## **Crash investigators:**

Thomas Ziller (h11739403) Mirza Gazdic (h1453681)

#### Overview

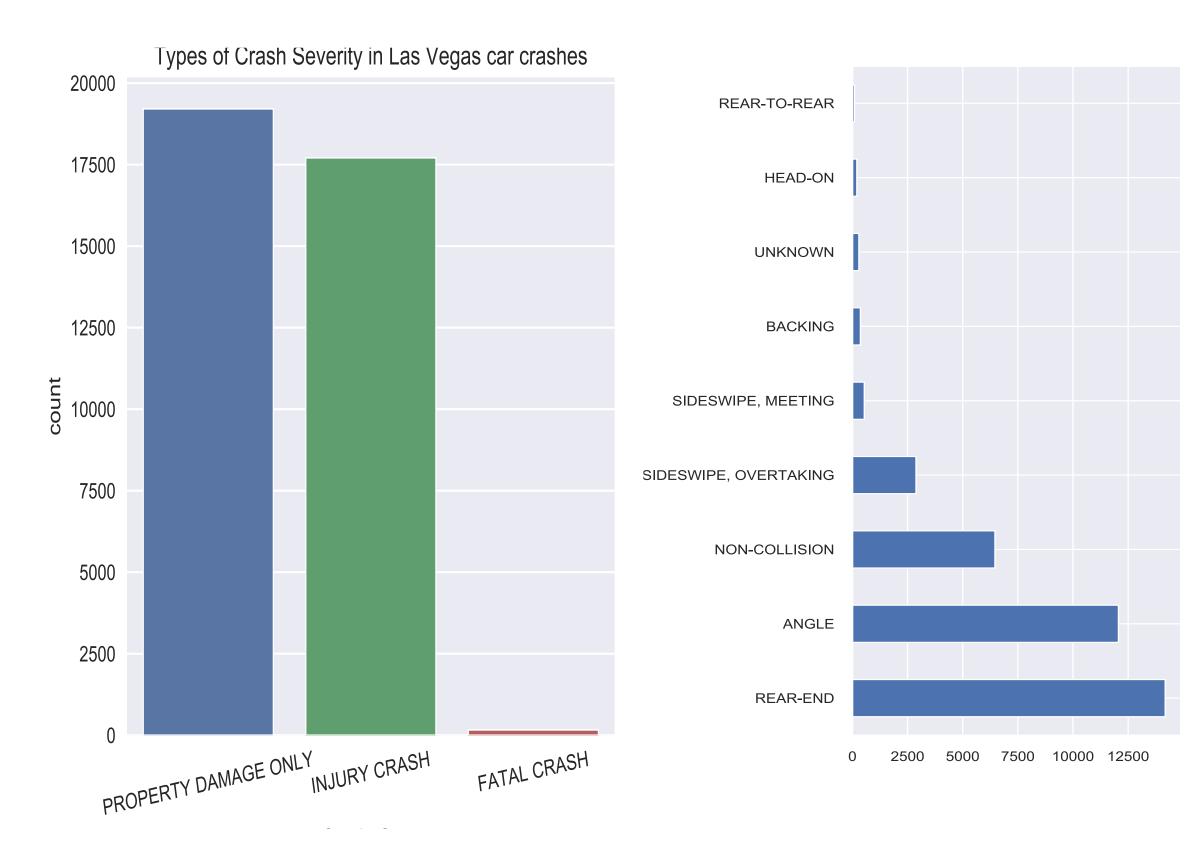
The purpose of this project was to investigate causes of car crash accidents, and to see if there is a possible correlation with some of the given variables. According to the "Association for safe international travel (AIRT.org)" nearly 1.25 million people die in road crashes each year which is on average 3,287 deaths a day. Car incidents and traffic crashes are one of the biggest reasons for sudden deaths in world. Death from car accidents are way higher than for any other means of transport.

