

REPORTING



FRENCH vehicles Accidents BETWEEN 2005-2019

Created BY:

BENHAKKI OthmanE

2022

Contents

[INTRODUCTION 2](#_Toc102032376)

[Project Content 2](#_Toc102032377)

[Project Presentation 3](#_Toc102032378)

[Tools Used 5](#_Toc102032379)

[Excel 5](#_Toc102032380)

[Talend 6](#_Toc102032381)

[PowerBI 7](#_Toc102032382)

[R Language 8](#_Toc102032383)

[Data Presentation 9](#_Toc102032384)

[Data Manipulation 11](#_Toc102032385)

[Preprocessing Stage 11](#_Toc102032386)

[Data Cleaning Stage 13](#_Toc102032387)

[Data Transformation 17](#_Toc102032388)

[Talend 17](#_Toc102032389)

[Data Reporting 18](#_Toc102032390)

[Crash Severity By State 18](#_Toc102032391)

[Crash Severity By Day/Night 20](#_Toc102032392)

[Crash Severity By Weather 21](#_Toc102032393)

[Crash Severity By Make 22](#_Toc102032394)

[Crash Severity By Vehicle Manufacturing Year 23](#_Toc102032395)

[Conclusion 24](#_Toc102032396)

# INTRODUCTION

## Project Content

The following project utilize a large dataset of car crashes *(about 4.2 million records)* outsourced from the French government official website for data accuracy and legal purposes.

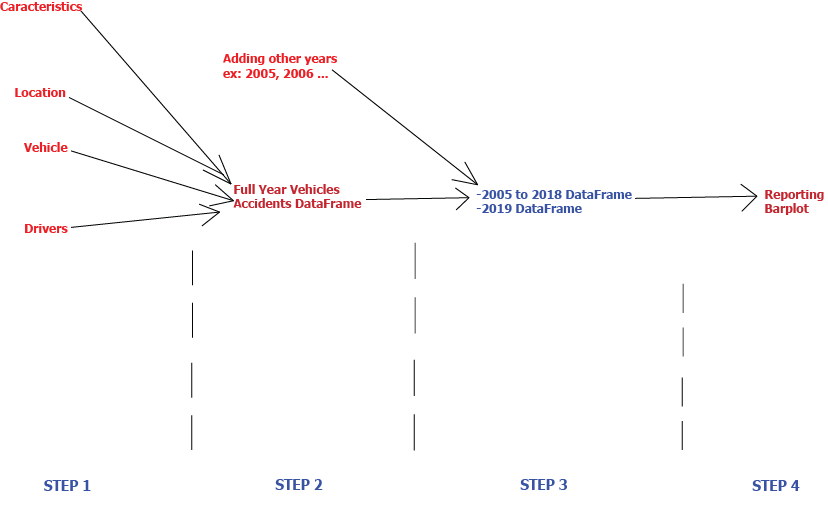
In this project, we will understand how we managed to unify and transform 56 database to a single dataset and then the steps followed to visualize useful information in both shapes text and graphs.

Before, we start detailing any further step we must clarify the project main study goal. The objective is to answer the impact of multiple factors on accidents and comparing the results with two time stamps 2005 to 2018 as reference and 2019 as a sample.

In addition, the questions are related to severity, vehicle category, road type and drivers sexe.

## Project Presentation

This project followed four main steps in order to achieve the given study goal.



We have every year record separately registered on dispersed files. The shared column between all four datasets of each year is Num\_Acc that represents our primary key, some important variables used:

* Grav: refers to the severity, it holds four different levels.
* Sexe: One for male and two for female.
* Num\_Acc: primary key represents a unique id for each record.
* Catv: vehicle category.
* Catr : describe the road type either highway, county highway…etc.

Right now, we have a global vision of what our dataset should carry as crucial information, we will apply several modification using R Studio. This tool is subject to be discussed in the next section.

The source of all datasets are from an official trusted governmental website:



The first step is to aggregate each year by the id. Then, the first dataset will regroup the following years (from 2005 to 2018) in a single dataset using rbind, and the second dataset will represent 2019 records only.

Now the challenge is to unify all datasets in one considering the large amount of files and the difference in delimitation between values in each table.

## Tools Used

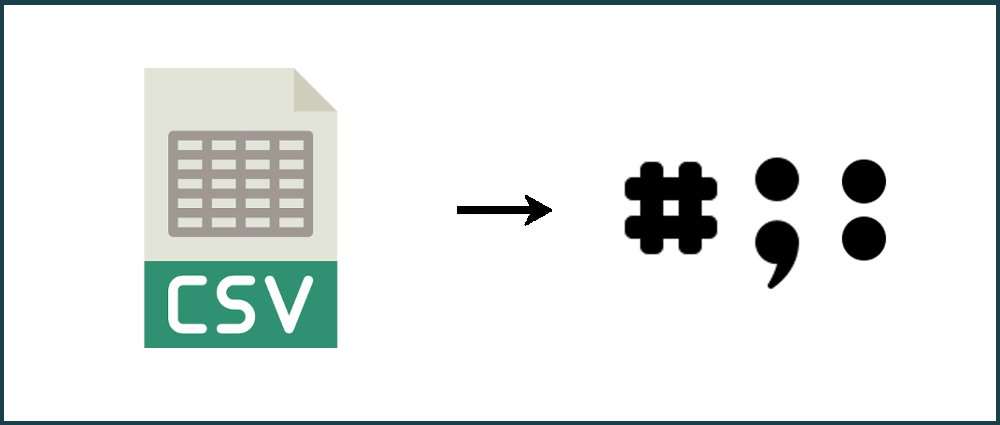
### R Language



R is an open-source programming language that is widely used as a statistical software and data analysis tool. R generally comes with the Command-line interface. R is available across widely used platforms like Windows, Linux, and macOS. Also, the R programming language is the latest cutting-edge tool.

