Course: *Data Wrangling (DATA 422)*

Course Tutors: *Giulio Valentino Dalla Riva, Thomas Li*

Project Title: *Death rate analysis in the USA due to drug addiction*

Team Members (Team Orion):

- *Neha Thakkar (51192498)*
- *Mrinal Jyoti Kumar (96475046)*
- *Siddharth Rana (45562544)*
- *Sudhanshu Kaushik (75074054)*

Executive Summary:

In this project, we are analyzing the drug related death due to different types of drug addiction across USA. The main intention of selecting this dataset is because we have analyzed in another dataset file which consists of the data for 230 countries in the world and we found that USA stands top among all in terms of death rate due to drug addiction. The dataset which we have chosen to analyze the USA dataset contains variables like drug type, gender, date of death, residence city, age etc. For doing the project analysis we have divided the project into two sections which are as below:

1) Processing Section – In this section the dataset was cleaned and merged with other cleaned dataset files.
2) Analysis Section – In this section the dataframe was visualized on the basis of many relationships with different variables and the visualized graphs were analyzed.

The primary intention of choosing this project to create awareness and analyze the vulnerability of drugs with respect to gender, age and drug type in USA. Also we have used all the necessary packages for both R and Julia in order to wrangle the data. The links of all the dataset files are shared below for reference.

Dataset links: https://ourworldindata.org/drug-use

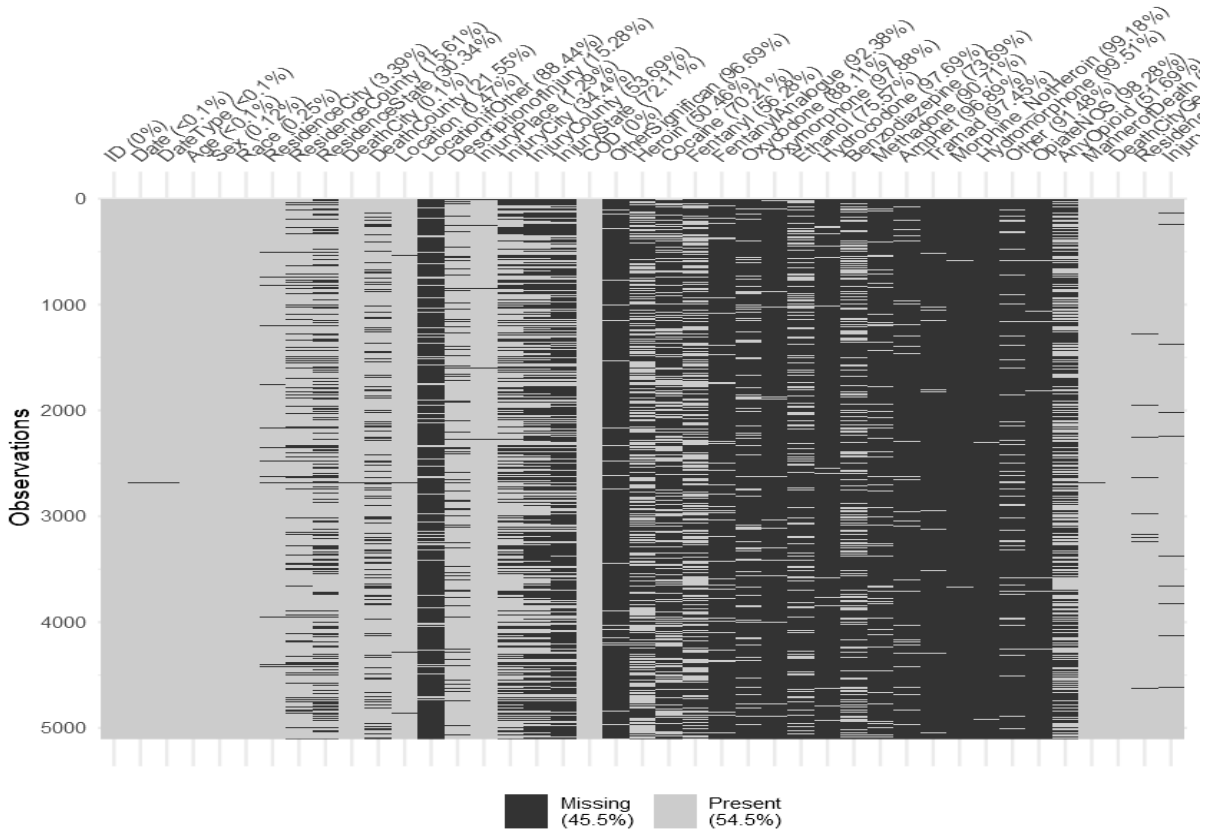
https://catalog.data.gov/dataset

Project Details:

In the project we have performed data wrangling stuff by cleaning the data, merging the data and scrapping the data. We have used R and Julia as the languages to perform data wrangling. The main aim of our project was to analyze the drug related death due to drug addiction across USA. By analyzing this we can make sure how to reduce the number of deaths in the future by implementing preventive actions. Nowadays drugs are being consumed by many people across the world and among all countries it's found that USA has the highest count of death due to drug addiction. The below steps will highlight all the findings and analysis done as part of our project.

Step 1: We have collected the dataset and analysed the messy and uncleaned dataset which we have found. We observed that there were many missing values in our data set which was the most challenging task for our project. The total number of missing values in our dataset file was close to 45% which was too high. The main challenge was which columns to remove from the dataframe and which one to retain. Hence, we did many team meetings just to understand what visualization graphs we have to show and based on that we made a decision of removing the unwanted columns from the loaded dataframe object.

The "vis_miss" function results of our initial dataframe object looks like the below screenshot. The black lines mean the data is missing and the while line means the data is available in the dataframe object.



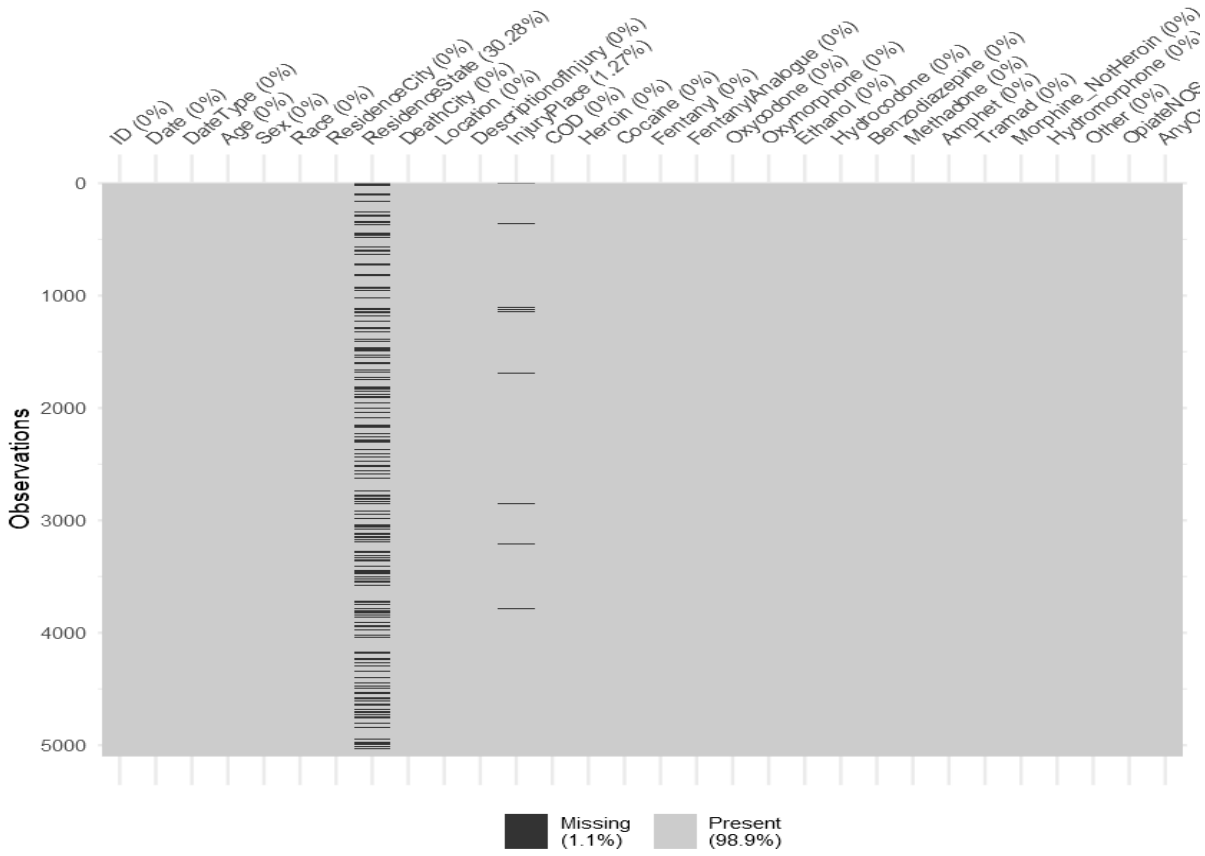
Step 2: In the next step, after we identified which columns have to be dropped from our dataframe object we have performed the necessary actions on our dataset to make it look meaningful than before. While doing so we used all the necessary packages of R and cleaned the dataframe. Mostly we used the tidyverse and dplyr packages to handle the dataframe.

