Lab 1 - PECARN TBI Data, Stat 215A, Fall 2024

Lab 1 - PECARN TBI Data

Stat 215A, Fall 2024

Introduction

Traumatic Brain Injury (TBI) is a signficant concern in pediatric healthcare, particulary when evaluatong childern who have head trauma. Identifying clinically important TBIs (ciTBI) is helpful for taking necessary futher medical intervention such as neurosurgical treatment or CT scans. The data we are using is from Pediatric Emergency Care Applied Research Network (PECARN), which is a dataset used to research acute injuries and illnesses among children in a wide range of demographics and institutions. In this lab, we aim to study this dataset and perform Exploratory Data Analysis to discover any patterns and insights for detecting the risk of Traumatic Brain Injuries in patients younger than 18. To diagnose patients with TBI, doctors must perform Computed Tomogorpahy (CT) scan. However, according to many studies, CT imaging of head-injured children has risks of radiation-induced malignancy. Most of patients with Minor Head Trauma (based on the Glasgow Coma Scale scores of 14-15) accounts for 40-60% of assessments, yet, less than 10% show signs of actual TBI. Therefore, creating a decision rule for identifying ciTBIs without excessive use of CT scans is the goal of this study.

We will first start with understanding the datasets and patterns in the features. Then, we will analyze the data cleaning process, as well as, justify the judgment calls made in this report.

Data

The dataset includes children under 18 years who presented with minor head trauma in emergency departments within 24 hours of injury and had Glasgow Coma Scale (GCS) scores of 14-15.

	PatNum	EmplType	Certification	InjuryMech	High_impact_InjSev	Amnesia_verb	LOCSeparate	Loc
0	1	3.0	3	11.0	2.0	0.0	0.0	92.0
1	2	5.0	3	8.0	2.0	0.0	0.0	92.0
2	3	5.0	3	5.0	2.0	NaN	NaN	92.0
3	4	5.0	3	6.0	1.0	91.0	0.0	92.0
4	5	3.0	3	12.0	2.0	91.0	0.0	92.0
...	...	...	...	...	...	...	...	...
43394	43395	5.0	3	8.0	2.0	0.0	0.0	92.0
43395	43396	5.0	3	6.0	1.0	91.0	0.0	92.0
43396	43397	5.0	3	7.0	1.0	0.0	0.0	92.0
43397	43398	5.0	1	8.0	2.0	0.0	0.0	92.0
43398	43399	5.0	90	8.0	3.0	0.0	0.0	92.0

Data Collection

Data was collected through standardized forms and follow-up phone surveys.

Data Cleaning

First, we will explore general features in categories using common knowledge. Then we can futher analayze specific questions with more features.

According to do documentation, I divided the dataset into subgroups with features that attribute to pre-condition, incidence, post-condition, and intervention. Pre-condition is any feature sets that describe about the patient irrelevant to the injury. For example, `Gender` and `Race` are attributes about the patient regardless of the injury. Incidence are variables that describe the injury - incidence that led to this analysis. Similarly, post-condition describes about the condition of patient after the injury. For example, `Seiz`, `Vomit`, `SFxPalp` describes wheter the patient had any post-traumatic seizure, vomit, or any palapble skull fractures after incident. Lastly, intervention is set of features that are written by ED. I grouped these features due to the subjective nature of diagnosis and the ability to self-express.

Pre-Condition (A description of patients before the injury):

- AgeTwoPlus
- Gender
- Ethnicity
- Race
- Drugs

Incidence (Relating to injury):

- InjuryMech
- High_impact_InjSev

Post-Condition (A description of patients after the injury):

- Amnesia_verb
- LOCSeparate
- LocLen
- Seiz
- SeizOccur
- SeizLen
- Vomit
- VomitNbr
- SFxPalp
- FontBulg
- SFxBas
- SFxBasHem

Intervention (Due to the subjective nature of ED and ability to self-express, it is necessary to compare differences between preverbal and verbal):

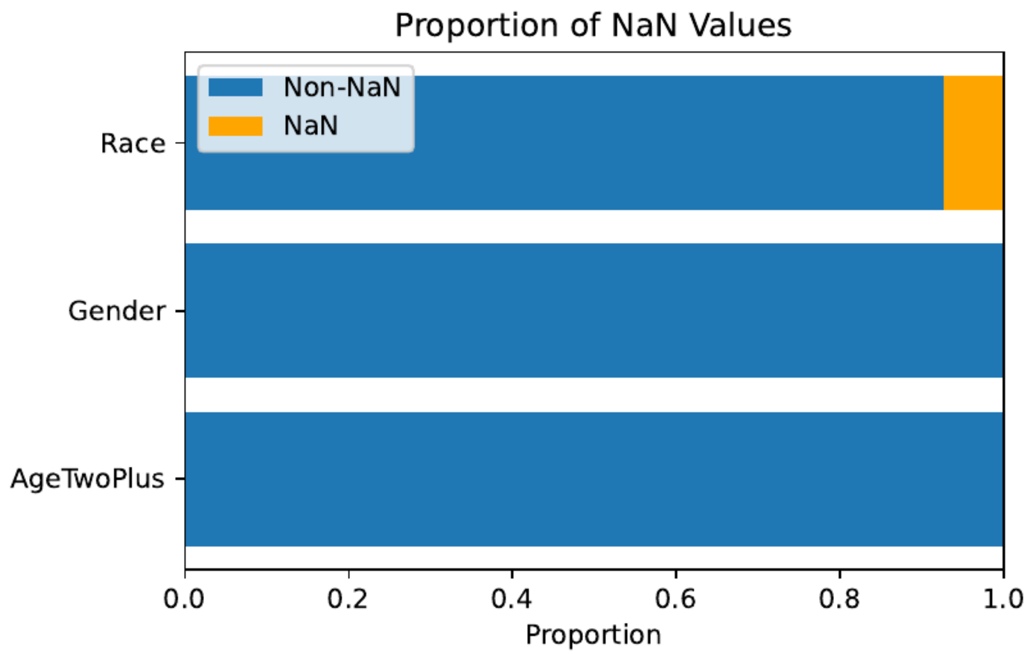
- ActNorm
- HA_verb
- HASeverity
- Intubated
- Paralyzed
- Sedated

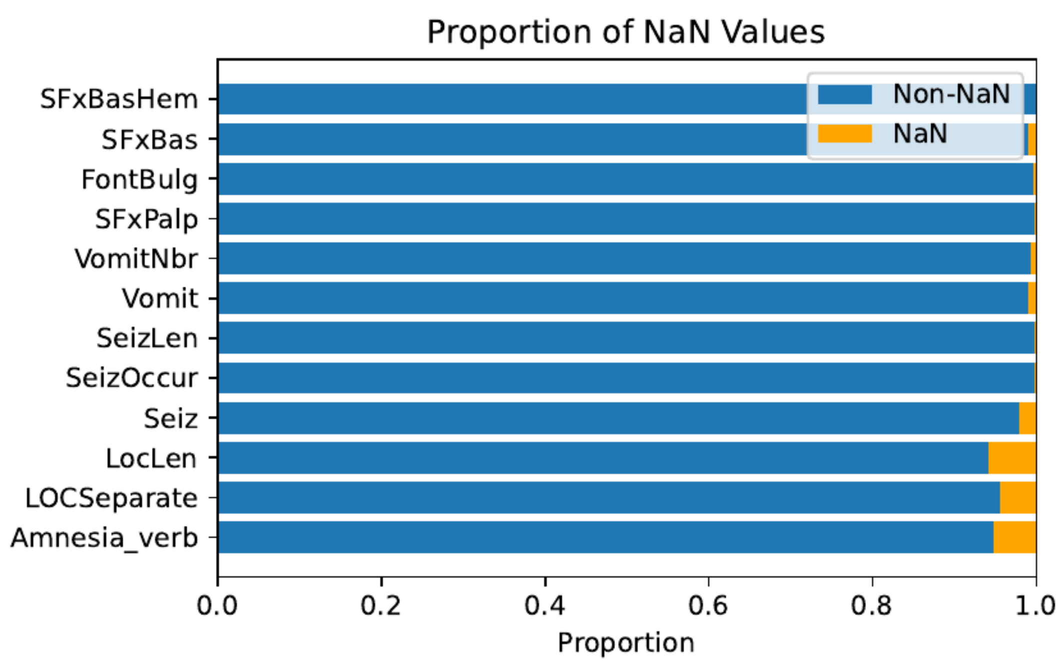
Two Groups (Major Head Trauma vs Minor Head Trauma):

