

## 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**.
3. 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

*Copyright © 2019 McGraw-Hill Education. All rights reserved. No reproduction or distribution without the prior written consent of McGraw-Hill Education.*

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*

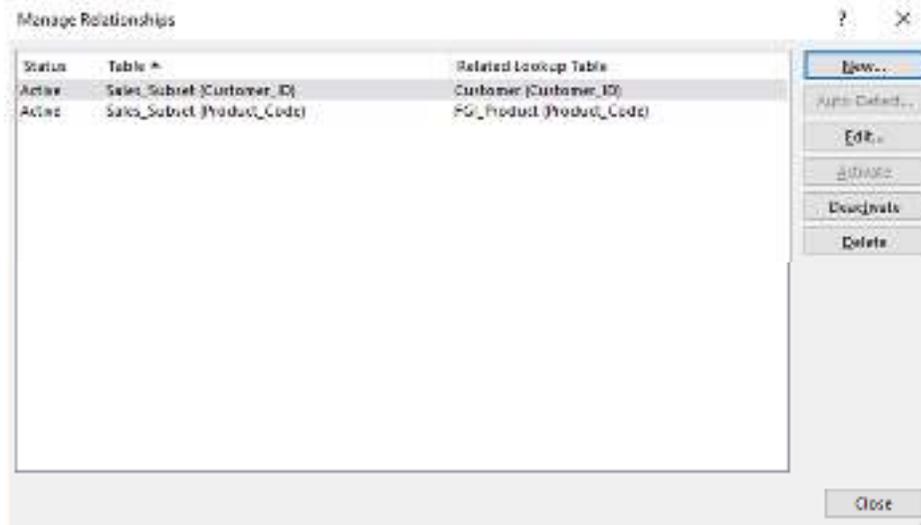
**(Level 2 Header) Part 2: Master the data: Prepare data for analysis in Excel**

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.



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

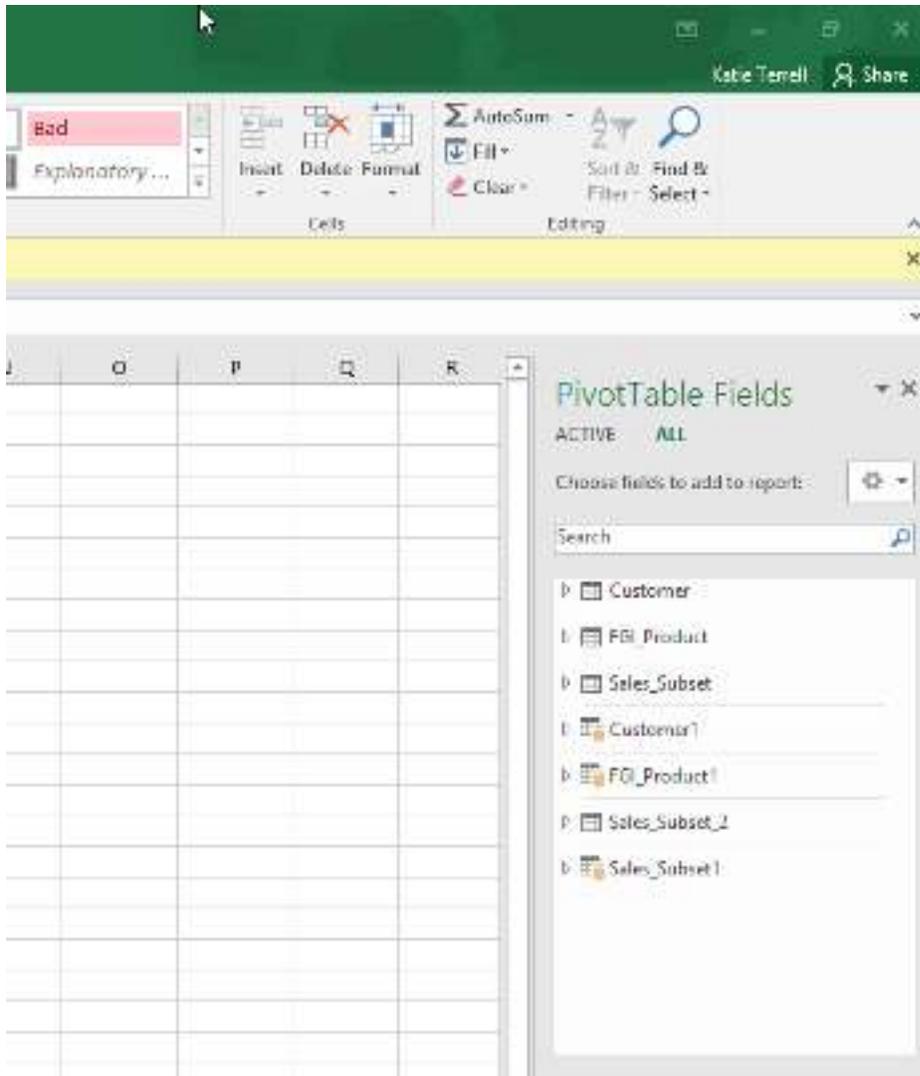
*KEY: open-ended question, no key provided*

