STAT 946: Case Studies in Data Science

Case Study 1: Air quality in Canada

Andy, Erick, Henry, Jason, Peter

2026-01-10

Contents

1 Datasets

1.1 Air Pollutant Data

1.1.1 Ontario:

The Ontario Air Pollutant Data was obtained from the following URL: https://www.airqualityontario.com/history/index.php.

It is published by the Government of Ontario.[1] We restrict the downloaded data to only contain the particular pollutant that we care about, which is PM2.5.

Each row of the dataset represents an observation by a station with multiple dates from January to December, recorded day by day.

For the columns, the following table demonstrates the meaning of each column.

Feature name	Definition
Station ID	A unique identifier for the station where PM2.5 measurements were recorded. It is an integer object.
Pollutant	Here, it is always assigned to fine particulate matter for PM2.5 measurement. It is a redundant column where all rows have the same value and could be further removed. It is a string object.
Date	The date of the measurements in Ontario time. It is represented in the format YYYY-MM-DD. It should be a date object.
HXX	Measurement of PM2.5 from XX:00 to XX:59, where XX represents 01 to 24. The value 24 is an exception, where H24 corresponds to 00:00 to 00:59. It should be a datetime object.

For this dataset, it is not difficult to obtain year-by-year data, as it is easy to download by year and each file contains all the dates for that year. However, if we want the full dataset from 2010 to 2024, further merging of the datasets is required.

1.1.2 British Columbia:

1.1.3 Alberta:

The Alberta Air Pollutant Data was obtained from the Alberta Data Management Platform, available at: https://datamanagementplatform.alberta.ca/Ambient

The data is operated and reported by community-based, multi-stakeholder organizations (commonly referred to as airsheds) as well as the Government of Alberta. These organizations are responsible for monitoring ambient air quality across Alberta and submitting station-level observations to the platform.

The data is subject to change at any time, as records may be further reviewed, corrected, or resubmitted by the reporting organizations. In addition, only data that has been officially submitted is available for download, and the displayed results depend on the selected search and filtering criteria.

Each row in the dataset represents a single observation from a monitoring station at a specific date and time for a given pollutant. Unlike the Ontario dataset, the Alberta data is stored in a vertical (long) format, where measurements for different parameters and timestamps are recorded as separate rows rather than as multiple hourly columns. The dataset includes, but is not limited to, the following key variables:

Feature name	Definition
Station ID	A unique identifier for the air monitoring station.
Station Name	The name of the monitoring station operated by an airshed or the government.
Parameter	The air pollutant or environmental parameter being measured (e.g., PM2.5).
DateTime	The date and time at which the observation was recorded.
Value	The measured concentration of the pollutant at the given time.
Units	The unit of measurement associated with the parameter.

From a technical perspective, collecting the Alberta data requires additional processing. Rather than downloading a single consolidated file, data must be retrieved station by station based on filtering criteria such as operator type and available parameters. To handle this efficiently and reproducibly, a Python-based web scraping program was used to automate the download process. The script iterates over all eligible stations, retrieves their corresponding datasets, and compiles a master table that serves as a directory of all stations meeting the project requirements.

1.1.4 Winnipeg, Manitoba:

The Winnipeg Air Quality Data was obtained from the City of Winnipeg Open Data Portal, available at: https://data.winnipeg.ca/Organizational-Support-Services/Air-Quality-deprecated-/f58p-2ju3/about_data

The dataset was published and maintained by the City of Winnipeg. The data records air quality measurements collected across Winnipeg, Manitoba, through a network of environmental sensors deployed throughout the city.

According to the data source, this dataset is currently marked as deprecated due to issues with the underlying data source, and it is no longer actively maintained. However, updates may resume in the future if a new data source is procured. Despite this designation, the dataset was last updated on November 26, 2025, and therefore remains valuable for historical analysis and recent trend exploration.

Each row in the dataset represents a single observation recorded at a specific location and time. Measurements are collected at 5-minute intervals, resulting in a large, high-frequency time-series dataset. The dataset includes the following key variables:

Feature name	Definition
ObservationID	A unique identifier for each recorded observation.
ObservationTime	The date and time when the measurement was recorded.
ThingID	An identifier associated with the sensing device.
LocationName	The name of the monitoring location within Winnipeg.
MeasurementType	The type of environmental measurement (e.g., PM2.5 particulates, temperature, humidity).
MeasurementValue	The recorded value of the measurement.
MeasurementUnit	The unit associated with the measurement value.
Location	Geographic coordinates of the monitoring station.
Point	Spatial point representation of the station location.

