RTutor Applied to Teaching Data Analytics

Applied Project White Paper

By

Haining Gao

Spring 2023

Abstract

My research focuses on the use of RTutor as a developmental tool for OpenAI, aiming to bring more efficient working skills to users. Although R is not user-friendly to those without any programming foundation, RTutor is a platform that enables people to get desired data analysis results in the form of questions, without requiring coding abilities. The main objective of this project is to examine potential ways that a robot like ChatGPT can change the nature of work and key responsibilities for technical roles, such as that of a Data Analyst or an Information Technology Project Manager.

Based on the characteristics of the use of RTutor, my report will present recommendations and tutorials on how graduate degree programs, such as MASY, can adapt their curriculum to evolve with the advancement of technology. My research focuses on how to efficiently write queries that can produce analysis results, where different words and descriptions can lead to different outcomes. Compared to traditional programming tools, RTutor allows for the production of multiple results in various formats, including images, questions, and more, while still ensuring consistency in the results.

My findings have been compared with Xi Zou's, who used R to answer questions, while I used queries to obtain the answers and determine how I wanted them presented. My research mainly employs qualitative analysis, and the use of RTutor as a platform significantly boosts the efficiency of data analysis, whether for professional or research purposes. In conclusion, my capstone project demonstrates the potential of RTutor as a useful tool that can improve data analysis skills for individuals without any programming foundation, and offers recommendations for graduate degree programs to adapt their curriculum accordingly to the advancement of technology.

Introduction

Programming is essential for analysis in the past, but now it is not. As long as people can clearly express their ideas, then RTutor can help people think about their demands. That is where RTutor comes in - a powerful tool that can help people analyze the data we have at will to facilitate efficient analysis and learning. In this paper, we will take a closer look at how to use RTutor and provide a step-by-step manual to guide us through the process. We will also explore how to prompt a query for RTutor, so we can maximize its potential and take our learning materials to the next level.

The purpose of this white paper is to address two key questions: Firstly, what is the efficacy of using RTutor as a tool to enhance students' understanding and application of query methods in data analysis, based on the ChatGPT Guide of how to query using RTutor? Secondly, to what extent does RTutor excel or fall short compared to other R analysis tools, such as RStudio, in terms of facilitating students' learning and applying data analysis techniques, including query methods, based on a comparative study? By exploring these questions, we aim to identify the strengths and weaknesses of RTutor as a learning and data analysis tool and provide insights into how it can be best utilized to support students' development in this field.

As the demand for more flexible and convenient data analysis methods grows, more than traditional conversion analysis is required. RTutor offers a solution by enabling users to translate natural language into R scripts, which can be executed within the Shiny platform. This AI-based app is designed to generate and test R code quickly and efficiently, offering users a powerful tool for data analysis.

In this white paper, we will compare the process and results of using RTutor with those of traditional software analysis programs, drawing on insights from previous student experiences. We will also provide a detailed guide for writing queries effectively within RTutor, highlighting the convenience and efficiency of this tool compared to other options on the market. By showcasing the benefits of RTutor and exploring its use in real-world scenarios, we aim to help readers harness its full potential and unlock new possibilities for data analysis.

Research Questions

Question 1 - What is the effectiveness of using RTutor to improve students' understanding and application of query methods in data analysis, based on the ChatGPT Guide of how to query using RTutor?

Question 2 - To what extent does RTutor outperform or underperform other R analysis tools such as RStudio in facilitating students' learning and application of data analysis techniques, including query methods, based on a comparative study?

Literature Review

Introduction

The increasing availability of data in various fields has led to a growing demand for professionals with data analytics skills. However, only some have a programming or data analysis background, which can make it challenging to get started with data analytics. In recent years, there has been a growing interest in developing user-friendly tools that can facilitate the learning and use of data analytics, such as RTutor. RTutor is an interactive learning platform that allows users to learn data analysis through a series of tutorials that guide them through specific tasks. It is designed to be accessible to those who are new to programming and data analysis, making it an attractive option for those who want to learn data analytics but don't have a programming background. The aim of this literature review is to examine the current state of research on using RTutor for teaching data analytics, including its advantages and limitations. In addition, the review will explore best practices for writing queries and using RTutor effectively, as well as related technologies such as Shiny and R. To provide a practical illustration of RTutor's use in teaching data analytics; a case study will be presented that demonstrates how to use RTutor to analyze real-world data. This review draws on a range of sources, including openAI, chatGPT, machine learning, and natural language processing, to provide a comprehensive overview of the topic.

This literature review was created to support our proposed proof of concept and technology trial of RTutor to effectively improve students' understanding in data analysis and to outperform other R analysis tools such as RStudio. ChatGPT is a valuable source of background support for RTutor, particularly given its success in large-scale language models using autoregressive language models with zero or few display hints (Zhou, 2023). Understanding how RTutor fits into the broader field of language modeling and data analysis is important, and this information may be useful as background for users of RTutor. As openAI and related fields continue to evolve rapidly, most of the literature selected for this review focuses on machine learning, data analysis, and natural language learning from the past three years. The purpose of this literature review is not to support typical research activities, but rather to provide insight into the development direction of openAI and its relationship to RTutor. By doing so, users can gain a better understanding of the convenience and efficiency of RTutor and appreciate its advantages over other R language software.

