

DATASET for type DATA

The dataset is composed of three distinct files, which aggregate processed data derived from open datasets of three cities: Dublin, San Francisco, and Valencia. The data has been mapped to a grid of 25x25m for Valencia and 50x50m for Dublin and San Francisco. The respective files are named DATA_ES_VLC.csv, DATA_IE_DUB.csv, and DATA_US_SFO.csv. The following sections provide the data dictionaries for each file.

Data Dictionary for “DATA_ES_VLC.csv”:

NAME	DESCRIPTION	DATA TYPE
VLC-ID	Unique identifier for each record	Integer
VLC-GEOM	Geometry of the area, likely in GeoJSON format (e.g., polygon coordinates)	Text
VLC-D/N	Indicator of day or night	Integer (1: Day, 0: Night)
VLC-HOSP	Represents the distance to the nearest hospital with 24-hour service	Float (may contain null values)
VLC-HOSP-24	Represents the distance to the nearest hospital	Float (may contain null values)
VLC-PS	Represents the distance to the Police stations. Values indicate the distance	Float (may contain null values)
VLC-RP	Average rent price	Float
VLC-AMI	Average monthly income	Float
VLC-TI	Represents traffic intensity and distance to the nearest street or road. The first value indicates the traffic intensity, and the second value indicates the distance to the nearest street or road	Text (requires data processing for numerical analysis)
VLC-SP	Sale price	Float
VLC-%-UNEMP	Percentage of unemployed individuals	Float
VLC-%-LOW-ED	Percentage of individuals with low educational attainment	Float
VLC-%-ADULT-POP	Percentage of adult population	Float
VLC-GR.AREAS	Represents the distance to the nearest green areas	Float (may contain null values)
VLC-C/P	Represents the distance to the nearest clubs and pubs	Float (may contain null values)
VLC-RE-CENTERS	Represents the distance to the nearest reintegration centres	Float (may contain null values)

Data Dictionary for “DATA_IE_DUB.csv”:

NAME	DESCRIPTION	DATA TYPE
DUB-ID	Unique identifier for each record	Integer
DUB-GEOM	Geometry of the area, likely in GeoJSON format (e.g., polygon coordinates)	Text
DUB-D/N	Indicator of day or night	Integer (1: Day, 0: Night)
DUB-HOSP	Represents the distance to the nearest hospital with 24-hour service	Float (may contain null values)
DUB-HOSP-24	Represents the distance to the nearest hospital	Float (may contain null values)
DUB-PS	Represents the distance to the Police stations. Values indicate the distance	Float (may contain null values)
DUB-AMI	Average monthly income	Float
DUB-TI	Represents traffic intensity and distance to the nearest street or road. The first value indicates the traffic intensity, and the second value indicates the distance to the nearest street or road	Text (requires data processing for numerical analysis)
DUB-SP	Sale price	Float
DUB-%-UNEMP	Percentage of unemployed	Float
VLX-%-LOW-ED	Percentage of individuals with low educational attainment	Float
DUB-%-ADULT-POP	Percentage of adult population	Float
DUB-GR.AREAS	Represents the distance to the nearest green areas	Float (may contain null values)
DUB-C/P	Represents the distance to the nearest clubs and pubs	Float (may contain null values)

Data Dictionary for “DATA_US_SFO.csv”:

NAME	DESCRIPTION	DATA TYPE
SFO-ID	Unique identifier for each record	Integer
SFO-GEOM	Geometry of the area, likely in GeoJSON format (e.g., polygon coordinates)	Text
SFO-D/N	Indicator of day or night	Integer (1: Day, 0: Night)
SFO-HOSP	Represents the distance to the nearest hospital with 24-hour service	Float (may contain null values)
SFO-HOSP-24	Represents the distance to the nearest hospital	Float (may contain null values)
SFO-PS	Represents the distance to the Police stations. Values indicate the distance	Float (may contain null values)
SFO-RP	Average rent price	Float
SFO-TI	Represents traffic intensity and distance to the nearest street or road. The first value indicates the traffic intensity, and the second value indicates the distance to the nearest street or road	Text (requires data processing for numerical analysis)
SFO-SP	Sale price	Float
SFO-%-UNEMP	Percentage of unemployed individuals	Float
SFO-%-ADULT-POP	Percentage of adult population	Float
SFO-GR.AREAS	Represents the distance to the nearest green areas	Float (may contain null values)
SFO-C/P	Represents the distance to the nearest clubs and pubs	Float (may contain null values)

