Contents

[Q. 1](#_Toc209738252)

[A. 2](#_Toc209738253)

[Q. 2](#_Toc209738254)

[A. 2](#_Toc209738255)

[Q. 2](#_Toc209738256)

[A. 2](#_Toc209738257)

[Q. 3](#_Toc209738258)

[A. 3](#_Toc209738259)

[Q. 4](#_Toc209738260)

[A. 4](#_Toc209738261)

[Q. 6](#_Toc209738262)

[A. 7](#_Toc209738263)

[Q. 10](#_Toc209738264)

[A. 10](#_Toc209738265)

## Q.

Act as an experienced machine learning and data analyst with strong analytical skills in studying, collecting, and understanding data, as well as identifying the potential sources of datasets, to successfully implement a high-performance machine learning model. Provide an accurate, clear, and practical response to help me in my below request:

Request: I am working with my team on a machine learning project to analyze residential fire incident datasets in Montreal, Quebec, Canada over the past few years. Based on the data shared with our team, I aim to identify the most dangerous and fire-prone residential geographical locations, with the ultimate goal of implementing proactive preventative measures for these locations. The dependent variable is planned to be either (high risk, medium risk, medium-low risk, low risk, or very low risk). Our team is currently studying and analyzing other datasets that could be combined with our existing dataset, including datasets of fire incidents from previous years, which should be included, and other important datasets that should be included.

Please provide a list of up to five other datasets that could be used in addition to the existing (joined) datasets. Please be sure to clearly indicate in what way how each of the suggested datasets may help us in to our project by providing more relevant data to achieve our goal of high-performance model. Provide your answer in a table with columns 1) for the dataset, 2) explaining the reasons for including it, 3) what important/key features are expected in this dataset that may be used in the project, 4) and the importance and/or significance of including this dataset in our project to improve the performance of a high-performance machine learning model (accuracy, precision, etc.). Please ask me questions, one by one, for a maximum of three questions to get a better and more accurate answer.

For more context about the data we already have, please note that, it is the dataset is Interventions by Montreal firefighters (SIM interventions) available in

https://app.geo.ca/en-ca/map-browser/record/2fc8a2b9-1556-410e-a118-c46e97e9f19e

## A.

Got it — thanks for sharing the context and the link to the **SIM interventions dataset**. Since you already have residential fire incident data (location, time, type of intervention, etc.), the key to improving your model will be to **combine it with other explanatory datasets** that capture **environmental, socioeconomic, and infrastructure-related risk factors**.

Before I build the table, I’d like to clarify one thing to make sure the datasets I suggest are the most relevant to your project:

**Question 1:**  
Do you want to focus only on **Montreal-specific datasets** (from open government portals, municipalities, etc.), or are you open to also using **Canada-wide datasets** (e.g., census, environment) that can be filtered down to Montreal?

## Q.

we are open to all datasets as long as its data , reliability, and features included in it are closely related and relevant and helpful in enhancing the performance of our ML model

## A.

Perfect — that gives more flexibility to combine **Montreal city–level data** with **Canada-wide structured datasets** that can be filtered down to Montreal neighborhoods. That will let us enrich your fire incident data with **contextual features** that can help predict risk.

Here’s my second clarification before I prepare the table:

**Question 2:**  
Do you and your team want the additional datasets to be **mainly structured tabular datasets** (e.g., CSVs with socioeconomic, demographic, building data, etc.), or would you also consider **geospatial datasets** (e.g., shapefiles, GIS layers with zoning, building footprints, etc.) that can be spatially joined with your fire incident data?

## Q.

open for both if serving the above purpose but preference to be given to structured tabular datasets

## A.

Excellent — thanks, that helps me tailor the suggestions. Since you prefer structured tabular datasets but are open to geospatial if highly useful, I’ll prioritize datasets that are **directly usable in ML pipelines** while also mentioning geospatial ones that can be converted to features (like density, zoning, distance).

Here’s my last clarification:

**Question 3:**  
Do you want the additional datasets to **only focus on fire-related and safety-related factors** (e.g., building age, fire hydrants, inspections, etc.), or do you also want to include **broader socioeconomic and environmental risk factors** (e.g., income levels, household size, climate variables) that research shows correlate with fire incidence and severity?