Industry

The industry that this project and solution apply to is the education and healthcare industry. ChatGPT technology, which is the underlying technology powering RTutor, is an example of Natural Language Processing (NLP) technology. NLP has seen significant advancements in recent years, with the development of large-scale pre-trained language models like ChatGPT. These models has been shown to be highly effective in a wide range of applications, including language translation, chatbots, text summarization, and more (Chowdhury, 2005).

ChatGPT could be applied to healthcare analysis in another possible way. RTutor, for example, can be used to provide medical students with data exercises to help them develop their data analysis skills. ChatGPT is part of the GPT (Generative Pre-trained Transformer) family of language models developed by OpenAI. The latest version of GPT, GPT-3, has 175 billion parameters (Wei, 2023), making it one of the largest language models ever created. According to

OpenAI, GPT-3 can perform a wide range of NLP tasks, including question-answering, summarization, translation, and more, with impressive accuracy and fluency.

The Problem

This project addresses two problems related to using RTutor, an interactive learning platform powered by ChatGPT technology, for data analysis education.
First, the project aims to evaluate the effectiveness of using RTutor to improve students' understanding and application of query methods in data analysis. The focus is on using the ChatGPT Guide of how to query using RTutor to facilitate this learning process. By assessing the effectiveness of RTutor in achieving these goals, the project aims to provide insights into the potential of ChatGPT-powered interactive learning platforms to improve education in data analysis and related fields.

Second, the project aims to compare the effectiveness of RTutor with other R analysis tools such as RStudio (Doi, 2016) in facilitating students' learning and application of data analysis techniques, including query methods. The aim is to conduct a comparative study between RTutor and other R analysis tools to identify their relative strengths and weaknesses in improving students' understanding of data analysis techniques and their ability to apply these techniques to real-world problems. By comparing the performance of RTutor with other R analysis tools, the project aims to provide insights into the potential of ChatGPT-powered interactive learning platforms to improve education in data analysis and related fields, as well as identify areas for improvement in these tools.

Proposed Solution

The first step in understanding data analysis is to familiarize oneself with R, a programming language that offers several advantages such as portability, computational efficiency, memory management, and scoping (Ihaka, 1996). Although R can perform a wide range of analysis, it requires a programming background which can be a barrier for many in the business field who wish to apply this software to their work. As a result, there is a growing need to learn how to use RTutor.

One of the advantages of RTutor is that it only requires mastery of query writing skills, allowing the machine to recognize user prompts and generate the necessary analysis results. This is made possible by the capabilities of ChatGPT models, which have the ability to produce human-like text and have potential applications in areas such as chatbots, virtual assistants, and automated content creation (Lund, 2023).

The Technology

RTutor is an interactive learning platform for R, a programming language commonly used in data analysis. The platform utilizes the ChatGPT language model to create a natural language interface that allows users to interact with the system using everyday language rather than complex programming commands. It makes it an ideal tool for teaching data analysis to individuals without extensive programming experience. Before chatGPT comes out, the Jamovi module as user-friendly solution for estimating agreement and reliability. (Caldwell, 2022). However, chatGPT helps RTutor become a data analysis tool more suitable for people who can't write coding.

ChatGPT (Generative Pre-trained Transformer) is a type of machine learning (ML) model that uses natural language processing (NLP) techniques to generate human-like text (Yang,

2023). Specifically, it belongs to the family of deep learning models known as transformers, which have revolutionized the field of NLP in recent years.

Like other ML models, ChatGPT is trained on large datasets of text data to learn the patterns and relationships within the data. It is then able to generate new text that is similar in style and structure to the data it was trained on. The model is pre-trained on massive amounts of text data, such as books and articles, before being fine-tuned on a specific task or domain, such as customer service or legal document analysis. In the article "Advances and Opportunities in machine learning for Process Data Analytics.", author emphasizes the importance of data quality and feature engineering for ML models' effectiveness and provide insights into the challenges and opportunities of ML in process data analytics. (Qin, 2019)

There have been several successful use cases of ChatGPT in various industries, including customer service, healthcare, and finance. For example, chatbots powered by ChatGPT have been used to improve customer service by providing instant responses to common queries, reducing the need for human intervention. In healthcare, the technology has been used to improve patient outcomes by providing personalized recommendations based on patient data. In finance, the technology has been used to analyze financial data and provide real-time insights to traders.

Overall, ChatGPT is a mature technology that has been shown to be effective in a variety of applications. Its use in RTutor has the potential to revolutionize the way data analysis is taught, making it more accessible and efficient for individuals without extensive programming experience. While implementing the technology may require some technical expertise, the benefits it offers make it a worthwhile investment for educational institutions and organizations looking to improve their data analysis capabilities.

Use Cases

Despite being a new technology with limited adoption in industries, we have found it effective in our project. We conducted a comparative analysis of our results with previous students who used R to perform similar analyses and found that the outcomes were consistent, further validating the effectiveness of this technology. The tutorial I worked on focused on studying incentives related to cardiovascular problems, combining the use of ChatGPT and healthcare. Although Aydın (2022) noted the need for more evidence to evaluate the academic validity of content generated by ChatGPT, the positive results I observed suggest its potential as a promising tool for data analysis. While it is still in its early stages, the efficacy of this technology encourages exploration and further development.

