CareerFoundry Data Analytics
September 30, 2025
Charlotte Lin

Achievement 6.2: Exploring Relationships

Urban Flood Risk Data: Global City Analysis 2025

Partial Answers from EDA

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?

- I found correlation between `drainage_density_km_per_km2` and `storm_drain_proximity_m` is weak (≈ -0.13) but directionally logical.
- My scatterplot confirmed variability, with denser networks generally closer to drains.
- This question is partially answered. Geographic mapping (e.g., Tableau by `city_name`, `admin_ward`) is still required to show clusters of sparse infrastructure.

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`?

- I found segments labeled `sparse_drainage` show lower density and higher proximity; `low_lying` corresponds to lower elevation.
- My categorical plots confirm expected physical differences.
- This question can be considered addressed and well-supported as EDA validates that the textual risk labels align with numeric measures.

3. **Topography & Rainfall**
    - How do low elevation and high rainfall intensity interact to influence segments with `low_lying` or `extreme_rain_history` risk labels?

- I found `extreme_rain_history` segments are tied to higher rainfall intensities and longer return periods; elevation correlation is weak overall (-0.18).
- The scatterplots showed the step-like bands in return periods, and categorical plots showed extreme_rain_history flags at higher intensities.
- This question is partially answered. Labels are valid, but combined elevation × rainfall effects need regression or interaction analysis.

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?

- The event tags are present but not yet parsed; intensity ↔ return period correlation (0.27) suggests flagged events may reflect extremes.
- I still need to split `risk_labels` into event vs. non-event groups.
- This question requires further analysis.

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?

- Some variables are skewed (esp. `storm_drain_proximity_m`, `historical_rainfall_intensity_mm_hr`), but categorical differences suggest clustering will yield meaningful subgroups once normalized.
- This question requires further analysis but EDA confirms feasibility.

From my EDA so far, here are the most promising strong relationships to explore further.

1. Rainfall Source & Rainfall Intensity

- Already strong patterns: ERA5 has systematically higher rainfall intensity distributions than gauges or blended sources.
- Extra visual: **violin plot** (shows distribution shape, not just box).

2. Storm Drain Type & Storm Drain Proximity

- Strong differences as open channels much farther away than inlets/manholes.
- Extra visual: **strip plot or swarm plot** overlaid on boxplot; shows raw density of values.
- Tells a clear infrastructure story.

3. Land Use & Drainage Density

- Industrial and residential segments consistently have denser drainage than green/water segments.
- Extra visual: **bar plot of means ± error bars**; may communicate group differences more directly than boxplots.

4. Soil Group & Elevation

- Clear stratification: Soil Group D higher elevation, A lower elevation.

- Extra visual: **grouped histogram** or **facet plots by soil group** to see how elevation distributions differ.

5. Risk Labels (flagged) & Numeric Variables

These relationships still need to be plotted:

- Segments tagged `low_lying` vs. not → compare elevation distributions.
- Segments tagged `sparse_drainage` vs. not → compare drainage density/proximity.
- Segments tagged `extreme_rain_history` vs. not → compare rainfall intensity/return period.
- Extra visuals: boxplots or violin plots for each numeric variable split by these tags.

Adjusted & New Questions

1. **Geospatial:** *How do drainage density and storm drain proximity vary across cities and land uses, and which geographic areas show the highest concentration of sparse infrastructure?*
2. **Infrastructure & Risk:** How do physical measures of drainage density and storm drain proximity compare between segments labeled as `sparse_drainage` and those not, and how does elevation distinguish `low_lying` from non-labeled segments?
3. **Topography & Rainfall:** Do segments with `extreme_rain_history` labels show systematically higher rainfall intensities and return periods, and how does elevation contribute to differentiating these segments?
4. **Temporal & Events:** How do segments with event tags (`event_YYYY-MM-DD`) differ from other segments in terms of rainfall intensity and return period distributions?
5. **Composite Risk Index:** Can clustering on elevation, drainage density, storm drain proximity, and rainfall intensity reveal distinct segment profiles (e.g., low-lying with sparse drainage vs. high-elevation with dense drainage), and how are these distributed by land use and city?
6. How do differences in **rainfall source** (ERA5 vs. gauges vs. blended) systematically affect intensity measures, and should these be controlled in risk modeling?
7. How do **land use categories** stratify infrastructure adequacy (e.g., industrial vs. green areas)?
8. Do `Unknown` categories in `storm_drain_type` or `rainfall_source` behave as true gaps, or do they skew analysis and require exclusion/imputation?

Hypotheses

H1: Segments with higher drainage density have systematically shorter storm drain proximities.

H2: Segments labeled `sparse_drainage` will have significantly lower drainage density and greater storm drain proximity than non-labeled segments.

H3: Segments labeled `low_lying` will have significantly lower elevation than non-labeled segments.

H4: Segments labeled `extreme_rain_history` will show significantly higher rainfall intensities and longer return periods than unflagged segments.

H5: Segments with event tags (`event_YYYY-MM-DD`) will show systematically higher rainfall intensities and longer return periods compared to untagged segments.

H6: A composite index combining elevation, drainage density, storm drain proximity, and rainfall intensity will produce distinct clusters, differentiating vulnerable low-lying, sparse-drainage segments from more resilient, high-density ones.