MODUL 9
SELF-SERVICE DATA PREPARATION IN SAS® VIYA®
Cleansing Data Using Data Quality Transforms

THEME DESCRIPTION

Students understand and are able to implement the SAS Data Preparation component to perform data transformation for the required analytical data quality through the SAS Data Studio and SAS Data Quality Transform features on the SAS Viya CAS platform (Cloud Analytics Services).

WEEKLY LEARNING OUTCOMES (SUB-LESSONS)

CLO-4-Sub-CLO-9:
Understand the concepts and technical aspects of the data science field and implement them well through effective analytical and modelling methodologies– C2.

Through the following learning steps**:**

1. Removing Data Inconsistencies Using Data Quality Transforms
2. Combining Data to Create an Analytical Base Table (ABT)
3. Filtering and Transposing Data for Better Insight

PRACTICUM SUPPORT

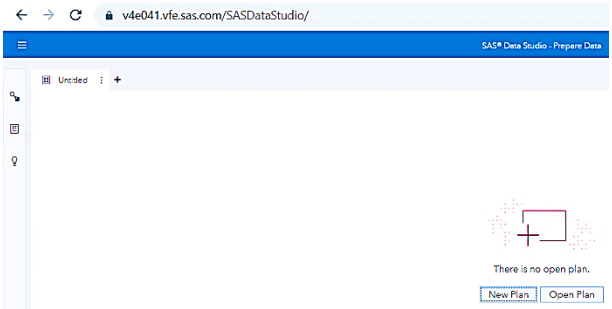
a. Windows Operating System
b. (any) Browser Application

PRACTICUM STEPS

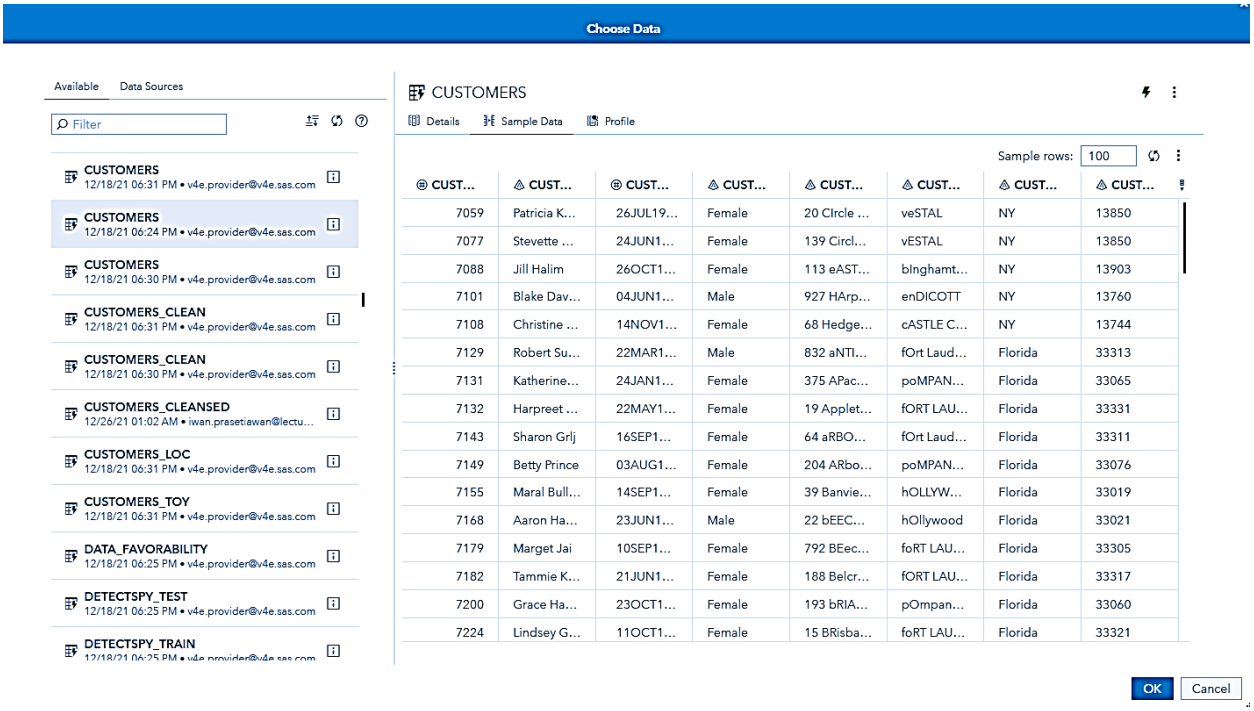
1. Removing Data Inconsistencies Using Data Quality Transforms

- This practice illustrates the use of the Data Quality transforms and the QKB definitions for cleansing data of anomalies and inconsistencies.
- The practice uses the CUSTOMERS table. The data **will be standardized, parsed,** and **placed** into a **desired format** in a target table.

a. Start, https://welcome.oda.sas.com
b. Open SAS®Studio, then open new tab for SAS® Drive to go to SAS Drive page.
c. Click (Show list of applications) and **select Prepare Data to open SAS Data Studio**.
d. Click New Plan.



e. Load the source table. Click the **Available tab**.

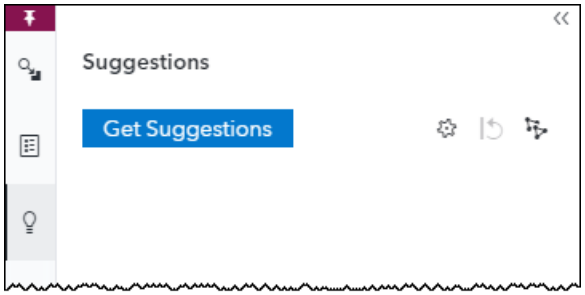


f. Click the CUSTOMERS table. Click OK to select the table as the source table.

CUSTOMER_NAME	CUSTOMER_ADDRESS	CUSTOMER_CITY	CUSTOMER_STATE
Patricia Kukahiko	20 CIrcle On The Green	veSTAL	NY
Stevette Welf	139 Circlebank dr	vESTAL	NY
Jill Halim	113 eAST STREET	bInghamton	NY
Blake Davenport	927 HArps Mill Road	enDICOTT	NY
Christine Deep	68 Hedgerow dr	cASTLE CREEK	NY
Robert Sullins	832 aNTIQUE LN	fOrt Lauderdale	Florida
Katherine Raybould	375 APache Ln	poMPANO BEACH	Florida

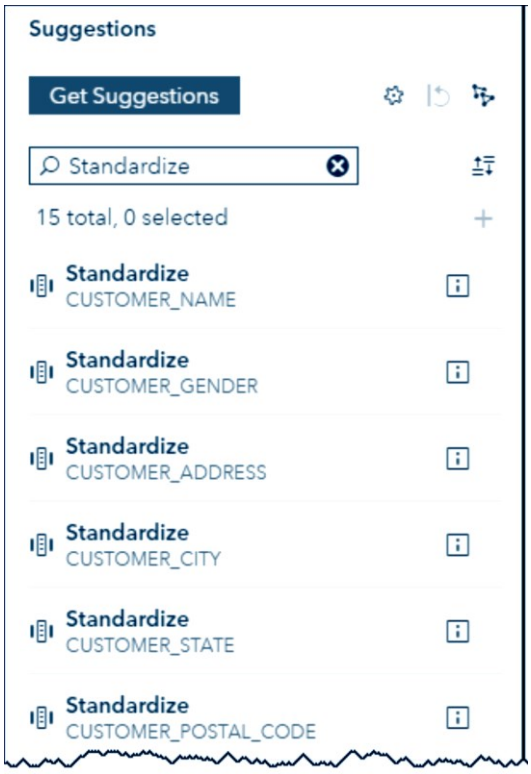
- You can see: many of the values for the CUSTOMER_ADDRESS, CUSTOMER_CITY, and CUSTOMER_STATE columns are cased and organized differently.
- To make sure that we have consistent data, we will first standardize the values in our character columns.

g. Click the Suggestions icon on the left toolbar.



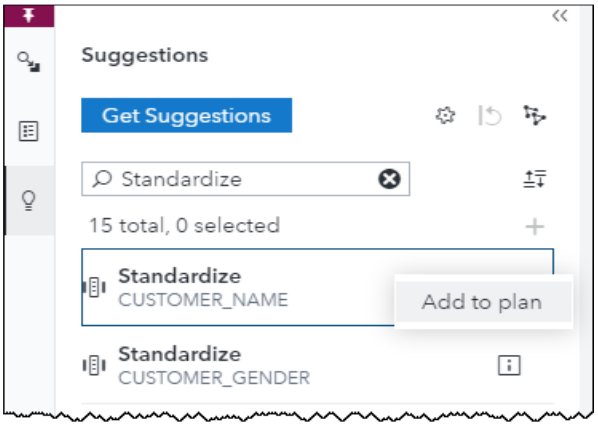
h. Click Get Suggestions.
Before doing any other cleansing operations, we want to standardize our data values.

i. In the Filter box, enter Standardize.