R was used in generating crucial graphs in the reporting phase, using barplot function we managed to answer all five questions and extract some relations ad new factors (subject to be discussed later).

# Data Presentation



The data used comes in different delimiters:

* Space: separator used mainly in header.
* Period: separator used also in header.
* Semi column: delimiter used in table content to separate the columns.

Why discussing this?

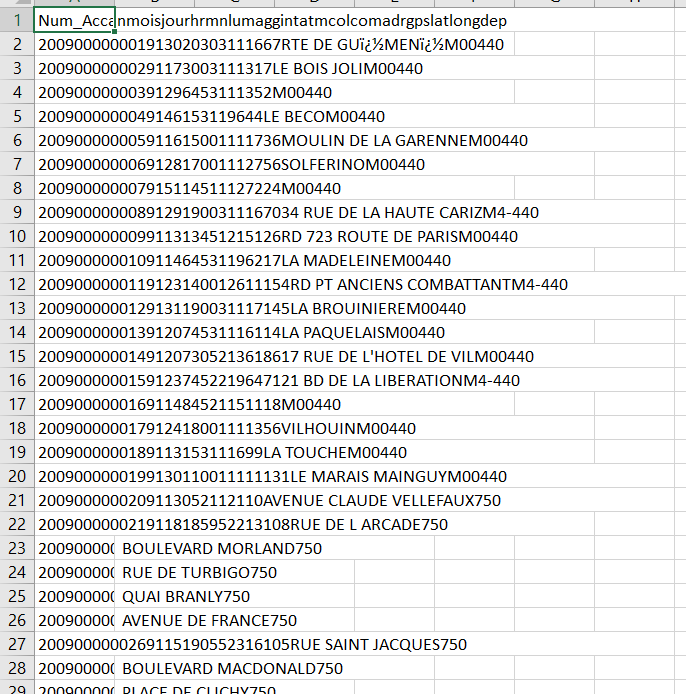
Due to reading issues, the CSV files reading phase can generate errors. To prevent it, we need to know each file delimiters. In order to specify it in the function of read.csv or read.csv2.

# Data Manipulation

## Preprocessing Stage

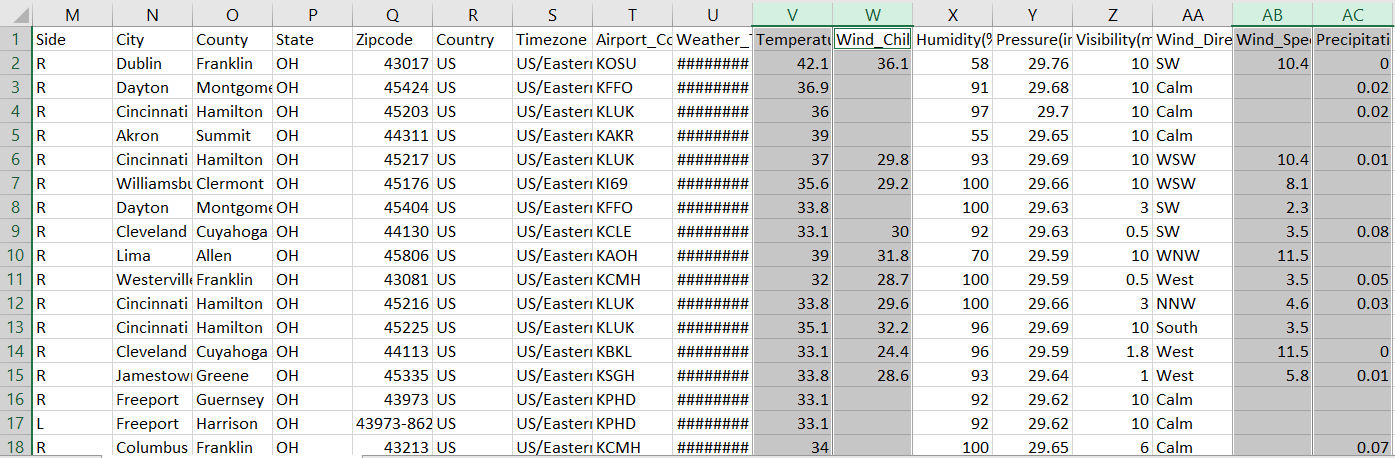
Exception:

The dataset of 2009 cannot be cleaned or used anyhow. This is due to data formatting, no separators were used which makes it impossible to identify columns values.

