

**Springboard—DSC Program**  
**Capstone Project 1 Milestone Report**  
**Determining the Likelihood and Type of**  
**Drug Abuse Visits to Emergency Departments in the US**  
**By Laura Eshee**  
**June, 2020**

## **Background**

According to The National Institute on Drug Abuse, “Addiction is a chronic disease characterized by drug seeking and use that is compulsive, or difficult to control, despite harmful consequences. The initial decision to take drugs is voluntary for most people, but repeated drug use can lead to brain changes that challenge an addicted person’s self-control and interfere with their ability to resist intense urges to take drugs. These brain changes can be persistent, which is why drug addiction is considered a "relapsing" disease—people in recovery from drug use disorders are at increased risk for returning to drug use even after years of not taking the drug.”<sup>1</sup>

Drug abuse occurs when a person takes a substance, whether illegal, prescribed or over the counter, for purposes other than those in which they are meant to be used, or when a person takes large quantities of the substance. Typically, the person is using the drug to alter his or her mood or feel better and not for a health reason.

Statistics for drug abuse include:

- Almost 21 million Americans have at least one addiction, yet only 10% of them receive treatment.<sup>2</sup>
- Drug overdose deaths have more than tripled since 1990.<sup>2</sup>
- From 1999 to 2017, more than 700,000 Americans died from overdosing on a drug.<sup>2</sup>
- More than 90% of people who have an addiction started to drink alcohol or use drugs before they were 18 years old.<sup>2</sup>
- Americans between the ages of 18 and 25 are most likely to use addictive drugs.<sup>2</sup>
- Alcohol and drug addiction cost the U.S. economy over \$600 billion every year.<sup>2</sup>
- During 2008–2011, an average of 1.1 million emergency department (ED) visits were made each year for drug poisoning, with a visit rate of 35.4 per 10,000 persons.<sup>3</sup>

- About one-quarter (24.5%) of drug-poisoning ED visits resulted in hospital admission.<sup>3</sup>

## **Client and Problem Statement**

Emergency Departments in hospitals nationwide must be prepared to accept patients that are suffering from drug abuse. They need to plan for enough staffing and supplies to handle the expected types and percentages of drug abuse related Emergency Department visits so that the departments can be prepared with enough staffing, medications, procedures, etc. to properly deal with these cases and have good outcomes.

## **Dataset**

The Drug Abuse Warning Network (DAWN) is a public health surveillance system that monitors drug abuse related visits to emergency departments in hospitals in large metro areas across the US. According to the Substance Abuse and Mental Health Services Administration (SAMHDA),

“A DAWN case is any ED visit involving recent drug use that is implicated in the ED visit. DAWN captures both ED visits that are directly caused by drugs and those in which drugs are a contributing factor, but not the direct cause of the ED visit. Annually, DAWN produces estimates of drug-related visits to hospital EDs for the nation as a whole and for selected metropolitan areas.

DAWN is used to monitor trends in drug misuse and abuse, identify the emergence of new substances and drug combinations, assess health hazards associated with drug abuse, and estimate the impact of drug misuse and abuse on the Nation’s health care system. DAWN relies on a longitudinal probability sample of hospitals located throughout the United States.

To be eligible for selection into the DAWN sample, a hospital must be a non-federal, short-stay, general surgical and medical hospital located in the United States, with at least one 24-hour ED. The dataset includes demographics, drugs involved in the ED visit (up to 16 drugs from 2004 through 2008 and up to 22 drugs from 2009 through 2011), toxicology confirmation, route of administration, type of case, and disposition of the patient following the visit.

Prepared DAWN Emergency Department National and Metro data tables are available on the DAWN website. The [DAWN website](#) also provides access to DAWN reports.”<sup>4</sup>

## Data Wrangling

The initial step in the project was to 'wrangle' the data, which is the process of taking raw data and transforming it into a format that is suitable for analysis.

First, the DataFrame was examined to see its size and column names. It has 284 columns, and 229,221 rows. Since the DataFrame is very large and somewhat unwieldy, the decision was made to remove the 'sdled' and 'CATID' columns. These columns won't be used in the analysis because their inclusion is beyond the scope of this project. The modified DataFrame has 84 columns.

Second, summary data for all columns was obtained. Below are the results:

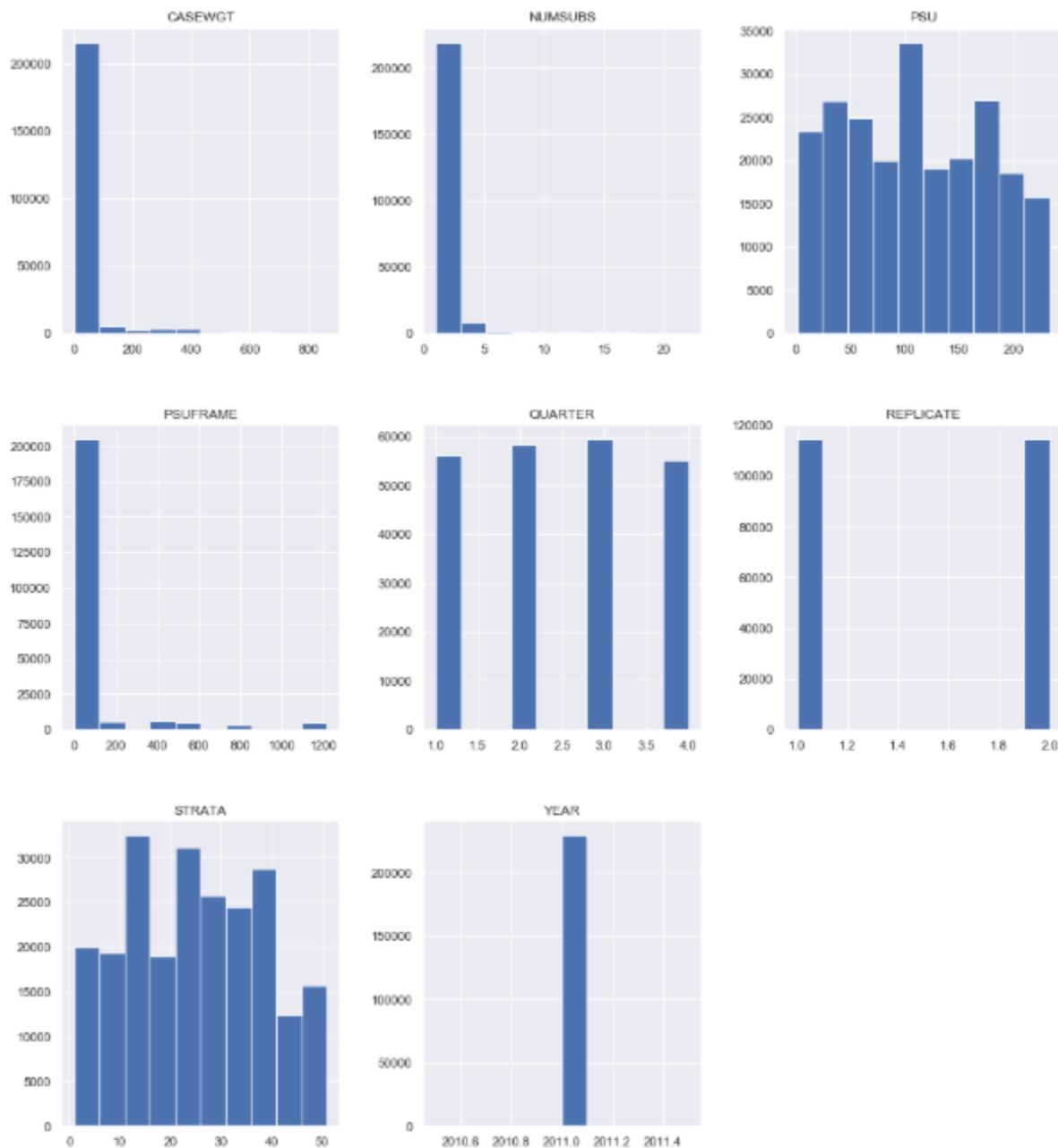
	STRATA	PSU	REPLICATE	CASEWGT	PSUFRAME	YEAR	QUARTER	NUMSUBS
count	229211.000000	229211.000000	229211.000000	229211.000000	229211.000000	229211.0	229211.000000	229211.000000
mean	24.681551	109.610839	1.500028	22.107901	73.348304	2011.0	2.495997	1.584104
std	13.331152	64.444097	0.500001	68.403862	211.852785	0.0	1.104806	1.163778
min	1.000000	1.000000	1.000000	0.938440	2.000000	2011.0	1.000000	1.000000
25%	13.000000	53.000000	1.000000	2.714999	8.000000	2011.0	2.000000	1.000000
50%	25.000000	109.000000	2.000000	4.190787	10.000000	2011.0	3.000000	1.000000
75%	35.000000	165.000000	2.000000	7.148615	17.000000	2011.0	3.000000	2.000000
max	51.000000	233.000000	2.000000	862.824350	1215.000000	2011.0	4.000000	22.000000

Next, the number of distinct values per column was counted. The results are below:

Feature	Count
CASEWGT	2,931
DRUGID_1	1,604
DRUGID_2	1,074
...	...
ALLABUSE	2
YEAR	1

Many of the variables in the DataFrame are categorical but are represented by numbers. In order to tell the data story better, it was decided to replace the numbers with the corresponding label. This replacement was done for all the variables including the 2600 drug names in the data glossary that was available with the dataset. Also, though there exists a shorter, cleaner way to perform the replacements than to do them one by one, due to time constraints, it was decided to use the method that is working rather than spend more time researching a more 'pythonic' method. Additionally, there were too many drug names to replace them using the same method as with the others. Therefore, it was decided to replace them by importing a dictionary and mapping the names to the numbers.

Then, subplot histograms of all quantitative variables were plotted. The resulting plots are below:



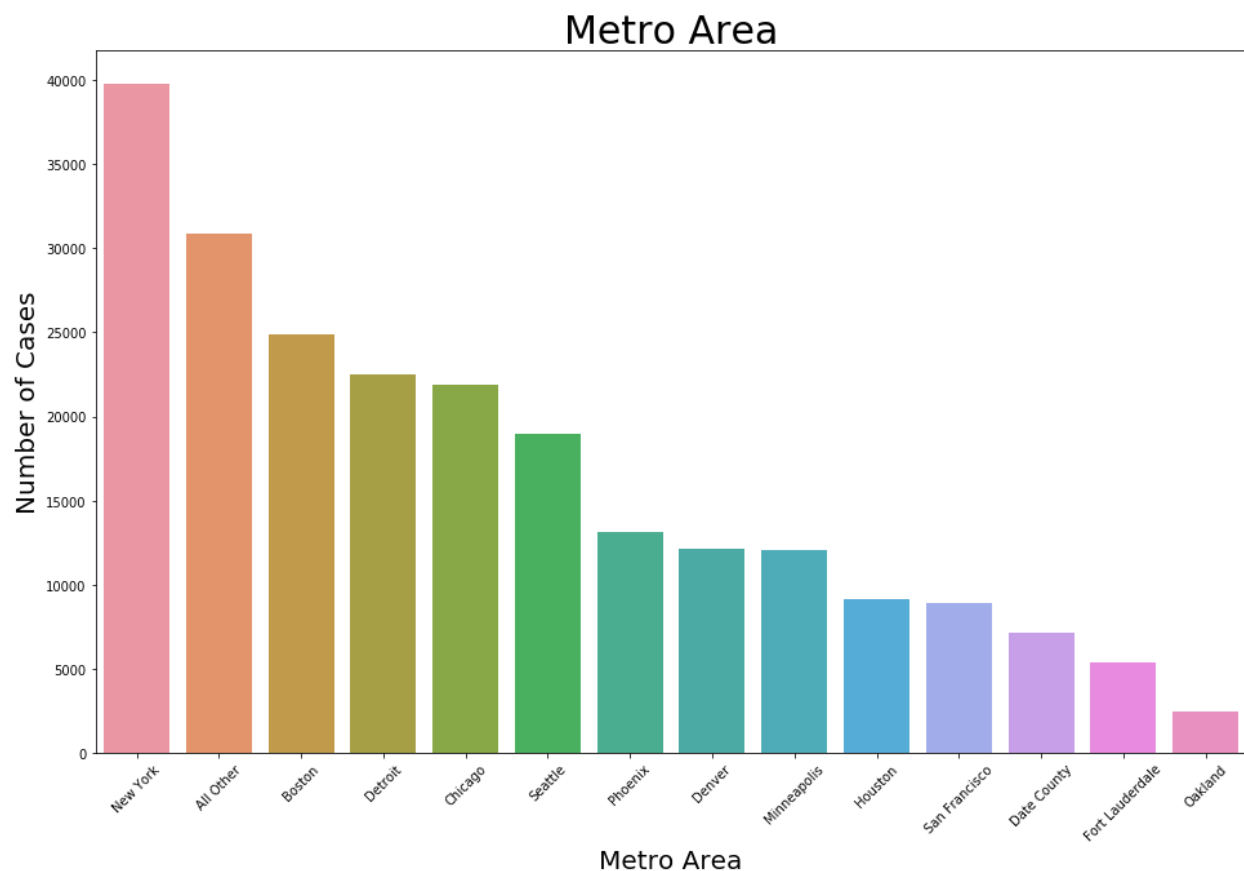
Finally, the DataFrame was checked for missing values. None were found.

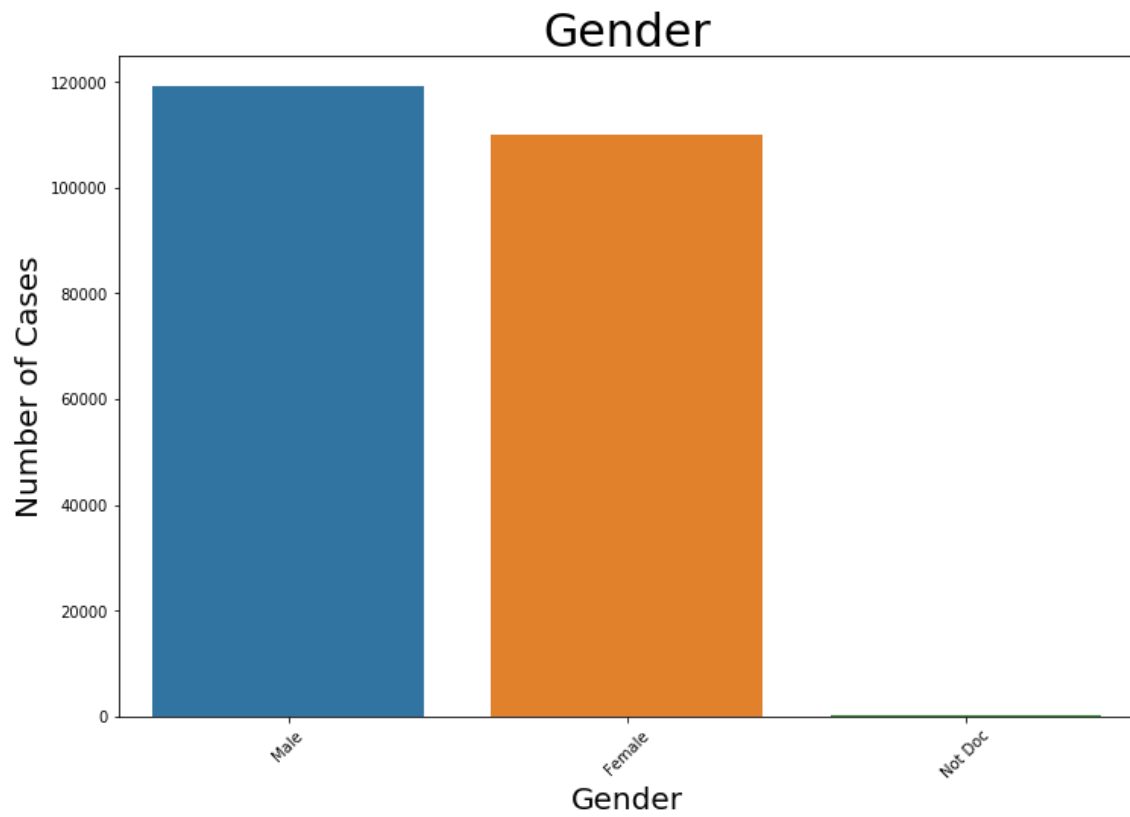
In summary, this data set has 86 columns and 229,211 rows. Most of the variables are categorical. In the original data set, the categorical variables were represented by numbers. Steps were taken to replace the numbers with their corresponding names. Since the DataFrame is very large and somewhat unwieldy the decision was made to remove the 'sdled' and 'CATID' columns. These columns won't be used in the analysis because their inclusion is beyond the

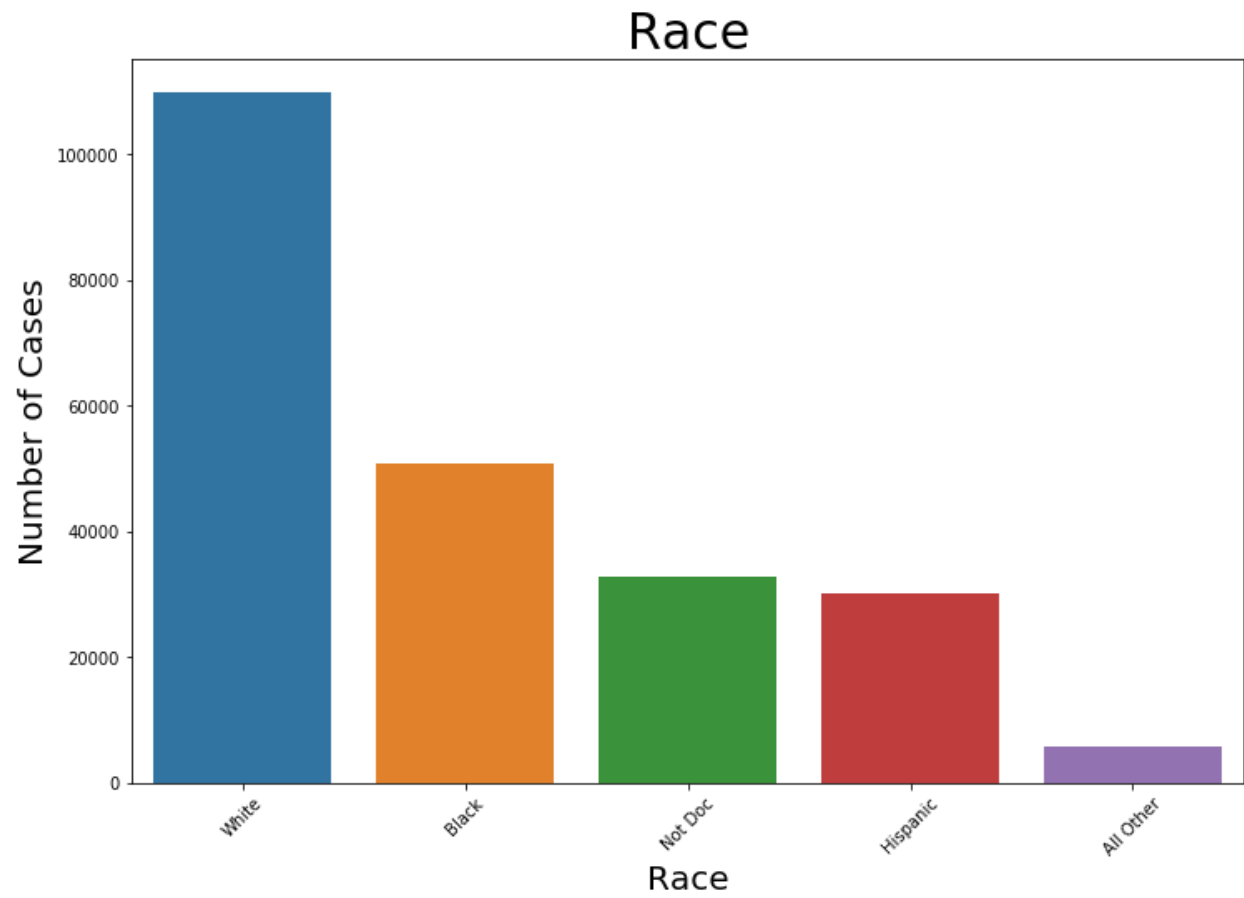
scope of this project. Additionally, the number of distinct values for each variable was plotted, subplot histograms of the quantitative variables were plotted, and it was determined that the data set did not include any missing values.

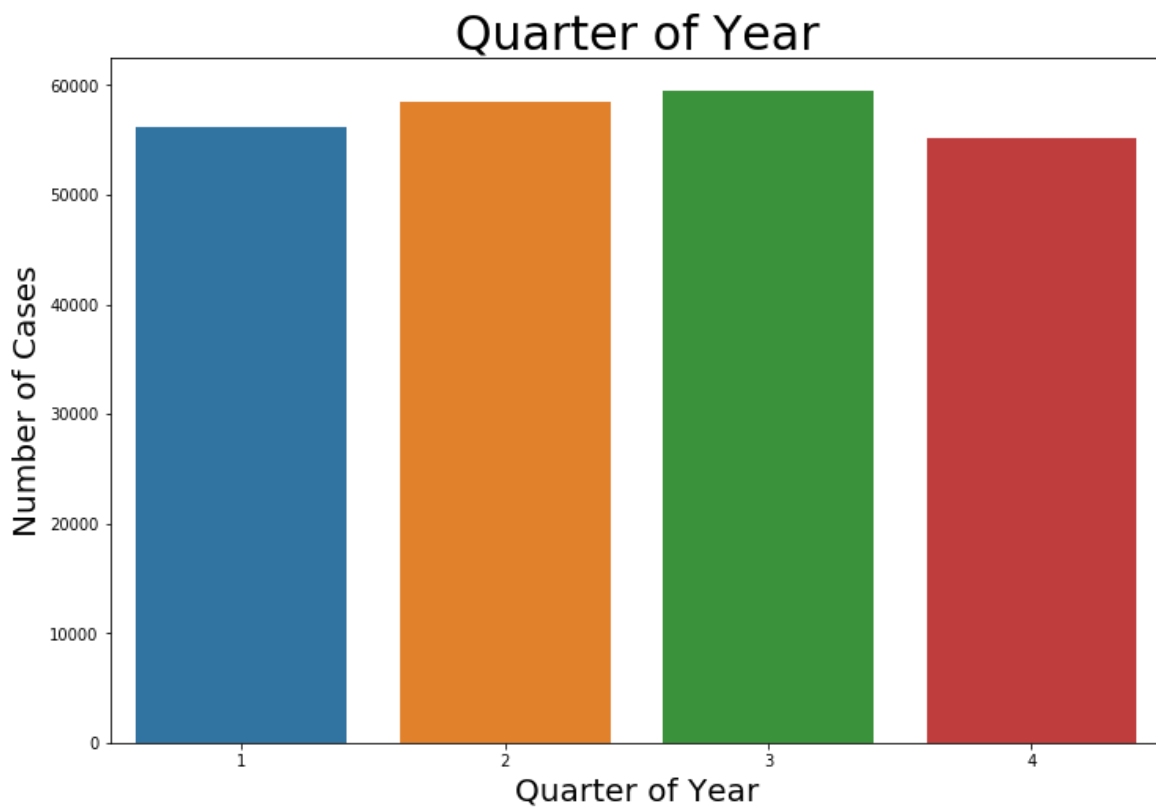
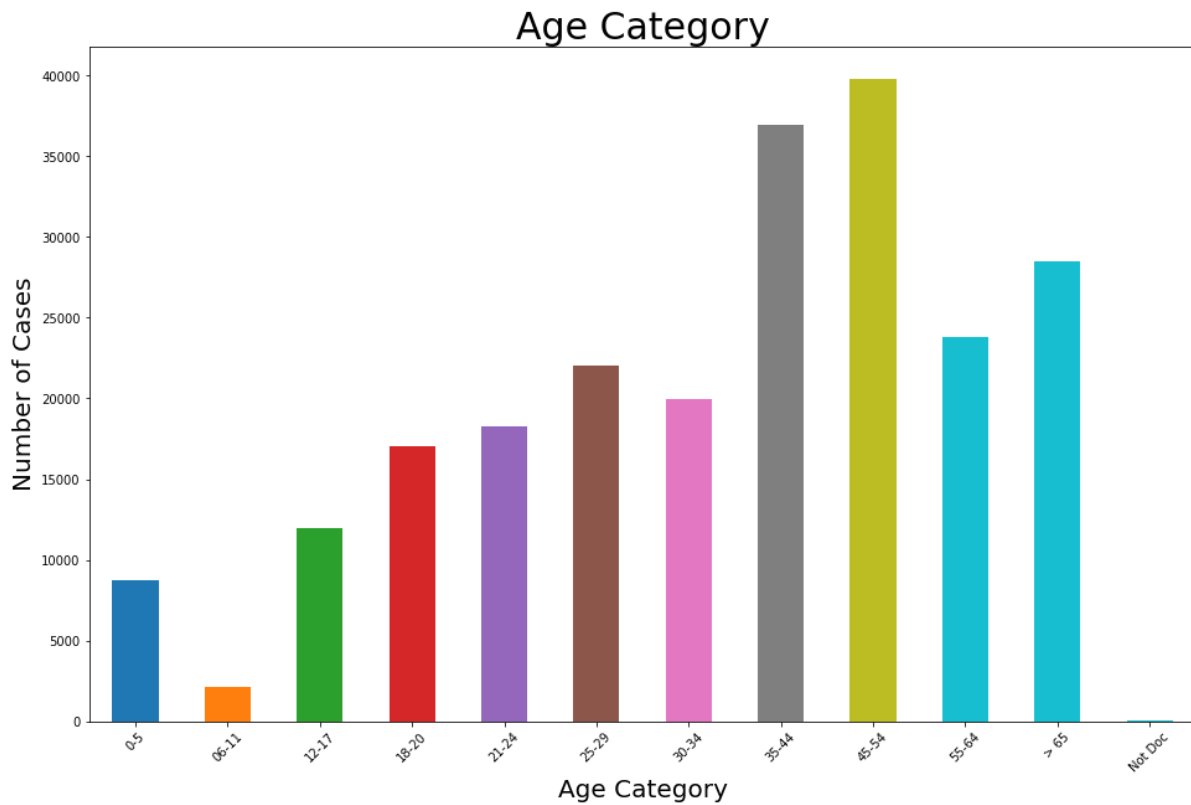
## Data Story

