanalyse durchgeführt, um die Struktur, die Verteilung und die Beziehungen im Datensatz zu untersuchen. Dabei werden verschiedene Arten von Diagrammen, Methoden und visuelle Techniken verwendet. Die Ergebnisse der Analyse dienen als Grundlage für weitere Entscheidungen im Laufe des Projekts. Weitere Informationen sind in folgendem Github Repository zu finden:
https://github.com/marvinvr/fhnw-cml1

Contents

1 Immobilien Standorte

Um einen ersten Einblick zu erschaffen, wurden alle Datenpunkte über die Schweizer Landkarte geplottet, um zu sehen, ob es in gewissen Regionen mehr Immobilien hat als in anderen.

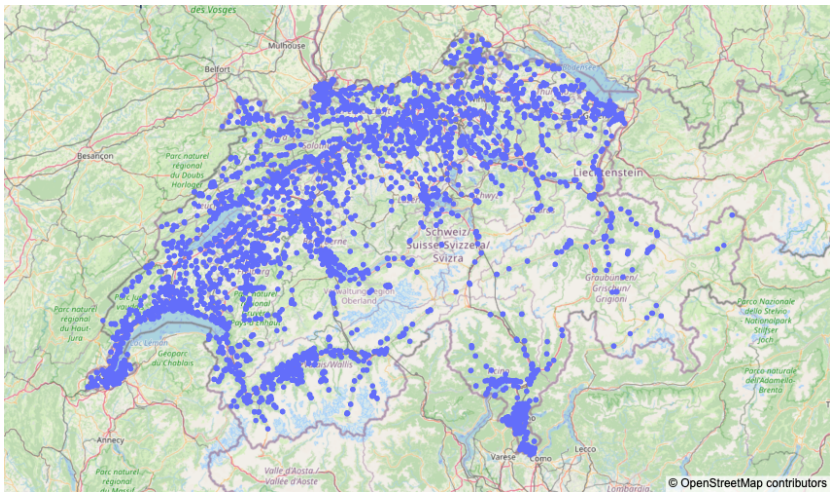

Figure 1: Standorte der Immobilien aus dem Datensatz

Man erkennt, dass die Immobilien im Westen dichter sind als im Osten, dementsprechend gibt es im Datensatz mehr Immobilien im Westen als im Osten. Das könnte für die später erstellten Vorhersagemodelle heissen, dass diese für Vorhersage der Immobilien im Westen genauer sind als im Osten.

2 Preisverteilung

2.1 Preisverteilung von allen Daten

Die Preise wurden mittels Boxplot untersucht, dieser gibt einen guten Einblick in die Verteilung der Preise und in den Ausreissern an.

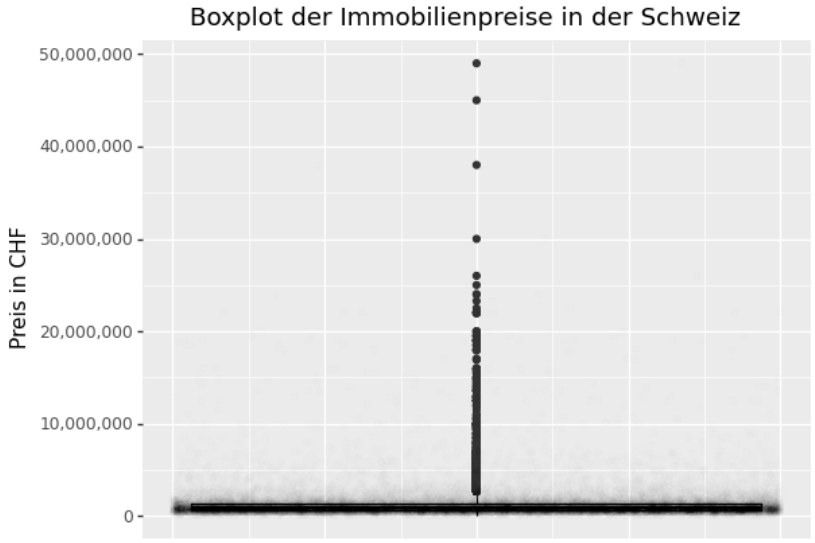

Figure 2: Boxplot der Immobilienpreise in der Schweiz

Im Plot (2) kann man aufgrund der Ausreisser nicht viel erkennen, deshalb wurde im nächsten Schritt die y-Achse limitiert.

