

# PlotQA: Reasoning over Scientific Plots

Nitesh Methani \*

Pritha Ganguly\*

Mitesh M. Khapra

Pratyush Kumar

Department of Computer Science and Engineering  
Robert Bosch Centre for Data Science and AI (RBC-DSAI)  
Indian Institute of Technology Madras, Chennai, India  
{nmethani, prithag, miteshk, pratyush}@cse.iitm.ac.in

## Abstract

Existing synthetic datasets (FigureQA, DVQA) for reasoning over plots do not contain variability in data labels, real-valued data, or complex reasoning questions. Consequently, proposed models for these datasets do not fully address the challenge of reasoning over plots. In particular, they assume that the answer comes either from a small fixed size vocabulary or from a bounding box within the image. However, in practice, this is an unrealistic assumption because many questions require reasoning and thus have real-valued answers which appear neither in a small fixed size vocabulary nor in the image. In this work, we aim to bridge this gap between existing datasets and real-world plots. Specifically, we propose PlotQA with 28.9 million question-answer pairs over 224,377 plots on data from real-world sources and questions based on crowd-sourced question templates. Further, 80.76% of the out-of-vocabulary (OOV) questions in PlotQA have answers that are not in a fixed vocabulary. Analysis of existing models on PlotQA reveals that they cannot deal with OOV questions: their overall accuracy on our dataset is in single digits. This is not surprising given that these models were not designed for such questions. As a step towards a more holistic model which can address fixed vocabulary as well as OOV questions, we propose a hybrid approach: Specific questions are answered by choosing the answer from a fixed vocabulary or by extracting it from a predicted bounding box in the plot, while other questions are answered with a table question-answering engine which is fed with a structured table generated by detecting visual elements from the image. On the existing DVQA dataset, our model has an accuracy of 58%, significantly improving on the highest reported accuracy of 46%. On PlotQA, our model has an accuracy of 22.52%, which is significantly better than state of the art models.

## 1. Introduction

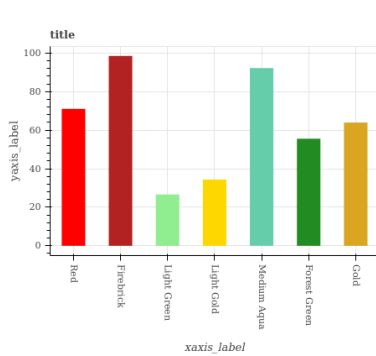
Data plots such as bar charts, line graphs, scatter plots, etc. provide an efficient way of summarizing numerical information. Recently, in [13, 12] two datasets containing plots and deep neural models for question answering over the generated plots have been proposed. In both the datasets, the plots are synthetically generated with data values and labels drawn from a custom set. In the FigureQA dataset [13], all questions are binary wherein answers are either Yes or No, (see Figure 1a for an example). The DVQA dataset [12], generalizes this to include questions which can be answered either by (a) fixed vocabulary of 1000 words, or (b) extracting text (such as tick labels) from the plot. An example question could seek the numeric value represented by a bar of a specific label in a bar plot (see Figure 1b). Given that all data values in the DVQA dataset are chosen to be integers and from a fixed range, the answer to this question can be extracted from the appropriate tick label.

While these datasets have initiated the research questions on plot reasoning, realistic questions over plots are much more complex. For instance, consider the question in Figure 1c, where we are to compute the average of floating point numbers represented by three bars of a color specified by the label. The answer to this question is neither in a fixed vocabulary nor can it be extracted from the plot itself. Answering such questions requires a combination of perception, language understanding, and reasoning, and thus poses a significant challenge to existing systems. Furthermore, this task is harder if the training set is not synthetic, but instead is sourced from real-world data with large variability in floating-point values, large diversity in axis and tick labels, and natural complexity in question templates.

To address this gap between existing datasets and real-world plots, we introduce the PlotQA<sup>1</sup> dataset with 28.9 million question-answer pairs grounded over 224,377 plots. PlotQA improves on existing datasets on three fronts. First,

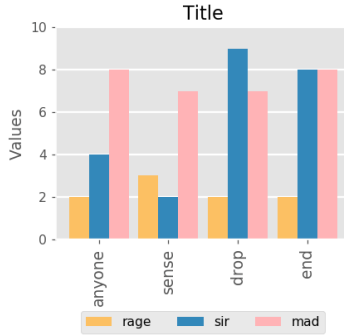
\*The first two authors have contributed equally

<sup>1</sup>Dataset can be downloaded from [bit.ly/PlotQA](http://bit.ly/PlotQA)



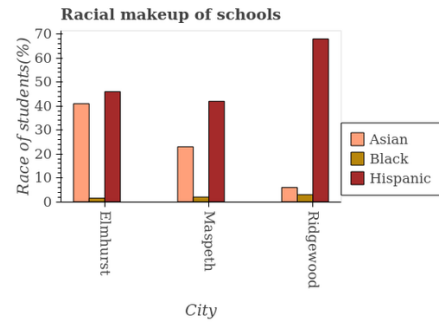
Q: Is Light Green the minimum?  
A: 1

(a) FigureQA



Q: What is the value of mad in drop?  
A: 7

(b) DVQA



Q: What is the average number of Hispanic students in schools? A: 51.67

(c) PlotQA

Figure 1: A sample {plot, question, answer} triplet from FigureQA, DVQA, and PlotQA (our) datasets.

| Answer Type      | Question Type                      |   |   |
|------------------|------------------------------------|---|---|
|                  | Structure                          | Data Retrieval  | Reasoning   |
| Yes/No           | Does the graph contain grids?      | Does the price of diesel in Barbados monotonically increase over the years? | Is the difference between the price of diesel in Angola in 2002 and 2004 greater than the difference between the price of diesel in Lebanon in 2002 and 2004? |
| Fixed vocabulary | How are the legend labels stacked? | What is the label or title of the X-axis ?                                  | In how many years, is the price of diesel greater than 0.6 units?   |
| Open vocabulary  | -                                  | What is the price of diesel in Lebanon in the year 2008?                    | What is the ratio of the price of diesel in Lebanon in 2010 to that in 2014?  |

Table 1: Sample questions for 9 different question-answer types in PlotQA. The example questions are with respect to the plot in Figure 2b. Note that there are no open vocabulary answers for *Structural Understanding* questions.

roughly 80.76% of the questions have answers which are not present in the plot or in a fixed vocabulary. Second, the plots are generated from data sourced from World Bank, government sites, etc., thereby having a large vocabulary of axis and tick labels, and a wide range in data values. Third, the questions are complex as they are generated based on 74 templates extracted from 7,000 crowd-sourced questions asked by workers on a sampled set of 1,400 plots. Questions are categorized into 9 (=3x3) cells based on the question type: ‘Structural Understanding’, ‘Data Retrieval’, or ‘Reasoning’ and the answer type: ‘Yes/No’, ‘From Fixed Vocabulary’, or ‘Out Of Vocabulary (OOV)’ (see Table 1).

We first evaluate three state of the art models on PlotQA, viz., SAN-VQA[36], Bilinear attention network (BAN) [16] and LoRRA [33]. Note that, by design none of these models are capable of answering OOV questions. In particular, SAN-VQA and BAN treat plot reasoning as a classification task and expect the answer to lie in a small vocabulary whereas in our dataset the answer vocabulary is prohibitively large (~5M words). Similarly, LoRRA assumes that the answer is present in the image itself as a text and the task is to just extract this region containing the text followed by OCR (optical character recognition). Again, such

a model will be unable to answer questions such as the one shown in Figure 1c, which form a significant segment of real-world use-cases and our dataset. As a result, these models give an accuracy of less than 8% on our dataset. On the other hand, existing models (in particular, SAN) perform well on questions with answers from a fixed vocabulary, which was the intended purpose of these models.

Based on the above observations, we propose a hybrid model with a binary classifier which given a question decides if the answer would lie in a small top- $k$  vocabulary or if the answer is OOV. For the former, the question is passed through a classification pipeline which predicts a distribution over the top- $k$  vocabulary and selects the most probable answer. For the latter (arguably harder questions), we pass the question through a pipeline of four modules: Visual element detection, Optical character recognition, Extraction into a structured table, and Structured table question answering. This proposed hybrid model significantly outperforms the existing models and has an aggregate accuracy of 22.52% on the PlotQA dataset. We also evaluate our model on the DVQA dataset where it gives an accuracy of 58%, improving on the best-reported result of SANDY [12] of 46%. In summary, we make two major contributions:

| Datasets | #Plot types | #Plot images | #QA pairs  | Vocabulary   | Avg. question length | #Templates                   | #Unique answers | Open vocab. |
|----------|-------------|--------------|------------|--|----------------------|------------------------------|-----------------|-------------|
| FigureQA | 4           | 180,000      | 2,388,698  | 100 colours from X11 colour set                      | 7.5                  | 15<br>(no variations)        | 2               | Not present |
| DVQA     | 1           | 300,000      | 3,487,194  | 1K nouns from Brown corpus                           | 12.30                | 26<br>(without paraphrasing) | 1576            | Not present |
| PlotQA   | 3           | 224,377      | 28,952,641 | Real-world axes variables and floating point numbers | 43.54                | 74<br>(with paraphrasing)    | 5,701,618       | Present     |

Table 2: Comparison between the existing datasets (FigureQA and DVQA) and our proposed dataset (PlotQA).

(1) We propose PlotQA dataset with plots on data sourced from the real-world and questions based on templates sourced from manually curated questions. The dataset exposes the need to train models for questions that have answers from an Open Vocabulary.

(2) We propose a hybrid model with perception and QA modules for questions that have answers from an Open Vocabulary. This model gives the best performance not only on our dataset but also on the existing DVQA dataset.

