ASSIGNMENT 2 - Data Visualization

SEMESTER 1, 2020/2021

CSC 3305, SECTION: 01

DATA SCIENCE

LECTURER: DR. RAINI HASSAN

Prepared by:

Name	Matric number
TASNIM RAFIA	1725826
HASAN TANVEER MAHMOOD	1725413
BILLAH SYED MASHKUR	1723387
K. M ZUBAIR	1722931

Due Date: 28th December 2020

Table Of Content

Introduction:

Drug is a chemical substance, usually of a known structure, which generates a biological effect when consumed. Drugs are basically known as medicine. In several nations, there are various governmental offices concerned with the regulation and monitoring of the manufacturing and use of drugs and the enforcement of different drug laws. Overall, the word drug addiction is considered as a crime. In this data visualization exercise, we will try to conclude the prices of drugs according to their weights in several years and consumers of drugs according to age and total value impacting on the economy.

The Research Question:

The prices of the drugs are evidently detectable according to the weights. It is assumed based on the prices of past few years that next year prices will be high or low. Predict the drugs consumers according to age and drug types. In this case, we are going have an inspection on the state:

"Does Drug Purchasing Depend on its Cost?"

The Research Objective:

- Identifying the highest price of the drugs.
- Identify the amount of consumed drug based on age and quantity.
- Comparative consumption of drugs over the years.

Data Set:

In this assignment we are trying to do our analysis by using real data sets. We have collected our data from Malaysian's open data portal "data.gov.my" under the National Anti-Drug Agency (AADK) Drug prices according to weights , Statistics by Age Classification and Type of Drugs and Total value of foreclosure rm by type of drug 2020 . In the datasets we ignored the 1st few rows which is the title of the dataset. Rest of the dataset has been shown below.

```
> # Read Dataset
> drug_prices <- read.csv("C:/Users/tanve/OneDrive/Desktop/harga dadah mengikut berat
x.csv", skip = 1)
> head(drug_prices)
```

```
> # Read Dataset
> drug_prices <- read.csv("C:/Users/tanve/OneDrive/Desktop/harga dadah mengikut berat  x.csv", skip = 1)
> head(drug_prices)
  Bil          Tahun Jenis.Dadah...Substans Harga.Purata Unit.Belian  X X.1 X.2 X.3
1   1           2014          Cannabis Herb   RM3,000.00        1 KG NA  NA  NA  NA
2   2           2015          Cannabis Herb   RM2,600.00        1 KG NA  NA  NA  NA
3   3           2016          Cannabis Herb   RM2,000.00        1 KG NA  NA  NA  NA
4   4           2017          Cannabis Herb   RM3,000.00        1 KG NA  NA  NA  NA
5   5           2018          Cannabis Herb   RM2,400.00        1 KG NA  NA  NA  NA
6   6 Half year 2019          Cannabis Herb   RM2,000.00        1 KG NA  NA  NA  NA
> |
```

Figure 1

```
total_value <- read.csv("C:/Users/tanve/OneDrive/Desktop/jumlah-nilai-rampasan-rm-mengikut-jenis-dadah-2020.csv", skip = 2)
head(total_value)
```

```
> head(total_value)
  TAHUN HEROIN..RM..JUTA. GANJA..RM..JUTA. CANDU.MENTAH..RM..JUTA. CANDU.MASAK..RM..JUTA. SYABU..RM..JUTA. PIL.AMPHETAMINE.TYPE.STIMULANTS..ATS..DAN.LAIN.LAIN...RM..JUTA.
1  2014           26.405           1.735                 2e-04               0.0020          218.200                                                45.380
2  2015           49.226           4.505                 1e-03               0.0020          145.967                                                59.291
3  2016           14.423           8.084                 1e-03               0.0130           95.120                                                52.460
4  2017           31.815           6.897                 0e+00               0.0040          110.779                                               110.692
5  2018           20.341           4.740                 0e+00               0.0004          383.568                                               119.028
6  2019           24.040           1.669                 2e-04               0.0073          272.964                                               107.521
> |
```

