CHAPTER 2 LABS - KEY

(Level 1 Header) Lab 2-1 Create a request for data extraction

Q1. Given that you are new and trying to get a grasp on Sláinte's operations, **list three questions related to sales** that would help you begin your analysis. For example, *how many products were sold in each state?*

Open-ended – no key provided.

Possible answers:

What is the highest selling product? How do quantities sold per product differ across states? What is average quantity of each product sold per day/per state/per month? What is the total quantity of each product sold per day/per state/per month?

Q2. Now **hypothesize the answers** to each of the questions. Remember, your answers don't have to be correct at this point. They will help you understand what type of data you are looking for. For example: *500 in Missouri, 6,000 in Pennsylvania, 4,000 in New York, etc.*

Open-ended – no key provided.

Q3. Finally, for each question, **identify the specific tables and attributes** that are needed to answer your questions. For example, to answer the question about state sales, you would need the [State] attribute which is most likely located in the [Customer] master table as well as a [Quantity Sold] attribute in a [Sales] table. If you had access to store or distribution center location data, you may also look for a [State] field there as well.

Open-ended – no key provided.

(Level 2 Header) Part 2: Generate a request for data

Now that you've identified the data you need for your analysis, complete a Data Request Form.

- 1. Open the Data Request Form
- 2. Enter your contact information.
- In the description field, identify the tables that you'd like to analyze, along with the time periods (e.g. past month, past year, etc.)
 Table Sales_Subset: Attributes: Customer_ID, Product_Code, Sales_Order_Quantity_Sold
 Table Customer_Table: Attributes: Customer_ID, Customer_St
- 4. Select a **frequency**. In this case this is a "One-off request".
- 5. Enter a request date (today) and a required date (one week from today)
- 6. Choose a **format** (spreadsheet).
- 7. Finally complete the **To be used in** box (internal analysis).
- 8. TAKE A SCREENSHOT (2-1) of your completed form.

(Level 2 Header) Part 3: Perform an analysis of the data

Q4. Take a moment and identify any attributes that you are missing from your original request that would be necessary to answer your original question of "How many products were sold in each state?".

Missing Customer_ID and Product_Code from the Sales_Subset table. Missing Customer_ID and Customer_St from Customer Table.

Q5. Evaluate your original questions and responses. Can you still answer the original question? *No*

Q6. Is there another question you could answer from the data Rachel provided?

Possible answers:

How many sales orders has each employee created? How many sales were created in the month of October? How much money was generated through sales for the entire period? How much money was generated through sales for the month of October?

END OF LAB

(Level 1 Header) Lab 2-2 Use PivotTables to de-normalize and analyze the data

(Level 2 Header) Part 1: Identify the Questions

Q1. Given Sláinte's request, **identify the data attributes and tables** needed to answer the question.

Sales_Subset: Product_Code, Sales_Order_Quantity_Sold

(Level 2 Header) Part 2: Master the data: Prepare data for analysis in Excel

Q2. When would it be a good idea to use a single table? Anytime all of the data you need are in a single table, there is no need to extract more than one table.

Alternative 2: Use the Excel Internal Data Model

1. **TAKE A SCREENSHOT (2-2a)** of the **Manage Relationships** window with both relationships created.

Manage F	Relationships		? ×
Status	Table *	Related Lookup Table	Liew
ation .	Salet Subret (Curtomer, D)	Customer (Customer, 10)	Augus Datact.
kist.	Sales_Subjet (Product_Code)	FG_hoduct(Product_Code)	Edit.
			≜da∧ata
			Descipate
			Delete
			Close

Q3. How comfortable are you with identifying primary key-foreign key relationships?

KEY: open-ended question, no key provided

Alternative 3: Merging the data into a single table using Excel Query Editor

13. Maximize the Query Editor window, and TAKE A SCREENSHOT (2-2b).

KEY Screenshot:

14	Climan - Cala	1 回面1	L) Deringen i Standardenser Standard			
1. cree	- IT, Fightmannets	w Annethington	Coll research and press		Column rates		Concernence and the Concernence of the second	CH Gaylette	
20.00		party second processory	PERMIT	1.4084		1.1.1.1. Y.A. ##		-m.	
		The second second	and the second s	-	- 77	10		1.0000000000000000000000000000000000000	
×	10	2001 Grown Multilary	ALC MARKET IS	10.00				There .	
	100	table included determinants	And insurance i	- April Market	- 11			104,348	
		THE COMPANY OF TAXABLE PARTY.	10.000	Section .				and the part later	
T	- M.	101 June professioner	1013 3 http://www.	parties				Jan Burner	
÷	- 44	Add company for	- paint in provide a protocial	(Annual Inc.	100				
	- P.	Add of the forestimation	PROTECTION 11	10000	100	100		August and August an	
		And I the Department	Alternational Processing	ineri.	1.00			forgetter	
	HC.	1011 Van 510 de ser-	A DE CARACTERI LA	100000	- 10	304		Repry Reference	
1.	- MC -	AND THE OWNER.	List operation	1000	- A.	100		Through Appro-	
10	- AL	inter Advise Sectors professor	1 To be Comprovided	(Second	1.78		(100 5.4)	Negat Scenes	
1	PE	and the second definition	All for the second	distanting in the second	- 16		100 M	decisional Composition	
14	10011	2021 Open Desembly	- Fill State Evolution strengt	Contract Con	- 10	. 104	(00) Instead 410	POLICIAL DURING	
1.	A	ANY Jone Derived	All time for parts	Concerning State	18	- 1. Ar		Partiel Control	
×.	100	per las bises	THE SECTION AND ADDRESS	(model)	37			An any Comment	
* e	*	# 1 . # # + + + + + + + + + + + + + + + + +	A Reference for the	a distante de la constante de la c	11			C. S. M. L. MARTIN, C.	
		MAR AN Average Section	The laws	200 C	*		100.00		
	11	474 - 17 K M -	NAMES OF A DESCRIPTION OF A DESCRIPTION OF A DESCRIPTIONO	N 4 T		0.000	AND DO DO		
		The late 1 and	the second second second second	10.00 C	14		200 Mar 1		
	w	and the stage	Contracting to Lower	Bar Sec			2.0 C		
	M	March Rose Sugar	Particle Applications	Barren (The Depute Loss		
	~	and sectors are set	A REPORT OF A REPORT	1. M			100 10		
- 8		TAXABLE AND	A. 1993 M. B. 1998 A.	ALC: NOTE: NO			200 Mar 1		
	w	and here and the	N. N. State and Street	Barrier .	1		COLUMN 1 IN THE REAL OF		
	10 C	March March 11, Descent and In-	Painten Berr	Fig. Sec. 6					
		and the second second	100 N 1 1 1 1 1 1 1	10000	-	10.0	AND INCOME.		
	-	THE REPORT OF A DESIGNATION OF A DESIGNATIONO OF A DESIGNATIONO OF A DESIGNATIONO OF A DESI	NO. 10 411	100000					
		When a loss in	Part of Sector Parts	Photo					
		had the ter from the	the second second	dia dia	- C				
		478 YO # 107 13	Rest of the line in	Sec. 1	-	1944	640 PM		
	2	T1 2000 C 200 C 200	Buck Barry	MIT 1874					
	÷	Table and the Table and	State of the second second	a lagla		-			
		My balance address	Welling and Selections		-				
		and on other states and the		Mark .					
	2	The second second second second	10.00						
	-	 Bit has a field of path and 	CHARGE STREET	h i m				-	
	÷.	And the second second second	The state base	la an					
48 (and the second second second	L D C D C C C C C C C C C C C C C C C C	10.00			100	~	

Q4. Have you used the Query Editor in Excel before? Double-click the [Sales_Subset] query and click through the tabs on the ribbon. Which options do you think will be useful in the future?

