Note: These sequential steps do not refer to specific functions. A single step may involve multiple function calls. The idea behind the steps is just to group the overarching tasks of the program together.

0) Set Constants

[One of the purposes of defining most of the constants up here at the top is to make it easy for Jiachen/others to find the constants they need to change in the future to adapt the model to new data]

1) Cluster NC data

a) load+aggregate NC data (including weather), grouping by sensor ID fields [and 'unit'?]

i) ??? Would it make sense to save the aggregation to .csv and add in functionality to continue the most recent date so that we don't need to re-query if we already have the data? (Would a stretch goal if we have time)

b) Encode and scale NC data

c) cluster NC data to get df of sensor id fields + cluster group number

d) Reload NC data + join cluster group num + aggregate, this time grouping by date, time, and clust_group_num

OUTPUT OF STEP1 = dataframe with 5 numbers (mean, stddev,min,max,update rate) per cluster group (shown below as c1, c2). 1 row per aggregated time period (below is by hour: 1 row per hour for every day in the date range. 30 days would be 30x24=720 rows)

NC DATA

Date	Hour	mean_0	std_0	min_0	max_0	urate_0	mean_1	std_1	min_1	max_1	urate_1	..._cn
2020-05-01	0	55.2	24.1	0	100	15	10	.1	2	18	1000	
2020-05-01	1	50.1	14.2	5	80	15	10	.1	2	18	1000	
...	...	...	...	...	...	...	...	...	...	...	...	
2020-05-01	23	37	19	1	64	15	5	.1	2	18	1000	
2020-	0	48	12	10	90	18	10	.1	2	18	1000	

05-02												

2) Model EC/NC relationship

a) Load+aggregate EC data, grouping by date, time, and sensor ID fields (no feature selection needed yet!)

EC DATA

uniqueID	date	hour	mean
AHU-01 SF Air Systems Energy AHU1_SF_VFD_PWR(kWh)	2020-05-01	16	4746.326782
AHU-01 SF Air Systems Energy AHU1_SF_VFD_PWR(kWh)	2020-05-01	17	4748.786743
...	...		
AHU-01 SF Air Systems Energy AHU1_SF_VFD_PWR(kWh)	2020-05-01	20	4753.706665
AHU-01 SF Air Systems Energy AHU1_SF_VFD_PWR(kWh)	2020-05-01	21	4756.166992

b) Also create second DF by aggregating further just using sensor ID fields for the entire time range (end result = 1 row per sensor)

uniqueId	mean	max	min	std	update_rate
AHU-01 SF Air Systems Energy AHU1_SF_VFD_PWR (kWh)					
AHU-02 SF Air Systems Energy AHU1_SF_VFD_PWR (kWh)					
...					
Rm 6203G EF-13 Floor 6 Energy EF_13_VFD_PWR(kWh)					
Rm 6206B EF-11 Floor 6 Energy EF_11_VFD_PWR(kWh)					

c) For each unique EC sensorID (i.e. each row in 2b_EC_data_df), create a Ridge Regression model using 2a_EC_data_df and step1_output_NC_data_df. Model is basically: Y=EC response of the unique EC sensor and Xn=NC data

These are all the coefficients from the Ridge Regression models								Unique EC SensorID
0	**1**	**2**	**3**		**17**	**18**	**19**	**uniqueID**
0.000037	-0.004377	0.0	-0.000041	...	5.876493	8.502804	20.087383	AHU-01 SF Air Systems Energy AHU1_SF_VFD_PWR(kWh)
0.000039	-0.004622	0.0	-0.000044	...	6.537176	8.851925	20.473544	AHU-02 SF Air Systems Energy AHU2_SF_VFD_PWR(kWh)

OUTPUT OF STEP2 = dataframe with EC sensor ID fields and all n coefficients from that unique EC sensor's Ridge Regression model. (a single row basically represents the results from a single Ridge Regression model)

3) Prep EC data for classification model

a) Load metadata and join with 2b_EC_data_df (inner join)

