Bundesweite Vermieterbefragung 2026

Bericht zur Datenaufbereitung

Prof. Alex Stomper, Prof. Gabriel Ahlfeldt, Dr. Christoph Scheuch

21. April 2026

1 Überblick

Dieser Bericht dokumentiert die Schritte der Datenaufbereitung und Auswertung der bundesweiten Vermieterbefragung 2026. Die Pipeline ist in vier R-Skripte gegliedert, die aufeinander aufbauen: Import und Labeling der Rohdaten, Filterung nach Qualitätskriterien, Erstellung von Post-Stratifikationsgewichten auf Ebene der Haus & Grund Landesverbände sowie gewichtete Auswertung und Export nach Excel. Ziel der Aufbereitung ist ein konsistenter, dokumentierter Datensatz, der Vergleichbarkeit mit der Vorjahreswelle herstellt, wo möglich, und die strukturellen Neuerungen des Fragebogens 2026 sauber abbildet.

Im Vergleich zur Welle 2025 ergeben sich mehrere inhaltliche Brüche, die in der Aufbereitung berücksichtigt werden mussten:

- Die Zuordnung der F2-Blöcke zu Ein-/Zweifamilien- bzw. Mehrfamilienhäusern wurde getauscht.
- Der Block zur Mietersuche erfasst nun die Informationsgrundlagen für die Mietpreisfestlegung, die gesamten Mieteinnahmen je Objekt wurden durch die Nettokaltmiete pro Quadratmeter der zuletzt vermieteten Wohnung ersetzt.
- Der Fragenblock zur kommunalen Wärmeplanung ist entfallen.
- Neu hinzugekommen sind Fragen zur Vermietung leerstehender Wohnungen sowie die Zusatzfragen ZF1 bis ZF4 zu hypothetischen Regulierungsmaßnahmen und Verhaltensreaktionen.

2 Import und Aufbereitung der Rohdaten

2.1 Umbenennung und Typisierung

Wir importieren die Rohdaten aus dem Excel-Export, welche technische Variablennamen aus dem Fragebogenwerkzeug (etwa `F2.2.1A1`) tragen. Diese werden in zwei Schritten bereinigt: Punkte werden durch Unterstriche ersetzt, anschließend erhalten die inhaltlich zentralen Blöcke sprechende Namen. Die Gebäudezählungen werden als `Zahl_EFH_ZFH`, `Zahl_WEG` und `Zahl_MFH` geführt, die zugehörigen Wohneinheiten-Blöcke (`F2.2.3A`, `F2.3.3A`, `F2.4.3A`) erhalten präfixierte Namen `WE_EFH_ZFH_1` bis

`WE_MFH_10`. Wichtig ist die Berücksichtigung des Rollentauschs zwischen 2025 und 2026: F2.2 bezieht sich nun auf Ein-/Zweifamilienhäuser, F2.4 auf Mehrfamilienhäuser.

Numerische Variablen werden explizit per als numerisch typisiert, wobei Postleitzahlen bewusst als Text erhalten bleiben, damit führende Nullen nicht verloren gehen. Textkategorische Variablen aus den Blöcken F7_1 und F7_2 werden als Faktoren mit expliziter Levels-Reihenfolge gemäß Codebuch kodiert. Dadurch bleibt die vorgesehene Antwortreihenfolge auch in Auswertungen und Exporten erhalten.

2.2 Labeling

Alle inhaltlich relevanten Variablen erhalten Variablen- und Wertelabels. Dies hat zwei Effekte: Erstens wird der Datensatz beim Export als Stata-Datei (`.dta`) selbsterklärend, zweitens können Auswertungen die Labels direkt für Tabellenüberschriften und Kategorienbeschriftungen nutzen. Gelabelt werden unter anderem Geschlecht, Beschäftigungsstatus, Bestand und Bestandstyp, Mietpreisänderungen und ihre Gründe, Mietanpassungsverhalten, Förderarten, Modernisierungshemmnisse, Konfliktarten, Gebäudemerkmale und die neuen Zusatzfragen ZF1 bis ZF4.

