ACCIDENT PROJECT

W205 Final Project

Angela Gunn

December 2015

TABLE OF CONTENTS

1 PROBLEM DESCRIPTION (ORIGINAL PROPOSAL)

I propose to examine the attributes of vehicular accidents in the US where distracted driving is involved. The National Highway Traffic and Safety Administration (NHTSA - http://www.nhtsa.gov/NASS) has a program called the National Automotive Sampling System (NASS) which provides a reusable resource to conduct data collection over the past 10 years. Within this program there are multiple datasets including the General Estimates System (GES) which I will use in this study. Additional databases like the Fatality Analysis Reporting System (FARS) and the Special Crash Investigations (SCI) are also very interesting, but will not be included (the scale of this could easily get out of hand for a single-person project).

While the GES includes all accidents, I will be looking only at those that involve distracted driving. I will attempt to include as many attributes that I deem interesting in my extraction of the data, but will concentrate on data that will assist in answering the following questions:

- With distracted driving, how often (and severe) are the injuries to the distracted driver
- With distracted driving, what events are most common (first harmful event, most harmful
- What locations are distracted driving accidents most common (highway, intersection,
- Which distractions (phone, eating, smoking, adjusting controls, etc.) cause the most

Challenges of this project beyond my learning curve of implementing and working with storing and retrieving data include the following highlights:

- The number of individual files to work with
- The files are in sas7dbat format, so must be converted (Python has a library that does this
- The attributes are all stored in codes; work will be required to make the codes meaningful
- The codes have changed over years, and table fields have been discontinued or merged into
- The modelling of meaningful data, particularly where data is presented in multiples (people,

The deliverable will be a method of querying the resulting distracted driving data. I would like to be able make an interface that will allow a user to query the data to retrieve a CSV file that is appropriate for use in Tableau to visualize the resulting information.

Data Source: ftp://ftp.nhtsa.dot.gov/GES/
User Manual: http://www-nrd.nhtsa.dot.gov/Pubs/812091.pdf

2 DATA SOURCE

Data Source: ftp://ftp.nhtsa.dot.gov/GES/

User Manual: http://www-nrd.nhtsa.dot.gov/Pubs/812091.pdf

The National Automotive Sampling System (NASS) General Estimates System (GES) was started in 1988. A survey of all police reported traffic accidents are reported on every year to help identify highway safety problem areas, provide consumer information initiatives and form a basis for costs and benefits of highway safety initiatives.

The survey takes a representative sample of the more than five million yearly police reported accidents. The fact these are police reported accidents is important, as there are many accidents that are not reported to police, likely due to minor damage and minimal personal injury.

This project looks at data starting in 2003, the first year distraction data was recorded. Only fields determined to potentially be of interest in the investigation of the impact of distracted driving are included. Within the variables included, data was either rolled up into a summary level as appropriate based on the identified needs of the project. The Meta Data section lists all fields and the detail of their meanings.

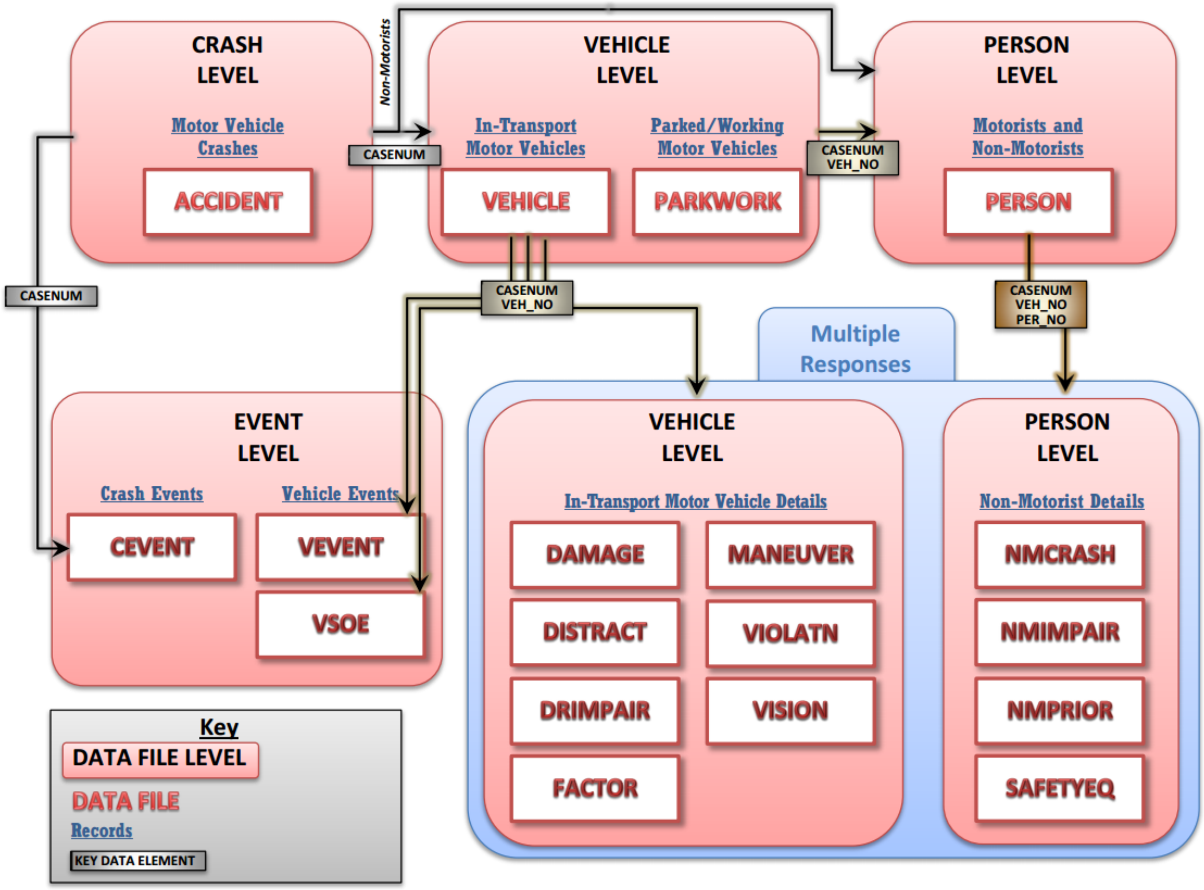
The NASS GES includes imputed data for fields where the data is missing or unknown. This project does not look at any of the imputed fields, opting to only use the original data fields. It is acknowledged there are an abundance of “Unknown” fields in all tables. However, the imputation methods used are not consistent over the years, and therefore analytical results would not be consistent.

The NASS GES files are available as SAS data files. Starting in 2009, the files are also available in text format.

Most data is stored as integers; use of the GES Analytical User’s Manual is necessary to interpret values in the tables. It must be noted that value meanings may change over the years – an 8 one year may have a value and the next year be indicate the value was not reported.

The full data set per year contains 18 different files. The below image, from the GES Analytical User’s Manual, depicts the relationship of the files.

GES Analytical User's Manual NASS GES SAS Data Files



GES Analytical User's Manual NASS GES SAS Data Files • 22

For this project, only the Accident, Vehicle, Person and Distract files are considered.

3 ARCHITECTURE

3.1 FILE STORAGE

Many of the files were available through the FTP site as text files, but the years 2002 – 2008, the files were only available in .sas7bat format. To address this, all files from the FTP site were retrieved. The .sas7bat files were converted to csv using the library SparkSQL SAS to perform a one-time conversion using the SASExport program.

(https://github.com/saurfang/spark-sas7bdat)

The raw data files are stored in S3, specifically at
https://s3-us-west-2.amazonaws.com/accident-project

These files are retrieved and stored in the project directory before being processed and saved to HDFS. See the appendices for details.

3.2 HIVE DATABASE

The data ultimately ends up in a Hive database.

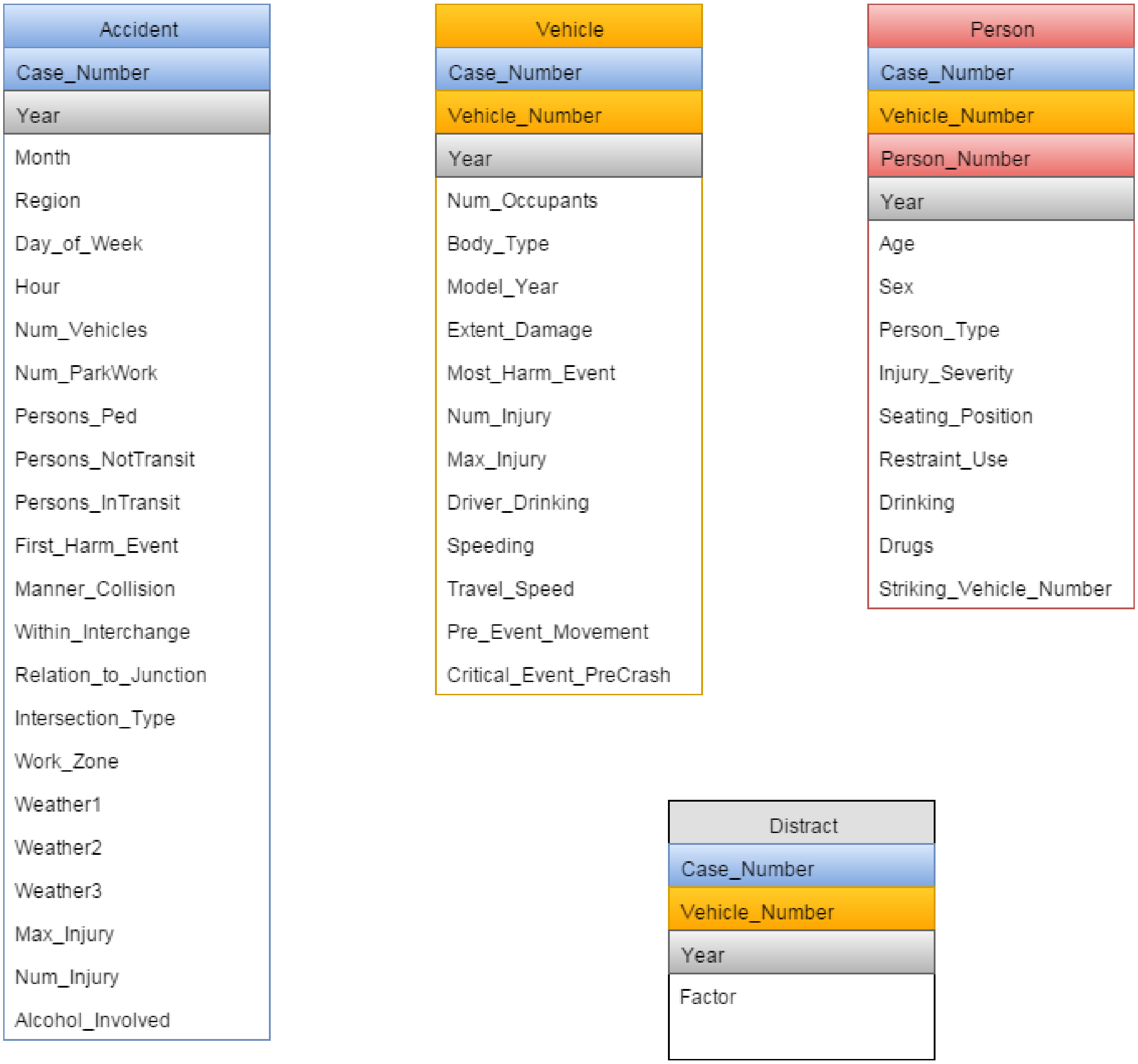
The database is named **accident_project.**

