**Analysis of Motor Vehicle Crashes in Texas in 2013**

By Hans-Jörg Stark

Table of Contents

[1 Task 3](#_Toc82796124)

[2 First Preview and getting familiar with the data 3](#_Toc82796125)

[2.1 Introduction to applied Python Tools 3](#_Toc82796126)

[2.2 First Findings 3](#_Toc82796127)

[3 Data Wrangling 8](#_Toc82796128)

[3.1 Working on temporal information 8](#_Toc82796129)

[3.2 Working on binary variables 8](#_Toc82796130)

[3.3 Thinning out variables with foreign keys 8](#_Toc82796131)

[3.4 Complete Dataset 8](#_Toc82796132)

[3.5 Optimised Dataset Version I 8](#_Toc82796133)

[3.6 Optimised Dataset Version II 9](#_Toc82796134)

[3.7 Statistical Information on Variables 9](#_Toc82796135)

[3.8 Sample Dataset of processed data 12](#_Toc82796136)

[3.9 Geospatial subset 14](#_Toc82796137)

[3.10 Final Check 14](#_Toc82796138)

[3.11 Data Wrangling in Tableau 16](#_Toc82796139)

[3.11.1 Time specific adaptions 16](#_Toc82796140)

[3.11.2 Crash related adaptions 17](#_Toc82796141)

[4 Visualizations 18](#_Toc82796142)

[4.1 Time related analysis 18](#_Toc82796143)

[4.2 Spatial related analysis 21](#_Toc82796144)

[4.3 Severity Type and Speed Class related analysis 24](#_Toc82796145)

[4.4 Medical Advisory related analysis 25](#_Toc82796146)

# Task

The assignment for this project is to create visualizations using data-set in CSV format on motor vehicle crashes in Texas from 2013.

# First Preview and getting familiar with the data

## Introduction to applied Python Tools

Since the data is unfamiliar to me and it comes with many columns and rows I start with a first basic exploration on the dataset. For this and also for the next part of Data Wrangling I continually developed a Python class that I will use throughout the next described steps. My goal is to create a class that I will be able to reuse in the future for similar data exploration tasks. Therefore, I am pursuing a sustainable and reusable coding in Python for future application

The mentioned class is “pdDataFrameTools.py”. I work with it particularly for this project and apply it in the Python file “analyseData.py”. Even if I am not going to describe all the steps I took in detail they can all be found and reproduced by running the file in a corresponding environment.

## First Findings

In order to get a feeling of the data to be analyzed I open the data in Python and put it in a pandas dataframe. The first findings are:

* The dataset contains of 171 columns and 5258 records.[[1]](#footnote-1)

Being efficient in the first analysis steps I created a subset of n=5 records. I discovered that many variables do not have values and are empty. Therefore, I chose the 5 sample records by design to have as many values possible in order to find out about the datatypes and if possible levels as well. Then I exported the samples both in regular tabular and also transposed format. The latter is easier to inspect because I felt more comfortable scrolling vertically than horizontally over all the 171 variables. The results can be found in the file “0\_firstImpressionMotorVehicleCrashes\_5\_Transposed.csv”.

The following facts were found:

* Many columns are void and can be omitted or dropped in order to have more of a manageable dataset that will also be mor performant to work with.
* Many columns hold – this is my interpretation – foreign keys in form of digits that I cannot interpret; thus, these columns can also be dropped
* There are many columns with “\_ID” ending – referring (my assumption) to secondary tables with more information that is at this point not available. These columns can also be dropped.
* There are quite a number of columns with the ending “\_Fl” that hold binary variables with the two level “Y” and “N” for “true” and “false”.
* Unfortunately most column-names are some kind of abbreviations I am not familiar with so I do not really know what they indicate.
* Many variables are part of an address or spatial reference in textual form
* Geographic information of point locations are available in the columns labelled “Latitude” and “Longitude”. These are of potential use for spatial analysis.

As a slightly more enhanced analysis I check the variables for the density of their values, i.e. how many missing values they have. The result can be found in the file “1\_nullValueExamination.csv”. I lists the variable name, the percentage of NULL or empty values and the percentage of these compared to the entire dataset (i.e. record count, n=5258).