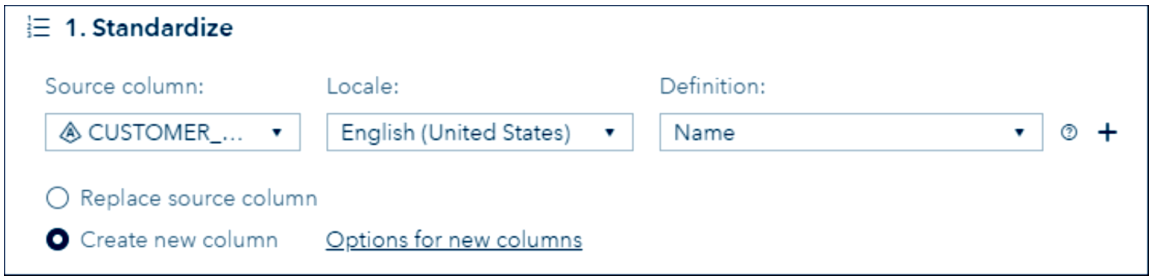


- There are six suggested standardizations for this data set.
- We will perform four only.

j. Right-click the Standardize CUSTOMER_NAME suggestion and select Add to plan to add the standardization to the data preparation plan.



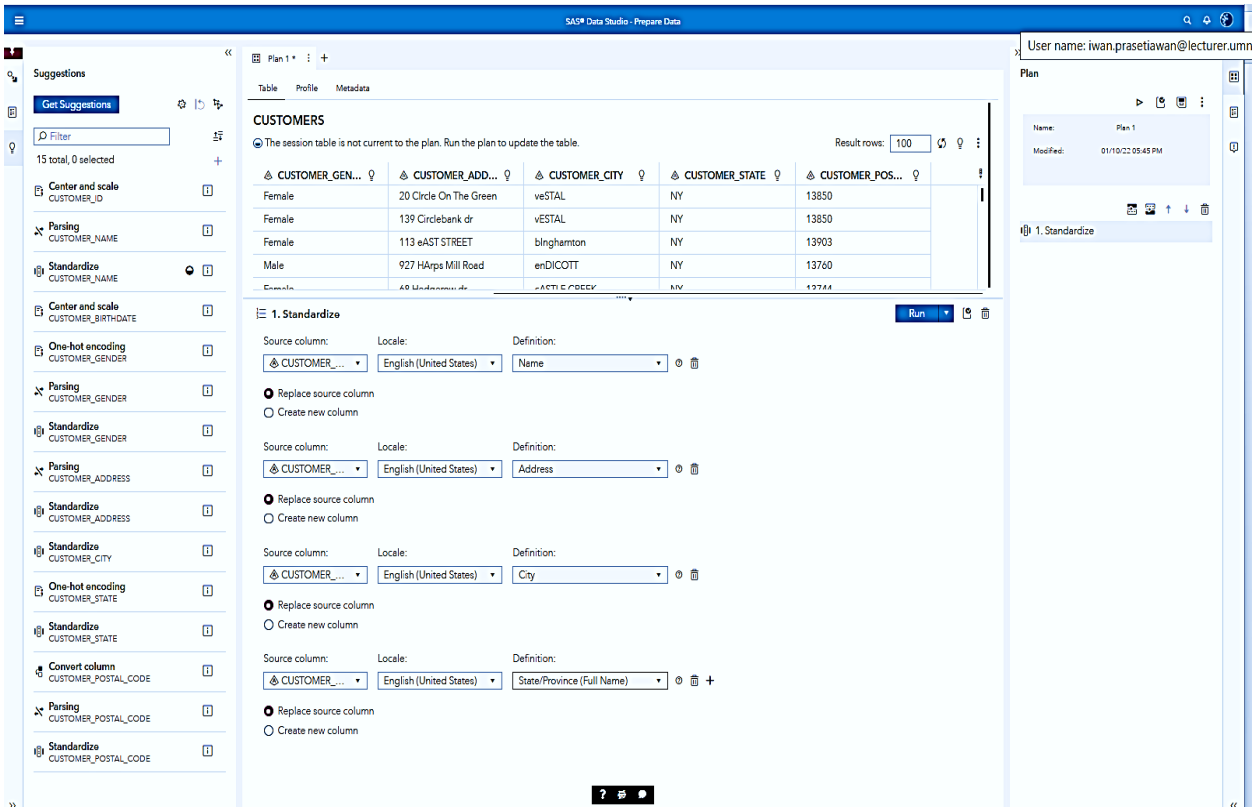
k. Configure the Standardize transform.



- Verify that the CUSTOMER_NAME column is selected as the source column.
- Verify that the Name definition is selected.
- From those definitions, the suggestion guessed the best definition to apply to the selected field.

- The best guess for the type of data in CUSTOMER_NAME was name information.
- The selected standardization definition from the English (United States) locale is called Name.
- Click Replace source column.

l. Add three more Standardize steps to the plan (click "**+**" next to CUSTOMER_NAME):
- the Standardize CUSTOMER_ADDRESS suggestion,
- the Standardize CUSTOMER_CITY suggestion, and
- the Standardize CUSTOMER_STATE suggestion.
- Verify that the CUSTOMER_ADDRESS column will be standardized with the Address definition
- Verify that the CUSTOMER_CITY column will be standardized with the City definition.
- Verify that the CUSTOMER_STATE column will be standardized with the State/Province (Full Name) definition.



m. Click Run and review the results.
- Partial List of Original Data: **data is inconsistent:**

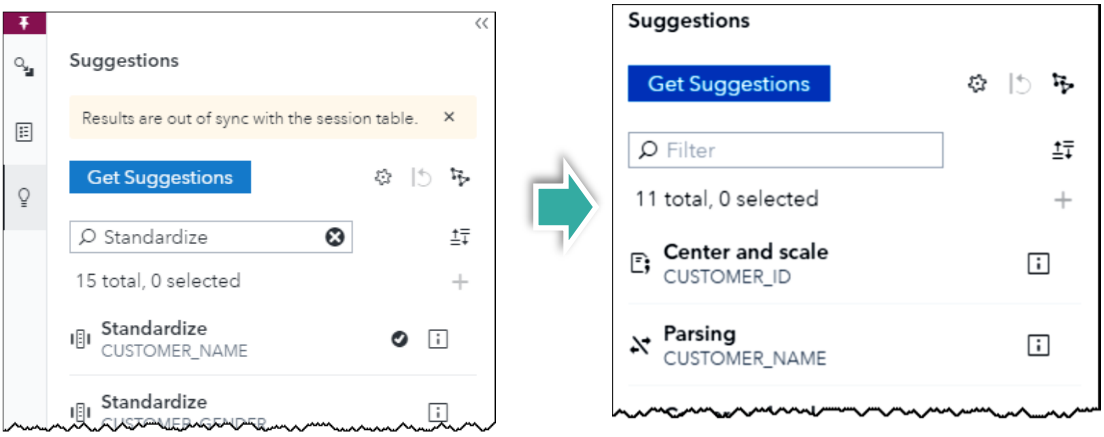
CUSTOMER_NAME	CUSTOMER_ADDRESS	CUSTOMER_CITY	CUSTOMER_STATE
Patricia Kukahiko	20 CIrcle On The Green	veSTAL	NY
Stevette Welf	139 Circlebank dr	vESTAL	NY
Jill Halim	113 eAST STREET	bInghamton	NY
Blake Davenport	927 HArps Mill Road	enDICOTT	NY
Christine Deep	68 Hedgerow dr	cASTLE CREEK	NY
Robert Sullins	832 aNTIQUE LN	fOrt Lauderdale	Florida
Katherine Raybould	375 APache Ln	poMPANO BEACH	Florida

- Partial List of Standardized Data:

CUSTOMER_NAME	CUSTOMER_ADDRESS	CUSTOMER_CITY	CUSTOMER_STATE
Patricia Kukahiko	20 Circle on the Green	Vestal	New York
Welf Stevette	139 Circlebank Dr	Vestal	New York
Jill Halim	113 East St	Binghamton	New York
Blake Davenport	927 Harps Mill Rd	Endicott	New York
Christine Deep	68 Hedgerow Dr	Castle Creek	New York
Robert Sullins	832 Antique Ln	Fort Lauderdale	Florida
Katherine Raybould	375 Apache Ln	Pompano Beach	Florida