With the data wrangling done, it's time to see what story the data tells. In order to do this, all the categorical variables were plotted. The plots are below:

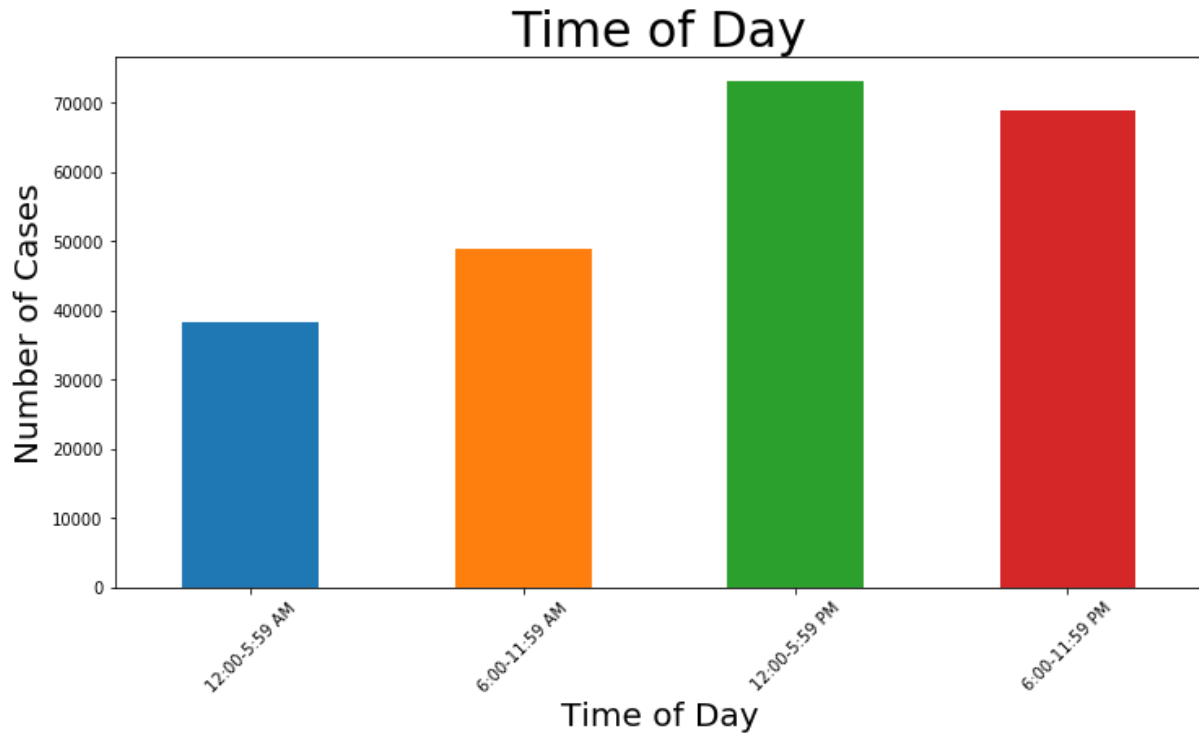




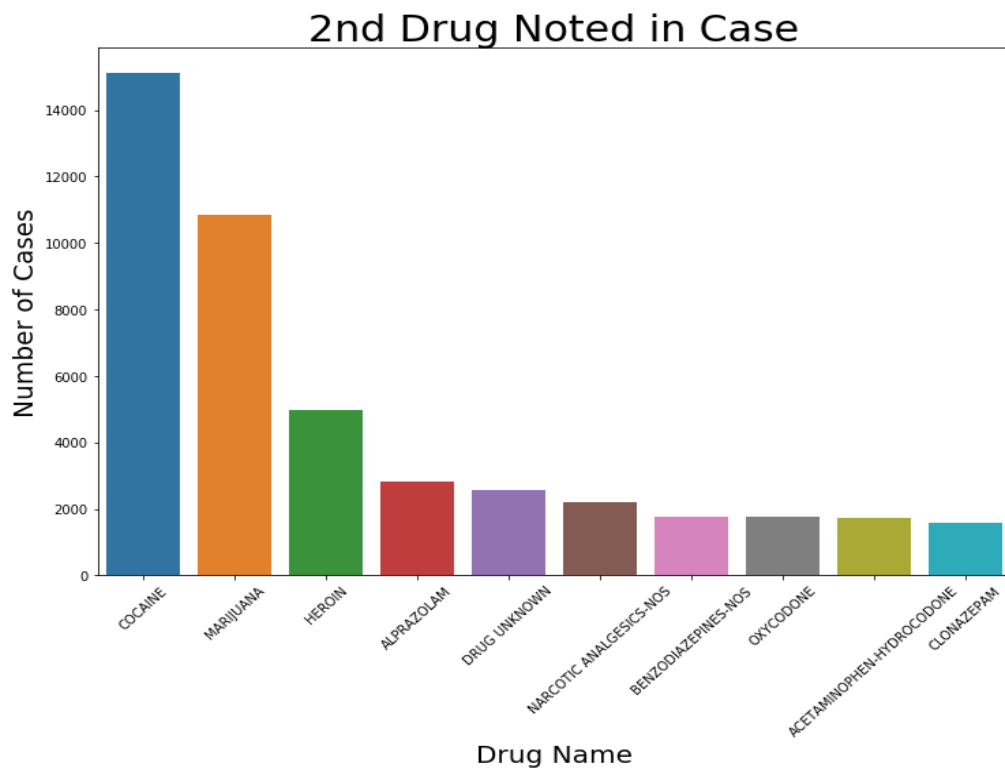
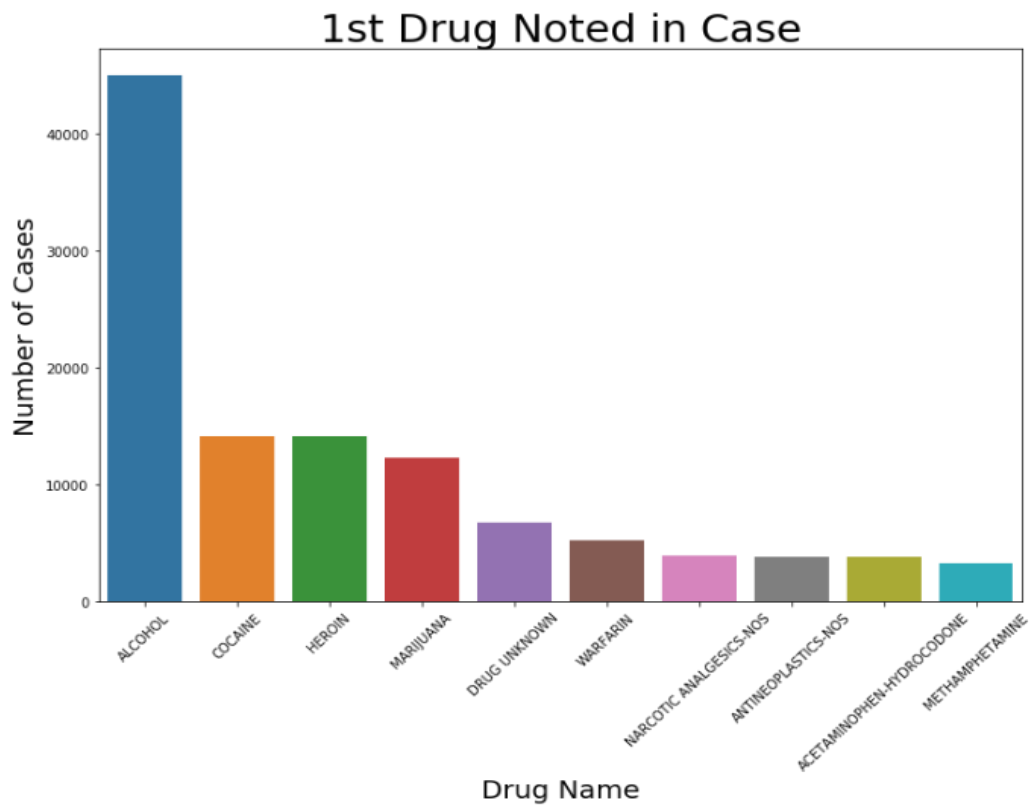


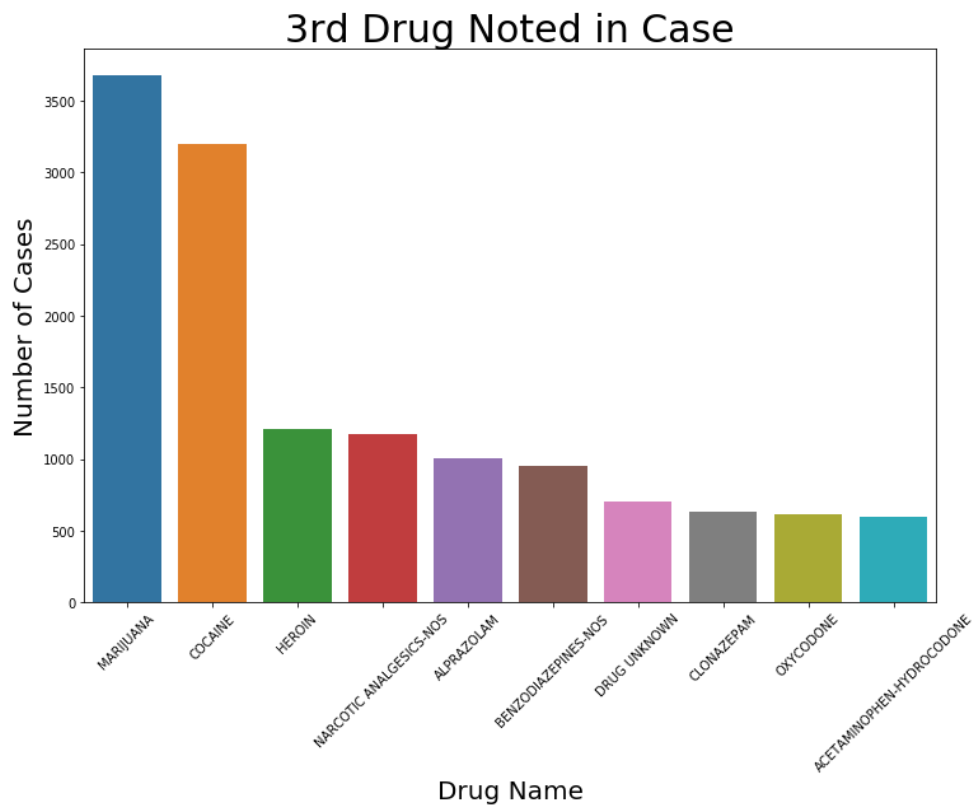




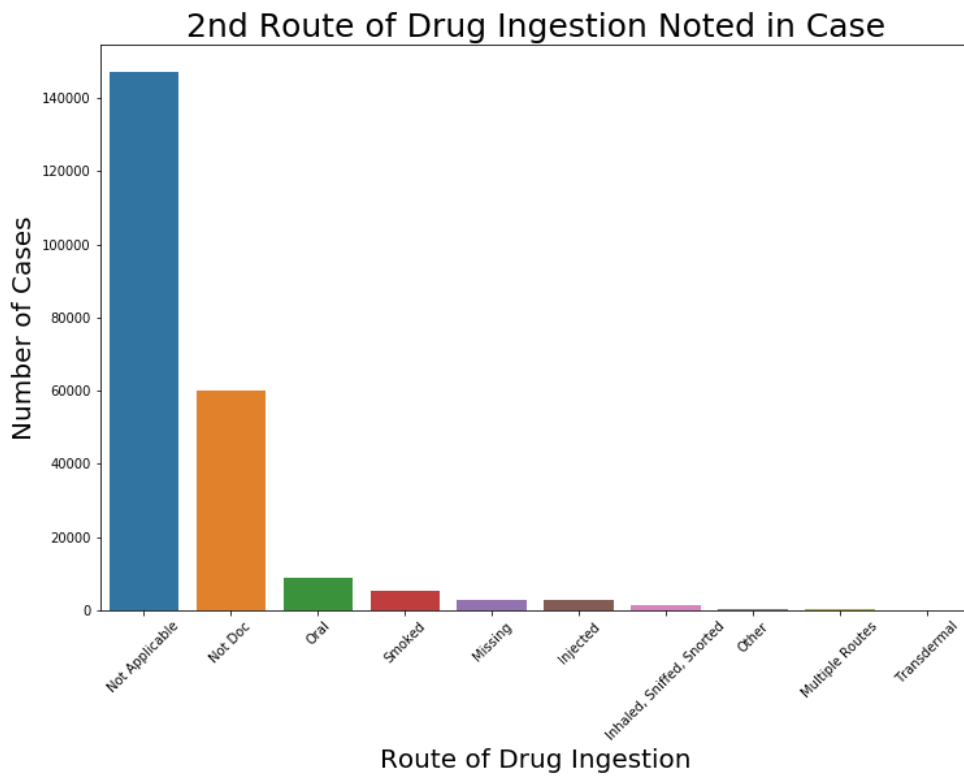
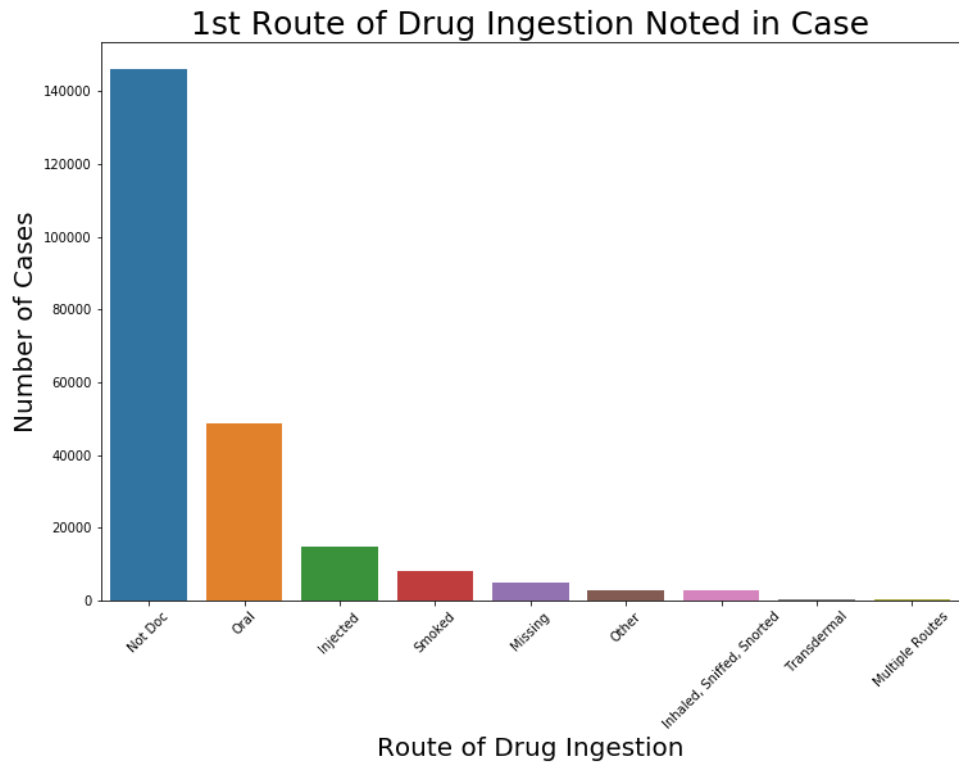


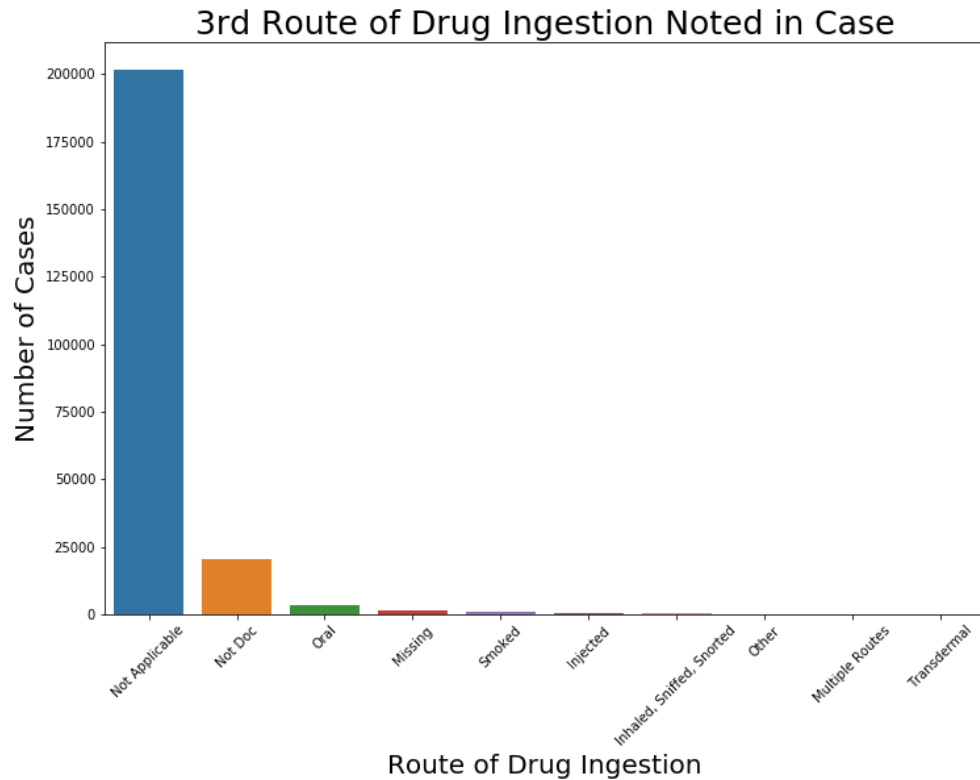
When plotting the distribution of the Drug(s) involved in the cases, it was decided to show only top 10 since there are > 2600 drugs possible. Otherwise, the graph would be too large and hard to read. Also, there were up to 22 substances involved in one case, and a plot was done for each instance. In order to be concise and deliver the main points of the content, the body of this paper does not contain all 22 of the graphs. The first few are included here, and the others are included in the appendix at the end of the document.



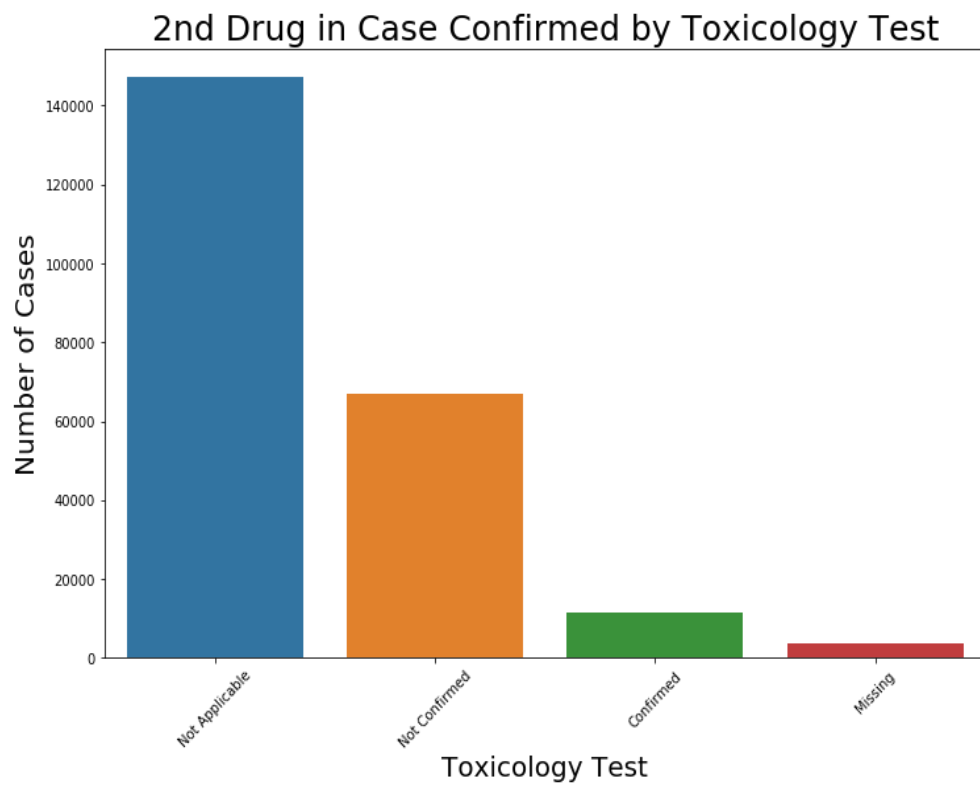
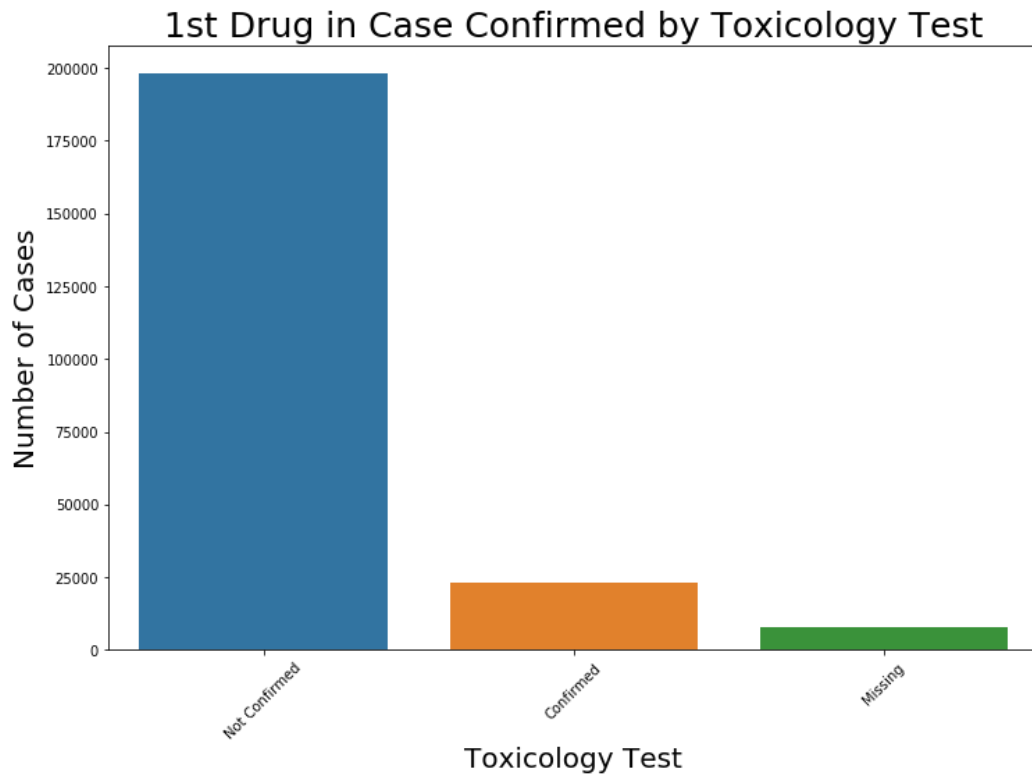


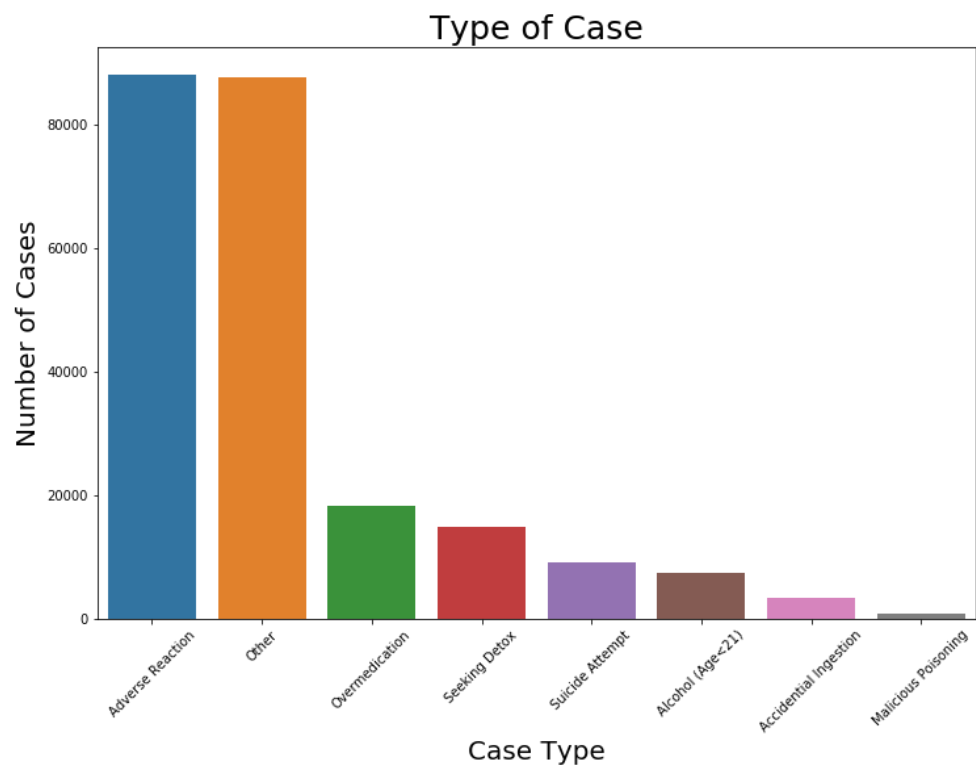
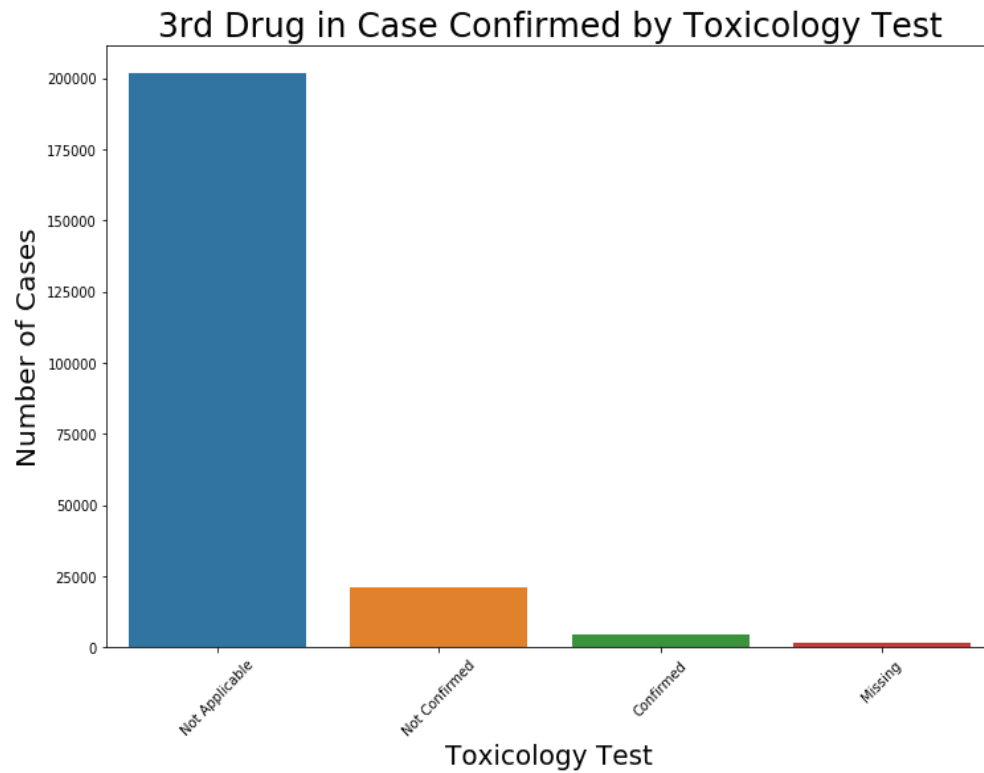
As with plotting the distribution of the Drug(s) involved in the cases, there were up to 22 routes of ingestion for substances involved in one case, and a plot was done for each instance. In order to be concise and deliver the main points of the content, the body of this paper does not contain all 22 of the graphs. The first few are included here, and the others are included in the appendix at the end of the document.

