INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 07 ISSUE: 07 | JULY - 2023 SJIF RATING: 8.176 ISSN: 2582-3930

PANDAS AI: A Step Towards GEN AI

Neha Jain
Marketing Analytics,
Accenture Technology,
Gurgaon, India
neha.t.jain@accenture.com

Aayushi Tayal
Marketing Analytics,
Accenture Technology, Gurgaon, India
aayushi.tayal@accenture.com

Abstract

In the era of Generative Artificial Intelligence, there is a widespread emphasis on employing conversational methods to address and resolve issues. Conversational approaches to problem-solving have become the primary focus for everyone, regardless of the nature of the problem at hand. With the release of ChatGPT and chatbots like technologies, developers are also having high hopes of having such libraries and tools in future which can make their life easier.

Conversational AI is a strategy enabling users to communicate with machines using language that is like that of humans. These programs will translate human language into relevant outcomes. Pandas AI has provided similar type of functionality for the most widely used Pandas library of Python.

Pandas AI is a python library that adds human like conversational ability to Pandas and can be used for Data Analysis, cleaning, manipulation, transformations, insights generation etc. This library is used in conjunction with existing Pandas python library.

Keywords—Pandas AI, Generative AI, Conversational AI, Artificial Intelligence

INTRODUCTION

Generative AI, a type of Artificial Intelligence that can create large numbers of data, for example images, text, videos etc. through the practice of selecting patterns from a set of available data and then use those knowledges to develop new and special outputs. With the recent breakthrough of CHATGPT, Companies are becoming focused on the development of Intelligent Systems that will acquire fundamental intelligence comparable to or higher than human intelligence and thus radically change our communication, work and innovation processes. Python has also developed some libraries which have gen AI capabilities and can help end users in various ways.

Pandas, a most widely used python library has been proved very beneficial for Data Analysis, cleaning, manipulation, transformation, and insights generation. Developers and Analyst used to spend a lot of time in writing codes and making the data ready for modelling purpose.

It is important to note that the size of data held by the software industry is continually growing due to factors such as the increasing use of the Internet, the proliferation of connected devices and generation of large scale datasets via scientific research, social media and other sources. This increase in data size requires some advanced data processing techniques which can help developers to generate the valuable insights quickly and efficiently.

Companies like Google, Facebook and Amazon are handling enormous amounts of data which is continuously increasing with their usage. For instance, Google is processing over 3.5-4 billion searches per day. Facebook having more than billions of active users is managing a large amount of data like the user profiles, their interactions with other users, posts, photos, their likes and dislikes etc. Similarly, e-commerce platforms like Amazon, Myntra and Flipkart are handling substantial volume of customer and product data. With the rise in multimedia content like high-definition photos, video and audio files whose size is comparatively very large than the text based data, the size of data is increasing more quickly than anticipated.

Take social media or any other e-commerce site as an example. When a person searches for a product or watches a video, they often see advertisements for related things. A user's account will still display the same sort of advertisement even if they view a product on

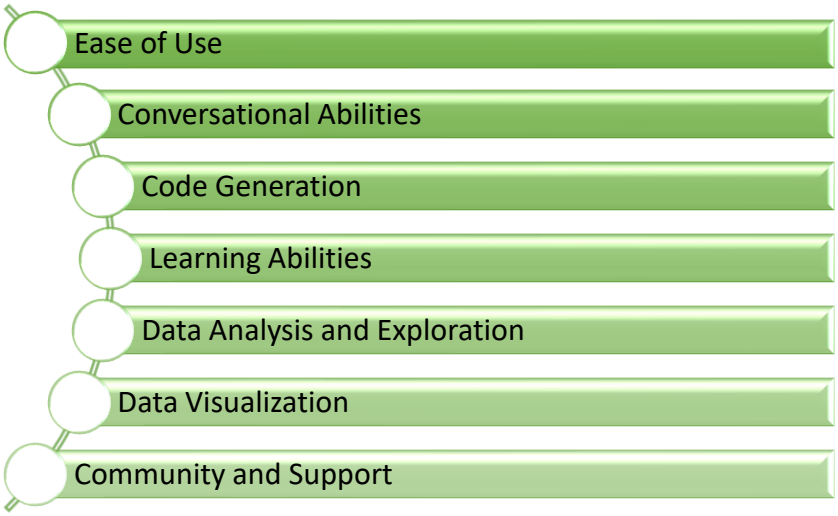
their browser and then switch to another social media network, such as Instagram or Facebook. This is how Data Analysis works. It examines client information, trend data, and then presents them with comparable adverts based on that analysis. YouTube and many other websites may also be used to examine the same pattern.

