Analytics and data visualization to the emergency number 123 in Bogotá.

Nicolas Barrera Boada.
Lina Maria Cubillos Vergara

nibarrerabo@unal.edu.co, licubillosv@unal.edu.co

Facultad de Ingeniería. Sede Bogotá.
Universidad Nacional de Colombia.

Abstract **—** In this project, an analytical and visual analysis will be carried out at the emergency number 123 in Bogotá. Through data visualization find heatmaps for hour and day of the week for differents types of emergency, also deduce there are no features for distinguish between a real and a fake call. In the analytical part realice forecasting to predict the expected volume of calls, also use anomaly detection to find outlier data based on the calls duration.

Keywords **—**emergency number, forecasting, anomaly detection.

I. Project Goal and Scope.

The purpose of this project is to make a descriptive analysis, analytical and visual, to observe the behavioral information on the 123 emergency number in Bogotá, in this way being able to know the major emergencies, the dates, times and/or locations in the one with the highest number of emergencies, relating false or canceled calls to the type of emergency, or the time or location at which they are made, among other analyzes that allow describing the behavior of the population. In addition to this given the current situation, you could see the change in behavior on emergency number 123 at the time of quarantine.

II. Literature Review.

Erwin A. Blackstone, Andrew J. Buck, Simon Hakim*(2005)* The authors, through a microeconomic approach, theory of the commons, make the approach to the use of emergency services, seeing this as a common good, examining 10 different strategies to combat false calls, focusing on calls for theft, they point out that in 2000 police responded to 36 million false calls at an estimated cost of $ 1.8 billion. [1]
for criminalistics W. Stan Crowder, Brent E. Turvey, (2018) fake calls can go a little further than an innocent game, they analisis different cities of the United States to face false emergency calls, which are classified as Accidental, Non-emergency Nuisance, and False Reports. It also makes a brief profile of the motivations of those who make the calls, noting that these can be part of criminal strategies:

"False 911 calls are made with a variety of criminal motivations in mind. The most common are Crime Concealment (e.g., alibi or diversion); Punishment or Revenge; Profit; Personal Sympathy; Sociopolitical; and Terroristic."

III. Case Study.

In the Kaggle virtual data science community, there is a data analysis competition for the 911 emergency line in Montgomery County, Pennsylvania [2]. This competition does not have a specific purpose and the analysis is open, the data includes emergency information such as the location, description, title, date and time at which the call occurred.

For these data there are 419 uploaded projects, which include cleaning, analysis and visualization of the same. In the analysis process there are both descriptive and predictive cases, seeking to model the types of accidents with the dates, the time of greatest number of calls, the location of the calls, among others. Similarly, the display shows heat maps, dispersion bars, etc. which visualice more nicely the data.

This case study is comparable to our project since the columns have similar information and therefore cleaning, analysis and visualization can be carried out on our data in the same way. However, our data includes extra columns such as final classification and priority of the call, which would allow for a richer analysis.

IV. Data Set.

The data source used in the case study will be open data, published on the website https://datosabierto.bogota.gov.co/dataset/b99881e0-8750-4fa8-af6c-3f30c9e38708?_external=True, which are in .csv format and all calls are reported from January 01, 2019 to December 31, 2019.

NAME	DESCRIPTION
DATE_INCIDENT	it is the date on which the call of the incident that is entered through the Unique Number of Security and Emergencies NUSE is registered.
DATE_HO	It is the date on which the movement of the ambulance to

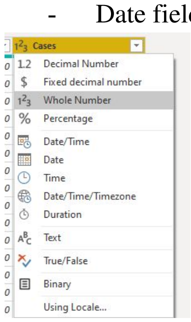
ME_DISP	the site of the emergency begins.
LOCALITY _CODE	It is the code of the 20 localities of the city of Bogotá according to the code belonging to each of them designated by the district.
LOCATIO	It is the locality where the incident happens.
AGE	The age of the patient.
UNIT	The description of the age if it is in hours, days, months or years.
GENDER	It is the distinction of gender of the patient
NETWORK	It is the location at the Bogota level of the health network which the incident was attended to
INCEDENT _TYPE	It is the initial description that typifies the command center of the Unique Number of Security and Emergencies NUSE
PRIORITY	It is the classification according to the priority of the incident.
MONTH	It is the month in which the incident occurred
FINAL CLASSIFIC ATION	It is the final classification that the regulatory center of urgencies and emergencies CRUE gives on the incident.