## Q.

of course other datasets to be included as long as they are reliable and related to the context of project and serving the above mentioned purpose

## A.

Perfect — thanks for confirming. Since you’re open to **fire-specific + broader socioeconomic/environmental factors**, here’s a table of **five additional datasets** (with explanations) that would significantly enrich your SIM fire intervention data and improve the predictive power of your ML model.

| **Dataset** | **Why Include It** | **Key Features Expected** | **Importance for Model Performance** |
| --- | --- | --- | --- |
| **Canadian Census (Statistics Canada – Census Profile by Dissemination Area)** | Socioeconomic and demographic factors are strongly linked to fire risk (e.g., overcrowding, income, age of population). | Population density, household income, education level, % renters vs. owners, household size, immigrant population. | Adds context to fire risk beyond physical incidents. Helps capture social vulnerability and lifestyle factors that increase fire likelihood or severity. Improves classification accuracy by reducing bias toward only historical incident data. |
| **Montreal Building/Property Assessment Roll (Rôle d’évaluation foncière)** | Fire risk is tied to building characteristics (age, type, material, value). Montreal publishes property-level data. | Building year of construction, building type (single-family, multi-unit), number of floors, material type, assessed value. | Critical for modeling **structural fire risk**. Older, low-value, or wooden-frame buildings may be at higher risk. Adds strong predictive features for location-level fire vulnerability. |
| **Weather & Climate Data (Environment and Climate Change Canada – Historical Climate Data)** | Weather conditions strongly influence fire likelihood and spread. Combining past incidents with weather data improves predictive power. | Temperature, humidity, wind speed, precipitation, extreme weather events at the daily/hourly level. | Improves temporal risk prediction. Fire-prone conditions (e.g., dry + windy days) can explain seasonal spikes. Increases precision in predicting “when” and “where” fires occur. |
| **Urban Infrastructure / Hydrant & Fire Station Locations (City of Montreal Open Data or GIS)** | Response time and available infrastructure affect severity of incidents. Areas with poor hydrant coverage or far from stations are higher risk. | Distance to nearest fire station, number of hydrants per km², hydrant functionality, street accessibility. | Important for **mitigation effectiveness**. Helps identify neighborhoods where fires escalate due to slower response or limited resources. Enhances recall for high-risk predictions. |
| **Insurance & Fire Safety Violations / Inspection Records (if available from City or provincial sources)** | Past safety code violations are strong predictors of future incidents. | Number of fire code violations, inspection dates, unresolved hazards, insurance claim density. | High predictive power for **chronic fire-prone properties**. Directly linked to preventable risk factors. Improves overall model robustness by reducing false negatives for “high risk” areas. |

✅ With these datasets, you’ll be combining **historical incident data** (what already happened) with **structural, social, and environmental predictors** (why it happened and where it’s more likely). This multi-source integration is exactly what boosts ML model performance (accuracy, precision, and generalization).

Would you like me to also:

1. Suggest **practical sources/links** (Montreal Open Data, StatCan, ECCC) for each dataset, or
2. Show how to **engineer features** from each dataset (e.g., distance to hydrant, income per capita, weather anomalies) for ML input?

## Q.

yes please

## A.

Great 👍 — let’s make this practical and directly useful for your Montreal fire risk project. Below I’ve expanded the table with **reliable sources/links** where possible, and I’ve added a **“Feature Engineering Ideas”** column to show how to actually turn each dataset into ML-ready features.

| **Dataset** | **Source / Where to Find** | **Why Include It** | **Key Features Expected** | **Feature Engineering Ideas for ML** | **Importance for Model Performance** |
| --- | --- | --- | --- | --- | --- |
| **Canadian Census (Statistics Canada – Census Profile by Dissemination Area)** | [Statistics Canada Census Profiles](https://www12.statcan.gc.ca/census-recensement/index-eng.cfm) | Socioeconomic and demographic factors strongly influence fire risk. | Population density, household income, education level, % renters vs. owners, household size, immigrant population. | Create **risk indices** (e.g., population density × % renters), normalize income by household size, flag areas with high overcrowding. | Adds context to fire risk beyond incidents; improves generalization and reduces bias. |
| **Montreal Building/Property Assessment Roll (Rôle d’évaluation foncière)** | [Ville de Montréal Open Data](https://donnees.montreal.ca/) | Building characteristics are directly tied to fire risk. | Year of construction, building type, number of floors, material type, assessed value. | Engineer **building age**, categorical encoding of building type, combine value with type (e.g., low-value wooden triplex). | Strong predictor of **structural fire risk**, boosts precision. |
| **Weather & Climate Data (Historical Climate Data – ECCC)** | [Environment & Climate Change Canada – Historical Climate](https://climate.weather.gc.ca/historical_data/search_historic_data_e.html) | Weather drives short-term and seasonal fire risk. | Temperature, humidity, precipitation, wind speed, extreme events. | Lag features (weather before fire), **heat index**, **drought index** (low precipitation + high temp), anomaly detection (e.g., unusually dry month). | Improves temporal predictions, helps explain spikes. |
| **Urban Infrastructure (Hydrant & Fire Station Locations)** | [Ville de Montréal Open Data – Fire Stations](https://donnees.montreal.ca/) & hydrant GIS data | Response time & infrastructure coverage affect severity. | Distance to fire stations, hydrant density, hydrant condition. | Compute **Euclidean or network distance** from centroid of each neighborhood to nearest station/hydrant. Ratio of hydrants to households. | Improves **recall** by identifying underserved areas. |
| **Fire Safety Violations / Insurance & Inspection Records** | If available: [Régie du bâtiment du Québec](https://www.rbq.gouv.qc.ca/) or Montreal fire prevention inspection datasets (may need request). | Violations are early indicators of high-risk buildings. | Fire code violations, inspection status, claims history. | Features: **count of violations per building/area**, flag unresolved hazards, rolling window of claims/incidents. | High predictive power for chronic high-risk locations, reduces false negatives. |

