KOMPETENZZENTRUM FÜR MATHEMATISCHE MODELLIERUNG IN MINT-PROJEKTEN IN DER SCHULE

KOMMS Reports Nr. 13 (2020)

Reports zur Mathematischen Modellierung
in MINT-Projekten in der Schule

KOMMS
Kompetenzzentrum für Mathematische Modellierung in MINT-Projekten in der Schule

Mathematische Modellierungswoche Dezember 2020 Themenschwerpunkt Digitaler Wandel

Patrick Capraro, Lynn Knippertz, Lisa Schneider

EUROPÄISCHER SOZIALFONDS
Mehr Chancen für Rheinland-Pfalz

EUROPÄISCHE UNION
Europäischer Sozialfonds

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

1 Die Felix-Klein-Modellierungswoche

Seit 1993 veranstaltet der Fachbereich Mathematik der TU Kaiserslautern jährlich die mathematischen Modellierungswochen. Die Veranstaltung erwuchs parallel zu der steigenden Relevanz angewandter mathematischer Forschungsgebiete, wie der Technomathematik und der Wirtschaftsmathematik. Sie soll dazu dienen, Schülerinnen und Schülern die Bedeutung mathematischer Arbeitsweisen in der heutigen Berufswelt, insbesondere in Industrie und Wirtschaft, begreifbar zu machen. Darüberhinaus bietet die Modellierungswoche den teilnehmenden Lehrkräften einen Einblick in die Projektarbeit mit offenen Fragestellungen im Rahmen der mathematischen Modellierung.

1.1 Partner und Finanzierung

Seit 2014 wird die Modellierungswoche vom Kompetenzzentrum für Mathematische Modellierung in MINT-Projekten in der Schule (KOMMS) organisiert, welches im selben Jahr gegründet wurde. Seit 2015 ist die Veranstaltung ein Teil des Projekts Schulentwicklung für mathematische Modellierung in MINT-Fächern (SchuMaMoMINT) und wird durch den Europäischen Sozialfonds (ESF), das Fraunhoferinstitut für Techno- und Wirtschaftsmathematik (ITWM), die TU Kaiserslautern sowie den Fachbereich Mathematik der TU Kaiserslautern finanziert. In diesem Zusammenhang gab es auch eine inhaltliche Weiterentwicklung der Modellierungswoche, innerhalb dessen der Aspekt einer Lehrkräftefortbildung stärker in den Vordergrund rückte.

Bereits zuvor war das Fraunhofer ITWM ein wichtiger Kooperationspartner und war an der Gestaltung der mathematischen Modellierungswoche wesentlich beteiligt. Im Jahr 2008 gründeten der Fachbereich Mathematik der TUK und das ITWM den Verein *Felix-Klein-Zentrum für Mathematik*, um die gemeinsamen Aktivitäten besser steuern zu können. Die Modellierungswoche wurde daraufhin in Felix-Klein-Modellierungswoche umbenannt.

1.2 Zielgruppe und Intention

Die Veranstaltung richtet sich an Schülerinnen und Schüler der gymnasialen Oberstufe in Rheinland-Pfalz, die ein außerordentliches Interesse an Mathematik und MINT-interdisziplinärem Arbeiten haben. Begleitet werden diese von Lehrkräften oder Referendaren ihrer Schule, welche die Projektarbeit beobachten und die Projektleiter bei der fachlichen Betreuung der Gruppen unterstützen.

Ziel ist es, sowohl bei den Schülerinnen und Schülern, als auch bei den Lehrkräften ein Bewusstsein dafür zu schaffen, wie das Arbeiten mit offenen Fragestellungen gelingen kann. Das KOMMS möchte dabei Wege aufzeigen, wie der Unterricht in den MINT-Fächern durch Konzepte wie *Forschendes Lernen* und *Eigenverantwortliche Projektarbeit* weiterentwickelt und bereichert werden kann.

1.3 Format und Durchführung

Die Modellierungswoche findet üblicherweise in Jugendherbergen in Rheinland-Pfalz statt. Zu Beginn werden am Sonntag Abend die Teilnehmer auf die verschiedenen Projektthemen aufgeteilt. Von Montag Morgen bis Donnerstag Abend wird inhaltlich gearbeitet. Am Freitag endet die Veranstaltung mit der Präsentation der Ergebnisse.

Meist wird jede Projektgruppe von einer Mitarbeiterin oder einem Mitarbeiter der TU Kaiserslautern sowie einer Lehrkraft betreut. Dabei haben die Schülerinnen und Schüler jedoch viele Freiheiten, sich selbst zu organisieren und auch den Arbeitsrhythmus zu gestalten.

Eine wesentliche Komponente dabei ist die Arbeitsumgebung in der Jugendherberge oder einer vergleichbaren Tagungseinrichtung:

- Jede Projektgruppe arbeitet in ihrem eigenen Seminarraum
- In den Pausen trifft man sich mit den anderen Gruppen, so dass es einen inhaltlichen Austausch gibt
- Es gibt wenige Faktoren, die vom Arbeitsthema ablenken; selbst abends arbeiten viele Gruppen noch an ihren Projekten weiter

1.4 Digitale Durchführung in Pandemiezeiten

Da 2020 während der Coronapandemie Präsenzveranstaltungen fast unmöglich waren, musste die ursprünglich für den Sommer geplante Modellierungswoche zunächst in den Dezember verschoben und schließlich in ein digitales Format umgewandelt werden. Die Projektwoche wurde als Onlinekurs im Lernportal OLAT durchgeführt, das auch für die Hochschullehre an der TUK genutzt wird. Die Teilnehmerinnen und Teilnehmer nutzten Videokonferenzen, um sich zu organisieren. Mit kollaborativen Programmierumgebungen konnten wir von unterschiedlichen Bildschirmen aus gemeinsam Code entwickeln und die Ergebnisse im Team mitverfolgen.

Auch wenn sicherlich die soziale Komponente des üblichen Formats unter einer digitalen Durchführung leidet, so zeigte sich dennoch, dass selbst unter den widrigen Umständen der Coronazeit die Modellierungswoche zum erfolgreichen Abschluss kommen kann.

