Examining the Visualization Practices of Data Scientists on Kaggle

Md Naimul Hoque
Stony Brook University

Darius Coelho
Stony Brook University

Klaus Mueller
Stony Brook University

Abstract

Kaggle, founded in 2010, has become the leading platform for data scientists to collaboratively explore and build data-based models, participate in competitions, and communicate with each other mostly through interactive notebooks and forum discussions. The growing community has a large userbase with 2.9M current users across 194 countries. This large community regularly produces a large number of solutions (Kernels) for problems and datasets posted on the website. These solutions tend to be of high quality as they are often commissioned by companies like Google, LinkedIn through competitions where users can win prizes for building the best data-based models. Visualization, an integral part of data science, is employed in a large portion of these kernels to either explore data or present results. In this project, we examine the content of these kernels to understand the visualization practices among Kaggle data scientists. Our work reveals insights about the libraries used, the most popular visual representations and the types of color palettes used by these data scientists.

Index Terms: Human-centered computing—Visualization—Empirical studies in visualization

1 Introduction

Visualization is an integral part of data science as it is used extensively in different phases of designing a data-based model and communicating its outcomes [4]. Real-world data today is often large in terms of number of data entries and dimensionality, it can also contain inconsistencies or noise. The use of appropriate visualizations allows data scientists to quickly explore the data and examine outliers or inconsistencies from the data. This process of finding patterns and trends in the data is known as Exploratory Data Analysis (EDA). EDA can be used to extract unique insights from data, take appropriate business decisions, learn relationships between variables, all of which contribute towards the design of a good data-based model.

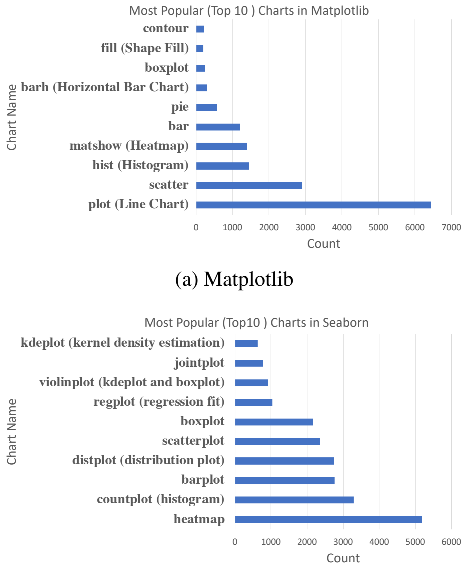
Companies such as Tableau and Spotfire provide visual EDA tools based on well established visualization guidelines, however a large number of data scientists today use python and jupyter notebooks to explore their data. Their visualization practices may not necessarily follow the well established visualization guidelines. Researchers have conducted studies on some aspects of these practices, however these studies consisted of a small group of users [1]. In this work, we examine the visualization practices of data scientists through the thousands of jupyter notebooks they post on the Kaggle[1] platform.

Organizations and individuals regularly post datasets and problem statements on Kaggle, some of these post are competitions that offer various rewards. The Kaggle community of 2.9 million users with varying degree of expertise in data science attempt to solve these problems and post their solutions to the website. These solutions are publicly accessible and receive upvotes from other users on the platform. We collect these solutions and extract information from them that can inform us about the visualizations they use. A similar approach was used to study the trends of people collaborating Github [2]. To the best of our knowledge, there is no study that analyzes the behaviour of users on Kaggle and that has motivated us to take on this project. Our study has the potential of bringing out the key visualization factors used by a large variety of users. In the remainder of this paper we discuss the methodology used to collect and analyze the Kaggle posts and the results of this analysis.

2 Dataset

To study the practices of data scientists on Kaggle, we collected 57,038 posts or kernels from kaggle. These kernels included various types of files including R, python, Jupyter Notebooks, Rmd, etc. but the majority of the kernels (50,193) were Jupyter Notebooks written in python (.ipynb) format thus we chose to only study the jupyter notebooks. Of this set of kernels, some were auto-generated by Kaggle bots, these kernels contain boilerplate code and are generated whenever a new competition is launched on Kaggle. We remove these kernels by filtering out those created by the user *kerneler* (Kaggles bot). Additionally, we filtered out kernels that had the keyword “tutorial” in their title as they are intended to teach other user about the functionality of libraries rather than explore a dataset. After this initial filtering we had 40,139 kernels remaining.

Our goal is to understand the visualization practices followed in the kernels, however not all kernels visualize the data. To select kernels with visualizations, we first extracted information about the packages used by the kernels and we observed that *matplotlib* (24,200 kernels) and *seaborn* (14,890 kernels) are the most popular visualization packages among the Kaggle users. It was interesting to see that less than 2000 kernels use *plotly*, a visualization packages known for its sophisticated, interactive, and complex plots which are not available in *matplotlib* and *seaborn*. Thus we chose to only



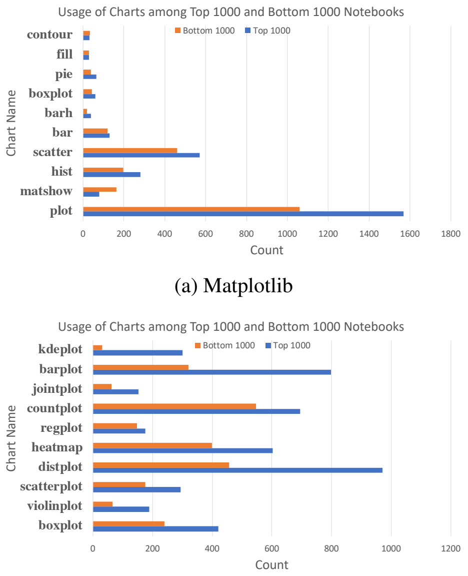
(a) Matplotlib

(b) Seaborn

Figure 1: The number of occurrences of the top 10 plots in (a) Matplotlib and (b) Seaborn

[1] kaggle.com

IEEE VIS 2019, 20-25 October, Vancouver, BC, Canada.



