# Introduction

Now you have been asked to analyze the May 2003 transactions. Unfortunately, these transactions are in two separate files, and the data is not as neat as the April 2003 transactions. You must perform additional steps to get the data into a state that you can analyze readily. Once you repair the transaction data, you extract the data to new tables. You can obtain information from several tables and combine it in a single view.

You should have imported the following source data using the import data wizard:

1. Acceptable \_Codes
2. Company\_Department
3. Employees
4. Transactions\_April
5. Transactions\_May\_1
6. Transactions\_May\_2
7. Unacceptable\_Codes

You can also use

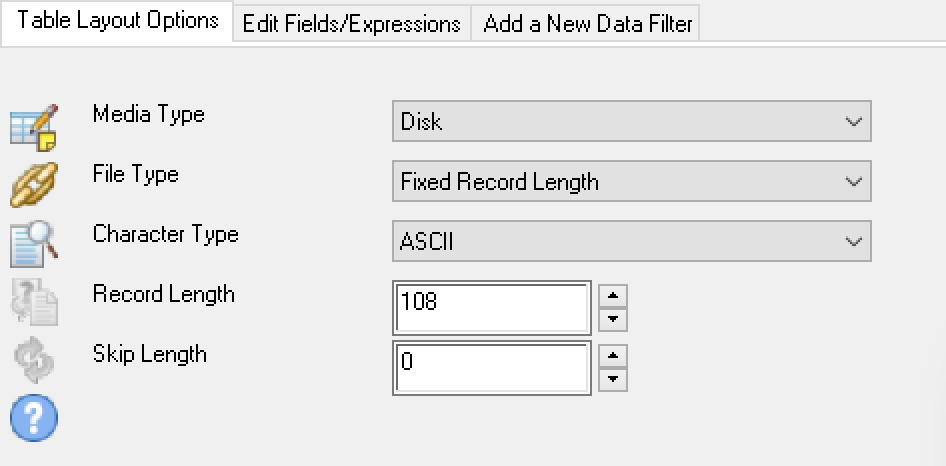
# Issues in This Lab

It was very important that you adjust the data types for certain columns in the tables. These changes are:

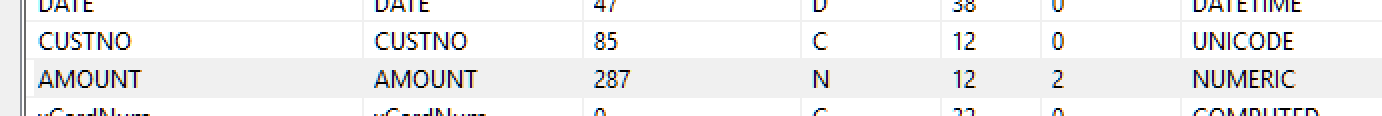
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Column Name | Data Type | Start | Length | Decimal | Notes |
| CardNum | UNICODE | 1 | 32 |  |  |
| CODES | UNICODE | 33 | 8 |  |  |
| DATE | DATETIME | 41 | 38 |  | PICTURE "YYYY-MM-DD" |
| CUSTNO | UNICODE | 79 | 12 |  |  |
| AMOUNT | NUMERIC | 91 | 12 | 2 |  |

1. The DESCRIPTON column in the May tables is of two different lengths. The placement of the column in the table also creates complications. You don’t have to extract this column as it makes the harmonization of the tables MUCH easier.

Figure 1 - Edit Table Layout - Table Layout Options



1. Also note the total length of each record also matters. There is a difference in record length between May 1 and May 2 transaction tables.
2. The length and precision of the column AMOUNT differs between May 1 and May 2. You should make sure these are the same length and level of precision – numeric 12,2 works.



1. The lab references ASCII, but UNICODE requires twice the space as ASCII. ASCII lengths should be doubled to accommodate the necessary UNICODE values.

## ACL Functions

One important method to ensure that you have the correct length is to use the ACL functions. For ASCII and UNICODE columns:

SUBSTR(ALLTRIM(COLUMN NAME),[DESIRED START], [DESIRED LENGTH])

For example:

SUBSTR(ALLTRIM(CARDNUM),1,16)

# Other Data Type Issues

Other tables have data type issues as well.

|  |  |  |
| --- | --- | --- |
| Table Name | Column | Final Data Type |
| CREDIT\_CARDS | CARDNUM | ASCII Text |
|  | DESCRIPTION | ASCII Text |

