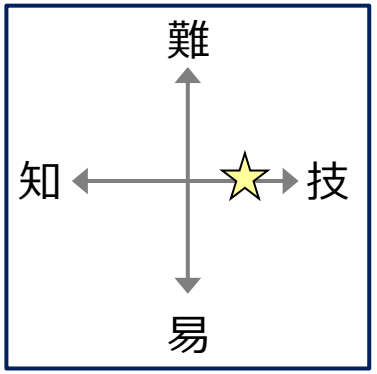
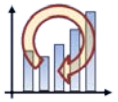
総務省　ICTスキル総合習得教材



［コース3］データ分析



3-2：データのクレンジングと可視化

http://www.soumu.go.jp/ict_skill/pdf/ict_skill_3_2.pdf

	1	2	3	4	5
［コース1］データ収集					
［コース2］データ蓄積					
［コース3］データ分析		●			
［コース4］データ利活用					

本講座の学習内容［3-2：データのクレンジングと可視化］

【講座概要】

- Excelのフィルター、ステータスバーを利用したデータチェックの方法を紹介します。
- Excel関数を利用した基本的なデータクレンジングの方法を説明します。
- Excel関数を利用したデータの整理、データセットの結合、データ集計の方法を示します。
- Excelのグラフ作成による可視化とグラフの使い分けを紹介します。

【講座構成】

実習

- [1] Excelにおけるデータチェック
- [2] Excel関数によるデータクレンジング
- [3] Excelにおける分析用データ確認と抽出
- [4] Excelにおけるデータセットの結合と集計
- [5] Excelにおけるデータの可視化

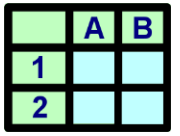
【学習のゴール】

- ✓ Excelのフィルター、ステータスバーを利用して、表記揺れ、異常値のチェックができる。
- ✓ Excel関数を利用して基本的なデータクレンジングができる。
- ✓ Excel関数を利用してデータ整理、集計ができる。
- ✓ Excelのグラフ作成で基本的な可視化ができる。

Excelによるデータクレンジング

◆Microsoft Excelを利用したデータクレンジングの基本操作を説明します。

- **データクレンジング**とは、**分析の障害となる異常値、重複データ等を取り除き、分析しやすい状態にすること**です。
  - 講座3-1でも紹介したように、データクレンジングにかかる時間が本格的な分析作業以上の時間となることがたびたびあります。データクレンジングを効率的に行うことはデータ分析、活用において重要です。
  - データクレンジングの必要はないデータであっても、本格的な分析前のデータ整理をはじめとして、より広義の「データの前処理」が必要となります。
- 本講座では、構造化（表形式）データのクレンジングの方法を紹介します。
  - 半構造化データ・非構造化データを構造化データへ変換するデータクレンジング・整理もあります。



- 一般に普及しているMicrosoft Excelを用いて、プログラミング不要で行えるデータクレンジングを紹介します。
  - Excelによるプログラミング不要のデータクレンジングは、技術面とコンピュータ環境面の制約が少ないため、組織・チーム内で依頼することが容易です。
  - 操作を例示したスクリーンキャプチャはMicrosoft Excel 2010で示していますが、Excel 2010以降であれば概ね同様の操作が実行できます。

【Excelにおけるデータクレンジングについて】

◆ Excelによるデータクレンジングには、ソフトウェアが広く普及しており、視覚に基づく直感的な操作がしやすいという長所がある一方で、作業プロセスとなるプログラムコードや作業記録となるログが自動で残らない短所があります。

この講座では、作業記録が残りやすいExcel関数を使ったデータクレンジングを説明します。

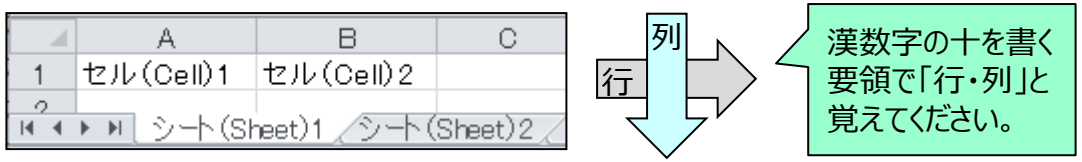
- データクレンジングの記録を残しておくことで、再度同じクレンジングを行う場合、途中からクレンジングの方法を変える場合に便利です。
- データ分析においては、他の人が行っても同じ分析結果を導出できる「客観性」や「再現可能性」が重要です。自分自身が理解するのみならず、他の人に説明できるように、他の人でも同じデータクレンジングが行えるように記録することが必要となります。
- 講座4-3で紹介するようにプログラミング言語を使ったデータ分析は意識せずとも、プログラムコードやログを残すことができます。

■ この講座では「【実習用データ】ICT3-2_データクレンジングと可視化.xlsx」を用いて実習を行います。
利用するExcelのシート番号は、各スライド右上の〔〕内に示します。 http://www.soumu.go.jp/ict_skill/dc/ict_3_2data.zip

通し番号の挿入

◆Excelにおけるデータクレンジングおよびデータ整理には通し番号を作っておくと便利です。

- Excelでは電子ファイル全体を**ブック**、ブック内の区切られたページを**シート**、各シート内の入力欄の枠を**セル**といいます。
  - Excelでは「A1」や「C3」と「列のアルファベット・行の番号」の組み合わせで表現されるセルの位置を「セルの番地（cell address）」や「セル番号」といいます。
- Excelでは水平線側を**行**、垂直線側を**列**と呼びます。
  - 講座2-1でも紹介したリレーショナルデータベースと呼ばれる構造化データの格納に適したデータベースにおいても、水平線を「行」、垂直線を「列」といいます。



- この講座では事例として、シート〔1〕にある文房具店の売上を示す構造化データをExcelでクレンジングするケースを考えます。
  - 構造化データ全体を「データセット」、行毎の個別の売上情報を「データレコード」と呼びます。
- 実習用データのシート〔1〕のように、元のデータセットに通し番号がない場合は、一番左に**［通し番号］**の列を作っておきます。
  - ［通し番号］は、行番号のIDとしても利用でき、データセットの全レコード（行）数を確認する場合にも、ソート（並び替え）を元に戻す場合においても、便利です。

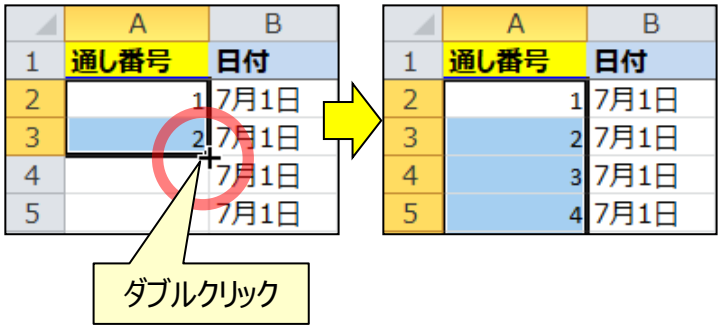
データセットの左端に空白の列を作り、1行目に「1」、2行目に「2」を入力して、入力した二つのセルを選択した状態で、「2」の右下の黒い四角をダブルクリックすると、最終行まで通し番号がつきます。

- 「1」のセルだけを選択した状態で、右下の黒い四角をダブルクリックしてしまうと、最終行まで「1」が並びますので、「2」まで含めて選択してからダブルクリックしてください。

通し番号を挿入したデータセット（10列目まで）

A	B	C	D	E	F
通し番号	日付	曜日	単価	数量	数量
1	7月1日	水	ボールペン黒	100	1
2	7月1日	水	鉛筆	80	5
3	7月1日	水	ボールペン赤	100	1
4	7月1日	水	ボールペン赤	100	2
5	7月1日	水	ノート	150	2
6	7月1日	水	はさみ	400	1
7	7月1日	水	はさみ	400	2
8	7月1日	水	はさみ	400	2
9	7月1日	水	はさみ	400	2
10	7月1日	水	ボールペン赤	100	1

Excelにおける通し番号のつけ方



欠損値・異常値のチェック

◆Excelにおいて、数値の欠損値・異常値の確認をする際は、「ステータスバー」が便利です。

- Excel画面下側の表示倍率を表す枠の部分を**ステータスバー**といい、ステータスバーを右クリックすることで、その表示内容を選択でき、欠損値・異常値の確認に利用できます。
  - データレコードにおいて一部の情報が利用できないものを欠損値または欠測値といいます。欠損値はセルが空白となるケースもあれば、「N/A」と文字列が記入されるケース、数値以外の文字列が入力できなかった場合では「-1」「9999」などの異常値が記入されるケースがあります。

