

AI-Powered Data Cleaning: Streamlining Processes for Smarter, Data-Driven Campus Decisions.

Presented by Lamis Ghoualmi, software analyst at the
Office of Innovative Technologies (OIT)

**TN Higher Education IT Symposium
2025**



Plan

- 1 Introduction
- 2 Data Cleaning
- 3 Artificial Intelligence (AI)
 - Large Language Model (LLM)
 - ChatGPT
- 4 The proposed approach
 - Data cleaning using AI
 - App Demonstration of data cleaning using AI
- 5 Conclusion and discussion
- 6 Useful links to build an AI app using Python



- In today's world, data is omnipresent.
- Analyzing data is considered a primary responsibility, not only in understanding the natural world but also in making profit and success across industries.

The topics listed highlight the diverse applications of data in various domains:

- **Retail:** Understand customer behavior, preferences, and trends.
- **Finance:** Predict stock market trends and assess investment risks.
- **Education:** Evaluate student performance, personalize learning paths, and optimize educational strategies.
- **Human Resources:** Streamline recruitment processes and assess employee satisfaction.
- **Transportation:** Optimize route planning.
- **Energy:** Monitor energy consumption, optimize power grid operations, and predict equipment failures.
- **Social Media:** Analyze user engagement and personalize content recommendations.
- **Healthcare:** Personalized medicine and treatment plans predict disease outbreaks and enhance patient care through data-driven insights.

Importance of Data analysis and data science in today's world

- Data analysis plays a critical role not only in driving corporate profitability but also in improving human lives and understanding the intricate evolution of the world.





The process of data analysis involves several steps:

- **Define Objective:** Clearly state the problem or objective of the analysis. Understand the goals and desired insights.
- **Data Cleaning:** Gather relevant data from various sources. Clean and preprocess the data.
- **Exploratory Data Analysis:** Perform EDA to gain insights into data characteristics using visualizations and summary statistics.
- **Machine Learning:** Apply statistical methods or machine learning algorithms for in-depth analysis based on the defined objective.
- **Interpretation and Communication:** Interpret the results in the context of the problem. Draw conclusions and communicate the findings effectively to stakeholders.

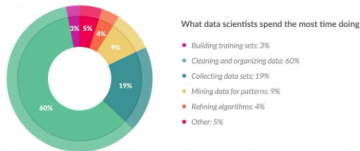
Data Cleaning

Data cleaning consumes **80 percent** of a data analyst's and scientist's time and resources.

Forbes

 A new survey of data scientists found that they spend most of their time **massaging** rather than mining or modeling data.  Still, most are happy with having the **sexiest job of the 21st century**. The survey of about 80 data scientists was conducted for the second year in a row by CrowdFlower, provider of a “data enrichment” platform for data scientists. Here are the highlights:

***Data preparation** accounts for about 80% of the work of data scientists*



Data scientists spend 60% of their time on cleaning and organizing data.

Importance of Data Cleaning



- Efficient data preparation is crucial for the successful analysis and prediction of valuable insights.
- Main data cleaning challenges:
 - Duplicate data.
 - Missing values.
 - Outliers.
 - Spelling errors in categorical data.

Challenges of Manual Data Entry



- Entering data manually introduces challenges such as typographical errors and inconsistent formatting.
- Cleaning these entries manually is time-consuming, involving meticulous inspection, rule creation, and quality assurance checks.

Example of Challenging categorical data Entry

Fruit Table: The unique value of the field name 'Fruit'

value
Banana
Straberries
Pineapple
Grapes
Apples
Strawberries
Aples
Grape
Grape

value
Grapes
Apples
Strawberries
Aples
Grape
Grap
Grappes
Appl

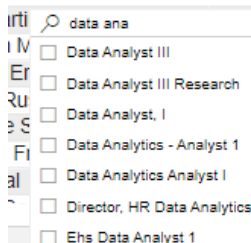
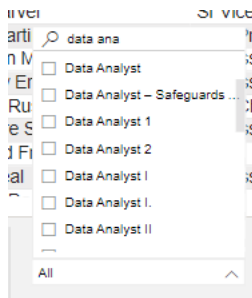
Example of Challenging categorical data Entry

HR table: The unique value of the field name 'Category'

value	value
Technology	
Finance	Sports
Healthcare	Heahcare
Education	Health
Sports	Sport
Heahcare	Tech
Health	Technolog
Sport	Education!
Tech	SportsS.
Technolog	Healthcar
	Educationnn