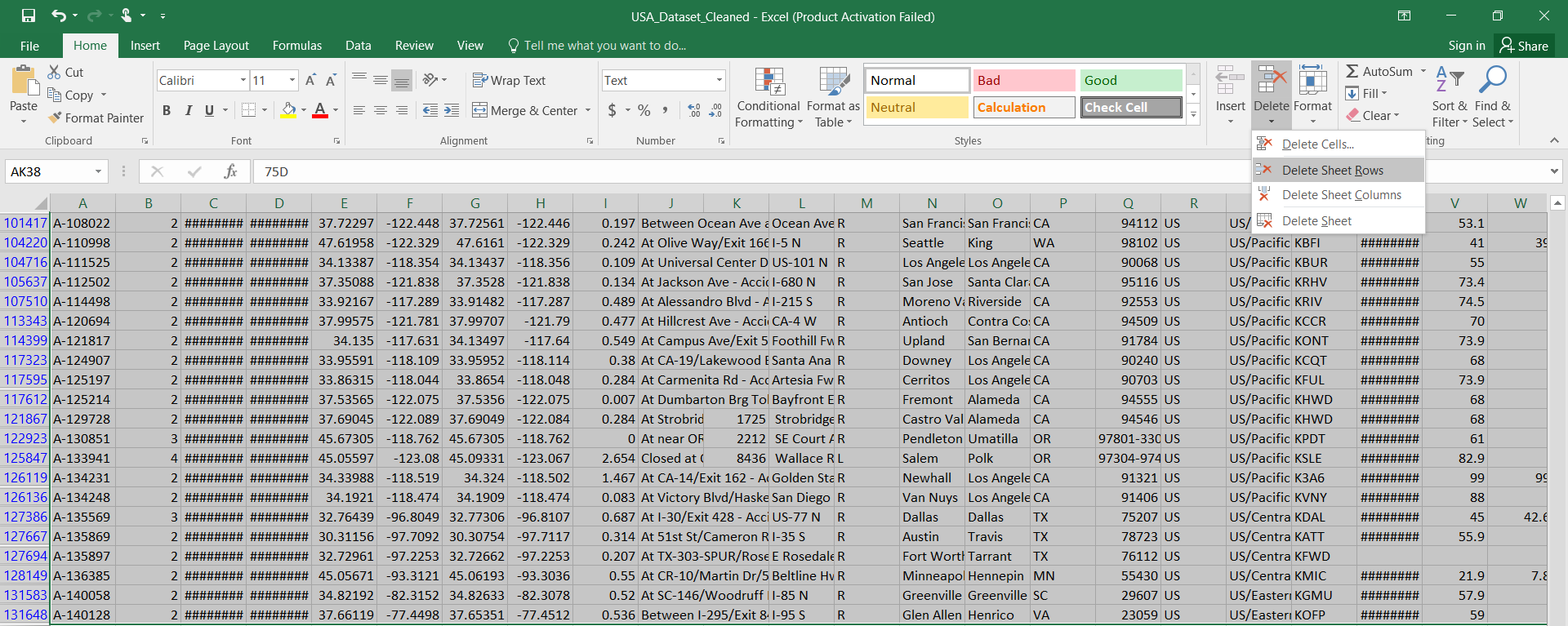
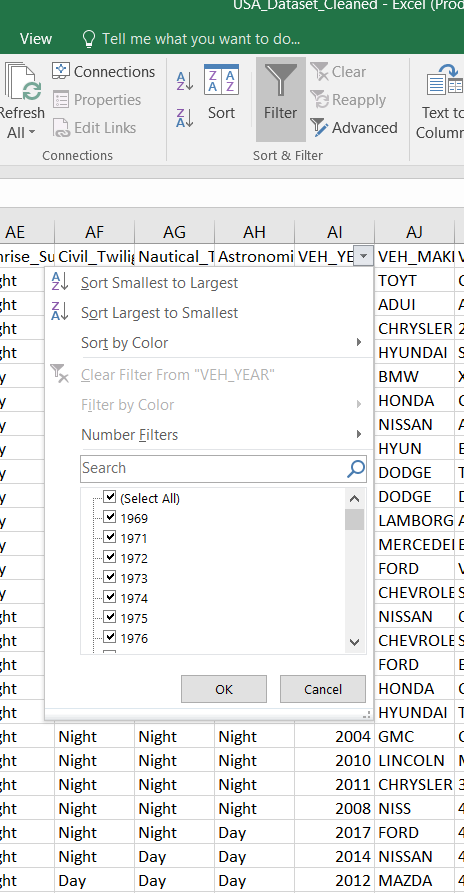


## Data Cleaning Stage

The first step in this phase is to remove unnecessary columns from each state for example Precipitation, Wind Direction, Pressure…etc. This means that we are keeping only the shared valuable columns as shown in the page 3 and 4.

