Chart Question Answering from Real-World Analytical Narratives

Anonymous ACL submission

Abstract

We present a new dataset for chart question answering (CQA) constructed from visualization notebooks. The dataset features real-world, multi-view charts paired with natural language questions grounded in analytical narratives. Unlike prior benchmarks, our data reflects ecologically valid reasoning workflows. Benchmarking state-of-the-art multimodal large language models reveals a significant performance gap, with GPT-4.1 achieving an accuracy of 69.3%, underscoring the challenges posed by this more authentic CQA setting.

1 Introduction

Data visualizations are an essential modality for communicating complex information about data. Alongside natural language, they serve as a key medium for communication across domains. As such, the ability to interpret and reason about visualizations is a crucial skill.

As multimodal large language models (MLLMs) evolve beyond simple perception tasks towards becoming visual assistants, there is growing interest in their ability to perform visual reasoning over structured data, including charts and other forms of data visualization. Tasks such as Chart Question Answering (CQA) have emerged for benchmarking a model's visualization reasoning capabilities.

In this work, we introduce a new dataset for CQA that aims to reflect the complexity of real-world data analysis. The dataset is constructed from student authored visualization notebooks, which combine explanatory analytical narrative with custom visualizations. Unlike existing CQA datasets, our dataset is grounded in ecologically valid analytical workflows. To situate this contribution, we first review prior work on visualization literacy and CQA. We then detail our data collection and question generation process, describing the structure and composition of the dataset. Finally, we report some initial benchmarking results using state-of-the-art MLLMs.

2 Related Work

Visualization Literacy datasets such as the visualization literacy assessment test (VLAT) (Lee et al., 2017) were initially created to assess human understanding of data visualizations. Recently, they have also been applied to probe the visualization literacy of MLLMs (Bendeck and Stasko, 2024). These manually curated datasets present small sets of charts paired with multiple-choice questions that probe the ability to perform specific analytic tasks such as retrieving values, identifying trends, or making comparisons. Whilst these tasks seem to mimic real-world analytical workflows (Amar et al., 2005), the hand-crafted design of these datasets limits their ability to accurately reflect the complexity of real-world visualization reasoning.

Chart Question Answering (CQA) is the task of answering a natural language question about a visualization image. CQA datasets are designed to benchmark the chart understanding capabilities of models. Early CQA benchmarks such as FigureQA (Kahou et al., 2018), DVQA (Kafle et al., 2018), and LEAF-QA (Chaudhry et al., 2020) used template-based questions and synthetically generated tasks. Again, these controlled settings are limited.

More recently, CQA datasets have moved toward real-world visualization images. Kim et al. (2020) and ChartQA (Masry et al., 2022) introduced chart images scraped from real-world reports and online sources. However, these datasets still only have questions that refer to a single chart, and do not include visualizations with multiple views or interactive elements. These datasets begin to reflect more realistic evaluation settings, but still do not completely capture visualization as done in-practice, where users often engage with visualizations that

have multiple views, such as dashboards or linked visualizations.

Some newer datasets begin to address this. CharXiv (Wang et al., 2024) includes charts composed of multiple sub views, although its questions still focus on one image. MultiChartQA (Zhu et al., 2025) allows questions to target multiple related visualizations, moving closer to the kinds of cross-chart reasoning analysts perform in practice. However, these datasets are still composed solely of static visualizations.

Another important distinction lies in how questions are generated. Some datasets, such as VLAT (Lee et al., 2017) and MultiChartQA (Zhu et al., 2025), rely exclusively on human-authored questions. While this approach ensures high-quality queries aligned with human reasoning, the scalability of dataset construction is limited. Conversely, other datasets like ChartQA (Masry et al., 2022) and CharXiv (Wang et al., 2024) adopt semi-automatic approaches, using models to produce questions alongside human validation, enabling larger datasets across more images.

Notably, previous datasets, whether template, human or machine-authored, are generated from the visualization image, caption, or from post hoc chart summaries. This often as a result of data collection processes that extract chart images in isolation, often scraped from online sources, removed from the surrounding analytical narrative. Due to the nature of source materials, this analytical context often does not exist at all and is left entirely implicit, available only from the visual context. The nature of these online sources may also raise copyright concerns due to the use of third-party images without explicit permission.

3 Methods

3.1 Data Collection

Our dataset is derived from literate visualization (litvis) notebooks, structured markdown documents that combine narrative analysis, code, embedded datasets, and inline visualizations (Wood et al., 2019). The notebooks were authored by undergraduate and postgraduate students as part of their final coursework for a 10-week data visualization module. These notebooks offer an ecologically valid window into real-world analytical practice: students independently selected datasets to analyze, posed research questions, and designed custom visualizations to explore those questions. These notebooks surface articulations of analytical reasoning that are typically left implicit in other sources of visualizations, providing a rich basis for question generation. See appendix D for an example notebook.

