DATA1002

Project Stage 2 Report

Crash Data Analysis

This report analyses multiple factors such as drugs, alcohol and speed limit to determine a possible relationship with crash frequency and severity

Contents

Part 1 – Overview of Data and Analysis

Domain Knowledge

Vehicle accidents are one of the most common accidents in the world, causing thousands of deaths and injuries annually. The state of the driver and environmental factors such as weather, time of day and lighting play a crucial role in determining the frequency and severity of these crashes. Understanding the trends between vehicle accidents and these issues is vital for the enactment of suitable policy and rules that can minimize these high-risk factors. That is why there are many public events supporting walking and public transport to minimize accidents and raise awareness of the problematic factors that cause these accidents. This is also supported by the government, as evident in the announcement of franchising to fund the improvement of public transport, as well as the collation and publication of road crash data for public viewing.

Consequently, this report will analyse road crash data and seek to identify and explain trends and factors that contribute to vehicle accidents.

Data Origins

The data that will be analyzed in this report was created by combining two vehicle accident datasets; one from Australia and the other from the US. The Australia one details crashes in South Australia and was accessed from the South Australian Government data Directory. The US dataset was accessed from the US Government's open data website and details crash statistics from Pennsylvania and the City of Pittsburgh.

These two datasets were combined into a single csv file using Python programming. In the combination process, any columns that were not common between the two datasets were disregarded. The remaining columns of our combined dataset retained the same meaning as the original columns, with only minor changes made to the values to make information consistent between the two datasets.

The original source of the Australian dataset can be found here:
https://data.sa.gov.au/data/dataset/road-crash-data

The original source for the US dataset can be found here:

https://catalog.data.gov/dataset/allegheny-county-crash-data

Analysis

Relationship between Crashes, Time, Drugs, Alcohol, Nation and Crash Severity

In this analysis, we were interested in the relationships between what time usually drug and alcohol related crashes occur. Therefore, we used the column **time, drug related, alcohol related** and for comparison we also included the column **Nation.** This analysis was conducted with Excel and Python.

The analysis' aim is to understand what time of the day there is higher number of accident due to alcohol, drug or both. There are also some charts about vehicle accidents not related to drug or alcohol for curious readers. The work can be seen in the *analysis.xlsx* file which contains 6 sheets: one with the data, 4 pivot table (each featuring certain attributes), and a sheet for visual analysis (slicer).

On the 4 sheets (**Drug & Alcohol, Just Drug, Just Alcohol, Neither Drug nor Alcohol**), you could see some numerical summaries and charts. However, for better comparison, it is recommended to look at the sheet **slicer**.

Here, you could also "play" with the charts and use the panels at the right. Using that, you are able to observe the number of crashes based on nation, month and day. We have displayed these graphs in figures 1-4 below. Generally, it is clear that most of the only alcohol related crashes happen in the evening or at night as supported through 4pm-11pm having the highest frequency of crashes for figures 1-3. Interestingly, drug related accidents (figure 3) occur during daytime between 9 am and 6 pm.