|  |  |  |
| --- | --- | --- |
|  | **Percent** | **nulls** |
| **Crash\_ID** | 0.0 | 0 |
| **Crash\_Fatal\_Fl** | 0.0 | 0 |
| **Cmv\_Involv\_Fl** | 0.0 | 0 |
| **Schl\_Bus\_Fl** | 0.0 | 0 |
| **Rr\_Relat\_Fl** | 0.0 | 0 |
| **Medical\_Advisory\_Fl** | 0.0 | 0 |
| **Amend\_Supp\_Fl** | 0.0 | 0 |
| **Active\_School\_Zone\_Fl** | 0.0 | 0 |
| **Crash\_Date** | 0.0 | 0 |
| **Crash\_Time** | 0.0 | 0 |
| **Case\_ID** | 12.8 | 673 |
| **Local\_Use** | 74.743 | 3930 |
| **Rpt\_CRIS\_Cnty\_ID** | 0.0 | 0 |
| **Rpt\_City\_ID** | 0.0 | 0 |
| **Rpt\_Outside\_City\_Limit\_Fl** | 0.0 | 0 |
| **Thousand\_Damage\_Fl** | 0.0 | 0 |
| **Rpt\_Latitude** | 79.194 | 4164 |
| **Rpt\_Longitude** | 79.194 | 4164 |
| **Rpt\_Rdwy\_Sys\_ID** | 0.0 | 0 |
| **Rpt\_Hwy\_Num** | 54.222 | 2851 |
| **Rpt\_Hwy\_Sfx** | 99.239 | 5218 |
| **Rpt\_Road\_Part\_ID** | 0.0 | 0 |
| **Rpt\_Block\_Num** | 16.546 | 870 |
| **Rpt\_Street\_Pfx** | 61.297 | 3223 |
| **Rpt\_Street\_Name** | 0.0 | 0 |
| **Rpt\_Street\_Sfx** | 32.37 | 1702 |
| **Private\_Dr\_Fl** | 0.0 | 0 |
| **Toll\_Road\_Fl** | 0.0 | 0 |
| **Crash\_Speed\_Limit** | 0.0 | 0 |
| **Road\_Constr\_Zone\_Fl** | 0.0 | 0 |
| **Road\_Constr\_Zone\_Wrkr\_Fl** | 0.0 | 0 |
| **Rpt\_Street\_Desc** | 75.979 | 3995 |
| **At\_Intrsct\_Fl** | 0.0 | 0 |
| **Rpt\_Sec\_Rdwy\_Sys\_ID** | 3.918 | 206 |
| **Rpt\_Sec\_Hwy\_Num** | 80.525 | 4234 |
| **Rpt\_Sec\_Hwy\_Sfx** | 99.62 | 5238 |
| **Rpt\_Sec\_Road\_Part\_ID** | 54.032 | 2841 |
| **Rpt\_Sec\_Block\_Num** | 34.234 | 1800 |
| **Rpt\_Sec\_Street\_Pfx** | 74.515 | 3918 |
| **Rpt\_Sec\_Street\_Name** | 0.0 | 0 |
| **Rpt\_Sec\_Street\_Sfx** | 32.18 | 1692 |
| **Rpt\_Ref\_Mark\_Offset\_Amt** | 95.607 | 5027 |
| **Rpt\_Ref\_Mark\_Dist\_Uom** | 95.607 | 5027 |
| **Rpt\_Ref\_Mark\_Dir** | 95.607 | 5027 |
| **Rpt\_Ref\_Mark\_Nbr** | 95.607 | 5027 |
| **Rpt\_Sec\_Street\_Desc** | 87.105 | 4580 |
| **Rpt\_CrossingNumber** | 99.924 | 5254 |
| **Wthr\_Cond\_ID** | 0.0 | 0 |
| **Light\_Cond\_ID** | 0.0 | 0 |
| **Entr\_Road\_ID** | 0.0 | 0 |
| **Road\_Type\_ID** | 58.558 | 3079 |
| **Road\_Algn\_ID** | 0.0 | 0 |
| **Surf\_Cond\_ID** | 0.0 | 0 |
| **Traffic\_Cntl\_ID** | 0.0 | 0 |
| **Investigat\_Notify\_Time** | 0.038 | 2 |
| **Investigat\_Notify\_Meth** | 0.095 | 5 |
| **Investigat\_Arrv\_Time** | 0.038 | 2 |
| **Report\_Date** | 0.0 | 0 |
| **Investigat\_Comp\_Fl** | 0.0 | 0 |
| **ORI\_Number** | 10.289 | 541 |
| **Investigat\_Agency\_ID** | 0.0 | 0 |
| **Investigat\_Area\_ID** | 62.933 | 3309 |
| **Investigat\_District\_ID** | 45.759 | 2406 |
| **Investigat\_Region\_ID** | 16.641 | 875 |
| **Bridge\_Detail\_ID** | 0.0 | 0 |
| **Harm\_Evnt\_ID** | 0.0 | 0 |
| **Intrsct\_Relat\_ID** | 0.0 | 0 |
| **FHE\_Collsn\_ID** | 0.0 | 0 |
| **Obj\_Struck\_ID** | 0.0 | 0 |
| **Othr\_Factr\_ID** | 0.0 | 0 |
| **Road\_Part\_Adj\_ID** | 0.0 | 0 |
| **Road\_Cls\_ID** | 0.0 | 0 |
| **Road\_Relat\_ID** | 0.0 | 0 |
| **Phys\_Featr\_1\_ID** | 0.0 | 0 |
| **Phys\_Featr\_2\_ID** | 0.0 | 0 |
| **Cnty\_ID** | 0.0 | 0 |
| **City\_ID** | 0.0 | 0 |
| **Latitude** | 17.954 | 944 |
| **Longitude** | 17.954 | 944 |
| **Hwy\_Sys** | 49.182 | 2586 |
| **Hwy\_Nbr** | 49.182 | 2586 |
| **Hwy\_Sfx** | 96.995 | 5100 |
| **Dfo** | 58.539 | 3078 |
| **Street\_Name** | 0.0 | 0 |
| **Street\_Nbr** | 56.771 | 2985 |
| **Control** | 58.558 | 3079 |
| **Section** | 58.558 | 3079 |
| **Milepoint** | 58.558 | 3079 |
| **Ref\_Mark\_Nbr** | 58.844 | 3094 |
| **Ref\_Mark\_Displ** | 58.844 | 3094 |
| **Hwy\_Sys\_2** | 92.126 | 4844 |
| **Hwy\_Nbr\_2** | 92.126 | 4844 |
| **Hwy\_Sfx\_2** | 99.677 | 5241 |
| **Street\_Name\_2** | 57.227 | 3009 |
| **Street\_Nbr\_2** | 100.0 | 5258 |
| **Control\_2** | 98.383 | 5173 |
| **Section\_2** | 98.383 | 5173 |
| **Milepoint\_2** | 98.383 | 5173 |
| **Txdot\_Rptable\_Fl** | 0.0 | 0 |
| **Onsys\_Fl** | 0.0 | 0 |
| **Rural\_Fl** | 0.0 | 0 |
| **Crash\_Sev\_ID** | 0.0 | 0 |
| **Pop\_Group\_ID** | 0.0 | 0 |
| **Located\_Fl** | 0.0 | 0 |
| **Day\_of\_Week** | 0.0 | 0 |
| **Hwy\_Dsgn\_Lane\_ID** | 58.558 | 3079 |
| **Hwy\_Dsgn\_Hrt\_ID** | 58.558 | 3079 |
| **Hp\_Shldr\_Left** | 58.558 | 3079 |
| **Hp\_Shldr\_Right** | 58.558 | 3079 |
| **Hp\_Median\_Width** | 58.558 | 3079 |
| **Base\_Type\_ID** | 58.558 | 3079 |
| **Nbr\_Of\_Lane** | 58.558 | 3079 |
| **Row\_Width\_Usual** | 58.558 | 3079 |
| **Roadbed\_Width** | 58.558 | 3079 |
| **Surf\_Width** | 58.558 | 3079 |
| **Surf\_Type\_ID** | 58.558 | 3079 |
| **Curb\_Type\_Left\_ID** | 58.558 | 3079 |
| **Curb\_Type\_Right\_ID** | 58.558 | 3079 |
| **Shldr\_Type\_Left\_ID** | 58.558 | 3079 |
| **Shldr\_Width\_Left** | 58.558 | 3079 |
| **Shldr\_Use\_Left\_ID** | 58.558 | 3079 |
| **Shldr\_Type\_Right\_ID** | 58.558 | 3079 |
| **Shldr\_Width\_Right** | 58.558 | 3079 |
| **Shldr\_Use\_Right\_ID** | 58.558 | 3079 |
| **Median\_Type\_ID** | 59.928 | 3151 |
| **Median\_Width** | 58.558 | 3079 |
| **Rural\_Urban\_Type\_ID** | 58.558 | 3079 |
| **Func\_Sys\_ID** | 58.558 | 3079 |
| **Adt\_Curnt\_Amt** | 58.558 | 3079 |
| **Adt\_Curnt\_Year** | 58.558 | 3079 |
| **Adt\_Adj\_Curnt\_Amt** | 58.558 | 3079 |
| **Pct\_Single\_Trk\_Adt** | 58.558 | 3079 |
| **Pct\_Combo\_Trk\_Adt** | 58.558 | 3079 |
| **Trk\_Aadt\_Pct** | 58.558 | 3079 |
| **Curve\_Type\_ID** | 91.86 | 4830 |
| **Curve\_Lngth** | 91.86 | 4830 |
| **Cd\_Degr** | 91.86 | 4830 |
| **Delta\_Left\_Right\_ID** | 92.868 | 4883 |
| **Dd\_Degr** | 91.86 | 4830 |
| **Feature\_Crossed** | 100.0 | 5258 |
| **Structure\_Number** | 100.0 | 5258 |
| **I\_R\_Min\_Vert\_Clear** | 100.0 | 5258 |
| **Approach\_Width** | 100.0 | 5258 |
| **Bridge\_Median\_ID** | 100.0 | 5258 |
| **Bridge\_Loading\_Type\_ID** | 100.0 | 5258 |
| **Bridge\_Loading\_In\_1000\_Lbs** | 100.0 | 5258 |
| **Bridge\_Srvc\_Type\_On\_ID** | 100.0 | 5258 |
| **Bridge\_Srvc\_Type\_Under\_ID** | 100.0 | 5258 |
| **Culvert\_Type\_ID** | 100.0 | 5258 |
| **Roadway\_Width** | 100.0 | 5258 |
| **Deck\_Width** | 100.0 | 5258 |
| **Bridge\_Dir\_Of\_Traffic\_ID** | 100.0 | 5258 |
| **Bridge\_Rte\_Struct\_Func\_ID** | 100.0 | 5258 |
| **Bridge\_IR\_Struct\_Func\_ID** | 100.0 | 5258 |
| **CrossingNumber** | 99.924 | 5254 |
| **RRCo** | 99.924 | 5254 |
| **Poscrossing\_ID** | 100.0 | 5258 |
| **WDCode\_ID** | 99.924 | 5254 |
| **Standstop** | 100.0 | 5258 |
| **Yield** | 100.0 | 5258 |
| **Incap\_Injry\_Cnt** | 0.0 | 0 |
| **Nonincap\_Injry\_Cnt** | 0.0 | 0 |
| **Poss\_Injry\_Cnt** | 0.0 | 0 |
| **Non\_Injry\_Cnt** | 0.0 | 0 |
| **Unkn\_Injry\_Cnt** | 0.0 | 0 |
| **Tot\_Injry\_Cnt** | 0.0 | 0 |
| **Death\_Cnt** | 0.0 | 0 |
| **MPO\_ID** | 29.593 | 1556 |
| **Investigat\_Service\_ID** | 100.0 | 5258 |
| **Investigat\_DA\_ID** | 100.0 | 5258 |
| **Investigator\_Narrative** | 100.0 | 5258 |

As the table shows there are some columns that are completely void and others that have a high percentage of empty or missing values.

