Covid-19 evolution in Santander, Colombia

Coursera Captstone project for IBM data science professional certification

By Jhon Padilla

Business Problem

- Santander is a region of Colombia. **In Colombia, covid-19 arrived in February 2020.** The government of Colombia decided to decree isolation measures in March 2020. One of the regions or departments of Colombia is Santander, which is one of the most important regions in the Northeast of the country.

- Now, **the number of cases in Santander has increased fastly** and it is important to do an analysis of how the pandemic behavior is over the people (by age, by city, by gender, etc.).

- A map with the cities and the number of Covid-19 cases is given to help travelers make decisions.

- Another aspect that should be improved is that Doctors now have to make decisions about what people should be attended at ICU (Intensive Care Units) or not; this situation is presented due to the limited number of ICU beds. Also, with this project, an ML algorithm for death probability calculation is developed based on parameters given in the data set. Such probability could be used by Doctors make decisions taking into account this parameter and other parameters on the patient's health.

- Finally, as the main city of Santander, Bucaramanga has the main healthcare centers and therer are people who have relatives sick from covid, they need to know what venues are near clinics and hospitals. Then, a map with venues near to main to these healthcare centers is provided for such situations.

Target Audience

Results of the project are useful for:

- People who need to travel to some city and need to know the number of Covid-19 cases in that place.
- People that needs to know what venues are near to healthcare centers in Bucaramanga (the main city of Santander).
- The statistical report it is important for authorities who need to make decisions over possible social, economic and health strategies. Finally, the report about the death probability is useful in medical decisions when the number of ICU beds is scarce.

Data Background

In Colombia, health ministry has a web page with open data at www.data.gov.co. In that page there are data from covid-19 pandemic in all regions and cities within Colombia. The data set has 174145 rows and 23 columns.

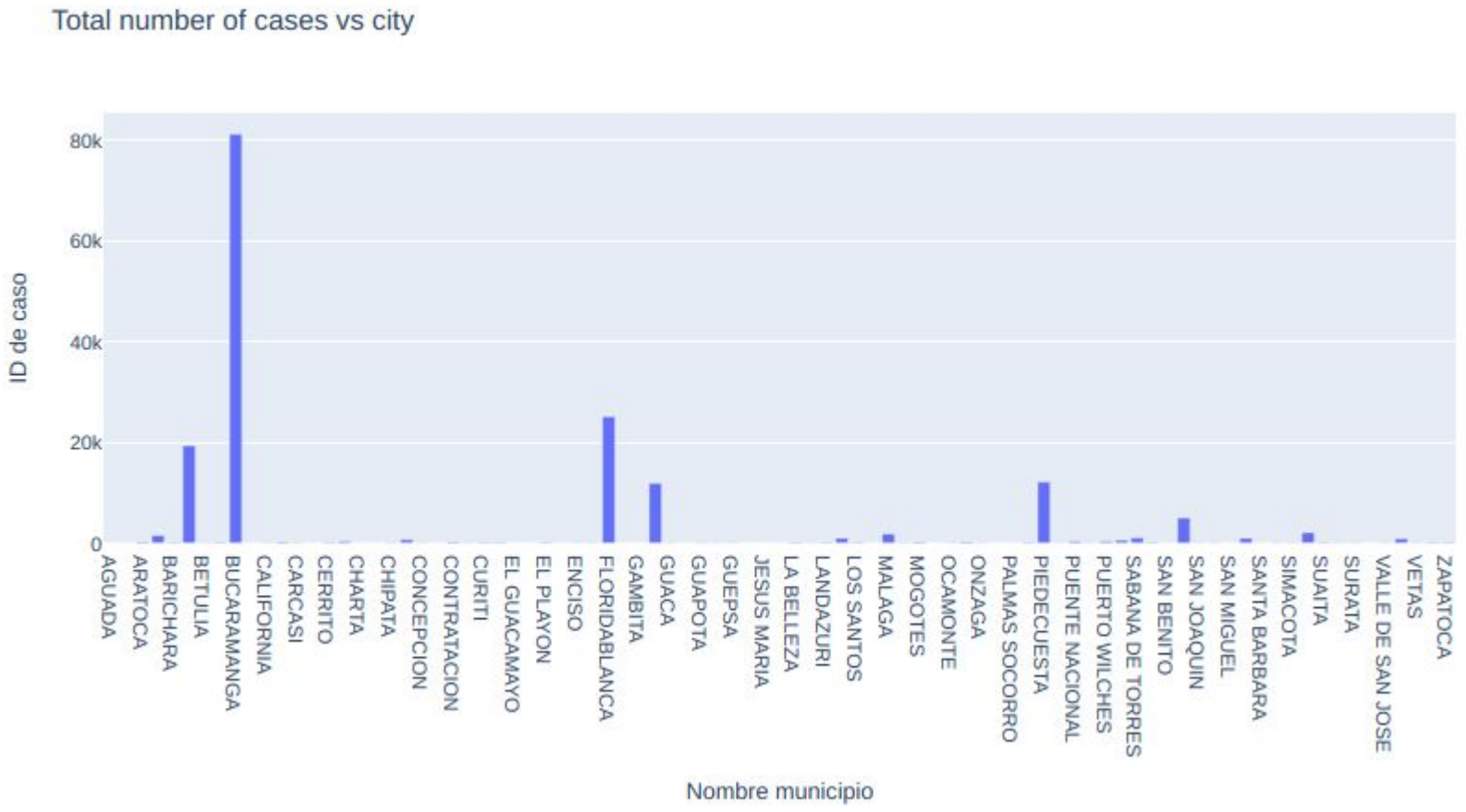
Columns of table:

Spanish name	Type	English translation
fecha reporte web	object	Web report date
ID de caso	int64	Case ID number
Fecha de notificación	object	Notification date
Código DIVIPOLA departamento	int64	Department Code
Nombre departamento	object	Department Name
Código DIVIPOLA municipio	int64	City code
Nombre municipio	object	City name
Edad	int64	Age
Unidad de medida de edad	int64	Age Cuantity
Sexo	object	Gender
Tipo de contagio	object	Type of contagion
Ubicación del caso	object	Case location
Estado	object	State
Código ISO del país	float64	Country ISO code
Nombre del país	object	Country name
Recuperado	object	Recovery
Fecha de inicio de síntomas	object	Symptom onset date
Fecha de muerte	object	Death date
Fecha de diagnóstico	object	Diagnosis date
Fecha de recuperación	object	Recovery date
Tipo de recuperación	object	Recovery type
Pertenencia étnica	float64	Ethnicity
Nombre del grupo étnico	object	Ethnic group

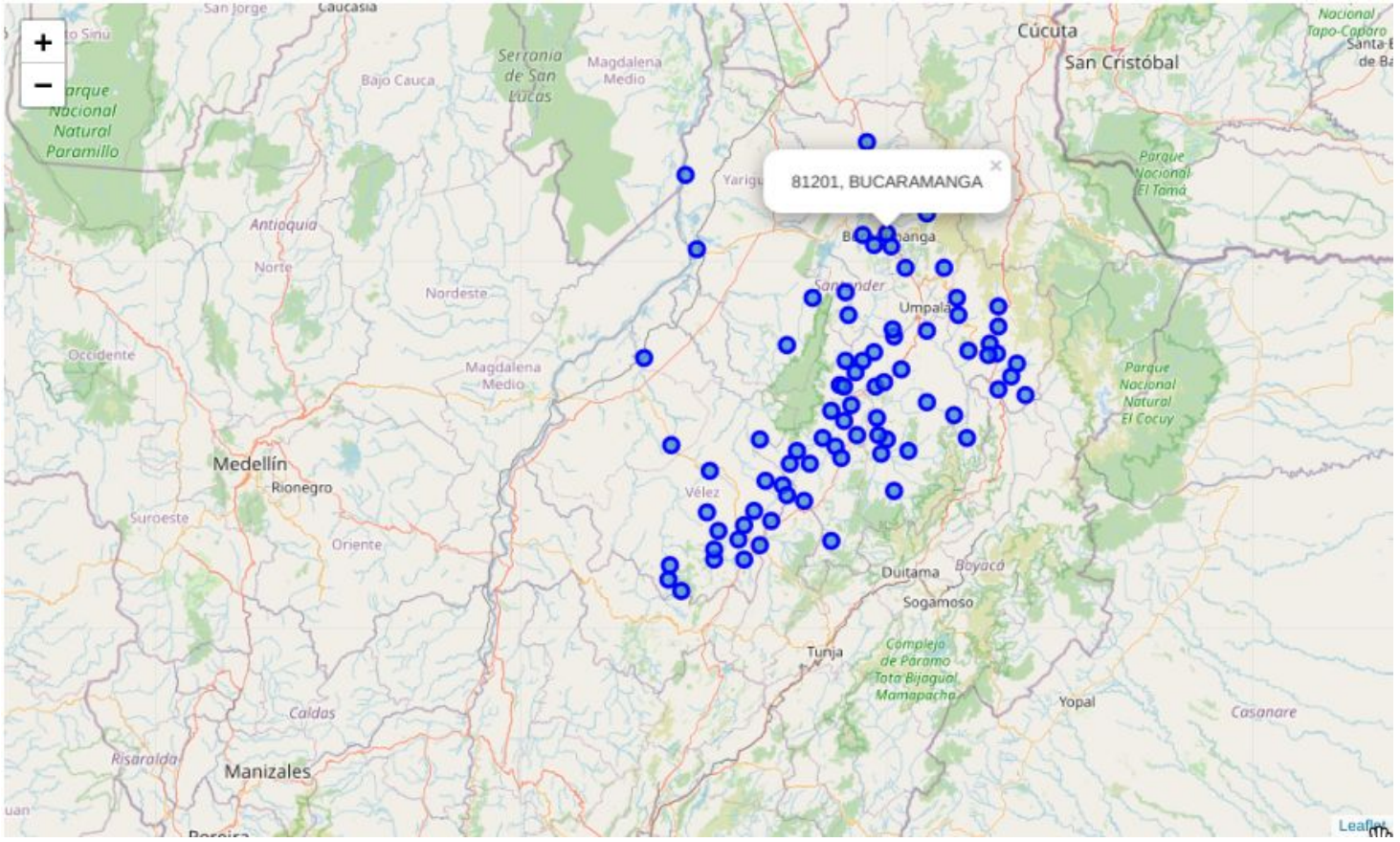
Methodology

- Exploratory Data Analysis
- ML model development
- Results and Discussion
- Conclusions

