CareerFoundry Data Analytics
September 25, 2025
Charlotte Lin

Achievement 6.1: Sourcing Open Data

Urban Flood Risk Data: Global City Analysis 2025

Data Source & Collection

The dataset used for this project is the Urban Flood Risk Data from Kaggle, which catalogs segments across global cities to assess rainfall-driven (pluvial) flood risk. Each record represents a spatial segment with variables including geographic coordinates, topography, land use, drainage infrastructure, rainfall intensity, and qualitative risk labels. The data was synthetically compiled from multiple global elevation and land datasets, rainfall sources, and infrastructure proximity metrics.

Brief Requirement

The Achievement 6 Brief requires selecting an open-source dataset that contains a geographic component along with sufficient categorical and continuous variables. The data must be recent (no more than 3 years old, or up to 10 years if no newer data is available), include at least 1,500 rows, and be suitable for exploratory, geospatial, and advanced analytical techniques. The Urban Flood Risk Data: Global City Analysis 2025 meets all of these requirements.

Data Limitation

- **Construction**: The dataset is pulled and modeled from global sources (e.g., DEMs, rainfall reanalysis) and does not represent direct field measurements.
- **Sentinel values**: Elevation includes `-3.0` as a placeholder for below-sea-level areas or where DEM coverage is incomplete.
- **Incomplete mapping**: Proximity and drainage density metrics are sometimes missing due to unmapped or poorly documented infrastructure networks.
- **Rainfall source gaps**: The `rainfall_source` field may contain `"None"` when no source is assigned, or values may be inconsistent across geographies (ERA5, gauges, blended sources, IMD).
- **Risk labels**: Labels can be multi-valued (pipe-delimited), event-specific (`event_YYYY-MM-DD`), or incomplete in representing the full spectrum of flood hazards.
- **Geographic abstraction**: The dataset includes latitude/longitude coordinates but lacks shapefiles or detailed boundary geometries for more advanced flooding spot analysis.

Why This Data

This dataset was chosen because it integrates **spatial, hydrologic, and infrastructure dimensions** into one table. It directly supports questions about urban resilience, climate adaptation, and infrastructure planning, which are all highly relevant to sustainability and strategic planning.

Data Cleaning Summary

- **Column name check**: all columns retained their lower case with underscore format from the raw dataset. No changes were needed.
- **Type conversion**: Checked numeric and categorical columns with `df.info()`; no conversions were required as all were already in correct formats.
- **Missing data handling**:
    - **Elevation** (`elevation_m`) → Missing values were imputed by taking the **median elevation within each city** (`city_name`). This preserves geographic context while preventing distortion from outliers. Sentinel values of `-3.0` (below sea level or DEM gaps) were left as-is, which serves as built-in flags for future analysis.
    - **Soil group** (`soil_group`) → Missing values were replaced with the **mode of the column**, which was category **"B"**.
    - **Drainage density** (`drainage_density_km_per_km2`) → Missing values were imputed by the **mean drainage density within each land use group** (`land_use`).
    - **Storm drain proximity** (`storm_drain_proximity_m`) → Missing values were imputed by the **mean storm drain proximity within each land use group** (`land_use`).
    - **Storm drain type** (`storm_drain_type`) → Missing values were replaced with `"Unknown"`.
    - **Rainfall source** (`rainfall_source`) → Missing values were replaced with `"Unknown"`.

  Imputation strategies (median by city, mean by land use, mode for soil group) preserved realistic distributions while filling missingness. This is confirmed because no variable experienced drastic shifts in mean or median, confirming that imputations were appropriate and non-distorting.

- **Sentinel value**: Recognized `-3.0` values in `elevation_m` as sentinel cases representing below-sea-level or DEM gaps; retained them for analysis.
- **Risk labels**: Reviewed the multi-valued risk_labels column (pipe-delimited); no transformations were applied during cleaning, though parsing into multi-hot indicators may be useful for later analysis.
- **Duplicates**: Verified none via uniqueness check; no duplicate segment IDs were found.
- **Export**: Saved cleaned dataset (2,962 rows × 17 columns).

