#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。

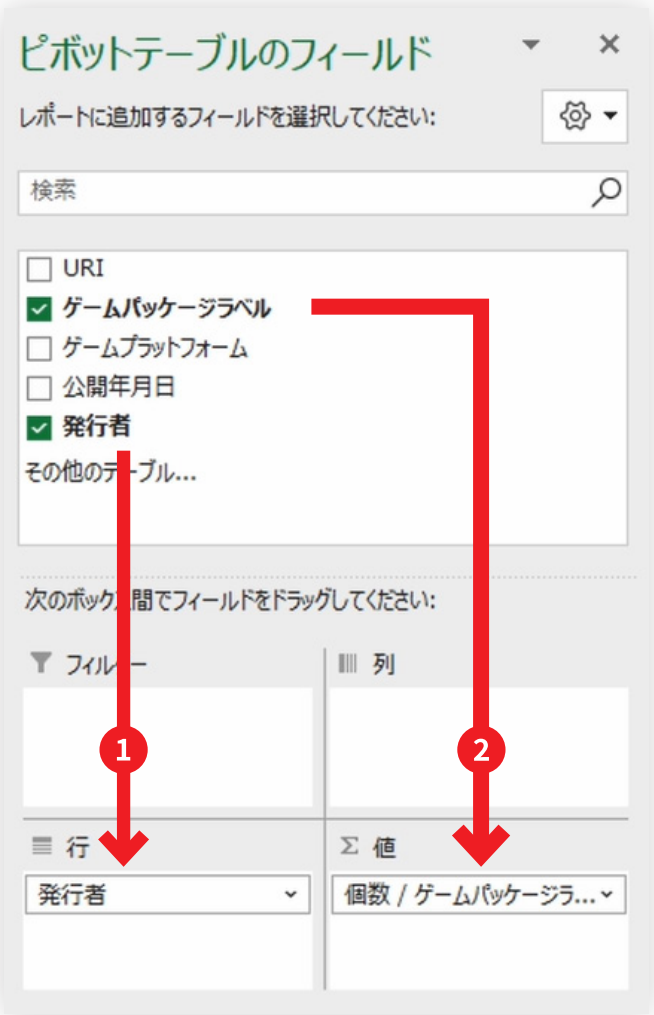
3.02 サンプルデータ<ゲームパッケージ.xlsx>をExcelで開きます。

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.artmuseums.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.artmuseums.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.artmuseums.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	Nintendo Switch	2017/12/28	ハムスター
5	https://mediaarts-db.artmuseums.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.artmuseums.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.artmuseums.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.artmuseums.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.artmuseums.go.jp/id/M740801	Moorhuhn Knights & Castles モーアフーン ナイツ アンド キ	Nintendo Switch	2017/12/28	Young Fun Studio
10	https://mediaarts-db.artmuseums.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.artmuseums.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.artmuseums.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.artmuseums.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.artmuseums.go.jp/id/M746591	カイジ〜絶望の鉄骨渡り〜 for Nintendo Switch ダウンロード	Nintendo Switch	2017/12/28	ソリッドスフィア株式会社
15	https://mediaarts-db.artmuseums.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.artmuseums.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.artmuseums.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.artmuseums.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.artmuseums.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.artmuseums.go.jp/id/M737863	Call of Duty : WWII + Destiny 2 デジタルデラックスバンドル	Xbox One	2017/12/22	Activision

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、❶「発行者」を「行」に、❷「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。



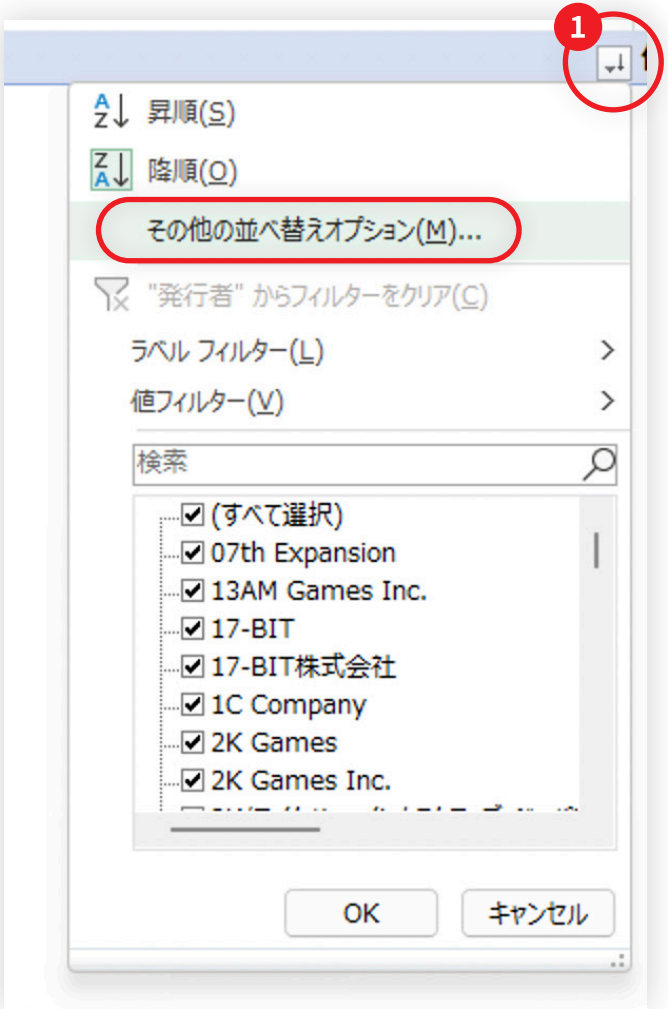
3. 分析例A：ゲーム会社別の発売数

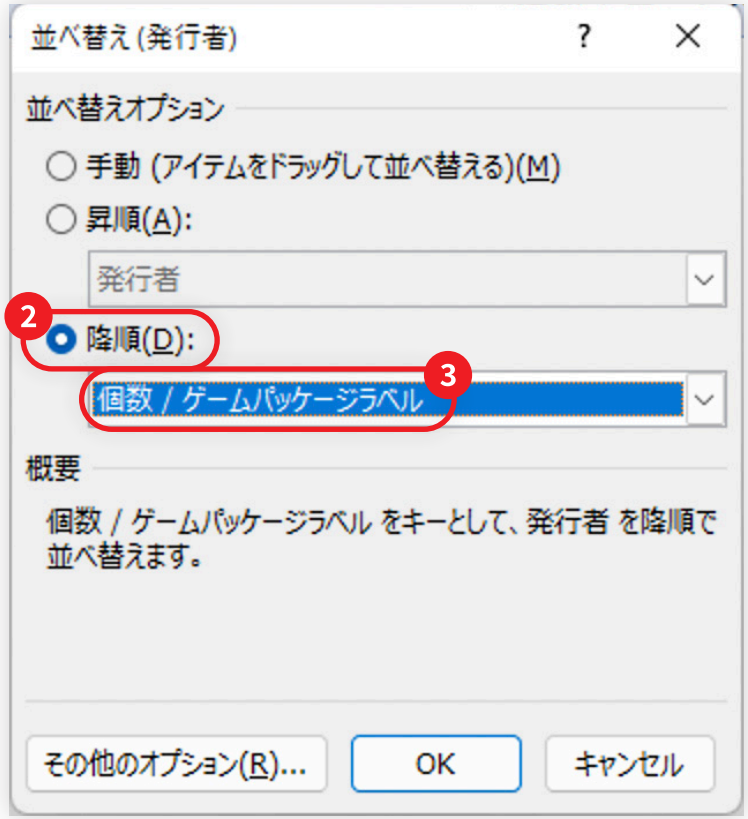
3.05 これでゲーム会社ごとにゲームの発売タイトルを集計することができました。

