Università degli Studi di Catania

Dipartimento di Economia e Impresa

Corso di Laurea Magistrale in Data Science for Management

ROAD CRASHES IN CHICAGO

Sebastiano Florio
1000023232

Big Data Analytics

Anno Accademico 2020/2021

This report is about Chicago Traffic Crashes (from the Chicago Data Portal). In my analysis I want' to ask to the question about the average age of the person involved, the number and type of injuries and how this is influenced by the different conditions and type of street, for example. To this aim, I have simplified the original dataset, that contained various field not relevant for my business questions or not compiled for privacy or other motivations.

BUSINESS QUESTIONS

1. How the external condition influenced the number of accidents?

1.1 How the number of crushes increase along the year? What is the year with most number of crashes?

1.2 Which is the season with the higher number of crashes? What are the days of the week and the part of the

day with an higher possibility of crash? And which climate condition?

1.3 Did the number of defect of the road infleunced the number of incidents? Which are the street and Beat with more crashes?

1.5 Did the condition of device infleunces the number of incidents? And in which type of street occurs more incidents?

2. What is the most frequent cause of an incident?

2.1 How this cause influence the type of injury?

3. What characteristics in common have the people involved in an accident?

3.1 The use of safety equipment can save the life of the peoples? And what are the percentages?

3.2 Which is the physical condition most frequent in an accident and how the gravity of the physical condition influenced the type of injury?

3.3 The ejection can be mortal?

3.4 There are difference in the average age person involved in an accident related to the part of the day?

DATASETS

The first dataset is ***Traffic_Crashes_-_Crashes.csv*** that shows information about each traffic crash on city streets within the City of Chicago limits and under the jurisdiction of Chicago Police Department (CPD). About half of all crash reports, mostly minor crashes, are self-reported at the police district by the driver(s) involved and the other half are recorded at the scene by the police officer responding to the crash. Many of the crash parameters, including street condition data, weather condition, and posted speed limits, are recorded by the reporting officer based on best available information at the time, but many of these may disagree with posted information or other assessments on road conditions. If any new or updated information on a crash is received, the reporting officer may amend the crash report at a later time. A traffic crash within the city limits for which CPD is not the responding police agency, typically crashes on interstate highways, freeway ramps, and on local roads along the City boundary, are excluded from this dataset.
The original dataset contained 510.000 rows and 49 columns, but I've deleted some field for example the Chicago Police Department report number that for privacy is blank in some crashes or other field not useful for my purpose. I've have used this first dataset to analyse the condition of

the street of the crash and to give a summary visualization of the number of crushes in the different Department of Chicago (in the first dashboard) field that I've selected are 14 and are:

Field Name	Description	Type
CRASH_RECORD_ID	This number serves as a unique ID in this dataset	STRING
CRASH_DATE	Date and time of crash as entered by the reporting officer	DATE
DEVICE_CONDITION	Condition of traffic control device (es. no controls, functioning properly)	STRING
WEATHER_CONDITION	Weather condition at time of crash	STRING
TRAFFICWAY_TYPE	Type of road (es. one-way, not divided)	STRING
ALIGNMENT	Street alignment at crash location (es. curve, straight)	STRING
ROAD_DEFECT	Defect of the street	STRING
PRIMARY_CONTRIBUTORY_CAUSE	The cause of the crashes	STRING
MOST_SEVERE_INJURY	The most severe injury of that crash	STRING
STREET_NAME	Name of the street of the crashes	STRING
BEAT_OF_OCCURENCE	Chicago Police Department Beat ID.	NUMBER
LATITUDE	Latitude of crashes	NUMBER
LONGITUDE	Longitude of crashes	NUMBER

I've also selected the date from 2017 to 2020, because 2021 is still underway and it would not allow me to have a fair analysis of the accident situation in relation to other years.

```
YEAR([CRASH_DATE])>=2017 and YEAR([CRASH_DATE])
<2021
```

The second dataset is *geo_export_ff119ed6-d009-447e-a3ff-351d7ff33e32.shp* a geospatial dataset that contain the boundaries of each Police Department and the relative Beat ID, so that I can use this information to join with the previous dataset and have a visualization of the different crashes for different Department.

The third dataset is *Traffic_Crashes_-_People.csv* that contains information about people involved in a crash and if any injuries were sustained. Each record corresponds to an occupant in a vehicle listed in the Crash dataset. Some people involved in a crash may not have been an occupant in a motor vehicle, but may have been a pedestrian or bicyclist… A crashes can have multiple involved person and hence have a one to many relationship between Crashes and Person dataset.

I've used this dataset to see the conditions of the involved persons before a crashes, the results of a crashes and the safety measure adopted. To this aim the original dataset (that contained 30 fields and 1.13M rows) has been simplified to 9 fields :

Field Name	Description	Type
PERSON_ID	Serves as a unique ID for the person in this dataset	STRING
PERSON_TYPE	Type of roadway user involved in crash (es. driver, bicycle)	STRING
CRASH_RECORD_ID	The id of the crash useful to do a join	STRING
SEX	Sex of the involved person	STRING
AGE	Age of the involved person	NUMBER

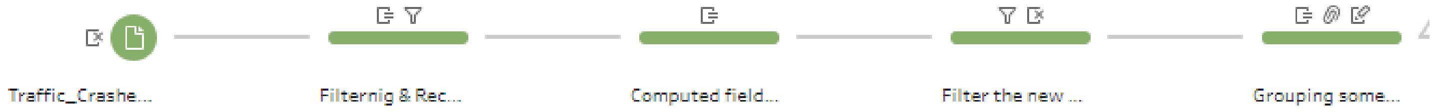
SAFETY_EQUIPMENT	Safety equipment used by vehicle occupant in crash, if any	STRING
EJECTION	Whether vehicle occupant was ejected or extricated from the vehicle as a result of crash	STRING
INJURY_CLASSIFICATION	Classification of injury (es. fatal, not evident)	STRING
PHYSICAL_CONDITION	Driver's apparent physical condition at time of crash, as observed by the reporting officer	STRING

*From the other field that can be important for my analysis I had to delete BAC_RESULT_VALUE, that is the value of the blood alcohol concentration because this value contains too many null values.

ETL

My ETL phase consist in filtering, compute new field, grouping, aggregate and join. I've done this phase with Tableau Prep.

PEOPLE CRASH DATASET



<u>1.Filtering</u>

I've managed the NULL values in the major number of cases by excluding that from the dataset, and also for the rows with the name Unknow and Other, that I've treated as null values, so I've excluded them. But in the dataset "Person" there are 2 value that can be Null or Unknow, this value are Ejection and Physical condition that for person that aren't in motor vehicle or that are pedestrian can be omitted.

<u>2.Computed Field</u>

I've recomputed the field **EJECTION** on the base of the present field EJECTION in a way that if the person is a pedestrian the null value can be admitted and I've substituted it with a NONE value, to distinguish them for the null values of the drivers, that I've interpretated as an unknow values and for this reason I've excluded them.

```
IF ([PERSON_TYPE]=="PEDESTRIAN") and
ISNULL([EJECTION]) then "NONE"
ELSE [EJECTION]
END
```

I've also recomputed the field **AGE**, because there were rows in which person type was "Driver" and age<16, so I've considered this rows as rows with a null value on the field AGE, and then I've removed them.

```
IF [PERSON_TYPE]=="DRIVER" AND AGE<16 THEN NULL
ELSEIF AGE<0 THEN NULL
ELSE [AGE]
END
```

3.Grouping

I’ve grouped the value of the field **SAFETY_EQUIPMENT** In only 4 values, so that I can easier analyse if the safety equipment is used or not. I’ve groped also the value of **INJURY_CLASSIFICATION** in 4 categories: FATAL,INCAPACITATING,NON INCAPACITATING and NOT EVIDENT. And also for **PHYSICAL_CONDITION** in which have collected the values of the type alcohol/drugs in a unique value and I’ve also removed the values that have a different meaning from the others

CRUSHES DATASET

1.Filtering

I’ve managed all the NULL values and the values OTHER, UNKNOW by excluding that from the dataset.

2.Computed Field

I’ve computed three field about the Date, that are:

SEASON

```
IF MONTH([CRASH_DATE])= 12 THEN 'winter' ELSEIF
MONTH([CRASH_DATE]) < 3 THEN 'winter' ELSEIF
MONTH([CRASH_DATE]) < 6 THEN 'spring' ELSEIF
MONTH([CRASH_DATE]) < 9 THEN 'summer' ELSEIF
MONTH([CRASH_DATE]) < 12 THEN 'autumn' END
```

DAY PART

```
IF DATEPART('hour',[CRASH_DATE])< 12 AND
DATEPART('hour',[CRASH_DATE])>=05 THEN 'Morning'
ELSEIF DATEPART('hour',[CRASH_DATE])>= 12 AND
DATEPART('hour',[CRASH_DATE])<17 THEN 'Afternoon'
ELSEIF DATEPART('hour',[CRASH_DATE])>= 17 AND
DATEPART('hour',[CRASH_DATE])<21 THEN 'Evening'
ELSE 'Nigh'
END
```

DAY TYPE

```
IF DATEPART('weekday',[CRASH_DATE]) = 1 OR
DATEPART('weekday',[CRASH_DATE]) = 7 THEN "weekend"
ELSE "weekday" END
```

3.Grouping

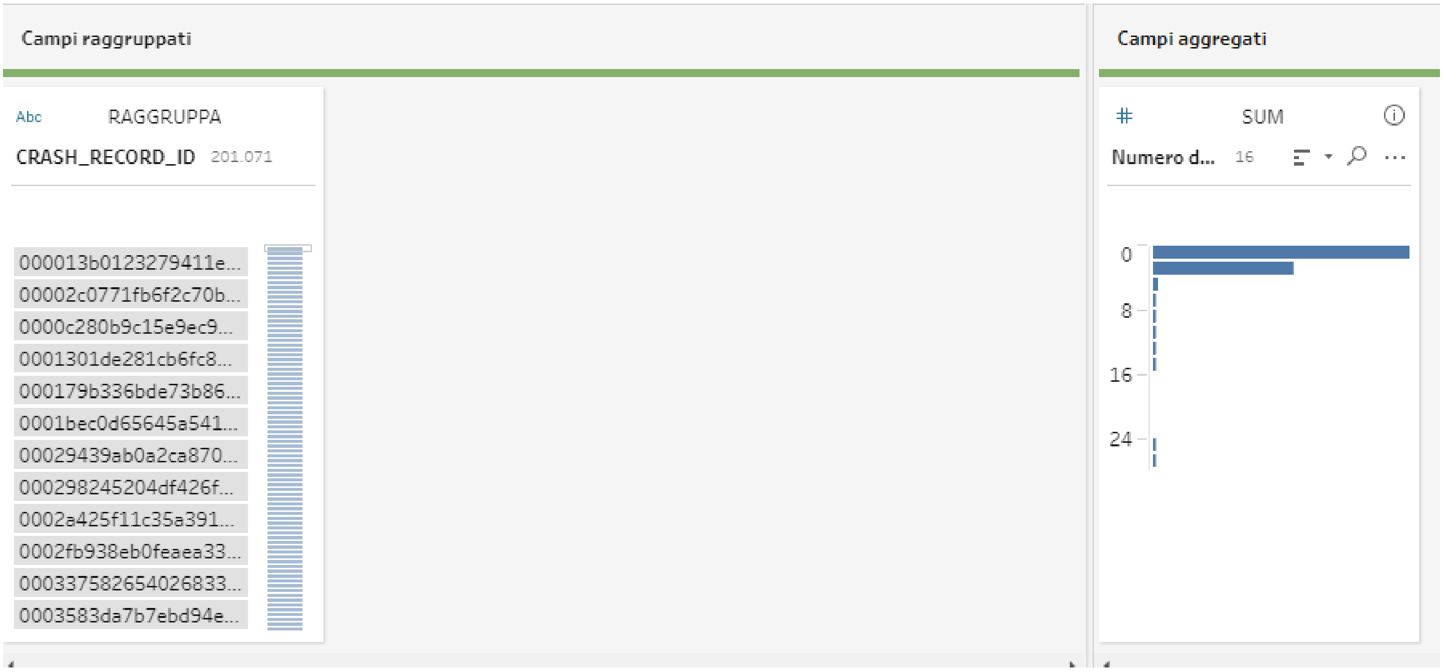
I've grouped the value of the field **WEATHER CONDITION** so that we have to distinguish only from blowing, clear, cloudy, rain, snow climatic conditions. **DEVICE CONDITION** in functioning improperly, properly, no controls, not functioning. And other grouping in **TRAFFICWAY TYPE**, **ALIGNMENT** (in curve, straight), **ROAD DEFECT**, **PRIMARY_CONTRIBUTORY_CAUSE** and **MOST_SEVERE_INJURY**.

GEOSPATIAL DATASET

I've only changed the type of the field **beat_num** from string to number so that I can perform a join with the field BEAT_OF_OCCURENCE of the Crushes dataset , because the string type start with 0 and the integer not, so I've repairred selected the type number to perform the join. But later I've reused the string type to use that field as a dimension to indicate the different District Police Departments of Chicago.