## 2. Related Work

**Datasets:** Over the past few years several large scale datasets for Visual Question Answering have been released. These include datasets such as COCO-QA [28], DAQUAR [23], VQA [1, 7] which contain questions asked over natural images. On the other hand, datasets such as CLEVR [11] and NVLR [35] contain complex reasoning based questions on synthetic images having 2D and 3D geometric objects. There are some datasets [14, 15] which contain questions asked over diagrams found in text books but these datasets are smaller and contain multiple-choice questions. FigureSeer [31] is another dataset which contains images extracted from research papers but this is also a relatively small (60,000 images) dataset. Further, FigureSeer focuses on answering questions based on line plots as opposed to other types of plots such as bar charts, scatter plots, *etc.* as seen in FigureQA [13] and DVQA [12]. There is also the recent TextVQA [33] dataset which contains questions which require models to read the text present in natural images. This dataset does not contain questions requiring numeric reasoning. Further, the answer is contained as a text in the image itself. Thus, *no* existing dataset contains plot images with complex questions which require reasoning and have answers from an Open Vocabulary.

**Models:** The availability of the above mentioned datasets has facilitated the development of complex end-to-end neural network based models ([36], [22], [37], [25], [30], [12], [33]). These end-to-end networks contain (a) encoders to compute a representation for the image and the question, (b) attention mechanisms to focus on important parts of the question and image, (c) interaction components to capture the interactions between the question and the image, (d)

OCR module to extract the image specific text and (e) a classification layer for selecting the answer either from a fixed vocabulary or from a OCR appended vocabulary.

## 3. The PlotQA dataset

In this section, we describe the PlotQA dataset and the process to build it. Specifically, we discuss the four main stages, *viz.*, (i) curating data such as year-wise rainfall statistics, country-wise mortality rates, *etc.*, (ii) creating different types of plots with a variation in the number of elements, legend positions, fonts, *etc.*, (iii) crowd-sourcing to generate questions, and (iv) extracting templates from the crowd-sourced questions and instantiating these templates using appropriate phrasing suggested by human annotators.

### 3.1. Data Collection and Curation

We considered online data sources such as World Bank Open Data, Open Government Data, Global Terrorism Database, *etc.* which contain statistics about various indicator variables such as fertility rate, rainfall, coal production, *etc.* across years, countries, districts, *etc.* We crawled data from these sources to extract different variables whose relations could then be plotted (for example, rainfall v/s years across countries, or movie v/s budget, or carbohydrates v/s food\_item). There are a total of 841 unique indicator variables (CO2 emission, Air Quality Index, Fertility Rate, Revenue generated, *etc.*) with 160 unique entities (cities, states, districts, countries, movies, food, *etc.*). The data ranges from 1960 to 2016, though not all indicator variables have data items for all years. The data contains positive integers, floating point values, percentages, and values on a linear scale. These values range from 0 to 3.50e+15.

### 3.2. Plot Generation

We included 3 different types of plots in this dataset, *viz.*, bar plots, line plots, and scatter plots. Within bar plots, we have grouped them by orientation as either horizontal or vertical. Figure 2 shows one sample of each plot type. Each of these plot types can compactly represent 3-dimensional data. For instance, in Figure 2b, the plot compares the diesel prices across years for different countries. To enable the development of supervised

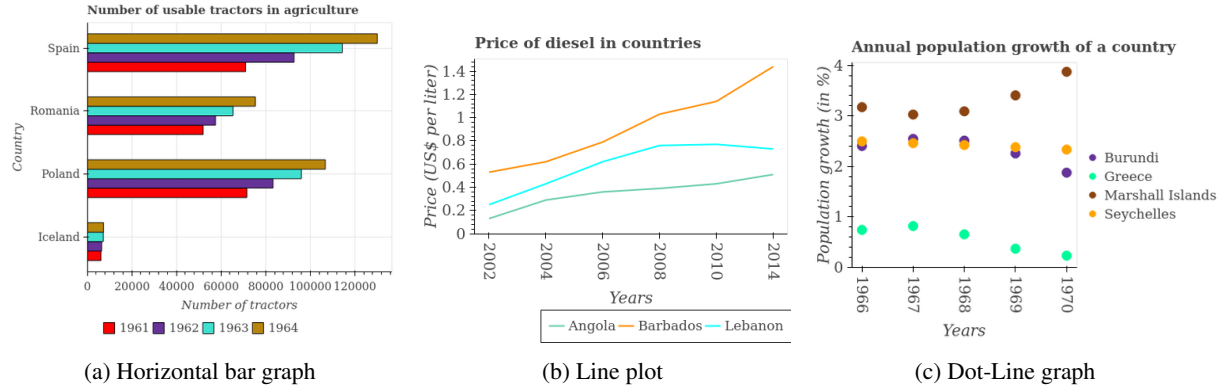


Figure 2: Sample plots of different types in the PlotQA dataset.

modules for various sub-tasks, we provide bounding box annotations for legend boxes, legend names, legend markers, axes titles, axes ticks, bars, lines, and title. By using different combinations of indicator variables and entities (years, countries, *etc.*) we created a total of 224,377 plots.

To ensure variety in the plots, we randomly chose the following parameters: grid lines (present/absent), font size, notation used for tick labels (scientific-E notation or standard notation), line style (solid, dashed, dotted, dash-dot), marker styles for marking data points (asterisk, circle, diamond, square, triangle, inverted triangle), position of legends (bottom-left, bottom-centre, bottom-right, center-right, top-right), and colors for the lines and bars from a set of 73 colors. The number of discrete elements on the  $x$ -axis varies from 2 to 12 and the number of entries in the legend box varies from 1 to 4.

### 3.3. Sample Question Collection by Crowd-sourcing

As the source data of PlotQA dataset is significantly richer in comparison to FigureQA and DVQA, we found it necessary to ask a larger set of annotators to create questions over these plots. However, creating questions for all the plots in our dataset would have been prohibitively expensive. We sampled 1,400 plots across different types and asked workers on Amazon Mechanical Turk to create questions for these plots. We showed each plot to 5 different workers resulting in a total of 7,000 questions. We specifically instructed the workers to ask complex reasoning questions which involved reference to multiple plot elements in the plots. We paid the workers USD 0.1 for each question.

### 3.4. Question Template Extraction & Instantiation

We manually analyzed the questions collected by crowd-sourcing and divided them into a total of 74 templates. These templates were divided into 3 question categories. These question categories along with a few sample templates are shown below. See Table 4 for statistics of dif-

ferent question and answer types in our dataset (please refer to the Supplementary material for further details).

**Structural Understanding:** These are questions about the overall structure of the plot and do not require any quantitative reasoning. Examples: “How many different coloured bars are there?”, “How are the legend labels stacked?”.

**Data Retrieval:** These questions seek data item for a single element in the plot. Examples: “What is the number of tax payers in Myanmar in 2015?”.

**Reasoning:** These questions either require numeric reasoning over multiple plot elements or a comparative analysis of different elements of the plot, or a combination of both to answer the question. Examples: “In which country is the number of threatened bird species minimum?”, “What is the median banana production?”.

We abstracted the questions into templates such as “In how many <plural form of X\_label>, is the <Y\_label> of/in <legend\_label> greater than the average <Y\_label> of/in <legend\_label> taken over all <plural form of X\_label>?”. We could then generate multiple questions for each template by replacing X\_label, Y\_label, legend\_label, *etc.* by indicator variables, years, cities *etc.* from our curated data. However, this was a tedious task requiring a lot of manual intervention. For example, consider the indicator variable “Race of students” in Figure 1c. If we substitute this indicator variable as it is in the above template, it would result in a question, “In how many cities, is the race of the students(%) of Asian greater than the average race of the students (%) of Asian taken over all cities?”, which sounds unnatural. To avoid this, we asked in-house annotators to carefully paraphrase these indicator variables and question templates. The paraphrased version of the above example was “In how many cities, is the percentage of Asian students greater than the average percentage of Asian students taken over all cities?”. Such paraphrasing for every question template and indicator variable required significant manual effort. Using this semi-automated process we generated

| Dataset Split | Plot Types |        |        |          | Question Types |                |            | Answer Types |              |             |
|---------------|------------|--------|--------|----------|----------------|----------------|------------|--------------|--------------|-------------|
|               | vbar       | hbar   | line   | dot-line | Structural     | Data-Retrieval | Reasoning  | Yes/No       | Fixed vocab. | Open vocab. |
| Train         | 52,463     | 52,700 | 25,897 | 26,010   | 871,782        | 2,784,041      | 16,593,656 | 784,115      | 3,095,774    | 16,369,590  |
| Validation    | 11,249     | 11,292 | 5,547  | 5,571    | 186,994        | 599,573        | 3,574,081  | 167,871      | 600,424      | 3,592,353   |
| Test          | 11,242     | 11,292 | 5,549  | 5,574    | 186,763        | 596,359        | 3,559,392  | 167,727      | 667,742      | 3,507,045   |

Table 3: Detailed Statistics for different splits of the PlotQA dataset.

| Answer (A) Type  | Question (Q) Type |                |           |
|------------------|-------------------|----------------|-----------|
|                  | Structure         | Data Retrieval | Reasoning |
| Yes/No           | 36.99%            | 5.19%          | 2.05%     |
| Fixed vocabulary | 63.01%            | 18.52%         | 15.92%    |
| Open vocabulary  | 0.00%             | 76.29%         | 82.03%    |

Table 4: Overall distribution of Q and A types in PlotQA.

| Dataset Split | #Images        | #QA pairs         |
|---------------|----------------|-------------------|
| Train         | 157,070        | 20,249,479        |
| Validation    | 33,650         | 4,360,648         |
| Test          | 33,657         | 4,342,514         |
| <b>Total</b>  | <b>224,377</b> | <b>28,952,641</b> |

Table 5: PlotQA Dataset Statistics

a total of 28,952,641 questions. This approach of creating questions on real-world plot data with carefully curated question templates followed by manual paraphrasing is a key contribution of our work. The resultant PlotQA dataset is much closer to the real-world challenge of reasoning over plots, significantly improving on existing datasets. Table 2 summarizes the differences between PlotQA and these existing datasets such as FigureQA and DVQA. Note that (a) the number of unique answers in PlotQA is very large, (b) the questions in PlotQA are much longer, and (c) the vocabulary of PlotQA is more realistic than FigureQA or DVQA.