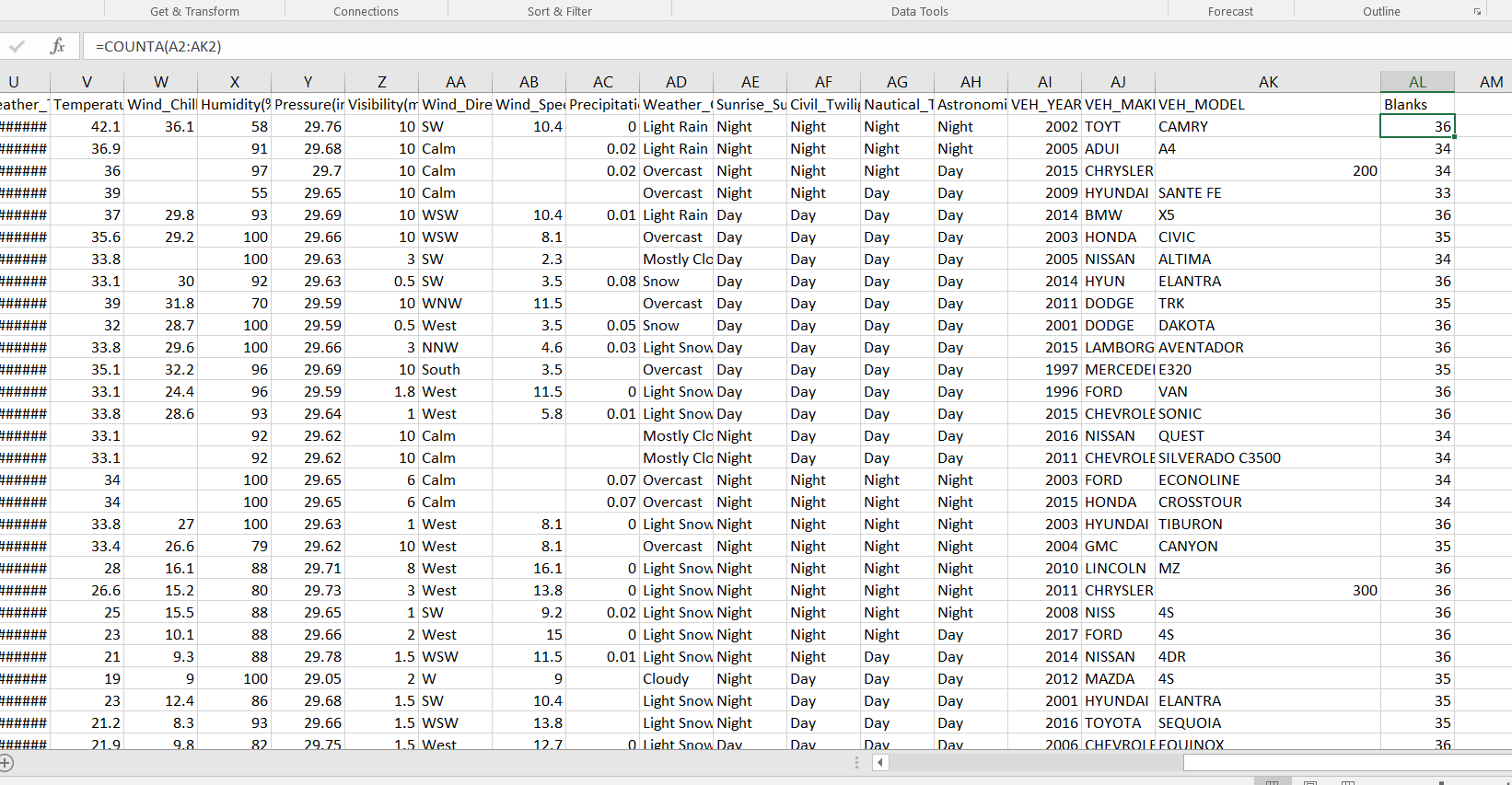


The second step is to remove rows with Nans/Unknown/Blanks/Illogical or ones that can make the result less trusty. To do that, we must select the target row then navigate to Data>Filter from Excel. The following action is to uncheck any unwanted value.



The third step is to remove the blanks caused by the previous delete operation. In order to successfully apply this modification, the following method is used:

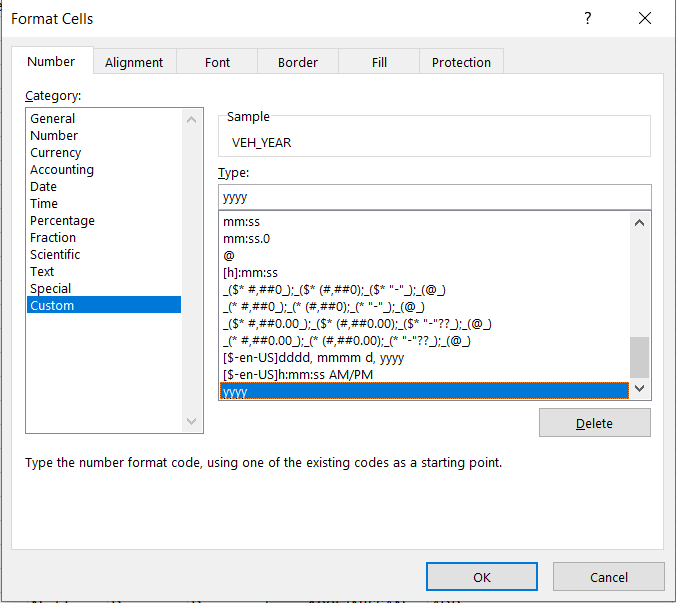
1. Adding a column named blanks that counts the number of columns filled within a row, if it is equal to zero then the row is totally empty. Otherwise, it must contain one or more values.



1. Then we navigate again to Data>Filter and we check zero. Now only the empty rows are showed. Finally, we apply a delete sheet rows from Home>Cells>Delete Sheet Rows.

The fourth step is to remove duplicate rows by mistake to prevent redundancy or falsifying the results.

The fifth and most crucial step is to change the format cells for each column depending on the type of values that holds.



Now we must have a full ready dataset for analysis use and exploration. This step guarantee the performance of this dataset as well as the prevention of misleading results.

**Exception:**

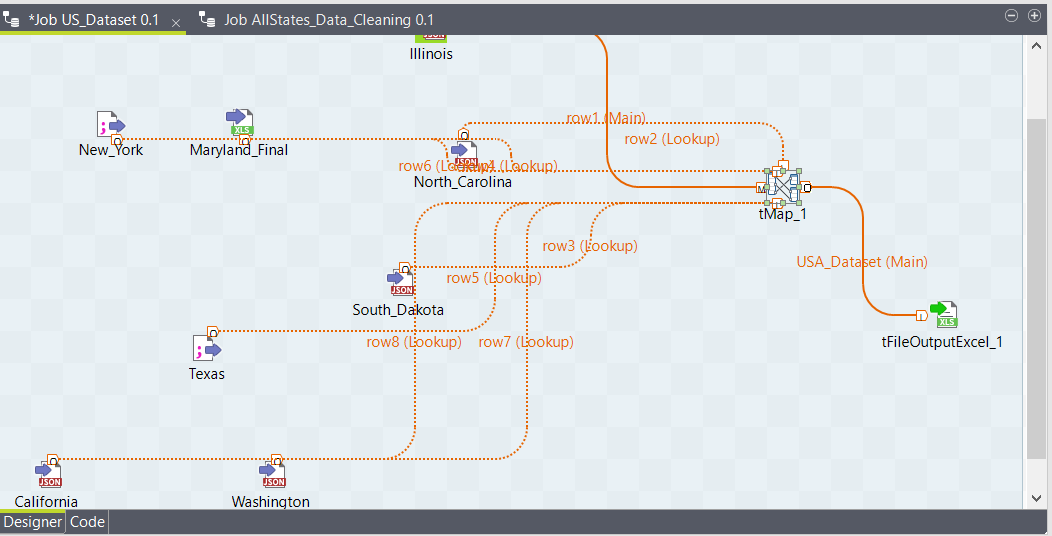
For KML, JSON, XML, RDF files are cleaned the same way using PowerBI due to visibility issues.