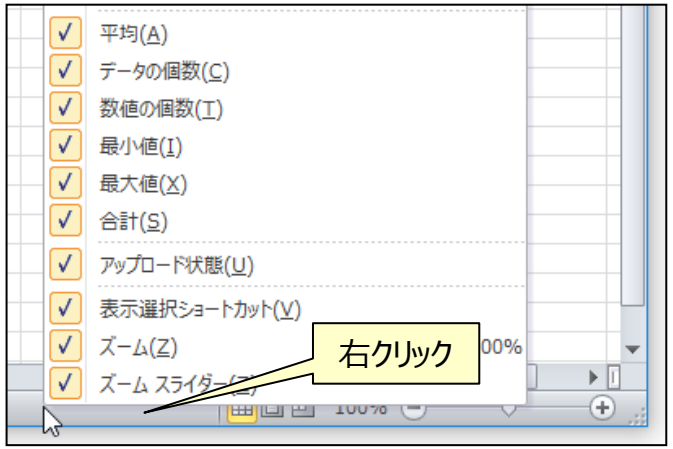
ステータスバーを右クリックし、表示項目の〈平均〉〈データの個数〉〈数値の個数〉〈最大値〉〈最小値〉〈合計〉にチェックを入れます。

- 異常値のチェックに利用するのは、主に〈データの個数〉〈数値の個数〉〈最大値〉〈最小値〉ですが、〈平均〉〈合計〉の表示もデータの確認に便利です。

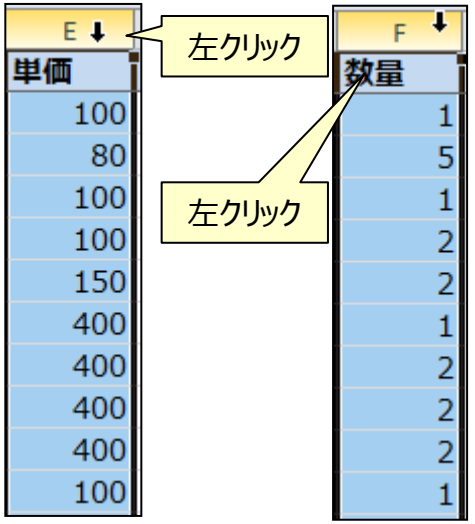
Excelの列頭のアルファベットを左クリックすることで、列全体を選択してからステータスバーを確認します。

- 例示の表のように1行目に［単価］［数量］などの変数名が入っている場合は、〈データの個数〉よりも〈数値の個数〉が1小さくなりますが、それ以上の差があれば、数値が入るべき列に文字入力があり、欠損値の可能性に気が付くことができます。
- ステータスバーの最大値や最小値が現実的な値になっているかを確認することで、簡潔な異常値のチェックができます。

ステータスバーの表示内容を選択



列全体を選択して異常値を確認



E列［単価］選択時のステータスバー

データの個数: 1032　数値の個数: 1000　最小値: 80　最大値: 400

〈データの個数〉よりも〈数値の個数〉が32小さく、
変数名以外に文字の入力が32あることに気がつきます。
最小値、最大値は現実的な値で問題はなさそうです。

F列［数量］選択時のステータスバー

データの個数: 1032　数値の個数: 999　最小値: -1　最大値: 9999

〈データの個数〉よりも〈数値の個数〉が3小さく、
変数名以外に文字の入力が33あることに気がつきます。
最小値、最大値は異常値であることにも気がつきます。

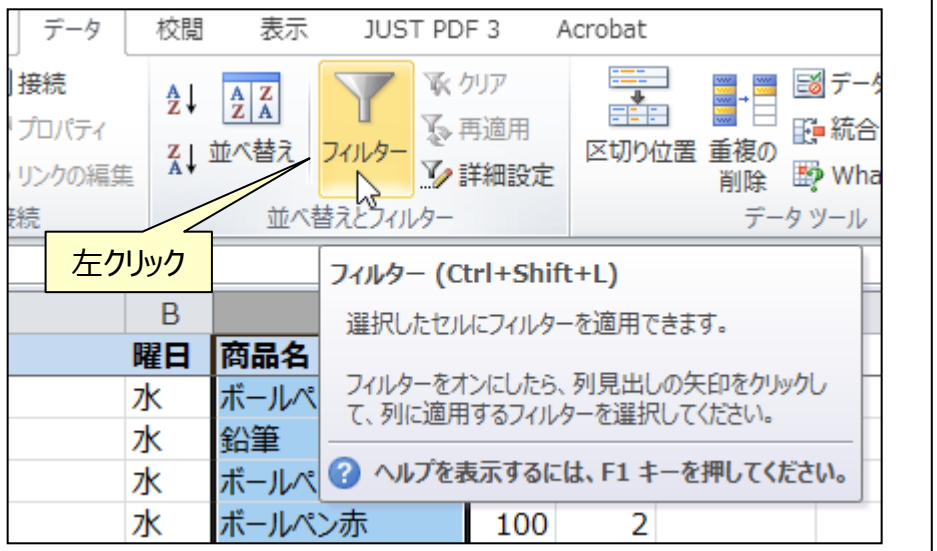
文字列の表記揺れのチェック

◆Excelの「フィルター」を使うと、文字列の表記揺れのパターンを簡単に確認できます。

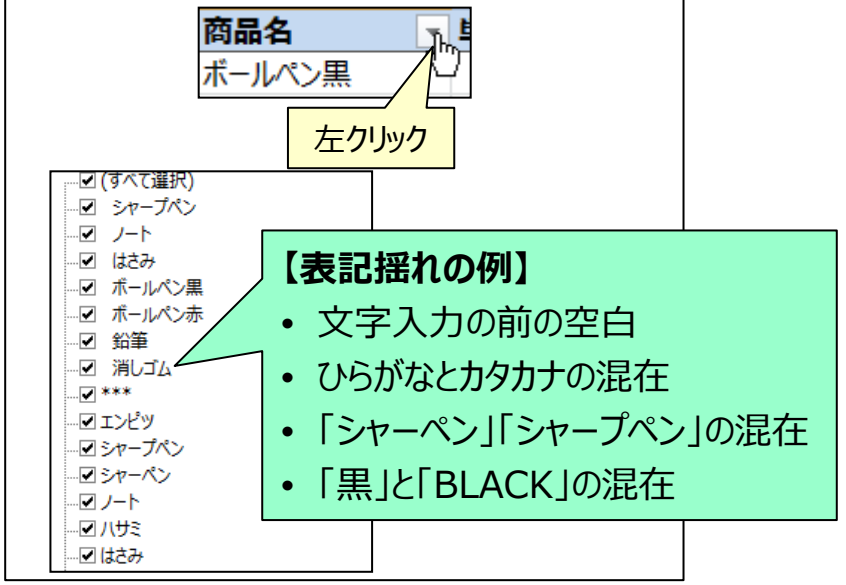
- ［商品名］の列には、「はさみ」と「ハサミ」が混在、文字のはじめに空白が入っているなどの表記揺れがあります。
  - 表記揺れはデータ集計・データ分析の障害となるため、文字列を統一する必要があります。
  - 「フィルター」の本来の用途は、指定した名称に一致する行を選択して抜き出すことですが、名称をリスト化してくれるために表記揺れの確認に利用できます。
- ひらがなとカタカナの混在、空白の挿入、半角全角の相異といった表記揺れの確認には**フィルター**が便利です。
  - Excelのフィルターは、半角と全角は区別する一方で、大文字と小文字は区別せず、大文字小文字の表記揺れはフィルターでは確認できません。

対象とする列の一部のセルを選択した状態で、Excel上部の「データ」タブにある「フィルター」を左クリックして、表示されたメニューからフィルターを左クリックしてください。表記揺れを確認したい列にある（下向きの三角▼）が入った四角をクリックすることで、文字列のリストが表示されます。

対象列を選択した状態で［データ］タブのフィルターをクリック



列頭の箱のボタンをクリック



□ 表記揺れの修正はプロセスが長いため、まずは異常値の確認と修正を行った後に説明します。

欠損値、異常値の置き換え

◆特定の条件に基づくセルの値の変換には、ExcelのIF関数が便利です。

- ExcelのIF関数は『=IF(条件式,条件を満たす場合の出力,条件を満たさない場合の出力)』とコンマで区分して入力することで、条件式で場合分けした出力ができます。
  - 括弧内に対象となる数値やセルを指定することで、定められた処理をするものを関数といいます。Excelではセルに「=」に続いて関数名を記入します。

IF関数で数値を転記し、数値でなければピリオド『NaN』を出力する場合

セルに『=IF(E2<10000,E2," NaN")』と入力すれば、対象セルのE2が考えられる上限の10000より小さい数値であればE2の値をそのまま出力し、文字を含め、それ以外なら『NaN』を出力することで、数値のみを転記できます。

- Excelの条件式において、「記号やスペースを含む全ての文字」はあらゆる数値より大きい値（無限大）として扱われます。このため、文字入力の可能性がある列においては、考えられる下限の0より大きいかを条件とする『=IF(E2>0,E2,"NaN")』ではなく、考えられる上限値（例えば10000）より小さいかを条件とする『=IF(E2<10000,E2,"NaN")』としてください。

数値のみの転記

D	E	F	G
商品名	単価	数量	改訂単価
ボールペン黒	100	1	100
鉛筆	80	5	80
ボールペン赤	100	1	100
ボールペン赤	100	2	100
ノート	150	2	150

