Homework Report

Name: Pranab Bhadani
Ph: 812-369-1845
Email: pranabbhadani@gmail.com

Question 1

What are the top three body parts most frequently represented in this dataset? What are the top three body parts that are least frequently represented?

Result:

Thus, the top three body parts that are most frequently represented in this dataset are **Head, Trunk lower** and **Finger**.

The top three body parts that are least frequently represented in this dataset are **25-25% of body burn, pubic region** and **Internal injuries**.

```
In [57]: df_merge[:3]['BodyPart']

Out[57]: 0            Head
         1    Trunk, lower
         2          Finger
         Name: BodyPart, dtype: object
```

```
In [71]: df_merge[len(df_merge)-4:]['BodyPart']

Out[71]: 22          Internal
         23      Not Recorded
         24       Pubic region
         25    25-50% of body
         Name: BodyPart, dtype: object
```

Approach used:

1. Fetched the csv file into a panda DataFrame.

2. Renamed the ‘body_part’ column name to ‘Code’. Reason for doing so is make a join with the BodyParts dataset for fetching the body part name associated with the code.

3. I found that there are records in the dataset whose age values are beyond 200 which is safe to assume as a corrupt entry in the dataset. Considering that, such records are really few (~5% of the total record) it is safe to discard it for further analysis.

4. After that grouped all the data based on the ‘code’ column applying ‘groupby’ method in dataframe. Then applied count () aggregate operation on the grouped column to find the frequency of each body parts.

5. The resultant data coming from groupby and count operation is then sorted using sort() method.

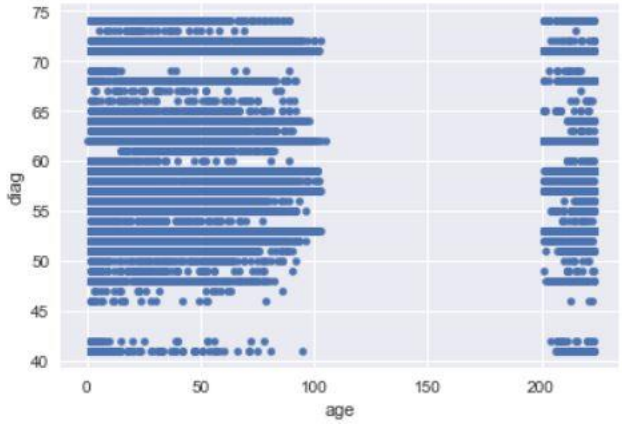
6. This dataset is then merged with the BodyParts dataset to fetch the name of the bodyPart w.r.t

the code.

7. Top 3 and the last 3 entries from the sorted data are then sliced out from the dataframe to come up with the final result.

Issue encountered:

1. First issue encountered in this analysis is regarding the age values. There are around 5% of the total records where age value is beyond 200, which is practically infeasible. Scatter plot points out this issue.



The above plot shows that there are records where age values are greater than 200. Since, such records are few, removing these will not affect our further analysis.

2. Second issue is regarding the 3 body parts least represented used in the dataset. Based on the result found, 'not recorded' turns out to be the third list body part which I believe cannot be kept in the category of body parts. So, I have fetched the 4th least represented body part and included that in my result.

Question 2

How many injuries in this dataset involve a skateboard? Of those injuries, what percentage were male and what percentage were female? What was the average age of someone injured in an incident involving a skateboard?

Result:

There are **495** injuries in the dataset that involve skateboard.
Out of these injuries **82.02**% were male and **17.98**% were female.
The average age of someone injured in an incident involving a skateboard is **17.89**

Approach used:

1. Fetched the dataset into a Panda dataframe.
2. From the NEISS Coding manual found that the code for skateboard is 1333. Filtered all the records that involved skateboard (code: 1333) in the prod1 or prod2 column and kept it in a new dataframe (df_skateboard)

3. To find the frequency of male and female, applied value_count() method on the ‘sex’ column of the filtered dataframe(df_skateboard).