# Cleansing Data

## Tasks

1. Import the May Transaction Data (p. 65) – document and submit your results

Figure 2 - Raw Data Import - May 2003 Tab 1

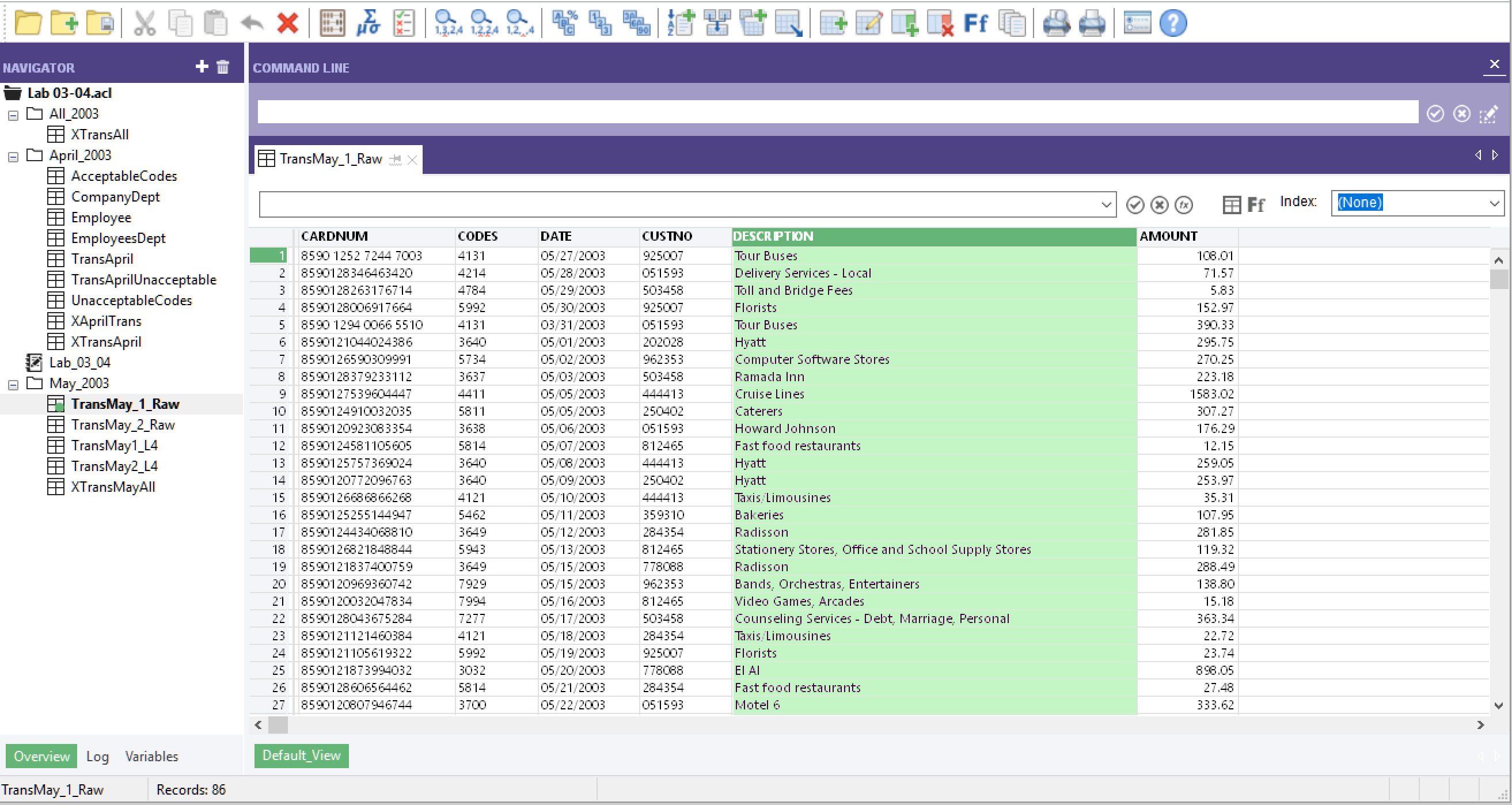
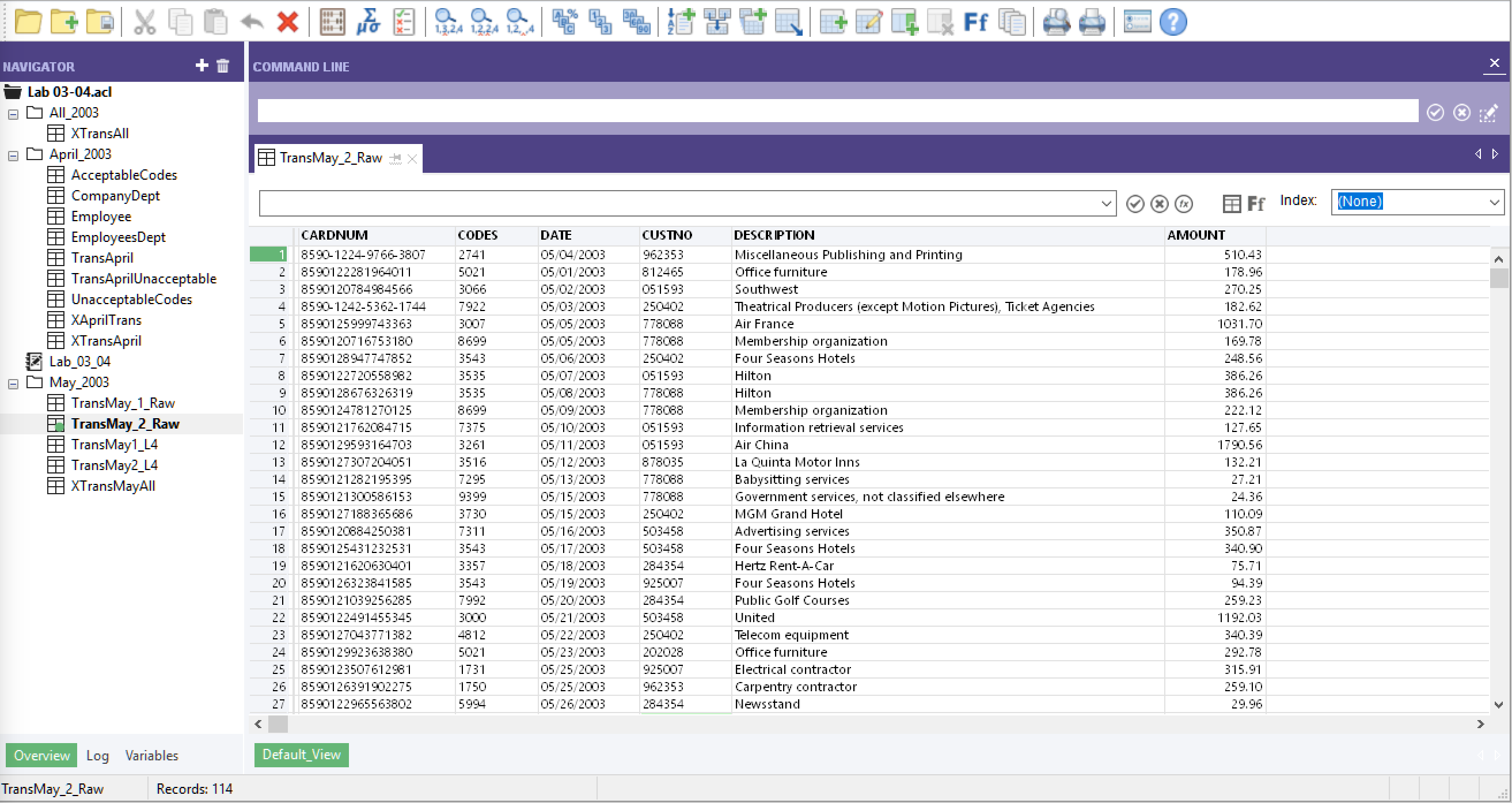
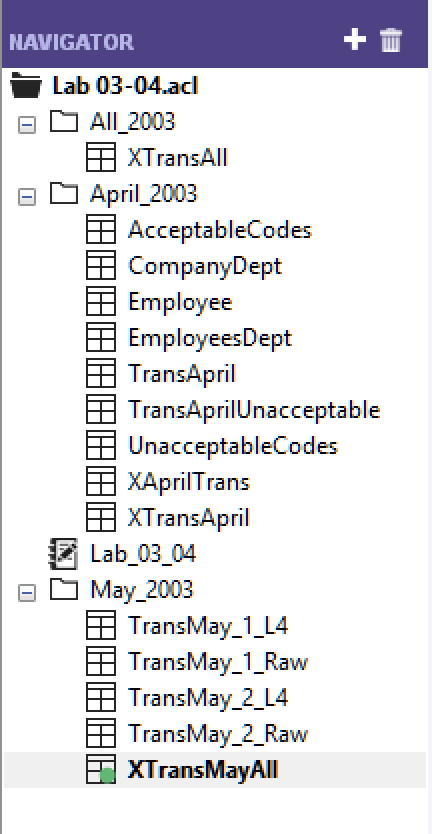


