beautiVis: An Annotated Visualization Dataset from Reddit's r/dataisbeautiful

Kylie Lin*
Georgia Institute of Technology

Sean Sheng-tse Ru*
Georgia Institute of Technology

Minsuk Chang†
Georgia Institute of Technology

Cindy Xiong Bearfield‡
Georgia Institute of Technology

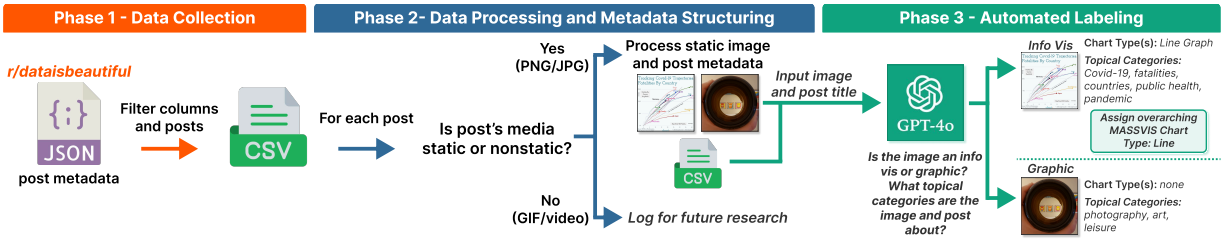

Figure 1: The three-phase process for curating the beautiVis dataset from the *r/dataisbeautiful* subreddit. We collected monthly JSONL archives and filtered out deleted or removed posts. Next, we processed the remaining content by extracting static images, cleaning post metadata (e.g., titles and unique identifiers) and organizing data into structured CSV files. Finally, we implemented an automated labeling pipeline that distinguishes visualizations from non-visual content, automatically annotates chart types and topical categories based on both image content and post titles, and maps identified visualizations to the MASSVIS taxonomy.

Abstract

Visualization researchers are increasingly leveraging artificial intelligence (AI) and machine learning (ML) models to design visualization tools that support complex analytic workflows. Effectively applying these models in visualization contexts involves fine-tuning to account for visualization-specific features, which in turn requires large, annotated visualization datasets. We present an annotated dataset, beautiVis, sourced from Reddit's *r/dataisbeautiful* subreddit, consisting of 52,836 information visualizations from 2012 to 2025. Each visualization is annotated with structured metadata including chart types (e.g., bar chart, line chart, scatter plot), topical categories (e.g., climate, politics, United States), and social engagement metrics (e.g., upvotes, downvotes). This dataset captures real-world visualization design practices across diverse domains over a 10+ year span, presenting researchers with opportunities to train models to learn the grammar of axes, gridlines, and glyph encodings present in visualizations and analyze temporal trends in data storytelling and community interaction on Reddit. The full dataset is publicly available at: https://osf.io/pkj7s and https://huggingface.co/datasets/beautiVis/beautiVis.

Keywords: Data visualization, Dataset, Annotations, Reddit, LLM

1 Introduction

As artificial intelligence (AI) continues to transform scientific and analytic workflows , researchers are increasingly turning to AI and machine learning (ML) models to design visualization tools that support complex data interpretation, reasoning, and decision-making (14). Yet, a fundamental challenge remains: most of these models are trained on natural images, which differ from data visualizations in both structure and the perceptual mechanisms they engage (11; 18).

Consider saliency prediction, a common task in both AI computer vision and visualization research (6; 32). While existing saliency models excel at predicting attention in photos and natural scenes, they fall short when applied to data visualizations (25; 34). This gap stems from a mismatch between the visual characteristics of natural images and the symbolic nature of data visualizations, likely influenced by a lack of quality training data. Therefore, bridging this gap requires not only visualization-specific models but also *large-scale, diverse visualization datasets* that reflect how people design, share, and interpret visualizations in the real world.

Such datasets remain scarce. Existing collections such as MASSVIS and ChartQA are limited in size, complexity, or scope (4; 24). As a result, progress in perceptual modeling, chart classification, and visualization recommendation systems has been constrained by a lack of representative training data (21). We posit that diverse and complex annotated datasets are essential for accelerating both theoretical and applied advances in the visualization community. Recent research underscores the promise of community-sourced visualization datasets, especially those that capture real-world usage and social feedback (1; 17). Social media platforms like Reddit offer a valuable opportunity: they contain diverse visualizations created by both experts and novices, span a wide range of domains, and include built-in signals of quality and engagement through votes, comments, and community moderation.

