

Analysis of Motor Vehicle Crashes in Texas in 2013

By Hans-Jörg Stark

Table of Contents

1	TASK	3
2	FIRST PREVIEW AND GETTING FAMILIAR WITH THE DATA	3
2.1	Introduction to applied Python Tools	3
2.2	First Findings	3
3	DATA WRANGLING	8
3.1	Working on temporal information	8
3.2	Working on binary variables	8
3.3	Thinning out variables with foreign keys	8
3.4	Complete Dataset	8
3.5	Optimised Dataset Version I	8
3.6	Optimised Dataset Version II	9
3.7	Statistical Information on Variables	9
3.8	Sample Dataset of processed data	12
3.9	Geospatial subset	14
3.10	Final Check	14
3.11	Data Wrangling in Tableau	17
3.11.1	Time specific adaptations	17
3.11.2	Crash related adaptations	18
4	VISUALIZATIONS	18
4.1	Time related analysis	18
4.2	Spatial related analysis	21
4.3	Severity Type and Speed Class related analysis	24
4.4	Medical Advisory related analysis	25

1 Task

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

2 First Preview and getting familiar with the data

2.1 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.

2.2 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

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

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

- 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”. It 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_FI	0.0	0
Cmv_Involv_FI	0.0	0
Schl_Bus_FI	0.0	0
Rr_Relat_FI	0.0	0
Medical_Advisory_FI	0.0	0
Amend_Supp_FI	0.0	0
Active_School_Zone_FI	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_FI	0.0	0
Thousand_Damage_FI	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_FI	0.0	0
Toll_Road_FI	0.0	0
Crash_Speed_Limit	0.0	0
Road_Constr_Zone_FI	0.0	0
Road_Constr_Zone_Wrkr_FI	0.0	0
Rpt_Street_Desc	75.979	3995
At_Intrstct_FI	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
Intrscat_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_Currt_Amt	58.558	3079
Adt_Currt_Year	58.558	3079
Adt_Adj_Currt_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.

3 Data Wrangling

3.1 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 `CrashDateFormatted` and contains the information of the concatenation of the fields `Crash Date` and `Crash Time`.

3.2 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`).

3.3 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`).

3.4 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`.

3.5 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 chooses 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.

3.6 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).

3.7 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_Currt_Amt	float64	36.0	287905.0	108705304.0	49887.70261587880	23847.0	58883.62991489030
Adt_Currt_Amt	float64	36.0	287905.0	108705304.0	49887.70261587880	23847.0	58883.62991489030
Adt_Currt_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
CrashDateFormatted	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_FI_01	int64	0	1	337	0.06409281095473560	0.0	0.24494148898094700
Ref_Mark_Displ	float64	-500.0	500.0	50.493999999999900	0.02333364140480570	0.0360000000000000	25.752659994529800
Ref_Mark_Nbr	object	None	None	None	None	None	None
Report_Date	object	None	None	None	None	None	None
Road_Constr_Zone_FI	boolean	None	None	None	None	None	None
Road_Constr_Zone_FI_01	int64	0	1	199	0.03784709014834540	0.0	0.19084447807263900
Road_Constr_Zone_Wrkr_FI	boolean	None	None	None	None	None	None
Road_Constr_Zone_Wrkr_FI_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_FI	boolean	None	None	None	None	None	None
Rpt_Outside_City_Limit_FI_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_FI	boolean	None	None	None	None	None	None
Rr_Relat_FI_01	int64	0	1	15	0.0028527957398250300	0.0	0.05334040133225790
Rural_FI	boolean	None	None	None	None	None	None
Rural_FI_01	int64	0	1	1205	0.2291745910992770	0.0	0.4203417673673010
Schl_Bus_FI	boolean	None	None	None	None	None	None
Schl_Bus_FI_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_FI	boolean	None	None	None	None	None	None
Thousand_Damage_FI_01	int64	0	1	4602	0.8752377329783190	1.0	0.33048058250743200
Toll_Road_FI	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

3.8 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_Currt_Amt					
Adt_Currt_Amt					
Adt_Currt_Year					
Amend_Supp_Fl	N	N	N	N	N
Amend_Supp_Fl_01	0	0	0	0	0
At_Intrstct_Fl	N	N	N	Y	N
At_Intrstct_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					
CrashDateFormatted	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).

3.9 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”.

3.10 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_Currt_Amt	58.558	3079
Adt_Currt_Amt	58.558	3079
Adt_Currt_Year	58.558	3079
Amend_Supp_Fl	0.0	0
Amend_Supp_Fl_01	0.0	0
At_Intrscct_Fl	0.0	0

At_Intrstct_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

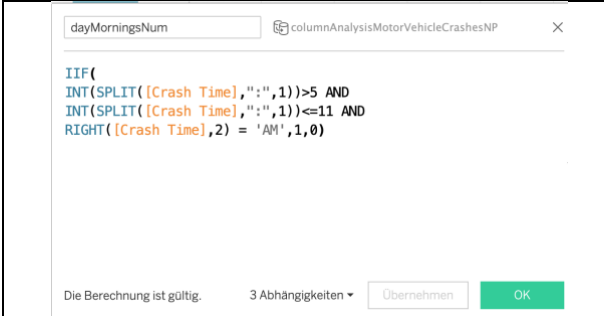
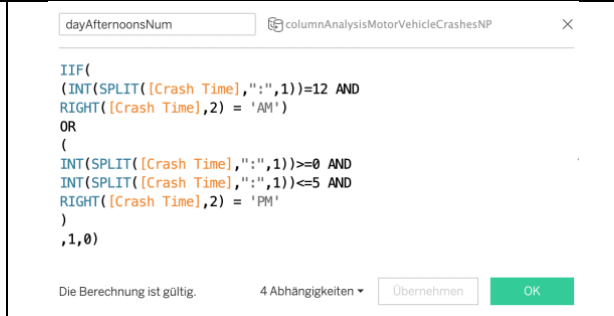
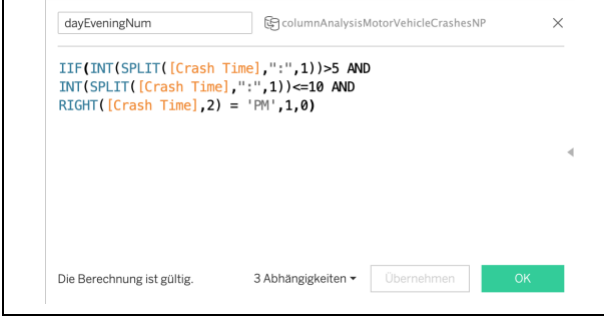
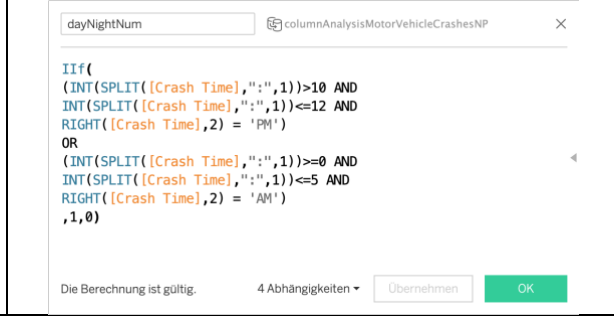
Private_Dr_Fl_01	0.0	0
Ref_Mark_Displ	58.844	3094
Ref_Mark_Nbr	58.844	3094
Report_Date	0.0	0
Road_Constr_Zone_Fl	0.0	0
Road_Constr_Zone_Fl_01	0.0	0
Road_Constr_Zone_Wrkr_Fl	0.0	0
Road_Constr_Zone_Wrkr_Fl_01	0.0	0
Roadbed_Width	58.558	3079
Row_Width_Usual	58.558	3079
Rpt_Block_Num	16.546	870
Rpt_Hwy_Num	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

