C533 - Data Integration with Alteryx

Import Data

1. You will use CUSTOMER_DATA table in the file Customer Data For Integration.sql.
2. Bring data into Alteryx from Microsoft SQL server using the Input tool.

Data Investigation

1. Examine the data types of the fields imported. Note that you can view both data and metadata in Alteryx.
2. Add "Field Summary" tool under Data Investigation to learn more about the data fields. (Use the built-in example in Alteryx to learn the tool). Explore the data using this tool. With the tool, it would be easy to answer the following questions. (These questions are only for exploration purpose. Real questions are in Part B of this assignment on Canvas)
    a. How many "999" are there in the "Birthdate"?
    b. How many unique values are there in the region field?
    c. What is the problem with the values in Region column?
    d. How many records have "W" as the region value?
    e. What is the median for the field AmtSpent
    f. What is the average (mean) for Purchases (number of purchases each customer made)?
    g. What are the duplicate SSNs?

Notes Before You Start Cleaning

In the pre-class demo, we filtered out invalid data. In other words, we discarded records if data in some fields were incorrect. In real world, we usually don't discard data rows with some invalid fields. We mark those invalid fields instead as practiced in the live session. In this homework, you are asked to mark invalid fields as Null values in Alteryx. When exported to Excel format, those fields will appear to be blank.

All values of 999 are considered placeholders of missing values in any column of this data set (even though 999 might be a legitimate value in some columns).

Clean Column BirthDate

1. Use the DateTime tool under Parse to convert the BirthDate to the correct data type. You need to specify the format that matches the incoming string field. Choose the appropriate one. 999 cannot be converted and you will get a number of conversion errors. It is OK to ignore the errors.
2. Add a Sort tool (Under Preparation) to sort the data by BirthDate. You will see some future dates.
3. Assuming that all future dates were recorded incorrectly by inadvertently adding 100 years. (If the dates were entered in Excel using 2-digit year format, Excel would automatically add 100 years for some dates, for example). Clean the data using the Formula tool (Under Preparation). Write a formula to subtract 100 years from a date if it is a future date (any date later than today or the time Alteryx workflow runs is considered future).
    a. Note: two functions can be used here: DateTimeNow() and DateTimeAdd(). Check Alteryx documentation on how to use the functions.

Calculate Actual Age

1. Check what problems the field Age has (Revealed by the Field Summary tool)?
2. Use a Formula tool to calculate the age of customers as of now. Name the new column "ActualAge". If the BirthDate value is Null, the actual age value should also be Null.
    a. Note: DateTimeDiff() function can be used.

Clean Column Region

1. The column should have only 5 values: Midwest, NorthEast, South, West and Alaska.
2. You can find all unique values in this column by using a tool under Preparation. Then you can write a formula on the column Region to get the result.

C533 – Data Integration Assignment @Jie Li and Michael Frutiger All rights reserved

3. Or, you can use the RegionCode column with values 1, 2, 3, 4, 5 to set values in the Region column to Midwest, NorthEast, South, West, and Alaska respectively.
    a. You can use if or switch function for this purpose. Check how to use switch function here: https://help.alteryx.com/20213/designer/conditional-functions

Clean Column Income

1. Turn 999 in the Income column into Null value.
    a. Note: use Null() to get the null value

Clean Columns Purchases and AmtSpent

1. Use a Formula tool to set Purchases value to be NULL (use null() in Alteryx) for Purchases of value 999
2. Use a Formula tool to set AmtSpent value to be NULL (use null() in Alteryx) for blank AmtSpent value if it is not already Null
3. If Purchases is 0 but AmtSpent is not, the data is not valid. In this case, you should set AmtSpent to be 0 as well.
4. If AmtSpent is 0 but Purchases is not, the data is not valid. In this case, you should set Purchases to be 0 as well.
5. Note: if statement needs be used extensively for this cleaning step.

Clean Column CustomerType

1. There should be 2 customer types: 1 for retail customer and 2 for corporate customer.
2. The values in CustomerTypes are a combination of 1, 2, R, C and some other values. 1 and 2 are the valid values. Set all R to 1 and C to 2. Other values (-1, 3, 18, etc.) are invalid and should be set to Null.

Clean Column SSN

1. In addition to some duplicates in SSN, there are other problems. Some values have less than 9 digits and some have more than 9. Some values have no dash (-), some have only one dash. Some values use underscore (_) instead of dash.
2. Use the formula tool to clean the values. If a value has 9 digits, it is considered a valid SSN. Invalid values should be set to Null.
    a. Note: it is recommended that you first remove all the dash (-) and underscore (_) , then check whether there are 9 digits in the SSN and go from there.
3. Set all the duplicate SSN to be Null assuming we cannot tell which one is correct.

Add Column Phone

1. Add phone numbers stored in Customer Phone Numbers.xlsx to the data.
2. First, import the data using the Input tool. Take a look at the data.
3. Use the Join tool to connect the two datasets. You might encounter an error when join the customer data. It is likely that the two customer fields in the two sources are stored with different data types. To fix this, use the Select tool to convert the data types to make them consistent before you do the join.

Double check your work

1. Use the Select tool to be sure that your output data has 12 columns. Of the original 11 columns, the Age column should not be included. The actualAge column should be included in the output. The remaining 10 columns should also stay. The Phone column is added from the join.
2. Set the data types properly. This is easy to do in the Select tool

Export the Cleaned Data

1. Export the clean data using an Output Data tool
2. The exported data should be saved in an Excel file named Customer Data Cleaned.xlsx.

Submit your work

1. Submit the Alteryx workflow and exported Excel file using “Data Integration and Preparation with Alteryx Part A” assignment
2. Answer the questions in using “Data Integration and Preparation with Alteryx Part B” assignment

C533 – Data Integration Assignment @Jie Li and Michael Frutiger All rights reserved