**A review of the road casualties in Dublin City Council**

Raúl Martín Sánchez

MSc in Data Analytics (Semester I), Integrated Assignment 1, CCT College Dublin

E-mail: [rmsryu@gmail.com](mailto:rmsryu@gmail.com) / [sbs22021@gmail.com](mailto:sbs22021@gmail.com)

Lectures: Marina Iantorno, Muhammad Iqbal, David McQuaid

Submission Data 8th April 2022

Abstract

In this paper, a detailed analysis of the roads collisions and casualties is presented to identify how different infrastructures and initiatives in Dublin City Council have an impact on roads collisions. The analysis of the proneness of an accident, distance to the fire brigade and ambulance services, GoSafe monitoring, and speed cameras on DCC roads, can influence the outcome and severity of it. Analysis of the road's collision is an important and difficult subject, multiple variables can affect the outcome of a collision, the majority of existing research on accident prediction is based on human error and behavioral patterns [1][2], including characteristics of the roadway, weather conditions, visibility among others.

To determine the concentration of roads collision, the K-means clustering criterion with euclidean distances was performed to assess the adequate locations and suitability of the fire brigade station locations and catchment areas. Distance to the closest ambulance service and the introduction of an estimated indicator of response time have been also included in the analysis.

The majority of the casualties have a low severity in terms of the criticality, given as a result of an imbalanced dataset to work with when it comes to categorizing the severity result. The use of different techniques for data preparation, resampling techniques, as well as the selection of the appropriate machine learning algorithms for categorizing the severity will be studied.

Keywords: Roads collisions, Statistical Analysis, Data Visualization, Cluster, Classification, Geospatial

Introduction

The number of roads collision and casualties in Dublin City Council are available via the map of collisions published by the Road Safety Authority for the years from 2005 to 2016, with only a subset of the features of the dataset available on the website. A request to both, the Research Department of the Road Safety Authority and Garda Síochána Analysis Services (GSAS) was placed to request access to it. In both cases, it was noted that due to GDPR restrictions the information could not be released.

One important missing feature on the public RSA dataset is the date and time, information such as weather conditions, climatology, and visibility as well as time-series analysis will need to be disregarded here.

However, the data collected using different techniques for combining the above-mentioned geospatial datasets gives enough raw data and features for the analysis.

RSA Dataset

Maps of collisions available on the RSA website with the URL <https://public.healthatlasireland.ie/rsa2/index.html> offer rudimentary access to the data. To obtain the XML used by the GIS map, it is required to zoom to the local view (GIS zoom level 6 to 8), and then move in the 4th direction to ensure covering the greater Dublin area to ensure all information relevant to DCC is downloaded. This process generates a list of 15 XML files of 1000 records each. To ensure it covered all areas of the interest in this study, several passes were done accepting any duplicated files and records that later on will be removed during the exploratory and data analysis phase [fig 2].

All 15 XML files were combined and converted to excel format with a total list of 35 columns. Documentation for those was extracted from the existing GIS map javascript (rsamapsite.js) which includes details, labels, and types of some of the most important features, but there was a significant number of undocumented columns, that after further analysis it was decided to exclude them (*refer to Data Preparation*). The final number of features after the removal of undocumented and merging datasets for analysis is 17: *'year', 'weekday', 'hour', 'splimit', 'gender', 'age', 'vehicle\_type', 'circumstances', 'latitude', 'longitude', 'total\_casualties', 'knn\_cluster', 'fbs\_distance\_m', 'fbs\_station', 'fbs\_estimated\_response\_time', 'is\_gosafe', 'safety\_index'* and the dependant variable *'severity'*.

The final dataset is the result of iterating on the [8] CRISP-DM process (except for the Deployment phase which was omitted for this assignment), the cycle was repeated 4 times in total, during the Evaluation stage. Each of them involves the analysis and feasibility of enriching the RSA dataset with new features given the low accuracy obtain for True Negative values seen in the confusion matrix. While the accuracy of the models achieves good results (>85), this could be misleading if the intention is to classify accidents where the outcome is serious or fatal (high severity). As a result of the imbalanced class distributions present in this dataset. It was important to identify which new features were available for inclusion, given the limitation of the existing dataset, as well the business understanding of those features and the possible impact on how would help to improve the final performance of the models. The 4 new features already mentioned above are:

* Safety Index: Analysis of proneness of accidents [5].
* Response Time Estimator
* Distance to closest Ambulance Station
* Monitored roads: GoSafe

Safety Index

This indicator, it’s calculated using the distance between each point and any other data points within a radius of 500 meters. To get the distance between two coordinate points on the earth, the haversine and spherical law of cosines equations also known as the [6] Haversine formula is used [cell 89][code 1].

Where:

The accumulation of accidents in the city center can be seen in the distribution of these results and the final safety index indicator

The safety index is normalized using MinMax scaler after outliers were removed. In this case, outliers were concentrated in the city center where there is a big agglomeration and condensation of the accidents [fig 6], by removing outliers in this area, it was not changed the significance of the index on them as the result index still keep the same city center area on the high range of the distribution of the safety index.

It was reviewed different scalers methods for the index including

1. Standard scaling
2. MinMaxScaler
3. MaxAbsScaler
4. RobustScaler
5. PowerTransformer
6. QuantileTransformer
7. QuantileTransformer
8. Normalizer

The rationale for using MinMax scaler is that it produces a positive number in a range between 0 and 1, where the highest number of accidents tends to the zero value and the lowest number of accidents gets will tend to 1. So the interpretation of the safety index becomes easy to read. As a convention, a safety index higher than 0.5 it’s considered unsafety or with higher proneness to accidents.

