CITADEL | CITADEL | Securities

APAC Datathon Report

Ariel Albert, Darryl Chan, Murtaza Pakawala,
Rachit Tamrakar

17 April 2023

Executive Summary

Philadelphia has one of the highest vehicular fatalities in the USA even though it has the lowest car ownership demographics for the city of its size. As such, we were given 5 years worth of vehicular accidents data to find novel insights and recommendations to improve road safety within the city.

Hence, for our group, we posed the question:

- What is the relationship between the impairment of the driver and the severity of the crashes?

This problem is important as the insights derived from finding the answer could reform or adjust the current way we punish impaired drivers. Additionally, it could tell us more about the behaviour of impaired or non-impaired drivers

Findings

Our group considered impaired by considering the state that the driver was in as the vehicle crashed. We included categories such as being fatigued, drunk or drugged as impaired.

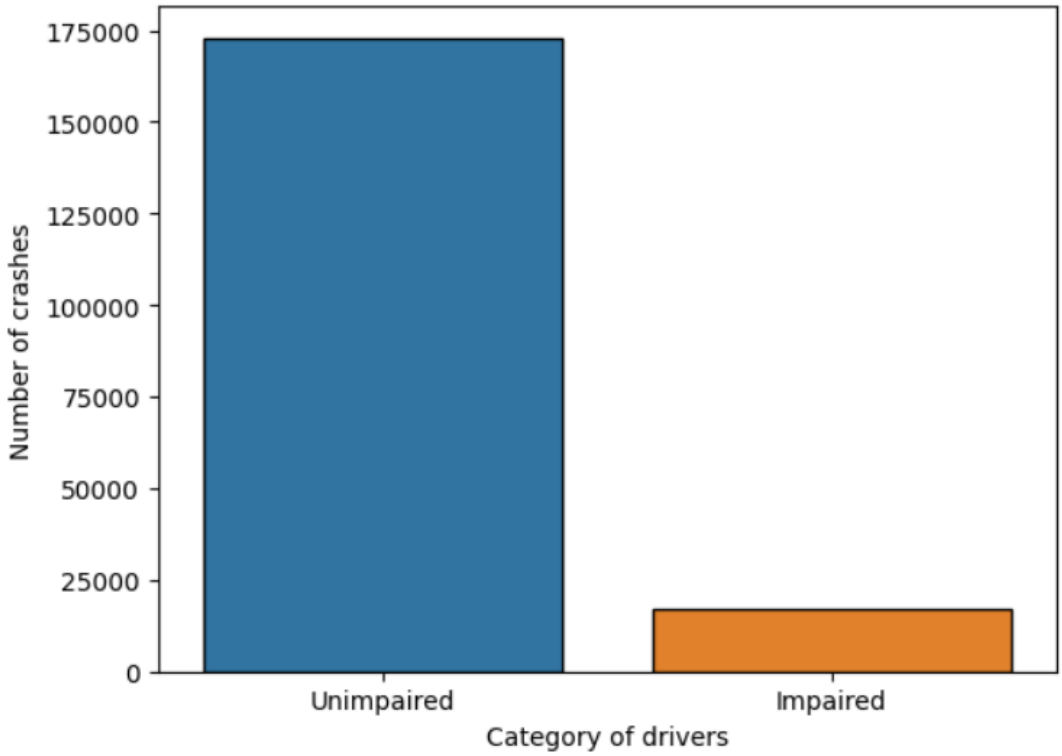

Figure 1: Number of crashes caused by unimpaired and impaired drivers

Our group found that the amount of unimpaired driver crashes to impaired drivers is about 172,950 to 16,995 which is roughly 10 : 1. Meaning that for every impaired driver crash, there are about 10 unimpaired driver crashes. Out of these Impaired drivers, we would like to find the distribution among the types of impairment