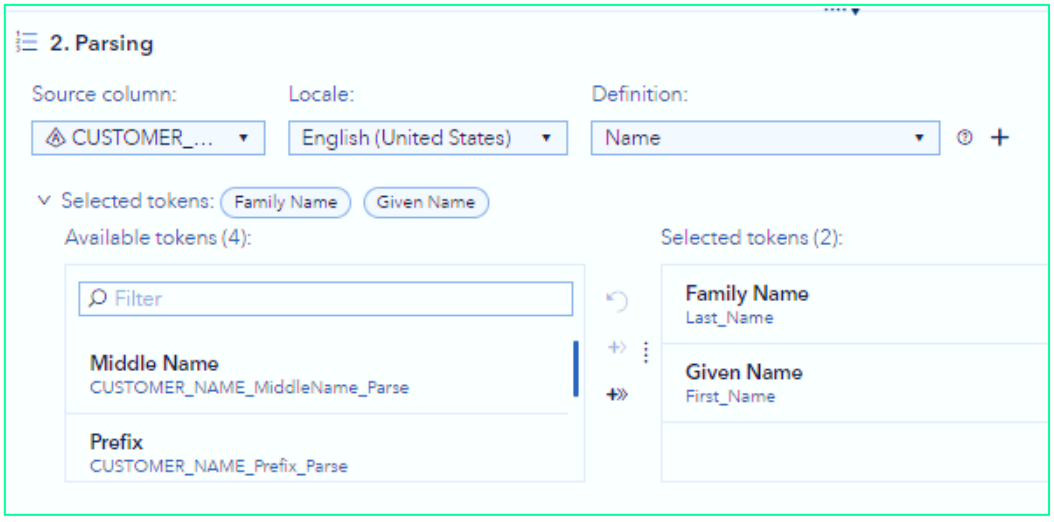
- Now, the standardized data is consistent.
- Consistent casings have been applied and values are displayed in consistent formats.
- For example, full state names are displayed instead of a mix between full state names and two-letter abbreviations.

n. Click the X next to the Filter box to clear the Standardize filter. Click Get Suggestions.



o. Double-click the **Parsing CUSTOMER_NAME suggestion** to add it to the plan.

- Recall that the Parsing transform enables us to break up a string into smaller semantic pieces called tokens.
- For example, a name string might have smaller pieces like a first name, middle name, or last name, which all have meaning separate from the larger value.
- Here, we are going to generate First_Name and Last_Name columns.



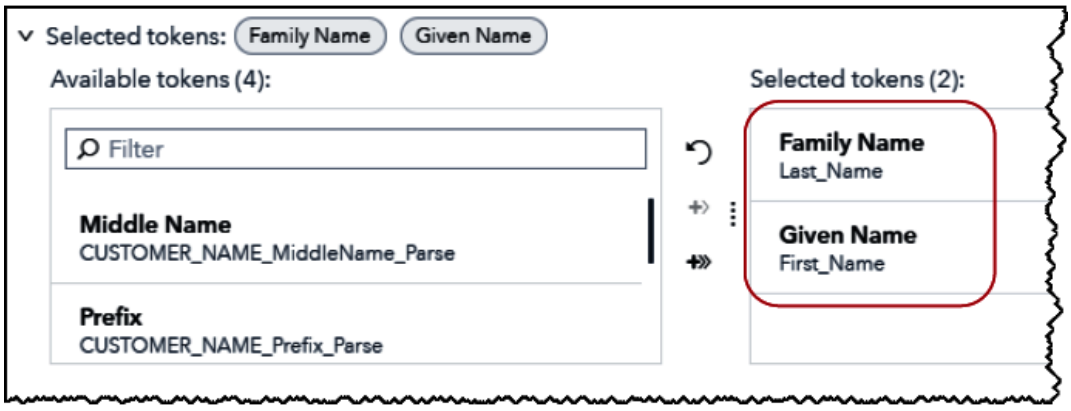
p. Configure the Parsing transform:

- Verify that CUSTOMER_NAME is selected from the Source column list.
- Verify that Name is selected in the Definition list.
- Double-click Family Name.
- Double-click Given Name to move the token to the Selected tokens list box.

q. The default names of the columns are CUSTOMER_NAME_FamilyName_Parse and CUSTOMER_NAME_GivenName_Parse. Let's **make these shorter**:

- Click Options for new columns. This enables us to change the new column defaults.
  1) Enter **Last_Name** in the Name of new column field to replace CUSTOMER_NAME_FamilyName
  2) Change the type to Varchar and the length to 50.
  3) Enter **First_Name** in the Name of new column field to replace CUSTOMER_NAME_GivenName
  4) Change the type to Varchar and the length to 50.

r. Click OK. The token names are changed.



s. Click Run and verify the results:

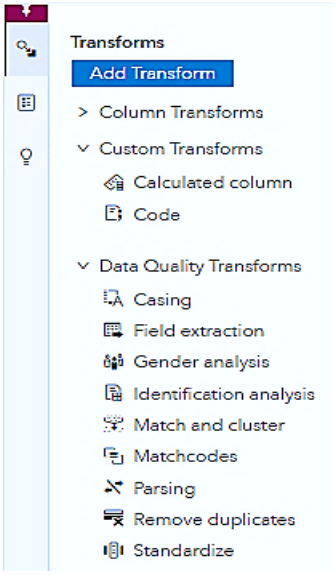
Last_Name	First_Name	CUSTOMER_NAME
Kukahiko	Patricia	Patricia Kukahiko
Stevette	Welf	Stevette Welf
Halim	Jill	Jill Halim
Davenport	Blake	Blake Davenport
Deep	Christine	Christine Deep
Sullins	Robert	Robert Sullins
Raybould	Katherine	Katherine Raybould

t. Add the Matchcodes transform to generate a match code for the CUSTOMER_NAME field.

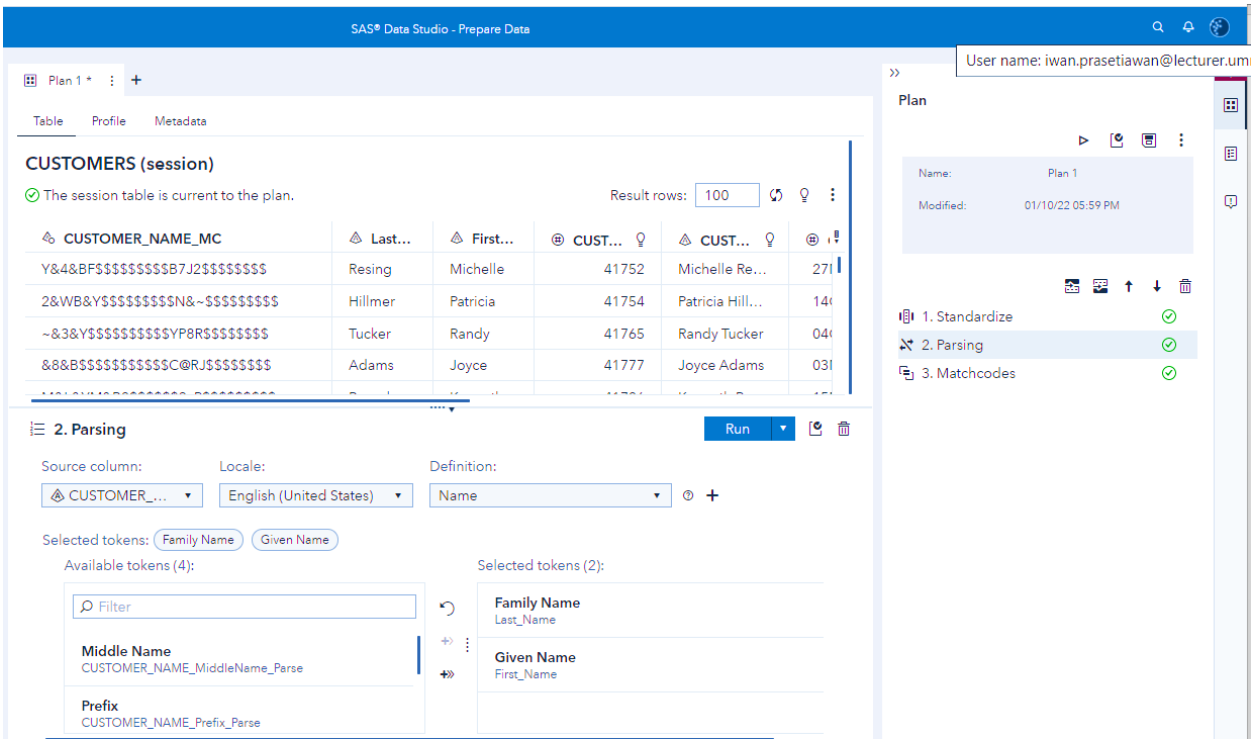
Note: There is no model evaluating the data for which columns are good candidates for match codes. Therefore, no match code suggestions are generated. We will add the Matchcodes transform from the Transforms menu.