# Data Transformation

## Talend

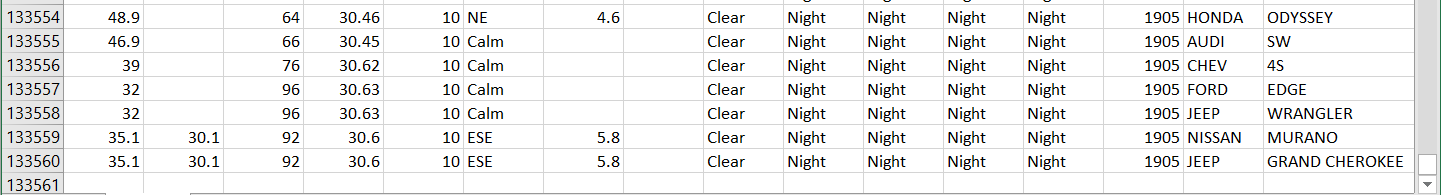
Merging data is the objective of this phase, thanks to Talend’s powerful component of data manipulation and mapping, we were able to merge data in a single excel file, after of course importing all states datasets to Talend.

Example:



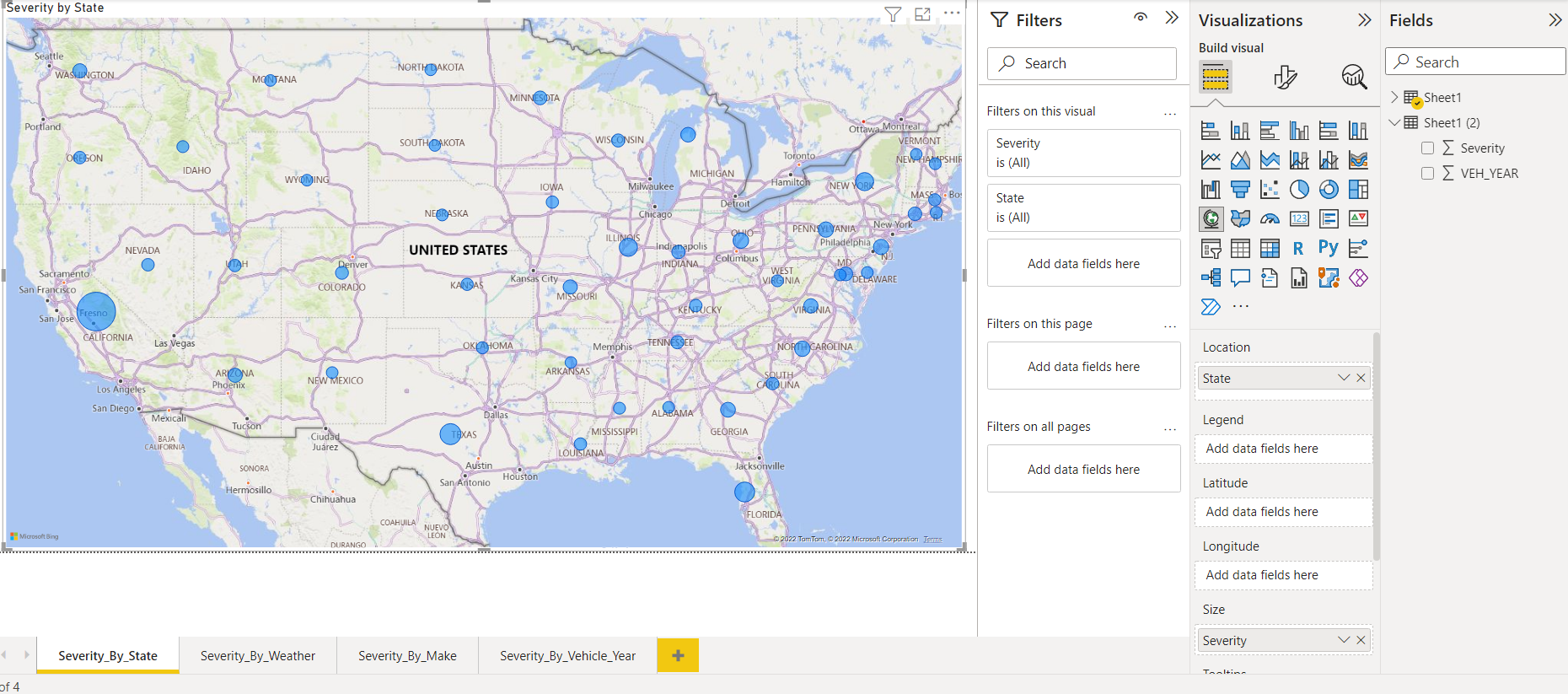
**Results**:

The total number of crashes before data cleaning were 2.4 million record and after we ended up with 133 560 record in the final excel file.