KEY: Open-ended question, no key provided.

Alternative 4: Use SQL queries in Access

1. TAKE A SCREENSHOT (2-2c).

KEY: Screenshot

4445.770	r - Lake order - Lake and	REAMER	STR - SHAR JIEM	aler	100 - 200	ers under sei kun	a president in the president in	CALL SET IN CONTRACTOR OF CONTRACTOR	- Casesar r	1.1.00 MILL-
50.00	10-12-04/19-02	Reel	2011	14	0 mile	Lub Gb	N. 354	Some of their de 100 billions are Verdinia	.84	3.223
1000	IPPess has	84	2011	2.	12 416	Table 244	5, 202	the End into colliderity of the synth	334	(2.34
100	IL AND DOD	2014	2017		10	Taby Give	8, 2943	Ofer here a 200 Wester (Arth, and a	362	2.788 -
.332	Hartoff loft	201	2857	1.24	91.202	de Ale	8, 200	125 Strate Collins Fred Line	10	7528
.2.4	1.41.41.41.41.91.0	Sec.	2011	2	O and	Fully Gird	N. 2003	Doug with with a shaked to see \$1.5.4 to	2	.7x25k
12.22	D-171-12 1000	25.41	2011		21.4 0	Later Ster 1	5,201	Last bill as it is 20 boolasts if all a	10	10.00
1000	141/14/22 16:00	2014	203.1		10 1.01	1.06.726	N. 2000	Witness Color#When these grade	HO-	20035 +
12224	Pileta bas	205	2857	100	11.100	lub/ab	\$1.200	InterNetwork That the Net A	10	.3633
-1221+	- D11410-24000	2634	2011	2	10,000	1.06.746	No 2020	Boarded statutes 2. Webshow Challe as C.	N/ 685	10220
1000	1011 (1002 1000	20	251		0.02	1.6.76	N: 200	Annual Andre and \$20 B hand it are the actions	. 9.	1100

(Level 2 Header) Part 3: Perform an analysis using PivotTables and Queries

6. TAKE A SCREENSHOT (2-2d)

KEY SREENSHOT:

×		•	- 8 ×
Ead Explanatory	Finant Delete Format Cells	Set ∑ AutoSum * Arr P ∑ Fil* Solid: Find & Char* Filter Select* Edting	ie Terrell (9, Share
			×
		R PivotTable Fi ACTIVE ALL Choose fields to addite Search D Customer L The Customer L The Product D Seles_Subset D The Customer D The Custome	-

1. TAKE A SCREENSHOT (2-2e)

Key screenshot:

Row Labels 💌 Sum of Sales_C	order_Quantity_Sold
Imperial IPA	61
Imperial Stout	258
IPA	235
Pale Ale	116
Stout	190
Wheat	119
Grand Total	979

1. TAKE A SCREENSHOT (2-2f)

Key screenshot:

2	Product_De: 👻	Total_Sales 👻	
	Imperial IPA	61	
	Imperial Stout	258	
	IPA	235	
	Pale Ale	116	
	Stout	190	
	Wheat	119	

2. Save your query as Total_Sales_By_Product and close your database.

(Level 2 Header) Part 4: Address and refine your results

Q5. If the owner of Sláinte wishes to identify which product sold the most, how would you make this report more useful?

Several possible answers. Some options include: sorting the data or filtering the data to view only the product associated with highest total_sales.

Q6. If you wanted to provide more detail, what other attributes would be useful to add as additional rows or columns to your report, or what other reports would you create?

Many possible answers. A good option would be to include Date data from the Sales_Subset table to do analysis on which product sells more based on months or seasons.

(Level 2 Header) Part 5: Communicate your findings

Let's make this easy for others to understand using visualization and explanations.

Q7. Write a brief paragraph about how you would interpret the results of your analysis in plain English? For example, which data points stand out?

Open-ended question, no solution provided.

Q8. In Chapter 4 we'll discuss some visualization techniques. Describe a way you could present this data as a chart or graph.

Open-ended question, no solution provided.

End of lab

(Level 1 Header) Lab 2-3 Resolve common data problems in Excel and Access

Q1. What do you expect will be major data quality issues with Lending Club's data?

Open-ended question, no key provided.

(Level 2 Header) Part 2: Master the Data

Q2. Given this list of attributes, what concerns do you have with the data's ability to predict answers to the questions you identified in Chapter 1?

Open-ended question, no key provided.

Q3. Is there anything in the data that you think will make analysis difficult? For example, are there any special symbols, non-standard data, or numbers that look out of place?

Open-ended question, no key provided.

Q4. What would you do to clean the data in this file?

Open-ended question, no key provided. The next section of the lab, "Let's identify some issues with the data..." introduces several of the items that need to be cleaned (or transformed).

Let's identify some issues with the data.

- There are many attributes without any data, and that may not be necessary.
- The [int_rate] values are written in ##.##%, but analysis will require #.####
- The [term] values include the word "months", which should be removed for numerical analysis.
- The [emp_length] values include "n/a", "<", "+", "year", and "years", which should be removed for numerical analysis
- Dates, including [issue_d], can be more useful if we expand them to show the day, month, and year as separate attributes. Dates cause issues in general because different systems use different date formats (e.g. 1/9/2009, Jan-2009, 9/1/2009 for European dates, etc.), so typically some conversion is necessary.

First, remove the unwanted data:

- 1. Save your file as "Loans2007-2011.xlsx" to take advantage of some of Excel's features.
- 2. Delete the first row that says "Notes offered by prospectus..."
- 3. Delete the last four rows that include "Total amount funded..."
- 4. Delete columns that have no values, including [id], [member_id], [url]
- 5. Repeat for any other blank columns or unwanted attributes.

The columns with the headers revol_bal_joint, sec_app_earliest_cr_line, sec_app_inq_last_6mths, sec_app_mort_acc, sec_app_open_acc, sec_app_revol_util, sec_app_open_act_il, sec_app_num_rev_accts, sec_app_chargeoff_within_12_mths, sec_app_collections_12_mths_ex_med, and sec_app_mths_since_last_major_derog can also be deleted.

Next, fix your numbers:

- 1. Select the [int_rate] column.
- 2. In the Home tab, go to the Number section and change the number type from **Percentage** to **General** using the drop-down menu.
- 3. Repeat for any other attributes with percentages.
- 4. TAKE A SCREENSHOT (2-3a) of your partially cleaned data file.

