Group 1 - 2019 Injury Report

R.S. & O.O.

Data Sourcing

The data for this report was collected from the United States Consumer Product Safety Commission's (CPSC) online databases. The access point for these databases was through a CPSC-provided Query Engine known as the National Electronic Injury Surveillance System (NEISS) Query Builder. The NEISS datasets provide stratified samples of hospital data based on emergency department size and geographic location. These samples are separated by their reported sex, so one can only request records for Males or Females due to the limitations of the Query Engine. Estimates are considered unstable when the estimate is for less than 1,200 data points, the number of records used is less than 20, or the calculated Coefficient of Variation for the dataset exceeds 33%. The Male and Female datasets queried for this report were for 20,000 records each and had a respective CV of 10%; well within the NEISS requirements for a reliable sample.

Data Stitching

Once the data for this report was sourced and locally stored, it had to be stitched into a single file for further analysis. Initially, this report was going to span National Injury Reports from 2015-2019, but after all of the data was stored in one location, it became clear that processing multiple GB's of unorganized information would quickly become unmanageable. It was therefore decided that this report would focus only on the most recent year of injury data across Males and Females in the USA; 5.6 MB of data totalling in 44,861 rows in Excel. After wrestling with various excel file managing services in python, R was selected as the data-management tool for this task. In 4 lines of code in R, the two individual excel files were bound into a single dataframe and saved back into local storage as an excel workbook, ready to be processed.

Data Pre-Processing

The programming language of choice for the data processing and analysis from this point onwards in the report was Python3. Leveraging the variety of statistical and machine learning libraries available (*numpy, pandas, datetime, etc*), the raw data was imported into the local programming environment for further analysis.

The first steps required was to format all columns appropriately. This was primarily important for the datetime data, as it was represented as a digit in Excel ie. 456223, and needed to be mapped from the relative TimeDeltaIndex to a specific datetime; the datetime library was very helpful in achieving this mapping.

Once the data was properly formatted with respect to type, it had to be appropriately decoded. Of the 25 columns provided in the raw dataset, 8 of them were strings encoded as numbers based on the NEISS Code Manual. In order for this report to provide any meaningful insights into the significance of the correlations between any of these dataset attributes, they had to be decoded into their human-readable form. The list below outlines the target columns for this decoding process:

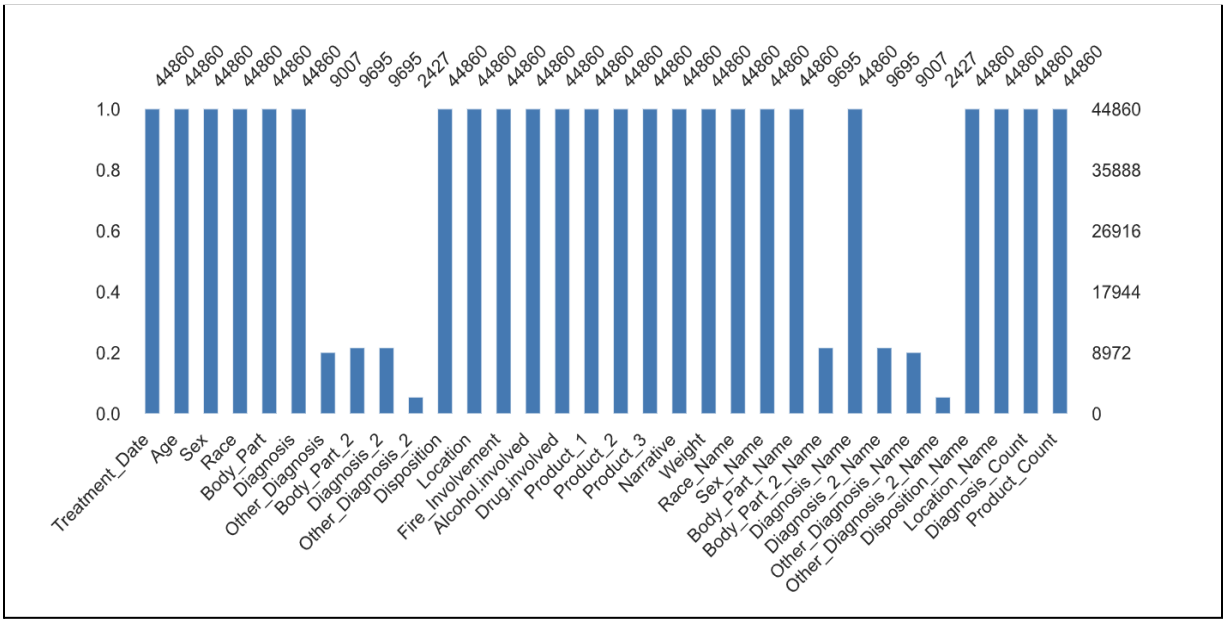
- Age (type formatting, encoding)
- Race (encoding, aggregation)
- Hispanic.Origin (encoding)
- Sex (encoding)

