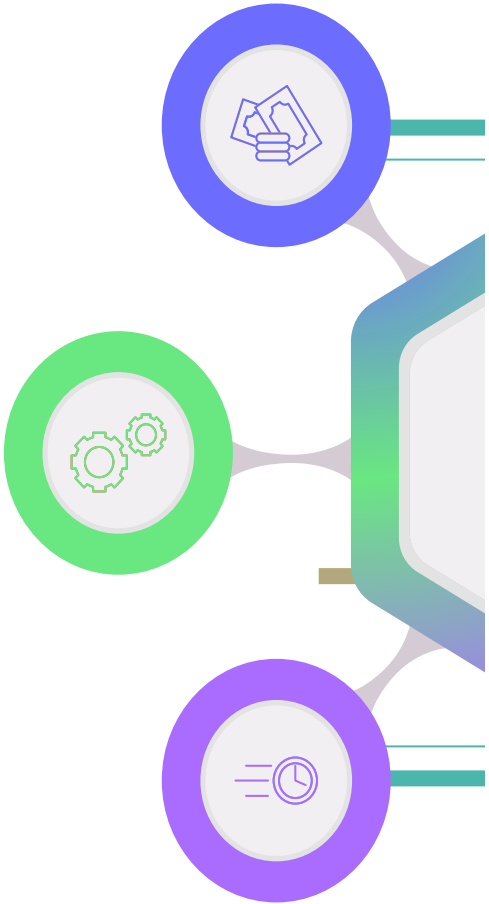
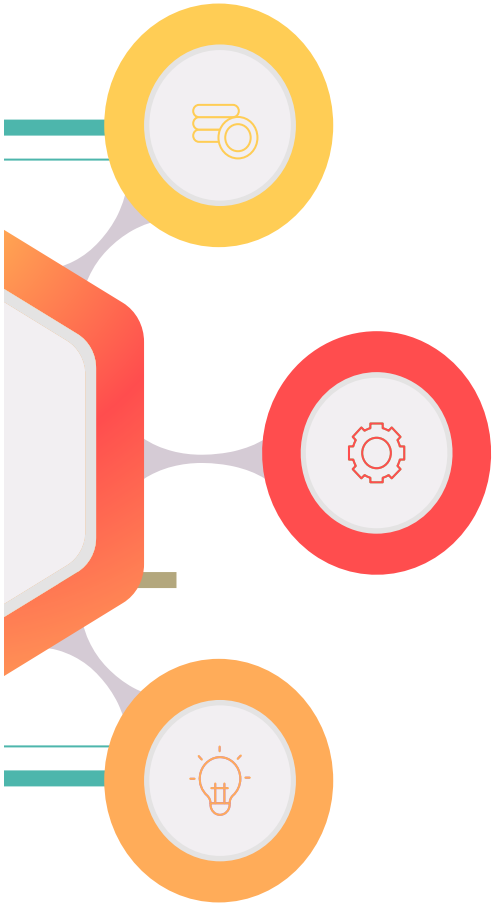


