**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***

1. Select a **frequency**. In this case this is a “One-off request”.
2. Enter a **request date** (today) and a **required date** (one week from today)
3. Choose a **format** (spreadsheet).
4. Finally complete the **To be used in** box (internal analysis).
5. **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***

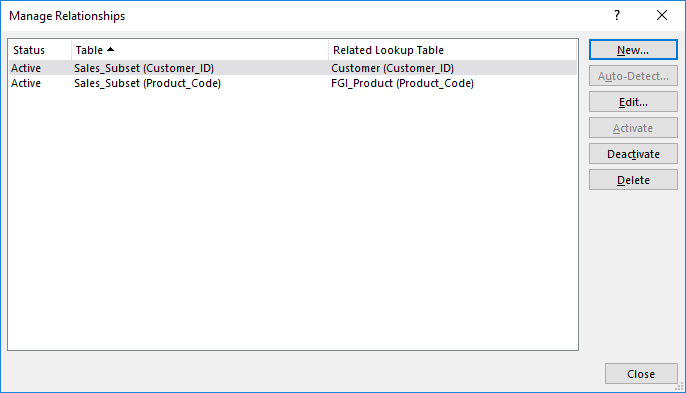
**(Level 2 Header) Part 2: Master the data: Prepare data for analysis in Excel**

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.



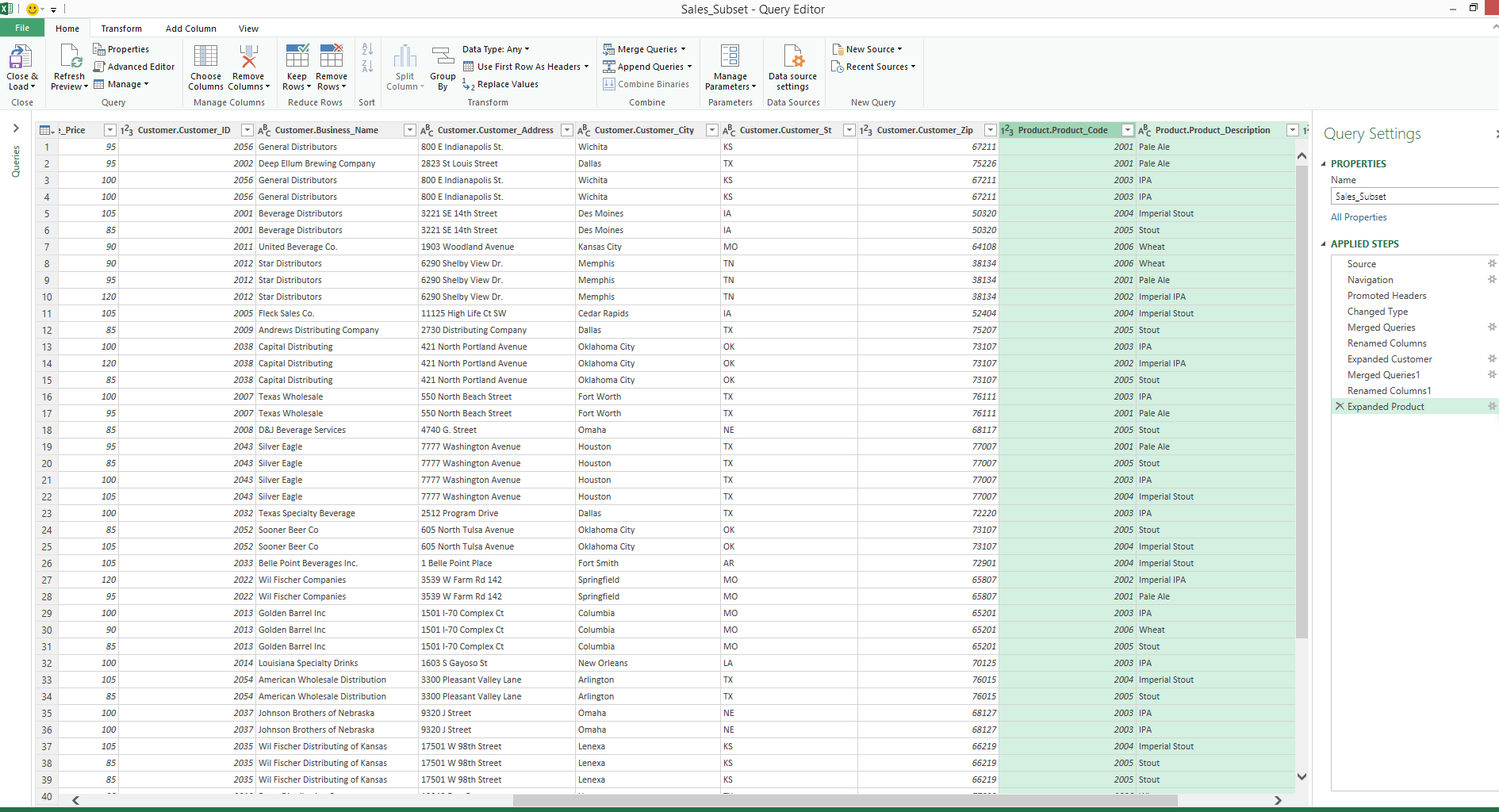
Q3. How comfortable are you with identifying primary key-foreign key relationships?