As noted before, the concentration of points in the inner circle of Dublin City center is making this estimator quite unbalanced. The histogram and plot with the chi-square distribution of the radius\_500 [fig 7][fig 12] represent a positive skew with a long tail in the positive direction, as seen in the previous plot for outliers [fig 6], it is easy to read from here that the farthest from the city center the lowest the accident proneness is.

Response time estimate

The impact of the response time on patients transported to the hospital was found to be significantly associated with pre-hospital trauma survival [3]. In 1997 emergency calls in urban areas in Ireland (85% of all emergency calls): 44% of calls received a response within nine minutes [4], in the year 2012, 44% of calls received a response of 11 minutes [fig 1][cell 66], an increase of 22.22%. In the case of Dublin City Council, an increase in the response time due to factors such as population and economic growth, infrastructures, and transportation services are challenging to accommodate the growing inflow of traffic [7].

Fire Brigade and Ambulance Call Outs for DCC are available at the following link <https://data.gov.ie/dataset/fire-brigade-and-ambulance/resource/e1c6721b-09c3-4e2a-bb27-b37ca75a9fed>, information on this data set has been merged with the RSA dataset to produce an estimate response time. The value is generated from the average response time for each fire brigade ambulance service for the year 2012 (a year that sits in the mid-range of the years covered in the RSA data set), and the average distance to the fire brigade stations. The result gives an indicative response-time/distance unit.

With this value, a simple rule of three was applied to the rest of the points to generate the estimated response time value [cell 77].

Response times and Distance to fire stations data both follow normal distribution hence the use of the mean as the estimator for this calculus [fig 4].

Go safe

GoSafe [9] is contracted by An Garda Síochána and the Department of Justice to provide and operate safety cameras in Ireland. The safety cameras are located in areas identified as Collision Prone Zones (CPZ’s) or Speed Enforcement Zones. Gosafe is an initiative to monitor the speed limit, as a risk factor for the accidents, however, there is still a high correlation of accidents occurring on the Gosafe locations, which indicates that other variables are risk factors of the accidents.

So, yet another indicator of the proneness of the accidents, however, they measure and have been created with a different purpose, let’s get the advantage of both and at the same time we can validate how accurate is the safety index compared to the GoSafe indicator.

The following comparison table puts together both indicators, goSave vs safety index.

* *P(A) = P(Accidents within zone)*
* *P(B) = P(Serious accidents within zone)*
* *P(C) = P(Fatal accidents within zone)*

|  |  |  |  |
| --- | --- | --- | --- |
|  | P(A) | P(B) | P(C) |
| ***Gosafe*** | 82.21% | 64.0% | 84.2% |
| ***Low prone to accidents***  ***Safety Index (<0.5)*** | 79.69% | 94.0% | 81.9% |
| ***High prone to accidents***  ***Safety Index (<0.5)*** | 20.31% | 6.0% | 18.1% |

From the table above, we can infer that the gosafe location is not related to the high proneness of the accidents, lower prone to accidents areas are a better indicator of the accidents, this can be seen on P(A) almost 80% of the accidents occur on areas with low value. This validates the selections of Gosafe locations, and the investment required to cover as much area as possible given the spread and randomness of this.

Distance to Fire brigade and ambulance services

Distance in meters to the closest fire brigade station calculating the shortest distance between 2 points on a sphere using the harvesine formula. Details of the station are available DCC website Dublin Fire Brigade. <https://www.dublincity.ie/residential/dublin-fire-brigade>

EDA

It is detailed below the process of data preparation of the RSA Dataset and as well the different techniques used to merge the rest of the data sets.

Data Preparation

After combining all the XML files it was produced a single excel file with 15.000 records on it. Names of the features have been mapped to include meaningful names to work with [cell 3]. All column names, category values, and labels have been added to a python class in the form of Enums (class) and Dictionaries available on *rsa/constants.py* file.

|  |
| --- |
| Collisions raw dataset description |
| RangeIndex: 15000 entries, 0 to 14999  Data columns (total 35 columns):  # Column Non-Null Count Dtype  --- ------ -------------- -----  0 id 15000 non-null int64  1 gps 15000 non-null object  2 splimit 15000 non-null int64  3 carrf 15000 non-null int64  4 carri 15000 non-null int64  5 gender 14203 non-null object  6 county 15000 non-null int64  7 pcycrf 15000 non-null int64  8 goodsri 15000 non-null int64  9 year 15000 non-null int64  10 pcycri 15000 non-null int64  11 goodsrf 15000 non-null int64  12 circumstances 15000 non-null object  13 pedri 15000 non-null int64  14 pedrf 15000 non-null int64  15 psvrf 15000 non-null int64  16 no\_unknown 15000 non-null int64  17 otherri 15000 non-null int64  18 vehicle 15000 non-null int64  19 vehicle\_type2 15000 non-null int64  20 vehicle\_type 15000 non-null int64  21 unknrf 15000 non-null int64  22 psvri 15000 non-null int64  23 no\_serious 15000 non-null int64  24 no\_notinjured 15000 non-null int64  25 unknri 15000 non-null int64  26 mcycrf 15000 non-null int64  27 mcycri 15000 non-null int64  28 no\_minor 15000 non-null int64  29 no\_fatal 15000 non-null int64  30 hour 15000 non-null int64  31 otherrf 15000 non-null int64  32 outcome 15000 non-null int64  33 age 15000 non-null object  34 weekday 15000 non-null int64  dtypes: int64(31), object(4)  memory usage: 4.0+ MB |

There was a list of undocumented columns created in the dataset with the suffix ***ri***or ***rf*** for the following categories *Car, Pedal Cycle Users, Goods Vehicle User, Pedestrians, PSV Users (Public Service Vehicles), Motor Cycle Users, and Unknown.* All columns type integer seems to contain counts related to the casualty record, However, It wasn’t clear what was counting, all ***rf*** columns have a value of 0, and concerning the ***ri*** columns, it was hard to relate the results with the rest of the information, [cell 12] hence it is preferred to remove these columns from the analysis.