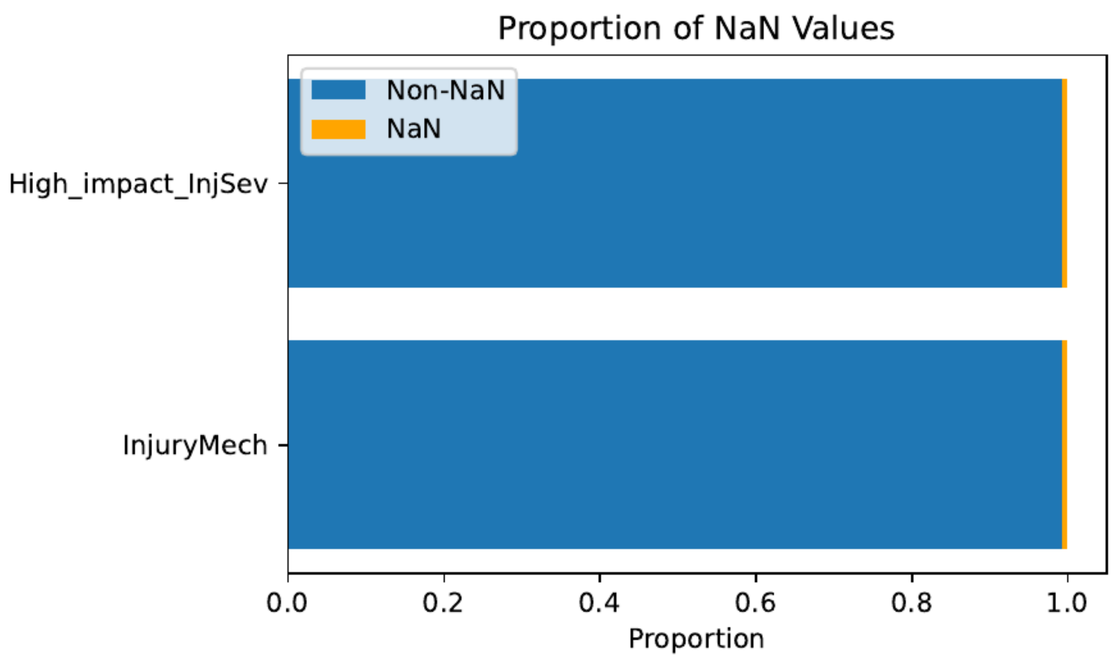
- GCSGroup

Label to determine the outcome or group by

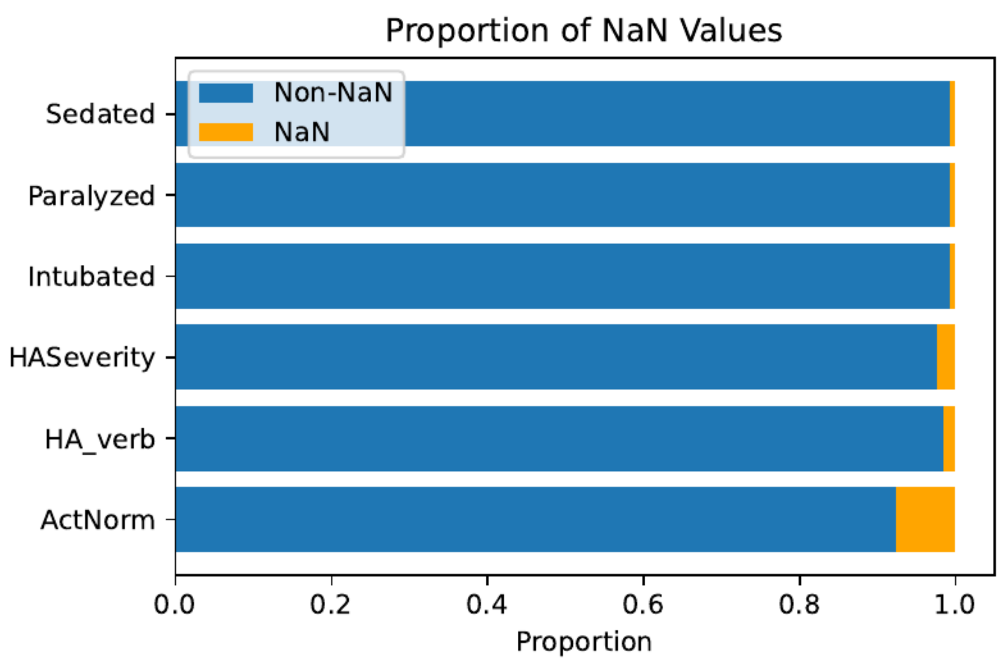
- PosIntFinal



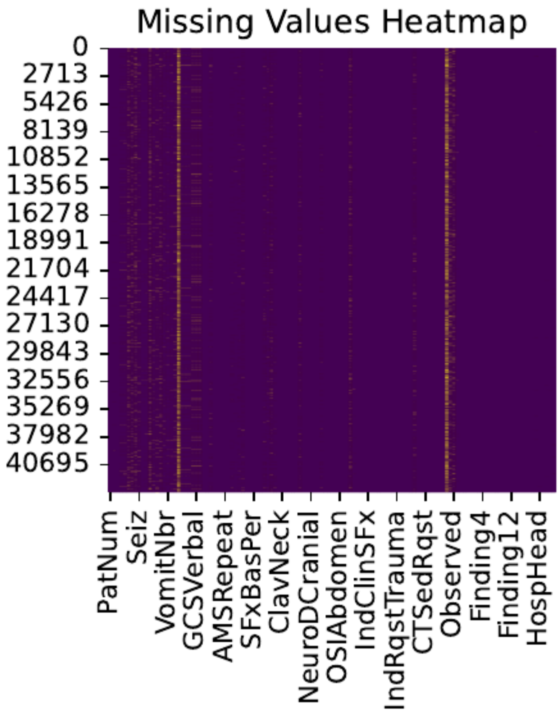




Proportion of NaN Values
Non-NaN
NaN
High_impact_InjSev
InjuryMech
0.0
0.2
0.4
0.6
0.8
1.0
Proportion



Proportion of NaN Values
Non-NaN
NaN
Sedated
Paralyzed
Intubated
HASeverity
HA_verb
ActNorm
0.0
0.2
0.4
0.6
0.8
1.0
Proportion



No columns had any missing values greater than 50% of the values, so instead of removing the whole feature, I decided to replace it using the mode of the feature (most common value). Additionally, I've included correlation values with the `PosIntFinal` which is binary value indicating whether the patient was diagnosed with ciTBIs. From the data report, clinically-important TBI was defined as having at least one of the following: (1) neurosurgical procedure performed, (2) intubated > 24 hours for head trauma, (3) death due to TBI or in the ED, (4) hospitalized for >= 2 nights due to head injury and having a TBI on CT.

We can see that top 4 correlated values are feature that are known as a result of lab. I've selected subset of features using judgment call (described above) to get a correlation values.

```
These are 4 strongly correlated values:
HospHeadPosCT    0.952243
HospHead         0.867533
Neurosurgery     0.508631
GCSTotal        -0.519716
Name: PosIntFinal, dtype: float64
```

```
These are 23 correlated values in order:
Intubated             0.392128
Sedated               0.295219
Paralyzed             0.270167
SFxBas                0.222757
SFxPalp               0.157692
LOCSeparate           0.152461
High_impact_InjSev    0.099557
Seiz                  0.086900
HA_verb               0.077288
Vomit                 0.071679
Amnesia_verb          0.071391
FontBulg              0.063583
HASeverity            0.026814
```

```
AgeTwoPlus           0.015662
Gender              -0.001215
Race                -0.003797
InjuryMech          -0.015431
VomitNbr            -0.048993
SeizLen             -0.049878
SeizOccur           -0.062200
LocLen              -0.107847
ActNorm             -0.167916
SFxBasHem           -0.219501
Name: PosIntFinal, dtype: float64
```

```
How much data was reduced after trimming using IQR:  0.7581050254614161
How much data was reduced after trimming using z-score:  0.9077858936841863

How much data was reduced after trimming using IQR:  0.03412521025830084
How much data was reduced after trimming using z-score:  0.272425631927003

How much data was reduced after trimming using IQR:  0.29652756975967187
How much data was reduced after trimming using z-score:  0.29652756975967187

How much data was reduced after trimming using IQR:  0.39546994170372585
How much data was reduced after trimming using z-score:  0.39546994170372585

How much data was reduced after trimming using IQR:  0.16018802276550148
How much data was reduced after trimming using z-score:  0.37523906080785274
```

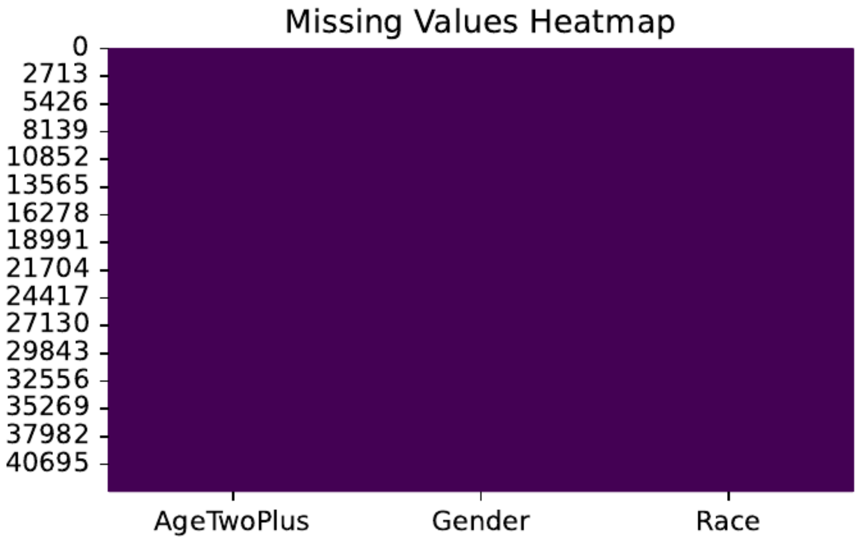
In general, if trimmed dataset is reduced by more that 10%, I will not trim the datasets. For any data sets that can be reduced using trimming, IQR trimming reduces less than z-score for the subgroups we've chose so I will use these set of cleaned data sets (replaced with mode and trimmed outlier with IQR).

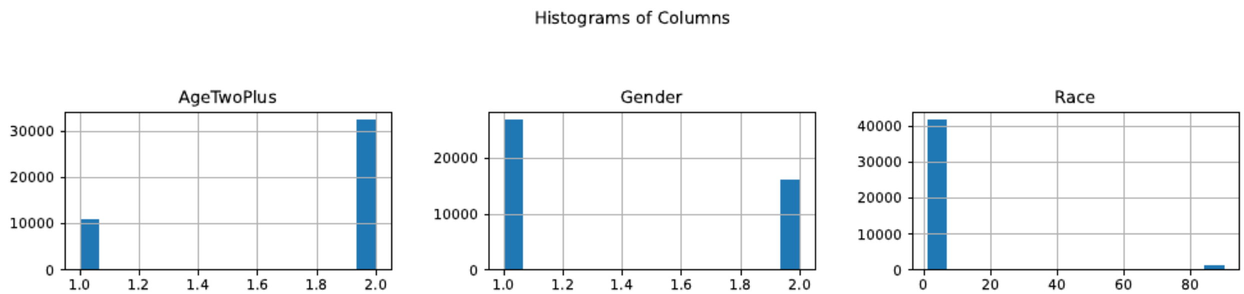
The reason we don't want too much trimming is so that we don't introduce much bias.

Data Exploration

```
Summary Statistics (Before Cleaning):
```

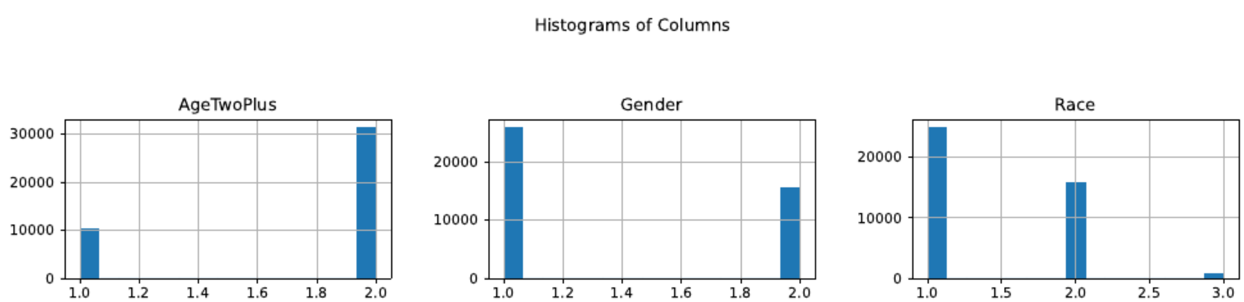
	count	mean	std	min	25%	50%	75%	max
AgeTwoPlus	43399.0	1.748750	0.433737	1.0	1.0	2.0	2.0	2.0
Gender	43399.0	1.376529	0.484521	1.0	1.0	1.0	2.0	2.0
Race	43399.0	4.165718	15.325210	1.0	1.0	1.0	2.0	90.0

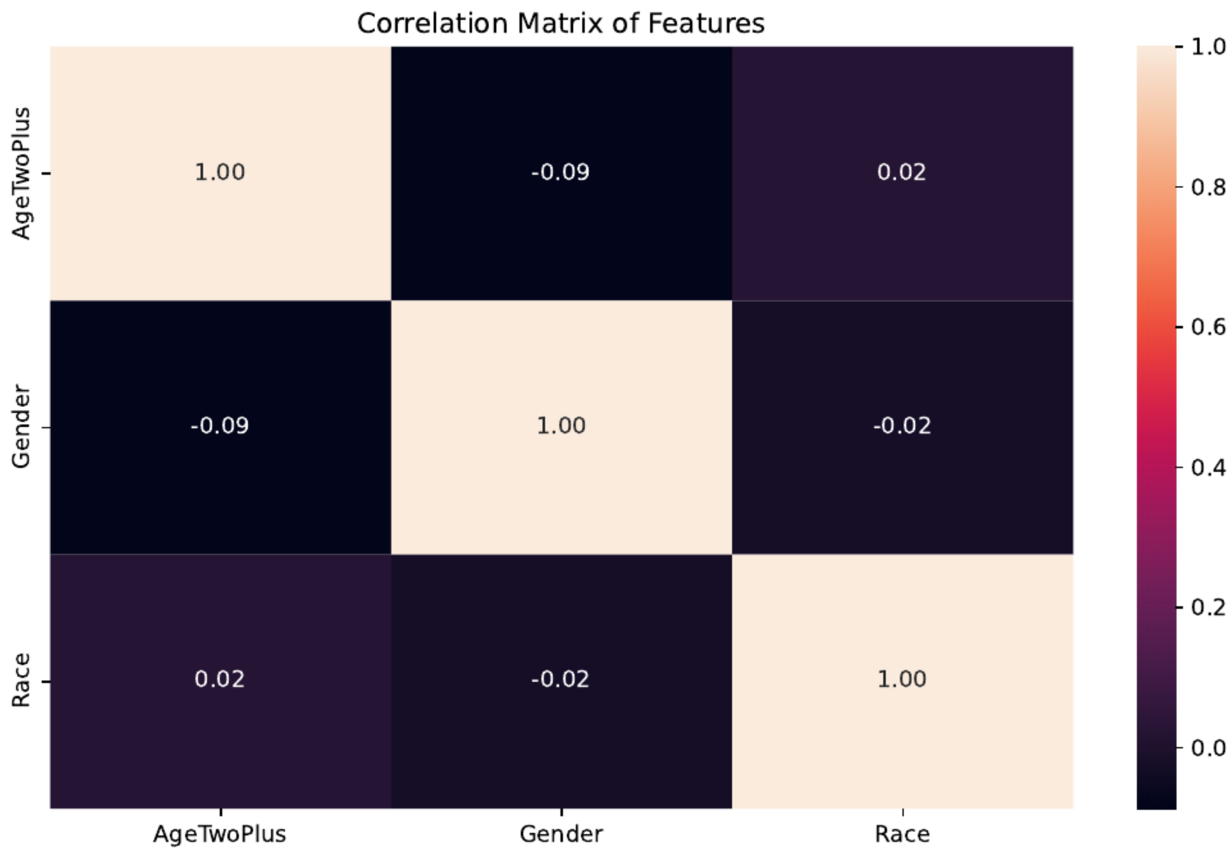




Summary Statistics (After Cleaning):