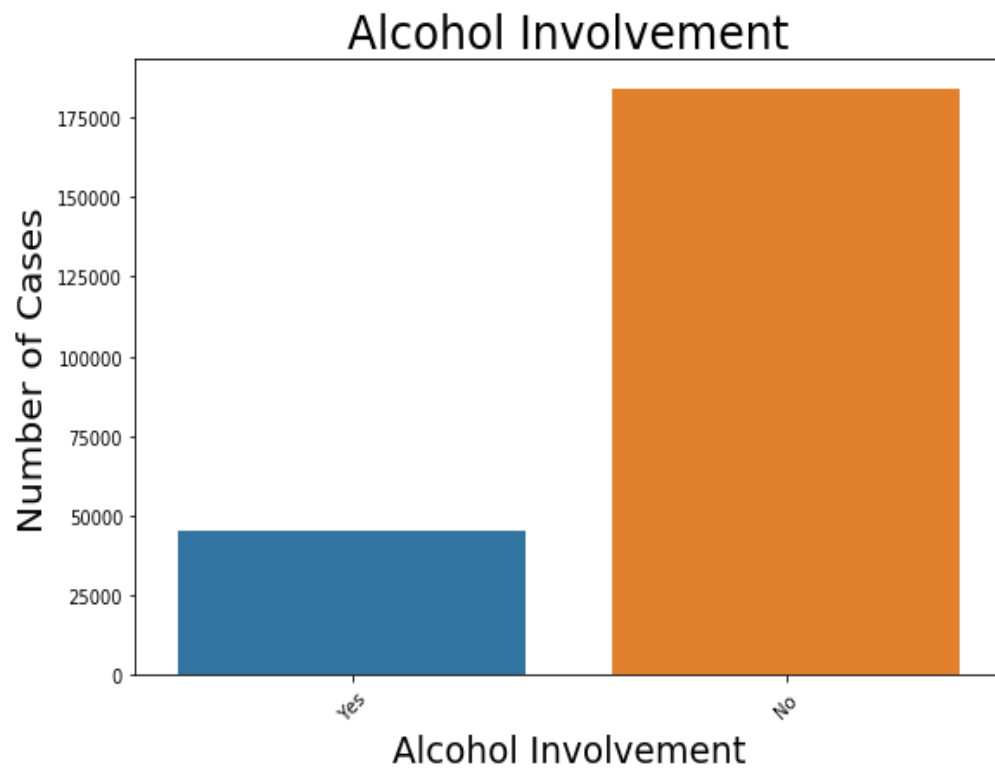
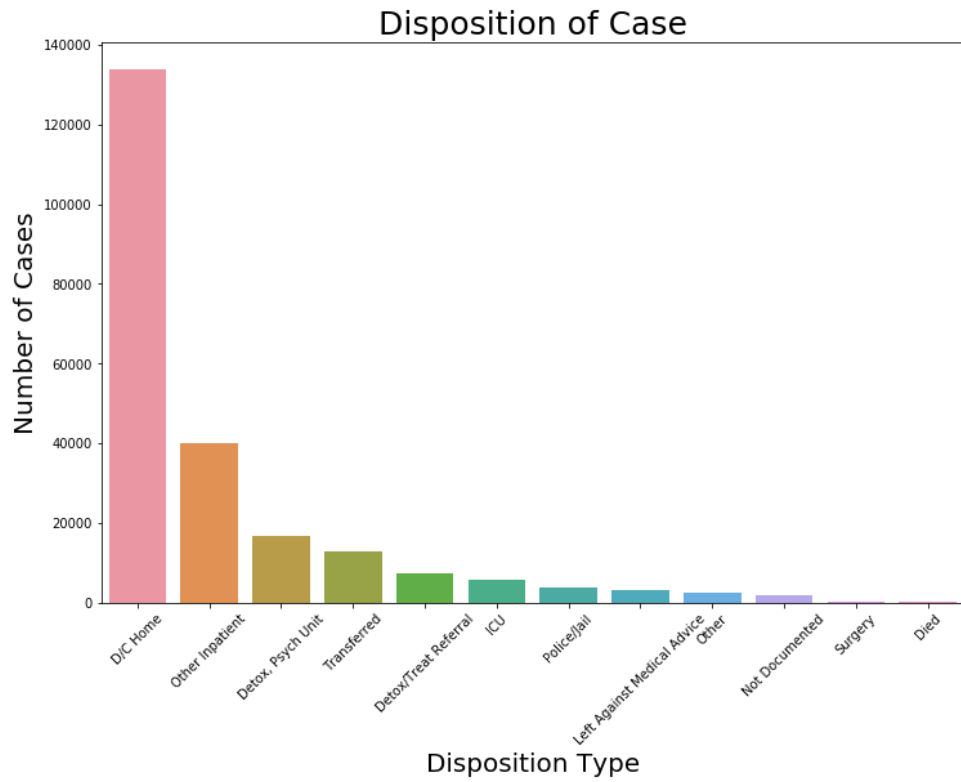




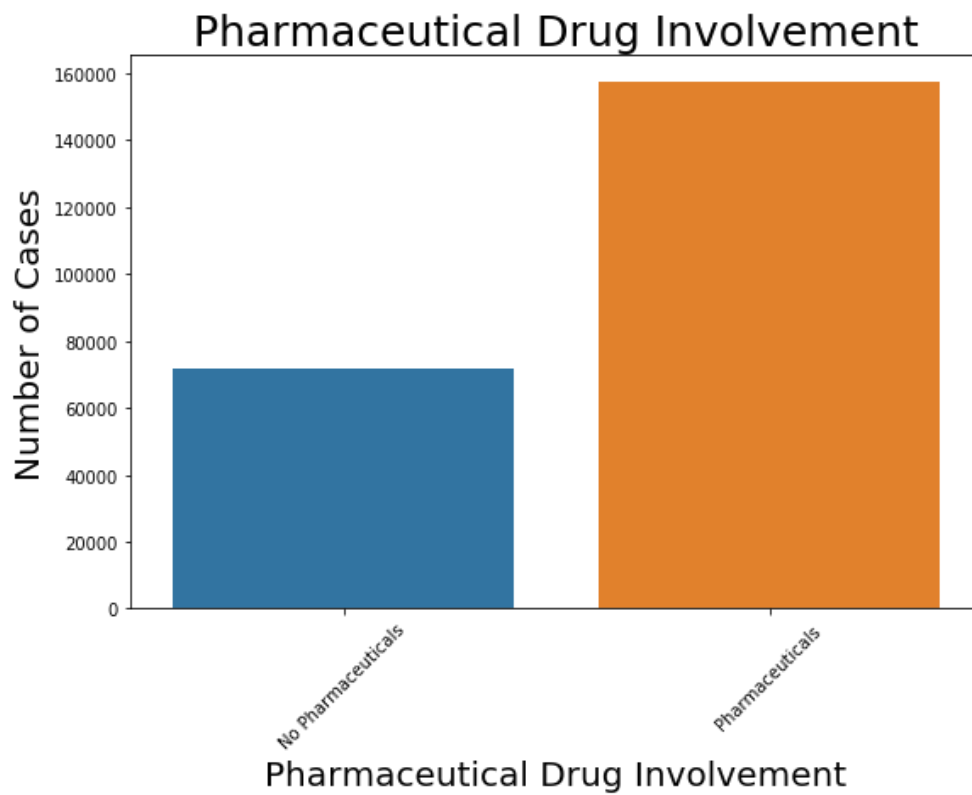
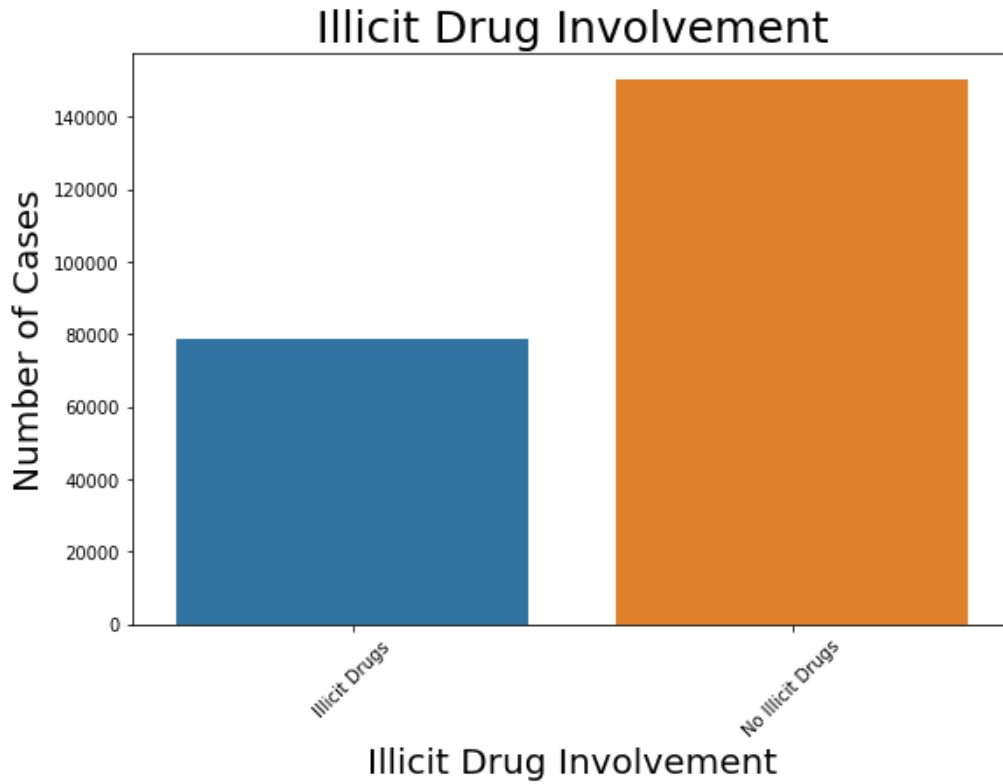
Again, as with plotting the distribution of the Drug(s) involved in the cases, whether a toxicology test was done for each of the up to 22 substances involved in one case was recorded, and a plot was done for each instance. In order to be concise and deliver the main points of the content, the body of this paper does not contain all 22 of the graphs. The first few are included here, and the others are included in the appendix at the end of the document.

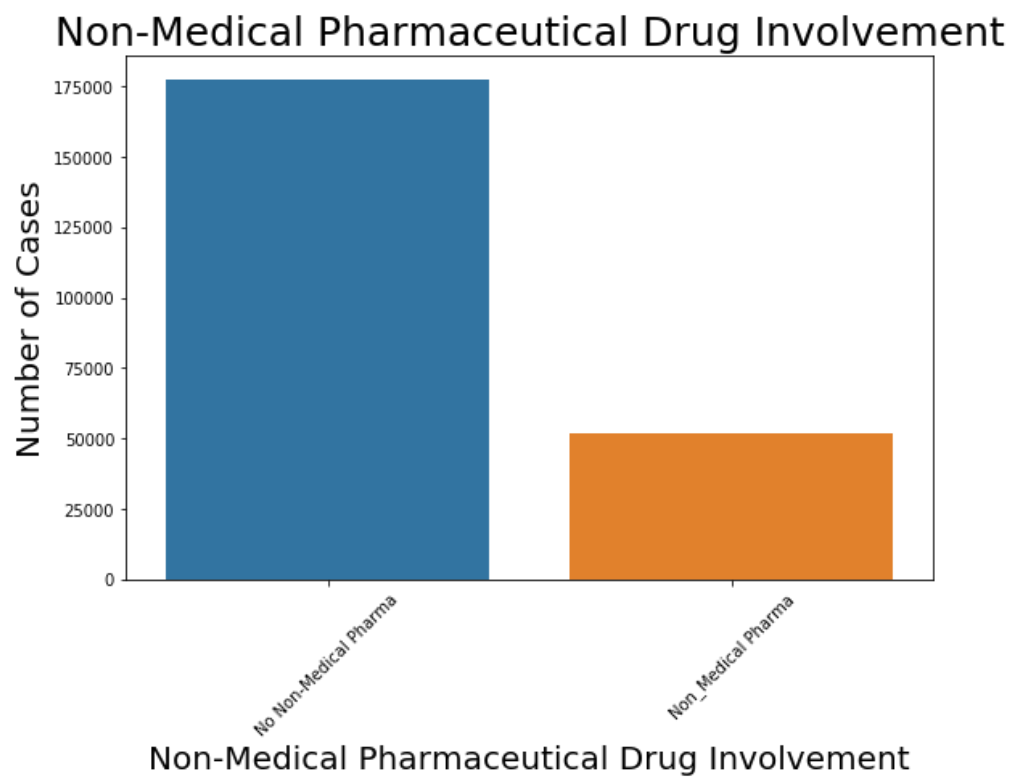














This data set has 85 variables. Eighty of them are categorical. These variables include demographic information on each case, such as metro area, age, sex and race. They also include the name of the drug(s) ingested, the route by which the drug(s) was ingested and whether a toxicology report was done. There are up to 22 drugs involved in a case. Also included in the data are the type of case, it's disposition, whether alcohol was involved, and whether the drugs were illicit, pharmaceutical or non-medical pharmaceutical.

The following observations were seen from the plots:

- The distribution of gender was fairly even.
- The predominate race involved in the cases was white.
- Most of the cases were between the ages of 35-54.
- The time of year is evenly distributed.
- The time of day that most cases occur is between 12:00 pm and 11:59 pm.
- Alcohol, cocaine, heroin and marijuana are involved in many of the cases.
- Most of the drugs were taken orally.
- Most were not confirmed by a toxicology test.
- Most of the cases were due to adverse reactions to the drugs.
- Most cases were discharged home.
- Alcohol was not involved in many of the cases.

- Most of the drugs involved were not illicit.
- Most of the drugs involved were pharmaceutical.
- Most of the cases did not involve non-medical pharmaceutical drugs.
- Most of the cases qualified as all misuse and abuse episodes, which means that the drugs were used for purposes that is not consistent with legal or medical guidelines.

## Inferential Statistics

The next step in the project was inferential statistics. Since this DataFrame has many variables, only a few were chosen for hypothesis testing. Also, since the variables are categorical, the Chi-Squared test for proportions and independence was used for the tests.

First, the locations of the cases were analyzed. Is there a significant difference in the frequency of cases in each of the cities?

- $H_0: p_1 = p_2 \dots = p_{15}$
- $H_a: p_1 \neq p_2 \dots \neq p_{15}$

The p-value was 0.0; therefore,  $H_0$  was rejected. It is reasonable to conclude that the distribution of the proportion of the location of the cases is not equal.

Second, the race of the cases was analyzed. Does the race of the patient vary significantly?

- $H_0: p_1 = p_2 \dots = p_{15}$
- $H_a: p_1 \neq p_2 \dots \neq p_{15}$

The p-value was 0.0; therefore,  $H_0$  was rejected. It is reasonable to conclude that the distribution of the proportion of race is not equal.

Next, the independence of disposition vs. case type was analyzed.

- $H_0$ : the population frequencies are equal to the expected frequencies
- $H_a$ : the population frequencies are not equal to the expected frequencies.

The p-value was 0.0; therefore,  $H_0$  was rejected. It is reasonable to conclude that the distribution of population frequencies are not equal to the expected frequencies.

Finally, the independence of sex vs. drugid\_1 was analyzed.

- $H_0$ : the population frequencies are equal to the expected frequencies
- $H_a$ : the population frequencies are not equal to the expected frequencies.

The p-value was 0.0; therefore,  $H_0$  was rejected. It is reasonable to conclude that the distribution of population frequencies are not equal to the expected frequencies.

## Modeling

Once the data story was completed, it was time to build the model.

Logistic Regression is the model most used for data sets that have many categorical variables. Logistic Regression is a classification technique that can be used to predict a qualitative response. In other words, these models predict the probability that a categorical variable such as gender belongs to a particular category. For example: if a person has long hair, what is the probability that he/she is female?

It was decided to use the Logistic Regression model on the full data set to determine the probability of an Emergency Department Drug Abuse visit being a particular case type. The results are in the table below:

### Logistic Regression Full Data Set

Class	Precision	Recall	F1-score	Support
Accidental Ingestion	0.88	0.01	0.02	1,301
Adverse Reaction	0.92	0.99	0.95	35,239
Alcohol (Age<21)	0.99	1.00	1.00	2,968
Malicious Poisoning	0.50	0.01	0.02	317
Other	0.87	0.93	0.90	35,051
Overmedication	0.74	0.72	0.73	7,259
Seeking Detox	0.77	0.51	0.61	5,937
Suicide Attempt	0.74	0.46	0.57	3,613

Note that the table above provides values for precision, recall, f1-score and support for each class (case type). The explanation for what these values represent is below:

- The precision represents how many are correctly classified among that class
- The recall means how many of this class were found over the whole number of elements of this class
- The f1-score is the harmonic mean between precision & recall
- The support is the number of occurrences of the given class in the dataset (there were 35,239 instances of Adverse Reactions).

As can be seen in the table, the data set is not balanced between the case types. This classifier does not do well with classes that have lower amounts of cases.

Therefore, in the Extended Analysis notebook, other approaches were explored with the goal of improving the results obtained in this baseline modeling notebook. The approaches included

combining resampling methods with Logistic Regression (LGR) and Random Forest Classifiers (RFC).

First, an attempt was made to run the full data set through SMOTE. SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to correct an imbalance problem. It aims to balance class distribution by randomly increasing minority class examples through replication. However, this operation did not work because the computer kept returning a memory error. As a result, it was decided to reduce the dataset by 50% and run the SMOTE process again. This time, the process ran successfully, and Logistic Regression was done on the more balanced reduced dataset. The results are in the table below:

#### Logistic Regression Reduced Data Set

Class	Precision	Recall	F1-score	Support
Accidental Ingestion	0.22	0.68	0.33	814
Adverse Reaction	0.97	0.84	0.90	22,024
Alcohol (Age<21)	0.99	1.00	1.00	1,855
Malicious Poisoning	0.09	0.75	0.16	198
Other	0.95	0.65	0.77	21,907
Overmedication	0.47	0.75	0.58	3,710
Seeking Detox	0.47	0.75	0.58	3,710
Suicide Attempt	0.47	0.72	0.57	2,258

Since the results of the Logistic Regression on the reduced dataset were not much improved, it was decided to use the Random Forest Classifier on the reduced dataset. The results are in the table below:

#### Random Forest Reduced Data Set

Class	Precision	Recall	F1-score	Support
Accidental Ingestion	0.82	0.04	0.08	814
Adverse Reaction	0.90	0.99	0.94	22,024
Alcohol (Age<21)	0.99	1.00	1.00	1,855
Malicious Poisoning	0.84	0.27	0.41	198
Other	0.84	0.93	0.89	21,907
Overmedication	0.71	0.66	0.68	4,537
Seeking Detox	0.80	0.44	0.57	3,710
Suicide Attempt	0.73	0.24	0.36	2,258

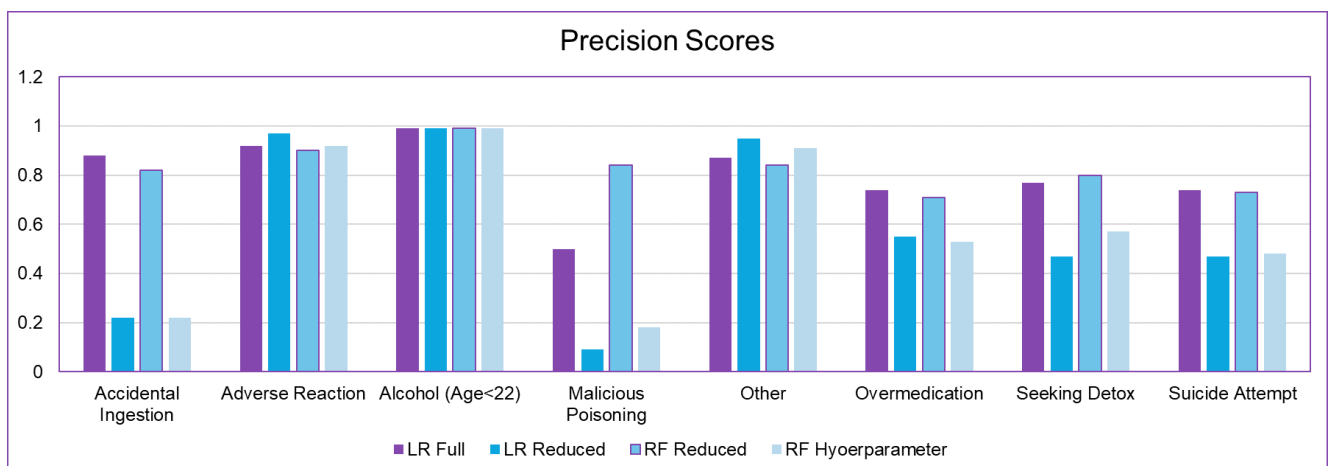
The results of the Random Forest Classifier on the reduced data set show that the model was overfitting the data. Therefore, it was decided to use the Randomized Search process to tune the hyperparameters and run the model again. The results are in the table below:

### Random Forest Reduced Data Set—Hyperparameter Tuning

Class	Precision	Recall	F1-score	Support
Accidental Ingestion	0.22	0.30	0.25	814
Adverse Reaction	0.92	0.92	0.92	22,024
Alcohol (Age<21)	0.99	1.00	1.00	1,855
Malicious Poisoning	0.18	0.40	0.25	198
Other	0.91	0.75	0.82	21,907
Overmedication	0.53	0.94	0.67	4,537
Seeking Detox	0.57	0.60	0.59	3,710
Suicide Attempt	0.48	0.39	0.43	2,258

Once again, the results were not much improved. Therefore, it would be preferable to do some more fine tuning and re-run the models. However, due to time constraints, it was decided to end the project. A table comparing the precision scores for each of the models is below. It shows the overfitting in the basic models and the effect of the adjustments in the tuned models.

### Precision Scores for Each Model



### Conclusion

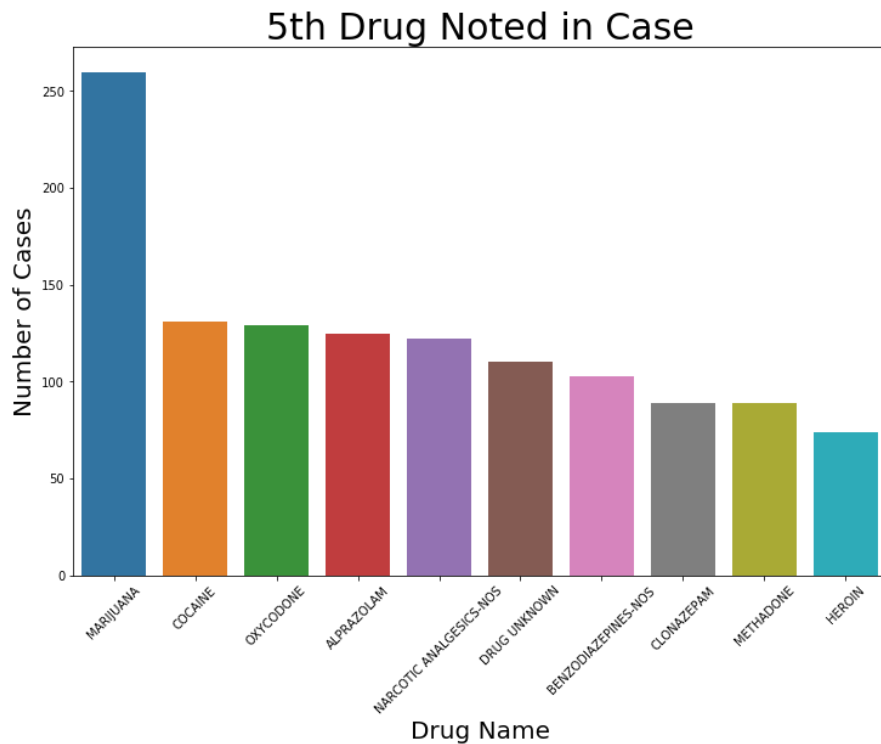
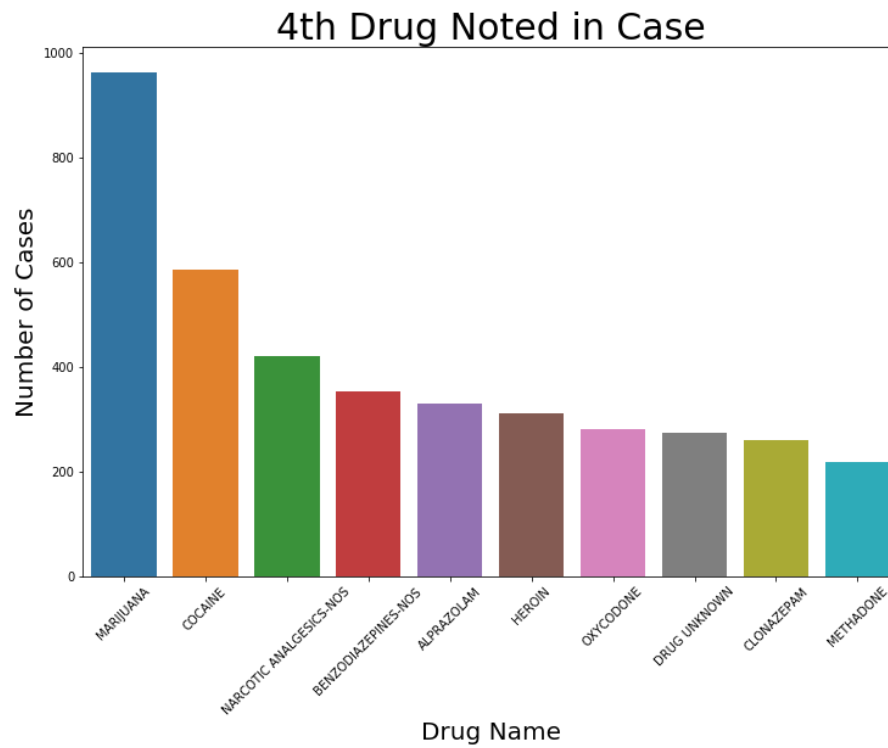
In summary, this milestone report shows the process used that prepared the data set for analysis, the plots done that show the story the data is telling, the inference testing that was done and the machine learning modeling that was performed.