Duplicate detection

It is known that data contains duplicated records. The removal of duplicated records was completed in 2 operations. Using column id, as the unique identifier of the casualty record, however, to verify these results, was also compared all column values across all the rows in the dataset using the duplicate function of the dataframe [cell 14] given a total of 4391 duplicated records that were removed from the dataset.

Columns analysis

The following transformations were applied to the columns

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Null** | **Transformation** | **Rationale** |
| gender | 578 | Artificially created a new type of gender for “Unknown” | ***[cell 17]*** Removing 5% of the dataset doesn’t justify removing those records, especially as the analysis is not interested in the distribution of the accidents by gender. |
| year | None | Format YYYY | ***[cell 19]*** Easier to read and interpret. |
| age | 642 | Replace with Unkown group | [cell 24] Same rationale as for gender. No reason to remove 6% of the dataset on this categorical value. |
| circumstances | 118 | Included in group “Other” | ***[cell 27]*** ditto |
| gps | None | Removed EPSG:900913  Replace latitude, longitude with EPS:4326 | ***[11] [cell 28]*** Standard GPS coordinates used in other datasets are in EPS:4326 format |
| vehicle | n/a | Removed | Undocumented column similar to vehicle\_type |
| vehicle\_type | None | Mapped undocumented categories to “Other” | ***[cell 32]*** Undocumented categories have been merged into the category “Other” |
| vehicle\_type2 | n/a | Removed | Undocumented column similar to vehicle\_type |
| total\_casualties | None | Sum of no\_minor, no\_serious and no\_fatal | ***[cell 35]*** Replace for no\_unkonwn, no\_notinjured, no\_minor, no\_serious, no\_fatal |
| outcome\_calculated | None | Map outcomes categories based on column values on no\_unkonwn, no\_notinjured, no\_minor, no\_serious, no\_fatal | ***[cell 35]*** To compare the original RSA column “outcome”. It was found that 6 record values didn’t match. As we have no evidence to justify if this is an error, it is best to stick to the count of casualties columns as recorded. |
| is\_fatal | None | Computed column | Helper for filters and other aggregations types |
| is\_dcc | None | Computed column | ***[cell 46] [code 2]*** For each location of the dataset it  Calculates if the point intersects with the polygon of the Dublin Committee Areas shapefile. |
| knn\_cluster | None | Computed | [cell 55][fig 11] Predicted value from the K-means cluster model indicating cluster zone. |
| fbs\_distance\_m | None | Computed | [cell 48] [code 3] Harvesine distance to the closest Firebrigade station |
| fbs\_station | None | Computed | [cell 48] [code 3] Harvesine distance to the closest Firebrigade station |
| fbs\_estimated\_response\_time | None | Computed | [cell 78] Calculated average of accidents to closest FB Station and average response time of ambulance service per station. This column was calculated using a simple rule of three methods. |

Ambulance Call Outs

This dataset is available on <https://data.smartdublin.ie> it contains measurements of the ambulance services response times for the year 2012. Data remediation, Null checks, and outliers removal using the IQR technique give an average response time of 11 minutes. [cells 59-78]

DCC Discovered Clusters Centers

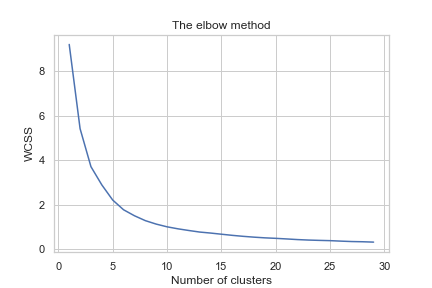
K-means clustering

As part of the data preparation, it has been generated a simple K-means cluster aggregation of the points with the intent of understanding how well the fire brigade station locations are distributed and prepared to cover emergency calls related to roads collision. The rationale is not trying to justify or not why a given station is located correctly or not, as there are other reasons to where they should be located, such as areas with high demographic, roads network, and as well distance to hospitals among others. The representation of how well the data points can be aggregated in the cluster can also help to understand how well the fire brigades are prepared to respond to the call-outs. Moreover, this is just an indicative reference to the reader as the K-means algorithm uses euclidian distance to refer to each point to the center of the calculated clusters, which may not be the optimal path when it comes to roads and traffic in Dublin. Before generating the model it was tested how well the data points can be aggregated on the cluster using the Hopkins statistic [12], which is a statistic that gives a value that indicates the clustering tendency, in other words: how well the data can be clustered by calculating the randomness of spatially distributed variables. In this case, it was used the latitude/longitude of the data points.

The results in the case of the RSA Dataset the score is 0.04 [code 3] which indicates there is a high tendency to cluster.

K-means model was calculated for the standard configuration values, given the high result score of the Hopkins indicator, there was no need to fine-tune the hyperparameters for cross-validation. That’s said, to find the optimum number of clusters it was calculated the sum of the squared errors (SSE) following the Elbow method to find the optimum number of clusters [13].

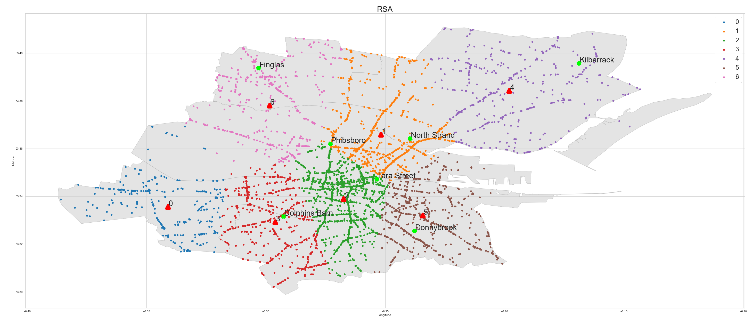
The ideal choice of the cluster in the RSA dataset is 7, which is when the Within cluster SSE (WCSS) errors start to flatten.