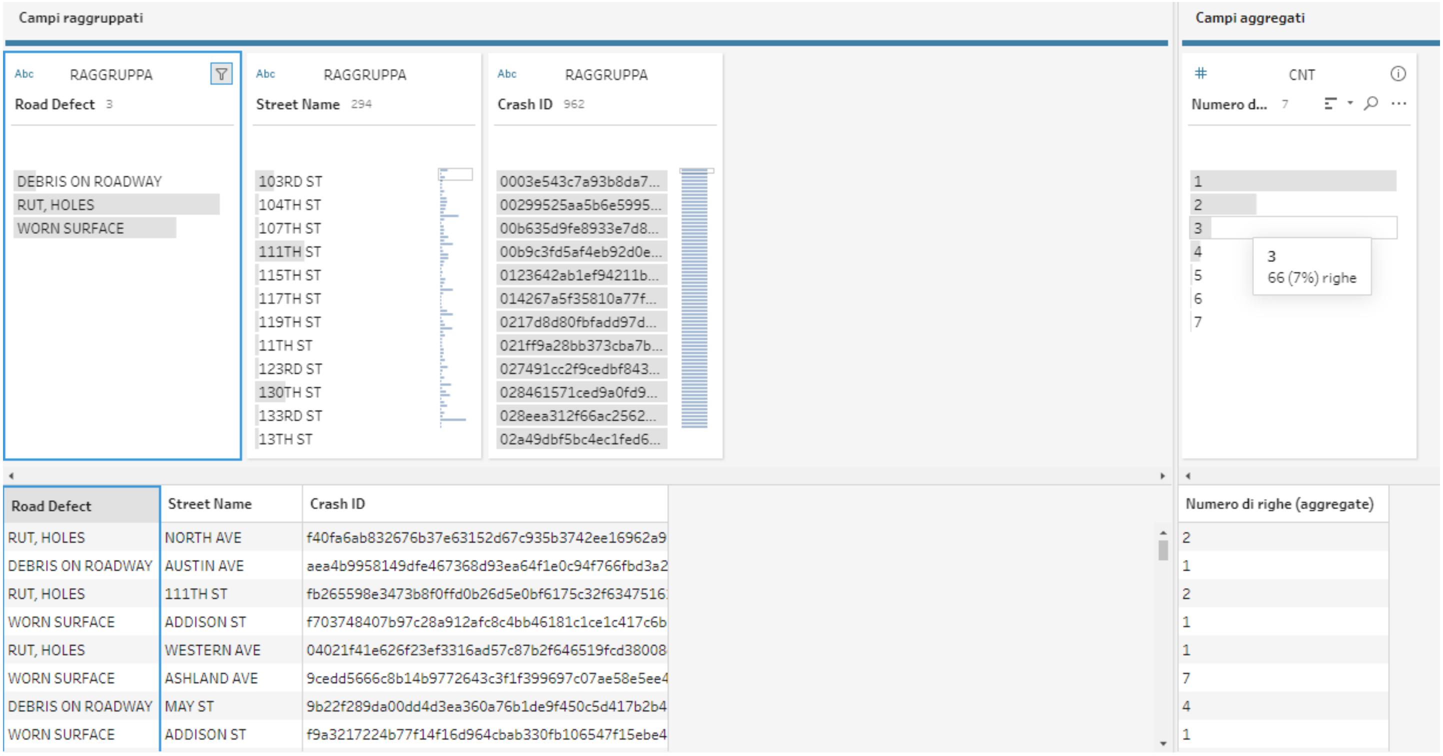
AGGREGATE DATASET

The first aggregate dataset was in the original Person dataset in which I've computed an aggregated field **INVOLVED** on the bases of the unique crash id, and I've count the number of occurrence of each crash id, this number represent the number of people involved in the crash identified by that id.

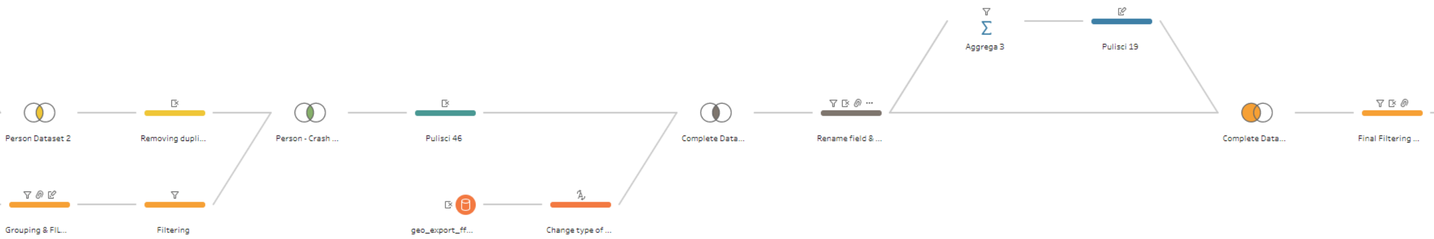


The second aggregate was on the Complete Dataset 1 from the third join, in which I've grouped for street name, crashes id to avoid the problem to count more defect for a street for a same crushes if in this crushes there are more people, and road defect in which I've considered only the negative occurrences (debris on roadway, rut, holes, worn surface) and to do this I've performed a filtering on this field.

```
NOT ((([Road Defect] == "NO DEFECTS") AND NOT
(ISNULL([Road Defect]))))
```



JOINS



<u>Person Dataset 2 :</u>

The first join was between the dataset resulting from the aggregate operations that contains the Crash ID and the Number of person involved and the original Person Dataset with an Inner Join in order to retain only the rows that matches. Then I've removed the only field that was in each of the two original dataset, that is Crash_record, that in the resulting join dataset appear two times.

<u>Person – Crash Dtaset :</u>

The second join was between the Person Dataset 2 (that contain the information of each person involved in a crash) and the dataset that contains information about that crashes. Because is possible that there are crashes in which no person is registered or person involved in crashes not in the period 2017-2020, I've performed also in this case the Inner Join. Also in this case I've removed the field Crash_record, that in the resulting join dataset appear two times.

<u>Complete Dataset 1 :</u>

The third join was between the Person – Crash Dataset and the dataset that contain the geospatial coordinates of each police departments, and I've computed this an inner join between the field BEAT_OF_OCCURENCES and beat_num of the two dataset.

<u>Complete Dataset 2 :</u>

The last join was between the dataset resulting from the aggregate operations that contains Crash ID , typology of defect, street name and number of defect and the Complete Dataset 1 with an Left Join in favour of the Complete Dataset 1 based on the clausule of Crash ID, because in the dataset aggregate we take in count only the crushes occurred in street with defect, but in the final dataset I want also crashes occurred in street with no defects . Also in this case then I've removed the duplicates and substituted the value null present in the field occurrence defects after the join with a numeric value 0.

FINAL DATASET

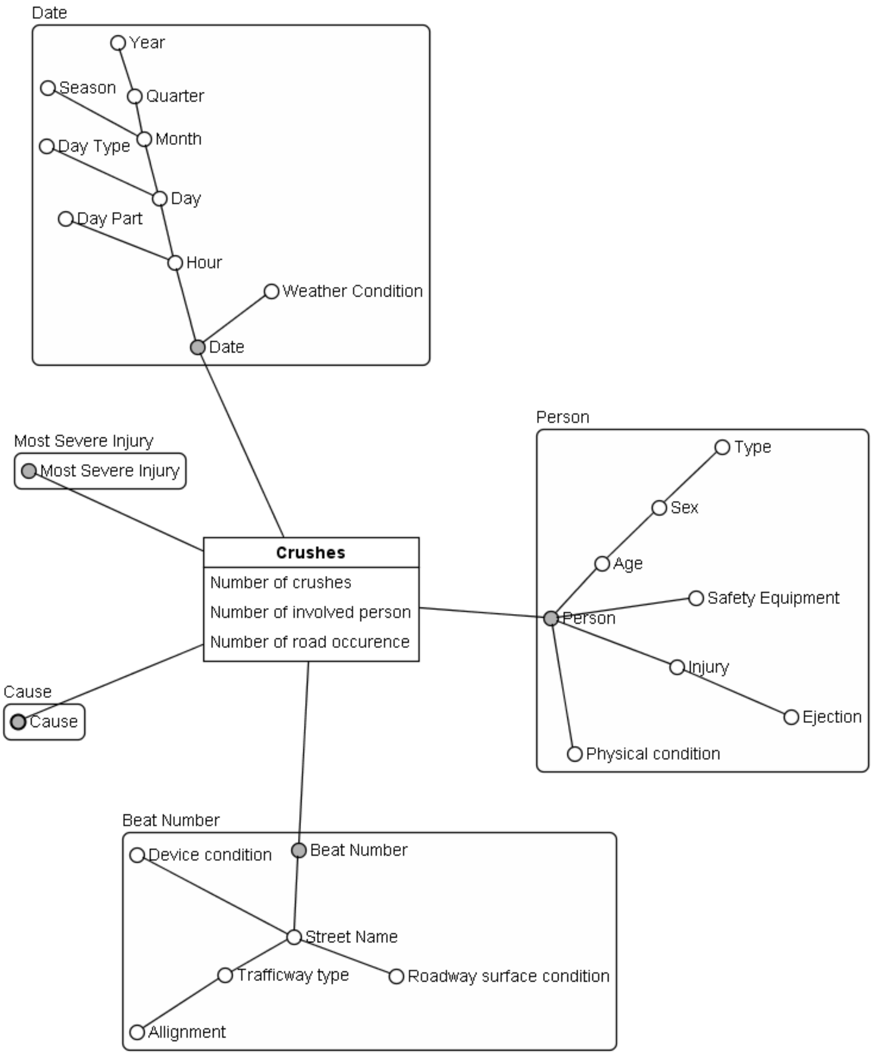
Occurences defect	Involved Person	Ejection	Day Type	Day Part	Season	Date	Device Condition	Trafficway Type	Weather Condition	Allignment	Roadway Surface ...
0	1	NONE	weekday	Nigh	spring	13/04/2020 22:50:00	FUNCTIONING PROPE...	DIVIDED	CLEAR	STRAIGHT	DRY
0	2	NONE	weekend	Evening	winter	23/02/2020 18:15:00	FUNCTIONING PROPE...	NOT DIVIDED	CLEAR	STRAIGHT	DRY
0	2	NONE	weekday	Afternoon	spring	05/05/2020 12:20:00	FUNCTIONING PROPE...	MORE WAY	RAIN	STRAIGHT	WET
0	2	NONE	weekday	Afternoon	spring	05/05/2020 12:20:00	FUNCTIONING PROPE...	MORE WAY	RAIN	STRAIGHT	WET
0	1	NONE	weekend	Morning	spring	02/05/2020 11:00:00	NO CONTROLS	DIVIDED	CLEAR	STRAIGHT	DRY
0	1	NONE	weekend	Evening	winter	23/02/2020 20:53:00	FUNCTIONING PROPE...	DIVIDED	CLEAR	STRAIGHT	DRY
0	1	NONE	weekday	Morning	spring	07/05/2020 06:20:00	FUNCTIONING PROPE...	MORE WAY	CLEAR	STRAIGHT	DRY
0	1	NONE	weekday	Morning	spring	07/05/2020 08:30:00	NO CONTROLS	PARKING LOT	CLEAR	STRAIGHT	DRY
0	2	NONE	weekday	Morning	spring	07/05/2020 09:08:00	NO CONTROLS	MORE WAY	CLEAR	STRAIGHT	DRY
0	2	NONE	weekend	Nigh	winter	08/02/2020 01:35:00	FUNCTIONING PROPE...	DIVIDED	CLEAR	STRAIGHT	DRY
0	2	NONE	weekend	Nigh	winter	08/02/2020 01:35:00	FUNCTIONING PROPE...	DIVIDED	CLEAR	STRAIGHT	DRY
0	2	NONE	weekend	Nigh	winter	08/02/2020 02:23:00	FUNCTIONING PROPE...	NOT DIVIDED	CLEAR	STRAIGHT	DRY
0	2	NONE	weekend	Nigh	winter	08/02/2020 02:23:00	FUNCTIONING PROPE...	NOT DIVIDED	CLEAR	STRAIGHT	DRY
0	2	NONE	weekend	Nigh	winter	08/02/2020 03:02:00	NO CONTROLS	NOT DIVIDED	CLEAR	STRAIGHT	DRY
0	1	NONE	weekend	Nigh	winter	08/02/2020 03:57:00	FUNCTIONING PROPE...	NOT DIVIDED	CLEAR	STRAIGHT	DRY

Road Defect	Cause	Street Name	Involved Units	Most Severe Injury	Latitude	Longitude	Person ID	Person Type	Crash ID	Sex	Age	Safety Equipment
NO DEFECTS	IMPROPER DRIVE	87TH ST	2	NOT EVIDENT	41,736044	-87,653404	0871921	DRIVER	af84fb5c8d996fcd3a...	M	37	USED
NO DEFECTS	IMPROPER DRIVE	DEVON AVE	2	NOT EVIDENT	41,997662	-87,700128	0848601	DRIVER	f25f09798b51603bde...	F	34	USED
NO DEFECTS	IMPROPER DRIVE	LOCKWOOD AVE	2	NOT EVIDENT	41,931414	-87,759020	0879679	DRIVER	49336aaca932f7935c...	M	26	USED
NO DEFECTS	IMPROPER DRIVE	LOCKWOOD AVE	2	NOT EVIDENT	41,931414	-87,759020	0879680	DRIVER	49336aaca932f7935c...	F	64	USED
NO DEFECTS	IMPROPER DRIVE	LAKE SHORE DR	2	NOT EVIDENT	41,791420	-87,580148	0880165	DRIVER	343f56c14e824e3e10...	M	61	USED
NO DEFECTS	IMPROPER DRIVE	MADISON ST	2	NOT EVIDENT	41,880499	-87,735596	0848686	DRIVER	91069a7cfa1cec819c...	M	32	USED
NO DEFECTS	IMPROPER DRIVE	COTTAGE GROVE AVE	2	NOT EVIDENT	41,780317	-87,606075	0880466	DRIVER	c210e677589588e64...	M	42	USED
NO DEFECTS	IMPROPER DRIVE	MERCHANDISE MART ...	2	NOT EVIDENT	41,888074	-87,634955	0880482	DRIVER	0ea6720e462e7d77d...	M	63	NOT USED
NO DEFECTS	IMPROPER DRIVE	CENTRAL AVE	2	NOT EVIDENT	41,947751	-87,766674	0880486	DRIVER	dcdd101792d3f2dd4...	M	49	USED
NO DEFECTS	IMPROPER DRIVE	PULASKI RD	2	NOT EVIDENT	41,749663	-87,721771	0837915	DRIVER	fef26188db886602ea...	M	44	USED
NO DEFECTS	IMPROPER DRIVE	PULASKI RD	2	NOT EVIDENT	41,749663	-87,721771	0837916	DRIVER	fef26188db886602ea...	M	70	USED
NO DEFECTS	IMPROPER DRIVE	HOWARD ST	2	NOT EVIDENT	42,019420	-87,690110	0837921	DRIVER	8d84e1048d01accf40...	M	36	USED
NO DEFECTS	IMPROPER DRIVE	HOWARD ST	2	NOT EVIDENT	42,019420	-87,690110	0837922	DRIVER	8d84e1048d01accf40...	F	49	USED
NO DEFECTS	IMPROPER DRIVE	SHERIDAN RD	2	INCAPACITATING	41,996052	-87,655423	0837935	DRIVER	a8405fd5fb5ec1a8b8...	M	62	NOT USED
NO DEFECTS	IMPROPER DRIVE	55TH ST	2	NOT EVIDENT	41,792945	-87,740197	0837948	DRIVER	944a74a0fd1cd3dc5f...	M	23	USED

Injury	Physical Condition	Beat	Geometry
NOT EVIDENT	NORMAL	613	Polygon
NOT EVIDENT	NORMAL	2412	Polygon
NOT EVIDENT	NORMAL	2514	Polygon
NOT EVIDENT	NORMAL	2514	Polygon
NOT EVIDENT	NORMAL	331	Polygon
NOT EVIDENT	NORMAL	1113	Polygon
NOT EVIDENT	NORMAL	312	Polygon
NOT EVIDENT	NORMAL	1831	Polygon
NOT EVIDENT	NORMAL	1633	Polygon
NOT EVIDENT	NORMAL	833	Polygon
NOT EVIDENT	NORMAL	833	Polygon
NOT EVIDENT	NORMAL	2411	Polygon
NOT EVIDENT	NORMAL	2411	Polygon
INCAPACITATING	NORMAL	2433	Polygon
NOT EVIDENT	NORMAL	813	Polygon

DIMENSIONAL FACT MODEL

The Dimensional Fact Model (DFM), proposed by Golfarelli & Rizzi is a graphical conceptual model that effectively supports the conceptual design of a data mart and creates an environment in which queries may be formulated intuitively. The conceptual representation generated by the DFM consists of facts, measures, dimensions, and hierarchies. I identified Crushes as the focus of interest of my analysis and Number of crashes (a field computed on Tableau) , number of person involved in each crashes, number of vehicle involved in each as the measures.



