MONITORING & EVALUATION RESOURCES
Bureau of Educational and Cultural Affairs
Evaluation Division

Data Cleaning for Substantive Analysis

Evaluation Division
BUREAU OF EDUCATIONAL AND CULTURAL AFFAIRS

The Bureau of Educational and Cultural Affairs' (ECA)'s Evaluation Division has been at the forefront of the Department of State's (DOS) monitoring and evaluation (M&E) efforts since 1999. Throughout its 20 years, the Evaluation Division has built a robust M&E system to ensure that ECA program staff and senior leadership benefit from timely performance data that can be utilized for evidence-based decision-making. The Evaluation Division's priority is to support the ECA Bureau's commitment to meeting and exceeding its programmatic goals by providing the data necessary to drive evidence-based decision-making throughout the Bureau.

For a complete listing of ongoing evaluation projects, an archive of completed reports, and resources for conducting evaluations, visit the ECA Evaluation Division website: https://eca.state.gov/impact/eca-evaluation-division

If you would like additional information or have any questions, please contact us at ECAevaluation@state.gov

Contents

INTRODUCTION

This guide is meant to serve as an introduction to survey data cleaning and analysis for individuals who are new to working with data or need a refresher. It provides tips and guidance on how to do basic data analysis of survey data. This guide is meant for those who:

1. Have limited experience with cleaning and analysis of survey data
2. Have less than 100 participants in the survey[1]
3. Are using Excel (or Google Sheets) as their main tool for cleaning and analysis

This guide is not meant to be exhaustive but rather provides the basic tools for the non-data scientist (or Monitoring and Evaluation (M&E) professional) to explore and understand data collected through a survey. However, it is not possible to provide examples that cover the breadth of programs or activities, nor is it exhaustive of the types of surveys and analysis available. Thus we encourage you to reach out to your Bureau of Educational and Cultural Affairs point of contact if you need further assistance.

IMPORTANT DEFINITIONS

Survey: a set of standardized questions (questionnaire) that are administered to selected individuals or groups of individuals. Surveys can be administered on paper, as an interview, online, or via telephone. You will choose the most appropriate type of survey depending on availability of respondents, security in the area, and/or time constraints.

Respondent: the individual taking the survey; the members of the group of people you would like information from, usually the beneficiaries of a program.

Data: (plural form of datum) data are generally understood to be pieces of information. Data can take many forms, including text, numbers, images, audio, and video. In this guide, we will be focusing on the first two types of data: text and numbers.

Data Point: a term for one unit or observation in a given data set (e.g. If 450 people take an evaluation survey for an exchange program, one person's response represents one data point/observation).

Data Set: a collection of separate pieces of information that are related to each other, for example the collection of responses to the questions on a survey from different individuals would be considered a data set. A data set is a collection of individual data points.

Variable: a variable is the opposite of a constant, and represents something that changes and can have more than one value. For example, in the mathematical equation $X^2=4$, the variable X can have more than one answer (2 or -2). In the example of the 450 who took the survey, a variable could be the age of the person who responded to the answer; some respondents might have the same age but more than likely, they will have wide variety of ages. There are different types of variables:

Quantitative variable: numerical variables such as counts, percent, or numbers. Generally, if you can add it, it's quantitative.

[1] The techniques in this guide can be used if you have more than 100 people who answered your survey, however, because most of the work is done manually, expect it to be fairly time consuming.

Qualitative variable: also called a **categorical variable** is a variable that describes data that are not numerical and fit into distinct, non-overlapping categories. For data analysis, qualitative variables are often assigned numbers, for example No = 0; Yes =1.

Continuing with the example of the exchange evaluation survey, the respondent's age is a quantitative variable but whether or not they spoke with an exchange alumni prior to applying (Yes/No) is a qualitative variable.

Outlier: observations that fall well outside of the general data (often quantified as any value more than three standard deviations away from the mean, or above the 99th percentile). For example, if the majority respondents range age 25-30 but you have a respondent who is aged 60, that respondent would be considered an outlier.

Illogical Value: a set of entries in a survey submission that seem contradictory (e.g. a respondent who reports that they are both male and pregnant, that they agree with both a statement and its opposite, etc.) These are to be reconciled or eliminated from analysis (see page 22 for tips on dealing with illogical values).

Typo: a mistaken entry or observation (e.g. on a scale of 1-10, the subject selects 100)

Sampling: the process by which a portion of a given population is selected for data collection and measurement. Sampling is borne both out of necessity and practicality. For example, as it is logistically infeasible to measure the political opinions of everyone in the United States, a representative sample of 1500 or more people could suffice as an indicator of trends in the general population.

Construct: a theoretical idea attempted to be measured or quantified (e.g. program satisfaction, skill learned, expertise in a subject area, envy)

Operational definition of a variable is the specific way in which it is measured in the survey. Different surveys may measure the same construct differently. For example, if asking for length of time at current job, the operational variable could be measured as a range (less than one year, 1 – 3 years, etc.) or in years and months (1 year and 3 months).

A Statistics Refresher

Mean: the arithmetic average of a given set of observations – Note: can be biased by outliers, for example if in your survey, you have 300 respondents who are aged 25 – 35 but you have 50 people who are 45 and above, the average age will be much higher as a result.

Median: the "middle" value in a given set of observations. To find the median, your numbers have to be listed in numerical order from smallest to largest, so that you may have to re-order your data and select the value in the "middle."

Variance: the degree to which data points/observation cluster around a mean; a high variance data set is distributed widely, varying greatly from the mean, while a low variance data set contains values that are all relatively close to the mean value.

Standard Deviation: the square root of the total variance in a data set, used as a standard by which to compare individual points/observations to the mean. The interval of three standard deviations in either direction from the mean in a normal distribution contains 99.7% of all observations (e.g. On an IQ test, the mean is generally thought to be 100 with a standard deviation of 15, so 99.7% of people would have an IQ between 55 and 145)

To put it all together:

Your dataset is the following list of values	13, 18, 13, 14, 13, 16, 14, 21, 13
The mean is	(13+18+13+14+13+16+14+21+13)/9=**15**
The median is the middle value from the ordered list of values	13, 13, 13, 13, **14**, 14, 16, 18, 21
The variance is	5.901
The standard deviation is	√5.901= **2.429**

What is data cleaning and why is it important?

Data cleaning is the process of identifying incomplete, inaccurate and/or irrelevant parts of a dataset and making corrections by modifying, replacing or deleting the ‘dirty’ data. Data analysis should take place only after the data have been cleaned as it is important that the results of your analysis be built on accurate and strong data. Data cleaning also has profound impacts on both the design process and analysis. Validating and cleaning data can inform the survey construction process, allowing for clearer and more concise designs in the future; likewise, analysis conducted on carefully cleaned data leads to clearer conclusions. Just remember that the quality of data analysis will never be better than the quality of data cleaning; data cleaning is a vital activity for your M&E efforts.

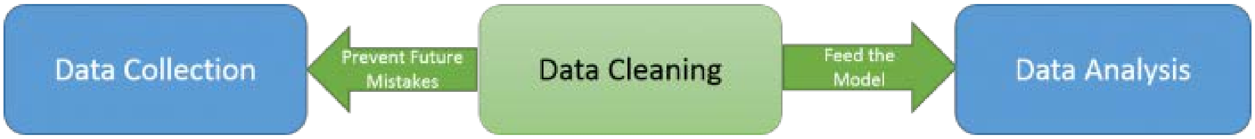


Figure 1: Mental Model for Data Cleaning

A central goal of working with data is maintaining its integrity and not altering it without intention. Enter data cleaning: by following clear guidelines and protocol, data can be effectively prepared for data analysis – ensuring that your measurements are meaningful. This document is designed to provide guidelines for this data cleaning process from planning and collection to analysis, and to help avoid simple errors when working with data. The following tips and advice assume that your survey data will be cleaned and analyzed in Excel (or Google Sheets). Note that Excel is not a statistical package, however it is most often the software used for data analysis.

If you are using an online survey tool (Google Forms, SurveyGizmo, SurveyMonkey, etc.) that delivers your survey data into Excel or Google Sheets, you can skip Step 0 and go straight to Step 1 in data cleaning and analysis.

STEP 0: ENTERING AND STORING YOUR DATA IN EXCEL

It is important to have an efficient central method of data storage and intake. While most online survey tools (such as SurveyMonkey, SurveyGizmo, Google Forms, etc.) allow for centralized storage, the storage process should be considered before the survey is designed. There should be a clear and direct path from survey distribution to a single file containing all collected data.

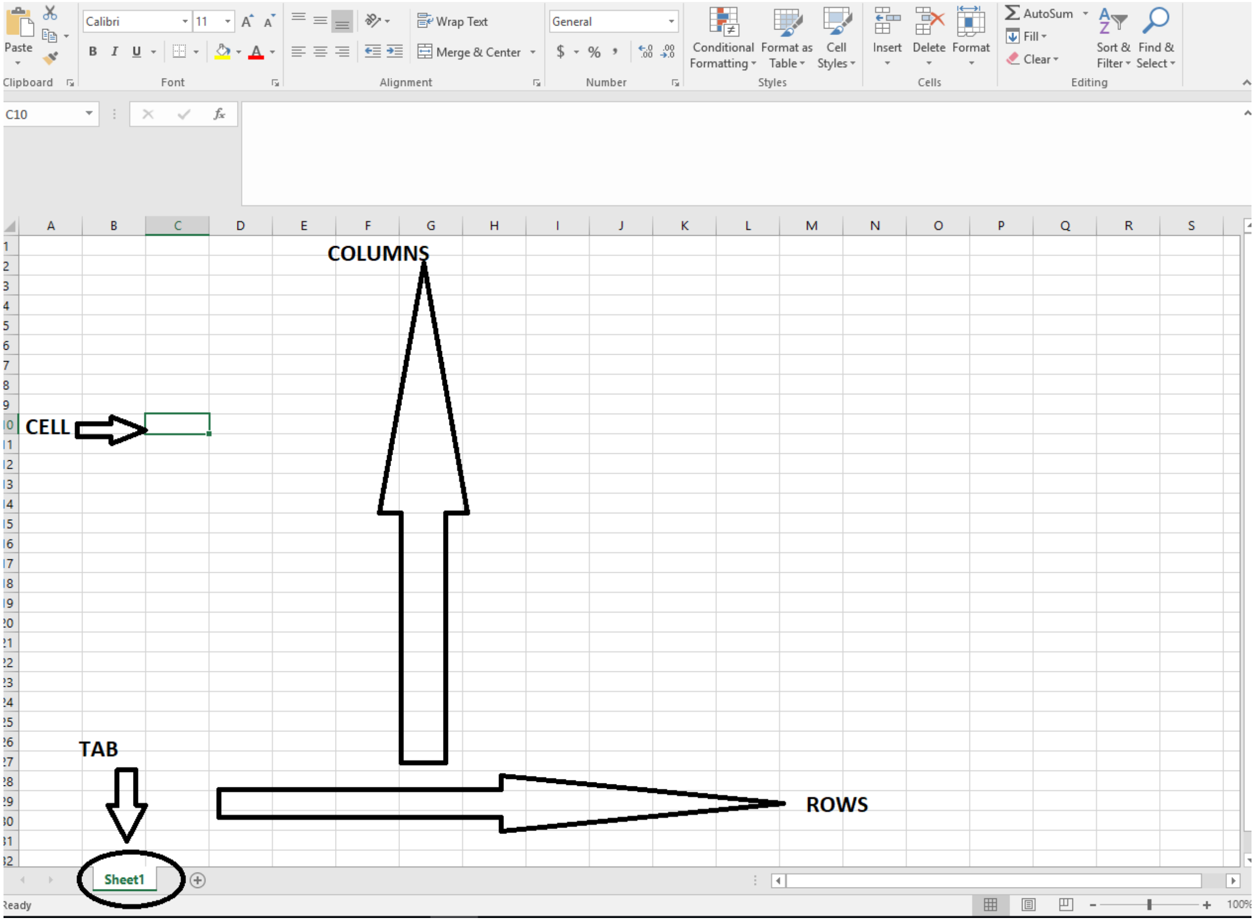
The good news is that since you are hand-entering your data into Excel, you can enter your data into the right format the first time and eliminate some of the work in the later steps. The not so good news is that manual data entry (besides being time-consuming) is very prone to error, so after data entry is complete, you should spot check a sample of the entries for errors.

A Brief Introduction To Microsoft Excel

When you open up Microsoft Excel, you will see a blank **worksheet.** This worksheet is part of the workbook. A **workbook** holds all of your worksheets and is simply another name for an Excel file. A blank Excel worksheet is composed of a series of vertical columns, horizontal rows, and individual cells (see Figure 1). You can select different worksheets by clicking on the tabs at the bottom of your workbook.

- **Columns** are alphabetized – A, B, C, D… - from left to right across the top
- **Rows** are numbered – 1,2,3,4… - from top to bottom down the left of the worksheet

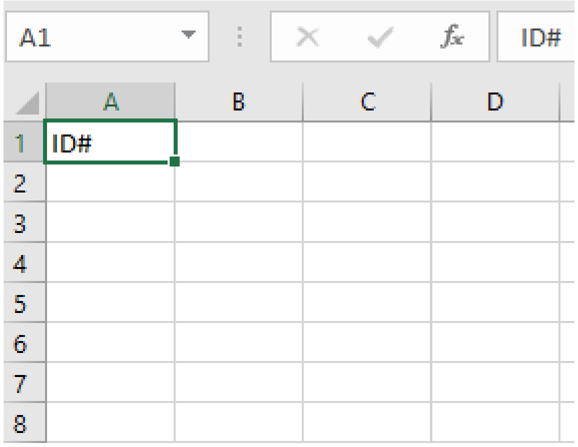
- **Cells** are individual boxes within the worksheet
- **Tabs** are at the bottom of the page and allow you to select different worksheets



Creating Colum Headers

The first column is for the ID number of each completed questionnaire. This is called a unique identifier[2]. Type the header “ID#” into the first cell in Cell A1 (Column A, Row 1) as in the figure below:

[2] A unique identifier is an assigned number that identifies each questionnaire. When assigning unique identifiers, you can use sequential numbers such as 001, 002, 003, etc. Write this number on the corner of each paper survey questionnaire, and enter that same number in the column labeled ID# followed by the data for that questionnaire. Later, this will allow you to find a particular questionnaire or remove identifying elements from your database.

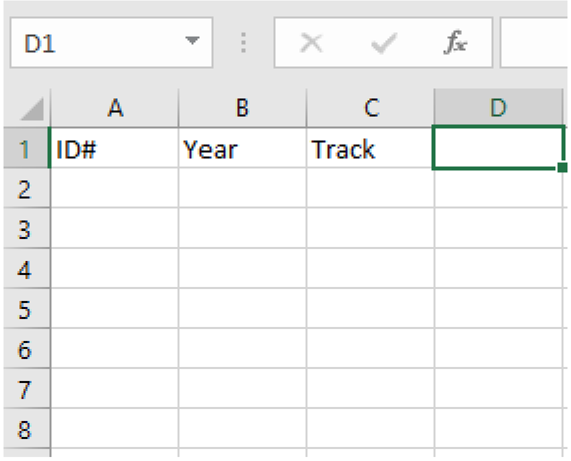


	A	B	C	D
1	ID#			
2				
3				
4				
5				
6				
7				
8				

Next, create column headers for each of the survey questions. Decide which kind of header will work better for you:

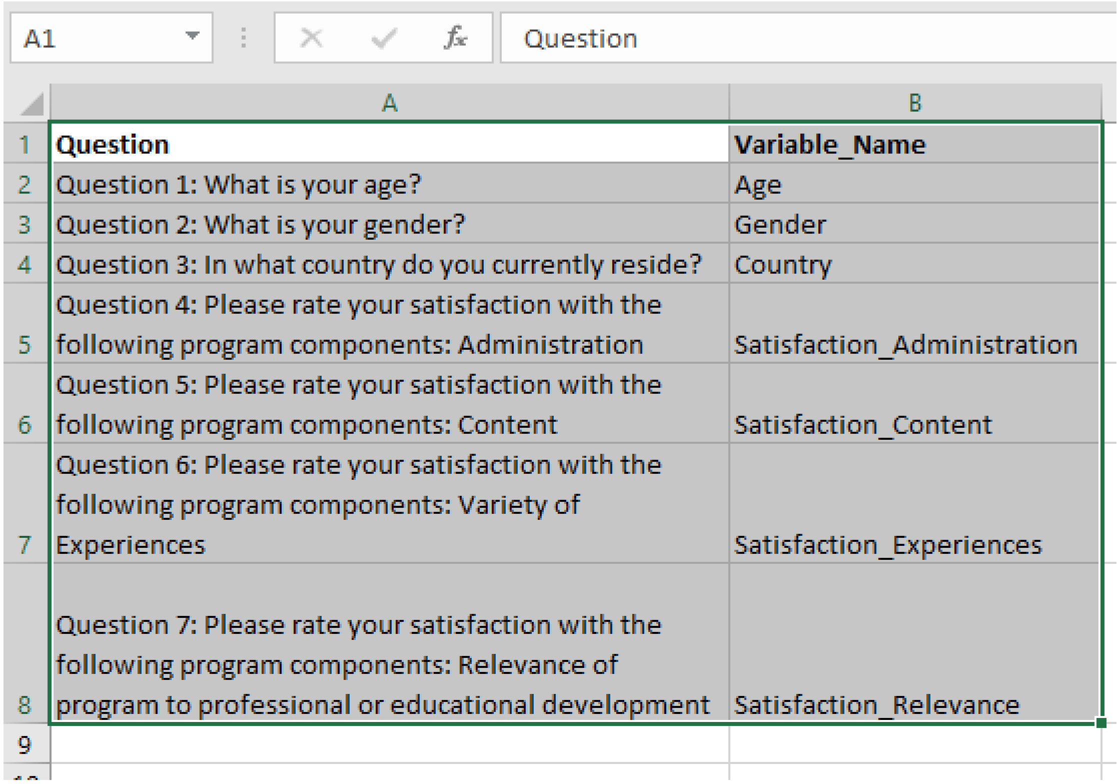
- Label the columns using **the number of each question** – Q1, Q2, Q3, Q4
- Use a **descriptive header** that encapsulates each question's meaning – for instance, if a question asks "Did you graduate from college?" the column header could be "College"
- You can also use a combination of the above two, for example: "Q2 – College"

Continue creating column headers until all questions have been labeled. Each header is entered into a separate column.



	A	B	C	D
1	ID#	Year	Track	
2				
3				
4				
5				
6				
7				
8				

Keep track of the header you give to each question. A good way to do this is to take a blank worksheet in a separate tab with two separate columns: one for the full printed question and in a separate question next to it, the header. This will be your **codebook.**



	A	B
1	**Question**	**Variable_Name**
2	Question 1: What is your age?	Age
3	Question 2: What is your gender?	Gender
4	Question 3: In what country do you currently reside?	Country
5	Question 4: Please rate your satisfaction with the following program components: Administration	Satisfaction_Administration
6	Question 5: Please rate your satisfaction with the following program components: Content	Satisfaction_Content
7	Question 6: Please rate your satisfaction with the following program components: Variety of Experiences	Satisfaction_Experiences
8	Question 7: Please rate your satisfaction with the following program components: Relevance of program to professional or educational development	Satisfaction_Relevance
9		

When entering your data, you may encounter some of the following issues:

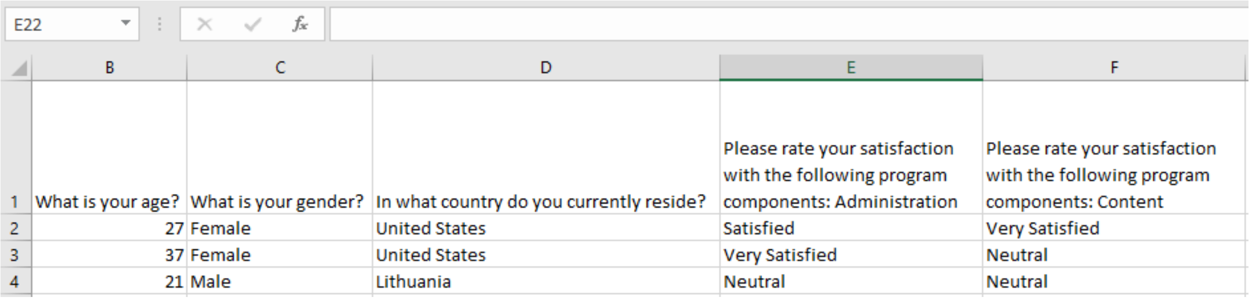
- An unanswered question: leave the cell for the unanswered question blank
- If two or more responses are selected for a question when only one is requested: treat this question as if it were not answered
- If an open-ended question has an incomplete answer: still enter the data that are given

Phew! Now that you've hand entered all the responses into the Excel spreadsheet, take a small number of questionnaires (ideally randomly selected) and go through the paper surveys and compare it with the data in the Excel spreadsheet to make sure the two documents match.

STEP 1: LOOK AT YOUR DATA

Now you have an excel sheet full of rows and rows of data (whether you've hand entered or had it pre-populated by a survey tool), the first step is taking a look at your data. Open your Excel spreadsheet and look at the data. Row 1 of your data should have the names of your variables (or corresponding questions)[3], for example "Age/Question 1: What is your age?" with each row corresponding to one person's response. Your data should look something like this:

[3] Some survey tools such as Survey Monkey will export two rows of variable names if there are matrix questions in the survey. An example of a matrix question would be to ask participants to rate their satisfaction level with different aspects of a training. In the case of matrix questions, you may have to do some relabeling to condense variable names into one row.