Figure 2

```
>age_classify <- read.csv("C:/Users/tanve/OneDrive/Desktop/statistik mengikut pengkelasan umur  jenis dadah 2.csv", skip = 3)
>head(age_classify)
```

```
> age_classify <- read.csv("C:/Users/tanve/OneDrive/Desktop/statistik mengikut pengkelasan umur  jenis dadah 2.csv", skip = 3)
> head(age_classify)
        X OPIAT. METHAMPHETAMIN..crystalline. GANJA METHAMPHETAMIN..tablets. ATS.. PIL.PSIKOTROPIK... LAIN.LAIN.... JUMLAH
1 REMAJA     46                         322    26                       27    74                 NA             NA    495
2  BELIA  5,082                      10,553   551                    1,826  2282                 13             60  20367
3 DEWASA  2,810                       2,893   178                      533   516                  1             18   6949
4 JUMLAH  7,938                      13,768   755                    2,386 2,872                 14             78 27,811
> |
```

Figure 3

Data Preprocessing:

Fortunately, in the dataset there are no major issues, no redundant data and no out lairs. But there are some columns and rows which do not have values. So, we have a minor task for data preprocessing. However, for ease of understanding we can translate our columns Bahasa Melayu to English.

```
#Change the Column Name to understand better
names (drug_prices) <- c("Bil", "Years", "Type of Drug", "Average Price", "Perchase Unit")
```

```
> #ignore unwanted/null value columns
> drug_prices <- drug_prices[,c("Bil", "Years", "Type of Drug", "Average Price","Perchase Unit")]

> #view data
> head(drug_prices)
```

```
> head(drug_prices)
  Bil          Years  Type of Drug Average Price Perchase Unit
1   1           2014 Cannabis Herb    RM3,000.00          1 KG
2   2           2015 Cannabis Herb    RM2,600.00          1 KG
3   3           2016 Cannabis Herb    RM2,000.00          1 KG
4   4           2017 Cannabis Herb    RM3,000.00          1 KG
5   5           2018 Cannabis Herb    RM2,400.00          1 KG
6   6 Half year 2019 Cannabis Herb    RM2,000.00          1 KG
>
```

Figure 4.1

In figure 4.1, we just ignore the four columns which are full of NA values.

```
> #Change the Column Name to understand better
> names (total_value) <- c("Year", "Heroin", "Ganja", "Raw opium", "Opium Cooking","Sayabu", "AMPHETAMINE-TYPE PILLS/Other ")

> head(total_value)
```

```
> head(total_value)
  Year Heroin Ganja Raw opium Opium Cooking  Sayabu AMPHETAMINE-TYPE PILLS/Other
1 2014 26.405 1.735     2e-04        0.0020 218.200                      45.380
2 2015 49.226 4.505     1e-03        0.0020 145.967                      59.291
3 2016 14.423 8.084     1e-03        0.0130  95.120                      52.460
4 2017 31.815 6.897     0e+00        0.0040 110.779                     110.692
5 2018 20.341 4.740     0e+00        0.0004 383.568                     119.028
6 2019 24.040 1.669     2e-04        0.0073 272.964                     107.521
>
```

Figure 4.2

```
> #Translate Column Name
> names (age_classify) <- c("Age Category", "Opiat", "Methamphetamin(crystal)", "Ganja", "Methamphetamin(tab)","ATS", "Psychotropic Pill", "Other", "Amount")

> #Replace all N/A to 0
> age_classify[is.na(age_classify)] = 0
```

```
Age Category Opiat Methamphetamin(crystal) Ganja Methamphetamin(tab)   ATS Psychotropic Pill Other Amount
     REMAJA     46                     322    26                  27    74                 0     0    495
      BELIA  5,082                  10,553   551               1,826  2282                13    60  20367
     DEWASA  2,810                   2,893   178                 533   516                 1    18   6949
     JUMLAH  7,938                  13,768   755               2,386 2,872                14    78 27,811
```