In this paper, we introduce a large-scale, open-source[1] dataset, beautiVis, derived from the *r/dataisbeautiful* subreddit (9). The dataset comprises 52,836 information visualizations from 2012 to 2025. Each visualization is associated with a Reddit post or thread and annotated with structured metadata, including chart types (e.g., bar chart, scatter plot), topical categories (e.g., climate, politics), and social interaction metrics (e.g., upvotes, downvotes). This dataset spans more than a decade of content and reflects evolving design practices and communicative strategies in the public sphere.

*Kylie Lin and Sean Sheng-tse Ru contributed equally to this work as co-first authors. email: klin368@gatech.edu, sru3@gatech.edu
†email: minsuk@gatech.edu
‡email: cxiong@gatech.edu

[1]beautiVis on Open Science Framework: https://osf.io/pkj7s/?view_only=098d28fdcb14454eb86833418491e60c
beautiVis on Hugging Face: https://huggingface.co/datasets/beautiVis/beautiVis

To construct the dataset, we collected monthly JSONL archives from *r/dataisbeautiful*, verified active posts, and extracted associated images. Through structured prompting of OpenAI's GPT-4o model (27), we generated high-level annotations such as chart type and topic. The resulting corpus provides a rich foundation for future research in AI-powered visualization analysis and generation, including saliency modeling, visualization recommendation, and automated alt-text generation. By bridging the gap between community-driven design and structured, machine-readable metadata, this dataset also offers a scalable foundation for developing human-centered visualization tools.

Contributions: We contribute (1) A 52,836-item visualization dataset with detailed metadata on visualization types and topics designed to support future visualization and machine learning research, and (2) A scalable, automated processing pipeline combining web scraping and advanced classification techniques.

2 Related Work

In this section, we review existing visualization datasets and automated visualization annotation methods.

2.1 Visualization Datasets and Metadata

Several visualization datasets have enabled large-scale analysis of visual design, reader perception, and automated techniques for visualization and analytics.

The MASSVIS dataset (3; 4) consists of 5,000+ static visualizations curated from government reports, news media, infographic sites, and scientific publications. It has supported research on visualization memorability and includes crowd-sourced memorability scores and metadata such as chart type and source domain. VisImages (10) is a corpus of 200,000 visualizations extracted from web-scraped images and alt-text. The dataset focuses on image-level visual features and supports research such as automatic chart classification. Vis30K (7) provides 30,000+ annotated visualizations primarily extracted from academic publications, offering metadata such as chart type, scientific domain, and associated captions. While valuable for studying scientific visualization conventions and context-specific design, they exhibit limited stylistic diversity compared to those found in public-facing media, such as news or social platforms.

Similarly, VizNet (13) and VizML (14) collect and analyze tens of thousands of visualizations generated on platforms like Plotly and Tableau Public. These datasets are structured and programmatically accessible, enabling large-scale machine learning analyses of design patterns, recommendation systems, and predictive modeling of user preferences. Yet, they primarily represent visualizations produced in structured analytics settings, which may not capture the diversity of visual storytelling styles used for broad public consumption.

Web-scraped datasets such as VizByWiki (20) and Beagle (2) have demonstrated the feasibility of large-scale collection and classification, with Beagle focusing on the automatic extraction of visualizations from webpages and VizByWiki drawing on Wikipedia content. Both focus heavily on vector-based formats (e.g., SVG) and often exclude raster-based charts, which are common in public-facing visualizations such as those shared on news or social media.

Our presented beautiVis dataset complements these existing resources by offering a diverse collection of real-world visualizations shared by the general public. We supplement existing datasets with richer metadata including static raster visualizations accompanied by rich metadata, such as chart type, data topics, and engagement metrics, across over a decade plus of content. These visualizations span a broad spectrum of purposes, including informative, persuasive, and exploratory designs intended for general audiences. To label beautiVis, we employed an automatic labeling pipeline, informed by a body of related literature which we turn to now.

2.2 Automated Visualization Extraction and Annotation

Recent advancements in automated visualization extraction have significantly improved the scalability and reliability of datasets. Early efforts like ReVision (31) relied on traditional computer vision methods, later evolving toward deep-learning-based frameworks such as ChartOCR (23), enhancing accuracy in text and numerical extraction. Other interactive systems such as ChartSense (15), which allowed extraction of chart data with user involvement, and approaches aimed at reverse-engineering visual encodings (29) provided valuable insights into chart interpretation.