2.3 Abgeleitete Variablen

Aus den Rohwerten werden mehrere Analysevariablen konstruiert:

- **Alter und Altersklassen**: aus `GeburtsjahrA1` wird `age = 2026 - Geburtsjahr` gebildet, anschließend in sechs Klassen gruppiert (unter 20, 20–40, 40–60, 60–80, 80–100, 100+).
- **Mieteinnahmen**: die präzise Eingabe `MieteinnahmenA1` wird in dieselben Spannen überführt wie die grobe Spannen-Angabe und mit dieser zusammengeführt, sodass eine einheitliche kategoriale Variable entsteht.
- **Gebäudebestand und Bestandstyp**: die Summe der drei Gebäudearten ergibt den Gesamtbestand, eine achtstufige Typologie klassifiziert Befragte nach der Kombination der gehaltenen Gebäudearten (kein Eigentum, nur MFH, nur WEG, nur EFH/ZFH, sowie alle Mischformen).
- **Wohneinheiten**: Summen pro Gebäudetyp aus den bis zu zehn erfassten Objekten, Durchschnitt pro Gebäude sowie Gesamtsumme; zusätzlich eine Spannen-Klassifikation (1–2, 3–5, 6–10, 11–15, mehr als 15 Wohneinheiten).
- **Baualtersklassen**: acht Klassen von „bis 1918" bis „2015 und später".
- **Eigentumsdauer**: kontinuierlich und in acht Klassen von unter 5 Jahren bis 60 Jahre und länger.
- **Gebäudetyp des ausgewählten Objekts**: aus der Variable `hv1` wird rekonstruiert, welcher Gebäudetyp (MFH, WEG, EFH/ZFH) für die detaillierten Folgefragen ausgewählt wurde.

2.4 Umgang mit „kA"-Angaben beim Bestand

Befragte, die die Gebäudezahl nicht beziffern konnten, aber die zugehörige „keine Angabe"-Flag gesetzt haben (`Zahl_*_kA == 1`), werden mit null Gebäuden kodiert. Das verhindert, dass sie aus den

Bestandsauswertungen herausfallen, und entspricht der inhaltlichen Lesart, dass diese Befragten in der jeweiligen Gebäudekategorie nichts halten.

3 Filterung

Aus den Zeitstempeln `START` und `END` werden zunächst `start_time`, `end_time`, `date` und `duration` (in Sekunden) abgeleitet, die anschließend für die zeitbasierten Filter genutzt werden. Die einzelnen Filter und ihre Auswirkungen sind:

Tabelle 1: Filterschritte

Filter	Begründung	Entfernt
`Vermietung == 1`	nur tatsächlich Vermietende	~7 %
Interviewdatum ab 21.01.2026	frühere Datensätze stammen aus Tests	0 %
`duration > 60` Sekunden	Ausschluss unrealistisch kurzer Antworten	~5 %
Alter zwischen 18 und 110 Jahren oder fehlend	Plausibilitätsgrenze	~<1% %
`HG_member == 1`	Fokus auf Haus & Grund-Mitglieder	~6 %
gültiger Vereinsname (`VereinA1` nicht fehlend)	Voraussetzung für die Gewichtung	~7 %

Durh die Filterschritte reduziert sich der Datensatz von 26,011 auf 20,136 (um etwa 23%). Wir filtern die Daten an dieser Stelle bewusst nicht nach `STATUS == 5` (vollständige Befragung), da für die Vereins- und Verbandsebene auch Teilbefragungen nützlich sind. Dieser Filter würde weitere 36% der Beocbachtungen ausschließen. Die Spalte `vollstaendig` gibt an, ob ein Fragebogen vollständig abgeschlossen wurde.

Abbildung Abbildung 1 zeigt die täglichen Fragebogeneingänge über die gesamte Feldzeit. Der auffälligste Befund ist ein starker Peak am 27. März 2026, an dem ein Vielfaches der üblichen Tageseingänge verzeichnet wurde. Wie Tabelle 2 zeigt, geht dieser Peak auf gezielte Mobilisierung durch wenige Ortsvereine zurück — allen voran Haus & Grund Waiblingen sowie Winnenden und Umgebung.

