IE University
Data Visualization

Individual Assignment

Arturo Ford Sosa

October 16, 2022

Subject + Dataset

The dataset chosen for this project contains information regarding all vehicular transit infraction tickets issued within the city of Madrid. This dataset was chosen because it contains enough dimensions to be able to produce something of analytical value to authorities who want to know more about subsets of infractions. In its raw form it is useful for basic analyses, but when expanded through ETL processing it can be used to obtain insights of more specific nature. Data is split by months in the original repository, so for purposes of illustrating trends and KPI's, all data available for this current year was used. Availability starts in July, so we have the data for 6 months. In total, we accumulate 355MB of information which accounts for 1.4 million rows.

The original datasets are composed of the following fields:

- CALIFICACION: 'LEVE', 'GRAVE', or 'MUY GRAVE' depending on severity of infraction.
- LUGAR: Description of approximate address. Contains high number of dirty rows.
- MES: An integer for month in course (6 for June, for example)
- ANIO: Year in course.
- HORA: Time of infraction registration, in HH.MM format. Unfit for visualizing without reformatting.
- IMP_BOL: Amount fined for infraction.
- DESCUENTO: Unknown, unexplained field.
- PUNTOS: Presumably, points stacked against driver's license of person committing the infraction.
- DENUNCIANTE: State entity in charge of applying sanctions to the person committing the infraction.
- HECHO-BOL: Type of infraction committed, of which there are a large number of distinct types.
- VEL_LIMITE: Speed limit (only if infraction is based on breaching speed limit).
- VEL_CIRCULA: Speed at which the infraction was committed.
- COORDENADA-X: Supposedly, coordinates of incident at axis X. However, this field was found to be completely broken, with most coordinates present pointing to France.
- COORDENADA-Y: Supposedly, coordinates of incident at axis Y. However, this field was found to be completely broken, with most coordinates present pointing to France.

For this dataset to be viable for use in long-term dashboarding, it is important to create a proper process that will gather data from the public repository and process it for internal storage and future usage. For purposes of this project, we have created an ETL process that reads all files for this year, concatenates them into a single dataset, and processes the data to ensure as much cleanliness and cohesion as possible. This process grabs the combined dataset and performs the following tasks:

1. Reformat 'HORA' and 'MES' fields: The field with time of incidence has a weird format that most programs won't recognize easily: 'HH.mm'. Notice the dot instead of a colon. For our usage we reset it to 'HH:mm:ss' for easy recognition by PowerBI and other tools. Furthermore, the month and year fields are combined into a single month field formatted to

contain the value for the first of the month in question. This way we don't have to reconfigure the values as dates in our data visualizations.

2. Drop unnecessary and dirty fields:
    a. LUGAR: Place is too specific to be used in aggregation. A recommendation for those in charge of the data source would be to limit this to a recognizable street name nearby or specific address, otherwise it doesn't work for our purposes.
    b. DESCUENTO: It is not described anywhere what this field is, so we discard it.
    c. ANIO: We combined the month and year fields already; this one becomes redundant.
    d. COORDENADA-X: Unfortunately, the coordinate information provided in this dataset does not remotely correspond with what is expected. Most of these points land somewhere in France or elsewhere, but none in Madrid. Therefore, we remove them.
    e. COORDENADA-Y: Same as COORDENADA-X.
3. Change datatypes as needed: VEL_LIMITE and VEL_CIRCULA are converted into numeric types for calculations in our dashboards.
4. New reference fields are created from the HECHO-BOL field with the intention of flagging infractions containing specific words and be able to group by these types with ease:
    a. ALCOHOL: Infractions related to alcohol usage while driving.
    b. VELOCIDAD: Infractions related to traffic speed violations.
    c. NEGLIGENTE: Infractions categorized as 'Negligent behavior'.
    d. TEMERARIA: Infractions categorizes as 'Reckless behavior'.
    e. ESTACIONAR: Infractions based on parking violations.
    f. OBEDECER: Infractions based on ignoring orders from authorities.
5. Remove whitespace in field names: A number of column names had whitespace in them, making ETL processing and visualization more error prone. We have removed these whitespaces.
6. Save new, polished dataset in new CSV file.