- Body_Parts (encoding)
- Location (encoding)
- Diagnosis (encoding)
- Disposition (encoding)

Data Pre-Processing

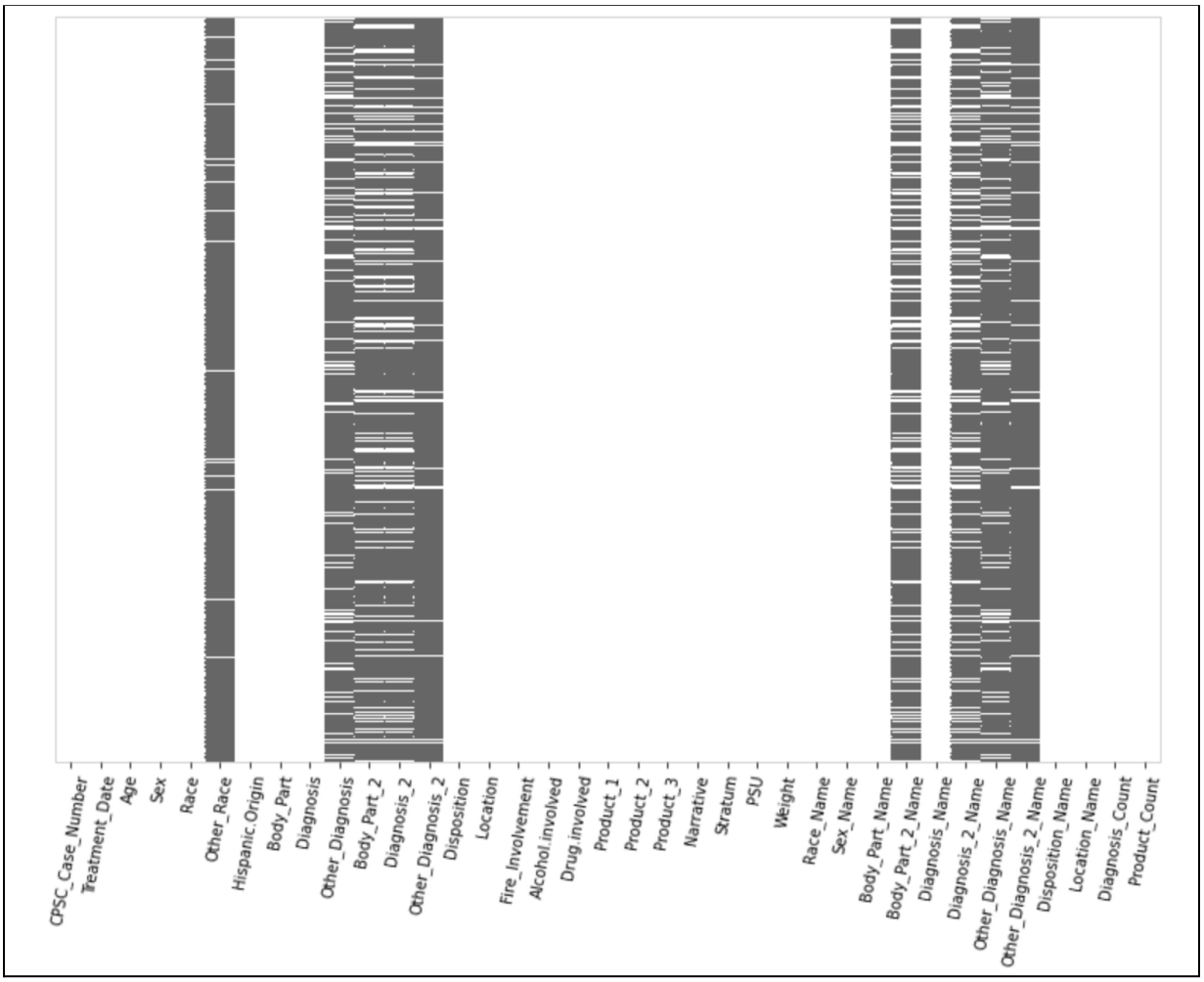
Once the data was successfully decoded from their numeric representations, visualizations were generated to support our comprehension of each variable's distribution and correlations with each other.