Transform – Quality Control

Modify data type fields

Quality Control:

- Numeric fields
- Date fields



V. Exploratory Data Analysis (EDA).

Counting the calls to the line 123, the Fig. 1 shows there is an average in almost all the month, but there is an atypical number of calls registered between february and april.

Figure 1: Number of calls to 123 in 2019

In the Fig. 2 identify there are a lot of incident types, in the rest of the EDA we will focus on those that supere a threshold of occurrence.

```
df['TIPO_INCIDENTE'].unique()

array(['lesiones personales', 'accidente de transito con heridos/muertos',
       'dolor toracico', 'inconsciente/paro cardiorrespiratorio',
       'herido con polvora', 'intento de suicidio', 'disparos', 'enfermo',
       'trastorno mental', 'dificultad respiratoria', 'heridos', 'caida',
       'intoxicaciones', 'convulsiones', 'incidente rescate acuatico',
       'sintomas gastrointestinales', 'accidente cerebro vascular',
       'quemaduras', 'rina', 'ideas de suicidio', 'embriaguez',
       'incendio estructural', 'patologia gineco - obstetrica',
       'incendio forestal', 'muerte natural', 'violencia sexual',
       'verificar situacion', 'abrir domicilio', 'sangrado vaginal',
       'incendio vehicular', 'accidente transito simple', 'rescates',
       'persona pidiendo auxilio', 'habitante de la calle', 'sin dato',
       'persona tendida en la via', 'electrocucion / rescate',
       'fuga de gas natural o propano', 'solicitud apoyo / desacato',
       'elemento caido y/o en peligro de caer', 'animal peligroso',
       'atraco / hurto en proces', 'muerto', 'explosion',
       'extraviados / desaparecidos', 'menor o persona abandonada',
       'rapto / secuestro', 'matpel materiales peligrosos', 'maltrato',
       'manifestacion / motin', 'prevencion', 'incendio con matpel',
       'amenaza de ruina', 'rescates montana', 'vehiculo recuperado',
       'hurto efectuado', 'venta o consumo alcohol u otro en menor',
       'delincuente capturado por civil', 'incendio',
       'hallazgo de explosivos', 'animal atrapado', 'vehiculo hurtado',
       'abejas', 'inundacion', 'deslizamiento',
       'persona o vehiculo sospechoso',
       'vehiculo abandonado / mal estacionado', 'psicologia', 'fallecido',
       'narcoticos', 'danos en servicios'], dtype=object)
```

Figure 2: Incident types

In the Fig. 3, observe some incident are more possible in men that in women, in this same way some diseases are exclusive for women.

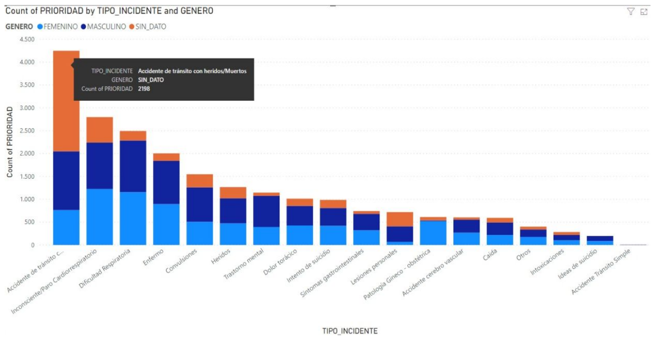


Figure 3: Incident type by gender

The Fig. 4, represent histogram of call with duration less than 30 minutes (85% data), showing us that majority of call have a duration between 0 and 10 minutes, that is because 123 is an emergency line and transfer those calls.

