Data Science 1 Projekt Report

Zahra Abdi Christopher Grosse Aylin Demir

1 Einleitung

In unserem Projekt untersuchen wir anhand der beiden Datensätze *World Happiness Report 2017* [1] und *The Human Freedom Index 2019* [2] wie die Werte Happiness (Glückseligkeit) und Freedom (Freiheit) korrelieren. Beide Datensätze geben die ermittelten Werte des Jahres 2017 auf einer Skala zwischen 0 bis 10 in Bezug auf das jeweilige Land wieder und erstellen über die ermittelten Werte jeweils eine entsprechende Rangliste.
Wir nehmen eine positive Korrelation der Werte Glückseligkeit und Freiheit an, wobei wir davon ausgehen, dass die Korrelation mittelstark ausgeprägt ist. Unsere Annahme stützen wir auf die Beobachtung, dass vor allem Glückseligkeit ein sehr subjektives Gefühl ist und sie zusätzlich relativ zur beobachteten Gruppe abhängt. Das bedeutet, dass eine Person A in der Gruppe X die mit Abstand glücklichste Person sein kann, jedoch genau diese Person A in der Gruppe Y die geringste Glückseligkeit aufweist. Die Cantril-Leiter [3], welche im von uns genutzten Datensatz *World Happiness Report* als Evaluationsmethode genutzt wurde, unterstreicht die Subjektivität von Glückseligkeit. Auch der Freiheitswert aus dem Datensatz *The Human Freedom Index* spiegelt das subjektive Empfinden über die persönlichen Freiheiten [4] wider, wobei zusätzlich als Ausgleich auch fest messbare Kenngrößen im Bereich der wirtschaftlichen Freiheit [5] einfließen [6]. Anhand der Clusteralgorithmen K-Means [7] und Affinity-Propagation [8] zeigen wir Gruppierungsmöglichkeiten der kombinierten Werte Glückseligkeit und Freiheit, um Schwachstellen und Optimierungsmöglichkeiten der statistisch starken Aussagekraft einer möglichen Korrelation aufzuzeigen.

2 Hauptteil

Im Folgenden beschreiben wir detailliert die wichtigsten Schritte unseres Projekts, welche die Aufbereitung und Zusammenführung der Datensätze, die Verifikation und Validierung des neuen Datensatzes, die Anwendung von zwei Clusteralgorithmen auf unseren Datensatz und die Auswertung der erworbenen Erkenntnisse sind.

Informationen zu den Datensätzen:

Information	Happiness Report	Freedom Report
Reihenanzahl - Länderanzahl	155	162
Spaltenanzahl - Kategorien	12	120
Anzahl verwendeter Spalten	3	6

2.1 Datenbereinigung und Zusammenführung

Für eine Zusammenführung der beiden Datensätze *World Happiness Report* und *The Human Freedom Index* ist zunächst eine Aufbereitung der Daten notwendig. Im Folgenden werden die projektrelevanten Spalten aus den Datensätzen aufgeführt.

Datensatz *World Happiness Report 2019*: Country, Happiness Rank, Happiness Score
Datensatz *The Human Freedom Index 2019*: year, ISO_code, countries, region, hf_score, hf_rank

Beide Datensätze enthalten zusätzlich Spalten, welche wir zunächst aufgrund einer zukünftigen Nutzung beibehalten oder wegen ihrer fehlenden Aussagekraft in Bezug auf unsere Annahme direkt aussortieren [9]. Aufgrund der zur Verfügung stehenden Informationen haben wir entschieden, die beiden Datensätze über die Spalte „Ländernamen“ zusammenzuführen, da so

eine eineindeutige Zuordnung möglich ist. Aus diesem Grund ist als erstes die Angleichung der Ländernamen beider Datensätze notwendig, damit keine Daten verloren gehen oder fehlerhaft zusammengefügt werden. Zur Überprüfung und Änderung der Daten importieren wir zunächst die beiden Datensätze in SQL, um sie mittels SQL-Anweisung zusammenzufügen [10]. Im nächsten Schritt werden die inkompatiblen Ländernamen mittels SQL-Anweisung extrahiert.
Insgesamt ist es möglich 147 Ländernamen so aufzubereiten, dass sie in einem gemeinsamen Happiness-Freedom-Korrelations-Datensatz (HFK-Datensatz) für unsere Datenanalyse zur Verfügung stehen. Ohne Bearbeitung sind 133 Ländernamen kompatibel. Bei 12 Ländernamen ist eine Korrektur beziehungsweise Anpassung notwendig [11]. 23 Ländernamen sind kompatibel, jedoch enthalten 15 keinen Glückseligkeits- beziehungsweise acht keinen Freiheitswert [12]. Zusätzlich ist es notwendig ein Komma innerhalb eines Ländernamens zu entfernen [13]. Für den Spezialfall Zypern nutzen wir zusätzliche Informationen, um eine Datenbasis zu generieren, welche die Zusammenführung der kombinierten Glückseligkeitswerte von Nordzypern und Zypern mit dem Freiheitswert von Zypern ermöglicht [14].

