031816

imported crashes2005\_2014,crashes2015, fatal2005\_2014, fatal2015

-to find out no of fatal crashes, ran query crashrptno\_killed\_05\_15:

SELECT crashrptno, killed from fatal2005\_2014

UNION

Select 1,2 from fatal2015

73 records

one record:

crashrptno  killed

1    2

-to find out which record it is, ran query "crashrptno=1\_fatal15" "crashrptno=1\_fatal05\_14"

no result

-to stitch together all data from 05 to 15, ran query "create\_crash\_05\_15", "create\_fatal\_05\_15":

CREATE TABLE crashl05\_15 AS SELECT \* FROM crash2015;

INSERT INTO crash05\_15 SELECT \* FROM crash2005\_2014;

CREATE TABLE fatal05\_15 AS SELECT \* FROM fatal2015;

INSERT INTO fatal05\_15 SELECT \* FROM fatal2005\_2014;

-to make tablenames consistent, removed plurals and "20"s

-redid the fatal crashes, ran query "no\_of\_fatal\_05\_15"

78 results

inconsistent with excel table final project, which has numbers of crashes, fatal crashes, and fatalities from MO highway patrol website

to find out which year is  missing, wanted to query each year, found the need to parse the years out of the date field, exported table crash05\_15, fatal05\_15

0321

"create table type\_str\_05\_15"

CREATE TABLE type\_str\_05\_15 AS SELECT crashrptno, crashtype, severity, AtStreet, OnStreet, injured, killed FROM crash05\_15 GROUP BY crashrptno

"crashtype\_w\_most\_crashes":

SELECT COUNT(crashrptno), crashtype FROM type\_str\_05\_15

GROUP BY 2

ORDER BY 1 DESC

"intersec\_w\_most\_crashes":

SELECT atstreet, onstreet, COUNT(crashrptno) FROM type\_str\_05\_15

GROUP BY 1, 2

ORDER BY 3 DESC

found that crash05\_15.csv has a line with no values.

deleted the line of record in crash05\_15 in SQLyog.

SELECT \* FROM crash05\_15

WHERE crashrptno = 0

then found that only the records in 2015 has no spaces, while the rest of the year has.for example, MO743 and MO 743. could influence the onstreet atstreet where most crashes happen. need to re-import.

redownloaded crashes2015. replaced two spaces, three spaces, 5,6 spaces, and space and ', and all '.

saved as crashes.txt

reimported crashes2015 into database crash. deleted crash15 and renamed crashes2015 into crash15. redid the stitch.

CREATE TABLE crashl05\_15 AS SELECT \* FROM crash15;

INSERT INTO crash05\_15 SELECT \* FROM crash05\_14;

then found that the spaces in fieldnames is messing me up. so replaced spaces with \_ in fieldnames by design table function in Navicat for all three tables: crash15, crash 05\_14, and crash05\_15

revised query "create table type\_str\_05\_15":

CREATE TABLE type\_str\_05\_15 AS SELECT crash\_rpt\_no, crash\_type, severity, At\_Street, On\_Street, injured, killed FROM crash05\_15 GROUP BY crash\_rpt\_no

revised query "crashtype\_w\_most\_crashes":

SELECT COUNT(crash\_rpt\_no), crash\_type FROM type\_str\_05\_15

GROUP BY 2

ORDER BY 1 DESC

ran query, result:

revised query "intersec\_w\_most\_crashes":

SELECT atstreet, onstreet, COUNT(crashrptno) FROM type\_str\_05\_15

GROUP BY 1, 2

ORDER BY 3 DESC

fatal05\_14(195) and fatal05\_15(208) doesn't have spaces between words. just going to leave them there.

LESSONS:

Keep record of how you cleaned the data, and keet consistent

Fieldnames: replace space with \_

Importing: .txt better; data type should be consistent

After importing: check intgrity, number of records,

0322

had Liz look over the data

add\_year\_field:

## Run this query to make sure that the YEAR( ) function works correctly on the `Date` field, because date is a function, so use ` as escape

SELECT `Date`, YEAR(`Date`), COUNT(\*)

FROM crash05\_15

GROUP BY 1, 2;

## this statement adds the field `crash\_year` as an integer field after the `Date` field:

ALTER TABLE crash05\_15 ADD crash\_year INT AFTER `Date`;

## this statement populates the `crash\_year` field with the results of the YEAR function:

UPDATE crash05\_15 SET crash\_year = YEAR(`Date`);

circumstances\_by\_year:

#to solve the problem of multiple records per crash, count(Distinct crash\_rpt\_no)

SELECT circumstances, COUNT(DISTINCT crash\_rpt\_no)

FROM crash05\_15

GROUP BY 1

ORDER BY 2 DESC

;

#to know the change over year, select crash\_year as well, and group by both circumstance and year

SELECT circumstances, crash\_year, COUNT(DISTINCT crash\_rpt\_no)

FROM crash05\_15

GROUP BY 1, 2

ORDER BY 3 DESC

crashtyype\_by\_year

intersections\_by\_year

on\_street\_by\_year

in\_street\_by\_year

exported those into spreadsheets, did pivot table and found a lot of things dropping.

0323

called Loretta, Missouri State Highway Patrol:

1. to get complete records, should set circumstance to blank and "all".

2. per circumstance per record, no records for pedestrians without circumstance

3. person involvement: pedestrian/driver present, means it's circumstance for pedestrian/driver present.

4. "Location", section 5 of the blank uniform crash report, has info about "?" miles before/after/at "?" intersecting street. How the officer choose the intersecting street is arbitrary. So, on\_street and at\_street is not completely precise and accurate.

5. "Driver" section, circumstance "Inattentative/distracted" has sub codes for communication device.

6. sent the csv with all fields from 2002 to 2014

reorganized the folders. put new csv into tables/wholedataset

to filter out columbia data

inserted a new line in crash.csv, fieldname 1,2,3...49

municipality field 15, columbia code 0610

CTRL C, CTRL V, into google sheet, and a new csv

found data in field 3 really wierd, became something like 2.01E + 12

date and time field were wierd, too. date field 07:00.0 and time 00:00.0

went back to the original csv, fix the data type, made the first line (field names) text, the first two fields text.

questions:

how many records don't have lat and long:

2014 data: 1524 records in total, 32 records don't have lat and lon

how many pedestrian records, circumstances by pedestrians

0324

on my own mac downloaded 2014, 2012, 2010

crash.csv

2014 1524 records, 1492 have lat and lon

2012 1556 records, 1483 have lat and lon

2008

filter by 15: 610, hide all fields, leaving 1, 24, 25. copy paste to lat.lon.xlxs

did so for 2014, 2012, 2008

downloaded all year data onto datalab2

desktop/wholedataset

2013 1562 crashes, filter columbia, leav 1,24,24, add to lat.lon

2011 1586 records, filter columbia, leave 1,24,24, add to lat.lon

2010 1848 records, filter columbia, leave 1,24,24, add to lat.lon

2009 2781 crashes,

2007 3158 crashes,

2006 3463 crashes,

Copied all into googlesheet columbia.lat.lon.06-14

Plotted in cartoDB, don’t know how to cluster analyze

Arcgis: no shapefile

Madi: remote connect to bombadil and use arcgis

Analyzed fatal circumstances on sequal pro localhost and datalib2

circumstance\_year.sql:

SELECT crash\_year, count(DISTINCT image\_no) from crash05\_15

WHERE severity = "fatal"

group by 1

;

SELECT circumstances, count(DISTINCT image\_no) from crash05\_15

WHERE severity = "fatal"

group by 1

ORDER BY 2 DESC

;

SELECT crash\_year, circumstances, count(DISTINCT image\_no) from crash05\_15

WHERE severity = "fatal"

group by 1, 2

Copy paste result to fatal\_circumstances\_year.xlxs, pivot table

Three biggest circumstances: alcohol, speed\_exceed\_limit, none

All decreasing by year.

