Proposal: Victorian Road Accidents

Abhisek Gautam (13679042), Bruno Quevedo (13500069), Ganesh Arunagiri (13662388), Priscilla Soenario (11685169), Priyanka Srinivasa (13684182), Sam Knowles (13762467)

01/09/2019

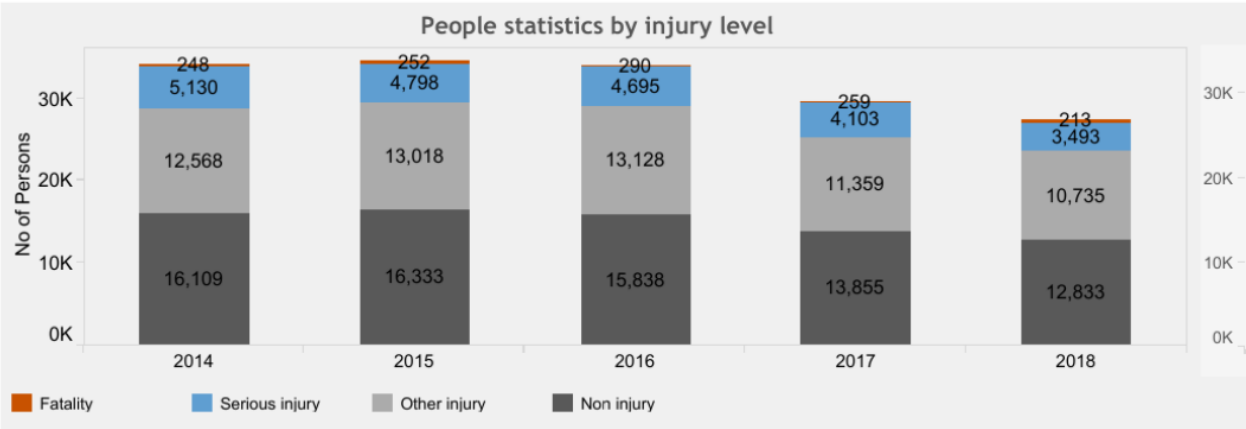
Statistical Thinking for Data Science

Background

Towards zero is the 2016-2020 road safety strategy for the state of Victoria. The strategy is a collaboration between multiple Victorian government bodies including the Transport Accident Commission, VicRoads, Victorian Police, the Department of Justice and Regulation and the Department of Health and Human Services to achieve zero road deaths in Victoria (TAC, 2019).

The short term objective of the collaboration is to reduce annual road fatalities to below 200 by 2020 and reduce serious injuries, which require hospital admission, by 15% over the 5 year period between 2016-2020 (TAC, 2019).

The Towards Zero campaign is centered upon the mantra of safer roads, safer speeds, safer vehicles and safer road users through improving infrastructure, focused education campaigns, increased monitoring and implementing innovations.



Since the implementation of the Towards Zero strategy, progress has been made to reduce both road fatalities and serious injuries (TAC, 2019). However, to ensure total deaths are below 200 by 2020, resourcing and policy need to be implemented efficiently, which this investigation aims to inform.

Weather Conditions as a Factor of Road Safety

Inclement weather conditions are widely regarded as a contributing factor to increased risk of road accidents. Conditions such as precipitation, high winds and temperature extremes may affect driver capabilities, traffic flow and vehicle performance via traction, stability and manoeuvrability (Baguley 2001; Goodwin 2002).

The relationship between weather and road accidents has been widely analysed in recent years with studies performed in multiple countries, including Belgium, France, Denmark and Finland (Karlaftis & Yannis 2010).

To accommodate changes in weather, governing authorities have provided adverse condition warnings (Roadsafety.transport.nsw.gov.au, 2019) and implemented road laws such as variable speed limits to influence driver behaviours.

Proposed Research Questions

To further investigate the reasons for high injury severity crashes and the effect of weather conditions on Victorian road accidents, the following research questions are being proposed:
1. What effect, if any, does weather has on the severity of road accidents?
2. Assuming wet conditions, how do driver age, vehicle quality and road condition impact accident severity?
3. How do seat belts and helmets impact on injury level?
4. What changes, if any, do our findings recommend to the Towards Zero campaign?

