Traffic accident analysis

Reynaldo Gil García

September 17, 2019

gilcu2@gmail.com

Abstract

The ALLSTATS19 data have the accidents happened in Manchester area between 1979 and 2004. The data was analyzed to find relevant facts. Also a clustering algorithm bases on connected components was applied to find the ways region with high accident occurrence.

1 Profiling

The data consist of 3 csv comma separated files with headers:

- Accidents: It is the primary source with 6224198 rows. For each accident has an identifier Accident Index and 31 more features . Some features allow to appreciate the seriousness of the accident like Accident_Severity, Number_of_Vehicles, Number_of_Casualties but for example is not explicit the number of fatal casualties. Other features are related to factors that can contribute to the accident like the Day of the week, Date, Time, Road_Type, Light_Condition, Weather_Condition, Road_Surface_Conditions. Other fields are about the position where the accident happened like Location_Easting, Location_Northing. There is also other official fields

- Vehicles: 10981968 rows that describe the vehicles involved in the accident. The field Accident_Index allow to join with the accident. Vehicle_Reference is an index of the vehicles involved in the accident. It is not an identifier of the vehicle so no way to discover if the same vehicle is involved in several accidents. Another fields that can considered factors are Vehicle_Type, Sex_of_Driver and Age_Band_of_Driver and Age_of_Vehicle. The rest of fields describe how this vehicle was involved in the accident. Information about brand and model can be so interesting from a research point of view but is not available.

- Casualties: Describe the people involved in the accident. It have the fields Accident_Index and Vehicle_Reference that allow to link with the accident and the vehicle. Casualty_Reference allow to identify the casualty in this accident There is no way to know if the same people is involved in several accidents neither. The other fields describe the casualties age, sex and how it was involved in the accident.

The document Brief guide to road accidents and safety data: Great Britain included with in the zip file introduce the data and link to additional documents. The document Road-Accident-Safety-Data-Guide-1979-2004.xls expose the codification of the fields. Also explain that they also use -1 to represent lack of information but there are fields with NULL and empty values.

We compute the maximum and minimum values of each field to know the domain. They are shown in Appendix 5.

2 Integration in Spark

We read each csv file to a DataFrame. The data can be integrated joining the different dataframes using the Accident_Index and Vehicle_Reference fields. Depending of the query we are interested different joins can be done. For example, to know the ages of drivers involved in in fatal accidents must be joined Accidents and Vehicles while for know the number of fatal casualties per accident Accident and Casualties must be joined. We create 3 dataframes with all the possible joins and use they as needed.

3 Discoveries

3.1 Absolute frequencies

First we try to find the possibles more important causes of fatal accidents. For this we count the frequency of the values of fields that we consider can contribute to the accidents: Day of the week, Road_Type, Light_Condition, Weather_Condition, Road_Surface_Conditions, Vehicle_Type, Sex_of_Driver and Age_Band_of_Driver and Age_of_Vehicle when the accident have fatal casualties (Accident_Severity field equal 1).

The results (See Appendix 5) don't show something interesting because reflect the normal frequency of events. For example, the most frequent fatal accidents occurs when light conditions are Day Light because the cars move more frequently in the day not because the Day Light increase the probabilities of accident. So, we need some way to take into account the universe, for example the total numbers of car moving during the day. As this information is very difficult to have, we can use the total of accident cars in Day Light as an approximation.

3.2 Relative frequencies

The analysis of the relative frequencies shows some interesting facts. For example fatal accidents are more probable on Sunday, in Dark conditions without light, with Snowing and strong winds, the way with oil, male drivers with more than 75 years, old cars and big good vehicles.

The total results are shown in appendix 5

4 Accidents hot spots

To find the region with high accident count we create a graph where to point are connected in the difference between its coordinates is one in any edge. Then the hot spots are the connected components of this graph.

The algorithm was implemented using Spark graphX and allow to filtering by the number of accidents that happen in a point to be considered.

5 Conclusions

The application of statistic techniques to the accident traffic data have allow to discover interesting

Appendix

Domain

We compute the maximum and minimum values of each field to know the domain.

Accidents:

- Accident_Index 197901A11AD14 2004984164804
- Location_Easting_OSGR 0 9999
- Location_Northing_OSGR 0 12137
- Longitude -0.000001 NULL
- Latitude 49.912761 NULL
- Police_Force 1 98
- Accident_Severity 1 3
- Number_of_Vehicles 1 192
- Number_of_Casualties 1 90
- Date 01/01/1979 NULL

- Day_of_Week 1 7
- Time 00:01 NULL
- Local_Authority_(District) 1 941
- Local_Authority_(Highway) 9999 W06000024
- 1st_Road_Class -1 6
- 1st_Road_Number -1 9999
- Road_Type -1 9
- Speed_limit 0 6
- Junction_Detail -1 9
- Junction_Control -1 4
- 2nd_Road_Class -1 6
- 2nd_Road_Number -1 9999
- Pedestrian_Crossing-Human_Control -1 2
- Pedestrian_Crossing-Physical_Facilities -1 8
- Light_Conditions -1 7
- Weather_Conditions -1 9
- Road_Surface_Conditions -1 5
- Special_Conditions_at_Site -1 5
- Carriageway_Hazards -1 3
- Urban_or_Rural_Area -1 3
- Did_Police_Officer_Attend_Scene_of_Accident -1 3
- LSOA_of_Accident_Location E01000001 W01001842