The resulting dataset has the following structure:

```
In [28]: processed.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1419372 entries, 0 to 244933
Data columns (total 15 columns):
 #   Column        Non-Null Count    Dtype
---  ------        --------------    -----
 0   CALIFICACION  1419372 non-null  object
 1   MES           1419372 non-null  object
 2   HORA          1419372 non-null  object
 3   IMP_BOL       1419372 non-null  float64
 4   PUNTOS        1419372 non-null  int64
 5   DENUNCIANTE   1419372 non-null  object
 6   HECHO-BOL     1419372 non-null  object
 7   VEL_LIMITE    190825 non-null   float64
 8   VEL_CIRCULA   190825 non-null   float64
 9   ALCOHOL       1419372 non-null  bool
 10  VELOCIDAD     1419372 non-null  bool
 11  NEGLIGENTE    1419372 non-null  bool
 12  TEMERARIA     1419372 non-null  bool
 13  ESTACIONAR    1419372 non-null  bool
 14  OBEDECER      1419372 non-null  bool
dtypes: bool(6), float64(3), int64(1), object(5)
memory usage: 116.4+ MB
```

If deployed into production by the municipal authorities, all that is required would be to finish programming a dynamic way to process any file (current version receives these specific 6 files) from our end. However, to ensure that we can gather as many insights as possible with the least number of errors, it is necessary to talk to our data providers to ensure that they clean their dataset in the future. Specifically, fixing the coordinate and place fields as well as eliminating whitespace in the column names. For convenience and example, a copy of the code for the ETL process is in Annex A of this report.

Target Audience + Context

Our client for this project is the Madrid City Hall Mobility Manager. Their task is to ensure that transportation and human movement throughout the city of Madrid is as fluent and efficient as it can be. Because of the nature of their work, they need to make decisions based on city planning, infrastructure, movement fluidity, and resources to ensure that their main goal of maintaining fluency and efficiency can be achieved. Which avenue should we close for this year's parade? What lane can we convert into a bike lane without turning this street into a daily traffic jam? Why is there a pattern of drinking and driving in this part of the city? These are just a few examples of the questions that they might ask themselves on a regular workday, and they are questions that can be answered with the right data. However, they might have problems extracting the exact values they need or structuring the information in a way that would show patterns or allow the establishing of a few KPIs. With the right structuring of their data inputs, we can build enough dimensions to analyze whatever is necessary to understand patterns that might be invisible otherwise.

Because of their institutional importance, they have access to all kinds of information relating to anything that can be measured with the daily movement of individuals. One such information is their infraction dataset. Every month, they obtain a report with all the infractions issued throughout the city. This dataset, however, is raw and likely coming from an automated source. By cleaning this through an ETL process, we can turn this dataset into something of multiple usage for infraction analysis. Considering the amount of dimensions and the structure we have available here, the possibilities for use cases are several. Perhaps they would like to know at what times do alcohol-related infractions occur, or what is the institution that serves the most infraction citations and tickets. Or even what percentage above the speed limit are most individuals committing infractions going at before being flagged down by authorities. All of these, and way more, can be answered by ordering the dataset in a particular way and graphing for it. For purposes of this report, we are going to build a dashboard which will provide statistics on some of the most common types of infractions, namely: Alcohol, Speed, Negligence, Recklessness, Parking, and Obeying Authority. The idea is to provide the first point of information for these types of infractions so that the municipal authorities can make the best informed decisions to curb the occurrence of these infractions as much as possible.