I’ve done the visualization of the DFM with Business Intelligence Modeler. These are the hierarchies of the model:

Date->Hour->Day->Month->Quarter->Year

Date->Wheather Condition

Date->Hour->Day part

Date->Hour->Day->Day type

Date->Hour->Day->Month->Season

Person->Age->Sex->Type

Person->Safety Equipment

Person->Injury->Ejection

Person->Physical condition

Beat Number->Street name->Defect

Beat Number->Street name->Trafficway type->Allignment

Beat number->Street name->Device condition

Beat number->Street name->Roadway surface condition

In tableau I’ve computed a calculated numerical field that is the measure of my DFM, Numebr of Crushes as by using an aggregate function countd, to count the number of distinct crushes id, related to a single crush.

```
COUNTD([Crash ID])
```

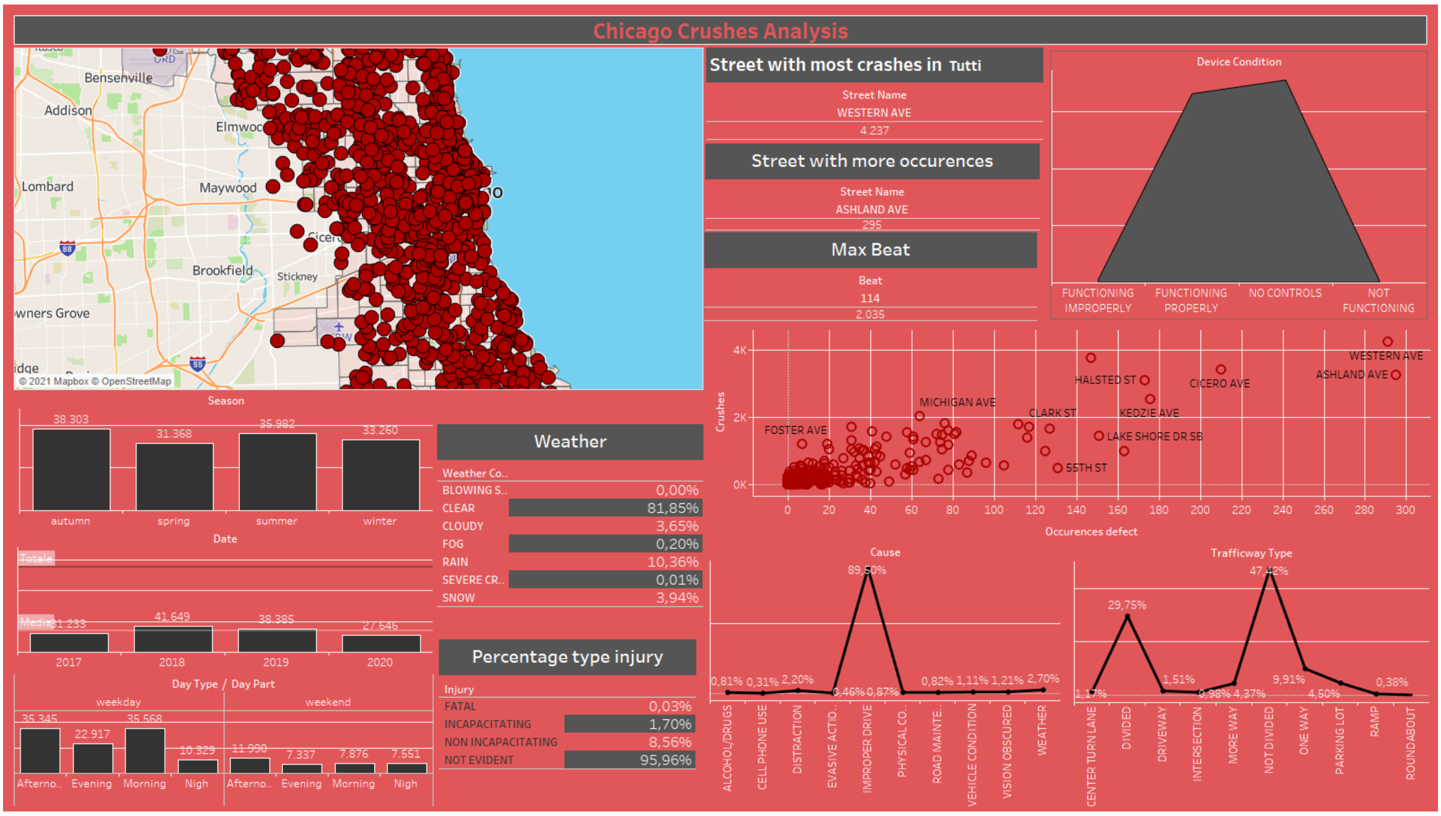
DASHBOARD

In order to answer the proposed business questions I’ve built two dashboards in the first dashboard I analyse
the number of crushes per area of Chicago, paying attention on the composition and condition of the streets, but also the number of crushes related to different year, different Season and Day Type and the related most common number of units involved.
In this dashboard every sheet can be used as a filter and a big space is taken by the maps, in which we can see the boundaries of each police department, the related number of crushes and the position of the crash in the map.
It contains the following sheets: *Date, Day, MAP, Street with more crushes, Street with ore occurrences, Season, Max Beat, Cause-TrafficWay,. Num units, Cake Wwather, Surface, Device Conditions.*

The second dashboard shows the information about the person involved in the crushes, paying attention to the average age, the relation between the use of safety equipment (and also ejection or cause of crash) and the typology of injury. Contains also the number of the typology of injury and all can be analysed in relation to the filter Sex, Person Type and Date that are all slider bar.

It contains the following sheets: *Avg Age, Cause injury, Person Type- Injury, physical condition, Safety Equipment, Involved Persons, Numbers.*

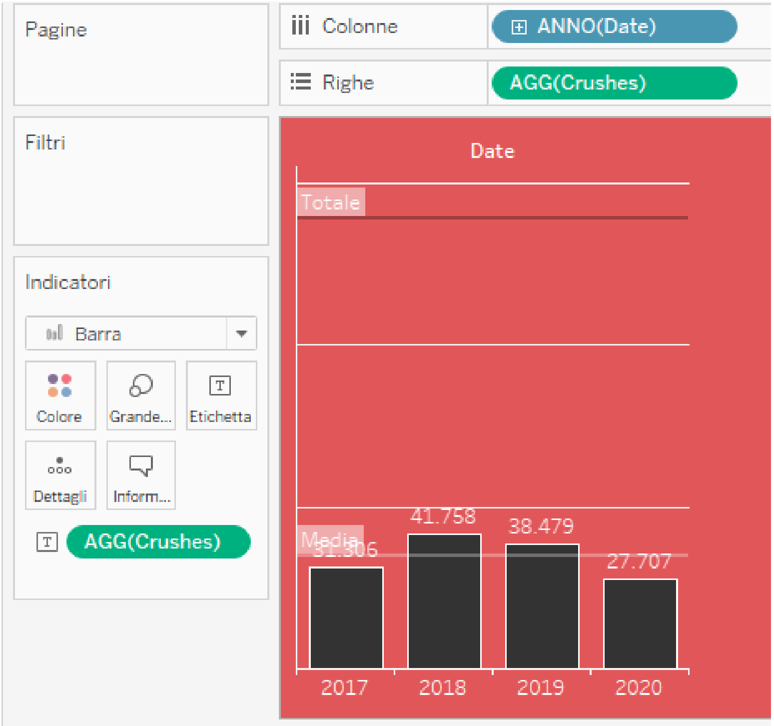
The principal colours of the dashboard is the red, because is related to a message of danger. In addition I’ve used only two different typologies of black and the white for the text. So that I can have a neat and clear dashboard.



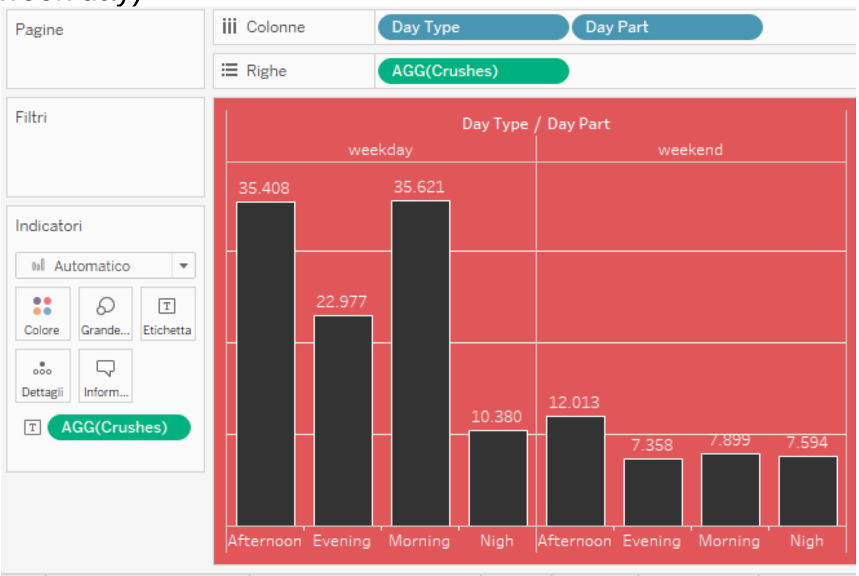
SHEETS

I've used 3 sheet with a similar composition and choice of graphics, because these three sheets are about the analysis of the number of crushes according to different aggregation of the sane dimension (Date).

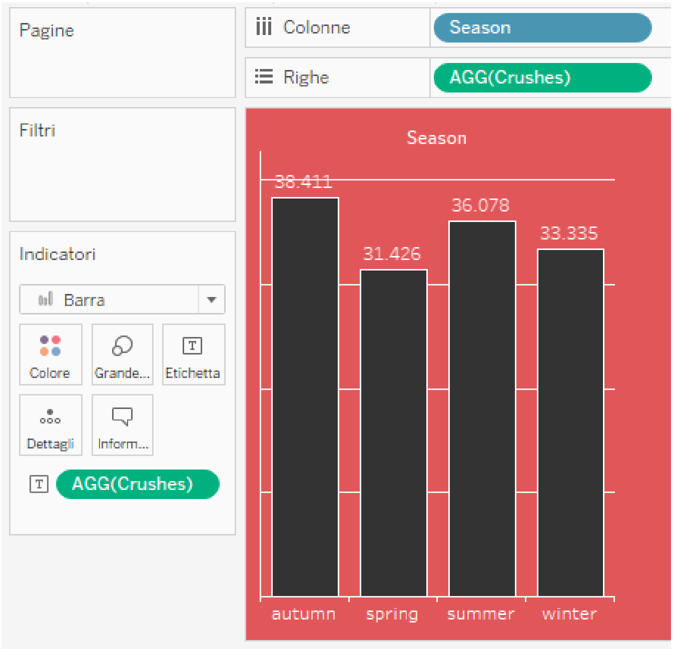
The **Date** sheet show the different number of crushes per year. So I've put in the columns the lower level of detail of the Date, that is ANNO(Date) and in the rows the relative number of crushes. I've used an histogram that immediately shows the differences of crushes between the year and I've re-moved the axes of number of crashes to a more simple visualization, we can visualize the number of crushes thanks to the label over each bin that indicate the total number of crashes for that year. I've also added two reference line so that we can see the average number of crashes in these 4 years and the total number of crashes.