When the setup script is executed, the final database will have 52 tables. accident, distract, person and vehicle are the ones used in this project. The other tables are there for the interest of the user only.

Table	Description
accident	The table with accident data as used by this project
accident_20xx	The base table for accident data for the year 20xx. This table will contain all data from the raw data file.
distract	The table with distraction data as used by this project
distract_20xx	The base table for distraction data for the year 20xx. This table will contain all data from the raw data file.
person	The table with person data as used by this project
person_20xx	The base table for person data for the year 20xx. This table will contain all data from the raw data file.
vehicle	The table with vehicle data as used by this project
vehicle_20xx	The base table for vehicle data for the year 20xx. This table will contain all data from the raw data file.

3.3 DATA SCHEMA

Database Name: Accident Project



Accident
Case_Number
Year
Month
Region
Day_of_Week
Hour
Num_Vehicles
Num_ParkWork
Persons_Ped
Persons_NotTransit
Persons_InTransit
First_Harm_Event
Manner_Collision
Within_Interchange
Relation_to_Junction
Intersection_Type
Work_Zone
Weather1
Weather2
Weather3
Max_Injury
Num_Injury
Alcohol_Involved

Vehicle
Case_Number
Vehicle_Number
Year
Num_Occupants
Body_Type
Model_Year
Extent_Damage
Most_Harm_Event
Num_Injury
Max_Injury
Driver_Drinking
Speeding
Travel_Speed
Pre_Event_Movement
Critical_Event_PreCrash

Person
Case_Number
Vehicle_Number
Person_Number
Year
Age
Sex
Person_Type
Injury_Severity
Seating_Position
Restraint_Use
Drinking
Drugs
Striking_Vehicle_Number

Distract
Case_Number
Vehicle_Number
Year
Factor

4 DATA RETRIEVAL AND STORAGE

4.1 PROCESS

When all project files have been cloned from the github repository, two commands are used to retrieve and store the necessary data:

```
$ chmod 755 start_up.sh
$ ./start_up.sh
```

The first line adds permissions to run the script. The second line actually executes the script. It will take about 30 minutes to complete all required processing.

It is assumed the platform has access to HDFS, HIVE and that the python libraries pandas, numpy and matplotlib are installed.

4.1.1 LOADING AND MODELLING

For each year (2002 – 2013), a script load_data_20xx.sh is executed. This script follows the following steps:

1) Retrieve the files from S3 (*TABLENAME*.TXT)
2) For each file, remove the header information (20xx_*tablename*.txt)
3) Remove the FILENAME.TXT files
4) Remove/Create HDFS file directories /user/w205/accident_project/2006
5) Move the text files 20xx_*tablename*.txt to the appropriate directory.

Note, the Parked data files are retrieved and processed but are ultimately not used in the project.

4.1.2 BASE TABLES

The next step is to create the base tables. There is one script 20xx_hive_base_*tablename*.sql for each base table. These scripts create tables in the HIVE database accident_project (which is created if not already existing) with the name *TABLENAME*_20xx with the fields from the raw data, using the original header information. These are external tables with the data stored in the related HDFS directory created in the Loading and Modelling step.

4.1.3 TRANSFORMING TO TARGET TABLES

The final step is to transform the base tables to the tables required for the project ("target tables"). Each year and table has its own UDF that takes the data from the base table and processes it into meaningful fields in the target tables. There are 12x4 = 48 distinct scripts, named 20xx*Table*.py

Once these tables are loaded into HDFS (directory /user/w205/accident_proejct/pyscripts/), the SQL scripts are executed to create and populate the target tables. There is one script, named hive_target_*tablename*.sql for each table. There is an insert statement for each year for each table in which the data for that specific year/table is inserted into a partition based on year. It should be noted that as a result of using partitions based on year, there is both a "file_year" and "year" field in the table. These fields should always be the same; the year field should be used for where clauses and grouping in queries as this will be indexed based on the partitions. The file_year field is kept in the table for consistency and developer paranoia.

4.1.3.1 UDF SCRIPTS

The UDF scripts (named 20xx*TableName*.py) take each table row from the select query from the insert statement and returns a new table row (in the form of a string). Where the value from the query needs to be modified to its string representation, data dictionary elements are used. In the case where multiple values are mapped to the same string, a separate function is written for readability and maintainability.

The value "NA" is used when the mappings defined do not match the value retrieved from the query.

4.1.4 HIVE VIEWS

The final step is creating views that will assist in distracted driving specific queries.

4.1.4.1 DISTRACT_ACCIDENTS

This is a view of the case number for all accidents involving distractions.

Fields: case_number
Tables: distract
Filter: distract.factor not in ('Not Distracted', 'Unknown')

4.1.4.2 DISTRACT_DRIVERS

This is a view of all persons who are distracted drivers

Fields: person.*
Tables: person, distract, distract_accidents
Filter: person.person_type = ‘Driver’) and dis distract.factor not in (‘Not Distracted’, ‘Unknown’)

4.1.4.3 NOT_DISTRACT_DRIVERS

This is a view of all persons who are not distracted drivers. Note that this will include persons who are drivers but are not distracted.’

Fields: person.*
Tables: person, distract, distract_accidents
Filters: if driver, distract.factor in (‘Not Distracted’); all persons who are not drivers

4.1.4.4 DISTRACT_VEHICLES

This is a view of all vehicles with distracted drivers.

Fields: vehicle.*
Tables: vehicle, distract, distract_accidents
Filters: distract.factor not in (‘Not Distracted’, ‘Unknown’)

4.1.4.5 NOT_DISTRACT_VEHICLES

This is a view of all vehicles where the drivers are not distracted.

Fields: vehicle.*
Tables: vehicle, distract, distract_accidents
Filters: distract.factor in (‘Not Distracted’)

4.2 QUERY SCRIPTS

Six scripts are available for creating CSV representations of the data, and one script is available for creating visualizations of specific fields.

4.2.1 TABLE_CREATE_CSV.PY

This script returns a CSV of all fields in one or two tables, as specified in the call parameters.

Usage:
$ SPARK-SUBMIT table_create_csv.py tablename1 [tablename2]

Valid table names:
Accident Person Vehicle Distract

CSV File:
tablename1.csv or tablename1_tablename2.csv

This script currently only supports four tables, and is case sensitive (bad design, but works for now).

A maximum of two tables is imposed for practicality. If an accident has 2 vehicles involved, and each vehicle has 2 people, and both drivers have 2 distractions, 6 rows would be returned – two for each driver and one for each additional passenger. This was found to be confusing and not as useful. The implementation was also more complicated.

All fields in the specified tables are returned; if two tables are used, case_number, year, vehicle_number and person number fields are included only once.

4.2.2 HISTOGRAM.PY

This script displays on screen the counts for a specified table field and produces a PNG file with a graphical view of the resulting histogram.

Usage:
$ SPARK-SUBMIT table_create_csv.py tablename field [year]

CSV File:
tablename_field.png or tablename_field_year.png

Valid table names:
Accident Person Vehicle Distract

This script currently only supports four tables, and is case sensitive (bad design, but works for now). Field name is not validated prior to attempting the select query call.

4.2.3 DISTRACT_ACCIDENTS_TIMECSV.PY

This script returns a CSV with accidents involving distractions over time.

Usage:
$ SPARK-SUBMIT distract_accidents_timeCSV.py

Fields returned:

Field Name	Description
year	The record year (as a date 01-01-yyyy)
total_count	The total number of accidents of all types in the specified year
distract_count	The total number of accidents that involved distracted driving in the specified year
percent_distract	The percentage of all accidents that were distracted driving related for the specified year (distract_count/total_count)
percent_change	The amount the percentage of distracted driving accidents changed over the previous year

4.2.4 DISTRACT_LEVELS_TIMECSV.PY

This script returns a CSV showing how the different types of distractions have changed over time.

