Interesting Findings and Lessons Learned

Observations on the Data

- There doesn't appear to be data before 2020-01-07
- Either different items are recording at different rates, or the data we have isn't sufficient to uniquely identify everything

Technical References

Simple intro on how to query in InfluxDB (All the official documentation is useful)
https://docs.influxdata.com/influxdb/v1.8/concepts/crosswalk/

- If you download influx from the above website (linked on left-hand pane) you can use a command line to connect to the server and issue commands via influxQL. Note that the address for connecting is listed in the Jupyter notebooks that UDL provided but it used quotes in the wrong places. Instead it should look like:

  ```
  influx -host 206.12.92.81 -port 8086 -username public -password public -database SKYSPARK
  ```

InfluxQL allows for Regular Expression (RegEx) pattern matching of strings. The documentation has some examples here:
https://docs.influxdata.com/influxdb/v1.8/query_language/explore-data/#regular-expressions

The specific syntax for the flavour of RegEx that InfluxQL uses can be here:
https://golang.org/pkg/regexp/syntax/

How Others Have Addressed

Papers

A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm (Relevant Topics: Feature mapping and analysis)

- https://www-sciencedirect-com.ezproxy.library.ubc.ca/science/article/pii/S0378778814006720?via%3Dihub
- This article discusses clustering buildings based on multiple criteria in order to provide a better energy rating than the Energy Star rating

Interesting Findings and Lessons Learned

- I think this is essentially what we need to do, just clustering on end uses rather than buildings

The identification of chemicals using clustering and extrapolation from an external database for electronic nose sensors (Relevant Topics: Feature mapping and analysis)

- https://www.jyi.org/2008-september/2017/11/5/the-identification-of-chemicals-using-clustering-and-extrapolation-from-an-external-database-for-electronic-nose-sensors
- This article discusses clustering chemicals based on scent, it also discusses the following:
  - Applying density based clustering instead of k-means
  - PCA to help improve the effectiveness of clustering
  - How they used PCA to on new data to either add it to an existing cluster, or start a new cluster if needed (this could be useful for making this easily expandable to other buildings)

Clustering in extreme learning machine feature space: (Relevant Topics: Feature mapping)

- http://www.intsci.ac.cn/users/jinxin/Mypapers/ELM-Neurocomputing-2013.pdf
  - Quick rundown on what an Extreme Learning Machine is: https://en.wikipedia.org/wiki/Extreme_learning_machine
  - Tutorial on how to code up an ELM: https://www.kaggle.com/robertbm/extreme-learning-machine-example
  - A package that can be used (maybe instead of self coding?): https://elm.readthedocs.io/en/latest/elm.html
- This article discusses using a neural network called an Extreme Learning Machine (ELM) to perform feature mapping before clustering
- It shows that performing ELM feature mapping or performing Non Negative Matrix Factorization (NMF) within the ELM feature space before clustering can significantly improve clustering accuracy

Instrumentation Data Analysis of a Large Dam: (Relevant Topics: Analysis)

- https://www.researchgate.net/publication/224294574_Data_Mining_Applied_to_the_Instrumentation_Data_Analysis_of_a_Large_Dam
- This article discusses how they analyzed the extensometers of a dam, the Methodology section has some interesting discussion on Factor Analysis, PCA, and Clustering

Cascaded Hidden Space Feature Mapping, Fuzzy Clustering, and Nonlinear Switching Regression on Large Datasets: (Relevant Topics: Fuzzy Clustering and potential feature mapping)

Interesting Findings and Lessons Learned

- https://ieeexplore-ieee-org.ezproxy.library.ubc.ca/stamp/stamp.jsp?tp=&arnumber=7886302
- I haven't fully completed reading this article, but it claims to outperform ELM and be more computationally efficient
- Can't find any Python packages doing this so probably not the ideal solution

Exploration of Missing/Problematic Data

Colour scheme:

- Nothing of Concern
- Likely not a problem, only address if causing problems or have time
- Issues that we will need to address

Items investigated:

- **Missing Information:** (Nothing of Concern)
  - It looks like the only time there is missing information in any of the columns it is when the groupRef=='weatherRef' which is expected
    - Meaning that it doesn't look like we have to deal with columns missing values
- **value column:** (probably ok, can just ignore these two sensors)
  - Only found 2 instruments with inconsistent datatypes
    - EF_06_FAIL_AL_BV
    - EF_11_FAIL_AL_BV
  - These two instruments don't look like they exist anymore (I haven't found any measurements from them since February)
  - There are some instruments that only ever have a value of 0 (Do these exist)
- **unit column:** (issues that we will need to addressed)
  - Found several unique instruments that have multiple different unit values (across all buildings)
    - Went through each one and identified what I think the units should be (Claudia double checked the ones I wasn't sure of)
    - We can probably just do a lookup table
  - Joined the SkySpark metadata csv to the queried data and found over 800 unique instruments in the Pharmacy building alone that have different units between the two databases (Just Pharmacy)
    - Need to ask Jiachen which one should be considered more reliable for units
  - There are only 2 observations with the unit degrees days C (does this exist? Should it just be C?)
- **navName column:** (Probably don't need to address, could make a stretch goal)

Interesting Findings and Lessons Learned

    - Found 21 instruments that have inconsistent navNames but look like they are the same instrument
        - None of these are in the Pharmacy building, and all look like minor data cleaning issues (probably something that will be caught with the data cleaning that is already being done)
- **equipRef, typeRef, siteRef, and groupRef columns:** (Nothing of Concern)
    - Nothing of concern identified
    - “Rusty Hat” and “Totem CSNM” exist in siteRef but do not exist in the original SkySpark database. Likely not an issue because they do not have any recent data.