4. Since now I have a series containing male and female frequency I have applied basic percentage formula to comeup with the answers.

5. To get the mean age of the people that got injured which involved skateboard, applied an aggregate function mean() on the ‘age’ column of the dataframe which gives me the result as 17.89

Question 3

What diagnosis had the highest hospitalization rate? What diagnosis most often concluded with the individual leaving without being seen?

Result:

Part1

Based on the analysis, **‘Fracture’** had the highest hospitalization rate. The following result includes the case of diag_other column in the dataset too:

```
In [25]: df_filter['Diagnosis'].value_counts()

Out[25]: Fracture                                          2078
         Internal organ injury                              731
         Other/Not Stated                                   676
         Contusions, Abrasions                              246
         Laceration                                         226
         Poisoning                                          106
         Ingested foreign object                             74
         Concussions                                         74
         Burns, thermal (from flames or hot surface)         56
         Strain or Sprain                                    54
         Dislocation                                         51
```

Part2

‘Laceration’ and **‘Pain’** diagnoses most often concluded with the individual leaving without being seen.

Approach used:

1. Fetched the data into a panda dataframe.
2. Replaced the ‘diag’ column with the name ‘Code’ to facilitate the join between this table and the diagnosis dataset.
3. Since the disposition code 4 and 2 signifies the case where the patient was hospitalized or transferred to another hospital. So, I have filtered all the rows from the main table which has disposition code as 2 or 4.
4. Next step was to count and sort the result to find the diagnosis(injuries) with highest frequency in the dataset. I have considered both ‘diag’ and ‘diag_other’ in my analysis. It turns out that ‘Fracture’ injury has the most number of case where the patient has been hospitalized.

5. In second part, Disposition code 6 gives me the case where diagnosis most often concluded with the individual leaving without being seen. So, filtered all the rows with disposition value as 6.
6. After that, sorted the filtered row on the basis of ‘diag’ and ‘diag_other’ to find the injury code with highest frequency.
   Turns out that **‘laceration’** and **‘Pain’** are the two highest injuries where the diagnosis concluded with the individual leaving without being seen.

Issue encountered:

The diagnosis caused by ‘other/not-stated’ had the highest frequency where the individuals left without being seen. So, I took the second highest case i.e ‘Laceration’ to be the diagnosis that satisfies the condition.

Question 4

Visualize any existing relationship between age and reported injuries.

Step 1

The first step to start the exploration is to find whether or not there is any correlation between the two variables: Age and Injuries. So, I have first made a correlation chart to gain some insight.

Correlation Table

```
df_age.corr()
```

Out[84]:

	CPSC Case #	psu	weight	age	diag	body_part	disposition	location	fmv	prod1	prod2
CPSC Case #	1.000000	0.006579	0.073466	0.040269	-0.009286	0.006943	0.012417	-0.009365	0.000550	-0.013374	-0.007592
psu	0.006579	1.000000	0.308727	0.066369	0.031617	-0.006640	-0.020401	0.002262	0.000976	-0.003859	0.002312
weight	0.073466	0.308727	1.000000	0.181417	-0.004368	-0.002385	-0.031574	-0.023806	0.000549	-0.009950	-0.016183
age	0.040269	0.066369	0.181417	1.000000	0.054587	-0.004451	0.223521	-0.185748	0.019206	-0.031824	0.015140
diag	-0.009286	0.031617	-0.004368	0.054587	1.000000	0.000842	0.035075	0.000699	-0.039395	0.011743	0.009428
body_part	0.006943	-0.006640	-0.002385	-0.004451	0.000842	1.000000	0.022449	-0.090288	0.020757	-0.081201	-0.016241
disposition	0.012417	-0.020401	-0.031574	0.223521	0.035075	0.022449	1.000000	-0.046075	0.026049	0.015218	0.012375
location	-0.009365	0.002262	-0.023806	-0.185748	0.000699	-0.090288	-0.046075	1.000000	-0.028546	0.085102	0.056845
fmv	0.000550	0.000976	0.000549	0.019206	-0.039395	0.020757	0.026049	-0.028546	1.000000	-0.033025	-0.028820
prod1	-0.013374	-0.003859	-0.009950	-0.031824	0.011743	-0.081201	0.015218	0.085102	-0.033025	1.000000	0.079463
prod2	-0.007592	0.002312	-0.016183	0.015140	0.009428	-0.016241	0.012375	0.056845	-0.028820	0.079463	1.000000