2.2 Preisverteilung Daten ohne Ausreisser

Figure 3: Boxplot der Immobilienpreise in der Schweiz, ohne Ausreisser

Nach dem Limitieren der y-Achse und nach dem Berechnen der Quantile und des Medians erkennt man relativ gut, dass 50 % der Daten sich zwischen 575'000.00 CHF und 1'415'000.00 CHF befinden. Der Median ist bei 860'000.00 CHF.

2.3 Untersuchung der Ausreisser

Als Ausreisser wurden im Boxplot alle Immobilien definiert, die einen Preis von über 2'675'000.00 CHF besitzen. Um zu schauen, ob gewisse Typen in den Ausreissern mehr vorhanden sind, als andere, wurden diese in einem Barplot dargestellt.

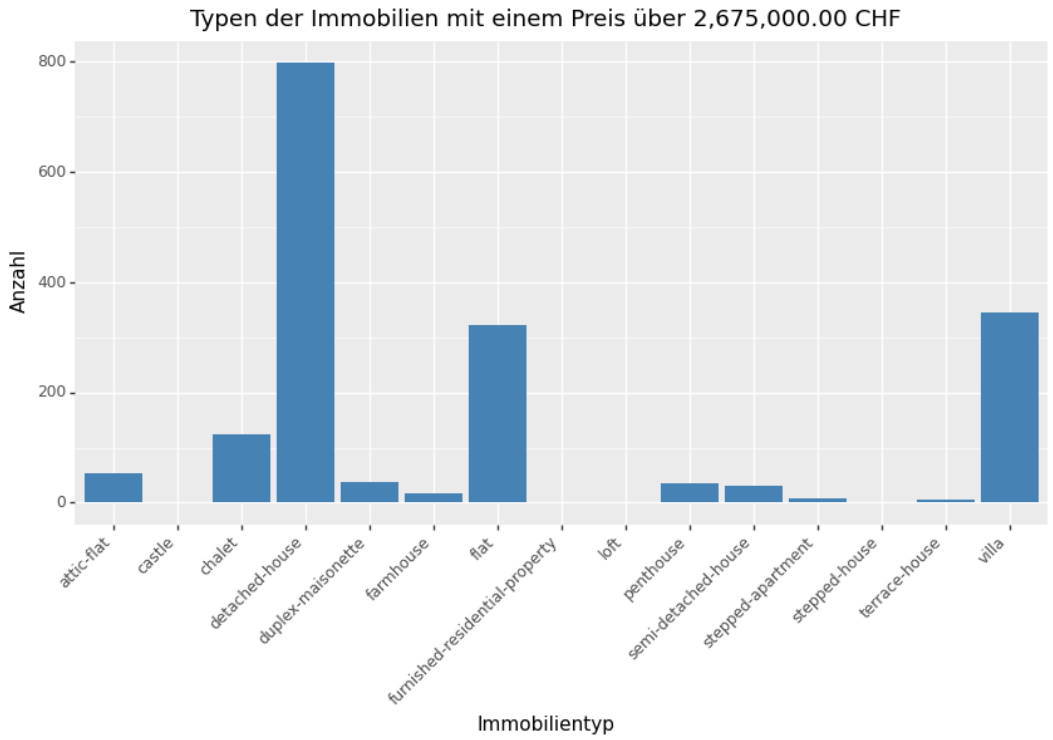

Figure 4: Immobilientypen der Ausreisser

Die meisten Ausreisser sind vom Typ «detached-house», mit ca. 800 Immobilien, gefolgt vom Typ «villa» mit ca. 350 Immobilien.