Data Sets

Of the plethora of data sets investigated, a select few remain which satisfied the criteria of a credible source, sufficient number of records, granularity at an individual level and relevant variables to address the proposed questions. The data sets identified are:

Road Accident Dataset

File Name	Description
ACCIDENT.CSV	Basic accident details, date, time, severity, injuries, etc.
ACCIDENT_LOCATION.CSV	Road accident details (road name, road type, etc.
PERSON.CSV	Person based details, age, sex etc
VEHICLE.CSV	Vehicle based data, vehicle type, make, etc.
ACCIDENT_EVENT.CSV	Sequence of events e.g.: left road, rollover, caught fire. etc.
ROAD_SURFACE_COND.CSV	Whether road was wet, dry, icy. etc.
ATMOSPHERIC_COND.CSV	Rain, wind, etc.
SUB_DCA.CSV	Detailed codes describing accident, etc.
ACCIDENT_NODE.CSV	Master location table (NB subset of accident table), latitude, longitude, etc.
ACCIDENT_CHAINAGE.CSV	Has detailed route and chainage data, etc.

Provided by VicRoads, the state road and traffic authority, the road accident data set details up to 130 variables for over 180 thousand accidents occurring between 2006 and 2019 in Victoria, Australia.

Weather Datasets

File Name	Description
STATION_X_RAINFALL.CSV	Daily rainfall, temperature and evaporation information provided by station
STATION_X.TXT	Station X location details, such as latitude and longitude

The weather data set is sourced from the Bureau of Meteorology, providing comprehensive daily statistics pertaining to weather conditions across stations in Victoria.

Supplementary Data Sets

Whilst not currently included in our core data set, we are considering a local government population data set from the ABS as a supplementary contextual layer to identify areas of anomalies with respect to frequency of crashes:

File Name	Description
ABS_REGIONAL_LGA2018_31082019194314233.csv	Annual local government population figures

Should our questions evolve or require such a layer, the annual observations will be merged with the core data set by matching on the year.

Data Merging

To address the research questions proposed, the data sets need to be merged on dimensions of time and location. With regards to time, both road crashes and weather datasets are documented on a day-basis and thus can be merged using the date variable. Merging on location is more difficult, as it will be necessary to calculate the distance between accidents and climate observatory stations, using latitude and longitude information, and then use the climate data from the closest observatory station in order to achieve better quality in the analysis.

Regression Modelling

The foremost question in our investigation, question 1, asks assuming wet road conditions, how do the driver's age, the vehicle's quality and the road's condition impact the severity of the accident. The road crash data set we have employed contains the following injury severity categorisation system, ordering injuries from most severe to least:

1. Fatal accident
2. Serious injury accident (at least one in accident sent to a hospital, possibly admitted)
3. Other injury accident
4. Non-injury accident The Ordinal Logistic Regression Model can be used to predict the behaviour of an ordinal dependent variable given one or more independent variables. Therefore, we believe this will be an appropriate and hopefully fruitful model to structure the answer of this question around.

We're cognisant, however, that a number of potential complications could arise from using such a model. For instance, the assumption that errors are homoscedastic is often violated (Williams 2008). We will look to address this and any other violation that arises with our model in the investigation.

Issues

Inevitably this project will invite a number of issues to be overcome or managed. A few complications, some of which have already been encountered, might include:

- Complexity of capturing and measuring weather variables
- Accurately assigning the most representative weather conditions to each road accident
- Multicollinearity occurring between weather conditions
- Accounting for potentially missing explanatory variables
- Interpreting our findings correctly and translating this into accurate and meaningful recommendations
- Recognising potential bias in the chosen data sets and our methodology

Unquestionably, this is merely a small sample of the numerous potential roadblocks that we're likely to incur.

Appendix

Reading Libraries

```
library(tidyverse)
library(geosphere) #to calculate geographical distances
```

Reading Data

```
accident <- read.csv("Datasets/Road Crashes/ACCIDENT.csv")
climate86018 <- read.csv("Datasets/Climate/86018.csv")
accident_node <-  read.csv("Datasets/Road Crashes/NODE.CSV")
```

