Analytics mindset

ETL

This user guide covers a series of cases regarding foundational (not comprehensive) procedures for extracting, transforming and loading (ETL) relevant data for analytics purposes. This case material is designed for any course that teaches about the ETL process. The cases are designed to be flexible so that they can be implemented in class, used as homework or as a student project or part of an exam.

It is recommended that students complete the ETL overview case first, followed by each case as listed in numerical order, as complexity increases as the case number increases. Depending on the background proficiency level of your students, you can easily select an individual case further in the series without requiring the students to complete the previous cases.

While other EYARC analytics mindset case materials walk students through the entire analytics mindset approach, these cases focus exclusively on the ETL process. The students are able to practice developing their ETL skills with rich, real-world data. As a reminder, an analytics mindset is the ability to:

- Ask the right questions
- Extract, transform and load relevant data (i.e., the ETL process)
- Apply appropriate data analytics techniques
- Interpret and share results with stakeholders

Below, each case and the associated files are presented in order. Additionally, for several cases, materials have been developed using Excel and Alteryx. The files for each software option are labeled as such.

ETL Overview case

This case provides an overview of fundamental considerations in the extract, transform and load (ETL) process, including extracting data, unique identifiers, joining (merging) data, common messy data problems and creating a repeatable ETL process. Students will read about these considerations and then answer a series of short questions to confirm their understanding of the reading. These questions can also be used as an assessment tool.

Case study and solutions

Analytics_mindset_case_studies_ETL_Overview.docx

Analytics_mindset_case_studies_ETL_Overview.pdf

Analytics_mindset_case_study_solutions_ETL_Overview.docx

Analytics_mindset_case_study_solutions_ETL_Overview.pdf

SCORE No. 06753-191US_21

Case 1: Identifying data problems

This case provides students with a small data set, in comma-separated file format. They are asked to review the data and identify problems in it that could cause problems when loading the data into a tool for analysis. Student are then asked to transform (clean) the data and address the identified issues.

Case study and solutions

Excel:

Analytics_mindset_case_studies_ETL_Case1_Excel.docx

Analytics_mindset_case_studies_ETL_Case1_Excel.pdf

Analytics_mindset_case_study_solutions_ETL_Case1_Excel.docx

Analytics_mindset_case_study_solutions_ETL_Case1_Excel.pdf

Alteryx:

Analytics_mindset_case_studies_ETL_Case1_Alteryx.docx

Analytics_mindset_case_studies_ETL_Case1_Alteryx.pdf

Analytics_mindset_case_study_solutions_ETL_Case1_Alteryx.docx

Analytics_mindset_case_study_solutions_ETL_Case1_Alteryx.pdf

Data set

Excel and Alteryx:

Initial data set:

Analytics_mindset_case_studies_ETL_Case1.csv

Modified data set:

Note that students need to modify the original data set to remove a comma in a field that does not have text qualifiers (or add the text qualifiers). We've provided this data set that was modified in Notepad.

Analytics_mindset_case_studies_ETL_Case1_modified.csv

Analytics workbooks

Excel:

Analytics_mindset_case_study_solutions_ETL_Case1.xlsx

Alteryx:

Analytics_mindset_case_study_solutions_ETL_Case1.yxzp

Analytics_mindset_case_study_solutions_ETL_Case1_AlteryxOutput.xlsx

Videos

Alteryx:

Note that a video has been prepared to address each identified issue and for how to produce the output file separately.

This user guide for the EYARC Access has removed video links for external distribution. See the user guide on the EYARC site for links.

SCORE No. 06753-191US_21

Case 2: Text extractions and unique identifiers

This case asks the students to perform text extraction from employee data (597 rows) using an employee code data field with multiple fixed-width components that relate to location, employee number, plant number and pay period frequency.

- For Excel, students are directed to perform text extractions using Text to Columns and are also provided suggested formulas by data field, including LEFT, MID, FIND and RIGHT.
- For Alteryx, students are not provided detailed direction for text extraction but can perform these procedures much more quickly than Excel using just one tool, RegEx.

Upon completing the text extraction, students are asked to create a unique identifier using the location and Plant ID.