We applied several filtering steps to ensure data quality. Submissions were excluded if they lacked visualizations, included personally identifiable information, lacked sufficient narrative, or otherwise failed to meet basic quality thresholds. After filtering, we retained 22 notebooks for further processing.

From each retained notebook, we extracted two primary sources of data: the analytical narrative written by the student, and the corresponding visualizations. Visualizations were captured by rendering each notebook in HTML and using a headless browser to take screenshots of the embedded figures. Interactive visualizations were present in many of the notebooks, a feature missing from many sources of visualizations in CQA. To partially capture these interactive dynamics, we developed a method for capturing some interactive views statically. For visualizations with discrete interactive controls, such as radio buttons or drop-down menus, we systematically enumerated all categorical options and recorded screenshots of each interactive view. This allowed us to collect multiple views of the same visualization, reflecting user-driven analytical actions that are absent in existing datasets. To prepare the narrative for question generation, we segmented the extracted content into chunks of at most 200 words.

3.2 Question Generation

We structured our dataset according to established analytical task taxonomies from visualization research to ensure that the questions in our dataset reflect realistic analytical goals. Specifically, we adopt the eight task categories defined in the VLAT (Lee et al., 2017), which were curated from prior task taxonomies by Amar et al. (2005) and Chen et al. (2009). These tasks are: Retrieve Value, Find Extremum, Find Correlations, Make Comparisons, Characterize Distribution, Determine Range, Find Anomalies, and Find Clusters.

Our question generation pipeline centers on the analytical narrative authored by students. This approach is inspired by Changpinyo et al.'s (2022) work in visual question answering (VQA), who demonstrate the viability of generating high-quality question-answer pairs from language context rather

Dataset	Visualizations			Questions	
	Real-World	# Chart Types	Multi/Interactive	Unanswerable	Narrative Context
LeafQA (2020)	✗	6	✗/✗	✗	✗
Kim et al. (2020)	∼	2	✗/✗	✗	✗
ChartQA (2022)	✓	3	✗/✗	✗	✗
CharXiv (2024)	✓	*unbounded*	✓/✗	✓	✗
MultiChartQA (2025)	✓	*unbounded*	✓/✗	✓	✗
Ours	✓	*unbounded*	✓/✓	✓	✓

Table 1: Comparison between our dataset and existing chart question-answering datasets, grouped by visualization and question characteristics.

than visual context. This approach allows us to generate meaningful, grounded questions using an LLM without parsing the chart images.

For each segment, we prompted an LLM to generate a question-answer pair grounded in the context. The prompt provided a short description of each task category with representative examples. The model was asked to extract a relevant quote from the narrative, use it to generate a question-answer pair, and classify the pair according to the task taxonomy. The quote extraction allows us to verify the fidelity of the pair later in our validation process.

We then prompted the LLM to generate multiple choice distractors. The model received the narrative context, question-answer pair, and task classification, and was instructed to generate three plausible but incorrect alternative answers. The distractors were designed to match the structure and domain of the correct answer. Additionally, we appended a fifth answer option: *"Cannot be determined from the visualization(s)"*. This serves both as a realistic distractor and also as a correct answer choice for some questions, which will be determined during the validation process. Full prompt templates are provided in appendix B.

This pipeline yielded an initial set of 429 multiple-choice QA pairs, each grounded in the analytical context and aligned to an analytical task. These pairs then underwent a rigorous manual validation process.

3.3 Human Validation

All 429 LLM-generated QA pairs underwent stringent human validation by a data visualization expert to ensure the quality and reliability of the dataset. Each pair was reviewed against a set of rejection criteria, targeting two primary sources of invalid questions: (1) misalignment with the available visualizations, and (2) quality issues arising from the narrative context or generation process.

The first criterion focused on visualization alignment. Some visualizations were unable to render due to the unavailability of the underlying datasets, and because our QA generation process operated on the narrative context alone, some generated pairs referred to visualizations that could not be recovered during our data collection pipeline. Any QA pair that could not be reliably related to at least one available visualization was excluded.

The second rejection criterion addressed the scope of the narrative context and generation quality. Some students describe aspects unrelated to analytical insights, such as dataset collection challenges, findings they found surprising, or general reflections. While these are interesting and valuable parts of the students' process, they are out of scope for this dataset and so QA pairs generated from this context were excluded.

During validation, we also explicitly associated each accepted QA pair with the specific views it referenced, as each notebook often included multiple charts. In some cases, questions required information that was only visible interactive views not captured, often tooltip values. When a question did relate to an available chart but remained unanswerable due to missing context, we retained it and assigned it "cannot be determined".

4 Dataset Analysis

Following validation, we retained 205 high-quality QA pairs, corresponding to 103 visualization images. 75 questions, 36.6%, have multiple visualization images or multiple views. 33 questions, 16.1% of questions are unanswerable. Table 1 provides a comparison of our dataset to previous work across key visualization and question characteristics.

