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

Appendices

The supplementary material is organised as follows: Section 1 presents the detailed statistics of the PlotQA dataset. Section 2 describes the methodology of knowledge graph creation from structured data. In Section 3, we provide sample plots from the PlotQA dataset. In Section 4, we list all the 74 question templates that were formulated from the crowd sourced questions. In Section 5, we further analyze our proposed pipeline and discuss some of the errors that occur in each stage.

1. PlotQA Statistics

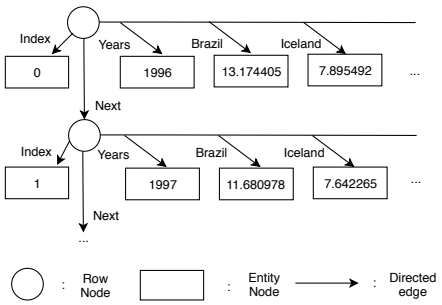
In this section, we present the detailed statistics of the PlotQA dataset. In particular, we report the dataset distribution according to the different plot types (vertical bar (vbar), horizontal bar (hbar), line and dot-line) and different question templates (structural understanding, data-retrieval, and reasoning) present in the PlotQA dataset. Table 1 presents the question distribution, categorized by their template type and detailed statistics of the different plots present in different dataset splits.

2. Construction of knowledge graph from structured data

Following [1], 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 1a is shown in Figure 1b. For reasoning on the knowledge graph, we adopted the same methodology as given in [1]. 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

(a) Semi-Structured Table



(b) Knowledge Graph

Figure 1: An example of the knowledge graph constructed from the semi-structured table.

3. Samples from the PlotQA dataset

Few examples of the {plot, question, answer} triplets from the PlotQA dataset are shown in Figure 3. For each of the plots, most of the question templates discussed in section 4 are applicable but depending on the context of the plot, their language varies from it's surface form.

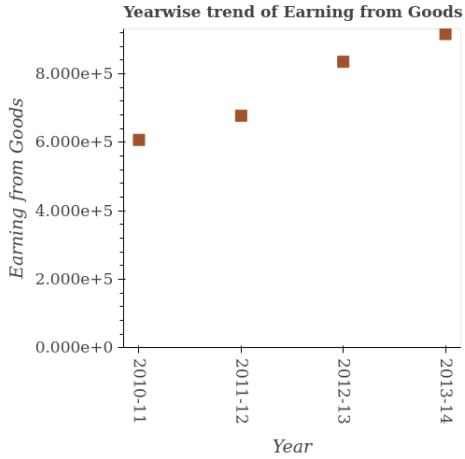
4. Question Templates

In this section, we present the 74 question templates which we have used for the question generation. Note that,

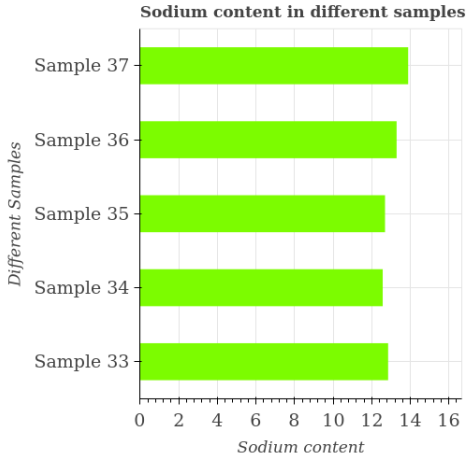
*The first two authors have contributed equally

Dataset	Plot Types				Question Types			Answer Types		
Split	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

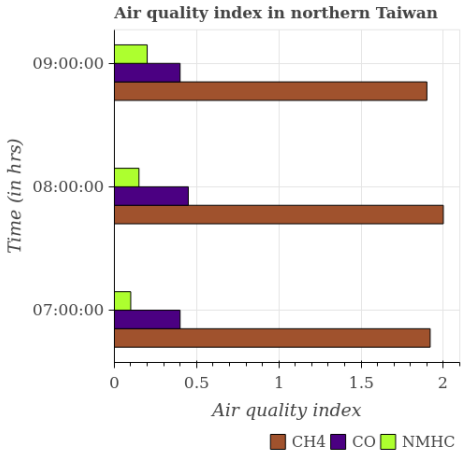
Table 1: Detailed Statistics for different splits of the PlotQA dataset.