🔑 **Next Steps for You & Your Team**

1. **Start with structured datasets** (Census, Building Roll, Weather) — all public, clean, and directly joinable.
2. **Add infrastructure datasets** (hydrants, fire stations) — may need some geospatial preprocessing but can be converted into tabular features.
3. **Push for inspections/violations dataset** — may require access requests, but even partial data is highly predictive.
4. Use **spatial joins** (by neighborhood, postal code, dissemination area) to merge all datasets with your SIM interventions dataset.

👉 Would you like me to draft a **step-by-step integration workflow** (from downloading → cleaning → feature engineering → merging → modeling) for these datasets, so you have a ready-made roadmap for your ML project?

## Q.

thanks but what other data sets that include data for the below factors which are directly or indirectly affect the level of risk residential fire occur in a specific resiedential area/home

Reasons to be explored - please list here

type of building

# of residents

location of supermarkets - incidents in the super lead to more risk

population age

average number fire incidents and # of total house income

vacant building

fire alarms, fire detectors

age of building

stages of fire development in residential room fires

Locations of Windows

Sizes of Windows

Firefighters knowledge & expereince (historical job/performance record)

gas cylinder mishandling

improper storage of goods

Close/Distance/Location to Forest ?

"building age,

number of extinguishers,

distance to hydrant,

neighborhood crime rate,

Owner has Phone ?"

Human negligence / distraction (cooking, unattended appliances, careless behavior) — ~35%. SIM reports a very high share of fires are linked to human negligence/distraction; their public materials emphasize this as the dominant driver. (Service de sécurité incendie de Montréal)

Smoking materials (cigarettes, ashtrays) — ~15% (major contributor to deaths and many fatal residential fires in Montréal). (Service de sécurité incendie de Montréal)

Electrical malfunction (wiring, outlets, faulty equipment) — ~12%. Increasing presence of small-electronics / batteries noted in SIM outreach. (Service de sécurité incendie de Montréal)

Heating equipment (space heaters, furnaces, chimneys) — ~8% (winter months increase). (Canadian Red Cross)

Candles / open flame (including religious candles, decorative) — ~6% (candles and other open flames are repeatedly identified by municipal prevention campaigns). (Service de sécurité incendie de Montréal)

age of building

follow fire code ? Recmmendations by building code for fires

Land Use Data

## A.

This is an excellent and very thoughtful list 👌 — you’re essentially mapping out a **multi-dimensional fire risk model** that combines **structural, social, behavioral, and environmental factors**. Not all of these have publicly available datasets at the same granularity, but many can be captured either **directly** (through open datasets) or **indirectly** (via proxies/engineered features).

Below I’ve grouped your factors into themes, listed possible datasets, and explained how they could help.

**🔹 1. Building Characteristics & Occupancy**

| **Factor** | **Possible Dataset(s)** | **Key Features** |
| --- | --- | --- |
| Type of building, Age of building, Number of residents, Vacant buildings | **Montreal Property Assessment Roll** (open data) + **Statistics Canada Census Housing Data** | Building type (single-family, duplex, triplex, apartments), year built, number of dwellings, occupancy rate, vacancy indicators. |
| Windows (location, size), Number of extinguishers, Distance to hydrant | **Building Permit Data** (City of Montreal, sometimes tied to plans) + **Urban Infrastructure datasets (hydrants, hydrant condition, fire stations)** | Window count/size (if available in permit/building footprint datasets), hydrant density/distance, safety equipment info if inspections available. |
| Fire alarms, fire detectors, compliance with fire code | **Fire inspection datasets** (if accessible from SIM or Régie du bâtiment du Québec) | Presence of alarms, extinguishers, compliance violations. |

**🔹 2. Socioeconomic & Demographic Factors**

| **Factor** | **Possible Dataset(s)** | **Key Features** |
| --- | --- | --- |
| Population age, Household income, Average number of residents | **Statistics Canada Census Profiles** | Median age, % elderly, household income, household size. |
| Neighborhood crime rate | **Montreal Police SPVM Open Data** | Reported crimes per neighborhood (proxy for socioeconomic stress, possible fire-setting). |
| Owner has phone (proxy: communication/digital access) | **Census – Technology Access Data (Internet, cell access, telephone service)** | % of households with telephone/internet access (indicator of ability to report incidents quickly). |