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

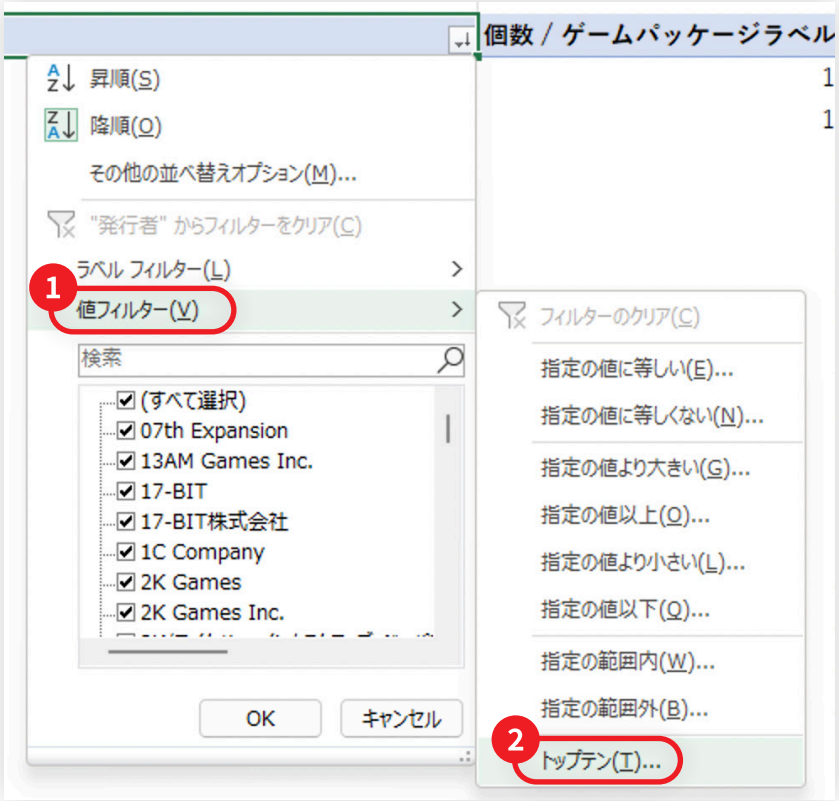
3.06 ❶フィルタのボタンを押して、❷並び替えの順序に「降順」を選び、❸並べ替えの対象に「個数 / ゲームパッケージ ラベル」を選びます。