Figure 3 - Raw Data Import - May 2003 Tab 2



1. Use Folders to Organize Your Project (p. 65) – document and submit your results



1. Compare and Correct Table Layouts (p. 66) – document and submit your results

Figure 4 - Correct Amount - Type, Length, and Precision

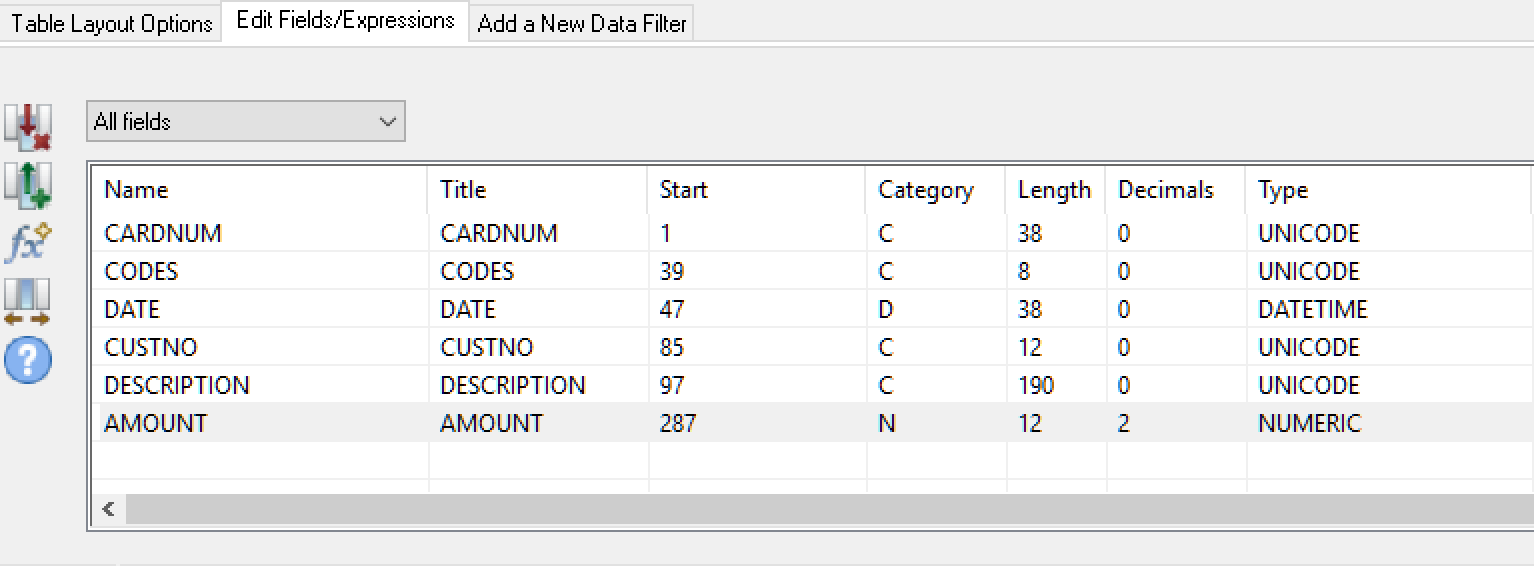
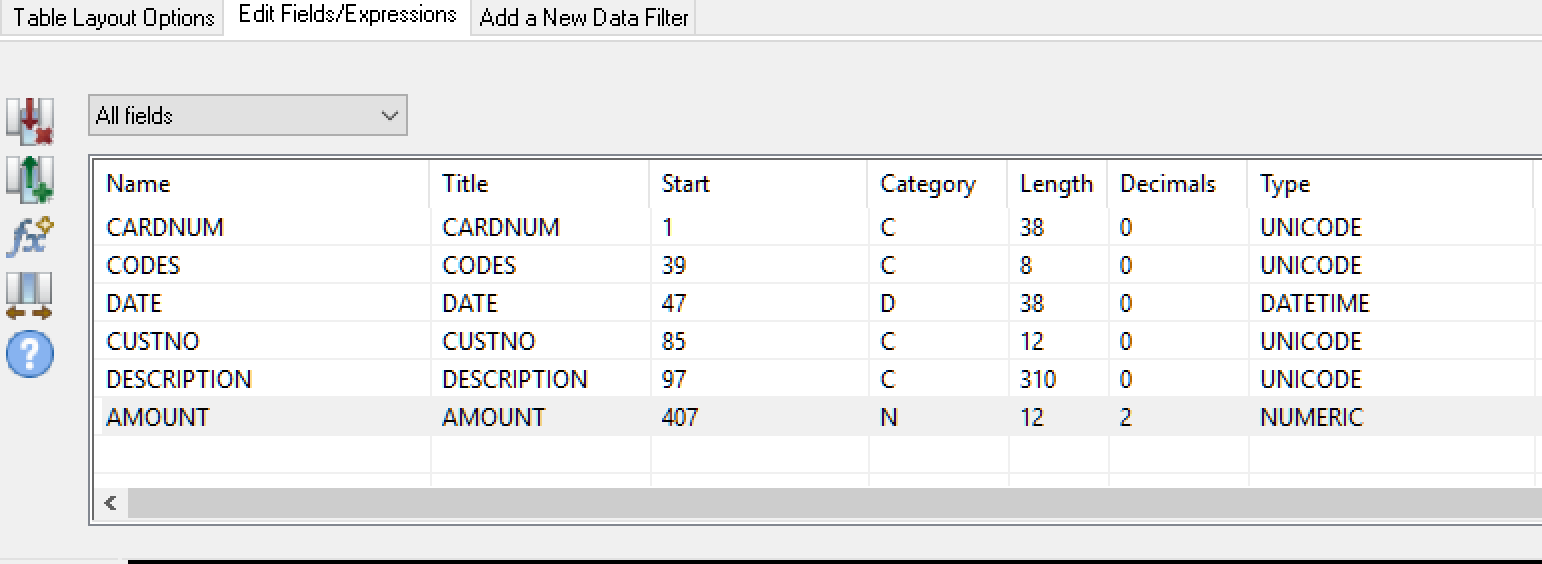