In this paper, we are going to discuss the way to converse with the data in a human-friendly language like we used to talk to our best friend. That's what Pandas AI does. Using Open AI LLM(Large Language Models) techniques, python has provided a library that is very helpful in conversing with the data in human like language.

PROBLEM STATEMENT

Currently for data loading, manipulation, visualization in form of charts, tables etc., data exploration such as finding patterns in data using Python Pandas, developers and analysts dedicate a significant amount of time in writing code. They typically write scripts in a programming language, which often requires an intermediate level of expertise. With any change in data, they perform the analysis and determine if the existing set of scripts will still function with the updated data. Furthermore, when there are changes in the data, the previously written set of manipulations and transformations may become outdated. As a result, this repetitive process involves extensive analysis, time, and effort from the developers, leading to delays in the overall process.

Using Gen AI capabilities of Pandas AI, it allows developers to interact with data in a conversational way and hence freeing them to focus more on other work. Pandas AI provides following benefits on top of other approaches:



HOW PANDAS AI WORKS?

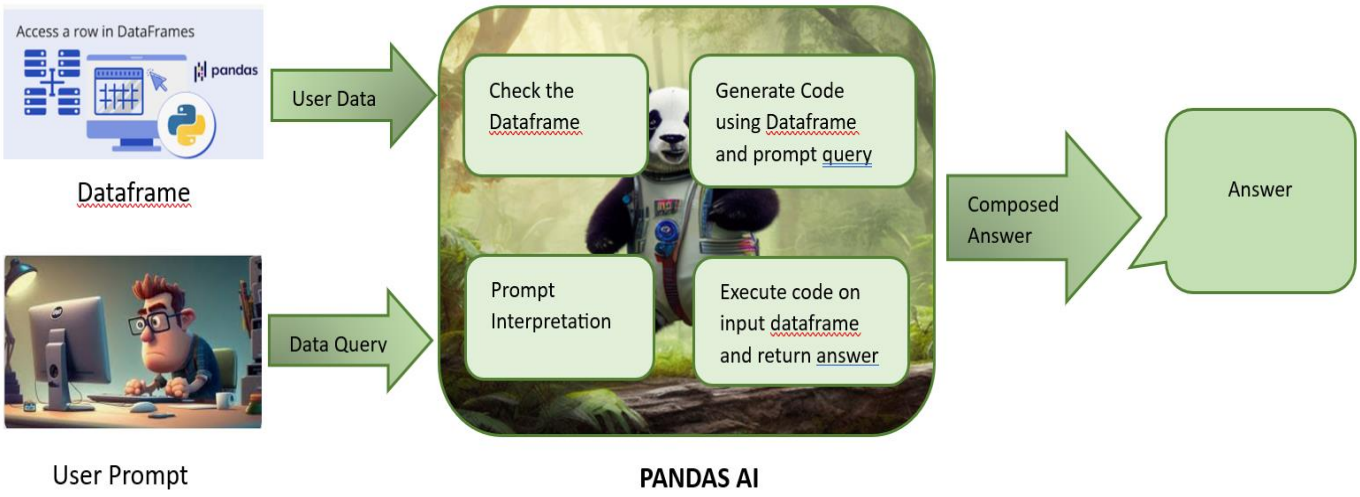
Pandas AI integrates generative artificial intelligence capabilities into Pandas, hence making data frames conversational. It is like a wrapper or an API on top of Pandas library providing the flexibility to ask questions from DataFrame in form of prompt and returning answers quickly.

Data scientists and Analyst used to spend a lot of time in data analysis, cleaning and manipulations. With the help of Pandas AI, this can be done without skimming through the data. They can now move forward with their data analysis. PandasAI should not be used in place of Pandas; rather, it should be utilized in addition to Pandas. You can pose queries to PandasAI, and it will provide responses in the form of Pandas DataFrames, saving time in coding, testing and debugging. It is used with OpenAI API.