Figure 1.0 - Missing Values Bar Chart



From the Missing Values Bar Chart above, certain trends in the data began to appear. Beyond the correlations between the raw data columns and their named counterparts, there seemed to be a pattern emerging between which variables were missing data.

Figure 1.1 - Missing Values Heatmap

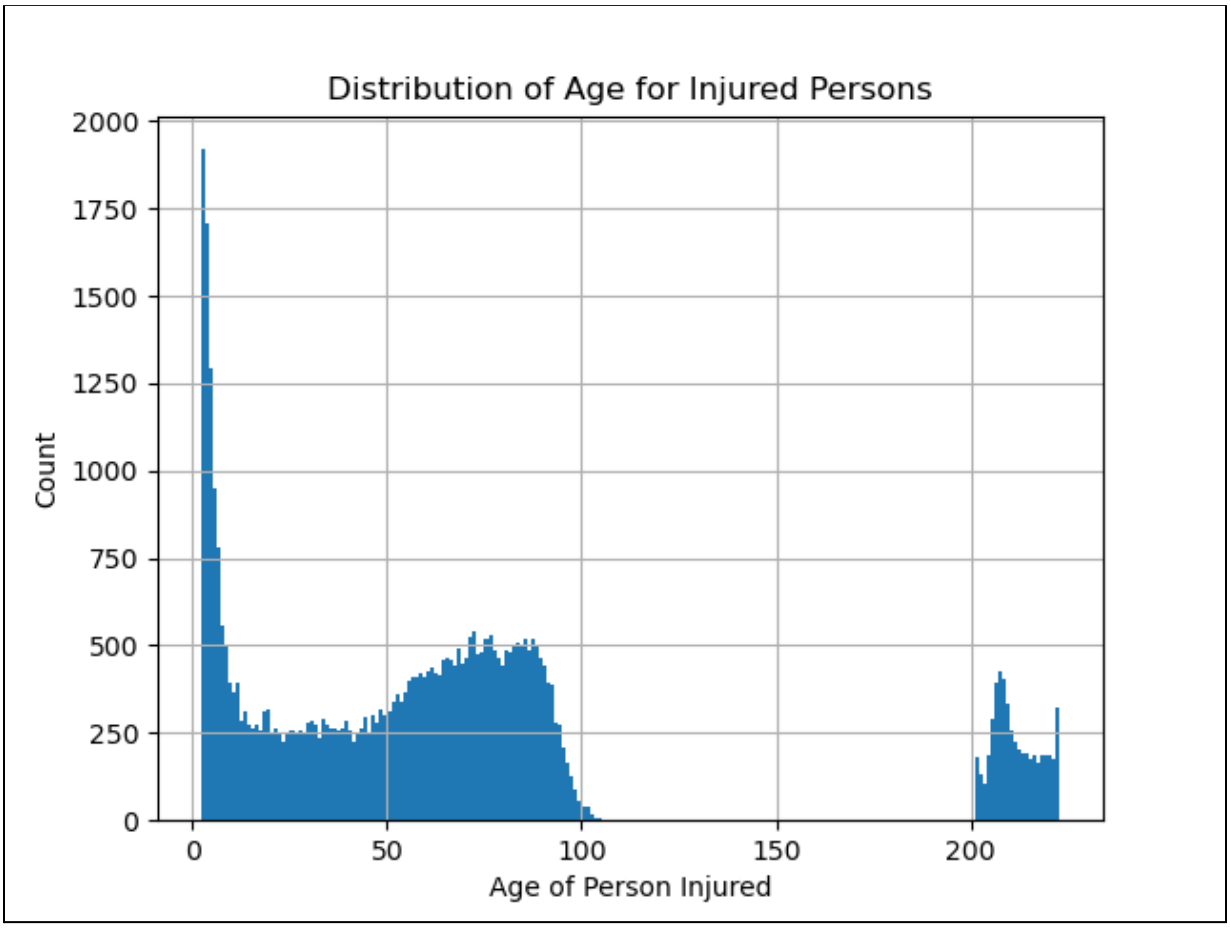


Leveraging the codebase available to us via the "*Data Mining for Business Analytics*" course materials, we were able to generate this shaded heatmap of the same dataset in order to view the aforementioned missing data trends from a new perspective. From this vantage point, it became clear that Body_Part_2 and Disgnosis_2 as well as Other_Diagnosis and Other_Diagnosis_2 had a 1-1 correlation with respect to missing or existing data points. This close correspondence between these variables led us to generate aggregations (*calculated columns*) from the dataframe to gain deeper insights into how these variables corresponded. These aggregations can be view in the Diagnosis Count and Product Count variables listed in the heatmap above, that are designed to keep track of the number of diagnosis and products involved with each row in the dataset.

Beyond the closely correlated variables under review, it was also noticed that Other Race and Hispanic Origin seemed to contain redundant and often missing data. We therefore decided to expand the Race category to accomodate for the few records with two Races listed as a new, unique race, and drop the columns "Other_Race" and "Hispanic_Origin". "PSU" and "Stratum"

were also dropped during this phase, not for lack of consistent data but because we could not identify a reliable definition for these variables in the NEISS documentation.

Figure 1.3 - Age vs Injury Histogram