Step 3: Even after removing the unwanted columns didn't solve all our problems as we saw in the vis_miss results that there were few other columns which had missing records and those columns were mostly the unique identifier column, age, sex, residence state columns. Hence we decided to remove the records from the unique identifier column, age and sex column as the number of missing records was minimal. Also we modified the records of the residence state column as "Unknown" as the number of missing records were minimal as well. We used the respective functions from the tidyverse and dplyr packages in R.

Step 4: We observed the vis_miss results after doing the above steps and the dataframe looks much cleaned. This was important with perspective of data analysis and visualization.

Step 5: In the next step wanted to show the zip code for the respective residence city column in the dataframe object just to show some relationship with the zip code. Hence we have obtained the zip code from the link (http://federalgovernmentzipcodes.us/download.html) and we have created a master dataset with the same information. Later we have joined the zipcode dataframe with our cleaned dataframe object using inner join function. Now our overall dataframe object has the zipcode column as well along with the other existing columns for our data analysis.

The below screenshot shows the final dataframe which we received and we can see that the missing values drastically reduced from 45% to 1.1 %.



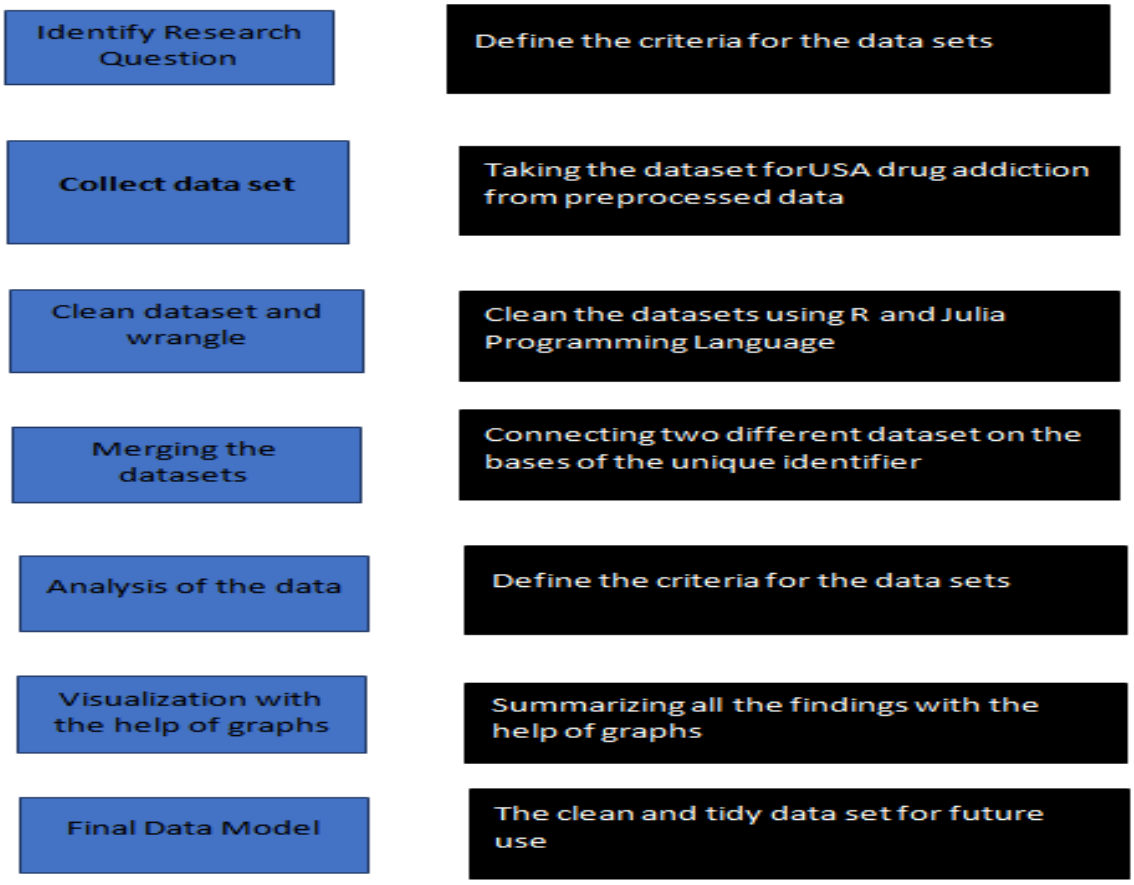
Step 6: We have divided the number of data visualization graphs for both Julia and R. Hence we have shown 4 graphs in Julia and 5 graphs in R. All these graphs show a significant relationship of variables and we have gained much information due to this.