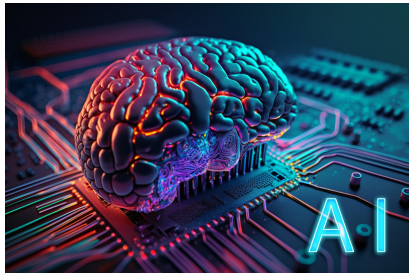
Example of more Challenging categorical data Entry

Real example of Employee Dashboard: Job title field name



Artificial Intelligence (AI)

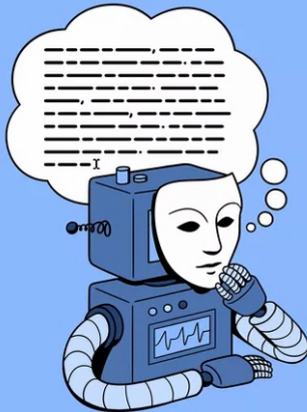
- Artificial intelligence (AI) refers to developing computer systems that can perform tasks that typically require human intelligence. These tasks include:
 - Learning from experience (machine learning).
 - Understanding natural language.
 - Recognizing patterns, solving problems, and adapting to changing environments.



Plan

- 1 Introduction
- 2 Data Cleaning
- 3 Artificial Intelligence (AI)
 - Large Language Model (LLM)
 - ChatGPT
- 4 The proposed approach
 - Data cleaning using AI
 - App Demonstration of data cleaning using AI
- 5 Conclusion and discussion
- 6 Useful links to build an AI app using Python

Large Language Model (LLM)



Large Language Model (LLM)

['lärj 'laŋ-gwij 'mä-dəl]

A deep learning algorithm that's equipped to summarize, translate, predict, and generate human-sounding text to convey ideas and concepts

Plan

- 1 Introduction
- 2 Data Cleaning
- 3 Artificial Intelligence (AI)
 - Large Language Model (LLM)
 - ChatGPT
- 4 The proposed approach
 - Data cleaning using AI
 - App Demonstration of data cleaning using AI
- 5 Conclusion and discussion
- 6 Useful links to build an AI app using Python

ChatGPT

- ChatGPT is a language model developed by OpenAI, specifically for the GPT-3.5 architecture.
- It is designed for natural language understanding and generation, allowing it to comprehend and produce human-like text.
- ChatGPT is trained on diverse internet text, making it capable of engaging in conversations, answering questions, and providing information.



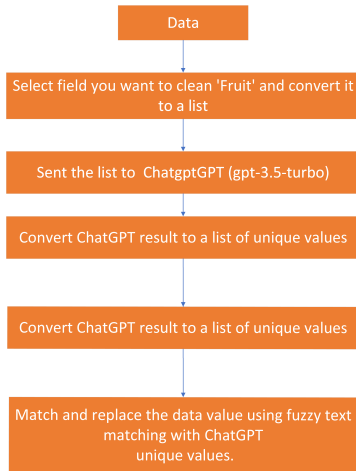
Plan

- 1 Introduction
- 2 Data Cleaning
- 3 Artificial Intelligence (AI)
 - Large Language Model (LLM)
 - ChatGPT
- 4 The proposed approach
 - Data cleaning using AI
 - App Demonstration of data cleaning using AI
- 5 Conclusion and discussion
- 6 Useful links to build an AI app using Python



- We propose an AI-driven solution aimed at efficiently cleaning categorical data, reducing both time and costs in data preparation.
- The proposed method utilizes ChatGPT to automatically correct spelling
- Then we use the fuzzy text matching method to compare the corrected words provided by ChatGPT and match them with our data values and we will replace them with the closes match to the results provided by ChatGPT.

The proposed approach: Data cleaning using AI



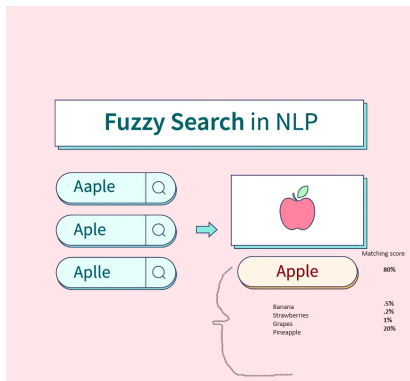
ChatGPT function

```
#
# Function to clean data using AI
def ai_clean(message):
    messages = [
        {
            "role": "system",
            "content": "Correct spelling errors. Standardize the words. Remove punctuation points. Convert everything to lowercase. "
        }
    ]

    #while True:
    if message:
        messages.append(
            {"role": "user", "content": message},
        )
        chat = openai.ChatCompletion.create(
            model="gpt-3.5-turbo", messages=messages, temperature=0.7, max_tokens=64, top_p=1
        )
        reply = chat.choices[0].message.content
        return reply
    #

def find_closest_match(word, choices):
    return process.extractOne(word, choices)[0]
#
```