However, challenges remain for stylized or complex visualizations common in informal web contexts. Specialized benchmarks such as ChartQA (24), DVQA (16), and PlotQA (26) were developed specifically to generate structured question-answer pairs aimed at evaluating visual comprehension and numeric reasoning capabilities of automated systems. ChartQA focuses on real-world chart images, requiring logical and visual reasoning. DVQA emphasizes understanding of synthetic bar charts through visual perception and numeric tasks, and PlotQA targets numeric reasoning using synthetic scientific plots. While these benchmarks support advancements in visualization comprehension tasks, they underscore the necessity of datasets with richer annotations and realistic complexity to address broader automation challenges effectively.

More recent approaches such as Chart2Vec (8) demonstrate the potential for embedding visualizations to support automated retrieval and recommendation, reinforcing the value of structured annotations in supporting advanced analytical tasks. These developments collectively emphasize the ongoing need for comprehensive datasets featuring diverse visual formats, detailed metadata, and realistic contexts to enhance automation and research capabilities.

3 Methodology

In this section, we describe our dataset curation process.

3.1 Data Collection

To collect posts and associated static images from the *r/dataisbeautiful* subreddit, we focused on the period from February 2012, when the community was founded, to January 2025, the most recent full month of data available at the time of collection. Because Reddit's API limits the number of accessible posts per category (e.g., New, Hot) to only 1000 posts per call, we found it impractical for large-scale scraping. Thus, we chose instead to make use of the `Arctic-shift` Reddit Documentation Download Tool (28), which allows bulk downloads of post data for a given subreddit or user.

We downloaded and parsed all post data from *r/dataisbeautiful*, skipping posts that had been deleted or removed by moderators for being off-topic, low-quality, or inappropriate. As a result, the dataset benefits from strong content curation through both automated filters and human moderators. To support temporal analyses and avoid API rate limits, we processed the data into monthly batches. We aggregated the processed data while retaining original post details, timestamps, and unique identifiers for traceability.

3.2 Data Processing and Metadata Structuring

Once the monthly files were downloaded, each record was parsed into a flat table and written to a CSV file. For every post, the original title was reformatted to remove punctuation and excessive whitespace. In a separate field, we assigned a unique identifier to each file based on its Reddit post upload date; the same string is later used as the filename for the corresponding image, guaranteeing one-to-one mapping between metadata and image.

We then utilized BeautifulSoup (30), a Python web scraping library, to only extract static images (PNG, JPG, JPEG, or SVG) with our file number naming scheme. Links that led to animated GIFs, videos, and other non-static media were logged for future analysis. Images successfully downloaded were then labeled.

3.3 Labeling Images

We created an automated labeling pipeline with OpenAI API (27), using the GPT-4o model.

Prompt Engineering and Generating Labels: To annotate the images at a large scale, we used the GPT-4o model through the OpenAI API to label the scraped images. Our initial approach simply asked the model to determine whether each image was a visualization or not, but this resulted in all images being incorrectly labeled as non-visualizations. To address this, we refined our prompt strategy to perform three key functions simultaneously:

1. filter visualizations from other content (e.g., photographs).
2. classifying the type of any detected visualizations.
3. tagging images with relevant topical categories derived from both the visual content and accompanying post titles.

The final prompt is included in the supplemental material and began with a precise definition of data visualizations as graphical representations of structured data, followed by 20 concrete examples of valid chart types (e.g., bar charts, scatter plots, choropleth maps) and 5 clear non-examples (e.g., photographs, illustrations, plain text). Crucially, we designed the prompt to *not* constrain responses to our provided list of chart types, allowing the model to identify chart types that we had not anticipated, while the examples served as anchoring references for common visualization forms. For over 50k images, we document the average cost for the labeling process using the OpenAI API to fall roughly between $75-100, with roughly 1500 input tokens per image and roughly 25-50 output tokens per input.

The model was instructed to output a list of chart types and a list of topical categories. If the list of chart types had only 'none', the image was labeled as a non-visualization. This compact format minimized token usage and simplified downstream parsing. Individual requests that fail indicate the image is corrupt, and the image is skipped. Any batches that run abnormally long or fail outright were rerun.

