Group 3 Final Report - Auto-Pedestrian Crashes

Nick Nemkov, Claudia Iwinski, Sikander Raja, Jacob Rodriguez, Tanuj Chintalapudi

CMPS-240-02

Executive Summary

We wanted to find out when an auto-pedestrian accident is most likely to occur. To do this, we downloaded a data set from Kaggle that contained information about auto-pedestrian accidents in Wayne County, Michigan from the years 2010 to 2018. All accidents recorded involved a pedestrian on foot and an opposing force such as a motor vehicle, bicycle, or other auto transportation. After we obtained the data set, we formed a few sub-questions to help answer our main question:

1. Do more auto-pedestrian accidents happen at an intersection?
2. Does the day of the week play a role in the frequency of the accidents occurring? What is the severity of the accidents?
3. Does the time of day have any correlation with auto-pedestrian crashes since it becomes harder to drive in darker lighting conditions?
4. As we go on into the future, there will be more people; thus, more people will be on the road. Will an increase of drivers on the road lead to an increase in auto-pedestrian crashes?
5. As the speed limit increases, does the number of auto-pedestrian crashes also increase? On average, how severe were the accidents?
6. Which common individuals were involved in the accidents (age and gender)?

Based on the results shown from the data visualizations:

1. Accidents are more likely to occur in places other than intersections. (Fig. 1)
2. Friday had the highest amount of accidents and possible injuries while Sunday had the least of all injury severities along with the least accidents occurring. (Fig. 3)
3. Most accidents happen between 3:00PM and 8:00PM. (Fig. 4)

4. No, an increase of people on the road over the years did not increase the number of accidents (Fig. 5)
5. No, the number of auto-pedestrian crashes did not substantially increase past 35 MPH and most of the accidents only resulted in possible injuries. (Figs. 6 and 7)
6. The most common victims in the accidents were males aged 20-25. (Figs. 8 and 9)

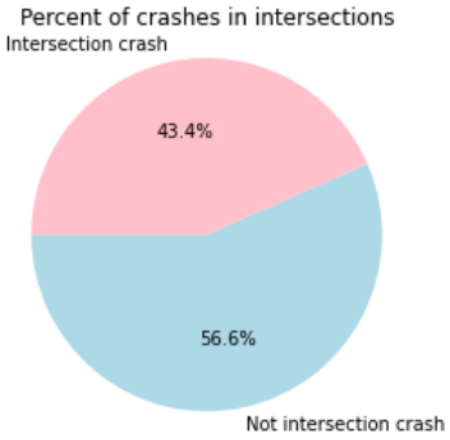
Data Cleaning

Cleaning the data set first started with the age column; some values couldn't be validated, replacing them with DOB invalid. 36% of the values in the age column were invalid, so these invalid DOBs became the mean age of 50 instead. In the second column, gender had a similar issue as the ages, with values not being assigned. 28% of the deals were uncoded or errors; if the values were replaced with male or female, it would've significantly skewed the results of this column. Instead, the invalid gender value rows were dropped entirely from the dataset. The third column, "Party Type," was removed because it contained only one unique value that wasn't a part of our analysis question. The speed limit column had some invalid values that couldn't be edited because values being replaced could inflate the data. These invalid values were all made into 0s and then removed. We decided to copy the primary data set for two other data sets since this cleaned data set removed necessary rows and data from our additional analysis.

Intersection Accidents and Hit-and-Run vs. Not-Hit-and-Run Accidents

Data was read from the original "Auto Pedestrian Crashes" data set from Kaggle by Syed Asim Ali Shah. Using the pd.read_csv command on 'ped_crashes.csv,' that was the name of the Kaggle file. From this data, another data frame was made for intersection crashes. Counting all the instances when intersection crashes were to happen and putting it into this new data frame.

After that, we made a Pie chart for the percent of crashes at intersections using matplotlib functions. This data showed that 2957 (43.4%) of crashes happen in intersections, while 3852 (56.6%) of crashes don't happen in intersections. From the data, an interesting finding between intersection crashes and hit-and-runs. Data from before was already clean from the primary Kaggle data set, lastly was putting the information into a Bar graph. Using the matplotlib function ax and title, we set x equal to intersection crashes and hue equal to hit and runs. The results showed fewer intersection crashes resulted in a hit and run, 1103 crashes, while none were 1501 crashes. Comparing the total none intersection crashes to a total of none intersection crashes that were a hit and run makes up about 39% of crashes. While the sum of intersection crashes to the sum of intersection crashes that were hit and runs 36% of intersection crashes. Meaning more crashes are likely to be hit and run if it is not happening in an intersection by 3%.



