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.

3 Exploratory Data Analysis

3.1 Top 5 provinces with lowest percentage of driver over the population

If we are interested to visit a city in which there is a low amount of risks regarding traffic, the first very basically approach could be search for cities with low number of drivers.
As explained we have a dataset from which we can count the total amount of activated drive licences in each Province.
Once we load these data, and we briefly perform so cleaning (for example keeping only B type, which are the car type in Italy) and transformations in order to get the count and then the percentage of driver over the all population.

	Province	% of driver
27	crotone	32.08
18	caltanissetta	34.31
29	enna	34.37
0	agrigento	34.39
21	catania	34.85

Figure 1: Top 5 Provinces with lower percentage of driver over the total population

Obviously, this approach gives an idea of which are the areas with less usage of cars. For example, even if a Province activate a large number of drive licences, does not mean that driver are still in this Province, and vice versa cities with low amount of drive licences could be not directly translated in a few number of driver (just think about small cities but with high number of point of interest).

3.2 Linechart with number of incident in last years, divided per region

Since we understand number of licences per Province is not enough: we can do better!
We can exploit data on incident in last years, from 2004 to 2018.
We can for example count the number of incidents happened for each year, in each region, creating a very explanatory pivot table.

REGIONE	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Abruzzo	1744	1528	1532	1373	1193	1350	1447	1274	1157	1154	1109	1036	1020	955	1035
Basilicata	279	282	276	263	239	293	373	320	280	300	284	285	283	250	331
Calabria	1521	1331	1352	1300	1268	1302	1344	1170	1120	1166	1083	1112	1222	1102	1234
Campania	2683	2750	2725	2841	2797	2747	2666	2253	2146	2048	2012	2048	2074	2072	2050
Emilia Romagna	5825	5624	5440	5504	4852	4698	5025	4424	4056	4020	3658	3772	3722	3731	3617
Friuli-Venezia Giulia	1267	1229	1249	1145	1057	1090	999	893	817	788	788	835	819	879	838
Lazio	4969	5149	5080	5026	5007	5333	5260	4797	4543	4323	4177	4177	4026	3875	3951
Liguria	2017	2179	2154	2023	1886	2042	2120	1962	1923	1925	1795	1604	1549	1644	1593
Lombardia	7538	7263	7199	7173	6658	6397	6644	5942	5737	5504	5624	5786	5789	5889	5992
Marche	1953	1898	2000	1725	1619	1534	1755	1564	1367	1145	1186	1157	1111	1238	1230
Molise	227	190	198	175	210	205	252	233	220	182	169	162	189	181	176
Piemonte	3765	3674	3505	3329	2931	2988	3173	3222	2878	2476	2599	2524	2561	2454	2575
Puglia	1749	1733	1970	1790	1841	1950	2061	1856	1621	1545	1514	1449	1613	1626	1549
Sardegna	918	895	965	851	805	941	906	819	793	835	821	850	808	774	870
Sicilia	2207	2099	2085	2077	2094	2407	2199	2165	1933	1925	1758	1693	1718	1725	1871
Toscana	3421	3428	3274	3227	3448	3509	3821	3568	3048	2961	3172	3057	3259	3086	3108
Trentino-Alto Adige	1286	1408	1411	1369	1154	1057	1045	1036	1186	1144	1142	1176	1169	1079	1170
Umbria	1041	1109	1116	1154	1086	1052	953	855	706	748	720	679	694	694	772
Valle d'Aosta	176	176	175	163	113	129	134	125	132	118	130	135	115	112	121
Veneto	4677	4497	4552	4548	4033	3662	3934	3667	3267	3013	3276	3180	3160	3153	3123

Figure 2: Number of incidents per region during last years

This table is not easily readable and understandable.
We decided to plot a line chart in which each line report number of incidents during years for a specific region.