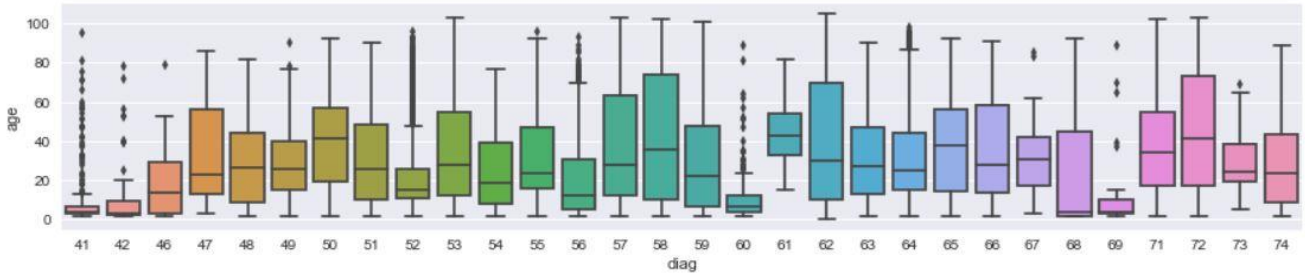
As from the above table we can see that the correlation between ‘age’ and ‘diag’ is very low with value being (0.054). So, this does not give much insight on any relationship between ‘age’ and the ‘injuries’.

Step2

To gain better understanding of the relation between the two parameters, I decided to build a box plot where the ‘x’ axis is diagnosis code and ‘y’ is age.

Result:

```
<matplotlib.axes._subplots.AxesSubplot at 0x174abfe6b38>
```



We can make certain inferences from the above box plot.

Inference:

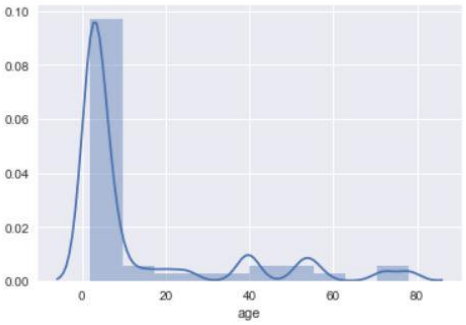
- Injuries code (41, 42, 60 and 69) that correspond to ingested foreign object, Aspirated foreign object, Dental Injuries and Submersion (including Drowning) respectively, happened more to lower age people (<20 years) i.e., based on the data we can infer that these injuries are more prone to children.
- If we consider the plot we can see that injuries with code (61 and 50) reported for older age people with mean age beyond 40. These codes correspond to Nerve Damage and Amputation.

Step 3

To further investigate the above trend (derived from box plot), I built a distplot (distribution plot) that would give more granular insight and clarity on the above observation.

For injury code 42, the distplot constructed with the help of matplotlib looks like this:

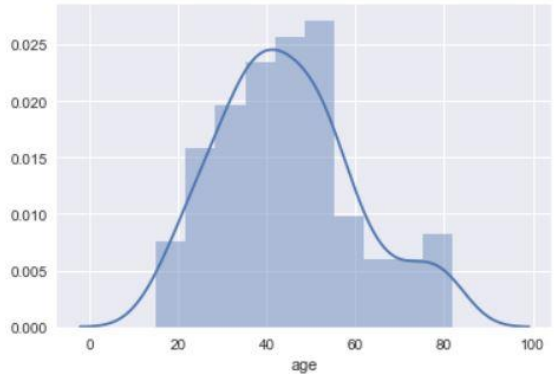
```
In [37]: df_age_42 = df_age[df_age['diag'] == 42]
         sns.distplot(df_age_42['age'],bins=10)
Out[37]: <matplotlib.axes._subplots.AxesSubplot at 0x173b33ff828>
```



This plot confirms that the injury code 42 (injury due to aspirated foreign object) is more likely to happen to children.