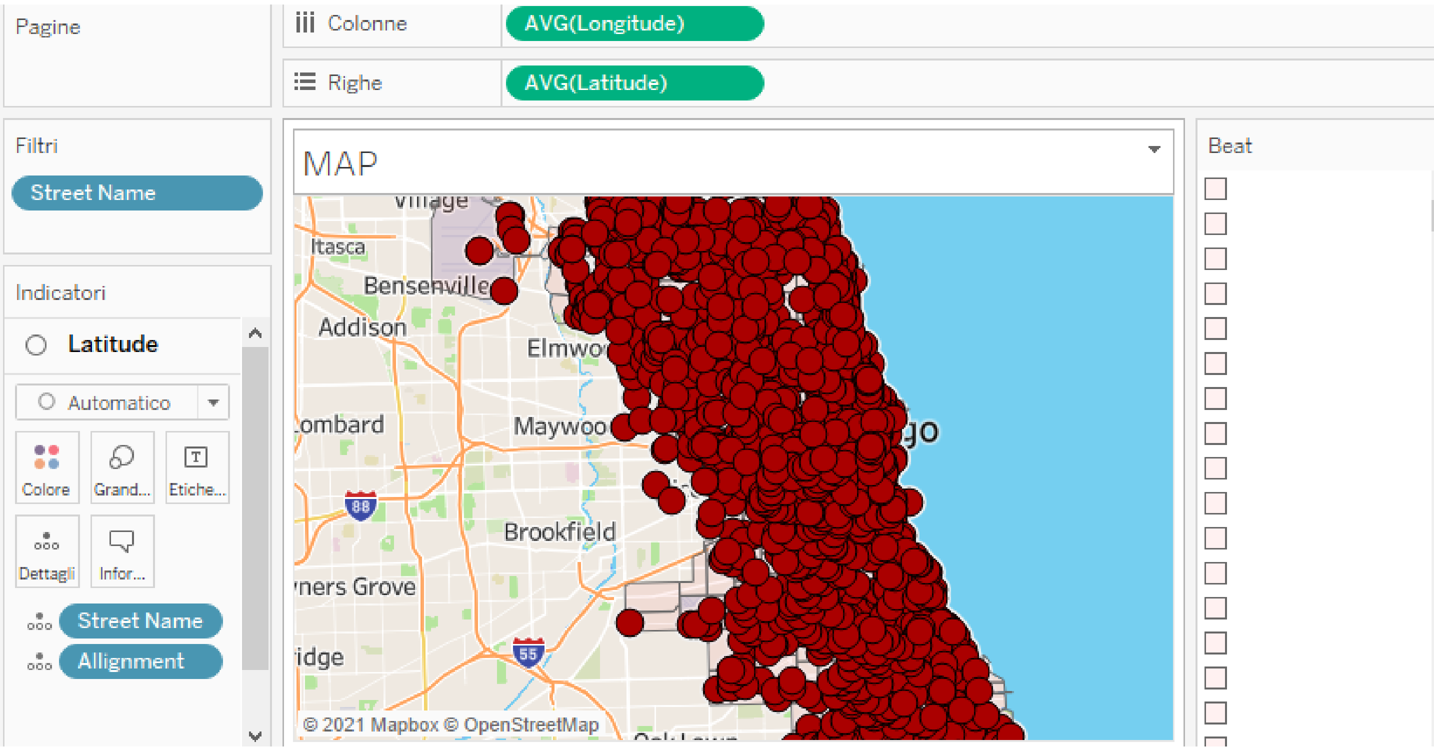
The **day** sheet is similar to the previous sheet but with the difference that I've shown by doing a Drill-Down an higher level of detail with Day part (afternoon, evening, morning, night) for the two different day type (weekend, week day).



Finally **Season.**



For the sheet map I've used the geospatial field and I've used the functionality of tableau to show also the street of that map, because I was interested not only to the beat number of each department id of Chicago but also on the position of the street, so a clear visualization of the street of Chicago is helpful. I've used for the different geospatial boundaries of the district the same of colour, with an opacity of the 5% because I'm not only interested in the number of crushes per beat (that I've inserted in details) but also on the position of the crashes, so is important to see also the map behind. The colour for the beat is again red, to indicate danger in that zones and also to specify to what zone the dataset is referred. On this level , I've added another level of indicator, but with the longitude and latitude of each street and also the name of the street.



In the sheet **Street with more crushes** is an information in a text form that display the name of the street with more occurrences of crushes. To retrieve this information a calculated field was created, which associates to the maximum number of crushes for street.

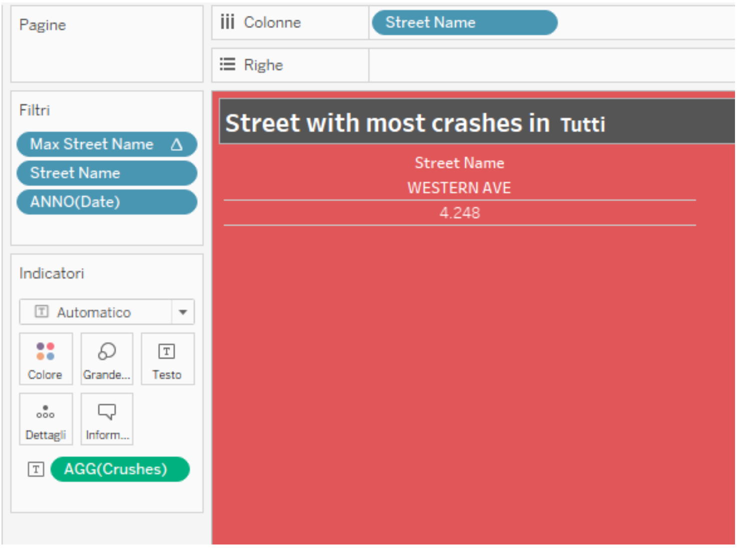
Max #Crushes

I risultati sono calcolati con Tabella (orizzontale).

```
WINDOW_MAX( ( [Crushes] ) )
```

Max Street Name

```
IF ( [Crushes] ) == [Max #Crushes]
THEN MAX( [Street Name])
END
```



The same is done for **street with more occurrences**, that indicate the name of the street with more occurrences of defect that have caused a crash reported by the agent:

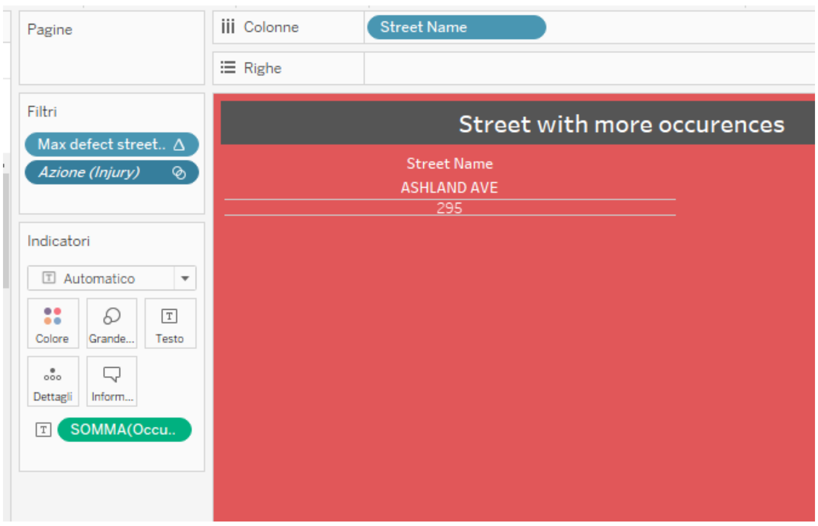
Max #Occurences

I risultati sono calcolati con Tabella (orizzontale).

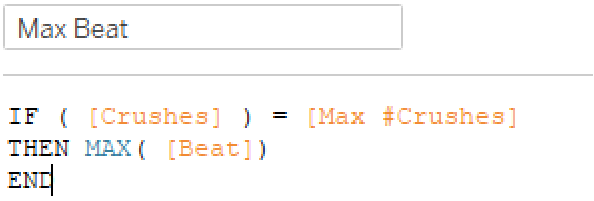
```
WINDOW_MAX(SUM([Occurences defect]))
```

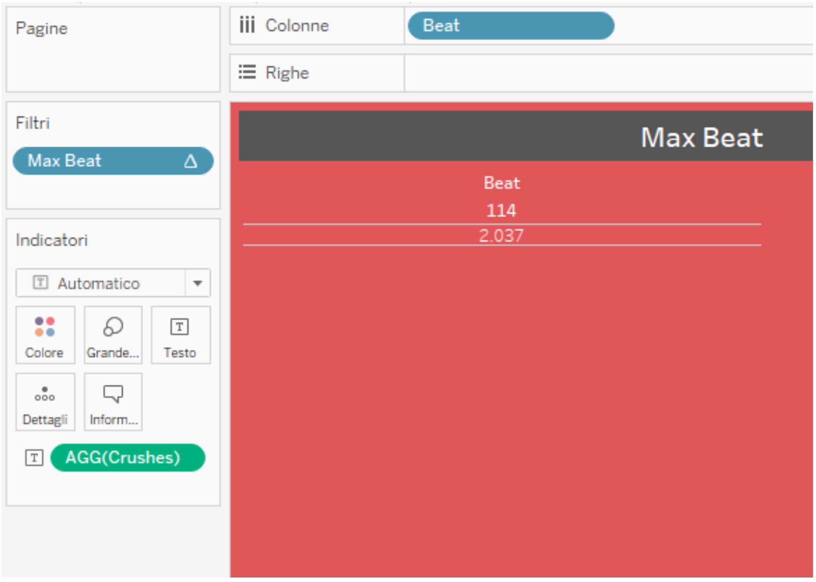
Max defect street name

```
IF SUM( [Occurences defect] ) == [Max #Occurences]
THEN MAX( [Street Name])
END
```

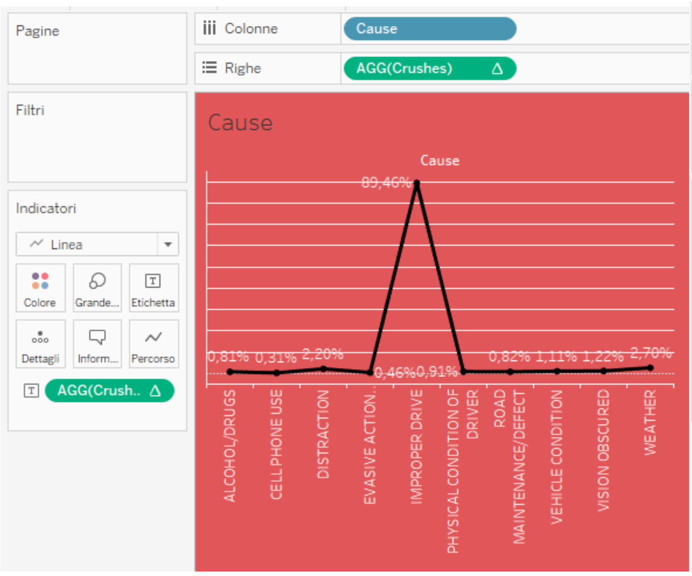


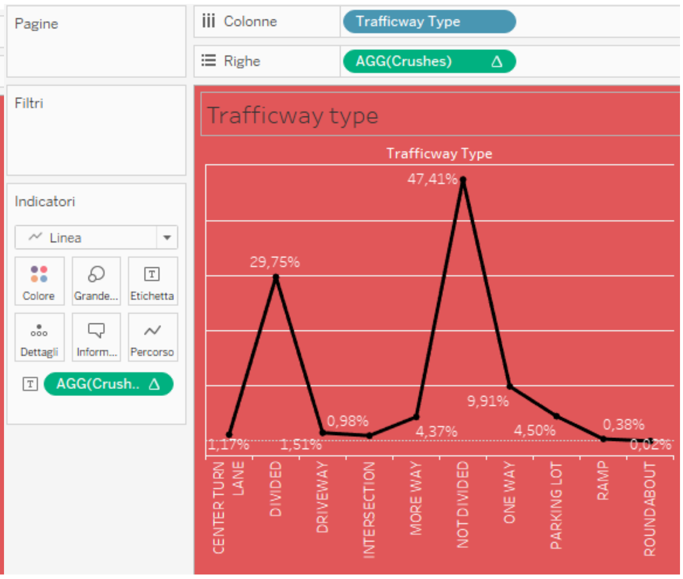
And also for **Max Beat** that indicate the beat with the higher number of crashes.





For **Cause** and **Trafficway** I've put in the rows the number of crashes in percentage and in the columns respectively the cause and the typology of trafficway.

