Final Assignment Report

A case of study on Italian Cities and Traffic Risk Analysis

Riccardo Pollo

May 21, 2021

1 Introduction and Problem definition

Italian peninsula is a beautiful place for spending vacation. There are lot of interesting venues to visit, lot of history, lot of thinks to do. Not only culture: in Italy there are lot of mountain in which it is possible to hiking, beaches where sunbathing and entertainment.
Suppose to be a travel agency and we need to reply to our clients about some questions about traffic in Italy. Our clients want to know more about drive risks on Italian roads.
Once we find some low-risk-cities, clients will be able to decide in which city spend their vacations.
Some of them will go by car, and we can suggest them to not only stay in the city they will choose, but also move to the neighbour cities if any. As data scientist we don't have any specific goal. Instead, it is asked to acquire information about risks create a report to explain the traffic situation in Italy to our clients, and they will decide autonomously based on our informations.
It is also asked to suggest cities according to the risks and their venues.
We scheduled some points to follow in order to successfully carry out the analysis.

- First, let's retrieve informations about provinces and drive licences per province.
- Let see how many people can drive, according to the population of each city and the total amount of active drive licenses.
- Retrieve informations about incidents in last years.
- Plot a line chart on the total number of incident cases in last year (from 2004 to 2018) per region.
- Which is the region with the larger number of incident from 2004 to 2018? We will plot a bar chart.
- Since we want to go "now", let's check the result according to the last year available (2018)
- Is there any correlation between number of incident and the number of drive licence over population density? We can check with a simple Pearson Correlation Coefficient.

- Now we want to see if some province are near to one other, in order to suggest to visit by car some near cities or venues.
- Suppose we want to divide Italy in three clusters: minimum risk, maximum risk, moderate risk. Let's plot it with *folium* library.
- Suppose we choose a couple of cities between which we need to choose our destination. to understand which kind of behaviour the population can have according to the entertainment, we want to exploit *Fourstroke* API which are the most popular category of venue for these cities.

2 Data Acquisition and Data Cleaning

In order to acquire data we explored lot of Italian government sites and some others. We have found lot of dataset, and after some comparison on which was the best ones, we decides on the following.

Provinces in Italy
`https://it.wikipedia.org/wiki/Province_d%27Italia`

List of active drive licences per region
`http://dati.mit.gov.it/catalog/dataset/patenti`

Incidents locations from 2004 to 2018
http://dati.mit.gov.it/catalog/dataset/localizzazione-incidenti-stradali-anni-2004-2018

In order to retrieve information about Provinces in Italy we found a Wikipedia page where Provinces are listed and for each Provinces some information are given in a table. From this table we selected as feature:

- Province Name
- Region
- Population number
- Area, in Km^2
- Density of population

We perform webscraping exploiting *BeautifulSoup* library to make it in an easier way and we collapse these informations in a *pandas* dataframe. Unfortunately we don't have coordinates for Provinces in this table, so we need to use *GeoPy* library to retrieve latitude and longitude for each province. Then we merge them into the previous created dataframe.
In order to obtain a count of the number of active drive licences we use the founded dataset that contains a *.csv* file for each Region, and inside this, each row represents a licence with the Province in which it has been activated, the Licence Type and the date of activation. We are not interested in this features, but we are interested in counting how many active "B" type licences each province has. So we first filter on "B" type, then we count licences for each provinces, and again we will join this information (*Number of Driver*) on the previous dataframe. We also add the percentage of driver over the total population in each province. Unfortunately for Palermo percentage is not correct, so we need to drop this city and not consider it.
The last dataset we use is the one regarding incidents in Italy from 2004 to 2008. It is a *.xlsx* file containing a table in which for each province, for each road, the number of incident is indicated for each year (one year per column). Again we need to sum incident per province, forgetting about count per road. Even if this dataset contains other informations we will only use these one.
Now all informations we need are ready to be used. We can perform our *Exploratory Data Analysis* tasks.