REPORT

EXPLORATORY ANALYSIS ON DRUG RELATED DEATHS

DATA SET LINK	https://www.kaggle.com/adarsh4u/drug-related-deaths
TEAM MEMBERS	Name: C.Diya SRN: PES1201700246 Email ID: diyasateesh96@gmail.com Contact Number: +91 8861969033
	Name: Namrata R SRN: PES1201700921 Email ID: namrata.ajjampur@gmail.com Contact Number: +91 7259601916

INTRODUCTION

Data set chosen:

https://www.kaggle.com/adarsh4u/drug-related-deaths

The above data set gives information about the drug induced deaths caused in a region over a span of 5 years, 2012-2018. It gives information about the different deaths causing the death, the immediate cause of death, the date and location of the death. Other than this it also specifies the age, gender, race and residence region of the casualties. The data consists of 4082 rows and 32 columns with each row giving details of the casualty.

The dataset specifies about 12 different drugs and has a column for other drugs in case the death was not caused due to the previous 12 drugs.

Why we chose this dataset?

This dataset has apt number of rows and columns to create a sample for analysis.

This is a pressing issue which needs to be discussed and dealt with. It is very important to analyse and make acute observations about each and every factor that may affect the number of deaths caused due to substance abuse and hence use this to reduce the proportion of deaths caused due to drugs.

Also, the dataset was not very well formatted, had many missing values and some repetitions. By cleaning this dataset, it would become easier to analyse and interpret from this data

AIM:

To predict the different factors that are most likely to cause drug induced deaths and hence use this to reduce the latter in the future.

Question to be asked:

Can we isolate a target group from test samples by analysing the chosen data frame based on various parameters? And if so, what are the different factors affecting it?

1) Which drug causes the greatest number of deaths in that region?
2) Most affected age group in that region
3) Percentage distribution of each race in that region
4) Change in number of drug induced deaths over a span of five years in that region
5) Categorizing drug induced deaths in that region based on gender.
6) Which year had the greatest number of deaths?
7) Isolating a group most affected by drugs based on age, gender, race
8) Analysis of specific cause of death.
9) Variation of drug induced deaths for each age group over time
10) Finding any anomaly in the age group, median age
11) What are the immediate causes of death?

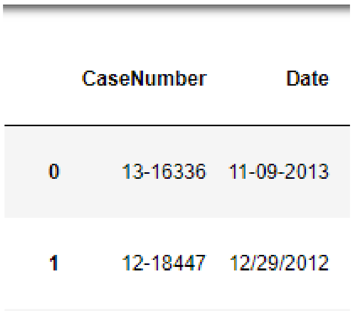
PROCESSING/CLEANING THE DATA

1) Formatting the 'Date' column:

On observation, it was found that the "Date" column in the data set did not have a standard format i.e., 01/14/2014 and 01-14-2014 (this is an example of the two different ways that the dates were entered into the data frame). It is important to keep it in a standardized format and follow a uniform format to make it easier and more efficient for further analysis of the based-on timeframe of death.

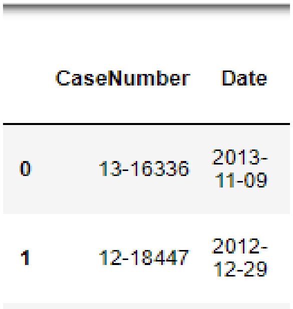
Therefore, we changed the date to a definite standard(conventional) format of yyyy-mm-dd, e.g. 2014-01-14.

BEFORE



	CaseNumber	Date
0	13-16336	11-09-2013
1	12-18447	12/29/2012

AFTER



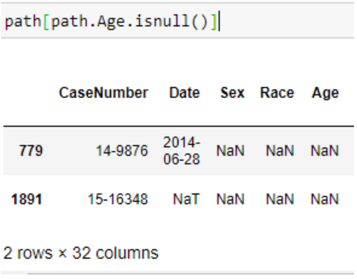
	CaseNumber	Date
0	13-16336	2013-11-09
1	12-18447	2012-12-29

2) Filling in the missing values for "Age" column:

We observed that there were two missing (NaN) values in the "Age" column. Following this, we replaced them by the median of the "Age" values of the entries. Using the median rather than the mean of the values would relatively decrease the skewness due to any present outliers or extreme values

The median for all the age values was found to be 42.0 and the NaN values of this column have been replaced by 42.0

BEFORE

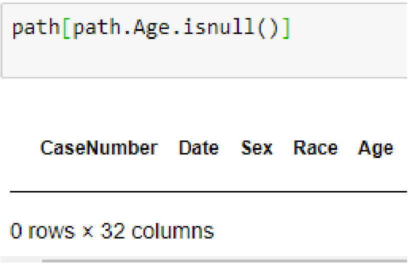


```
path[path.Age.isnull()]
```

	CaseNumber	Date	Sex	Race	Age
779	14-9876	2014-06-28	NaN	NaN	NaN
1891	15-16348	NaT	NaN	NaN	NaN

2 rows × 32 columns

AFTER



```
path[path.Age.isnull()]
```

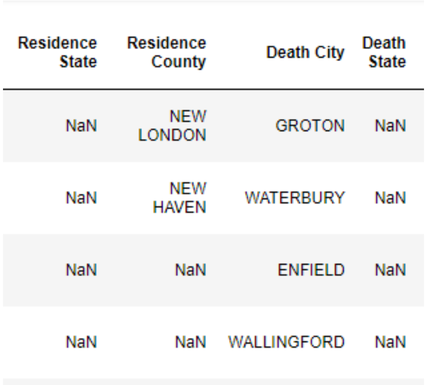
CaseNumber	Date	Sex	Race	Age

0 rows × 32 columns

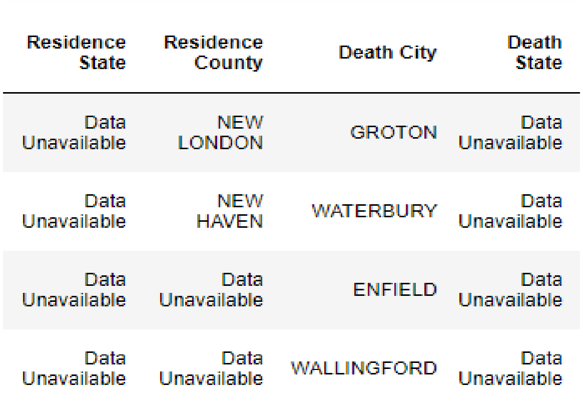
3) Dealing with missing values of the qualitative data:

To enhance the visual representation and to make the data set more comprehensive, the missing quantitative values in columns like Resident Location, County, City etc. were filled “Data Unavailable” to make it easier for the reading and inferring the data set.

BEFORE **AFTER**



Residence State	Residence County	Death City	Death State
NaN	NEW LONDON	GROTON	NaN
NaN	NEW HAVEN	WATERBURY	NaN
NaN	NaN	ENFIELD	NaN
NaN	NaN	WALLINGFORD	NaN

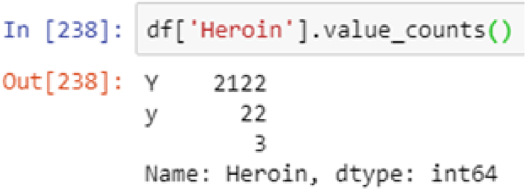


Residence State	Residence County	Death City	Death State
Data Unavailable	NEW LONDON	GROTON	Data Unavailable
Data Unavailable	NEW HAVEN	WATERBURY	Data Unavailable
Data Unavailable	Data Unavailable	ENFIELD	Data Unavailable
Data Unavailable	Data Unavailable	WALLINGFORD	Data Unavailable

4) Standardizing the format and cleaning rows for each drug:

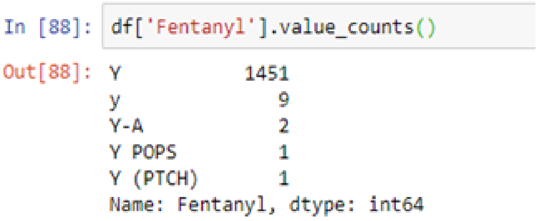
It can be seen that for each column specific to a drug, if the drug has caused the death then ‘Y’ has been entered, but on further observation of the entries of these columns, most of them weren’t in this standard format i.e. some had entries such as ‘Y- ‘, ‘Y ‘,’y’,’   ’ and various other strings. Hence, to make the dataset more uniform for extraction of information and subsequent analysis, we decided to clean these columns such that each entry had the value of either ‘YES’ or ’NO’ depending on whether the user consumed the drug or not.

