Statistical Computing - DLBDBSC01

Einleitung

Die wohltätige Organisation WealthyGrowth möchte mit Kampagnen das Wirtschaftswachstum in Entwicklungsländern fördern. Konkret bedeutet das die Steigerung des GDP per capita (Bruttoinlandsprodukt pro Kopf). Für diese Kampagnen wurden drei Ansätze aufgestellt, deren Potenzial in dieser schnellen Analyse untersucht wird. Der erste Ansatz hat das Ziel einer Schulpflicht, dadurch soll eine höhere Bildung der breiten Bevölkerung erreicht werden. Der zweite Ansatz möchte den Anteil der erwerbstätigen Frauen erhöhen, da somit mehr Arbeitskraft zur Verfügung stehen würde. Der dritte Ansatz hat eine Erhöhung der Lebenserwartung als Ziel, weil dadurch ebenfalls mehr Arbeitskraft verfügbar wäre.

Diese Datenanalyse soll eine erste Aussage über das Potenzial der einzelnen Ansätze liefern und Empfehlungen für das weitere Vorgehen aussprechen. Umgesetzt wird dies mithilfe der frei verfügbaren Datensätze der Weltbank. Aus diesem werden relevante Indikatoren ausgewählt und die Daten durch Bereinigen oder Interpolieren von fehlenden Werten für die Analyse vorbereitet. Für die Interpolation wird nur eine einfache lineare Regression verwendet, damit möglichst schnell erste Ergebnisse vorliegen. Der in den Ansätzen vermutete Einfluss verschiedener Faktoren auf das Wirtschaftswachstum wird mithilfe von Korrelationsanalysen validiert. Basierend darauf entstehen die Empfehlungen für die Kampagnen. Im Folgenden ist die Umsetzung genauer dokumentiert.

Quellcode, Daten und Ergebnisse sind zusätzlich in diesem GitHub Repository zu finden: https://github.com/Patzold/Statistical-Computing-DLBDBSC01

Datenbeschaffung

Auswahl der Indikatoren

Die folgenden Details zu den jeweiligen Indikatoren beruhen auf den Angaben der Weltbank zu den jeweiligen Datensätzen (data.worldbank.org). Da alle Kampagnen die Steigerung des GDP per capita als Ziel haben, wird dieser Indikator natürlich in jedem Fall benötigt. Der Indikator ist nicht inflationsbereinigt (aktuelle US-Dollar / current US$) und in tausend angegeben. In jeder Hypothese ist eine Variable/Einflussfaktor auf die Zielgröße GDP per capita vorgeschlagen. Somit ist GDP per capita immer die abhängige Variable und wird zur späteren Korrelationsanalyse bei jeder Hypothese benötigt.

Der erste Ansatz verfolgt den Einflussfaktor Bildung. Ein dafür wichtiger Indikator ist die Einschulungsrate (scholl enrollment). Diesen gibt es nach Bildungsgrad aufgeschlüsselt, also primary, secondary und tertiary, sowie die Unterteilung in gross oder net Prozentangaben. Die net enrollment rate bezieht sich auf die Einschulungsrate von Kindern im Schulalter des jeweiligen Landes, bei der Bruttoangabe gibt es keine Selektierung nach Alter. Zunächst einmal verwenden wir die „Gross enrollment ratio“, da diese nicht nach Altersgruppen filtert und somit einen umfassenderen Einblick in die Bildungslevel der breiten Bevölkerung eines Landes gibt. Da sich die Hypothese explizit auf die höhere Bildung bezieht, fällt der Indikator für primary education (kann mit Grundschulbildung übersetzt werden) raus. Die Indikatoren für die Einschulungsrate der secondary education und tertiary education werden beide zur Beurteilung der Hypothese verwendet. Neben der Einschulungsrate gibt auch noch andere Indikatoren für den Bildungsbereich, z.B. die Abschlussquote. Allerdings ist die Abschlussquote im Vergleich zur Einschulungsrate weniger passend für die im Ansatz vorgeschlagene Maßnahme der Schulpflicht. Eine Schulpflicht umfasst schließlich nur den Schulbesuch, verpflichtet aber i.d.R. nicht zu einem Schulabschluss. Gleiches gilt für weitere Indikatoren wie die Progressionsrate oder dem Alphabetisierungsgrad. Nur der Indikator für Staatsausgaben für Bildung, gemessen in Prozent vom Bruttoinlandsprodukt, kann nützlich sein um einen umkehrten Kausalzusammenhang

zwischen Wirtschaftswachstum und Schulbildung etwas genauer zu beleuchten. Aus diesem Grund ist auch er Teil der weiteren Datenverarbeitung.