Table 2 provides a breakdown of question types in the dataset by visualization task. The observed imbalance reflects the natural distribution of an-

Task	Count	GPT-4.1	Qwen2.5-VL-32B	Qwen2.5-VL-7B
All	205	69.27%	56.59%	31.71%
Retrieve Value	68	76.47%	55.88%	25.00%
Find Extremum	55	69.09%	60.00%	36.36%
Find Correlations	22	72.73%	54.55%	27.27%
Make Comparisons	22	50.00%	59.09%	50.00%
Characterize Distribution	15	66.67%	46.67%	20.00%
Determine Range	12	75.00%	58.33%	41.67%
Find Anomalies	9	44.44%	55.56%	33.33%
Find Clusters	2	100.00%	50.00%	0.00%

Table 2: Accuracy by task type for GPT-4.1 and Qwen2.5-VL models. The top row reports overall accuracy across all tasks, followed a task breakdown, ordered by task frequency.

alytical strategies employed by students in their projects. Tasks such as Retrieve Value and Find Extremum are most common, suggesting a strong emphasis on identifying specific data points or extreme values. Conversely, higher-order tasks like Find Clusters or Find Anomalies are relatively rare.

5 Model Evaluation

We evaluated the performance of two state-of-the-art vision-language models on our dataset: OpenAI's proprietary GPT-4.1 (OpenAI, 2025) and Alibaba's open-weight Qwen2.5-VL models at two parameter scales (7B and 32B) (Bai et al., 2025). Each model was presented with the question and corresponding visualization(s) and tasked with selecting the correct answer from the five multiple-choice options.

As shown in Table 2, GPT-4.1 achieved the highest accuracy at 69.27%, outperforming both versions of Qwen2.5-VL. The 32B variant of Qwen2.5-VL attained a moderate accuracy of 56.59%, while the 7B variant lagged significantly at 31.71%. This performance disparity underscores the impact of model scale on complex visual question answering tasks. Appendix C provides some examples from our dataset alongside GPT4.1's responses.

Table 2 presents model accuracy broken down by question type. GPT-4.1 demonstrates consistently strong performance across most tasks, exceeding 66% accuracy in five of the eight categories. It performs particularly well on Retrieve Value and Determine Range, tasks that rely on precise visual extraction, suggesting strong literal comprehension of chart elements. However, its performance drops on more interpretive tasks such as Make Comparisons (50.00%), perhaps indicating challenges with contextual or higher-order reasoning. Interestingly, Qwen2.5-VL-32B outperforms GPT-4.1 on these two tasks, despite trailing on most others, suggesting possible strengths in certain visual discrimination tasks. The 7B variant of Qwen2.5-VL performs substantially worse across nearly all categories, aside from Make Comparisons, where it matches GPT-4.1's performance.

Caution is however warranted when interpreting results for less frequent task types such as Find Anomalies and Find Clusters, which contain relatively few questions. Despite this, the overall trends suggest that performance differences across task types are meaningful, and that structured taxonomies offer useful insight into the capabilities and limitations of current MLLMs in chart understanding.

6 Conclusion

Our dataset introduces a more realistic and ecologically grounded benchmark for chart question answering, reflecting how visualizations are created and interpreted in practice. By capturing analytical narratives, multiple and interactive views, it challenges current models in ways prior datasets do not. Initial evaluations highlight substantial performance gaps, pointing to the need for models with deeper reasoning and contextual understanding of visual data. We observe significant variance in model performance across task types, suggesting that certain forms of visual reasoning remain especially challenging. We intend release the dataset publicly to support future research. We hope this dataset fosters future research toward more capable and context-aware multimodal systems.

Ethics Statement

This study and its data collection procedures were formally approved by our university's Research Ethics Committee. Upon receiving approval, we contacted graduates of the program to inform them about the study's aims and potential contributions. We obtained explicit informed consent from those who agreed to participate, specifically for the use of their coursework in our research. The dataset exclusively comprises submissions from students who voluntarily provided permission for their materials to be processed and released as part of this research.

Limitations

While our dataset offers a more ecologically grounded benchmark for CQA, it has several limitations. Firstly, the task distribution is imbalanced, with lower-level tasks like Retrieve Value more common and higher-order tasks like Find Clusters underrepresented. Future work could curate a more balanced set to cover a wider range of reasoning types. Secondly, the dataset includes only 205 validated question–answer pairs. This limited size reflects our emphasis on rigorous human validation to ensure alignment between questions, narratives, and visualizations. Our methodology could be extended to larger corpora of visualization notebooks to create a more expansive dataset. Finally, all questions are in English. While the tasks are conceptually broad, some formulations may not generalize well across languages. Future efforts could explore multilingual extensions by incorporating narratives from other languages.

References

R. Amar, J. Eagan, and J. Stasko. 2005. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 111–117. ISSN: 1522-404X.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL Technical Report. *arXiv preprint*. ArXiv:2502.13923 [cs].

Alexander Bendeck and John Stasko. 2024. An Empirical Evaluation of the GPT-4 Multimodal Language Model on Visualization Literacy Tasks. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–11. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

Soravit Changpinyo, Doron Kukliansy, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. 2022. All you may need for VQA are image captions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1947–1963, Seattle, United States. Association for Computational Linguistics.