Figure 4.3

```
> #Translate row Bahasa to English
> age_classify$`Age Category` <- str_replace_all(age_classify$`Age Category`,
"REMAJA\\s?", "Teenagers")
> age_classify$`Age Category` <- str_replace_all(age_classify$`Age Category`, "BELIA\\s?",
"Youth")
> age_classify$`Age Category` <- str_replace_all(age_classify$`Age Category`,
"DEWASA\\s?", "Adults")
> age_classify$`Age Category` <- str_replace_all(age_classify$`Age Category`,
"JUMLAH\\s?", "Amount")
```

```
> age_classify
  Age Category Opiat Methamphetamin(crystal) Ganja Methamphetamin(tab)   ATS Psychotropic Pill Other Amount
1    Teenagers    46                     322    26                  27    74                 0     0    495
2        Youth 5,082                  10,553   551               1,826  2282                13    60  20367
3       Adults 2,810                   2,893   178                 533   516                 1    18   6949
4       Amount 7,938                  13,768   755               2,386 2,872                14    78 27,811
> |
```

Figure 4.4

In 4.3 and 4.4 we make all null values to 0 and translate rows from Bahasa Melayu To English . Also, rename the columns from Bahasa Melayu to English to understand columns better.

```
> #Unique Drug name
> unique(drug_prices$'Type of Drug')
```

```
> unique(drug_prices$'Type of Drug')
 [1] "Cannabis Herb"                    "Heroin No.3"
 [3] "Opium"                            "Cocaine (Salt / Powder)"
 [5] "Kratom (Leaf)"                    "Kratom ( Liquid)"
 [7] "Methamphetamine (Crystalline)"    "Methamphetamine ( Pills / Tablets) "
 [9] "Ecstasy"                          "Ketamine"
[11] "Benzodiazepines / Erimin 5"       "Codeine"
```

Figure 4.5

This figure 4.5 showed the unique name of the drugs. So that we can easily identify the unique drugs.

```
> library(stringr)
> drug_prices$`Average Price (RM)` <- str_replace_all(drug_prices$`Average Price (RM)`, "RM\\s?", " ")
> head(drug_prices)
```

```
> head(drug_prices)
  Bil          Years  Type of Drug Average Price (RM) Perchase Unit
1   1           2014 Cannabis Herb           3,000.00          1 KG
2   2           2015 Cannabis Herb           2,600.00          1 KG
3   3           2016 Cannabis Herb           2,000.00          1 KG
4   4           2017 Cannabis Herb           3,000.00          1 KG
5   5           2018 Cannabis Herb           2,400.00          1 KG
6   6 Half year 2019 Cannabis Herb           2,000.00          1 KG
>
```

Figure 4.6

In figure 4.6 we replace all the rows which contain the value with RM. So we remove RM and keep the value only to analyses data better.

Data Wrangling:

In this phase, we are going to structurize the datasets into a linear form to understand and observe the data more swiftly and easily. Some unnecessary information or prefixes or suffixes may occur in the datasets which may assist in leaving the whole data unstructured or unorganized. We will do that part here which is called data wrangling:

```
> #removing extra white spaces where necessary
> drug_prices$'Bil' <- gsub(' ','',drug_prices$'Bil')
> drug_prices$'Years' <- gsub(' ','',drug_prices$'Years')
> drug_prices$'Average Price (RM)' <- gsub(' ','',drug_prices$'Average Price (RM)')


> #view data
> head(drug_prices)
```

```
> head(drug_prices)
  Bil         Years  Type of Drug Average Price (RM) Perchase Unit
1   1          2014 Cannabis Herb           3,000.00          1 KG
2   2          2015 Cannabis Herb           2,600.00          1 KG
3   3          2016 Cannabis Herb           2,000.00          1 KG
4   4          2017 Cannabis Herb           3,000.00          1 KG
5   5          2018 Cannabis Herb           2,400.00          1 KG
6   6 Halfyear2019 Cannabis Herb           2,000.00          1 KG
```

Figure 5.1

```
> #removing unnecessary info from the front of the years column
> drug_prices$'Years' <- gsub('Halfyear','',drug_prices$'Years')

> #view data
> head(drug_prices)
```

```
  Bil Years  Type of Drug Average Price (RM) Perchase Unit
1   1  2014 Cannabis Herb           3,000.00          1 KG
2   2  2015 Cannabis Herb           2,600.00          1 KG
3   3  2016 Cannabis Herb           2,000.00          1 KG
4   4  2017 Cannabis Herb           3,000.00          1 KG
5   5  2018 Cannabis Herb           2,400.00          1 KG
6   6  2019 Cannabis Herb           2,000.00          1 KG
>
```