# Data Wrangling

## Working on temporal information

Before cleaning up the dataset and getting rid of unnecessary columns/variables the information on date and time are being brought into proper format so that for temporal analysis they can be used appropriately. This is done with the self-created method “properDateFormat()” (line 43 in “analyseData.py”). The result is a new variable of type date called “CrashDateFormated” and contains the information of the concatenation of the fields “Crash Date” and “Crash Time”.

## Working on binary variables

As mentioned before there are some variables with “Y”/”N” values. Those are being turned into numerical values 0 (=”N) and 1 (=”Y”) for further investigation (line 50 in “analyseData.py”).

## Thinning out variables with foreign keys

In the dataset there are a number of columns with the ending “\_ID” that contain of foreign keys (assumption). These are of no value for this analysis without having any meaning or indication on how to use them for analysis. Therefore at this stage they are being removed to a thinned out new dataset (line 56 ff. in “analyseData.py”).

## Complete Dataset

In a next step I tried to create a dataset that contains only of records with completely filled variables and has no empty or null values (line 67 in “analyseData.py”). The result is an empty dataset which indicates that the entire remaining dataset after the processing of 3.1 to 3.3 is still containing null values in an amount that it cannot be thinned out without losing all the records. In other words: there is no completely filled record among the entire data. The result of this step can be found in “2\_fullyFilledDataFrameMotorVehicleCrashes.csv”.

## Optimised Dataset Version I

In order to get an optimised version as sort of compromise of a dataset I created a method that extracts the top n populated records. n can be passed as parameter to the implemented method and a dataframe of size n is returned (line 72 ff. in “analyseData.py”). I chose for n=500 indicating the top 10%. With the resulting dataset I check if there are still completely empty columns and if so these are being dropped because they are not needed.

The result of this step can be found in “4\_mostPopulatedDataFrameWOEmptyColsMotorVehicleCrashes.csv”. Since it is possible that depending on n also important information like spatial location that will be needed for spatial analysis is being dropped I check if among the remaining not empty variables “Latitude” and “Longitude” are still available. If they are not they shall not be dropped from the original data

(line 83 ff. in “analyseData.py”).

A derived dataset is now produced that choses from the original data only those variables that are not empty for n best records. The resulting dataset is exported into the file “5\_exportBestPossibleDataset\_%i.csv” (%i as placeholder for n 🡪 5\_exportBestPossibleDataset\_500.csv).

The problem with this approach is that the number of variables being kept for further analysis is dependent on n. The higher n, the lower the drop-out rate of empty variables; the lower n the higher the probability that more variables are being dropped.

## Optimised Dataset Version II

An alternative and more robust version of thinning out the dataset is to drop all columns from the original dataset that have more than n % empty values. In 2.2 a list of all columns with the percentage of empty values was provided. According to this overview the threshold for n can be chosen (line 98 ff. in “analyseData.py”). In this example the threshold was set arbitrarily to n=40% which means that a variable, that will stay in the resulting dataset, must have at least a completeness rate of 60% or may not have more than 40% empty values. The resulting dataset is exported into the file “6\_exportOptimisedDataset\_Threshold\_%iPercent.csv” (%i as placeholder for n).