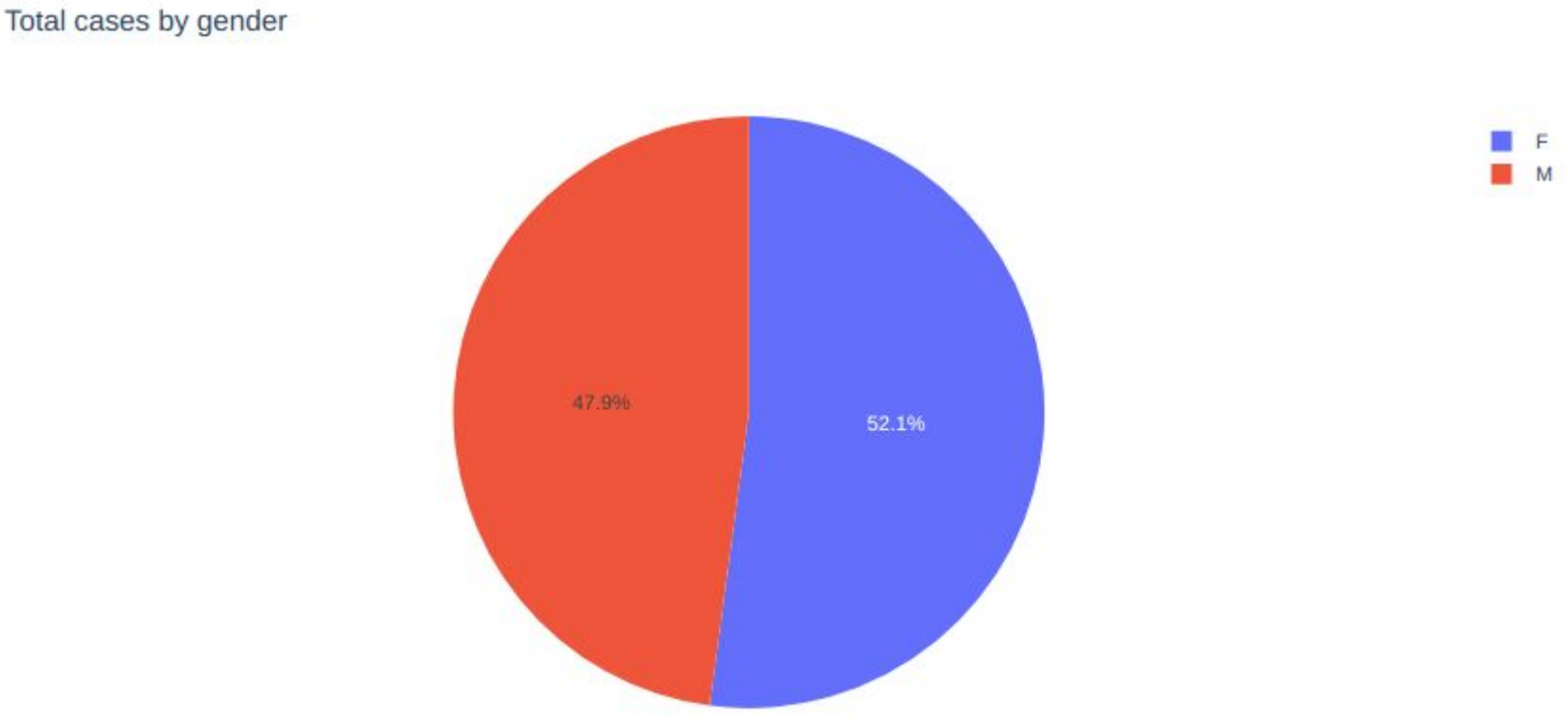
Exploratory Data Analysis



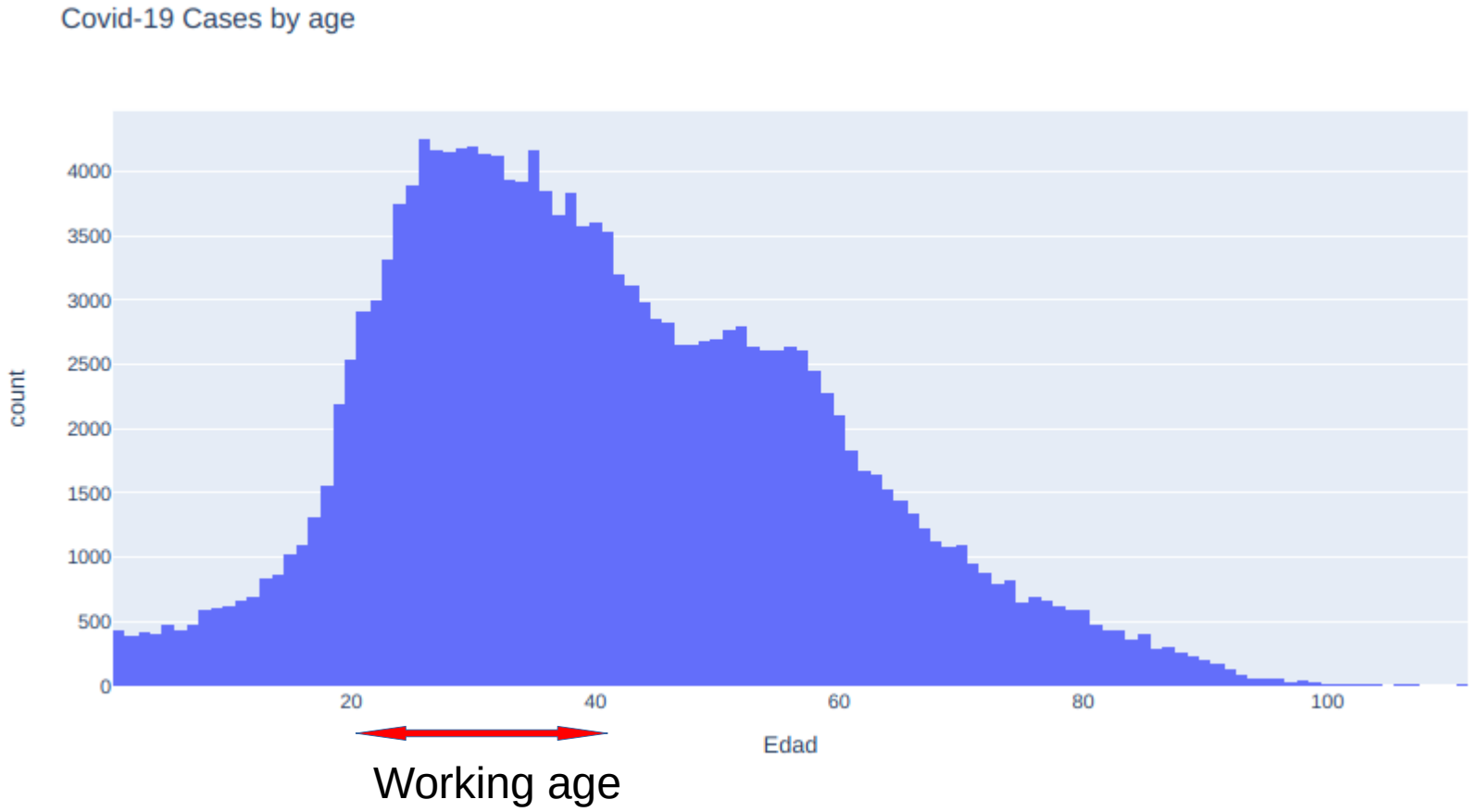
Number of Cases per city map



