Automating the Data Science Analysis process with AI using Fundamental Components of Data Science

by

Madeline E. Kaufman

Data Science Capstone

Arizona State University

DAT 490

Professor (or Dr.) Marko Samara

February 2024

Table of Contents

Executive Summary

The primary objective of this project is to automate the data science analysis process by leveraging Artificial Intelligence (AI). By generalizing the fundamental components of data science this AI technology can accommodate a diverse range of datasets. These functions cover a broad spectrum of data science operations including exploratory data analysis to determine the proper analysis to perform. By doing so, I will try to remove the manual, repetitive tasks often associated with data preprocessing, analysis, and interpretation, making the process more efficient and accessible.

At the core of this approach is the development of AI algorithms capable of performing a comprehensive suite of data science operations. This includes, but is not limited to, exploratory data analysis (EDA), which is crucial for understanding the underlying characteristics of the data, identifying patterns, outliers, and anomalies, and determining the most appropriate analytical techniques to apply. The AI system is designed to automatically conduct EDA, leveraging machine learning and statistical models to extract insights without explicit human direction for every step.

A key advantage of automating the data science analysis process through AI is the potential for scalability and adaptability. As the system encounters a diverse range of datasets and analysis scenarios, it can learn and improve over time, enhancing its accuracy and efficiency. This learning process not only refines the system's analytical capabilities but also its ability to fine-tune analysis parameters and interpret results in a contextually relevant manner.

Project Plan

Research Questions and Hypothesis

Question One: What are the fundamental components of data science and is it feasible to distill these components into a set of generic, reusable functions that an AI can utilize?

Research Question One Hypothesis

Hypothesis: Creating a comprehensive library of specific but generic reusable data science functions, encompassing graph generation and advanced data analysis techniques, will significantly improve the efficiency and standardization of data science processes across various scenarios.

The development of a versatile and inclusive toolkit will lead to measurable improvements in how data science tasks are performed, particularly in terms of time efficiency and consistency in different use cases.

Question Two: Is it possible to develop a comprehensive set of guidelines for each fundamental data science component, providing context and directives on their application in various data science scenarios? In layman's terms: Can we create an "instruction manual" to teach an AI when best when/why/how we would want to use different functions in different data science use cases.

Research Question Two Hypothesis

Hypothesis: An extensive, user-friendly manual detailing the application of fundamental data science components will enhance the AI system's ability to understand and apply these components effectively, resulting in more accurate and context-appropriate data science solutions.

The creation of a detailed manual will bridge the gap between technical functionality and practical application, leading to a more informed AI system that can make better decisions regarding the selection and application of data science tools in various scenarios.

Question Three: After the creation of a library of data science functions and corresponding usage instructions, can an advanced system be designed to automate the workflow of diverse data science tasks through AI-driven orchestration?

Research Question Three Hypothesis

Hypothesis: By integrating a library of data science functions and user guidelines into an AI system and refining it with various data inputs, the system will achieve a high level of automation in data science tasks, thereby increasing the overall efficiency, accuracy, and scalability of data science workflows.

The combination of a well-developed function library, comprehensive guidelines, and iterative refinement with diverse data will enable the AI system to autonomously perform a wide range of data science tasks more effectively, leading to advancements in how data analysis processes are conducted.

The Data

The Data I will use to create this system will come from all of my knowledge learned while taking courses at ASU, along with other researched data from official sources. I will also be using data that the business has collected to test on the software.

Further data for testing will be gathered on various data providing platforms like Kaggle. An example of one of these datasets is Amazon Sales Data. The goal here will be to find datasets which have been previously analyzed by other people, what part of the data they chose to analyze and with which methods. This will be crucial in testing performance and accuracy of the designed AI model. By comparing outputs and analysis techniques it will also help with improving the AI to be able to train it on which data analysis to perform on which kind of data. Because of the size and length of time required to complete the project I will choose a few data analysis techniques to specifically test.

https://www.kaggle.com/datasets/thedevastator/unlock-profits-with-e-commerce-sales-data

Methodology

Research Question One

I will be trying to compile and create specific but generic reusable data science functions that can be utilized by an AI system. This library will encompass a wide range of functions, including various types of graph generation (using tools like matplotlib, Plotly, ggplot, seaborn), and advanced data analysis techniques for example linear and multilinear regression, hypothesis

testing, logistic regression, etc. The objective is to create a robust toolkit that simplifies and standardizes data science processes across different scenarios.

Research Question Two

I will create an extensive, user-friendly manual or set of guidelines that detail the application of fundamental data science components in various scenarios. This manual will then become the baseline for prompting the AI system, enabling it to understand the context, purpose, and methodology for applying specific data science functions and tools. The focus will be to bridge the gap between technical functionality and practical application, making it easier for the AI to make informed decisions about which tools to use in different data science tasks and scenarios.

Research Question Three

After compiling the user prompting I will feed in various data and design/refine the AI system to be capable of automating a wide array of data science tasks by intelligently orchestrating the use of the developed library of data science functions and guidelines. This system aims to enhance efficiency, accuracy, and scalability in data science workflows, paving the way for more sophisticated and autonomous data analysis processes.

The final step will be to test the system for outputs and accuracy. To compile different components of a complete data analysis on data.

Campaign Implementation

To implement my ambitious project of revolutionizing data science through AI, I will embark on a phased approach.

The first phase involves the development of a comprehensive library of reusable data science functions. This library, encompassing tools for graph generation and advanced analysis techniques, will form the backbone of our AI's analytical capabilities. I will source testing data from diverse platforms, including our internal datasets and external sources like Kaggle, to ensure robustness and versatility.

Simultaneously, I will develop an extensive manual that will serve as an instructional guide for the AI, detailing the practical application of these data science components across various scenarios. This guide aims to bridge the gap between technical functionality and real-world usage, ensuring that our AI can make informed decisions about tool selection and application.

In the final phase, I will integrate these components into our AI system, focusing on refining its ability to automate a wide array of data science tasks. This phase will involve rigorous testing using both internal and external datasets, including complex scenarios such as the analysis of Amazon Sales Data. The goal here is to enhance the AI's capability in handling diverse data types and structures, from boolean values to categorical data.

Throughout the implementation, I will continually assess the system's performance, focusing on efficiency, accuracy, and scalability. The successful execution of this campaign will enhance our AI's ability to visualize and analyze legal data. This project aligns perfectly with our mission to make legal information accessible and comprehensible, thereby democratizing data understanding through advanced AI solutions.

- Enhanced AI capability for intuitive, visual representation of complex legal data.
- Improved efficiency and accuracy in data science tasks.
- Positioning the company as a frontrunner in AI-driven data analysis innovation.
- Broad applicability across industries needing data-driven insights.

Literature Review

Automated data analytics is a transformative approach in the field of data science, reshaping how data is analyzed and interpreted. This evolving paradigm alleviates the burden of repetitive and mundane tasks from data professionals, potentially enabling them to focus on more intricate and innovative aspects of their work. This literature review aims to delve into the intricacies of automated data analytics, exploring its definition, benefits, practical implementation, and real-world examples. By illuminating these facets, the review seeks to provide a comprehensive understanding of automated data analytics and its significant role in enhancing business operations.