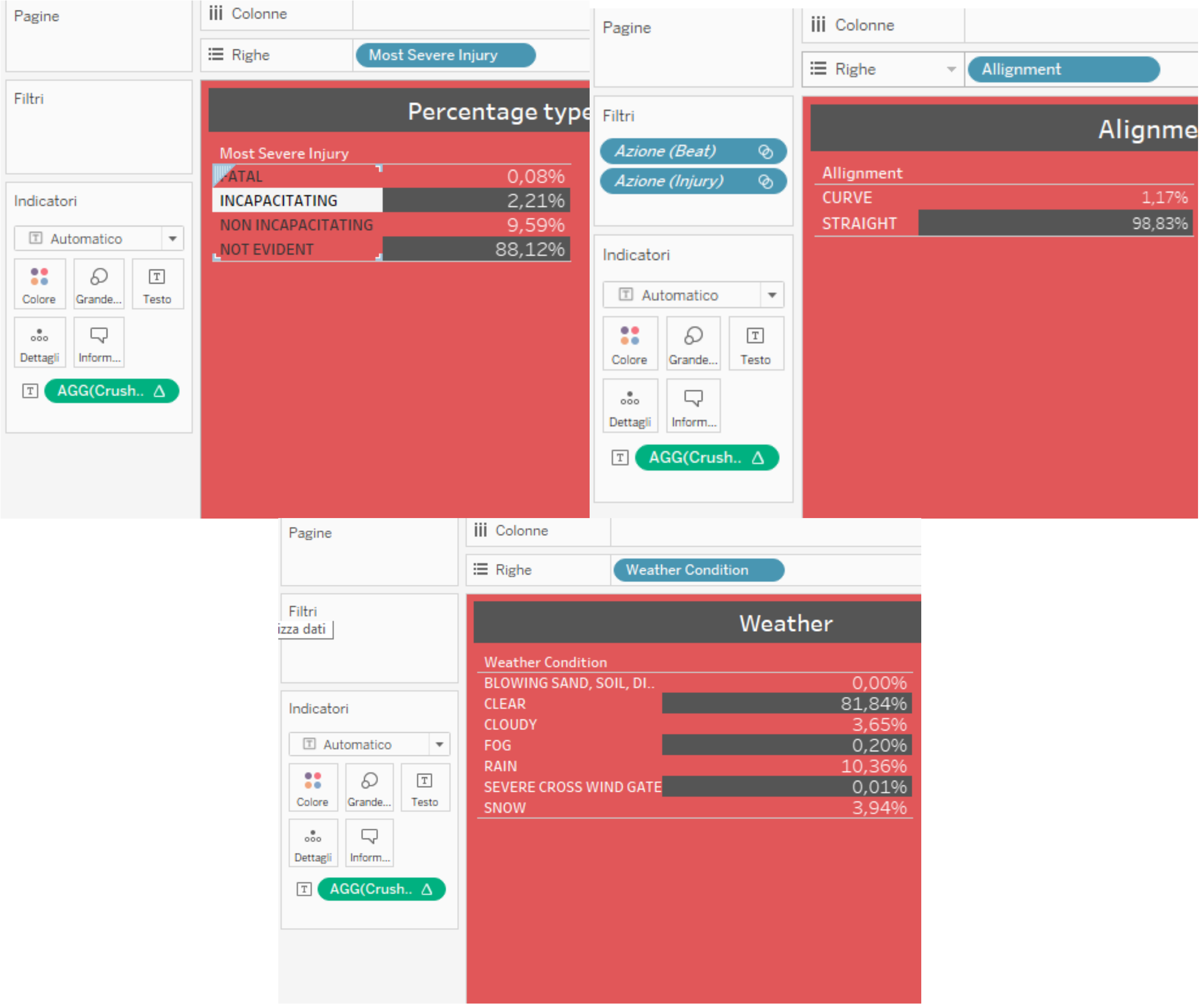




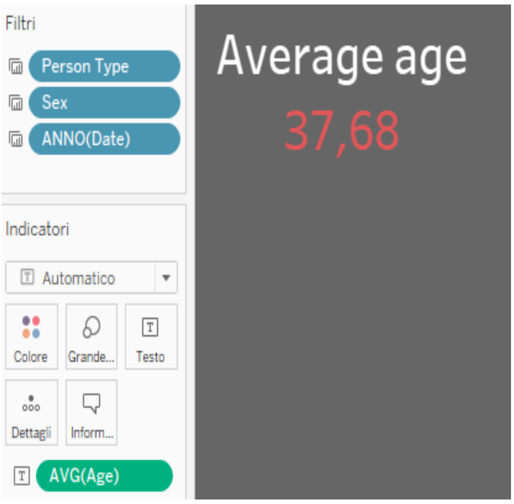
In the sheet **Device Conditions** I've used as graph the area in correspondence of the number of the conditions of the device of the streets of a crashes, to see what is the most common situation of device condition in as crush.

The sheet **Crushes – Defect** is a scatterplot in which I've put in the columns the number of crushes, and in the rows the number of defect. The observations are the streets. Thanks to the scatterplot we can see if at the increase of the number of defect, there is an increase of the number of crushes.

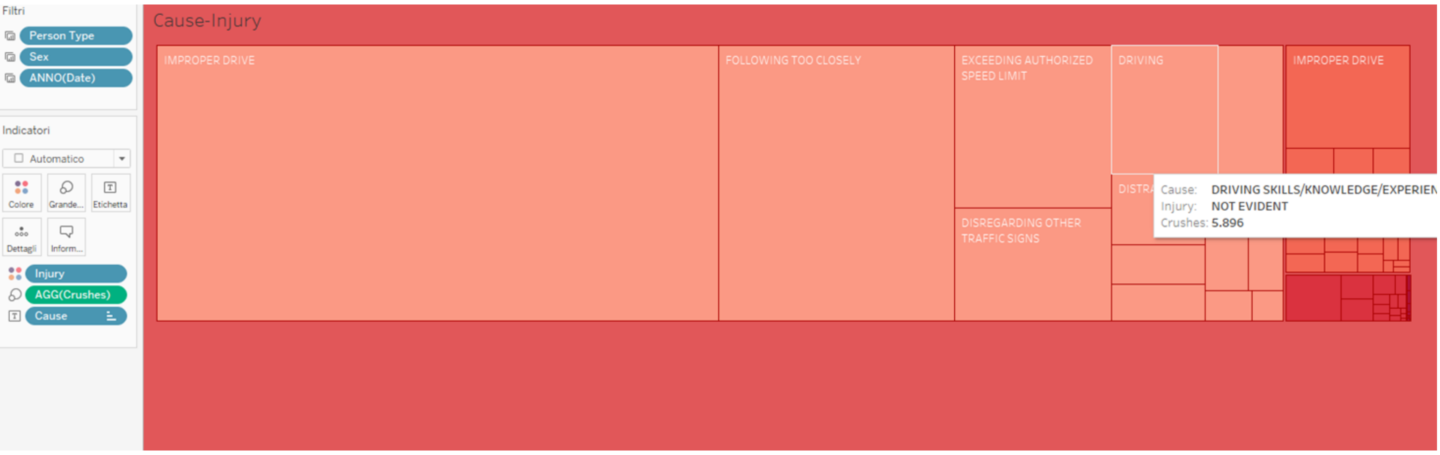
For the first dashboard I've finally three different tables that show respectively the percentage of typologies of injury relative to the total of crushes, the percentage of crashes for the curves or straight and the percentage of crashes for the different weather conditions.



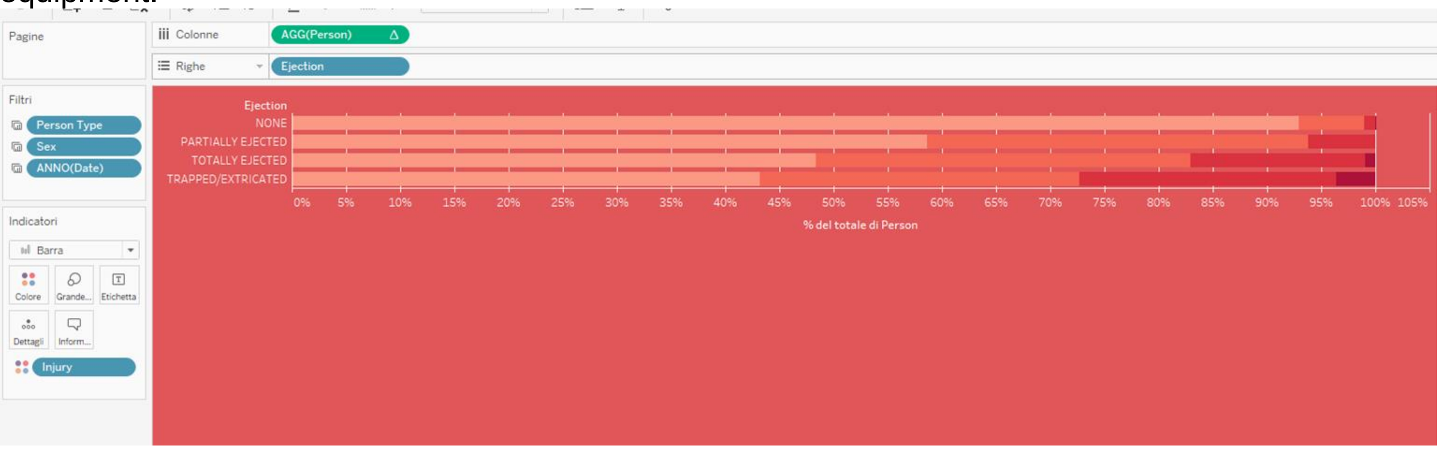
For the second Dashboard we have **Avg Age** that is a text label with the indication of average age of the people involved in accidents in Chicago per year, person Type and sex. I've used a different colour background that match with the use of colour of the other graphs, indeed I've used the scale of black of the bin of the histogram for example.

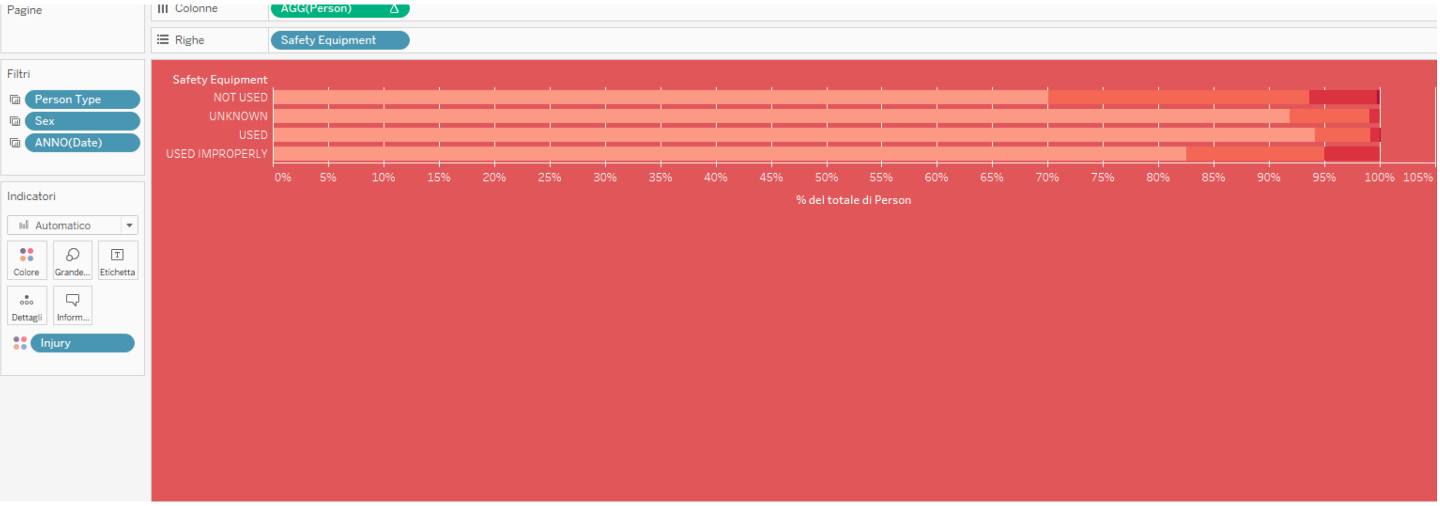


The sheet **Cause-Injury** is a tree-map that express the most common cause of accident per typology of injury. Colour intensity, as well as size are used as pre attentive attributes.

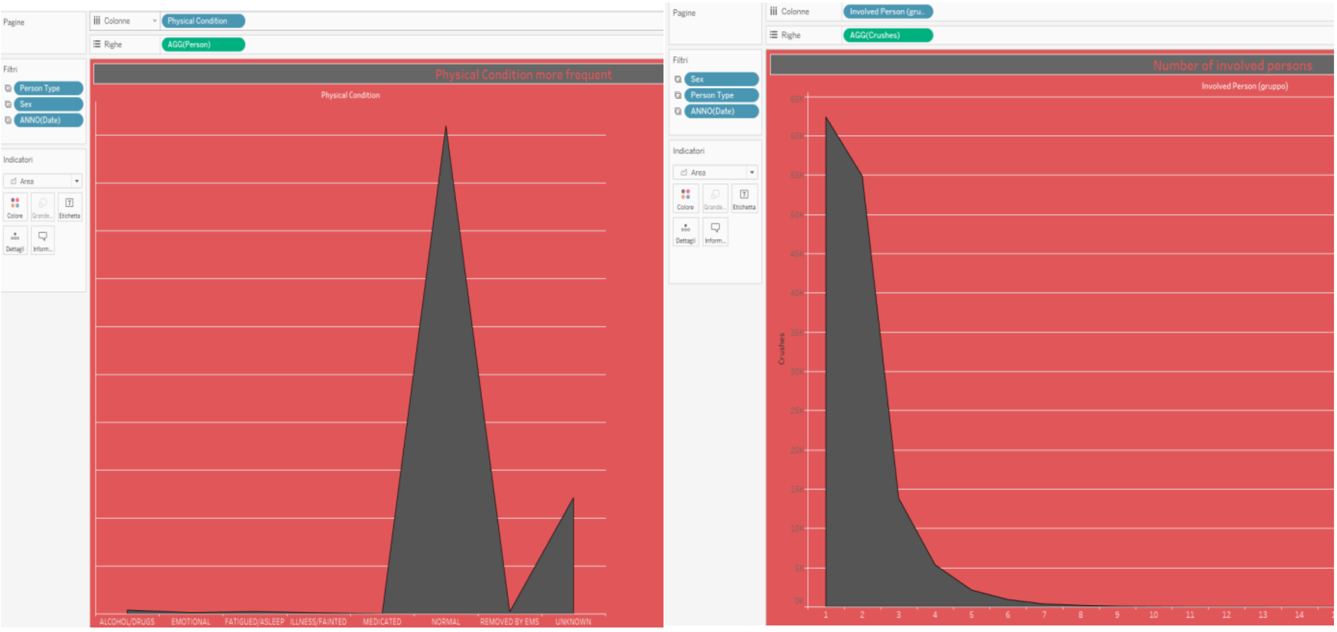


There are two sheet (**Safety Equipment – Injury** e **Person Type-Injury-Ejection**) with the same structure: a bar-plot with a percentage of row that in the first case express the percentage of each typology of injury per the final situation of ejection and in the second case the typology of injury per the presence or not of the safety equipment. I've added a computed field that is Person that count the number of Person involved. I've selected this typology of visualization because in the case of ejection and safety equipment there is a relevant difference in the number of occurrence between the values, so is more interesting to see how the state of ejection or the state of safety equipment influenced the type of injury, rather then the number of person involved related to the value of ejection or safety equipment.