Figure 5.2

In figure 5.2, we can notice that the string 'Half Year' is eliminated as it was data which is not useful to us. So we structurized and data wrangled it.

```
>#Now lets structurize all the values in a consistent capitalization
> drug_prices$'Bil' <- str_to_sentence(drug_prices$'Bil')
> drug_prices$'Years' <- str_to_sentence(drug_prices$'Years')
> drug_prices$'Type of Drug' <- str_to_sentence(drug_prices$'Type of Drug')
> drug_prices$'Average Price (RM)' <- str_to_sentence(drug_prices$'Average Price (RM)')
> drug_prices$'Perchase Unit' <- str_to_sentence(drug_prices$'Perchase Unit')
> #view the final data
> head(drug_prices)
```

```
> head(drug_prices)
  Bil Years  Type of Drug Average Price (RM) Perchase Unit
1   1  2014 Cannabis herb           3,000.00          1 kg
2   2  2015 Cannabis herb           2,600.00          1 kg
3   3  2016 Cannabis herb           2,000.00          1 kg
4   4  2017 Cannabis herb           3,000.00          1 kg
5   5  2018 Cannabis herb           2,400.00          1 kg
6   6  2019 Cannabis herb           2,000.00          1 kg
>
```

Figure 5.3

Finally, in figure 5.3 we used data consistent capitalization to form a standard and consistent form of a specific data. Therefore, there will be no multiple forms of a single data.

```
#Removing extra decimal places and reducing it to a constant 2 decimal place

> total_value$Heroin <- format(round(total_value$Heroin, 2))
> total_value$Ganja <- format(round(total_value$Ganja, 2))
> total_value$'Raw opium' <- format(round(total_value$'Raw opium', 2))
> total_value$'Opium Cooking' <- format(round(total_value$'Opium Cooking', 2))
> total_value$'Sayabu' <- format(round(total_value$'Sayabu', 2))
> total_value$'AMPHETAMINE-TYPE PILLS/Other ' <-
format(round(total_value$'AMPHETAMINE-TYPE PILLS/Other ', 2))


> #view the final data
> head(total_value)
```

#Before wrangling:

```
> head(total_value)
  Year Heroin Ganja Raw opium Opium Cooking  Sayabu AMPHETAMINE-TYPE PILLS/Other
1 2014  26.41 1.735     2e-04        0.0020 218.200                       45.380
2 2015  49.23 4.505     1e-03        0.0020 145.967                       59.291
3 2016  14.42 8.084     1e-03        0.0130  95.120                       52.460
4 2017  31.82 6.897     0e+00        0.0040 110.779                      110.692
5 2018  20.34 4.740     0e+00        0.0004 383.568                      119.028
6 2019  24.04 1.669     2e-04        0.0073 272.964                      107.521
```

#After wrangling:

```
  Year Heroin Ganja Raw opium Opium Cooking Sayabu AMPHETAMINE-TYPE PILLS/Other
1 2014  26.41  1.74         0          0.00 218.20                        45.38
2 2015  49.23  4.50         0          0.00 145.97                        59.29
3 2016  14.42  8.08         0          0.01  95.12                        52.46
4 2017  31.82  6.90         0          0.00 110.78                       110.69
5 2018  20.34  4.74         0          0.00 383.57                       119.03
6 2019  24.04  1.67         0          0.01 272.96                       107.52
> 
```

