

メディア芸術データベースで学ぶ データサイエンス

- 解説書 -

データサイエンスを学ぶ学生・教員の皆さまへ

政府のAI戦略のもと、文理を問わず、すべての高等教育機関の学生が数理・データサイエンス・AIを学ぶことができる環境の整備が進められています。さらに2022年には「我が国が強みを有する文化産業等におけるAI 利活用の促進」が具体目標の一つとして掲げられ*、マンガやアニメーションなどのコンテンツ領域においても、データやAI 活用の人材育成に期待が寄せられています。

そこで本教材は、これからデータサイエンスを学ぼうとする学習者が、楽しみながらデータに向き合い、そこから得られる発見を共有する喜びを体験していただけるよう、文化庁が公開している「メディア芸術データベース」から、マンガ・アニメーション・ゲーム分野のデータを用いて開発されました。

教材はオンラインで公開され、いつでも、誰でも、無償で利用いただけます。ポイントをわかりやすく解説した動画も御用意しています。初学者向けデータサイエンスの教材づくりや講義をご担当の教員の方々、自学習用の教材を求める学生の皆さまをはじめとして広くご活用いただければ幸いです。

* 『AI 戦略2022』(内閣府、令和4年4月22日)



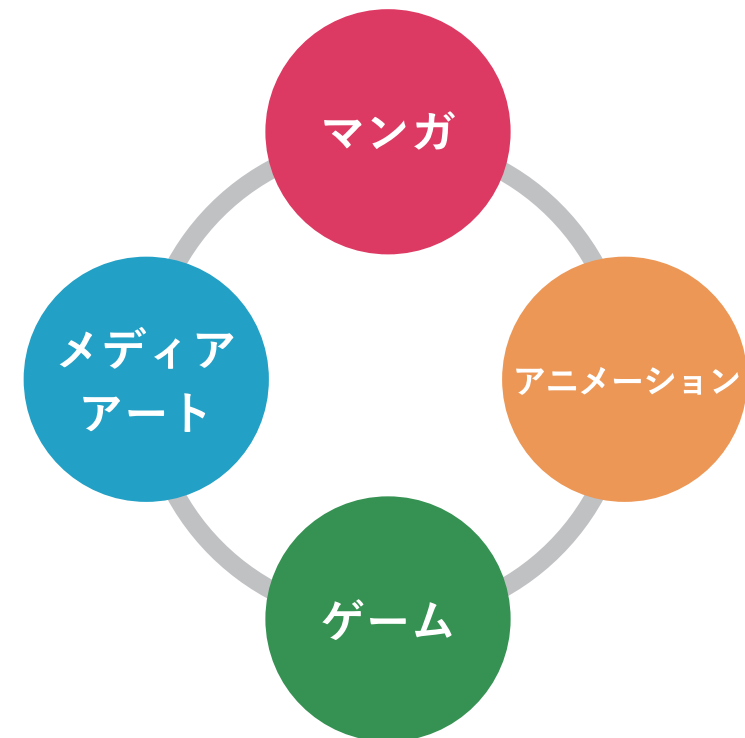
メディア芸術データベースとは

文化庁では平成22年度から、日本国内でこれまでに創造されてきたメディア芸術に関する調査を行ってきました。その成果を活用して構築されているのが、メディア芸術の作品に関する情報を広く一般に公開・提供するメディア芸術のデータ基盤「メディア芸術データベース」です。

マンガ・アニメーション・ゲーム・メディアアートの4つの分野について作品情報や複数の連携機関における所蔵情報をデータベースとして整備し、それらを横断的に検索できるウェブサイトを公開しています。さらに、ウェブ上での情報流通に便利なデータ提供サービスを行っています。

メディア芸術データベースは、メディア芸術の作品が次の世代へと引き継がれていくために、いつでも誰でもどんな時でも、作品や知りたい情報に容易にたどり着けるようにするためのデータ基盤となることを目指しています。

- 各分野のデータ提供件数(2023年3月現在)
- マンガ分野 52万件以上
- アニメーション分野 13万件以上
- ゲーム分野 4.9万件以上
- メディアアート分野 1.4万件以上



教材一覧

本教材シリーズは、データサイエンス初学者を対象に、Excel でできるデータ分析の基礎とビジュアライゼーションについて学べる全5コマで構成されています。

ウェブサイト上にテキストとイントロダクション動画が公開されていますので、どなたでも自由にブラウザからアクセスし、いつでも、どこからでも、すべて無償で御利用いただけます。

#1:メディア芸術 × データサイエンスって誰のため?何のため?



メディア芸術データベースとは何か、データサイエンスをすることで何ができるかについて解説します。文化振興、ビジネス、学術研究など、自分の専門分野や興味関心に応じて、データサイエンスを学ぶ意義への理解を促します。

#2:データビジュアライズにチャレンジしよう～初級編



マンガ分野を対象に、Excelのピボットテーブルを用いた集計・分析について解説しながら、データをビジュアライズすることの面白さを理解します。データはあらかじめメディア芸術データベースから抽出したサンプルデータセットを使います。

#3:データビジュアライズにチャレンジしよう～応用編



ゲーム分野を対象に、データから特徴量を設計し、分析を進めます。自分の興味関心や仮説立証のための分析と、データビジュアライズを通じて新たな発見が得られることを体験していきます。基礎的なテキスト解析にもチャレンジします。

#4:MADB のデータは自由に使ってもいいの?



データを扱う上で留意すべき事項について、著作権を中心に学びます。著作物やオープンデータとは何かについて解説したうえで、メディア芸術データベースだけでなく、外部のデータを組み合わせるさらに高度な分析に向けて準備します。

#5:データビジュアライズからの発見を伝えよう



本教材を用いた分析を通じて自分が発見したことを伝えるためのプレゼンテーションを作成します。相手を想定したうえで、明確でわかりやすく、説得力のあるデータビジュアライズを作成するための配色や表現の工夫について学びます。

サンプルデータセット

テキストで解説されているデータ分析を実践いただくためのサンプルデータセットを.xlsx形式および.csv形式で提供しています。このデータセットは、メディア芸術データベースから抽出したデータを成形したものです。また、学習者にとってもなじみやすい作品を含めること、データサイエンス初学者の方でも扱いやすいデータボリュームとすることを目的に、1990年～2017年までを対象期間としています。

- ・ **週刊少年誌.xlsx** (週刊少年誌.csv)
- ・ **月刊少女誌.xlsx** (月刊少女誌.csv)
- ・ **テレビアニメ.xlsx** (テレビアニメ.csv)
- ・ **ゲームパッケージ.xlsx** (ゲームパッケージ.csv)

サンプルデータセットは、MADB Labウェブサイトよりダウンロードいただけます。



https://mediag.bunka.go.jp/madb_lab/lod/ds/



#1

メディア芸術 × データサイエンスって
誰のため？何のため？

1. 導入

1.01 メディア芸術データベースでデータサイエンスをしよう

- データサイエンスで分析の対象となるデータは、数字の羅列に限りません。自分たちにとって身近なマンガ、アニメーション、ゲーム、メディアアートなどもデータサイエンスの手法を使って分析することが可能です。本教材では、メディア芸術データベース(以降、MADB)を題材に、楽しみながらデータサイエンスを学んでいくことを目的にしています。

1.02 メディア芸術データベースについて

- メディア芸術データベースは、文化庁が広く一般に提供しているメディア芸術のデータベースです。マンガ、アニメーション、ゲーム、メディアアートの4分野における作品情報や所蔵情報をデータベースとして整備し、メディア芸術作品へのアクセスおよびその保存・利活用の要となるデータを提供しています。
- メディア芸術データベースは、メディア芸術作品が次の世代へと引き継がれていくことを目的に、誰でもいつでも作品情報やアクセス情報にたどり着けるようにするためのデータ基盤として開発されました。

1.03 文化の保護・振興とデータベース

- 文化芸術基本法では、国がメディア芸術保護のための各種施策を講ずることを定めています。その一環として開発・公開されているのがメディア芸術データベースです。メディア芸術データベースによって作品情報やアクセス情報の提供を行うことで、メディア芸術作品を次の世代へと引き継いでいくことに繋がります。
- それだけでなく、データサイエンスの手法を用いてメディア芸術を分析することで、新たな作品の創造や作品の分析を通じた新たな発見にも繋がります。

2. データベースをデータサイエンスすることでなにができるか?

2.01 文化振興目的

- メディア芸術作品は、何も無いところから誰かが突然ひらめいて生み出されるものではありません。データベースを分析し、ルーツを知るとは、クリエイターがより深く既存のコンテンツを理解し、新たな作品を創作できることにつながります。

2.02 ビジネス目的

- これまで作られてきた作品を分析すれば、その傾向から今後の流行を予測することができるかもしれません。ヒットする作品を考える上で、データベースから得られる知見を利用してみましょう。

2.03 学術研究目的

- メディア芸術作品は、時代の流行を取り入れて作られています。データベースのデータを分析することで、過去の時代を分析し、そこから当時の世相を分析することができます。

2.04 (文化の持つ価値を)社会(に)還元(する)目的

- メディア芸術作品を活用した地域おこしが各地で盛んに行われています。データベースのデータから地域に関連するデータを探してみましょう。

3. MADB紹介

3.01 代表的な項目紹介

- メディア芸術データベースにはどのような項目があるのでしょうか。代表的な項目について確認をしてみましょう。

分野	共通項目	個別の項目
マンガ	タイトル、作者名 など	発行者名(出版社)、単行本シリーズ(ID)、ISBN、作者(ID)、ページ数、価格など
アニメーション		発行者名(放送局)、アニメシリーズ(ID)、放送開始時刻、放送枠時間など
ゲーム		発行者名(メーカー)、出版者名、価格、公開年月日、レーティングなど
メディアアート		キーワード、公開年月日、保管場所、素材など

3.02 データのカバー範囲

- メディア芸術データベースにはどれくらいの量のデータが登録されているのでしょうか。

2023年3月現在

分野	データ件数	データの範囲
マンガ	527,353	1874年～2023年
アニメーション	136,768	1963年～2019年
ゲーム	49,684	1961年～2018年
メディアアート	14,335	1935年～2020年

3.03 分野を越えたビッグデータの分析から新たな発見を

- こうした大量のデータが収められていることによってどのようなことが実現できるでしょうか。異なる時代の作品同士を比較することで、時代ごとの傾向の違いや、時代を越えて通底する特長を明らかにすることができます。
- また、マンガ、アニメーション、ゲーム、メディアアートという分野を横断して作品の広がりや、分野ごとの特長を明らかにすることも可能となります。

4. データベースをどのように見るか?

- 4.01 データベース内のデータから知見を得るには、1つ1つのデータに注目し丁寧に読み込む方法と、データ全体を俯瞰しデータ間の関係性や位置づけを確認する方法があります。
- 4.02 本教材では、特に後者のアプローチに焦点を当て、データを俯瞰的に見ることや、定量的な手法でデータの特徴を分析することを学びます。
- 4.03 またグラフや図などでデータの傾向を明らかにする可視化の手法についても取り上げます。



#2

データビジュアライズに
チャレンジしよう～初級編



1. ビジュアライズの意義

- 1.01 人間はデータを見て、それを直接理解することは困難です。特に、たくさんの特徴量をもつデータは、そのままでは複雑すぎて手に負えません。そこで、例えば数値を棒グラフにすれば、数値を直感的に比較できるようになり、とても理解しやすくなります。ビジュアライズを行う意義は、人間の理解を助けることにあります。
- 1.02 ビジュアライズする際には、多くの場合データを2次元で表現します。例えば、棒グラフであれば、比較したい項目を横軸に取り、その数値を縦軸に取ってデータを表現します。こうしたグラフを描くためには、そもそも元のデータを集計したりして、複雑さを抑えた状態に変換する必要があります。これは、データに対してなんらかの観点を与えた状態であると言えるでしょう。つまり、ビジュアライズをするということは、データをどのような観点で見ると決めて、その観点とセットで結果を共有するということになります。

2. データを見てみよう

- 2.01 今回ビジュアライズの題材とするのは、1990年から2017年までの少年マンガ雑誌4誌のデータです。
- 週刊少年ジャンプ、週刊少年サンデー、週刊少年マガジン、週刊少年チャンピオン
- 2.02 MADBから抽出したデータをクレンジングして使いやすい形にしたものを用意してあります。
- 2.03 ではサンプルデータ<週刊少年誌.xlsx>をエクセルで開いてみましょう。

3. このファイルは

3.01 (何行ある) 113,282件のデータがある。

3.02 (どういう単位でデータが入っている)一冊のマンガ雑誌の中の掲載されている作品をそれぞれ一つのレコードとして保存してある。

マンガタイトル	各話タイトル	雑誌番号名	
DADA!	trouble.10 悪いヤツほどよくドジる	週刊少年サンデー 1990年 表示号数1	} 雑誌一冊分
拳児	第4部 第29話 宿命のライバル	週刊少年サンデー 1990年 表示号数1	
...	
燃える!お兄さん	☆ボーナスの行方の巻	週刊少年ジャンプ 1990年 表示号数1	} 雑誌一冊分
聖闘士星矢	●伝説の聖闘士の巻	週刊少年ジャンプ 1990年 表示号数1	
...	
4P田中くん	第170話 打球の行方	週刊少年チャンピオン 1990年 表示号数1	
魔界学園	第23話 三人の助っ人	週刊少年チャンピオン 1990年 表示号数1	
...	
名門!第三野球部	第108話 勇気をください!	週刊少年マガジン 1990年 表示号数1	
ブレイクショット	第127話◆第三の球、ふたたび!!	週刊少年マガジン 1990年 表示号数1	
...	

3.03 (属性は何がある) URI/雑誌タイトル/マンガタイトル/各話タイトル/作者名/開始ページ/終了ページ/雑誌巻号名/ページ数/公開年月日/出版日/価格

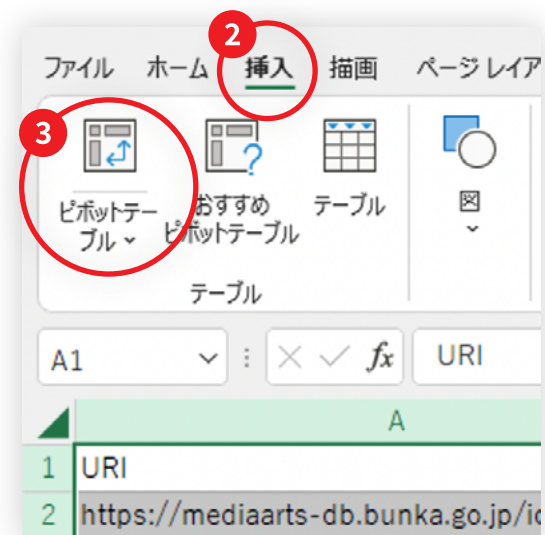
URI	雑誌 タイトル	マンガ タイトル	各話 タイトル	作者名	開始 ページ	終了 ページ	雑誌 巻号名	ページ 数	公開 年月日	価格
-----	------------	-------------	------------	-----	-----------	-----------	-----------	----------	-----------	----

4. ピボットテーブルの挿入

4.01 このデータでは、あるマンガタイトルについての情報は、掲載された雑誌ごとに分かれているので、マンガタイトルごとの情報を見なければ、集計をする必要があります。こうした操作をするには、ピボットテーブルを使うのが便利です。

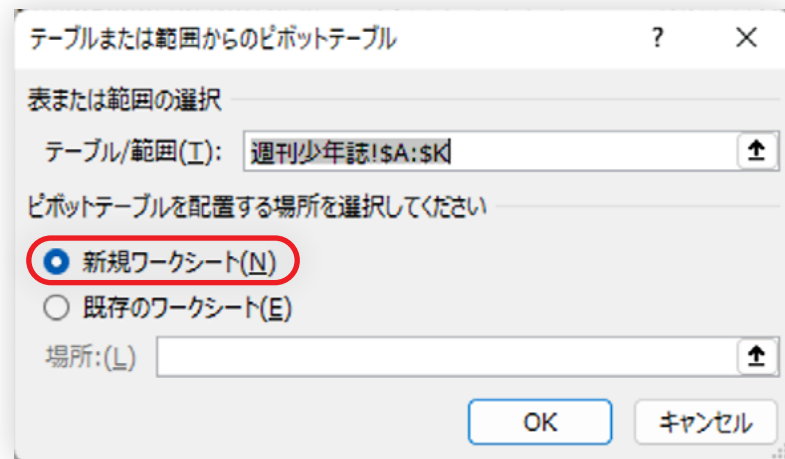
4.02 ①データ全体を選んで(左上の三角をクリック)、②「挿入」から③「ピボットテーブル」を選びます。

	A	B
1	URI	雑誌タイトル
2	https://mediaarts-db.bunka.go.jp/id/M578292	週刊少年サンデー
3	https://mediaarts-db.bunka.go.jp/id/M578292	週刊少年サンデー
4	https://mediaarts-db.bunka.go.jp/id/M578292	週刊少年サンデー
5	https://mediaarts-db.bunka.go.jp/id/M578292	週刊少年サンデー
6	https://mediaarts-db.bunka.go.jp/id/M578292	週刊少年サンデー



4. ピボットテーブルの挿入

4.03 「新規ワークシート」を選びます。



4. ピボットテーブルの挿入

4.04 このような画面になります。

The screenshot shows a spreadsheet application interface. On the left, a task pane titled "ピボットテーブル1" (PivotTable1) provides instructions: "レポートを作成するには、[ピボットテーブルのフィールドリスト] からフィールドを選択してください" (To create a report, select fields from the [PivotTable Fields List]). Below the text are icons representing a data table and a pivot table. The main spreadsheet area shows columns A through N and rows 1 through 21. On the right, the "ピボットテーブルのフィールド" (PivotTable Fields) task pane is open. It contains a search bar and a list of fields with checkboxes: URI, 雑誌タイトル (Magazine Title), マンガタイトル (Manga Title), 各話タイトル (Episode Title), 作者名 (Author Name), 開始ページ (Start Page), 終了ページ (End Page), and 雑誌番号 (Magazine Number). Below this list, there are sections for "フィルター" (Filter) and "列" (Column Labels), and "行" (Row Labels) and "Σ 値" (Summarize With). At the bottom of the task pane, there is a checkbox for "レイアウトの更新を保留する" (Retain layout updates) and an "更新" (Refresh) button. The spreadsheet's status bar at the bottom shows "準備完了" (Ready), "アクセシビリティ: 利用不可" (Accessibility: Not Available), and a zoom level of 100%.

