DAV/DGVFM
Jahrestagung
2025

Nora Valiente Bauer, Munich Re
Sarah Hogekamp, Gen Re

Der SDC Anonymisierungsprozess im Kontext aktuarieller Anwendungen

Bonn, 30.04.2025

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Agenda

Disclaimer: Bei den im Anschauungsfall verwendeten Anonymitätsniveaus und -verfahren handelt es sich nicht um eine allgemeine, juristische Bewertung sondern um eine für den fiktiven Anwendungsfall beispielhaft getroffene Entscheidung zur Veranschaulichung eines Anonymisierungsprozesses.

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Mitglieder

- Dariush Sadeghi-Yam
- Karsten de Braaf
- Eva Odenkirchen
- Dr. Christoph Falkenau
- Nora Valiente Bauer
- Sarah Hogekamp
- Marta Gamnitzer
- Erwin Hedtke (Ausgetreten)

Ziele

- ✓ AG Unterlagen
- ✓ Notebooks in R, Python
- ✓ Ergebnisbericht
- ❑ Webinar
- ❑ Veröffentlichung Artikel
- ❑ Bestandteil der Ausbildung

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

„*Information is the oil of the 21st century, and analytics is the combustion engine*.“

– Peter Sondergaard, Senior Vice President and Global Head of Research at Gartner, Inc.

Aktuelle Situation für Unternehmen

- Die Verwendung personenbezogener Daten wird in der EU durch die Datenschutz-Grundverordnung (DSGVO) geregelt.
- Jedes Unternehmen und jede Institution, die personenbezogene Daten in der EU verarbeitet, muss die DSGVO einhalten, auch wenn sie sich an einem anderen Ort befindet.
- Die Erwägungsgrund 26 besagt, dass anonymisierte Daten von dessen Vorschriften der DSGVO ausgenommen sind: „Die Grundsätze des Datenschutzes sollten daher nicht für anonyme Informationen gelten,…“

Ziele

- Erzeugung von anonymen Daten gemäß der DSGVO
- Dabei sollten so wenig Informationen wie möglich während des Datenanonymisierungsprozess verloren gehen

DSGVO und Erwägungsgrund 26

- Eine Definition von Anonymisierung ist nicht in der DSGVO verankert
- Erwägungsgrund 26: Anonyme Daten sind Informationen, die sich *„Die Grundsätze des Datenschutzes sollten daher nicht für anonyme Informationen gelten, dh. für Informationen, die sich nicht auf eine identifizierte oder identifizierbare natürliche Person beziehen, oder personenbezogene Daten, die in einer Weise anonymisiert worden sind, dass die betroffene Person nicht oder nicht mehr identifiziert werden kann.“*