2 Themenschwerpunkt *Digitaler Wandel*

Der Titel passte aus zwei Gründen gut in die aktuelle Zeit. Zum einen mussten sich die Teilnehmerinnen und Teilnehmer pandemiebedingt an digitalen Arbeitsmethoden orientieren, um sich mit ihrem Team und mit den Betreuerinnen und Betreuern der TUK austauschen zu können. Zum anderen ging es inhaltlich um Themen, die für junge Menschen hochaktuell sind, nämlich unser Leben mit modernen Medien und vernetzen elektronischen Geräten, sowie die großen Datenmengen, die von diesen Technologien gesammelt werden. In zwei Projektteams gingen die Teilnehmerinnen und Teilnehmer zwei spannenden Fragen nach. Diese waren:

1. Mit Data Science zum perfekten Raumklima
2. Wie sicher sind unsere Daten in sozialen Netzwerken? Was Social-Media-Kanäle über uns wissen

Aus mathematischer Sicht handelt es sich bei beiden Fragestellungen um Themen aus dem Bereich Data Science. Es wurde mit großen Datenmengen gearbeitet (Datentabellen mit 30 000 bzw. 50 000 Zeilen), was bereits erste technologische Hürden bot. Nach dem Sichten des Datensatzes mussten Fehler bereinigt werden (z.B. offensichtliche Ausreißer in der Temperaturmessung auffinden und beseitigen) und schließlich konnten mit statistischen Methoden Modelle aufgestellt werden, um die gegebenen Fragen zu beantworten.

Die Projekte folgten damit einem klassischen Ablauf eines Data Science Projekts:

1. Daten sichten und auswählen (*welche Daten sollen verwendet werden*)
2. Daten bereinigen (*offensichtliche Fehler beheben*)
3. Daten transformieren (*Erzeugung neuer Daten durch Umrechnen oder Umstrukturieren der bestehenden Daten*)
4. Data Mining (*Suche nach Informationen in den transformierten Daten*)
5. Interpretation

3 Ergebnisse der Projektgruppen

3.1 Mit Data Science zum perfekten Raumklima

3.1.1 Problemstellung

In unserer digitalisierten Zeit ist es ein Leichtes, zu einem gegebenen Problem jede Menge Daten und Messwerte zu erzeugen. Elektronische Sensoren und automasierte Archivierungsprozesse erlauben es uns, Daten in großer Menge zu erheben und in digital zu speichern. Wenn es jedoch darum geht, aus den riesigen Zahlenkolonnen die wertvollen Informationen herauszuziehen, ist mathematischer Sachverstand gefragt.

Aus dem Wohnzimmer einer Wohnung wurden über einen Zeitraum von mehreren Monaten Daten über das Raumklima gesammelt. Dazu wurde sowohl die Raumtemperatur als auch die Luftfeuchtigkeit gemessen. Zusätzliche Temperatursensoren an der Balkontür sollen dazu dienen, die Zeiträume zu erkennen, in denen die Tür geöffnet war. Anhand der Daten wollen wir untersuchen, wie das perfekte Lüftungsverhalten in dieser Wohnung aussehen könnte. Dabei gibt es verschiedene Dinge zu beachten:

- Aufgrund der Dämmung kann es zu Feuchtigkeitsbildung an den Fenstern und Außenwänden kommen, weswegen die Luftfeuchtigkeit nicht zu hoch werden darf.
- Im Winter sollte darauf geachtet werden, dass nicht zu viel Wärme durch das Lüften verloren geht.
- Im Sommer sollte durch eine geschickte Auswahl der Uhrzeiten zum Lüften die Temperatur in der Wohnung so gering wie möglich gehalten werden.
- Der Vorschlag für ein gutes Lüftverhalten soll auch nach Kriterien der Praktikabilität gewählt werden.

3.2 Beschaffenheit der Daten

Von August bis Dezember wurden Temperatur- und Luftfeuchtigkeitsdaten aus einem Wohnzimmer gesammelt. Es erfolgten automatisierte Messungen in Intervallen von 5 Minuten. Erhoben wurde dabei

- ein Zeitstempel,
- die Raumtemperatur und die Luftfeuchtigkeit am Balkonfenster,
- die Temperatur an der Balkontür und
- die Temperatur in der Dichtung der Balkontür.

Abbildung 1: Eine automatisierte Temperaturerfassung (Foto: Patrick Capraro).

```
timestamp,temp0,temp1,temp2,humidity
2020-08-04 23:32:15.645260,26.0,26.0,25.875,54.0
2020-08-04 23:35:05.962821,26.0,26.0,25.875,53.0
2020-08-05 07:55:05.596606,23.0,22.5,22.25,56.0
2020-08-05 08:00:05.516615,26.0,22.625,22.312,48.0
2020-08-05 08:05:05.436624,30.0,22.75,22.375,41.0
2020-08-05 08:10:08.156649,30.0,22.812,22.437,41.0
2020-08-05 08:15:05.036637,30.0,22.937,22.5,41.0
2020-08-05 08:20:05.197323,29.0,23.0,22.625,41.0
2020-08-05 08:25:05.036760,29.0,23.0,22.687,42.0
2020-08-05 08:30:04.876604,31.0,23.312,22.812,38.0
```

Abbildung 2: Ein Ausschnitt aus den Rohdaten.

Die Daten wurden mit einem Raspberry Pi (siehe Abbildung 1) gesammelt und in einer csv-Datei abgelegt (siehe Abbildung 2). An 6 Tagen wurde über das Lüften Protokoll geführt.

3.2.1 Vorüberlegungen

Aus den Temperaturdaten, die an der Balkontür abgenommen werden, kann rekonstruiert werden, zu welchen Zeitpunkten gelüftet wurde (das entspricht der dritten Komponente des Data Science Prozesses: *Daten transformieren*). Von diesem Punkt an können die Schülerinnen und Schüler erfassen, welche Auswirkungen das Lüften hat. Hier kann vor allem ermittelt werden, wie stark die Temperatur oder die Luftfeuchtigkeit ab- bzw. zunehmen und ob die Stärke dieser Veränderung mit anderen Daten in Zusammenhang steht, z.B. mit der Uhrzeit, dem Datum oder der Temperaturdifferenz zwischen innen und außen.

Aus solchen Daten kann schließlich eine Kosten-Nutzen-Rechnung aufgebaut werden. Der Nutzen liegt in der Absenkung der Luftfeuchtigkeit und eventuell auch in einer Temperaturabsenkung im Sommer. Die Kosten sind ein Temperaturverlust im Winter bzw. ein Temperaturanstieg im Sommer. Eventuell kann auch die Tätigkeit des Lüftens an sich als Kostenfaktor betrachtet werden, da man ein Interesse daran hat, nicht ständig die Fenster und Türen öffnen und schließen zu müssen.