Key Screenshot:

	kra	her	R I PR	e langt at	Tem J M	CHI SA	ina Yang)	E Developer -	nayi.	ooliao	losar Fie	ы <u>9</u> 74		690 (1911) - 190 (1912) -			
-	Nica Hican Frend		64 E 4	-171 -1751 - 0	1000			Yaa Tat. Yaqa & Carar	Ser.			Gerif Same P		Normal Classic Coll	Ees Dys	a Stativiza	Goog (hope)
	UNCON			1.978		L	Nerrg P		4	04111	- F						SHI .
0	0.4	E F	×.	X 0.15	50												
11		R.	C 2	D		or 1 a		1.1	34	2	T los	M	1 9	U.	12	1 2	
I. k	an arer ha	ndes en	anced a	terri i	nt rete in	utel me gred	a aub are	deemp title e	mp ler	normal cu	Annual I	In verificati	c brue e	i loan statu	sympt p	d sees	putpose
2	45000	15000	15000	60 month	0.1223	306.61 C	CL	MANAGED 1	years	TENT	700	00 Source V	e Dep	Lf Fully Pald :	1	D	dept cos
1	10400	10453	10400	22 month	0.0622	321.00 A	A3	Truck Driv B	ynata	MORTCA	c 300	00 Not Veril	1 Dec-	14 Charged C:	1		credit, ce
41	23425	21435	23425	60 month	0.1567	\$16.06 D	01	Frugtarar 6	Shate	TENT	625	Source V	e Dec	14 Fully Pale :	T		inacit_ta
5	7250	7682	7650	28 marit	0.1368	260.2 0	C7	Technole 4	19541	TV26	300	00 Source V	a Dec	14 Charged C	1		Qual_cas
0	5830	5600	3900	28 year 18	0.1368	328.33 C	C3	Adam 5p41	D ⁺ years	PENE.	020	D Source V	Dia .	14 Fully PARE	ŧ.		dest. ine
ŧ.	125.00	12802	129.0	63 man H	0.1/34	\$15.06 D	04	Surry Sal 1	C+ years	WORICA	c 120	00 Veniled	Un	14 In Grace Pr	6		1.0
8	11041	100.00	10.84	Sherne H	0.1359	3521.6	-85	hear, he	41312	45.41	900	10 Venhed	· Det	14 Chinged Co	(i))		delet_eur
6	121/1	200.0	23375	to minth	0.7154	4 Pr 200	- 65	Harrigal 1	te poars	WRITE	100	D Source V	000	14 Dunest	a .		delt cor
12	1.081	1,103	1428	so on the	ILTSER.	10.05.05 11	1.5	ties up the	te years	VE REPORT	1 20	bollory is	inec	to sully wat	π.		dett_cor
ei l	2981	10.01	2.48	es nomb.	0.0154	39.30 B	8.	Manufaction	1 8741	VE KING	1 200	PL SOURCE V	LIN.	to subject	1		hane in

Then, remove any words from numerical values:

- 1. Select the [term] column.
- 2. Use Find & Replace (Ctrl+H or Home > Editing > Find & Select > Find & Replace) to find the words "months" and "month" and replace them with a null/blank value "". Important: Be sure to include the space before the words and go from the longest variation of the word to the shortest. In this case if you replaced "month" first, you would end up with a lot of values that still had the letter "s" From "months".
- 3. Now select the [emp_length] column and find and replace the following values:

Original value	New value
na or n/a	0
< 1 year	0
1 year	1
2 years	2
3 years	3
4 years	4
5 years	5
6 years	6
7 years	7
8 years	8
9 years	9
10+ years	10
, (comma)	(blank)

This can be done either with Find and Replace or with a False VLookup. The n/a cells have nonprintable characters in them, so the =CLEAN function will be useful for ensuring the n/a values are found in their cells.

H	5.	e - 1	R	- •2									
File	Han	ne Inser	t Page L	avout	Formulas	Data	Rovice	View	Developer	Fuzzy L	oskup Pr	we Fi	wet 🛛 Tel
th	P ap rm		Calibit	• 1	· · A ·	¢ = =	- */	Be	Viap Test	Gent	al.	8)	
Pasto	Copy	et Painter	B I U	• Elle	<u>ðr - A</u>	• = =	= 53	豆 臣)	vlerge & Center	- 5 -	% t. 3	a.	Conditional Fi
	Clipboard	η.		Font		n	34	ignment		4	Number	- 15	
ją:		1.1.1.1.1.1	1 3	6									
4	A	8	c	D		Ŧ	G	н	1	1	ĸ	1	м
t le	oan_amn fi	unded_a fi	unded_arte	rm	int_rate	installmer	grade	sub_gra	deemp_title e	mp_leng	home_ow	annua	Lin verificatio
2	15000	15000	15000	60	0.1239	335.64	c	CI	MANAGES	10	RENT	78	000 Source Ve
Ē	10400	10400	10400	36	0.0699	321.08	A	EA	Truck Driv	8	MORTGAC	58	000 Not Varifi
4	21425	21425	21425	60	0.1559	516.30	D	DI	Programm	6	RENT	63	800 Source Vo
5	7650	7650	7650	36	0.1365	260.2	c	C3	Technical	0	RENT	50	000 Source Ve
6	9600	9600	9600	36	0.1365	325.53	C.	C3	Admin Spi	10	RENT	69	000 Source Ve
7	12800	12800	12800	60	0.1714	319.08	D	04	Senior Sal	10	MORTGAG	125	000 Verified
5	10000	10000	10000	36	0.1199	332.1	в	85	Investmer	8	RENT	90	000 Verified
9	21075	21075	210/5	60	0.2199	581.95	E	ES .	Paralegal	10	MORIGAC	55	000 Source Ve
Q.:	17000	17000	17000	36	0.1365	578.22	c	C3	Deputy sh	10	MORTGAG	75	000 Verified
1	2500	2500	2500	36	0.1199	83.03	в	85	Manufacts	0	MORTGAC	89	000 Source Ve
12	5250	5250	5250	30	0.1144	172.98	в	84	Store Man	2	RENT	26	000 Not Venfi

4. **TAKE A SCREENSHOT (2-3b)** of your partially cleaned data file, showing the [term] column.

Q5. Why do you think it is useful to reformat and extract parts of the dates before you conduct your analysis? What do you think would happen if you didn't?

Open-ended question, no key provided.

Q6. Did you run into any major issues when you attempted to clean the data? How would you resolve those?

Open-ended question, no key provided.

END OF LAB

(Level 1 Header) Lab 2-4 Generate summary statistics in Excel

Because every question in this lab is open-ended, there is no key provided.

