

18BCE0272, Nitin Ranjan

DATA VISUALISATION DIGITAL ASSIGNMENT 2

Question.

Take any application dataset of your choice.

1. List all the attributes
2. Describe the type of each attributes.
3. Identify any two graph or plot suitable for the dataset.
4. For the identified graph, write the following
 - Data type
 - Mark
 - Channel
 - Task
 - Scalability
5. Plot the graph with detailed information.

Dataset:

The dataset I have picked up is the data of monthly sale of medicines between January 2014 and October 2019.

It is available on - <https://data.world/liz-friedman/covid-19-impact-on-education>

Or, I have uploaded the dataset here on GitHub: <https://github.com/NitinR2510/Data-Visualisation-Class/blob/main/Assignment%202/salesmonthly.csv>

Exploratory Data Analysis:

The dataset has 6 columns

```
> head(pd)
  date M01AB M01AE N02BA N05B N05C
1 31-01-2014 127.69 99.090 152.100 354 50
2 28-02-2014 133.32 126.050 177.000 347 31
3 31-03-2014 137.44 92.950 147.655 232 20
4 30-04-2014 113.10 89.475 130.900 209 18
5 31-05-2014 101.79 119.933 132.100 270 23
6 30-06-2014 112.07 94.710 122.900 323 23
```

Date -> The date on which data was collected

Just for understanding, let me enlist the drugs each code stands for.

M01AB -> Pain Relievers

M01AE -> Anti-inflammatory drugs

N02BA -> non-steroidal anti-inflammatory drugs

N05B -> Anxiolytics

N05C -> Sedatives

What type of data is present?

Most of the data is thus quantitative except for date which is a qualitative data type.

All the columns represent sales. Sales data is named, ordered, has proportionate intervals between different numbers i.e. in terms of absolute number of goods sold and revenue generated by the same and can accommodate an absolute zero – Ratio (measurement).

Date however is ordinal. We have no absolute zero for the same. However, in our case, it might actually be treated as ratio, because the first data point serves as the origin of the data.

What graphs can be useful here?

Now, we are dealing with sales of 6 different commodities over a period of time.

One of the simplest plots that can be used for the purpose is a line graph. 5 lines that independently depict the sales figure against the dates on which the data was collected. This would help in visualizing –

1. Independent sales figure each day for all drug categories.
2. Pattern in which the sales varied over time.
3. A comparison between the sales patterns of various drug categories involved.

A scatter plot is a good choice for the

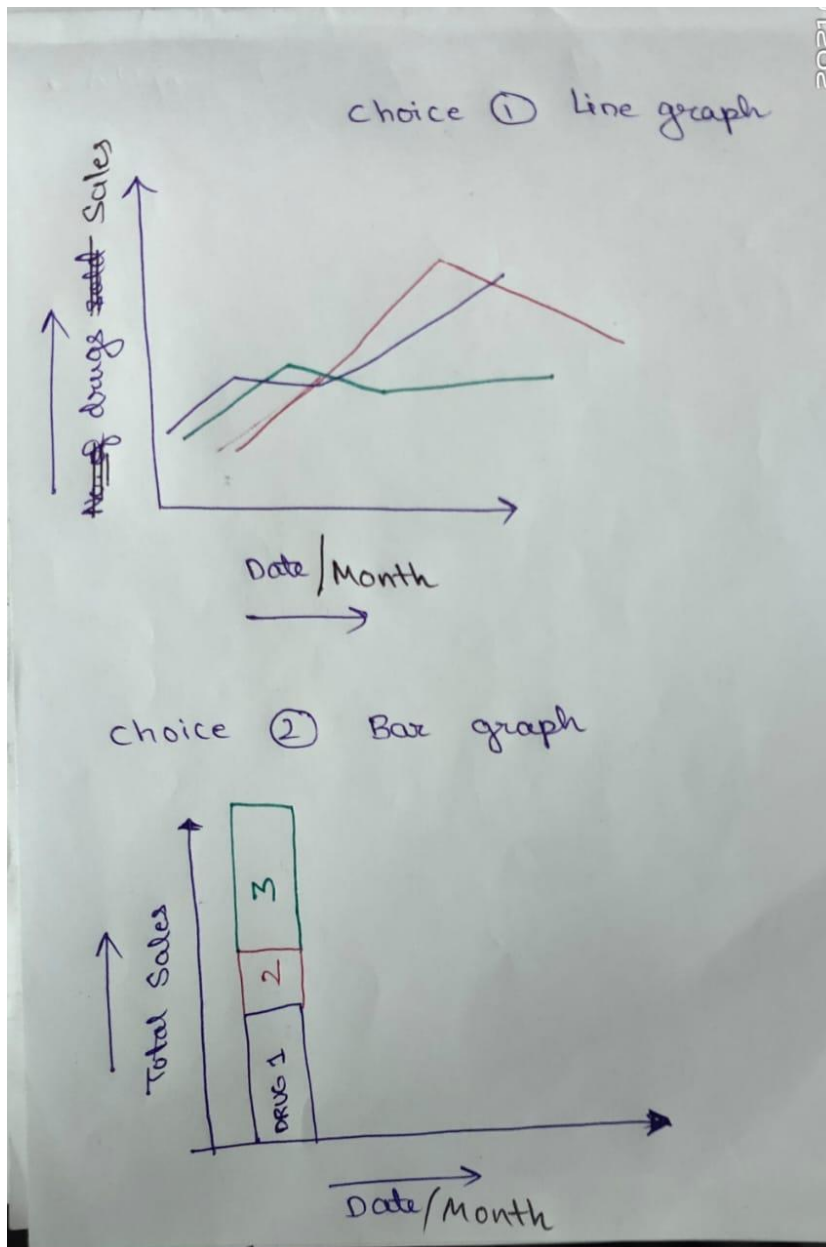
The other choice might be a segmented/grouped bar graph.