Data Transformation

```
glimpse(accident)
```

```
## Observations: 181,605
## Variables: 28
## $ ACCIDENT_NO          <fct> T20060000010, T20060000018, T20060000022,...
## $ ACCIDENTDATE         <fct> 13/01/2006, 13/01/2006, 14/01/2006, 14/01...
## $ ACCIDENTTIME         <fct> 12.42.00, 19.10.00, 12.10.00, 11.49.00, 1...
## $ ACCIDENT_TYPE        <int> 1, 1, 7, 1, 1, 1, 4, 4, 1, 2, 4, 1, 2, 1,...
## $ Accident.Type.Desc   <fct> Collision with vehicle, Collision with ve...
## $ DAY_OF_WEEK          <int> 6, 6, 7, 7, 7, 7, 1, 1, 2, 2, 2, 2, 3, 3,...
## $ Day.Week.Description <fct> Friday, Friday, Saturday, Saturday, Satur...
## $ DCA_CODE             <int> 113, 113, 190, 130, 121, 116, 171, 171, 1...
## $ DCA.Description      <fct> RIGHT NEAR (INTERSECTIONS ONLY)          ...
## $ DIRECTORY            <fct> MEL, MEL, MEL, MEL, MEL, MEL, VCS, MEL, M...
## $ EDITION              <fct> 40, 40, 40, 40, 40, 40, 9, 40, 40, 40, 40...
## $ PAGE                 <fct> 91A, 91, 169, 88, 169, 146, 704, 103, 90,...
## $ GRID_REFERENCE_X     <fct> G, H, C, J, G, G, A, B, E, D, E, F, D, D,...
## $ GRID_REFERENCE_Y     <int> 7, 8, 11, 8, 5, 2, 4, 5, 11, 3, 9, 5, 11,...
## $ LIGHT_CONDITION      <int> 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 1, 3, 1,...
## $ Light.Condition.Desc <fct> Day, Day, Day, Day, Day, Day, Day, Dark S...
## $ NODE_ID              <int> 43078, 29720, 203074, 55462, 202988, 2774...
## $ NO_OF_VEHICLES       <int> 3, 2, 1, 2, 2, 2, 1, 1, 3, 1, 1, 3, 1, 2,...
```

```
## $ NO_PERSONS           <int> 6, 4, 2, 2, 3, 2, 1, 1, 5, 2, 2, 4, 2, 2,...
## $ NO_PERSONS_INJ_2     <int> 0, 0, 1, 1, 0, 1, 1, 1, 2, 0, 1, 1, 1, 1,...
## $ NO_PERSONS_INJ_3     <int> 1, 1, 0, 0, 3, 0, 0, 0, 2, 1, 0, 2, 0, 0,...
## $ NO_PERSONS_KILLED    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ NO_PERSONS_NOT_INJ   <int> 5, 3, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1,...
## $ POLICE_ATTEND        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1,...
## $ ROAD_GEOMETRY        <int> 1, 2, 5, 2, 5, 1, 5, 5, 2, 5, 5, 1, 5, 5,...
## $ Road.Geometry.Desc   <fct> Cross intersection, T intersection, Not a...
## $ SEVERITY             <int> 3, 3, 2, 2, 3, 2, 2, 2, 2, 3, 3, 2, 2, 3,...
## $ SPEED_ZONE           <int> 60, 70, 100, 80, 50, 100, 100, 70, 80, 60...
```

```
glimpse(climate86018)
```

```
## Observations: 4,988
## Variables: 11
## $ station           <int> 86018, 86018, 86018, 86018, 86018, 86018, 86...
## $ YYYY.MM.DD        <fct> 2006-01-01, 2006-01-02, 2006-01-03, 2006-01-...
## $ daily_rain        <dbl> 0.0, 9.6, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,...
## $ daily_rain_source <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ max_temp          <dbl> 28.0, 22.0, 21.0, 20.0, 24.0, 24.0, 26.0, 29...
## $ max_temp_source   <int> 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, ...
## $ min_temp          <dbl> 21.0, 13.5, 14.5, 13.5, 9.5, 14.5, 14.5, 13....
## $ min_temp_source   <int> 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, ...
## $ evap_pan          <dbl> 4.8, 5.8, 6.4, 4.2, 6.2, 6.6, 6.4, 6.8, 3.0,...
## $ evap_pan_source   <int> 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, ...
## $ metadata          <fct> "name=CAULFIELD (RACECOURSE)                ...
```

```
#Adding station latitude and longitude to all rows (to calculate distance between the accident and clim
climate86018$station_lat <- rep(-37.8795,nrow(climate86018))
climate86018$station_long <- rep(145.0368,nrow(climate86018))

#Converting fct to date
accident$ACCIDENTDATE <- as.Date(accident$ACCIDENTDATE, "%d/%m/%Y")
climate86018$YYYY.MM.DD  <- as.Date (climate86018$YYYY.MM.DD, "%Y-%m-%d")

#Rename date columns (to be used as primary key between accidents and climate data)
accident <- accident %>% rename (DATE = ACCIDENTDATE)
climate86018 <- climate86018 %>% rename (DATE = YYYY.MM.DD)
```