One key attribute of our data set is the person who was injured by a consumer product. The data set includes several demographic indicators, such as age, race and gender. When examining the data we found instances of age that were beyond 200 years. This caused us to look at a distribution of ages across the entire dataset:

Note that the majority of ages fall between zero and 100 years, which meets our expectations for the general population. Then there is a group that starts at 200 years, and appears to be approximately 25 years wide. This was a significant number of cases, not just a couple errors. This caused us to go to the NEISS Data coding dictionary and find out if there was perhaps a valid meaning to ages greater than 200 years. Finding the age description informed us that for children below the age of 2 years, the data is recorded in months, with 200 years representing less than one month old and 223 years representing 23 months old. With this information, we had two options to address the data:

1) Replace the data for children under 2 with the average age between 0 and 2 years, which is one year. This would be easy to do, but we realize that injuries to young children may not follow a uniform distribution of ages.
2) Since the age field is a numeric value, we could replace the 200+ value with a fractional number of years, so 206 (6 months old) would be replaced with 0.5 years.
3) We could remove the data from the data set, which would mean the loss of 5094 affected records, out of our total dataset of 44,860 records.

The team noticed that young children represent a disproportionate number of injuries given the number of years included in this category. This has pushed us to use method 2, with the expectation that this data will be useful in the future.