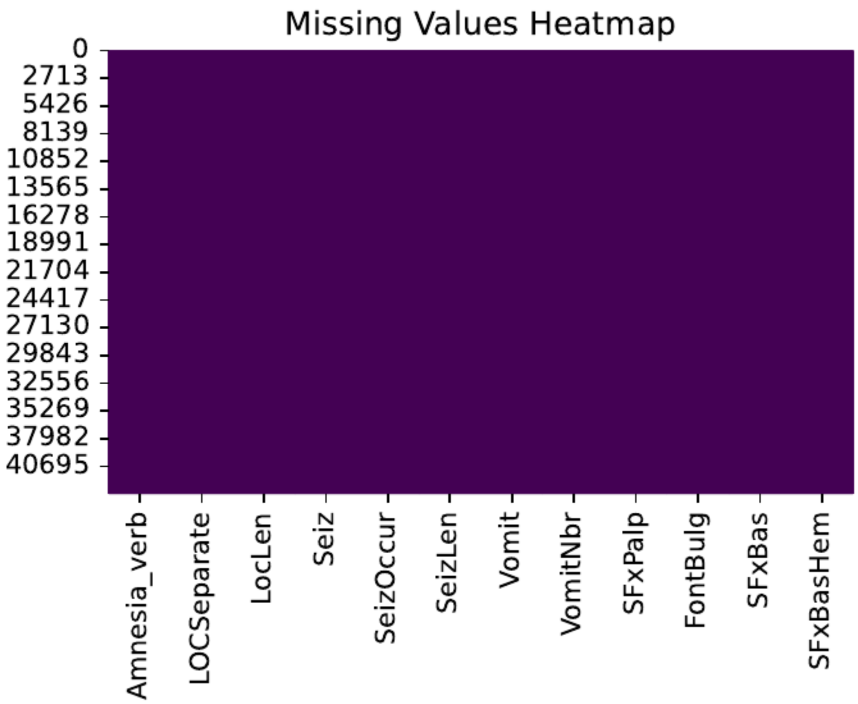
	count	mean	std	min	25%	50%	75%	max
AgeTwoPlus	41918.0	1.750775	0.432569	1.0	2.0	2.0	2.0	2.0
Gender	41918.0	1.376235	0.484446	1.0	1.0	1.0	2.0	2.0
Race	41918.0	1.422587	0.533942	1.0	1.0	1.0	2.0	3.0

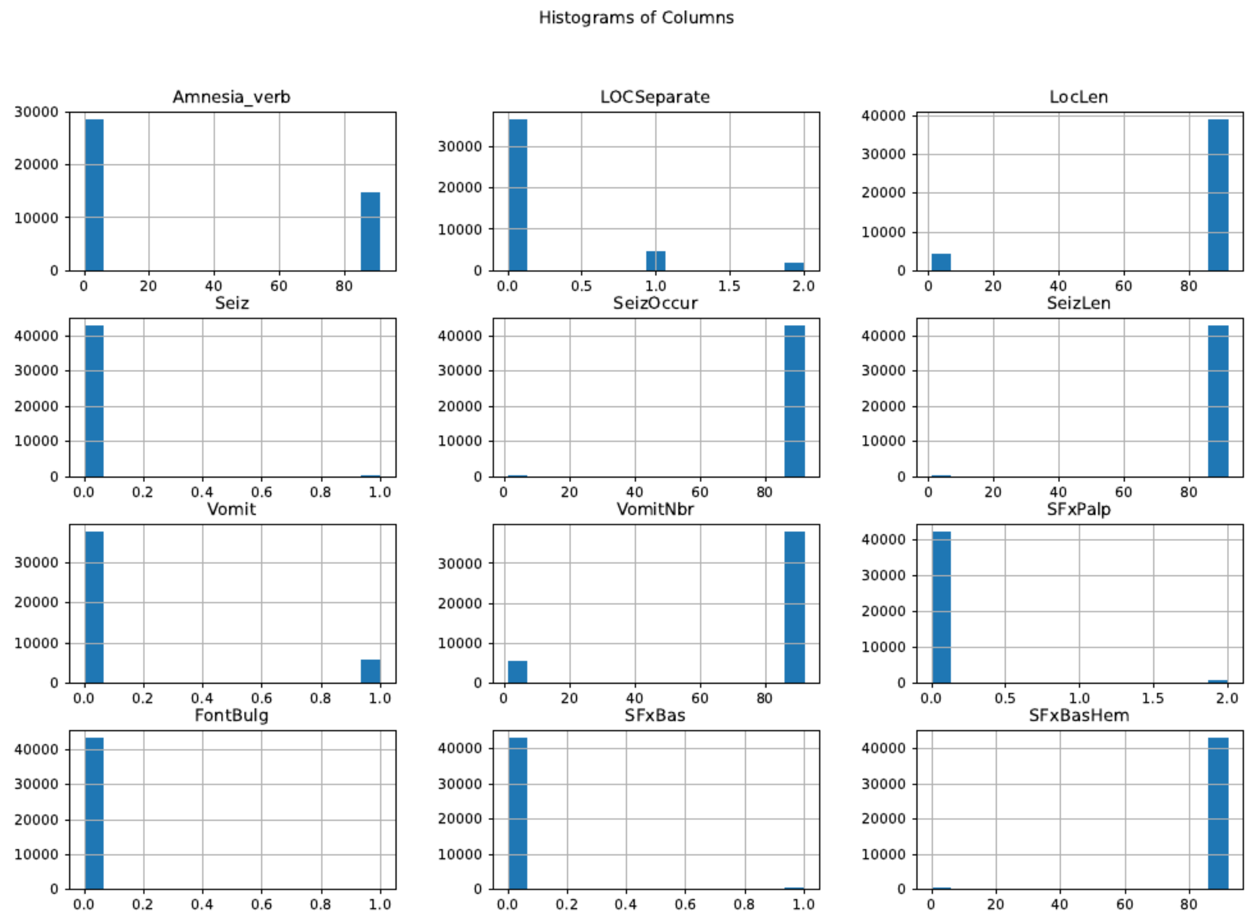




`Summary Statistics (Before Cleaning):`

	count	mean	std	min	25%	50%	75%	max
Amnesia_verb	43399.0	31.148229	43.068433	0.0	0.0	0.0	91.0	91.0
LOCSeparate	43399.0	0.204014	0.505093	0.0	0.0	0.0	0.0	2.0
LocLen	43399.0	83.133413	26.786687	1.0	92.0	92.0	92.0	92.0
Seiz	43399.0	0.013894	0.117054	0.0	0.0	0.0	0.0	1.0
SeizOccur	43399.0	90.892509	9.941777	1.0	92.0	92.0	92.0	92.0
SeizLen	43399.0	90.987673	9.513010	1.0	92.0	92.0	92.0	92.0
Vomit	43399.0	0.133736	0.340372	0.0	0.0	0.0	0.0	1.0
VomitNbr	43399.0	80.579322	29.959698	1.0	92.0	92.0	92.0	92.0
SFxPalp	43399.0	0.050185	0.304455	0.0	0.0	0.0	0.0	2.0
FontBulg	43399.0	0.000830	0.028790	0.0	0.0	0.0	0.0	1.0
SFxBas	43399.0	0.009125	0.095087	0.0	0.0	0.0	0.0	1.0
SFxBasHem	43399.0	91.165419	8.697251	0.0	92.0	92.0	92.0	92.0