Data Protection Working Party

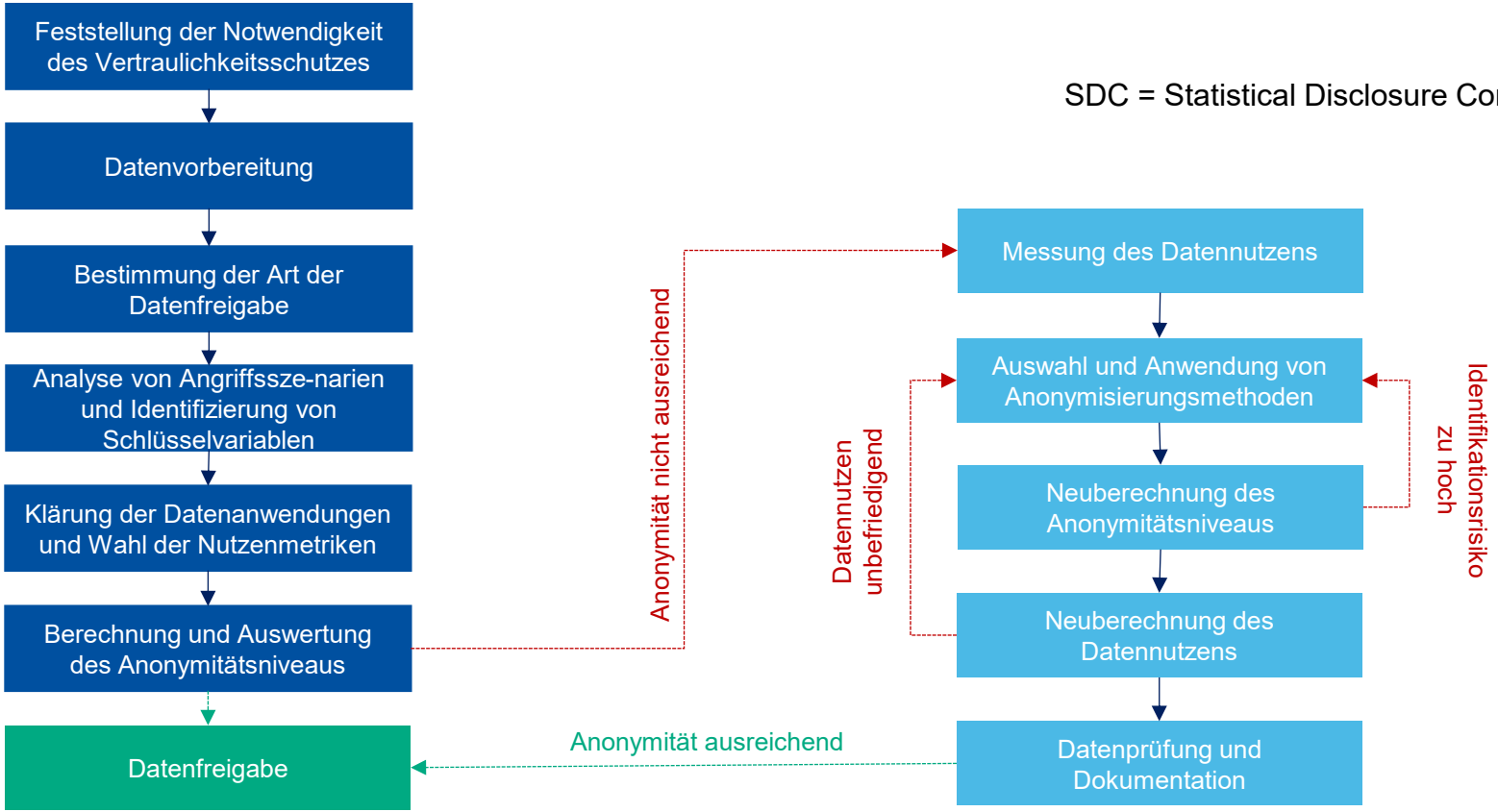
Es werden Risiken definiert, um die Robustheit einer Anonymisierungstechnik festzustellen:

- ***Singeling out*:** Möglichkeit, einige oder alle Datensätze zu isolieren, die eine Person im Datensatz identifizieren.
- ***Linkability*:** Fähigkeit, mindestens zwei Datensätze zu verknüpfen, die dieselbe betroffene Person oder eine Gruppe von betroffenen Personen betreffen (entweder in derselben Datenbank oder in zwei verschiedenen Datenbanken)
- ***Inference*:** Möglichkeit, den Wert eines Attributes mit großer Wahrscheinlichkeit aus den Werten einer Reihe anderer Attribute abzuleiten.

Es wird nicht spezifiziert wie der Datenanonymisierungsprozess aussehen soll und mittels welcher Methoden die Risiken minimiert werden können.

SDC-Anonymisierungsprozess

SDC = Statistical Disclosure Control



DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Aufgabe

Ein Aktuar hat einen Datensatz vor sich den er anonymisieren möchte/muss zur Weitergabe für eine externe Firma. Der Datensatz bestehet aus 50 Personen aus dem Jahre 2012 in den USA.

Daten

Kategoriale Variablen	
Vorname	48 Kategorien
Nachname	49 Kategorien
Beruf	35 Kategorien
Postleitzahl	50 Kategorien
Geschlecht	2 Kategorien
Geburtstag	50 Kategorien
ICD-Code	20 Kategorien

Metrische Variablen	
Einwohner	Einwohnerzahl
PurchPowCapita	Kaufkraft pro Kopf
PurchPowHousehold	Kaufkraft pro Haushalt
Gehalt	Pro Person

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Schritt 1 — Notwendigkeit des Vertraulichkeitsschutzes

- Bestätige sinnvolle rechtliche Rahmenbedingungen, die für den Datensatz und seine gewünschte Anwendung zutreffen.
- Teile die Variablen ein:
  - ***Direkte Identifikatoren****:* Attribute, die eine eindeutige Identifikation von Personen ermöglichen, wie z. B. Name oder Ausweisnummer.
  - ***Quasi Identifikatoren***: Attribut oder Menge von Attributen, das für sich allein keine Identifizierung ermöglicht, aber durch Hinzunahme anderer Daten zur Identifizierung führen kann.
  - ***Sensible Attribute***: Datenkategorien oder Informationen, die besonders schützenswert sind, weil sie potenziell schädlich oder sensibel für die betroffene Person sein können, wenn sie kompromittiert oder unangemessen verwendet werden

In unserem Beispiel:

- Es handelt sich um personenbezogene Daten die zu anonymisieren sind
- ***Direkte Identifikatoren:*** Vorname, Nachname,
- ***Quasi Identifikatoren*****:** Geschlecht, Geburtstag, Postleitzahl, Beruf, Einwohner, PurchPowCapita, PurchPowHousehold
- ***Sensible Attribute*****:** Einkommen, ICD-Code

DAV Deutsche Aktuarvereinigung e.V.