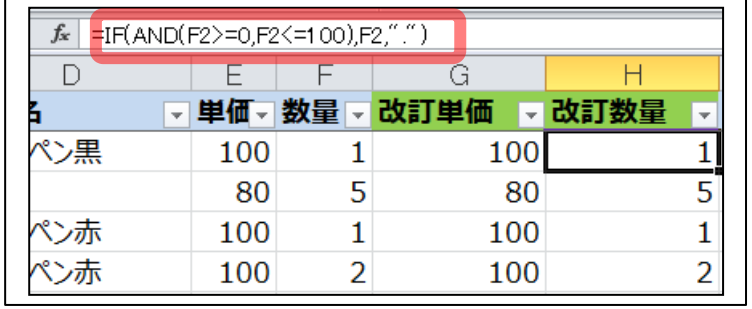
- 欠損値は、プログラミング言語や分析ソフトウェアでの利用も考慮して、『NA』『NULL』『NAN』『.』で表しますが、この講座では『NaN』に置き換えます。

IF関数で-1や9999といった異常値も除き、0以上100以下のみ数値を出力する場合

セルに『=IF(AND(F2>=0,F2<=100),F2,"NaN")』と入力すれば、対象セルのF2が0以上100以下ならE2の値をそのまま出力し、そうでなければ『NaN』を出力することで、数値のみを転記できます。

- 条件式の中に入っているANDは、両方満たす場合の「かつ」を表すExcel関数で、コンマで区切ることで複数の条件を与えることができます。また、どちらかを満たす場合の「または」を表すORというExcel関数もあります。
- Excelの条件式では「より大きい(>)」「より小さい(<)」の記号の後ろにイコール(=)を入れることで、「以上(>=)」「以下(<=)」となります。

0以上100以下の数値の転記



D	E	F	G	H
名	単価	数量	改訂単価	改訂数量
ペン黒	100	1	100	1
	80	5	80	5
ペン赤	100	1	100	1
ペン赤	100	2	100	2

関数を入力後、そのセルの右下の黒い■をダブルクリックすると、下側の列にも同じように関数が入ります。

表記揺れの統一（1）PHONETIC関数の利用

◆PHONETIC関数は、ひらがな、カタカナの表記揺れ統一に利用できます。

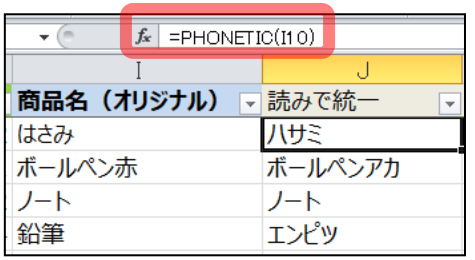
- 元の商品名を右側が空白の列にコピーしてから、**１列ずつ右に変換していく形で表記揺れを補正**していきます。
  - 2行目でExcel関数を作った後は、セルの右下の黒い四角■をダブルクリックして、列の最下段まで同じ関数を反映させます。

PHONETIC（フォネティック）関数：文字列の読み仮名をカタカナで出力

セルに『=PHONETIC(I2)』と入力すれば、対象セルI2のフリガナを出力します。

- 初期設定において、PHONETIC関数はひらがな、半角カタカナを全角カタカナで出力します。
- PHONETIC関数は、例外的に関数の出力を引き継げません。例えば、A1に「黒」や「クロ」と入力し、B1に『=PHONETIC(A1)』と入力すれば、「クロ」と出力しますが、C1に『=PHONETIC(B1)』と入力してもB1の「クロ」の出力を引き継がず、C1は空白となります。このため、PHONETIC関数は変換プロセスの最初など、値そのものを括弧内に指定する必要があります。

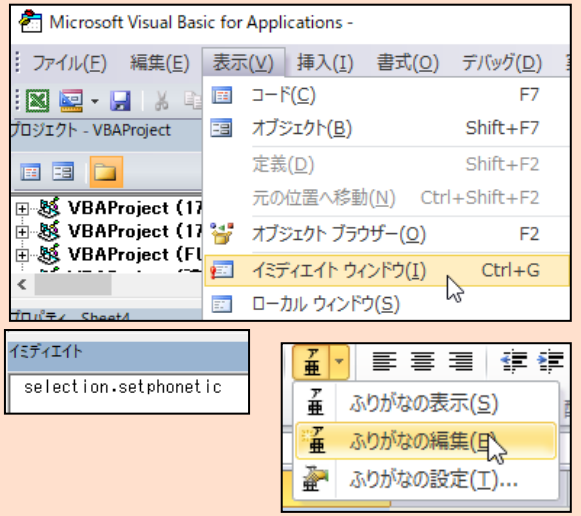
読み仮名（カタカナ）で統一



【PHONETIC関数における漢字の読み仮名】

◆ PHONETIC関数は、ひらがなとカタカナの統一には常に利用できますが、漢字の読み仮名については、Excelに実際に漢字変換して入力した設定が反映され、他のファイルからコピー&ペーストしたり、CSVを読み込んだ場合は漢字の読み仮名がつかず、漢字のまま出力されてしまいます。

◆ CSV等から読み込んだ漢字に一括して、標準的な読み仮名をつける場合は、読み仮名を付けたい範囲を選択し、「Alt +F11」でVisual Basicを表示し、表示のボタンから「イミディエイト ウィンドウ」を選択し、表示された欄に『selection.setphonetic』と入力し、Enterを押してください。



- フリガナの誤り等は、Excelのメニューの「ふりがなの編集」から変更できます。「ふりがなの設定」ではひらがな表示への変更も可能です。

表記揺れの統一（2）TRIM･UPPER関数の利用

◆TRIM関数は空白の除去、UPPER関数は大文字への統一に利用できます。

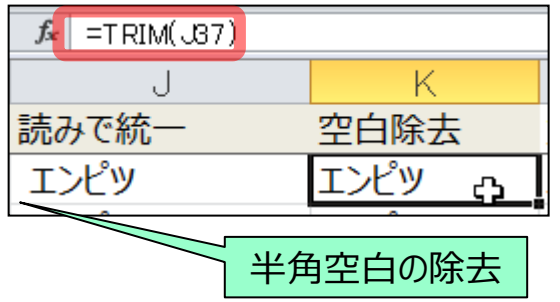
- **TRIM関数**は、文字列の前と後にある全角および半角の空白を除去して出力します。

TRIM（トリム）関数：文字列の始めと終わりの空白を削除して出力

セルに『=TRIM(J2)』と入力すれば、対象セルJ2の前後の空白を除去します。

- TRIM関数は全角の空白、半角の空白をともに除去します。
- TRIM関数は単語内で複数の空白が続く場合は、一つの空白にまとめるため、文字内に空白がある場合は、空白が全てなくなるわけではありません。
  （例）「 ノ　　ート 」→「ノ ート」

空白除去による表記揺れの統一



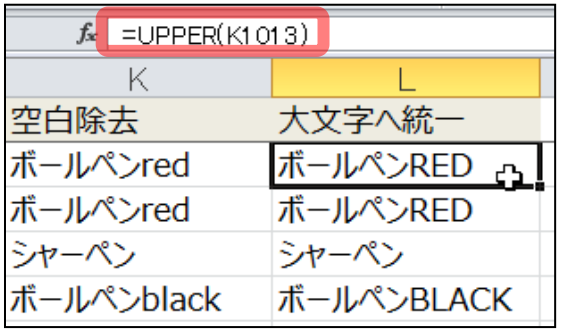
- **UPPER関数**は、英字の小文字を大文字に統一して出力します。

UPPER（アッパー）関数：文字を全て大文字に変更して出力

セルに『=UPPER(K2)』と入力すれば、対象セルK2のアルファベットを全て大文字で統一します。

- Excelの集計において、一般に全角と半角は区別する一方で、大文字と小文字は区別しません。しかし、視覚的な統一感、他のプログラムでの利用可能性を考えれば、大文字と小文字は統一している方が良いです。
- UPPER関数の代わりにLOWER関数を利用すれば、アルファベットを小文字で統一することができます。

アルファベットの大文字への統一



- 半角と全角の混在がある場合は、**ASC関数**で半角に統一するか、**JIS関数**で全角に統一して下さい。
  - 今回のデータクレンジングでは、データチェック時に半角・全角の不統一がなかったため、ASC関数の利用は省略しています。

表記揺れの統一（3）SUBSTITUTE関数の利用

◆SUBSTITUTE関数は、文字の置き換えに利用できます。

- ExcelのSUBSTITUTE関数は『=SUBSTITUTE(対象となるセルの番地,“置き換え元の文字列”,“置き換え後の文字列”)』と、コンマで区切り、引用符で文字列を区切って指定します。

SUBSTITUTE（サブスティチュート）関数：文字を置き換えて出力

セルに『=SUBSTITUTE(N2,“シャーペン”,“シャープペン”)』と入力すれば、対象セルN2の「シャーペン」という文字列を「シャープペン」に置き換えます。

