

データサイエンスをはじめましょう
- Data Science for All -

鈴木寛 (Hiroshi Suzuki)

2023-09-11

目次

この文書について

データサイエンスを始めましょう。

データサイエンスは、広い意味をもったことばで、一口に、学び始めると言っても、さまざまな始め方があると思います。本書では、そのひとつを提案するとともに、共に学んでいきたいと願って、書き始めました。

いろいろな方々や、利用方法を想定して、導入のような内容や、本書の中心をなす、Rでのプログラミングを利用しなくても、ダッシュボードなどを利用することにより、データをさまざまな角度から見る経験をする、スクリプトや、テンプレートを使って、一部を置き換えることで、データサイエンスを経験すること、基本的な R でのプログラミングを学んで、自分で、簡単なプログラムを書いて、分析をすること、さらに、こんなときは、どうしたら良いかなど、少し詳しい説明などを含んだ部分などです。

また、順次、例を提供する、ブログのようなものも書いていきたいと思います。この書を利用するために「データサイエンスを教えてくださいませんか」と、教える方のサポートも、書いていく予定です。

みなさんも一緒にデータサイエンスを学んでみませんか。

著者について

著者は、大学の学生の時以来、数学を学び、大学で教え、2019 年春に退職。それ以来、少しずつ、データサイエンスを学んでいます。

幸運にも、2019 年 9 月の日本数学会教育委員会主催教育シンポジウムで、「文理共通して行う数理・データサイエンス教育」という題で、話す機会が与えられ、その後、あることが契機となり、2020 年度から、毎年、冬学期（12 月から 2 月）に、大学院一般向け（分野の指定なし）の授業、「研究者のためのデータ分析（Data Analysis for Researchers）」を担当しています。複数の教員で担当しますが、基本的な部分は、わたしが教え、応用について、他の先生がおしえています。受講生は 20 人程度で、殆どが、外国人。それも、多国籍で、多くても一国から三人程度。英語で教えています。

この授業の関係で親しくなった経済学の先生に依頼されて、いくつかのレベルの経済学の授業で、2・3 時限、世界銀行の世界開発指標などを使った分析の入門を、特別講師として教えさせていただきました。これらのコースは日本語で教えましたが、日本語で書かれ

た一般の学生向けのデータサイエンスの教科書、学習教材が不足していることも感じました。

2023 年 3 月に京都大学数理解析研究所で開かれた数学教育の会で、話す機会が与えられ、大学などで全学生向けに提供する「数理・データサイエンス・AI」について講演し、生成系 AI の進化も踏まえて、どのようなことを学ぶ機会とすれば良いかについても、考えました。

「データサイエンスをはじめましょう」も、このような背景の下で書かれたものです。

コースで利用したものや、講演記録ににご興味のある方は、このページの下にもリンクがある、著者のホームページを参照してください。

コンピュータ言語について

データサイエンスには、コンピュータを使います。コンピュータに指示をして、データを使いやすく変形したり、計算をしてもらったり、グラフを書いてもらったりします。そのためにコンピュータとのやりとりをする言語が必要です。さまざまな言語（プログラミング言語と呼びます）が、使われますが、データサイエンスで、一番、使われているのは、Python と、R です。

「データサイエンスをはじめましょう」では、統計解析のために開発された R を使います。いずれは、python についても触れたいと思いますが、プログラミングの経験がない方も含めて、最初にデータサイエンスを学ぶには、R は最適だと考えています。また現在は、R Studio IDE (integrated development environment, 統合開発環境) で、R を簡単に使うことができます。さらに、簡単なプログラムであれば、Posit Cloud で試したり、共有することも可能です。また、再現性 (Reproducibility) や、なにを実行しているのかの説明を同時に記述すること (Literate Programming) は、非常に重要ですが、その記述も、R Markdown や、Quarto によって、可能になっています。これが、Excel や、Google Spread Sheet ではなく、R を推奨する理由でもあります。この文書も、R Markdown の一つの形式の、bookdown を利用しています。最後に、Bookdown に関連して、膨大な数の、参考書も、無償で提供されており、オンラインで読むことができることも、R をお薦めする理由です。

ただし、日本語の学習教材は、まだ十分とは言えない状況です。この文書を書き始めたのも、すこしでも、お役に立つことができればとの、気持ちが背景にあります。

もう一つ付け加えると、高校の教科書でも、一部、Python のスクリプトが使われ、大学で、全学生向けに、データサイエンスのコースを提供するときにも、Excel を使うか、Python を使うかが R と比較するとより一般的のように見えます。しかし、それは、教える側の都合が影響しているのではないのでしょうか。急いでコースを提供するため、身近な Excel から始めることにしたり、データサイエンスなら、情報科学の教員がコースをデザインせざるを得ないことから、汎用性が高く、かつデータサイエンスでも、中心的な役割を果たす Python を選択すると言ったことです。

しかし、すべての人が、データサイエンスを学ぶ必要があるならば、文系の先生、そして、純粋数学を生業（なりわい）としてきたわたしのようなものも含めて、教えることに関わることは、とても大切だと思います。教員も学ぶ必要があるからです。R は、社会科学系の分野で、最近特に使われています。分野ごとには、SPSS や、Stata や、MATLAB などが使われているかもしれませんが、これらは、有償で、大学では、使えたとしても、一般の学生が卒業後も、これらのソフトを使えることは稀です。また、研究においても、最近では、さまざまな分野で、R がより一般的になっているとも聞きます。

百聞は一見にしかずで、例を見ていただくのがよいと思いますが、R では、非常に短い命令で、対話的に、すぐ結果が得られます。テンプレートを使って、一部だけ書き換えて使うことも可能で、その人のレベルにあった利用が可能だと思います。さらに、Shiny のような、ダッシュボード形式のものも、利用可能ですし、learnr のように、対話型の、練習問題を提供することも可能です。

少しずつ紹介していきたいと思います。

言語について

ご覧の通り、本書は、日本語で書かれています。用語は、英語、あるいは、英語を追記、または、英語をカタカナにしただけのものを使用する可能性が大きいですが、説明は、極力、日本語で書いていく予定です。

しかし、基本的に、コード（プログラムの記述）には、日本語を使わないで書いていく予定です。とくに、初心者にとっては、二バイト文字で表現される日本語を含むコードの扱いは、負担になることが多いからです。最近では、コードの中で日本語を使用しても、ほとんど、問題は起きないようになっています。そうであっても、世界の人の共通言語として、プログラム言語を学んでいくときには、日本語を使わないことは意義があると思います。日本語を使わないことで、世界中の人たちから、アイディアを学び、あるときは協力し、例を提供して、作業をしていくことが可能だからです。

少し慣れてきて、日本語のデータなどを扱うときには、コードにも日本語を使う必要が生じてきますから、日本語の利用についても、追って説明していきます。APPENDIX ?? を参照してください。

最初は、みなさんも、変数（variable）や、オブジェクト（object）に名前をつけるときは、半角英数を使い、日本語は、使わないようにすることをお勧めします。自分には難しいと感じる時は、ローマ字を半角英数で使うのはいかがでしょうか。

PDF、ePub 版について

この文書は、PDF 版と、ePub 版も作成しています。しかし、扱いが異なるので、ある程度完成するまでは、あまり更新しない予定です。いずれ、これらも、更新したものを公開できると良いのですが。試験公開版は、下のリンクにあります。

- PDF 版
- ePub 版

参考

この電子書籍以外にも、データ・サイエンスについて幾つかの文書をインターネット上に公開しています。わたしのホームページにリストしてありますので、ご興味のあるかたは、参考にしてください。

- データサイエンスを学びませんか・データサイエンス教育

第1章

はじめに

1.1 データサイエンスとは

データサイエンスとはどのようなものでしょうか。いくつかの定義を紹介しますが、新しい分野で、非常に広い範囲の人たちが、データサイエンスに関係していることから、誰でもが受け入れられる定義することは、難しいように思います。

簡単に表現すると「データを活用するための科学」かなと、わたしは、考えています。皆さんが、これから、データサイエンスを学びながら、自分だったらどのように表現するか、考えてください。

「データを活用するための科学」には、三つのことが含まれています。「データ」「活用するため」「科学」。一つ一つ、厳密な定義は難しいですが、データは、これから、皆さんがたくさん出会いますので、それまで置いておきましょう。「活用するため」と書きましたが、課題をみつけたり、その解決のための意思決定の根拠をさがしたりということでしょうか。分野は、さまざまですから、表現もまちまちかもしれません。最後に「科学」これも、簡単ではありませんが、わたしは、二つのことが大切だと考えています。一つは、反証可能性 (falsifiability^{*1}) です。もう一つは、再現可能性 (reproducibility^{*2}) です。反証可能性は、自分はそう信じるというような主観ではなく、正しいかどうかチェックすることができるという意味です。再現可能性は大体理解できると思いますが、データから得られる事実を、他のひとが操作をしても、同じ結果が得られるということです。そのためには、その「操作」が明確になっており、合理性も確保されていなければいけませんね。

これら二つの性質は、「活用するため」に、共通の客観的な基盤を保証するということだと思います。データサイエンスは、根拠を明確にした議論 (evidence based, fact based) のために、欠かせないことで、多様な価値観の多様なひとたちが合意して課題に立ち向かっていくためには、必要不可欠なことだと思います。そのいみでも、これから、データサイエンスは、ますます、重要度がましくしていくと思います。

^{*1} カール・ポパー (Karl Raimund Popper, 1902-1994) が科学論において提唱した考え方]

^{*2} 科学はすべて再現可能とは言えません。宇宙の成り立ちや、生物進化などは、再現可能ではありませんが、事実を確認しそれを根拠とした議論を考えると、可能な限り確保することが、反証可能性にもつながります。

もう少し、踏み込んで、データサイエンスについて書いてみましょう。実際に、どのような作業が重視されているのかも、含まれています。

Data Science: データ (Data) を活用して課題を発見・探求し、適切な解決策を探る意思決定のための科学 (Decision Science) で、エンピリカル (Empirical Study) すなわち、理論ではなく、実証性を特徴とする。データから得られる特徴を表示するとともに、数理モデルを適用し・機械学習などで評価し・アルゴリズムを策定する数理的思考を通して得られた結果を、可視化などによってコミュニケーションをおこない、共有し、他者の意見を聞き理解する努力をしながら、さらに課題について、あらたにデータを活用して考え、検証し、適切な解決策がもたらす新たな課題も予測しながら、調整をはかること。

上に書いたのは、ひとつの表現ですが、AI*³ に、表題の「データサイエンスとは」と聞いてみると、次のような答えをしてくれました。

ChatGPT (2023.6.3)

データサイエンス (Data Science) は、データを解析し、洞察を得るための科学的な手法やツールの開発と応用を含む、多様なスキルセットを組み合わせた領域です。データサイエンスは、統計学、機械学習、データマイニング、データ可視化などの技術や手法を利用して、大量のデータからパターンやトレンドを抽出し、意思決定に役立つ洞察を導き出すことを目指します。

データサイエンスのプロセスは、データの収集、前処理、分析、モデリング、評価、展開などのステップで構成されます。最初に、関連するデータを収集し、必要な形式に整理するために前処理を行います。次に、統計的な手法や機械学習アルゴリズムを使用してデータを分析し、予測モデルやパターンを発見します。モデルのパフォーマンスを評価し、必要に応じて改善を行い、最終的に洞察をビジネスや科学の現場に適用することが目標です。

データサイエンスは、さまざまな分野で活用されています。例えば、ビジネス分野では、マーケティングキャンペーンの最適化、顧客セグメンテーション、需要予測などに利用されます。医療分野では、病気の診断や治療計画の最適化、新薬の開発などに応用されます。また、交通や都市計画、環境保護、金融など、さまざまな領域でデータサイエンスの手法が活用されています。

データサイエンスには、数学や統計学、プログラミング、データベース管理、機械学習、ビジュアル化などの幅広いスキルが必要です。また、データの解釈やビジネス理解、コミュニケーション能力も重要な要素です。データサイエンティストやデータアナリストは、これらのスキルを駆使して、データを活用し、意思決定のサポートや新たな知見の発見に貢献します。

ちょっと日本語がぎこちない感じもするので、英語での問いと答えも書いておきましょう。

*3 どんどん進化していますから、答えを得た日付も加えておきます。一般的なことがらについてまとめて答えてくれるので、その意味で便利です。

What is datascience?

Data science is a multidisciplinary field that involves using scientific methods and tools to analyze data and extract insights. It encompasses a wide range of skills, including statistics, machine learning, data mining, and data visualization, to explore and understand complex data sets.

At its core, data science involves collecting, organizing, cleaning, and processing large volumes of data to uncover patterns, trends, and correlations. It applies various statistical techniques and algorithms to gain meaningful insights and make informed decisions. Data scientists employ exploratory data analysis, hypothesis testing, and predictive modeling to extract valuable information from data and generate actionable recommendations.

Data science finds applications in numerous industries and domains. It is widely used in business for market research, customer segmentation, demand forecasting, and personalized recommendations. In healthcare, data science aids in disease diagnosis, treatment optimization, and drug discovery. Other areas where data science is applied include transportation and urban planning, environmental conservation, finance, and social sciences.

To excel in data science, one needs a solid foundation in mathematics, statistics, programming, and data manipulation. Proficiency in programming languages such as Python or R is crucial for data manipulation and building models. Strong analytical skills, critical thinking, and the ability to communicate findings effectively are also essential for data scientists.

Data science is a rapidly evolving field, driven by advancements in technology and the increasing availability of data. It plays a crucial role in leveraging data assets to gain insights and make data-driven decisions in various industries.

学際的分野（multidisciplinary field）という言葉も使われています。かなりの広がりをもった分野です。

1.1.1 なぜ、データサイエンスが最近注目されているのか

では、なぜ、最近、急に、データサイエンスが注目されるようになったのでしょうか。最初にも少し書きましたし、みなさんが、少しずつ学びながら、考えていただくのが一番よいと思いますが、いくつか理由があると思います。

個人的には、インターネットの普及や、さまざまなセンサーの普及にも助けられて、膨大なデータが得られるようになり、さらに、コンピュータの進化により、大きなデータであっても、それらを、扱うことができるようになったこと。それによって、複雑な問題であっても、その解決と意思決定に関して、データを使って根拠を明確にした（evidence based）議論を支援することが可能になったことだと考えています。

付け加えておかないといけないと考えていることがもう一つあります。それは、問題解決と書きましたが、課題と向き合うという表現でもよいと思います。これまでは、因果関係が単純な課題については、根拠を示して、ある程度議論できましたが、実際の世の中の問題は、非常に多くの要素が相互に絡み合っており、それらを、調整すると、結果にも、さまざまな変化が生じるため、調整も非常に困難だということも挙げられると思います。

それだけ複雑な課題と向き合うには、データを丁寧にみる必要があると、その中で、さまざまな指標に関して評価値が高いものを選択することが可能になって来たということです。

そのことが、もう一歩進んで、AI（Artificial Intelligence 人工知能）を支える技術に使われています。ここで使った、ChatGPT の GPT は、Generative Pre-trained Transformer の略ですが、LLM（Large Language Model）によって、自然なことを生成できるよう技術が実用段階に入っています。たとえば、「データサイエンスが最近注目されているのは...」とすると、この後に続く言葉として、評価値が高いものを、選択してくれる。それを、さらに、調整して、応答することが可能になっているのです。

少し長いですが、もう少し、踏み込んで説明していますので、ChatGPT の応答を下に挙げておきましょう。

ChatGPT (2023.6.3)

データサイエンスが最近注目されている理由は、以下の要素によるものです。

1. ビッグデータの増加: 近年、デジタル技術の進歩により、膨大な量のデータが生成されるようになりました。ウェブサイトのアクセスログ、ソーシャルメディアの投稿、センサーデータ、トランザクションデータなど、多様なソースから大量のデータが利用可能です。データサイエンスは、これらのビッグデータを活用して洞察を得ることができるため、需要が高まっています。
2. 技術の進歩: データサイエンスを支える技術の進歩も、その注目を高めています。クラウドコンピューティングの発展やオープンソースのツールとライブラリの普及により、データの収集、処理、分析が容易になりました。さらに、機械学習や深層学習のアルゴリズムの改善や高速化も進んでおり、より高度な予測モデルやパターンの抽出が可能になっています。
3. ビジネス価値の認識: データは企業にとって貴重な資産であり、データを活用することで競争優位を得ることができます。データサイエンスを活用することで、マーケットのニーズやトレンドを理解し、効果的なマーケティング戦略を立てることができます。顧客行動の予測やセグメンテーションを行い、顧客満足度や収益性を向上させることも可能です。
4. 知識と人材の需要: データサイエンスの分野での知識とスキルを持つ人材への需要が高まっています。企業や組織は、データサイエンスの専門家を雇用して、データの分析と活用を行いたいと考えています。また、データサイエンスは学問としても注目を浴びており、大学や研究機関ではデータサイエンスに関連す

るプログラムや研究が増えています。

これらの要素により、データサイエンスはビジネスや科学の分野で重要な役割を果たしており、その注目度が高まっています。以下に、その具体的な理由をいくつか挙げます。

5. 創新と競争力の追求: データサイエンスは新たなビジネスモデルや製品の開発に貢献します。データを活用して市場のトレンドや消費者のニーズを洞察し、それに基づいて新しいサービスや製品を開発することが可能です。企業はデータを駆使して競争力を高め、革新的なソリューションを提供することを求めています。
6. 問題解決と意思決定の支援: データサイエンスは複雑な問題を解決し、意思決定をサポートする上で重要なツールとなっています。データの分析により、事実に基づいた洞察が得られ、ビジネス上の課題や機会を特定することができます。データサイエンスの手法はリスク評価や予測、最適化などにも応用され、効果的な戦略立案や意思決定を支援します。
7. パーソナライズされた体験の提供: データサイエンスは顧客の行動や好みを理解し、パーソナライズされた体験を提供するための手段としても利用されます。顧客データや行動履歴を分析し、個別のニーズや関心に基づいたオファーや推奨を行うことで、顧客満足度や忠誠度を向上させることができます。
8. 社会的なインパクトと公共政策への応用: データサイエンスは社会的な課題や公共政策の分析にも役立ちます。例えば、交通渋滞の予測や都市計画の最適化、犯罪予測や防犯対策の強化などにデータサイエンスが活用されています。データ駆動のアプローチは社会のさまざまな側面において効果的な解決策を提供する可能性があります。

これらの要素により、データサイエンスはビジネスや科学の分野で重要な役割を果たしており、その注目度が高まっています。以下に、その具体的な理由をいくつか挙げます。

1. 自動化と効率化の追求: データサイエンスは、業務プロセスの自動化や効率化にも貢献します。データの分析と予測モデルの活用により、タスクの自動化や最適なリソースの割り当てが可能となります。これにより、人的な作業負担を軽減し、生産性を向上させることができます。
2. 組織内のデータ文化の普及: データサイエンスの注目度の高まりに伴い、企業や組織においてデータ文化の重要性が認識されています。データサイエンスの手法やアプローチを組織全体で活用し、データに基づいた意思決定や問題解決を行うことが求められています。組織内でのデータの収集、管理、分析の能力向上に取り組むことで、競争力を高めることができます。
3. 専門化と職業の多様化: データサイエンスの注目度の高まりに伴い、関連する職業や専門化も増えています。データサイエンティスト、データアナリスト、

機械学習エンジニアなどの役割が求められるようになりました。また、データサイエンスを支えるツールやプラットフォームの開発も進んでおり、それに関連する職業やスキルも増加しています。

総じて、データサイエンスの注目度が高まっている理由は、データの増加、技術の進歩、ビジネス価値の認識、創新と競争力の追求、問題解決と意思決定の支援、パーソナライズされた体験の提供、社会的なインパクトと公共政策への応用、自動化と効率化の追求、組織内のデータ文化の普及、専門化と職業の多様化など、さまざまな要素が重なっています。

データサイエンスは、ビジネスや科学の領域でデータの価値を最大限に引き出すための力強いツールとなっており、今後もその重要性和注目度は高まっていくことが予想されます。

1.1.2 なぜ、一人ひとりが、データサイエンスを、学ぶ必要があるのでしょうか

専門家に任せてはいけないのでしょうか。

個人的には、次のように考えています。非常に広い範囲の課題に関係している課題に、データサイエンスが使われるだけでなく、その背後にある、単純な、因果関係ではない、考え方を、理解する必要があることが挙げられると思います。

さらに、わたしが特に大切だと考えているのは、個人のコンピュータでも、十分に利用可能であること、オープンデータとか、パブリックデータと言われる、誰でも利用できるデータが、使いやすい形で、提供されているということです。

ChatGPT (2023.6.3)

一人ひとりがデータサイエンスを学ぶ必要がある理由は、以下のような点が挙げられます。

1. データリテラシーの重要性: データは現代社会においてますます重要な役割を果たしています。データに基づく意思決定や問題解決は、個人の日常生活から組織や社会の活動まで広がっています。データサイエンスを学ぶことで、データを扱うスキルや知識を身につけることができます。データを理解し、適切に分析して活用する能力は、個人の能力向上や自己啓発にも繋がります。
2. 個人の成長とキャリアの展望: データサイエンスは求人市場で高い需要があり、競争力のあるキャリアの一つとなっています。データサイエンスのスキルを持つことで、自身のキャリアの成長や就業機会の拡大が期待できます。データサイエンスの知識やスキルは、さまざまな職業や業界で役立つだけでなく、起業や独立の道も開ける可能性があります。
3. 自己啓発と問題解決能力の向上: データサイエンスは、問題解決や意思決定において非常に有用なツールです。データサイエンスを学ぶことで、情報を分析

し洞察を得る力や、パターンを見つける力、予測や最適化の能力を養うことができます。これにより、個人の問題解決能力や意思決定力が向上し、より効果的な選択や行動ができるようになります。