## Statistical Information on Variables

From the remaining dataset – reduced set of variables but still complete set of records – for each column some statistic information is computed and documented in “7\_columnAnalysisMotorVehicleCrashesNP.csv”. The variables are sorted alphabetically (line 110 ff. in “analyseData.py”).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **name** | **type** | **min** | **max** | **sum** | **mean** | **median** | **stdev** |
| **Active\_School\_Zone\_Fl** | boolean | None | None | None | None | None | None |
| **Active\_School\_Zone\_Fl\_01** | int64 | 0 | 1 | 8 | 0.0015214910612400200 | 0.0 | 0.03898031692743520 |
| **Adt\_Adj\_Curnt\_Amt** | float64 | 36.0 | 287905.0 | 108705304.0 | 49887.70261587880 | 23847.0 | 58883.62991489030 |
| **Adt\_Curnt\_Amt** | float64 | 36.0 | 287905.0 | 108705304.0 | 49887.70261587880 | 23847.0 | 58883.62991489030 |
| **Adt\_Curnt\_Year** | float64 | 2014.0 | 2014.0 | 4388506.0 | 2014.0 | 2014.0 | 0.0 |
| **Amend\_Supp\_Fl** | boolean | None | None | None | None | None | None |
| **Amend\_Supp\_Fl\_01** | int64 | 0 | 1 | 247 | 0.04697603651578550 | 0.0 | 0.21160766677447500 |
| **At\_Intrsct\_Fl** | boolean | None | None | None | None | None | None |
| **At\_Intrsct\_Fl\_01** | int64 | 0 | 1 | 1451 | 0.27596044123240800 | 0.0 | 0.44703946556054000 |
| **Cmv\_Involv\_Fl** | boolean | None | None | None | None | None | None |
| **Cmv\_Involv\_Fl\_01** | int64 | 0 | 1 | 333 | 0.06333206542411560 | 0.0 | 0.2435824277890220 |
| **Control** | float64 | 1.0 | 3631.0 | 1511112.0 | 693.4887563102340 | 286.0 | 875.823695332062 |
| **CrashDateFormated** | datetime64[ns] | None | None | None | None | None | None |
| **Crash\_Date** | object | None | None | None | None | None | None |
| **Crash\_Fatal\_Fl** | boolean | None | None | None | None | None | None |
| **Crash\_Fatal\_Fl\_01** | int64 | 0 | 1 | 32 | 0.006085964244960060 | 0.0 | 0.07778223400124400 |
| **Crash\_ID** | int64 | 13056580 | 15731655 | 70308384453 | 13371697.309433200 | 13367782.5 | 175707.57689484500 |
| **Crash\_Speed\_Limit** | int64 | -1 | 85 | 211702 | 40.26283758082920 | 40.0 | 18.205988299999100 |
| **Crash\_Time** | object | None | None | None | None | None | None |
| **Day\_of\_Week** | object | None | None | None | None | None | None |
| **Death\_Cnt** | int64 | 0 | 2 | 34 | 0.006466337010270060 | 0.0 | 0.08477402802055910 |
| **Dfo** | float64 | 0.0 | 877.328 | 258938.50700000000 | 118.77913165137600 | 30.8345 | 180.71997229966700 |
| **Hp\_Median\_Width** | float64 | 0.0 | 463.0 | 50091.0 | 22.988067921064700 | 12.0 | 35.94818029552030 |
| **Hp\_Shldr\_Left** | float64 | 0.0 | 30.0 | 10646.0 | 4.88572739788894 | 4.0 | 4.301458639678890 |
| **Hp\_Shldr\_Right** | float64 | 0.0 | 24.0 | 14484.0 | 6.647085819183110 | 9.0 | 4.410085746088910 |
| **Hwy\_Nbr** | object | None | None | None | None | None | None |
| **Hwy\_Sys** | object | None | None | None | None | None | None |
| **Incap\_Injry\_Cnt** | int64 | 0 | 5 | 160 | 0.030429821224800300 | 0.0 | 0.2022938781231950 |
| **Investigat\_Arrv\_Time** | object | None | None | None | None | None | None |
| **Investigat\_Comp\_Fl** | boolean | None | None | None | None | None | None |
| **Investigat\_Comp\_Fl\_01** | int64 | 0 | 1 | 4715 | 0.896728794218334 | 1.0 | 0.3043417152456880 |
| **Investigat\_Notify\_Meth** | object | None | None | None | None | None | None |
| **Investigat\_Notify\_Time** | object | None | None | None | None | None | None |
| **Latitude** | float64 | 25.87355104 | 36.29784522 | 133020.50416475000 | 30.83460921760540 | 30.295072125 | 1.9356882972552800 |
| **Located\_Fl** | boolean | None | None | None | None | None | None |
| **Located\_Fl\_01** | int64 | 0 | 1 | 4314 | 0.8204640547736780 | 1.0 | 0.38383695745955600 |
| **Longitude** | float64 | -106.59705190000000 | -93.72825139 | -420752.39555925 | -97.53184876199570 | -97.12924858 | 2.483453686705740 |
| **Median\_Width** | float64 | 0.0 | 455.0 | 36856.0 | 16.914180816888500 | 3.0 | 32.850643950992100 |
| **Medical\_Advisory\_Fl** | boolean | None | None | None | None | None | None |
| **Medical\_Advisory\_Fl\_01** | int64 | 0 | 1 | 43 | 0.008178014454165080 | 0.0 | 0.09007040276656480 |
| **Milepoint** | float64 | 0.003 | 73.247 | 26637.537000000000 | 12.22466131252870 | 9.739000000000000 | 10.404947191277900 |
| **Nbr\_Of\_Lane** | float64 | 2.0 | 14.0 | 10032.0 | 4.603946764570900 | 4.0 | 1.9925487987135300 |
| **Non\_Injry\_Cnt** | int64 | 0 | 47 | 10494 | 1.99581589958159 | 2.0 | 1.8175223447882300 |
| **Nonincap\_Injry\_Cnt** | int64 | 0 | 4 | 751 | 0.14282997337390600 | 0.0 | 0.45025190886163400 |
| **ORI\_Number** | object | None | None | None | None | None | None |
| **Onsys\_Fl** | boolean | None | None | None | None | None | None |
| **Onsys\_Fl\_01** | int64 | 0 | 1 | 2724 | 0.5180677063522250 | 1.0 | 0.4997209736749540 |
| **Pct\_Combo\_Trk\_Adt** | float64 | 0.0 | 61.7 | 18088.1 | 8.301101422670970 | 5.3 | 8.703583625858750 |
| **Pct\_Single\_Trk\_Adt** | float64 | 0.7 | 33.0 | 9571.8 | 4.39274896741622 | 3.4 | 3.0949487441579300 |
| **Poss\_Injry\_Cnt** | int64 | 0 | 8 | 1428 | 0.2715861544313430 | 0.0 | 0.6773847497513960 |
| **Private\_Dr\_Fl** | boolean | None | None | None | None | None | None |
| **Private\_Dr\_Fl\_01** | int64 | 0 | 1 | 337 | 0.06409281095473560 | 0.0 | 0.24494148898094700 |
| **Ref\_Mark\_Displ** | float64 | -500.0 | 500.0 | 50.493999999999900 | 0.02333364140480570 | 0.036000000000000000 | 25.752659994529800 |
| **Ref\_Mark\_Nbr** | object | None | None | None | None | None | None |
| **Report\_Date** | object | None | None | None | None | None | None |
| **Road\_Constr\_Zone\_Fl** | boolean | None | None | None | None | None | None |
| **Road\_Constr\_Zone\_Fl\_01** | int64 | 0 | 1 | 199 | 0.03784709014834540 | 0.0 | 0.19084447807263900 |
| **Road\_Constr\_Zone\_Wrkr\_Fl** | boolean | None | None | None | None | None | None |
| **Road\_Constr\_Zone\_Wrkr\_Fl\_01** | int64 | 0 | 1 | 88 | 0.016736401673640200 | 0.0 | 0.12829429017356400 |
| **Roadbed\_Width** | float64 | 20.0 | 318.0 | 172847.0 | 79.32400183570450 | 76.0 | 35.81617729293280 |
| **Row\_Width\_Usual** | float64 | 50.0 | 800.0 | 468676.0 | 215.0876548875630 | 180.0 | 126.67340176728800 |
| **Rpt\_Block\_Num** | object | None | None | None | None | None | None |
| **Rpt\_Hwy\_Num** | object | None | None | None | None | None | None |
| **Rpt\_Outside\_City\_Limit\_Fl** | boolean | None | None | None | None | None | None |
| **Rpt\_Outside\_City\_Limit\_Fl\_01** | int64 | 0 | 1 | 1016 | 0.19322936477748200 | 0.0 | 0.3948688788676490 |
| **Rpt\_Sec\_Block\_Num** | object | None | None | None | None | None | None |
| **Rpt\_Sec\_Street\_Name** | object | None | None | None | None | None | None |
| **Rpt\_Sec\_Street\_Sfx** | object | None | None | None | None | None | None |
| **Rpt\_Street\_Name** | object | None | None | None | None | None | None |
| **Rpt\_Street\_Sfx** | object | None | None | None | None | None | None |
| **Rr\_Relat\_Fl** | boolean | None | None | None | None | None | None |
| **Rr\_Relat\_Fl\_01** | int64 | 0 | 1 | 15 | 0.0028527957398250300 | 0.0 | 0.05334040133225790 |
| **Rural\_Fl** | boolean | None | None | None | None | None | None |
| **Rural\_Fl\_01** | int64 | 0 | 1 | 1205 | 0.2291745910992770 | 0.0 | 0.4203417673673010 |
| **Schl\_Bus\_Fl** | boolean | None | None | None | None | None | None |
| **Schl\_Bus\_Fl\_01** | int64 | 0 | 1 | 18 | 0.003423354887790030 | 0.0 | 0.0584147626815977 |
| **Section** | float64 | 1.0 | 24.0 | 9756.0 | 4.477283157411660 | 3.0 | 3.9298033324250700 |
| **Shldr\_Width\_Left** | float64 | 0.0 | 60.0 | 18096.0 | 8.304726938962830 | 8.0 | 7.866880901053050 |
| **Shldr\_Width\_Right** | float64 | 0.0 | 44.0 | 24813.0 | 11.387333639284100 | 10.0 | 8.83594717833426 |
| **Street\_Name** | object | None | None | None | None | None | None |
| **Street\_Name\_2** | object | None | None | None | None | None | None |
| **Street\_Nbr** | float64 | 2.0 | 41944.0 | 10141653.0 | 4461.791904971400 | 2600.0 | 5170.989359176460 |
| **Surf\_Width** | float64 | 12.0 | 218.0 | 124193.0 | 56.995410738871000 | 48.0 | 24.9669566209884 |
| **Thousand\_Damage\_Fl** | boolean | None | None | None | None | None | None |
| **Thousand\_Damage\_Fl\_01** | int64 | 0 | 1 | 4602 | 0.8752377329783190 | 1.0 | 0.33048058250743200 |
| **Toll\_Road\_Fl** | boolean | None | None | None | None | None | None |
| **Toll\_Road\_Fl\_01** | int64 | 0 | 1 | 56 | 0.010650437428680100 | 0.0 | 0.10265968043708900 |
| **Tot\_Injry\_Cnt** | int64 | 0 | 8 | 2339 | 0.4448459490300490 | 0.0 | 0.8197330099333780 |
| **Trk\_Aadt\_Pct** | float64 | 0.9 | 64.7 | 27655.300000000000 | 12.691739329968000 | 9.7 | 9.858237858600270 |
| **Txdot\_Rptable\_Fl** | boolean | None | None | None | None | None | None |
| **Txdot\_Rptable\_Fl\_01** | int64 | 0 | 1 | 4504 | 0.8565994674781290 | 1.0 | 0.35051417393408400 |
| **Unkn\_Injry\_Cnt** | int64 | 0 | 5 | 878 | 0.16698364397109200 | 0.0 | 0.4231705380561870 |