- Click the Transforms icon.
- Expand the Data Quality grouping.
- Double-click Matchcodes.
- Select CUSTOMER_NAME as the source column.
- Change the name of the new column to CUSTOMER_NAME_MC.
- Verify that the locale is English (United States).
- Select the Name definition from the definition list.
- Verify that Sensitivity is set to 85.
- Click Run and review the results.

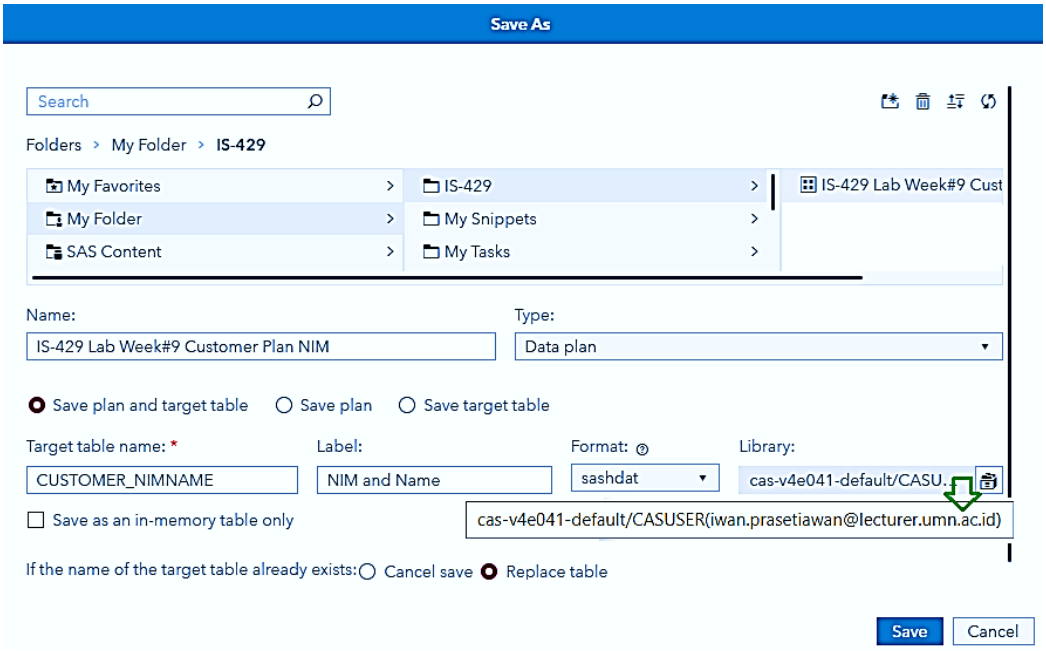
CUSTOMER_NAME_MC	CUSTOMER_NAME
3&3&2&3@$$$$$$$N&~$$$$$$$$	Patricia Kukahiko
4~&M&~$$$$$$$$L&WM$$$$$$$	Stevette Welf
2&W&B$$$$$$$$$C7WW$$$$$$$	Jill Halim
8&M&BM&Y~$$$$$$MW&3$$$$$$$$	Blake Davenport
8&M$$$$$$$$$$$JY74$$$$$$$	Christine Deep
4&W&B4$$$$$$$$M@M$$$$$$$$	Robert Sullins
Y&&M&&W8$$$$$$$J&~2$$$$$$$	Katherine Raybould



- Match codes can be useful when you want to join tables with no primary key or when clustering fields.
- For better accuracy when performing fuzzy joins using match codes, always use more than one or two fields, such as name, address, city, and state.
- For example, you could join two tables where the CUSTOMER_NAME, CUSTOMER_ADDRESS, and CUSTOMER_STATE match codes are the same.
- Click (Options) in the upper right corner of the Plan pane. Click save as.



- Save your Plan as "IS-429 Lab Week#9A Plan" and target table name "CUSTOMER_NIMNAME" :

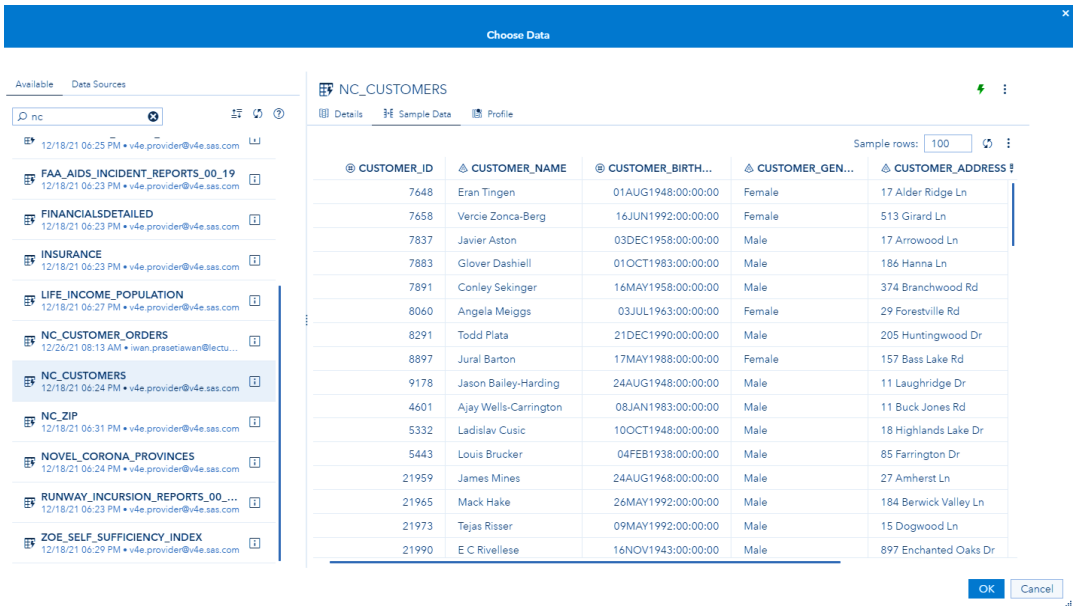


- **Screenshotted your plan and your table output as Output Figure A.**
- **Click Generate Code**, the system will automatically generate the code of your plan in SAS format and place it directly into the "download" folder in your workspace with a .txt file format.
- **Rename your plan code generated as "IS-429 Lab Week#9A Plan NIM Name.sas"**
- Select Close to close the plan.

2. Combining Data to Create an Analytical Base Table (ABT)

- The following practice illustrates the use of the Join transform to combine two tables to create an analytical base table (ABT).
- The NC_CUSTOMERS and RETAIL_ORDERS tables are joined based on the CUSTOMER_ID primary key column.

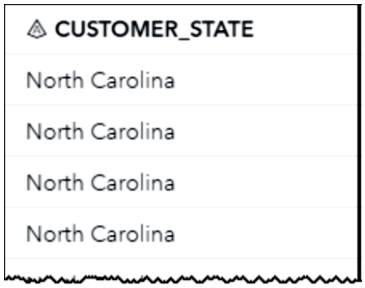
g. Re-run the command in points 1.a to 1.e above.



The NC_CUSTOMERS table contains information about customers

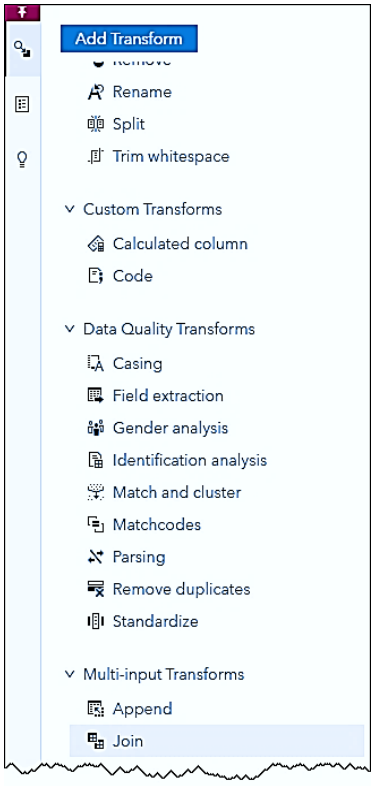
.

h. Scroll to the right and review the CUSTOMER_STATE column.