Der zweite Ansatz befasst sich mit dem Anteil erwerbstätiger Frauen. Die Geburtenrate (fertility rate) wird hier als entsprechender Indikator verwendet, dieser ist in Geburten pro Frau angegeben. Alternative Indikatoren wie die Arbeitslosenquote berücksichtigen nicht für die Hypothese relevante Faktoren wie den Mutterschaftsurlaub und werden daher nicht verwendet (International Labour Organization, 1998).

Die dritte Hypothese verfolgt eine Erhöhung der Lebenserwartung, da dieser Indikator bereits existiert verwenden wir auch die life expecancy at birth, welche in Jahren angegeben wird. Eine Möglichkeit um die Lebenserwartung zu steigern, ist eine bessere Qualität der Gesundheitsversorgung für die breite Bevölkerung. Konkret bedeutet dies eine Erhöhung des Anteils an DPT-Immunisierungen von Kindern. DPT ist eine Abkürzung für Diphtherie, Keuchhusten & Tetanus. Kindern, denen diese Immunisierungen fehlen sind anfälliger für die entsprechenden Infektionen, welche auch langfristige gesundheitliche Folgen haben können. Für diesen Anteil existiert bereits ein Indikator als Prozentangabe, welcher auch verwendet wird. Des Weiteren beschreibt der Zugang zu Elektrizität auch die Qualität der Gesundheitsversorgung mit, da praktisch alle medizinischen Geräte Strom benötigen. Hierfür stellt die Weltbank ebenfalls einen Indikator zu Verfügung, Prozent der Bevölkerung angegeben wird. Die Daten der Weltbank sind unter data.worldbank.org zu finden und werden als CSV-Datei heruntergeladen. Neben der Datei für die Daten an sich sind im Download zusätzlich noch zwei Dateien mit Metadaten, eine Datei für den Indikator und eine Datei für die Länder und Aggregate (siehe nächstes Kapitel) enthalten.

Datenbereinigung

Da es sich um tabellarische Daten handelt bietet sich die Python-Bibliothek Pandas mit der DataFrame Datenstruktur für solche Anwendungsfälle bestens an. Der Aufbau der CSV-Dateien ist für alle Indikatoren identisch, wir betrachten im Folgenden als Beispiel die heruntergeladenen Daten für GDP per capita. Die ersten vier Zeilen der CSV-Datei besteht aus Informationen zur Datenquelle (World Development Indicators) und dem letzten Änderungsdatum. Auch wenn diese Informationen recht nützlich sein können, werden sie für die weitere Auswertung nicht benötigt und nur die darauffolgenden Zeilen verwendet.

Die darauffolgenden Zeilen bestehen aus einer Auflistung aller Länder und aggregierten Regionen, mit dem jeweiligen Indikator Wert von 1960 bis 2023. Falls für ein Jahr kein Wert vorliegt, handelt es sich um ein leeres Feld. Neben den Spalten für Landesname und den Jahren gibt es auch noch welche für den Landescode, Indikatorname (bei jeder Spalte gleich) und Indikatorcode (ebenfalls bei jeder Spalte gleich). Da Indikatorname & -code überall gleich sind, bereits aus dem Dateinamen bekannt und im weiteren Verlauf der Datentransformation nicht benötigt, werden sie aus beiden Spalten entfernt.

Wie bereits erwähnt sind Felder für die Jahre, in denen kein Wert des Indikators für ein Land verfügbar ist, leer. In Pandas wird dies durch eine NaN Value (in neueren Versionen auch als NA) dargestellt. Um einen ersten Überblick über die Verteilung und Menge der NaN Werte zu bekommen, wird für jede (Länder)Spalte sowohl die Anzahl der dort vorhandenen NaN Werte, als auch deren relativen Anteil an allen Werten des Landes in einer Tabelle ausgeben. Diese Tabelle ist jeweils sehr lang und im GitHub repository zu finden. Alle Länder haben mindestens einen NaN Wert. Erklärbar ist dies, da eine (nicht so benannte) Spalte für das Jahr 2023 existiert, für das Jahr selbst aber noch keine Werte vorliegen. Schließlich ist es auch noch nicht vorbei und daher kann noch keine Werte für einen Indikator, der sich auf das gesamte Jahr beziehen, feststehen. Die Gründe für warum Werte fehlen könnten unterschiedlich sein. Manche jüngeren Länder existierten schlichtweg für einen Teil der Zeitspanne noch nicht, andere erheben den entsprechenden Indikator gar nicht oder stellen ihn nicht öffentlich zur Verfügung.

