

[QuadReal Property Group](https://www.quadreal.com/)

# Indoor Air Quality Sensor Data

Indoor Air Quality (IAQ) sensors provide various environmental measurements for QuadReal Properties.

## 

# Problem

From time to time, sensors may not report data to the cloud. This can happen for many, many reasons, including but not limited to:

* Power loss
* Misconfiguration
* Physical tampering
* Mechanical or Electrical fault
* Network or Communication issues

The primary objective of this challenge is to predict the missing IAQ sensor data from outages. An accurate model of the sensors can be used to:

* Fill gaps of missing values from outages
* Anomaly detection to detect and alert relevant operators

# Data

## Sensors Information

The sensor datasheet can be found here:

* <http://www.atlasenleo.com/download/support/ATLASEN_AT04_datasheet_en_v9.pdf>

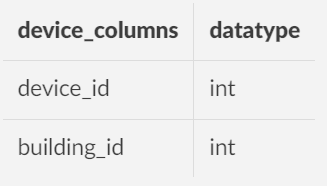
This document includes the meanings and sensitivity of the various outputs, among other information.

Further, acceptable concentrations for airborne substances can be found in the acceptable\_concentrations.png.

## Data Format

There are two datasets:

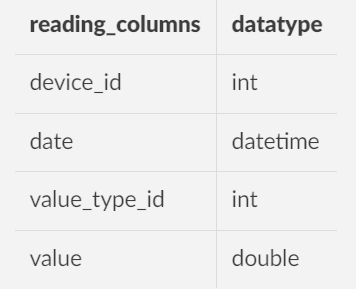
1. devices.csv. This dataset contains two columns, device\_id and building\_id, both integer values.



These are the unique identifiers representing IAQ sensors, and buildings.

1. sampled\_readings.csv.

This dataset contains the following columns



* device\_id is the device identifier from devices.csv
* date is a datetime value
* value\_type\_id is an integer value indicating the value represented

The following table provides the value types.



* Finally, value is the numerical value for the data, which will be the output of the model.

# Additional Information

The following information may be helpful:

* **Book time with us to ask any questions regarding the challenge:**
  + <https://outlook.office365.com/owa/calendar/QuadRealOfficeHour@quadreal.com/bookings/s/PtuxQhbSqkGUL8_LnvB6PQ2>
* It will be useful to add additional features to the data.
  + For instance, buildings have Heating, Ventilation, and Air Conditioning (HVAC) systems that run on schedules. They can also exhibit strong seasonal trends, changes in air quality on the basis of occupancy, etc.
  + Adding features derived from the timestamp (day\_of\_week, hour\_of\_day, etc.) may be of use.
* The building\_id identifier is provided to identify sensors within the same building.
  + While not co-located, incorporating information from other sensors within the same building may improve the accuracy of gap filled data.
  + You may also consider using this information to categorize different buildings based on their behavior.
  + How to use (and whether to use) building\_id is entirely to the discretion of the analyst.

## 

# Expected Outcome

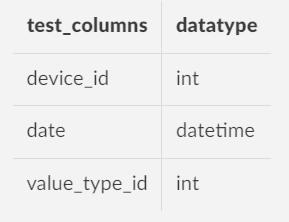
The expected outcome of this project is to design a comprehensive model(s) that can be effectively transitioned into production. Output of the model should help:

* **Primary Objective**: Predicting missing values in Indoor Air Quality (IAQ) sensor data for occasions of outage
* **Secondary Objective**: Detecting anomalies with dynamic threshold that adapts to trends and seasonalities

## Judging

A portion of the data has been withheld as a test set for evaluating the predictive model, provided in test.csv file.

For the test data, the IAQ sensor value must be predicted. Mean squared error will be the judging metric of the challenge.



## What to submit?

1. **submission.csv** with predicted values for the test set. The submission should include the predicted values corresponding to the given device\_id, date, and value\_type\_id. The predicted values should be placed in a separate column, value. Correctness of prediction will account for 70% of the final score.
2. **Notebook** with full process including data pre-processing to model training and inference on test dataset. The notebook will be judged according to [the General Judging criteria](https://drive.google.com/file/d/19jba2OUGWGsr-2PPRx3px4FFOE42OIGf/view?usp=drive_link), and will account for 30% of the final score.

(Optional) If you’d like to be eligible for the $500 Interac Most Innovative Solution Award, check out the details in our [participant guide](https://docs.google.com/document/d/1TrUlcToyEzFcHDB8dyXYGxyXRe3BdE_VBlhf-V5zfps/edit#heading=h.gfvxp7dgpqm2)!

## Where to submit?

* Compile all files included of your project into a **ZIP file** and include it as part of your submission on [our Devpost platform](https://cxc-powered-by-ey.devpost.com/).
  + If the ZIP file exceeds Devpost size limit (>35 MB), upload it to a Google Drive or Github repository and include a link in your submission.
* One submission per team is needed, but include all team member info in the submission.

# Benefits to Students

Engaging in this project provides students with a unique opportunity to tackle real-world challenges in the domain of environmental data analysis. The benefits include:

* **Practical Application of Data Science Techniques**: Students will gain hands-on experience applying data science techniques to address issues related to IAQ sensor data. This includes preprocessing, feature engineering, and ML predictive modeling.
* **Understanding IOT Sensor Time Series Data Challenges**: Dealing with real sensor data introduces students to challenges such as missing values, anomalies, and data gaps, enhancing their ability to handle practical data scenarios.
* **Application of Feature Engineering**: Incorporating additional features derived from timestamp data and exploring the use of building identifiers allows students to apply feature engineering strategies to improve model accuracy.

Overall, participation in this project offers students a rich learning experience, aligning with the industry demand for data scientists capable of addressing complex challenges in environmental data analysis.

# Prizes

* 1st Place: $500
* 2nd Place: $250
* 3rd Place: $100

Finalists will be invited to chat with company representatives to elaborate on how they approached the problem.