Summative Portfolio: L4

Steve Crawshaw

12/29/22

Table of Contents

# 1. Introduction

My name is Steve Crawshaw. I am a project manager for Bristol City Council and I have undertaken the L4 Apprenticeship (Data Analyst) to improve my knowledge and skill in data analytics with a view to undertaking the L7 apprenticeship and achieving a role as a data scientist.

I have worked for Bristol City Council (BCC) since 1998 in essentially the same role, although I was seconded to another organisation between 2013 and 2016. My main role now is managing a network of air quality monitors and the data that they generate. This will be the subject on which I will focus for the summative portfolio.

# 2. Employer: Bristol City Council

[Bristol City Council](https://en.wikipedia.org/wiki/Bristol_City_Council) is a large unitary local authority in the South West of England.

## 2.1 Goals, Vision and Values

Bristol City Council’s [Corporate Strategy](https://www.bristol.gov.uk/policies-plans-strategies/corporate-strategy) outlines a vision of driving an inclusive, sustainable and healthy city of hope and aspiration where everyone can share the city’s success. It also describes the activities required by law.

The Corporate Strategy’s main priorities are informed by 5 key principles.

* Development and delivery
* Environmental sustainability
* Equality and inclusion
* Resilience
* World-class employment

It’s also arranged around 7 main themes:

* **Children and young people:** A city where every child belongs and every child gets the best start in life, whatever circumstances they were born into.
* **Economy and skills:** Economic growth that builds inclusive and resilient communities, decarbonises the city and offers equity of opportunity.
* **Environment and sustainability:** Decarbonise the city, support the \* recovery of nature and lead a just transition to a low-carbon future.
* **Health, care and wellbeing:** Tackle health inequalities to help people stay healthier and happier throughout their lives.
* **Homes and communities:** Healthy, resilient, and inclusive neighbourhoods with fair access to decent, affordable homes.
* **Transport and connectivity:** A more efficient, sustainable, and inclusive connection of people to people, people to jobs and people to opportunity.
* **Effective development organisation:** From city government to city governance: creating a focussed council that empowers individuals, communities, and partners to flourish and lead.

Bristol City Council is currently a mayoral - led authority. The current Mayor, [Marvin Rees](https://en.wikipedia.org/wiki/Marvin_Rees) has set out values for the organisation as shown below.

|  |
| --- |
| Bristol City Council’s values |

## 2.2 How BCC Uses Data

BCC is a large and complex organisation dealing with a wide range of functions, from managing highways and planning applications to looking after vulnerable people. It follows that multiple systems and approaches exist for managing data across the organisation, many of which have evolved over time and have not been centrally planned or managed.

A recently published [“Data, Insight and Information Strategy”](https://democracy.bristol.gov.uk/documents/s64321/DII%20Strategy%20Final.pdf) sets out the strategic direction and objectives for the council in this area:

 The strategy aims to deliver these objectives by developing a single analytics team to deliver insights across the organisation and by consolidating the disparate and distributed datasets across the council into a corporate data lake with analysis being done on one data analytics platform (Power BI). Several projects are under way to deliver this change.

## 2.3 Data Architecture

No specific document exists about BCC’s data architecture, however there is an information asset taxonomy, which is summarised in the diagram below. This was developed in 2017 by an officer who has now left BCC so the information may be somewhat outdated.

|  |
| --- |
| Information Asset Taxonomy |

Much of this data architecture is not of direct relevance to my work as it pertains to other departments and functions. In my own work area, the main data architecture is summarised as follows:

* 182 passive samplers (Diffusion tubes) providing monthly concentrations of pollutants (NO2)
* MS Access database to hold diffusion tube and monitoring network meta data on site
* A network of 8 real time air quality monitors (analysers)
* 4G Telemetry to connect these devices to:
* A proprietary communications software suite (Envista Commcentre)
* A database client and analytics program (Envista ARM)
* A SQL server database hosted on Azure
* An open data (OD) platform provided by Opendatasoft
* FME and FME Server processes to Extract, Transform and Load (ETL) air quality data to the open data platform
* Dashboards, visualisation and analytics delivered through the OD platform
* Bespoke reporting and ETL pipelines delivered through R, sourcing data through the OD portal and the Envista database

## 2.4 Security Standards and Policies

The over - arching Information Governance Framework outlines roles and responsibilities, policies and procedures, along with best practice and standards for managing the Council’s information assets. This has been developed to take account of the standards set by external organisations, such as the NHS in respect of the transition of Public Health to the Council and the requirements of the Public Sector Network (PSN) Code of Connection (CoCo).