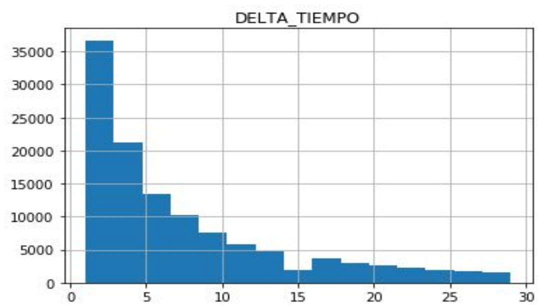


Figure 4:: Call duration histogram

Fig. 5, visualize in a easy way the weekday and hour were there is more activity in the emergency line, the less activity occurs between 11 pm and 5 am, and major activity occurs in two differents range, in the morning between 8 am and 11 am, and in the after between 7pm and 9 pm. Filtering this data by the type of incident is possible to observe the differents schedule where is more possible to occur.

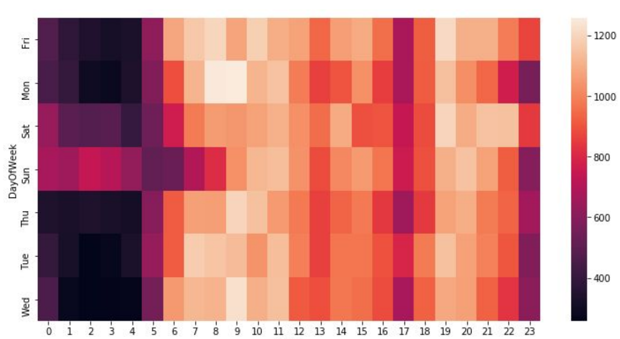


Figure 5: HeatMap number of call by weekday and hour

Searching for a discriminant between real and fake calls, plot different features subject to that discriminant. In Fig. [6-8] observe that there isn't a significant difference between real and fake calls for incident type, location, month, etc. that means isn't possible find a discriminator, a possibility is that there is a wrong final emergency classification, part of this study is to propose a better classification for unreal calls.