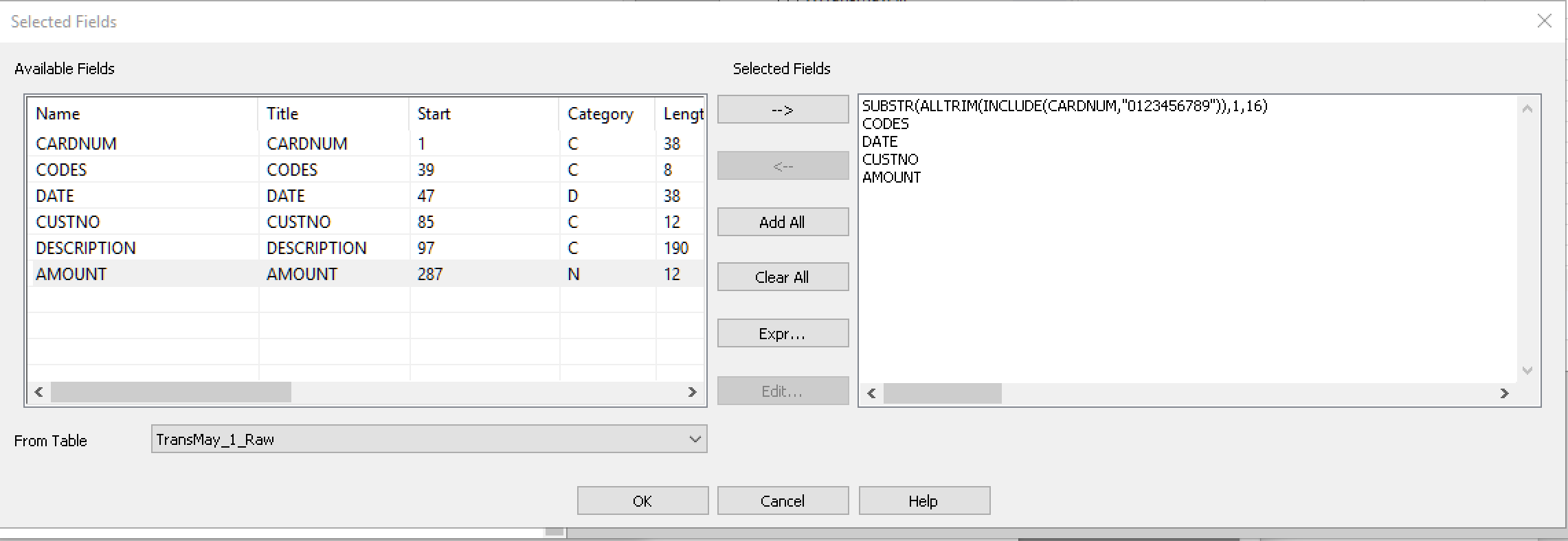
Figure 5 - Correct Amount - Type, Length, and Precision



Notice the length of DESCRIPTION column is different between the two tables. Amount was larger with a length of 18. It was changed to 12 and this should be sufficient.

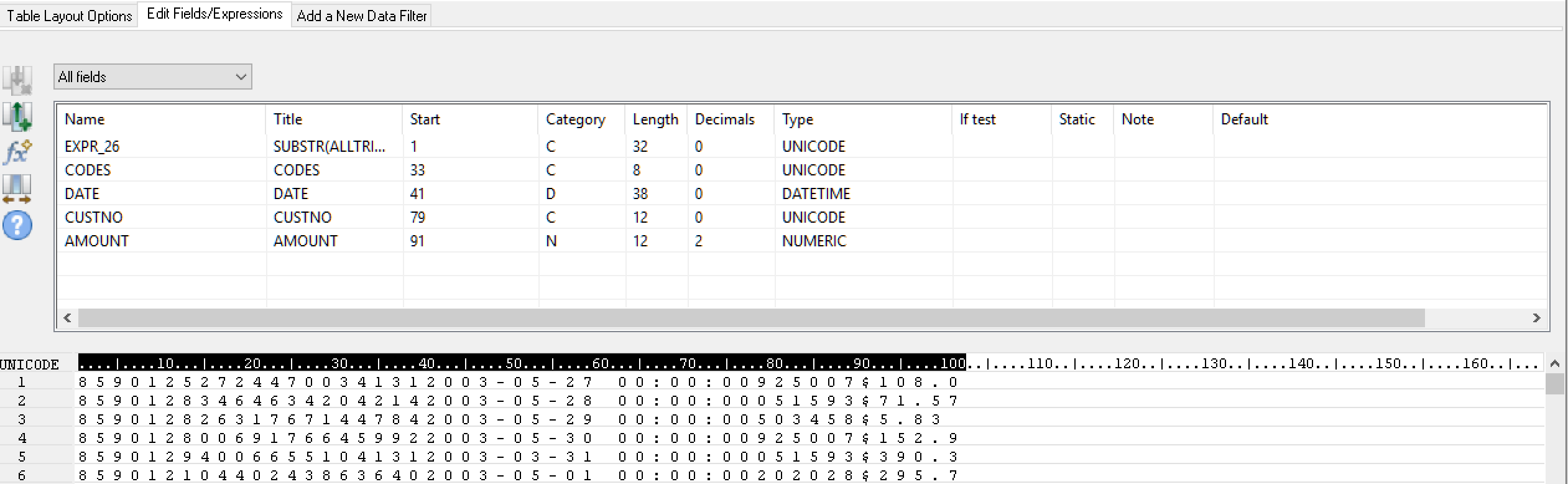
1. Combine and Verify the Tables (p. 67) – document and submit your results

