A Study on Vehicles' accidents
in the United Kingdom

Laura C. Carles G.

Universitat Politecnica de Catalunya

June 13, 2018

Structure

Our work this semester

1. Improvements to the Naive Bayes Classifier
2. Accidents data preprocessing
3. Multivariate Data Visualization techniques

Structure

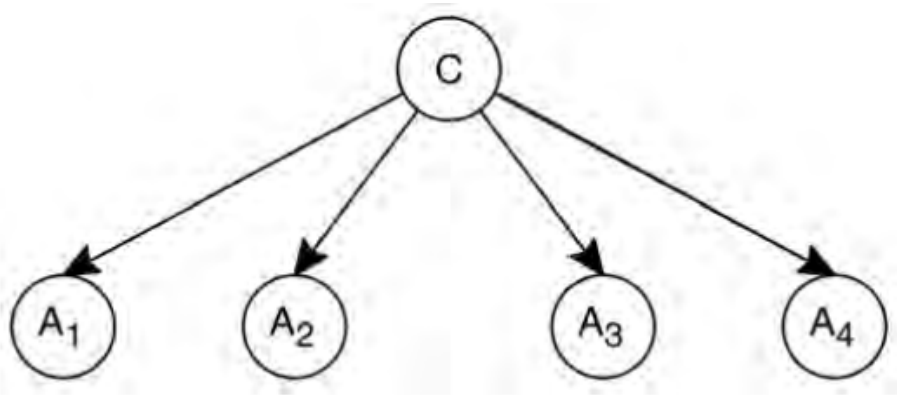
Our work this semester

1. Improvements to the Naive Bayes Classifier
2. Accidents data preprocessing
3. Multivariate Data Visualization techniques

Naive Bayes

Brief introduction

- C: Class to be estimated
- $A_x, x \in X$, where
  - X the number of attributes.
  - A_x attribute.



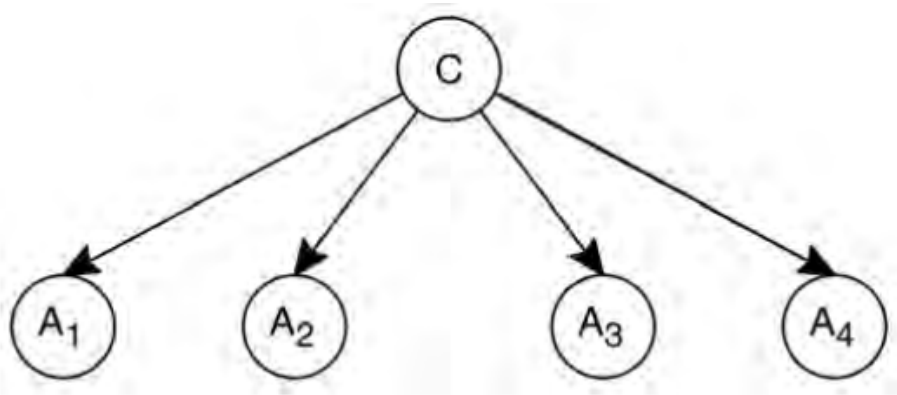
The estimated class can be computed as the class with maximum probability between classes given a_i.

$$c(E) = argmax_{c \in C} P(c) \prod_{i=1}^{n} P(a_i|c) \quad (1)$$

Naive Bayes

Brief introduction

- C: Class to be estimated
- $A_x, x \in X$, where
  - X the number of attributes.
  - A_x attribute.



The estimated class can be computed as the class with maximum probability between classes given a_i.

$$c(E) = argmax_{c \in C} P(c) \prod_{i=1}^{n} P(a_i|c) \quad (1)$$

Conditional independence assumption.

Independence assumption

What if data is dependent?

1. *IRL* data is not independent

Independence assumption

What if data is dependent?

1. *IRL* data is not independent
2. Nevertheless, although being a simple approach, **it works quite well**

Independence assumption

What if data is dependent?

1. *IRL* data is not independent
2. Nevertheless, although being a simple approach, **it works quite well**

The big question

Is there a way to improve Naive Bayes to take into account attribute dependencies?

Independence assumption

What if data is dependent?

1. *IRL* data is not independent
2. Nevertheless, although being a simple approach, **it works quite well**

The big question

Is there a way to improve Naive Bayes to take into account attribute dependencies?

More importantly...

If so, **would that give us better predictions?**

Improving NB

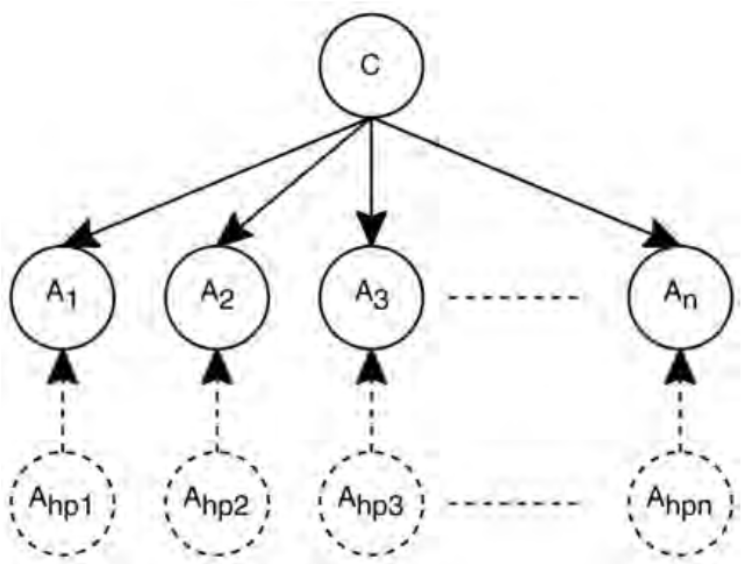
Many different extensions on NB

- Tree Augmented Naive Bayes (TAN)
- NbTree
- Averaged one-dependence estimators (AODE)
- Selective Bayesian Classifiers (SBC)
- **Hidden Naive Bayes (HNB)**

Hidden Naive Bayes

Its inner workings

- C: Class to be estimated
- $A_i, i \in I$, where
  - N the number of attributes.
  - A_i attribute.
  - A_{hpi} Weight dependency as hidden parent.



Predicted class can be computed as the class with maximum probability between classes given a_i and the relationship with all other attributes a_{hpi} .

$$c(E) = argmax_{c \in C} P(c) \prod_{i=1}^{n} P(a_i | a_{hp_i}, c) \tag{2}$$

Hidden Naive Bayes

Its inner workings II

- Taking the influences from other factors as a hidden parent:

$$P(a_i|a_{hp_i}, c) = \sum_{j=1;j\neq i}^{n} W_{ij} * P(a_i|a_j, c)$$

Hidden Naive Bayes

Its inner workings II

- Taking the influences from other factors as a hidden parent:

$$P(a_i|a_{hp_i}, c) = \sum_{j=1;j\neq i}^{n} W_{ij} * P(a_i|a_j, c)$$

- Weights are expressed as the conditional mutual information between variables:

$$W_{ij} = \frac{Ip(A_i; A_j|C)}{\sum_{j=1,j\neq i}^{n} Ip(A_i; A_j|C)}$$

Hidden Naive Bayes

Its inner workings III

- The conditional mutual information can be calculated as:

$$\mathrm{Ip}(\mathrm{A}_i; A_j|C) = \sum_{a_i,a_j,c} P(a_i, a_j, c) log \frac{P(a_i,a_j|c)}{P(a_i|c)P(a_j|c)}$$

Hidden Naive Bayes

Its inner workings III

- The conditional mutual information can be calculated as:

$$\mathrm{Ip}(A_i; A_j|C) = \sum_{a_i,a_j,c} P(a_i, a_j, c) log \frac{P(a_i,a_j|c)}{P(a_i|c)P(a_j|c)}$$

Computing probabilities

- Avoiding zero frequency:

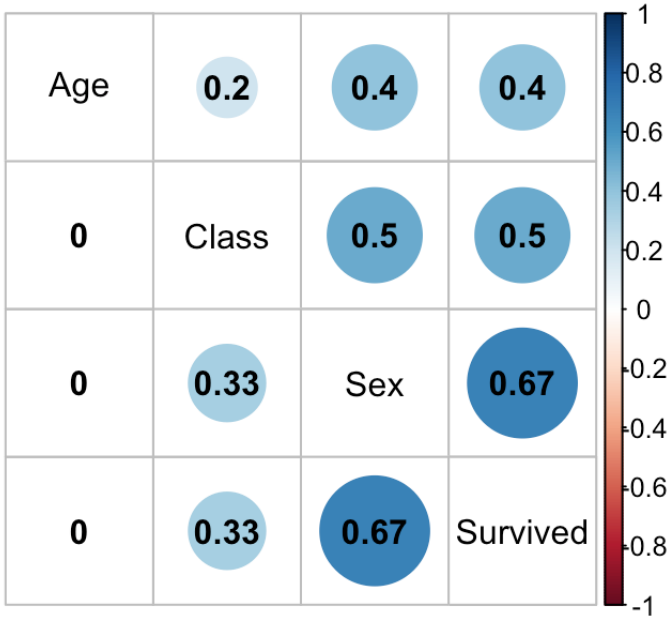
$$P(c) = \frac{F(c) + \frac{1}{\#classes}}{\#train + 1}$$

$$P(a_i|a_j, c) = \frac{F(a_i, a_j, c) + \frac{1}{\#val_i}}{F(a_j, c) + 1}$$

Hidden Naive Bayes

Results

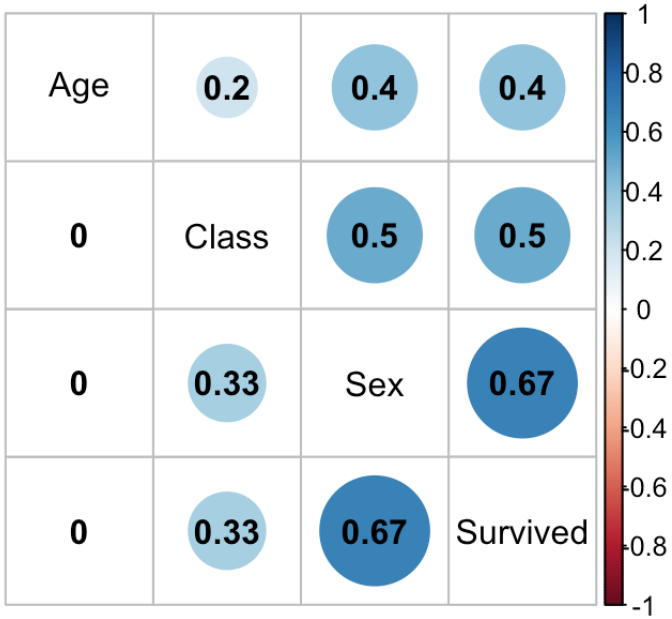
- Titanic dataset



Hidden Naive Bayes

Results

- Titanic dataset

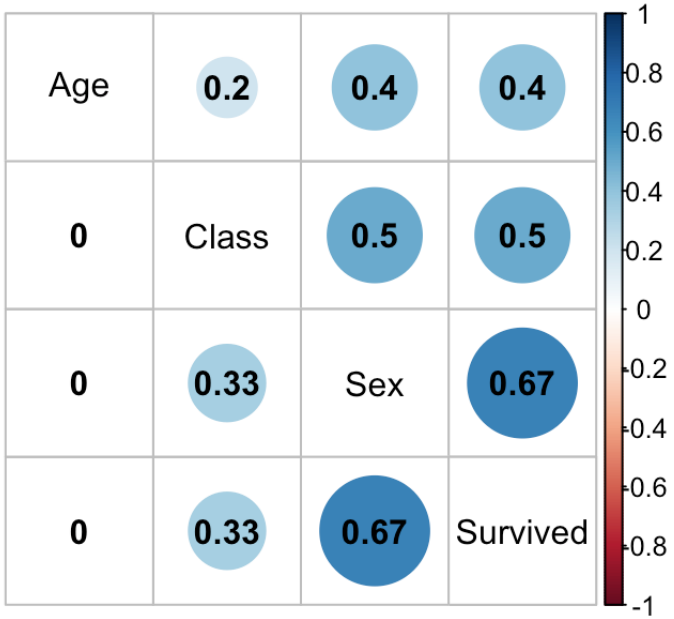


No real independence between variables!

Hidden Naive Bayes

Results

- Titanic dataset



No real independence between variables!

Attr.	NB (%)	HNB (%)
Class	41	46
Sex	73	78
Age	60	**95**
Survived	73	76

Structure

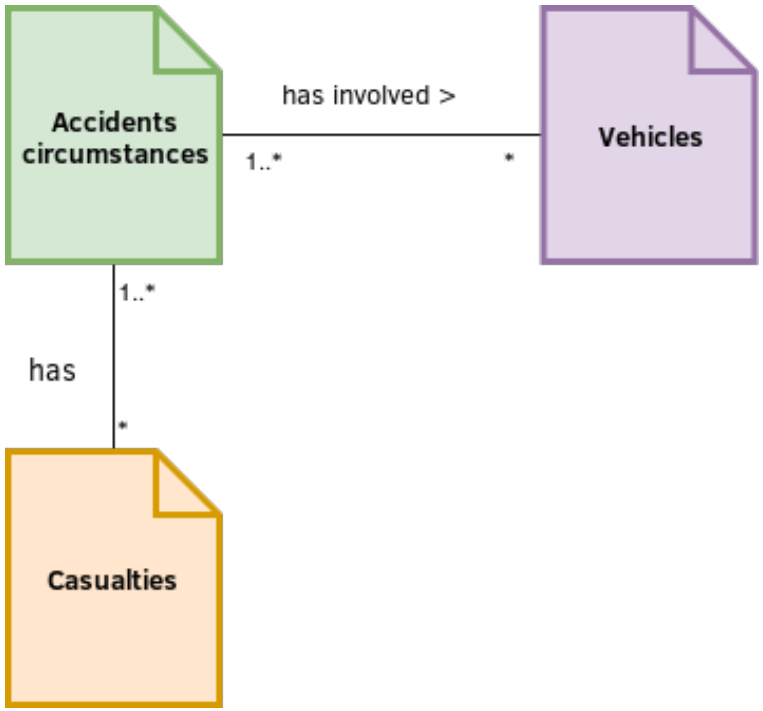
Our work this semester

1. Improvements to the Naive Bayes Classifier
2. Accidents data preprocessing
3. Multivariate Data Visualization techniques

First glance at data

Files

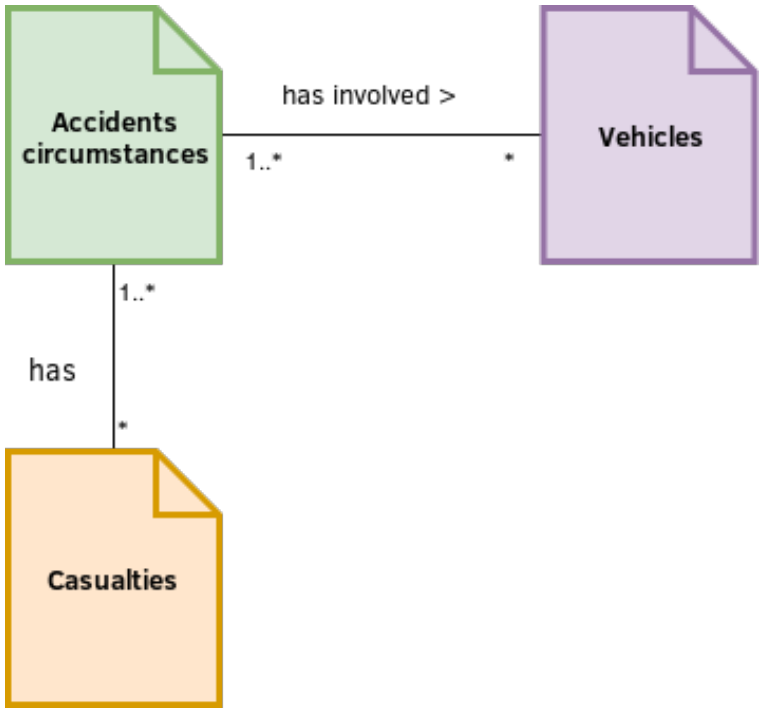
1. **Accident circumstances**, general circumstances and main information.
2. **Vehicles involved**.
3. **Casualties**, injured and death.