5. 分析例A:掲載話数が多いマンガ10位

5.01 このデータの期間で掲載数が多かったマンガを可視化してみます。まず、右上の項目一覧から、①「マンガタイトル」を「行」のフィールドにドラッグ&ドロップします。次に、同じく②「マンガタイトル」を「値」にドラッグ&ドロップします。

このスクリーンショットは、Excelのピボットテーブル作成の最初のステップを示しています。右側の「ピボットテーブルのフィールド」タスクパネで、「マンガタイトル」がリストアップされています。この項目が「行」のフィールドにドラッグ＆ドロップされています。赤い丸と矢印で「1 ドラッグ&ドロップ」と注釈されています。

このスクリーンショットは、Excelのピボットテーブル作成の2番目のステップを示しています。最初のステップで設定された「マンガタイトル」の行に加え、もう一度「マンガタイトル」が「値」のフィールドにドラッグ＆ドロップされています。赤い丸と矢印で「2 ドラッグ&ドロップ」と注釈されています。

5. 分析例A:掲載話数が多いマンガ10位

5.02 こうすることで、各マンガのタイトルごとに、個数を集計することができます。
セルの幅を調節するようになります。

The screenshot shows an Excel spreadsheet with a pivot table. The pivot table is located in the range A3:H3. The rows of the pivot table are labeled with manga titles, and the columns are labeled with counts. The pivot table is set to show 'Manga Title' as rows and 'Count' as columns. Two red circles highlight the 'Manga Title' and 'Count' headers.

行ラベル	個数 / マンガタイトル
[イラスト]	1
[シボトラ]	130
「マガジンと僕。」	1
「若一カードバトルで勝負です!!」	1
『お兄さん』のできるまで	1
・・・のマル秘エピソードをマンガ化!!	33
★★★(三ツ星)のスペシャルリテ	49
★SANTAI★	12
めティンクル ティンクル ♀アイドル☆スター	49
090 -えこといっしょ-	60
0の宇宙 ブラックホール誕生を予言した男	1
1000年の疾風	15
15の夜 ~cocoon・音姫	1
17の夜 ~DIFFUSE・歌乱	1
18の夜 ~いじめる	2
19の夜 AIDS編	2
20の夜	1
21 2×2	108

The pivot table fields task pane on the right shows the following settings:

- Report to add fields: マンガタイトル
- Filter: URI, 雑誌タイトル, 各話タイトル, 作者名, 開始ページ, 終了ページ
- Row Labels: マンガタイトル
- Column Labels: 個数 / マンガタイトル

5. 分析例A:掲載話数が多いマンガ10位

5.03 上位10件を見るために、これを並び替えましょう。①フィルタ機能のボタンをクリックします。②「並び替えオプション」で③「個数/マンガタイトル」を選び、最後に④「降順」を選択して、⑤OKボタンをクリックします。すると掲載話数順にタイトルを並べることができます。

The image shows a sequence of three screenshots illustrating the steps to sort manga titles by the number of chapters in descending order. A red arrow points from the 'OK' button in the final screenshot back to the 'Sort Options' dialog in the first screenshot.

Step 1: The main application window shows a list of manga titles. A red circle labeled '1' highlights the filter button (a downward arrow icon) in the top right corner of the list.

Step 2: A dialog box titled '並び替え (マンガタイトル)' (Sort (Manga Title)) is open. A red circle labeled '2' highlights the 'その他の並び替えオプション(M)...' (Other Sort Options (M)...) button.

Step 3: The '並び替えオプション' (Sort Options) section of the dialog is shown. A red circle labeled '3' highlights the '個数 / マンガタイトル' (Number of Chapters / Manga Title) option in the dropdown menu.

Step 4: The '並び替えオプション' (Sort Options) section is shown. A red circle labeled '4' highlights the '降順(D):' (Descending (D):) radio button.

Step 5: The '並び替えオプション' (Sort Options) dialog is shown. A red circle labeled '5' highlights the 'OK' button.

Background Data Table:

行ラベル	個数 / マンガタイトル
こちら葛飾区亀有公園前派出所	1306
はじめの一步	1176
名探偵コナン	1009
ONE PIECE	889
MAJOR	748
NARUTO-ナルト-	721
BLEACH	710
...	651
強の弟子ケンイチ	582
の空	571
のごとく!	566
...	558
TAIL	546
少年の事件簿	512
の騎士	502
ハンド編	488
テイルドレン	479

5. 分析例A:掲載話数が多いマンガ10位

5.04 ここから上位十件をグラフにしていきたいと思います。まず、フィルタ機能で上位十件だけを表示します。①「値フィルター」のメニューから②「トップテン」を選びます。そして③表示する項目数に10を入力します。これで上位10件だけを表示することができます。

ラベル	個数 / マンガタイトル
こちら葛飾区亀有公園前	1306
はじめの一步	1176
探偵コナン	1009
ONE PIECE	889
MAJOR	748
NARUTO-ナルト-	721
BLEACH	710
銀魂	651
史上最強の弟子ケンイチ	582
あひるの空	571
総計	8363

トップテンフィルター (マンガタイトル)

表示

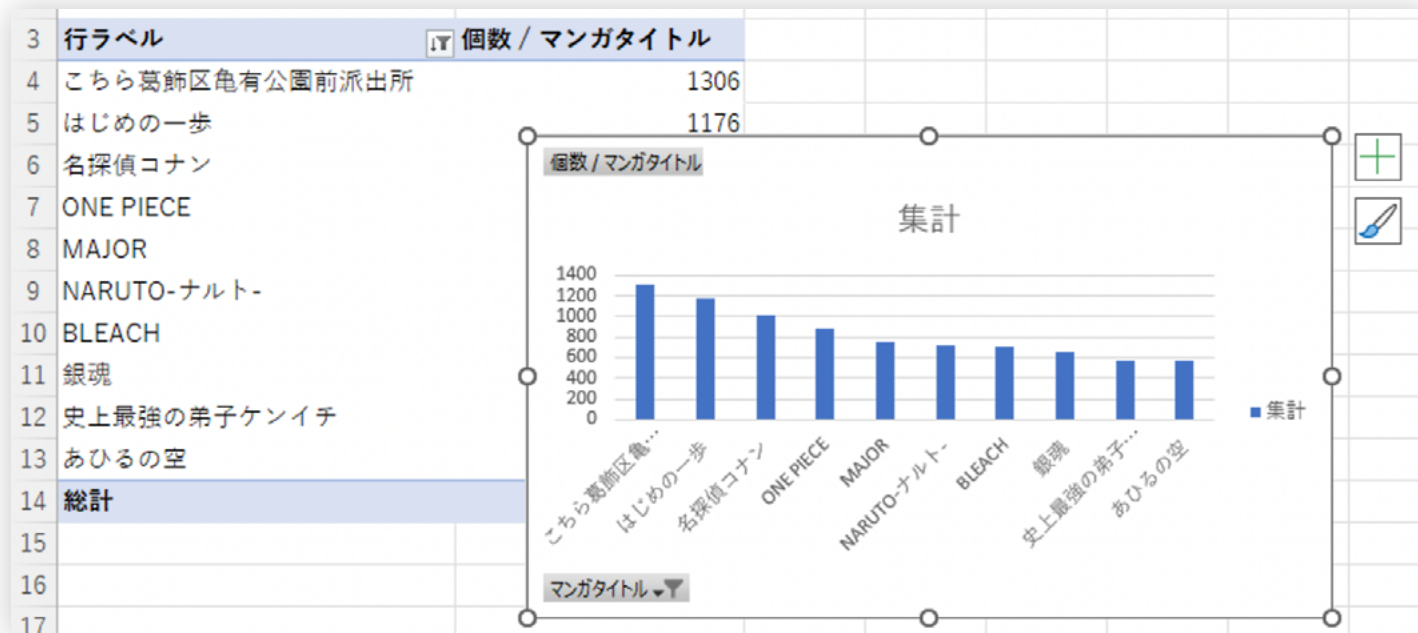
個数 / マンガタイトル の 上位 10 項目

OK キャンセル

行ラベル	個数 / マンガタイトル
こちら葛飾区亀有公園前派出所	1306
はじめの一步	1176
名探偵コナン	1009
ONE PIECE	889
MAJOR	748
NARUTO-ナルト-	721
BLEACH	710
銀魂	651
史上最強の弟子ケンイチ	582
あひるの空	571
総計	8363

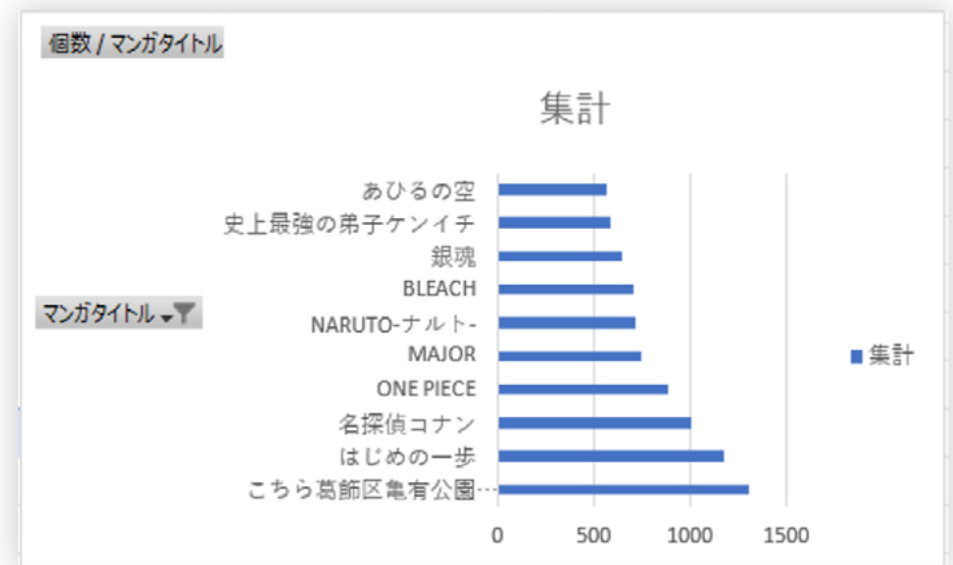
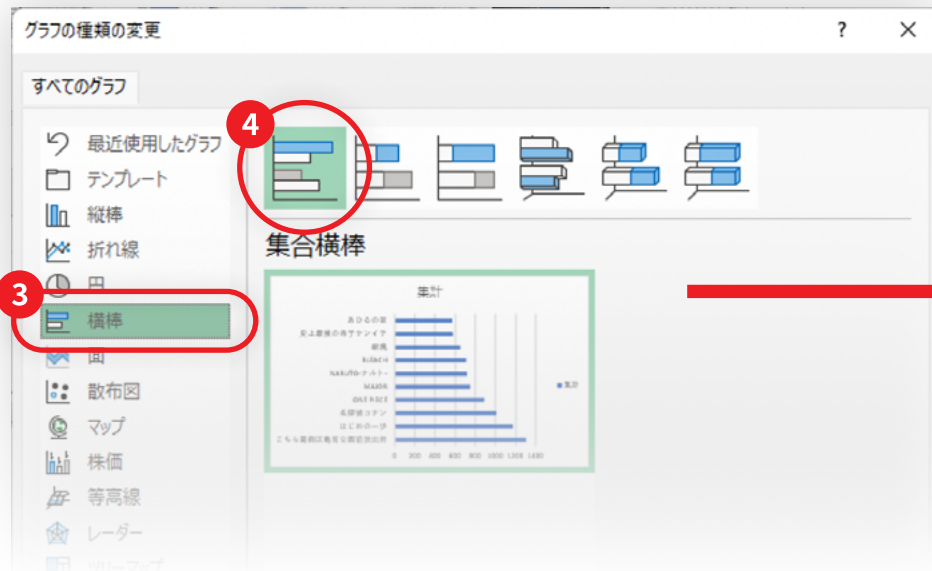
5. 分析例A:掲載話数が多いマンガ10位

5.05 リボンタブの①「ピボットテーブル分析」から②「ピボットグラフ」を選ぶことで、これをグラフにすることができます。



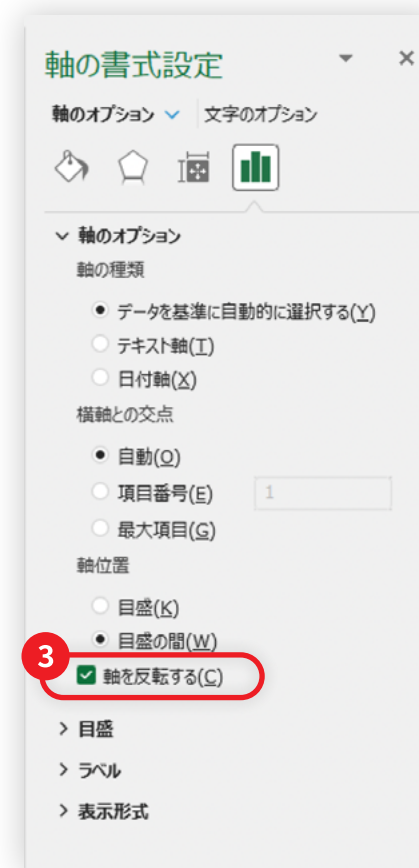
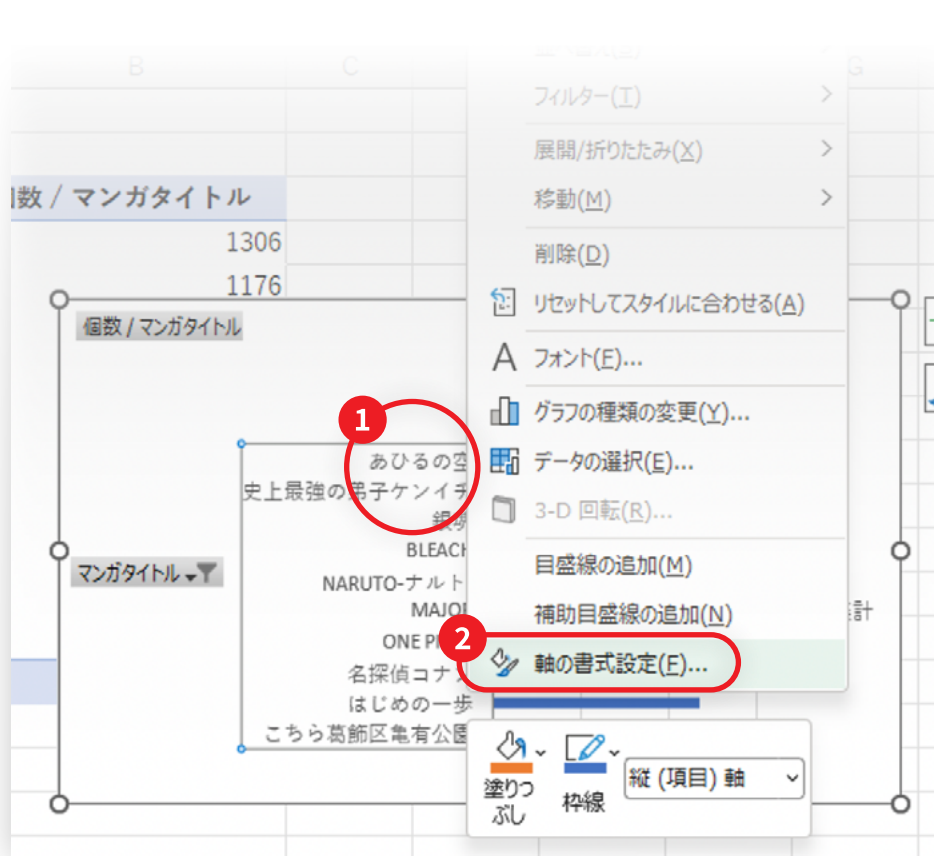
5. 分析例A:掲載話数が多いマンガ10位

5.06 これをもう少し整形して見やすくします。グラフを選んだ状態で①リボンタブの「デザイン」を選び、②「グラフの種類の変更」から③「横棒」、④「集合横棒」を選びます。するとグラフが横向きにはなるのですが、今度は上下が反転してしまっています。