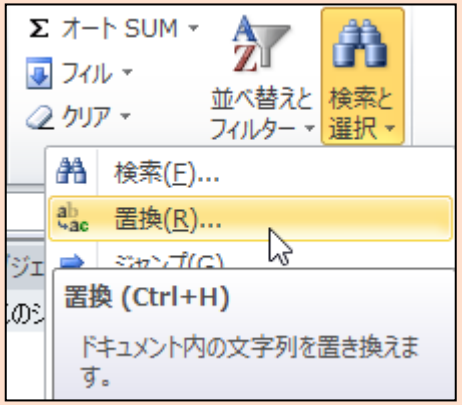
- 削除したい文字列がある場合は『=SUBSTITUTE(N2,“[削除対象文字列]”,“”)』とすることで、文字列を削除できます。

置き換えによる表記揺れの統一

【Excelメニューの検索・置換機能を利用したデータクレンジングについて】

- ◆ Excelの「検索と置換」から置換機能を使うことでも、表記揺れの統一作業は可能ですが、Excel関数を使う場合に比べて、作業手順や置換内容が分かりにくくなります。
- ◆ 本講座のようにExcel関数を使って置換すると、作業手順、置換内容を簡単に確認することができ、事後的に追加するデータレコードがある場合やデータクレンジングの方法を変更したい場合でも対応が容易です。

それでもExcelの置換機能を使う場合は、置換した文字のリスト、指定範囲を別シート等に記録するようにしましょう。



クレンジング完了の確認と値での貼り付け

◆データクレンジング後（改訂後）の列を関数との関係が切れた「値」で貼り付けます。

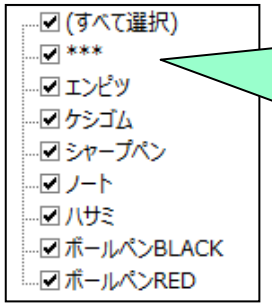
- 異常値や表記揺れの改訂が完了したことを、ステータスバーやフィルターから確認します。

クレンジング前のF列［数量］の選択時のステータスバー

クレンジング後のH列［改訂数量］の選択時のステータスバー

［改訂数量］は、-1や9999といった［数量］の異常な値を一つずつ、文字列の『NaN』に変更し「数値の個数」は2減少し、最大値、最小値も正常な範囲にあることを確認できます。

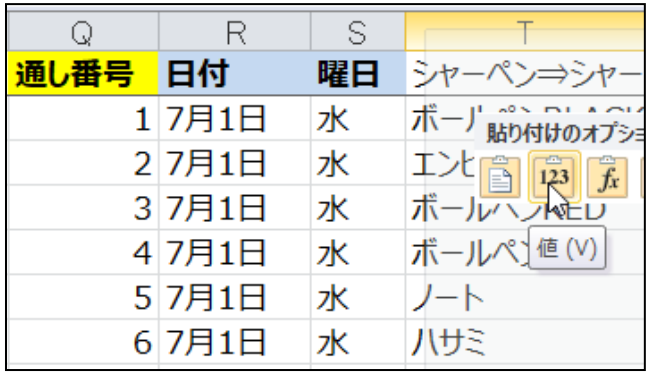
O列より、商品名リストの表記統一を確認



日付の区切りの「***」と各商品の統一された名称のみが表示されており、表記揺れが解消していることを確認できます。

- 分析等に利用する列の貼り付け先を決め、元のデータセットと1列以上空けるか別シートに**「値」で貼り付け**ます。
  - 「値での貼り付け」は、コピーしてから右クリックメニューで指定します。値で貼り付けると、計算や変換に利用したExcel関数との関係が切れるため、改めてExcel上のデータ分析が可能になるとともに、他の分析用のプログラム言語、ソフトウェアでも利用できます。
  - セルの色や表示形式も貼り付けたい場合は、いったん「Ctrl+V」の通常の貼り付けで書式等を含めて貼り付けた範囲に、改めて「値」で貼り付けます。
  - Excel上で1列をあけると、視覚的にも区切りが明らかになることに加えて、フィルター等でも別のデータセットとして認識されます。

クレンジング完了の列を値で貼り付け



表記揺れと異常値を除いた改訂済データセット（1列をあけて貼り付け）

P	Q	R	S	T	U	V
	通し番号	日付	曜日	改訂商品名	改訂単価	改訂数量
	1	7月1日	水	ボールペンBLACK	100	1
	2	7月1日	水	エンピツ	80	5
	3	7月1日	水	ボールペンRED	100	1
	4	7月1日	水	ボールペンRED	100	2
	5	7月1日	水	ノート	150	2
	6	7月1日	水	ハサミ	400	1

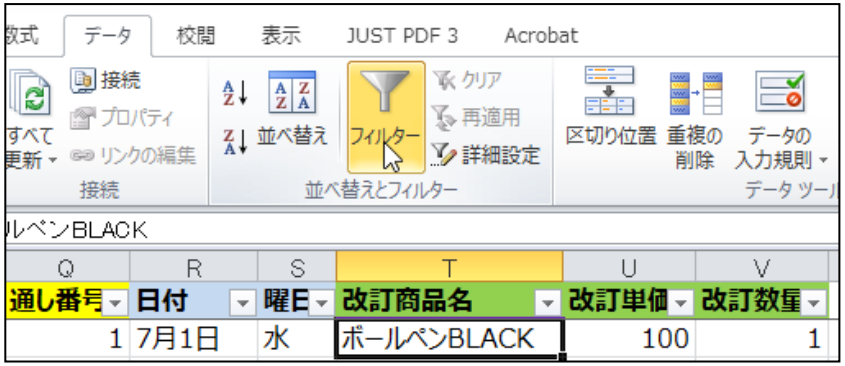
フィルターによる不要データレコード（行）の非表示

シート〔3〕クレンジング後の抽出表示
T列とV列のフィルターによる抽出表示

◆Excelのフィルター機能を使って、分析に利用するデータレコード（行）のみを表示します。

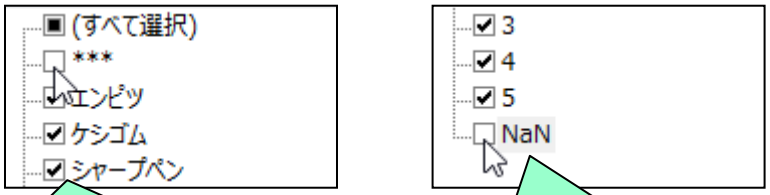
- **フィルター本来の用途**として、区切り文字や欠損値のリストのチェックマークを外し、**表示対象から除外**します。
  - 本講座では、これまでに「表記揺れの確認」のためにフィルターの表示を利用しました。
  - 異常値または欠損値があり、正しい数値が分からない場合の対処方法は、大別して二種類あります。一つの対処方法は、利用が困難なデータレコードを分析対象とするデータセットから除去すること、もう一つの対処方法はもっともらしい値を入力することです。この講座では、より簡単な対処方法として、異常値・欠損値のデータレコードを分析対象とするデータセットから除去する対処方法を示します。

改訂データセットに対してフィルターを利用



Q	R	S	T	U	V
通し番号	日付	曜日	改訂商品名	改訂単価	改訂数量
1	7月1日	水	ボールペンBLACK	100	1

区切りの「***」と異常値・欠損値の「NaN」を除外して表示



［改訂商品名］のフィルターから、日付の区切りに相当する「***」の左側のチェックマークを外します。

［改訂数量］のフィルターから、異常値を変換した『NaN』の左側のチェックマークを外します。

- データセットにおける区切り文字や欠損値のデータレコードが非表示となっていることを確認します。
  - フィルターによって非表示の行（データレコード）がある場合は、Excelの左側の行番号を示す文字が青色になります。
  - Excelの左側の行表示やの［通し番号］の値が飛んでいることで、オリジナルのデータセットから非表示になっている行があることが分かります。

フィルターをかける前（全ての列の表示）

	Q	R	S	T	U	V
1	通し番号	日付	曜日	改訂商品名	改訂単価	改訂数量
34	33	7月1日	水	ボールペンBLACK	100	1
35	34	7月1日	水	ハサミ	400	NaN
36	35	***	***	***	NaN	NaN
37	36	7月2日	木	エンピツ	80	1
38	37	7月2日	木	エンピツ	80	1

フィルターをかけた後（区切り行、欠損値の非表示）

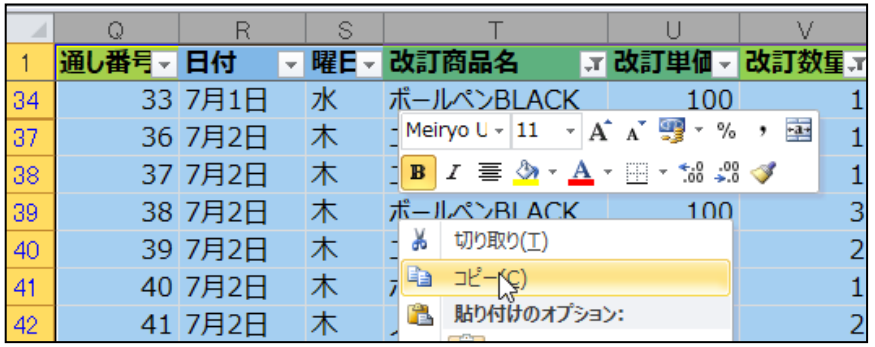
	Q	R	S	T	U	V
1	通し番号	日付	曜日	改訂商品名	改訂単価	改訂数量
34	33	7月1日	水	ボールペンBLACK	100	1
37	36	7月2日	木	エンピツ	80	1
38	37	7月2日	木	エンピツ	80	1
39	38	7月2日	木	ボールペンBLACK	100	3
40	39	7月2日	木	エンピツ	80	2