	B	C	D	E	F
1	What is your age?	What is your gender?	In what country do you currently reside?	Please rate your satisfaction with the following program components: Administration	Please rate your satisfaction with the following program components: Content
2	27	Female	United States	Satisfied	Very Satisfied
3	37	Female	United States	Very Satisfied	Neutral
4	21	Male	Lithuania	Neutral	Neutral

Now, look at the first five rows of data, the bottom five rows and five random rows. Ask yourself whether the data you are looking at make sense? Are there names in the names columns, numbers in the age column, or is there different data in the column? If there seems to be a misalignment of the values with the column name/title (for example, you see a list of cities in a column labeled State), and you hand-entered the data, you may want to go back to Step 0 and re-enter the data. Otherwise, make note of any anomalies; these will be addressed further on in the cleaning process.

STEP 2: CLEAN YOUR DATA

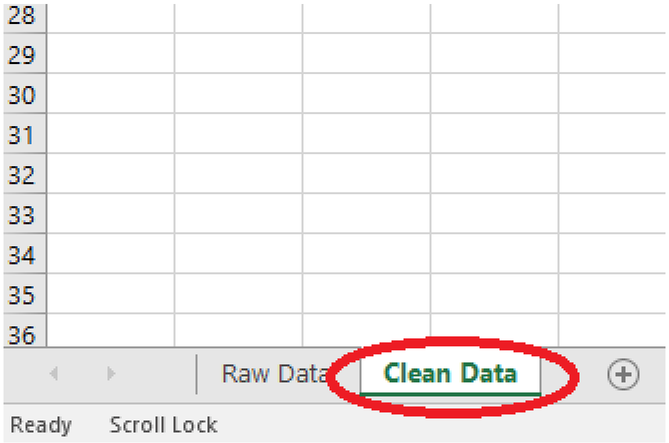
No data are perfect (if they are, then you should be very worried about tampering), so as a first step when cleaning your data, you will need to identify all the issues/mistakes/errors in your data. Some issues/mistakes/errors such as typos can easily be corrected, others may so extensive that you will need to remove these data entirely. The following first walks you through how to identify different types of mistakes/errors in your data, then as a second step, shows you how to gauge the extent and severity of these mistakes/errors in your data, and finally how to correct them.

When data is first collected and entered into a particular format (normally Excel, .csv, Google Sheet), it is known as **"raw"** data before it is properly cleaned and formatted. These raw data can sometimes include typos, illogical values, duplicate entries, and other undesirable data points in no particular format. **It is important that a complete copy of the original raw data be saved during the data cleaning and analysis step for reference, so if you make a mistake in later steps you can backtrack**. Maintaining a copy of the raw data also allows others to go back and replicate and/or do additional analyses.

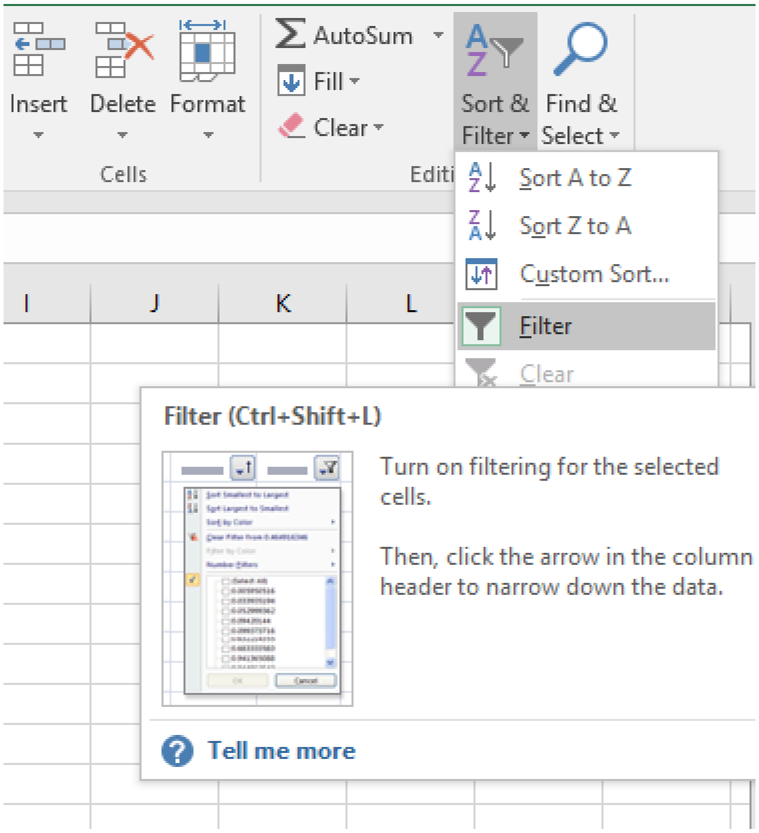
Note: if your data came from a survey tool such as Google Forms, SurveyGizmo, etc. (where the questions were checkboxes, etc.) you may not need to do as much cleaning as if the data were hand-entered from a paper survey. However, you should pay special attention to questions where the values were typed in by the survey respondents (for example: age).

To begin the process of cleaning your data, follow the below steps:

1. Rename the worksheet with all the raw data "Raw Data."
2. Select all of the spreadsheet of the raw data by using "CTRL+A", copy and then paste into a new blank worksheet. Name that sheet "Clean Data"



3. Now add a filter to your data, you do this by selecting the first row (the one with the column headers) and clicking and the ‘Sort & Filter’ button the in upper right hand corner

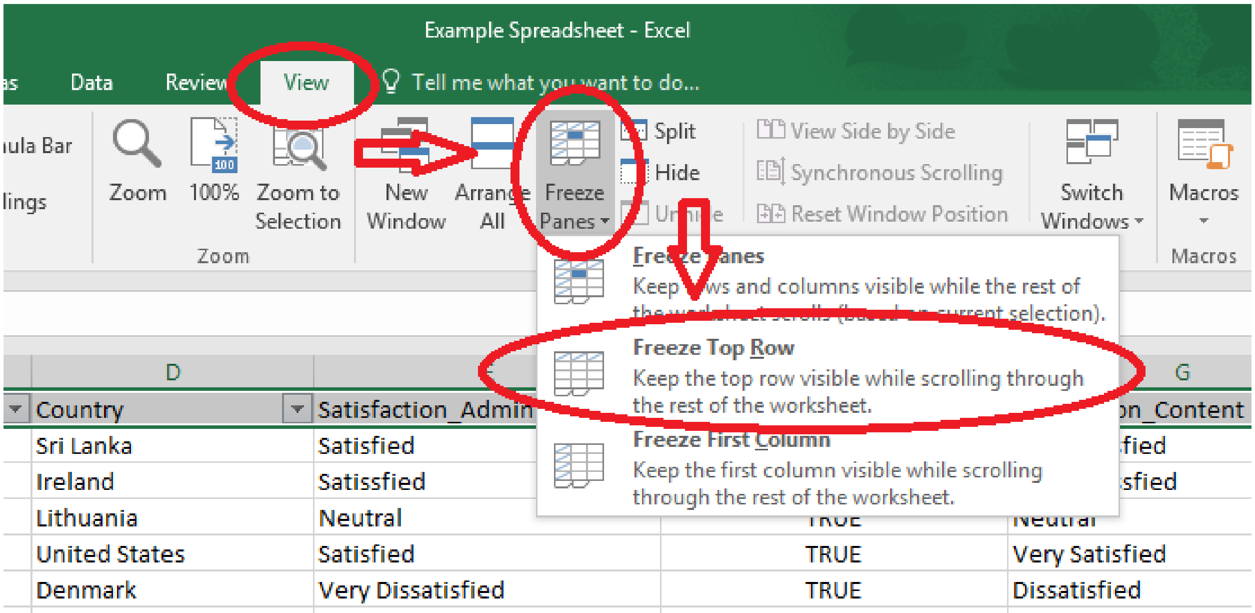


Turn on filtering for the selected cells.

Then, click the arrow in the column header to narrow down the data.

The next step is going to be to “Freeze” the first row with all the column headers/variable names are so that when you scroll through your data, you are able to always see the column headers/variable names. Follow the next 4 steps:

1. Select the first row (Row A)
2. Go to ‘View’
3. Click on “Freeze Panes”
4. Select “Freeze Top Row”



Freeze Panes
Keep rows and columns visible while the rest of the worksheet scrolls (based on current selection).

Freeze Top Row
Keep the top row visible while scrolling through the rest of the worksheet.

Freeze First Column
Keep the first column visible while scrolling through the rest of the worksheet.

Country	Satisfaction_Admin		on_Content
Sri Lanka	Satisfied		fied
Ireland	Satissfied		sfied
Lithuania	Neutral	TRUE	Neutral
United States	Satisfied	TRUE	Very Satisfied
Denmark	Very Dissatisfied	TRUE	Dissatisfied

\
As mentioned above, raw data often contains all sorts of ‘errors’ and/or ‘mistakes’ so the following next steps will outline what these issues can be and how to fix them. As you go through the process of cleaning your data, you will want to make sure you update your codebook (review Step 0 for reference on what a codebook is) with any new columns, changes, etc. that you make to the data.

2.1 Typos

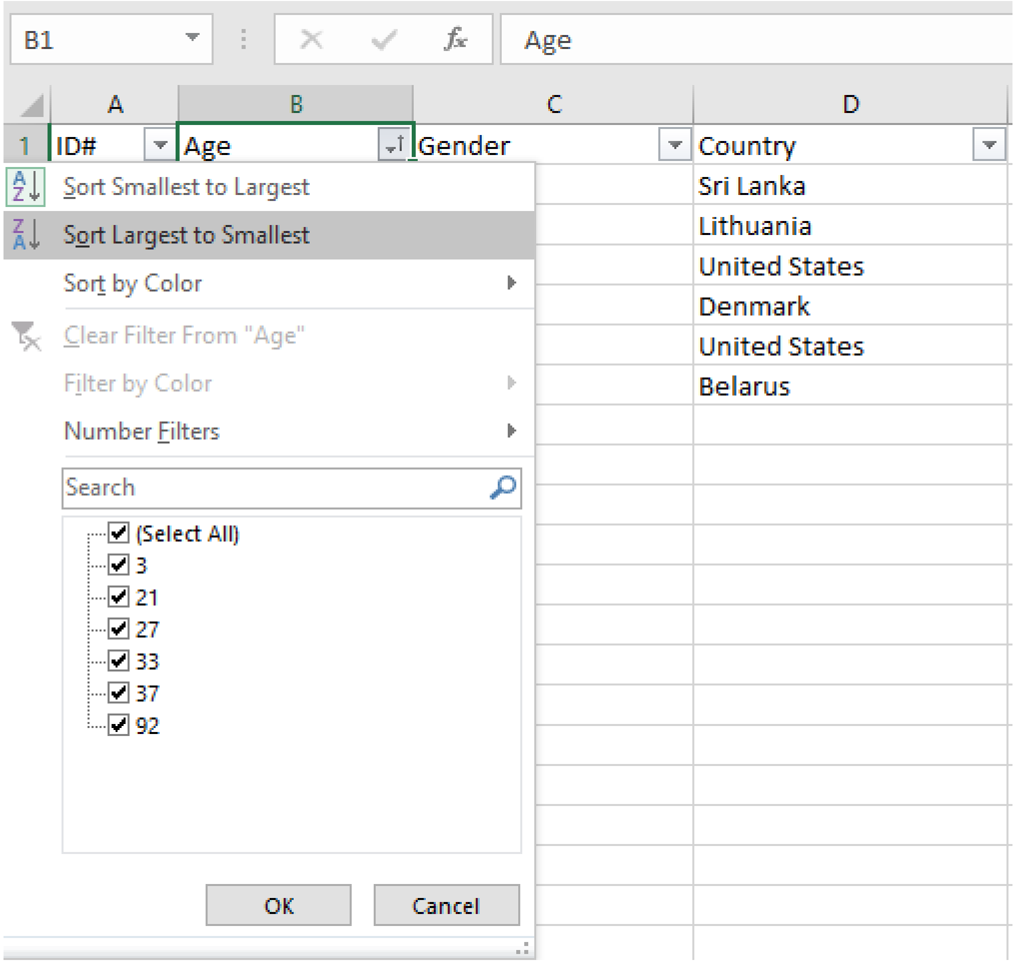
A typo is typically a spelling error. In the case of data collected from surveys, it is a data entry that is either clearly misspelled or can reasonably be assumed (based on the question or range of acceptable values) to be a mistake.

The following two examples show how you would go about identifying a typo, in the next section, we will discuss how to address typos in your data cleaning using the Excel ‘AND’ and ‘OR’ functions.

Example 2.1.1:
In the following example, the survey asked participants to answer, on a scale of 1 to 100, how satisfied they were with a program. With the help of your codebook, you know that for the corresponding variable *ProgramSatisfaction*, acceptable values range on a scale from 1 to 100, so a value of 999 is clearly a typo.

ID	ProgramSatisfaction
11234156	66
23456443	999
34523463	27
32457534	32
45363434	100

Example 2.1.2:
In the following example, you have asked survey participants to enter their age. Once you’ve sorted your data, you can quickly see if there are any values outside of the norm, the program works with young professionals (25-34) so someone being the age of 3 or 92 participating (let alone being able to write for the 3-year old!) in the program is highly unlikely.



B1 | Age

ID#	Age	Gender	Country
			Sri Lanka
			Lithuania
			United States
			Denmark
			United States
			Belarus

This "eyeballing" method works if you have a low number of survey responses as in the example above where you have 6 observations. In a case where you have 50+ observations, it would be time consuming to review all observations from the dropdown menu for each response. If you have a low number of potential values such as "Female", "Male" or a range of 1-100, you can use a combination of Excel functions[4] as outlined below:

[4] A function is predetermined formula that helps perform common mathematical functions. Functions can save you the time of writing lengthy formulas. Each function has a specific order, called syntax, which must strictly followed for the function to work correctly

1. All functions begin with the equals (=) sign
2. After the equals sign, define the function name (.e.g. Sum)
3. Add one or more arguments – numbers, text, or cell references – enclosed by parentheses. If there is more than one argument, separate each by a comma.
4. An example of a function that calculates the average of numbers in a range of cells (B3 through B10 and C3 through C10) is: =AVG(B3:B10, C3:C10)
5. Some tips on using formulas:
    a. When using a function, always keep text value in quotes "Female", "Male", "Country", etc.
    b. Lowercase/Uppercase matters: if during manual data entry someone entered "female" instead of "Female", Excel will recognize these as different values

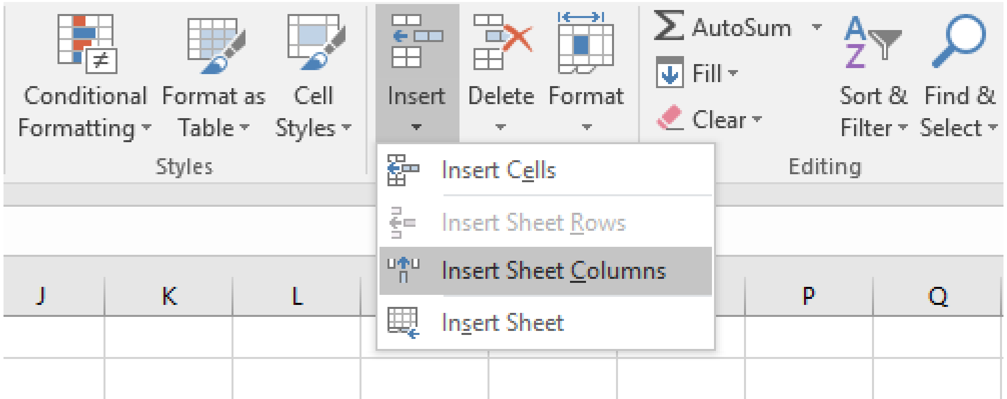
THE EXCEL "AND" FUNCTION

The first function you need to know is the "AND" Function, which allows you to test whether the values that were entered meet certain conditions (such as >1 but <100). The AND function tests the conditions you give it and returns TRUE if all conditions are evaluated to be TRUE, if not the function returns FALSE

- The function follows a particular formula: AND(logical1, logical2, …)
- Logical1, logical2, etc. are the conditions you want to test and get a TRUE or FALSE statement back. You can test just one condition or several
- Now let's look at some examples of the formula that show how to use the AND functions in Excel

Formula	**Description**
=AND(A2="Female")	Returns TRUE if A2 contains "Female", Returns False otherwise
=AND(B2>=1, B2<=100)	Returns TRUE if B2 is equal or greater than 1 **and** if B2 is equal or lesser than 100 (that is B2 is between 1 and 100), Returns FALSE otherwise
=AND(A2="Female", B2>=1, B2<=100)	Returns TRUE if A2 contains "Female" **and** if B2 is equal or greater than 1 **and** B2 is equal or lesser than 100, Returns FALSE otherwise

Let's look at an example of how you would use the AND function when cleaning data. Select a variable (select the whole column), and insert a column next to it. Give it a descriptive name such as Test_ProgramSatisfaction.

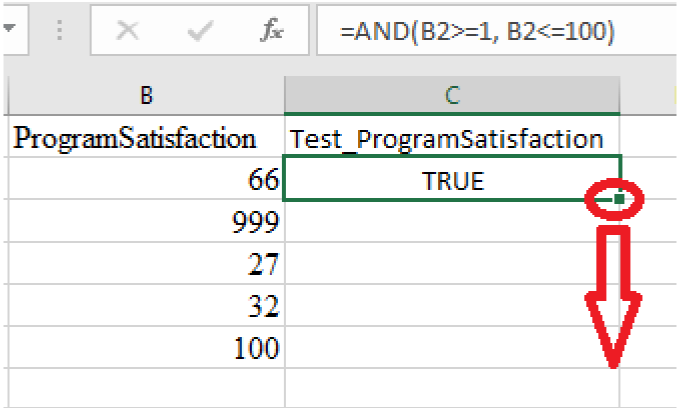


Using Example 1, we're going to use the AND function to test whether the values for the variable *ProgramSatisfaction* are within the range of 1 to 100 (reminder: the survey asked participants to answer, on a scale of 1 to 100, how satisfied they were with a program so anything not in that range could be a typo).

In this case, the *ProgramSatisfaction* variable is in column B and the test column is C *(Test_ProgramSatisfaction),* and we've told excel using the AND function to test the two logical statements:

i) B2>=1: if the value in B2 is greater or equal to 1 and;
ii) B2<=100: if the value in B2 is lesser or equal to 100.

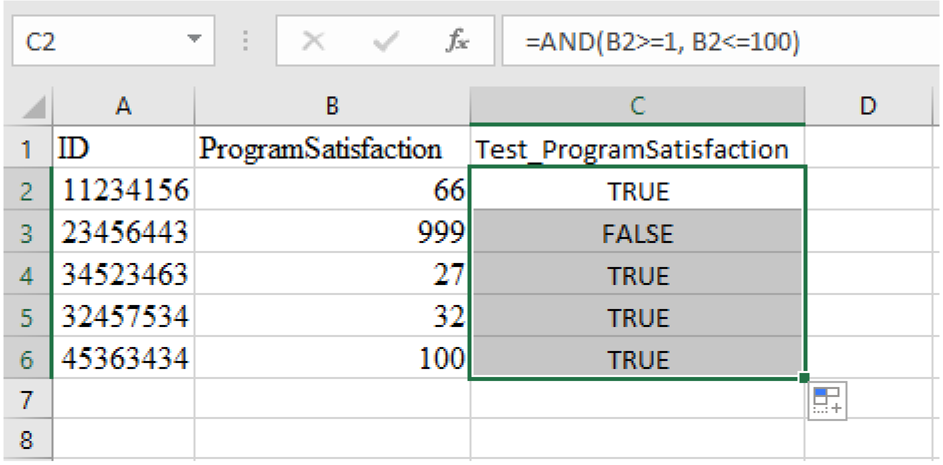
Excel will return a TRUE statement if the value B2 is between 1 and 100; otherwise we'll get a FALSE statement.



=AND(B2>=1, B2<=100)

B	C
ProgramSatisfaction	Test_ProgramSatisfaction
66	TRUE
999	
27	
32	
100	

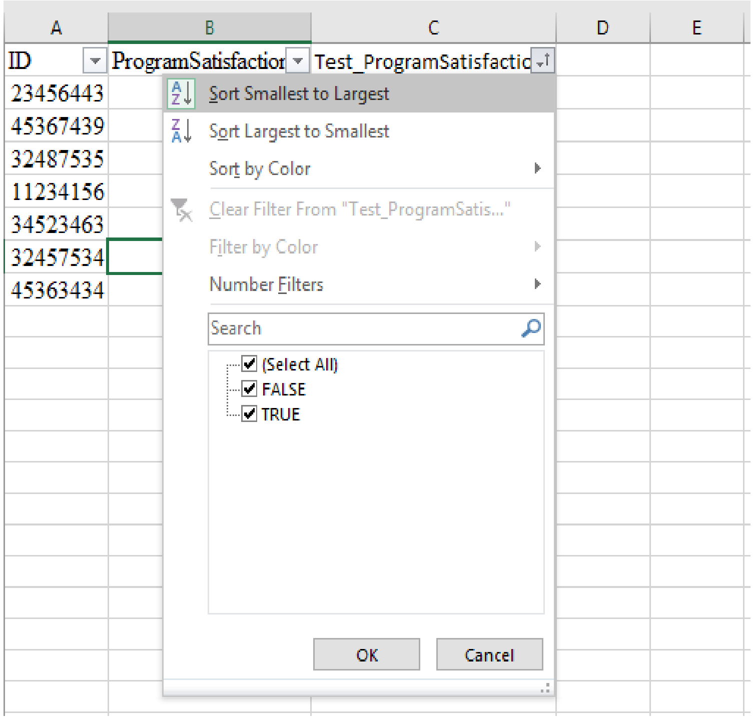
To apply the same function to the rest of the cells, you can either 1) double click on the lower right corner (circled in red) and it will fill out the rest of the column as long as there are other data in the column next to it or 2) click on the lower right hand corner and drag it down. And TADA! You can now see that you have one value that does not meet the conditions set out earlier.



