

JUAN FELIPE ZAMBRANO BAEZA

SAMUEL PATINO LUCUMI

PROJECT ETL

link of the database:

<https://www.kaggle.com/datasets/nathaniellybrand/chicago-car-crash-dataset>

summary : this data, released on Data.gov that is The Home of the U.S. Government's Open Data Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

contains all car crash data from 2001+ which happened in Chicago. The dataset contains the date of crash, the speed limit, the weather conditions, the injuries, the latitude and longitude and much more information about the crash in the city.

we as a group chose this dataset because we could see that it was a very large data set which contained a lot of information.

which is more than 700,000 rows and 42 columns. this data that this dataset contains is from 2001 to June 2023 of all accidents reported in the city of Chicago which is one of the largest cities in the United States.

for this reason it was also decided to work in a group of 2 people since it was a very large dataset and it would be very tiring to be done by one person. and between the group we can decide how we will approach the data set because it contains a lot of information and we can perform different types of transformation with it and also because it contains some null values and undelivered information.

2.-After carrying out the analysis of the data set and all its columns and rows as a group, we have decided that we will use the following columns for our data extraction, transformation and loading project:

- Weather condition
- Crash type
- Street no
- Street direction
- Street name
- injuries total
- injuries fatal
- injuries incapacitating

- injuries non incapacitating
- injuries reported not evident
- injuries unknown

- Crash hour
- Crash day of week
- Crash month
- Latitude and longitude
- location

We think that each of these is important since they are the main factors by which an accident occurs. We must bear in mind that each one is different and plays a role when a person drives a car, in this case in Chicago, which normally has a humid climate. The speed depends on each person and his own criteria when driving, so it is a variable that changes a lot. Injuries and deaths are a variable that occurs as a consequence at the time of an accident, it is a variable that we believe is important to analyze. Finally, the time and the days are protagonists since people drive more in specific hours and on specific days.

The main objective is analyze the relationship between climate conditions, driving speed, injuries and deaths, time of occurrence, and geographical locations in Chicago road accidents, with the goal of identifying patterns, trends, and potential influencing factors to enhance road safety and accident prevention strategies

-columns to be eliminated and why

-lighting condition:we as a group and with the objective that we have visualized, we do not believe it is convenient to include this column since it does not contribute to us.

-**first crash type**:this column has been eliminated as we do not believe it is convenient to know the place where the vehicle was impacted

-**trafficway type**:this column has been eliminated

-**lane cnt**:this column was eliminated since most of its data is null

-**alignment**::we as a group and with the objective that we have visualized, we do not believe it is convenient to include this column since it does not contribute to us.

-**roadway surface cond** this column was deleted as this column's data type is object

-**road defect**:this column was deleted as this column's data type is object

-**intersection related i**:this column was eliminated since most of its data is null

-**not right of way**:this column was eliminated since most of its data is null

-**hit and run**:this column was eliminated since most of its data is null

-**date police notified** we eliminate since knowing what day the police were notified about the accident is not part of our objective to predict some patterns.

-**prim contributory cause**::we as a group and with the objective that we have visualized, we do not believe it is convenient to include this column since it does not contribute to us.

-sec contributory cause:we as a group and with the objective that we have visualized, we do not believe it is convenient to include this column since it does not contribute to us.

beat of occurrence his column has been eliminated as we do not believe it is convenient to know the beat of occurrence of the accidents

-photos taken:this column was eliminated since most of its data is null and void and it would not help us to know if photos were taken of the accident or not.

-statements taken:this column was eliminated since most of its data is null and void and it would not help us to know if the driver's declaration was taken.

dooring:this column has been eliminated as we do not believe it is convenient to know if is dooring