END OF LAB

(Level 1 Header) Lab 2-5 – College Scorecard Extract and Data Preparation

(Level 2 Header) Part 2: Master the Data

1. Take a screenshot (1)

Screenshot Key:

Sector Area	1. 1.6 1	H Y Fren	1000		图	21	Nomel	A45	Loca	jacosi .	And a start of the second s
Thumber a B T H		日日 ゴエ Pressioner	3100 -	取在 2	an and in	-	17655, 111	Tables active	1000	S- VARCAL	nota
comparise of the		Tangana		1.	10	_			we.		
	CILL RD COULD	IN & CONTRUCT & AND CARLON	ILUM VACUUE	LUDDED W	FILUARS .	aria.	ALC: NO.	AN NONMERAD ?	SAVALU R	PERSONAL STRUCTURE	A DUC FRY DA
				NUMBER OF							
A	in the second	9 (c (1.5))	2	1	12-21	20		24	1 3	2 1	+ V
	LEN CAUN N.C.				the local party lines with	a second second	the second second	and the second se	a designed and the second states	a i Daerolaj, Jieri	. V
N HO, KNINN STRANGA WHITE .		M. JARARDANANS A HEADER	CALL UNDER	A Sala	AND NO.		ALC: NO.	المدد الاصرامي	HEN LAR UN		
n no, kana sinanakan Johnet. Na 1745 sesta Akalih serang San	44,120,21.78,475	ALL ACCORDING AT LODIE	CALL COMES I	A LANA	CANELINA,	14.1	2425 N 1913	المناجعة (10 من 10 من 10 من 10 من 10 من	10. 2110 3	A ROY DAT	10000
n 10. Karal Lington and Lington 180 Million Alexandria (Marin 1993) International Lington (Marine)	44,155 E	NUMERAL AND ADDRESS OF THE STATE	Carlo Caelo, I Di Ol, Unite D Di Ol, Unite D	With Lands of With Lands With Lands	Centrum Sector 2 Charles		840 <u>5</u> 1913 2423207	MALAYUR LUAL Ng Sakay Calas Ng Sakay Calas	10. 2110 3	A ROY DAT	10000
n 100 koma utaying kabupatèn ku 1980 kabupatèn Neletik kabupatèn ku 1980 kabupatèn di Abupatèn Kabupatèn 1980 kabupatèn kabupatèn Kabupatèn Kab	an in the second se Second second second Second second	Martin Martin and States and S	Callat Calls_1 Droll_atriay_1 Droll_atriay_1 Ar data_roll_ar	Maria Carlos Maria Carlos National Carlos	CANELINA CANELINA CAREAGO CAREAGO		anda ja intu Matan Aga Antan Jase Lati Matan	une of the second s	100 2 10 10 10 100 2 10 0 10 100 10 10 10	2.9.00 07.000 1.91 0.000	/Maginalities
n met kom stratisk konstantisk 1980 – Maria Arabitisk, sversk frans 1980 – Maria Arabitisk, sversk fransk 1987 – Frank Stratisk, storagen fra 1997 – Statisk stratisk, storagen fra 1997 – Statisk stratisk, sversk fransk	an (150 m) fairt a fair State da a a Maria State da a a Maria State da a	N. Parkara I. Andreas and A. Markara and A. Mark	Callet Coulds II Droll, United Sci Droll, United Sci Droll, Droll, Droll Market Science Weiters Weiter	NA LUNA NA LUNA NG LUNA NG LUNA NG LUNA NG LUNA	CARLONA 2 BO CO 2 CACINANI CACINANI CACINANI NO BONE		andra je inter Jerovanski Stanistick Stanistick Stanistick	une an on ourse ne abhair cha an chir seac y an a mar a seac an an a ga na seac an	NEAL COLOR	n di seri ne merinari ne merinari	ZORIJANCIJE INIJA UMA
P. TRACKS IN THE RECEIPTION CONTINUES. (B) A Contract A contract on the processing of the contract of the c	an the string of the Sector String (Sector String String (Sector String) (Sector String) (Sector String)	Al Development of the set of t		NA CURA NA CUR	Carriera 2 de los 2 Constantes 2 de los 2 Constantes 2 de los 2 Constantes 2 de los 2 Constantes		Allo ja Muo 24 Conto (27 1 Martin (27 1 Martin (27 Martin (27 20 Martin (27 20 Martin (27)	ung jarun ung j Rejaran da an Ann Ang yang ta Managan da ang yang Karan da ang kang ta		a Aprilan Der Michaels Der Michaels A Appendiction	zowycania Maz wys
n met kom sing transverse och met 1960 f. V. Landa Ale Millinger og fyrand 1967 f. Brandski at State och film for 1967 f. Brandski at State och film for 1967 f. Brandski at State och film for	an the string of the Sector String (Sector String String (Sector String) (Sector String) (Sector String)	Al Development of the set of t		NA CURA NA CUR	Carriera 2 de los 2 Constantes 2 de los 2 Constantes 2 de los 2 Constantes 2 de los 2 Constantes		Allo ja Muo 24 Conto (27 1 Martin (27 1 Martin (27 Martin (27 20 Martin (27 20 Martin (27)	ung jarun ung j Rejaran da an Ann Ang yang ta Managan da ang yang Karan da ang kang ta		a Aprilan Der Michaels Der Michaels A Appendiction	zowycania Maz wys

Q1. By looking through the data in the text file, what do you think the delimiter is?

Comma

2. Take a screenshot (2)

Screenshot Key:

日	•9		C	B												Nonici - Sace	
16	6 H	ee 1	6.51	1.9	etym	- Famile	a Data	See.	West 1	Discopera	Fugli	ikin D	van Piar	Q 14	er alle yee		
	From			1. C	E	New Year	Shaw Qar Barr Tele Sacra Su Sacra Su	- 14	Concdar	dir. Sh	24 <u>519</u> 54 ⁵⁰⁰	Ter T	Care Roorry Assarch	Terts Columns		nter Side	0 = n Car
110		•11]	8	×.	6	0,028											
41	Α.	. 0	ī.	0	c	T .	r	G.	1.01.1	17	line d	ĸ	-1	- 54	N	0	
1	NITE	INSTINC	្រភ	7	STARN	CONTRO	DI CORASIO	ADM_RAT	T.SAT_4WS	lans	USDS_WS	UGDS_B/Z	LIGDS_HIS	UGDS_ASI	UGDS_AIA	LIGES_NH	KD5_294
2	100154	Aleberte	No	inal	AL.		1 11	0.5258	5 027	/1205	11111	0.7151	0.0055	0.0010	0.0024	0.0019	D
\$	100553	Universi	. Un	ningh	AL		1 15	0.8565	110/	11385	0.5922	0.25	002265	0.0538	0.0002	0.0007	0.0068
4	100690	Amnidas	LINC	nogom	AL SP		2 20	NULL	NULL	291	0.299	0,4192	0.0055	0.0084	0	U	0
5	100706	Universit	ty Hu	texilie	31		1 10	0.8203	1219	5451	0.6988	-0.1255	0.0882	0,0376	0,0148	0.0002	0.0172
1	100724	Alabama	SMo	nteom	AL		1 19	0.5320	179 7	2011	0.0133	0.4208	0.0121	0,0010	0.001	0.0006	0.0008
ž.	100751	The Unit	e Tu	calcos	AL.		1 16	0.5105	11/18	23051	0.7325	0.1110	0.0345	0.0306	0.0035	0.0009	0.0061
d	100760	Control	al Ale	ounde.	AL		1 1	NULL	NULL	1592	0.7255	0.2513	0.0044	0.0025	0.0044	Ų	V
)	100312	Athons i	t.Ath	DITS	AL		1 22	NULL	NULL	2991	0.7823	0.12	0.0191	0.0058	0.0157	0.001	0.0174
n.	100330	Autorn	h Me	nigom	M.		1 18	0.8017	7 1009	4304	0.5328	0.3375	0,0078	D.0221	0.0044	0.0016	0.0297
1	101111	Auburn	hou	11110	24.		1 1/	0.6343	1215	20514	0.0507	0.0704	0.0348	0,0337	0.0074	D	0
2		Demonal		Sec. 1.			2 23	0.5552	1131	1130	0.7983	0.1104	0.0235	0.051/	0.0000	U	0.0061

