Big Data Project

**Team Members**

Khaled Said Badawy Shebl, Section: 2

Hussein Ata Hussein, Section: 2

Ahmed Sahrawy Shehata, Section: 1

Abanon Amin, Section: 1

David Mounir, Section: 2

**Under Supervision**

Dr. Sherine Rady

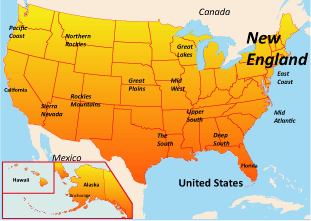
TAs: Mariam Mohamed

Hazem Marwan

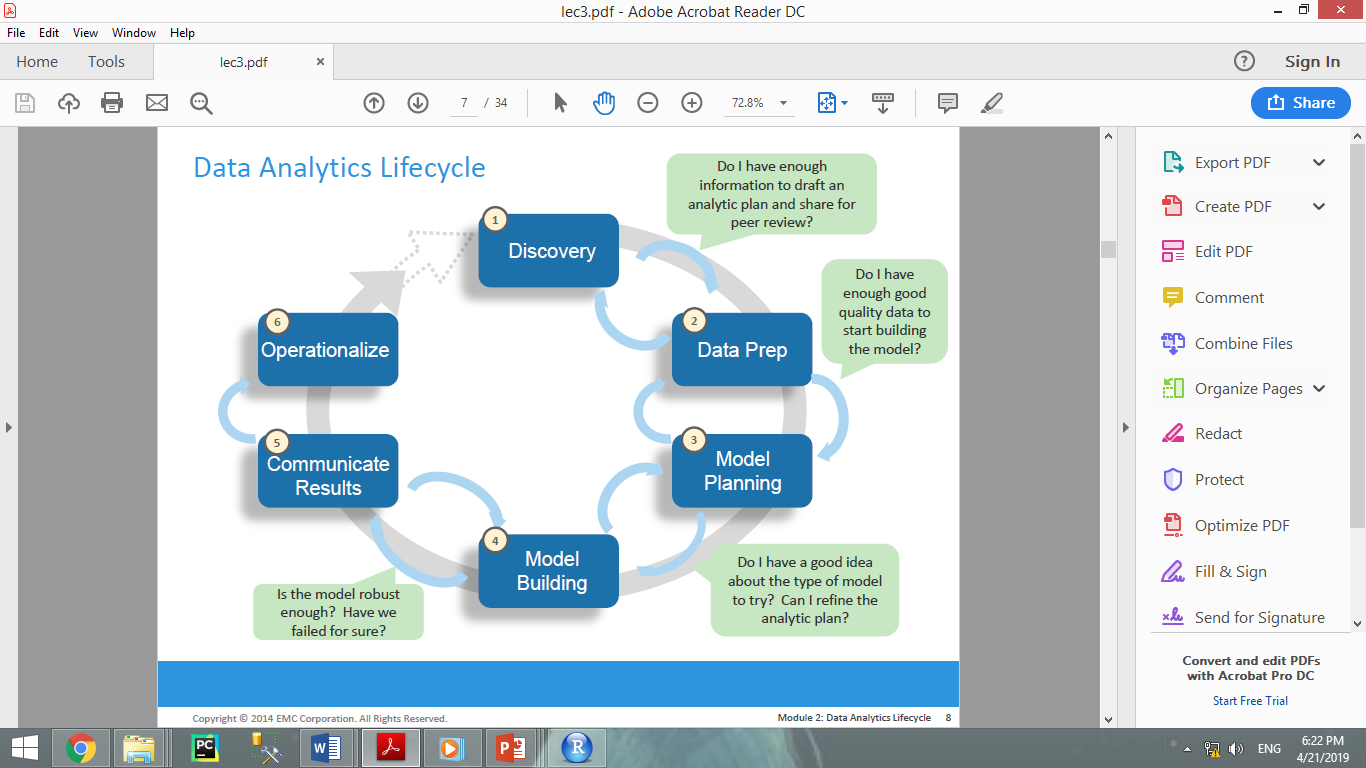
Main Idea: Predict manner of death (Accident, Pending or Natural) for one that

takes drugs**.**

Dataset name: Accidental Drug Related Deaths 2012-2018 in New England







**Phase 1: Discovery**

1. **Business Domain:** Predict manner of death (Accident, Pending or Natural) for one that

takes drugs**.**

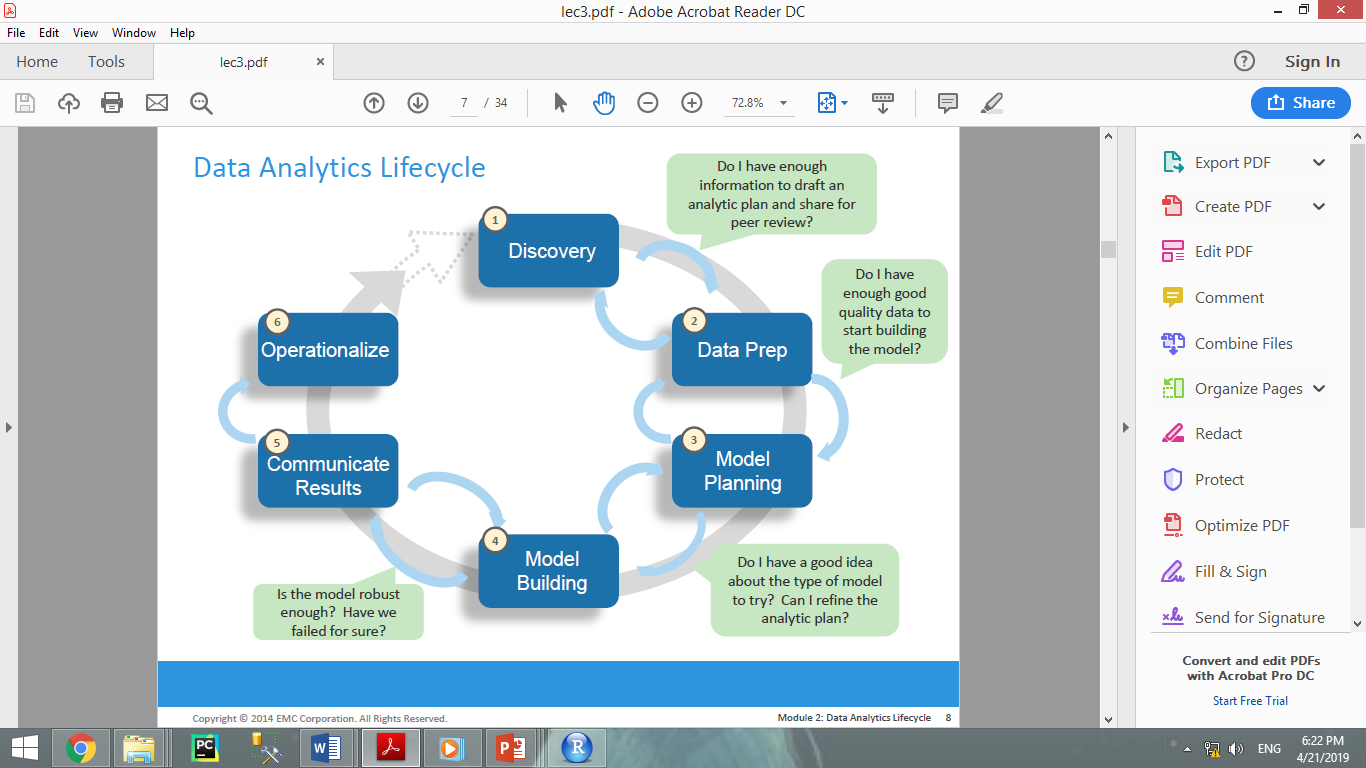
1. **Learn from the past:** There is no previous attempts to solve this problem.
2. **Resources:**
3. **Technologies:** R language.
4. **Data:** Accidental Drug Related Deaths 2012-2018 data set.
5. **People:** Team members.
6. **Time:** 2 weeks.
7. **Have sufficient resources to attempt the project:** Yes we have.

1. **Main problem:** People takes different drugs and the actual reason of death is not correctly identified.
2. **Stakeholders:** Doctors and society.
3. **Objective:** Helpdoctors to know accidental drug related deaths.
4. **Initial Hypotheses:** H0: Manner of death is accident.

H1: Manner of death is pending.

H2: Manner of death is natural.

1. **Tools:** RStudio.



**Phase 2: Data Preparation**

**Dataset description:** A listing of each accidental death associated with drug overdose in Connecticut (Connecticut is the southernmost state in the New England region of the United States) from 2012 to 2018**.**

Data are derived from an investigation by the Office of the Chief Medical Examiner which includes the toxicity report, death certificate.