Tabelle 2: Top-5 Vereine am 27. März 2026

Verein	n
Haus & Grund Waiblingen, Winnenden und Umg. e.V.	170
Haus & Grund Berlin-Neukölln e.V.	35
Haus & Grund Dortmund e.V.	33
Haus & Grund Berlin-Schöneberg/Friedenau	15
Haus & Grund Stuttgart	12

Um zu prüfen, ob die am Peak-Tag eingegangenen Fragebögen qualitativ auffällig sind, wurden Interviewdauer, Straightlining-Rate und inhaltliche Streuung verglichen. Abbildung 2 zeigt die Verteilung der Interviewdauer für den Peak-Tag und die übrige Feldzeit, jeweils für alle Interviews und für vollständig abgeschlossene. Die Kurven verlaufen in beiden Facetten weitgehend deckungsgleich; der Median beider Gruppen liegt in einem plausiblen Bereich, ohne dass die Peak-Tag-Interviews systematisch kürzer ausfallen.

Tabelle 3: Straightlining-Anteil und Streuung bei ZF-Items. Straightlining bezeichnet das Ankreuzen derselben Antwortkategorie über alle Items einer Batterie hinweg (SD = 0).

	n	Anteil Straightlining	Median SD
Übrige Feldzeit	19276	3.0%	1.37
27. März 2026	847	3.6%	1.41

Auch beim Antwortverhalten innerhalb der ZF-Itembatterien ergeben sich keine Auffälligkeiten (Tabelle 3): Der Anteil der Straightliner — also Befragter, die über alle Items hinweg dieselbe Kategorie ankreuzten (SD = 0) — ist am Peak-Tag nicht erhöht, und die mediane Streuung unterscheidet sich kaum von der übrigen Feldzeit. Auf Basis dieser Prüfungen wurde der Peak-Tag nicht aus der Analyse ausgeschlossen.

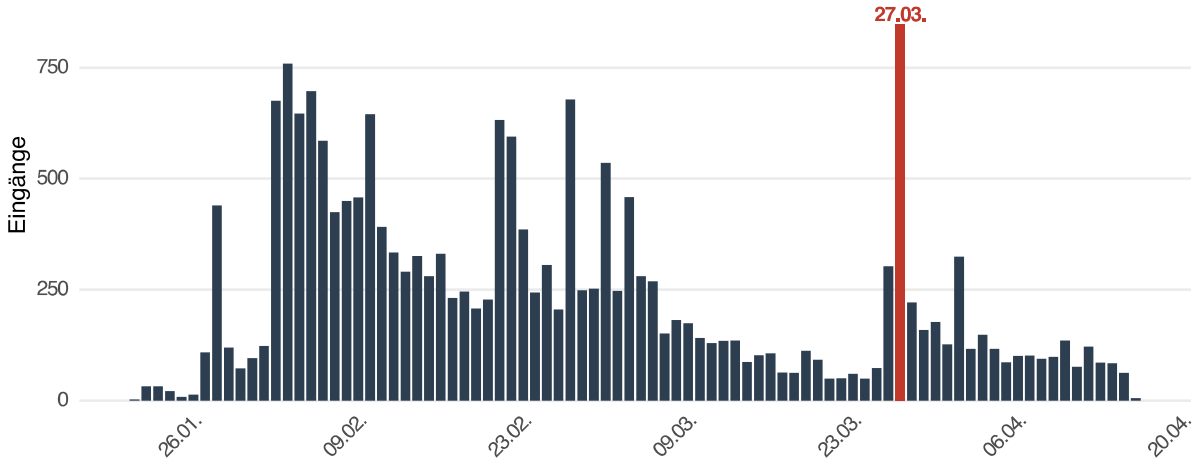

Abbildung 1: Tägliche Fragebogeneingänge über die gesamte Feldzeit. Der Peak am 27. März 2026 geht auf gezielte Mobilisierung durch einzelne Ortsvereine zurück.