## Sample Dataset of processed data

For easy inspection a sample on n=5 (could be changed to any number) records of the resulting dataset is exported to a CSV file: “8\_sampleOptimisedDataFrameMotorVehicleCrashes\_%i.csv” (%i as placeholder for n), cf. line 118 in “analyseData.py”. The data is available in standard tabular and transposed format, sorted alphabetically by column names:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **13359487** | **13568672** | **13141256** | **13161804** | **13183463** |
| **Active\_School\_Zone\_Fl** | N | N | N | N | N |
| **Active\_School\_Zone\_Fl\_01** | 0 | 0 | 0 | 0 | 0 |
| **Adt\_Adj\_Curnt\_Amt** |  |  |  |  |  |
| **Adt\_Curnt\_Amt** |  |  |  |  |  |
| **Adt\_Curnt\_Year** |  |  |  |  |  |
| **Amend\_Supp\_Fl** | N | N | N | N | N |
| **Amend\_Supp\_Fl\_01** | 0 | 0 | 0 | 0 | 0 |
| **At\_Intrsct\_Fl** | N | N | N | Y | N |
| **At\_Intrsct\_Fl\_01** | 0 | 0 | 0 | 1 | 0 |
| **Cmv\_Involv\_Fl** | N | N | Y | N | N |
| **Cmv\_Involv\_Fl\_01** | 0 | 0 | 1 | 0 | 0 |
| **Control** |  |  |  |  |  |
| **CrashDateFormated** | 2013-07-09 09:30:00 | 2013-11-16 00:00:00 | 2013-02-12 17:05:00 | 2013-02-26 13:35:00 | 2013-03-05 14:45:00 |
| **Crash\_Date** | 7/9/2013 | 11/16/2013 | 2/12/2013 | 2/26/2013 | 3/5/2013 |
| **Crash\_Fatal\_Fl** | N | N | N | N | N |
| **Crash\_Fatal\_Fl\_01** | 0 | 0 | 0 | 0 | 0 |
| **Crash\_Speed\_Limit** | 15 | 60 | 15 | 0 | 35 |
| **Crash\_Time** | 9:30 AM | 12:00 AM | 5:05 PM | 1:35 PM | 2:45 PM |
| **Day\_of\_Week** | TUE | SAT | TUE | TUE | TUE |
| **Death\_Cnt** | 0 | 0 | 0 | 0 | 0 |
| **Dfo** |  |  |  |  |  |
| **Hp\_Median\_Width** |  |  |  |  |  |
| **Hp\_Shldr\_Left** |  |  |  |  |  |
| **Hp\_Shldr\_Right** |  |  |  |  |  |
| **Hwy\_Nbr** |  |  |  |  |  |
| **Hwy\_Sys** |  |  |  |  |  |
| **Incap\_Injry\_Cnt** | 0 | 0 | 0 | 0 | 0 |
| **Investigat\_Arrv\_Time** | 10:15 AM | 9:45 AM | 5:10 PM | 2:19 PM | 3:04 PM |
| **Investigat\_Comp\_Fl** | Y | N | Y | Y | Y |
| **Investigat\_Comp\_Fl\_01** | 1 | 0 | 1 | 1 | 1 |
| **Investigat\_Notify\_Meth** | DISPATCHED | DISPATCHED | DISPATCHED | PHONE | ATASCOSA COUNTY S.O. |
| **Investigat\_Notify\_Time** | 9:35 AM | 9:15 AM | 5:08 PM | 2:19 PM | 2:52 PM |
| **Latitude** |  |  |  |  |  |
| **Located\_Fl** | N | N | N | N | N |
| **Located\_Fl\_01** | 0 | 0 | 0 | 0 | 0 |
| **Longitude** |  |  |  |  |  |
| **Median\_Width** |  |  |  |  |  |
| **Medical\_Advisory\_Fl** | N | N | N | N | N |
| **Medical\_Advisory\_Fl\_01** | 0 | 0 | 0 | 0 | 0 |
| **Milepoint** |  |  |  |  |  |
| **Nbr\_Of\_Lane** |  |  |  |  |  |
| **Non\_Injry\_Cnt** | 2 | 0 | 1 | 2 | 0 |
| **Nonincap\_Injry\_Cnt** | 0 | 0 | 0 | 0 | 0 |
| **ORI\_Number** |  |  | TX2400500 | TX0710200 |  |
| **Onsys\_Fl** | N | N | N | N | N |
| **Onsys\_Fl\_01** | 0 | 0 | 0 | 0 | 0 |
| **Pct\_Combo\_Trk\_Adt** |  |  |  |  |  |
| **Pct\_Single\_Trk\_Adt** |  |  |  |  |  |
| **Poss\_Injry\_Cnt** | 0 | 0 | 0 | 0 | 0 |
| **Private\_Dr\_Fl** | Y | N | N | N | N |
| **Private\_Dr\_Fl\_01** | 1 | 0 | 0 | 0 | 0 |
| **Ref\_Mark\_Displ** |  |  |  |  |  |
| **Ref\_Mark\_Nbr** |  |  |  |  |  |
| **Report\_Date** | 7/10/2013 | 11/24/2013 | 2/12/2013 | 2/26/2013 | 3/11/2013 |
| **Road\_Constr\_Zone\_Fl** | N | N | N | N | N |
| **Road\_Constr\_Zone\_Fl\_01** | 0 | 0 | 0 | 0 | 0 |
| **Road\_Constr\_Zone\_Wrkr\_Fl** | N | N | N | N | N |
| **Road\_Constr\_Zone\_Wrkr\_Fl\_01** | 0 | 0 | 0 | 0 | 0 |
| **Roadbed\_Width** |  |  |  |  |  |
| **Row\_Width\_Usual** |  |  |  |  |  |
| **Rpt\_Block\_Num** |  |  |  |  |  |
| **Rpt\_Hwy\_Num** |  |  |  |  |  |
| **Rpt\_Outside\_City\_Limit\_Fl** | N | Y | N | N | Y |
| **Rpt\_Outside\_City\_Limit\_Fl\_01** | 0 | 1 | 0 | 0 | 1 |
| **Rpt\_Sec\_Block\_Num** |  |  |  |  |  |
| **Rpt\_Sec\_Street\_Name** | NOT REPORTED | 9 12 | LAREDO COMMUNITY COLLEGE ROADWAY | VISCOUNT BLVD EB | NOT REPORTED |
| **Rpt\_Sec\_Street\_Sfx** |  |  |  |  |  |
| **Rpt\_Street\_Name** | PARKING LOT | D 12 | LAREDO COMMUNITY COLLEGE ROADWAY | AIRWAY BLVD NB | BLUNZER |
| **Rpt\_Street\_Sfx** |  |  |  |  | RD |
| **Rr\_Relat\_Fl** | N | N | N | N | N |
| **Rr\_Relat\_Fl\_01** | 0 | 0 | 0 | 0 | 0 |
| **Rural\_Fl** | N | Y | N | N | Y |
| **Rural\_Fl\_01** | 0 | 1 | 0 | 0 | 1 |
| **Schl\_Bus\_Fl** | N | N | N | N | N |
| **Schl\_Bus\_Fl\_01** | 0 | 0 | 0 | 0 | 0 |
| **Section** |  |  |  |  |  |
| **Shldr\_Width\_Left** |  |  |  |  |  |
| **Shldr\_Width\_Right** |  |  |  |  |  |
| **Street\_Name** | PARKING LOT | E D 12 | LAREDO COMMUNITY COLLEGE ROADWAY | AIRWAY BLVD NB | BLUNZER RD |
| **Street\_Name\_2** |  |  |  |  |  |
| **Street\_Nbr** |  |  |  |  |  |
| **Surf\_Width** |  |  |  |  |  |
| **Thousand\_Damage\_Fl** | Y | Y | N | Y | Y |
| **Thousand\_Damage\_Fl\_01** | 1 | 1 | 0 | 1 | 1 |
| **Toll\_Road\_Fl** | N | N | N | N | N |
| **Toll\_Road\_Fl\_01** | 0 | 0 | 0 | 0 | 0 |
| **Tot\_Injry\_Cnt** | 0 | 0 | 0 | 0 | 0 |
| **Trk\_Aadt\_Pct** |  |  |  |  |  |
| **Txdot\_Rptable\_Fl** | N | Y | N | Y | Y |
| **Txdot\_Rptable\_Fl\_01** | 0 | 1 | 0 | 1 | 1 |
| **Unkn\_Injry\_Cnt** | 0 | 1 | 0 | 0 | 1 |