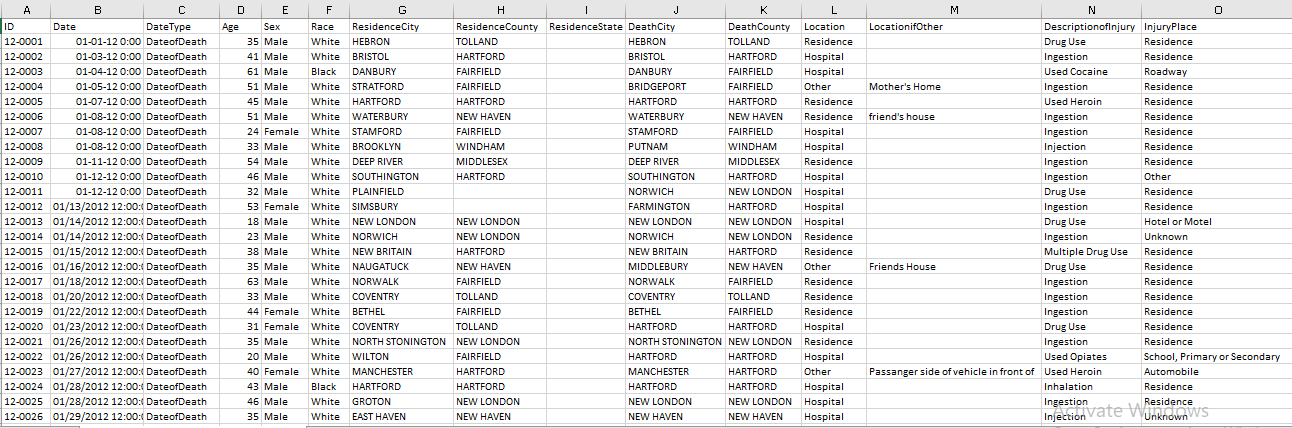
**Date source:**

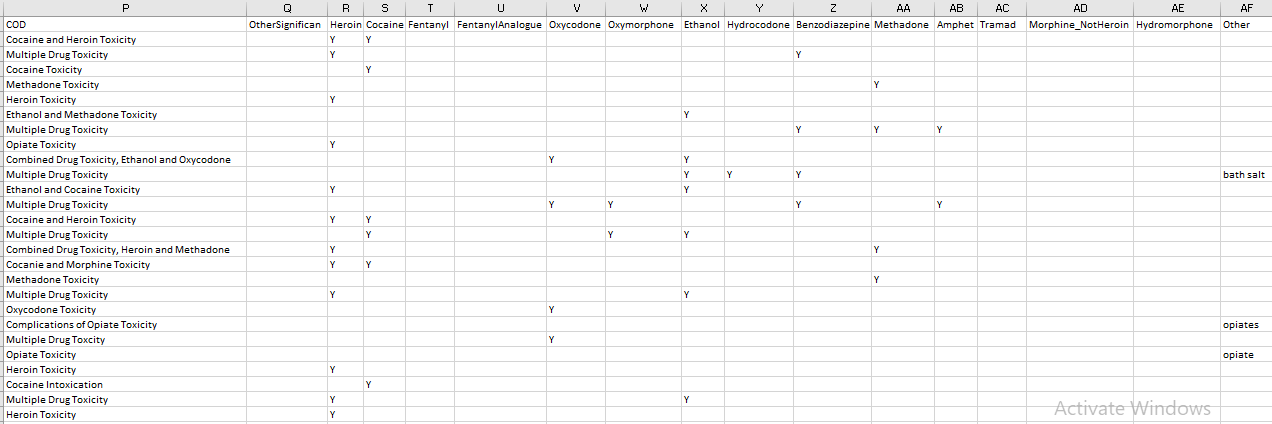
https://catalog.data.gov/dataset/accidental-drug-related-deaths-january-2012-sept-2015

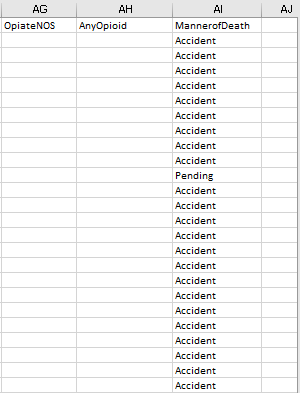
**Columns description:**

|  |  |
| --- | --- |
| **Column** | **Description** |
| ID | State ID |
| Date | Date of death or date of reporting state |
| DateType | Type of date (Reporting date or Death date) |
| Age | Age of state |
| Sex | Gender of state |
| Race | Race of state |
| ResidenceCity | City where the state is living in |
| ResidenceCounty | Country where the state is living in |
| ResidenceState | State where the state is living in |
| DeathCity | City where the state died |
| DeathCounty | Country where the state died |
| Location | Where the state is died(Residence place or Hospital or other) |
| LocationifOther | If location of death is other location not hospital or residence place |
| DescriptionofInjury | What type of overdose which causes the state injury(Ingestion, Used Heroin, Multiple Drug Use, Used Opiates). |
| InjuryPlace | where is the state has an injury(Residence  place, Hotel or Motel, Bar or Night Club , Unknown, other) |
| InjuryCity | City where the state has an injury |
| InjuryCounty | Country where the state has an injury |
| InjuryState | State where the state has an injury |
| COD | State where the state has an injury.  COD :- is an indicative measure of the amount of oxygen that can be consumed by reactions in a measured solution |
| OtherSignifican | Observations about the state such as another Illness(such as Coronary Artery Disease). |
| Heroin | Overdose of heroin |
| Cocaine | Overdose of cocaine |
| Fentanyl | Overdose of Fentanyl |
| FentanylAnalogue | Overdose of FentanylAnalogue |
| Oxycodone | Overdose of Oxycodone |
| Oxymorphone | Overdose of Oxymorphone |
| Ethanol | Overdose of Ethanol |
| Hydrocodone | Overdose of Hydrocodone |
| Benzodiazepine | Overdose of Benzodiazepine |
| Methadone | Overdose of Methadone |
| Amphet | Overdose of Amphet |
| Tramad | Overdose of Tramad |
| Morphine\_NotHeroin | Overdose of Morphine Not Heroin |
| Hydromorphone | Overdose of Hydromorphone |
| OpiateNOS | Opiates cover a huge variety of drugs(Y for yes or Null for No) |
| AnyOpioid | Overdose of AnyOpioid |
| Other | Any other type of dugs |
| MannerofDeath | Death type(During drug abuse “pending” or Accident or natural). |
| DeathCityGeo | Death city latitude and longitude |
| InjuryCityGeo | Injury city latitude and longitude |
| ResidenceCityGeo | Residence city latitude and longitude |

**Read data from .csv file to perform preprocessing on data**







**1)Read data set**



**2)Check if each column contains null values**

For example: COD column



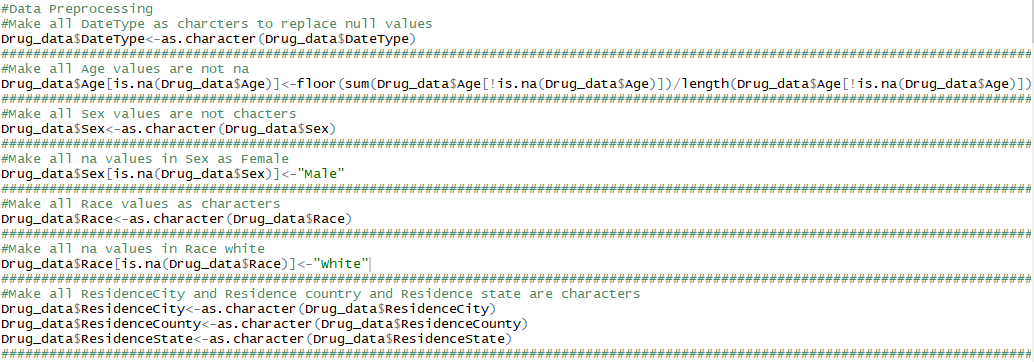
Output : **0** then there is no null values.

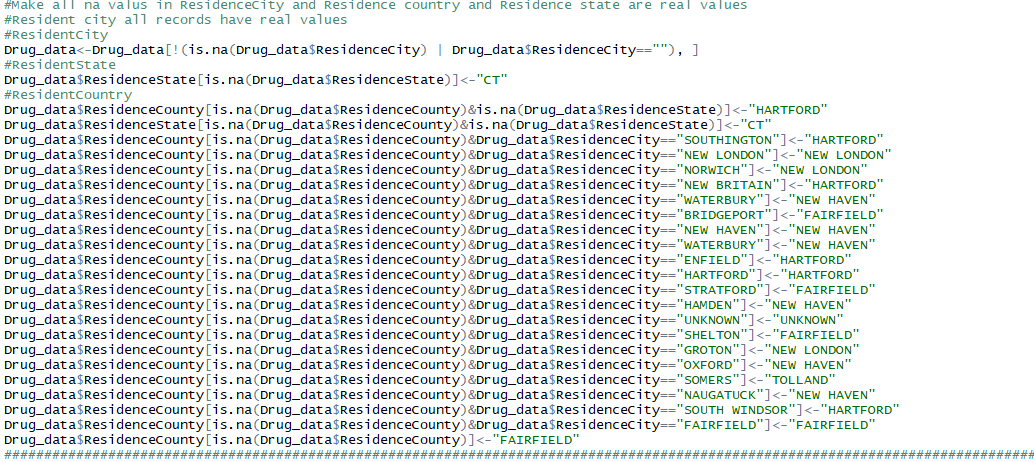
ResidenceCity column



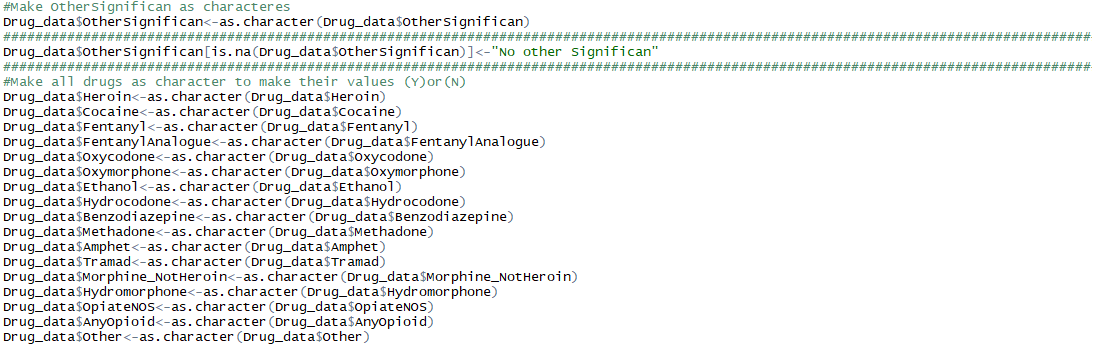
Output : **172** then there is 172 null values.