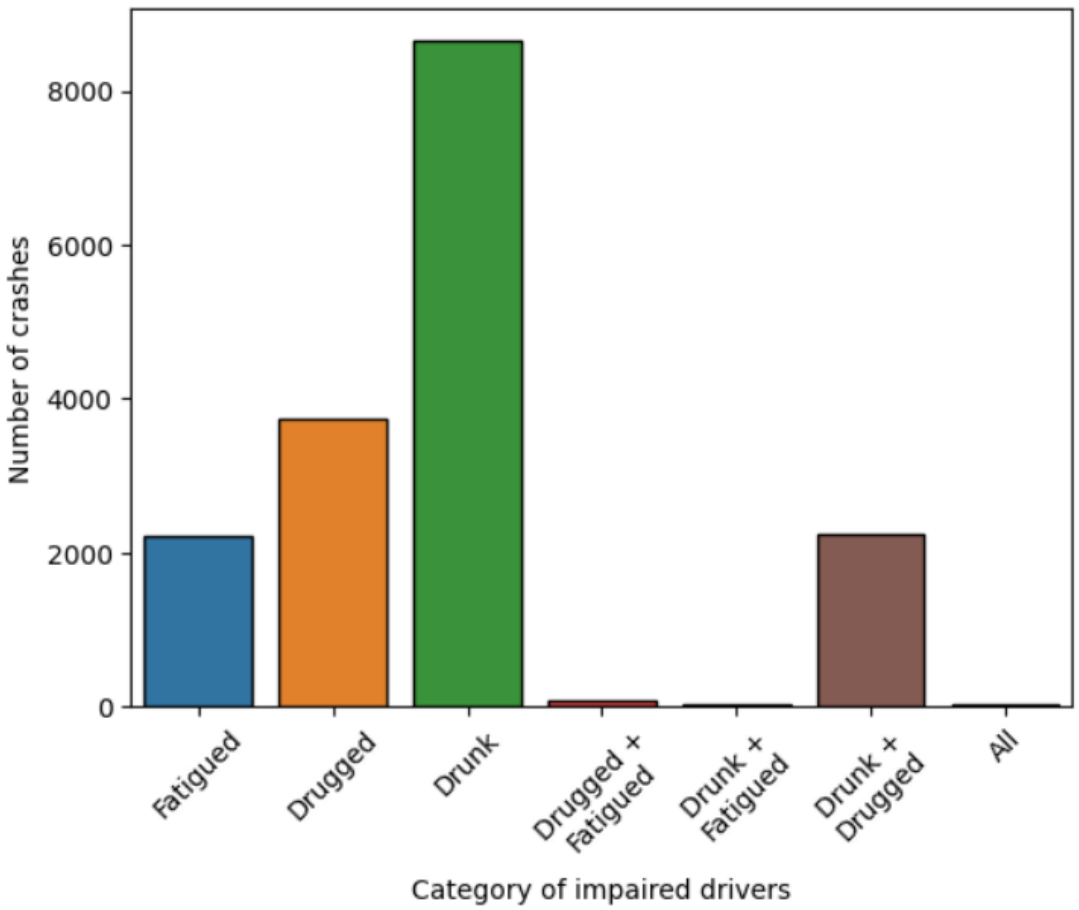

Figure 2: Number of crashes caused by impaired drivers according to their categories

It seems that among the impairments, being drunk is the most common and it is followed by being drugged then fatigued is tied with being drunk and drugged. However, we can see that the sample of combination of being drugged and fatigued, or drunk and fatigued or every category is very small.

Now our team wants to inquire about the severity of the vehicle damage done if the person is impaired. In the dataset given, we are given details on the severity of the damage done to the vehicle after the crash in order of their severity:

- None (Scratches)
- Minor (Bumps, still drivable)

- Moderate (May be undrivable)
- Severe (Car is disabled)

However, to see if an impaired driver is more likely to cause more severe damage to the car, we randomly undersampled the data from the unimpaired drivers by 10 : 1 to match with the number of impaired drivers. This is done to make a more meaningful visualisation see the relative difference.