3.11 Data Wrangling in Tableau

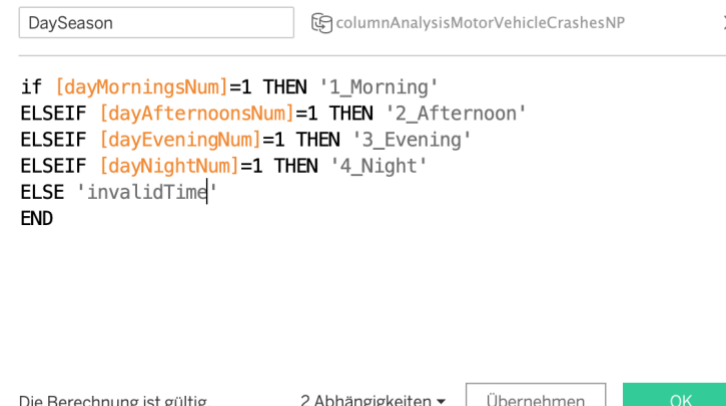
3.11.1 Time specific adaptations

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:



3.11.2 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:

CrashSeverityType

columnAnalysisMotorVehicleCrashesNP

```
IF [Tot Injry Cnt] = 0 AND [Death Cnt] = 0
THEN "noPhysicalHarm"
ELSEIF [Tot Injry Cnt] > 0 AND [Death Cnt] = 0
THEN "crashInjured"
ELSEIF [Death Cnt] > 0
THEN "crashDeaths"
END
```

Die Berechnung ist gültig. 5 Abhängigkeiten Übernehmen OK

crashSeverityTypeNum

columnAnalysisMotorVehicleCrashesNP

```
IF [Tot Injry Cnt] = 0 AND [Death Cnt] = 0
THEN 0
ELSEIF [Tot Injry Cnt] > 0 AND [Death Cnt] = 0
THEN 1
ELSEIF [Death Cnt] > 0
THEN 2
END
```

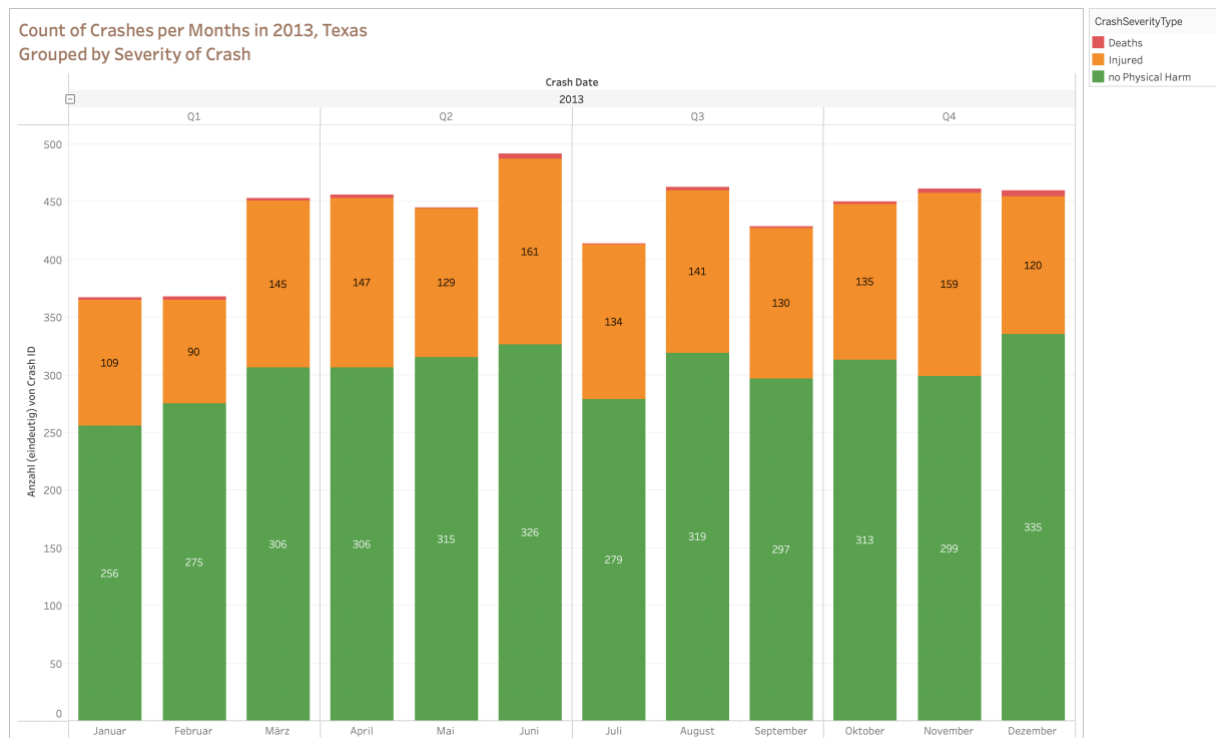
Die Berechnung ist gültig. Übernehmen OK