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. LEAF-QA: Locate, Encode & Attend for Figure Question Answering. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3501–3510. ISSN: 2642-9381.

Yang Chen, Jing Yang, and William Ribarsky. 2009. Toward effective insight management in visual analytics systems. In *2009 IEEE Pacific Visualization Symposium*, pages 49–56. ISSN: 2165-8773.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: Understanding Data Visualizations via Question Answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5648–5656. ISSN: 2575-7075.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2018. FigureQA: An Annotated Figure Dataset for Visual Reasoning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net.

Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. 2020. Answering Questions about Charts and Generating Visual Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–13, New York, NY, USA. Association for Computing Machinery.

Sukwon Lee, Sung-Hee Kim, and Bum Chul Kwon. 2017. VLAT: Development of a Visualization Literacy Assessment Test. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):551–560. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2025. GPT-4.1 (April 14 version).

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. CharXiv: Charting

Gaps in Realistic Chart Understanding in Multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pages 113569–113697. Curran Associates, Inc.

Jo Wood, Alexander Kachkaev, and Jason Dykes. 2019. Design Exposition with Literate Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):759–768.

Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. 2025. MultiChartQA: Benchmarking vision-language models on multi-chart problems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11341–11359, Albuquerque, New Mexico. Association for Computational Linguistics.

A Task Information

Task Name	Description	Pro Forma Abstract	Examples (Q → A)
Retrieve Value	Given a set of specific cases, find attributes of those cases.	What are the values of attributes {X, Y, Z, ...} in the data cases {A, B, C, ...}?	What was the price of a barrel of oil in February 2015? → $50 What is the average internet speed in Japan? → 15.3 Mbps What is the weight of the person who is 165.1 cm tall? → 60 kg
Find Extremum	Find data cases possessing an extreme value of an attribute.	What are the top/bottom N data cases with respect to attribute A?	In which month was the price of a barrel of oil the lowest in 2015? → August Which country has the fastest average internet speed in Asia? → South Korea What is the height of the tallest person among the 85 males? → 198 cm
Determine Range	Find the span of values of an attribute within a set.	What is the range of values of attribute A in a set S of data cases?	What was the price range of a barrel of oil in 2015? → $38 to $60 What is the range of average internet speeds in Asia? → 4.3 Mbps to 15.3 Mbps What is the weight range among the 85 males? → 52 kg to 90 kg
Characterize Distribution	Characterize the distribution of a quantitative attribute.	What is the distribution of values of attribute A in a set S of data cases?	How is the distribution of taxi passenger ratings characterized? → Skewed to the left What is the distribution pattern of student grades in the dataset? → Approximately normal distribution centered around 75%
Find Anomalies	Identify anomalies within a set of data cases.	Which data cases in a set S of data cases have unexpected/exceptional values?	Which individual's height deviates most from the others? → 210 cm Which city's metro system deviates most from the trend? → Beijing
Find Clusters	Find clusters of similar attribute values.	Which data cases are similar in value for attributes {X, Y, Z, . . . }?	Describe any groups of individuals who share similar height and weight characteristics. → A group is clustered around 176 cm in height and 70 kg in weight. What patterns of similarity can you find among metro systems based on number of stations and system length? → Several metro systems are clustered around 300 stations and 200 km length.
Find Correlations	Determine relationships between two attributes.	What is the correlation between attributes X and Y in a set S?	What is the relationship between height and weight? → Negative linear How does ridership relate to stations? → Positive correlation Trend in coffee prices over 2013? → Increasing
Make Comparisons	Compare sets of cases with respect to an attribute.	How do data cases compare with respect to attribute A?	Apple vs Huawei market share? → Apple's is larger Ratings between 4.6–4.8 and 4.2–4.4? → 4.6–4.8 has more Shanghai vs Beijing ridership? → Shanghai's is higher

B Prompts

Prompt: QA Generation

```
You are a data visualization expert and question-generation assistant.

Given the following TEXT:

{ANALYTICAL CONTEXT}

Your task is to generate between 3 and 10 QUESTION-ANSWER pairs based on the TEXT,
and assign each one to the most appropriate TASK listed below.

Only generate questions if the information in the TEXT is clearly related to a task.

{TASK INFORMATION}

### Output Instructions:
- For each QA pair, include:
    - The direct **quote** from the TEXT
    - The **question**
    - The **answer**, which should be concise and suitable for a multiple choice test
    - The **most appropriate TASK** name from the list
- Only generate a question if it fits into one of the tasks.
- Do not repeat questions
- Prefer fewer, high-quality questions
- Avoid yes/no or true/false answers.
- Output must be a JSON list of dictionaries, like this:

```json
[
  {"quote": "Example quote", "q": "Example question?", "a": "Answer.", "task": "Retrieve Value"},
  ...
]
```
```