# Data Reporting

## Crash Severity By State



The following map shows the severity of crashes by states, the wider the circle, the more severe the crash is. California is the state with deadly to highly dangerous crashes especially the city of Fresno, which is widely known for illegal street racing.

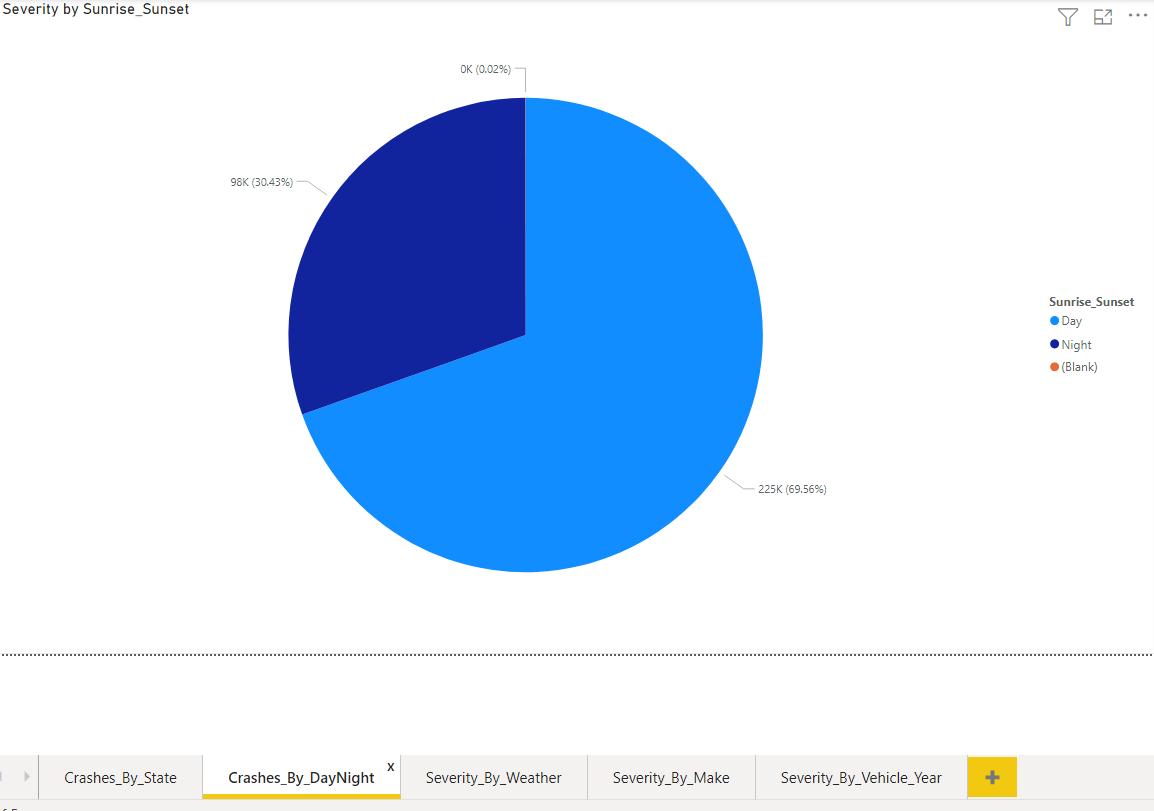
Otherwise, the other states have a severity ranging from low to mid (from fender bender to totaled vehicles with no/low injuries)

The states with severity below four are:

* Florida
* Ohio
* New Mexico
* North/South Carolina
* North/South Dakota
* Georgia
* Alabama
* Mississippi
* Texas
* Washington
* Idaho

This clarify that each state have a different rate of severity. This makes us wonder about some circumstances like weather, day/night and car features (like airbags, emergency-braking systems…etc.)

