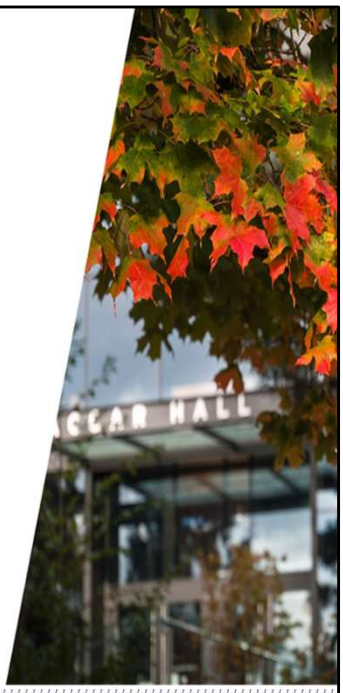
FOSTER Accounting

INTRO TO EXTRACTING, TRANSFORMING AND LOADING DATA

ACCTG 522

Class 2



Welcome!

Class Agenda

- Review and ETL Overview
- ETL Cases in the FSB Remote Labs
- Conclusion and look ahead

Review: ETL Overview

The Analytical Mindset

The Analytical Mindset is defined as the ability to work with data to apply appropriate analytics and interpret and share insight with stakeholders.

Why do we care about Data Quality?

Analytics quality (and insight) is driven by data quality.

Garbage in = Garbage Out !

Why do we care about Extract, Transform & Load (ETL)?

The Extract, Transform & Load or ETL Process is the first task required to complete the analytical process.

- In most cases, data required for the analytical process is stored in a database and must be **extracted** from that database and **loaded** into analytics software.

Why do we care about Extract, Transform & Load (ETL)?

In most cases, the data will also have to be transformed, or modified to apply analytics tools

- this can include **calculating** additional variables (like multiplying price and quantity to get sales),
- **cleaning** the data due to differences in formatting,
- **preparing** and **joining** datasets,
- and **transposing** or **pivoting** the data.

Identifying data issues

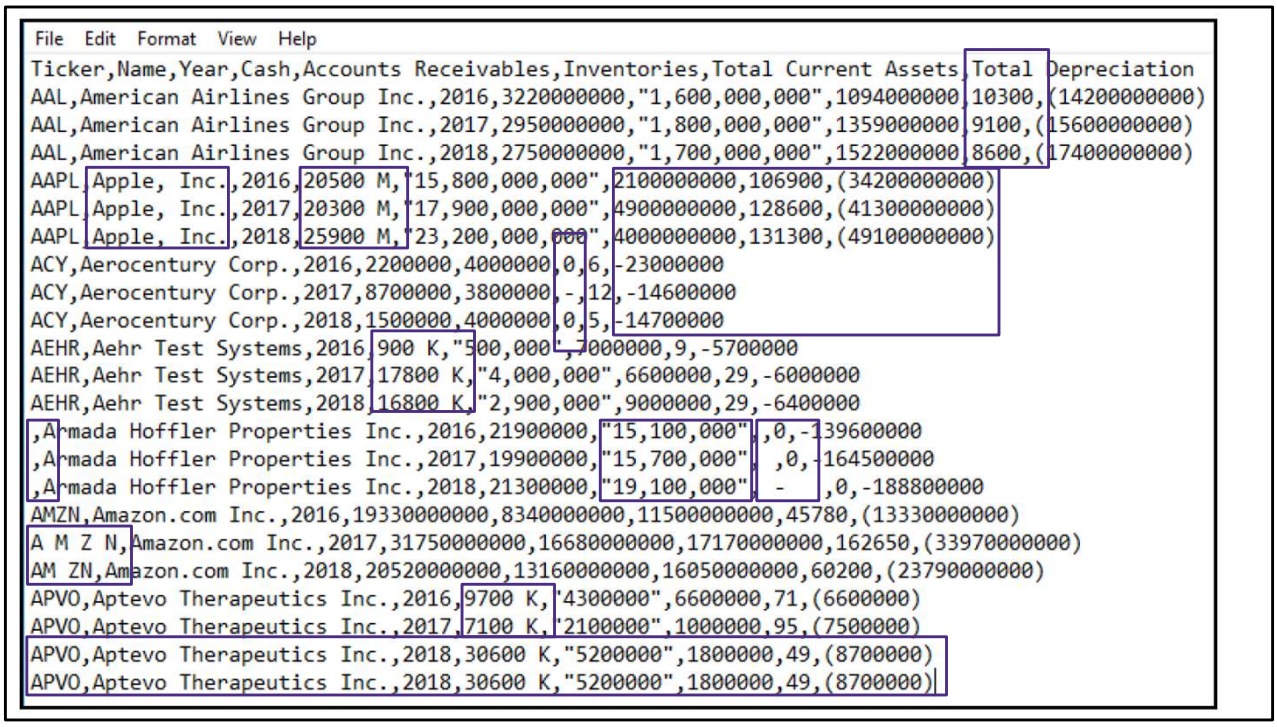
Revision.

There are 9 data quality issues with the file:

- Identify the issues and why
- Propose how they could be solved

File Edit Format View Help

```
Ticker,Name,Year,Cash,Accounts Receivables,Inventories,Total Current Assets,Total Depreciation
AAL,American Airlines Group Inc.,2016,3220000000,"1,600,000,000",1094000000,10300,(14200000000)
AAL,American Airlines Group Inc.,2017,2950000000,"1,800,000,000",1359000000,9100,(15600000000)
AAL,American Airlines Group Inc.,2018,2750000000,"1,700,000,000",1522000000,8600,(17400000000)
AAPL,Apple, Inc.,2016,20500 M,"15,800,000,000",2100000000,106900,(34200000000)
AAPL,Apple, Inc.,2017,20300 M,"17,900,000,000",4900000000,128600,(41300000000)
AAPL,Apple, Inc.,2018,25900 M,"23,200,000,000",4000000000,131300,(49100000000)
ACY,Aerocentury Corp.,2016,2200000,4000000,0,6,-23000000
ACY,Aerocentury Corp.,2017,8700000,3800000,-,12,-14600000
ACY,Aerocentury Corp.,2018,1500000,4000000,0,5,-14700000
AEHR,Aehr Test Systems,2016,900 K,"500,000",7000000,9,-5700000
AEHR,Aehr Test Systems,2017,17800 K,"4,000,000",6600000,29,-6000000
AEHR,Aehr Test Systems,2018,16800 K,"2,900,000",9000000,29,-6400000
,Armada Hoffler Properties Inc.,2016,21900000,"15,100,000",,0,-139600000
,Armada Hoffler Properties Inc.,2017,19900000,"15,700,000", ,0,-164500000
,Armada Hoffler Properties Inc.,2018,21300000,"19,100,000", -   ,0,-188800000
AMZN,Amazon.com Inc.,2016,19330000000,8340000000,11500000000,45780,(13330000000)
A M Z N,Amazon.com Inc.,2017,31750000000,16680000000,17170000000,162650,(33970000000)
AM ZN,Amazon.com Inc.,2018,20520000000,13160000000,16050000000,60200,(23790000000)
APVO,Aptevo Therapeutics Inc.,2016,9700 K,"4300000",6600000,71,(6600000)
APVO,Aptevo Therapeutics Inc.,2017,7100 K,"2100000",1000000,95,(7500000)
APVO,Aptevo Therapeutics Inc.,2018,30600 K,"5200000",1800000,49,(8700000)
APVO,Aptevo Therapeutics Inc.,2018,30600 K,"5200000",1800000,49,(8700000)
```



File Edit Format View Help

```
Ticker,Name,Year,Cash,Accounts Receivables,Inventories,Total Current Assets,Total Depreciation
AAL,American Airlines Group Inc.,2016,3220000000,"1,600,000,000",1094000000,10300,(14200000000)
AAL,American Airlines Group Inc.,2017,2950000000,"1,800,000,000",1359000000,9100,(15600000000)
AAL,American Airlines Group Inc.,2018,2750000000,"1,700,000,000",1522000000,8600,(17400000000)
AAPL,Apple, Inc.,2016,20500 M,"15,800,000,000",2100000000,106900,(34200000000)
AAPL,Apple, Inc.,2017,20300 M,"17,900,000,000",4900000000,128600,(41300000000)
AAPL,Apple, Inc.,2018,25900 M,"23,200,000,000",4000000000,131300,(49100000000)
ACY,Aerocentury Corp.,2016,2200000,4000000,0,6,-23000000
ACY,Aerocentury Corp.,2017,8700000,3800000,-,12,-14600000
ACY,Aerocentury Corp.,2018,1500000,4000000,0,5,-14700000
AEHR,Aehr Test Systems,2016,900 K,"500,000",7000000,9,-5700000
AEHR,Aehr Test Systems,2017,17800 K,"4,000,000",6600000,29,-6000000
AEHR,Aehr Test Systems,2018,16800 K,"2,900,000",9000000,29,-6400000
,Armada Hoffler Properties Inc.,2016,21900000,"15,100,000",,0,-139600000
,Armada Hoffler Properties Inc.,2017,19900000,"15,700,000",,0,-164500000
,Armada Hoffler Properties Inc.,2018,21300000,"19,100,000",-,0,-188800000
AMZN,Amazon.com Inc.,2016,19330000000,8340000000,11500000000,45780,(13330000000)
A M Z N,Amazon.com Inc.,2017,31750000000,16680000000,17170000000,162650,(33970000000)
AM ZN,Amazon.com Inc.,2018,20520000000,13160000000,16050000000,60200,(23790000000)
APVO,Aptevo Therapeutics Inc.,2016,9700 K,"4300000",6600000,71,(6600000)
APVO,Aptevo Therapeutics Inc.,2017,7100 K,"2100000",1000000,95,(7500000)
APVO,Aptevo Therapeutics Inc.,2018,30600 K,"5200000",1800000,49,(8700000)
APVO,Aptevo Therapeutics Inc.,2018,30600 K,"5200000",1800000,49,(8700000)
```

