

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)

2023-06-04

目次

この文書について

データサイエンスを始めてみませんか。

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。

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。

コンピュータ言語について

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。特に、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、簡単なものであれば、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 によって、可能になっています。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もできることも、R をお勧めする理由です。

ただし、日本語のものは、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。

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、日本語の扱いは、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思います。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。

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できていますから、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。

PDF、ePub 版について

実は、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、ほとんど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

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。

上に書いたのは、ひとつの表現ですが、AI にも表題の「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。

ちょっと日本語がごちゃごちゃい感じもするので、英語での問いと答えも書いておきましょう。

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.

領域 (multidisciplinary field) という言葉も使われていますから、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。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、根拠をしめして、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言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言うことです。

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、自然なことばを生成できるような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、応答と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。
9. 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。

10. 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、人的な作業負荷を軽減し、生産性を向上させることができます。
11. 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。
12. 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はなく、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、パブリックデータとか、オープンデータと言われる、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

この電子ブックでは、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す。

1.2 「データサイエンスを始めましょう」の特徴

1.2.1 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。

1.2.2 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、それが可能になって来たこと。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。

1.2.3 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から理由の一つです。

もちろん、日本の課題から目を逸らすわけではありません。世界の中の日本を意識し、さらに、日本の課題にも目を向けていきたいと思います。

1.2.4 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ればと願っています。

1.3 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

世界銀行のデータで、どのようなことができるか、外観してみます。

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。

2.3 第二部 基本

R の基本、および、`tidyverse` パッケージの基本を、いくつかの、基本的なパブリックデータを使いながら、学びます。

2.4 第三部 国際機関のデータの活用

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。

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

一つ一つのステップについて、より詳しく学びます。

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。

第 3 章

最初のデータ

データを分析するときには、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. 必要なら整形 Transform data
5. 視覚化 Visualize data
6. データを理解 Understand data

いろいろな視覚化を行い、そのデータからさまざまなことを理解する部分が中心だと思えます。

3.1 R のパッケージを活用

3.1.1 準備 Setup

まず、世界銀行 (World Bank) の、世界開発指標 (WDI: World Development Indicators) の一つの、GDP について、みてみましょう。GDP にも何種類かの尺度があるのですが、次のものを見てみます。

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

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。

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に基本的なツール (パッケージ) tidyverse を使います。これらの使い方などは、あとから、説明します。

```
library(tidyverse)
#> -- 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
#> -- 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 con
library(WDI)
```

3.1.2 データ取得 Import data

いよいよ、データを取得します。それを、df_gdp と名前をつけます。

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

3.1.3 データ構造の確認

```
df_gdp
#> # A tibble: 16,492 x 13
#>   country iso2c iso3c year      gdp status lastupdated
#>   <chr>    <chr> <chr> <dbl>    <dbl> <lgl> <date>
#> 1 Afghanistan AF AFG 2015 2.00e10 NA 2022-12-22
#> 2 Afghanistan AF AFG 2011 1.82e10 NA 2022-12-22
#> 3 Afghanistan AF AFG 2014 2.06e10 NA 2022-12-22
#> 4 Afghanistan AF AFG 2013 2.06e10 NA 2022-12-22
#> 5 Afghanistan AF AFG 2012 2.02e10 NA 2022-12-22
#> 6 Afghanistan AF AFG 2007 9.72e 9 NA 2022-12-22
#> 7 Afghanistan AF AFG 2010 1.56e10 NA 2022-12-22
#> 8 Afghanistan AF AFG 2009 1.22e10 NA 2022-12-22
#> 9 Afghanistan AF AFG 2008 1.02e10 NA 2022-12-22
#> 10 Afghanistan AF AFG 2003 4.54e 9 NA 2022-12-22
#> # i 16,482 more rows
#> # i 6 more variables: region <chr>, capital <chr>,
#> # longitude <dbl>, latitude <dbl>, income <chr>,
#> # lending <chr>
```

```
df_gdp %>% glimpse()
#> Rows: 16,492
```

```
#> Columns: 13
#> $ country      <chr> "Afghanistan", "Afghanistan", "Afghani~
#> $ iso2c        <chr> "AF", "AF", "AF", "AF", "AF", "AF", "A~
#> $ iso3c        <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AF~
#> $ year         <dbl> 2015, 2011, 2014, 2013, 2012, 2007, 20~
#> $ gdp          <dbl> 19998156214, 18190410821, 20550582747,~
#> $ status       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
#> $ lastupdated  <date> 2022-12-22, 2022-12-22, 2022-12-22, 2~
#> $ region       <chr> "South Asia", "South Asia", "South Asi~
#> $ capital      <chr> "Kabul", "Kabul", "Kabul", "Kabul", "K~
#> $ longitude    <dbl> 69.1761, 69.1761, 69.1761, 69.1761, 69~
#> $ latitude     <dbl> 34.5228, 34.5228, 34.5228, 34.5228, 34~
#> $ income       <chr> "Low income", "Low income", "Low incom~
#> $ lending      <chr> "IDA", "IDA", "IDA", "IDA", "IDA", "ID~
```