Usage:
$ SPARK-SUBMIT distract_levels_timeCSV.py

Fields returned:

Field Name	Description
year	The record year (as a date 01-01-yyyy)
factor	The specific distraction
factor_count	The number of times the distraction was recorded in the specified year
year_all_factor_count	The total number of all distractions recorded in the specified year
year_percent_all_factors	The percentage the given distraction over all distractions in the specified year
year_percent_change	The amount the percentage of the given distraction changed over the previous years

4.2.5 PERSON_INJURIESCSV.PY

This script returns a CSV showing the injuries of the various people involved in the accidents.

Usage:
$ SPARK-SUBMIT person_injuriesCSV.py

Fields returned:

Field Name	Description
person_type	The type of person for this record; distracted drivers are returned as 'DISTRACTED Driver'
injury_severity	The severity of the injury
number_persons	The number of people of the given person type that had the specified injury
percent_within type	The number of people with the specified injury over all injuries for that person type.

4.2.6 PRECRASH_EVENTCSV.PY

This script returns a CSV showing the event that occurred just prior to the occurrence of the accident, by distracted and non-distracted drivers

Usage:
$ SPARK-SUBMIT precrash_eventCSV.py

Fields returned:

Field Name	Description
driver_attention	Specifies if this record is for 'Not Distracted' or 'Distracted' drivers
critical_event_precrash	The precrash event for this record.
number_vehicles	The number of vehicles for the given driver attention and specified precrash event.
is_distracted_count	The count of all vehicles for the given driver attention (all precrash events)
percent_vehicles	The percentage of vehicles over all vehicles for the given driver attention for the specified precrash event (number_vehicles/is_distracted_count)

4.2.7 VEHICLE_DAMAGECSV.PY

This script returns a CSV showing extent of vehicle damage, by distracted and non-distracted drivers.

Usage:
$ SPARK-SUBMIT vehicle_damageCSV.py

Fields returned:

Field Name	Description
driver_attention	Specifies if this record is for 'Not Distracted' or 'Distracted' drivers
extent_damage	The extent of damage to the vehicle
number_vehicles	The number of vehicles for the given driver attention and specified extent of damage.
percent_vehicles	The percentage of vehicles over all vehicles for the given driver attention for the specified extent of damage

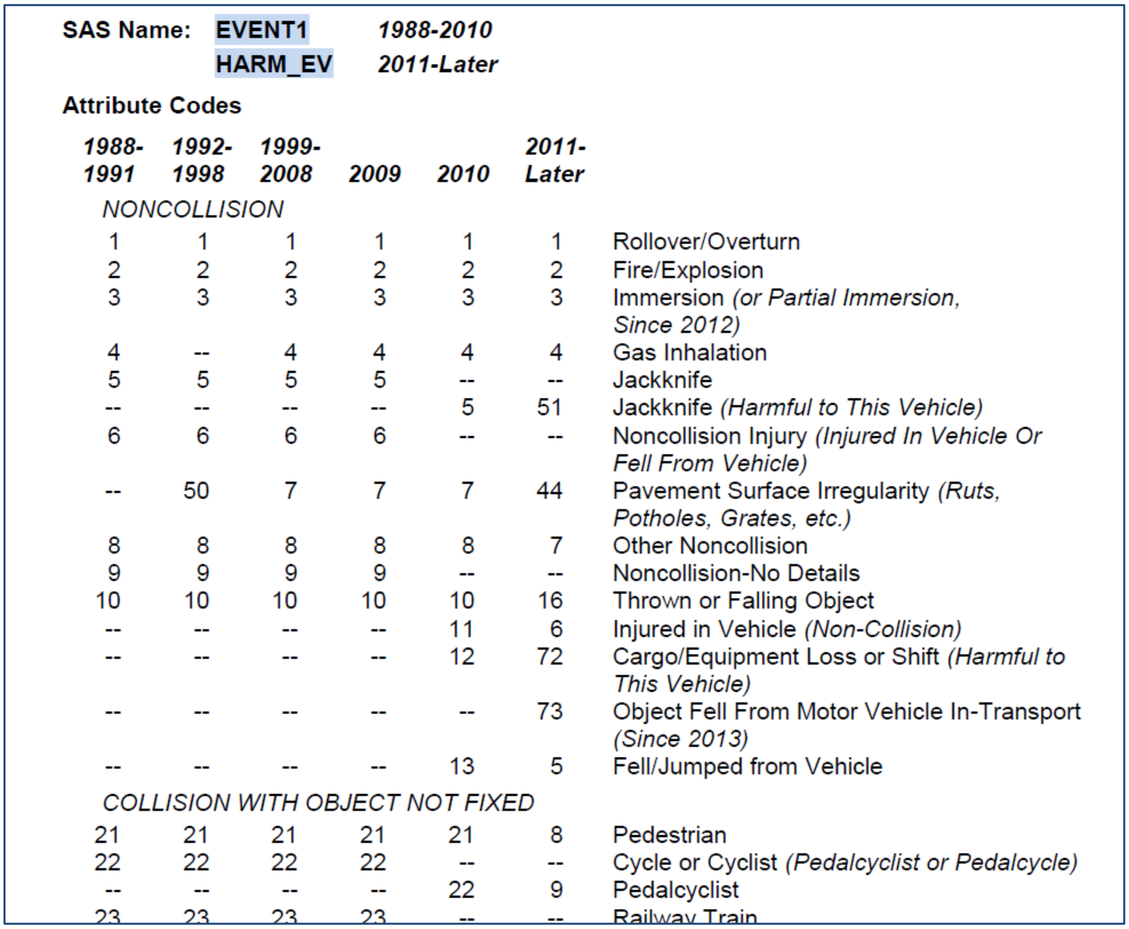
5 LIMITATIONS & SCALING UP

This project is not without limitations.

5.1 DATA SOURCE

Each year in the NASS GES raw data demands special attention. It cannot be assumed that the numeric values for one year will be the same ones used the next year. The field names, and even the order the fields are provided in the raw data can also change.

For example, the first harmful event field has a field name change in 2011. It also has four different interpretations of codes for the periods in this project. While many stay the same, others are added, and others given different descriptions. Each case must be evaluated separately. This means, even when new data files arrive, the data cannot easily be added to the project database without due diligence in cleaning the data.



SAS Name: **EVENT1** ***1988-2010***
HARM_EV ***2011-Later***

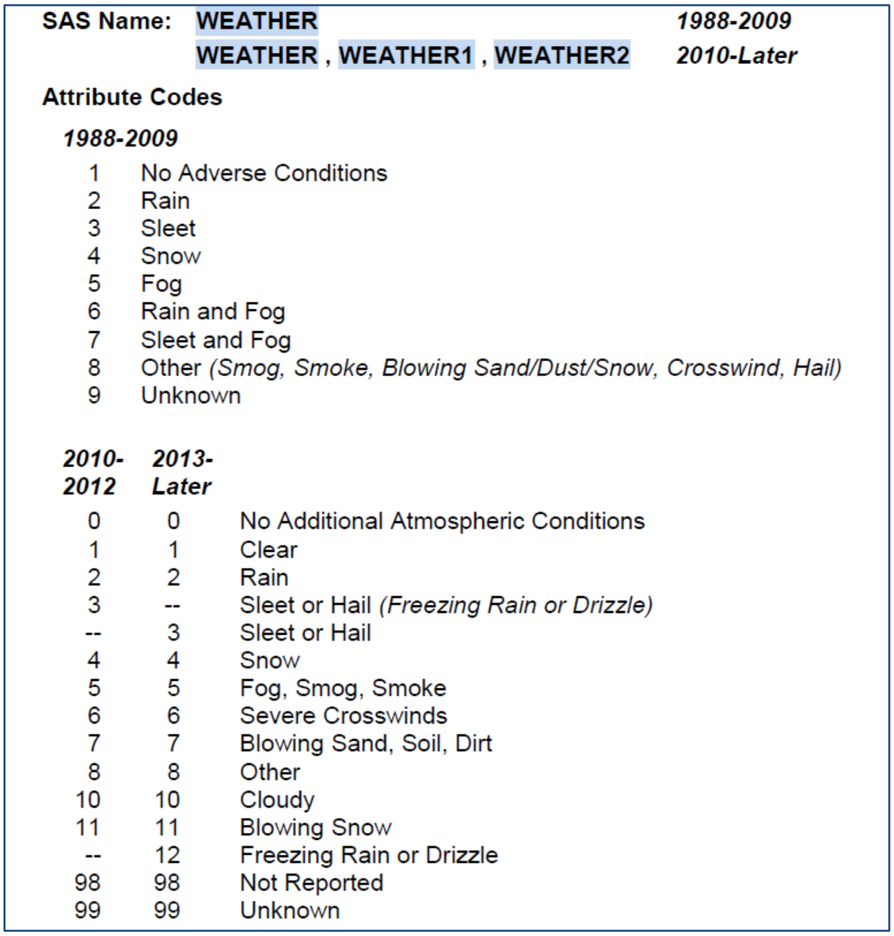
Attribute Codes

1988-1991	***1992-1998***	***1999-2008***	***2009***	***2010***	***2011-Later***	
NONCOLLISION						
1	1	1	1	1	1	Rollover/Overturn
2	2	2	2	2	2	Fire/Explosion
3	3	3	3	3	3	Immersion *(or Partial Immersion, Since 2012)*
4	--	4	4	4	4	Gas Inhalation
5	5	5	5	--	--	Jackknife
--	--	--	--	5	51	Jackknife *(Harmful to This Vehicle)*
6	6	6	6	--	--	Noncollision Injury *(Injured In Vehicle Or Fell From Vehicle)*
--	50	7	7	7	44	Pavement Surface Irregularity *(Ruts, Potholes, Grates, etc.)*
8	8	8	8	8	7	Other Noncollision
9	9	9	9	--	--	Noncollision-No Details
10	10	10	10	10	16	Thrown or Falling Object
--	--	--	--	11	6	Injured in Vehicle *(Non-Collision)*
--	--	--	--	12	72	Cargo/Equipment Loss or Shift *(Harmful to This Vehicle)*
--	--	--	--	--	73	Object Fell From Motor Vehicle In-Transport *(Since 2013)*
--	--	--	--	13	5	Fell/Jumped from Vehicle
COLLISION WITH OBJECT NOT FIXED						
21	21	21	21	21	8	Pedestrian
22	22	22	22	--	--	Cycle or Cyclist *(Pedalcyclist or Pedalcycle)*
--	--	--	--	22	9	Pedalcyclist
23	23	23	23	--	--	Railway Train

5.2 SUMMARIZED DATA

In order to simplify the data, elements were rolled up into more general categories. For example, with First Harmful Event, data was rolled up into a general ‘Non-Collision’ category. This loses some detail in the information, but with respect to distracted driving, it was determined knowing if the non-collision was specifically due to a rollover or pavement surface was not important.