Now l plotted the distribution curve for injury code 61.

```
In [41]: df_age_61 = df_age[df_age['diag'] == 61]
         sns.distplot(df_age_61['age'],bins=10)
Out[41]: <matplotlib.axes._subplots.AxesSubplot at 0x173b2ef4ba8>
```

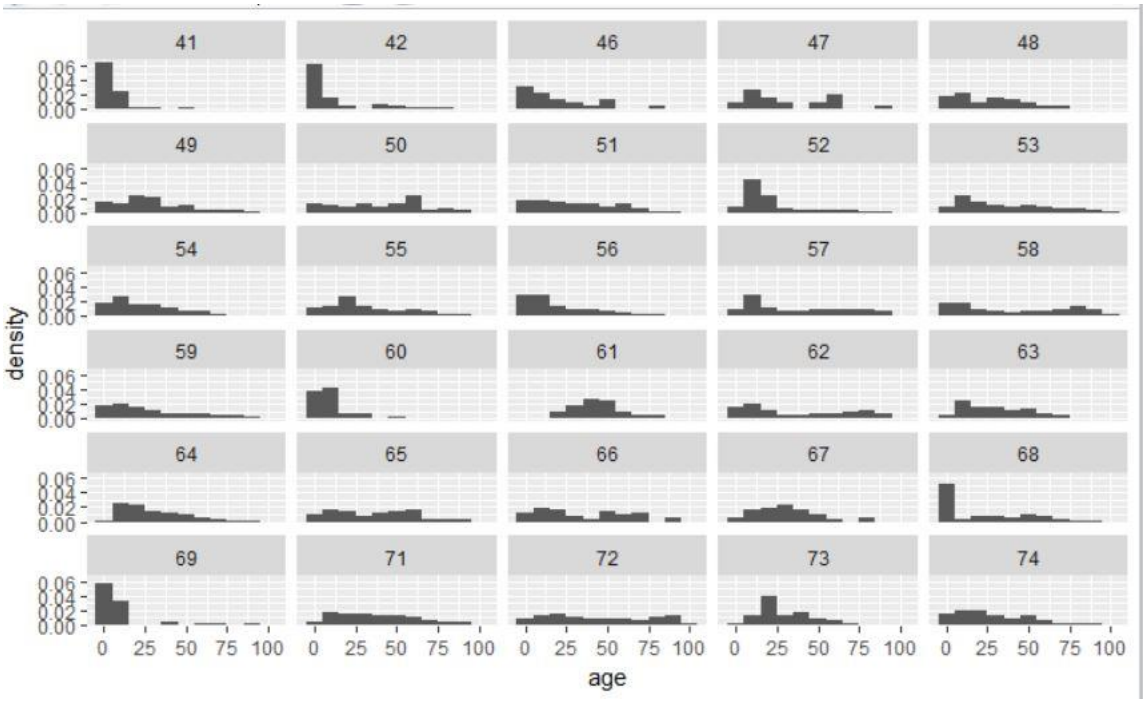


This plot depicts the same things as suggested by the box-plot, this shows that the injury (61) i.e. Nerve Damage are more frequent for older people with meanage > 40.

Step 4

Finally, I constructed facetgrid to discover any relationships between age and the reported injuries. This approach will ‘facet’ the entire graph on the basis of column ‘diag’ (reported injuries). For each grid I made a distribution plot on the basis of age.

Result looks like the following:



Conclusion

As evident from the above facetgrid plot of 'diagnosis' and the age distribution, we can weakly conclude that, except a few, the injuries are mostly reported for children and tend to decline as the age increases.