In this sample there may be empty columns – but over the entire dataset they are filled to at least the percentage that was defined (cf. 3.6 ).

## Geospatial subset

For specific geospatial and spatio-temporal analysis in a Geographic Information System a separate dataset is created that contains only information on the Crash ID, Date and Time and Longitude and Latitude (line 122 ff. in “analyseData.py”). This data can be found in “subSetGeoMotorVehicleCrashes.csv”.

## Final Check

As a final check on lines 129 in “analyseData.py” the remaining data is checked for Null Values. The result may not have any variable with a percentage of Null values above the defined threshold! The result is available in “9\_finalExaminationThinnedDf.csv”.

|  |  |  |
| --- | --- | --- |
|  | **Percent** | **nulls** |
| **Active\_School\_Zone\_Fl** | 0.0 | 0 |
| **Active\_School\_Zone\_Fl\_01** | 0.0 | 0 |
| **Adt\_Adj\_Curnt\_Amt** | 58.558 | 3079 |
| **Adt\_Curnt\_Amt** | 58.558 | 3079 |
| **Adt\_Curnt\_Year** | 58.558 | 3079 |
| **Amend\_Supp\_Fl** | 0.0 | 0 |
| **Amend\_Supp\_Fl\_01** | 0.0 | 0 |
| **At\_Intrsct\_Fl** | 0.0 | 0 |
| **At\_Intrsct\_Fl\_01** | 0.0 | 0 |
| **Cmv\_Involv\_Fl** | 0.0 | 0 |
| **Cmv\_Involv\_Fl\_01** | 0.0 | 0 |
| **Control** | 58.558 | 3079 |
| **CrashDateFormated** | 0.0 | 0 |
| **Crash\_Date** | 0.0 | 0 |
| **Crash\_Fatal\_Fl** | 0.0 | 0 |
| **Crash\_Fatal\_Fl\_01** | 0.0 | 0 |
| **Crash\_ID** | 0.0 | 0 |
| **Crash\_Speed\_Limit** | 0.0 | 0 |
| **Crash\_Time** | 0.0 | 0 |
| **Day\_of\_Week** | 0.0 | 0 |
| **Death\_Cnt** | 0.0 | 0 |
| **Dfo** | 58.539 | 3078 |
| **Hp\_Median\_Width** | 58.558 | 3079 |
| **Hp\_Shldr\_Left** | 58.558 | 3079 |
| **Hp\_Shldr\_Right** | 58.558 | 3079 |
| **Hwy\_Nbr** | 49.182 | 2586 |
| **Hwy\_Sys** | 49.182 | 2586 |
| **Incap\_Injry\_Cnt** | 0.0 | 0 |
| **Investigat\_Arrv\_Time** | 0.038 | 2 |
| **Investigat\_Comp\_Fl** | 0.0 | 0 |
| **Investigat\_Comp\_Fl\_01** | 0.0 | 0 |
| **Investigat\_Notify\_Meth** | 0.095 | 5 |
| **Investigat\_Notify\_Time** | 0.038 | 2 |
| **Latitude** | 17.954 | 944 |
| **Located\_Fl** | 0.0 | 0 |
| **Located\_Fl\_01** | 0.0 | 0 |
| **Longitude** | 17.954 | 944 |
| **Median\_Width** | 58.558 | 3079 |
| **Medical\_Advisory\_Fl** | 0.0 | 0 |
| **Medical\_Advisory\_Fl\_01** | 0.0 | 0 |
| **Milepoint** | 58.558 | 3079 |
| **Nbr\_Of\_Lane** | 58.558 | 3079 |
| **Non\_Injry\_Cnt** | 0.0 | 0 |
| **Nonincap\_Injry\_Cnt** | 0.0 | 0 |
| **ORI\_Number** | 10.289 | 541 |
| **Onsys\_Fl** | 0.0 | 0 |
| **Onsys\_Fl\_01** | 0.0 | 0 |
| **Pct\_Combo\_Trk\_Adt** | 58.558 | 3079 |
| **Pct\_Single\_Trk\_Adt** | 58.558 | 3079 |
| **Poss\_Injry\_Cnt** | 0.0 | 0 |
| **Private\_Dr\_Fl** | 0.0 | 0 |
| **Private\_Dr\_Fl\_01** | 0.0 | 0 |
| **Ref\_Mark\_Displ** | 58.844 | 3094 |
| **Ref\_Mark\_Nbr** | 58.844 | 3094 |
| **Report\_Date** | 0.0 | 0 |
| **Road\_Constr\_Zone\_Fl** | 0.0 | 0 |
| **Road\_Constr\_Zone\_Fl\_01** | 0.0 | 0 |
| **Road\_Constr\_Zone\_Wrkr\_Fl** | 0.0 | 0 |
| **Road\_Constr\_Zone\_Wrkr\_Fl\_01** | 0.0 | 0 |
| **Roadbed\_Width** | 58.558 | 3079 |
| **Row\_Width\_Usual** | 58.558 | 3079 |
| **Rpt\_Block\_Num** | 16.546 | 870 |
| **Rpt\_Hwy\_Num** | 54.222 | 2851 |
| **Rpt\_Outside\_City\_Limit\_Fl** | 0.0 | 0 |
| **Rpt\_Outside\_City\_Limit\_Fl\_01** | 0.0 | 0 |
| **Rpt\_Sec\_Block\_Num** | 34.234 | 1800 |
| **Rpt\_Sec\_Street\_Name** | 0.0 | 0 |
| **Rpt\_Sec\_Street\_Sfx** | 32.18 | 1692 |
| **Rpt\_Street\_Name** | 0.0 | 0 |
| **Rpt\_Street\_Sfx** | 32.37 | 1702 |
| **Rr\_Relat\_Fl** | 0.0 | 0 |
| **Rr\_Relat\_Fl\_01** | 0.0 | 0 |
| **Rural\_Fl** | 0.0 | 0 |
| **Rural\_Fl\_01** | 0.0 | 0 |
| **Schl\_Bus\_Fl** | 0.0 | 0 |
| **Schl\_Bus\_Fl\_01** | 0.0 | 0 |
| **Section** | 58.558 | 3079 |
| **Shldr\_Width\_Left** | 58.558 | 3079 |
| **Shldr\_Width\_Right** | 58.558 | 3079 |
| **Street\_Name** | 0.0 | 0 |
| **Street\_Name\_2** | 57.227 | 3009 |
| **Street\_Nbr** | 56.771 | 2985 |
| **Surf\_Width** | 58.558 | 3079 |
| **Thousand\_Damage\_Fl** | 0.0 | 0 |
| **Thousand\_Damage\_Fl\_01** | 0.0 | 0 |
| **Toll\_Road\_Fl** | 0.0 | 0 |
| **Toll\_Road\_Fl\_01** | 0.0 | 0 |
| **Tot\_Injry\_Cnt** | 0.0 | 0 |
| **Trk\_Aadt\_Pct** | 58.558 | 3079 |
| **Txdot\_Rptable\_Fl** | 0.0 | 0 |
| **Txdot\_Rptable\_Fl\_01** | 0.0 | 0 |
| **Unkn\_Injry\_Cnt** | 0.0 | 0 |