Flash drive went bad, saved onto mac desktop.

To do:

add field 10.44 leftTheScene, 10.38 DistractionCd

Is there a pattern, are some of them , a particular officer , geo-analysis

Create a new table for those don’t have lat and lon,

CREATE TABLE… AS

SELECT

Data cleaning, create a new field for intersection info formated as revofnizeable by geocoding system: street amdbersent street

bring the new table to geo code just for the new table

Do the geo coding

Look at where the points are, if there is concentrated area, if evenly distritubed across the city, don;t need to worry about including them

Extension of arcGis- Spatial analyst

merge map of lat&lon, and map of those without lat&lon in one shapefile.

VIRGINIA AVE & HITT ST

For mapping, one table each year. Plot one year at a time

Get shapefile from boonecounty road’s file

April 1

Check whether circumstance as all or blank influence the data output

<https://www.mshp.dps.missouri.gov/TR15Map/index.jsp>

April 6

create a database crash\_jasmine in localhost at bombadil

reimport the wholedataset

-copy and paste the column names from Columbia 2015 (for crash, vehicle and person )into previous data in ultraedit (because they don't have fieldnames!!!)

-filter out columbia data in previous years

crash 2015, person 2015,      date YMD,-

crash 2014, MDY, /

record count:

Crashes

2015: 1380

2014: 137431

2013: 139306

2012: 137420

2011: 142973

2010: 151404

2009: 153020

2008: csv: 155890    sql import failed

2007: 166115

2006: 167239

2005: 175151

persons

2014: 348347

2013: 355356

2012: 351811

2011: 364,306

2010: 387,435

2009: 391,603

2008: 394656

2007: 413432

2006: 340,198

2005: 343,619

Vehicles

2015: 2700

2014: import failed Lost connection 93,301 93,300

Csv: 151163

2013: csv 251221

2012: import failed:

[Err] [Row1] [Imp] 2006 - MySQL server has gone away

[Err] [Row1] [Imp] INSERT INTO `vehicles2012`

csv248442

180,623 180,622

2011: csv203655

[Err] [Row1] [Imp] 2006 - MySQL server has gone away

[Err] [Row1] [Imp] INSERT INTO `vehicles2012`

2010: 273918

2009: 261267

2008: 243645

2007: 306384

2006: 306384

2005: 244520

To solve the problem of failed imports:

Google stackoverflow,

Set global max\_allowed\_packet=64\*1024\*1024 and import as vehicles2012-2

Didnt work

Downloaded CSVsplitter, slice the vehicles2012.csv into smaller chunks vehicles2012-000(150000), and vehicles2012-001(9000)

[Err] [Row1] [Imp] 1406 - Data too long for column 'DriverLicenseClass' at row 1

[Err] [Row1] [Imp] INSERT INTO `vehicles2012-001`

Crashes2008 into 000 and 001, 000 importaed correctly, 001 didnt

[Err] [Row1] [Imp] 1406 - Data too long for column 'AtLocStreet' at row 1

[Err] [Row1] [Imp] INSERT INTO `crashes2008-001`

SET @@global.sql\_mode= '';

Then tried vehicles2014 didn’t work

Create table crashes2008\_liz like crashes2008

LOAD DATA LOCAL INFILE 'C:/Users/yhy6f/Desktop/clean/2008\_De-Personalised\_Extract/crashes2008.csv'

INTO TABLE crashes2008\_liz

FIELDS TERMINATED BY ','

ESCAPED BY ''

LINES TERMINATED BY '\r\n'

IGNORE 1 LINES;

DELETE FROM vehicles2014

[SQL]LOAD DATA LOCAL INFILE 'C:/Users/yhy6f/Desktop/clean/2014\_De-Personalised\_Extract/vehicles2014.csv'

INTO TABLE vehicles2014

FIELDS TERMINATED BY ','

ESCAPED BY ''

LINES TERMINATED BY '\r\n'

IGNORE 1 LINES;

Affected rows: 151163

Time: 4.157s

DELETE FROM vehicles2012

LOAD DATA …….

Crashes2014 date field is off. The date field in csv files in clean folder is already messed up

Re downloaded 2014 data, copied and pasted crash table into clean/2014\_depersonlised\_ extract, replaced the old crashes.csv,

Loaded 2014 crashes using the same.

Loaded all data into crashes, persons, vehicles

In the raw data that downloaded from loretta’s link, vehicles2007 is the same as vehicles2006

Skipped 2007, need to load vehicles 2007.

To remoce “ï»¿” in personid in persons:

update persons

set personid = replace(personid, 'ï»¿', '');

The changeing of code could be because I opened the file in mac before. LESSON learned.

To find out crashes involving device distractions

SELECT persons.crashid, crashes.municipality

FROM persons INNER JOIN crashes

ON persons.crashid=crashes.crashid

Taking forever toload, froze everytime

04082016

SELECT count(\*)

FROM crashes

WHERE crashes.crashid IN

(SELECT crashid FROM persons)

1640482

Diff from count(\*) in crashes

SELECT \*

from persons

INNER JOIN crashes

ON persons.crashid = crashes.crashid

where crashes.Municipality="0610"

Worked! Took forever tho

CREATE TABLE 610person

AS (SELECT \*

from persons

INNER JOIN crashes

ON persons.crashid = crashes.crashid

where crashes.Municipality="0610")

[Err] 1060 - Duplicate column name 'CrashId'

Liz:

SELECT person.\*

Vs

SELECT \*

SELECT contribcircCD1

FROM persons

INNER JOIN crashes

ON persons.crashid = crashes.crashid

WHERE crashes.municipality = "0610"

Found that CPD changed policy regarding how crashes are reported

If both cars can leave the scene, no one needs medical treatment, no legal violations, parties will be asked to file report online themselves.

Self report webpage:

<https://www.gocolumbiamo.com/CMS/WebForms/form.php?formid=17>

041116

Contaced IT-self-call about the webpage, hasn’t got back to me:

573-874-7500

Liz helped fix query create\_610person query

SELECT person.\*

Vs

SELECT \*

CREATE TABLE 610person

AS (SELECT persons.\*

   from persons

   INNER JOIN crashes

   ON persons.crashid = crashes.crashid

   where crashes.Municipality="0610")

65867 records in 610person

SELECT count(\*)

FROM 610crash

INNER JOIN 610person

on 610crash.crashid = 610person.CrashId

83135

no\_of\_610person.crashid\_NOT\_IN\_610crash.CrashId

SELECT count(610person.crashid)

FROM 610person

WHERE 610person.crashid NOT IN (SELECT 610crash.CrashId FROM 610crash)

Output: 0

no\_of\_610person.crashid\_NOT\_IN\_610crash.CrashId

SELECT count(610crash.crashid)

FROM 610crash

WHERE 610crash.crashid NOT IN (SELECT 610person.CrashId FROM 610person)

Output: 3462

SELECT count(crashid), crashyear

FROM 610crash

GROUP BY 2

ORDER BY crashYear

3469 2005

3462 2006

3158 2007

2906 2008

2780 2009

1847 2010

3170 2011

3112 2012

1561 2013

Check back in Sequel Pro localhost crash database (which was imported from spreadsheets downloaded online)

SELECT count(DISTINCT ImageNo), crash\_year

FROM crash05\_15

GROUP BY 2

ORDER BY crash\_year

3448 2005

3461 2006

3158 2007

2903 2008

2779 2009

1847 2010

1584 2011

1556 2012

1561 2013

1524 2014

1381 2015

Something wrong with 2012 and 2013. Add distinct in bombadil navicat query

SELECT count(DISTINCT crashid), crashyear

FROM 610crash

GROUP BY 2

ORDER BY crashYear

3469 2005

3462 2006

3158 2007

2906 2008

2780 2009

1847 2010

1585 2011

1556 2012

1561 2013

1524 2014

1380 2015

Remove duplicates from 610crash

CREATE TABLE 610crash\_liz

AS (SELECT DISTINCT \*

   FROM 610crash);

DROP TABLE 610crash

RENAME TABLE 610crash\_liz TO 610crash

CREATE TABLE 610person\_liz

AS (SELECT DISTINCT \*

   FROM 610person);

DROP TABLE 610person

RENAME TABLE 610person\_liz TO 610person

CREATE TABLE crashes\_NoDup

AS (SELECT DISTINCT \*

   FROM crashes);

DROP TABLE crashes;

RENAME TABLE crashes\_NoDup TO crashes;

CREATE TABLE persons\_NoDup

AS (SELECT DISTINCT \*

   FROM persons);

DROP TABLE persons;

RENAME TABLE persons\_NoDup TO persons;

SELECT count(crashid), count(DISTINCT crashid), CRASHYEAR

FROM 610crash

GROUP BY 3

##What’s in 610crash but not in 610person?