But, we cannot make a strong conclusion that injuries decline as the age increases, as we saw some evidences of the 'higher-age' injuries.

Question 5

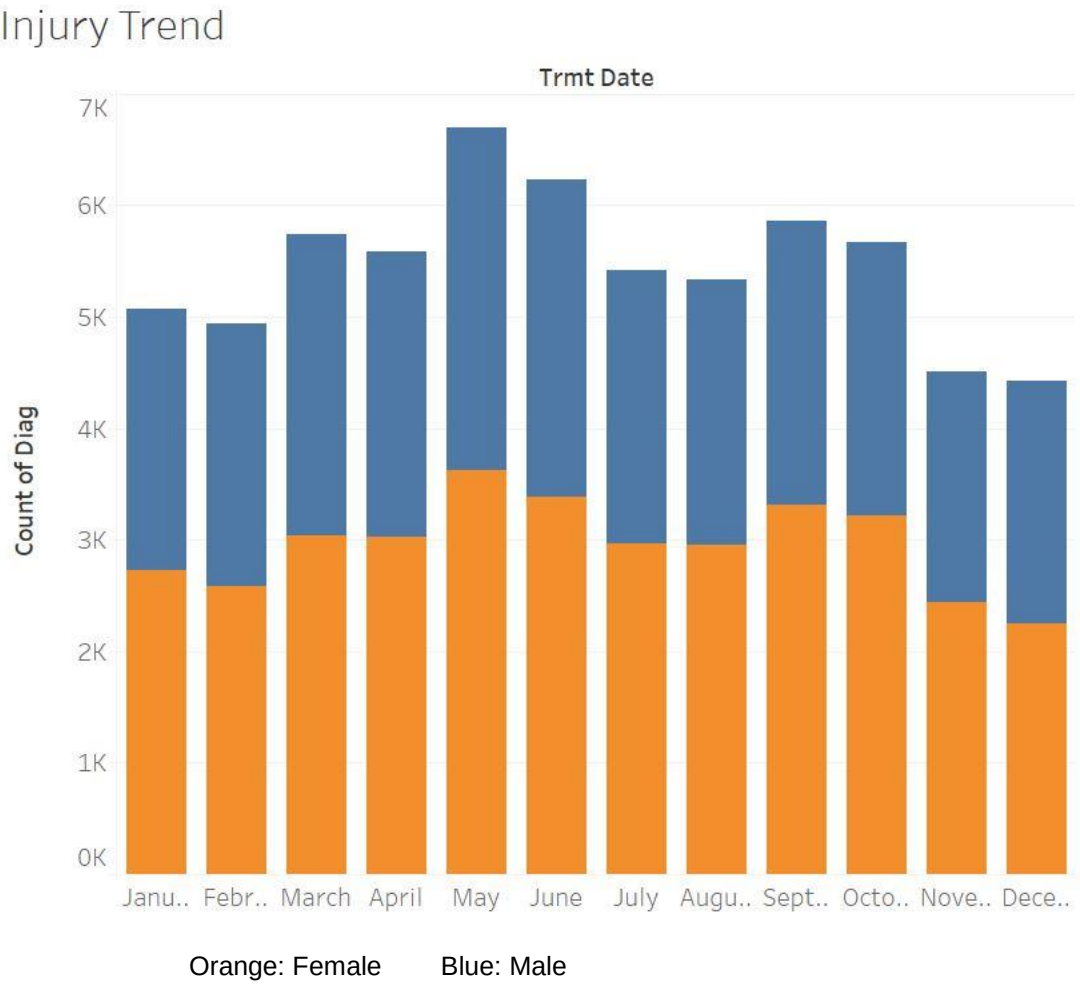
Investigate the data however you like and discuss any interesting insights you can find in the data.

On further investigating the dataset, I found few interesting result which I am going to highlight here:

1st Insight

My first approach is to see if there is any interconnect between any of the parameters in the dataset with respect to time.