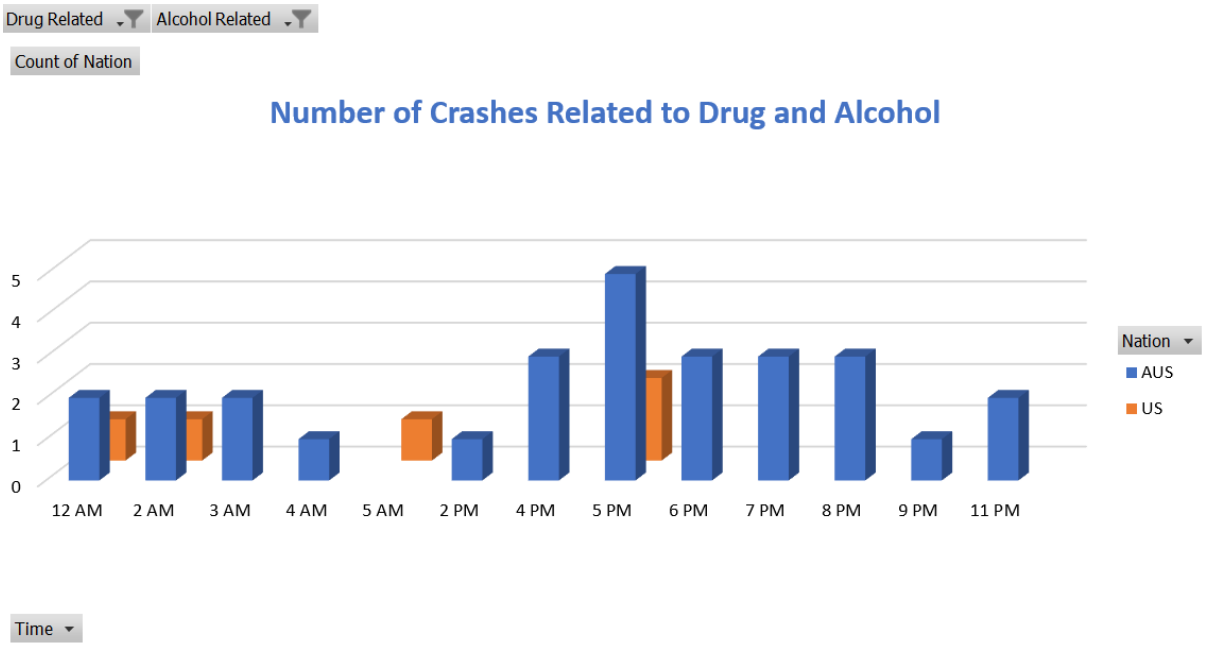


Figure 1. Number of Crashes Related to Drug and Alcohol

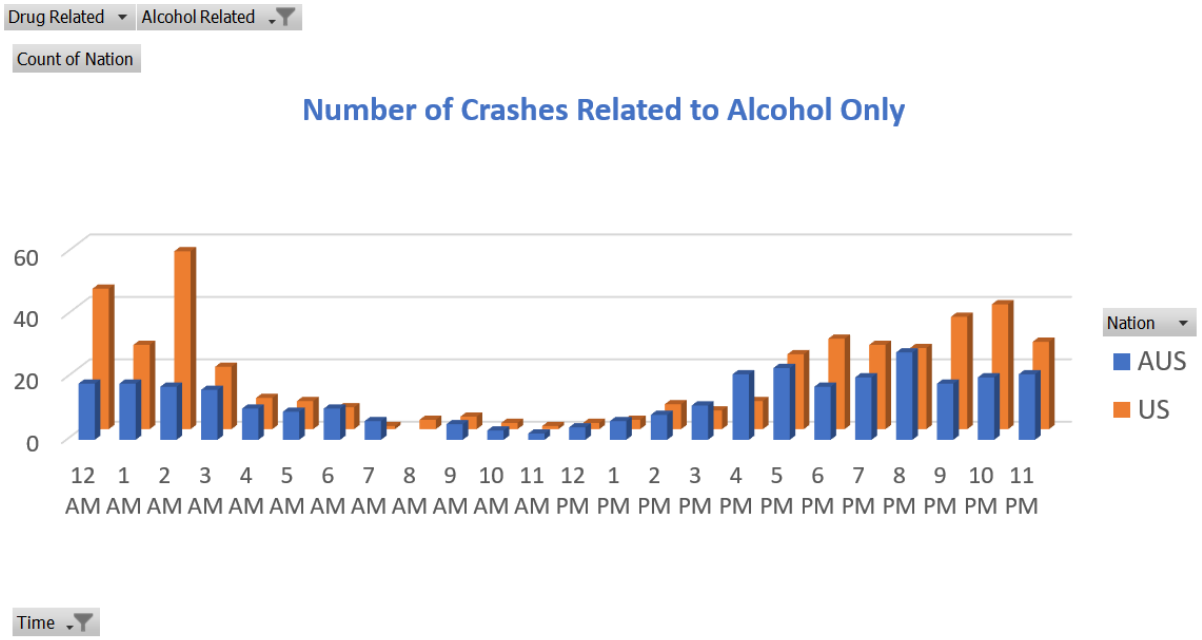


Figure 2. Number of Crahses Related to Alcohol Only

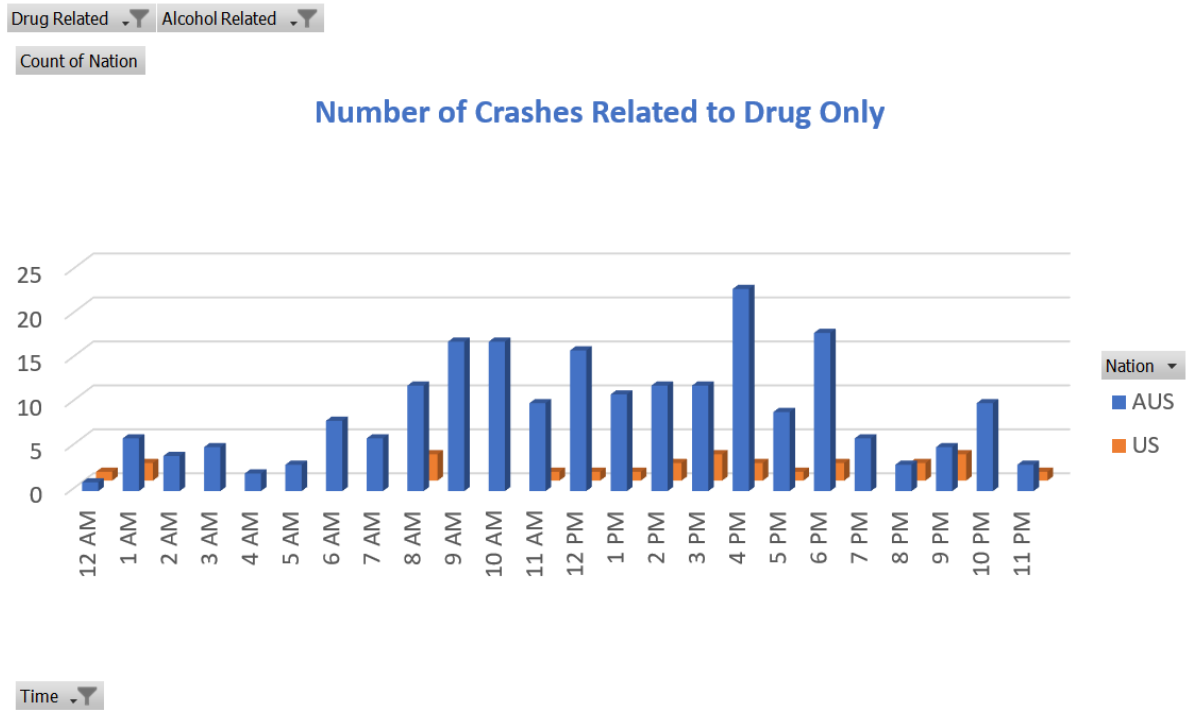


Figure 3. Number of Crashes Related to Drugs Only

It is also clear that most of the vehicle accidents are not due to alcohol or drug consumption as you can see in figure 4, which has a larger frequency of crashes compared to that of figures 1-3. These crashes also usually occur daylight with the highest numbers at 8 am and in the afternoon. The

assumption could be that those hours are the time of the day when traffic jam and high number of cars are present due to work time/end of work.