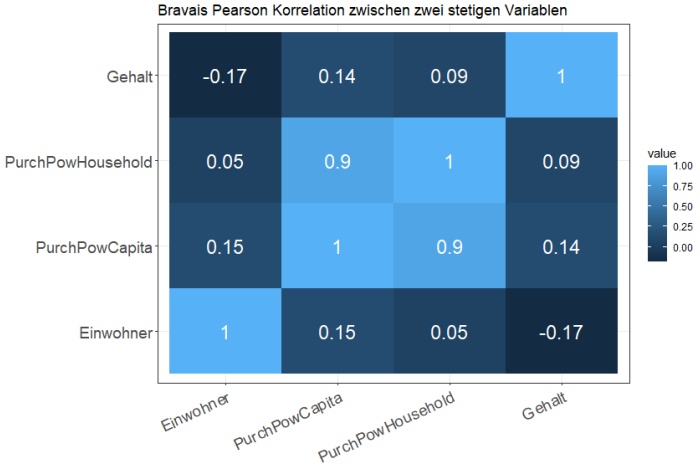
DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

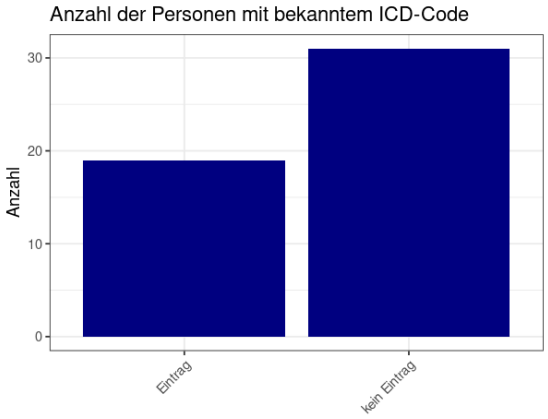
Schritt 2 — Datenvorbereitung

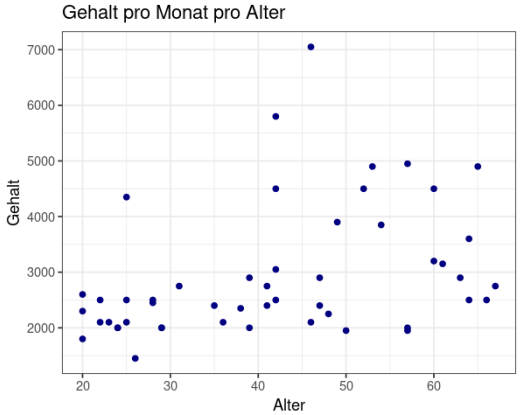
- Daten in gewünschte Datenstruktur bringen, bereinigen und erste deskriptive Analysen durchführen.
- Entfernung von direkten Identifikatoren
- Vorgehensweisen und Best Practices unterscheiden sich nicht wesentlich von der Datenvorbereitung bei klassischen aktuariellen- oder Data Science- Analysen. Besonders zu beachten sind:
  - Konsolidierung von Variablen mit ähnlichen Informationen
  - Analyse von Attributs Abhängigkeiten

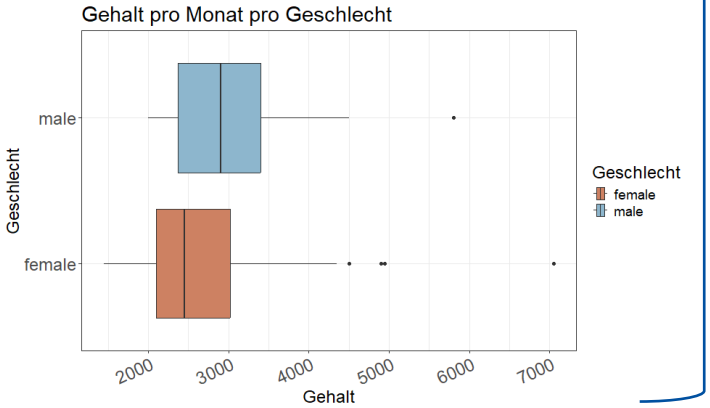
In unserem Beispiel:

- Entfernung von: Vorname, Nachname
- Korrelationsplots
- Untersuche sensibles Attribut genauer