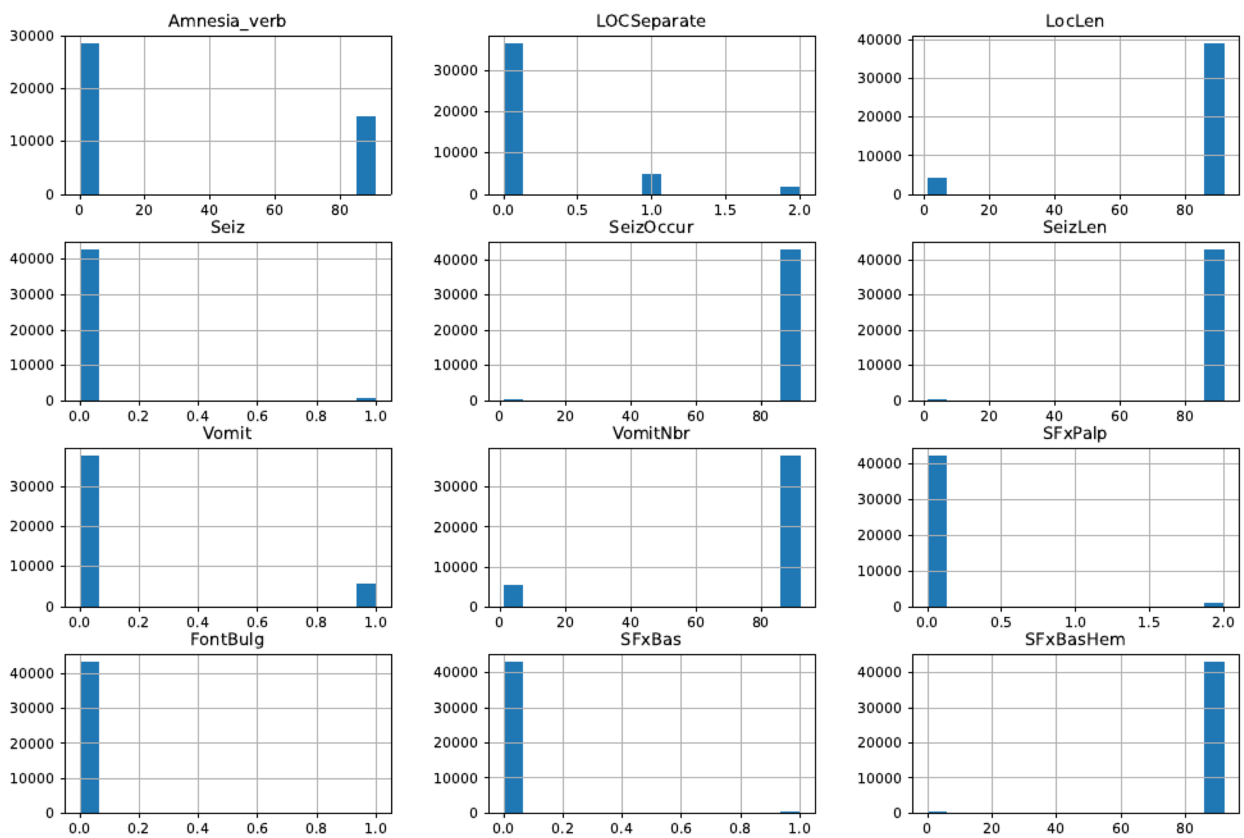


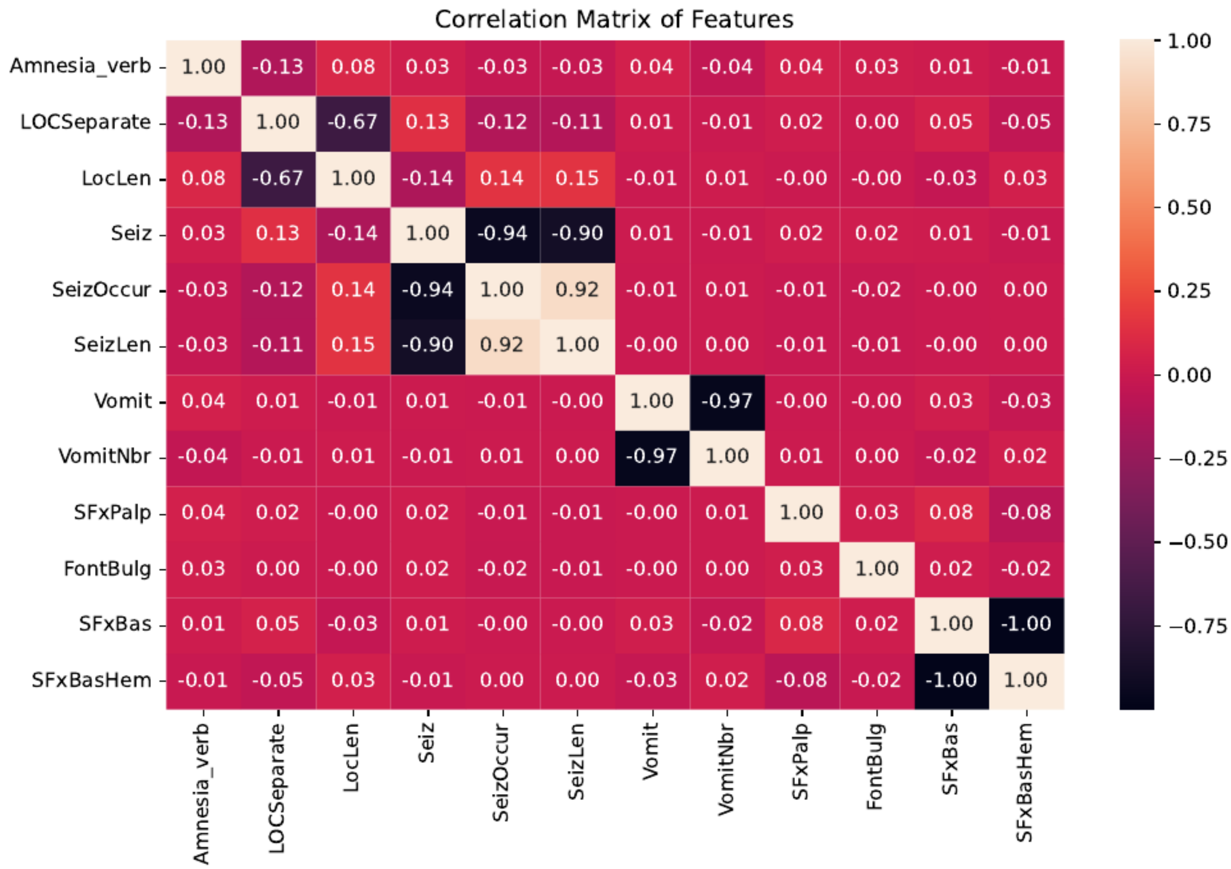


Summary Statistics (After Cleaning):

	count	mean	std	min	25%	50%	75%	max
Amnesia_verb	43399.0	31.148229	43.068433	0.0	0.0	0.0	91.0	91.0
LOCSeparate	43399.0	0.204014	0.505093	0.0	0.0	0.0	0.0	2.0
LocLen	43399.0	83.133413	26.786687	1.0	92.0	92.0	92.0	92.0
Seiz	43399.0	0.013894	0.117054	0.0	0.0	0.0	0.0	1.0
SeizOccur	43399.0	90.892509	9.941777	1.0	92.0	92.0	92.0	92.0
SeizLen	43399.0	90.987673	9.513010	1.0	92.0	92.0	92.0	92.0
Vomit	43399.0	0.133736	0.340372	0.0	0.0	0.0	0.0	1.0
VomitNbr	43399.0	80.579322	29.959698	1.0	92.0	92.0	92.0	92.0
SFxPalp	43399.0	0.050185	0.304455	0.0	0.0	0.0	0.0	2.0
FontBulg	43399.0	0.000830	0.028790	0.0	0.0	0.0	0.0	1.0
SFxBas	43399.0	0.009125	0.095087	0.0	0.0	0.0	0.0	1.0
SFxBasHem	43399.0	91.165419	8.697251	0.0	92.0	92.0	92.0	92.0

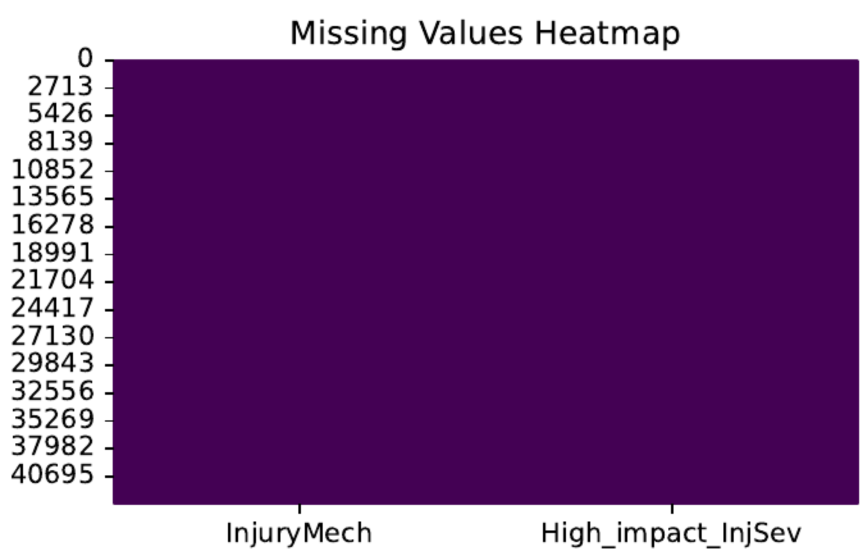
Histograms of Columns

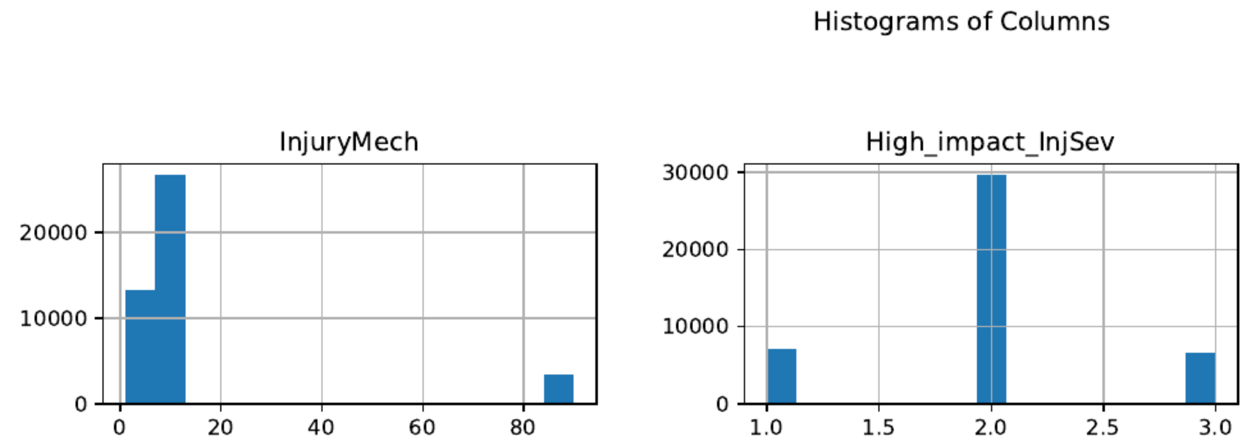




Summary Statistics (Before Cleaning):

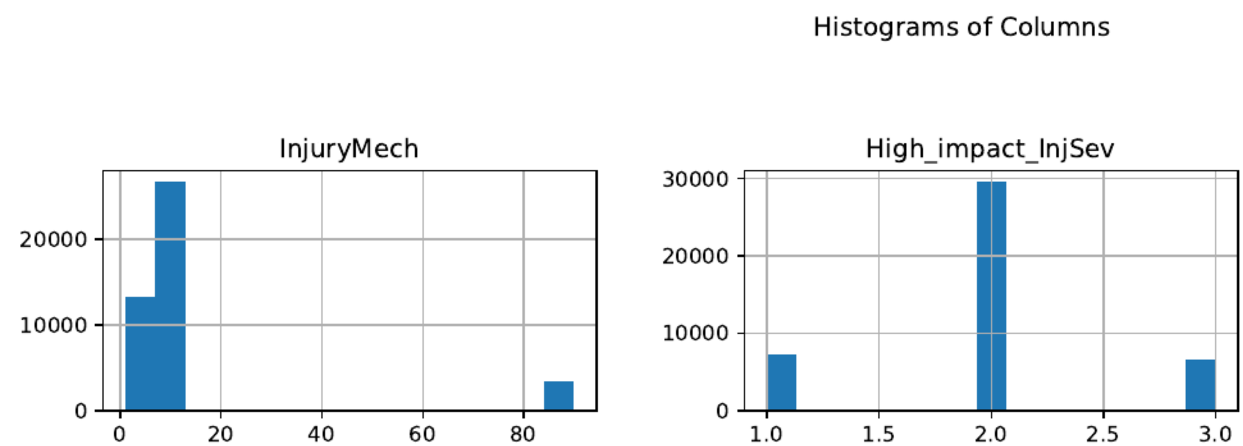
	count	mean	std	min	25%	50%	75%	max
InjuryMech	43399.0	13.864190	22.623717	1.0	6.0	8.0	10.0	90.0
High_impact_InjSev	43399.0	1.986659	0.563684	1.0	2.0	2.0	2.0	3.0

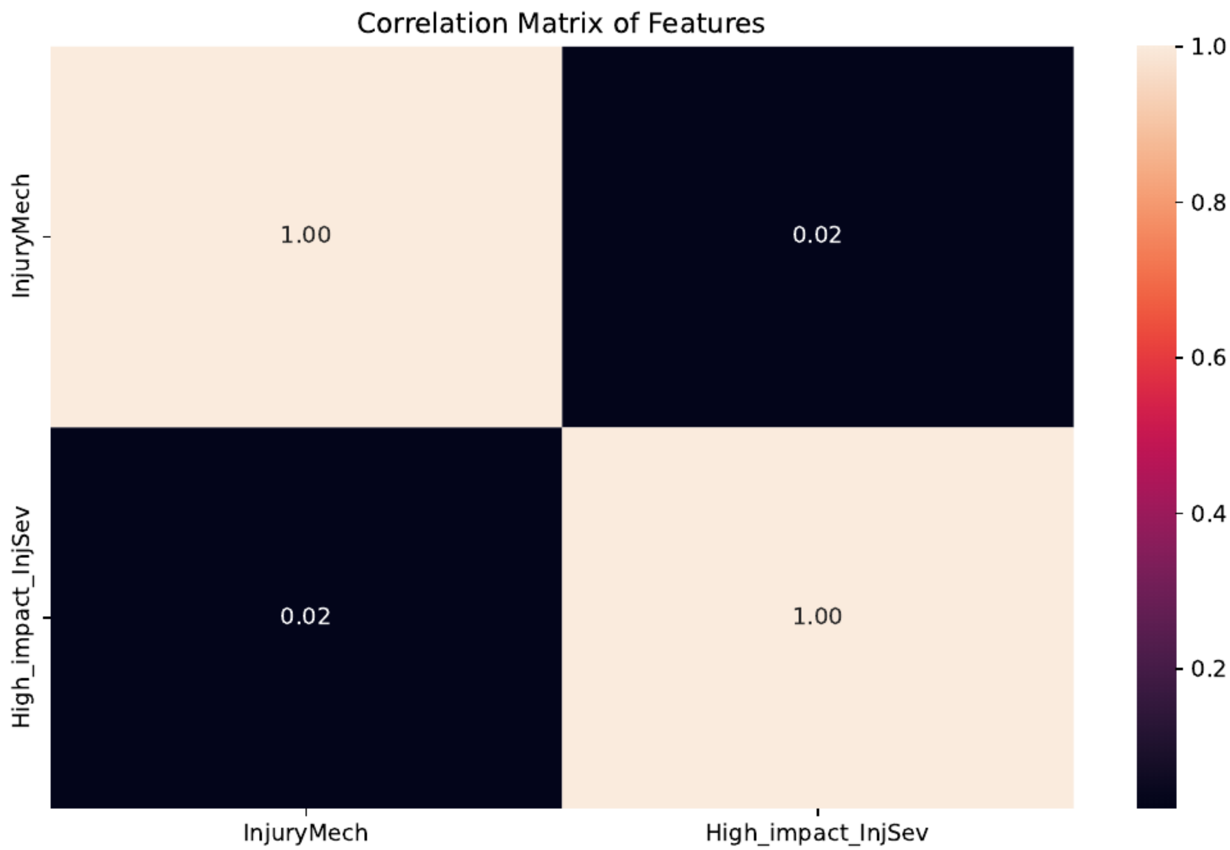




Summary Statistics (After Cleaning):

