

Overview

How can I find and summarize tennis player data?

Goals

By the end of this case, you will be able to efficiently query data in Excel while utilizing common operations such as filtering, indexing, matching, and conditional functions. You will also learn several key Excel functions, including:

- [VLOOKUP\(\)](#)
- [HLOOKUP\(\)](#)
- [INDEX\(\)](#)
- [MATCH\(\)](#)
- [SUMPRODUCT\(\)](#)
- [SUMIF\(\)](#)
- [COUNTIF\(\)](#)
- [COUNTIFS\(\)](#)

Introduction

Business Context. You are an analyst for the United States Tennis Association (USTA) and have been tasked with analyzing the results of a local tournament that took place in 2020 in the state of Pennsylvania. Your findings could potentially be used by senior business executives to make changes to the way the tour is run in subsequent years.

Business Problem. You need to set up an Excel workbook to efficiently find specific information and calculate various summary statistics about the player data.

Analytical Context. Your boss has provided you with an Excel workbook file ([data.xlsx](#)), which contains information for all tennis matches played during the tournament, as well as historical information about player rankings. You will work on it to calculate the required figures.

Columns and Rows

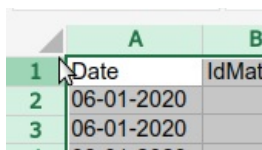
It's rows and columns all the way

As you've seen in previous cases, Excel's formatting is based on intersecting rows and columns. One thing to note is that the first row is often considered special and is known as the **header row**, which gives names to the data contained in each column. The `data.xlsx` file has several sheets with information about matches, players, and counties. The `Matches` sheet holds information about particular matches, such as the player names, internal numeric IDs, the date of the match, etc.

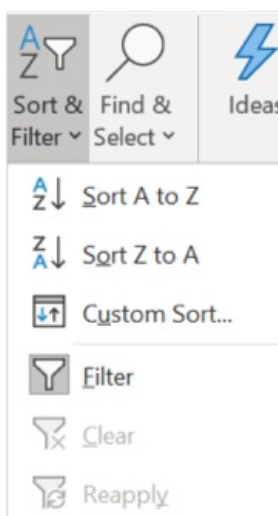
Filtering

Filters help us manually locate the data we need without altering it. To create filters in any given sheet, follow these steps:

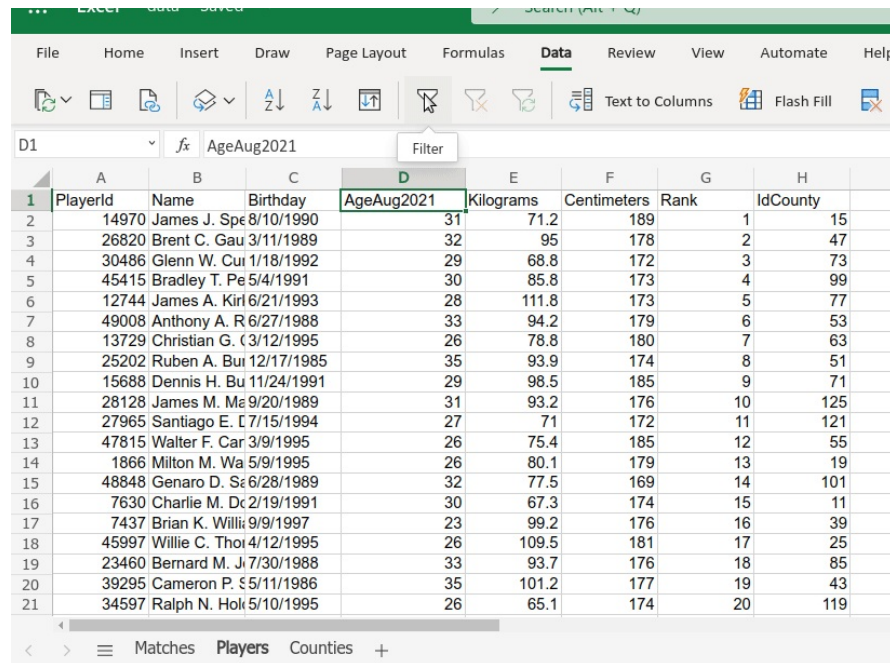
1. Select the whole table you want to filter. A good shortcut, if the only thing on your sheet is the table, is to click the little triangle in between column A and row 1, which will select the whole sheet:



2. In the Home tab of the Toolbar, click the "Sort & Filter" button. This will open a drop-down menu. Click on "Filter".



In the image below, we can see how to filter Column D with the header AgeAug2021 in the Players worksheet so that we only keep players who are 30 years old. Doing this will add a little arrow on the header row for all columns. These arrows allow you to filter the table to find the exact data you need:



	A	B	C	D	E	F	G	H
	PlayerId	Name	Birthday	AgeAug2021	Kilograms	Centimeters	Rank	IdCounty
2	14970	James J. Spe	8/10/1990	31	71.2	189	1	15
3	26820	Brent C. Gau	3/11/1989	32	95	178	2	47
4	30486	Glenn W. Cui	1/18/1992	29	68.8	172	3	73
5	45415	Bradley T. Pe	5/4/1991	30	85.8	173	4	99
6	12744	James A. Kirl	6/21/1993	28	111.8	173	5	77
7	49008	Anthony A. R	6/27/1988	33	94.2	179	6	53
8	13729	Christian G. C	3/12/1995	26	78.8	180	7	63
9	25202	Ruben A. Bui	12/17/1985	35	93.9	174	8	51
10	15688	Dennis H. Bu	11/24/1991	29	98.5	185	9	71
11	28128	James M. Ma	9/20/1989	31	93.2	176	10	125
12	27965	Santiago E. C	7/15/1994	27	71	172	11	121
13	47815	Walter F. Car	3/9/1995	26	75.4	185	12	55
14	1866	Milton M. Wa	5/9/1995	26	80.1	179	13	19
15	48848	Genaro D. Se	6/28/1989	32	77.5	169	14	101
16	7630	Charlie M. De	2/19/1991	30	67.3	174	15	11
17	7437	Brian K. Willi	9/9/1997	23	99.2	176	16	39
18	45997	Willie C. Thor	4/12/1995	26	109.5	181	17	25
19	23460	Bernard M. Ji	7/30/1988	33	93.7	176	18	85
20	39295	Cameron P. S	5/11/1986	35	101.2	177	19	43
21	34597	Ralph N. Hol	5/10/1995	26	65.1	174	20	119

Exercise 1

How many players are exactly 18 years old?

Lookups

VLOOKUP & HLOOKUP

Filtering is one way to narrow down our results; however, the process of extracting the data we want is still very manual. How can we extract this data in a more automated way and use it elsewhere?

Enter the VLOOKUP() and HLOOKUP() functions. These names are short for “Vertical Lookup” and “Horizontal Lookup”. Since our data is in vertical tables, we will use VLOOKUP(). (HLOOKUP() works the same as VLOOKUP(), just with the axes flipped, so it should be a piece of cake after this).

One example of something we can look up is the name of a player given his internal ID. Let’s create a new column in the worksheet Matches with the name of the winner.

The VLOOKUP() function has three mandatory arguments and one optional argument:

1. The cell containing the value we will look up to help us find more information.
2. The table in which to look for the value. This range of cells *must* have the value we are looking for in its first column.
3. The number of the column where we can find the value we need, counting from the first column in our range.
4. Optionally, we can tell the function we want an exact match by putting the FALSE keyword at this position

Here’s an illustrative example:

Formula bar: `=VLOOKUP(F3;A1:C5;3;FALSE)`

	A	B	C	D	E	F	G
1	ID	Name	Subject				
2	123	Myriam	Algebra			ID	Subject
3	456	James	Statistics			789	Engineering
4	789	Camilla	Engineering				
5	101112	Adam	Data analytics				
6							
7							
8							

Now let’s walk through each of the above for our specific goal:

1. The value we want to look up is contained in cell D2.

2. For the range in which we need to look for the player name, we will use the sheet *Players*, and we will take the range of cells starting with A2 and ending with B346. Note that we start our range in column A since it is the one that contains the player IDs. We end it in column B because it contains the players' names, but we could also end it in any other column after K if we wanted data that was farther to the right.
3. Column B contains our desired data; since column A contains our lookup value, B is the 2nd column.
4. We want an exact match, so we use FALSE as the last argument.