Automated Data Analytics: Definition and Scope

Automated data analytics is defined as the application of technology to perform data analysis tasks with minimal human intervention. This concept extends beyond mere task automation; it encompasses a range of applications from automating entire data processes and business intelligence dashboards to developing self-governing machine learning models. The scope of automated analytics is broad, encompassing tasks like data discovery, preparation, replication, and maintenance, and therefore reducing the traditional data handling methodologies and opening new avenues for efficient data management.

Benefits of Automated Data Analytics

Automated data analytics brings forth significant advantages, among them being speeding up the reporting process. By automating parts of the data pipeline, the time needed

from request to delivery of analytical products is drastically reduced. This automation not only conserves time but also translates into substantial cost savings, as it can lessen the workload on data professionals. Furthermore, automation can enhance process reliability and accuracy, mitigating the risks associated with human errors and building more robust and full-proof systems.

When and How to Automate Data Analytics

Automation should be considered when the task in question is of high value, repetitive, and prone to errors. The decision to automate should be based on a task's potential to resolve business issues and its capacity to save time. The implementation ranges from partial automation, which aids existing procedures, to full automation, where decisions are made in real-time without need for human intervention.

Examples and Case Studies of Automated Analytics

Real-world applications of automated analytics are diverse and impactful. For instance, automating data collection streamlines the process of acquiring and refreshing data, freeing up resources for more strategic tasks. Automated reporting and dashboards simplify the creation of essential business insights, while business intelligence automation enables the efficient analysis of various business metrics. Additionally, the automation of machine learning models and big data processes demonstrates the advanced capabilities of AI in predicting trends and detecting anomalies, thereby providing businesses with a competitive edge in data-driven decision-making.

Data Analysis Automation is still an emerging field as is AI, AI Research and AI Development. There are still many mysteries as to the power of technology and how it will impact all aspects of life as the future unfolds. A common misunderstanding and something that researchers are looking into is whether or not data scientists feel their jobs are at risk. This is a theme with almost any and all talks about AI and job security. Most came to the same conclusion that if anything automation will only help and not harm their job security. A good example of this is yes, AI could automate processes and become very good at it, but at the same time a human will still need to process, understand, and implement/make decisions based on those data analysis. The world is still just learning about AI and its current capabilities.

It's important to note that the goal of this project is not to create a software that will replace the job of a human but rather create a software that is accessible to the commonwealth. Not everyone has knowledge of how to perform data analysis and data science yet this could potentially bridge the gap to better help and educate individuals on certain data. I believe that if used for good, AI could also potentially help tremendously in many different aspects especially to those who don't have direct access to information or might lack the ability to understand said information. A prime example is our software AskAbe which was initially built so that the average person without a law degree could start understanding the laws in which our country is built. Data analysis automation would further this goal, providing people with the tools they need to simplify complex information.

Final Research Questions

What are the fundamental components of data science and is it feasible to distill these components into a set of generic, reusable functions that an AI can utilize?

Goal: Develop a comprehensive library of generic, reusable data science functions that can be effectively utilized by AI systems. This library would encompass a wide range of functions, including various types of graph generation (using tools like matplotlib, plotly, ggplot, seaborn), and advanced data analysis techniques (like linear and multilinear regression, hypothesis testing, logistic regression). The objective is to create a robust toolkit that simplifies and standardizes data science processes across different scenarios.

Is it possible to develop a comprehensive set of guidelines for each fundamental data science component, providing context and directives on their application in various data science scenarios? In layman's terms: Can we create an "instruction manual" to teach an AI when best when/why/how we would want to use different functions in different data science use cases.

Goal: Create an extensive, user-friendly manual or set of guidelines that detail the application of fundamental data science components in various scenarios. This manual would serve as a reference for the AI system, enabling it to understand the context, purpose, and methodology for applying specific data science functions and tools. The focus is on bridging the gap between technical functionality and practical application, making it easier for AI to make informed decisions about which tools to use in different data science tasks and scenarios.

After the creation of a library of data science functions and corresponding usage instructions, can an advanced system be designed to automate the workflow of diverse data science tasks through AI-driven orchestration?

Goal: Design an advanced AI-driven system capable of automating a wide array of data science tasks by intelligently orchestrating the use of the developed library of data science functions and guidelines. This system aims to enhance efficiency, accuracy, and scalability in data science workflows, paving the way for more sophisticated and autonomous data analysis processes.

Exploratory Data Analysis

For my exploratory data analysis I compiled a list of all potential functions to be used within the data science automation AI. Although I didn't end up getting to use and experiment with all of them, I think this is a good list and hopefully will be able to expand and make the project better in the future to add more complex data analysis tasks.

Exploratory Data Analysis refers to the crucial process of performing initial investigations on data to discover patterns to check assumptions with the help of summary statistics and graphical representations.

Importing Python Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from mpl_toolkits.basemap import Basemap as Basemap
from matplotlib.colors import rgb2hex, Normalize
from matplotlib.patches import Polygon
from matplotlib.colorbar import ColorbarBase
```

Reading Dataset

```
data = pd.read_csv("data_file_name.csv")
```

For timing sake I will read in the data manually for each different data set. Generalizing this in the future so that any data set can be read into the machine.

Analyzing Data

- shape() : number of observations and features
- head() : display top 5 observations
- tail() : last 5 observations
- info() : understand the data type and information about data
- nunique() : check for duplication
- isnull() check for missing values
    - isnull().sum() : number of missing record in each column
    - (data.isnull().sum()/(len(data)))*100 : calculate percentage of missing values in each column

Data Reduction

- Drop Columns that Aren't useful: df.drop()
- Drop Rows missing values: df.dropna()
- Missing Data?

Feature Engineering

Goal of feature engineering is to create meaningful data from raw data.
Creating features:

For example if we have a Soda dataset and the 'Brand' and 'Type' are both in the same column, we would split this and create a new column Brand, Type. So if there was one column let's say the data values were Diet Coke by Coca-Cola, Diet Pepsi by Pepsi we would split these to make a column ['Brand'] and the data values would be Coca-Cola, Pepsi and a new column ['Soda Type'] with data values Diet Coke, Diet Pepsi

Data Cleaning/Wrangling

- Data Type Conversion
- "Address Matching" - AZ vs. az vs. Arizona (cell values same meaning but different format - need to standardize this.

This is very important for the AI system as the model will get confused without standardization unless specifically taught in each case in which these three differences mean the same thing.

Statistics Summary

Statistics summary gives a high-level idea to identify whether the data has any outliers, data entry error, distribution of data such as the data is normally distributed or left/right skewed

describe()

- Count
- Mean
- Standard Deviation
- Median
- Mode
- Min/Max
- Range

Separating data types for analysis - this will greatly help the AI know exactly which columns are of what data type for easier analysis