概要 (`summary(df_gdp)`) からもある程度わかります。

```
df_gdp %>% summary()
#>      country          iso2c          iso3c
#> Length:16492      Length:16492      Length:16492
#> Class :character  Class :character  Class :character
#> Mode  :character  Mode  :character  Mode  :character
#>
#>
#>
#>      year          gdp          status
#> Min.   :1960      Min.   :8.824e+06      Mode:logical
#> 1st Qu.:1975      1st Qu.:2.441e+09      NA's:16492
#> Median :1990      Median :1.784e+10
#> Mean   :1990      Mean    :1.162e+12
#> 3rd Qu.:2006      3rd Qu.:2.158e+11
#> Max.   :2021      Max.    :9.651e+13
#>
#>      lastupdated      region          capital
#> Min.   :2022-12-22      Length:16492      Length:16492
#> 1st Qu.:2022-12-22      Class :character  Class :character
#> Median :2022-12-22      Mode  :character  Mode  :character
#> Mean   :2022-12-22
#> 3rd Qu.:2022-12-22
#> Max.   :2022-12-22
#>
```

```

#>   longitude      latitude      income
#> Min.      :-175.22  Min.      :-41.286  Length:16492
#> 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   :3472     NA's    :3472
#>   lending
#> Length:16492
#> Class :character
#> Mode  :character
#>
#>
#>
#>

```

国のリストをみてみましょう。

```

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"
```

3.1.4 必要なら整形 Transform data

変数が多いので、日本の部分だけみてみます。

```
df_gdp %>% filter(country == "Japan")
#> # A tibble: 62 x 13
#>   country iso2c iso3c year    gdp status lastupdated
#>   <chr>   <chr> <chr> <dbl>  <dbl> <lgl>   <date>
#> 1 Japan   JP    JPN   2021  4.94e12 NA    2022-12-22
#> 2 Japan   JP    JPN   2020  5.04e12 NA    2022-12-22
#> 3 Japan   JP    JPN   2019  5.12e12 NA    2022-12-22
#> 4 Japan   JP    JPN   2018  5.04e12 NA    2022-12-22
#> 5 Japan   JP    JPN   2017  4.93e12 NA    2022-12-22
#> 6 Japan   JP    JPN   2016  5.00e12 NA    2022-12-22
#> 7 Japan   JP    JPN   2015  4.44e12 NA    2022-12-22
#> 8 Japan   JP    JPN   2014  4.90e12 NA    2022-12-22
#> 9 Japan   JP    JPN   2013  5.21e12 NA    2022-12-22
#> 10 Japan  JP    JPN   2012  6.27e12 NA    2022-12-22
#> # i 52 more rows
#> # i 6 more variables: region <chr>, capital <chr>,
#> #   longitude <dbl>, latitude <dbl>, income <chr>,
#> #   lending <chr>
```