Prompt: Answer Choices Generation

```
You are creating a multiple choice question about data visualization.

Given the following context:
Context: {ANALYTICAl CONTEXT}

We have a question and answer pair:
Question: {QUESTION}
Correct Answer: {ANSWER}

Generate 3 **plausible but incorrect** answer choices. These should:
- Be related to the same context
- Be in the same format as the correct answer
(e.g. numerical with the same units, textual with similar length)
- Be different from the correct answer
- Be wrong
- DO NOT make answers that are along the lines of cannot be determined/don't know/can't tell

Output as only a Python list: ["a1", "a2", a3"]
```

Prompt: Model Evaluation

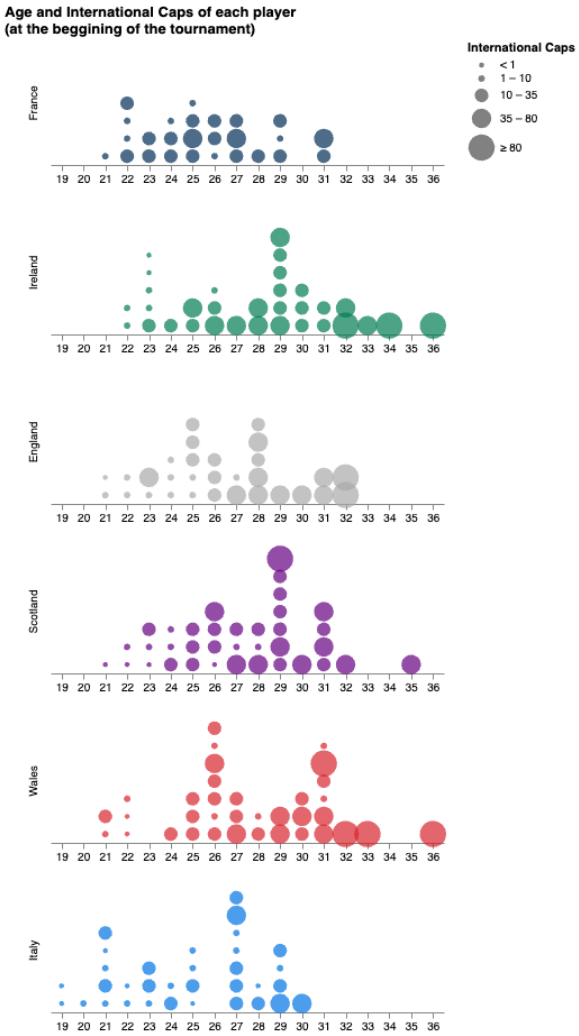
```
Question: {QUESTION}

Answer choices: {ANSWER CHOICES}

Please respond with ONLY the letter (A, B, C, D or E) corresponding to your answer.
```

C Examples from the Dataset

Faceted Views

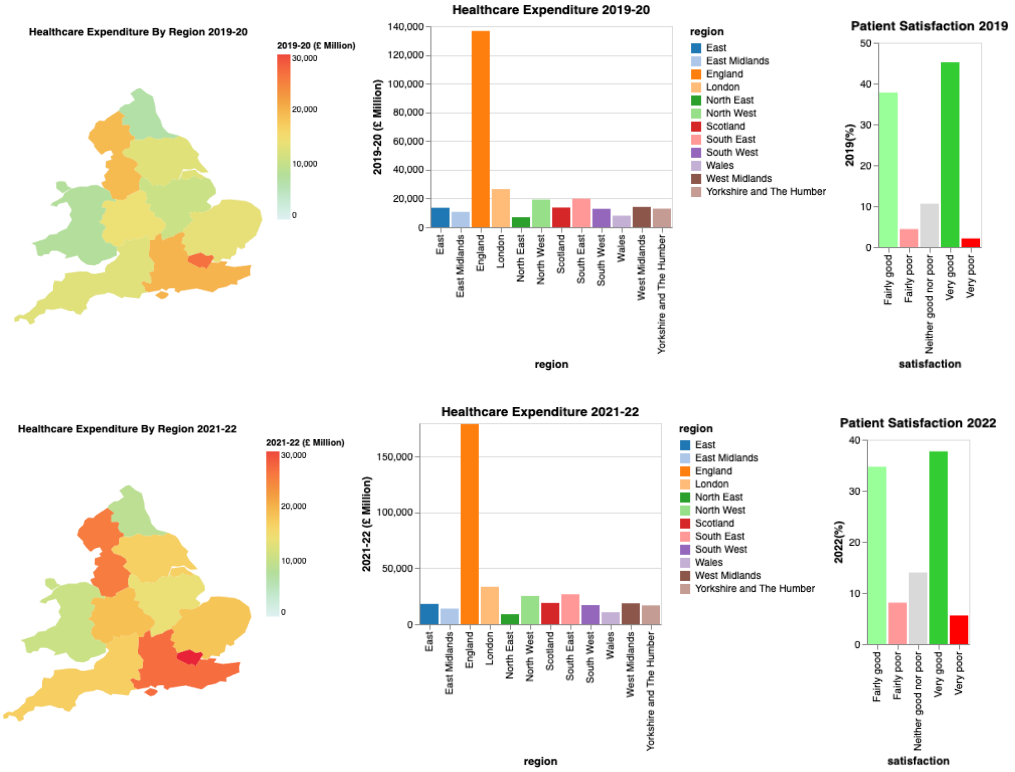


Retrieve Value: What is the range of ages in the France rugby team?
Answers: 14 years, 10 years, 8 years, 15 years, Cannot be determined from the visualization(s)
GPT 4.1: 8 years