Vehicles:

- Vehicle_Reference 0 201
- Vehicle_Type -1 113
- Towing_and_Articulation -1 5
- Vehicle_Manoeuvre -1 18
- Vehicle_Location-Restricted_Lane -1 9

- Junction_Location -1 -1
- Skidding_and_Overturning -1 5
- Hit_Object_in_Carriageway -1 11
- Vehicle_Leaving_Carriageway -1 8
- Hit_Object_off_Carriageway -1 10
- 1st_Point_of_Impact -1 4
- Was_Vehicle_Left_Hand_Drive? -1 -1
- Journey_Purpose_of_Driver -1 -1
- Sex_of_Driver -1 3
- Age_Band_of_Driver -1 11
- Engine_Capacity_(CC) -1 99999
- Propulsion_Code -1 9
- Age_of_Vehicle -1 99
- Driver_IMD_Decile -1 10
- Driver_Home_Area_Type -1 3

Casualties

- Casualty_Reference 0 991
- Casualty_Class 1 3
- Sex_of_Casualty -1 2
- Age_Band_of_Casualty -1 11
- Casualty_Severity 1 3
- Pedestrian_Location -1 10
- Pedestrian_Movement -1 9
- Car_Passenger -1 2
- Bus_or_Coach_Passenger -1 4
- Pedestrian_Road_Maintenance_Worker -1 0
- Casualty_Type -1 113
- Casualty_Home_Area_Type -1 3

Frequencies

- Field: Day_of_Week

  3 13486

  2 13977

  4 14088

  1 14540

  5 15300

  7 17618

  6 18524

- Field: Light_Conditions

  -1 32

  7 725

  5 931

  6 17659

  4 27043

  1 61143

- Field: Weather_Conditions

  -1 31

  6 177

  3 474

  9 625

  7 1126

  5 1873

  4 2421

  8 3871

  2 12566

  1 84369

- Field: Road_Surface_Conditions

  -1 151

  5 192

  3 558

  4 1828

  2 37053

  1 67751

- Field: Sex_of_Driver

  -1 254

  3 3804

  2 24839

  1 153534

- Field: Age_Band_of_Driver

  1 167

  2 417

  3 1434

  11 4083

  -1 5749

  10 6895

  9 13822

  8 21792

  4 25733

  5 27673

  7 31893

  6 42773

- Field: Age_of_Vehicle

  17 364

  16 546

  15 1006

  14 1458

  13 2104

  12 2944

  11 3573

  10 4166

  9 4519

  8 4876

  7 5086

  6 5321

  5 5603

  4 5993

  3 6038

2 6601

1 7255

-1 114356

- Field: Vehicle_Type

  18 5

  16 16

  -1 53

  105 72

  10 140

  17 181

  103 333

  20 456

  3 475

  108 928

  110 981

  2 1205

  90 2400

  21 2889

  106 3509

  11 4620

  1 6433

  19 10665

  104 15989

  113 16127

  109 114954

Relative frequencies

- Field: Day_of_Week

  3 0.015375915246623479

  4 0.015818460484726675

  2 0.016002573784493636

  5 0.01637833307106591

  6 0.017494302829744093

  7 0.01965816352717754

  1 0.020955628546350996

- Field: Light_Conditions
  7 0.013729760439352335
  1 0.01380652885352949
  -1 0.019464720194647202
  4 0.020075765730520168
  5 0.024775134387141413
  6 0.04952505419738787

- Field: Weather_Conditions
  9 0.011391182313594693
  3 0.01239539748953975
  2 0.014214337182719727
  8 0.015039843346297153
  6 0.01600361663652803
  -1 0.01615424700364773
  1 0.017844033641171413
  5 0.02011016030149135
  7 0.021678025489969582
  4 0.023396277469607065

- Field: Road_Surface_Conditions
  3 0.012620436965666983
  -1 0.013153310104529617
  4 0.014432680388766512
  1 0.01696968270042152
  2 0.01815607966272116
  5 0.022403733955659276

- Field: Sex_of_Driver
  3 0.007859374225219418
  2 0.009861187479281342
  -1 0.0110051993067591
  1 0.019297852002541475

- Field: Age_Band_of_Driver
  2 0.006420916481892092
  -1 0.006872913857248061
  1 0.008040055847094506

3 0.009036543175645445

6 0.01634565521531989

7 0.016724051317924253

4 0.016905792564552168

5 0.01790600721209686

8 0.018405731858419785

9 0.019745291172332022

10 0.023135332467645767

11 0.03197988627285117