4 Visualizations

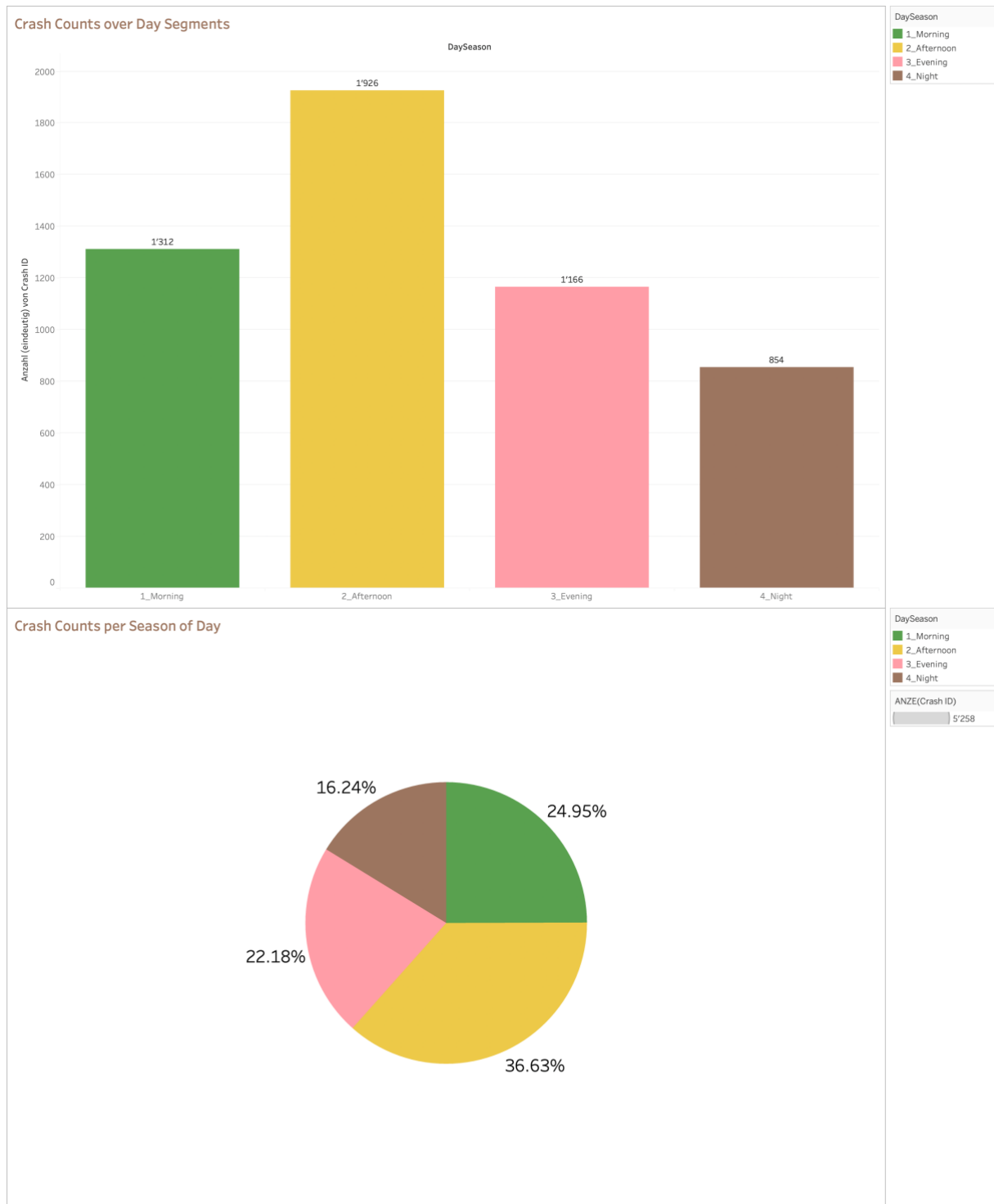
The next section provides an overview of the conducted analyses in Tableau and QGIS.

4.1 Time related analysis

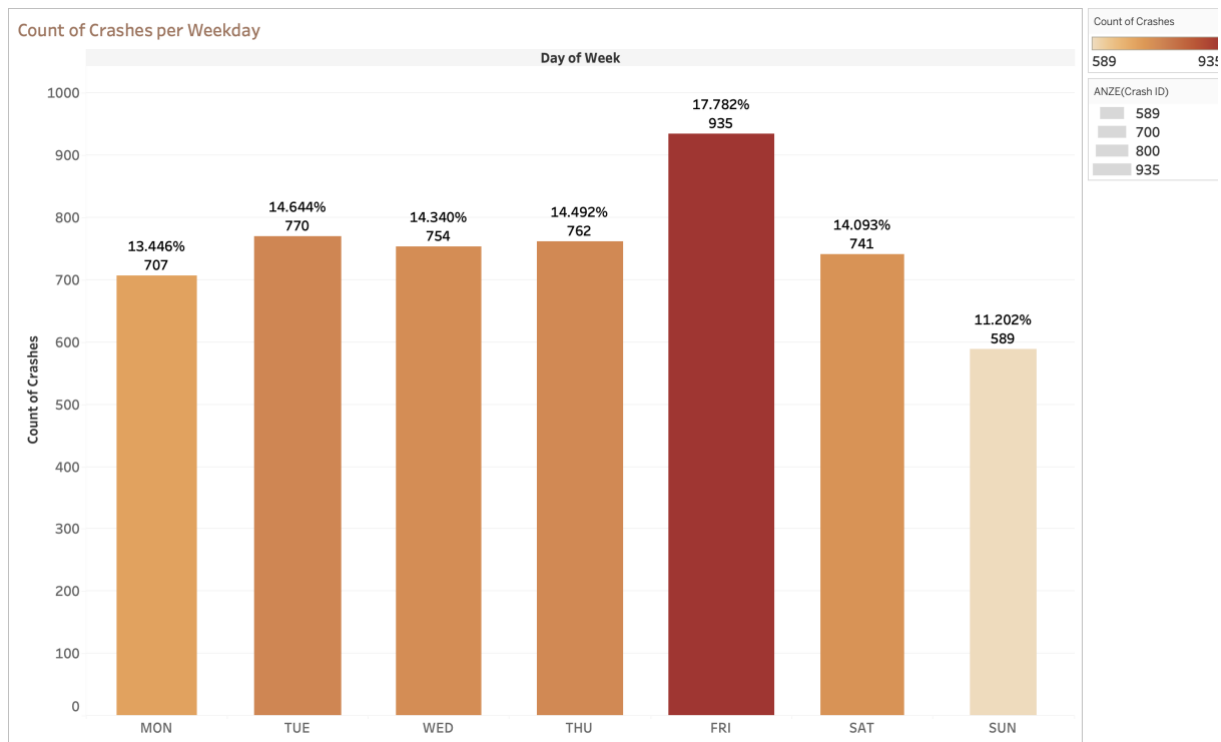
This paragraph shows different time related graphs and analyses.



