Week 3
Let’s Talk About Transforms

ISTA 322 - Data Engineering

Last week - recap

- Data types and data structures
- Python refresher
- Introduction to Google colab / Jupyter notebooks

Middle out

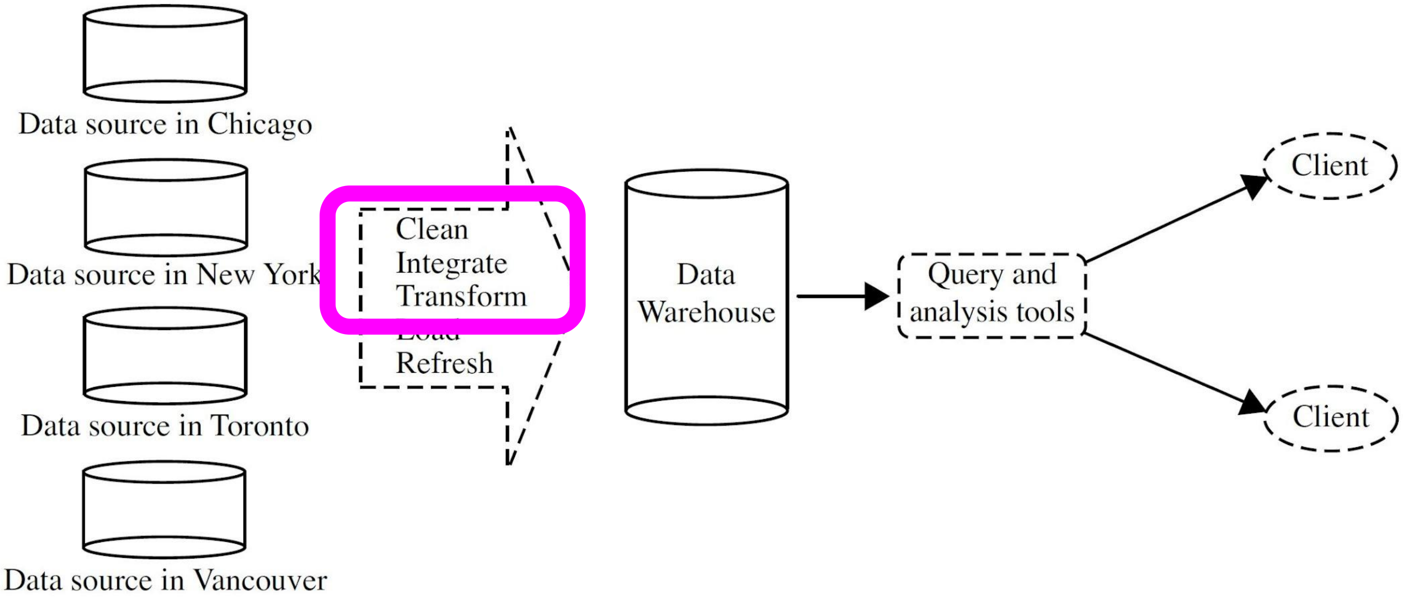


Figure 1.6 Typical framework of a data warehouse for *AllElectronics.*

Starting with the T in ETL

- Why?
- You all know *some* python
- Most of you have done some kind of data transforms
- It’s the core of ETL
- The other ends are a bit more of making technology work for you (sorta)
- Overall, easier to build from the T in ETL

Data transforms

- Transforms are key to making data useful
- Two general purposes of transforms
  - Fixing errors
  - Making data useable / relevant
- We’ll cover major examples within each
  - Solid foundation / framework
  - You’ll need to learn more throughout your careers

  * Importance of “learning to learn”
- Depends a lot on existing data quality and infrastructure

Fixing errors

- Existing data can be filled with errors
- Lack of precision
    - Data was entered wrong
    - Missing values
    - Bad fills
- Unnecessary
    - Too many columns
    - End user doesn’t actually need
    - Long mappings
- Repetitive
    - Duplication?
    - Not normalized?

