Roberto Alcalde Diego, Alyssa Gladding, Darren Hoffmann-Marks

05/03/2019

Locating Traffic Accidents in the Town of Revere

Link To Github Repository: https://github.com/darren68/course-2019-spr-proj/tree/master/darren68_gladding_ralcalde

1. Introduction

We partnered with the city of Revere to analyze data sets showing traffic accident reports so that we could identify causes, hotspots and potentially design a way to prevent future accidents. In order to perform this analysis, we computed statistics that could be gathered from the provided data and transformed the state plane coordinates into latitude and longitude for geospatial mapping. We set the following goals for the project:

- Geospatially map the traffic accidents and display them in an interactive format with filter options.
- Analyze the evolution over time of the accidents within the city and consider potential causes while taking into account the accident severity, the key causes and consequences, and the area/neighborhood within Revere. By doing this, the end goal was to find the impact that different policies had on the number of accidents in different areas of the city, and based on these results, repeat or avoid them in the rest of the town, now or in the future.
- Compare the traffic safety of Revere with other towns of similar demographics in eastern MA.

The main goal of the project was to enable our partner from Revere to investigate what changes in the town over the past 20 years may have had an impact on traffic safety.

2. Original Data Sets & Milestone 1

In order to perform these tasks our Spark! partner provided us with the following data sets, which we converted into Mongo Collections: (Here are placed the collections, the original versions can be found in http://datamechanics.io/?prefix=darren68_gladding_ralcalde/)

Traffic Accidents in Revere and Stats on all other MA towns: (all provided by Spark partner) darren68_gladding_ralcalde.Revere2000

darren68_gladding_ralcalde.Revere2001to2016

darren68_gladding_ralcalde.Revere2014

darren68_gladding_ralcalde.Revere2015

darren68_gladding_ralcalde.Revere2016

darren68_gladding_ralcalde.AllTowns1990to2016

These data sets contain all the accidents in Revere over the past 18 years, taking into account causes for the accident (road condition, type of drive way, weather condition…) the consequences (injuries, fatalities…) and the location of the accidents (State Plain MA coordinates). Additionally, we have the total number of accidents in each MA city over the years 1990-2016.

From these data sets we developed the following collections:

Getting the coordinates of the accidents into Latitude and Longitude format: darren68_gladding_ralcalde.Revere2001to2016LatLng

Selecting and grouping by road conditions to understand the effect of this variable: darren68_gladding_ralcalde.RoadConditionsEffectRevere2016

Preparing the dataset of accidents to be an input for our cluster algorithm (selecting only coordinates and year): darren68_gladding_ralcalde.2001To2016XYyear

To understand the way cities of MA have evolved in a general way that scales with population and demographics across towns we computed the percentage change per year of each city (selection and projections): darren68_gladding_ralcalde.AllTownsRateofChange2010To2016

3. Optimization Algorithms, Statistical Analysis - Milestone 2

Optimization algorithms: One of the key goals was to identify which policies could have been successful or not in changing the number of accidents in different areas of the town. In order to properly analyze the different changes in the number of accidents within the city, we decided to cluster the accidents of the 2001 to 2016 interval and analyze how each cluster had evolved. The end goal of this was to enable us to examine how each cluster changed every year, and find out what specific measures or phenomena may have triggered these changes, similarly, it would allow policy makers to test if their past policies had any impact at all. To do this, as stated above, we transformed the data set into the "darren68_gladding_ralcalde.2001To2016XYyear" data set, which simply had the coordinates and the year where the accident took place.