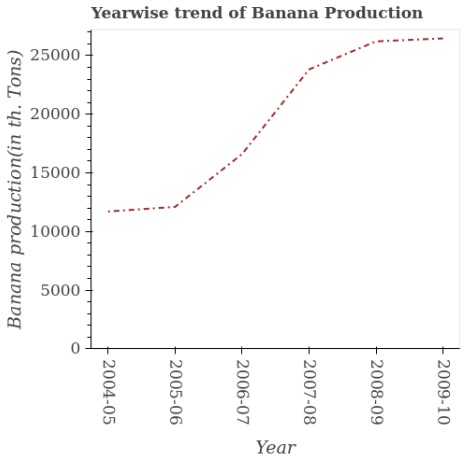
Q1: What is the difference between the highest and the second highest amount of earnings from goods?
A: $0.9e + 5$



Q2: What is the ratio of the sodium content in Sample 37 to that in Sample 33?
A: 1.086



Q3: What is the average air quality index value for NHMC per hour?
A: 0.167

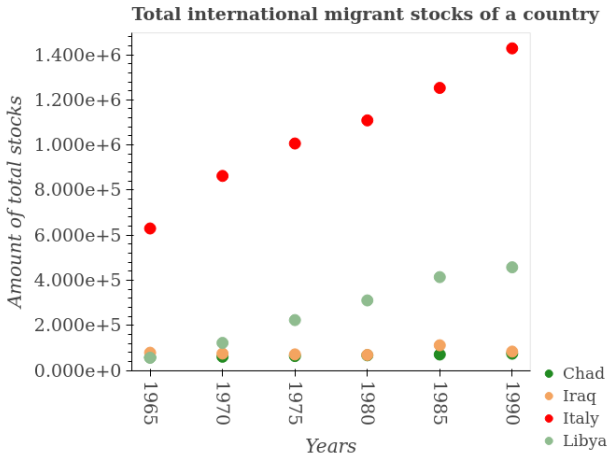


Q4: Does the amount of banana production monotonically increase over the years?
A: Yes

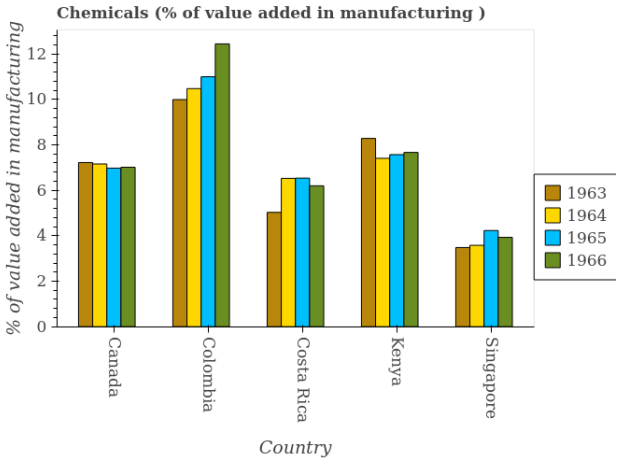
Figure 2: Sample {plot, question, answer} triplet present in the PlotQA dataset.

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.

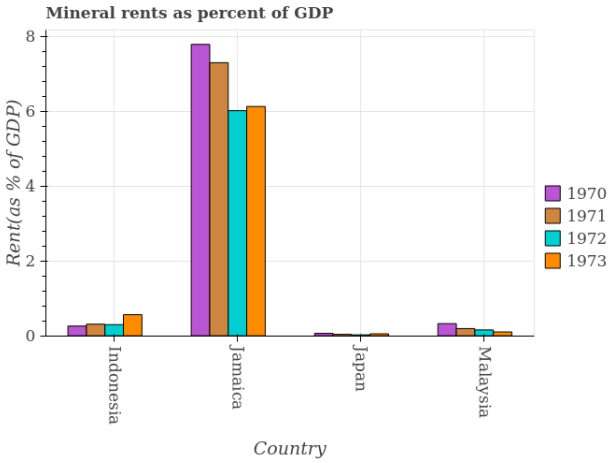
1. **Structural Understanding** :
    1. Does the graph contain any zero values?
    2. Does the graph contain grids ?
    3. Where does the legend appear in the graph ?
    4. How many legend labels are there?
    5. How are the legend labels stacked?
    6. How many <plural form of X_label> are there in