```
cat_cols=data.select_dtypes(include=['object']).columns
num_cols = data.select_dtypes(include=np.number).columns.tolist()
print("Categorical Variables:")
print(cat_cols)
print("Numerical Variables:")
print(num_cols)
```

Data Visualization

Univariate Analysis

Analyzing the data by taking one variable at a time to show data patterns

Data Visualization Might Include

Categorical Variable

- Count Plot
- Bar Chart
- Pie Plot

Numerical Variable

- Histogram, Box Plot, Density Plot

Bi-Variate Analysis

Analyzing two variables to understand the relationship between them, if they are related and if the variable is independent and dependent.

Data Visualizations Might Include

Categorical Variable

- Stacked bar chart - for categorical

Numerical Variable

- Pair Plots
- Scatter Plots

Multivariate Analysis

Uses multiple variables (more than 2)

Data Visualizations Might Include

- Heat Maps

Ways to draw charts and plots in python:

Matplotlib is a Python 2D plotting library used to draw basic charts.

Seaborn is also a python library built on top of Matplotlib that uses short lines of code to create and style statistical plots from Pandas and Numpy

Choosing Type of Data Analysis

Regression Analysis

Describes the relationship between a set of independent and dependent variables.

Linear Regression

Continuous Dependent Variables

Linear regression is a fundamental statistical method used in data science to understand and predict the relationship between two variables. Let's say we're analyzing the impact of advertising spend on sales revenue. In this scenario, advertising spend is the independent variable (predictor), and sales revenue is the dependent variable (outcome we're trying to predict).

Modeling the Relationship:

Linear regression will help us model this relationship. The technique assumes that the relationship between the advertising spend (X) and sales revenue (Y) can be described by a straight line (linear). This line can be expressed by the equation $Y = aX + b$, where 'a' represents

the slope of the line (how much sales increase with each unit increase in advertising) and 'b' represents the y-intercept (the expected sales when the advertising spend is zero).

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Advanced Types of linear regression (not going to be implemented with this version but something to consider for future implementation)

Ridge Regression

Lasso Regression

Partial Least Squares (PLS) Regression

Nonlinear Regression

Continuous Dependent Variables

Nonlinear regression is a type of regression analysis used when data fits a model better than a linear approach. It is used to model complex relationships between a dependent variable and one or more independent variables.

Example and Modeling the Relationship:

Suppose you're studying the growth rate of bacteria under varying temperatures. Here, the temperature is the independent variable, while the bacteria growth rate is the dependent variable. Unlike linear regression, the relationship between these variables might not be a straight line. Instead, it could follow a curve, like a logistic growth curve.

In nonlinear regression, this relationship might be modeled with an equation like Y = a / (1 + be^(-cX)), where 'a', 'b', and 'c' are parameters of the model, and 'e' is the base of natural logarithms. This equation represents a logistic curve, where the growth rate initially increases rapidly with temperature, then levels off.

Multi-Linear Regression

Continuous Dependent Variables

Multivariate linear regression is an extension of linear regression, where you predict a dependent variable based on multiple independent variables.

Example and Modeling the Relationship:

Consider you're analyzing the factors that affect house prices. Here, the house price is the dependent variable, and the independent variables could be the size of the house, the number of bedrooms, the age of the house, and so on.

In multivariate linear regression, the relationship between the dependent variable (Y) and independent variables (X1, X2, X3, ..., Xn) is modeled as Y = a0 + a1X1 + a2X2 + ... + anXn + b. Here, 'a0, a1, ..., an' represent the coefficients that quantify the impact of each independent variable on the dependent variable, and 'b' is the intercept. The model tries to fit a hyperplane to the data points in this higher-dimensional space.

Other Types of Regression to Note:

1. Binary Logistic Regression
    a. Categorical Dependent Variables
2. Ordinal Logistic Regression
    a. Categorical Dependent Variables
3. Nominal Logistic Regression
    a. Categorical Dependent Variables
4. Poisson Regression
    a. Count Dependent Variables

K-Nearest Neighbors (KNN)

KNN is used for classification and regression. It classifies a data point based on how its neighbors are classified. KNN finds the 'k' nearest data value in the dataset and averages their values (for regression) or picks the most common category (for classification).

Naive Bayes

Naive Bayes is a probabilistic classifier based on applying Bayes' theorem with strong independence assumptions between features.

NEURAL NETS

Neural networks are a set of algorithms, modeled loosely after the human brain, designed to recognize patterns. They interpret sensory data through machine perception, labeling, and clustering.

Example and Modeling: In image recognition, a neural network learns to identify and classify objects in images by processing the data through interconnected layers of neurons.

Clustering

Clustering Methods

- Connectivity-based Clustering (Hierarchical clustering)
- Centroids-based Clustering (Partitioning methods)
- Distribution-based Clustering
- Density-based Clustering (Model-based methods)
- Fuzzy Clustering
- Constraint-based (Supervised Clustering)

Clustering Algorithms

- K-Means clustering
- Mini batch K-Means clustering algorithm
- Mean Shift
- Divisive Hierarchical Clustering
- Hierarchical Agglomerative clustering
- Gaussian Mixture Model
- DBSCAN
- OPTICS
- BIRCH Algorithm

Sentiment Analysis

Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text to determine whether the writer's attitude is positive, negative, or neutral.

Other functions/Assessments to consider

- Regularization
- Bias-Variance Tradeoff
- Cross-Validation
- Maximum Likelihood Estimation (MLE) & Optimization
- Hypothesis Testing
- Probability
- Train/Test Models
- Confidence

Data Visualizations

Plot Types

- Scatter
- Bar
- Stem
- StackPlot
- Stairs
- Histogram
- Box plot
- Errorbar
- Hexbin
- Pie
- Contour
- Streamplot
- 3d Graphs
- Geographic Maps

Example With Selected testing data Amazon Sales Report, Height vs. Weight, Real Estate to show EDA in action and examples of some regression methods.

Methodology

Research Question One - What are the fundamental components of data science and is it feasible to distill these components into a set of generic, reusable functions that an AI can utilize?