Um ein Land zum Treffen von Aussagen verwenden zu können, sollte ein Mindestmaß an Informationen vorhanden sein. Länder mit zu vielen leeren Werten werden von der weiteren

Datenanalyse ausgeschlossen. Ein passendes Threshold für ab wann zu viele Werte fehlen festzulegen ist schwierig. Es müssen genügend Werte vorhanden sein, damit ein Trend und damit auch eine aussagekräftige Lineare Regression für die Interpolation fehlender Werte möglich ist. Hinzu kommt, dass bei der Linearen Regression nicht alle Werte von der Regression genutzt werden können, da eine Unterteilung in Trainings- und Validierungsdaten stattfindet. Gleichzeitig sollten aber auch nicht zu viele Länder ausgelassen werden, da so interessante Informationen für die spätere Korrelationsanalyse verloren gehen könnten. Wir legen einen Grenzwert von 70% fest, bedeutet wenn für mehr als 70% der Jahre keine Werte verfügbar sind kann das entsprechende Land bei der weiteren Analyse nicht berücksichtigt werden und wird in diesem Schritt entfernt. Bei weit verbreiteten Indikatoren wie dem GDP oder dem Zugang zu Elektrizität fallen jeweils nur 12 und 5 Länder weg, bei der Einschulungsrate für die tertiäre Bildung sind es 72. Der Grenzwert von 70% ist bereits die höchstmögliche Grenze für eine aussagekräftige Regression, bei der Interpretation der Korrelationsergebnisse mit den beiden Indikatoren der Einschulungsrate sollte allerdings der (relativ) höhere Anteil an fehlenden Ländern berücksichtigt werden. Eine Auflistung der entfernten Länder für jeden Indikator ist im GitHub repository zu finden.

Des Weiteren bestehen die Zeilen nicht nur aus Ländern, sondern auch aus Zusammenfassungen (Aggregaten) nach z.B. Regionen oder ökonomischen Unterteilungen. So gibt es beispielsweise das Low income group aggregate (LIC), oder das Middle East and North Africa regional aggregate (MEA). Für diese Datenanalyse sind hauptsächlich Länder und ökonomische Verhältnisse interessant. Zusammenfassungen von Ländern können die Ergebnisse verfälschen, da die Analyse dann zum Teil auf Mittelwerten (oder anderen Gewichtungen) einzelner Bereiche basiert und bestimmte Regionen oder Länder effektiv häufiger vorkommen als andere. Jedoch können die Zusammenfassungen auch Vorteile bringen, da sie in der Regel für alle Jahre Werte verfügen und so fehlende Länder oder ungenaue interpolierte Werte kompensieren können. Basierend auf diese Abwägung wird das world aggregate (WLD), low income, lower middle income und low and middle-income group aggregate (LIC, LMC, LMY), sowie das middle income und upper middle income group aggregate (MIC & UMC) verwendet. Alle anderen Zusammenfassungen werden aus der Tabelle entfernt. Die nun entstandene Tabelle mit den bereinigten Daten wird als neue CSV-Datei gespeichert. Die einzelnen Schritte haben so jeweils eine Art Zwischenergebnis, welches im nächsten Schritt einfach geladen werden kann.

Interpolation von Werten

Wie im letzten Kapitel bereits beschrieben, steht nicht immer für jedes Land und jedes Jahr ein Wert zur Verfügung. Die fehlenden Werte werden durch eine lineare Regression interpoliert. Dafür wird für jeden Indikator aller Länder durchgegangen. Das Ergebnis des Arbeitsschritts wird dann wieder als neue CSV-Datei abgespeichert. Die im letzten Arbeitsschritt bereinigten Daten werden nun verwendet. Zuerst werden die Felder entfernt, die keine Indikatorwerte sind (Landesname & -code). Auch das Feld für das Jahr 2023 wird entfernt, da hier noch kein Wert feststeht (siehe vorheriges Kapitel für Begründung). Zwar könnte auch für das Jahr 2023 ein Wert interpoliert werden. Im Vergleich zu anderen Jahren und Indikatoren gäbe es für dieses Jahr dann ausschließlich interpolierte Werte, die immer eine gewisse Ungenauigkeit aufweisen und somit die Verlässlichkeit der späteren Analysen verringern. Durch Weglassen dieses Jahres geht außerdem kein realer Informationsgehalt verloren, da wie gesagt noch keine realen Indikatorwerte für 2023 vorliegen. Sollten ein Indikatorwert kleiner oder gleich Null sein (kann durch Transformationen die in den nächsten Abschnitten beschrieben werden auftreten), werden diese mit nahe Null (0,0001) ersetzt. Dies ist unter anderem für die Evaluierungsmetriken erforderlich.