Figure 5.4

In this figure (5.4), we can observe that the values are more organized and structured now as they have been rounded with 2 decimal places only. This helps to manipulate the data more easily in a structured manner. Notice that the column 'Raw opium' has the values approximate to ~ 0, so we can calculate the value to be 0 for better results.

```
#removing ',' sign to manipulate with data more efficiently in case of calculation

> age_classify$'Opiat' <- gsub(',',",age_classify$'Opiat')
> age_classify$'Methamphetamin(crystal)' <-
gsub(',',",age_classify$'Methamphetamin(crystal)')
> age_classify$'Methamphetamin(tab)' <- gsub(',',",age_classify$'Methamphetamin(tab)')
> age_classify$'ATS' <- gsub(',',",age_classify$'ATS')
> age_classify$'Amount' <- gsub(',',",age_classify$'Amount')


> #view the final data
> head(age_classify)
```

```
> head(age_classify)
  Age Category Opiat Methamphetamin(crystal) Ganja Methamphetamin(tab)  ATS Psychotropic Pill Other Amount
1    Teenagers    46                     322    26                  27   74                0     0    495
2        Youth  5082                   10553   551                1826 2282               13    60  20367
3       Adults  2810                    2893   178                 533  516                1    18   6949
4       Amount  7938                   13768   755                2386 2872               14    78  27811
> age_classify
```

Figure 5.5

In the above figure (5.5), we can see that removing the ',' sign could make the data more useful in order to make calculations. Other than that, the dataset looks clean and structured overall.

Rewrite the data:

```
#Now lets re-write the whole datasets in new files

> df <- data.frame(age_classify)
> write.csv(df,"C:\\Users\\mashk\\Desktop\\age_classify.csv", row.names = FALSE)

> df <- data.frame(drug_prices)
> write.csv(df,"C:\\Users\\mashk\\Desktop\\drug_prices.csv", row.names = FALSE)

> df <- data.frame(total_value)
> write.csv(df,"C:\\Users\\mashk\\Desktop\\total_value.csv", row.names = FALSE)
```

Data Analysis:

1. Age:

In order to enable data visualization better and more efficiently, we prepared some charts. The chart here (chart 1) is showing the separation of drug consumption based on ages. The greatest number of consumers are youth generation; whereas, the least is teenagers.

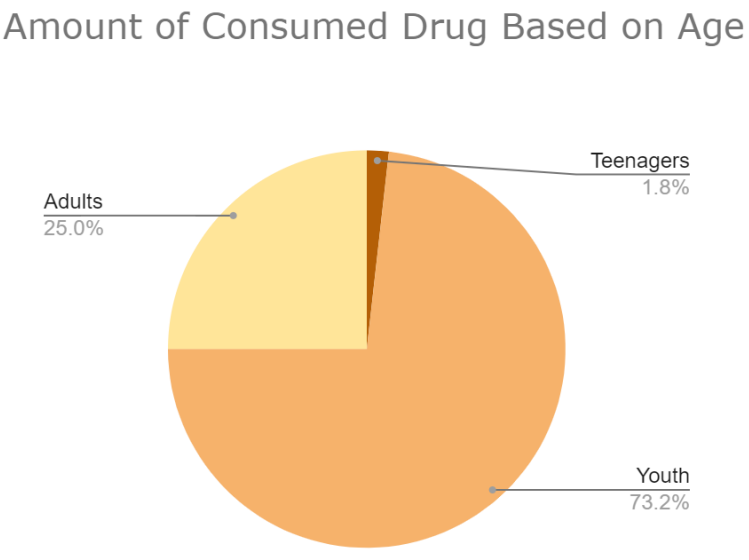


Figure 6.1

2. Amount:

The chart below (figure 6.2) is a bar chart that shows the total amount of drugs sold out over the years based on quantity. Ecstasy is the most popular drug so far in all generations that might be highlighted from the graph.