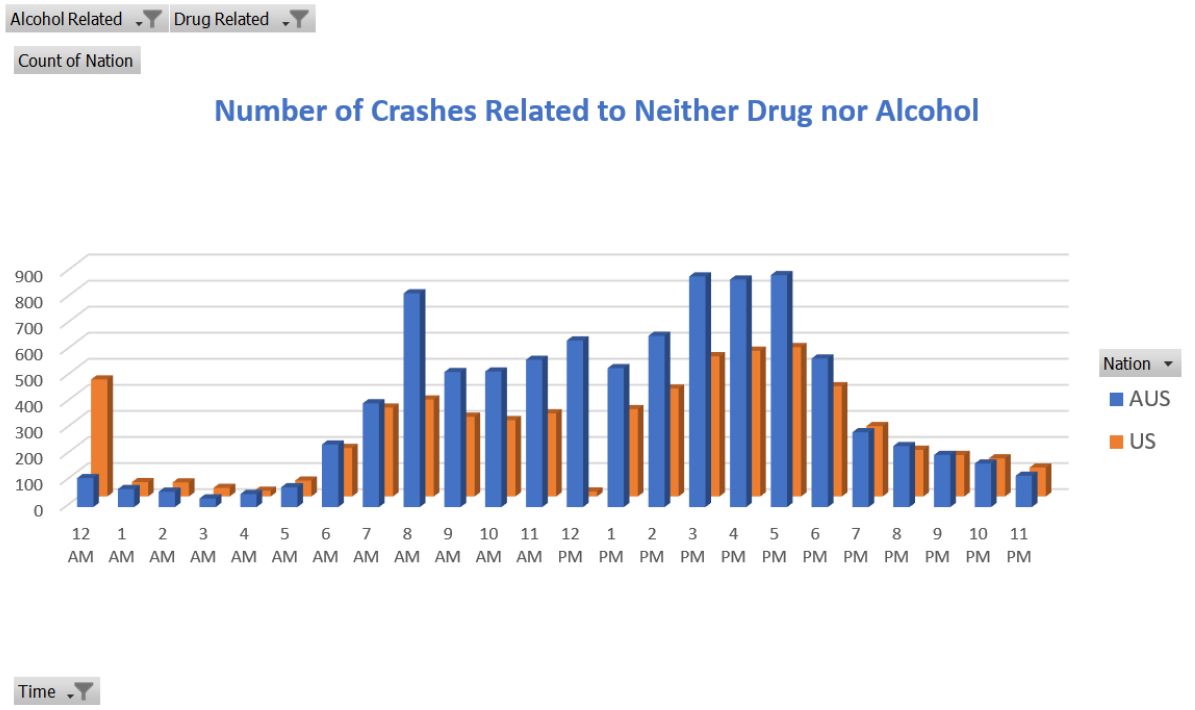


Figure 4. Number of Crashes Related to Neither Drugs nor Alcohol

It is also important to mention that the US and AUS bar chart are not for comparing the two countries since our dataset have way more AUS items which could lead to bias conclusion. The reason for displaying the two is to see whether they have different trends, however the limited US data shows that they have similar patterns in figure 1-4.

For further observation, use the panels and see how the charts changes.

To extend further on these graphs, we created numerical summaries which finds the average crashes for each crash severity depending on the presence of alcohol and drugs. To do this, we created a program in python called ‘summaries.py’ which used the ‘statistics’ package to find the mean and standard deviation.

For fatal injuries, the numerical summaries created were:

```
Statistics for Fatal injuries:
The mean number of Fatal injuries for crashes involving alcohol and drugs is: 0.15151515151515152
The standard deviation of Fatalities for crashes involving alcohol and drugs is: 0.36410954062720957

The mean number of Fatal injuries for crashes involving alcohol is: 0.032670454545454544
The standard deviation of Fatal injuries for crashes involving alcohol and drugs is: 0.2004567314974432

The mean number of Fatal injuries for crashes involving drugs is: 0.036885245901639344
The standard deviation of Fatal accidents for crashes involving alcohol and drugs is: 0.1888674583364159

The mean number of Fatal injuries for crashes not involving alcohol and drugs is: 0.005208002540489044
The standard deviation of Fatal injuries for crashes involving alcohol and drugs is: 0.07791300252446984
```

For serious injuries, the numerical summaries created were:

```
Statistics for Serious injuries:
The mean number of Serious injuries for crashes involving alcohol and drugs is: 0.18181818181818182
The standard deviation of Seriousities for crashes involving alcohol and drugs is: 0.3916747259003201

The mean number of Serious injuries for crashes involving alcohol is: 0.07954545454545454
The standard deviation of Serious injuries for crashes involving alcohol and drugs is: 0.30534697877546413

The mean number of Serious injuries for crashes involving drugs is: 0.21721311475409835
The standard deviation of Serious accidents for crashes involving alcohol and drugs is: 0.5499259393385045

The mean number of Serious injuries for crashes not involving alcohol and drugs is: 0.03416957764369641
The standard deviation of Serious injuries for crashes involving alcohol and drugs is: 0.20715368777178433
```

For minor injuries, the numerical summaries created were:

```
Statistics for Minor injuries:
The mean number of Minor injuries for crashes involving alcohol and drugs is: 0.7272727272727273
The standard deviation of Minorities for crashes involving alcohol and drugs is: 0.801277389263827

The mean number of Minor injuries for crashes involving alcohol is: 0.31676136363636365
The standard deviation of Minor injuries for crashes involving alcohol and drugs is: 0.6085710893989229

The mean number of Minor injuries for crashes involving drugs is: 0.9221311475409836
The standard deviation of Minor accidents for crashes involving alcohol and drugs is: 1.1027845261539553

The mean number of Minor injuries for crashes not involving alcohol and drugs is: 0.3500158780565259
The standard deviation of Minor injuries for crashes involving alcohol and drugs is: 0.6622085998208821
```

From these summaries we notice that across all crash severities, the presence of alcohol and drugs has a significantly larger mean compared to just having alcohol, drugs or neither, which is as expected. We also see that for alcohol and drugs, the standard deviation is relatively low compared to the mean which indicates a small relative spread. This increases the reliability of our conclusion.