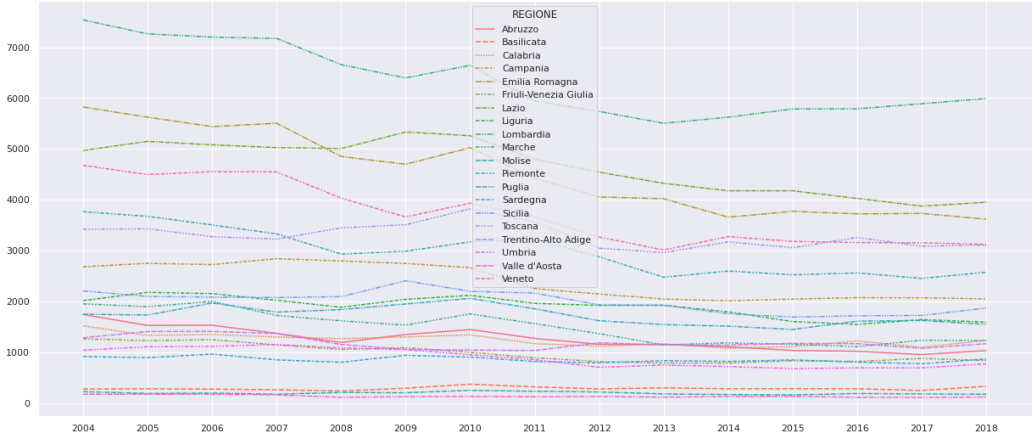

Figure 3: Number of incidents per region during last years

This plot is very interesting, because it is suggesting that for all regions, the incidents trend is decreasing. This could be thank to technologies on vehicles (brake, sensors etc.), to a more careful population, or more skilled traffic engineers.
There are three region with a low number of incidents, and this low value has been maintained during last years: Umbria, Molise and Basilicata.

3.3 Barchart showing average count of incidents per region

Go ahead, and we are interesting to see the average (according to the years) number of incident in each region. So, for each region, we evaluate the formula:

$$avg_{inc} = count(incident)/(2018 - 2004)$$

These values are plotted in bar chart in figure 3.3.

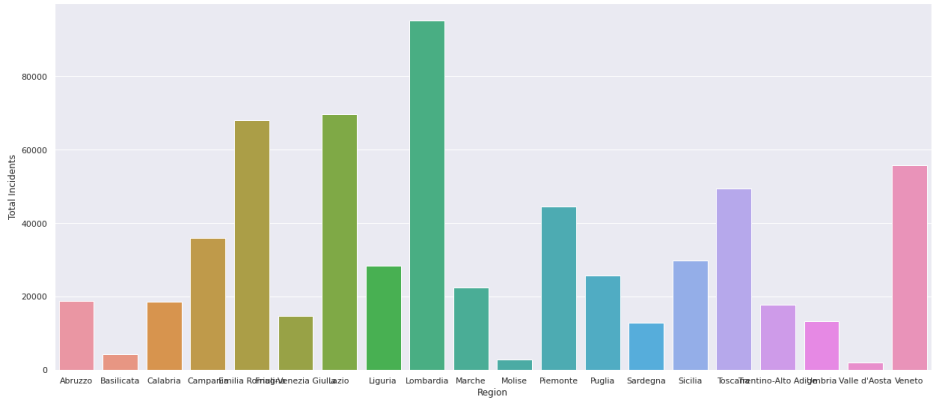

Figure 4: Average number o incident per year in each Region.

The Region with the highest number of incident is Lombardia, followed by Lazio and Emilia Romagna.
Is is true to conclude that this are the Region in which we have the largest number of incidents?
Probably this is true, but these are not the dangerous ones: these are also the bigger Regions in Italy.
What if we divide the formula for the area of each Region? We will obtain the bar chart in figure 3.3.