C2 | =AND(B2>=1, B2<=100)

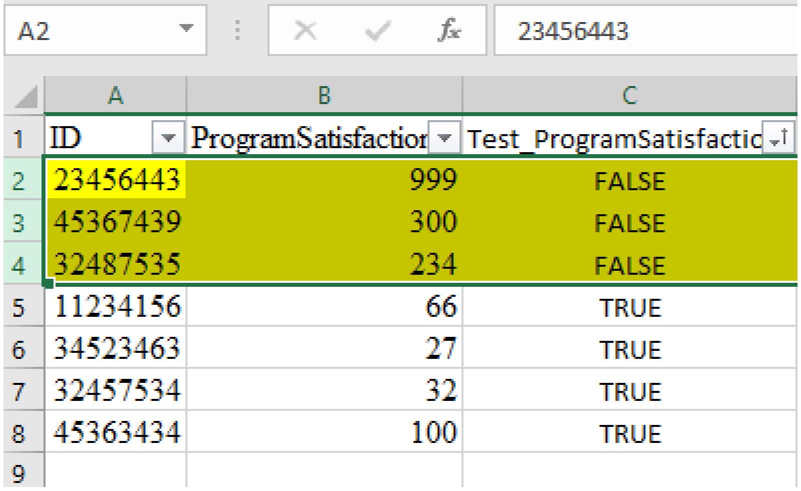
	A	B	C	D
1	ID	ProgramSatisfaction	Test_ProgramSatisfaction	
2	11234156	66	TRUE	
3	23456443	999	FALSE	
4	34523463	27	TRUE	
5	32457534	32	TRUE	
6	45363434	100	TRUE	
7				
8				

You can now use the filter to sort all the values that do not meet the conditions.



ID	ProgramSatisfactio	Test_ProgramSatisfactio
23456443		
45367439		
32487535		
11234156		
34523463		
32457534		
45363434		

Now highlight all rows (i.e. observations) that have a 'FALSE' value



	A	B	C
1	ID	ProgramSatisfactio	Test_ProgramSatisfactio
2	23456443	999	FALSE
3	45367439	300	FALSE
4	32487535	234	FALSE
5	11234156	66	TRUE
6	34523463	27	TRUE
7	32457534	32	TRUE
8	45363434	100	TRUE
9			

Now let's fix the typos:

1. Use the filter again to sort the test column to bring all the typos to the top.
2. Insert a new column
3. Copy the column you are looking at (in this case *ProgramSatistisfaction*) into the new column
4. Rename it "*ColumnName*_Fixed" where *ColumnName* is the name of the variable you are looking at (*ProgramSatisfaction_Fixed)*
5. Fix the typos in this new column: in this case, since there is no way to know what the original value was (or what the participant intended to enter), you will leave the value blank
6. Update your codebook to include the newly created column, the corresponding question and acceptable values

Remember you want to make the steps of cleaning your data as clear as possible for yourself and anyone who might be looking at your data later. That is why instead of simply deleting the typos here, we are taking the additional step of creating a separate column that shows that the ‘typos’ were removed/left blank

THE EXCEL “OR” FUNCTION

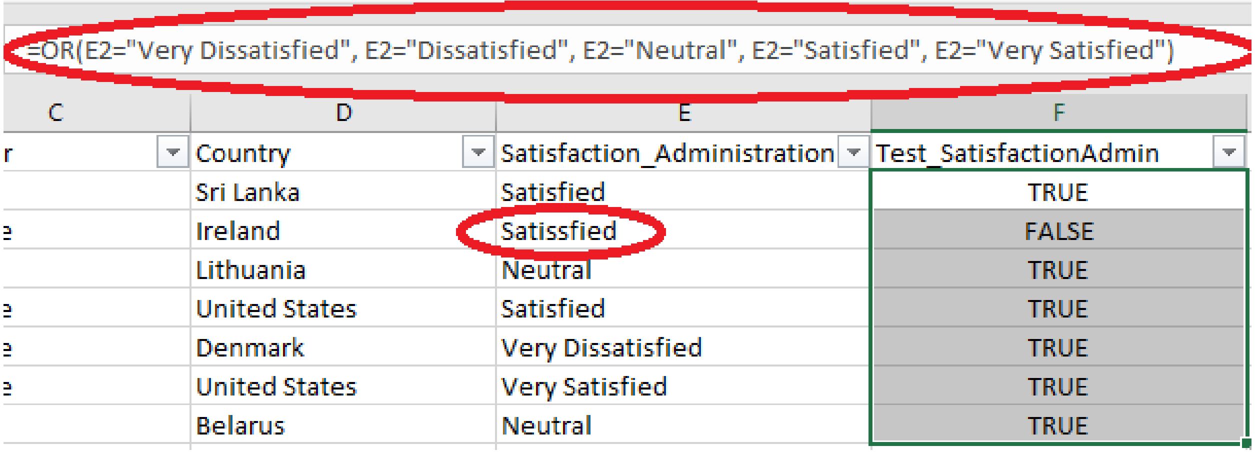
The second function you need to know is the “OR” function, which allows you to test whether any of the listed conditions is true. The OR function will return TRUE if **at least one** of the conditions is true and FALSE if all the listed conditions is false

- The function follows a particular formula: =OR(logical1, logical2, …)
- Logical1, logical2, etc. are the conditions you want to test and get a TRUE or FALSE statement back. You can test just one condition or several. Note that you have to repeat the “Cell#=” as part of each condition

Now let’s look at some examples of the formula that show how to use the OR functions in Excel

Formula	**Description**
=OR(A2=“Female”)	Returns TRUE if A2 contains “Female”, Returns False otherwise. When only one condition is given, the OR function behaves similarly to the AND function.
=OR(A2= “Female”, A2= “Male”)	Returns TRUE if A2 contains “Female” **or** “Male”, FALSE otherwise
=OR(B2>=1, B2<=100)	Returns TRUE if B2 is equal or greater than 1 **or** B2 is equal or lesser than 100 (that is B2 is between 1 and 100), Returns FALSE otherwise
=OR(A2= “ ”, B2= “ ”)	Returns TRUE if A2 **or** if B2 blank **or** both, Returns FALSE otherwise

Insert a blank column next to the variable you want to clean and give it a descriptive name such as *Test_SatisfactionAdmin*. For the example below, we are testing whether the values for the variable *Satisfaction_Administration* all meet the conditions of falling into the distinct categories of “Very Dissatisfied”, “Dissatisfied”, “Neutral”, “Satisfied”, and “Very Satisfied” as seen in the picture below. This is one example of how the OR function can be used to identify typos.



```
=OR(E2="Very Dissatisfied", E2="Dissatisfied", E2="Neutral", E2="Satisfied", E2="Very Satisfied")
```

C	D	E	F
r	Country	Satisfaction_Administration	Test_SatisfactionAdmin
	Sri Lanka	Satisfied	TRUE
e	Ireland	Satissfied	FALSE
	Lithuania	Neutral	TRUE
e	United States	Satisfied	TRUE
e	Denmark	Very Dissatisfied	TRUE
e	United States	Very Satisfied	TRUE
	Belarus	Neutral	TRUE

Now let’s fix the typos:

1. Use the filter again to sort the test column to bring all the typos to the top.
2. Insert a new column
3. Copy the column you are looking at (in this case *Satistisfaction_Administration*) into the new column
4. Rename it "*ColumnName*_Fixed" where *ColumnName* is the name of the variable you are looking at (*Satisfaction_Administration_Fixed)*
5. Fix the typos in this new column
6. Update your codebook to include the newly created column, the corresponding question and acceptable values

Remember you want to make the steps of cleaning your data as clear as possible for yourself and anyone who might be looking at your data later. That is why instead of simply fixing the typos in the original column, we created a new column with a header that makes it clear that any issues/typos have been fixed.

In some cases, it will be easy to fix the typo when it's clear what the correct value is, as in the example of 'Satissfied' it's pretty clear that the correct value is 'Satisfied'. In other cases, such as when the survey participant has entered '999' when asked to rate program satisfaction on a scale of 1 to 100, they could have meant to enter '99' or '9' or anything else. In such cases, unless you have the ability to go back to participants (time-consuming), you may have to remove such values from your dataset.

2.2 Mindless Entries

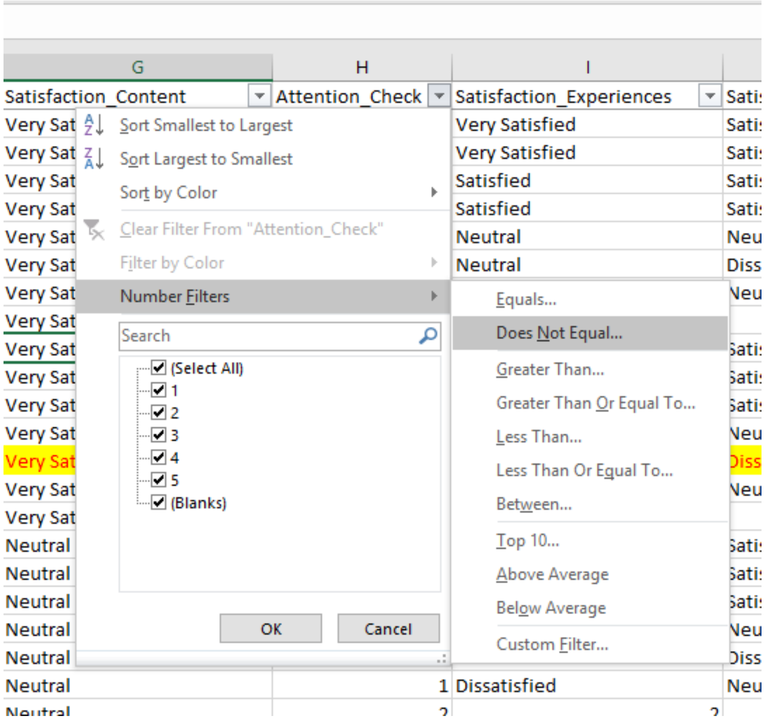
Typically, survey participants are not passionate about the quality of the data, they just want to get through the survey. For this reason, although some data may appear to be complete, these responses could be meaningless if answered without effort (e.g. a program participant answering "Strongly Agree" to every answer just to finish quickly). Their response adds no insight or value to analysis or the program and should be eliminated from the dataset, even though the response looks complete and genuine at first glance. Checking the data for simple or contradictory answers is necessary.

Additionally, adding in an attention check to your survey could also help. An example of an attention check question could be:

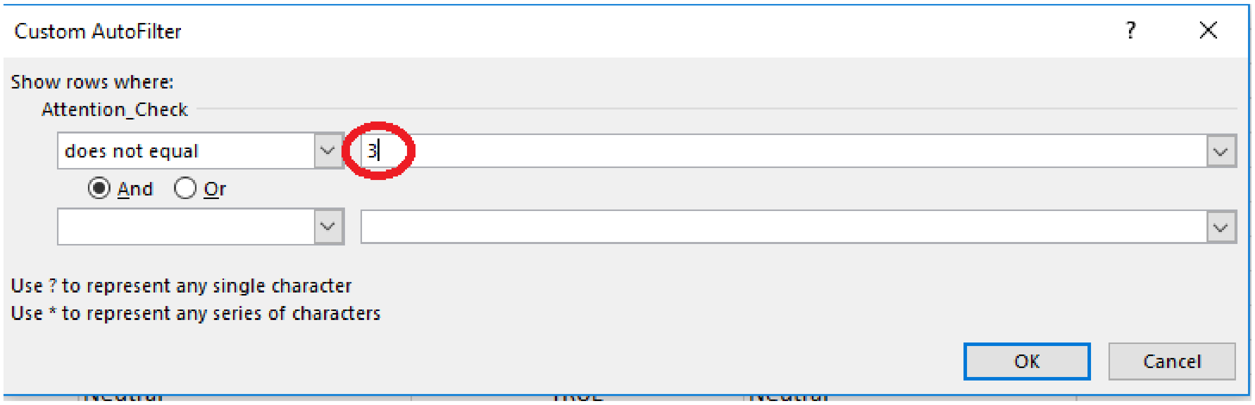
Please select 3 if you're paying attention:
1 2 3 4 5 6 7 8 9
This is a quick, but effective method to validate data as participants respond. If you have an attention check question in your survey, you would use this question to identify survey participants who were not paying attention.

To identify participants who were not paying attention (i.e. did not select 3), in Excel you will use the filter button →

1. Go to 'Number Filters' → 'Does Not Equal'



2. In the dialog box for ‘does not equal’ enter 3 --> Click OK


Custom AutoFilter

Show rows where:

Attention_Check

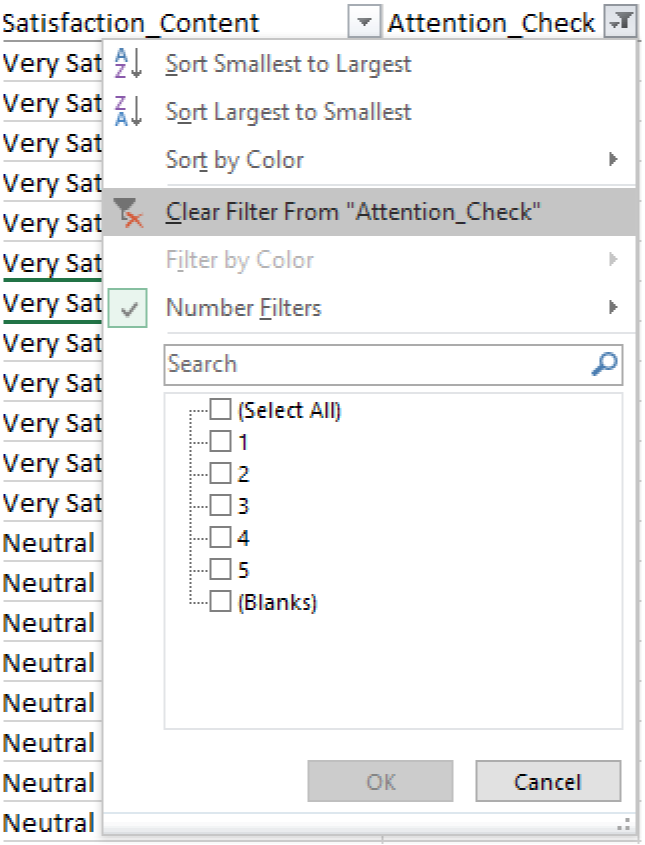
does not equal | 3

● And ○ Or

Use ? to represent any single character

Use * to represent any series of characters

3. You will be left with just those data points where the participants were not paying attention.
4. Highlight these data points. We will come back to them and remove them in the last step of our data cleaning in section 2.7.
5. Clear the filter so that you can now see all the values again



Satisfaction_Content	Attention_Check
Very Sat	
Very Sat	
Very Sat	
Very Sat	
Very Sat	
Very Sat	
Very Sat	
Very Sat	
Very Sat	
Very Sat	
Very Sat	
Very Sat	
Neutral	
Neutral	
Neutral	
Neutral	
Neutral	
Neutral	
Neutral	
Neutral	

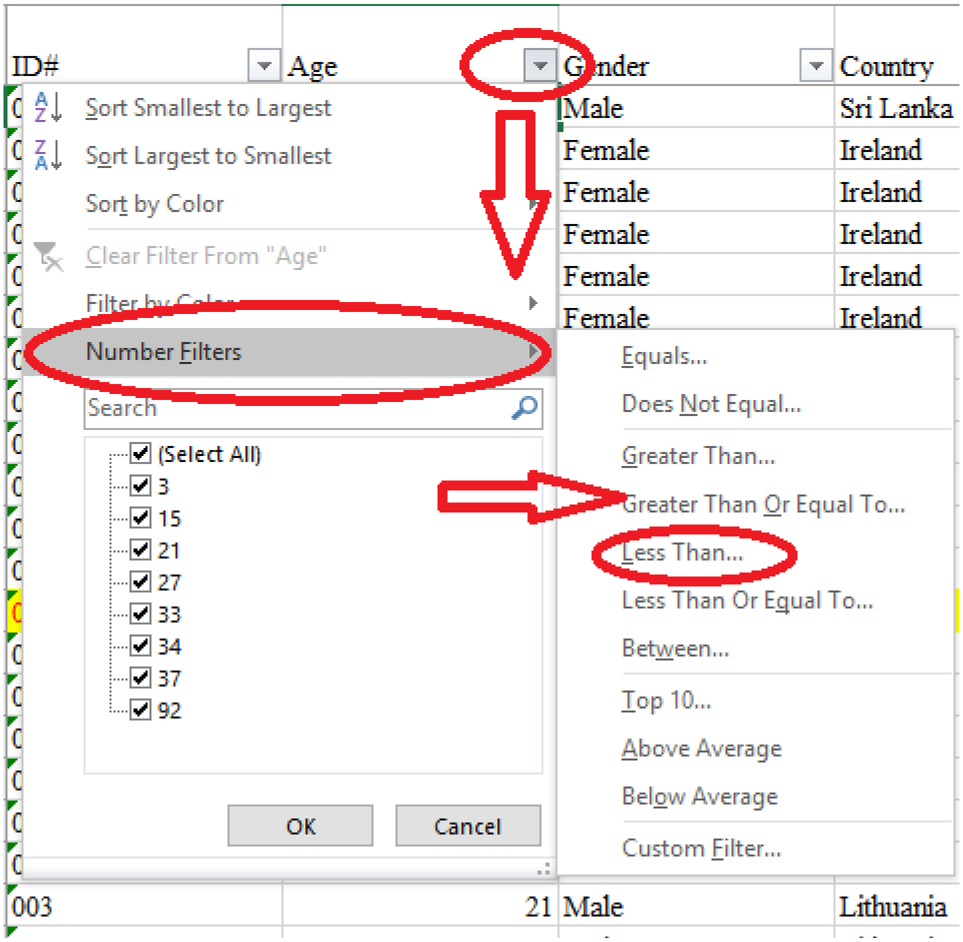
If you did not have an attention check in your questionnaire, not to worry, the process of data cleaning should help eliminate some of the mindless entries.

2.3 Illogical Values

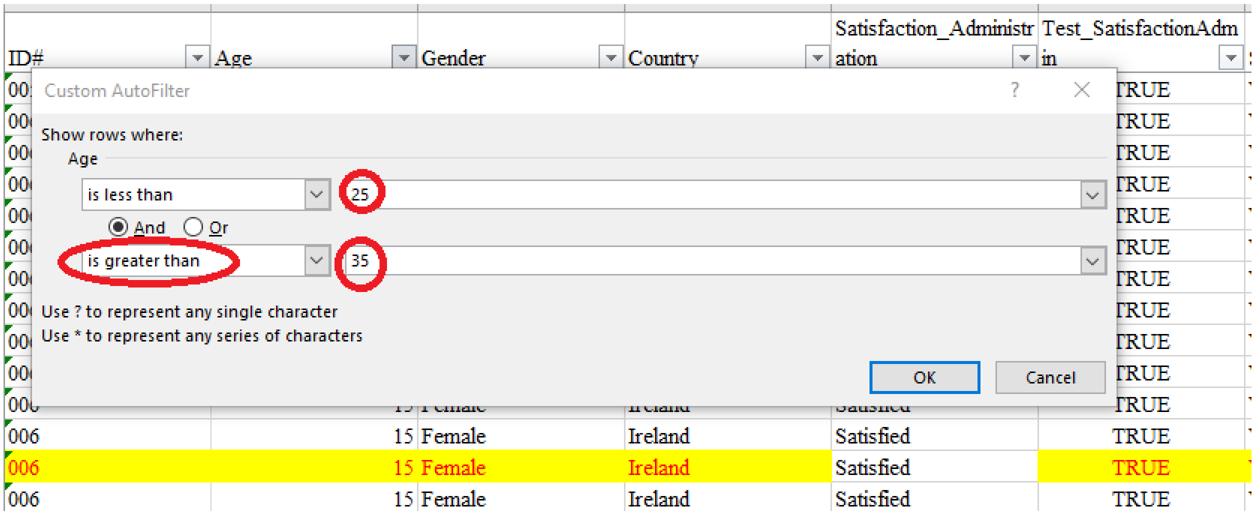
If you have a program that is focused on young professionals (25 – 35), for example a program called Young Leaders, it is unlikely that any of your participants would be younger than 25 or older than 35. If you do have a survey respondent select that they are 45 you know this is not possible, this is called an illogical value. These illogical values call the entire participant response in to question, and may need to be eliminated from the data.

Example 2.3.1:
If you have a situation such as all your participants are 25 - 35, you would use the filter button to identify the illogical values:

1. Go to ‘Number Filters’ → ‘Less than…’



2. In the dialog box, for ‘is less than’ enter “25”
3. In the second drop down, select ‘is greater than’ and enter “35”


Custom AutoFilter

Show rows where:

Age

is less than 25

● And ○ Or

is greater than 35

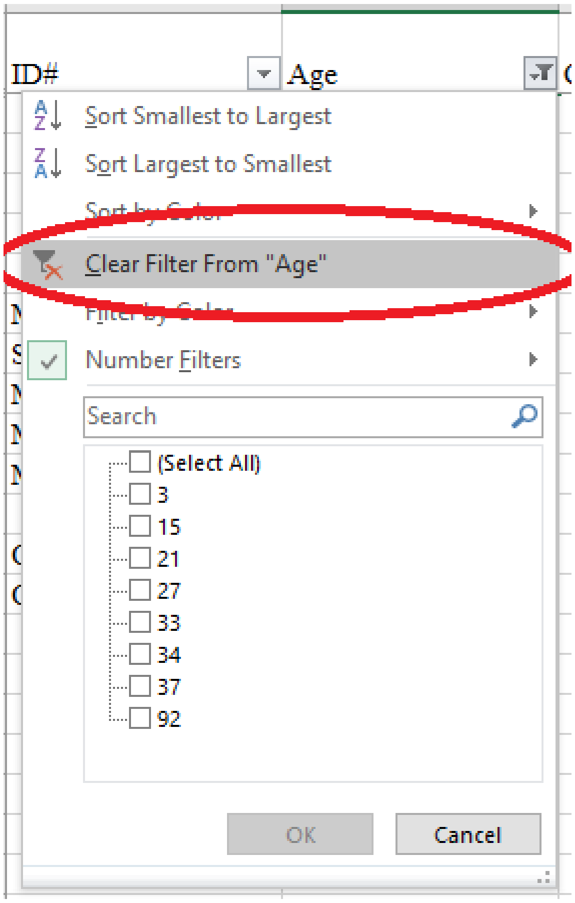
Use ? to represent any single character

Use * to represent any series of characters

4. This leaves you with all the illogical values (i.e. participants who selected a number less than 25 and older than 35). Highlight all those data points. We will come back to these in our last step of data cleaning.