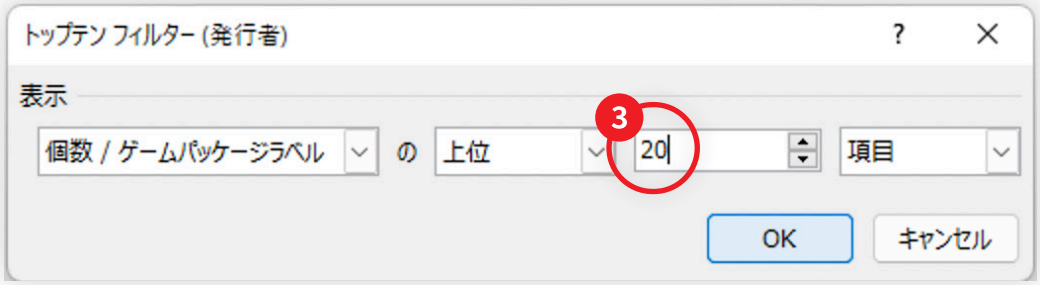




3. 分析例A：ゲーム会社別の発売数

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





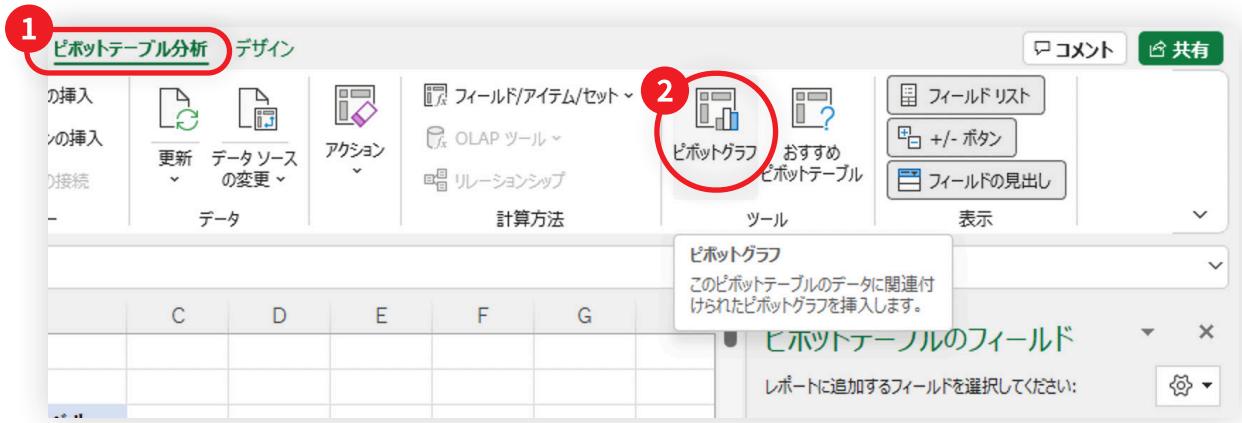
3. 分析例A：ゲーム会社別の発売数

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**

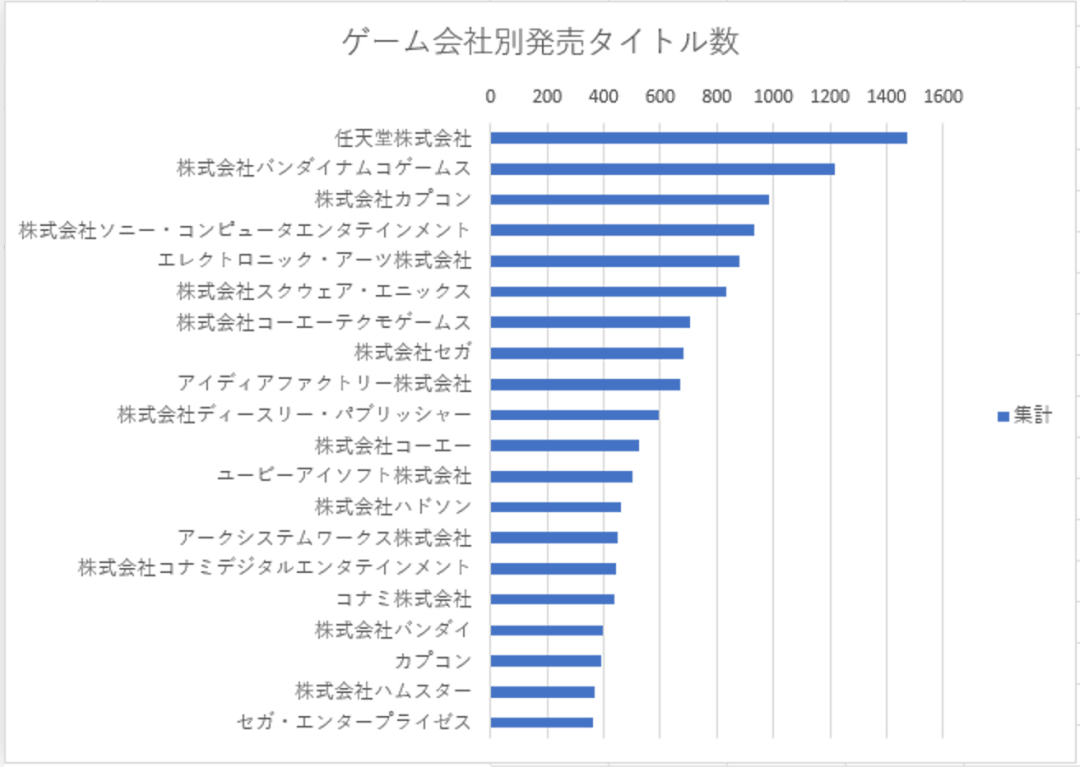
3. 分析例A：ゲーム会社別の発売数

3.09 これをグラフにします。❶リボンタブの「ピボットテーブル分析」の❷「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。

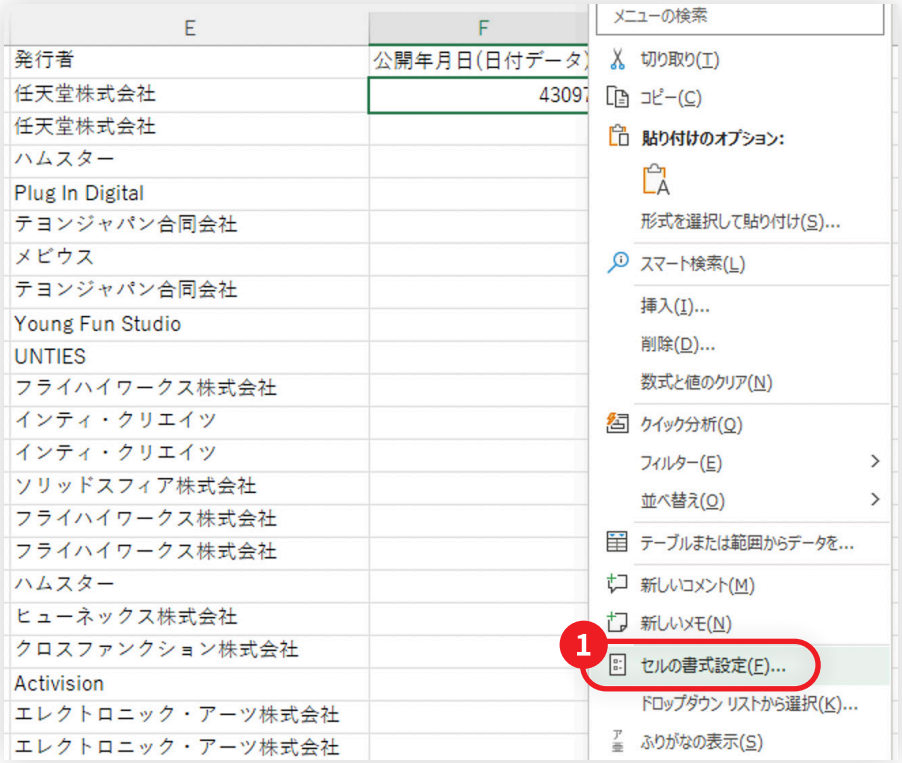
D	E	F	G
公開年月日	発行者	公開年月日(日付データ)	
2017-12-28	任天堂株式会社		
2017-12-28	任天堂株式会社		
2017-12-28	ハムスター		
2017-12-28	Plug In Digital		
2017-12-28	テヨンジャパン合同会社		

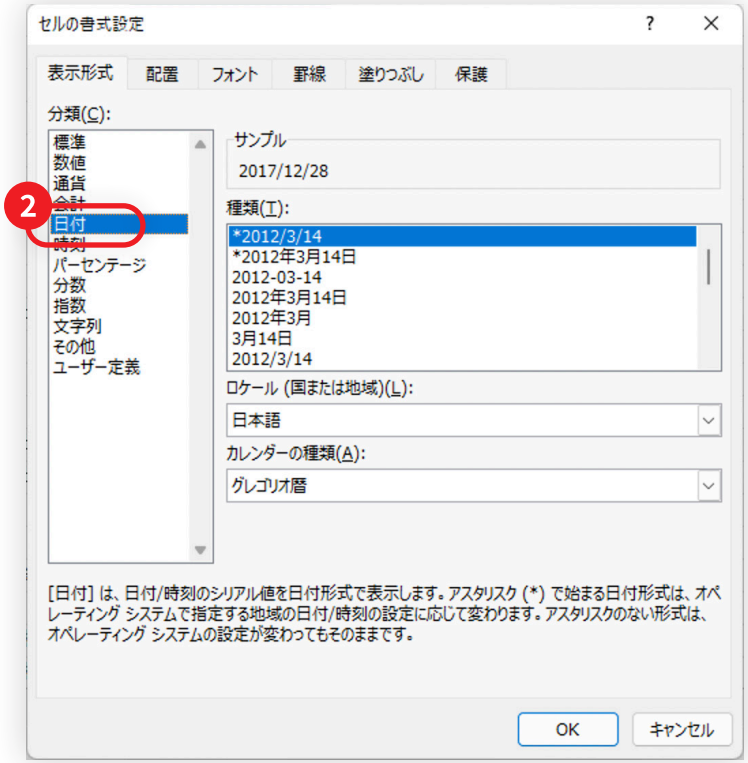
4.04 Excelの関数を使って文字列で書かれた日付を、日付専用のデータに変換します。「DATEVALUE」関数を使います。

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。❶「書式」を❷「日付」に変更してみましょう。





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	パック・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	パンプレスト	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、ひとまずそのままで進めましょう。