In addition to rolling up data, some data elements were split out. Weather on the accident table was split into three separate fields in 2010. The choice was to combine the fields into one field as was used prior to 2010, or to move to the three field format. The three field format was selected to be forward compatible.



SAS Name:		
	WEATHER	*1988-2009*
	WEATHER , WEATHER1 , WEATHER2	*2010-Later*

Attribute Codes

1988-2009

1	No Adverse Conditions
2	Rain
3	Sleet
4	Snow
5	Fog
6	Rain and Fog
7	Sleet and Fog
8	Other *(Smog, Smoke, Blowing Sand/Dust/Snow, Crosswind, Hail)*
9	Unknown

2010-2012	***2013-Later***	
0	0	No Additional Atmospheric Conditions
1	1	Clear
2	2	Rain
3	--	Sleet or Hail *(Freezing Rain or Drizzle)*
--	3	Sleet or Hail
4	4	Snow
5	5	Fog, Smog, Smoke
6	6	Severe Crosswinds
7	7	Blowing Sand, Soil, Dirt
8	8	Other
10	10	Cloudy
11	11	Blowing Snow
--	12	Freezing Rain or Drizzle
98	98	Not Reported
99	99	Unknown

5.3 CHOICE OF DATABASE

The HIVE database was selected as it was not foreseen that a relational database would be necessary. The biggest issue with the choice of having everything stored as strings is one spelling mistake can distort the data queried. Strings were hardcoded in the python UDF files. Given there is one file for each year for each table, it was very possible for errors to be made. Relational databases would have better supported using look-up or reference tables. The tables could be indexed by year and code so one spot would have the logic instead of the many scripts; updating to new years would also be easier. It would just be a matter of inserting the raw data into the target tables and inserting rows into the look-up tables for the field mappings.

5.4 SCALING UP & FUTURE WORK

One of the developer's reasons for selecting HIVE as the database was for the learning experience. If time and resources allow, it would be interesting to see the data using a SAS platform.

More work to make the data more accessible to users who do not have SQL knowledge or access to the infrastructure required to run this project.

Of course one of the biggest scaling up work that could be done is to add more years and more tables. Data collection began in 1988, and distraction data collection in 2002. There are about 18 tables/files per year, so really only the surface of the data available was scratch with this project.

Another opportunity is to add additional datasets from other sources. France has similar open data on accidents; one of the biggest challenges with this particular data set is translating the data from French to English. As a side note, the visualizations created from the France data set is quite interesting: http://www.r-bloggers.com/inter-relationships-in-a-matrix/

6 TABLE META DATA

This document lists the tables from the Accident Project Database. It includes the list of elements for each table and interpretation of the possible values within each element.

Data is originally from the National Automotive Sampling System (NASS) General Estimates System (GES). http://www.nhtsa.gov/NASS
Information on the GES data can be found here:

http://www.nhtsa.gov/Data/National+Automotive+Sampling+System+(NASS)/NASS+General+Estimates+System

Data elements from the original files have been filtered to only elements deemed useful for the purposes of looking at distracted driving. The Project elements may be directly related to an element in GES, or may be a result of combining or splitting a data element in GES into a more useful format.

Summary of the difference in table size:

Table	GES elements	Project elements
Accident	53	21
Vehicle	72	18
Person	32	15
Distract	8	5

The Accident_Project database has 4 tables: Accident, Vehicle, Person and Distract. Common to all tables are the below columns:

- case_number
- file_year
- year

6.1 ACCIDENT TABLE

The Accident table contains general information about the police reported accident. Records are identified by case_number, which is unique for all cases.

6.1.1 CASE_NUMBER

The unique number for each accident.

6.1.2 FILE_YEAR

The year the file was taken from.

6.1.3 MONTH

The month the accident occurred.

Value
January
February
March
April
May
June
July
August
September
November
December

6.1.4 DAY_OF_WEEK

The day of the week the accident occurred.

Value
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Unknown

6.1.5 REGION

The region where the accident occurred.

Value	Notes
Northeast	Included states: PA, NJ, NY, NH, VT, RI, MA, ME, CT
Midwest	Included states: OH, IN, IL, MI, WI, MN, ND, SD, NE, IA, MO, KS
South	Included states: MD, DE, DC, WV, VA, KY, TN, NC, SC, GA, FL, AL, MS, LA, AR, OK, TX
West	Included states: MT, ID, WA, OR, CA, NV, NM, AZ, UT, CO, WY, AK, HI

6.1.6 NUM_VEHICLES

The number of vehicles involved that were in transit

6.1.7 NUM_PARKWORK

The number of vehicles that were not in transit (parked or working vehicles).

The information is only available 2005 and later. 999 is used if the value is not available.

6.1.8 PERSONS_PED

The number of people involved who are pedestrians.

The information is only available 2011 and later. 999 is used if the value is not available.

6.1.9 PERSONS_NOTTRANSIT

The number of people who are non-motorists. This field is more inclusive than persons_ped.

A non-motorist is defined as a pedestrian, a cyclist, an occupant of a motor vehicle not in-transport, a person riding a horse, an occupant of an animal drawn conveyance, person associated with non-motorist conveyance (e.g., baby carriage, skate board, wheelchair), or an other non-motorist (e.g., person outside a trafficway, person in a house).

6.1.10 PERSONS_TRANSIT

The number of people who are in a vehicle in transit at the time of the accident.

The information is only available 2011 and later. 999 is used if the value is not available.

6.1.11 FIRST_HARM_EVENT

This element describes the first event that resulted in damage or injuries.

Value	Notes
Non-Collision	Rollover, Fire/Explosion, Jackknife, and other accidents that do not involve a collision.
Collision – Object Not Fixed	Collision with a pedestrian, cyclist, train, animal or other objects that are not in a fixed position.
Collision – Object Fixed	Collision with a building, bridge structure, traffic barrier, curb, tree or other non-mobile object.
Collision – Vehicle in Transport	Collision with a motor vehicle; prior to 2010 these events were classified as Collision – Object Not Fixed.
Unknown	Not reported or Unknown.

6.1.12 MANNER_COLLISION

This element describes the orientation of two motor vehicles when the first harmful event occurred.

Value
Front-to-Rear
Front-to-Front
Rear-to-Rear
Angle
Sideswipe – Same Direction
Sideswipe – Opposite Direction
Unknown

6.1.13 WITHIN_INTERCHANGE

This element identifies if the accident is in presence of an interchange.

Value
Yes
No
Unknown

6.1.14 RELATION_TO_JUNCTION

This element identifies the accident location with respect to the junction. It correlates with the within_interchange element.

Value
Non-Junction
Intersection
Intersection Related
Interchange Area
Driveway Access
Entrance or Exit Ramp Related
Rail Grade Crossing
Crossover Related
Through Roadway
Other Location Within Interchange Area
Unknown

6.1.15 INTERSECTION_TYPE

This element identifies the different types of intersections.

Value
Not an Intersection
Four-Way
T-Intersection
Y-Intersection
Traffic Circle
Roundabout
Five-Point or More
Unknown

6.1.16 WORK_ZONE

This element identifies if the accident occurred in a work zone (construction area).

Value
No
Yes

6.1.17 WEATHER1, WEATHER 2, WEATHER 3

These elements record the prevailing atmospheric conditions.

Prior to 2010, only one field was recorded, but from 2010 and later three separate fields were recorded. Efforts have been made to map the single records from prior to 2010 into multiple fields where possible. For example, if the 2009 weather value was "Rain and Fog," it would be transposed into the accident_project database as weather1 = "Rain" and weather2 = "Fog."

Value
Clear
Rain
Sleet or Hail
Snow
Fog or Smog or Smoke
Other
Unknown

6.1.18 MAX_INJURY

This element indicates the most severe injury recorded in the accident. The numbers indicate the suggested order of severity that has been used by NASS since 2001.

Value
01 Fatal
02 Incapacitated
03 Non-Incapacitating
04 Possible Injury
05 Injured, Unknown Severity
06 No Injury
07 Died Prior
08 Unknown if Injured
09 No Person Involved

6.1.19 NUM_INJURY

The number of people with injuries in 1 – 5 of the max injury list.

Special cases:

98 : No Person Involved in the Crash
99 : All Persons in Crash are Unknown if Injured

6.1.20 ALCOHOL_INVOLVED

This element indicates if alcohol was involved in the accident.

Note: “No – Not Applicable” is used only if the accident involves only passengers of in-transport motor vehicles, occupants of motor vehicles not in-transport or unknown occupant types who are in an in-transport motor vehicle where there is no driver present.

Value
Yes – Alcohol Involved
No – No Alcohol Involved
No – Not Applicable
Unknown

6.1.21 YEAR

This field is the same as “file_year”; it is included as part of the partitioning of the table.

6.2 VEHICLE TABLE

The Vehicle table contains data about vehicles in transport, including information about the driver and precrash data. Records are identified by case_number and vehicle_number, which together are unique.