Erkenntnisse

- Nach Anonymisierung behalte Struktur der Korrelationen und Zusammenhänge
- Besonders sensibles Attribut sollte weiterhin interpretierbar sein

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Schritt 3

Bestimmung der Art der Datenfreigabe

- Anonymisierungsstrategie hängt von der Art der Datenfreigabe ab (Unternehmensintern vs öffentliche Freigabe)

⇒ Art der Freigabe beeinflusst das Re-identifizierungsrisiko

- Eruiere Möglichkeiten um das Risiko zu minimieren:
  - Begrenze den zugriff auf Teams
  - Zugriff über Data Center des eigenen Unternehmens
  - Begrenze das Teilen mit externen Firmen oder Organisationen auf berechtigte Personen

Firma

Richtet Data Center innerhalb des eigenen Unternehmens ein

Externe Firma

Muss zustimmen die Daten auf ein Data Center zu übertragen

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Schritt 4

Analyse von Angriffsszenarien und Bestimmung von Schlüsselvariablen

- Nutze eine Übersicht von möglichen Angriffsszenarien, berücksichtige besonders
  - den Einsatz von externen Datenquellen
  - Mögliche Dritter sowie die damit verbundenen Kosten und den erforderlichen Zeitaufwand
  - Technologischen Stand und Entwicklung um langfristigen Schutz der Daten sicher zustellen
- Unterscheidung zwischen zwei Arten von Informationen zur Messung von Datenschutzrisiken:
  - ***Allgemeine Informationen***: verleihen den anonymisierten Daten nutzen
  - ***Spezifische Informationen***: ermöglichen es dem Angreifer die Privatsphäre einzelner Personen zu verletzen
- Verschiedene Angriffsszenarien:
  - ***Prosecuter Attack:***
    - Angreifer hat Hintergrund Informationen über eine Person aus dem Datensatz und nutzt diese, um im offen gelegten Daten und zusätzliche Informationen zu erfahren
  - ***Journalist Attack:***
    - Angriff der darauf abzielt, durch den Zugriff auf andere Informationsquellen über eine oder mehrere Personen, diese zu re-identifizieren. Angreifer weiß jedoch nicht sicher ob die bestimmte Person im Datensatz enthalten ist
  - ***Marketer Attack:***
    - Angreifer zielt darauf ab so viele Personen wie möglich aus dem Datensatz zu re-identifizieren, auch wenn dies bedeutet dass einige von ihnen fälschicherweise identifiziert werden
- NIST Adversial Machine Learning beschreibt konkret technische Durchführungen von Angriff Szenarien

DAV Deutsche Aktuarvereinigung e.V.

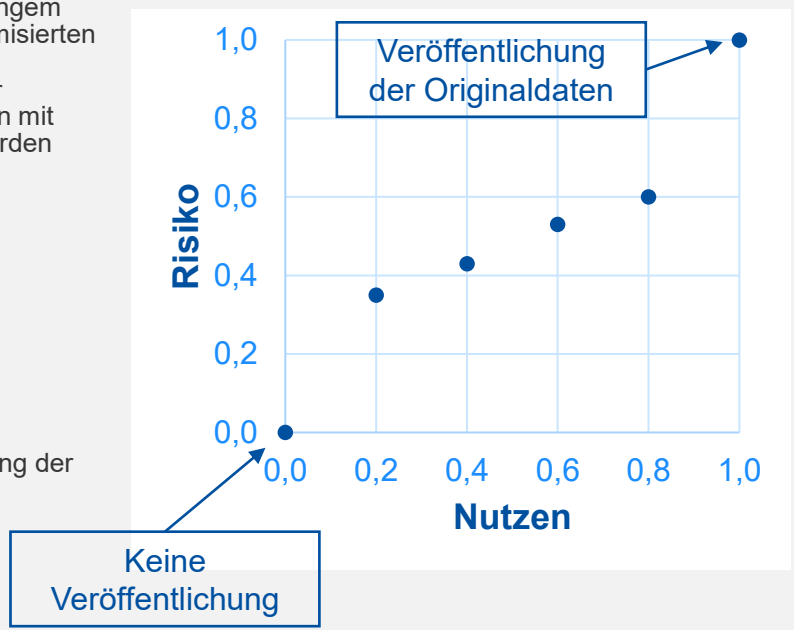
DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Schritt 5

Klärung der Datenanwendung und Wahl der Nutzen Metriken

- Anonymisierungsstrategie ist ein Kompromiss aus geringem Re-Identifizierungsrisiko und hohen Nutzen der anonymisierten Daten
- Idealerweise erhält man vom Nutzer eine Definition der gewünschten Datengranularität pro Variable zusammen mit einer Minimum Granularität, die nicht unterschritten werden sollte.