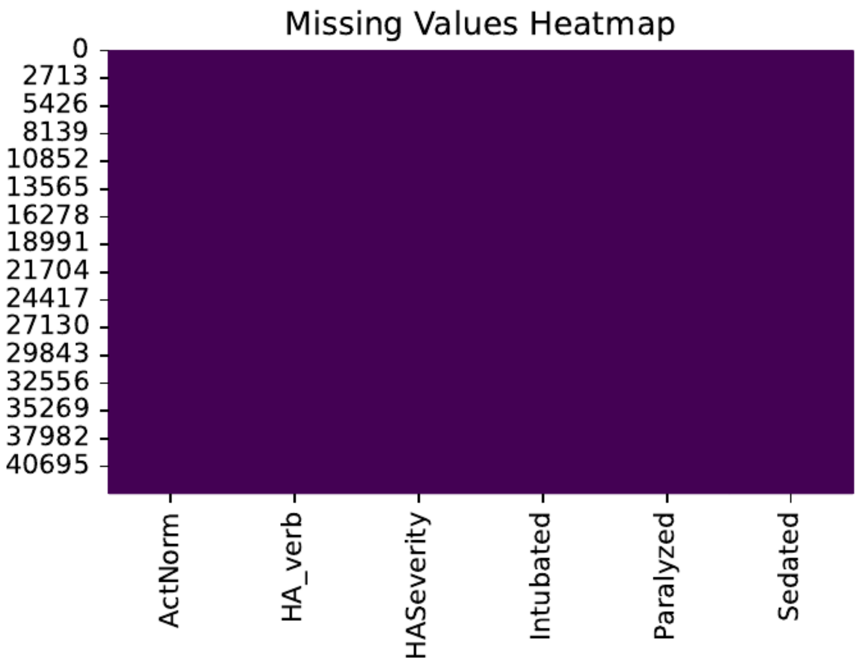
	count	mean	std	min	25%	50%	75%	max
InjuryMech	43399.0	13.864190	22.623717	1.0	6.0	8.0	10.0	90.0
High_impact_InjSev	43399.0	1.986659	0.563684	1.0	2.0	2.0	2.0	3.0

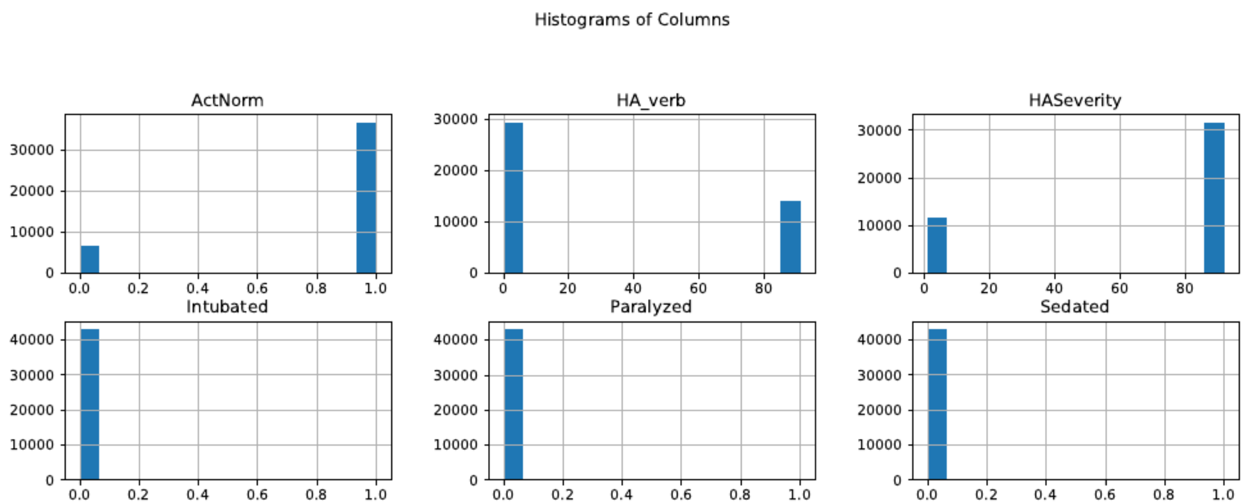




`Summary Statistics (Before Cleaning):`

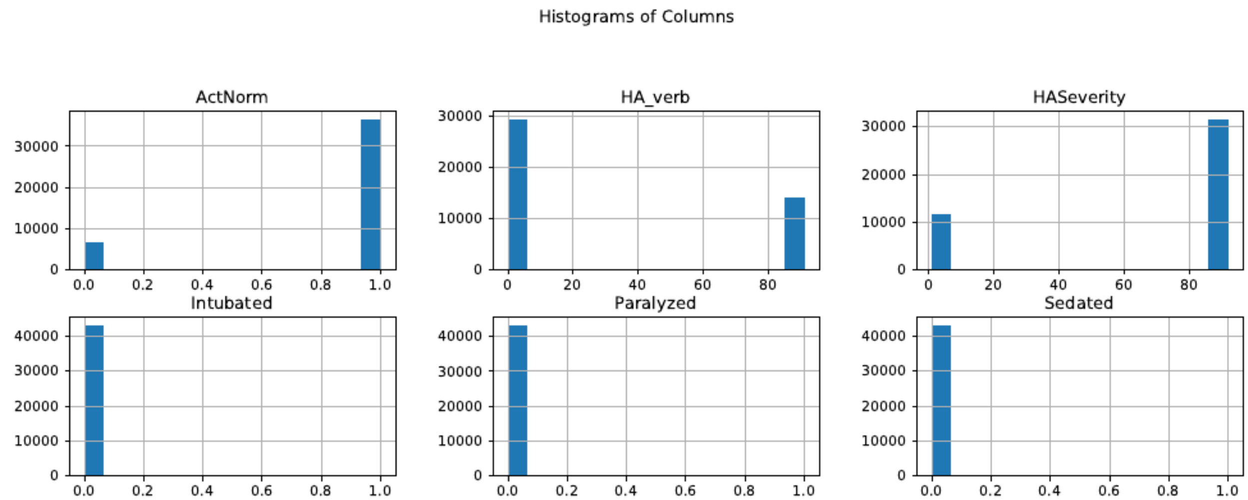
	count	mean	std	min	25%	50%	75%	max
ActNorm	43399.0	0.843430	0.363399	0.0	1.0	1.0	1.0	1.0
HA_verb	43399.0	29.776400	42.385174	0.0	0.0	1.0	91.0	91.0
HASeverity	43399.0	67.505864	40.173334	1.0	2.0	92.0	92.0	92.0
Intubated	43399.0	0.005023	0.070697	0.0	0.0	0.0	0.0	1.0
Paralyzed	43399.0	0.003134	0.055892	0.0	0.0	0.0	0.0	1.0
Sedated	43399.0	0.004931	0.070048	0.0	0.0	0.0	0.0	1.0

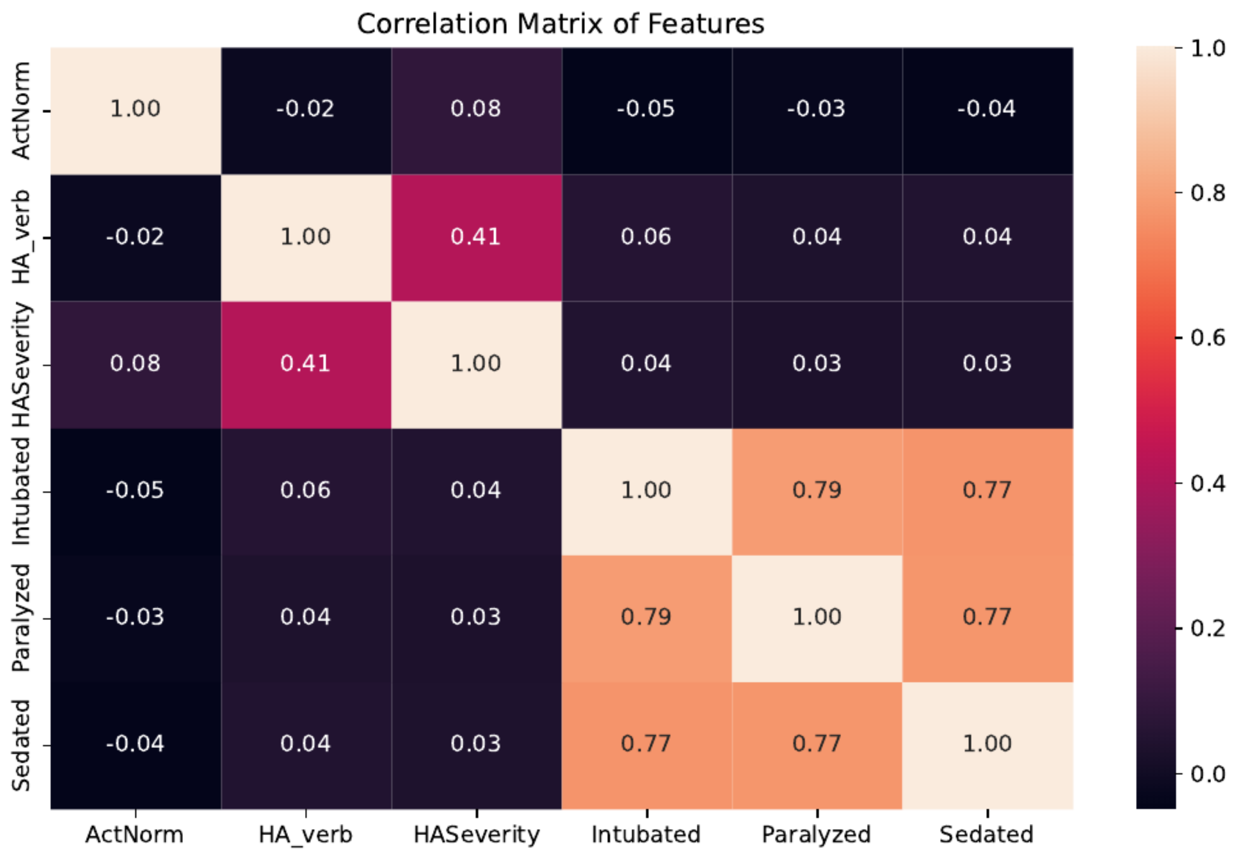




Summary Statistics (After Cleaning):

	count	mean	std	min	25%	50%	75%	max
ActNorm	43399.0	0.843430	0.363399	0.0	1.0	1.0	1.0	1.0
HA_verb	43399.0	29.776400	42.385174	0.0	0.0	1.0	91.0	91.0
HASeverity	43399.0	67.505864	40.173334	1.0	2.0	92.0	92.0	92.0
Intubated	43399.0	0.005023	0.070697	0.0	0.0	0.0	0.0	1.0
Paralyzed	43399.0	0.003134	0.055892	0.0	0.0	0.0	0.0	1.0
Sedated	43399.0	0.004931	0.070048	0.0	0.0	0.0	0.0	1.0