6.2.1 CASE_NUMBER

The unique number for each accident.

6.2.2 VEHICLE_NUMBER

The consecutive number assigned to the vehicle; numbers start at 1.

6.2.3 FILE_YEAR

The year the file was taken from.

6.2.4 NUM_OCCUPANTS

The number of occupants in the vehicle.

Special cases:

0 : None
96 : Ninety-six or More (2009 – Later)
97 : Unknown (2010 only)
99 : Unknown (2009 – Later)
999 : Unknown (2003 – 2008)

6.2.5 BODY_TYPE

This element identifies the classification of vehicle.

Value	Notes
Automobile	Includes convertibles, sedans, hatchbacks, wagons.
Automobile Derivative	Includes limousine, auto-based pickup (El Camino)
Utility Vehicle	Includes Landover, Bronco, Landcruiser, Bronco II, etc.
Van-Based Light Truck	Includes minivans, standard vans and other van type vehicles.
Light Conventional Truck (pickup style)	Includes pickup style cab, <= 10,000 LBS
Other Light Truck	Includes cab chassis based (dump or tow truck), and other light conventional trucks (<= 10,000 LBS)
Bus	Includes school buses and other buses.
Medium or Heavy Trucks	Includes truck-tractors, and single unit straight trucks (>= 10,000 LBS)
Motor Home	Includes chassis mounted and medium/heavy truck based motor homes
Motored Cycles, Mopeds, All-Terrain Vehicle	Includes motorcycles, mopeds, scooters, ATVs
Farm or Construction Vehicle	Includes farm equipment and construction equipment (like graders)
Other Vehicle	Includes motor vehicles not included in previous categories
Unknown	Unknown or not reported.

6.2.6 MODEL_YEAR

The manufacturer's model year of vehicle.

Special cases:

7777 : Not Reported (2010)
9998 : Not Reported (2011 – Later)
9999 : Unknown

6.2.7 EXTENT_DAMAGE

This element records the amount of damage sustained by the vehicle on an operational damage scale.

Value
No Damage
Minor Damage
Functional Damage
Disabling Damage
Unknown

6.2.8 MOST_HARM_EVENT

This element describes the first event that resulted in damage or injuries.

Value	Notes
Non-Collision	Rollover, Fire/Explosion, Jackknife, and other accidents that do not involve a collision.
Collision – Object Not Fixed	Collision with a pedestrian, cyclist, train, animal or other objects that are not in a fixed position.
Collision – Object Fixed	Collision with a building, bridge structure, traffic barrier, curb, tree or other non-mobile object.
Collision – Vehicle in Transport	Collision with a motor vehicle; prior to 2010 these events were classified as Collision – Object Not Fixed.
Unknown	Not reported or Unknown.

6.2.9 NUM_INJURY

The number of people who were injured in this vehicle.

Special cases:

0 : No Person Injured
98 : No Person in Vehicle
99 : All Persons in the Vehicle are Unknown if Injured

6.2.10 MAX_INJURY

This element indicates the most severe injury recorded in the accident. The numbers indicate the suggested order of severity that has been used by NASS since 2001.

Value
01 Fatal
02 Incapacitated
03 Non-Incapacitating
04 Possible Injury
05 Injured, Unknown Severity
06 No Injury
07 Died Prior
08 Unknown if Injured
09 No Person Involved

6.2.11 DRIVER_DRINKING

This element indicates if the driver was drinking, according to police-reported alcohol involvement.

Value
Alcohol Involved
No Alcohol Involved
No Driver Present
Unknown

6.2.12 SPEEDING

This element records if the driver's speed was related to the crash as indicated by law enforcement.

Value
No
Yes
No Driver Present
Unknown

6.2.13 TRAVEL_SPEED

The speed the vehicle was travelling prior to the accident as reported by investigating officer.

Special cases:

- 0 : Stopped
- 997 : Speed Greater than 151 mph (2009 – Later)
- 998 : Not Reported (2009 – Later)
- 999 : Unknown

6.2.14 PRE_EVENT_MOVEMENT

This element identifies the vehicle's activity prior to the driver's realization of the impending critical event or just prior to impact if the driver took no action.

Value
No Driver Present
Going Straight
Decelerating in Road
Accelerating in Road
Starting in Travel Lane
Stopped in Road
Passing or Overtaking Another Verhicle
Disabled or Parked in Travel Lane
Leaving a Parking Position
Entering a Parking Position
Turning Right
Turning Left
Making a U-turn
Baking Up
Negotiating a Curve
Changing Lanes
Merging
Successful Corrective Action to Previous Critical Event
Other
Unknown

6.2.15 CRITICAL_EVENT_PRECRASH

This element describes the critical event which made the accident imminent.

Value	Notes
Vehicle Loss of Control	Includes flat tire, stalling, poor road conditions and excessive speed for conditions
Vehicle Travelling on Lane Edge	Includes any movement that results in vehicle leaving travel lane or road edge
Vehicle Turning at Junction	Includes turning, any direction, at an intersection
Vehicle Crossing Intersection	Vehicle was passing through the intersection
Vehicle Decelerating	Vehicle was decelerating
Other Motor Vehicle in Lane	Includes any situation where another vehicle is travelling in this vehicle's lane
Other Motor Vehicle Encroaching Into Lane	Includes any situation where another vehicle is encroaching into this vehicle's lane.
Pedestrian, Pedalcyclist or Other Non-Motorist in Road	Includes any action of a non-motorist including being in the roadway or approaching the roadway
Animal in Road	Includes any action of an animal including being in the roadway or approaching the roadway
Object in Road	Includes any action of an object including being in the roadway or approaching the roadway
Other Event	Other critical precrash event
Unknown	Unknown or not reported.

6.2.16 ACCIDENT_CATEGORY & ACCIDENT_TYPE

These two elements describe the crash itself. The category gives a broad label to the type of accident and the accident type gives a more detailed look.

Accident_Category	Notes
No Impact	
Single Driver	
Same Trafficway, Same Direction	
Same Trafficway, Opposite Direction	
Changing Trafficway or Turning	
Intersecting Paths (Vehicle Damage)	"T-Bone" type accidents
Miscellaneous	Includes accident_type "Backing of Vehicle"
Unknown	

Accident_Type	Notes
No Impact	
Right Roadside Departure	Includes driving off, losing control or avoidance maneuvers.
Left Roadside Departure	Includes driving off, losing control or avoidance maneuvers.
Forward Impact	Vehicle frontal area is impacted
Rear End	Vehicle rear area is impacted
Sideswipe or Angle	Includes strikes in all sides
Head-On	Front-to-Front collision
Turn Across Path	Includes all turning directions
Turn Into Path	Includes all turning directions
Straight Paths (T-Bone)	Includes strikes in all sides
Backing or Other	Miscellaneous types.
Unknown	

6.2.17 YEAR

This field is the same as “file_year”; it is included as part of the partitioning of the table.

6.3 PERSON TABLE

The Person table includes all motorist and non-motorist data. Records are identified by case_number, vehicle_number and person_number which together are unique.

6.3.1 CASE_NUMBER

The unique number for each accident.

6.3.2 VEHICLE_NUMBER

The consecutive number assigned to the vehicle; numbers start at 1.

6.3.3 PERSON_NUMBER

The consecutive number assigned to persons in the vehicle, starting at 1.

6.3.4 FILE_YEAR

The year the file was taken from.

6.3.5 AGE

The age of the person.

Special cases:

997 : Not Reported (2010)
998 : Not Reported (2011 – Later)
999 : Unknown

6.3.6 SEX

The sex of the person involved.

Value
Male
Female
Unknown

6.3.7 PERSON_TYPE

This element describes the person involved.

Value	Notes
Driver	
Passenger	
Occupant – Unknown	Occupant of a vehicle, but in unknown capacity.
Occupant – Not in Transport	Occupant of a vehicle not in transport
Occupant – Non-Motor Vehicle	Occupant of a non-motored vehicle
Pedestrian	
Cyclist	
Persons in or on Buildings	
Other or Unknown Non-Occupant or Motorist	

6.3.8 INJURY_SEVERITY

The severity of injury to this person. The numbers indicate the suggested order of severity that has been used by NASS since 2001.

Value
01 Fatal
02 Incapacitated
03 Non-Incapacitating
04 Possible Injury
05 Injured, Unknown Severity
06 No Injury
07 Died Prior
08 Unknown if Injured
09 No Person Involved

6.3.9 SEATING_ROW & SEATING POSITON

The seat the person was in.

Seating Row
Non-Motorist
Front Seat
Second Seat
Third Seat
Fourth Seat
Other Seat Location
Unknown

Seating Position
Non-Motorist
Left Side
Middle
Right Side
Other
Sleeper Section of Cab
Cargo Area
Trailing Unit
Riding on Vehicle Exterior
Unknown

6.3.10 RESTRAINT_USE

This element indicates of the person was using a seating restraint.

Seating Position
Belt Restraint
Not Applicable
No Restraint Used
Child Restraint
Helmet Used
Other
Unknown

6.3.11 DRINKING

This element indicates if alcohol was involved for this person and reflects the judgement of law enforcement.

Value
No Alcohol Involved
Alcohol Involved
Unknown

6.3.12 DRUGS

This element indicates if drugs were involved for this person and reflects the judgement of law enforcement.

Value
No Drugs Involved
Drugs Involved
Unknown

6.3.13 STRIKING_VEHICLE_NUMBER

In the case the person is not an occupant of a motor vehicle, this element identifies the vehicle, if any, that hit the person.

Special cases:

0 : Occupant of Motor Vehicle
999 : Unknown

6.3.14 YEAR

This field is the same as “file_year”; it is included as part of the partitioning of the table.

6.4 DISTRACT TABLE

The Distract table identifies each driver distraction as a separate record. Records are identified by case_number and vehicle_number, which together are unique. Note the person number is not on these records.