Data collection for Winnipeg is relatively straightforward. All records are integrated into a single, centralized table that can be downloaded directly from the data portal without the need for station-by-station retrieval or automated scraping.

1.2 Meteorological Data

1.2.1 Canada:

The Historical Climate Data (Canada) dataset was obtained from the following URL: https://climate.weather.gc.ca/. This dataset is sourced from the official Environment and Climate Change Canada (ECCC) archive.

It provides comprehensive historical weather observations from meteorological stations across Canada. The data includes key climatic variables such as air temperature, precipitation (rain and snow), wind speed, wind direction, and humidity.

Feature name	DataType	Definition
Longitude (x)	Float	The geographic longitude of the station in decimal degrees (North American Datum 1983 - NAD83). Negative values indicate West.
Latitude (y)	Float	The geographic latitude of the station in decimal degrees (NAD83). Positive values indicate North.
Station Name	String	The official name of the climate observation station (e.g., "TORONTO CITY").
Climate ID	String / Int	A unique 7-digit identifier assigned by the Meteorological Service of Canada. (Note: Often stored as a string to preserve leading zeros).

Feature name	DataType	Definition
Date/Time (LST)	DateTime	The full timestamp of the observation in ISO 8601 format (YYYY-MM-DD HH:MM). Crucial: This is always Local Standard Time (LST). It does not adjust for Daylight Saving Time.
Year	Int	The year of the observation.
Month	Int	The month of the observation (1-12).
Day	Int	The day of the month (1-31).
Time (LST)	Time	The hour of the observation (00:00 to 23:00). Based on Local Standard Time.
Temp (°C)	Float	Dry Bulb Temperature. The temperature of the ambient air in degrees Celsius.
Dew Point Temp (°C)	Float	The temperature to which the air would have to be cooled (at constant pressure) to become saturated. Used to calculate humidity.
Rel Hum (%)	Float	Relative Humidity. The ratio of the actual amount of water vapor in the air to the maximum amount it can hold at that temperature, expressed as a percentage.
Wind Dir (10s deg)	Float	Wind Direction. The direction from which the wind is blowing.

Feature name	DataType	Definition
Wind Spd (km/h)	Float	Wind Speed. The average wind speed over the 2 minutes ending at the time of observation, in kilometers per hour.
Visibility (km)	Float	The greatest distance at which a black object of suitable dimensions can be seen and recognized against the horizon sky.
Stn Press (kPa)	Float	Station Pressure. The atmospheric pressure measured at the station's elevation (not corrected to sea level), in kilopascals (1 kPa = 10 mb).
Wind Chill	Float	An index indicating how cold the weather feels to the average person due to the combined effect of cold temperature and wind. Only calculated when Temp $<$ 0°C.
Weather	Category	A textual description of the "Present Weather" observed at the station. (See detailed breakdown below).

Detailed Breakdown: The Weather Column The Weather column in ECCC data is a composite string. It is constructed by combining Intensity, Descriptor, and Precipitation/Obstruction.

1. Sky Conditions (Cloud Cover)

   Clear: Sky is clear (0/10 to 1/10 cloud cover).

Mainly Clear: Mostly clear (1/10 to 4/10 cloud cover).

Mostly Cloudy: More clouds than clear sky (5/10 to 9/10 cloud cover).

Cloudy: Overcast (10/10 cloud cover).

2. Precipitation (Can be Light, Moderate, or Heavy)

   Rain: Liquid precipitation.

   Drizzle: Very small liquid drops (slower falling).

   Snow: Frozen precipitation (flakes).

   Snow Grains: Very small, white, opaque grains of ice.

   Ice Crystals: Tiny ice needles or plates (diamond dust).

   Ice Pellets: Frozen raindrops (sleet).

   Hail: Solid balls of ice (diameter > 5mm).

3. Descriptors (Modifiers)

   Showers: Precipitation is intermittent (e.g., Rain Showers, Snow Showers).

   Freezing: Liquid precipitation freezing on contact (e.g., Freezing Rain, Freezing Drizzle, Freezing Fog).

   Blowing: Wind raising snow/sand to a height interfering with visibility (e.g., Blowing Snow).

   Drifting: Wind moving snow near the ground (e.g., Drifting Snow).

   Thunderstorms: Electrical storms (can be combined: Thunderstorm with Rain).

4. Obstructions to Vision

   Fog: Visibility reduced to less than 1 km by water droplets.

   Mist: Visibility reduced (1 km to 10 km) by water droplets.

   Haze: Visibility reduced by dry particles (dust/pollutants).