The framework consists of the following areas:

1. IGF Principles
2. The Information Governance Strategy.
3. Appropriate Information Governance Responsibilities.
4. Information Asset Ownership
5. An Information Governance Structure.
6. Effective Information Governance policies and procedures.
7. An Information Asset Register
8. An Information Risk Register
9. Information Governance communication and training.

There are a number of policies and procedures in the framework which deliver the outcomes of the Information Governance Strategy, including:

1. Instant Messaging Policy
2. Acceptable Use Policy
3. Training, Awareness and Development Procedure
4. Agile or Teleworking Policy
5. Logical Access Control Policy
6. Physical Access Control Policy
7. Information Security Incident Reporting Policy
8. Subject Access Request Policy

All the policies and procedures are hosted on a “metacompliance” platform, which manages access and control of the policies and ensures that relevant staff have read and agreed to the policies.

# 3. My Role: Air Quality Project Manager

My official job title is “Project Manager”. This is currently under review, partly because the extent of actual project management activity is quite limited. The majority of my time is spent managing a network of air quality monitors and the data that arises from the network.

## 3.1 Key Deliverables

The deliverables for which I am responsible are summarised as follows:

1. Data capture rates exceeding 85% for continuous monitoring, 75% for passive.
2. Monthly data cleaning (ratification) of continuous data.
3. Annual reporting of air quality data.
4. Calculation of Key Performance Indicators (KPIs) for air quality.
5. Developing or revising KPIs as necessary.
6. Ad hoc analysis and summaries of air quality data to support other programmes.
7. Ensuring all relevant air quality data are published on the Open Data Portal.
8. Ensuring all relevant technical guidance is followed in relation to air quality management.
9. Responding to relevant requests from stakeholders for air quality data and analysis.
10. Delivery of specific projects such as “Slow the Smoke”

## 3.2 Key Skills and Knowledge

The key skills and knowledge for my role are as follows:

### 3.2.1 Knowledge

* A good understanding of the legal framework for air quality management in the UK.
* Familiarity with the relevant technical guidance on assessing air quality.
* Knowledge of air quality policy and interactions with other domains like transport.
* An understanding of the development management (land use planning) process in the UK including Environmental Impact Assessment (EIA).
* An understanding of the principles of open data and legal framework for public access to data.

### 3.2.2 Skills

* Processing and analysing medium sized (up to 10 million observations) data sets
* SQL (SQL server, ODSSQL and MS Access)
* Excel
* HTML, Javascript and CSS for web development of open data products
* R - packages openair, openaq, sf, timetk, fastverse, deweather and tidyverse are relevant to the air quality domain
* [FME](https://www.safe.com/) and FME Server for automating web services and data integration
* Time series analysis
* Network telemetry: IP, analogue, wireless, Teltonika RMS
* Technical report writing and comprehension of technical reports relating to the domain
* Communication skills - ability to report technical information to non - specialists
* Project management for small and medium sized projects
* Technical skills related to installation, maintenance and quality control of air monitoring instruments
* Negotiation skills for contract management and securing outcomes in the planning process

## 3.3 Strengths and Weaknesses

### 3.3.1 Strengths

* Long experience in the domain and a good level of skills and knowledge
* Strong work ethic
* Motivation to deliver and improve services and air quality
* Wide range of contacts internally and externally
* Collaborative approach to working

### 3.3.2 Weaknesses

* Wide, rather than deep data skill set
* Limited understanding of statistical theory and advanced analysis
* No exposure to team environment of other data analysts so lacking peer support
* Isolated from corporate data analytics functions
* Limited employer incentive to improve skills

## 3.4 Areas for Improvement

* Improved understanding of statistical theory and learning
* Power BI (corporate data analytics tool)
* Azure (corporate data platform)
* Python and PANDAS for comprehensive tooling of data science operations

# 4. Portfolio Projects

I have identified four projects which align with business and apprenticeship progress review objectives. They are focussed on creating data products which will enhance business processes and increase confidence in our data and analysis. They are summarised below.

Objective 1 Objective 2

Objective 3 Objective 4

I am able to work on these projects concurrently and I believe they are all achievable by the end of the apprenticeship.