2.4 Preisverteilung der Immobilientypen

Um die genaue Verteilung einzelner Immobilientypen zu sehen, wurde ein Dichte-Plot für die wichtigsten Typen erstellt.

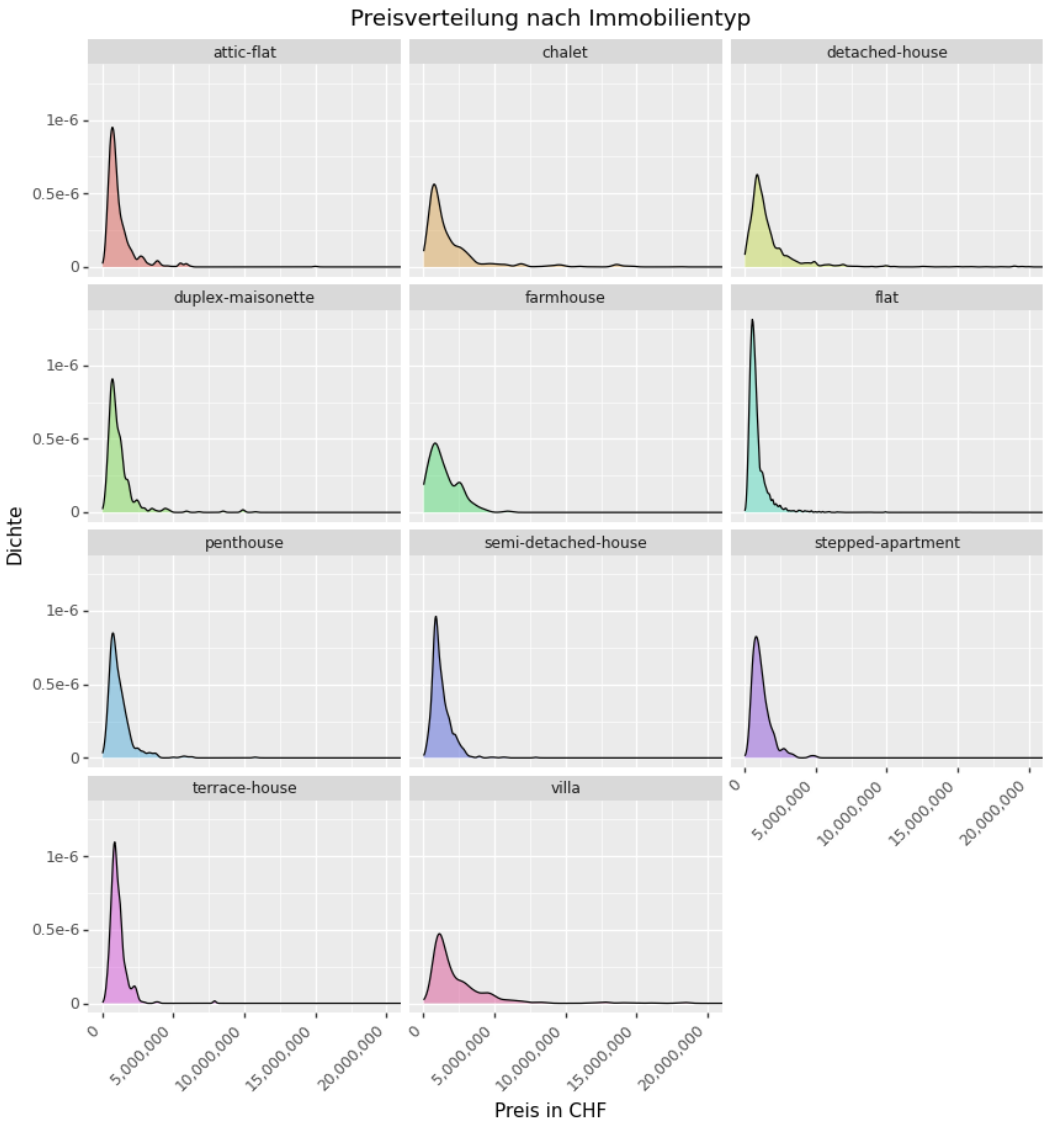

Figure 5: Preisverteilung nach Immobilientyp

Anhand dieser Verteilungen kann man passende Transformationen herausfinden, die für das Trainieren von Modellen nützlich sein könnten.
Man würde zum Beispiel folgende Transformationen anwenden:
Linksschiefe Verteilung: Log-Transformation oder Box-Cox Transformation
Rechtsschiefe Verteilung: Box-Cox Transformation oder Ansatz mit Freiheitsgraden.

Zusätzlich wurde ebenfalls ein Dichte Plot für alle Immobilientypen gemacht.

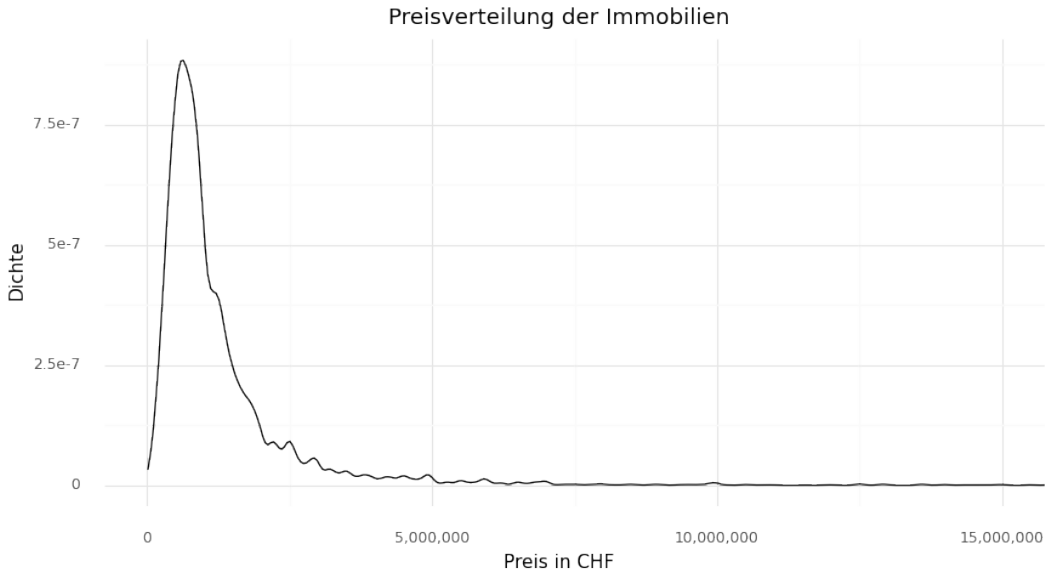

Figure 6: Preisverteilung der Immobilien

Hier erkennt man, dass die meisten Immobilien zwischen 0.00 CHF - 2'500'000.00 CHF liegen, dies ist wichtig zu sehen, da die Modelle in diesem Abschnitt die meisten Datenpunkte haben und somit diesen Bereich am besten verstehen wird.

3 Abhängigkeit der Attribute

3.1 Korrelationsmatrix

Eine Korrelationsmatrix ist eine Tabelle, welche die Korrelationen zwischen verschiedenen Attributen darstellt. Jeder Eintrag in der Matrix repräsentiert die Korrelation zwischen zwei Variablen. Die Diagonal Werte in der Matrix sind immer 1, da jede Variable mit sich selbst perfekt korreliert. Die Korrelationsmatrix wird hier verwendet, um Beziehungen zwischen Attributen zu identifizieren, diese werden anschliessend in den nächsten Schritten genauer untersucht.