専門家に任せることも重要ですが、一人ひとりがデータサイエンスの基本的な概念や手法を理解することは有益です。以下にその理由を説明します。

4. データの信頼性と誤解の回避: データは私たちの日常生活やビジネスの意思決定に影響を与える重要な要素です。しかし、データの解釈や分析方法によっては誤解を招くこともあります。データサイエンスの知識を持つことで、データの信頼性を評価し、適切な解釈を行う能力が向上します。これにより、誤った結論や誤解に基づく意思決定を避けることができます。
5. 自己管理とデータプライバシーの保護: 近年、個人のデータの重要性和プライバシーの問題が注目されています。データサイエンスを学ぶことで、自身のデータの管理方法やプライバシーの保護について理解し、適切な対策を取ることができます。データの収集や利用に対して主体的に関与し、自己管理の意識を持つことが重要です。
6. クリティカルシンキングと問題解決能力の向上: データサイエンスの学習は、クリティカルシンキングや問題解決能力を向上させる助けとなります。データを扱い、分析するためには、情報を適切に評価し、論理的に考える必要があります。データサイエンスの学習は、証拠に基づく意思決定や論理的な議論のスキルを磨くのに役立ちます。

データサイエンスの専門家に相談することは重要ですが、一人ひとりが基本的なデータサイエンスの知識やスキルを身につけることで、自己啓発や問題解決能力の向上、データの信頼性の評価、プライバシーの保護などが可能となります。データサイエンスの学習は、個人の成長やキャリアの展望にもプラスの影響を与えることが期待されます。

1.1.3 AI の活用

「データサイエンスをはじめましょう」では、AI も活用していきたいと思います。クリティカルシンキングのことも、書かれていますが、どうしても、人は、バイアスのかかった考えかたをしてしまいます。それは、経験も知識も限られていますし、単純な答えを受け入れたい傾向^{*4}ということもあるかもしれません。

そのために、まず、仲間に聞いてみる。そして、AI に聞いてみる。それらを踏まえて、さらに、課題や、問いを精緻なものにしていく。それが、まず、第一に大切だと考えるからです。

さらに、最近の AI は、人間の普通の言葉を理解して、コード（コンピュータに伝える言

^{*4} 私はこれを、単純化バイアスと呼んでいます。遺伝子に組み込まれているのかもしれませんが。しかし、それに抗うのも、人間としてのたいせつな生き方のように思います。

葉)を書くこともできるようになっています。そのような支援を有効に使うことで、プログラミングの訓練を受けていない人でも、データサイエンスを学ぶことができるようになって来ています。

翻訳にも活用することが可能で、長い文章の要約もかなり上手にしてくれるようになってきます。

AIはまだ、発展途上で、ここに書いたことも、少し経つと、古くなってしまふことでしょう。しかし、少し長いスパンで考えると、AI研究の目標値として、掲げられているのは、AGI (Artificial General Intelligence)、汎用人工知能。(人間が実現可能な)あらゆる知的作業を理解・学習・実行することができる人工知能ということです。しかし、課題も多くあります。課題について、AIに聞いてみると次のような回答をしてくれました。

Q. 現在の General Artificial Intelligence の課題をまとめてください。

GPT4 (2023.8.8)

現在の一般人工知能 (General Artificial Intelligence, AGI) には、以下のような主要な課題があります。

1. 理解と推論の限界: 現在の AI は、人間のように世界を理解したり、複雑な推論を行う能力をまだ持っていません。これは、言語理解、感情理解、一般的な常識理解など、多くの問題領域に影響を与えます。
2. 教育データのバイアス: AI は学習データからパターンを学びますが、そのデータが偏っている場合、AI も偏った結果を出す可能性があります。これは、AI が人間の先入観や社会的な偏見を反映する可能性を示しています。
3. 説明可能性と透明性の欠如: AI の決定過程はしばしば「ブラックボックス」とされ、その理由は明らかにされていません。これは、AI の決定がなぜ行われたのか、どのような情報に基づいて行われたのかを理解するのを難しくしています。
4. 安全性とエチカルな問題: AI が自己学習や自己改良を行う能力が増すにつれて、その結果が予測不可能になる可能性があります。また、AI が人間の倫理観や法律を理解し、それに従うようにすることも大きな課題です。
5. リソースの要求: 大規模な AI モデルの訓練は、大量の計算リソースと電力を必要とします。これは環境に対する影響や、AI 開発の費用効率に関わる問題です。

たしかに、どれも、難しそうな課題ですね。みなさんは、これから、どのように、AI を活用し、どのように、向き合われるのでしょうか。

1.2 Gapminder

Gapminder はオラロスリング (Ola Rosling) とアンナロスリング (Anna Rosling Rönnlund) とハンスロスリング (Hans Rosling) が設立した組織です。

ハンス・ロスリングは、FACTFULNESS (ファクトフルネス) 10 の思い込みを乗り越え、データを基に世界を正しく見る習慣の著者です。

Gapminder: <https://www.gapminder.org>

一番上にテストあります。- もしかしたら、あなたの世界観は、間違っているかもしれません。(You are probably wrong about - upgrade your worldview)

Bubble Chart: [https://www.gapminder.org/tools/#\\$chart-type=bubbles&url=v1](https://www.gapminder.org/tools/#$chart-type=bubbles&url=v1)

Dollar Street: <https://www.gapminder.org/dollar-street>

Data: <https://www.gapminder.org/data/>

1.2.1 Factfulness (ファクトフルネス) Hans Rosling

実際のデータから、現実を見ていないと、非常に歪んだ世界観に毒されているかもしれません。ハンス・ロスリングは、公衆衛生が専門のスウェーデンの医師で、アフリカでも年間も働いた経験ももっている方で、ひとは、さまざまなバイアス (先入観) により、間違った根拠のもとで世界を見ているといい、統計的な資料から、息子のオラと協力して、非常にわかりやすい、バブルチャートでの時代の移り変わりを表現し、サイトに載せています。また、それだけでは、実際の生活が見えにくいという理由から、息子の奥さんのアンナさんの提案で、ダラー・ストリートというプロジェクトもしています。この程度の経済状態の暮らしがどのようなものかを、写真などで、紹介するものです。

先生から教わった知識は、先生が勉強した時代にはある程度正しかったかもしれないが、世界は変化している。自分の経験から、判断すると、その先入観から、多くの間違いを犯すというような指摘もしています。

1. 分断本能を抑えるには、大半の人がどこにいるかを探そう。
2. ネガティブ本能を抑えるには、悪いニュースのほうが広まりやすいことを覚えておこう。
3. 直線本能を抑えるには、直線もいつかは曲がることを知ろう、
4. 恐怖本能を抑えるには、リスクを計算しよう。
5. 過大視本能を抑えるには、数字を比較しよう。
6. パターン化本能を抑えるには、分類を使おう。

7. 宿命本能を抑えるには、ゆっくりとした変化でも変化していることを心に留めよう。
8. 単純化本能を抑えるには、ひとつの知識がすべてに応用できないことを覚えておこう。
9. 犯人捜し本能を抑えるには、誰かを責めても、問題は解決しないと肝に銘じよう。
10. 焦り本能を抑えるには、小さな一歩を重ねよう。

(ファクトフルネスから)

データサイエンスはこれらのいくつかを克服するひとつの方法であるように見えます。

ハンス・ロスリングは、すでに 2017 年 2 月 7 日に亡くなっていますが、心配している 5 つのリスクとして、挙げているのは、感染症の世界的な流行、金融危機、世界大戦、地球温暖化、極度の貧困。どれも示唆に富んでいるように見えます。

Youtube にたくさん、ビデオも出ていますが、二つだけリンクを載せておきます。

- How not to be ignorant about the world | Hans and Ola Rosling
– バイアスを意識しながら (日本語の字幕がついています)
- The best stats you've ever seen, Hans Rosling

1.2.2 参考文献

- 「私はこうして世界を理解できるようになった」ハンス・ロスリング、ファニー・ヘルエスタム著、青土社 (ISBN978-4-7917-7217-9, 2019.10.10) “How I learned to understand the world” by Hans Rosling with Fanny Haergestam の翻訳
- 「ファクトフルネス - 10 の思い込みを乗り越え、データを基に世界を正しく見る習慣」ハンス・ロスリング、オーラ・ロスリング、アンナ・ロスリング・ロランド著、日経 BP 社 (ISBN978-4-8222-8960-7, 2019.1.15)

1.3 「データサイエンスを始めましょう」の特徴

1.3.1 学習者として想定しているのは

高等学校を卒業したひとたちを対象と想定して、書いていこうと思います。

1.3.2 オープン・パブリックデータの活用

データサイエンスは、広い分野ですが、ここでは、オープンデータとか、パブリックデータと言われるものを主として活用していきます。

すでに、書いたように、インターネットに繋がったコンピュータがあれば、だれでも、データサイエンスを学び、データから智を得ることが可能になって来ています。特に、世界

に目をむけると、すばらしいサイトがたくさんあり、国際機関などが、膨大なデータを提供しているので、まずは、それを活用したいと思います。

1.3.3 世界のデータをみること

日本のデータも、使っていこうと思いますが、まずは、世界の中の課題をみる必要があります。扱いやすい、世界のデータがたくさんあることも理由の一つです。

むろん、日本の課題から目を逸（そ）らすわけではありません。世界の中の日本を意識し、日本の課題にも目を向けていきたいと思っています。

1.3.4 目標としていること

ここで扱う内容は限られていますが、データサイエンスの基本を身につけることで、ここで、取り上げる、オープンデータ、パブリックデータだけでなく、さまざまな課題にデータを通して、向き合うことができると考えています。

ここまで学べばというゴールはありません。日常的に、データを通して、課題に向き合う習慣が身についていけばと願っています。

1.4 学習方法について

インターネット上で公開していますので、さまざまな方法で学んでいただきたいと思います。個人で学ぶことも可能で、実際に、それが可能なように、書いていく予定です。

しかし、おすすめは、何人かのグループ、または、大学などで一緒に学ぶことです。それは、データサイエンスの目的でもある、問いを持ち、課題に取り組んでいくためには、さまざまな視点からの意見や、考え方が必要だからです。異なる見方をたいせつにする訓練にもなります。

ひとつのグラフから、それぞれが違うことを発見することも多くあります。それを経験しながら、共に考えていく経験が貴重だと思います。

さらに、データの背景にあることを、想像したり、情報を得るために、グループの一員として、AI に加わってもらうことも、有効だと思います。できれば、複数の AI に質問をするのが良いでしょう。

さまざまな課題に、向き合うときに、グループのメンバーや、AI は、多様な意見を言ってくれることと思います。

データを元にした議論の訓練をすることで、根拠を明確にした説明をする訓練もすることができます。そのためにも、グループで学ぶことは有益です。

コードも、AI は教えてくれますが、聞き方が適切ではないと、間違った回答が得られることもあります。それも、グループで議論することで、聞き方を工夫していくことも可能

だと思います。

人の前で、発表する練習も、一連の学びの中で経験していくことをお勧めします。ぜひ、みなさんにあった、学びの場を協力して創っていきましょう。

1.5 参考

- 対話型 AI - Chat Bot について
 - AI の使い方や例について、書いてあります。参考にしてください。
- Data Analysis for Researchers 2022
 - オープンデータを用いた、データ分析の授業のデジタルブック

第 2 章

学ぶ内容

2.1 データサイエンス入門

具体的なデータを利用して、データサイエンスとは、どのようなものかをみていながら、ここで学ぶことの概要を紹介します。オープン・パブリックデータとしては、世界銀行のデータを使います。さまざまなデータが公開され、簡単に取得できるようになっている現状も紹介します。コードの詳細には、こだわらず、データサイエンスの実際について、雰囲気を感じていただければと思います。

2.2 第一部 パブリックデータ

世界のさまざまな、パブリックデータの紹介をし、ダッシュボードと呼ばれる機能を活用して、データをみることをします。

世界銀行の世界開発指標 (WDI)、国際連合 (UN Data)、OECD、日本のデータ (e-Stat) を外観します。

ここでは、R は使わず、サイトが提供するデータを探したり、サイト内でグラフを作成したり、データを取得するには、どのような方法があるかなどを紹介したいと思います。

これらの機関内の機能を、ホームページ閲覧ソフト (Google Chrome, Edge, Safari など) を使うだけで、かなりの情報が得られることを、経験していただければと思います。

2.3 第二部 基本

R の基本を学びます。R は、もともと、統計解析ソフトとして、開発されたもので、さまざまな分野の研究者によって利用され、また、それぞれの分野に必要な機能を、パッケージという形で開発して発展してきた言語です。非常に多くのひとたちが、開発に加わったために、痒いところに手が届く、多くの機能を、パッケージによって使うことができるようになりました。しかし、他方、統一性は十分ではなく、少し複雑な作業を実行するための、プログラミング言語としての機能も十分ではないという欠点も生じました。

わたしの理解では、それを一気に解決したのが、Hadley Wickham 等、その後、RStudio そして、現在の、Posit に引き継がれた、`tidyverse` というパッケージ群の開発です。他の研究者も、`tidyverse` の開発思想を受け継ぎ、発展させる形で、開発をしています。

そこで、R の起動とともに、最初に読み込まれる、Base R など、基本パッケージに、`tidyverse` を加えたものを基本として、極力、これらだけで、基本を学んでいきたいと思います。実際には、他のさまざまな便利なパッケージを使うことも、有用ですが、それは、後に回して、`tidyverse` を中心に学んでいきます。

`tidyverse` により、R は、プログラミング言語としても、一つの優秀な言語となったと思います。コーディングや、プログラミングと言われる、一つ一つのステップを構築し、それを繋げていくことを、学んでいきたいと思います。

もう一つ追加しておくのは、R Markdown の活用です。この「データサイエンスをはじめましょう」も、R Markdown の一つの形式、bookdown を使って書いています。

データサイエンスを学ぶ上で、わたしが必要かつ不可欠と考えているのが、再現性 (Reproducibility) と、なにを実行しているのかの説明を同時に記述すること (Literate Programming) です。コードとともに、その結果を、その下に出力し、かつ、そのコードの説明も加え、さらに、それによって、何がわかるかも、同時に書いていくことは、データサイエンスの核となすもので、それによって、データサイエンスの目的を達成することができると考えているからです。

データサイエンスでは、最後のコミュニケーションまでがひとつのまとまりです。他の人に聞いてもらうために発表したり、読んでもらうために、レポートを作成することも、一連の流れに加えることが必須だと思います。

指導してくださる方がいるときは、そのレポートをみてもらって、評価してもらったり、アドバイスを受けることは不可欠でしょう。それには、そのレポートに、コードとともに結果も書かれており、さらに、それは、何のためで、そこから、何が得られるのか書かれていることも必要です。

R Markdown の活用も、ともに学んでいきたいと思います。

2.4 第三部 国際機関などのデータの活用

R を使って、第一部で概観したデータを実際に分析する手法を学びます。

国際機関などの公的機関では、さまざまなデータを提供していますが、それぞれに特徴があり、データの形式や、データ取得の方法が異なります。それらを、少しずつ説明しながら、それぞれのデータを、すでに学んだことを応用しながら、分析する実際を経験していきたいと思います。

世界銀行の、世界開発指標 (World Development Indicators) が、一番整っているので、まずは、世界開発指標から学びますが、世界銀行の他のデータや、国際連合のデータ、他の国際連合の機関が提供しているデータや、経済開発協力機構 (OECD) や、Our World

in Data、Euro Stat などと共に、日本のデータである、e-Stat の使い方も学びたいと思います。

国際機関だけではなく、他にも、オープン・パブリックデータを提供しているところがたくさんあります。少しずつその利用方法も含めて、紹介していきたいと思います。

2.5 第四部 探索的データ分析 Exploratory Data Analysis

データを分析していくには、基本的なステップがありますが、その一つ一つのステップについて、より詳しく学びます。

これまでに、紹介できなかったいくつかの手法についても、紹介していきたいと思います。

2.6 第五部 分析例

実際の分析例を加えていきたいと思います。

2.7 付録

技術的なコメントなど、幾つかのトピックについて書いていきます。

だいたい、このような構成を考えています。

第3章

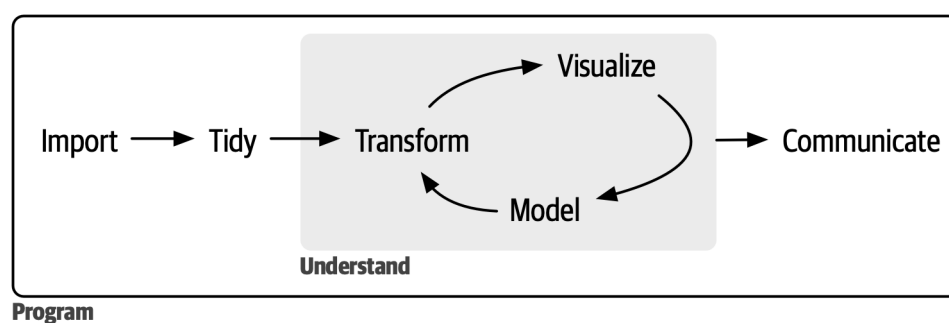
はじめてのデータサイエンス

3.1 データサイエンスの実際

データから情報を得るときには、大体次のような手順をとります。

1. 準備 Setup
2. データを取得 Import data
3. データ構造の確認 View data
4. 必要に応じて整形 Transform data
5. 視覚化 Visualize data
6. データを理解 Understand data
7. レポートなどにまとめる Communicate data

下の図は R for Data Science に掲載されている図です。よく、表現されていると思います。詳細は、少しずつ説明します。



はじめにに書きましたが、基本的には、問いをもちデータを取得し、視覚化などを通して、データを理解し、さらに問いを深めるサイクルが、データサイエンスの核だと思います。

R を使った分析の一つの例を、見て行きます。一つ一つのコード（コンピュータ・プログラム）の簡単な説明は、加えますが、あまりそれに捉われず、「データサイエンスとは何か？」を考えながら、雰囲気を味わってください。

3.2 Rのパッケージを活用

3.2.1 準備 Setup

世界銀行（World Bank）の、世界開発指標（WDI: World Development Indicators）の一つの、GDP（Gross Domestic Product 国内総生産）のデータから始めます。GDPにも何種類かの尺度がありますが、次のものを見てみます。

- NY.GDP.MKTP.CD: GDP (current US\$)*¹

NY.GDP.MKTP.CD は、データコードと言われるもので、世界開発指標（WDI）には、一つずつ決まっています。

World Development Indicators のサイトの下にある、Data Themes（テーマ）からテーマを選択し、下にスクロールすると、Code をみることができます。ちなみに、ここで利用する NY.GDP.MKTP.CD: GDP (current US\$) は、テーマ Economy（経済）の、一番上にあります。

経済用語の英語はよく知らないという方は、ブラウザ（Edge, Google Chrome, Safari など）の翻訳機能を使うのも良いでしょう。ただ、そのページの対話型の機能（interactive function）を利用するときは、翻訳機能を OFF にする必要がある場合もありますので、注意してください。

エラーメッセージを調べるときなどに、英語のほうが情報がたくさん得られますから、言語を、英語に変更しておきます。

R には、WDI のデータを取得する R のツール（パッケージ）WDI がありますから、それを使います。また、データを取り扱うための基本的なツール（パッケージ）tidyverse を使いますので、次のコードで、これらを読み込みます。# 以下はコメント（簡単な説明を書きました）#> 以下は、コードを実行すると、表示される情報（出力）です。以下同様です。

```
Sys.setenv(LANG = "en") # 言語を英語に
library(tidyverse)      # tidyverse パッケージを読み込みます
#> -- Attaching core tidyverse packages ---- tidyverse 2.0.0 --
#> v dplyr      1.1.2      v readr      2.1.4
#> v forcats    1.0.0      v stringr   1.5.0
#> v ggplot2    3.4.2      v tibble    3.2.1
#> v lubridate  1.9.2      v tidyr     1.3.0
#> v purrr      1.0.1
```

*¹ GDP（Gross Domestic Product）とは、ある国のある期間（通常は1年）における、その国で生産されたすべての最終財・サービスの市場価値の総額を指します。これは国内総生産とも呼ばれます。GDPは、その国の経済力や活力を測る指標の1つとして広く用いられています。WDIのGDP（Current USD）は、各国のGDPを米ドルで表したものであり、通貨の価値が異なっても比較可能な形で国際比較ができます。ただし、GDPはあくまで市場価格を基準としているため、非市場活動や自然災害などの影響を受ける場合がある点に注意が必要です。

```
#> -- Conflicts ----- tidyverse_conflicts() --
#> x dplyr::filter() masks stats::filter()
#> x dplyr::lag() masks stats::lag()
#> i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
library(WDI) # WDI パッケージを読み込みます
```

データを保存する場所を作成しておくことをお勧めします。保存しておくときは、このディレクトリを使います。

```
dir.create("./data")
```

3.2.2 データ取得 Import data

データを取得します。少し時間がかかります。取得したデータに、`df_gdp` などと、わかりやすい名前をつけます。`df` は data frame の略で、R で標準的なデータの形式です。

```
df_gdp <- WDI(country = "all",
              indicator = c(gdp = "NY.GDP.MKTP.CD"),
              extra = TRUE)
```

このコードで、全ての国の GDP を取得できます。GDP の値は、`NY.GDP.MKTP.CD` という名前の列にありますが、覚えやすいように、`gdp` という名前に変更しておきます。`extra = TRUE` とすることによって、それぞれの国についての情報などが追加されます。

3.2.3 データ構造の確認

最初の数行だけを見るには、`head(df_gdp)` とします。

```
head(df_gdp)
#> # A tibble: 6 x 13
#>   country iso2c iso3c year   gdp status lastupdated region
#>   <chr>   <chr> <chr> <dbl> <dbl> <lgl> <date>   <chr>
#> 1 Afghan~ AF    AFG    1963 7.51e8 NA    2023-07-25 South~
#> 2 Afghan~ AF    AFG    1962 5.47e8 NA    2023-07-25 South~
#> 3 Afghan~ AF    AFG    1961 5.49e8 NA    2023-07-25 South~
#> 4 Afghan~ AF    AFG    1960 5.38e8 NA    2023-07-25 South~
#> 5 Afghan~ AF    AFG    2003 4.54e9 NA    2023-07-25 South~
#> 6 Afghan~ AF    AFG    2002 3.85e9 NA    2023-07-25 South~
#> # i 5 more variables: capital <chr>, longitude <dbl>,
#> #   latitude <dbl>, income <chr>, lending <chr>
```