   Smoke: Visibility reduced by smoke (wildfires/industrial).

Common Combined Examples in the Dataset:

Rain, Fog (Raining and Foggy)

Moderate Snow (Snowing with moderate intensity)

Heavy Rain , Thunderstorm

Blowing Snow (Wind speed is high, visibility is low due to snow)

Using scripting tools like Python, leverage Station_Inventory.csv provided by ECCC to filter target site IDs and time ranges, then combine this with ECCC's fixed-format URLs to automate bulk CSV downloads and data merging.

Traffic Data

1.2.2 Statistics Canada:

The Statistics Canada Traffic Flow data was obtained from the following URL: https://www150.statcan.gc.ca/n1/pub/71-607-x/71-607-x2022018-eng.htm It is published by Statistics Canada.

We downloaded the dataset as a CSV export and kept the file in its original format without modification.

Each row of the dataset represents a single traffic camera with its location and metadata (e.g., camera identifier, source/jurisdiction, and road name). For the columns, the following table demonstrates the meaning of each column.

Feature name	Definition
WKT	Camera location geometry in WKT format: `POINT (longitude latitude)` (degrees).
CSDUID	Census Subdivision identifier associated with the camera location. Treat as an ID (string) even if stored numerically.
traffic_camera	Unique identifier for the traffic camera (one row per camera in this file).
traffic_source	Data source/jurisdiction providing the camera feed (e.g., a province or municipality).
camera_road	Road or corridor name associated with the camera, as provided by the source system.
xYYYY_MM_DD	Daily vehicle count for the calendar date `YYYY-MM-DD` for the given camera. Column names follow the pattern `xYYYY_MM_DD` (e.g., `x2022_02_02`). This file includes 1368 daily columns spanning 2022-02-02 to 2025-10-31. Values are non-negative counts (stored as numeric; may appear as decimals due to missing values). Missing/blank values indicate no data available for that camera on that date.

1.2.3 Government of Canada:

The Traffic Flow (Débit de circulation) data was obtained from the following URL: https://open.canada.ca/data/en/dataset/c77c495a-2a4c-447e-9184-25722289007f. It is produced by Québec's Ministry of Transport and Sustainable Mobility (MTMD).

We downloaded the dataset as a CSV export and kept the file in its original format without modification.

Each row of the dataset represents one **traffic section** (a segment of the Québec road network). For the columns, the following table demonstrates the meaning of each column.

Feature name	Definition
ide_sectn_trafc	Unique identifier for the traffic section.
num_sectn_trafc	Traffic section number (the same number is used for both sides of divided roads).
des_debut_sous_route	Description of the start of the sub-route.
des_fin_sous_route	Description of the end of the sub-route.
rtss_debut_chaing	Start RTSS and chainage (in metres) for the traffic section.
rtss_fin_chaing	End RTSS and chainage (in metres) for the traffic section.
annee_en_cours	Aggregated values for the current year, including DJMA (annual average daily traffic), DJMH (winter AADT), DJME (summer AADT), percentage of heavy vehicles, and the 30th highest hour.
anneex	Aggregated values for year "x", including DJMA (annual average daily traffic), DJMH (winter AADT), DJME (summer AADT), and the percentage of heavy vehicles.
dat_debut_sectn_trafc	Start date of the traffic section (format: YYYYMMDD).

Feature name	Definition
rtss_debut	Start RTSS for the traffic section. RTSS is a 14-character alphanumeric identifier in the MTQ linear reference system, structured as Route [99999], Tronçon [99], Section [999], Sous-route [9x9x].
val_chang_debut	Start chainage of the traffic section (in metres).
rtss_fin	End RTSS for the traffic section. RTSS is a 14-character alphanumeric identifier in the MTQ linear reference system, structured as Route [99999], Tronçon [99], Section [999], Sous-route [9x9x].
val_chang_fin	End chainage of the traffic section (in metres).
djma_annee_x	Reference year for DJMA (annual average daily traffic) for year “x”.
val_djma_annee_x	DJMA (annual average daily traffic) value for year “x” (unit: vehicles/day).
djme_annee_x	Reference year for DJME (summer average daily traffic) for year “x”.
val_djme_annee_x	DJME (summer average daily traffic) value for year “x” (unit: vehicles/day).
djmh_annee_x	Reference year for DJMH (winter average daily traffic) for year “x”.
val_djmh_annee_x	DJMH (winter average daily traffic) value for year “x” (unit: vehicles/day).
cam_annee_x	Reference year for percentage of heavy vehicles for year “x”.
val_cam_annee_x	percentage of heavy vehicles value for year “x” (unit: %).
val_30e_heure	Design hour (30th highest hour): estimate of the maximum ‘normal’ hourly traffic flow for the year.