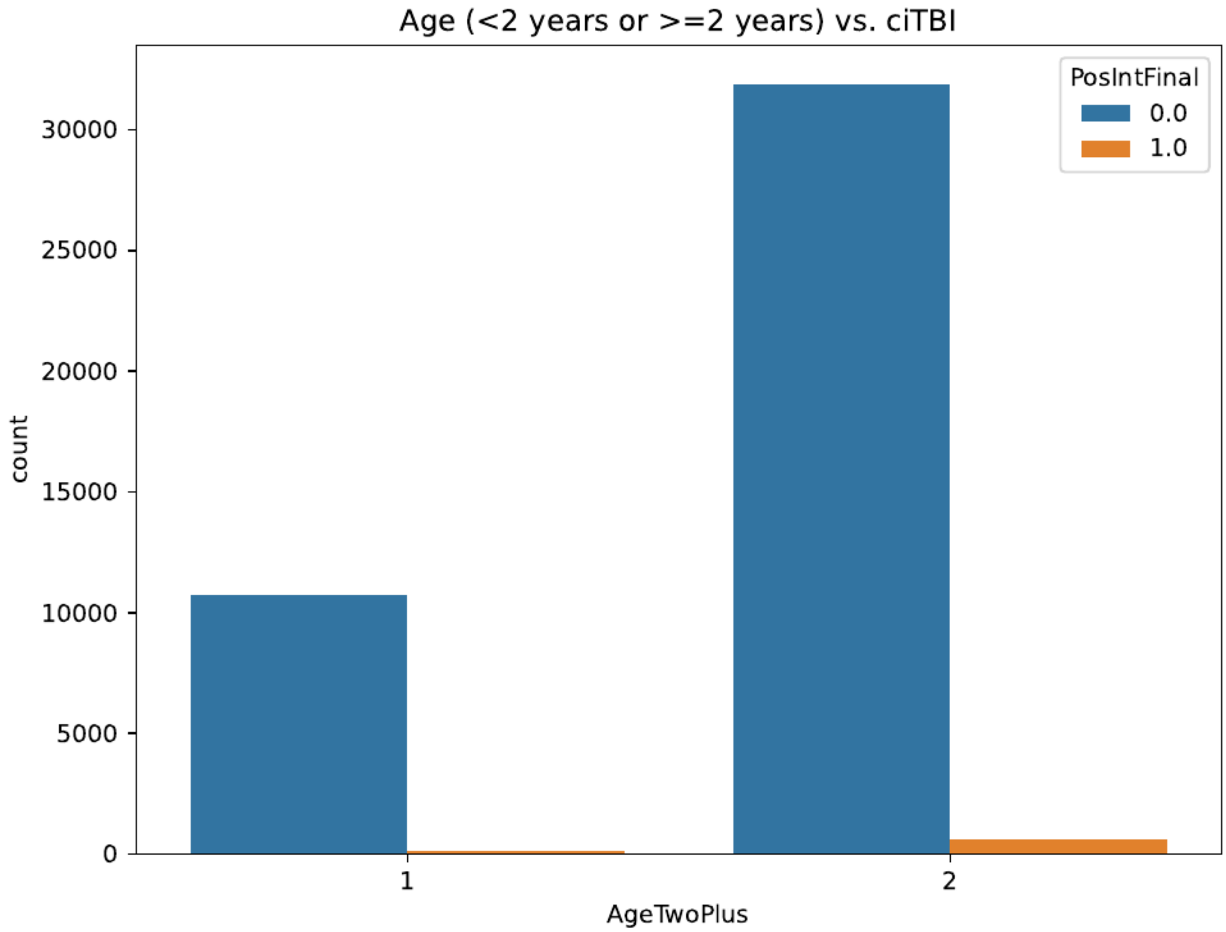


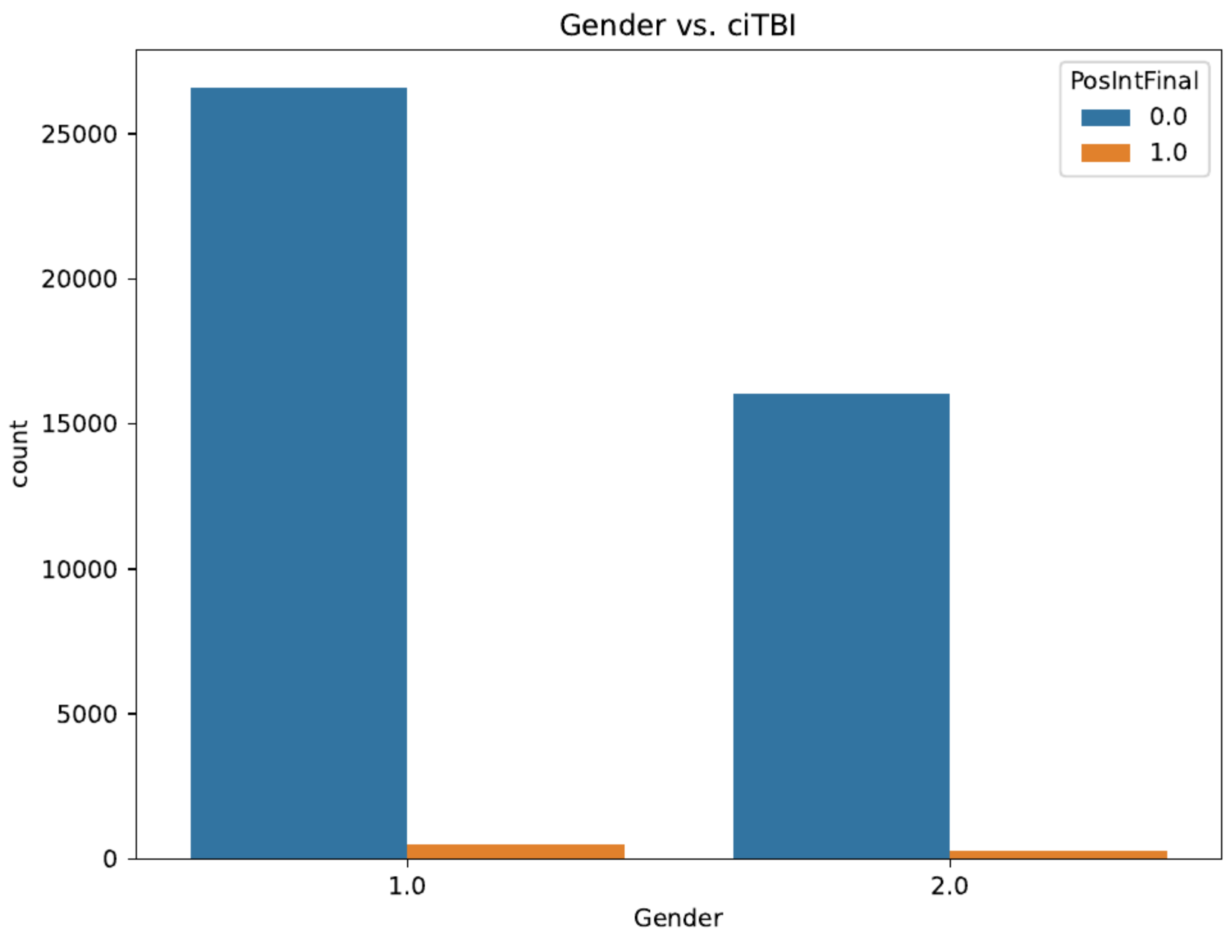


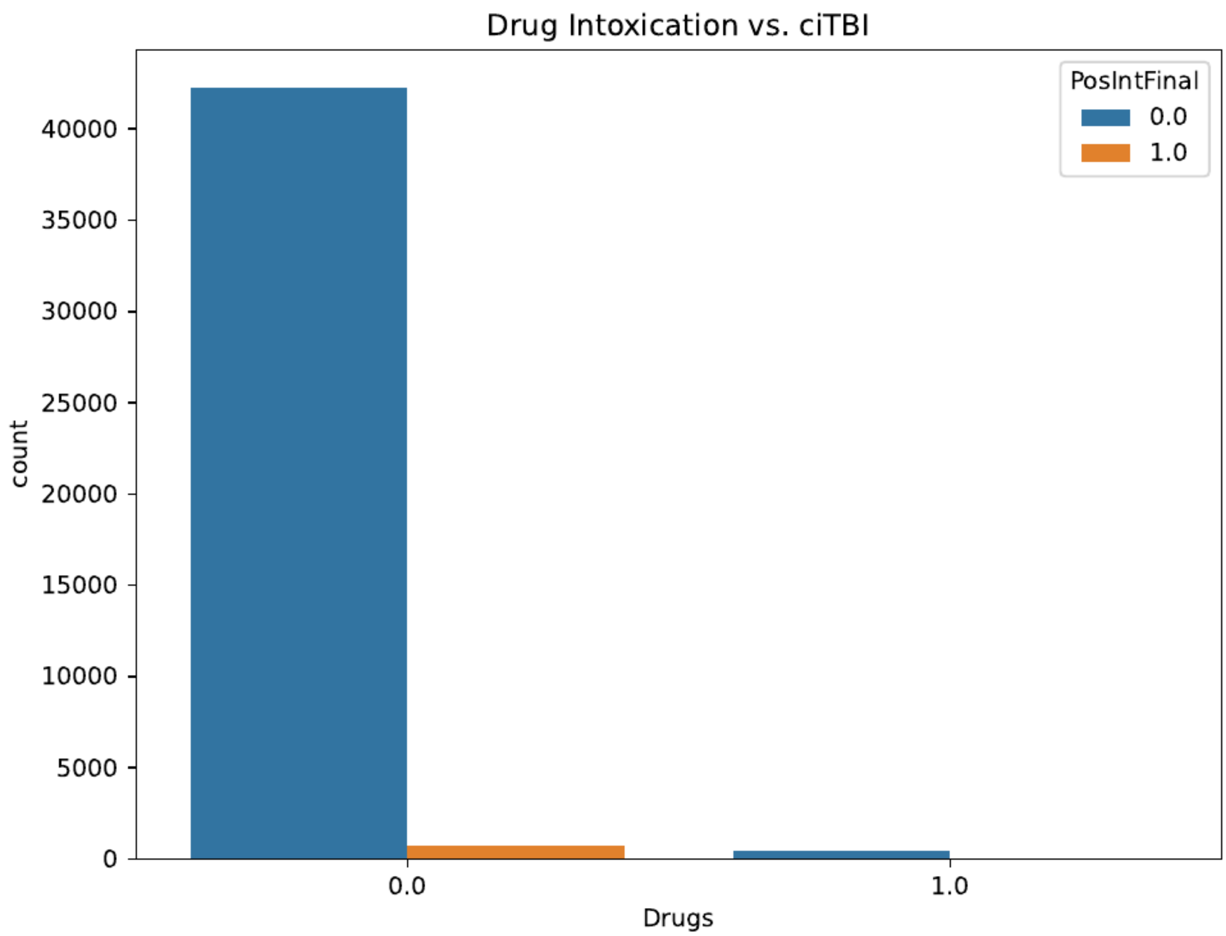
Findings

First finding

Can we predict ciTBI (`PosIntFinal`) based on injury mechanism and injury severity? A chi-squared test is a statistical hypothesis test used in the analysis of contingency tables. I used chi-square contignece function to compute the chi-square statistic and p-value for the hypothesis test of independence of the observed frequencies in the contingency table observed. If the p-value is less than 0.05, we say that it is statistically significant and can asumme that the two observed frequencies are not independet. Higher values indicate that there are more relationships.




Gender vs. ciTBI
PosIntFinal
0.0
1.0
25000
20000
15000
10000
5000
0
count
1.0
2.0
Gender




```
Chi-square test for AgeTwoPlus vs ciTBI: chi2 = 10.350189010132969, p-value = 0.0012946142872781068
Chi-square test for Gender vs ciTBI: chi2 = 0.04429724363595672, p-value = 0.8333015632206818
Chi-square test for Drugs vs ciTBI: chi2 = 43.030281799667264, p-value = 5.389911598157804e-11
```

Thus, according to this test, we found that for pre-condition, use of drug is more correlated to ciTBIs.