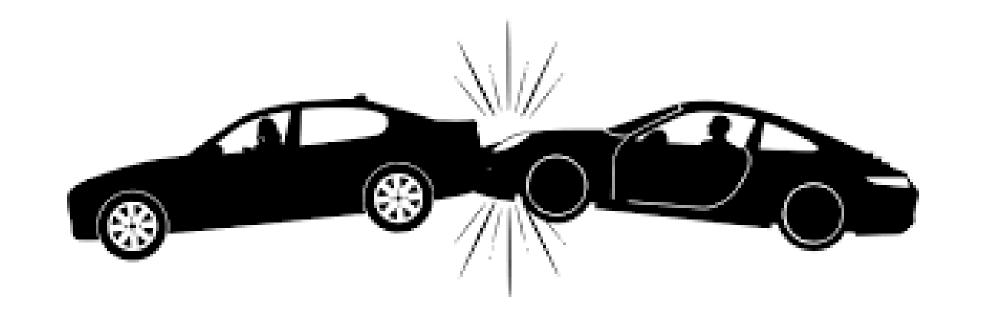
Additionally, when looking on the analysis of car crashes we did not manage to find any significant dataset with relevant variables trying to see the background of car accidents. In our dataset we have more interesting variables, which could lead to the further analysis and possible reasons for these accidents.

We are hoping to see that factors that impair drivers actions such as drug/alcohol use, bad road/weather, increase in speed, etc. have a bigger effect on the outcomes of the crash.

# Research questions

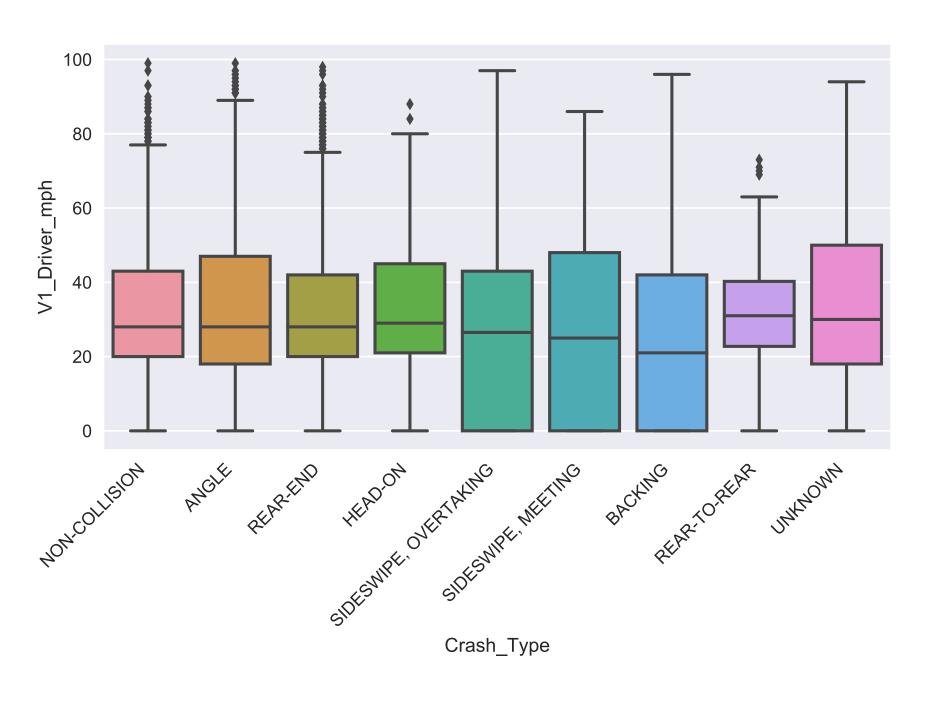
- 1. Does crash severity (property damage, injury crash, fatal crash) depend on the variables such as speed of the driver, state of the driver and drivers actions? Are there also any impacts on crash types?
- 2. Could the different crash type (angular, sideswiping, rear end, head-on crash) correlate with the speed of the driver, its state, road factors and weather as well?