Data Profile

This data profile is constructed based on the final cleaned dataset.

	dtype	n_unique	n_missing	pct_missing
segment_id	object	2962	0	0.0
city_name	object	63	0	0.0
admin_ward	object	91	0	0.0
latitude	float64	2962	0	0.0
longitude	float64	2962	0	0.0
catchment_id	object	180	0	0.0
elevation_m	float64	2229	0	0.0
dem_source	object	5	0	0.0
land_use	object	9	0	0.0
soil_group	object	4	0	0.0
drainage_density_km_per_km2	float64	855	0	0.0
storm_drain_proximity_m	float64	1750	0	0.0
storm_drain_type	object	5	0	0.0
rainfall_source	object	5	0	0.0
historical_rainfall_intensity_mm_hr	float64	866	0	0.0
return_period_years	int64	6	0	0.0
risk_labels	object	195	0	0.0

The cleaned dataset contains 2,962 records across 17 variables, with a mix of categorical, numeric, and spatial attributes. A variable-level profile shows the following:

- Identifiers and Spatial Attributes
  - `segment_id` uniquely identifies each record.
  - `latitude` and `longitude` are continuous numeric fields with 2,962 unique values, confirming that every segment has its own coordinate pair.
  - `city_name`, `admin_ward`, and `catchment_id` provide hierarchical geographic groupings with 63 cities, 91 wards, and 180 catchments represented.

- Topographic and Environmental Features

    - `elevation_m` is numeric with 2,229 unique values, ranging from sentinel values (-3.0) up to 266.7 m.
    - `slope_percent` (not shown in the table but present in earlier checks) provides terrain gradient information.
    - `soil_group` is categorical with 4 categories; missing values were filled with the modal group (“B”).
    - `land_use` is categorical with 9 categories, covering residential, industrial, green space, and others.

- Infrastructure Metrics
    - `drainage_density_km_per_km2` has 855 unique values, reflecting variability in stormwater infrastructure coverage.
    - `storm_drain_proximity_m` has 1,750 unique values, with a skew toward shorter distances but extending up to ~752 m.
    - `storm_drain_type` includes 5 categories after cleaning (`CurbInlet`, `Manhole`, `GratedInlet`, `OpenChannel`, and `Unknown`).

- Rainfall and Hazard Labels
    - `rainfall_source` is categorical with 5 categories, dominated by ERA5 reanalysis but supplemented by gauges, blended sources, IMD, and `"Unknown"`.
    - `historical_rainfall_intensity_mm_hr` is numeric with 866 unique values, ranging up to 150 mm/hr.
    - `return_period_years` is numeric (integer) with 6 distinct values, representing common hydrologic recurrence intervals (2–100 years).
    - `risk_labels` is categorical with 195 unique multi-value combinations (pipe-delimited), covering conditions like `ponding_hotspot`, `sparse_drainage`, and event tags (`event_YYYY-MM-DD`).

- Completeness
    - No missing values remain in the dataset (`n_missing = 0` for all variables).
    - All imputations were successful, making the dataset analysis-ready.

Descriptive Statistics of Variables

Numerical Variables

	latitude	longitude	elevation_m	drainage_density_km_per_km2	storm_drain_proximity_m	historical_rainfall_intensity_mm_hr	return_period_years
count	2962.000000	2962.000000	2962.000000	2962.000000	2962.000000	2962.000000	2962.000000
mean	19.392549	31.712840	37.594541	6.292750	123.072485	43.815057	19.734976
std	24.447844	79.530794	37.999568	2.110028	103.819179	25.224690	25.185680
min	-36.999038	-123.292949	-3.000000	1.270000	0.200000	5.400000	2.000000
25%	6.579197	-43.120010	9.240000	4.752500	51.125000	25.800000	5.000000
50%	23.760790	36.890411	26.170000	6.280000	100.450000	37.900000	10.000000
75%	37.886114	101.701955	58.047500	7.670000	160.650000	55.575000	25.000000
max	55.821219	174.911271	266.700000	12.070000	751.700000	150.000000	100.000000

Note: a larger view of this table can be found in the 6.1 Jupyter Notebook.