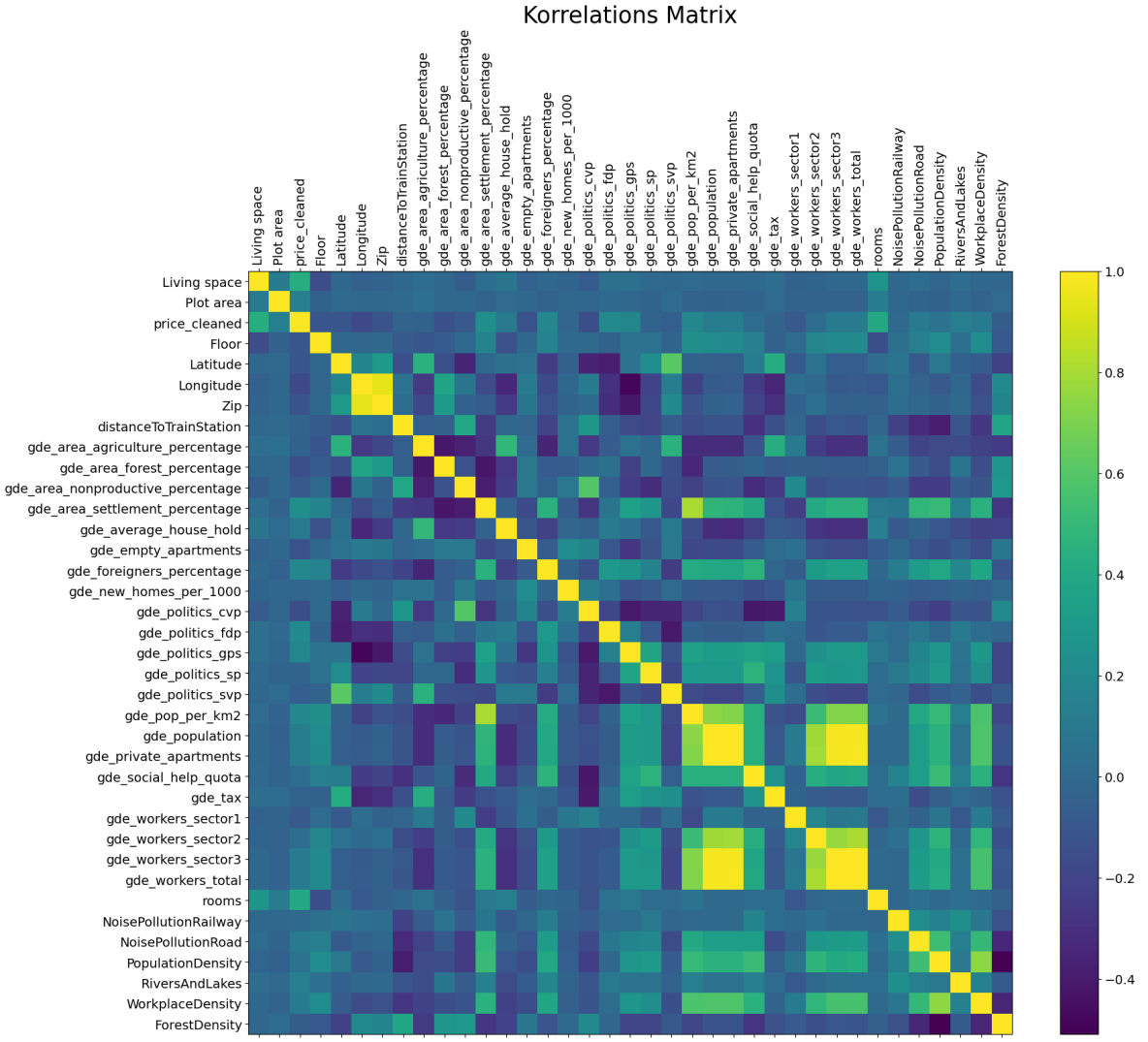

Figure 7: Korrelationsmatrix

Anhand der Farben der einzelnen Felder kann man erkennen, wie sehr verschiedene Attribute miteinander korrelieren. Einige Korrelationen sind von vor hinein klar und ergeben Sinn, ohne dass man diese weiter untersuchen muss, wie die Korrelation zwischen «gde_population» und «gde_workers_total» oder die Korrelation zwischen «gde_population» und «gde_private_apartments». Einige Korrelationen, welche nicht direkt erklärbar sind, werden in den nächsten Schritten untersucht. Um dies zu untersuchen, wurden alle Attribut-Paare genommen, welche eine Korrelation von über 0.7 haben. Da die Werte 1.0 für eine perfekte positive Korrelation und 0 für keine Korrelation sprechen, spricht man bei 0.7 bereits von einer starken positiven Korrelation. Da in den Daten keine starken negativen Korrelationen enthalten sind, werden nur positive Korrelationen untersucht, die nicht direkt erklärbar sind.