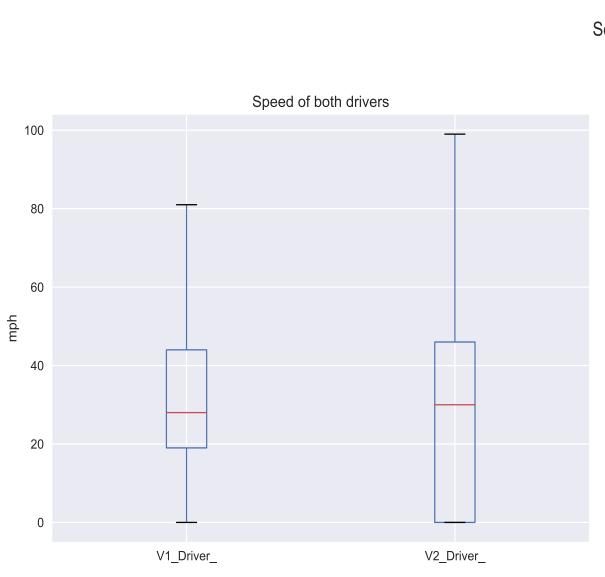




## Dataset(s)

For our research we will be using a data set on Las Vegas car crashes acquired directly from the "Open Data Las Vegas" site. This data set is available in the CSV format, but it had to be cleaned and reshaped as it was messy.





## **Steps**

First we loaded the dataset and imported the tools we will be using for processing. After that we scanned our data set for duplicates, and found we had no duplicates. However, we than scanned it for missing data and found a big problem there as a lot of our columns had more than 90% of missing data. We proceeded by dropping most of these columns as they were also not relevant for the research. The ones that had missing values but were relevant we imputed with backward and forward fill method. Furthermore, renaming and reorganization of data set had to be done, as some of the columns were not clear enough. Lastly, after the data was organized and processed properly, we proceeded with analyzing, plotting heatmaps in order to show correlation of our columns with our main variables.

### **Tools**

Working with this data set required also different libraries, some of them that we used were csv, numpy, pandas. For the visualization purposes we used seaborn and matplotlib.pyplot. They were chosen on the basis of familiarity as well as big interphase, so that our analysis would be more effective.

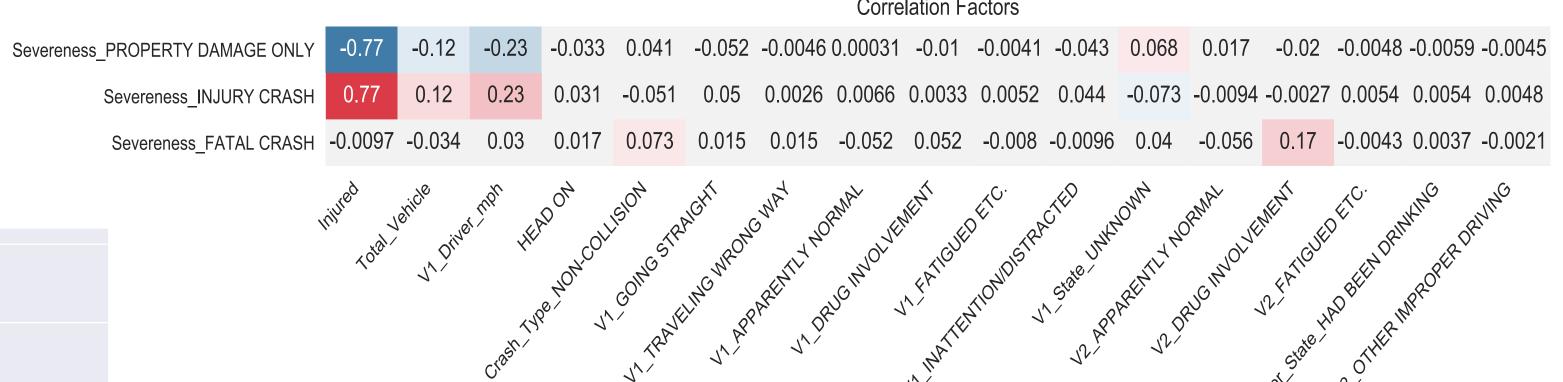
### Structure of the notebook

It was proposed to us to have a certain structure, which we stuck to throughout the whole project. Here are the titles of each section of our notebook:

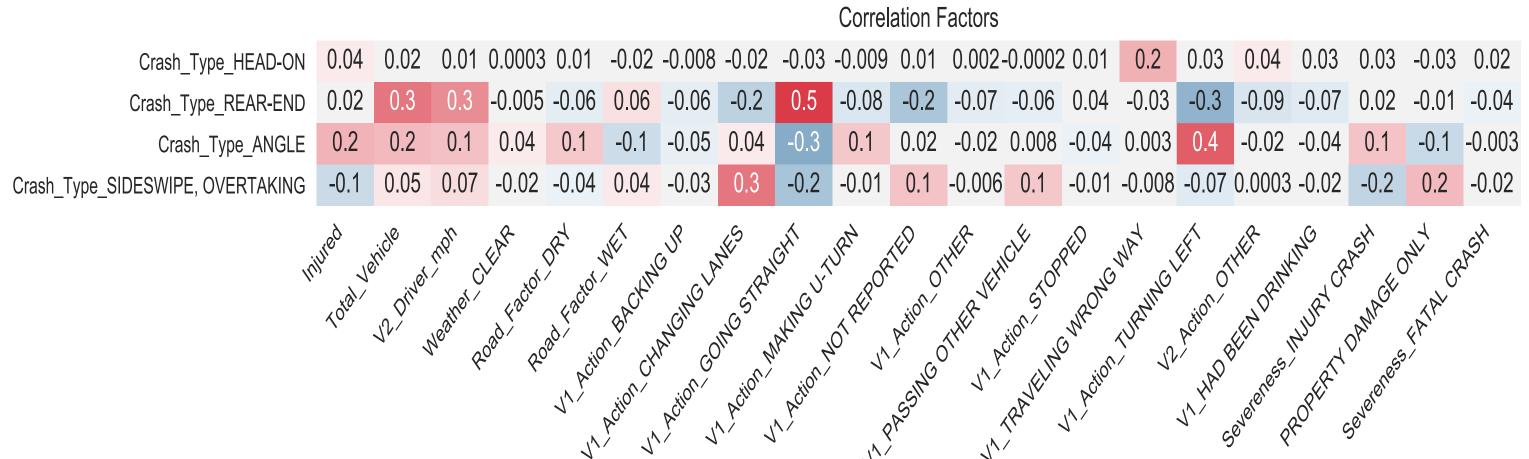
Team - Research introduction - Data source - Data Inspection - Data Cleansing - Data Analysis/Visualization - Conclusion and future impacts

#### Results

1. Does crash severity depend on different variables? By looking at the correlation with significant variables taken we can see that the crash severity depends mostly on the negative effects on the driver (speed, state, type of a crash ).



2. Could the different crash type correlate with the speed of the driver, its state, weather, etc.? The column crash type seems to correlate with variables such as drivers action, its speed, but not so much the weather or drivers state.



#### 3. Limitations and Lessons Learned

We faced many limitations during our project. One of them would be a dataset with too many categorical variables. Furthermore, too many missing values created additional problem for significant analysis. Additionally filling those values with any imputation method lowers the reliability of the results. Even though, the dataset comes from a reliable data source, it is always necessary to check for missing values and duplicates! In the end it would be very useful to have additional dataset from Las Vegas with more numerical values, thus increasing the opportunities for the analysis and visualization.

### References

- •https://opendataportal-lasvegas.opendata.arcgis.com/datasets/traffic-crashes
- https://www.asirt.org/safe-travel/road-safety-facts/
- •https://stackoverflow.com/

0.45 0.30 0.15 0.00 -0.15 -0.30