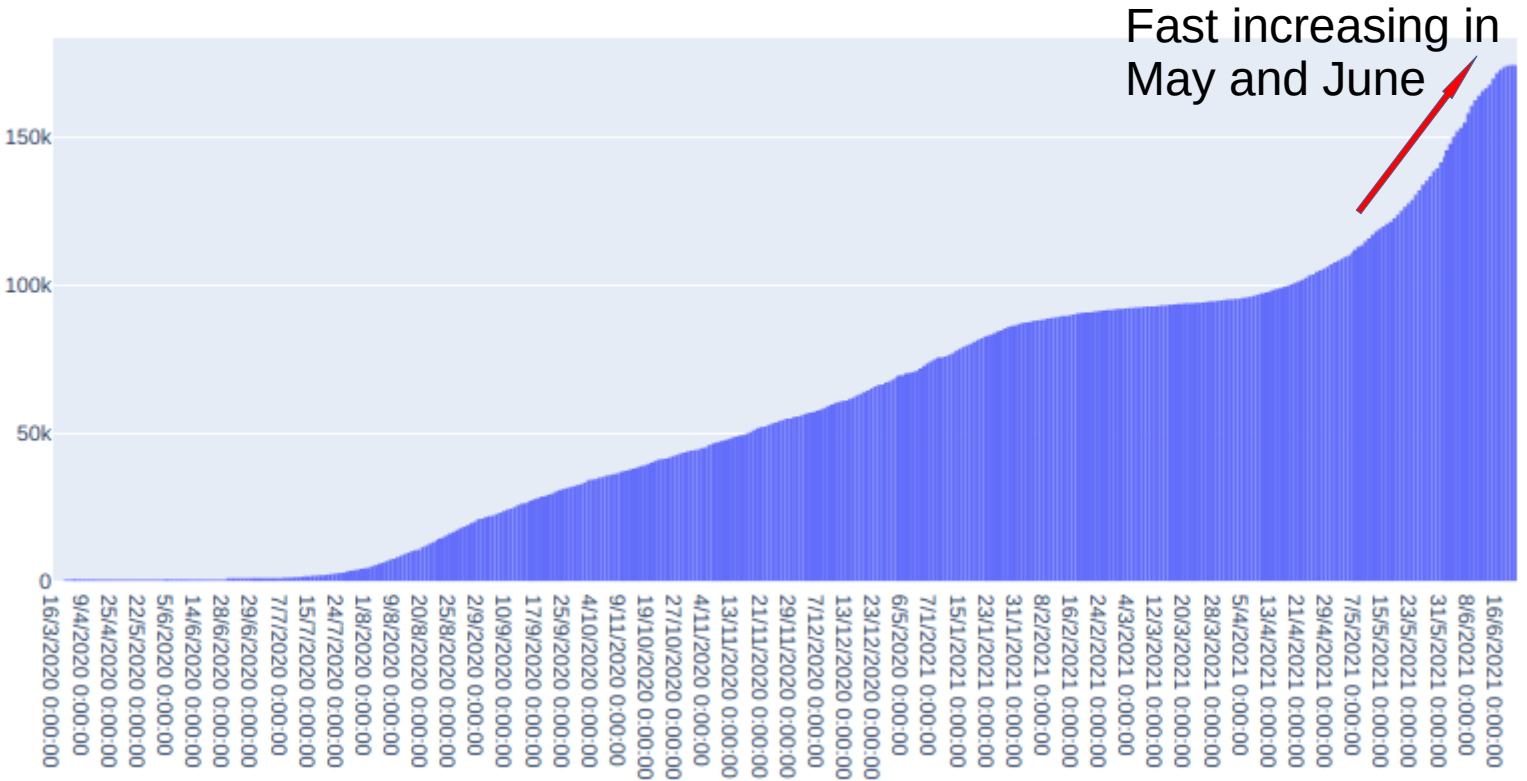
Number of cases by gender



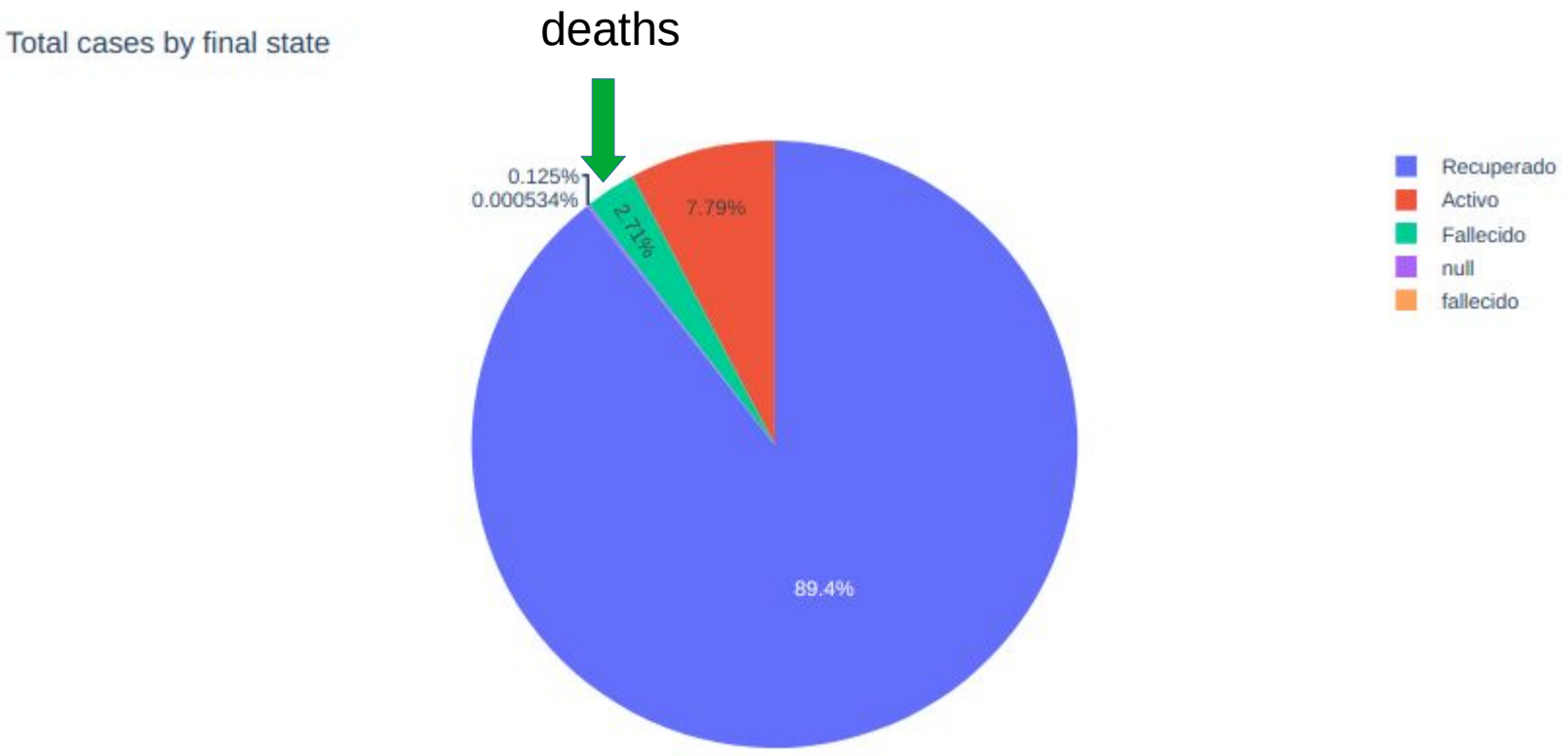
Distribution of cases by age