## 4. Proposed Model

Existing models for VQA are of two types: (i) read the answer from the image (as in LoRRA) or (ii) pick the answer from a fixed vocabulary (as in SAN and BAN). Such models work well for datasets such as DVQA where indeed all answers come from a fixed vocabulary (global or plot specific) but are not suited for PlotQA with a large number of OOV questions. Answering such questions involves various sub-tasks: (i) detect all the elements in the plot (bars, legend names, tick labels, etc), (ii) read the values of these elements, (iii) establish relations between the plot elements, e.g., creating tuples of the form {country=Angola, year=2006, price of diesel = 0.4 }, and (iv) reason over this structured data. Expecting a single end-to-end model to be able to do all of this is unreasonable. Hence, we propose a multi-staged pipeline to address each of the sub-tasks.

We further note that for simpler questions which do not require reasoning and can be answered from a small fixed size vocabulary, such an elaborate pipeline is an overkill. As an illustration consider the question ‘‘How many bars are there in the image?’’. This does not require reasoning and can be answered based on visual properties of the image. For such questions, we have a simpler *QA-as-classification* pipeline. As shown in Figure 3, our overall model is thus

a hybrid model containing the following elements: (i) a binary classifier for deciding whether the given question can be answered from a small fixed vocabulary or needs more complex reasoning, and (ii) a simpler *QA-as-classification* model to answer questions of the former type, and (iii) a multi-staged model containing four components as described below to deal with complex reasoning questions.

### 4.1. Visual Elements Detection (VED)

The data bearing elements of a plot are of 10 distinct classes: the title, the labels of the  $x$  and  $y$  axes, the tick labels or categories (e.g., countries) on the  $x$  and  $y$  axis, the data markers in the legend box, the legend names, and finally the bars and lines in the graph. Following existing literature ([4], [12]), we refer to these elements as the *visual elements* of the graph. The first task is to extract all these visual elements by drawing bounding boxes around them and classifying them into the appropriate class. To this end, we can either apply object detection models such as RCNN, Fast-RCNN [6], YOLO [27], SSD [21], *etc.* or instance segmentation models such as Mask-RCNN [9]. Upon comparing all methods, we found that Faster R-CNN [29] model along with Feature Pyramid Network(FPN) [20] performed the best and hence we used it as our VED module.

### 4.2. Object Character Recognition (OCR)

Some of the visual elements such as title, legends, tick labels, *etc.* contain numeric and textual data. For extracting this data from within these bounding boxes, we use a state-of-the-art OCR model [34]. We crop the detected visual element to its bounding box, convert the cropped textual image into gray-scale, resize and deskew it, and then pass it to an OCR module. Existing OCR modules perform well for machine-written English text, and indeed we found that a pre-trained OCR module<sup>2</sup> works well on our dataset.

<sup>2</sup><https://github.com/tesseract-ocr/tesseract>

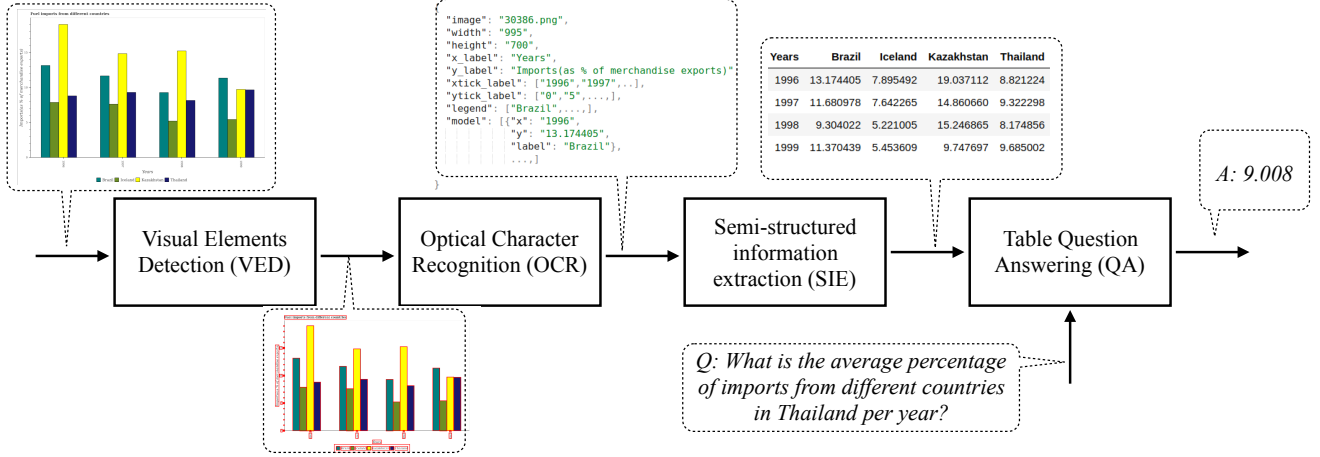


Figure 3: Our proposed multi-staged modular pipeline for QA on scientific plots.

### 4.3. Semi-Structured Information Extraction (SIE)

The next stage of extracting the data into a semi-structured table is best explained with an example shown in Figure 3. The desired output of SIE is shown in the table where the rows correspond to the ticks on the  $x$ -axis (1996, 1997, 1998, 1999), the columns correspond to the different elements listed in the legend (Brazil, Iceland, Kazakhstan, Thailand) and the  $i, j$ -th cell contains the value corresponding to the  $x$ -th tick and the  $y$ -th legend. The values of the  $x$ -tick labels and the legend names are available from the OCR module. The mapping of legend name to legend marker or color is done by associating a legend name to the marker or color whose bounding box is closest to the bounding box of the legend name. Similarly, we associate each tick label to the tick marker whose bounding box is closest to the bounding box of the tick label. For example, we associate the legend name Brazil to the color “Dark Cyan” and the tick label 1996 to the corresponding tick mark on the  $x$ -axis. With this we have the 4 row headers and 4 column headers, respectively. To fill in the 16 values in the table, there are again two smaller steps. First we associate each of the 16 bounding boxes of the 16 bars to their corresponding  $x$ -ticks and legend names. A bar is associated with an  $x$ -tick label whose bounding box is closest to the bounding box of the bar. To associate a bar to a legend name, we find the dominant color in the bounding box of the bar and match it with a legend name corresponding to that color. Second, we need to find the value represented by each bar. We extract the height of the bar using bounding box information from the VED module and then search for the  $y$ -tick labels immediately above and below that height. We then interpolate the value of the bar based on the values of these bounding ticks. With this we have the 16 values in the cells and thus have extracted all the information from the plot into a semi-

structured table. The output of each stage is discussed in the supplementary material.

### 4.4. Table Question Answering (QA)

The final stage of the pipeline is to answer questions on the semi-structured table. As this is similar to answering questions from the WikiTableQuestions dataset [26], we adopt the same methodology as proposed in [26]. In this method, the table is converted to a knowledge graph and the question is converted to a set of candidate logical forms by applying compositional semantic parsing. These logical forms are then ranked using a log-linear model and the highest ranking logical form is applied to the knowledge graph to get the answer. Note that with this approach the output is computed by a logical form that operates on the numerical data. This supports complex reasoning questions and also avoids the limitation of using a small answer vocabulary for multi-class classification as is done in existing work on VQA. There are recent neural approaches for answering questions over semi-structured tables such as [24, 8]. Individually these models do not outperform the relatively simpler model of [26], but as an ensemble they show a small improvement of only (1-2%). To the best of our knowledge, there is only one neural method [19] which outperforms [26], but the code for this model is not available which makes it hard to reproduce the results.

## 5. Experiments

### 5.1. Train-Valid-Test Splits

By using different combinations of 841 indicator variables and 160 entities (years, countries, etc), we created a total of 224,377 plots. Depending on the context and type of the plot, we instantiated the 74 templates to create mean-



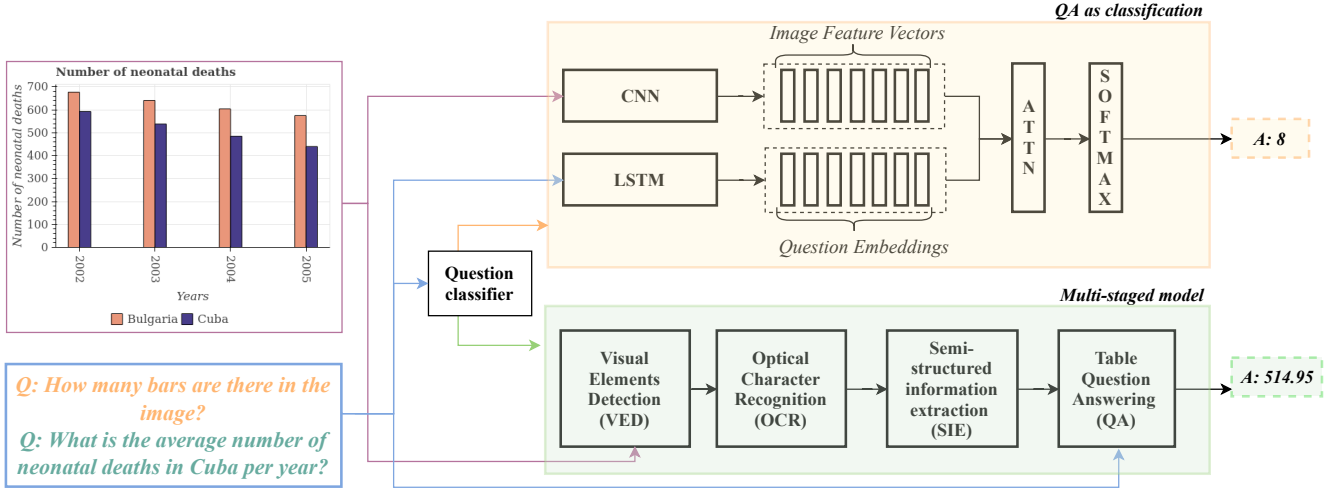


Figure 4: Our proposed model containing (i) a question classifier for deciding whether the question can be answered from a fixed vocabulary (orange) or needs more complex reasoning (green), (ii) *QA-as-classification* model to answer questions of the former type, and (iii) *multi-staged model* as a pipeline of perception and QA modules for answering complex questions.