BEFORE



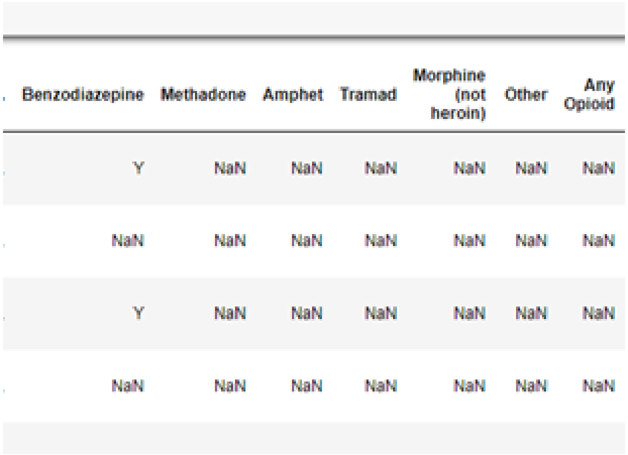
```
In [238]: df['Heroin'].value_counts()

Out[238]: Y     2122
          y       22
                   3
          Name: Heroin, dtype: int64
```



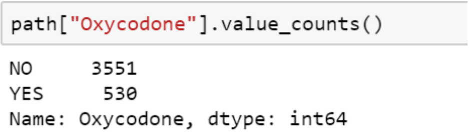
```
In [88]: df['Fentanyl'].value_counts()

Out[88]: Y            1451
         y               9
         Y-A             2
         Y POPS          1
         Y (PTCH)        1
         Name: Fentanyl, dtype: int64
```



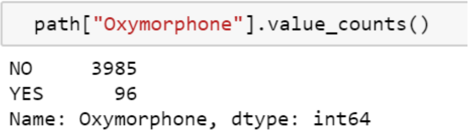
Benzodiazepine	Methadone	Amphet	Tramad	Morphine (not heroin)	Other	Any Opioid
Y	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN
Y	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN

AFTER



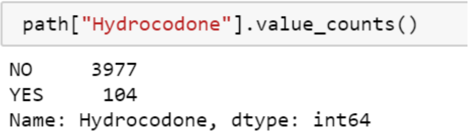
```
path["Oxycodone"].value_counts()

NO     3551
YES     530
Name: Oxycodone, dtype: int64
```



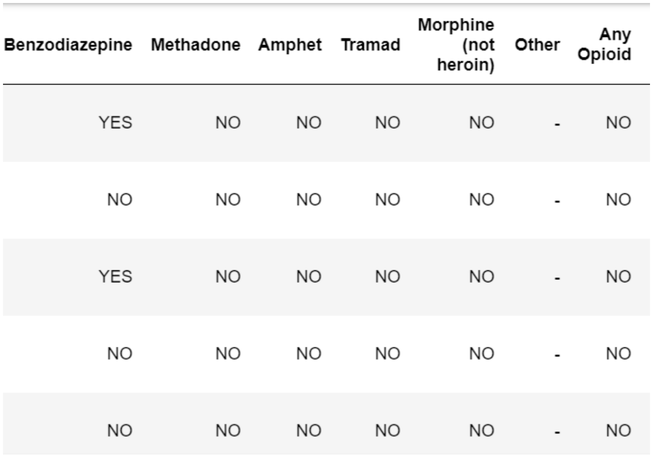
```
 path["Oxymorphone"].value_counts()

NO     3985
YES      96
Name: Oxymorphone, dtype: int64
```