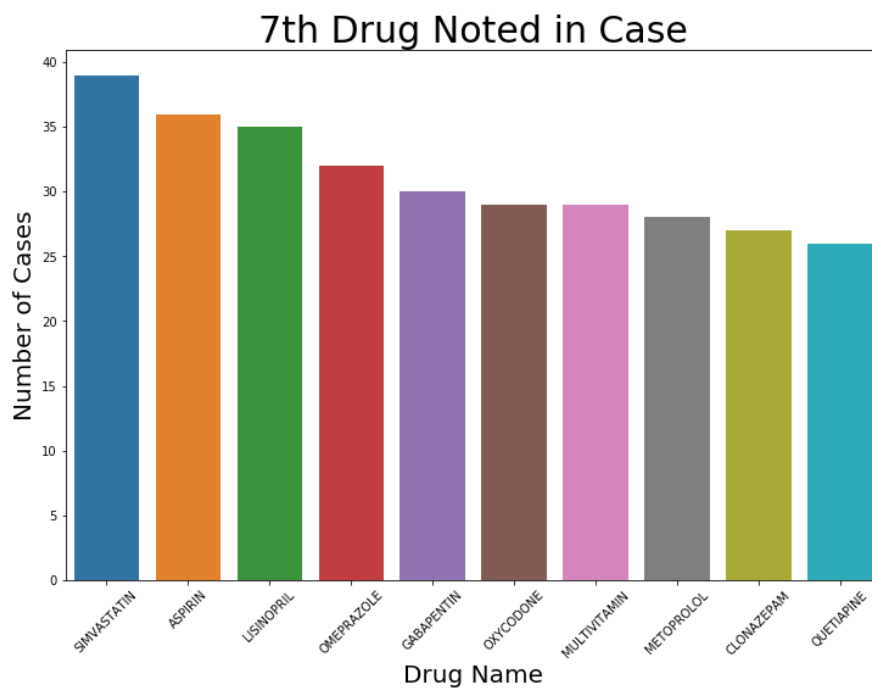
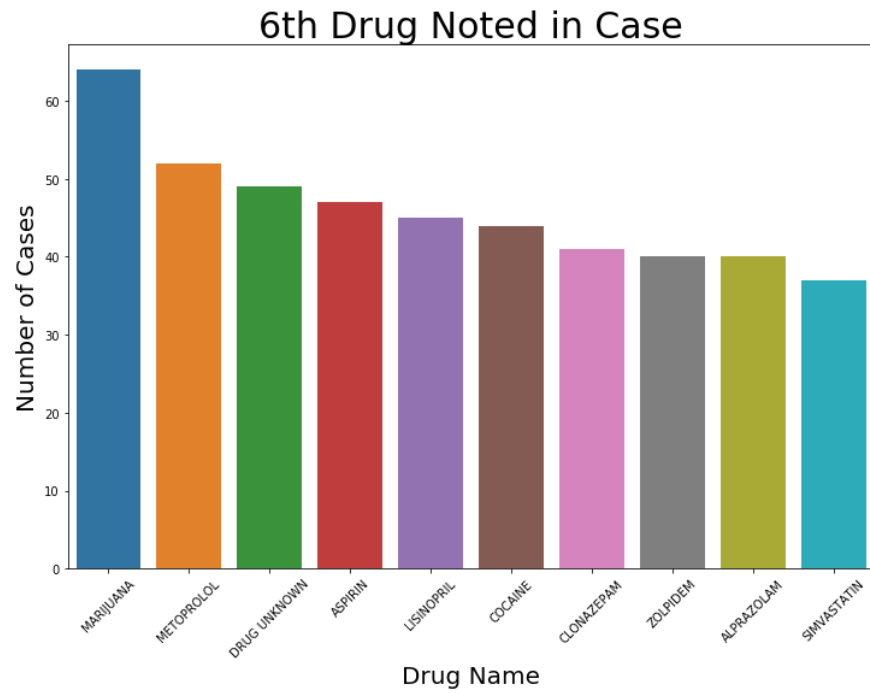
## Footnotes:

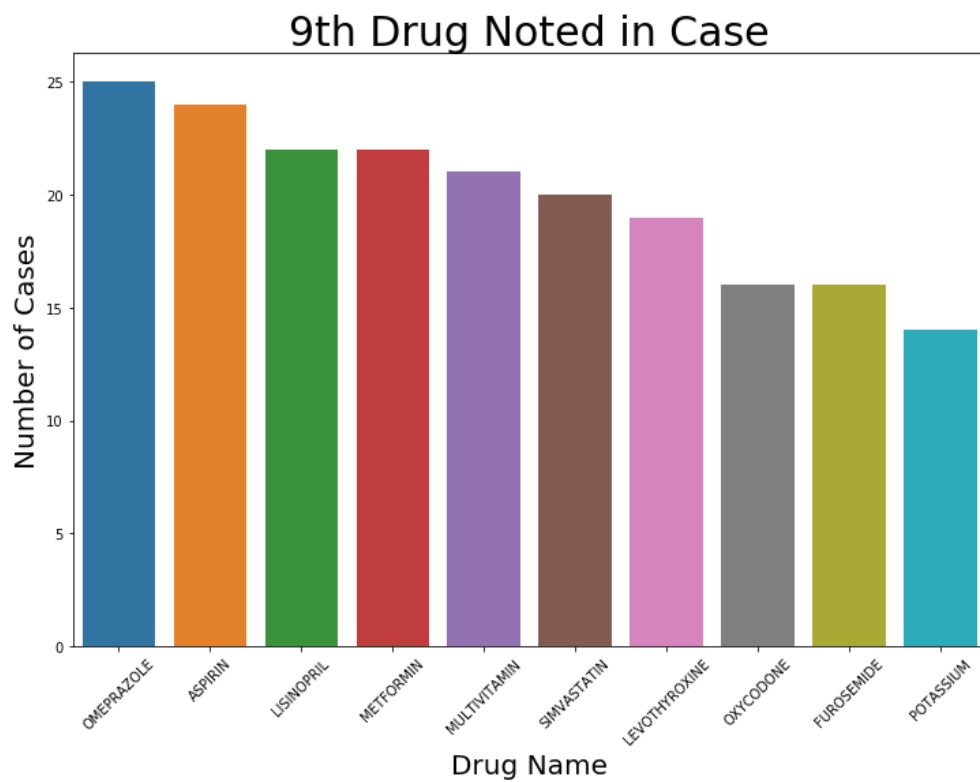
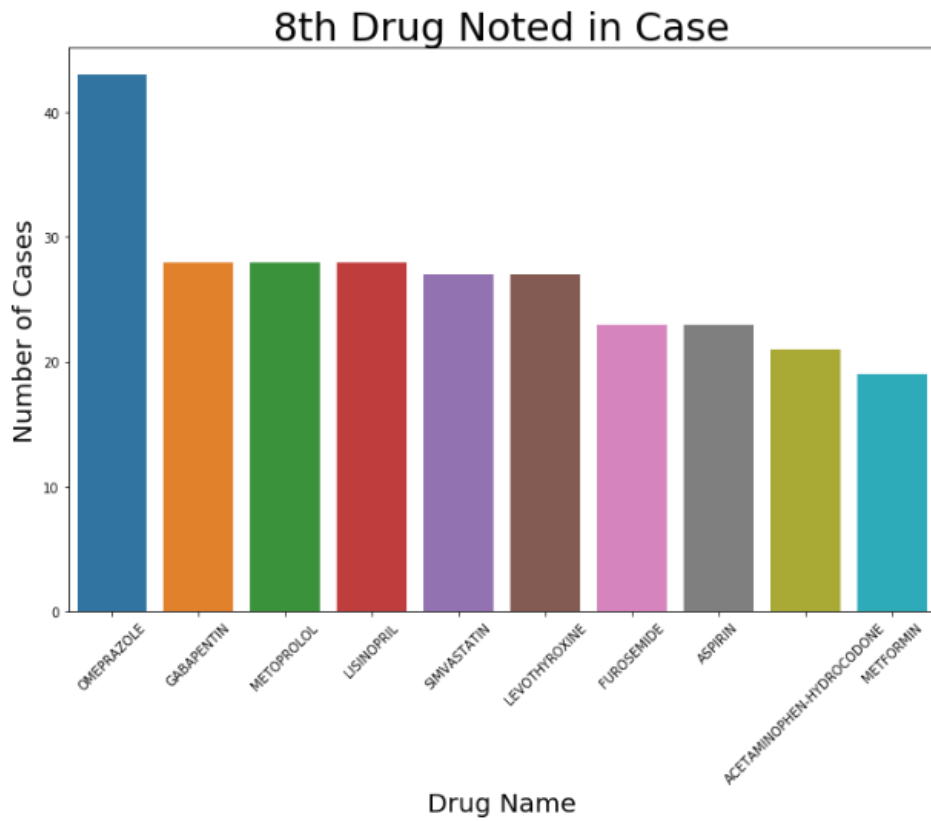
1. 'Understanding Drug Use and Addiction', *National Institute on Drug Abuse*, June, 2018, <https://www.drugabuse.gov/publications/drugfacts/understanding-drug-use-addiction>, (accessed December 14, 2019)
2. Yerby, Nathan, 'Statistics on Addiction in America', *Addiction Center*, December 5, 2019, <https://www.addictioncenter.com/addiction/addiction-statistics/>, (accessed December 14, 2019)
3. Albert, Michael, M.D., M.P.H.; McCaig, Linda F, M.P.H.; Uddin, Sayeedha, M.D., M.P.H., 'Emergency Department Visits for Drug Poisoning: United States, 2008-2011', *Centers for Disease Control and Prevention*, April, 2015, <https://www.cdc.gov/nchs/products/databriefs/db196.htm>, (accessed December 14, 2019)
4. 'Drug Abuse Warning Network 2011 (DAWN-2011-DS0001)', *Substance Abuse & Mental Health Services Administration*, 2011, <https://www.datafiles.samhsa.gov/study-dataset/drug-abuse-warning-network-2011-dawn-2011-ds0001-nid13747>, (accessed December 14, 2019)

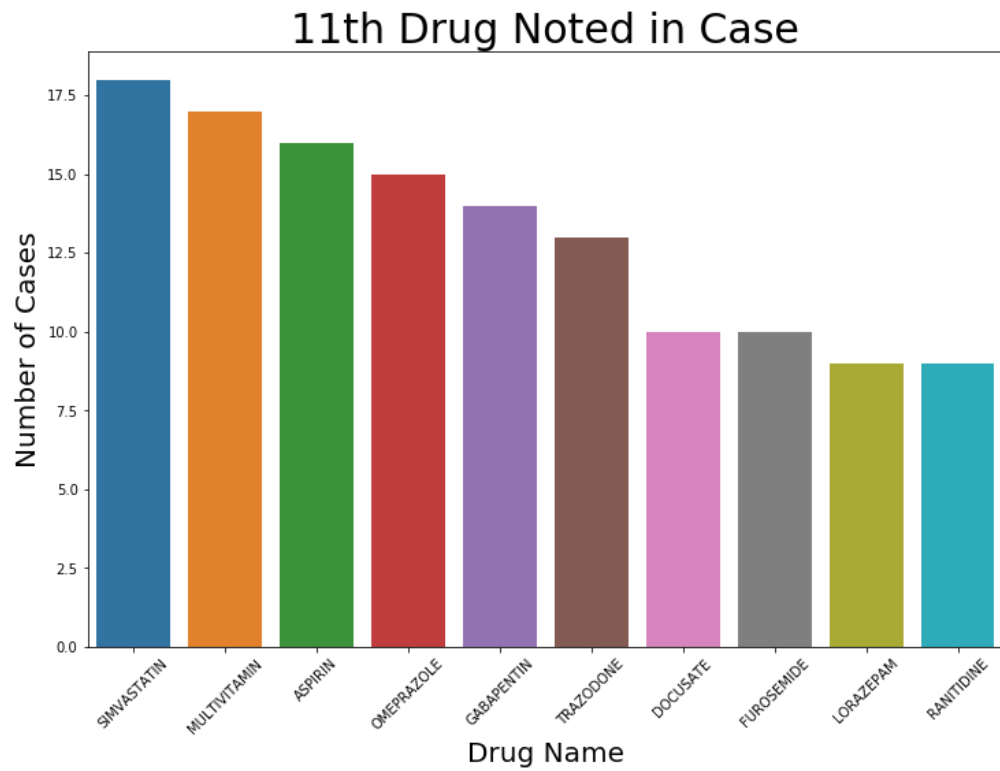
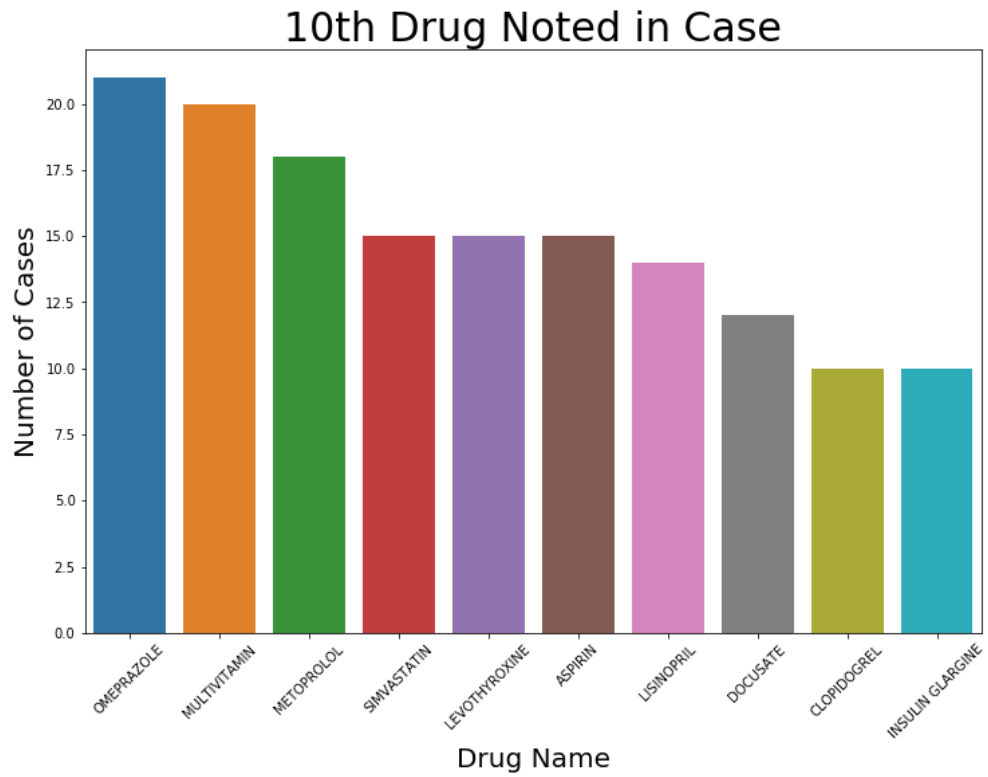


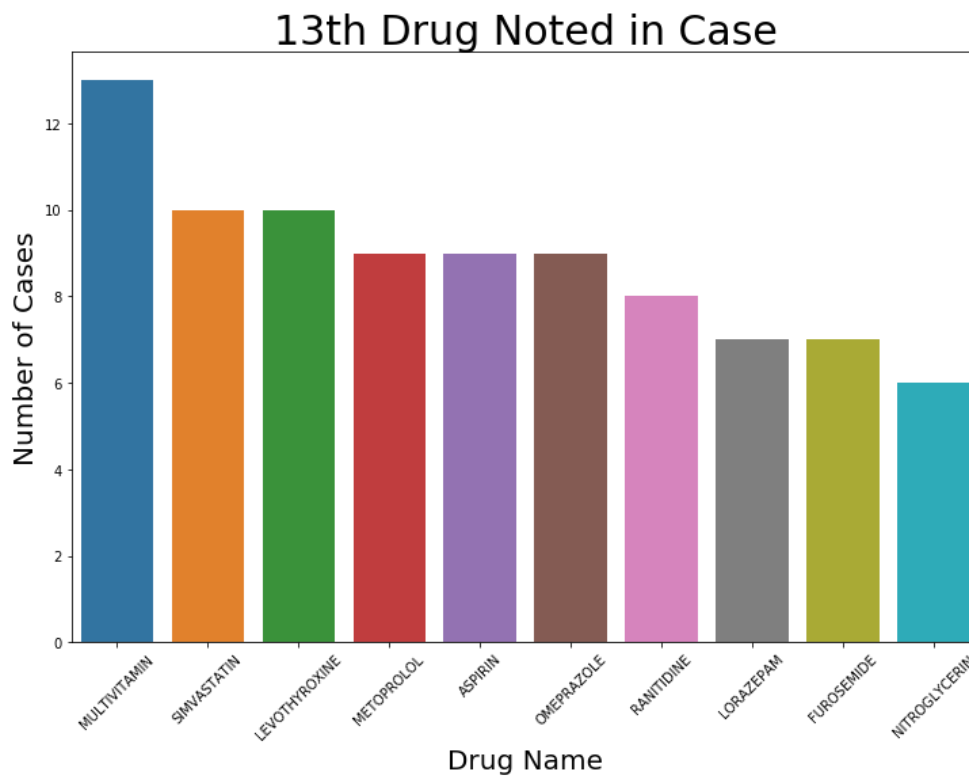
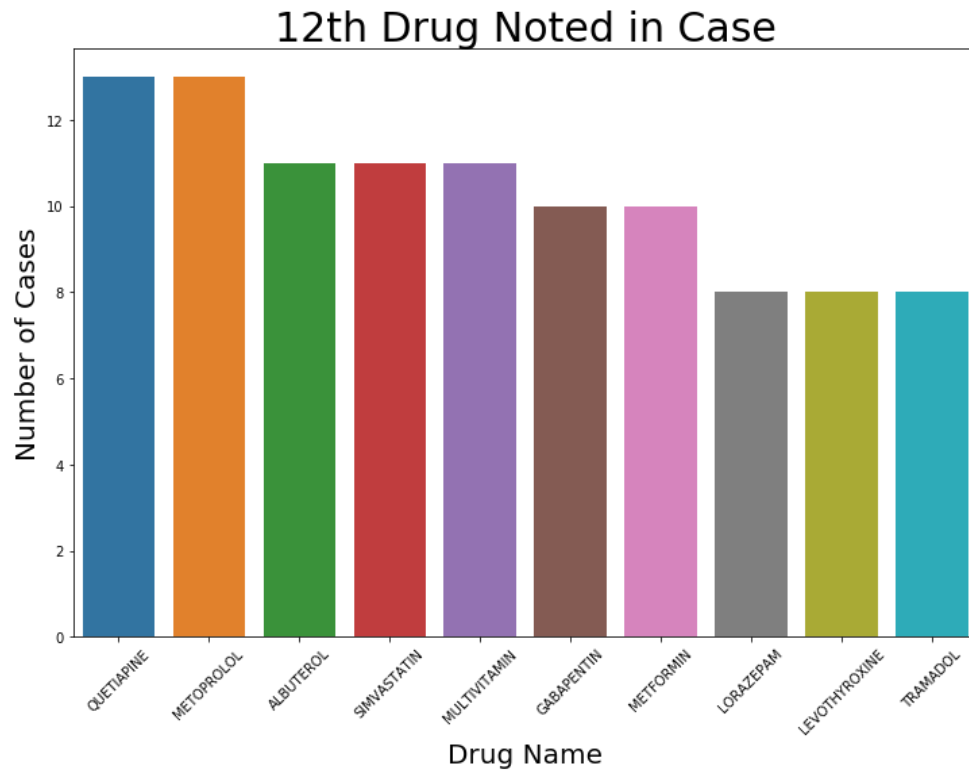
## Appendix A – Graphs of Top 10 Drugs Involved in Cases

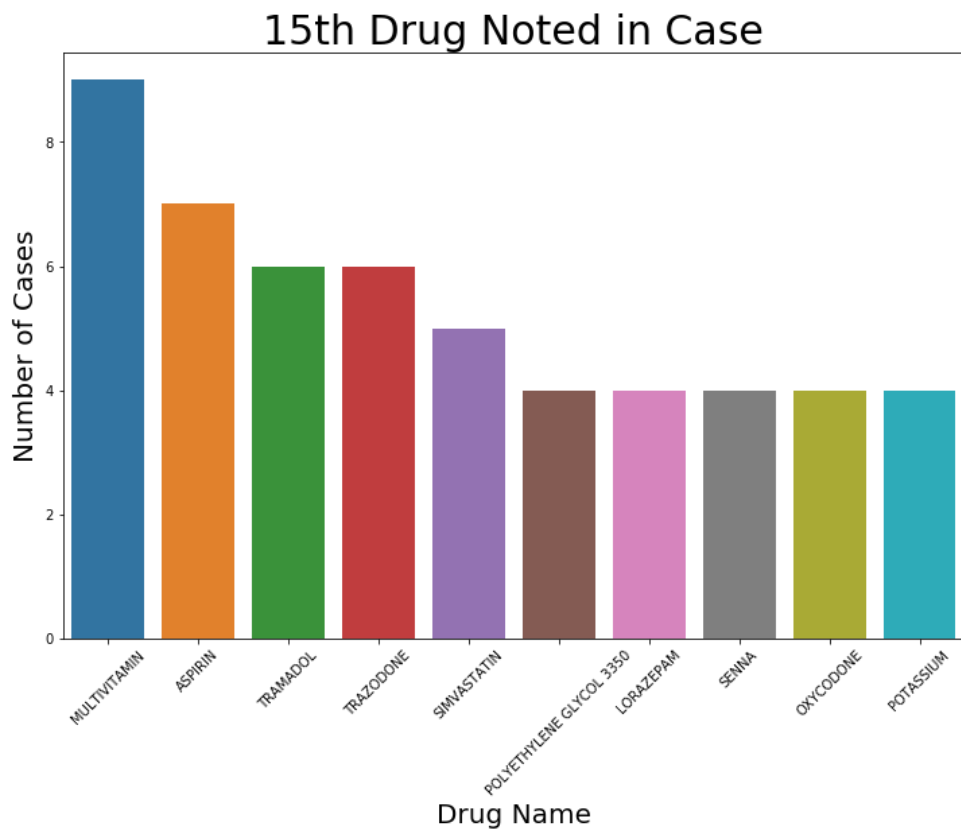
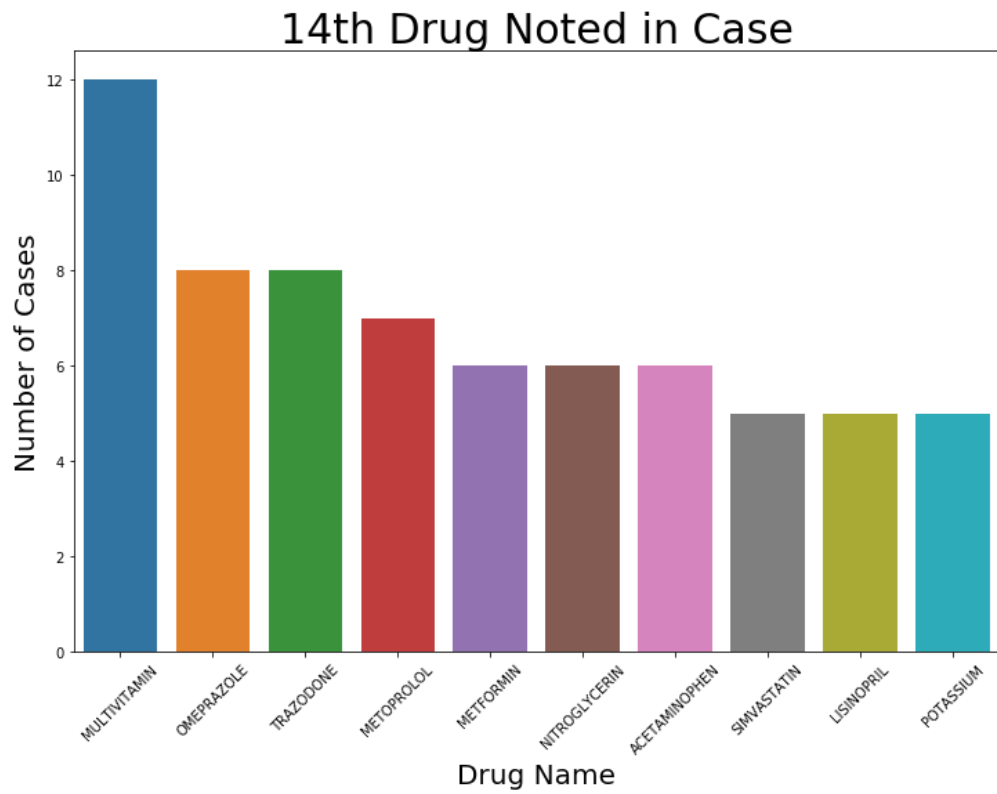