- Maße für den Datennutzen
  - ***Allgemeine Maße***:
    z.B. Anzahl fehlender oder geänderter Werte
  - ***Statistische Maße***:
    z.B. Mittelwert, Korrelation
  - ***Visualisierung***:
    z.B. Histogramme, Box-Plots
  - ***Regressionsmodelle***:
    Bei bekanntem Endnutzen kann die Beibehaltung der Struktur untersucht werden.



DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Schritt 6 — Berechnung und Auswertung des Anonymitätsniveaus

- Zielwert des Anonymität Niveaus sollte im Verhältnis zu der Art der Datenfreigabe und den identifizierten Angriffsszenarien stehen
- Berechne das Risiko einer Re-Identifizierung: K-Anonymität, L-Diversität, T-Ähnlichkeit

K-Anonymität:
Es haben mindestens k-Personen die gleiche Kombination von Quasi-Identifikatoren

Geschlecht	Beruf	Einkommen
Männlich	Handwerker	2500,00
Männlich	Handwerker	2500,00
Männlich	Handwerker	2500,00

3-Anonym: Die Äquivalenzklasse „männlich, Handwerker“ hat 3 Einträge. Sie ist somit 3 anonym. ABER: wenn jemand weiß, dass ein Mann ein Handwerker ist, ist er sicher, dass er 2500 Dollar verdient.

L-Diversität:
Jede Gruppe von K-Anonymität hat mindestens L verschiedene Werte für die sensible Information

Geschlecht	Beruf	Einkommen
Männlich	Handwerker	2300,00
Männlich	Handwerker	3500,00
Männlich	Handwerker	2300,00

2- Divers: für die gleiche quasi-identifizierende Gruppe gibt es verschiedene Einkommen, so dass man nicht direkt auf ein Einkommen schließen kann.

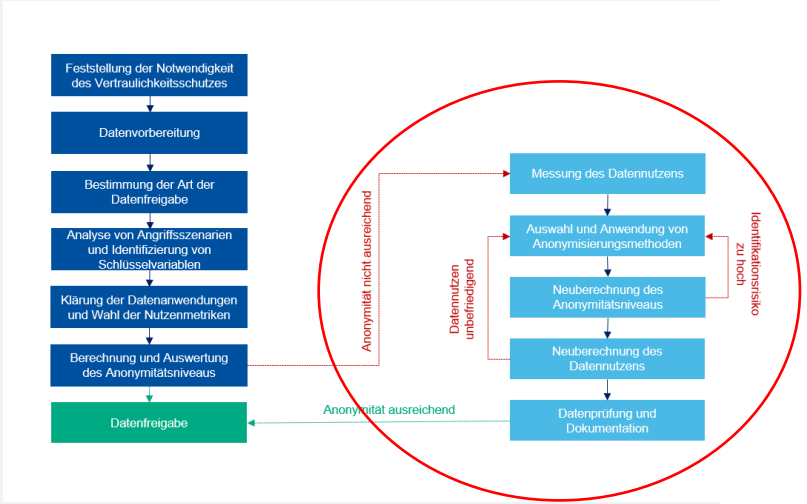
DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Schritt 6 — Berechnung und Auswertung des Anonymitätsniveaus

- Zielwert: ***2-Diversität***

Quasi-Identifikatoren	
Beruf	35 Kategorien
Postleitzahl	50 Kategorien
Geschlecht	2 Kategorien
Geburtstag	50 Kategorien



- Ist-Wert:
  - Postleitzahl und Geburtstag haben jeweils 50 Kategorien bei 50 Datensätzen
  - Bei einer unveränderten Veröffentlichung wäre der Datensatz ***1-Anonym*** und ***1-Divers***

➡ Die Anonymität ist nicht ausreichend. Die Schritte 7 bis 11 müssen durchgeführt werden zur Erhöhung des Anonymitätsniveaus

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

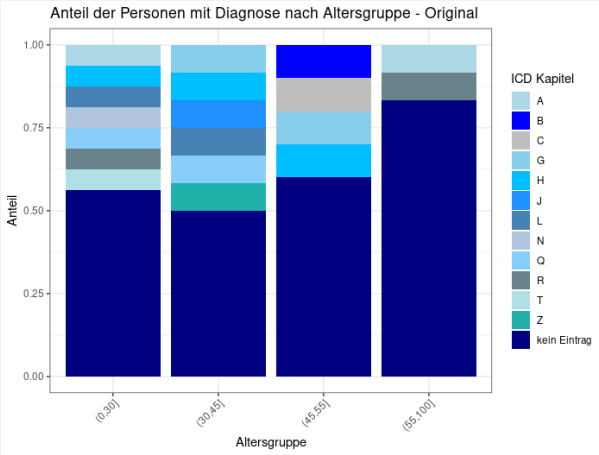
Schritt 7 — Messung des Datennutzens

- ***Allgemeine Maße:***
  - In den Originaldaten fehlen 0 Werte.
  - Es wurden 0 Werte geändert.
- ***Statistische Maße***:
  - Mittelwert