First glance at data

Files

1. **Accident circumstances**, general circumstances and main information.
2. **Vehicles involved**.
3. **Casualties**, injured and death.



No direct join of data!

First glance at data

Dimensionality

	# Individuals	# Variables
Accident	136.621	32
Vehicles	252.500	24
Casualties	181.384	16
Total	**570.505**	**72**

First glance at data

Dimensionality

	# Individuals	# Variables
Accident	136.621	32
Vehicles	252.500	24
Casualties	181.384	16
Total	**570.505**	**72**

Pretty large dataset

First glance at data

Missing values

At least 57.000 rows with missing data!

Main issues

to be dealt with

1. There is no direct join between files.
2. Dataset is considerably large.
3. Many missing values.
4. Very complex variables
   - Mainly categorical
   - More than 7 different categories

Preprocessing needed

Dimensionality reduction

Preprocessing - Dimensionality reduction

Removing variables

- **Objectively**: mostly empty
- **Subjectively**: uninteresting for us or not crucial.

Preprocessing - Dimensionality reduction

Removing variables

- **Objectively**: mostly empty
- **Subjectively**: uninteresting for us or not crucial.
  - Day of week - can be derived from date.
  - Local Authority District & Highway - mostly empty.
  - Road number & class.
  - Pedestrian Crossing Human Control - Only 1% had human control.
  - Police force
  - Urban or rural Area.
  - ...

Preprocessing - Dimensionality reduction

Missing values treatment

Still, many rows with still missing values.

	# rows	# NA	# rows NA	% rows NA
Accidents	136.575	1.939	1.083	0,79
Vehicles	252.500	97.971	81.748	**32,38**
Casualties	181.384	2.898	2.862	1,5

Preprocessing - Dimensionality reduction

Missing values treatment

Still, many rows with still missing values.

	# rows	# NA	# rows NA	% rows NA
Accidents	136.575	1.939	1.083	0,79
Vehicles	252.500	97.971	81.748	**32,38**
Casualties	181.384	2.898	2.862	1,5

Decisions taken

1. Remove rows in *Casualties* and *Vehicles* with NA.
2. Remove rows with unimputable values *(Latitude, Longitude, Time, Speed Limit)*.
3. Impute with Knn *Accidents*.

Feature processing

Preprocessing - Feature processing

Too many possible values that can be reduced to TRUE/FALSE or GOOD/BAD.

Simplifiable variables

- Light Conditions
- Weather Conditions
- Surface Conditions
- Special Conditions at site
- Carriageway Hazards

Preprocessing - Feature processing

Example

Weather

1. ✓ Fine no high winds
2. ✗ Raining no high winds
3. ✗ Snowing no high winds
4. ✗ Fine + high winds
5. ✗ Raining + high winds
6. ✗ Snowing + high winds
7. ✗ Fog or mist
8. ✗ Other

$\rightarrow$

1. Good weather (1)
2. Bad Weather [2 - 8]

Preprocessing - Feature processing

Example

Road Surface Conditions

1. ✓ Dry
2. ✗ Wet or damp
3. ✗ Snow
4. ✗ Frost or ice
5. ✗ Flood over 3cm
6. ✗ Oil or diesel
7. ✗ Mud

$\rightarrow$

1. Good Road Surface Conditions (1)
2. Bad Road Surface Conditions [2 - 7]

Preprocessing - Feature processing

Example

Road Surface Conditions

1. ✓ Dry
2. ✗ Wet or damp
3. ✗ Snow
4. ✗ Frost or ice
5. ✗ Flood over 3cm
6. ✗ Oil or diesel
7. ✗ Mud

$\rightarrow$

1. Good Road Surface Conditions (1)
2. Bad Road Surface Conditions [2 - 7]

Many more...

Preprocessing - Feature processing

Vehicles type reduction

20 different types of vehicles!

1. Pedal cycle
2. Motorcycle 50cc and under
3. Motorcycle 125cc and under
4. Motorcycle over 125cc and up to 500cc
5. Motorcycle over 500cc
6. Taxi/Private hire car
7. Car
8. Minibus (8 - 16 passenger seats)
9. Bus or coach (17 or more pass seats)
10. Ridden horse
11. Agricultural vehicle
12. Tram
13. Van / Goods 3.5 tonnes mgw or under
14. Goods over 3.5t. and under 7.5t
15. Goods 7.5 tonnes mgw and over
16. Mobility scooter
17. Electric motorcycle
18. Other vehicle
19. Motorcycle - unknown cc
20. Goods vehicle - unknown weight

Preprocessing - Feature processing

Vehicles type reduction

Example of Motorcycles grouping

- Motorcycle 50cc and under
- Motorcycle 125cc and under
- Motorcycle over 125cc and up to 500cc
- Motorcycle over 500cc
- Mobility scooter
- Electric motorcycle

$\rightarrow$

- Motorcycle

Preprocessing - Feature processing

Vehicles type reduction

Example of Cars grouping

- Van / Goods 3.5 tonnes mgw or under
- Taxi/Private hire car
- Car

$\rightarrow$

- Car

Preprocessing - Feature processing

Vehicles type reduction

Example of Cars grouping

- Van / Goods 3.5 tonnes mgw or under
- Taxi/Private hire car
- Car

$\rightarrow$

- Car

And more...

Preprocessing

Files join

Preprocessing - Files join

Steps

1. Correspondence of many to one. (Many Vehicles involved to one accident)

Preprocessing - Files join

Steps

1. Correspondence of many to one. (Many Vehicles involved to one accident) $\rightarrow$ **Count number of vehicles per accident.**

Preprocessing - Files join

Steps

1. Correspondence of many to one. (Many Vehicles involved to one accident) $\rightarrow$ **Count number of vehicles per accident.**
2. Aggregation applied to:
   - # Male drivers.
   - # Female drivers.
   - # Male casualties.
   - # Female casualties.
   - ...

Preprocessing - Files join

Steps

1. Correspondence of many to one. (Many Vehicles involved to one accident) $\rightarrow$ **Count number of vehicles per accident.**
2. Aggregation applied to:
   - # Male drivers.
   - # Female drivers.
   - # Male casualties.
   - # Female casualties.
   - ...
3. Computed the age disparity between
   - Drivers.
   - Vehicles.

Preprocessing - Files join

Steps

1. Correspondence of many to one. (Many Vehicles involved to one accident) $\rightarrow$ **Count number of vehicles per accident.**
2. Aggregation applied to:
   - # Male drivers.
   - # Female drivers.
   - # Male casualties.
   - # Female casualties.
   - ...
3. Computed the age disparity between
   - Drivers.
   - Vehicles.
4. Vehicles type joining.

Preprocessing - Files join

Steps

1. Correspondence of many to one. (Many Vehicles involved to one accident) $\rightarrow$ **Count number of vehicles per accident.**
2. Aggregation applied to:
   - # Male drivers.
   - # Female drivers.
   - # Male casualties.
   - # Female casualties.
   - ...
3. Computed the age disparity between
   - Drivers.
   - Vehicles.
4. Vehicles type joining.
5. Worst casualty in accident *(Fatal? Slightly injured?)*
6. ...

Preprocessing - Files join

Aggregation examples

Max age disparity example

- Driver 1: 60 years old. (MAX)
- Driver 2: 18 years old. (MIN)
- Driver 3: 30 years old.

Preprocessing - Files join

Aggregation examples

Max age disparity example

- Driver 1: 60 years old. (MAX)
- Driver 2: 18 years old. (MIN)
- Driver 3: 30 years old.
- Disparity: MAX - MIN = 42 years.

