**McNulty Project Plan**

Kelly Jones

**Topic:** 2015 Traffic Fatalities

**Data Source:** National Highway Traffic Safety Administration (NHTSA): Fatality Analysis Reporting System (FARS), accessed via Kaggle

**Links:** [Kaggle Page](https://www.kaggle.com/nhtsa/2015-traffic-fatalities) | [Download Data](https://www.kaggle.com/nhtsa/2015-traffic-fatalities/downloads/2015-traffic-fatalities.zip/3) | [Summary Table](https://www-fars.nhtsa.dot.gov/Main/index.aspx) | [FARS Encyclopedia](https://www-fars.nhtsa.dot.gov/Help/Terms.aspx) | [FARS Acronyms](https://www-fars.nhtsa.dot.gov/Help/acronyms.aspx)

**Tables To Use:**

|  |  |  |
| --- | --- | --- |
| **WILL USE** | | |
| person.csv | ST\_CASE | Unique crash identifier by year |
|  | STATE | State where crash occurred |
|  | VEH\_NO | Consecutive for vehs in crash. Pair with ST\_CASE |
|  | PER\_NO | Consecutive for ppl in crash. Pair with ST\_CASE and VEH\_NO |
|  | AGE | In years, up to 120 |
|  | SEX | 1 male, 2 female, 8/9 |
|  | PER\_TYP | 1 Driver, 2 Passenger – remove all others |
|  | INJ\_SEV | Injury severity – 0 to 4 (fatal) increase in severity, also 5 6 8 9 |
|  | SEAT\_POS | Bunch of categories – will need to connect to vehicle damage area in order to be useful |
|  | REST\_USE | Restraint use – seatbelts, helmet (for motorcycle), several others – look at descriptives. May just do 07 (“none used”) vs all others. Appendix helpful. |
|  | REST\_MIS | Misuse of restraint system or helmet. (Q: does this include not using it?) |
|  | AIR\_BAG | Several deployment options |
|  | EJECTION | 0 not, 1 totally, 2 partially, 3 unknown degree, 7+ |
|  | EXTRICAT | 0 not extricated/NA, 1 extricated, 9 unknown |
|  | DRINKING | Whether alcohol was involved for this person, per judgment of law enforcement |
|  | DRUGS | Drugs involved? 0 no, 1 yes, 8/9 ? |
|  | HOSPITAL | Transported to first treatment facility (0 none, 1 EMS air, 2 law enforcement, etc.) – probably shouldn’t use, too closely tied to mortality |
|  | DOA | 0 na, 7 at scene, 8 en route, 9 ? |
| **vehicle.csv** | ST\_CASE | Unique crash identifier by year |
|  | STATE | State where crash occurred |
|  | VEH\_NO | Consecutive for vehs in crash. Pair with ST\_CASE |
|  | NUMOCCS | Number of occupants in the vehicle |
|  | HIT\_RUN | Diff kinds of hit & run, including perpetrator and victim – may not be useful |
|  | REG\_STAT | Vehicle’s state of registration |
|  | BODY\_TYP | Body type – v. granular, almost 100 options, could be useful to group them |
|  | MOD\_YEAR | Vehicle model year |
|  | VIN | Vehicle’s VIN |
|  | TOW\_VEH | Identifies any attached trailing units |
|  | J\_KNIFE | Whether this vehicle jackknifed at any point |
|  | GVWR | Gross Vehicle Weight Rating, if applicable – may be useful with BODY\_TYP |
|  | HAZ\_INV | Hazardous material involvement (whether carrying hazmat) – 1 NO, 2 YES |
|  | BUS\_USE | Type of bus service – 0 if not used as bus |
|  | TRAV\_SP | Speed vehicle was traveling prior to crash (estimate by investigator after the fact) – may want to make categorical, since up to 151 is scalar |
|  | UNDERIDE | Any underride or override (0 if not) – often fatal |
|  | ROLLOVER | 0 none, 1 2 or 9 for diff types |
|  | IMPACT1 | Initial contact point, lots of ways measured |
|  | DEFORMED | Extent of damage: 0 none, 2 minor, 4 functional, 6 disabling, 8/9 unreported/unknown |
|  | TOWED | Vehicle removal – 1 driven away, 2-3 towed, 4 abandoned, 5 not towed, 8/9 unrptd/unknwn |
|  | M\_HARM | “Most Harmful Event” (at vehicle level, unlike First HE) – long, specific list |
|  | VEH\_SC1, VEH\_SC2 | Vehicle-level related crash factors, e.g, tires and wheels, headlights, horns |