2.2 Test und Verifikation der Datenqualität

Die erste Verifikation des HFK-Datensatzes zeigt hinsichtlich der Ländernamen eine 100 prozentige Korrektheit, jedoch wurden nur 96 anstatt der erwarteten 147 Länder angezeigt. Die Ursache des Problems entsteht beim Einlesen der CSV-Datei *The Human Freedom Index* in die SQL-Datenbank, da diese Spalten enthält, in denen nicht für alle Länder ein entsprechender Wert vorliegt. Die fehlenden Informationen wurden von den Autoren mittels Bindestrichzeichen dargestellt. Um das korrekte Einlesen zu gewährleisten, besteht die Notwendigkeit, die Bindestriche durch einen numerischen Wert zu ersetzen, sodass für die SQL-Datenbank der Datentyp-Konflikt innerhalb einer Spalte zwischen String- und Double-Werten ausgeräumt ist. Wir ersetzen alle Bindestriche durch den Wert 00, da so weiterhin die Unterscheidung von den korrekten 0-Werten gewährleistet ist. Zusätzliche Tests und eine vollständige Verifikation des HFK-Datensatzes versichern, dass die Zusammenführung eine verlässliche Datengrundlage zur Analyse bietet.

2.3 Anwendung von K-Means und Affinity Propagation auf die Daten

Für die Visualisierung und Anwendung der Clusteralgorithmen nutzen wir Python.

Abbildung 1:
Datenpunktewolke mit Regressionsgerade

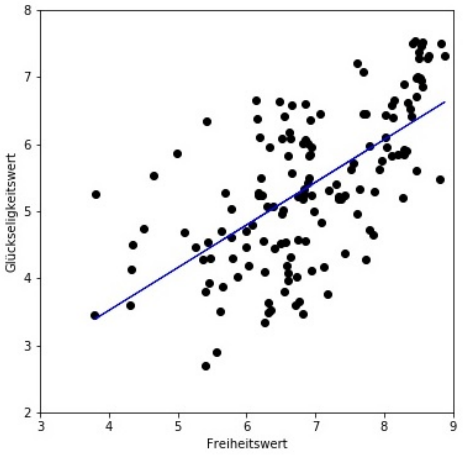
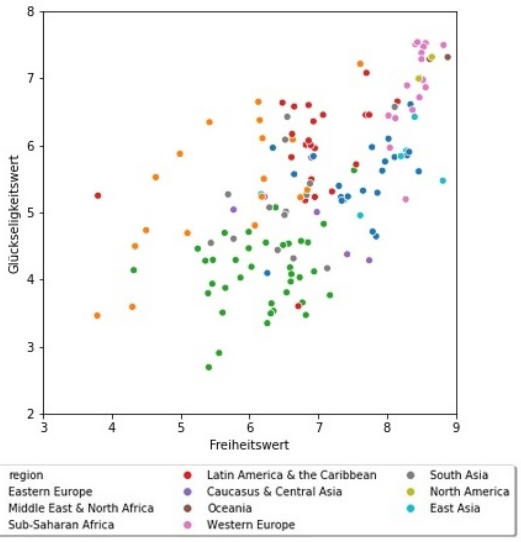


Abbildung 2:
Datenpunktewolke in Regionen unterteilt



Eine erste Analyse der Punktewolke mit der Regressionsgeraden bestätigt die von uns getroffene Annahme der positiven Korrelation zwischen Glückseligkeit und Freiheit.