Step 7: We also wanted to perform some bit of web scrapping from web links. Hence we have fetched the most common reasons of drug addictions and common treatment of drug addiction from the web text and converted it as a dataframe object. This information gives an extra bit of knowledge from the information which we have gathered after analysing the graphs plotted. The web scrapped web-links are: https://sunrisehouse.com/cause-effect/reasons-addiction/ and https://sunrisehouse.com/assessment-diagnose/drug-addiction/

Step 8: Finally we have used the "write_csv" function to write the final model into a csv file. The final model is a csv file (cleaned_data.csv) with number of rows as 5099 and number of columns as 36.

Project Design:

In this section the complete project flow has been shown starting from identifying the research question which was to find the dataset until obtaining the final model. The blue boxes are the process steps and the black boxes are the tasks done with respect to that step.



Project Scope:

In this section we have defined the usage of this project finding in future after analysing our graphs. The below are the few project scopes.

- The final model would be useful for future use in terms of accuracy and reliability
- The cleaned dataset would be useful for determining the cause of death
- It also helps the observer to save the costs of the specific survey in regards to the drug related death as it is the clean data
- The project results could be used for further analysis and for future predictions
- The governed bodies could implement some necessary actions in order to reduce the drug death cause and the easy accessibility of drugs

Project Strategic:

In this section we have highlighted the strategies which we have implemented throughout the project work during individual or team work.

- Strategy was made on - How to clean the untidy dataset from 45% to less number
- Strategy was made on - How to show relationship among different fields
- Strategy was made on - Who will perform the peer reviews for whichever tasks
- Strategy was made on - Whom to assign what task to work in parallel

- Strategy was made on - When challenges were found, how to tackle the same with inputs from each one of the team members by brainstorming the issue
- Strategy was made on - How to improvise on the designed code by following the process, data ethics and code ethics

It was only due to effective strategy we came up with lots of ideas and designed the final model. Also we could find many challenges during the project work and also the resolution for the same which we have mentioned below in the succeeding section.

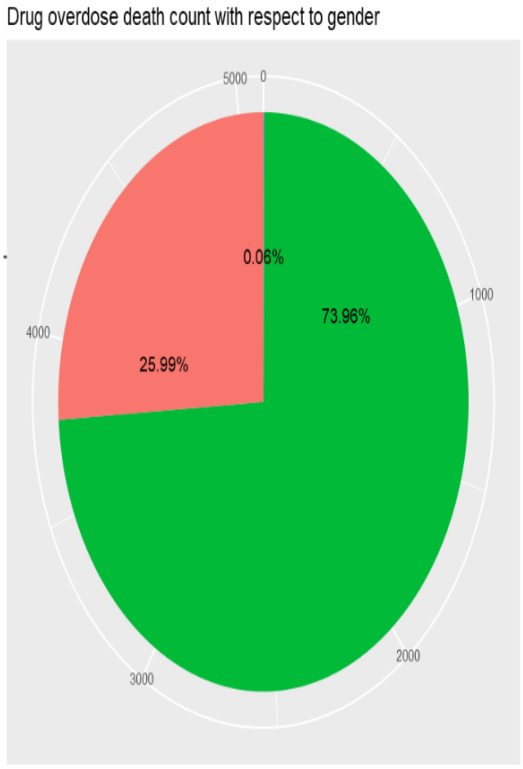
Data Visualization:

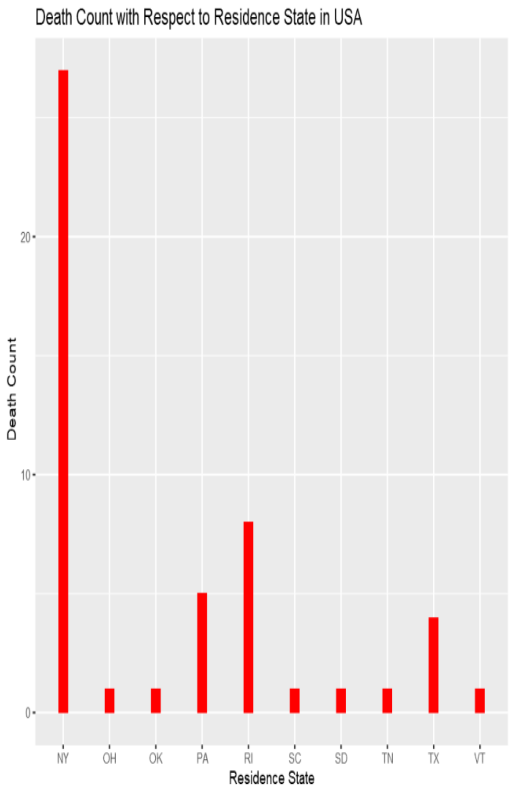
In this section we have shown the various graphs which we have plotted in both R and Julia.

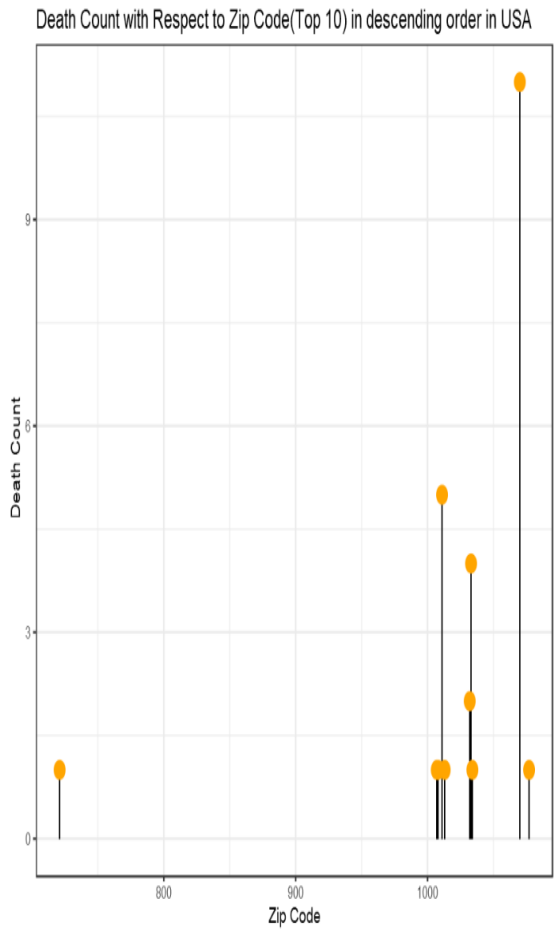
The various graphs are listed below:

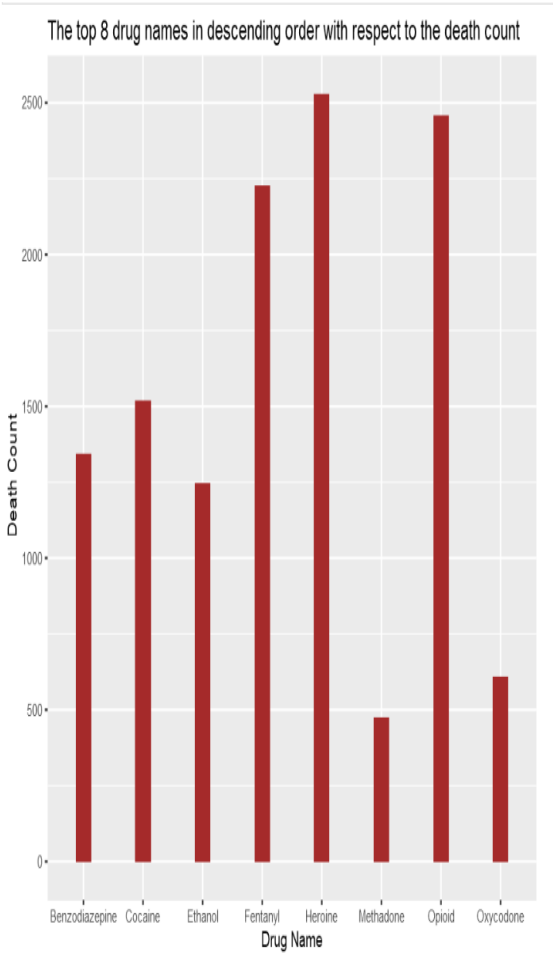
- Pie Chart – Death Count Percentage With Respect To Gender In USA
- Bar Graph/Line Graph – Death Count With Respect To Different Age Groups In USA
- Bar Graph - Death Count With Respect To Various Drugs Being Used In USA
- Bar Graph – Death Count With Respect To Years
- Bar Graph – Death Count In Years With Respect to Heroine drug
- Bar Graph – Death Count With Respect To Residence State In USA
- Bar Graph – Death Count With Respect To ZIPCode In USA For Highest Deaths