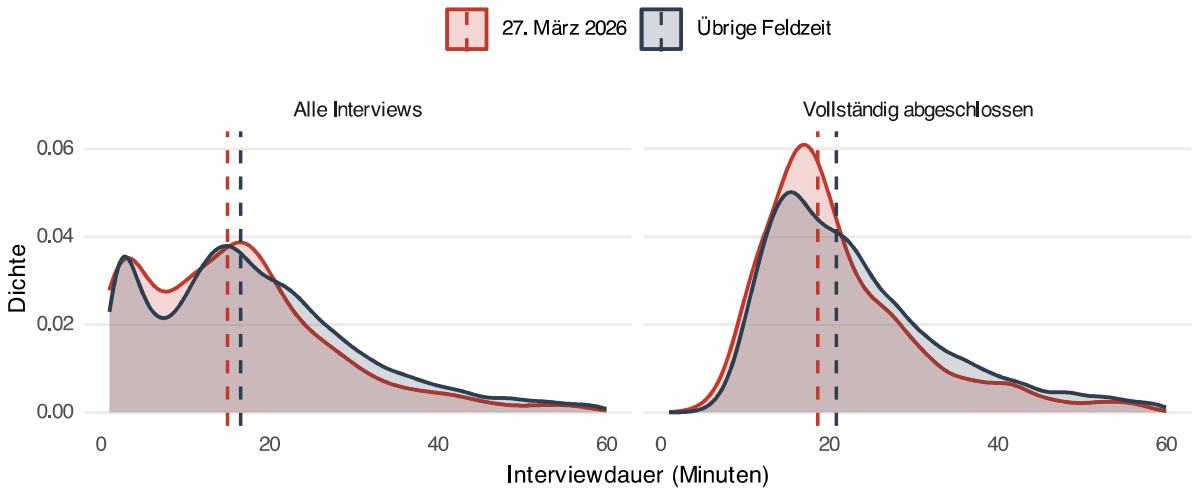

Abbildung 2: Verteilung der Interviewdauer am Peak-Tag und in der übrigen Feldzeit (Werte über 60 Minuten ausgeblendet). Gestrichelte Linien zeigen den Median je Gruppe.

4 Erstellung der Gewichtung

4.1 Mapping Verein auf Landesverband

Die Vereinsliste wird an den gefilterten Datensatz angespielt. Zwei Beobachtungen ohne validen Vereinsnamen gehen dabei verloren. Eine manuelle Korrektur sorgt dafür, dass der Verein Haus & Grund Berlin-Steglitz unter dem Dachverband Haus & Grund Berlin e.V. geführt wird, da die Mitgliederliste nur den Dachverband enthält. Ein Sanity-Check stellt sicher, dass alle Landesverbände aus dem Vereinsmapping auch in der Mitgliederliste vorhanden sind.

4.2 Berechnung der Post-Stratifikationsgewichte

Die Gewichte werden auf Ebene der Landesverbände berechnet. Grundlage ist die aktuelle Mitgliederliste. Für jeden Landesverband wird der Anteil an der Gesamtmitgliedschaft und der Anteil an der Stichprobe berechnet. Das Rohgewicht ist der Quotient beider Anteile; Landesverbände, die in der Stichprobe unterrepräsentiert sind, erhalten Gewichte größer als eins, überrepräsentierte Landesverbände kleiner als eins.

Es werden zwei Gewichtsvariablen erzeugt, die unterschiedliche Analysezwecke bedienen:

- **`gewicht`** (normalisiert): auf einen Mittelwert von eins renormalisiert, sodass gewichtete Fallzahlen im Mittel den ungewichteten entsprechen. Geeignet für Anteilsschätzungen und Mittelwertvergleiche auf Stichprobenebene.
- **`gewicht_exp`** (Expansion): skaliert die Stichprobe direkt auf die Gesamtpopulation, sodass die Summe der Gewichte der Gesamtmitgliederzahl N entspricht. Geeignet für Hochrechnungen auf die Grundgesamtheit.

Sei L die Menge der Haus & Grund Landesverbände, N_ℓ die Mitgliederzahl von Landesverband $\ell \in L$ laut Mitgliederliste und n_ℓ die Zahl der gültigen Fälle aus ℓ in der Stichprobe nach Filterung. Weiter sei

$$N = \sum_{\ell \in L} N_\ell, \qquad n = \sum_{\ell \in L} n_\ell$$

die Populations- bzw. Stichprobengröße über alle berücksichtigten Landesverbände. Der Populations- und der Stichprobenanteil eines Landesverbands sind

$$p_\ell = \frac{N_\ell}{N}, \qquad s_\ell = \frac{n_\ell}{n}.$$

Das Rohgewicht eines Falls i in Landesverband $\ell(i)$ ist der Quotient der Anteile,

$$\tilde{w}_i = \frac{p_{\ell(i)}}{s_{\ell(i)}} = \frac{N_{\ell(i)}/N}{n_{\ell(i)}/n}.$$