SELECT 610crash.crashid

FROM 610crash

WHERE 610crash.crashid NOT IN (SELECT 610person.CrashId

                              FROM 610person)

Output:

All 2006 crashid. Probably didn’t load 2006 person

Check back persons, lack persons2006 data.

LOAD DATA LOCAL INFILE 'C:/Users/yhy6f/Desktop/clean/2006\_De-Personalised\_Extract/persons2006.csv'

INTO TABLE persons

FIELDS TERMINATED BY ','

ESCAPED BY ''

LINES TERMINATED BY '\r\n'

IGNORE 1 LINES

Output: affected 0 rows

INSERT INTO persons

SELECT \* FROM persons2006

Output:

Affected rows: 340198

Time: 5.331s

No of records in persons;

3694509

Re-create table610person

drop table 610person

CREATE TABLE 610person

AS (SELECT persons.\*

   from persons

   INNER JOIN crashes

   ON persons.crashid = crashes.crashid

   where crashes.Municipality="0610")

Reran the remove duplicate query:

CREATE TABLE 610person\_liz

AS (SELECT DISTINCT \*

   FROM 610person);

DROP TABLE 610person;

RENAME TABLE 610person\_liz INTO 610person;

Reran “610crash.crashid\_NOT\_IN\_610person.Cashid”

SELECT 610crash.crashid

FROM 610crash

WHERE 610crash.crashid NOT IN (SELECT 610person.CrashId

                              FROM 610person)

Same output 3456 results

Search for one of the crashid 0060142574

Not in persons, not in persons2016

Search in csv file, its 60142574 insread of 0060142574.

So crashes2016 and persons2016 are not in the same format. IN other tables crashid are in 006 format.

Cleaner.py (Didn’t work.)

import csv

csvfile = open('./2006\_De-Personalised\_Extract/persons2006.csv', 'r')

outfile = open('./2006\_De-Personalised\_Extract/persons2006\_zfilled.csv', 'w')

# Now a DictReader and DictWriter

# DictReader and DictWriter are imported libraries

reader = csv.DictReader(csvfile)

writer = csv.DictWriter(outfile, reader.fieldnames)

# DictWriter writes to outfile

#reader.fieldname refers to the headers

# Write headers

writer.writeheader()

# Clean and write the data to output

for row in reader:

   row["CrashId"]=row["Crash"].zfill(11)

   writer.writerow(row)

041216

##make crashid the same digits:

Lpad\_crashid\_persons:

select crashid, lpad(crashid, 10, '0')

from persons

where length(crashid) = 8

##Reran the query create\_610person:

[SQL]drop table 610person;

Affected rows: 0

Time: 0.011s

[SQL]

CREATE TABLE 610person

AS (SELECT persons.\*

   from persons

   INNER JOIN crashes

   ON persons.crashid = crashes.crashid

   where crashes.Municipality="0610");

Affected rows: 57247

Time: 5365.647s

## reran 610crash.crashid\_NOT\_IN\_610person.Cashid

Same output 3456 results

This means select lpad doesn’t change the column, only returns the padded column

##Make a copy of persons

##create\_persons\_copy

CREATE TABLE persons\_copy LIKE persons;

INSERT INTO persons\_copy

SELECT \* FROM persons

##Use lpad to autofill those crashid with only 8 digits

##lpad\_crashid\_persons\_1

UPDATE persons SET crashid = LPAD(crashid, 10, '0')

##create\_610person

[SQL]drop table 610person;

Affected rows: 0

Time: 0.006s

[SQL]

CREATE TABLE 610person

AS (SELECT persons.\*

   from persons

   INNER JOIN crashes

   ON persons.crashid = crashes.crashid

   where crashes.Municipality="0610");

Affected rows: 64442

Time: 5386.044s

#### reran 610crash.crashid\_NOT\_IN\_610person.Cashid

Zero results

##no\_of\_crashes\_by\_year\_610person

SELECT count(DISTINCT 610person.CrashId), 610crash.crashYear

FROM 610crash, 610person

WHERE 610crash.crashid = 610person.CrashId

GROUP BY 2

3469 2005

3462 2006

3158 2007

2906 2008

2780 2009

1847 2010

1585 2011

1556 2012

1561 2013

1524 2014

1380 2015

##no\_of\_crashes\_by\_year\_610person\_injured\_fatal

SELECT count(DISTINCT 610person.CrashId), 610crash.crashYear

FROM 610crash, 610person

WHERE 610crash.crashid = 610person.CrashId

AND 610person.PersonalInjuryLevel NOT IN ("5","U")

GROUP BY 2

826 2005

873 2006

830 2007

763 2008

841 2009

745 2010

634 2011

632 2012

551 2013

545 2014

537 2015

##no\_of\_persons\_by\_year\_610person

SELECT count(DISTINCT 610person.personId), 610crash.crashYear

FROM 610crash, 610person

WHERE 610crash.crashid = 610person.CrashId

GROUP BY 2

7090 2005

7209 2006

8519 2007

8058 2008

7900 2009

4997 2010

4318 2011

4316 2012

4159 2013

4131 2014

3745 2015

##no\_of\_persons\_by\_year\_610person\_injured\_fatal

SELECT count(DISTINCT 610person.personId), 610crash.crashYear

FROM 610crash, 610person

WHERE 610crash.crashid = 610person.CrashId

AND 610person.PersonalInjuryLevel NOT IN ("5","U")

GROUP BY 2

1151 2005

1202 2006

1155 2007

1071 2008

1163 2009

1059 2010

896 2011

861 2012

765 2013

757 2014

741 2015

##comm\_device\_by\_year\_610person

SELECT count(DISTINCT 610person.CrashId), 610crash.crashYear

FROM 610crash, 610person

WHERE 610crash.crashid = 610person.CrashId

AND (DistractionCd1 IN ("5","6","7","8") OR

    DistractionCd2 IN ("5","6","7","8") OR

    DistractionCd3 IN ("5","6","7","8") OR

    DistractionCd4 IN ("5","6","7","8"))

GROUP BY 2

18 2005

31 2006

23 2007

42 2008

46 2009

24 2010

24 2011

22 2012

27 2013

21 2014

32 2015

##CREATE\_610person\_injured

CREATE TABLE 610person\_injured

AS (SELECT \*

FROM 610person

WHERE PersonalInjuryLevel NOT IN ("5","U"))

##comm\_device\_by\_year\_610person\_injured

SELECT count(DISTINCT 610person\_injured.CrashId), 610crash.crashYear

FROM 610crash, 610person\_injured

WHERE 610crash.crashid = 610person\_injured.CrashId

AND (DistractionCd1 IN ("5","6","7","8") OR

    DistractionCd2 IN ("5","6","7","8") OR

    DistractionCd3 IN ("5","6","7","8") OR

    DistractionCd4 IN ("5","6","7","8"))

