ULTIMATE

GUIDE TO Cleaning Data WITH EXCEL AND GOOGLE SHEETS

CHRIS RAFTER

A complete guide to cleaning and preparing data for analysis using Excel™ and Google Sheets™

Produced by:

Table of Contents

ANALYSTS

Chapter 1: Introduction

So you just got handed a new data file. It's tempting to just load it up into your favorite visualization tool. But your first stop should be to determine the quality of your data.

The truth is, most data has at least a few data quality problems. The data may have been collected recently, or maybe it came from an application.

You'd have good reason to check it's quality before proceeding.

Data with quality issues can often operate just fine in its native application. It could be a duplicate record that nobody ever accesses, they might not even know it's there.. The other reason is that most application data is looked at a small sliver at a time. One account or customer at a time. Rarely does anyone export the entire dataset and look at it in aggregate. Over time, duplicate and inaccurate records build up and are rarely purged.

Poor data quality is the kryptonite of good reporting and credible analytics. If your data isn't of adequate quality, at worst you won't be able to proceed any further. At best, others may question your conclusions if you can't show the right attention to data quality.

In this book, I'll use terms like "Company" and "Business". But these techniques really apply to any organization that works with data. Kindly insert schools/governments/ not-for-profits/religious organizations/political campaigns/etc. as needed.

I'll also use "Customers " a lot, but substitute your term of choice for the people

and persons you interact with. You may call them subscribers, members, voters, students, associates or citizens. They’re all Person data types.

Chapter 2: Why Clean Data?

Data cleansing is the process of spotting and correcting inaccurate data. Organizations rely on data for many things, but few actively address data quality. Whether it's the integrity of customer addresses or ensuring invoice accuracy. Ensuring effective and reliable use of data can increase the intrinsic value of the brand. Business enterprises must assign importance to data quality.

A data driven marketing survey conducted by Tetra data found that 40% of marketers do not use data to its full effect. Managing and ensuring that the data is clean can provide significant business value.

Improving data quality can eliminate problems like expensive processing errors, manual troubleshooting, and incorrect invoices. Data quality is also a way of life because important data like customer information is always changing and evolving.

Business enterprises can achieve a wide range of benefits by cleansing data and managing quality which can lead to lowering operational costs and maximizing profits.

Who are the heroes who allow the organization to seize and enjoy all these benefits? I affectionately refer to these poor souls as **PWCD's**, or **People Who Clean Data**[1].

These brave people, and hopefully you are reading this because you hope to be one of them, are the noblest. They often get little recognition even though they clean up the messes of hundreds, if not thousands of other people every day. They are the noble janitors of the data world. And I salute them.

Top 5 Benefits of Data Cleaning

Improve the Efficiency of Customer Acquisition Activities

Business enterprises can significantly boost their customer acquisition and retention efforts by cleansing their data regularly. With the high throughput of the prospecting and lead process, filtering, cleansing, enriching having accurate data is essential to its effectiveness. Throughout the marketing process, enterprises must ensure that the data is clean, up-to-date and accurate by regularly following data quality routines. Clean data can also ensure the highest returns on email or postal campaigns as chances of encountering outdated addresses or missed deliveries are very low. Multi-channel customer data can also be managed seamlessly which provides the enterprise with an opportunity to carry out successful marketing campaigns in the future as they would be aware of the methods to effectively reach out to their target audience.

Improve Decision Making Processes

The cornerstone of effective decision making in a business enterprise is data. According to Sirius Decisions, data in an average B2B organization doubles every 12-18 months and though the data might be clean initially, errors can creep in at any time. In fact, in nearly all businesses where data quality is not managed, data quality decay is constantly at work. Each time new records are added; duplicates may be created. Things happening outside your organization, like customers moving and changing emails and telephone numbers will, over time, degrade data quality.

Yet the majority of enterprises fail to prioritize data quality management, or even acknowledge they have a problem! In fact, many of them don’t even have a

record of the last time quality control was performed on their customer's data. More often than not they merely discard or ignore data they believe to be of poor quality, and make decisions through other means. Here you can see that data quality is a massive barrier toward digital transformation and business intelligence, much less every company's desire to become more Data Driven.

Accurate information and quality data are essential to decision making. Clean data can support better analytics as well as all-round business intelligence which can facilitate better decision making and execution. In the end, having accurate data can help business enterprises make better decisions which will contribute to the success of the business in the long run.

Streamline Business Practices

Eradicating duplicate and erroneous data can help business enterprises to streamline business practices and avoid wasteful spending. Data cleansing can also help in determining if particular job descriptions within the enterprise can be changed or if those positions can be integrated somewhere else. If reliable and accurate sales information is available, the performance of a product or a service in the market can be easily assessed.

Data cleansing along with the right analytics can also help the enterprise to identify an opportunity to launch new products or services into the market at the right time. It can highlight various marketing avenues that the enterprises can try. In practically any other business process you can name, decisions are made every day, some large, but many small. It is this systematic pushing of high-quality information down the chain of command, into the hands of individual contributors that helps them improve decisions made at all levels of the organization. Called Operational Intelligence, it is used more commonly for quick lookups and to inform the thousands of decisions that are made every day inside the organization.

Increase Productivity

Having a clean and properly maintained enterprise dataset can help organizations ensure that the employees are making the best use of their time and resources. It can also prevent the staff of the enterprise from contacting customers with out-of-date information or create invalid vendor files in the system by conveniently helping them to work with clean records thereby maximizing the staff's efficiency and productivity. High quality data helps reduce the risk of fraud, ensuring the staff has access to accurate vendor or customer data when payments or refunds are initiated.

Increase Revenue

Business enterprises that work on improving the consistency and increasing the accuracy of their data can drastically improve their response rates which results in increased revenue. Clean data can help business enterprises to significantly reduce the number of returned mails. If there are any time sensitive information or promotions that the enterprise wants to convey to their customers directly, accurate information can help in reaching the customers conveniently and quickly.

Duplicate data is another aspect which can be effectively eradicated by data cleansing. According to Sirius Decisions, the financial impact of duplicate data is directly proportional to the time that it remains in the database.

Duplicate data can significantly drain the enterprise's resources as they will have to spend twice as much on a single customer. For example, if multiple mails are sent to the same customer, they might get annoyed and might completely lose interest in the enterprise's products and services.

Chapter 3: How Spreadsheets Became The #1 BI Tool in the World

Microsoft Excel and Google Sheets are the first choice of many users when it comes to handling large amounts of data. They're readily available, easy to learn and support universal file formats. When it comes to using a spreadsheet application like Excel or Google Sheets, the point is to present data in a neat, organized manner which is easy to comprehend. They're also on nearly everyone's desktop, and were probably the first data-centric software tool any of us learned.

In this eBook, we are going to tell you some of the tips as to how to clean and prep up your data using Excel and Google Sheets, and make it accurate and consistent, and make it look elegant, precise, and user-friendly.

Risks of Cleaning Data In Spreadsheets

 While spreadsheet tools are quite adequate for many small to mid-level data chores, there are some important risks to be aware of. Spreadsheets are desktop-class, file-oriented tools which means their entire data contents are stored in volatile RAM while in use and on disk while you're not using them. That means that between saves, the data is stored in RAM, and can be lost.

Spreadsheet tools also lack any auditing, change control, and meta-data features that would be available in a more sophisticated data cleaning tool. These features act as backstops for any unintended user error. Caution must be exercised when using them as multiple hours of work can be erased in a microsecond.

Unnoticed sorting and paste errors can also tarnish your hard work. If the data saves to disk while in this state, it can be very hard, if not impossible, to undo the damage and revert to an earlier version.

Spreadsheets also lack repeatable processes and automation. If you spend 8 hours cleaning a data file one month, you'll have to repeat nearly all of those steps the next time another refreshed data file comes along. More purpose-designed tools like Inzata Analytics allow you to record and script your cleaning activities via automation. Data is also staged throughout the cleaning process, and rollbacks are instantaneous. You can set up data flows that automatically perform cleaning steps on new, incoming data. Basically, this lets you get out of the data cleaning business almost permanently.

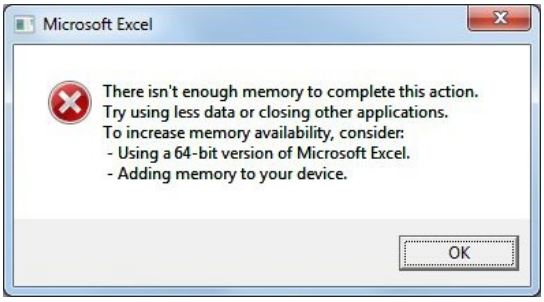
Performance and Size Limits in Spreadsheet Tools

Most folks don’t bother to check the performance limits in Spreadsheet tools before they start working with them. That’s because the majority won’t run up against them. However, if you start to experience slow performance, it might be a good idea to refer to the limits below to measure where you are and make sure you don’t start stepping beyond them. Like I said above, spreadsheet tools are fine for most small data, which will suit the majority of users.

Excel Limits

Excel is limited to 1,048,576 rows by 16,384 columns in a single worksheet.

- A 32-bit Excel environment is subject to 2 gigabytes (GB) of virtual address space, shared by Excel, the workbook, and add-ins that run in the same process.
- 64-bit Excel is not subject to these limits and can consume as much memory as you can give it. A data model’s share of the address space might run up to 500 – 700 megabytes (MB), but could be less if other data models and add-ins are loaded.



Google Sheets Limits

- Google Spreadsheets are limited to 5,000,000 cells, with a maximum of 256 columns per sheet. (Which means the rows limit can be as low as 19,231, if your file has a lot of columns!)
- Uploaded files that are converted to the Google spreadsheets format can't be larger than 20 MB, and need to be under 400,000 cells and 256 columns per sheet.

In real-world experience, running on midrange hardware, Excel can begin to slow to an unusable state on data files as small as 50mb-100mb. Even if you have the patience to operate in this slow state, remember you are running at redline. Crashes and data loss are much more likely!

If you believe you will be working with larger data, why not check out a tool like Inzata, designed to handle profiling and cleaning of larger datasets?

Chapter 4: Which is better? Google or Excel?

Either tool can be very effective. It really comes down to which one you're most comfortable with. Excel came first, and Google Sheets has tried very hard to emulate its major features.