データの構造を見るときには、`str(df_gdp)` もよく使われます。今度は、列が縦に並んで表示されます。

```

str(df_gdp)
#> spc_tbl_ [16,758 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
#> $ country      : chr [1:16758] "Afghanistan" "Afghanistan" "Afghanistan" "Afgha
#> $ iso2c         : chr [1:16758] "AF" "AF" "AF" "AF" ...
#> $ iso3c         : chr [1:16758] "AFG" "AFG" "AFG" "AFG" ...
#> $ year          : num [1:16758] 1963 1962 1961 1960 2003 ...
#> $ gdp           : num [1:16758] 7.51e+08 5.47e+08 5.49e+08 5.38e+08 4.54e+09 ...
#> $ status        : logi [1:16758] NA NA NA NA NA NA ...
#> $ lastupdated: Date[1:16758], format: "2023-07-25" ...
#> $ region        : chr [1:16758] "South Asia" "South Asia" "South Asia" "South As
#> $ capital        : chr [1:16758] "Kabul" "Kabul" "Kabul" "Kabul" ...
#> $ longitude     : num [1:16758] 69.2 69.2 69.2 69.2 69.2 ...
#> $ latitude      : num [1:16758] 34.5 34.5 34.5 34.5 34.5 ...
#> $ income        : chr [1:16758] "Low income" "Low income" "Low income" "Low inco
#> $ lending       : chr [1:16758] "IDA" "IDA" "IDA" "IDA" ...
#> - attr(*, "spec")=
#> .. cols(
#> ..   country = col_character(),
#> ..   iso2c = col_character(),
#> ..   iso3c = col_character(),
#> ..   year = col_double(),
#> ..   gdp = col_double(),
#> ..   status = col_logical(),
#> ..   lastupdated = col_date(format = ""),
#> ..   region = col_character(),
#> ..   capital = col_character(),
#> ..   longitude = col_double(),
#> ..   latitude = col_double(),
#> ..   income = col_character(),
#> ..   lending = col_character()
#> .. )
#> - attr(*, "problems")=<externalptr>

```

概要 (summary(df_gdp)) からもある程度わかります。

```

summary(df_gdp)
#>   country      iso2c      iso3c
#> Length:16758      Length:16758      Length:16758
#> Class :character  Class :character  Class :character
#> Mode  :character  Mode  :character  Mode  :character
#>
#>

```

```

#>
#>
#>      year      gdp      status
#> Min.   :1960   Min.   :8.825e+06   Mode:logical
#> 1st Qu.:1975   1st Qu.:2.523e+09   NA's:16758
#> Median :1991   Median :1.843e+10
#> Mean   :1991   Mean   :1.207e+12
#> 3rd Qu.:2007   3rd Qu.:2.244e+11
#> Max.   :2022   Max.   :1.006e+14
#>      NA's      :3393
#> lastupdated      region      capital
#> Min.   :2023-07-25   Length:16758   Length:16758
#> 1st Qu.:2023-07-25   Class :character   Class :character
#> Median :2023-07-25   Mode  :character   Mode  :character
#> Mean   :2023-07-25
#> 3rd Qu.:2023-07-25
#> Max.   :2023-07-25
#>
#> longitude      latitude      income
#> Min.   : -175.22   Min.   : -41.286   Length:16758
#> 1st Qu.: -15.18   1st Qu.:  4.174   Class :character
#> Median :  19.54   Median : 17.277   Mode  :character
#> Mean   :  19.16   Mean   : 18.740
#> 3rd Qu.:  50.53   3rd Qu.: 39.715
#> Max.   : 179.09   Max.   : 64.184
#> NA's   :3528     NA's   :3528
#> lending
#> Length:16758
#> Class :character
#> Mode  :character
#>
#>
#>
#>

```

国のリストをみてみましょう。とても長いリストの中には、地域名も含まれています。

```

df_gdp |> distinct(country) |> pull()
#> [1] "Afghanistan"
#> [2] "Africa Eastern and Southern"
#> [3] "Africa Western and Central"
#> [4] "Albania"

```

```
#> [5] "Algeria"
#> [6] "American Samoa"
#> [7] "Andorra"
#> [8] "Angola"
#> [9] "Antigua and Barbuda"
#> [10] "Arab World"
#> [11] "Argentina"
#> [12] "Armenia"
#> [13] "Aruba"
#> [14] "Australia"
#> [15] "Austria"
#> [16] "Azerbaijan"
#> [17] "Bahamas, The"
#> [18] "Bahrain"
#> [19] "Bangladesh"
#> [20] "Barbados"
#> [21] "Belarus"
#> [22] "Belgium"
#> [23] "Belize"
#> [24] "Benin"
#> [25] "Bermuda"
#> [26] "Bhutan"
#> [27] "Bolivia"
#> [28] "Bosnia and Herzegovina"
#> [29] "Botswana"
#> [30] "Brazil"
#> [31] "British Virgin Islands"
#> [32] "Brunei Darussalam"
#> [33] "Bulgaria"
#> [34] "Burkina Faso"
#> [35] "Burundi"
#> [36] "Cabo Verde"
#> [37] "Cambodia"
#> [38] "Cameroon"
#> [39] "Canada"
#> [40] "Caribbean small states"
#> [41] "Cayman Islands"
#> [42] "Central African Republic"
#> [43] "Central Europe and the Baltics"
#> [44] "Chad"
#> [45] "Channel Islands"
```

```
#> [46] "Chile"
#> [47] "China"
#> [48] "Colombia"
#> [49] "Comoros"
#> [50] "Congo, Dem. Rep."
#> [51] "Congo, Rep."
#> [52] "Costa Rica"
#> [53] "Cote d'Ivoire"
#> [54] "Croatia"
#> [55] "Cuba"
#> [56] "Curacao"
#> [57] "Cyprus"
#> [58] "Czechia"
#> [59] "Denmark"
#> [60] "Djibouti"
#> [61] "Dominica"
#> [62] "Dominican Republic"
#> [63] "Early-demographic dividend"
#> [64] "East Asia & Pacific"
#> [65] "East Asia & Pacific (excluding high income)"
#> [66] "East Asia & Pacific (IDA & IBRD countries)"
#> [67] "Ecuador"
#> [68] "Egypt, Arab Rep."
#> [69] "El Salvador"
#> [70] "Equatorial Guinea"
#> [71] "Eritrea"
#> [72] "Estonia"
#> [73] "Eswatini"
#> [74] "Ethiopia"
#> [75] "Euro area"
#> [76] "Europe & Central Asia"
#> [77] "Europe & Central Asia (excluding high income)"
#> [78] "Europe & Central Asia (IDA & IBRD countries)"
#> [79] "European Union"
#> [80] "Faroe Islands"
#> [81] "Fiji"
#> [82] "Finland"
#> [83] "Fragile and conflict affected situations"
#> [84] "France"
#> [85] "French Polynesia"
#> [86] "Gabon"
```

```
#> [87] "Gambia, The"
#> [88] "Georgia"
#> [89] "Germany"
#> [90] "Ghana"
#> [91] "Gibraltar"
#> [92] "Greece"
#> [93] "Greenland"
#> [94] "Grenada"
#> [95] "Guam"
#> [96] "Guatemala"
#> [97] "Guinea"
#> [98] "Guinea-Bissau"
#> [99] "Guyana"
#> [100] "Haiti"
#> [101] "Heavily indebted poor countries (HIPC)"
#> [102] "High income"
#> [103] "Honduras"
#> [104] "Hong Kong SAR, China"
#> [105] "Hungary"
#> [106] "IBRD only"
#> [107] "Iceland"
#> [108] "IDA & IBRD total"
#> [109] "IDA blend"
#> [110] "IDA only"
#> [111] "IDA total"
#> [112] "India"
#> [113] "Indonesia"
#> [114] "Iran, Islamic Rep."
#> [115] "Iraq"
#> [116] "Ireland"
#> [117] "Isle of Man"
#> [118] "Israel"
#> [119] "Italy"
#> [120] "Jamaica"
#> [121] "Japan"
#> [122] "Jordan"
#> [123] "Kazakhstan"
#> [124] "Kenya"
#> [125] "Kiribati"
#> [126] "Korea, Dem. People's Rep."
#> [127] "Korea, Rep."
```



```
#> [128] "Kosovo"
#> [129] "Kuwait"
#> [130] "Kyrgyz Republic"
#> [131] "Lao PDR"
#> [132] "Late-demographic dividend"
#> [133] "Latin America & Caribbean"
#> [134] "Latin America & Caribbean (excluding high income)"
#> [135] "Latin America & the Caribbean (IDA & IBRD countries)"
#> [136] "Latvia"
#> [137] "Least developed countries: UN classification"
#> [138] "Lebanon"
#> [139] "Lesotho"
#> [140] "Liberia"
#> [141] "Libya"
#> [142] "Liechtenstein"
#> [143] "Lithuania"
#> [144] "Low & middle income"
#> [145] "Low income"
#> [146] "Lower middle income"
#> [147] "Luxembourg"
#> [148] "Macao SAR, China"
#> [149] "Madagascar"
#> [150] "Malawi"
#> [151] "Malaysia"
#> [152] "Maldives"
#> [153] "Mali"
#> [154] "Malta"
#> [155] "Marshall Islands"
#> [156] "Mauritania"
#> [157] "Mauritius"
#> [158] "Mexico"
#> [159] "Micronesia, Fed. Sts."
#> [160] "Middle East & North Africa"
#> [161] "Middle East & North Africa (excluding high income)"
#> [162] "Middle East & North Africa (IDA & IBRD countries)"
#> [163] "Middle income"
#> [164] "Moldova"
#> [165] "Monaco"
#> [166] "Mongolia"
#> [167] "Montenegro"
#> [168] "Morocco"
```

```
#> [169] "Mozambique"  
#> [170] "Myanmar"  
#> [171] "Namibia"  
#> [172] "Nauru"  
#> [173] "Nepal"  
#> [174] "Netherlands"  
#> [175] "New Caledonia"  
#> [176] "New Zealand"  
#> [177] "Nicaragua"  
#> [178] "Niger"  
#> [179] "Nigeria"  
#> [180] "North America"  
#> [181] "North Macedonia"  
#> [182] "Northern Mariana Islands"  
#> [183] "Norway"  
#> [184] "Not classified"  
#> [185] "OECD members"  
#> [186] "Oman"  
#> [187] "Other small states"  
#> [188] "Pacific island small states"  
#> [189] "Pakistan"  
#> [190] "Palau"  
#> [191] "Panama"  
#> [192] "Papua New Guinea"  
#> [193] "Paraguay"  
#> [194] "Peru"  
#> [195] "Philippines"  
#> [196] "Poland"  
#> [197] "Portugal"  
#> [198] "Post-demographic dividend"  
#> [199] "Pre-demographic dividend"  
#> [200] "Puerto Rico"  
#> [201] "Qatar"  
#> [202] "Romania"  
#> [203] "Russian Federation"  
#> [204] "Rwanda"  
#> [205] "Samoa"  
#> [206] "San Marino"  
#> [207] "Sao Tome and Principe"  
#> [208] "Saudi Arabia"  
#> [209] "Senegal"
```

```
#> [210] "Serbia"
#> [211] "Seychelles"
#> [212] "Sierra Leone"
#> [213] "Singapore"
#> [214] "Sint Maarten (Dutch part)"
#> [215] "Slovak Republic"
#> [216] "Slovenia"
#> [217] "Small states"
#> [218] "Solomon Islands"
#> [219] "Somalia"
#> [220] "South Africa"
#> [221] "South Asia"
#> [222] "South Asia (IDA & IBRD)"
#> [223] "South Sudan"
#> [224] "Spain"
#> [225] "Sri Lanka"
#> [226] "St. Kitts and Nevis"
#> [227] "St. Lucia"
#> [228] "St. Martin (French part)"
#> [229] "St. Vincent and the Grenadines"
#> [230] "Sub-Saharan Africa"
#> [231] "Sub-Saharan Africa (excluding high income)"
#> [232] "Sub-Saharan Africa (IDA & IBRD countries)"
#> [233] "Sudan"
#> [234] "Suriname"
#> [235] "Sweden"
#> [236] "Switzerland"
#> [237] "Syrian Arab Republic"
#> [238] "Tajikistan"
#> [239] "Tanzania"
#> [240] "Thailand"
#> [241] "Timor-Leste"
#> [242] "Togo"
#> [243] "Tonga"
#> [244] "Trinidad and Tobago"
#> [245] "Tunisia"
#> [246] "Turkiye"
#> [247] "Turkmenistan"
#> [248] "Turks and Caicos Islands"
#> [249] "Tuvalu"
#> [250] "Uganda"
```

```
#> [251] "Ukraine"
#> [252] "United Arab Emirates"
#> [253] "United Kingdom"
#> [254] "United States"
#> [255] "Upper middle income"
#> [256] "Uruguay"
#> [257] "Uzbekistan"
#> [258] "Vanuatu"
#> [259] "Venezuela, RB"
#> [260] "Vietnam"
#> [261] "Virgin Islands (U.S.)"
#> [262] "West Bank and Gaza"
#> [263] "World"
#> [264] "Yemen, Rep."
#> [265] "Zambia"
#> [266] "Zimbabwe"
```

今回は下のように、`|>`（パイプと呼びます）で繋げてコードを書きました。

```
df_gdp |> distinct(country) |> pull()
```

最初は、データ、その中の、異なる国を選択して、書き出してくださいというものです。

これは、

```
pull(distinct(df_gdp, country))
```

と同じです。どんどん、かっこの中に入れ子になって複雑になるので、一つ一つのステップを、順に書いたものが、最初のものになります。

```
df_gdp |> head()
```

```
df_gdp |> str()
```

なども可能です。かっこの中に最初に入るものが直前のもの、ここでは、データになっています。

3.2.4 必要に応じて整形 Transform data

変数が多いので、日本の部分だけ `filter` を使って選択します。country が Japan と一致する場合のみを選択するときは、`==` を使います。数値ではないので、引用符をつけます。半角を使ってください。

```
df_gdp |> filter(country == "Japan")
#> # A tibble: 63 x 13
#>   country iso2c iso3c year      gdp status lastupdated
#>   <chr>    <chr> <chr> <dbl>   <dbl> <lgl>   <date>
```

```
#> 1 Japan JP JPN 2022 4.23e12 NA 2023-07-25
#> 2 Japan JP JPN 2021 5.01e12 NA 2023-07-25
#> 3 Japan JP JPN 2020 5.05e12 NA 2023-07-25
#> 4 Japan JP JPN 2019 5.12e12 NA 2023-07-25
#> 5 Japan JP JPN 2018 5.04e12 NA 2023-07-25
#> 6 Japan JP JPN 2017 4.93e12 NA 2023-07-25
#> 7 Japan JP JPN 2016 5.00e12 NA 2023-07-25
#> 8 Japan JP JPN 2015 4.44e12 NA 2023-07-25
#> 9 Japan JP JPN 2014 4.90e12 NA 2023-07-25
#> 10 Japan JP JPN 2013 5.21e12 NA 2023-07-25
#> # i 53 more rows
#> # i 6 more variables: region <chr>, capital <chr>,
#> # longitude <dbl>, latitude <dbl>, income <chr>,
#> # lending <chr>
```

```
df_gdp |> filter(country == "Japan") |> head(2)
#> # A tibble: 2 x 13
#>   country iso2c iso3c year    gdp status lastupdated
#>   <chr>    <chr> <chr> <dbl>  <dbl> <lgl>   <date>
#> 1 Japan   JP    JPN  2022 4.23e12 NA    2023-07-25
#> 2 Japan   JP    JPN  2021 5.01e12 NA    2023-07-25
#> # i 6 more variables: region <chr>, capital <chr>,
#> # longitude <dbl>, latitude <dbl>, income <chr>,
#> # lending <chr>
```

2行目の、gdp の、4.940878e+12（この文書では、幅の都合で、4.9e+12 と表示されているかもしれませんが）は、Scientific notation とされるもので、

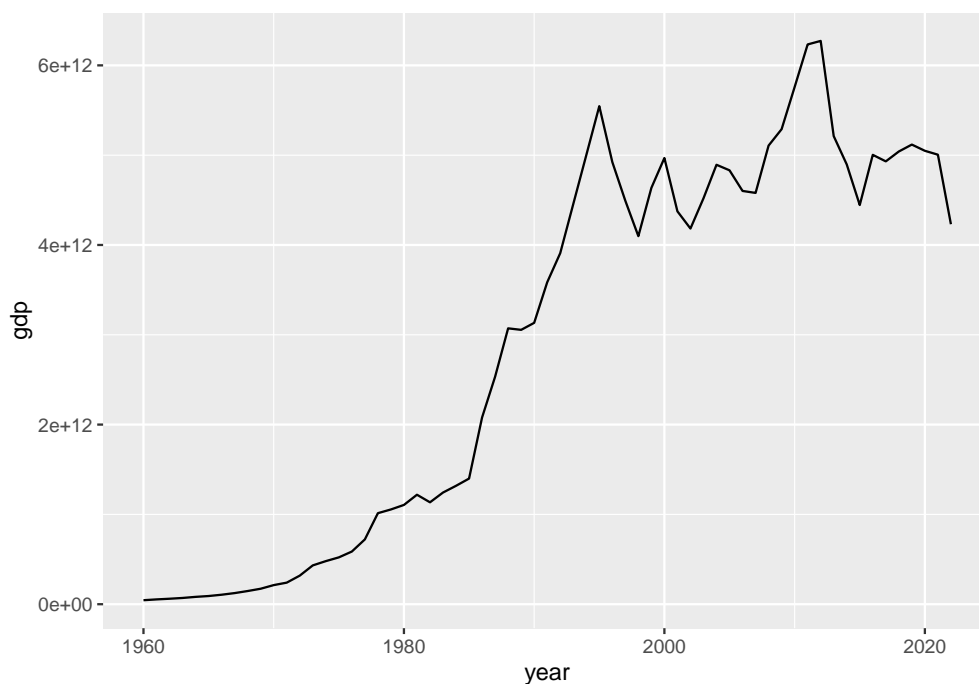
$$4.940878 \times 10^{12} = 4,940,887,800,000$$

を意味します。e+3 は千 (thousand)、e+6 は百万 (million)、e+9 は、10 億 (billion)、e+12 は、兆 (trillion) ですから、日本の、2021 年の GDP は、約 5 兆ドルとなります。

3.2.5 視覚化 data visualization

3.2.5.1 Fig 1. 日本の GDP の経年変化を表す折線グラフ (line graph)

```
df_gdp |> filter(country == "Japan") |>
  ggplot(aes(x = year, y = gdp)) + geom_line()
```



```
df_gdp |> filter(country == "Japan") |>
  ggplot(aes(x = year, y = gdp)) + geom_line()
```

日本を選択したときに、それに名前をつけた新しいオブジェクトを作り、それを使うこともできますが、名前がどんどん増えるので、パイプを使って、コードを続けて書いていく方法をとっています。

```
ggplot(aes(x = year, y = gdp)) + geom_line()
```

の部分が、グラフを描く部分で、「x 軸を、year、y 軸を、gdp として、それを、折線グラフで描いてください」というコードです。

Warning: [38;5;238mRemoved 1 row containing missing values

などと表示される場合がありますが、それは、値がない（missing または値が NA, not available）年があることを言っています。データがない年を最初から削除してこくことも可能です。

3.2.6 データの理解 Understand data

上の折線グラフを使った、視覚化によって見えてくるものがいくつもありますね。どんなことがわかりますか。気づいたこと（observation）をあげてみましょう。

コードを描くことではなく、この部分が、データサイエンスの核の部分です。気づいたこと、疑問点を列挙してみましょう。

急激に増加しているとき、増加減少が繰り返している時、全体としては、1995 年ごろからはあまり、増加していないように見えることなどがわかりますね。それぞれのピークや、

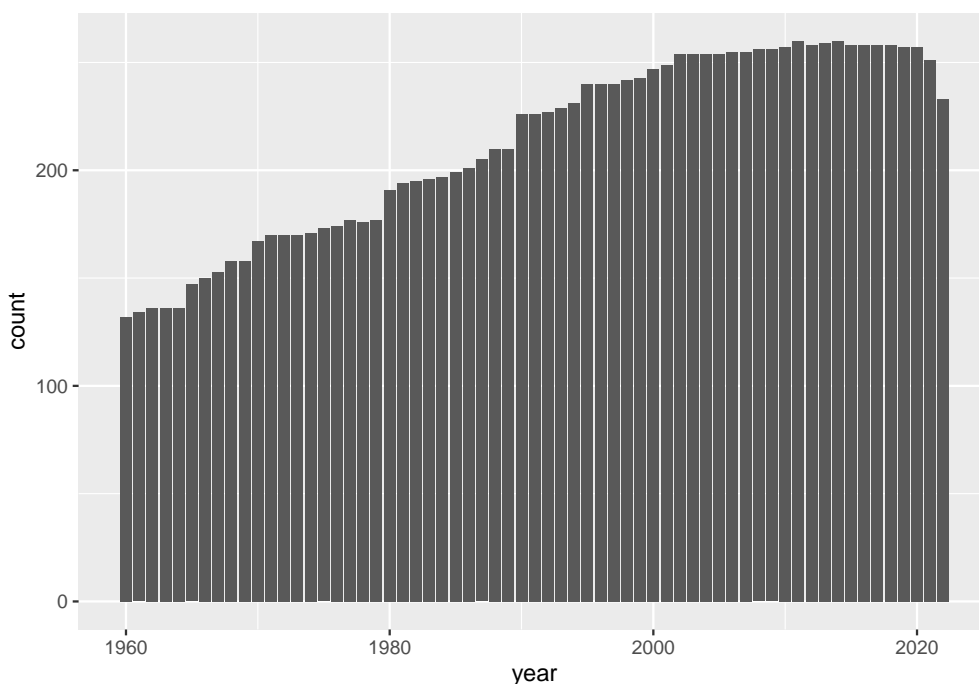
下落は、なにがあったのかも気になりませんか。これは、世界的な傾向なのでしょうか。日本だけでしょうか。他にも似た傾向の国があるのでしょうか。将来はどうなっていくのでしょうか。などなど。

3.2.7 さまざまな視覚化

3.2.7.1 Fig 2. 各年ごとのデータの数

`summary(df_gdp)` で、データ自体は、1960 年から 2022 年までのようですが、年によって、どの程度データがあるか、調べてみます。

```
df_gdp |> drop_na(gdp) |> ggplot(aes(x = year)) + geom_bar()
```



```
df_gdp |> drop_na(gdp) |> ggplot(aes(x = year)) + geom_bar()
```