Making data useable / relevant

- Different clients will have different needs
- Need to transform data into a format that is useful
- Business intelligence
  - Aggregate daily statistics
  - Make key metrics to display
  - Filtering
- Data science
  - Extract data from strings
  - Scale & alter units
  - Binning/grouping
  - New features
- Of course there are more, but not covering them all

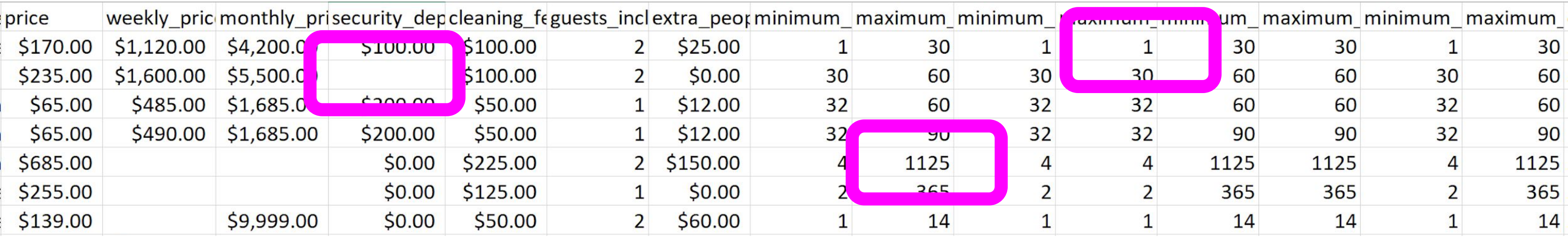
One ETL to rule them all

- Couple notes
- Don’t think you’re going to set up just one ETL
- Just because you work with a DE doesn’t mean you won’t have to do an ETL
- ETL doesn’t have to fix everything
- We’re focused on a general framework here!

Transforms - errors

Data sources can be filled with errors

- Bad defaults might need to be fixed
  - Fill empty cells with 9999
  - Allow too wide of range on data input
- Just allowing empty cells
  - Could be fine
  - Might need to fill

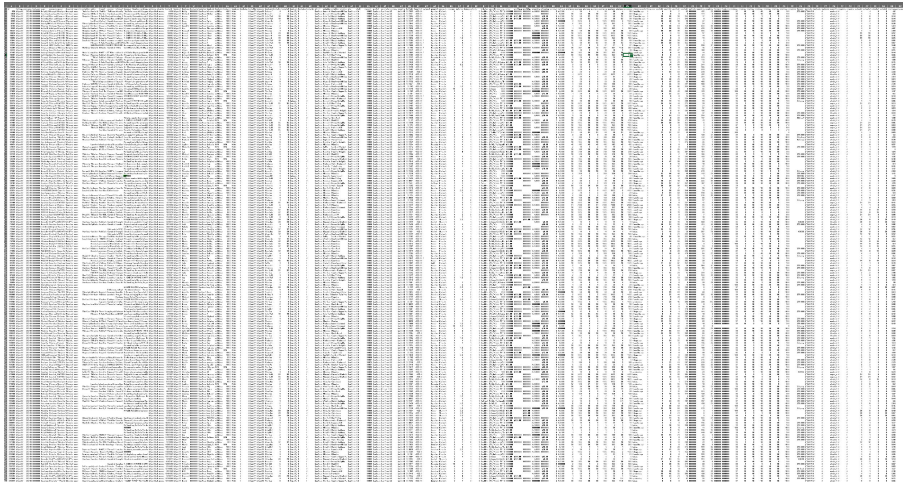


price	weekly_pric	monthly_pri	security_dep	cleaning_fe	guests_incl	extra_peop	minimum_	maximum_	minimum_	maximum_	minimum_	maximum_	minimum_	maximum_
$170.00	$1,120.00	$4,200.00	$100.00	$100.00	2	$25.00	1	30	1	1	30	30	1	30
$235.00	$1,600.00	$5,500.00		$100.00	2	$0.00	30	60	30	30	60	60	30	60
$65.00	$485.00	$1,685.00	$200.00	$50.00	1	$12.00	32	60	32	32	60	60	32	60
$65.00	$490.00	$1,685.00	$200.00	$50.00	1	$12.00	32	90	32	32	90	90	32	90
$685.00			$0.00	$225.00	2	$150.00	4	1125	4	4	1125	1125	4	1125
$255.00			$0.00	$125.00	1	$0.00	2	365	2	2	365	365	2	365
$139.00		$9,999.00	$0.00	$50.00	2	$60.00	1	14	1	1	14	14	1	14