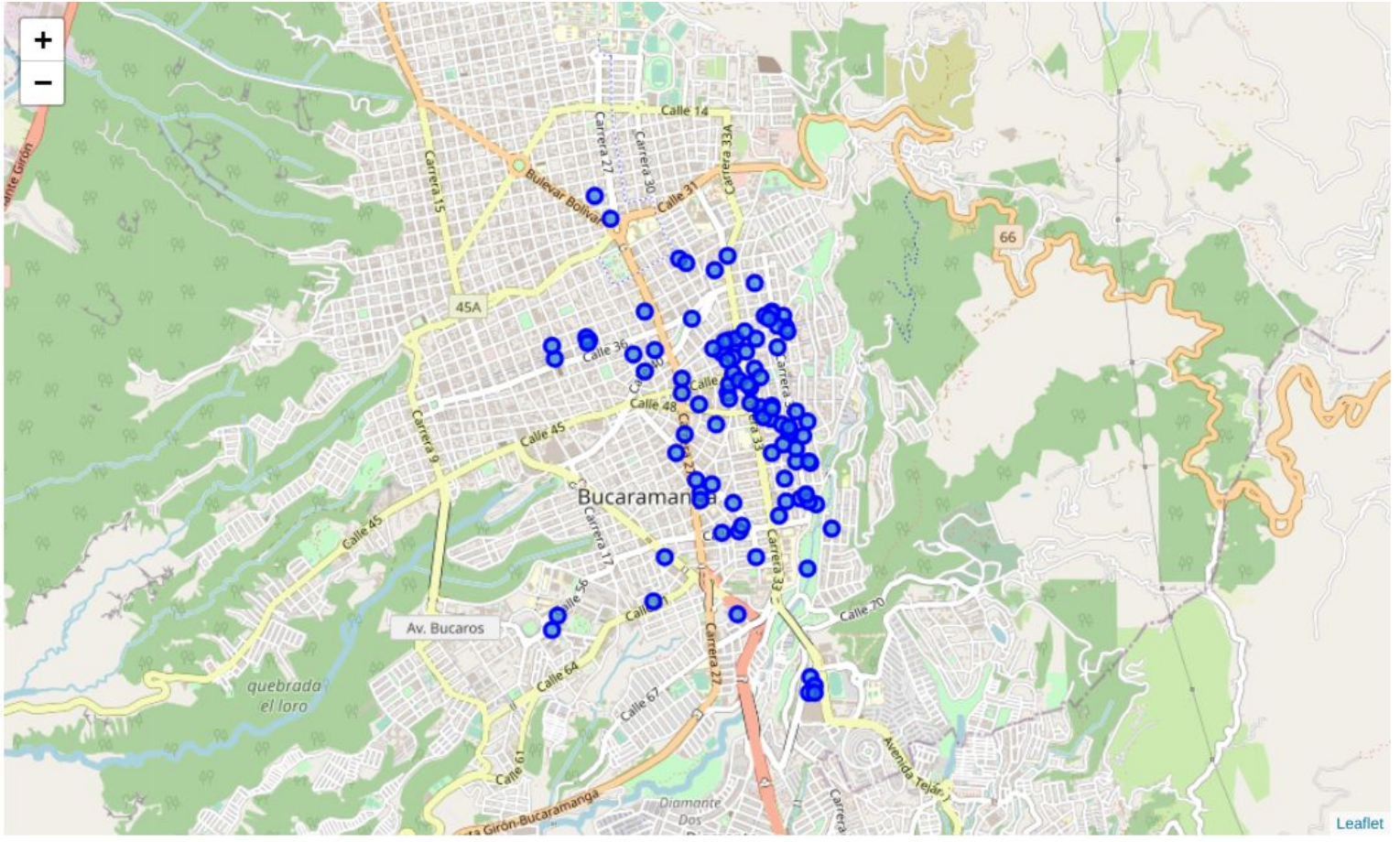
Number of cases over the time



Analysis of deaths



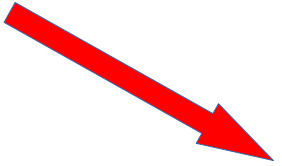
Venues near to healthcare centers in Bucaramanga city



ML model Development

- Data Cleaning

	Código DIVIPOLA municipio	Edad	Sexo	Pertenencia étnica	Recuperado
0	68276	24	M	6.0	Recuperado
1	68001	39	F	6.0	Recuperado
2	68001	49	F	6.0	Recuperado
3	68001	33	F	6.0	Recuperado
4	68001	80	F	6.0	Recuperado