- 3. To ensure that you captured all of the data through the extraction from the .txt file, we need to validate it. Validate the following check sums:
 - You should have 7,704 records (rows).
 - Compare the attribute names (column headers) to the attributes listed in the data dictionary. Are you missing any, or do you have any extras?

• The average SAT score should be 1,059.07 (this is leaving NULL values as NULL).

Q2. In the check sums, you validated that the average SAT score for all of the records is 1,059.07. When we work with the data more rigorously, several tests will require us to transform NULL values. If you were to transform the NULL SAT values into 0, what would happen to the average (would it stay the same, decrease, or increase)?

The average would decrease

How would that change to the average impact the way you would interpret the data?

It would inaccurately represent a very low SAT average across all schools (Correct Answer)

Do you think it's a good idea to replace NULL values with 0s in this case?

No

- 4. To avoid the issues with NULL, blanks, and 0s, we will remove all of the records that contain NULL values in either SAT_AVG or C150_4. Do so.
- 5. Perform a =COUNT() to verify the amount of records that remain after removing all records associated with NULL values in SAT_AVG or C150_4. 1,271 records should remain.
- 6. Take a screenshot (3)

Key Screenshot:

	A DESCRIPTION OF THE OWNER OWNER OF THE OWNER OWNE OWNER OWNE OWNER OWNE		4 				-					07.64			30 a ja	isa ny	9F)					
Ē	- Kar Anda	1,454	1.011.04	+ u - x			10-	Pre-	ine . Na La das	-		-	je.		terest COLLER	8.8 1976	122.9-1 (ika: Text	- 16 A	a CALL	An t	
	of the second	45			1			ant at		¥ .		1.		0.0			1.64	4				-12
	0.00	1.12	4	WITH-																		
	the design of the local division in the	1.1. 60			- Park	less:	-40-52	AND CO					No.		P.S.M.						-N	
3	240 AUTO		citra	e conce	in.	ao	Barry Barry	1° 194.1	210			CLUB-	1285, 440	NLC 11	C KOK ALC	CY 30	10000	VR.	UDATE NO	12 F # 1816 F # 1	ALC: NO	111
1	100244				1		E.V.DI	120	-				100		1				aprile Sa		A TRUE	£
3	120/20/20 10:00:0		420		5	24		1.04.	10	3.415	1010	1197	1.982	1040	6 A	- 2		0.90	2.54		1.040	
1	100	T. HEALT	ALC:		1	10	100	10.	die	Maria	24.22	5.045	1.00	20.1	2		1 V	United	10.0	A VIGE MULL	1.000	
1	1001Cank				1	.5	L.* 008	50	0.1	33114	11-0	Loga	1.23	0.000	30.16	E-134	1.301	0.001	13:1	102 M 1	THEA.	
3	29W17 Sec. 6	5 V els q	- 0k			25	1.22%	44	40.	5.017	3 100	1.1004	1.1982	4,540	1000	81014	1	0.944	Alla V	10764	SALLA.	
3	LOLD LONG	of an interest	6-1.e.		1	10	3.40	34.		. mrs	61.2	8.004s	\$ 340	2.00/1	a Male	Rouge	122	VALE	Acres	DOM NOL	LORD Y	
3	200 Channe		1 2		15	0	1.142	CD .	100-	3/14	25.42	LOH	1 290	1.2.	10.00	E.	E 201	0.200	122.04	SHARE MELL	.731	
1	Three are to	sillin	14		1	22	1.000	NTT.	244	3 775	1000	6.1684	1.000	A lefter	1000	1.000	12.91	0.0001	1.142	A153 MULT	1.412	
1	and dee	W.Sedua	d un		L	12	0.6.2	2.0.	-110	140 at	AILM	C.D.J.	V Alla	U.S.S.V.	4 1	10.000	1.000	V.Ster.	A.L.M.	30000 M011	1200	
1	120.00 3.4	Side I and	eW0		1	22	2/12	05	34.	3497	23.10	E.MA	1.301		1. 11.01	E-CO.	1.300	1.23.1	3.20 cit	SYME METER	13037	
9	200112 Currow		. 76		2	-24	\$ 23.79	441	287	15.Wet	\$ 0.00	1.161	1.0	-2.001	N 3	1.4154	63.17	1	1005 51	10VM	2010	
1			de a l							1.1.1	1000	1.1.1	and see a	1.000	1000	and an all		in the second second				

Your data is now ready for the test plan. This lab will continue in chapter 3.

END OF LAB

(Level 1 Header) Lab 2-6 Comprehensive Case: Dillard's Store Data: How to Create an E-R Diagram

Questions for Parts 1-3 are all open-ended, no key provided.

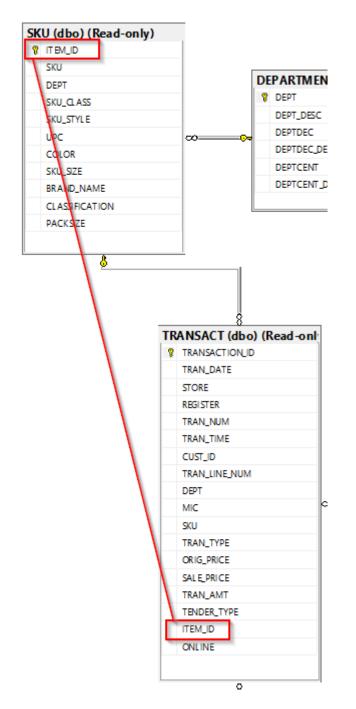
(Level 2 Header) Part 4: Address and Refine Results

Q3. What is the primary key for the TRANSACT table? What is the primary key for the SKU table?

ITEM_ID – Correct Answer for SKU table TRANSACTION_ID – Correct Answer for TRANSACT table

Q4. How do we connect the SKU database to the TRANSACT table? How do we join tables from two different related tables?

Tables are joined by relating the foreign and primary keys. The TRANSACT table has a foreign key from the SKU table, so the relationship between the two characters is the joining of TRANSACT.ITEM_ID and SKU.ITEM_ID.



END OF LAB

(Level 1 Header) Lab 2-7 Comprehensive Case: Dillard's Store Data: How to Preview Data From Tables in a Query

(Level 2 Header) Part 1: Identify the Questions

Q1. How would a view of the entire database or certain tables out of that database allow us to get a feel for the data?

