Finally, it is important to note that data cleaning and parsing shouldn't be considered a one-time job, but a reoccurring process. As you get to know your data, it often sparks new ideas and ways in which to transform the information. Once in the analysis stage, you may need to return to wrangling in order to parse the data differently after you've noticed an interesting insight or a stubborn roadblock.

Table 12.2.2: Summary Table of Excel Tools and Functions for Cleaning

	What it does	Location/Syntax
Text to Columns	Takes text in one or more cells and splits it across multiple cells.	Found on the Data Tab.
TRIM	Removes all spaces from a text string except for single spaces between words.	TRIM(text)
Find and Replace	Locates particular number of text string and replaces it with something else.	Found on the Home Tab, in the Editing section
MID	Returns a specific number of characters from a text string, starting at the position you specify, based on the number of characters you specify.	MID(text, start_num, num_chars)
FIND	Locates one text string within a second text string and returns the number of the starting position of the first text string from the first character of the second text string.	FIND(find_text, within_text, [start_num])
LOOKUP	Searches a row or column for a value.	LOOKUP(lookup_value, lookup_vector, [result_vector])

Skill Check Answers

1. 9 columns **2.** pet **3.** Rows 3 and 6 **4. a.** January: Marcel Canton; February: Becca Harold; March: William Benton; **b.** William Benton **5.** Answers could include student IDs or SSN numbers

12.2 Exercises

✔ CONCEPT CHECK

1. Data wrangling is the process of acquiring _____ data, cleaning it, organizing it, and transforming it so that the data is _____.

2. In Excel, the _____ function allows you to extract a defined number of characters from a string of text.

3. In Excel, the _____ function can find information in an array of data based on a specified identifier.

4. True or False: Excel reads a space as a character.

5. True or False: Pivot tables can summarize the information in a large data set into table.

PRACTICE

For each data file snippet, determine the delimiter that would be used to separate the data into columns and determine how many columns the data would be separated into.

6. Date,Time,PurchaseTotal,PaymentMethod
   01072021,1132,$523.45,Credit
   02072021,1623,$723.11,Paypal
   02072021,1704,$312.77,Credit

7. CustID;Age;Birthdate;Membership;AutoRenew;Flavor
   021789;44;07161976;Premium;Yes;Raspberry
   022893;30;01011990;Basic;Yes;Chocolate
   023113;35;03301985;Premium;No;Vanilla

8. ItemName Size Inventory Price LastReordered
   PinkShirtSS S 24 $18.99 07/23/21
   PinkShirtLS M 15 $19.99 03/12/21
   GreenShirtSS L 12 $18.99 05/15/21

9. Title,Author,Page Count,Word Count,Star Rating,Copies Sold (Millions),List Price
   A Grand Scheme,Lynn Minning,277,73682,4.2,1.73,$15.99
   The Princess Pickle,Susie Roadings,298,82546,3.8,2.77,$16.50
   The Scout,Allison Feist,308,81620,4.3,3.85,$14.99

Describe the inconsistencies in the data in each spreadsheet and explain the steps that should be taken to clean up the data.

10.

	A	B	C	D	E
1	Date	Time	Customer Name	Email Address	Subscription
2	11-10-2022	0732	Emmet Murphy		Premium
3	10.12.2022	1157	Susan Franklin	Sfranks@emails.com	Basic
4	12-10-2022	2005	Kyla Jones	no@no.no	Basic
5					
6	13-10-2022	1053	Ed	ed@edssite.co	Premium
7	10.13.2022	1442	Stephen Moore	SMoore@yahoos.com	Basic

11.

	A	B	C	D	E
1	Chapter	Status	Page Count	Approved	Editor
2	1	Finalized	72	No	Susan
3	2	Draft	55	No	Kyle
4	3	Editing	58	No	Antonio
5	3	Editing	58	No	Antonio
6	5	Draft	73		Susan
7	6	Finalized	60	Yes	Kyle
8	7	Finalized	63	Yes	Antono
9	8	Draft	59	No	Susan

12. Which of the following data will Excel find if you run the command "Find *ing" and select the "Match case" option?

a. swing **b.** JUMPING

c. Ingress **d.** ingested

13. Which of the following data will Excel find if you run the command "Find s?n"?

a. sand **b.** spin

c. sin **d.** stain

Determine the output of each Excel function.

14. MID("Once upon a time, there lived a lovely princess",6,6)

15. MID("The sum of twelve and twenty is thirty-two.",12,10)

16. Suppose the string "11/20/22 The Pillow Store $76.34" is in cell A5. What will the function MID(A5,10,FIND("$",A5)-10) return?

17. Suppose the string "051121 Okra Express $76.34" is in cell B3. What will the function MID(B3,8,FIND("$",B3)-8) return?