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

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

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

***KEY Screenshot:***



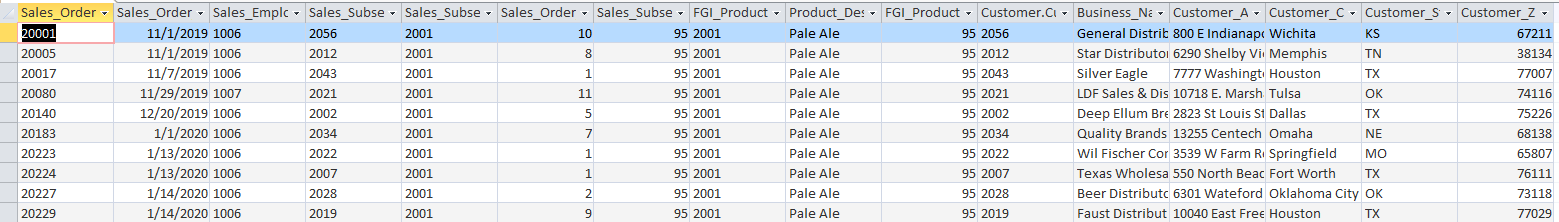
Q4. Have you used the Query Editor in Excel before? Double-click the [Sales\_Subset] query and click through the tabs on the ribbon. Which options do you think will be useful in the future?