Figure 6 - Extract and Transform Data in One Step

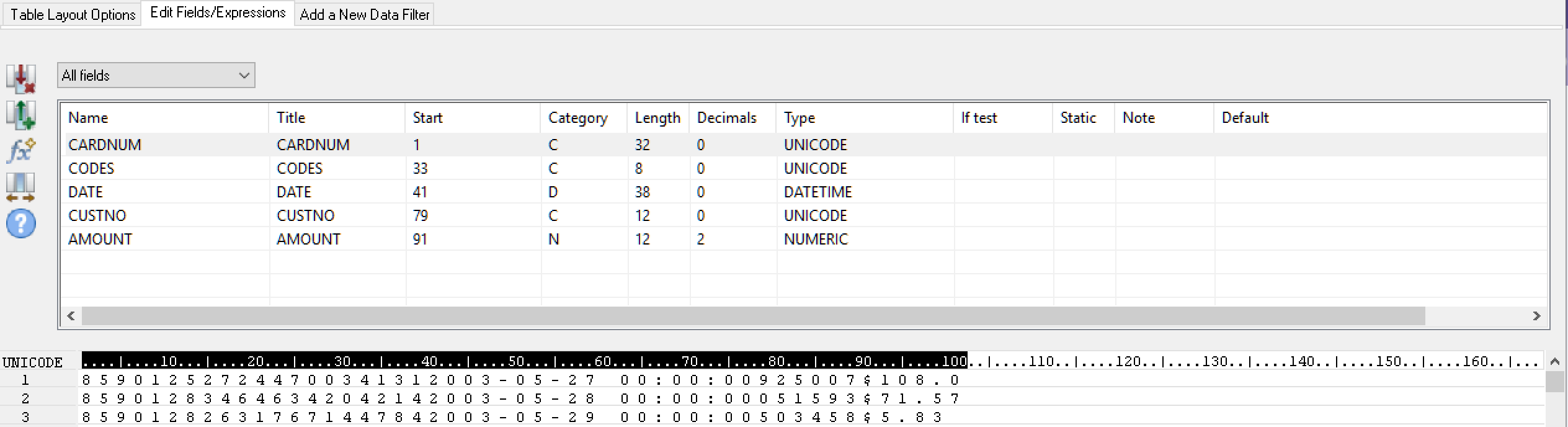


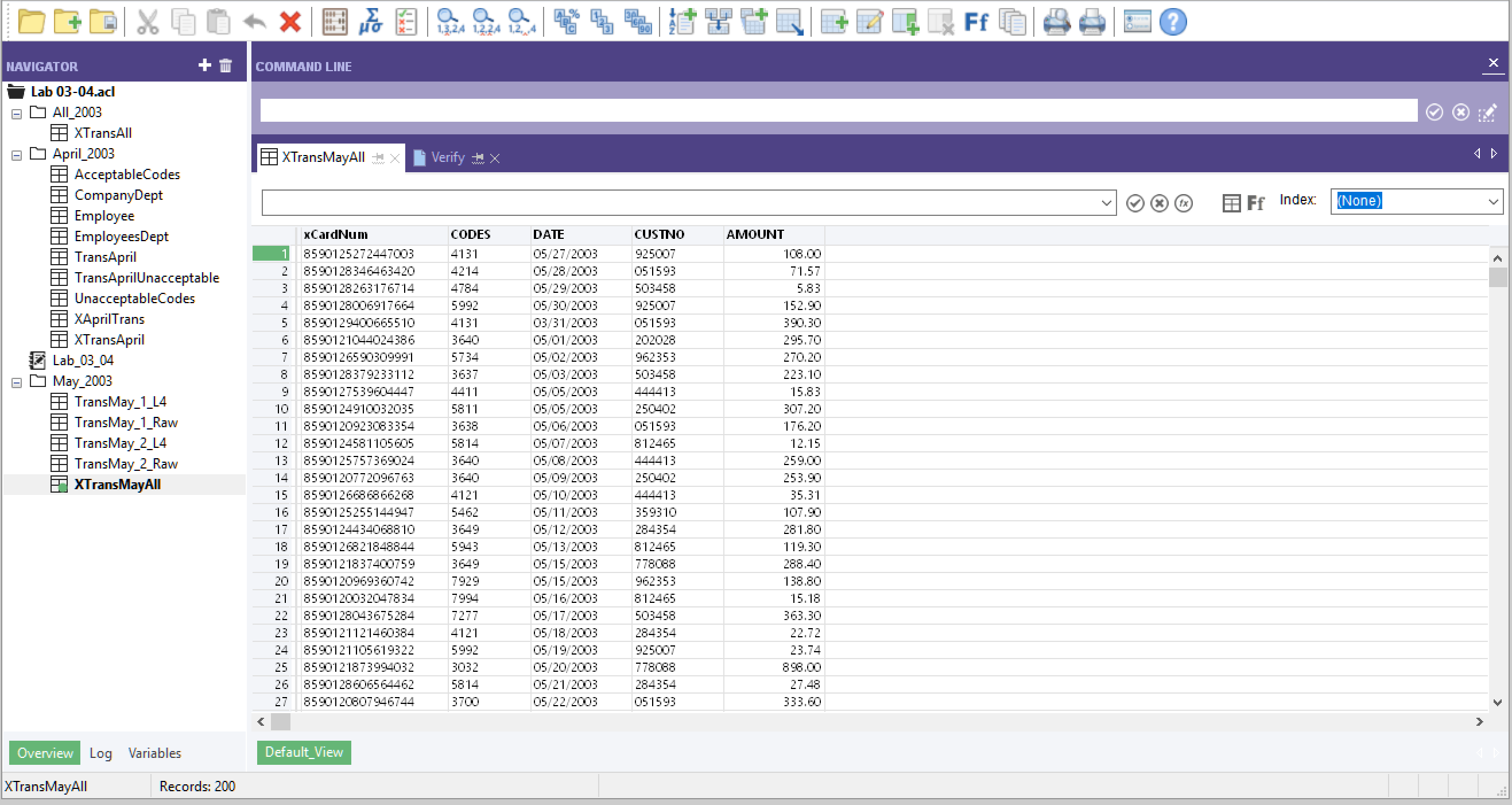
You can use ACL functions to transform the data as you extract it:

SUBSTR(ALLTRIM(INCLUDE(CARDNUM,"0123456789")), 1, 16)

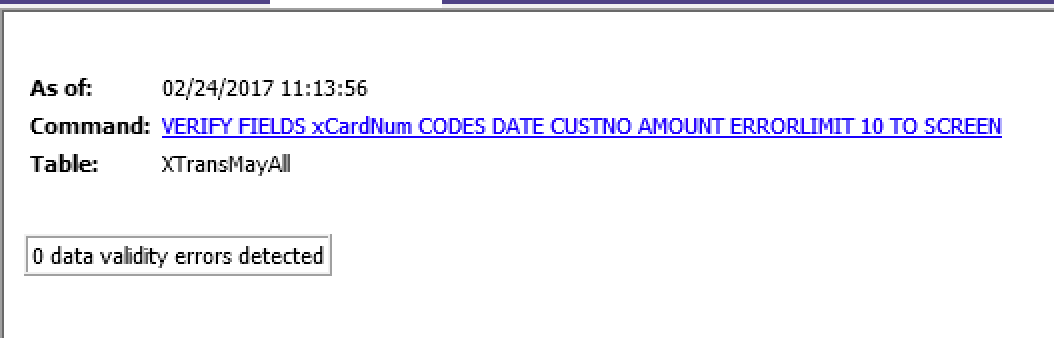


Edit the table layout to rename the columns

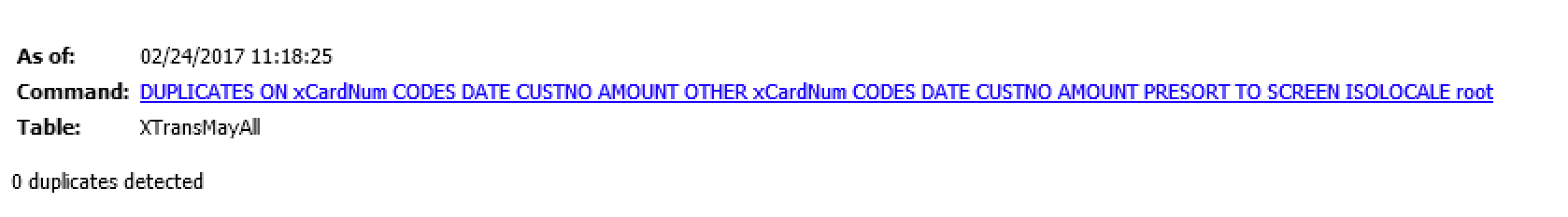




1. Verify the new table (p. 68) – document and submit your results

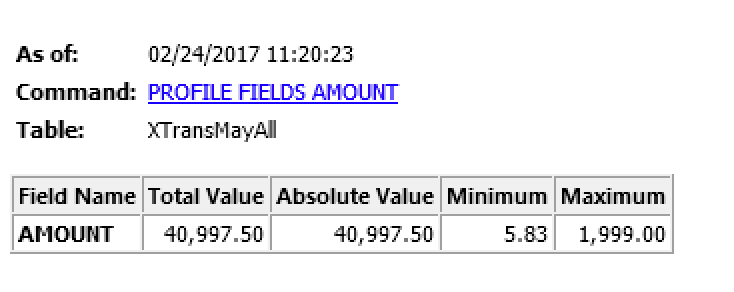


1. Check the table for duplicates (p. 69) – document and submit your results



## Exercises

1. Run the Profile command on the Trans\_May\_All table. Note the highest and lowest transaction values – document and submit your results