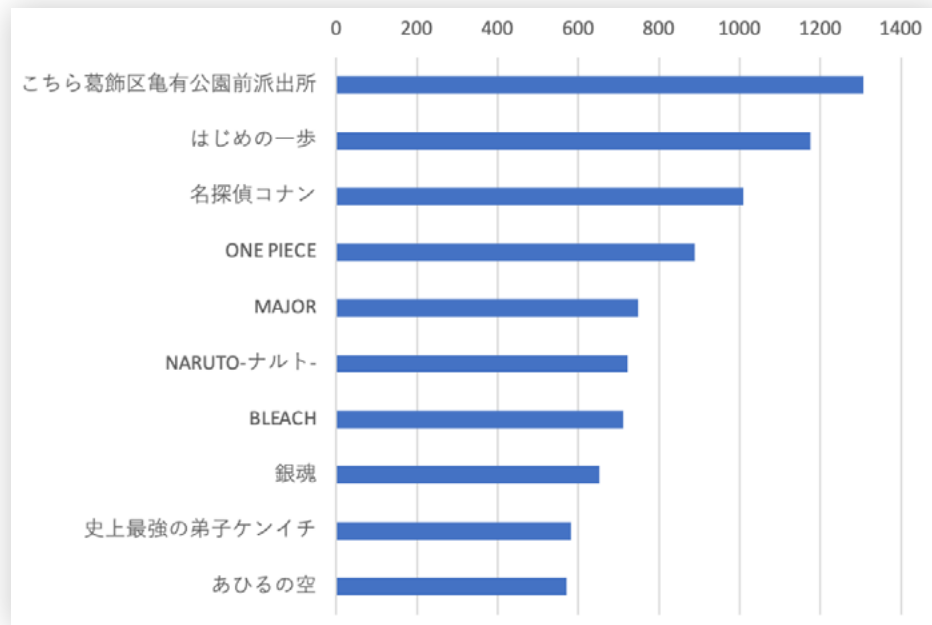
5. 分析例A:掲載話数が多いマンガ10位

5.07 これを反転するには、軸の設定を行います。①マンガのタイトルが書いてある辺りを右クリックして、②「軸の書式設定」を選んで、③メニューの中の「軸を反転する」にチェックを入れます。あとはタイトルは「集計」と書いてあるだけなのでここでは消してしましましょう。



5. 分析例A:掲載話数が多いマンガ10位

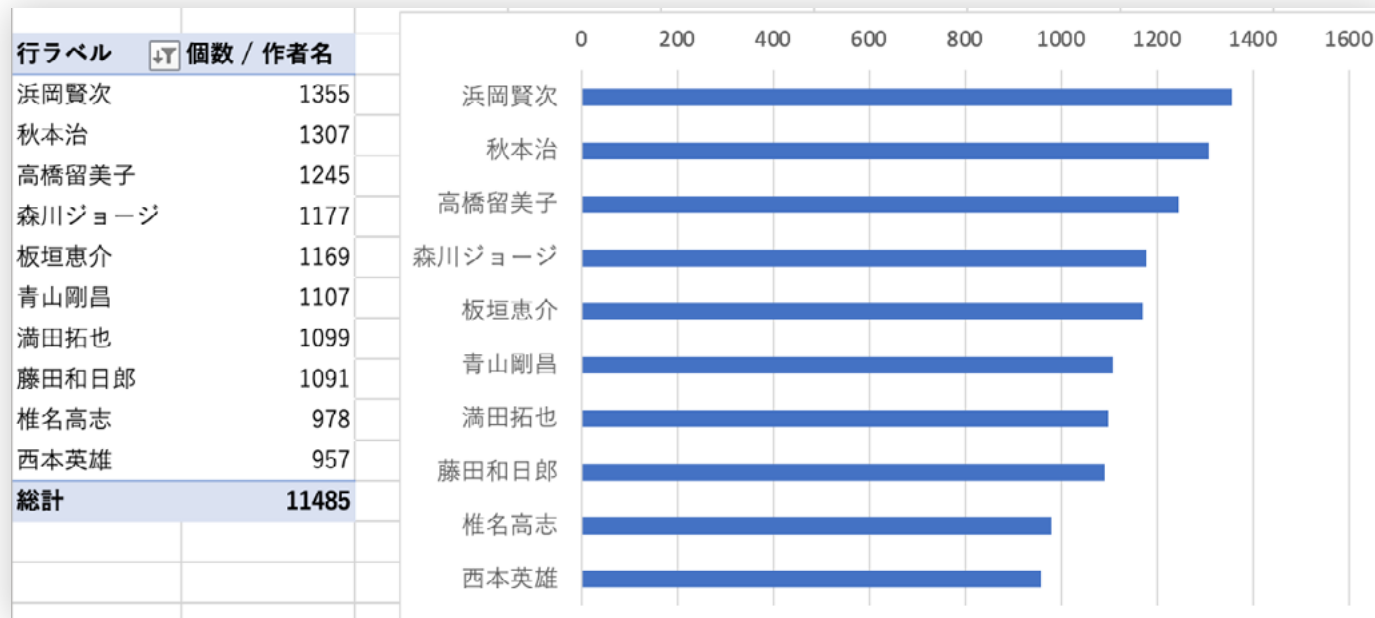
5.08 以上で以下のようなグラフが得られます。一位は「こち亀」でした。こち亀は1990年より前から連載しているので、それを含めたらもっと多くなったはずですね。



6. 分析例B:掲載話数が多い作者10位

6.01 基本的な操作は上記と同じですが、最初に、ピボットテーブルの「行」に「作者名」を選びます。あとは同様に、以下のグラフを得ることができます。

6.02 1位になったのは「浦安鉄筋家族シリーズ」などで知られる「浜岡賢次」さんでした。作品とは違うランキングになるのが興味深いですね。



7. 分析例C: 掲載ページ数が多い作者10位

- 7.01 これまでは元のデータから直接作れるグラフでしたが、最後は一つ自分で特徴となるデータを作り出してそれをグラフ化してみましょう。
- 7.02 作者ごとの掲載ページ数の合計を集計してみます。元データには、マンガの一話分の雑誌の中での位置である「開始ページ」「終了ページ」の情報があるので、そこからページ数を計算します。
- 7.03 まず、元データのシートの、データが入っている部分の右に、①「掲載ページ数」という列のタイトルを付けます。

H	I	J	K	L
雑誌巻号名	ページ数	公開年月日	価格	① 掲載ページ数
週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	
週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	
週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	
週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	
週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	
週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	

7. 分析例C:掲載ページ数が多い作者10位

7.04 そしてその下に、ページ数を計算して入れていきます。①掲載するページ数は、「(終了ページ)-(開始ページ)+1」で計算することができます。

E	F	G	H	I	J	K	L
作者名	開始ページ	終了ページ	雑誌巻号名	ページ数	公開年月日	価格	掲載ページ数
吉田聡	125	142	週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	=G2-F2+1
椎名高志	143	148	週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	
上田悦	209	214	週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	
藤原芳秀	221	238	週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	
竜崎遼児	257	274	週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	
高橋留美子	29	44	週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	

7.05 そして、計算式を下にコピーします。

E	F	G	H	I	J	K	L
作者名	開始ページ	終了ページ	雑誌巻号名	ページ数	公開年月日	価格	掲載ページ数
吉田聡	125	142	週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	18
椎名高志	143	148	週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	6
上田悦	209	214	週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	6
藤原芳秀	221	238	週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	18
竜崎遼児	257	274	週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	18
高橋留美子	29	44	週刊少年サンデー 1990年 表示号数1	414	1990/1/1	190	16

7. 分析例C:掲載ページ数が多い作者10位

7.06 ここまでできたら、ピボットテーブルを再度作りましょう。そして、①「行」に「作者名」を、新たに追加した②「掲載ページ数」を追加します。すると、「掲載ページ数」は数値なので、その合計が集計されます。

行ラベル	合計 / 掲載ページ数
[マンガ・イラスト]中邑みつのり	21
[共作]	11
[作・演出・特撮]丹波鉄心	223
AR-V	6
Boichi	428
CLAMP	5154
KAITO	1140
KANB	5
Mr.オモキ	10
ONE	18
RAN	38
SP☆なかてま	112
S松江名君	4
tatta	66
THE SEIJI	1244
UMA	424
アイハラシゲ	80

ピボットテーブルのフィールド

レポートに追加するフィールドを選択してください:

検索

- 雑誌巻号名
- ページ数
- 公開年月日
- 価格
- 掲載ページ数
- その他のフィールド...

ドラッグ&ドロップ

フィールド

行

作者名

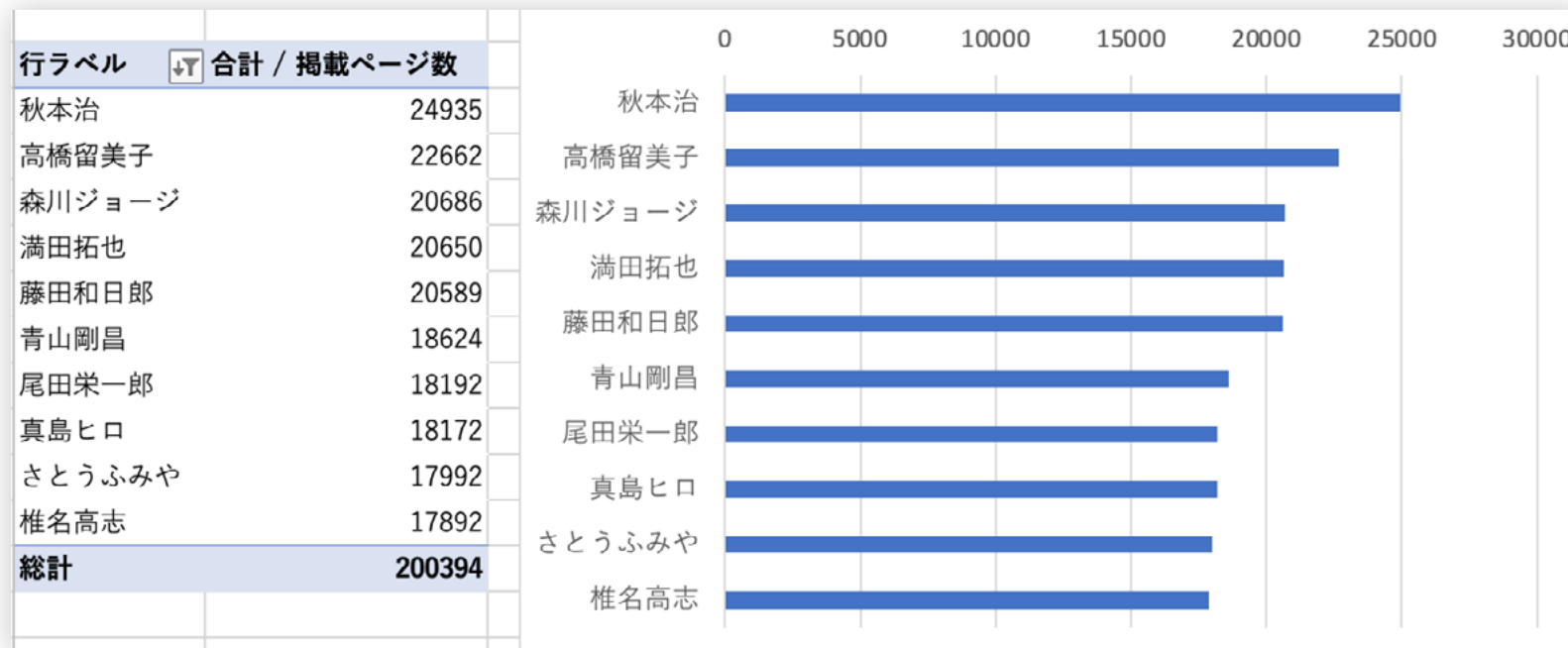
列

Σ 値

合計 / 掲載ページ数

7. 分析例C:掲載ページ数が多い作者10位

7.07 あとはこれまでと同じように、「合計 / 掲載ページ数」でソートして、上位10件を取り出して、グラフを描きましょう。これで作者別のページ数のグラフを描くことができました。ここでの一位はやはりこち亀の「秋本治」さんでした。



#3

データビジュアライズに
チャレンジしよう～応用編



1. 特徴量エンジニアリング

1.01 前回は、もともとのデータを集計するだけで出来る分析をしました。しかし、より高度な分析をするためには、データから特徴量*を自分で設計しなくてはならないことも多いです。今回はそうした処理を必要とする分析をしていきましょう。

2. データを見てみよう

2.01 今回ビジュアライズの題材とするのは、1990年から2017年までのコンシューマゲームの発売タイトル一覧です。

2.02 36,618件のデータがある。

2.03 一つのゲームタイトルをそれぞれ一つのレコードとして保存してある。

ゲームパッケージラベル	ゲームプラットフォーム	出版日	発行者
がんばれゴエモン外伝 きえた黄金キセル	ファミリーコンピュータ	1990-01-05	コナミ株式会社
ヒーロー集合!!ピンボールパーティ	ゲームボーイ	1990-01-12	株式会社ジャレコ
ワールドボウリング	ゲームボーイ	1990-01-13	株式会社アテナ
ZOOM!	メガドライブ	1990-01-13	セガ・エンタープライゼス
SOLARSTRIKER	ゲームボーイ	1990-01-26	任天堂株式会社
ガイフレーム	PCエンジン	1990-01-26	日本コンピュータシステム株式会社
...

2.04 データの列の項目は「URI」「ゲームパッケージラベル」「ゲームプラットフォーム」「公開年月日」「発行者」の5つです。

URI	ゲームパッケージラベル	ゲームプラットフォーム	公開年月日	発行者

* 特徴量:データの属性とその値のこと

3. 分析例A:ゲーム会社別の発売数

3.01 まずは基本の分析をして前回のおさらいをしましょう。ゲーム会社ごとに発売タイトルを集計してどの会社が多くゲームを出しているかをグラフにしてみましょう。

3.02 サンプルデータ<ゲームパッケージ.xlsx>をExcelで開きます。

3.03 全データを選んでピボットテーブルを挿入しましょう。

	A	B	C	D	E
1	URI	ゲームパッケージラベル	ゲームプラットフォーム	公開年月日	発行者
2	https://mediaarts-db.bunka.go.jp/id/M757067	マリオパーティ100 ミニゲームコレクション ダウンロード版	ニンテンドー3DS	2017-12-28	任天堂株式会社
3	https://mediaarts-db.bunka.go.jp/id/M757066	マリオパーティ100 ミニゲームコレクション パッケージ版	ニンテンドー3DS	2017-12-28	任天堂株式会社
4	https://mediaarts-db.bunka.go.jp/id/M744641	アケアカNEOGEO ザ・キング・オブ・ファイターズ '96 ダウンロード版	Nintendo Switch	2017-12-28	ハムスター
5	https://mediaarts-db.bunka.go.jp/id/M743286	The Next Penelope ダウンロード版	Nintendo Switch	2017-12-28	Plug In Digital
6	https://mediaarts-db.bunka.go.jp/id/M751087	タロミア ダウンロード版	Nintendo Switch	2017-12-28	テヨンジャパン合同会社
7	https://mediaarts-db.bunka.go.jp/id/M740219	L.F.O. -Lost Future Omega- ダウンロード版	Nintendo Switch	2017-12-28	メビウス
8	https://mediaarts-db.bunka.go.jp/id/M754741	ヒューマン フォール フラット ダウンロード版	Nintendo Switch	2017-12-28	テヨンジャパン合同会社
9	https://mediaarts-db.bunka.go.jp/id/M740801	Moorhuhn Knights & Castles モーアフーン ナイツ アンド キャッスル	Nintendo Switch	2017-12-28	Young Fun Studio
10	https://mediaarts-db.bunka.go.jp/id/M765592	不思議の幻想郷TOD -RELOADED- ダウンロード版	Nintendo Switch	2017-12-28	UNTIES
11	https://mediaarts-db.bunka.go.jp/id/M750301	スライムの野望 ダウンロード版	Nintendo Switch	2017-12-28	フライハイワークス株式会社
12	https://mediaarts-db.bunka.go.jp/id/M756678	マイティガンヴォルト バースト ダウンロード版	Nintendo Switch	2017-12-28	インティ・クリエイツ
13	https://mediaarts-db.bunka.go.jp/id/M756679	マイティガンヴォルト バースト 体験版; ダウンロード版	Nintendo Switch	2017-12-28	インティ・クリエイツ
14	https://mediaarts-db.bunka.go.jp/id/M746591	カイジ〜絶望の鉄骨渡り〜 for Nintendo Switch ダウンロード版	Nintendo Switch	2017-12-28	ソリッドスフィア株式会社
15	https://mediaarts-db.bunka.go.jp/id/M765994	魔女と勇者III ダウンロード版	ニンテンドー3DS	2017-12-27	フライハイワークス株式会社
16	https://mediaarts-db.bunka.go.jp/id/M766005	魔神少女 エピソード3 -勇者と愚者- ダウンロード版	ニンテンドー3DS	2017-12-27	フライハイワークス株式会社
17	https://mediaarts-db.bunka.go.jp/id/M744316	アーケードアーカイブス フロントライン オンライン配信版	プレイステーション4	2017-12-26	ハムスター
18	https://mediaarts-db.bunka.go.jp/id/M754404	ハロー・レディ! -Superior Dynamis- 体験版 オンライン配信版	プレイステーション Vita	2017-12-25	ヒューネックス株式会社
19	https://mediaarts-db.bunka.go.jp/id/M743618	Unbox: Newbies Adventure オンライン配信版	プレイステーション4	2017-12-22	クロスファンクション株式会社
20	https://mediaarts-db.bunka.go.jp/id/M737863	Call of Duty : WWII + Destiny 2 デジタルデラックスバンドル ダウンロード版	Xbox One	2017-12-22	Activision
21	https://mediaarts-db.bunka.go.jp/id/M741113	Need for Speed アルティメットバンドル ダウンロード版	Xbox One	2017-12-22	エレクトロニック・アーツ株式会社

3. 分析例A:ゲーム会社別の発売数

3.04 ピボットテーブルのフィールドの、①「発行者」を「行」に、②「ゲームパッケージラベル」を「値」にドラッグアンドドロップします。

ピボットテーブルのフィールド

レポートに追加するフィールドを選択してください:

検索

- URI
- ゲームパッケージラベル
- ゲームプラットフォーム
- 公開年月日
- 発行者

その他のテーブル...