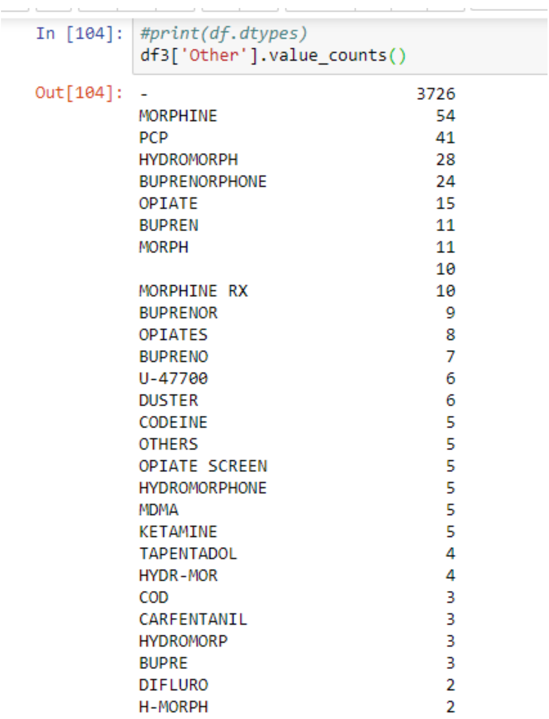
```
 path["Hydrocodone"].value_counts()

NO     3977
YES     104
Name: Hydrocodone, dtype: int64
```



Benzodiazepine	Methadone	Amphet	Tramad	Morphine (not heroin)	Other	Any Opioid
YES	NO	NO	NO	NO	-	NO
NO	NO	NO	NO	NO	-	NO
YES	NO	NO	NO	NO	-	NO
NO	NO	NO	NO	NO	-	NO
NO	NO	NO	NO	NO	-	NO

5) Ensuring no drug is repeated in the ‘Other’ column



```
In [104]: #print(df.dtypes)
          df3['Other'].value_counts()

Out[104]: -                               3726
          MORPHINE                          54
          PCP                               41
          HYDROMORPH                        28
          BUPRENORPHONE                     24
          OPIATE                            15
          BUPREN                            11
          MORPH                             11
                                            10
          MORPHINE RX                       10
          BUPRENOR                           9
          OPIATES                            8
          BUPRENO                            7
          U-47700                            6
          DUSTER                             6
          CODEINE                            5
          OTHERS                             5
          OPIATE SCREEN                      5
          HYDROMORPHONE                      5
          MDMA                               5
          KETAMINE                           5
          TAPENTADOL                         4
          HYDR-MOR                           4
          COD                                3
          CARFENTANIL                        3
          HYDROMORP                          3
          BUPRE                              3
          DIFLURO                            2
          H-MORPH                            2
```

Along with columns for various drugs that have caused the death, there is another column in the data frame that includes the drugs that haven’t been mentioned earlier. On further looking into the values of this column, it was observed that morphine was entered a number of times in the ‘Other’ column although there is a column specific to morphine. For such data entries, the row under morphine was updated to “YES” and the other column was made empty. Also, within the ‘Other’ column, Morphine was entered in different formats, such as MORPH, morphine, Morphine, MORPHINE. All these cases were considered and to bring in uniformity, the entire row was capitalized as it is case sensitive.