Normalization and Post Processing: Despite these precautions, the model's outputs required post-processing to correct typographical variations/errors (e.g., "cholopleth" → "choropleth"), standardize unexpected variations of some chart type names ("line chart" → "line graph"), and clean unwanted characters. We also used a cleaning script to ensure that the output matched to our intended list format.

Chart Types Consolidation: The initial labeling process identified over 600 chart types. Following established practices (12), we adopted a modified version of the MASSVIS categorization system to group chart types into 12 overarching categories (see Table 2) with an additional *Other* category for unclassifiable cases. This categorization serves three key purposes: it enables meaningful statistical comparisons across visualization types by reducing sparsity, aligns with perceptual and cognitive distinctions in how readers interpret visual encodings, and facilitates comparison with prior visualization research that uses similar taxonomies. To implement this, we provided GPT-4o with both our complete list of unique chart types and detailed documentation of the MASSVIS categories, prompting it to generate a mapping dictionary between the two.

3.4 Validation of Model-Generated Annotations

To assess GPT-generated annotations, we conducted a manual validation of a stratified sample of 300 images from the dataset. The sample included 100 images that our model had classified as *non-visualizations* and 200 classified as *visualizations*. Two trained research assistants independently coded the 300 images via the same annotation protocol used in the automated process. Disagreements between coders were discussed and resolved through consensus, resulting in 100% agreement in the final human annotations. These human-verified labels were treated as the ground truth for subsequent comparison with labeling from the automated pipeline.

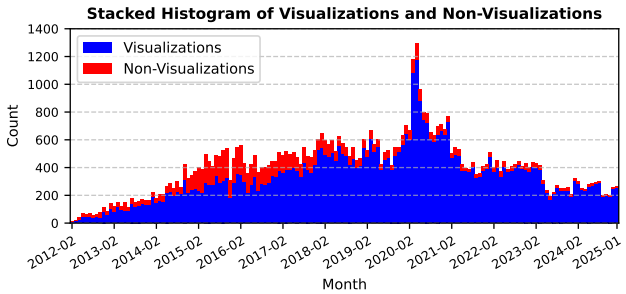

Figure 2: The monthly distribution of posts submitted to r/dataisbeautiful from February 2012 to January 2025. **Blue** represents **Visualizations**: posts with a valid data visualization image. **Red** represents **Non Visualizations**: posts where the image is either absent or does not depict a data visualization.

Among the 100 images that our model labeled as non-visualizations, human coders identified seven as visualizations. Among the 200 images that our model labeled as visualizations, human coders identified one as not a visualization and twelve that were misclassified in terms of chart type. Based on these comparisons, the inter-rater reliability between our model and the human coders was **Cohen's** $\kappa = 0.939$, indicating *very high agreement* (19). This result demonstrates that our model was highly accurate in differentiating visualization content from non-visual materials and classifying the chart type. The discrepancies in chart-type classification involved visualization types such as *cartogram*, *choropleth map*, *line chart*, *box plot*, *area chart*, *waffle chart*, *icon array*, and several instances involving *multiple visualizations* (e.g., combinations of bubble chart, bar chart, and map; or network, map, and bar chart).

These results suggest that while GPT-4o achieves high reliability in identifying visualizations, minor discrepancies persist in differentiating among fine-grained visualization types, particularly in chart composite, maps, and less common formats. Overall, the validation indicates that our model's automated annotations are robust for large-scale visualization identification, with modest room for improvement in subtype labeling.

4 Results

We present our annotated dataset, detailing its structure, chart types, high-level categories, and key statistics from our analysis.

4.1 Dataset Composition

Overall, GPT-4o classified 52,836 images as visualizations and 10,044 images as non-visualizations, while 48 images were identified as corrupted. We organized the data into separate CSV files for visualizations and non-visualizations, along with corresponding image folders to support analysis and exploration. The visualization CSV files contain columns that reflect the three processing phases. First, the original metadata (prefixed with *json*) fields extracted from JSON files, such as post titles, authors, URLs, and engagement metrics. Second, the post-processed columns (prefixed with *pp*) contain cleaned titles and image filenames. Lastly, the GPT-generated label columns (prefixed with *gpt*) provide the identified chart types, their standardized cleaned versions, high-level topical categories, and the corresponding MASSVIS taxonomy groupings. The non-visualization CSV files include the original metadata and two GPT-processed fields: chart type (labeled none) and high-level category (if applicable). Prefix labels were omitted.