たとえば、gdp のところに、4.940878e+12 とあるのは、Scientific notation と言われるもので、

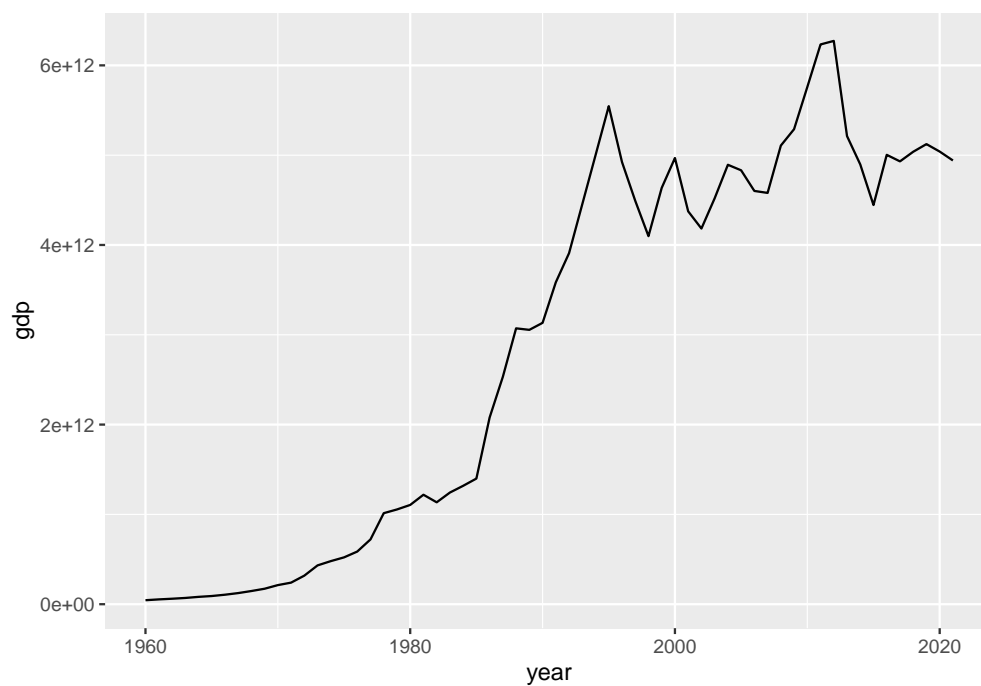
$$4.940878^{12} = 4,940,887,800,000$$

を意味します。最初に current US\$ と、書いてあるように、現在の USD に換算した数値です。

3.1.5 視覚化 data visualization

日本の GDP の経年変化を折線グラフ (line graph) でみてみましょう。

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

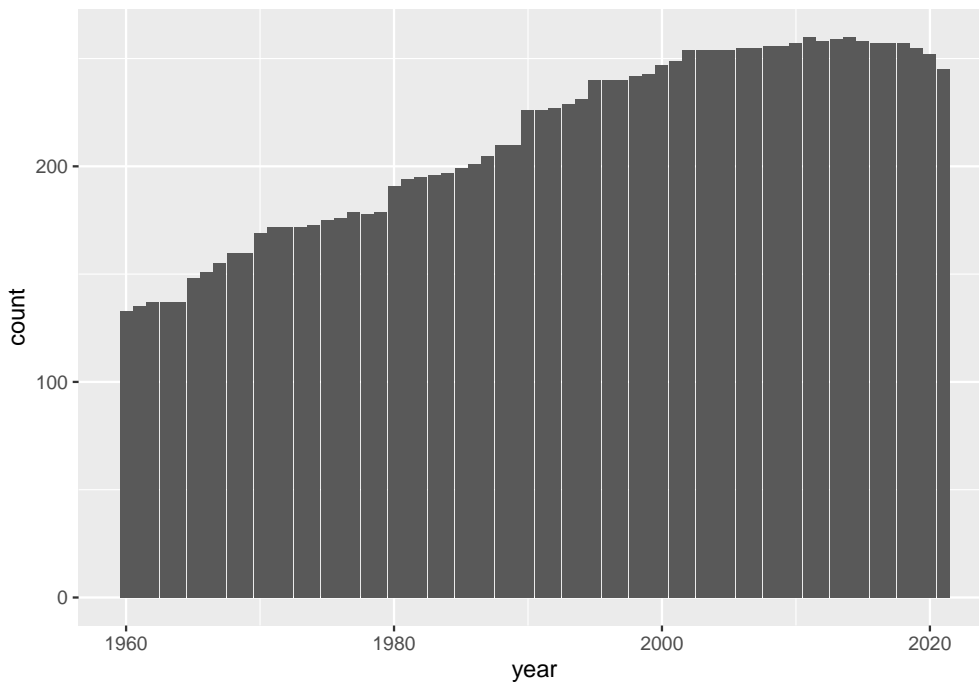


3.1.6 データを理解 Understand data

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。気づいたことをあげてみましょう。

3.1.7 さらなる分析

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



最新の 2021 年のデータはすべてあるわけではなさそうですが、大きい順に並べてみましょう。

```
df_gdp %>% filter(year == 2021) %>% drop_na(gdp) %>% arrange(desc(gdp))
#> # A tibble: 245 x 13
#>   country      iso2c iso3c  year    gdp status lastupdated
#>   <chr>        <chr> <chr> <dbl>  <dbl> <lgl>  <date>
#> 1 World        1W    WLD   2021 9.65e13 NA    2022-12-22
#> 2 High income  XD    <NA>   2021 5.98e13 NA    2022-12-22
#> 3 OECD members OE    OED   2021 5.83e13 NA    2022-12-22
#> 4 Post-demogr~ V4    PST   2021 5.50e13 NA    2022-12-22
#> 5 IDA & IBRD ~ ZT    IBT   2021 3.80e13 NA    2022-12-22
#> 6 Low & middl~ XO    LMY   2021 3.64e13 NA    2022-12-22
#> 7 Middle inco~ XP    MIC   2021 3.58e13 NA    2022-12-22
#> 8 IBRD only    XF    IBD   2021 3.55e13 NA    2022-12-22
#> 9 East Asia &~ Z4    EAS   2021 3.09e13 NA    2022-12-22