Transforms - errors

Data sources can be filled with errors

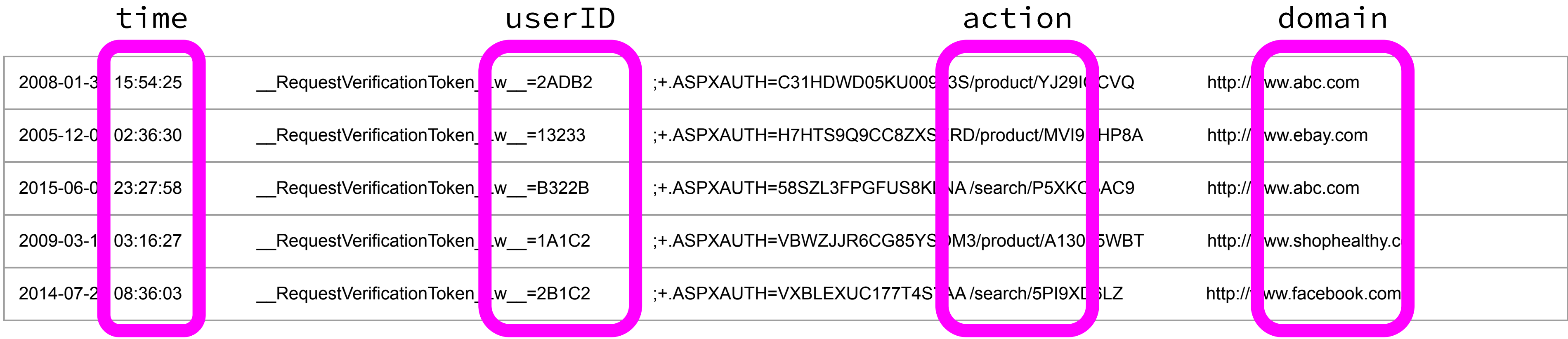
- Bad defaults might need to be fixed
  - Fill empty cells with 9999
  - Allow too wide of range on data input
- Just allowing empty cells
  - Could be fine
  - Might need to fill
- Excessive data!
  - 106 columns
- Unnecessary data
  - Remove identifying info?



Transforms - Making data useable / relevant

Many ways to accomplish this

- Extract data from strings
    - Splitting columns to parse information
    - Regex fun times **‘[0-9]{5}$’**
    - Data type conversions

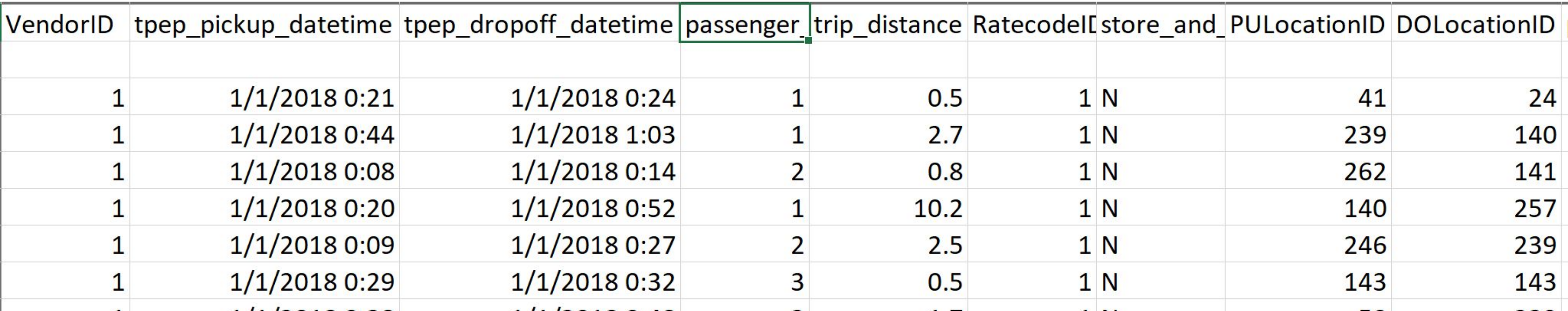

time
userID
action
domain

2008-01-3 15:54:25	__RequestVerificationToken_.w__=2ADB2	;+.ASPXAUTH=C31HDWD05KU009 3S/product/YJ29I CVQ	http:// ww.abc.com
2005-12-0 02:36:30	__RequestVerificationToken_.w__=13233	;+.ASPXAUTH=H7HTS9Q9CC8ZXS RD/product/MVI9 HP8A	http:// ww.ebay.com
2015-06-0 23:27:58	__RequestVerificationToken_.w__=B322B	;+.ASPXAUTH=58SZL3FPGFUS8KI NA /search/P5XKC AC9	http:// ww.abc.com
2009-03-1 03:16:27	__RequestVerificationToken_.w__=1A1C2	;+.ASPXAUTH=VBWZJJR6CG85YS M3/product/A130 5WBT	http:// ww.shophealthy.c
2014-07-2 08:36:03	__RequestVerificationToken_.w__=2B1C2	;+.ASPXAUTH=VXBLEXUC177T4S AA /search/5PI9XD LZ	http:// ww.facebook.com