## Crash Severity By Day/Night

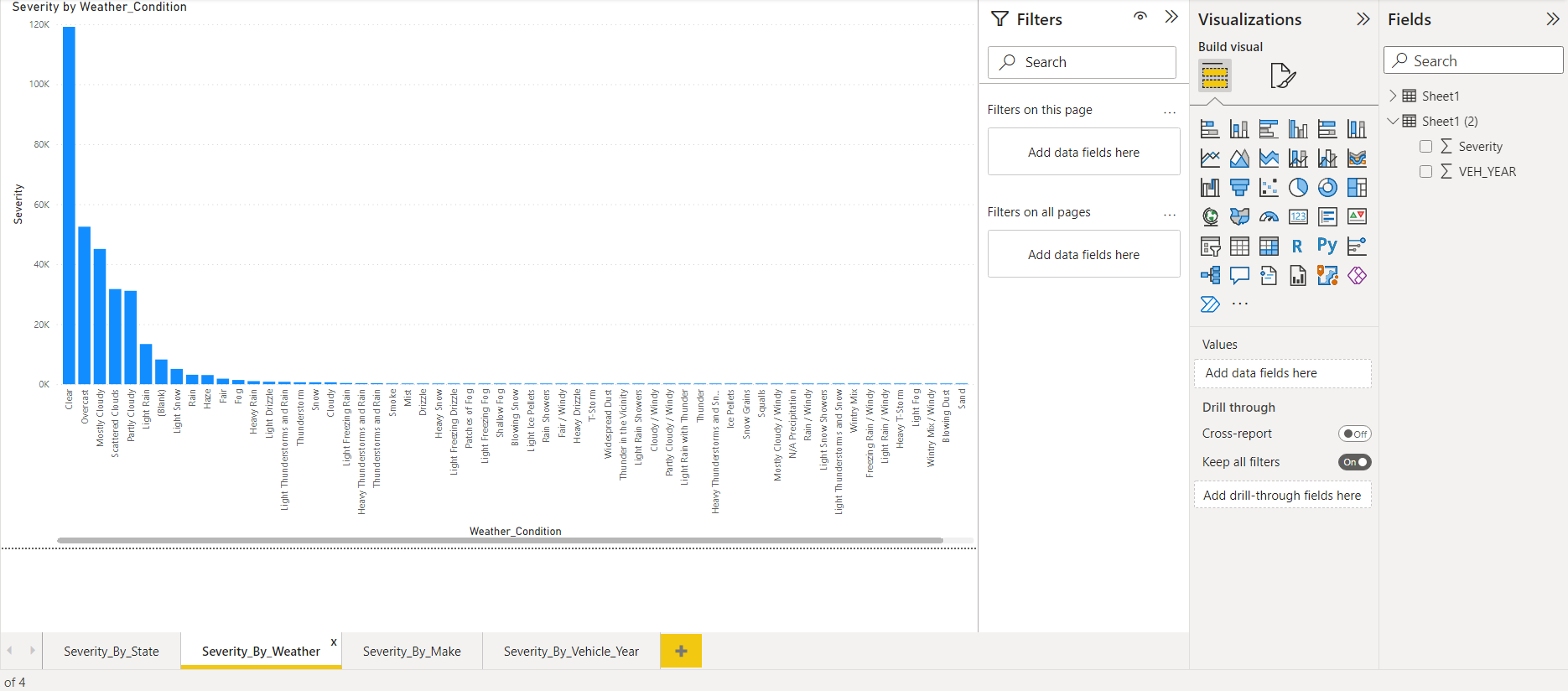


The US traffic during the day is relatively higher compared to it at night, this leads to more accidents even though the day. This pie graph ensure the validity of the previous statement with a total value of 69.58% of car accidents occurs during the day and 30.43% at night. We can state that accidents at the day are 2.3X times more than at night.

Right now we extracted two crucial information about this dataset

* California is the highest state in severity of accidents
* Crashes are willing to happen at day 2.3X more than at night

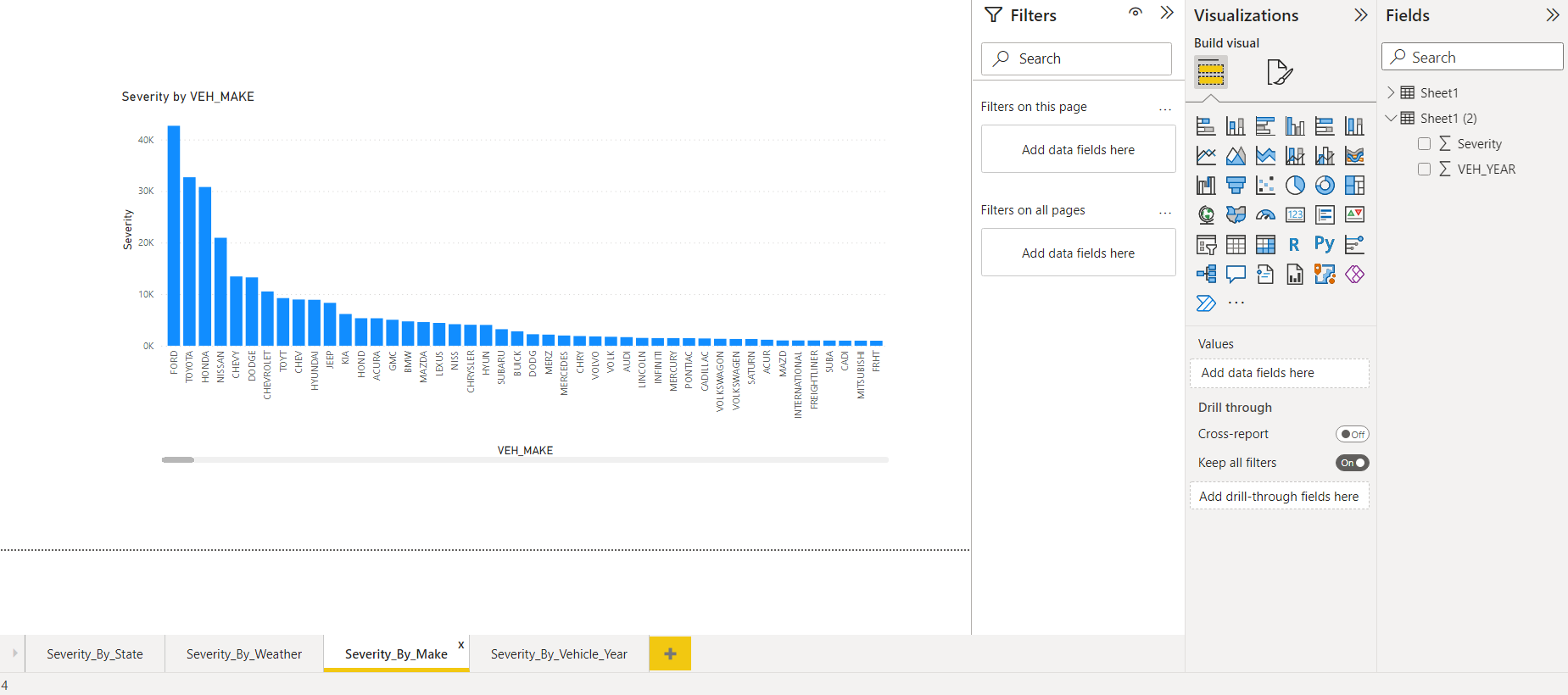
## Crash Severity By Weather



Everyone believes that bad weather is a leading factor to a severe crash. This chart proves that this statement is not always true, more than 120 thousand crashes happened during a relatively clear weather followed by cloudy (ranging from 55 thousand to 35 thousand accidents) and then a rainy weather (ranging from 15 thousand to 4 thousand accidents). However, other weather conditions counts too but they have a low to very low rate of crashes like snowing.