ingful {question, answer} pairs for each of the plots. We created train (70%), valid (15%) and test (15%) splits (Table 5). The dataset and the crowd-sourced questions can be downloaded from the link: [bit.ly/PlotQA](http://bit.ly/PlotQA)

## 5.2. Models Compared

We compare the performance of the following models:

- **IMG-only:** This is a simple baseline where we just pass the image through a VGG19 and use the embedding of the image to predict the answer from a fixed vocabulary.
- **QUES-only:** This is a simple baseline where we just pass the question through a LSTM and use the embedding of the question to predict the answer from a fixed vocabulary.
- **SAN** [36]: This is an encoder-decoder model with a multi-layer stacked attention [2] mechanism. It obtains a representation for the image using a deep CNN and a representation for the query using LSTM. It then uses the query representation to locate relevant regions in the image and uses this to pick an answer from a fixed vocabulary.
- **SANDY** [12]: This is the best performing model on the DVQA dataset and is a variant of SAN. Unfortunately, the code for this model is not available and the description in the paper was not detailed enough for us to reimplement it.<sup>3</sup> Hence, we report the numbers for this model only on DVQA (from the original paper).
- **LoRRA** [33]: This is the recently proposed model on the TextVQA dataset. It concatenates the image features extracted from pre-trained ResNet-152 [10] model with the re-

gion based features extracted from Faster-RCNN [5] model. It then reads the text present in the image using a pre-trained OCR module and incorporates an attention mechanism to reason about the image and the text. Finally, it does multi-class classification where the answer either comes from a fixed vocabulary or is copied from the text in the image.

- **BAN** [16]: This model exploits bilinear interactions between two groups of input channels, *i.e.*, between every question word (GRU [3] features) and every image region (pre-trained Faster-RCNN [29] object features). It then uses low-rank bilinear pooling [17] to extract the joint distribution for each pair of channels. BAN accumulates 8 such bilinear attention maps which are then fed to a two-layer perceptron classifier to get the final joint distribution over answers from a fixed vocabulary.

- **Our Model:** This proposed model shown in Figure 4 with two model paths. The training data for the binary classification is generated by comparing the performance of the individual models: For a given question, the label is set to 1 if the performance of QA-as-classification model is better than the multi-stage pipeline, and 0 otherwise. We use an LSTM to represent the input question and then perform binary classification on this representation.

## 5.3. Training Details

**SAN:** We used an existing implementation of SAN<sup>4</sup> for the initial baseline results. Image features are extracted from the last pooling layer of VGG19 network. Question features are the last hidden state of the LSTM. Both the LSTM

<sup>3</sup>We have contacted the authors and while they are helpful in sharing various details, they do not have access to the original code now.

<sup>4</sup><https://github.com/TingAnChien/san-vqa-tensorflow>

hidden state and 512-d image feature vector at each location are transferred to a 1024-d vector by a fully connected layer, and added and passed through a non-linearity (tanh). The model was trained using Adam [18] with an initial learning rate of 0.0003 and a batch size of 128 for 25,000 iterations.

**Our model:** The binary question classifier in the proposed model contains a 50-dimensional word embedding layer followed by an LSTM with 128 hidden units. The output of the LSTM is projected to 256 dimensions and this is then fed to the output layer. The model is trained for 10 epochs using RMSProp with an initial learning rate of 0.001. Accuracy on the validation set is 87.3%. Of the 4 stages of the multi-stage pipeline, only two require training, *viz.*, Visual Elements Detection (VED) and Table Question Answering (QA). As mentioned earlier, for VED we train a variant of Faster R-CNN [20] with FPN using the bounding box annotations available in PlotQA. We trained the model with a batch size of 32 for 200,000 steps. We used RMSProp with an initial learning rate of 0.004. For Table QA, we trained the model proposed in [26] using questions from our dataset and the corresponding ground truth tables.

## 5.4. Evaluation Metric

We used accuracy as the evaluation metric. Specifically, for textual answers (such as India, CO2, etc.) the model’s output was considered to be correct only if the predicted answer exactly matches the true answer. However, for numeric answers with floating point values, an exact match is a very strict metric. We relax the measure to consider an answer to be correct as if it is within 5% of the correct answer.

## 5.5. Human Accuracy on PlotQA dataset

To assess the difficulty of the PlotQA dataset, we report human accuracy on a small subset of the Test split of the dataset. With the help of in-house annotators, we were able to evaluate 5,860 questions grounded in 160 images. Human accuracy on this subset is found to be 80.47%. We used the evaluation metric as defined in section 5.4. Most human errors were due to numerical precision as it is difficult to find the exact value from the plot even with a 5% margin.

## 6. Observations and Results

**1. Evaluating models on PlotQA dataset (Table 7):** The baselines IMG-only and QUES-only performed poorly with an accuracy of 4.84% and 5.35% respectively. Existing models (SAN, BAN, LoRRA) perform poorly on this dataset. In particular, BAN and LoRRA have an abysmal accuracy of less than 1%. This is not surprising given that both models are not designed to answer OOV questions. Further, the original VQA tasks for which BAN was proposed does not have any complex numerical reasoning questions as found in PlotQA. Similarly, LoRRA was designed only for text based answers and not for questions

requiring numeric reasoning. Note that we have used the original code [32] released by the authors of these models. Given the specific focus and limited capabilities of these existing models it may even seem unfair to evaluate these models on our dataset but we still do so for the sake of completeness and to highlight the need for better models. Lastly, our model gives the best performance of 22.52% on the PlotQA dataset.

**Ablation Study of proposed method:** Table 8 presents the details of the ablation study of the proposed method for each question type (structural, data retrieval, reasoning) and each answer type (binary, fixed vocabulary, OOV). QA-as-classification performs very well on Yes/No questions and moderately well on Fixed vocab. questions with a good baseline aggregate accuracy of 7.76%. It performs poorly on Open vocab. question, failing to answer almost all the 3,507,045 questions in this category. On the other hand, the QA-as-multi-stage pipeline fails to answer correctly any of the Yes/No questions, performs moderately well on Fixed vocab. questions, and answers correctly some of the hard Open vocab. questions. Our model combines the complementary strengths of QA-as-classification and QA-as-multi-stage pipeline achieving the highest accuracy of 22.52%. In particular, the performance improves significantly for all Fixed Vocab. questions, while retaining the high accuracy of QA-as-classification on Yes/No questions and QA-as-multi-stage pipeline’s performance on Open vocab. We acknowledge that the accuracy is significantly lower than human performance. This establishes that the dataset is challenging and raises open questions on models for visual reasoning.

**2. Analysis of the pipeline** We analyze the performance of VED, OCR and SIE modules in the pipeline.

**VED:** Table 9 shows that the VED module performs reasonably well at an Intersection Over Union (IOU) of 0.5. For higher IOUs of 0.75 and 0.9, the accuracy falls drastically. For instance, at IOU of 0.9, dotlines are detected with an accuracy of under 20%. Clearly, such inaccuracies would lead to incorrect table generation and subsequent QA. This brings out an interesting difference between this task and other instance segmentation tasks where the margin of error is higher (where IOU of 0.5 is accepted). A small error in visual element detection as indicated by mAP scores of 75% is considered negligible for VQA tasks, however for PlotQA small errors can cause significantly misaligned table generation and subsequent QA. We illustrate this with an example given in Figure 5. The predicted red box having an IOU of 0.58 estimates the bar size as 760 as opposed to ground truth of 680, significantly impacting downstream QA accuracy.

**OCR:** We evaluate the OCR module in standalone/oracle mode and pipeline mode in Table 10. In the oracle mode, we feed ground truth boxes to the OCR model whereas in



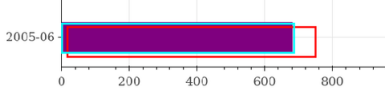


Figure 5: Ground-truth (cyan) and predicted (red) boxes.

| Model     | DVQA (TEST)   | DVQA (TEST-NOVEL) |
|-----------|---------------|-------------------|
| SAN       | 32.1%         | 30.98%            |
| SANDY-OCR | 45.77%        | 45.81%            |
| Our Model | <b>57.99%</b> | <b>59.54%</b>     |

Table 6: Accuracy of different models on DVQA dataset.

| Models   | IMG  | QUES | BAN  | LoRRA | SAN  | Our Model    |
|----------|------|------|------|-------|------|--------------|
| Accuracy | 4.84 | 5.35 | 0.01 | 0.02  | 7.76 | <b>22.52</b> |

Table 7: Accuracy (in %) of different models on PlotQA.

the pipeline model we perform OCR on the output of the VED module. We observe only a small drop in performance from 97.06% (oracle) to 93.10% (after VED), which indicates that the OCR module is robust to the reduction in VED module’s accuracy at higher IOU as it does not depend on the class label or the exact position of bounding boxes.