**[Fig 11] Elbow Method for determining the number of clusters**

As expected the stations are not located in the ideal centers of the cluster [fig 10], and we could argue that Phibsoro, North Strand, and Tara Street stations are too close to each other fighting for the same areas. We can also see in [fig 6] (*Radius 500 outliers on DCC map*), that the outliers where the highest accident concentration occurs, perfectly fit cluster center number 1 (orange color), in contrast to the actual stations closest to the city center.

For cluster number 3 (*red color on the map*) there seems to be a left alone and not covered by the stations. There is however one station outside of DCC, Tallagh station, that covers this area, and while it looks like a big area to cover for that station it’s also worth noting that the Phoenix Park is taking big of the space to it.



**[fig 10] Map of casualties and optimal cluster classification.**

After completing this model, a new column was included in the RSA dataset indicating membership to each of the 7 clusters generated by the K-means model [cell 55].

Data Dictionary Final

After all the previous steps it was generated the following data frame.

|  |
| --- |
| DCC Datarame final |
| Int64Index: 5245 entries, 0 to 5327  Data columns (total 23 columns):  # Column Non-Null Count Dtype  --- ------ -------------- -----  0 id 5245 non-null int64  1 year 5245 non-null int64  2 weekday 5245 non-null int64  3 hour 5245 non-null int64  4 splimit 5245 non-null int64  5 gender 5245 non-null int64  6 age 5245 non-null int64  7 county 5245 non-null int64  8 vehicle\_type 5245 non-null int64  9 circumstances 5245 non-null int64  10 latitude 5245 non-null float64  11 longitude 5245 non-null float64  12 total\_casualties 5245 non-null float64  13 outcome\_calculated 5245 non-null int64  14 is\_fatal 5245 non-null bool  15 severity 5245 non-null float64  16 knn\_cluster 5245 non-null int64  17 fbs\_distance\_m 5245 non-null float64  18 fbs\_station 5245 non-null int64  19 fbs\_estimated\_response\_time 5245 non-null float64  20 is\_gosafe 5245 non-null bool  21 radius\_500m 5245 non-null int64  22 safety\_index 5245 non-null float64  dtypes: bool(2), float64(7), int64(14)  memory usage: 911.7 KB |

Data Dictionary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Column** | **Name** | **Description** | **Source(s)** | **Type** | | **Options** |
|  | **Qualitative / Quantitative** | **Categorical Discr / Contin** |
| 0 | id | Unique identifier | RSA | Numeric | N/A | Enter number |
| 1 | year | Age group | RSA | Qualitative | Discrete | 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016 |
| 2 | weekday | Day of the week | RSA | Qualitative | Categorical | 1: "Sunday", 2: "Monday", 3: "Tuesday", 4: "Wednesday", 5: "Thursday", 6: "Friday",7: "Saturday" |
| 3 | hour | Hour of the day | RSA | Qualitative | Categorical | 1: "7am to 10am", 2: "10am to 4pm", 3: "4pm to 7pm", 4: "7pm to 11pm", 5: "11pm to 3am", 6: "3am to 7am" |
| 4 | splimit | Speed limit | RSA | Quantitative | Discrete | 10, 30, 40, 50, 60, 80, 100, 120 |
| 5 | gender | Gender | RSA | Qualitative | Categorical | 0: "Unknown", 1: "Male", 2: "Female |
| 6 | age | Group of age | RSA | Qualitative | Categorical | 1: "G1: 0-9", 2: "G2: 10-14", 3: "G3: 15-17", 4: "G4: 18-20", 5: "G5: 21-24", 6: "G6: 25-34", 7: "G7: 35-44", 8: "G8: 45-54", 9: "G9: 55-64", 65: "G11: 65 and Over", 99: "Unknown" |
| 7 | county | County | RSA | Qualitative | Categorical | Dublin |
| 8 | vehicle\_type | Vehicle\_Type | RSA | Qualitative | Categorical | 1: "Bicycle", 2: "Motorcycle", 3: "Car", 4: "Goods vehicle", 5: "Bus",6: "Other" |
| 9 | circumstances | Circumstances | RSA | Qualitative | Categorical | 1: "Pedestrian",2: "Single vehicle only", 3: "Head-on conflict", 4: "Head-on right turn", 5: "Angle, both straight", 6: "Angle, right turn", 7: "Rear end, straight", 8: "Rear end, right turn", 9: "Rear end, left turn", 10: "Other" |
| 10 | gps | Easting, Northing EPSG:900913 | RSA | Quantitative | Continuos | Pair values (Easting, Northing). EPSG the world is between 85.06°S and 85.06°N. and -180° left to 180° right |
| 11 | latitude | Latitude | RSA | Quantitative | Continuos | EPSG:4326 |
| 12 | longitude | Longitude | RSA | Quantitative | Continuos | EPSG:4326 |
| 13 | total\_casualties | Sum of total casualties involved in the same accident | Calculated | Quantitative | Discrete |  |
| 14 | outcome\_calculated | Outcome\_Calculated | Calculated | Qualitative | Categorical | 1: "Fatal", 2: "Serious", 3: "Minor",4: "Not Injured" |
| 15 | is\_fatal | Is fatal accident | Calculated | Qualitative | Categorical | True, False |
| 16 | is\_dcc | The location is within DCC Committee Area | DCC Committee Area | Qualitative | Categorical | True, False |
| 16 | severity | Severity | Calculated | Qualitative | Categorical | 0: "Low", 1: "High" |
| 17 | knn\_cluster | KNN asigned cluster | Calculated | Qualitative | Categorical | 1,2,3,4,5,6,7,8 |
| 18 | fbs\_distance\_m | Distance in meters to closest Firebrigade station | Calculate Firebrigade and ambulance | Quantitative | Continuos |  |
| 19 | fbs\_station | Associated Firebrigade station | Calculate Firebrigade and ambulance | Qualitative | Categorical | 0:'Tara Street', 1:'Donnybrook', 2:'Dolphins Barn',3:'Phibsboro', 4:'North trand', 5:'Finglas', 6:'Kilbarrack', 7:'Tallaght', 8:'Rathfarnham', 9:'Blanchardstown', 10:'Skerries', 11:'Balbriggan', 12:'Dun Laoghaire',13:'Swords' |
| 20 | fbs\_estimated\_response\_time | Ambulance estimated response time for the incident | Calculate Fire brigade and ambulance | Quantitative | Continuos | in minutes |
| 21 | is\_gosafe | The location is within the Gosafe zone | Calculated An Garda Síochaná | Qualitative | Categorical | True, False |
| 22 | radius\_500m | Count of accidents within 500m | Calculated | Quantitative | Discrete |  |
| 23 | safety\_index | Safety Index | Calculated | Quantitative | Continuos | Range[0,1] |