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	

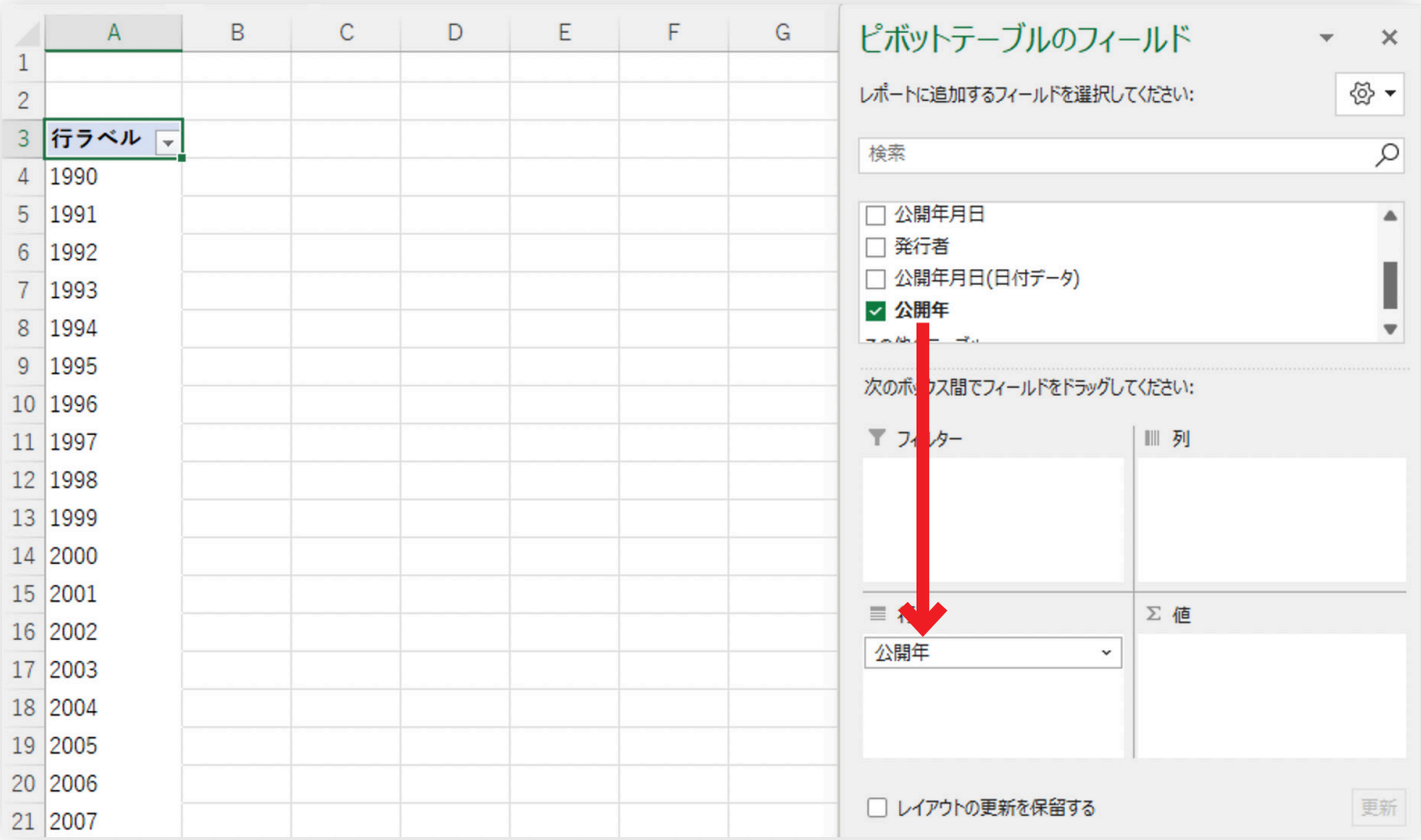
4.12 更に全体にコピーします。これでゲームの発売日を年別に集計する準備が整いました。

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

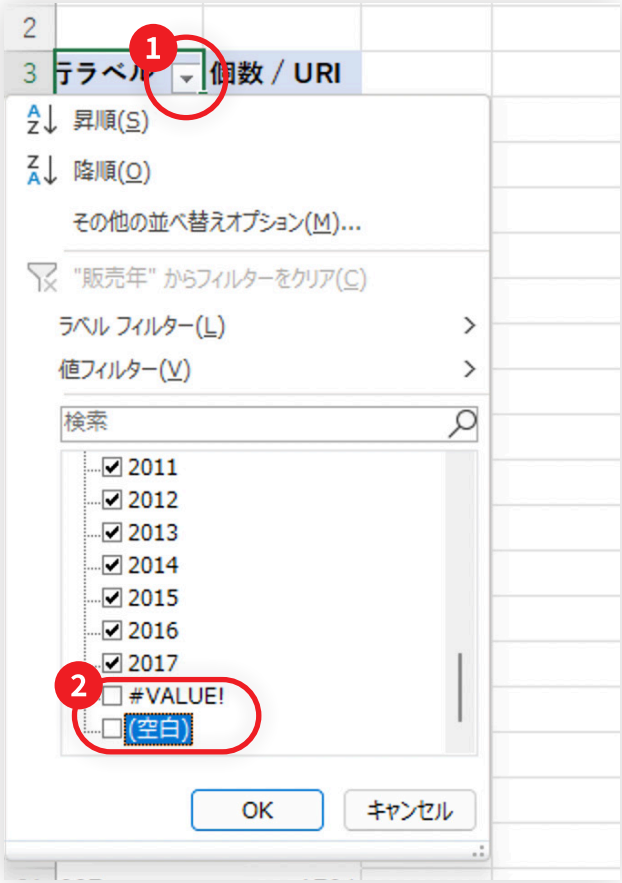
4.13 シート全体を選択してピポットテーブルを挿入します。そうすると前回とは違って、追加した列の分も使えるようになっています。