The final data after modification of the age column with method 2 shows that the age range now meets expectations for the oldest persons, while also demonstrating that young children represent a significant portion of the public who suffers injuries from consumer products

Descriptive Analysis

Figure 2.0 - Injuries by Date

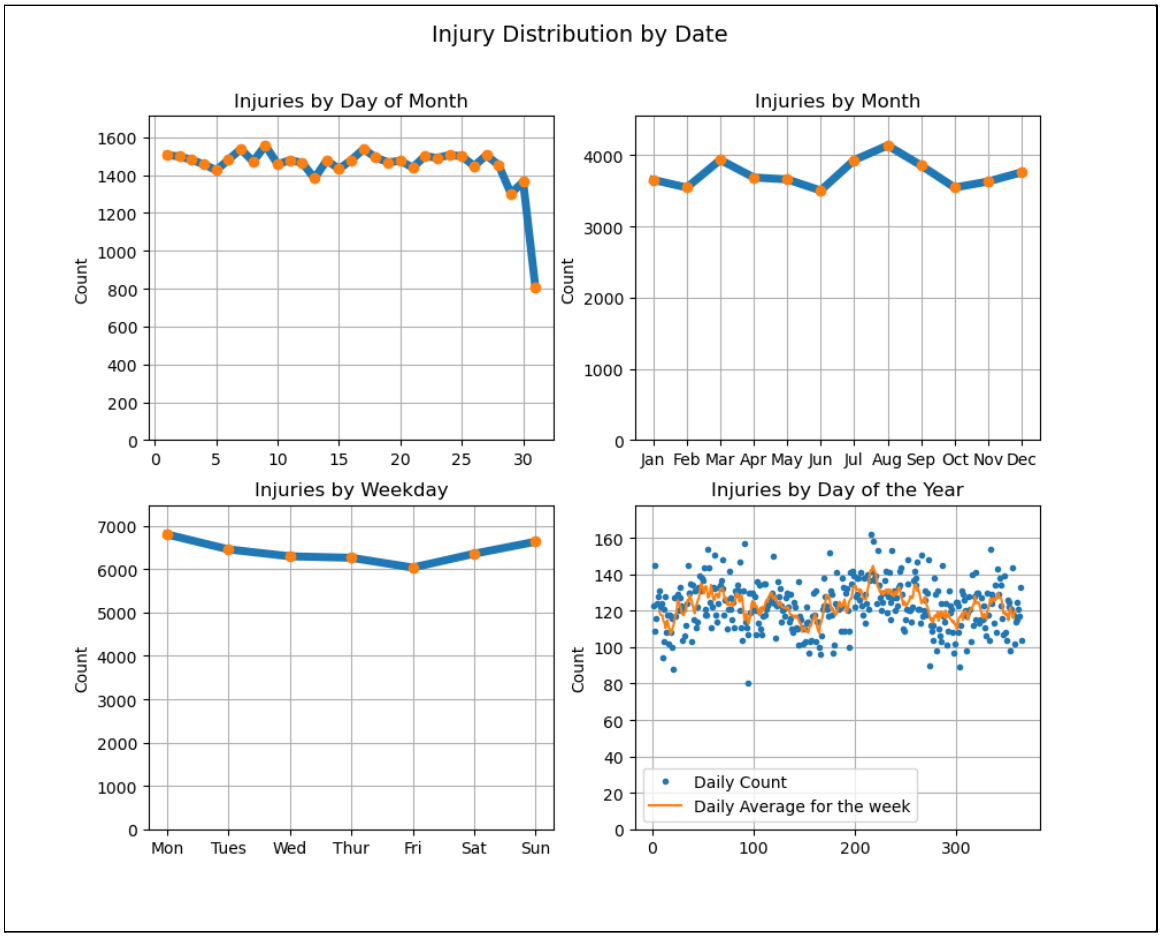
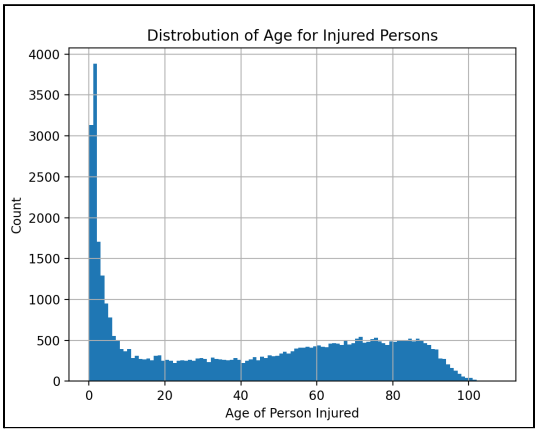
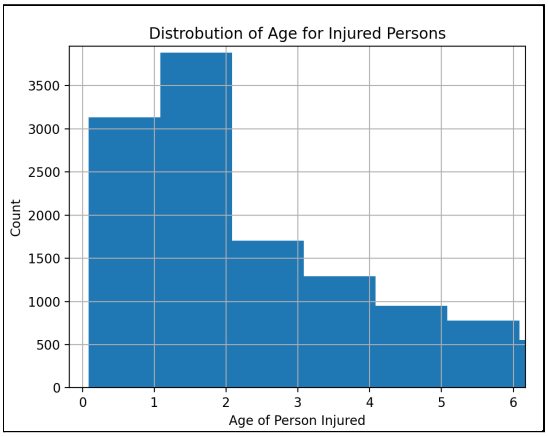