Open-ended question, no key provided.

Q2. What types of data would you guess that Dillard's, a retail store, gather that might be useful? How could Dillard's suppliers use this data to predict future purchases?

Open-ended question, no key provided. Possible answers include: sales data (sales orders, sales order dates, items sold), customer data (what each customer purchases, where they live), inventory (retail price, cost, category), etc.

TAKE A SCREENSHOT OF YOUR RESULTS (2)

KEY Screenshot

1L		Universit (MIL)*	44.18															
	alact top 18 1																	
	from transact.																	
	and distance																	
	THEFT IN A	Timilals	204	60/0		tier, tak	9,100	THALINE		×.			98,845	345,9955	TRUN, MIT	19409-744		01.14
	THEOR ROUT	Tenjak more	140	11	14	1118	10001111	Tangungan T	101	185	1160626		10.10	143,7455 8.0	Tion, MIT 20.95	DAIX	154MA	.01
	THEFT IN A	Tim, Jaki struct of 204-0.0		11 28										MAR PRES MIT	TRAN, MIT 26.06 (17.11		tháphia troni fe	8
	THEOR ROUT	Tenjak more	140	11 20 40	14	1118	10001111		101	185	1160626		10.10	143,7455 8.0	Tion, MIT 20.95	DAIX	154MA	8
	THEOR DON D	Tim, Jaki struct of 204-0.0	140	11 28	4 3 01 22	1118	162542344		196 965	188 300	17889828 0050925 2421023	1	41100 01:00	MAR PRES MIT	TRAN, MIT 26.06 (17.11	CARK GAM	tháphia troni fe	11 11 11
	THEORY OF HEARING	Ton JAK strate o 2040.0 State o	142 140 140	11 20 40	4 3 01 22	1118 1139 1646	183847344 163547344 11565000		195 955 267	111 100 114	17889828 0050925 2421023	1	81.10 61.00 10.00	224,344 8.8 1111 121	TRAN_MAT 28.08 17.117 2:50	SAIX SAM OLFD	1709010 17090170 (757900)	11 11 11
	The card of the same	Torn, J.W.S. attracts of 201+01-01 301+01-01 potentiol	140 140 140	11 20 12 20 12	a X UI	1118 1139 1646 1008	1825-234 1525-234 1158-5800 1884-5880		開始が開	111 100 114 270	17889628 0050405 2431023 0020513	1	81.10 64.00 10.00 34.00	34.5,7455 8.8 1737 331 4.8	TRONG MIT 28.08 17.37 2.50 8.00	DAIDX BAAH OLFD BAAH	1546540 17206170 (757600) 17362008	N
1	Transian Dong B Transian Dong B 214440000 214440000 214440000 214440000 214440000 214440000 214440000 214440000	Tion, 1.9 K atruct of 2014/0.01 2014/0.01 2014/0.01 2014/0.01	140 140 140 140	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	4 3 01 22	1118 1139 1946 1008 1234	180821111 16062384 11565000 186270888 146506885		19.55% 19.25	11 20 計算版	TTRABLE OCTOBER PADIELLE MERINER STREPS	1	41.10 (1-0) 10:00 31.00 10:00	245,745 8.0 1111 121 121 121 121 121	TRAN, MAT 18.06 17.317 2.50 8.55 1.96	DAIDS BARK OLFO BARK OLFO	rajenie rzajenie (SZNU) rzajenie krytych SZNU	
1	Weise Character Versie Character Presson 20160000 20160000 20160000 20160000 20160000 20160000 20160000 20160000 20160000 20160000	Total (14) (5 2014-01-01 2014-01-01 2014-01-01 2014-01-01 2014-01-01 2014-01-01 2014-01-01 2014-01-01	142 148 149 140 140 140	10-10-00	4 3 10 2 4 3 11	1118 109 1646 1008 129 129 121 1028	Tablattin HOSATINA HISSATINA HISSATINA HISSATINA HISSATINA HISTATINA HISTATINA		1993年1993年	11 治計算術設備	7780848 6050925 2404233 8429933 6778732 4624763 8728433	1	91.00 (1-0) (1-0) 34.00 (2-0) 22.00 (0.0)	94.5,7465 94.8 17.17 255 8.8 155 125 34.00	TRUN, MIT 26.06 17.317 2:25 8.05 3.95 7.28 26.06	DATA BANK OLFO BANK OLFO DATA	TAMANU UTUMI NI UTUMI NI UTUMI NI MPRITIT NI TUMI NI UTUMI NI UTUMI NI	
	Control (Control (Contro) (Contro) (Contro) (Contro) (Contro) (Contro) (Contro) (Contro)	New 1915 2014/0.01 2014/0.01 2014/0.01 2014/0.01 2014/0.01 2014/0.01 2014/0.01 2014/0.01	142 140 140 140 140 140 140	1201000	4 3 11 3 4 3 4 3 11 11 11	1114 109 1646 1008 129 129 121 1028 1040	100001111 162562384 115850000 1884/18880 168000009 115600047 187502660 15500047		·····································	11 第三十二十二十二十二十二十二十二十二十二十二十二十二十二十二十二十二十二十二十二	1788808 0050405 340403 8020413 0178702 4014105 0788833 7455400	1	91.10 (1-0) (1-0) 34.00 (2-0) 22.00 (0)10 (0)10 (1-0)(1445,74555 24.00 17.17 125 125 125 24.00 126	TRAN, ANT 20.04 (17.17 2:20 8:06 3:06 1:20 2:20 2:20 3:20 3:20 3:20 3:20 3:20 3	DAUX SAM OLFO SMR OLFO DAUX DAUX DAUX	1548642 17284110 (757842) 17385420 6296747 6552420 17386946 17386946	
	Weise Character Versie Character Presson 20160000 20160000 20160000 20160000 20160000 20160000 20160000 20160000 20160000 20160000	Total (14) (5 2014-01-01 2014-01-01 2014-01-01 2014-01-01 2014-01-01 2014-01-01 2014-01-01 2014-01-01	142 148 149 140 140 140	10-10-00	4 3 10 2 4 3 11	1118 109 1646 1008 129 129 121 1028	Tablattin HOSATINA HISSATINA HISSATINA HISSATINA HISSATINA HISTATINA HISTATINA		1993年1993年	11 治計算術設備	7780848 6050925 2404233 8429933 6778752 4624765 8728433		91.00 (1-0) (1-0) 34.00 (2-0) 22.00 (0.0)	94.5,7455 94.8 17.17 255 8.8 155 125 34.00	TRUN, MIT 26.06 17.317 2:25 8.05 3.95 7.28 26.06	CAUX BAAM OLF0 BAAM CUP0 DAAM CAUX	TAMANU UTUMI NI UTUMI NI UTUMI NI MPRITIT NI TUMI NI UTUMI NI UTUMI NI	

Q3. What do you think 'P' and 'R' represent in the TRAN_TYPE table? How might transactions differ if they are represented by 'P' or 'R'.