Zum Interpolieren wird eine Lineare Regression verwendet. Wie der Name bereits suggeriert, nimmt dieses Verfahren den statistischen Zusammenhang der Einfluss- und Zielgrößen als linear an. Ein lineares Modell (also eine Gerade) wird so angepasst, dass der Fehler der Geraden zu allen Datenpunkten des Trainingsdatensatzes minimal ist. Alle Indikatoren liegen in Form kontinuierliche Variablen vor. Konkret wird die ordinary least squares (OLS) Methode von Scikit-Learn genutzt. Das auf Deutsch als Methode der kleinsten Quadrate bekannte Verfahren minimiert

die quadrierte Differenz zwischen der Regressionsgeraden und den Datenpunkten. Die Regression hat die Annahme, dass es einen (linearen) Trend bei der Entwicklung des jeweiligen Indikators gibt, welcher sich zeitlich fortsetzt. Somit kann der Zeitpunkt (spricht das jeweilige Jahr) als unabhängige Variable und der jeweilige Indikatorwert (beispielsweise das Bruttoinlandsprodukt pro Kopf) als abhängige Variable betrachtet werden.

Um die Qualität der linearen Regression beurteilen zu können, berechnen wir das Bestimmtheitsmaß R^2. Der bestmögliche Wert ist 1.0, was eine perfekte Vorhersage bedeuten würde. Ein Wert von 0.0 oder sogar negativ bedeutet eine imperfekte Vorhersage. Gerade wenn eine lineare Regression auf Daten mit nicht-linearen Zusammenhang angewendet wird, kann der R^2 Score extreme negative Werte (bis zu $-\mathrm{Inf}$) annehmen und eine Auswertungsstatistik verunreinigen. Extrem negative Werte werden daher mit 0.0 ersetzt (Scikit-learn developers, 2023). Zusätzlich wird auch noch der Mean Absolute Percentage Error (MAPE), der Mean Absolute Error (MAE) und der Root Mean Squared Error (RMSE) berechnet. Laut Hyndman und Athanasopoulos (2018) ist letzteres, der mittlere absolute Fehler, ist die Summer der absoluten Fehler aller Datenpunkte geteilt durch die Anzahl aller Datenpunkte. Er besitzt also die gleiche Skalierung wir die zugrunde liegenden Daten, was die Interpretation der Genauigkeit der linearen Regression eines bestimmten Datensatzes einfach macht. Der Root Mean Squares Error, also die Wurzel der mittleren quadratischen Abweichung, besitzt ebenfalls die gleiche Einheit wie der Datensatz. Effektiv führ eine Minimierung des mittleren absoluten Fehlers zu Vorhersagen in Richtung des Medians und eine Minimierung des RMSE zum Mittelwert. Jedoch eignen sich diese Fehlermetriken nicht um die Genauigkeit von mehreren linearen Regressionen, deren Daten unterschiedliche Skalierungen verwenden, zu vergleichen oder zu verallgemeinern. Dafür ist der Mean Absolute Percentage Error (MAPE) besser geeignet, da dieser sensitiv für relative Fehler ist und beispielsweise durch eine globale Skalierung der Zielvariable nicht verändert wird (Scikit-learn developers, 2023). Allerdings hat der Mean Absolute Percentage Error auch einige Nachteile, unter anderem entwickelt er extreme Werte, wenn die zugrunde liegenden Daten sich zu sehr Null annähern (Hyndman und Athanasopoulos, 2018). Daher sollten wir zur abschließenden Qualitätsbeurteilung der Regression die Gesamtheit der Metriken betrachten, nicht nur einzelne. Wie zur Evaluierung von Machine Learning Modellen üblich, teilen wir die verfügbaren Daten in Trainings- und Validierungsdaten (oftmals auch Testdaten genannt). Da es sich bei den Datensätzen um Zeitreihen handelt, besteht der Validierungsdatensatz nicht aus zufällig gewählten Datenpunkten, sondern aus zehn aufeinander folgenden Jahren. Dabei sind die jüngsten Jahre nicht Teil der Validierungsdaten. Wegen der Coronapandemie und weiteren Ereignissen ist der angenommene Trend hier besonders stark verfälscht und der Einfluss des untersuchten Faktors auf die Zielvariable deutlich geringer. Daher sind die letzten zehn Jahre noch Teil der Trainingsdaten und die zehn Jahre davor bilden den Validierungsdatensatz. An dieser Stelle sei noch erwähnt, dass die Gefahr der Überanpassung (overfitting) des Regressionsmodells auf die Trainingsdaten bei der linearen Regression geringer ist als bei anderen Regressionsmethoden. Komplexe Größen wie das Bruttoinlandsprodukt haben sehr viele Einflussfaktoren und folgen in der Realität nur selten einem linearen Verlauf, können aus dem Grund auch nur begrenzt durch eine Regressionsgerade perfekt beschrieben werden. Das ist auch einer der Gründe, weshalb wir keine multiple lineare Regression durchführen: das Modell kann komplexe Zusammenhänge nicht abbilden, der Aufwand wäre unverhältnismäßig. Bei einer ausführlicheren Datenanalyse kann durch eine andere Regressionsmethoden wie k-nearest neighbor oder XGBoost eine höhere Genauigkeit der interpolierten Werte erreicht werden.