棒グラフ (bar graph) を使います。gdp の値が、欠損値 (NA: not available) のデータを削除してから、グラフを描きます。

3.2.7.2 2022 年の GDP の降順での表示 (1)

最新の 2022 年のデータはすべてあるわけではなさそうですが、gdp の値が大きい順に並べてみましょう。

```
df_gdp |> filter(year == 2022) |> drop_na(gdp) |> arrange(desc(gdp))
#> # A tibble: 233 x 13
#>   country      iso2c iso3c year      gdp status lastupdated
#>   <chr>        <chr> <chr> <dbl>   <dbl> <lgl>   <date>
#> 1 World       1W     WLD   2022 1.01e14 NA     2023-07-25
#> 2 High income XD     <NA>   2022 6.15e13 NA     2023-07-25
```

```
#> 3 OECD members OE OED 2022 5.96e13 NA 2023-07-25
#> 4 Post-demogr~ V4 PST 2022 5.60e13 NA 2023-07-25
#> 5 IDA & IBRD ~ ZT IBT 2022 4.04e13 NA 2023-07-25
#> 6 Low & middl~ XO LMY 2022 3.87e13 NA 2023-07-25
#> 7 Middle inco~ XP MIC 2022 3.82e13 NA 2023-07-25
#> 8 IBRD only XF IBD 2022 3.76e13 NA 2023-07-25
#> 9 East Asia & Z4 EAS 2022 3.07e13 NA 2023-07-25
#> 10 Upper middl~ XT <NA> 2022 3.01e13 NA 2023-07-25
#> # i 223 more rows
#> # i 6 more variables: region <chr>, capital <chr>,
#> # longitude <dbl>, latitude <dbl>, income <chr>,
#> # lending <chr>
```

3.2.7.3 2021 年の GDP の降順での表示 (2)

最初に、World と表示され、グループや、カテゴリーのデータもあるようですから、それを、まず、削除することが必要です。region の列を見ると、World などは、Aggregates となっているので、そのようなものを削除すればよさそうです。数値の大きい順に並べたいので、desc 降順 (descending order) にします。

```
df_gdp |> filter(year == 2022, region != "Aggregates") |>
  drop_na(gdp) |> arrange(desc(gdp))
#> # A tibble: 184 x 13
#>   country      iso2c iso3c year      gdp status lastupdated
#>   <chr>        <chr> <chr> <dbl>    <dbl> <lgl>    <date>
#> 1 United Stat~ US    USA  2022 2.55e13 NA    2023-07-25
#> 2 China      CN    CHN  2022 1.80e13 NA    2023-07-25
#> 3 Japan      JP    JPN  2022 4.23e12 NA    2023-07-25
#> 4 Germany    DE    DEU  2022 4.07e12 NA    2023-07-25
#> 5 India      IN    IND  2022 3.39e12 NA    2023-07-25
#> 6 United King~ GB    GBR  2022 3.07e12 NA    2023-07-25
#> 7 France     FR    FRA  2022 2.78e12 NA    2023-07-25
#> 8 Russian Fed~ RU    RUS  2022 2.24e12 NA    2023-07-25
#> 9 Canada     CA    CAN  2022 2.14e12 NA    2023-07-25
#> 10 Italy      IT    ITA  2022 2.01e12 NA    2023-07-25
#> # i 174 more rows
#> # i 6 more variables: region <chr>, capital <chr>,
#> # longitude <dbl>, latitude <dbl>, income <chr>,
#> # lending <chr>
```

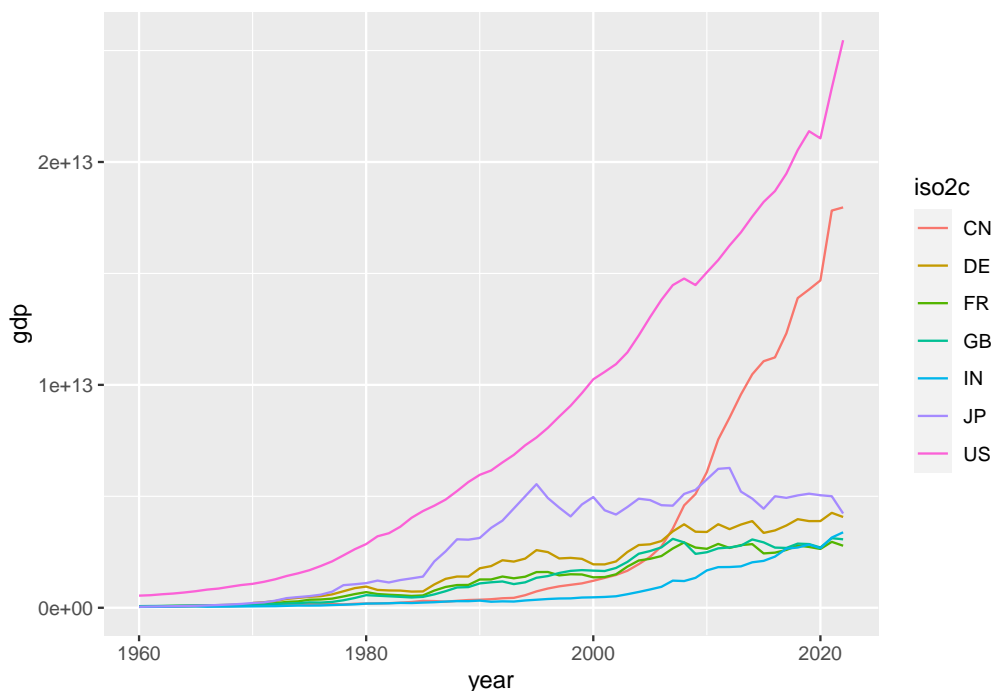
これは、グラフではありませんが、これも一つの可視化と考えられないことはありません。

上位7カ国は、United States, China, Japan, Germany, India, United Kingdom, France

であることがわかりました。このあと、Russian Federation, Canada, Italy と続き、でこまでが、2022 年の GDP が 2 兆ドルを越している国となります。

3.2.7.4 Fig 3. 2022 年時の GDP 上位 7 カ国の GDP 経年変化

```
df_gdp |> filter(iso2c %in% c("US", "CN", "JP", "DE", "IN", "GB", "FR")) |>
  ggplot(aes(x = year, y = gdp, col = iso2c)) + geom_line()
#> Warning: Removed 10 rows containing missing values
#> (`geom_line()`).
```



```
df_gdp |> filter(iso2c %in% c("US", "CN", "JP", "DE", "IN", "GB", "FR")) |>
  ggplot(aes(x = year, y = gdp, col = iso2c)) + geom_line()
```

ここでは、最初に、`filter` を使って、7 カ国のデータを選択しています。そのときには、`%in%` として、国名を、combine するといういみで、`c()` とひとまとめにします。数字ではなく、文字なので、引用符で囲んでいます。この場合は、single quote でも構いませんが、半角を使ってください。

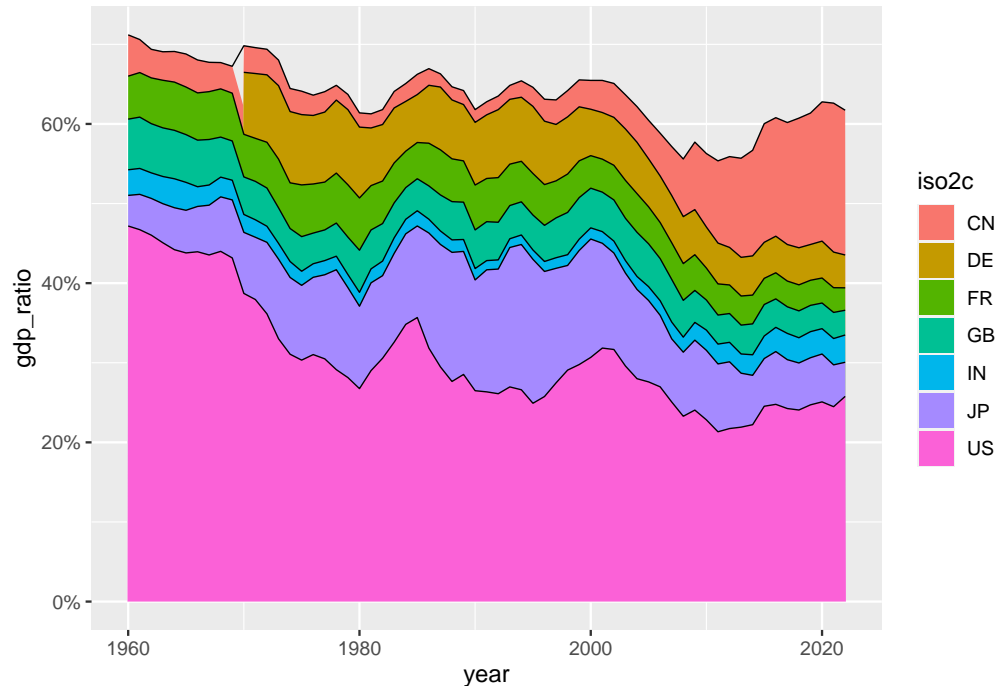
このグラフからは、どのようなことがわかりますか。気づいたことを書いてみましょう。

もう少し、このようなグラフをみてみたいというような、メモも大切です。

3.2.7.5 Fig 4. 世界の GDP における割合 (1)

```
df_gdp |>
  filter(region != "Aggregates") |> drop_na(gdp) |>
  group_by(year) |> mutate(gdp_ratio = gdp/sum(gdp)) |> ungroup() |>
```

```
filter(iso2c %in% c("US", "CN", "JP", "DE", "IN", "GB", "FR")) |>
ggplot(aes(x = year, y = gdp_ratio, fill = iso2c)) + geom_area() +
geom_line(col = "black", position = "stack", linewidth = 0.3) +
scale_y_continuous(labels = scales::percent_format(accuracy = 1))
```



まず、下の部分が新しいですが、ここでは、年毎にグループにして、その上で、新しい `gdp_ratio` という名前の列を追加し、その `gdp` の値を、`gdp` 合計で割っています。すなわち、世界の、GDP における割合が計算されています。

```
group_by(year) |> mutate(gdp_ratio = gdp/sum(gdp)) |> ungroup() |>
```

下の部分では、`geom_area` を使って、`fill=iso2c` により、`iso2c` ごとに、違う色を塗って、`position = "stack"` により、積み上げ型の、グラフを描き、境目がわかりやすいように、0.3 の太さの黒の線を描いてください。また、`y` 軸は、小数点以下を省いたパーセント表示に変えてください。というコードです。

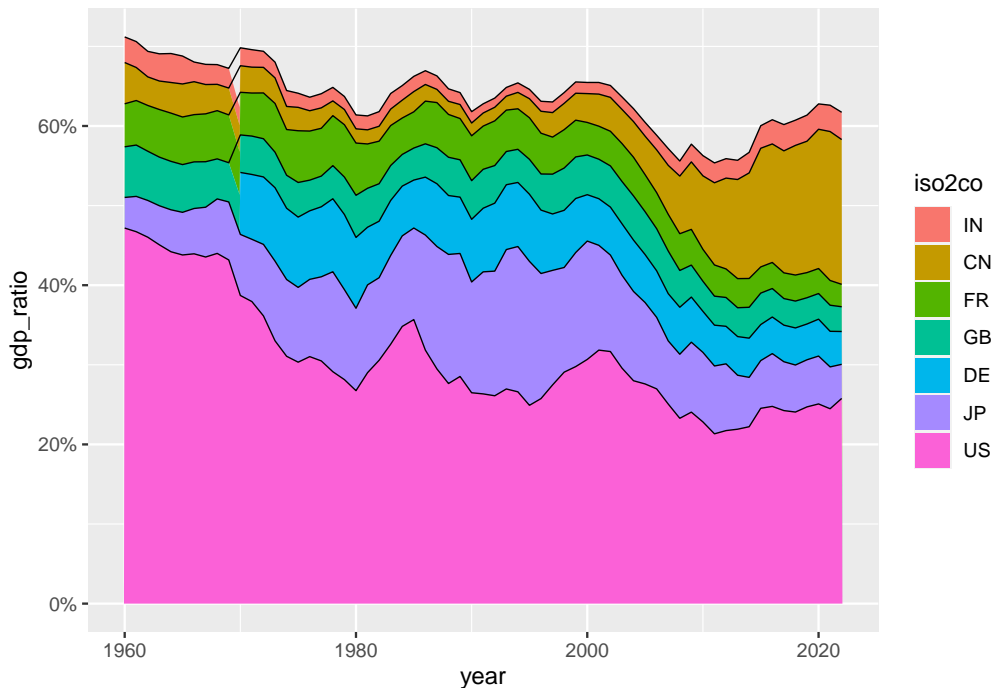
```
ggplot(aes(x = year, y = gdp_ratio, fill = iso2c)) + geom_area() +
geom_line(col = "black", position = "stack", linewidth = 0.3) +
scale_y_continuous(labels = scales::percent_format(accuracy = 1))
```

3.2.7.6 Fig 4. 世界の GDP における割合 (2)

これは、上から、`iso2c` のアルファベットの順番になっていますが、積み上げの順序を変更することもできます。

```
df_gdp |>
filter(region != "Aggregates") |> drop_na(gdp) |>
```

```
group_by(year) |> mutate(gdp_ratio = gdp/sum(gdp)) |> ungroup() |>
filter(iso2c %in% c("US", "CN", "JP", "DE", "IN", "GB", "FR")) |>
mutate(iso2co = factor(iso2c, levels = c("IN", "CN", "FR", "GB", "DE", "JP", "US"))) |>
ggplot(aes(x = year, y = gdp_ratio, fill = iso2co)) + geom_area() +
geom_line(col = "black", position = "stack", linewidth = 0.3) +
scale_y_continuous(labels = scales::percent_format(accuracy = 1))
```



これらは、世界全体の GPT における割合です。

どのようなことがわかりますか。

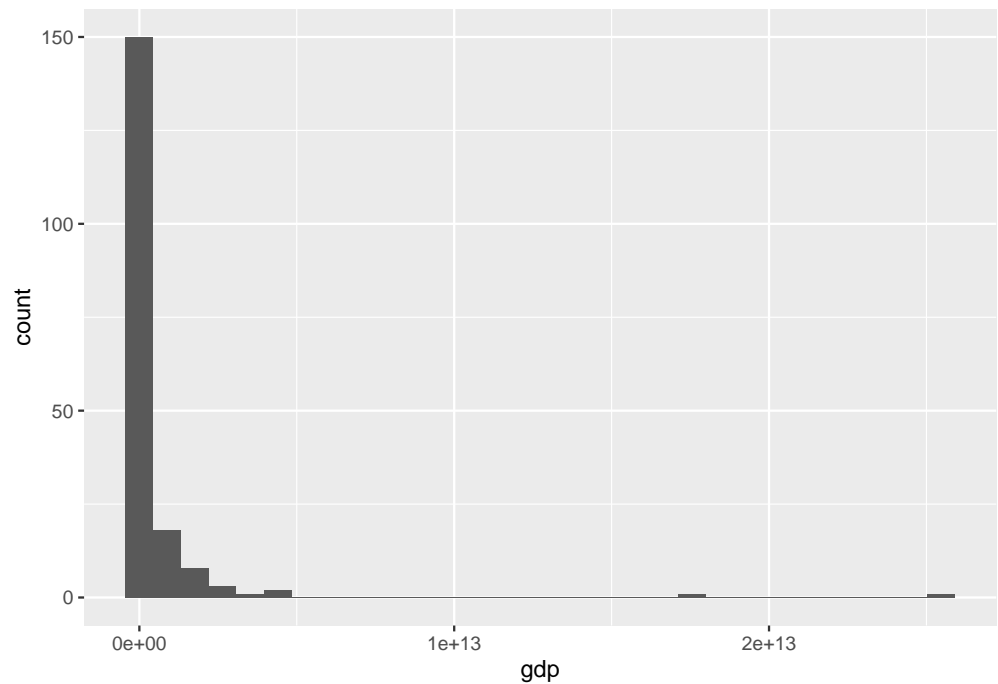
主要国で、60%～70% を占めていることがわかります。それぞれの国や、幾つかの国の影響力や、その時代による変化も、ある程度みることができるようになります。

気づいたこと、疑問に思ったことなどを、書き出してみてください。

GDP が大きな国と、小さな国があるのはわかりますが、それは、どのように分布しているのでしょうか。

3.2.7.7 Fig 5. 2022 年の世界の国の GDP の分布 (1)

```
df_gdp |> drop_na(gdp) |>
  filter(year == 2022) |> filter(region != "Aggregates") |>
  ggplot(aes(gdp)) + geom_histogram()
#> `stat_bin()` using `bins = 30`. Pick better value with
#> `binwidth`.
```

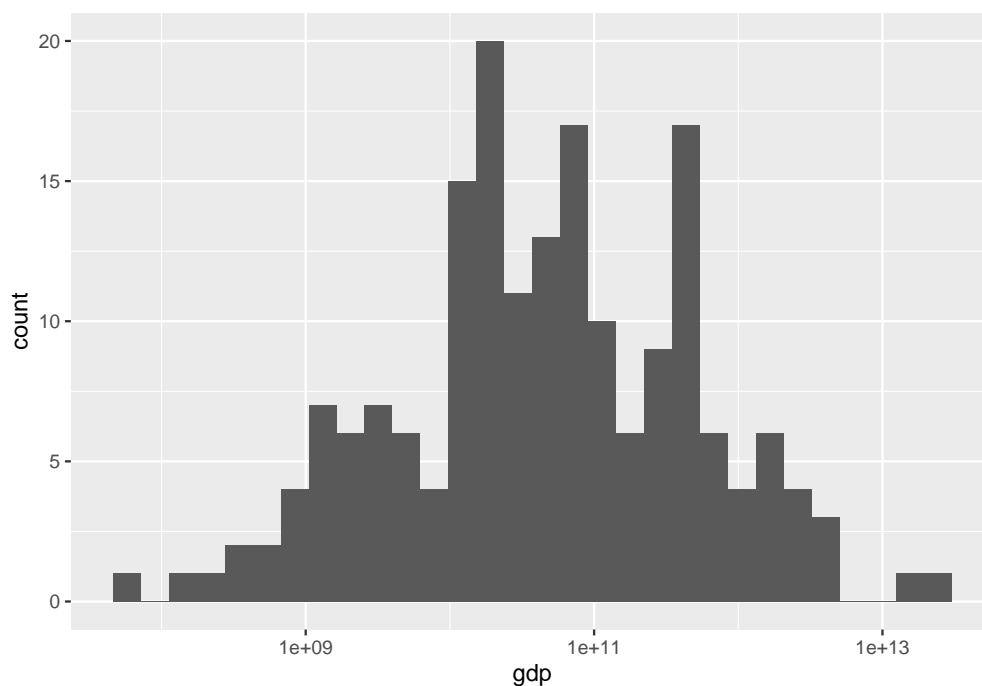


小さいところに集中していることがわかりますが、`scale_x_log10()` を加え、対数軸をとってみます。

$\log_{10}(1000) = 3$, $\log_{10}(1000000) = 6$, $\log_{10}(1000000000) = 9$ などになります。

3.2.7.8 Fig 6. 2022 年の世界の国の GDP の分布 (2)

```
df_gdp |> drop_na(gdp) |>
  filter(year == 2022) |> filter(region != "Aggregates") |>
  ggplot(aes(gdp)) + geom_histogram() + scale_x_log10()
#> `stat_bin()` using `bins = 30`. Pick better value with
#> `binwidth`.
```

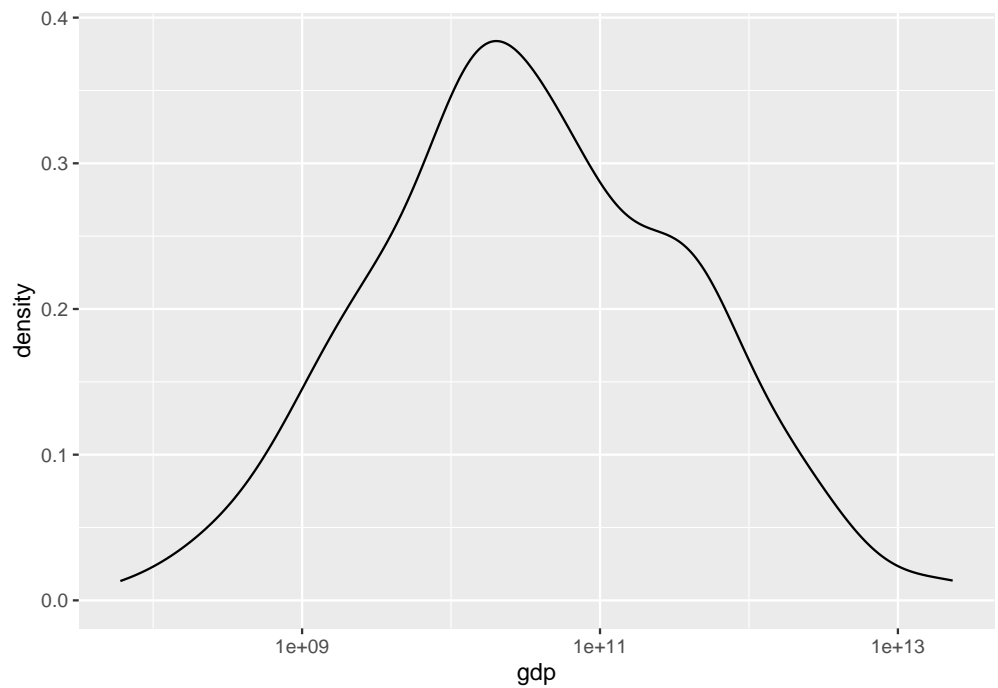


幅を変更したり、分ける個数を変更するには `binwidth = 0.5` や、`bins = 20` を、`geom_histogram()` のかっこの中に入れます。

