Technical report – Data Journalism project

Letizia Monti

Subject of the article

The article is about the accidents that happened in the mountains of Trentino during 2019 and the exploration of some features such as the causes and the seriousness of the accidents, the activities done when the accidents occurred and the sex of the rescued people. Also, the variation of the accidents over a range of time of a year was considered. The aim was to provide information on the most frequent kind of accidents in the mountains, the period with more injured people and to discover if there exist types of accidents more associated to serious injuries.

Data

I chose to consider the data on the activity of the *Soccorso Alpino Trentino*, the organization that deals with the rescue and the recovery of injured people in the mountain territory, in the underground environment and in the arduous areas. This kind of data are not open but have to be requested to the organization. In my case, the data were provided after making a request to the press office of the *Soccorso Alpino Trentino*, specifying the purpose and the kind of data needed.

Data cleaning

Both the data cleaning and data analysis operations are reported in the attached coding file.

For the purpose of my analysis I performed some data cleaning operations to check for duplicated rows and missing values. Since the original data were on the operations of the *Soccorso Alpino Trentino*, it was very common that for the same accident more than one *stazione* on the territory was involved and so there were recorded two or more operations of different *stazioni* for the same person. For this reason, I decided to create a reduced dataset that had a unique injured individual per row.

The values of the column relative to the physical state of the injured people was modified substituting numbers with their labels, e.g. 0 with *illesi*, 1 with *feriti leggeri* etc. A new variable relative to the mountain community in which injuries happened was added, according to the reported place of the accident.

Data analysis

Descriptive analysis was performed to get absolute and relative frequencies of the variables of interest: sex and physical state of injured people, causes and activities during the accident, helicopter intervention, date, and place of the accidents.

The relationship between the causes of the accidents and the physical state of the injured people and the activities done during the accidents and the physical state were explored to see if there were some causes or activities more associated with certain physical states than others. In this case, the activities and the causes with less than 10 cases were not considered.

The distribution of the accidents over time was analysed by grouping it by week and by distinguishing helicopter rescue and land teams operations. In particular, the kind of accidents that required the helicopter intervention were further analysed to get how they were spatially distributed and to what kind of injuries they were related.

Finally, the distribution of the accidents over the month of the year and the mountain community was explored to discover which were the most frequented areas of Trentino depending on the period of the year. Supposing a relationship between seasonal injuries and tourism, the data regarding the arrival of foreigners by mountain community were related to the number of foreigners accidents by mountain community to see if there most tourist areas were also the ones with more injuries.

Data visualization

It was performed with the Python library *Plotly*. The following types of plot were used:

- *Treemap*: to visually represent the distribution of the nationality of the foreign injured people, the activities done when the accidents occurred and the causes of the accidents. Having each category a shape size depending on its frequency allow the user to immediately understand which are the most common foreign nationalities, activities or causes of the accidents. A palette of different colours was used to have as much contrast as possible. Also, each shape is provided with the category name and the relative frequency to help the user to distinguish between shapes of similar size.
- *Line chart*: to represent the number of accidents, of helicopter rescue operations and land teams operations over time and mountain community. It is the best way to represent time-series data. Hovering on the line gives the dates corresponding to the week number in the x-axis.

- *Heatmap*: to visualize the relationship between activities and physical state, and causes and physical state. Thanks to the continuous colour palette, the user can immediately see which are the most frequent associations of variables, depending on the colour's tone.
- *Choropleth map*: to represent the number of accidents, the number of helicopter interventions and the number of foreigners' injuries by mountain community. A continuous colour palette is used to simplify the comparison between different areas. Hovering text is provided with mountain communities names and frequencies.

For completeness, all the plots are provided with additional information, i.e. absolute and relative frequencies, when hovering over data points.

In addition, two tables were provided to gather data of the physical state of injured people and of the arrivals of tourists in 2019 in Trentino's mountain communities. They were used to enrich provided graphical representations by offering numerical information about what described in the article.