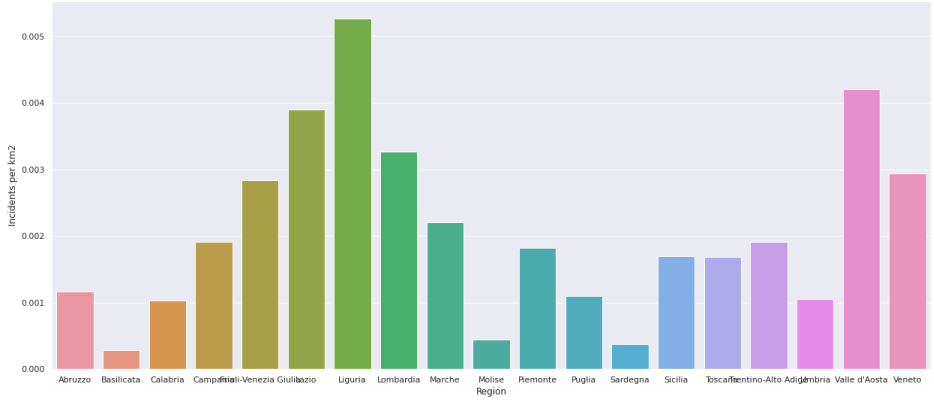

Figure 5: Average number o incident per year in each Region taking into account the area.

The conclusion changes: now Liguria is the "most" dangerous Region. This is confirmed by our knowledge on the domain. This region is a very popular tourist destination. Highway has lot of curves, and a famous internal road (called "via Aurelia") guests lot of traffic, especially during summer, with lot of blind curves and it has lot of incidents every year.
Molise and Basilicata confirm their position we have seen from the line chart above. Sardegna can be added to the selected "low risk Regions", while Umbria is a little higher, similar to Abruzzo.

3.4 Comparing previous average bar chart with 2018 values

What if we zoom in considenring only 2018?
Does the average value found has been respected?

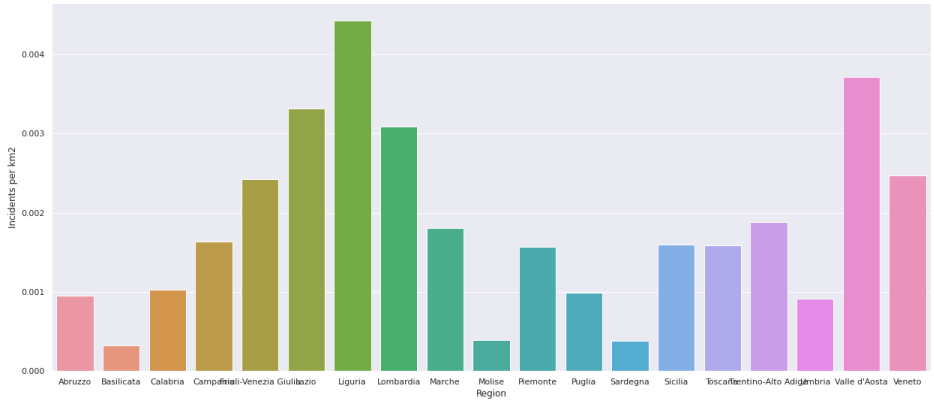

Figure 6: Average number o incident in 2018 in each Region taking into account the area.

Plotting the same bar chart filtering on 2018 measurements.
Resulting bar chart seems to suggest that mean values persists. Liguria is the most dangerous, Basilicata, Molise, and Sardegna the least.

3.5 Correlation between the number of drive licences per density of population and the number of incident in a Region in 2018

Our intuition suggest that the higher the percentage of driver over the all population is, the highest the number of incidents in a Region is. So, our hypothesis is that a correlation exists between these measure.
We still consider only data regarding 2018, evaluating the Pearson Correlation Coefficient between the two measures.
Results is -0.58.
It seems that a negative correlation exists for those two variables. This is not very intuitive. We need to go deep and analyse what it means. Density is the number of people over a certain Area. If Area increase, density decrease. We can suppose that some region has an higher area, so density decrease. If density decrease, drive licences over density increase. But if number of people doesn't change, number of drive licences should be basically the same as number of incidents, while a bigger area means an higher density. That's what probably this negative correlation could be explained.
This results suggest that, using our data, we must be careful on considering the number of drive licences, because seems not to be an meaningful feature.