Zur detaillierteren Analyse unseres HFK-Datensatzes verwenden wir die beiden Clusteralgorith-

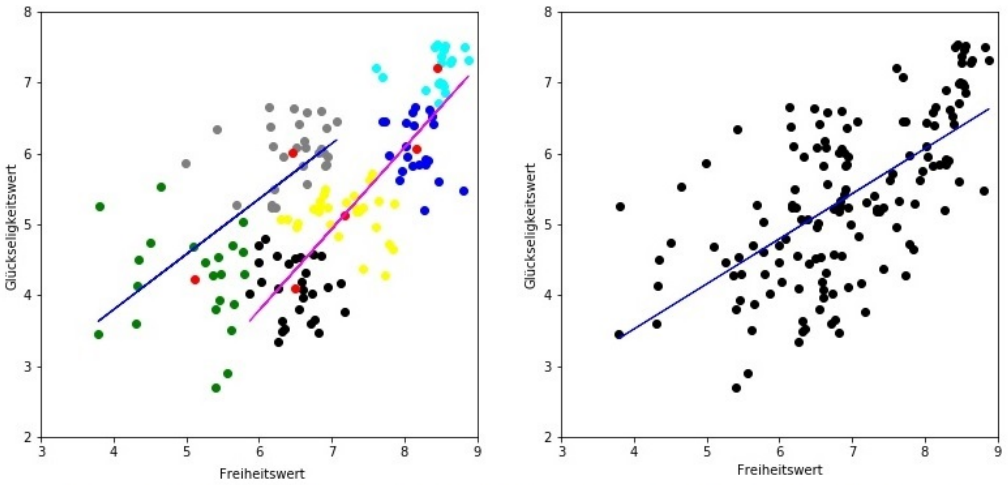
men K-Means [15] und Affinity-Propagation [16].

2.3.1 K-Means

Der eingesetzte K-Means-Algorithmus [17] unterteilt die Daten in die von uns vorgegebene Anzahl von 6 Cluster, welche automatisch durch den Algorithmus initialisiert und optimiert werden.

Abbildung 3:
K-Means-Algorithmus mit Regressionsgeraden durch die entstandenen Cluster grau und grün sowie blau, cyan, gelb und schwarz

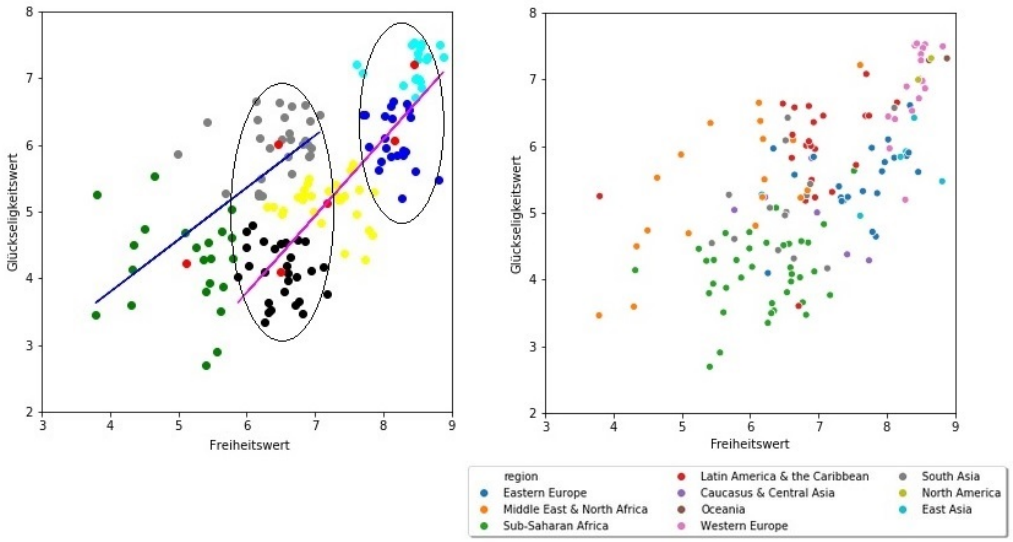
Abbildung 1:
Datenpunktewolke mit Regressionsgerade zum Vergleich



Im Vergleich zur reinen Datenpunktewolke (rechts) ermöglicht die Abgrenzung durch die Cluster (links) detailliertere Analysen. Wir interpretieren die Daten durch zwei Geraden mit geringfügig veränderter Steigung jedoch mit einem erheblich unterschiedlichen Werteniveau. Der Grund des Unterschiedes ist durch die vorhandenen Daten nicht direkt erklärbar.