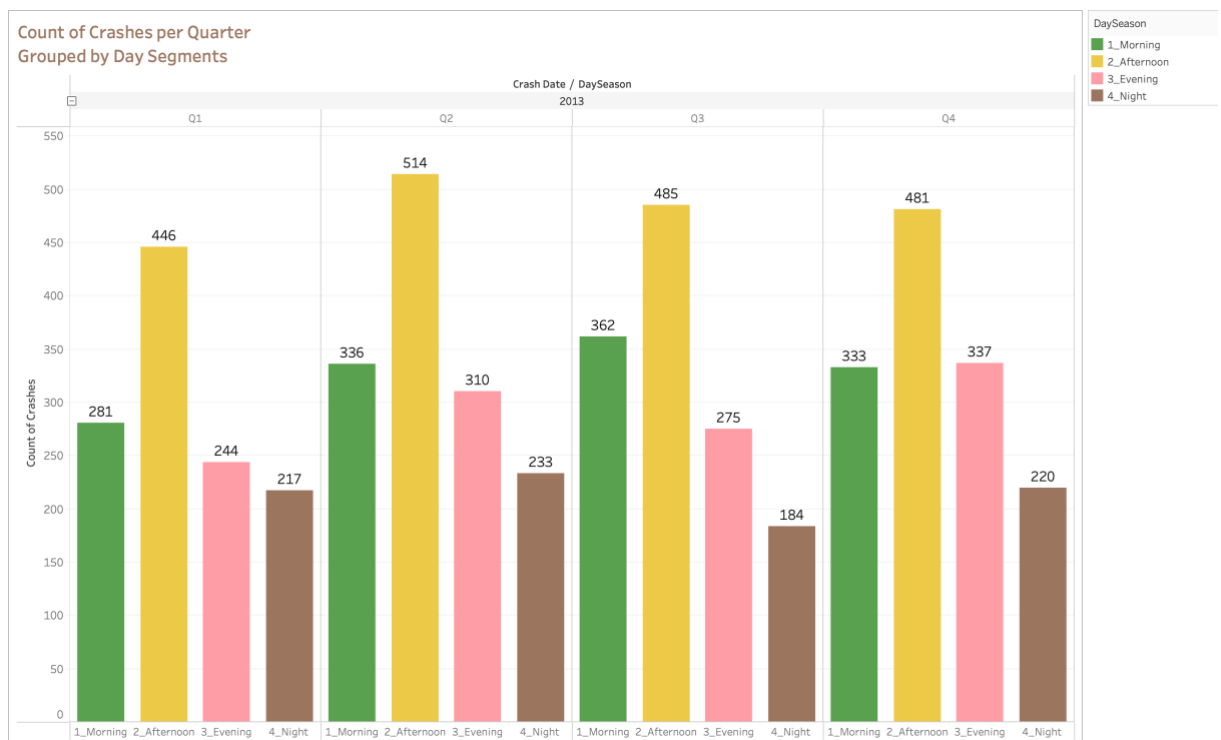
There seem to be no extreme monthly outlier in the aggregation of the count of crashes. Interestingly the highest number is in June when the days are the longest. Which might not be expected. An explanation might be that during the warm(er) season there is more traffic and thus a higher probability of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



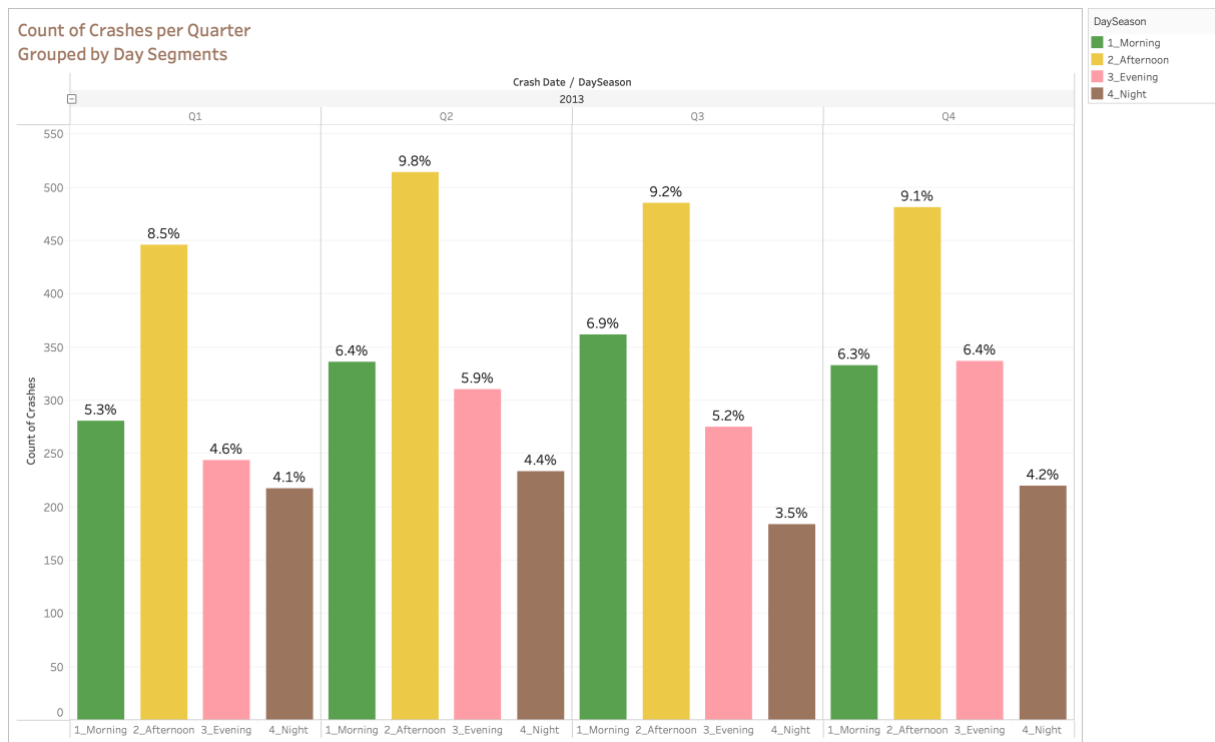
The highest number of crashes is in the afternoons and the lowest during the nights. This might be caused by the total traffic in general.



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.