The first stage in my project methodology is to create the conceptual framework. The primary goal at this stage will be to develop a deep understanding of existing data science functions and their diverse applications. This foundational knowledge is crucial for the effective integration of these functions into the AI system. To achieve this, comprehensive research will be conducted. This research will delve into various data science functions, with a specific focus on graph generation, regression analyses, hypothesis testing, and other relevant methods. The overall research will aim to cover a broad spectrum of sources, including academic journals, industry reports, and case studies. The output will be a well-structured conceptual framework. The framework will outline the types of data science functions that are most relevant and beneficial for the project. It will serve as a guiding document for the subsequent development stages. This process has been started and began outlining in the EDA.

The next stage is the development of Data Science Functions. The aim here is to develop a versatile and comprehensive library of generic, reusable data science functions. This will be achieved by programming a variety of functions using Python and several data visualization libraries such as matplotlib, Plotly, ggplot, and seaborn. A key focus will be to ensure that each function is adaptable and can be applied to different datasets and scenarios, enhancing the utility

and flexibility of each particular one. The output will be a full program of code that is a comprehensive library of data science functions. This library will be accompanied by thorough documentation and examples of use cases, facilitating easy adoption and application in various contexts.

The third and final stage of tackling this research question is evaluation and refinement of data science functions. The objective is to rigorously test and refine the developed functions. This will also be merged with research question 2 and 3 as the library and functions will need evaluation and fine-tuning at each step of this process. This will involve applying the functions to a diverse range of datasets, such as those available on Kaggle, as well as internal business data of our company. This testing will particularly be crucial for evaluating the performance and adaptability of the functions in real-world scenarios. The final output is a refined and robust library of data science functions. This library will be ready for seamless integration into the broader AI system, having been tested and optimized for practical use. As the process continues to develop over time this step will be revisited as needed to make sure that the data science library is working properly and efficiently by being generic enough to handle a variety of different data and not generic enough that it can't perform what it needs to.

Research Question Two - Is it possible to develop a comprehensive set of guidelines for each fundamental data science component, providing context and directives on their application in various data science scenarios? In layman's terms: Can we create an "instruction manual" to teach an AI when best when/why/how we would want to use different functions in different data science use cases.

The first stage of the second research question is manual development and prompt engineering of the AI model. This step in short is like teaching someone who knows nothing about data science, all about it. This will start by creating a detailed and comprehensive manual that guides the AI system's interaction with the data science components. This is called prompt engineering. This involves the development of clear “guidelines and instructions” like a manual that I will then use to train a new AI model using OpenAI’s LLM. These guidelines will detail the circumstances under which various functions should be used, their intended purposes, and the methodologies for their application in different scenarios. The output is a comprehensive manual that serves as an 'instruction book' for the AI system, guiding its interaction with the data science components.

The next stage will be to integrate a Feedback Loop into the AI model design. This will ensure the AI system effectively utilizes the manual and can adapt based on its usage. This involves integrating feedback mechanisms into the system. These mechanisms will allow the AI to learn from its interactions with the manual and to adapt its understanding and application of the data science tools over time. The final output is an AI system that is not static but continually improves its understanding and application of data science tools based on ongoing interaction with the manual.

Research Question Three - After the creation of a library of data science functions and corresponding usage instructions, can an advanced system be designed to automate the workflow of diverse data science tasks through AI-driven orchestration?

The final step will be complete system development and AI integration. The objective here is to develop an complete AI system capable of autonomously performing data science

tasks. This will be achieved by integrating the library of data science functions into an AI framework. The “system” (program) will utilize machine learning techniques to enable autonomous selection and application of the right data science tools. The output from this stage is hopefully an initial version of the AI system, which is ready for preliminary testing.

The next step is to vigorously test the AI system's performance in real-world scenarios and refine it based on the findings. This involves using various datasets, such as Amazon Sales Data, to test the system's ability to autonomously perform data analysis. The results will be compared against established benchmarks to assess its performance. This will provide valuable insights into the system's performance, including identification of areas that require improvement. In turn we can go back to research question two and research question three (stage three) to refine the data science functions as well as better train the AI model.

Lastly, the final step will lean toward scalability and adaptability testing. In this step I will break down the full process to ensure the AI system is capable of handling large datasets and can adapt to new, unforeseen scenarios. The model will be tested with increasingly complex and large datasets. Additionally, it will be introduced to new types of data and analysis scenarios to assess its adaptability. The final output is a scalable and adaptable Data Analysis AI, that will hopefully be capable of handling a range of data science tasks across various contexts.

Modeling Techniques to be used

Prompt Engineering

Prompt engineering is the process of designing and refining inputs given to an AI model. In particular - language models. This is to hopefully help the AI produce the most accurate and relevant outputs wanted. The prompt essentially guides the AI when given an user's input or

question, so that it can formulate a response. This response of the model depends heavily on how this prompt is structured. The importance of prompt engineering when developing this software is crucial to its performance success. Effective prompt engineering will significantly enhance the AI's overall performance. More specifically, it will be critical to tailor the prompt to guide the model toward producing the desired output in our case complete data analysis. This process will involve experimenting with different phrasings, structures, and types of information included in the prompt. More specifically to this means providing context and specific instructions within the prompt to narrow down the AI's focus to data science functionality and analysis.