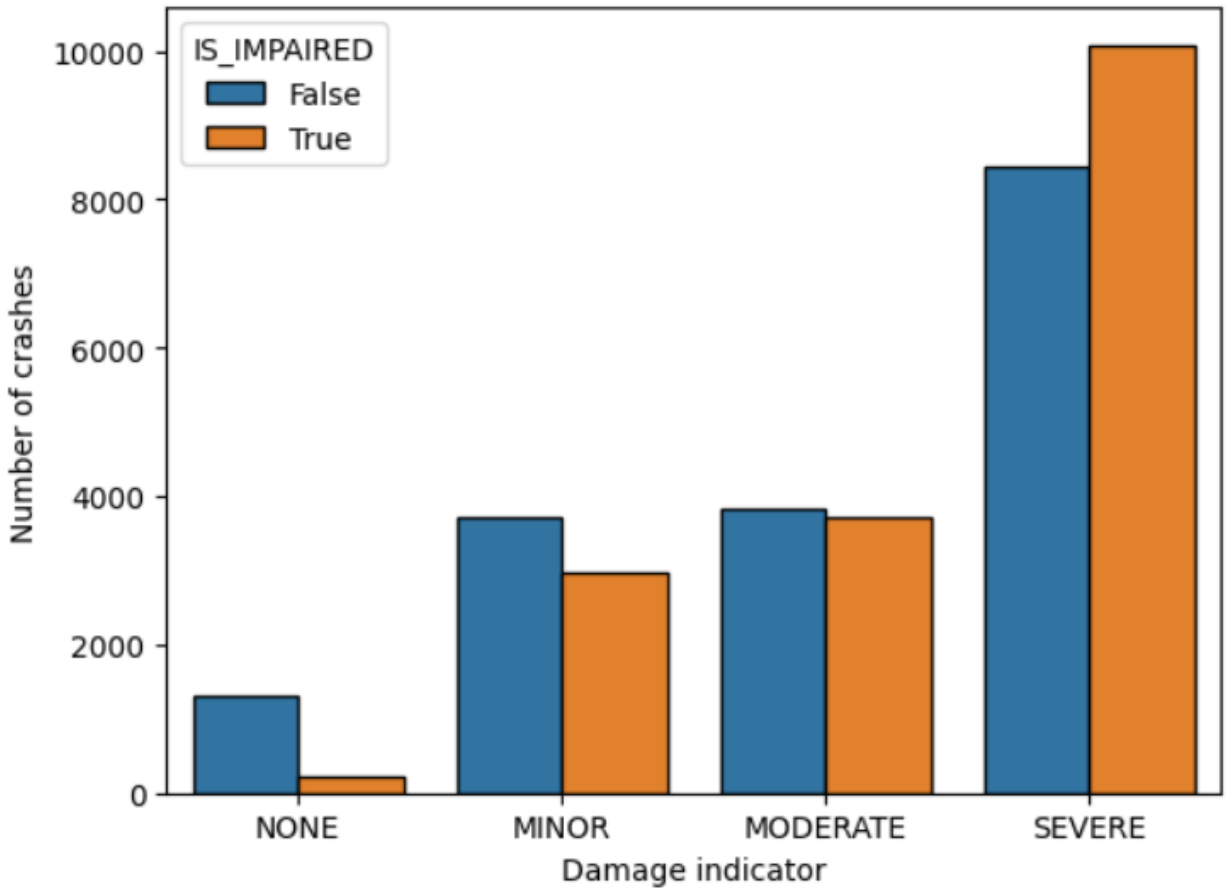

Figure 3: Impaired and unimpaired drivers number of crashes according to their severity after random undersampling

Here in figure 3 we can see that after undersampling, we are able to see the relative difference more clearly. It is clearly seen that if a driver is impaired it is more likely that the vehicle will at least suffer minor damages and that they are very minimal. Additionally there are also more severe crashes from impaired drivers. However, it seems that regardless of the impairments of the drivers, the most severe damage is the highest among the two categories with more crashes from the impaired drivers.

Statistical test

From this visualisation, it seems quite apparent that impaired drivers are more likely to cause more severe damage, the reason could be that they have less control of their vehicles due to their impairment.

Next we create the contingency table so that it can be used in the chi-square test for homogeneity.

	None	Minor	Moderate	Severe	Total
Impaired	259	3627	4543	11772	20201
Unimpaired	1448	4317	4880	10417	21062
Total	1707	7944	9423	22189	41263

Figure 4: Contingency table of undersampled unimpaired crashes and impaired crashes according to their vehicle damage

We construct the hypothesis test and the test statistic.