4.2 Visualization Types and High-Level Categories

Our initial classification identified over 600 chart types, some of which were duplicate or very similar visualizations with different names. This was due to inconsistent terminology, excessive details (e.g., "3d bar chart" vs plain "bar chart"), and overly specific hybrid labels (e.g., "animated political spectrum chart"). Table 1

Table 1: Most Frequent Chart Types and Categories in the Dataset

Top Chart Types		Topical Categories	
Chart Type	Count	Category	Count
bar chart	12,253	trends	13,100
line graph	10,660	geography	6,047
choropleth map	9,497	demographics	5,555
scatter plot	3,803	united states	4,120
area chart	2,198	health	4,116
heatmap	1,889	politics	3,814
sankey diagram	1,835	sports	3,754
bubble chart	1,773	data analysis	2,949
network diagram	1,389	finance	2,918
pie chart	1,308	covid-19	2,904
donut chart	689	history	2,616
histogram	584	economics	2,611
map	296	statistics	2,478
box plot	290	technology	2,410
slope chart	254	economy	2,308
radar chart	225	social media	2,251
violin plot	217	environment	1,940
waffle chart	203	countries	1,771
chord diagram	190	elections	1,758
dot plot	182	population	1,726

Table 2: Categorizing Our Dataset with Labels from MASSVIS

Type	Count	Type	Count
Bar	12,499	Grid & Matrix	1,947
Line	11,011	Trees & Networks	1,588
Maps	10,329	Distribution	1,125
Point	5,760	Table	38
Circle	2,257	Text	6
Area	2,216	Other	932
Diagrams	2,008		

summarizes the most frequent chart types as well as the top high-level categories. Table 1 also presents the frequency distribution of processed chart types, with bar charts being the most popular. Similar to the chart types, the high-level categories had variations, like "United States" vs. "USA", which artificially inflates the number of unique labels. The most frequent high-level thematic categories were *trends* (13,100), *geography* (6,047), and *demographics* (5,555). The files listing counts for all chart types and high-level categories are included in the supplemental material.

Table 2 provides the MASSVIS chart type categories and their counts. The most frequent MASSVIS chart types were *Bar* (12,499), *Line* (11,011), and *Maps* (10,329), whereas *Table* (38) and *Text* (6) significantly lack presence. The *Other* category had 982 images, which were specialized or unconventional types that didn't fit into the standard taxonomy (e.g., 'mouse movement tracking visualization').

5 Discussion & Future Work

Next, we discuss limitations in the dataset curation and opportunities for future work building off of the BeautiVis dataset.

5.1 Scalability, Label Quality, and Manual Verification

Our pipeline combines web scraping with LLMs to enable scalable, automated dataset generation. While effective, future work can further refine this approach to enhance label quality and coverage. Some scraped images were blurry, incomplete, or were missing context. GPT-4o occasionally produced false positives and false negatives when classifying visualizations.

In addition, the level of detail in LLM-generated labels presents both advantages and challenges. By giving GPT-4o the flexibility to identify a broad set of visualization types other than what was listed in our prompt, we encountered cases of inconsistent terminology, excessive details, and overly specific hybrid labels. This highlights the challenge of achieving a balance between flexibility and consistency and emphasizes the need to improve prompt engineering for the automated labeling process to significantly improve data quality (36). Due to the limitations of LLMs, manual verifications are important. Human reviewers can effectively address errors, contextual ambiguity, and other inconsistencies generated by LLMs. Future work can focus on developing a systemic approach for integrating manual validation efficiently along with automated labels to ensure data quality. Future work could integrate chain-of-thought prompting (35) to encourage intermediate reasoning steps, potentially reducing errors and improving the consistency of automated labeling.

5.2 Platform-Specific Limitations

Relying exclusively on the r/dataisbeautiful subreddit introduces several platform-specific limitations. First, the subreddit has strict moderation rules that shape the content and types of visualizations collected. For example, submitted posts should contain the subreddit's qualifying visualizations, provide a direct link to original sources, use clear and unbiased titles, comply with anti-plagiarism guidelines, and follow weekly posting restrictions (e.g., political topics are allowed only on Thursdays and personal data only on Mondays). This results in most posts being removed by moderators, with many submitters having to resubmit their posts multiple times.