Figure 2.1 - Age Distributions





Each data record includes a date when the injury occurred. One way that we wanted to examine the data was to look at how records as a function of different measures of time. To do this we used the datetime() functions within python to extract months and days from the dataset. Then then used these new date and time to look for any unexpected trends in the data that might required deeper investigation. We started with the general hypothesis that no 'day of the month', 'month of the year', 'day of the year', or 'day of the week' is dramatically more likely to have injuries than other days. We then broke out those date/time designators and then used group-by() data aggregation to count() records within each of these groups. For the days of the year, we noticed a significant amount of noise from one day to the next, so we implemented a smoothing function to show us any structure in that data.

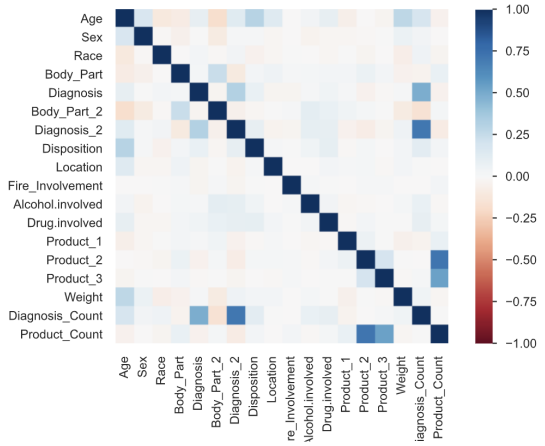
We found that in general, our hypothesis was confirmed, no date or month is significantly more dangerous than others, but there are some details worth noting:

1) There is an indicator that suggests the 29/30/31st days of the month appear to be less likely to have an injury than other days. We realized that not all months have more than 28 days, so those days of the month are less likely to occur, likely accounting for the lower number of injuries.
2) There does appear to be more accidents in some months, accidents in August appears to be approximately 10% more likely than in June or October. This could be related to weather, vacation patterns, school, or other factors.
3) Similarly, it appears that accidents are more likely on some days of the week, with Friday being the safest day, with about 15% fewer accidents happening compared to Monday.