Below are some of the points about Pandas AI:

- Pandas AI uses cutting-edge Open AI Large Language Models (LLM).
- One of the input required by this function is a DataFrame, which may be readily given by reading any kind of file or by manually generating it. This is the same data that a user want to analyze.

- Another input required by this function is a user prompt which consist of the actual query that user wants to perform on the input data. This query is in a human like conversational language.
- Pandas AI also has the flexibility to return answer to user query in a conversational way.
- It is also capable of executing complex join queries and drawing insights in form of bar graphs.



- The incoming DataFrame is first inspected internally by Pandas AI, which takes DataFrame columns and some of its starting rows into account.
- Post that, Prompt question entered by user is interpreted in a machine-like language.
- Code generation takes place after analysis of input DataFrame and interpretation of prompt question.
- Once the code is generated successfully based on the prompt interpretation, it is executed on the input DataFrame. Based on the user settings about conversational flag, answer is returned to the end user.
- Pandas AI also has the capability to show the code generated on the basis of prompt to end user.

COMPARISON BETWEEN TRADITIONAL APPROACH AND PANDAS AI

Data analysts and scientists used to invest a lot of effort on cleaning, manipulating, and analyzing data. This can be accomplished without skimming the data with the use of Pandas AI. Hence, making their life easy.

Let's take the example of drawing a histogram of Top 10 Countries according to their population:

Traditional Approach:

```
# -*- coding: utf-8 -*-
"""
Created on Fri Jun  2 01:40:02 2023

@author: aayushi.tayal
"""

import pandas as pd
import matplotlib.pyplot as plt

data = pd.read_csv(r"C:\Users\aayushi.tayal\Downloads\PopulationData.csv",encoding = 'unicode_escape')

colours = ['Red','Blue','Green','Purple','Orange','Yellow','Pink','Grey','Cyan','Olive','Brown','purple']

data = data.sort_values('Population',ascending=False)

# Select the top 10 rows
top_10_rows = data.head(10)
plt.bar(top_10_rows['Country'],top_10_rows['Population'],color=colours)
plt.xticks(rotation=90)
plt.xlabel('Country')
plt.ylabel('Population')
plt.title('Population by Country')
plt.show()
```

Pandas AI Code:

```
In [2]: import pandas as pd
   ...: from pandasai import PandasAI
   ...: from pandasai.llm.openai import OpenAI
   ...: llm = OpenAI(api_token=key)
   ...: pandas_ai = PandasAI(llm)
   ...:
   ...: data = pd.read_csv(r"C:\Users\aayushi.tayal\Downloads\PopulationData.csv",encoding = 'unicode_escape')
   ...: pandas_ai.run(data, prompt='Select Top 10 Countries according to Population and Plot the histogram according to
population showing for each the population, using different color for each bar', is_conversational_answer=False)
```

If we'll look at number of lines of code after doing the basic setup(Importing libraries and reading dataframe) in both approaches, then 9 lines of code is used in traditional approach while Pandas AI has achieved the same results in a single line.

Also, in the traditional approach, we need to take care about the colors palette, but Pandas AI itself take care of all these things.

Capability	Traditional Approach	Pandas AI
Lines of Code	Multiple	1
Color Settings in Histogram	Manual	Automated
Conversational Answer	No	Yes (Flag can be set to True/False)
Code Generation	No	Yes. Code executed in the backend can be shared with the response.
Efforts	High	Low
Automated Data Cleaning	No	Yes
Intelligent Data Visualization	No	Yes
Handling Large Datasets	Limited	Efficiently
Automated Feature Engineering	No	Yes

As you can see, PandasAI offers several features that Pandas lacks, such as automated data cleaning, feature engineering, and intelligent data visualization.

These additional capabilities can significantly streamline your data analysis tasks and empower you to derive deeper insights from your data.

While Pandas is an incredibly powerful and widely used library, PandasAI takes data analysis to the next level by integrating Artificial Intelligence algorithms and automation into the process. It provides a more efficient and intuitive way to handle large datasets, automate repetitive tasks, and unlock hidden patterns in your data.