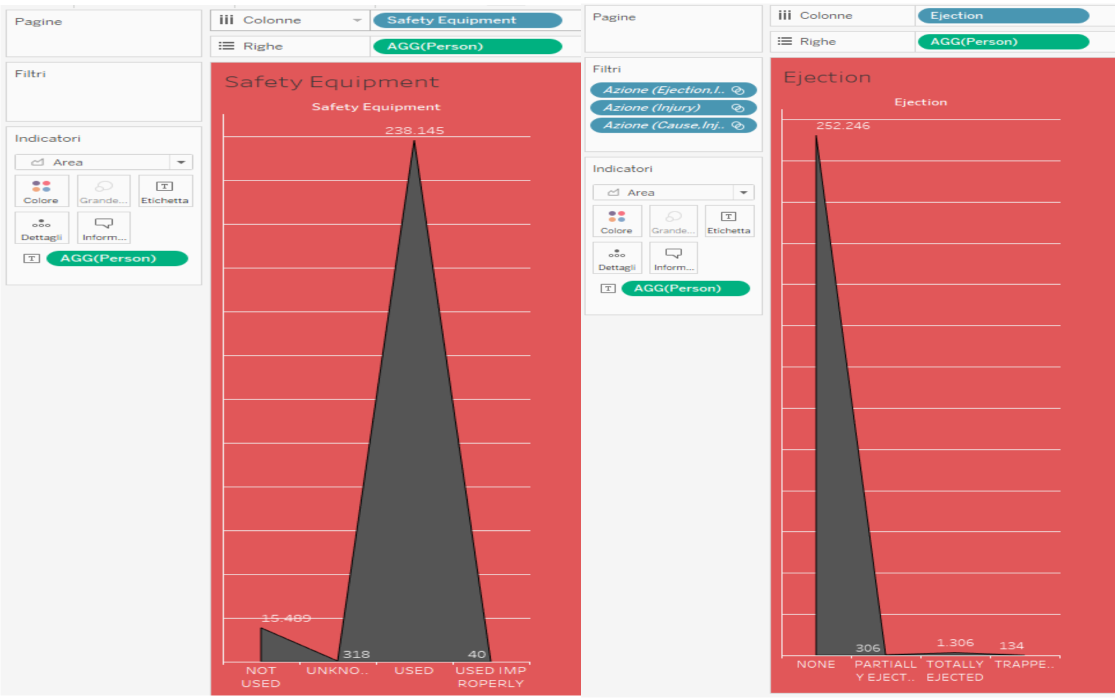




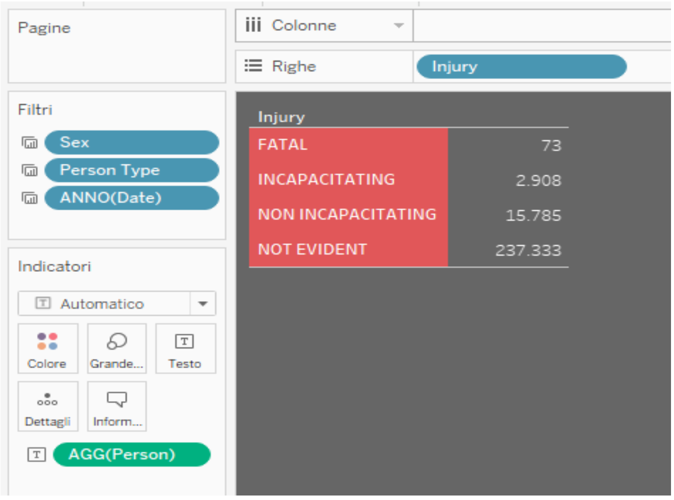
In the sheet **Physical condition** and **Involved Persons** I've used a plot that show the area in correspondence of the number of physical condition registered for a person involved in a crashes, to see what is the more common physical condition. And for the second plot in correspondence of the occurrence of number of person involved in a crash, to see what is the more common value for the number of person involved in a crash.



I've used the same graph for the sheets Ejection and Safety Equipment, that are sheets in help of the sheets *Safety Equipment – Injury* e *Person Type-Injury-Ejection*, to display also the number of people that have used or not respectively safety equipment and have been extracted or not. These sheet are **Safety Equipment** and **Ejection** respectively.



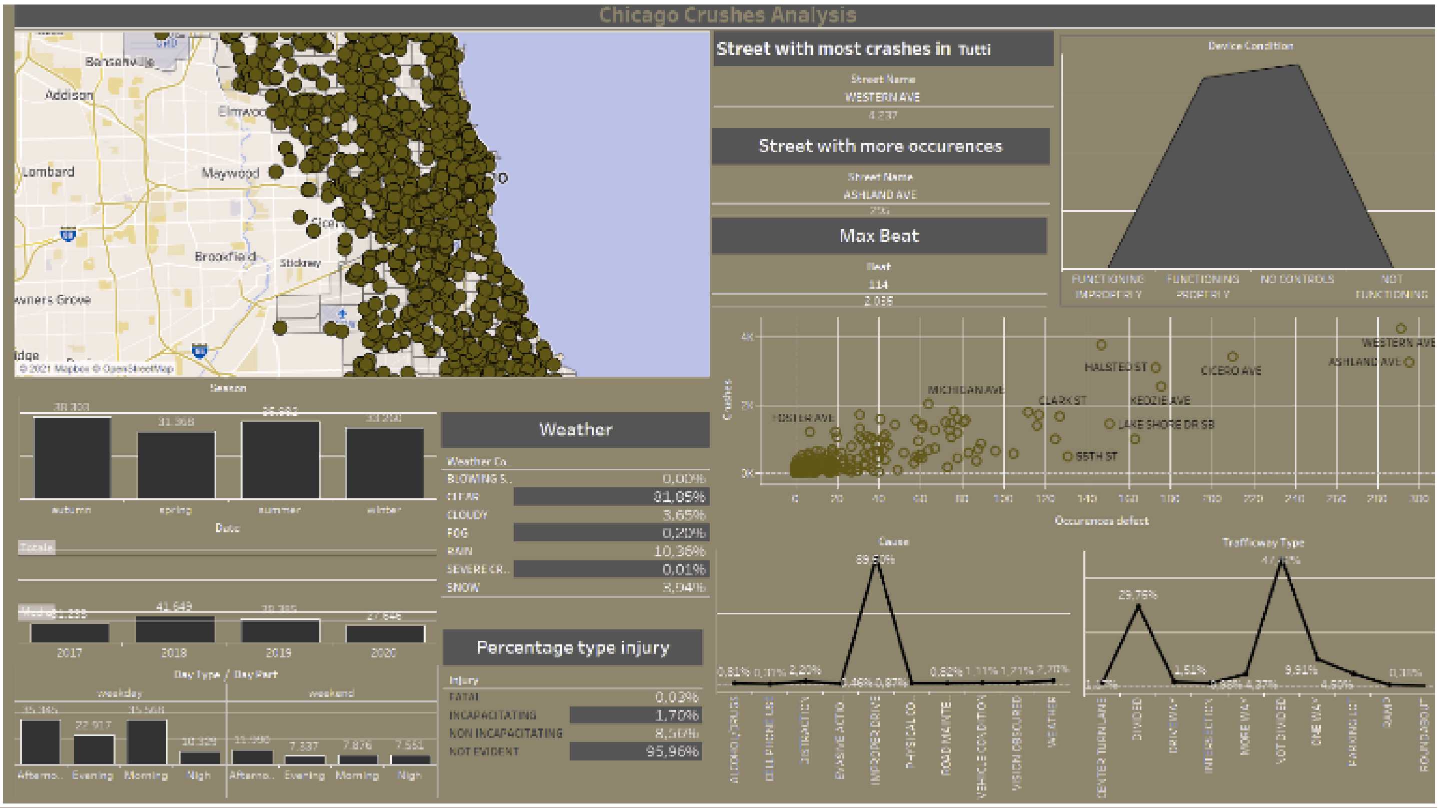
In the sheet **Numbers** I've shown a table that show the number of person per typology of injury.

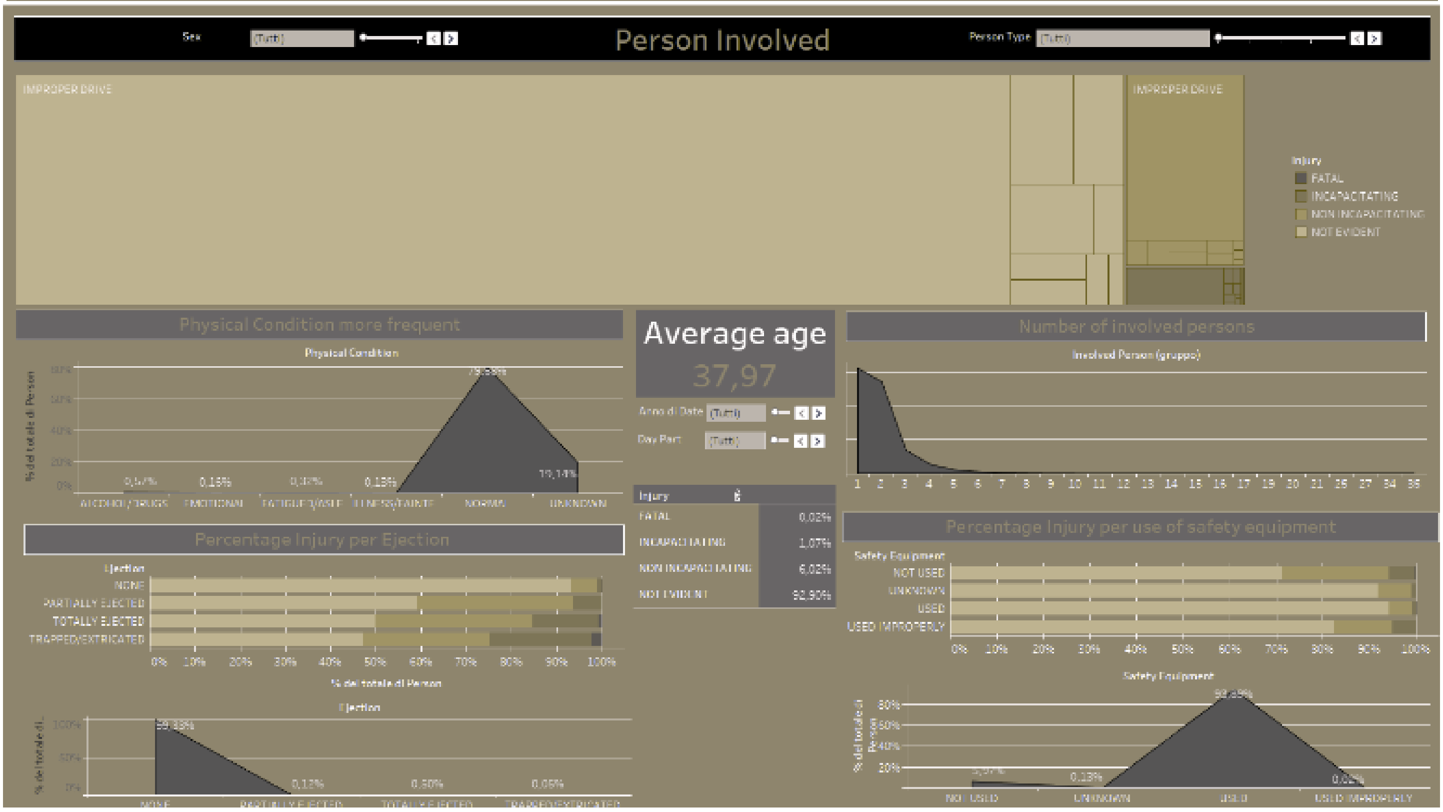


COLOW BLINDESS CHECK

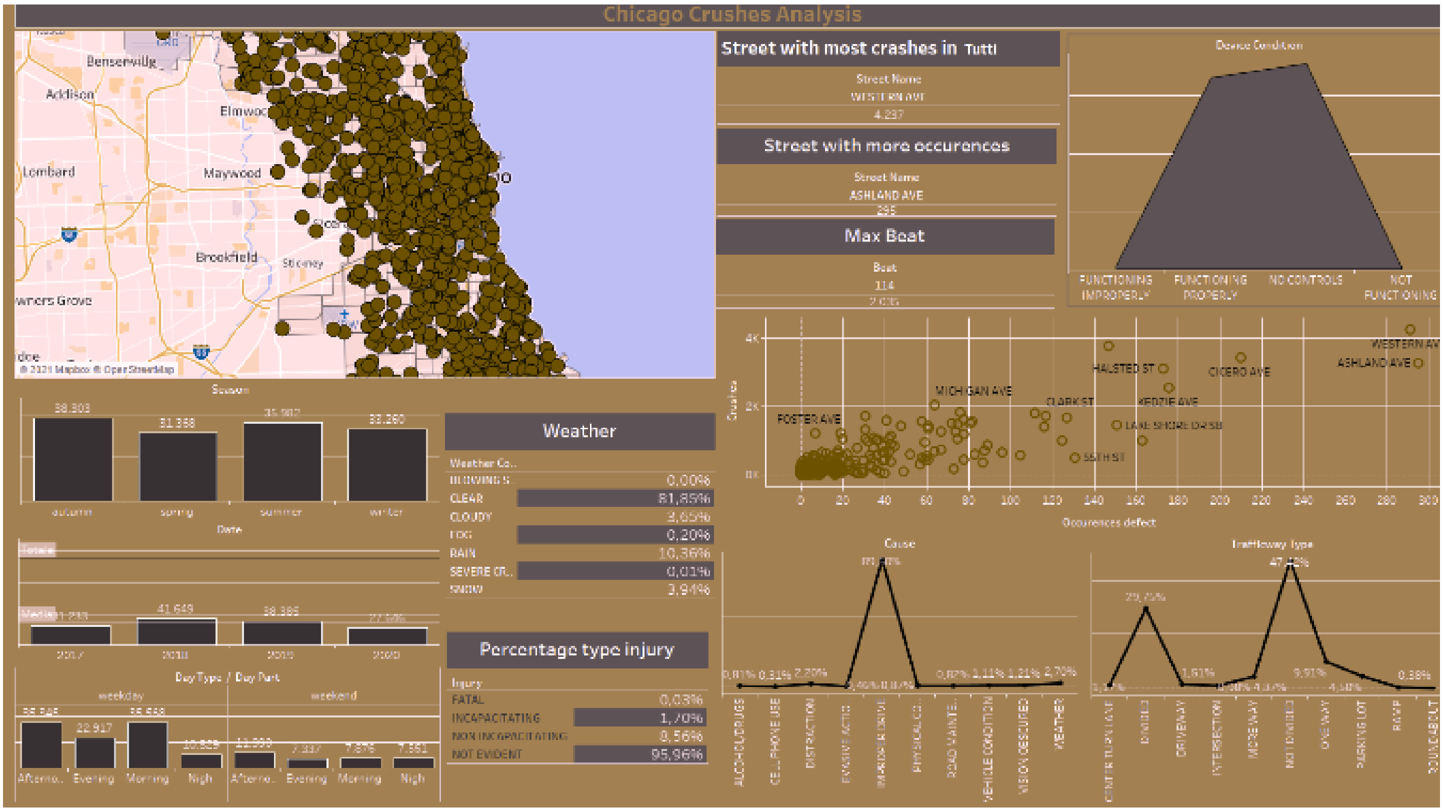
My dashboards mainly use two colours, red and gray, and white for text. I've performed a check for colour blindness in order to be sure that the dashboards can be properly seen and understood by colour blind people : for Red-Blind, Green-Blind, Blue-Blind and Blue Cone Monochromacy and Monohromacy/Achromatopsia views by using Coblis - Color Blindness Simulator and they were successful.