DATASET for type TWT

This dataset contains tweets collected through web scraping and analyzed using natural language processing (NLP) algorithms and neural networks. The aim is to identify and classify tweets that discuss gender-based violence or they have some component of feelings related to violence or fear in the cities of Valencia and Toluca. Below is a description of each column in the dataset.

Data Dictionary for TWT_ES_VLC.csv

NAME	DESCRIPTION	DATA TYPE
VLC-USER	Unique identifier for each user	Text
VLC-TIME	Timestamp of the tweet, including date and time	Datetime
VLC-TWEETS	Text content of the tweet	Text
VLC-REPLY	Number of replies to the tweet	Float (may contain null values)
VLC-RETWEETS	Number of retweets	Float (may contain null values)
VLC-LIKES	Number of likes on the tweet	Float
VLC-LAT	Latitude associated with the tweet's location	Float
VLC-LON	Longitude associated with the tweet's location	Float
VLC-CLASIF	Classification of the tweet regarding gender-based violence (1: Related, 0: Not related)	Integer
VLC-PROB	Probability that the tweet is related to gender-based violence (calculated by the NLP model)	Float
VLC-VEC	Feature vector generated from processing the tweet's text	Float

Data Dictionary for TWT_MX_TOL.csv

NAME	DESCRIPTION	DATA TYPE
TOL-USER	Unique identifier for each user	Text
TOL-TIME	Timestamp of the tweet, including date and time	Datetime
TOL-TWEETS	Text content of the tweet	Text
TOL-REPLYS	Number of replies to the tweet	Float (may contain null values)
TOL-RETWEETS	Number of retweets	Float (may contain null values)
TOL-LIKES	Number of likes on the tweet	Float
TOL-LAT	Latitude associated with the tweet's location	Float
TOL-LON	Longitude associated with the tweet's location	Float
TOL-CLASIF	Classification of the tweet regarding gender-based violence (1: Related, 0: Not related)	Integer
TOL-PROB	Probability that the tweet is related to gender-based violence (calculated by the NLP model)	Float
TOL-VEC	Feature vector generated from processing the tweet's text	Float

DATASET for type MAP

The dataset presented here results from participatory mapping activities carried out in the cities of Toluca and Valencia. The description of each column allows the anonymized user to identify the point in coordinates where he/she specifies the perception of insecurity, harassment, or both. The dataset includes the date of participation.

Data Dictionary for MAP_ES_VLC.csv

NAME	DESCRIPTION	DATA TYPE
VLC-ID	Unique identifier for each record.	Integer
VLC-COD-PART	Participant code.	Integer
VLC-INCID	Type of reported incident (e.g., Harassment, Insecurity, Both).	Text
VLC-OBS-INS	Additional observations about insecurity.	Text
VLC-DATE	Date of the report.	Date
VLC-LON	Geographic longitude of the reported location.	Float
VLC-LAT	Geographic latitude of the reported location.	Float
VLC-OBS-HAR	Additional observations about street harassment .	Text

Data Dictionary for MAP_MX_TOL.csv

NAME	DESCRIPTION	DATA TYPE
TOL-ID	Unique identifier for each record.	Integer
TOL-COD-PART	Participant code.	Integer
TOL-INCID	Type of reported incident (e.g., Harassment, Insecurity, Both).	Text
TOL-OBS-INS	Additional observations about insecurity.	Date
TOL-DATE	Date of the report.	Text
TOL-LON	Geographic longitude of the reported location.	Float
TOL-LAT	Geographic latitude of the reported location.	Float
TOL-OBS-HAR	Additional observations about street harassment .	Text