6.4.1 CASE_NUMBER

The unique number for each accident.

6.4.2 VEHICLE_NUMBER

The consecutive number assigned to the vehicle; numbers start at 1.

6.4.3 FILE_YEAR

The year the file was taken from.

6.4.4 FACTOR

This element identifies what best describe the driver’s attention to driving.

Value	Notes
Not Distracted	
Looked but Did Not See	
By Other Passengers	
By a Moving Object in Vehicle	
Talking or Listening to Cellular Phone	
Manipulating Cellular Phone	
Adjusting Climate or Audio Controls	
Adjusting Other Controls	
Using or Reaching other Devices	
Distracted by Outside Person or Event	
Eating or Drinking	
Smoking	
Inattention – Lost in Thought	Includes sleepiness, daydreaming, and general carelessness
Other Distraction	
Unknown	

6.4.5 YEAR

This field is the same as “file_year”; it is included as part of the partitioning of the table.

7 APPENDIX

7.1 LINUX FILE SYSTEM

205Project/

```
        start_up.sh
```

205Project/accident_project/raw_data/

```
        2002_accident.txt
        2002_distract.txt
        2002_person.txt
        2002_vehicle.txt
        2003_accident.txt
        2003_distract.txt
        2003_person.txt
        2003_vehicle.txt
        2004_accident.txt
        2004_distract.txt
        2004_person.txt
        2004_vehicle.txt
        2005_accident.txt
        2005_distract.txt
        2005_parked.txt
        2005_person.txt
        2005_vehicle.txt
        2006_accident.txt
        2006_distract.txt
        2006_parked.txt
        2006_person.txt
        2006_vehicle.txt
        2007_accident.txt
        2007_distract.txt
        2007_parked.txt
        2007_person.txt
        2007_vehicle.txt
        2008_accident.txt
        2008_distract.txt
        2008_parked.txt
        2008_person.txt
        2008_vehicle.txt
        2009_accident.txt
        2009_distract.txt
        2009_parked.txt
        2009_person.txt
        2009_vehicle.txt
        2010_accident.txt
        2010_distract.txt
        2010_parked.txt
        2010_person.txt
        2010_vehicle.txt
        2011_accident.txt
        2011_distract.txt
        2011_parkwork.txt
        2011_person.txt
        2011_vehicle.txt
        2012_accident.txt
        2012_distract.txt
        2012_parkwork.txt
        2012_person.txt
        2012_vehicle.txt
        2013_accident.txt
        2013_distract.txt
        2013_parkwork.txt
        2013_person.txt
        2013_vehicle.txt
```

205Project/loading_and_modelling/

```
        2002_hive_base_tables.sql
        2003_hive_base_tables.sql
        2004_hive_base_tables.sql
        2005_hive_base_tables.sql
        2006_hive_base_tables.sql
        2007_hive_base_tables.sql
        2008_hive_base_tables.sql
        2009_hive_base_tables.sql
        2010_hive_base_tables.sql
        2011_hive_base_tables.sql
        2012_hive_base_tables.sql
        2013_hive_base_tables.sql
        load_data_2002.sh
        load_data_2003.sh
        load_data_2004.sh
        load_data_2005.sh
        load_data_2006.sh
        load_data_2007.sh
        load_data_2008.sh
        load_data_2009.sh
        load_data_2010.sh
        load_data_2011.sh
        load_data_2012.sh
        load_data_2013.sh
```

205Project/query_scripts/

```
        distract_accidents_timeCSV.py
        distract_levels_timeCSV.py
        histogram.py
        person_injuriesCSV.py
        precrasth_eventCSV.py
        table_create_csv.py
        vehicle_damageCSV.py
```

205Project/transforming/

```
2002Accident.py
2002Distract.py
2002Person.py
2002Vehicle.py
2003Accident.py
2003Distract.py
2003Person.py
2003Vehicle.py
2004Accident.py
2004Distract.py
2004Person.py
2004Vehicle.py
2005Accident.py
2005Distract.py
2005Person.py
2005Vehicle.py
2006Accident.py
2006Distract.py
2006Person.py
2006Vehicle.py
2007Accident.py
2007Distract.py
2007Person.py
2007Vehicle.py
2008Accident.py
2008Distract.py
2008Person.py
2008Vehicle.py
2009Accident.py
2009Distract.py
2009Person.py
2009Vehicle.py
2010Accident.py
2010Distract.py
2010Person.py
2010Vehicle.py
2011Accident.py
2011Distract.py
2011Person.py
2011Vehicle.py
2012Accident.py
2012Distract.py
2012Person.py
2012Vehicle.py
2013Accident.py
2013Distract.py
2013Person.py
2013Vehicle.py
hive_target_accident.sql
hive_target_distract.sql
hive_target_person.sql
hive_target_vehicle.sql
hive_views.sql
```

7.2 HDFS FILE SYSTEM

user/w205/accident_project/

```
2002/
      ACCIDENT/2002_accident.txt
      DISTRACT/2002_distract.txt
      PERSON/2002_person.txt
      VEHICLE/2002_vehicle.txt
2003/
      ACCIDENT/2003_accident.txt
      DISTRACT/2003_distract.txt
      PERSON/2003_person.txt
      VEHICLE/2003_vehicle.txt

2004/
      ACCIDENT/2004_accident.txt
      DISTRACT/2004_distract.txt
      PERSON/2004_person.txt
      VEHICLE/2004_vehicle.txt
2005/
      ACCIDENT/2005_accident.txt
      DISTRACT/2005_distract.txt
      PERSON/2005_person.txt
      VEHICLE/2005_vehicle.txt
2006/
      ACCIDENT/2006_accident.txt
      DISTRACT/2006_distract.txt
      PERSON/2006_person.txt
      VEHICLE/2006_vehicle.txt
2007/
      ACCIDENT/2007_accident.txt
      DISTRACT/2007_distract.txt
      PERSON/2007_person.txt
      VEHICLE/2007_vehicle.txt
2008/
      ACCIDENT/2008_accident.txt
      DISTRACT/2008_distract.txt
      PERSON/2008_person.txt
      VEHICLE/2008_vehicle.txt
2009/
      ACCIDENT/2009_accident.txt
      DISTRACT/2009_distract.txt
      PERSON/2009_person.txt
      VEHICLE/2009_vehicle.txt
2010/
      ACCIDENT/2010_accident.txt
      DISTRACT/2010_distract.txt
      PERSON/2010_person.txt
      VEHICLE/2010_vehicle.txt
2011/
      ACCIDENT/2011_accident.txt
      DISTRACT/2011_distract.txt
      PERSON/2011_person.txt
      VEHICLE/2011_vehicle.txt
2012/
      ACCIDENT/2012_accident.txt
      DISTRACT/2012_distract.txt
      PERSON/2012_person.txt
      VEHICLE/2012_vehicle.txt
2013/
      ACCIDENT/2013_accident.txt
      DISTRACT/2013_distract.txt
      PERSON/2013_person.txt
      VEHICLE/2013_vehicle.txt
```

```
user/w205/accident_project/data/
       ACCIDENT/                                      PERSON/
              year=2002/000000_0                             year=2002/000000_0
              year=2003/000000_0                             year=2003/000000_0
              year=2004/000000_0                             year=2004/000000_0
              year=2005/000000_0                             year=2005/000000_0
              year=2006/000000_0                             year=2006/000000_0
              year=2007/000000_0                             year=2007/000000_0
              year=2008/000000_0                             year=2008/000000_0
              year=2009/000000_0                             year=2009/000000_0
              year=2010/000000_0                             year=2010/000000_0
              year=2011/000000_0                             year=2011/000000_0
              year=2012/000000_0                             year=2012/000000_0
              year=2013/000000_0                             year=2013/000000_0
       DISTRACT/                                      VEHICLE/
              year=2002/000000_0                             year=2002/000000_0
              year=2003/000000_0                             year=2003/000000_0
              year=2004/000000_0                             year=2004/000000_0
              year=2005/000000_0                             year=2005/000000_0
              year=2006/000000_0                             year=2006/000000_0
              year=2007/000000_0                             year=2007/000000_0
              year=2008/000000_0                             year=2008/000000_0
              year=2009/000000_0                             year=2009/000000_0
              year=2010/000000_0                             year=2010/000000_0
              year=2011/000000_0                             year=2011/000000_0
              year=2012/000000_0                             year=2012/000000_0
              year=2013/000000_0                             year=2013/000000_0
```