**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:**





1. TAKE A SCREENSHOT (2-2e)

*Key screenshot:*

Row Labels	Sum of Sales_Order_Quantity_Sold
Imperial IPA	61
Imperial Stout	258
IPA	235
Pale Ale	116
Stout	190
Wheat	119
<b>Grand Total</b>	<b>979</b>

1. TAKE A SCREENSHOT (2-2f)

**Key screenshot:**

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.

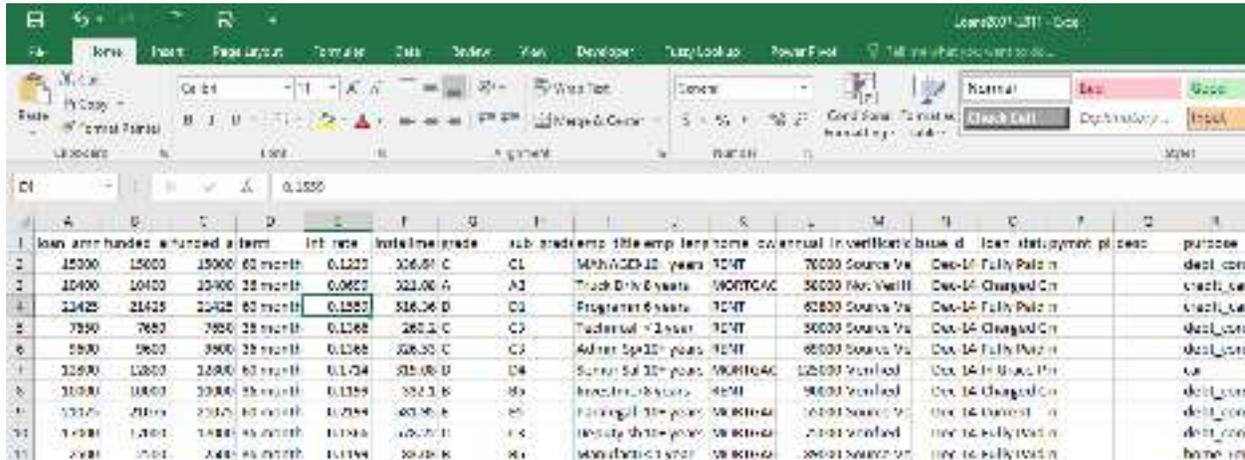
*The columns with the headers revol\_bal\_joint, sec\_app\_earliest\_cr\_line, sec\_app\_inq\_last\_6mths, sec\_app\_mort\_acc, sec\_app\_open\_acc, sec\_app\_revol\_util, sec\_app\_open\_act\_il, sec\_app\_num\_rev\_accts, sec\_app\_chargeoff\_within\_12\_mths, sec\_app\_collections\_12\_mths\_ex\_med, and sec\_app\_mths\_since\_last\_major\_derog can also be deleted.*

Next, fix your numbers:

*Copyright © 2019 McGraw-Hill Education. All rights reserved. No reproduction or distribution without the prior written consent of McGraw-Hill Education.*

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:**



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.