- Apple has a comma in its name, Apple, Inc., but the name does not have a text qualifier.
- Armada Hoffler Properties Inc. is missing the ticker symbol.
- The ticker for Amazon.com Inc. (AMZN) has spaces in some instances, but not in others.
- Aptevo Therapeutics Inc. has a duplicate row of information.
- For the Cash category:
- Apple, Inc.’s cash is listed in millions using a number, followed by a space and an M.
- Aehr Test Systems’ cash is listed in thousands

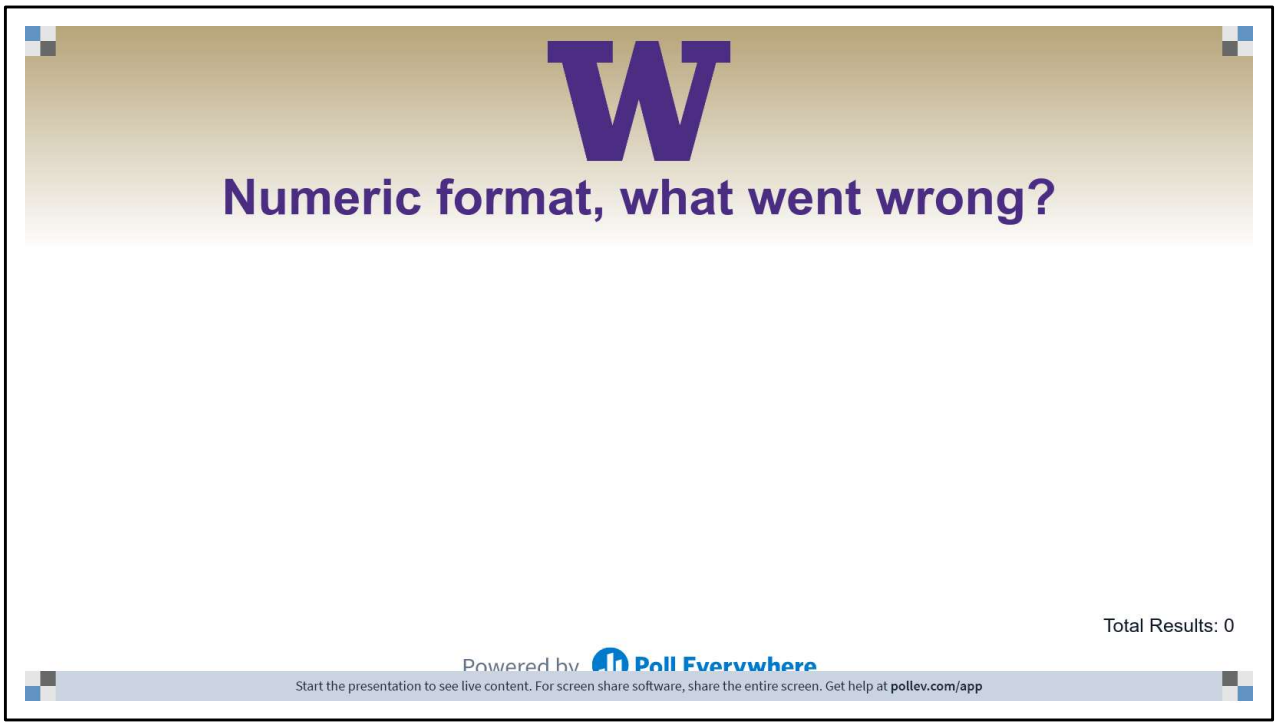
using a number, followed by a space and a K.
•Aptevo Therapeutics Inc.'s cash is listed in thousands using a number, followed by a space and a K.
•For Accounts Receivables, some amounts are shown as text and include commas in the number (commas within a text qualifier are appropriate, but this is important to note for purposes of loading the data).
•For Inventories, null values are displayed in various forms as text, including a zero, a blank, a hyphen (-), a hyphen with spaces or a space.
•For Total Current Assets, none of the amounts are listed in millions.
•For Depreciation, some numbers use parentheses for credits (or negatives) and others use a minus sign.

ETL Cases in the FSB Remote Labs

Revision – Data Quality

Open discussion (2 mins with classmate)

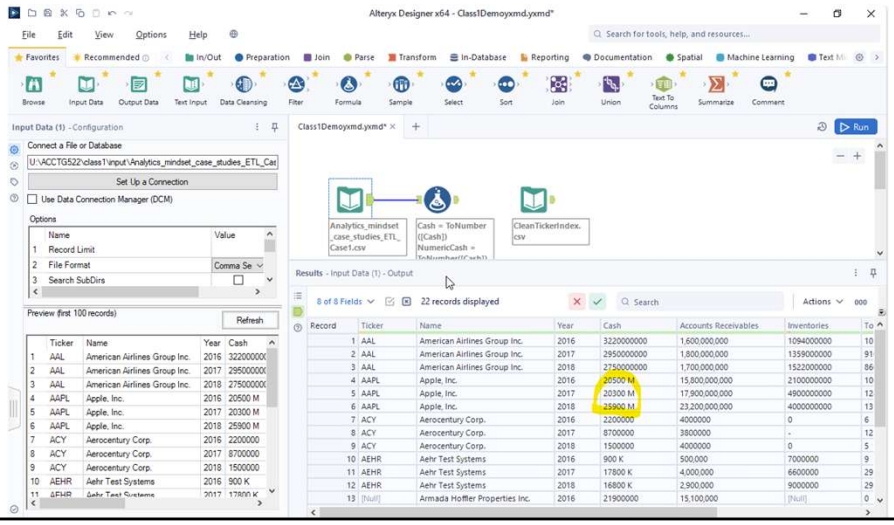
One lesson from the formatting exercise we undertook was that we need to be careful with converting string variables to numbers – what went wrong?



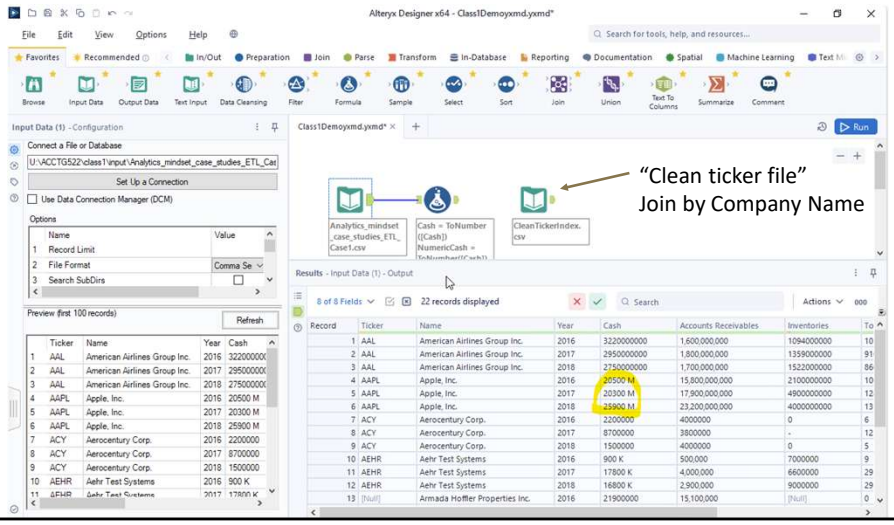
Poll Title: Do not modify the notes in this section to avoid tampering with the Poll Everywhere activity.
More info at polleverywhere.com/support

Numeric format, what went wrong?
https://www.polleverywhere.com/free_text_polls/r2W3O8apKA9FaBriBMSHa