#> 10 Upper middl~ XT    <NA>   2021 2.71e13 NA    2022-12-22
#> # i 235 more rows
#> # i 6 more variables: region <chr>, capital <chr>,
#> #   longitude <dbl>, latitude <dbl>, income <chr>,
#> #   lending <chr>
```

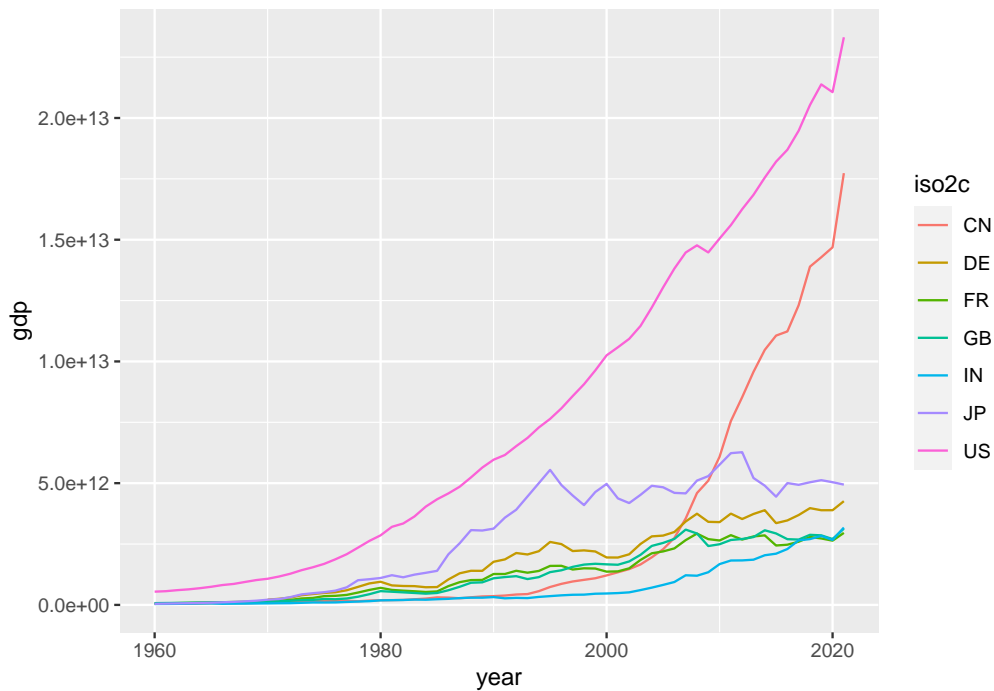
国以外のグループでも数値があるようですから、国だけを選んでみます。それには、region のところの Aggregates 以外を選択します。

```
df_gdp %>% filter(year == 2021, region != "Aggregates") %>%
  drop_na(gdp) %>% arrange(desc(gdp))
#> # A tibble: 196 x 13
#>   country      iso2c iso3c year    gdp status lastupdated
#>   <chr>        <chr> <chr> <dbl>  <dbl> <lgl>  <date>
#> 1 United Stat~ US    USA  2021 2.33e13 NA    2022-12-22
#> 2 China      CN    CHN  2021 1.77e13 NA    2022-12-22
#> 3 Japan      JP    JPN  2021 4.94e12 NA    2022-12-22
#> 4 Germany    DE    DEU  2021 4.26e12 NA    2022-12-22
#> 5 India      IN    IND  2021 3.18e12 NA    2022-12-22
#> 6 United King~ GB    GBR  2021 3.13e12 NA    2022-12-22
#> 7 France     FR    FRA  2021 2.96e12 NA    2022-12-22
#> 8 Italy      IT    ITA  2021 2.11e12 NA    2022-12-22
#> 9 Canada     CA    CAN  2021 1.99e12 NA    2022-12-22
#> 10 Korea, Rep. KR    KOR  2021 1.81e12 NA    2022-12-22
#> # i 186 more rows
#> # i 6 more variables: region <chr>, capital <chr>,
#> #   longitude <dbl>, latitude <dbl>, income <chr>,
#> #   lending <chr>
```