Einwohner	PP Capita	PP Household
10.882	22.142	46.388

  - Darstellung der Bravais Pearson Korrelation zwischen allen stetigen Variablen in der Datenvorbereitung.
- ***Visualisierung***:
  - Darstellung von Zusammenhängen verschiedener Variablen als Histogramme oder Box-Plots in der Datenvorbereitung.

- ***Weitere Optionen***: z.B. könnte man für die Versicherungswirtschaft Kopfschäden, Profile, Tafeln und Diagnosen nach Altersgruppe vergleichen



DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Schritt 8 — Auswahl und Anwendung von Anonymisierungsmethoden

Auswahl

1. Bestimmung der Quasi-Identifikatoren zur Anonymisierung: PLZ, Geburtstag, Beruf, *Einwohner, PP Capita, PP Household (Rückschlüsse auf PLZ möglich?)*
2. Veränderung dieser Attribute zur Minimierung des Identifikationsrisikos

Nicht-Pertubativ: Unterdrückung oder Ersatz bestimmter Werte durch ein spezielles Zeichen (z.B. Asterix *)

- ***Generalisierung***: Veränderung von Werten zur Reduktion der Genauigkeit.
- ***Suppression***: Unterdrückung aller Instanzen eines seltenen Wertes.

Pertubativ: Ersetzung originaler Werte durch *synthetische Daten*

- ***Random Noise***: Addition oder Multiplikation von Werten zu numerischen Merkmalen

	Generalisierung	Suppression	Random Noise
PLZ (5-Steller)	1- bis 4-Steller	Nach Einwohnerzahl	-
Geburtstag	Altersgruppen	Hohe/niedrige Alter	+/- n Tage
Beruf	Blue-/White-Collar	Seltene Berufe	-
Stetige Variablen	Runden	Ausreißer entfernen	+/- n

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Schritt 8

Auswahl und Anwendung von Anonymisierungsmethoden

Anwendung

- Geburtstag → Generalisierung → Altersgruppe („<50“, „50+“)
- Beruf → Generalisierung → Berufskategorie (Blue-/White Collar)
- Geschlecht → keine Anonymisierung
- PLZ → Suppression → Spalte wurde gelöscht. Selbst auf 1-Steller-Ebene bleiben bei den 50 Datensätzen 9 Level übrig.

Anony-misierte Daten

Kategoriale Variablen

~~Vorname~~	~~48 Kategorien~~	
~~Nachname~~	~~49 Kategorien~~	
Beruf	35 Kategorien	2 Kategorien
~~Postleitzahl~~	~~50 Kategorien~~	
Geschlecht	2 Kategorien	
Geburtstag	50 Kategorien	2 Kategorien
ICD-Code	20 Kategorien	

Metrische Variablen

Einwohner	Einwohnerzahl	+/- 1%
PP Capita	Kaufkraft pro Kopf	+/- 1%
PP Household	Kaufkraft pro Haushalt	+/- 1%
Gehalt	Pro Person	

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Schritt 9

Neuberechnung des Anonymitätsniveaus

Geschlecht	Alters-gruppe	Berufs-kategorie	Anzahl Datensätze	Levels (Gehalt)	Levels (ICD10Code)
weiblich	<50	blue collar	5	5	2
weiblich	<50	white collar	18	11	9
weiblich	50+	blue collar	6	4	2
weiblich	50+	white collar	6	5	3
männlich	<50	blue collar	5	4	4
männlich	<50	white collar	6	5	4
männlich	50+	blue collar	4	4	2

- Die Anzahl der Datenpunkte in jeder Äquivalenzklasse ist mindestens 4. Der Datensatz ist ***4-Anonym.***
- Bei der Betrachtung der sensiblen Attribute finden sich in jeder Äquivalenzklasse mindestens 4 unterschiedliche Gehälter 2 unterschiedliche ICD-Codes. Der Datensatz ist ***2-Divers.***

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Schritt 10 — Neuberechnung des Datennutzens