Since the dataset captures the reported date of injury, I found an interesting insight that 'May is the most injury prone month'. I used Tableau to plot the following histogram that shows the relationship between different months and the report number of injuries.



Orange: Female Blue: Male

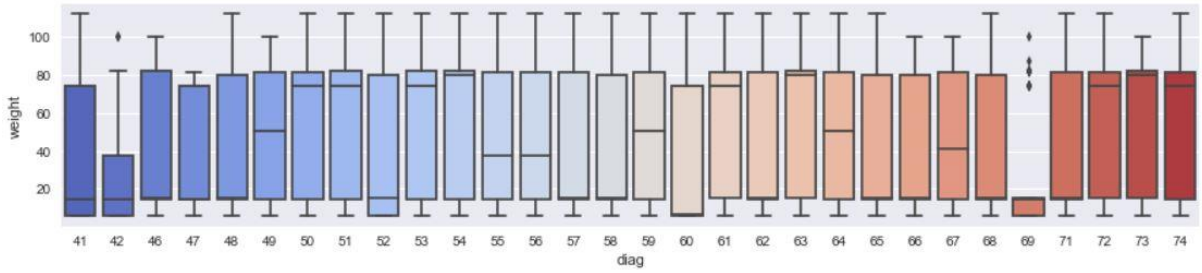
We can see that most of the injury were reported in May. This can be attributed to the fact that May and June are summer months and as we have seen earlier that major distribution of injury are reported for people of lesser age. This observation can be attributed to the summer break where children move out more and hence are more prone to injuries.

2nd insight:

The second most interesting insight was the association between weight and reported injuries

I have used a boxplot to show the relationship between this two factors.

```
<matplotlib.axes._subplots.AxesSubplot at 0x179fedd70b8>
```



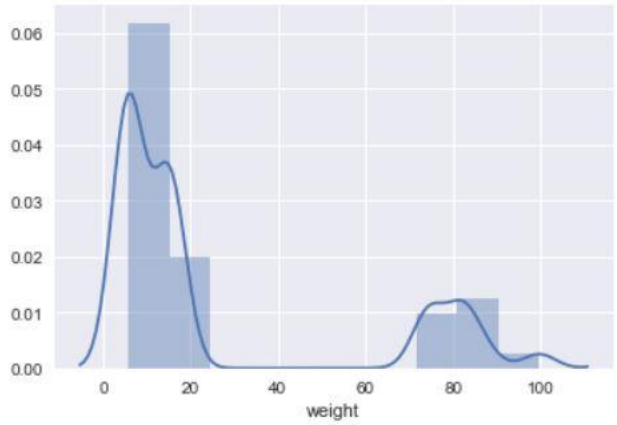
Conclusion

As we can see from the above plot, most of the injuries are uniformly distributed over the entire weight range, but there are two injuries that are mostly reported for people with lower weight (less than 40 units)

So, the injury #42 (aspirated foreign object) and #69 (submersion) are more likely to happen to ‘lighter’ people.

The above conclusion can also be verified from the following distribution plots for injury #42 and #69.

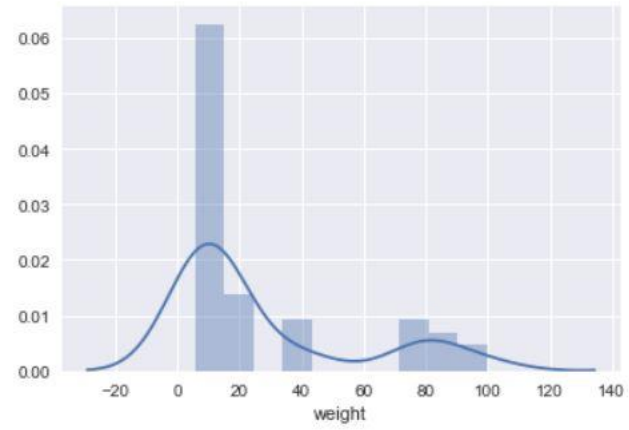
```
In [14]: sns.distplot(df_age_69['weight'],bins=10)
Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x179843c2e80>
```