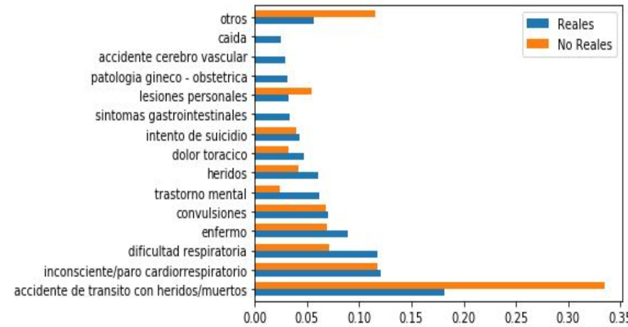


Figure 6: Incident type for real and unreal call

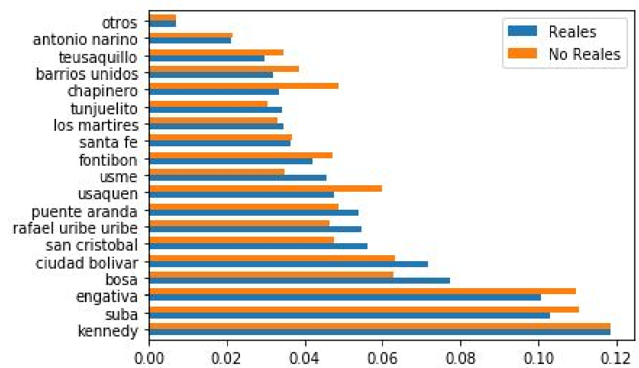


Figure 7: Location for real and unreal call

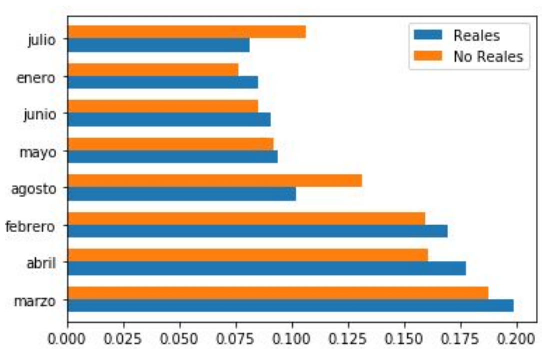


Figure 8: Month for real and unreal call

The Bogota Health Observatory [4], classifies canceled and fake call as unreal and bad use of the 123 line, and transferred as real call, bringing closer to canceled calls some of those there aren't bad use of the line but doesn't require transfer, specially suicidal attempt or mental diseases. Also a common mistake is the wrong labeling of the call in the final classification.

VI. Analytical Data Analysis.

1. Forecasting

1.1 Applying prediction to the data with Power BI - Forecasting (Exponential Smooth).

Power BI was used as a tool for predicting data, the model used for prediction is Exponential smoothing

Exponential smoothing is a rule of thumb technique for smoothing time series data using the exponential window function. Whereas in the simple moving average the past observations are weighted equally, exponential functions are used to assign exponentially decreasing weights over time. It is an easily learned and easily applied procedure for making some determination based on prior assumptions by the user, such as seasonality. Exponential smoothing is often used for analysis of time-series data.

A first analysis was performed on the 2019 dataset, in which the FINAL CLASSIFICATION field determines the action taken on each call, this record exists for the months of January to August 2019, for the following months the entity did not continue recording this information.

1.2 Setting up the prediction

When applying the data prediction, it was initially configured for 3 months, with a 95% confidence interval, filtering the data that only records a transfer of the user who registered the call.

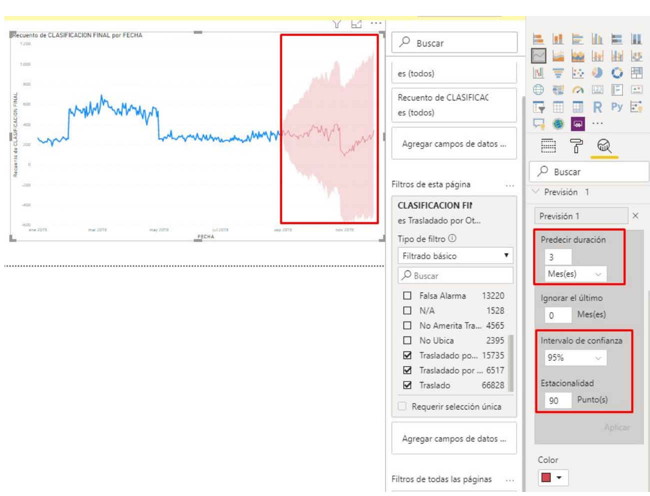


Figure 9: Forecasting setting up

When applying the data prediction for the following 3 months, it was obtained that the highest volume of calls that require the attention of the 123 staff, occurs for Wednesday, October 23, 2019, with a maximum expected number of 403 calls.

Figure 10: Forecasting volume calls

It is important to highlight that to apply data prediction, you must have historical information to generate valuable data and the applied model may have a higher degree of reliability, for this reason we make a prediction for the volume of calls from 2014 to 2019.

1.3 Call volume from 2014 to 2019

The dataset from 2014 to 2019 has data on the location, type of incident and number of incidents presented on each date

ID	ANIO	MES	COD_LOCALIDAD	LOCALIDAD	COD_INCIDENTE	INCIDENTE	CANT_INCIDENTES
20140119535	2014	1	19	CIUDAD BOLIVAR	535	INFORMACIÓN CONFIDENCIAL	37
20140104535	2014	1	4	SAN CRISTOBAL	535	INFORMACIÓN CONFIDENCIAL	35
20140118535	2014	1	18	RAFAEL URIBE URIBE	535	INFORMACIÓN CONFIDENCIAL	33
20140109535	2014	1	9	FONTIBON	535	INFORMACIÓN CONFIDENCIAL	35