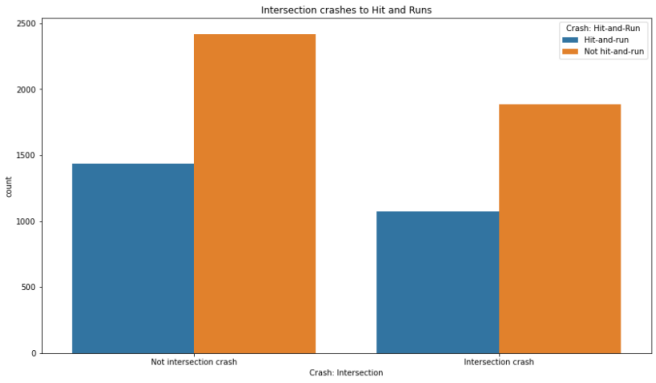


Figure 1: Occurrences of Accidents in Intersections Figure 2: Occurrences of Hit-and-Runs to Intersection accidents

Accidents by Day of Week (including injury severity)

Using the Pedestrian Auto Crashes dataset, the team constructed two counters for the total accidents. This was done by specifying a python command to count all unique values in the selected column. The first counter grouped the number of accidents by the day of the week. Out

of all the days, Friday had the most accidents at 778 cases. Sunday had the least amount of accidents at 488. The second counter added up the number of accidents based on the injury severity. Out of 4898 total accidents the most common severity was Possible Injury (1814 cases), and the least common is Fatal Injury (373 cases). Next, in order to see the pattern in the given data, bar graphs were plotted using first the accident cases by the day of the week. The next graph was a grouped bar plot displaying the accident cases per day of the week, each day having five subsections showing the total counts of accident injury severity for that day. It can be concluded that Sunday has the least amount of total accidents and the injury count is much less than for the other days of the week.