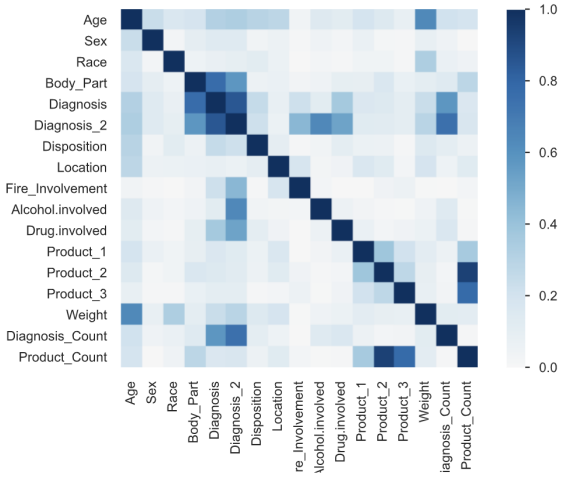
Analyzing the correlations between each variable, we utilized the pandas Pearson and Phi-k correlation matrices to identify both the linear and categorical (non-linear) dependencies that existed in our dataset.

Figure 2.2 - Correlation Matrices

Pearson Correlation Matrix



Phi-K Correlation Matrix

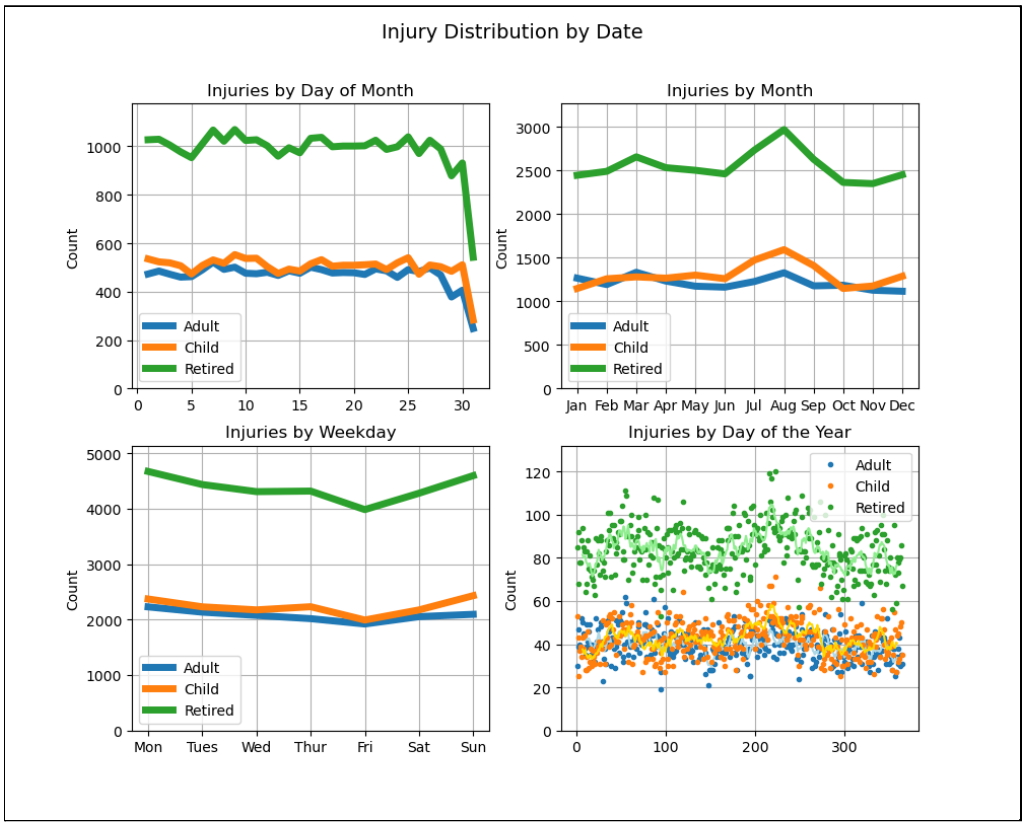


To do a rough exploration between age groups, the team split the dataset into three groups, persons under 18 years of age (children), people above 65 years (retirement age), and people between those ages (adults). The goal was to look for any significant differences in the timing of accidents between those different groups. The same metrics were used as the general population, and the three groups were plotted on the same plots.