Das Thema knüpft in mehrerer Hinsicht an aktuelle Themen aus dem Alltagsleben der Schülerinnen und Schüler an:

1. Umwelt- und Klimaschutz: Allein aus ökologischer Sicht besteht ein großes Interesse, die Heizungswärme im Winter nicht unnötig durch das Fenster entweichen zu lassen.
2. Internet der Dinge: Die Datenmodelle können dazu genutzt werden, häusliche Aufgaben (teilweise) zu automatisieren. Ein Alarmsystem (in Form einer Warnlampe oder sogar einer Email direkt an das Smartphone) könnte den Nutzer warnen, dass er wieder lüften muss.
3. Sensible Daten erfassen: Aus den Sensormesswerten lassen sich verschiedene Informationen über das Verhalten der Bewohner rekonstruieren. Man kann ermitteln, zu welchen Uhrzeiten gekocht wird und möglicherweise auch an welchen Tagen und zu welchen Zeiten die Wohnung unbewohnt ist. In einer Zeit, in der viele Haushaltsgeräte mit dem Internet verbunden sind, können solche Sensormessdaten dazu verwendet werden, Nutzerprofile anzulegen.

Speziell der dritte Punkt lässt viel Spielraum zur Interpretation der Daten. Hier war es hilfreich, dass der Projektbetreuer in der Wohnung lebte, aus der die Daten stammten, da er die Vermutungen der Gruppe bestätigen oder widerlegen konnte. Bei genauerer Betrachtung der Daten waren viele Hinweise auf den Alltag in der Wohnung erkennbar, beispielsweise zu welchen Zeiten gekocht wurde (starker Anstieg der Luftfeuchtigkeit in einem charakteristischen Zeitfenster), aber auch die Urzeiten, zu denen morgens das Zimmer genutzt wurde (leichter Anstieg in der Luftfeuchtigkeit mit deutlichen Unterschieden zwischen Werktag und Wochentag).

Zum Arbeiten mit den Daten ist die Nutzung einer höheren Programmiersprache sinnvoll. Python beispielsweise verfügt mit der Bibliothek Pandas über hervorragende Data Science Tools. Hier kann die csv-Datei als sogennanter Data Frame geladen und bearbeitet werden.

Da sich csv-Dateien auch sehr einfach von Tabellenkalkulationsprogrammen einlesen lassen, ist es natürlich auch denkbar, mit Excel oder einer vergleichbaren Software zu arbeiten. Hinderlich ist hier die schiere Fülle an Daten mit fast 30 000 Zeilen. Falls das Nutzen einer Programmiersprache dennoch nicht in Frage kommt, könnte hier auch eine Hybridvariante denkbar sein. Man könnte die Tabelle Wochenweise in unterschiedliche Dateien aufspalten (mit jeweils 2016 Zeilen), die sich einfacher mit einer Tabellenkalkulation bearbeiten lassen. Ist der gesamte Datensatz sehr groß, könnte das Aufteilen der Daten automatisiert über eine Programmiersprache erfolgen (beispielsweise durch den Betreuer des Projekts oder einen externen Helfer).

3.2.2 Ergebnisse der Projektgruppe

Die Schülerinnen und Schüler erhielten eine kurze Einführung in die Programmiersprache Python mit der Bibliothek Pandas. Im Vordergrund stand das gezielte Auswählen von Daten mit Hilfe sogenannter Masken. Damit konnten beispielsweise bestimmte Zeiträume (eine Stunde, ein Tag, eine Woche) selektiert werden, oder man konnte nach bestimmten Events suchen (wann änderte sich die Temperatur um mehr als 2 Grad innerhalb von 15 Minuten?). Außerdem wurde die Visualisierung der Daten mit Hilfe von Schaubildern thematisiert.

Nach dem Programmiercrashkurs verbrachte die Gruppe einige Zeit damit, die Daten zu sichten und jene Tage etwas genauer unter die Lupe zu nehmen, für die ein Lüftungsprotokoll existierte. Aus den resultierenden Beobachtungen konnten die ersten Vermutungen

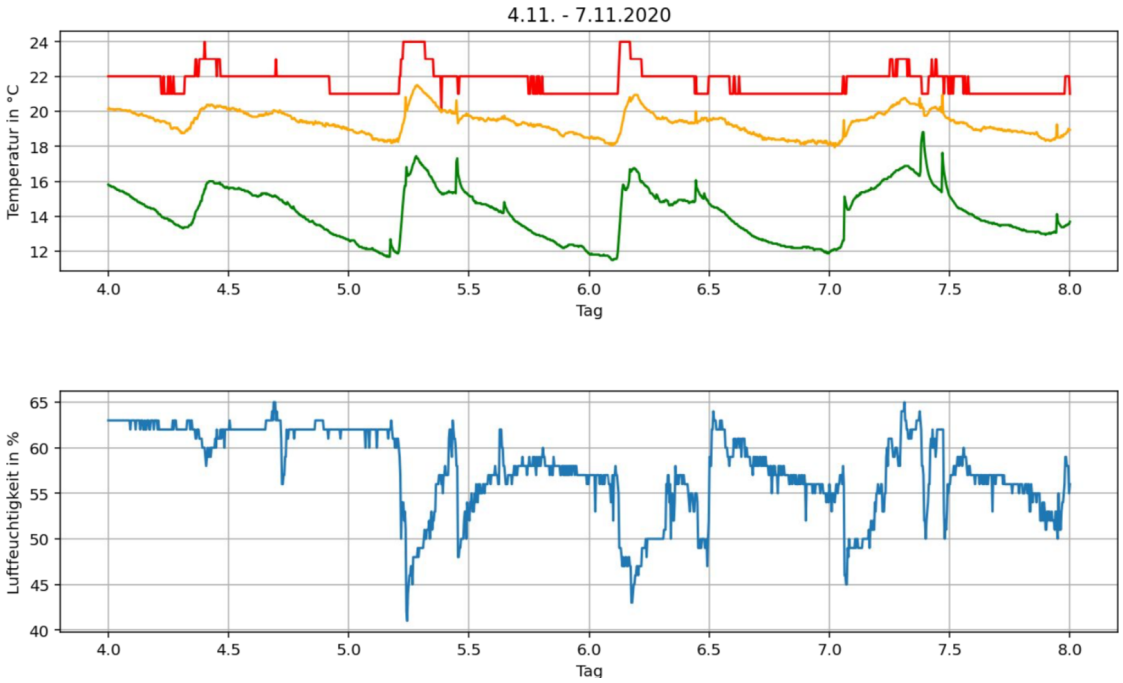

Abbildung 3: Charakteristischer Verlauf über 4 Tage. Ganze Werte auf der Zeitachse entsprechen Mitternacht. Man erkennt, dass oft zu den selben Zeiten gelüftet wurde: Früh am Morgen und kurz vor Mittag.

über das Verhalten des Systems aufgestellt werden. Es wurden mehrere Vermutungen darüber geäußert, unter welchen Bedingungen ein Lüften mehr oder weniger sinnvoll ist.