1. The bar graph is not as good a choice as the line graph as it shall lead to a strain upon the client if the client wishes to understand the data of sales in absolute terms.
2. However, it is a very good choice if the aim of the organization is compare the absolute sales per day among all drug categories.

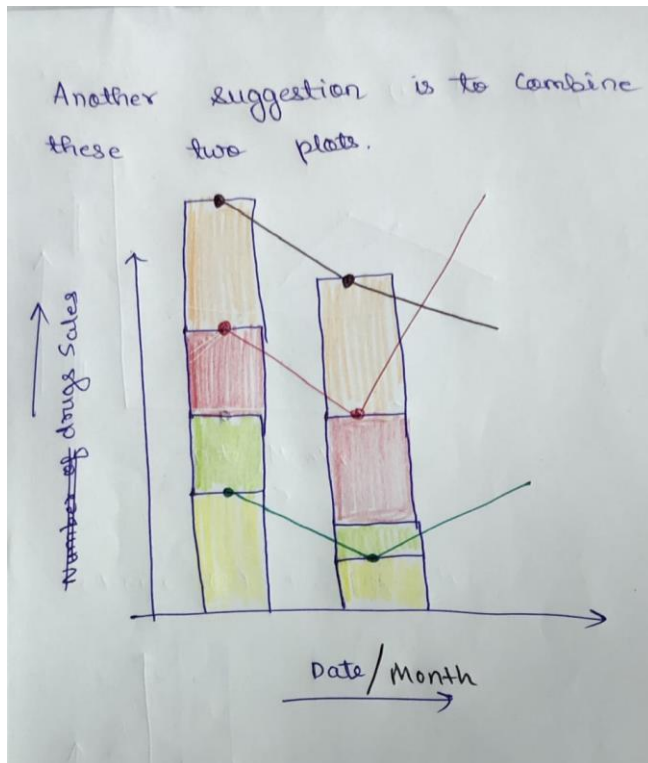
The following table explores both the graphs in detail