Descriptive Analysis of continuous variables

The following features of the RSA dataset are continuous variables

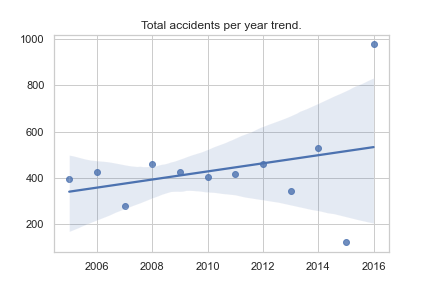
|  |  |  |
| --- | --- | --- |
|  | **fbs\_distance\_m** | **fbs\_estimated\_response\_time** |
| **count** | 5245 | 5245 |
| **mean** | 1425.042069 | 11.611601 |
| **std** | 831.92276 | 5.816285 |
| **min** | 5.778145 | 0.042097 |
| **25%** | 827.168846 | 7.563209 |
| **50%** | 1283.919247 | 11.027802 |
| **75%** | 1827.718788 | 15.268004 |
| **max** | 4912.301418 | 39.797644 |

|  |  |  |
| --- | --- | --- |
|  | **latitude** | **longitude** |
| **count** | 5245 | 5245 |
| **mean** | 53.3485 | -6.265413 |
| **std** | 0.022583 | 0.035264 |
| **min** | 53.30215 | -6.385915 |
| **25%** | 53.33159 | -6.284762 |
| **50%** | 53.34528 | -6.264842 |
| **75%** | 53.36117 | -6.245282 |
| **max** | 53.41045 | -6.143288 |

|  |  |  |
| --- | --- | --- |
|  | radius\_500m | safety\_index |
| **count** | 5245 | 5245 |
| **mean** | 89.857388 | 0.305352 |
| **std** | 71.80604 | 0.246756 |
| **min** | 1 | 0 |
| **25%** | 34 | 0.113402 |
| **50%** | 65 | 0.219931 |
| **75%** | 135 | 0.460481 |
| **max** | 292 | 1 |

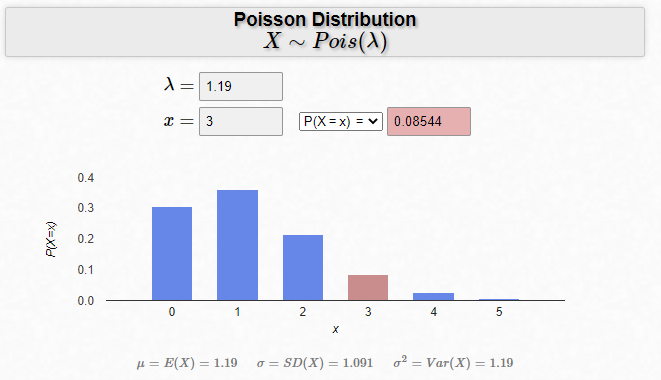
The 4 above variables all follow a Normal distribution [fig 4], [fig 5]

Except for the case of safety\_index, it is a chi-square distribution [fig 7].

The total number of accidents per year is increasing with the time as we can see in the trend regression line calculated below

**[fig 15] Total accidents per year trend.**

The daily number of accidents in Dublin for the years of study in this data set is 1.19 [cell 122], and the trend it’s growing. Using the probability mass function for Poisson, for a lambda value of 1.19 the probability of 3 accidents per day is 8%.



Data Visualization

A variety of plots has been used for anlysis of the RSA dataset. Will be described below all figures included in this study. But first, we must establish some conventions used for them.

Colour Palette

The RSA dataset is mainly composed of categorical variables (see data dictionary for reference) in most cases categories do not exceed 10 values. The luminosity and variate of the tab10 color palette have been set as the default for all the plots.