Daraus ergeben sich die beiden finalen Gewichte. Das normalisierte Gewicht skaliert auf Mittelwert eins,

$$w_i = \frac{\tilde{w}_i}{\bar{\tilde{w}}}, \qquad \bar{\tilde{w}} = \frac{1}{n}\sum_{j=1}^{n} \tilde{w}_j,$$

und erfüllt nach Konstruktion $\sum_{i=1}^{n} w_i = n$ sowie $\bar{w} = 1$. Das Expansionsgewicht skaliert direkt auf die Gesamtpopulation,

$$w_i^{\mathrm{exp}} = \frac{N_{\ell(i)}}{n_{\ell(i)}},$$

und erfüllt nach Konstruktion $\sum_{i=1}^{n} w_i^{\mathrm{exp}} = N$.

Für eine gewichtete Kennzahl — etwa den Anteil einer Kategorie A oder den Mittelwert einer numerischen Variable x — ergibt sich mit beiden Gewichten dasselbe Ergebnis,

$$\hat{p}_A = \frac{\sum_{i=1}^{n} w_i \, \mathbf{1}\{i \in A\}}{\sum_{i=1}^{n} w_i}, \qquad \hat{\bar{x}} = \frac{\sum_{i=1}^{n} w_i \, x_i}{\sum_{i=1}^{n} w_i},$$

da die Normalisierung sich in Zähler und Nenner hebt.

4.3 Kontrolle

Nach der Gewichtsberechnung werden drei automatische Checks durchgeführt: die Summe von `gewicht` muss der Fallzahl n entsprechen, die Summe von `gewicht_exp` muss der Gesamtpopulation N entsprechen, und der gewichtete Stichprobenanteil je Landesverband muss mit dem Populationsanteil übereinstimmen. Alle drei Checks bestehen.

Abbildung 3 zeigt den Vergleich der Stichproben- und Populationsanteile je Landesverband. Verbände, bei denen beide Punkte nah beieinander liegen, erhalten ein Gewicht nahe eins; deutliche Abstände

signalisieren stärkere Korrekturen. Die Gewichtung korrigiert strukturelle Verzerrungen durch unterschiedliche Beteiligungsquoten der Landesverbände, sodass Anteile und Mittelwerte die Relationen der Mitgliederpopulation widerspiegeln.