EDA – Exploratory Data Analysis

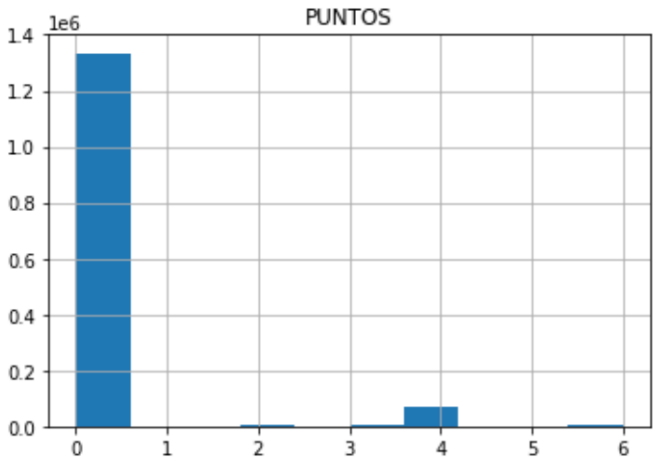
Once we have finished processing the data, we can perform some EDA on the resulting dataset to verify that there are no consistency problems and that the data behaves as expected, including cleanliness and a lack of null values where possible. When querying for basic statistical data on the numeric fields in our dataset, we get the following:

```
In [40]: processed.describe()
Out[40]:
             IMP_BOL     PUNTOS  VEL_LIMITE  VEL_CIRCULA
count  1419372.0 1419372.0    190825.0     190825.0
mean       118.5       0.2        71.2         82.2
std         72.7       0.9        12.5         13.3
min         30.0       0.0        20.0         33.0
25%         90.0       0.0        70.0         75.0
50%         90.0       0.0        70.0         78.0
75%        200.0       0.0        70.0         94.0
max       1000.0       6.0        90.0        160.0
```

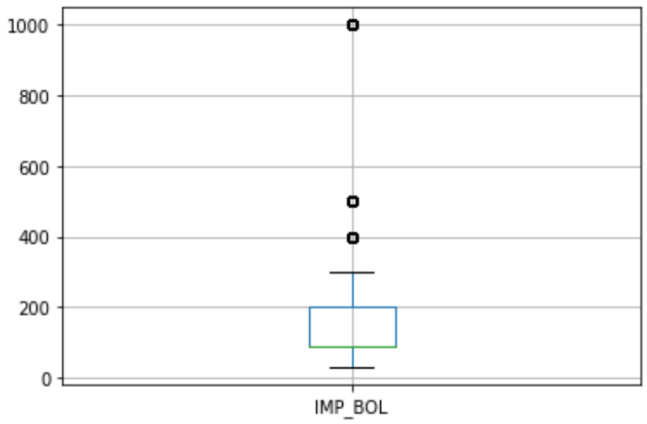
There results are consistent with what is expected of a clean dataset of this nature. All fields contains the correct minimums and maximums and the ranges of values are within the expected. The only thing that could be considered questionable is how the count of rows is different between fields, but this is something we can answer by checking for nulls:

```
In [32]: processed.isnull().sum()
Out[32]:
CALIFICACION          0
MES                   0
HORA                  0
IMP_BOL               0
PUNTOS                0
DENUNCIANTE           0
HECHO-BOL             0
VEL_LIMITE      1228547
VEL_CIRCULA     1228547
ALCOHOL               0
VELOCIDAD             0
NEGLIGENTE            0
TEMERARIA             0
ESTACIONAR            0
OBEDECER              0
dtype: int64
```

By looking at the above query of the dataset, we can see that nulls are not present anywhere in except for the VEL_* fields. This, however, can easily be explained by the fact that those fields are only populated when the infraction that took place is speeding-related. Therefore, we can conclude that the dataset is clean enough to proceed with further analysis. For example, we can check how many driver license points are issued commonly based on the severity of the infraction:



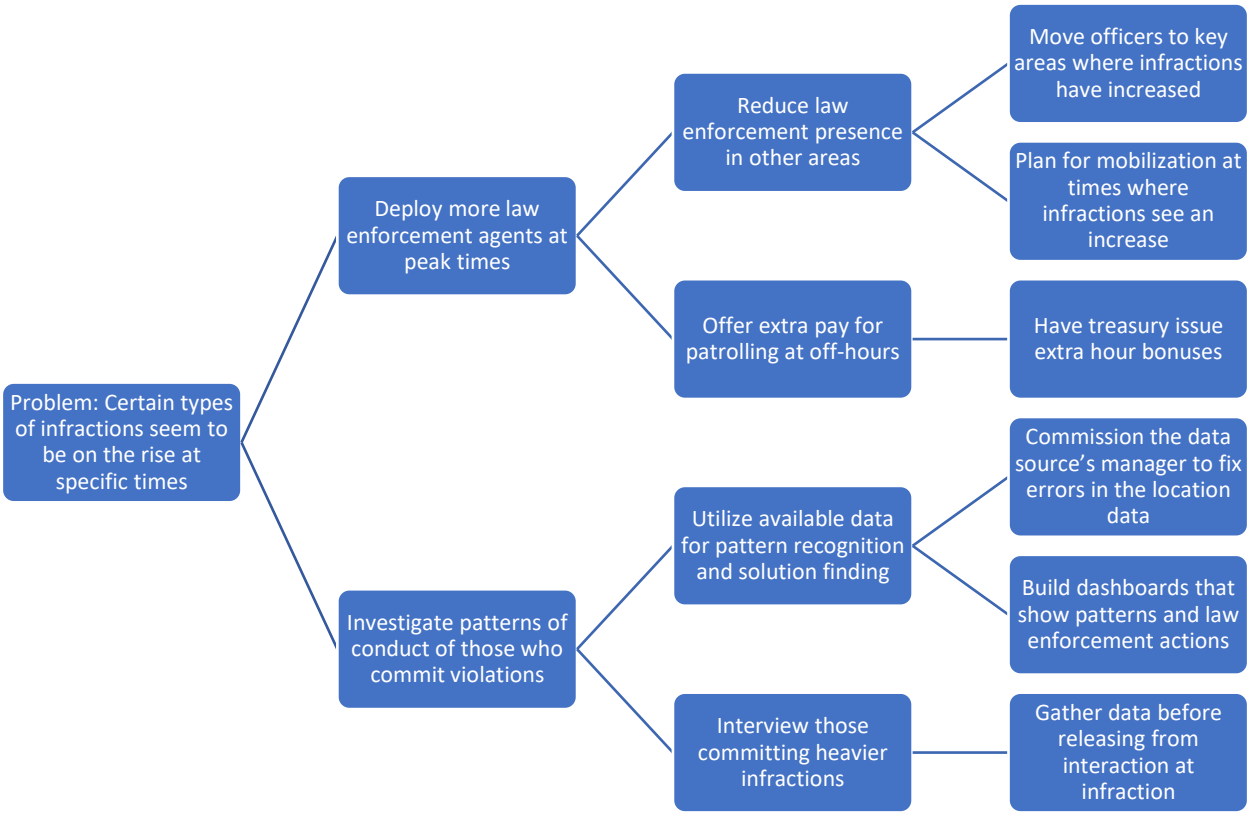
When plotting a histogram of points, surprisingly enough, it seems that the vast majority of tickets issued do not give any points against the driver at all. This probably means that for the most part the infractions are minor and don't require a threatening sanction against the person committing the infraction. This can also be seen when analyzing the cost of said infractions against the driver, as seen below:



This graph shows the same story as the one before, where minor infractions account for the vast majority of the total. Rarely do any tickets issued go above 400-euro fines, they are all considered outliers.

SPSI – MECE


Situation
•No cohesive dataset to check for specific infraction types; little readily-available information.
•Authorities are unsure on how well they are doing with enforcing traffic laws
Problem
•Certain types of infractions seem to be on the rise at specific times
•Authorities have no way of knowing how effective they are when compared by department.
Solution
•Build a dashboard with statistical data regarding particular types of infractions across time.
•Build a second dashboard with comparative information between departments, including KPIs and benchmarks.
Impact
•Time-based statistical data enables law enforcement to focus resources at more relevant timeframes and maximize impact.
•The state can better utilize resources by better understanding how effective every authority is at enforcing traffic laws.



Storyboard & KPI

Based on the information available, there are a number of stories that we can tell the municipal authorities to better tackle traffic infractions:

- Based on the MECE analysis used, we can create graphs that show how certain types of crime occur throughout the day. Based on this we can identify trends. For example, if alcohol-related infractions occur mostly in early hours of the morning, we can come up with the story that drivers take their cars after enjoying drinks out with friends. If we notice these kinds of infractions have been on the rise, then we need to identify a trend that would tell why greater alcohol consumption is happening. This would allow us to take action by targeting the root cause and aim for the pattern to drop to under a pre-established benchmark.
    - Chart idea: Amount of fines issued by minute of day, grouped by type of fine. One for every type of infraction studied. Hypothesis is that this might show a pattern, but it might also show that infractions occur at random times throughout the day.
    - Clustered bars displaying the three types of fines over the months. This way we can see if heavier infractions got more common or not.
- Following this logic, we can also measure when heavier infractions (such as negligent and reckless) take place as well. This would allow us to benchmark an 'allowed' amount per month, which could be used as a KPI to see how effective law enforcement is at curbing these infractions. The story is that the more authorities