***KEY: Open-ended question, no key provided.***

**Alternative 4: Use SQL queries in Access**

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

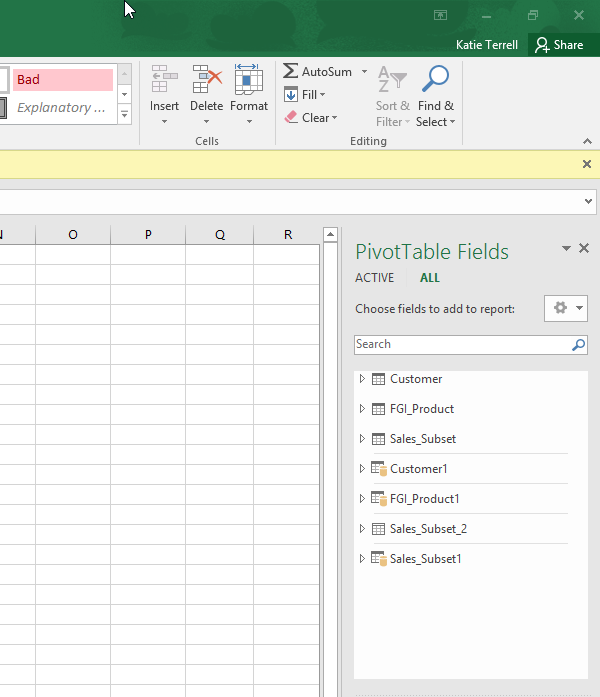
***KEY: Screenshot***

******

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

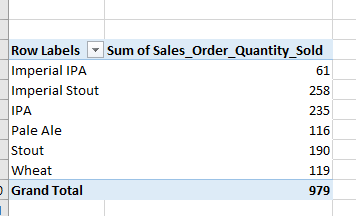
1. **TAKE A SCREENSHOT (2-2d)**

***KEY SREENSHOT:***

****

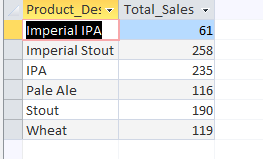
1. **TAKE A SCREENSHOT (2-2e)**

***Key screenshot:***

****

1. **TAKE A SCREENSHOT (2-2f)**

***Key screenshot:***

****

1. 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:

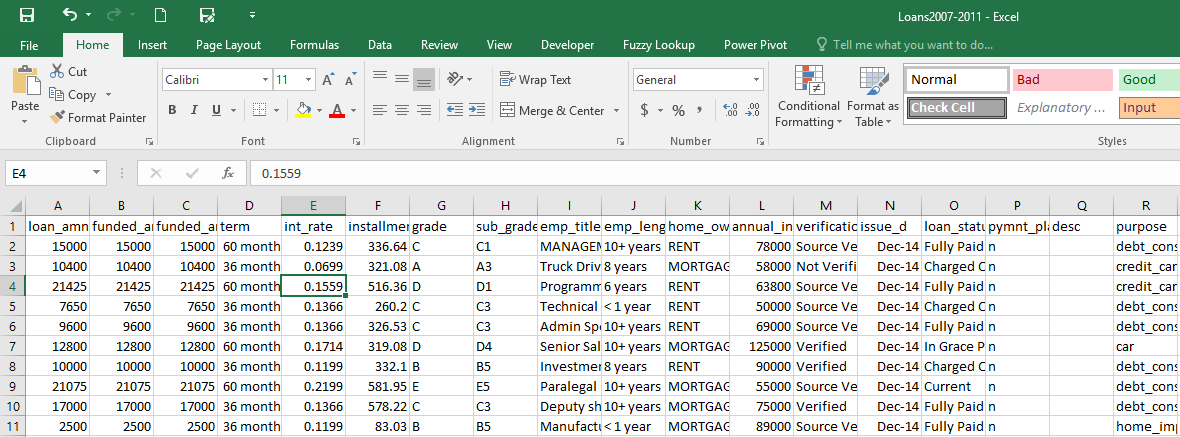
1. Save your file as “Loans2007-2011.xlsx” to take advantage of some of Excel’s features.
2. Delete the first row that says “Notes offered by prospectus…”
3. Delete the last four rows that include “Total amount funded…”
4. Delete columns that have no values, including [id], [member\_id], [url]
5. Repeat 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:***



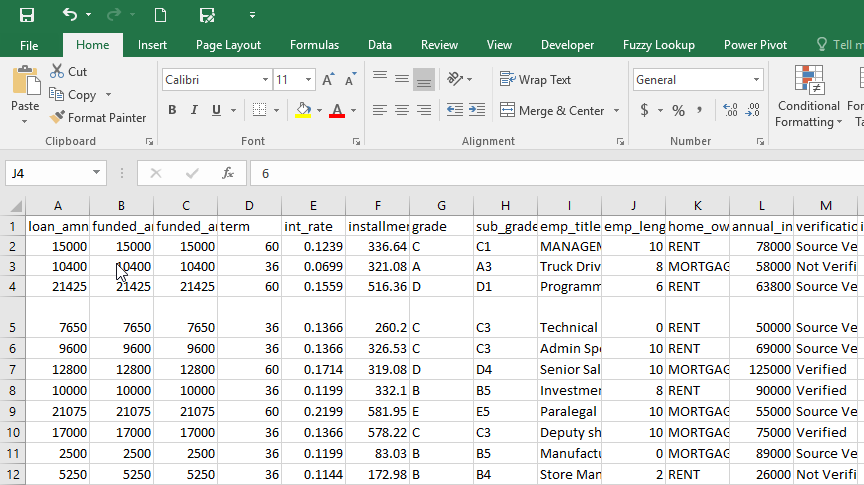
Then, remove any words from numerical values:

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.***

1. **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.***

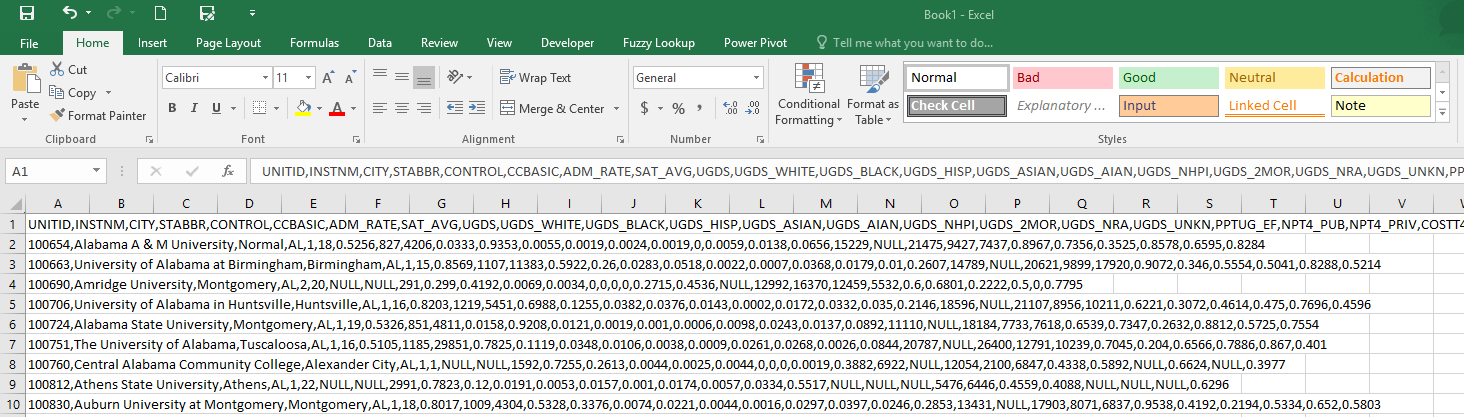
**END OF LAB**

**(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:***

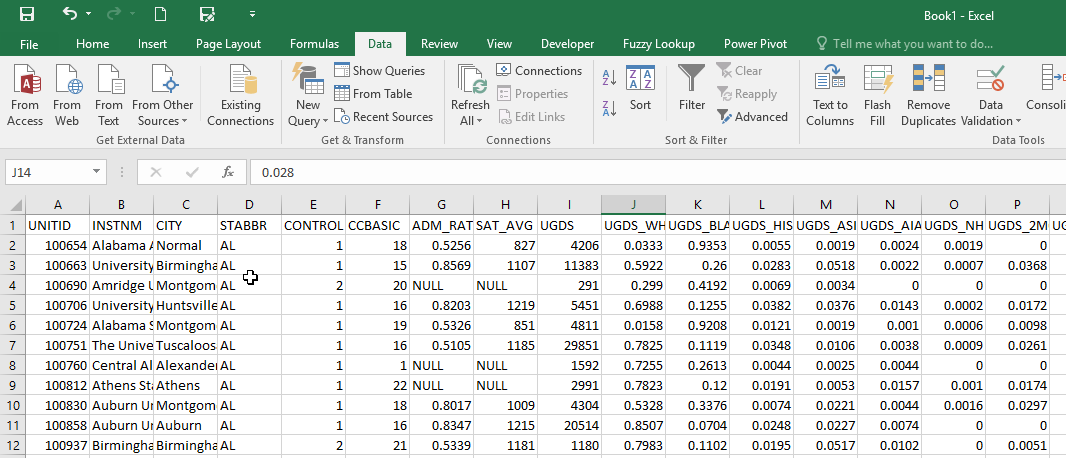
****

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

***Comma***

1. **Take a screenshot (2)**

***Screenshot Key:***

****

1. 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?