Feature name	Definition
index_agreg	Flag indicating whether an aggregated historical data file is available for this section (O=Yes, N=No).
index_sectn	Flag indicating whether annual report files for permanent counting sites are available (O=Yes, N=No).
index_donnees	Flag indicating whether hourly data files (average hourly by day of week) are available (O=Yes, N=No).
url_index_agregees	URL linking to the aggregated historical data file (PDF), when available.
url_index_section	URL linking to annual report data files (PDF/XLS), when available.
url_index_donnees	URL linking to hourly data files (XLS), when available.
objectid	Unique internal identifier.

1.3 International Data

1.3.1 Beijing:

1.3.2 Shenzhen:

This Dataset was obtained from the following URL: https://www.microsoft.com/en-us/research/project/urban-air/ It is published by Urban Air Project (Urban Computing Team, Microsoft Research)

The Dataset is comprised of six parts of data that were collected over a period of one year (from 2014/05/01 to 2015/04/30), named city data, district data, air quality station data, air quality data, meteorological data and weather forecast data, respectively. This dataset covers 4 major Chinese cities (Beijing, Tianjin, Guangzhou and Shenzhen) and 39 adjacent cities within 300 kilometers to them. Each city is associated with a geo-location denoted by (latitude, longitude), containing a set of districts. In total, there are 2,891,393 air quality records, 1,898,453 (real-time) meteorology records, and 910,576 weather forecast records. Air quality is recorded at 437 air quality stations every hour. The real-time meteorological data are collected at a district (or city) level every hour. Weather forecast has a district (or city) level record of two coming days, with a temporal granularity of 3 hour, or 6 hour, or 12 hour. The feature definitions of each part of data are described as follows.

1.3.2.1 City Data (`city.csv`)

Feature Name	Definition
City ID	A 3-digit number representing the unique identifier for a city.
Chinese Name	The name of the city in Chinese.
English Name	The Chinese Pinyin corresponding to the city's Chinese Name.
Latitude	The latitude coordinate of the city center (chosen as the town hall).
Longitude	The longitude coordinate of the city center (chosen as the town hall).
Cluster ID	Indicates the city cluster: `1` for 'Cluster A' (near Beijing) and `2` for 'Cluster B' (near Guangzhou).

1.3.2.2 District Data (`district.csv`)

Feature Name	Definition
District ID	A 5-digit number representing the district; the first 3 digits correspond to the City ID.
Chinese Name	The name of the district in Chinese.
English Name	The Chinese Pinyin corresponding to the district's Chinese Name.
City ID	The identifier of the city to which the district belongs.

1.3.2.3 Air Quality Monitoring Station Data (`station.csv`)

Feature Name	Definition
Station ID	A 6-digit number representing the station; the first 3 digits correspond to the City ID.
Chinese Name	The name of the station in Chinese.
English Name	The Chinese Pinyin corresponding to the station's Chinese Name.
Latitude	The latitude coordinate of the station.
Longitude	The longitude coordinate of the station.
District ID	The identifier of the district to which the station belongs; used to link meteorology and forecast data.

1.3.2.4 Air Quality Data (`airquality.csv`)

Feature Name	Definition
Station ID	The identifier for the air quality monitoring station.
Time	The timestamp of the air quality record.
PM25	Concentration of PM2.5 in ug/m^3.
PM10	Concentration of PM10 in ug/m^3.
NO2	Concentration of Nitrogen Dioxide in ug/m^3.
CO	Concentration of Carbon Monoxide in mg/m^3.

Feature Name	Definition
O3	Concentration of Ozone in ug/m^3.
SO2	Concentration of Sulfur Dioxide in ug/m^3.

1.3.2.5 Meteorology Data (`meteorology.csv`)

Feature Name	Definition
ID	Corresponds to either the District ID or City ID.
Time	The timestamp of the meteorological record.
Weather	A code (0-16) representing weather conditions (e.g., 0=Sunny, 1=Cloudy, 2=Overcast, 3=Rainy, etc.).
Temperature	The temperature in Celsius (°C).
Pressure	The surface pressure in hPa.
Humidity	The relative humidity percentage.
Wind Speed	The speed of the wind in meters per second (m/s).
Wind Direction	A code (0-24) representing wind direction (e.g., 0=No wind, 1=East, 2=West, 3=South, 4=North, etc.).