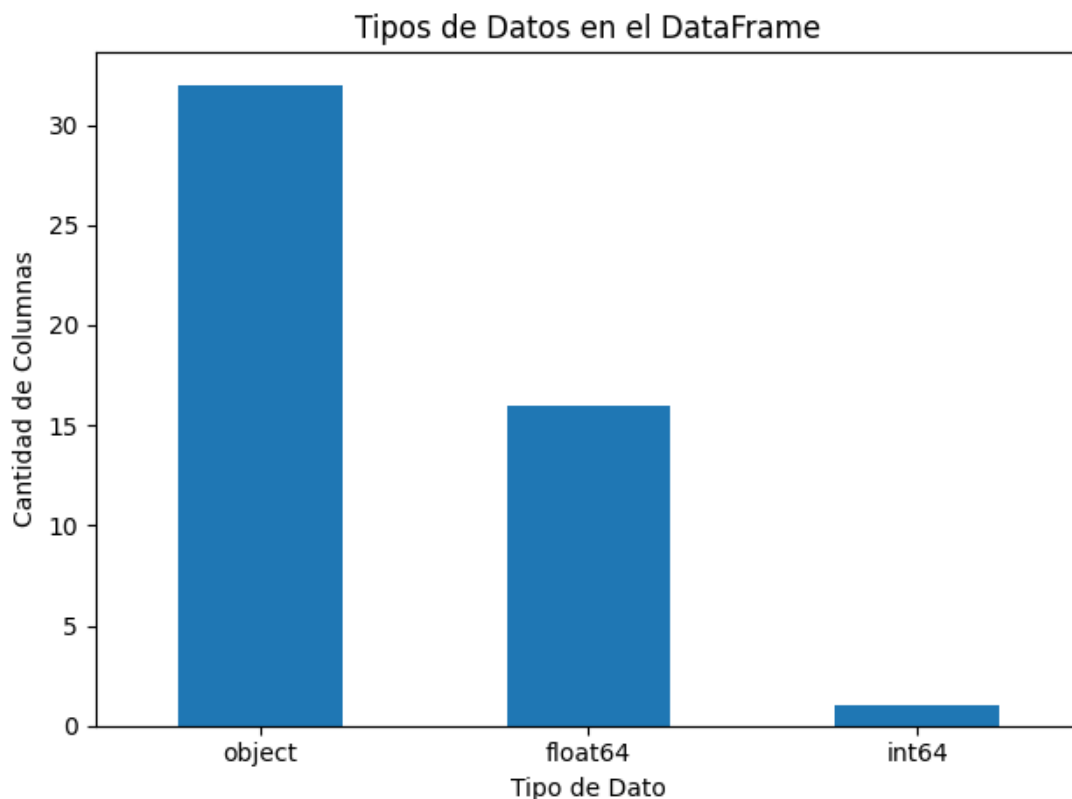
-work zone:this column was eliminated since most of its data is null

-work zone type:this column was eliminated since most of its data is null

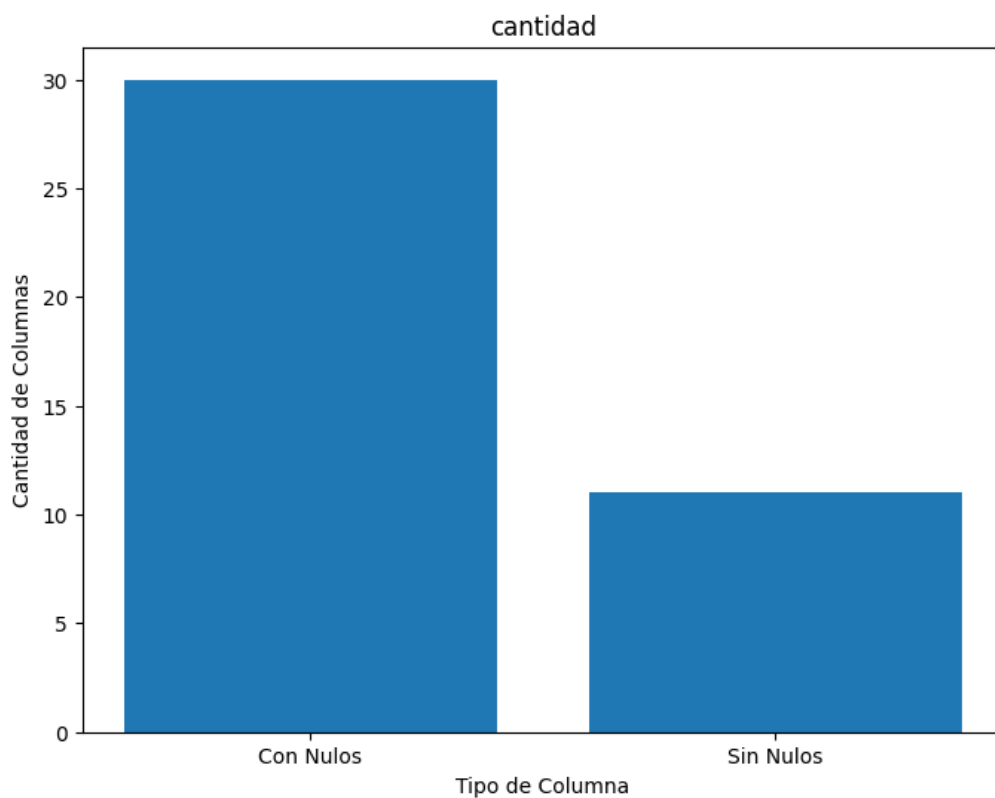
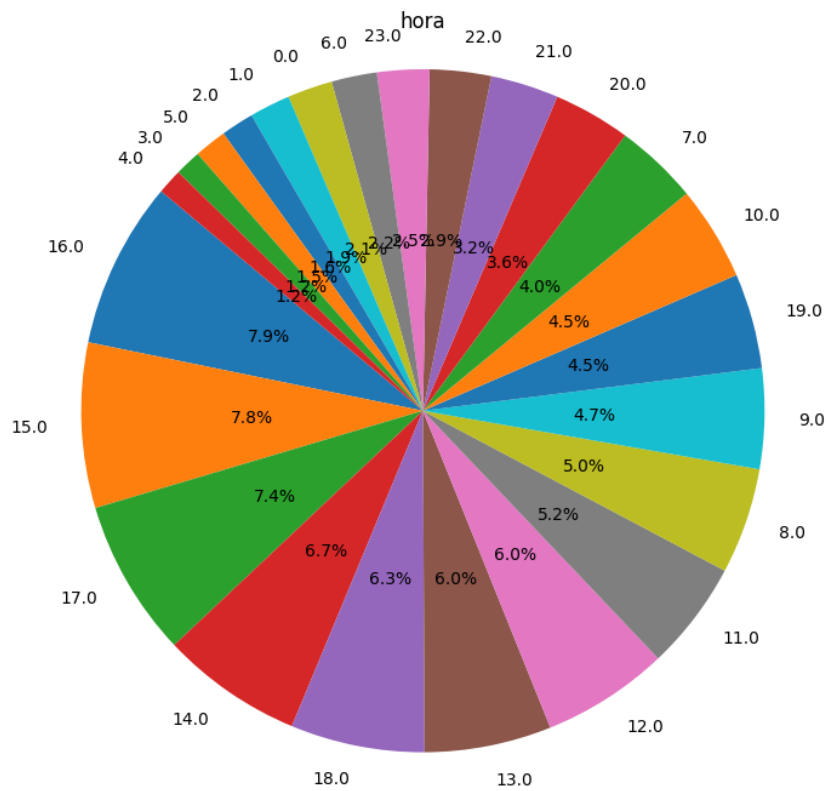
-workers present:this column was eliminated since most of its data is null
nu units