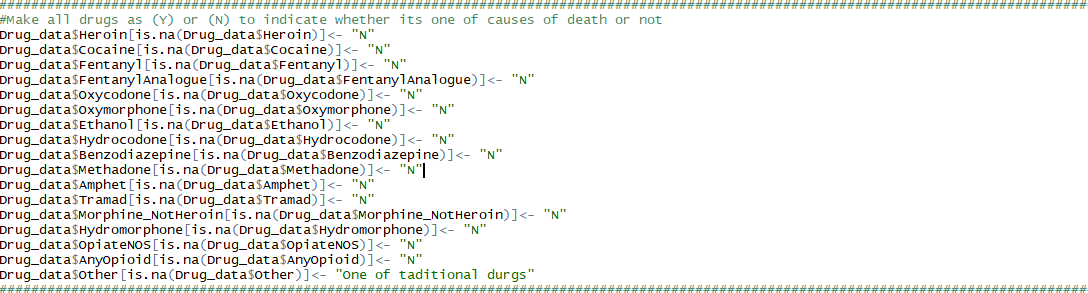
And So on for all columns and we observe columns which have null values to perform preprocessing on these columns.

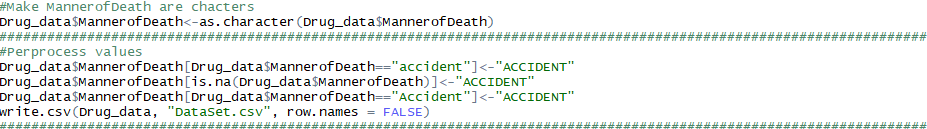












**Note :**

For null values for age we calculate floor of average for all rows and put that value instead the null value.

**3)Delete columns that have values will not affect our goal:**

DeathCityGeo, InjuryCityGeo, ResidenceCityGeo, InjuryCity, InjuryCounty and InjuryState.

**4)Delete some rows**

Delete rows that have (ResidenceCity or DeathCity) are null because our goal is to study accidents because of drugs based on location which is not available at those rows.

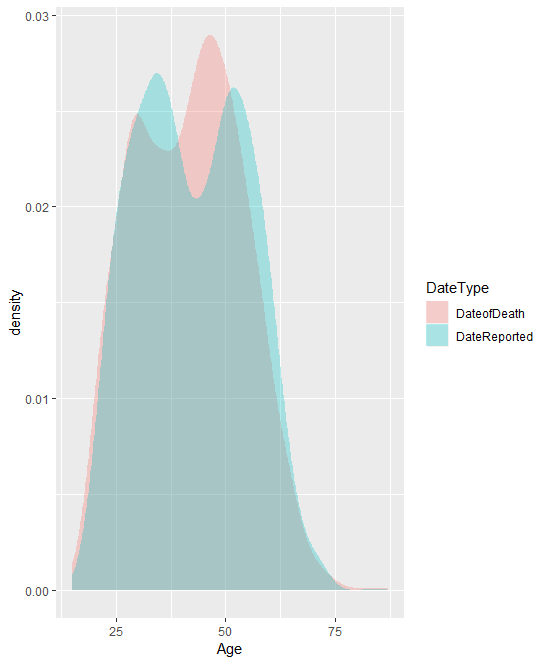




The next step is data visualization:

1)Detect relation between age and Date type



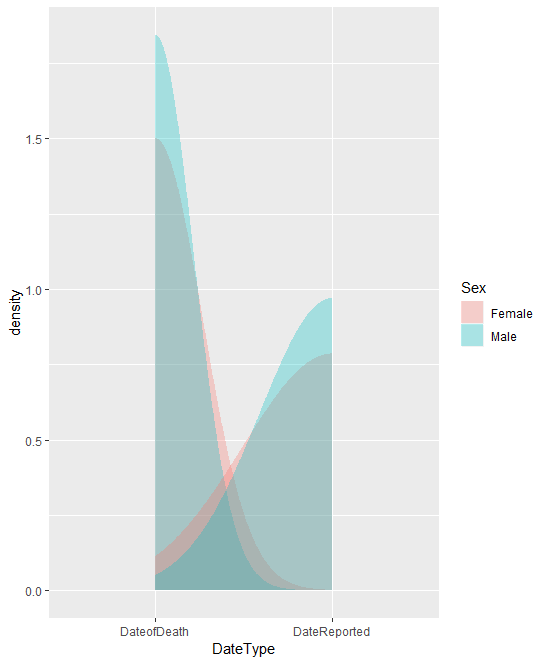


Observations :-

1. People who are between (45-50) are die because of drugs.

2)Detect relation between sex and date type



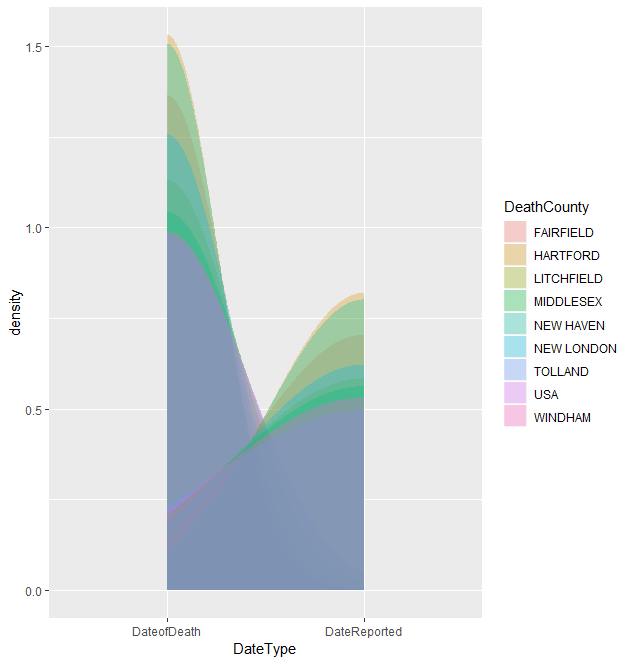


Observations :-

1)Number of males who die because of drugs are greater than number of females.

3) Detect relation between date type and country



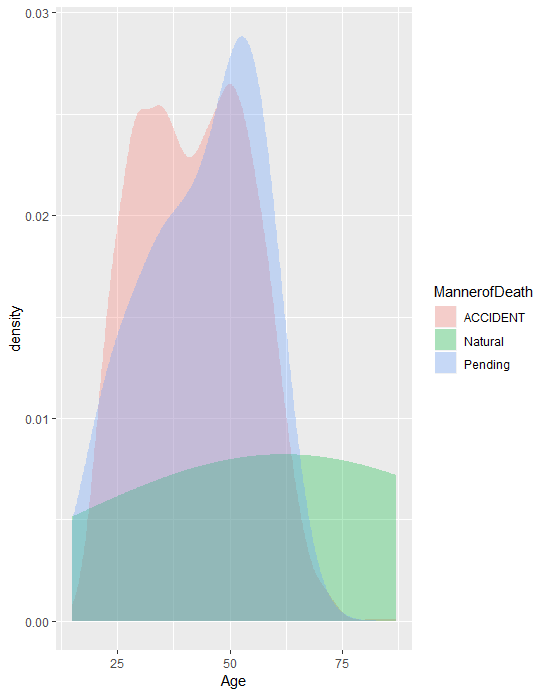


Observations :-

1)HARTFORD Country have the highest number of people who die because of drugs.

4)Detect relation between age and manner of death



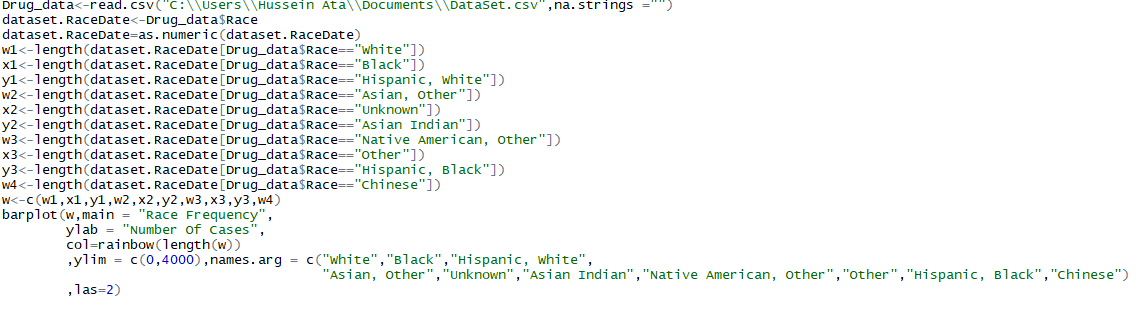


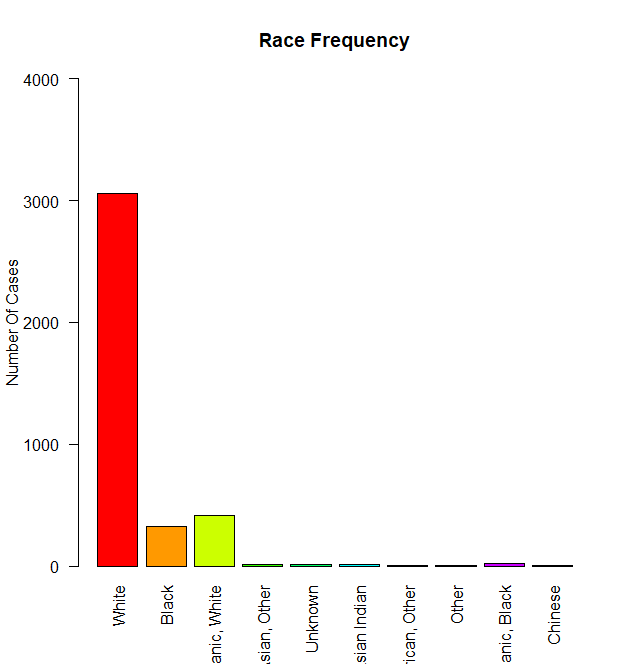
Observations :-

1)People who are between (50-55) years old die during taking drugs.