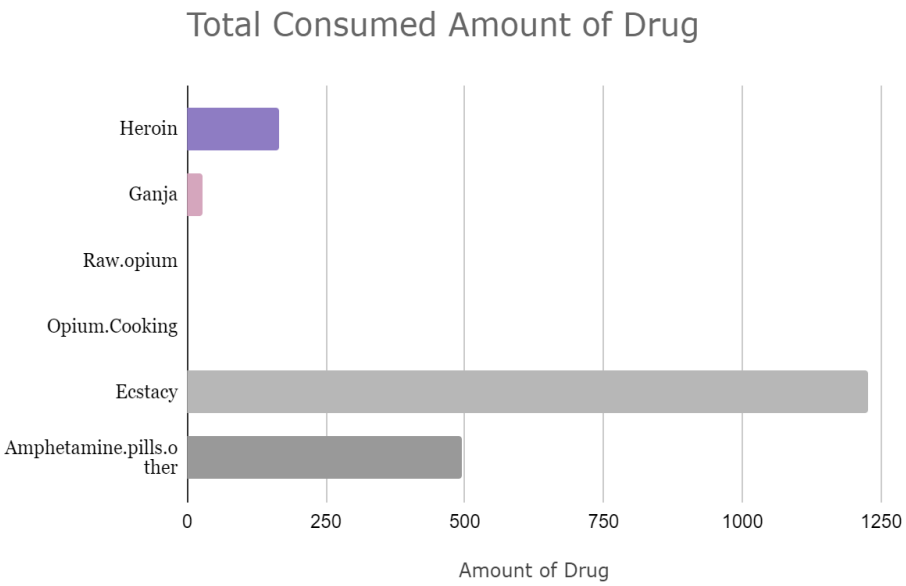


Figure 6.2

3. Years:

The line graph (6.3) here will indicate the consumption of drugs over the years. Heroine is the most, and Opium is the least consumed drug among all of them.

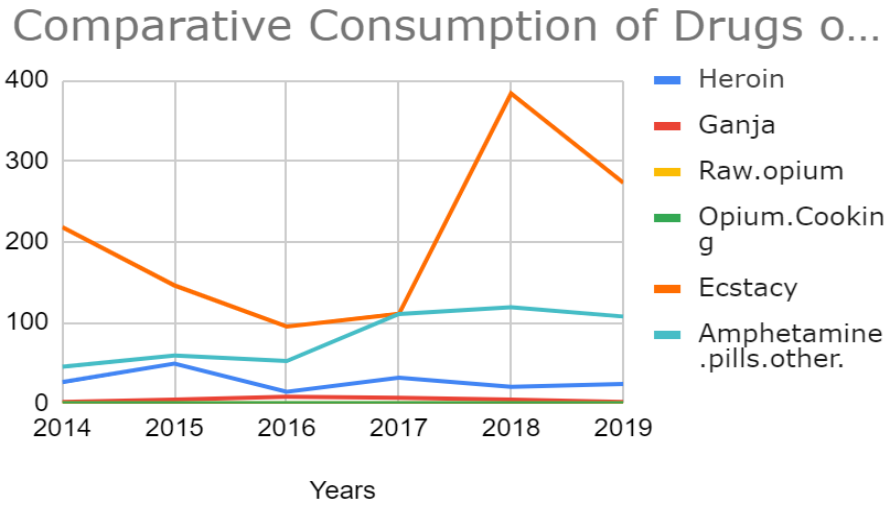


Figure 6.3

4. **Price:**

From 2014 until 2019, the most expensive drug is Cocaine. The least costly one is Ecstasy. Cocaine was most expensive in the mid time of 2014 and 2019. The cost of Methamphetamine (crystalline) had an unstable price.