To verify the effectiveness of the prediction model, the dataset that records data from 2014 to 2019 was used, the information from 2014 to 2018 was taken as a base, the comparison of the real data for the year 2019 was made vs. what was shown in the prediction for the same year and the following data was obtained

Real data 2019

FECHA	CANT_INCIDENTES
martes, 1 de enero de 2019	214322
viernes, 1 de febrero de 2019	221572
viernes, 1 de marzo de 2019	253118
lunes, 1 de abril de 2019	227265
miércoles, 1 de mayo de 2019	237343
sábado, 1 de junio de 2019	222936
lunes, 1 de julio de 2019	223707
jueves, 1 de agosto de 2019	240354
domingo, 1 de septiembre de 2019	236694
martes, 1 de octubre de 2019	236292
viernes, 1 de noviembre de 2019	254017

Data prediction for 2019

FECHA	forecastValue	confidenceHighBound	confidenceLowBound
martes, 1 de enero de 2019	211995	234040	189949
viernes, 1 de febrero de 2019	231907	256564	207249
viernes, 1 de marzo de 2019	247637	274655	220619
lunes, 1 de abril de 2019	237995	267183	208806
miércoles, 1 de mayo de 2019	253889	285097	222680
sábado, 1 de junio de 2019	227713	260818	194607
lunes, 1 de julio de 2019	237366	272265	202467
jueves, 1 de agosto de 2019	246323	282928	209718
domingo, 1 de septiembre de 2019	246293	284528	208058
martes, 1 de octubre de 2019	260551	300349	220753
viernes, 1 de noviembre de 2019	256554	297856	215252
domingo, 1 de diciembre de 2019	274275	317028	231521

When verifying the model's prediction, it is evident that the data is within the real threshold for calls made in 2019.

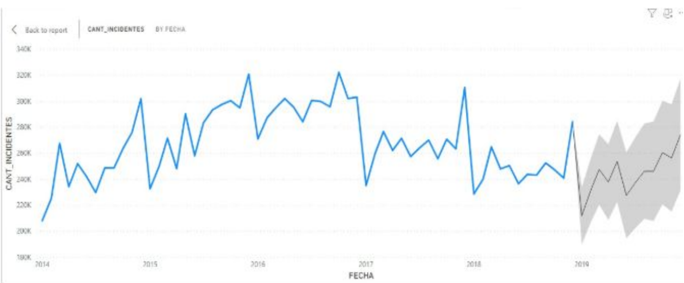


Figure 11: Forecasting 2019

The image shows that the data prediction was within the estimated range for 2019 due to the historical dataset information.

2. Anomaly Detection

Despite the mentioned about the no existence of features or combinations of them that allow us to discriminate between real and unreal calls, we observe that there are defined behaviors in the data that in change allow us identify anomaly detection.

Anomaly detection refers to identification of items or events that do not conform to an expected pattern, detecting anomalies, perturbations of normal behavior indicate a presence of intended or unintended induced attacks, defects, faults, and such. Dynamic systems having numerous components in perpetual motion where the "normal" behavior is constantly redefined. [5]

This kind of analysis could be used in many situations:

- Detect early changes in emergency behaviors (traffic accident, crime, suicide attempt, diseases, etc.)
- Identify call duration for different types of emergency and extreme times behaviour.
- Observe the effectiveness of politics to reduce differents types of emergencies in local o global way.
- etc.

In this work will focus on the second item, this problem is interesting because the expected use of the line 123 doesn't depend only in the numbers of calls but also in his occupancy time. In the Fig. 4 represent the histogram of calls duration with less than 30 minutes, this data represent 85% of total data. The restant 15% is considered outlier data, and this is composed by noise (uninteresting data) and anomaly (sufficiently interesting data).

To detect the noise and the anomaly data, first we fit the data using a probabilistic method, after determine a threshold for the outliers, given that the diverse types of emergency have dissimilar behaviour this method must be applied for each one, in Fig. 12 and Fig. 13 we see an example.