Diese Vermutungen wurden anschließend durch gezielte Datenanalysen quantifiziert. Es stellte sich heraus, dass die Uhrzeit das wichtigste Kriterium ist. Zu allen Jahreszeiten fiel beim Lüften in den frühen Morgenstunden die Luftfeuchtigkeit am stärksten ab. Im Sommer hatte dies den zusätzlichen Effekt, dass die Raumtemperatur leicht gesenkt werden konnte. Selbst im Winter zeigte sich aus einer Kosten-Nutzen-Betrachtung, dass das Lüften am Morgen sinnvoll ist. Lüften am Nachmittag hingegen birgt im Sommer die Gefahr, dass die Luftfeuchtigkeit noch weiter zunimmt. Hier kam die Gruppe zum Schluss, dass eine bessere Auswertung des Datenmaterials vor allem dann möglich wäre, wenn weitere Umweltinformationen (Temperatur und Luftfeuchtigkeit außerhalb der Wohnung) zur Verfügung stehen würden.

Weiterhin konnte gezeigt werden, dass häufiges Stoßlüften bessere Werte erzielt als Dauerlüften. Interessant ist auch die Tatsache, dass die Gruppe charakteristische Tagesverläufe aufzeigen konnte, die darauf hindeuteten, dass die Wohnung verlassen war. Typisch waren hier – neben einer fehlenden Temperaturschwankung an der Balkontür – überaus konstante Luftfeuchtigkeitswerte. Diese waren an den meisten anderen Tagen stark schwankend.

Neben den reinen Data Mining Prozessen verbrachte die Gruppe auch viel Zeit mit der Visualisierung der Daten. Durch sorgfältige Aufbereitung der Schaubilder konnten die wesentlichen Informationen für die Präsentation ansprechend dargestellt werden.

3.3 Wie sicher sind unsere privaten Daten in sozialen Netzwerken? Was Social-Media-Känale über uns wissen

3.3.1 Problemstellung

Facebook, Instagram, WhatsApp – diese Social-Media-Kanäle sind im heutigen Leben kaum mehr wegzudenken und werden von Menschen auf aller Welt genutzt. Doch gerade die scheinbar selbstverständliche Nutzung dieser Angebote zeigt Problematiken auf: Durch die kostenlose Nutzung werden persönliche Informationen (z.B. Interessen, gesundheitliche Verfassung oder politische Meinung) gesammelt und an Unternehmen verkauft. Solche Unternehmen generieren aus den gesammelten Daten personalisierte Werbung, verbreiten diese über Social-Media-Kanäle gezielt an deren Nutzer/innen und können so ihre Umsätze enorm steigern. Das Alter von Nutzer/innen in einem sozialen Netzwerk ist so eine relevante persönliche Information, die zur gezielten Adressierung von Werbung genutzt wird: Eine Person über 30 interessiert sich weniger für die neueste Hitsingle eines Rappers und dafür eher für Kitaplätze, während das Interesse einer Person unter 20 eher umgekehrt ist. Manche Nutzer/innen möchten sich personalisierten Werbeangeboten entziehen und möglichst keine Spuren im Netz hinterlassen. Sie vermeiden daher die Angabe von persönlichen Informationen oder geben falsche oder unvollständige Informationen von sich preis – aber reicht das aus? In diesem Projekt soll untersucht werden, wie gut das Alter eines Nutzers/einer Nutzerin in einem sozialen Netzwerk vorhergesagt werden kann, auch wenn das Alter der Person nicht explizit angegeben ist. Konkret sollen die Schüler/innen anhand eines vorgegebenen Datensatzes die Altersverteilung der Nutzer/innen analysieren und Regeln für eine Vorhersage des Alters von Nutzer/innen aufstellen, die ihr Alter falsch angegeben haben. Zusätzlich sollen sie untersuchen, wie weitere Informationen über Nutzer/innen gewonnen werden können.

3.3.2 Vorüberlegungen

Als Datengrundlage dient ein realer Datensatz der Website Friendster. Friendster verzeichnet seine Gründung im Jahr 2002 und hatte bis zu 350 Millionen Nutzer (Klanert, 2014). Die Website gilt als Vorbild für weitere Social-Media-Kanäle wie zum Beispiel Facebook. Aktuell ist Friendster nicht mehr zu erreichen, aber die Daten des Netzwerks sind gespeichert worden und können im Internetarchiv[1] heruntergeladen werden. Der Datensatz zeigt neben der Nutzer-ID weitere Informationen zu Geschlecht, Beziehungsstatus, Interessen, Alter, Freunde, Freundes-Freunde, Alter der Freunde und Alter der Freundesfreude. Die Analyse der Altersverteilung der Nutzer/innen gelingt z.B. durch eine geeignete Visualisierung der Nutzerdaten mittels eines Histogramms. Kleinere Nutzergruppen können mithilfe ungerichteter Graphen visualisiert werden, die die Nutzer als Knoten und die Verbindungen zwischen den Nutzern als Kanten abbilden. Ziel der Analyse ist das Herausfiltern von Auffälligkeiten (z.B. sehr alte oder sehr junge Personen). Zusätzlich können hier weitere Informationen des Datensatzes visualisiert werden, um sich einen Überblick über die vorliegenden Daten zu verschaffen. Für die Altersschätzung können unterschiedliche Heuristiken herangezogen werden. In der Mathematik werden Heuristiken im Allgemeinen als Verfahren beschrieben, die für ein Problem eine Lösung liefern, die aber nicht notwendigerweise optimal ist (Marti & Reinelt, 2011). Eine Vorhersage des exakten Alters gelingt beispielsweise durch die Berechnung von Lagemaßen der Altersverteilung der Freunde oder der

[1] https://archive.org/details/archive-team-friendster.