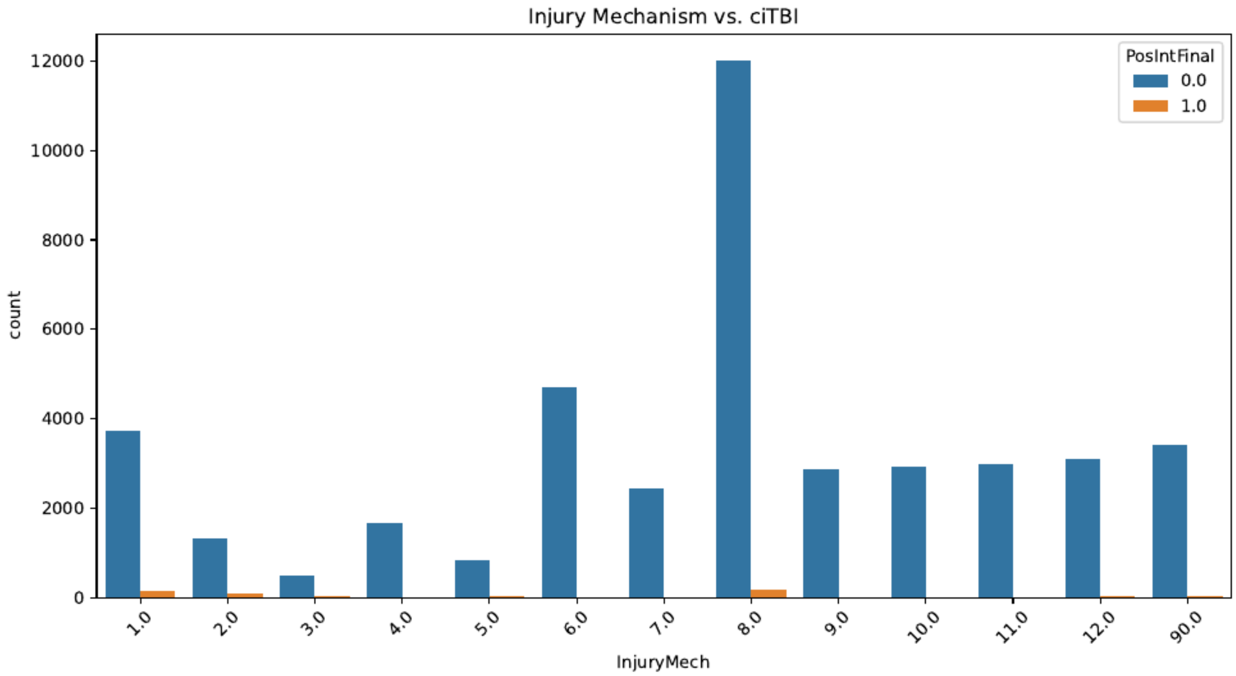
Second finding

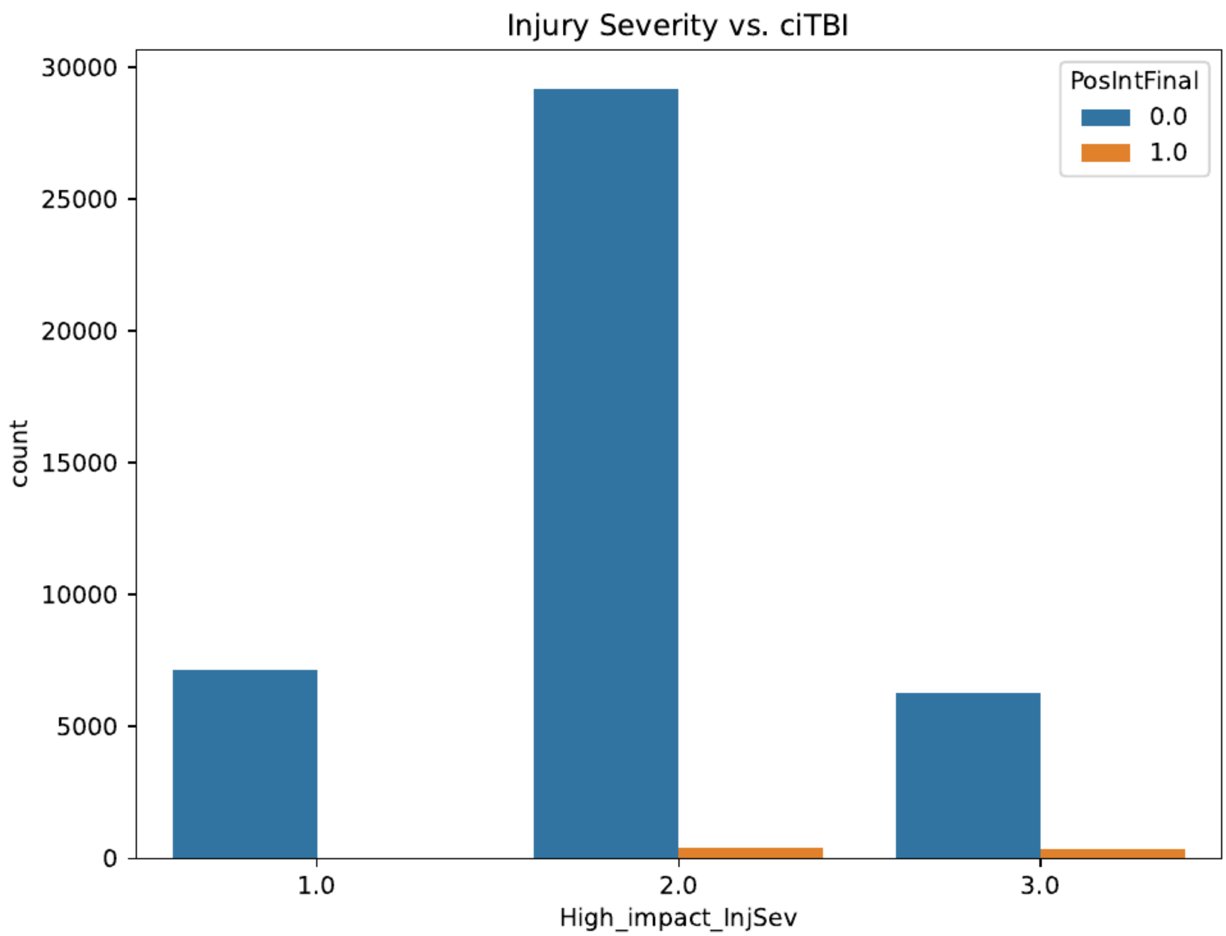
Which inury mechanism and severity correlate to higher chance of ciTBIs?

PosIntFinal InjuryMech	0.0	1.0
1.0	3747	163
2.0	1326	107
3.0	518	38
4.0	1671	30
5.0	853	48
6.0	4710	23
7.0	2451	4
8.0	11998	186
9.0	2891	17
10.0	2950	29
11.0	2991	25

PosIntFinal	0.0	1.0
InjuryMech		
12.0	3119	39
90.0	3411	54

PosIntFinal	0.0	1.0
High_impact_InjSev		
1.0	7161	27
2.0	29199	403
3.0	6276	333





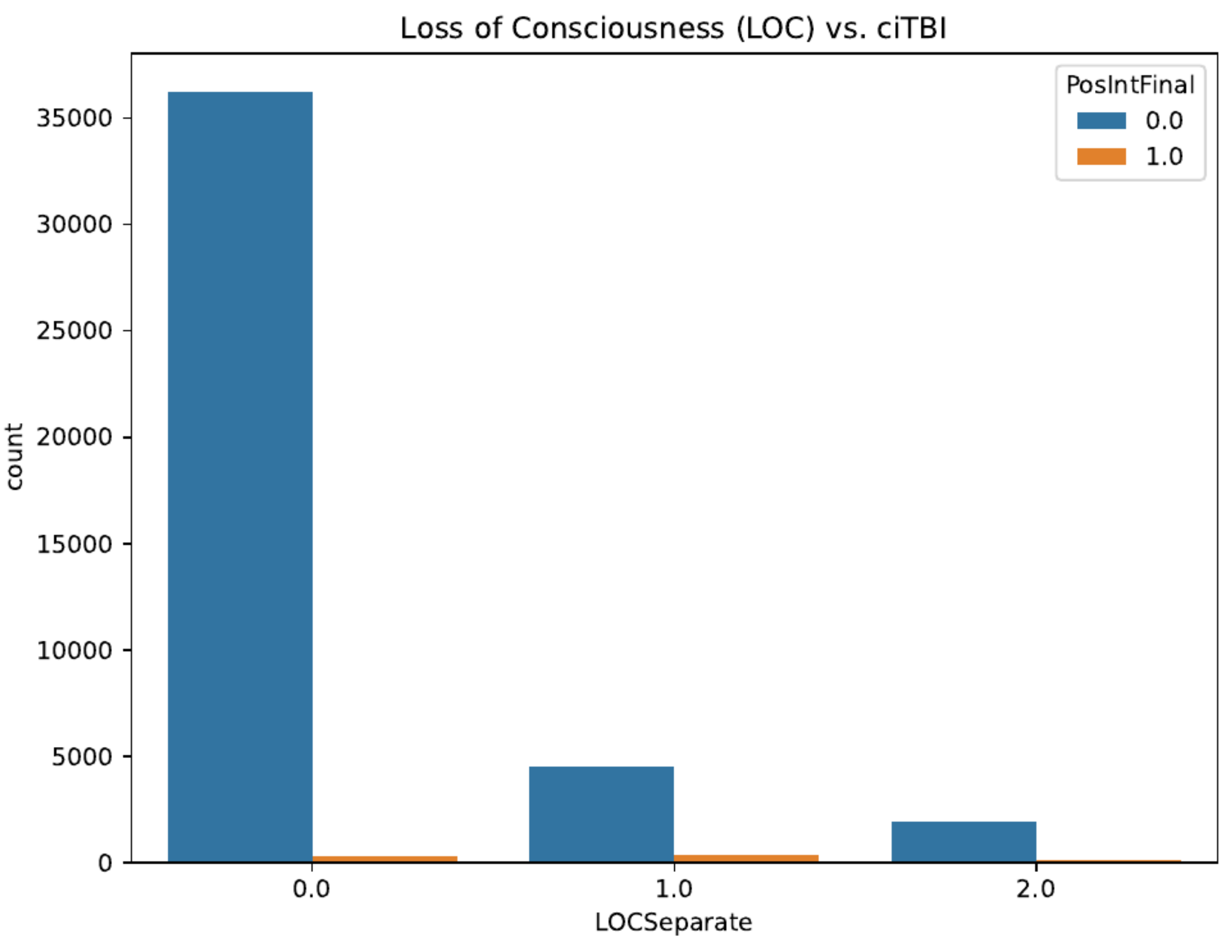

```
InjuryMech
1.0     0.041688
2.0     0.074669
3.0     0.068345
4.0     0.017637
5.0     0.053274
6.0     0.004859
7.0     0.001629
8.0     0.015266
9.0     0.005846
10.0    0.009735
11.0    0.008289
12.0    0.012350
90.0    0.015584
Name: PosIntFinal, dtype: float64
High_impact_InjSev
1.0    0.003756
2.0    0.013614
3.0    0.050386
Name: PosIntFinal, dtype: float64
```

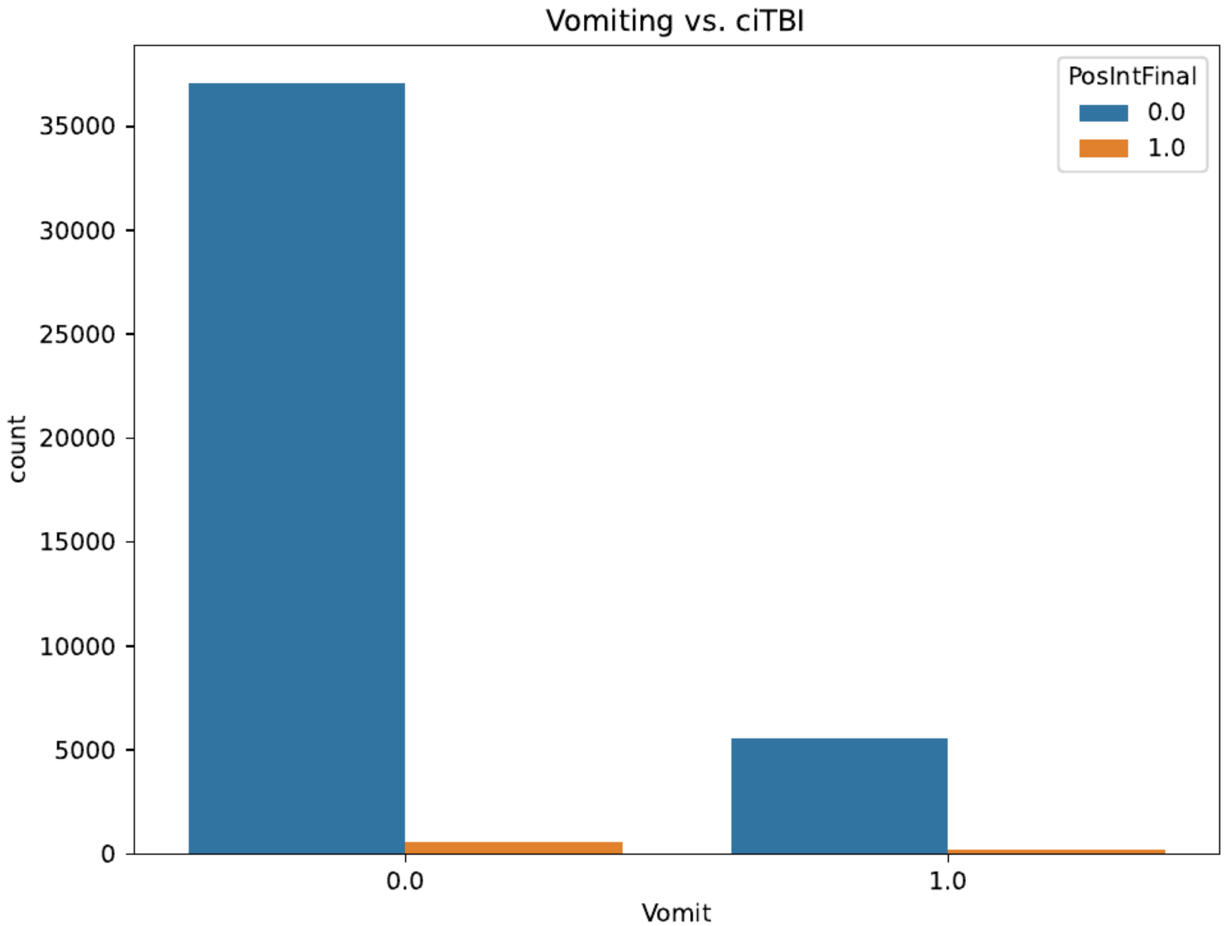
In this finding, inury mechanism 2 (Pedestrian struck by moving vehicle) and severity 3 (High) accounts for most ciTBI patients.