The most visible difference between the two is that Excel is mainly used as a desktop app, and Google Sheets is mainly used as a cloud application (i.e., accessed in a browser).

This has both pros and cons:

Desktop App (Excel)

- More robust UI and ribbon, more right-click options ● Takes up local disk space ● Quicker user experience, but computations can sometimes be slower ● Opening and saving files takes slightly longer ● Like any desktop app, it can crash and there is a risk of losing work ● Opening multiple local files is faster ● Files always saved to local storage ● More wizards and troubleshooting available

Cloud App (Google Chrome or Excel Online)

- App is in the cloud, requires lower desktop computer resources
- Browser interface is less robust, and can sometimes be less responsive than a desktop app
- Less right-click and context-window functionality
- Calculations and queries on larger datasets are often faster, since it executes on cloud servers.
- Much less likely to crash.
- Work is constantly saved in the cloud, your files are always updated.
- Collaboration and sharing are slightly easier

For this book, we will alternate between using screenshots of both tools. Most of the operations are identical and we will point out where special instructions for one tool or the other are required.

Chapter 5: Types of Data Quality Problems

Now it's time to start thinking about the specific quality issues within our data. There is a whole field built around multiple aspects of Data Quality Measurement and Management. Because it was written by data nerds, of course it has a fancy title: THE SIX PRIMARY DIMENSIONS FOR DATA QUALITY ASSESSMENT.

(BTW, that right there is an 8-word title for a concept that's only six-words long. Hooray for Data Nerds!) ● Completeness ● Uniqueness ● Timeliness ● Validity ● Accuracy ● Consistency But this is not a book about data quality, it's a book for people wanting to clean data with Excel and Google Sheets. So I'm not going to explore the entire world of data quality dimensions.

I am going to zero in on the most common problems you're likely encounter with dirty data, and show you the fastest possible way to:

1. Spot them, and
2. Evaluate them, and
3. Fix them

Most Common Data Problems

I've picked the most common data problems you're likely to see in the average organization. There might be other problems not on this list, but the ones below are the most common. If you do come up with a new one, drop us a line. We're always on the hunt for new species of data problem!

While you're reading these below, focus on remembering how to *spot* them. We'll get to fixing them in a later chapter.

Missing values

Many times because the data has not been entered in the system correctly, or certain files may have been corrupted, the data has several missing variables. For example, if an address does not include a zip code at all, the remaining information can be of little value, since the geographical aspect of it would be hard to determine.

Null values

Some systems exporting data will output a "NULL" or "0" value when there is a blank field. These will end up in your data file, and they are equivalent to a blank field, however if you check only for Blank values, you might be undercounting the real number of missing values. There may be other conventions used by your system, such as multiple "0"s, "xxxxx''s, so watch for patterns of those as well.

Partial or Incomplete Values

Sometimes there is data in a field, but it's not all the data you're expecting. Incomplete values can be as simple as an initial instead of a name in the Last Name field. These are most often caused by human data entry error but they

could also come from failed writes or a mis-mapping of fields.

Another example might be where a single data field contains an incomplete value (such as a phone number without the area code). The other type of incomplete value is an incomplete record. For example, a Customer Address record that is missing a Zip code. It's important to differentiate between the two because each one must be addressed

Duplicates

Multiple copies of the same records take a toll on the computation and storage, but may also produce skewed or incorrect insights when they go undetected. One of the key problems could be human error — someone simply entering the data multiple times by accident — or it can be an algorithm that has gone wrong.

A remedy suggested for this problem is called "data deduplication". This is a blend of human insight, data processing and algorithms to help identify potential duplicates based on likelihood scores and common sense to identify where records look like a close match.

Mis-Formatted Data (Data in The Wrong Format)

If the data is stored in inconsistent formats, the systems used to analyze or store the information may not interpret it correctly. For example, if an organization is maintaining the database of their consumers, then the format for storing basic information should be pre-determined. Name (first name, last name), date of birth (US/UK style) or phone number (with or without country code) should be saved in the exact same format. It may take data scientists a considerable amount of time to simply unravel the many versions of data saved.

Text Encoding Artifacts

As data is moved and imported from system to system and written to files, text encoding can vary. This can cause strange, non-English symbols to appear amongst your data instead of familiar apostrophes and quotation marks.

Unstructured Data

Exactly like it sounds, data with no structure. No columns, no headings, it barely fits in a spreadsheet. Just a big file of text information. If you downloaded all of the comments ever made on Facebook and threw them into one file, that would be some unstructured data. Unfortunately there's no way to go about cleaning unstructured data before you've structured it. It's a whole other topic.

Dataset with Delimiter and Offset Issues

This type of data problem is practically invisible in a CSV file but looks absolutely horrible in the spreadsheet, but luckily it is not all that difficult to fix. It occurs when the delimiter structure of your input file is off. It can be cause by a single comma or semicolon in the wrong place. This misplaced delimiter will cause the spreadsheet to misinterpret the data structure. If you've ever loaded a file that looked fine for the first 500 rows, then suddenly everything was shifted by one column for the rest of the file and everything after it looked like gibberish, this error was likely the culprit.

Chapter 6: Things You Must Do *Before* Cleaning Data

Determine What Cleaning Your Data Needs

First, understand what kind of data you have. The easiest way to do this is to load it up into your favorite text editor or spreadsheet application. If the data is CSV or other delimited file, it should load fine into either tool. I prefer a spreadsheet like Excel or Google Sheets because it's slightly easier to view CSV data in columns and rows. (If it won't even load or loads funny, start with a basic text editor until you determine what is wrong.)

Data Profiling

The process of studying the data to determine what is wrong is known as *Data Profiling*, and it's important that you do it before any cleaning.

When I was a student, I had a teacher who constantly told us to the whole test before you start filling in answers. Of course, many of us scoffed at that. Then on our next quiz she fiendishly put in the 2nd to last question: "*Do not answer the first 18 questions on this test. Any mark, including erased marks will result in a zero score*." I wasn't the first to get to question #19, but I saw other students furiously erasing the entire sheet before giving up in frustration. I learned a lesson that day: Survey the whole pool, don't just jump in.

Data cleaning is no different. The number one mistake beginners make is that they don't profile data before jumping into cleaning, they spot an issue and right away start cleaning it, then they spot another, then another.

Then, after they're a few hours in, they realize that there are ***way*** more issues

than they realized, and some are systemic, and they've been creating new issues in the data, so they have to start over. Data Profiling can be as simple as looking over the data to more complex statistical tests to check for Uniformity. Don't skip it.

Get data into a Wide Format (aka Square Format)

This is just to get the data ready to work with. A wide format is pretty simple. It means that the data is composed entirely in rows and columns (doesn't matter how many, as long as it's roughly square shaped). Your columns should ideally contain headings (the title of the column). Do your best to put these in, they'll help you stay organized. You want to avoid gaps between rows or any extraneous information that might be in the spreadsheet.

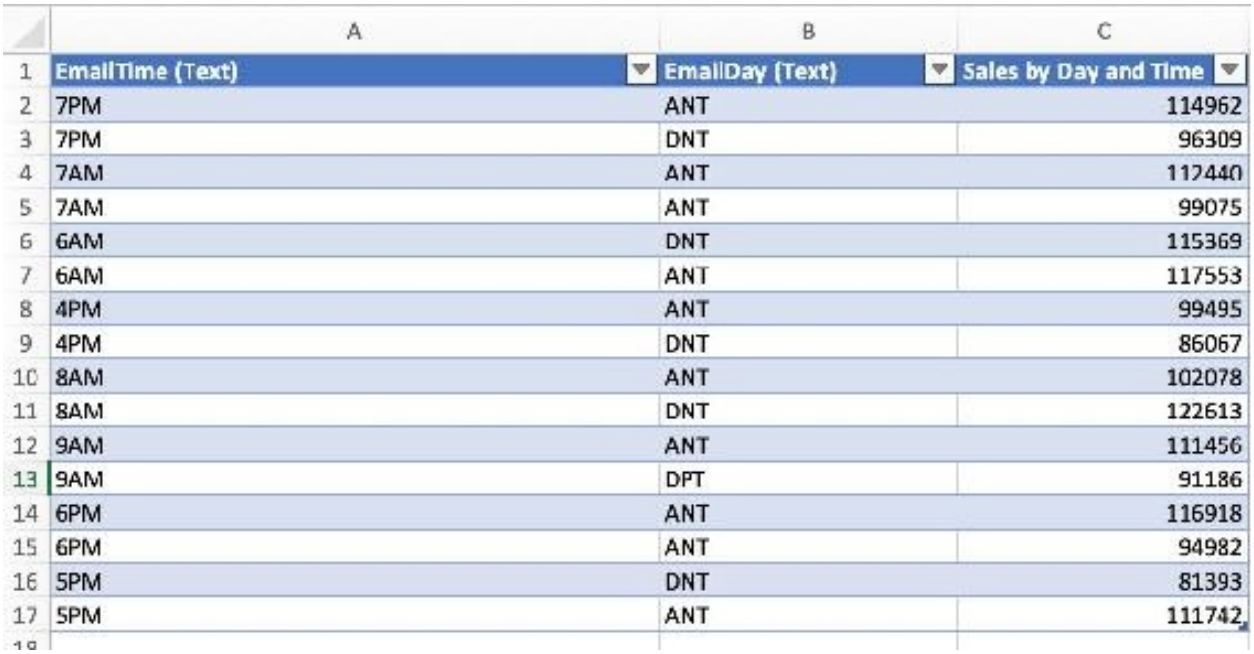
You can also take the opportunity to remove any extraneous data and/or formatting. I find it's easier to start out with unformatted data due to the conditional formatting we might apply later. It's also easier on the eyes.

This not only makes your work look neat and clean, it also gives you some virtual boundaries as you're working with the data that will prevent errors. Using the various functions provided in Excel and Google Sheets, it is possible to convert non-tabular data into tabular data.

Profiling Data with Tables

One of the quickest ways to profile data, whether it has headers or not, is to use the spreadsheet tool's Format as Table function. Excel's is right there in the ribbon, Google Sheets calls theirs "Filter Views" and it's under the 'Data' menu. Select all of the columns you wish to include as a table (which should be all of them). Clicking "Format as Table" will create a new table in the same spot

on your worksheet. The table will have column headings and allow you to perform additional functions on each individual column with a few clicks.