Thus, our final formula is:

```
=VLOOKUP(D2,Players!$A$2:$B$346,2,FALSE)
```

We can then fill the formula down and see the names of the winners of all matches:

	A	B	C	D	E	F
1	Date	IdMatch	DurationM	IdWinner	IdLoser	WinnerName
2	06-01-2020	333	121	14970	45415	James J. Spears
3	06-01-2020	332	107	26820	30486	Brent C. Gauthier
4	06-01-2020	331	184	14970	12744	James J. Spears
5	06-01-2020	330	116	30486	49008	Glenn W. Cunningham
6	06-01-2020	329	134	45415	31665	Bradley T. Peace
7	06-01-2020	328	89	26820	25202	Brent C. Gauthier
8	06-01-2020	327	165	14970	962	James J. Spears
9	06-01-2020	326	92	30486	3867	Glenn W. Cunningham
10	06-01-2020	325	142	12744	25105	James A. Kirk
11	06-01-2020	324	92	49008	46494	Anthony A. Rodriguez
12	06-01-2020	323	206	13729	31665	Christian G. Chang
13	06-01-2020	322	69	25202	15448	Ruben A. Burr
14	06-01-2020	321	148	15688	45415	Dennis H. Burt
15	06-01-2020	320	96	26820	11293	Brent C. Gauthier
16	06-01-2020	319	90	14970	1866	James J. Spears
17	06-01-2020	318	112	28128	30486	James M. Manuel
18	06-01-2020	317	149	14970	27965	James J. Spears
19	06-01-2020	316	140	30486	48848	Glenn W. Cunningham
20	06-01-2020	315	65	14970	16817	James J. Spears
21	06-01-2020	314	94	30486	29329	Glenn W. Cunningham

Exercise 2

Create a new column in the *Matches* worksheet with the names of the losers of each match.

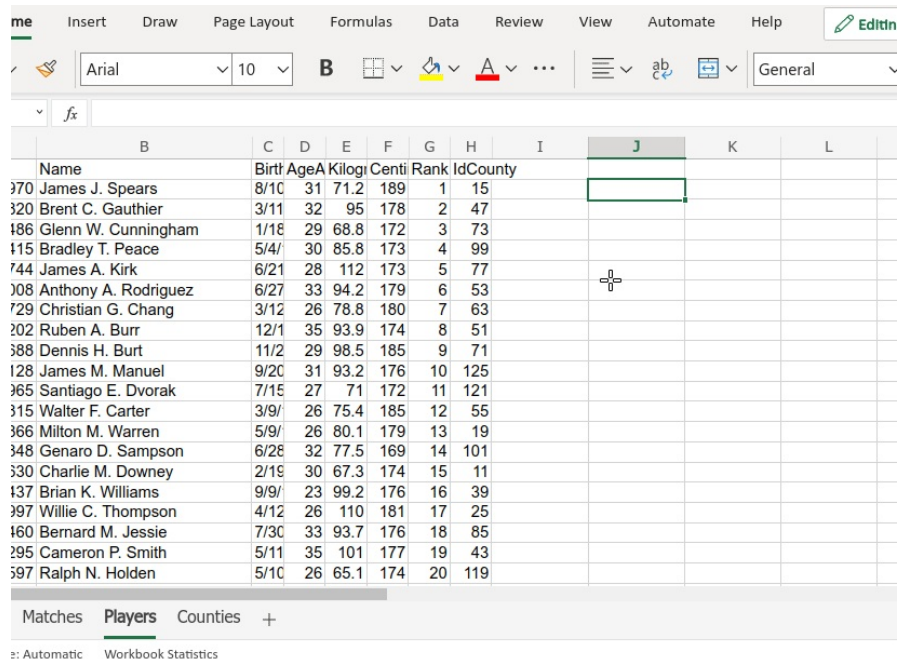
Other Ways to Search

Now we know how to use `VLOOKUP()` and `HLOOKUP()` to find information in a table. But they have some limitations - first, they can only look for data in a column that is *after* the initial column in a range. Second, they will return only the first match in the range every time, so if we want the data after the first match, we are out of luck. This means `VLOOKUP()` is best suited for tables where each row has a *unique identifier*, and that identifier is in the very first column.