**🔹 3. Behavioral & Human Risk Factors**

| **Factor** | **Possible Dataset(s)** | **Key Features** |
| --- | --- | --- |
| Human negligence/distraction (cooking, unattended appliances), Smoking, Electrical malfunction, Heating equipment, Candles/open flame | **SIM Fire Incident Reports (existing)** — cause categories are usually recorded. | Incident cause breakdown: % cooking, % electrical, % smoking, % heating. Can be aggregated by area to model local behavioral risks. |
| Gas cylinder mishandling, Improper storage of goods | **Municipal hazardous material permits/inspection records** (if accessible). | Gas use permits, storage violations, reported hazmat incidents. |

**🔹 4. Environmental & Land Use**

| **Factor** | **Possible Dataset(s)** | **Key Features** |
| --- | --- | --- |
| Close to forest/green areas | **Montreal Land Use GIS Data / CanVec Land Cover (Natural Resources Canada)** | Distance to forest, density of trees, wildland-urban interface zones. |
| Land use zoning (residential, commercial, mixed use, industrial) | **Montreal Land Use & Zoning Data (open data)** | Type of land parcel, zoning category, mixed-use vs. strictly residential. |
| Location of supermarkets (and similar high-footfall sites) | **Business Listings / Points of Interest Data (Montreal Open Data, or private sources like OpenStreetMap)** | Location of supermarkets, shops, restaurants; foot traffic density. |

**🔹 5. Operational & Firefighter Factors**

| **Factor** | **Possible Dataset(s)** | **Key Features** |
| --- | --- | --- |
| Firefighter knowledge & experience (job/performance history) | **Internal SIM workforce/training records** (may not be public, but could be requested for collaboration). | Training level, years of service, response performance. |
| Stages of fire development in residential room fires | **NFPA / NIST fire research datasets** (experimental fire dynamics, simulations) | Not Montreal-specific, but could provide synthetic features (time to flashover, growth rates by building type). |

**🔹 6. Fire Code & Safety Compliance**

| **Factor** | **Possible Dataset(s)** | **Key Features** |
| --- | --- | --- |
| Follow fire code? | **Régie du bâtiment du Québec (RBQ) / Montreal Inspection Data** | Compliance status, code violations, enforcement history. |
| Building permits / retrofits | **Montreal Open Data (permits)** | Renovation type, retrofits (sprinklers, alarms). |

✅ **Takeaway:**  
You already have SIM incident-level data (causes, interventions). By combining it with:

* **Census (population, income, age, occupancy)**
* **Property/building datasets (age, type, vacancy, permits)**
* **Inspection/compliance (alarms, extinguishers, violations)**
* **Infrastructure (hydrants, stations, land use, forest proximity)**
* **Behavioral causes (from SIM’s own cause fields)**

…you can capture nearly all the risk factors you listed. Some datasets (like firefighter performance or window sizes) may be hard to get publicly, but proxies exist (e.g., building age/type often correlates with window design and code compliance).

👉 Do you want me to now **map these datasets into a prioritized acquisition plan** (i.e., which ones you should start collecting first because they give the biggest performance boost to your ML model, and which ones can be added later)?

## Q.

I want you to now map these all datasets into a prioritized acquisition plan (i.e., which ones you should start collecting first because they give the biggest performance boost to your ML model, and which ones can be added later) and provide links to access and download the dataset for all of them

## A.

Great — below is a prioritized **acquisition plan** that maps every dataset we discussed (and a few important extras) into a clear order: **start-with / add-soon / add-if-possible / specialized**. For each item I give: the dataset name, a direct authoritative link (where available), why it’s important for your Montreal residential-fire risk model, the key features you should extract, and a short note about acquisition effort or access requirements.

I’ve prioritized by **expected predictive value → ease of acquisition → ability to join (spatial/temporal)** so you can get fast gains and then layer on higher-effort but valuable sources.

Important: the SIM interventions dataset you already shared is baseline and should be the anchor for joins and labeling. ([Open Government Canada](https://open.canada.ca/data/en/dataset/2fc8a2b9-1556-410e-a118-c46e97e9f19e?utm_source=chatgpt.com))