Spatial Coordinates

- `latitude`: spans from -37.0 to 55.8, reflecting the dataset's global coverage across both hemispheres.
- `longitude`: spans from -123.3 to 174.9, again confirming wide global distribution.

Elevation (`elevation_m`)

- Ranges from -3.0 (sentinel for below sea level or DEM gaps) to 266.7 m.
- Mean: 37.6 m; Median: 26.2 m → slightly right-skewed.
- Majority (75%) of values are below 58 m, highlighting a concentration in lower-elevation urban areas.

Drainage Density (`drainage_density_km_per_km2`)

- Values between 1.27 and 12.07, with a mean of 6.29.
- Median: 6.28, very close to the mean → relatively balanced distribution.
- 75% of observations are under 7.67, indicating most cities have moderate drainage coverage.

Storm Drain Proximity (`storm_drain_proximity_m`)

- Wide range: 0.2 m to 751.7 m.
- Mean: 123.1 m; Median: 100.5 m → distribution is **right-skewed** with some extreme outliers.
- 75% of values are under 160.7 m, showing most segments are within short distance of storm drains.

Historical Rainfall Intensity (`historical_rainfall_intensity_mm_hr`)

- Range: 5.4 to 150 mm/hr.
- Mean: 43.8 mm/hr; Median: 37.9 mm/hr.
- Upper quartile (75%) reaches 55.6 mm/hr, indicating frequent moderate-to-high rainfall events in the dataset.

Return Period (`return_period_years`)

- Discrete values from 2 to 100 years.
- Median: 10 years, Mean: ~19.7 years, showing skew toward higher return periods.
- The quartiles (5, 10, 25) reflect standard hydrologic recurrence intervals.

Categorical Variables

- `storm_drain_type`: Curb Inlet (28%), Manhole (25%), Grated Inlet (22%), Open Channel (19%), Unknown (6%).
- `rainfall_source`: ERA5 (45%), LocalGauge (17%), Blended (15%), IMD (12%), Unknown (11%).
- `Soil_group`: B (42%), C(28%), D (20%), A (10%)
- `Land_use`: was already complete and unchanged.

Ethical Considerations

The dataset is based on real data but because it was synthetically put together, there could be a level of privacy risks. The aggregation in geospatial visualization needs to be managed carefully to avoid presenting exact points as real “flood hotspots.” In addition, gaps in infrastructure or rainfall data may not reflect actual absence and therefore must be contextualized to avoid misinterpretation. The data cleaning steps I took offered justification for how I addressed this and the final statistics showed that my cleaning did not skew the data. Finally, equity considerations are important. Interpretations should avoid attributing risk to communities without acknowledging systemic infrastructure and planning gaps.

CareerFoundry Data Analytics
September 25, 2025
Charlotte Lin

Questions for Data Analysis

1. **Geospatial Visualization**
   - How do average drainage density and storm drain proximity vary across cities and wards, and where are the geographic clusters of low-infrastructure segments located?
   - Rationale: Answerable with `city_name`, `admin_ward`, and lat/long; Tableau and Python can both map these spatial differences.

2. **Infrastructure & Risk**
   - To what extent are sparse drainage and greater storm drain proximity associated with the likelihood of a segment being labeled a `ponding_hotspot`?
   - Rationale: Supports logistic regression in Python and visual hotspot comparisons in Tableau.

3. **Topography & Rainfall**
   - How do low elevation and high rainfall intensity interact to influence segments with `low_lying` or `extreme_rain_history` risk labels?
   - Rationale: Can be modeled with multiple regression or visualized with Tableau scatterplots colored by labels.

4. **Time-Series / Temporal Analysis**
   - Do segments tagged with specific flood event dates (`event_YYYY-MM-DD`) exhibit systematically higher historical rainfall intensity or longer return periods compared to untagged segments?
   - Rationale: Fits the “temporal” dimension in the brief, combining `risk_labels` with rainfall metrics.

5. **Composite Risk Index**
   - When combining elevation, drainage density, storm drain proximity, and rainfall intensity into a composite index, can clustering techniques identify the most vulnerable groups of segments?
   - Rationale: Supports Python clustering (k-means, hierarchical) and Tableau cluster plots or heatmaps.