Note: You should think about what the appropriate range given the timeframe you are surveying in is. If you are surveying alumni of a program 2 years after its end then you should adjust the top of your age range, since participants who were 35 two years ago will now be 37 or even 38. If you are surveying within 1-2 months, some participants may have turned 36.

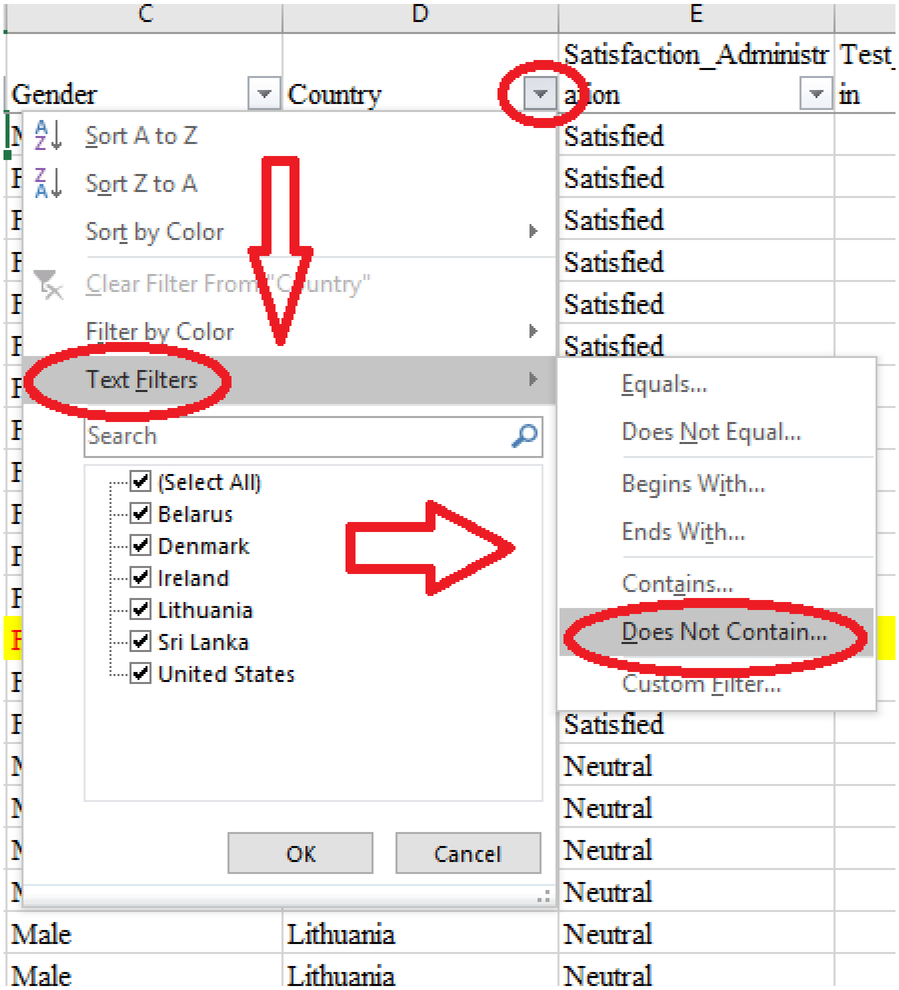
5. Clear the Filter



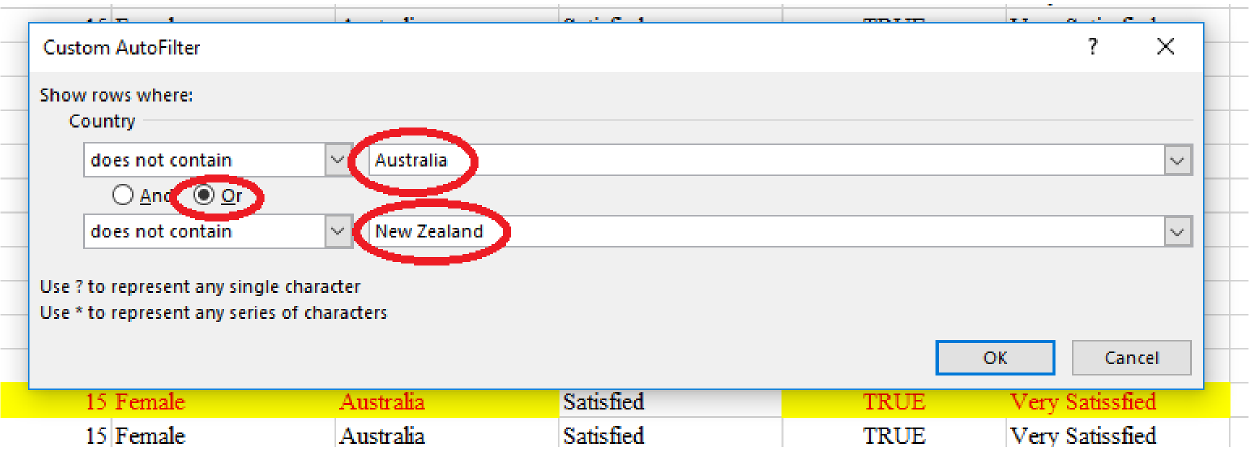
Note: The 'Number Filters' features several different options: less than, greater than, does not equal, between, which can all be used to identify illogical <u>numeric</u> values. In the example 2.3.2, we will see how to identify illogical <u>text</u> values.

<u>Example 2.3.2</u>:
Now consider if you have a program focused on participants from Australia and New Zealand. If you have participants who said they were from the United States or any other country, this would be an illogical value, you would use the text filter button to identify the illogical values:

1. Go to 'Text Filters' → 'Does Not Contain…'



2. In the dialog box, for ‘does not contain’ enter “Australia” → Check the ‘Or’ button and use the dropdown to select ‘does not contain’ and enter ‘New Zealand’



Show rows where:

Country

does not contain | Australia

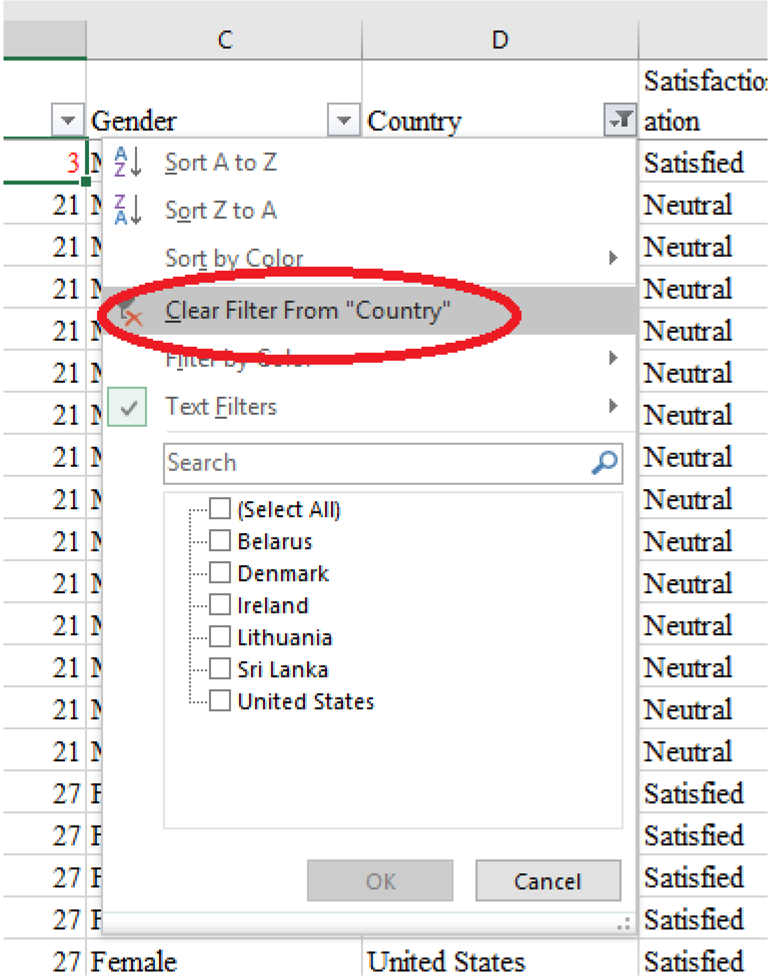
○ And ● Or

does not contain | New Zealand

Use ? to represent any single character
Use * to represent any series of characters

3. This leaves you with all the illogical values (i.e. participants who selected a country that is not in line with the focus countries). Highlight all those data points as we will come back to these later in the last step of our data cleaning in section 2.7.

4. Clear the Filter



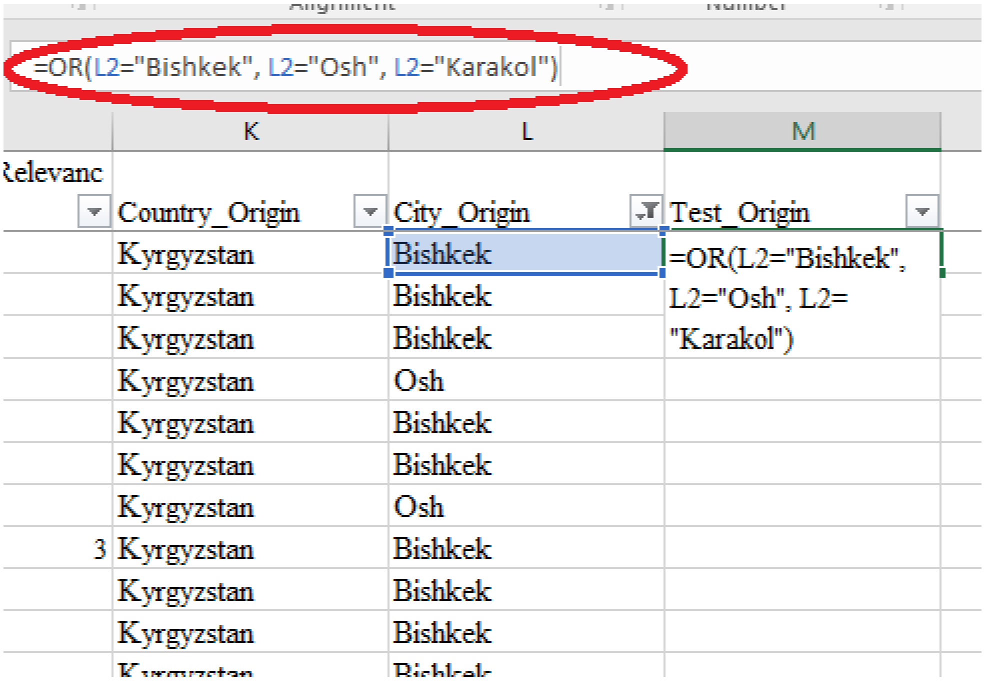
Note: You can only use the drop down (text or numeric) filters for up to two values (as you can see from the above pop-up box). If you have more than two valid values, for example participants come from the UK (England, Ireland and Scotland), you can either do it for up to two values at a time or use the 'OR' function we learned about in section 2.1.

Example 2.3.3:
In this example, say you know that 50 people participated in your program, 40 answered your survey, and there were 5 participants from Kyrgyzstan—4 women and 1 man. Of the Kyrgyzstani participants, you see data like this:

ID	Country of Origin	Gender	City of Origin
12342135	Kyrgyzstan	M	Osh
12352346	Kyrgyzstan	F	Bishkek
12352366	Kyrgyzstan	F	Astana
12645456	Kyrgyzstan	F	Bishkek
12366789	Kyrgyzstan	F	Osh

Since Astana is in Kazakhstan, not Kyrgyzstan, either there is a Kazakh female who mistakenly put her home country as Kyrgyzstan, or a Kyrgyz female indicated the wrong city. In this example, you would use the OR function introduced in the typos section to identify these illogical values.

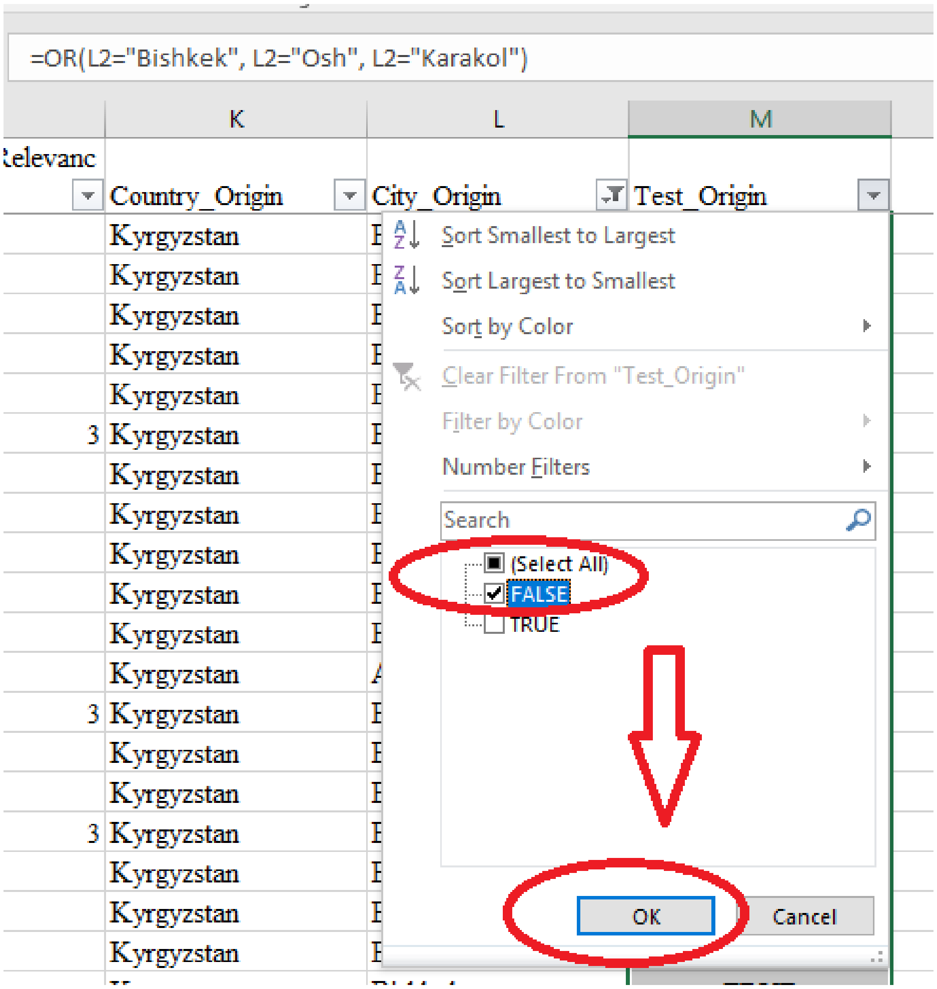
Insert a blank column next to the variable you want to clean and give it a descriptive name such as *Test_Origin*. We will test whether the values for the variable *Country of Origin* and *City of Origin* correspond. In the example below, Excel will return "TRUE" if the variable *City_Origin (Cell L2)* has either Bishkek, Osh, Karakol, etc. as values and "FALSE" if not.



=OR(L2="Bishkek", L2="Osh", L2="Karakol")

Relevanc	Country_Origin	City_Origin	Test_Origin
	Kyrgyzstan	Bishkek	=OR(L2="Bishkek", L2="Osh", L2= "Karakol")
	Kyrgyzstan	Bishkek	
	Kyrgyzstan	Bishkek	
	Kyrgyzstan	Osh	
	Kyrgyzstan	Bishkek	
	Kyrgyzstan	Bishkek	
	Kyrgyzstan	Osh	
3	Kyrgyzstan	Bishkek	
	Kyrgyzstan	Bishkek	
	Kyrgyzstan	Bishkek	
	Kyrgyzstan	Bishkek	

1. Drag the formula down
2. Use the filter button to identify the data points that do not meet the specified conditions (e.g. where Excel returned 'FALSE')



```
=OR(L2="Bishkek", L2="Osh", L2="Karakol")
```

3. Highlight all illogical values. We will remove these in the last step of our data cleaning process in section 2.7.

2.4 Duplicates

If you have only 50 participants, and you have 55 complete responses, it is likely that five people may have duplicate responses. Understanding who is answering your survey questions –the age of the participants, their gender, and other identifying factors will help you identify values that could potentially be duplicates.

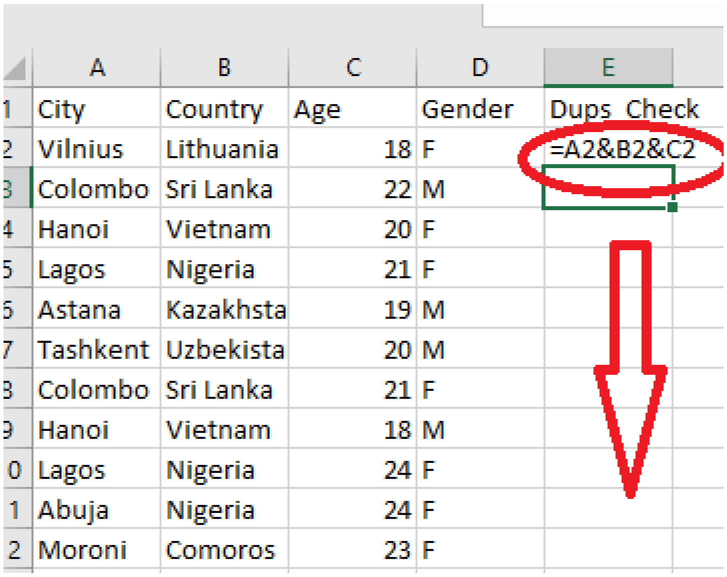
You know that 50 people participated in your program, 40 answered your survey, and there were 5 participants from Kyrgyzstan—3 women and 2 men. Of the Kyrgyzstani participants, you see data like this:

ID	Gender	V1	V2	V3	V4	V5	V6	V7
11664346	M	1	1	1	3	1	2	4
52364664	M	2	5	7	2	1	1	1
98798237	M	1	1	1	2	1	2	4
21357239	F	4	4	4	4	4	5	6
34257693	F	7	7	6	5	2	1	1

You can see that there should be two males—not 3-- and that the third and first males' answers are extremely similar. Since the data do not match what you know about participation in the program and the answers are very similar between the two males, it is likely that both of those responses came from the same participant. If the data were less neatly organized, or displayed as text, it would be much more difficult to identify this duplicate set of answers.