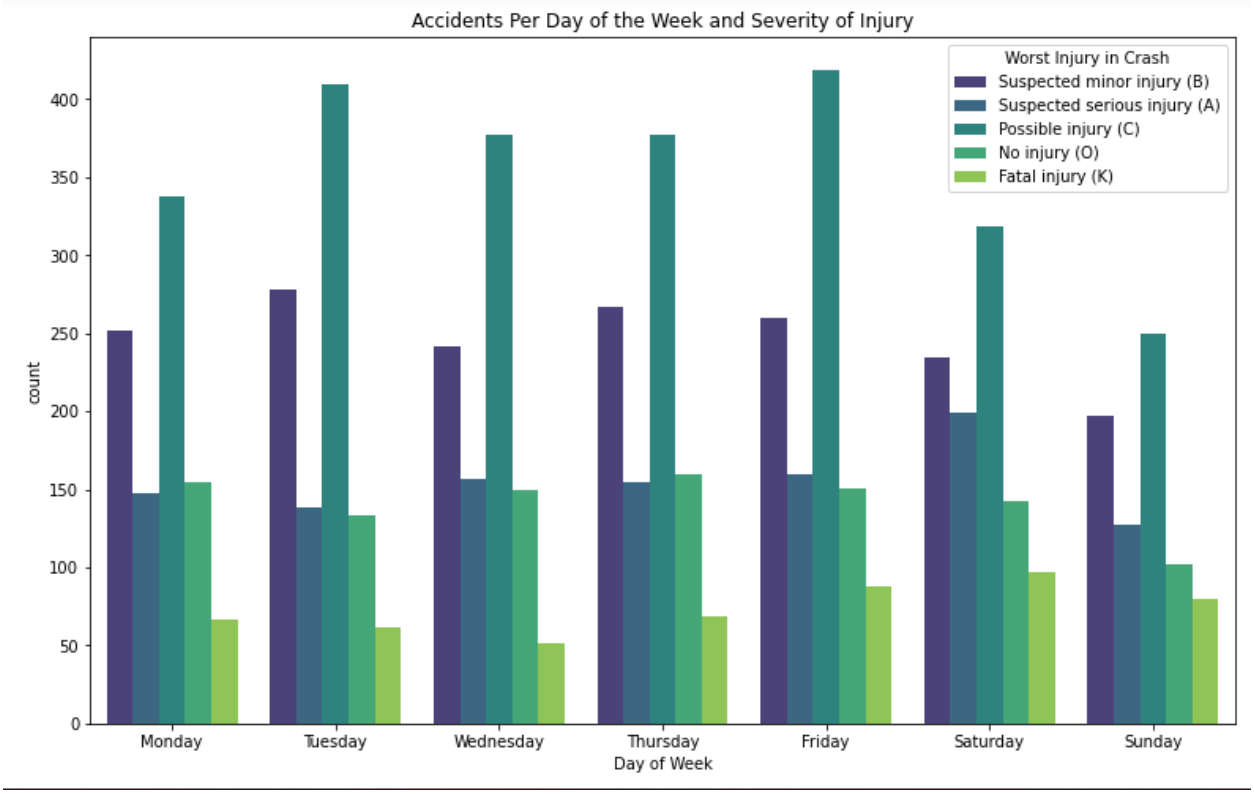


Figure 3: Visualization of Total Accident Injury Severity per Day of the Week

Accidents by Time of Day

Using the “Time of Day " column from the auto pedestrians crashes, we did a “line plot” to visualize in which intervals of time crashes are most likely to occur and which part of intervals of time is most likely not to occur. From “figure 4**”** we concluded that the time interval that accidents are most likely to occur is between 3:00PM - 8:00PM and the time interval that accidents that are most likely not to occur is between 3:00AM - 6:00AM.

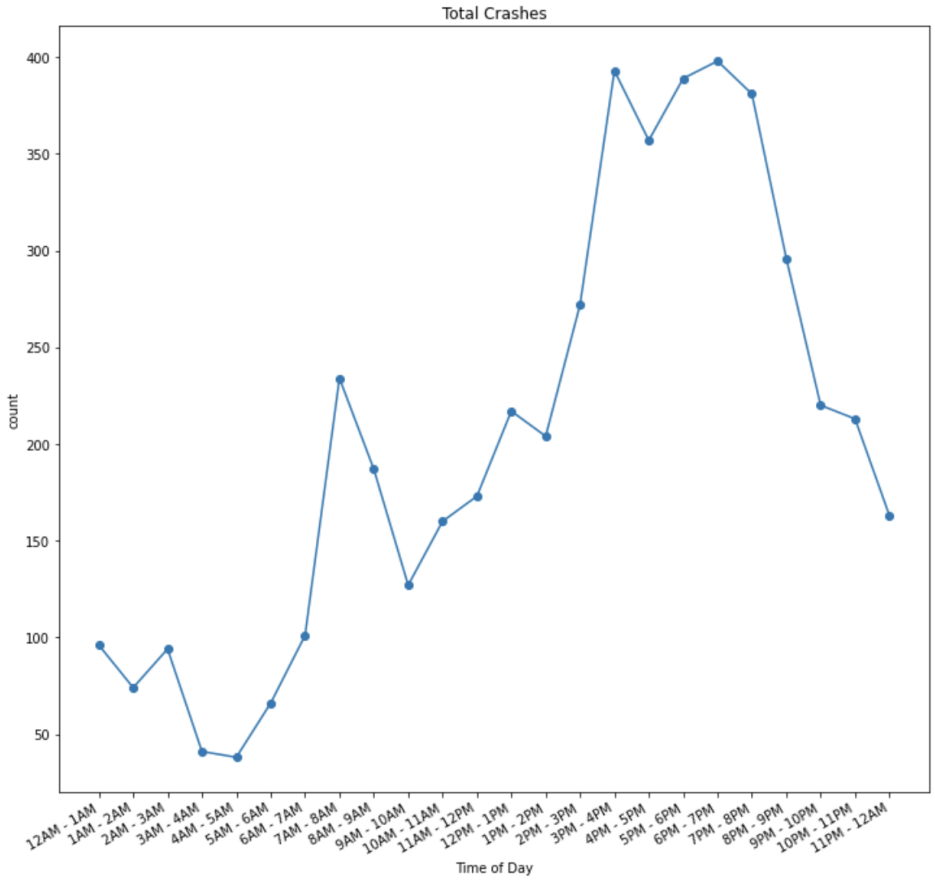


Figure 4: Visualization for Occurences of Accidents by “Time of Day”