# Text Mining on US News University Ranking Dataset

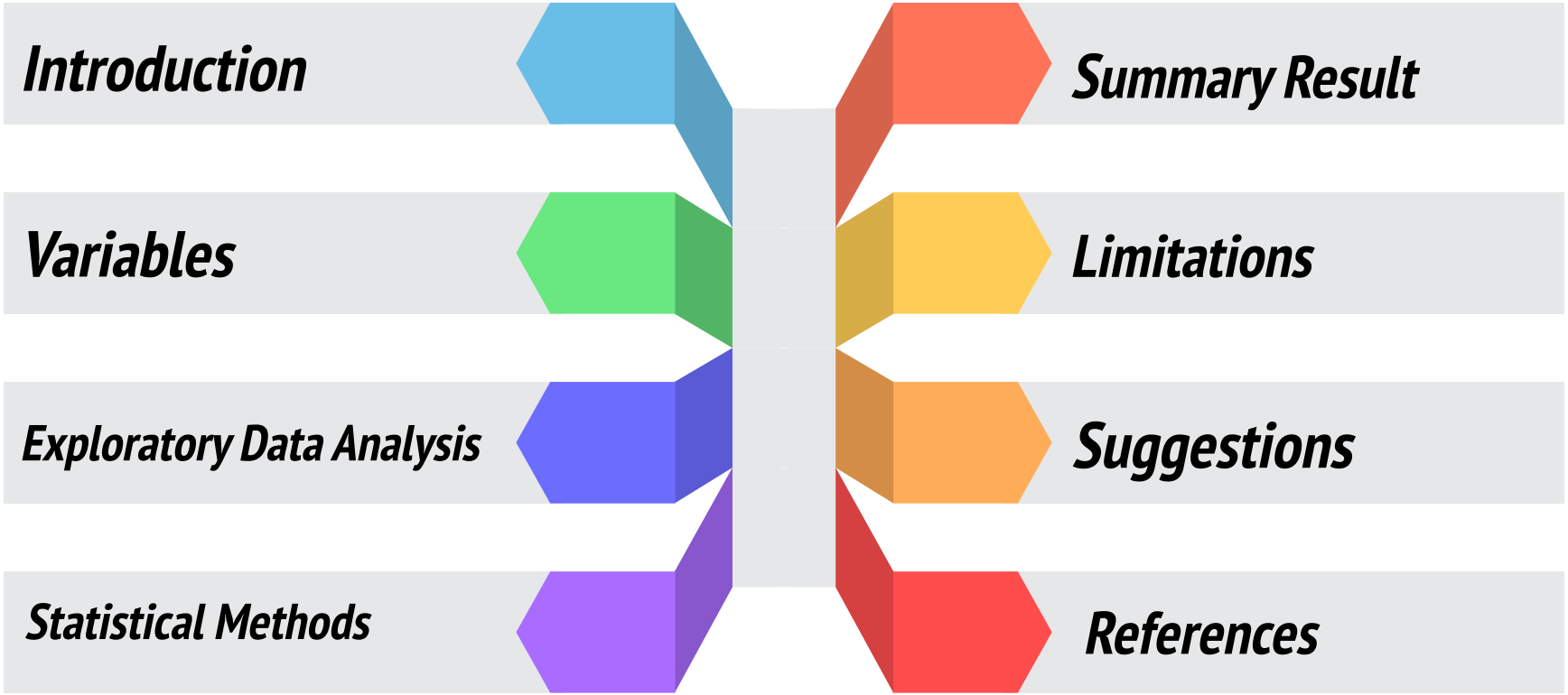
Stats 141XP Final Project


**Group Member:**

Kaizen Schanz
Thi Nguyen
Cynthia Wu
Minghao Li
Najia Pan
Yupei Hu

# CONTENTS


Introduction
Variables
Exploratory Data Analysis
Statistical Methods
Summary Result
Limitations
Suggestions
References

# ABSTRACT

This project aims to develop a method for calculating the difference and similarity scores between two strings, specifically the names of schools or metrics used by US News. The constantly changing naming conventions employed by US News creates issues when compiling longitudinal datasets and performing aggregations. The proposed method will provide recommendations for top matches and a standardized name that a given school or metric name should be recoded to. This approach will enable researchers to accurately and efficiently compile and analyze longitudinal data without being affected by the constantly changing naming conventions used by US News. The success of this project will facilitate the analysis of education-related data, which is critical for making informed decisions and policies related to education.

# INTRODUCTION

The problem is that College rankings uses inconsistent naming conventions for schools and metrics from year to year, making it difficult to compile longitudinal datasets and perform aggregations. This inconsistency in naming conventions creates issues when trying to analyze education-related data, which is essential for making informed decisions and policies related to education. Without a standardized method for calculating the difference and similarity between two strings, researchers are forced to spend a significant amount of time manually identifying and recording school and metric names. This process is time-consuming and error-prone, and it hinders the ability to conduct accurate and efficient data analysis. Therefore, there is a need for an automated method that can provide recommendations for top matches and a standardized name for school and metric names, to enable researchers to perform accurate and efficient data analysis.

# DATASETS AND VARIABLES

The present study involves the analysis of two distinct datasets - raw and cleaned - each of which pertains to school names and years. The primary objective of this investigation is to employ data cleaning techniques to merge the raw dataset with the cleaned dataset, and subsequently develop three distinct categories in the resultant CSV file, namely match, unmatch, and new school. Our analysis revealed that simple string matching algorithms were inadequate for achieving optimal results. We also found that even though some schools had high matching scores, they were two different schools in reality.Thus, we explored several sophisticated data analysis and cleaning methodologies. Our findings are presented in the subsequent sections.