Merging Data

```
#Creating mergedFiles object, which contains accident, accident location and climate data
mergedFiles <- merge(accident,accident_node,by="ACCIDENT_NO")
mergedFiles <- merge(mergedFiles, climate86018,by="DATE")
```

Calculating Distance

```
#Adding dist_station_crash column with distance between accident location and climate observatory
mergedFiles <- mergedFiles %>%
    mutate(dist_station_crash = distHaversine(cbind(station_long, station_lat), cbind(Long, Lat)))

glimpse(mergedFiles)
```

```
## Observations: 183,580
## Variables: 52
## $ DATE                 <date> 2006-01-01, 2006-01-01, 2006-01-01, 2006...
## $ ACCIDENT_NO          <fct> T20060017796, T20060001270, T20060005504,...
## $ ACCIDENTTIME         <fct> 17.00.00, 01.15.00, 16.30.00, 16.09.00, 0...
## $ ACCIDENT_TYPE        <int> 1, 2, 1, 1, 2, 1, 4, 1, 5, 1, 4, 9, 1, 1,...
## $ Accident.Type.Desc   <fct> Collision with vehicle, Struck Pedestrian...
## $ DAY_OF_WEEK          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ Day.Week.Description <fct> Sunday, Sunday, Sunday, Sunday, Sunday, S...
## $ DCA_CODE             <int> 132, 106, 130, 137, 103, 160, 181, 121, 1...
## $ DCA.Description      <fct> RIGHT REAR.                          , VEH...
## $ DIRECTORY            <fct> MEL, MEL, MEL, MEL, MEL, MEL, MEL, MEL, M...
## $ EDITION              <fct> 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 4...
## $ PAGE                 <fct> 91, 100A, 432, 17, 382, 41, 2M, 1B, 243, ...
## $ GRID_REFERENCE_X     <fct> A, F, A, H, A, A, F, N, A, A, A, E, J, F,...
## $ GRID_REFERENCE_Y     <int> 2, 6, 12, 6, 6, 2, 2, 3, 2, 8, 4, 8, 2, 1...
## $ LIGHT_CONDITION      <int> 1, 3, 1, 1, 3, 3, 1, 1, 1, 3, 1, 1, 1, 1,...
## $ Light.Condition.Desc <fct> Day, Dark Street lights on, Day, Day, Dar...
## $ NODE_ID.x            <int> 43133, 43466, 202061, 25970, 202189, 2020...
## $ NO_OF_VEHICLES       <int> 2, 2, 5, 2, 1, 2, 1, 2, 1, 2, 1, 1, 2, 2,...
## $ NO_PERSONS           <int> 3, 3, 9, 4, 2, 2, 1, 5, 1, 5, 2, 3, 6, 3,...
## $ NO_PERSONS_INJ_2     <int> 2, 0, 1, 0, 0, 1, 0, 2, 0, 1, 1, 1, 0, 0,...
## $ NO_PERSONS_INJ_3     <int> 0, 1, 0, 1, 1, 0, 1, 2, 1, 0, 1, 0, 1, 1,...
## $ NO_PERSONS_KILLED    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ NO_PERSONS_NOT_INJ   <int> 1, 2, 8, 3, 1, 1, 0, 1, 0, 4, 0, 2, 5, 2,...
## $ POLICE_ATTEND        <int> 2, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2,...
## $ ROAD_GEOMETRY        <int> 2, 2, 5, 5, 5, 5, 5, 1, 2, 2, 5, 2, 1, 2,...
## $ Road.Geometry.Desc   <fct> T intersection, T intersection, Not at in...
## $ SEVERITY             <int> 2, 3, 2, 3, 3, 3, 3, 2, 3, 2, 2, 2, 3, 3,...
## $ SPEED_ZONE           <int> 60, 60, 70, 60, 60, 50, 60, 50, 100, 100,...
## $ NODE_ID.y            <int> 43133, 43466, 202061, 25970, 202189, 2020...
## $ NODE_TYPE            <fct> I, I, N, N, N, N, N, I, I, I, N, I, I, I,...
## $ AMG_X                <dbl> 2521356, 2511353, 2443593, 2496707, 24747...
## $ AMG_Y                <dbl> 2392466, 2373238, 2378882, 2420305, 24350...
## $ LGA_NAME             <fct> CASEY, FRANKSTON, GEELONG, MORELAND, HUME...
## $ Lga.Name.All         <fct> CASEY, FRANKSTON, GEELONG, MORELAND, HUME...
## $ Region.Name          <fct> METROPOLITAN SOUTH EAST REGION, METROPOLI...
## $ Deg.Urban.Name       <fct> MELB_URBAN, MELB_URBAN, LARGE_PROVINCIAL_...
## $ Lat                  <dbl> -37.96874, -38.14215, -38.08966, -37.7181...
## $ Long                 <dbl> 145.2431, 145.1295, 144.3570, 144.9626, 1...
## $ Postcode.No          <int> 3802, 3199, 3214, 3046, 3429, 3019, 3142,...
## $ station              <int> 86018, 86018, 86018, 86018, 86018, 86018,...
## $ daily_rain           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ daily_rain_source    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ max_temp             <dbl> 28, 28, 28, 28, 28, 28, 28, 28, 28, 28, 2...
## $ max_temp_source      <int> 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 2...
## $ min_temp             <dbl> 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 2...
## $ min_temp_source      <int> 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 2...
## $ evap_pan             <dbl> 4.8, 4.8, 4.8, 4.8, 4.8, 4.8, 4.8, 4.8, 4...
## $ evap_pan_source      <int> 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 2...
## $ metadata             <fct> name=CAULFIELD (RACECOURSE)                ...
## $ station_lat          <dbl> -37.8795, -37.8795, -37.8795, -37.8795, -...
## $ station_long         <dbl> 145.0368, 145.0368, 145.0368, 145.0368, 1...
## $ dist_station_crash   <dbl> 20656.739, 30347.481, 64067.465, 19107.91...
```