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

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	loan_amt	funded_a	funded_a	term	int_rate	installmer	grade	sub_grade	emp_title	emp_len	home_ow	annual_in	verification
2	15000	15000	15000	60	0.1239	335.64	C	C1	MANAGER	10	RENT	78000	Source Ver
3	10400	10400	10400	36	0.0699	321.08	A	A3	Truck Driv	8	MORTGAG	58000	Not Verifi
4	21425	21425	21425	60	0.1559	515.30	D	D1	Programmer	6	RENT	63800	Source Ver
5	7650	7650	7650	36	0.1365	200.2	C	C3	Technical	0	RENT	50000	Source Ver
6	9600	9600	9600	36	0.1365	325.53	C	C3	Admin Spl	10	RENT	69000	Source Ver
7	12800	12800	12800	60	0.1714	319.08	D	D4	Senior Sal	10	MORTGAG	125000	Verified
8	10000	10000	10000	36	0.1199	332.1	B	B5	Investmer	8	RENT	90000	Verified
9	21075	21075	21075	60	0.2199	581.95	L	L5	Paralegal	10	MORTGAG	55000	Source Ver
10	17000	17000	17000	36	0.1365	578.22	C	C3	Deputy sh	10	MORTGAG	75000	Verified
11	2500	2500	2500	36	0.1199	83.03	B	B5	Manufact	0	MORTGAG	89000	Source Ver
12	5250	5250	5250	30	0.1144	172.98	B	B4	Store Man	2	RENT	26000	Not Verifi

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.*

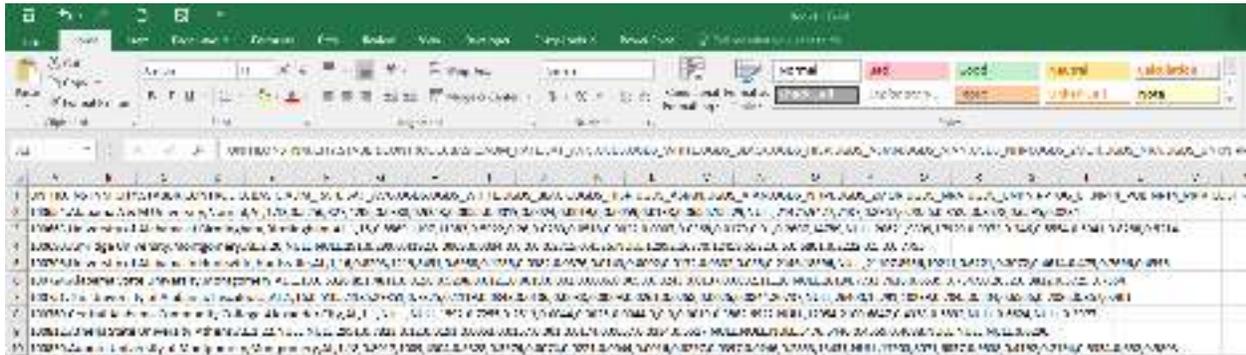
**END OF LAB**

## (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:

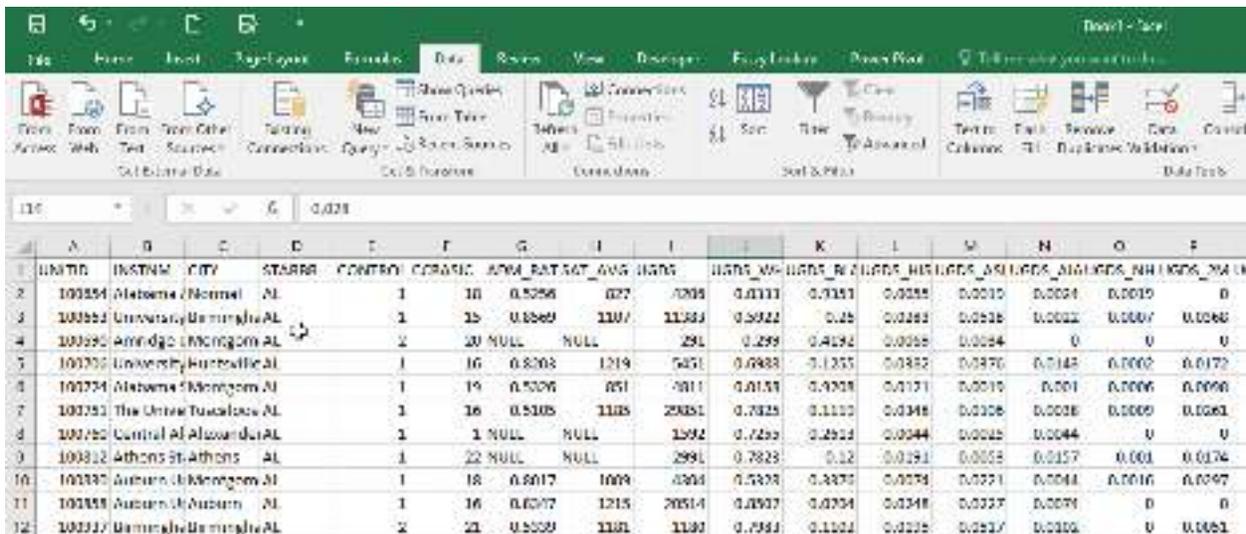


Q1. By looking through the data in the text file, what do you think the delimiter is?