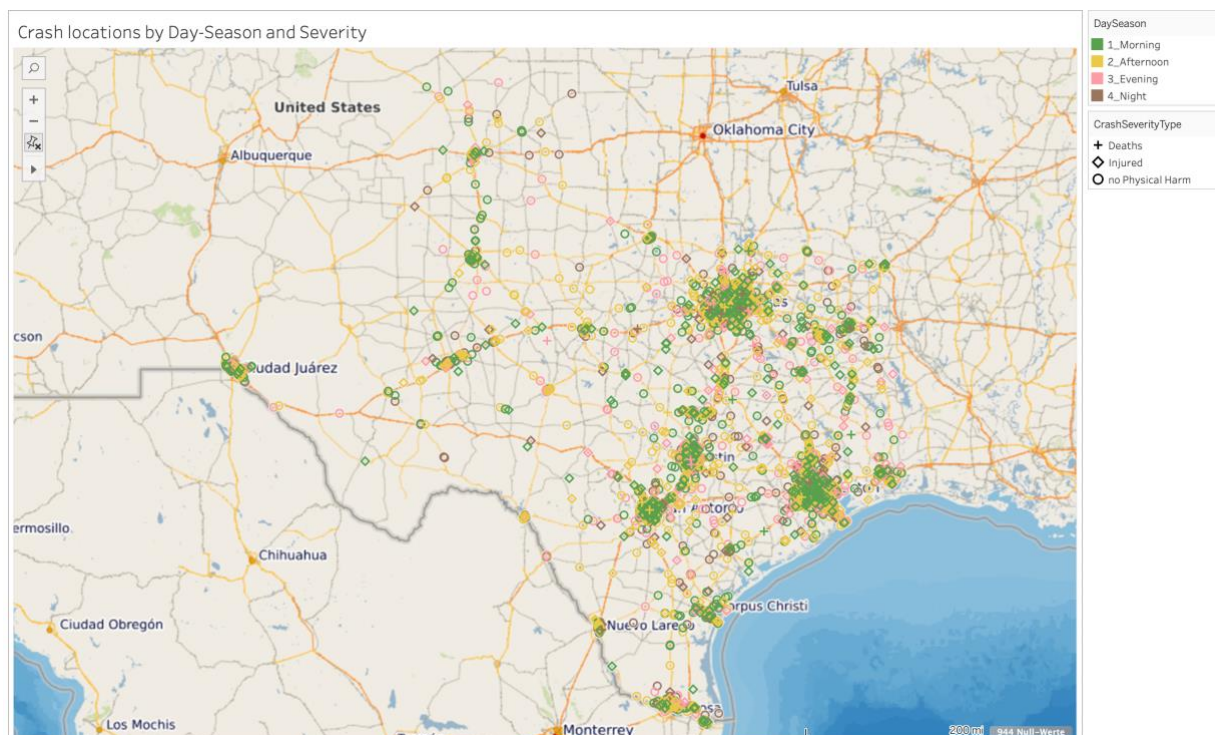


The figure shows the analysis of the number of accidents per quarter, grouped by day segments. At first glance, the distribution of values across all quarters looks very similar. On closer inspection, however, it is noticeable that the ratio between the incidents in the morning and in the evening varies 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).

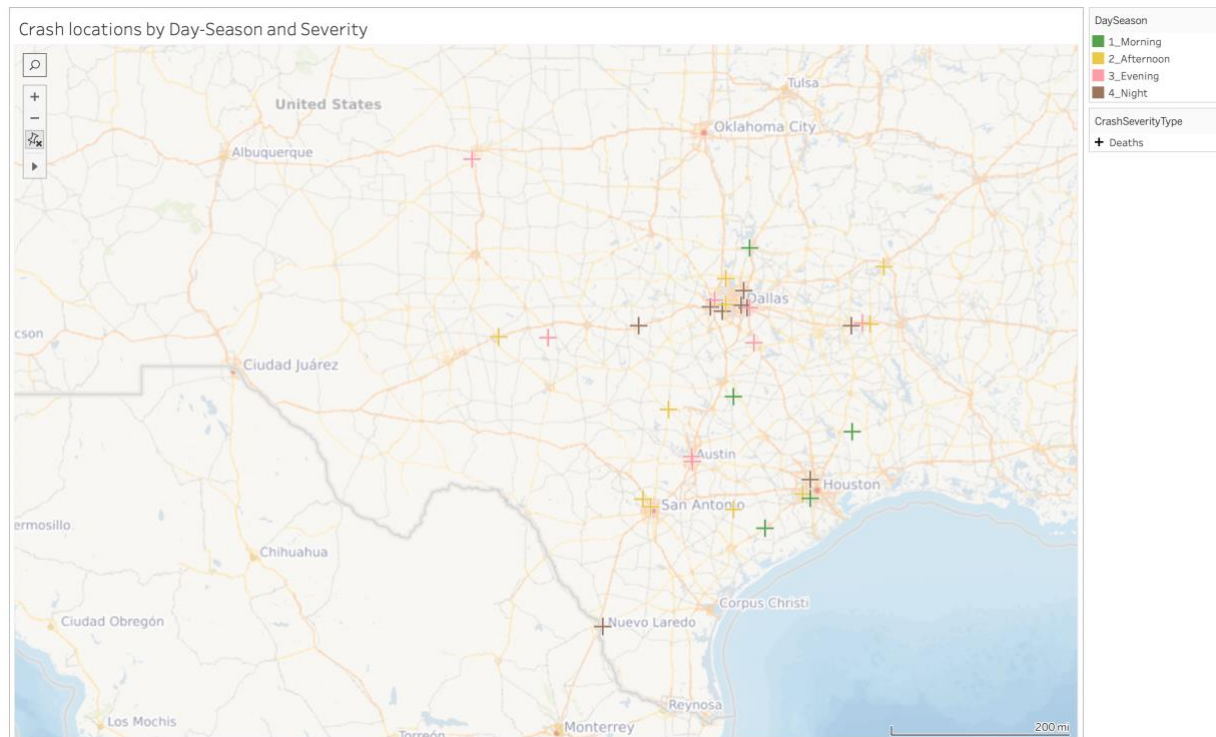


4.2 Spatial related analysis

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.