3.2.7.9 Fig 7. 2021 年の世界の国の GDP の分布（3）

また、密度曲線に変えるには、`geom_density` を使います。

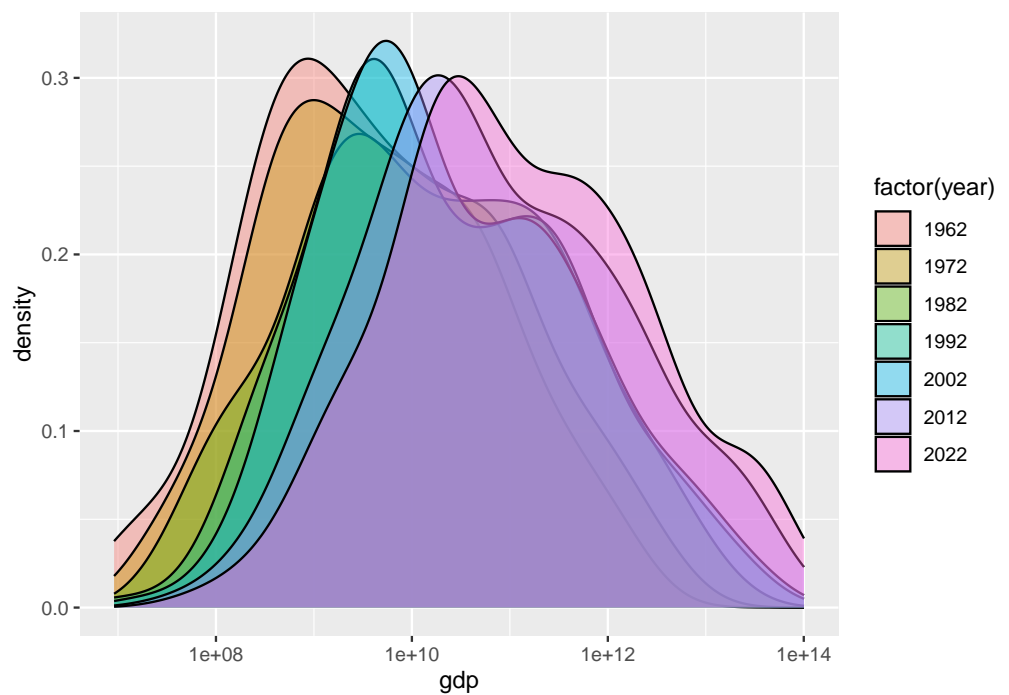
```
df_gdp |> drop_na(gdp) |>  
  filter(year == 2021) |> filter(region != "Aggregates") |>  
  ggplot(aes(gdp)) + geom_density() + scale_x_log10()
```



3.2.7.10 Fig 8. 2022 年までの世界の国の GDP の分布（4）

これは、2022 年のデータですが、10 年ごとの、density の変化をしてみます。alpha の値は透明度です。

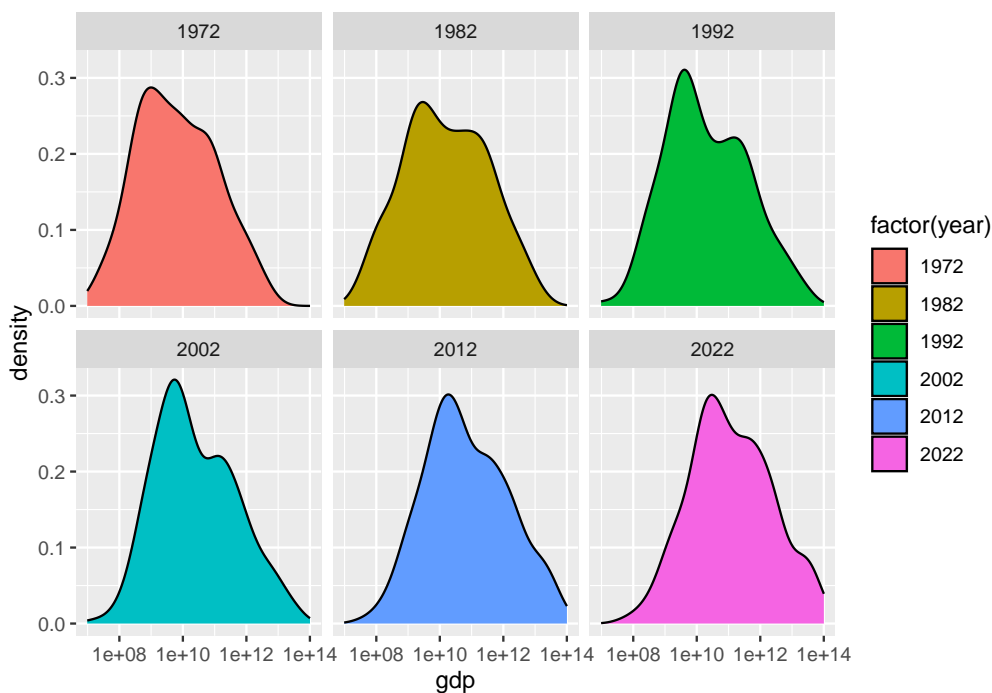
```
df_gdp |> drop_na(gdp) |>
  filter(year %in% c(1962, 1972, 1982, 1992, 2002, 2012, 2022)) |>
  ggplot(aes(gdp, fill = factor(year))) + geom_density(alpha = 0.4) + scale_x_log10()
```



3.2.7.11 Fig 9. 2022 年の世界の国の GDP の分布 (5)

少しみにくいので、分けてみます。

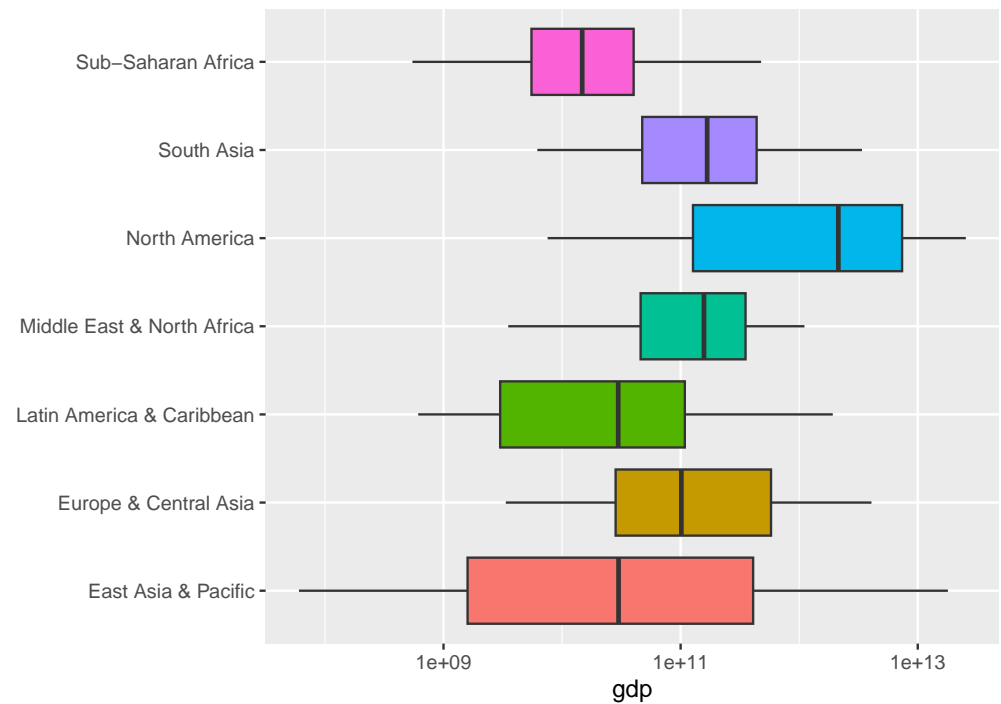
```
df_gdp |> drop_na(gdp) |>
  filter(year %in% c(1972, 1982, 1992, 2002, 2012, 2022)) |>
  ggplot(aes(gdp, fill = factor(year))) +
  geom_density() + scale_x_log10() + facet_wrap(~year)
```



3.2.7.12 Fig 10. 地域ごとの GDP の分布

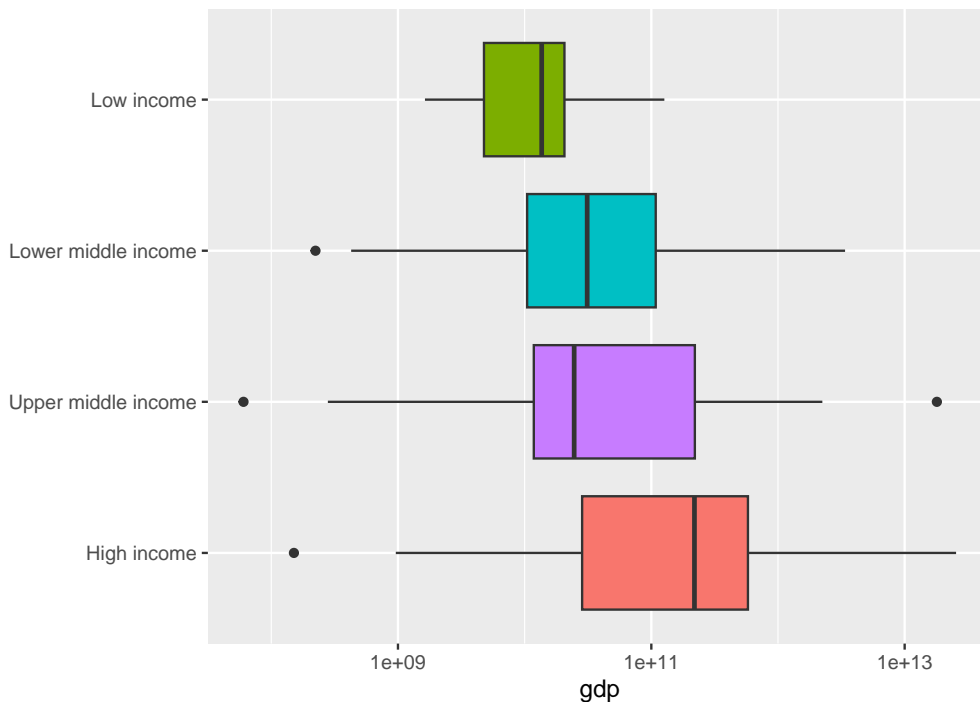
いくつかのグループごとに分布をみても可能です。それには、箱ひげ図 (Boxplot) が有効です。箱ひげ図では、そのグループの国を値の大きさの順にならべて、四分分し、その、真ん中の二つが箱の部分に対応しています。また、真ん中の太い線は、中央値 (median) を表しています。詳しくは、後ほど説明します。

```
df_gdp |> drop_na(gdp) |> filter(region != "Aggregates") |>
  drop_na(region) |> filter(year %in% c(2022)) |>
  ggplot(aes(gdp, region, fill = region)) +
  geom_boxplot() + scale_x_log10() + labs(y = "") +
  theme(legend.position = "none")
```



3.2.7.13 Fig 11. 収入の多寡による分類ごとの GDP 分布

```
df_gdp |> drop_na(gdp) |> filter(region != "Aggregates") |>
  drop_na(income) |> filter(year %in% c(2022)) |>
  mutate(level = factor(income, c("High income", "Upper middle income", "Lower mi
ggplot(aes(gdp, level, fill = income)) +
  geom_boxplot() + scale_x_log10() + labs(y = "") +
  theme(legend.position = "none")
```

これからも、いろいろなことがわかりますね。点は、外れ値を表しています。外れ値についても、きちんと決まっています。収入の多寡（Income Level）のグループは、GNI per Capita という、一人当たりの国民総所得（GNI を人口で割ったもの）をもとに、世界銀行が決めているものです。

3.2.7.14 世界地図の準備

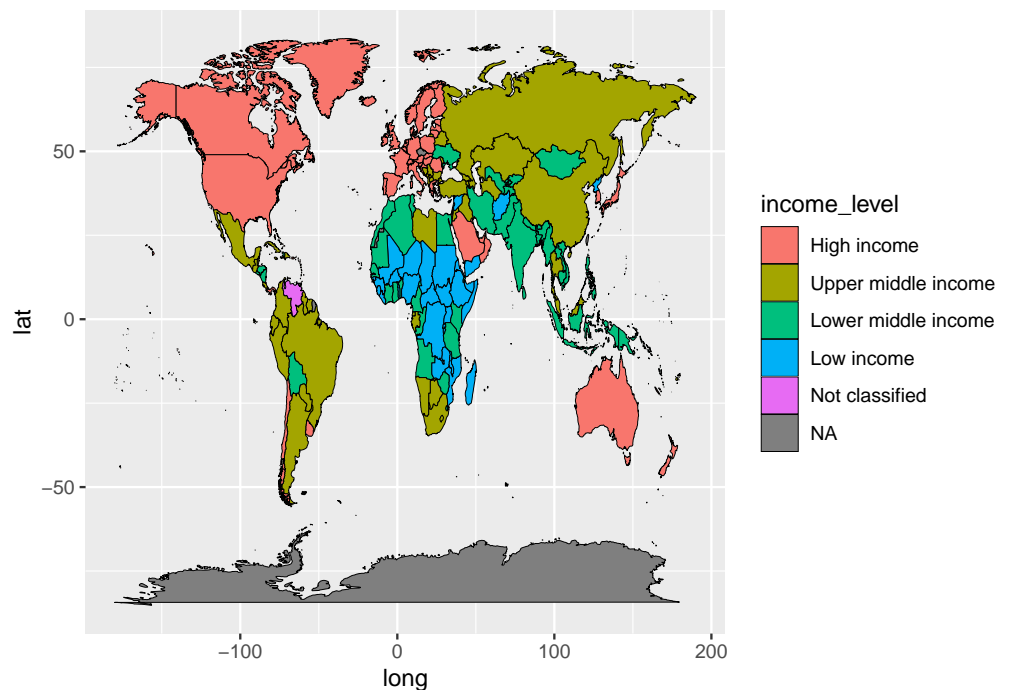
地図で、国の収入の多寡（income level）をみてみましょう。

```
library(maps)
gdp_short <- df_gdp |> filter(year == 2022, region != "Aggregates") |>
  select(iso2c, gdp, income)
map_world <- map_data('world')
map_gdp <- map_world |>
  mutate(iso2c = iso.alpha(region, n=2)) |>
  left_join(gdp_short, by = "iso2c")
head(map_gdp)
#>      long      lat group order region subregion iso2c gdp
#> 1 -69.89912 12.45200     1     1  Aruba      <NA>    AW  NA
#> 2 -69.89571 12.42300     1     2  Aruba      <NA>    AW  NA
#> 3 -69.94219 12.43853     1     3  Aruba      <NA>    AW  NA
#> 4 -70.00415 12.50049     1     4  Aruba      <NA>    AW  NA
#> 5 -70.06612 12.54697     1     5  Aruba      <NA>    AW  NA
#> 6 -70.05088 12.59707     1     6  Aruba      <NA>    AW  NA
#>      income
```

```
#> 1 High income
#> 2 High income
#> 3 High income
#> 4 High income
#> 5 High income
#> 6 High income
```

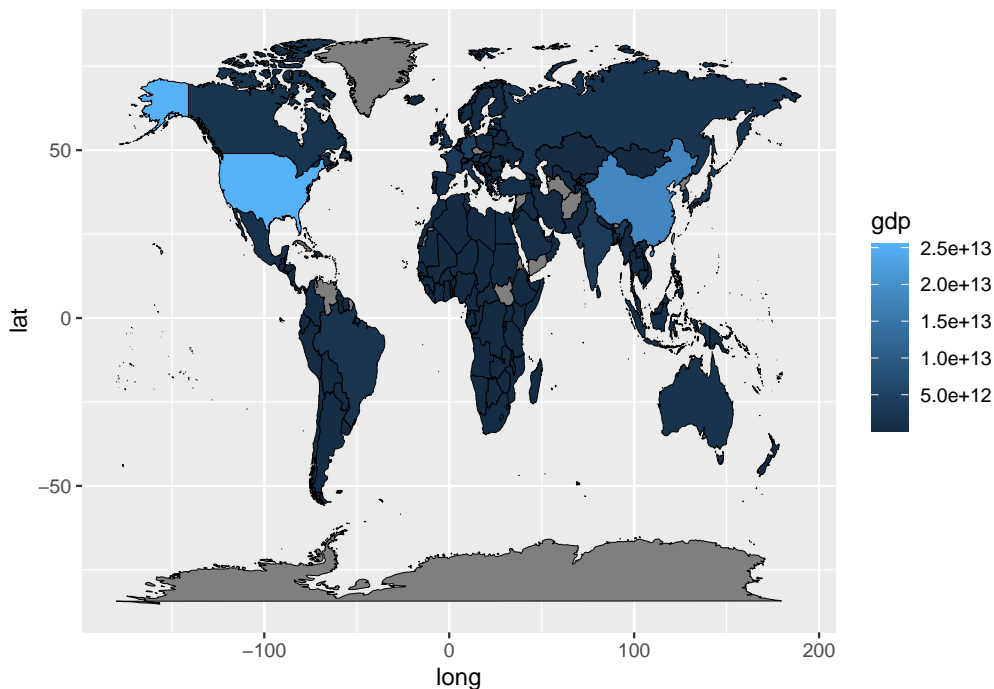
3.2.7.15 Fig 12. Income Level による色分け地図

```
map_gdp |> mutate(income_level = factor(income, levels = c("High income", "Upper
ggplot() +
  geom_map(aes(long, lat, map_id = region, fill = income_level), map = map_world,
#> Warning: Using `size` aesthetic for lines was deprecated in ggplot2
#> 3.4.0.
#> i Please use `linewidth` instead.
#> This warning is displayed once every 8 hours.
#> Call `lifecycle::last_lifecycle_warnings()` to see where
#> this warning was generated.
#> Warning in geom_map(aes(long, lat, map_id = region, fill =
#> income_level), : Ignoring unknown aesthetics: x and y
```



3.2.7.16 Fig 13. GDP による色分け地図

```
map_gdp |>
  ggplot() +
  geom_map(aes(x=long, y=lat, map_id = region, fill = gdp), map = map_world, col = "black", size = 0.1)
#> Warning in geom_map(aes(x = long, y = lat, map_id = region,
#> fill = gdp), : Ignoring unknown aesthetics: x and y
```



3.3 練習

1. それぞれのグラフから、わかったこと、問いなどを列挙してみましょう。
2. Fig 1 の Japan の部分を他の国や、グループ（World など）に変えてみてください。何がわかりますか。
3. Fig 3 の iso2c で選択する、国を変更してください。何がわかりますか。
4. Fig 4 または Fig 5 の iso2c の部分を他の国に変更してください。何がわかりますか。
5. Fig 5 または Fig 6 の、区間の幅や、数を変更してみてください。何がわかりますか。
6. Fig 7, Fig 8, Fig 9 の年を変更してみてください。何がわかりますか。
7. Fig 12, Fig 13 の年を変更してみてください。何がわかりますか。

3.4 プロジェクト

他のデータで、同様のことをしてみましょう。

1. 最初に、gdp = "NY.GDP.MKTP.CD" としましたが、GNI per capita, Atlas method

(current US\$): NY.GNP.PCAP.CD に変えてみましょう。

```
df_gnppcap <- WDI(country = "all",  
  indicator = c(gnppcap = "NY.GNP.PCAP.CD"),  
  extra = TRUE)
```

2. World Development Indicators のサイトの下にある、Data Themes (テーマ) から自分が調べたいテーマを選び、そのテーマから、データコードを取得して、同様の分析をしてみてください。データがあまりない場合もありますので、ある程度データが多いものを選択することをお勧めします。

3.5 まとめ

今回は、経済指標である、GDP を使いました。データサイエンスについて少しずつ、学んでいきます。

コードの説明は、簡単にしかしていませんから、理解するのは難しいと思いますが、いくつかのことは、ご理解いただけたと思います。

- それほど、長くはない、コードで、データを見ていく。R は対話型 (interactive) のプログラミング言語と言われています。
- コードに続けて、結果が表示されるので、コードと出力の対応が見やすい。また、コメントや説明も併記することができる。これは、RMarkdown という形式の中で、コードを書いていることによるものです。RMarkdown は、再現性 (reproducibility) と、プログラム・コードの内容をコンピュータにわかるようにでなく、人間にもわかるように記述する (Literate Programming) を実現しています。
- 視覚化 (Visualization) によって、わかることが色々ある。また、視覚化の方法もたくさんあり、いろいろな見方をすることで、データについての理解が深まっていく。
- 視覚化を通して、データを理解すること、問いを持ち、他の視覚化などを用いて、さらに、理解を深めることがたいせつ。
- 理解したことを元にして、さらに、そのデータ、または、他のデータを使って、新たな発見をしていく。

統計的な指標も用いますが、それらによって、新しい知識を生み出すとも表現しますが、そのような営み全体が、データサイエンスの核をなす部分だと思っています。

第 4 章

世界開発指標とオープンデータ

最初の例として紹介した、はじめてのデータサイエンスでは、世界銀行（World Bank）の世界開発指標（WDI）のひとつである、GDP（Gross Domestic Product 国内総生産）のデータを使いました。「データサイエンスをはじめましょう」では、世界開発指標をこれからも、頻繁に活用しますので、世界開発指標が、どのようなものか確認しておきましょう。

また、世界開発指標に代表されるオープンデータによって誰でも世界の課題を知り、向き合うことができるようになっていきます。そのオープンデータは、どのようなものなのかも、見ていきたいと思います。

4.1 世界開発指標（WDI）

世界開発指標は、貧困撲滅と開発援助とに取り組んでいる金融機関である世界銀行が提供している、オープンデータの一つで、世界銀行は、他にも、たくさんのデータを提供しています。

世界開発指標は、英語では、World Development Indicators（WDI）と呼びますので、以下、省略形の WDI を使います。

4.1.1 概要

まず、簡単に概要を述べておきましょう。

World Development Indicators（WDI）は、世界銀行が開発に関する各国間比較可能なデータの集大成である 1400 の時系列指標（the World Bank's premier compilation of cross-country comparable data on development; 1400 time series indicators）からなり、それらは、テーマ別にも分類されています。

また、さまざまな形式（CSV, Excel など）で、データを検索し、ダウンロードできるように整理されており、さらに、API（Application Program Interface：アプリケーションのプログラムを使ってデータを利用するための仕様）も整備されており、「データサイ

エンスをはじめましょう」で利用するコンピュータ言語 R でも、簡単に、検索やデータのダウンロードが可能なパッケージが提供されています。

4.1.2 三つの入り口

- 世界開発指標 (WDI) (World Development Indicators)
- 世界銀行オープンデータ (World Bank Open Data)
- 世界開発指標(WDI)の統計表 (World Development Indicators Statistical Tables)

三つのサイトのリンクを書きました。一つ一つみていきましょう。

まず、英語を読むことが苦手な場合は、ブラウザ (Google Chrome, Edge, Safari などのホームページ閲覧ソフト) の自動翻訳機能を使うのも良いでしょう。しかし、自動翻訳機能を利用していると、リンク先に飛べなかったり、いくつかの機能が使えないこともありますので、正常に機能しないときは、自動翻訳機能を一旦 OFF にして利用してください。

4.1.3 世界開発指標 (WDI)

まずは、このサイトがおすすめです。上の方に以下のように書かれています。

The World Development Indicators is a compilation of relevant, high-quality, and internationally comparable statistics about global development and the fight against poverty. The database contains 1,400 time series indicators for 217 economies and more than 40 country groups, with data for many indicators going back more than 50 years.