分析に利用するデータレコード（行）の抽出

◆不要な行を除去して、分析に使うデータレコードだけを別シートに貼り付けます。

- フィルターによって分析対象とするデータレコード（行）のみを表示した状態で、それらの列を全て選択し「コピー」をクリックすると、**非表示の行は点線で区切られてコピー範囲が表示**されます。

フィルターをかけた状態で「コピー」を選択



	Q	R	S	T	U	V
1	通し番号	日付	曜日	改訂商品名	改訂単価	改訂数量
34	33	7月1日	水	ボールペンBLACK	100	1
37	36	7月2日	木	[illegible]		1
38	37	7月2日	木	[illegible]		1
39	38	7月2日	木	ボールペンBLACK	100	3
40	39	7月2日	木	[illegible]		2
41	40	7月2日	木	[illegible]		1
42	41	7月2日	木	[illegible]		2

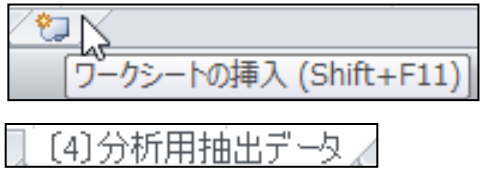
非表示の行をコピー対象に含めない点線での表示

	Q	R	S	T	U	V
1	通し番号	日付	曜日	改訂商品名	改訂単価	改訂数量
34	33	7月1日	水	ボールペンBLACK	100	1
37	[illegible]	[illegible]	木	エンピツ	80	1
38	37	[illegible]	[illegible]	[illegible]	[illegible]	1
39	38	[illegible]	[illegible]	[illegible]	[illegible]	3
40	39	7月2日	木	エンピツ	80	2
41	40	7月2日	木	ボールペンRED	100	1
42	41	7月2日	木	ノート	150	2

非表示の行が点線で区切られて表示

- Excelのシート群の右側にあるボタンをクリックして新規作成したシートに、コピーしたデータレコードを貼り付けます。
  - シート内で行（データレコード）の一貫性を確保することに加えて、データクレンジング作業と分析作業を区分するために別シートに貼り付けます。
  - クレンジング前データと抽出データの行の対応、除外された行を分かりやすくするためにも、講座の冒頭部で［通し番号］をつけていました。

新規作成したシートへの必要なデータレコードのみを貼り付け



	A	B	C	D	E	F
1	通し番号	日付	曜日	改訂商品名	改訂単価	改訂数量
34	33	7月1日	水	ボールペンBLACK	100	1
35	36	7月2日	[illegible]	[illegible]	[illegible]	1
36	37	7月2日	[illegible]	[illegible]	[illegible]	1
37	38	7月2日	[illegible]	[illegible]	[illegible]	3
38	39	7月2日	[illegible]	[illegible]	[illegible]	2
39	40	7月2日	木	ボールペンRED	100	1

分析対象から外した行（データレコード）を除外して貼り付けることができます。

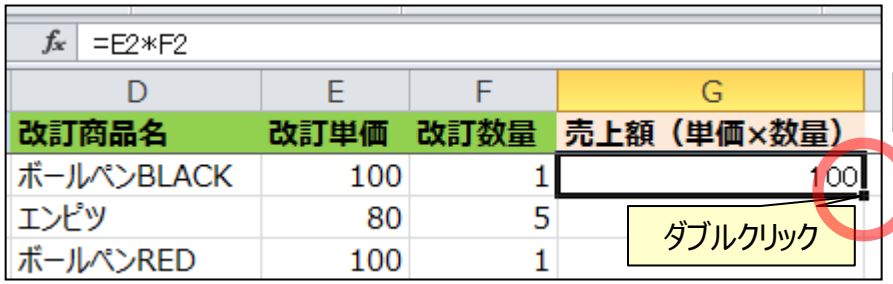
- 以上でデータクレンジングと分析用データレコードの抽出を終え、以降ではデータ集計を行います。

全体の合計値の導出：SUM関数

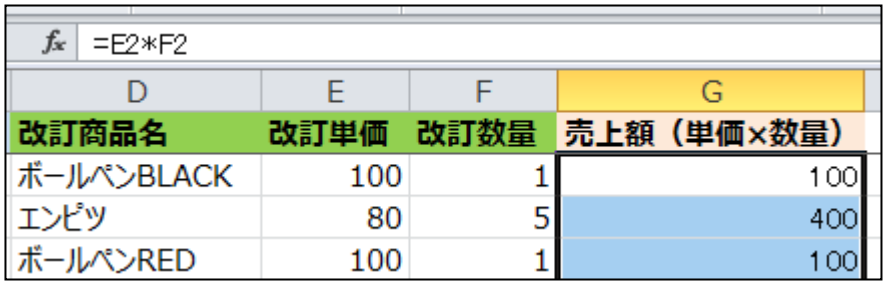
◆売上額の列に対してSUM関数を用いて、合計値を導出します。

- 個々の行にある単価と数量を掛け合わせて［売上額］を作成します。G2セルに『=E2*F2』と入力します。
  - データクレンジングの完了後はExcel上のピボットテーブルを使うケースもあれば、統計分析用のプログラミング・ソフトウェアで処理することもあります。Excelのピボットテーブルを利用したデータ処理は講座3-3、プログラミングソフトRを使ったデータ処理は講座4-3で示します。
- G2セル選択時に右下に表示される■をダブルクリックすることで、G列の下部に同じルールの計算を適用します。

G2セルにおける売上額（単価×数量）の導出



右下の■をダブルクリックすることで全行に適用



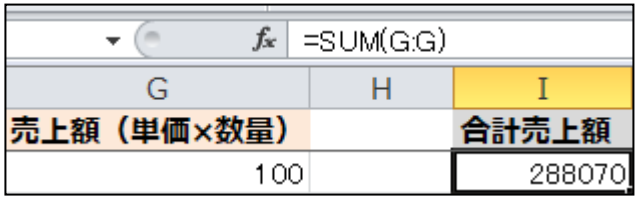
- 売上の列全体を範囲指定した際のステータスバーの合計値および合計値を導出する**SUM関数**によって、［合計売上額］を把握できます。
  - SUM関数では文字列は除外して計算されるため、変数名を含んで範囲指定をすることができます。Excelでは(左端の列:右端の列)の形で範囲指定ができるので、G列全体の合計値を導出する場合は、『=SUM(G:G)』と指定してください。

SUMIF（サムイフ）関数：対象範囲の合計値を算出

ステータスバーにおける売上高に関する表示

=

SUM関数での合計値の出力



□ 続いて、日付別や商品別に売上を集計するために、各項目のリストを作成します。

「重複の削除」による項目名リスト作成

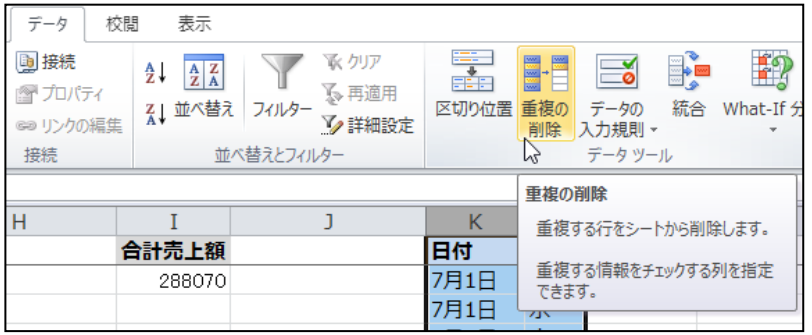
◆条件別の集計値を導出するために、「重複の削除」によって、項目名リストを作成します。

- Excel関数で条件別の集計値を導出するためには、集計対象とする項目名をリストアップする必要があります。
  - 条件別の集計値は講座3-3で示すピボットテーブルでも導出できますが、ここではExcel関数を利用した条件別の集計値の導出を示します。
- 項目名リストを作成するためにB列とC列の［日付］と［曜日］をK列とL列へ、2列あけてD列の［改訂商品名］をO列に貼り付け、各列を選択した状態でExcelのメニューの「データ」タブにある**「重複の削除」**をクリックします。