**Prioritized acquisition plan — datasets, links, rationale, features, effort**

**Priority A — Collect first (highest impact, usually public & tabular)**

1. **SIM interventions (ground-truth incident data)** — *anchor dataset (if not already fully downloaded).*
   * Link: Interventions by Montreal firefighters (Open Canada). ([Open Government Canada](https://open.canada.ca/data/en/dataset/2fc8a2b9-1556-410e-a118-c46e97e9f19e?utm_source=chatgpt.com))
   * Why: Source of incident labels (fire occurrence, type/cause, location, date/time, severity). Essential for supervised learning.
   * Key features: geolocation, date/time, incident cause, units deployed, severity codes, building address/type (if present).
   * Effort: You already have it — ensure full historical range and clean geocodes.
2. **Statistics Canada — Census data (Dissemination Area or DA-level profiles)**
   * Link: Census Profile / WDS download (2021 and other years). ([Statistics Canada](https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/download-telecharger.cfm?Lang=E&utm_source=chatgpt.com))
   * Why: Provides socioeconomic & demographic predictors strongly correlated with fire risk (income, age, household size, % renters, phone/internet access proxies).
   * Key features: median household income, household size, % renters vs owners, population density, age distribution, % with phone/internet.
   * Effort: Public CSV/API; join by DA or postal-code mapping; high value for model generalization.
3. **Montreal Property Assessment / Evaluation Roll (property-level building attributes)**
   * Link: Ville de Montréal / evaluation roles (property assessment pages & EvalWeb). ([Montréal](https://montreal.ca/demarches/consulter-les-roles-devaluation-fonciere?utm_source=chatgpt.com))
   * Why: Direct building-level structural features — year built, building type, assessed value, number of dwellings. Strong structural risk signal.
   * Key features: year of construction, building type (single-family, multiplex, apartment), number of units, assessed value, lot/building footprint (if present).
   * Effort: Public (municipal portal / EvalWeb); may be in XML/CSV — needs processing to join to addresses.
4. **Montreal / Canada hydrant locations & fire station locations (infrastructure)**
   * Links:
     + Fire hydrants (Canada Open Data / Montreal coverage). ([Open Government Canada](https://open.canada.ca/data/en/dataset/cb4de65e-138b-4936-9d5c-2d9a0bc9b4ce?utm_source=chatgpt.com))
     + Fire stations (Ville de Montréal Open Data). ([Montreal Open Data](https://donnees.montreal.ca/dataset/casernes-pompiers?utm_source=chatgpt.com))
   * Why: Response capability features (distance to nearest station, hydrant density) strongly affect severity and are predictive of risk escalation.
   * Key features: point coords for hydrants and stations, hydrant attributes (if pressure/condition present), station coverage polygons.
   * Effort: Public geospatial points — compute network/euclidean distances, hydrant-per-household ratios per area.
5. **Building permits & renovation records (Montreal permits dataset)**
   * Link: Montreal open data — building/ construction permits. ([Open Data Montreal](https://opendatamontreal.ca/data/building-permits?utm_source=chatgpt.com))
   * Why: Permit history shows renovations (e.g., electrical, heating upgrades), code compliance, and retrofits (sprinklers/alarms). Good proxy for safety retrofits and window changes.
   * Key features: permit type, date, work description, address, permit outcome.
   * Effort: Public; needs text parsing to extract relevant permit types (electrical, heating, renovation).

**Priority B — Add next (high impact, some spatial processing or access requests)**

1. **Environment and Climate Change Canada — Historical weather (station-level hourly/daily)**
   * Link: ECCC historical data / climate data extraction tool. ([Climate Data Canada](https://climate.weather.gc.ca/?utm_source=chatgpt.com))
   * Why: Weather drives seasonal/temporal risk and supports lag features (dry/windy days → higher short-term risk). Improves temporal precision.
   * Key features: daily/hourly temp, humidity, wind speed, precipitation; heat/drought anomalies.
   * Effort: Public CSV downloads by station — map incidents to nearest station/time window; compute lag/rolling features.
2. **SIM / Montreal fire inspection & code-violation records (inspections, alarms, extinguisher presence)**
   * Link: may require request to SIM or RBQ; RBQ general pages for safety code & complaint service. ([RBQ](https://www.rbq.gouv.qc.ca/?utm_source=chatgpt.com))
   * Why: Direct indicators of non-compliance (missing smoke detectors, unresolved hazards). Very high predictive power for chronic problem properties.
   * Key features: inspection date, violation type, resolved/unresolved flags, number of violations per address.
   * Effort: Often not fully public — you may need an access request to municipal SIM or RBQ, or look for partial datasets on Montreal open-data portals.
3. **SPVM / Montreal crime & incidents (neighborhood crime rates)**
   * Link: SPVM crime profiles & Montreal crime open data. ([SPVM](https://spvm.qc.ca/en/Fiches/Details/Crime-profiles?utm_source=chatgpt.com))
   * Why: Crime rates often correlate with social disorder and risks (e.g., arson, property neglect). Useful as a socioeconomic/behavioral proxy.
   * Key features: crime counts by type, geolocation / neighborhood, temporal patterns.
   * Effort: Public open data; requires aggregation to DA/postal areas.
4. **Points-of-Interest (POI) — supermarkets, restaurants, commercial locations (OpenStreetMap or Montreal business registry)**
   * Link: OpenStreetMap or Montreal business/POI datasets; Montreal open data has business listings. (OSM + donnes.montreal.ca) ([Montreal Open Data](https://donnees.montreal.ca/en?utm_source=chatgpt.com))
   * Why: Proximity to high-footfall locations (supermarkets, restaurants) changes ignition risk (commercial cooking incidents, higher foot traffic) and exposure. Also useful to model mixed-use risk.
   * Key features: POI type, coordinates, hours (if available).
   * Effort: OSM export or municipal business lists; spatial join to compute counts/distances.