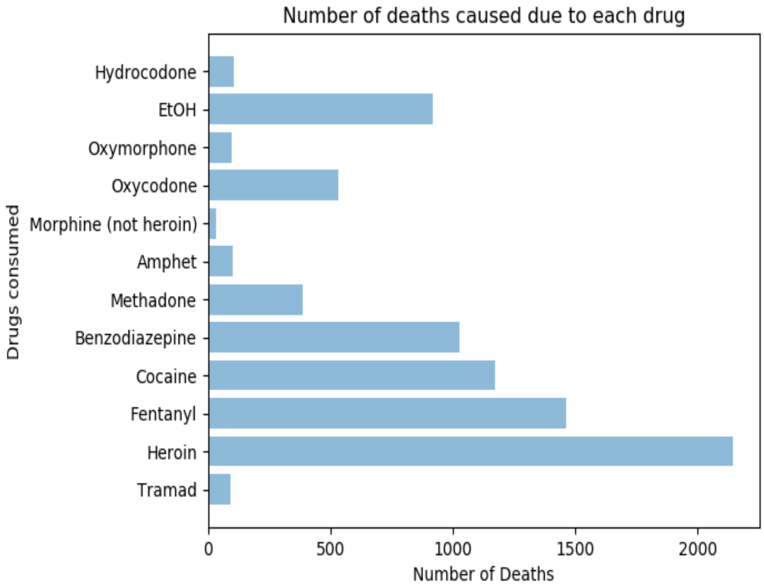
EXPLORATORY/DESCRIPTIVE ANALYSIS

1) In order to gain knowledge on the number of deaths caused to the consumption of a specific drug among the various drugs, we decided to plot bar graph which gives a comparative proportion of deaths caused by each drug that has been consumed.

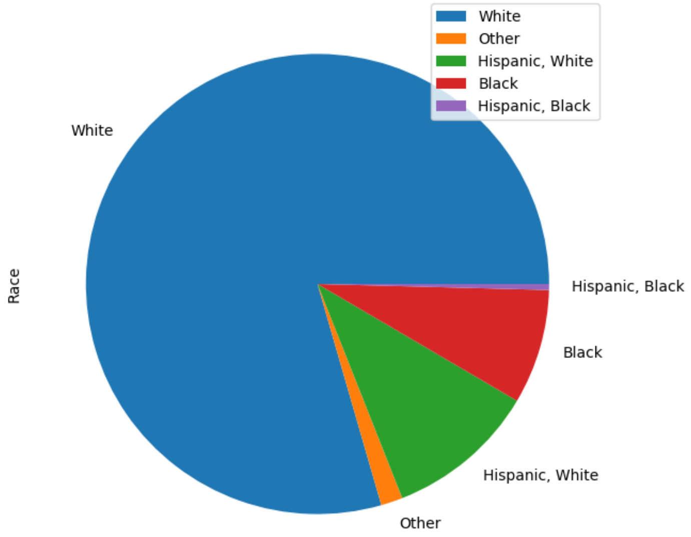
From the bar graph of Number of deaths VS Drugs Consumed

Brief Analysis/Insight:

We can infer that the drug that caused the greatest number of deaths in this region is HEROIN.



Thus, the insight that we have gained from this graph is number of deaths associated to each drug that has been consumed by the users in this region.