GROUP BY 2

4 2005

4 2006

3 2007

4 2008

6 2009

2 2010

5 2011

6 2012

5 2013

5 2014

8 2015

No of 911 calls:

https://www.como.gov/PSJC/About\_Us/Contact/index.php?font=75

Public Safety Joint Communications Center

573-874-7400

Custodian of records: Katie

Called Lydia about:

No of self-reported crashes

Document of the accident policy in November 2009

##injuries\_by\_year

SELECT count(DISTINCT 610person.PersonId),610crash.crashYear

FROM 610crash, 610person

WHERE 610crash.crashid = 610person.CrashId

AND 610person.PersonalInjuryLevel IN (2,3,4)

GROUP BY 2

893 2005

970 2006

929 2007

848 2008

979 2009

896 2010

757 2011

733 2012

662 2013

623 2014

652 2015

##fatalities\_by\_year

SELECT count(DISTINCT 610person.PersonId),610crash.crashYear

FROM 610crash, 610person

WHERE 610crash.crashid = 610person.CrashId

AND 610person.PersonalInjuryLevel = "1"

GROUP BY 2

9 2005

15 2006

8 2007

8 2008

10 2009

2 2010

5 2011

5 2012

6 2013

9 2014

8 2015

##crashes\_involving\_injuries\_fatalities\_by\_year

SELECT count(DISTINCT 610person.CrashId),610crash.crashYear

FROM 610crash, 610person

WHERE 610crash.crashid = 610person.CrashId

AND 610person.PersonalInjuryLevel NOT IN ("5","U")

GROUP BY 2

826 2005

873 2006

830 2007

763 2008

841 2009

745 2010

634 2011

632 2012

551 2013

545 2014

537 2015

##Dumped no of crashes, injuries, fatalities, and crashes involving injuries and fatalities into google sheet:

https://docs.google.com/spreadsheets/d/13HKK4z52mqsrXufP8PmkA2mpCakbBF6z1ecKmHypvAw/edit?usp=sharing

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| year | crashes | injuries | fatalities | crashes\_involving\_injuries\_fatalities |
| 2005 | 3469 | 893 | 9 | 826 |
| 2006 | 3462 | 970 | 15 | 873 |
| 2007 | 3158 | 929 | 8 | 830 |
| 2008 | 2906 | 848 | 8 | 763 |
| 2009 | 2780 | 979 | 10 | 841 |
| 2010 | 1847 | 896 | 2 | 745 |
| 2011 | 1585 | 757 | 5 | 634 |
| 2012 | 1556 | 733 | 5 | 632 |
| 2013 | 1561 | 662 | 6 | 551 |
| 2014 | 1524 | 623 | 9 | 545 |
| 2015 | 1380 | 652 | 8 | 537 |

Pulled some numbers from MSHP website online mapping system:

|  |  |  |  |
| --- | --- | --- | --- |
| year | crashes | injuries | fatalities |
| 2005 | 3469 | 893 | 9 |
| 2006 | 3462 | 970 | 15 |
| 2007 | 3158 | 929 | 8 |
| 2008 | 2906 | 848 | 8 |
| 2009 | 2780 | 979 | 10 |
| 2010 | 1847 | 896 | 2 |
| 2011 | 1585 | 757 | 5 |
| 2012 | 1556 | 733 | 5 |
| 2013 | 1561 | 662 | 6 |
| 2014 | 1524 | 623 | 9 |
| 2015 | 1382 | 651 | 8 |

Something wrong with no of crashes and injuries in 2015. Might be because MSHP updated the system

##went back to sequel pro database crash table crash05\_15, which was imported from spreadsheets downloaded online in march

##no\_of\_crashes\_by\_year

Saved in crash/0413

SELECT count(DISTINCT imageno), crash\_year

FROM crash05\_15

GROUP BY 2

|  |  |
| --- | --- |
| 2005 | 3448 |
| 2006 | 3461 |
| 2007 | 3158 |
| 2008 | 2903 |
| 2009 | 2779 |
| 2010 | 1847 |
| 2011 | 1584 |
| 2012 | 1556 |
| 2013 | 1561 |
| 2014 | 1524 |
| 2015 | 1381 |

##no\_of\_injuries\_by\_imageno

SELECT imageno, (sum(injured)/count(imageno)) AS no\_of\_injuries

FROM crash05\_15

WHERE injured > 0

GROUP BY 1

QUESTION:

Haven’t been able to generate no of injuries and fatalities per year

##

SELECT \* FROM crash05\_15

WHERE crash\_Year = 2015

##Export into CSV file, named as “MSHP\_2015”, saved into Crash/clean

##Import MSHP\_2015 into bombadil.crash\_jasmine.mshp\_2015

##mshp\_2015.imageno\_NOT\_IN\_crashes2015.crashid

SELECT mshp\_2015.ImageNo FROM `mshp\_2015`

WHERE mshp\_2015.ImageNo NOT IN

(SELECT crashes2015.CrashId FROM crashes2015)

Output：

2150005196

So crashid 2150005196 is not in my database, but is in the online mapping system

Question:

Should I include it?

If I do, it’s hard to get this record in line with others because this one is from spreadsheet downloaded online, doesn’t have as many fields. Lack lat long, and distraction. Not going to analyze these.

If not, does it affect the accuracy?

But the 2015 data is not complete anyway.

##search “2150005196” in the raw file “crash” in the 2015 zip, returned no result. It’s lacking in the dataset that Loretta sent.

## crashes2015.crashid\_NOT\_IN\_mshp\_2015.imageno

SELECT crashes2015.CrashId FROM crashes2015

WHERE crashes2015.CrashId NOT IN

(SELECT mshp\_2015.ImageNo FROM mshp\_2015)

Output： none

##circum\_alcohol\_by\_year

SELECT count(DISTINCT 610person.crashid), crashYear

FROM 610person, 610crash

WHERE 610person.crashid = 610crash.CrashId

AND (ContribCircCd1 = "18"OR

ContribCircCd2 = "18"OR

ContribCircCd3 = "18"OR

ContribCircCd4 = "18"OR

ContribCircCd5 = "18"OR

ContribCircCd6 = "18"OR

ContribCircCd7 = "18"OR

ContribCircCd8 = "18"OR

ContribCircCd9 = "18"OR

ContribCircCd10 = "18"OR

ContribCircCd11 = "18"OR

ContribCircCd12 = "18")

GROUP BY 2

Output

169 2005

188 2006

179 2007

135 2008

142 2009

132 2010

105 2011

111 2012

104 2013

113 2014

82 2015

##circum\_alcohol\_by\_year\_injured\_fatal

SELECT count(DISTINCT 610person.crashid), crashYear

FROM 610person, 610crash

WHERE 610person.crashid = 610crash.CrashId

AND 610person.PersonalInjuryLevel NOT IN ("5", "U")

AND (ContribCircCd1 = "18"OR

ContribCircCd2 = "18"OR

ContribCircCd3 = "18"OR

ContribCircCd4 = "18"OR

ContribCircCd5 = "18"OR

ContribCircCd6 = "18"OR

ContribCircCd7 = "18"OR

ContribCircCd8 = "18"OR

ContribCircCd9 = "18"OR

ContribCircCd10 = "18"OR

ContribCircCd11 = "18"OR

ContribCircCd12 = "18")

GROUP BY 2

52 2005

52 2006

43 2007

39 2008

35 2009

34 2010

19 2011

33 2012

24 2013

21 2014

18 2015

##circum\_drug\_by\_year

SELECT count(DISTINCT 610person.crashid), crashYear

FROM 610person, 610crash

WHERE 610person.crashid = 610crash.CrashId

AND (ContribCircCd1 = "19"OR

ContribCircCd2 = "19"OR

ContribCircCd3 = "19"OR

ContribCircCd4 = "19"OR

ContribCircCd5 = "19"OR

ContribCircCd6 = "19"OR

ContribCircCd7 = "19"OR

ContribCircCd8 = "19"OR

ContribCircCd9 = "19"OR

ContribCircCd10 = "19"OR

ContribCircCd11 = "19"OR

ContribCircCd12 = "19")