ADVANTAGES OF PANDAS AI

1. **Conversational AI**

   Pandas AI is a Python library that enhances Pandas with generative AI capabilities to make dataframes conversational. It generates Python code that operates on your data and provides you with the answers you want using large language models (LLMs) like GPT. Using this library, developers will be able to chat with their data in a human friendly language.

2. **Reduced Development Efforts and Faster response**

   Data scientists and Analyst used to spend a lot of time in writing business logic. With the help of Pandas AI, this can be achieved very quickly and without spending much efforts. Hence making the life of Developers and Analyst easy.

3. **Less Dependency on Python Experts**

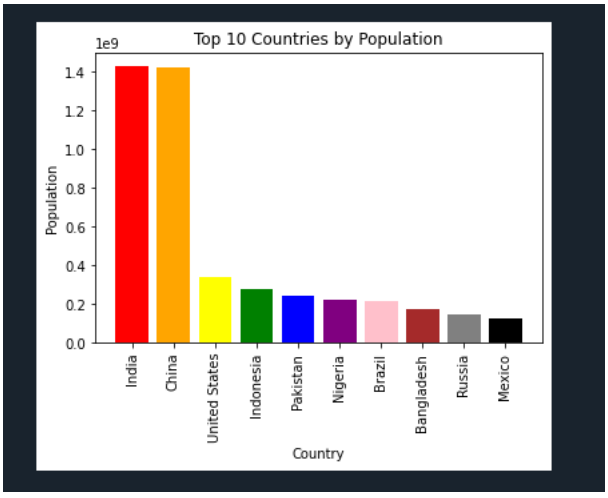
   Since, Pandas AI has the capability of generative AI using which developers can talk to their datasets in a human like language, So, it's not a prerequisite to write lines of code. Hence a newbie having basic understanding of python will be able to proceed with the process.

4. **Ability to generate insights in form of different types of graphs without writing the actual Code**

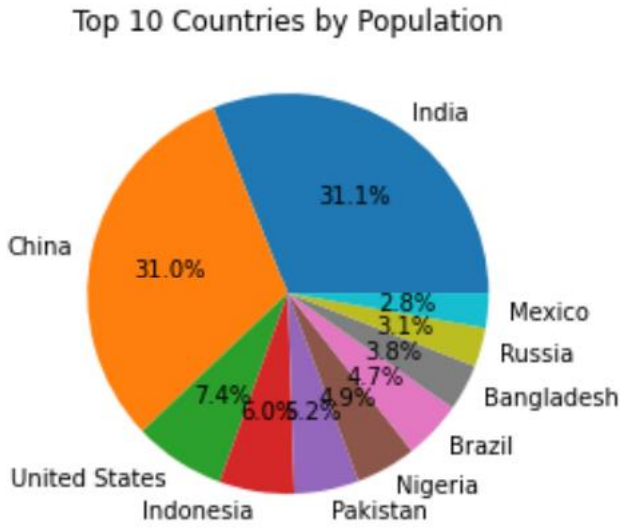
   Pandas AI can help us to generate various types of graphs very quickly. Type of graph insights can be mentioned in the prompt question and Pandas AI will automatically generate it.

   For example:

   For Histogram type of graph, the prompt can be mentioned as: "*Select Top 10 Countries according to Population and* ***Plot the Histogram*** *according to population showing for each the population, using different color for each bar*". The generated graph will be visible like this:



The same view can be generated by in form of pie chart, the prompt can be changed to: "*Select Top 10 Countries according to Population and Plot the **Pie chart** according to population showing for each the population, using different color for each bar*". Below is the generated view:



5. Automation of Data cleaning and Preprocessing

Since Pandas AI has the conversational Ability and the flexibility to perform all the CURD operations and complex queries, So, it has become very easy to analyze, clean and manipulate data. User can change the prompt question and can see the results quickly.

USE CASES OF PANDAS AI

Let's dive into some real-life use cases where this tool can truly shine.