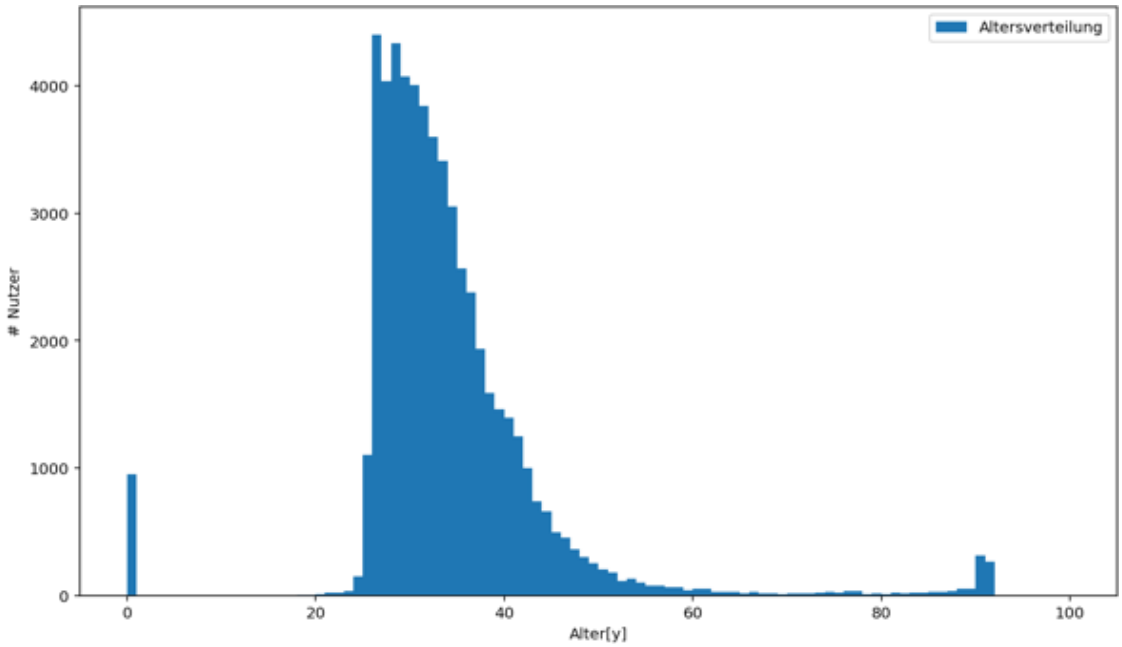

Abbildung 4: Altersverteilung der Nutzer/innen Friendster.

Freundes-Freunde. Die Vorhersageregel zum arithmetischen Mittel könnte lauten: Das Alter des Nutzers entspricht dem auf eine ganze Zahl gerundeten arithmetischen Mittel des angegebenen Alters der Freunde (bzw. der Freundes-Freunde). Analog können Vorhersageregeln zu weiteren Lagemaßen (Median, Modus) aufgestellt werden. Möchte man das Alter eines Nutzers innerhalb eines Intervalls vorhersagen, so eignet sich die Berechnung unterschiedlicher Streumaße, wie Spannweite, Quartilsabstand und Standardabweichung. Die Vorhersageregel zum Quartilsabstand könnte lauten: Das Alter des Nutzers liegt innerhalb des Quartilsabstands des angegebenen Alters der Freunde (bzw. der Freundes-Freunde). Analog können die Vorhersageregeln zu weiteren Streumaßen generiert werden. Möchte man das Alter eines Nutzers innerhalb eines vorgegebenen Intervalls vorhersagen, so kann der gefilterte Datensatz als „wahre Datenmenge“ angenommen und die relative Häufigkeit q berechnet werden, in diesem Datensatz Nutzer/innen mit einem Alter innerhalb eines festgelegten Intervalls, z.B. [20,30]. anzutreffen. Mit einer Wahrscheinlichkeit von q entspricht dann das Alter eines zufälligen Nutzer(s)/in dem Alter im Intervall [20,30]. Die Regel zur Altersvorhersage könnte hier lauten: Mit der Wahrscheinlichkeit q wird dem Nutzer ein Alter zwischen 30 und 40 Jahren zugeordnet. Mit der Wahrscheinlichkeit $1 - q$ ist der Nutzer nicht zwischen 30 und 40 Jahren alt. Diese Heuristiken können auch für weitere Informationen aus dem Datensatz verwendet werden. Das Projekt bietet Einblicke hinter die Kulissen von Social-Media-Kanälen und regt dazu an, das eigene Nutzerverhalten im Netz kritisch zu hinterfragen.

Ergebnisse der Projektgruppe In einem ersten Schritt hat die Gruppe den gegebenen Datensatz aus Friendster genauestens analysiert: Schaut man sich die Altersverteilung der Nutzer/innen in einem Histogramm genauer an (Abbildung 4), wirken die Werte bei 0 und über 90 Jahren unnatürlich und erwecken den Eindruck, dass diese Nutzer/innen ein falsches Alter angegeben haben.

Auf Basis dieser Beobachtung hat sich die Gruppe die Frage gestellt, wie man Falschangaben von Nutzer/innen herausfiltern kann. Gibt es Altersangaben, die offensichtlich falsch

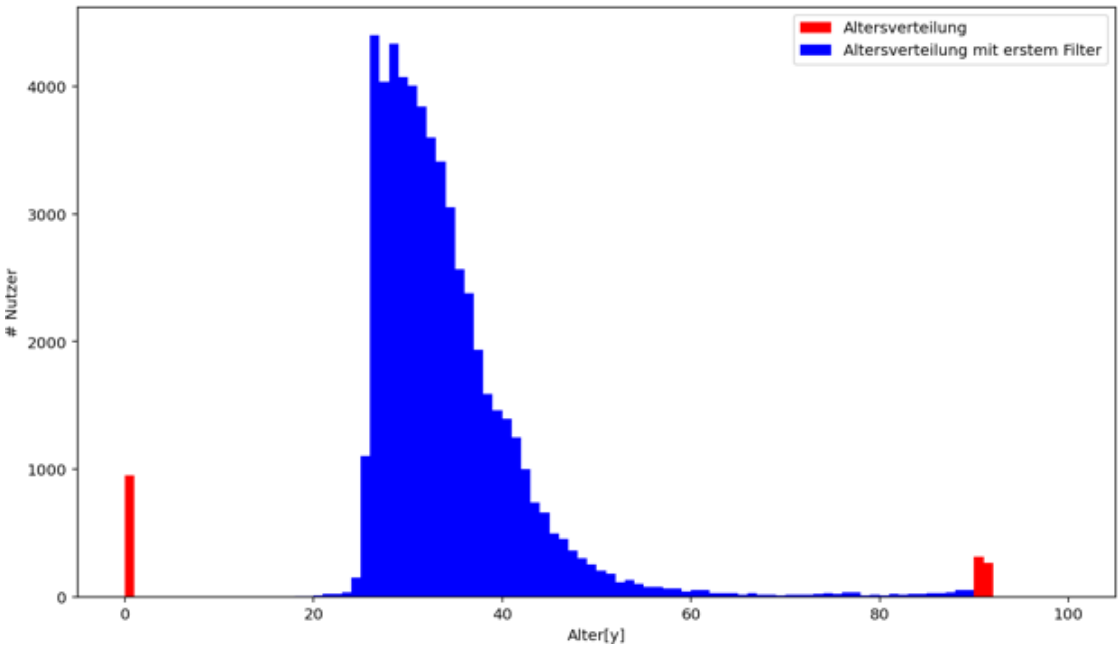

Abbildung 5: Erster Filter.

sind? Lassen sich Zusammenhänge zwischen weiteren Angaben, wie z.B. die Geschlechtsverteilung, die Sexualität und die Verteilung der Interessen sowie des Beziehungsstatus herstellen?