**Priority C — Add when feasible (specialized, private, or research datasets)**

1. **Land cover / forest proximity / wildland-urban interface (NRCan / CanVec / Canadian Forest Service)**
   * Link: Canadian Forest Service / CNFDB (forest fire databases) & land cover products. ([cwfis.cfs.nrcan.gc.ca](https://cwfis.cfs.nrcan.gc.ca/ha/nfdb?utm_source=chatgpt.com))
   * Why: For properties near large green spaces / forest patches, wildfire/vegetation fire risk increases. Useful for edges of city and summer risk.
   * Key features: landcover type, tree canopy density, distance to forest/wildland.
   * Effort: Public spatial data; compute proximity metrics.
2. **Hazardous materials permits / gas cylinder registration / business storage permits (municipal or provincial)**
   * Link: RBQ and municipal permitting pages; may require records requests (RBQ, borough-level). ([RBQ](https://www.rbq.gouv.qc.ca/en/online-services/complaint-report-and-declaration/public-safety/?utm_source=chatgpt.com))
   * Why: Areas with more permitted hazardous storage or frequent hazmat complaints have higher localized risk of storage-related fires.
   * Key features: permit type, address, storage volumes, inspection history.
   * Effort: Often not fully public; expect requests or FOI.
3. **Insurance / claims aggregates (private: CatIQ, IBC, or insurer datasets)**
   * Link: Insurance Bureau of Canada (IBC) insights; CatIQ (subscription) — public summaries exist. ([IBC](https://www.ibc.ca/news-insights/news/2024-shatters-record-for-costliest-year-for-severe-weather-related-losses-in-canadian-history-at-8-5-billion?utm_source=chatgpt.com))
   * Why: Claims density / cost per area helps validate and weight risk and reveal under-reporting in incident datasets.
   * Key features: claims counts, insured loss, claim type by geography (if available).
   * Effort: Usually paid/subscription or access-controlled; use if budget/partnership available.
4. **NIST / NFPA / academic fire dynamics datasets (stages of fire growth, flashover times)**
   * Link: NFPA / NIST fire research resources; scientific data (fire spread datasets). ([Nature](https://www.nature.com/articles/s41597-024-03436-4?utm_source=chatgpt.com))
   * Why: Provides mechanistic understanding (growth rates, flashover times) and can be used to generate synthetic features or priors for stage-of-fire modeling.
   * Key features: time-to-flashover under different fuels, growth curves by room type/material.
   * Effort: Public research — needs translation into model features (e.g., expected time-to-critical by building-type).
5. **Firefighter workforce / training & historical performance records (SIM internal HR data)**
   * Link: likely internal to SIM; request needed.
   * Why: Experience and training levels in a particular station/shift can affect response quality and outcomes; useful if you can access anonymized aggregate performance stats.
   * Key features: years of service, certified trainings, average response times by crew, historical outcomes.
   * Effort: Internal data — collaboration/partnership with SIM required; sensitive access controls.

**Quick recommended acquisition sequence & actions (concrete roadmap)**

1. **Immediate (week 0–2):** Export / confirm full SIM interventions history (clean geocodes & timestamps). ([Open Government Canada](https://open.canada.ca/data/en/dataset/2fc8a2b9-1556-410e-a118-c46e97e9f19e?utm_source=chatgpt.com))
2. **Short term (week 1–3):** Download Census DA profiles (StatCan API/CSV) and dissemination area boundaries for spatial joins. ([Statistics Canada](https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/download-telecharger.cfm?Lang=E&utm_source=chatgpt.com))
3. **Short term (week 1–3):** Download Montreal property assessment/eval roll (EvalWeb / municipal pages). ([Montréal](https://montreal.ca/demarches/consulter-les-roles-devaluation-fonciere?utm_source=chatgpt.com))
4. **Short term (week 1–3):** Download hydrant & fire station point datasets and building permits (Montreal Open Data). ([Open Government Canada](https://open.canada.ca/data/en/dataset/cb4de65e-138b-4936-9d5c-2d9a0bc9b4ce?utm_source=chatgpt.com))
5. **Medium term (week 3–6):** ECCC weather historical data (station mapping & lag features). ([Climate Data Canada](https://climate.weather.gc.ca/?utm_source=chatgpt.com))
6. **Medium term (week 3–8):** SPVM crime datasets, POI (OSM/business), landcover proximity. ([Open Data Montreal](https://opendatamontreal.ca/data/criminal-acts?utm_source=chatgpt.com))
7. **Medium–long term (month 2+):** Request SIM/RBQ inspection & violations data; request access or partnership for firefighter performance data; consider CatIQ/IBC for insurance data if budget allows. ([RBQ](https://www.rbq.gouv.qc.ca/?utm_source=chatgpt.com))
8. **Ongoing:** Add NIST/NFPA research-derived features (synthetic) and regularly refresh weather & recent incident data.