References

Al Hassan, Y. and Barker, D. (1999). The impact of unseasonable or extreme weather on traffic activity within Lothian region, Scotland. Journal of Transport Geography, 7(3), pp.209-213.

Baguley, C. (2019). The importance of a road accident data system and its utilisation.. [online] Pdfs.semanticscholar.org. Available at: https://pdfs.semanticscholar.org/2d8e/c35ac2bd46cc41934d84e1ee2024da6dbec1.pdf [Accessed 1 Sep. 2019].

Bergel-Hayat, R., Debbarh, M., Antoniou, C. and Yannis, G. (2013). Explaining the road accident risk: Weather effects. Accident Analysis & Prevention, 60, pp.456-465.

Goodwin, L. (2019). Weather Impacts on Arterial Traffic Flow. [online] Citeseerx.ist.psu.edu. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.587.9563&rep=rep1&type=pdf [Accessed 1 Sep. 2019].

Ops.fhwa.dot.gov. (2019). How Do Weather Events Impact Roads? - FHWA Road Weather Management. [online] Available at: https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm [Accessed 1 Sep. 2019].

Roadsafety.transport.nsw.gov.au. (2019). Bad weather - Drivers - Staying safe - NSW Centre for Road Safety. [online] Available at: https://roadsafety.transport.nsw.gov.au/stayingsafe/drivers/badweather/index.html [Accessed 1 Sep. 2019].

TAC. (2019). Home - Towards Zero | Road Safety Victoria. [online] Available at: https://www.towardszero.vic.gov.au/ [Accessed 1 Sep. 2019].

Williams, R. (2008). Ordinal regression models: Problems, solutions, and problems with the solutions. Available at: https://www.stata.com/meeting/germany08/GSUG2008.pdf [Accessed 1 Sep. 2019].

Yannis, G., Karlaftis, G. (2010). Weather Effects on Daily Traffic Accidents and Fatalities: A Time Series Count Data Approach . [online]. Available at: https://www.nrso.ntua.gr/geyannis/wp-content/uploads/geyannis-pc102.pdf Accessed 1 Sep. 2019].