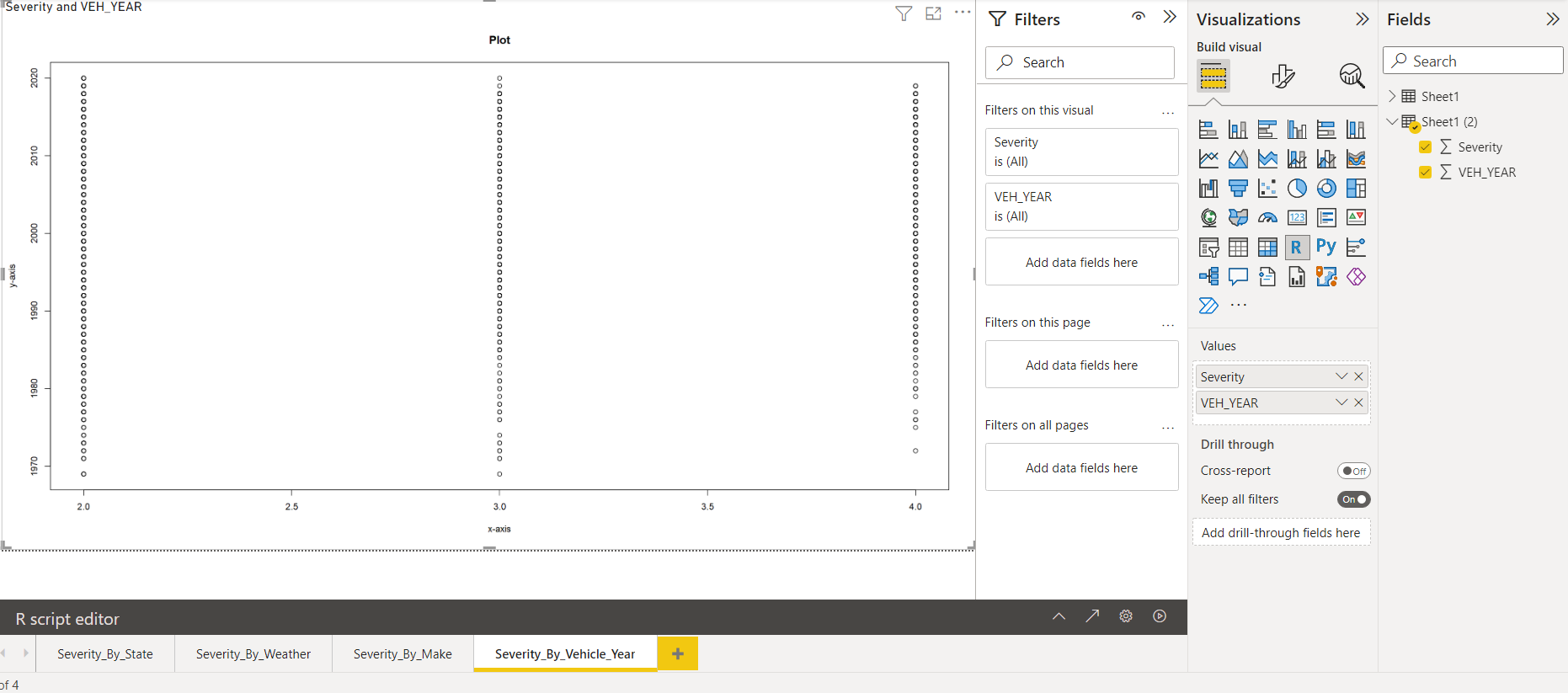
Since most accidents happens at the day during a clear weather, we can relate this with speed, since on a very bad weather condition pretty much every car driver slows down which leads to a lower rate of accidents. In the other side, clear conditions might help car drivers to accelerate due to a clear vision, which unfortunately leads to even more severe road crashes.

## Crash Severity By Make



Japanese cars are widely known for their low cost and cheap quality of build, which can cause severe damage during an accident, compared to European cars that were built using premium/solid materials plus the integration of high-end technologies. Also American cars have severe accidents rate, due to the fact that most of them have big engines, mostly V8 which can encourage drivers to speed on public roads. This graph clearly marks down five most severe crashes that are related mostly to American and Japanese brands (Ford, Toyota, Honda, Nissan and Chevy), were European cars are lower in severity and in the total number of crashes.

## Crash Severity By Vehicle Manufacturing Year



The graph shown proves a new statement, the severity of level 2 and 3 are pretty much occurred on every car generation, but we can clearly identify the non-existent of cars below 1975 on car crashes of level 4, this can be related to many factors, but most importantly speed, because old cars have relatively lower speed which lowers the severity of crashes. We can conclude that very old cars (below 1975) have a maximum severity of level 3 (totaled cars but no physical injuries).

# Conclusion

This study helped to extract five crucial criteria’s that exists on every severe crash. First, large state with a hot weather during the whole year exactly like California. Second, a crash is more likely to occur 2.3X times more during the day than at night. Third, clear or good weather conditions raises this rate too. Fourth, cheap overall built cars are the source of raising the severity of an accident. Fifth and last, slower cars have lower fatality rate, so speed is also a leading factor. To conclude, those five components are the factors of having a deadly to high danger accidents on US public roads.

**Large States + Daytime + Clear Weather Conditions + Cheap Built Cars + Speed = Severe Accidents**