グラフではありませんが、これも一つの視覚化とも考えられないことはありません。

上位7カ国の GDP の推移を書いてみましょう。

```
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()`).
```



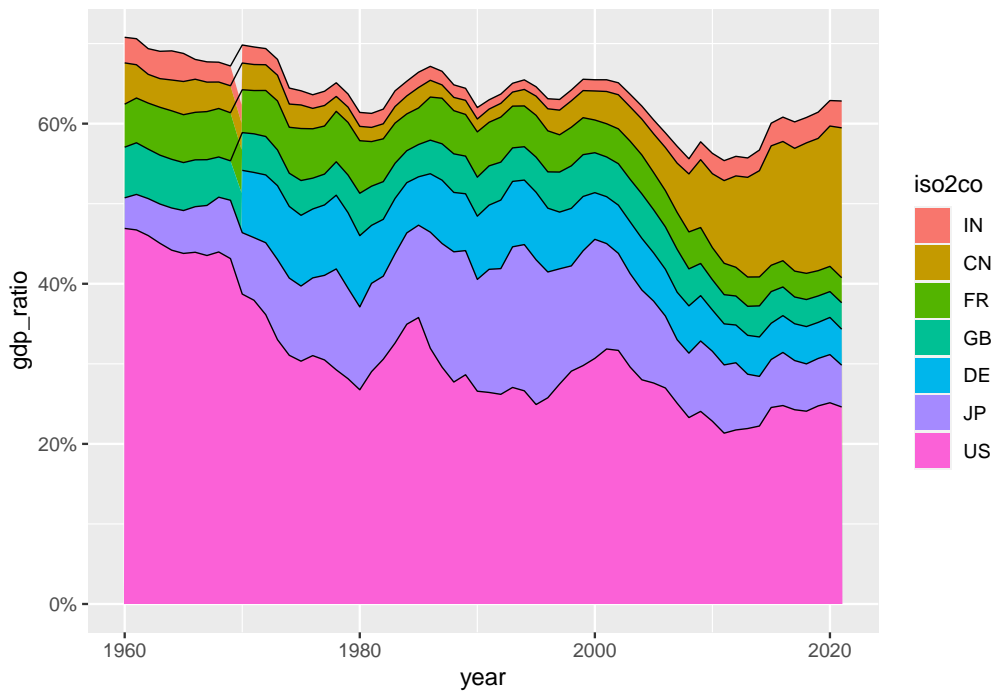
ここからも幾つかのことがわかるかと思います。気づいたことを書いてみましょう。

```
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 = "dodge") +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1))
```



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、少し変更すると下のようになります。

```
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() +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1))
```