In addition, the subreddit only allows users to interact in English; this includes posts, text within visualizations, and comments. Further, topic-wise (as shown in Table 2) 'united states' is #4 on the list of top topical categories. This suggests that the discussions are largely US-centric, similar to a broader pattern in HCI research where roughly 73% of CHI paper findings come from Western samples, with 45.8% from the US alone (22). This US-centric focus may bias the themes, data sources, and design methods of posts in the subreddit toward American contexts, limiting the generalization of the dataset to a global population. Future work should incorporate visualizations and community interactions from more geographically and culturally diverse platforms and domains to offer a more representative understanding of global visualization practices.

5.3 Opportunities for Future Insight

The beautiVis dataset presents several opportunities for future research, which we outline below.

Social Computing Insights: A key strength of the dataset is its social interaction data, including upvotes, downvotes, and the number of comments associated with visualizations. These metrics allow for analysis of social dynamics and temporal trends. For instance, looking at the visualization metadata, there are a total of 76,047,285 upvotes and only 3,908 downvotes, providing the insight that the *r/dataisbeautiful* subreddit is largely a place where users receive positive feedback on posted visualizations. Additionally, it is possible to examine how the amount of upvotes associated with high-level categories changes over time. As an example, Figure 3 shows five categories that consistently appeared in the top five most upvoted categories ("politics," "geography," "demographics," "united states," and "social media") and how they change over time.

Political visualizations were prominent in the early years of the subreddit, peaking around major election cycles. Relative popularity declined after 2017 but gradually recovered in subsequent years, likely reflecting renewed engagement with data-driven political discourse. In parallel, visualizations about *social media* and *demographics* gained traction, indicating a growing public interest in how digital platforms and population patterns shape social life.

Across all years, the subreddit exhibits a clear U.S.-centric orientation. Visualizations categorized under *United States* consistently received high levels of engagement, underscoring the predominance of American data sources and topics within the community. This

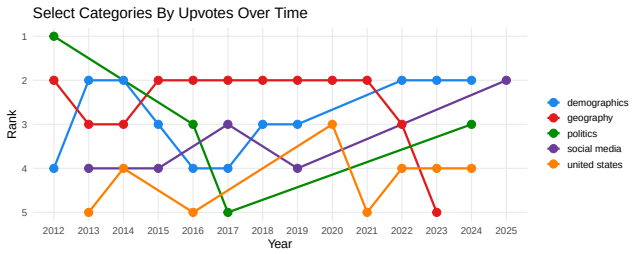

Figure 3: Select high-level categories and their rank in terms of total upvotes from 2012 to 2025. Years where top terms do not appear indicate that their ranking dropped below the top-5 for that year.

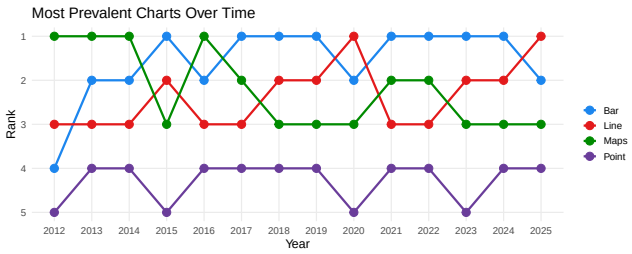

Figure 4: The top four most popular chart types and their relative popularity from 2012 to 2025.

trend may reflect both the platform's English-speaking user base and the broader cultural prominence of U.S.-related data narratives in online discourse. Overall, the temporal patterns in topic popularity reveal how collective interests within *r/dataisbeautiful* evolve in response to sociopolitical contexts, technological developments, and the community's demographic composition.

Future enhancements beautiVis could capture additional social interaction metrics, such as the content of posted comments to allow for a deeper understanding of how visualizations posted are received by the community. These social metrics can guide our understanding of online user engagement patterns, the evolution of visualization techniques, and public reception of these techniques.

Temporal Insights: Another strength of the beautiVis dataset is its temporal metadata. Since posts are organized by date, future studies can analyze observable shifts that tie to historical events. For example, in March and April of 2020, there was a notable spike in posts in the subreddit focused on the COVID-19 pandemic. Examining the most popular chart types over time (as in Figure 4) reveals that bar charts, line plots, maps, and point charts were consistently the most popular charts from 2012 to 2025; however, while maps were the first most popular in the subreddit's earlier years, they dropped to second and third starting in 2015. This suggests an overall preference for common chart types, accompanied by a gradual shift from geographically grounded narratives using maps toward data-driven comparisons and temporal analyses using bars and lines. By analyzing the high-level topical categories column and comments associated with the posts, future research efforts can gain contextual insights on various historical events.