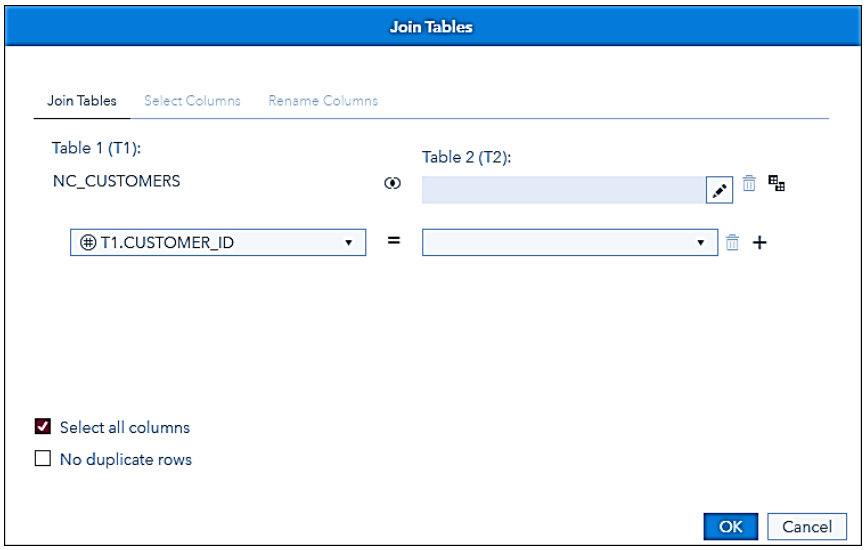


Every customer in the NC_CUSTOMERS table lives in North Carolina.

i. Double-click Join under Multi-input Transforms

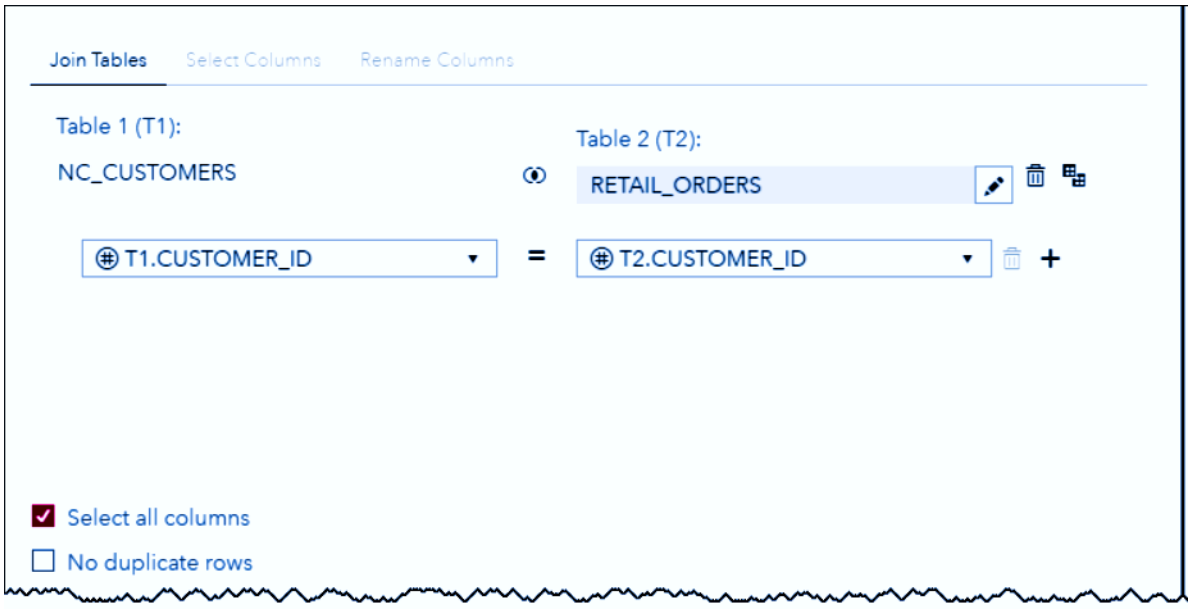


- The Join Tables window appears:

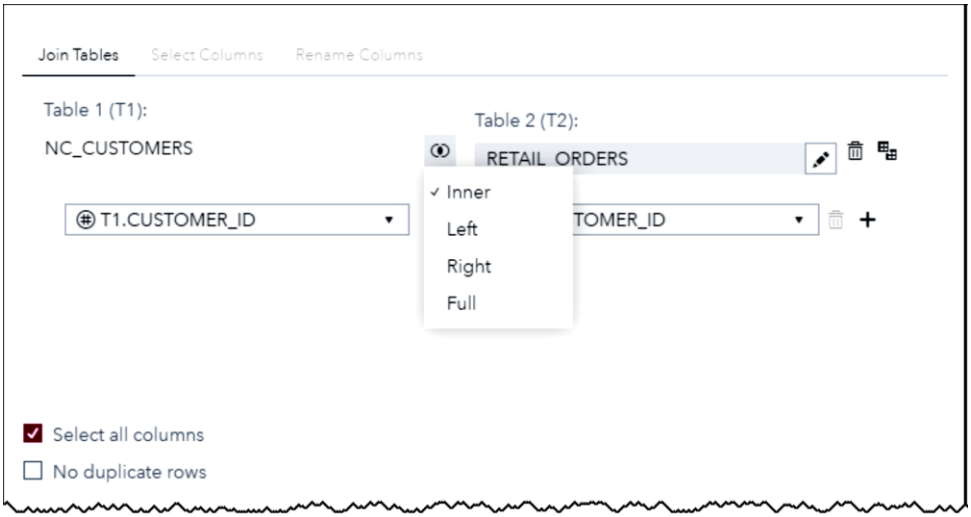


j. Configure the join
   - Click (Edit) to select Table 2.
   - Select RETAIL_ORDERS on the Available tab.

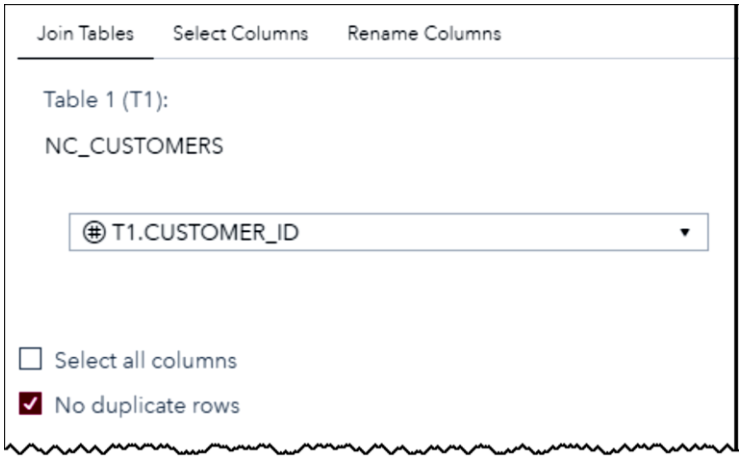
   - Notice that the RETAIL_ORDERS table has a CUSTOMER_ID column. This will be used in the join.
   - Click OK.



k. The join automatically filled in the Table 2 column based on intuitive processing.
l. Click (Click to select the join type) and verify that Inner is selected as the join type.

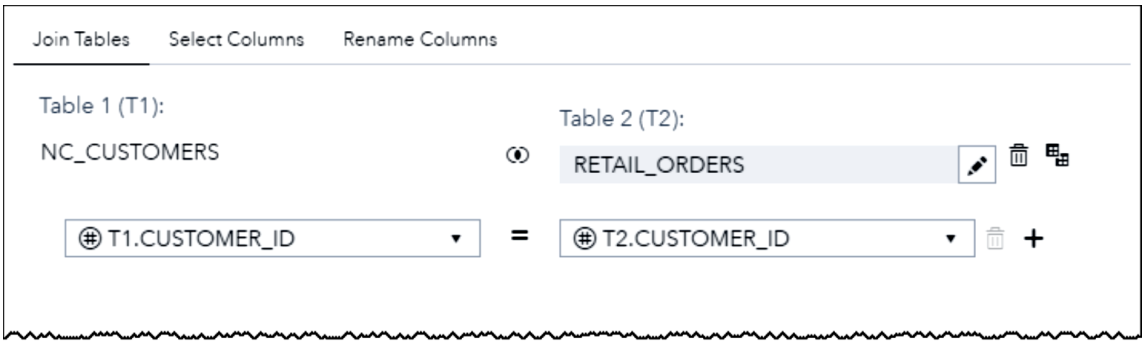


m. Clear the Select all columns check box to remove the default.
n. Select the No duplicate rows check box.