2) People who are between ((30-35) , (47-52)) years old die because of accidents related to drugs.

5) Detect race of the highest number of people that taking drugs

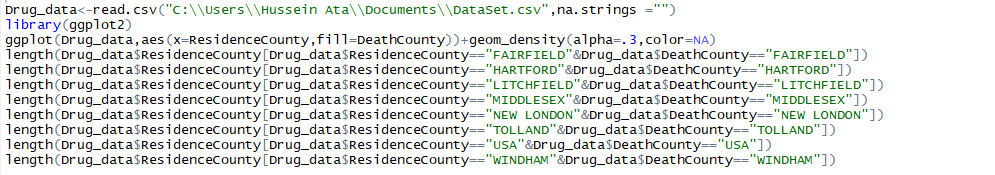


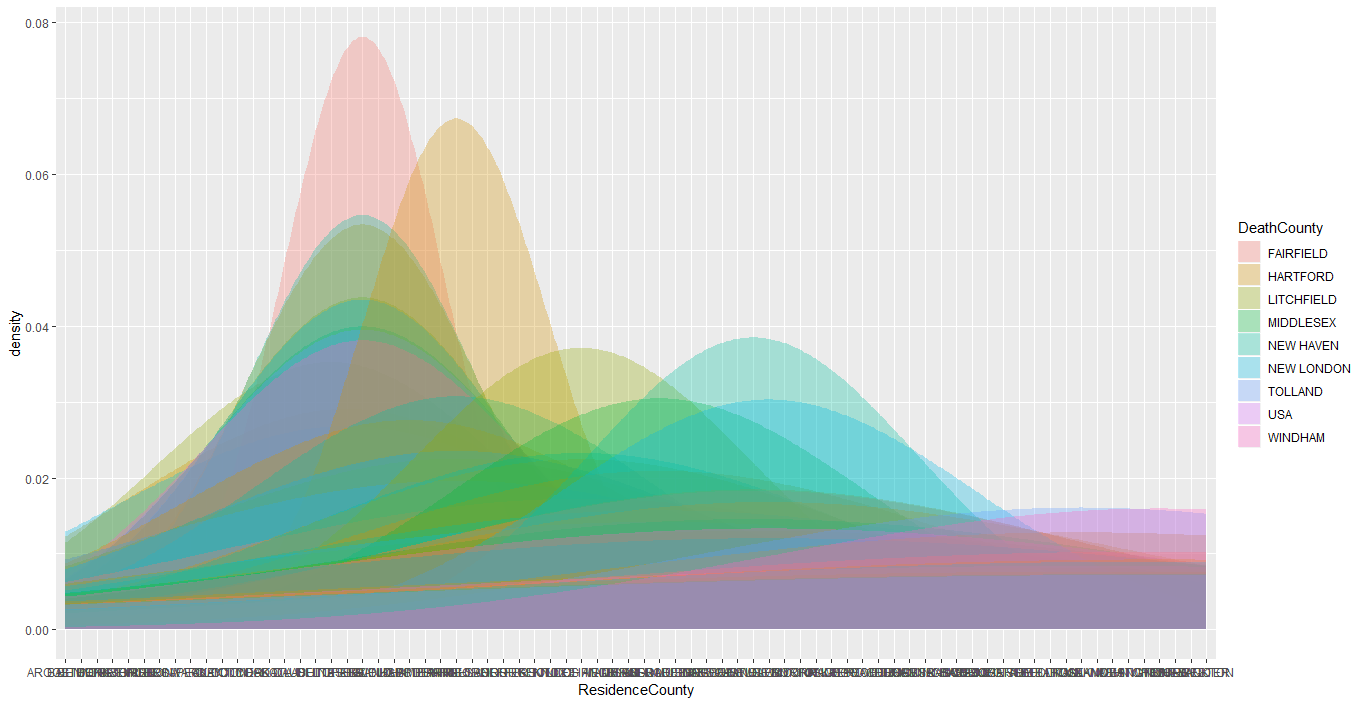


Observations :-

1)White people are the most people that takes drugs.

6)Detect relation between resident country and death country

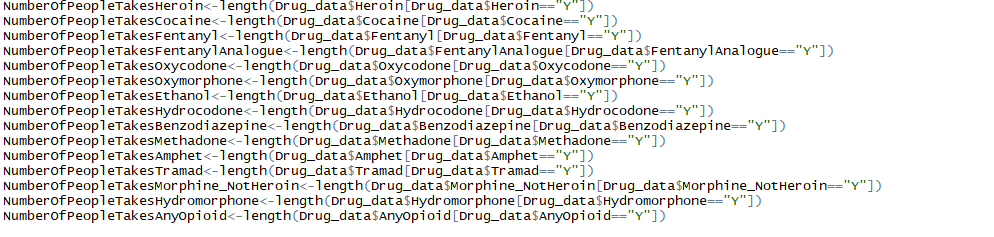


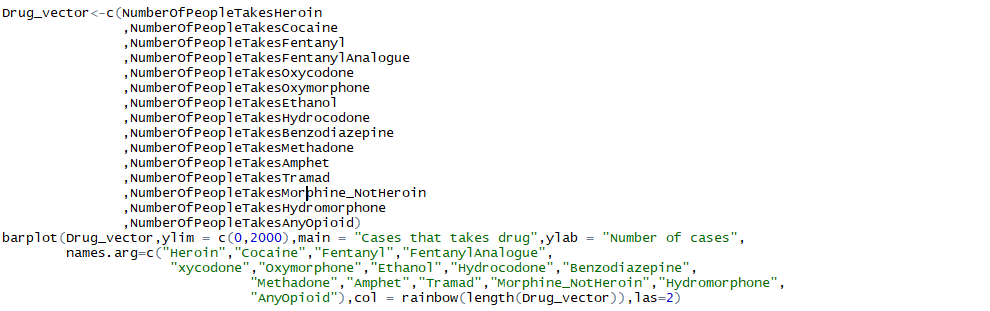


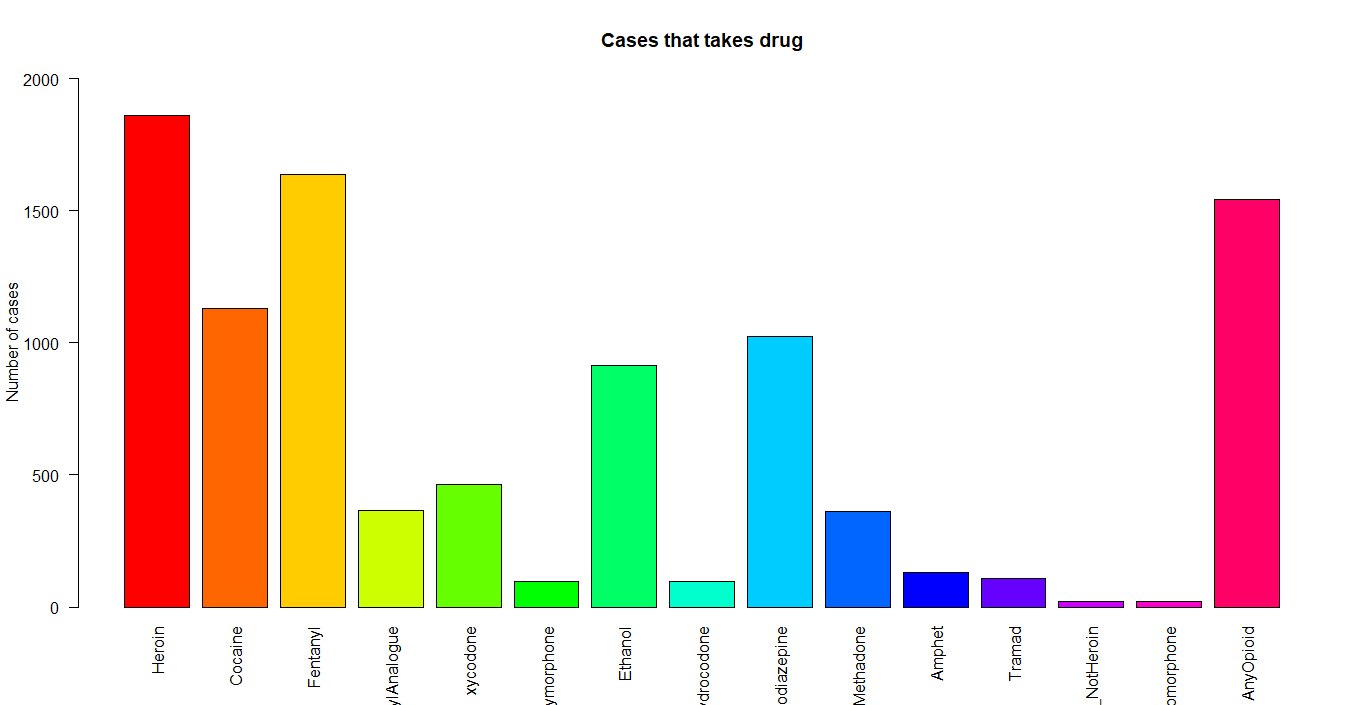
Observations :-

1)The highest country that is resident count and death country at the same time is (HARTFORD).