Yearly Accidents

In order for us to answer one of our sub-questions, “As we go on into the future, there will be more people; thus, more people will be on the road. Will an increase of drivers on the road lead to an increase in auto-pedestrian crashes?”, we decided to illustrate our findings in a

line graph. Before this, the team constructed a counter for the total accidents that happened each year from the years 2010-2018. A specific python command allowed us to count all unique values in the selected column of "Crash Year". From this, it displayed the total count of accidents that occurred each year, with the year 2018 having the most with 822 accidents and 2016 being the year with the least, having only 651 accidents. Next, we concluded that a line graph would provide us with more optimal results to answer our question. After interpreting our line graph, we believed that there is no significant trend occurring due to it being very scattered. There is no definite increase or decrease as the years went on.

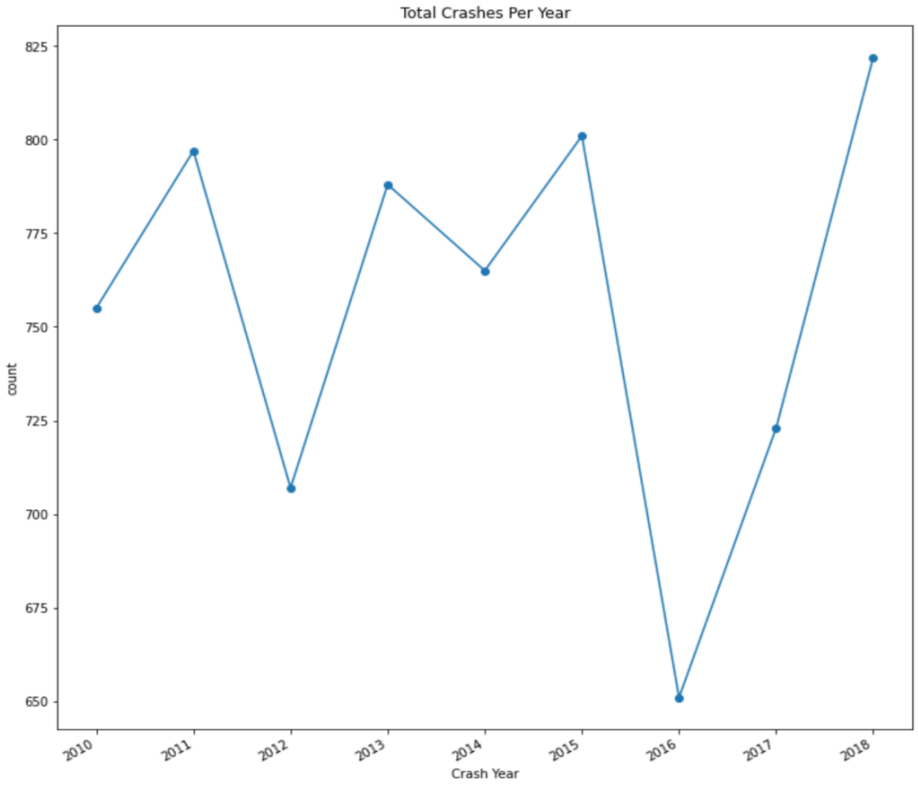


Figure 5: Visualization for the Number of Accidents Each Year

Accidents per Speed Limit in Area

Before developing any graphs showing the relationship between the speed limit at the crash site and the number of accidents, we first had to use the *astype* function to convert all the values in the original data set from strings to integers. Next we used the *unique* and *sorted* functions to get the unique speed limit values sorted in ascending order. Then we plotted these values using both a bar plot and a density plot to show the distribution and modes in the data. Since this data set is focused on auto pedestrian crashes, the data is skewed towards low speed areas that you would most likely find pedestrians in. There are also two smaller peaks at 55 and 75 MPH, but almost nothing at 60 or 65 MPH. We could infer that the majority of Wayne County's highways have speed limits of either 55 or 75 MPH.