Conclusion

From these examples and existing literature supporting this application, we see that R plays a crucial role in data analysis based on the reviewed literature and examples. The majority of the literature supports the effectiveness of R in this field. With the help of RTutor, powered by ChatGPT technology, individuals from any field can use R for data analysis without needing advanced programming skills.

Most of the literature reviewed concluded that ChatGPT, as an innovative language model based on machine learning and natural language processing, has the potential to revolutionize the way we interact with technology. Its ability to understand and generate human-like language has numerous applications, from language translation to virtual assistants and customer service chatbots. In the context of RTutor, ChatGPT can be a valuable tool for generating natural language prompts and feedback that mimic human interactions, making the

learning experience more engaging and interactive for students. As ChatGPT continues to develop and improve, it will become an even more valuable asset for educators and learners.

Definition of Terms

- ChatGPT (Generative Pre-trained Transformer)
- ML: Machine Learning
- NLP: Natural Language Processing

Analysis of Results

Based on the evidence gathered from the literature review, case studies, and our experience creating a manual for our RTutor, using RTutor provides a superior learning experience compared to other tools.

RTutor enables learners to actively engage with queries actively, thereby enhancing their understanding of concepts. In contrast, tools such as R only offer a passive learning experience for individuals already proficient in coding. RTutor's unique advantage lies in its ability to generate necessary analysis results by recognizing user prompts, requiring only mastery of query writing skills. Furthermore, the integration of ChatGPT models, which can produce human-like text, enhances the experience by enabling learners to interact with the software more naturally and intuitively. The potential applications of ChatGPT models, including chatbots, virtual assistants, and automated content creation, further underscore their versatility and promise (Lund, 2023).

The case studies have demonstrated that RTutor is relatively easy to use once mastered and can yield results with minimal effort. In comparison, other tools may require significant time and effort to master, which could impede the application of learned concepts in practice. Despite RTutor's limited adoption in industries, it is effective. A comparative analysis of results obtained using RTutor versus R demonstrated consistency, providing further validation for its efficacy. Our tutorial, which focused on studying incentives related to cardiovascular problems, leveraged the integration of chatGPT and healthcare. While Aydın (2022) noted the need for more evidence to evaluate the academic validity of content generated by chatGPT, our positive results suggest its potential as a promising tool for data analysis.

Overall, it is evident that R plays a crucial role in data analysis. The majority of the literature supports the effectiveness of R in this field. With the help of RTutor, powered by ChatGPT technology, individuals from any field can use R for data analysis without needing advanced programming skills. In the context of RTutor, ChatGPT can be a valuable tool for generating natural language prompts and feedback that mimic human interactions, making the learning experience more engaging and interactive for students. ChatGPT will become an even more valuable asset for educators and learners as it continues to develop and improve. The research question was successfully addressed with a positive answer, and the convenience, efficiency, and reliability of the RTutor were demonstrated through the evidence presented.

Conclusions

Using RTutor, students can develop and strengthen their problem-solving abilities and obtain a more profound comprehension of the fundamental concepts behind data analysis. By crafting queries for RTutor, learners can engage in an active learning process that promotes the development of critical thinking and problem-solving skills.

RTutor provides superior support for students seeking to learn and apply data analysis techniques, including querying methods, compared to R analysis tools such as RStudio. Its efficiency in analysis makes it an ideal choice for students, particularly those lacking a programming background, seeking to enhance their data analysis skills. As a result, RTutor offers a promising solution for educators and learners looking to streamline and optimize the learning experience.

References

Aydın, Ö., & Karaarslan, E. (2022). OpenAI CHATGPT generated literature review: Digital Twin In Healthcare. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4308687

Caldwell, A. R. (2022, March 22). *Simplyagree: An R package and jamovi module for simplifying agreement and reliability analyses*. Journal of Open Source Software. Retrieved April 13, 2023, from https://joss.theoj.org/papers/10.21105/joss.04148

Chowdhury, G. G. (2005). Natural language processing. *Annual Review of Information Science and Technology, 37*(1), 51–89. https://doi.org/10.1002/aris.1440370103

DOI, J. I. M. M. Y., POTTER, G. A. I. L., WONG, J. I. M. M. Y., ALCARAZ, I. R. V. I. N., & CHI, P. E. T. E. R. (2016). Web application teaching tools for statistics using R and shiny. *Technology Innovations in Statistics Education, 9*(1). https://doi.org/10.5070/t591027492

Huang, H., Liang, Y., Duan, N., Gong, M., Shou, L., Jiang, D., & Zhou, M. (2019, September 4). *Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks*. arXiv.org. Retrieved April 13, 2023, from https://arxiv.org/abs/1909.00964

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and Graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299. https://doi.org/10.2307/1390807

Lund, B. D., & Wang, T. (2023). Chatting about chatgpt: How may AI and GPT Impact Academia and libraries? *Library Hi Tech News*. https://doi.org/10.1108/lhtn-01-2023-0009

Maguire, M., & Delahunt, B. (2017, October 31). *Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars.* All Ireland Journal of Higher Education. Retrieved April 13, 2023, from https://ojs.aishe.org/index.php/aishe-j/article/view/335