Interestingly, across all crash severities we also see that the mean for crashes involving just drugs is consistently higher than the mean for crashes involving just alcohol. This may be due to drugs have a larger mental effect compared to alcohol, which limits the ability of the driver to make good decisions.

Another interesting thing to note is that for minor injuries, the mean for crashes not involving alcohol and drugs (0.35) is slightly larger compared to the mean for crashes involving just

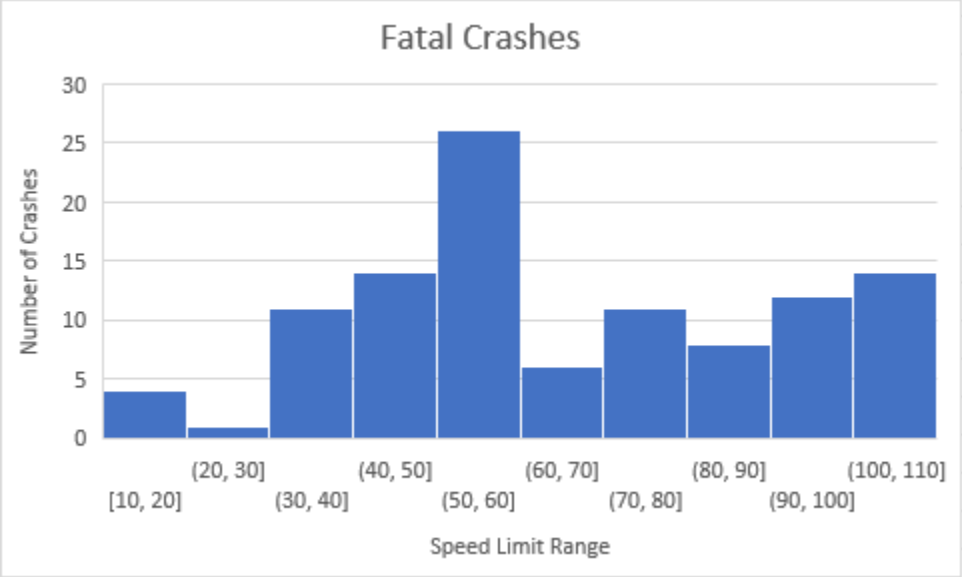
alcohol (0.32). However, the standard deviation for these values is relatively larger compared to the mean which limits the reliability of this conclusion. This trend is also not reflected across the other crash severities.

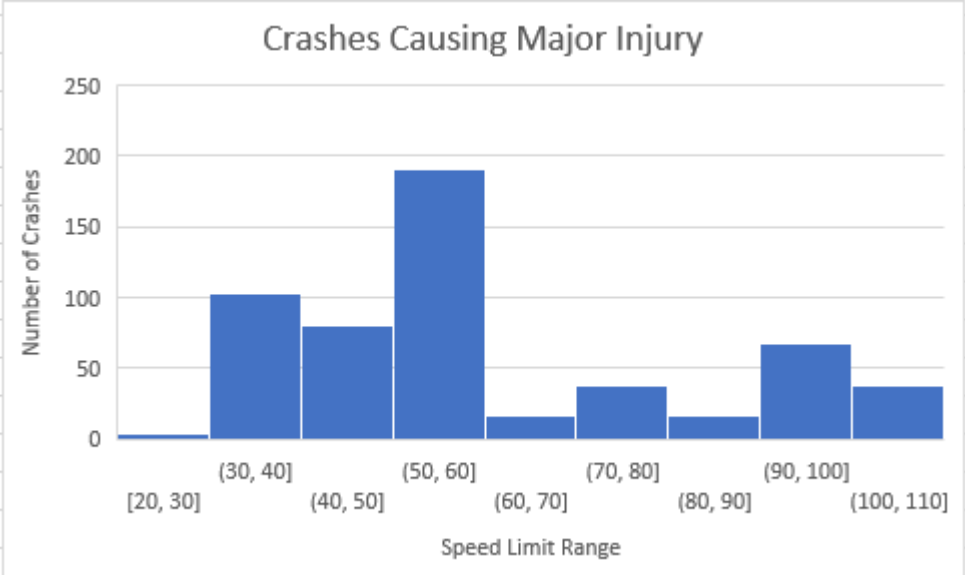
Relationship between Speed Limit and Crash Severity