**Notes on joining, feature engineering & likely model gains**

* **Spatial join keys:** addresses (normalized), postal codes (FSA + LDU), dissemination area / census tract are safest for privacy and consistent joins. Use DA when address-level joins are messy. ([Statistics Canada](https://www150.statcan.gc.ca/n1/en/catalogue/92-169-X?utm_source=chatgpt.com))
* **Temporal alignment:** join weather and event-lag features (1–7 day lags), and aggregate socioeconomics at a stable time window (census year).
* **Proxies where direct data missing:** window size/location → use building permits & building age/type; number of residents → household size from census; human negligence categories → aggregate SIM cause fields per area.
* **Likely performance improvement:** adding structural (property roll) + socioeconomic (Census) + infrastructure (hydrants/stations) typically yields the largest AUC/accuracy gains in urban-fire risk models — these are the highest-priority sources.

**Quick links summary (clickable)**

* SIM interventions (Interventions by Montreal firefighters) — Open Canada. ([Open Government Canada](https://open.canada.ca/data/en/dataset/2fc8a2b9-1556-410e-a118-c46e97e9f19e?utm_source=chatgpt.com))
* Statistics Canada — Census Profile / WDS (download service). ([Statistics Canada](https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/download-telecharger.cfm?Lang=E&utm_source=chatgpt.com))
* Montreal Property Assessment / EvalWeb pages (property evaluation roll). ([Montréal](https://montreal.ca/demarches/consulter-les-roles-devaluation-fonciere?utm_source=chatgpt.com))
* Fire hydrants — Canada Open Data (Montreal coverage). ([Open Government Canada](https://open.canada.ca/data/en/dataset/cb4de65e-138b-4936-9d5c-2d9a0bc9b4ce?utm_source=chatgpt.com))
* Fire stations — Ville de Montréal Open Data. ([Montreal Open Data](https://donnees.montreal.ca/dataset/casernes-pompiers?utm_source=chatgpt.com))
* Building permits (Montreal Open Data). ([Open Data Montreal](https://opendatamontreal.ca/data/building-permits?utm_source=chatgpt.com))
* Environment & Climate Change Canada — Historical Climate Data & extraction tool. ([Climate Data Canada](https://climate.weather.gc.ca/?utm_source=chatgpt.com))
* SPVM crime data / crime profiles. ([SPVM](https://spvm.qc.ca/en/Fiches/Details/Crime-profiles?utm_source=chatgpt.com))
* OpenStreetMap / Montreal Open Data (POI, hydrant imports). ([OpenStreetMap](https://wiki.openstreetmap.org/wiki/Montr%C3%A9al/Imports/Bornes_d%27incendie?utm_source=chatgpt.com))
* Canadian Forest Service / CNFDB & landcover data (forest proximity). ([cwfis.cfs.nrcan.gc.ca](https://cwfis.cfs.nrcan.gc.ca/ha/nfdb?utm_source=chatgpt.com))
* RBQ (Régie du bâtiment du Québec) — safety code & complaints (inspection entry point). ([RBQ](https://www.rbq.gouv.qc.ca/?utm_source=chatgpt.com))
* Insurance / CatIQ / IBC (aggregates & industry data — typically paid/partner). ([CatIQ](https://public.catiq.com/category/catastrophes/?utm_source=chatgpt.com" \o "Category: Catastrophes))

If you’d like, I can now **(pick one)**:

1. Draft the **exact API/download commands and quick Python snippets** to fetch and preview each public dataset (addresses, code samples to read CSVs / shapefiles and do the first joins), **or**
2. Produce a **compact prioritized feature-engineering checklist** mapping each dataset to 8–12 final ML-ready features (exact column names and transformation suggestions), ready to drop into your pipeline.

Which of the two would you like me to do next?

## Q.

what if it is just a quick project for one month or less and people working on it as part time small project - all team members already have other FT jobs and dedicating few hours a week to work on this project. What are the 5 most important, reliable and relevant datasets in all your answers to be considered (in addition to the Montreal SIM interventions dataset). Provide table with dataset name, its source and link to access and download, its importance/necessity to be included (must be included, should be included, preferred to be included)

## A.

Got it ✅ — since this is a \*\*short timeline (≤1 month, part-time project)\*\*, the focus should be on \*\*datasets that are quick to obtain, already structured, and directly relevant\*\* to fire risk prediction in residential areas.