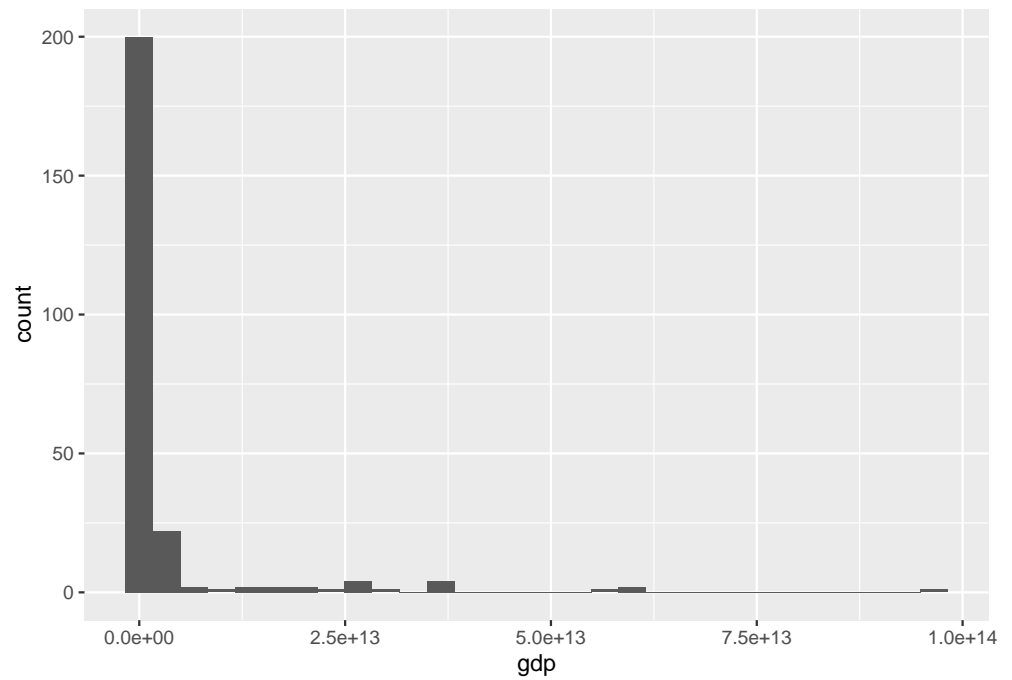



これらは、世界全体の GPT における割合です。主要国で、60%～70% を占めていることがわかりますが、それぞれの国や、幾つかの国の影響力でしょうか、ある程度みることができるように見えます。

いろいろなことが見えてくるように思います。

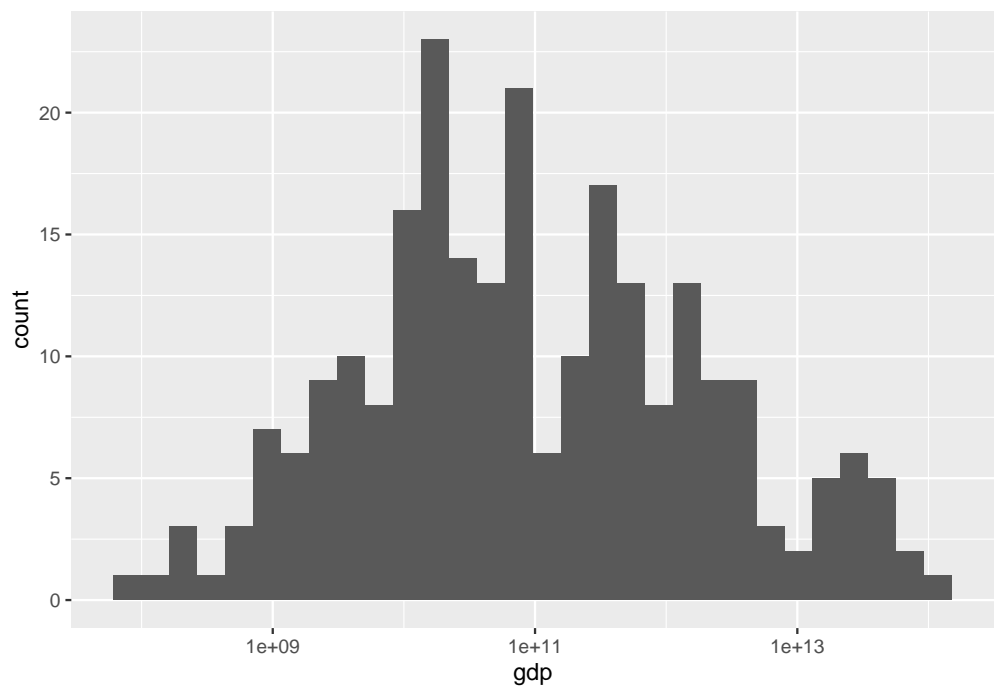
GDP が大きな国と、小さな国があるのはわかりますが、それは、どのように分布しているのでしょうか。

```
df_gdp %>% drop_na(gdp) %>%
  filter(year == 2021) %>%
  ggplot(aes(gdp)) + geom_histogram()
#> `stat_bin()` using `bins = 30`. Pick better value with
#> `binwidth`.
```



