Student ID: 20302634

MEDA2002 Data Visualization and Interpretation Journal week 3

Saen-Anan Bunyasiwa[a,1]

[a]Data Science Major, Faculty of Science and Engineering, Curtin University; [1]Completed Introduction to Data Visualization with ggplot2 via DataCamp®

This manuscript was compiled on September 12, 2020

This week's report was compiled on LaTeX using Overleaf platform. The contents for this week discussion will be about Rstudio and the modules inside which are useful for data cleaning, data wrangling and data visualization.

RStudio | ggplot | Visualization | tidyverse |

RStudio is one of the most used tool for Data Scientist and is also powerful if use wisely. Therefore, mastering the tool might add value to the work. In this case, using it to visualize the data is relatively easy if the user already know how it handle the data.

In this journal, will cover the basics of importing the data and the basics of data visualization using `ggplot2` module and the composition of elements in the chart.(1)

Basics of `ggplot2`

(2)Using `ggplot2` is easy compared to python or other programming language due to its structure of creating the chart. One of the most interesting things about ggplot is that the structure is built accordingly as the grammar of chart building. Layers of chart were built. Starting off with the base chart with aesthetics.

```
library(ggplot2)
ggplot(data = mpg, aes(...))
```

Then, we have geom layer which satisfies the types of chart we want to build. For example, I want to build a histogram. I will just start off by the base layer then the histogram layer.

```
ggplot(data = mpg, aes(...) +
geom_histogram(binwidth = 1)
```

There are almost 20 types of geom that I can add to the plot. However, there is one distinct geom that make me interested. It is `geom_jitter` which add a little bit of random variation each point of data and prevent overplotting.

Aesthetics mapping. Mapping the aesthetics is quite challenging since it need a lot of adjustments for an unseen data. The common types of aesthetics are alpha(opacity), color, size, x, y, linetype, etc. Each geom have their own aesthetics configuration. For example, if I want a scatterplot with half opaque and size a little bit bigger, I can do this

```
ggplot(data = mpg, aes(x = cty, y = hwy) +
geom_point(aes(alpha = 0.5, size = 1.5))
```

Regression line, hline, vline. ggplot has its own geom for the line which is `geom_line` or a line associated with a model `geom_smooth`. For instance, I want to add the linear regression model to the plot

```
ggplot(data = mpg, aes(x = cty, y = hwy) +
geom_point(aes(alpha = 0.5, size = 1.5)) +
geom_smooth(method = 'lm')
```

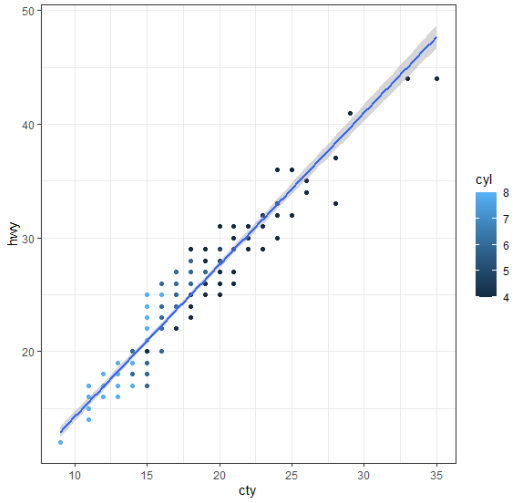

Fig. 1. Figure ran as code in **section Regression line, hline, vline**

Stats. Sometimes, the input data needed some transformation before we can visualiza it or sometimes we just transform on our needs. For example, plotting the frequency, we can use `stat_bin`.

Theme. Making theme from scratch is difficult. From my experience, as I have finished the course to making the theme myself. However, ggplot offer basic theme which are, for example, `theme_bw` or `theme_minimal`. However, to make an intuitive dashboard or chart as seen on TV or in the news report require making the thme from scratch, if you want to make it in RStudio.

Theoretical Study

After I know the basics of making the chart in RStudio and ggplot. Now, I can learn in the lecture.

Visualizing flow. The flow of making visualization, according to the book (3), is straightforward. The raw data needs to be tidy up then the cycle of —transform, model, and visualize will loop until satisfied. Then, we can setup the communication between the visualizer and the audience.

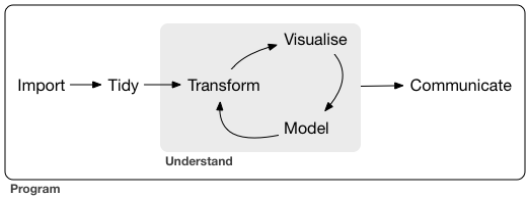

Fig. 2. The workflow of data visualization (3)

Tidyverse. Using `tidyverse` to wrangle the data is widely use amongst R users due to its compactness in cleaning the data. Useful tools inside tidyverse are —`dplyr`, `readr`, etc. Also, functions like `mutate`, `pivot_longer` is so easy to understand and manipulate as we wish. Then, we can visualize it.