3.6 Get nearest provinces for each one and visualize it in a map

Even if it has nothing to do with the risk, we want to analyse Provinces in order to be able to suggest to our client nearest Provinces to visit if they want.
One good approach to do this is to use DBScan, using coordinates as features. In this case "density" has the meaning of "Proximity".
We tune hyperparameters (epsilon first) according to our perception of "nearness", and the result is te one on figure 3.6.

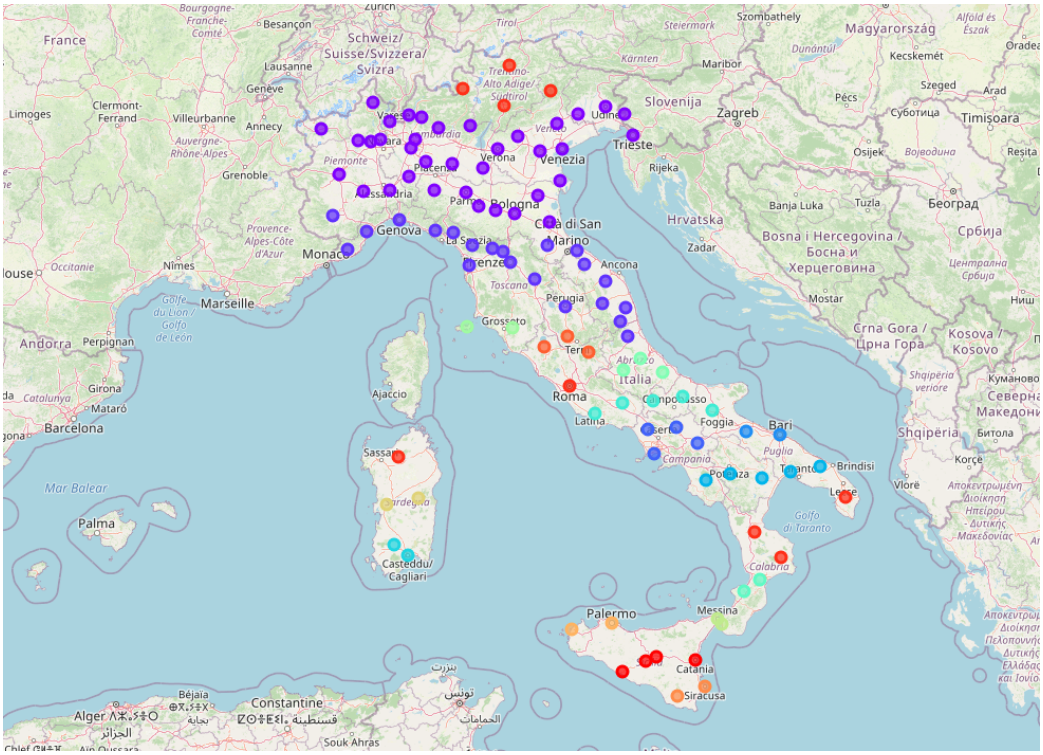

Figure 7: Nearest Provinces.

Lot of Provinces can be considered as "easily reachable". A client could organize a tour to visit all the provinces in cluster.
Only few province are not connected to any one other: one of these is Rome, which is bigger enough to be visited alone.

3.7 Cluster Italian Region according to traffic risk and visualize it

Go back to our risk analysis. What if we want to recognize exactly 3 type of region: low risk, moderate and higher.
We use K-means to do it automatically, and the only feature is the "*Incidents per* km^2".
That's the result!