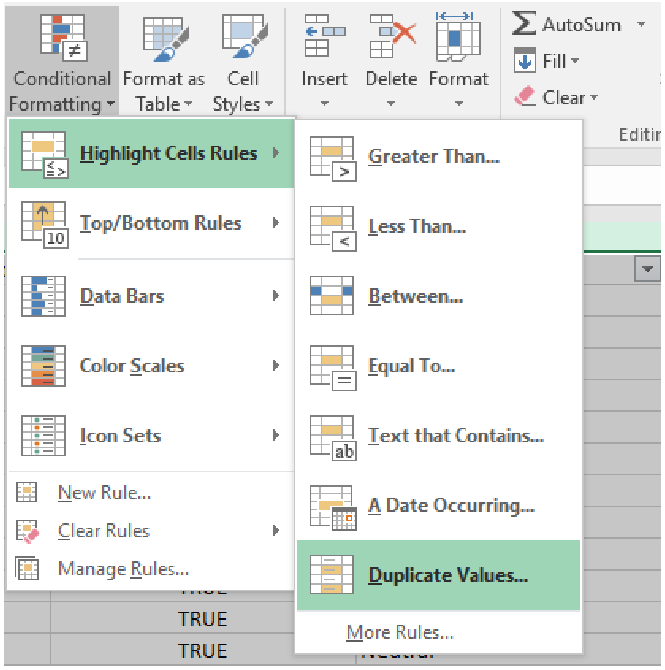
To identify duplicate values in Excel:

1. Create a new column called 'Dups_Check' and combine values from different columns such as city, state, age, and gender to flag records you should look at more closely
   - In order to do this, you can use the '&' operator, which allows you to combine values from different columns.
   - To combine values from cells A2, B2 and C2, you would write the following formula: "=A2&B2&C2"

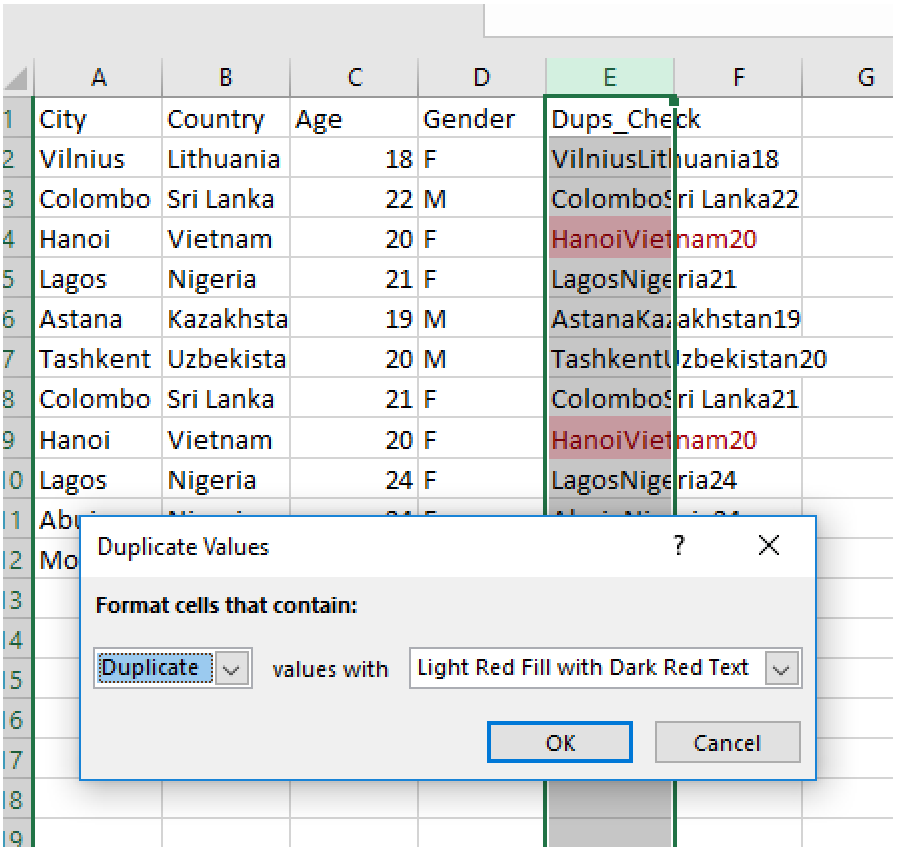


	A	B	C	D	E
1	City	Country	Age	Gender	Dups_Check
2	Vilnius	Lithuania	18	F	=A2&B2&C2
3	Colombo	Sri Lanka	22	M	
4	Hanoi	Vietnam	20	F	
5	Lagos	Nigeria	21	F	
6	Astana	Kazakhsta	19	M	
7	Tashkent	Uzbekista	20	M	
8	Colombo	Sri Lanka	21	F	
9	Hanoi	Vietnam	18	M	
0	Lagos	Nigeria	24	F	
1	Abuja	Nigeria	24	F	
2	Moroni	Comoros	23	F	

   - Either drag down or double-click on the lower right hand corner to auto-fill the remaining columns
2. Select the column 'Dups_Check'
3. Go to 'Conditional Formatting' → 'Highlight Cell Rules' → Duplicate Values



4. A dialog box will come up – this is just to select what type of formatting you would like (leave as default) and select OK. All duplicates will be in red formatting.



	A	B	C	D	E
1	City	Country	Age	Gender	Dups_Check
2	Vilnius	Lithuania	18	F	VilniusLithuania18
3	Colombo	Sri Lanka	22	M	ColomboSri Lanka22
4	Hanoi	Vietnam	20	F	HanoiVietnam20
5	Lagos	Nigeria	21	F	LagosNigeria21
6	Astana	Kazakhsta	19	M	AstanaKazakhstan19
7	Tashkent	Uzbekista	20	M	TashkentUzbekistan20
8	Colombo	Sri Lanka	21	F	ColomboSri Lanka21
9	Hanoi	Vietnam	20	F	HanoiVietnam20
10	Lagos	Nigeria	24	F	LagosNigeria24
11	Abu				
12	Mo				

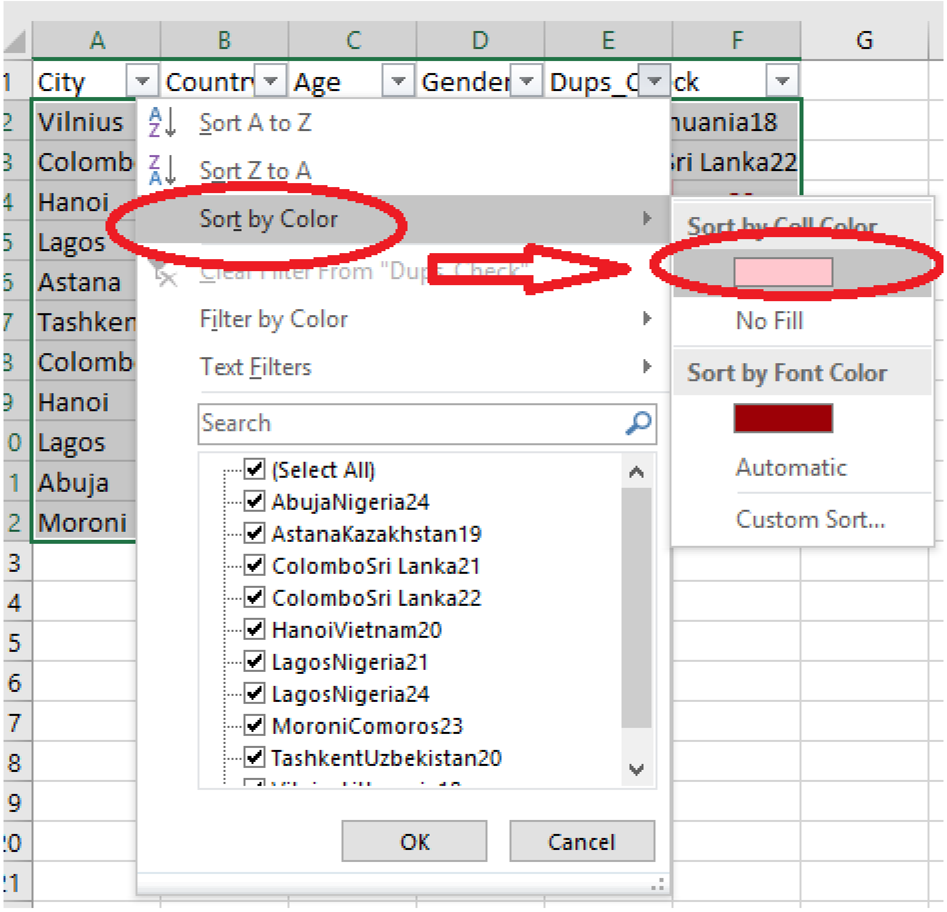
Duplicate Values

Format cells that contain:

Duplicate values with Light Red Fill with Dark Red Text

OK Cancel

5. Use the filter button to keep only the duplicate values: Go to ‘Sort by Color’ → ‘Sort by Cell Color’.



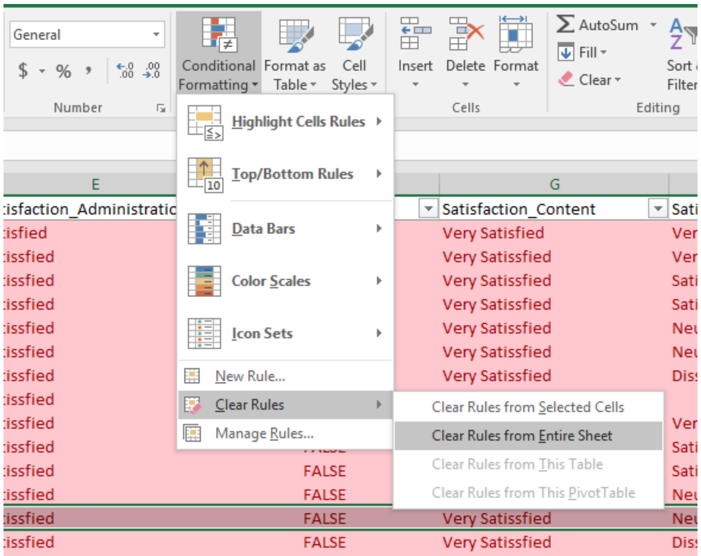
6. Next review the values across the different rows to identify duplicate values such as in the example below. Determining whether values are duplicates is going to be based on your judgement and your observation of the data.
   - In the table below, there are three different points: Age, Gender and State as well as the remaining responses to the questions that are similar. We can reasonably assume that these are duplicate values

ID	Age	Gender	State	V1	V2	V3	V4	V5	V6	V7
11664346	32	M	ID	1	1	1	3	1	2	4
52364664	27	M	KY	2	5	7	2	1	1	1
98798237	32	M	ID	1	1	1	2	1	2	4
21357239	28	F	SD	4	4	4	4	4	5	6
34257693	31	F	CA	7	7	6	5	2	1	1

   - In the table below, we have a similar example but we only have one identifying characteristic (Gender) as well as the remaining responses to the questions that are similar. Here, it is more difficult to determine whether these are duplicate values because it is possible (though unlikely) that two random males surveyed had the same answers

ID	Gender	V1	V2	V3	V4	V5	V6	V7
11664346	M	1	1	1	3	1	2	4
52364664	M	2	5	7	2	1	1	1
98798237	M	1	1	1	2	1	2	4
21357239	F	4	4	4	4	4	5	6
34257693	F	7	7	6	5	2	1	1

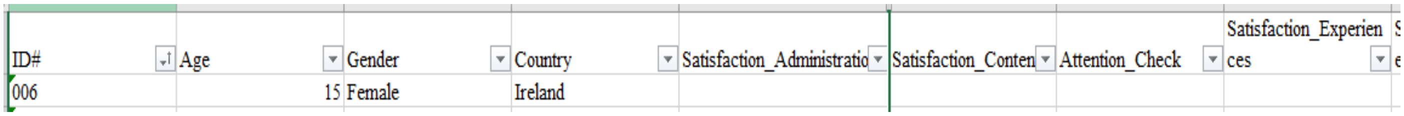
7. For those data points that you identify as duplicates, go ahead and select all but one of the duplicates (if you have 3 duplicates, select 2) and highlight them (this won't immediately show up because of the conditional formatting but it is happening)
8. Once you have identified and highlighted the duplicates, clear the conditional formatting: Click on the 'Conditional Formatting' button → 'Clear Rules' → 'Clear Rules from Entire Sheet'
9. Again do not worry, we will deal with duplicate values in the last step of our data cleaning in section 2.7.



2.5 Missing values

Missing values or incomplete responses make data more variable and less reliable. Since it is impossible to determine why a response was halted prematurely, or why a question in a survey simply was not answered, it is difficult to determine if the responses reflect some greater truth like program satisfaction. While most data analysis software will not report incomplete responses, some may report responses where only one or two questions are missing. Whether or not to include these responses is a judgment call for the researcher; if the rest of the answer appears valid, and the dataset is not large enough where they could afford to simply eliminate the response from analysis, then it may be worthwhile to retain the response. **However, the missing value should not be added in to any analysis of that question as a response of "0" or N/A. It should simply appear as though there is one less data point when analyzing that answer.**

Generally you do not have to do anything about missing values, though it is still worthwhile to keep the data for analysis even if the responses are incomplete. However, in some cases, the survey participant answered so few questions that the data are not worthwhile: for example in a case like this where the participant entered only their demographic data (age, gender, country, etc.) but did not answer any of the substantive questions. In such cases, it is best to remove these observations:



ID#	Age	Gender	Country	Satisfaction_Administratio	Satisfaction_Conten	Attention_Check	Satisfaction_Experiences	S e
006	15	Female	Ireland					

In your data cleaning process, you will want to highlight these rows and remove them later in section 2.7.

2.6 Outliers

An outlier is typically defined as any value that is 3 standard deviations away from the mean, meaning that it is an extremely rare observation to see. For example, say I recently created a dataset of how long it takes for messages to reach different destinations over the internet. All of the times were in the range from 0.5 to 0.8 seconds, except for three. The other three were all over 5,000 seconds. This is a major red flag that something has gone wrong in the production of the data. In this particular case an error in the code I wrote caused some failures to continue counting while all other messages were being sent and received.

Outliers can either be caused by faulty data collection like this example, or they can simply be incredibly rare, but valid values. In either case, these values can lead you astray during the data analysis process—especially if you are using the mean. Median-based measures are less susceptible to this outlier effect. It is a good idea to take a look at the largest and smallest values and ensure they are in a reasonable range. In the data validation stage of data cleaning, simply noting outliers and documenting them adds information to later analysis.

Calculating simple descriptive statistics of your data is a simple way to begin this analysis. By calculating the mean, variance, standard deviation and median of numeric survey responses for each scale, you can begin to gain general insights regarding response patterns. These descriptive statistics help to identify the effect of outliers on the mean and other measures. In this way, you can determine whether or not it is useful to analyze the data with outliers included or eliminate outliers entirely from analysis.

The following describes the different Excel functions that you can use to calculate simple descriptive statistics of your data (note that this is only applicable to numeric variables), which can be used to identify outliers.

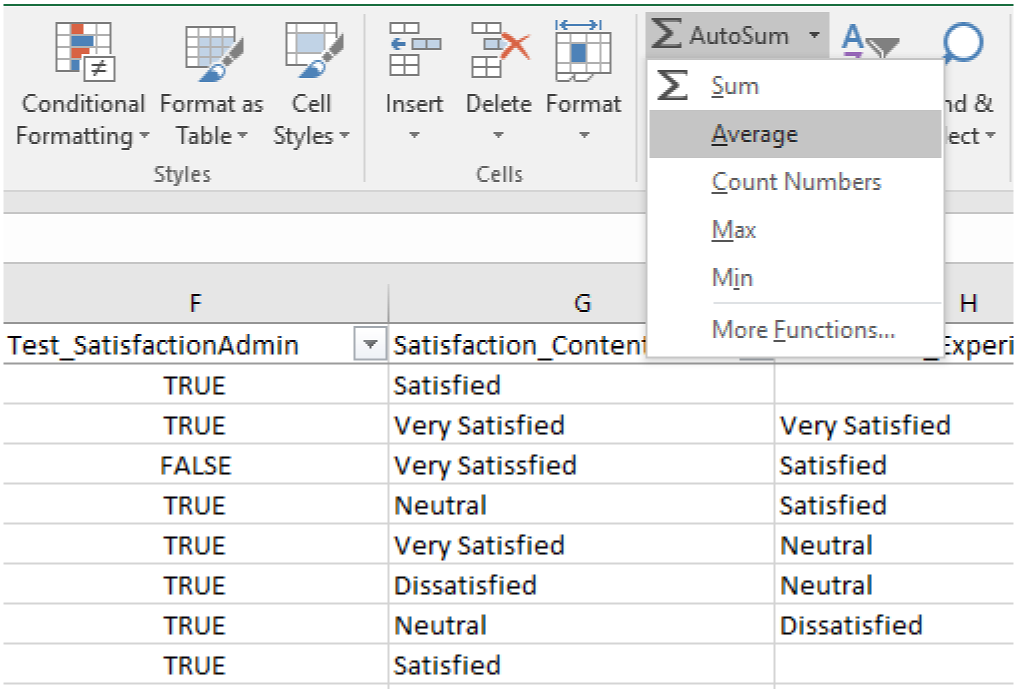
We will use primarily the following functions, which follow the same formula as previous Excel functions reviewed above.

Descriptive Statistic	**Formula**	**Description**
Mean	=AVERAGE(B2:B8)	Returns the mean (or average) of cells B2 through B8
Minimum Value	=MIN(B2:B8)	Returns the lowest value in the cells B2 through B8
Maximum Value	=MAX(B2:B8)	Returns the largest value in the cells B2 through B8
Standard Deviation	=STDEV.P(B2:B8)	Returns the standard deviation for cells B2 through B8.
Median	=MEDIAN(B2:B8)	Returns the median for cells B2 through B8.

Now, we are going to calculate these statistics for the numeric variables in our data:

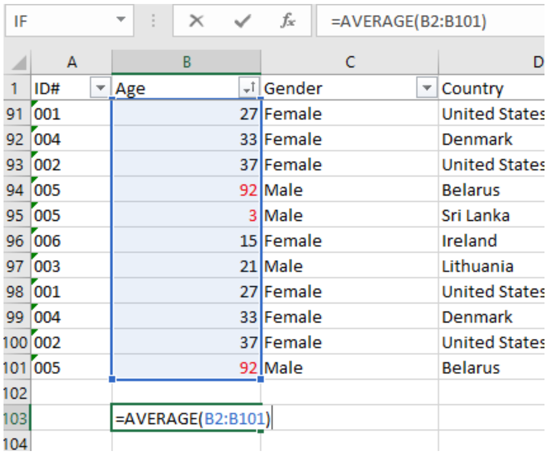
1. Scroll all the way down to the bottom of your dataset, and

2. Select a cell two rows below your last data point
3. For the Average, Min, and Max functions:
    a. Go to the top right hand corner
    b. Click on the AutoSum Button
    c. Select Average, Min or Max



F	G	H
Test_SatisfactionAdmin	Satisfaction_Content	_Experi
TRUE	Satisfied	
TRUE	Very Satisfied	Very Satisfied
FALSE	Very Satissfied	Satisfied
TRUE	Neutral	Satisfied
TRUE	Very Satisfied	Neutral
TRUE	Dissatisfied	Neutral
TRUE	Neutral	Dissatisfied
TRUE	Satisfied	

Your formula should look something like this (Excel will automatically select the data for you):



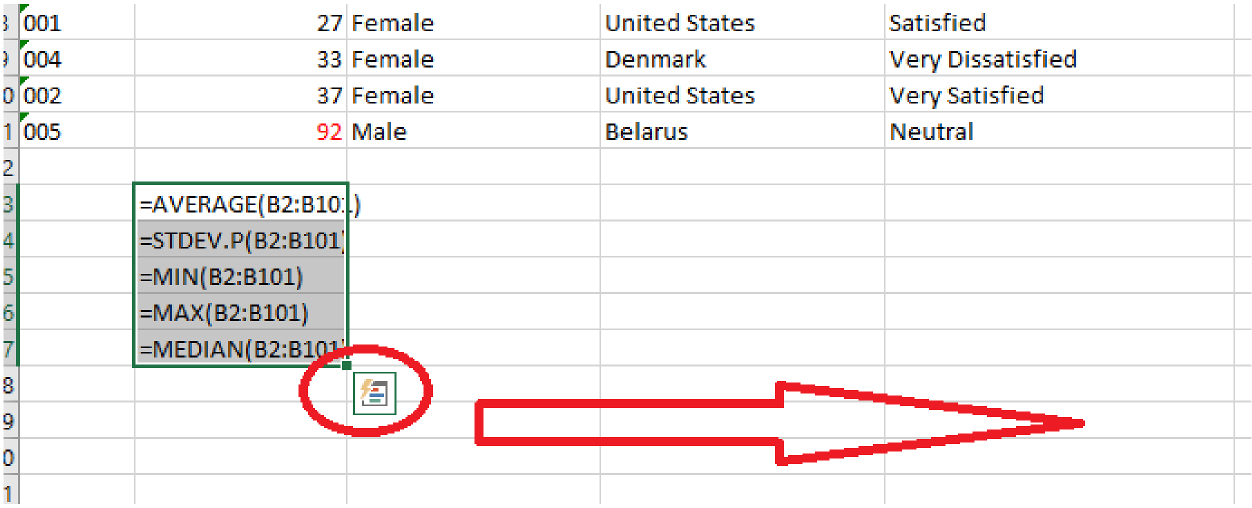
=AVERAGE(B2:B101)

	A	B	C	D
1	ID#	Age	Gender	Country
91	001	27	Female	United States
92	004	33	Female	Denmark
93	002	37	Female	United States
94	005	92	Male	Belarus
95	005	3	Male	Sri Lanka
96	006	15	Female	Ireland
97	003	21	Male	Lithuania
98	001	27	Female	United States
99	004	33	Female	Denmark
100	002	37	Female	United States
101	005	92	Male	Belarus
102				
103		=AVERAGE(B2:B101)		
104				

4. Hit Enter
5. To calculate the median and standard deviation, you'll need to type those in
    a. Select a cell right below your last formula
    b. Type in the formula and replace the cell numbers (in red) using those completed by the AutoSum button for the other formulas.
        i. =STDEV.P(B2:B8)
        ii. =MEDIAN(B2:B8)

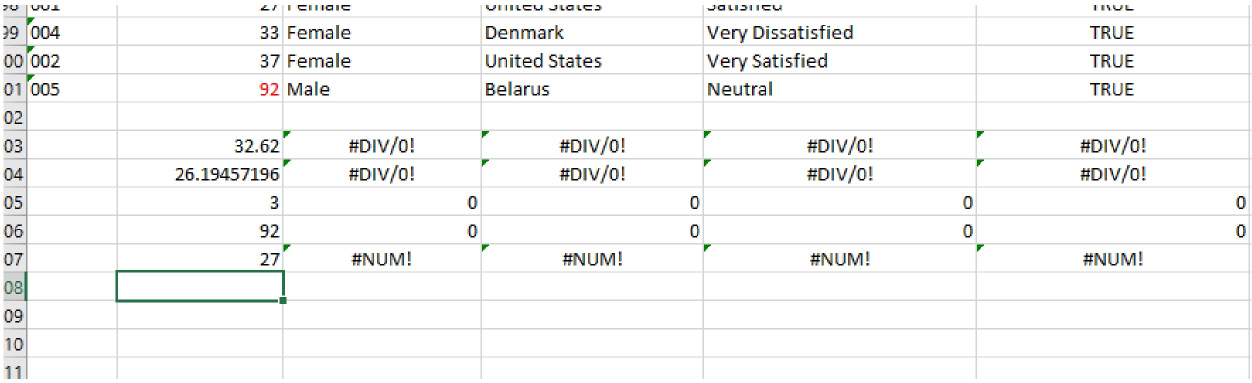
c. Hit Enter

Once you have all four formulas for the different statistics for the first variable, rather than hand-entering for each variable, you can just drag the formula across the worksheet (this is called the autofill feature):



3	001	27	Female	United States	Satisfied
9	004	33	Female	Denmark	Very Dissatisfied
0	002	37	Female	United States	Very Satisfied
1	005	92	Male	Belarus	Neutral
2					
3		=AVERAGE(B2:B101)			
4		=STDEV.P(B2:B101)			
5		=MIN(B2:B101)			
6		=MAX(B2:B101)			
7		=MEDIAN(B2:B101			
8					
9					
0					
1					

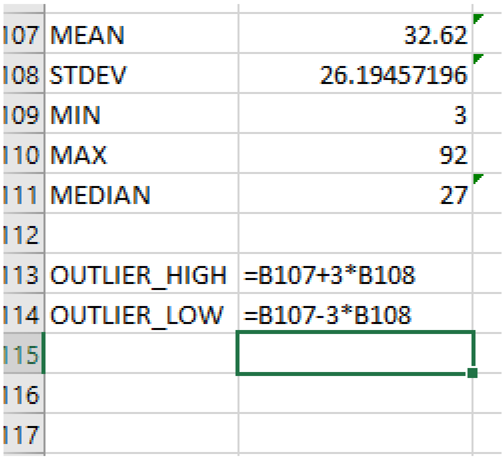
You might get something like this but do not panic! Just delete the formulas for the columns that have text (non-numeric) values as you see below with the #DIV/0 and the #NUM! Values.