$$H_0 : The\ distribution\ of\ vehicle\ damage\ between\ impairments\ is\ the\ same$$

$$H_1 :\ H_0\ is\ not\ true$$

We denote the categories as such:

$$i\ =\ \{1 : Impaired,\ 2 : Unimpaired\}$$

$$j\ =\ \{1 : None,\ 2 : Minor,\ 3 : Moderate,\ 4 : Severe\}$$

$$O_{i,j} =\ Observed\ value\ samples\ in\ the\ ith\ and\ jth\ category$$

$$E_{i,j} =\ Expected\ value\ samples\ in\ the\ ith\ and\ jth\ category$$

$$p_{i,j} =\ Probability\ of\ a\ sample\ falling\ in\ ith\ and\ jth\ category$$

We find the total number

$$Total,\ n\ = \sum_i \sum_j O_{i,j}$$

So we estimate the probabilities $\hat{p}$ as such:

$$Under\ H_0\ p_{i,k} = p_{i,j}\ \forall j,k$$

$$estimated\ probabilities,\ \hat{p}_{i,*} = \frac{\sum_j O_{i,j}}{n}\ and\ \ \hat{p}_{*,j} = \frac{\sum_i O_{i,j}}{n}$$

$$E_{i,j} = n \cdot \hat{p}_{i,*} \cdot \hat{p}_{*,j}$$

$$\frac{(E_{i,j} - O_{i,j})^2}{E_{i,j}} \sim \chi^2(1)$$

Since we estimate the probabilities, and for each row we have 4 categories so we have 3 independent estimators and for each column we have 2 categories and so we only have 1 estimator and in total we have 3 by 1 independent estimators and so our test statistics degree of freedom will be 3.

$$Test\ statistic,\ T = \sum_i \sum_j \frac{(E_{i,j} - O_{i,j})^2}{E_{i,j}} \sim \chi^2(3)$$

Next we recreate the contingency table with the expected observations wrapped in parentheses.

	None	Minor	Moderate	Severe	Total
Impaired	259 (835.7)	3627 (3889)	4543 (4613)	11772 (11326)	20201
Unimpaired	1448 (873.1)	4317 (4055)	4880 (4810)	10417 (10863)	21062
Total	1707	7944	9423	22189	41263

Figure 5: Contingency table of undersampled unimpaired crashes and impaired crashes according to their vehicle damage with the expected observations