次のボックス間でフィールドをドラッグしてください:

フィルター	列
行	Σ 値
発行者	個数 / ゲームパッケージラ...

①

②

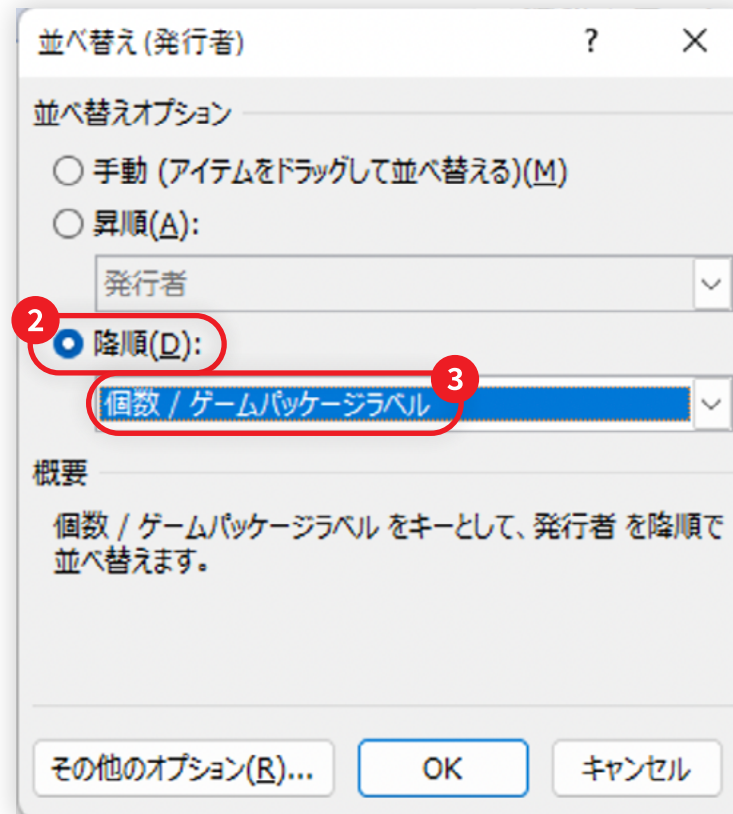
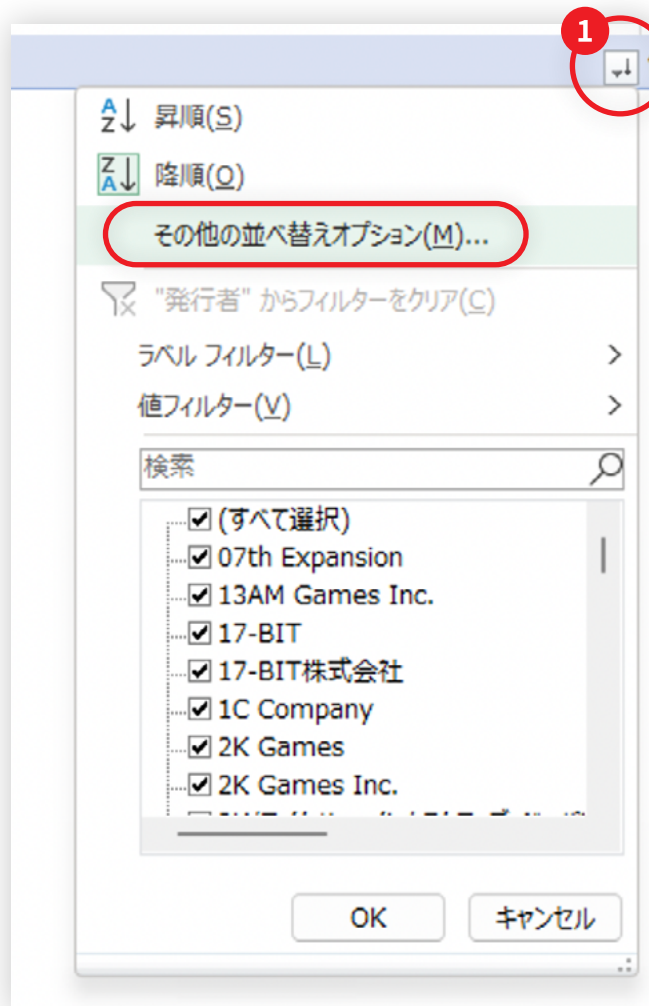
3. 分析例A:ゲーム会社別の発売数

3.05 これでゲーム会社ごとにゲームの発売タイトルを集計することができました。

3	行ラベル	▼ 個数 / ゲームパッケージラベル
4	07th Expansion	10
5	13AM Games Inc.	1
6	17-BIT	1
7	17-BIT株式会社	1
8	1C Company	1
9	2K Games	3
10	2K Games Inc.	18
11	2K/テイクツー・インタラクティブ・ジャパン	2
12	3goo	2
13	505 Games Ltd.	1
14	5pb.	70
15	5pb.,Genterprise	1
16	9003inc.	1
17	ACCESS SOFTWARE INCORPORATED	2
18	ACE Team,W-Russell	1
19	Activision	11
20	Activision Publishing, Inc	1
21	Activision Publishing, Inc.	10
22	Activision, Inc	3

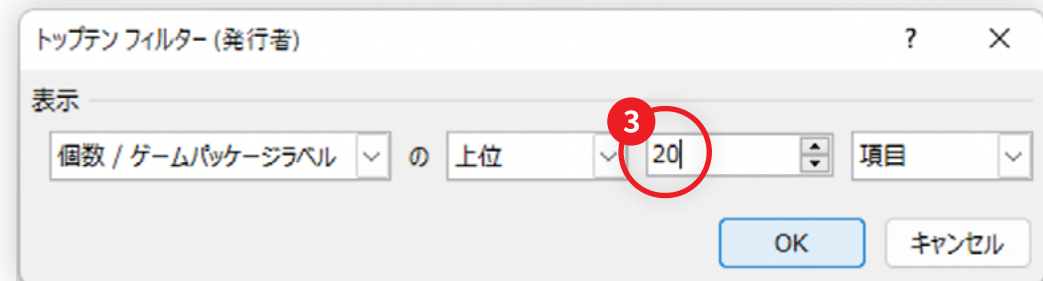
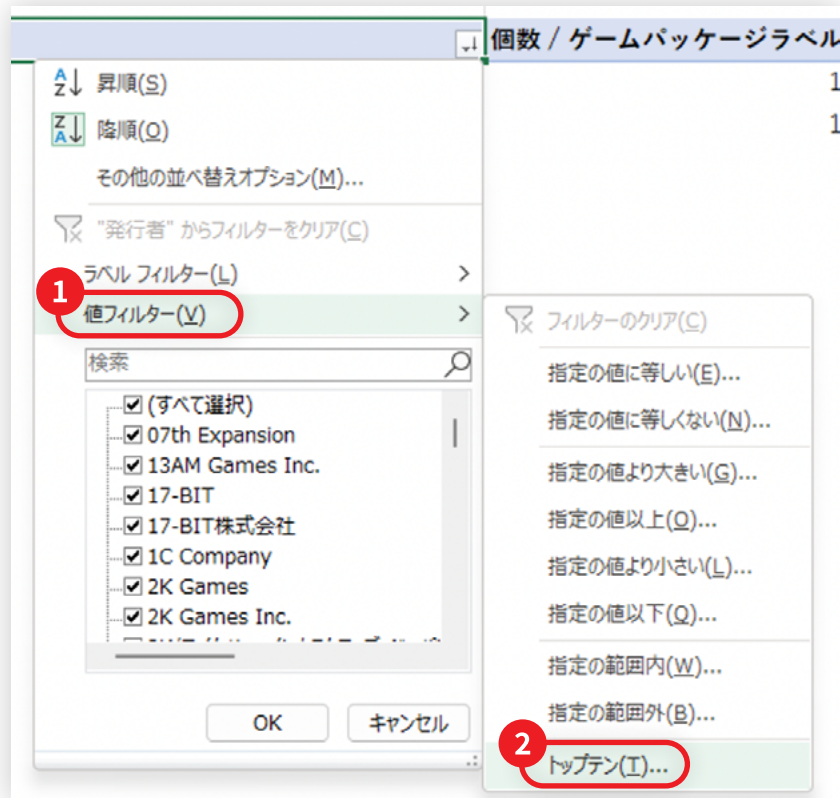
3. 分析例A:ゲーム会社別の発売数

3.06 ①フィルタのボタンを押して、②並び替えの順序に「降順」を選び、③並び替えの対象に「個数 / ゲームパッケージラベル」を選びます。



3. 分析例A:ゲーム会社別の発売数

3.07 上位20件に絞ってみましょう。①「値フィルター」のメニューから②「トップテン」を選び、③項目数に「20」と入力します。



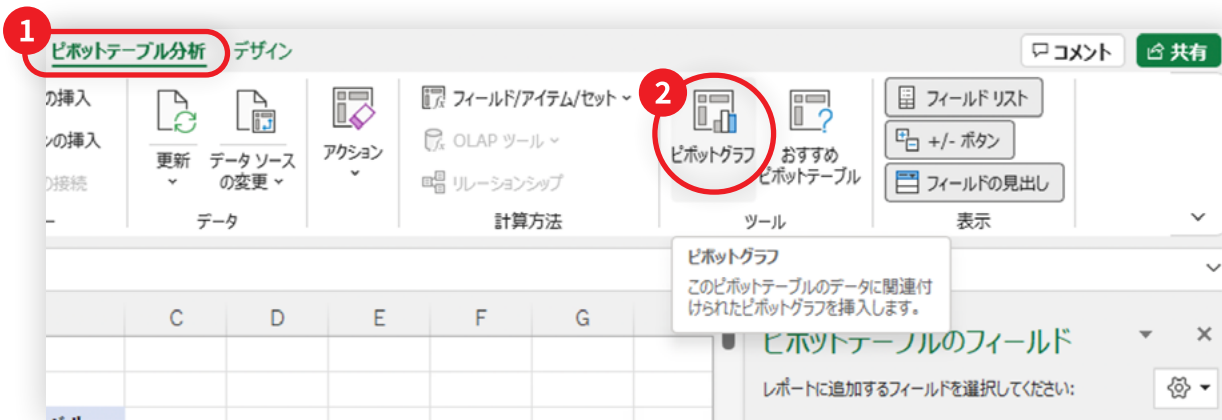
3. 分析例A:ゲーム会社別の発売数

3.08 これで上位20件を表にまとめることが出来ました。

3	行ラベル	個数 / ゲームパッケージラベル
4	任天堂株式会社	1475
5	株式会社バンダイナムコゲームス	1216
6	株式会社カプコン	986
7	株式会社ソニー・コンピュータエンタテインメント	931
8	エレクトロニック・アーツ株式会社	881
9	株式会社スクウェア・エニックス	831
10	株式会社コーエーテクモゲームス	706
11	株式会社セガ	683
12	アイディアファクトリー株式会社	669
13	株式会社ディースリー・パブリッシャー	597
14	株式会社コーエー	523
15	ユービーアイソフト株式会社	500
16	株式会社ハドソン	461
17	アークシステムワークス株式会社	446
18	株式会社コナミデジタルエンタテインメント	442
19	コナミ株式会社	438
20	株式会社バンダイ	397
21	カプコン	393
22	株式会社ハムスター	365
23	セガ・エンタープライゼス	359
24	総計	13299

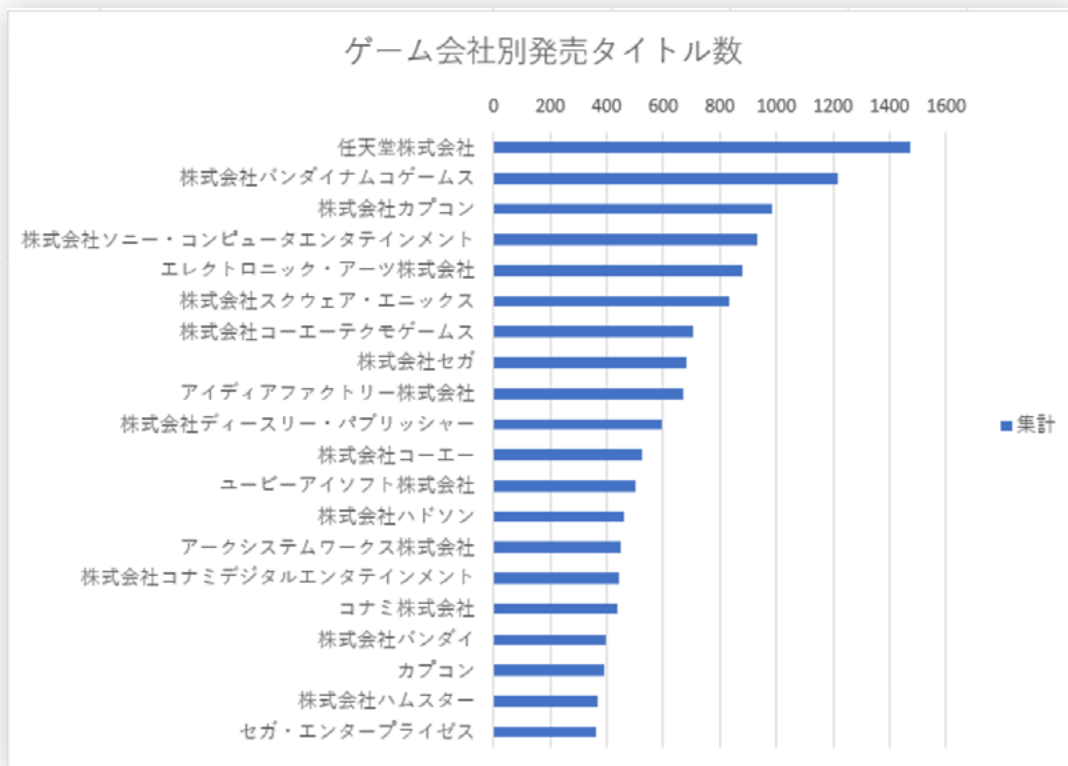
3. 分析例A:ゲーム会社別の発売数

3.09 これをグラフにします。①リボンタブの「ピボットテーブル分析」の②「ピボットグラフ」を選びます。



3. 分析例A:ゲーム会社別の発売数

3.10 これをグラフにします。リボンタブの「ピボットテーブル分析」の「ピボットグラフ」を選びます。



3.11 グラフを見ると ゲーム会社別発売タイトル数の 一位は任天堂で、二位はバンダイナムコゲームスでした。よく見ると、三位の「株式会社カプコン」は、下の方に「カプコン」もあるので、表記揺れがありますね。「株式会社バンダイ」も下の方にあって、これは合併前なので表記ゆれではありませんが、バンダイナムコゲームスと合わせると任天堂より多そうです。

4. 分析例B:年別発売タイトル数

4.01 続いて、特徴量エンジニアリングが必要な例として、発売タイトル数を年別にまとめてみましょう。そしてそれをさらにゲームプラットフォーム別にまとめることで、どのゲーム機が流行していたのかをビジュアライズしてみましょう。

4.02 まずはゲームの公開年月日を年にまとめます。しかし実はこの「公開年月日」のデータは文字列なので、日付データではなく、扱いにくい形になっています。これを日付データに直していきます。

4.03 元データの右に「公開年月日(日付データ)」というタイトルの列を追加します。

D	E	F	G
公開年月日	発行者	公開年月日(日付データ)	
2017-12-28	任天堂株式会社		
2017-12-28	任天堂株式会社		
2017-12-28	ハムスター		
2017-12-28	Plug In Digital		
2017-12-28	テヨンジャパン合同会社		

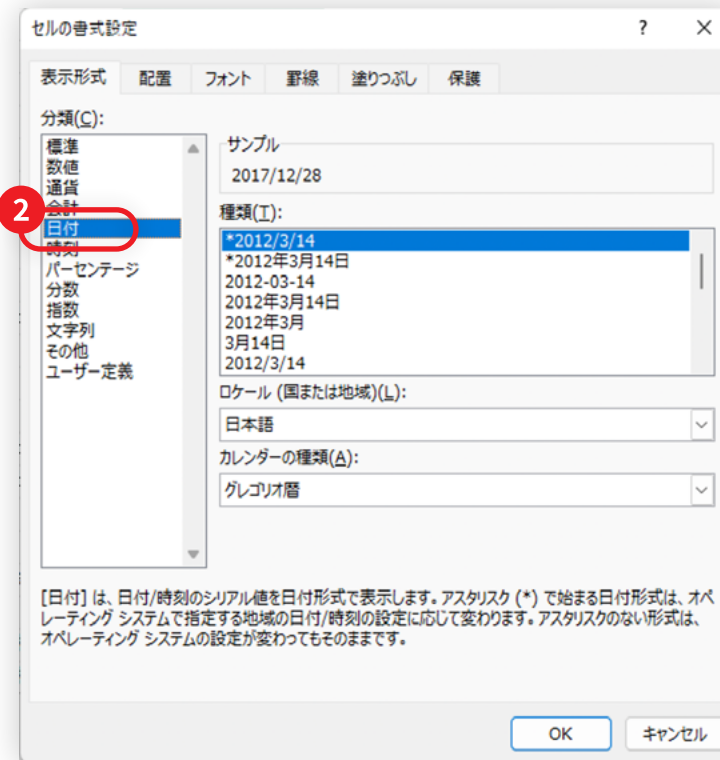
4.04 Excelの関数を使って文字列で書かれた日付を、日付専用のデータに変換します。「DATEVALUE」関数を使います。