	A	B	C
1	EmailTime (Text)	EmailDay (Text)	Sales by Day and Time
2	7PM	ANT	114962
3	7PM	DNT	96309
4	7AM	ANT	117440
5	7AM	ANT	99075
6	6AM	DNT	115369
7	6AM	ANT	117553
8	4PM	ANT	99495
9	4PM	DNT	86067
10	8AM	ANT	102078
11	8AM	DNT	122613
12	9AM	ANT	111456
13	9AM	DPT	91186
14	6PM	ANT	116918
15	6PM	ANT	94982
16	5PM	DNT	81393
17	5PM	ANT	111742

The Benefits of Tables (Excel)/Filter Views (Sheets)

With a table you can:

- Sort the entire table by individual columns (while avoiding common sort errors)
- Filter the data displayed in the table by values inside columns
- Quickly view the range of values in each column to detect NULLs, Blanks, and outlier values.

Doing this across the dataset is great for getting a quick understanding of the data’s condition.

Getting Started with Tables for Data Profiling

Simply select the columns in the dataset you want to format as table (usually this is all of them) and click “Format as Table”. (In Google Sheets go to “Data”, select “Filter Views”)

A dialogue will pop up confirming the range you want to format, and there will be a checkbox “My table has headers”. If you leave this checked, Excel will use the 1st row in the dataset as the header values and not count it as part of the dataset. If your data does not have headers, uncheck this. But you recall what I said about headers right? Get some.

Now you’ll see the table created and Row 1 will contain your headers with some additional buttons. If your data did not have headers, Excel will still create a header row in Row 1 with generic header names (i.e., Column 1, Column 2, etc.)

In all cases, the first row will now contain some additional functions accessible when you click a cell in that row.

Continue exploring the data in the table. Use the value selectors and filters to understand how common different values are in your data. Scroll through the list of unique values to spot outliers and slight variances in the data. You can also spot duplicates fairly easily as the table will Alphabetize the discrete set of values so you can. Although it might be tempting to fix errors within the data at this point, this is purely an exploratory phase.

You can start a small scratchpad or separate worksheet to jot down the issues you’re seeing. Don’t worry about determining how widespread the issue is across the data, you’ll get to that later. For now, just focus on cataloging problems by type.

Using Tables is a great way to explore and get familiar with a dataset, since you can easily sort, filter and view the unique values in each column. It's one of the first things I do to a new data set first and really speeds up your exploration, especially with really large ones.

One major advantage of tables (over a standard worksheet) is that sort and filter operations will always affect the entire dataset, not just the selected column. This is a huge safety net in avoiding sort errors. It eliminates the possibility that you sort your dataset and forget to include a column. It also saves time not having to select the entire dataset every time.

Looking over the filtering options for every column is a great tactic for catching alternate spellings of words that might mean the same thing and makes it simple to fix these by filtering to only them.

For example, if you encounter a category column with only 3-4 possible values, but then see an alternate spelling of one of the values, you can zoom to it and correct it easily by copying the correct values over the bad ones.

Once you're done using the functions of the table, you have the option of leaving the data in Table format or removing the table format if you want to do other things to the data.

Of course, if you're planning on exporting your final dataset into some delimited or text format, you're going to want to remove the Table structure since it's a spreadsheet-only thing and not supported by TXT or CSV file formats.

Removing/Reversing Table

To remove the Table format, you would think you could click the table button again and select "Remove Table", but there is no such choice. Instead Excel wants you to select the table, then click the contextual "Table" choice at the top

of the ribbon, then select “Convert to Range”. This will remove the table features, but oddly, will leave the table formatting. It will also remove any filters and leave the data in the last sorted state

Chapter 6: Categorizing Different Data Quality Problems

One question PWCD[2] usually ask themselves at this point is “Do I need other data to clean up this dataset?”

PWCD, generally like to be able to clean up data without having to go searching for a whole lot of other values. I don’t think it’s because they’re lazy, (After all, these are people who I’ve seen write and compile a 100-line Python script just to avoid manually deleting commas in 20 fields).

No, I don’t think it’s that. I think it’s a desire for efficiency and accuracy. Chasing down other data can be time consuming, and sometimes impossible. Bringing other data into the fold also means inheriting any problems that new data has.

No sir, if I can infer or interpolate a field without getting out of this chair, damn sure I’m going to do it. (OK, maybe I am a little lazy).

It’s now time to divide and prioritize your data’s problems into two categories:

Type 1: Things you can *usually* fix from the original data file *(“I’ll have it done before lunch.”)*	Type 2: Problems that require additional information to fix *(“Probably not gonna happen.”)*
• Misformatted data	• Missing Data

• Text encoding artifacts • Delimiter and offset issues	• Null Values • Unstructured data
Problems that could fall into either category *("Uh. I'll need to get back to you.")*	
• Partial or incomplete values • Duplicates	

These are general guidelines, nothing with data is ever absolute. However, misformatted data, assuming the data is otherwise complete and you know how the correct formatting should look, can be easy to correct with a few keystrokes.

Example: Phone Numbers (555)-###-####

And

555-###-####

Both are complete and accurate phone numbers, they are just formatted differently. Simply pick one format and change all of values to conform to that format (don't worry-we'll show you how in a later chapter). This can be done without needing to pursue additional information, all that you need to fix it is within your source data.

However, if a customer name record has the LastName but is missing the FirstName, and you don't know their first name, there's really nothing you can do to fix that. You'll have to find somebody who knows the correct first name -

which can be a challenge indeed.

You might also need to make some informed decisions through this process about what is worth fixing and what is not. If your analysis plans for the data later on don't really rely on the First Name value, then great news: Your work is done!

With cleaning data there is always a trade-off between effort and results. Your goal should never be perfection, "good enough" is certainly good enough.

You may have noticed I slipped a 3rd category in there, problems that may or may not be fixable with just the source data. Well, it really does depend on your data and the nature of the problems.

Duplicate records are a perfect example. If every field in the duplicate record is the same value as the other duplicate record, meaning they are 100% identical, then you can usually delete one, keep the other, and move on with life.

However, what if the two records are identical except for one value, say email address, which is different.

Now, which one do you delete?

Without more information, you run the risk of deleting the "good" one. So again, time to go find someone who knows which email address is correct.

Now let's throw another angle in.

What if your records contain a "Last Updated" date stamp. One of the records was last updated 2 months ago, the other was last updated 6 years ago. Does that make it easier to decide? It sure does.

Partial or incomplete values are similar, it depends on how much data is missing.

Did someone just forget to type in a few letters in a State Name, so you have a value like: “Oklahom”

You’d probably be justified in changing that to “Oklahoma” and getting on with life.

However, if the missing data is not something you can logically interpolate or infer, then you need to go get more data.

There are two other techniques that you can use to fix Type 2 problems without obtaining additional data, if you’re lucky: ***Interpolation & Inference***. These are by no means going to be 100% accurate, but you can infer and sometimes interpolate missing values to produce a complete dataset.

Inference

Inference is based on logically inferring what the missing value would be based on other values in the record. Sometimes other values in the row can give clues as to what a missing value might be.

Example

If you know Zip Code, you can infer City and State with great accuracy. However, the inverse is not true, since Zip Codes are a subset of Cities and Cities are a subset of State.

Interpolation

Interpolation is most often used with numeric data and time series, and involves using mathematical formulas to compute or approximate the “most likely” value to put into a field. Again, sticking with the “good enough” doctrine, this is not going to be 100% perfect, but you can often get close enough.

Example

Let's say you have a dataset of daily temperature readings for six months. We know while there could be a wide range between the highest and lowest temperatures in this range, they usually don’t differ that much from day to day.

You could create a formula that takes the temperature from one day ago and one day to the future of the missing value, and averages them. You could make this even simpler by calculating the average temperature per week and using that value to fill in missing records during each week.

Your accuracy with interpolation depends greatly on the data and your knowledge of it. If you are not experienced with a dataset, try to speak with someone who is. They might be able to give you some rules you could use to simulate missing values.

End Goals for the Data

Finally determine what your end goals are for this data.

Will you keep it in a Spreadsheet for analysis?

OR

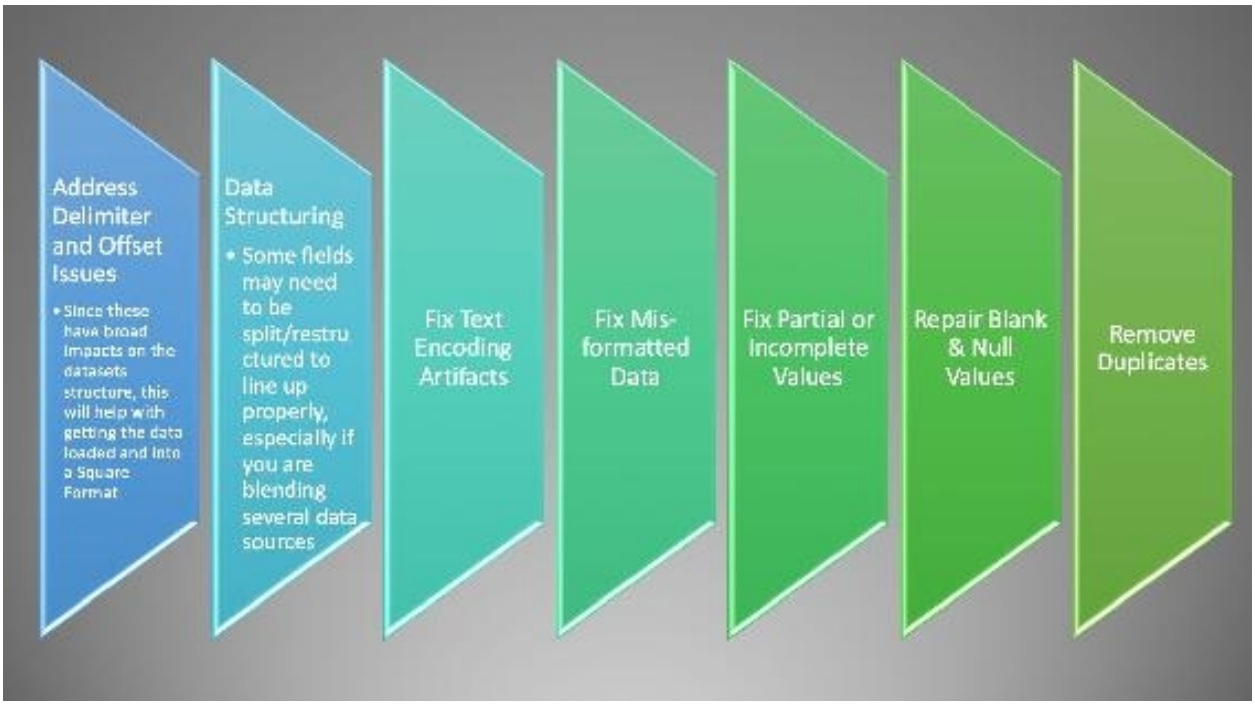
Are you planning to load it into another tool or program for analysis?