Revision – Data Quality



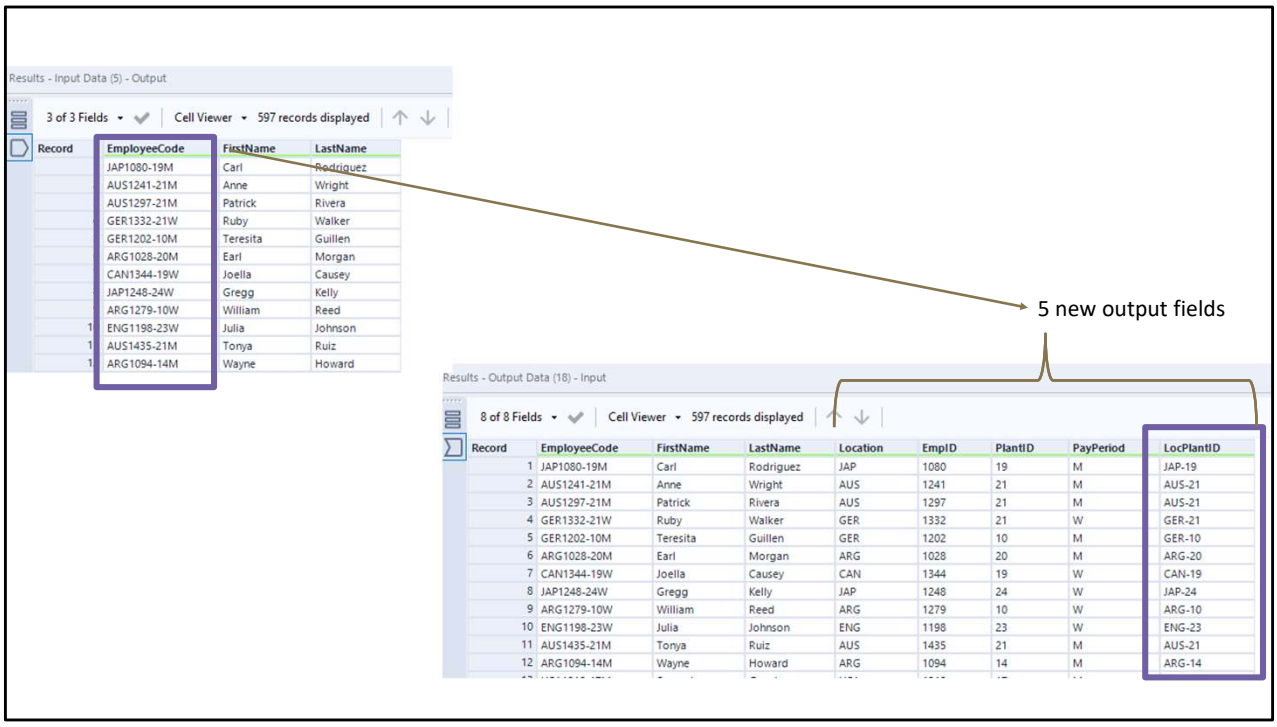
Alternative – Data Quality



ETL Cases

Our first ETL case “Text Extraction and Unique Identifiers” we will:

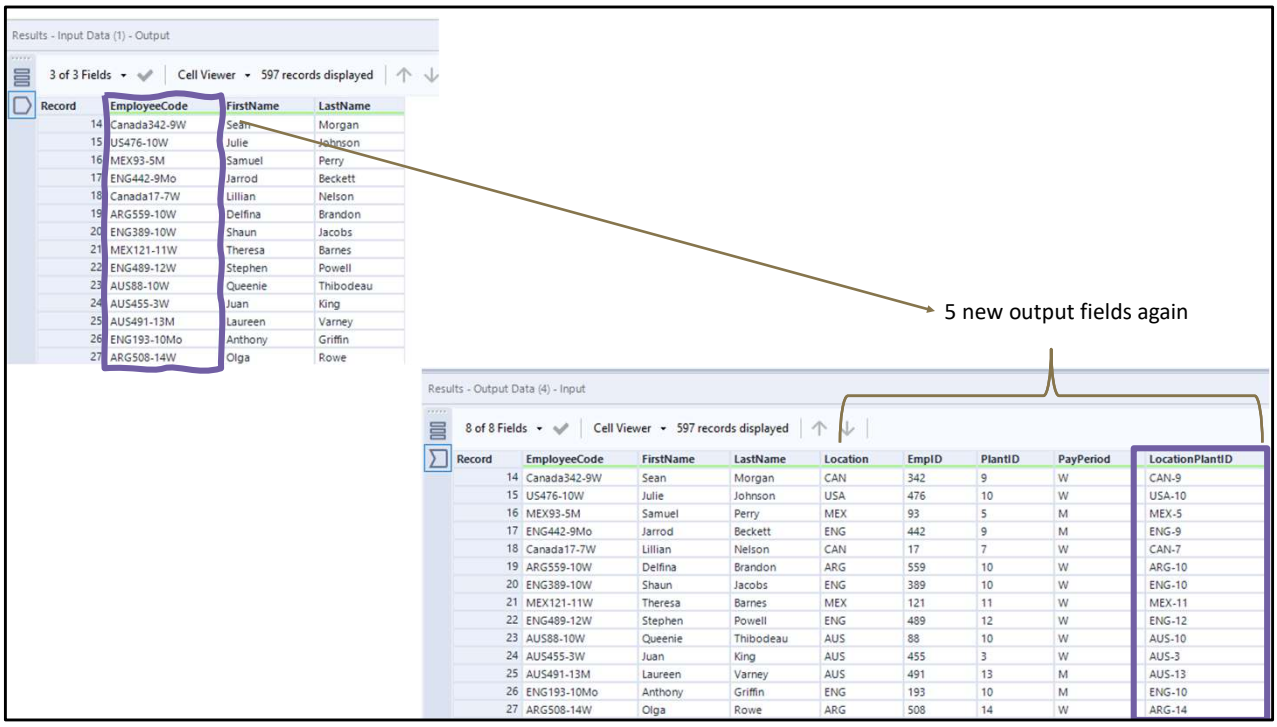
- Prepare data for a join by “Location-PlantID”
- The information is in “EmployeeCode”


Results - Input Data (5) - Output
3 of 3 Fields
Cell Viewer
597 records displayed
Record
EmployeeCode
FirstName
LastName
JAP1080-19M Carl Rodriguez
AUS1241-21M Anne Wright
AUS1297-21M Patrick Rivera
GER1332-21W Ruby Walker
GER1202-10M Teresita Guillen
ARG1028-20M Earl Morgan
CAN1344-19W Joella Causey
JAP1248-24W Gregg Kelly
ARG1279-10W William Reed
1 ENG1198-23W Julia Johnson
1 AUS1435-21M Tonya Ruiz
1 ARG1094-14M Wayne Howard
5 new output fields
Results - Output Data (18) - Input
8 of 8 Fields
Cell Viewer
597 records displayed
Record
EmployeeCode
FirstName
LastName
Location
EmpID
PlantID
PayPeriod
LocPlantID
1 JAP1080-19M Carl Rodriguez JAP 1080 19 M JAP-19
2 AUS1241-21M Anne Wright AUS 1241 21 M AUS-21
3 AUS1297-21M Patrick Rivera AUS 1297 21 M AUS-21
4 GER1332-21W Ruby Walker GER 1332 21 W GER-21
5 GER1202-10M Teresita Guillen GER 1202 10 M GER-10
6 ARG1028-20M Earl Morgan ARG 1028 20 M ARG-20
7 CAN1344-19W Joella Causey CAN 1344 19 W CAN-19
8 JAP1248-24W Gregg Kelly JAP 1248 24 W JAP-24
9 ARG1279-10W William Reed ARG 1279 10 W ARG-10
10 ENG1198-23W Julia Johnson ENG 1198 23 W ENG-23
11 AUS1435-21M Tonya Ruiz AUS 1435 21 M AUS-21
12 ARG1094-14M Wayne Howard ARG 1094 14 M ARG-14

ETL Cases

Our second ETL case “Advanced ETL Text Extraction and Unique Identifiers” we will:

- Prepare data for a join by “Location-PlantID”
- The information is in “EmployeeCode”
- But this time, clean the “EmployeeCode” as it is **messy**!


Results - Input Data (1) - Output
3 of 3 Fields
Cell Viewer
597 records displayed
Record
EmployeeCode
FirstName
LastName
14 Canada342-9W Sean Morgan
15 US476-10W Julie Johnson
16 MEX93-5M Samuel Perry
17 ENG442-9Mo Jarrod Beckett
18 Canada17-7W Lillian Nelson
19 ARG559-10W Delfina Brandon
20 ENG389-10W Shaun Jacobs
21 MEX121-11W Theresa Barnes
22 ENG489-12W Stephen Powell
23 AUS88-10W Queenie Thibodeau
24 AUS455-3W Juan King
25 AUS491-13M Laureen Varney
26 ENG193-10Mo Anthony Griffin
27 ARG508-14W Olga Rowe
5 new output fields again
Results - Output Data (4) - Input
8 of 8 Fields
Cell Viewer
597 records displayed
Record
EmployeeCode
FirstName
LastName
Location
EmpID
PlantID
PayPeriod
LocationPlantID
14 Canada342-9W Sean Morgan CAN 342 9 W CAN-9
15 US476-10W Julie Johnson USA 476 10 W USA-10
16 MEX93-5M Samuel Perry MEX 93 5 M MEX-5
17 ENG442-9Mo Jarrod Beckett ENG 442 9 M ENG-9
18 Canada17-7W Lillian Nelson CAN 17 7 W CAN-7
19 ARG559-10W Delfina Brandon ARG 559 10 W ARG-10
20 ENG389-10W Shaun Jacobs ENG 389 10 W ENG-10
21 MEX121-11W Theresa Barnes MEX 121 11 W MEX-11
22 ENG489-12W Stephen Powell ENG 489 12 W ENG-12
23 AUS88-10W Queenie Thibodeau AUS 88 10 W AUS-10
24 AUS455-3W Juan King AUS 455 3 W AUS-3
25 AUS491-13M Laureen Varney AUS 491 13 M AUS-13
26 ENG193-10Mo Anthony Griffin ENG 193 10 M ENG-10
27 ARG508-14W Olga Rowe ARG 508 14 W ARG-14

Why Alteryx?