Original Content Insights: Finally, future work could also leverage the dataset's substantial volume of Original Content (OC) posts. These posts, marked with '[OC]' in their titles, refer to ones containing visualizations created and shared directly by their original authors. The subreddit requires OC creators to explain their visualization methods, the data sources used, and the specific tools or software used to create the visualization in their post's comment section. This transparency encourages high-quality submissions, promotes ethical sharing behaviors, provides valuable context, and inspires interactions like discussion and feedback.

6 Conclusion

Our automated approach and dataset set a foundation for future datasets and methods by introducing a reproducible and improvable pipeline. The ability to generate labels and other annotations for information visualizations can be used for a wide variety of tools and studies, such as building real-world visualization search engines with vector embeddings (8). The visualizations in the dataset can also be used to generate large-scale question-answer pairs with the structured labels and metadata (16; 24; 26). The dataset itself can support benchmarking of LLMs for chart understanding (5), reasoning capabilities (35), and alignment with human interpretations (33).

Acknowledgments

This work was supported in part by the National Science Foundation awards III-2453462, IIS-2534219, and IIS-2237585. The authors would like to thank Ibaad Sayeed for helping with the work.

References

[1] B. Bach, E. Freeman, A. Abdul-Rahman, C. Turkay, S. Khan, Y. Fan, and M. Chen. Dashboard design patterns. *IEEE transactions on visualization and computer graphics*, 29(1):342–352, 2022. 1

[2] M. Beagle, E. Hoque, and M. Kay. Beagle: Automated extraction and interpretation of visualizations from the web. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–12. ACM, 2018. doi: 10.1145/3173574.3173745 2

[3] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva. Beyond memorability: Visualization recognition and recall. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):519–528, Jan 2016. doi: 10.1109/TVCG.2015.2467732 2

[4] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2306–2315, Dec 2013. doi: 10.1109/TVCG.2013.234 1, 2

[5] V. S. Bursztyn, J. Hoffswell, E. Koh, and S. Guo. Representing charts as text for language models: An in-depth study of question answering for bar charts. In *2024 IEEE Visualization and Visual Analytics (VIS)*, pp. 266–270, 2024. doi: 10.1109/VIS55277.2024.00061 5

[6] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand. Where should saliency models look next? In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pp. 809–824. Springer, 2016. 1

[7] J. Chen, Y. Wang, A. Wang, Y. Wang, H. Zhang, K. Xu, X. Wang, Y. Wu, and H. Qu. Vis30k: Figures and tables from ieee visualization publications. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1039–1049, Feb 2021. doi: 10.1109/TVCG.2020.3028945 2

[8] Q. Chen, Y. Chen, R. Zou, W. Shuai, Y. Guo, J. Wang, and N. Cao. Chart2vec: A universal embedding of context-aware visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1234–1244, Jan 2023. doi: 10.1109/TVCG.2022.3209410 2, 5

[9] r. community. r/dataisbeautiful subreddit. https://www.reddit.com/r/dataisbeautiful/, 2025. 1

[10] D. Deng, Y. Wu, X. Shu, J. Wu, S. Fu, W. Cui, and Y. Wu. Visimages: A fine-grained expert-annotated visualization dataset. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):464–474, Jan 2022. doi: 10.1109/TVCG.2021.3114776 2

[11] S. L. Franconeri, L. M. Padilla, P. Shah, J. M. Zacks, and J. Hullman. The science of visual data communication: What works. *Psychological Science in the public interest*, 22(3):110–161, 2021. 1

[12] J. Heer, M. Bostock, and V. Ogievetsky. A tour through the visualization zoo. *Communications of the ACM*, 53(6):59–67, Jun 2010. doi: 10.1145/1743546.1743567 3

[13] K. Hu, M. Bakker, S. Li, T. Kraska, and C. Hidalgo. Viznet: Towards a large-scale visualization learning and benchmarking repository. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–12. ACM, 2019. doi: 10.1145/3290605.3300892 2