Knowing what you ultimately want to do with the data will help guide how you perform certain functions. With Excel, and Google Sheets, there are certain features and functions that you’ll want to avoid.

Chapter 7: Planning Your Data Cleaning Tasks - What's the Right Order?

Now that you've jotted down the various problems observed with the data, (and don't worry if you missed any - more tests are coming), it's time to decide what order in which to tackle them.

In general, the following order is preferred and provides the most efficient way to work. Some problems need to be fixed before others.



The above order is merely a guideline, your mileage may vary with your particular datasets. The step 'Remove Duplicates' is at the end because it is (usually) the only step that usually involves deleting records, and we might need

those records to perform other tasks.

However if your dataset contains many duplicates, you may wish to move that step up to avoid doing cleanup work on rows that will ultimately be thrown away.

Chapter 8: Cleaning Data: Getting Started

Now that you have your list of data issues identified and prioritized, you're ready to start the actual cleaning process.

What follows is a list of how-to techniques that address each of the Data Problems we discussed above. They are roughly in sequential order, but ultimately you will determine the correct order, depending on your task prioritization from earlier. You may wish to do some of these in Table[3] mode.

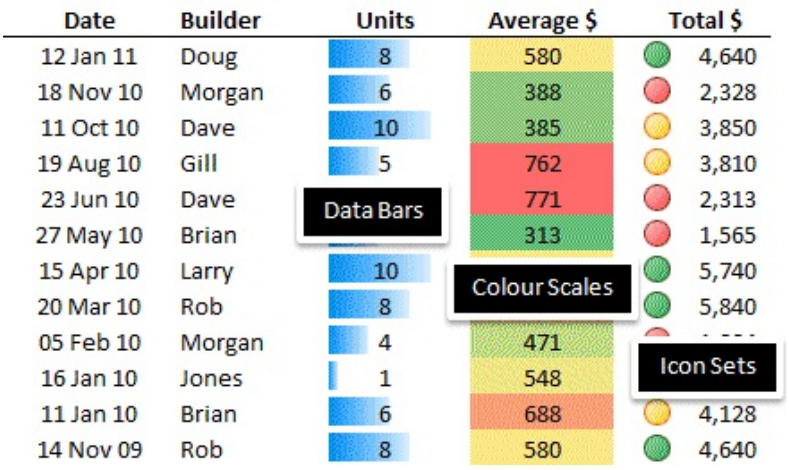
Also note, many techniques are multi-taskers and can be used to address multiple data problems, so you might come back to them again and again as you move through your process.

The Best Way to Deal with Data Errors: Highlight Them

If the data in your Excel sheet is full of errors, you can address them in two ways. You can use conditional formatting or use the "Go to special" to highlight with the errors.

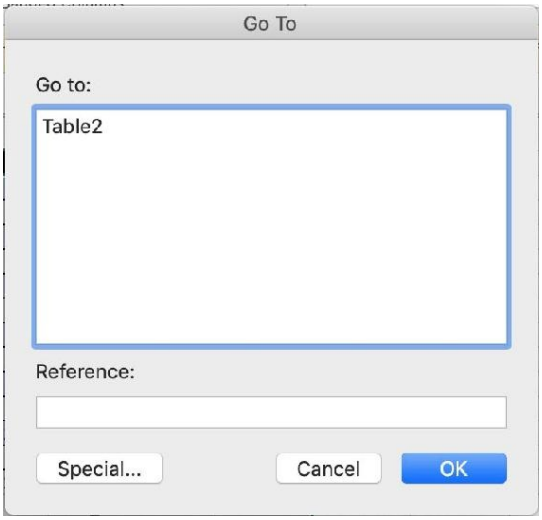
Conditional formatting – 1. Select the data set, 2. Go to home, 3. Select conditional formatting and select new rule, 4. Select "Format only cells that contain" in the new formatting rule dialogue box, 5. In the rule description, select errors from the drop down.

Next you have to select the format and click OK. This will make sure that the error is highlighted in the selected dataset.



Go to Special—Select the data set, Press F5, which will open the Go to Dialogue box. Next, click on special button at the bottom left. Select the formulas and uncheck all the options except for “Errors”. This will select the cells which have

errors. Now, you can manually highlight the errors and decide how to correct them.



Removing NonPrinting Characters

Certain parts of the text may contain leading, trailing, or multiple embedded space characters or nonprinting characters that can cause problems when you sort, filter, search, *etc.* These unnecessary characters make your text hard to understand and cause problems for the reader. It also makes your data look awkward, rather than elegant. You can make use of the trim, clean, and substitute functions to get rid of these unwanted and nonprinting characters.

You can trim by selecting the cells where you want to remove spaces and click TRIM. You can also clean text with **=CLEAN(text)** .

The Power of PASTE > VALUES

I’m going to mention the PASTE SPECIAL > VALUES feature both at the beginning and end of this section, because it has usefulness throughout your data cleaning process and especially at the end.

Paste Values will basically take whatever is displayed in a cell on the screen, and make that the new value. It’s very handy if you’ve used a function or formula to create a new field and you want to lock those new values. It’s especially helpful if you want to use a column of values based on formulas to re-sort your dataset and avoids circular reference errors.

To use it, just create a new empty column directly next to the one you want to clone, copy your source column and select Paste Special > Values from the menu. You are then free to delete the original column with the formulas and the source columns if you no longer need them. Now that new column will no longer change if the input values are change or are deleted.

Remember that except for the Undo feature, Paste Values is a one-way trip, there is no way to get back to the formulas from a column that contains values, so be sure you don’t need them. If you’re unsure, sometimes it’s easier to switch to a new worksheet, paste your values there as a new worksheet, that way you can always go back and tweak your formulas if needed.

Get Rid of Unnecessary Spacing

No one wants to look at a messy spreadsheet; a disorganized Excel or Google Sheet which looks like a complicated maze to navigate and makes collecting the data in it seem like a difficult chore.

Unwanted spaces make the presentation of data look erratic, so it’s best to remove them and adhere to the uniform pattern of spacing. Maintaining a uniform pattern of spacing, preferably leaving a row between lines of text, so that the data doesn’t look too crowded.

One way to get rid of the extra spaces is using the TRIM function. The EXCEL Trim function takes the cell reference as the input and removes leading and trailing spaces and the spaces between the words, if there are any.

Conversion from Text to Numbers

Sometimes there is a possibility that when you import your data from the text files to the external databases, the numbers might be stored as text and this could be an issue if you want use those cells for the calculations. There is also a possibility that some of the records contain an apostrophe before the number to make it a text value. This would also create an issue with the calculations. You can get rid of this problem and convert the numbers which are stored as text into numbers with this simple trick.

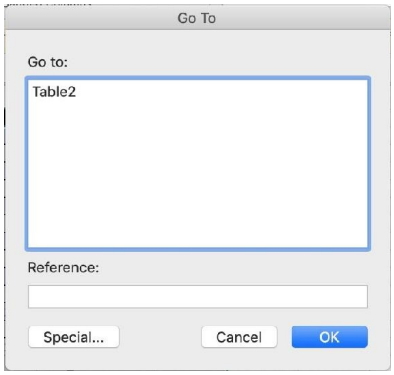
In any cell, which is blank, type 1, select the cell and press Control+C. Then select the cell which you want to convert to numbers. Select paste. For Excel, you can paste by using the keyboard shortcut Alt+E+S. Now, in the Paste Dialogue box, select multiply. Next click OK. Your numbers which are in the form of a text will be converted back to numbers.

How to Address Missing Data

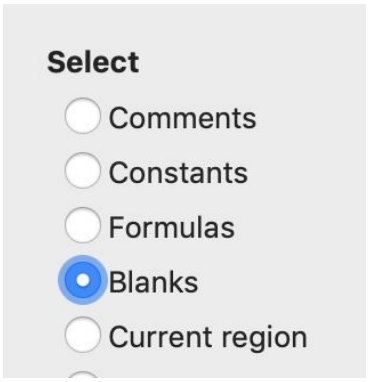
It is often said that prevention is better than cure. This is true in the case of blank cells. They can create a big problem if they do not get treated beforehand. You might face a problem with blank cells in a data set which is used to create reports or dashboards. If there is a small data set, you can fill the blank cells with 0 or Not Available or just highlight it.

But if you are working on a huge data set, doing it manually can take you hours. And for that, you can select and populate all the blank cells at once. This is how you can do it with Search Special.

Select the entire data set and press F5 which will open the "Go to dialogue box".



Next, click on special button which will open the go to special dialogue box. Go to the dialogue box, select blank and click OK. Again, go to the dialogue box and select blank, which will select all the blank cells in your data set. Now it is as simple as it can get. Enter "Not Available" or " 0" or anything you want to in these cells and press Control+Enter. And tada, you are done.


Select
Comments
Constants
Formulas
Blanks
Current region

Correct Use of Upper Case

A lot of users consider it unnecessary to follow the rules of capitalization in an Excel or Google Sheet. They couldn’t be more wrong. To ensure a professional and clean appearance to the reader, make sure the first letter of the words in headings and the words at the start of a new sentence of text are capitalized. Also ensure that the entire text of the headings is in bold. This makes it look like a certain degree of effort has been put into the creation of the spreadsheet, and the data wasn’t just haphazardly typed in and then forgotten.

The correct use of capitalization improves the presentation of your spreadsheet, and makes it look neater and more readable.