In einem ersten Filter wird das ungewöhnliche bzw. unnatürliche Alter der Nutzer/innen herausgefiltert (Abbildung 5).

Nachdem die Altersangaben herausgefiltert wurden, die offensichtlich falsch angegeben sind, schaute sich die Gruppe den Datensatz genauer an und untersuchte die Altersverteilung bzgl. der Geschlechter sowie die Verteilung der Interessen. Das Diagramm in Abbildung 6 veranschaulicht die Interessensverteilung der jeweiligen Altersgruppen. Es wird deutlich, dass „Freunde“ in nahezu allen Altersgruppen auf hohes Interesse stößt, während „Relationship with Women“ sowie „Dating with Women“ in der Altersgruppe 46–50 auf wenig Interesse stößt. Im Gegensatz dazu hat diese Altersgruppe – im Vergleich zu den anderen Altersgruppen – ein hohes Interesse an „Dating Men“ sowie „Relationship with Men“.

Eine weitere Angabe, die eine Rolle bei der Schätzung des Alters spielt, ist die sexuelle Orientierung der Nutzer/innen (Abbildung 7). Auch hier wird vor allem der Unterschied von der Altersgruppe 46–50 zu den anderen Altersgruppen deutlich: Der Anteil der homosexuellen Orientierung ist in dieser Altersgruppe höher als in anderen Altersgruppen.

In einem weiteren Histogramm wird die Altersverteilung nach Beziehungsstatus untersucht (Abbildung 8). Dieses Diagramm veranschaulicht, dass der am meisten angegebene Beziehungsstatus in allen Altersgruppen „single“ ist.

Nun wird der Sonderfall betrachtet, wenn Nutzer/innen keine Angabe zum Beziehungsstatus gemacht haben (Abbildung 9): Vor allen bei den Profilen, die als Alter 0 angegeben haben, ist der Beziehungsstatus unbekannt.

Neben den persönlichen Angaben spielt auch das Durchschnittsalter der Freunde 1. (direkte Freunde) und 2. (Freundes-Freunde) Grades eine Rolle (Abbildung 10). Im Folgenden wird

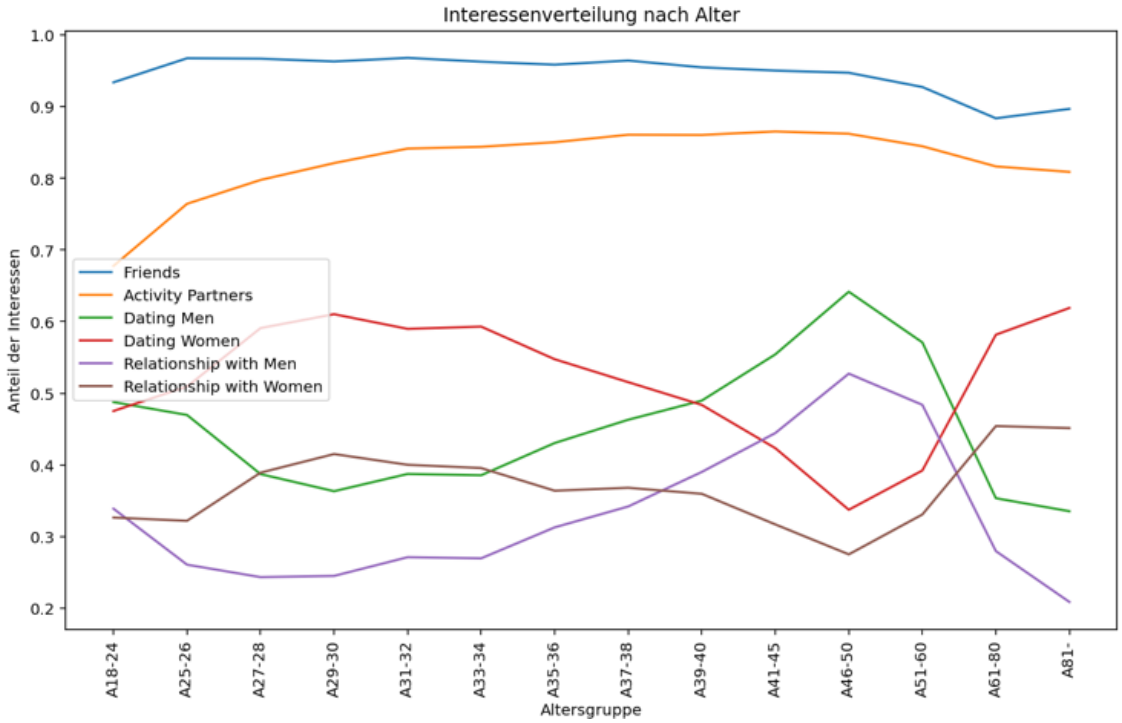

Abbildung 6: Interessensverteilung nach Altersgruppen.

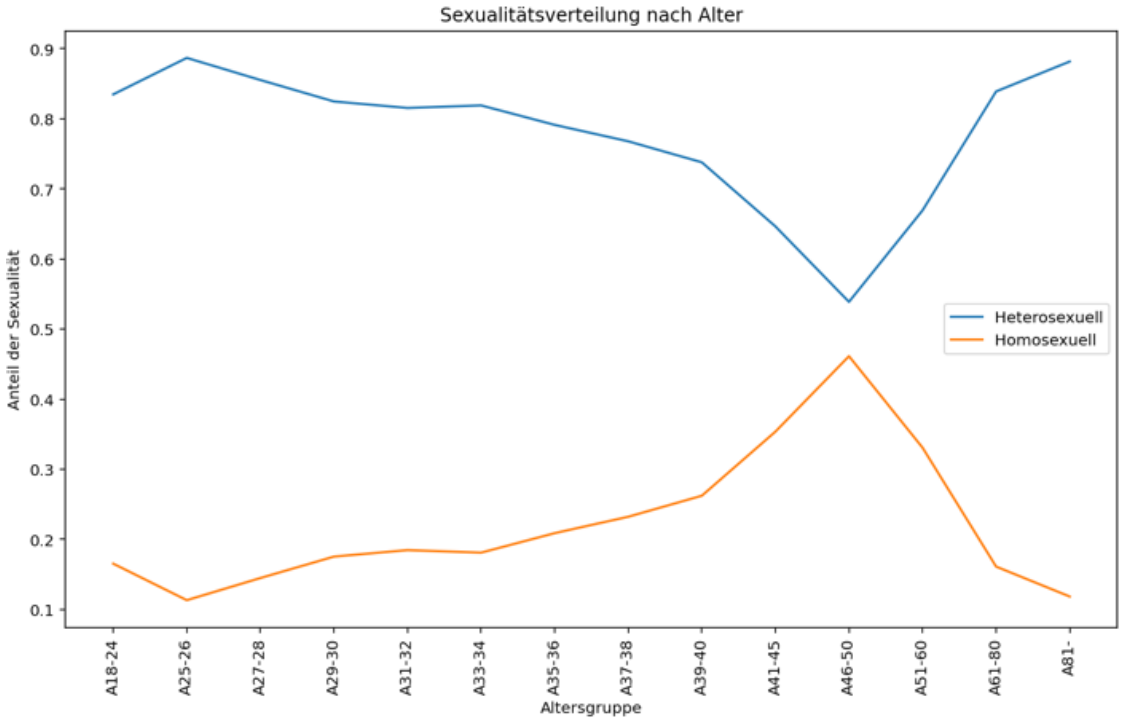

Abbildung 7: Verteilung der sexuellen Orientierung nach Altersgruppen.