Alteryx is cutting-edge software that is excellent for the ETL process.

Working with cutting-edge software allows you to:

- Become adaptable and resilient to new software.
- Impress recruiters with practical discussions of cutting-edge software.

To FSB remote labs

We will use **Alteryx on the FSB remote labs** to solve both cases.

I recommend **saving screenshots** of your work in Alteryx (I've added blank slides if you want to save them in this ppt document, ctrl+m adds more slides)

The Tools We will use:


Input Data Tool | Alteryx Help


Text To Columns Tool | Alteryx Help

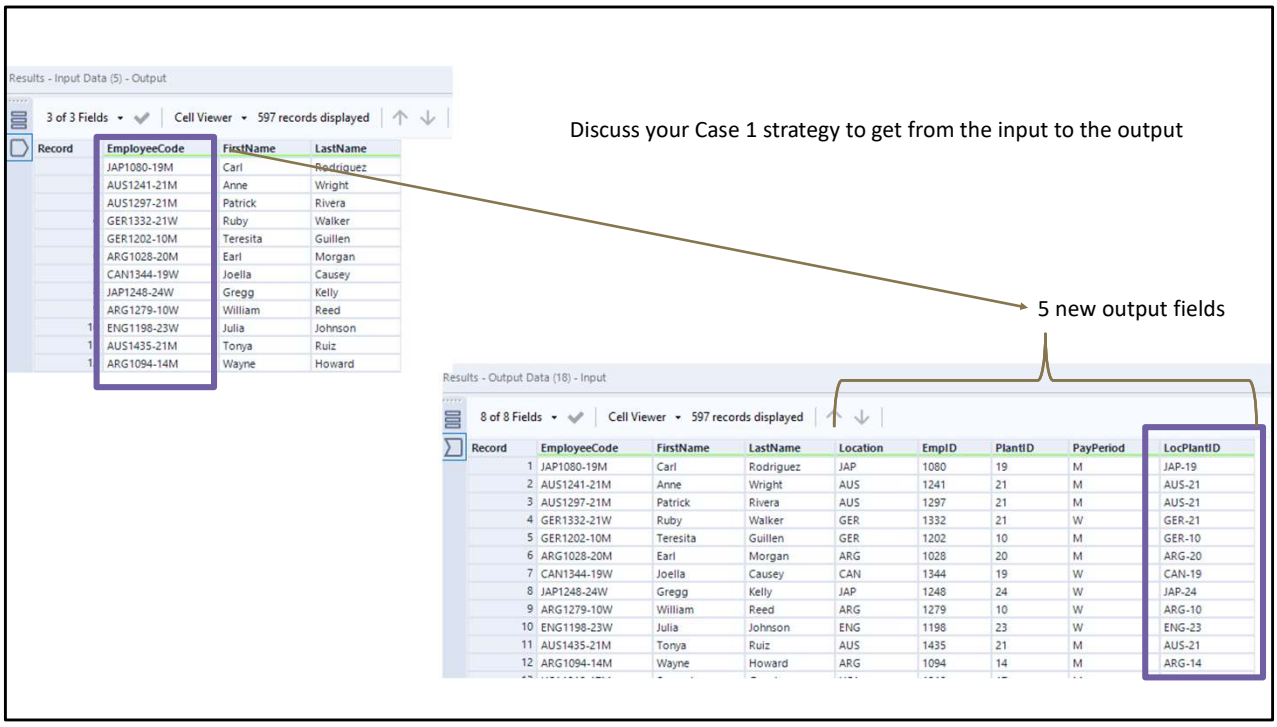

Formula Tool | Alteryx Help

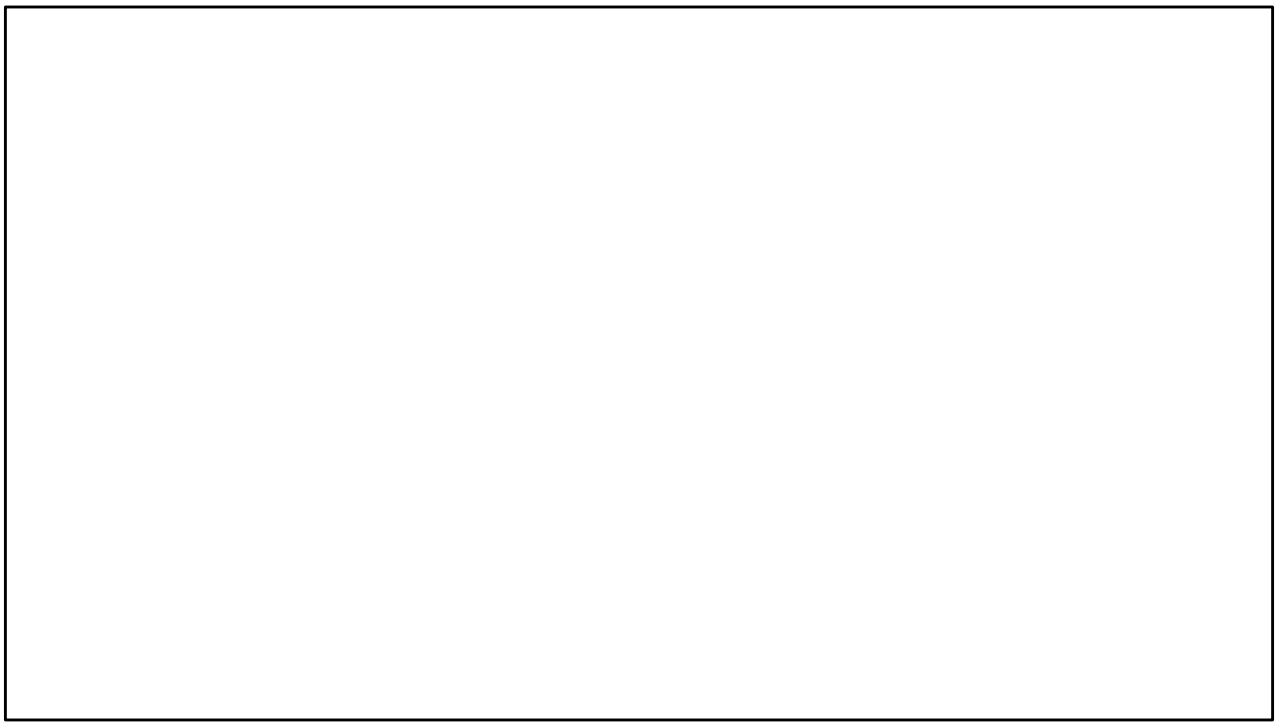

RegEx Tool | Alteryx Help

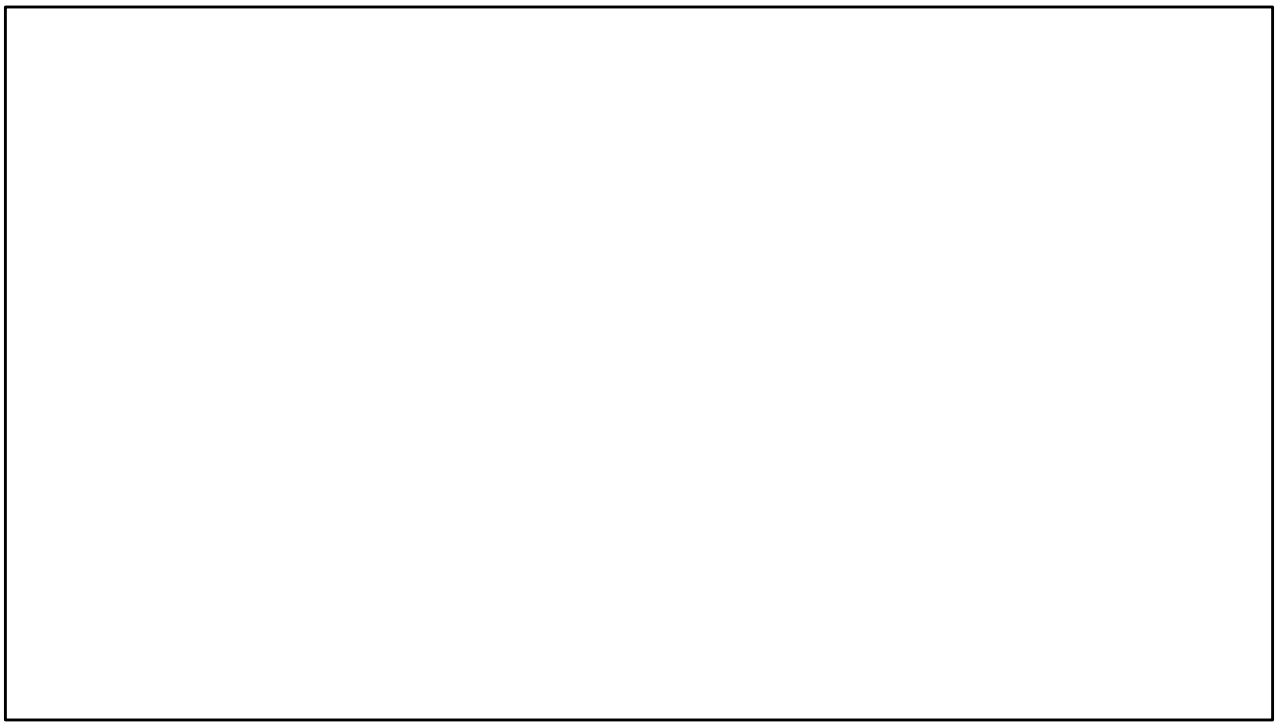

Output Data Tool | Alteryx Help

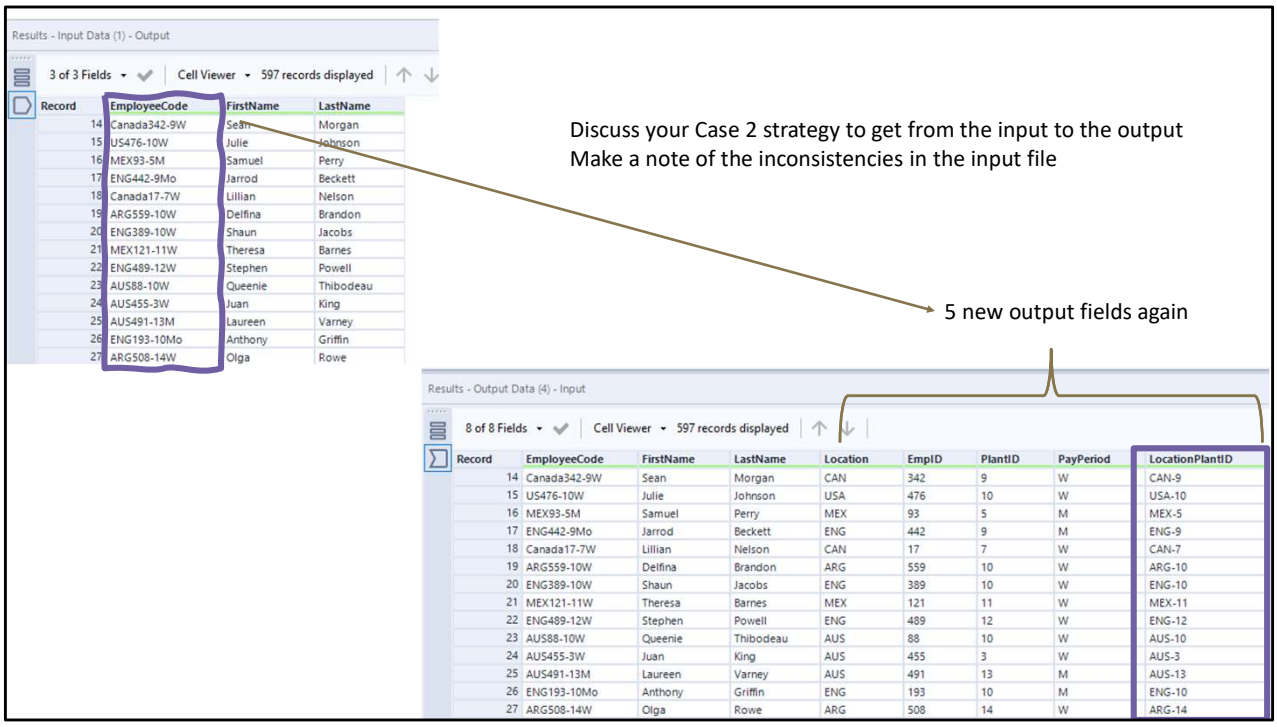
Lab Agenda

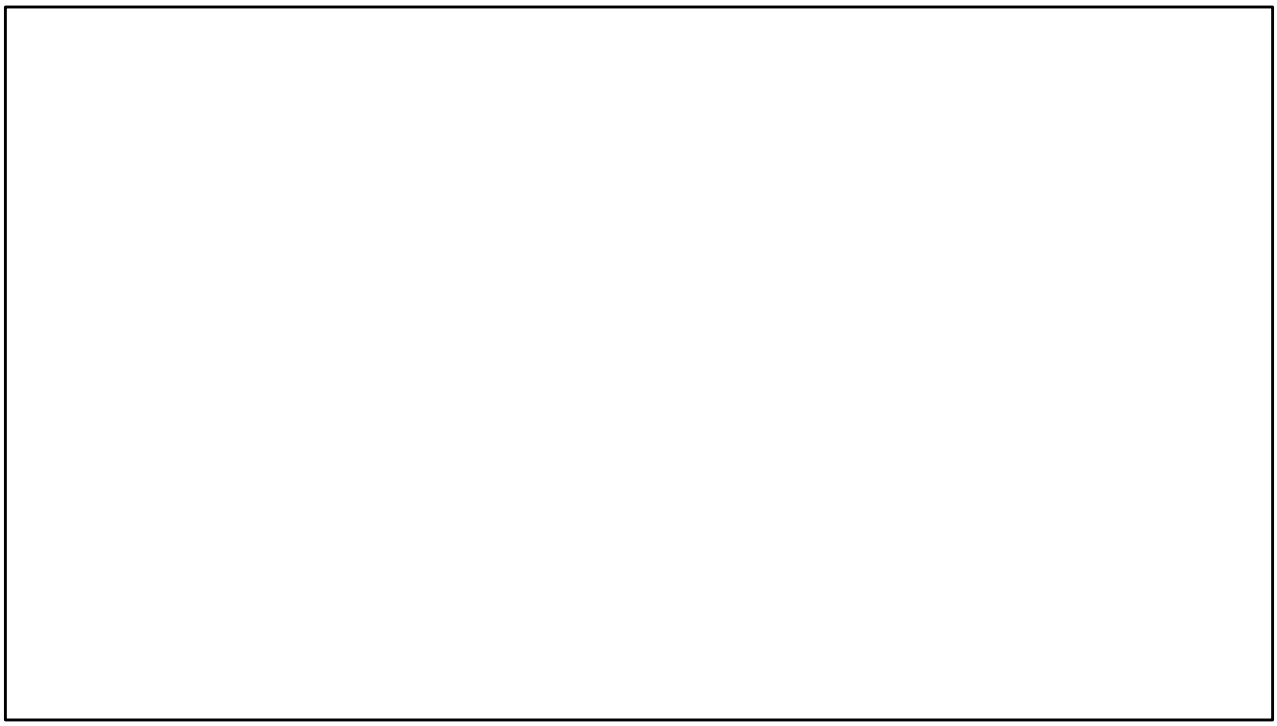
1. Solve ETL Case 2
2. Discussion of ETL strategy for messy ETL Case 3 (find the inconsistencies)
3. Solve ETL Case 3