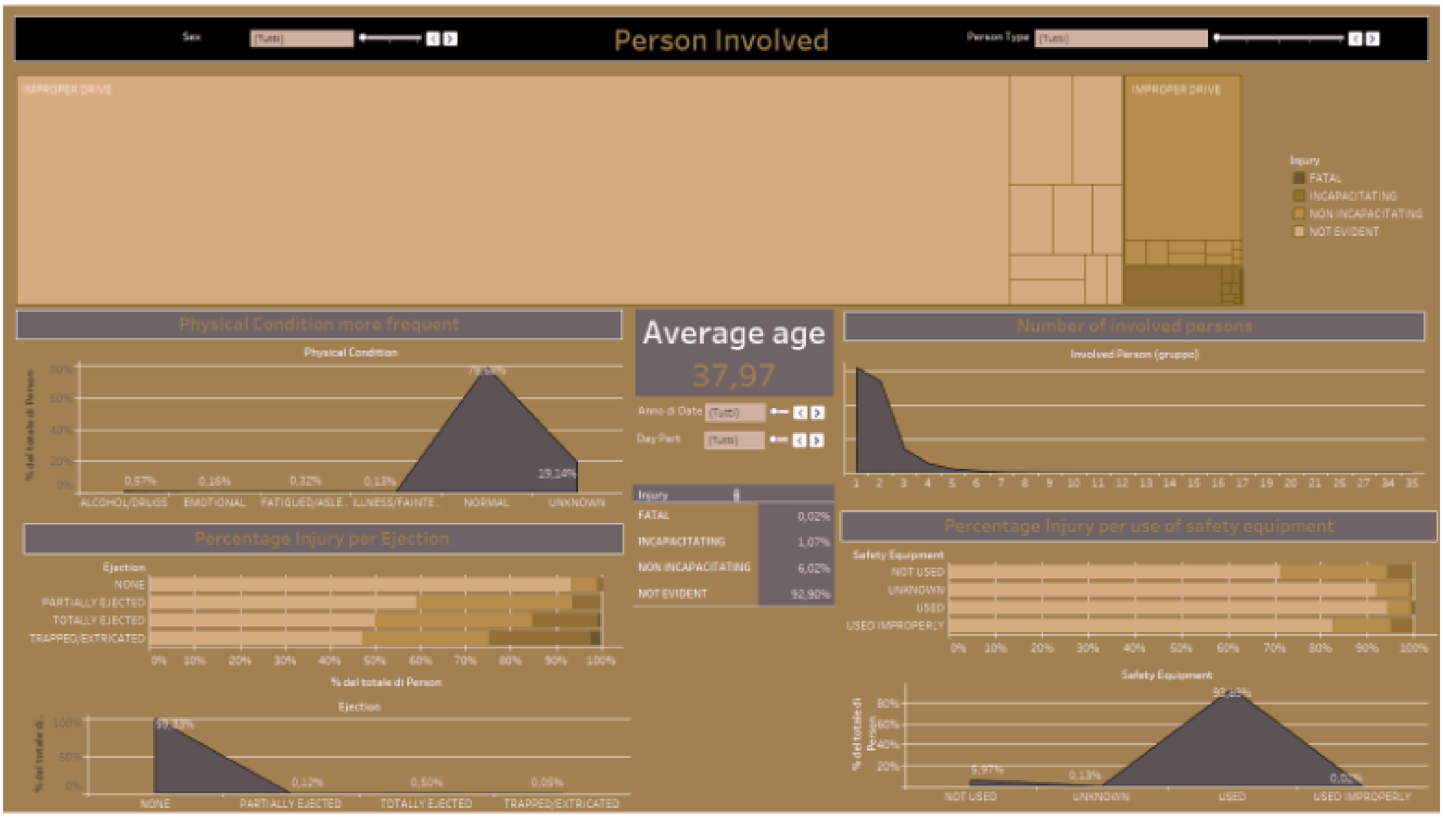
DICROMATIC VIEW : RED-BLIND



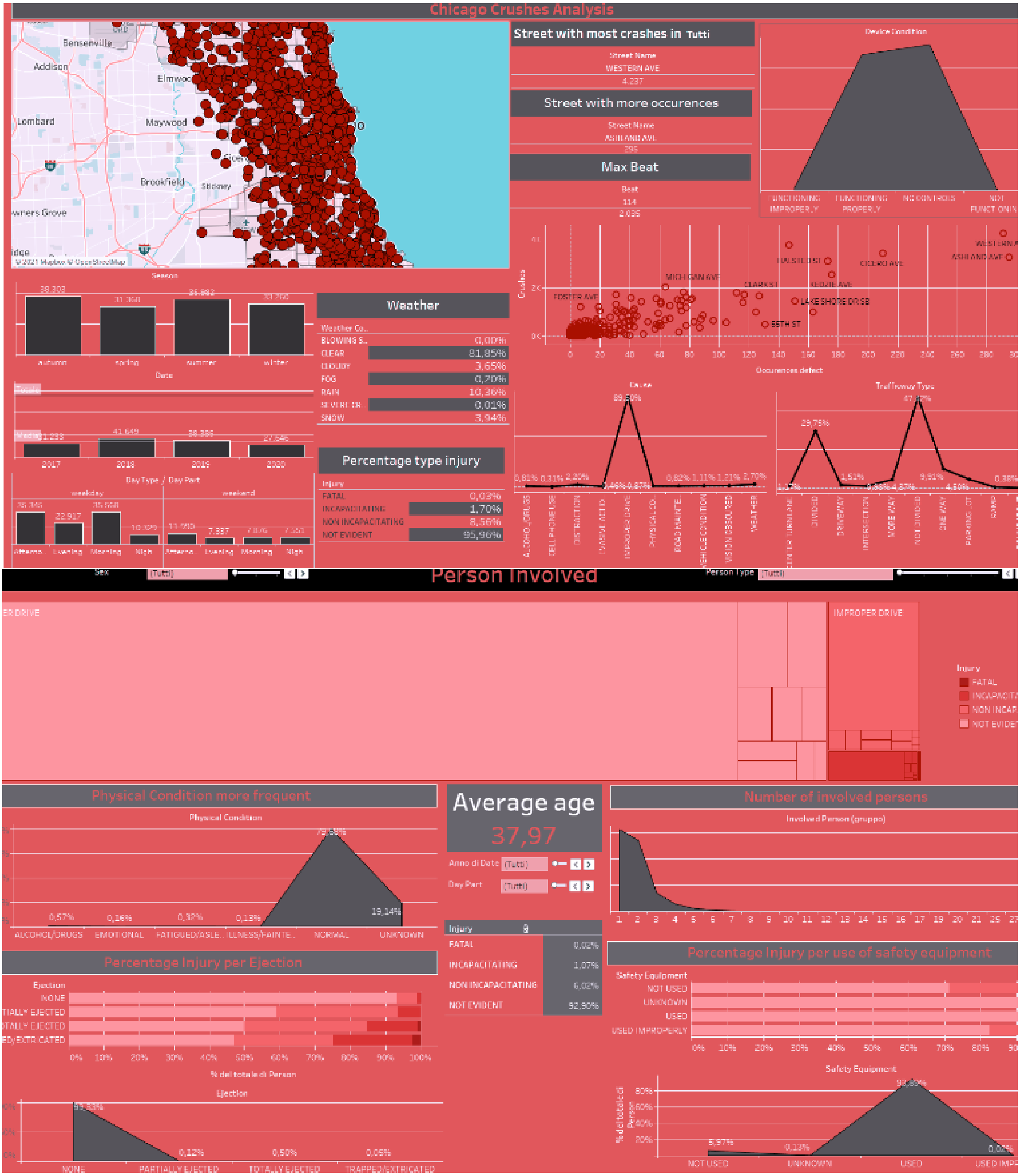


<u>DICROMATIC VIEW: GREEN-BLIND</u>

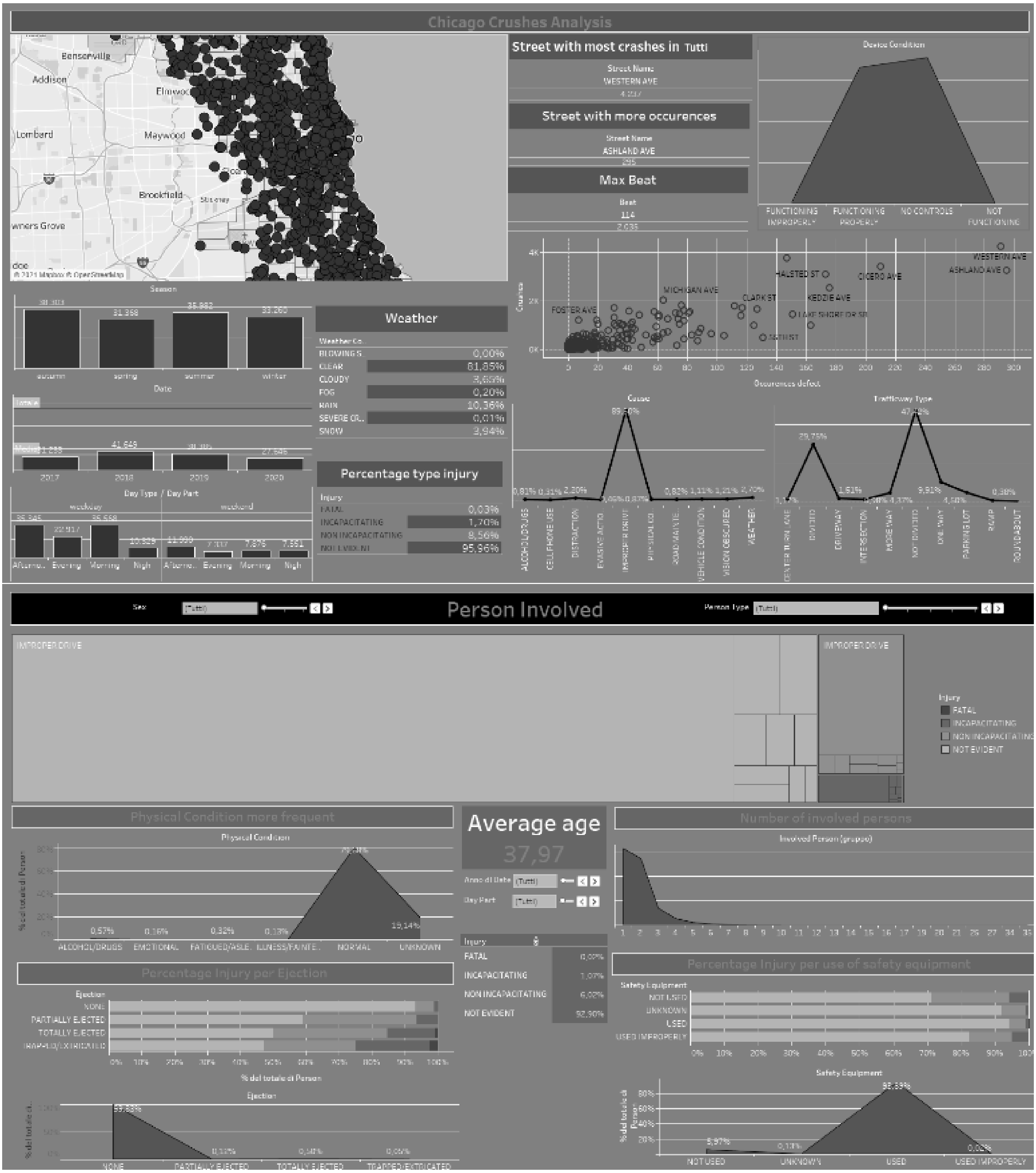




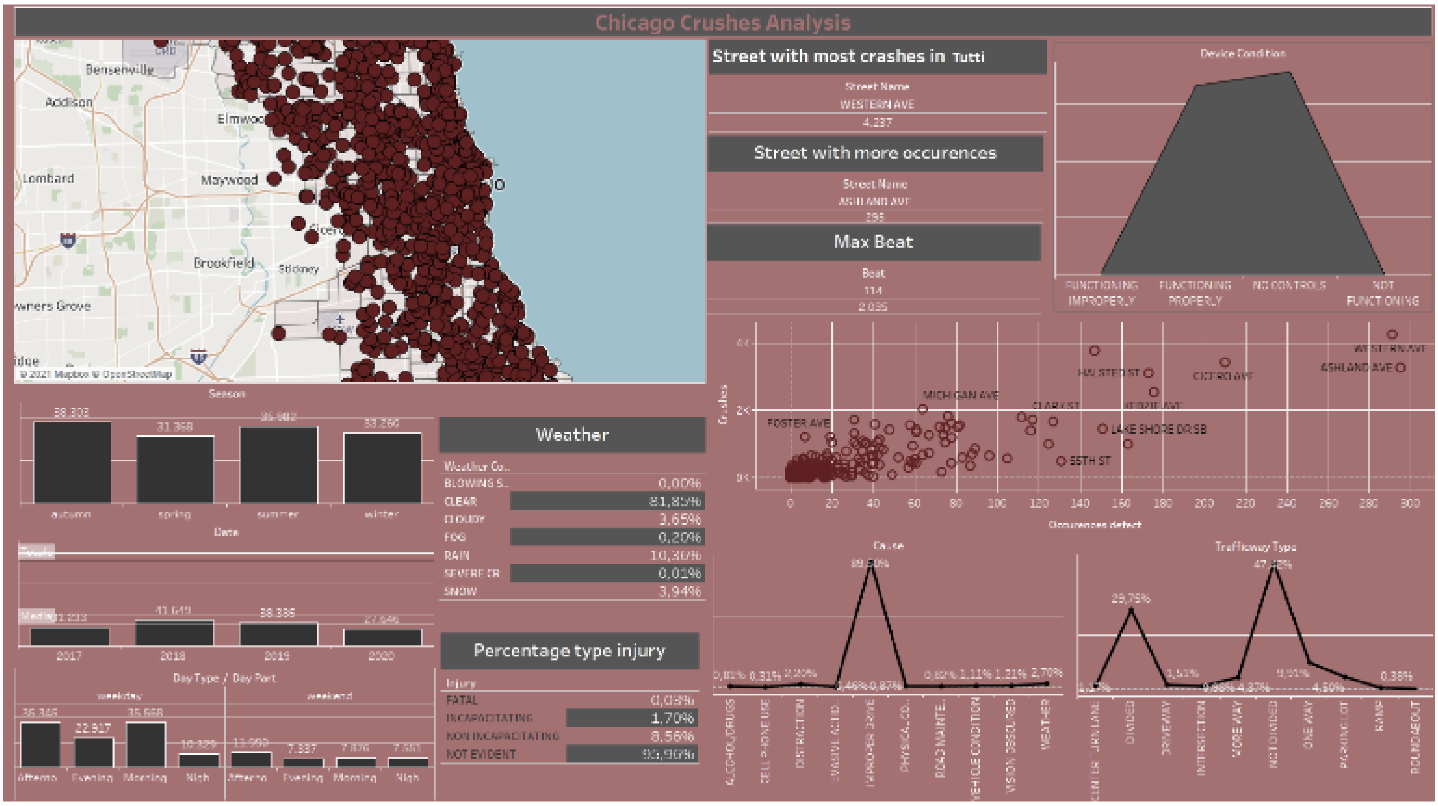
DICROMATIC VIEW : BLUE-BLIND

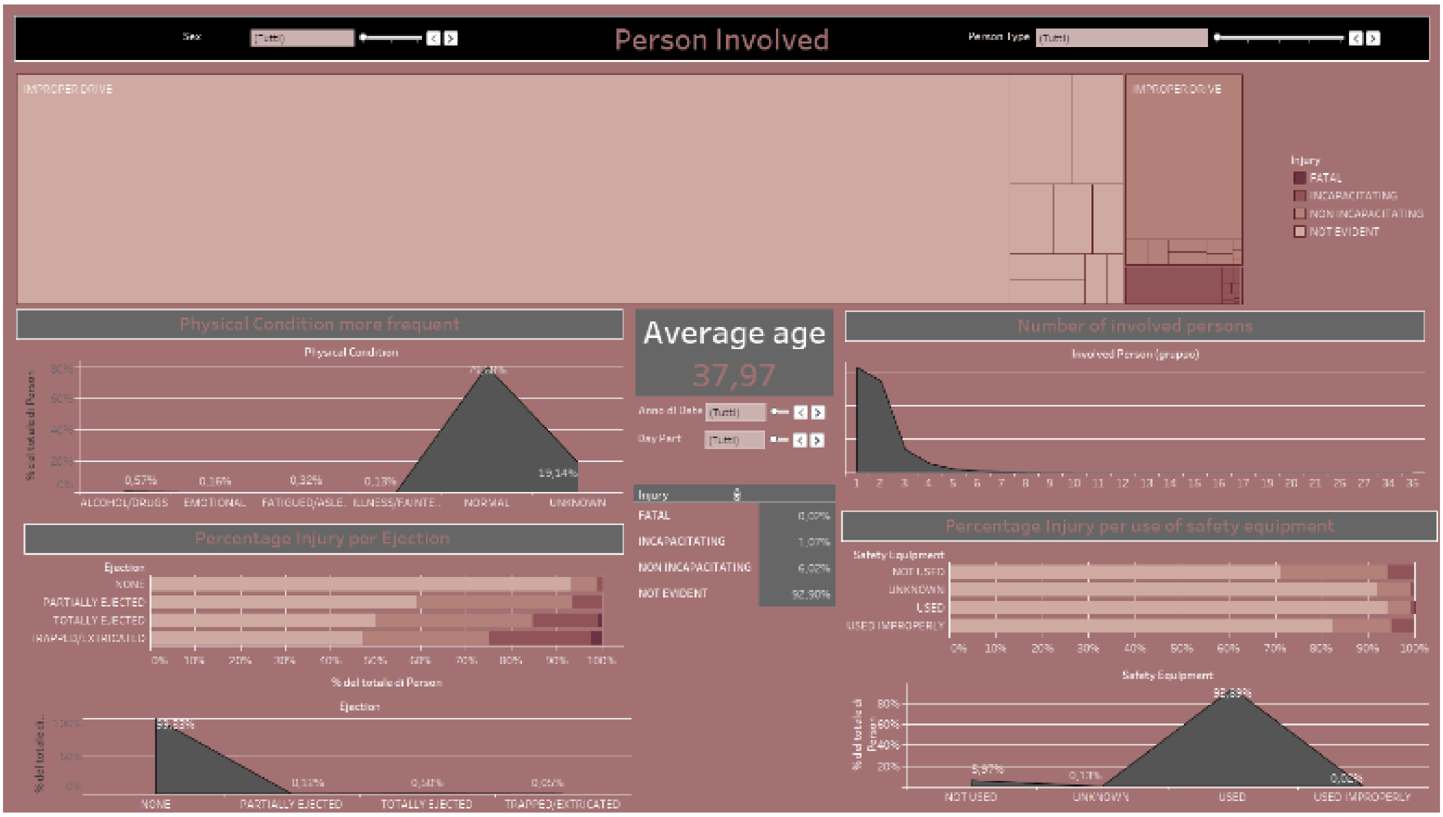


MONOCHROMATIC VIEW: ACHROMATOPSIA



MONOCHROMATIC VIEW : BLUE CLONE MONOCHROMACY





ANSWERS

1. How the external condition influenced the number of accidents?

1.1 How the number of crushes increase along the year? What is the year with most number of crashes?

1.2 Which is the season with the higher number of crashes? What are the days of the week and the part of the

day with an higher possibility of crash? And which climate condition?

1.3 Did the number of defect of the road infleunced the number of incidents? Which are the street and Beat with more crashes?

1.5 Did the condition of device infleunces the number of incidents? And in which type of street occurs more incidents?

My reports shows a subset of the total crushes registered in Chicago between 2017 and 2020. From the first Dashboard we can see that the number of crushes per year obviously not follow a precise trade, in the 2018 and 2019 we have the maximum number of crashes, with the 2018 that have in absolute the higher number of crashes. In these four years in Chicago in average there were 38.820 crashes per year. Thanks to the use of the table “Pecentage type injury” we can see that the 2020 is the year with the higher crashes in which the most several injury is fatal.

The seasons have more or less the same values of crushes, but in total the autumn is the season with an higher value of crushes, followed by the summer. But the situation change if we look at the hours of the day, indeed in the night hours the situation change and the season with the highest number of crashes became the summer, this is a predictable results, because in summer is possible that the number of veichles or pedestrian present in the street are higher until later than the other seasons and also the possibility of crashes increases.
The weekday obviously are the day with an higher number of crashes because there is a difference of scale between the two part of the week, and the afternoon and the morning are the hour with the higher number of crushes, also because I’ve considered all the type of crushes not only the car crushes. The higher number of crashes occurs when the weather is clear .

In the four years the street with more occurrences of defect is the Ashland Ave with 295 reported defect, but the street with an higher number of crushes is the Western Ave. So from this point of view seems that the number of defect don’t influenced the number of crashes; but if we look at the scatterplot in the first dashboard we can see that the higher observations for number of crashes and number of defects are Western Ave and Ashland Ave, so also the first have an high number of defect and also the second have an high number of crash. In particular we can see an high correlation between occurrence of defect and number of crashes. The department of Chicago with the higher number of crashes is the Beat 114 with 2037 crashes.

Another important aspect that we can see from the first dashboard is that the higher number of crashes occurs when the street presents no device.
With the 47,41% of the occurrences the not divided streets are that with the highest possibility of crash, followed by the divided street, one way and parking lot. And for the 98,83% of the cases the crashes occurs in straight.

These last questions answer how important it is to have infrastructures with functioning roads and devices in good condition in order to reduce the number of accidents.

2. What is the most frequent cause of an incident?

2.1 How this cause influence the type of injury

The most common cause of an accident are the improper drive (following too closely, disregarding signals and exceeding authorized speed limit) with the 89,46% of the occurrences followed by weather(2,70%) and distraction(2,20%). Improper drive is the principal cause for all the typologies of injury. For non incapacitating injury also physical condition(2,36%) of the driver is the principal cause of crash. For incapacitating and Fatal injuries alcohol is the third major cause of injury.

3. What characteristics in common have the people involved in an accident?

3.1 The use of safety equipment can save the life of the peoples? And what are the percentages?

3.2 Which is the physical condition most frequent in an accident and how the gravity of the physical condition influenced the type of injury?

3.3 The ejection can be mortal?

3.4 There are difference in the average age person involved in an accident related to the part of the day?

From the second dashboard we can see that the use of safety equipment brings the higher percentage of accident in which the person have not evident injury, the situation is more critical in not used safety equipment and used improperly for which the percentage of several injury increases in an obvious way. For driver, not using safety equipment, the probability of a incapacitating injury goes from 0,69% (case of used safety equipment) to 2,26 % and there is a considerable increasing of the fatal injury from 0,01% to 0,24%. The situation is more critical for the bicyclists and pedestrians, in which the use of safety equipment increases the probability of not evident injury but the probability of incapacitating and fatal injury remain high. This show that persons in bicycle or pedestrians are more in risk when occurs in a crash. This obviously depends on typology of vehicles with which they collide. In general the 94,14% of the person use the safety equipment, as is reported after the crash.