Abbildung 4:
Entspricht Abbildung 3 mit zusätzlichen Markierungen von Artefakten bezüglich der Korrelation

Abbildung 2:
Datenpunkte in Regionen unterteilt



Es ist erkennbar, dass Bestimmte Cluster bei gleichen Freiheitswerten eine stark erhöhte Glückseligkeitswerte aufweisen. Im Vergleich beider Abbildungen weisen diese Cluster jeweils gehäuft Länder einer bestimmten Region auf. Dies verbessert die Möglichkeiten die Anomalien durch zusätzliche Daten zukünftig erklären zu können.

2.3.2 Affinity-Propagation

Der Affinity-Propagation-Algorithmus [18] ermittelt die Anzahl der Cluster und die Clusterzentren anhand der vorhandenen Daten und der Gruppierungsmöglichkeiten selbstständig. Selbst

eine extreme Variation [19] der Startpunkte führt zu gleichen Ergebnissen.

Abbildung 5:
Affinity-Propagation-Algorithmus zur Clusterung mit Regressionsgeraden der ursprünglichen Datenpunktewolke (Cyan) und Regressionsgerade durch die Clusterzentren (Magenta)

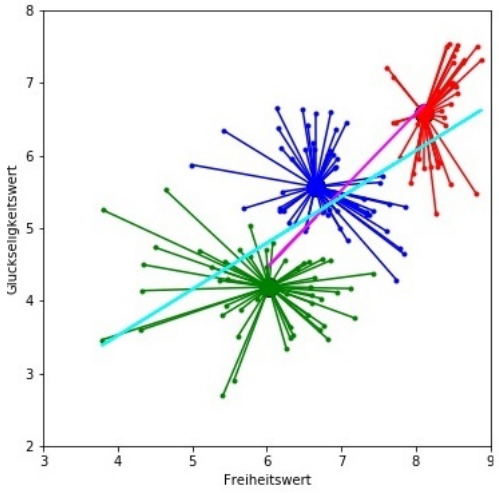
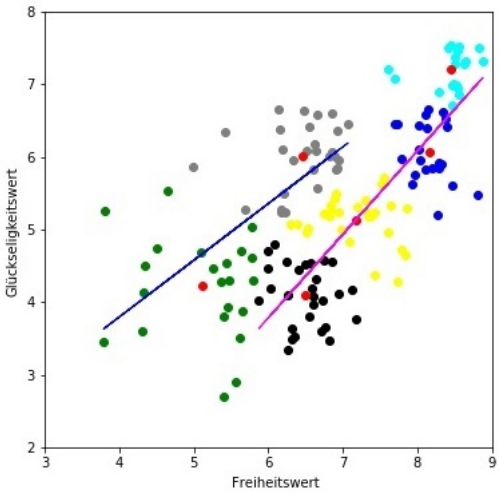


Abbildung 3:
K-Means-Algorithmus zum Vergleich



Im Vergleich zeigt sich, dass schon der unterschiedliche Anstieg der Geraden Interpretationsspielräume ermöglicht, aber vor allem trägt die Wahl des Algorithmus kombiniert mit der Clusteranzahl zu verschiedenen Sichtweisen und Interpretationen bei.

3 Evaluation

Durch die Verwendung der Clusteralgorithmen und deren mehrmalige Ausführung verfestigt sich unsere Annahme der positiven Korrelation zwischen Glückseligkeit und Freiheit, jedoch deckt sie auch Anomalien auf und bestärkt uns in der Annahme, dass die Korrelation nicht so stark sein könnte, wie sie durch den Anstieg der Regressionsgeraden wiedergegeben wird.
Weiterführende Forschungen des Projektes richten sich auf die Erweiterung der Daten durch Informationen wie das BIP der Länder und andere Einflussfaktoren, welche innerhalb der Cluster gehäuft auftreten oder fehlen, um die Korrelation zwischen Glückseligkeit und Freiheit exakter bestimmen, sie gegebenenfalls mittels regionaler Unterschiede differenzieren und ihren Grenznutzen definieren zu können.