項目名リスト用の列の配置

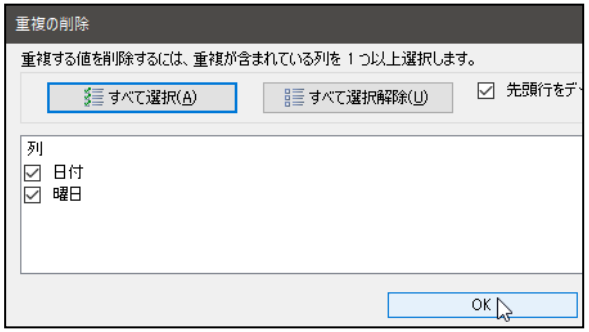
K	L	M	N	O
日付	**曜日**			**改訂商品名**
7月1日	水			ボールペンBLACK
7月1日	水			エンピツ
7月1日	水			ボールペンRED
7月1日	水			ボールペンRED
7月1日	水			ノート

K列とL列を選択して「重複の削除」をクリック



- 表示されたダイアログボックスの設定を変えずに「OK」をクリックすると、それぞれの項目名リストが表示されます。
  - 「重複の削除」は、本来、指定した変数の組み合わせから重複データを除外するための機能ですが、項目名リスト作成にも利用できます。
  - 「データレコードのIDが一致している」などの重複記入があるデータセットの場合、「重複の削除」はデータクレンジングにおいても利用します。

ダイアログボックスで「OK」をクリック



重複削除に関する情報表示

「重複の削除」は記録が残らないため、
操作内容を記録しておくことが望ましいです。

［日付と曜日］と［商品名］に関して、
重複と漏れのない項目名リストが完成

［日付］と［曜日］と［改訂商品名］のリスト

K	L	M	N	O
日付	**曜日**			**改訂商品名**
7月1日	水			ボールペンBLACK
7月2日	木			エンピツ
7月3日	金			ボールペンRED
7月4日	土			ノート
7月5日	日			ハサミ
7月6日	月			ケシゴム
7月7日	火			シャープペン
7月8日	水			

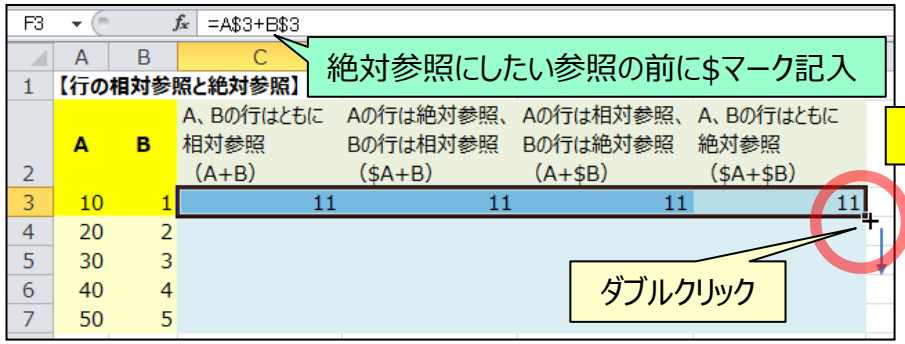
【参考】Excelにおける相対参照と絶対参照

◆Excelにおけるセルの参照には、相対参照と絶対参照の2種類があります。

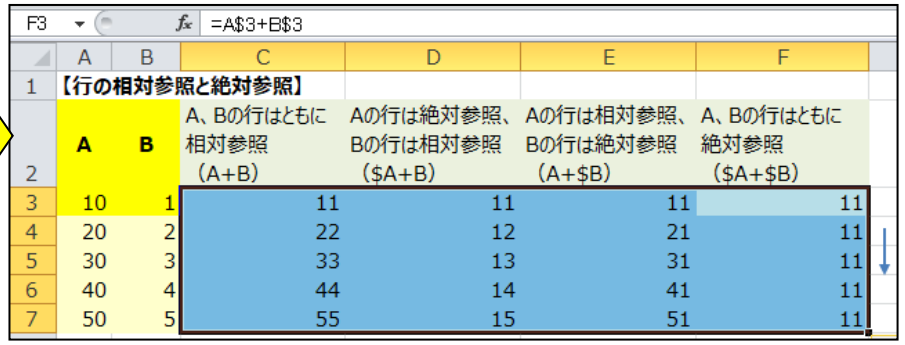
- 次のVLOOKUP関数の照合対象範囲の設定の仕方に関連して、Excelの**相対参照**と**絶対参照**を紹介します。
- Excelには、セルのコピーや入力の引き延ばしに応じて、**演算や関数が参照するセルが対応する形で移動する相対参照**と**参照するセルが固定されている絶対参照**があります。
  - Excelの標準設定は相対参照であり、絶対参照を指定する操作をしなければ、自動的に相対参照になっています。
  - 絶対参照を行うには、演算やExcel関数で他のセルを参照しているセルの記入に$（ドル記号）を挿入します。
- 実習用Excelのシート〔参考〕では、「行」「列」「行・列」の3種で、相対参照と絶対参照の違いを確認できます。

シート〔参考〕のC3～F3セルを選択した状態で、F3の右下に表示される■をクリックして確認してください。

シート〔参考〕の行の相対参照と絶対参照



相対参照と絶対参照の組み合わせによる出力



シート〔参考〕のJ列およびP列の青いセルの範囲を選択して、水色の範囲に入力を引き延ばして、出力を確認してください。

- 演算や関数で参照するセルの指定の前に**「$（ドル記号）」を記入すると、絶対参照**になり、セルのコピーや入力の引き延ばしをしても、参照するセルが変わらず固定されています。
  - 横の行のみ絶対参照にする場合は「A$1」、縦の列のみ絶対参照にする場合は「$A1」という形式で指定し、行と列の双方を固定する場合は「A1」と固定したい行または列の前に$を記入します。

データセットの結合（1）VLOOKUP関数の利用

◆VLOOKUP関数を使って、データセットの結合や変数の追加ができます。

- シート〔4〕と〔5〕のAA・AB列には日付別天気データがあり、日付・曜日のリストの右側に天気のデータを結合します。
- 天気データを日付データに結合するためには、Excelの**VLOOKUP関数**を利用します。
  - 今回の実習用データでは、K列は重複のない日付順になっているのでAB列の該当部分を貼り付けることでもデータセットを結合できますが、VLOOKUP関数は、照合する文字に重複があったり、照合する文字が順不同になっている項目であってもデータセットを結合できます。
- ExcelのVLOOKUP関数は『=VLOOKUP(照合する文字,照合対象範囲,照合対象範囲内の表示する列目,近似値の可否)』と、コンマで区切って指定することで、「照合する文字」が「照合対象範囲」の一列目が一致した場合に、「照合対象範囲内の表示する列目」の文字列を表示します。

M2のセルに『=VLOOKUP(K2,AA2:AB36,2,FALSE)』と入力すると、7月1日の天気データとして「晴れ」と表示されます。

- Excel関数における範囲の指定は、「左上のセルの番地：右下のセルの番地」として指定します。「AA2:AB36」という指定で、6月29日～8月2日の天気データが記載された35行2列の長方形の範囲指定となります。なお「AA:AB」とセルの番地の数値の記入を抜けば、列全体の指定になります。
- VLOOKUP関数の第1項目の「K2」にある文字「7月1日」と第2項目の「AA2:AB43」の範囲の1列目の「6月29日～8月2日」の文字列を照合し、第4項目によって近似値を許可せずに完全一致の文字列がある行において、第3項目の「2」列目の文字をセルに出力します。
- 近似値の可否は、許可する場合は「TRUE」、許可しない場合は「FALSE」と入力します。データセットの結合では、原則として「FALSE」にしてください。

VLOOKUP関数を使用して対応する「天気」を出力

M2 =VLOOKUP(K2,AA2:AB43,2,FALSE)

	J	K	L	M	N
1		日付	曜日		
2		7月1日	水	晴れ	
3		7月2日	木		
4		7月3日	金		
5		7月4日	土		

照合する文字の列　関数の出力の列

照合対象範囲の［日付］と［天気］のリスト

AB4 晴れ

	Z	AA	AB	AC	AD
1		日付	天気		
2		6月29日	くもり		
3		6月30日	くもり		
4		7月1日	晴れ		
5		7月2日	くもり		

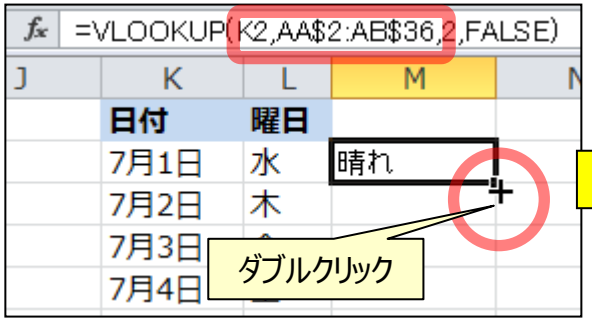
照合対象範囲の1列目　照合対象範囲の2列目

データセットの結合（2）VLOOKUP関数の指定

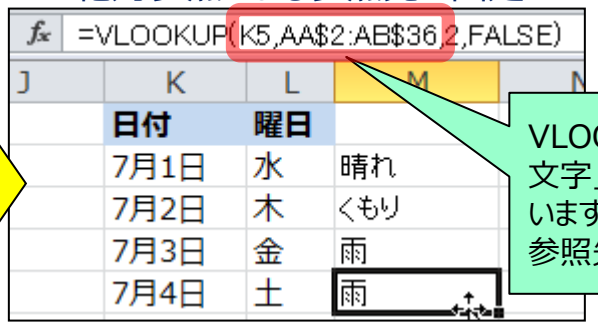
◆VLOOKUP関数の照合対象範囲は、絶対参照での指定が便利です。

- VLOOKUP関数を利用したM2のセルの『=VLOOKUP(K2,AA2:AB36,2,FALSE)』の第2項目の「照合対象範囲」の行を絶対参照に変更すべく『=VLOOKUP(K2,AA**$**2:AB**$36**,2,FALSE)』と$マークを挿入します。
  - 今回の実習用データでは、結合・追加したい変数が天気データの一種類のみであるため、照合対象範囲の列側を絶対参照で固定する必要はありません。もし、結合・追加したい変数が複数ある場合は、第2項目は行・列ともに絶対参照で照合対象範囲を固定して、第3項目の数字を増やすことで複数の変数を結合・追加します。
- M2セルの右下の■をダブルクリックするか、下側へドラッグすることによって、3行目以下のセルも同様に入力します。