INDEX() + MATCH()

A more flexible (and all-around better) way to look up information is with the combination of the `INDEX()` and `MATCH()` functions. Using a combination of these, you can look into columns *before* the column with your initial value and look for values *beyond* the first that satisfy your set of conditions. The only downside is that these are a little harder to grasp at first, but as we progress, you will soon become comfortable with them.

`INDEX()` gives you the result of a value in a range given the range and the position of the value in that range. It also works for horizontal ranges, and it even works for ranges with multiple columns and rows! Its arguments are the cell range where you can find the value and the position of the relevant cell in that range.



The screenshot shows the Microsoft Excel interface. The ribbon at the top includes 'me', 'Insert', 'Draw', 'Page Layout', 'Formulas', 'Data', 'Review', 'View', 'Automate', and 'Help'. The 'Formulas' tab is active, showing the formula bar with `=INDEX(Sheet1!$B:$L,MATCH(J1,$B:$B,0),0)`. Below the formula bar, a table of player statistics is displayed. The table has columns for Name, Birth, Age, Kilogr, Centi, Rank, Id, and County. The data is sorted by Rank in ascending order. The 'Players' tab is selected in the bottom navigation bar.

	B	C	D	E	F	G	H	I	J	K	L
	Name	Birth	Age	Kilogr	Centi	Rank	Id	County			
370	James J. Spears	8/10	31	71.2	189	1	15				
320	Brent C. Gauthier	3/11	32	95	178	2	47				
186	Glenn W. Cunningham	1/18	29	68.8	172	3	73				
115	Bradley T. Peace	5/4/	30	85.8	173	4	99				
744	James A. Kirk	6/21	28	112	173	5	77				
108	Anthony A. Rodriguez	6/27	33	94.2	179	6	53				
729	Christian G. Chang	3/12	26	78.8	180	7	63				
202	Ruben A. Burr	12/1	35	93.9	174	8	51				
388	Dennis H. Burt	11/2	29	98.5	185	9	71				
128	James M. Manuel	9/20	31	93.2	176	10	125				
365	Santiago E. Dvorak	7/15	27	71	172	11	121				
315	Walter F. Carter	3/9/	26	75.4	185	12	55				
366	Milton M. Warren	5/9/	26	80.1	179	13	19				
348	Genaro D. Sampson	6/28	32	77.5	169	14	101				
330	Charlie M. Downey	2/19	30	67.3	174	15	11				
137	Brian K. Williams	9/9/	23	99.2	176	16	39				
397	Willie C. Thompson	4/12	26	110	181	17	25				
160	Bernard M. Jessie	7/30	33	93.7	176	18	85				
295	Cameron P. Smith	5/11	35	101	177	19	43				
597	Ralph N. Holden	5/10	26	65.1	174	20	119				