## 4.1 Project 1: A data processing pipeline for the statutory reporting of air quality data

### 4.1.1 Introduction

Bristol City Council monitors air quality across the city and reports data to government under the Local Air Quality Management (LAQM) regime. This is a statutory duty under the Environment Act (1995) for councils to manage air quality to ensure the quality of the air meets legal limits for regulated pollutants.

In addition to this duty, there is also a requirement to provide monitoring and evaluation of the operation of a Clean Air Zone (CAZ) in Bristol. The CAZ is the main mechanism by which traffic pollution (nitrogen dioxide or NO2) will be reduced to comply with the legal limits.

The two reporting regimes are similar, but different in some respects. For LAQM, the reporting is directly to the government and takes the form of an Annual Status Report (ASR) to be submitted by June 30th of each year. In this ASR are summary tables of air quality data from all our monitoring which are required in a format determined by a template document. For CAZ reporting BCC is required to submit a spreadsheet of summary data in a set format to the Institute for Transport Studies at Leeds, who are the body that is conducting research on all the CAZs in England on behalf of the government.

The aim of this project is to develop a process for reporting for both of these purposes that limits the need for manual data entry.

### 4.1.2 The LAQM regime

Local authorities are required to review and assess air quality in their areas and report to government annually. Where concentrations of regulated pollutants exceed legal limits “Air Quality Objectives” (AQO), councils must declare an air quality management area (AQMA) and publish a plan for dealing with the breaches of the AQOs.

Several urban local authorities experience breaches in the AQO for nitrogen dioxide, a traffic pollutant. The AQO is set at 40 gm-3 measured as annual mean. In Bristol, this is the only pollutant for which exceedences are measured.

The relevant guidance for LAQM, [LAQM.TG(22)](https://laqm.defra.gov.uk/wp-content/uploads/2022/08/LAQM-TG22-August-22-v1.0.pdf) prescribes the monitoring regime that must be followed for each pollutant, including how data is monitored, processed and reported. The process used in this project will adhere to the requirements of LAQM.TG(22).

Further prescriptive guidance is provided by Defra for air quality reporting in the form of templates for the [report](https://laqm.defra.gov.uk/wp-content/uploads/2022/05/ASR_Template_England_2022_v1.0.docx) itself and for the [tables](https://laqm.defra.gov.uk/wp-content/uploads/2022/05/ASR_Table_Template_England_2022_v1.0.xlsb) that should be included in the report.

For this project, the data process will focus only on tables prefixed “A” in the spreadsheet referenced. These are the tables that contain monitoring data. Other tables in the report require manual data entry as they need data that is sourced from other officers and departments in the council so will not be addressed in this pipeline.

In addition to these tables, there is a requirement to include thematic maps in the report which show concentrations of pollutants at monitoring sites. The process will also produce spatial datasets which can be used in a GIS to plot the maps or will produce graphical output directly for inclusion in the report as image files.

### 4.1.3 CAZ Reporting

The purpose of the CAZ is to reduce concentrations of one pollutant (NO2) to legal levels in the shortest time possible. Therefore the reporting is limited to NO2. However, the reporting on this single pollutant requires greater granularity than the LAQM reporting. This is partly because advanced techniques for analysing changes in time series air quality data will be used to assess the impact of a specific policy intervention (the CAZ) and therefore the hourly time series data need to be reported, not just the summary statistics. This analysis will be done by ITS using the [`AQEval’ R package](https://cran.r-project.org/web/packages/AQEval/index.html) and the data for Bristol will be supplied by providing access to the open data portal [dataset for continuous air quality data](https://opendata.bristol.gov.uk/explore/dataset/air-quality-data-continuous/information/?disjunctive.location).

The reporting for diffusion tube data consists of providing data in a pre - formatted template quarterly. The data is a combination of meta - data for the measurement site, e.g. location, height, distance from a road and the “raw” monthly concentrations from the diffusion tube. The measurement technique for diffusion tubes and continuous analysers is described later on.

### 4.1.4 Proposed analysis and processing techniques

#### Summary

The overall process will be controlled using the [targets R package](https://books.ropensci.org/targets/). This is a “Make” - like pipeline tool for Statistics and data science in R to help maintain a reproducible workflow. Targets learns how the pipeline fits together, skips costly runtime for tasks that are already up to date, runs only the necessary computation, supports implicit parallel computing, abstracts files as R objects, and shows tangible evidence that the results match the underlying code and data.