For example in the model, changing a prompt from:

"*Imagine you're trying to predict your grade in school based on how many hours you study. You plot this on a graph, with the amount of study hours on the x-axis and grades being on the y-axis. You might see on the graph that the more you study, the better your grades tend to be, and then you might draw a straight line through these points. This line is basically what linear regression is - it's like drawing a line through points on a graph to predict stuff.*"

to be more like:

Linear regression is a fundamental statistical method used in data science to understand and predict the relationship between two variables. Let's say we're analyzing the impact of advertising spend on sales revenue. In this scenario, advertising spend is the independent variable (predictor), and sales revenue is the dependent variable (outcome we're trying to predict).

Data Collection: We first will collect historical data showing how different levels of advertising spend have corresponded to sales revenue.

Modeling the Relationship: Linear regression will help us model this relationship. The technique assumes that the relationship between the advertising spend (X) and sales revenue (Y) can be described by a straight line (linear). This line can be expressed by the equation $Y = aX + b$, where 'a' represents the slope of the line (how much sales increase with each unit increase in advertising) and 'b' represents the y-intercept (the expected sales when the advertising spend is zero).

The Analysis and Prediction: By applying linear regression, we calculate the values of 'a' and 'b' that best fit our historical data. This allows us to predict the sales revenue for any given level of advertising expenditure. For example, if our model finds that every $1,000 increase in advertising leads to a $5,000 increase in sales, a business can use this information to make informed decisions on their advertising investments.

Assessing the Accuracy: It's important to assess the accuracy of our model. Looking at metrics like R-squared, which tells us how much of the variance in sales can be explained by our advertising expenditure. A higher R-squared indicates a better fit of our model to the data.

Assumptions and Limitations: Linear regression is powerful, but it has its limitations. It assumes a linear relationship and is sensitive to outliers. It also doesn't account for complex scenarios where other variables might influence the relationship we're studying.

WIll dramatically affect the AI capabilities and functionality. In the first example, the AI system will likely generate what's called "Hallucinations" - this is when it is given some context but not

enough to fully understand the topic in hand. A hallucination would be an output generated by the AI that is unrelated to the data provided at all and will be highly inaccurate and even potentially not related to linear regression analysis at all.

Whereas in the second example, this is a highly detailed explanation of linear regression with examples and specific context clues. This will better guide the AI model and train it to fully understand exactly what to do and each step along the way. This would provide a much better output that is more likely to be accurate. It will also more likely be able to perform linear regression on a broader range of different data sets and types because it has specific requirements for performing linear regression like x and y variables (two variables compared to one another). It also provides context to assessing the accuracy of performing such an analysis leading to being able to perform a more complete and in-depth linear regression analysis.

Machine Learning

Machine learning involves training algorithms to make predictions or take actions based on data. Implementing machine learning will allow the AI system to learn from data, adapt to new circumstances, and improve over time without needing to be programmed for every specific scenario. For example by using ML the AI system can be trained on historical datasets where it can learn to identify patterns and relationships between different variables. This will teach the AI how to process raw data, deal with inconsistencies, and extract meaningful features that are essential for making predictions. The AI system will also learn to apply the most appropriate predictive models based on the data characteristics and the analysis objectives. It can gain the ability to discern which functions are likely to yield the most accurate predictions in different scenarios. The AI system will not only be trained on making these predictions but also

interpreting the results, understanding the significance of different predictive variables and the confidence in its predictions. The AI will then hopefully be able to provide insights and actionable information from the predictive analysis, not just raw predictions. It will also allow for implementing the mechanisms needed for ongoing learning of the AI from new data, so it can adapt to changing patterns in data over time.

Data Visualizations & Analyses

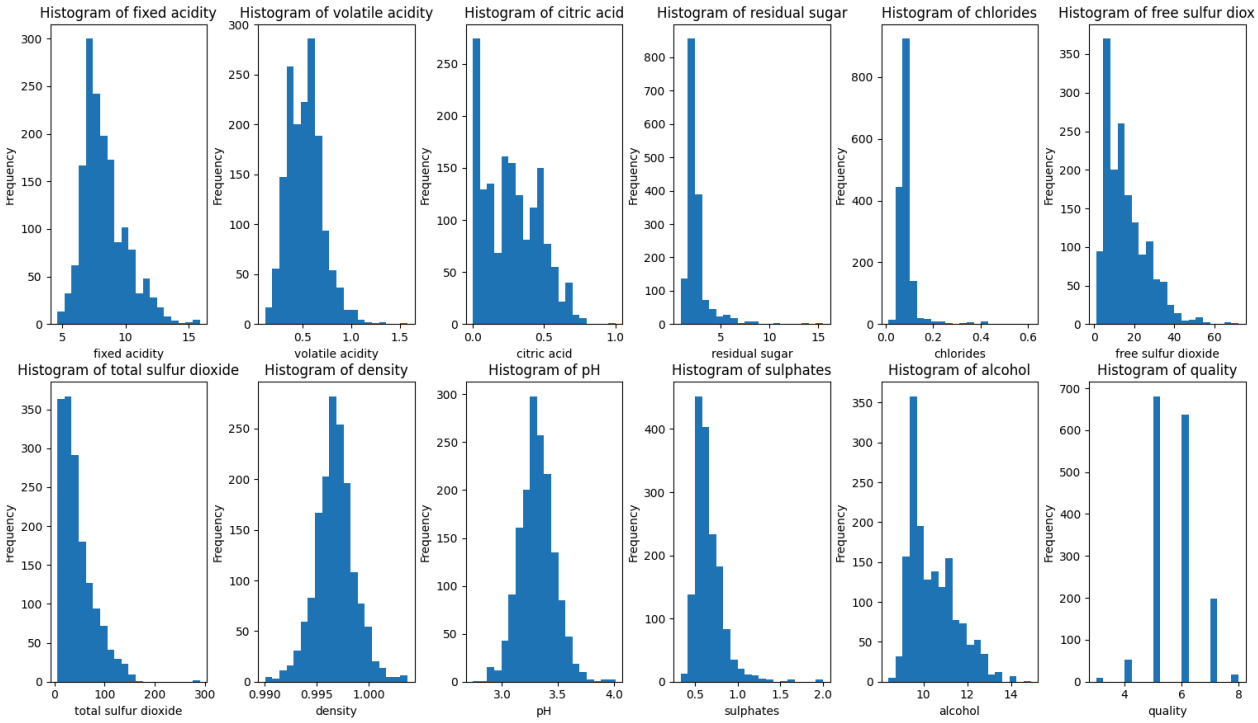
In the development of this Data Analysis AI a big step is creating data visualizations to perform analysis on. There are a variety of data visualizations on different datasets provided. This is to showcase the versatility of the system being created. The visualizations themselves were created using the generic functions made and utilizing AI. It is important to note that the data analysis summary provided for each individual dataset was written by me. Herein lies another developmental issue with this system where it will be able to produce a full report, the data visualizations must be interpreted and analyzed by a human. For a more in-depth explanation see the *Recommendations Section*.

The visualizations provided, while informative, exhibit several areas where improvements are necessary to enhance their interpretability and professional presentation. The biggest and most notable is the absence of descriptive titles on the graphs. Titles serve to immediately inform the viewer of the context and content of the data being presented. Additionally, some of the visualizations do not fit within the boundaries of the output, leading to truncated axes and labels, which can obscure critical information and diminish the utility of the graph. These issues will hopefully be addressed when the system is fully hooked up and

functional. These graphs were all made from generic functions, so in order to keep them usable for a variety of data, implementing things like axis labels and graph titles have proven challenging. My hope is that when each program and functions are routed that this will be easily fixable.

Also, at the very end there are a few “visualization” examples of exploratory data analysis outputs. This includes pandas.head(), pandas.describe(), and pandas.info() all on the dataset Red Wine Quality. The reasoning for including these in the data visualization section is to show how I found a solution to incorporating an output of these functions into a report style output. The goal was to be able to display a readable version of these functions in a way that can be easily incorporated into a data analysis report. Most of these functions return a data frame or series or nothing at all. Capturing and saving these in a picture format allows seamless integration into a complete single file data analysis report. Although they are quite small when displayed here, the goal was to create dynamic zoomable pictures to ensure the capture of all the data. Hopefully I will be able to do this in the final display of the output.