絶対参照での入力の引き延ばし



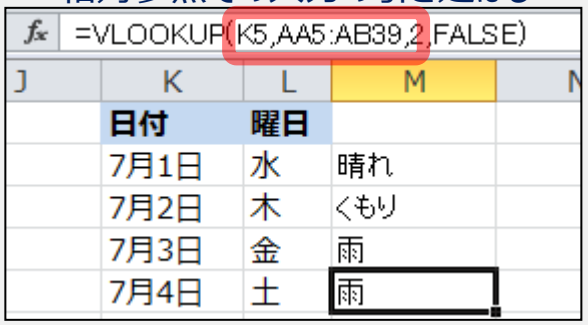
絶対参照による参照先の固定



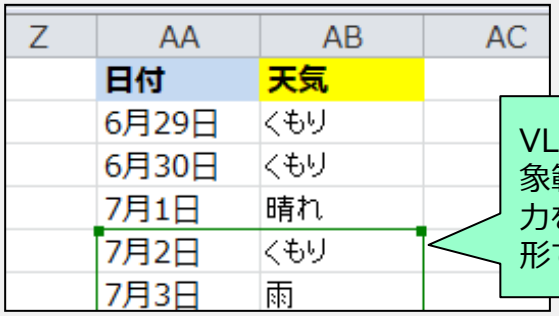
VLOOKUP関数の第1項目の「照合する文字」は、相対参照で参照先が移動していますが、第2項目の「照合対象範囲」の参照先は絶対参照で固定されています。

□ VLOOKUP関数の第2項目を相対参照のままで入力を引き延ばすと、照合対象範囲の参照先がずれてしまいます。

相対参照での入力の引き延ばし



相対参照による参照先のずれ



VLOOKUP関数の、第2項目の「照合対象範囲」が相対参照のままでは、M列の入力を引き延ばした際に、参照先も対応する形で移動して、ずれてしまいます。

項目名別合計値の導出：SUMIF関数

◆SUMIF関数は、項目別の合計値を導出できます。

- ExcelのSUMIF関数は『=SUMIF(照合対象範囲,照合する文字,合計対象範囲)』と、コンマで区切って指定することで、「照合対象範囲」と「照合する文字」が一致した行に関して、「合計対象範囲」の合計値を導出します。
- 日付別の合計売上額を導出するために、「照合対象範囲」をデータセットの日付のB列、「照合する文字」を項目名リストの重複のない日付のK列、「合計対象範囲」をデータセットの売上額のG列に設定します。

SUMIF（サムイフ）関数：「照合対象範囲」と「照合する文字」が一致した行で合計値を算出

「照合対象範囲」および「合計対象範囲」の最上段、最下段の行を絶対参照で指定し、『=SUMIF(B$2:B$998,K2,G$2:G$998)』と入力します。

2行目の日付別合計額が正しく表示されれば、セルの右下の■をダブルクリックするか、下側に延ばして3行目以下も同様に入力します。

- 照合対象範囲や合計対象範囲を列全体とする場合は、絶対参照とは別の指定方法として『=SUMIF(B:B,K2,G:G)』と指定する方法もあります。相対参照による問題は、指定セルの移動によって参照先の行が変わりますが、列全体を指定すると変化する行自体がありません。

［日付別合計額］の表示

=SUMIF(B$2:B$998,K2,G$2:G$998)

K	L	M	N
日付	曜日	結合：天気	日付別合計額
7月1日	水	晴れ	11020
7月2日	木	くもり	10450

代替的な指定方法：=SUMIF(B:B,K2,G:G)

- 同様の手順で項目名リストの商品名の右側に、商品別販売数および商品別合計数を導出します。

商品別売上数は、先頭行に「照合する文字」を「ボールペンBLACK」のセル番地N2とする『=SUMIF(D$2:D$998,N2,F$2:F$998)』と入力し、「ボールペンBLACK」の売上数の表示後にセル右下の■をダブルクリックします。

商品別合計額は、先頭行に「照合する文字」を「ボールペンBLACK」のセル番地N2とする『=SUMIF(D$2:D$998,N2,G$2:G$998)』と入力し、「ボールペンBLACK」の合計額の表示後にセル右下の■をダブルクリックします。

［商品別売上数］［商品別合計額］の表示

=SUMIF(D$2:D$998,O2,F$2:F$998)

O	P	Q
改訂商品名	商品別売上数	商品別合計額
ボールペンBLACK	471	47100
エンピツ	339	27120
ボールペンRED	164	16400
ノート	303	45450

項目名別平均値の導出：AVEREGEIF関数

◆AVERAGEIF関数は、項目別の平均値を導出できます。

- ExcelのAVERAGEIF関数は『=AVEREGEIF(照合対象範囲,照合する文字,平均対象範囲)』と、コンマで区切って指定することで、照合対象範囲と照合文字が一致した行に関して、平均対象範囲の平均値を導出します。
- ［曜日別平均売上額］および［天気別平均売上額］を導出するために、S列に「月、火、水、木、金、土、日」の［曜日］、U列に「晴れ、くもり、雨」の［天気］の項目を文字列で記入し、項目名別リストの合計売上額をAVERAGEIF関数で指定します。

AVERGEIF（アベレージイフ）関数：「照合対象範囲」と「照合する文字」が一致した行で平均値を算出

［曜日別平均売上額］は、先頭行に照合する文字を「月」のセル番地S2とする『=SUMIF(L$2:$L$32,S2,N$2:N$32)』と入力し、「月」の平均売上額の表示後にセル右下の■をダブルクリックします。

［天気別平均売上額］は、先頭行に照合する文字を「晴れ」のセル番地S2とする『=SUMIF(M$2:M$32,U2,N$2:N$32)』と入力し、「晴れ」の平均売上額の表示後にセル右下の■をダブルクリックします。

［曜日別平均売上額］および［天気別平均売上額］の表示

T2　=AVERAGEIF(L$2:L$32,S2,N$2:N$32)

	R	S	T	U	V
1		曜日	曜日別平均売上額	天気	天気別平均売上額
2		月	8418	晴れ	10859
3		火	9303	くもり	8901
4		水	10568	雨	7857
5		木	9994		
6		金	8868		
7		土	9843		
8		日	7668		