世界開発指標は、世界的な開発と、貧困との闘いに関連する、高品質で、国際的に比較可能な統計をまとめたものです。データベースには、217 の経済と 40 以上の国グループの 1,400 の時系列指標が含まれており、多くの指標のデータは 50 年以上前にさかのぼることができます。

この文章の下には、テーマ別のアイコンがあり、そこから、それぞれのテーマについてみていくことができるようになっています。さらにその下には、最近のニュースやお薦めの記事などが掲載されています。

4.1.3.1 テーマ (さまざまなトピック・課題)

1. 貧困と不平等 (Poverty and Inequality) : 貧困、繁栄、消費、所得分配
2. 人々 (People) : 人口動態、教育、労働、健康、ジェンダー
3. 環境 (Environment) : 農業、気候変動、エネルギー、生物多様性、水、衛生
4. 経済 (Economy) : 成長、経済構造、所得と貯蓄、貿易、労働生産性
5. 国家と市場 (States and Markets) : ビジネス、株式市場、軍事、通信、輸送、テクノロジー

6. グローバルリンク (Global Links) : 債務、貿易、援助への依存、難民、観光、移住

4.1.3.2 それぞれのテーマについて

一つテーマを選んで中身をみてみましょう。テーマを選んで、中身をみてみると、一番上に、最近のトピックなどが書かれ、次に、そのテーマについての概要や課題が書かれています。データの集め方や、どのように標準化するかなどについての説明もあります。

その下に、注目の指標 (Featured Indicators) として、主要な WDI の指標のリストがあり、さらにその下には、データについての説明などが書かれています。

4.1.3.3 注目の指標 (Featured Indicators)

この指標も、いくつかのグループに分けられていますが、それぞれの指標には、以下のような情報があります。

- Indicator 指標の名称
 - マウスを当てると、より詳しい概要をみることができます。
- Code 指標コード
 - R の WDI パッケージなど、API を利用して、データを取得するときに必要なコードです。
 - このコードをクリックすると、World (世界) に関する、この指標のデータをグラフが表示されます。世界に対応するデータがない場合もあります。
- Time coverage どの程度の期間のデータがあるかの概要
 - 1960 年から 10 年刻みでデータがどの程度あるかを確認することができます。
- Region coverage 地域ごとのデータの収集状況の概要
 - ラテンアメリカ、南アジア、サハラ砂漠以南のアフリカ、ヨーロッパと中央アジア、中東と北アフリカ、東アジアと太平洋諸国、北アメリカごとに、データがどの程度揃っているかを確認することができます。
- Get data データの取得
 - API、XLS、CSV、DataBank とあり、データを、取得するための情報や、Excel、コンマで区切られてテキストデータが取得可能で、また、DataBank では、World Bank のデータの、表や、グラフ、地図、各データの情報などを調べることができます。この DataBank の利用については、別の項で説明します。

4.1.3.4 備考

はじめてのデータサイエンスで利用した、GDP の指標は、テーマから経済 (Economy) を選択すると、Featured Indicator (注目の指標) の一番上にあります。まずは、GDP

について、上に書いた、Indicator、Code、Time Coverage、Region Coverage、GetDataを確認することをお勧めします。

このサイトにあるのは、基本的な指標だけです。すべての指標についての情報を得るときには、この次とその次の項目を見てください。しかし、全体では、1400 も指標があり、何をみたらよいか、かえって混乱してしまうかもしれません。まずは、このような基本的なデータで、使い方を把握してから、興味のあるデータを調べるのが良いと思います。

また、基本的な指標を調べると、その指標のページから関連した指標についても調べることができます。

WDI は、すべての年、すべての国についてのデータがあるわけではありません。前もって、期間や、地域ごとにどの程度データがあるかを見ておくことはとても有効です。

それぞれの指標についての概要や特定の国についての情報は、次の項目から見ることも可能です。このページに掲載されている指標で興味をもったものについて、指標の名称や指標コードをコピー・ペーストなどで、記録しておくことをお勧めします。

4.1.4 世界銀行オープンデータ

世界銀行のオープンデータ全体にアクセスできるサイトです。

その下には、いくつかのグラフや、最近のニュースやトピックが書かれており、さらにその下には、More Resources（さらに...）といくつかの項目があります。Open Data Catalog、Data Bank、などと共に、World Development Indicators もあります。この World Development Indicators (WDI) を選択すると、最初に紹介したサイトに飛ぶことができます。また、Data Bank を選択すると、WDI の説明に登場した、ダッシュボードに飛ぶことができ、WDI だけでなく、他のデータについても、調べることができます。

上の検索窓の下に、Country（国）と Indicators（指標）と書いてあります。これらから、WDI について調べることができます。

4.1.4.1 国別サイト

Country（国別）のサイトを選択すると、国のリストが出てきます。

J のところに、Japan（日本）がありますから、選択してみてください。日本のさまざまな指標と小さなグラフが出てきます。

Indicator（指標）、Most Recent Values（直近の値）、Trend（傾向）が表示されます。指標をクリックすると、大きなグラフが出てきます。その指標のサイトですから、そこで、その指標についての他の国の状況などを確認することができます。そこでは、選択した指標と似た指標が選択できたり、もっと詳しいことをしらべるデータバンク（Data Bank）へのリンクもあります。このサイトから、データをダウンロードすることもできるようになっています。

また、国別サイトの右の方には、地域や、経済的状况などによって、グループに分けてある帯もあります。その一番下には、World（世界）もありますから、世界全体について見てみたり、6段階の収入（GNI per Capita：一人当たりの国民総生産）の階級に分けた階級ごとの値を見ることもできるようになっています。

4.1.4.2 指標別サイト

Indicators（指標）を選択すると、トピックに分けて、指標が並んでいます。

よく見ると、一番上に、Featured Indicators（特徴的な指標）と、All Indicators（すべての指標）とあり、最初に開いているのは、特徴的な指標の方であることがわかります。すべての指標（All Indicators）の方を開けると、よりたくさんの指標を見ることができます。

トピックは、以下のものに分かれています。

- Agriculture & Rural Development 農業と農村開発
- Aid Effectiveness 援助の有効性
- Climate Change 気候変動
- Economy & Growth 経済と成長
- Education 教育
- Energy & Mining エネルギーと鉱業
- Environment 環境
- External Debt 対外債務
- Financial Sector 金融セクター
- Gender 性別
- Health 健康
- Infrastructure インフラ
- Poverty 貧困
- Private Sector 民間部門
- Public Sector 公共部門
- Science & Technology 科学技術
- Social Development 社会開発
- Social Protection & Labor 社会的保護と労働
- Trade 貿易
- Urban Development 都市開発

それぞれの指標を選択すると、グラフが表示される画面が出てきます。それは、上で国別のところから選択したものと同じです。

はじめてのデータサイエンスで、GDP を調べるときには、Economy and Growth のトピックにある、GDP (current US\$) の WDI コード NY.GDP.MKTP.CD を指定して、データを取得しました。この、Indicator Code（WDI コード）は、Details（詳細）を見ると、その指標の概要とともに、書いてあります。そのデータコードは、そのページの上の URL にも表示されています。

WDI という R のパッケージを使って、データを読み込みました。そのときに必要だったのが、この Indicator Code (WDI コード) でした。あとで、詳しく調べてみたい指標がありましたら、その指標名 (Indicator Name) と、WDI コード (Indicator Code) を、使えるように、メモなどに貼り付けて (Copy-Paste) おくことを、お勧めします。

4.1.5 世界開発指標 (WDI) の統計表

三つ目の入り口の説明をしましょう。英語名は、World Development Indicators Statistical Tables となっています。WDI の一番下の、Useful Resources (資料) の中にもありますし、それぞれの、テーマの中にもリンクがあります。

世界開発指標 (WDI) の統計表のリンクを開くと、WDI が7つのグループに分けられています。

1. World View 世界の姿
2. Poverty and Shared Prosperity 貧困と富の分配
3. People 人々
4. Environment 環境
5. Economy 経済
6. States and Market 国と市場
7. Global Links 世界の繋がり

WDI のテーマとほぼ一致しています。それぞれのテーマを選択すると、その説明とともに、さらに項目に分かれており、そのデータを見たり、データを PDF または Excel 形式で、ダウンロードして利用できるようになっています。

4.1.6 課題

世界開発指標 (WDI) の、データで、調べてみたいデータコードをいくつか見つけて、書き出してください。あとで利用しやすいように、データ名と、WDI コードをコピーして、メモ帳などに貼り付けておくと良いでしょう。さらに、そのコードに関する情報やリンクを、一緒に記録としておくと、あとで便利です。

4.2 オープンデータ (Open Data)

世界開発指標 (WDI) のように、公開されているデータを、オープンデータと呼びます。世界銀行などの、国際機関や、国際的に活動する非営利団体、それぞれの国の公的機関が、膨大なデータを公開 (オープンに) し、誰でも利用できるようになってきています。

パブリックデータ (Public Data) とも呼ばれますが、正確な定義があるわけではなく、各

機関によって、少しずつ考え方が違う面もあります。しかし、そのリーダーとでもいうべき、世界銀行は、オープンデータについて、厳密に定義をしています。

世界銀行のオープンデータ (Open Data) の定義を見てみましょう。

4.2.1 オープンデータの定義 (Open Data Defined)

オープンデータという言葉は、厳密な意味を持っています。データまたはコンテンツは、出所が明示されオープンという性質が維持されれば、誰でも自由に利用、再利用、再配布できるものを言います。

1. データは法的にオープンでなければなりません。つまり、パブリックドメインに置かれ、最小限の制限で自由に使用できなければなりません。
2. データは技術的にオープンでなければなりません。つまり、誰でも自由に使える一般的なソフトウェアツールを使ってデータにアクセスし、機械で読み取ることが可能な電子フォーマットで提供されていなければなりません。パスワードやファイアウォールによる制限を受けずに、公共のサーバーで、だれでもアクセスできなければなりません。また、オープンデータを見つけやすくするために、さまざまな組織がオープンデータカタログを作成し管理して必要があります。

上の定義で使われている、一つ一つの言葉について、詳細は述べませんが、最初に、初めてのデータサイエンスで、例を示し、この章で、WDI を例にとって説明してきましたので、みなさんも、基本的な部分は、ご理解いただいたと思います。

4.2.2 オープンデータの意義

一旦、足を止めて、みなさんに考えていただきたいと思います。

1. オープンデータは誰に対してオープン（公開）になっているのでしょうか。
2. 公的 (Public) データというとき、公的とはどのような意味でしょうか。
3. なぜ、膨大な公的データのオープン化が進み、たいせつにされているのでしょうか。

4.2.3 オープンデータの利活用

さまざまな国際機関では、データを、Excel 形式や、CSV (Comma Separated Values) 形式などで、提供する以外に、ダッシュボード形式で、グラフを生成するなどして、データの可視化をある程度できるようにしています。さらに、コンピュータのアプリケーションでデータを直接取得できるように、API (Application Program Interface) を提供しています。

世界開発指標 (WDI) は、この章で説明してきたように、世界の状況やそれぞれの国についての、1400 余のさまざまな指標についてのデータを、長期間にわたって (1960 年ごろ

から) 提供している基本的なデータベースで、最初に調べてみることをお勧めします。さまざまな課題を理解するためにも、たいせつだと思います。

さらに、WDI は、この膨大な指標についてのデータすべてを統一した形式で提供するという特徴があります。したがって、複数の指標についての関係を調べることも、容易にできます。実は、データサイエンスを始めると、違ったデータを一つにまとめて分析することは、技術的にそれなりに困難があります。それが、WDI では、ほとんどないのです。そのために、非常に使いやすいデータベースになっています。

データを、ダウンロードして、読み込むことが必要になりますが、API が整備されていると、分析ソフト（「データサイエンスをはじめましょう」では R を中心的に使いますが）に直接、取り込むことができるので、データの出所も明確で、他の人も同じように取得できれば、データ自体の、Reproducibility（再現可能性）も担保できます。はじめてのデータサイエンスでも、すべてのコード（コンピュータプログラム）が書かれていたと思います。基本的には、これを、たとえば、コピー、ペーストすれば、皆さんのコンピュータでも同じ結果が得られるということです。これについては、まだ、みなさんは、実感が持てないかもしれませんね。これから、少しずつ、説明していきます。

プログラミング言語 R を使い始める前に、データに慣れることも大切だと思いますから、次には、さまざまな、オープンデータについて、簡単な説明と共に、付随している、ダッシュボードの利用の仕方も説明しながら、みていきたいと思います。

第 5 章

データサイエンスノートブック

データサイエンスの記録について書きます。また、そのために、必要なツールについても、少しだけ説明します。

この「データサイエンスをはじめましょう」では、R で、自分でコードを書き、R Markdown や、Quarto に記録し、データサイエンスを進めていくことを目指しています。しかし、第一部では、まず、データを見ることに慣れるために、それぞれのサイトが提供する、ダッシュボードを使ってデータを見ていくことを、紹介します。

そのときにも、ノートを作成し、記録をとっていくことは、とても大切です。その説明を少しだけ書きます。

5.1 再現性のために記録すべきこと

根拠を明確にする（evidence based）ことが、データサイエンスにおいて、必須であることは、すでに、書きました。これから、データを見て行きますが、そのときに、基本的な情報を、記録をしておくことをお勧めします。それが、今後のためにも有用ですし、その習慣をつけることが大切だからです。いくつか、記録すべき項目を書いておきます。

1. 日付：そのデータを調べた日付を書いておきます。サイトの内容が変更になる場合もあります。
2. データ名：もし、そのデータを特定するデータコードがあれば、それも記録しておきましょう。
3. データリンク：インターネット上のアドレスです。ブラウザ（Google Chrome、Edge、Safari などのホームページ閲覧ソフト）の上の窓に、URL（Universal Resource Locator）が表示されますから、それを記録しておきましょう。データ自体の URL を取得できる場合もありますが、そのデータが置かれている、ページ（Website）の URL だけが、取得でき、ダウンロードボタンでダウンロードする形式になっている場合もあります。その場合、右クリックや、Ctrl+Click で、データの、URL が取得できる場合もありますが、できない場合もあります。

- データをダウンロードした時は、そのファイル名と、ダウンロードした日付も忘れずにかいておくことをお勧めします。
4. メタデータ：また説明しますが、データには、データについてのデータ（これをメタデータと言います）が付いていることが多いです。最初からすべて記録する必要はありません。上の、データリンクがあれば、必要な時に、戻ることができます。しかし、データの定義や、変数の定義、データの収集方法などによっては、自分が求めているものではなかったり、データ自体がオリジナルデータではない場合もあります。最初から、詳細を記録する必要はありませんが、注意をはらうことを習慣にしておくことは大切です。
 5. ダッシュボードを設定したパラメーター（何を意味するかは少しずつ説明して行きます）：ていねいに、記録するのは、大変ですが、再現には、パラメーター情報は、必要です。場合によっては、そのリンクが提供され、そのリンクを使って、同じものが再現できる場合もあります。あるいは、埋め込み (embed) するための iframe link というものが提供されている場合もあります。もし、それがあれば、記録しておいてください。実際には、HTML 文書に埋め込んでも、そのままでは表示できない場合もありますが、設定値が含まれていますので、少し慣れれば、再現することも可能です。
 6. コメント：そのデータからわかったこと、疑問点、さらに知りたいことなど。少しでも書いてあると、あとで、とても便利です。このようなものが、データサイエンスの核でもありますから、ぜひ、記録しておいてください。

5.2 技術的なこと

5.2.1 ブラウザーの言語と様々な翻訳機能

1. ブラウザーの言語：実は、コンピュータのシステム言語が関係しますが、ほとんどの場合、システムの言語は変更せずに、ブラウザの言語を変更できるようになっています。
 - Windows でも、Mac でも、Google Chrome が使えますから、Google Chrome で説明すると、「Google Account を管理」から、設定できます。
 - Google Public Data Explorer と検索してみてください。すると、日本語の場合には、ほとんど出てきませんが、英語だと膨大なデータがあります。このサイトの場合には、右上に言語とでますから、言語を英語に変更すると、たくさんのデータを見ることができます。検索エンジンの言語によって、表示されるものが、大きく変化しますから、わたしは、Google のアカウントを複数使って、それぞれで、違う言語設定にしています。言語を変更できるようになると、検索の世界がとても広がります。
2. ブラウザーの翻訳機能：最近のブラウザには、翻訳機能が付属しており、簡単に切り替えることができます。しかし、ブラウザによって、設定方法が異なります

ので、調べて、いつでも使えるようにしておいてください。必要に応じて、翻訳機能の ON/OFF ができるととても便利です。

3. DeepL などのアプリの翻訳機能：最近、ブラウザーの翻訳機能の質も向上しているので、不要かもしれませんが、わたしは、DeepL も併用しています。サイト上でも、翻訳ができますし、アプリをダウンロードして、設定を確認すると、翻訳したい箇所を選択すれば、すぐ翻訳してくれるショートカット機能があります。
4. ChatGPT など AI の翻訳機能：最近、LLM (Large Language Model) の発達で、様々な AI による、翻訳の質も非常に向上しています。ここでは、ChatGPT と書きましたが、他にも様々な AI で、翻訳が可能です。長い文章は、字数を指定して要約をしてもらうことも可能ですから、慣れると世界が広がって行きます。

5.2.2 画像

1. グラフのダウンロード：グラフ (graph, chart とも言います) は、画像になっていますから、あとで、利用する場合は、リンクを取得して、そのリンクで同じものを開くことができる場合もありますが、ダウンロードして保存しておくほうが安全です。ダウンロード方法が書いてあったり、右クリックまたは、Ctrl+ クリックで、ダウンロードできる場合が多いと思います。また、ダウンロードしたあとに、ノートに貼り付けておくことができれば、そのほうがあとで利用するときに便利です。取得したサイトの URL や、取得した日付も記録しておくことをお勧めします。
2. 画面収録 (Screen Capture)：ダウンロードできない場合、その方法が見つからない場合は、画面収録も一つの方法です。Windows, Mac によって方法が異なりますから、あらかじめ調べておくとい良いでしょう。

5.2.3 データファイル

データファイルは、CSV (comma separated values カンマで区切られたテキストデータ) などのテキストデータ、Excel ファイル、または、これらを、圧縮したり、いくつかのファイルをまとめて圧縮したりしてある場合があります。以下、少しだけ、注意点や、確認すべき点を書いておきます。

1. 圧縮されている場合の解凍方法を確認しておくこと。Windows か Mac でも違いますから、解凍方法を確認しておいてください。
2. CSV が一般的ですが、他にも、区切り文字が、スペースだったり、TAB だったり、縦棒だったり、様々な形式があります。R を使うようになれば、どの形式であっても、読み込めますし、変換することも可能ですが、二種類の問題が一般的です。
 - 上に書いた区切り文字の違い
 - Encoding の問題。こちらは、日本語などを含むファイルではよく起こる問題です。いわゆる文字化けが生じて中身が読めない場合があります。

- すべての対策を書くことはできませんが、区切り文字の違いは、Excel の機能でも、解決できます。Microsoft Office も高額なソフトですから、持っていないという場合は、機能限定ですが、Online 版は無料で使えますから、試してみてください。Google Spreadsheet でよみこいむことができる場合もあります。

5.3 まとめ

本書で紹介する、R を使い始めれば、統一した方法で解決できる課題もありますが、記録を取るということは、基本的ですから、簡単に書きました。

上のような記録を何に書くかということは書きませんでした。基本的には、記録を取ることが大切で、何に記録することはあまり、問題ではありません。しかし、リンクを貼り付けて、すぐ開くことができたり、画像を貼り付けたりが、できると便利でしょう。

使い慣れたものを使ってください。できれば、どこでも使えるような Cloud 型のサービスがお勧めです。他の方におすすめを聞くのも良いかもしれません。

次からは、実際に、オープンデータのサイトに行って、データを見ることを経験して行きたいと思います。

第Ⅰ部

第一部オープンデータ

第 6 章

オープンデータ

6.1 概要

すでに、オープンデータについて、説明をし、その例として、世界銀行の世界開発指標 (World Development Indicators (WDI)) の説明をし、その指標の一つである、GDP (Gross Domestic Product 国内総生産) を例にとって、はじめてのデータサイエンスというタイトルで、データサイエンスの実際を見ました。

第一部では、世界銀行や、経済協力開発機構 (OECD)、国際連合 (United Nations) などの、国際機関のオープンデータや、日本のデータとして、政府機関統計 (e-Stat) の紹介をします。

さらに、実際に、それぞれのサイトのオープンデータを、ダッシュボード (dashboard) 呼ばれる対話型 (Interactive) の機能の使い方を紹介しながら、データを見ていきます。

また、実際のデータの取得 (ダウンロード) にも触れ、それぞれの期間が提供する API (アプリケーションプログラムインターフェース) の活用についても、簡単に触れていきます。

第二部で、R を使ったデータサイエンスについて学びますが、その例においても、いくつかのオープンデータを用いる機会がありますので、第一部は、その準備としての位置付けです。オープンデータの分析のより詳しい説明は、第三部で行います。

R を使ったデータサイエンスを、早く学びたい方は、この第一部をスキップして、第二部から読んでくださって構いません。しかし、みなさんの中には、R を使うところまでできるかは自信がないけれど、実際のオープンデータを見て、データサイエンスではどのようなことを考えるのかを体験してみたいという方もおられると思い、第一部を書いています。