Graph	Mark	Channel	Task	Scalability
Line graph	Solid dots, solid squares and solid triangles to identify the points on the plane where the recorded data-point lies.	<ol style="list-style-type: none"> 1. Colour shall behave as a channel to distinguish lines and hence medicinal categories. 2. The height of the data point from the axis that represents month depicts the weight of the data point or in our case, the value of the sales. 	<p>The graph successfully achieves task abstractions of overview, zoom and filter i.e. the viewer can get an overview of the data, pick up the item of interest and ignore the rest. However, the task of details on demand is only partially met and its fulfillment depends on the detail the user is seeking.</p> <p>The graph however perfectly illustrates the pattern of sales of various drugs.</p>	<p>A very large number of categorical values may make the visualization look cluttered.</p> <p>In case of line graph, there is only one possible category – the name of medicines and the total number of values it can take is 5.</p> <p>So, this should not have any effect on the current visualization.</p>
Bar graph (segmented/grouped)	Small lines separate regions in bars.	<ol style="list-style-type: none"> 1. Colour acts as a channel to differentiate categories of medicine. 2. Gaps between the bars differentiate date/months. 3. Size of the bars is clearly the third channel that will provide the viewer with a rough quantitative information. 	<p>The graph achieves overview, zoom and filter. Again, as far as the task abstraction of details on demand is concerned, the graph achieves it partially.</p> <p>The graph is however meant to compare daily sales among drug groups and that is achieved quietly nicely by the same.</p>	<ol style="list-style-type: none"> 1. In case of bar graph, there are two possible categories – the name of medicine and the date. 2. A very large number of medicinal categories may make the visualization look overwhelmed with information and further abstraction may be needed. <p>Hence, not very scalable. However, this should not affect the current dataset.</p>

Rough Estimate of how it should look like:



Just a suggestion: As already discussed, neither of the graphs are very scalable. However, since in the current dataset, there are not a lot of categorical variables or values involved, a hybrid graph of lines and bars may be used.



But this is even less scalable as compared to its parent curves and hence, it is rejected.

Modelling the curves in R

Let me now code these curves in R.

Line Graph

CODE

```
library(ggplot2)
```

```
library(dplyr)
```

```
pd = read.csv("C:/Users/lenovo/Documents/GitHub/Data-Visualisation-Class/Assignment  
2/salesmonthly.csv")
```

```
pd = na.omit(pd)
```

```
#line graph
```

```
png(file = "sales.png")
```

```
plot(pd$M01AB, type="o", pch = 19, col="blue", xlab="Month from Jan2014", ylab = "Sales of Drugs",  
main = "Drug Sales (Feb2014- Oct2019)@18BCE0272", ylim = c(0,250), lty=6)
```

```
lines(pd$M01AE, type="o", pch = 18, col = "maroon", lty = 5)
```

```
lines(pd$N02BA, type="o", pch = 19, col = "green", lty = 4)
```

Nitin Ranjan, 18BCE0272

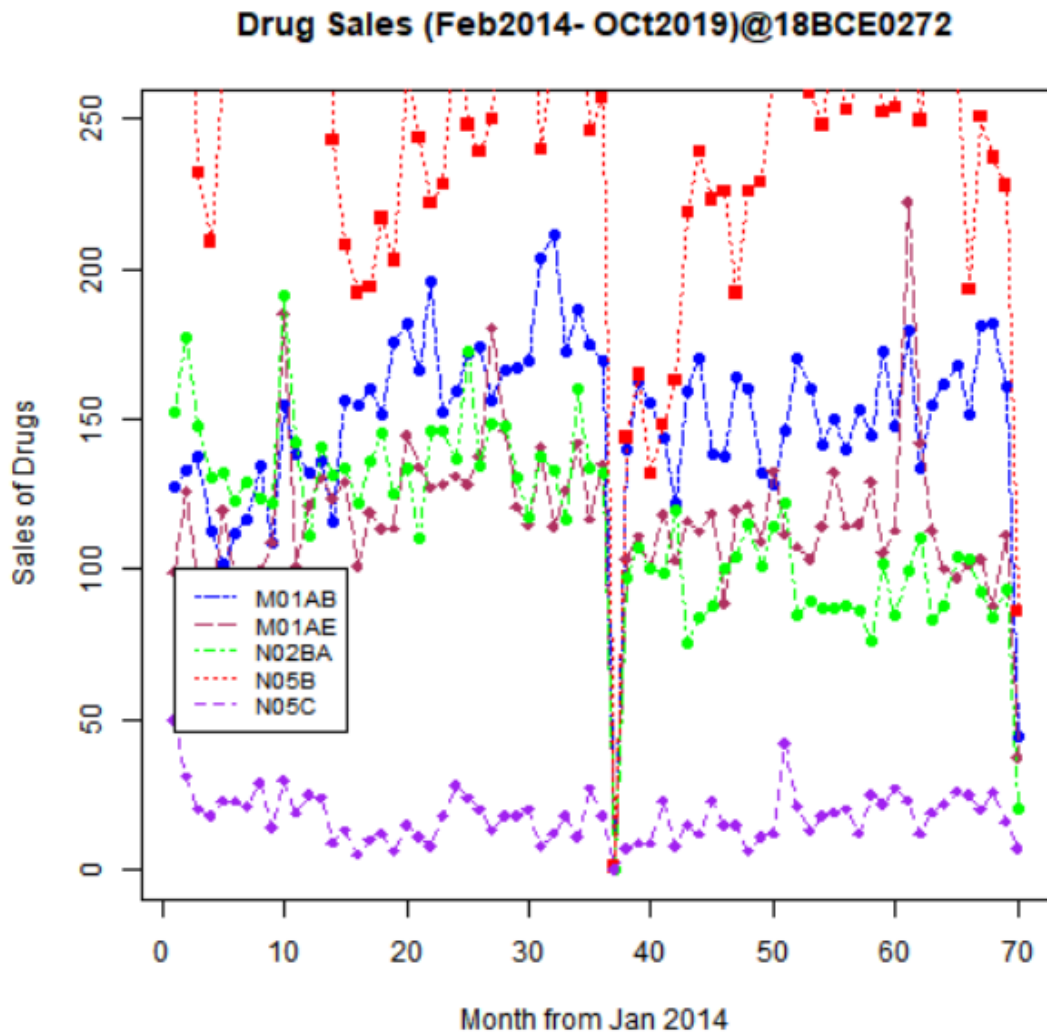
Data Visualisation DA 2

```
lines(pd$N05B, type="o", pch = 15, col = "red", lty = 3)
```

```
lines(pd$N05C, type="o", pch = 18, col = "purple", lty = 2)
```

```
legend(1,100, legend = c("M01AB","M01AE", "N02BA","N05B","N05C"),  
col=c("blue","maroon","green","red","purple"), lty=6:2, cex=0.8)
```

```
dev.off()
```