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。



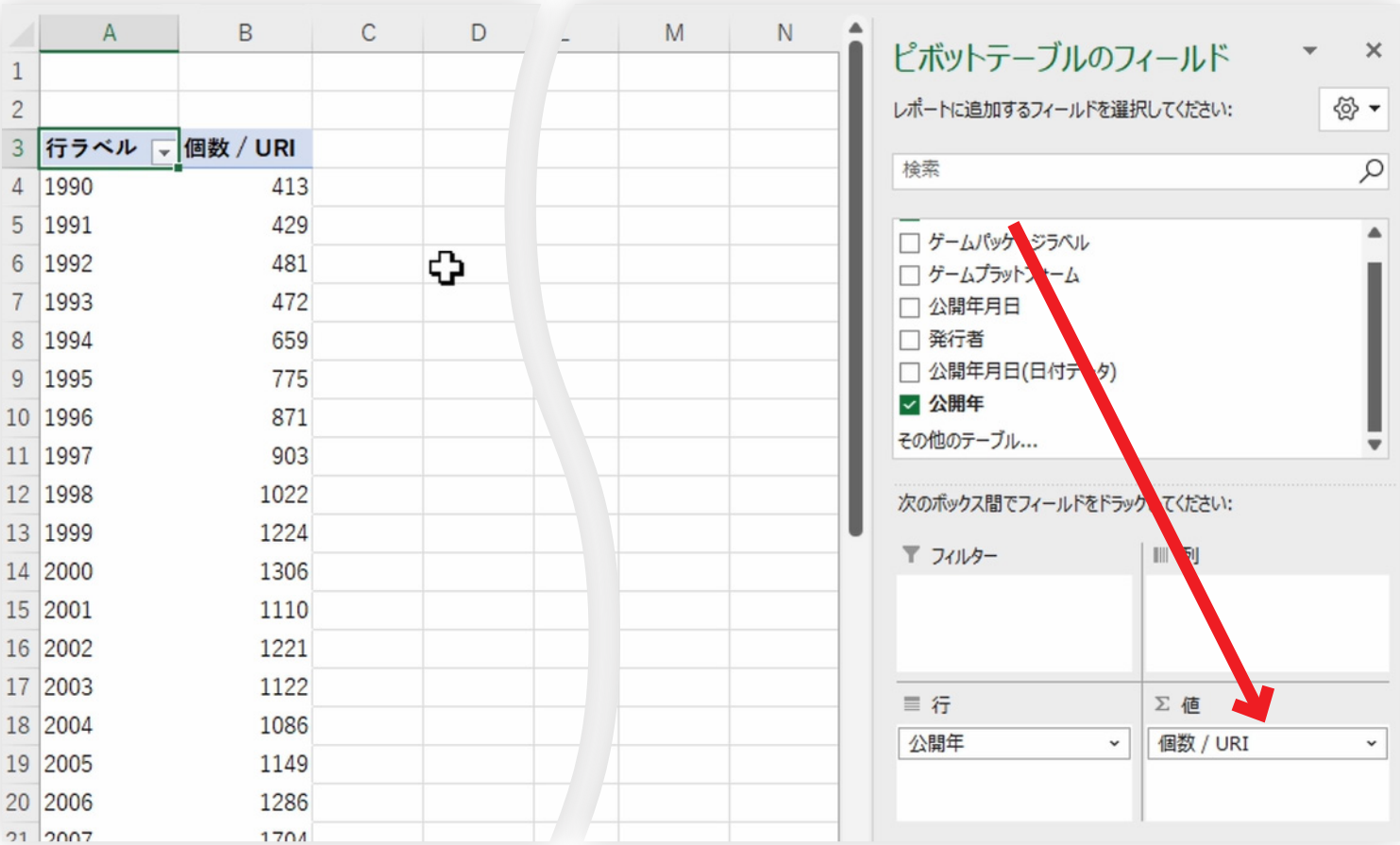
4. 分析例B：年別発売タイトル数

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。❶「行ラベル」のところにあるフィルタから、❷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

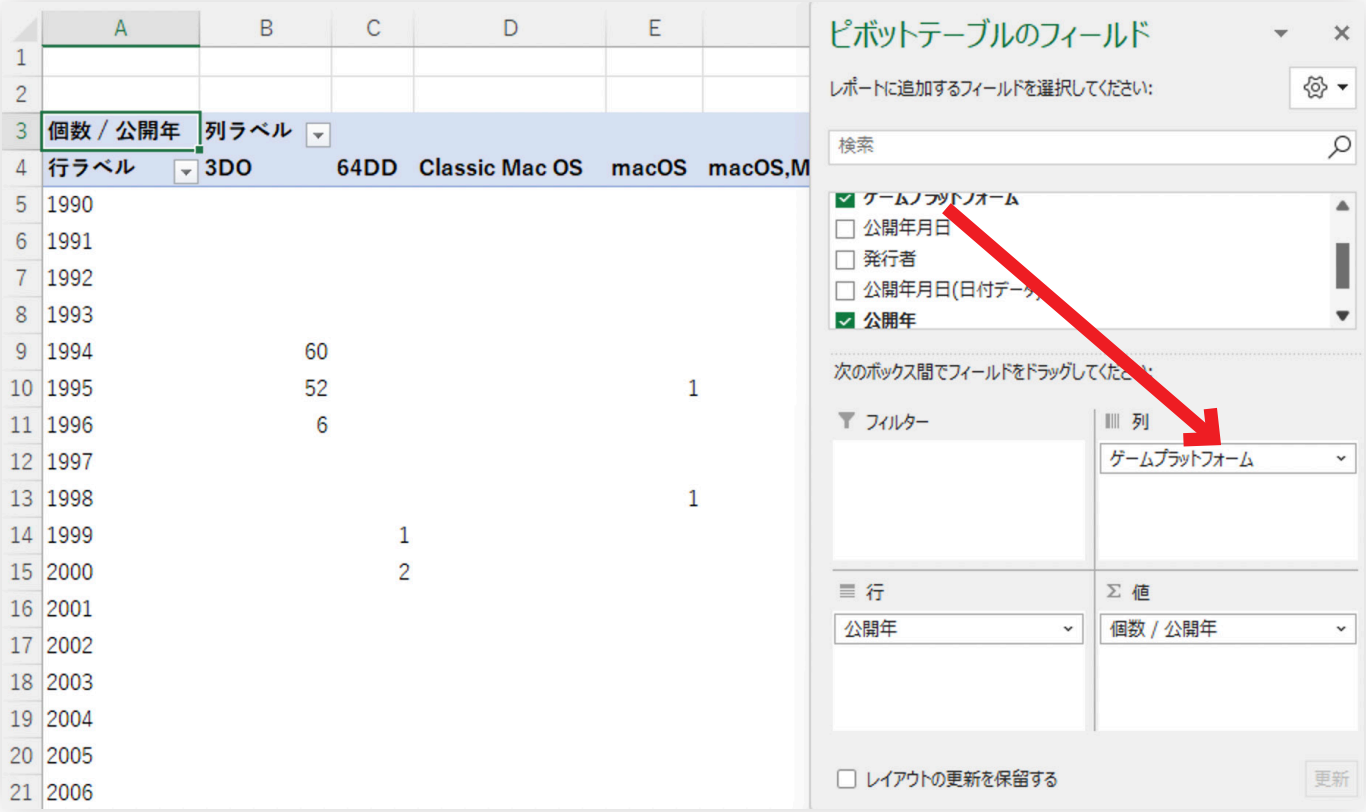
4.16 ここで、「URI」(「ゲームパッケージラベル」でもよい)を「値」に入れると個数をカウントできるので、まず年別の発売 タイトル 数を集計することが出来ます。



4. 分析例B：年別発売タイトル数

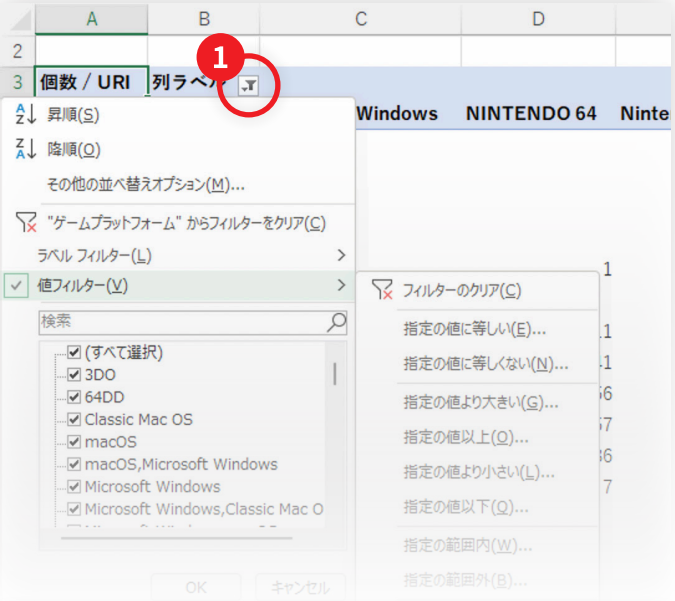
4.17 この結果だけ見ると、ゲームの発売数は年々増えているのが分かります。

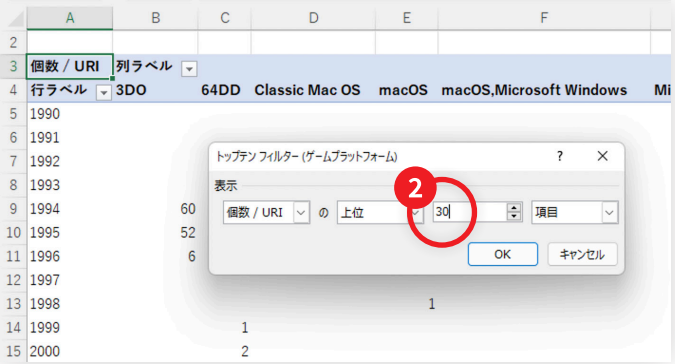
4.18 そしてピボットテーブルの新たな使い方ですが、こんどは「ゲームプラットフォーム」を「列」にドラッグアンドドロップします。すると、ゲームのプラットフォームごとに発売タイトル数を分けて集計することが出来ます。