Bei Betrachtung der Entwicklung der Indikatoren einiger Länder lässt sich ein exponentielles Wachstum feststellen. Um für Länder mit nicht-linearer (z.B. exponentieller) Werteentwicklung eine lineare Regression mit möglichst hoher Genauigkeit durchzuführen, werden die Daten vorher transformiert. Um ein exponentielles Wachstum in ein lineares Wachstum zu verwandeln, wird ein logarithmische Transformation genutzt. Konkret wird der natürliche Logarithmus, also der Logarithmus zur Basis der Eulerschen Zahl verwendet. Nach der Interpolation wird die Exponentialfunktion der Daten berechnet und somit die logarithmische Transformation umgekehrt.

Natürlich ist es aufwendig & fehleranfällig für jedes Land manuell zu entscheiden, ob ein exponentielles Wachstum vorliegt. Daher führen wir für jedes Land eine lineare Regression durch, einmal mit logarithmischer Transformation und einmal ohne. Als interpolierte Werte werden dann die Werte genutzt, welche aus der Regression mit dem besseren R^2 Score stammen. Dadurch erzielen wir ein besseres Ergebnis bei den Regressionen und erhöhen somit auch die Aussagekraft der folgenden Korrelationsanalysen. Um einen Eindruck über die Gesamtgenauigkeit der Regressionen und damit auch über die Aussagekraft der kommenden Korrelationsanalysen zu bekommen, schauen wir uns eine Übersicht der Ergebnisse der Qualitätsbegutachtung an. Da es unmöglich ist die Ergebnisse für um die 200 Regressionen pro Indikator übersichtlich darzustellen, wird in der Tabelle nur der Median von den Ergebnissen aller Länder ausgewiesen. Die Ergebnisse für den Validierungsdatensatz sind mit „(val)“ gekennzeichnet.

Indikator	R^2	**MAE**	**MAPE**	**RMSE**	**MAE** (val)	**MAPE** (val)	**RMSE** (val)
Zugang Elektrizität	0.645	0.013	0.002	0.017	0.013	0.001	0.016
Geburtsrate	0.829	0.14	0.065	0.165	0.181	0.125	0.191
BIP pro Kopf	0.768	295.263	0.299	414.516	538.312	0.164	697.918
Immunisierung	0.495	0.16	0.037	0.216	0.12	0.025	0.143
Lebenserwartung	0.925	0.028	0.006	0.035	0.038	0.007	0.039
Einschulungsrate Sekundär	0.802	0.211	0.037	0.248	0.208	0.035	0.247
Einschulungsrate Tertiär	0.791	0.755	0.192	0.926	0.993	0.125	1.316

Tabelle 1: Median der Ergebnisse der Evaluierungsmetriken. Quelle: Eigene Darstellung.