GROUP BY 2

Output

25 2005

31 2006

19 2007

21 2008

22 2009

24 2010

14 2011

14 2012

9 2013

17 2014

13 2015

##circum\_drug\_by\_year\_injured\_fatal

SELECT count(DISTINCT 610person.crashid), crashYear

FROM 610person, 610crash

WHERE 610person.crashid = 610crash.CrashId

AND 610person.PersonalInjuryLevel NOT IN ("5", "U")

AND (ContribCircCd1 = "19"OR

ContribCircCd2 = "19"OR

ContribCircCd3 = "19"OR

ContribCircCd4 = "19"OR

ContribCircCd5 = "19"OR

ContribCircCd6 = "19"OR

ContribCircCd7 = "19"OR

ContribCircCd8 = "19"OR

ContribCircCd9 = "19"OR

ContribCircCd10 = "19"OR

ContribCircCd11 = "19"OR

ContribCircCd12 = "19")

GROUP BY 2

15 2005

12 2006

8 2007

11 2008

8 2009

11 2010

7 2011

7 2012

3 2013

7 2014

7 2015

##crashtype\_no\_of

SELECT crashType, count(CrashId)

FROM 610crash

GROUP BY 1

ORDER BY 2 DESC

07(MV in transport) 19384

03(Fixed object) 3027

09(parked vehicle) 1338

05(pedestrian) 338

10(overturning) 322

01(animal) 285

02(pedalcycle) 248

04 136

11 100

13 15

16 14

06 6

15 5

17 5

18 3

14 2

##crashtype\_fatal

SELECT 610crash.crashType, count(DISTINCT 610crash.crashid)

FROM 610crash, 610person\_fatal

WHERE 610crash.crashid = 610person\_fatal.CrashId

GROUP BY 1

ORDER BY 2 DESC

03(fixed object) 40

07(MV in transport) 25

05(pedestrian) 10

10(overturning) 6

02(pedalcycle) 2

11(other non collision) 1

01(animal) 1

##crashtype\_fatal\_by\_year

SELECT count(610crash.crashid), 610crash.crashType, 610crash.crashYear

FROM 610crash, 610person\_fatal

WHERE 610crash.crashid = 610person\_fatal.CrashId

GROUP BY 2, 3

ORDER BY 1 DESC

7 03 2006

6 03 2009

5 07 2006

5 03 2008

5 03 2013

4 07 2005

3 03 2005

3 03 2007

3 07 2007

3 07 2009

3 03 2015

2 05 2006

2 10 2012

2 05 2011

2 05 2014

2 07 2012

2 07 2014

2 07 2015

2 03 2014

1 01 2008

1 05 2009

1 10 2014

1 02 2007

1 11 2005

1 02 2013

1 05 2015

1 07 2008

1 03 2010

1 03 2011

1 03 2012

1 10 2007

1 10 2008

1 05 2005

1 10 2010

##no\_of\_personinvolvement\_injured\_fatal

SELECT personinvolvementcode, count(personid)

FROM 610person\_injured\_fatal

GROUP BY 1

ORDER BY 2 DESC

1driver 7710

2passenger 2586

3pedestrian 329

4pedalcyclist 196

##no\_of\_personinvolvement\_injured

SELECT personinvolvementcode, count(distinct personid)

FROM 610person

WHERE personalinjurylevel IN (2,3,4)

GROUP BY 1

1driver 5859

2passenger 2570

3pedestrian 319

4pedalcyclist 194

##no\_of\_personinvolvement\_fatal

SELECT personinvolvementcode, count(personid)

FROM 610person\_fatal

GROUP BY 1

ORDER BY 2 DESC

1(driver) 57

2(passenger) 16

3(pedestrian) 10

4(pedalcyclist) 2

##no\_of\_pedestrians

SELECT count(personid), personinvolvementcode

FROM 610person

WHERE personinvolvementcode = 3;

Output: 383

##no\_of\_pedestrians\_injured\_fatal

SELECT count(personid), personinvolvementcode

FROM 610person\_injured\_fatal

WHERE personinvolvementcode = 3

Output: 329

##no\_of\_pedestrians\_fatal

SELECT count(personid), personinvolvementcode

FROM 610person\_fatal

WHERE personinvolvementcode = 3

Output: 10

##no\_of\_persons\_by\_year\_injured\_fatal

SELECT count(DISTINCT 610person.personId), 610crash.crashYear

FROM 610crash, 610person

WHERE 610crash.crashid = 610person.CrashId

AND 610person.PersonalInjuryLevel NOT IN ("5","U")

GROUP BY 2

year no of persons injured or fatal

2005 1151

2006 1202

2007 1155

2008 1071

2009 1163

2010 1059

2011 896

2012 861

2013 765

2014 757

2015 741

question mark on this query: not in line with the added number of injuries and fatalities. Didn’t use the result

041416

##CREATE TABLE 610crash\_without\_lat\_long

CREATE TABLE exonerations = agate.Table.from\_csv('exonerations-20150828.csv')

AS (SELECT crashid, OnLocStreet, AtLocStreet

FROM 610crash

WHERE LandedLongitude = 0)

export the table into excel for cleaning and geocode

Create table 0414 put table 610crash\_without\_lat\_long into 0414

Mapping

<http://support.esri.com/en/knowledgebase/techarticles/detail/44610>

Clean data:

create new field for crashes without lat long

format:

Geocode

batchgeo

04182016

Meeting with Mike,decided on elements in infographic:

1. Map
2. Injuries
3. Fatalities
4. Top 5 circumstances
5. the number of STARS reported, self-reported crashes and 911 calls regarding collision

For element 4)

The previous analysis (circumstance.xlsx in folder 0322) based on spreadsheets downloaded from online mapping system showed that the top 5 known circumstances are:

None 20642

**Distracted/Inattentive 7624**

**Failed to Yield 5134**

**Following Too Close 4079**

**Too Fast for Conditions 2031**

**Improper Lane Usage/Change 1704**

To check in the newly-imported crash database in bombadil:

**##distracted\_no\_of\_crashes:**

SELECT count(DISTINCT 610person.crashid)

FROM 610person, 610crash

WHERE 610person.crashid = 610crash.CrashId

AND (ContribCircCd1 = "21"OR

ContribCircCd2 = "21"OR

ContribCircCd3 = "21"OR

ContribCircCd4 = "21"OR

ContribCircCd5 = "21"OR

ContribCircCd6 = "21"OR

ContribCircCd7 = "21"OR

ContribCircCd8 = "21"OR

ContribCircCd9 = "21"OR

ContribCircCd10 = "21"OR

ContribCircCd11 = "21"OR

ContribCircCd12 = "21")

Output: **7625**

**##failed\_to\_yield\_no\_of\_crashes:**

SELECT count(DISTINCT 610person.crashid)

FROM 610person, 610crash

WHERE 610person.crashid = 610crash.CrashId

AND (ContribCircCd1 = "17"OR

ContribCircCd2 = "17"OR

ContribCircCd3 = "17"OR

ContribCircCd4 = "17"OR

ContribCircCd5 = "17"OR

ContribCircCd6 = "17"OR

ContribCircCd7 = "17"OR

ContribCircCd8 = "17"OR

ContribCircCd9 = "17"OR

ContribCircCd10 = "17"OR

ContribCircCd11 = "17"OR

ContribCircCd12 = "17")

Output: **5135**

**##following\_too\_close\_no\_of\_crashes:**

SELECT count(DISTINCT 610person.crashid)

FROM 610person, 610crash

WHERE 610person.crashid = 610crash.CrashId

AND (ContribCircCd1 = "9"OR

ContribCircCd2 = "9"OR

ContribCircCd3 = "9"OR

ContribCircCd4 = "9"OR

ContribCircCd5 = "9"OR

ContribCircCd6 = "9"OR

ContribCircCd7 = "9"OR

ContribCircCd8 = "9"OR

ContribCircCd9 = "9"OR

ContribCircCd10 = "9"OR

ContribCircCd11 = "9"OR

ContribCircCd12 = "9")