Answers will vary, but P represents Purchase and R represents Return.

Q4. What benefit can you gain from selecting only the top few rows of your data, particularly from a large dataset?

Answers will vary, but some possible solutions include getting a quick glance at the data without having to wait for the query to run if it's a large dataset.

(Level 1 Header) Lab 2-8 Comprehensive Case: Dillard's Store Data: Connecting Excel to a SQL Database

Q1. What can you do in Excel that is much more difficult to do in other data management programs?

Open-ended question, no key provided.

Q2. Since most accountants are familiar with Excel, name three data management functions you can do easier in Excel than any other program? How does that familiarity help you with your analysis?

Open-ended question, no key provided.

1. Take a screenshot of the PivotTable.

Row Labels 💌	Count of STORE
AL	9
AR	10
AZ	17
CA	3
со	8
FL	44
GA	12
IA	5
ID	2
IL	3
IN	3
KS	6
KY	6
LA	15
MO	10
MS	6
MT	2
NC	15
NE	3
NM	6
NV	5
NY	1
ОН	15
OK	10
SC	8
TN	10
тх	61
UT	11
VA	6
WY	1
Grand Total	313

Q3.

Reference your PivotTable and find which state has the highest number of Dillard's stores. Which states have the fewest? How many stores are there across the country?

Texas has the highest number of stores, New York and Wyoming have the lowest. There are 313 stores across the country.

Q4. Counting the number of stores per state is one example of how the data that has been loaded from SQL Server into Excel can become useful information through a PivotTable. What are other ways that you could organize the STORE data in a PivotTable to come up with meaningful information?

Open-ended question, no key provided.

Q5. Joins are made based on their Primary Key – Foreign Key relationship. Looking at the ER Diagram or the dataset, which two columns form the relationship between the TRANSACT and STORE tables?

Transact.ITEM_ID = Store.ITEM_ID

Q6. Looking at the first several rows of data, compare the amounts in ORIG_PRICE, SALE_PRICE, TRAN_AMT. What do you think tran_amt represents?

The total transaction amount, taking into account discounts.

Q7. What are the means for each of the attributes?

ORIG_PRICE: 53.9857

SALE_PRICE: 35.461

TRAN_AMT: 27.83595

Q8. The mean from TRAN_AMT is lower than the means for both ORIG_PRICE and SALE_PRICE, why do you think that is? (Hint: it is not an error).

The TRAN_AMT not only takes into account discounts, but also is negative when the transaction is a return.

(Level 2 Header) Part 5: Address and Refine Results

Q9. How does doing a query within Excel allow quicker and more efficient access and analysis of the data?

Open-ended question, no key provided. Possible responses include not having to export the query results from the database.

Q10. Is 15 days of data sufficient to capture the statistical relationship among and between different variables. What will Excel do if you have over 1 million rows? There are statistical programs such as SAS and SPSS that allow for transformation and statistical analysis of bigger datasets.

Open-ended question, no key provided. Possible responses include – that 15 days of data may be sufficient for a snapshot, but more data would make for stronger statistical analysis. If Excel has over 1 million rows, it will cut off the results at the 1,048,576th row.

(Level 2 Header) Part 2: Master the Data and Part 3: Perform an Analysis of the Data

For Step 8, the query the students should run should be:

SELECT State, COUNT(Cust_ID) FROM Transact INNER JOIN Customer ON Transact.Cust_ID = Customer.Cust_ID GROUP BY State

Q1. How many different states are listed?

65 states with values, 67 states if NULL values are included

Q2. Why are there so many more states listed than 50?

Answers will vary. Possible answers include that non-state territories are listed in the state field of the Customer table, and there could be inconsistency in the state field of the Customer table that would create duplicate state values.

Q3. What do you assume the Other, XX, blank, and Null states represent? If you were to analyze this data to learn more about the amount of customers from different places have shopped at Dillard's, what would you do with this data – group it, leave it out, leave it alone? Why?

Answers will vary. The Other, XX, blank, and Null states probably represent Customers who didn't indicate their state when they made their purchase. There is inconsistency in the way unknowns have been recorded in the data. Leaving the data with Other, XX, blank, or Null out of the analysis is likely the best solution if you are doing analysis based on geographic data – it is not meaningful if there isn't a geographic location attached to the record.

(Level 1 Header) Lab 2-9 Comprehensive Case: Dillard's Store Data: Joining Tables

(Level 2 Header) Part 1: Identify the Questions

Q1. If we wanted to join the TRANSACT and the CUSTOMER tables, what fields (or variables) would we use to join them?

The two tables are linked by the Customer.Cust_ID primary key in the Customer table and the Transact.Cust_ID foreign key in the Transact table.

Q2. Because most accountants are familiar with Excel, name three data management functions you can do easier in Excel than any other program. How does that familiarity help you with your analysis?

Answers will vary. Possible answers include pivoting data, working with long calculations, etc.

(Level 2 Header) Part 2: Master the Data and Part 3: Perform an Analysis of the Data

Step 8 has the students create a query that will show how many customers have shopped at Dillard's, grouped by their respective states (using the entire dataset). The query is:

SELECT STATE, COUNT(TRANSACT.CUST_ID) AS Number_Of_Customers
FROM CUSTOMER
INNER JOIN TRANSACT
ON TRANSACT.CUST_ID = CUSTOMER.CUST_ID
GROUP BY STATE

Possible modifications – you could count any field in the Transact table, it does not have to be Transact.Cust_ID. The alias could also be different (whatever you would like to rename the field).

Q3. How many different states are listed?

67 records are returned, indicating there are more than just the 50 states listed.

Q4. Why are there so many more states listed than 50?

The table lists a few options for when the employee didn't gather the customer's state information (blanks, NULL, Other, XX). The table also lists territories (such as PR for Puerto Rico). Several of the other options are acronyms for branches of armed forces (AE, AP, AA).

Q5. What do you assume the Other, XX, blank, and Null states represent? If you were to analyze these data to learn more about the number of customers from different places have shopped at Dillard's, what would you do with these data: group them, leave them out, leave them alone? Why?

You can assume that the Other, XX, blank, and Null state values represent customers who did not provide their state information to the employee. Answers will vary for the second part of this question, and it depends on the analysis the student wishes to do.