Find Extremum: Which team has the narrowest age range?
Answers: France, Ireland, Scotland, Wales, Cannot be determined from the visualization(s)]
GPT 4.1: France

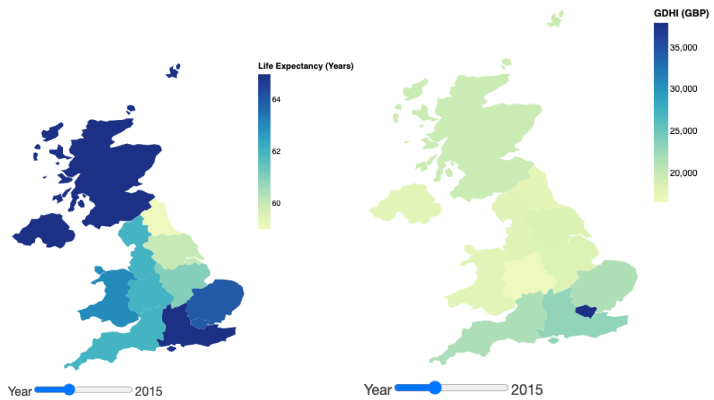
Make Comparisons: How does the age range of the France rugby team compare to that of Wales?
Answers: France's range is wider than Wales', France's range is the same as Wales', France's range is narrower than Wales', France's range is 7 years less than Wales', Cannot be determined from the visualization(s)
GPT 4.1: France's range is narrower than Wales'

Multiple Images



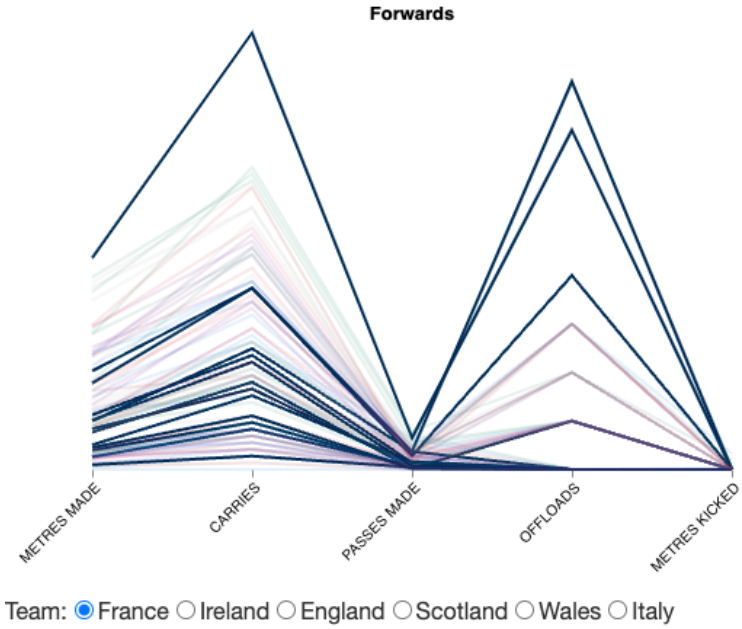
Find Correlations: What is the relationship between healthcare expenditure and patient satisfaction between 2019 and 2022?
Answers: Patient satisfaction remained relatively stable despite increased expenditure., Healthcare expenditure declined, leading to decreased patient satisfaction., Patient satisfaction increased with increased expenditure., Despite increased expenditure, patient satisfaction declined., Cannot be determined from the visualization(s)
GPT 4.1: Patient satisfaction remained relatively stable despite increased expenditure.

Multiple Images



Retrieve Value: What is the life expectancy and GDHI of Northern Ireland?
Answers: 65 years and £20,916, 65 years and £17,916, 60 years and £27,916, 75 years and £15,916, Cannot be determined from the visualization(s)
GPT 4.1: 65 years and £20,916

Interactive View, Cannot be determined



Find Anomalies: Which French forwards have unusually high offload numbers compared to other forwards?
Answers: Gael Fickou and Damian Penaud, Gregory Alldritt and Antoine Dupont, Cyril Baille and Francois Cros, Cyril Baille and Gregory Alldritt, Cannot be determined from the visualization(s)
GPT 4.1: Cyril Baille and Gregory Alldritt

D Example Literate Visualization Notebook

Exploring the 2022 Six Nations Championship

What are the research questions that your data visualization will help you to answer?

