Melanie R Loaiza
Logical data flow
CS689 — Final Project

Box (Boston transform notes):
1. Drop columns _fulltext, _id
2. dateFormat convert (violdttm)
3. viol_level from " ** **** to critical and not critical
4. Add new values : actions (no actions taken) , legalowner na set not applicable
5. Drop null values in dates and locations (address , city , state and zip)
6. Filter violation descriptions (mf_violdesc) and map (legal_reference, detection_method, violation_category, common_cause.
7. Random data :
   7.1 record_date set random dates after violdttm
   7.2 score set values between 0-14 and 14-27 (viol_level)
8. sourcesystem set to BostonOpenData

Box (NYC transform notes):
1. dateFormat convert (inspection_date)
2. Filter critical_flag critical and not critical values
3. Map actions (closed , re-open, violationscited, noviolations)
4. Remove null values (buildings)
5. Random data:
   5.1 legalowner
   5.2 viol_status (pass or fail) depending on the actions
   (Pass = re-open or noviolations , else fail)
6. sourcesystem set to NewYorkOpenData
7. Add state set to New York

Flow nodes:
- BostonOpenData — Extract — (Food Establishment Inspections) → Import data from API — Load → bostonData dataFrame — Transform — Python → bostonData Transformed — PosgreSQL → Boston inspectionStaging — Load
- NYOpenData — Extract — (DOHMH Restaurant Inspections) → Import data from csv — Load → nyData dataFrame — Transform — Python → nyData Transformed — PosgreSQL → NYC inspectionStaging — Load
- Normalized restaurantNames cuisineDescription
  - distinct restaurants with IDs → Boston restaurantStaging
  - distinct restaurants with IDs → NYC restaurantStaging → Insert new / Check updates / Update old restaurants → distinct restaurants → restaurantDIM SCD Type 2
  - distinct restaurants with IDs → inspectionFact_staging — Transform → Identify distinct Violation Descriptions / Update CASE for NY violations → violationStaging → Overwrite changes in existing violations / Insert new ones → distinct violations → violationDIM SCD Type 1
  - → Identify distinct dates / Normalization → dateStaging → Insert date values → distinct dates → datetDIM SCD Type 0 — Load