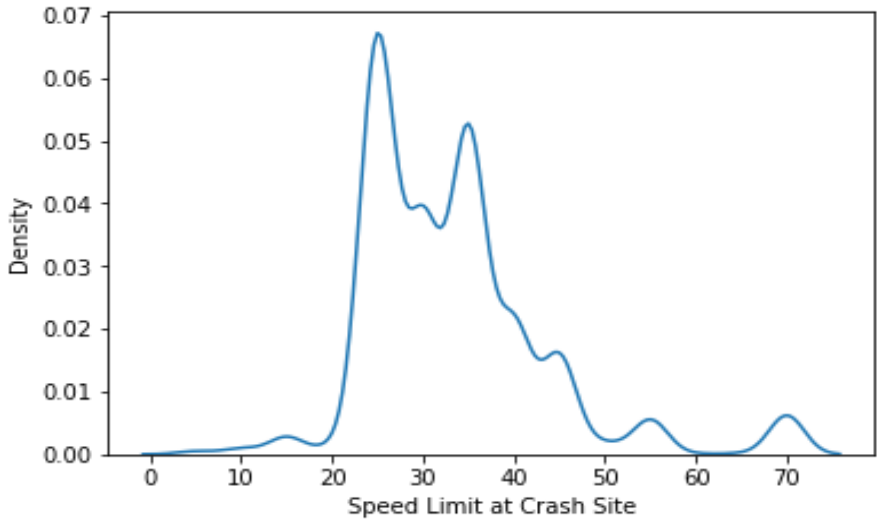


Figure 6: Visualization for the Number of Accidents at Each Speed Limit

Severity of Injuries Compared to Area Speed Limit

Since the prerequisite preparation of data was already completed in the previous section, not much more had to be done here. We decided to use a heat map to show the relationship between the severity of injuries and the speed limit at the crash site. Most of the accidents occurred between 25 and 35 MPH with only possible injuries. 55 and 75 MPH show a higher

occurrence of severe injuries compared to minor ones but occur much less frequently than accidents at lower speeds.