D	E	F	G
公開年月日	発行者	公開年月日(日付データ)	
2017-12-28	任天堂株式会社	=DATEVALUE(D2)	
2017-12-28	任天堂株式会社	DATEVALUE(日付文字列)	
2017-12-28	ハムスター		
2017-12-28	Plug In Digital		
2017-12-28	テヨンジャパン合同会社		

4. 分析例B:年別発売タイトル数

4.05 数字になりましたが、これであっています。Excelでは日付は1900年1月1日からの通し番号として扱われています。①「書式」を②「日付」に変更してみましょう。

E	F	メニューの検索
発行者	公開年月日(日付データ)	切り取り(I)
任天堂株式会社	4309	コピー(C)
任天堂株式会社		貼り付けのオプション:
ハムスター		形式を選択して貼り付け(S)...
Plug In Digital		スマート検索(L)
テヨンジャパン合同会社		挿入(I)...
メビウス		削除(D)...
テヨンジャパン合同会社		数式と値のクリア(N)
Young Fun Studio		クイック分析(Q)
UNTIES		フィルター(E) >
フライハイワークス株式会社		並べ替え(Q) >
インティ・クリエイツ		テーブルまたは範囲からデータを...
インティ・クリエイツ		新しいコメント(M)
ソリッドスフィア株式会社		新しいメモ(N)
フライハイワークス株式会社		セルの書式設定(E)...
フライハイワークス株式会社		ドロップダウンリストから選択(K)...
ハムスター		ふりがなの表示(S)
ビューネックス株式会社		
クロスファンクション株式会社		
Activision		
エレクトロニック・アーツ株式会社		
エレクトロニック・アーツ株式会社		



4. 分析例B:年別発売タイトル数

4.06 その上で、数式を下にコピーしましょう。これで文字列データだった公開年月日を日付データに変換することが出来ました。

D	E	F
公開年月日	発行者	公開年月日(日付データ)
2017-12-28	任天堂株式会社	2017/12/28
2017-12-28	任天堂株式会社	2017/12/28
2017-12-28	ハムスター	2017/12/28
2017-12-28	Plug In Digital	2017/12/28
2017-12-28	テヨンジャパン合同会社	2017/12/28
2017-12-28	メビウス	2017/12/28
2017-12-28	テヨンジャパン合同会社	2017/12/28
2017-12-28	Young Fun Studio	2017/12/28
2017-12-28	UNTIES	2017/12/28
2017-12-28	フライハイワークス株式会社	2017/12/28
2017-12-28	インティ・クリエイツ	2017/12/28
2017-12-28	インティ・クリエイツ	2017/12/28
2017-12-28	ソリッドスフィア株式会社	2017/12/28
2017-12-27	フライハイワークス株式会社	2017/12/27
2017-12-27	フライハイワークス株式会社	2017/12/27
2017-12-26	ハムスター	2017/12/26



4. 分析例B:年別発売タイトル数

4.07 このとき、一部の行ではエラーが発生します。ほとんどの行には年・月・日のデータが有るのですが、年と月までの情報しかないものと、年までのデータしか無いものが紛れています。このうち、月までの情報があるものはエラーにはならないのですが、年だけのデータの場合はエラーになってしまいます。ExcelのDATEVALUE関数の仕様では、月までの情報がある場合は、月の1日目として扱われています。

1991-02-01	ビデオシステム株式会社	1991/2/1	1991
1991-02	徳間書店インターメディア株式会社	1991/2/1	1991
1991-02	株式会社ホームデータ	1991/2/1	1991
1991-02	ビクター音楽産業株式会社	1991/2/1	1991
1991-02	ビクター音楽産業株式会社	1991/2/1	1991
1991-01-29	株式会社ナムコ	1991/1/29	1991
1991-01-26	株式会社セガ・エンタープライゼス	1991/1/26	1991
1991-01-25	ハドソン	1991/1/25	1991
1991-01-25	バック・イン・ビデオ	1991/1/25	1991
1991-01-25	株式会社タイトー	1991/1/25	1991
1991-01-25	株式会社タイトー	1991/1/25	1991
1991-01-25	日本コンピュータシステム株式会社	1991/1/25	1991
1991-01-18	タイトー	1991/1/18	1991
1991-01-18	HUDSON SOFT	1991/1/18	1991
1991-01-11	株式会社タイトー	1991/1/11	1991
1991-01-08	株式会社ハル研究所	1991/1/8	1991
1991-01-08	データイースト株式会社	1991/1/8	1991
1991-01-05	コナミ株式会社	1991/1/5	1991
1991-01-03	セガ・エンタープライゼス	1991/1/3	1991
1991-01	株式会社ジャスト	1991/1/1	1991
1991	イマジニア株式会社	#VALUE!	#VALUE!
1991	クエスト	#VALUE!	#VALUE!
1991	SQUARE	#VALUE!	#VALUE!
1990-12-29	バンプレスト	1990/12/29	1990

エラーにはなっていない

エラーになっている

4. 分析例B:年別発売タイトル数

- 4.08** 現実にはこのように不完全なデータが含まれていることはよくあり、こうした問題に対処してデータを使えるようにすることを「データクレンジング」と呼びます。データ分析においてデータクレンジングの占める割合は大きく、人によっては実際の分析をする作業は3割でデータクレンジングが7割と言ったりします。
- 4.09** さて、ここでエラーが出ているデータを修正をすることも出来ますが、このままでも分析自体は可能なので、ひとまずそのまま進めましょう。
- 4.10** 次に、日付データから年を取り出しましょう。これも元データの右に列を追加していきます。列のタイトルは「公開年」としましょう。

D	E	F	G
公開年月日	発行者	公開年月日(日付データ)	公開年
2017-12-28	任天堂株式会社	2017/12/28	
2017-12-28	任天堂株式会社	2017/12/28	
2017-12-28	ハムスター	2017/12/28	
2017-12-28	Plug In Digital	2017/12/28	
2017-12-28	テヨンジャパン合同会社	2017/12/28	

4. 分析例B:年別発売タイトル数

4.11 日付データから年の部分を取り出します。関数には「YEAR」を使います。

E	F	G
発行者	公開年月日(日付データ)	公開年
任天堂株式会社	2017/12/28	=YEAR(F2)
任天堂株式会社	2017/12/28	YEAR(シリアル値)
ハムスター	2017/12/28	

4.12 更に全体にコピーします。これでゲームの発売日を年別に集計する準備が整いました。

E	F	G
発行者	公開年月日(日付データ)	公開年
任天堂株式会社	2017/12/28	2017
任天堂株式会社	2017/12/28	2017
ハムスター	2017/12/28	2017
Plug In Digital	2017/12/28	2017
テヨンジャパン合同会社	2017/12/28	2017
メビウス	2017/12/28	2017
テヨンジャパン合同会社	2017/12/28	2017

4. 分析例B:年別発売タイトル数

- 4.13 シート全体を選択してピボットテーブルを挿入します。そうすると前回とは違って、追加した列の分も使えるようになっています。
- 4.14 年別に集計するのですから、まずは「公開年」を「行」に入れてみましょう。するとこうなります。

行ラベル						
1990						
1991						
1992						
1993						
1994						
1995						
1996						
1997						
1998						
1999						
2000						
2001						
2002						
2003						
2004						
2005						
2006						
2007						

ピボットテーブルのフィールド

レポートに追加するフィールドを選択してください:

検索

- 公開年月日
- 発行者
- 公開年月日(日付データ)
- 公開年

次のボックス間でフィールドをドラッグしてください:

フィルター	列

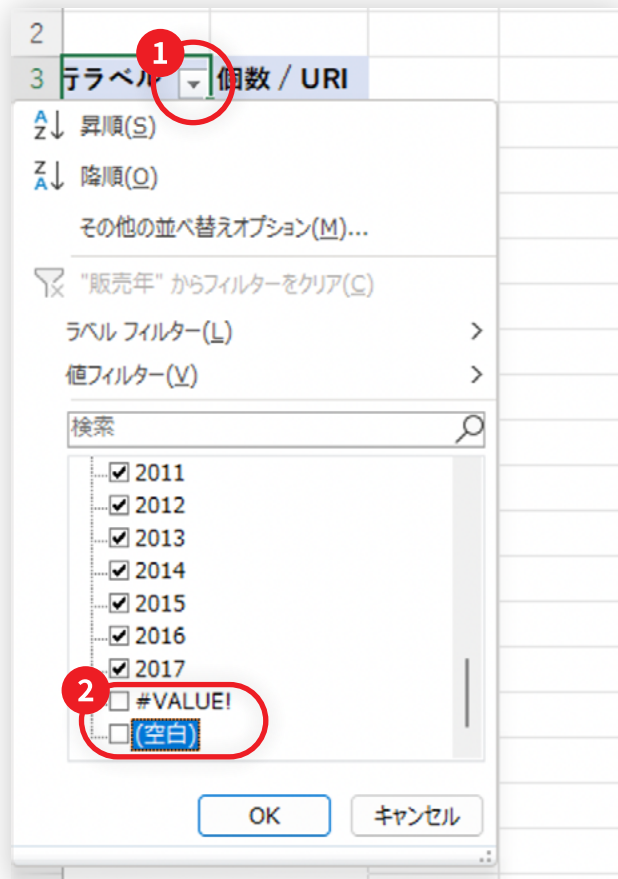
行ラベル	Σ 値
公開年	

レイアウトの更新を保留する 更新

4. 分析例B:年別発売タイトル数

4.15 発売年の下のところを見ると、エラーになっていたところも集計されているのが分かります。ここでは、エラーになっているものを除外してみましょう。

- ①「行ラベル」のところにあるフィルタから、②エラーに該当するところのチェックを外すことで、除外することが出来ます。ついでに「空白」の行も除外しておきましょう。



4. 分析例B: 年別発売タイトル数

4.16 ここで、「URI」(「ゲームパッケージラベル」でもよい)を「値」に入れると個数をカウントできるので、まず年別の発売 タイトル 数を集計することができます。

行ラベル	個数 / URI
1990	413
1991	429
1992	481
1993	472
1994	659
1995	775
1996	871
1997	903
1998	1022
1999	1224
2000	1306
2001	1110
2002	1221
2003	1122
2004	1086
2005	1149
2006	1286
2007	1704

ピボットテーブルのフィールド

レポートに追加するフィールドを選択してください:

検索

- ゲームパッケージラベル
- ゲームプラットフォーム
- 公開年月日
- 発行者
- 公開年月日(日付データ)
- 公開年

その他のテーブル...

次のボックス間でフィールドをドラッグしてください:

フィルター	列
行	Σ 値
公開年	個数 / URI

4. 分析例B:年別発売タイトル数

4.17 この結果だけ見ると、ゲームの発売数は年々増えているのが分かります。

4.18 そしてピボットテーブルの新たな使い方ですが、こんどは「ゲームプラットフォーム」を「列」にドラッグアンドドロップします。すると、ゲームのプラットフォームごとに発売タイトル数を分けて集計することが出来ます。

個数 / 公開年	列ラベル				
行ラベル	3DO	64DD	Classic Mac OS	macOS	macOS, M
1990					
1991					
1992					
1993					
1994	60				
1995	52			1	
1996	6				
1997					
1998				1	
1999			1		
2000			2		
2001					
2002					
2003					
2004					
2005					
2006					

ピボットテーブルのフィールド

レポートに追加するフィールドを選択してください:

検索

- ゲームプラットフォーム
- 公開年月日
- 発行者
- 公開年月日(日付データ型)
- 公開年

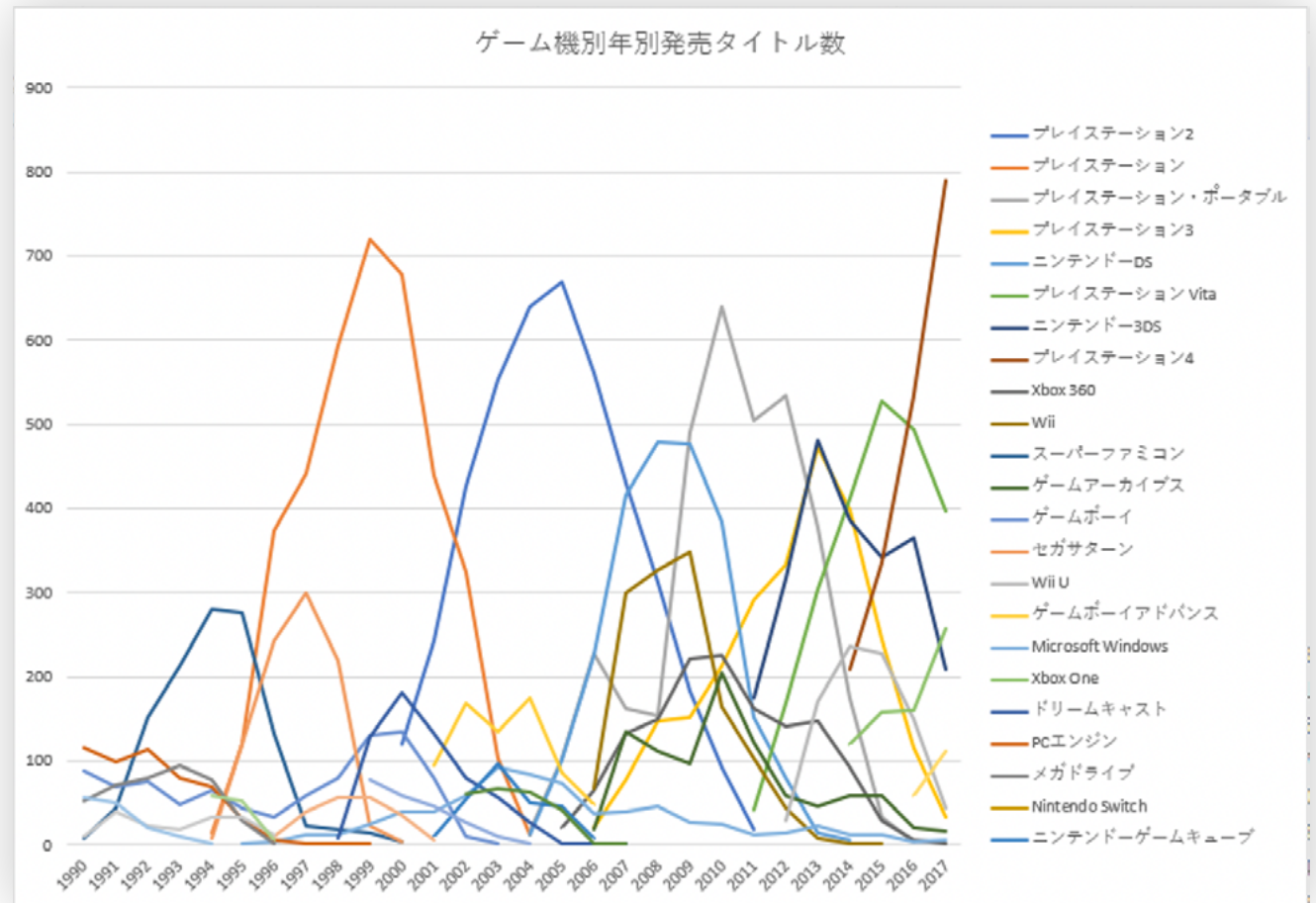
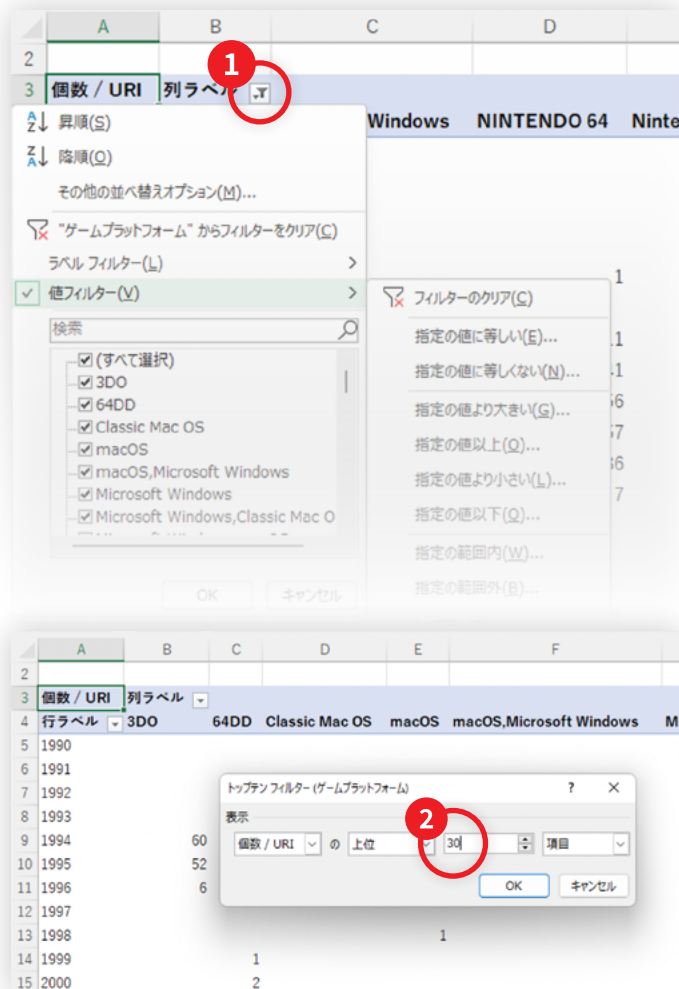
次のボックス間でフィールドをドラッグしてください:

フィルター []	列 ゲームプラットフォーム
行 公開年	Σ 値 個数 / 公開年

レイアウトの更新を保留する 更新

4. 分析例B: 年別発売タイトル数

4.19 集計するゲームプラットフォームは、①列ラベルのフィルタを使って制限することができます。ここでは②上位30件としてみました。そしてこれを折れ線グラフでビジュアライズするとこうなります。



4. 分析例B:年別発売タイトル数

- 4.20** 各ゲームプラットフォームが、盛り上がっては衰退していく様がよく見えます。
- Nintendo Switchは、まだ販売されたばかりで2017年までのデータしかありませんが、このあと右肩上がりすることになります。
 - プレイステーションとプレイステーション2の波は似ていて、その次に来ているのはプレイステーション3ではなくてプレイステーション・ポータブルであるように見えます。

5. 分析例C:テキスト分析

- 5.01** 少し趣向を変えて、テキストに注目した分析をしてみましょう。ここでテキスト分析に使えるのは「ゲームパッケージラベル」ですね。これはゲームのタイトルに相当します。
- 5.02** ゲームやファンタジー小説でいわゆる「属性」を表す言葉として使われている「火」「水」「風」「土」という言葉のうち、ゲームのタイトルに一番登場するのはどれでしょうか？
- 一番登場しなさそうなのは土ですけど、他は分かりませんね。

5. 分析例C:テキスト分析

5.03 ではこれを集計してみましょう。この集計は、ピボットテーブルを離れて、エクセルの関数で行います。

- ① COUNTIF関数を使います。
- ② 以下のように書くと「火」が含まれるタイトルを数え上げることが出来ます。

	A	B	C	D	E	F
1						
2						
3			=COUNTIF(元データ!\$B:\$B,"*火*")			
4						

- ③ COUNTIF関数の1つ目の引数は、データの場所を指しています。ここでは、後々のことを考えて絶対参照にしています。

fx | =COUNTIF(元データ!\$B\$2:\$B\$36619,"*火*")

シート名
ゲームパッケージラベルの範囲

	A	B	C	D
1	URI	ゲームパッケージラベル	ゲームプラットフォーム	公開年月日
2	https://mediaarts-db.bunka.go.jp/id/M757067	マリオパーティ100 ミニゲームコレクション ダウンロード版	ニンテンドー3DS	2017-12-28
3	https://mediaarts-db.bunka.go.jp/id/M757066	マリオパーティ100 ミニゲームコレクション パッケージ版	ニンテンドー3DS	2017-12-28
4	https://mediaarts-db.bunka.go.jp/id/M744641	アケアカNEOGEO ザ・キング・オブ・ファイターズ '96 ダウンロード版	Nintendo Switch	2017-12-28
5	https://mediaarts-db.bunka.go.jp/id/M743286	The Next Penelope ダウンロード版	Nintendo Switch	2017-12-28
6	https://mediaarts-db.bunka.go.jp/id/M751087	タロミア ダウンロード版	Nintendo Switch	2017-12-28
7	https://mediaarts-db.bunka.go.jp/id/M740219	L.F.O. -Lost Future Omega- ダウンロード版	Nintendo Switch	2017-12-28
8	https://mediaarts-db.bunka.go.jp/id/M754741	ヒューマンファール フラット ダウンロード版	Nintendo Switch	2017-12-28
9	https://mediaarts-db.bunka.go.jp/id/M740801	Moorhuhn Knights & Castles モアファンナイツ アンド キャ	Nintendo Switch	2017-12-28
10	https://mediaarts-db.bunka.go.jp/id/M765592	不思議の幻想園TOD -RELOADED- ダウンロード版	Nintendo Switch	2017-12-28
11	https://mediaarts-db.bunka.go.jp/id/M756301	スライムの新選 ダウンロード版	Nintendo Switch	2017-12-28

- ④ 2つ目の引数は、「火」という文字を含む文字列を意味しています。「*(アスタリスク)」はワイルドカードと言って、何が入っても入らなくてもいいということを表します。

fx | =COUNTIF(元データ!\$B\$2:\$B\$36619,"*火*")

↑
数え上げるもの

"*火*" : 「火」という文字を含む文字列
***** : 任意の0文字以上の文字列

(例) 以下のどれも該当する

- ああ火ああ
- 火ああ
- あ火
- 火

- ⑤ 結果は以下のようになりました。50個のゲームタイトルが「火」を含んでいるようです。

	A	B	C	D	E	F
1						
2						
3			50			
4						

5. 分析例C:テキスト分析

5.04 他の文字でもやってみましょう。ここで、毎回関数を書き直すのは面倒なので、検索する文字の方も参照を使って書くことにします。

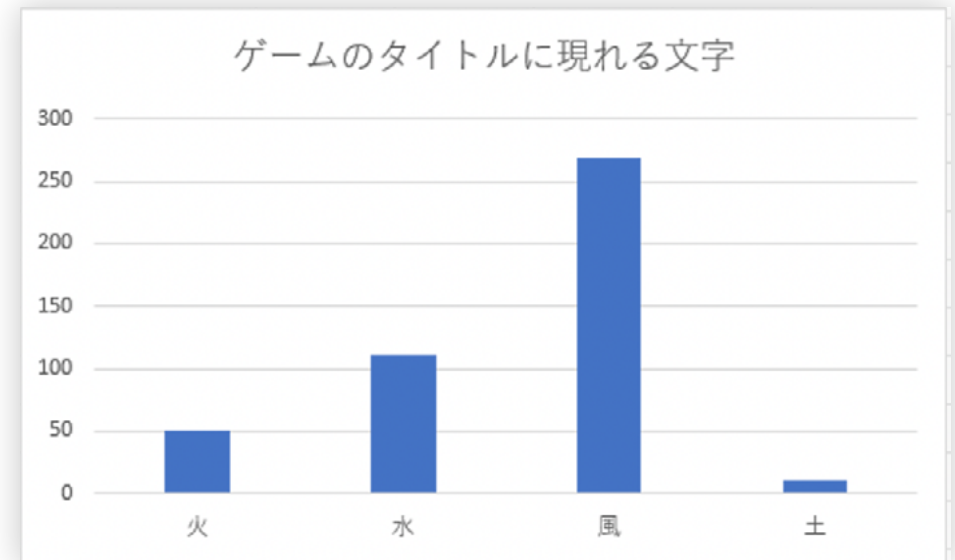
① 文字列は「&(アンド)」で結合することが出来ます。参照先のアドレスと*(アスタリスク)を文字列で結合します。以下のようにします。

	A	B	C	D	E	F
1						
2						
3		火	=COUNTIF(元データ!\$B:\$B,"*"&B3&"*")			
4						

② では他の文字でもやってみましょう。「火」「水」「風」「土」の数を数えてみます。

火	50
水	111
風	269
土	10

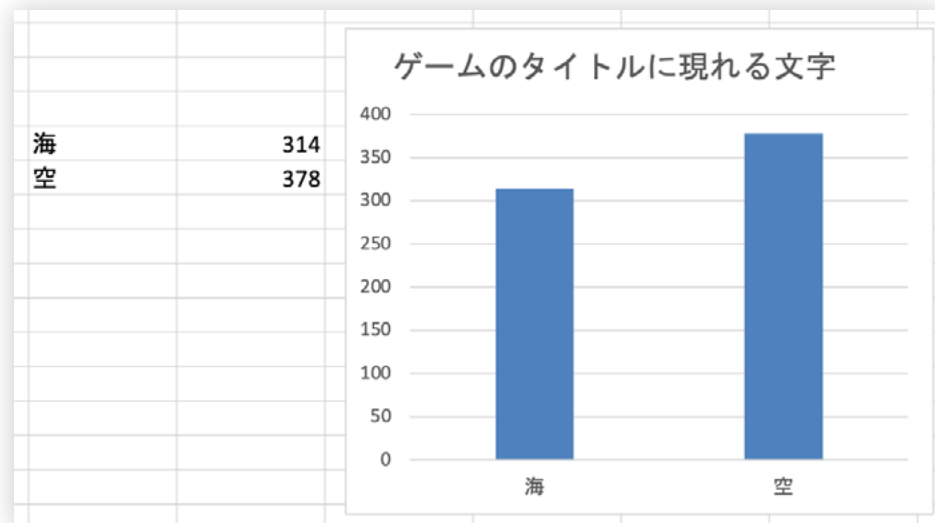
③ 一番多いのは風でした。土は予想通りすごく少ないですね。グラフにもしてみましょう。棒グラフがいいですね。



5. 分析例C:テキスト分析

5.05 他の例でも試してみましょう。

① 「海」と「空」ではどちらが多いでしょうか。僅差で空の方が多いという結果になりました。



5. 分析例C:テキスト分析

5.06 では最後に、先程作った公開年のデータを使って、テキスト分析と時系列分析を組み合わせてみましょう。COUNTIFS関数を使うと、検索条件を複数指定することが出来ます。片方は今やったテキストの検索を行って、もう一つの条件で年を指定しましょう。

① ここでは、「竜」と「ドラゴン」が、それぞれどれぐらい出てくるかを時系列に沿って集計してみます。

② まず、年と検索する文字列を表のラベルに設定します。

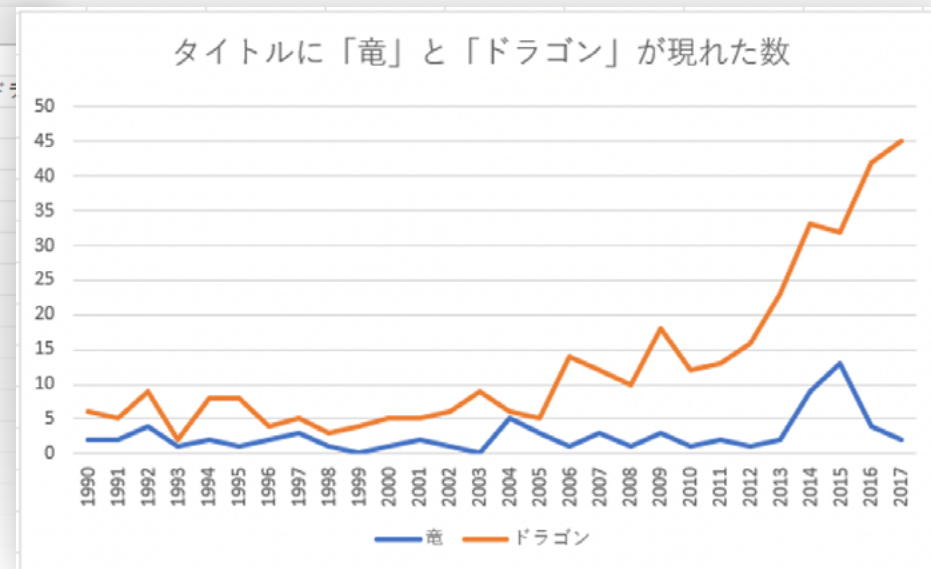
	A	B	C	D	E
1					
2			竜	ドラゴン	
3		1990			
4		1991			
5		1992			
6		1993			
7		1994			
8		1995			
9		1996			
10		1997			
11		1998			
12		1999			
13		2000			
14		2001			
15		2002			
16		2003			
17		2004			
18		2005			
19		2006			
20		2007			
21		2008			

③そして、COUNTIFSを使って、縦横両方の検索条件にヒットするものを数え上げます。

	A	B	C	D	E	F	G	H	I	J
1										
2			竜	ドラゴン						
3		1990	=COUNTIFS(元データ!\$G:\$G,\$B3,元データ!\$B:\$B,"*"&C\$2&"*")							
4		1991	COUNTIFS(検索条件範囲1, 検索条件1, [検索条件範囲2, 検索条件2], [検索条件範囲3, ...])							
5		1992								

④結果は以下のとおりです。基本的にドラゴンの方が多くてさらにドラゴンは最近増えていますね。

	A	B	C
1			
2			竜
3		1990	2
4		1991	2
5		1992	4
6		1993	1
7		1994	2
8		1995	1
9		1996	2
10		1997	3
11		1998	1
12		1999	0
13		2000	1
14		2001	2
15		2002	1
16		2003	0



5. 分析例C:テキスト分析

5.07 これでもいいのですが、一年間に発売されるゲームの数は増えているので、全体としてこれらの文字がよく出てくるようになってきているのかはこれでは分かりません。そこで、ゲーム数全体から、「ドラゴン」「竜」が含まれているものの割合を計算します。

5.08 まず、発売されるゲームの年ごとの総数を調べます。年の列の横に総数の列を作ります。

	総数	竜	ドラゴン
1990		2	6
1991		2	5
1992		4	9
1993		1	2
1994		2	8
1995		1	8
1996		2	4
1997		3	5
1998		1	3
1999		0	4
2000		1	5
2001		2	5
2002		1	6
2003		0	9
2004		5	6
2005		3	5
2006		1	14
2007		3	12
2008		1	10

5. 分析例C:テキスト分析

5.09 元データの「発売年」の列のデータには年の情報が入っていますので、その年が何回出現するかを数えることで販売されたタイトルの数を調べることができます。COUNTIF関数を使って集計するには、列の全データから、ある年に一致するものをカウントします。あとは数式を下にコピーしましょう。

	総数	竜	ドラゴン
1990	=COUNTIF(元データ!G:G,\$B3)		
1991	COUNTIF(範囲, 検索条件)		5
1992		4	9
1993		1	2
1994		2	8
1995		1	8
1996			

	総数	竜	ドラゴン
1990	413	2	6
1991	429	2	5
1992	481	4	9
1993	472	1	2
1994	659	2	8
1995	775	1	8
1996	871	2	4
1997	903	3	5
1998	1022	1	3
1999	1224	0	4
2000	1306	1	5
2001	1110	2	5
2002	1221	1	6
2003	1122	0	9
2004	1086	5	6
2005	1149	3	5
2006	1286	1	14
2007	1704	3	12
2008	1728	1	10

5. 分析例C:テキスト分析

5.10 次に、総数と言葉の出現数を使って割合を計算します。①「竜(割合)」と「ドラゴン(割合)」という列を追加します。そして、②各出現数を総数で割ります。

	総数	竜	ドラゴン	竜(割合)	ドラゴン(割合)
1990	413	2	6		
1991	429	2	5		
1992	481	4	9		
1993	472	1	2		

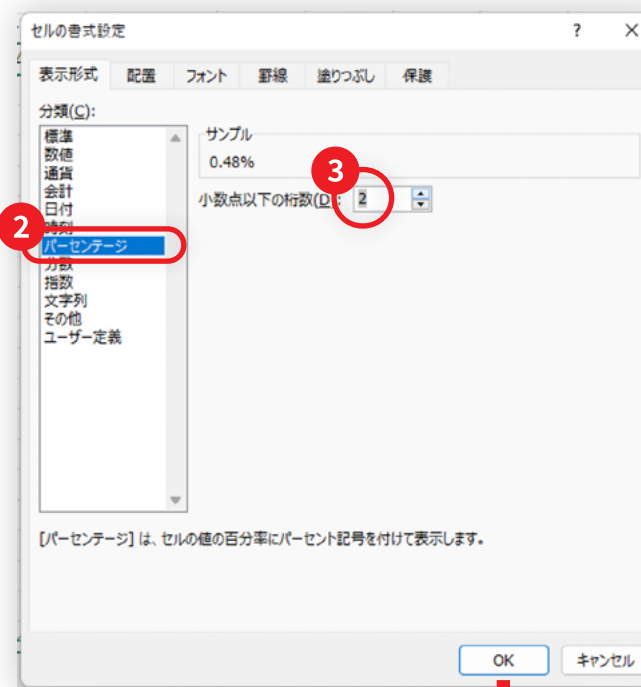
	総数	竜	ドラゴン	竜(割合)	ドラゴン(割合)
1990	413	2	6	=D3/C3	
1991	429	2	5		
1992	481	4	9		
1993	472	1	2		

	総数	竜	ドラゴン	竜(割合)	ドラゴン(割合)
1990	413	2	6	0.004842615	=E3/C3
1991	429	2	5		
1992	481	4	9		
1993	472	1	2		

5. 分析例C:テキスト分析

5.11 このままでもいいのですが、少数の桁はこれほど細かくは必要ないので、パーセンテージ表記にしてみます。①「セルの書式設定」を選んで、②「パーセンテージ」を選び、③「小数点以下の桁数」を2とします。

	総数	竜	ドラゴン	竜(割合)	ドラゴン(割合)
1990	413	2	6	0.004842615	0.014527845
1991	429	2	5		
1992	481	4	9		
1993	472	1	2		
1994	659	2	8		
1995	775	1	8		
1996	871	2	4		
1997	903	3	5		
1998	1022	1	3		
1999	1224	0	4		
2000	1306	1	5		
2001	1110	2	5		
2002	1221	1	6		
2003	1122	0	9		
2004	1086	5	6		
2005	1149	3	5		
2006	1286	1	14		
2007	1704	3	12		
2008	1728	1	10		