Matches **Players** Counties +

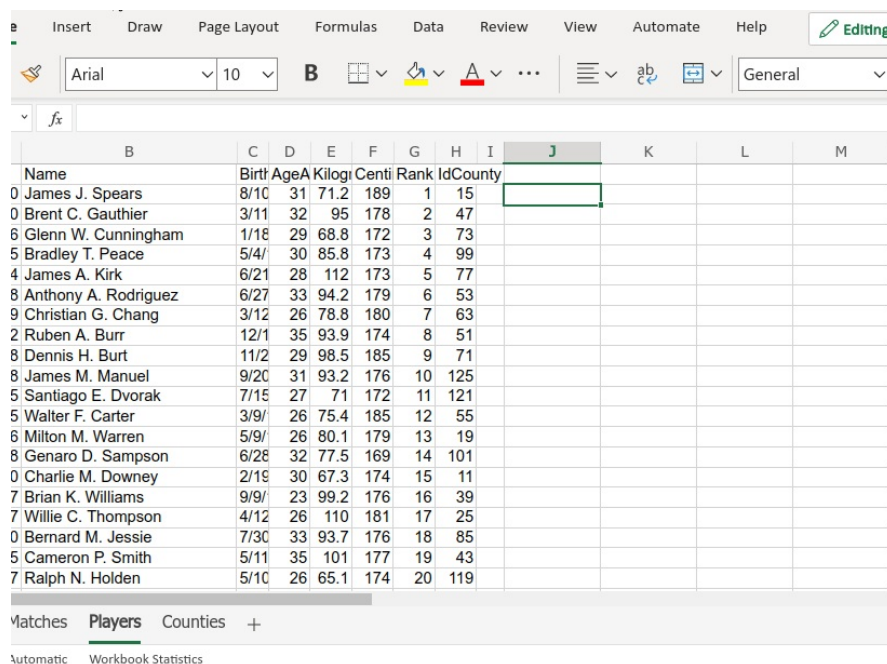
Automatic Workbook Statistics

This enables us to find the value of a particular cell in a range by looking at its position in that range. But how can we find the value of a cell if we don't know its position?

`MATCH()` gets the position of any value in a horizontal or vertical range. It needs three arguments:

1. The value to look for.
2. The horizontal or vertical range in which to look for it.
3. (Optionally) The number 0 if you want to find an exact match, or 1 if an approximate match is enough. This is important! For most cases, you probably want to use 0 for an exact match.

`MATCH()` will return the position of your value in the list if it exists:



	B	C	D	E	F	G	H	I	J	K	L	M
	Name	Birth	Age	Kilogr	Centi	Rank	Id	County				
0	James J. Spears	8/10	31	71.2	189	1	15					
0	Brent C. Gauthier	3/11	32	95	178	2	47					
6	Glenn W. Cunningham	1/18	29	68.8	172	3	73					
5	Bradley T. Peace	5/4	30	85.8	173	4	99					
4	James A. Kirk	6/21	28	112	173	5	77					
8	Anthony A. Rodriguez	6/27	33	94.2	179	6	53					
9	Christian G. Chang	3/12	26	78.8	180	7	63					
2	Ruben A. Burr	12/1	35	93.9	174	8	51					
8	Dennis H. Burt	11/2	29	98.5	185	9	71					
8	James M. Manuel	9/20	31	93.2	176	10	125					
5	Santiago E. Dvorak	7/15	27	71	172	11	121					
5	Walter F. Carter	3/9	26	75.4	185	12	55					
6	Milton M. Warren	5/9	26	80.1	179	13	19					
8	Genaro D. Sampson	6/28	32	77.5	169	14	101					
0	Charlie M. Downey	2/19	30	67.3	174	15	11					
7	Brian K. Williams	9/9	23	99.2	176	16	39					
7	Willie C. Thompson	4/12	26	110	181	17	25					
0	Bernard M. Jessie	7/30	33	93.7	176	18	85					
5	Cameron P. Smith	5/11	35	101	177	19	43					
7	Ralph N. Holden	5/10	26	65.1	174	20	119					

Matches **Players** Counties +

Automatic Workbook Statistics

Now, how do we combine these two functions in order to do a full lookup? In place of the positional arguments of `INDEX()`, we can use `MATCH()`. So, instead of telling `INDEX()` in what row to look for the cell, we let `MATCH()` find the row for us.

For instance, let's find the winner's ranking for each match:

	A	B	C	D	E	F	G	H
1	Date	IdMatch	Duration	IdWinner	IdLoser	WinnerName	LoserName	WinnerRank
2	06-01-2020	333	121	14970	45415	James J. Spears	Bradley T. Peace	1
3	06-01-2020	332	107	26820	30486	Brent C. Gauthier	Glenn W. Cunningham	2
4	06-01-2020	331	184	14970	12744	James J. Spears	James A. Kirk	1
5	06-01-2020	330	116	30486	49008	Glenn W. Cunningham	Anthony A. Rodriguez	3
6	06-01-2020	329	134	45415	31665	Bradley T. Peace	Walter J. Parker	4
7	06-01-2020	328	89	26820	25202	Brent C. Gauthier	Ruben A. Burr	2
8	06-01-2020	327	165	14970	962	James J. Spears	Andrew M. Shank	1
9	06-01-2020	326	92	30486	3867	Glenn W. Cunningham	Kenneth J. Broyle	3
10	06-01-2020	325	142	12744	25105	James A. Kirk	Donald M. West	5
11	06-01-2020	324	92	49008	46494	Anthony A. Rodriguez	Ronald V. Wright	6
12	06-01-2020	323	206	13729	31665	Christian G. Chang	Walter J. Parker	7
13	06-01-2020	322	69	25202	15448	Ruben A. Burr	Rocco T. Hayden	8
14	06-01-2020	321	148	15688	45415	Dennis H. Burt	Bradley T. Peace	9
15	06-01-2020	320	96	26820	11293	Brent C. Gauthier	David C. Sandus	2
16	06-01-2020	319	90	14970	1866	James J. Spears	Milton M. Warren	1
17	06-01-2020	318	112	28128	30486	James M. Manuel	Glenn W. Cunningham	10
18	06-01-2020	317	149	14970	27965	James J. Spears	Santiago E. Dvor	1
19	06-01-2020	316	140	30486	48848	Glenn W. Cunningham	Genaro D. Sampa	3
20	06-01-2020	315	65	14970	16817	James J. Spears	Charles B. Berns	1
21	06-01-2020	314	94	30486	29329	Glenn W. Cunningham	Antonio B. Englar	3

The formula we write in cell H2 is

```
=INDEX(Players!$A$2:$H$346,MATCH(D2,Players!$A$2:$A$346,0),7)
```

Let's break it down:

- `MATCH(D2,Players!A2:A346,0)` gets the position of the value in cell D2 (the winner's ID in the Matches worksheet) in the Players worksheet.
- `=INDEX(Players!A2:H346,<RESULT FROM MATCH>,7)` takes the row number that the `MATCH()` function generated and passes it as an argument to the `INDEX()` function, which retrieves the value in the 7th column of the Players worksheet (the ranking) that corresponds to that row number. Here, `<RESULT FROM MATCH>` is saying that what goes here is the result from the `MATCH()`. Enclosing text between a `<` and a `>` is a customary way of inserting comments when talking about formulas or codes in computer science. For instance, you could also say that the syntax of the `INDEX()` function is `=INDEX(<range>,<row number>,<column number>)`.

Exercise 3

Find all the losers' ranks.

Note: The advantage of using this approach, with `INDEX()` and `MATCH()` instead of `VLOOKUP()`, is that we can look for values in any row or column, which means that the position of the column with the unique identifier does not matter; so it's more flexible.

Summarizing

Summarizing data

Imagine we want to know how a player's age affects their chances of winning a match. We can find the answer to these questions using **aggregation functions**, which take several values and give us a single result. You have already seen some basic examples of this in previous cases with functions like `SUM()` and `AVERAGE()`.

Another useful function is **SUMPRODUCT()**. `SUMPRODUCT()` takes two ranges of the same size, calculates the product of corresponding cells in each range (i.e. the product of the first cells in each range, the product of the second cells in each range, etc.), then sums up all of these products.

One of the most common uses of this function is for finding a weighted average. You are probably familiar with weighted averages from your grade school days - your grade was divided into several deliverables, each of which had a different weight, and your final grade was decided by the weighted average of all the deliverables.

The syntax to calculate the weighted average is as follows:

```
=SUMPRODUCT(<range of weights>,<range to be  
averaged>)/SUM(<range of weights>)
```

A weighted average is different from a normal average - don't confuse them!

Exercise 4

Find the weighted average of the amount of winner serve points with the minutes played in each match as the weight. How do you interpret the result?

Conditional Functions

Sometimes you will only want to average values for a particular player, a particular country, or for any other particular *condition*. **Conditional functions** allow you to do just that - you provide them conditions, and they ignore any value along a range that does not fulfill that condition.

AVERAGEIF () is one of these functions and takes two mandatory arguments and one optional one:

1. The range on which to check the criteria or condition.
2. The condition to check.
3. (Optional) The range on which to average. If this is not provided, then the operation is carried out on the range given in the first argument.

For example, let's use AVERAGEIF () to find the average duration of matches played where a player from Bradford County won. You can see the formula in cell A1 of the AverageDuration sheet:

```
=AVERAGEIF(Matches!J2:J1463,"=Bradford County",Matches!C2:C1463)
```

Conditions are given in double quotes. If you wanted to check if the values in a range are above 5, you would type ">5" as the condition. If you want to check if the values are less than or equal to 30, you would type "<=30". In the example above, we check the condition "=Bradford County" along column J, which holds the county code for the winner of each match. We also gave the optional third argument with column C, which holds the number of minutes in the match.