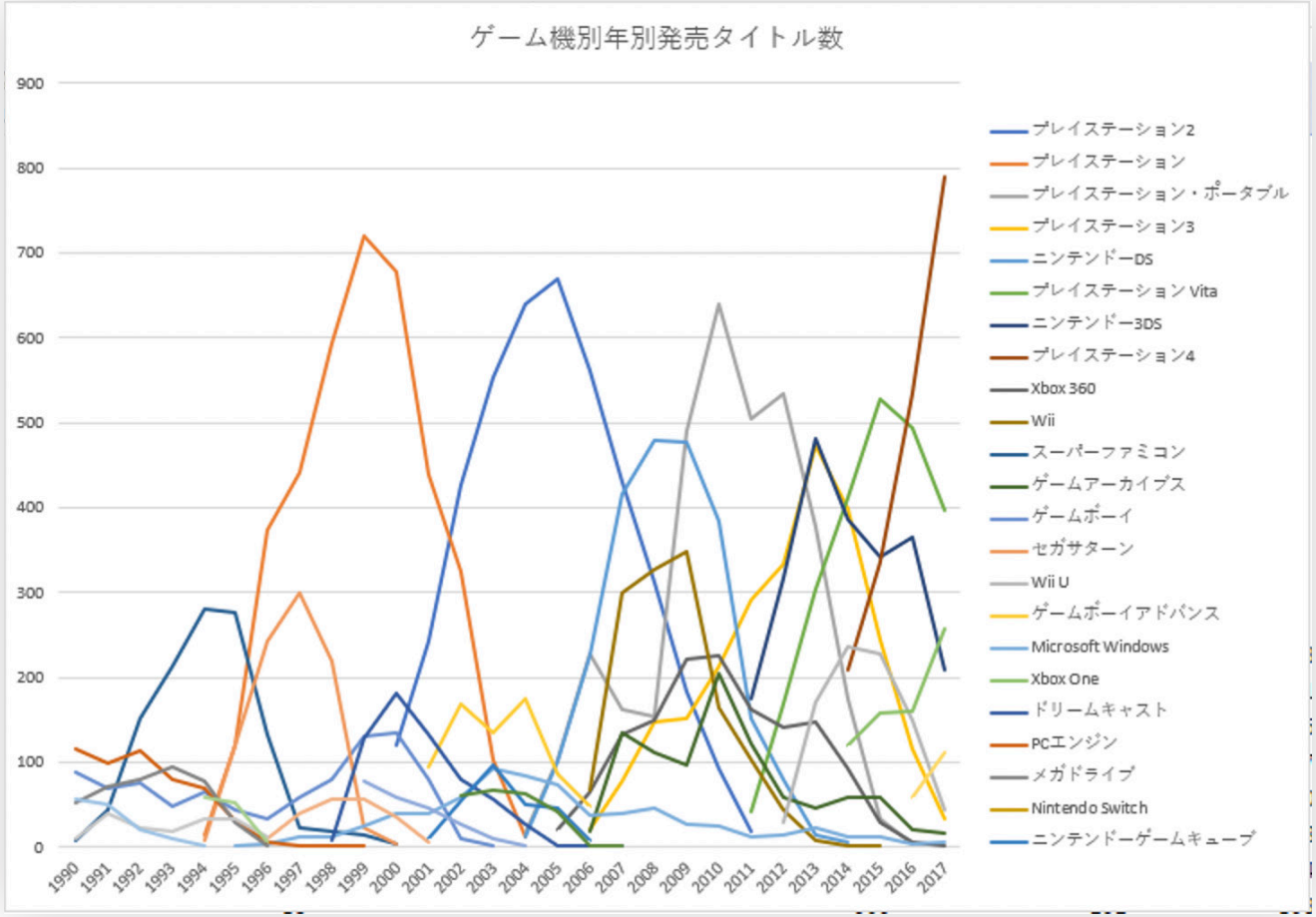


4. 分析例B：年別発売タイトル数

4.19 集計するゲームプラットフォームは、❶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❷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そうなのは「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。

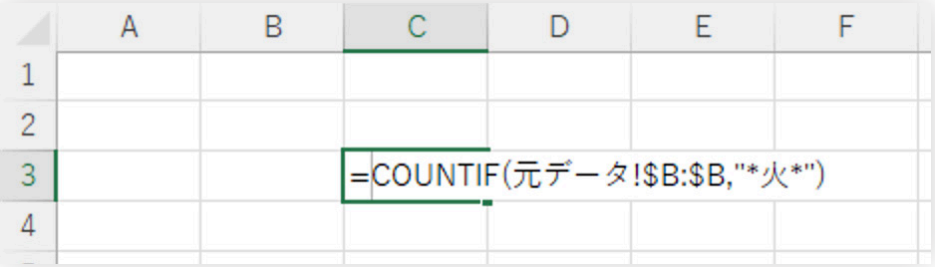
5.02 ゲームやファンタジー小説でいわゆる「属性」を表す言葉として使われている「火」「水」「風」「土」という言葉のうち、ゲームのタイトルに一番登場するのはどれでしょうか？