|  | FIRE\_EXP | Binary – if fire related to the crash in this vehicle |
|  | DEATHS | Fatalities in this vehicle ☹ |
|  | DR\_DRINK | Binary – drunk driver for this vehicle (contrast to accident rpt of # total drunk drivers) |
|  | L\_STATE | State of issue for driver’s license |
|  | DR\_ZIP | Driver’s zip code |
|  | L\_TYPE | Driver’s license type – not, full, intermed, permit |
|  | L\_STATUS | Driver’s license status – none, valid, suspended, revoked, expired, permit |
|  | L\_ENDORS | Compliance with CDL endorsements |
|  | L\_COMPL | Type of license possessed (or not) by driver for the class of vehicle driven during crash – good summary field for above details? |
|  | DR\_HGT | Driver’s height (inches) |
|  | DR\_WGT | Driver’s weight (pounds) |
|  | PREV\_ACC | # previous crashes by this driver in last 5 yrs |
|  | PREV\_SUS | # previous license suspensions/revocations for this driver in last 5 yrs |
|  | PREV\_DWI | # prev DWI convictions in last 5 yrs |
|  | PREV\_SPD | # prev speeding convictions in last 5 yrs |
|  | PREV\_OTH | # prev other moving violations in last 5 yrs |
|  | SPEEDREL | Accident was speeding related, per officers – 2,3,4,5, mean yes; 0 no; 8/9 unknown |
|  | DR\_SF1, DR\_SF2, DR\_SF3, DR\_SF4 | Driver-level related factors expressed by investigating officer; looooong, granular list |
|  | VNUM\_LAN | Number of travel lanes |
|  | VSPD\_LIM | Speed limit in 5 mph increments up to 80 |
|  | VPROFILE | Level, hillcrest, uphill, etc |
|  | VSURCOND | Roadway surface condition – dry, wet, snow, sand, water, oil, mud/dirt/gravel, etc |
| accident.csv | ST\_CASE | Unique identifier by year |
|  | STATE |  |
|  | PEDS | Number of non-occupants – can make this binary as indicator of whether there were people involved who were not in a MV (expect higher survival in that case – fatality was likely the unprotected pedestrian/bicyclist) |
|  | COUNTY | County code via GSA geographical codes |
|  | CITY | City code via GSA geographical codes |
|  | MONTH |  |
|  | DAY |  |
|  | DAY\_WEEK |  |
|  | YEAR |  |
|  | HOUR |  |
|  | MINUTE |  |
|  | LATITUDE |  |
|  | LONGITUD |  |
|  | HARM\_EV | “First Harmful Event” (applies to crash – “Most Harmful Event” is per vehicle) |
|  | MAN\_COLL | Manner of collision – orientation of vehicles if involved in First Harmful Event, e.g., rear-end, head-on, sideswipe (0 if not a collision) |
|  | LGT\_COND | Light condition, e.g. daylight, dusk, dark but lighted |
|  | WEATHER, WEATHER 1, WEATHER2 | Identifies up to 2 weather values; “WEATHER” is derived from the other 2 coded elements |
|  | SCH\_BUS |  |
|  | NOT\_HOUR, NOT\_MIN | Hour & min of EMS notification |
|  | ARR\_HOUR, ARR\_MIN | Hour & min of EMS arrival at crash scene |
|  | CF1, CF2, CF3 |  |
|  | DRUNK\_DR | # of drunk drivers involved – virtually never >2 |
|  | FATALS | Number of fatally injured persons in crash ☹ |
| **MAY USE** | | |
| Cevent |  | Every event chronologically |
| Vevent |  | Every event in order by vehicle |
| Vsoe |  | Simplified vevent |
| Damage |  | One record per damaged area |
| Distract |  | Driver distractions – one record per distraction |
| Drimpair |  | Driver impairments – one record per impairment |
| Factor |  | Vehicle circumstances which may have contributed to crash – one record per circumstance |
| Violatn |  | Violations charged to drivers – one record per violation |
| Nmcrash |  | Non-motorist contributing circumstances (e.g., pedestrians, bicyclists) – one record per action |
| Vindecode |  | More detailed info on vehicle based on VIN – one record per vehicle |

\*\*\* Do I need vehicle information from Parkwork to account for those drivers/passengers?