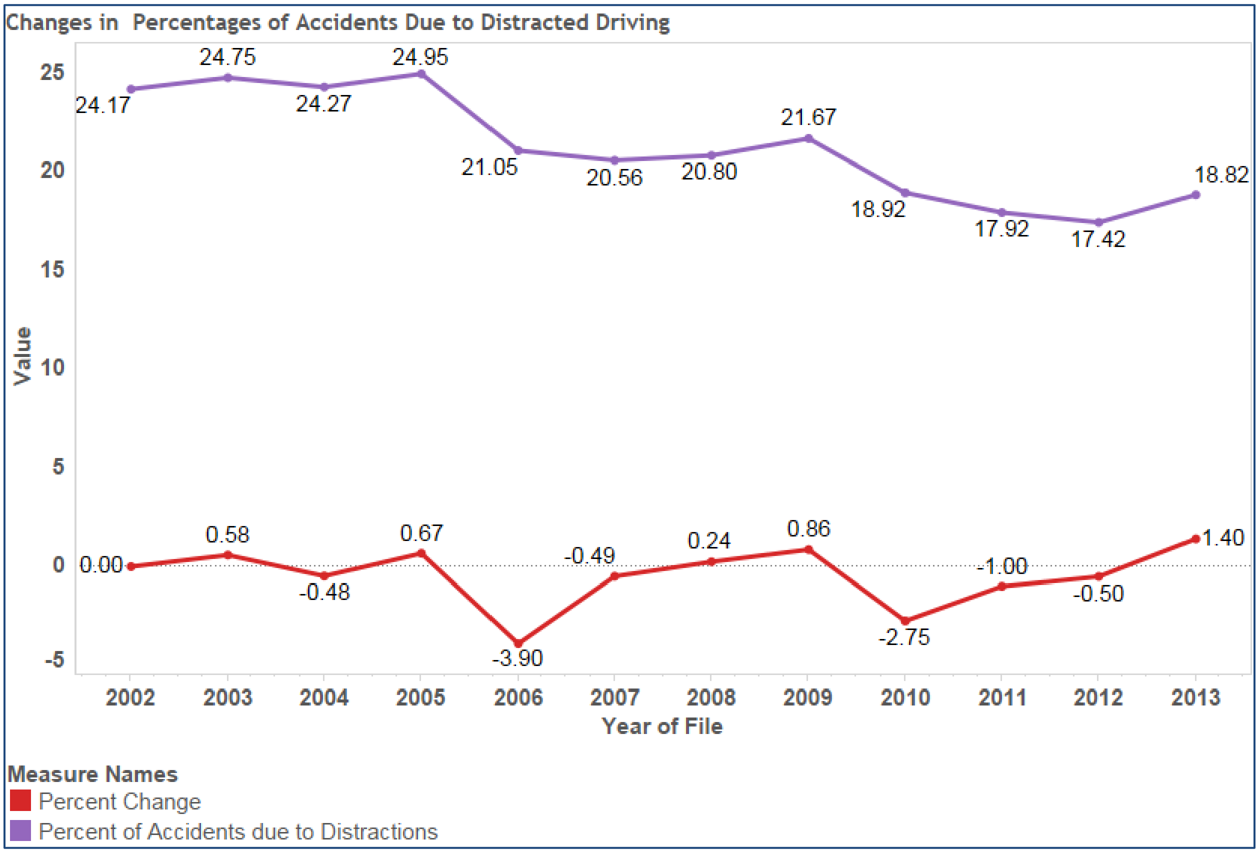
```
user/w205/accident_project/pyscripts/
       2002Accident.py              2006Accident.py               2010Accident.py
       2002Distract.py              2006Distract.py               2010Distract.py
       2002Person.py                2006Person.py                 2010Person.py
       2002Vehicle.py               2006Vehicle.py                2010Vehicle.py
       2003Accident.py              2007Accident.py               2011Accident.py
       2003Distract.py              2007Distract.py               2011Distract.py
       2003erson.py                 2007Person.py                 2011Person.py
       2003Vehicle.py               2007Vehicle.py                2011Vehicle.py
       2004Accident.py              2008Accident.py               2012Accident.py
       2004Distract.py              2008Distract.py               2012Distract.py
       2004Person.py                2008Person.py                 2012Person.py
       2004Vehicle.py               2008Vehicle.py                2012Vehicle.py
       2005Accident.py              2009Accident.py               2013Accident.py
       2005Distract.py              2009Distract.py               2013Distract.py
       2005Person.py                2009Person.py                 2013Person.py
       2005Vehicle.py               2009Vehicle.py                2013Vehicle.py
```

7.3 SOME VISUALIZATIONS

In response to the questions I asked in my proposal, here are some visualizations that attempt to answer those questions.

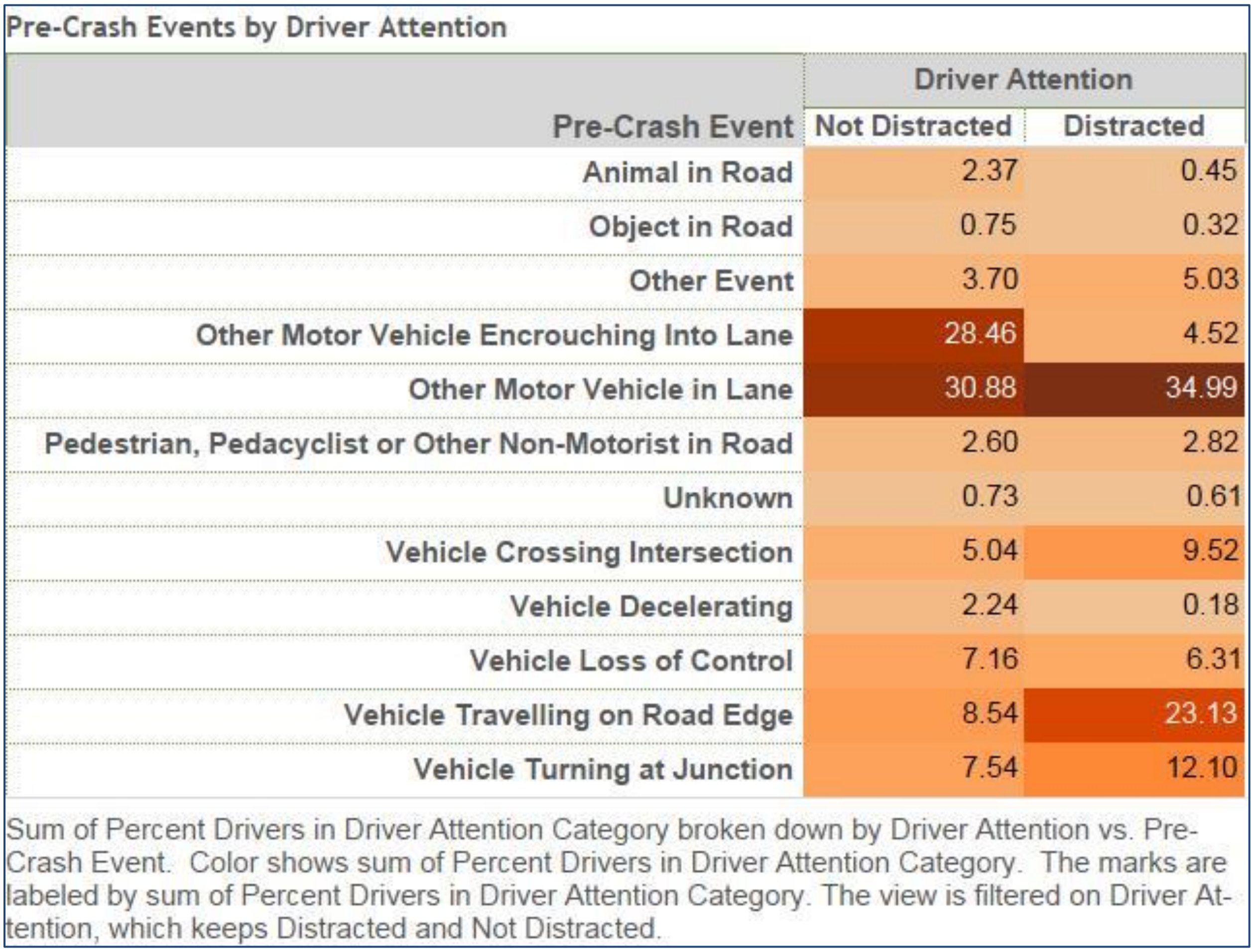
7.3.1 CHANGES IN PERCENTAGES OF ACCIDENTS DUE TO DISTRACTED DRIVING

This graph shows the percentage of accidents in the data set attributed to distracted driving and the percent change over time. We can see the percentage of distracted driving is about 5 percentage points lower in 2013 than it was in 2002. 2006 and 2010 were years of large decreases in distracted driving. 2013 had the largest year-over-year increase.



7.3.2 PRE-CRASH EVENTS BY DRIVER ATTENTION

This chart shows the different percentages of driver attention over the various pre-crash events. Of note, non-distracted drivers have a much larger portion of accidents with another motor vehicle encroaching into lane. This is likely the distracted driver, since distracted drivers have more accidents with their vehicle travelling on the road edge, meaning their vehicle wondered outside of their driving lane, and potentially into another driver's lane.