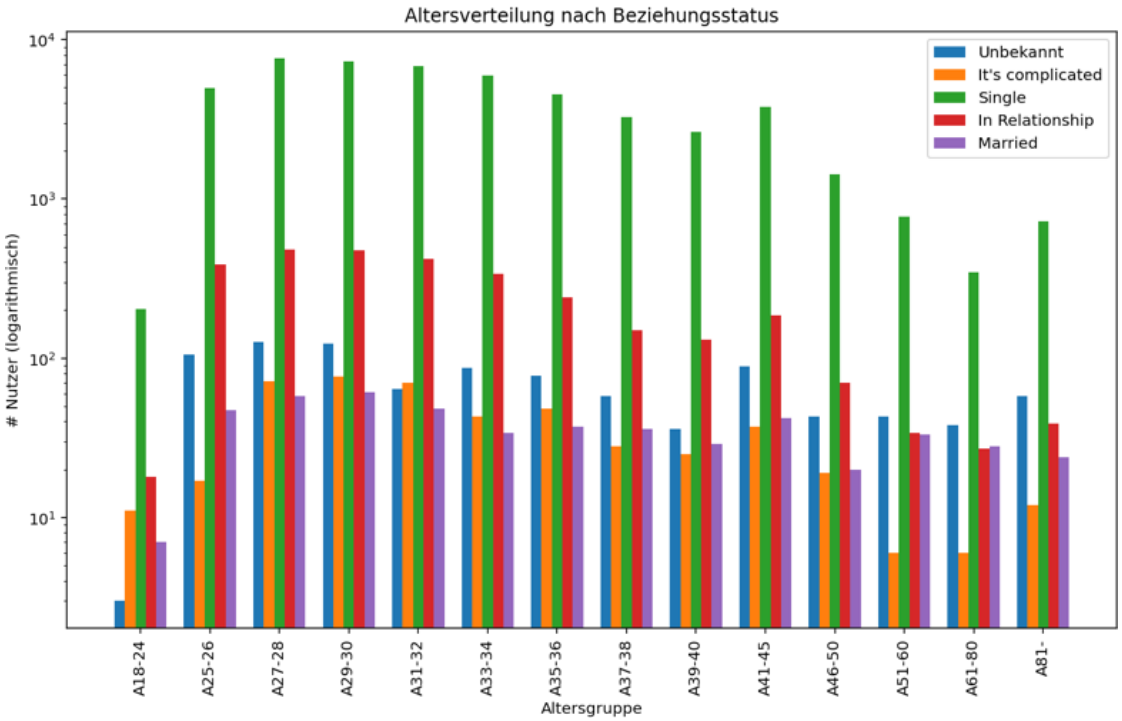

Abbildung 8: Altersverteilung nach Beziehungsstatus.

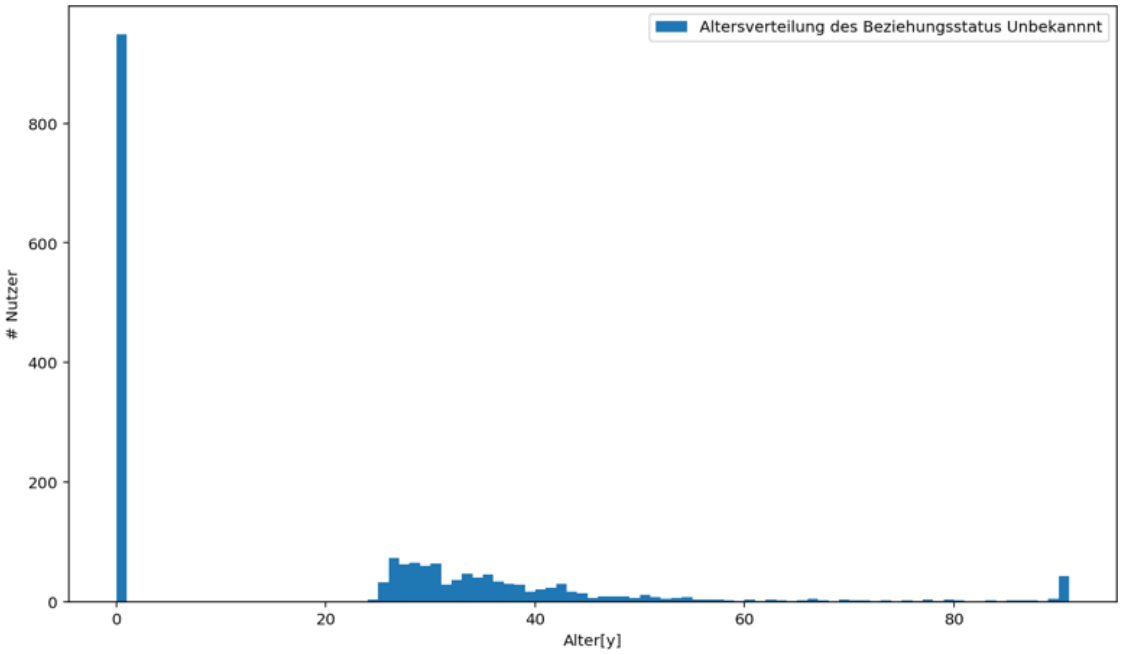

Abbildung 9: Altersverteilung bei Beziehungsstatus „Unbekannt“.

