Why Visualize Data When Coding? Preliminary Categories for Coding in Jupyter Notebooks

Tasha Settewong, Natanon Ritta,
Chaiyong Ragkhitwetsagul, Thanwadee Sunetnanta
Faculty of ICT, Mahidol University
Nakhon Pathom, Thailand

Raula Gaikovina Kula, Kenichi Matsumoto
Nara Institute of Science and Technology (NAIST)
Nara, Japan

***Abstract*—Data visualization becomes a crucial component in data analytics, especially data exploration, understanding, and analysis. Effective data visualization impacts decision-making and aids in discovering and understanding relationships. It leads to benefits in data-intensive software development tasks e.g., feature engineering in machine learning-based software projects. However, it is unknown how visualizations are used in competitive programming. The idea of this paper is to report early results on what visualizations are prevalent in competitive programming. Grandmasters are the highest level reached in competitions (novice, expert, master, and grandmaster). Analyzing the visualizations of 7 high-rank competitors (i.e., Grandmaster) in Kaggle, we identify and present a catalog of visualizations used to both tell a story from the data, as well as explain the process and pipelines involved to explain their coding solutions. Our taxonomy includes nine types from over 821 visualizations in 68 instances of Jupyter notebooks. Furthermore, most visualizations are for data analysis for distribution (DA Distribution), and frequency (DA Frequency) are most used. We envision that this catalog can be useful to better understand different situations in which to employ these visualizations.**

***Index Terms*—data visualization, machine learning competition, data analysis**

I. Introduction

Data visualization has significantly contributed to the success of data-intensive projects e.g., data analytics, data science, and machine learning-based software projects [1]. Data visualization has previously been used in the past to present information from the data. However, with the current trend of Big data, data visualization becomes more than a graph showing. It involves various practices in data analytics, especially machine learning-based software ranging from data understanding, data exploring, data pre-processing, feature engineering, machine learning model training, and model interpretation. Thus, the presenting of big data tends to require efficient data visualization techniques to obtain insight and knowledge in order to make use of the data together with its implicit patterns and relationships to enhance the easiness of information comprehension [2]–[4].

Existing research explains how data visualization is commonly used as a proxy in decision-making, that must be relies on the interpretation from the use of data visualization [1], [5]–[7]. Specifically, data visualization is adopted to leverage their comprehensive understanding of vast amounts of data as well as identify and discover patterns and trends in such a way that is efficient and attractive for human cognition, such as visualizing information of the recent pandemic situation (i.e., Covid-19) showing infected cases across several countries [7], and assisting in the model building process to the capture of relationships of data in the computational model [1].

Although there has been much work in understanding data visualization, the extent to how programmers use visualization in competitive programming is unknown. In the context of competitive programming, Kaggle[1] provides a platform for machine learning-based competitions, which requires competitors to perform intensive data analysis practices. Competitors are required to develop data analysis notebooks in the form of the Jupyter notebooks, which include the use of data visualization to present their data understanding, analysis, and model training. Popular notebooks, however, are not only notebooks that train a highly accurate model but also notebooks that contain nice data storytelling using effective data visualization techniques.

In this paper, we present early results of a manual analysis of visualizations used in competitive programming. We empirically study and categorize those visualizations to examine visualization types that are used by notebook grandmasters. Our study performs on the KGTorrent [8] and aims to answer the following research questions:

- *RQ1: What type of visualizations are used by the grandmaster users?* First, we would like to explore visualization types that are mostly used by the grandmaster competitors.
- *RQ2: Do different competitors use different types of visualizations?* Second, we want to study whether the grandmaster competitors use the same visualization types in a competition.

We particularly focus on the notebooks from competitors who have the notebook grandmaster tier, which is the highest ranking in the competitions. The competitors who achieve this grandmaster tier must present their expertise in data analysis and communicate their analysis approach to voters (e.g., using visualization). Our study analyzes a total of 68 notebooks containing 821 visualizations overall. The results from this preliminary study can help us further develop an approach to understanding the factors that lead to the use of

[1]https://www.kaggle.com/

To be published in the 29th Asia-Pacific Software Engineering Conference (APSEC 2022), Virtual Event.

Fig. 1: Summary of the research approach to perform our study

effective and impactful visualizations. Our main contribution is nine visualization categories (i.e., Distribution, Frequency, Map, Percentage, Statistics, Train Model, Test Model, Image Model, and Feature Importance) that are used in competitive programming. Our visualizations are available at https://github.com/NAIST-SE/VizJupyterNotebooks.