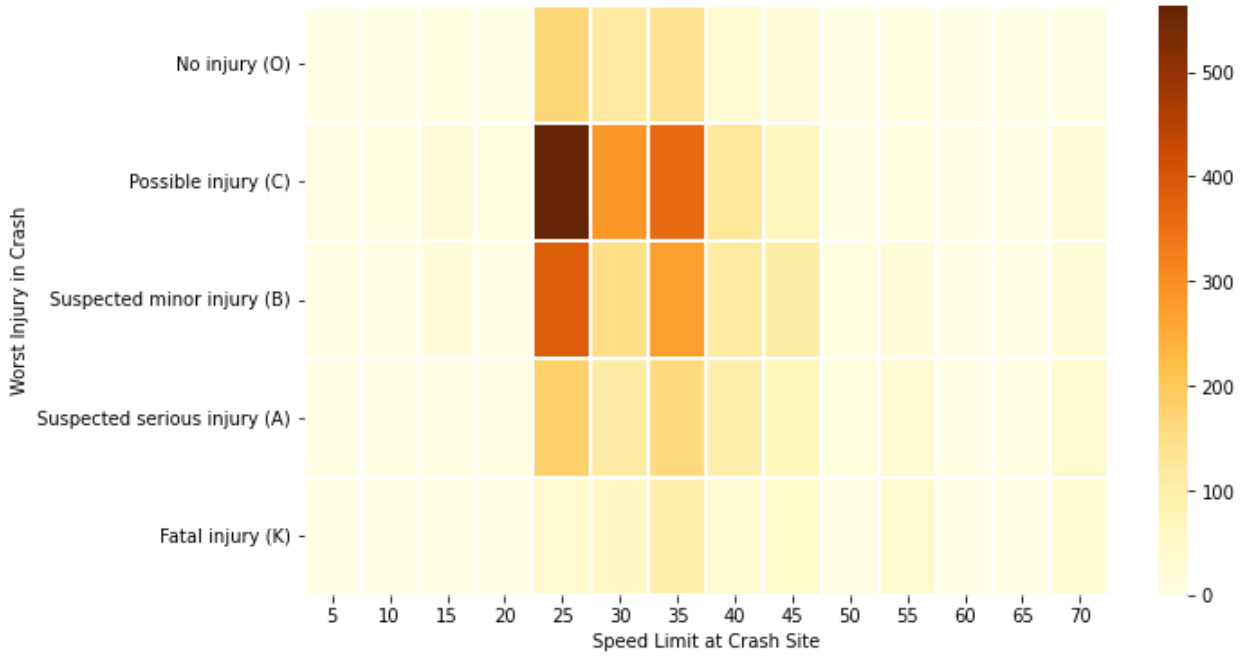


Figure 7: Visualization for the Severity of Injury at Each Speed Limit

Accident Victims by Age and Gender

Using the Columns “Person Age” and “Person Gender” from the dataset, it can be possible to discover the type of victims commonly present in most of the dataset accidents. The data was cleaned beforehand, with the “Person Gender” column being cleared of all values known as “Uncoded & Errors”. The “Person Age” column had all its “DOB invalid” values replaced with the mean person age in the dataset, 50 years old. To create a gender counter, the use of the *value_counts()* command was implemented. An age counter was performed much the same way. A pie chart was used to visualize the two genders, both categorical variables, as fractions of the total number of accidents. It can be seen that the division is about 60% male to

40% being involved in the accident. The total accidents by age were displayed with a histogram, showing that the most common accident victims were people of age 20 to 25.

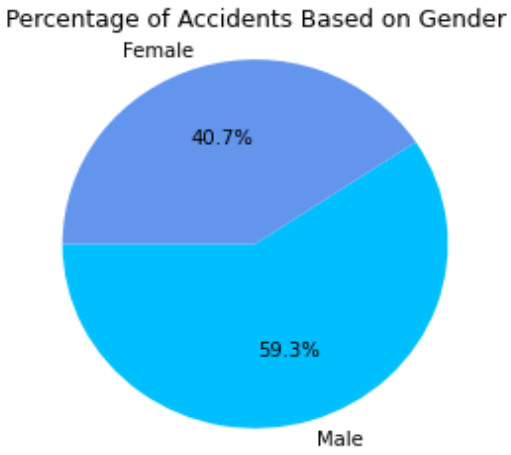


Figure 8: Gender of Accident Victims

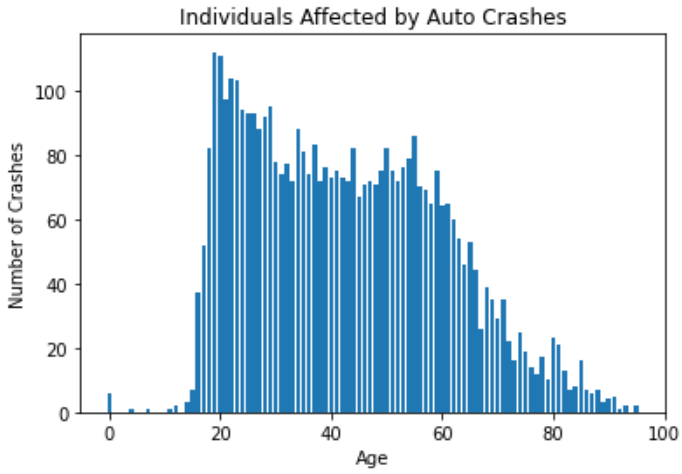


Figure 9: Number of Accidents Compared to Victim's Age

Conclusions

Based on our findings, auto-pedestrian crashes are most likely to occur with men under the age of 30 on Friday between the hours of 3 PM and 8 PM at speeds between 25 and 35 MPH with only possible pedestrian injuries sustained. Despite the increased likelihood of auto-pedestrian crashes in people under the age of 30, the total number of crashes per year has stayed mostly the same from 2010-2018 in Wayne County. On average, most of the auto-pedestrian crashes occurred at low speeds that you would find in a typical neighborhood with crosswalks. However, we saw from the data that auto-pedestrian crashes were less likely to occur in intersections with 43.4% of accidents occurring in intersections, and the other 56.6% of accidents occurring elsewhere. An increase of drivers on the road doesn't seem to lead to an increase in auto-pedestrian crashes.

References

Shah, S. A. A. (2022, October 20). *Auto pedestrians crashes*. Kaggle. Retrieved from https://www.kaggle.com/datasets/syedasimalishah/auto-pedestrians-crashes