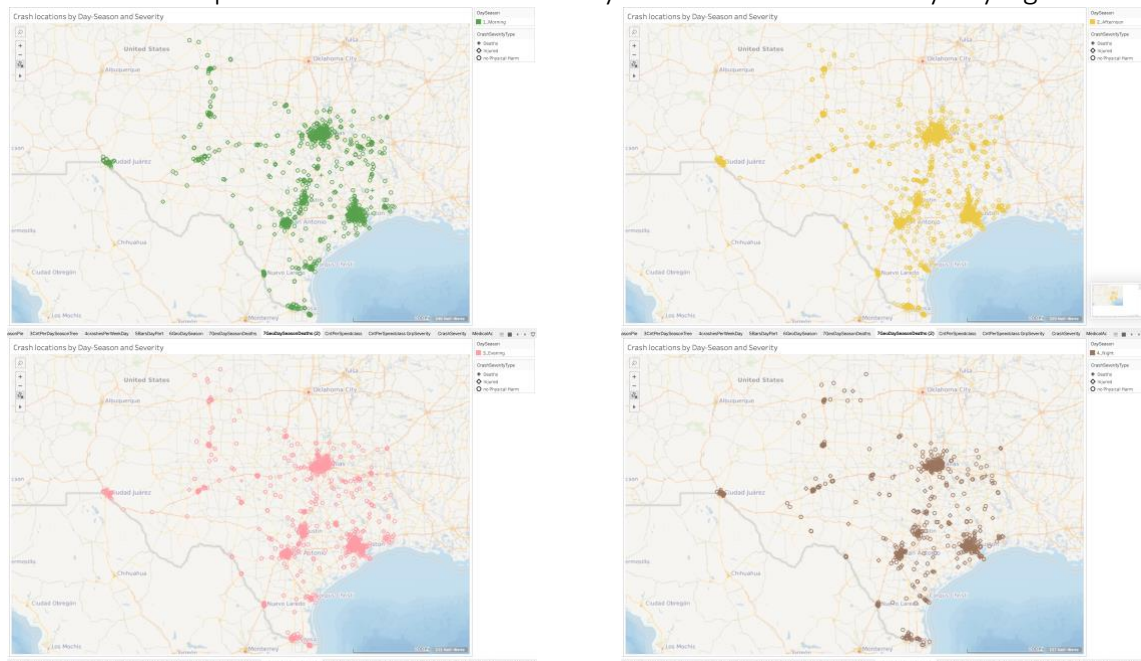


The map shows only all available locations of crashes with recorded deaths, classified by day segments. The highest densities are around urban areas. But also, on major streets or network connections crashes can be found.



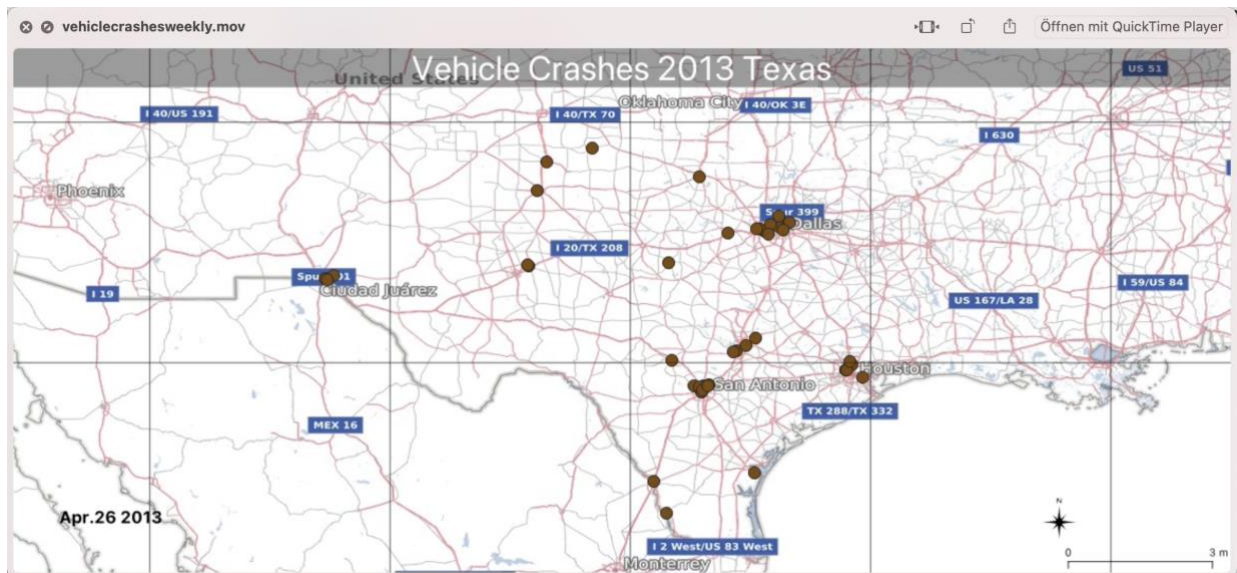
The map shows only the location of where crashes with recorded deaths occurred, classified by day segments. It is interesting to see, that deadly crashes at night are mostly found in cities or higher populated areas. On the other hand, deadly crashes in the morning are nearly all outside of cities. Deadly 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.