Unlike Microsoft Word, Excel doesn't have a Change Case button for changing capitalization. However, you can use the UPPER, LOWER, or PROPER functions to automatically change the case of existing text to uppercase, lowercase, or proper case. Functions are just built-in formulas that are designed to accomplish specific tasks—in this case, converting text case.

- UPPER will convert the target cell to all UPPER CASE
- LOWER will convert the target to all lower case.
- PROPER will convert the text to Proper Case (1st letter of each word is capitalized).

Merging Cells and Wrapping Text

When you're writing long phrases or sentences that tend to spill over the borders of the text box, all you need to do is select the number of cells you wish to contain the text inside and merge them. Then, select all the text, and use the option of "wrap text."

This will neatly contain the text within the boundaries set by you and will prevent them from spilling over borders. It makes your spreadsheet look much more organized and neat and makes the text easier for the reader to find in one place, rather than having to scroll to read the entire sentence.

So, select the number of tiles that you want to put the text in and merge them. Select the text and click "wrap text".

A great way to clean data quickly: Find and Replace:

Find and replace is indispensable when it comes to data cleansing. For example, you can select and remove all zeros, change references in formulas, find and change formatting, and so on.

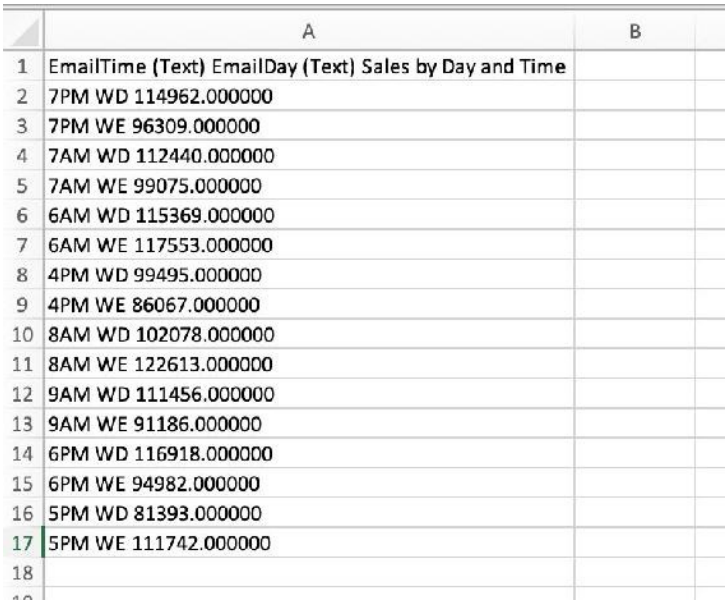
In an Excel sheet, finding a particular cell with a particular unit and changing it, especially when they are many, can be a herculean task.. If you want to remove all the 2's or all the 5's, you can Go to home, click find and select and click REPLACE by the shortcut Control+H. A find and replace dialogue box will open, where you have to register in the "Find what" and "Replace with" and then you click the replace all.

Find and Replace is very powerful. In fact Google Sheets supports Regular Expressions, which allow you to fine tune your formulas with wildcards. However, be sure to back up regularly and validate your results before moving on. There is no easy way to step backward.

Parse Data Using Text to Column

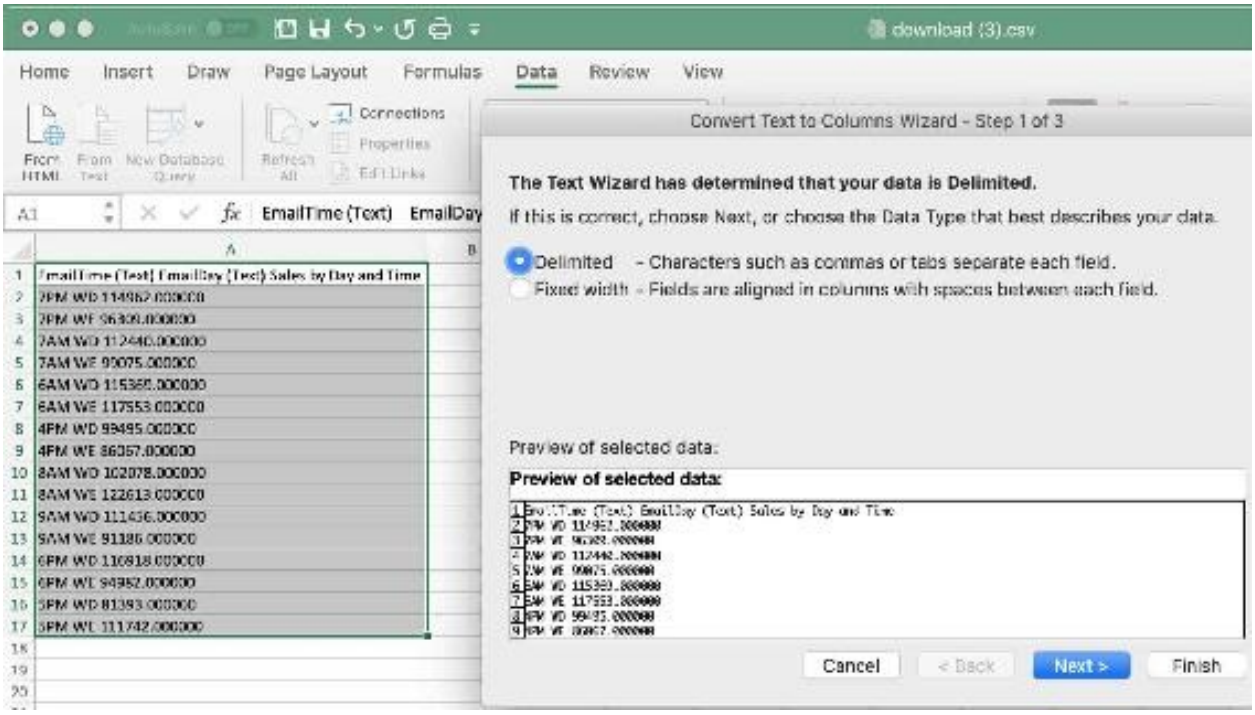
A very common issue is finding several distinct fields of data crammed into a single spreadsheet field. This makes it impossible to filter or properly organize your data.

To resolve this problem, you can parse and split this text into various cells by using the Text to Column function in Excel. To parse the data, go to Data and select Text to Column, which opens the text to column box.

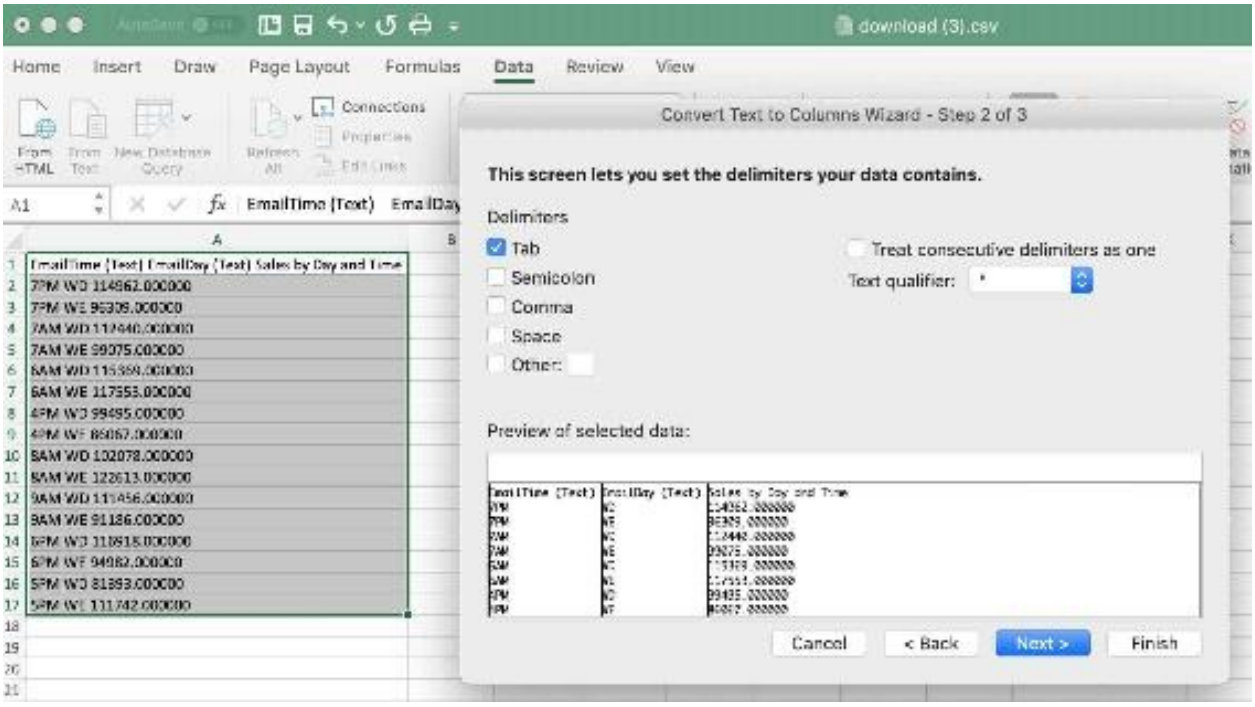


	A	B
1	EmailTime (Text) EmailDay (Text) Sales by Day and Time	
2	7PM WD 114962.000000	
3	7PM WE 96309.000000	
4	7AM WD 112440.000000	
5	7AM WE 99075.000000	
6	6AM WD 115369.000000	
7	6AM WE 117553.000000	
8	4PM WD 99495.000000	
9	4PM WE 86067.000000	
10	8AM WD 102078.000000	
11	8AM WE 122613.000000	
12	9AM WD 111456.000000	
13	9AM WE 91186.000000	
14	6PM WD 116918.000000	
15	6PM WE 94982.000000	
16	5PM WD 81393.000000	
17	5PM WE 111742.000000	
18		