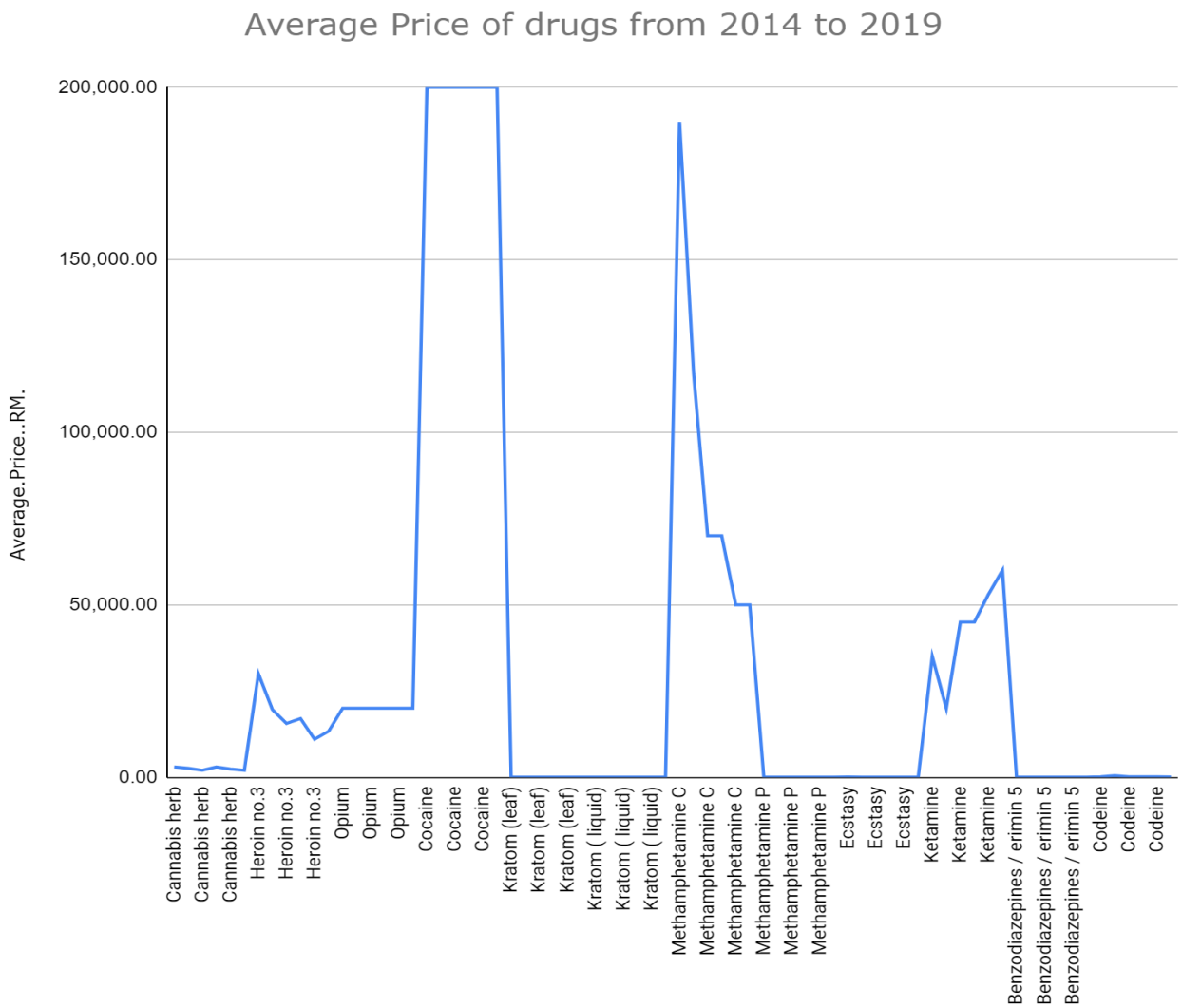


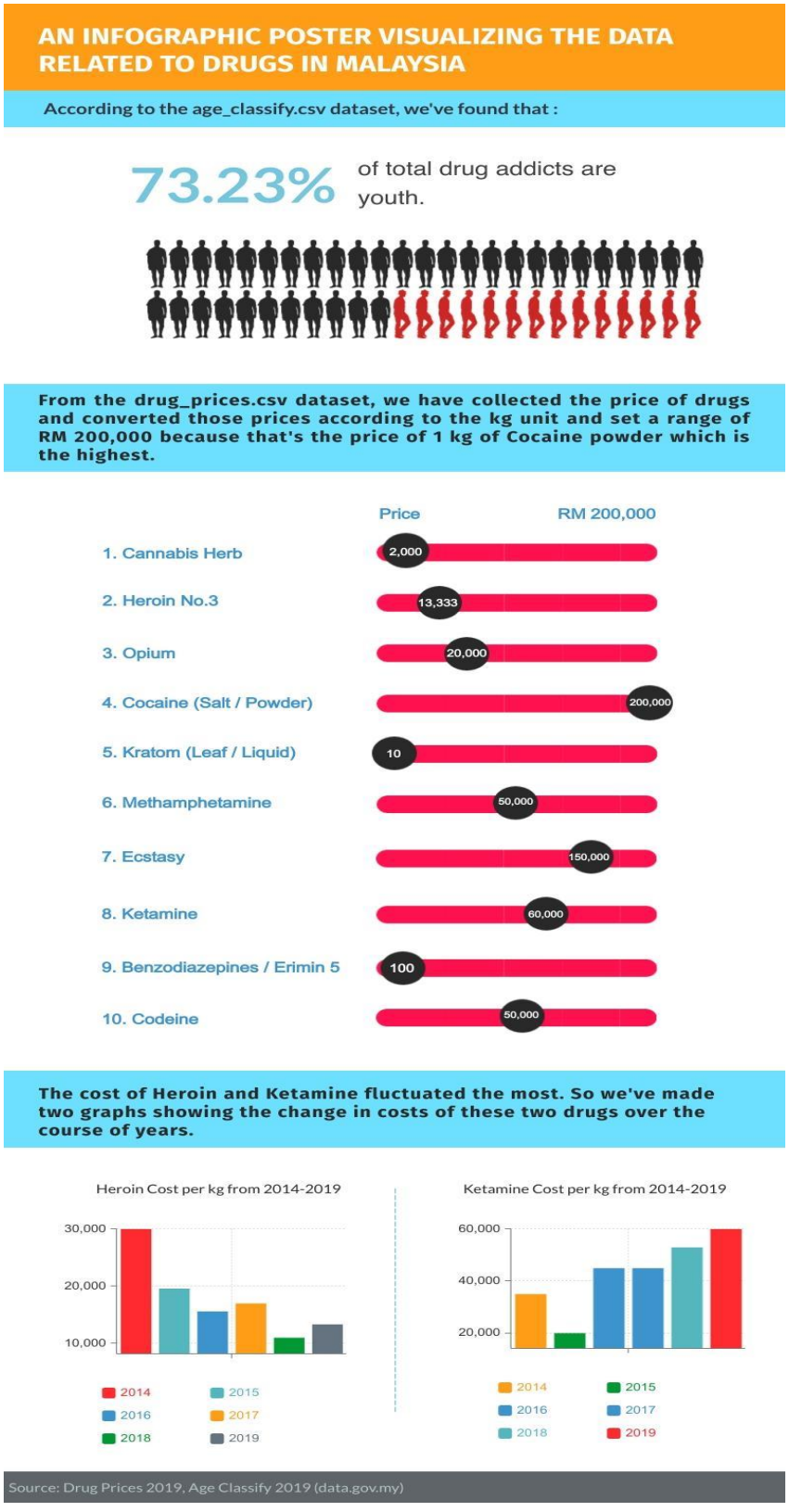
Figure 6.4

Finding:

From the analyzed data, it is observable that, the less the price of drugs, the more it indicates its consumption. For example, from graph 4, it is noticeable that Ecstasy is offered in exchange for a smaller amount of money. Also, chart 3 shows that the most consumed drug product is ecstasy. Thus, we can reach the conclusion that **Purchasing of Drug depends on its cost.**

Data Visualization:

Judging by the analysis, it can be seen that most of the drug addicts are youth and adults and the price of these drugs vary from time to time. But the cost of drugs might affect its consumption. We've used several charts in our infographic poster to support the notion of our project.



References:

- https://www.data.gov.my/data/ms_MY/dataset/jumlah-nilai-rampasan-mengikut-jenis-dadah/resource/98754e4d-fbc5-4862-8870-d4505940dda9

- https://www.data.gov.my/data/ms_MY/dataset/statistik-mengikut-pengkelasan-umur-dan-jenis-dadah/resource/4a5453c3-0dd9-46ea-ad4b-be15431214b1

- https://www.data.gov.my/data/ms_MY/dataset/harga-purata-dadah-mengikut-berat-dan-jenis-dadah/resource/ec5a3fb1-ff8b-4bec-87ab-001b021b5ea0

❑ **Work Distribution:**

Completed by:	**Tasks**
TANVEER MAHMOOD HASAN	Introduction, Research objective, Dataset , Data Preprocessing.
TASNIM RAFIA	Research question, Research objective, Data Analysis, Finding.
BILLAH SYED MASHKUR	Data Wrangling, Rewrite dataset.
K. M. ZUBAIR	Data Visualization, Infographic Poster.