## Data Wrangling in Tableau

### Time specific adaptions

Now that the data has been prepared for use in Tableau there are still some more data wrangling options of interest.

Firstly, the time information is classified into four classes: Morning, Afternoon, Evening and Night seasons. This process of dummy coding for this classification is conducted directly in Tableau with the option to create a new computed field. The following screenshots show the definitions of the four mentioned classes:

|  |  |
| --- | --- |
|  |  |
| Definition of Morning Day-Season:  06:00 – 11.59 a.m. | Definition of Afternoon Day-Season:  00:00 – 05.59 p.m. |
|  |  |
| Definition of Evening Day-Season:  06:00 – 10.59 p.m. | Definition of Night Day-Season:  11:00 p.m. – 05.59 a.m. |

The next step imputes four levels for a categorical variable called “DaySeason” to express in verbal form the time-season of the day when the crash occurred:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

### Crash related adaptions

It is definitively interesting to classify the crashes according to their severity. There seems to be no direct classification on the severity of the accidents. I therefore chose to implement three classes as indicator for some kind of crash-severity:

1. Crashes without physical damage or harm
2. Crashes with injured people
3. Crashes that caused people to die

The implementation in Tableau is as follows, numerical and nominal:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

# Visualizations

The next section provides an overview of the conducted analyses in Tableau and QGIS.

## Time related analysis

This paragraph shows different time related graphs and analyses.

|  |
| --- |
|  |
| There seem to be no extreme monthly outlier in the aggregation of the count of crashes. Interestingly the highest number is in June when the days are the longest. Which might not be expected. An explanation might be that during the warm(er) season there is more traffic and thus a higher probability of an accident. |

|  |
| --- |
|  |
|  |
| The highest number of crashes is in the afternoons and the lowest during the nights. This might be caused by the total traffic in general. |

|  |
| --- |
|  |
| The highest number of crashes is on Fridays. This could be because it is the end of the traditional working-week, so people might be tired and less attentive. On Sundays there is the lowest number, probably because the traffic is lower. |

|  |
| --- |
|  |
| The figure shows the analysis of the number of accidents per quarter, grouped by day segments. At first glance, the distribution of values across all quarters looks very similar. On closer inspection, however, it is noticeable that the ratio between the incidents in the morning and in the evening varies over the quarters: While it is practically balanced in the 4th quarter, it diverges in the 3rd quarter. This becomes more visible when the graph is labelled with the percentage shares (see figure below). |
|  |

## Spatial related analysis

This paragraph shows some spatially related graphs and findings. For these analyses the information on Latitude and Longitude is essential to draw the following maps. According to my previous research I found that of the available data 944 records (i.e. 17.95%) do not have spatial information. Although this is quite a significant number the spatial analysis of the remaining data shows some interesting findings. As backdrop map a Web Map Service (WMS) from OpenStreetMap was chosen and integrated into Tableau.

|  |
| --- |
| Ein Bild, das Karte enthält.  Automatisch generierte Beschreibung |
| The map shows only all available locations of crashes with recorded deaths, classified by day segments. The highest densities are around urban areas. But also, on major streets or network connections crashes can be found. |

|  |
| --- |
|  |
| The map shows only the location of where crashes with recorded deaths occurred, classified by day segments. It is interesting to see, that deadly crashes at night are mostly found in cities or higher populated areas. On the other hand, deadly crashes in the morning are nearly all outside of cities. Deadly crashes in the afternoons and evenings seem to be equally distributed between rural and urban areas. |

The next four maps show the locations of deadly crashes individualized by daysegments.

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

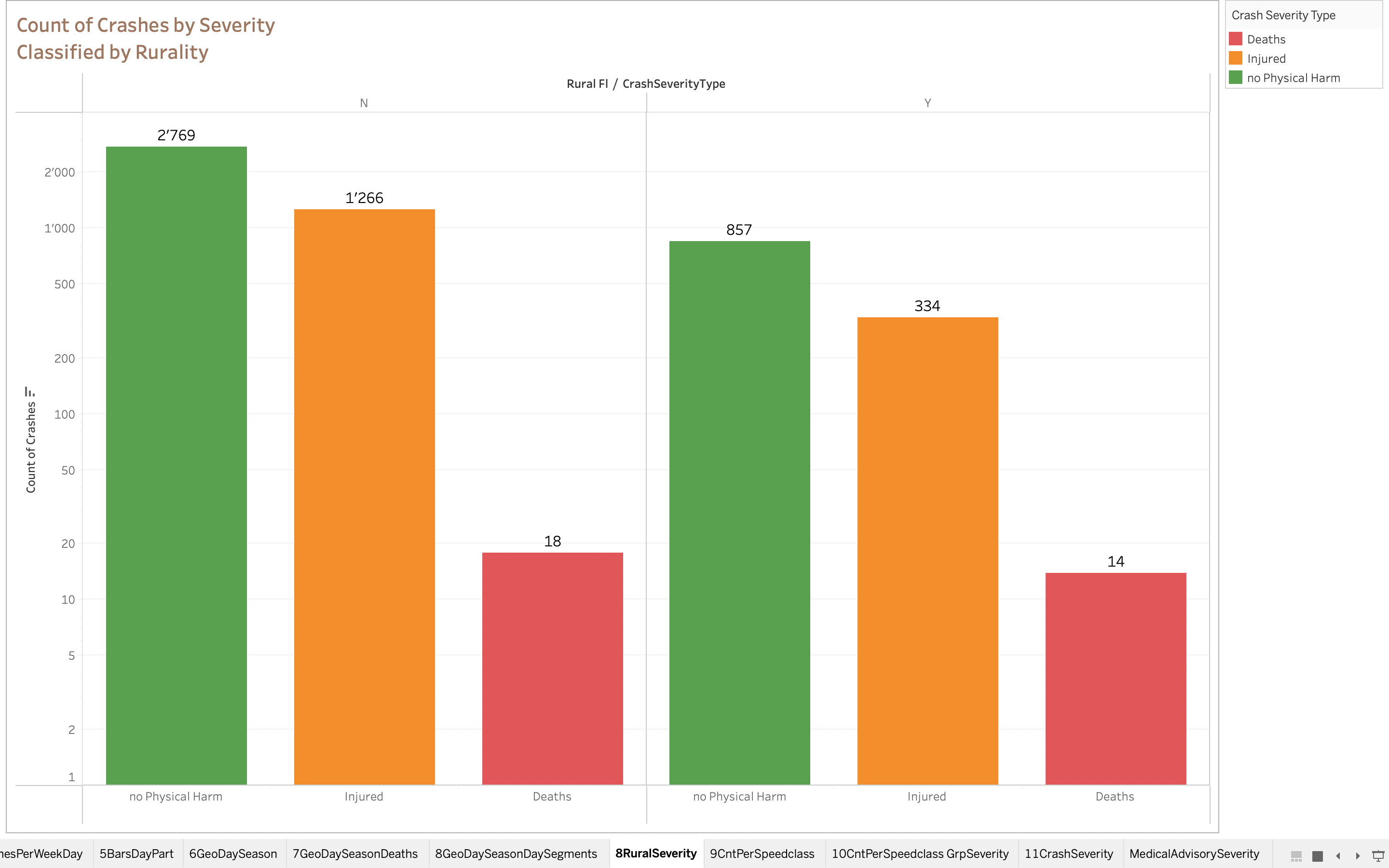
The comparison of the four maps show that for all day segments most of the crashes happened in the densely populated areas. In rural areas there are differences where the crashes occurred.