There is only one exception when it was required more color combinations. During the analysis of the clustering tendency, for the higher number of cluster values above 10, it was decided to choose the following selected colors from the xkcd palette with high contrast compared with tab10 palette for improving visualization: <https://xkcd.com/color/rgb/>

|  |  |
| --- | --- |
| xkcd:dark blue |  |
| xkcd:dark orange |  |
| xkcd:dark green |  |
| xkcd:dark red |  |
| xkcd:dark purple |  |

Data Visualization Tools

Matplotlib and Seaborn plots

Available on plots.py and in the Jupiter notebook file the following plots have been created:

* Histogram and distributions plots. [fig 1,4,5,7, 12]
* Crosstab visualizations Feature by Vehicle Type [fig 12-16]
* Box plots [Fig 1], [fig 3]
* Map visualizations [Fig 6, 9 10]

Machine Learning

This study aims to analyze the severity of the accidents and generate a model that can predict successfully the classification of the severity of an accident. As indicated in the data dictionary severity column is a category column of 2 values, high and low. High aggregates outcomes type serious and fatal. And a value of severity low for outcome types minor. The rationale for this column and dependent variable for the study has to do with the imbalance of data between minor outcome types and the rest. Even after this combination of the data sets it still results in an unbalanced dataset. To deal with this [14],[15] it is required to infer new data to the minority of the high severity class. This technique it’s called data sampling, and in our case over-sampling the minority category it’s done using Synthetic Minority Oversampling Technique. Random oversampling was also attempted but it was producing a higher level of overfitting in the final model.

To compare the different augmentation techniques and the severity outcome of an accident it was used three machine learning models, Random Forests (RF), Naïve Bayes, and Support Vector Machines (SVM). With python libraty sklearn.

For the 3 models, it was used different hyperparameter values for the cross-validation using GridSearchCV for the calculation of the best accuracy. The dataset was previously scaled using a standard scaler because of the variance of the continuous variables. Overall the 3 models have performed well, b

Results

|  |  |  |  |
| --- | --- | --- | --- |
|  | **accuracy** | **precision** | **recall** |
| RF | 0.940175348 | 0.945360825 | 0.935714286 |
| NB | 0.709128417 | 0.688405797 | 0.775510204 |
| SVC | 0.886539453 | 0.868217054 | 0.914285714 |

Conclusions

Roads collisions and casualties are a complicated subject of study that depends on many different variables which on some occasions fall into random events prediction models. In this paper, it has been studied a different view on how Dublin City Council and its infrastructure and initiatives have an impact on the outcome. The study offers to the reader a different approach to understanding how accidents can be classified out of the traditional human behaviors factors, climatology, and light conditions. The result gives a satisfactory level of confidence and accuracy in the three models proposed, and the techniques applied to accomplish them.

References

1. D. Mishra and R. Pandey, “Study of cab driver characteristics and accident proneness,” Technology, vol. 10, no. 4, pp. 82–95, 2019
2. Abdulhafedh, A. (2017) Road Crash Prediction Models: Different Statistical Modeling Approaches. Journal of Transportation Technologies, 7, 190-205.
3. Mohammed-Najeeb Mahama et al. (2018) ‘Emergency response time and pre-hospital trauma survival rate of the national ambulance service, Greater Accra (January – December 2014)’, BMC Emergency Medicine, 18(1), pp. 1–7. doi: 10.1186/s12873-018-0184-3.
4. Breen N, Woods J, Bury G, et alA national census of ambulance response times to emergency calls in IrelandEmergency Medicine Journal 2000;17:392-395
5. Chandar, S. et al. (2020) ‘Road Accident Proneness Indicator Based On Time, Weather And Location Specificity Using Graph Neural Networks’. Available at: https://search.ebscohost.com/login.aspx?direct=true&db=edsarx&AN=edsarx.2010.12953&site=eds-live (Accessed: 1 April 2022). Pag 3
6. Ariyanto, R. et al. (2019) ‘A Web and Mobile GIS for Identifying Areas within the Radius Affected by Natural Disasters Based on OpenStreetMap Data’, International Journal of Online & Biomedical Engineering, 15(15), pp. 80–95. doi: 10.3991/ijoe.v15i15.11507.
7. Department Of Transports, ‘Transport Trends 2020 An Overview of Ireland’s Transport Sector’, p 3. Available at: https://igees.gov.ie/wp-content/uploads/2021/03/Transport-Trends-2020.pdf
8. Azevedo, A. and Santos, M.F., 2008. KDD, SEMMA and CRISP-DM: a parallel overview. IADS-DM. (accessed 2 April 2022).
9. Health Safety Authority, go Safe Leading by Example, Available at: <https://www.hsa.ie/eng/vehicles_at_work/driving_for_work/vehicle_risk_management_business_case_studies/go_safe_case_study_dec_2017.pdf>
10. Beinecke, J., Heider, D. Gaussian noise up-sampling is better suited than SMOTE and ADASYN for clinical decision making. BioData Mining 14, 49 (2021). <https://doi.org/10.1186/s13040-021-00283-6>
11. EPSG:4326 Dataset: v10.059, EPSG Geodetic Parameter Dataset, maintained by the Geodesy Subcommittee of the IOGP Geomatics Committee. Available at: <https://epsg.org/crs_4326/WGS-84.html>
12. Hao, R. et al. (2021) ‘A Promising Approach: Artificial Intelligence Applied to Small Intestinal Bacterial Overgrowth (SIBO) Diagnosis Using Cluster Analysis’, Diagnostics (2075-4418), 11(8), p. 1445. doi: 10.3390/diagnostics11081445.
13. Lingbo Liu et al. (2018) ‘Fast Identification of Urban Sprawl Based on K-Means Clustering with Population Density and Local Spatial Entropy’, Sustainability, 10(8), p. 2683. doi: 10.3390/su10082683.
14. Michael Oladunjoye, A Comprehensive Analysis of Handling Imbalanced Dataset. (2021). International Journal of Advanced Trends in Computer Science and Engineering, 10(2), 454–463. <https://doi.org/10.30534/IJATCSE/2021/031022021>
15. Jacqueline Beinecke and Dominik Heider (2021) ‘Gaussian noise up-sampling is better suited than SMOTE and ADASYN for clinical decision making’, BioData Mining, 14(1), pp. 1–11. doi: 10.1186/s13040-021-00283-6.