3.2 Korrelation zwischen «Longitude» und «Zip» (0.94)

Die Korrelation zwischen «Longitude» und «Zip» beträgt 0.94, dementsprechend spricht man hier von einer sehr starken positiven Korrelation, was man im folgenden Plot sehen kann.

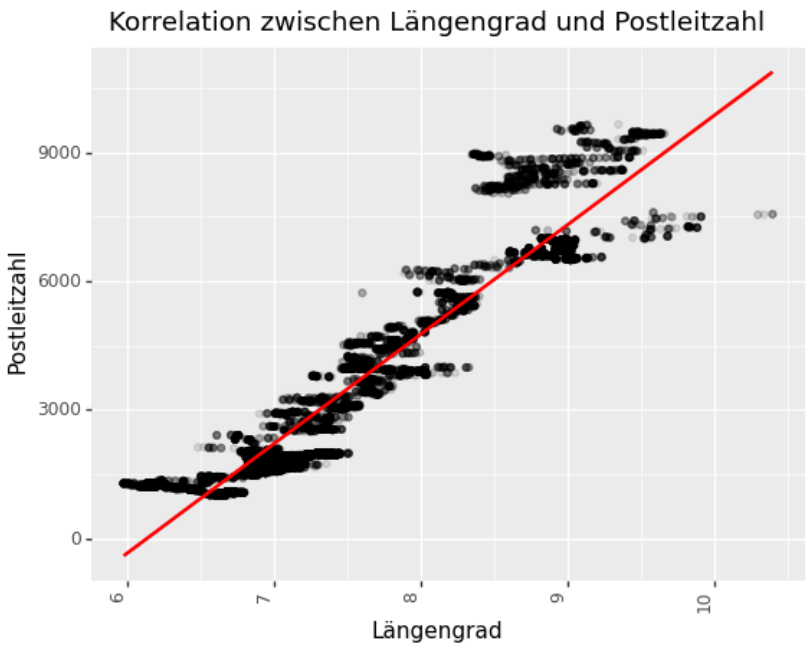

Figure 8: Korrelation zwischen Langengrad und Postleitzahl

Der Grund für diese Korrelation liegt an der Art, wie Postleitzahlen in der Schweiz vergeben wurden.

Figure 9: Postleitzahlen in der Schweiz

Wie man sehen kann, wurden diese entlang der Längengrade vergeben, welche die positive Korrelation erklären.

3.3 Korrelation zwischen «WorkplaceDensityL» und «gde_population» (0.74)

Die Korrelation zwischen «WorkplaceDensityL» und «gde_population» beträgt 0.74, hier spricht man von einer starken positiven Korrelation.