- 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。

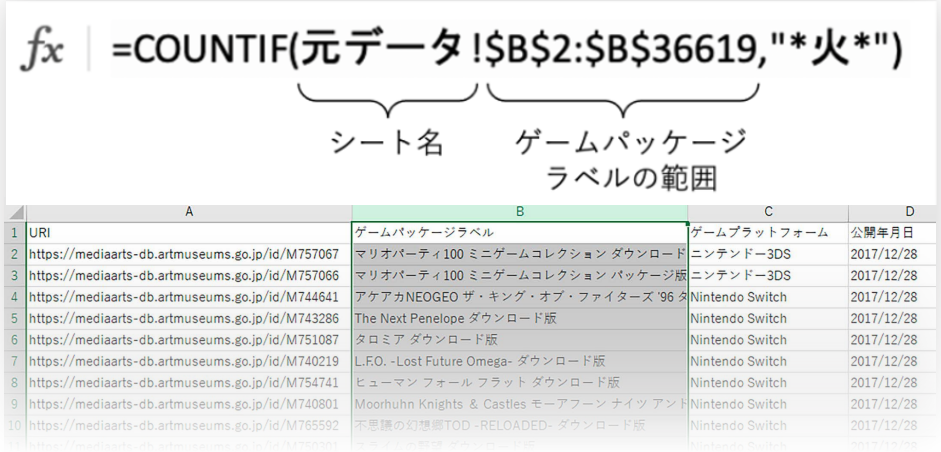
5. 分析例C：テキスト分析

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。

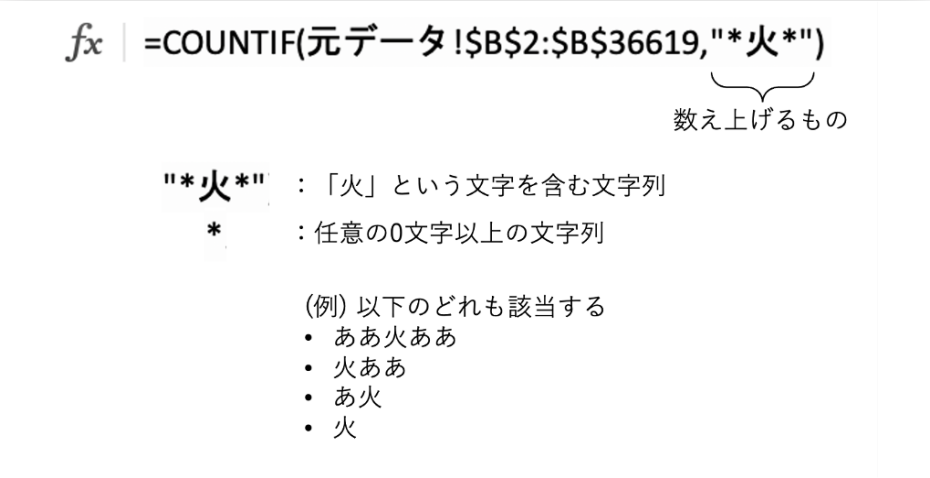


③ COUNTIF関数の1つ目の引数は、データの場所を指しています。ここでは、後々のことを考えて絶対参照にしています。

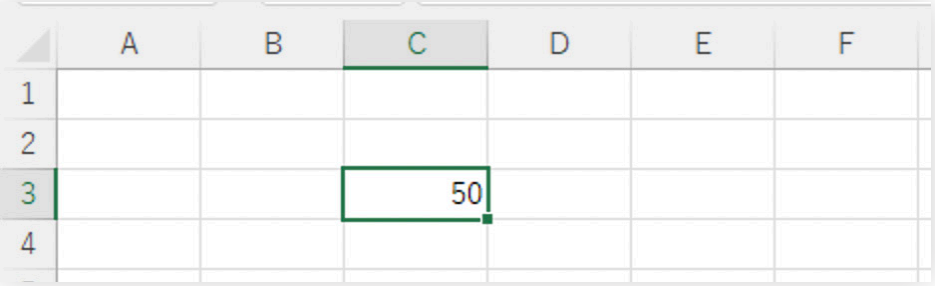


	A	B	C	D
1	URI	ゲームパッケージラベル	ゲームプラットフォーム	公開年月日
2	https://mediaarts-db.artmuseums.go.jp/id/M757067	マリオパーティ100 ミニゲームコレクション ダウンロード	ニンテンドー3DS	2017/12/28
3	https://mediaarts-db.artmuseums.go.jp/id/M757066	マリオパーティ100 ミニゲームコレクション パッケージ版	ニンテンドー3DS	2017/12/28
4	https://mediaarts-db.artmuseums.go.jp/id/M744641	アケアカNEOGEO ザ・キング・オブ・ファイターズ '96 ダ	Nintendo Switch	2017/12/28
5	https://mediaarts-db.artmuseums.go.jp/id/M743286	The Next Penelope ダウンロード版	Nintendo Switch	2017/12/28
6	https://mediaarts-db.artmuseums.go.jp/id/M751087	タロミア ダウンロード版	Nintendo Switch	2017/12/28
7	https://mediaarts-db.artmuseums.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.artmuseums.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
9	https://mediaarts-db.artmuseums.go.jp/id/M740801	Moorhuhn Knights & Castles モーアフーン ナイツ アンド	Nintendo Switch	2017/12/28
10	https://mediaarts-db.artmuseums.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

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。



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。



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。

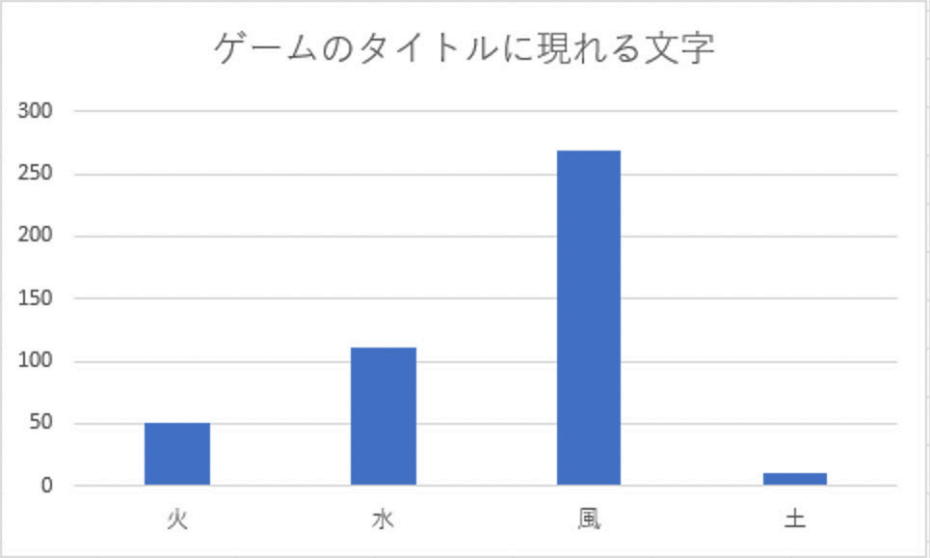
① 文字列は「&(アンド)」で結合することが出来ます。参照先のアドレスと*(アスタリスク)を文字列で結合します。以下のようにします。

	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

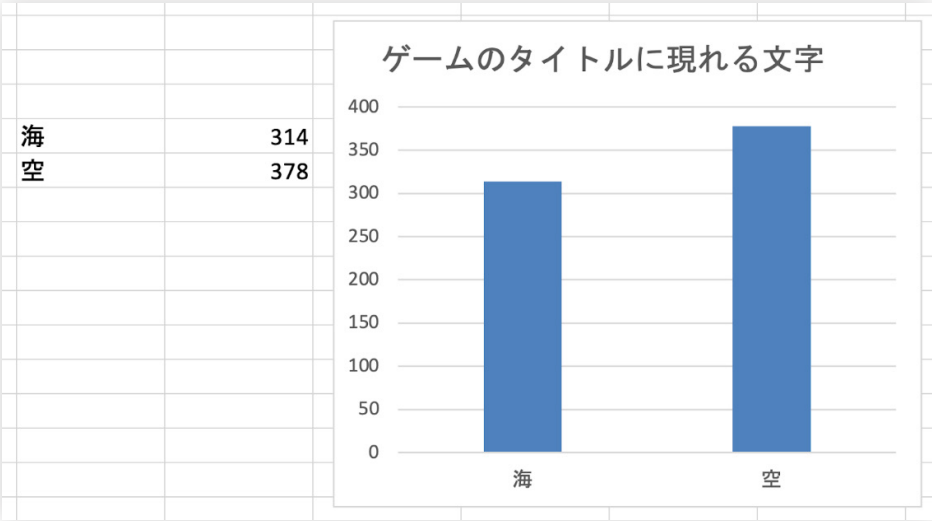
③ 一番多いのは風でした。土は予想通りすごく少ないですね。グラフにもしてみましょう。棒グラフがいいですね。