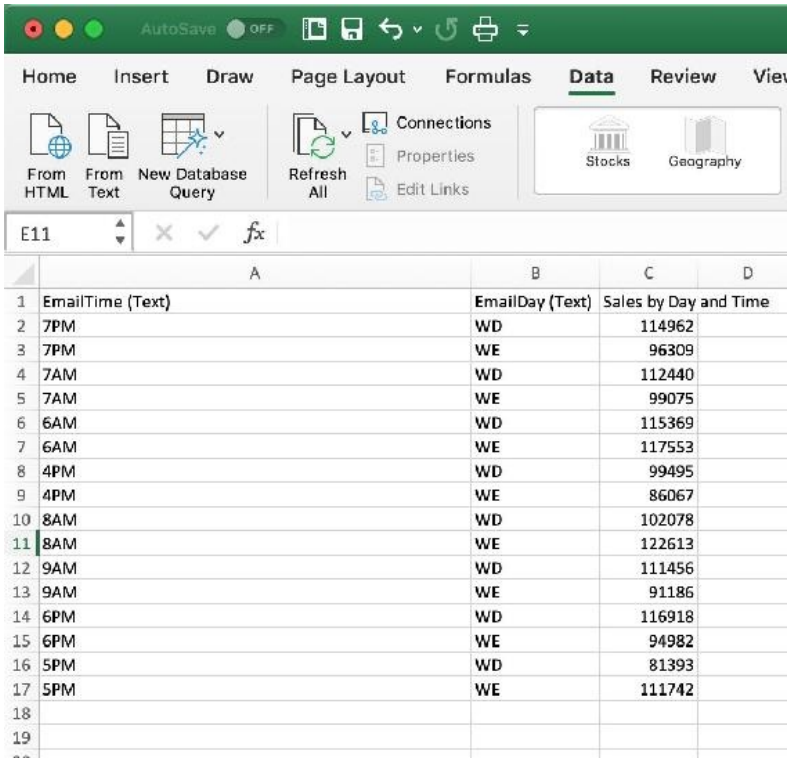
Starting dataset, multiple values jammed together in one column.



Next, select the data type and click next.

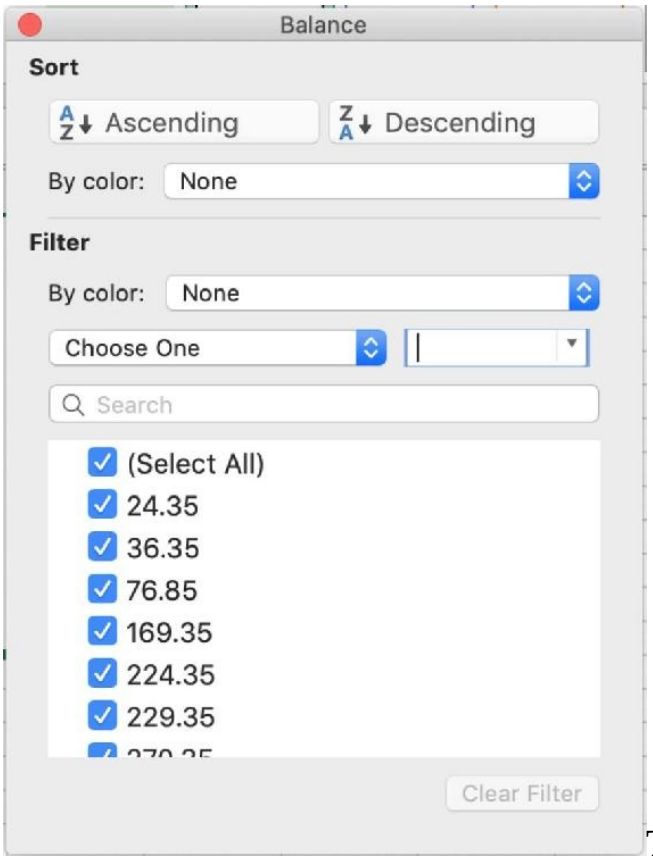


Next, select Delimiter, the character which separates your data. Now, select the data format and also the destination cell. Click finish and your work is done!



	A	B	C	D
1	EmailTime (Text)	EmailDay (Text)	Sales by Day and Time	
2	7PM	WD	114962	
3	7PM	WE	96309	
4	7AM	WD	112440	
5	7AM	WE	99075	
6	6AM	WD	115369	
7	6AM	WE	117553	
8	4PM	WD	99495	
9	4PM	WE	86067	
10	8AM	WD	102078	
11	8AM	WE	122613	
12	9AM	WD	111456	
13	9AM	WE	91186	
14	6PM	WD	116918	
15	6PM	WE	94982	
16	5PM	WD	81393	
17	5PM	WE	111742	
18				
19				

Check for non-standard values in the data



The quickest and easiest way to do this is with a Table/Filter View. Just go into the column and select the drop-down arrow. In the window that appears, all of the different discrete values in the column will show up, regardless of how many fields the column contains.

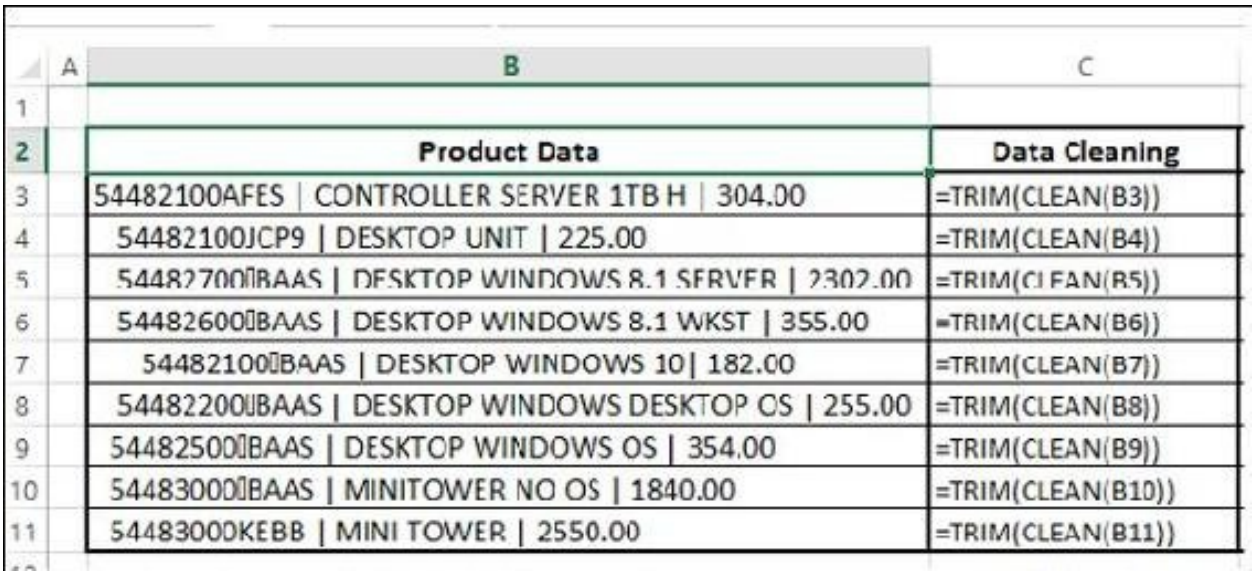
This is super helpful with Attribute columns where you're only expecting a few values to show up. It's a lot shorter and easier to scroll through to spot non-standard values.

If you have any, first click "Select All" to unselect everything, then select just the non-standard value(s) to filter and highlight them in your table.

When you click OK, you will see all the rows containing just the non-standard values in your data.

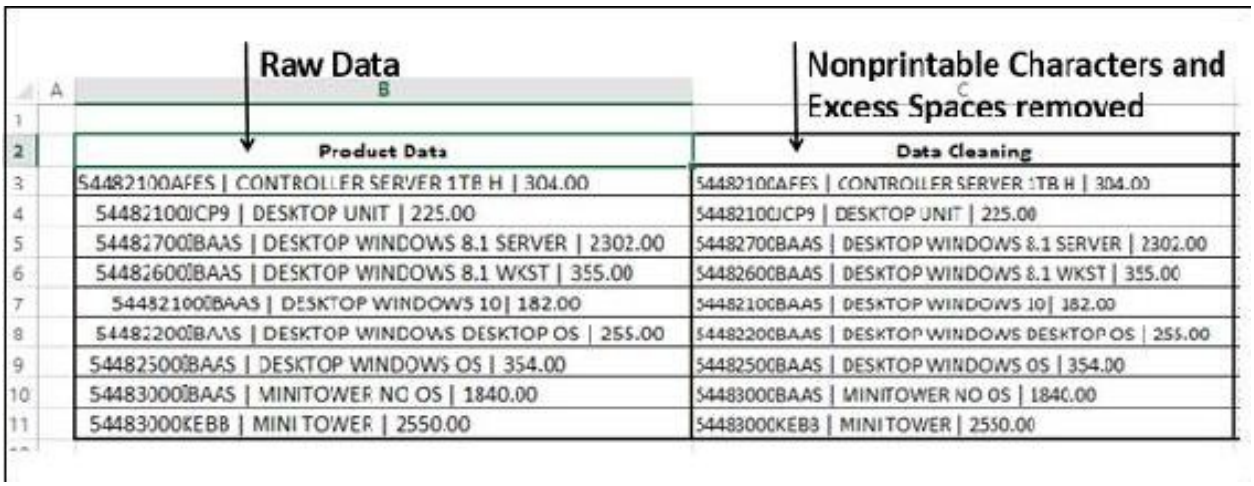
Use TRIM to remove spaces

- =TRIM: Remove extra spaces
- =PROPER: Makes first letter in each word uppercase
- =CLEAN: Removes all non-printable characters from text
- =VALUE: Converts text to number
- =TEXT: Converts number or text to new format



	A	B	C
1			
2		**Product Data**	**Data Cleaning**
3		54482100AFES | CONTROLLER SERVER 1TB H | 304.00	=TRIM(CLEAN(B3))
4		54482100JCP9 | DESKTOP UNIT | 225.00	=TRIM(CLEAN(B4))
5		54482700□BAAS | DESKTOP WINDOWS 8.1 SERVER | 2302.00	=TRIM(CLEAN(B5))
6		54482600□BAAS | DESKTOP WINDOWS 8.1 WKST | 355.00	=TRIM(CLEAN(B6))
7		54482100□BAAS | DESKTOP WINDOWS 10| 182.00	=TRIM(CLEAN(B7))
8		54482200□BAAS | DESKTOP WINDOWS DESKTOP OS | 255.00	=TRIM(CLEAN(B8))
9		54482500□BAAS | DESKTOP WINDOWS OS | 354.00	=TRIM(CLEAN(B9))
10		54483000□BAAS | MINITOWER NO OS | 1840.00	=TRIM(CLEAN(B10))
11		54483000KEBB | MINI TOWER | 2550.00	=TRIM(CLEAN(B11))