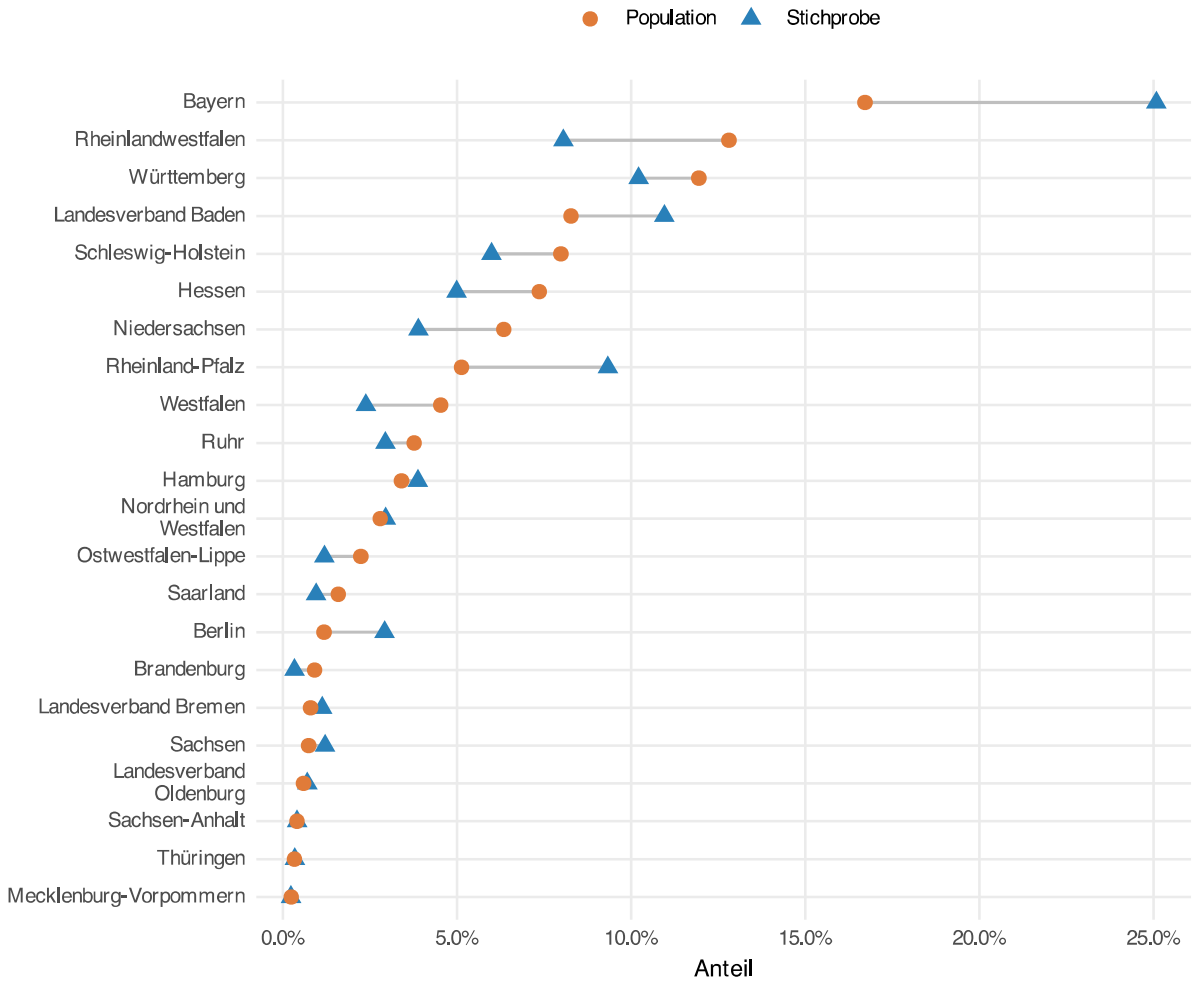

Abbildung 3: Stichproben- und Populationsanteile je Haus & Grund Landesverband. Verbände, bei denen die Punkte auseinanderfallen, erhalten stärkere Gewichtungskorrekturen.

5 Auswertung und Export

5.1 Illustration: Normalisiertes vs. Expansionsgewicht

Die beiden Gewichtsvariablen liefern für Anteilsschätzungen stets dasselbe Ergebnis — sie unterscheiden sich ausschließlich in der Skala der Absolutzahlen. Abbildung 4 zeigt dies am Beispiel der Wohnbestandsgröße: Die ungewichteten Anteile weichen von den gewichteten ab, weil Landesverbände mit hoher Beteiligung in der Stichprobe überrepräsentiert sind; `gewicht` und `gewicht_exp` korrigieren dies identisch. Abbildung 5 macht den Skalenunterschied sichtbar — `gewicht` summiert auf die Stichprobengröße n, `gewicht_exp` auf die geschätzte Grundgesamtheit N.

