# Task 2: Dataset Description

The Retail Sales Index data, collected by the Office of National Statistics (ONS), and Met Office weather data can be accessed here:

* Accessed 12/12/2023: <https://www.metoffice.gov.uk/research/climate/maps-and-data/uk-and-regional-series>
* Accessed 16/12/2023: <https://www.ons.gov.uk/businessindustryandtrade/retailindustry/datasets/retailsalesindexreferencetables/current>

Both datasets are licensed under the Open Government Licence 3.0 (<https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>). This permits users to copy, alter and share the information, so long as a license-link and attribution are provided.

The *MetOffice* provides thorough documentation on its data collection: <https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/haduk-grid> The data recorded at regional weather stations is used to *interpolate* values for a 1km x 1km data grid for the entire UK land surface, using multiple-regression with inverse-distance weighted interpolation (<https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/methods>). Accuracy is monitored through the calculation of root-mean square errors (RMSE) by comparing real and predicted values (<https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/faq#faq1>). While these datasets do undergo rigorous error-checking, the inherent limitations of interpolation may inevitably introduce some error.

The ONS constructs the Retail Sales dataset for every month using the Monthly Business Survey (MBS-RSI). Detailed documentation about the quality checks and methodology can be found here: <https://www.ons.gov.uk/businessindustryandtrade/retailindustry/methodologies/retailsalesindexrsiqmi> As many financial institutions depend on this data, the ONS invests many resources into making this data as reliable as possible. The data is collected from businesses using an electronic survey. Furthermore, the ONS uses “chasing” tactics by trained employees to ensure a high monthly response rate. A system of automated checks is implemented to “flag” figures above a certain threshold.

The RSI dataset provided by the ONS comes in the form of multiple tables in Excel sheets. The *tables* we will use are the *KPSA1*, the seasonally-adjusted **percentage change in retail sales volume** compared to *the same month a year earlier*, and *KPSA3*, which stores percentage change from *the previous to the current month*. Each percentage-change is represented as a floating-point number. This project will evaluate the effect of the weather on these *two* “target” vectors, in order to get an overview of the impact of weather on both long-term and short-term patterns in consumer spending. Using *seasonally-adjusted* data helps to isolate the effect of weather on sales by accounting for the impact of holidays. Each table consists of 421 rows indexed from January 1988 to October 2023, with 10 columns representing sales volume for different retail sectors. For this project, the column named *All retailing excluding automotive fuel* will be selected to exclude impact of the volatility of fuel prices.

The weather data consists of 7 tables each one of which represents a different “feature” or weather parameter. Each *row* in these tables is indexed by a year, while each column represents a month/season. The cells contain the values representing the parameter measured (floating-point numbers). The units used are: degrees Celsius (temperature), number of days (days with frost and more than 1mm precipitation), hours (sunshine), and millimetres (precipitation). Each table comprises a different number of rows/years, ranging from starting in 1836 to 1960 to 2023. However, as we only have data for retail sales volume from January 1988 to October 2023, 421 rows must eventually be extracted from the restructured data.