Result:



	Product Data	Data Cleaning
3	54482100AFES | CONTROLLER SERVER 1TB H | 304.00	54482100AFES | CONTROLLER SERVER 1TB H | 304.00
4	54482100JCP9 | DESKTOP UNIT | 225.00	54482100JCP9 | DESKTOP UNIT | 225.00
5	54482700BAAS | DESKTOP WINDOWS 8.1 SERVER | 2302.00	54482700BAAS | DESKTOP WINDOWS 8.1 SERVER | 2302.00
6	54482600BAAS | DESKTOP WINDOWS 8.1 WKST | 355.00	54482600BAAS | DESKTOP WINDOWS 8.1 WKST | 355.00
7	54482100BAAS | DESKTOP WINDOWS 10| 182.00	54482100BAAS | DESKTOP WINDOWS 10| 182.00
8	54482200BAAS | DESKTOP WINDOWS DESKTOP OS | 255.00	54482200BAAS | DESKTOP WINDOWS DESKTOP OS | 255.00
9	54482500BAAS | DESKTOP WINDOWS OS | 354.00	54482500BAAS | DESKTOP WINDOWS OS | 354.00
10	54483000BAAS | MINITOWER NO OS | 1840.00	54483000BAAS | MINITOWER NO OS | 1840.00
11	54483000KEB8 | MINI TOWER | 2550.00	54483000KEB3 | MINI TOWER | 2550.00

Use the @CONCATENATE and @SPLIT functions to restructure records

These two functions are super-handy if you have data that needs to be split or combined. @CONCATENATE will create a single column from multiple columns and values, you can even insert text in between the concatenated values. @SPLIT does - you guessed it - the exact opposite, it lets you define rules for splitting columns. These two are a situation where you probably want to create a new column and use Paste Values to paste just the result values of the new columns. That way you can delete the original columns.

Spell Check

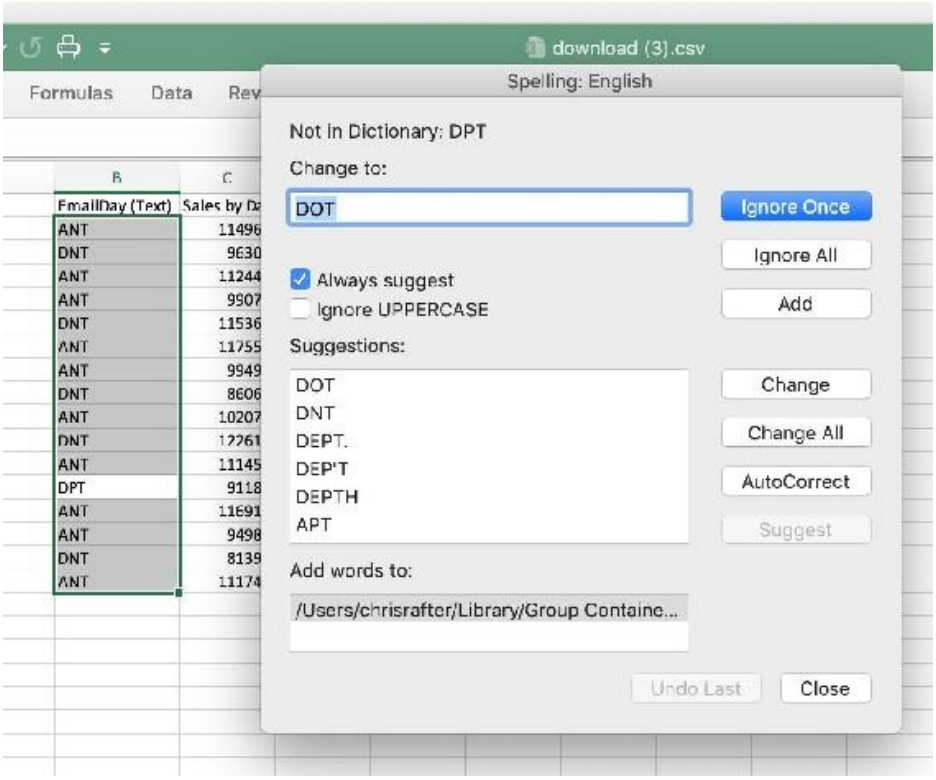
The worst way to present an Excel or Google Sheet is to have it filled with spelling errors. This not only makes your work appear hurried and sloppy, it is also a cause of major irritation and inconvenience for the reader. Realize that not all your data is going to benefit from a spell-check, so use it only on columns that contain words that would normally be found in a dictionary. However, if you regularly clean data with a lot of the same terms, you can also hack the spell check slightly by putting these terms into a Custom Dictionary in Excel. This will let you search through the data and find non-standard, “fuzzy” matches of the correct term, and hopefully correct them. This seems to work only with alphabetic values, not numbers.

To spell check, select the cell from where you want to start checking. Press F7 on your keyboard OR you can also click the spelling button on the review tab in the proofing group which will perform the check on the worksheet.

This technique is helpful if you regularly see the same data quality issues, and can be faster than Find and Replace.

Example

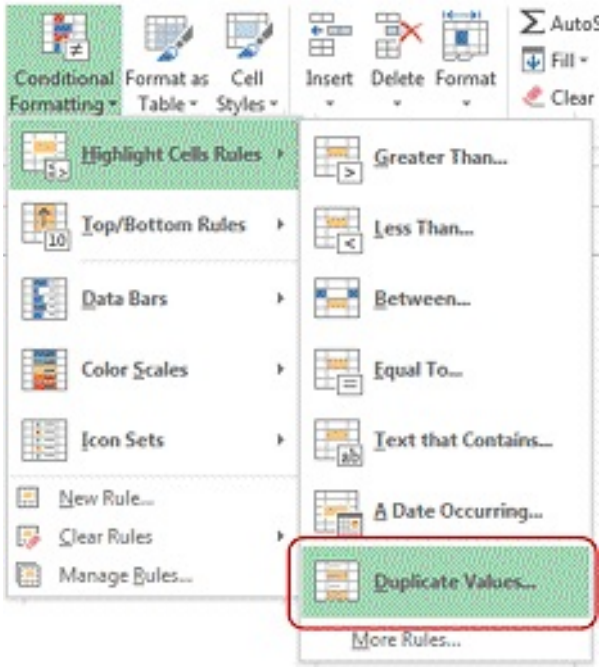
You have an attribute column with 2 acceptable correct values, “ANT” and “DNT”. If you add these terms to the Custom Dictionary, then run spell check on that data column, it will highlight any deviations and give you the option to change it to one of the acceptable values.


download (3).csv
Formulas
Data
Rev
B
C
EmailDay (Text)
Sales by Da
ANT 11496
DNT 9630
ANT 11244
ANT 9907
DNT 11536
ANT 11755
ANT 9949
DNT 8606
ANT 10207
DNT 12261
ANT 11145
DPT 9118
ANT 11691
ANT 9498
DNT 8139
ANT 11174
Spelling: English
Not in Dictionary: DPT
Change to:
DOT
Ignore Once
Ignore All
Always suggest
Ignore UPPERCASE
Add
Suggestions:
DOT
DNT
DEPT.
DEP'T
DEPTH
APT
Change
Change All
AutoCorrect
Suggest
Add words to:
/Users/chrisrafter/Library/Group Containe...
Undo Last
Close

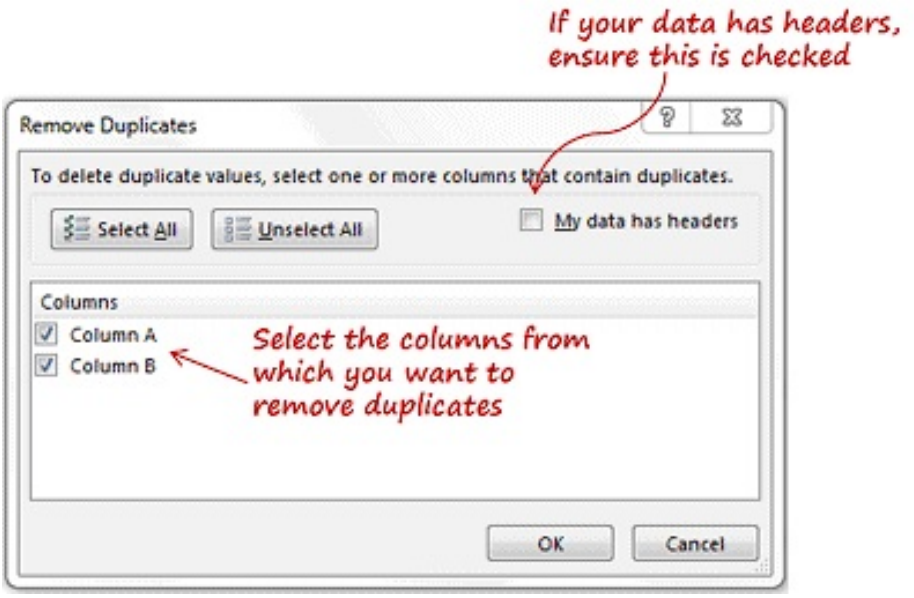
Getting Rid of Duplicates

Duplicates are never good; definitely not in your Excel. They'll inflate your counts and make just about any calculation inaccurate. You can do two things to remove the duplicates – ***Highlight It*** or ***Delete It.***

- Highlight Duplicate Data:
    - Select the data and Go to Home –> Conditional Formatting –> Highlight Cells Rules –> Duplicate Values.
    - Specify the formatting and all the duplicate values get highlighted.