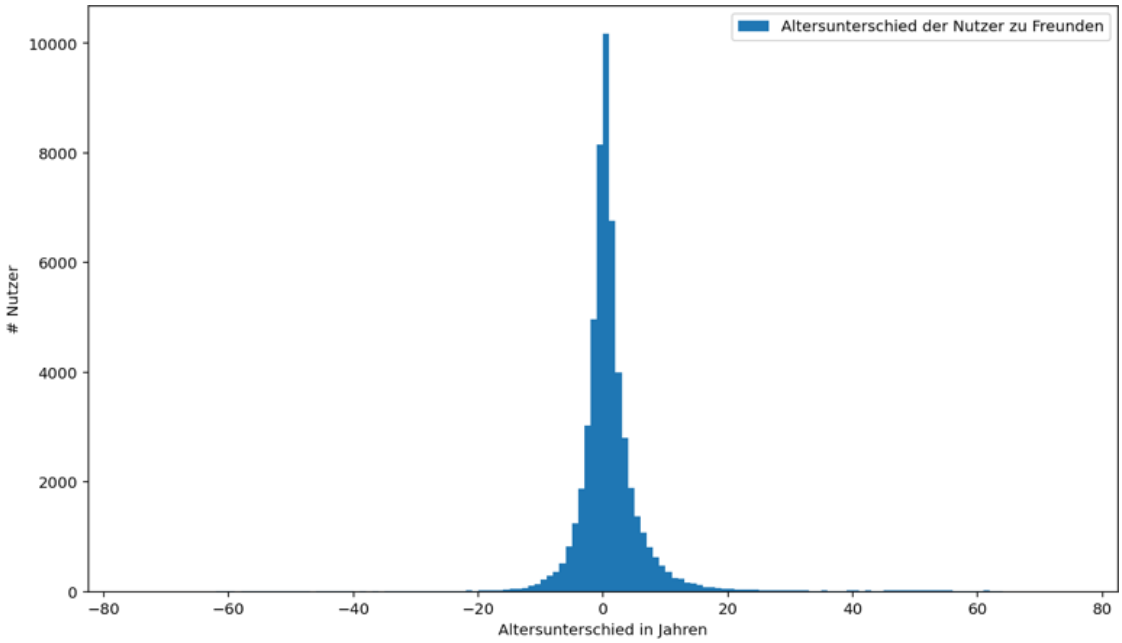

Abbildung 10: Durchschnittsalter Freunde 1. und 2. Grades.

die Übereinstimmung des Freundesalters mit dem eigenen Alter untersucht. Je größer das Intervall gewählt wird, desto höher ist die Übereinstimmung (Tabelle 1).

Intervallgröße	Übereinstimmung
0 Jahre	17,2%
1 Jahr	44,1%
2 Jahre	60,3%
3 Jahre	70,8%
4 Jahre	77,6%
5 Jahre	82,3%
6 Jahre	85,7%
7 Jahre	88,1%
8 Jahre	89,9%
9 Jahre	91,3%
10 Jahre	92,3%

Tabelle 1: Übereinstimmung des Freundesalter mit dem eigenen Alter.

Nachdem der Datensatz analysiert wurde, entwickelte die Gruppe ein Vorgehen, mit dem das Alter der Nutzer/innen bestmöglich geschätzt werden kann, auch wenn keine oder falsche Angaben zum Alter getätigt wurden. Ausgangspunkt für eine „gute“ Schätzung sind folgende Kriterien: Anzahl der Freunde, das mittlere Alter der Freunde, die sexuelle Orientierung sowie der Beziehungsstatus. Dabei wären weitere Kriterien wie Interessen sowie die Anzahl Freundes-Freunde möglich gewesen, spielen aber eine ungeordnete Rolle. Mit dem entwickelten Verfahren werden 33% der Altersgruppen exakt richtig geschätzt. Im Mittel kann das Alter auf drei Jahren geschätzt werden. Abbildung 11 veranschaulicht die Altersverteilung vor und nach der Filterung und Schätzung des Alters. Es wird deutlich: 6,695% aller Angaben bzgl. des Alters sind falsch und werden durch Schätzungen ersetzt.

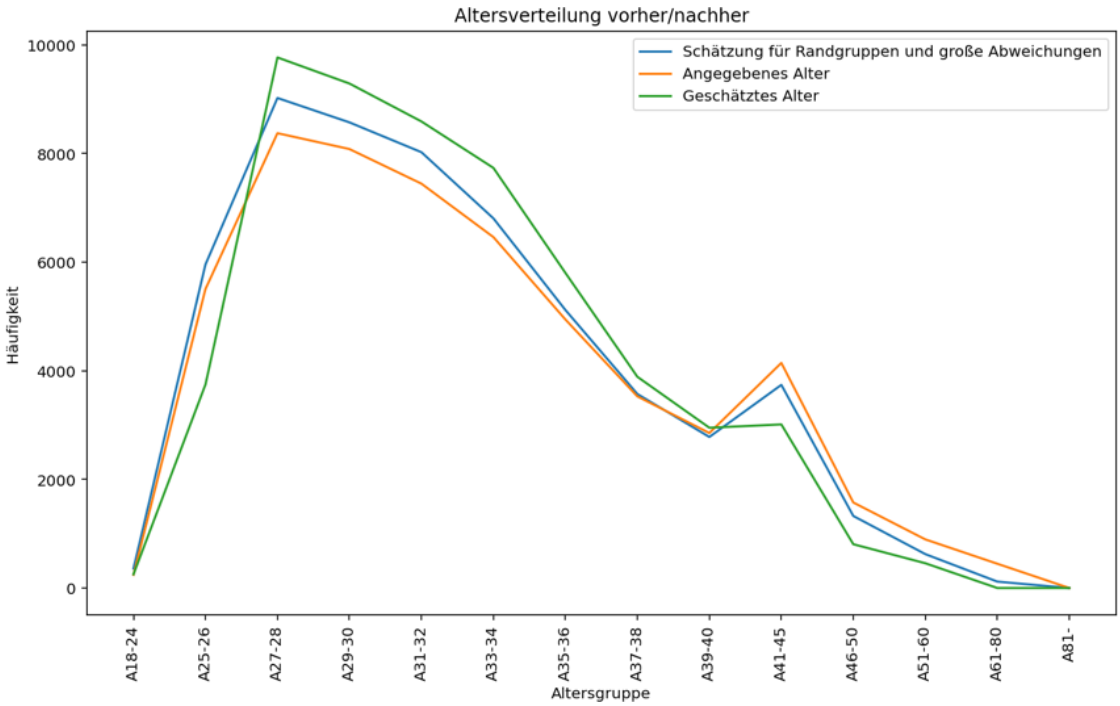

Abbildung 11: Altersverteilung vor/nach der Filterung und Schätzung des Alters.

3.3.3 Quellen

Klanert, J. (2014). Das Online-Verhalten der User wird immer komplexer. Zugriff auf http://www.focus.de/digital/experten/klanert/wandel-durch-das-internet-das -online-verhalten-der-user-wird-immer-komplexer_id_4078023.html (18.12.2020)

Marti, R. & Reinelt, G. (2011). The linear ordering problem. exact and heuristic methods in combinatorial optimization. Springer Verlag: Berlin; Heidelberg.