98	001	27	Female	United States	Satisfied	TRUE
99	004	33	Female	Denmark	Very Dissatisfied	TRUE
00	002	37	Female	United States	Very Satisfied	TRUE
01	005	92	Male	Belarus	Neutral	TRUE
02						
03		32.62	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
04		26.19457196	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
05		3	0	0	0	0
06		92	0	0	0	0
07		27	#NUM!	#NUM!	#NUM!	#NUM!
08						
09						
10						
11						

NEXT, we are going to identify the outliers by identifying any values that are 3 standard deviations or more below or above the mean using the statistics that we have already calculated using the formulas above. We will do this by first calculating two values as in the example below:

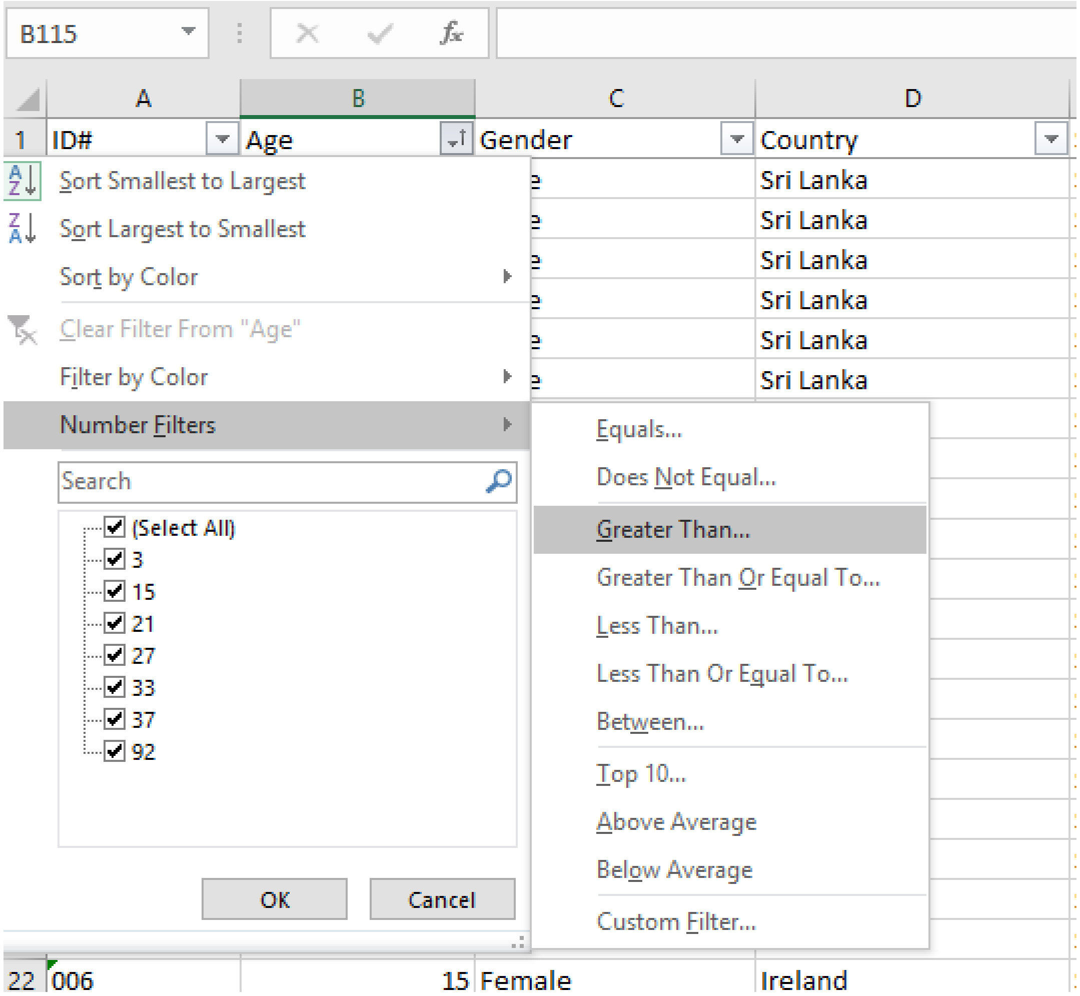
1. '*Outlier_high*': the sum of the mean and three times the standard deviation (MEAN+3*STDEV); any value greater this value is clearly an outlier
2. '*Outlier_low*': three times the standard deviation subtracted from the mean (MEAN+3*STDEV); any value less than this value is an outlier (in some cases, you might get a negative number in which case you do not need to worry about outliers in the low end of the distribution)



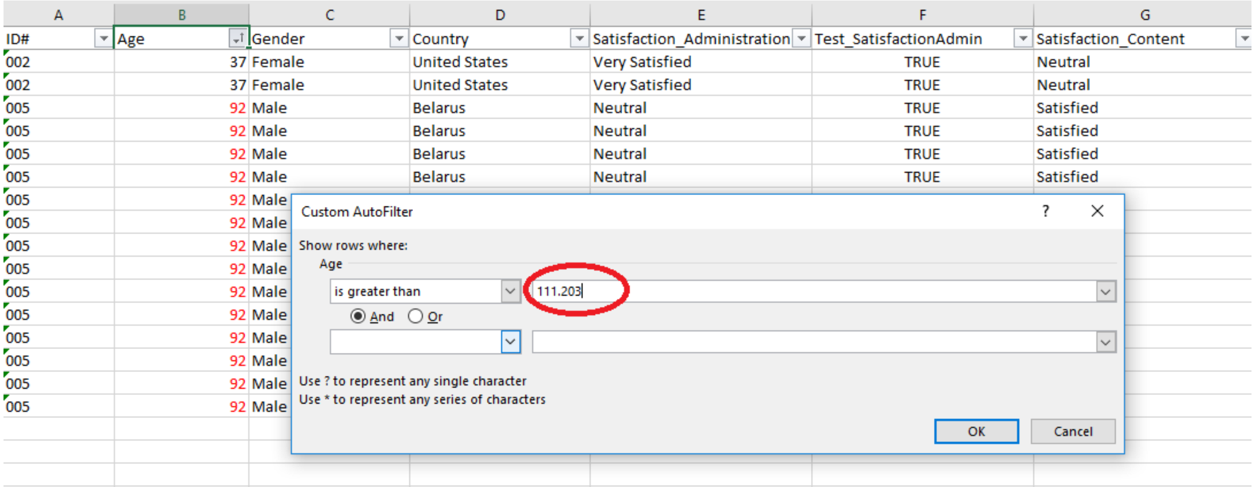
107	MEAN	32.62
108	STDEV	26.19457196
109	MIN	3
110	MAX	92
111	MEDIAN	27
112		
113	OUTLIER_HIGH	=B107+3*B108
114	OUTLIER_LOW	=B107-3*B108
115		
116		
117		

Next, we will identify the outliers using the ‘Sort’ button

1. Go to the Filter button
2. Select ‘Numbers Filter’
3. Select ‘Greater Than…’



4. You’ll get a pop up window like the one below. Enter the number you got for *Outlier_high*



ID#	Age	Gender	Country	Satisfaction_Administration	Test_SatisfactionAdmin	Satisfaction_Content
002	37	Female	United States	Very Satisfied	TRUE	Neutral
002	37	Female	United States	Very Satisfied	TRUE	Neutral
005	92	Male	Belarus	Neutral	TRUE	Satisfied
005	92	Male	Belarus	Neutral	TRUE	Satisfied
005	92	Male	Belarus	Neutral	TRUE	Satisfied
005	92	Male	Belarus	Neutral	TRUE	Satisfied
005	92	Male				
005	92	Male				
005	92	Male				
005	92	Male				
005	92	Male				
005	92	Male				
005	92	Male				
005	92	Male				
005	92	Male				
005	92	Male				

Custom AutoFilter

Show rows where:

Age

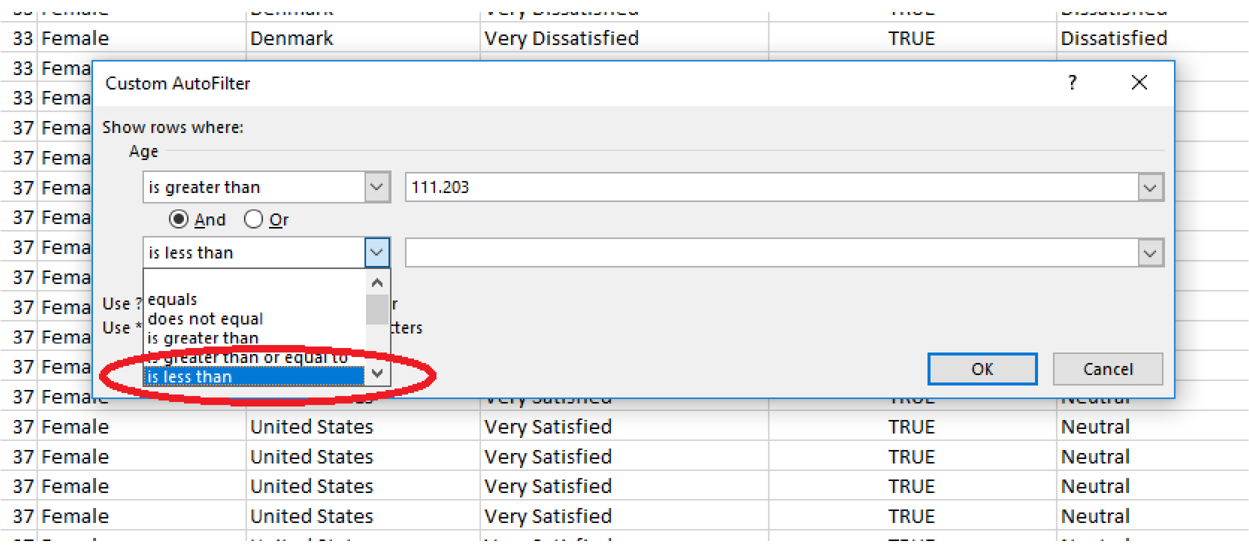
is greater than 111.203

● And ○ Or

Use ? to represent any single character

Use * to represent any series of characters

5. If *Outlier_low* is a positive value (i.e. non-negative; if it's negative skip this step and go to step 6), in the second box, scroll down to 'less than'



33	Female	Denmark	Very Dissatisfied	TRUE	Dissatisfied
33	Fema				
33	Fema				
37	Fema				
37	Fema				
37	Fema				
37	Fema				
37	Fema				
37	Fema				
37	Fema				
37	Fema				
37	Fema				
37	Female				
37	Female	United States	Very Satisfied	TRUE	Neutral
37	Female	United States	Very Satisfied	TRUE	Neutral
37	Female	United States	Very Satisfied	TRUE	Neutral
37	Female	United States	Very Satisfied	TRUE	Neutral

Custom AutoFilter

Show rows where:

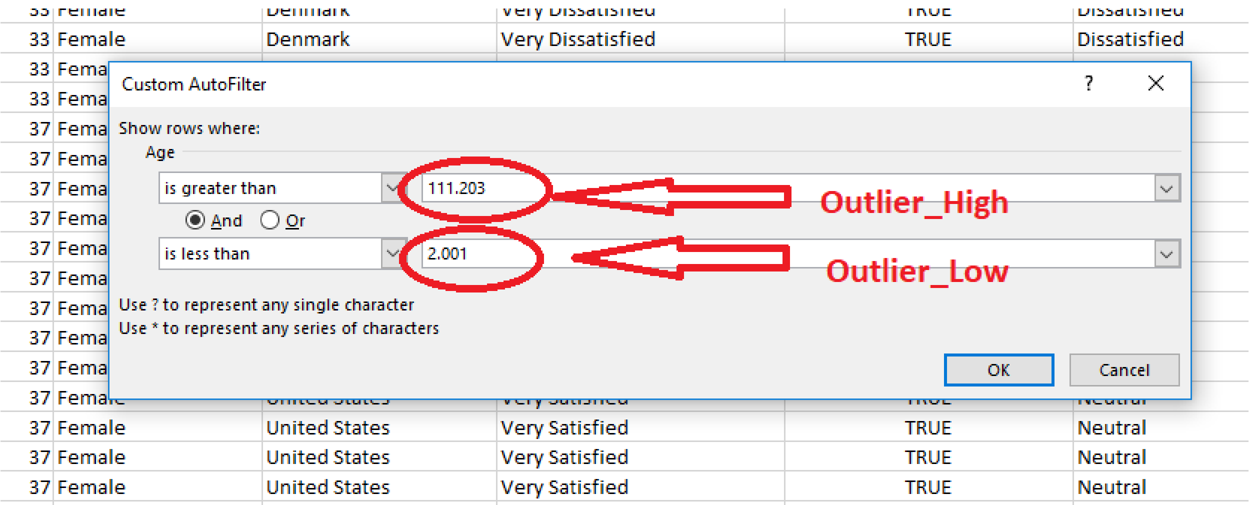
Age

is greater than 111.203

● And ○ Or

is less than

6. Enter the value for *Outlier_low* into the corresponding box and click ok



33	Female	Denmark	Very Dissatisfied	TRUE	Dissatisfied
33	Female	Denmark	Very Dissatisfied	TRUE	Dissatisfied
33	Fema				
33	Fema				
37	Fema				
37	Fema				
37	Fema				
37	Fema				
37	Fema				
37	Fema				
37	Fema				
37	Fema				
37	Fema				
37	Female	United States	Very Satisfied	TRUE	Neutral
37	Female	United States	Very Satisfied	TRUE	Neutral
37	Female	United States	Very Satisfied	TRUE	Neutral
37	Female	United States	Very Satisfied	TRUE	Neutral

Custom AutoFilter

Show rows where:

Age

is greater than | 111.203

● And ○ Or

is less than | 2.001

Use ? to represent any single character

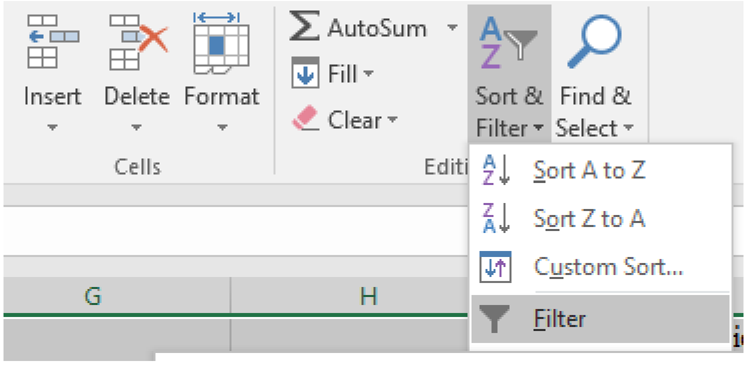
Use * to represent any series of characters

7. Now you should have only values that are 3 standard deviations above or below the mean, these are the outliers (If there are no values, no problem!).
8. Highlight these values (for each different type of issue, use a different color, when reviewing your data at the end of the cleaning process, it'll make it easier to remember what color corresponds to what issue)
9. Repeat the above steps for numeric variables.
10. And finally, we now move to the last step of data cleaning.

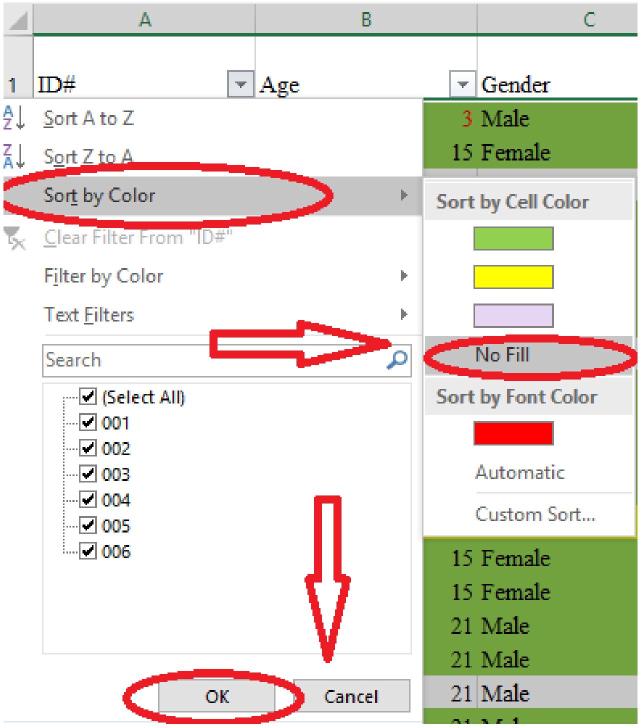
2.7 The very last step of data cleaning

Phew! That was a lot of work! The rule of thumb is that for any data work you will most likely spend 90% of the time cleaning it and 10% on the analysis, but that 90% is vitally important because otherwise the data that you are analyzing is full of errors and are not valid.

1. Copy your whole worksheet using CTRL+A and CTRL+C
2. Insert a new worksheet. Name this worksheet 'Final'
3. Copy this the data into the new worksheet using CTRL+V
4. Apply a Filter to Row 1 (the headers/variable names)



5. Using any of the variables, select 'Sort by Color' → 'No Fill'. This will move all of your highlighted rows to the bottom of the excel sheet and shows you how many data points you would have left if you delete all the data points with problems.



6. Selecting how many observations to delete should be based on your judgment. Hopefully you kept track of which color meant what and you can remove data points in the following order:
    a. Duplicates
    b. Illogical values
    c. Observations with significant missing values
    d. Outliers

At each step, go back and check whether you still have a decent number of observations to do analysis (if you have 30 survey responses and if by deleting all of the above, you are left with only 15, you might want to stop after removing duplicates)

7. Delete the original columns with the typos (remember you should have a new column _fixed with the corrected values)
8. If you have elected to keep some of the data with issues, remove highlighting.

Now you're ready to do some analysis!!!

STEP 3: ANALYZE YOUR DATA

3.1 Descriptive Statistics

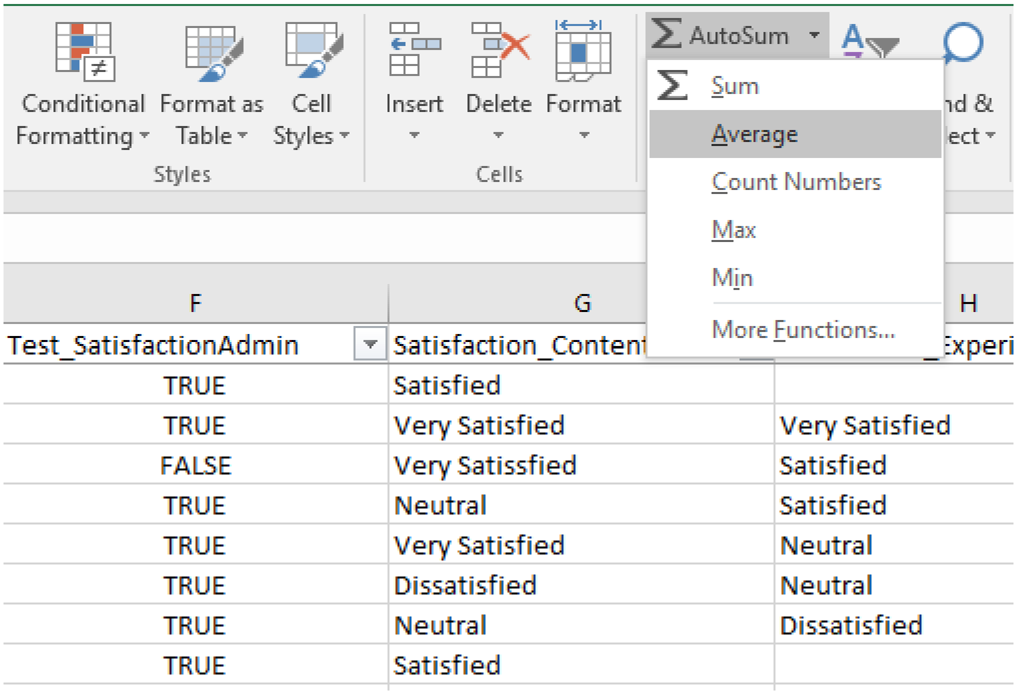
The following describes the different Excel functions that you can use to calculate simple descriptive statistics of your data (note that this is only applicable to numeric variables).

We'll use primarily the following functions, which follow the same formula as previous Excel functions reviewed above:

Descriptive Statistic	Formula	Description
Mean	=AVERAGE(B2:B8)	Returns the mean (or average) of cells B2 through B8
Minimum Value	=MIN(B2:B8)	Returns the lowest value in the cells B2 through B8
Maximum Value	=MAX(B2:B8)	Returns the largest value in the cells B2 through B8
Standard Deviation	=STDEV.P(B2:B8)	Returns the standard deviation for cells B2 through B8.
Median	=MEDIAN(B2:B8)	Returns the median for cells B2 through B8.