graphics

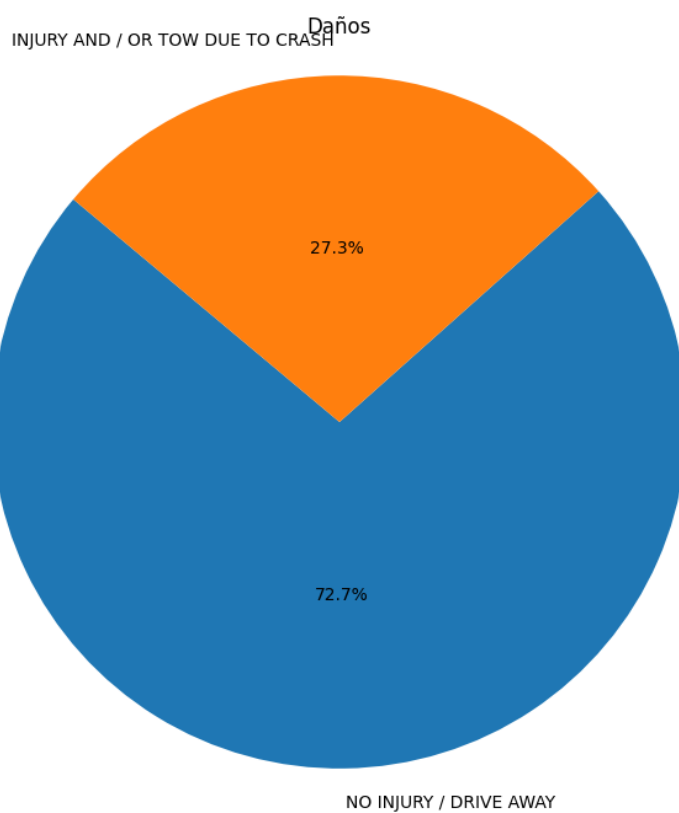
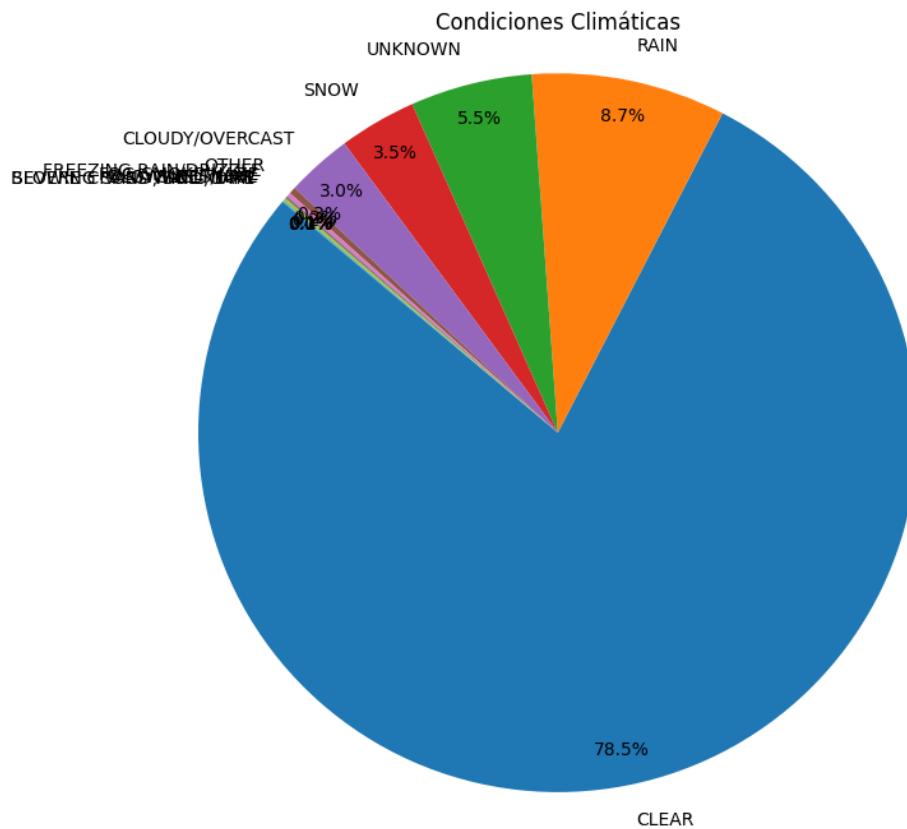
-This graph helps us to understand in a visual way which are the types of data that are most frequently presented.