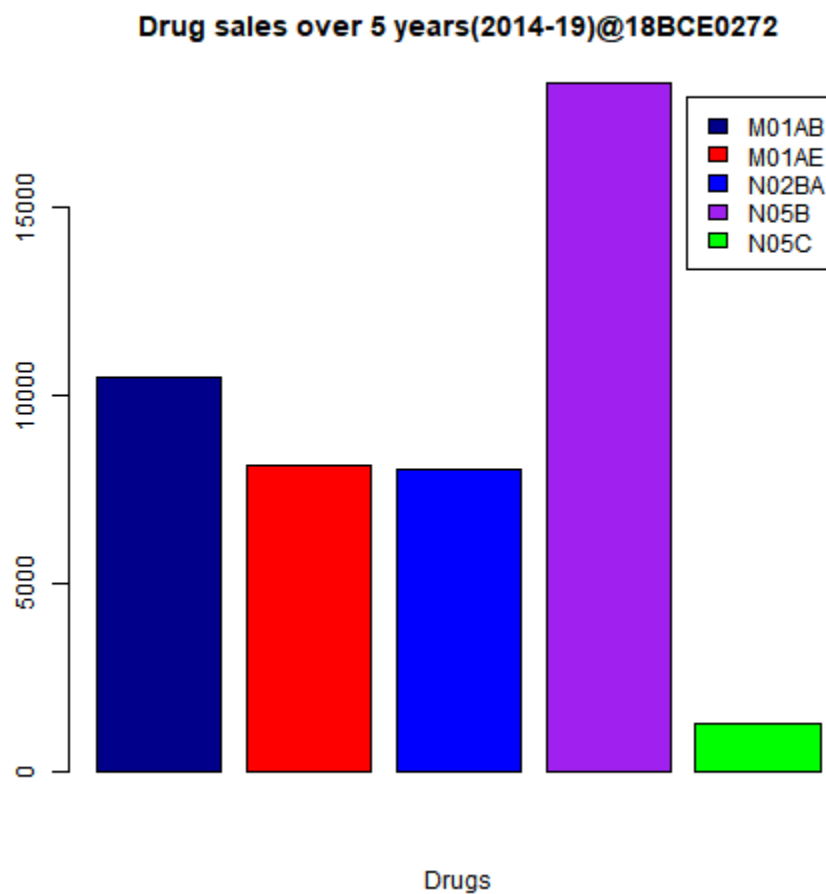


The bar graph

Let us first compare total sales of all drug groups.