Use the provided pivot table to answer the questions.

18. The following pivot table displays a partial inventory of a clothing store, where the columns indicate the season the item of clothing is intended for.

Style and Color	Fall	Spring	Winter	Grand Total
Cardigan		**28**		**28**
Blue		8		8
Ivory		10		10
Pink		9		9
Purple		1		1
Coat	**14**		**14**	**28**
Black	6		5	11
Blue	4		4	8
Green	4		5	9
Jacket	**23**	**22**		**45**
Black	12	7		19
Blue	6	7		13
Gray	5	8		13
Grand Total	**37**	**50**	**14**	**101**

a. What is the total inventory of each type of clothing?

b. Which seasons are the clothing items intended for?

c. What color cardigans are in stock?

d. How many items in stock are intended for spring?

19. The following pivot table displays a summary of houses that are listed for sale by a real estate agent, where the columns indicate the number of bathrooms the house has.

Style and Bedrooms	1	2	2.5	Grand Total
Cape Cod	**16**	**24**		**40**
2	4			4
3	12	18		30
4		6		6
Colonial	**3**	**24**	**5**	**32**
2	2	1		3
3	1	13		14
4		10	5	15
Farmhouse	**6**	**14**		**20**
2	2			2
3	4	12		16
4		2		2
Ranch	**10**	**18**		**28**
2	2			2
3	8	12		20
4		6		6
Grand Total	**35**	**80**	**5**	**120**

a. What styles of house are listed as for sale by this real-estate agent?

b. How many Cape Cod style houses listed for sale have three bedrooms?

c. How many houses listed for sale have two bathrooms?

d. How many farmhouses are listed for sale?

20. The following spreadsheet is comprised of two tables containing order and customer information for an online art store. Determine the identifier that should be used to join these data tables together using the LOOKUP function in Excel.

	A	B	C	D	E	F	G
1	**Order Information**				**Customer Information**		
2	**Online Request**	**ID**	**Item**		**ID**	**Name**	**Sign-up Date**
3	202	1939	Kittens		1159	Abigail Berry	01-09-2020
4	203	1321	Flowers		1752	Adam Diaz	02-12-2019
5	204	1551	Springtime		1099	Ali Richardson	14-04-2020
6	205	1304	Kittens		1337	Daniel King	10-06-2018
7	206	1882	Ducks		1919	Jade Davis	12-02-2020
8	207	1099	Nighttime		1939	James Knight	05-04-2019
9	208	1182	Dusk		1321	Liam Smith	17-05-2020
10	209	1159	Eclipse		1551	Martha MacDonnell	02-11-2018
11	210	1529	Sunset		1529	Mira Ward	29-03-2021
12	211	1752	Flowers		1882	Olivia Skywell	11-08-2019
13	212	1337	Nighttime		1182	Omar Torres	12-02-2021
14	213	1919	Eclipse		1304	Skylar Davis	25-12-2021

21. The following partial spreadsheet is comprised of two tables containing book inventory and author information for an online bookstore. Determine the identifier that should be used to join these data tables together using the LOOKUP function in Excel.

	A	B	C	D	E	F	G
1	**Inventory**				**Author Information**		
2	**Inventory ID**	**Price**	**Book Title**		**Author ID**	**Author**	**Book Title**
3	175946854	15.99	Attack at Dusk		27789	L.M.Queen	Glossary of Slant Rhymes
4	175946855	14.79	A Speck of Thyme		27790	Diane Lufthaus	Attack at Dusk
5	175946856	12.99	Glossary of Slant Rhymes		27791	Patricia L. Black	A Midnight Meeting
6	175946857	14.99	Whining over Wine		27792	Kristine Phantom	A Sunset of Sparks
7	175946858	15.59	A Midnight Meeting		27793	T. Lewis Walsh	Return to Sender
8	175946859	13.99	A Chance of Snow		27794	Elena Miller	Whining over Wine
9	175946860	12.99	Last Year's Mistake		27795	Sophia M. Chen	A Chance of Snow
10	175946861	14.99	The Two Faces of Torrance		27796	L. Noah Song	A Speck of Thyme
11	175946862	15.99	Return to Sender		27797	Maxim Mandal	The Dashing Duchess
12	175946863	15.99	A Sunset of Sparks		27798	Alexander C. Ghosh	Last Year's Mistake
13	175946864	15.99	Where the River Ends		27799	Tomas Williams	Where the River Ends
14	175946865	15.99	The Dashing Duchess		27800	John Mark Castillo	The Two Faces of Torrance

WRITING & THINKING

22. Explain why it would be useful to have two different pivot tables of the same data set in a presentation.