**SIE:** We now evaluate the performance of the SIE module. We consider each cell in the table to be a tuple of the form {row header, column header, value } (e.g., {Poland, 1964, 10000 tractors}). We consider all the tuples extracted by the SIE module with the tuples present in the ground truth table to compute the F1-score. Even though Table 9 suggests that the VED model is very accurate with a mAP@0.5 of 96.43%, we observe that the F1-score for table extraction is only 0.68. This indicates that many values are not being extracted accurately due to the kind of errors shown in Figure 5 where the bounding box has a high overlap with the true box. We thus need better plot VED modules which can predict tighter bounding boxes (higher mAP at IOU of 0.9) around the plot’s visual and textual elements. Inaccurate VED module leads to erroneous tables which further affects the downstream QA accuracy.

- In summary, a highly accurate VED for structured images is an open challenge to improve reasoning over plots.

**3. Evaluating new models on the existing DVQA dataset (Table 6):** The proposed model performs better than the existing models (SAN and SANDY-OCR) establishing a new SOTA result on DVQA. The higher performance of the proposed hybrid model in comparison to SAN (in contrast to the PlotQA results) suggests that the extraction of the structured table is more accurate on the DVQA dataset. This is because of the limited variability in the axis and tick labels and shorter length (one word only) of labels.

| Accuracy (in %)             |                 |              |                |              |
|-----------------------------|-----------------|--------------|----------------|--------------|
| Model (Agg. acc.)           | Q Type \ A type | Structural   | Data Retrieval | Reasoning    |
| Human                       | Yes/No          | 99.77        | 100            | 76.51        |
| Baseline (80.47)            | Fixed vocab.    | 99.29        | 83.31          | 59.97        |
|                             | Open vocab.     | NA           | 87.58          | 58.01        |
| QA as classification (7.76) | Yes/No          | 91.12        | 97.32          | 62.75        |
|                             | Fixed vocab.    | 66.85        | 30.76          | 16.03        |
|                             | Open vocab.     | NA           | 0.00           | 0.00         |
| Multistage                  | Yes/No          | 0.00         | 0.00           | 0.00         |
| Pipeline (18.46)            | Fixed vocab.    | 42.12        | 16.07          | 7.24         |
|                             | Open vocab.     | NA           | 57.39          | 14.95        |
| Our Model (22.52)           | Yes/No          | <b>91.12</b> | <b>97.32</b>   | <b>62.75</b> |
|                             | Fixed vocab.    | <b>66.86</b> | <b>22.64</b>   | <b>7.95</b>  |
|                             | Open vocab.     | NA           | <b>57.39</b>   | <b>14.95</b> |

Table 8: Ablation study of proposed method on PlotQA. Note that there are no open vocab. answers for *Structural Understanding* question templates (see Table 1).

| Class          | AP@0.5        | AP@0.75       | AP@0.9        |
|----------------|---------------|---------------|---------------|
| Title          | 100.00%       | 78.83%        | 0.22%         |
| Bar            | 95.84%        | 94.30%        | 85.54%        |
| Line           | 72.25%        | 62.04%        | 37.65%        |
| Dotline        | 96.30%        | 95.14%        | 18.07%        |
| X-axis Label   | 99.99%        | 99.99%        | 99.09%        |
| Y-axis Label   | 99.90%        | 99.90%        | 99.46%        |
| X-tick Label   | 99.92%        | 99.74%        | 96.04%        |
| Y-tick Label   | 99.99%        | 99.97%        | 96.80%        |
| Legend Label   | 99.99%        | 99.96%        | 93.68%        |
| Legend Preview | 99.95%        | 99.94%        | 96.30%        |
| <b>mAP</b>     | <b>96.43%</b> | <b>92.98%</b> | <b>72.29%</b> |

Table 9: VED Module’s Accuracy on PlotQA dataset

|              | Oracle        | After VED     |
|--------------|---------------|---------------|
| Title        | 99.31%        | 94.6%         |
| X-axis Label | 99.94%        | 95.5%         |
| Y-axis Label | 98.43%        | 97.07%        |
| X-tick Label | 94.8%         | 91.38%        |
| Y-tick Label | 93.38%        | 88.07%        |
| Legend Label | 98.53%        | 91.99%        |
| <b>Total</b> | <b>97.06%</b> | <b>93.10%</b> |

Table 10: OCR Module Accuracy on the PlotQA dataset.

## 7. Conclusion

We introduce the PlotQA dataset to reduce the gap between existing synthetic plot datasets and real-world plots and question templates. Analysis of existing VQA models on PlotQA reveals that they perform poorly for Open Vocabulary questions. This is not surprising as these models were not designed to handle complex questions which require numeric reasoning and OOV answers. We propose a hybrid model with separate pipelines for handling (i) simpler questions which can be answered from a fixed vocabulary and (ii) complex questions with OOV answers. For

OOV questions, we propose a pipelined approach that combines visual element detection and OCR with QA over tables. The proposed model gives state-of-the-art results on both the DVQA and PlotQA datasets. Further analysis of our pipeline reveals the need for more accurate visual element detection to improve reasoning over plots.

## References

- [1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: visual question answering. In *ICCV*, 2015.
- [2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.
- [3] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734, 2014.
- [4] M. Cliche, D. S. Rosenberg, D. Madeka, and C. Yee. Scatteract: Automated extraction of data from scatter plots. In *ECML PKDD*, 2017.
- [5] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He. Detectron. <https://github.com/facebookresearch/detectron>, 2018.
- [6] R. B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015.
- [7] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.
- [8] T. Haug, O. Ganea, and P. Grnarova. Neural multi-step reasoning for question answering on semi-structured tables. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, pages 611–617, 2018.
- [9] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988, 2017.
- [10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.
- [11] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.
- [12] K. Kafle, S. Cohen, B. L. Price, and C. Kanan. DVQA: understanding data visualizations via question answering. *CoRR*, abs/1801.08163, 2018.
- [13] S. E. Kahou, A. Atkinson, V. Michalski, Á. Kádár, A. Trischler, and Y. Bengio. Figureqa: An annotated figure dataset for visual reasoning. *CoRR*, abs/1710.07300, 2017.
- [14] A. Kembhavi, M. Salvato, E. Kolve, M. J. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.
- [15] A. Kembhavi, M. J. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*, 2017.
- [16] J. Kim, J. Jun, and B. Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 1571–1581, 2018.
- [17] J. Kim, K. W. On, W. Lim, J. Kim, J. Ha, and B. Zhang. Hadamard product for low-rank bilinear pooling. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.
- [18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
- [19] J. Krishnamurthy, P. Dasigi, and M. Gardner. Neural semantic parsing with type constraints for semi-structured tables. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1516–1526, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.
- [20] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 936–944, 2017.
- [21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector. In *ECCV*, 2016.
- [22] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016.
- [23] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014.
- [24] A. Neelakantan, Q. V. Le, M. Abadi, A. McCallum, and D. Amodei. Learning a natural language interface with neural programmer. *CoRR*, abs/1611.08945, 2016.
- [25] H. Noh and B. Han. Training recurrent answering units with joint loss minimization for VQA. *CoRR*, abs/1606.03647, 2016.
- [26] P. Pasupat and P. Liang. Compositional semantic parsing on semi-structured tables. In *ACL*, 2015.
- [27] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6517–6525, 2017.
- [28] M. Ren, R. Kiros, and R. S. Zemel. Image question answering: A visual semantic embedding model and a new dataset. *CoRR*, abs/1505.02074, 2015.
- [29] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Sys-*

*tems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015.

- [30] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *NIPS*.
- [31] N. Siegel, Z. Horvitz, R. Levin, S. K. Divvala, and A. Farhadi. Figureseer: Parsing result-figures in research papers. In *ECCV*, 2016.
- [32] A. Singh, V. Natarajan, Y. Jiang, X. Chen, M. Shah, M. Rohrbach, D. Batra, and D. Parikh. Pythia-a platform for vision & language research. In *SysML Workshop, NeurIPS*, volume 2018, 2018.
- [33] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [34] R. Smith. An overview of the tesseract ocr engine. In *ICDAR*, 2007.
- [35] A. Suhr, M. Lewis, J. Yeh, and Y. Artzi. A corpus of natural language for visual reasoning. In *ACL*, 2017.
- [36] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola. Stacked attention networks for image question answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 21–29, 2016.
- [37] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

## Appendices

The supplementary material is organised in the following manner: Section 8 describes the methodology of knowledge graph creation from structured data. In Section 9, we further analyse our proposed pipeline and discuss the errors in each stage. In Section 10, we provide sample plots from the PlotQA dataset. In Section 11, we list all the 74 question templates that were formulated from the crowd sourced questions.

### 8. Construction of knowledge graph from structured data

Following [26], we convert the semi-structured table into a knowledge graph which has two types of nodes *viz.* row nodes and entity nodes. The rows of the table become row nodes, whereas the cells of each row become the entity nodes in the graph. Directed edges exist from the row nodes to the entity nodes of that column and the corresponding table column header act as edge-labels. An example of knowledge graph of the semi-structured table given in Figure 6a is shown in Figure 6b. For reasoning on the knowledge graph, we adopted the same methodology as given in [26]. The questions are converted to a set of candidate logical forms by applying compositional semantic parsing. Each of these logical forms is then ranked using a log-linear model and the highest ranking logical form is applied to the knowledge graph to get the final answer.

| Years | Brazil    | Iceland  | Kazakhstan | Thailand |
|-------|-----------|----------|------------|----------|
| 1996  | 13.174405 | 7.895492 | 19.037112  | 8.821224 |
| 1997  | 11.680978 | 7.642265 | 14.860660  | 9.322298 |
| 1998  | 9.304022  | 5.221005 | 15.246865  | 8.174856 |
| 1999  | 11.370439 | 5.453609 | 9.747697   | 9.685002 |

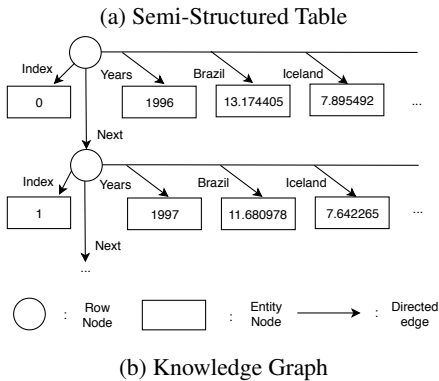


Figure 6: An example of the knowledge graph constructed from the semi-structured table.