[14] K. Hu, D. Orghian, and C. Hidalgo. Vizml: A machine learning approach to visualization recommendation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–12. ACM, 2019. doi: 10.1145/3290605.3300358 1, 2

[15] D. Jung, W. Kim, H. Song, J. Jeong, L. Lee, B. Kim, and M. Seo. Chartsense: Interactive data extraction from chart images. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–10. ACM, 2017. doi: 10.1145/3025453.3025957 2

[16] K. Kafle, B. Price, S. Cohen, and C. Kanan. Dvqa: Understanding data visualizations via question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5648–5656. IEEE, 2018. doi: 10.1109/CVPR.2018.00592 2, 5

[17] T. Kauer, A. Srinivasan, and E. Bertini. Public life of data: Investigating reactions to visualizations on reddit. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–14. ACM, 2021. doi: 10.1145/3411764.3445608 1

[18] C. Knittel, J. Awuah, S. Franconeri, and C. X. Bearfield. Gridlines mitigate sine illusion in line charts. In *2024 IEEE Visualization and Visual Analytics (VIS)*, pp. 246–250. IEEE, 2024. 1

[19] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pp. 159–174, 1977. 3

[20] A. Lin, J. Wu, H. Zhang, Y. Wu, and W. Cui. Vizbywiki: Mining data visualizations from the web. In *Proceedings of the World Wide Web Conference*, pp. 873–882. ACM, 2018. doi: 10.1145/3178876.3186136 2

[21] K. Lin, S. S.-T. Ru, D. N. Rapp, H. Guan, and C. X. Bearfield. What makes a visualization visually complex? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, 7 pages. ACM, New York, NY, USA, 2025. doi: 10.1145/3706599.3719983 1

[22] S. Linxen, C. Sturm, F. Brühlmann, V. Cassau, K. Opwis, and K. Reinecke. How weird is chi? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pp. 1–14. ACM, Yokohama, Japan, 2021. doi: 10.1145/3411764.3445488 4

[23] J. Luo, X. Li, J. Wang, and Y. Yang. Chartocr: Data extraction from chart images via a deep hybrid framework. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 1917–1926. IEEE, 2021. doi: 10.1109/WACV48630.2021.00196 2

[24] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *ACL Findings*, pp. 1–12, 2022. 1, 2, 5

[25] L. E. Matzen, M. J. Haass, K. M. Divis, Z. Wang, and A. T. Wilson. Data visualization saliency model: A tool for evaluating abstract data visualizations. *IEEE transactions on visualization and computer graphics*, 24(1):563–573, 2017. 1

[26] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar. Plotqa: Reasoning over scientific plots. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 1527–1536. IEEE, 2020. doi: 10.1109/WACV45572.2020.9093439 2, 5

[27] OpenAI. Chatgpt (gpt-4o) api. `https://chatgpt.com/?model=gpt-4o`, 2025. 2, 3

[28] Photon Reddit Archive. Arctic-shift reddit data download tool. `https://arctic-shift.photon-reddit.com/`, 2025. 2

[29] Z. Poco and J. Heer. Reverse-engineering visualizations: Recovering visual encodings from chart images. *Computer Graphics Forum*, 36(3):353–363, Jun 2017. doi: 10.1111/cgf.13193 2

[30] L. Richardson. Beautiful soup: Html parsing library for python. `https://www.crummy.com/software/BeautifulSoup/`, 2023. Version 4.12.2. 2

[31] M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer. Revision: Automated classification, analysis and redesign of chart images. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pp. 393–402. ACM, 2011. doi: 10.1145/2047196.2047247 2

[32] S. Shin, S. Chung, S. Hong, and N. Elmqvist. A scanner deeply: Predicting gaze heatmaps on visualizations using crowdsourced eye movement data. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):396–406, 2022. 1

[33] H. W. Wang, J. Hoffswell, S. M. T. Thane, V. S. Bursztyn, and C. X. Bearfield. How aligned are human chart takeaways and llm predictions? a case study on bar charts with varying layouts, 2024. 5

[34] Y. Wang, W. Wang, A. Abdelhafez, M. Elfares, Z. Hu, M. Bâce, and A. Bulling. Salchartqa: Question-driven saliency on information visualisations. In *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 1–14, 2024. 1

[35] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, Jan 2022. doi: 10.48550/arXiv.2201.11903 4, 5

[36] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, Feb 2023. doi: 10.48550/arXiv.2302.11382 4