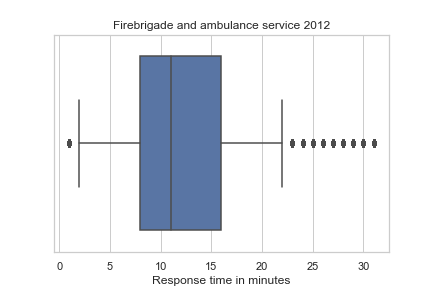
Jupiter Notebook References

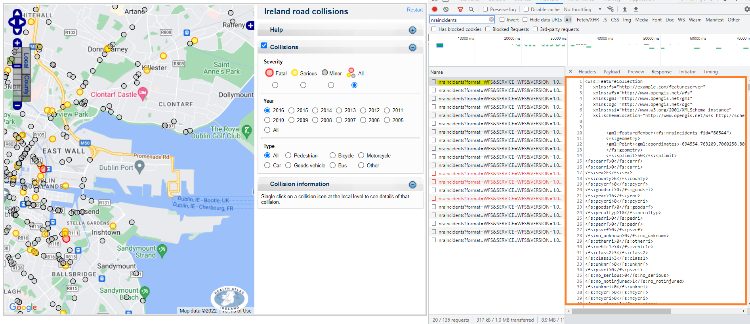
1. [cell 3] map XML columns to RSA columns
2. [cell 12] plot comparing ri columns with vehicle\_type
3. [cell 14] duplicate detection
4. [cell 17] gender mapping
5. [cell 19] year format yyyy
6. [cell 24] age group nulls mapping
7. [cell 27] circumstances null mapping
8. [cell 28] gps format conversation to latitud / longitude
9. [cell 32] vehicle\_type map types.
10. [cell 46] Calculate is\_dcc column
11. [cell 48] Calculate the closest FB station and distance in meters.
12. [cell 52] Hopkins statistic calculation
13. [cell 55] Calculate cluster membership with K-means prediction model.
14. [cell 66] 2012 results for ambulance services. Quartile 44.
15. [cell 77] get\_estimated\_response\_time
16. [cell 89] Calculate radius\_500 column.
17. [cell 122] Poisson distribution number of accidents 1 day.

RSA library References

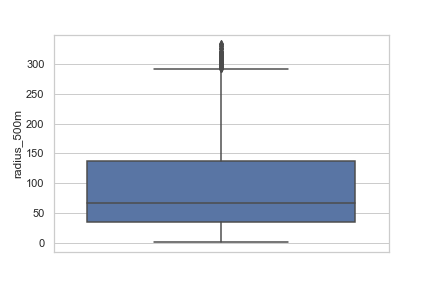
1. [code 1] ***total\_number\_accident\_on\_radius***, geo.py, line 124
2. [code 2] ***geo\_is\_dcc***: Calculate intersection with DCC Committee Areas. geo.py, line 25
3. [code 3] ***geo\_distance\_to\_closest\_fire\_station:*** geo.py, line 102

Table of figures

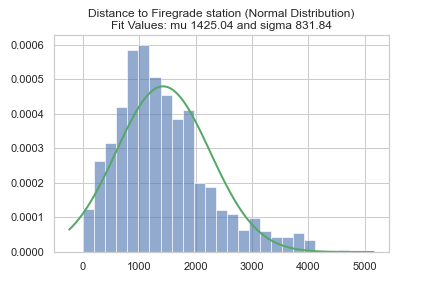
1. [Fig 1]: Response time year 2012
2. [fig 2]: RSA Map XML data download with Chrome extensions.