### 9. Some failure cases

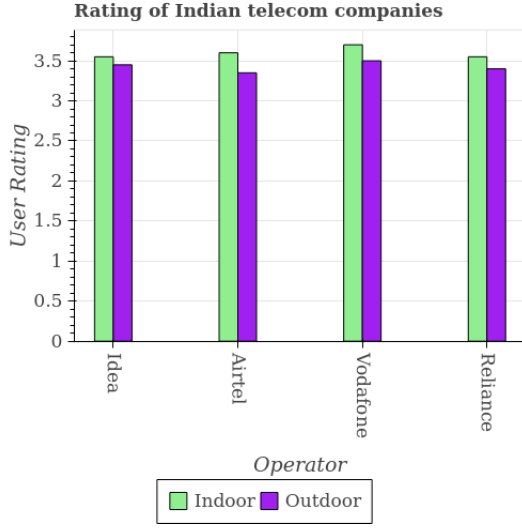
In this section, we visualize the output of each stage in our multistage pipeline and show some interesting failure cases with the help of an example.

- **VED Stage:** Although the bounding boxes predicted by Faster R-CNN coupled with Feature Pyramid Network (FPN) fit reasonably well at an IOU of 0.5, it is not acceptable as the values extracted from these bounding boxes will lead to incorrect table generation and subsequent QA. Example: In Figure 7b, consider the bar representing the “Indoor User Rating” value for “Vodafone”. The overlap between the ground-truth box (blue) and the predicted box (red) is higher than 0.5 but the values extracted from the detected box is 4.0 as opposed to the actual value which is 3.73. Another interesting failure case is shown in Figure 9b. There are multiple overlapping data points and the model is able to detect only one of the points. This leads to incomplete table generation as shown in Figure 10b where the values for Liberia for the years 2008, 2009 and 2010 could not be extracted. This small error might be acceptable for other VQA tasks but for PlotQA these small errors will escalate to multiple incorrect answers.

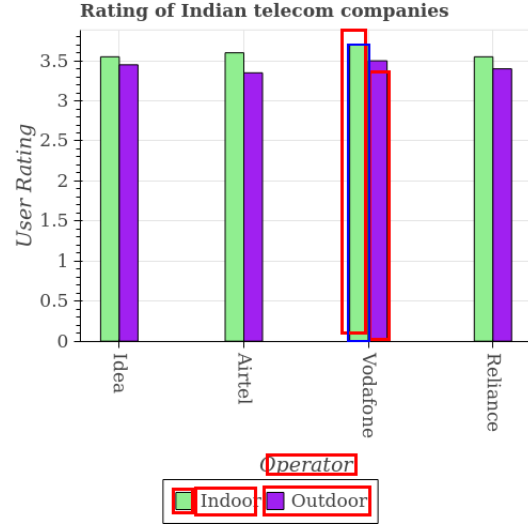
- **OCR stage:** A slight misalignment in the bounding boxes predicted by VED module causes significant errors while extracting the text. Example: In Figure 7b, consider the box enclosing the legend label “Indoor”. The rightmost edge of the predicted bounding box is drawn over the letter “r”, which makes the OCR module incorrectly recognize the text as “Indoo”. A similar error is made while performing OCR on the X-axis title which is read as “Dperator” instead of “Operator”. Consider another example where there are misaligned bounding boxes on axes tick-labels as shown in Figure 9b. The values extracted are 200B, -2009 and -100 as opposed to the ground-truth values 2008, 2009 and 100. This slight error leads to incorrect column name in the subsequent generated tables (Figure 8b and Figure 10b) and incorrect answers to all the questions pertaining to these labels as shown in Table 11 and Table 12.

- **SIE stage:** Figure 8a shows the oracle table which is generated by using the ground-truth annotations and Figure 8b shows the table generated after passing the plot image through the different stages of our proposed multistage pipeline. It is evident from the generated table that the errors propagated from the VED and the OCR stage has lead to an incorrect table generation.

- **QA stage:** In Table 11 and Table 12 we compare the answer predictions made by different models with the ground-truth answer on randomly sampled questions. Note that, our proposed model combines the complementary strengths of both, QA-as-classification and QA as multistage pipeline, models.



(a) Input plot image



(b) Few examples of the predicted bounding boxes

Figure 7: Errors made by the VED stage (highlighted in red).

| Operator | Indoor | Outdoor |
|----------|--------|---------|
| Idea     | 3.55   | 3.45    |
| Airtel   | 3.62   | 3.38    |
| Vodafone | 3.73   | 3.51    |
| Reliance | 3.57   | 3.41    |

(a) Oracle table generated using ground-truth annotations

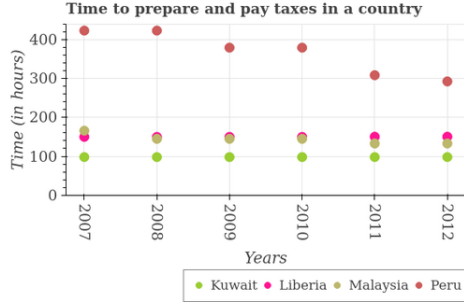
| Dperator | Indoo | Outdoor |
|----------|-------|---------|
| Idea     | 3.53  | 3.45    |
| Airtel   | 3.60  | 3.40    |
| Vodafone | 4.0   | 3.35    |
| Reliance | 3.62  | 3.40    |

(b) Generated Semi-Structured Table

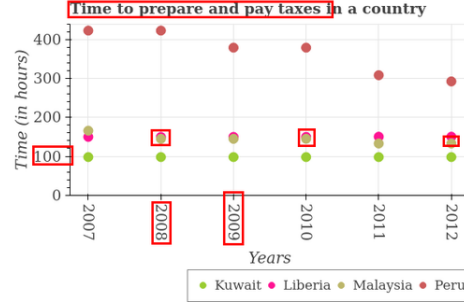
Figure 8: Errors made by the OCR and SIE stage (highlighted in red). Note that most of these errors have been propagated from the VED stage.

| Question   | Ground Truth | QA as classification | Multistage Pipeline | Our Model |
|--|--------------|----------------------|---------------------|-----------|
| Q1. What is the average indoor user rating per operator?   | 3.62         | 500                  | 3.54                | 3.54      |
| Q2. What is the total user rating for Vodafone in the graph?   | 7.24         | 5                    | 7.35                | 7.35      |
| Q3. What is the label or title of the X-axis?  | Operator     | Years                | Dperator            | Dperator  |
| Q4. What is the indoor user rating of Airtel?  | 3.62         | 0                    | 3.60                | 3.60      |
| Q5. How many groups of bars are there?   | 4            | 4                    | 4                   | 4         |
| Q6. What is the ratio of indoor user rating of Airtel to that of the outdoor user rating of Vodafone?                          | 1.03         | No                   | 3.35                | 3.35      |
| Q7. What is the difference between the highest and lowest outdoor user rating?   | 0.13         | 1.5                  | 0.10                | 0.10      |
| Q8. Does "Indoor" appear as one of the legend-labels in the graph?   | Yes          | Yes                  | 1.0                 | Yes       |
| Q9. For how many operators are the indoor user rating greater than the average outdoor user rating taken over all operators?   | 4            | 4                    | 3.4                 | 4         |
| Q10. Is the sum of outdoor user rating in Reliance and Idea greater than the maximum outdoor user rating across all operators? | Yes          | Yes                  | 6.85                | Yes       |

Table 11: Answers predicted by different models on the sample questions.



(a) Input plot image



(b) Few examples of the predicted bounding boxes

Figure 9: Errors made by the VED stage (highlighted in red).

| Years | Kuwait | Liberia | Malaysia | Peru |
|-------|--------|---------|----------|------|
| 2007  | 98     | 150     | 166      | 424  |
| 2008  | 98     | 150     | 145      | 424  |
| 2009  | 98     | 150     | 145      | 380  |
| 2010  | 98     | 150     | 145      | 380  |
| 2011  | 98     | 150.5   | 133      | 309  |
| 2012  | 98     | 150.5   | 133      | 293  |

(a) Oracle table generated using ground-truth annotations

| Years | Kuwait | Liberia | Malaysia | Peru |
|-------|--------|---------|----------|------|
| 2007  | 100-   | 150     | 165      | 420  |
| 2008  | 100-   | -       | 155      | 420  |
| 2009  | 100-   | -       | 155      | 380  |
| 2010  | 100-   | -       | 153      | 383  |
| 2011  | 100-   | 155     | -        | 310  |
| 2012  | 100-   | 155     | -        | 295  |

(b) Generated Semi-Structured Table

Figure 10: Errors made by the OCR and SIE stage (highlighted in red). Note that most of these errors have been propagated from the VED stage.

| Question  | Ground Truth | QA as classification | Multistage Pipeline | Our Model  |
|---|--------------|----------------------|---------------------|------------|
| <b>Q1.</b> How are the legend-labels stacked?   | horizontal   | horizontal           | horizontal          | horizontal |
| <b>Q2.</b> What is the time (in hours) required to prepare and pay taxes in Kuwait in 2008?   | 98           | 100                  | 100-                | 100-       |
| <b>Q3.</b> What is the difference between in time (in hours) required to prepare and pay taxes in Kuwait in 2009 and the time (in hours) required to prepare and pay taxes in Liberia in 2008?                                      | -52          | -0.5                 | 100-                | 100-       |
| <b>Q4.</b> What is the difference between the highest and the lowest time (in hours) required to prepare and pay taxes in Peru?   | 131          | 500                  | 125                 | 125        |
| <b>Q5.</b> Is it the case that every year the sum of time (in hours) required to prepare and pay taxes in Peru and Kuwait is greater than the sum of the time (in hours) required to prepare and pay taxes in Liberia and Malaysia? | Yes          | Yes                  | 320                 | Yes        |

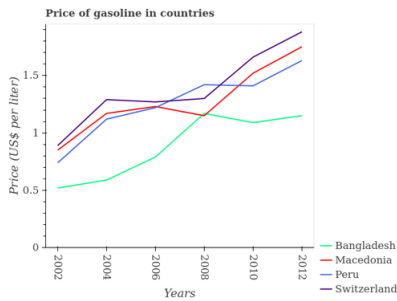
Table 12: Answers predicted by different models on the sample questions.