Comma

#### 2. Take a screenshot (2)

Screenshot Key:



- 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?

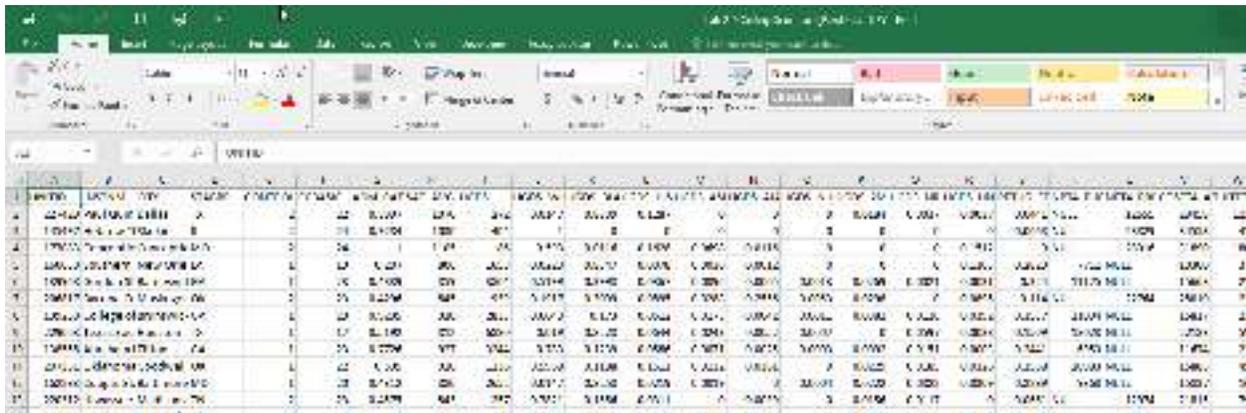
*It would inaccurately represent a very low SAT average across all schools (Correct Answer)*

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:**



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**

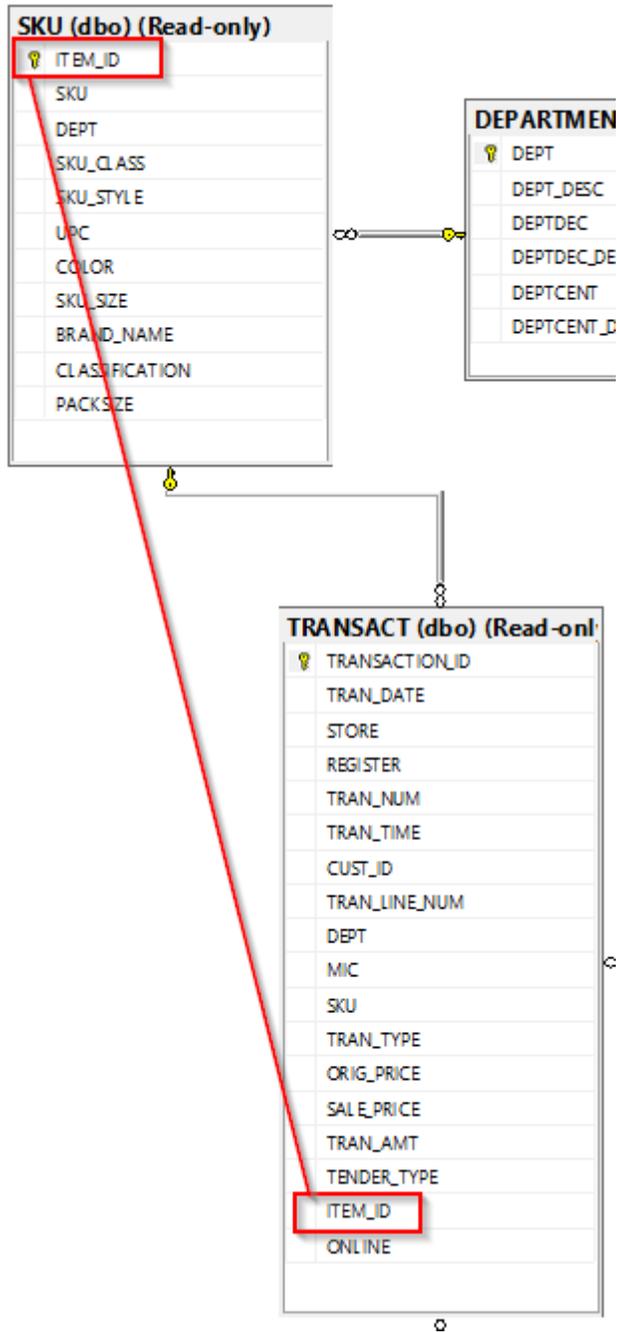
Q3. What is the primary key for the TRANSACT table? What is the primary key for the SKU table?