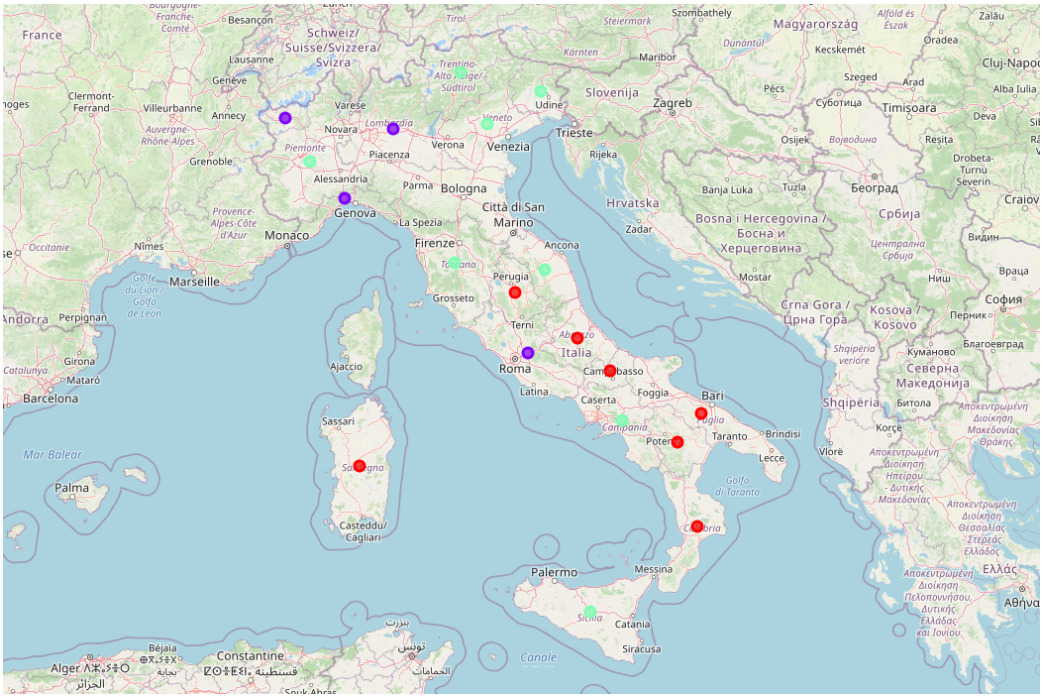

Figure 8: Region division according to their risk.

According to the map violet Regions have an higher risk, cyan Regions have a moderate risk, and red ones are the safest.

3.8 Get information about venues for Provinces in most attractive Regions

Suppose a client decide we want a region with the sea, and with lower risk. We are still undecided between Sardinia, Calabria and Puglia.
Let's take the provinces in those region (we know the domain so this information will be given) and choose only between provinces connected to at least 1 other province (remember the previous cluster). From this we want to retrieve information about the type of venues in there (using foursquare API).

Province	count	categories
cagliari	3	Bar, Ice Cream Shop
catanzaro	3	Pizza Place
nuoro	3	Mountain
oristano	3	Church, Pizza Place
reggio calabria	5	Café
sud sardegna	3	Miscellaneous Shop
vibo valentia	5	Pizza Place

Figure 9: Most Popular venue per Province.

4 Results & Conclusion

Now we are able to enquiry lot of question and retrieve all the answer we want.
For example, suppose we are interested in try the real italian Pizza. We want to have a great choice of Pizza Place. Catanzaro has 4 Pizza Places, and Vibo Valentia 5. Catanzaro and Vibo Valentia are also easily reachable as we discovered in previous clustering based on DBScan. They are in Calabria, one of the region with lower number of incidents per area as shown in previous bar chart.
One of these should be our choice!

5 Future work

Data we used are a good sample of the real world, but more informations can be exploited in order to obtain more fitted analysis in order to suggest location for vacations.
Could be very interesting having more data, collect a sample of choices of some clients, in order to use places as label of each sample, and using rank to some factors (from 0 to 5, how do you want to have the sea? how many points to restaurants? how many points to low traffic risk? etc.).
We can try to develop and compare different model (such as regression, decision tree or a simple ANN) to predict the best place for each client, deploy a software that using information categorizing a client can suggest places that he can choose according to its will.