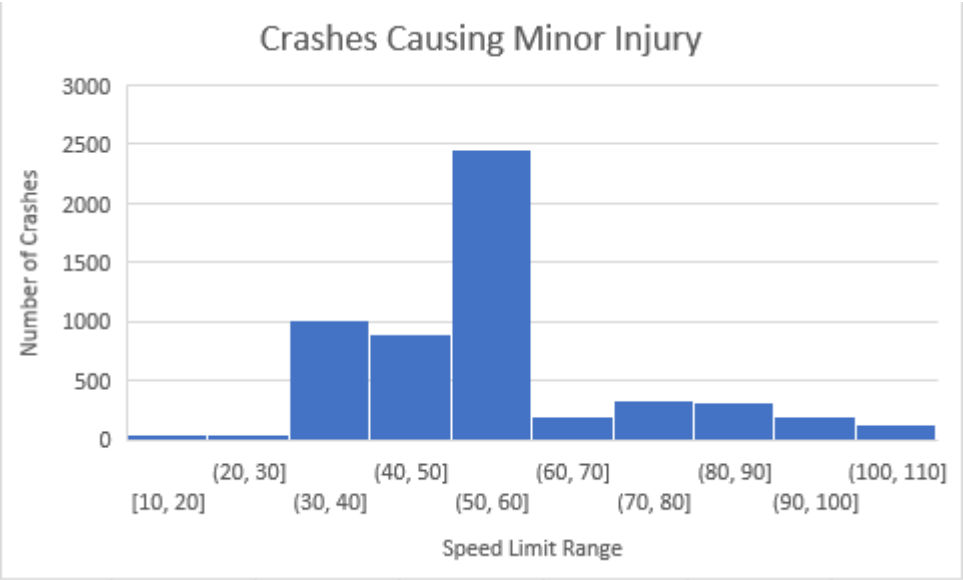
Speed is one of the vital variables influencing crashing consequences. In this section, the relationship between **speed limit** and **crash severity** will be investigated. Python is used for data filtering and MS Excel is used for creating the charts.

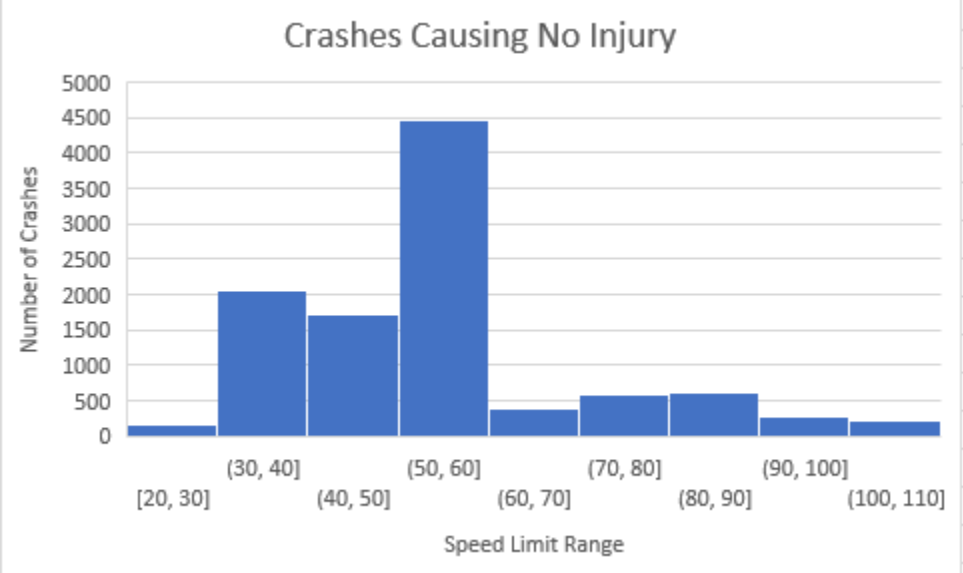
To determine whether the severer crashes are with higher speed limits, 4 bar charts are made from filtered data.

In the charts, most fatal crashes and crashes causing major injury have speed limits in a higher range. From calculation, the average speed limit for the 4 types of crashes are 70.4, 65.5, 59.9 and 59.3. Clearly, this higher speed limits are associated with higher crash severities. However, the fatal crashes and the ones caused major injury are significantly fewer than other types of crashes. This conclusion will be more reliable if gaps between crash numbers are smaller.