Red Wine Quality Data Set & Visualizations



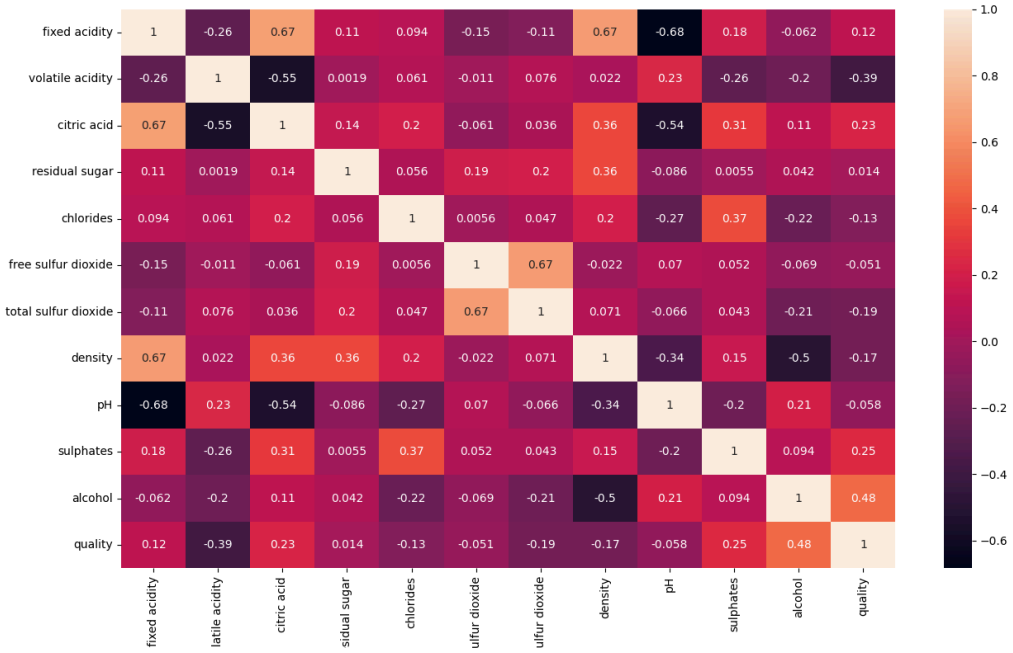
Collection of Histograms

The image shows a collection of histograms, each representing the distribution of different physicochemical variables and quality ratings of wine.

- **Fixed Acidity**: This histogram shows a right-skewed distribution, indicating that most wines have lower fixed acidity, with fewer wines having high fixed acidity.
- **Volatile Acidity:** The distribution of volatile acidity is also right-skewed, similar to fixed acidity, with most wines having lower volatile acidity.
- **Citric Acid**: The histogram for citric acid displays a right-skewed distribution. Most wines have lower citric acid content, with a small number of wines showing higher citric acid levels.
- **Residual Sugar**: The distribution is highly right-skewed, with a large peak at the lower end, indicating that the majority of wines have low residual sugar content.

- **Chlorides**: The histogram for chlorides is heavily right-skewed, with most wines having low chloride content.
- **Free Sulfur Dioxide**: This variable shows a right-skewed distribution, with the majority of wines having lower levels of free sulfur dioxide.
- **Total Sulfur Dioxide**: Again, a right-skewed distribution is observed. Most wines have a lower concentration of total sulfur dioxide.
- **Density**: The histogram shows a somewhat normal distribution but slightly left-skewed, indicating that there are a few wines with lower density than the mode.
- **pH**: The pH levels of wine seem to have a roughly normal distribution, with the majority of wines clustered around a central pH value.
- **Sulphates**: The distribution of sulphates is right-skewed, with most wines having lower concentrations of sulphates.
- **Alcohol**: The alcohol content histogram is right-skewed, showing that most wines have a lower alcohol percentage, with fewer wines having higher alcohol content.
- **Quality**: The quality rating histogram indicates that most wines fall within the middle quality categories, with fewer wines rated at the extremes (low or high quality).

Each histogram provides insight into the distribution of these variables within the dataset of Red Wine Quality. Skewness in the distributions suggests that there are outliers that may affect the mean and standard deviation of these variables. For example, in the case of residual sugar, the skewness might imply that while most red wines have a low residual sugar content, there are some wines with exceptionally high residual sugar levels.



Heatmap

The heatmap of Red Wine Quality represents a correlation matrix for various physicochemical properties of Red Wine. Each cell displays the correlation coefficient between each of the properties on the corresponding row and column. The correlation coefficient is a statistical measure that expresses the extent to which two variables are linearly related to one another, ranging from -1 to 1. A value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

The colors on the heatmap range from dark purple (indicating a strong negative correlation) to dark red (indicating a strong positive correlation), with lighter shades of each color indicating weaker correlations. Cells that are closer to white indicate a very weak or no correlation. The

diagonal from the top left to the bottom right shows perfect correlation (1.0) as it represents the correlation of each variable with itself.

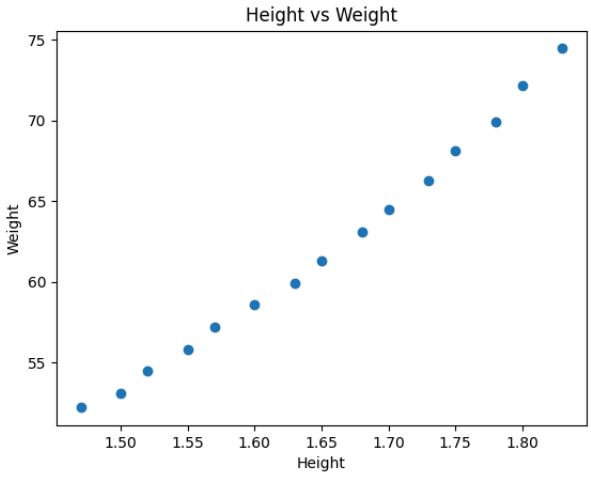
Some notes and analysis on different aspects of the heatmap:

- There is a strong negative correlation (-0.68) between fixed acidity and pH, suggesting that as the fixed acidity increases, the pH level decreases. This makes sense because wine is generally high in acidity.
- There is a significant positive correlation (0.67) is noted between citric acid and fixed acidity, indicating that wines with higher fixed acidity tend to also have higher citric acid levels.
- There is a very strong positive correlation (0.67) between free sulfur dioxide and total sulfur dioxide. This is expected as free sulfur dioxide contributes to the total sulfur dioxide content.
- A notable negative correlation (-0.5) is observed between alcohol content and density. This implies that wines with higher alcohol content tend to have lower density, which is consistent with the physical properties of ethanol.
- There is a moderately positive correlation (0.48) between alcohol and quality, indicating that higher alcohol levels may be associated with higher quality ratings in red wine.
- There's a positive correlation (0.25) between sulphates and quality, suggesting that sulphates might play a role in the perceived quality of red wine.
- The correlation coefficient is -0.39 between volatile acidity and quality, indicating a moderate negative relationship. This infers that as the volatile acidity in wine increases, the quality rating tends to decrease.

- A strong negative correlation (-0.55) is present with Citric Acid and Volatile Acidity, indicating that wines with higher levels of citric acid tend to have lower levels of volatile acidity.

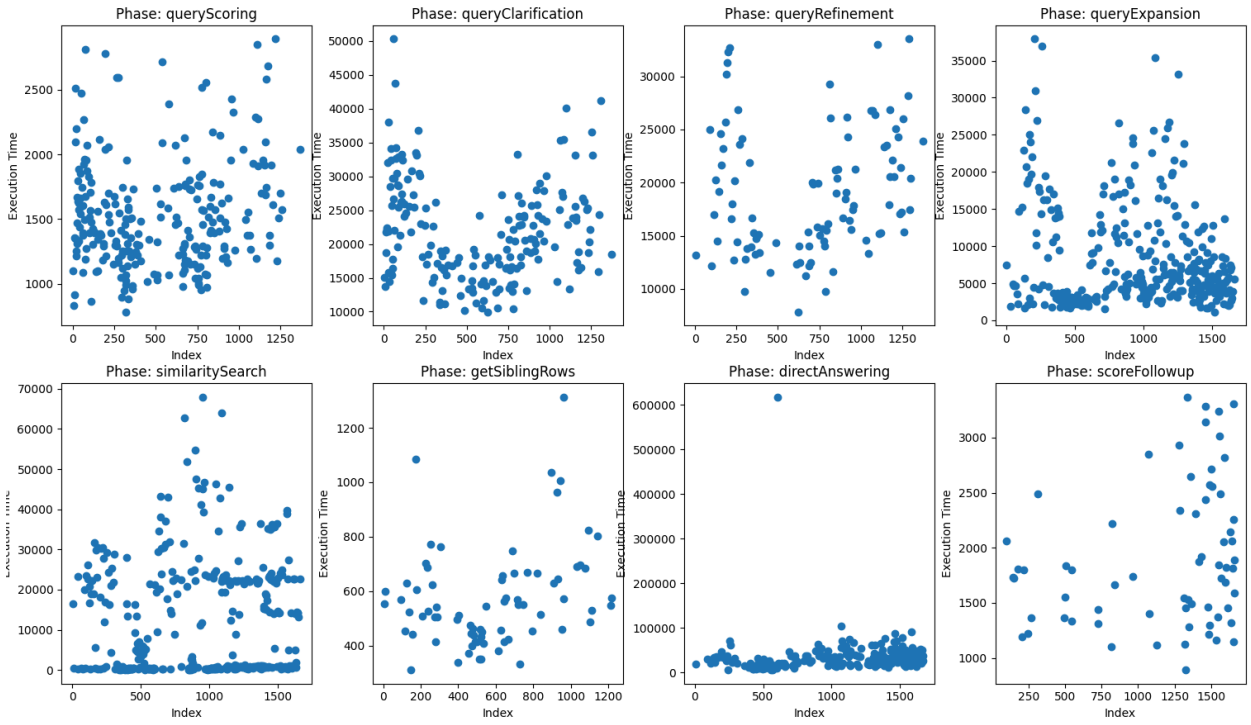
The colors on the heatmap range from dark purple (indicating a strong negative correlation) to dark red (indicating a strong positive correlation), with lighter shades of each color indicating weaker correlations. Cells that are closer to white indicate a very weak or no correlation. The diagonal from the top left to the bottom right shows perfect correlation (1.0) as it represents the correlation of each variable with itself.

Height vs. Weight Simple Data



This scatter plot represents the relationship between height (in meters) and weight (in kilograms) of a set of individuals or observations. From the plot, we can observe that there appears to be a positive correlation between height and weight, indicating that as height increases, weight tends to increase as well. This is a common physiological pattern because taller individuals typically have more body mass, contributing to greater weight. The relationship between the two variables appears to be linear, as the points seem to form a line when going from the bottom left to the top right of the plot. The trend line (if drawn) would likely have a positive slope, starting from the lower left corner of the plot and moving towards the upper right corner, reflecting the positive correlation between height and weight.

Ask Abe API Debugging Execution Log Data



Here are multiple scatterplots provided of the eight unique api_phases discovered by the ai and plotted against execution_time.

Phase: queryScoring

The scatter plot shows a broad range of execution times with no clear trend or pattern based on the index. This suggests that the execution time for the query scoring phase is quite random which would make sense as the time it takes to score questions asked to abe can be of wide variety.

Phase: queryClarification