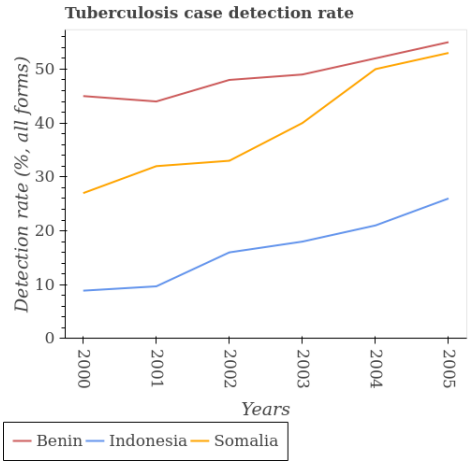
Q5: What is the difference between two consecutive major ticks on the Y-axis?
A: $2.000\text{e} + 5$



Q6: In how many cases, is the number of bars for a given year not equal to the number of legend labels?
A: 0



Q7: 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



Q8: What is the total tuberculosis detection rate in Indonesia?
A: 101

Figure 3: Few more sample {plot, question, answer} triplet present in the PlotQA dataset.

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 <Y_label> of/in <legend_label1> in $<$ i^{th}x_tick> and the <Y_label> of/in <legend_label2> in $<$ j^{th}x_tick> ?
16. What is the average <Y_label> of/in <legend_label> per <singular form of X_label> ?
17. In the year $<$ i^{th}x_tick>, what is the difference between the <Y_label> of/in <legend_label1> and <Y_label> of/in <legend_label2> ?
18. What is the difference between the <Y_label> of/in <legend_label1> and <Y_label> of/in <legend_label2> in $<$ i^{th}x_tick> ?
19. In how many <plural form of X_label>, is the <Y_label> greater than <N> units ?
20. Do a majority of the <plural form of X_label> between $<$ i^{th} x_tick> and $<$ j^{th} x_tick¿ (inclusive/exclusive) have <Y_label> greater than N <units> ?
21. What is the ratio of the <Y_label> in $<$ i^{th} x_tick> to that in $<$ j^{th} x_tick> ?
22. Is the <Y_label> in $<$ i^{th} x_tick> less than that in $<$ j^{th} x_tick> ?
23. In how many <plural form of X_label>, is the <Y_label> of/in <legend_label> greater than <N> <units>?
24. What is the ratio of the <Y_label> of/in <legend_label1> in $<$ i^{th} x_tick> to that in $<$ j^{th}x_tick>?

25. Is the <Y_label> of/in <legend_label> in < i^{th} x_tick> less than that in < j^{th} x_tick> ?
26. Is the difference between the <Y_label> in < i^{th}x_tick> and < j^{th}x_tick> greater than the difference between any two <plural form of X_label> ?
27. What is the difference between the highest and the second highest <Y_label> ?
28. Is the sum of the <Y_label> in < i^{th}x_tick> and < $(i+1)^{th}$x_tick> greater than the maximum <Y_label> across all <plural form of X_label> ?
29. What is the difference between the highest and the lowest <Y_label> ?
30. In how many <plural form of X_label>, is the <Y_label> greater than the average <Y_label> taken over all <plural form of X_label> ?
31. Is the difference between the <Y_label> of/in <legend_label1> in < i^{th}x_tick> and < j^{th}x_tick> greater than the difference between the <Y_label> of/in <legend_label2> in < i^{th}x_tick> and < j^{th}x_tick> ?
32. What is the difference between the highest and the second highest <Y_label> of/in <legend_label> ?
33. What is the difference between the highest and the lowest <Y_label> of/in <legend_label> ?
34. 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> ?
35. Is it the case that in every <singular form of X_label>, the sum of the <Y_label> of/in <legend_label1> and <legend_label2> is greater than the <Y_label> of/in <legend_label3> ?
36. Is the sum of the <Y_label> of/in <legend_label1> in < i^{th}x_tick> and < j^{th}x_tick> greater than the maximum <Y_label> of/in <legend_label2> across all <plural form of X_label>?
37. Is it the case that in every <singular form of X_label>, the sum of the <Y_label> of/in <legend_label1> and <legend_label2> is greater than the sum of <Y_label> of <legend_label3> and <Y_label> of <legend_label4> ?

5. 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 4b, 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 6b. 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 7b 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.