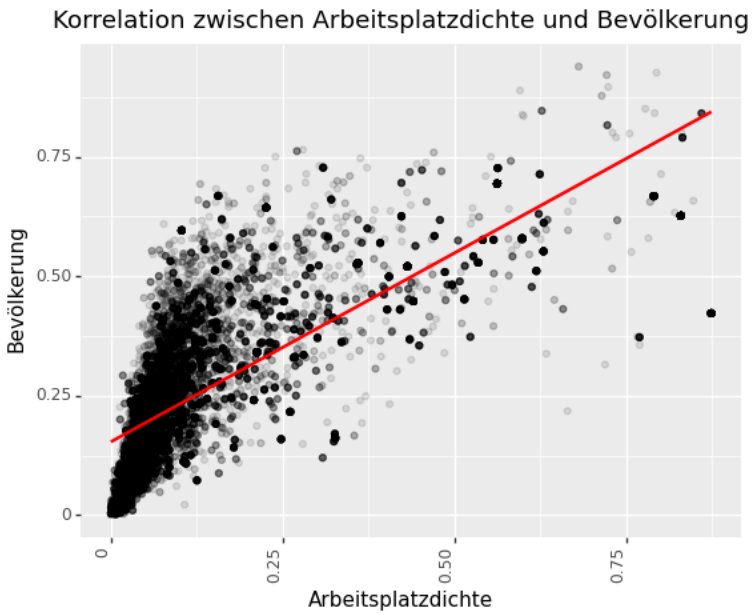

Figure 10: Korrelation zwischen Arbeitsplatzdichte und Bevölkerung

Im Gegensatz zur Korrelation von «Longitude» und «Zip» (8) erkennt man hier, dass die rote Gerade nicht so gut zu den Punkten passt wie bei 8. Dennoch ist hier eine positive Korrelation zu erkennen, was bedeutet, dass es eine Beziehung zwischen zwei diesen zwei Attributen gibt, bei der eine Erhöhung von einem Attribut in der Regel mit einer Erhöhung vom anderen Attribut einhergeht und umgekehrt.

3.4 Korrelation zwischen «gde_area_settlement_percentage» und «gde_pop_per_km2» (0.81)

Die Korrelation zwischen «WorkplaceDensityL» und «gde_population» beträgt 0.81. Diese starke positive Korrelation lässt sich auch im folgenden Plot schön darstellen.

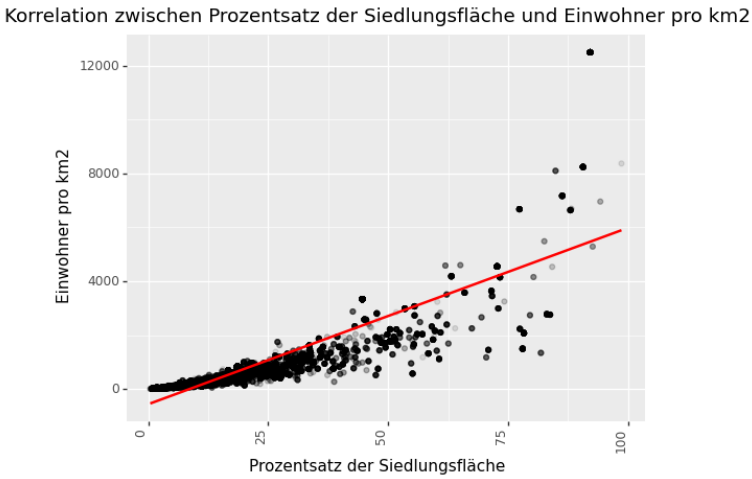

Figure 11: Korrelation zwischen Prozentsatz der Siedlungsflache und Einwohner pro km2

4 Rückblick

Grundsätzlich sind die Daten nach einer ausführlichen Bereinigung sehr nützlich. Dank einer ausführlichen Datenanalyse können nun mehrere Eigenarten bestimmt werden:

- Die Datenpunkte sind sehr ungleich auf die verschiedenen Immobilientypen verteilt.
- Es gibt grosse Schwankungen in den Preisen.
- Die Daten sind sehr unregelmässig auf die Schweiz verteilt.
- Korrelationen zwischen bestimmten Attributen wurden erkannt.
- Mithilfe von Variablentransformationen kann die Verteilung verbessert werden.