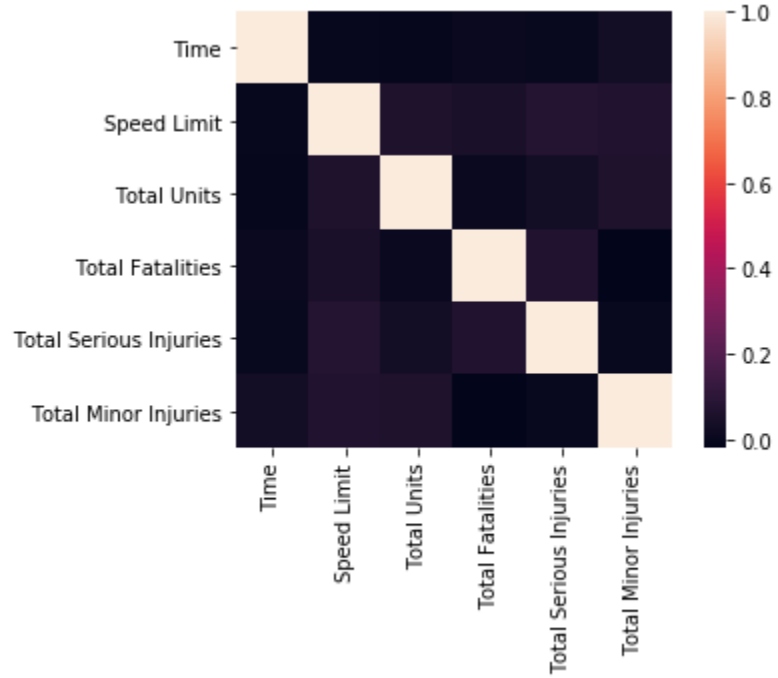


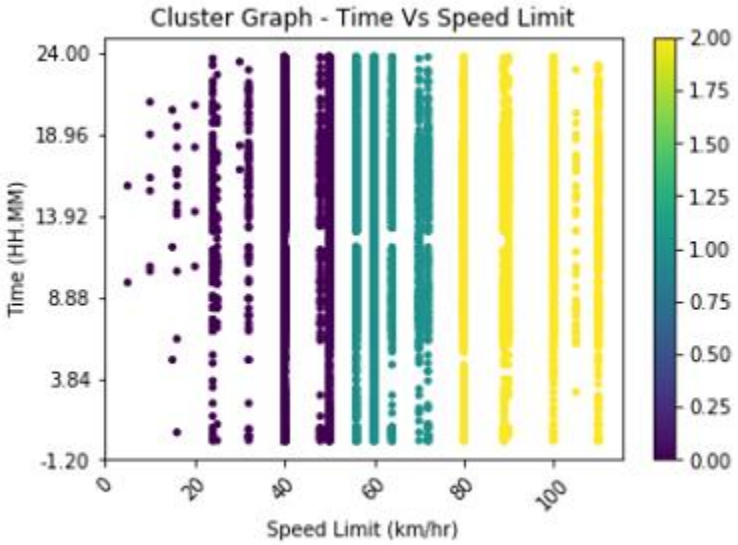

Crashes Causing Minor Injury
Number of Crashes
3000
2500
2000
1500
1000
500
0
[10, 20]
(20, 30]
(30, 40]
(40, 50]
(50, 60]
(60, 70]
(70, 80]
(80, 90]
(90, 100]
(100, 110]
Speed Limit Range


Crashes Causing No Injury
Number of Crashes
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
[20, 30]
(30, 40]
(40, 50]
(50, 60]
(60, 70]
(70, 80]
(80, 90]
(90, 100]
(100, 110]
Speed Limit Range

Relationship between Speed Limit, Time and Crash Severity

Extending on this, we wanted to investigate the relationship between time and speed limit. As a result, we used the 'sci-kit learn' package in Python to create a scatter plot of speed limit vs time to identify any linear relationships or correlations between the two variables. A heatmap and clustered scatter plot can be seen below.





By looking at this plot, we observed that there is no obvious trend or correlation. The data points are spread out and randomly dispersed across all values.

After deducing that there was no correlation, we wanted to cluster the data points to determine any distinct categories. Namely, we created 3 centroids for the data, hoping that these three centroids would correlate to the three crash severities and provide us further insight into such crashes and the scenarios in which they occur (fatal, serious, minor).

However, after creating the cluster graph seen above, we observed that three distinct clusters were not created. These clusters do not seem to have any relation to the three crash severities and so we concluded that speed limit and time did not have a multivariate relationship with crash severities.

This is further reflected in the heat map, where we see that all the squares indicating the relationship between time, speed limit and the different crash severities are very dark, indicating a very weak correlation.

Relationship between Month and Crash severity

Our aim was to determine if there was a relationship between certain months and the magnitude of the different crash severities (fatal, serious and minor injuries).

For this, we created a Python program called 'summaries.py' that used the 'statistics' package to create numerical summaries. These summaries included the mean and standard deviation for each crash, the total crashes for that month as well as determining the top 3 months with the highest means.

For fatal crashes, the 3 months with the highest average fatalities were April, July and August. Their mean and standard deviation can be seen below.

```
Mean Fatal crashes for  April: 0.0104083266661329063
Standard deviation of Fatal for  April: 0.11624704212325589
Total amount of Fatal crashes: 13
```

```
Mean Fatal crashes for  July: 0.009900990099009901
Standard deviation of Fatal for  July: 0.09904762623474865
Total amount of Fatal crashes: 13
```

```
Mean Fatal crashes for  August: 0.010431154381084 84
Standard deviation of Fatal for  August: 0.11451246899886898
Total amount of Fatal crashes: 15
```

For major injuries, the 3 months with the highest average major injuries were March, April and July. Their mean and standard deviation can be seen below.

```
Mean Serious crashes for  March: 0.047843665768194 07
Standard deviation of Serious for  March: 0.24853325596237028
Total amount of Serious crashes: 71
```

```
Mean Serious crashes for  April: 0.0488390712570056
Standard deviation of Serious for  April: 0.25946940536095486
Total amount of Serious crashes: 61
```

```
Mean Serious crashes for  July: 0.05483625285605483
Standard deviation of Serious for  July: 0.2705738450699702
Total amount of Serious crashes: 72
```

For Minor injuries, the 3 months with the highest average minor injuries were March, November and December. Their mean and standard deviation can be seen below.

```
Mean Minor crashes for  March: 0.3894878706199461
Standard deviation of Minor for  March: 0.7029827650024858
Total amount of Minor crashes: 578
```

```
Mean Minor crashes for  November: 0.3783592644978784
Standard deviation of Minor for  November: 0.7131252516586228
Total amount of Minor crashes: 535
```

```
Mean Minor crashes for  December: 0.366445916114790 27
Standard deviation of Minor for  December: 0.6808499484205293
Total amount of Minor crashes: 498
```

Originally, we hypothesized that the average for each crash severity would be highest around the holiday months. For minor crashes, we that the results reflect this hypothesis, where the months with the highest crashes occur around the latter half of the year, except for March. The reason why minor crashes support this trend is likely due to the congestion around holiday times and thus the slow speeds one would travel. It is these slow speeds that would be responsible for the high amount of minor crashes. However, we see that this trend is not reflected for major and fatal crashes.

For major crashes and fatal crashes, the months with the highest averages occur during the early to middle parts of the year. An explanation is that early-middle of the year holds the peak activity for many people as the activities for work or school begin to heighten. However, conclusions for fatal crashes can be deemed unreliable due to the low amount of data collected (4, 15 and 14 for March, August and October respectively). We also note that for major and fatal crashes, the standard deviations are relatively high compared to the average which indicates high spread in the data. Consequently, there may be great variation for our analysis on major and fatal crashes.