Each operation will be assigned to a function. Targets will ensure that the functions run in the right order, and only when all pre - conditions are satisfied.

#### Data Sources

The data needed for the analysis are in four separate repositories:

1. A corporate Azure SQL server database hosting BCC’s continuous air quality data
2. A Microsoft Access database for the diffusion tube data and monitoring site meta data
3. The council’s [open data portal](https://opendata.bristol.gov.uk/explore/?sort=modified)
4. UK Government’s [Air Information Resource](https://uk-air.defra.gov.uk/data/) which holds data for the air monitoring sites run by government as part of the Automated Urban and Rural Network (AURN). There are two of these sites in Bristol.

Three of these repositories will be used in the analysis. The open data portal will not be used as it is planned to cease operation of the contract with the provider in 2022. This is somewhat sub optimal as the open data portal contains all the relevant data needed for the analysis and data is easily accessed using a powerful and well documented API.

The data from databases will be accessed using a combination of the DBI, dplyr and odbc packages. For the data that is hosted on the government’s website, the openair package provides a helpful function called importAURN. Once the data is available in the R session, they will be manipulated using base R and tools from the tidyverse metapackage to produce dataframes to populate the tables in the required format. The dataframes will be exported as csv files and also pasted directly into the spreadsheet template.

For the generation of spatial output, the sf and osm packages will be used to create shapefiles for use in ArcGIS and output graphics as png files.

The targets package will control the processing pipeline and ensure efficient operation of the data manipulation steps.

## 4.2 Project 2: To develop a process for monthly diagnostics and QA reporting from air monitors and telemetry devices

## 4.3 Project 4: Comparing performance of low – cost sensors with reference method instruments

### 4.3.1 Introduction

The Slow the Smoke (StS) project is a Citizen Science project funded by Defra’s Air Quality Grant. It aims to test engagement approaches based on citizen monitoring of air quality with a particular focus on emissions from domestic solid fuel burning (wood burning stoves etc.).

Bristol City Council leads the project and I am the project manager. We have two partners, the University of the West of England (UWE), and Knowle West Media Centre (KWMC). UWE lead on the technical aspects of air quality and writing the final report. KWMC lead on the outreach activities.

The combination of citizen science, community engagement and behavioural surveys is intended to identify effective approaches to influencing behaviour in relation to domestic emissions to air.

Ten citizen scientists have self selected in the study area which is a ward in the city centre called Ashley. This ward was selected because we have evidence that there is a higher level of solid fuel burning than average.

The citizen scientists have each been given a “low cost” air sensor that monitors particulate matter (PM). PM is fine dust in the air, including smoke. There are two fractions of PM that are important for health; PM10 (aerodynamic diameter < 10 μ) and PM2.5 (aerodynamic diameter < 10 μ). The devices deployed monitor both of these fractions using a light - scattering approach, where the diffusion of laser light is a function of the concentration of PM in the sampled air.

Because these devices do not directly measure concentration of PM, but use a proxy measure, they are not as accurate as “approved” measuring instruments used to assess compliance with air quality objectives. It is therefore necessary to attempt to characterise the performance of these devices in relation to approved or “reference method” devices by co - locating the low cost sensors with reference method instruments and comparing measurements.

### 4.3.2 Co - location Study

There are three monitoring sites which measure PM in Bristol. A map and summary data for the sites are shown below.

| Co - location: Potential Sites |
| --- |
| Site description | Pollutants |
| AURN St Pauls - 452 |
| The monitoring station is located within a self contained air conditioned unit within the car park of a day nursery. The monitoring station is approximately 30 metres south east of Wilder Street which is a lightly trafficked urban back street. The surrounding area is primarily residential with some commercial premises in the immediate vicinity. | PM2.5, PM10 |
| Parson Street School - 215 |
| The Parson Street School monitoring site has been operating since 2002 and was selected as a roadside site that represents residential exposure to relatively high concentrations of traffic generated pollution. The sample inlet is approximately 1 metre from ground level and four metres from the kerb of Bedminster Road where traffic queues for the traffic lights | PM2.5 |
| Temple Way - 500 |
| Roadside monitoring station beside multi-lane section of the A4044 (Temple Way). This is a major route in and out of Bristol city centre. | PM10 |

The co - location study used two of these sites; Parson Street School and Temple Way. This is because these sites are both operated by the council. Using the AURN St Pauls site would have required permissions from the Environment Agency which would have been a time consuming and uncertain process.