The process of clustering the locations:

- Approach: We utilized the K-means algorithm to map each accident's location to a cluster mean point.
- Challenges: Having about 13,000 data points, where the X,Y coordinates were provided as strings, trying to bring all the data into an array that had to be iterated over multiple times until a mean point set was found, translated into runtimes of hours and often maxing out our computers' memories.
- Solution: To ease the process of assigning the cluster locations, we decided to hardcode the starting mean points in locations that were random but distributed across all coordinates. Secondly, instead of clustering all accidents of all years, we picked a subset of years distributed across all 15 years, and clustered the accidents of those years instead. Then we did a run over all the years and linearly mapped each accident to its closest cluster mean.
- Output: We decided to have each clustering run for each year be placed in its own collection, and thus ended up having 15 collections with tuples of the form:
    - (Coordinates of mapped Cluster Mean), ((Coordinates of the Accident), year)
    - Name of the collection for 2001: darren68_gladding_ralcalde.Clusters2001

Statistical Analysis: We performed three sets of statistical analyses, looking to answer the following changes:

1 - How has the city evolved? i.e. how did the overall number of yearly accidents change over the years from 2002 to 2016. The results of this analysis can be found in: darren68_gladding_ralcalde.MainStats

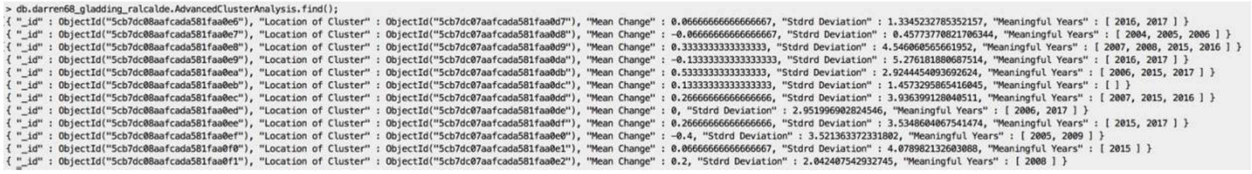
2 - How has the severity of accidents evolved over time? (i.e. partitioning the accidents between Non-Injury, Non-Fatal Injury, and Fatal Injury) The results of this analysis can be found in: darren68_gladding_ralcalde.SeverityStats

3 - The most important, complex, and meaningful question was: For each predetermined area of Revere (the clusters found by the optimization algorithm), did any of those areas have a significant change in the number of accidents recorded between consecutive years? We analyze these changes and used the standard deviation of the changes to detect any years where there were significant changes. The results for this analysis can be found in:

darren68_gladding_ralcalde.ClusterEvolution

darren68_gladding_ralcalde.AdvancedClusterAnalysis

Below we can find the last of these resulting collections:

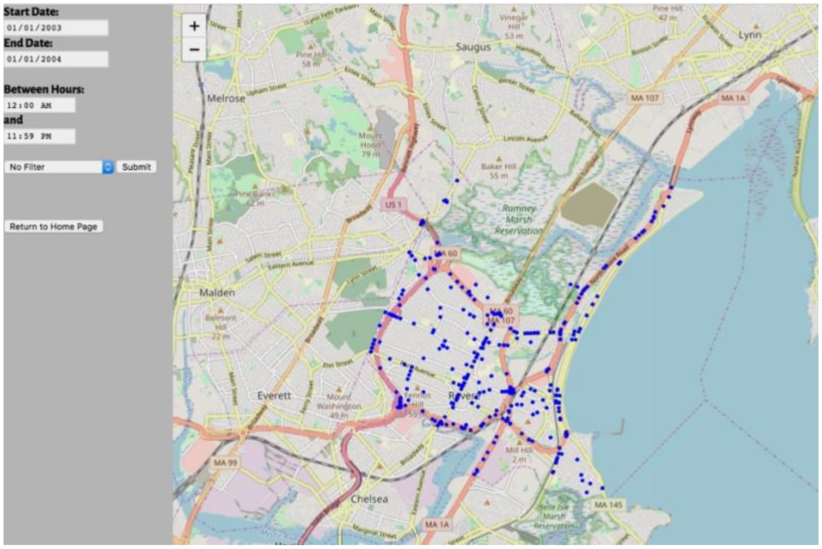


```
> db.darren68_gladding_ralcalde.AdvancedClusterAnalysis.find();
{ "_id" : ObjectId("5cb7dc08aafcada581faa0e6"), "Location of Cluster" : ObjectId("5cb7dc07aafcada581faa0d7"), "Mean Change" : 0.06666666666666667, "Stdrd Deviation" : 1.3345232785352157, "Meaningful Years" : [ 2016, 2017 ] }
{ "_id" : ObjectId("5cb7dc08aafcada581faa0e7"), "Location of Cluster" : ObjectId("5cb7dc07aafcada581faa0d8"), "Mean Change" : -0.06666666666666667, "Stdrd Deviation" : 0.45773770821706344, "Meaningful Years" : [ 2004, 2005, 2006 ] }
{ "_id" : ObjectId("5cb7dc08aafcada581faa0e8"), "Location of Cluster" : ObjectId("5cb7dc07aafcada581faa0d9"), "Mean Change" : 0.3333333333333333, "Stdrd Deviation" : 4.546060565661952, "Meaningful Years" : [ 2007, 2008, 2015, 2016 ] }
{ "_id" : ObjectId("5cb7dc08aafcada581faa0e9"), "Location of Cluster" : ObjectId("5cb7dc07aafcada581faa0da"), "Mean Change" : -0.13333333333333333, "Stdrd Deviation" : 5.276181880687514, "Meaningful Years" : [ 2016, 2017 ] }
{ "_id" : ObjectId("5cb7dc08aafcada581faa0ea"), "Location of Cluster" : ObjectId("5cb7dc07aafcada581faa0db"), "Mean Change" : 0.5333333333333333, "Stdrd Deviation" : 2.9244454093692624, "Meaningful Years" : [ 2006, 2015, 2017 ] }
{ "_id" : ObjectId("5cb7dc08aafcada581faa0eb"), "Location of Cluster" : ObjectId("5cb7dc07aafcada581faa0dc"), "Mean Change" : 0.13333333333333333, "Stdrd Deviation" : 1.4573295865416045, "Meaningful Years" : [ ] }
{ "_id" : ObjectId("5cb7dc08aafcada581faa0ec"), "Location of Cluster" : ObjectId("5cb7dc07aafcada581faa0dd"), "Mean Change" : 0.26666666666666666, "Stdrd Deviation" : 3.936399128040511, "Meaningful Years" : [ 2007, 2015, 2016 ] }
{ "_id" : ObjectId("5cb7dc08aafcada581faa0ed"), "Location of Cluster" : ObjectId("5cb7dc07aafcada581faa0de"), "Mean Change" : 0, "Stdrd Deviation" : 2.951996902824546, "Meaningful Years" : [ 2006, 2017 ] }
{ "_id" : ObjectId("5cb7dc08aafcada581faa0ee"), "Location of Cluster" : ObjectId("5cb7dc07aafcada581faa0df"), "Mean Change" : 0.26666666666666666, "Stdrd Deviation" : 3.5348604067541474, "Meaningful Years" : [ 2015, 2017 ] }
{ "_id" : ObjectId("5cb7dc08aafcada581faa0ef"), "Location of Cluster" : ObjectId("5cb7dc07aafcada581faa0e0"), "Mean Change" : -0.4, "Stdrd Deviation" : 3.521363372331802, "Meaningful Years" : [ 2005, 2009 ] }
{ "_id" : ObjectId("5cb7dc08aafcada581faa0f0"), "Location of Cluster" : ObjectId("5cb7dc07aafcada581faa0e1"), "Mean Change" : 0.06666666666666667, "Stdrd Deviation" : 4.078982132603088, "Meaningful Years" : [ 2015 ] }
{ "_id" : ObjectId("5cb7dc08aafcada581faa0f1"), "Location of Cluster" : ObjectId("5cb7dc07aafcada581faa0e2"), "Mean Change" : 0.2, "Stdrd Deviation" : 2.042407542932745, "Meaningful Years" : [ 2008 ] }
```

4. Results and Visualizations of the Statistical Analysis & Mapping - Milestone 3

As we can see from the screenshot above, our analysis allowed us to find the most volatile clusters, in terms of change per year and which years had the biggest change. (last column of each row – Meaningful Years) These results raise questions, what happened in those areas in such years? Looking at cluster 2, we can now ask, what happened in 2004, 2005 and 2006? Did any new policies help reduce accidents? Was it fortuitous? Was any route within the cluster under repairs? All these questions can now be asked for each year with specific meaning within each cluster, by our Spark! partner to the authorities in charge of traffic safety and infrastructure development.