7)Determine what type of drug that is most taken



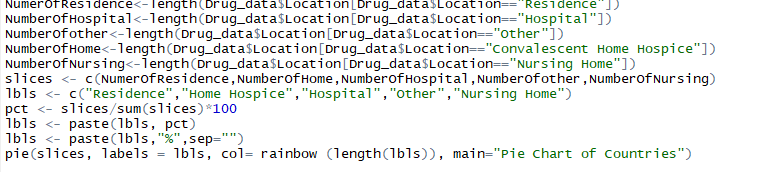


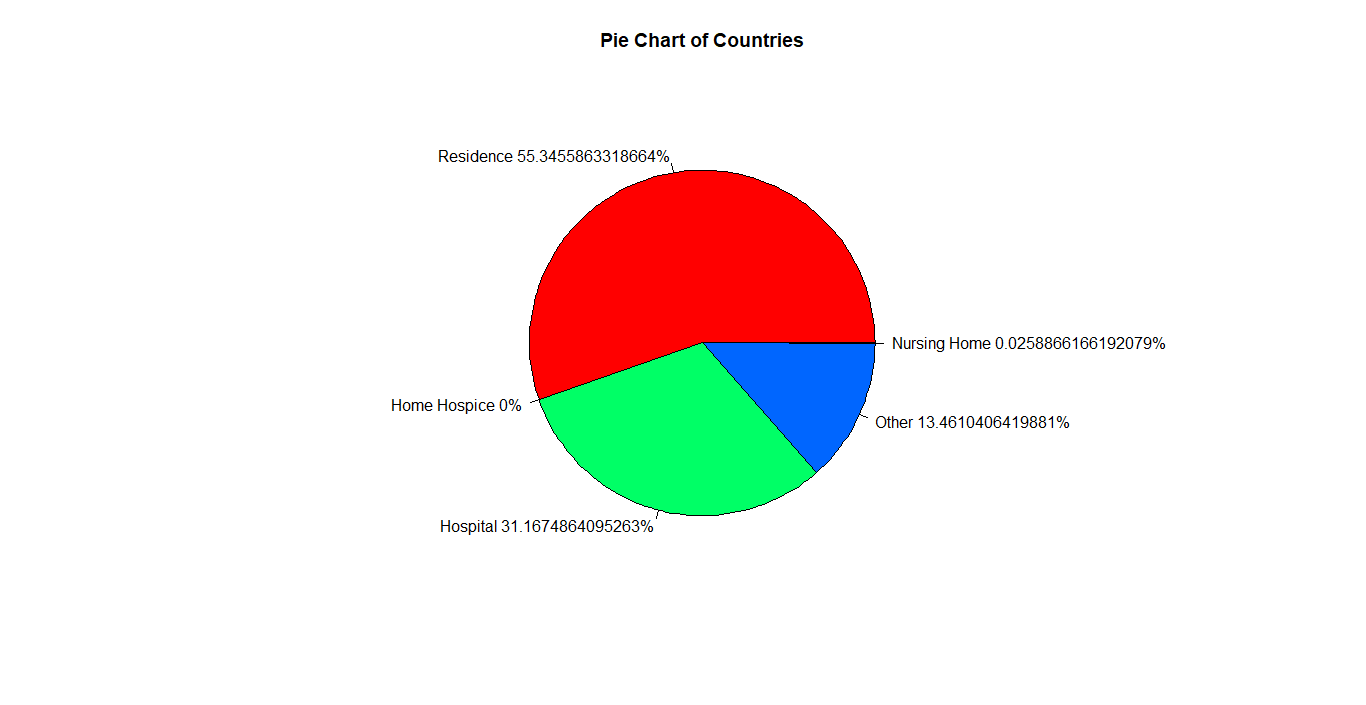


Observations :-

1)The most used drug is (Heroin then Any Opioid).

8)Determine percentage of each location that person die in.



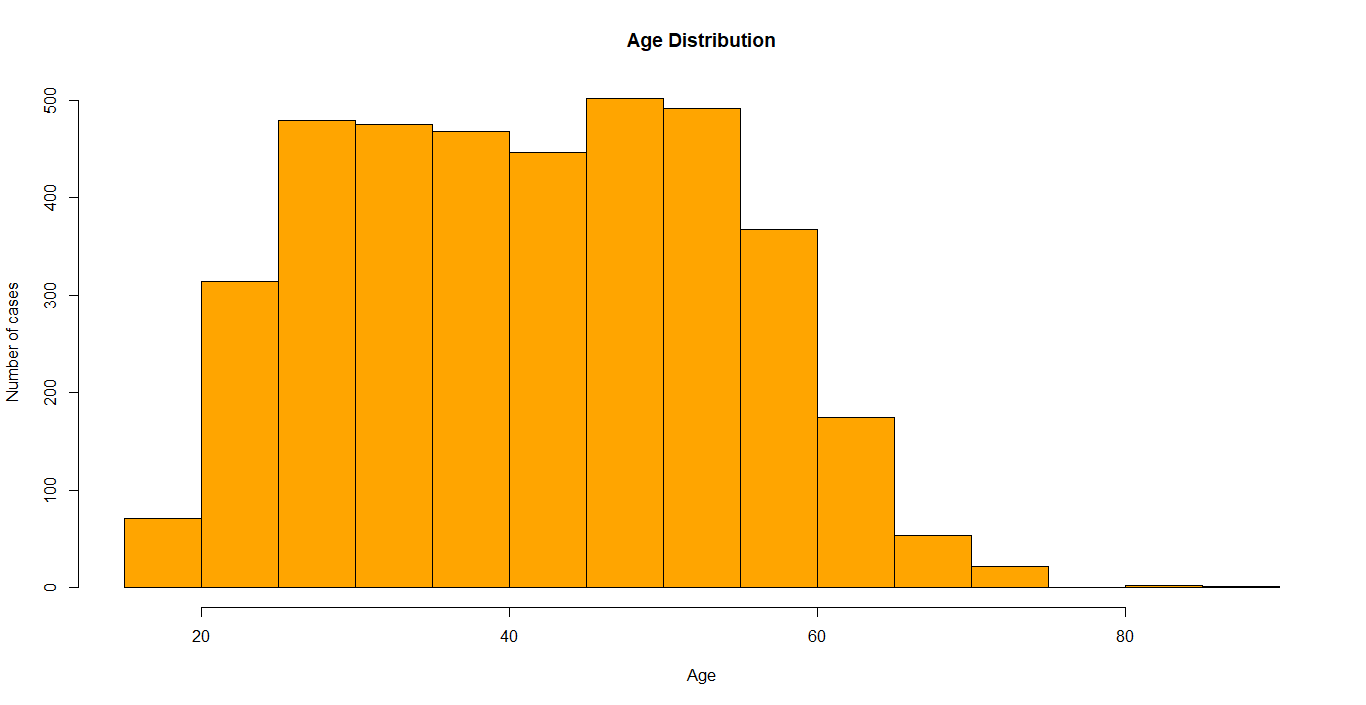


Observations :-

1)Most of people that takes drugs die in residence location.

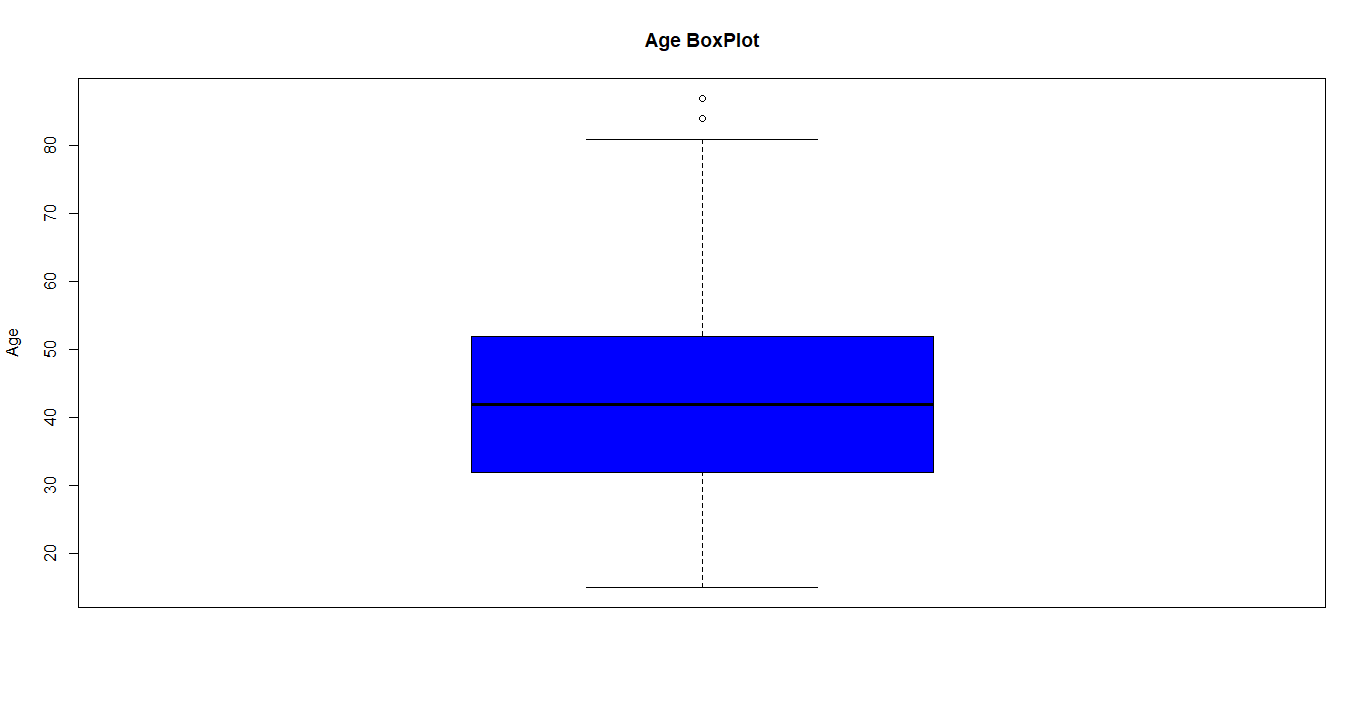
9)Know distribution of age.