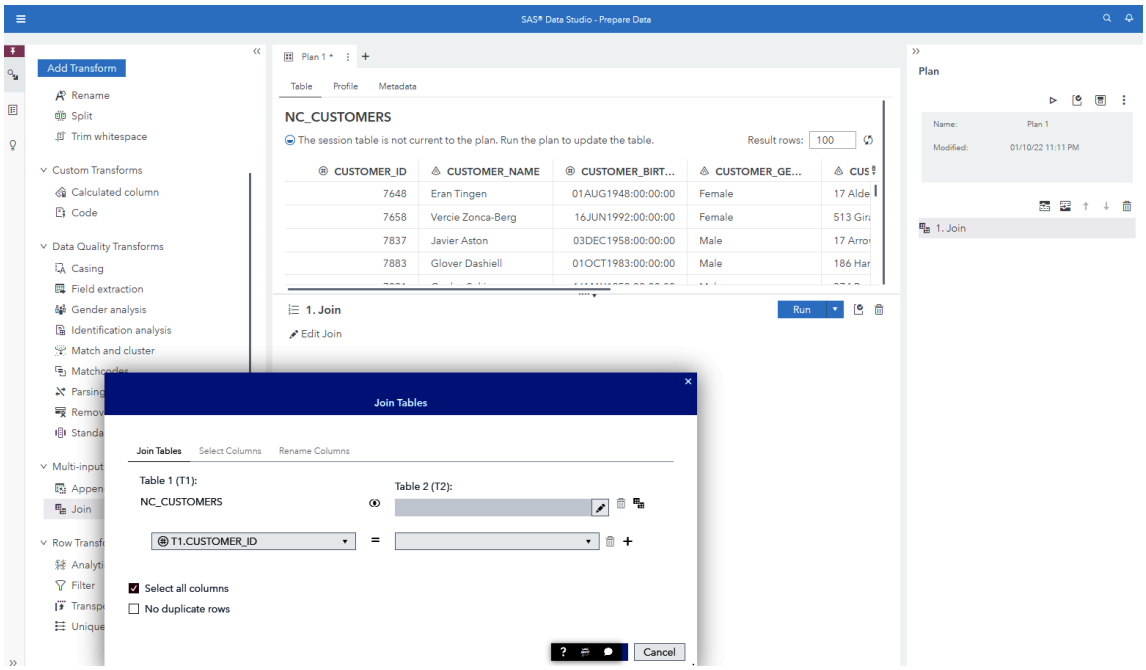
o. Click the Select Columns tab. Double-click T2.CUSTOMER_ID_1 to move the column to the Available items list box.

p. Because this is the column that we are joining on, we do not need two of the same column in the target table.

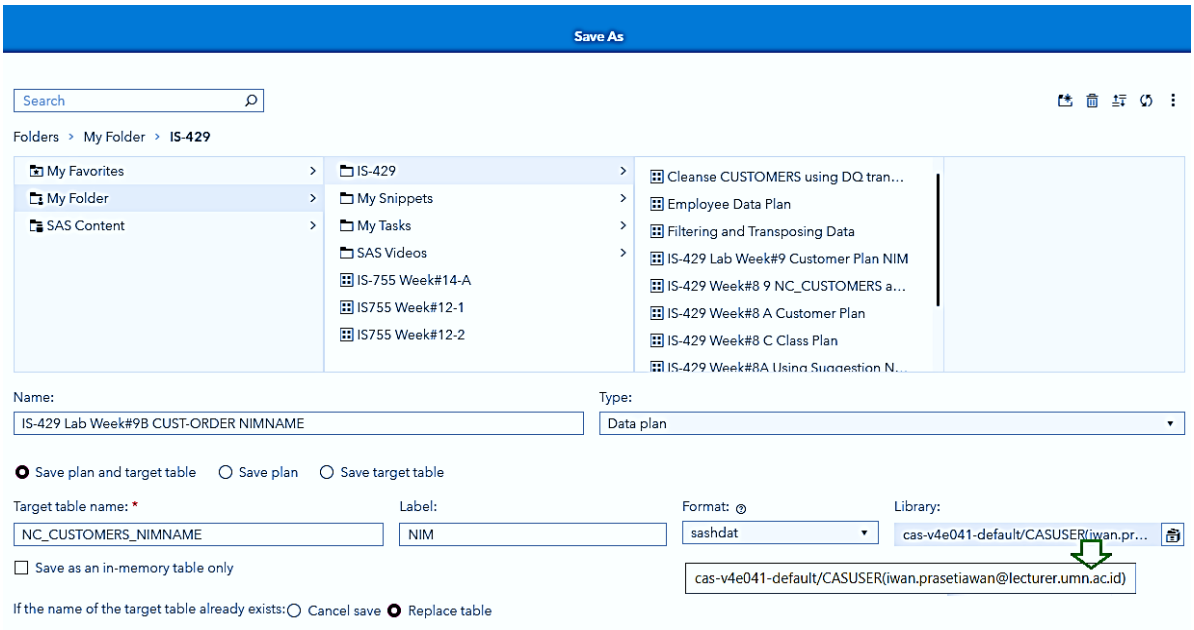


q. Click OK. Click Run and verify the results.

r. Click (Options) in the upper right corner of the Plan pane. Click Save As:



s. **Save your plan as "IS-429 Lab Week9B Cust-Order NIMNAME" at your CAS library by name of your new table as NC_CUSTOMER_NIMNAME.**



- **Screenshotted your plan and your table output as Output Figure B.**
- **Click Generate Code**, the system will automatically generate the code of your plan in SAS format and place it directly into the "download" folder in your workspace with a .txt file format.
- **Rename your plan code generated as "IS-429 Lab Week#9B Plan NIM Name.sas"**
- Select Close to close the plan.

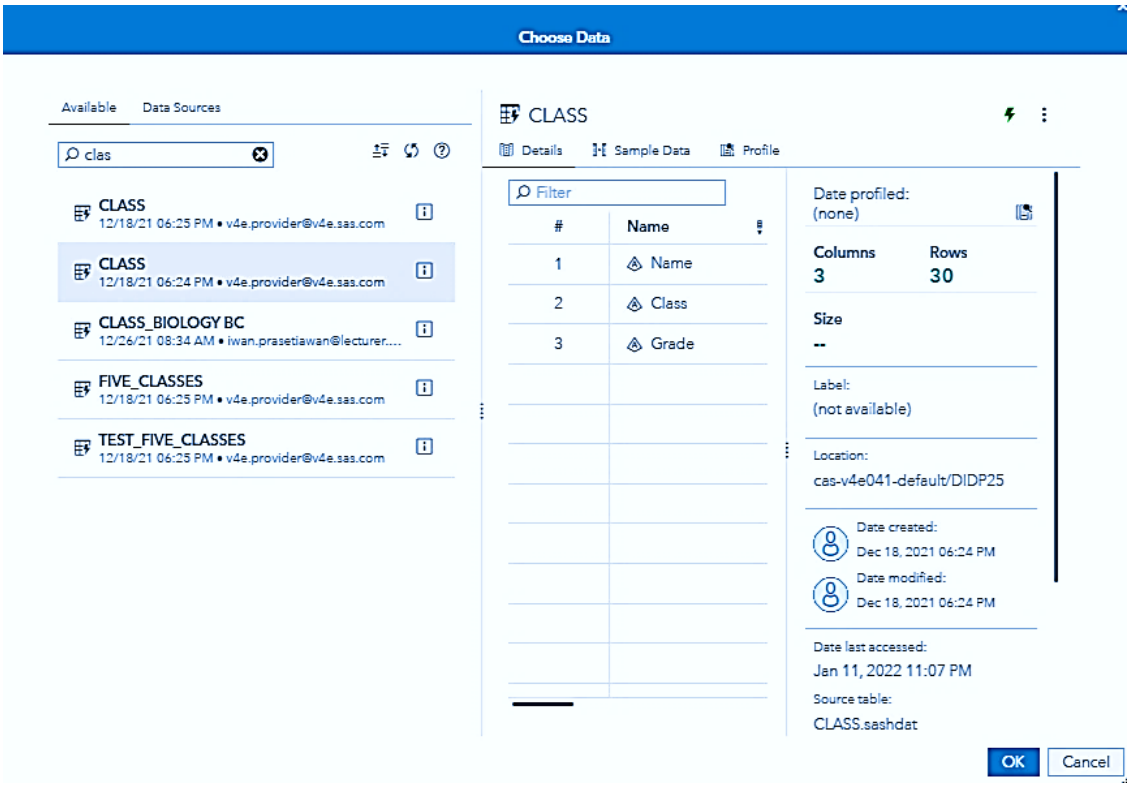
t. Return to SAS Drive.

3. Filtering and Transposing Data for Better Insight

- This practice illustrates how to filter rows and transpose data into a desired output format to support data modeling, reporting, and analytics.

a. Re-run the command in points 1.a to 1.e above.