Here’s a \*\*prioritized shortlist of 5 datasets\*\* (in addition to the Montreal SIM interventions dataset you already have):

| **Dataset** | **Source & Link** | **Importance / Necessity** | **Why Include** |
| --- | --- | --- | --- |
| **Census Profile, Montréal (demographics, households, income, age distribution, housing type, vacancy, etc.)** | Statistics Canada 2021 Census: Census Profile Montréal  <https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=E> | **Must be included** | Provides population density, household size, income, building occupancy, and age structure → strong predictors of residential fire risk. |
| **Open Data Montréal – Building Footprints & Attributes (includes age of building, type of building, residential/commercial use, # floors, etc.)** | Ville de Montréal Open Data: Building Data | **Must be included** | Crucial for linking fire incidents to building characteristics (age, size, type, residential density). |
| **Fire Hydrant & Water Infrastructure Locations** | Ville de Montréal Open Data: Fire Hydrants | **Should be included** | Distance to hydrants is a strong determinant of fire response effectiveness and potential damage. Easy to spatially join with incidents. |
| **Socioeconomic Indicators (low-income prevalence, unemployment, housing conditions)** | Statistics Canada – Neighbourhood-level socioeconomic data (e.g., Canadian Index of Multiple Deprivation, SDG Hub datasets) | **Should be included** | Adds context for human-behavior-driven fire risks (neglect, smoking, heating practices). Helps capture risk patterns not explained by physical building features. |
| **Land Use & Zoning (residential vs commercial proximity, vacant land/buildings)** | Ville de Montréal Open Data: Zoning | **Preferred to be included** | Helps identify high-risk land use clusters (dense residential, proximity to commercial activity, vacant structures prone to fire). |

🔥 \*\*Summary for your short project\*\*:

1. \*\*Census data (StatsCan)\*\* and \*\*Building data (Montréal Open Data)\*\* are \*\*top priority\*\* → these will give the biggest boost to model accuracy.

2. \*\*Hydrant data\*\* and \*\*Socioeconomic indicators\*\* are second tier but still quick wins.

3. \*\*Zoning data\*\* is useful but optional for a 1-month project if time runs short.

Would you like me to also suggest a \*\*practical step-by-step workflow\*\* for how to quickly \*\*merge these datasets with your SIM interventions data\*\* so your team doesn’t lose time figuring out the integration?

## Q.

https://donnees.montreal.ca/ville-de-montreal/batiments

https://donnees.montreal.ca/ville-de-montreal/bornes-fontaines

https://donnees.montreal.ca/ville-de-montreal/zonage

Above links in your answer are broken links. I receive below error message

404 Not Found

The requested URL was not found on the server. If you entered the URL manually please check your spelling and try again.

## A.

Thanks — good catch, and sorry for the broken links. Those earlier paths were slightly off. Below are **correct, working dataset pages** on Montréal’s open-data portal (exact pages where you can download CSV / GeoJSON / SHP). I’ve included a one-line download tip for each.

| **Dataset** | **Working link (click to open)** | **Quick download tip** |
| --- | --- | --- |
| Bâtiments (building footprints / 2.5D roofs) | <https://donnees.montreal.ca/dataset/batiment-2d>. ([Montreal Open Data](https://donnees.montreal.ca/dataset/batiment-2d)) | On the page look under **Ressources** → download **SHP** or **GPKG** (or use the API/GeoJSON resource if you prefer). |
| Bornes d'incendie (fire hydrants) | <https://donnees.montreal.ca/fr/dataset/geolocalisation-des-bornes-fontaines>. ([Montreal Open Data](https://donnees.montreal.ca/fr/dataset/geolocalisation-des-bornes-fontaines/resource/75d2f1e2-b88d-4784-87df-5e7fc1e491a9)) | Use the **Télécharger (CSV)** button or the **API** link shown on the page to get CSV/GeoJSON. |
| Plan d'urbanisme — Densité de construction (zoning / density layer) | <https://donnees.montreal.ca/dataset/plan-urbanisme-densite>. ([Montreal Open Data](https://donnees.montreal.ca/dataset/plan-urbanisme-densite)) | Under **Ressources** you can download **GeoJSON** or **SHP** (GeoJSON is fastest for quick joins). |
| Plan d'urbanisme — Limites de hauteur (zoning / height limits) | <https://donnees.montreal.ca/dataset/plan-urbanisme-limites-hauteur>. ([Montreal Open Data](https://donnees.montreal.ca/dataset/plan-urbanisme-limites-hauteur)) | Page provides **GeoJSON** and **SHP** resources — use GeoJSON for ease of use in Python. |
| Montréal Open Data home (search other datasets) | <https://donnees.montreal.ca/> (portal search) ([Montreal Open Data](https://donnees.montreal.ca/?utm_source=chatgpt.com)) | Use the search box on the portal to find related layers (permits, hydrants updates, building permits, inspections). |

If you want, I can **immediately** paste a tiny Python snippet (requests + geopandas / pandas) that downloads one of these GeoJSON/CSV resources directly (so your team can copy–paste and fetch every file in minutes). Which format do you prefer for downloads — **CSV**, **GeoJSON**, or **SHP**?