**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



TRANSACTION_ID	TRANSACTION_DATE	STORE_ID	REGION_ID	QUANTITY	TRANSACTION_AMOUNT	TRANSACTION_AMOUNT_CUST_ID	TRANSACTION_AMOUNT_DEPT	MERCHANDISE_ID	TRANSACTION_TYPE	ORIGINAL_PRICE	SALES_PRICE	TRANSACTION_AMOUNT	TRANSACTION_AMOUNT_CUST_ID	ITEM_ID	ORDER_ID
1	2014-01-01	140	11	40	1111	10001111	1	100	P	27.75	27.75	1111	10001111	10001111	10001111
2	2014-01-01	140	11	20	1111	10001111	1	100	P	27.75	27.75	1111	10001111	10001111	10001111
3	2014-01-01	140	11	100	1111	10001111	1	100	P	27.75	27.75	1111	10001111	10001111	10001111
4	2014-01-01	140	11	20	1111	10001111	1	100	P	27.75	27.75	1111	10001111	10001111	10001111
5	2014-01-01	140	11	44	1111	10001111	1	100	P	27.75	27.75	1111	10001111	10001111	10001111
6	2014-01-01	140	11	72	1111	10001111	1	100	P	27.75	27.75	1111	10001111	10001111	10001111
7	2014-01-01	140	11	81	1111	10001111	1	100	P	27.75	27.75	1111	10001111	10001111	10001111
8	2014-01-01	140	11	10	1111	10001111	1	100	P	27.75	27.75	1111	10001111	10001111	10001111
9	2014-01-01	140	11	8	1111	10001111	1	100	P	27.75	27.75	1111	10001111	10001111	10001111
10	2014-01-01	140	11	27	1111	10001111	1	100	P	27.75	27.75	1111	10001111	10001111	10001111

- 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
CO	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
OH	15
OK	10
SC	8
TN	10
TX	61
UT	11
VA	6
WY	1
<b>Grand Total</b>	<b>313</b>

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 to Discussion Questions**

1. 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?
7. 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.

- i. 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
- l. 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

<b>SECTION 1: TO BE COMPLETED BY CUSTOMER/REQUESTOR</b>			
<b>Name</b>	Vernon Richardson	<b>Contact Number</b>	479-345-0537
		<b>Email Address</b>	vjricha@uark.edu
<b>Description of Information Required</b> <i>(Please include dates/timeframes for any analysis, and other specific indicators/categories required in the data)</i>	From the College Scorecard data, I need the following data items for each year and unique identifier (UNITID) of the dataset: <ul style="list-style-type: none"> <li>a. UNITID – a unique identifier for the institution</li> <li>b. STABBR – State postcode</li> <li>c. COSTT4_A – Average cost of attendance</li> </ul>		
<b>Purpose/Context</b> <i>(what the data is required for)</i>	I am trying to determine if different regions of the country have significantly different costs of attendance.		
<b>Frequency</b> <i>(circle as appropriate)</i>	<i>One Off Request / Other</i>	<i>Annually</i> X	<i>Term</i>
<b>Request Date</b>	7/15/2020	<b>Required Date</b>	7/28/2020
<b>Format Required</b> <i>(Table, Spreadsheet, Word, etc.) – please specify</i>	Tab Delimited File	<b>Customer</b> <i>(if not requestor)</i>	
<b>To be used in</b> <i>(presentation, report, etc.) – please specify</i>	Report to University of Arkansas administration	<b>Intended Audience</b> <i>(if appropriate)</i>	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.