Figure 2.3 - Injury vs Date

Group 1 - 2019 Injury Report

R.S. & O.O.



This figure provides a couple insights:

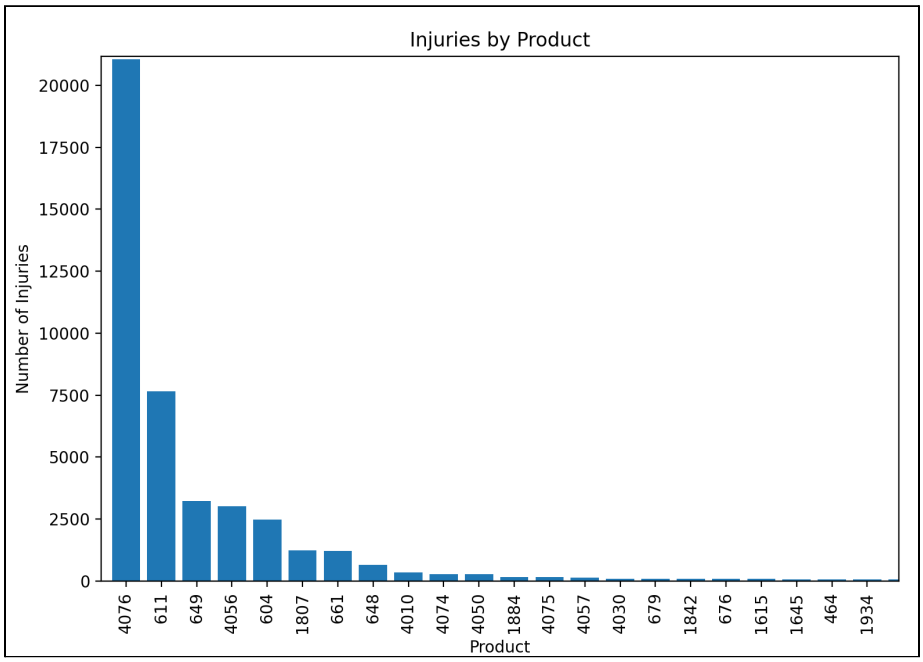
1) Accidents are approximately twice as likely for people over the age for 65 astray are for people below the age of 65.
2) The Adult group appears to sho less variation across days of the week, and months of the year when compared to the other two data groups.

An initial assessment was made to highlight the products that are associated with the highest number of injuries across all records. The team felt that there may be some products that are more dangerous than others, so the data was aggregated for each product, and the products were sorted by the number of injuries.

The results were surprising, with one product contributing to more than 20,000 injuries across a dataset of 44,860 records. This suggests a few things:

1) The dataset may be highly skewed towards the circumstances surrounding this product
2) A small number of products account, approximately 10, account for the vast majority of injuries in the dataset.
3) The team may need to account for the fact that of the 332 total products accounted with an injury, the majority of records are associated with fewer than 10 projects.

Figure 2.4 - Injuries vs Product



Observing these injuries by their frequency of body parts, we see the following trends:

Figure 2.5 - Injuries vs. Body Parts

Group 1 - 2019 Injury Report

R.S. & O.O.

Value	Count	Frequency (%)
Head	12907	28.8%
Face (including eyelid, eye area and ...	5298	11.8%
Trunk, lower	5209	11.6%
Trunk, upper (not including shoulders)	3155	7.0%
Shoulder (including clavicle, collarbone)	1763	3.9%
Knee	1670	3.7%
Not recorded	1409	3.1%
Leg, lower (not including knee or ankle)	1373	3.1%
Foot	1164	2.6%
Mouth (including lips, tongue and teeth)	1086	2.4%
Other values (16)	9826	21.9%