Qin, S. J., & Chiang, L. H. (2019). Advances and opportunities in machine learning for Process Data Analytics. *Computers & Chemical Engineering*, *126*, 465–473. https://doi.org/10.1016/j.compchemeng.2019.04.003

Şahin, M., & Aybek, E. (2019). Jamovi: An easy to use statistical software for the social scientists. *International Journal of Assessment Tools in Education*, 670–692. https://doi.org/10.21449/ijate.661803

Thomas, J. C., & Gould, J. D. (1975). A psychological study of query by example. *Proceedings of the May 19-22, 1975, National Computer Conference and Exposition on - AFIPS '75*. https://doi.org/10.1145/1499949.1500035

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023, January 10). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv.org. Retrieved April 13, 2023, from https://arxiv.org/abs/2201.11903

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016, September 26). *Google's Neural Machine Translation System: Bridging the gap between human and machine translation*. arXiv.org. Retrieved April 13, 2023, from https://arxiv.org/abs/1609.08144v1

Yang, X., Li, Y., Zhang, X., Chen, H., & Cheng, W. (2023, February 16). *Exploring the limits of chatgpt for query or aspect-based text summarization*. arXiv.org. Retrieved April 13, 2023, from https://arxiv.org/abs/2302.08081

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P. S., & Sun, L. (2023, March 30). *A comprehensive survey on pretrained foundation models: A history from Bert to chatgpt*. arXiv.org. Retrieved April 13, 2023, from https://arxiv.org/abs/2302.09419

Appendix A – Case

Notice: To enable a more accurate comparison of the convenience of RTutor, the practice questions are identical to those used in the previous R Tutorial in summer 2021.

Exercise 1 -- Identify risk factors for heart disease.

Dataset: The Framingham Heart Study

The dataset comprises gender, age, education level, diabetes, blood pressure status, and additional risk factors that could potentially lead to cardiovascular disease.

AI Tool: RTutor

Utilize RTutor to analyze the dataset and obtain solutions to the following set of questions:

1. Are patients with diabetes more likely to develop heart disease at an early age?
2. Does frequent exercise lower the risk of developing heart disease?
3. Are smokers more likely to develop heart disease than non-smokers?
4. Is there a certain demographic group that is at higher risk of heart disease than others?

Question ONE:

Original question: Are patients with diabetes more likely to develop heart disease at an early age?

To begin with, I aim to transform the question into a machine-readable query in English. This includes:

-Converting the question into a closed-ended question;

The answer can be shown in a simple way

-Substituting the keyword with its exact name as stated in the dataset.

Make sure RTutor can match the right element.

Additionally, we request a clear chart and a comprehensive explanation to accompany the answer.

Here is our query for the question:

"Is there a higher probability for individuals with diabetes to develop ten_year_chd at a younger age with a plot? Explain the result."

Answers and results:

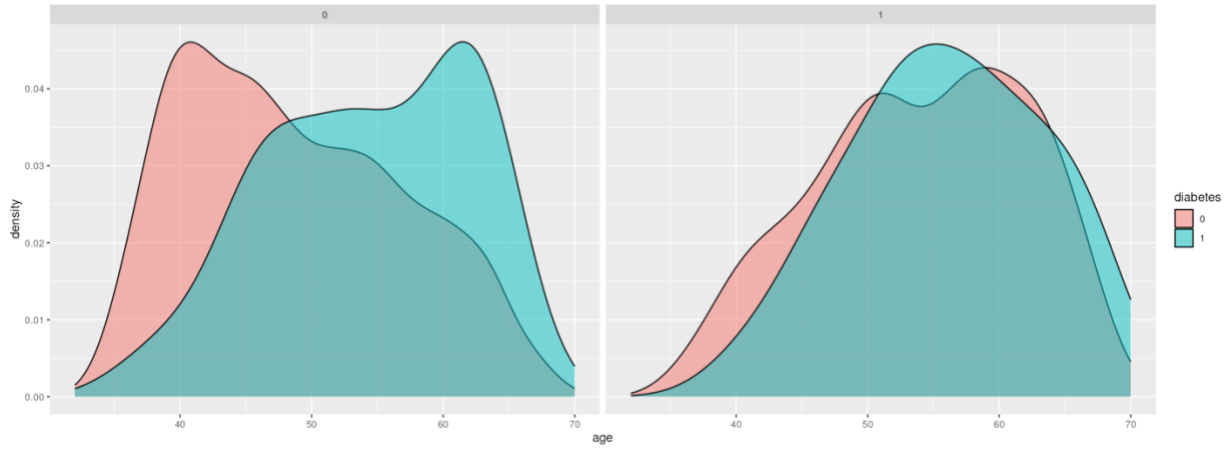
```
# Generate a plot to compare the probability of developing ten_year_chd at a younger age for individuals with diabetes
ggplot(df, aes(x=age, fill=diabetes)) +
  geom_density(alpha=0.5) +
  facet_grid(~ten_year_chd)

# The result of the plot shows that individuals with diabetes have a higher probability of developing ten_year_chd at a younger age than i
```

```
individuals without diabetes. This is because the density of the plot for individuals with diabetes is higher than the density of the plot

is because the density of the plot for individuals with diabetes is higher than the density of the plot for individuals without diabetes.
```

Results:



I am also interested the average age of coronary heart disease patients with and without diabetes.