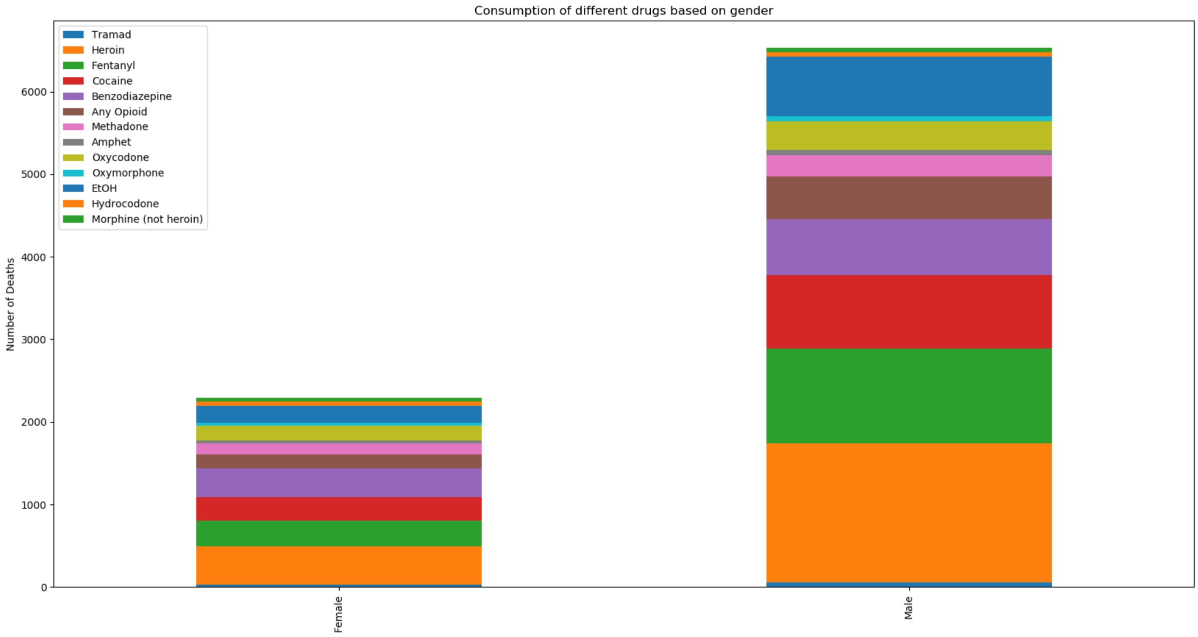


2)Considering the contribution of the race factor to the number of drugs induced deaths, we decided to use a pie chart to show the proportion of people of each race in that region who were affected.

We used 5 major categories as shown.

Brief Analysis/Insight:

It can be clearly seen that the most affected race (most number of casualties) is the White race. But it must be kept in mind that this is specific to that region and hence the general distribution or proportion of people for each group of race will vary.



3)Firstly, to try to answer the basic questions on number of deaths caused due to each drug and going one step further by sub dividing it based on gender, we analysed the data using stacked bar graph (x axis – gender, stacks – different drugs, y – number of deaths). To create this graph, using the data frame, we created sub data frames for each gender, i.e. Male and Female. Finally stored the data into a nested dictionary with keys being the drugs and within that the gender, storing values of number of deaths.

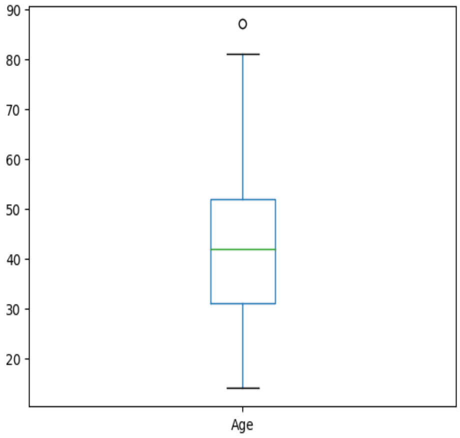
Brief Analysis/Insight:

1. Male overall shows higher number of deaths by a significant amount, more than double in total.

2. Of all the drugs, the greatest number of deaths caused due to Heroin for both male and female.

3. It can be seen that the next to heroin for Male is Fentanyl, whereas for female it is benzodiazepine

4)In order to find any extreme case or any anomaly in the "Age" column, we plotted a boxplot using the ages of the people that succumbed to drug induced deaths.



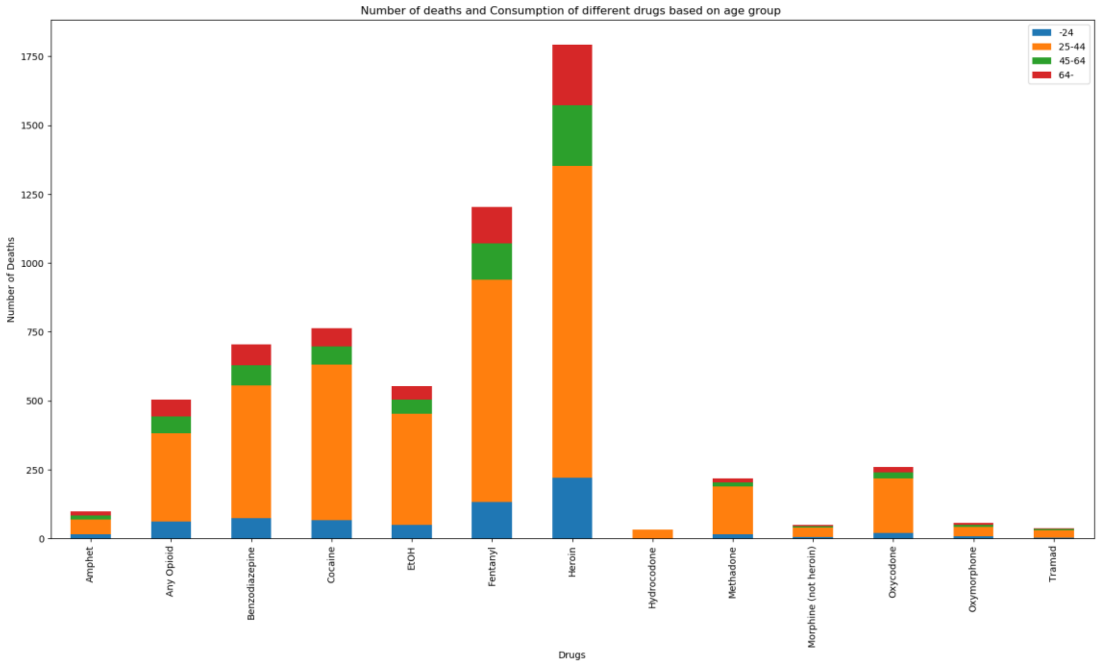
Brief Analysis/Insight:

1. There is one evident outlier (anomaly) i.e., at the age 87. However, it is not unlikely that this case cannot occur, due to which we decided that it would be best to only detect it, but not replace it.

2. The median age can be observed from the graph and is observed to be roughly 42 years.

5) Few other questions, trying to relate age with factors like drugs causing deaths and number of deaths can be answered by using a stacked bar plot.

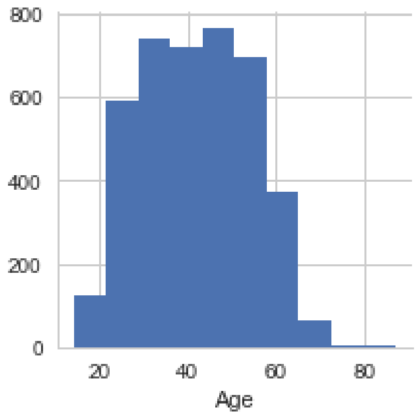
Instead of checking individually for each age, we grouped the ages into 4 categories (<=24, 25-44, 45-64,>64). Similar to the previous bar plot, we created a nested dictionary storing values of number of deaths in each age group, for every drug.

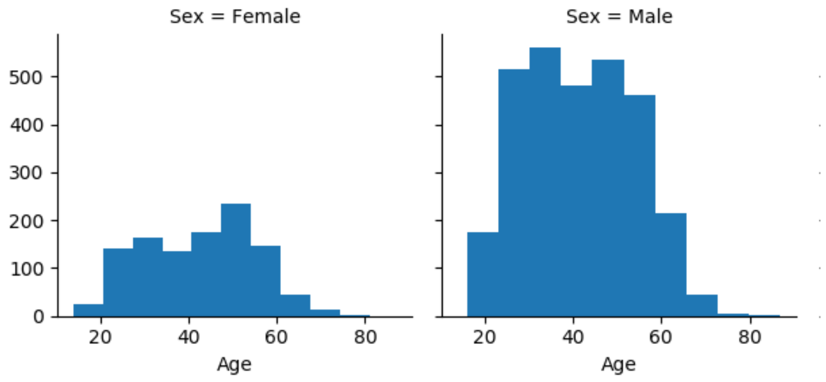


Basic Analysis/Insight: 1) it can be distinctly seen that for each and every drug, the most effected age group is from 25-44.