Transforms - Making data useable / relevant

Working with event level data

- Binning
  - ‘short’, ‘medium’, ‘long’ trips from trip distance
- Aggregating
  - Number of trips per time, popular locations

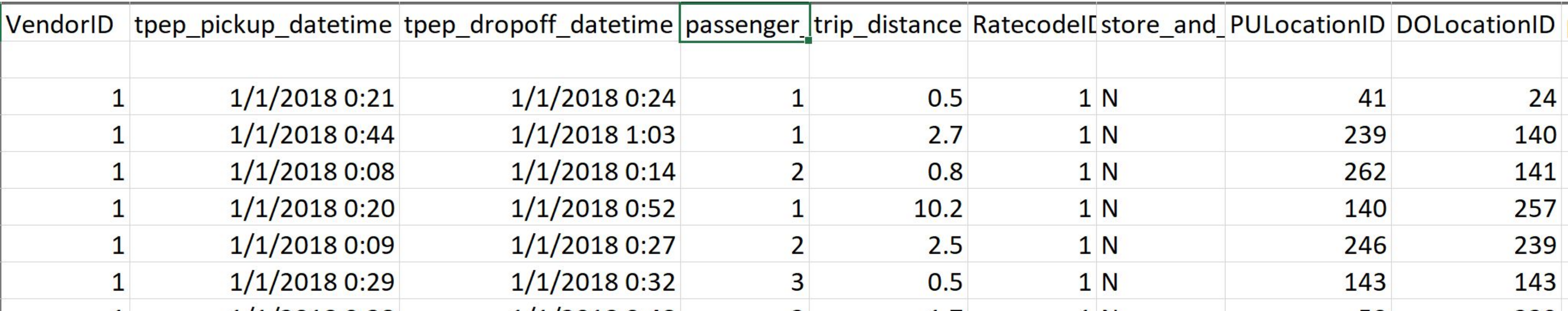


VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_	trip_distance	RatecodeI	store_and_	PULocationID	DOLocationID
1	1/1/2018 0:21	1/1/2018 0:24	1	0.5	1	N	41	24
1	1/1/2018 0:44	1/1/2018 1:03	1	2.7	1	N	239	140
1	1/1/2018 0:08	1/1/2018 0:14	2	0.8	1	N	262	141
1	1/1/2018 0:20	1/1/2018 0:52	1	10.2	1	N	140	257
1	1/1/2018 0:09	1/1/2018 0:27	2	2.5	1	N	246	239
1	1/1/2018 0:29	1/1/2018 0:32	3	0.5	1	N	143	143

Transforms - Making data useable / relevant

Working with event level data

- Altering units
- Calculating useful metrics
  - Trip time
  - Speed of trip
  - Speed at different times



VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_	trip_distance	RatecodeID	store_and_	PULocationID	DOLocationID
1	1/1/2018 0:21	1/1/2018 0:24	1	0.5	1	N	41	24
1	1/1/2018 0:44	1/1/2018 1:03	1	2.7	1	N	239	140
1	1/1/2018 0:08	1/1/2018 0:14	2	0.8	1	N	262	141
1	1/1/2018 0:20	1/1/2018 0:52	1	10.2	1	N	140	257
1	1/1/2018 0:09	1/1/2018 0:27	2	2.5	1	N	246	239
1	1/1/2018 0:29	1/1/2018 0:32	3	0.5	1	N	143	143

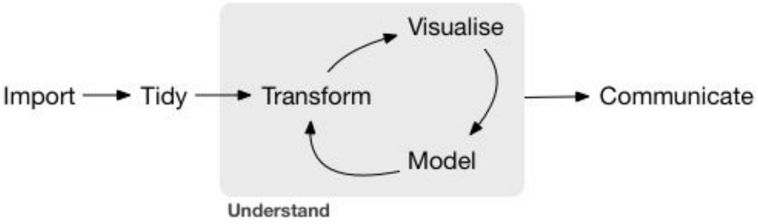
Checking your data

You should constantly test the ‘integrity’ of your data by asking questions and checking expectations

- General external formats
  - Right number of digits in zip/phone/social/ID/etc
- Bounded in reality
  - Age in 0-115
  - Webpage had more views than clicks
  - Only positive for some measures
- Appropriate data structure
  - Right number of rows and columns
  - Correct data types

Wrapping up

- The actual actions depend on the data ***and*** what you’re going to use it for
- You’ll still need to do a lot of this even if you’re not a DE
  - Lots of my other DS classes have data cleaning lessons!



- Don’t be afraid to use a subset to work on
- Let’s go actually do this