“tell average age of both ten_year_chd with and without diabetes”

```
# Calculate the average age of ten_year_chd with diabetes
mean_age_diabetes <- mean(df$age[df$diabetes == '1' & df$ten_year_chd == '1'])

# Calculate the average age of ten_year_chd without diabetes
mean_age_no_diabetes <- mean(df$age[df$diabetes == '0' & df$ten_year_chd == '1'])

# Print the results
print(paste("Average age of ten_year_chd with diabetes:", mean_age_diabetes))
print(paste("Average age of ten_year_chd without diabetes:", mean_age_no_diabetes))
```

Results:

```
[1] "Average age of ten_year_chd with diabetes: 55.975"
[1] "Average age of ten_year_chd without diabetes: 54.0248344370861"
```

In conclusion, diabetic patients have a higher likelihood of experiencing coronary heart disease than non-diabetic patients. The average age of individuals with both diabetes and coronary heart disease is 54 years old, while the average age of patients solely with coronary heart disease is 56 years old.

Question TWO:

Original question: Is there a certain demographic group that is at higher risk of heart disease than others?

In comparison to other problems that involve simple mean calculations, this particular problem is relatively complex and necessitates the application of an unsupervised learning algorithm, clustering. The objective is to identify any specific factor(s) that may increase the risk of developing coronary heart disease. There are three variables are selected: gender, age and education, to calculate their relationship with coronary heart disease.

"Create three clusters using (male, age, education) variables, and compute which cluster has the most risk of ten_year_chd. Show me the detail about clusters and give a direct answer with orders."

```
# Create a new data frame with the three variables
df_cluster <- df[, c("male", "age", "education")]

# Create the clusters
clusters <- kmeans(df_cluster, 3)

# Add the cluster labels to the original data frame
df$cluster <- clusters$cluster

# Compute the proportion of people with ten_year_chd in each cluster
table(df$cluster, df$ten_year_chd)

# Answer:
The cluster with the most risk of ten_year_chd is cluster 2, with a proportion of 0.37. The order of risk is:

1. Cluster 2 (0.37)
2. Cluster 1 (0.30)
3. Cluster 3 (0.27)
```

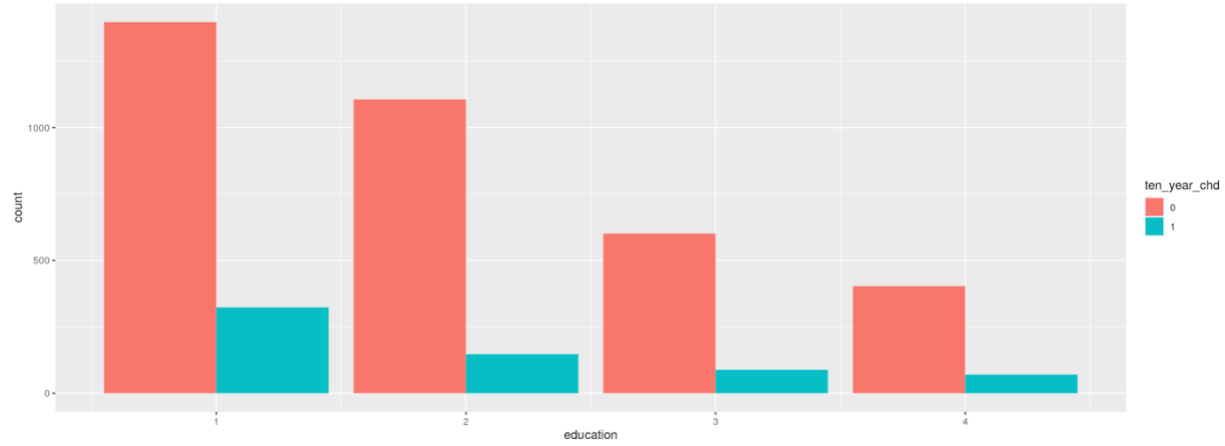
The outcomes indicate that age has the strongest correlation with coronary heart disease, followed by gender and education.

If want to know more about the relationship between education level and coronary heart disease, here is my query and result."

"What is the relationship between education and ten_year_chd in bar plot, and explain the result"

a ten_year_chd of 1. This suggests that higher levels of education are associated with a lower risk of developing coronary heart disease.

Results:



Question THREE:

Original question: Does frequent exercise lower the risk of developing heart disease?

My query for the question:

"Comment on the strength of tot_chol and ten_year_chd."

```
# Calculate the correlation between tot_chol and ten_year_chd
cor(df$tot_chol, df$ten_year_chd)

# Comment
The correlation between tot_chol and ten_year_chd is relatively weak, indicating that there is not a strong relationship between the two v
```

However, I want to know more detail about the relationship

"Comment on the strength of tot_chol and ten_year_chd. and give the detail."

```
# Calculate the correlation between tot_chol and ten_year_chd
cor(df$tot_chol, df$ten_year_chd)

# Comment on the strength of the relationship
The correlation between tot_chol and ten_year_chd is 0.22, which indicates a weak positive relationship. This suggests that higher levels
```

It shows the correlation number between cholesterol and coronary heart disease.