**Join Method:**

* ~~Start with person.csv and keep only drivers and passengers (PER\_TYPE = 1, 2)~~
* ~~Use ST\_CASE to merge on accident.csv data~~
* ~~Use ST\_CASE and VEH\_NO to merge on vehicle.csv data~~
* ~~Should have one row per person~~

**Data Quality:**

* **~~Maybe Add:~~** ~~Related to junction, RELJCT1 (binary) and RELJCT2 (type of junction). Route signing var ROUTE (interstate, US highway, etc.)~~
  + **~~Decision:~~** ~~Yes – added latter two~~
* **~~Dropping Nulls:~~** ~~May want to drop DRINKING and DRUGS because half of dataset is not reported or unknown – as interim step, imputed those values as 0.2 (think it’s unlikely – belligerent inebriation would likely have been noted by officers)~~
  + **~~Decision:~~** ~~Dropped~~
* **~~Drunk Driver:~~** ~~Create dummy? Current DRUNK\_DR is total number of drunk drivers in accident, which is always 3 or less – but important thing is probably diff from 0 to 1. Also have DR\_DRINK which measures whether driver of that vehicle was drunk~~
  + **~~Decision:~~** ~~Created binary~~
* **~~Bucketing:~~** ~~Features like UNDERIDE and BODY\_TYP could benefit from bucketing similar types/outcomes together – as second feature or just a recode? How to do hierarchy? Also L\_STATUS~~
  + **~~Decision:~~** ~~Created binaries for several features~~
* **Maybe Drop:**
  + IMPACT1: Weird scale – only useful if connected to where person is sitting
  + DEFORMED: Lots of nulls, may not tell us much that’s actionable
  + PREV\_SUS: Values are suspicious. How can someone’s license be suspended/revoked 75 times in 5 years? (Consider other PREV\_ driver factors – same 2485 cases missing for most)
  + DR\_SF1 (to 4): Driver features weirdly split across 4, plus same 3070 cases missing
* **~~Speed:~~** ~~What to do with 40,000 missing travel speeds? May need to use “speed a factor in the crash” feature instead (SPEEDREL)~~
  + **~~Decision:~~** ~~Drop it. Use other feature.~~

**Presentation:**

* Local tie-in: zero plan
  + WA in middle (?) of pack nationally in 2015 – confirm exact rank
  + Model: National data (50K rows instead of 316 of full data) – can help us identify the most important factors in fatal collisions
  + Some comparisons btwn WA and national avg, e.g.:  
    *Format: Maybe double-stacked horizontal 100% bars with filter to toggle? With text of best 3 states and worst 3 states in that category? OR vertical bars with WA in orange, rest in blue, ranked best to worst? (harder for dist)*
    - % accidents with drunk drivers
    - % accidents involving buses
    - Distribution of weather
    - Fatalities per 100,000 residents – sep viz, sep data source (pre-agg)
    - % out of state drivers – implications for signage/consistency
    - % school bus accidents
    - % with unsecured loads causing accidents (do I have that?)
* Re: data quality, look at reporting requirements for fatality accidents vs. non
* Map of event locations? Drivers far from home?

**Extension Ideas for Data:**

* **Response Time:** Use notification hour/min vs. arrival hour/min to calculate response time and keep that in the model? (Had dropped all 4 fields: NOT\_HOUR, NOT\_MIN, ARR\_HOUR, ARR\_MIN)
* **Car Out of State:** Compare state of vehicle registration to state of accident and flag it. (Had dropped REG\_STAT from model.) Could also do L\_STATE (driver’s license state), also dropped. If I could get center point (in lat/long) of each zip code, could do distance from license zip code to accident spot – how far from home were people?
* **Over Speed Limit:** Compare TRAV\_SP to VSPD\_LIM

**Miscellaneous:**

* “…half of all fatal car-versus-truck accidents are classified as underride truck accidents.” ([Source](https://www.chicagolawyer.com/truck-accident-lawyer/underride-and-override-accidents/))