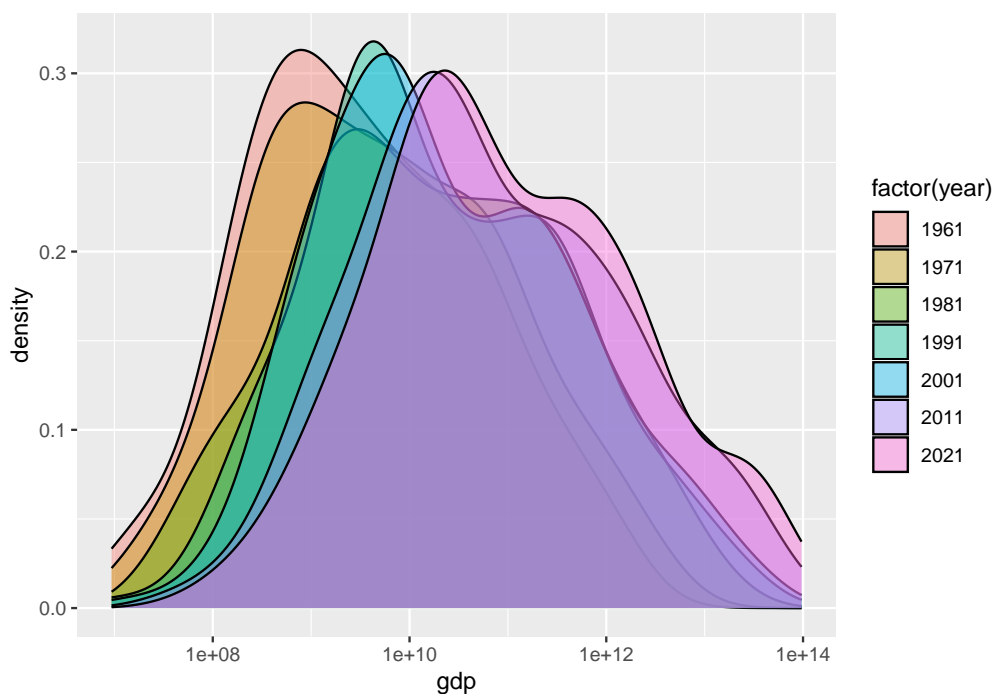
小さいところに集中していることがわかりますが、対数軸をとってみます。

```
df_gdp %>% drop_na(gdp) %>%
  filter(year == 2021) %>%
  ggplot(aes(gdp)) + geom_histogram() + scale_x_log10()
#> `stat_bin()` using `bins = 30`. Pick better value with
#> `binwidth`.
```