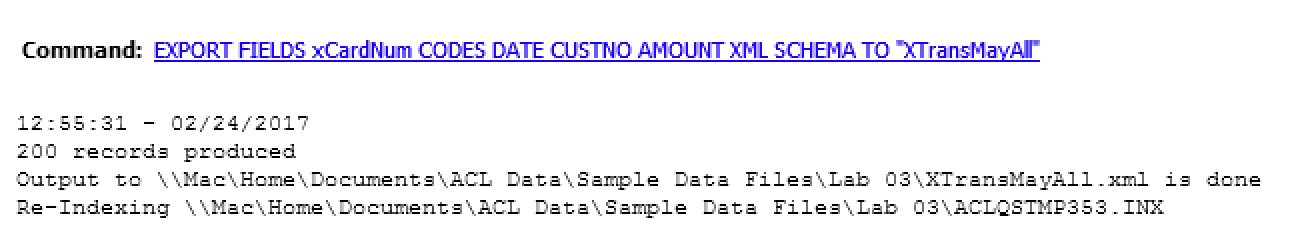


1. Check for duplicates in the CUSTNO field. Classify the table on the same field. Are the results significant? – document and submit your results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | |  |  |
|  | | |  | | |

CUSTNO 50348 has the highest percentage of charges and the largest expense amount.

1. Export a table to XML that lists the employee name, card number, and total dollar value of associated transactions per-card in May. – document and submit your results



1. View the file in Internet Explorer or another browser that supports XML viewing. – document and submit your results

## 

## Tasks

1. Isolate and Correct Validity Errors (p.70) – document and submit your results

See Step 4 above.

1. Check the format the card numbers (p. 70) – document and submit your results

See Step 4 above.

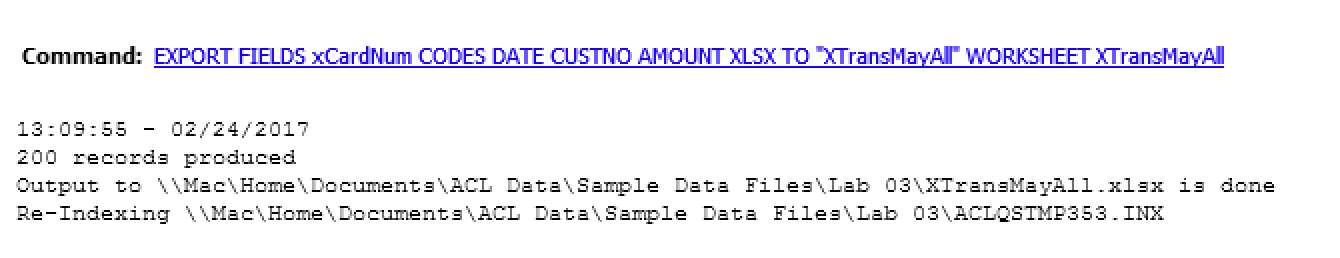
1. Create a 16-digit card number field (p. 71) – document and submit your results

See Step 4 above.

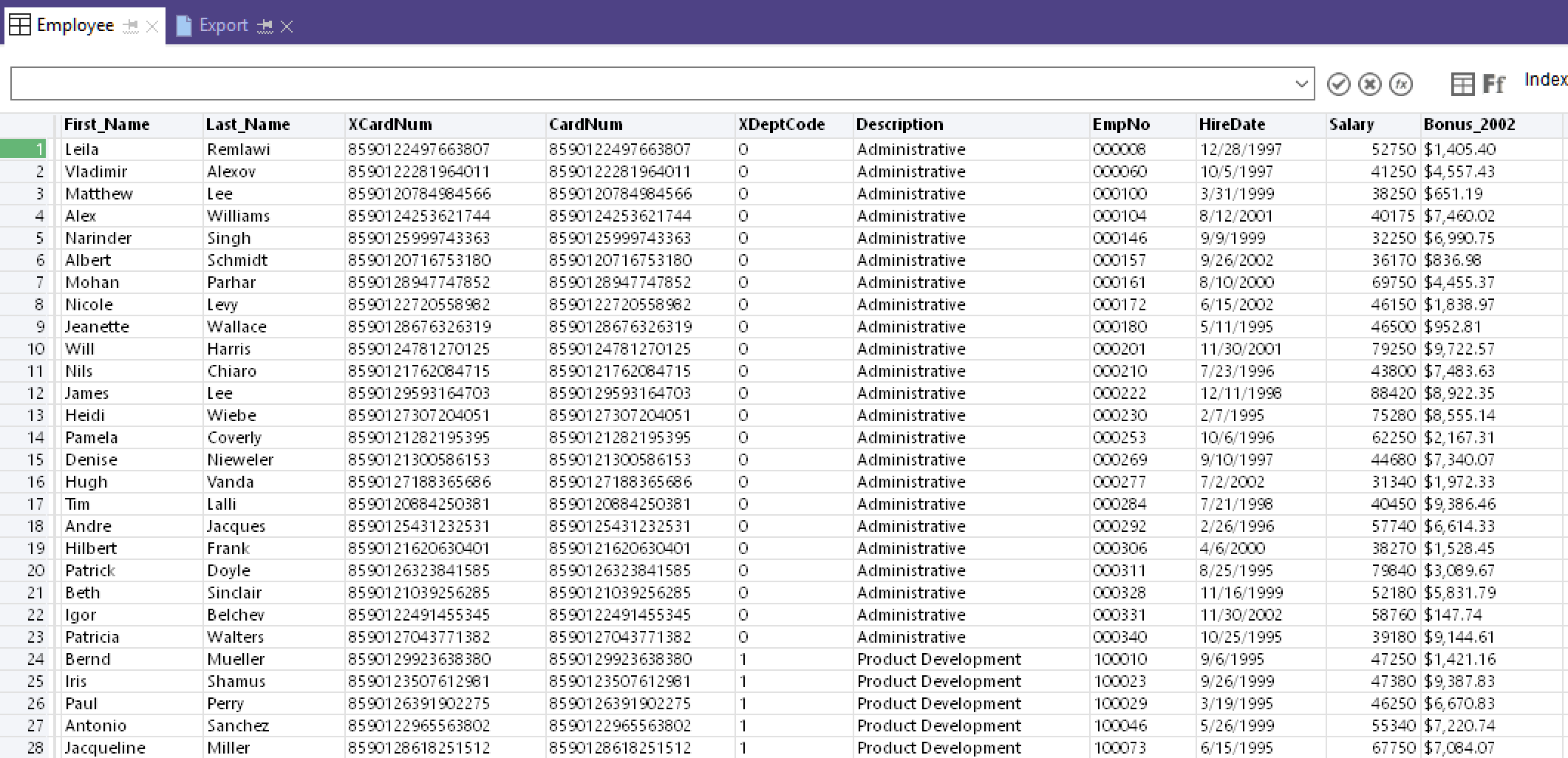
1. Extract the corrected records to a new table (p. 72) – document and submit your results