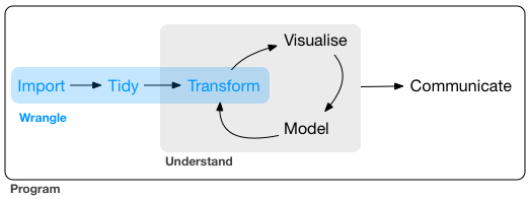

Fig. 3. The wrangle part in the workflow (3)

Questions raised

After one week of study, I have some questions that I need to find answer for myself and my colleagues.

Question 1. How many preset themes are there? Which do I prefer?

Answer: There are several preset themes.

Table 1. Preset themes offered by ggthemes (4)

Themes	features
1. theme_tufte	Minimal Ink theme
2. theme_wsj	Wall Street Jounal design
3. theme_economist	The economist design
4. theme_classic	White background, no grid
Et cetera	...

Preferably, it depends on the data and the audience. My default setting would be tufte.

Question 2. How to make geo-map beautifully like Tableau in week2?

Answer: I have tried to make geo-map. Unfortunately, I can make my best only using `theme_map` with US muder data. I think, in the future, I can improve myself by making my own theme from scratch and make it beautiful like in Tableau. [`Fig.4`]

Question 3. What is a geom?

Answer: The word "geom" stands for geometrical objects. It represent the data points whether it is discrete or continuous or even a 2D distribution. Also, differentiating the geom makes an advantage of layers in the plot. It makes it easier to interpret an complicated plot code.

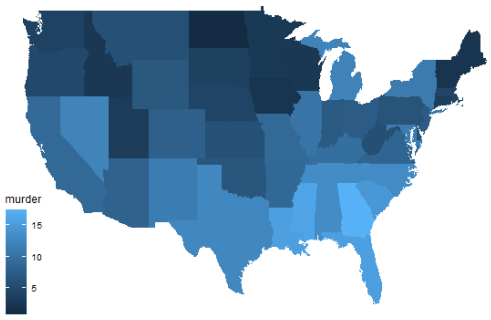

Fig. 4. Murder rate in USA geographically.

Question 4. Is there any limitations in `theme()`?

Answer: According to (5), setting the elements of the chart need to be set outside the theme customization code, therefore it might be confusing on the user in a complex graph with many elements and layers within graphs. Moreover, minor customizations of the element in theme is quite hard like LaTeX. Without proper baseline knowledge, will result in a poor theme.

Question 5. Is Tableau better than `ggplot`?

Answer: This question has always comes across my mind when I am working on visualizing the data. So I gathered some answer. The results were that it depend on economic status. If the company I am working has a license to Tableau, using Tableau might be a better idea due to its easy and intuitive design. On the other hand, RStudio and ggplot are an open source platform. Another consideration is the project/work I am working on. If I need an embed into notebook with all the explanation and presentation in RMarkdown. Then, using RStudio would be better. But for a fast and simple task, using Tableau is easier. Here is a figure that I think can help me and the reader of this journal decide.

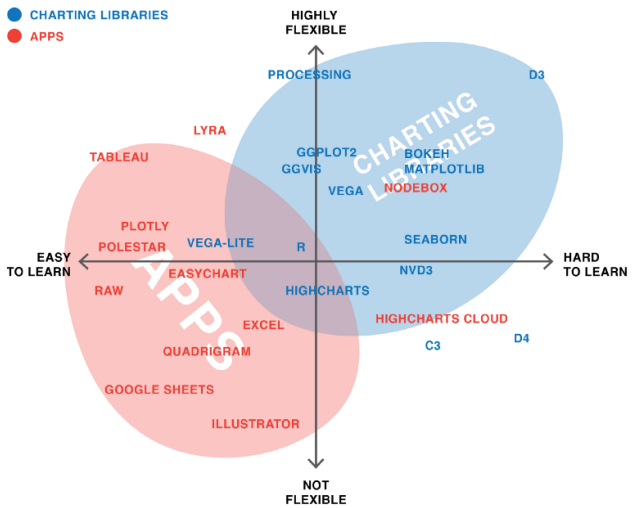

Fig. 5. Graph of flexibility and difficulty to learn

References

1. DataCamp introduction to data visualization with ggplot2 (https://learn.datacamp.com/courses/introduction-to-data-visualization-with-ggplot2) (2020) Accessed: 1 September 2020.
2. RStudio ggplot cheatsheet (https://rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf) (2015) Accessed: 10 September 2020.
3. H Wickham, G Grolemund, *R for data science: import, tidy, transform, visualize and model data*. (OReilly), (2017).
4. GITHUB gatherings of all preset themes (https://yutannihilation.github.io/allYourFigureAreBelongToUs/ggthemes/) (2020) Accessed: 12 September 2020.
5. RPubs customizing themes in gg (https://rpubs.com/mclaire19/ggplot2-custom-themes) (2020) Accessed: 12 September 2020.