1. **Financial Data Analysis**

Financial data analysis often involves working with large and complex datasets, such as stock market data or financial statements. PandasAI can handle these datasets effortlessly, allowing you to perform in-depth analysis, detect anomalies, and make data-driven investment decisions with confidence. The automated feature engineering and visualization capabilities of PandasAI can also help uncover hidden patterns and trends in financial data, enabling you to gain a competitive edge.

```
import pandasai as pdai

# Load stock market data
stock_data = pdai.read_csv('stock_data.csv')
# Calculate rolling mean of stock prices
stock_data['Rolling Mean'] = pdai.rolling_mean(stock_data['Close'], window=30)
# Visualize stock prices and rolling mean
pdai.plot_line_chart(stock_data, x='Date', y=['Close', 'Rolling Mean'])
```

2. **Customer Segmentation**

Understanding your customers and their behavior is crucial for businesses in various industries. With PandasAI, you can easily segment your customer base based on various attributes and characteristics, such as demographics, purchase history, or browsing behavior. By leveraging the automated feature engineering capabilities of PandasAI, you can extract valuable insights and create targeted marketing campaigns to improve customer satisfaction and drive revenue growth.

```
import pandasai as pdai

# Load customer data
customer_data = pdai.read_csv('customer_data.csv')
# Perform customer segmentation based on purchase history and demographics
customer_segments = pdai.segment_customers(customer_data, features=['Purchase History', 'Age', 'Gender'])
# Visualize customer segments
pdai.plot_pie_chart(customer_segments, labels='Segment', values='Count')
```

3. **Healthcare Analytics**

In the healthcare industry, analyzing vast amounts of patient data is essential for making informed medical decisions and improving patient outcomes. PandasAI can streamline the analysis process, allowing healthcare professionals to extract valuable insights from electronic health records, clinical trial data, or medical imaging data. The ability to handle large datasets and automate certain data cleaning and feature engineering tasks makes PandasAI a valuable tool in healthcare analytics.

```
import pandasai as pdai

# Load patient data
patient_data = pdai.read_csv('patient_data.csv')
# Perform analysis on patient data
average_heart_rate = pdai.mean(patient_data['Heart Rate'])
diabetes_patients = pdai.filter(patient_data, condition="'Diabetes == 'Yes'"')
# Visualize average heart rate
pdai.plot_bar_chart(x=['All Patients', 'Diabetes Patients'], y=[average_heart_rate, len(diabetes_patients)])
```

4. **Data Transformations**
The process of converting or manipulating data from one form to another and making it suitable for further analysis, visualization and modelling purpose takes a huge amount of time. Depending on the data changes, developers also need to spend time in checking the code and writing it again. Since Pandas AI has the capability of reading, cleaning, updates and transforming it, So, this process has become comparatively easy and any change in the transformation will require change in the prompt question only.

5. **Data Validation**
One of the crucial aspect of Data management is data validation which ensures that the data entered into the system is accurate, consistent and meets the standards of the data. It also ensures that the data is complete which helps us in maintaining the quality of data.

There are various types of validation which are required to maintain the data quality like dirty data checks, Length validation, Regex validation, Format Validation etc. Pandas AI is helpful in doing these validations. This can be further applied to Healthcare and Marketing analytics.

6. **Insights Generation for HealthCare, Marketing Industries and Supply Chain Management**
As mentioned in the benefits of PandasAI, it is very easy to generate various types of insights like Pie Chart, Histogram, Line chart etc. These can be proved very beneficial for Healthcare Industry for drawing various insights like drug usage according to the historical trend, generating insights of various diseases over time. These types of insights help us in analyzing the trend, Future prediction, and planning.

Similarly for a marketing industry, we can analyze the customers data and generate the insights on the purchase history for a product. These insights can be very helpful for predicting the future planning for the product.

7. **Knowledge Management for Students and Professionals**
With the Generative AI feature of Pandas AI, it also has the capability of generating the code that has been executed for a particular prompt question.

Using this feature of the library, Students and Professionals can increase their knowledge base. They can generate the code to solve any problem, analyse it and learn. Pandas AI is currently in the development mode and not preferred for the production applications, but using this feature, developers can generate the code for any feature or functionality and use it in their code base.