Fuzzy text matching

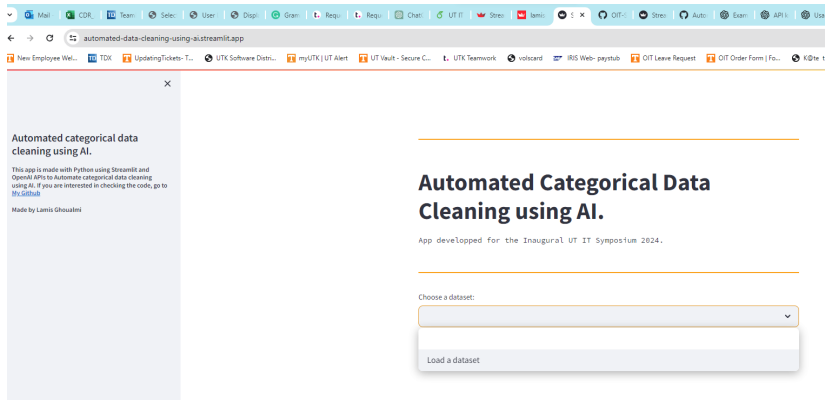


Plan

- 1 Introduction
- 2 Data Cleaning
- 3 Artificial Intelligence (AI)
 - Large Language Model (LLM)
 - ChatGPT
- 4 **The proposed approach**
 - Data cleaning using AI
 - **App Demonstration of data cleaning using AI**
- 5 Conclusion and discussion
- 6 Useful links to build an AI app using Python

App Demonstration

Link to The app <https://automated-data-cleaning-using-ai.streamlit.app/>.



Conclusion

- In this project, we introduced an innovative approach for cleaning categorical data and correcting spelling errors introduced by humans during data collection.
- Our proposed use of ChatGPT to correct the spelling errors, and then employs a text-matching algorithm to replace the data with the correct words.
- The testing phase demonstrated highly promising results.
- Data cleaning based on AI has the potential to reduce costs and time, providing data analysts and scientists with more dedicated time to focus on the primary objective which is getting valuable insights from the data.

Interested in building an AI App using Python ?

Useful links to build an AI app using Python as a programming language:

- **OpenAI:** is a Python API that allow developers to integrate artificial intelligence (AI), more specifically ChatGPT models into their applications, which is capable of natural language understanding and generation. Link to OpenAI playground: <https://platform.openai.com/examples>.
- **Streamlit:** is a Python library used for creating web applications with minimal effort. It is designed to make it easy for data scientists and developers to turn data scripts into shareable web apps. Link to Streamlit <https://streamlit.io/>.
- If you are interested to use ChatGPT in your daily life.
Link: <https://chat.openai.com/>.
- If you are interested to use UT Verse (University proper ChatGPT with protected data): <https://oit.utk.edu/ai/ut-verse/>.
- Link to my app script <https://github.com/lamisghoualmi/Automated-data-cleaning-using-AI/blob/main/newVersion.py>.

Interested in building an AI App using Python ?

When building a data cleaning method using AI for a specific data topic, consider the following steps: (It could be great for a research project.)

- **Data Topic:** Focus on developing the tool for a more specific data type, such as environmental data, pharmaceutical data, medical data, etc.
- **Adjust ChatGPT 3.5 Parameters:** Adapt the parameters of ChatGPT 3.5 to align with the nuances and intricacies of the chosen data domain.
- **Utilize and compare with pre-existing data cleaning methods:** integrate other pre-existing data cleaning techniques.
- **Compare Methodologies:** Compare the accuracy of these methods in handling specific challenges within the domain. This comparative analysis helps identify the most effective approach for ensuring data accuracy and reliability.

Link to see [Scary AI Youtube video](#).



- **Contact:** *Lamis Ghoualmi*, Software analyst at Applications: Information Management & Analytics at OIT, email: lgoualm@utk.edu. Manager: *Kelly Stroud*, IT manager of Information Management & Analytics team, Applications at OIT, email: kstroud4@utk.edu