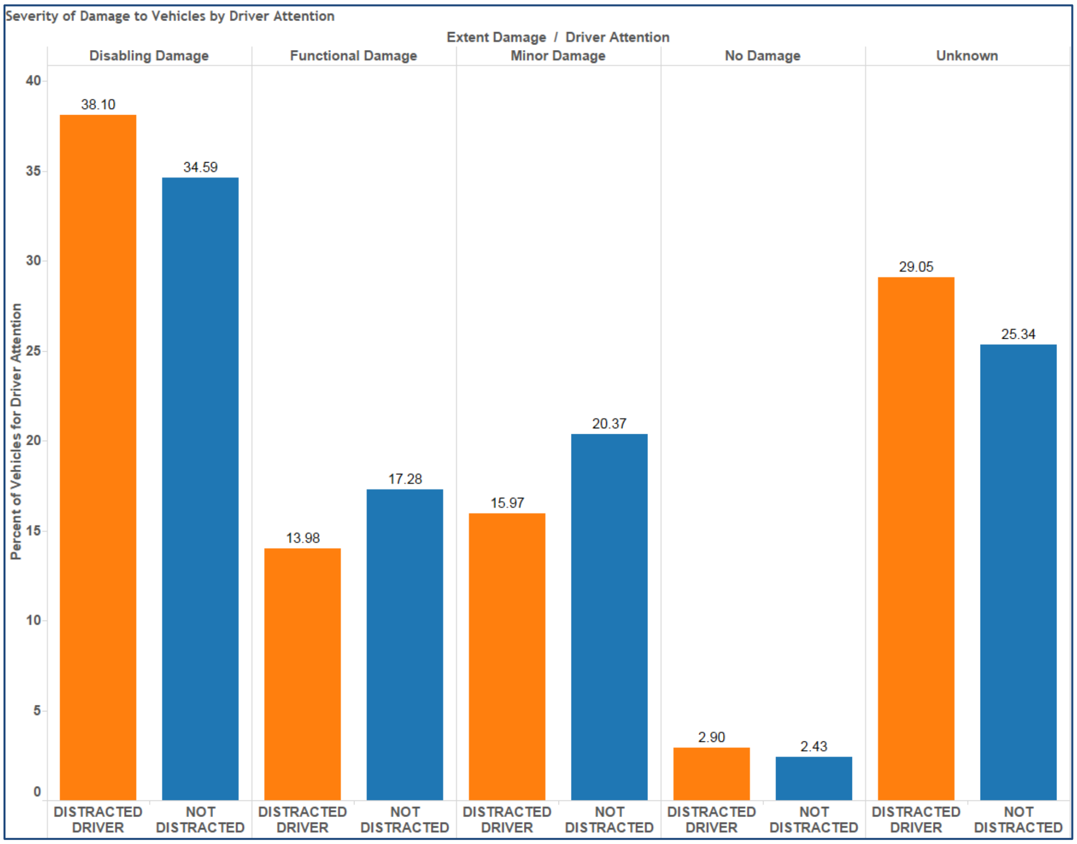
Pre-Crash Events by Driver Attention

	Driver Attention	
Pre-Crash Event	Not Distracted	Distracted
Animal in Road	2.37	0.45
Object in Road	0.75	0.32
Other Event	3.70	5.03
Other Motor Vehicle Encrouching Into Lane	28.46	4.52
Other Motor Vehicle in Lane	30.88	34.99
Pedestrian, Pedacyclist or Other Non-Motorist in Road	2.60	2.82
Unknown	0.73	0.61
Vehicle Crossing Intersection	5.04	9.52
Vehicle Decelerating	2.24	0.18
Vehicle Loss of Control	7.16	6.31
Vehicle Travelling on Road Edge	8.54	23.13
Vehicle Turning at Junction	7.54	12.10

Sum of Percent Drivers in Driver Attention Category broken down by Driver Attention vs. Pre-Crash Event. Color shows sum of Percent Drivers in Driver Attention Category. The marks are labeled by sum of Percent Drivers in Driver Attention Category. The view is filtered on Driver Attention, which keeps Distracted and Not Distracted.

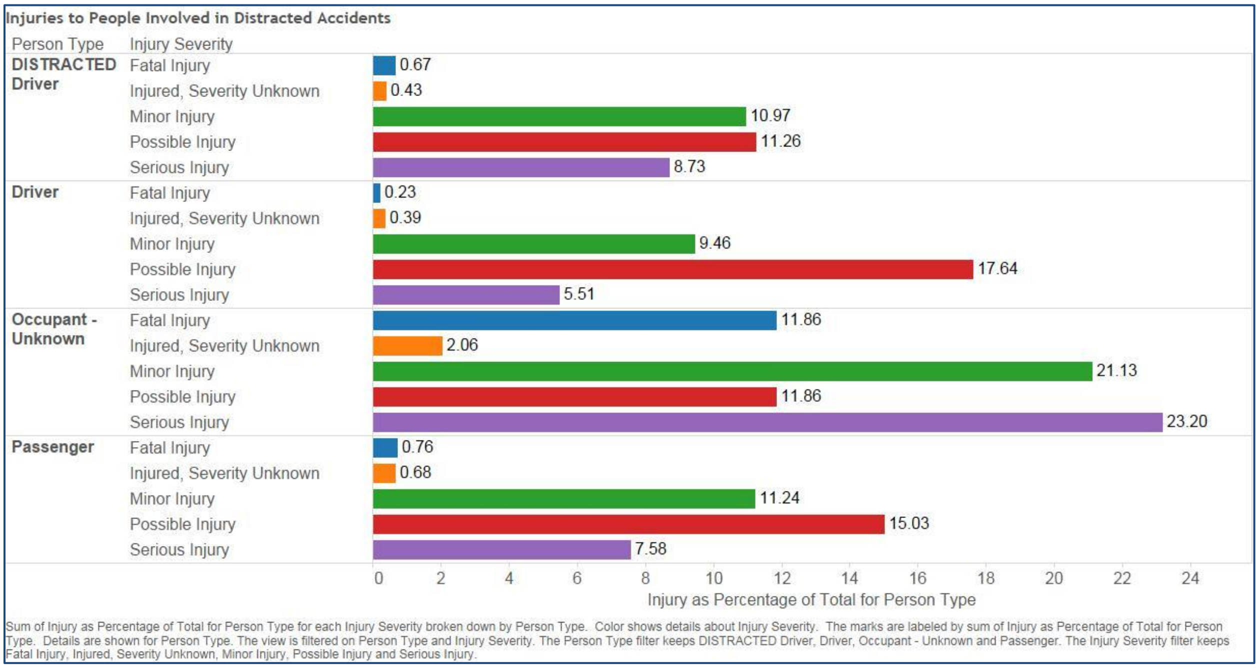
7.3.3 SEVERITY OF DAMAGE TO VEHICLES BY DRIVER ATTENTION

This graph compares the extent of damage to vehicles by the attention of the driver. There isn't much really interesting in this view; statistical analysis with a statistical package would validate if there is any significance here.



7.3.4 INJURIES TO PEOPLE INVOLVED IN DISTRACTED ACCIDENTS

This visual shows the severity of injuries to the different types of people. It has been filtered to show only four specific person types. This is another situation where statistical analysis in a statistical package like R would be interesting to see if there are any significant comparisons. It also highlights the frustrations seen in the data with many “Unknown” data values. Perhaps if imputed data was used here, the information would be more revealing.



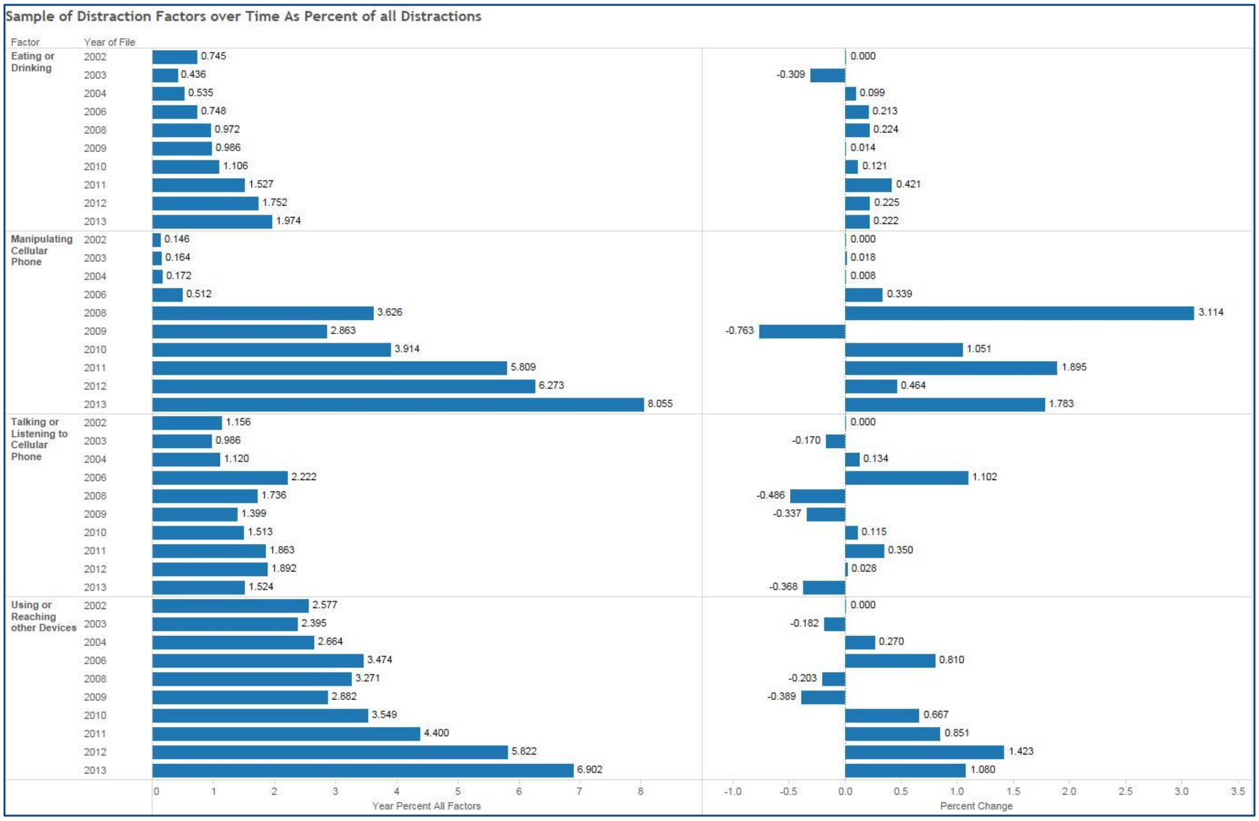
Injuries to People Involved in Distracted Accidents

Sum of Injury as Percentage of Total for Person Type for each Injury Severity broken down by Person Type. Color shows details about Injury Severity. The marks are labeled by sum of Injury as Percentage of Total for Person Type. Details are shown for Person Type. The view is filtered on Person Type and Injury Severity. The Person Type filter keeps DISTRACTED Driver, Driver, Occupant - Unknown and Passenger. The Injury Severity filter keeps Fatal Injury, Injured, Severity Unknown, Minor Injury, Possible Injury and Serious Injury.

7.3.5 SAMPLE OF DISTRACTION FACTORS OVER TIME

This final visual shows the change of distractions over time for a small subset of all available distractions. These were selected because they were deemed to be the most interesting.

This visual clearly shows an increase in the proportion of distracted driving due to manipulating a cell phone or reaching for a device, but also shows that actually talking or listening to a call has not really changed over time.



Sample of Distraction Factors over Time As Percent of all Distractions