II. Background

In this section, we explain the background of our study, which is the related work on data visualizations and the Kaggle platform.

a) Data visualization: Data visualization is the method of representing data and information in graphical form with visual elements using technologies in a more understandable way. It is significantly essential to transforming the data into a form that is more accessible, understandable, and interactable. Many studies used data visualization to leverage their comprehensive understanding of data [9], [10]. For example, Abela [11] categorized the chart into four major types based on the purpose user would like to show e.g., comparison, relationship, distribution, and composition. Shakir Khan [5] adopts data visualization to gain its advantage for exploring the countries' dataset to provide a holistic and interpretive view of the world. Crapo et al. [1] adopt data visualization to support model building by assisting the modeler in discovering and understanding relationships within the data which can lead directly to the capture of relationships in the computational model if the visualization and modeling tools are well integrated. In addition, Islam et al. [6] make use of data visualization to assist an engineer in addressing the presence of large clusters of mutual dependence that have been considered an issue preventing understanding, testing, maintenance, and reverse engineering. Furthermore, data visualization is recently used in order to help in understanding different aspects of COVID-19, such as displaying information on cases and death totals for different countries [7]. In particular, Dong et al. [12] used the dataset to investigate how data scientists perform data cleansing.

b) The Kaggle Platform: Kaggle is an online community platform for people who are interested in data science and machine learning. It allows users to collaborate and compete with others in data science competitions to solve challenges. The Kaggle notebooks are a computational environment that enables reproducible and collaborative analysis, which can be differentiated into two types; scripts, and notebooks. We focus on the notebook which is a Jupyter Notebook consisting of a sequence of cells, where each cell can be marked as either textual description or code execution. Mostly, data visualization in Kaggle is used in data exploration to be the guild line for building a model, feature engineering, and model interpretation. Each user's (i.e., competitor) performances are ranked based on their contribution to four platforms which are (1) Competitions, (2) Notebooks, (3) Datasets, and (4) Discussion. The Kaggle progression system[2] is used to determine competitor ranking which consists of 5 performance tiers; (1) Novice, (2) Contributor, (3) Expert, (4) Master, and (5) Grandmaster. A tier promotion can be achieved according to the quality and quantity of work that the competitors contribute and voting. Recently, Wang et al. [13] considered highly-voted notebooks on Kaggle as a proxy for well-documented notebooks and categorized the documentation in the markdown cells which cover a broad range of topics and purposes into nine types.

III. Data Collection and Processing

Figure 1 shows our research methodology. It consists of the setting up of KGTorrent, constructing the dataset of our study (data collection), extracting visualizations, classifying visualizations, and performing data analysis.

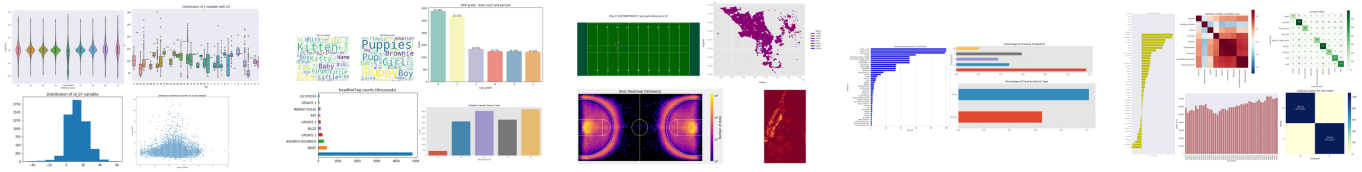
We use the KGTorrent as the source of our data. KGTorrent [8] is a dataset, provided by Quaranta et al., containing Python Jupyter notebooks with rich metadata retrieved from the Kaggle platform. It also contains a database for recording those metadata referring to the notebooks and the activities of Kaggle users on the platform, e.g., users' actions, competitions, and code kernels. KGTorrent contains the information of over five million Kaggle users and over two thousand competitions.

a) Target Competitors: In this study, we focus on the `Notebook Grandmaster` tier achieving 15 notebook gold medals. Each medal is awarded to popular notebooks which are measured by the number of upvotes on those notebooks which require 50 votes. We selected to focus on notebook grandmasters because they mostly have data visualization in their notebooks for data exploration and can be ranked based on the upvotes of the notebooks.

We downloaded the KGTorrent package from the Zenodo repository[3]. The dataset requires us to use MySQL to import the whole metadata of notebooks and competitors.