Finally animated visualizations for the crashes over the entire year 2013 can be found in the assignment directory in the file “vehiclecrashesweekly.mov” as animation aggregated by weeks and “vehiclecrashesdaily.mov” aggregated by days.

Ein Bild, das Karte enthält.

Automatisch generierte Beschreibung

Finally, some information about rurality of the crashes:



The figure shows on the left no rural areas and on the right rural areas and the count of crashes per severity type. Mind you for better readability the scale is logarithmic! There seem to be no significant differences between crashes in rural and non-rural areas.

## Severity Type and Speed Class related analysis

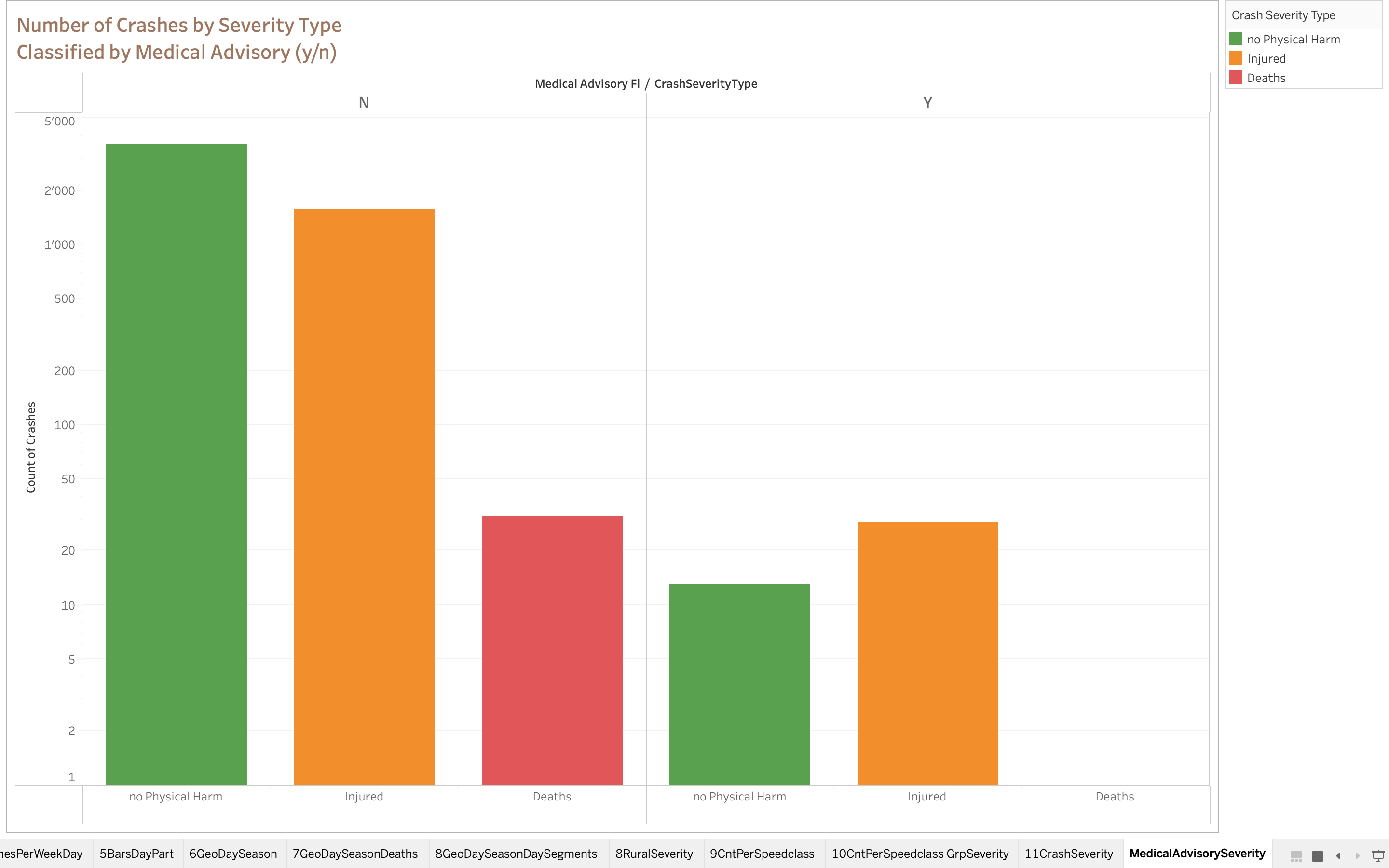
This paragraph shows different analyses based on the speed limit classes and the derived variable of severity of crash.

|  |
| --- |
|  |
| The graph shows the distribution of crash counts by speed limit classes. Very interestingly the highest number is not where people are allowed to drive the fastest but in some sort of middle-ranged speed class around 30 mph. Also, there is a significant number of crashes with speed limit class 0. Either 0 stands for a null or unknown value or this number could indicate rear-end-collisions.  The distribution of the global crash severity can be found in the following pie chart: |

|  |
| --- |
|  |
| The figure shows the distribution of crash counts by speed limit classes and grouped by the severity of the crashes. As severity there have been three classes implemented (cf. 3.11.2). Mind you: For better comparability, the scales were individually adapted!  There is a significant threshold between speed limit classes 25 and 30!  While most of the crashes with no physical harm or only injured people occur in the speed limit classes 30 to 45 the highest number of deaths in crashes occurs in the speed limit class 55! Thus, the conclusion may be drawn that speed has a significant influence on the severity of the crash in terms of injuries. |

## Medical Advisory related analysis

The final chart is on finding out about a correlation between medical advisory – which is available in the dataset as Boolean variable – and the severity type of crashes.



The scale in this figure is for better comparability again logarithmic. Interestingly there seems to be significant differences between crashes with or without medical advisories. Surprisingly there are no deaths recorded where there was medical advisory involved!

1. Most of the figures and findings I present can be verified and found in the file “\_analysis.log” that I created during the Python data wrangling process as a log-file of my processes [↑](#footnote-ref-1)