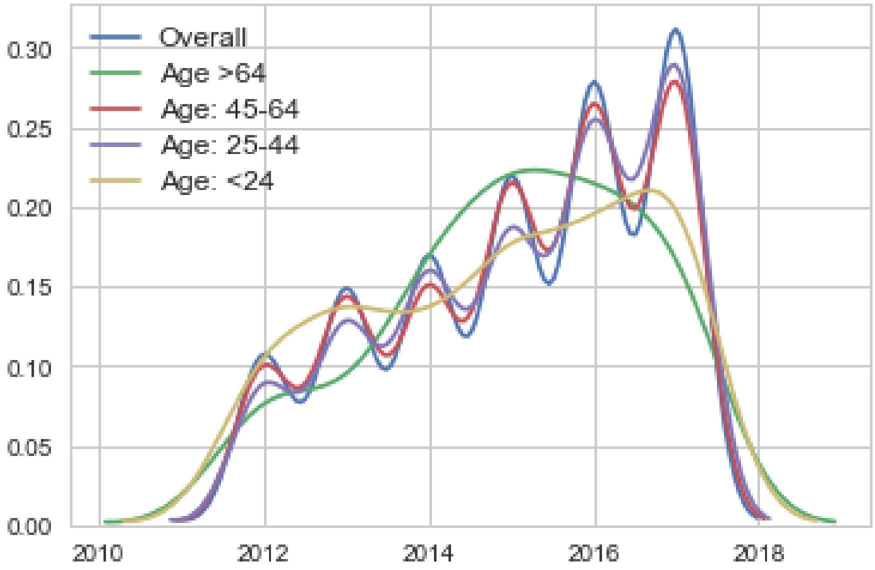
2) For each drug the proportion of age groups effected is pretty similar with 25-44 highest and >64 the lowest.

3) A clear picture can be drawn about which drug has cause most number of deaths, the order of drugs causing a number of deaths by the bar plot.





6) Another important factor to be analysed is the time-frame. It is important to know if there has been any improvement over time, or if in fact it is getting worse. Now to be more specific, we created line plots for the number of deaths caused over the years, and for better comparative study , did this for each age group and overall for the entire data-frame



Brief Analysis/Insight:

1. The overall behaviour is mainly affected by the age groups 25-44, 45-64

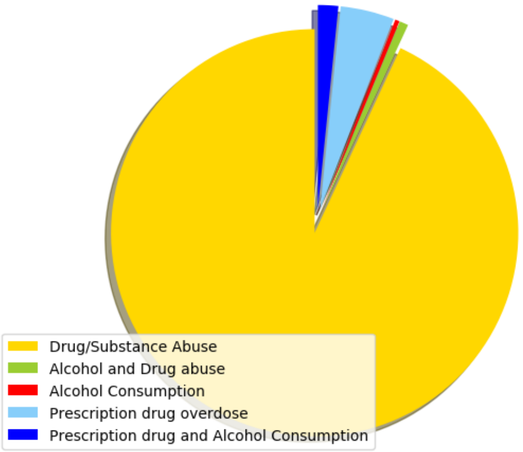
2. For the age group above 64, the highest number of deaths was somewhere in 2015, but on the positive side over the years it has decreased.

3. For age groups less than 24, there is a gradual and steady increase in the number of deaths, although compared to the other age groups, it is much lesser

4. Finally, for the age groups 25-44 and 45-64, there has been a constant rise in number of deaths, the rise being much steeper than that for the other two age groups, especially, the last one year there is a steep rise / increase in number of deaths

7) We plotted a graph to show the comparison between the different immediate causes of death.

Immediate cause of death



Brief Analysis/Insight:

1. Drug/Substance abuse is responsible for the greatest number of deaths.

2. However, it can be inferred that there appears to be a notable proportion of deaths due to prescription drugs. This proves that users at times misuse their medications.

Conclusion

For the dataset chosen by us, we have observed that the drug that is responsible for the most number of deaths is Heroin.

Compared to other age groups, the ratio of then number of deaths caused by drugs is significantly higher for age group 25-44 years.

Although, the greatest number of deaths were caused due to recreational drugs, we observed a notable number of deaths caused due to prescribed medications. From this, we can deduce that a lot of people misuse prescription drugs.

Unfortunately, over the years the number of drug induced deaths have increased considerably.

Thus, it is important to work with the target group in order to reduce the impact of drug usage.

Our Takeaway from this assignment:

This exercise gave us the opportunity to work with a particular data set and go in depths to read it and extract necessary information.

We learnt to interpret every aspect of the given data and provide an appropriate visual representation of the various factors using graphs that compared different variables from which we were able to deduced important patterns and facts.

Thus, working on this assignment has provided us a basic understanding of the overall process of restructuring(cleaning), analysing and coming to an apt conclusion from the insights drawn from the data set.