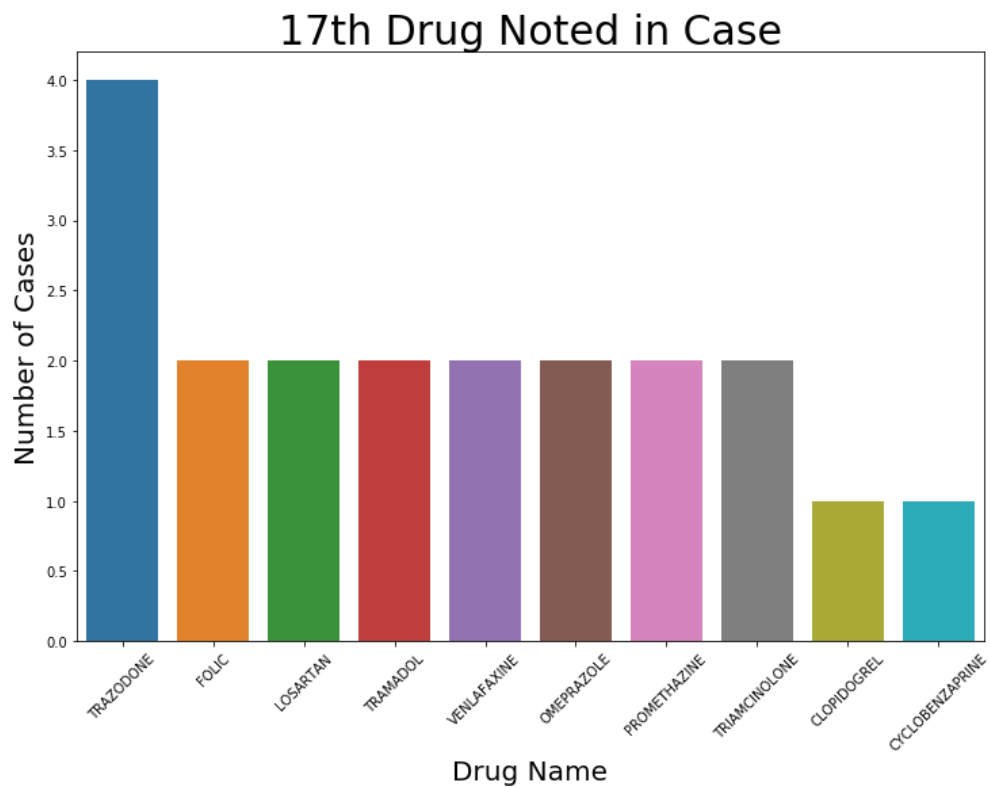
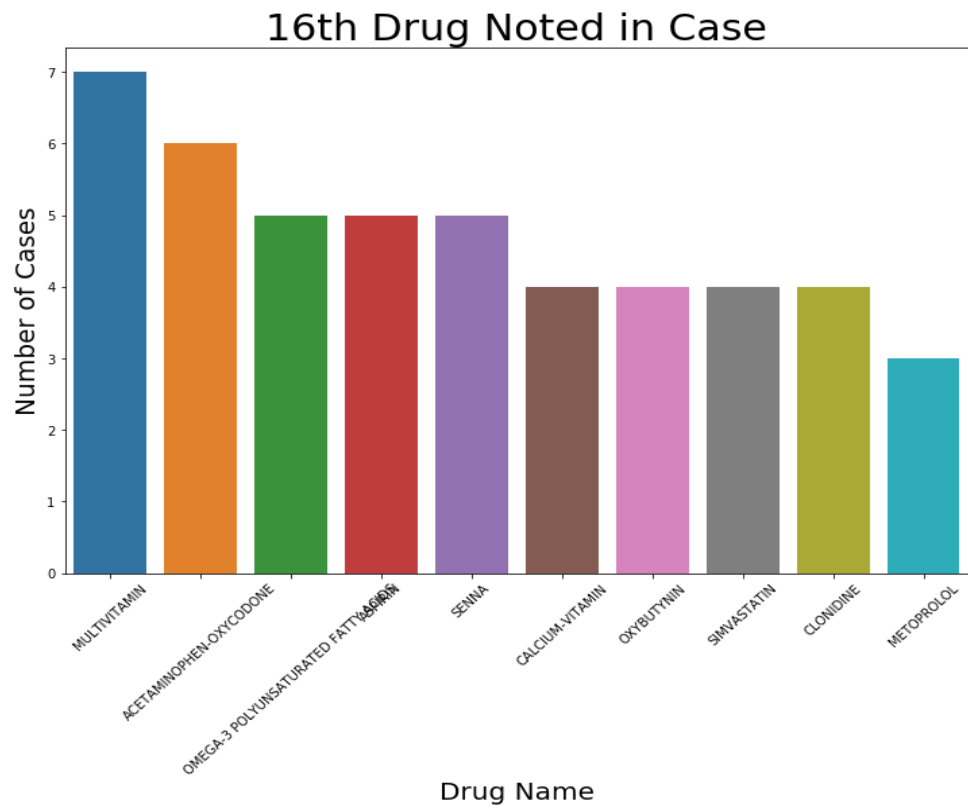


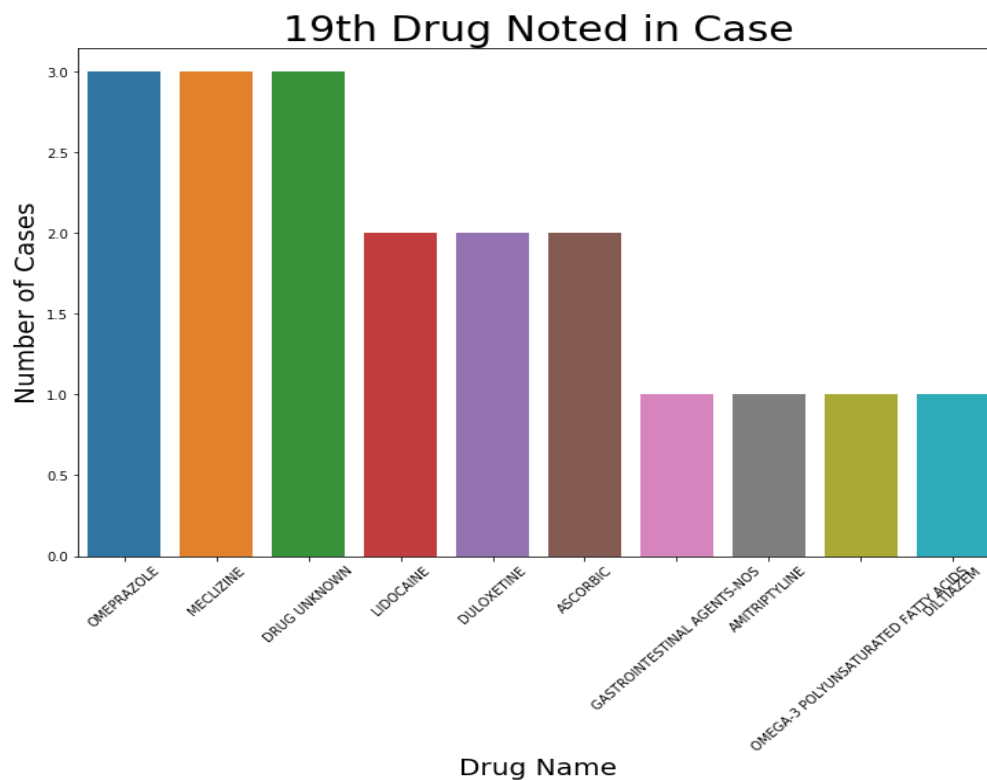
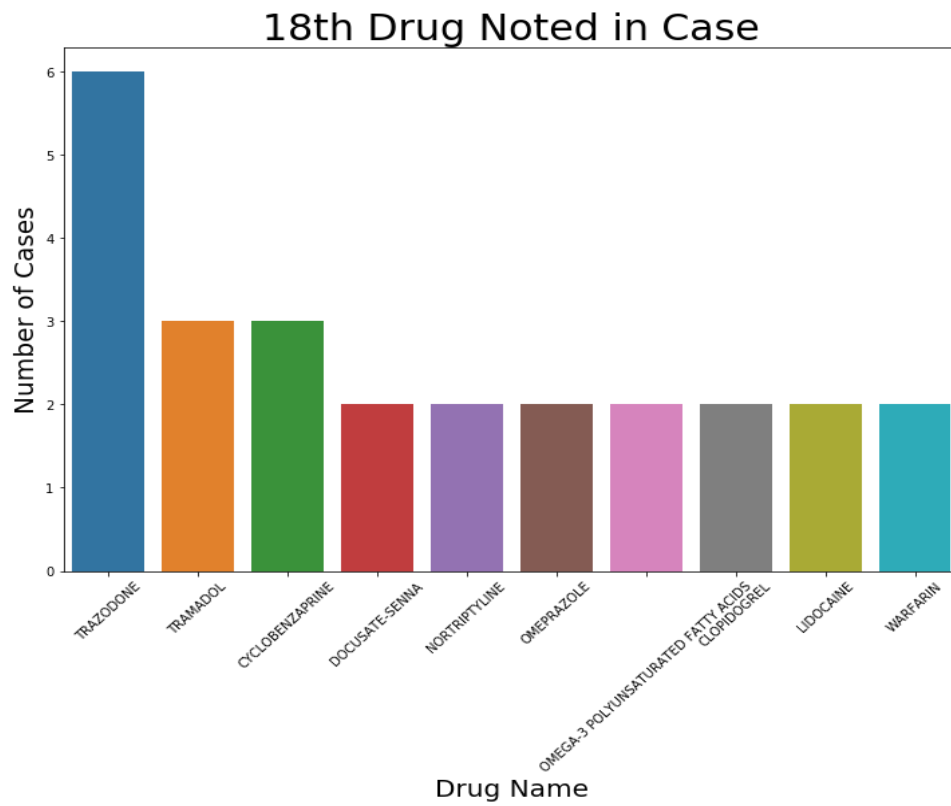




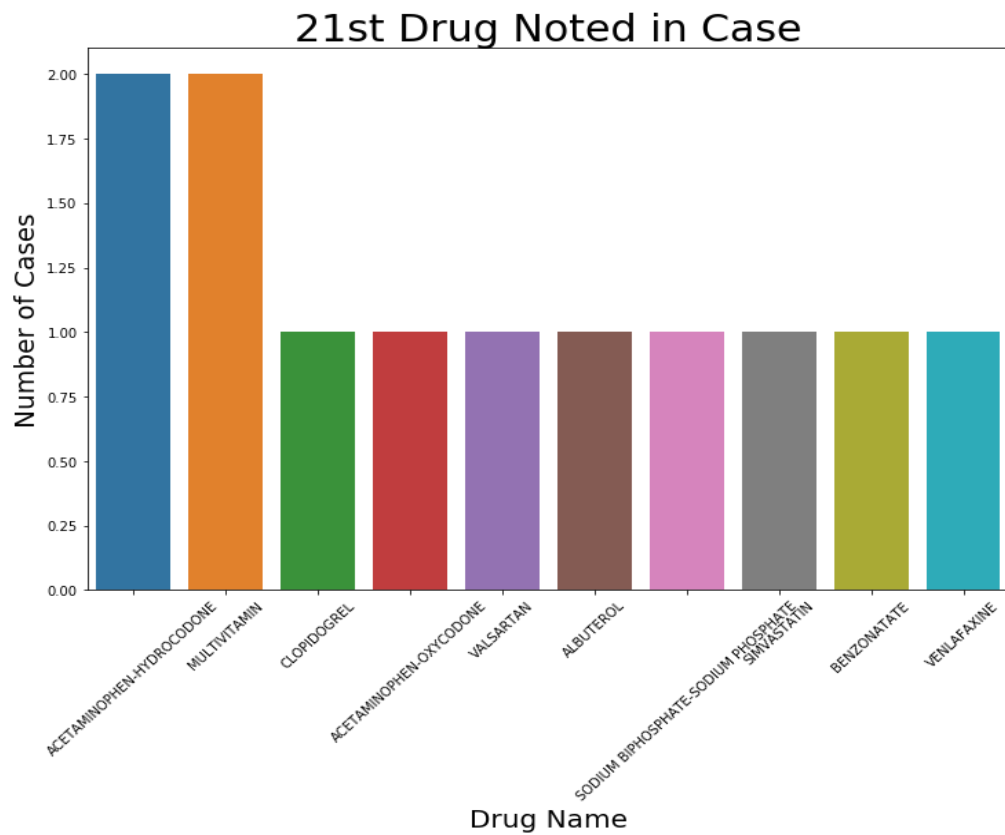
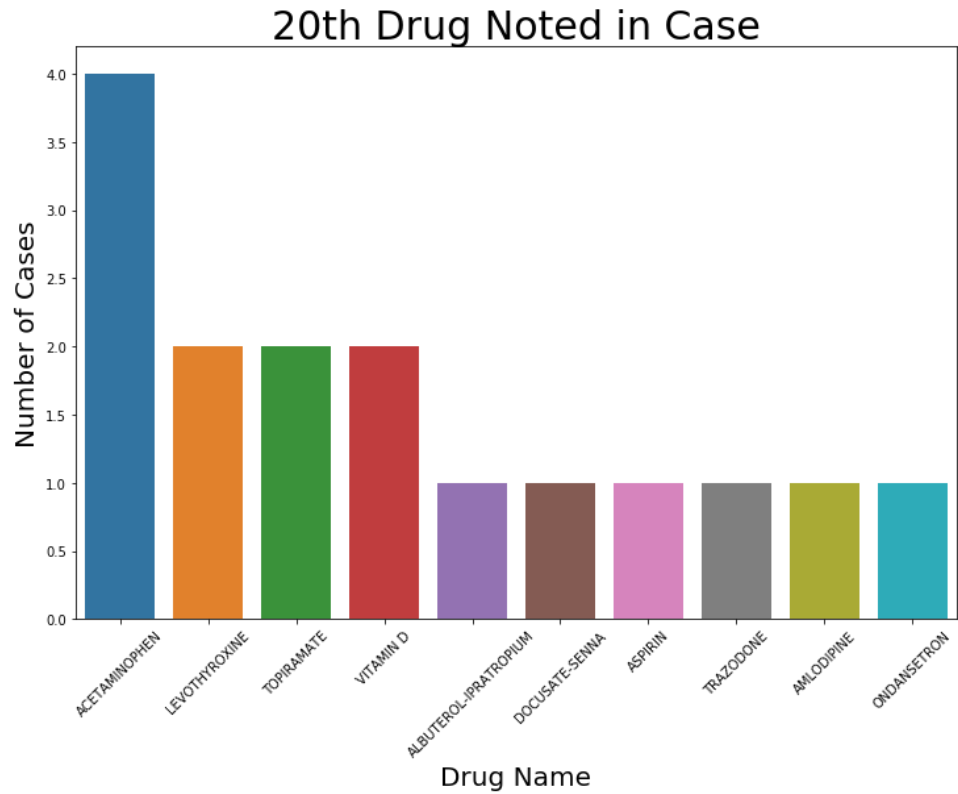


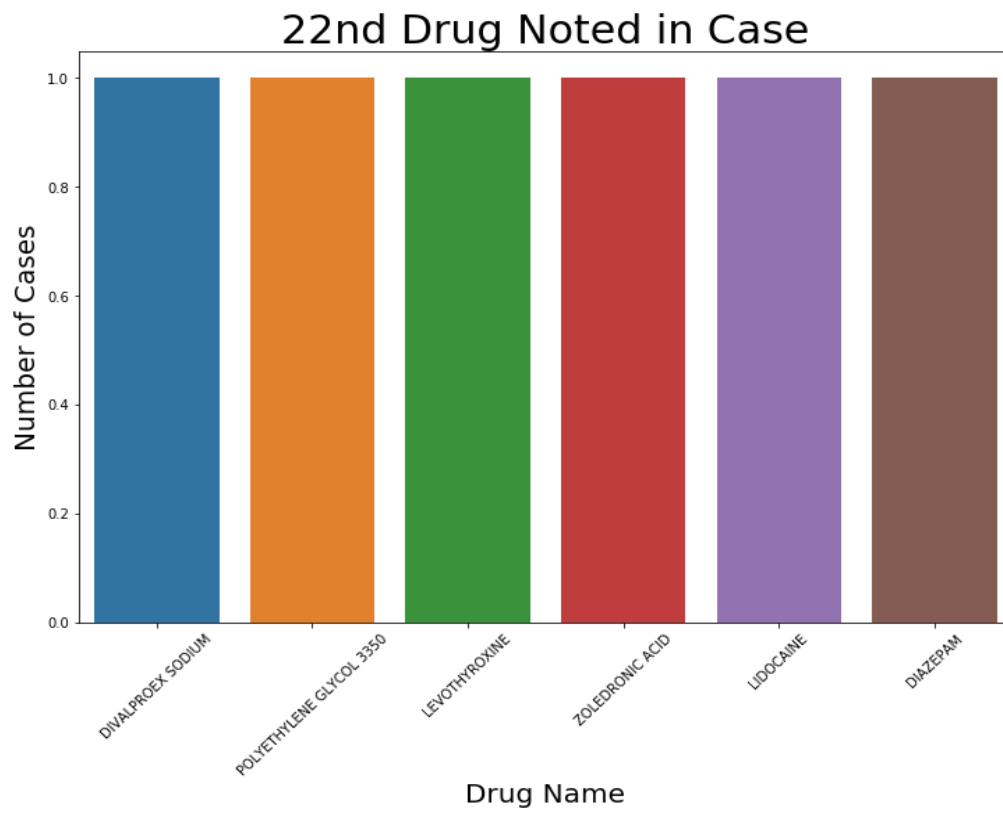




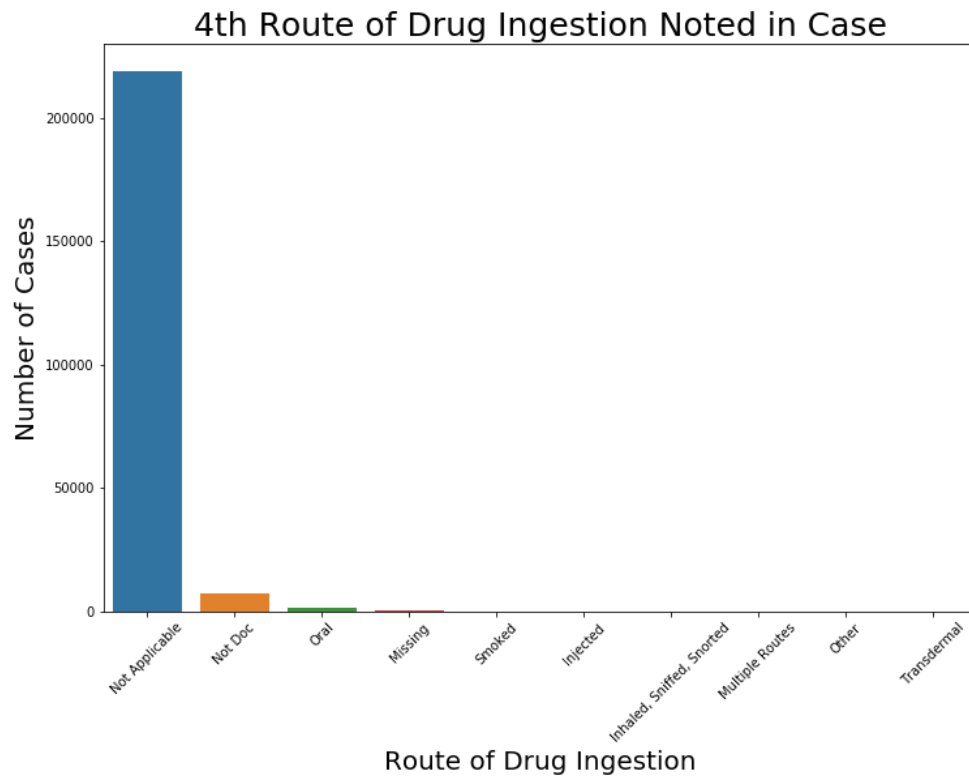


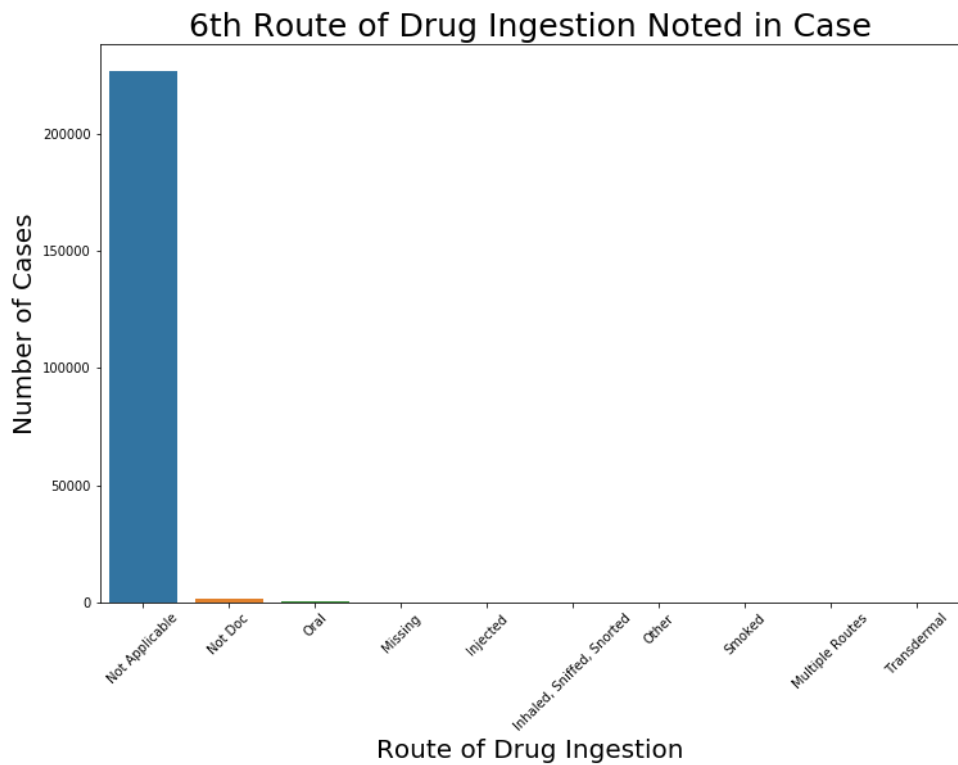
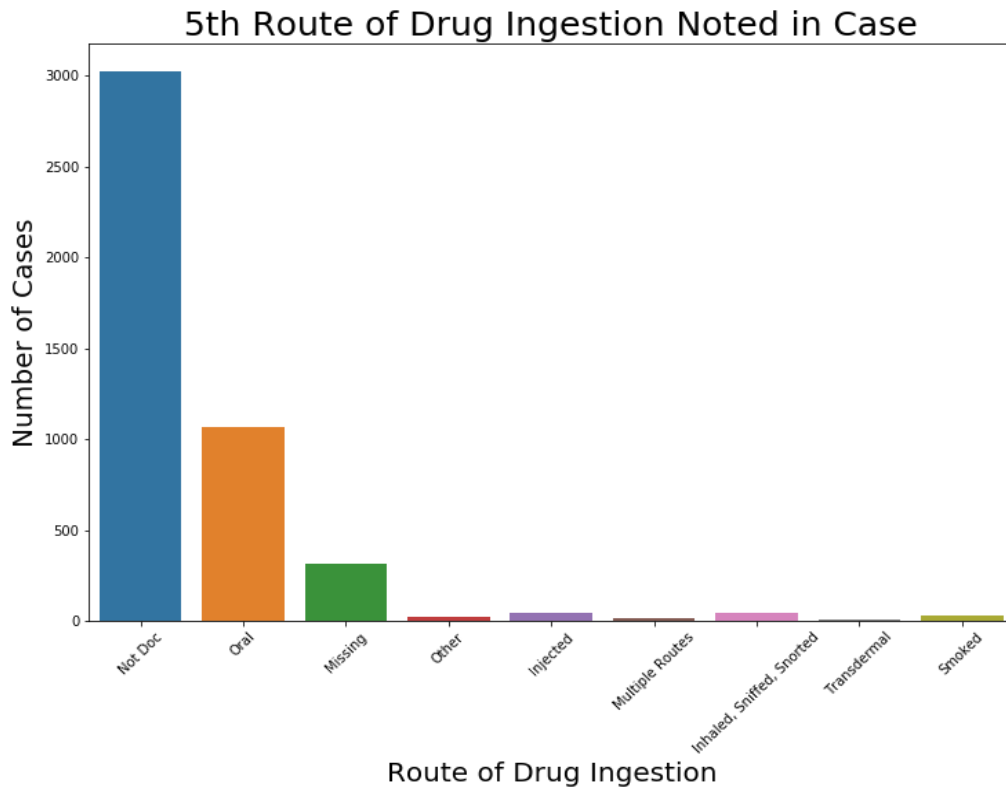