[2]https://www.kaggle.com/progression

[3]https://zenodo.org/record/4468523#.YxTN5qxBxb8



(a) Distribution (DA) (b) Frequency (DA) (c) Map (DA) (d) Percentage (DA) (e) Statistics (DA)

(f) Train Model (ML) (g) Test Model (ML) (h) Image Model (ML) (i) Feature Importance (ML) (j) Others

Fig. 2: Sample visualizations of each classification type

A. Data Processing

To filter data for our study, we have two criteria for selecting competitors in this experiment. First, based on ranking, we select competitors whose tier is notebook grandmaster. Second, the competitors are required to participate in 9 or 10 competitions. Consequently, for each competitor, the top 10 upvoted-notebook are retrieved. As Table I shows, our dataset has in total of 7 notebook grandmasters, 68 notebooks, and 821 visualizations. Overall, we found that, on average, the notebook grandmasters in our dataset have 118 visualizations. The smallest and largest number of visualizations are 34 and 214 visualization, respectively.

TABLE I: The number of visualizations of each competitor

Competitor	#notebooks	#visualizations
1	10	132
2	10	133
3	10	138
4	10	214
5	9	66
6	9	34
7	10	104
total	68	821

IV. Categories of Visualizations

In this section, we discuss our approach and results of the classification.

A. Systematic Classification Approach

We used manual coding for developing our categories. Similar to thematic analysis, we assigned three of the authors to first manually code 20 visualizations, and discuss between each other regarding each visualization. In the end, our grouping is based on three components of purpose and then features of the visualization. For the purpose, we decided upon two purposes e.g., data analysis (DA), and machine learning model interpretation (ML). Furthermore, we then decided on the categories based on three features of the visualization, which are:

- the graph type (e.g., bar chart, histogram, and box plot)
- the title and keywords contained in the source code used for constructing a visualization (e.g., percentage, correlation)
- the data label showing in the graph (e.g., y-axis and x-axis)

As part of the coding, the first three authors first coded 50 samples together. Once the first author was confident, they went on to manually classify the rest of the visualizations.

TABLE II: Visualization classification type

Type	Purpose	Description
Frequency	DA	Visualization that count the frequency of data
Distribution	DA	Showing the data distribution e.g., violin plot, box plot, histogram
Statistics	DA	Visualizing statistics data e.g., correlation, confusion matrix
Percentage	DA	Visualization that shows percentage of data
Map	DA	Displaying data on a geographical map to demonstrate the spatial relationships in data
Train Model	ML	Visualization that is generated in training model process
Test Model	ML	Visualization produced during the test of a model
Image Model	ML	Showing the image data
Feature Importance	ML	Visualizing the feature importance of model
Others	-	Other visualization that cannot fall into any classification type

B. Visualization Categories from Jupyter Notebook

Figure 2 shows an example visualizations of each classification type used by the notebook grandmaster. Furthermore, Table II shows a brief description of types and its description. Note that in our preliminary study, the visualization types

TABLE III: Most Frequent Association sets of Visualizations used by each competitor

Left_Hand_Side	Right_Hand_Side	Support	Confidence	Lift
(‘competitor1’,)	(’Distribution (DA)’, ’Train Model (ML)’, ’Frequency (DA)’)	0.074	0.5	6.8
(‘competitor2’,)	(’Statistics (DA)’, ’Distribution (DA)’, ’Feature Importance (ML)’, ’Frequency (DA)’)	0.044	0.3	6.8
(‘competitor3’,)	(’Feature Importance (ML)’, ’Distribution (DA)’)	0.088	0.6	3.709
(‘competitor4’,)	(’Image Model (ML)’, ’Frequency (DA)’)	0.074	0.5	3.4
(‘competitor5’,)	(’Test Model (ML)’,)	0.088	0.667	2.519
(‘competitor6’,)	(’Test Model (ML)’,)	0.088	0.667	2.519
(‘competitor7’,)	(’Map (DA)’,)	0.044	0.3	1.569

used in our study have been defined based on the basic characteristics of the visualization. We now discuss each of the categories.

*a) **Distribution DA**:* Figure 2a shows an example of the visualization. We classify visualization in this type by the purpose of doing data analysis, and the graph tends to show the distribution of data. The graph type in this category is a histogram, box plot, and violin plot. For instance, the visualization use histogram with a title with the keyword ’Distribution of id_07 variable’.