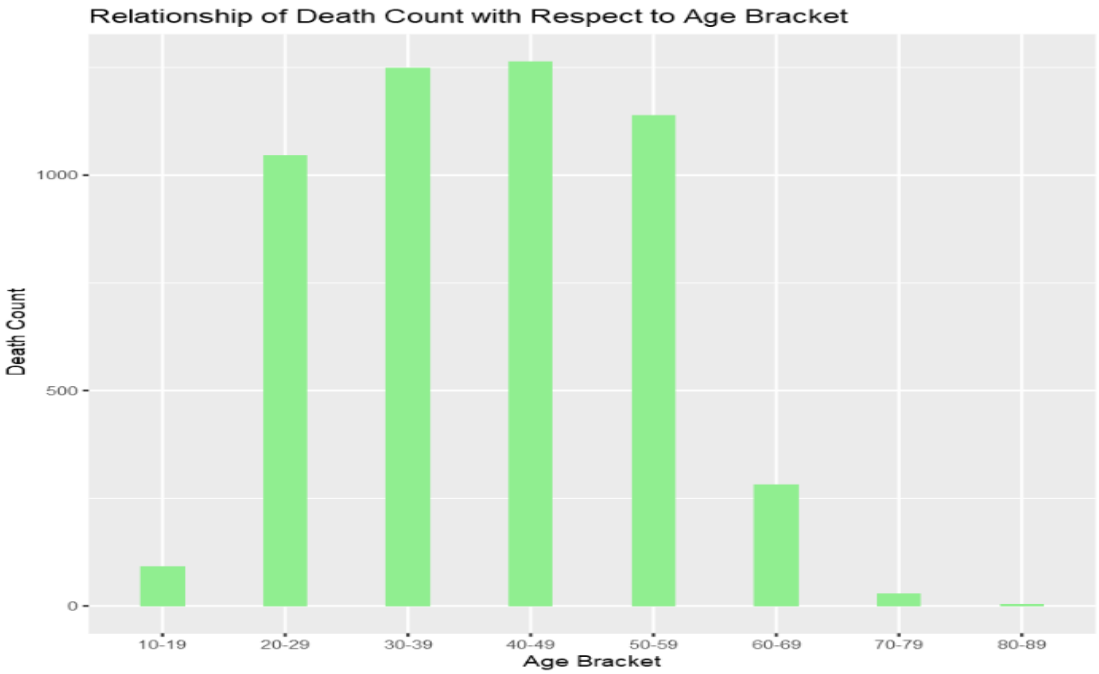
Graphs in R -



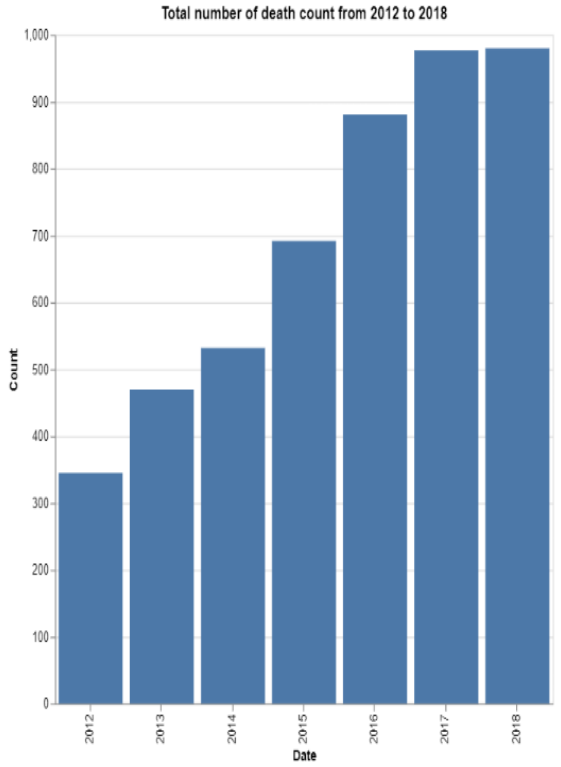


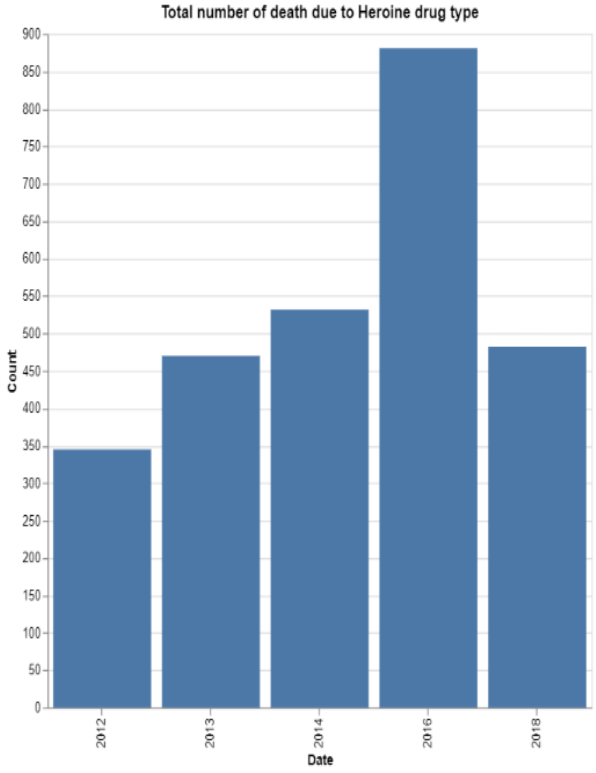

Death Count with Respect to Zip Code(Top 10) in descending order in USA
Death Count
9
6
3
0
800
900
1000
Zip Code

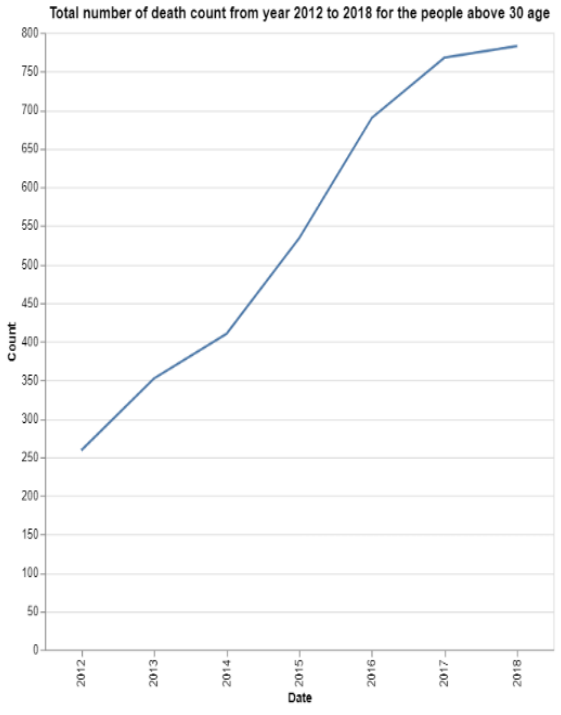

The top 8 drug names in descending order with respect to the death count
Death Count
2500
2000
1500
1000
500
0
Benzodiazepine
Cocaine
Ethanol
Fentanyl
Heroine
Methadone
Opioid
Oxycodone
Drug Name

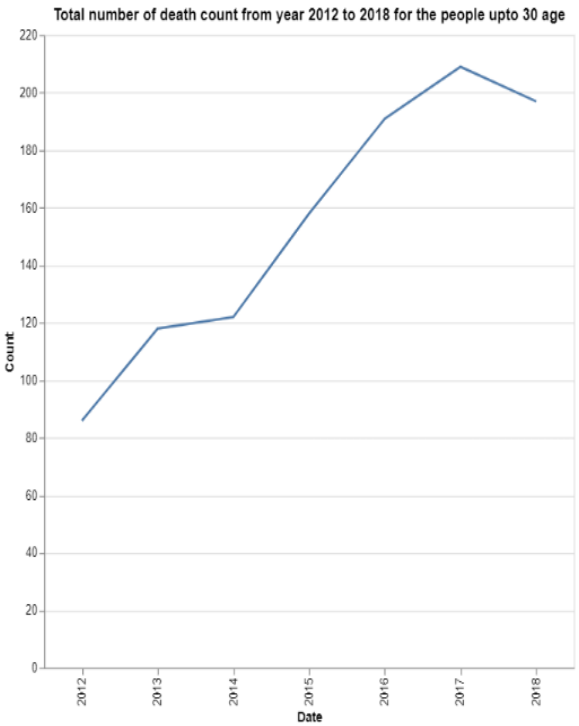

Relationship of Death Count with Respect to Age Bracket
Death Count
1000
500
0
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89
Age Bracket

Graphs in Julia -









Data Ethics:

As part of the data ethics we have made sure that the datasets which we have downloaded are not from the copy writing sites. All these datasets are from the government sites which we have provided in our referential links. We have also made sure that we don't temper with the personal data like the unique identifier records in our dataset. Maintaining data ethically is very important as part of data wrangling project.