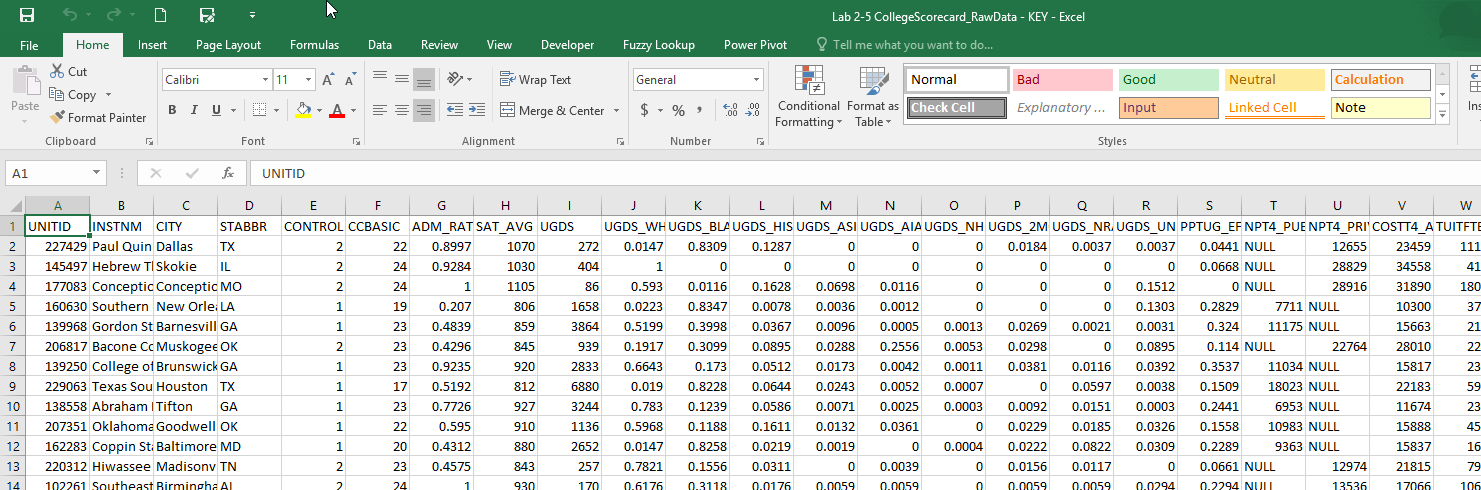
***It would inaccurately represent a very low SAT average across all schools (Correct Answer)***

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

***No***

1. 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.
2. 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.
3. **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**

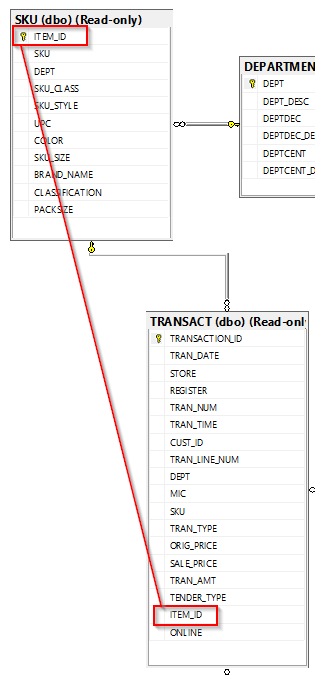
Q3. What is the primary key for the TRANSACT table? What is the primary key for the SKU table?