また、実際に、オープンデータを活用して、ある課題について調べるときには、基本的なオープンデータについての知識は有用です。どこに、そのようなデータがあるかを見つけることがたいせつであるとともに、データサイエンスはなんと言っても、理論ではなく、実証的なものですから、実際のデータに触れながら、学んでいくのがたいせつだと考えているからです。

第一部の学びを通して、こんなふうに、データがたくさん、公開されているのか、と、そんな感触を持っていただければと思います。

6.2 さまざまな機関のオープンデータ

すでに少しだけ紹介した世界銀行以外にも、多くの機関がデータを提供しています。いずれも、使いやすくなってきています。少しずつ、いくつかのデータベースに、アクセスして、できれば、API の利用の仕方も、習得して、データを調べることに少しずつ慣れていただければと思います。

下のリストは、私が個人的に、何度か使ったことのある、データベースです。「データサイエンスをはじめましょう」でも、少しずつ消化していきます。他にも、たくさんのデータベースがありますので、ぜひ、調べてみてください。

- 国際連合 UNdata
- 経済協力開発機構 OECD data、
- 世界格差データベース（World Inequality Database）WID、
- 欧州連合の統計局 Eurostat、
- データで見る私たちの世界 Our World in Data

6.2.1 日本のデータ

日本では、それぞれの政府機関でデータを膨大なデータを公開していますが、次のサイトで、それらをまとめて公開しています。

- e-Stat：政府統計の総合窓口 - 統計で見る日本
- ダッシュボード：対話型形式でのデータの視覚化

6.2.2 持続的開発目標（SDGs）データ

持続的開発目標（Sustainable Development Goals（SDGs））は、皆さんもご存知かと思います。2015年に国際連合で定められ、2030年までの達成目標が掲げられています。様々なサイトがありますから、ご存知かと思いますが、SDGsは、17の分野に分かれており、それぞれにさらに細かい分類があり、その一つ一つに、数値目標が掲げられています。数値目標だけに、依存することは、十分ではないかもしれませんが、少なくとも、その数値目標の達成を検証するために、データが公開されています。

自分が興味をもっている課題に絞って、各サイトでどのようなデータが提供されているかをみていくのもよいと思います。

- 世界銀行：[https://databank.worldbank.org/source/sustainable-development-goals-\(sdgs\)](https://databank.worldbank.org/source/sustainable-development-goals-(sdgs))
- 国連：<https://unstats.un.org/sdgs/dataportal/database>

- データで見る私たちの世界：Our World in Data SGD Tracker: <https://sdg-tracker.org/>
- SDG Index: <https://dashboards.sdgindex.org/map>

6.3 オープンデータについて

オープンデータについては、すでに、世界銀行の定義をご紹介しました。パブリックデータをいう言葉も使われます。それらの違いなども含めて、AI に聞いてみましょう。

Q. オープンデータ（Open Data）とか、パブリックデータ（Public Data）と言われるものは、どのようなものですか。定義のようなものと、いくつかの例、さらに、重要視されてきている背景を説明してください。

GPT4. (2023.8.15)

オープンデータ（Open Data）とは、誰でも自由に利用、再配布、再利用することができ、制限がない、または限定的な制限だけが存在するデータを指します。オープンデータは一般的には無償で提供され、機械可読形式（たとえばCSV、JSON など）で利用可能であり、再利用のためのライセンスが明確にされています。

一方、**パブリックデータ（Public Data）**は公共の手段で取得されたデータ、または公共の利益のために公開されるべきデータを指します。これは政府や公共機関、研究機関などが収集したデータで、しばしば公共の利益のために公開されます。

オープンデータとパブリックデータの主な違いは、オープンデータが再利用と再配布の観点から「オープン」であることを強調するのに対し、パブリックデータは単に公開されているだけで、再利用や再配布については必ずしも明示されていないことです。

例えば、以下のようなものがあります：

1. 気象データ：天気予報、気温、風向、風速、降水量など
2. 交通データ：交通渋滞、駐車場の利用状況、公共交通機関の運行状況など
3. 統計データ：人口、雇用、健康、教育などのデータ
4. 環境データ：汚染レベル、森林破壊、気候変動などのデータ
5. 地理データ：地図、土地利用、地質学的情報など

オープンデータとパブリックデータが重要視されてきた背景には、以下の要素があります：

1. **透明性と説明責任**：データが公開されることで、政府や企業の行動についての透明性が高まり、説明責任が強化されます。
2. **イノベーションと経済的価値**：データが自由に利用できることで、新たなビジネスモデルや製品、サービスの開発が促進され、経済的価値が創出されます。

3. **市民参加:** データへのアクセスが提供されることで、市民が社会的・政治的な決定プロセスに参加する機会が増えます。
4. **社会的・環境的課題の解決:** 大量のデータを解析することで、社会的・環境的な課題をより良く理解し、解決策を見つけることが可能になります。

以上のように、オープンデータとパブリックデータは、透明性、説明責任、イノベーション、市民参加、社会的・環境的課題の解決などの観点から重要性を増しています。

6.3.1 オープンデータの未来

みなさんは、どのように考えますか。まだ、発展途上で、特に、パブリックデータについては、国によって、考え方も一定ではないように見受けられますが、これから、どのような世界を目指していくのか、データサイエンスは、どのような目的のために必要なのか。みなさんは、なんのために勉強しているのか、上の一つ一つの項目をもとに、周囲のみなさん、友人などと、ディスカッションをしながら、考えていただければと思います。

いずれもたいせつな項目だと思いますが、項目3の市民参加は、公開されているデータをわたしたち市民が利用して初めて可能になることです。項目4の、社会的・環境的課題の解決には、私たち市民も責任をもって、取り組むべきだということが基盤にあるのだと思います。みなさんは、どう考えられますか。

6.4 ダッシュボード

はじめに、オープンデータを利用するときには、対話型のダッシュボードは、助けとなります。ダッシュボードとはどのようなものか、AIに説明してもらいましょう。

Q. データを対話的にみたり視覚化したりする、ダッシュボード (dashboard) について説明してください。

ChatGPT (2023.08.15)

ダッシュボード (dashboard) は、情報やデータを一元化し、対話的に表示・分析・視覚化できるインタラクティブなユーザーインターフェースです。主にビジネスや組織内で使用され、重要な指標や情報をリアルタイムに監視・分析するために利用されます。

ダッシュボードは、さまざまなデータソースからの情報を収集し、それを一つの場所にまとめて表示します。これにより、膨大なデータを迅速に把握し、傾向やパターンを特定することが可能になります。例えば、売上データ、顧客の統計情報、ウェブ解析データ、ソーシャルメディアの指標など、さまざまなデータソースからの情報を一つのダッシュボードに表示することができます。

ダッシュボードは、視覚化ツールを使用してデータをグラフやチャート、表などの形式で表示します。これにより、データを直感的に理解しやすくし、視覚的なパターンや相関関係を把握することができます。例えば、売上の推移を折れ線グラフで表示したり、地域別

の売上を地図上に表示したりすることができます。

ダッシュボードは、利用者が必要な情報に素早くアクセスできるように設計されています。ユーザーは、ダッシュボード上の要素をクリックしたり、フィルタを適用したり、グラフをズームしたりすることで、詳細な情報を取得したり特定の視点でデータを分析したりすることができます。

また、ダッシュボードはリアルタイムな情報の監視にも利用されます。データソースが更新されるたびに、ダッシュボード上の情報も自動的に更新されます。これにより、組織内の関係者が最新の情報にアクセスし、迅速な意思決定を行うことができます。

ダッシュボードは、情報の可視化と分析によって、効率的なデータ駆動型の意思決定を支援する役割を果たします。組織内のさまざまなレベルの利用者が、ビジネスのパフォーマンスを把握し、問題や機会を特定し、戦略やアクションプランを立案するために活用します。

Q. 例や、ダッシュボードのサイトを教えてください。

ChatGPT (2023.8.15)

もちろんです！以下にいくつかのダッシュボードの例と、一般的に使用されるいくつかのダッシュボードサイトをご紹介します。

ダッシュボードの例:

1. Google Analytics: ウェブサイトのトラフィックやユーザー行動などの情報をリアルタイムに監視・視覚化するためのダッシュボードです。
2. Salesforce: セールスや顧客関連のデータを統合し、ビジネスのパフォーマンスを追跡するためのダッシュボードです。
3. Tableau: データの視覚化と分析に特化したダッシュボード作成ツールであり、さまざまな業界で利用されています。

ダッシュボードのサイト:

1. Klipfolio (<https://www.klipfolio.com/>)
2. Geckoboard (<https://www.geckoboard.com/>)
3. DashThis (<https://dashthis.com/>)
4. Databox (<https://databox.com/>)
5. Microsoft Power BI (<https://powerbi.microsoft.com/>)

これらのサイトでは、ダッシュボード作成ツールやサービスを利用して、カスタムのダッシュボードを作成することができます。それぞれのサイトには、詳細な情報やデモが提供されている場合もありますので、ぜひご覧ください。

上で、ダッシュボードのサイトとあるのは、これらのサイトのツールを使って、ダッシュボードを作ることができるというものです。少し、ダッシュボードの利用に慣れてきたら、挑戦してみてもいいかもしれません。

6.5 API を利用したパッケージ

それぞれのオープンデータを検索したり、読み込んだりするための R のパッケージの紹介をします。上に書いたように、第二部で利用するためです。詳しい使い方は、第三部で紹介します。

基本的には、以下の項目について紹介します。

- R の（公式）パッケージサイト：パッケージも含め、R の管理をしている CRAN の公式サイトの情報です。この下の情報も、ほとんどの場合、このリンク先に掲載されています。
 - 資料（Materials）：README（はじめにお読みください）などで、基本情報が書かれています。
 - マニュアル（Manual）：利用者用説明書です。パッケージで利用可能な関数（命令、データ）などの情報がすべて書かれてあります。それぞれの関数（命令）に関しては、R Studio 内の Help（ヘルプ検索窓）からも利用可能です。
 - 使い方の例（Vignette）：開発者のサイト（GitHub（バージョン管理システムの支援サイト）など）にある場合もありますが、最近は、公式パッケージサイトに、Vignette として使い方の例が掲載されている場合が増えてきています。すべて理解しようとせず、使い方の例からまずはみてみるのが有効です
-

第 7 章

世界銀行（World Bank）

7.1 概要

世界銀行は、貧困削減と持続的成長の実現に向けて、途上国政府に対し融資、技術協力、政策助言を提供する国際開発金融機関です。2030 年までに極度の貧困をなくし、各国の下位 40 パーセントの人々の所得を引き上げて繁栄の共有を促進するという 2 つの目標を掲げています。

世界銀行のサイトを見てみると、下の方にはグラフも出ており、さまざまな統計データを提供することに力を入れていることがわかります。

すでに、世界開発指標については、簡単に説明しました。ここでは、世界銀行のサイトの中のデータについて、見ていくとともに、ダッシュボード（dashboard）の活用、データの取得方法や、API（Application Program Interface）を用いて、R でデータを検索したり、取得するパッケージの紹介を簡単にします。

7.2 データベース

7.2.1 三つのサイト

- 世界銀行オープンデータ
- データカタログ
- 世界開発指標（WDI）

一つ一つみていきましょう。

英語を読むことが苦手な場合は、ブラウザ（Google Chrome, Edge, Safari などのホームページ閲覧ソフト）の自動翻訳機能を使うのも良いでしょう。しかし、自動翻訳機能を利用していると、リンク先に飛べなかったり、いくつかの機能が使えないこともありますので、正常に機能しないときは、自動翻訳機能を一旦 OFF にして利用してください。

7.2.2 世界銀行オープンデータ

世界銀行のオープンデータ全体にアクセスできるサイトです。上の検索窓の下に、Country (国) と Indicators (指標) と書いてあります。

その下には、いくつかのグラフや、最近のニュースやトピックが書かれており、さらにその下には、More Resources (さらに...) といくつかの項目があります。Open Data Catalog、Data Bank、などと共に、World Development Indicators もあります。この World Development Indicators (WDI) を選択すると、上の三つのサイトの三番目に飛びます。また、Data Bank を選択すると、三つのサイトの二番目に飛びます。

最初に書いた、Country (国) と Indicators (指標) から説明しましょう。

7.2.2.1 国別サイト

Country (国別) のサイトを選択すると、国のリストが出てきます。

J のところに、Japan (日本) がありますから、選択してみてください。日本のさまざまな指標とグラフが出てきます。

Indicator (指標)、Most Recent Values (直近の値)、Trend (傾向) が表示されます。指標をクリックすると、大きなグラフが出てきます。その指標のサイトですから、そこで、その指標についての他の国の状況などを確認することができます。そこでは、選択した指標と似た指標が選択できたり、もっと詳しいことを調べるデータバンク (Data Bank) へのリンクもあります。このサイトから、データをダウンロードすることもできるようになっています。

また、国別サイトの右の方には、地域など、グループに分けてある帯もあります。その一番下には、World (世界) もありますから、世界全体について見てみたり、収入の階級に分けたサイトの指標を見ることもできるようになっています。

7.2.2.2 指標別サイト

Indicators (指標) を選択すると、トピックに分けて、指標が並んでいます。

よく見ると、一番上に、Featured Indicators (特徴的な指標) と、All Indicators (すべての指標) とあり、最初に開いているのは、特徴的な指標の方であることがわかります。すべての指標の方を開けると、よりたくさんの指標を見ることができます。

トピックは、以下のものに分かれています。

- Agriculture & Rural Development 農業と農村開発
- Aid Effectiveness 援助の有効性
- Climate Change 気候変動
- Economy & Growth 経済と成長

- Education 教育
- Energy & Mining エネルギーと鉱業
- Environment 環境
- External Debt 対外債務
- Financial Sector 金融セクター
- Gender 性別
- Health 健康
- Infrastructure インフラ
- Poverty 貧困
- Private Sector 民間部門
- Public Sector 公共部門
- Science & Technology 科学技術
- Social Development 社会開発
- Social Protection & Labor 社会的保護と労働
- Trade 貿易
- Urban Development 都市開発

それぞれの指標を選択すると、グラフが表示される画面が出てきます。それは、上で国別のところから選択したものと同じです。

GDP を調べるときには、NY.GDP.MKTP.CD という、コードを指定して、データを取得しました。この、Indicator Code（WDI コード）は、Details（詳細）を見ると、その指標の概要とともに、書いてあります。そのデータコードは、そのページの上の URL にも表示されています。

はじめてのデータサイエンスでは、WDI という R のパッケージを使って、データを読み込みました。そのときに必要だったのが、この Indicator Code でした。あとで、詳しく調べてみたい指標がありましたら、その、WDI コード（Indicator Code）を、あとで、使えるように、メモなどに貼り付けて（Copy-Paste）おくことを、お勧めします。

7.2.3 オープンデータカタログ（Open Data Catalog）

The Data Catalog is designed to make World Bank's development data easy to find, download, use, and share. It includes data from the World Bank's microdata, finances and energy data platforms, as well as datasets from the open data catalog. There are different ways to access and download datasets.

データカタログは、世界銀行で編纂した開発に関するデータを簡単に検索、ダウンロード、使用、共有できるように設計されています。これには、世界銀行のマイクロデータ、財務、エネルギーデータプラットフォームからのデータ、およびオープンデータカタログからのデータセットが含まれています。データセットにアクセスしてダウンロードするには、さまざまな方法があります。

世界銀行（World Bank）で編纂したり、他の機関から提供を受けたデータがリストされ

ています。

一番上には、Search Box (検索窓) があり、その下には、Featured (特徴的な、またはお薦め) とあり、いくつかのトピックが並んでいます。右に、スクロールするとさらにいくつかのトピックを見ることができます。その中でも、上で述べた世界開発指標 (WDI) もありますし、Covid-19 (コロナウイルス感染症) 関連のデータもあります。

それぞれの、トピックに、関連のデータがリストされています。

7.2.4 世界銀行 (World Bank)

このページの最初にも書きましたが、簡単にまとめておきましょう。

- 世界銀行 (World Bank) : <https://www.worldbank.org>
- 世界銀行について (Who we are) :
 - 極度の貧困状態の削減 (To end extreme poverty) : 2030 年までに、極度の貧困状態にある世界人口の割合を 3% に削減する。By reducing the share of the global population that lives in extreme poverty to 3 percent by 2030.
 - 繁栄を共に享受 (To promote shared prosperity) : すべての国の最貧困層の 40% の人々の所得を増加させることによって共栄を促進。By increasing the incomes of the poorest 40 percent of people in every country.
- 世界銀行オープンデータ (World Bank Open Data) : <https://data.worldbank.org>
 - Data Bank, World Development Indicators, etc.

日本について : <https://data.worldbank.org/country/japan?view=chart>

7.2.5 世界開発指標 (World Development Indicator)

すでに紹介しましたが、簡単にまとめておきます。

- World Development Indicators (WDI) : 世界銀行が開発に関する各国間比較可能なデータの集大成である 1400 の時系列指標 (the World Bank's premier compilation of cross-country comparable data on development; 1400 time series indicators)
 - テーマ別 (Themes) : 貧困と格差、人間、環境、経済、国家と市場、グローバルリンク集 (Poverty and Inequality, People, Environment, Economy, States and Markets, Global Links)
 - オープンデータとデータバンク (Open Data & DataBank) : Explore data, Query database
 - すべてのデータおよびメタデータを Excel または CSV 形式で、一括してダウンロードすることもできるようになっています。Bulk Download:
 - コンピュータを使って読み込む場合のデータの仕様が書かれています。API (Application Program Interface) Documentation

1400 ものデータがありますから、すべてのデータやメタデータをダウンロードすれば、すぐ、データを調べることができるわけではありません。基本的なことをおく必要がありますね。上にもリンクのある、データカタログから、世界開発指標 (World Development Indicators) を選択すると、次のサイトにリンクがついています。

このページからは、Databank (ダッシュボード) へのリンクなどの他、Excel ファイルや、CSV ファイルで、メタデータを取得することもできるようになっています。WDI 全部のリストもここで見るができます。このファイルから探すのが最適とは言えませんが、そのようなファイルを持っておくことは便利です。

7.3 ダッシュボード (Dashboard)

世界銀行のダッシュボードには二種類あります。一つは、それぞれの指標についてのダッシュボード、もう一つは、データバンク (DataBank) です。

7.3.1 World Bank アカウント

ダッシュボードは、リンクにアクセスすれば、アカウントなどを作成せず、すぐに使えます。しかし、ダッシュボードを使って、グラフを作成したり、自分用の、データを作成したりした場合には、その結果を、保存をしたり、リンク (iframe 形式) を、文書に埋め込んだりする必要を感じる場合があります。その場合には、一般用アカウントを作成する必要があります。このサイトの Sign Up から、アカウントを作成してください。

7.3.2 指標毎のダッシュボード

はじめてのデータサイエンスで使った、GDP (Current US\$)、データコード NY.GDP.MKTP.CD についてみましょう。

いくつかの方法があります。

1. 世界銀行オープンデータ から、指標 (Indicator) を選択し、その中の Economy & Growth (経済と成長) の中から、GDP (current US\$) を選択すると、ダッシュボードが現れ、世界の GDP の推移のグラフが表示されます。
2. 世界銀行オープンデータ から、国 (Country) を選択し、例えば、J から、Japan を選択し、Economics (経済) の指標の GDP (current US\$) を選択すると、ダッシュボードが現れ、日本の GDP の推移のグラフが表示されます。指標によっては、上の、Theme (テーマ) や、Topic (トピック) から選ぶ必要がある場合もあります。
3. 世界開発指標 (WDI) の Data Theme (テーマ) の中の、ECONOMY (経済) の中から、GDP (current US\$) を選択すると、ダッシュボードが現れ、世界の GDP の推移のグラフが表示されます。
4. もし、WDI コード (この場合は、NY.GDP.MKTP.CD) を知っていれば、このコードを、世界銀行オープンデータの検索窓に入れて検索すると、上のダッシュボードのページが表示されます。

最初は、選択した指標について、世界か、日本など選択した国の、折れ線グラフ (Line Graph) が表示されていると思います。そして、下の方に、国のリストがあり、その一番下には、地域のリストが続いています*1。また、Line (折れ線グラフ) と書いた右には、Bar (棒グラフ) と、Map (地図) とあります。

その右には、Also Show (追加) とあり、Aggregates (総合)、Same Region (同じ地域)、Similar Values (近い値)、Highest Values (最高値)、Lowest Values (最小値) を表示することもできます。表示しないときは、None を選択します。

その右には、Share (共有) と、Details (詳細) があります。詳細には、その指標についての、詳しい説明があります。すべてを理解することはできないかもしれませんが、翻訳機能も使って、確認しておくことをお勧めします。詳細には、データコードも書かれています。記録しておく、次に同じ指標のデータを探すときに便利です。共有からは、Web Page や、SNS に埋め込む、iframe link を取得することができます。

右の方の帯には、関連する指標がリストされ、さらに、Download (ダウンロード)、Data Bank (データバンク)、WDI Tables (統計表) へのリンクがあります。

7.3.2.1 使い方

具体例としては、GDP (Current US\$) の世界 (World) のグラフが表示されているとします。上の検索窓に GDP (Current US\$) と入っていると思います。

- 国や地域の追加
 - 英語で入力しますから、グラフの下、国や地域名のところから、追加したい、国や地域名を選んで、書き出しておいてください。一つずつ追加するときは、コピーするのも良いかもしれません。
 - 日本をグラフに追加するときは、Japan ですから、検索窓に、Japan と入れて少し待ちます。すると、窓の下に、Japan が表示されますから、それを選択 (クリック) します。すると、日本のグラフに変わります。India も加えてください。次に、検索窓に、United と入力すると、United Kingdom, United States, United Arab Emirates が表示されますから、United Kingdom を選択してください。Japan, India, United Kingdom の三つの国のグラフが同時に表示されます。これによって、いくつかの国のその指標における経年変化を比較することができます。
 - いろいろな国や地域を加えてみてください。たとえば、GDP を指標として、United States を加えると、United States の GDP の値が大きいため、他の国のグラフは下の方に重なり合うようになります。Afghanistan を加えると、ある年以前のデータがなかったり、値がとても小さいために、X 軸に張り付くようになってしまったりします。みやすいグラフを表示するには、どのような