［曜日別平均売上額］の出力より、水曜日が最も平均売上額が高く、日曜日が最も平均売上額が低いことが分かります。

［天気別平均売上額］は、晴れの日で10,859円、くもりの日で8,901円、雨の日で7,857円です。

時間の経過に関するデータの可視化：折れ線グラフ

◆時間の経過に関するデータの可視化は、折れ線グラフが適しています。

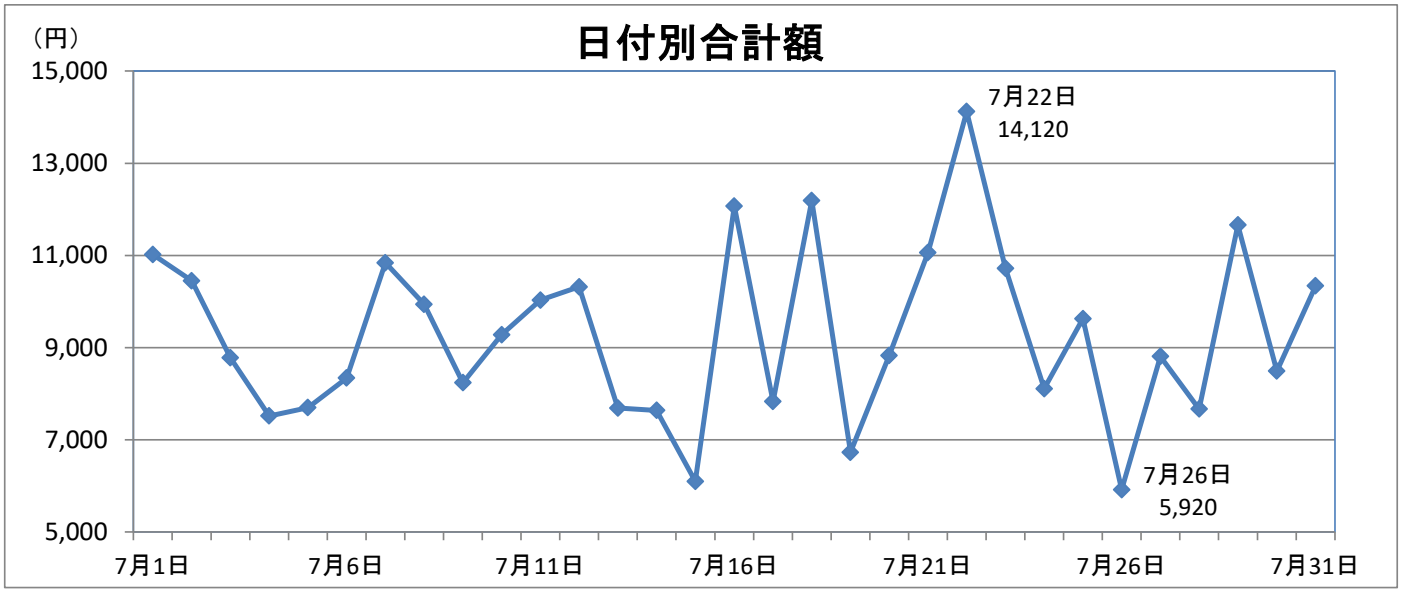
- 推移などの**時間の経過に関するデータの可視化は折れ線グラフ**が適しており、［日付別合計額］を可視化します。
  - 集計表も可視化の一種ですが、画像で把握できるグラフでの可視化の方が理解しやすく、印象に残りやすくなります。

グラフの横軸にしたい日時の情報と数値のデータを縦に並べ、範囲を選択した状態で、Excelの「挿入」タブのグラフ「折れ線」内にある「折れ線」または「マーカー付き折れ線」をクリックします。

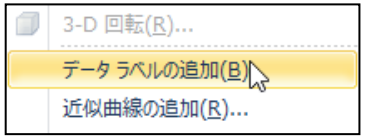
集計表による可視化

日付	日付別合計額
7月1日	11,020
7月2日	10,450
7月3日	8,780
7月4日	7,520
7月5日	7,700
7月6日	8,340
7月7日	10,840
7月8日	9,940
:	:

折れ線グラフによる推移の可視化



- 右上のグラフでは最大値と最小値のデータラベルを折れ線グラフ内に記入しています。
  - グラフ内の各データを表すマーカーを右クリックして「データラベルの追加」をクリックすると、グラフ内にデータの値が表示されます。
  - 全てのマーカーが選択された状態で「データラベルの追加」を選択することで、一括してデータラベルを入れることができます。
  - 表示されたデータラベルを右クリックし、「データラベルの書式設定」を選択することで、データラベルの表示内容や区切り方を選択できます。



割合に関するデータの可視化：円グラフ

◆割合に関するデータの可視化は、円グラフが適しています。

- 割合の**構成表示は円グラフ**が適しており、［商品別合計額］の構成を円グラフで可視化します。
- 円グラフは割合が大きい順に表示すると分かりやすいため、まず［商品別合計額］が大きい項目順に並べ替えます。

並べ替えたい変数を指定して、「データ」タブの「並べ替え」をクリックして、「降順（大きい方から降りていく順）」に並べ替えます。

Excelの機能における「並べ替え」による「商品別合計額」の降順の並び替え

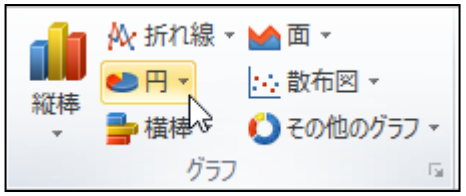
「並べ替え」は記録が残らないため、操作内容を記録しておくことが望ましいです。

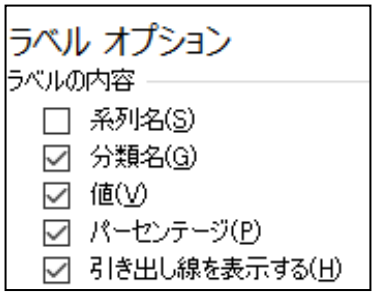
割合を区分する商品名の情報と割合を表したい商品別合計額の情報を縦に並べ、範囲を選択した状態でExcelの「挿入」タブのグラフ「円」をクリックします。

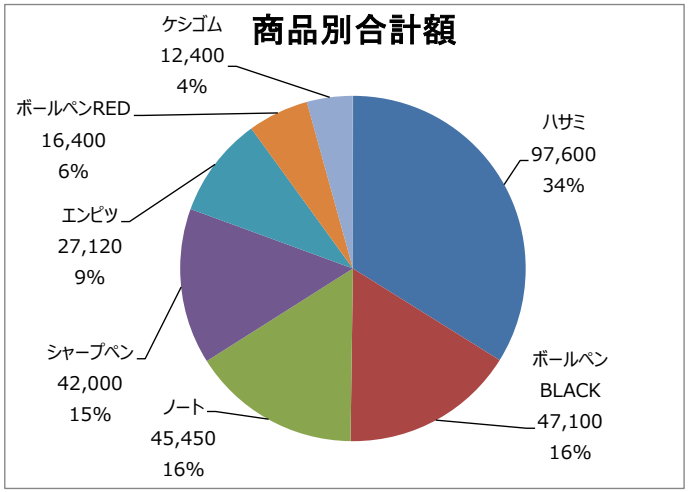
- 円グラフを右クリックして「データラベルの追加」をクリックし、「値」にチェックすると額が表示され、「パーセンテージ」にチェックすると割合が表示されます。
- データラベルの「引出線を表示する」によって、グラフの外側の白地の部分にラベルの情報を記入できます。

「商品別合計額」が大きい順に並べ替えたデータに基づく円グラフの作成

商品名	商品別合計額
ハサミ	97,600
ボールペンBLACK	47,100
ノート	45,450
シャープペン	42,000
エンピツ	27,120
ボールペンRED	16,400
ケシゴム	12,400

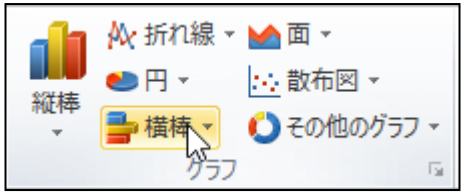






項目間の差を示すデータの可視化：棒グラフ

◆水準や項目間の差に関する表示は、棒グラフが適しています。

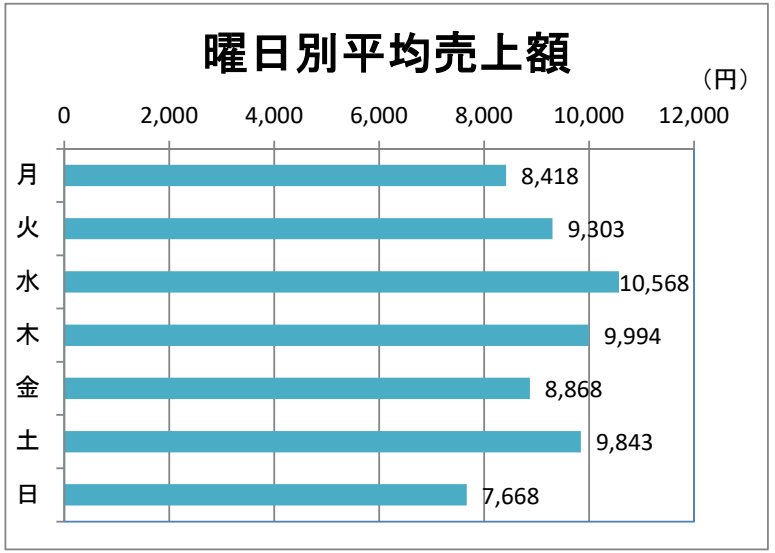


- 異なる**項目の水準の差を示す可視化は棒グラフ**が適しています。
  - 分類としたい項目のテキストと水準を表したい数値を縦に並べ、範囲を選択した状態でExcelの「挿入」タブのグラフ「縦棒」または「横棒」クリックします。
    - ここでは例示として、横棒のグラフを使って［曜日別平均売上額］の可視化を説明します。
- 可視化においては、作成者が強調したい点を図表に入れ込むことができるケースがあります。
  - 営業資料などの作成者の主観的な強調が許容されるケースもあれば、学術資料などの客観的・画一的な表示が望ましいケースもあります。
- 下記の3種の図表が持っている情報量は同じですが、情報の分かりやすさ、印象は異なります。
  - グラフの最大値、最小値といった表示範囲を変更し、縮尺を変えるだけでも、グラフの印象が変わります。

表による可視化

曜日	曜日別 平均売上額
月	8,418
火	9,303
水	10,568
木	9,994
金	8,868
土	9,843
日	7,668

客観的・画一的な棒グラフ



尺度を変更し、強調表示をした棒グラフ