$$realised\ test\ statistic \approx 965.3$$

$$p - value = P(T > t) = P(\chi^2(3) > 965.3) \approx 0$$

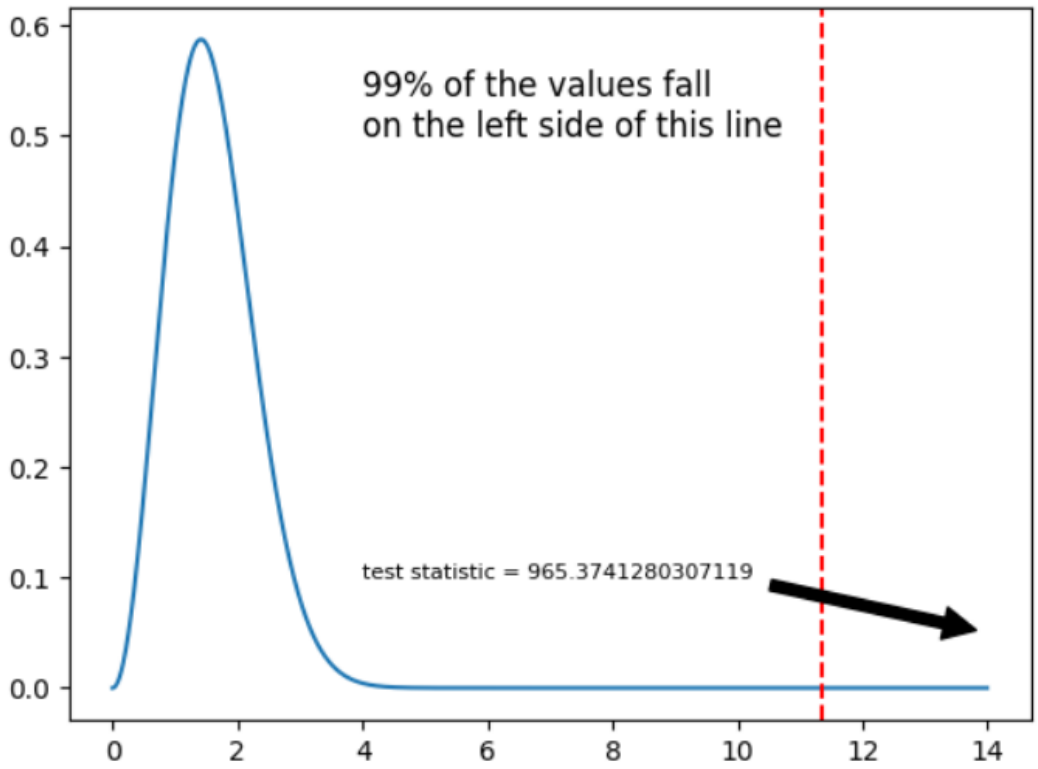

Figure 6: plot of $\chi^2(3)$ distribution with labels

As our test statistic falls in the critical region where $\alpha = 0.01$ we can reject H_0 that the distribution of vehicle damage across impairment is the same at 0.01 significance level.

Hence, by the chi-square homogeneity test, we can conclude that there is a statistically significant difference in the two distributions. Additionally from the bar plot, we can see that the crash from an impaired driver is more likely to cause more vehicular damage.

Conclusion and recommendation

In Philadelphia, Pennsylvania it is already illegal to drive under the influence of alcohol or drugs and the penalty includes fines, re-education or even jail. Additionally, offenders may be required to install an ignition interlock system where the vehicle is shut off if the driver has a high-alcohol in his blood level. However, alcohol-related traffic fatalities still make up 30% of all vehicular fatalities (Pennsylvania Department of Transportation, 2023). So we can see that despite 10 : 1 unimpaired to impaired crash cases, alcohol-related traffic fatalities make up a disproportionate amount of fatalities. Furthermore, in figure 2, drinking related accidents make up most of the impaired crash cases.

Perhaps, more vehicular police presence can be placed near bars to dissuade drunks from driving under influence. Additionally, they can also be placed near high-accident areas. Police presence has shown to be able to deter traffic violations (National Institute of Justice, 2020).

Technical Expositions

Wrangling & Cleaning:

We used three datasets:

- **crash_info_general.csv**
- **crash_info_flag_variables.csv**
- **crash_info_vehicles.csv**

We want to investigate if there is a correlation between the impairment of the driver to the severity of the crash. This would be done by measuring the **DAMAGE_INDICATOR** from **crash_info_vehicles.csv**.

We have 101 variables and we will remove duplicate observations, but we want to select data from crash_info_general.csv that are relevant to the question. Such as : **CRN**, **SECONDARY_CRASH**. **CRN** is the primary key for all the three datasets and we will also remove **TRUE** values for **SECONDARY_CRASH** from **crash_info_general.csv** as including them may be considered double counting.

Then we noticed that **SECONDARY_CRASH** was only added more recently, so earlier data had many missing values and we filled it with **FALSE.** Then we also removed **DEER_RELATED** accidents as they are irrelevant.

Finally, some crashes had unknown or missing **DAMAGE_INDICATOR,** we also removed these as well as they were not useful for our question. Then we labelled the impaired driver crashes if they are either fatigued, drunk or drugged, those that do not fall under these categories we labelled them as unimpaired.