Preprocessing - Files join

Aggregation examples

Max age disparity example

- Driver 1: 60 years old. (MAX)
- Driver 2: 18 years old. (MIN)
- Driver 3: 30 years old.
- Disparity: MAX - MIN = 42 years.

Watch out when interpreting results. When disparity is 0, it may be because there is only one vehicle involved.

Preprocessing - Files join

Aggregation examples

Max age disparity example

- Driver 1: 60 years old. (MAX)
- Driver 2: 18 years old. (MIN)
- Driver 3: 30 years old.
- Disparity: MAX - MIN = 42 years.

Watch out when interpreting results. When disparity is 0, it may be because there is only one vehicle involved.

Accident type joining decision

1. Car + Car $\rightarrow$ Car Accident
2. Motorcycle + Motorcycle $\rightarrow$ Motorcycle Accident
3. Car + Motorcycle $\rightarrow$ Mixed
4. ...

Preprocessing Summary

Preprocessing

Main steps summary

1. Dimensionality reduction

Step applied

1. Interesting vars. selection
2. 57k rows with NA in one variable - **Removed var.**

Preprocessing

Main steps summary

1. Dimensionality reduction
2. Feature processing

Step applied

1. Interesting vars. selection
2. 57k rows with NA in one variable - **Removed var.**
3. Simplification of categorical vars

Preprocessing

Main steps summary

1. Dimensionality reduction
2. Feature processing
3. Missing Values treatment

Step applied

1. Interesting vars. selection
2. 57k rows with NA in one variable - **Removed var.**
3. Simplification of categorical vars
4. Imputation only on accidents with Knn (1083 rows, 0.79%)

Preprocessing

Main steps summary

1. Dimensionality reduction
2. Feature processing
3. Missing Values treatment
4. Joining

Step applied

1. Interesting vars. selection
2. 57k rows with NA in one variable - **Removed var.**
3. Simplification of categorical vars
4. Imputation only on accidents with Knn (1083 rows, 0.79%)
5. Aggregation of vehicles and casualties.

Result

Preprocessing result

Separated files dimensionality

	# Ind	# Vars	% Info. ret.	% Info. lost
Accidents	136.575	17	53,125	46,875
Vehicles	252.500	5	20,83	79,17
Casualties	181.384	5	31,25	68,75

Preprocessing result

Separated files dimensionality

	# Ind	# Vars	% Info. ret.	% Info. lost
Accidents	136.575	17	53,125	46,875
Vehicles	252.500	5	20,83	79,17
Casualties	181.384	5	31,25	68,75

Final dimensionality (joined)

	# Ind.	# Vars
Final ds.	136.575	31

Conclusions

1. A good and thorough preprocessing of a dataset this large takes a lot of time.

Conclusions

1. A good and thorough preprocessing of a dataset this large takes a lot of time.
2. Must take many decisions.
3. Missing data. Not always imputable.

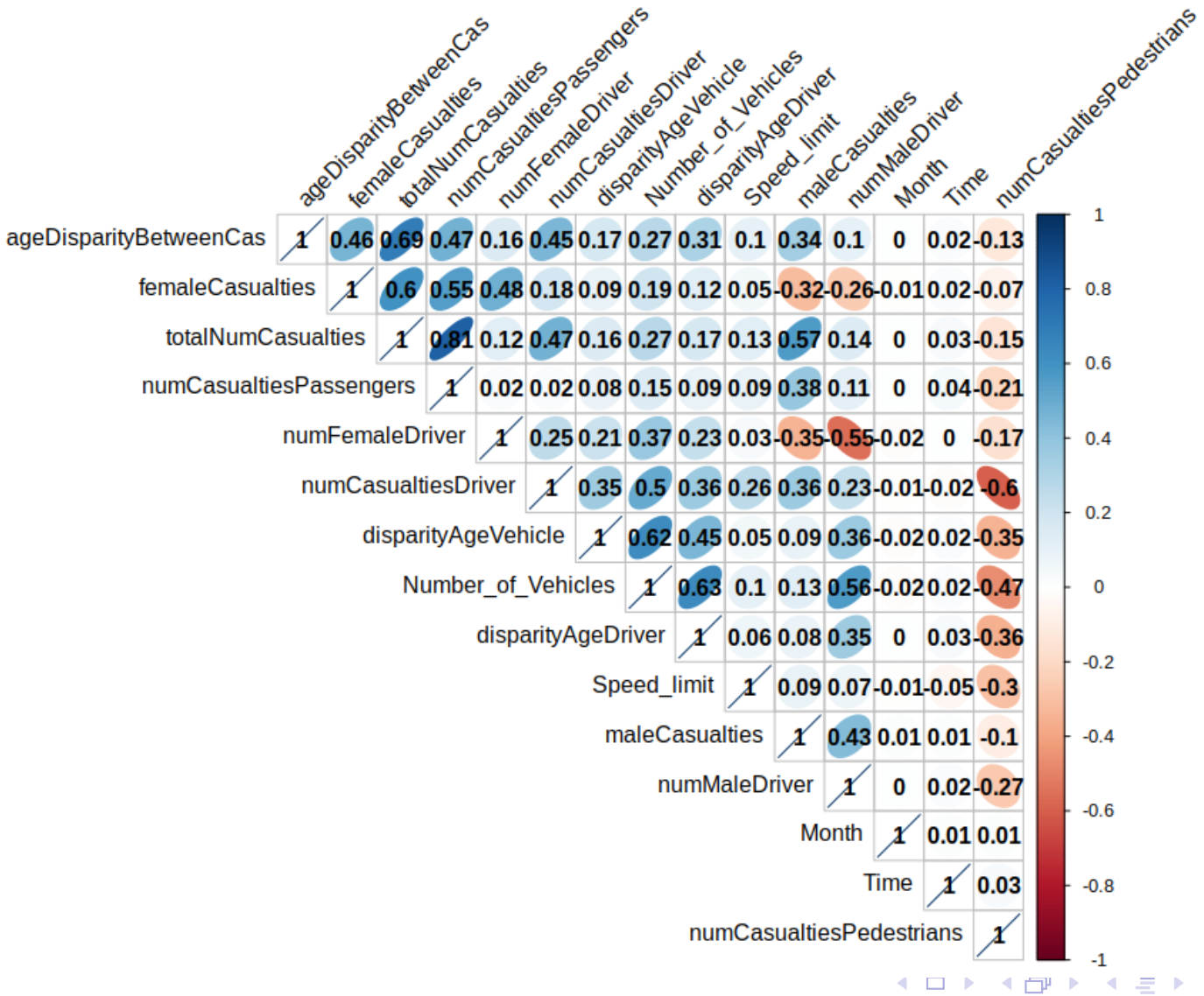
Conclusions

1. A good and thorough preprocessing of a dataset this large takes a lot of time.
2. Must take many decisions.
3. Missing data. Not always imputable.
4. Many source files $\rightarrow$ High difficulty on managing the data.
5. Joining is not direct.

The greater the number of variables, the more complex the decisions to take about it.