***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?

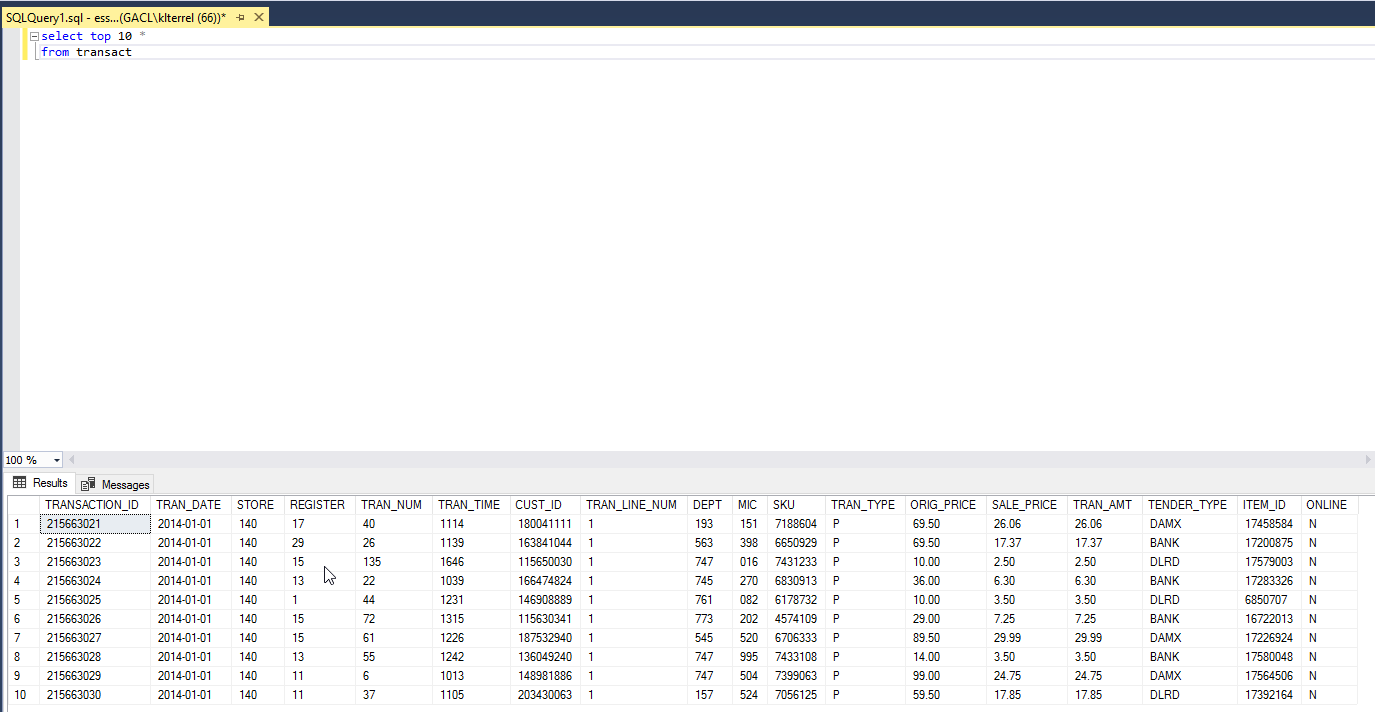
***Open-ended question, no key provided.***

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***

****

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**

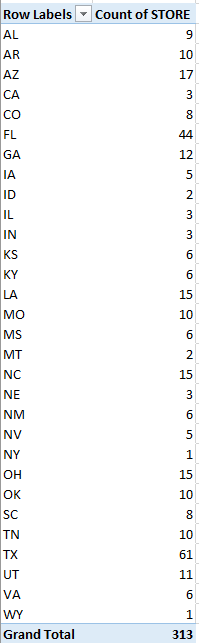
Q1. What can you do in Excel that is much more difficult to do in other data management programs?

***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.

****

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.***