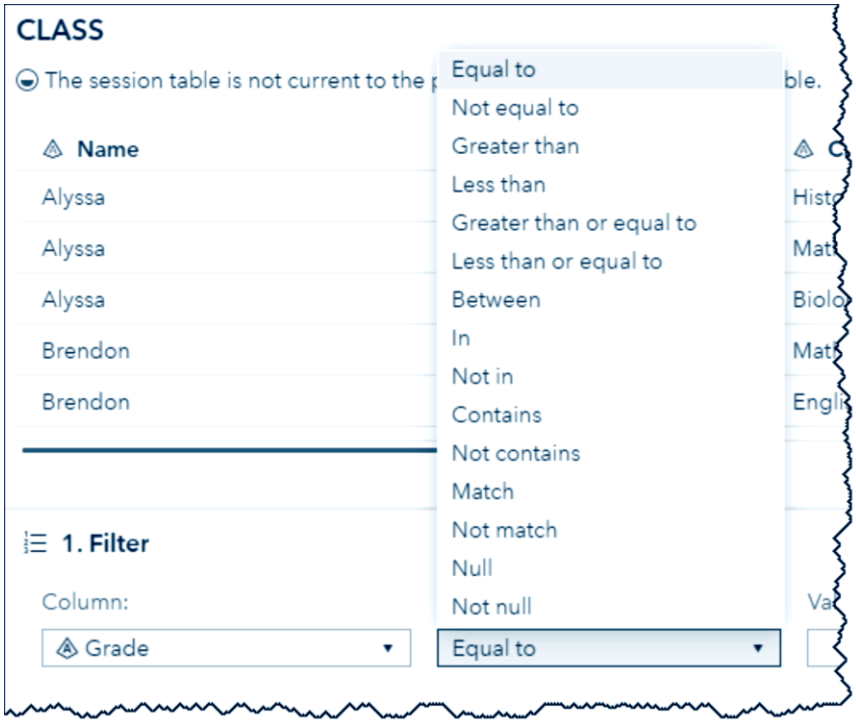
b. Click the Available tab. Click the CLASS table and select OK.



- The CLASS table has 30 rows. Each row contains a student's name, the class that the student is in, and the grade the student received in that class.
- First, we will explore the table by filtering the data. Then we will transpose the table to get one row per class.

c. Double-click Filter under Row Transforms. Configure the Filter transform.

- Select Grade in the Column list
- Click the down arrow to view the list of operators

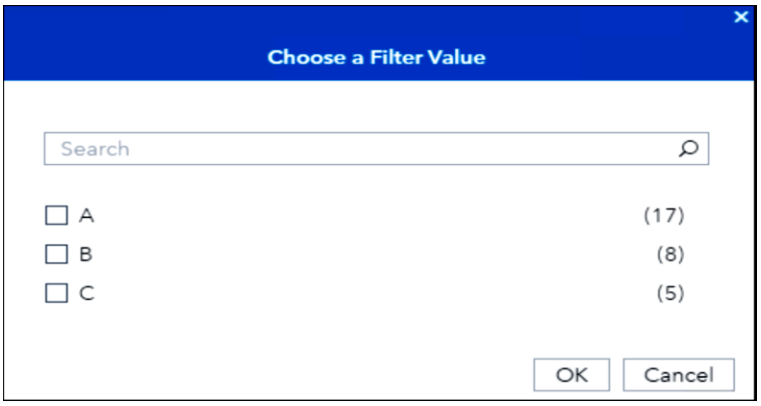


- There are many operators in place for the Filter transform.

d. Click In.

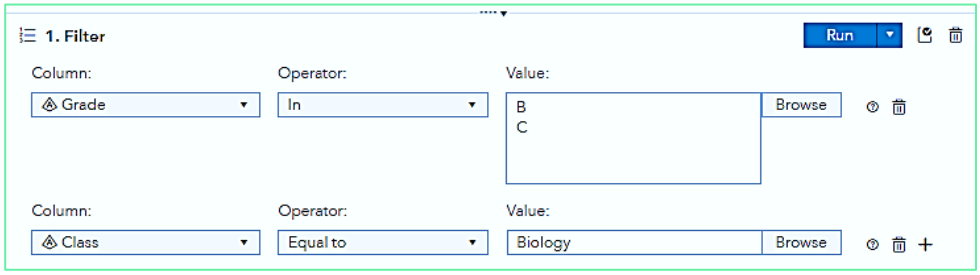
- The IN operator enables us to show all values that match one value in a list.
- The Filter transform adds an implicit AND between every unique filter condition that we configure. If we tried to filter for Grade = B and Grade = C, we would get no rows because a student cannot have a grade equal to both B and C.
- Therefore, the IN operator enables us to do the operation that we want, which is to return students where the grade is equal to B or C.

e. Click Browse.
Browse retrieves a list of the unique values in the selected column and their frequencies



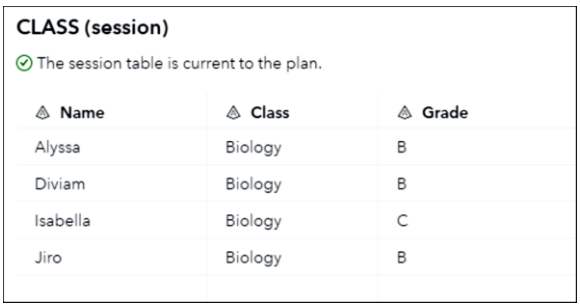
f. Click the check boxes next to B and C.
We want to check which students got grades other than an A this semester and in which classes.

g. Click OK. Click "+" (Add) to add another filter.
The school has been concerned with grades in the sciences. We will filter the B and C grade rows to see which of these grades were in the Biology class:
1) Select Class in the Column list.
2) Verify that Equal to is selected in the Operator list.
3) Click Browse.
4) Click Biology



- The filter now looks for rows where Grade is B or C and Class is Biology.
- There is an implicit AND between conditions in the Filter transform

h. Click Run and verify the results.



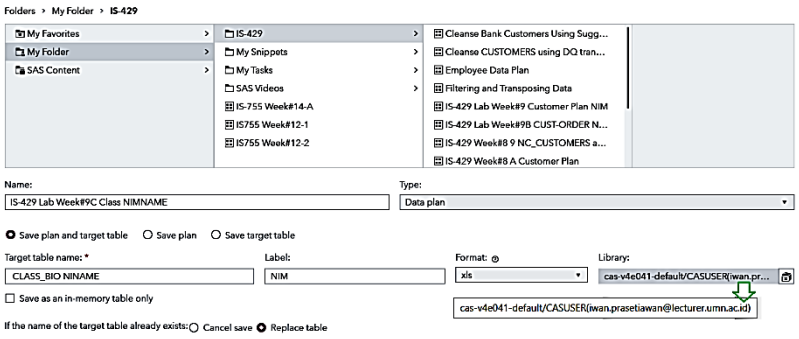
CLASS (session)

The session table is current to the plan.

Name	Class	Grade
Alyssa	Biology	B
Diviam	Biology	B
Isabella	Biology	C
Jiro	Biology	B

Only four students received a grade of B or C in Biology.

i. **Save your plan as "IS-429 Lab Week#9C Class NIMNAME" at your CAS library by name of your new table as CLASS_BIO NIMNAME with spreadsheet format.**



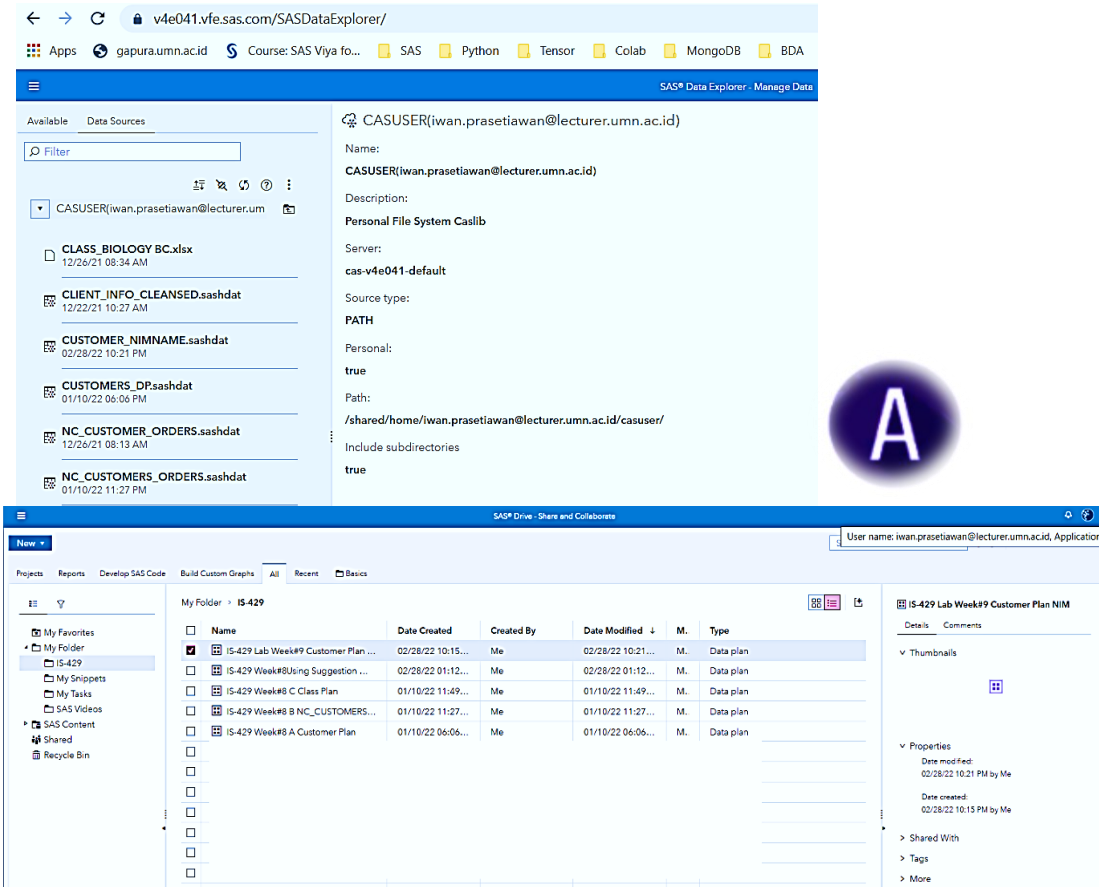
- **Screenshotted your plan and your table output as Output Figure C.**
- **Click Generate Code, rename your plan code generated as "IS-429 Lab Week#9C Plan NIM Name.sas"**