Multivariate Plotting

Involving variables correlation



Applying the HNB classifier

Attribute as class	% accuracy NB	% accuracy HNB
Accident Severity	67.05	**79.04**
Road Type	8.15	**82.89**
Speed Limit	11.64	**66.75**
Road conditions	78.31	**83.46**
Weather conditions	82.39	82.29

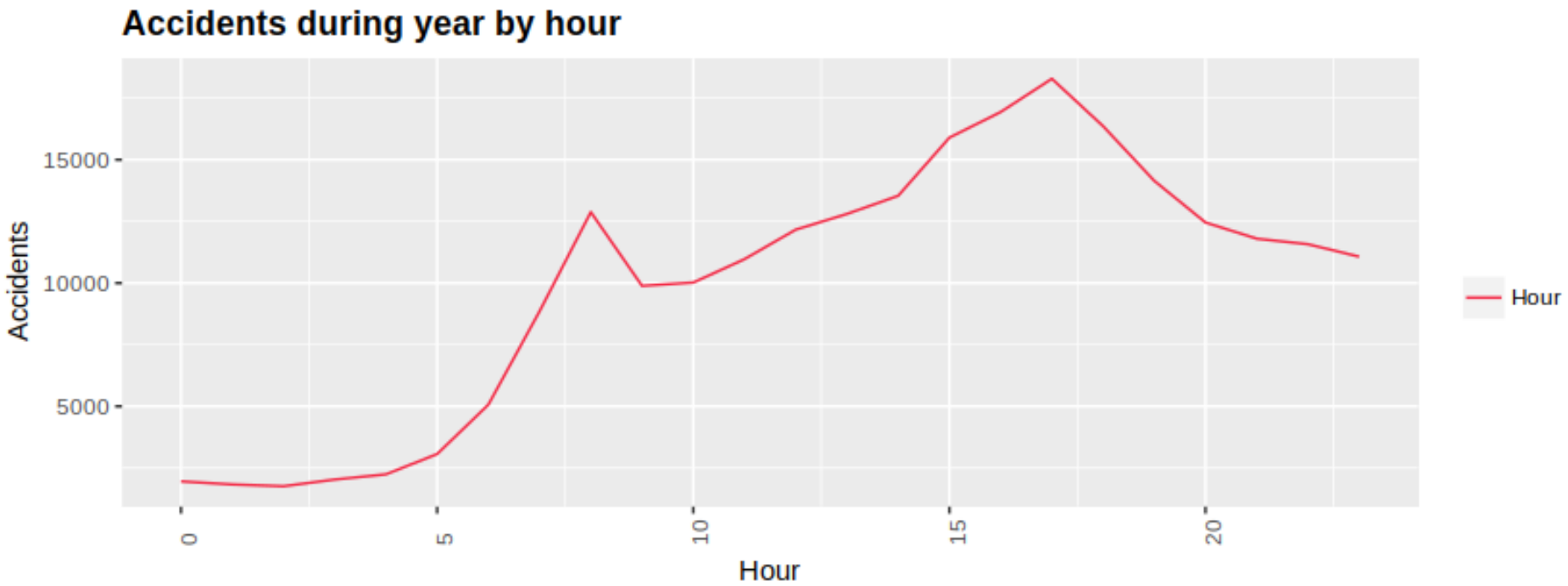
Structure

Our work this semester

1. Improvements to the Naive Bayes Classifier
2. Accidents data preprocessing
3. Multivariate Data Visualization techniques

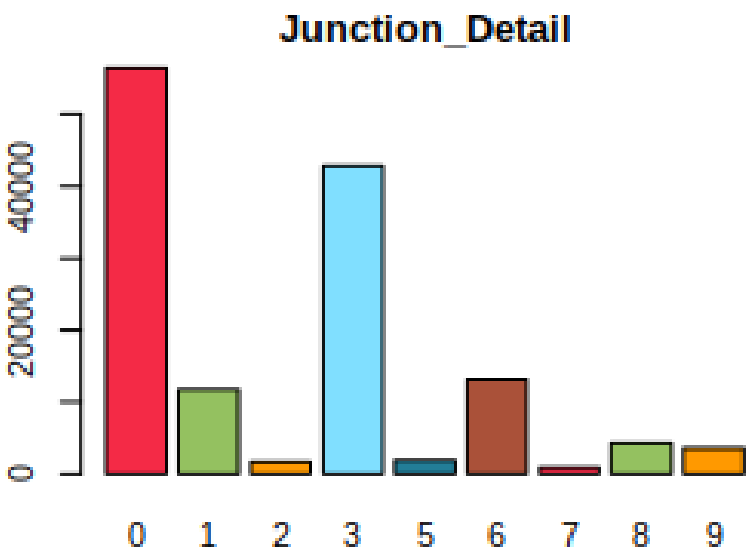
Simple Plotting

Accidents per hour evolution



Simple Plotting

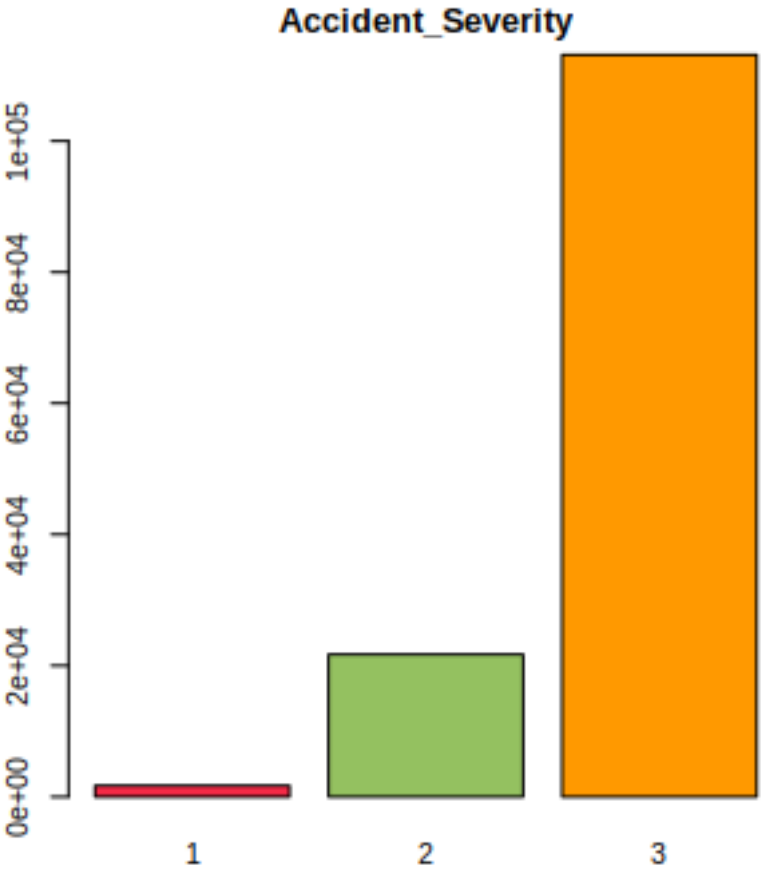
Histograms



1. Not at junction or within 20 metres
2. Roundabout
3. Mini-roundabout
4. T or staggered junction
5. Slip road
6. Crossroads
7. More than 4 arms (not roundabout)
8. Private drive or entrance
9. Other junction