hit their KPIs, the more effective law enforcement is, and infractions should remain lower.
    - Create a chart that shows how prevalent these types of infractions are over the months. This way we can identify cyclical patterns and find proper benchmarks to keep these numbers low.
- Since we also have the information by authority department, we can conduct comparative analyses that would show how effective each department is at tackling their respective infraction rate. Further analysis would be required for this, but authorities could establish metrics of success and measure themselves by those through this dashboard. All that would be required would be to find a 'fair' benchmark. The story here is that all departments must keep each other at a high standard in order to collectively succeed. We first see where every department is standing at present, then we compare. If we identify that the trend is that departments are not working as good as they should, then we go looking for a metric that all departments can comply with in the future, such as requiring a percentage of all tickets issued by them. The actionable would be implementing these measures as department policy to see how they influence overall infraction statistics after.
    - Total money inflow by infraction type
    - Total money inflow over the months.
    - Count of infractions by department
    - Total points issued this year, could be useful for benchmarking driver well-behavior.

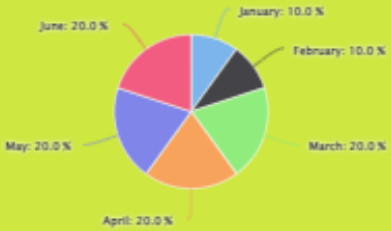
Sketch/wireframe and visuals

Dashboard 1: Infraction analysis

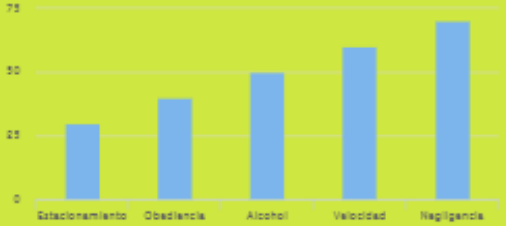
Amount of fines over the months, totals:
Option 1: Pie chart
Option 2: Scatterplot
x: Month
y: Fine count

Amount of alcohol-related fines over time of day, same as total amount of fines on the left.
Option 1: Bar chart
Option 2: Scatterplot
x: Time of day
y: Fine count

This would go top left



A graph like this would go in the other three corners



Amount of speed-related fines over time of day, same as alcohol-related above

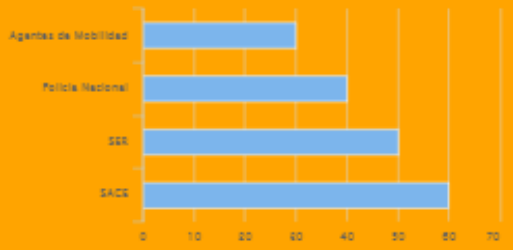
Negligent and Reckless infractions, in clustered columns grouped by count per month

Dashboard 2: Department aggregates and at-large analysis

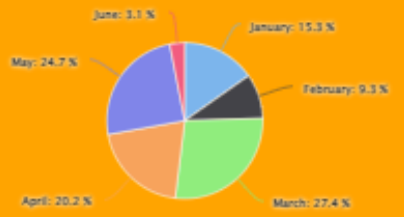
Total money inflow by infraction type in 2022

Total infraction cash inflow by month

This would be used for infractions by department



And this one could be used for total infraction points per month



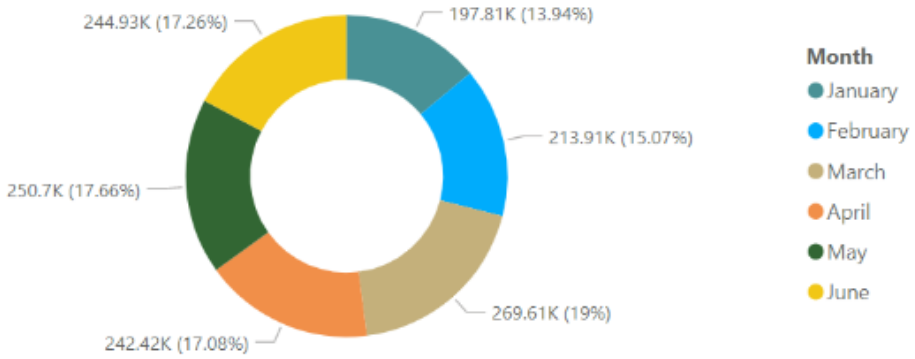
Infractions by department:
Option1: Pie chart
Option 2: Bar chart