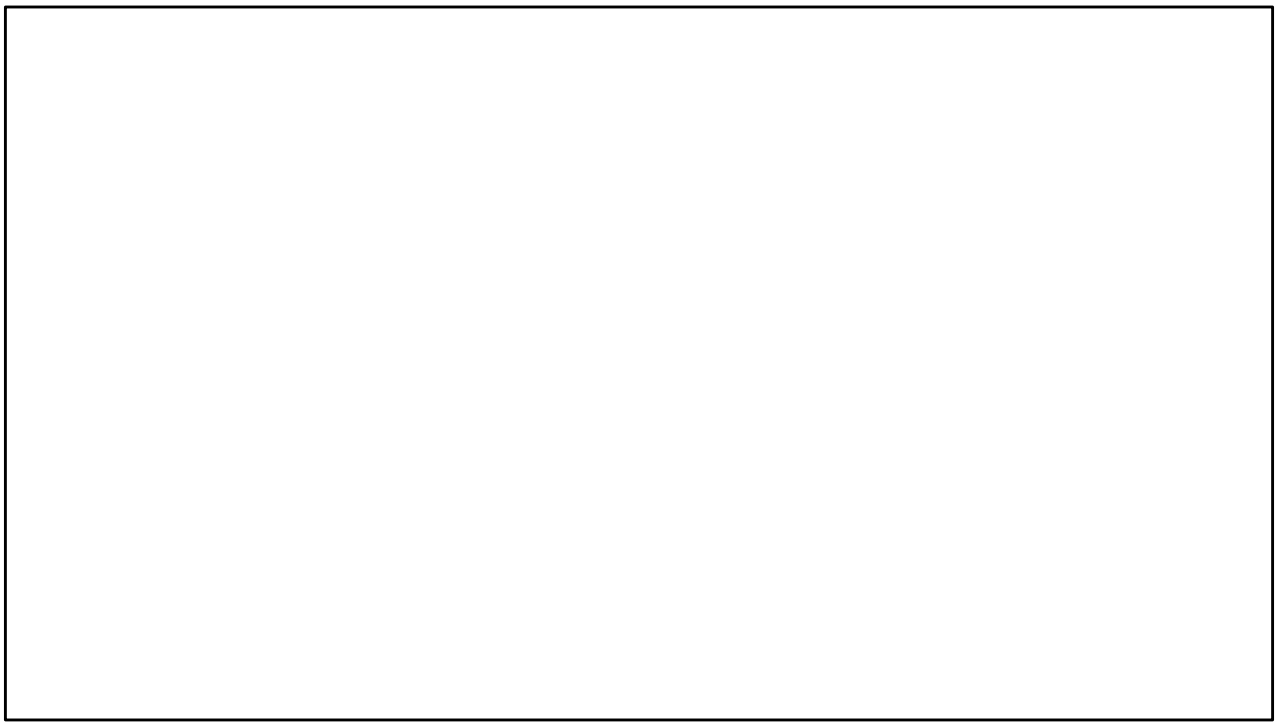

Results - Input Data (5) - Output
3 of 3 Fields
Cell Viewer
597 records displayed
Record
EmployeeCode
FirstName
LastName
JAP1080-19M
Carl
Rodriguez
AUS1241-21M
Anne
Wright
AUS1297-21M
Patrick
Rivera
GER1332-21W
Ruby
Walker
GER1202-10M
Teresita
Guillen
ARG1028-20M
Earl
Morgan
CAN1344-19W
Joella
Causey
JAP1248-24W
Gregg
Kelly
ARG1279-10W
William
Reed
ENG1198-23W
Julia
Johnson
AUS1435-21M
Tonya
Ruiz
ARG1094-14M
Wayne
Howard
Discuss your Case 1 strategy to get from the input to the output
5 new output fields
Results - Output Data (18) - Input
8 of 8 Fields
Cell Viewer
597 records displayed
Record
EmployeeCode
FirstName
LastName
Location
EmpID
PlantID
PayPeriod
LocPlantID
1 JAP1080-19M Carl Rodriguez JAP 1080 19 M JAP-19
2 AUS1241-21M Anne Wright AUS 1241 21 M AUS-21
3 AUS1297-21M Patrick Rivera AUS 1297 21 M AUS-21
4 GER1332-21W Ruby Walker GER 1332 21 W GER-21
5 GER1202-10M Teresita Guillen GER 1202 10 M GER-10
6 ARG1028-20M Earl Morgan ARG 1028 20 M ARG-20
7 CAN1344-19W Joella Causey CAN 1344 19 W CAN-19
8 JAP1248-24W Gregg Kelly JAP 1248 24 W JAP-24
9 ARG1279-10W William Reed ARG 1279 10 W ARG-10
10 ENG1198-23W Julia Johnson ENG 1198 23 W ENG-23
11 AUS1435-21M Tonya Ruiz AUS 1435 21 M AUS-21
12 ARG1094-14M Wayne Howard ARG 1094 14 M ARG-14

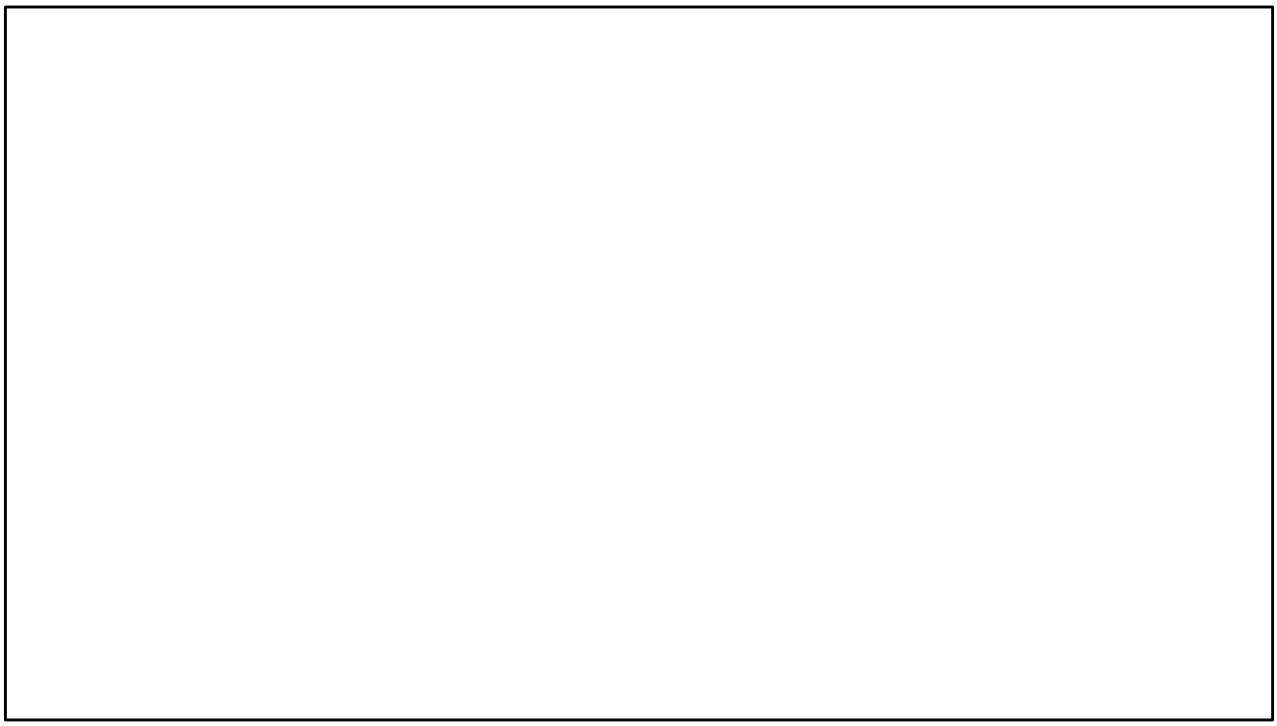





Results - Input Data (1) - Output
3 of 3 Fields
Cell Viewer
597 records displayed
Record
EmployeeCode
FirstName
LastName
14 Canada342-9W Sean Morgan
15 US476-10W Julie Johnson
16 MEX93-5M Samuel Perry
17 ENG442-9Mo Jarrod Beckett
18 Canada17-7W Lillian Nelson
19 ARG559-10W Delfina Brandon
20 ENG389-10W Shaun Jacobs
21 MEX121-11W Theresa Barnes
22 ENG489-12W Stephen Powell
23 AUS88-10W Queenie Thibodeau
24 AUS455-3W Juan King
25 AUS491-13M Laureen Varney
26 ENG193-10Mo Anthony Griffin
27 ARG508-14W Olga Rowe
Discuss your Case 2 strategy to get from the input to the output
Make a note of the inconsistencies in the input file
5 new output fields again
Results - Output Data (4) - Input
8 of 8 Fields
Cell Viewer
597 records displayed
Record
EmployeeCode
FirstName
LastName
Location
EmpID
PlantID
PayPeriod
LocationPlantID
14 Canada342-9W Sean Morgan CAN 342 9 W CAN-9
15 US476-10W Julie Johnson USA 476 10 W USA-10
16 MEX93-5M Samuel Perry MEX 93 5 M MEX-5
17 ENG442-9Mo Jarrod Beckett ENG 442 9 M ENG-9
18 Canada17-7W Lillian Nelson CAN 17 7 W CAN-7
19 ARG559-10W Delfina Brandon ARG 559 10 W ARG-10
20 ENG389-10W Shaun Jacobs ENG 389 10 W ENG-10
21 MEX121-11W Theresa Barnes MEX 121 11 W MEX-11
22 ENG489-12W Stephen Powell ENG 489 12 W ENG-12
23 AUS88-10W Queenie Thibodeau AUS 88 10 W AUS-10
24 AUS455-3W Juan King AUS 455 3 W AUS-3
25 AUS491-13M Laureen Varney AUS 491 13 M AUS-13
26 ENG193-10Mo Anthony Griffin ENG 193 10 M ENG-10
27 ARG508-14W Olga Rowe ARG 508 14 W ARG-14







Extra RegEx examples

Dates:

- Invoice Date: 03/15/2025
- Payment Due: 04/14/2025

Invoices:

- INV-100234
- INV-567890
- Reference: INV-456

Extra RegEx examples

Firm Identifiers:

- CIK: 0000320193
- Ticker: TSLA
- Ticker: AAPL

Matching dollar amounts and accounts:

- Total: $12,450.00
- Late Fee: $25
- Refund: ($1,200)

Extra RegEx examples

Phone numbers, email, credit card numbers:

- 206-123-4567
- real_person@example.com
- Potential fraud detected: 4532-9876-1234-5678

ETL Cases, Continued

Today we will examine “Joining Data”:

- Joining data is another **fundamental step** in the ETL process, as in most cases, developing insight will require data from different databases.
- Joining data can sometimes be referred to as merging data, blending data, or combining data.

Why is joining data important?