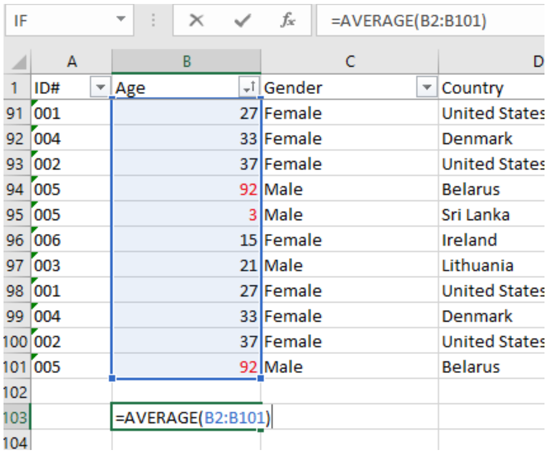
Now, we are going to calculate these statistics for the numeric variables in our data:

1. Scroll all the way down to the bottom of your dataset, and
2. Select a cell two rows below your last data point
3. For the Average, Min, and Max functions:
    a. Go to the top right hand corner
    b. Click on the AutoSum Button
    c. Select Average, Min or Max



Test_SatisfactionAdmin	Satisfaction_Content	_Experi
TRUE	Satisfied	
TRUE	Very Satisfied	Very Satisfied
FALSE	Very Satissfied	Satisfied
TRUE	Neutral	Satisfied
TRUE	Very Satisfied	Neutral
TRUE	Dissatisfied	Neutral
TRUE	Neutral	Dissatisfied
TRUE	Satisfied	

Your formula should look something like this (Excel will automatically select the data for you):

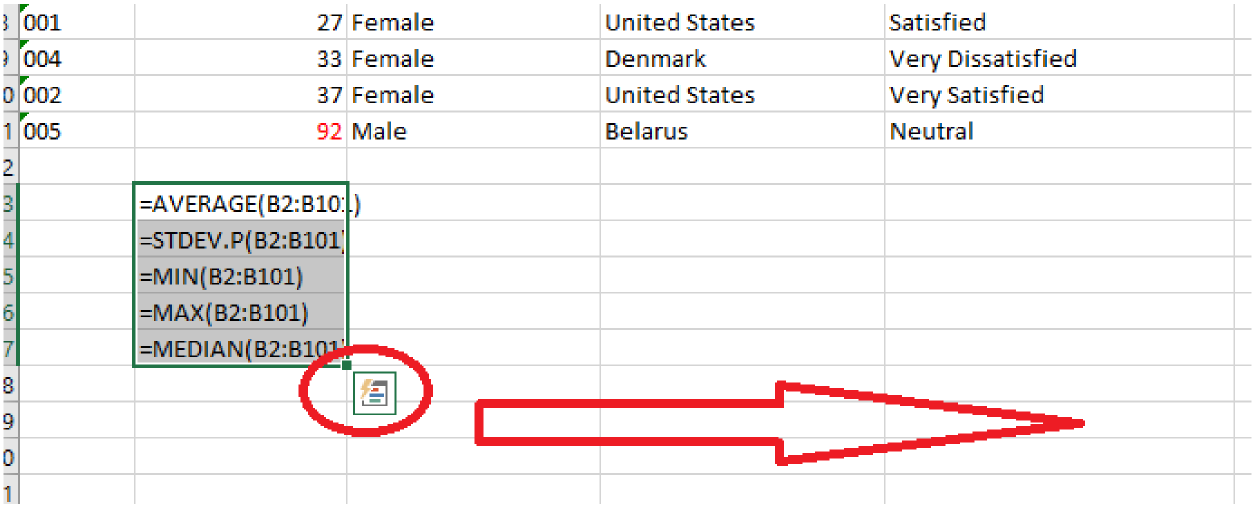


IF | =AVERAGE(B2:B101)

	A	B	C	D
1	ID#	Age	Gender	Country
91	001	27	Female	United States
92	004	33	Female	Denmark
93	002	37	Female	United States
94	005	92	Male	Belarus
95	005	3	Male	Sri Lanka
96	006	15	Female	Ireland
97	003	21	Male	Lithuania
98	001	27	Female	United States
99	004	33	Female	Denmark
100	002	37	Female	United States
101	005	92	Male	Belarus
102				
103		=AVERAGE(B2:B101)		
104				

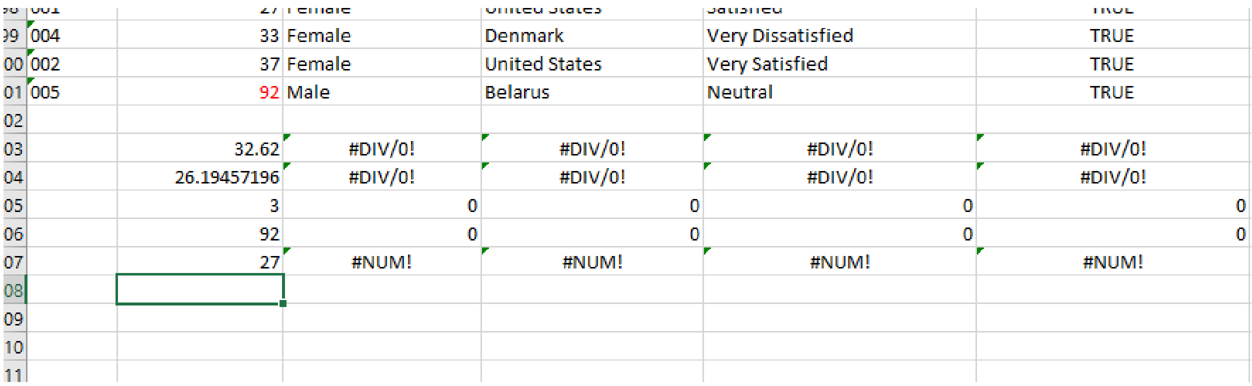
4. Hit Enter
5. To calculate the median and standard deviation, you'll need to type those in
   a. Select a cell right below your last formula
   b. Type in the formula and replace the cell numbers (in red) using those completed by the AutoSum button for the other formulas.
      i. =STDEV.P(B2:B8)
      ii. =MEDIAN(B2:B8)
6. Hit Enter

Once you've got all four formulas for the different statistics for the first variable, rather than hand-entering for each variable, you can just drag the formula across the worksheet (this is called the autofill feature):



001	27	Female	United States	Satisfied
004	33	Female	Denmark	Very Dissatisfied
002	37	Female	United States	Very Satisfied
005	92	Male	Belarus	Neutral
	=AVERAGE(B2:B101)			
	=STDEV.P(B2:B101)			
	=MIN(B2:B101)			
	=MAX(B2:B101)			
	=MEDIAN(B2:B101)			

You might get something like this but do not panic! Just delete the formulas for the columns that have text (non-numeric) values:

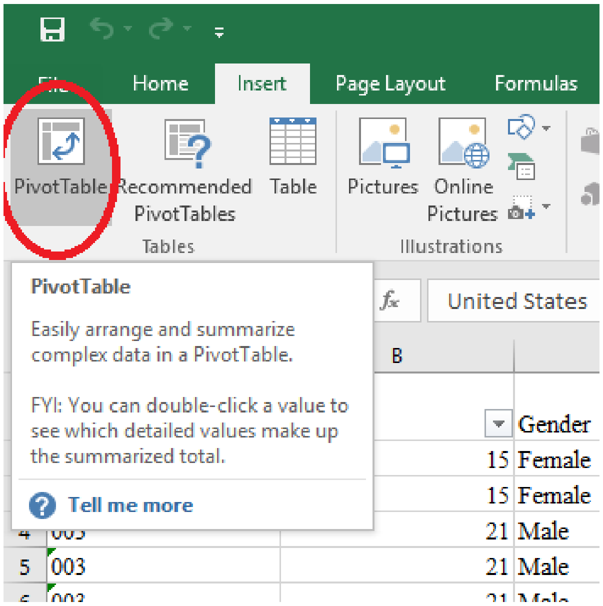


98	001	27	Female	United States	Satisfied	TRUE
99	004	33	Female	Denmark	Very Dissatisfied	TRUE
00	002	37	Female	United States	Very Satisfied	TRUE
01	005	92	Male	Belarus	Neutral	TRUE
02						
03		32.62	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
04		26.19457196	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
05		3	0	0	0	0
06		92	0	0	0	0
07		27	#NUM!	#NUM!	#NUM!	#NUM!
08						
09						
10						
11						

3.2 Counts and sub-counts

Now we are going to learn about another Excel feature called a pivot table, which allows you to reorganize and summarize selected rows and columns of data in a spreadsheet or database table to obtain a desired report. It is one of the most powerful features of Excel.

1. Select any cell in the source data
2. On the Insert tab of the ribbon, click the PivotTable button



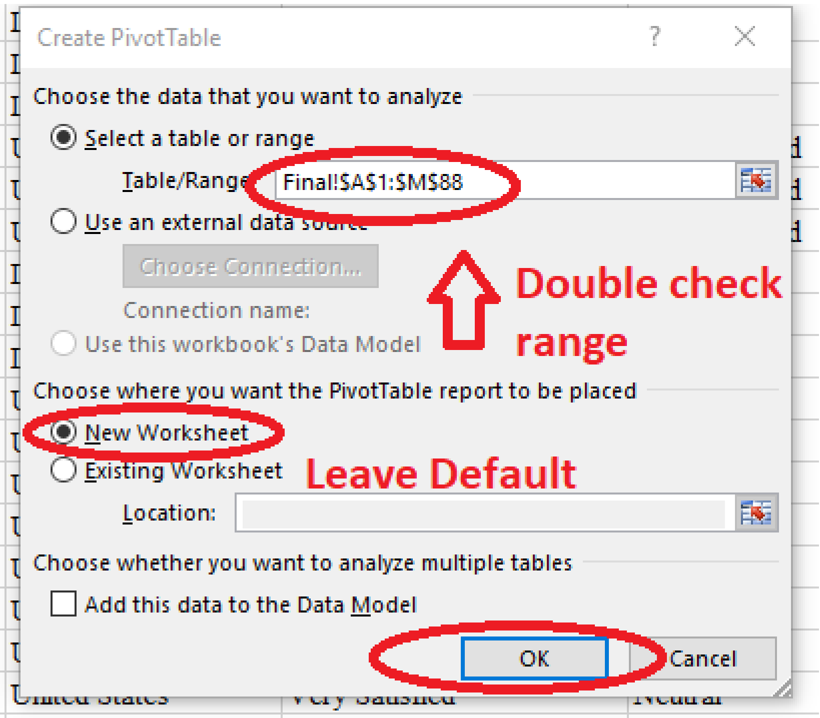
PivotTable

Easily arrange and summarize complex data in a PivotTable.

FYI: You can double-click a value to see which detailed values make up the summarized total.

Tell me more

3. A dialog box called ‘Create PivotTable’ comes up. The default is to select all the data in the worksheet (if you have basic statistics at the bottom of the worksheet, make sure to remove these before the pivot table) and to insert the pivot table in a new worksheet; Click OK



Create PivotTable

Choose the data that you want to analyze

● Select a table or range

Table/Range: Final!$A$1:$M$88

○ Use an external data source

Choose Connection...

Connection name:

○ Use this workbook's Data Model

Choose where you want the PivotTable report to be placed

● New Worksheet

○ Existing Worksheet

Location:

Choose whether you want to analyze multiple tables

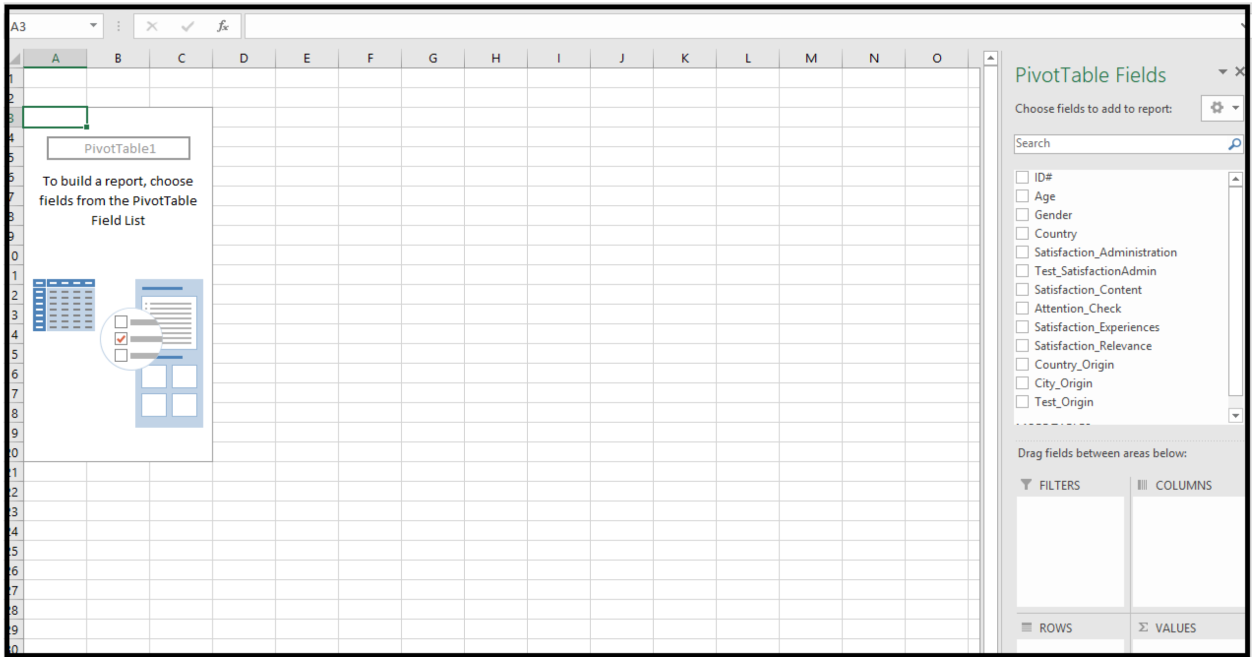
□ Add this data to the Data Model

OK Cancel

Double check range

Leave Default

4. Your blank pivot table will look like this:



To build a report, choose fields from the PivotTable Field List

PivotTable Fields

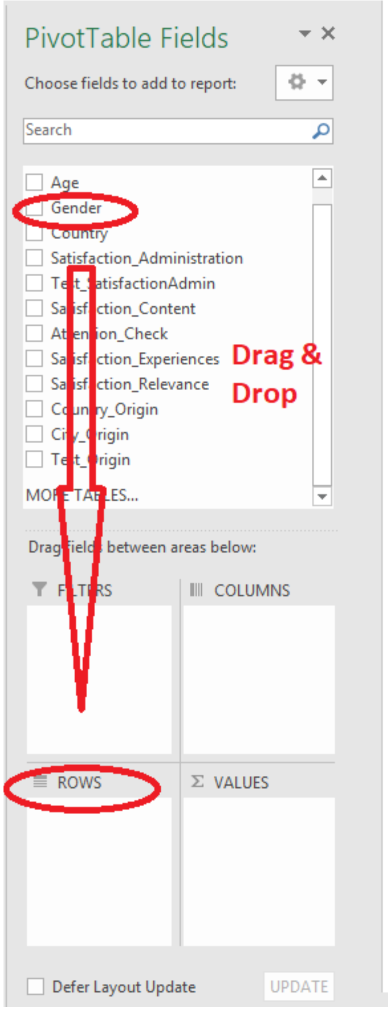
Choose fields to add to report:

Search

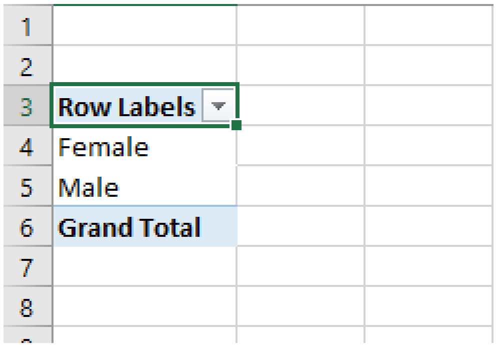
□ ID#
□ Age
□ Gender
□ Country
□ Satisfaction_Administration
□ Test_SatisfactionAdmin
□ Satisfaction_Content
□ Attention_Check
□ Satisfaction_Experiences
□ Satisfaction_Relevance
□ Country_Origin
□ City_Origin
□ Test_Origin

Drag fields between areas below:

5. Now, say you want to know how many males and females responded to your survey
    a. In the right hand corner, drag and Drop ‘Gender’ from the PivotTable Fields and drop it under ‘Rows’

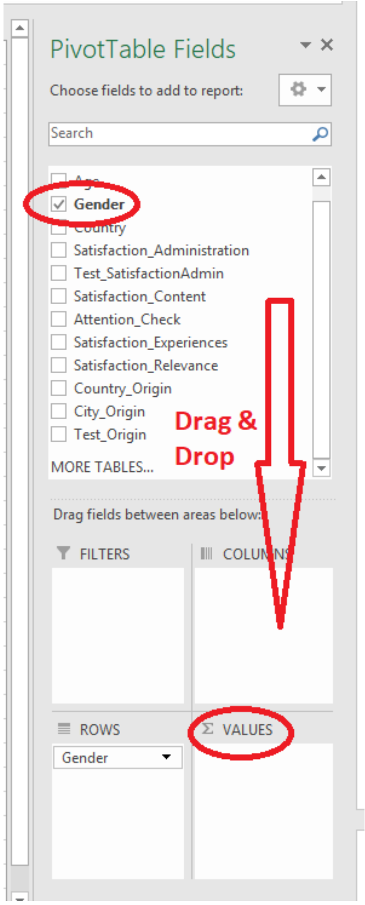


b. You should get something like this in the left hand corner

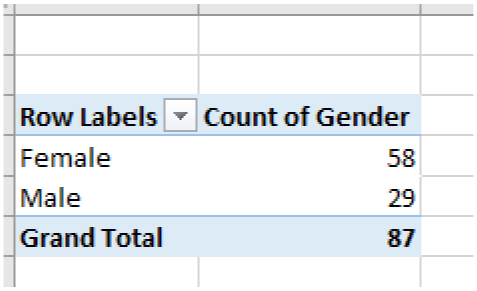


1			
2			
3	**Row Labels** ▼		
4	Female		
5	Male		
6	**Grand Total**		
7			
8			

c. Next, in the right hand corner, you'll repeat the drag and drop step with 'Gender' into the Values Box

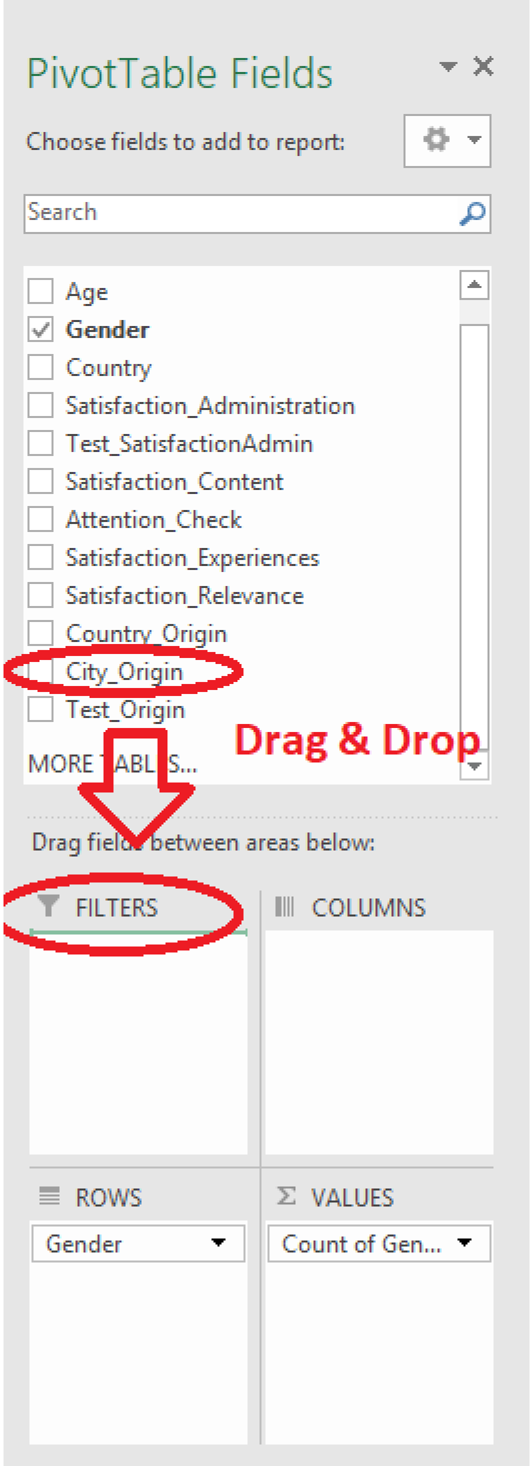


d. Now you should have a table with the counts Male and Female survey respondents in the upper right hand corner

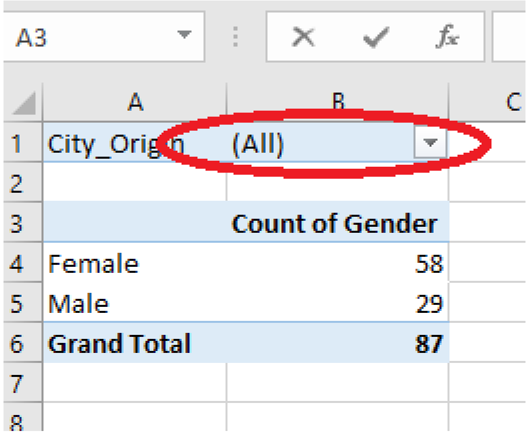


Row Labels	Count of Gender
Female	58
Male	29
Grand Total	**87**

6. Now say you want to know the number of males and females who responded to your survey who are from Bishkek. Basically, you want to look at a sub-group of the group you surveyed.
   a. Drag and drop the variable *City_Origin* into the ‘Filters’ box in the PivotTable Fields