10)Get outliers of Age using Boxplot

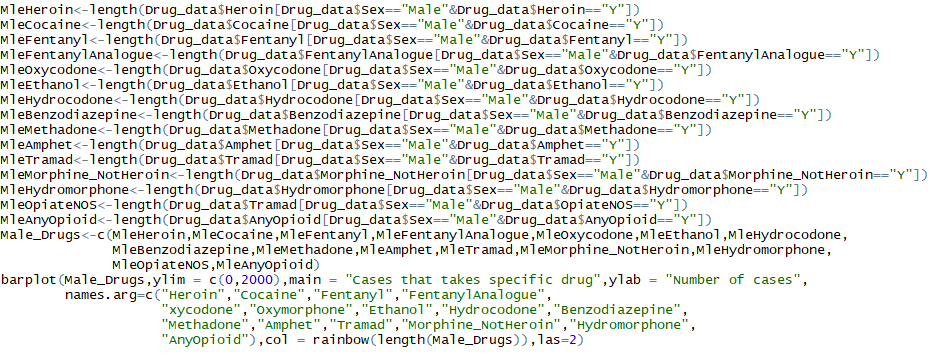


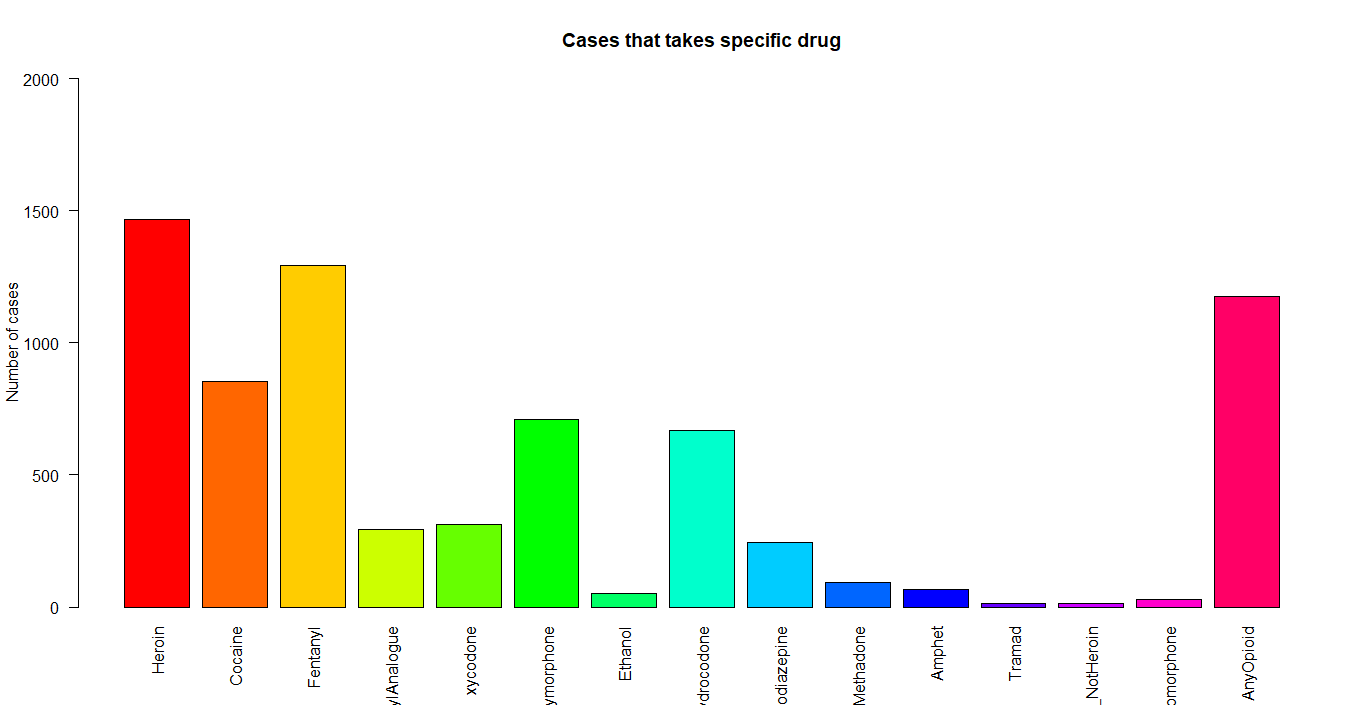


Observations :-

1)Mean age approximately 40 years old and outliers for age greater than 80 years old.

11)Determine the most common drug between males.



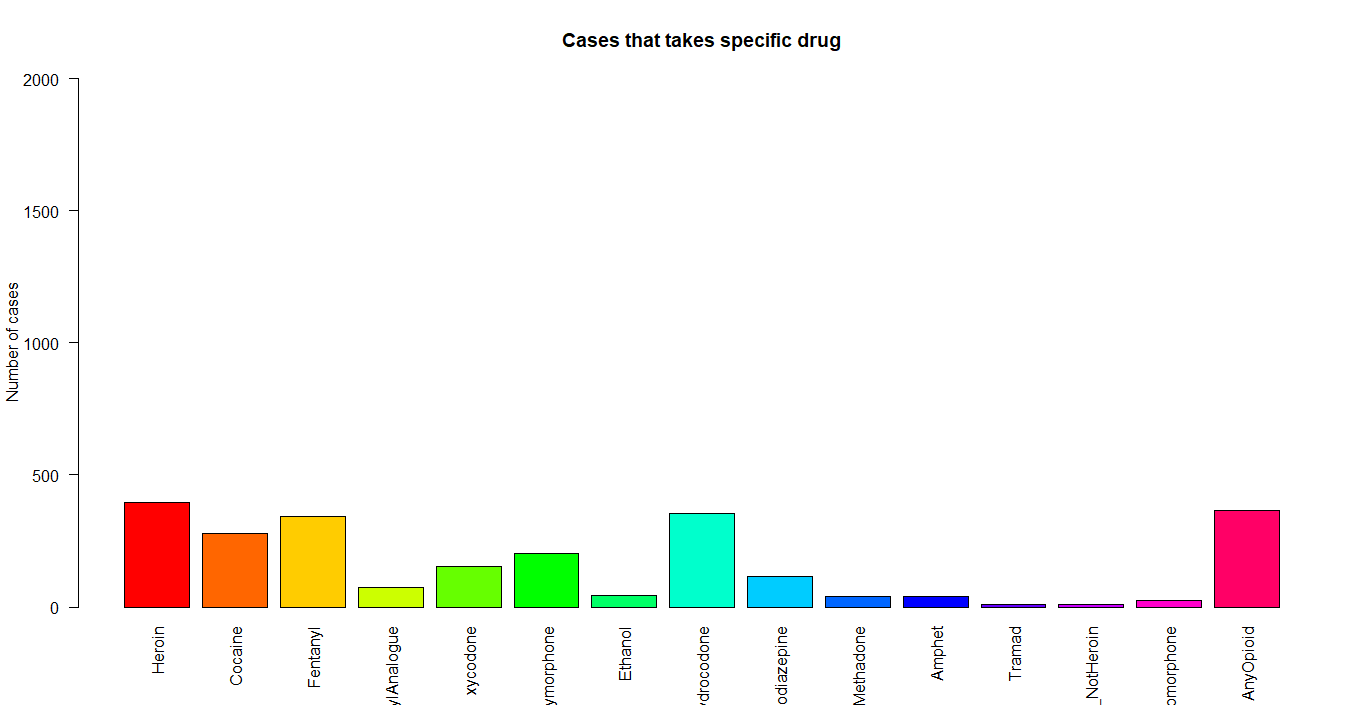


Observations :-

1)The most common drug between males is (Heroin).

12)Determine the most common drug between Females.