Case study and solutions

Excel:

Analytics_mindset_case_studies_Case2_Excel.docx

Analytics_mindset_case_studies_Case2_Excel.pdf

Analytics_mindset_case_study_solutions_Case2_Excel.docx

Analytics_mindset_case_study_solutions_Case2_Excel.pdf

Alteryx:

Analytics_mindset_case_studies_Case2_Alteryx.docx

Analytics_mindset_case_studies_Case2_Alteryx.pdf

Analytics_mindset_case_study_solutions_Case2_Alteryx.docx

Analytics_mindset_case_study_solutions_Case2_Alteryx.pdf

Data set

Excel:

Analytics_mindset_case_studies_ETL_Case2_Excel.xlsx

Alteryx:

Analytics_mindset_case_studies_ETL_Case2_Alteryx.xlsx

Analytics workbooks

Excel:

Analytics_mindset_case_study_solutions_ETL_Case2.xlsx

Alteryx:

Analytics_mindset_case_study_solutions_ETL_Case2.yxzp

Analytics_mindset_case_study_solutions_ETL_Case2_AlteryxOutput.xlsx

Videos

Alteryx:

This user guide for the EYARC Access has removed video links for external distribution. See the user guide on the EYARC site for links.

Case 3: Advanced text extractions and unique identifiers

This case asks the students to perform text extraction from employee data (597 rows) using an employee code data field with variable components that relate to location, employee number, plant number and pay period frequency.

- For Excel, text extractions are performed using formulas, including MID, MIN, FIND, LEFT, RIGHT, IF, ISNUMBER, VALUE and CONCATENATE (note that formulas are not suggested for each data field in this case specifically but rather just comprehensively at the end of the case instructions as an overall hint for students to consider which should foster their research and critical thinking). As the formulas are more complex, the solution set includes how to videos to provide for instruction if desired.
- For Alteryx, students are not provided detailed direction for text extraction but can perform these procedures much more quickly than Excel using just one tool, RegEx.

Upon completing the text extraction, students are asked to create a unique identifier using the location and Plant ID.

Case study and solutions

Excel:

Analytics_mindset_case_studies_Case3_Excel.docx

Analytics_mindset_case_studies_Case3_Excel.pdf

Analytics_mindset_case_study_solutions_Case3_Excel.docx

Analytics_mindset_case_study_solutions_Case3_Excel.pdf

Alteryx:

Analytics_mindset_case_studies_Case3_Alteryx.docx

Analytics_mindset_case_studies_Case3_Alteryx.pdf

Analytics_mindset_case_study_solutions_Case3_Alteryx.docx

Analytics_mindset_case_study_solutions_Case3_Alteryx.pdf

Data set

Excel:

Analytics_mindset_case_studies_ETL_Case3_Excel.xlsx

Alteryx:

Analytics_mindset_case_studies_ETL_Case3_Alteryx.xlsx

Analytics workbooks

Excel:

Analytics_mindset_case_study_solutions_ETL_Case3.xlsx

Alteryx:

Analytics_mindset_case_study_solutions_ETL_Case3.yxzp

Analytics_mindset_case_study_solutions_ETL_Case3_AlteryxOutput.xlsx

Videos

Excel:

This video discusses how each formula works utilizing the Evaluate formula under the Formula tab in the Excel ribbon.

This user guide for the EYARC Access has removed video links for external distribution. See the user guide on the EYARC site for links.

Alteryx:

This user guide for the EYARC Access has removed video links for external distribution. See the user guide on the EYARC site for links.

Case 4: Joining data

Students are provided with general ledger journal entry data (789 rows) on one tab in the data file along with other important accounting relevant data sets on other tabs that relate to the journal entry data, including the chart of accounts, source information, business unit information and preparer information. The students are asked to join all of this data.

- For Excel, the students are provided with suggested formulas, including INDEX/MATCH. Students are told that their final deliverable will be used for future ETL work so the process needs to be easily repeatable. This consideration prevents students from creating any new data fields. One join requires a unique identifier and therefore students are required to create the unique identifier within a formula, and ultimately, a formula with an array, to complete their work.
- For Alteryx, students can perform a series of joins to easily complete this work. By default, the work in Alteryx is easily repeatable.