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。

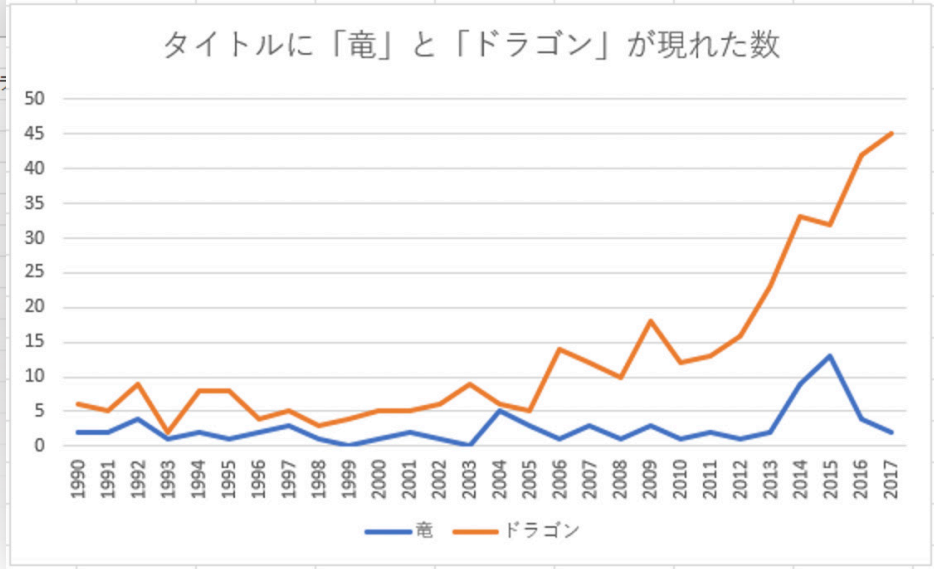
	A	B	C	D	E
1					
2			竜	ドラゴン	
3		1990			
4		1991			
5		1992			
6		1993			
7		1994			
8		1995			
9		1996			
10		1997			
11		1998			
12		1999			
13		2000			
14		2001			
15		2002			
16		2003			
17		2004			
18		2005			
19		2006			
20		2007			
21		2008			

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	ドラ
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の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

	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。❶「竜(割合)」と「ドラゴン(割合)」という列を追加します。そして、❷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		

❶

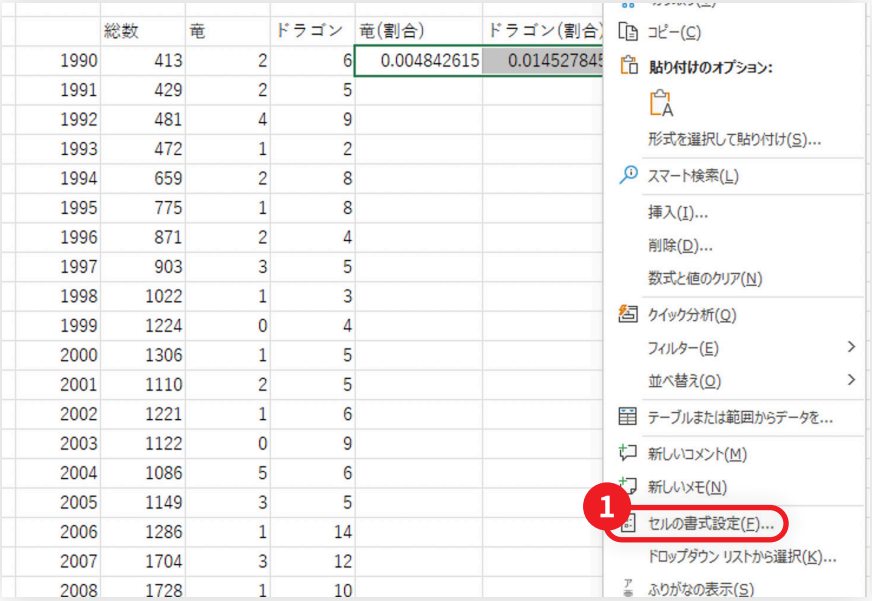
	総数	竜	ドラゴン	竜(割合)	ドラゴン(割合)
1990	413	2	6	=D3/C3	
1991	429	2	5		
1992	481	4	9		
1993	472	1	2		

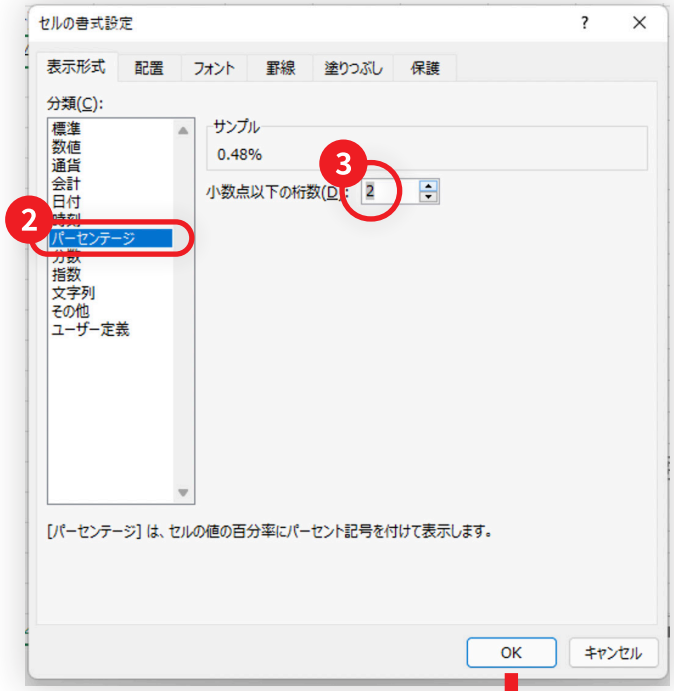
❷

	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。❶「セルの書式設定」を選んで、❷「パーセンテージ」を選び、❸「小数点以下の桁数」を2とします。



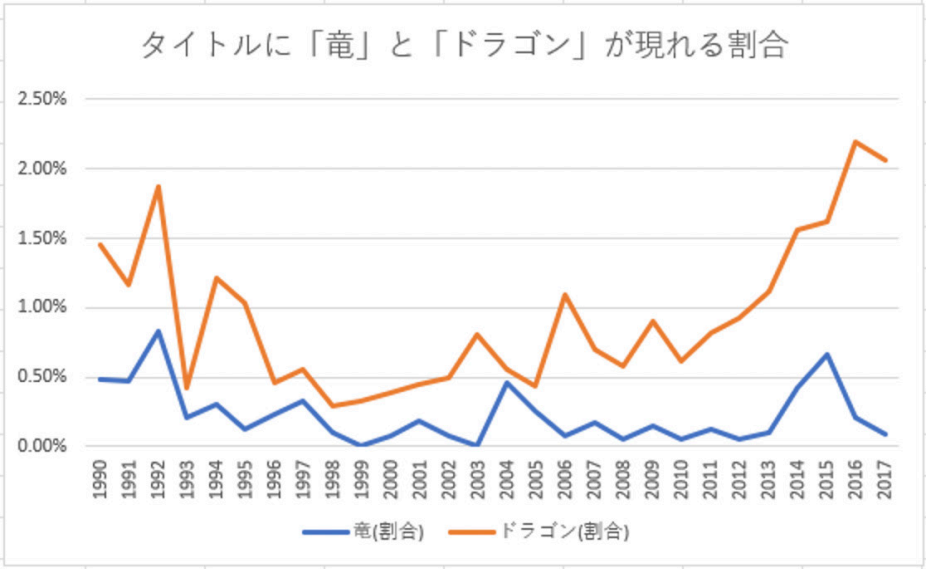


	総数	竜	ドラゴン	竜(割合)	ドラゴン(割合)
1990	413	2	6	0.48%	1.45%
1991	429	2	5		
1992	481	4	9		
1993	472	1	2		

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。

	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ませんね。

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。