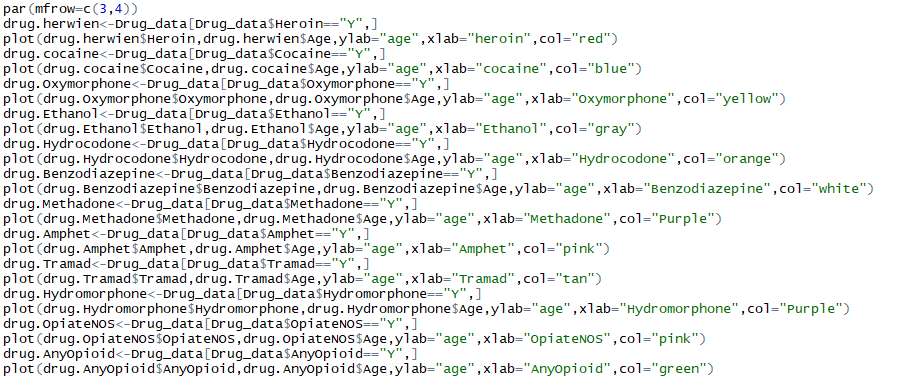
Observations :-

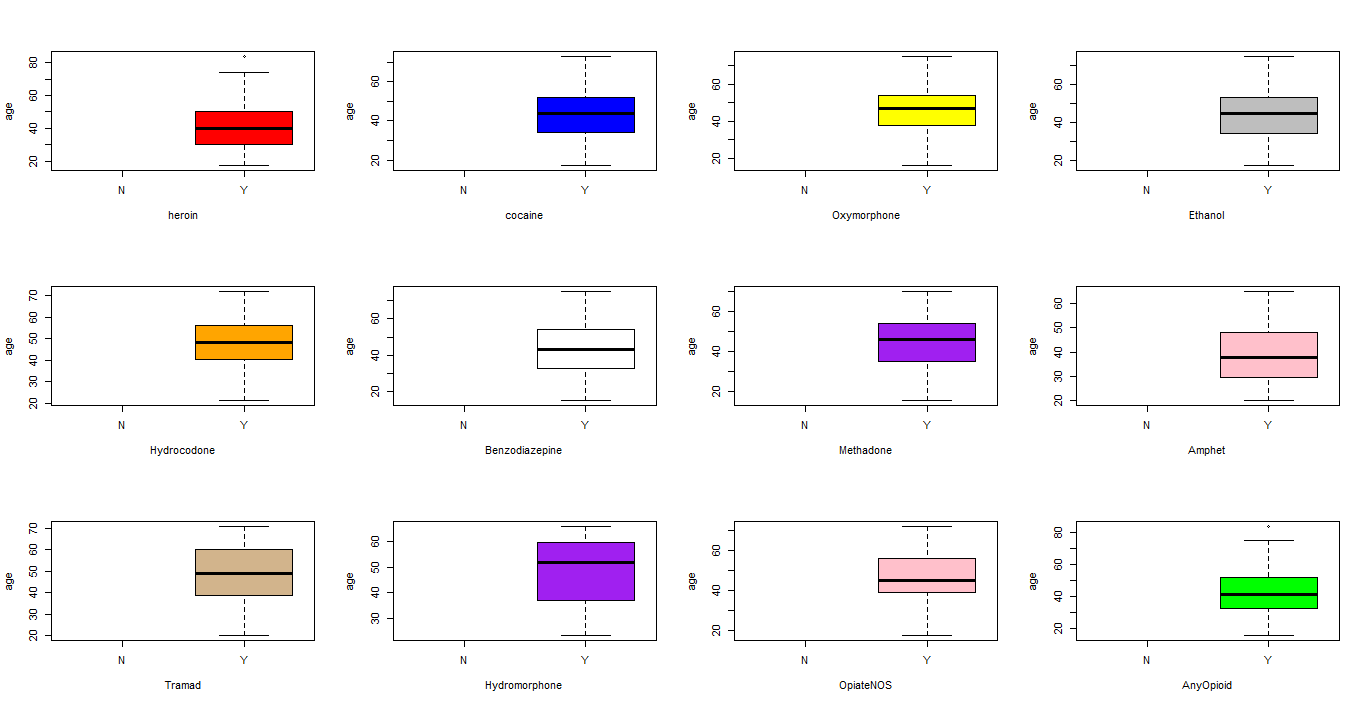
1)The most common drug between females is (Heroin or any opioid).

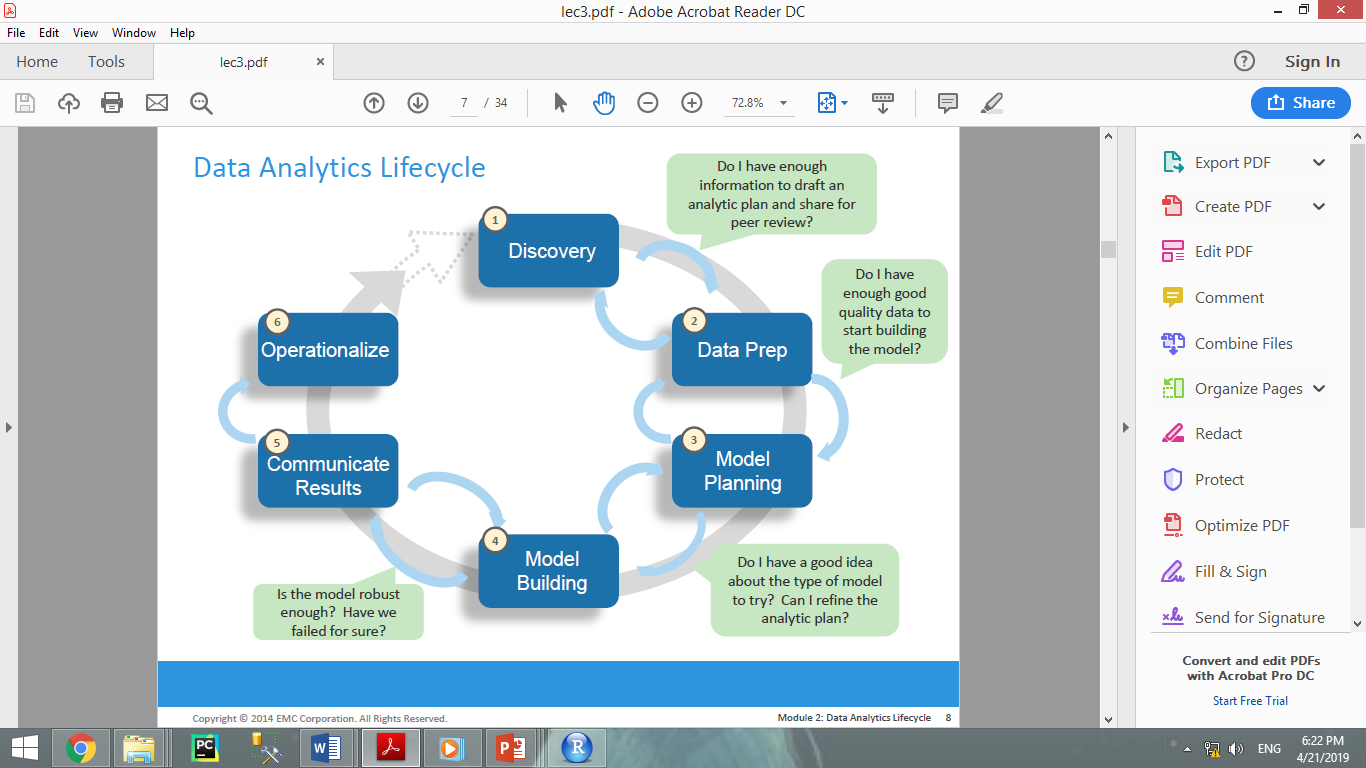
Note

It’s clearly obvious that number of males who take heroin is greater than number of females who take heroin or any opioid.

13) Determine median for age of people who take specific type of drug.







**Phase 3: Model planning**

**Hypothesis:**

H0: Manner of death is accident.

H1: Manner of death is pending.

H2: Manner of death is natural.

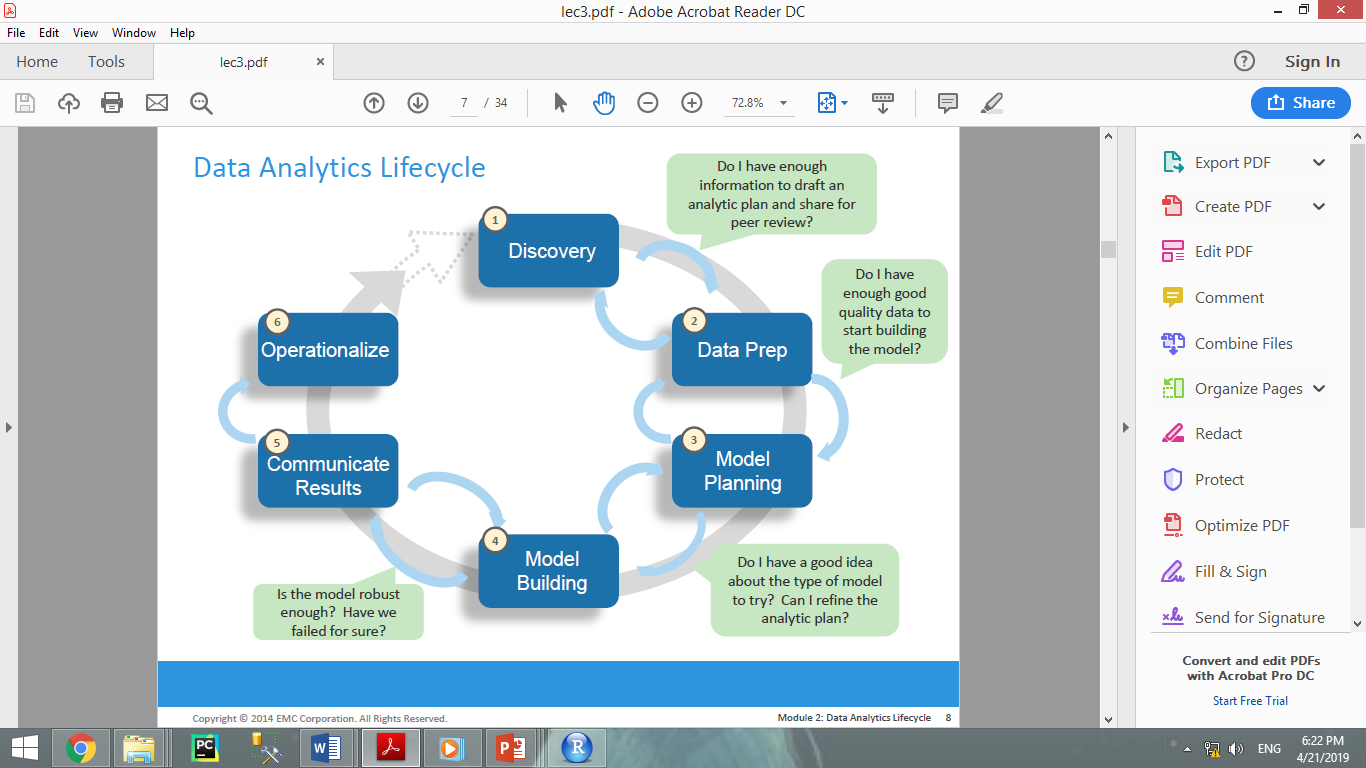
**Methods:**

Apriori and Naïve Bayesian classifier.

**variables inter-dependencies**

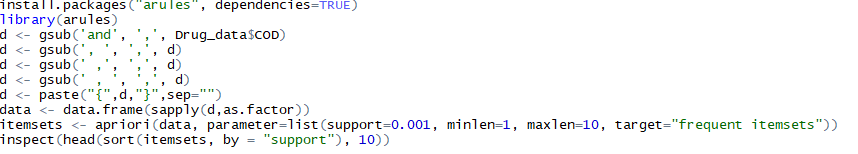
Manner of death depends of all other attributes.