The [airrohr SDS011 fine dust sensors](https://sensor.community/en/sensors/airrohr/) require a wifi signal in order to push data to a server. The Bristol City Council sites did not previously have wifi access available. The telemetry at our air monitoring sites was a combination of 3G modems and analogue land lines in 2021. In order to accommodate the co - location study and also for the purposes of virtualising our data collection machine, I procured, configured and installed Teltonika RUT950 4G LTE routers at all of our monitoring sites. This enabled 4G TCP/IP access to the data loggers or instruments at all of the sites, and also provided a wifi hotspot to enable the SDS011 sensors to send data.

The physical installation of the SDS011 at both sites was complete in early May 2022.

#### Colocation Study: Concept

The aim of the colocation study is to compare the performance of the low cost sensors with the performance of reference method instruments measuring the same pollutant. The method of implementing this comparison was to collect hourly data for the two co - located devices and establish the linearity of the response using a linear model to report coefficients and r2.

Within this study it was not possible to compare the responses of multiple low cost sensors with each other as there was not a budget to purchase additional devices for this purpose.

#### Reference Method Equipment

Continuous Ambient Air Quality Monitoring Systems (CAMS) are certified by the Environment Agency under their MCERTS scheme. This certifies the quality regimes and equipment for environmental permit holders. Local authorities are required to use MCERTS (or equivalent) equipment for monitoring air quality under the LAQM (Local Air Quality Management) regime. The certification and approval process ensures that the measurement standard provides data of an acceptable quality. In addition to the specification of the equipment, a local authority is required to adhere to the calibration and maintenance requirements for the equipment as set out in the relevant guidance [LAQM.TG(16)](https://laqm.defra.gov.uk/documents/LAQM-TG16-April-21-v1.pdf) and the [Local Site Operator (LSO) manual](https://uk-air.defra.gov.uk/assets/documents/reports/empire/lsoman/lsoman.html) for sites that are part of, or affiliated to the national AURN (Automated Urban and Rural Network).

The equipment used in this colocation study is the Met One BAM 1020 Continuous Particulate Monitor, hereafter referred to as “BAM 1020”. The instrument works by drawing a sample of air through a filter tape every hour. The deposited PM is then exposed to a source of radiactive Carbon 14 on one side of the filter tape. A beta radiation detector is on the other side of the tape and measures the attenuation of the beta radiation through the sampled filter. The attenuation of the beta radiation is a function of the deposited PM mass on the filter tape. Because the flow rate of the sampled air is known, the concentration in μgm-3 can be calculated. Hourly concentrations are recorded, either on an internal or external data logger. These data are regularly polled by a central telemetry system.

#### Colocation Sites: Temple Way

The [Temple Way site](https://opendata.bristol.gov.uk/pages/aqcontinuoussites/?q=siteid:500) is affiliated to the national monitoring network. This means that the site is owned by Bristol City Council, but the management of the data and the oversight of the QA regime is done by the Environment Agency’s contractors. PM10 and NOx are measured at this site and summary metadata are provided below.

#### Colocation Sites: Parson Street School

Oxides of Nitrogen (NOx) have been measured at [Parson Street School](https://opendata.bristol.gov.uk/pages/aqcontinuoussites/?q=siteid:215) for many years. The enclosure is close to the roadside of a busy, queuing road and represents exposure of schoolchildren and school staff. In recognition of the need to understand exposure to PM2.5 at a roadside site the monitoring station was updated with a BAM 1020 in 2021. The BAM 1020 when configured for monitoring PM2.5 includes an additional particle size cut off filter and also incorporates a heated inlet to drive off volatile compounds from the sampled air. The summary metadata for the site is shown below.