Exercise 5

Find the average duration in minutes for matches played where a player from Fulton County won.

More Conditional Functions

More conditional functions include `SUMIF()`, `COUNTIF()`, and `COUNTIFS()`. `SUMIF()` and `COUNTIF()` have similar syntax to `AVERAGEIF()`. For `COUNTIFS()`, you can include up to 127 conditional statements! It is not necessarily good practice to do this, as it would be difficult to keep track of so many conditions in a single function. The formula is:

```
=COUNTIFS(criteria_range1, criteria1, criteria_range2, criteria2,...)
```

Where the ... represents the inclusion of additional conditions.

Use this formula to calculate the number of matches that lasted longer than 120 minutes where the winner had fewer than 60 winner serve points:

```
=COUNTIFS(Matches!C:C, ">120", Matches!L:L, "<60")
```

Which should yield a count of 153.

Exercise 6

Calculate the number of matches won by a player from Franklin County where the match lasted longer than 100 minutes and there were more than 80 winner serve points.

Further Reading

There are two additional functions related to `VLOOKUP`: `HLOOKUP` (introduced earlier) and `XLOOKUP`.

▼ **Read more about `HLOOKUP`, and try an exercise, here.**

`HLOOKUP(lookup_value, table_array, row_index_num, [range_lookup])`

- `lookup_value`: The value to search for in the first row of the table.
- `table_array`: The range of cells that contains the data.
- `row_index_num`: The row number in the `table_array` from which to return a value.
- `range_lookup`: This is optional. If set to `TRUE` or omitted, it will find an approximate match. If set to `FALSE`, it will find an exact match.

For instance, `=HLOOKUP("Banana", A1:D4, 3, FALSE)` will search for "Banana" in the first row of the range A1:D4 and return a value from the third row. Since "FALSE" was used for the `range_lookup`, an exact match is returned.

To try it out, use the Matches sheet from the Tennis data sheet and find the duration of the match on row 1173.

▼ [Click here for the answer.](#)

```
=HLOOKUP("DurationMinutes", A1:K1463, 1173, FALSE)`
```

103 Minutes

▼ **Read more about `XLOOKUP`, and try an exercise, [here](#).**

XLOOKUP(lookup_value, lookup_array, return_array, [if_not_found], [match_mode], [search_mode])

- lookup_value: The value to search for.
- lookup_array: The range or array to search.
- return_array: The range or array from which to return values.
- [if_not_found]: Optional. Value to return if no match is found.
- [match_mode]: Optional. Specifies the match mode. Default is 0 (exact match). When an exact match cannot be found, #N/A will be returned. -1 is for an exact match or next smallest item, while 1 is for an exact match or next largest item.
- [search_mode]: Optional. Specifies the search mode. Default is 1 (search first-to-last). -1 is for a reverse search, last-to-first.

For instance, =XLOOKUP("Banana", A1:A10, B1:B10, "Not found") will search for "Banana" in the range A1:A10 and return the corresponding value from B1:B10. If an exact match cannot be found, "Not found" will be returned.

XLOOKUP is more flexible than HLOOKUP and VLOOKUP as it allows searching in any direction and can return values based on multiple criteria. It's available in Excel 365 and Excel 2019. If you have an older version of Excel, you may not have access to XLOOKUP.

Use the Players sheet from the Tennis data, and look up the name of the player who has ID 12839.

▼ [Click here for the answer.](#)

```
=XLOOKUP(12839, A:A, B:B)
```

Gabriel L. Brandon

Conclusions & Takeaways

In this case, we learned how to look for information in Excel using a variety of methods like filtering and looking up data with `VLOOKUP()`, `HLOOKUP()`, and `INDEX() + MATCH()`.

We also learned about applying aggregation functions to our data in order to extract more meaning out of it (with functions such as `SUMPRODUCT()` and `AVERAGEIF()`).

Here is a short summary document for review. You can download by clicking [HERE](#)