*b) **Frequency DA**:* In this category, as figure 2b shows, the visualization display the frequency of the data for data analysis with the graph type count plot, bar plot, and word cloud, such as the count plot with title ’Adoption speed classes rates’ and the data labels showing in the graph are count for the y-axis and adoption speed for the x-axis.

*c) **Map DA**:* As figure 2c shows, the map category contains the visualizations that display the data on a geographical map. For instance, the visualization is a heat map that shows the number of shots of women’s basketball dedicated to each location on the basketball field.

*d) **Percentage DA**:* This percentage category demonstrates in figure 2d, consisting of the visualization that shows the percentage of data e.g., the bar graph with the title ’Percentage of fraud by device type’ and the x-axis shows the percentage along with device type for the y-axis.

*e) **Statistics DA**:* Figure 2e express an example of the statistics category. The visualization in this category is dedicated to displaying the statistics data, such as the visualization with the keyword correlation and confusion matrix in the title.

*f) **Train Model ML**:* The train model category consists of the visualizations that have a purpose for model interpretation in the training process e.g., the line graph shows the training accuracy.

*g) **Test Model ML**:* The visualization that intends to show model interpretation or the result in the testing process is dedicated to this category, for instance, the graph with the keyword submission or prediction.

*h) **Image Model ML**:* The image model category comprises the visualization that shows the image that is being used in the model e.g., the visualization shows the result of image recognition from the model with its labeling as the title of visualization shown in figure 2h.

*i) **Feature Importance ML**:* The visualization in this category shows the importance of input features to the model. The visualization that is classified in this category has the keyword as feature importance in its title, as shown in figure 2i.

*j) **Others**:* The visualization that cannot classify into the types stated above is contained in this Others category due to the lack of information to interpret the visualization.

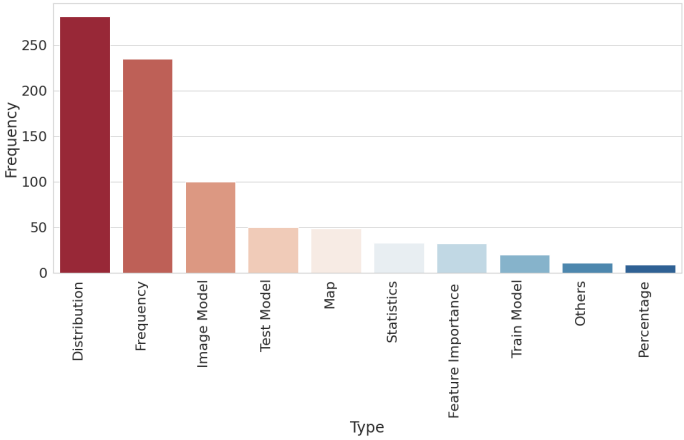


Fig. 3: The frequency of each visualization type

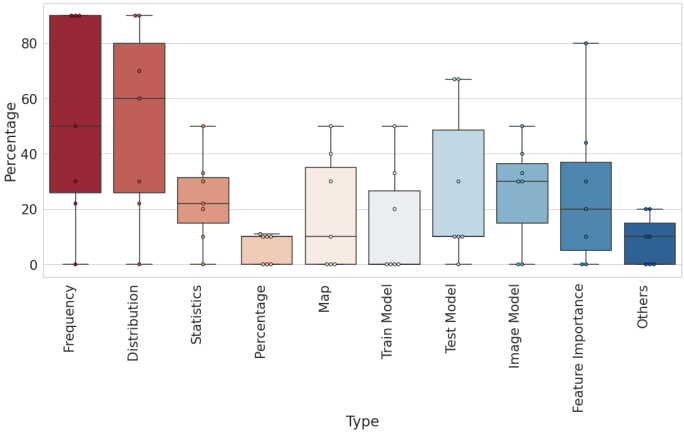


Fig. 4: The percentage of each visualization type

V. Empirical Evaluation

Using our classification, we can answer our RQs.

A. *RQ1: What type of visualizations are used for users?*

Table III is a result of applying association rule mining. Each competitor tends to have their own style of using a set of

visualization in common. We can see that most competitors use different sets of visualizations, and that there is not common sets between them. For example, competitor 1 frequently uses the distribution, training model and frequency visualizations, while competitor 7 only uses Map DA in their visualizations.

Furthermore, we also count the frequency of each visualization type to examine which type is used the most. As a result shown in Figure 3, visualization types that are being used the most are DA-Distribution(282), DA-Frequency(235), and Image Model(100), respectively.