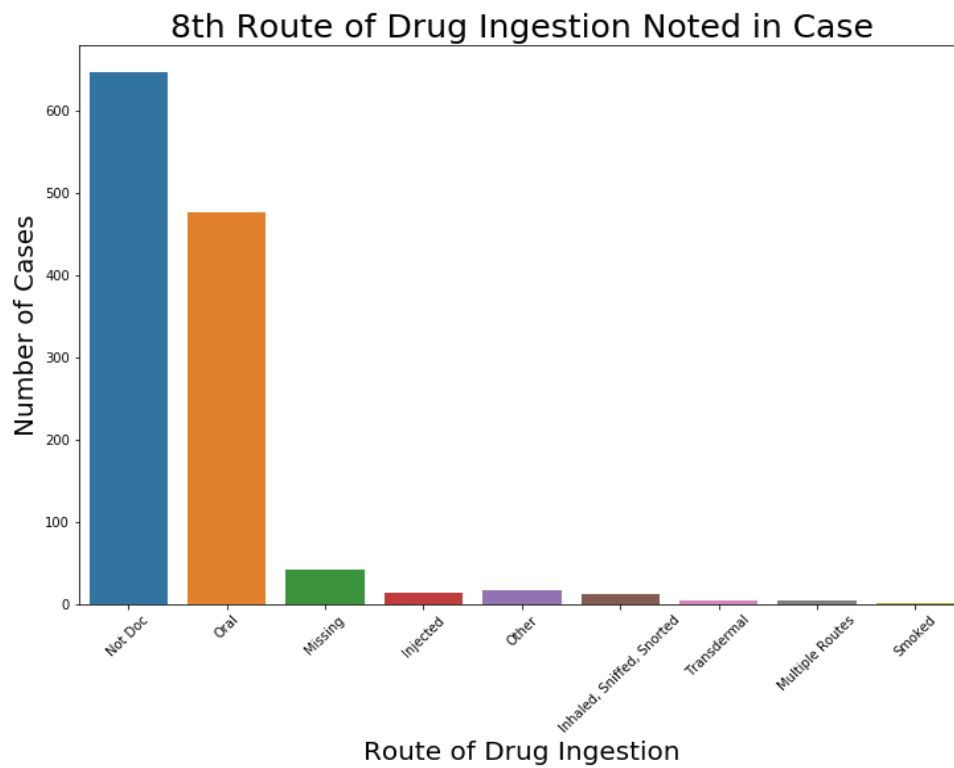
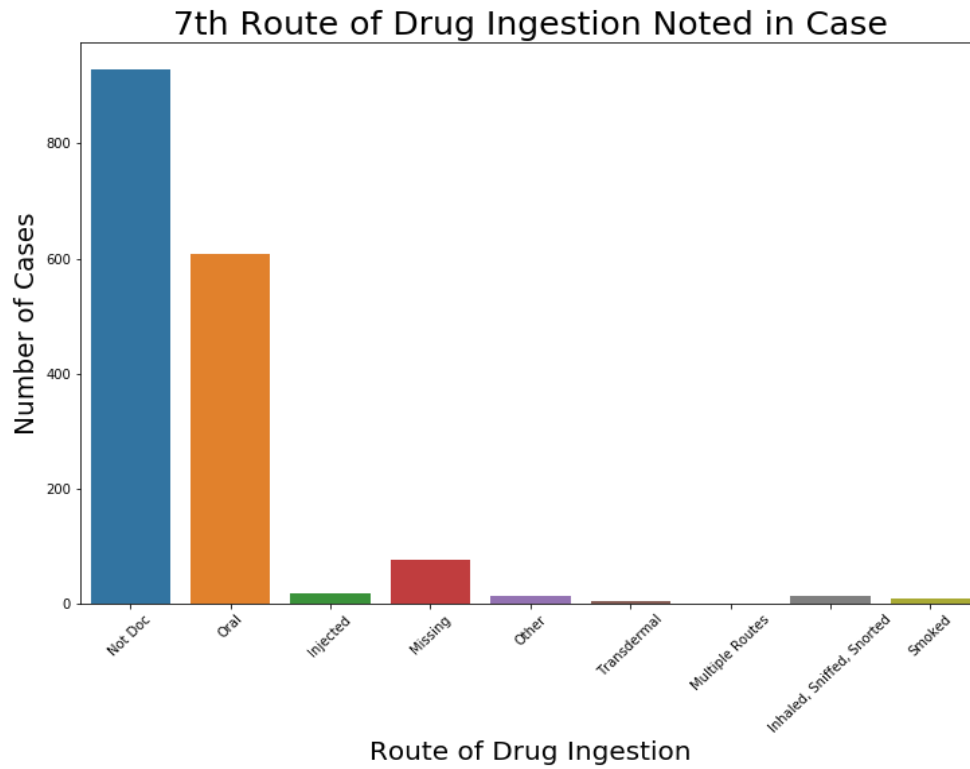


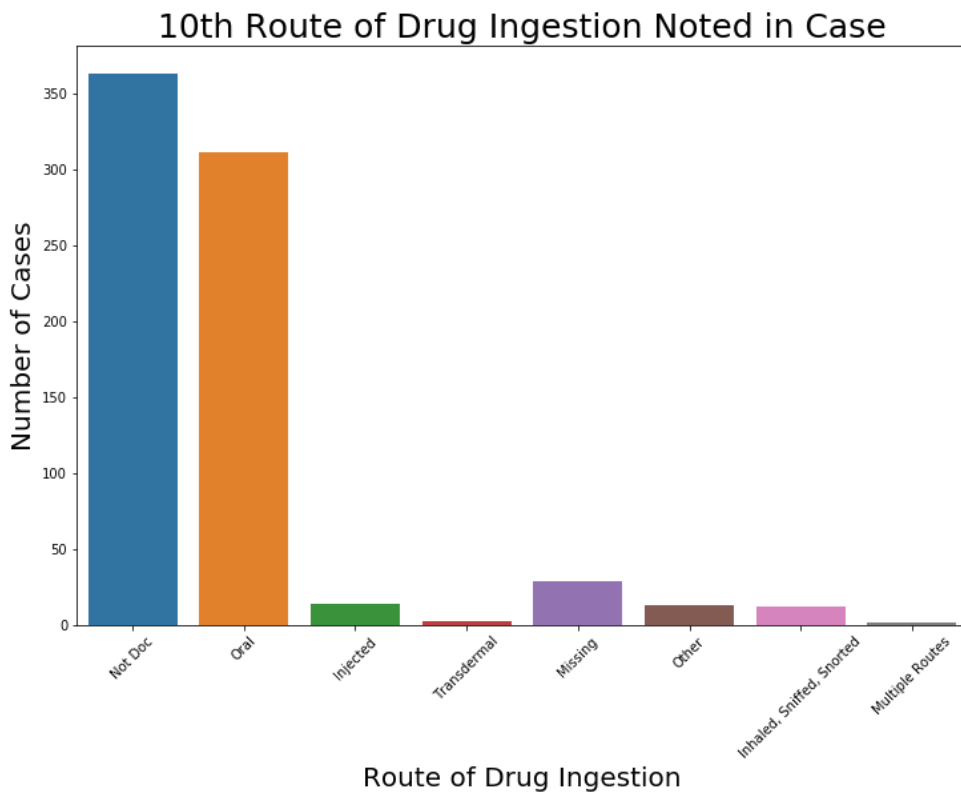
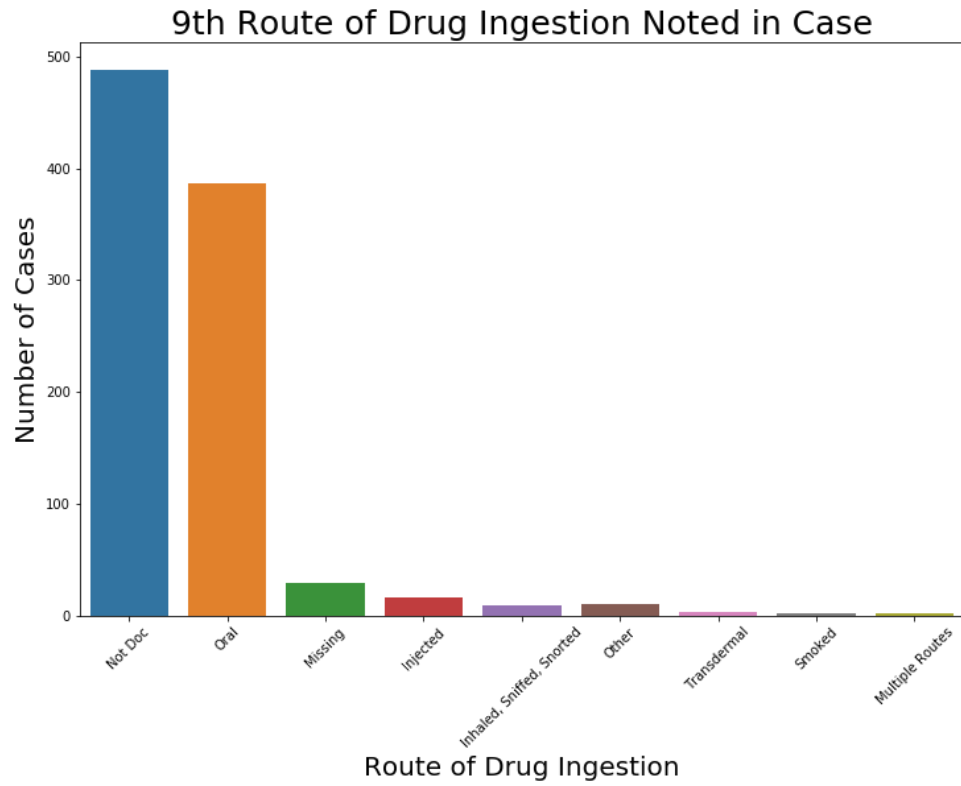


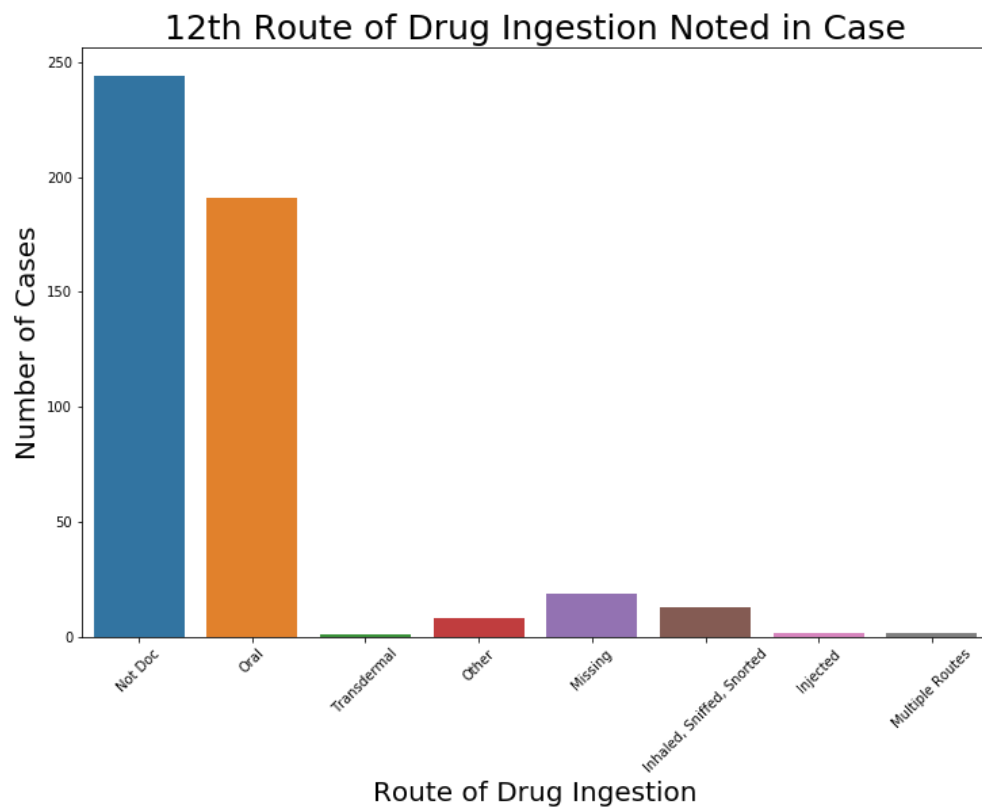
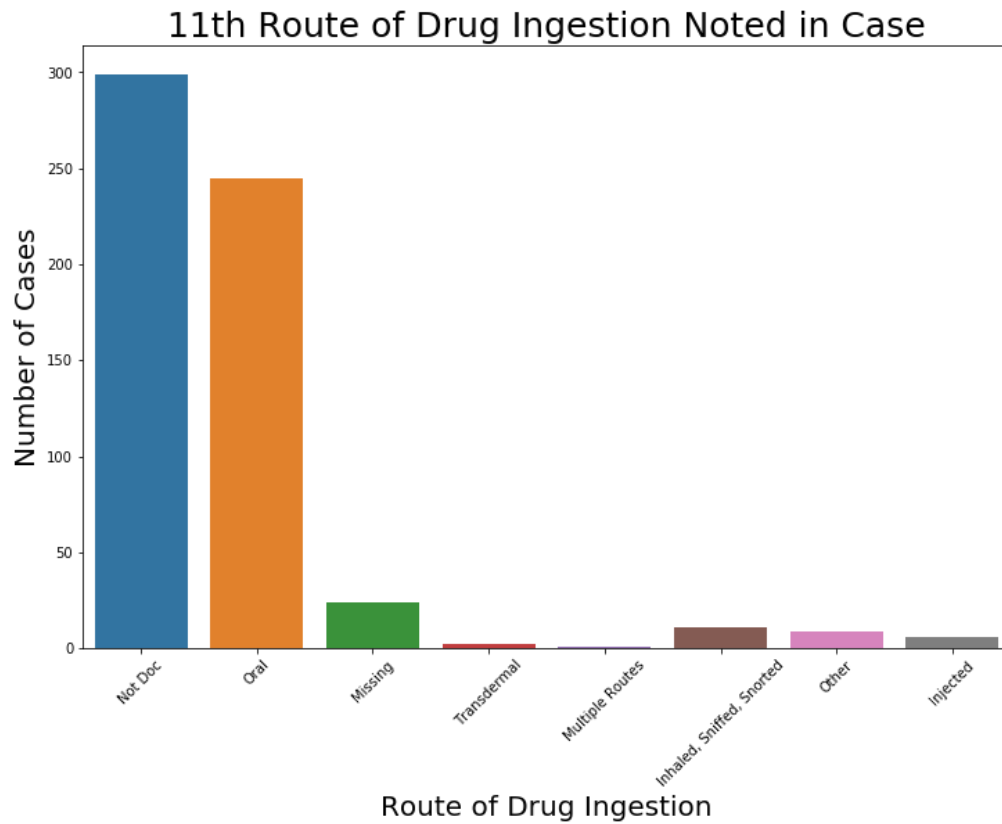
## Appendix B – Graphs the Routes of Ingestion Involved in Cases



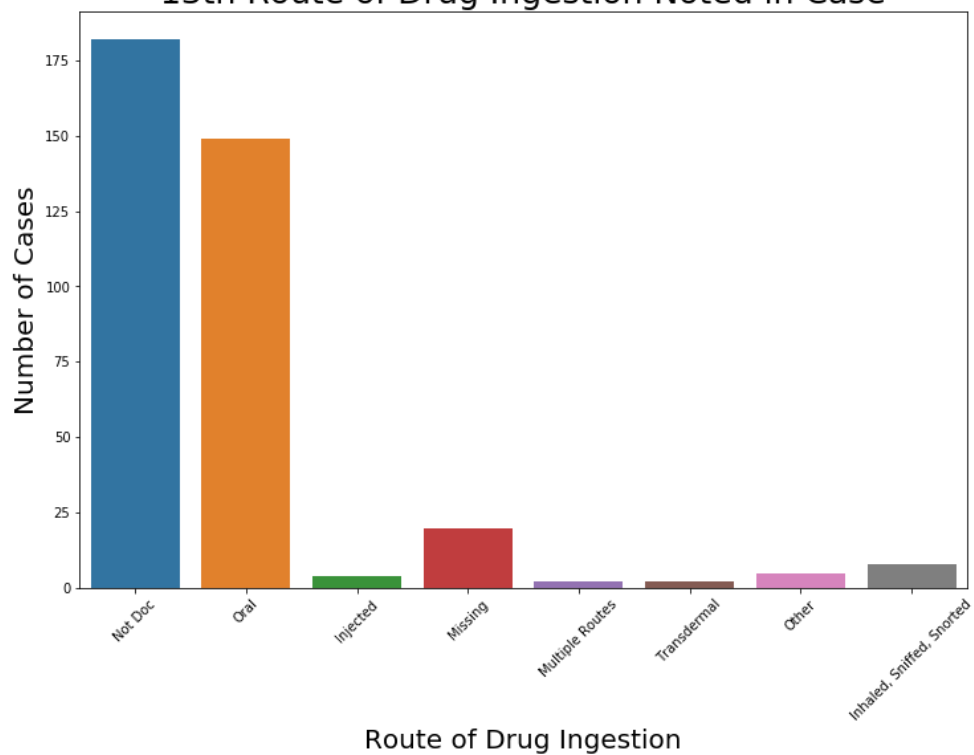




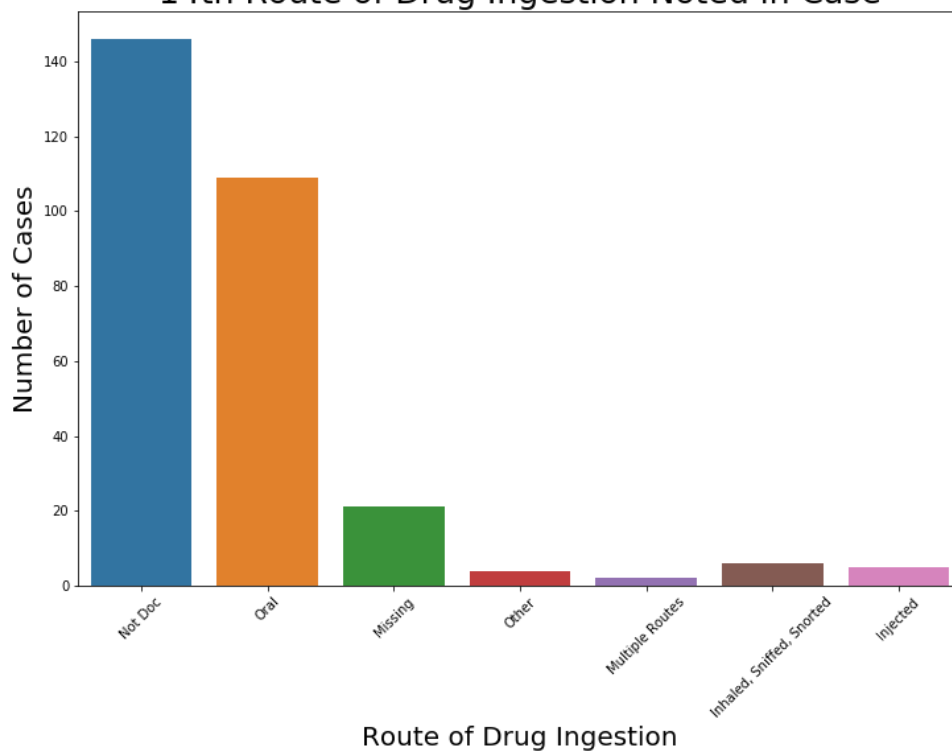




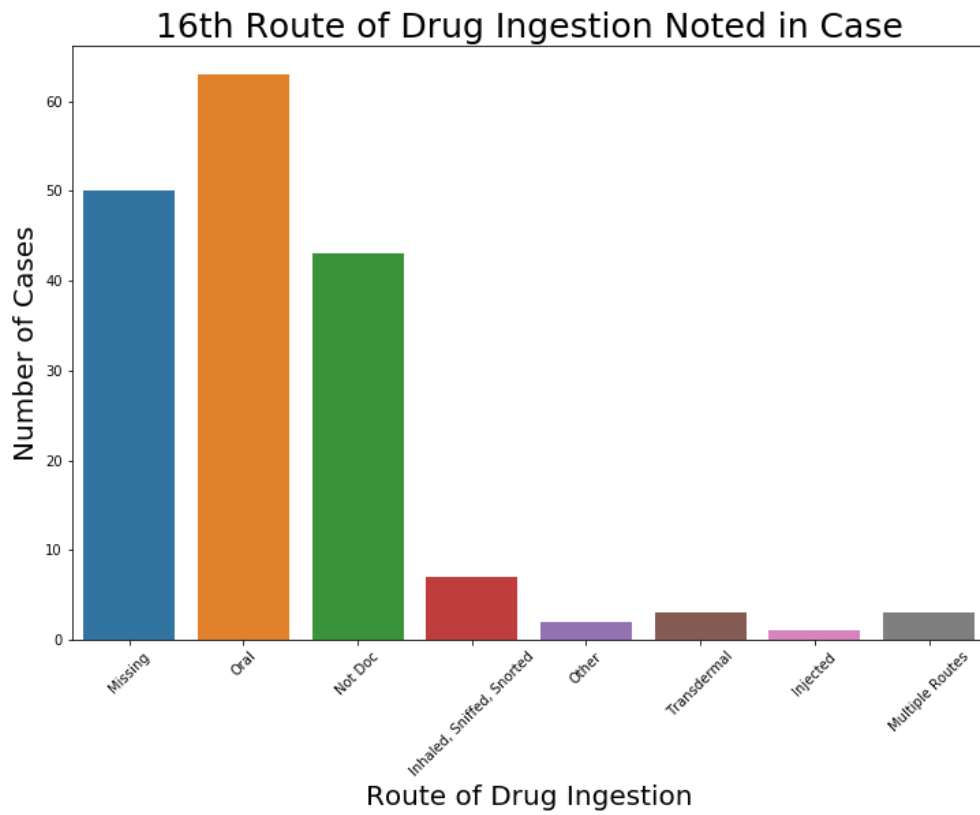
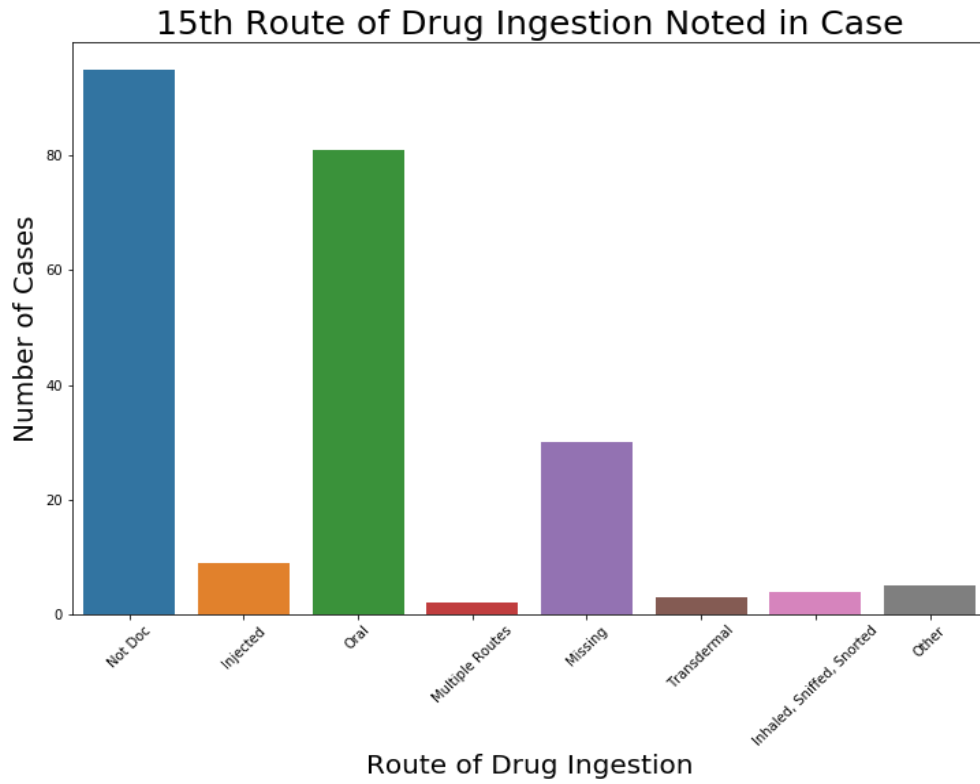
### 13th Route of Drug Ingestion Noted in Case

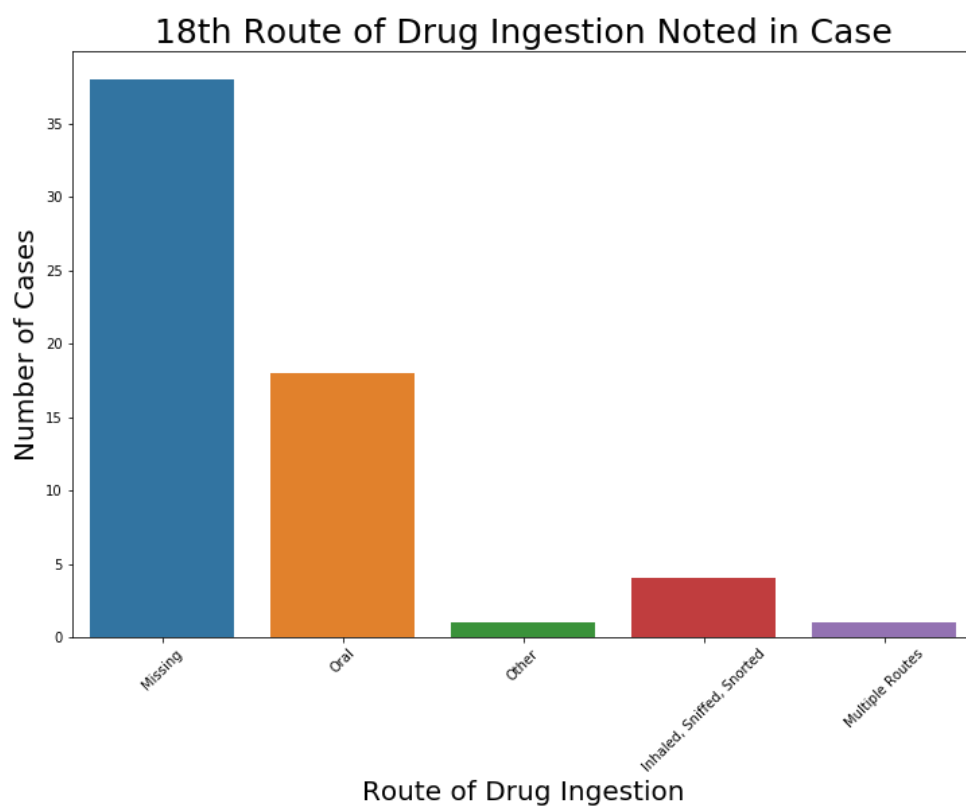
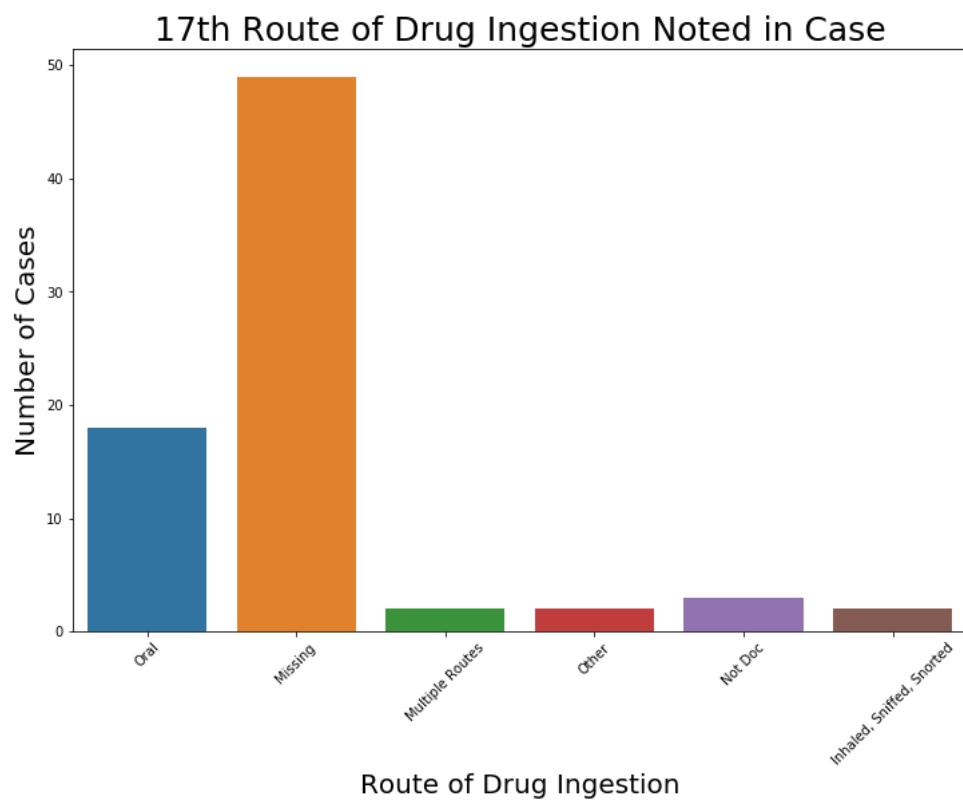