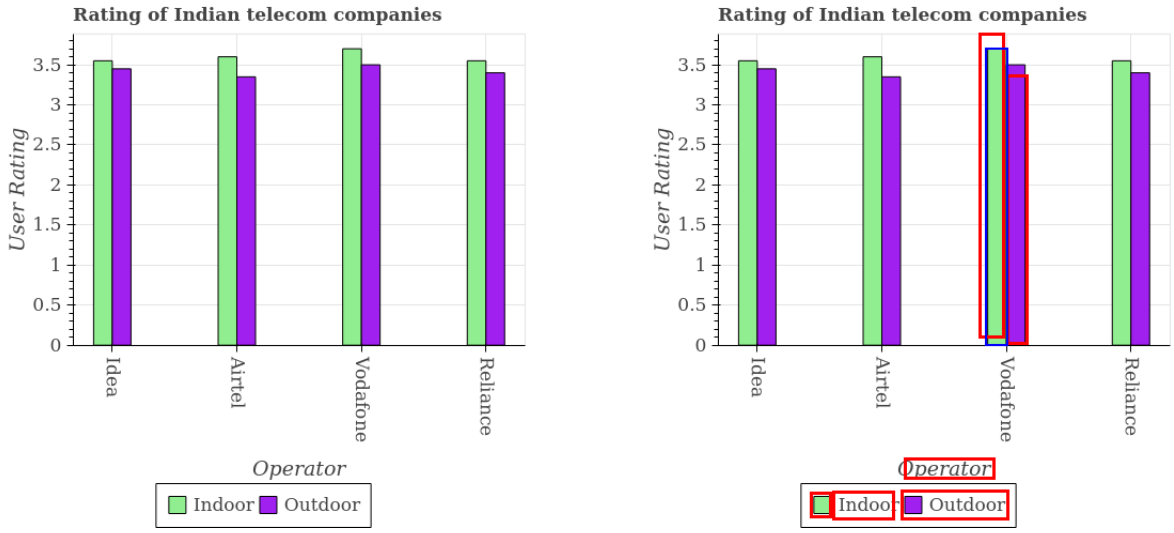
- **OCR stage:** A slight misalignment in the bounding boxes predicted by VED module causes significant errors while extracting the text. Example: In Figure 4b, consider the box enclosing the legend label "Indoor". The rightmost edge of the predicted bounding box is drawn over the letter "r", which makes 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 6b. 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 5b and Figure 7b) and incorrect answers to all the questions pertaining to these labels as shown in Table 2 and Tale 3.

- **SIE stage:** Figure 5a shows the oracle table which is generated by using the ground-truth annotations and Figure 5b 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 2 and Table 3 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.

References

[1] P. Pasupat and P. Liang. Compositional semantic parsing on semi-structured tables. In *ACL*, 2015.



(a) Input plot image

(b) Few examples of the predicted bounding boxes

Figure 4: 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

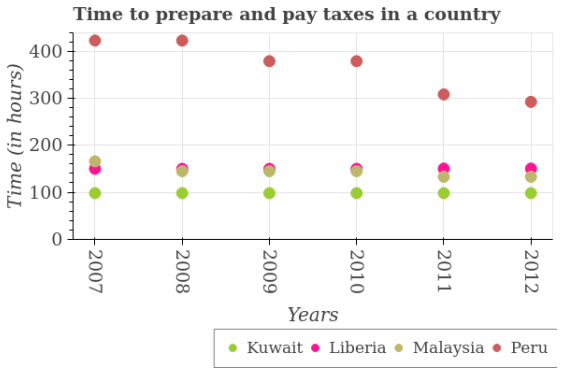
Dperator	Indoo	Outdoor
Idea	3.53	3.45
Airtel	3.60	3.40
Vodafone	4.0	3.35
Reliance	3.62	3.40

(b) Generated Semi-Structured Table

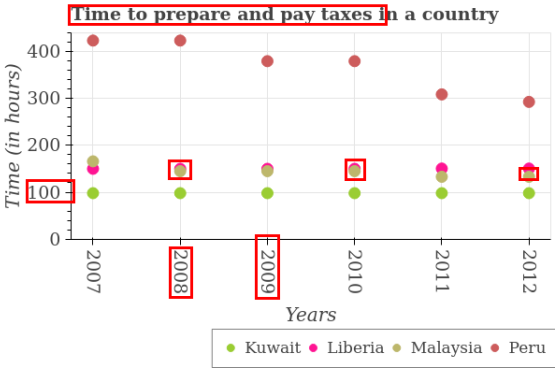
Figure 5: Errors made by the OCR and SIE stage (highlighted in red). Note that most of these errors have been propagated from the VED stage.

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 2: Answers predicted by different models on the sample questions.



(a) Input plot image



(b) Few examples of the predicted bounding boxes

Figure 6: 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
200B	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 7: 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. How are the legend-labels stacked?	horizontal	horizontal	horizontal	horizontal
Q2. What is the time (in hours) required to prepare and pay taxes in Kuwait in 2008?	98	100	100-	100-
Q3. 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-
Q4. 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
Q5. 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 3: Answers predicted by different models on the sample questions.