Here, the execution times are also widely spread without a distinct pattern. It is good to see that a majority of query clarification questions are answered in the lower region of the execution time. Reminder: queryClarification execution_time is the amount of time that it takes for follow_up questions to be asked by the AI and answered by the user.

Phase: queryRefinement

The data points are spread out across the plot. Similar to the previous plots, this could indicate that because of the variety of questions asked and the complexity of needing to be refined to be a better question.

Phase: similaritySearch

The execution time shows a high variability with several outliers indicating instances of significantly higher execution times. These outliers may indicate specific queries that are much more complex or difficult to process than others.

Phase: getSiblingRows

Execution times are relatively lower compared to other phases but show a clear spread across the index, indicating variability that does not appear to be directly related to the index.

Phase: directAnswering

The scatter plot indicates a significant outlier with an extremely high execution time compared to all others. This outlier might suggest an exceptional case where the direct answering phase took

an unusually long time to execute. This is really interesting to us as a business because it shows that the AI spent a very long time searching for an answer in our database.

Phase: scoreFollowup

The execution times for the score follow-up phase are mostly clustered within a lower range, with a few moderate outliers. The majority of the execution times are below 2000, with no discernible pattern based on the index.

From these scatter plots, we can infer that the execution time varies considerably across different phases and does not show a clear trend with respect to the index. For a more detailed analysis, statistical measures such as correlation coefficients could be calculated, or a time-series analysis could be conducted if the index represents a sequential order of execution. Additionally, examining the outliers in more detail could provide insights into what causes the execution time to spike in those instances.

Data Function transformed into usable pictures

df.describe()

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.0	1599.0	1599.0	1599.0	1599.0	1599.0	1599.0	1599.0	1599.0	1599.0	1599.0	1599.0
mean	8.319637273295501	0.5278205128205129	0.2709756097560976	2.5388055034395	0.08746654128049279	15.874921826141339	46.46779237023139	0.9967466791744841	3.3111131357473416	0.6581488430206917	10.422983114446529	5.6360225140712945
std	1.7410963181276953	0.17905970415353537	0.19480113740531057	1.4099280595972798	0.04706531020100901	10.460156969609725	32.895324478299074	0.0018873339538425554	0.15433664690354277	0.16950697959010996	1.0656675818473946	0.8075694397734705
min	4.6	0.12	0.0	0.9	0.012	1.0	6.0	0.99007	2.74	0.33	8.4	3.0
25%	7.1	0.39	0.09	1.9	0.07	7.0	22.0	0.9956	3.21	0.55	9.5	5.0
50%	7.9	0.52	0.26	2.2	0.079	14.0	38.0	0.99675	3.31	0.62	10.2	6.0
75%	9.2	0.64	0.42	2.6	0.09	21.0	62.0	0.997835	3.4	0.73	11.1	6.0
max	15.9	1.58	1.0	15.5	0.611	72.0	289.0	1.00369	4.01	2.0	14.9	8.0

df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #  Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0  fixed acidity         1599 non-null   float64
 1  volatile acidity      1599 non-null   float64
 2  citric acid           1599 non-null   float64
 3  residual sugar        1599 non-null   float64
 4  chlorides             1599 non-null   float64
 5  free sulfur dioxide   1599 non-null   float64
 6  total sulfur dioxide  1599 non-null   float64
 7  density               1599 non-null   float64
 8  pH                    1599 non-null   float64
 9  sulphates             1599 non-null   float64
 10 alcohol               1599 non-null   float64
 11 quality               1599 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

df.head()

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5.0
1	7.8	0.88	0.0	2.6	0.098	25.0	67.0	0.9968	3.2	0.68	9.8	5.0

Full Run Through of Model

There are three almost completed AI models. The first one that performs exploratory data analysis. The second, performs the actual decision of the call. And the third, decides what to label the graph title and each axis. There are a few things that are needed as human input to the machine which are the dataset, as well as an option brief description to include for the exploratory data analysis. I have tested with and without and it performs about the same, however for more complex datasets the AI will be able to perform more in depth analyses with the data description provided. For example, in the api debugging analysis (which I am still working out getting the proper output as the analysis on unique column variables has been quite challenging to implement) although the exploratory data analysis is given to the AI, it has trouble specifically selecting to perform analysis by grouping unique values within columns.

Analysis of AI Model

The AI model designed to automate the data science analysis process demonstrates promising performance, particularly in the realm of statistical analysis. It appears to handle a wide array of statistical tasks with accuracy and efficiency, from basic descriptive statistics and

data visualization to more complex inferential statistics based on unique values within a singular column. The model's ability to adapt to various datasets and automatically select the appropriate statistical methods suggests a robust understanding of data characteristics and analysis requirements. Preliminary results indicate that for most statistical analyses, the model not only speeds up the analysis process but also maintains a high level of precision and reliability in its outputs. This performance is indicative of the model's advanced machine learning algorithms and its capacity to learn from data, making it a valuable tool for data scientists and analysts seeking to streamline their workflow and enhance the quality of their insights.

Challenges

When creating data science analyses with artificial intelligence, several significant challenges arised, primarily related to integrating various AI functions and fine-tuning prompts for optimal performance. One major hurdle was the difficulty in seamlessly connecting the outputs of one function or the artificial intelligence as inputs for another. This issue stemmed from discrepancies in data formats, scales, and contextual interpretations. The AI is meant to be generalized to perform on all different types of data and subsequently made it challenging to provide inputs for the different functions. Additionally, fine-tuning prompts for AI models, especially in natural language processing tasks, demanded a deep understanding of the model's architecture and training data. Crafting prompts that effectively communicated the task at hand while avoiding biases or ambiguities, required iterative experimentation and a nuanced grasp of the language model's capabilities. These challenges underscore the complexity of developing AI-driven data science analyses, necessitating advanced technical skills, creativity, and a strategic approach to problem-solving.

Recommendations and Next Steps

