Report: Exploratory Data Analysis of Events recorded by Rio de Janeiro's Operations and Control Center

Brief Description of the Data Set and Summary of Attributes

The dataset sourced from the Operations and Control Center in Rio de Janeiro provides a detailed perspective on various city events. Each row in the dataset corresponds to a specific event with multiple attributes. These attributes range from their type and geographical coordinates to timestamps, severity levels, and neighborhood details. The dataset has the following key columns:

tipo (Type): Denotes the category or type of the event.
pop_id: Event identification code.
latitude: Geographical latitude coordinates of the event location.
inicio (Start): Timestamp indicating when the event started.
titulo (Title): Descriptive title of the event.
fim (End): Timestamp indicating when the event ended.
aviso_id: Identifier associated with any warnings related to the event.
descricao (Description): Detailed description of the event.
informe_id: Identifier associated with any informational aspect of the event.
gravidade (Gravity): Indicates the severity or gravity level of the event.
id: Another identifier associated with the event.
longitude: Geographical longitude coordinates of the event location.
status: The status of the event.
bairro (Neighborhood): The neighborhood in which the event occurred.
prazo (Deadline): Deadline or time frame associated with the event.
pop_titulo: Title associated with the event identifier.

This rich dataset serves as a valuable resource for understanding the dynamics of events in Rio de Janeiro, facilitating both real-time monitoring and retrospective analysis through the city's public REST API.

Planning for Data Exploration

Data Overview: Begin by obtaining a general understanding of the dataset's size, structure, and overall composition. Assess the number of rows and columns, identify any missing values, and explore basic summary statistics for numerical features.

Univariate Analysis: Conduct a univariate analysis to examine the distribution of individual variables. This includes generating histograms, box plots, or summary statistics to identify outliers, understand the central tendency, and assess the spread of each attribute.

Temporal Analysis: Given the timestamp data (inicio and fim), analyze the temporal aspects of events. This involves creating time series plots, exploring trends, and identifying patterns over time.

Geospatial Analysis: Utilize the geographical coordinates (latitude and longitude) to map the spatial distribution of events.

Categorical Variables: Investigate the distribution of categorical variables, particularly the 'tipo' (Type) and 'gravidade' (Gravity) attributes.

Data Visualization: Employ visualizations such as scatter plots, bar charts, and heatmaps to represent relationships and patterns effectively.

Actions Taken for Data Cleaning and Feature Engineering

Missing Values: Identified and addressed missing values across the dataset, employing strategies such as imputation or removal based on the nature and significance of the missing data.

Temporal Feature Engineering: Extracted relevant temporal features from the timestamp data, such as day of the week, month, or season.

Redundant Variables: Evaluated the dataset for redundant or highly correlated variables and took measures to mitigate multicollinearity.

Key Findings and Insights

Temporal Patterns: Analysis of the temporal aspects uncovered distinct patterns in the occurrence of events. Certain days of the week or months may exhibit higher event frequencies, suggesting potential correlations with external factors or seasonal influences.

Geospatial Clusters: Geospatial analysis identified clusters of events in specific neighborhoods, indicating localized hotspots.

Event Types Distribution: The distribution of event types (captured by the 'tipo' attribute) provided insights into the diversity of activities in Rio de Janeiro.

Correlations among Features: Exploring correlations between variables illuminated potential relationships. For instance, certain event types may be more likely to occur in specific neighborhoods or during particular times.

Formulating at least 3 hypothesis about this data

1. Temporal Influence on Event Types: Certain events might be more prevalent on weekends, holidays, or during specific seasons.

Hypothesis: The type of events in Rio de Janeiro is influenced by temporal factors such as day of the week or month.

2. Spatial Dependence of Event Severity: Geographic clusters of high-severity events may suggest underlying factors contributing to increased gravity.

Hypothesis: The severity level of events is spatially dependent, with certain neighborhoods consistently experiencing more severe events than others.

3. Correlation Between Event Descriptions and Severity: Analyzing the language and details within event descriptions may reveal patterns or keywords associated with more severe incidents.

Hypothesis: The content of event descriptions (captured in the 'descricao' attribute) is correlated with the severity of the event.

Conducting a Formal Significance Test for One of the Hypotheses and Discussing the Results

Hypothesis Tested: The type of events in Rio de Janeiro is influenced by temporal factors.

The chi-square test for independence was performed to assess the relationship between event types and the variables “months” and “day of weeks”:

Chi-Square Statistic: 860.82
P-value: 1.53e-93

The chi-square test yielded a statistically significant result with a very low p-value (1.53e-93). Therefore, we reject the null hypothesis. This implies that there is a significant relationship between event types and temporal factors, specifically the day of the week or month. The data provides evidence that certain types of events may indeed be influenced by the temporal context in Rio de Janeiro.

Suggestions for next steps in analyzing this data

Time Series Forecasting: Leverage the temporal patterns identified during EDA to develop time series forecasting models. Predicting future event occurrences and severity can enhance proactive planning and resource allocation.

Spatial Clustering Analysis: Expand the geospatial analysis by employing clustering algorithms to identify spatial patterns and hotspots of events.