Conclusion

Combining this with the conclusions from our Excel 'slicer' graphs, we can infer that during 9am – 6pm, crashes will cause the most injury. 9am-6pm is the time period in which most drug related accidents occur, and as drug related accidents result in the highest average injuries especially for minor injuries (which has an average of 0.92), these would be some of the most dangerous hours. During this time, we would expect these crashes to be in the speed limit range of 50-60, as this is the mode for which most minor crashes occur. As March, November and December are the months where minor crashes occur the most, we would expect 9am-6pm to be even more dangerous.

To remedy this, the government could enforce stricter policy between March, November and December between 9am-6pm. This could involve deploying more police to catch speeding as well as more frequent drug checks due to its positive correlation with more serious accidents.

Part 2 – Techniques of Analysis

Relationship between Crashes, Time, Drugs, Alcohol, Nation and Crash Severity

First of all, we created a python file (named *new_file.py*) to create a new file from the *final.csv* and named it *analysis.csv* (the python file and *final.csv* file must be in the same directory for successful running). The python file rounded the values of column **time** (rounded the minutes, so if a crash occurred at 12:13 the program rounded it to 12.10), and instead of using time period (am and pm) we converted them to numbers only (e.g. from 1:00 pm to 13:00). This process mainly used split(), endswith(), startswith() function and conditional statements within a for loop. We did this to simplify the analysis later on.

Since a csv file cannot hold multiple sheets due to possible data loss, we converted the *analysis.csv* file to xlsx file. Opening this file (*analysis.xlsx*) we inserted four pivot tables for the four aspects we were curious about (crashes related to only drug, crashes related to only alcohol, crashes related to both alcohol and drug and crashes related to neither of them).

From those four tables, pie charts and bar charts could be created and designed. From that point, we implemented a new technique: slicer. We copied the 4 different bar charts and parsed them into a new sheet (slicer). We inserted 2 slicers, the panels at the right of the charts. However, inserting them is not enough, since in standard mode one slicer could only affect one bar chart. Therefore, we have to connect the slicer to the other charts we wanted to use on by right clicking on the created slicer and clicking on the “Report Connection”.

To create the numerical summaries, we made a Python program called ‘summaries.py’. Within this program were two functions, with the function being used for this particular analysis called alcoholdrugCrash(). This function takes in a string as an input which corresponds to the crash severity the user wants summaries for.

The function then reads through the data file and extracts relevant information on the crash severity. Within this function, a dictionary is created that is used to categorise these crash statistics as involving drugs, alcohol etc. The values for these dictionary keys are lists which have values appended to them. As the function reads through the values, the function will append the appropriate crash statistic to the dictionary key to which it belongs using if statements. For example, if the user wanted summaries for fatal crashes, the string input would be ‘Fatal’.

```
if drugs == "Y" and alcohol == "Y":


    if crashType == "Fatal":
        crash_dict["DA"].append(fatal)
    elif crashType == "Serious":
        crash_dict["DA"].append(serious)
    elif crashType == "Minor":
        crash_dict["DA"].append(minor)

elif drugs == "N" and alcohol == "Y":
    if crashType == "Fatal":
        crash_dict["A"].append(fatal)
    elif crashType == "Serious":
        crash_dict["A"].append(serious)
    elif crashType == "Minor":
        crash_dict["A"].append(minor)

elif drugs == "Y" and alcohol == "N":
    if crashType == "Fatal":
        crash_dict["D"].append(fatal)
    elif crashType == "Serious":
        crash_dict["D"].append(serious)
    elif crashType == "Minor":
        crash_dict["D"].append(minor)
```

The code block above illustrates this idea. Here, we used dictionary keys “DA” to stand for Drugs and Alcohol, “A” for alcohol, “D” for drugs etc.

Once all the information had been appended to the appropriate dictionary key, we iterated through each key and used the statistics package of Python to find the mean and standard deviation. The code used to achieve this can be seen below.

```
for i, k in crash_dict.items():

    if len(k) == 0:

        continue

    mean = statistics.mean(k)
    sd = statistics.stdev(k)
```

Why we chose Excel and the 'Statistics' package in this analysis

Throughout this analysis, we realized that we are unable to work on the csv file since it cannot save multiple work sheets, and as soon as we saved the work, it actually disappeared. Yet, we insisted to work on excel because of the practicality of the slicer. With only one click, we are able to see a completely different picture of the dataset. Also, in our previous project, we only used python, hence we would like to work with a different tool for a short time.

Strength of analysis:

Using Excel provides an interactive way to do analysis and compare different datasets with one another. With some clicks we can create different graphical summaries and it does not require a high level of skill in technology. Unlike in python, both the dataset and the analysis could be in one Excel file which simplifies the work in terms of storing analysis. Additionally, Excel provides descriptions and panels for their accessories. If we want to use functions, the program gives us a short help to understand how to use it or even ordering numbers does not require any coding skill. The numerical summaries we did create with Python provide us a quantitative understanding of the differences between alcohol, drugs etc. as well as allowing us to directly determine the spread.

Limitations of analysis:

Unfortunately, it is hard to do complex analysis with Excel. With Excel, it is hard to create multiple combination of different charts, while in python we are able to merge different graphical analysis. When it comes to numerical analysis, one needs to do certain number of steps to acquire certain results. Therefore, it is suggested to use python NumPy, Sci-kit or statistics libraries to calculate numerical analysis such as spread, standard deviation, p-value etc. Also, when graphing in python, we could detect any incorrect item in the dataset more easily and

correct them. In Excel, by having a large dataset, we may not able to detect all the incorrect items which could alter the analysis.

Relationship between Speed Limit and Crash Severity

In this analysis, speed limit values were written to 4 independent files using ‘separate.py’. The Python ‘csv’ library is used to read the cleaned dataset and write to new csv files. The ‘reader’ function is used to read data values into lists, then write the stored values to their assigned values.