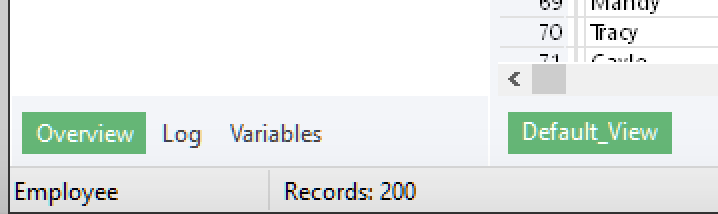
See Step 4 above.

1. Export the Combined Transaction Table to Excel (p. 73) – document and submit your results



1. Create an Error-Free employee record table (p. 73) – document and submit your results

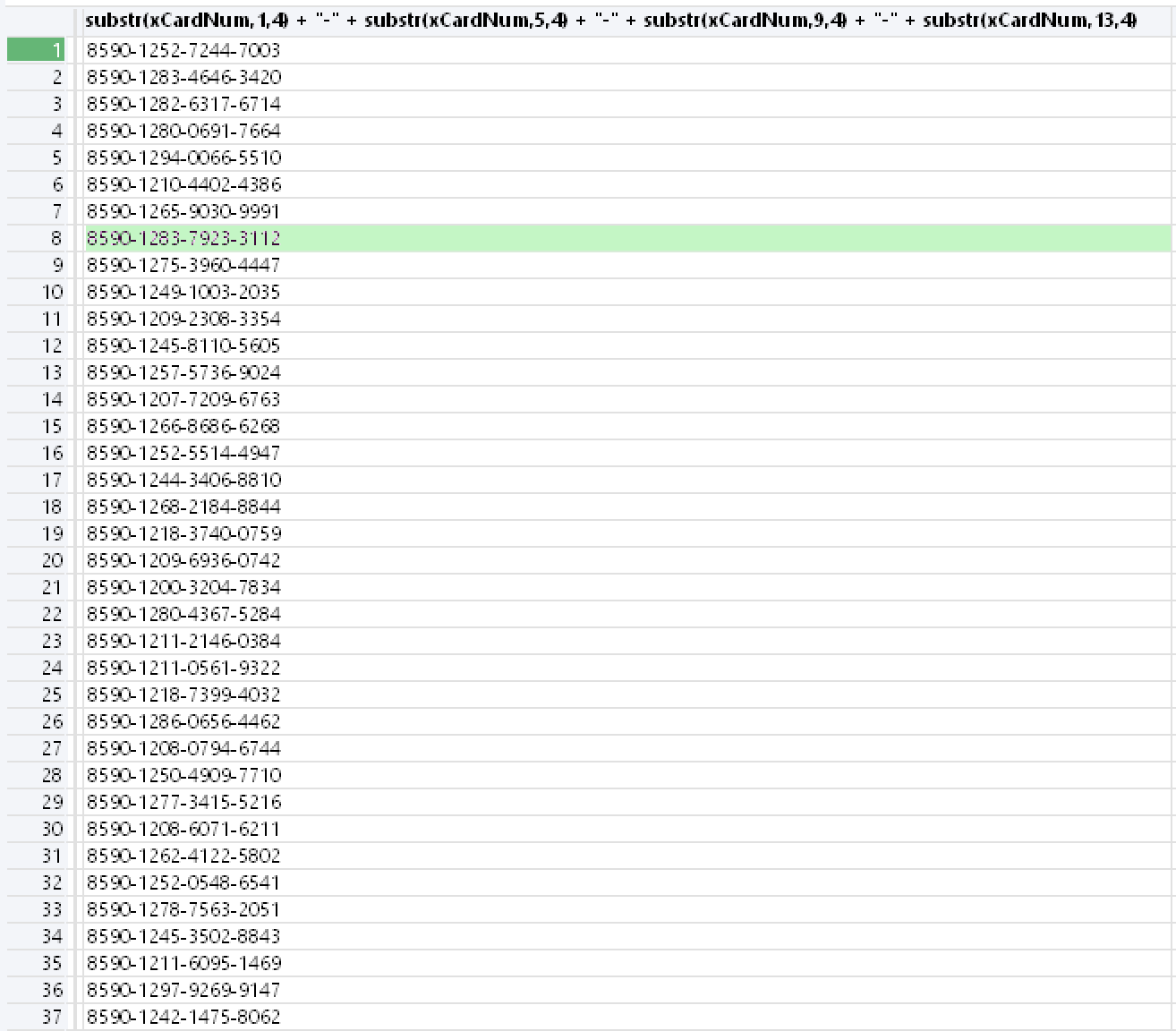




## Exercises

1. Suppose you wanted all card numbers to have the form: xxxx-xxxx-xxxx-xxxx. What strategy would you follow to create a computed field like that? (HINT: SUBSTR() and string + “-“) – document and submit your results.

SUBSTR(xCardNum,1,4) + "-" + SUBSTR(xCardNum,5,4) + "-" + SUBSTR(xCardNum,9,4) + "-" + SUBSTR(xCardNum,13,4)



1. In this tutorial, you have worked through several examples of how to perform analyses using ACL. What other analyses could you perform with the data? – document and submit your results.

Benford Analysis – although the number of records is small.

1. If you had the information available, checking the Social Security Number (SSN) for each employee would help determine whether each record was unique. How would you check for duplicate Social Security Numbers? How would you ensure that all Social Security Numbers were in the format of one long string of digits: xxxxxxxxx? (HINT: ACL Functions) – document and submit your results.

Use the same process as before. ACL functions: SUBSTR(ALLTRIM(INCLUDE(SSN,"0123456789")), 1, 9)

Social Security numbers are nine characters in length – all numbers. Be sure to store SSN as UNICODE.