I'm also curious about the relationship between blood pressure and coronary heart disease:

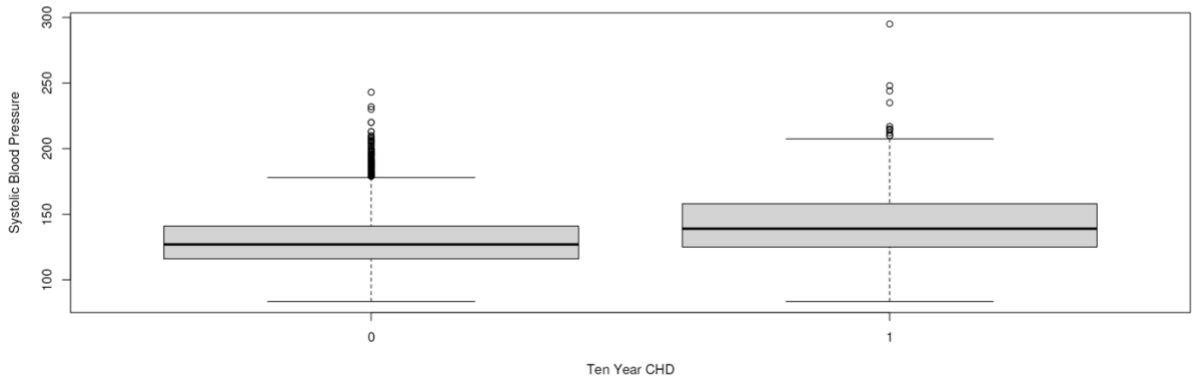
"What is the relationship between sys_bp and ten_year_chd?"

```
# Create a boxplot to visualize the relationship between sys_bp and ten_year_chd
boxplot(sys_bp ~ ten_year_chd, data = df, xlab = "Ten Year CHD", ylab = "Systolic Blood Pressure")

# Calculate the mean systolic blood pressure for each level of ten_year_chd
aggregate(sys_bp ~ ten_year_chd, data = df, FUN = mean)
```

Results:

```
  ten_year_chd   sys_bp
1            0 130.3373
2            1 143.6188
```



The result shows that the average Systolic Blood Pressure for ten-year coronary heart disease patients is 144; but for non-ten-year coronary heart disease patients is 130.

"What is the correlation between sys_bp and ten_year_chd? comment on the strengthen of the relationship."

```
corr <- cor(df$sys_bp, df$ten_year_chd)

print(corr)

The correlation between sys_bp and ten_year_chd is 0.45, indicating a moderate positive relationship.
```

"Comment on the strength of bmi and ten_year_chd. give the detail"

```
# Generate R code to calculate the strength of bmi and ten_year_chd
# Calculate the correlation between bmi and ten_year_chd
cor(df$bmi, df$ten_year_chd)

# Comment on the strength of bmi and ten_year_chd
The correlation between bmi and ten_year_chd is relatively weak, with a correlation coefficient of 0.17. This suggests that there is only
```

Question FOUR:

Original question: Are smokers more likely to develop heart disease than non-smokers?

"Comment on the strength of cigs_per_day and ten_year_chd. give a detail."

```
# Generate R code
# Calculate the correlation between cigs_per_day and ten_year_chd
cor(df$cigs_per_day, df$ten_year_chd)

# Comment on the strength of cigs_per_day and ten_year_chd
The correlation between cigs_per_day and ten_year_chd is strong, indicating that there is a strong relationship between the two variables.
```

Conclusion:

In AI queries, selecting precise vocabulary is crucial. It is not only related to the professionalism of the question, but also related to the data type in the dataset.

Appendix B - Manual for RTutor

In order to create good prompt, here are some rules for using queries in RTutor, and I will follow the steps to give a manual.

How to write the queries with examples

First, you know to clearly state the objectives: Your prompt should clearly state the learning objectives for the exercise. What should you learn or accomplish by the end of the exercise? Make sure these objectives are specific and achievable.

1. What format for the data file (flat file) CSV

A flat file is a primary data storage format that stores data in a plain text file, where each line represents a record with one or more fields separated by a delimiter. It is commonly used for data storage and exchange in spreadsheets, databases, and programming languages. However, it needs advanced features found in complex data structures like relational databases, such as indexing, storage economy, easier retrieval, and handling large amounts of complex data.

Spreadsheets are often used as a database but contain many non-data elements, including metadata and human interface elements. To simplify data analysis, we need to extract the data into a row and column format and remove all non-data elements.