In terms of representing the data, bar charts are used in this scenario. The initial idea was to use line chart to show all the data values. However, this method is only applicable to the first two sub datasets. For enormous number of values, a line chart is not able to show any useful information. The bar charts are made directly by MS Excel. Each chart is made from all the values in the corresponding files.

Bar chart was also considered, but no obvious trend can be seen when the values are in ascending order. Finally, we found that a distribution chart is most suitable for this case.

Relationship between Speed Limit, Time and Crash Severity

The csv used for this is a changed document from the original cleaned file. The changes made include:

- Change of time to decimals, so that it can be plotted on a cluster scatter graph
- Change of speeds to be non-rounded (they were rounded before), to provide more accurate results in the graph

```
data = pd.read_csv("time_altered_analysis.csv")

dataset = data[['Time', 'Speed Limit', 'Total Units', 'Total Fatalities', 'Total Serious Injuries', 'Total Minor Injuries']]
cor = dataset.corr()
sns.heatmap(cor, square = True)
```

A heatmap was created to understand the correlation between variables with the columns in question including:

- Time
- Speed Limit
- Total Units
- Total Fatalities
- Total Serious Injuries
- Total Minor Injuries

```
model = KMeans(3)
model.fit(dataset)
clust = model.predict(dataset)
center = model.cluster_centers_
kmeans = pd.DataFrame(clust)
dataset.insert((dataset.shape[1]),'kmeans',kmeans)
```

Three clusters were then created of the data in the dataset and their centers defined.

```
ax = plt.figure().add_subplot(111)
scatter = ax.scatter(dataset['Speed Limit'],dataset['Time'], c=kmeans[0],s=10)
ax.set_xlabel('Speed Limit (km/hr)')
ax.set_ylabel('Time (HH.MM)')
ax.set_title('Cluster Graph - Crashes in Time Vs Speed Limit')
start, end = ax.get_ylim()
ax.yaxis.set_ticks(np.arange(start, end, 5.04))
ax.yaxis.set_major_formatter(FormatStrFormatter('%.2f'))
plt.colorbar(scatter)
plt.xticks(rotation=45)
```

Finally the graph was then plotted with changes to the form of plotting including:

- Setting the x-axis label
- Setting the y-axis label
- Setting the title
- Minimizing the tens of thousands of variable notches for time to only six notches.

- Change y-axis to have two decimal points to mirror time format
- Rotate x-axis label values by 45 degrees to make them more readable

There are some strengths to cluster analysis using k-means

- K-means can produce tighter clusters
- Instances can change clusters whenever the centers/centroids are recomputed

While there are also some limitations

- It's difficult to predict the number of clusters there are
- The order of the data can alter the final graph
- It's extremely sensitive to the scale of the data

Relationship between Month and Crash Severity

For this, we created a Python program called ‘summaries.py’ which heavily utilized the package ‘statistics’. Within this program are two functions, with the function ‘crashSummary()’ being the main function used for this analysis.

This function takes in a string value which determines the crash severity you want a summary for. For example, if you wanted summaries for fatal crashes, your input would be ‘Fatal’ and for minor it would be ‘Minor’ etc.

This function reads through the dataset and uses a dictionary called ‘fatal_dict’ to store the required values. Its key corresponds to each month, and the value for each key is a list containing the values for the crash severity chosen.

Depending on the string input given by the user, the function utilizes if statements to determine what crash severity statistics to append to each dictionary list as seen below.

```
if month in fatal_dict:
    if crashType == "Fatal":
        fatal_dict[month].append(fatal)
    elif crashType == "Serious":
        fatal_dict[month].append(serious)
    elif crashType == "Minor":
        fatal_dict[month].append(minor)
else:
    if crashType == "Fatal":
        fatal_dict[month] = [fatal]
    elif crashType == "Serious":
        fatal_dict[month] = [serious]
    elif crashType == "Minor":
        fatal_dict[month] = [minor]
```

Once it read through the file and collected the relevant information, the function uses the ‘statistics’ package to create numerical summaries. In particular, the statistics.mean() and statistics.stdev() functions are used to calculate the mean and SD. Once the mean is calculated,

the value for the particular dictionary key is assigned to the mean. We then use list comprehension to find the dictionary keys with the highest means. This can be seen below.

```
for i,k in fatal_dict.items():
    total = sum(k)
    mean = statistics.mean(k)
    sd = statistics.stdev(k)

    print("Mean {} crashes for {}: {}".format(crashType, i, mean))
    print("Standard deviation of {} for {}: {}".format(crashType, i, sd))
    print("Total amount of {} crashes: {}".format(crashType, total))
    print()

    if mean > highestMean:
        highestMean = mean
        highestMonth = i

    if total > highestCrashCount:
        highestCrashCount = total
        highestCrashMonth = i

    fatal_dict[i] = mean

highestList = sorted(fatal_dict, key=fatal_dict.get, reverse=True)[:3]
```

Initially, we tried finding the average and standard deviation without the use of the 'statistics' library. However, it proved to be difficult and time consuming, so we searched up alternatives online. It was here that we learnt about the 'statistics' library and employed it for our analysis.

The reasons why we chose to create such numerical summaries is due to the low amount of fatal, major and minor crashes recorded. This is evident in the very low averages for these crashes in our results section, with a large majority of them being below 0.5 When graphing this, we found it very difficult to distinguish the differences between the different crash severities, so our solution was to create numerical summaries to scrutinize and examine the data. By creating these summaries, we are able to mathematically quantify our claims and conclusions which would have been otherwise impossible with graphs due to the small magnitude of the data.

However, a limitation is that compared to graphs, it is more difficult to determine any trends in the data. The visualizations created by graphs can help determine any strange anomalies in the

data as well as being far easier to interpret. Making graphs in say something like Excel is also a less time-consuming process compared to writing out specialized code to create summaries.