Output: **4079**

**##too\_fast\_for\_conditions\_no\_of\_crashes:**

SELECT count(DISTINCT 610person.crashid)

FROM 610person, 610crash

WHERE 610person.crashid = 610crash.CrashId

AND (ContribCircCd1 = "5"OR

ContribCircCd2 = "5"OR

ContribCircCd3 = "5"OR

ContribCircCd4 = "5"OR

ContribCircCd5 = "5"OR

ContribCircCd6 = "5"OR

ContribCircCd7 = "5"OR

ContribCircCd8 = "5"OR

ContribCircCd9 = "5"OR

ContribCircCd10 = "5"OR

ContribCircCd11 = "5"OR

ContribCircCd12 = "5")

Output: **2031**

**##improper\_lane\_usage\_no\_of\_crashes:**

SELECT count(DISTINCT 610person.crashid)

FROM 610person, 610crash

WHERE 610person.crashid = 610crash.CrashId

AND (ContribCircCd1 = "13"OR

ContribCircCd2 = "13"OR

ContribCircCd3 = "13"OR

ContribCircCd4 = "13"OR

ContribCircCd5 = "13"OR

ContribCircCd6 = "13"OR

ContribCircCd7 = "13"OR

ContribCircCd8 = "13"OR

ContribCircCd9 = "13"OR

ContribCircCd10 = "13"OR

ContribCircCd11 = "13"OR

ContribCircCd12 = "13")

Output:**1704**

Dumped numbers into google sheet “top 5 known circumstances”

Got data of 911 calls regarding collision from joint comunication

Uploaded the excel spreadsheet “collision report jasmine 2005-2015”  in google drive.

Dumped the no into google sheet “no of 911calls regarding collisions” <https://docs.google.com/spreadsheets/d/1-au0SdH6SSJMuPnwaSQxm9hvxktKT9WMeq9wI0NipQ0/edit#gid=0>

Make element 5) out of this chart

|  |  |  |  |
| --- | --- | --- | --- |
| year | no of 911 calls regarding collisions | no of STARS report | no of self-report |
| 2005 | 6095 | 3469 |  |
| 2006 | 6009 | 3462 |  |
| 2007 | 6194 | 3158 |  |
| 2008 | 5659 | 2906 |  |
| 2009 | 5364 | 2780 |  |
| 2010 | 5738 | 1847 |  |
| 2011 | 5492 | 1585 |  |
| 2012 | 5401 | 1556 |  |
| 2013 | 5613 | 1561 |  |
| 2014 | 5838 | 1524 |  |
| 2015 | 5887 | 1380 |  |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

0420

Created folder 0414, put file 610crash\_without\_lat\_long into 0414

Added address, added city, state, country into

Saved 610crash\_without\_lat\_long\_tested.csv into 0414

Tested CST, CRD, on batchgeo, need to remove CST

needs MO

Wrote geoclean.py, added city, state, country into Lat\_long.csv

IS:

Search “,IS”

50060050,CST RANGELINE,IS 70 E,RANGELINE & IS 70 E,Columbia,Missouri,United State, didn’t work

50060050,CST RANGELINE,IS 70 E,RANGELINE St & IS 70 E,Columbia,Missouri,United State,

Didn’t work

50060050,CST RANGELINE,IS 70 E,RANGELINE St & IS 70,Columbia,Missouri,United State, didn’t work

50060050,CST RANGELINE,IS 70 E,RANGELINE St & Interstate 70,Columbia,Missouri,United State, didn’t work

50060050,CST RANGELINE,IS 70 E,RANGELINE St & I-70,Columbia,Missouri,United State, didn’t work

50060050,CST RANGELINE,IS 70 E,RANGELINE St & Interstate 70 Business Loop,Columbia,Missouri,United State, didn’t work

0421

Questions:

Are there any patterns in fatal crashes?

Circumstances?

Crashtype?

Fixed object: drunk? Drug? Circumstance?

Correlation between crashtype and circumstances?

SELECT 610crash.crashid,610crash.OnLocStreet, 610crash.AtLocStreet

FROM 610crash,610person\_fatal

WHERE 610crash.CrashId=610person\_fatal.crashid

AND Crashtype = "03"

Found repetitive crashid

So there is problem with query ##crashtype\_fatal

SELECT 610crash.crashType, count(DISTINCT 610crash.crashid)

FROM 610crash, 610person\_fatal

WHERE 610crash.crashid = 610person\_fatal.CrashId

GROUP BY 1

ORDER BY 2 DESC

03 37

07 22

05 9

10 6

02 2

01 1

11 1

Made changes in 0413.xlsx

Made stacked bar chart of fatal crashtypes in highchart

|  |
| --- |
| <http://cloud.highcharts.com/show/abydaw>  <iframe class='highcharts-iframe' src='//cloud.highcharts.com/embed/abydaw' style='border: 0; width: 100%; height: 500px'></iframe> |

##fatal\_crashtype=03\_OnStreet\_AtStreet

SELECT 610crash.crashid,610crash.OnLocStreet, 610crash.AtLocStreet

FROM 610crash,610person\_fatal

WHERE 610crash.CrashId=610person\_fatal.crashid

AND Crashtype = "03"

##fatal\_crashtype=03\_circumstances

SELECT 610crash.crashid,610person\_fatal.ContribCircCd1,610person\_fatal.ContribCircCd2, 610person\_fatal.ContribCircCd3,610person\_fatal.ContribCircCd4, 610person\_fatal.ContribCircCd5

FROM 610crash,610person\_fatal

WHERE 610crash.CrashId=610person\_fatal.crashid

AND Crashtype = "03"

Found a lot of 18 (alcohol) and 13(improper lane usage/change)  in contributing cirumstances

##circum\_longer\_than\_5

SELECT \* FROM 610person

WHERE ContribCircCd5 IS NOT NULL

AND ContribCircCd5 != ""

The most contribcircd of all crashes are 9.

##fatal\_crashtype=3\_alcohol

SELECT count(distinct 610crash.crashid)

FROM 610crash,610person\_fatal

WHERE 610crash.Crashid=610person\_fatal.crashid

AND Crashtype = "03"

AND (ContribCircCd1 = "18"OR

ContribCircCd2 = "18"OR

ContribCircCd3 = "18"OR

ContribCircCd4 = "18"OR

ContribCircCd5 = "18"OR

ContribCircCd6 = "18"OR

ContribCircCd7 = "18"OR

ContribCircCd8 = "18"OR

ContribCircCd9 = "18"OR

ContribCircCd10 = "18"OR

ContribCircCd11 = "18"OR

ContribCircCd12 = "18")

Output: 16

##fatal\_crashtype=3\_lane

SELECT count(distinct 610crash.crashid)

FROM 610crash,610person\_fatal

WHERE 610crash.Crashid=610person\_fatal.crashid

AND Crashtype = "03"

AND (ContribCircCd1 = "13"OR

ContribCircCd2 = "13"OR

ContribCircCd3 = "13"OR

ContribCircCd4 = "13"OR

ContribCircCd5 = "13"OR

ContribCircCd6 = "13"OR

ContribCircCd7 = "13"OR

ContribCircCd8 = "13"OR

ContribCircCd9 = "13"OR

ContribCircCd10 = "13"OR

ContribCircCd11 = "13"OR

ContribCircCd12 = "13")

Output: 15

Data-lib2.crash

##fixted\_object\_circumstances

SELECT count(DISTINCT IMAGE\_NO), circumstances

FROM crash05\_15

WHERE Crash\_Type = "Fixed Object"