*1 アルファベットの昇順になっていますが、そのリストの Country (国) の右にある 型のマークを押すと に変わり、降順になります。

ものの比較をするかも重要になってくるのがわかると思います。

- 追加した国や地域を消すのは、単にその国名の、右に表示される x マークを選択すれば良いですし、Delete キーでも消去できます。
- 実は、Japan など一つの国を加えてからあとは、下の国名などのリストから、追加したい国を選択すると、追加されていきます。検索窓に入れるよりも簡単だと思います。

- Also Show の活用

- 上に書いたように、Aggregates (総合)、Same Region (同じ地域)、Similar Values (近い値)、Highest Values (最高値)、Lowest Values (最小値) を追加できます。たとえば、Japan (日本) だけを残しておいて、Same Region を選択すると、いくつかの国の値が、薄く表示されます。地域名をみると、これは、East Asia & Pacific (東アジアと太平洋地域) であることが推測されます。
- そこで、East Asia & Pacific を選択すると、その地域の総計の値が追加されます。

7.3.2.2 データダウンロード

R のパッケージを用いたデータのダウンロードについてはあとから述べますが、指標毎のダッシュボードの右の帯のダウンロードからも、データをダウンロードすることができます。

CSV (Comma Separated Values)、XML (Extensible Markup Language)、EXCEL (Microsoft Excel Spreadsheet) と三つの形式でデータを取得できます。それぞれ、R など読み込むことができます。ただし、CSV と、XML は、Zip 形式の圧縮ファイルになっています。EXCEL は、三つのシート (Data, Metadata - Countries, Metadata - Indicators) が一つの、ブックになっています。Metadata (メタデータ) は、データのデータで、データについての情報が収められています。

7.3.3 データバンク (DataBank)

上で説明した指標毎のダッシュボードの右の帯からも、データバンクのダッシュボードが開きますが、世界銀行オープンデータの下にある、データバンクを選択すると、リストが表示されます。ここで、ある程度選択してから、ダッシュボードを表示させることもできます。一番上に、World Development Indicators (世界開発指標) がありますから、選択してください。あとから、他のものに変更することも可能です。

表が表示され、左の方に、Variables (変数)、Layout (レイアウト)、Styles (形式)、Save (保存)、Share (共有)、Embed (埋め込み) とあり、右上には、Table (表)、Chart (グラフ)、Map (地図)、Metadata (メタデータ) とあり、その下には、Preview (表示) などとなっていると思います。

膨大なデータを選択し、形式を整えて、ダウンロードしたり、グラフを表示したりすることができます。

7.3.3.1 例1

まずは、一つ目の例として、GDP (Current US\$) の値を大きい方から国順に並べてみましょう。

1. Variable の Database で、World Development Indicators を選択します。
2. Countries では、上に、All、Countries、Aggregates とありますから、Countries を選択します。
3. 一番左のチェックボックス Select All (すべて選択) を選びます。これで国が全部選択されました。現在ですと、Selected 217 と出ています。
4. Series は、いくつか選択されている可能性がありますから、X マークを選択して、まずは、全部選択を解除し、次に、GDP (current US\$) にチェックを入れます。Selected 1 となっていることを確認してください。
5. Time の一番左のチェックボックスから、Select All にしてください。現在では、Selected 63 と表示されました。
6. 右の方に、Selections have been modified ... (選択が変更されました ...) と出ますので、Apply Changes (変更を適用) を選びます。
7. 上の Layout タブを選択し、Time を Column (列)、Country を Row (行)、Series を Page (ページに指定します)

ここまでで、Table に、GDP (current US\$) についての表が表示されていることを確認してください。

これを書いている時点では、2022 年が最も新しいデータで、2022 の年のところをクリックすると、最初は、国名のアルファベット順になっていると思いますが、それが、その指標の値の、昇順、降順に変更できます。ここでは、GDP が大きな値の順に並べたいので、降順にします。

10 位までの国は、United States, China, Japan, Germany, India, United Kingdom, France, Russian Federation, Canada, Italy となっているかと思います。

リンク

7.3.3.2 例2

2022 年の値で、10 位までの国を選択して、折れ線グラフや、棒グラフなどを書いてみましょう。

1. Variables の Countries の X で選択を解除し、上の 10 カ国を選んでください。

2. 右の方に、Selections have been modified ... (選択が変更されました ...) と出ますので、Apply Changes (変更を適用) を選びます。すると表が 10 カ国のものになります。
3. 上の、Styles を選択し、Chat Type & Style で、Line (折れ線グラフ) を選択すると、しばらくして、10 カ国の、折れ線グラフを表示させることができます。色も変更することも可能です。
4. Chat Type & Style で Horizontal Bar を選択し、Layout で、

リンク

以下では、変数 (Variables) の選び方と、グラフ (Chart) について簡単に説明します。

7.3.3.3 変数 (Variables)

Database、Country、Series、Time とあり、それぞれの右に、Available と Selected とあります。

どのデータベースについて、国を選択し、系列を選び、期間を選択するという形式になっています。

Database の左の三角印を選択すると、データベース名が確認できます。現在は、World Development Indicators が選択されています。ここで、他のデータベースに変更することも可能ですが、まずは、そのままにしましょう。

次は、Countries (国) です。All、Countries、Aggregates と一番上にあります。国だけを表示するか、地域やグループを表示するか、すべてを表示するかを選択できます。

簡単のために、まずは、Countries (国) を選択しましょう。

国がいくつか選択されている場合もありますから、X マークをクリックして、すべて選択を消去し、国を選んでみましょう。GDP のところで経験したように、GDP の多い方から、United States、China、Japan、Germany、India、United Kingdom、France を選択してみましょう。

7.3.3.4 GDP per capita (constant 2015 US\$)

実質 GDP (2015 年を基準にしたもの) を、総人口で割った値。アメリカ合衆国、英国、ドイツ、フランス、日本、中国、日本、ロシア、ウクライナの 2021 年における比較棒グラフ - リンク

年次変化を示す折線グラフ -

7.3.3.5 Central government debt, total (% of GDP)

2020 年の政府の負債 (GDP の百分率) - リンク

政府の負債 (GDP の百分率) の年次変化を示す折線グラフ

7.3.3.6 CO2 emissions (metric tons per capita)

CO2 排出量 (1 人あたりのメートルトン) - リンク

CO2 排出量 (1 人あたりのメートルトン) の年次変化の折線グラフ

7.3.3.7 Military expenditure (% of GDP)

2021 年の軍事費 (GDP の%) - リンク

軍事費 (GDP の%) の年次変化

7.3.3.8 Military expenditure (current USD)

2021 年の軍事費 (現在の米ドル)

軍事費の年次変化

7.3.3.9 Proportion of seats held by women in national parliaments (%)

2021 年、国会で女性が占める議席の割合 (%) - リンク

国会で女性が占める議席の割合 (%) の年次変化

7.3.4 世界のさまざまな課題から見る

7.4 API

世界銀行 (World Bank) の API を利用した R のパッケージを二つ紹介します。

7.4.1 WDI

Search and download data from over 40 databases hosted by the World Bank, including the World Development Indicators ('WDI'), International Debt Statistics, Doing Business, Human Capital Index, and Sub-national Poverty indicators.

世界開発指標 (「WDI」)、国際債務統計、Doing Business、人的資本指数、準国家貧困指標など、世界銀行が主催する 40 以上のデータベースからデータを検索してダウンロードします。

- R のパッケージサイト : <https://CRAN.R-project.org/package=WDI>
- 資料 (Materials) : <https://cran.r-project.org/web/packages/WDI/readme/README.html>
- マニュアル (Manual) : <https://cran.r-project.org/web/packages/WDI/WDI.pdf>

- 使い方の例：<https://vincentarelbundock.github.io/WDI/>

7.4.2 wbstats

Programmatic Access to Data and Statistics from the World Bank API

世界銀行 API からのデータと統計へのプログラムによるアクセス

- R のパッケージサイト：<https://CRAN.R-project.org/package=wbstats>
- 資料 (Materials)：README
- マニュアル (Manual)：<https://cran.r-project.org/web/packages/wbstats/wbstats.pdf>
- 使い方の例 (Vignette)：<https://cran.r-project.org/web/packages/wbstats/vignettes/wbstats.html>

7.5 Google Public Data Explorer

Google のパブリックデータ探索 (Public Data Explorer) サイトの紹介とともに、世界開発指標をこれを使って見てみたいと思います。

Google で Public Data を検索すると、おそらく、次のサイトが見つかると思います。

<https://www.google.com/publicdata/directory?hl=ja&dl=ja#!>

これは、日本語サイトで、2023 年 9 月現在では、データの提供元の数が 7 となっています。また、

右上に、言語とありますから、それで English (United States) を選択すると下のリンクに飛びます。ここには、45 と書かれています。

https://www.google.co.jp/publicdata/directory?hl=en_US&dl=en_US#!

英語版で使うことをお勧めします。例えば、上で見た World Development Indicators (世界開発指標) は、英語版だけでなく、日本語版にもありますが、中身を見てみると、国によってデータがなかったり、少し古いデータまでしかなかったりなどあるようです。どの指標の、どのデータとすべてを挙げることはできませんが、英語版を使った方が安全だと思います。指標について、英語で意味するものがよくわからない時は、ブラウザーの翻訳機能を使って見当をつけるのも良いでしょう。

英語版には、上に例が出ていると思います。自動的にスライドしますが、一番最初は、世界開発指標で、私が確認したときは、Living Longer with Fewer Children (子供の数が少ないと長生き) という表題になっています。このグラフをクリックしてください。

7.5.1 例 1 WDI: Living Longer with Fewer Children

グラフをクリックすると、このようなページが表示されると思います。

左の帯には、Public Data、World Development Indicators と書かれ、いろいろな項目が並んでいます。また、よく見ると、以下のように書かれています。

X 軸：Life Expectancy (生まれた時点の平均寿命)

Y 軸：Fertility Rate (出生率)

色：Region (地域)

サイズ：Population (人口)

下：矢印と 2017 の数

と出ているかと思います。何を書いてあるか確認してください。違っていても構いません。

これで大体理解できたと思いますが、これが、WDI のデータを元にして、2017 年時点での、平均寿命と、出生率の散布図で、丸の大きさと、人口を表し、色で地域を表しています。

下の矢印を押すと、1960 から始まって、どのように変化したかを見ることができます。Gapminder のところで、書いた、バブルチャートと言われるもので、ハンス・ロスリングが Google に管理を依頼したと言われています。

左には、Compare by という項目があります。それを開くと、Region、Lending Type、Income Level とあります。Region に Color と書いてあると思いますが、Income Level の右のプルダウンメニューを押し、一番下にある、Color by This を選択すると、右上の、凡例 (Legend) と言われるものが、Income Level に変わります。動かしてみるとわかりますが、右下に、High Income がまとまっていると思います。日本は、見つかりますか。

Country List のところの、日本にチェックを入れると、Japan と表示されますから、すぐ見つかります。マウス (ポインター) を丸に近づけると国をみることもできます。High Income でも、Fertility Rate が 2 に近い国もあることがわかります。どんな国がそうになっていますか。

この辺りでやめておきます。

右上には、折れ線グラフと、棒グラフと、地球のマークと、散布図のマークがあります。現在は、散布図が使われています。その右には、ギアマークと、リンクのようなマークがあります。

ギアマークで、X 軸や、Y 軸を対数にしたり、リンクから、このグラフのリンクを取得することもできます。リンクは二種類ありますが、上が、通常のリンク、下は、iframe リンクと言われるものです。

7.5.2 例 2 GDP per Cap vs CO2 per Cap in Log-Log

今度は、日本語でも、英語でも良いですが、一人当たりの GDP を X 軸に、一人当たりの、CO2 排出量を Y 軸にとり、対数にして、表示してみましょう。

X 軸：1 人あたり GDP (実質値: 2010 年基準、米ドル表示) - 対数

Y 軸：1 人あたりの CO2 排出量 - 対数

サイズ：人口

色：地域

リンク

7.5.3 まとめ

例から始めましたので、二つの WDI の指標を使い、さらに、人口や、地域など、他の指標も、一つのグラフに含めたものを見てきました。最初にすごいものから始めてしまいましたが、大雑把には、次のようなものになっています。

- 折れ線グラフ (line graph)：一つの指標の時系列での変化。
 - いくつかの国や地域などについての値を用いて、色で区別して表示することも可能です。
- 棒グラフ (bar graph)：一つの指標をいくつかの項目 (国や、地域など) を表示
 - 特定の年の、特定の指標の値を、いくつかの国や地域について表示
- 色付き世界地図 (choropleth map)：カテゴリーごとに、色を変えて、地図上に表示
 - 個人の収入の多寡などのグループ (income level) ごとに色を変えて、地図上に表示することなどが可能です。
- 散布図 (scatter plot)：二つの指標の関係性を表示
 - さらに、点などの大きさにも他の指標の情報を加えたり、色などで、カテゴリーごとの情報を加えることも可能です。

最後に一番最初に挙げた、Living Longer with Fewer Children で何種類の情報が表示されているか見てみましょう。同時に、数値か、カテゴリー (グループ) かも書いておきます。

1. Life Expectancy (生まれた時点の平均寿命)：数値 - X 軸
2. Fertility Rate (出生率)：数値 - Y 軸
3. Country (国)：カテゴリー：点

4. Population (人口) : 数値 : 点の大きさ

5. Region (地域) : カテゴリー : 点の色

6. Year (年) : (離散的 : とびとびの) 数値 : 一枚ごとのスライド

ほかにはありますか。普通は二つの指標しか表せないように思いますが、ここでは、6つの情報が入っていますね。

いろいろと調べてみませんか。

なかなか素晴らしいですね。ただ、2014年にプロジェクトがスタートしてから、2017年、2020年と更新されていますが、更新の頻度はあまり高くないように見えます。例としては、十分機能していると思いますが。

第 8 章

OECD（経済協力開発機構）

8.1 概要

OECD（経済協力開発機構）はヨーロッパ諸国を中心に日・米を含め 38 ヶ国の先進国が加盟する国際機関です。OECD は国際マクロ経済動向、貿易、開発援助といった分野に加え、最近では持続可能な開発、ガバナンスといった新たな分野についても加盟国間の分析・検討を行っています。

https://www.meti.go.jp/policy/trade_policy/oecd/index.html

8.2 データベース

8.3 OECD

OECD Data: <https://data.oecd.org/>

Topic と Country に分かれて、調べられ、検索も可能です。

8.3.1 テーマ別

- Agriculture
 - Agricultural output
 - Agricultural policy
 - Fisheries
 - Sustainable agriculture
- Development
 - Development resource flows
 - Official development assistance (ODA)

- Economy
- Education
- Energy
- Environment
- Finance
- Government
- Health
- Innovation and Technology
- Jobs
- Society

8.3.2 例

OECD Chart: Inflation (CPI), Total / Food / Total less food, less energy, Annual growth rate (%), Monthly, Jul 2023

OECD Chart: Quarterly GDP, Total, Percentage change, previous period, Quarterly, Q1 2023

8.4 ダッシュボード

OECD.Stat <https://stats.oecd.org/Index.aspx>

8.5 API

第 9 章

United Nations（国際連合）

9.1 概要

9.2 データベース

UNdata: <https://data.un.org>

膨大なリストがあります。さまざまな機関からデータが提供されていることもあり、WDI のように、規格が統一されてはいません。

9.3 ダッシュボード

いくつかのトピックについては、ダッシュボードがあります。

9.4 API

一部のデータのための API があります。

第 10 章

Our World in Data

10.1 概要

サイト：<https://ourworldindata.org>

様々なトピックについてデータからのレポートがあり、そのデータも見ることができるようになっています。さらに、使いやすい API も提供されています。

一番上の帯の Browse by Topic からトピック探することができます。

- Population and Demographic Change 人口動態
 - Population Growth
 - Age Structure
 - Gender Ratio
 - Life Expectancy
 - Child and Infant Mortality
 - Fertility Rate
 - Urbanization
 - Migration
- Health 健康
- Energy and Environment エネルギーと環境
- Food and Agriculture 食料と農業
- Poverty and Economic Development 貧困と開発経済
- Education and Knowledge 教育と知識
- Innovation and Technological Change 技術開発とイノベーション

- Living Conditions, Community and Wellbeing 生活環境、共同体と幸せ
- Human Rights and Democracy 人権と民主主義
 - **LGBT+ Rights**
 - **Women's Rights**
 - **Child Labor**
 - **Human Rights**
 - **Democracy**
 - **Corruption**
- Violence and War 暴力と戦争
 - **Biological and Chemical Weapons**
 - **War and Peace**
 - **Military Personnel and Spending**
 - **Terrorism**
 - **Nuclear Weapons**
 - **Violence against children and children's rights**
 - **Homicides**

これら以外にも、コロナウイルス感染症や、持続的開発目標（SDGs）についてのサイトなどもあります。

10.2 レポート

10.2.1 Population Growth

人口において、日本は、1950 年に何番目で、2021 年には何番目かわかりますか。上のダッシュボードからもわかります。

10.2.2 War and Peace

1945 年以降のデータが主ですが、世界の紛争などの死者数などを知ることができます。古いデータも少しだけあります。

<https://ourworldindata.org/war-and-peace#the-decline-of-wars-between-great-powers>

10.3 データベース

owid: <https://ourworldindata.org/>

10.4 ダッシュボード

10.4.1 使い方

1. Chart グラフ

- いくつかの国（変更可能な場合もあります）についての年次変化を表す折れ線グラフや棒グラフなど

2. Choropleth 色付き地図

- 右上のプルダウンメニューからは、地域（アフリカ、北アメリカ、南アメリカ、アジア、ヨーロッパ、オセアニア）を選択して拡大することができます。
- それぞれの国にマウスまたはポインターを重ねると、その国の状況を見ることができます。

3. Table 表

- データの一部が表となっています。降順・昇順などの並べ替えも可能です。

4. Sources 出典

- 変数の簡単な定義とともに、World Bank などもとのデータの出典へのリンクも含まれています。

5. Download ダウンロード

- グラフやデータをダウンロードすることができるようになっています。

6. Share 共有

- 共有リンクや、ホームページに埋め込む iframe リンクなどが付いています。

10.4.2 Covid-19 Data Explorer

リンク

日本を加えてあります。

10.4.3 Causes of Death

死亡原因についての分析のページです。

リンク<https://ourworldindata.org/causes-of-death>

10.4.4 Causes of death, World, 2019

10.4.5 Share of Deaths from major causes of Japan, 2020

10.5 API

- パッケージの公式サイト: <https://CRAN.R-project.org/package=owidR>
 - はじめにお読みください (ReadMe) : <https://cran.r-project.org/web/packages/owidR/readme/README.html>
 - マニュアル (Manual) : <https://cran.r-project.org/web/packages/owidR/owidR.pdf>
 - 使い方の例 (Vignette) : Create and Analyse a Dataset
- owidR: Import Data from Our World in Data

第 11 章

e-Stat 日本の政府統計

11.1 概要

基本的なデータベースは、「社会・人口統計体系」から見ることができます。一覧表示も可能ですが、まずは、リンクから階層形式で見ると、都道府県別データと、市町村別データに分かれています。また、それぞれが、基礎データと社会統計指標に分かれ、さらに、それぞれが、

11.2 データベース

- e-Stat：政府統計の総合窓口 - 統計で見る日本
- ダッシュボード：対話型形式でのデータの視覚化

たくさんの統計が掲載されています。ダッシュボードは、そのうちのいくつかを見やすいように、グラフによって視覚化されています。

まず、e-Stat 政府統計の総合窓口に行くと、右上の若葉マークに使い方が書かれていますから、それを読むのがお勧めです。ここには、詳細は書きません。

まず「すべて」（政府統計一覧）では、すべての統計情報を見ることができ、検索もできます。左の帯に、データベースと、ファイルとありますが。基本的に、API を使って、データを読み込むことができるようになっているのが、データベースです。

その下には、統計分類（大分類）として、トピックごとに分類されたものを調べるものがあります。これは、トップページの「分野」と大体対応しています。その下には、組織で調べる、統計の種類で絞り込み、政府統計名で絞り込み、提供周期で絞り込み、調査年で絞り込み、調査月で絞り込み、50 音で絞り込み、統計表フォーマットで絞り込み、集計地域区分で絞り込みと続きます。

まずは、統計の種類から、基幹統計を見てみましょう。ここからも見るすることができます。全部で 61 のデータベース、64 のファイルと書かれて、国勢調査や、人口推計、労働力調査、家計調査、学校基本調査などが並んでいます。

11.2.1 家計調査

家計調査をみてみましょう。家計調査についての説明がありますが、その下の、二人以上の世帯、年次を見てみます。費目分類 010 品目分類（2020 年改定）（総数：金額）の DB（データベース）をみてみましょう。

表が現れます。

左の方にある、表示項目選択を開きます。

時間軸：すべて解除して、最新の 2022 のものだけ選択します。

左の方にある、レイアウト設定を開きます。

ページ上部：表章項目、世帯区分、時間軸（年次）

列：費目分類

行：地域分類

確定させてもとに戻ると、それぞれの項目のそれぞれの地域の平均消費額がわかります。一番上には全国平均が載っています。

たとえば、二人以上の世帯の、この調査における平均世帯人数は、2.91 人、消費支出 3,490,383 円、牛乳の支出 15,001 円など出ています。

11.2.1.1 補足

統計表としてではありませんが、家計調査からのランキングのようなものも、公開されています。ご興味のあるかたはみてください。

家計調査（二人以上の世帯） 品目別都道府県庁所在市及び政令指定都市（※）ランキング（2020 年（令和 2 年）～2022 年（令和 4 年）平均）

ただし、このランキングの下になっているデータを探そうとすると、なかなか大変です。

11.3 ダッシュボード

何種類かのグラフを見ることができると同時に、右の方に、時系列表とあり、これから、データをダウンロードできるようにもなっています。

人口ピラミッド

人口ピラミッドを見てみましょう。1920 年からの人口ピラミッドを見ることができます。また、日本の全体だけではなく、都道府県、市町村、世界の国々についても見るができます。ピラミッドの形が変わっていく様子、戦争などで、急激に変化する様子なども、見ることができます。

また、それぞれの個々のグラフのデータをダウンロードすることができます。全体のデータを一括ダウンロードできると良いと思うのですが。

11.4 API

11.5 estatapi 利用概要

CRAN estatapi URL: <https://CRAN.R-project.org/package=estatapi>

11.5.1 アプリケーション ID の設定

```
appId <- " " # 私のものは、英数40文字
```