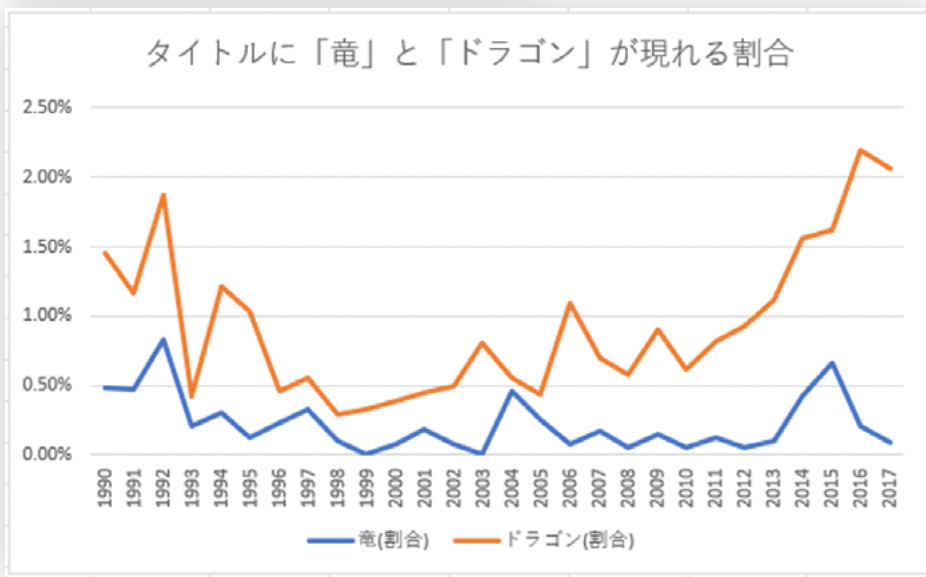


	総数	竜	ドラゴン	竜(割合)	ドラゴン(割合)
1990	413	2	6	0.48%	1.45%
1991	429	2	5		
1992	481	4	9		
1993	472	1	2		

5. 分析例C:テキスト分析

5.12 あとはデータを下にコピーすると集計は終了です。グラフにもしてみます。

年	総数	竜	ドラゴン	竜(割合)	ドラゴン(割合)
1990	413	2	6	0.48%	1.45%
1991	429	2	5	0.47%	1.17%
1992	481	4	9	0.83%	1.87%
1993	472	1	2	0.21%	0.42%
1994	659	2	8	0.30%	1.21%
1995	775	1	8	0.13%	1.03%
1996	871	2	4	0.23%	0.46%
1997	903	3	5	0.33%	0.55%
1998	1022	1	3	0.10%	0.29%
1999	1224	0	4	0.00%	0.33%
2000	1306	1	5	0.08%	0.38%
2001	1110	2	5	0.18%	0.45%
2002	1221	1	6	0.08%	0.49%
2003	1122	0	9	0.00%	0.80%
2004	1086	5	6	0.46%	0.55%
2005	1149	3	5	0.26%	0.44%
2006	1286	1	14	0.08%	1.09%
2007	1704	3	12	0.18%	0.70%
2008	1728	1	10	0.06%	0.58%



5.13 前回は、もともとのデータを集計するだけで出来る分析をしました。しかし、より高度な分析をするためには、データから特徴量を自分で設計しなくてはならないことも多いです。今回はそうした処理を必要とする分析をしていきましょう。

5.14 これを見ると、2000年前後に竜もドラゴンも両方少なくなっている時期がありますね。

5.15 竜やドラゴンはファンタジーの象徴のような存在なので、もしかするとこの時期は、王道のファンタジーとは違った世界観の作品を各メーカーが作ろうとしていたのかもしれないね。

5.16 このように、テキストを分析をすることで、みんなが好むものや、そのトレンドの変化などが見えてきます。

データを使ったビジュアライズ、楽しんでいただけたでしょうか。ここで提示したのは、このデータでできる分析の一部でしかありません。ぜひ皆さんも面白い分析を考えてみてください。

#4

MADBのデータは
自由に使ってもいいの？



#4 MADBのデータは自由に使ってもいいの？

多くのデータ同様、MADBのデータも他のデータと組み合わせて利用することがあると思います。そこで、MADBを含め様々なデータの利用に際して注意すべき点を著作権に関してみていきましょう。

なお、著作権の他にも、他人のデータベースやデータセットを利用する場合に法律的な観点から気を付けるべき点が存在している場合があります。学習や学術研究の範囲でデータを扱っている場合には適用を受けないことが多いですが、気になる場合は調べたり専門機関・専門家に相談したりしてみましょう。



1. データをダウンロードしたり、改変すると著作権侵害になる？

1.01 原則：他人の著作物は、無断で利用すると著作権侵害になります。

※あくまで原則なので、例外もあります。

1.02 「著作物」でないもの「利用」に当たらない行為などは問題ない、という点もやや重要です。

1.03 データサイエンスの授業との関係で考える要注意ケース

要注意ケース1：情報解析に該当しない利用をしようとしている

要注意ケース2：データの内容が著作物(画像、文章など)

要注意ケース1と2両方に該当する場合は特に要注意です。

2. 著作物とは

2.01 著作物は、「思想や感情の創作的な表現」を指すというのが一般的な定義です。

2.02 この定義で言う「創作的」というのは独創性とか、高い美的価値などではなく、非常に創作性が低いレベルの表現、作品でも著作物とされることがあります。誰かが気軽にとった写真とか、ネットに投稿したつぶやきなどであっても、著作物とされる可能性があります。

2.03 「表現」はわかりやすいところでは、言語による表現であれば字句の選択のような外形的な要素も指しています。ただ、それだけを指しているわけではなく、文章の構成や内容の取捨選択なども表現の一種とみなされることがあります。つまり「他人の文章でもちょっと言葉をおき換えればもう別の著作物だから、何をしても著作権侵害にならない」となるとは限らない点に注意が必要です。

3. 利用とは

3.01 コピー、改変、ネットでの公開、公の上映や上演などを指します。

4. データと著作権

4.01 MADBに収納されているデータはメタデータと呼ばれるような情報です。作品そのものや作品の内容ではなく、作品についての、公開年月日とか、出版者名などのデータです。

4.02 一般的にメタデータは著作物に該当しない場合が多いです。データの多くは思想や感情の表現ではなく、事実そのものであったり、表現であっても創作的な表現ではなかったりするためです。そこで、データを勝手にコピーしたり、改変したとしても、著作権侵害になることは稀です。

4.03 そこで、MADBのデータの大半は自由に使っても著作権侵害にならないでしょう。ただ、あえて言えば作品のタイトルには、長くて、十分に個性が感じられて、それだけで(中身などは関係なく)著作物と見なせるものがあるかも知れません。

5. オープンデータと著作権

- 5.01 データの中には著作物にあたるものもあります。また、MADBに収納されている個々のデータはおおよそ著作物ではないと上に述べましたが、データの取捨選択・配列やデータベースの構成など、個々のデータ項目とは別の側面について著作権が認められることがあります。データセット、データベースを利用する時にはその点にも注意が必要な場合があります。
- 5.02 オープンデータは、改変や商業利用を含め多様な利用を誰でもできるように許諾しており、利用の条件も一定範囲のものに限られているようなデータを指します。データセットが著作物であっても、このような許諾がついたオープンデータならかなりいろいろな使い方ができます。
- 5.03 MADBも、こうしたオープンデータの一つです。具体的には、政府標準利用規約(第2.0版)と呼ばれる利用規約の対象になっています。この規約では、データの提供者の明記、改変した場合はその旨の明記、などの条件はあるものの、利用者や利用の用途などは制限がなく、誰でも、商業利用なども含めて利用できるようになっています。
- 5.04 文化庁はデータベースの構築主体ですから、データベースの構成やデータの取捨選択について創作性があれば、それについての著作権は文化庁の著作権になるでしょう。ですが、政府標準利用規約(第2.0版)に従っておけば、文化庁の著作権を侵害することにはなりません。
- 5.05 世界的にも、政府が保有するデータや作成するデータはこのようなオープンデータとして提供する動きが広がっています。データを公開した上で、誰にでも使えるように許諾を与えます。



#5

データビジュアライズからの
発見を伝えよう



この教材のまとめとして、メディア芸術データベースを活用したデータ分析を通じて、皆さんが発見したことを伝えるためのプレゼンテーションを作成します。

プレゼンテーションする相手を想定したうえで、明確でわかりやすく、説得力のあるデータのビジュアライゼーション(表現)の工夫についても学びます。

また、授業やグループ学習でこの教材を使う場合には、プレゼンテーションをロールプレイで行い、想定される質問への対応方法についても実践します。



1. 構成を考える

1.01 まず、プレゼンテーションの全体構成を考えます。相手に伝えたい内容をわかりやすく論理的に説明するための構成の例としては、次のような4つの手順が考えられます。

- ① データ分析の背景と目的
- ② 想定した仮説
- ③ データ分析結果
- ④ データ分析結果の考察

1.02 4つの手順に従って、具体的にテキスト(言葉)で説明を書き出してみましょう。

① データ分析の背景と目的

なぜこのテーマで分析しようと考えたのか、背景を説明します。また、この分析により何が明らかにしようとしているのか、目的を明示します。

② 想定した仮説

この分析によってどのような結果になることを予想しているのか。また、その結果は、何に対してどのような効果や影響をもたらすことが期待できるかについて述べます。

③ データ分析結果

データ分析の手順と結果を説明します。分析対象としたデータをどこから入手したものであるのか、そしてデータ項目、対象期間など対象としたデータの範囲についても明示しましょう。コマ3で学習したように、入手したデータそのものを分析したのではなく、そのデータを用いて特徴量エンジニアリングを行った際は、その旨もまとめておきましょう。

次に、分析結果をビジュアライズしたグラフや図式について、そこから得られた発見がより効果的、魅力的に伝わるような工夫をしましょう。具体的な工夫については「2. データビジュアライゼーションの工夫や配慮」で説明します。

④ データ分析結果の考察

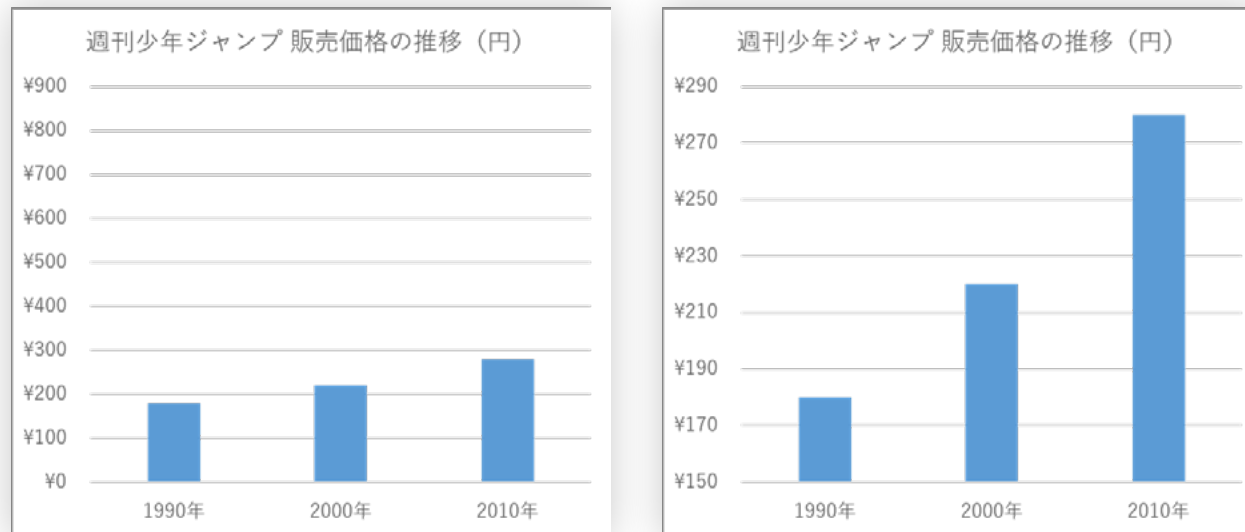
③の分析結果から何が分かったか②の仮説に対して、想定していた通りだったと言えるのかどうか、考察をまとめます。なお、必ずしも仮説通りの結果が得られなくてもよい、ということを忘れずに。たとえ、結果が当初考えた通りでなくても、「仮説とは違う」ということを新たに発見できたこととなります。さらに、その考察から、①のテーマに対して何がいえるのか、まとめとなる結論を導き出します。

2. データビジュアライゼーションの工夫や配慮

2.01 ここでは、分析結果をわかりやすく効果的に伝えるためのコミュニケーション技術としてのデータビジュアライゼーションについて、どのような工夫ができるか、またどのような点に気を付けるべきかについて解説します。

2.02 なぜ、データビジュアライゼーションは大事なのでしょう。その理由を考えるためにいくつかの例をみてみましょう。以下に示す2つのグラフをみてみましょう。それぞれのグラフから何がわかりますか。2つのグラフを比較してみた際に、その表現方法にはどのような違いがありますか。

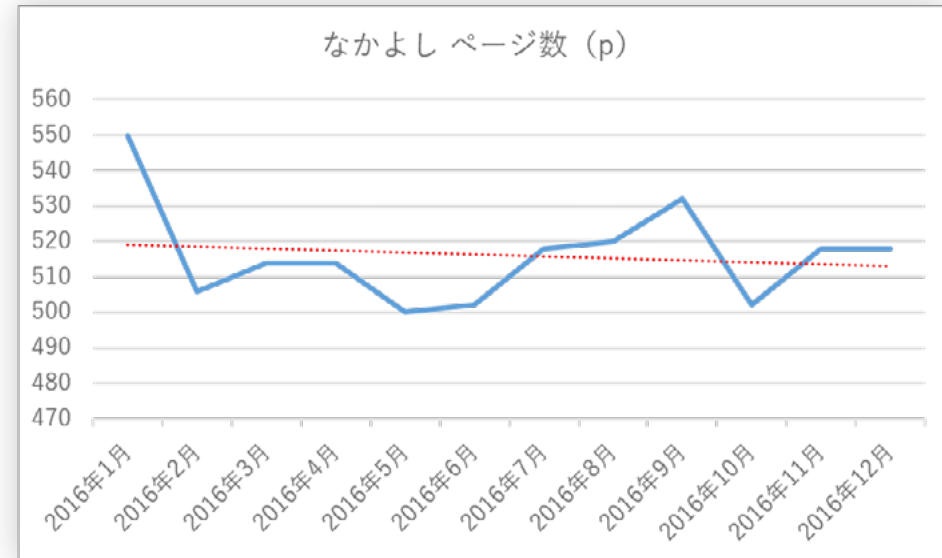
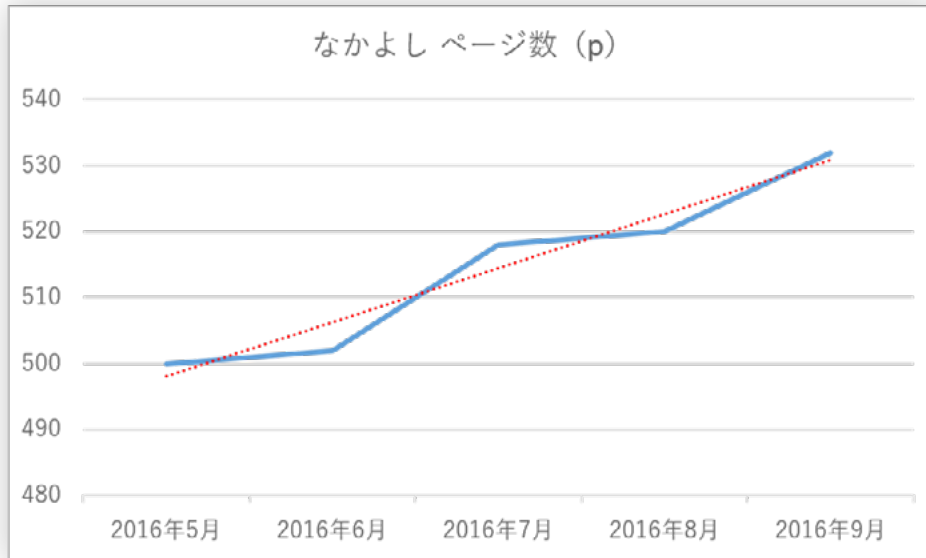
例1)



- 同じデータを示しているのに、左側は価格推移の変動がゆるやかに、右側は価格が大幅に上がっているように見えます。
- 左側の目盛のつけ方がちがうことがわかります。

2. データビジュアライゼーションの工夫や配慮

例2)



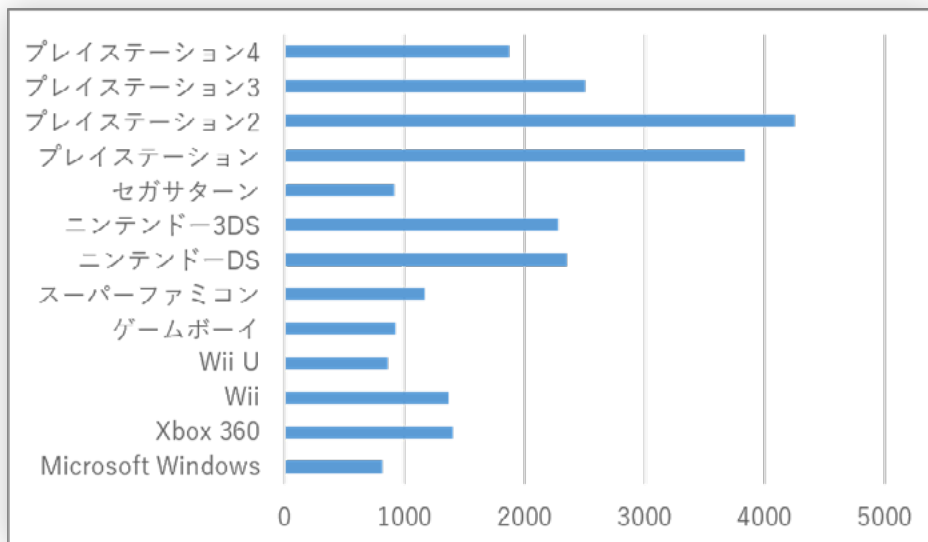
→ 同じ少女月刊誌の総ページ数データを示しているのに、左はページ数が増えているように見えます。右のグラフからはページ数はやや減っていることがわかります。こうした違いは、分析対象としたデータの期間が異なることから発生してしまいます。

