--- Data & File Descriptions ---

A) Raw data files

Raw data includes the data measured by measurement devices/sensors from the field/lab (all csv in this project) in "Mahdavi & Siegel (2021) Indoor Air" (particulate matter (PM) phase). Due to data privacy, only a representative sample of the raw data (meaning one from each type) is shared (in the "Raw Data Files" folder).

See "Table 1.pdf" at the end of this document, "Data Pipeline and Code Configuration.jpg", and "list_of_files.xlsx" files to understand more about raw data.

B) Pre-processed files

These files have been already processed by other team members (using STATA or Python). They have a format of either ".dta", ".csv", or ".xlsx" (depending on how stored from STATA or python). They play the role of raw files in this project, but still called as pre-processed file data as they don't come from purely raw data (e.g., sensors or lab sheet completions). Table 2 has a list of these files (coming at the end of this document).

C) Code files

Codes include all ".py" or ".ipynb" that read raw data or previously processed data to generate processed data, figures, calculations, or statistical analyses. There are four different code types used in this project:

- **Processing codes:** which process from raw data (stored in the "Processing Codes" folder).
- **Generic code:** which does generic processing such as path file name correction or labelling categorical parameters (not shown in "Data Pipeline and Code Configuration.jpg").
- **Calculation codes:** which present numeric or statistical results from the processed data (stored in the main "Code" folder).
- **Visualization codes:** which aim to plot figures. This may come with additional processing of data to make the dataframe compatible for plotting (stored in the "Plotting Codes" folder).

See "Data Pipeline and Code Configuration.pdf", and "list_of_files.xlsx" files to see the list of code blocks. (Two generic codes titled "notion_correction.py" and "labels_all.py" is not shown in "Data Pipeline and Code Configuration.pdf" but it is called by many code blocks).

D) Processed data files

Processed data includes the data generated after processing in data pipelines in "Mahdavi & Siegel (2021) IA". See "Data Pipeline and Code Configuration.pdf", and "list_of_files.xlsx" files to check the processed data lists and blocks. "Data Pipeline & Code Configurations.jpg" illustrates the entire data pipeline processing that generates processed data using codes from the raw data or previously processed data. Table 3 (at the end of this document) has a list of all these files.

A separate code block ("df_summary.ipynb") also presents a summary of some processed files. For data privacy purposes, not all the processed dataframes have been presented. Processed files are not shared as spreadsheet files either.

E) Plots

Plots include all the figures presented in “Mahdavi & Siegel (2021) IA” (PM Phase) from the processed data. See “Data Pipeline and Code Configuration.pdf”, and “list_of_files.xlsx” for more information. The plots aren’t presented in separate files (e.g., jpg) but illustrated in the same code file generating them (in Jupyter). See “Plotting Codes” for more information.

F) Other files

Any other file (mostly guidelines or descriptions) not classified above. A full list is available in “list_of_files.xlsx” (sheet “other_summary_files”) provided in the repository.

Table 1 – Raw Data Summary (Mahdavi & Siegel (2021) IA – PM Phase)

Item #	Raw Data Files & Name Conventions	Raw Data File Description	Variables/Columns	Variables/Columns Explanation
1	filters_weight_gain_master.xlsx	Spreadsheet that includes all gravimetric analyses results from weighting of filters pre and post deployment plus post-extraction.	SN	Filter serial number: 1-90
			site	Site at which filter was deployed: 1-21
			round	Round in which filter was deployed: 1-4 (proxy for seasons)
			filter_size	Dimensions of filter (20×24×1 inch)
			Ext_loc	Location where filter was extracted (Institute for Dust Analysis, or Blue Heaven Technologies)
			m_pre_d	Mass of filter prior to deployment
			m_post_d	Mass of filter after the deployment (with the dust collected in it)
			m_post_ext	Mass of filter after dust extraction from filter.
2	3-month_schedule_all.xlsx	Spreadsheet with deployment and collection of filters in the studied sites' HVAC systems.	site	Site at which filter was deployed: 1-21
			date_start	Date on which filter was deployed in the site
			time_start	Time (in hh:mm) at which the filter was deployed in the site
			round	Round in which filter was deployed: 1-4 (proxy for seasons)
			date_end	Date on which filter was collected from the site
			time_end	Time (in hh:mm) at which the filter was collected from the site
3	filters_selected.xlsx	List of filters sent to Blue Heaven Technologies Inc., for further testing.	site	Site at which filter was deployed: 1-21
			date_install	Date on which filter was deployed in the site
			filter_type	Type of filters: MERVs 8, 8E, 11, and 14
			duration	No. of days filter was in service in the HVAC of the studied homes

Table 2 – Pre-processed Data Summary (Mahdavi & Siegel (2021) IA – PM Phase)