Injury #69

```
df_age_42 = df_age[df_age['diag'] == 42]
sns.distplot(df_age_42['weight'],bins=10)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x17986800c18>
```



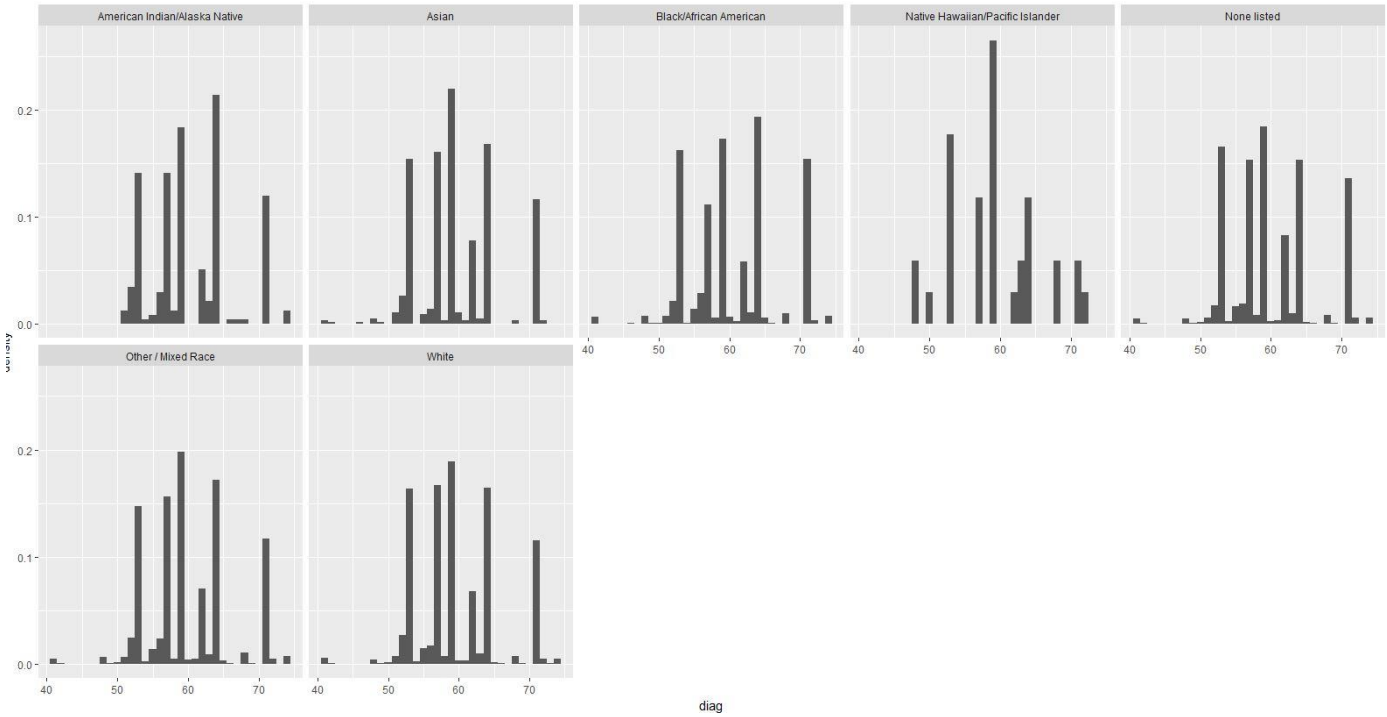
Injury #42

The above plots further corroborate the previous claims

3rd insight

The other interesting insight was the relationship between race and type of injury.

I plotted the following facet graphs, faceting based on ‘race’ and finding the distribution of injuries across ‘races’.



Conclusion:

Based on the above plots, we can certainly say that all races are most prone to submersion (injury #69) except American Indian/Alaska Native and Black/African American who are more prone to laceration (injury #59).