Solutions Manual – Chapter 2

Solutions to Discussion Questions

- The information needs only to be entered once and changes or edits only need to be done in one file versus multiple files. It won't take up unnecessary space (which is expensive), take up unnecessary processing to run reports to ensure that there aren't multiple versions of the truth, and will not increase the risk of data entry errors.
- 2. Relational databases are designed to support business processes across the organization, which results in improved communication across functional areas and more integrated business processes.
- 3. Relational databases all connect with each other by use of the primary and foreign key. That makes data analysis very easy to do since you can readily join the tables and run the requested data analysis.
- 4. Relational databases can be designed to aid in the placement and enforcement of internal controls and business rules in ways that flat files cannot. Due to the nature of the primary key/foreign key, both a primary key and a foreign key must line up with each other before any business can be transacted. If there is no supplier in the approved supplier file, it is not possible to process a purchase order without linking to the approved supplier file.
- 5. The data dictionary is a centralized repository of descriptions for all of the data attributes of the data set. Attributes of a data dictionary for each field might include a variable name, a brief description, whether the field is made up of numbers or text or alphanumerics, the size (or number of digits) of the field, whether it serves as a primary or foreign key and notes, etc.
- 6. Before extracting the data, it is important to be able to answer these questions:
 - a. What is the purpose of the data request? What do you need the data to solve? What business problem will it address?
 - b. What risk exists in data integrity (e.g., reliability, usefulness)? What is the mitigation plan?
 - c. What other information will impact the nature, timing and extent of the data analysis?
- The analyst needs to know what data is available, how it comes, what it includes, and how reliable the data is to be able to answer the central question which was the reason for the analysis.
- 8. The more frequent the requested report, the more the database administrator will set it up for automatic extraction and delivery. It may also be a question of how often the data changes. If the data is updated weekly and the data is extracted daily, that may not make any sense.
- 9. The database administrator is most familiar with the data and may be able to help the analyst get the data needed to address the question. There also might be some sensitivities to who gets what data to ensure that the data gets to the intended analyst and audience.
- 10. The impact of transforming data to work with NULL, N/A and zero values in the dataset might have an impact on programs like Excel.
 - a. Transforming NULL and N/A values into blanks.
 - i. The COUNT and AVERAGE functions would not include these fields in their computation for these variables.
 - b. Transforming NULL and N/A values into zeroes.

- The COUNT and AVERAGE functions would incorporate these zeroes and would be included in their computation for these variables. It would have an impact particularly on the computation of the average since it would have the value of zero.
- c. Deleting records that have NULL and N/A values from your dataset.
 - i. The COUNT and AVERAGE functions would not include these fields in their computation for these variables. If they are deleted all of the other fields and variables would be deleted as well, thus having a bigger impact on the overall dataset.

Solutions to Problems

Problem 2.1

Attributes needed from the College Scorecard data to compare the cost of attendance across types of institutions (public, private non-profit, private for-profit) would include:

- a. CONTROL 1 = Public. 2 = Private nonprofit. 3 = Private for-profit
- b. COSTT4_A Average cost of attendance

Problem 2.2

Attributes needed from the College Scorecard data to compare SAT scores across types of institutions (public, private non-profit, private for-profit) would include:

- a. CONTROL 1 = Public. 2 = Private nonprofit. 3 = Private for-profit
- b. SAT_AVG average equivalent SAT of students admitted
- c. UNITID a unique identifier for the institution

Problem 2.3

Attributes needed from the College Scorecard data to compare levels of diversity across types of institutions (public, private non-profit, private for-profit) would include:

- a. CONTROL 1 = Public. 2 = Private nonprofit. 3 = Private for-profit
- b. UNITID a unique identifier for the institution
- c. UGDS enrollment of undergraduate certificate/degree-seeking students
- d. UGDS_WHITE total share of enrollment of undergraduates who are white
- e. UGDS_BLACK total share of enrollment of undergraduates who are black
- f. UGDS_HISP total share of enrollment of undergraduates who are Hispanic
- g. UGDS_ASIAN total share of enrollment of undergraduates who are Asian
- h. UGDS_AIAN total share of enrollment of undergraduates who are American Indian/Alaska Native
- i. UGDS_NHPI total share of enrollment of undergraduates who are Native Hawaiian/Pacific Islander

- j. UGDS_2MOR total share of enrollment of undergraduates who are two or more races
- k. UGDS_NRA total share of enrollment of undergraduates who are non-resident aliens
- I. UGDS_UNKN total share of enrollment of undergraduates whose race is unknown

Problem 2.4

Attributes needed from the College Scorecard data to compare completion rate across types of institutions (public, private non-profit, private for-profit) would include:

- a. CONTROL 1 = Public. 2 = Private nonprofit. 3 = Private for-profit
- b. UGDS enrollment of undergraduate certificate/degree-seeking students
- c. UNITID a unique identifier for the institution

Problem 2.5

Attributes needed from the College Scorecard data to compare the percentage of students who receive federal loans at universities above and below the median cost of attendance across all institutions (public, private non-profit, private for-profit) would include:

- a. CONTROL 1 = Public. 2 = Private nonprofit. 3 = Private for-profit
- b. PCTFLOAN Percent of all federal undergraduates receiving a federal student loan
- c. COSTT4_A Average cost of attendance
- d. UNITID a unique identifier for the institution

Problem 2.6

Attributes needed from the College Scorecard data to compare the percentage of students who receive federal loans at universities above and below the median cost of attendance across all institutions (public, private non-profit, private for-profit) would include:

- a. UNITID a unique identifier for the institution
- b. STABBR State postcode
- c. COSTT4_A Average cost of attendance

Problem 2.7

Name	Vernon Richardson	Contact Number	479-345-0537						
Name		Email Address	vjricha@uark.edu						
Description of Information Required (Please include dates/timeframes for any analysis, and other specific indicators/categories required in the data)	From the College Scorecard data, I need the following data items for each year and unique identifier (UNITID) of the dataset: a. UNITID – a unique identifier for the institution b. STABBR – State postcode c. COSTT4_A – Average cost of attendance								
Purpose/Context (what the data is required for)	I am trying to determine if different regions of the country have significantly different costs of attendance.								
Frequency (circle as appropriate)	One Off Request / Other	Annually X	Term						
Request Date	7/15/2020	Required Date	7/28/2020						
Format Required (Table, Spreadsheet, Word, etc.) – please specify	Tab Delimited File	Customer (if not requestor)							
To be used in (presentation, report, etc.) – please specify	Report to University of Arkansas administration	Intended Audienc appropriate)	e (if Chancellor, President, Provost of the University of Arkansas						

Problem 2.8

Diversity can be determined by a number of different dimensions. The College Scorecard data seems to have information on race including the following fields:

UGDS – enrollment of undergraduate certificate/degree-seeking students UGDS_WHITE – total share of enrollment of undergraduates who are white

UGDS_BLACK – total share of enrollment of undergraduates who are black UGDS_HISP – total share of enrollment of undergraduates who are Hispanic UGDS_ASIAN – total share of enrollment of undergraduates who are Asian UGDS_AIAN – total share of enrollment of undergraduates who are American Indian/Alaska Native UGDS_NHPI – total share of enrollment of undergraduates who are Native Hawaiian/Pacific Islander UGDS_2MOR – total share of enrollment of undergraduates who are two or more races UGDS_NRA – total share of enrollment of undergraduates who are non-resident aliens UGDS_UNKN – total share of enrollment of undergraduates whose race is unknown

Depending on the focus of the report, it may make sense to capture broader dimensions of diversity rather than knowing the population of each individual race category. In that case, it may make sense to combine categories to have less categories.

Problem 2.9

You would first need to calculate the median cost of attendance at universities and determine which universities are above and below that median. You may need to do this for each year included in the analysis as the cost of attendance changes from year to year. Once this is done, you can compute the percentage of students who receive federal loans at each university and compare them for those both above and below the median cost of attendance.