Assumptions & Observations :

- Zip source file sent over on some frequency
- Data from US only
- Zillow listing end date is not available. Assume all listings in zillow dataset are active
- Dataset doesn't have data for all 50 states
- Format of all rentals data(even from sources other than Zillow) & sales data are same(even from sources other than MLS)
    - JSON - rentals
    - CSV - REAL ESTATE

Final Aggregated Table & Mapping with source data:

FINAL Aggregated Table:	Zillow Master Table:	MLS Master Table:
real_estate_gold	rent_df_master	sale_df_master
id	id	mls_id
listing_type	home_status	'ForSale' (***hardcoded***)
source	source	source
home_status	'Active' (***hardcoded***)	status
market	market	market
home_type	home_type	property_type
address	address	address
city	city	city
postal	postal	zip_or_postal_code
region	region	state_or_province
price	price	price
sqft	sqft	square_feet
beds	beds	beds

baths	baths	baths
year_built	year_built	year_built
load_ts	load_ts	load_ts

Future Work :

- Automate unzip and load from zip source file
- read only .json or .csv files
- File Name as a column and check if already existing
- Incremental load with last updated value from source
- Logs
- Notifications
- Add source file name to master table to prevent same file being loaded twice, causing duplicate data
- Put together a single python file for daily scheduled run