Challenges Encountered:

In any project there are certain challenges which have to be taken care. In our project also we have encountered certain challenges which took us some time to handle. This has been tackled by group discussion and effective strategies were implemented. The challenges which we encountered are listed below:

- While cleaning the original dataset with 45% untidy records
- While handling the Date Column in the dataset, as the data type for the same was in "custom" type and not in "date" type
- While showing the different ages into age brackets to show in the bar graph as part of our analysis

Challenges Resolution:

The above challenges had to be resolved so that we can analyze our dataset and come up with some conclusion. Hence we strategized diligently during our group discussions and we came up with certain solutions in order to move forward. The following are the solutions to resolve the challenges which we encountered:

- As part of the first challenge we have dropped the unwanted columns from the dataset which were not part of our analysis. Also for certain columns wherever we found blank records we have removed or replaced the records
- The data type of the date column was in custom type; hence we were not able to use it while converting into year format. For resolving this issue we have changed the data type of the date column from custom to date type in excel and then used the dataset by converting into dataframe object.
- As part of the analysis we had to convert the ages into respective age bucket. For this we had to make some age groups and then used a loop to fetch each age number and put that into respective age bucket.

Conclusion:

We can conclude that various drugs are being used in huge amount in many countries. USA is among the top of all the countries in the world with maximum consumption. After analyzing the datasets for our project we have come up with the following conclusions which are as follows:

- Male population have more death count compared to the females due to drug addiction in USA
- People with age group greater than 30 years are more into drug addiction (40-49 age group is the most vulnerable)
- 'Heroine' is the most commonly used drug in USA
- New York State has more cases of drug addiction compared to other states in USA
- The usage of drugs has been increasing every year i.e. people are getting more addicted irrespective of the consequences

The main intention of carrying out this project was to show the analyzed statistics like mentioned above so that the respective govern bodies can take the necessary actions to prevent the same. Several hypothesis were predicted like the age group of less than 30 years would be the most vulnerable group with regards to the drug death but it was proven incorrect when we analyzed the graphs on the basis of drug death with respect to the age group.

Best Practices:

As part of the best practices performed in our project, we have maintained a centralized public repository (Github) where we have committed all our project related files. The github is used as the source to store, update and maintain our project related file systems. Also as part of group conversations and interactions we have used Keybase account to do the same on a consistent basis.

Every team member of our project was part of the peer review process for any new code design, existing code updation and any other file review. We made sure that each file is reviewed by atleast two reviewers at a time to maintain the data quality.

Github link: https://github.com/data-wrangler19/myproject.git

Packages Used:

R	JULIA
tidyverse	Queryverse
dplyr	Vegalite, VegaDatasets
ggplot2	dataFrames, query
visdat, skimr, rvest	statistics, dates

The above packages were used throughout the project while working with R and Julia module of the project.

File To Be Referred:

We have attached the following files in our project folder in Github repository for your reference.

Coding Files -

- Data_Wrangling_Project_R : This file has all the code related to our main dataset on USA using R
- Data_Wrangling_Project_Julia : This file has all the code related to our main dataset on USA using Julia
- USA_Analysis_DeathRate_R: This file has all the code related to our all countries dataset using R
- Web-Scrapping_R: This file has all the code related to the web scrapping done in the html files using R

Uncleaned Data Files –

- Accidental_Drug_Related_Deaths_2012-2018: This is our main uncleaned dataset file for USA
- death-rates-from-drug-use-disorders: This is our all countries dataset file
- Zipcode: This is our prepared zipcode dataset file

Cleaned Data Files –

- cleaned_data: This is our cleaned data model file which is in .csv format

Text Files –

- Professional Diary : This diary consists of the elaborated work done on the basis of days during our project work as a team
- Data Wrangling Project Report: This is a ".pdf" file which has the project report file
- Data_Wranging_Presentation: This ".ppt" file consists of our project work presentation

***Note**: Please refer to the respective ".ipythn" file to find the code comments for the respective lines of code designed.*