GROUP BY 2

ORDER BY 1 DESC

669 Too Fast for Conditions

645 Alcohol

632 Distracted/Inattentive

532 None

292 Speed Exceeded Limit

269 Improper Lane Usage/Change

138 Physical Impairment

112 Vehicle Defects

109 Improper Turn

98 Drugs

98 Unknown

81 Overcorrected

76 Other

43 Improper Backing

42 Driver Fatigue/Asleep

28 Violation Signal/Sign

26 Following Too Close

25 Wrong Side (Not Passing)

18 Vision Obstructed

17 Improper Passing

15 Animal(s) in Roadway

14 Failed to Yield

7 Object/Obstruction in Roadway

5 Wrong Way (One-way)

3 Failed to Secure Load/Improper Loading

3 Improper Start from Park

2 Improperly Stopped on Roadway

2 Failed to Use Lights

1 Improperly Parked

1 Improper Riding/Clinging to Vehicle Exterior

1 Improper Signal

1 Improper Towing/Pushing

##MV\_in\_transport\_circumstances

SELECT count(DISTINCT IMAGE\_NO), circumstances

FROM crash05\_15

WHERE Crash\_Type = "Motor Vehicle in Transport"

GROUP BY 2

ORDER BY 1 DESC

17974 None

6269 Distracted/Inattentive

4856 Failed to Yield

4034 Following Too Close

1256 Improper Lane Usage/Change

1175 Violation Signal/Sign

1156 Too Fast for Conditions

712 Improper Turn

595 Unknown

547 Alcohol

419 Improper Backing

305 Improper Passing

236 Vehicle Defects

197 Speed Exceeded Limit

155 Wrong Side (Not Passing)

154 Vision Obstructed

142 Other

113 Improperly Stopped on Roadway

97 Physical Impairment

81 Drugs

55 Improper Start from Park

42 Improper Signal

33 Wrong Way (One-way)

30 Object/Obstruction in Roadway

29 Improperly Parked

22 Driver Fatigue/Asleep

22 Overcorrected

9 Failed to Secure Load/Improper Loading

6 Failed to Use Lights

2 Animal(s) in Roadway

1 Improper Riding/Clinging to Vehicle Exterior

1 Improper Towing/Pushing

1 Failed to Dim Lights

##cir\_count(\*)

SELECT ContribCircCd12, COUNT(\*)

FROM 610person\_fatal

GROUP BY 1

ORDER BY 1

##fatal\_crashtype=03\_speed

SELECT count(distinct 610crash.crashid)

FROM 610crash,610person\_fatal

WHERE 610crash.Crashid=610person\_fatal.crashid

AND Crashtype = "03"

AND (ContribCircCd1 = "4"OR

ContribCircCd2 = "4"OR

ContribCircCd3 = "4"OR

ContribCircCd4 = "4"OR

ContribCircCd5 = "4")

Output: 22

##CREATE TABLE circumstances\_fatal

CREATE TABLE circumstances\_fatal

(

crashid INT NOT NULL,

ContribCircCd INT  NOT NULL,

ColNum INT NOT NULL

);

-- this would be the first insert

INSERT INTO circumstances\_fatal (crashid, ContribCircCd, ColNum)

SELECT crashid, ContribCircCd1 as ContribCircCd, 1 as ColNum

FROM 610person\_fatal

WHERE ContribCircCd1 is not NULL;

[SQL]

-- this would be the first insert

INSERT INTO circumstances\_fatal (crashid, ContribCircCd, ColNum)

SELECT crashid, ContribCircCd1 as ContribCircCd, 1 as ColNum

FROM 610person\_fatal

WHERE ContribCircCd1 is not NULL;

[Err] 1366 - Incorrect integer value: '' for column 'ContribCircCd' at row 4

##Changed ContribCircCd INT  NOT NULL,into varchar(50) NOT NULL

[Err] 1264 - Out of range value for column 'crashid' at row 79

##Changed crashid INT  NOT NULL,into crashid VARCHAR(50) NOT NULL,

Worked!

But no of crashes in the new table is 76, as against 78 in 610person\_fatal

##fatal\_cir\_no\_of\_crashes

SELECT ContribCircCd,count(DISTINCT crashid)

from circumstances\_fatal

GROUP BY 1

ORDER BY 2 DESC

37

4 24

13 19

18 18

17 8

22 8

5 5

20 4

14 3

21 2

U 2

19 2

9 2

3 1

1 1

26 1

28 1

##crashtype03\_no\_of

SELECT count(\*)

FROM 610crash

WHERE Crashtype = "03"

##fatal\_safetydevice

SELECT SafetyDeviceCd1, count(\*)

FROM 610person\_fatal

GROUP BY 1

ORDER BY 2 DESC

2 36

5 20

1 11

7 7

U 6

8 4

13 1

pedestrianschoolInfoInd and pedestrianaction: most are blank

datalib2

##fatal\_fixedobject\_cir

SELECT count(DISTINCT IMAGENO), circumstances

FROM fatal05\_15

WHERE CrashType = "FixedObject"

GROUP BY 2

ORDER BY 1 DESC

SpeedExceededLimit 27

Alcohol 22

ImproperLaneUsage/Change 17

Drugs 7

PhysicalImpairment 4

TooFastforConditions 4

Overcorrected 2

VehicleDefects 1

ViolationSignal/Sign 1

ImproperPassing 1

None 1

##fatal\_cir

SELECT count(DISTINCT IMAGENO), circumstances

FROM fatal05\_15

GROUP BY 2

ORDER BY 1 DESC

37 SpeedExceededLimit

36 Alcohol

28 None

25 ImproperLaneUsage/Change

12 Drugs

11 FailedtoYield

7 TooFastforConditions

7 Distracted/Inattentive

6 PhysicalImpairment

5 Overcorrected

3 WrongWay(One-way)

2 Unknown

2 VehicleDefects

2 FailedtoUseLights

2 FollowingTooClose

1 ImproperlyStoppedonRoadway

1 ImproperPassing

1 ViolationSignal/Sign

1 WrongSide(NotPassing)

Numebrs are different because query

##create table circumstances fatal

Inserts contribcircd2 every time

##fatal\_cir\_no\_of\_crashes

SELECT ContribCircCd,count(DISTINCT crashid)

from circumstances\_fatal

GROUP BY 1

ORDER BY 2 DESC

64

18 24

4 24

13 19

17 8

22 8

19 7

20 5

5 5

14 3

28 3

21 3

9 2

U 2

1 2

26 2

3 1

##fatal\_fixedobject\_cir

SELECT ContribCircCd,count(DISTINCT circumstances\_fatal.crashid)

from circumstances\_fatal

INNER JOIN 610crash

ON 610crash.crashid = circumstances\_fatal.crashid

WHERE 610crash.crashtype="03"

GROUP BY 1

ORDER BY 2 DESC

30

4 22

18 16

13 15

19 5

5 3

20 3

28 2

1 1

##fatal\_safetydevice

SELECT SafetyDeviceCd1, count(\*)

FROM 610person\_fatal

GROUP BY 1

ORDER BY 2 DESC

2 36

5 20

1 11

7 7

U 6

8 4

13 1

##fatal\_safetydevice\_driver

|  |
| --- |
| SELECT SafetyDeviceCd1, count(\*) |
| FROM 610person\_fatal |
| WHERE PersonInvolvementCode=1 |
| GROUP BY 1 |
| ORDER BY 2 DESC |

2 26

5 17

7 7

U 3

8 2

1 1

13 1

0423

Question with 0421.xlsx

Fatal\_cir results different from datalib2

SELECT count(\*)

FROM fatal05\_15

208

SELECT DISTINCT\*

FROM fatal05\_15

206

So there are two duplicate records, but shouldn’t influence the result of fatal\_cir

Added depersonalized2007.zip into crash, vehicle.csv and vehicles2007.csv in clean

Imported vehicle2007.csv into bomdadil.crash\_jasmine

##correct\_ve

DELETE FROM vehicles2007;

LOAD DATA LOCAL INFILE 'C:/Users/yhy6f/Desktop/clean/2007\_De-Personalised\_Extract/vehicles.csv'

INTO TABLE vehicles2007

FIELDS TERMINATED BY ','

ESCAPED BY ''

LINES TERMINATED BY '\r\n'

Created 610vehicle

Created folder agate in folder crash

Copied all csv files into agate

Jupyter notebook Crash

crashes2008 = agate.Table.from\_csv('crashes2008.csv')

crashes2008\_err = crashes2008.where(lambda row: len(row) > 49)

for rows in crashes2008\_err:

   print row

Error:

ValueError: Row 65003 has 50 values, but Table only has 49 columns.