Infraction points issued per month

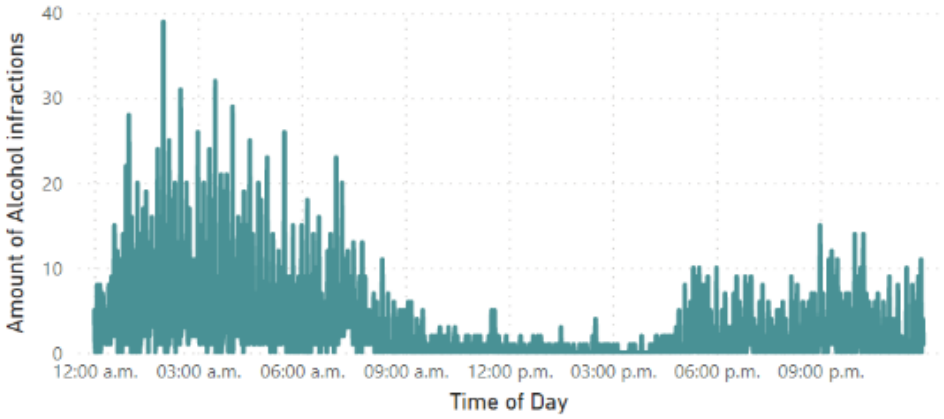
GOBIERNO DE ESPAÑA
MINISTERIO DE TRANSPORTES, MOVILIDAD Y AGENDA URBANA

Infraction Analysis Dashboard

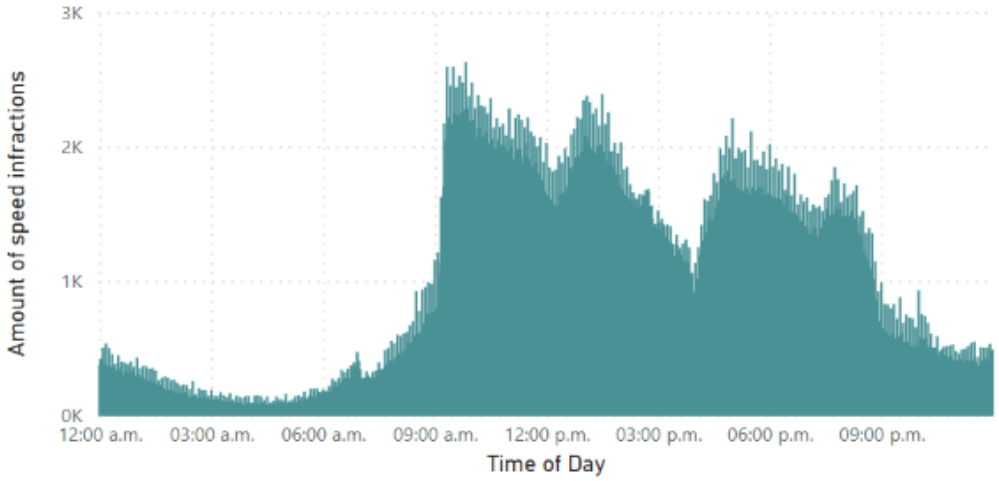
Total amount of fines, grouped by months