The most frequent physical condition in a person that is involved in a crash is Normal with a percentage of 98,16% of the person. The second more common situation is Alcohol/Drugs.
The condition in which the person is after a crash is relevant because influenced the typology of injury.
When high levels of alcohol are discovered in people after an accident, then the percentage of serious injuries (fatal + incapacitating) rises to 6,18%.
When people are diagnosed with an emotional state after an accident, then the rate of serious injury (fatal + incapacitating) is 0.97%.
When people are diagnosed with an fatigued physical state after an accident, then the rate of serious injury (fatal + incapacitating) is 1,73 %.
When people are diagnosed with an illness physical state after an accident, then the rate of serious injury (fatal + incapacitating) is 17,34%.
When people are diagnosed with an normal physical state after an accident, then the rate of serious injury (fatal + incapacitating) is 0,96%.
When people after an accident are medicated then the rate of serious injury (fatal + incapacitating) is 2,33%.
When people after an accident are removed by ems then the rate of serious injury (fatal + incapacitating) is 26,48%.

So if we not considering the condition of medicated and removed by ems the physical condition after a crash with the higher percentage of more serious injury is illness/fainted.
The situation have can be analysed for different Person Type.
For Drivers the physical condition most frequent remain “Normal” and the second is “Alchool/Drugs” but a difference occurs between Male(0,77%) and Female(0,49%), with the male that have an higher probability to use alcohol or drugs. These situation changes in according to the typologies of injury, for a fatal injury for example the percentage of Alcohol Drugs increase until to 3,03% for male and 0% for female, the most common physical condition now is Removed by ems, followed by normal, (and illness/fainted and alcohol/drugs only for the male). For incapacitating injury the percentage of alcohol/drugs is the third more high.
For driver Night is the part of the day with the higher number of fatal injury and the higher concentration of alcohol/drugs and this can also be seen by the tree map because in correspondence of night the rectangle relative to the fatal injury increases in size.
Bicycle have a percentage of fatal injury higher than driver, and this type of injury are more concentrated in morning. The most frequent physical condition also for the bicycle is normal, followed by alcohol/drugs.
Pedestrian have the higher percentage of fatal injury and incapacitating injury. So we can say that in a crash a pedestrian have in the 17,87% of the cases the probability to develop a serious injury. The more frequent physical condition is always normal,followed by removed by ems and alcohol/drugs. The physical condition with the higher percentage of fatal injury is (besides remove with ems) alcohol and drugs with a probability of fatal injury of 1,02%.
For Pedestrian afternoon is the part of the day with the higher percentage of fatal injury. But in the morning the percentage of incapacitating injury increases up to the 42,86%.
The risk of ejection after a crash can be relevant for the life of the person, indeed when a person is ejected the percentage of serious injury increase from 0,98% (incapacitating) and 0,02% (fatal) of the case of no ejection up to 16,84% and 0,76%. The situation is worst when the person remain trapped in the remains of the vehicles(respectively 25% and 2,88%).
The night is the part of the day with the higher percentage of fatal and incapacitating injury(0,06%).
The average age of the people involved in a crash is 40,56 years old. With the female that in percentage are younger in relation to male. The average age of fatal injury is 43,11 years old. The average age of incapacitating injury is 40,78 years old. For non incapacitating injury is 38,87 years old. For not evident injury 40,65 years old.
The youngest person involved in crashes are bicycle, the older are driver. The crashes with the younger average age are in night, instead in afternoon and morning there is the oldest.

CONCLUSION

The study helped us to derive factors that are responsible for accidents and how we can reduce the severity of the injury, seeing at the part of the day, the condition of the street, the weather, the use of safety equipment and more. The analysis shows indeed that the most serious death is happening in night and in the period of summer, with a state of illness or alcohol/drugs. And in general the person more at risk in a crash are the pedestrian. Finally, this study recommends infrastructure which can help to reduce accidents, indeed the situation in which there are no device, damage on street are the situation of more crashes.