## 10. Samples from the PlotQA dataset

Few examples of the {plot, question, answer} triplets from the PlotQA dataset are shown in Figure 12. For each of the plots, most of the question templates discussed in section 11 are applicable but depending on the context of the plot, their language varies from it's surface form.

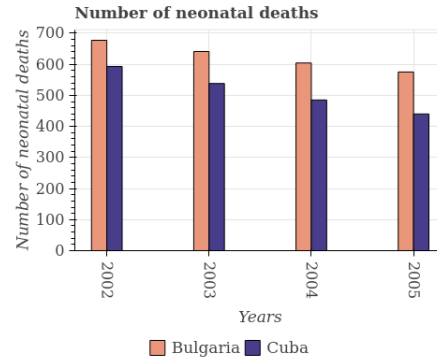
## 11. Question Templates

In this section, we present the 74 question templates which we have used for the question generation. Note that, not all question templates are applicable to each and every type of plot. Also depending on the context of the plot, the question varies from the template's surface form.



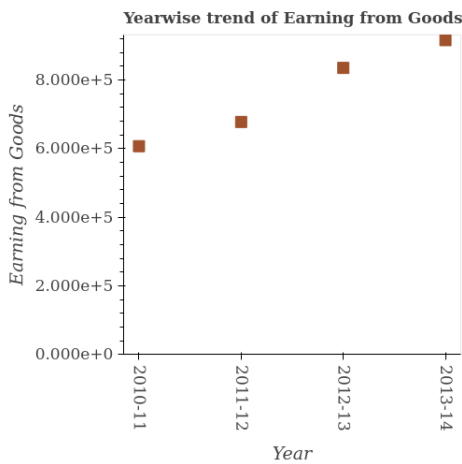
**Q1:** What is the difference between price of gasoline in Switzerland and price of gasoline in Macedonia in 2008?

A: 0.15



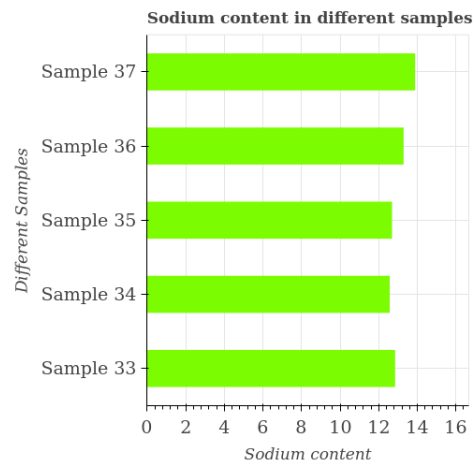
**Q2:** In how many years is the number of neonatal deaths in Cuba greater than 500?

A: 2



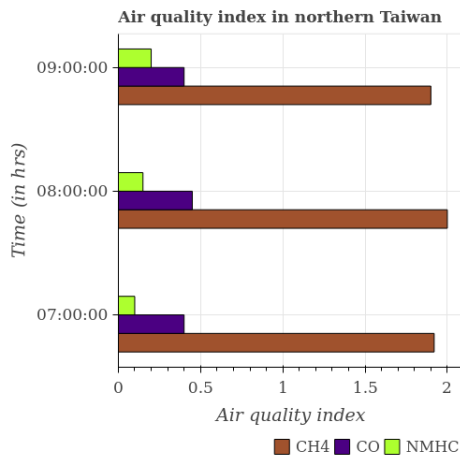
**Q3:** What is the difference between the highest and the second highest amount of earnings from goods?

A:  $0.9e + 5$



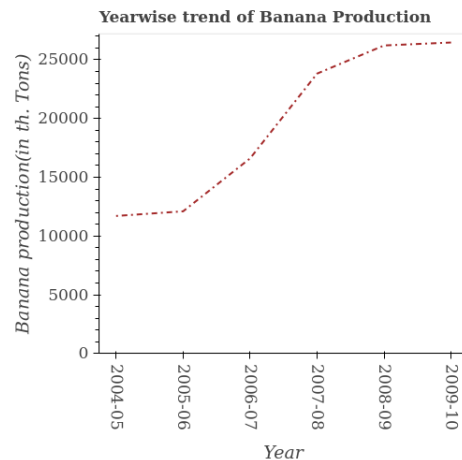
**Q4:** What is the ratio of the sodium content in Sample 37 to that in Sample 33?

A: 1.086



**Q5:** What is the average air quality index value for NHMC per hour?

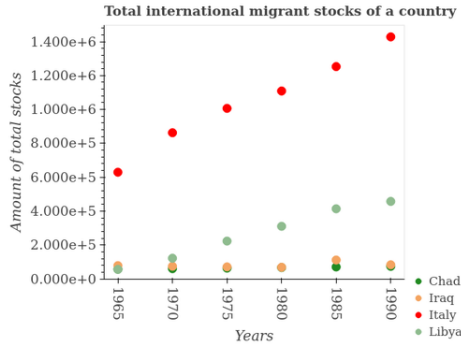
A: 0.167



**Q6:** Does the amount of banana production monotonically increase over the years?

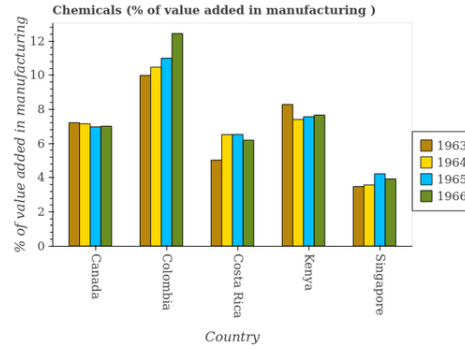
A: Yes





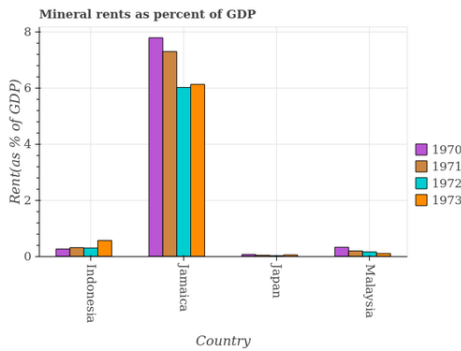
**Q7:** What is the difference between two consecutive major ticks on the Y-axis?

**A:** 2.000e+5



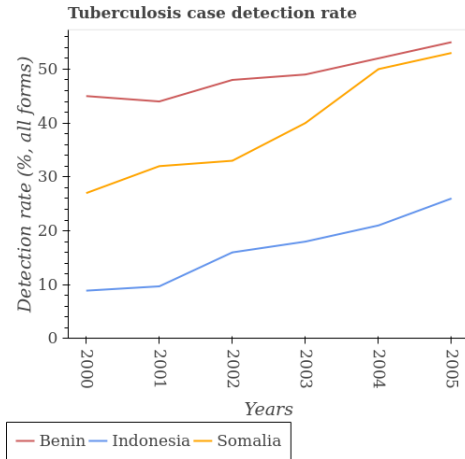
**Q8:** In how many cases, is the number of bars for a given year not equal to the number of legend labels?

**A:** 0



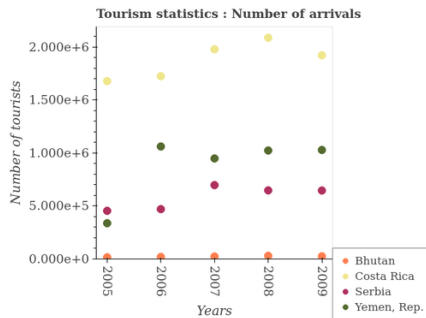
**Q9:** In how many countries, is the mineral rent (as % of GDP) in 1970 greater than the average mineral rent (as % of GDP) in 1970 taken over all countries?

**A:** 1



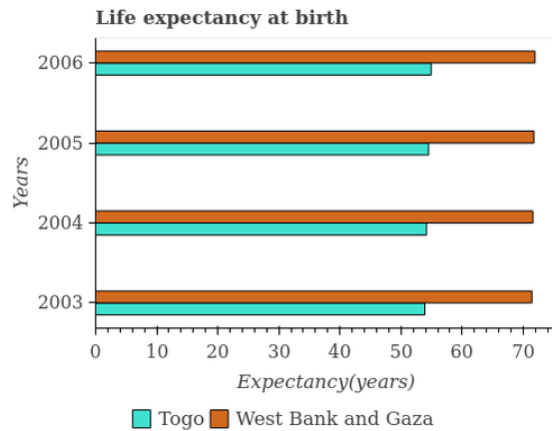
**Q10:** What is the total tuberculosis detection rate in Indonesia?