CODE

```
>  
> png(file="Total-Sales.jpg")  
> counts <- c(sum(pd$M01AB), sum(pd$M01AE),sum(pd$N02BA), sum(pd$N05B), sum(pd$N05C)  
+ )  
> barplot(counts, main="Drug sales over 5 years(2014-19)@18BCE0272",  
+ xlab="Drugs", col=c("darkblue","red","blue","purple","green"),legend =  
c("M01AB","M01AE","N02BA","N05B","N05C"))  
> dev.off()
```



#segmented bar graph

```
png(file="salesasbar.png")
```

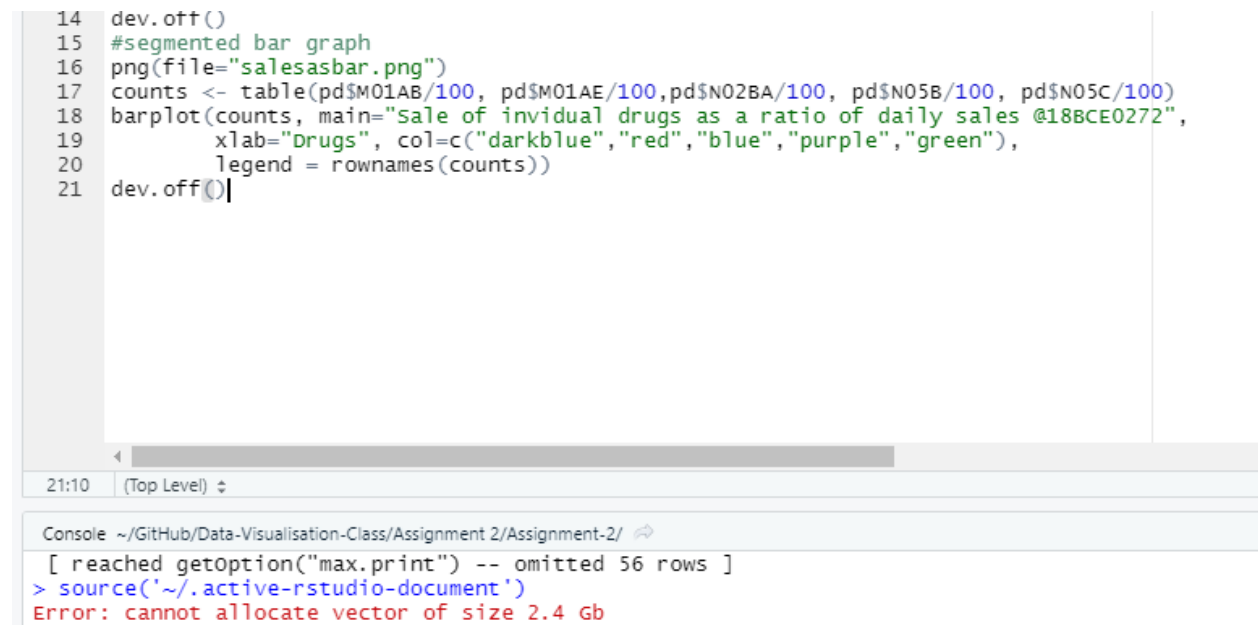
```
counts <- table(pd$M01AB/100, pd$M01AE/100, pd$N02BA/100, pd$N05B/100, pd$N05C/100)
```

```
barplot(counts, main="Sale of individual drugs as a ratio of monthly sales @18BCE0272",
```

```
        xlab="Drugs", col=c("darkblue","red","blue","purple","green"),
```

```
        legend = rownames(counts))
```

```
dev.off()
```

The image shows a screenshot of the RStudio interface. The top pane displays R code for creating a segmented bar plot. The code includes comments and function calls like 'png()', 'table()', 'barplot()', and 'dev.off()'. The bottom pane shows the console output, which includes a message about reaching the maximum print limit and a red error message: 'Error: cannot allocate vector of size 2.4 Gb'.

```
14 dev.off()
15 #segmented bar graph
16 png(file="salesasbar.png")
17 counts <- table(pd$M01AB/100, pd$M01AE/100, pd$N02BA/100, pd$N05B/100, pd$N05C/100)
18 barplot(counts, main="Sale of individual drugs as a ratio of daily sales @18BCE0272",
19         xlab="Drugs", col=c("darkblue","red","blue","purple","green"),
20         legend = rownames(counts))
21 dev.off()
```