1.3.2.6 Weather Forecast Data (`weatherforecast.csv`)

Feature Name	Definition
ID	Corresponds to either the District ID or City ID.
Forecast Time	The time at which the forecast was issued.
Future Time	The future time for which the weather is predicted.
Temporal Granularity	The updating frequency/interval of the forecast (3, 6, or 12 hours).
Weather	Predicted weather condition code; same definitions as in Meteorology data.
Up Temperature	The upper bound/high temperature forecast.
Bottom Temperature	The lower bound/low temperature forecast.

Feature Name	Definition
Wind Level	The median value representing the wind level (e.g., 3.5 is used for level 3-4).
Wind Direction	Predicted wind direction code; same definitions as in Meteorology data.

1.3.3 1.4.3 KnowAir:

The KnowAir dataset was obtained from the following URL: https://zenodo.org/records/15614907. It is published by Wang et al. (2025).

The dataset contains air quality data collected from two major, densely populated regions in China, recorded hourly from 2016 to 2023: - The Beijing-Tianjin-Hebei and Surrounding Areas BTHSA, with data from 228 monitoring stations. - The Yangtze River Delta YRD, with data from 127 monitoring stations.

The dataset contains 2 csv files that describe the coordiates of the monitoring stations and 2 NetCDF (.nc) files that contain the hourly air quality data for each region. The NetCDF files are collections of aligned 2-D arrays that all share the same coordinates (time x station ID). Each 2-D array represents a different air quality variable, such as PM2.5 concentration, Ozone, etc. The following tables demonstrates the meaning of each variable in the 4 files.

1.3.3.1 Dataset dictionary (`dataset_bthsa.nc` and `dataset_yrd.nc`)

Feature	Definition
Coordinates:	
time	datetime64: 2016-01-01T00:00:00.000000000 to 2023-12-31T23:00:00.000000000
station	
Air Quality Variables from CNEMC:	
PM2.5	Fine particulate matter
O3	Ozone
Meteorological Variables from ERA5 Reanalysis:	
t2m	2-meter air temperature
d2m	2-meter dew point temperature
tp	Total precipitation
sp	Surface pressure
blh	Boundary layer height
msdwswrf	Mean surface downward short-wave radiation flux
u100	U-component of wind at 100m
v100	V-component of wind at 100m

1.3.3.2 station Dictionary (`station_bthsa.csv and station_yrd.csv`)

Feature	Definition
station_id	station id
station_name	chinese name
city	chinese name of the city the station is in
city_en	english name of the city the station is in
lon	longitude of the station
lat	latitude of the station

Acknowledgements

We would like to express our sincere gratitude to Dr. Lysy and Teaching Assistant Daniel Zhang for their valuable guidance and supervision throughout this course. Their advising was instrumental in the successful completion of this case study.

We also acknowledge the various government agencies and organizations listed in the references for providing open-access datasets. We are grateful for the permission to utilize these resources for academic purposes.

Finally, we acknowledge the use of generative AI tools, specifically ChatGPT and Gemini, in the preparation of this report. These tools were utilized to refine grammatical fluency and ensure clarity of expression, allowing us to draft our initial thoughts in our own words before polishing. They also assisted in debugging code and formatting citations. However, these technologies served solely as supportive tools; the domain knowledge, analytical logic, and core ideas presented in this work remain the original intellectual product of the authors.

References

[1] Government of Ontario. Air quality ontario: Historical pollutant data, 2024. Accessed: 2024-05-22.

[2] Government of British Columbia. Bc air quality: Find station data map, 2024. Accessed: 2024-05-22.

[3] Government of Alberta. Access air data, 2024. Accessed: 2024-05-22.

[4] City of Winnipeg. Winnipeg open data: Air quality, 2024. Accessed: 2024-05-22.

[5] Environment and Climate Change Canada. Historical climate data, 2024. Accessed: 2024-05-22.

[6] Statistics Canada. Canadian vehicle survey: Interactive tool, 2022. Accessed: 2024-05-22.

[7] Government of Canada. Open government portal: Traffic datasets, 2024. Accessed: 2024-05-22.

[8] Gouvernement du Québec. Débit de circulation (traffic flow), 2024. Accessed: 2024-05-22.

[9] Harvard Dataverse. Beijing air quality dataset, 2024. Accessed: 2024-05-22.

[10] Microsoft Research. Urban air: Real-time fine-grained air quality analysis, 2024. Accessed: 2024-05-22.

[11] Shuo Wang. Knowair: Pm2.5-gnn dataset, 2020. Accessed: 2024-05-22.