### 14th Route of Drug Ingestion Noted in Case

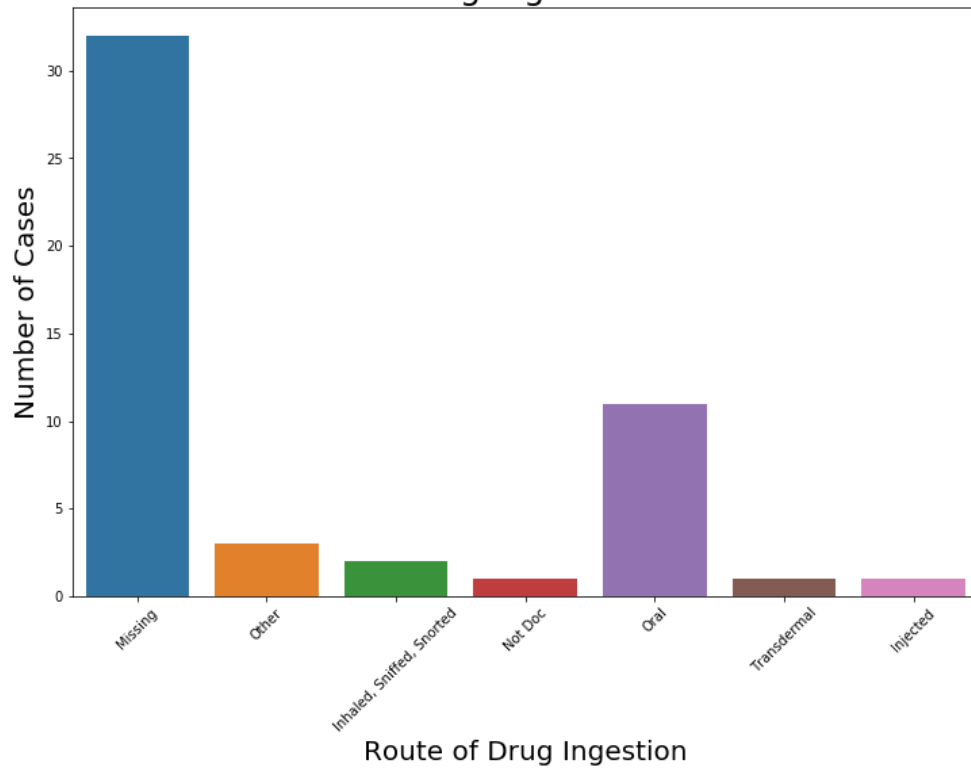




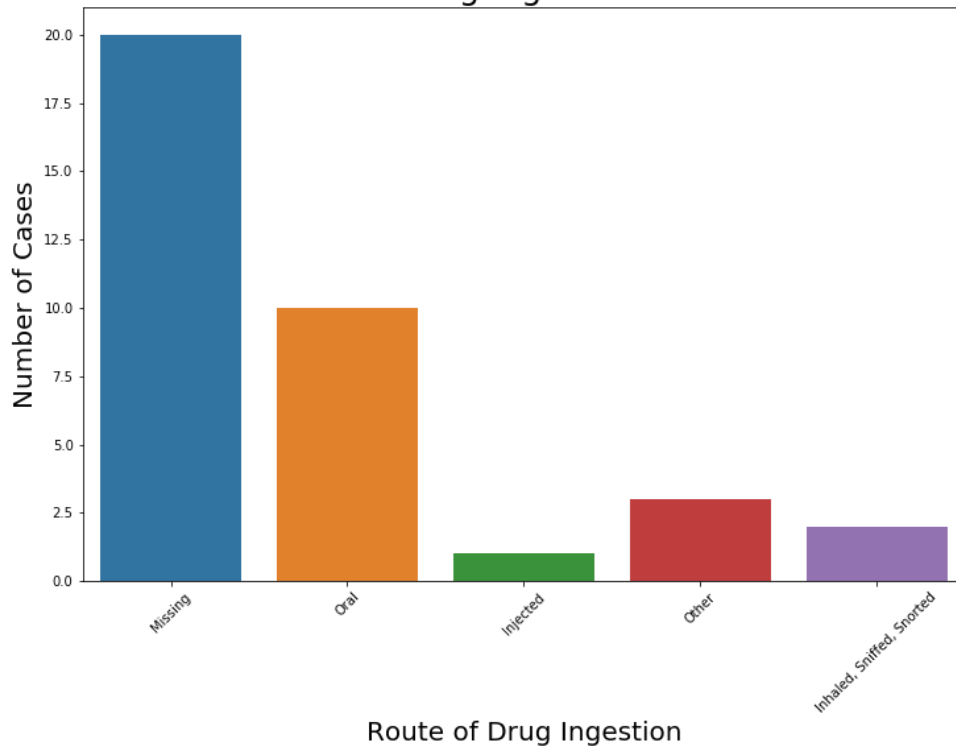




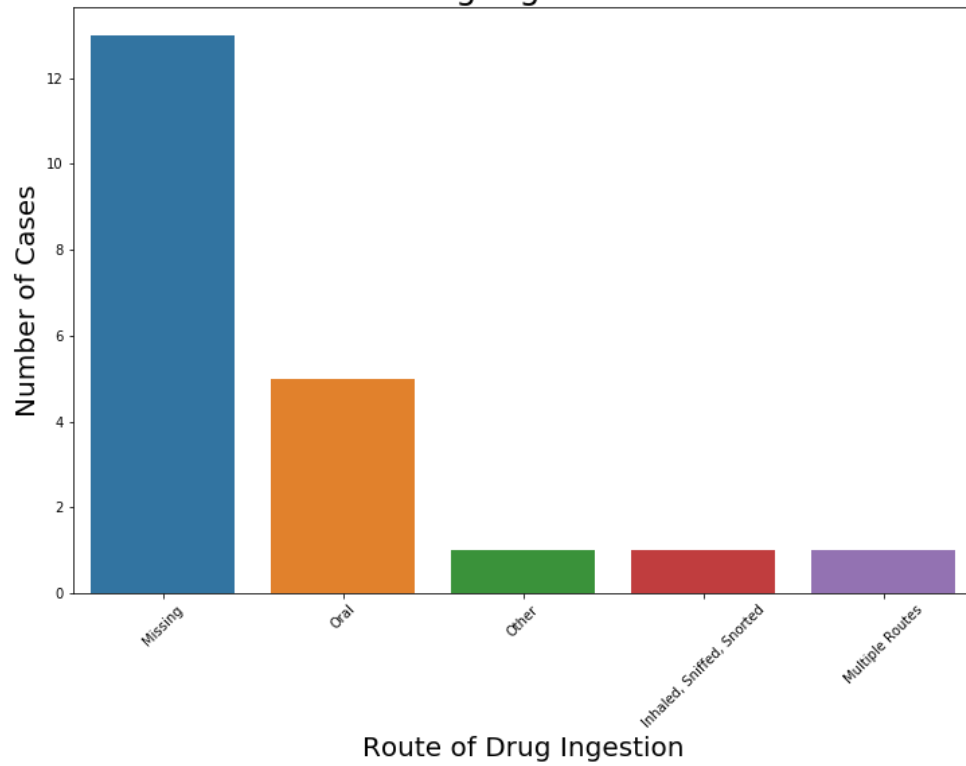
19th Route of Drug Ingestion Noted in Case



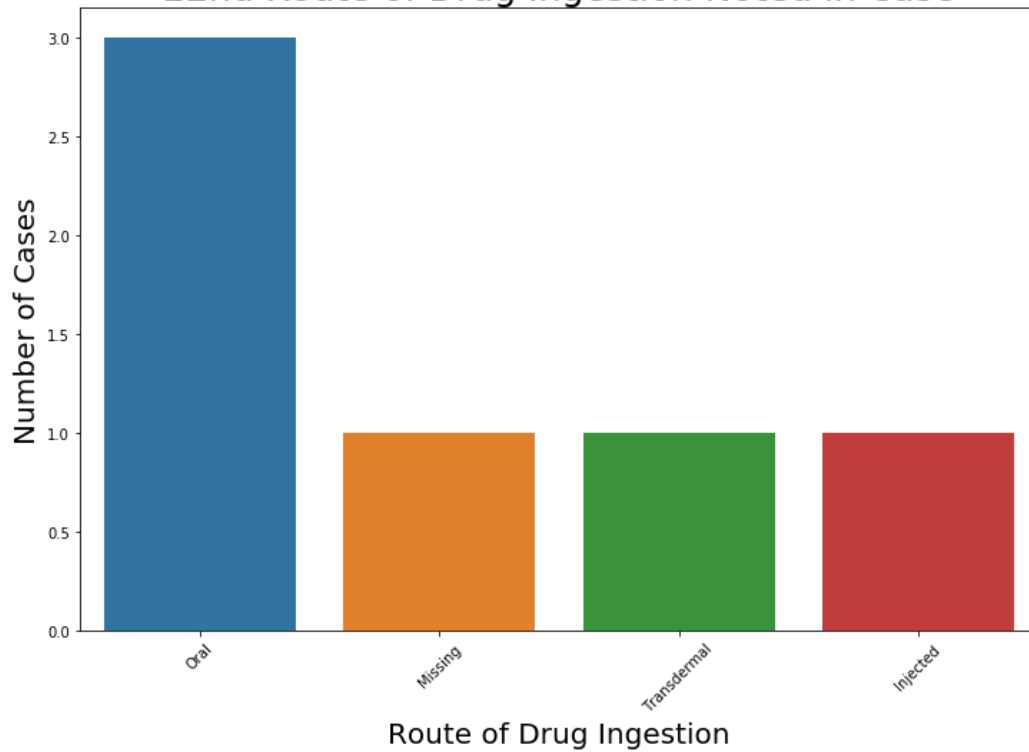
20th Route of Drug Ingestion Noted in Case



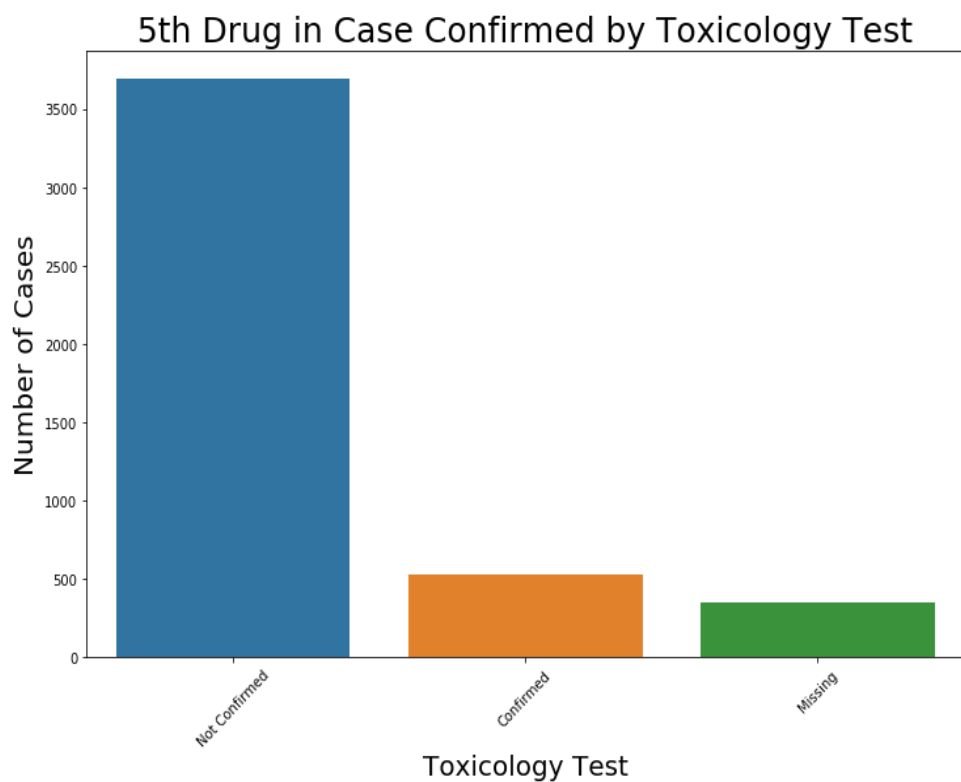
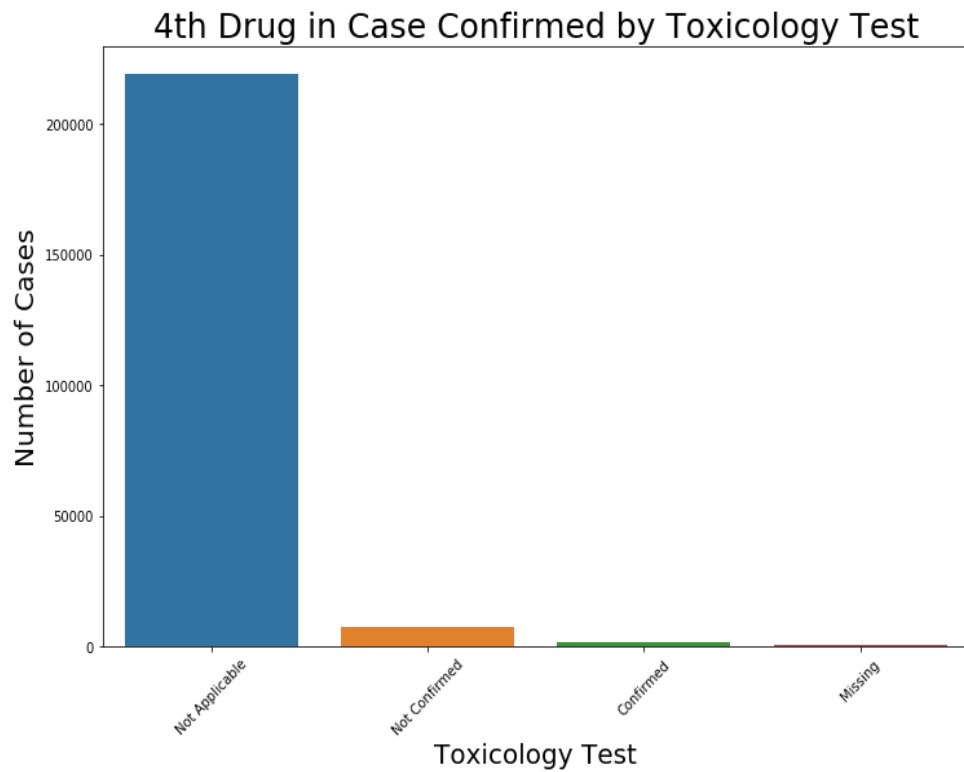
21st Route of Drug Ingestion Noted in Case

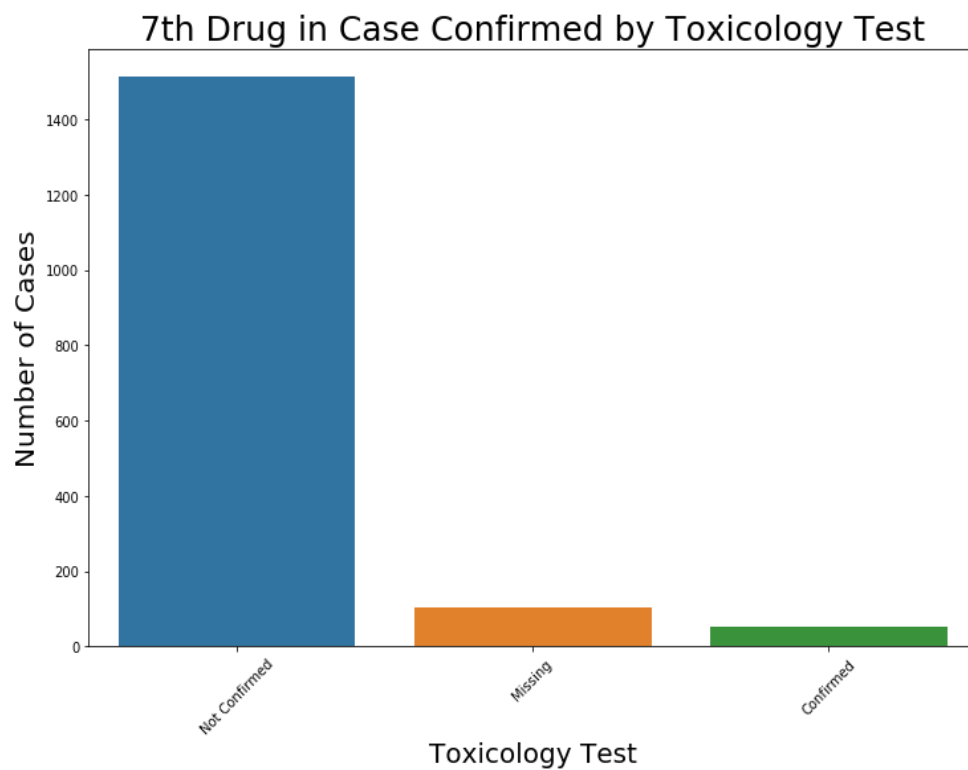
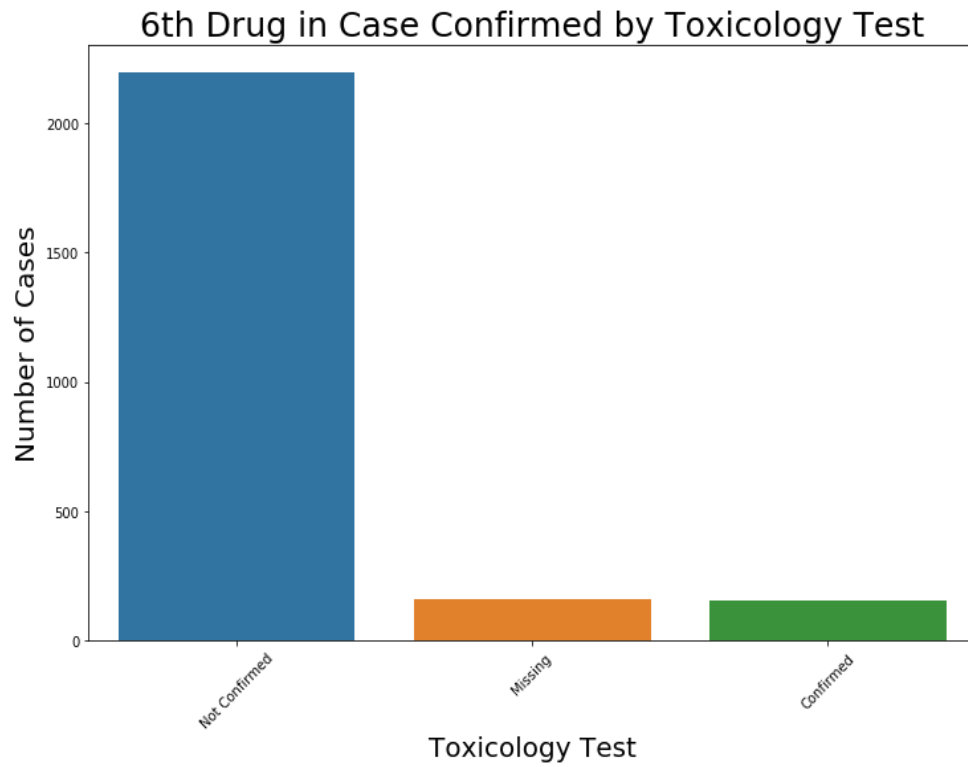


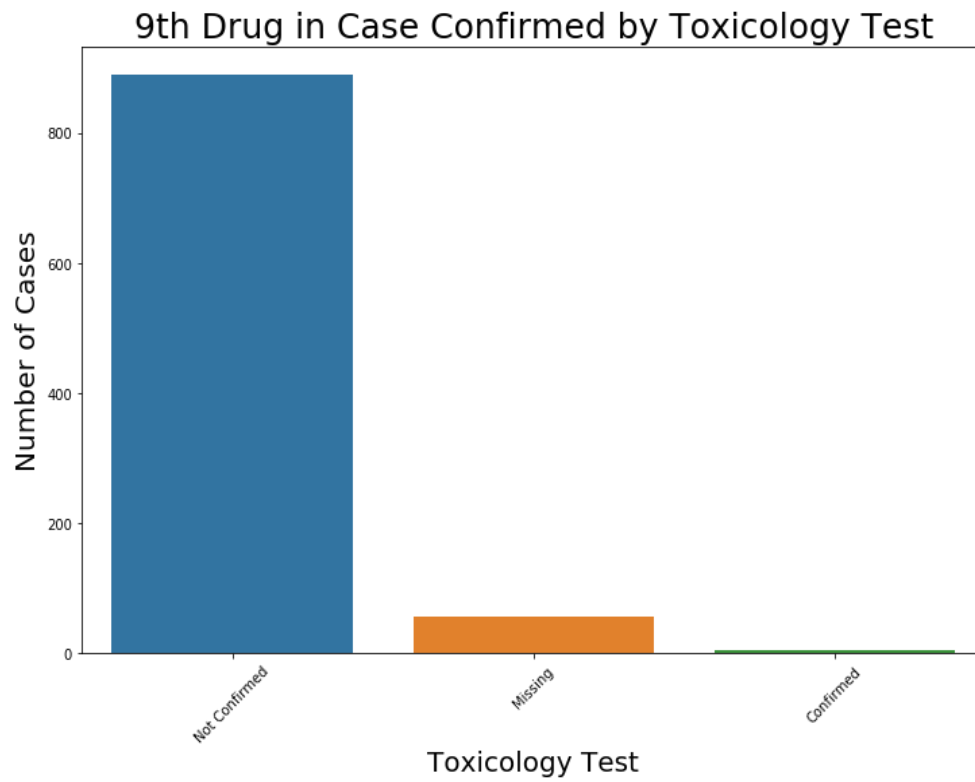
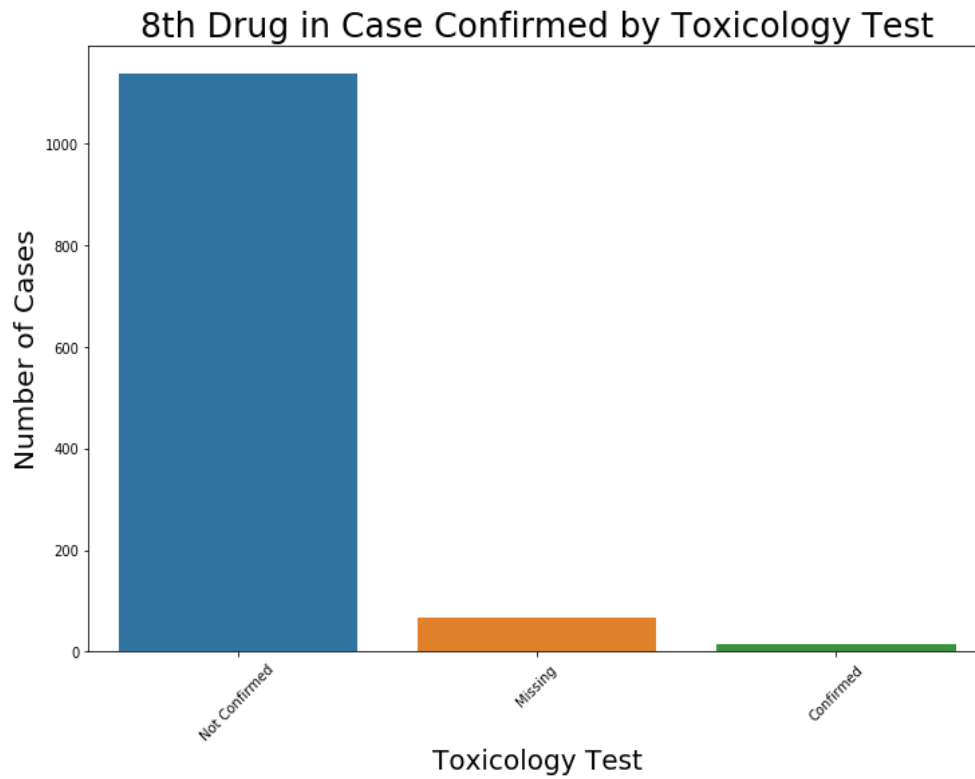
22nd Route of Drug Ingestion Noted in Case