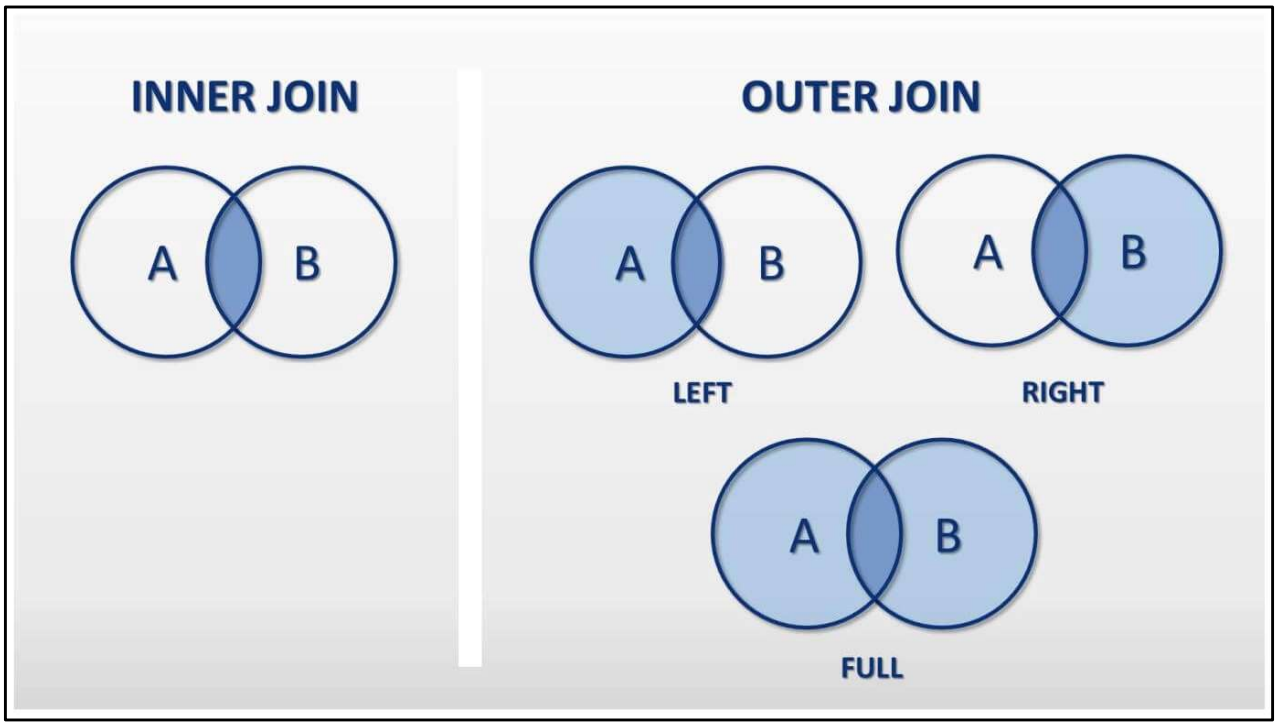
Joining data is part of the transforming step in ETL.

- Incorrect joins are a **data quality issue**, incorrect joins can:
  - **crash** (or really slow down) our programs!
  - have us **lose** important data.
  - Have us generate **incorrect conclusions/insight** from our analysis.

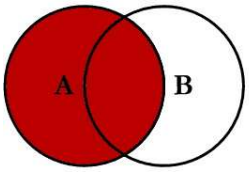
Joining using unique keys

In all cases we will use unique keys to join the data.

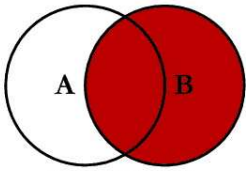
- All the unique keys are described in the case
- One of them was tricky, can you remember which one?


INNER JOIN
A
B
OUTER JOIN
A
B
LEFT
A
B
RIGHT
A
B
FULL

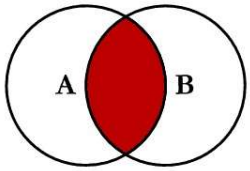
SQL JOINS



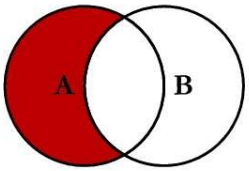

```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
```



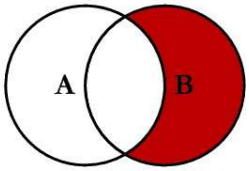

```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
```



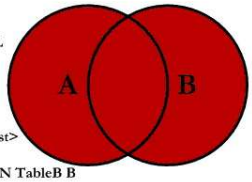

```
SELECT <select_list>
FROM TableA A
INNER JOIN TableB B
ON A.Key = B.Key
```



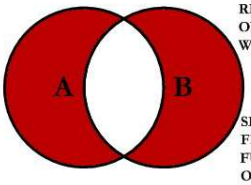

```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
WHERE B.Key IS NULL
```




```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
```




```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
```




```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
OR B.Key IS NULL
```

Joining using unique keys

In all joins we will use “one to many” matches.

- Joining the large information source (“JELineItems”) to the smaller set of information in the other data sheets.

Why join across multiple software?

Because joining data is so fundamental, we want to work towards becoming software agnostic.

Working with cutting-edge software allows you to:

- Become adaptable and resilient to **new software**.
- Impress recruiters with practical discussions of cutting-edge software.

The tools we will use:

Alteryx Joins:

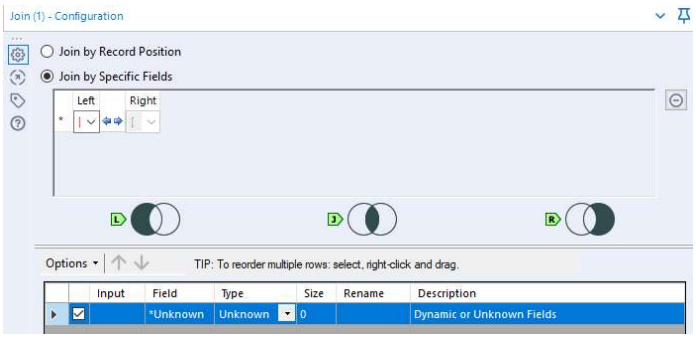


Tableau Relationships

JELineItems+ (Analytics_mindset_case_studies_ET...

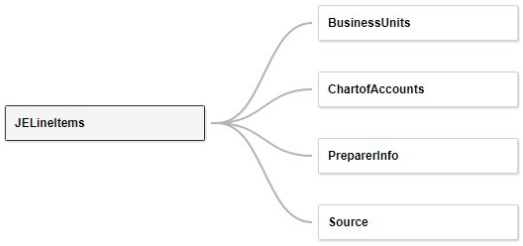
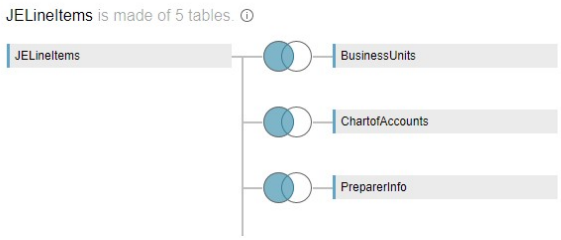
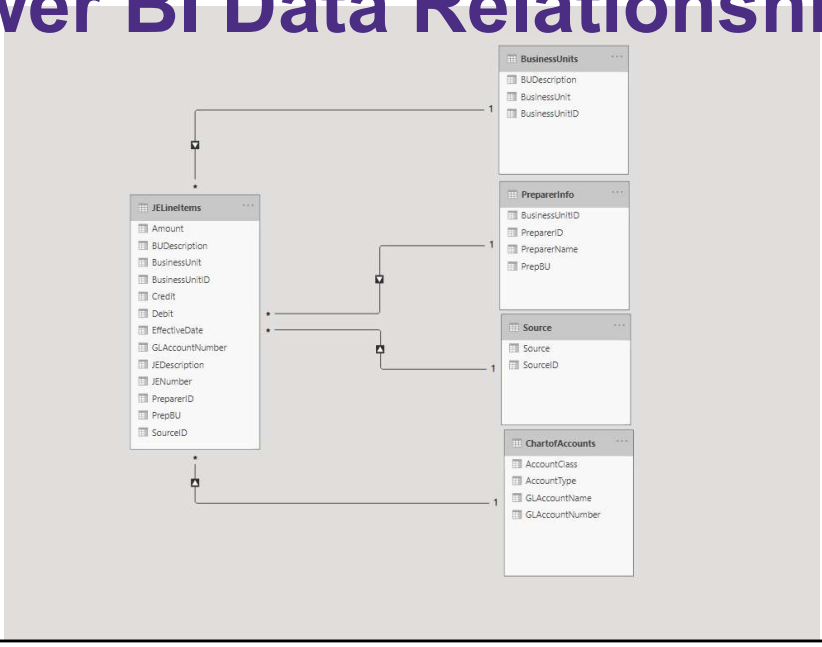