Item #	Raw Data Files & Name Conventions	Pre-Processed Data File Description	Variables/Columns	Variables/Columns Explanation
1	tsp_mass_error.dta	Error/uncertainties associated with TSP mass collected on the filters.	SN	Filter serial number: 1-90
			round	Round in which filter was deployed: 1-4 (proxy for seasons)
			site	Site at which filter was deployed: 1-21
			ft	Filter type: MERVs 8, 8E, 11, and 14 (labelled as 1-4)
			TSP_mass_error	TSP mass uncertainty associated with the filter
2	eff_effectiveness.csv	Efficiency of HVAC filters deployed in this study per size bin.	site	Site at which filter was deployed: 1-21
			bin	Bin size for the efficiency
			filter_type	Filter type: MERVs 8, 8E, 11, and 14 (labelled as 1-4)
			filter_condition	New (not yet deployed), old (loaded after deployment)
			filter_eff	Filter efficiency associated with the bin size
			mu_filter_eff	Measurement uncertainty in "filter_eff"
3	Filtration volume by day.dta	Filtration volume measured by heating, cooling, and fan-only modes in the course of filter deployments	day	Day # since the beginning of the project (first filter deployment)
			method	Method by which system runtime was measured (motor on-off sensor, supply duct, pressure sensor etc.,)
			filter_type	Type of filters: MERVs 8, 8E, 11, and 14
			fan_cool_volume	Filtration volume by cooling mode
			heat_volume	Filtration volume by heating mode
			Fan_only_volume	Filtration volume by fan only mode
			total_volume_cf	Total filtration volume (in cf)
4	Filtration_volume_erros.xlsx Filtration_volume_values.xlsx	21×4 array-like spreadsheets correspond to site no. and filter type.	N/A	
5	natl_psd_master.xlsx	See {repository name} Table 2 natl_psd_master.xlsx and natl_d_master.xlsx for more info.	N/A	
6	dc1700_append.dta (× 21)	DC1700 data collected from all 21 studied sites recording time-series particle number concentration.	small_CountPerFoot3	Small bin size (>0.5 µm) number concentration of particles
			large_CountPerFoot3	Large bin size (>2.5 µm) number concentration of particles
			time	Date and time recorded by DC1700 sensors

Table 3 – Processed Data Summary (Mahdavi & Siegel (2021) IA – PSD Phase)

Item #	Processed File Name	Processed File Description	Variables/Columns	Variables/Columns Explanation
1	pm_master.xlsx	A master spreadsheet including all PM relevant data	SN	Bin size of PSD (µm)
			site	The site where the filter was deployed to collect dust: 1-20
			round	The round where filter was deployed (proxy for season): 1-4
			ft	Filtre types - 1: MERV8, 2: MERV8E, 3: MERC11, 4: MERV14
			TSP Concentration	TSP Concentration measured by quantitative filter forensics technique
			TSP Concentration Error	Uncertainty in TSP concentration
			PM10	PM_{10} concentration measured by quantitative filter forensics technique
			PM10 Error	Uncertainty in PM_{10} concentration
			PM2.5	$PM_{2.5}$ concentration measured by quantitative filter forensics technique
			PM2.5 Error	Uncertainty in $PM_{2.5}$ concentration
			TSP mass	Dust mass collected on the HVAC filters (as total suspended particles)
			TSP mass error	Uncertainty in TSP mass
			filtration volume	Filtration volume of the air passing through HVAC filter
			filtration volume error	Uncertainty in filtration volume
			PM2.5 Fr	Fraction of $PM_{2.5}$ in the HVAC filter dust
			PM10 Fr	Fraction of PM_{10} in the HVAC filter dust
			eff_2.5	HVAC filter efficiency of $PM_{2.5}$
			eff_2.5_error	Uncertainty in HVAC filter efficiency of $PM_{2.5}$
			eff_10	HVAC filter efficiency of PM_{10}
			eff_10_error	Uncertainty in HVAC filter efficiency of PM_{10}
			Ext_loc	Location where dust extraction from HVAC filter took place (IDA or BHT)
2	sn_summary.xlsx	Essential information of filters (e.g., SN, site, round)	SN, site, round, ft, and Ext_loc	See pm_master.xlsx for more info
3	pm_d_drop.xlsx	Important information from PM data merged with: 1) d-value of the dust samples collected from HVAC filters deployed in the studied sites; and 2) building properties of the studied sites	Several columns from pm_master.xlsx PLUS	See pm_master.xlsx for more info
			Dx (0), Dx (10), Dx (25), Dx (50), Dx (75), Dx (90), Dx (100)	Size cutoffs for minimum, 10^{th}, 25^{th} , median, 75^{th}, 90^{th}, and maximum percentiles
			building_type	Building types: Semi-detached, deteched, townhouse, condos.
			basement_apartment	Where or not a basement apartment existed (Y/N)
			floor_area	Net floor are in the studied home/suite
			ceiling_height	Ceiling height summed over all floors
			volume	Home total volume
			no_occupants	Number of occupants living in the studied home.
			no_pets	Number of pets existing in the studied home.
			pet_type	Pet type: Cat, dog, hamster
			evidence_pm_source	Where or not a visual evidence of smoking existing in the home (Y/N)

			p_air_cleaner	Where or not a visual evidence of air cleaner existing in the home (Y/N)
			height_ave	Average height of the home (Ceiling heights divided by No. of floors)
4	dc_1700.xlsx dc_1700_agg.xlsx	Particle number concentration data	site	The site where the filter was deployed to collect dust: 1-20
			round	The round where filter was deployed (proxy for season): 1-4
			DC 0.5-2.5	Small channel particle number concentration
			DC > 2.5	Large channel particle number concentration
			stat	Percentiles: 10^{th}, 50^{th} (median), and 90^{th} (dc_1700_agg.xlsx) only