To further support the analysis of accidents in the town, we performed a mapping of the accidents in the town, and built it within our web-based application allowing for the user to filter through the desired years. Below, a screenshot of the mapping for years 2002-2003 can be found:



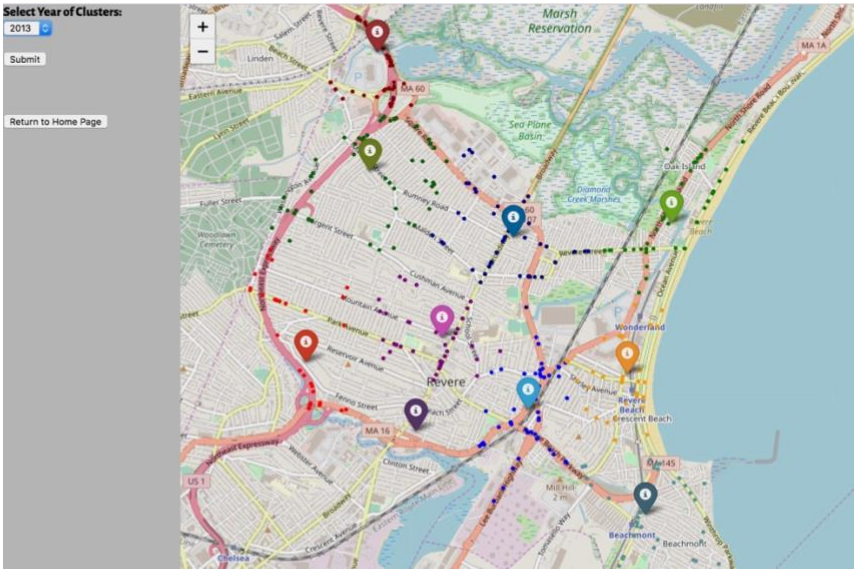
Start Date:
01/01/2003
End Date:
01/01/2004

Between Hours:
12:00 AM
and
11:59 PM

No Filter

As we can see above, the 2002 – 2003 map shows the main hotspots for accidents in the town, which are the entries and exits of the MA 60 and MA 107 routes and Broadway road. We also enabled some filtering methods to further examine the accident distribution in the town.

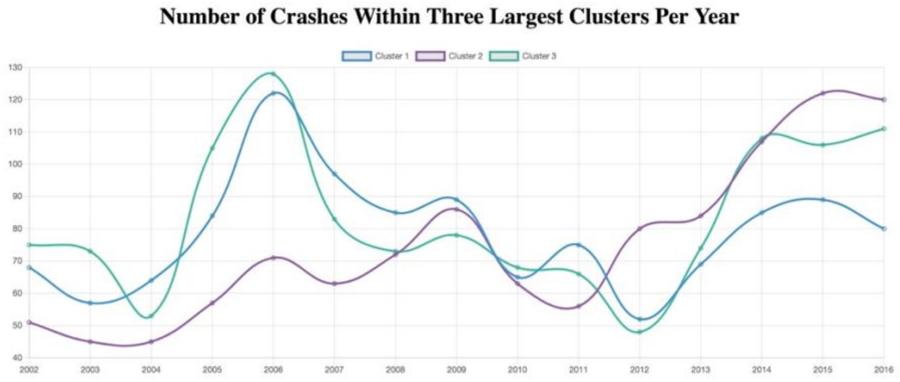
We also performed a mapping that includes our clusters, so that they become meaningful in terms of their location: - Again, we enabled some interactivity for the user (our Spark partner) to extract whatever information can be best leveraged for policy making.