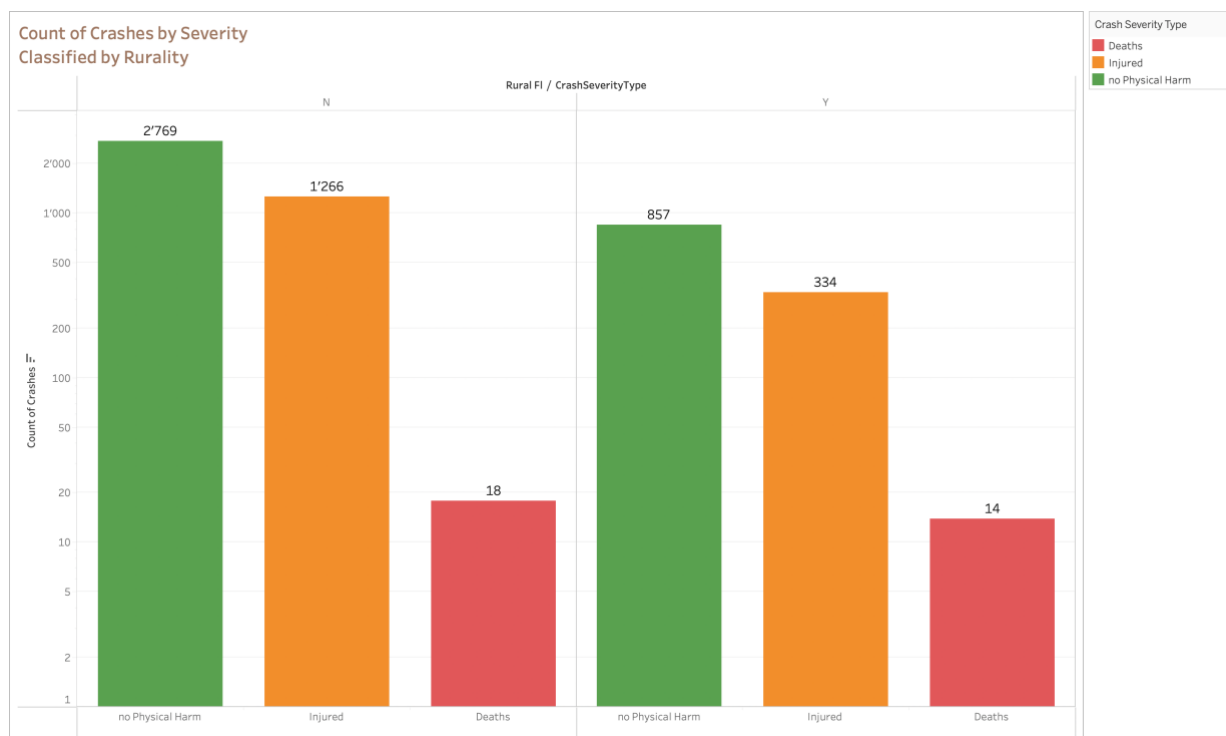


The comparison of the four maps show that for all day segments most of the crashes happened in the densely populated areas. In rural areas there are differences where the crashes occurred.

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.



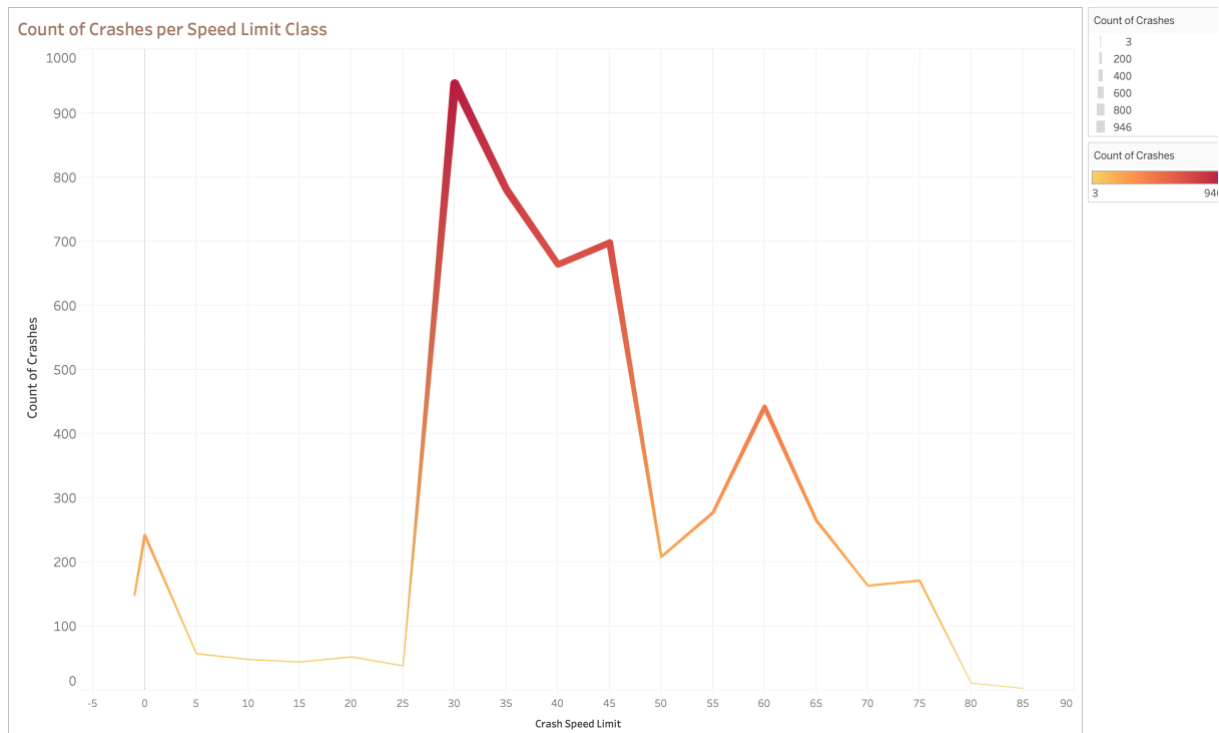
Finally, some information about rurality of the crashes:



The figure shows on the left non-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.

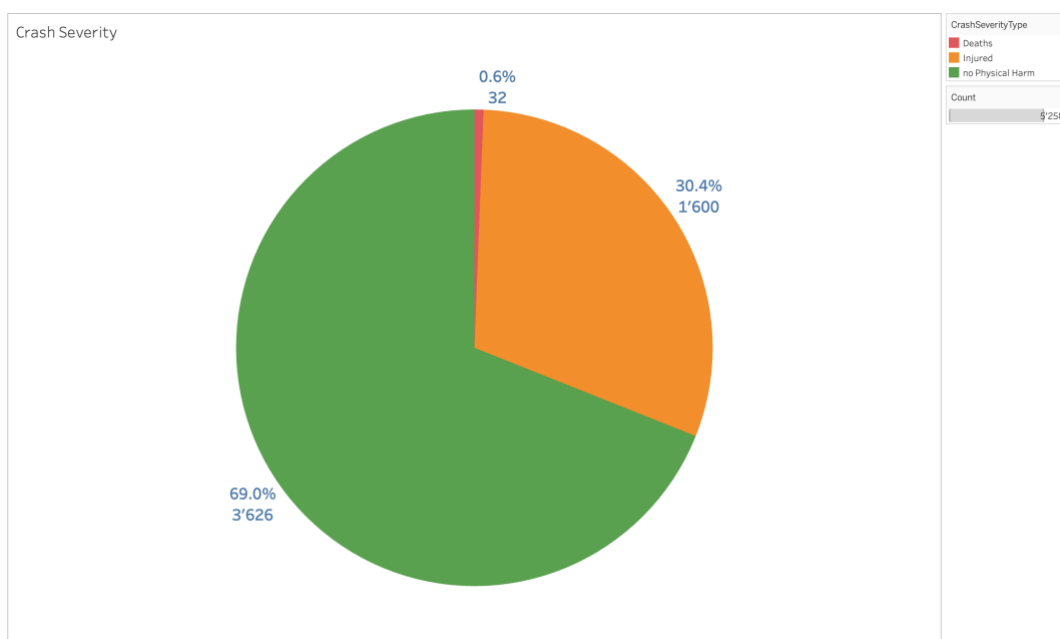
4.3 Severity Type and Speed Class related analysis