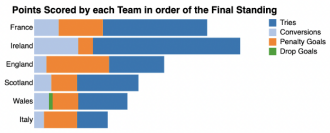
Research Questions

- How does the experience and structure of each team vary?
- In what ways did each position play differently? Is there a distinction between the play of backs and forwards?
- What are the differences in the way that each team played?

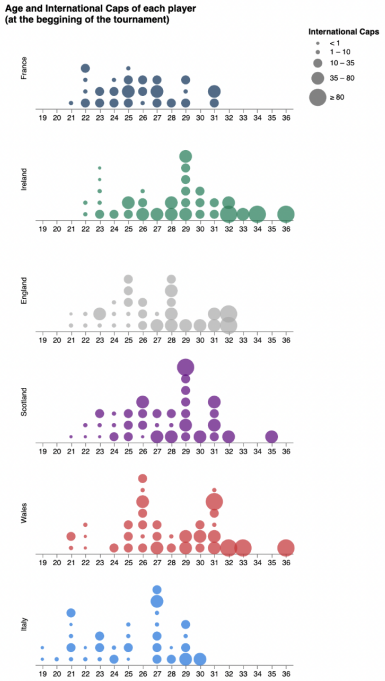
1. The Visualization

Insert your visualization here.

1.1 Context: Final Standing and Points Scored



1.2 Team Experience and Structure

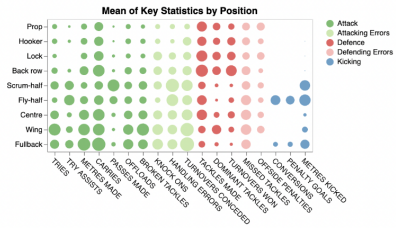


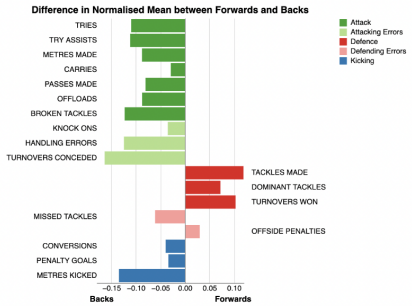
467

1.3 Player Statistics by Position

Contextual note:

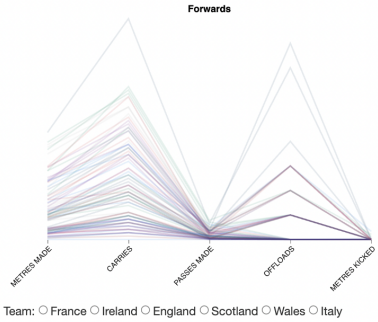
The forward positions are Prop, Hooker, Lock, and Back Row (which consists of two Flankers and a Number 8). The back positions are Scrum Half, Fly Half, Centre, Wing and Fullback.

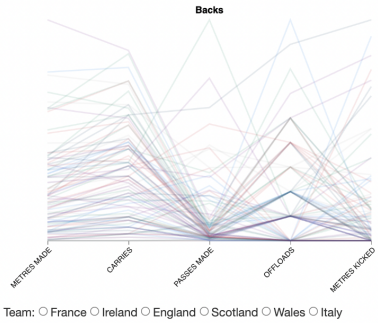




468

1.4 Player Statistics by Team





2. Insights

What has your visualization allowed you to discover about your data that help you answer your research questions? Identify a maximum of the 3 most important insights that result directly from your visualization.

2.1 Insight One - Team Experience and Structure

From an initial look at visualisation 1.2, there are some similarities in the general age and experience structure of each team: the older players tend to hold the most experience, as would be expected, whilst there are several younger players with fewer caps, and the mode age of each team tends to be in the mid-to-late twenties. However, there are some interesting differences between each team.

All teams played some younger, less experienced players in this championship, with each team having several players under 25 with fewer than 10 international caps. France, who won the grand slam, are the team with the fewest inexperienced players, having played no players with no previous caps. This is in stark contrast to Italy, the losing team, who clearly have the youngest, least experienced team. It is important to note that these are only the players who actually played and not the entire training squad, so this does not necessarily mean that France will lack depth in the future, but it will be interesting to see which new players they choose to cap in the lead up to the 2023 Rugby World Cup.

As mentioned, the mode of each team tends to be in the mid-to-late twenties, but there is a clear difference in the distribution surrounding this mode. Teams like Ireland and Scotland, both with a mode of 29, show a peak that stands out compared to the rest of the age distribution. Conversely, France has two modes at 22 and 25, and the general distribution of ages looks quite uniform.

In terms of the width of each distribution, Ireland and Wales show a similar structure of a wide range of ages, with a tail of a few very experienced players over the age of 31. Again, this contrasts with the uniform distribution of France, which has the narrowest range of ages of only 10 years compared to Wales' 15 years. Looking again towards the 2023 World Cup, this may indicate that France's squad is much more resilient against retirement and injuries than the other teams, whereas Ireland and Wales could be at risk of losing their key experience on the pitch before 2023 with their exposed tail of older players who are more likely to retire.