Tableau Joins



Power BI Data Relationships



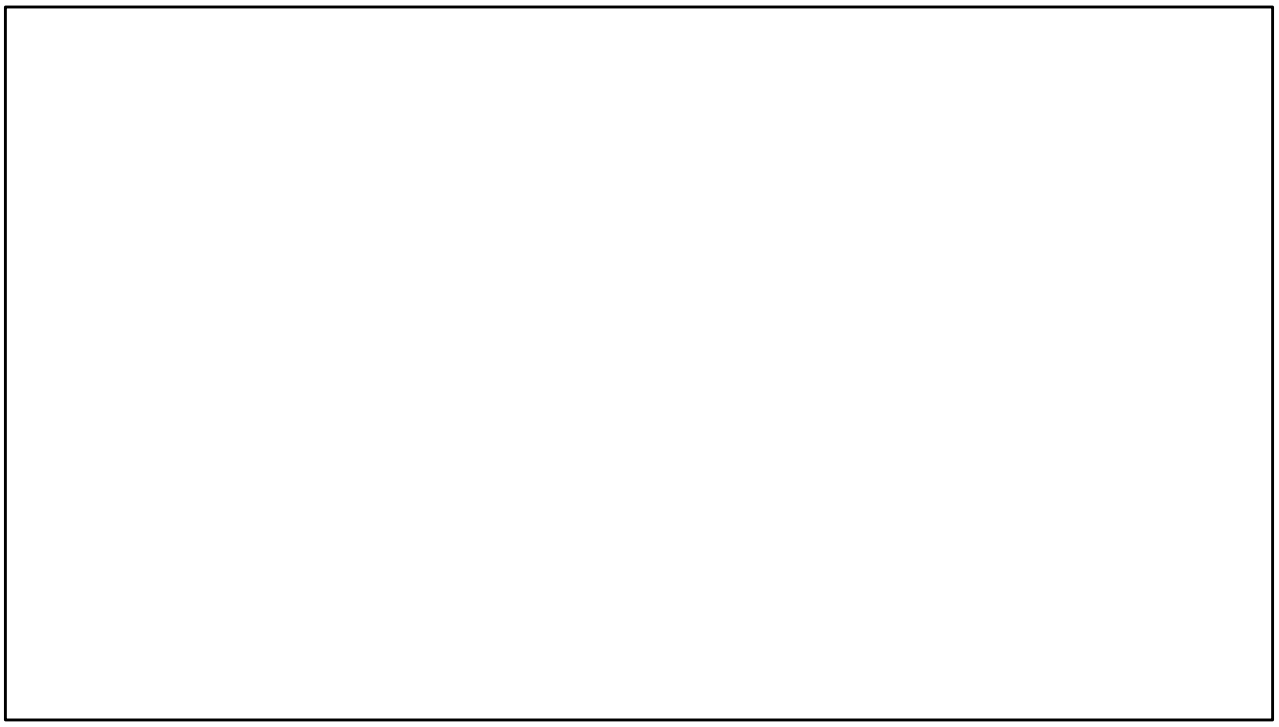
To FSB remote labs

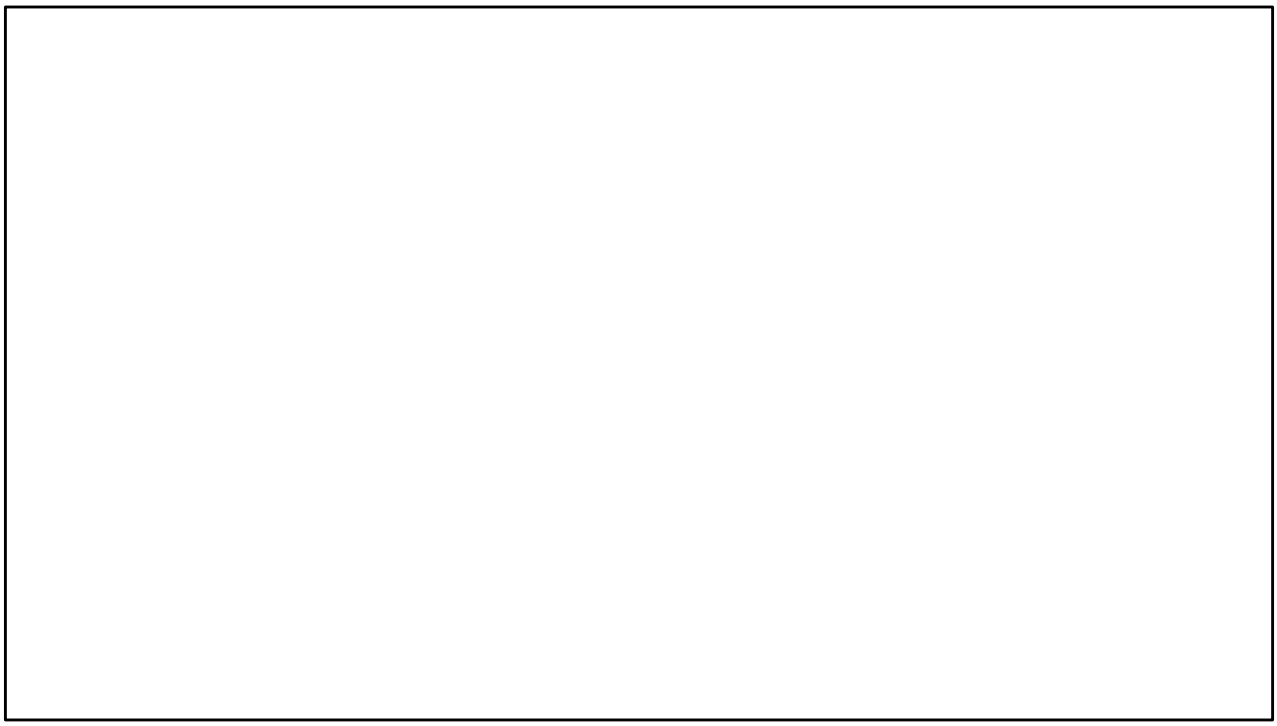
We will begin in Alteryx.

Again I recommend **saving screenshots** of your work in Alteryx (I've added blank slides if you want to save them in this ppt document, ctrl+m adds more slides)

Lab Agenda

1. Case Demonstration in Alteryx
2. Then in Tableau (time permitting)
3. Then in Power BI (time permitting)





Conclusion and look-ahead

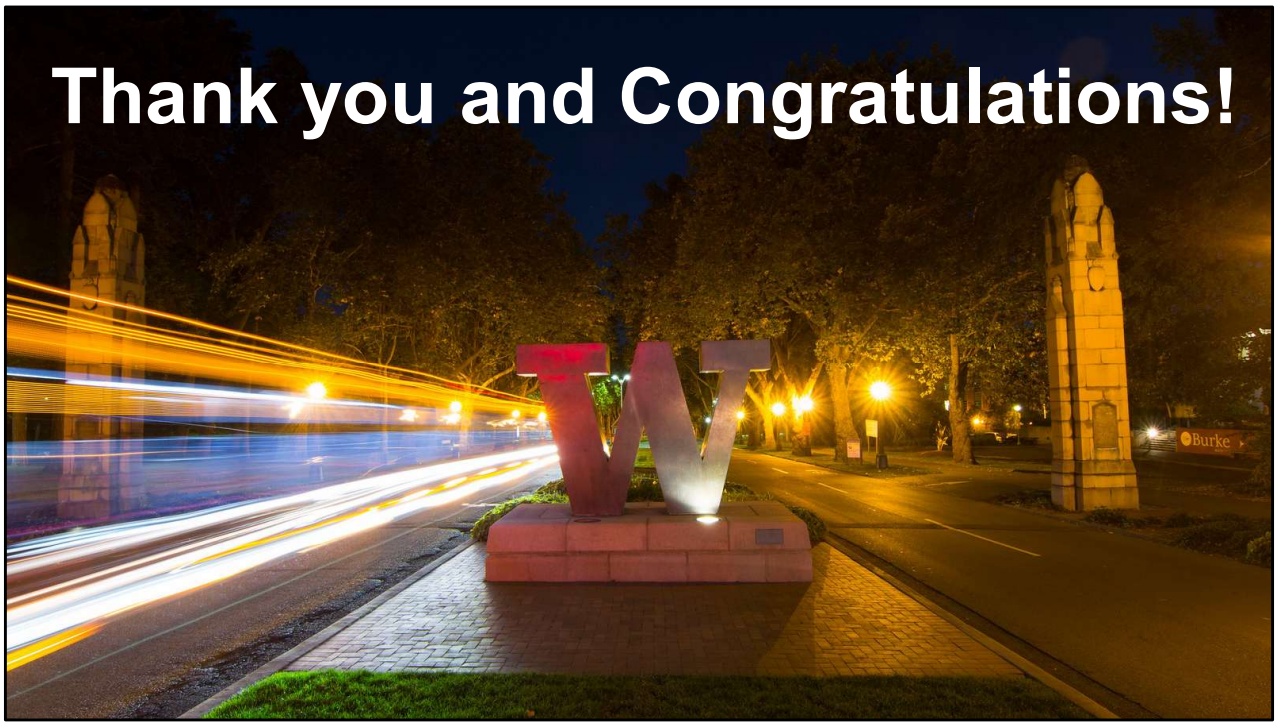
ETL is the first step in the analytical process – we cannot undertake data analytics without data!

ETL is especially important when the data needs to be cleaned or transformed for use in analytics.

Conclusion and look-ahead

Wednesday:

- Cases: Enron Emails Case – RegEx Heavy


Thank you and Congratulations!
Burke