#### Colocation Configuration

The low cost sensors were co - located inside the cages of the monitoring sites. Parson Street was installed on 29th March 2022 and Bristol Temple Way was installed on 1st May 2022. The photographs below show the detail of the colocated devices at each monitoring site.

|  |  |  |
| --- | --- | --- |
| |  | | --- | | Temple Way | |  |

|  |  |  |
| --- | --- | --- |
| |  | | --- | | Parson Street | | Co - located instruments |

#### Data Sources

Data from one low cost sensor and from both reference instruments are published in near real time on the council’s open data portal. The BAM 1020 data are available through the [air-quality-data-continuous](https://opendata.bristol.gov.uk/explore/dataset/air-quality-data-continuous/information/?disjunctive.location) dataset and the Parson Street data are available through the [luftdaten\_pm\_bristol](https://opendata.bristol.gov.uk/explore/dataset/luftdaten_pm_bristol/table/?disjunctive.sensor_id&q=sensor_id+%3D+71552&sort=date&location=22,51.43267,-2.60496&basemap=jawg.streets) dataset. This is a dataset that is a geographical subset of the sensor.community [archive](https://archive.sensor.community/) focussed on Bristol. In addition, data are aggregated to give an hourly mean value for both PM10 and PM2.5. Custom functions were written in R to access these files and import the data as data frames. The data for the low cost sensor at Temple Way are not published on the council’s open data portal. This is because the sensor did not register properly on the sensor.community portal. The data are however available through a [combination of online](https://api-rrd.madavi.de/csvfiles.php?sensor=esp8266-6496445) .csv and .zip files.

#### Data Processing Pipeline

Initially the data processing pipeline was split into three R scripts, one for retrieving and preparing the data, one for plotting, and one for modelling the data. However, while developing this approach, it became apparent that there were significant problems with this method. Tracking the operations across three different scripts became confusing, and testing and debugging was sub - optimal and resulted in errors when running the scripts if they weren’t called in order. Researching this issue identified a possible solution, which was eventually used for this project. The [targets](https://books.ropensci.org/targets/) R package by Will Landau is a “Make” - like pipeline tool for data science in R. It helps to maintain a reproducible workflow, minimising repetition, running only necessary code, and forces a functional approach to code and pipeline development.

##### Retrieving and Preparing data for Modelling

Data for the two BAM 1020 instruments and the Parson Street low cost sensor are retrieved using a purpose built R function which uses endpoints from the [Opendatasoft API](https://opendata.bristol.gov.uk/api/v2/console).

For the Temple Way sensor, it was not possible to register this on the sensor.community website. A different approach was developed to retrieve the data, which is available through the [Madavi API](https://api-rrd.madavi.de/csvfiles.php?sensor=esp8266-6496445) as a combination of csv and zip files. This operation entails retrieving a number of files in different formats and joining the data into one dataframe.

Once data are retrieved from the online repositories, they are combined into a nested dataframe (or tibble in the tidyverse vernacular) with functions from the purrr package such that columns are created that can be used in the subsequent modelling process. The resulting nested dataframe is a 2 row, four column tibble grouped by the site ID with list columns containing the raw hourly data, the prepared source data for the model, and a pivoted (wide) dataset used in the plotting and modelling functions.

The approach of nesting the model data within a tibble enables a concise iterative approach to developing the modelling pipeline and means that the modelling process itself generates a tibble with list columns that contain the output of the regression model, as well as plots which explain the data, and model output.

##### Plotting Functions

Functions were developed to create a range of visualisations of both the raw and processed data and model outputs.

1. plot\_drift() is a function based on the ggplot2 library which takes the prepared model data tibble and a site ID and returns a ggplot object (plot) showing how the pollutant concentrations from the low cost sensor diverge from the reference method measurements over time. The plot uses the geom\_smooth() geom with the smoothing set to “gam” in order to show a smoothed trace.
2. plot\_scatter() is another ggplot based function which generates a scatter plot for co - located instruments. It also incorporates functions from the ggside library (extension for ggplot2) which enables side - plots for the scatter plot showing the distribution of data points for each axis. Further it includes stat\_cor() and stat\_regline\_equation() from the ggpubr library to show correlation statistics and the equation derived from the linear model on the chart. This function therefore provides much useful information about the relationship between the two datasets on a single chart.
3. prep\_timeplot() is a helper function that takes the prepared data for modelling and processes it for use in another function. The function subsets the data, renames columns, pivots longer and pads the time series such that the data is ready for the plot\_time\_series() function.
4. plot\_time\_series() takes the data from step 3 above and a string representing an interval (e.g. “hour”) and plots a time series chart using the ggplot2 library.
5. Other helper functions which format and save plots.