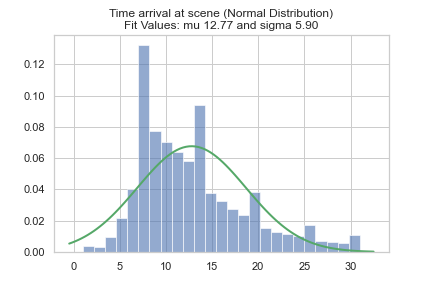
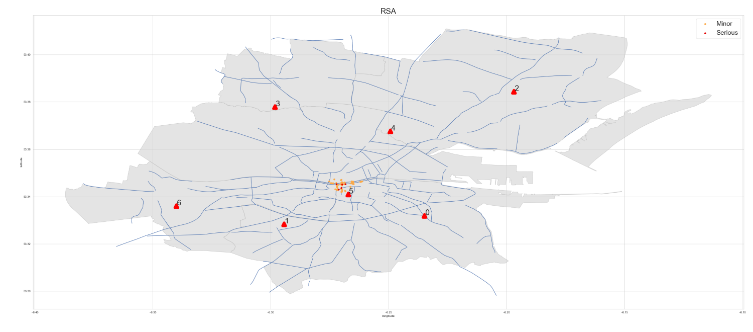
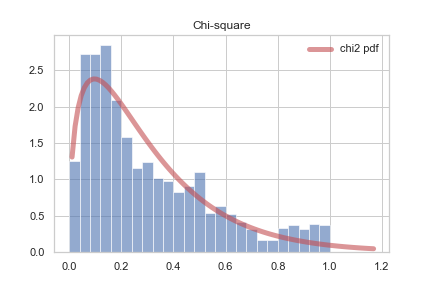
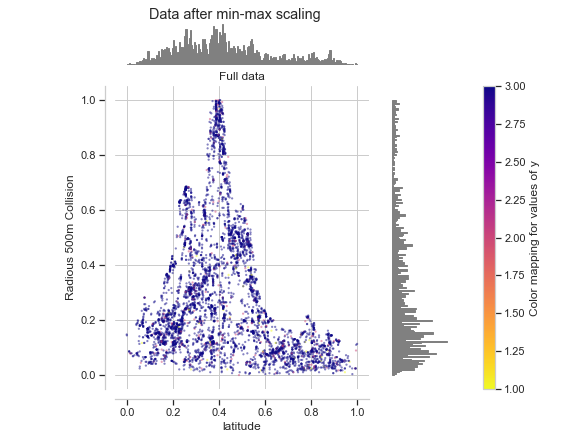
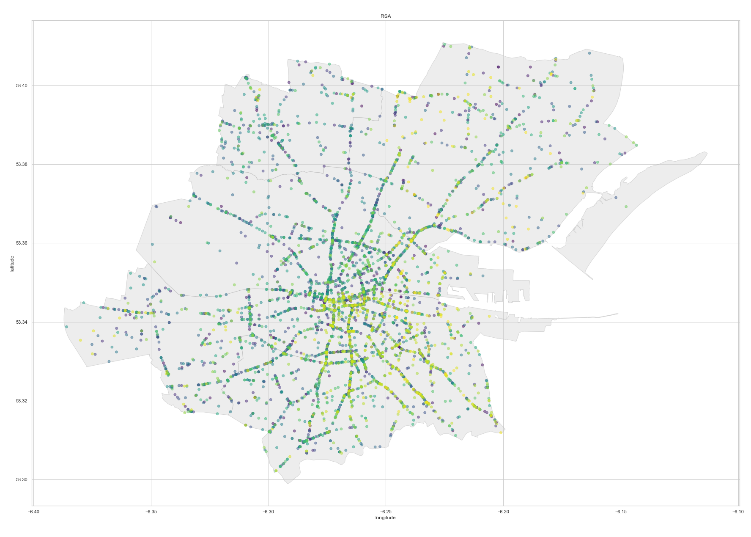
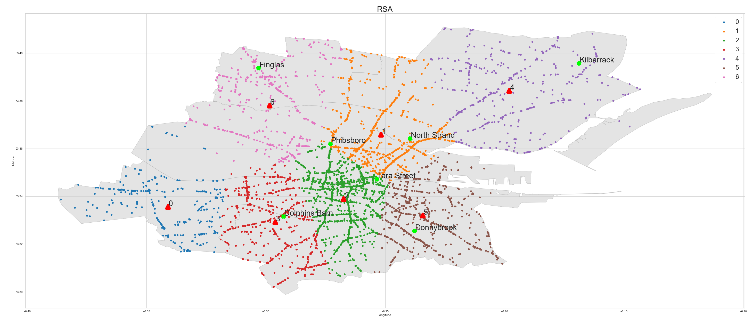
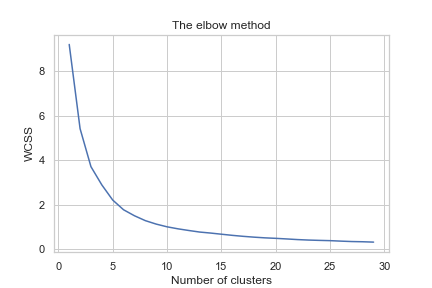
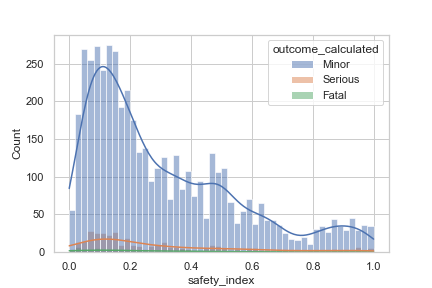
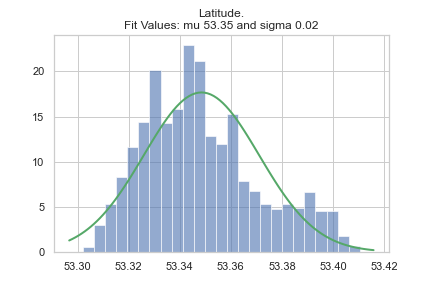
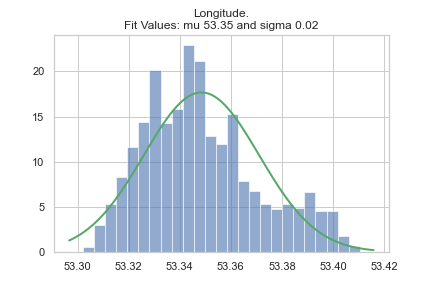
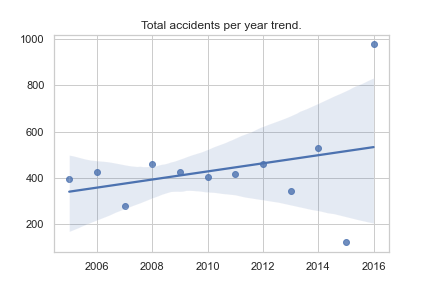


1. [fig 3]: Radius 500 / Safety index outliers plot



1. [fig 4] Distance to Firebrigade station histogram and normal distribution



1. [fig 5] Response time of ambulance service histogram and normal distribution  
   
2. [fig 6] Radius 500 outliers on DCC map  
   
3. [fig 7] Safety Index and Chi-Square distribution  
   
4. [fig 8] Safety index min max scaler plot
5. [fig 9] RSA map of casualties for DCC.  
   
6. [fig 10] Map of casualties and optimal cluster classification.  
   
7. [fig 11] Elbow Method for determining the number of clusters  
   
8. [fig 12] Safety index by outcome type
9. [fig 13] Latitude histogram with Normal distribution  
   
10. [fig 14] Longitude histogram with Normal distribution  
    
11. [fig 15] Accidents trend  
    
12. [fig 16] Heatmap correlation plot RSA Features