Simple Plotting

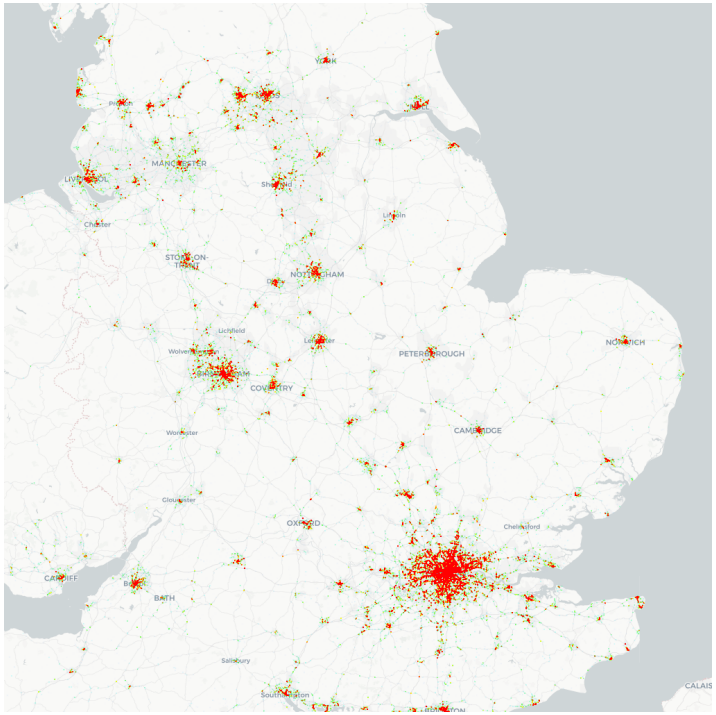
Histograms



1. Slight
2. Serious
3. Fatal

Multivariate Plotting

Heatmap representation

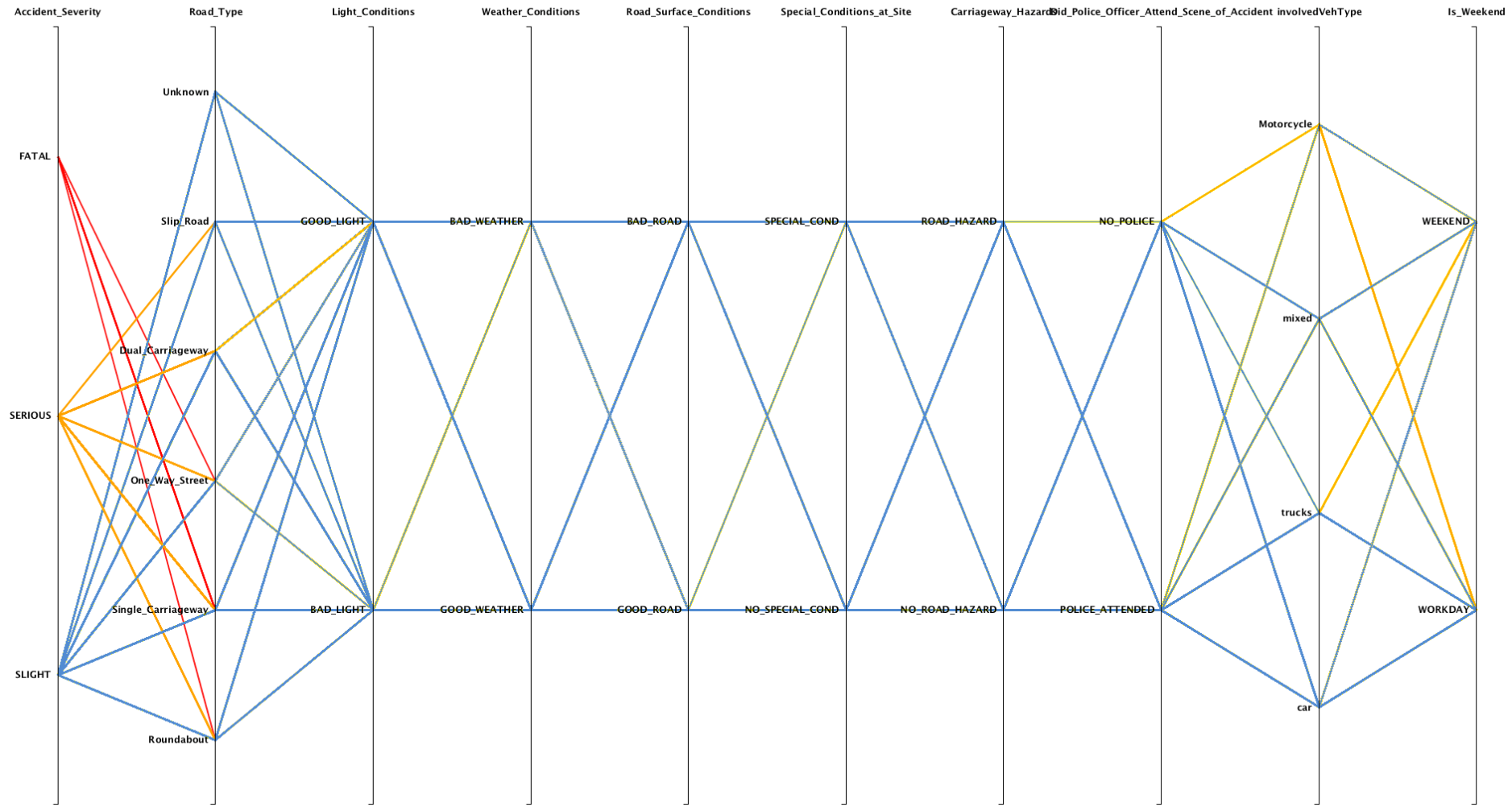


Density of accidents remains in the cities!

Multivariate Plotting

KNIME parallel sets

Parallel Coordinates Plot

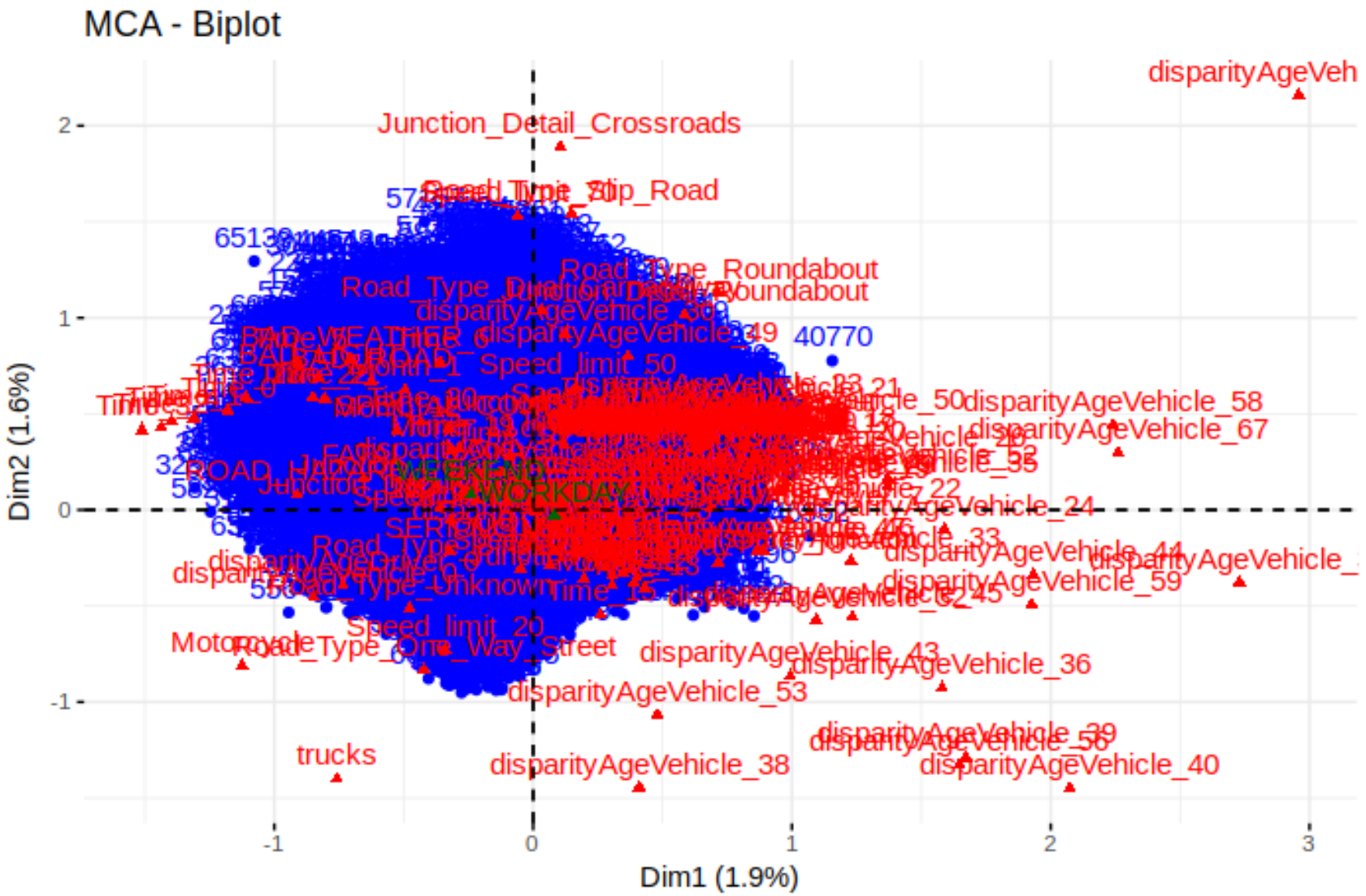


Mixed type attributes

- Principal Components analysis (PCA)
- Multiple Correspondance analysis (MCA)

Beyond Plotting

MCA - Results



- Low Inertia with the factorial plane (3%).
- Dimensionality reduction, retaining up to 90% of inertia.
- Homogeneous cloud of individuals in the biplot. No feasible explanation at first. Clustering will be necessary to distinguish clusters.