1. Devise a strategy for analyzing your May transactions that closely mirrors the strategy you followed for the April transactions. Find the unacceptable transactions for this set of transactions. – document and submit your results

Develop a relation to the acceptable and unacceptable tables.

1. For your own projects, think of various approaches that improve the quality of your data before you import it into ACL. – document and submit your results

Plan, plan, clean, and clean

1. What are **two ways** to test your data after you import it into ACL. – document and submit your results

Column length, column format, validity checking

## Further Analysis

1. Does the cardholder have transactions occurring on weekends or holidays?

|  |  |
| --- | --- |
|  | CDOW(DATE,3) |

1. Do an unusual number of transactions have rounded values, such as $1000, $200, or $450?

8590128906141710 05/02/2003 503458 1300 1300.00 Linea Rafael

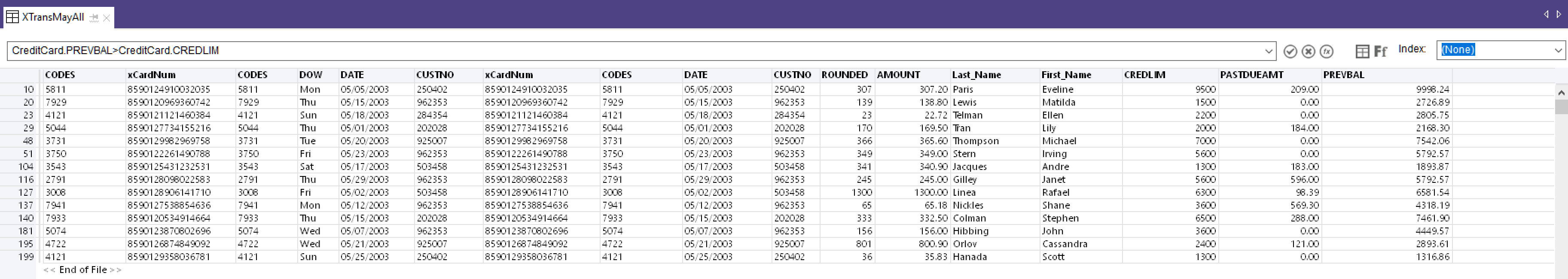
Is one transaction unusual?

1. Are there credits charged back to a card number, and what are the explanations for these credits?

No. None for May. There were chargebacks in April.

1. Does the cardholder regularly come close to or exceed his or her credit limit?

CreditCard.PREVBAL > CreditCard.CREDLIM



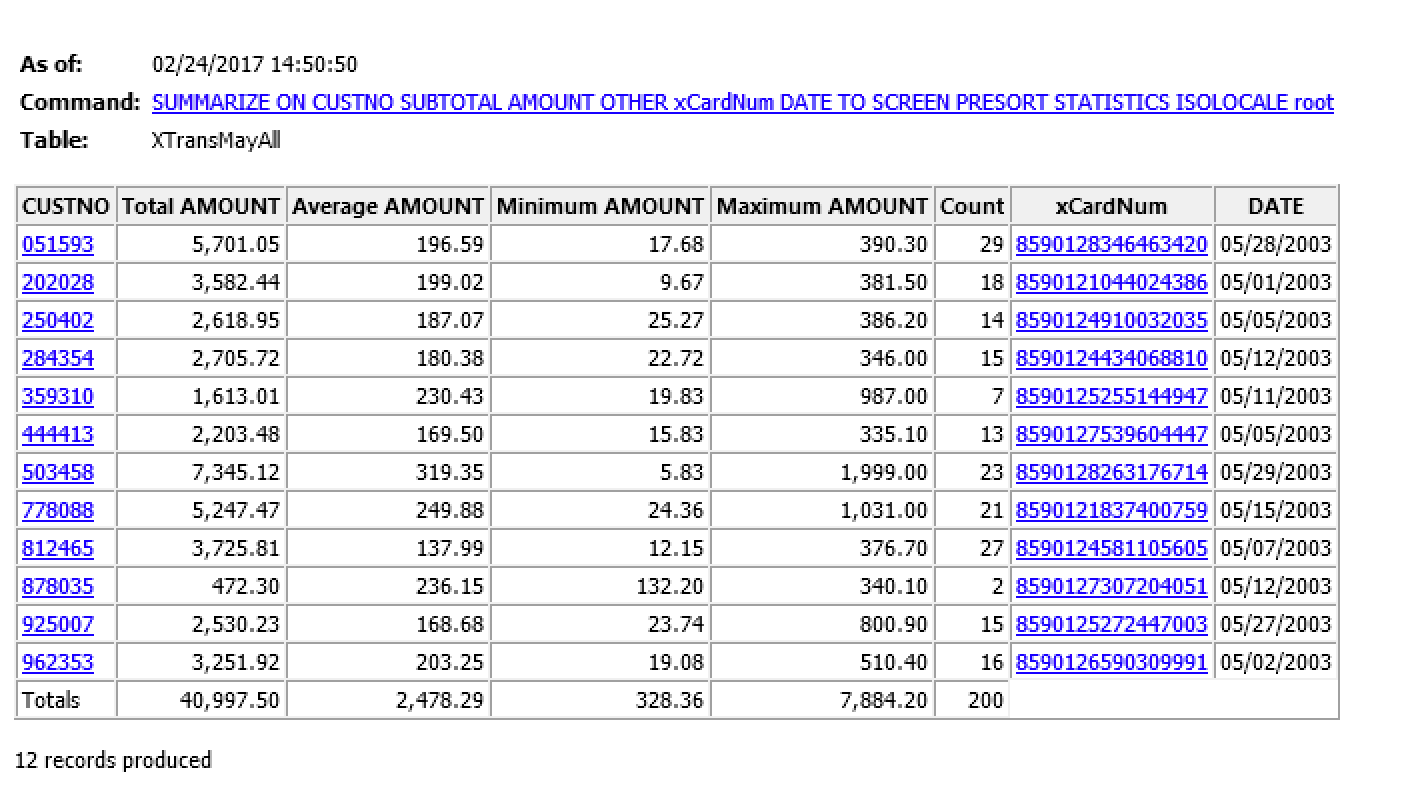
1. Has there been a sudden increase in the cardholder’s purchases?

To answer this question, you would need to compare April with May.

|  |  |
| --- | --- |
|  | Totals Spend: 178,748.43 Transactions: 481 |
|  | Total Spend: 137,750.93  Transactions: 281 |
|  | Total Spend: 40,997.50 Transactions: 200 |

April’s spend was greater than May’s

1. Is the cardholder responsible for charging other people’s expenses?



No, but multiple employee charged the same customers for expenses.

1. Which five or ten employees spent the most? To which customers did they charge their expenses?