Summary: We identify nine types of visualization, from which the data analysis distribution (DA Distribution), data analysis for the frequency (DA Frequency) are being the most used.

B. RQ2: Do different competitors use different types of visualizations?

According to Table III, each user prefer different type of visualizations. However, some common set of visualization is used by competitors. For instance, competitors 1, 2, and 3 all use the distribution (DA) visualization between themselves. Figure 4 shows that the median of the percentage of type DA-Distribution and DA-Frequency are above other types of visualization classification, which is 60 and 50 percent, respectively.

Summary: No, it seems that, similar to RQ1, most competitors use visualizations to understand the frequency and distribution of the data.

VI. Limitations

Internal validity refers to factors that could have affected our findings. We encountered an issue with KGTorrent. There is some data inconsistency between the database and the dataset. Some notebook file names in the database cannot be found in the dataset. Additionally, we rely on manual visualization classification. We, however, acknowledge that the agreement rate should be determined. However, the dataset is small in this preliminary study.

Threats to external validity concern the generalizability of our findings. This is a small, preliminary study, conducted on only 68 competition notebooks and contains 821 visualizations from Kaggle. Our future work thus requires us to expand the scope of our study by increasing the sample size.

VII. Conclusion and Future Challenges

This study investigated the extent to which set is commonly used in Kaggle competition. The study has been performed on data 68 Jupyter notebooks from KGTorrent [8]. We identified and classified visualizations manually, which has 821 visualizations in total. To achieve the objective of our study, we came up with these research questions (1) What type of visualizations are used for users? (2) Do different competitors use different types of visualizations? Our experimental results reveal that each user prefers to use a different type of visualization. Nevertheless, the common set of visualizations that are being used is DA-Distribution and DA-Frequency.

This paper is the first step toward our goal that we aim to develop a visualization recommendation tool based on competition challenges on the Kaggle platform. Additionally, we need to perform an in-depth analysis of how visualization is used by grandmasters differs from novices and how competitors use discussion features in each competition. Furthermore, our findings from this preliminary study support our motivation in developing a tool for visualization recommendations for data analytics and machine learning-based programming in order to support data analytics practitioners at all levels of expertise.

Acknowledgment

This work has been supported by JSPS KAKENHI Grant Numbers JP8H04094, JP20K19774, and JP20H05706.

References

[1] A. W. Crapo, L. B. Waisel, W. A. Wallace, and T. R. Willemain, "Visualization and the process of modeling: A cognitive-theoretic view," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 2000.

[2] C. Ware, *Information visualization: perception for design*. Morgan Kaufmann, 2019.

[3] F. Post, G. Nielson, and G.-P. Bonneau, "Data visualization: The state of the art," 01 2003.

[4] L. Wilkinson, "Playfair's commerical and political atlas and statistical breviary," *Psychometrika*, 2007.

[5] S. Khan, "Data visualization to explore the countries dataset for pattern creation." *International Journal of Online & Biomedical Engineering*, 2021.

[6] S. S. Islam, J. Krinke, and D. Binkley, "Dependence cluster visualization," in *Proceedings of the 5th International Symposium on Software Visualization*. Association for Computing Machinery, 2010.

[7] J. L. D. Comba, "Data visualization for the understanding of covid-19," *Computing in Science & Engineering*, 2020.

[8] L. Quaranta, F. Calefato, and F. Lanubile, "Kgtorrent: A dataset of python jupyter notebooks from kaggle," in *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*, 2021.

[9] I. Khan and K. Malik, "Modal parameter identification of bridge based on large scale data sets," in *Proceedings of the International Conference on High Performance Compilation, Computing and Communications*. Association for Computing Machinery, 2017.

[10] T. J. Jankun-Kelly and K.-L. Ma, "A spreadsheet interface for visualization exploration," in *Proceedings of the Conference on Visualization '00*. IEEE Computer Society Press, 2000.

[11] A. Abela, "Chart suggestions - a thought starter," 2009, accessed: 2022-10-20. [Online]. Available: https://extremepresentation.typepad.com/files/choosing-a-good-chart-09.pdf

[12] H. Dong, S. Zhou, J. L. Guo, and C. Kästner, "Splitting, renaming, removing: A study of common cleaning activities in jupyter notebooks," in *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, 2021.

[13] A. Y. Wang, D. Wang, J. Drozdal, X. Liu, S. Park, S. Oney, and C. Brooks, "What makes a well-documented notebook? a case study of data scientists' documentation practices in kaggle," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2021.