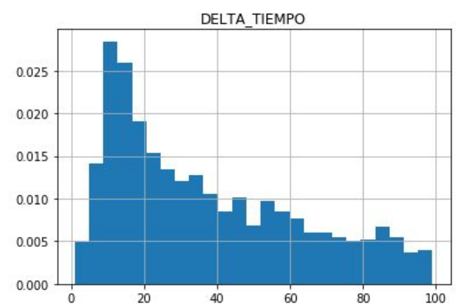


Figure 12: Histogram call duration for mental disorder

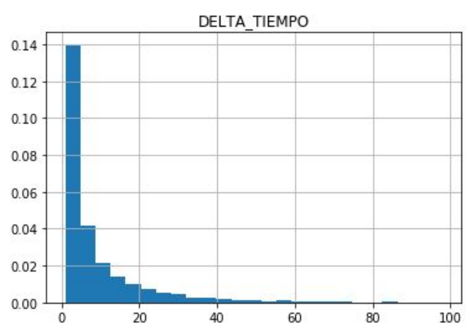


Figure 13: Histogram call duration for traffic accident

Using the proposed method for the type mental disorder, fitting to a general pareto distribution and assuming probability threshold of 95% for noisy outlier and 85% for anomaly outlier, obtains that for calls duration over 280 minutes are noisy data and over 156 minutes are anomaly data.

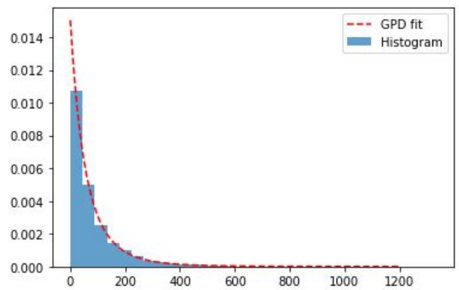


Figure 14: Probabilistic fitting for mental disorder

Now for the type traffic accident the best fitting is reached with an exponential fit, assuming the same threshold noisy data is for calls duration over 29 minutes and anomaly data over 19 minutes.

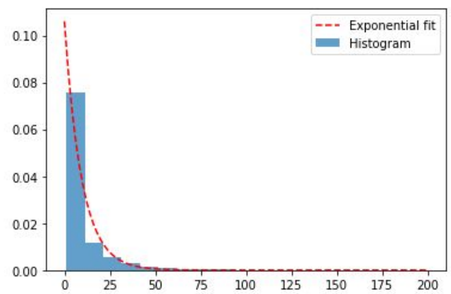


Figure 15: Probabilistic fitting for traffic accident

Comparing normal versus anomaly data there is a reduced difference in some features like classification, priority, month, etc. An important aspect is that despite being only 5% of percentage of the total data, the percentage of consumed time for anomaly is 10%, this amerites to review in a meticulous way this anomalous data.

VII. Conclusions

In the data visualization observe there is an patron in the type of emergency for the hour and day of the week, this data could be used for the design of politics.We couldn't identify features for discriminate between real and fake calls.

In the call duration visualization we observe there is a majority of data between 0 and 10 minutes, using this data to determine outliers we identify anomaly and noisy data.

The nature of the dataset did not allow the prediction of data on the type of incident to be applied since the final classification is not available for the period from 2014 to 2019.

Prediction of the volume of calls for 2019 was made based on historical data from 2014 to 2018 in order to evaluate the effectiveness of the applied model, obtaining an acceptable margin of error and close to the actual data.

References

[1] Erwin A. Blackstone, Andrew J. Buck, Simon Hakim(2005) "Evaluation of alternative policies to combat false emergency calls",Evaluation and Program Planning, Volume 28, Issue 2

[2] Emergency Calls-911: https://www.kaggle.com/mchirico/montcoalert#911.csv

[3] W. Stan Crowder, Brent E. Turvey,(2018). "Chapter 4 - False 9-1-1 Calls", in False Allegations, Editor(s): Brent E. Turvey, John O. Savino, Aurelio Coronado Mares, Academic Press, 2018.

[4] http://saludata.saludcapital.gov.co/osb/

[5] https://medium.com/sciforce/anomaly-detection-another-challenge-for-artificial-intelligence-c69d414b14db