Select Year of Clusters: 2013

To enable our web application server to execute quickly, we put CSVs of the computed clusters with their state plane coordinates converted to latitude and longitude in datamechanics.io. This saves the server a lot of time since it now doesn't have to convert every pair of coordinates to latitude and longitude.

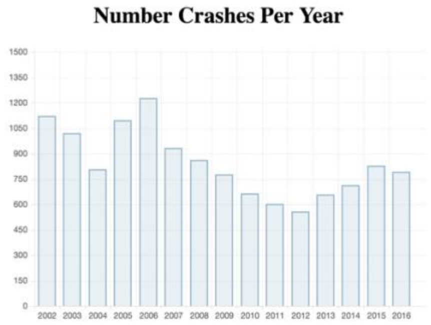
To further exploit the clusters, we performed a visualization of the evolution of the clusters over time, which can be found below:

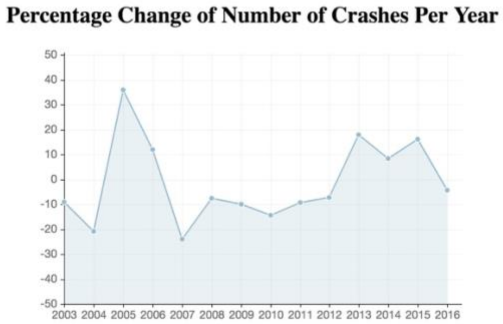


The most important take away comes from cluster 2, where we can see how in 2006 it kept traffic accidents low while the other largest clusters did not. It is worth asking now, what happened differently around cluster 2, and how it can be replicated.

Lastly, we performed a series of visualizations of trends over time, which can be found below:

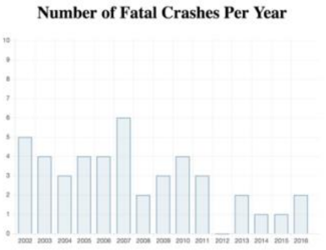
Evolution of Traffic Safety in Revere:

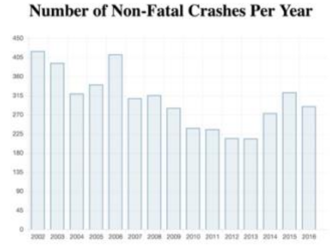


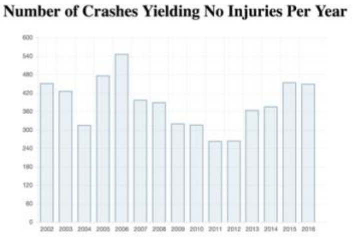


As we can see, the overall number per year has decreased, however, we can see how by the last 3 year, the number has increased from the all-time low in 2012, not only do these graphs show the overall trend, but enable us to understand if the changes per cluster are cluster specific or common to the entire town.

Evolution of Traffic Accident Severity:


Number of Fatal Crashes Per Year


Number of Non-Fatal Crashes Per Year


Number of Crashes Yielding No Injuries Per Year

The graphs above show the evolution of the numbers of accidents on each severity over the studied period of time. While the statistics agree with the general number of crashes per year, it is clear that there is a notorious improvement in terms of fatality overall. It would be worth asking, whether any policy helped this, or simply cars getting safer was the key driver.

5. **Future Work**

There are many areas to further explore to keep making Revere safer, for example, our main partner for this project was Revere, not Massachusetts, so it would be worth exploring only those areas where Revere can truly make an impact, discarding state routs for example. Also, based on our results, if some of these changes in certain areas can be attributed to some polices, this can provide us with data on the effect (in terms of accidents avoided) of a policy, and how much it costs, which could translate to an optimization problem where the impact of different policies can be projected over different areas of the town while trying to minimize the investment from the town.

-