# **Dataset: Claims**

1. **Based on dataset 1, can we say which factors could be used to identify fraudulent claims?**

To answer this question, Data understanding is needed. The first chart shows the distribution of damage according to the level of education. Fraud was detected in major damages across all levels of education, but more likely in "Graduate".

Chart, pie chart

Description automatically generated

The second visualization represents the gender age distribution by accident types. While fraud was not detected in the majority of "no collisions", it is observed that the average age of men is higher than that of women. Only the number of men detected in "front collisions" is higher, while others are evenly distributed. This also applies to those who are not identified as fraud.

Chart, bar chart

Description automatically generated

Third, the relationship is shown between fraud and claim type. Fraud was detected in approximately 28% of vehicle accidents. Fraud was also detected in two other types of accidents, with 27%.

Chart, bar chart, waterfall chart

Description automatically generated

The fourth table represents the relationship between fraud detection and month as a customer. Fewer fraud was detected in longer-term collaborations. In cases identified as fraud, the customer relationship is 183 months, while in those who are not identified as fraud, this period is 349 months. So, the fraud rate in long-term collaborations is lower, and the fraud rate in short-term deals is higher.

Chart, bar chart

Description automatically generated

After tableau, Parallel Coordinates are done to visualize relationships and patterns between features, representing “yes” as green (2) and “no” as yellow (1):

A picture containing text

Description automatically generated

This shows the main difference in features between fraud or not fraud. Vehicle, injury, and property claim that compose the total claim, annual policy, month, and age, all of them decrease when having a fraud. It is important to highlight that the colors here were chosen being the darkest one (green) as the “yes” since it is the focus.

Finally, for the last visualization, UMAP, based on the relationships discovered previously, the subsets can be defined also using the correlation heatmap:

Chart, treemap chart

Description automatically generated

The correlation is high in the following: injury, vehicle and property claim with total claim amount; month as a customer with age. These also have causation, meaning one causes another. For instance, having more months as a customer can mean that the age of the customer is greater. These features are not going to be removed, but placed in different subsets since it is considered necessary for the analysis:

Text

Description automatically generated

Starting with Subset1:

UMAP1

Subset2:

Subset3:

Subset4:

Subset5:

In conclusion, the main factors that identify fraudulent claims are:

# **Dataset: Collisions**

1. **Based on dataset 2, can we say which factors contribute to severe vehicle damage?**

To answer this question there are several tasks to do before. First it is necessary to understand the dataset using Tableau. Two dashboards are created containing similar attributes, those describing weather and road and those about the driver and vehicles.

Chart

Description automatically generated

Since the focus are damages labeled as “Severe”, the first visualization is created to verify how is the balancing of the dataset. It can be seen that “Severe” damages only compose 7% of the complete dataset, while “Light” has 93%. It is important to highlight that if a machine learning model was going to be done, it will be necessary to balance the data before using it.

String parameters are applied to those similar attributes, such as Weather with Road and Point of impact with Manoeuvre.

Graphical user interface, application

Description automatically generated

On the second dashboard, the characteristics between the driver and vehicles are shown. It is important to highlight that the colors of all the visualizations is uniform having purple as “Light” damages, and “Severe” as green. Additionally, since the focus of the analysis is about “Severe” damages, a filter was applied:

Chart, bar chart

Description automatically generated

Graphical user interface, application

Description automatically generated

So, there are some important insights:

1. Most of the “Severe” collisions have the same causes than “Light”, for instance generally “single carriageway” has more collisions.
2. Most of the “Severe” collisions occur with Daylight, when driving forward and with dry weather.
3. The number of vehicles involved were 2 more likely, having more crashes with males involved in their 20s-30s. The speed with more collisions is between 25-35 km/hr.

After this, a Parallel Coordinates visualization is done to find out about the relationships between the variables where the green color represents “light”, while the pink symbolizes “Severe”:

A picture containing chart

Description automatically generated

The features are similar for both damages type, but engine capacity, age and number of vehicles are different. “Severe” tends more to less engine capacity and from 2 to 6 vehicles. It is important to highlight that the colors were adjusted, since “Severe” had less collisions, it had the clearer color.

The Third visualization is UMAP, in where four subsets were implemented revised from the previous visualization analysis. This process started in Tableau when the visualizations were done with into attributes that formed a logical group, confirmed with the Parallel Coordinates. Then correlation and causation are revised before defining the subsets:

Chart

Description automatically generated

In this case, since the correlation is below 0.5 and there are not that many attributes, no features are going to be eliminated, so the final subsets where the following:

Text

Description automatically generated

The four UMAPS are shown, starting with Subset1:

Chart, scatter chart

Description automatically generated

There is no clear information that can be seen since both type of damages is similar having more collision when the vehicle is going ahead and turning.

Subset2:

Chart, scatter chart

Description automatically generated

The main clusters have the following features:

Text, table

Description automatically generated

Subset3:

Chart, scatter chart

Description automatically generated

The results of the clusters were compiled in:

Text, table

Description automatically generated

Subset4:

Graphical user interface, scatter chart

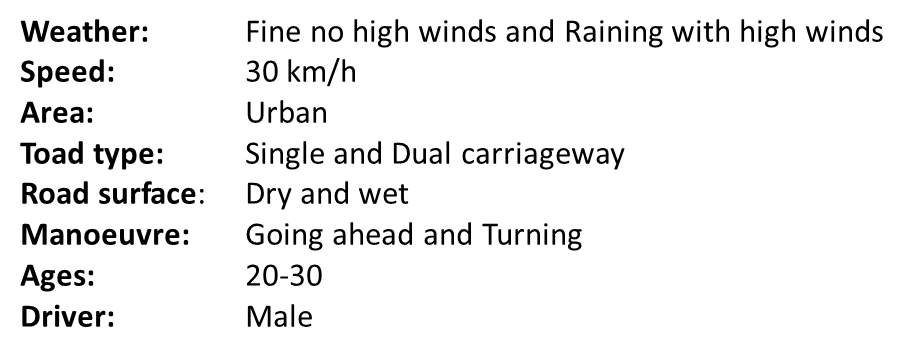
Description automatically generated

Again, it is difficult to see the features relationships, but looking closely:

Table

Description automatically generated

In conclusion, it can be said that the main factors for a “Severe Damage were:



Finally, the main reason for the choice of the visualizations is that with Tableau easy to use and easy to analyze when getting to know the data. Parallel coordinates help to review the relationships in a single visualization with all the features.