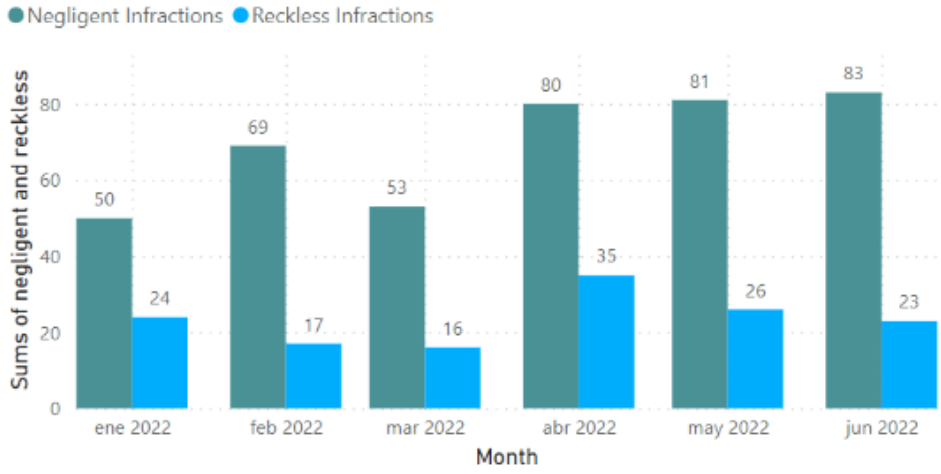
Amount of alcohol-related fines over time of day



Amount of speed-related fines over time of day



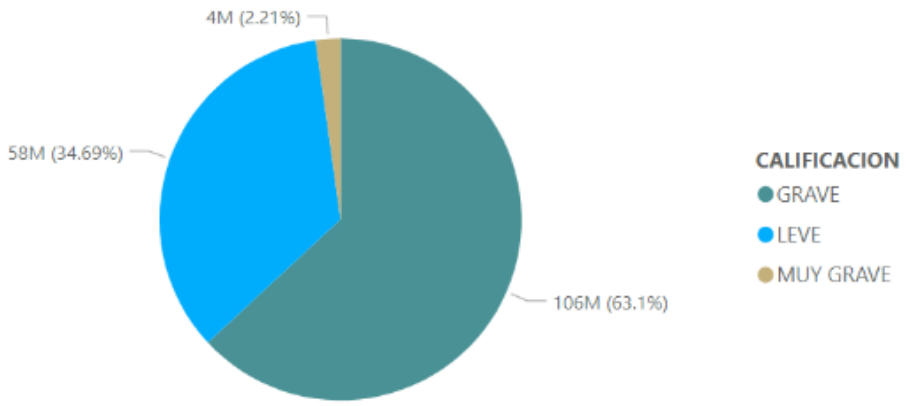
Negligent and Reckless infractions in 2022



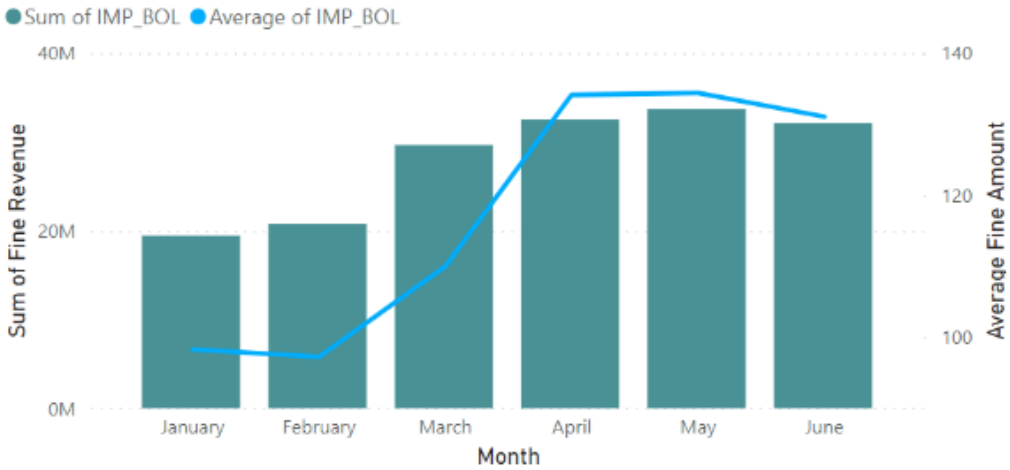
GOBIERNO DE ESPAÑA | MINISTERIO DE TRANSPORTES, MOVILIDAD Y AGENDA URBANA

Aggregates and At-large Analysis Dashboard

Total money inflow y infraction type, 2022



Total infraction cash inflow by month, with average infraction fine amount



Infraction amounts by department

Sum of Infraction Points imposed on drivers