Clustering Approach

- No time based division $\Longrightarrow$ dealing with the whole dataset.

Clustering Approach

- No time based division $\Longrightarrow$ dealing with the whole dataset.
- Too much data for a hierarchical clustering

Clustering Approach

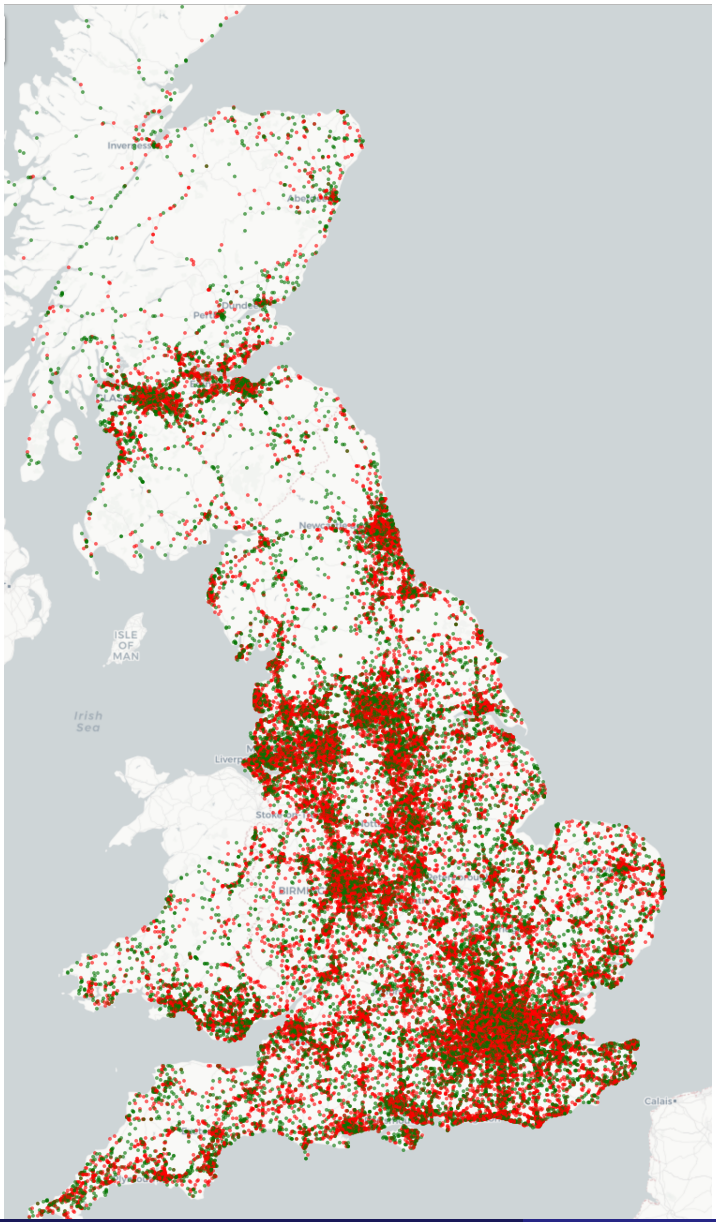
- No time based division $\Longrightarrow$ dealing with the whole dataset.
- Too much data for a hierarchical clustering
- Alternative methods: Iterative kmeans

Clustering Approach

- No time based division $\Longrightarrow$ dealing with the whole dataset.
- Too much data for a hierarchical clustering
- Alternative methods: Iterative kmeans
- Discovery # num clusters?

Beyond Plotting

Clustering

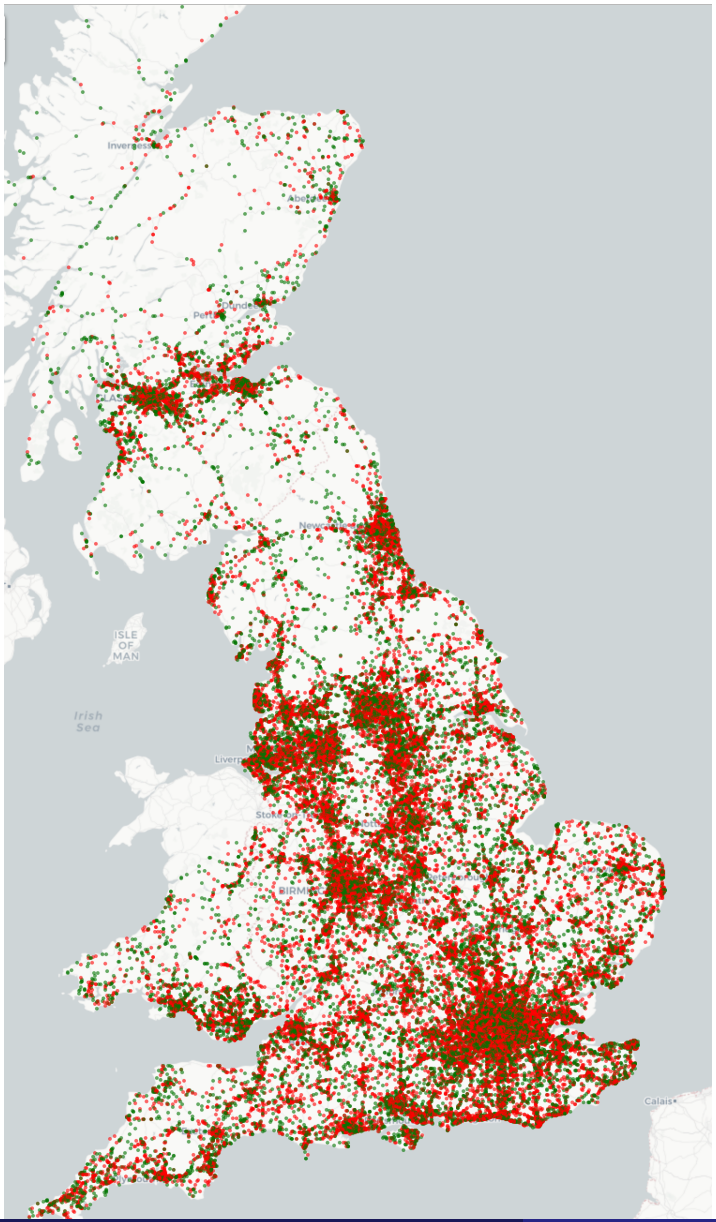


Identified clusters:

- Mixed vehicle type, junction based. No special conditions. Different vehicle age.

Beyond Plotting

Clustering

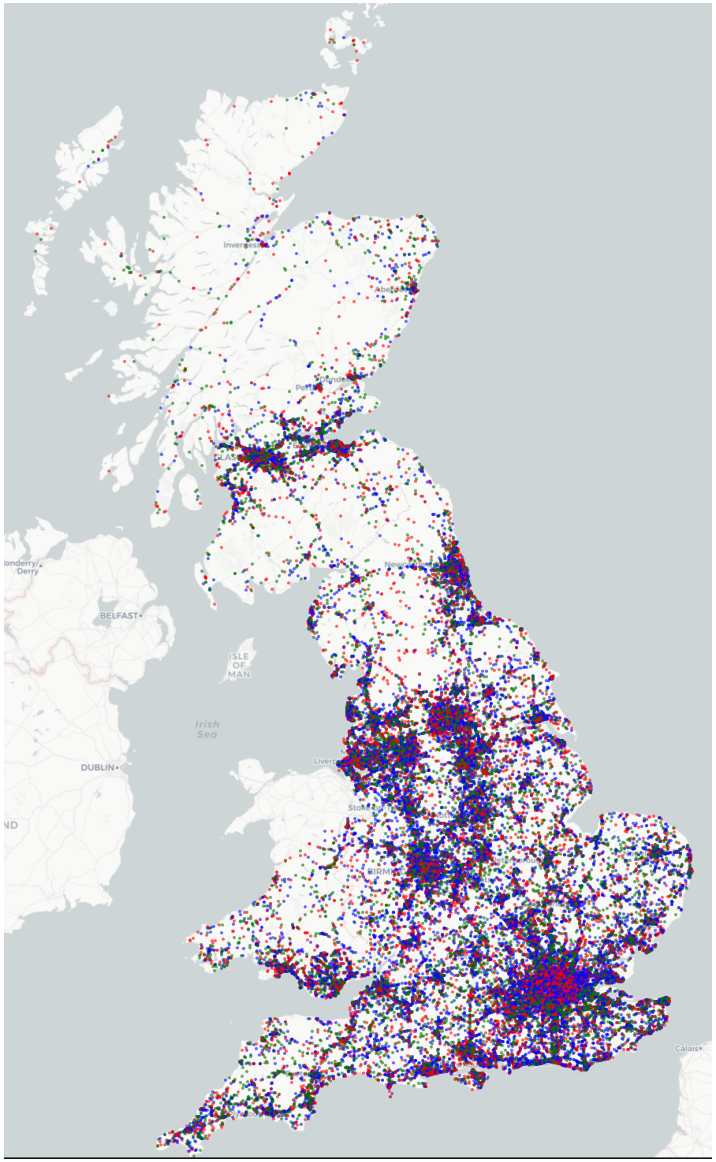


Identified clusters:

- Mixed vehicle type, junction based. No special conditions. Different vehicle age.
- Low speed, truck / motorcycles involved. Accident happened arround midnight.

Beyond Plotting

Clustering

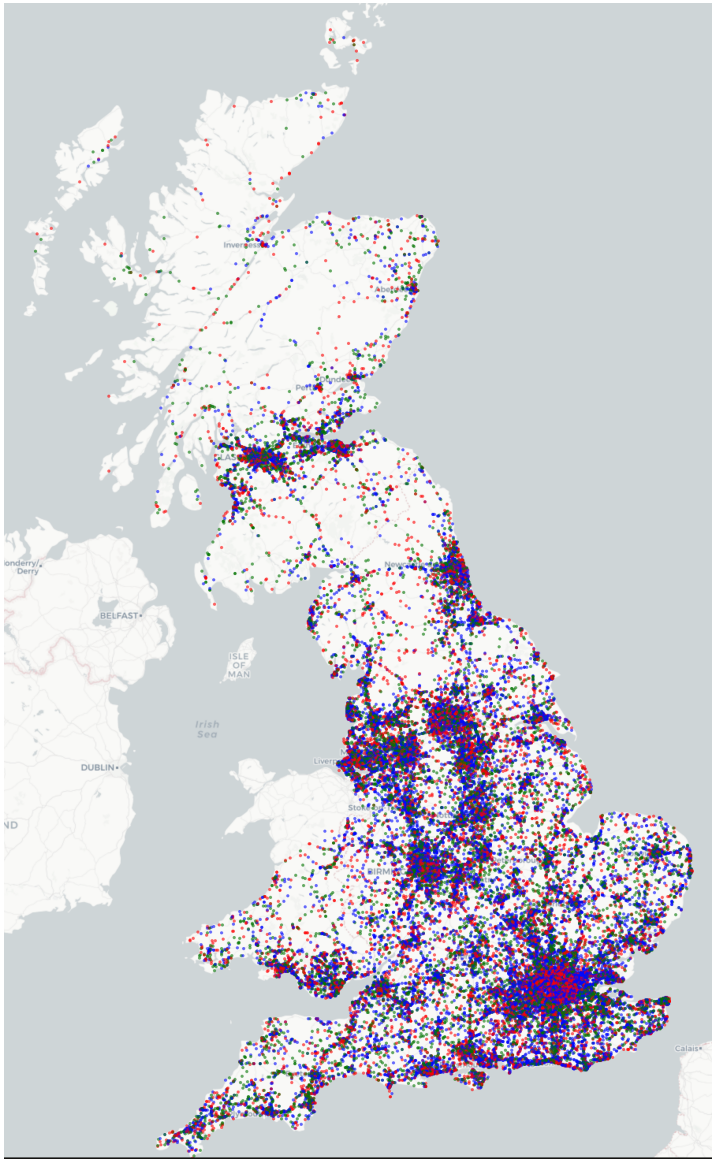


Identified clusters:

- Low speed, truck / motorcycle mixed with good conditions.

Beyond Plotting

Clustering

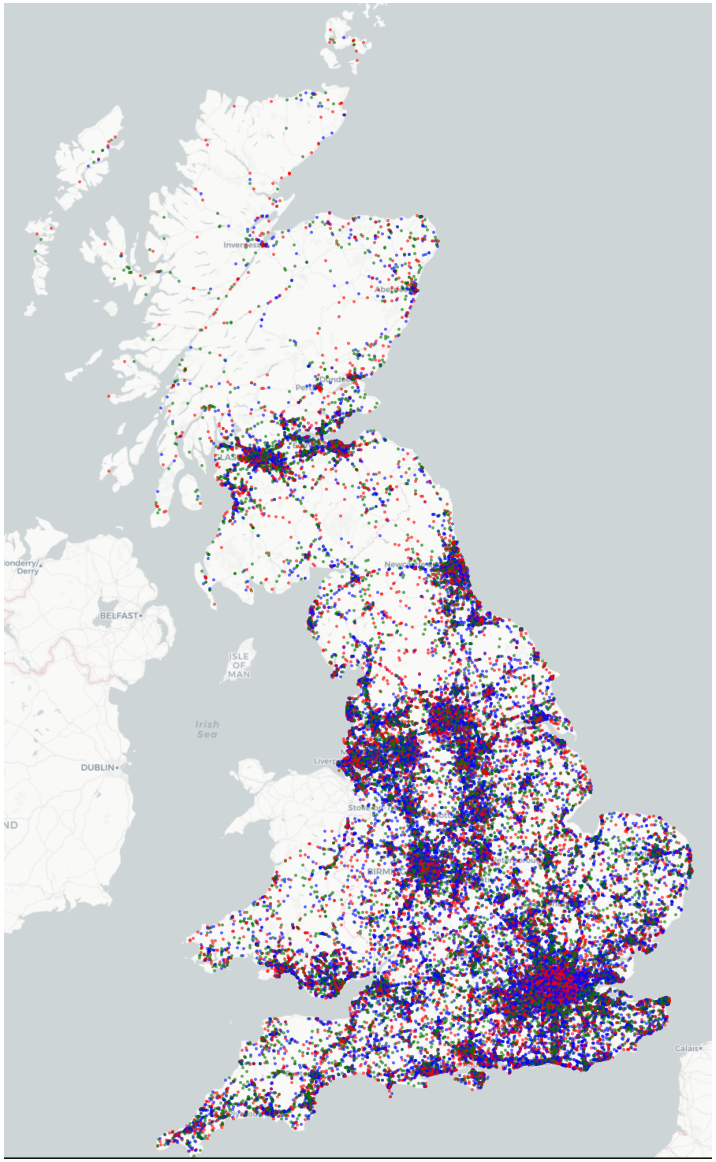


Identified clusters:

- Low speed, truck / motorcycle mixed with good conditions.
- High speed, winter based with bad conditions.

Beyond Plotting

Clustering



Identified clusters:

- Low speed, truck / motorcycle mixed with good conditions.
- High speed, winter based with bad conditions.
- Mixed type, near junction, vehicle age disparity.

Possible future work

Looking back, we would change a couple of things and add others.

To change

1. Change the age disparity $\rightarrow$ Has many problems and it dirtied our MCA.
2. Reselect important variables $\rightarrow$ i.e. Day of week (recomputed afterwards).

Possible future work

Looking back, we would change a couple of things and add others.

To change

1. Change the age disparity $\rightarrow$ Has many problems and it dirtied our MCA.
2. Reselect important variables $\rightarrow$ i.e. Day of week (recomputed afterwards).

To add

1. Time series.
2. Explore other clustering methods.
3. Explore powerful tools and frameworks for Data Viz. (i.e. Tableau)
4. Focus on vehicles, not accidents circumstances
5. Focus on casualties profiles

"If we torture the data long enough, it will confess"

- Darrell Huff

How to Lie With Statistics

1954