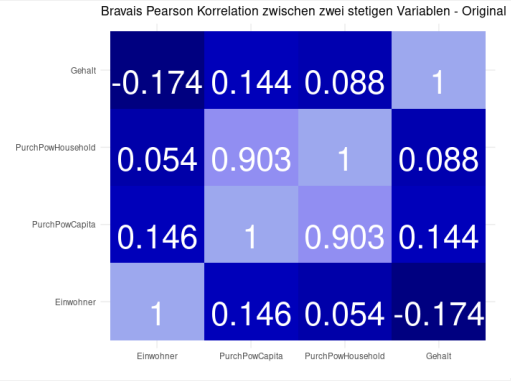
Allgemeine Maße

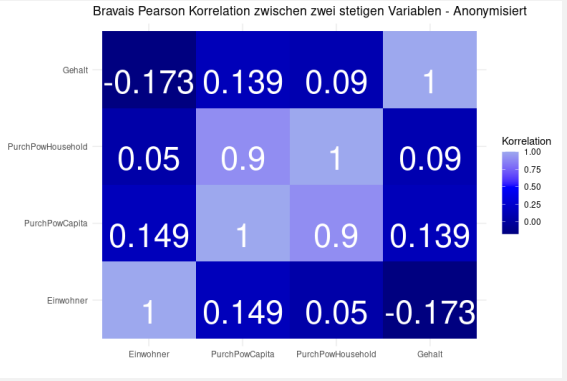
- Datensätze mit fehlenden Werten: 50 (100%)
- Datensätze mit geänderten Werten: 50 (100%)

➢ In jedem Datensatz wurden Werte gelöscht und geändert

Statistische Maße

Mittelwert	Original	Anony-misiert
Einwohner	10.882	10.875
PP Capita	22.142	22.149
PP Household	46.388	46.429





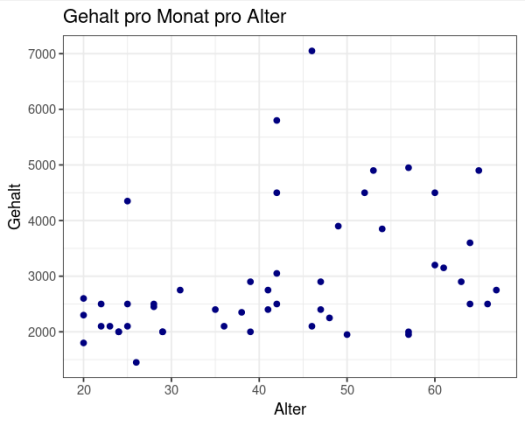
DAV Deutsche Aktuarvereinigung e.V.

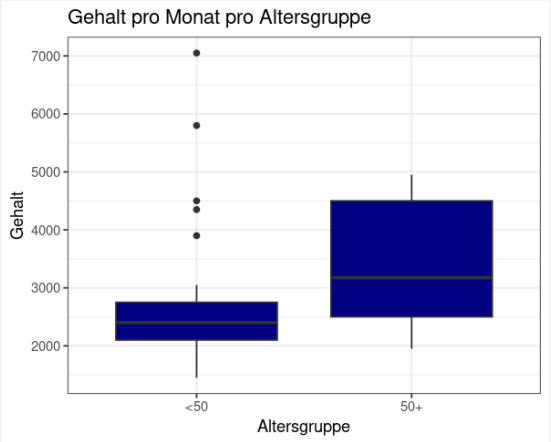
DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Schritt 10 Neuberechnung des Datennutzens

Visualisierung

- Die Gehaltsverteilung nach Geschlecht ist unverändert
- Die Gehaltsverteilung nach Alter wurde generalisiert:





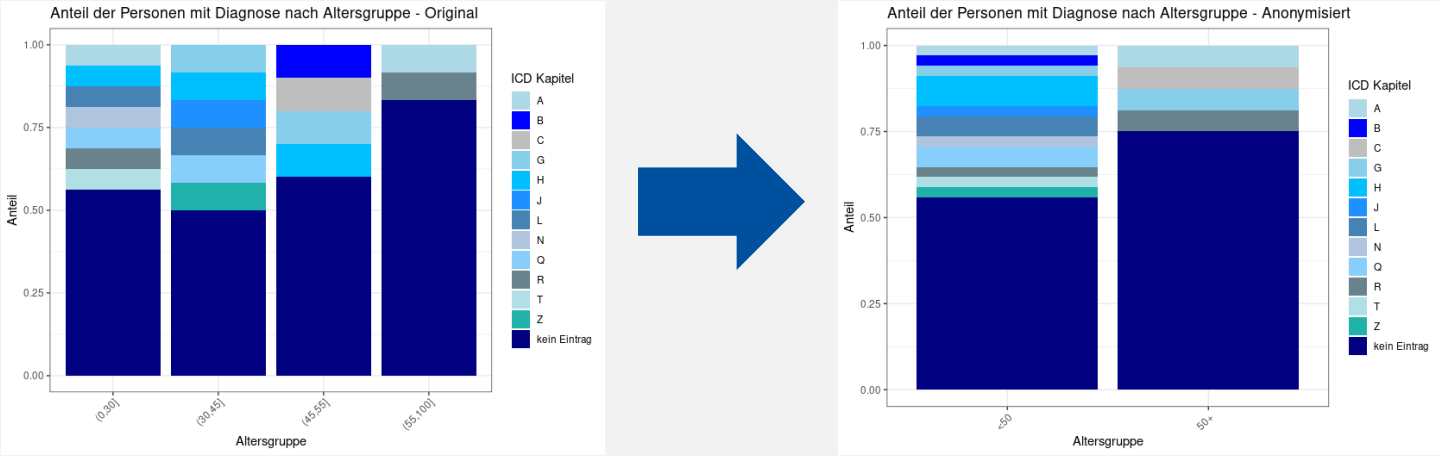
DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Schritt 10 — Neuberechnung des Datennutzens

Versicherungswirtschaft

- Kopfschäden, Profile, Tafeln
- Diagnosen nach Altersgruppe



DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Schritt 11 **Datenprüfung und Dokumentation**

Datenprüfung

- Datenstruktur überprüfen und Konsistenz der Variablenbeziehungen sicherstellen
- Verteilung der Variablen auf Anomalien untersuchen (z.B. negative Gehaltswerte)
- ggfs. Abweichungen minimieren zwischen früheren und aktuellen Datensatzversionen

Dokumentation

Interne Dokumentation

- Anonymisierungsprozess inkl. Methoden und zugehörige Parameter dokumentieren
- Risikometriken vor und nach Anonymisierung dokumentieren
- Erlaubt spätere Reproduktion + Beleg bei Überprüfung

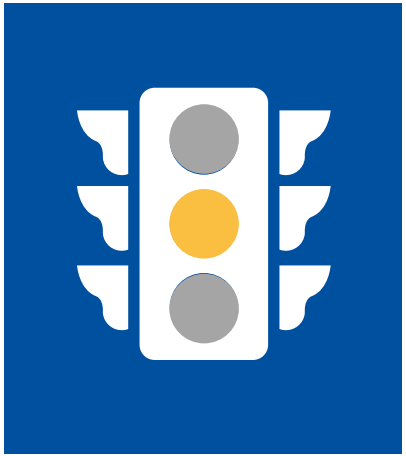
Externe Dokumentation

- Informiert Nutzer über Anonymisierung
- Beschreibt Einschränkungen in Nutzung
- Grobe Beschreibung der Methoden
- Sicherstellung aktueller Metadaten des Datensatzes (Variablenbeschreibung, Labels)

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Schritt 12 Datenfreigabe



Als letzter Schritt erfolgt die tatsächliche Datenfreigabe, **im Einklang mit der gewählten Art der Freigabe in Schritt 3.**

- Die Versicherungswirtschaft arbeitet an vielen Stellen mit sensiblen, personenbezogenen Daten.
- Der **SDC-Anonymisierungsprozess** ist ein Toolkit zur Aufbereitung von Datensätzen. Ziel ist es, Daten nutzbar zu machen, Informationen nicht zu verlieren und gleichzeitig die Gefahr einer Offenlegung des Personenbezugs zu reduzieren.
- Die aktuelle **Gesetzgebung**, die **Art der Veröffentlichung** und die geplante **Datenverwendung** müssen bei der Durchführung des Anonymisierungsprozesses berücksichtigt werden.
- Weiterführende Informationen zur **Anonymisierung**: AG Unterlagen, Notebooks, Ergebnisbericht

- Bei der Anonymisierung werden die Datenelemente eines originalen Datensatzes angepasst. Es ist jedoch möglich, dass das Offenlegungsrisiko dadurch nicht hinreichend reduziert wird.
- Alternative: **Synthetische Daten**
  - Idee: Verwendung der statistischen Eigenschaften eines Datensatzes zur Kalibrierung eines Modelles, welches neue „synthetische“ Daten erzeugt.
  - Erster Einblick im Ergebnisbericht der AG Anonymisierung
  - ➢ *Start einer neuen AG zur Synthetisierung von Daten*

DAV/DGVFM
Jahrestagung
2025

Vielen Dank für Ihre Aufmerksamkeit

Nora Valiente Bauer
nvalientebauer@munichre.com
Munich Re

Sarah Hogekamp
sarah.hogekamp@genre.com
Gen Re