It is important to note that students should identify that there are records that will not match across all data sets involving the PreparerName field.

Case study and solutions

Excel:

Analytics_mindset_case_studies_Case4_Excel.docx

Analytics_mindset_case_studies_Case4_Excel.pdf

Analytics_mindset_case_study_solutions_Case4_Excel.docx

Analytics_mindset_case_study_solutions_Case4_Excel.pdf

SCORE No. 06753-191US_21

		Alteryx: **Analytics_mindset_case_studies_Case4_Alteryx.docx** **Analytics_mindset_case_studies_Case4_Alteryx.pdf** **Analytics_mindset_case_study_solutions_Case4_Alteryx.docx** **Analytics_mindset_case_study_solutions_Case4_Alteryx.pdf**
	Data set	Excel: **Analytics_mindset_case_studies_ETL_Case4_Excel.xlsx** Alteryx: **Analytics_mindset_case_studies_ETL_Case4_Alteryx.xlsx**
	Analytics workbooks	Excel: **Analytics_mindset_case_study_solutions_ETL_Case4.xlsx** Alteryx: **Analytics_mindset_case_study_solutions_ETL_Case4.yxzp** **Analytics_mindset_case_study_solutions_ETL_Case4_AlteryxOutput.xlsx**
	Videos	Excel: The video provides an overview of INDEX/MATCH and its benefits compared with VLOOKUP, as well as discusses how each formula works utilizing the Evaluate formula under the Formula tab in the Excel ribbon. This user guide for the EYARC Access has removed video links for external distribution. See the user guide on the EYARC site for links. Alteryx: This user guide for the EYARC Access has removed video links for external distribution. See the user guide on the EYARC site for links.

Case 5: Using VBA to transform data

This is a two-part case that asks students to perform simple and more complex macros by coding in the Visual Basic Applications (VBA) language within Excel. In the first part of the case, students are provided with donor information for a non-profit organization (180 donors) and are asked to program a macro to format this data, as well as conditionally remove data, based on certain criteria. In the second part of the case, students are provided with two different sets of tax depreciation data for a company's assets that has been generated in a disaggregated format that is not ideal for management's review. Students are asked to program a macro to modify the data presentation where the first data set is easier to modify as it is presented in a more uniform format while the second data set is more difficult as it is presented in a variable format.

Case study and solutions

Excel:

Analytics_mindset_case_studies_Case5_Excel.docx

Analytics_mindset_case_studies_Case5_Excel.pdf

Analytics_mindset_case_study_solutions_Case5_Excel.docx

Analytics_mindset_case_study_solutions_Case5_Excel.pdf

Data sets

Analytics_mindset_case_studies_ETL_Case5Part1.xlsx

Analytics_mindset_case_studies_ETL_Case5Part2.xlsx

Analytics workbooks

Excel (macro-enabled workbooks):

Analytics_mindset_case_study_solutions_ETL_Case5Part1.xlsm

Analytics_mindset_case_study_solutions_ETL_Case5Part2.xlsm

SCORE No. 06753-191US_21

Data

The data for this case is real-world data, with the exception of Case 5 Part 1, that includes a variety of data elements, data file types and file sizes appropriate for each case. You are welcome to utilize the data sets separately as desired.

Analytics tools

We have designed these cases to be flexible with respect to the tools that can be used to perform the analyses. We provide several of the cases and the solutions using both Excel and Alteryx. A potential approach to these cases would be to do a case first in Excel and then perform it in Alteryx so students can understand the functionality and power of Alteryx more in depth. The cases can also be adapted to be completed using other technology tools. Files are named with a reference to the technology tool, as appropriate.