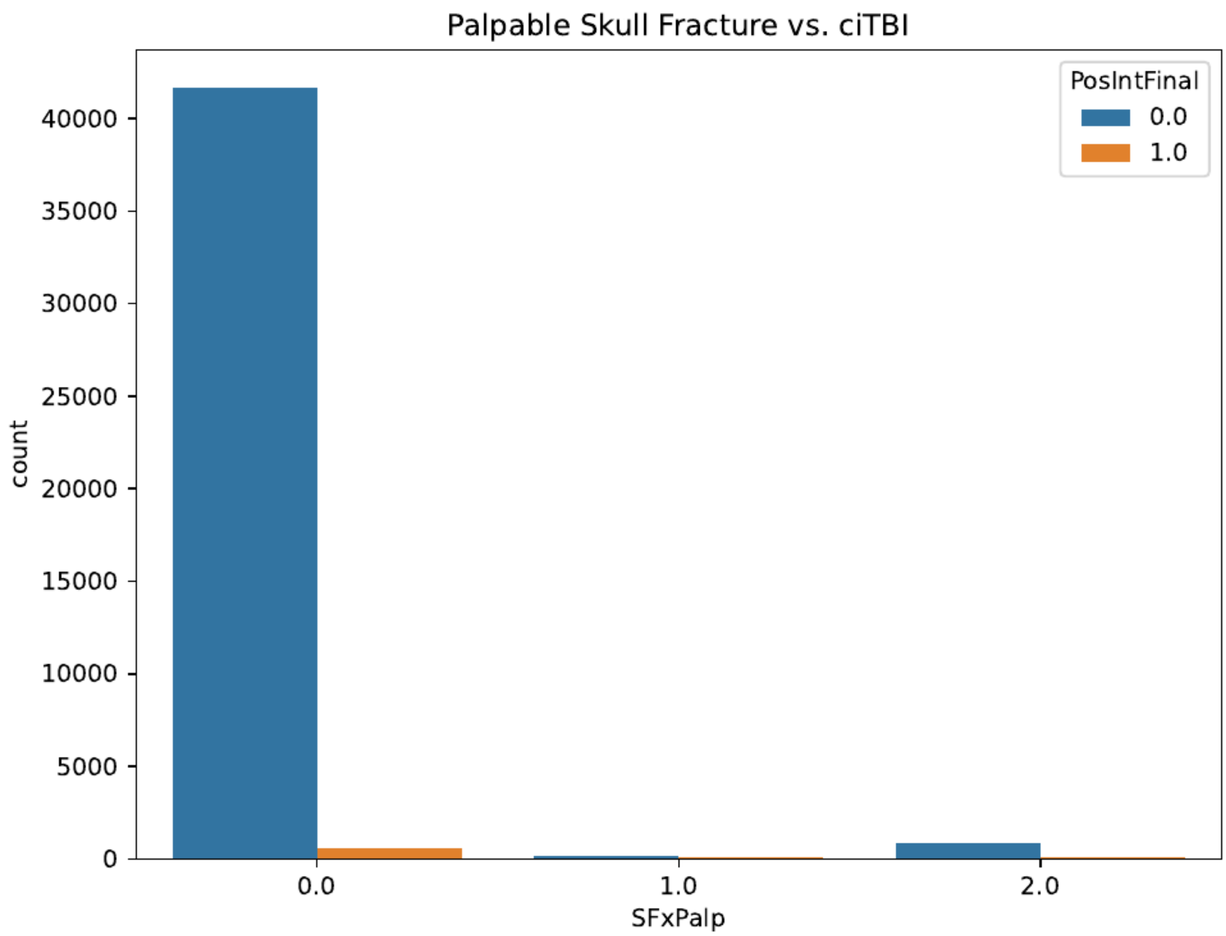
Third finding

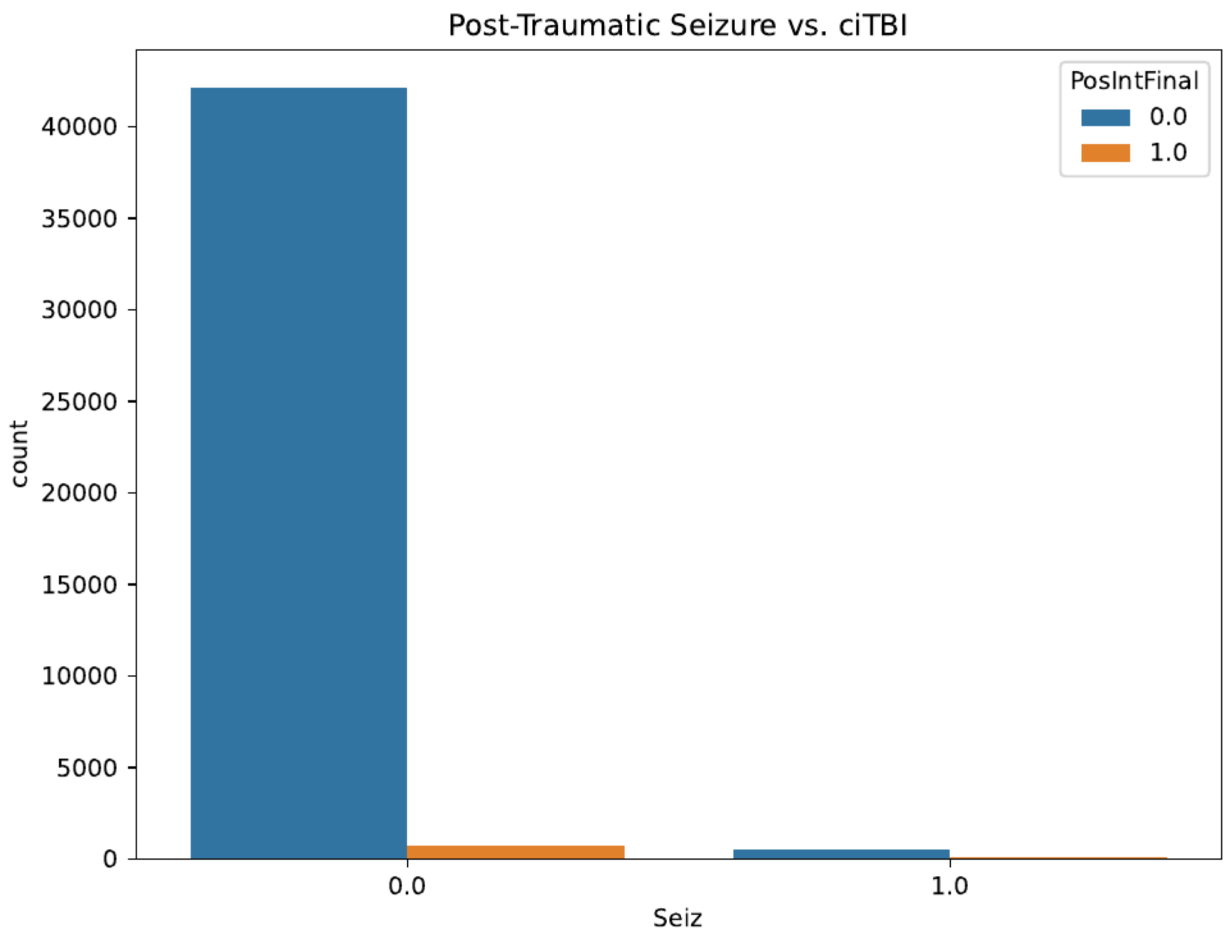
From the post-condition features, can we accurately predict the outcome of ciTBIs? In other words, we want to see what feature sets can accurately discern the potential of having ciTBIs.

	Amnesia_verb	LOCSeparate	LocLen	Seiz	SeizOccur	SeizLen	Vomit	VomitNbr	SFxPalp	FontBulg
0	0.0	0.0	92.0	0.0	92.0	92.0	0.0	92.0	0.0	0.0
1	0.0	0.0	92.0	0.0	92.0	92.0	1.0	3.0	0.0	0.0
2	0.0	0.0	92.0	0.0	92.0	92.0	0.0	92.0	1.0	0.0
3	91.0	0.0	92.0	0.0	92.0	92.0	0.0	92.0	0.0	0.0
4	91.0	0.0	92.0	0.0	92.0	92.0	1.0	1.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...
43394	0.0	0.0	92.0	0.0	92.0	92.0	0.0	92.0	0.0	0.0
43395	91.0	0.0	92.0	0.0	92.0	92.0	0.0	92.0	0.0	0.0
43396	0.0	0.0	92.0	0.0	92.0	92.0	0.0	92.0	0.0	0.0
43397	0.0	0.0	92.0	0.0	92.0	92.0	0.0	92.0	0.0	0.0
43398	0.0	0.0	92.0	0.0	92.0	92.0	0.0	92.0	0.0	0.0




Vomiting vs. ciTBI
PosIntFinal
0.0
1.0
count
35000
30000
25000
20000
15000
10000
5000
0
0.0
1.0
Vomit


Palpable Skull Fracture vs. ciTBI
PosIntFinal
0.0
1.0
count
40000
35000
30000
25000
20000
15000
10000
5000
0
0.0
1.0
2.0
SFxPalp



	precision	recall	f1-score	support
0.0	0.98	1.00	0.99	12779
1.0	0.20	0.00	0.01	241
accuracy			0.98	13020
macro avg	0.59	0.50	0.50	13020
weighted avg	0.97	0.98	0.97	13020

	Coefficient
LOCSeparate	1.088662
Vomit	0.991379
SFxPalp	1.303787
Seiz	1.125431

Reality Check

- Do a reality check. What reality could you compare your cleaned data to?
- Clearly state your assumptions and explain why this reality check is useful.
- Does your cleaned data pass the reality check or are there issues? Discuss.

Stability Check

Take one of your findings and present a perturbed version. How does this affect your finding? Add a before and after plot here.

Discussion

- Did the data size restrict you in any way? Discuss some challenges that you faced as a result of the data size.
- Address the three realms: data / reality, algorithms / models, and future data / reality.
- Where do the parts of the lab fit into those three realms?
- Do you think there is a one-to-one correspondence of the data and reality?
- What about reality and data visualization?

Conclusion

- You should make attempts to connect your findings/analysis back to the domain problem in every section of this report, but here in the conclusion, you can reiterate your main points and provide overarching remarks on the PECARN data as it relates to the domain problem

Academic honesty statement

Please address to Bin.

Collaborators

Bibliography