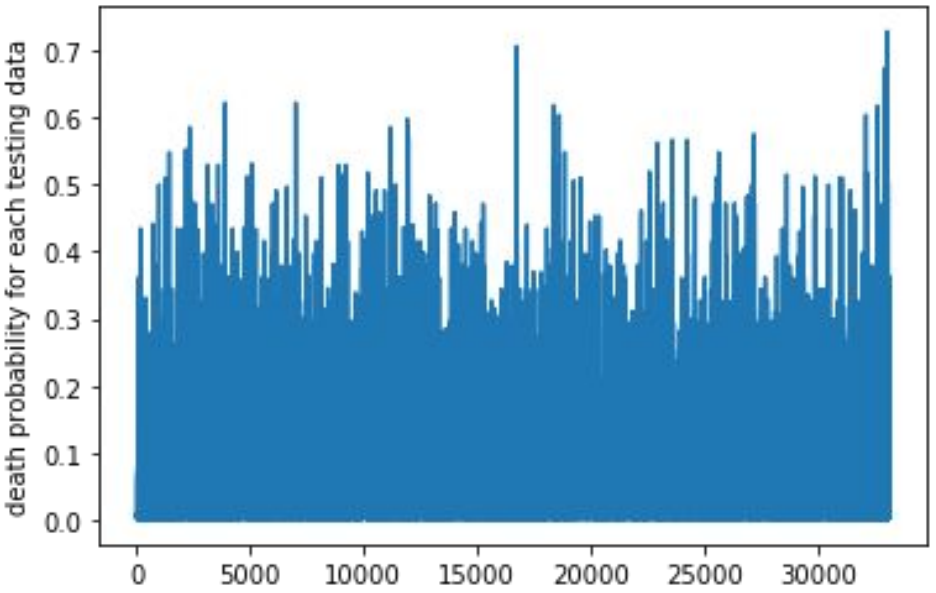
	Código DIVIPOLA municipio	Edad	Sexo	Pertenencia étnica	Recuperado
0	68276	24	0	6.0	1.0
1	68001	39	1	6.0	1.0
2	68001	49	1	6.0	1.0
3	68001	33	1	6.0	1.0
4	68001	80	1	6.0	1.0

ML model development

- The Machine Learning model that was selected is Logistic Regression.
- Such algorithm can calculate the probability of occurrence of a given event based on several parameters given as input.
- In our case, **the event is the death of a patient**.
- **Input parameters** to Logistic Regression model are: city code, age, gender and ethnicity
- **The output is the "Recuperado" column** (Recovery or final state in this situation).
- **Steps:**

1. Input parameters were normalized
2. Data were divided into training set and test set.
3. Then, Logistic Regression algorithm was fitted
4. predict_proba algorithm was executed to return estimates for all classes, ordered by the label of classes. So, the first column is the probability of class 0, P(Y=0|X), and second column is probability of class 1, P(Y=1|X). Finally, the probability of death was graphed for all cases.

Logistic Regression model output



Results and discussion

- In the data analysis performed in this project, we first obtain the relationship between the cities and the number of cases and we observe that the main cities, which have higher population, have the majority of cases. These cities are: Bucaramanga (81201 cases), Floridablanca (25193 cases), Barrancabermeja (19395 cases), Piedecuesta (12169 cases), Girón (11951 cases). All cities, except Barrancabermeja, are part of the Metropolitan Area of Bucaramanga.

- Also, about the relationship between the number of cases and the age, it is clear that most cases are located between 20 and 40 years old, this may be because these people are of working age and also are students and young people with a hectic social life.

- If we see the comparison by gender, the number of cases for both, male and female, are similar, with 52,1% for females and 47,9% for males.

- Another aspect evaluated was the number of cases that led to death. The percentage of deaths is 2,71%. Thus, the number of deaths are 103849 in Colombia and 4719 in Santander, which is an important number compared with other countries in the world.

- About the number of cases over the time, we can see that in the last two months (May and June), the number of cases was increased fastly. This is consequence of two main aspects: the holly week vacations in April, and the protests ocurred due to economic recession caused by social isolation and the closure of businesses for long periods of time decreed by the government.

- Finally, in this project, a Machine Learning model was developed to make predictions about the death probability based on parameters such as: age, city, ethnicity and gender. This model was developed with a Logistic Regression algorithm. The output of the algorithm is the probability of death. This results should be taken as another tool but not the final answer to this question because health parameters that doctors knows about patient should be the main reasons. However, this probability could be useful when the doctors should make ethical decisions when ICU beds are scarce.

Conclusions

- In this project,an important situation that is happening in Colombia with the increasing of covid-19 cases, was studied with base in the data set that is feeded by Colombia government and is located at www.data.gov.co web page.

- First, an exploratory data analysis was developed, which showed important results about the distribution by cities, the cases distributed by age and gender, and the number of cases that finish in death. Besides, the number of cases over the time.

- Also, another two tools were developed: a map with the number of cases per city, which could be useful for travelers, and another map with he main venues around the health centers area at Bucaramanga, the capital city of Santander, which could be useful for people with relatives hosted in Clinics and hospitals in that city.

- Finally, in this project, a Machine Learning model was developed to make predictions about the death probability based on parameters such as: age, city, ethnicity and gender.