Die Regressionen haben durchweg einen gutes R^2 Ergebnis und somit eine hohe Verlässlichkeit. Der Root Mean Squared Error und der Mean Absolute Error sind (wie bereits ausgeführt) in der Einheit des jeweiligen Indikators. Der Median des mittleren absoluten Fehlers der Interpolation für alle Länder des Indikators BIP pro Kopf liegt für die Validierungsdaten also bei 538 tausend US-Dollar. Bei der Lebenserwartung beträgt er ca. 0,04 Jahre (was sehr genau ist) und bei der Geburtenrate liegen wir 0,181 Kinder pro Frau daneben. Der Mean Absolute Percentage Error lässt sich als relative Fehlerangabe zwischen Datensätzen vergleichen. Der Fehler liegt bei Trainings- und Validierungsdaten größtenteils bei um die 10%, was für unsere Bedürfnisse vollkommen akzeptabel ist.

Die nun vollständigen Daten werden in neuen CSV-Dateien gespeichert und stehen so im nächsten Schritt zur Verfügung. Alle Ergebnisse der Evaluierungsmetriken sind ebenfalls in einer CSV-Datei gespeichert und im GitHub repository zu finden, sodass auch im Nachhinein die Genauigkeit der Regressionen für jedes Land von jedem Indikator verfügbar ist.

Korrelationsanalysen

Da wir nun vollständige Indikatorwerte haben, können wir uns wieder den aufgestellten Hypothesen zuwenden. Diese formulieren mögliche Einflüsse auf das Wirtschaftswachstum, welches gesteigert werden soll. Mithilfe von Korrelationsanalysen können wir einen ersten Plausibilitätstest vornehmen.

Zunächst einmal laden wir von zwei Indikatoren, die untersucht werden sollen, die im vorherigen Schritt gespeicherten Daten. Von jedem Indikator können wir nur die Länder verwenden, die bei beiden Indikatoren vorhanden sind. Nur zur Erinnerung: bei der Bereinigung von Daten haben wir Länder, bei denen zu viele Werte gefehlt haben, aus der jeweiligen Tabelle entfernt. Welche Länder dazu gezählt haben, war nicht für alle Indikatoren gleich. Außerdem werden ungültige Werte von der Korrelationsanalyse ausgeschlossen. Ungültige Werte sind fachlich nicht mögliche Werte, wie z.B. ein negatives Bruttoinlandsprodukt. Diese Werte können bei der Interpolation entstehen, wenn die Regressionsgerade ein starkes Wachstum nachbildet und aus diesem Grund eine starke Steigung besitzt. Für Jahre die lange vor diesem Wachstum in der Vergangenheit liegen kann dann ein negativer Wert vorhergesagt werden. Im Kapitel zur Interpolation wurde

bereits besprochen, dass der Indikator für das Bruttoinlandsprodukt pro Person oftmals ein exponentielles Wachstum aufweist. Daher werden wir diesen Indikator nun wieder logarithmisch Transformieren, bevor die Korrelation mit einem anderen Indikator bestimmt wird. Zum Bestimmen der Korrelation wird der Pearsonsche Korrelationskoeffizient (Pearson correlation coefficient, auf Deutsch auch Produkt-Moment-Korrelation genannt) verwendet. Dieser Korrelationskoeffizient kann Werte zwischen -1 und +1 als Ergebnis haben. Ein Ergebnis von 0 bedeutet keine Korrelation, ein Ergebnis von -1 oder +1 bedeutet eine exakte lineare Korrelation, die positiv oder negativ sein kann (The SciPy community, 2023; Bamberg, G., Baur, F. & Krapp, M., 2022). Mögliche Korrelationen lassen sich auch durch ein Streudiagramm grafisch genauer untersuchen. Anhand der Indikatoren BIP pro Kopf und Lebenserwartung ist dies beispielhaft einmal dargestellt:

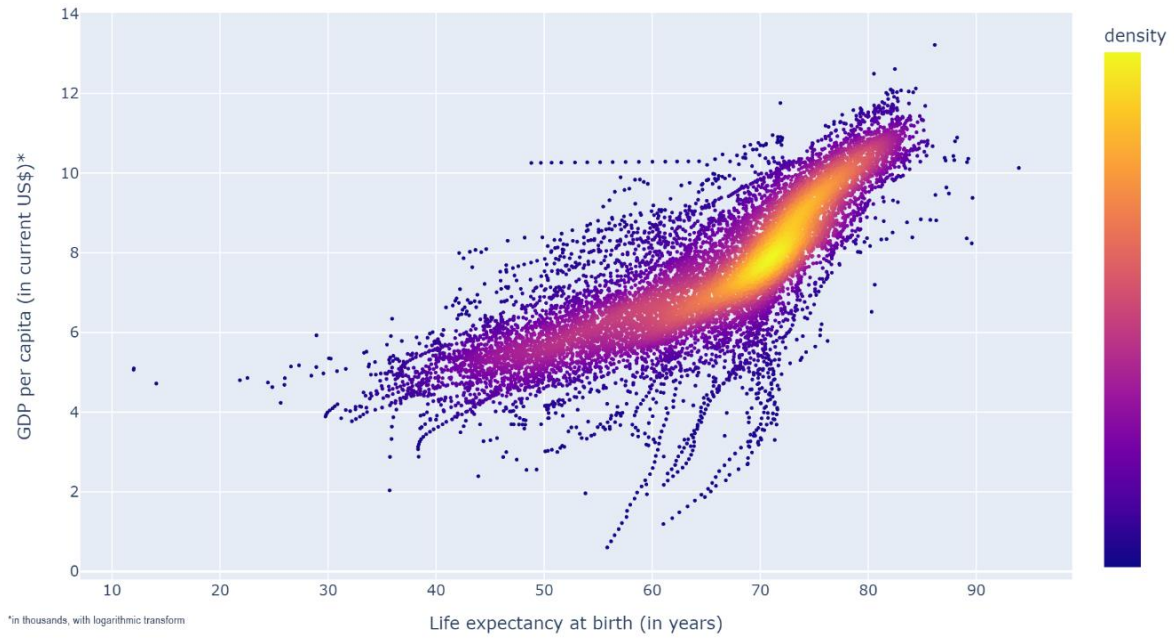

Abbildung 1: Streudiagramm für die beiden Variablen Bruttoinlandsprodukt pro Kopf und Lebenserwartung. Quelle: Eigene Darstellung.

Beginnen wir mit der ersten These: durch eine höhere Bildung der breiten Bevölkerung kann das Wirtschaftswachstum verbessert werden. Das Ziel wäre eine Schulpflicht für weiterführende Bildung. Eine Korrelation zwischen der Einschulungsrate der höheren Bildungsstufen und dem GDP per capita würde diese Hypothese unterstützen. Die Korrelation zwischen dem GDP per capita und der Einschulungsrate für die sekundäre Bildungsstufe beträgt 0.729, was ein starker positiver linearer Zusammenhang ist. Die Korrelation zwischen dem GDP per capita und der Einschulungsrate für die tertiäre Bildungsstufe beträgt 0.747, was ein starker positiver linearer Zusammenhang ist. Wenn ein größerer Teil der Bevölkerung eine Schule mit höherem Bildungsniveau besucht, steigt auch das Bruttoinlandsprodukt pro Kopf – und umgekehrt. Die Korrelationsanalyse unterstützt also die erste Hypothese.

Der zweite Ansatz möchte durch mehr erwerbstätige Frauen und damit mehr verfügbare gesamte Arbeitskraft Wirtschaftswachstum erwirken. Ziel ist die Stellung der Frau im Beruf zu stärken und für mehr Erwerbstätigkeit von Frauen zu werben. Eine Korrelation zwischen einer geringeren Geburtenrate, welche zu einem höheren Anteil an erwerbstätigen Frauen führt, und einem höheren GDP per capita würde diesen Ansatz unterstützen. Die Korrelation zwischen der Geburtenrate und dem GDP per capita beträgt -0.708, was ein starker negativer linearer Zusammenhang ist. Wenn die Geburtenrate sinkt, dementsprechend der Anteil der erwerbstätigen Frauen steigt, steigt ebenfalls das Bruttoinlandsprodukt pro Kopf – und umgekehrt. Die Hypothese wird also ebenfalls von der Korrelationsanalyse unterstützt.

Der dritte Ansatz möchte mit einer Erhöhung der Lebenserwartung, welche an die gesamt verfügbare Arbeitskraft gekoppelt ist, ein Wirtschaftswachstum bewirken. Eine Korrelation zwischen einer höheren Lebenserwartung und einem höheren GDP per capita würde diesen Ansatz unterstützen. Die Korrelation zwischen der Lebenserwartung und dem GDP per capita beträgt 0.787, was ein starker positiver linearer Zusammenhang ist. Wenn die Lebenserwartung steigt, steigt ebenfalls das Bruttoinlandsprodukt pro Kopf – und umgekehrt. Auch bei dem dritten Ansatz unterstützt die Korrelationsanalyse die Hypothese. Um die Lebenserwartung zu steigern werden eine Erhöhung des Anteils an DPT-Immunisierungen von Kindern und ein verbreiteter Zugang zu Elektrizität. Die Korrelation zwischen der Lebenserwartung und der DPT-Immunisierung beträgt 0.713, die Korrelation zwischen der Lebenserwartung und dem Zugang zu Elektrizität liegt bei 0.764. Beides sind starke positive Zusammenhänge. Aus diesem Grund scheinen die vorgeschlagenen Maßnahmen geeignet, um die Lebenserwartung zu steigern.