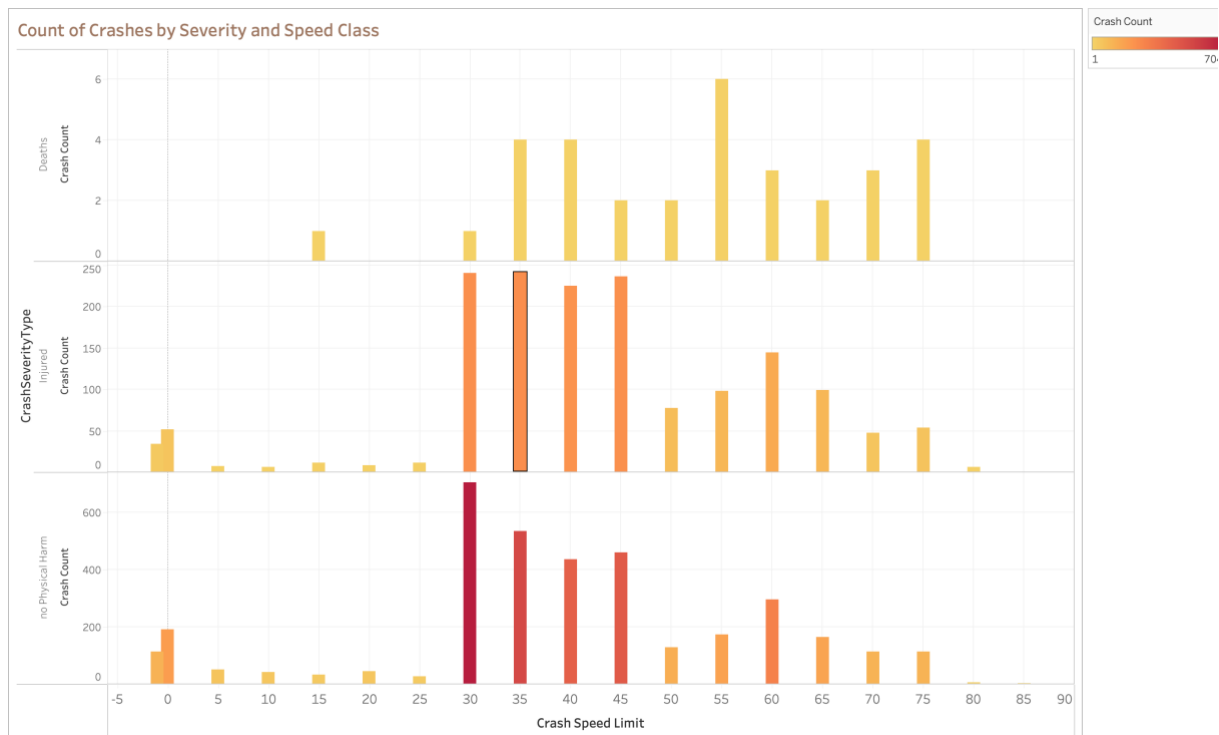
This paragraph shows different analyses based on the speed limit classes and the derived variable of severity of crash.



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:





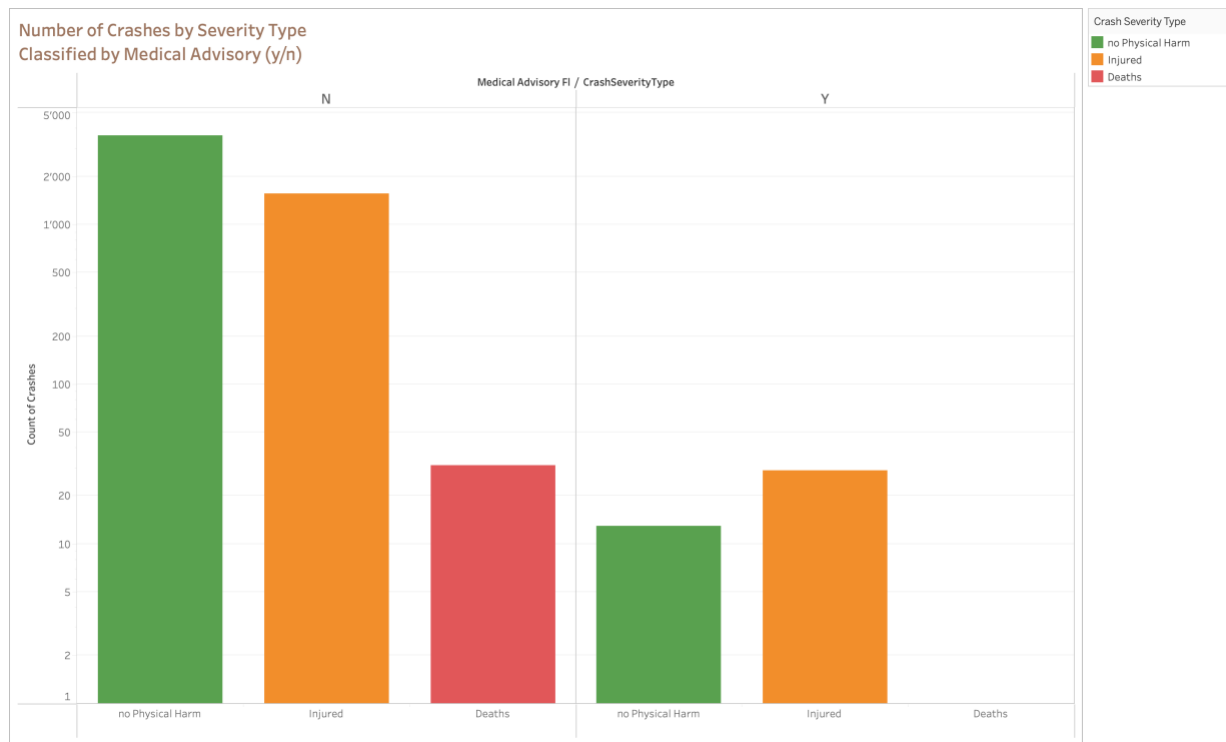
The figure shows the distribution of crash counts by speed limit classes and grouped by the severity of the crashes. As severity there have been three classes implemented (cf. 3.11.2). Mind you: For better comparability, the scales were individually adapted!

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.

4.4 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!