Data Analysis AI systems have become indispensable tools in extracting insights from a variety of datasets. It is essential to incorporate more advanced features like data visualization analysis and sophisticated data science techniques. However, enhancing AI capabilities introduces complex ethical considerations. Balancing these technological advancements in this AI with ethical obligations ensures that this AI system is not only powerful but also responsible and equitable.

Integrating the ability for the AI to analyze data visualizations would create a significant leap forward in understanding and interpreting complex datasets. This capability would allow the AI to derive insights from graphical data, expanding its analytical reach. However, concerns arise regarding transparency and bias. It is crucial that the algorithm developed for this purpose will have a clear decision-making process and will be trained on diverse datasets to minimize this bias. It is also ideal to improve and produce better data visualizations. Data visualizations are essential in facilitating the understanding and decision-making made with data analysis. This includes making sure the visualization is fully shown in the output, there are proper labels including titles and specified metrics. There is a responsibility to ensure that these visualizations do not mislead or misrepresent the data. Improving the AI system to ensure more detailed visualizations will hopefully lead to less misrepresented data.

It will also be beneficial to include more complex data science analysis techniques and tools like clustering. Incorporating a clustering algorithm will enhance the AI's ability to identify patterns and similarities within the data. Another important part would be to improve the data cleaning processes. This is essential for accuracy and reliability. However, ethical considerations

include the potential for data manipulation or exclusion that could lead the AI to skew results. Maintaining integrity during data cleaning requires clear guidelines to the AI and oversight to ensure that all data is treated fairly and without bias.

The incorporation of advanced data analysis capabilities into AI systems presents significant opportunities to enhance their utility and effectiveness. However, these advancements must be pursued with a strong commitment to ethical principles to ensure that the development and use of this AI in data analysis contributes positively to society. By adhering to the recommendations outlined above, it is possible to continue advancing a Data Analysis AI system that is not only technologically advanced but also responsible, transparent, and inclusive.

Source Code

Below is an explanation of the different source code files attached with this presentation. For security and confidentiality reasons, certain files are excluded which include a config.py file which hosts personal information like my OpenAI API Key and database passwords. These are needed for a user to be able to be able to run the files and produce an output. If the files are to be run by other users, they will need to provide these on their own personal device.

function.py / functions.ipynb

This python file contains all of the functions used for Exploratory Data Analysis and the production and output of various graphs. These functions have all been generalized to support

different datasets and can be called for testing in the main() function. However, the only file that needs to be run to produce an output is the runner.py file.

prompts.py / prompts.ipynb

This python file contains all of the prompts used to generate the ai model. These prompts have been specifically created for this particular AI project. They have been iterated over multiple times to produce the best possible output. There are three different prompts, and thus three different AI models. One, which performs exploratory data analysis, which is fed a variety of the EDA function outputs. The second model receives the output from the first AI model, it then makes a decision on which data analysis to perform. This output is then returned and validated by a pydantic model and calls the function used to generate the graph. There are a few issues with this model where sometimes it will specify the correct data analysis to perform (ie. Correlation Analysis, or Linear Regression) and sometimes it will simply state the analysis as being the function to call or graph to perform the analysis (ie. Heatmap, Scatterplot). I'm not necessarily sure why this happens, and it will need to be looked into more in depth. The last AI model (prompt) is fed the 2nd model's output. This prompt is solely to label the graphs Titles and Axis'. Again, like the 2nd prompt this one varies with title output, sometimes it includes measurements (ie. inches, centimeters) and sometimes it does not. It also sometimes has creative names for titling the graphs and sometimes it does not. These are all common issues and can hopefully be fixed with further testing, but it is the best I could get it to for the deadline of this project. It is also not fully set up, as I was having a ton of issues with hooking this up with both calling the function, as well as using multiple different inputs (having the AI call specific inputs

for each graph. As well as title labels). I hope to continue to build upon what I have created and improve it even more.

runner.py / runner.ipynb

This is the python file that controls all of the AI models and outputs. This is the only file that should be run to produce an output of the models. It is connected to each separate python file. If the user chooses they can change the dataset and also change the exploratory_data_analysis function to include a brief description of the dataset that will be part of the input to the AI system. This is also where the PDF of complete analysis is provided and should include exploratory data analysis as well as analysis chosen and reasoning. This output will also first save the graph, and then include the graph in the PDF. It still needs some work on the formatting and I am considering switching to produce a jupyter markdown file which can then be converted to a PDF later. This was more of an experimentation toward the end of this project to see if it would be feasible to produce a report style document with all of the information and output produced by the AI.

utilityFunction.py / utilityFunction.ipynb

This python file contains all of the utility functions needed to connect to OpenAI, Supabase Database, and Postico local database.

References

(n.d.). Kaggle: Your Machine Learning and Data Science Community. Retrieved February 11, 2024, from https://www.kaggle.com/

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014, December 22). *Power to the People: The Role of Humans in Interactive Machine Learning | AI Magazine*. AAAI. Retrieved February 11, 2024, from https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2513

Introduction to Automated Data Analytics (With Examples). (2022). Keboola. Retrieved February 11, 2024, from https://www.keboola.com/blog/automated-data-analytics

NEUMEISTER, L. (2023, June 22). *Lawyers submitted bogus case law created by ChatGPT. A judge fined them $5000*. AP News. Retrieved February 11, 2024, from https://apnews.com/article/artificial-intelligence-chatgpt-fake-case-lawyers-d6ae9fa79d0542db9e1455397aef381c

Shyr, J., & Spisic, D. (2017, November 9). YouTube: Home. Retrieved February 11, 2024, from https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.1318