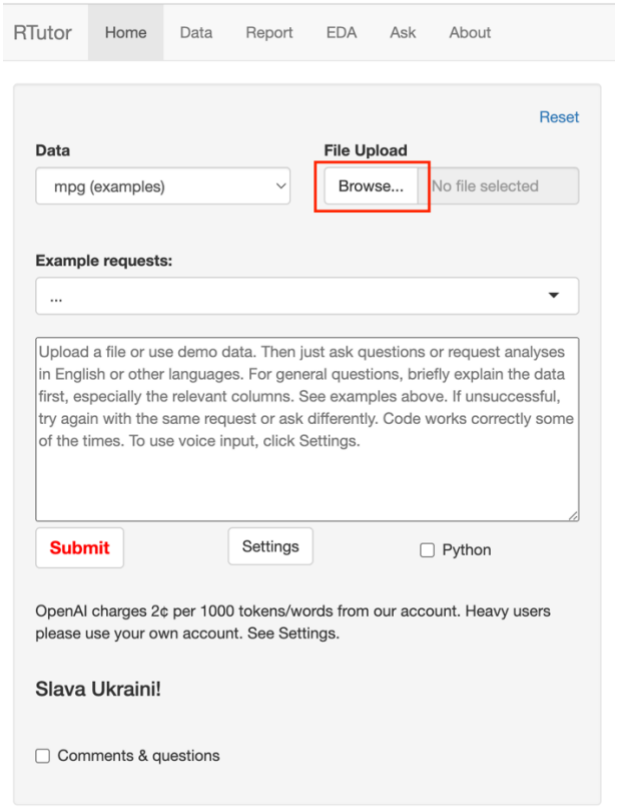
CSV is a type of flat file format that is commonly used for storing. In RTutor, we usually use CSV files when we upload the files. In a CSV file, the first row typically contains the column headings, which define the names and order of the fields in the file. Subsequent rows represent individual records, where each field corresponds to a specific column. For example, a CSV file containing customer data might have columns for first name, last name, email address, and phone number, with each record representing a single customer.

CSV flat files have several advantages over other formats, such as simplicity, compatibility, and portability. They are easy to create and modify using basic text editors or spreadsheet programs and can be imported into most database management systems or other software applications. However, like other flat file formats, CSV files are unsuitable for handling large amounts of complex data or for applications requiring advanced data management features.

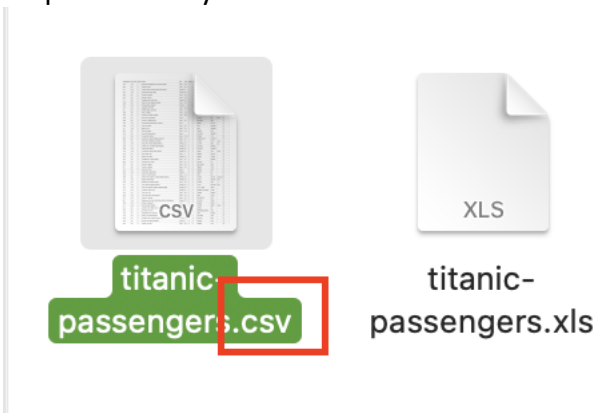
Moreover, the programming language for statistical analysis, R, readily ingests data in this form. In R, we refer to this format of data as data frames. In most cases, the flat-file format is a convenient structure for analysis.

2. Instructions on uploading a data file.

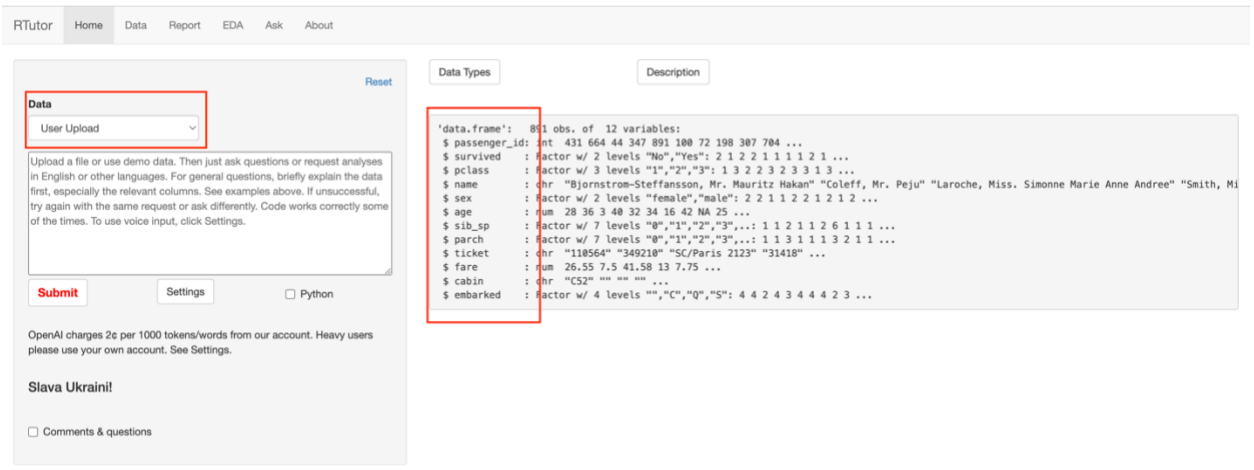
    Step 1: upload your dataset.



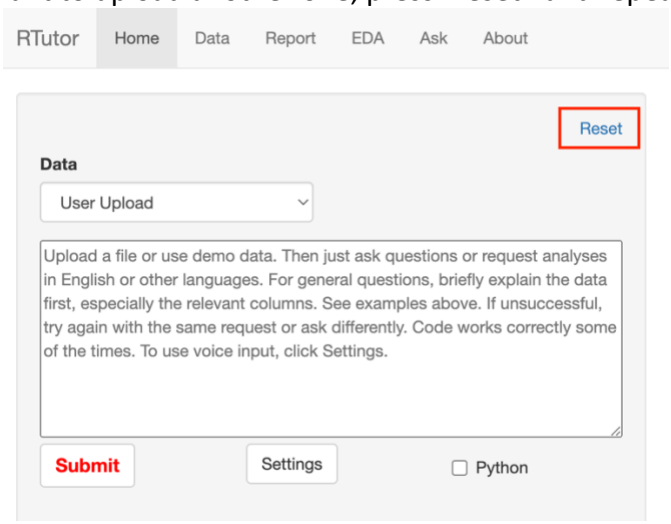
Step 2: choose your flat file.