2. データビジュアライゼーションの工夫や配慮

- 2.03** これらの例からわかるように、データは示し方によって伝わり方や伝わる内容が、簡単に変わってしまいます。それが意図しない内容として伝わるが無いように留意し、正しく伝えるためにできる工夫や配慮について十分に検討しましょう。
- 2.04** もうひとつ、気を付けるべきことがあります。自分が伝えたいことにあうように、結果を歪めるようなデータの伝え方や切り取り方をするような操作は絶対に行ってはいけません。こうした行為は、事実と異なる内容から誤った認識を相手に与えたり、またその内容をもとに何らかの判断や意思決定がなされたりした場合に、損害を生じさせてしまう危険があります。ビジネスにおいても、研究においても、倫理上の観点からこのような操作はあってはならないことであると理解しましょう。

3. 具体的なビジュアライゼーションの工夫

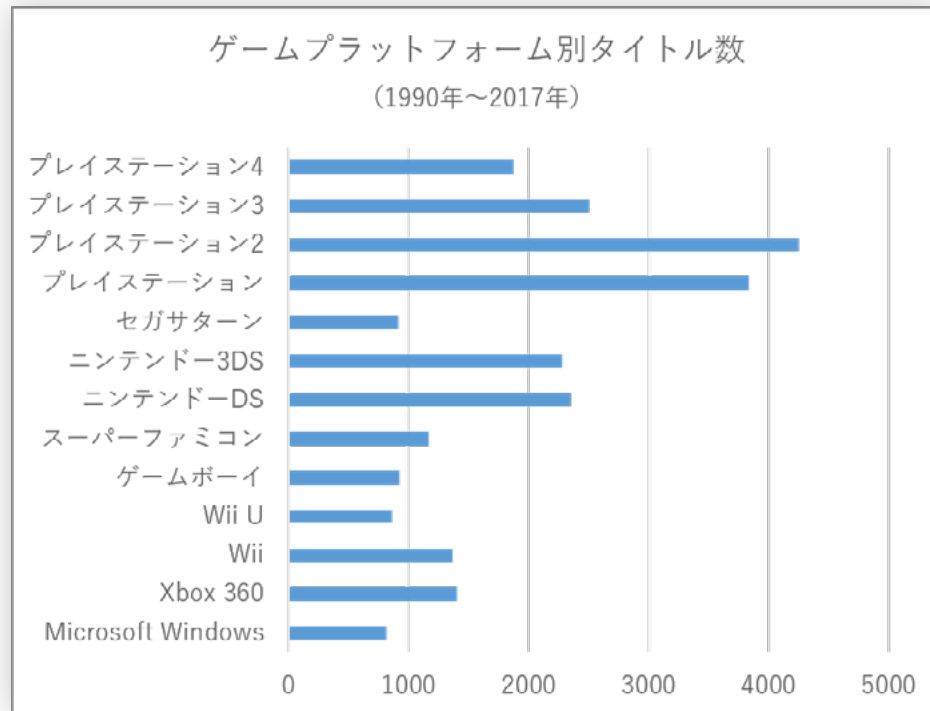
3.01 ここでは、サンプルデータセットを利用して得られた分析結果をまとめたグラフを用いて、具体的なビジュアライズの工夫について検討してみます。次のこのグラフを改善するとしたらどんなポイントがあるか、考えてみましょう。



3. 具体的なビジュアライゼーションの工夫

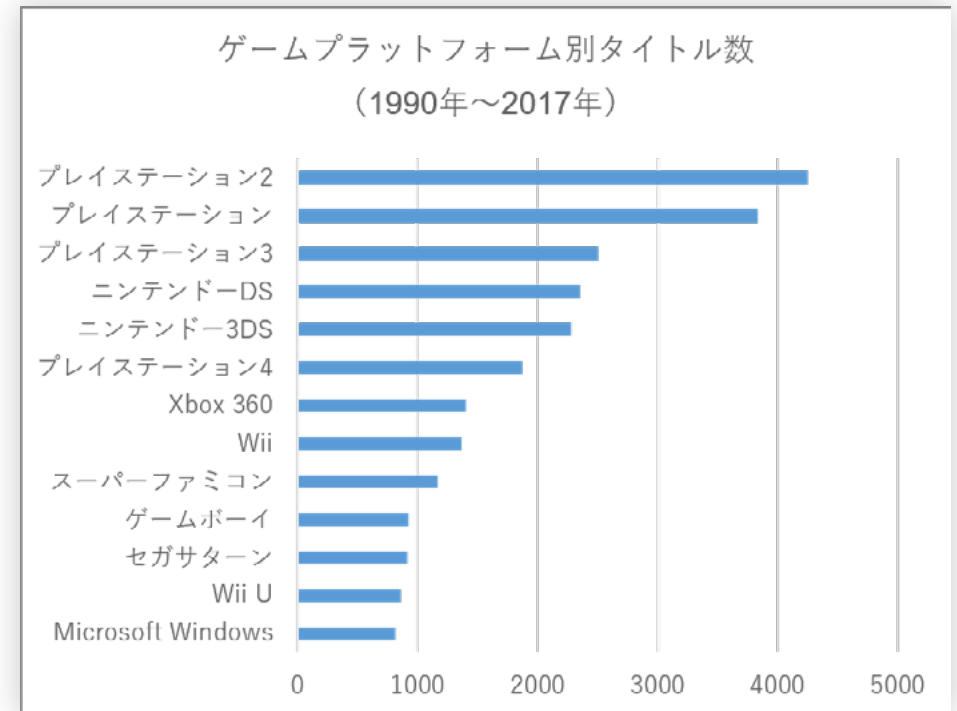
3.02 改善点の例

① 分析のタイトルをつける>>ゲームプラットフォーム別タイトル数(1990年～)



→ 何を示したグラフなのかを明確に示すことは、相手が結果を理解する第1歩になります。

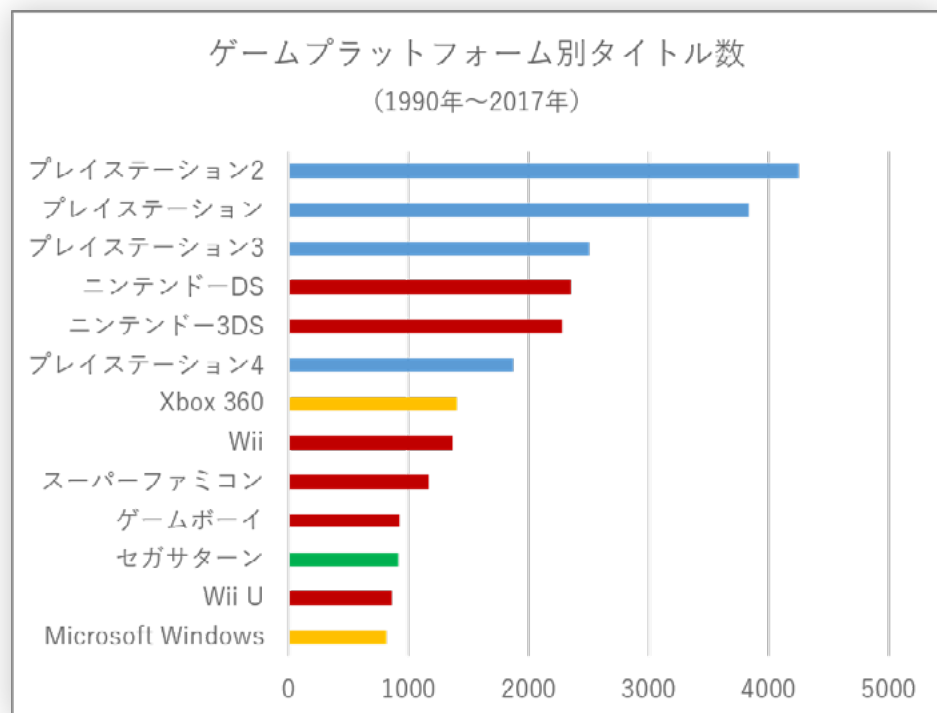
② タイトル数が多い順に並べてみる



→ すっきり見えるようになりました。しかし、これだけでは何を伝えようとしているのか、分析の観点があいまいです。

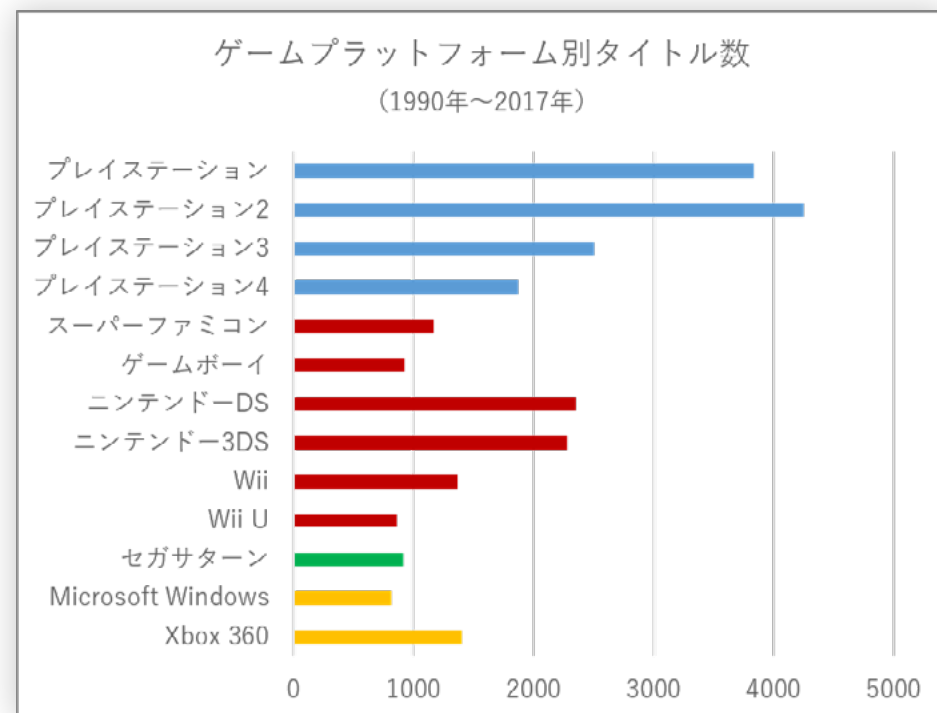
3. 具体的なビジュアライゼーションの工夫

③ プラットフォームのメーカー別にグラフの軸を色分けしてみる



→ これでどのメーカーのゲームプラットフォームからより多くのタイトルが世に出ているのかを表現しやすくなります。でもまだ煩雑にみえます。

④ グラフの並び順もメーカー別、発売順に工夫してみます。



→ これでメーカー別のタイトル数を比較することができるようになりました。

このように、同じデータのビジュアライズでも、分析の目的や観点にあわせて、さまざまな表現の工夫によって、より伝わりやすくすることができるようになります。

3. 具体的なビジュアライゼーションの工夫

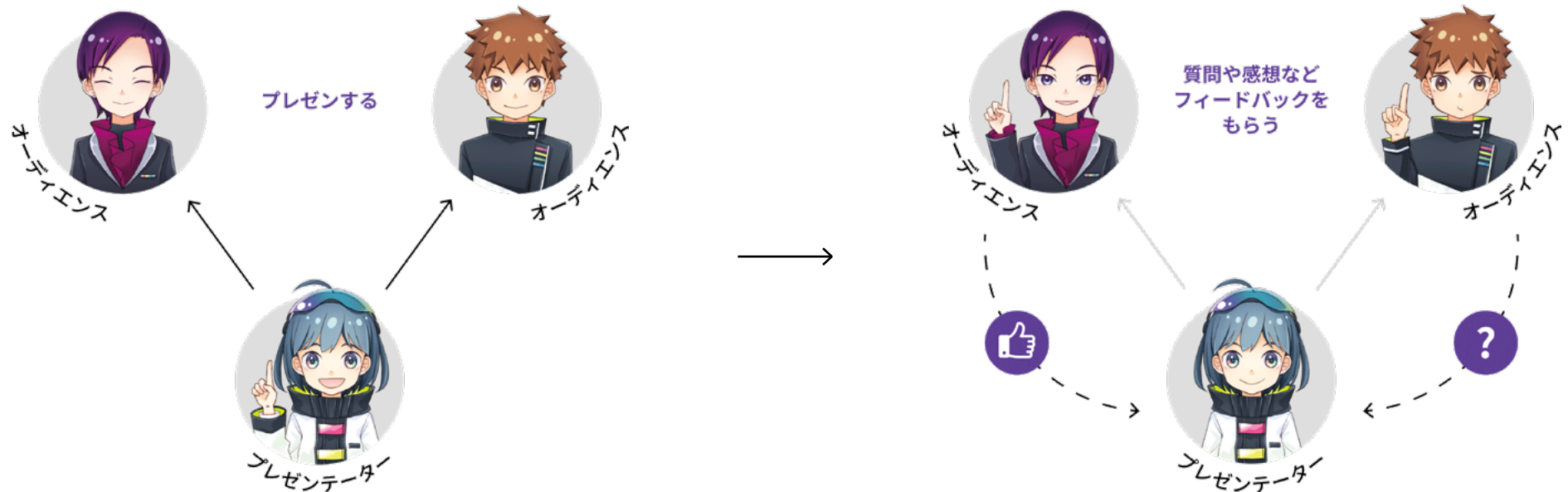
3.03 コマ2・3で作成したグラフや、自分がこれまでに行った分析結果について、わかりやすく魅力的に伝えるための自分なりの工夫や改善をしてみましょう。なお、その際には、以下の2点についてチェックするようにしましょう。

- プレゼンテーションの目的、仮説を検証するためのグラフとして結果から得られた発見を正しく表現するものになっているでしょうか？
- データの見え方は事実を伝えていますか、自分にとって都合のよいようにデータを見せていませんか？

4. プレゼンテーションの実施

4.01 1でまとめたテキスト(言葉)と、ビジュアライズしたグラフを、スライド形式の資料などにまとめプレゼンテーションを完成させましょう。そのうえで、実際にデータビジュアライズが意図したとおり相手に伝わるかどうかを評価するために、模擬プレゼンテーションを実施しましょう。授業やグループワークで行う場合には、3人程度のグループをつくりましょう。個人学習の場合には、周りの友人や家族などに協力してもらいましょう。

- ① 自分が作成したスライドを用いてプレゼンテーションします
- ② 1人はプレゼンター、残る2人はオーディエンス役となる
- ③ プレゼンテーションに対する質疑応答を行う
- ④ 3人のプレゼンが終了したらお互いのプレゼンテーションについて評価する



5. データビジュアライゼーションでできること(まとめ)

ここまでデータ分析とデータビジュアライゼーションについて学んできました。メディア芸術データベースには、今回皆さんが使ったサンプルデータだけでなく、豊富なデータがそろっています。ぜひ、データ分析とビジュアライズのスキルを磨いて、自分がみつけた新しい発見を通じて、メディア芸術の世界、そして社会をもっと豊かなものにしていきましょう!



※50音順

教材企画・開発・監修

岡田龍太郎(武蔵野大学データサイエンス学部データサイエンス学科助教/国際大学GLOCOM 研究員(併任))
菊地映輝(国際大学GLOCOM 研究員/講師)
小林奈穂(国際大学GLOCOM 主幹研究員/研究プロデューサー)
榊原直樹(清泉女学院大学人間学部 文化学科 専任講師)
波木井丈瑠(国際大学GLOCOM リサーチアシスタント)
kakeami(第2回メディア芸術データベース活用コンテストビジュアルクリエイション部門最優秀事例作品制作者)
渡辺智暁(国際大学GLOCOM 主幹研究員/教授)

テキストデザイン

本宮亜里・渡邊英弘(Makaira Art&Design/MAD)

教材動画制作

大谷友次郎・小島岳・高橋直秀(株式会社アウスタ)

声の出演

アム/清水梨々香・ソウキ/橘諒・リョウ先生/名取舞

イラスト

木野陽

制作協力

ミハラテツヤ

進行管理

桜井陽子・森由紀(大日本印刷株式会社)

※本教材に使用したデータセット項目等については、第2回メディア芸術データベース活用コンテストビジュアルクリエイション部門最優秀事例である「マンガと学ぶデータビジュアライゼーション」を参照して開発されています。

詳しくはこちらをご参照ください。

「マンガと学ぶデータビジュアライゼーション」(kakeami氏による作品サイト)

<https://kakeami.github.io/viz-madb/>

政府標準利用規約(第2.0版)、(後継バージョンがある場合は、それも可)に基づき利用を許諾する。



メディア芸術
データベース



Agency for Cultural Affairs, Government of Japan