Location if other depends on location

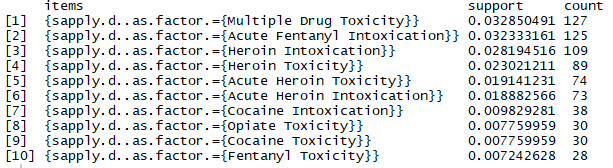
s

**Phase 4: Model Building**

**Apriori**



Result



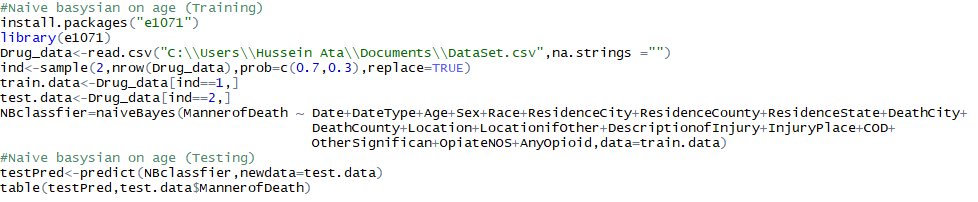
Observation

The most frequent COD is {Multiple Drug Toxicity} with support 0.032850491.

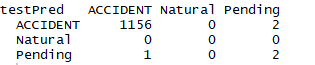
**Naïve Bayesian classifier**

Note :

We split dataset to 70% training and 30% for testing.



Result

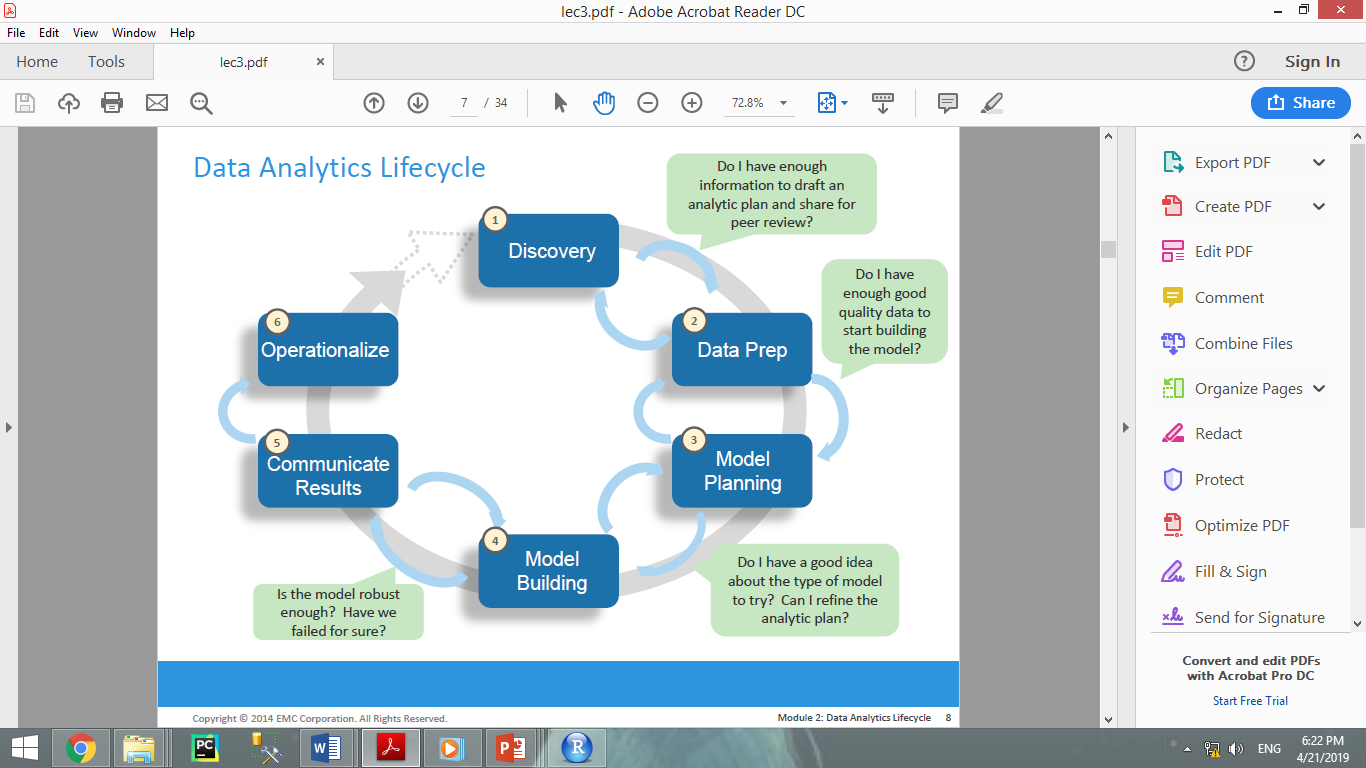


Accuracy

((1156+2)/(1156+2+1+2))\*100 = 99.74%.

Error rate

((1+2)/(1156+2+1+2))\*100 = 0.258%.



**Phase 5: Communicate Results**

Interpret the results :

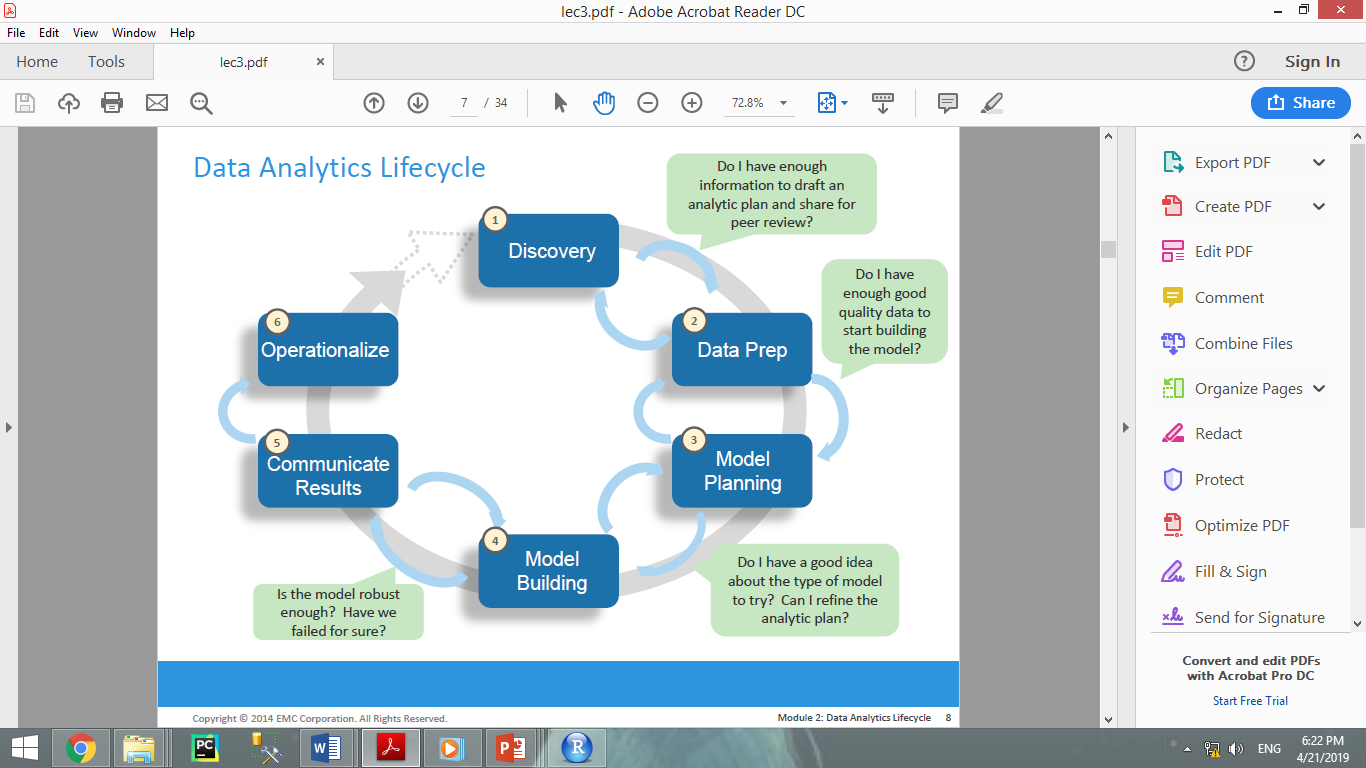
* Most people who take drugs die in accidents because of drugs.

Compare to IH’s from Phase 1:

* Initial hypnosis is not rejected.

Identify key findings :

1. People who are between (45-50) are die because of drugs.
2. Number of males who die because of drugs are greater than number of females.
3. HARTFORD Country have the highest number of people who die because of drugs.
4. People who are between (50-55) years old die during taking drugs.
5. People who are between ((30-35) , (47-52)) years old die because of accidents related to drugs.
6. White people are the most people that takes drugs.
7. The highest country that is resident count and death country at the same time is (HARTFORD).
8. The most used drug is (Heroin then Any Opioid).
9. Most of people that takes drugs die in residence location.
10. Mean age approximately 40 years old and outliers for age greater than 80 years old.
11. The most common drug between males is (Heroin).
12. The most common drug between females is (Heroin or any opioid).



**Phase 6: Operationalize**

Benefits:

* Now we are able to classify people who die because of drugs and know manner of death for

new given case and determine the most frequent drugs taken people in New England.

* We discover that many people die in accidents because of drugs then doctors need to be able

to focus on drug examination in case of accidents.

**References**

* <http://www.learnbymarketing.com/tutorials/naive-bayes-in-r/>
* https://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html
* https://catalog.data.gov/dataset/accidental-drug-related-deaths-january-2012-sept-2015