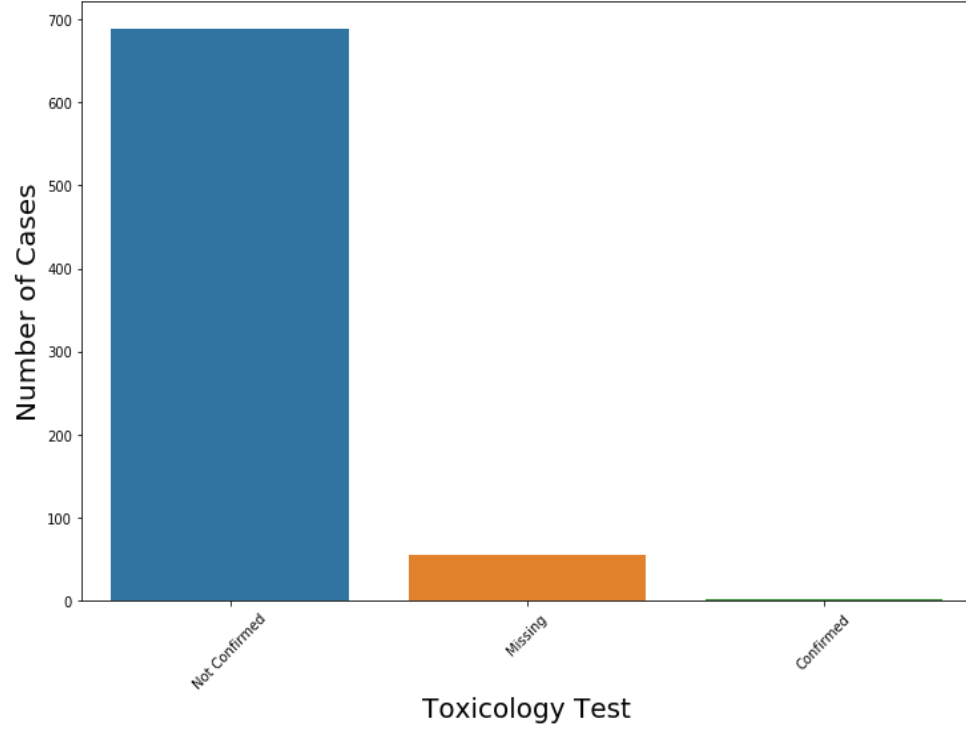
## Appendix C – Graphs of Whether a Toxicology Test was Done



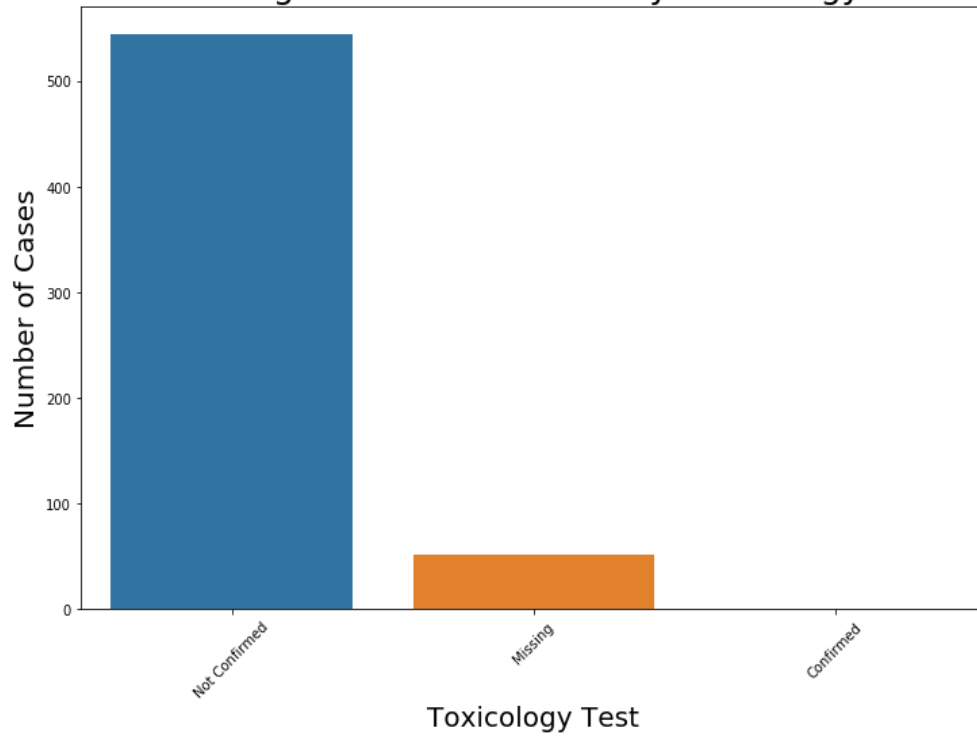




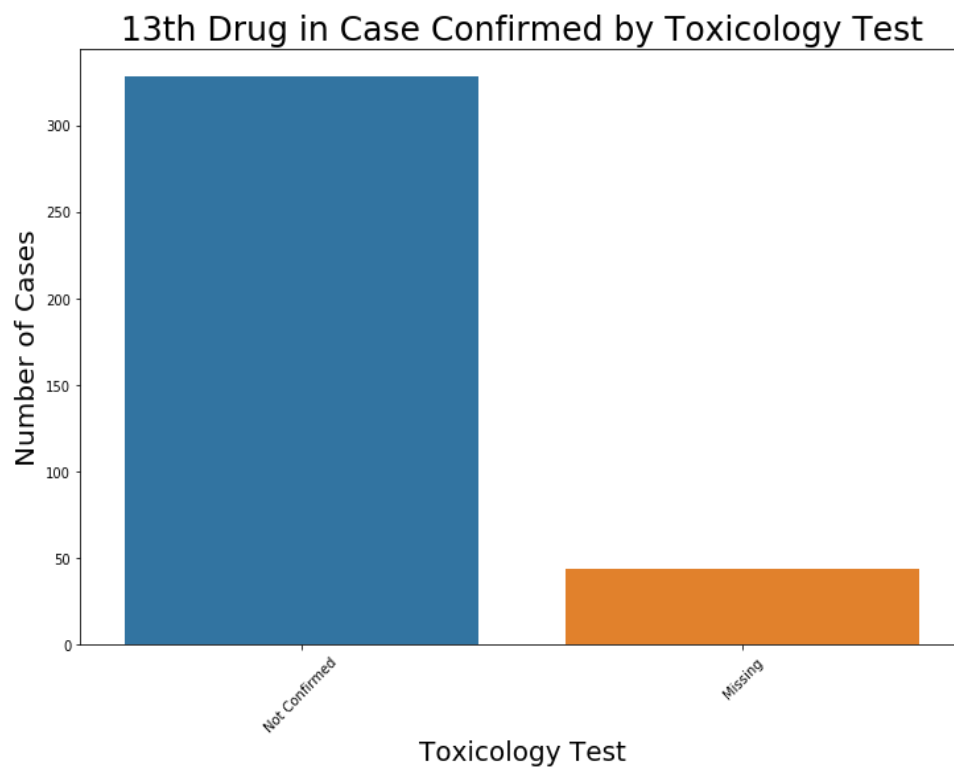
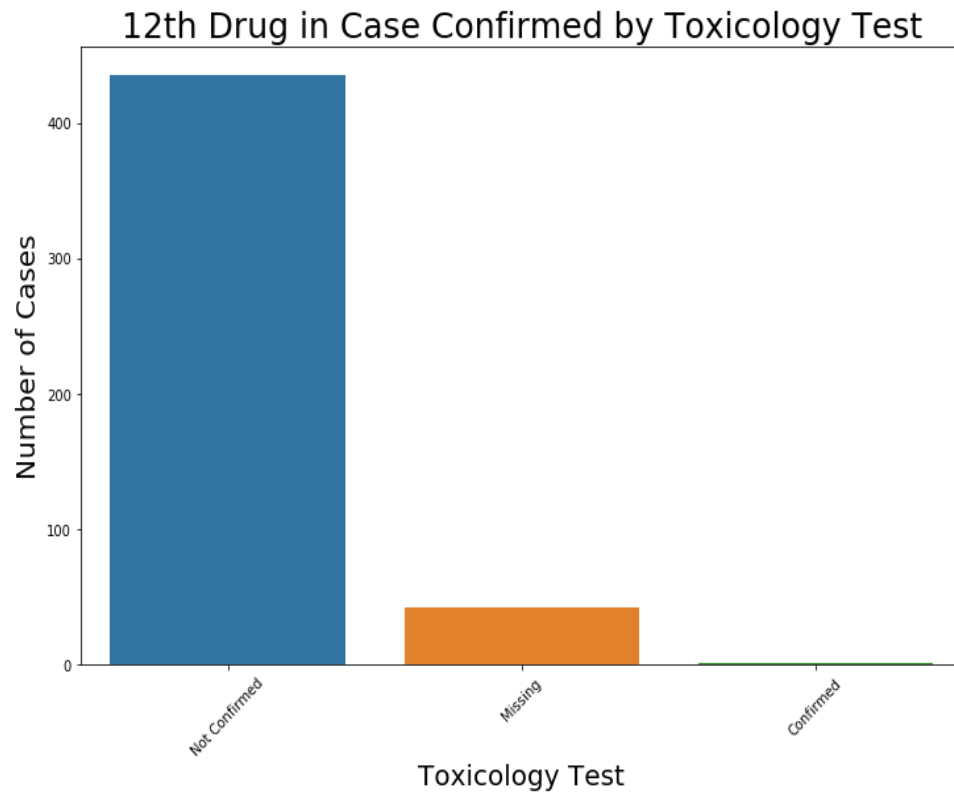
10th Drug in Case Confirmed by Toxicology Test

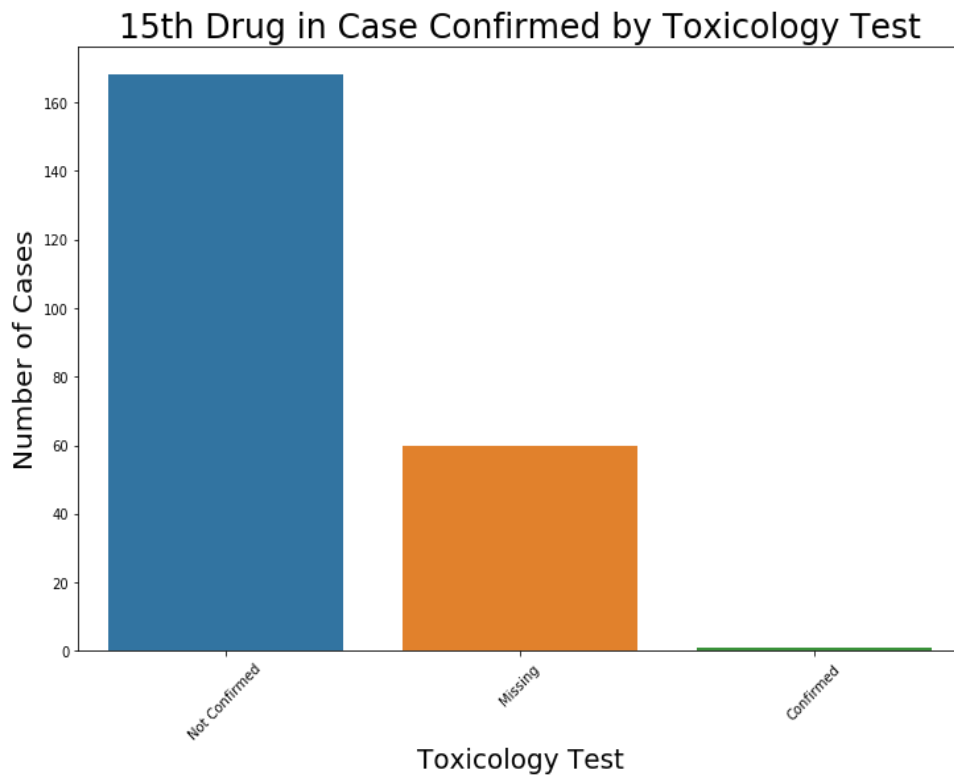
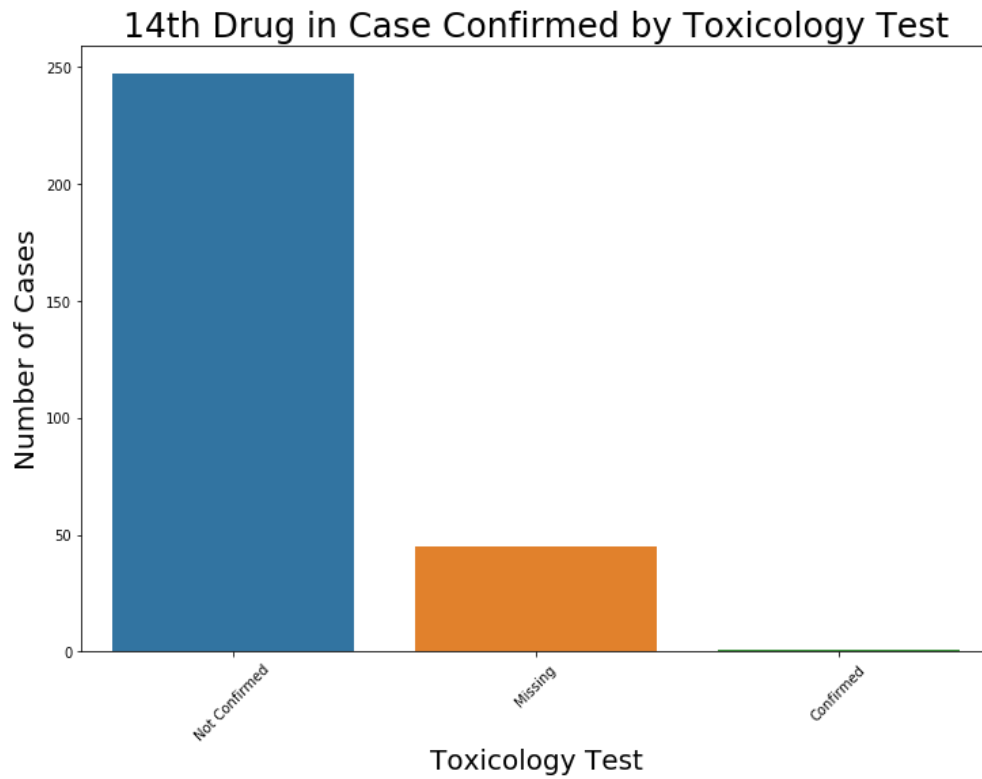


11th Drug in Case Confirmed by Toxicology Test

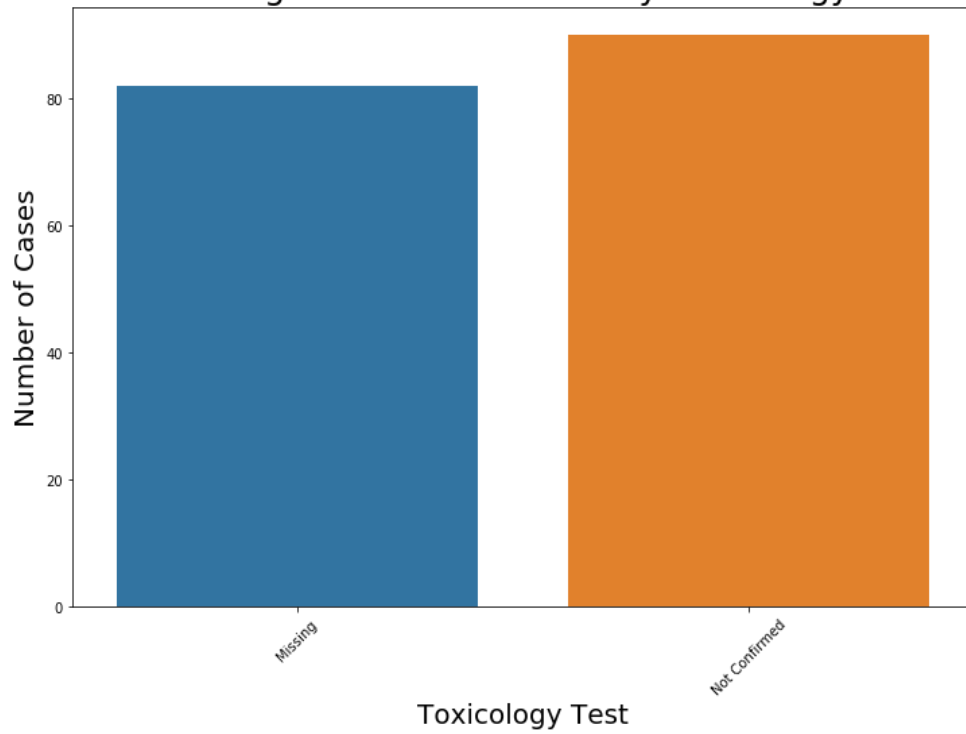




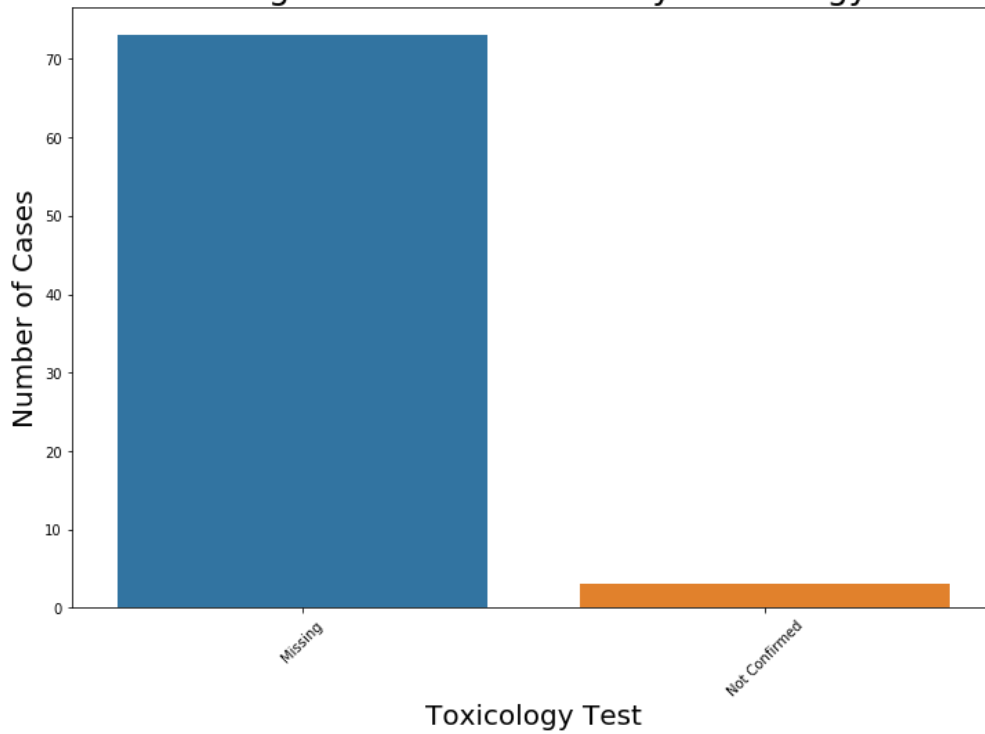


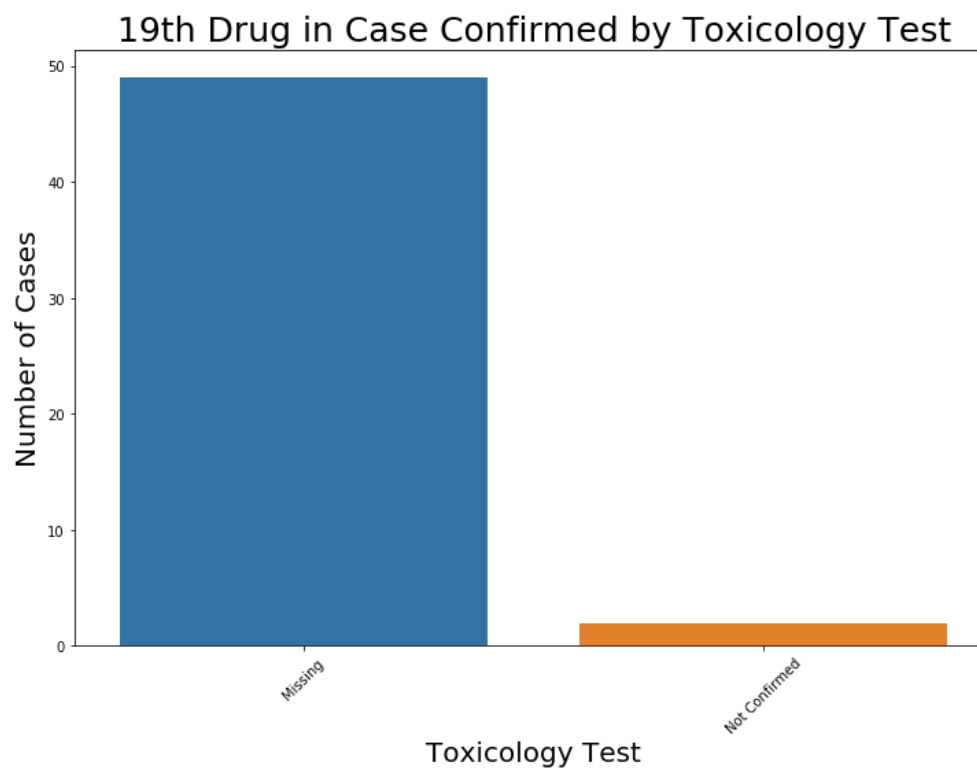
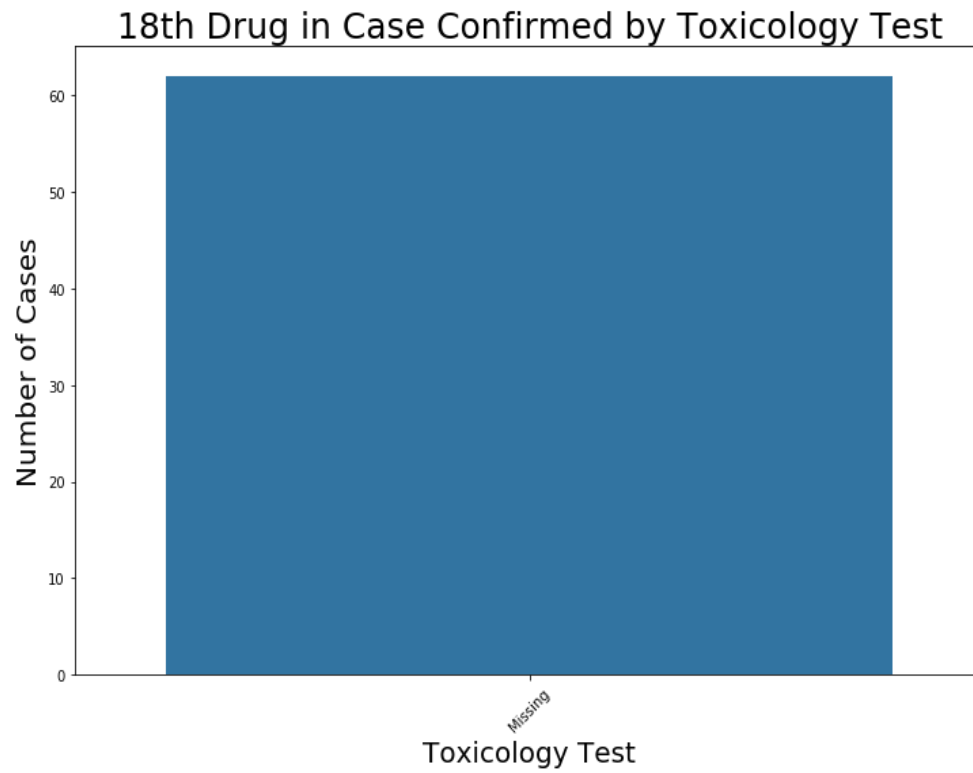


16th Drug in Case Confirmed by Toxicology Test

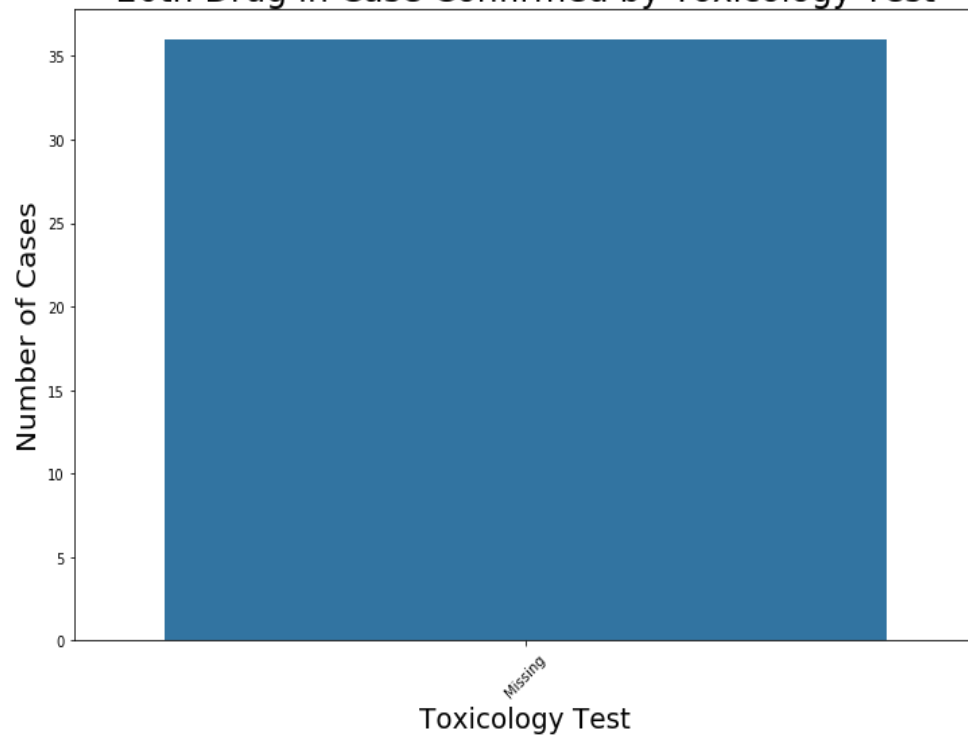


17th Drug in Case Confirmed by Toxicology Test





20th Drug in Case Confirmed by Toxicology Test



21st Drug in Case Confirmed by Toxicology Test