4 Fazit

Die Verwendung von Clusteralgorithmen eröffnet zusätzliche Interpretationsmöglichkeiten der Daten. Hierbei ist zu beachten, dass der angewandte Algorithmus starke Auswirkungen auf die Interpretationsspielräume ausübt. Die einseitige, ungefilterte Verwendung oder Interpretation der Ergebnisse eines einzigen Algorithmus kann zu fehlerhaften Schlussfolgerungen oder Unterbeziehungsweise Überbewertung des Informationsgehalts der Daten und somit zu einer Verzerrung der Wirklichkeit führen. Zur Vermeidung einer möglichen Verzerrung empfehlen wir jeden Algorithmus mehrfach auf die Daten anzuwenden, sodass das Ergebnisspektrum abschätzbar ist. Zusätzlich sollten verschiedene Algorithmen zur Datenauswertung genutzt werden, um einen wissenschaftlichen und fachgerechten Gesamtüberblick möglicher Ergebnisse zu gewährleisten und sichere Aussagen treffen zu können.

Relevante Links

Github: https://github.com/aylin-d/ZACData_DataScience1
Freedom Index 2019: https://www.kaggle.com/gsutters/the-human-freedom-index?select=hfi_cc_2019.csv

World Happiness: https://www.kaggle.com/unsdsn/world-happiness?select=2017.csv

Literatur

[1] *World Happiness Report: 2017.csv.* https://www.kaggle.com/unsdsn/world-happiness?select=2017.csv, 24.6.20.

[2] *The Human Freedom Index: hfi-cc-2019.csv.* //www.kaggle.com/gsutters/the-human-freedom-index?select=hfi_cc_2019.csv, 24.6.20.

[3] *World Happiness Report.* https://www.kaggle.com/unsdsn/world-happiness, 24.6.20.

[4] *The Human Freedom Index 2017, ab Seite 15.* https://object.cato.org/sites/cato.org/files/human-freedom-index-files/2017-human-freedom-index-2.pdf, 24.6.20.

[5] *The Human Freedom Index 2017, ab Seite 15.* https://object.cato.org/sites/cato.org/files/human-freedom-index-files/2017-human-freedom-index-2.pdf, 24.6.20.

[6] *The Human Freedom Index 2017.* https://object.cato.org/sites/cato.org/files/human-freedom-index-files/2017-human-freedom-index-2.pdf, 24.6.20.

[7] *Data Science 1 Projekt Report Anhang, K-means Algorithmus.* https://github.com/aylin-d/ZACData_DataScience1/blob/master/ZACData_Anhang.pdf.

[8] *Data Science 1 Projekt Report Anhang: Affinity Propagation Algorithmus.* https://github.com/aylin-d/ZACData_DataScience1/blob/master/ZACData_Anhang.pdf.

[9] *Data Science 1 Projekt Report Anhang: Tabelle1 und Tabelle2.* https://github.com/aylin-d/ZACData_DataScience1/blob/master/ZACData_Anhang.pdf.

[10] *Data Science 1 Projekt Report Anhang: SQL-Code.* https://github.com/aylin-d/ZACData_DataScience1/blob/master/ZACData_Anhang.pdf.

[11] *Data Science 1 Projekt Report Anhang: Tabelle3.* https://github.com/aylin-d/ZACData_DataScience1/blob/master/ZACData_Anhang.pdf.

[12] *Data Science 1 Projekt Report Anhang: Tabelle4.* https://github.com/aylin-d/ZACData_DataScience1/blob/master/ZACData_Anhang.pdf.

[13] *Data Science 1 Projekt Report Anhang: Tabelle5.* https://github.com/aylin-d/ZACData_DataScience1/blob/master/ZACData_Anhang.pdf.

[14] *Data Science 1 Projekt Report Anhang: Tabelle6.* https://github.com/aylin-d/ZACData_DataScience1/blob/master/ZACData_Anhang.pdf.

[15] *Data Science 1 Projekt Report Anhang: K-means Algorithmus.* https://github.com/aylin-d/ZACData_DataScience1/blob/master/ZACData_Anhang.pdf.

[16] *Data Science 1 Projekt Report Anhang: Affinity Propagation Algorithmus.* hhttps://github.com/aylin-d/ZACData_DataScience1/blob/master/ZACData_Anhang.pdf.

[17] *The Most Comprehensive Guide to K-Means Clustering You'll Ever Need.* https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/, 24.6.20.

[18] *Demo of affinity propagation clustering algorithm.* https://scikit-learn.org/stable/auto_examples/cluster/plot_affinity_propagation.html, 24.6.20.

[19] *DataScience1_ZACData Jupyter Notebook.* https://github.com/aylin-d/ZACData_DataScience1/blob/master/DataScience1_ZACData1.ipynb.