Went back to crashes2008, row 65005 (including header), all cells are moved to the right, from field “Troop District” : “[D”

Vehicles

2015: 2700

2014: import failed Lost connection 93,301 93,300

Csv: 151163

2013: csv 251221

2012: import failed:

[Err] [Row1] [Imp] 2006 - MySQL server has gone away

[Err] [Row1] [Imp] INSERT INTO `vehicles2012`

csv248442

180,623 180,622

2011: csv203655

[Err] [Row1] [Imp] 2006 - MySQL server has gone away

[Err] [Row1] [Imp] INSERT INTO `vehicles2011`

Vehicles2011: row

Abnormalty starts from field Vehicle model and VehicleColorCd1

Vehicle model,  VehicleColorCd1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| MTN | GRN | NA |  | 1 | 3 | N |
| flat bed | one ton | BRO | TAN |  | 1 | 20 |
| Durango | BLU | NA |  | 1 | 3 | N |

So I deleted flatbed and one ton,

crashes = agate.TableSet([

       agate.Table.from\_csv('crashes2005.csv'),

       agate.Table.from\_csv('crashes2006.csv'),

       agate.Table.from\_csv('crashes2007.csv'),

       agate.Table.from\_csv('crashes2008.csv'),

       agate.Table.from\_csv('crashes2009.csv'),

       agate.Table.from\_csv('crashes2010.csv'),

       agate.Table.from\_csv('crashes2011.csv'),

       agate.Table.from\_csv('crashes2012.csv'),

       agate.Table.from\_csv('crashes2013.csv'),

       agate.Table.from\_csv('crashes2014.csv'),

       agate.Table.from\_csv('crashes2015.csv'),

   ],

   ['2005', '2006', '2007','2008','2009','2010','2011','2012','2013','2014','2015'],

   key\_name='year'

)

crashes\_table = crashes.merge()

Stopped at crashes2008

ValueError: Row 119931 has 50 values, but Table only has 49 columns.

After creating 610vehicle,

##

|  |
| --- |
| SELECT 610person\_fatal.SafetyDeviceCd1, count(610vehicle.CrashId) |
| FROM 610person\_fatal, 610vehicle |
| WHERE 610person\_fatal.crashid = 610vehicle.CrashId |
| AND 610person\_fatal.PersonInvolvementCode= "1" |
| AND 610vehicle.VehicleBodyTypeCd IN ("10", "12") |
| GROUP BY 1 |
| ORDER BY 2 DESC |

|  |  |
| --- | --- |
| SafetyDeviceCd1 | no of crashes |
| 7 | 6 |
| 8 | 2 |
| 13 | 1 |

0507

AgencyORI:  <http://www.mshp.dps.missouri.gov/MSHPWeb/Publications/Reports/Agency_Status_2012.pdf>

SELECT DISTINCT AgencyORI

FROM

MO0100400 MUPD

MO0100200 CPD

MO0100000 BOONE COUNTY SO

MOMHPFF00

MOMHPQQ00

MOMHPCC00

MOMHPBB00

MOMHP0022

How many reports in total?

SELECT count(DISTINCT crashid)

FROM 610crash

25228

How many reports filed by MUPD and Boone County Sheriff?

SELECT count(DISTINCT crashid)

FROM 610crash

WHERE AgencyOri IN ("MO0100400", "MO0100000")

2932

0509

Wrote a script to check the column numbers

Check\_csv.py:

import csv

with open('crashes2010.csv') as f:

   reader = csv.reader(f, delimiter =",")

   i = 0

   for row in reader:

       i = i+1;

       if len(row) > 49:

           print i

(csvkit) Yes-MacBook-Pro:data jasmineyehan$ python check\_csv.py

91056

95167 MISSOURI RVR, C, , NA,

99242 BRIDGE A3047 1 MISSOURI RVR, B, , NA,

100991 BRIDGE 3 CST SPRING RD, R,,NA,

111807  LIT PINE,,NA

113617 RVR, C,,NA,

114522 PVT 3312 HIGHWAY 19, CUBA, MISSOURI, 65453,

115630 PVT UEHLIN QUARRY, 4949 STATE ROUTE T,N

117033 BRIDGE A3292 3 MISSOURI RVR, C,,NA,

121360  CEDAR C,,NA,

124401 CEDAR C,,NA

147971 7 MISSOURI RVR, C,,NA,1

148540  MISSOURI RVR, K,,NA,

148715 CEDAR C,,NA,

148747 MISSOURI RVR, K,,NA,

2011:

99786 MISSOURI RIVER, C,,NA,

100201  M,NA,NA,60,0.01,U ,NA,CST NORTH OF MARSHFIELD ST,,NA

108522

109898

113279

114203

117036

118994

119552

119860

122204

122648

125371

139351

139874

142588

142694

2012:

(csvkit) Yes-MacBook-Pro:data jasmineyehan$ python check\_csv.py

71832

97606

97891

101504

101863

102700

103706

106084

107437

109908

111870

113500

114037

118458

120105

134813

135071

2013:

90740

102974

135329

135663

135937

137688

138189

2014:

**new-line character seen in unquoted field - do you need to open the file in universal-newline mode?**

2015: no

2008: all

2007: all

2006:

119933

137182

139570

145905

2005:no result

Print len(row) in persons:

persons2005: 39

persons2006:

 File "check\_csv.py", line 6, in <module>

   for row in reader:

\_csv.Error: new-line character seen in unquoted field - do you need to open the file in universal-newline mode?

Persons2007: 39

In vehicles:

>118

vehicles2006:

87798

Removed the , between Ram 1500 and Laramie

vehicles2007:

34212:  removed the , between S-10 and BLU

61142: removed the , between 7400 and 6x4 and SFA

|  |
| --- |
|  |

vehicles2008：

139589: removed the , between Street Sob and Motorcycle

vehicles2009:

217976

Removed the , between RED and BLK

vehicles2010:

217976

corrected the wrong vehicles2010.csv which had the sm data as vehicles2009.csv

Reran check\_csv.py

217976

Copied line 217900 to 218030

Returned no results

vehicles2012.csv:

51011

85458

117396

174874

Traceback (most recent call last):

 File "check\_csv.py", line 6, in <module>

   for row in reader:

\_csv.Error: field larger than field limit (131072)

Couldn’t find anything wrong

51011

3120012370V2

Deleted the “, “ before LLV

85458

3120011393V2

Deleted the “,” before FELT

117396

3120006086V1

Deleted the “,” before BGE

174874

Deleted the “,” before BLK

vehicles2013:

2097

79240

118617

209499

209499

Crashid 3130029174V1

Original：

3130029174V1 ,3130029174,NA,NA  ,2011,U   ,YARD SPORT, YS 700XLT 4WD,GRN,BLK,1,3,11,N,4 ,NA,NA,N,NA,N,NA,  ,N,11,N,NA,NA, ,NA,NA,NA  ,NA,1,N,IL,1 ,D ,NA,1 ,NA,N ,V,N,NA,NA,N ,NA,12,NA,N ,N ,NULL,,,,,1 ,36,,,,,,,,,,,,,,12,,,,,,,,,,1 ,2 ,3 ,4 ,12,13,14,15,,,,,,,,,,,,,,,,,,42        ,,,,,,,,7 ,,,,

Deleted the “,U”