2.2 Insight Two - The Role of Different Positions

Visualisation 1.3 reveals the importance of position in rugby — each position has a certain role to fulfil on the pitch, and this is especially clear when split into forwards and backs.

The forwards play much more of a defensive role compared to the backs, having a higher normalised mean in only four categories: tackles made, dominant tackles, turnovers won, and offside penalties. Excluding the offside penalties, these are the three most physical jobs on the pitch. In defence, backs tend to miss more tackles than forwards, likely due to being more isolated when defending than forwards. Whilst it may seem that forwards do fewer jobs than backs, it is important to note that there are no set-piece statistics included (the scrum and lineout), which is a key job for the forwards. Perhaps their lack of dominance in the categories shown highlights the importance of the forwards' role in the set-piece.

In terms of specific positions within the forwards, we can see that the back row is much more agile in attack than the other forward positions, on average making more metres, carries, and breaking more tackles. This is likely related to their looser position in the scrum. It is interesting that not a single lock player made a try assist in the entire championship. Again, this is illustrative of their positioning in the set-piece — they are stuck in the middle of the scrum, and are often in charge of the lineout, so would not be in the right position to assist a try. This is also shown in that they scored fewer tries on average than the other forwards.

Looking now towards the backs, we see that they dominate in the attacking categories, making the most metres, scoring the most tries and breaking the most tackles. As a result, backs also dominate in terms of the attacking errors made, because, as demonstrated by the other statistics, they have more time on the ball in attack, and are also likely to be involved in more complex attacking plays than the forwards.

In terms of specific backs positions, there are a few players with very specific roles that are demonstrated in the visualisation. The scrum-half on average makes far more passes than any other position on the pitch, because they have to recycle the ball at the back of each ruck. The fly-half makes the second-highest number of passes on average, demonstrating their role as a playmaker. It is also interesting to see that fullbacks conceded no offside penalties in the championship, highlighting their different role in defence compared to the other players. Fullbacks usually stand behind the defensive line to cover kicks and breakthrough tackles, so would never be in the position to be offside.

A role that is exclusive to the backs is kicking. In terms of kicking on the pitch, scrum-halves, fly-halves and fullbacks kick much more often than centres or wingers. Position kicks (conversions and penalties) are most often taken by the fly-half, but are taken by the fullback on some teams, hence the two dots in those categories.

One overall trend that is quite compelling is turnovers conceded. There is a trend running from prop to fullback, with turnovers conceded increasing down the list. This demonstrates nicely the difference in role and positioning of the different positions — wingers and fullbacks are much more likely

2.2 Insight Three - Team Style of Play

As established, forwards and backs have very different roles so visualisation 1.4 is split into forwards and backs, so that the different style of play between each team can be established more clearly. Data are normalised over all players, however, so comparisons can be drawn between the two graphs. Also, only key attacking statistics are included, as this is where the style of play will be demonstrated the most.

Firstly, looking at the forwards. We can see that the metres made and carries roughly tend to decrease as we go down the table. This makes sense because teams who lose more often will spend less time in attack than winning teams. However, there are a few key differences between France and the rest of the teams. Most of France's forwards actually make a similar number or fewer metres made than forwards of other teams, apart from Gregory Alldritt who by far makes the most carries of not only any forward, but of any player in the whole tournament. In terms of offloads, forwards in all teams besides France make a similar, fairly low number, but France's forwards, in particular prop Cyril Baille and back row Gregory Alldritt, make far more offloads than the other forwards, with numbers rivalling the backs. As explored in the previous section, it is unusual for players in these positions to make so many plays in attack. If the forwards are able to keep the ball alive by successfully offloading, it allows the attacking team to maintain momentum and thus break the defensive line, scoring more tries. This illustrates that France perhaps have a different style of play in the forwards with more attacking flair that helped them to win the grand slam.

Now, looking at the backs. Compared to the forwards plot, the lines are much less clustered at each axis, again, demonstrating that the forwards have much more specific roles than the
backs, who tend to make more varied plays. The messiness of the lines makes it harder to gather any overall trends about style of play in the backs, but there still are some comparisons to be drawn between the teams, especially focussing again on Grand Slam winners, France.

Despite winning all of their games, France's backs actually make fewer metres and carries compared to the other teams, but their scrum-half, Antoine Dupont, kicks the highest number of metres of any player, despite not being a position kicker. This demonstrates that France often choose not to hold onto possession, rather kicking it away on their own terms and letting the other team attack, a bold tactic which clearly worked to their advantage. This shows the confidence that France have in their defence. Conversely, the second team in the table, Ireland, make substantially more carries and kick fewer metres, suggesting that they are more inclined to maintain possession of the ball in attack and run it up, perhaps playing a more structured game. This difference in tactic is reflected in visualisation 1.1, where it can be seen that despite coming second, Ireland actually scored the most points on the pitch, particularly gaining points from tries. Once more, it will be interesting to see how these tactics develop as we look towards the World Cup.