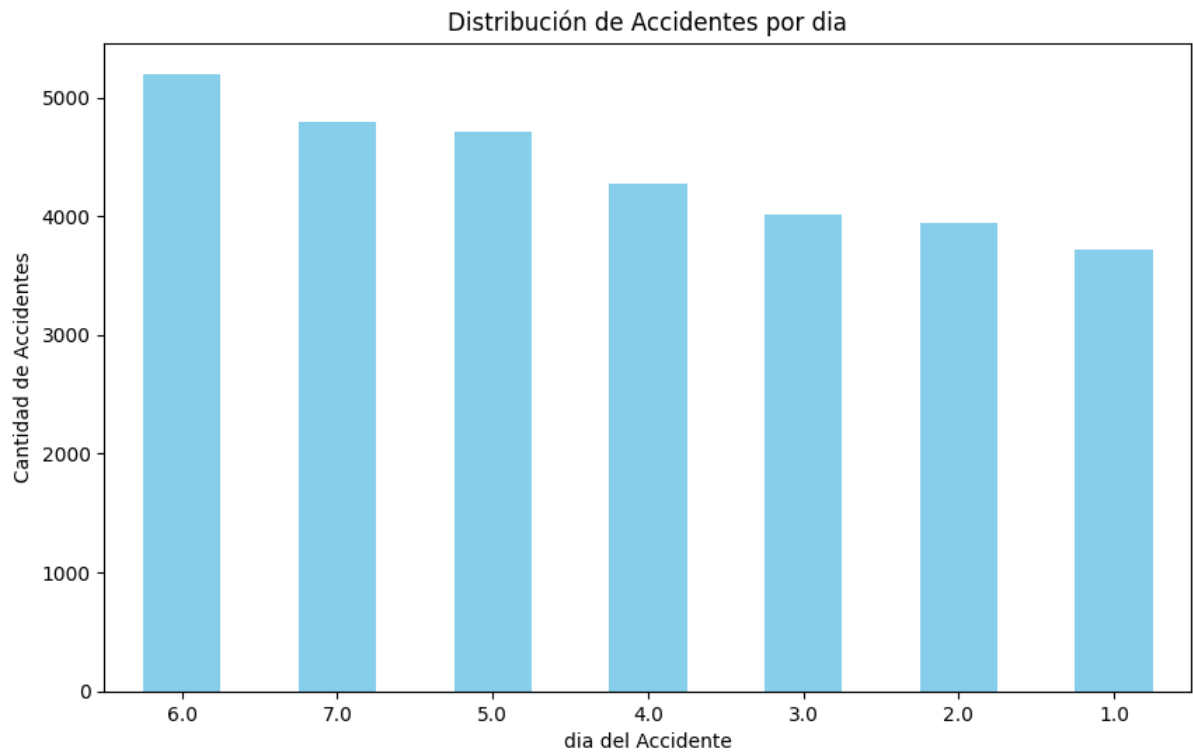
-He graph shows in a visual way the percentage of accidents according to the time of day.



-This graph was taken from the weather conditions column to understand under which conditions accidents occur.



-This graph helps us to understand in a visual way on which day there are more accidents.



-This graph helps us to understand in a visual way in which month there are more accidents.