Alteryx

This software is one of the leading self-service data analytics programs currently available. It is a powerful tool for the ETL process and data analytics, and is especially known for its spatial analytics tools. “Alteryx allows a single user to access various data sources, clean and prepare data, perform a variety of analyses and then deploy the results for consumption and to operationalize the insights discovered. It boasts visual workflows and an intuitive drag-and-drop interface that can eliminate the need to write code.”[1]

The workflows allow the user to visually understand the steps that the data goes through for transformation or analysis. This workflow also provides a clear audit trail and creates a repeatable process. When using an Alteryx workflow, the initial data sources remains intact. Alteryx performs the transformation and generates a new output file, which helps maintain the integrity of the data, allows mistakes to be easily fixed and allows the same workflow to be used on multiple datasets (with the same fields and properties). Alteryx can process very large data sets very quickly. As the workflow is running, the user can see where any errors may be occurring and can monitor progress as percentage of completion statistics are calculated throughout. As soon as the data is processed, a report is generated that shows if there are any errors in any fields (e.g., trailing white spaces) that may create problems for data analysis, as well as some basic visual analytics that show descriptive information about the processed data.

Access: A free trial of Alteryx and academic licensing is available at: https://www.alteryx.com/why-alteryx/alteryx-for-good. Alteryx will provide individual licenses for students and faculty as well as lab licenses for use in a computer lab or classroom. However, it is important to note that Alteryx only works on PCs at this time.

Community and training:

- Alteryx is intuitive and easy to learn. It boasts an active user community (https://www.alteryx.com/community) that openly shares workflow examples to help users identify potential solutions. Like other analytics tools, there are always many ways to organize data to achieve

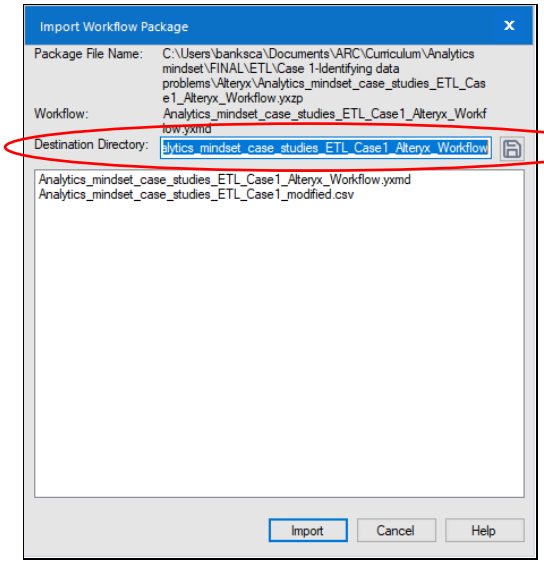
[1] Alteryx, Inc. 2017 10-k: https://www.sec.gov/Archives/edgar/data/1689923/000119312518073878/d530988d10k.htm

SCORE No. 06753-191US_21

the desired solutions. By reviewing approaches of other data scientists, you and your students can develop more advanced skills and perspectives.

- Also within the Alteryx community, there are easy to follow tutorials as part of the Alteryx Academy (https://community.alteryx.com/t5/Alteryx-Academy/ct-p/alteryx-academy). You can learn most of the necessary ETL functions in a few hours using these tutorials.
- Further, there is a set of tutorials designed with the proficient Excel user in mind that walks you through basic Excel functions using Alteryx instead. As users become more advanced, they can participate in the Alteryx "weekly challenges" (https://community.alteryx.com/t5/Weekly-Challenge/bd-p/weeklychallenge), which provide difficult data analytics problems for the user community to solve.

Note that for Alteryx workflow solution files, these have been provided as a packaged workflow (.yxzp file type [Options > Export Workflow >]) that is a single, zipped file that includes the Alteryx workflow and all of its dependencies (e.g., input file and naming convention for output file). Students are asked to submit their completed workflow as their deliverable using this file type as well, including their name in the file, to allow the workflow to be easily executed. When this packaged workflow is opened, you will be asked to provide a file directory where you want the files to be saved before you are able to run the workflow.



Annotation

One of the nice features of this tool is that it allows the user to document each step of the workflow through the use of annotation. The case requirements do not include requirements for students to annotate each step of their workflow as written although annotations have been provided in the solution set. You can add a requirement with a description of the level of detail that you would like to see as desired.

How to videos

"How to" videos have been prepared to accompany cases as appropriate, with links provided in this user guide to allow access to these videos directly. These links have NOT been included in the case solutions file. These videos can help instructors learn the ETL steps themselves and can be provided directly to students to supplement in-class instruction and facilitate their completion of the case requirements.