Flood Disaster Data Analysis Report

1. Dataset Description

1.1 Source

The dataset used for this project is urban_pluvial_flood_risk_dataset.csv, sourced from a geospatial hydrology dataset collection.
It contains information about 2,963 urban segments across multiple global cities, focusing on flood risk indicators such as elevation, land use, soil type, and storm drainage.

1.2 Columns

The dataset includes 17 attributes, categorized as geographic, hydrologic, infrastructural, and environmental:

- segment_id: Unique segment identifier
- city_name: City in which the segment is located
- admin_ward: Local administrative division
- latitude, longitude: Geographic coordinates
- catchment_id: Catchment area code
- elevation_m: Elevation in meters
- dem_source: Source of elevation model
- land_use: Type of land use (Residential, Roads, Industrial, etc.)
- soil_group: Soil classification (A–D based on infiltration)
- drainage_density_km_per_km2: Drainage coverage ratio
- storm_drain_proximity_m: Distance to nearest storm drain
- storm_drain_type: Type of drainage (CurbInlet, Manhole, Grated, etc.)
- rainfall_source: Source of rainfall data
- historical_rainfall_intensity_mm_hr: Average rainfall intensity (mm/hr)
- return_period_years: Statistical flood recurrence period
- risk_labels: Flood risk category (monitor, low_lying, extreme_rain_history, etc.)

1.3 Data Quality

The dataset is clean and consistent, with minimal missing values.
Numeric columns (elevation, rainfall intensity, drainage density) are within realistic ranges, and categorical attributes have uniform naming.
It is suitable for machine learning classification and geospatial flood analysis.

2. Operations Performed

- Initialized a PySpark Session and loaded the dataset using Spark’s read.csv() function.
- Verified schema and previewed records using .printSchema() and .show().
- Handled missing values using fillna() and string replacement for nulls.
- Filtered records to analyze high-risk zones, low-lying areas, and drainage deficiencies.
- Computed summary statistics (mean, min, max) for hydrologic parameters.
- Grouped data by soil_group, land_use, and storm_drain_type to explore flood risk patterns.
- Generated visualizations including bar charts, pie charts, scatter plots, and a correlation heatmap.
- Classified and interpreted risk_labels to identify critical flood-prone regions.

3. Key Insights

- Elevation: Ranges from -3 m to 266.7 m, with low-lying zones (<10 m) showing higher flood vulnerability.
- Soil Type: Groups B and C dominate the dataset (1,460+ records), while Group D soils (low infiltration) contribute most to flood risk.
- Land Use: Residential (827) and Roads (599) make up nearly 50% of flood-prone areas.
- Risk Labels:
    - Monitor– 67% of records (general observation zones)
    - Low-lying – 13%
    - Extreme rain history – 5%
    - Ponding hotspots – 2%

Patterns Observed

- Low elevation + Soil D + No storm drain → high flood susceptibility.
- Commercial and road zones experience repetitive flood events.
- Storm drain type “None” strongly correlates with high-risk labels.

4. Recommendations

- Infrastructure Improvement:
  Develop drainage in the 178 unserved zones with no storm drains.
  Prioritize open channels and grated inlets in high-risk districts.
- Urban Planning:
  Limit construction in low-lying residential and commercial zones.
  Encourage use of permeable pavements to increase infiltration.
- Flood Monitoring:
  Focus continuous monitoring on “monitor” and “low_lying” zones with rainfall > 80 mm/hr.
- Data Utilization:
  Integrate this dataset with satellite rainfall and soil moisture indices for early flood prediction.

5. Conclusion

The Urban Pluvial Flood Risk dataset provides a comprehensive view of how geography, land use, and infrastructure interact to shape urban flood risks.
Findings highlight that poor drainage, impervious land use, and low elevation are the dominant flood drivers.
Cities such as Manila, San Francisco, and Philadelphia show recurring flood risk clusters.
These insights can guide policy decisions, city planning, and resilient infrastructure development to mitigate urban flooding.