**A:** 101



**Q11:** Is it the case that in every year, the sum of the number of tourists in Costa Rica and Serbia greater than the number of tourists in Bhutan?

**A:** Yes



**Q12:** How many bars are there on the 2<sup>nd</sup> tick from the top?

**A:** 2

Figure 12: Sample {plot, question, answer} triplet present in the PlotQA dataset.

## 1. Structural Understanding :

1. Does the graph contain any zero values?

2. Does the graph contain grids ?

3. Where does the legend appear in the graph ?
4. How many legend labels are there?
5. How are the legend labels stacked?
6. How many <plural form of X\_label> are there in the graph?
7. How many <figure-type>s are there?
8. How many different colored <figure-type>s are there?
9. How many groups of <figure-type>s are there?
10. Are the number of bars on each tick equal to the number of legend labels?
11. Are the number of bars in each group equal?
12. How many bars are there on the  $i^{th}$  tick from the left?
13. How many bars are there on the  $i^{th}$  tick from the right?
14. How many bars are there on the  $i^{th}$  tick from the top?
15. How many bars are there on the  $i^{th}$  tick from the bottom?
16. Are all the bars in the graph horizontal?
17. How many lines intersect with each other?
18. Is the number of lines equal to the number of legend labels?

## 2. Data Retrieval :

1. What does the  $i^{th}$  bar from the left in each group represent?
2. What does the  $i^{th}$  bar from the right in each group represent?
3. What does the  $i^{th}$  bar from the top in each group represent?
4. What does the  $i^{th}$  bar from the bottom in each group represent?
5. What is the label of the  $j^{th}$  group of bars from the left?
6. What is the label of the  $j^{th}$  group of bars from the top?
7. Does the <Y\_label> of/in <legend-label> monotonically increase over the <plural form of X\_label> ?
8. What is the difference between two consecutive major ticks on the Y-axis ?
9. Are the values on the major ticks of Y-axis written in scientific E-notation ?
10. What is the title of the graph ?

11. Does <legend\_label> appear as one of the legend labels in the graph ?
12. What is the label or title of the X-axis ?
13. What is the label or title of the Y-axis ?
14. In how many cases, is the number of <figure\_type> for a given <X\_label> not equal to the number of legend labels ?
15. What is the <Y\_value> in/of < $i^{th}$  X\_tick> ?
16. What is the <Y\_value> of the  $i^{th}$  <legend\_label> in < $i^{th}$  X\_tick> ?
17. Does the <Y\_label> monotonically increase over the <plural form of X\_label> ?
18. Is the <Y\_label> of/in <legend\_label1> strictly greater than the <Y\_label> of/in <legend\_label2> over the <plural form of X\_label> ?
19. Is the <Y\_label> of/in <legend\_label1> strictly less than the <Y\_label> of/in <legend\_label2> over the <plural form of X\_label> ?

## 3. Reasoning :

1. Across all <plural form of X\_label>, what is the maximum <Y\_label> ?
2. Across all <plural form of X\_label>, what is the minimum <Y\_label> ?
3. In which <X\_label> was the <Y\_label> maximum ?
4. In which <X\_label> was the <Y\_label> minimum ?
5. Across all <plural form of X\_label>, what is the maximum <Y\_label> of/in <legend\_label> ?
6. Across all <plural form of X\_label>, what is the minimum <Y\_label> of/in <legend\_label> ?
7. In which <singular form of X\_label> was the <Y\_label> of/in <legend\_label> maximum ?
8. In which <singular form of X\_label> was the <Y\_label> of/in <legend\_label> minimum ?
9. What is the sum of <title> ?
10. What is the difference between the <Y\_label> in < $i^{th}$  x\_tick> and < $j^{th}$  x\_tick> ?
11. What is the average <Y\_label> per <singular form of X\_label> ?
12. What is the median <Y\_label> ?
13. What is the total <Y\_label> of/in <legend\_label> in the graph?
14. What is the difference between the <Y\_label> of/in <legend\_label> in < $i^{th}$  x\_tick> and that in < $j^{th}$  x\_tick> ?

15. What is the difference between the  $\langle Y\_label \rangle$  of/in  $\langle legend\_label1 \rangle$  in  $\langle i^{th}x\_tick \rangle$  and the  $\langle Y\_label \rangle$  of/in  $\langle legend\_label2 \rangle$  in  $\langle j^{th}x\_tick \rangle$  ?
16. What is the average  $\langle Y\_label \rangle$  of/in  $\langle legend\_label \rangle$  per  $\langle singular\ form\ of\ X\_label \rangle$  ?
17. In the year  $\langle i^{th}x\_tick \rangle$ , what is the difference between the  $\langle Y\_label \rangle$  of/in  $\langle legend\_label1 \rangle$  and  $\langle Y\_label \rangle$  of/in  $\langle legend\_label2 \rangle$  ?
18. What is the difference between the  $\langle Y\_label \rangle$  of/in  $\langle legend\_label1 \rangle$  and  $\langle Y\_label \rangle$  of/in  $\langle legend\_label2 \rangle$  in  $\langle i^{th}x\_tick \rangle$  ?
19. In how many  $\langle plural\ form\ of\ X\_label \rangle$ , is the  $\langle Y\_label \rangle$  greater than  $\langle N \rangle$  units ?
20. Do a majority of the  $\langle plural\ form\ of\ X\_label \rangle$  between  $\langle i^{th}x\_tick \rangle$  and  $\langle j^{th}x\_tick \rangle$  (inclusive/exclusive) have  $\langle Y\_label \rangle$  greater than  $\langle N \rangle$  units ?
21. What is the ratio of the  $\langle Y\_label \rangle$  in  $\langle i^{th}x\_tick \rangle$  to that in  $\langle j^{th}x\_tick \rangle$  ?
22. Is the  $\langle Y\_label \rangle$  in  $\langle i^{th}x\_tick \rangle$  less than that in  $\langle j^{th}x\_tick \rangle$  ?
23. In how many  $\langle plural\ form\ of\ X\_label \rangle$ , is the  $\langle Y\_label \rangle$  of/in  $\langle legend\_label \rangle$  greater than  $\langle N \rangle$  units ?
24. What is the ratio of the  $\langle Y\_label \rangle$  of/in  $\langle legend\_label1 \rangle$  in  $\langle i^{th}x\_tick \rangle$  to that in  $\langle j^{th}x\_tick \rangle$  ?
25. Is the  $\langle Y\_label \rangle$  of/in  $\langle legend\_label \rangle$  in  $\langle i^{th}x\_tick \rangle$  less than that in  $\langle j^{th}x\_tick \rangle$  ?
26. Is the difference between the  $\langle Y\_label \rangle$  in  $\langle i^{th}x\_tick \rangle$  and  $\langle j^{th}x\_tick \rangle$  greater than the difference between any two  $\langle plural\ form\ of\ X\_label \rangle$  ?
27. What is the difference between the highest and the second highest  $\langle Y\_label \rangle$  ?
28. Is the sum of the  $\langle Y\_label \rangle$  in  $\langle i^{th}x\_tick \rangle$  and  $\langle (i + 1)^{th}x\_tick \rangle$  greater than the maximum  $\langle Y\_label \rangle$  across all  $\langle plural\ form\ of\ X\_label \rangle$  ?
29. What is the difference between the highest and the lowest  $\langle Y\_label \rangle$  ?
30. In how many  $\langle plural\ form\ of\ X\_label \rangle$ , is the  $\langle Y\_label \rangle$  greater than the average  $\langle Y\_label \rangle$  taken over all  $\langle plural\ form\ of\ X\_label \rangle$  ?
31. Is the difference between the  $\langle Y\_label \rangle$  of/in  $\langle legend\_label1 \rangle$  in  $\langle i^{th}x\_tick \rangle$  and  $\langle j^{th}x\_tick \rangle$  greater than the difference between the  $\langle Y\_label \rangle$  of/in  $\langle legend\_label2 \rangle$  in  $\langle i^{th}x\_tick \rangle$  and  $\langle j^{th}x\_tick \rangle$  ?
32. What is the difference between the highest and the second highest  $\langle Y\_label \rangle$  of/in  $\langle legend\_label \rangle$  ?
33. What is the difference between the highest and the lowest  $\langle Y\_label \rangle$  of/in  $\langle legend\_label \rangle$  ?
34. In how many  $\langle plural\ form\ of\ X\_label \rangle$ , is the  $\langle Y\_label \rangle$  of/in  $\langle legend\_label \rangle$  greater than the average  $\langle Y\_label \rangle$  of/in  $\langle legend\_label \rangle$  taken over all  $\langle plural\ form\ of\ X\_label \rangle$  ?
35. Is it the case that in every  $\langle singular\ form\ of\ X\_label \rangle$ , the sum of the  $\langle Y\_label \rangle$  of/in  $\langle legend\_label1 \rangle$  and  $\langle legend\_label2 \rangle$  is greater than the  $\langle Y\_label \rangle$  of/in  $\langle legend\_label3 \rangle$  ?
36. Is the sum of the  $\langle Y\_label \rangle$  of/in  $\langle legend\_label1 \rangle$  in  $\langle i^{th}x\_tick \rangle$  and  $\langle j^{th}x\_tick \rangle$  greater than the maximum  $\langle Y\_label \rangle$  of/in  $\langle legend\_label2 \rangle$  across all  $\langle plural\ form\ of\ X\_label \rangle$  ?
37. Is it the case that in every  $\langle singular\ form\ of\ X\_label \rangle$ , the sum of the  $\langle Y\_label \rangle$  of/in  $\langle legend\_label1 \rangle$  and  $\langle legend\_label2 \rangle$  is greater than the sum of  $\langle Y\_label \rangle$  of  $\langle legend\_label3 \rangle$  and  $\langle Y\_label \rangle$  of  $\langle legend\_label4 \rangle$  ?