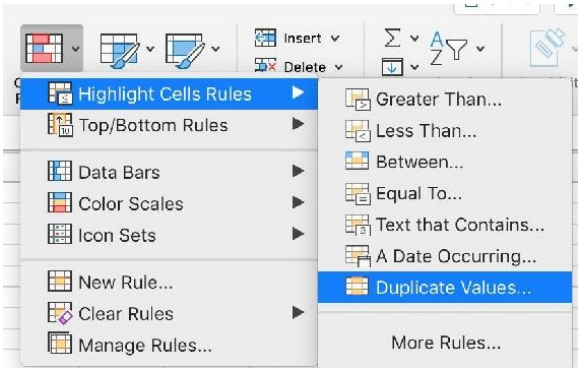


- Delete Duplicates in Data:
    - Select the data and Go to Data –> Remove Duplicates.
    - If your data has headers, ensure that the checkbox at the top right is checked.
    - Select the Column(s) from which you want to remove duplicates and click OK.



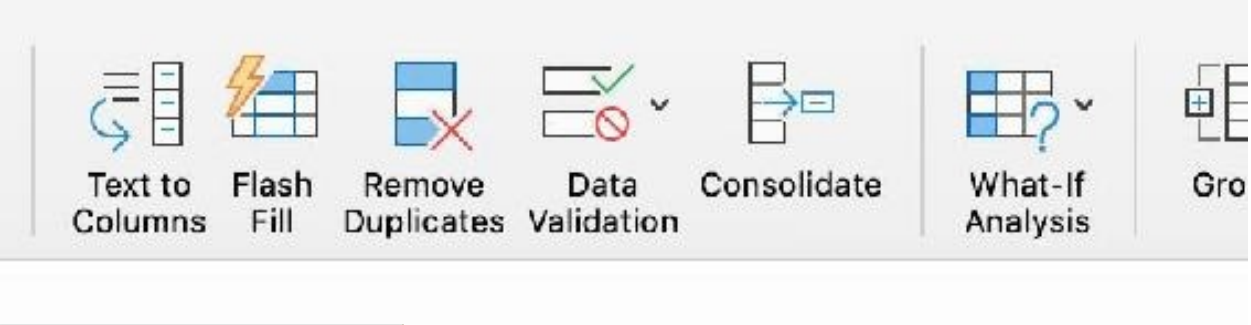
This removes duplicate values from the list. If you want the original list intact, copy-paste the data at some other location and then do this.

Select the data, click Home, go to conditional formatting, highlight cells rules and click duplicate values. Make sure that you specify the formatting and your work will be done.



You can also delete the duplicates in data. Simply, go to Data and then Remove Duplicates. But note, this will only delete duplicates where *every field in the data is identically duplicated.* In real world business data, most duplicate records are not 100% identical, so this might not get all the duplicates you would want to target.

In case of data with headers, the checkbox at the top right corner should be checked. Next, click the columns from where you wish to remove duplicates and click OK to delete the duplicates.



Prepare data for output with PASTE VALUES

Now, if you're truly finished and your data looks good, time to clean up all those formula cells and get data ready for output as raw values. You're going to use Paste Values one last time to create a completely new worksheet that contains only the values. That way you can still revisit formulas if you have to adjust something later, and you also keep a record of all your formulas if you need to repeat the cleaning process with a new dataset in the future.

Exporting Data from Spreadsheets to CSV, TXT Files

If you're going to analyze or load your data someplace else, you're going to want to use a more universal file format like Comma-Separated-Values, Pipe-Delimited, or plain ASCII Text.

Spreadsheet tools offer this option under the Save As… or Download As… feature. Do note that while you may have multiple sheets in your spreadsheet, only the active one can be saved in this manner, so make sure all of your target data is on the same sheet.

Chapter 9: Final Considerations

Cleaning data comes with a lot of complications and most people have a problem with data cleaning because frankly, it gets boring and monotonous after some time. However it is a critical function that can dramatically improve decision accuracy and business performance.

We know that the job of data analysts is especially tough. We hope the techniques in this book will help you on your journey toward quality data, and that you will progress swiftly.

Data cleaning isn't as difficult and frustrating as it might seem if you have the right training and strategy as to how to deal with data.

When you have the task of managing data, keeping on top of consistency and accuracy are two underlying jobs you have to deal with every day. These steps should help make it easier to create a daily protocol. Once you have completed your data cleaning process, you can confidently move forward using the data for deep, operational insights now that your data is accurate and reliable.

Appendix: A Few Expert Tips for Data Analysts

The First Step - Plan:

They say that the best way to do a particular job is to plan about it at the infancy stage. If you do not plan and get done with your work, that might be a possibility that you have to go back to the beginning and plan it again. There are many times when statisticians and data analysts jump to the final decision, even before having data, and that is where the problem starts. They sometimes even decide as to which statistical test will be used on the data, which is getting ahead of themselves.

Further, as a data analyst, you should consider the variables which are important, the interactions which should be interrogated and the statistical package which will be used for analysis. You, as a data analyst, should know that each package has its own advantages so that you can arrange the data accordingly from its infancy. Planning is not a small task. It involves everything from collecting your data to making sure that it is in the right format and is capable of answering your questions at the end.

You never know which direction you will take while making your research. It is better to do a pilot study for the data that you plan to fit into the categories that you have designed and for that, you need to collect some data. So again, you will have to go back and start from scratch. You again need to plan everything from the beginning until and unless you know how your study will progress. The only thing which matters is that if you have planned your data in a proper manner,

your outcome will be very satisfying. But you should still be prepared for a few surprises!

Collection of data:

The main thing you should keep in mind while doing collection of data is whether you will get the data from an outside source, or you will collect the data yourself. When you get the data from an outside source, and if the data is a little messy, there are not a lot of things that you can do about it, other than cleaning the data yourself. But when it comes to collecting the data yourself, you can decide on some of the standards before you get started and it will save you by going through a lot of pain in clearing the data.

In case you have a data set which is not big enough and can fit on a single Excel worksheet, then it is better for you to keep your entire data on a single worksheet. Consider that you enter your data across multiple worksheets, now you will have to go through the pain of making a few mistakes and then sorting your data can be much more painful.

Further, you might have to start with your data set again. You might also be asked to arrange the data so that each and every column is a single variable like height, weight, inside net measurement and much more or each row should correspond to a single sample like patient, test-tube, customer.

Hence, it is better to format your data in this manner from the starting. You will also need to sort your data by different columns and to restore the original order, hence, the best way to go about it is to create a unique ID column, numbered in consecutive integers.

Here's a little secret for you; your very own Excel also has a data entry form, which is built in and, we are sure, a lot of you might not know about it. You can use this to enter your data quickly and easily. Try this feature and you can thank us later.

Keep These Expert Tips In Mind

There are two different things, data cleaning and organizing it. Data cleaning can be done by anybody but cleaning the data quickly, efficiently and organizing it is an art. When you start with your data collection or data entry, it is better to organize your Excel workbook so that you can work on it, without losing your sanity. Don't worry, if you do not know how to do it, we will tell you a few simple steps.

Firstly create a worksheet for your data which is raw, another sheet for cleaning in progress, and the third sheet for the cleaned data, so that you can analyze it. There are a lot of advantages when you make 3 worksheets in your Excel, one of it being that you can always view your data in the stages, as and when they come. If by any chance you discover an error in the later worksheets, you can always trace it back, and figure out where the error was caused. In this way, you will not feel irritated and frustrated by the clutter on your Excel sheet.

The next part is data cleaning and we emphasize that you do not clean your data in the same worksheet where you store your data. Hence, there are other sheets which you will need in your Excel workbook which includes code sheet, note sheet, spreadsheets 1 2 3 4 and many more so that you can clean your data in different columns.

You can always use the Find and Replace feature and replace that single word in the entire workbook. It is also critical to remove the trailing and the spaces which can create havoc on your analysis, and for that Excel has a few formulae which includes trim, clean and substitute.

There is not any waiting time and you can remove all the non-printing characters from the entire workbook in just a minute, even if the workbook is small or big. You can also remove duplicates, find and replace, standardize the case of your

text data such as LOWER UPPER and PROPER.

A SCREENSHOT OF WORKBOOK WITH THREE WORKSHEETS

Code, Calculate and Convert your Data:

These are the very critical 3C's, which will help you to get your data set analysis ready. It is very likely that when you collected your data, you might have entered them as codes like [small, medium, large] or [1,2,3]. Hence, you should know what these codes mean and it is very necessary to make notes of the same. You should enter the code in a code worksheet and make sure that you explain the same in the notes worksheet.

Further, some data is collected and some is not. There is also a possibility that sometimes the data which has been collected has to be rounded, and sometimes it has to be kept in categories. For example, weight may be measured in kilograms or grams and rounded to the nearest decimal places. This can be done by ROUND, ROUNDUP and ROUNDDOWN. All of this basically depends on what you want to do with your data, how do you want to analyze it, so that you can perform calculations.

What you can do is that you can also create a new worksheet known as calculated data where you can create composite variables, such as the body mass index, convert the numerical variables into categories, and then round your data. As we have told you earlier that some of your data must also have been stored as text. Now you will have to convert it into integers and this is where you should learn how to use VLOOKUP and HLOOKUP.

Using this, your task will become very simple. VLOOKUP can be used very effectively in the following way: For example, you can have entries like grade 1 grade 5 grade 7 and you want to convert it into normal integers of 1, 5 and 7. In this case, you can use VLOOKUP and make your work quite easy.

A PICTURE OF CODE CALCULATE AND CONVERT

The Most Important Rule - The Integrity of Data:

Sometimes, there is a possibility that data may be entered wrongly. For instance, the age of a particular person while being 10 may be written as 100. Learn the formula to use MIN, MAX, AVERAGE, COUNT and for the text entries COUNTIF will help you to know how many entries of small medium or large you have in your variable. If you have an empty set you can use COUNTBLANK.

In case you want to identify statistical outliers in your numerical variables, you can always use the QUARTILE function of Excel. It will be better for you to use this function at an earlier stage so that you can decide what to do with this as it is the make and break deal. Now there can be an issue that there is a negative number in both the age as well as the height column, which is obviously an error and you need to sort it out. It is also possible that the minimum, mean and the maximum Heights and ages can be incorrect. So in these cases, you can use descriptive statistics which can quickly identify the errors before you move any further and start your analysis.

---

[1] “People with Crappy Data” is an alternate interpretation coined by some of my clients.

[2] “People Who Clean Data”, “People with Crappy Data”

[3] To remove Table Mode, go to “Table” at the top and select the button “Convert to Range”. This will remove all table filtering and reset your data to basic rows and columns.