(a) Matplotlib

(b) Seaborn

Figure 2: The number of occurrences of the top 10 plots in (a) Matplotlib and (b) Seaborn in the top and bottom 1000 rated kernels.

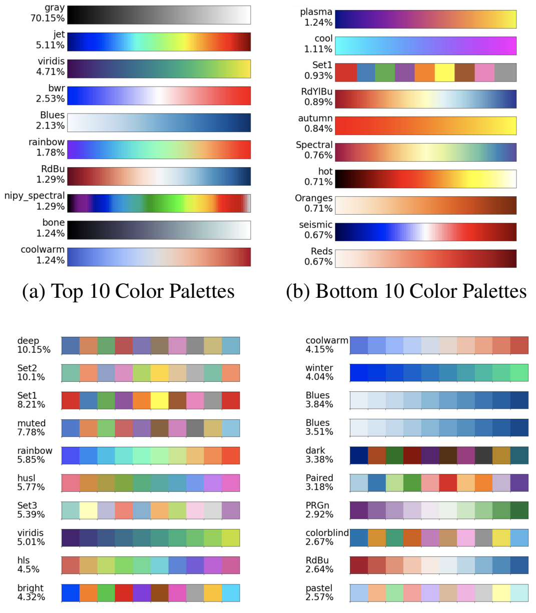
analyze kernels that use the two popular libraries - *matplotlib* and *seaborn*, this left us with 32,020 kernels. Finally, we used *meta-kaggle*[2], a dataset maintained by Kaggle, in our analysis to acquire meta-data related to the kernels, it tells us the number of views and upvotes each kernel recieved.

3 Insights

We only analyzed kernels that used the *matplotlib* and *seaborn* libraries, thus we extract information from the kernels by examining the function calls from these libraries. First we examined the type of charts employed by the kaggle users, this is shown in figure 1. We found that for *matplotlib* the most popular plots (functions) are the *plot* (line chart), *scatter plot*, *histogram*, etc. while the most used charts with *seaborn* are *Heatmap*, *countplot*, *barplot*, *distplot* etc. From figure 1 we can see that *plot* (line chart), which is commonly used to bring out trends from data, is very popular among the users. We also see that plots such as *scatter plot*, *heatmap*, and *regression fit*, which are commonly used to visualize bi-variate relationship, are frequently used. A fair amount of kernels use *histogram*, *distplot*, and *boxplot* to show the summary of data. Notably, pie charts, which are often criticized for not showing visual differences between entities[3], are not that popular among the users in Kaggle.

From the *meta-kaggle* dataset, we know the number of views, comments, and up-votes each notebook got. We use the number of up-votes as a measure of goodness of kernels and try to find out how top rated kernels (those with higher up-votes) differ from the kernels that did not get good ratings in terms of their use of visualization. We sort the kernels by their up-votes and take the first 1000 as top rated kernels and the bottom 1000 as the lowest rated kernels. Figure 2 shows the comparison between top rated and lowest rated kernels for each visualization packages. Each bar represents the count of the particular chart usage in that category. From Figure 2 we can see that although the distribution of visualization usage between top rated and lowest rated kernels are identical, top rated kernels use twice as many visualizations than the lowest rated kernels.

Color is a visual variable that is used often to convey categorical (hue) or continuous (intensity) values and has been studied in detail in visualization research [3]. We investigated Kaggle kernels to find patterns and trends followed by the kaggle users when they use color and color palettes. We found that *red* is the most common color that users employ when they are not using the default color. Also, *blue*, *green*, *black* are fairly common. Thus, we observe that users tend to use the basic RGB colors more often and are reluctant towards using custom colors. In addition to the individual colors, we studied the color palettes used by Kaggle users. Seaborn users employ fairly bright color palettes with these palettes being more suitable for categorical data. On the other hand matplotlib users employ continuous palettes with the *grey* and *jet* (rainbow color map) being the most frequent.



(a) Top 10 Color Palettes

(b) Bottom 10 Color Palettes

(c) Top 10 Color Palettes

(d) Bottom 10 Color Palettes

Figure 3: The most and least used color palettes in Matplotlib (a) & (b) and Seaborn (c) & (d)

4 Conclusion and Future Work

In this preliminary study, we have found some interesting trends showing how data scientists use visualization in their data-based storytelling (notebooks). Particularly interesting was the relatively high selection of the rainbow colormap which has been shown to have significant perceptual shortcomings. In the future, we plan to investigate this kaggle dataset in more detail. We plan on investigating the type of data shown in each chart, the colors used in these charts, and how they compare to the standards set by the visualization researchers. Additionally, we plan on investigating the order in which these charts are presented, pairing this with the rating would inform us of the best approaches for explaining an analysis to a wide audience. We believe that this work would help guide aspiring data scientists to use visualization in there solutions more effectively.

References

[1] A. Batch and N. Elmqvist. The interactive visualization gap in initial exploratory data analysis. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):278–287, Jan 2018.

[2] A. Lima, L. Rossi, and M. Musolesi. Coding together at scale: Github as a collaborative social network. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[3] S. Silva, B. S. Santos, and J. Madeira. Using color in visualization: A survey. *Computers & Graphics*, 35(2):320–333, 2011.

[4] S. S. Skiena. *The Data Science Design Manual*. Springer, 2017.

[2] https://www.kaggle.com/kaggle/meta-kaggle

[3] https://www.businessinsider.com/pie-charts-are-the-worst-2013-6