Let's take the example of drawing a pie chart:

Below is the Pandas AI code for generating a Pie chart view of Top 10 countries according to their population:

```
   ...: import pandas as pd
   ...: from pandasai import PandasAI
   ...: from pandasai.llm.openai import OpenAI
   ...: llm = OpenAI(api_token=key)
   ...: pandas_ai = PandasAI(llm)
   ...:
   ...: data = pd.read_csv(r"C:\Users\aayushi.tayal\Downloads\PopulationData.csv",encoding =
'unicode_escape')
   ...: pandas_ai.run(data, prompt='Select Top 10 Countries according to Population and Plot the Pie Chart
according to population showing for each the population, using different color for each bar',
is_conversational_answer=False, show_code=True)
```

If you'll notice in the above code, we have setup a flag with the name show_code to True for generating the code. Below is the code generated by the library:

```
In [6]: import pandas as pd
   ...: import matplotlib.pyplot as plt
   ...:
   ...: # Creating the dataframe
   ...: data = {'Country': ['China', 'Indonesia', 'Pakistan', 'India', 'United States'],
   ...:         'Birth Rate': [15.9, 21.6, 18.7, 21.6, 12.1],
   ...:         'Population': [6402133538, 277534122, 240485658, 240485658, 7298971515]}
   ...: df = pd.DataFrame(data)
   ...:
   ...: # Sorting the dataframe by population in descending order and selecting top 10 countries
   ...: top_10 = df.sort_values('Population', ascending=False).head(10)
   ...:
   ...: # Plotting the pie chart
   ...: plt.pie(top_10['Population'], labels=top_10['Country'], autopct='%1.1f%%')
   ...: plt.title('Top 10 Countries by Population')
   ...: plt.show()
```

This can also help the users in understanding the most optimized way of implementing any functionality.

8. **Database integration for automated SQL Queries**
Since Pandas AI works on dataframe which is a tabular kind of structure having multiple rows and columns, So it is valid for database tables as well. We can create a python script to integrate database with python console and process SQL queries in Human friendly language.

DATA CACHING

PandasAI also has the flexibility to store the result of previously executed queries in a file named cache.db stored in root directory of the project. The cache is a SQLite database, and can be viewed using any SQLite client. This is useful for below reasons:

1. It allows the user to quickly retrieve the results of a query without having to wait for the model to generate a response.
2. It cuts down on the number of API calls made to the model, reducing the cost of using the model.

Default Mode

By Default, Cache is enabled**.**

Disabling the Cache

The cache can be disabled by setting the cache parameter to False when creating the PandasAI object:

pandas_ai = PandasAI(llm, cache=False)

Clearing the Cache

The cache can be cleared by deleting the cache.db file. The file will be recreated automatically when the next query is made. Alternatively, the cache can be cleared by calling the clear_cache() method on the PandasAI object:

pandas_ai.clear_cache()

DATA SECURITY AND PRIVACY

Since, PandasAI is internally driven by OpenAI. Currently, there are no rules and regulations for the same. But there is a certain feature offered within the library which can take care of the Security aspect. It allows you to anonymize and enforce privacy to your data so that actual data is not sent to LLM provider.

In case we don't want to enforce the privacy but want to anonymize it, then it will cause the data to be masked with some random numbers before sending it to LLM provider.

CONCLUSION

PandasAI is a game-changer in the world of data analysis. With its advanced AI capabilities and seamless integration with Pandas, it empowers data analysts and scientists to tackle complex tasks more efficiently and effectively. Whether you're handling large datasets, automating feature engineering, or visualizing data, PandasAI is your go-to tool.

REFERENCES

https://levelup.gitconnected.com/introducing-pandasai-the-generative-ai-python-library-568a971af014

https://github.com/gventuri/pandas-ai

https://medium.com/@fareedkhandev/pandas-ai-the-future-of-data-analysis-8f0be9b5ab6f

https://www.kdnuggets.com/2023/05/pandas-ai-generative-ai-python-library.html

Python ETL vs. ETL Tools. If you're researching ETL solutions you… | by Sean Knight | Towards Data Science

https://www.geeksforgeeks.org/pandas-ai/

https://www.analyticsinsight.net/pandas-ai-is-it-the-future-of-data-analysis/