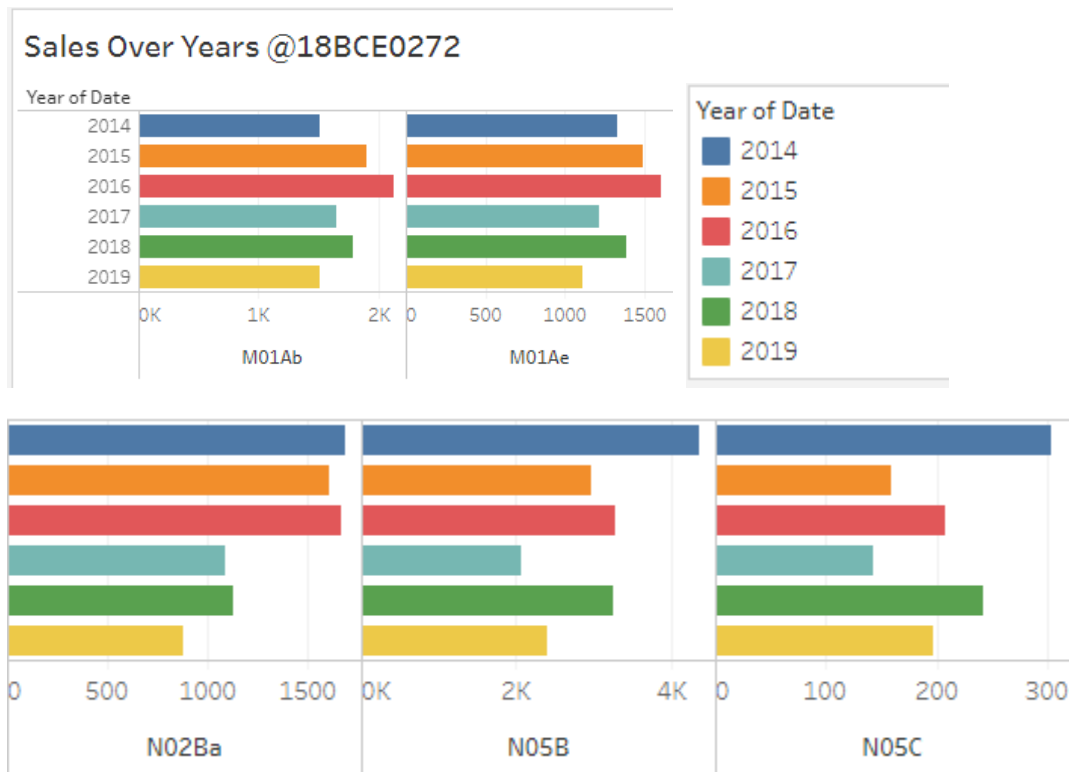
21:10 (Top Level) ↕

Console ~/GitHub/Data-Visualisation-Class/Assignment 2/Assignment-2/ ↗

```
[ reached getOption("max.print") -- omitted 56 rows ]
> source('~/.active-rstudio-document')
Error: cannot allocate vector of size 2.4 Gb
```

So, apparently, my system is not strong enough to create a vector that large.

Instead of a segmented bar graph to represent a lot of independent bars, we can compare overall yearly sales of each drug, abstracting the category and reducing data visible to the client and yet providing with a rough picture of the sales.



RESULT

- Clearly enough, anxiolytics were the highest selling drugs and their sales have consistently remained high, and that might point out to the fact there was
 - A large scale therapeutic need for people suffering from anxiety
 - However, this might not mean that the number of anxiety cases themselves were very high.
- The user can successfully compare patterns of sale of various drugs in a given period of time and the pattern of the sale of each drug over the years.
- I was unable to create a segmented bar graph due to system inefficiencies.

CONCLUSION

Thus, the chosen dataset was studied, the data explored and visualized and the results were proposed and demonstrated.

Submitted on 19th May 2021

By

Nitin Ranjan

18BCE0272