##### Modelling Functions

The model data tibble previously created is used as input to a function to run a linear model on the data for each site. A script to import custom themes for plots is imported at the top of the \_targets.R file. The make.model.output.tbl() function takes the data tibble and creates a new output tibble with the following features:

1. a list column (model\_obj) which is the model object derived from the lm() function comparing the low - cost (dependent variable) and reference (independent variable) device data for each site and pollutant
2. a list column containing the linear coefficients (intercept and slope) from each model object using the tidymodels package.
3. a list column, perf, which holds a tibble of performance metrics for each site derived by the glance() function from the broom package. Glance accepts a model object and returns a tibble with exactly one row of model summaries. The summaries are typically goodness of fit measures, p-values for hypothesis tests on residuals, or model convergence information.

#### Summary Tables

The performance metrics for each model are extracted and tabulated into a range of formats for display. the gtsummary package produces a nicely formatted html table and the performance package from the easystats suite of statistical tools is used to generate a comprehensive html dashboard for each model showing the performance of the model including several plots.

| name | command |
| --- | --- |
| daily\_csv\_urls | get.daily.csv.urls(data\_root\_url = "https://api-rrd.madavi.de/data\_csv/",   sensor = "esp8266-6496445") |
| parson\_st\_sds\_hr\_tbl | get.parson.st.sds.hr.tbl(start\_date) |
| ref\_tbl | get.ref.tbl(start\_date) |
| zip\_file\_urls | get.zip.file.urls(data\_root\_url = "https://api-rrd.madavi.de/data\_csv/",   sensor = "esp8266-6496445", start\_date) |
| daily\_files\_tbl | get.daily.files.tbl(daily\_csv\_urls) |
| monthly\_files\_tbl | get.monthly.files.tbl(zip\_file\_urls) |
| temple\_way\_sds\_hr\_tbl | make.temple.way.sds.hr.tbl(daily\_files\_tbl, monthly\_files\_tbl) |
| combined\_long\_tbl | make.combined.long.tbl(ref\_tbl, temple\_way\_sds\_hr\_tbl,   parson\_st\_sds\_hr\_tbl) |
| model\_data\_tbl | make.model.data.tbl(combined\_long\_tbl) |
| drift\_plot\_215\_gg | plot.drift.site.gg(model\_data\_tbl, site = 215) |
| model\_output\_tbl | make.model.output.tbl(model\_data\_tbl) |
| timeplot\_tbl | prep.timeplot.tbl(model\_data\_tbl) |
| write\_model\_data\_tbl | write.model.data.tbl(model\_data\_tbl) |
| drift\_plot\_500\_gg | plot.drift.site.gg(model\_data\_tbl, site = 500) |
| save\_drift\_plot\_215\_gg | save.ggplot(drift\_plot\_215\_gg) |
| scatter\_gg\_215 | model\_output\_tbl[model\_output\_tbl$siteid == 215, ]$plot[[1]] |
| scatter\_gg\_500 | model\_output\_tbl[model\_output\_tbl$siteid == 500, ]$plot[[1]] |
| dashboard\_500 | make.dashboard(model\_output\_tbl, siteid = 500) |
| model\_perf\_tbl | make.model.perf.tbl(model\_output\_tbl) |
| dashboard\_215 | make.dashboard(model\_output\_tbl, siteid = 215) |
| model\_perf\_tbl\_gt | make.model.perf.tbl.gt(model\_output\_tbl) |
| sp\_plot\_tbl | prep.sp.plot.tbl(timeplot\_tbl) |
| time\_series\_hour\_gg | plot.time.series.gg(timeplot\_tbl, interval = "hour") |
| time\_series\_day\_gg | plot.time.series.gg(timeplot\_tbl, interval = "day") |
| save\_drift\_plot\_500\_gg | save.ggplot(drift\_plot\_500\_gg) |
| save\_scatter\_gg\_215 | save.ggplot(scatter\_gg\_215) |
| save\_scatter\_gg\_500 | save.ggplot(scatter\_gg\_500) |
| transpose\_model\_perf\_tbl | transpose.model.perf.tbl(model\_perf\_tbl) |
| save\_model\_perf\_tbl\_gt | save.model.perf.tbl.gt(model\_perf\_tbl\_gt) |
| save\_png\_pm25 | save.png.summaryplot(sp\_plot\_tbl, "pm2.5") |
| save\_png\_pm10 | save.png.summaryplot(sp\_plot\_tbl, "pm10") |
| save\_gg\_time\_series\_hour\_gg | save.ggplot(time\_series\_hour\_gg) |
| save\_gg\_time\_series\_day\_gg | save.ggplot(time\_series\_day\_gg) |