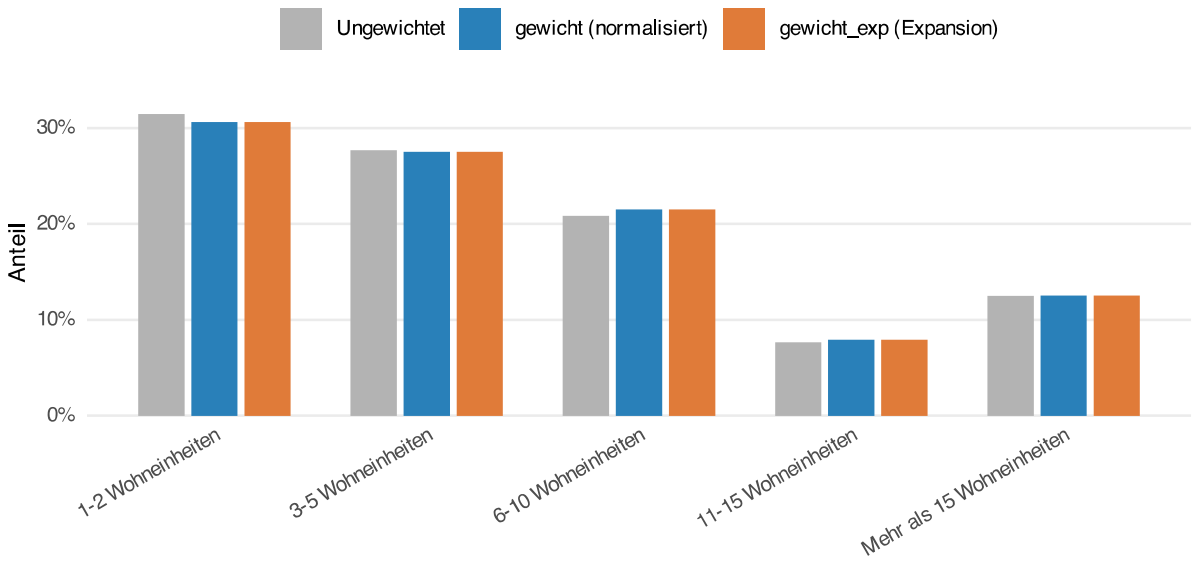

Abbildung 4: Verteilung der Wohnbestandsgröße: ungewichtet, normalisiert gewichtet und expansionsgewichtet im Vergleich. Für Anteilsschätzungen sind beide Gewichte äquivalent.

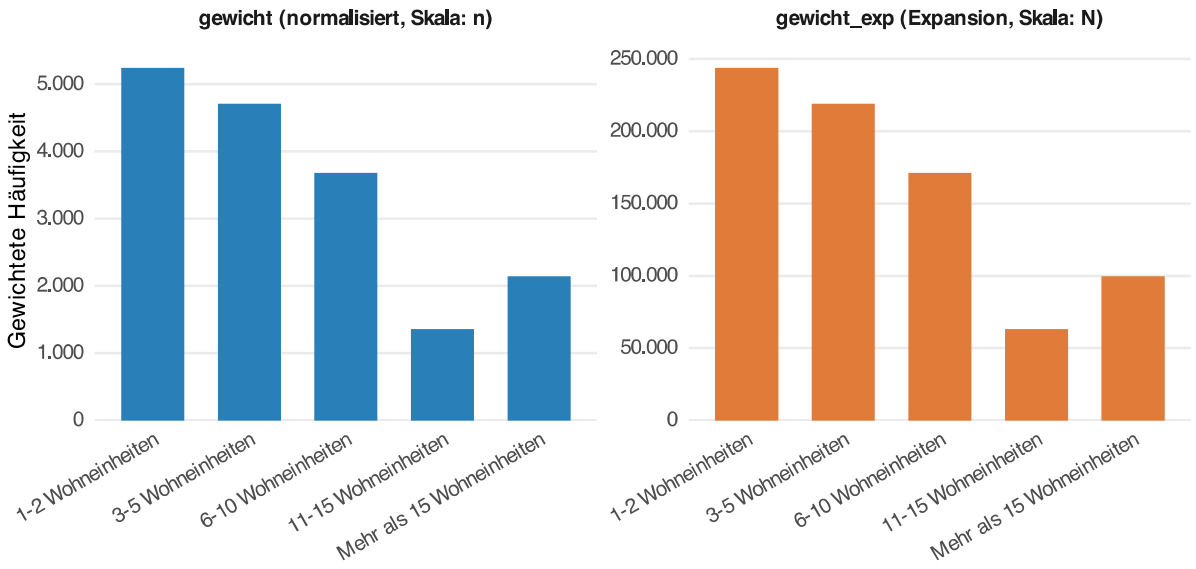

Abbildung 5: Absolutzahlen je Gewichtsvariante. `gewicht` summiert auf n (Stichprobengröße), `gewicht_exp` hochgerechnet auf N (Gesamtmitgliedschaft). Die relativen Verhältnisse sind in beiden Fällen identisch.

Die gewichteten Auswertungen werden direkt in thematisch gegliederte Excel-Arbeitsmappen mit den Reitern Soziodemografie, Bestand, Miete, Investitionen, Konflikte, Gebäude und Zusatzfragen exportiert. Zusätzlich zur Gesamtauswertung wird für jeden Landesverband eine eigene Arbeitsmappe erzeugt. Die Auswertungen verwenden standardmäßig das normalisierte Gewicht `gewicht`; für Hochrechnungen steht `gewicht_exp` zur Verfügung.

Parallel wird der aufbereitete Datensatz als Stata-Datei gespeichert, damit er für weitergehende Analysen außerhalb von R zur Verfügung steht. Offene Textfelder werden vor dem Export entfernt.