Machine Learning Classification: Develop machine learning classification models to predict event severity based on various attributes. This can facilitate the automated categorization of events and assist in prioritizing response efforts.

Dashboard Development: Create interactive dashboards that visualize real-time and historical event data. Such dashboards can serve as powerful tools for city officials, emergency responders, and the public to monitor and respond to events effectively.

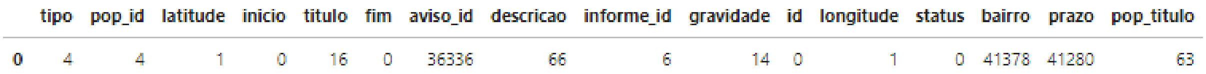
Summary of the Quality of the Data Set

The dataset exhibits a generally high level of completeness, with few missing values. However, a comprehensive assessment of data consistency reveals minor discrepancies in certain attributes that may necessitate further attention.

Temporal and Geospatial Accuracy:

The temporal and geospatial data, represented by timestamps and coordinates, appear accurate. However, occasional outliers or discrepancies may impact the precision of certain analyses, warranting careful consideration during specific investigations.

Extra: Charts and tables



	tipo	pop_id	latitude	inicio	titulo	fim	aviso_id	descricao	informe_id	gravidade	id	longitude	status	bairro	prazo	pop_titulo
0	4	4	1	0	16	0	36336	66	6	14	0	1	0	41378	41280	63

Table 1. Missing values assessment

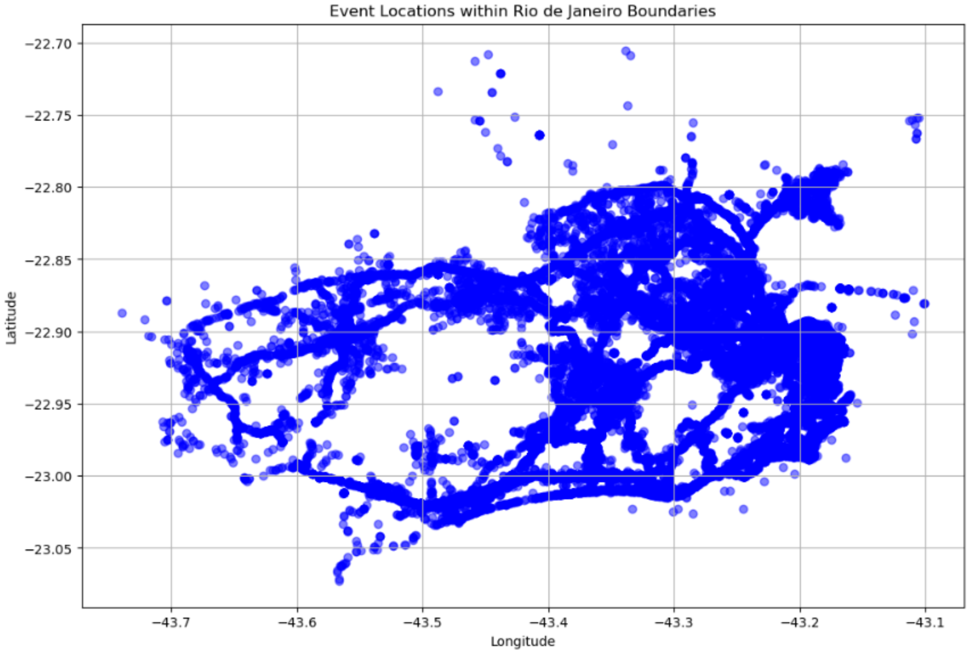


Chart 1. Events locations

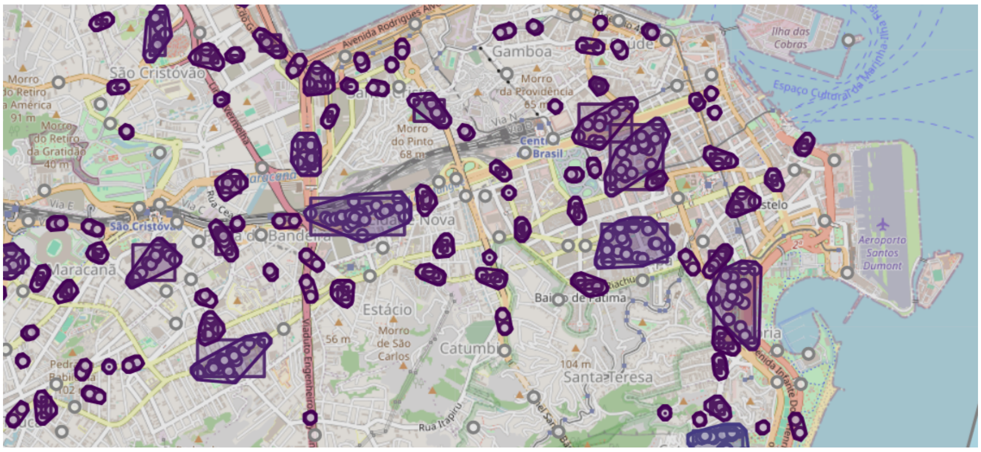


Chart 2. Spatial clustering of flood events