Um das Potential der Ansätze beurteilen zu können sind Korrelationsanalysen hilfreich. Sie haben gezeigt, dass bei allen drei Ansätzen ein mindestens starker Zusammenhang zwischen den relevanten Faktoren besteht. Allerdings ist es bei der Interpretation wichtig, sich in Erinnerung zu rufen, was Korrelationsanalysen nicht zeigen. So kann die Hypothese über die Beeinflussung der Variablen untereinander, also welche Variable die unabhängige und welche die abhängige ist, nicht bewiesen werden. Dies lässt sich anhand der letzten Hypothese verdeutlichen: hier wird davon ausgegangen, dass eine höhere Lebenserwartung zu mehr Wirtschaftswachstum führt. Allerdings kann die Kausalreihenfolge auch umgekehrt sein, also das aufgrund von Wirtschaftswachstum und den dadurch gestiegenen Wohlstand sich die Gesundheitsversorgung verbessert hat, was zu einer höheren Lebenserwartung führt. Ähnliche Hypothesen, bei denen die Kausalreihenfolge umgekehrt ist, lassen sich auch für die anderen Ansätze aufstellen. So kann ein höherer Wohlstand der breiten Bevölkerung die Einschulungsrate erhöhen, da mehr Familien die Bildung ihrer Kinder finanzieren können und weniger Zwang besteht, so früh wie möglich mit der Erwerbsarbeit zu starten.

Zusammenfassung

Die Kampagnenansätze der Organisation WealthGrowth wurden in einer ersten Datenanalysen untersucht. Basierend auf bereinigten und interpolierten Daten der Weltbank wurde mithilfe von Korrelationsanalysen der Einfluss von vorgeschlagenen Faktoren auf das Wirtschaftswachstum untersucht.

Eine höhere Lebenserwartung korreliert am stärksten mit gesteigertem Wirtschaftswachstum. Historisch betrachtet besteht eine starke Korrelation zwischen einem größeren Anteil an DPT-Immunisierungen bei Kindern, sowie dem Zugang zu Elektrizität für einen größeren Anteil der Bevölkerung und einer längeren Lebenserwartung.

Eine höhere Bildung der breiten Bevölkerung weist die nächsthöchste Korrelation auf. Besonders effektiv ist die tertiäre Bildungsstufe, aber auch bei der sekundären Bildungsstufe ist die Korrelation stark. Bei der Analyse wurden die Indikatoren der jeweiligen Einschulungsrate verwendet, da diese die vorgeschlagene Maßnahme der Schulpflicht am besten beschreiben.

Die Stellung der Frau im Beruf stärken ist ebenfalls eine geeignete Maßnahme um das Wirtschaftswachstum anzukurbeln, die Korrelation ist hier auch stark.

Die Schlussfolgerungen aus Korrelationsanalysen sind mit Vorsicht zu genießen, da die keine Kausalzusammenhänge beweisen können. Die Richtigkeit dieser Datenanalyse kann durch verbesserte Regressionsmethoden bei der Interpolation weiter verbessert werden.

Literaturverzeichnis

International Labour Organization. (1998). *Guidelines concerning treatment in employment and unemployment statistics of persons on extended absences from work, endorsed by the Sixteenth International Conference of Labour Statisticians.* *https://www.ilo.org/public/english/bureau/stat/download/guidelines/exleave.pdf*

Scikit-learn developers. (2023). *sklearn.metrics.r2_score.* https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html#sklearn.metrics.r2_score

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice.* OTexts.

Scikit-learn developers. (2023). *sklearn.metrics.mean_absolute_percentage_error.* https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_percentage_error.html#sklearn.metrics.mean_absolute_percentage_error

The SciPy community. (2023). *scipy.stats.pearsonr* https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html

Bamberg, G., Baur, F. & Krapp, M. (2022). *Statistik (19. Aufl.)*. De Gruyter Oldenbourg.