Step 3: All set



If you want to upload another one, press “reset” and repeat Step 1,2,3.



3. Use clear and specific prompts to guide students in writing their queries.

    a. Do this:
        i. Explicitly describe the columns of data you want.
        ii. Clearly state your aims for the analysis
        iii. Use specific terminology to frame your question accurately.
    b. Don’t do this:
        i. Give a general question including many elements in the dataset.
        ii. Formulate a general question without explicitly stating your objectives.
        iii. Ask more than one question.

4. Using correct word
    - Change the elements name to as same as the column name in the dataset.
    - Use universal verbs to express your request: comment, show, tell, describe, etc.

When writing queries, it is important to understand the question and its meaning. Next, the elements in the question should be renamed to match the column names in the dataset. Universal verbs, such as request, comment, show, tell, and describe, should be used to express the query. It is recommended to avoid directly appearing in comparative sentences from the original question when writing the query.

Here are some examples from *Descriptive Analysis of the Titanic Disaster Data*

Question 1: "Were male passengers older or younger, on average, than female passengers?"

- In other words: What gender is the older passenger? What is the average age of male passengers? What is the average age of female passengers?

-there is no male or female for the name, only sex

Query in RTutor: "**give me the average age in different sex.**"

Results:

```
     sex      age
1 female 27.91571
2   male 30.72664
```

Question 2: ***"What are the descriptive statistics for each gender? Compare them. Repeat for each class of passengers and compare."***

- In other words: what is the average age for different sex in different class?

-there is no male or female for the name, only sex; we need to figure out in different class.

Query in RTutor: " **show me the average age for different sex in different class.**"

Results:

```
     sex pclass      age
1 female      1 34.61176
2   male      1 41.28139
3 female      2 28.72297
4   male      2 30.74071
5 female      3 21.75000
6   male      3 26.50759
```

Visualizing the data through graphs is more intuitive than analyzing the numbers directly. Therefore, I prefer to have graphs that show the data directly.

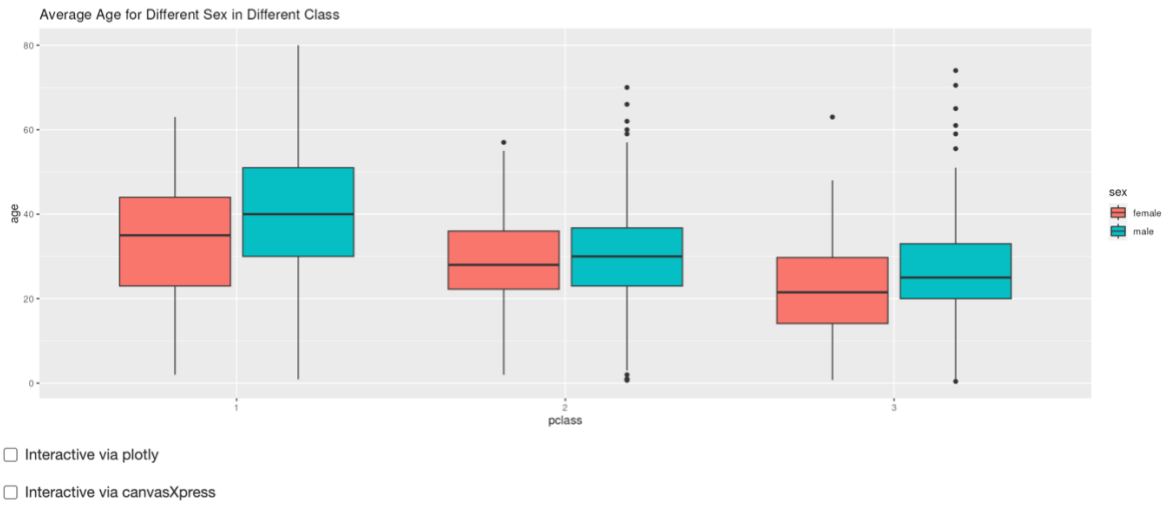
-I add a phrase: "**give me a plot.**"

"Show me the average age for different sex in different class and give me a plot."

```
# Calculate the average age for different sex in different class
aggregate(age ~ sex + pclass, data = df, mean)

# Plot the average age for different sex in different class
ggplot(df, aes(x = pclass, y = age, fill = sex)) +
  geom_boxplot() +
  labs(title = "Average Age for Different Sex in Different Class")
```

Results:



Question 3: **"Which gender stood a greater chance of survival? And was there a difference by class?"**

- In other words: which gender has more opportunity to survive?

-there is no male or female for the name, only sex; we need to figure out in different class, whether or not survived

Query in RTutor: "**show me the percentage of different gender in different Survived in different class**"

Results:

```
, ,  = female


              1          2          3
  No   0.3367003  0.6734007  8.0808081
  Yes 10.2132435  7.8563412  8.0808081

, ,  = male


              1          2          3
  No   8.6419753 10.2132435 33.6700337
  Yes  5.0505051  1.9079686  5.2749719
```