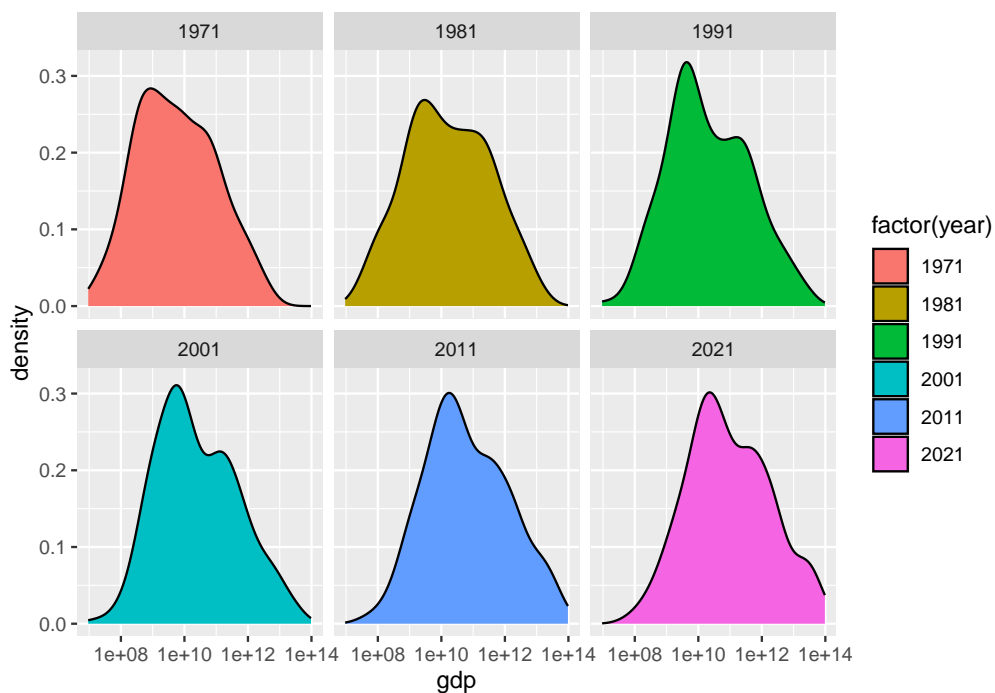


これは、2021 年のデータですが、変化を見ることもできるでしょうか。

```
df_gdp %>% drop_na(gdp) %>%
  filter(year %in% c(1961, 1971, 1981, 1991, 2001, 2011, 2021)) %>%
  ggplot(aes(gdp, fill = factor(year))) + geom_density(alpha = 0.4) + scale_x_log10()
```

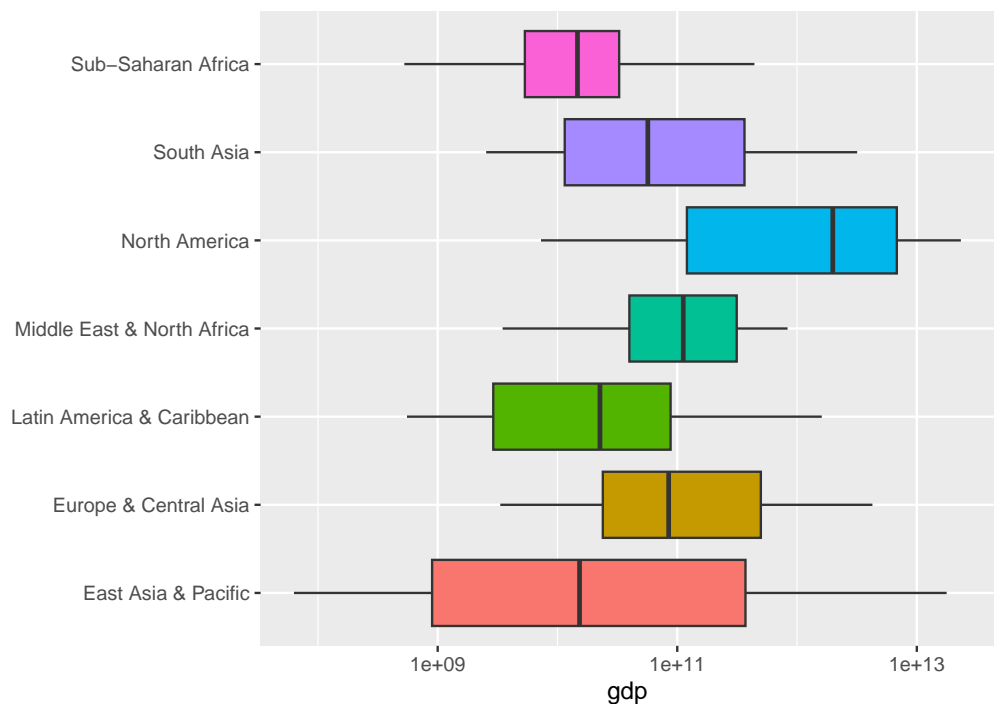


```
df_gdp %>% drop_na(gdp) %>%
  filter(year %in% c(1971, 1981, 1991, 2001, 2011, 2021)) %>%
  ggplot(aes(gdp, fill = factor(year))) +
  geom_density() + scale_x_log10() + facet_wrap(~year)
```



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、Boxplot が有効です。

```
df_gdp %>% drop_na(gdp) %>%
  filter(region != "Aggregates") %>%
  drop_na(region) %>%
  filter(year %in% c(2021)) %>%
  ggplot(aes(gdp, region, fill = region)) +
  geom_boxplot() + scale_x_log10() + labs(y = "") +
  theme(legend.position = "none")
```



```
df_gdp %>% drop_na(gdp) %>%
  filter(region != "Aggregates") %>%
  drop_na(income) %>%
  filter(year %in% c(2021)) %>%
  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")
```