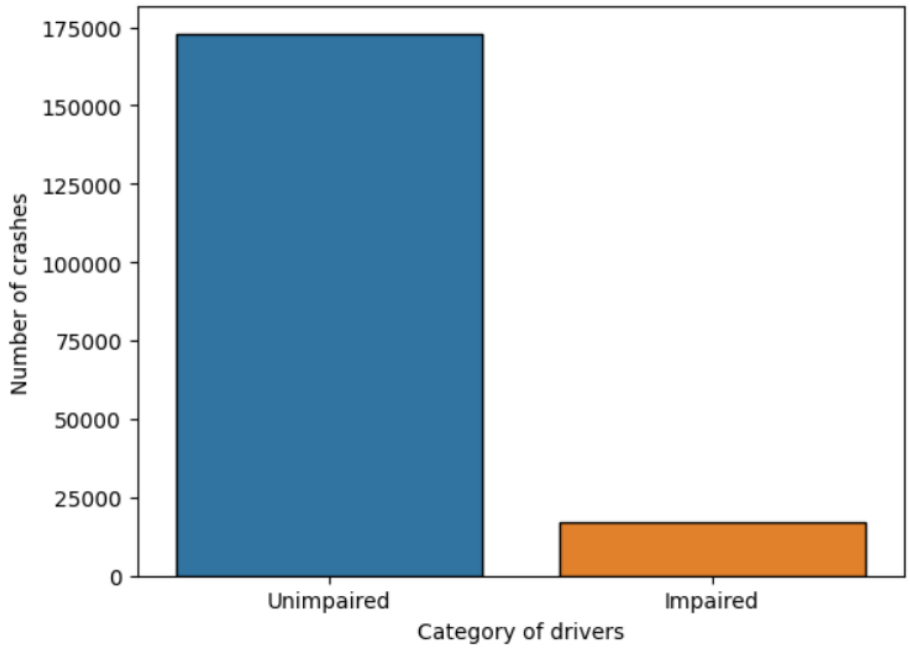

Figure 1: Number of crashes caused by unimpaired and impaired drivers

After labelling, we found that the unimpaired driver crashes outnumber the impaired 10 to 1. So we did a random sample to create a more equal number of crashes between the two categories so that they have equal weights.

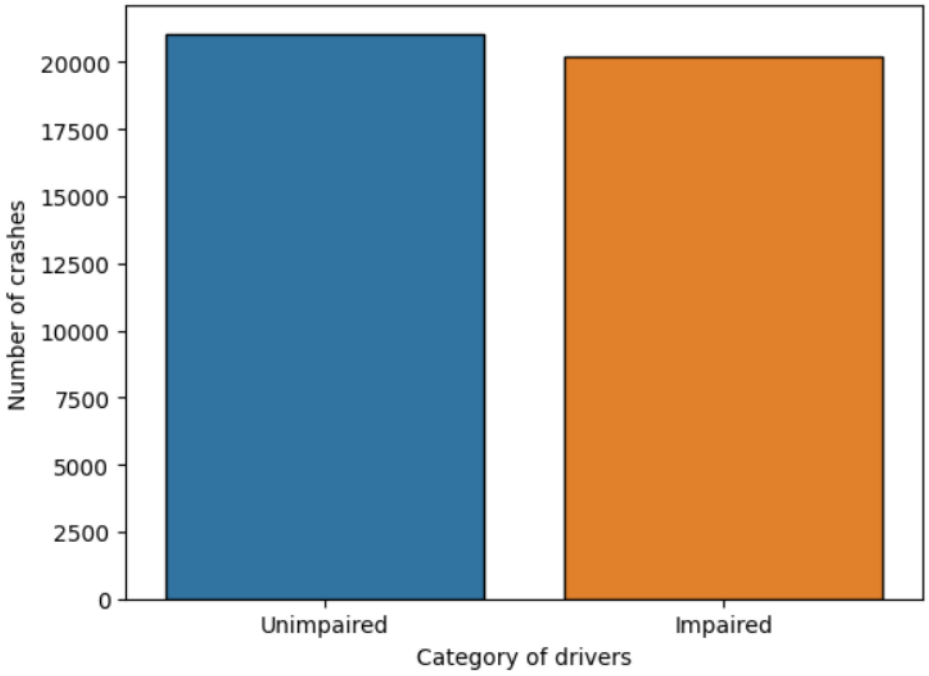

Figure 7: Number of crashes caused by unimpaired and impaired drivers after undersampling

With these, we are able to create the contingency table and we are able to create a hypothesis test to answer our question using the chi-square homogeneity test.

	None	Minor	Moderate	Severe	Total
Impaired	259 (835.7)	3627 (3889)	4543 (4613)	11772 (11326)	20201
Unimpaired	1448 (873.1)	4317 (4055)	4880 (4810)	10417 (10863)	21062
Total	1707	7944	9423	22189	41263

Figure 5: Contingency table of undersampled unimpaired crashes and impaired crashes according to their vehicle damage with the expected observations

References

1. National Institute of Justice. “Effect of High-Visibility Enforcement on Motor Vehicle Crashes.” *National Institute of Justice*, June 2020, nij.ojp.gov/topics/articles/effect-high-visibility-enforcement-motor-vehicle-crashes.

2. Pennsylvania Department of Transportation. “Impaired Driving.” *Pennsylvania Department of Transportation*, 2023, www.penndot.pa.gov/travelinpa/safety/trafficsafetyanddrivertopics/pages/impaired-driving.aspx.