➲ Finally, today's practicum is over, collect all the screenshots you produce into a word format file as "IS-429 Lab Week#9 Screenshots NIM.doc/docx"

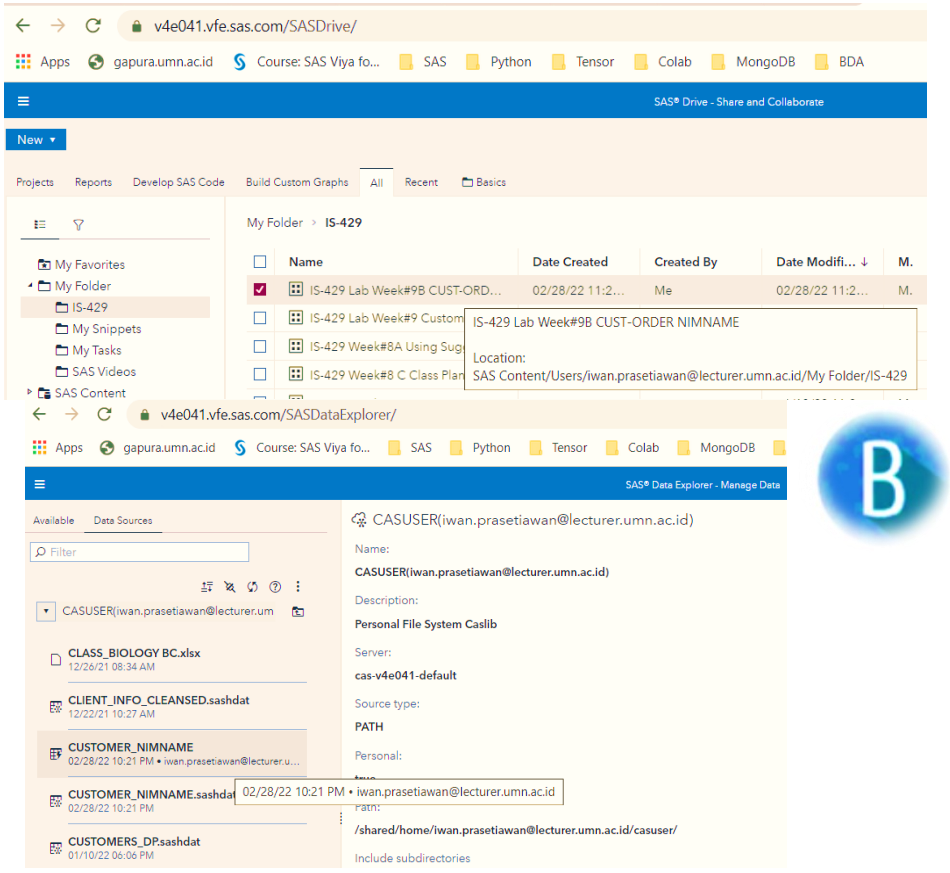
➲ **Zipped your word file and SAS files and submit immediately today to e-Learning IS-429 Practicum Week#9 with the naming format IS-429 BDA Week#9 NIM yourName.zip.**

RESULTS/ OUTPUT

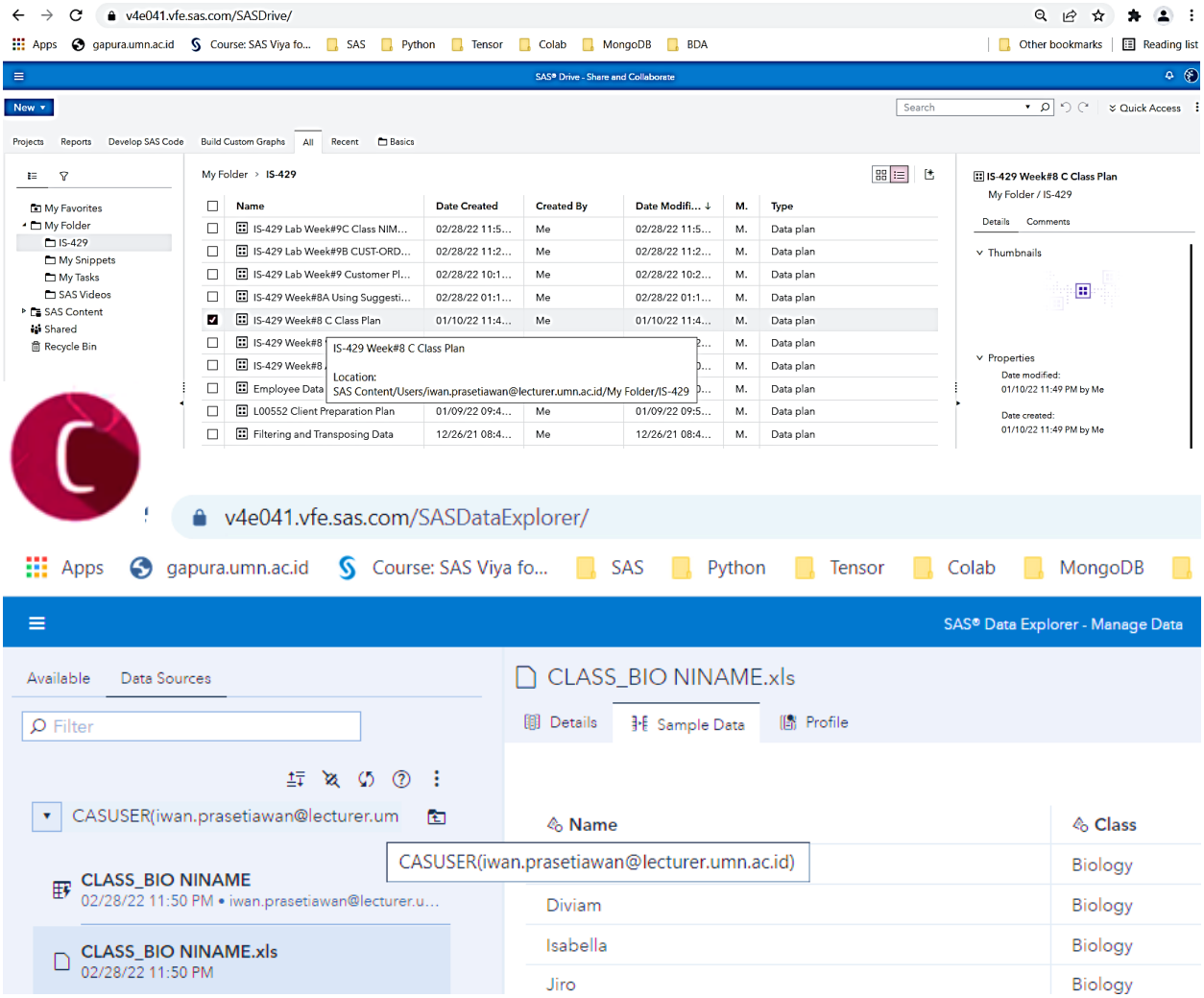
A. Removing Data Inconsistencies Using **Data Quality Transforms**: **Customer table**



B. Combining Data to Create an **Analytical Base Table (ABT)**



C. Filtering and Transposing Data for Better Insight



REFERENCE

1. Anna Yarbrough. 2020. Introduction to Data Curation for SAS® Data Scientists Course Notes. SAS Institute Inc. Cary, NC, USA.
2. SAS Institute Inc. 2020. SAS® Viya® Programming: Getting Started. SAS Institute Inc. Cary, NC, USA.
3. Johnny Starling, Erin Winters, and Anna Yarbrough. 2020.Self-Service Data Preparation in SAS® Viya® Course Notes. . SAS Institute Inc. Cary, NC, USA
4. SAS® Support | Documentation
5. Other additional references are excerpts from various Online Learning/websites.