	2b_EC_data_df					metadata				
uniqueId	**mean**	**max**	**min**	**std**	**update_rate**	**unit**	**energy**	**power**	**water**	**sensor**
AHU-01 SF Air Systems Energy AHU1_SF_VFD_PWR(kWh)						kWh	yes_energy	no_power	no_water	yes_sensor
AHU-02 SF Air Systems Energy AHU1_SF_VFD_PWR(kWh)						kWh	no_energy	no_power	no_water	yes_sensor
...						kWh	no_energy	no_power	no_water	yes_sensor
Rm 6203G EF-13 Floor 6 Energy EF_13_VFD_PWR(kWh)						kWh	yes_energy	no_power	no_water	yes_sensor
Rm 6206B EF-11 Floor 6 Energy EF_11_VFD_PWR(kWh)						kWh	yes_energy	no_power	no_water	yes_sensor

b) Apply feature selection function(s) to the joined EC+metadata (by joining with NRCan labels)

	2b_EC_data_df					metadata				training data			
uniqueId	**mean**	**max**	**min**	**std**	**update_rate**	**energy**	**power**	**water**	**sensor**	**ALEX-NRCanLabelGuess**	**isGas**	**equipNew**	**navNew**
AHU-01 SF Air Systems Energy AHU1_SF_VFD_PWR(kWh)						yes_energy	no_power	no_water	yes_sensor	4_Auxillary_Motors	no_gas	Air_Equip	Energy
AHU-02 SF Air Systems Energy AHU1_SF_VFD_PWR(kWh)						no_energy	no_power	no_water	yes_sensor	4_Auxillary_Motors	no_gas	Air_Equip	Energy

c) Encode and scale the EC+metadata

	2b_EC_data_df					Encoding after feature selection				NRCan labels
uniqueId	**mean**	**max**	**min**	**std**	**update_rate**	**isGas_yes_gas**	**energy_no_energy**	**energy_yes_energy**	**equipRef_Air_Equip**	**ALEX-NRCanLabelGuess**
AHU-01 SF Air Systems Energy AHU1_SF_VFD_PWR(kWh)						0.0	0.0	1.0	1.0	4_Auxillary_Motors
AHU-02 SF Air Systems Energy AHU1_SF_VFD_PWR(kWh)						0.0	0.0	1.0	1.0	4_Auxillary_Motors

d) Join the model coefficients from step2 output to the EC+metadata

OUTPUT OF STEP4 = dataframe with EC sensor ID fields, selected EC features, model coefficients

	All of the encoded/scaled columns that passed feature selection									**NRCan labels**	**These are all the coefficients from the Ridge Regression models**			
uniqueId	**mean**	**max**	**min**	**std**	**update_rate**	**isGas_yes_gas**	**energy_no_energy**	**energy_yes_energy**	**equipRef_Air_Equip**	**ALEX-NRCanLabelGuess**	**0**	**1**	**...**	**19**
AHU-01 SF Air Systems Energy AHU1_SF_VFD_PWR(kWh)						0.0	0.0	1.0	1.0	4_Auxillary_Motors				
AHU-02 SF Air Systems Energy AHU1_SF_VFD_PWR(kWh)						0.0	0.0	1.0	1.0	4_Auxillary_Motors				

4) Classification model

a) Run classification model on output from step 4

b) ?

c) PROFIT!

OUTPUT OF STEP5 = dataframe with EC sensor ID fields and end-use group

These are all the sensorID fields (i.e. combined they uniquely identify an EC sensor)						**Our awesome final results**
groupRef	***equipRef***	***siteRef***	***typeRef***	***navName***	***unit***	**EndUse**
Pharmacy Air Systems	*AHU-20*	*Pharmacy*	*AHU20_SF_VFD_PWR(kWh)_TL*	*Energy*	_	Space Heating
Pharmacy	*Cooling*	*Pharmacy*	*CHWP_P5A_VFD*	*5*	_	Space

Hydronic Systems	*Plant P-5A*		*_PWR(kWh)_TL*			Cooling
...	...	...	...	...	...	...
Pharmacy Utilities	*AHU-01 SF*	*Pharmacy*	*AHU1_SF_VFD_ PWR(kWh)*	*Energy*	*kWh*	Space Heating