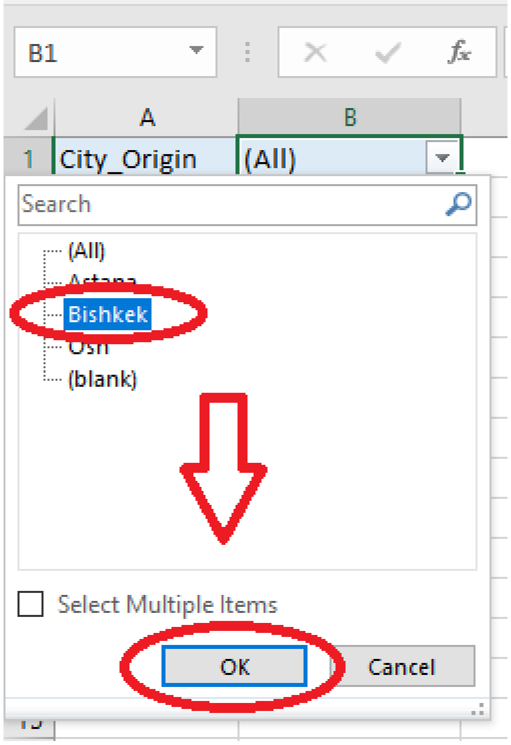


b. Now you should have something that looks like this with your new variable at the top *City_Origin*. The ‘All’ tells you that right now the table includes all values for *City_Origin*



A3

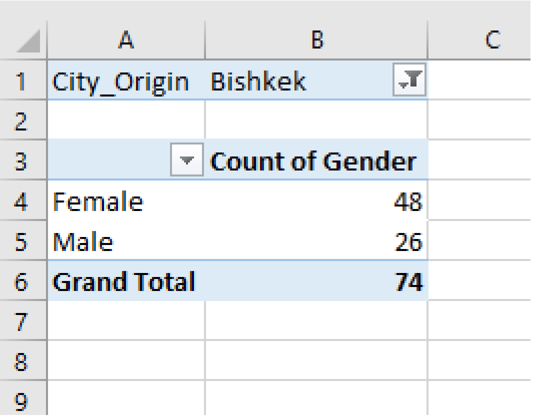
	A	B	C
1	City_Origin	(All)	
2			
3		Count of Gender	
4	Female	58	
5	Male	29	
6	**Grand Total**	**87**	
7			
8			



B1

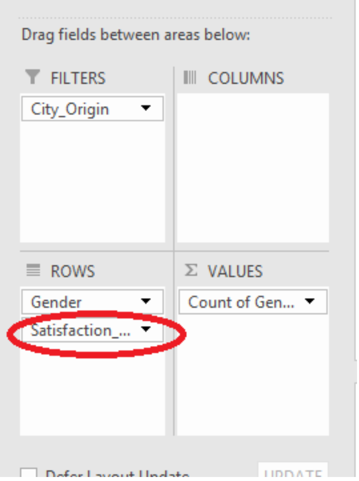
	A	B
1	City_Origin	(All)

c. Select Bishkek and this should automatically update your table



	A	B	C
1	City_Origin	Bishkek	
2			
3		Count of Gender	
4	Female	48	
5	Male	26	
6	**Grand Total**	**74**	
7			
8			
9			

d. Now you're looking at just the numbers of males and females respondents who are from Bishkek



Drag fields between areas below:

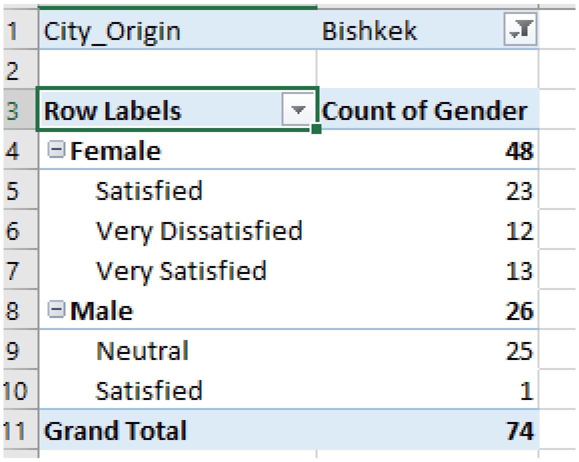
FILTERS: City_Origin

COLUMNS:

ROWS: Gender, Satisfaction_...

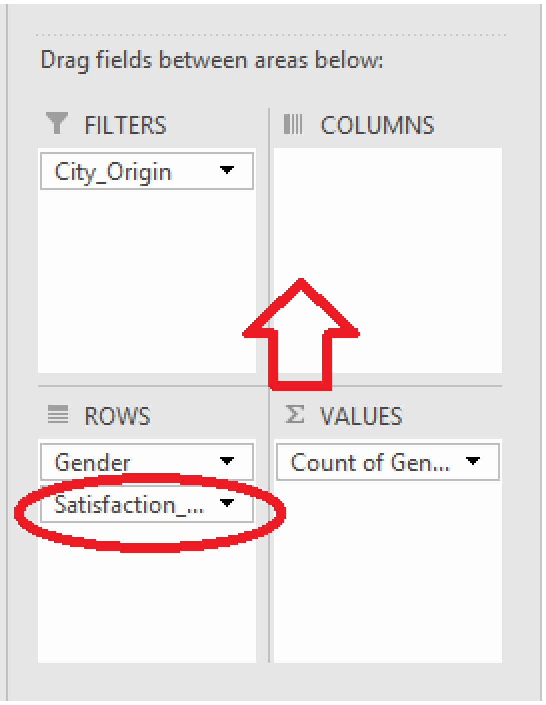
VALUES: Count of Gen...

7. Now you are interested in knowing whether females and males from Bishkek responded differently about their satisfaction with the administration of the program
   a. You'll drag and drop the variable you're interested in, in this case *Satisfaction_Administration* into the Rows box right beneath *Gender*
   b. This should update your table automatically



1	City_Origin	Bishkek
2		
3	**Row Labels**	**Count of Gender**
4	**⊟Female**	**48**
5	Satisfied	23
6	Very Dissatisfied	12
7	Very Satisfied	13
8	**⊟Male**	**26**
9	Neutral	25
10	Satisfied	1
11	**Grand Total**	**74**

c. You can also update the way the orientation of the table by moving any of your variables to the 'Columns' box



Drag fields between areas below:

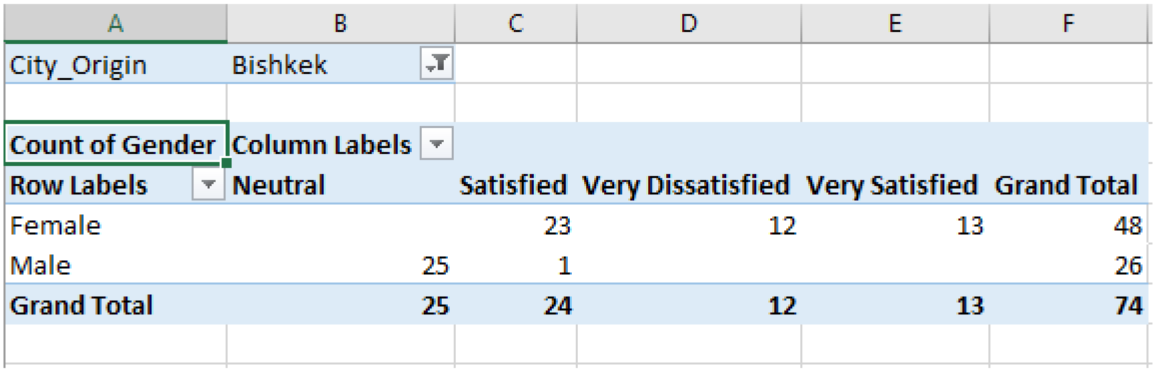
FILTERS: City_Origin

COLUMNS

ROWS: Gender, Satisfaction_...

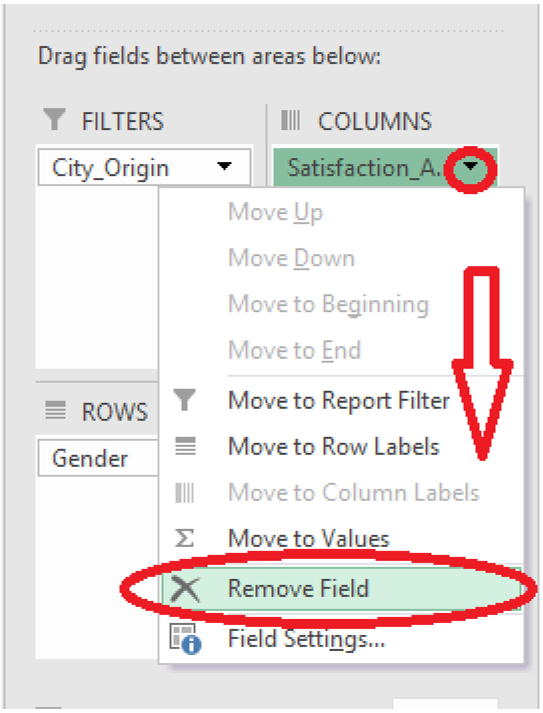
VALUES: Count of Gen...

d. Simply moving your variables to the 'Columns' box gives you a different view



A	B	C	D	E	F
City_Origin	Bishkek				
Count of Gender	Column Labels				
Row Labels	Neutral	Satisfied	Very Dissatisfied	Very Satisfied	Grand Total
Female		23	12	13	48
Male	25	1			26
Grand Total	25	24	12	13	74

e. If you want to go back to the very first table, you can just remove the variables from the different boxes by selecting ‘Remove field’ from the drop down and your table continues to automatically update.

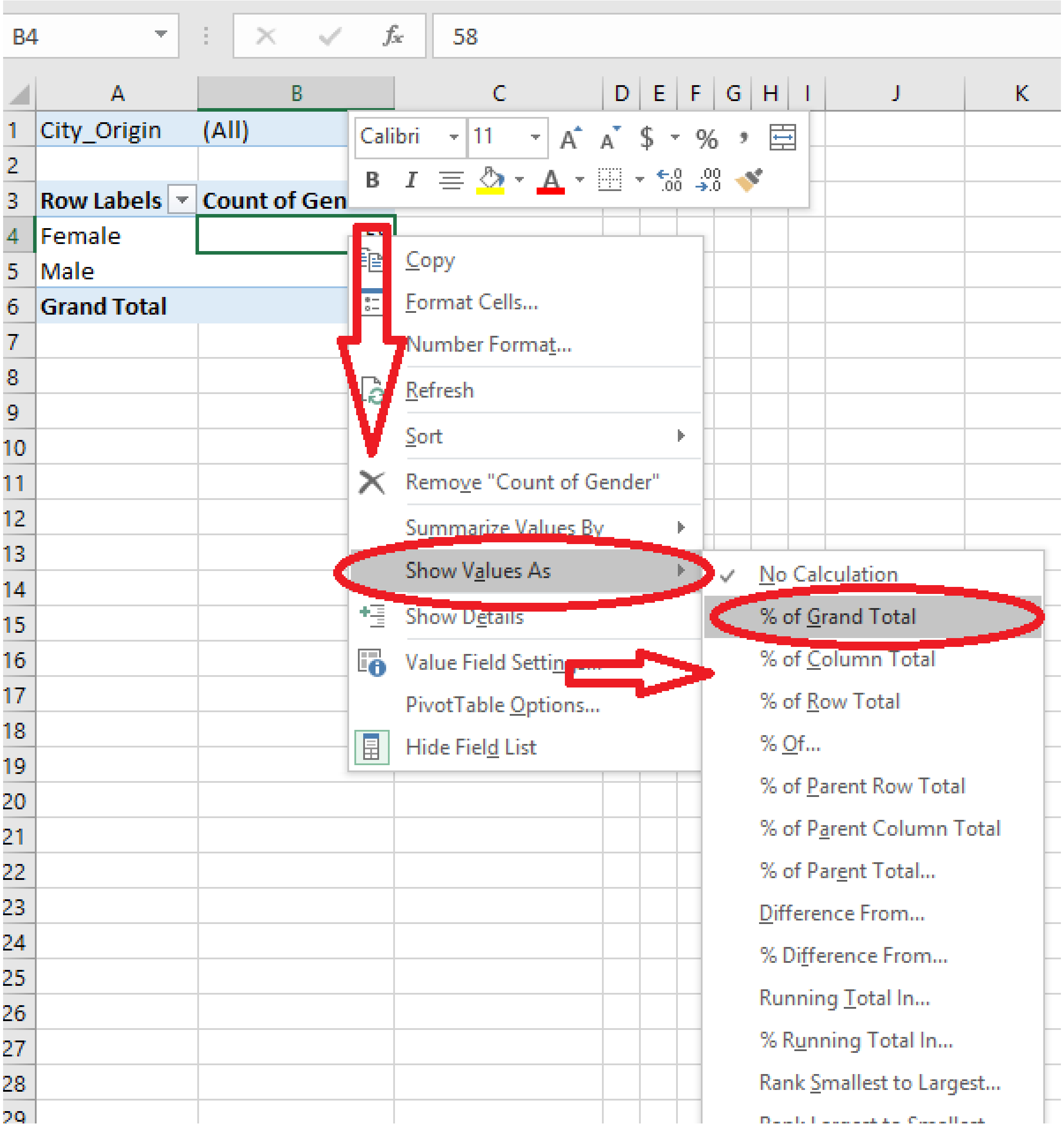


Now you know how to slice and dice your data for counts and sub-counts.

3.3 Percentages

So now you’re interested in understanding the percentage of women and men who responded to your survey.

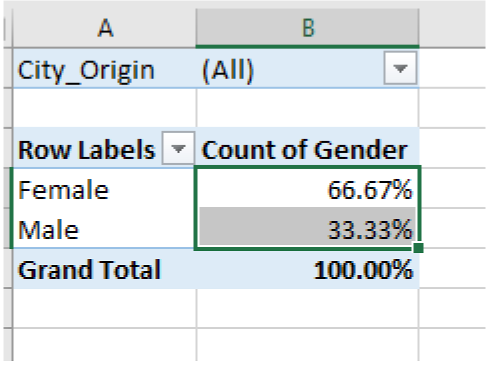
1. Select the value(s) you would like to see as percentages → Select ‘Show Values As’ → % of Grand Total



B4 58

	A	B
1	City_Origin	(All)
2		
3	Row Labels	Count of Gen
4	Female	
5	Male	
6	Grand Total	

2. And then Excel automatically updates the values, Tada!

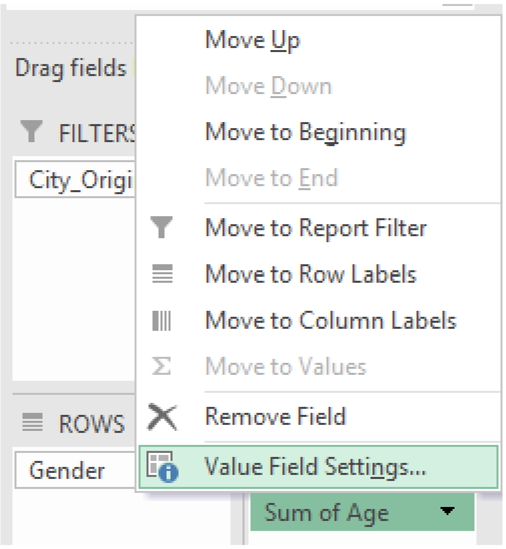


A	B
City_Origin	(All)
Row Labels	Count of Gender
Female	66.67%
Male	33.33%
Grand Total	100.00%

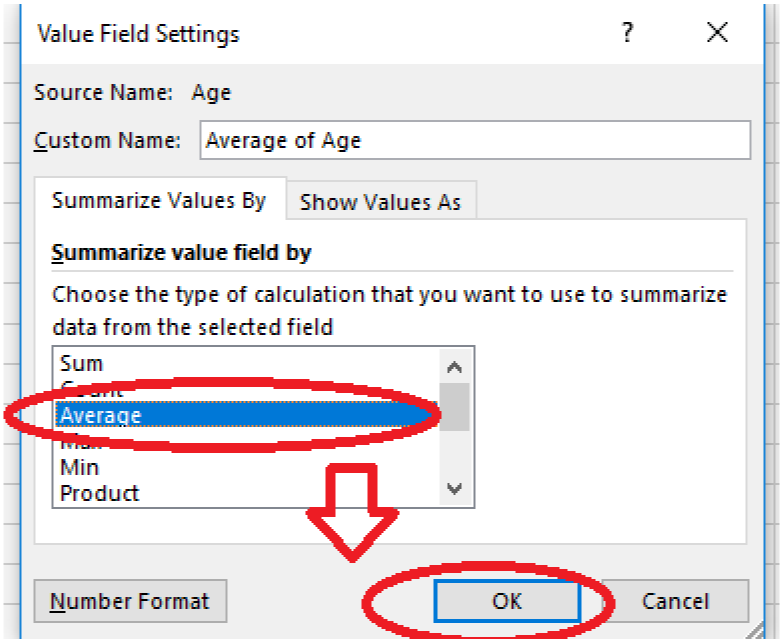
3.4 Averages

Now you're interested in knowing what the average age for different groups of respondents (males and females)

1. Drag and drop 'Age' into the 'Values' Box (the default for excel is to add all numeric values)
2. Click on the dropdown for Age → Click on 'Value Field Settings…'



3. In the dialog box, select 'Average' → Click OK



Value Field Settings

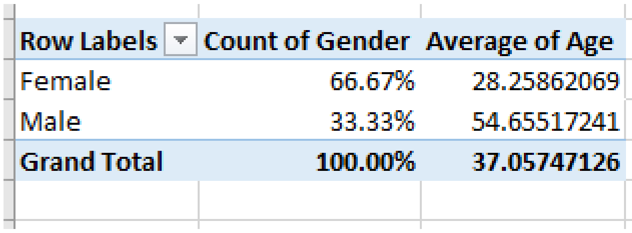
Source Name: Age

Custom Name: Average of Age

Summarize value field by

Choose the type of calculation that you want to use to summarize data from the selected field

4. Now you have a column with the average age of male, female as well as the average age for all participants



Row Labels	Count of Gender	Average of Age
Female	66.67%	28.25862069
Male	33.33%	54.65517241
Grand Total	**100.00%**	**37.05747126**

As you can see, Excel Pivot Tables are infinitely flexible: you can add several rows, columns with different permutations

CONGRATULATIONS! YOU MADE IT TO THE END! YOU ARE NOW AN EXCEL WHIZ!!!

ADDITIONAL RESOURCES TO LEARN ABOUT EXCEL[5]

ExcelJet: https://exceljet.net/

Excel Easy: https://www.excel-easy.com

UCLA Institute for Digital research and Education: https://stats.idre.ucla.edu

UCLA Institute for Digital research and Education Tips for Excel:
https://stats.idre.ucla.edu/other/mult-pkg/faq/general/tips-for-excel/

[5] This does not represent an endorsement, but are rather resources that members of the Evaluation Division have consulted and found useful in their work.

SOURCES

Bailey, B. (2017, June 14). Data Cleaning 101. Retrieved January, 2018 from https://towardsdatascience.com/data-cleaning-101-948d22a92e4

Broomell, S. (August-December, 2016) Empirical Research Methods, (class, Carnegie Mellon University, Department of Social and Decision Sciences, Pittsburgh, PA,).

Bruns, D. (2014, July 30). Excel Pivot Tables. Retrieved January, 2019, from https://exceljet.net/things-to-know-about-excel-pivot-tables

Cheusheva, S. (2018, August 23). Using logical functions in Excel: AND, OR, XOR and NOT. (2018, August 23). Retrieved January, 2019, from https://www.ablebits.com/office-addins-blog/2014/12/17/excel-and-or-xor-not-functions/#excel-and-function

Columbia Center for Teaching and Learning, QMSS e-Lessons Quantitative Methods in Social Sciences. Retrieved December, 2018, from http://ccnmtl.columbia.edu/projects/qmss/measurement/validity_and_reliability.html

GCFLearnFree.org. Excel XP – Inserting and Deleting Rows and Columns. Retrieved January, 2018 from https://edu.gcfglobal.org/en/excelxp/inserting-and-deleting-rows-and-columns/1/

Data Science: An introduction/Definitions of Data. Retrieved December, 2018, from https://en.wikibooks.org/wiki/Data_Science:_An_Introduction/Definitions_of_Data

Excel Easy. Pivot Tables in Excel. Retrieved January, 2019, from https://www.excel-easy.com/data-analysis/pivot-tables.html

Statistics How To. Elementary Statistics for the rest of us!. Retrieved December, 2018, from https://www.statisticshowto.datasciencecentral.com/

Stapel, E. Mean, Median, Mode, and Range. Retrieved January, 2019, from https://www.purplemath.com/modules/meanmode.htm

World Bank DIME Wiki. Checklist: Data Cleaning. Retrieved August, 2018, from https://dimewiki.worldbank.org/wiki/Checklist:_Data_Cleaning

World Bank DIME Wiki. Data Cleaning. Retrieved August, 2018, from https://dimewiki.worldbank.org/wiki/Data_Cleaning

Qualtrics. What is a Survey? Retrieved January, 2019, from https://www.qualtrics.com/experience-management/research/survey-basics/

Quartz. Quartz/bad-data guide. Retrieved January, 2018, from https://github.com/Quartz/bad-data-guide/blob/master/README.md#there-are-inexplicable-outliers