# DATASETS

- Cleaned dataset – contains standard school names and metric names
  – records data from 2014 to 2022

| Number of Rows | Number of Columns |
|---|---|
| 84885 | 7 |

# DATASETS

- Raw dataset – contains unstandardized school names and metrics name

  – records data from 2014 to 2023

| Number of Rows | Number of Columns |
|---|---|
| 122341 | 9 |

# DATASETS

- 2023 dataset – contains only data of the latest issue year
  - a subset of the raw dataset

| Number of Rows | Number of Columns |
|---|---|
| 14092 | 9 |

# Target Variables

| Cleaned dataset | | Raw dataset |
|---|---|---|
| ` Institution name` | ⟷ | ` Name ` |
| ` Metric Name ` | ⟷ | ` Metric.description ` |
| ` Ranking Type ` | ⟷ | `School.Type` |

* Ranking Type - School.Type is only targeted when identifying regional schools

# Exploratory data analysis

- The cleaned dataset only include universities that appear in the National Universities ranking.
- Any schools outside of the National Universities ranking have been excluded from the cleaned dataset
- National Universities ranking is a comprehensive list of the best colleges and universities in the United States

# Exploratory data analysis

- Number of unique school/metric names of 2023 data

| Unique school names | Unique metrics |
| --- | --- |
| 443 | 34 |

# Exploratory data analysis

- Number of unique school/metric names of cleaned data

| Unique school names | Unique metrics |
| --- | --- |
| 430 | 42 |

# Exploratory data analysis

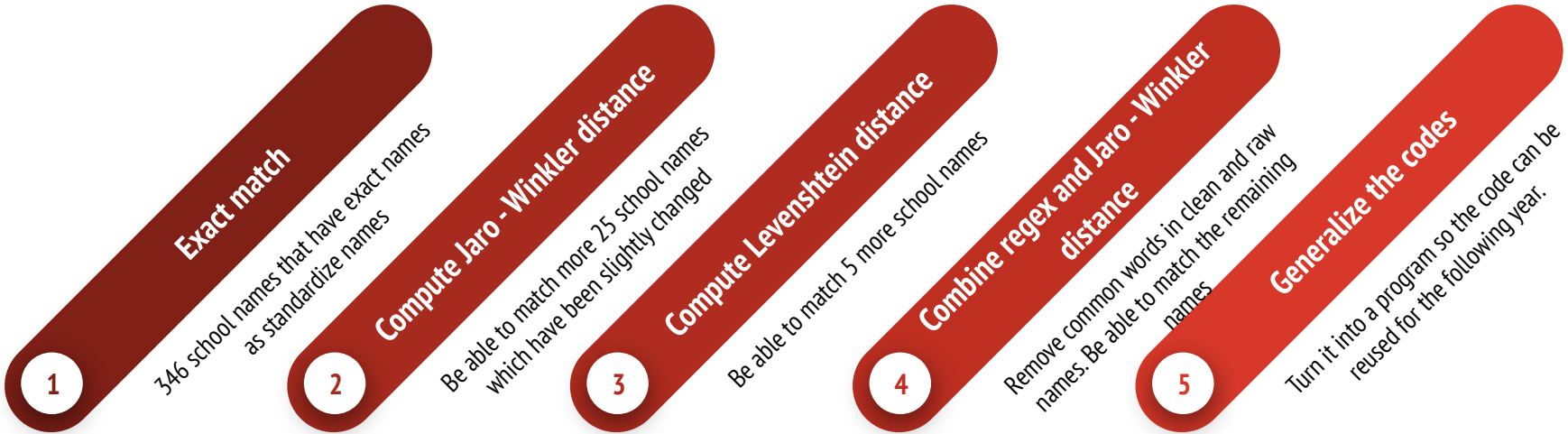
- Number of missing values of the target variables:

| ` Name ` | ` Metric.description ` |
|---|---|
| 0 | 0 |

# Exploratory data analysis

- **Partial identical**: Some schools, such as Touro College and Touro University, are identical but recorded as different suffixes.
- **Preposition issues:** In the raw dataset, school names are more likely to include prepositions such as **"the", "at",** and **"of"**, while in the clean dataset is not.

# STATISTICAL METHOD



# STATISTICAL METHOD

***Round1: Cleaning dataset and match exactly school names***

During the initial data cleaning process, the unique school names were extracted from the cleaned dataset and sorted alphabetically. As a result, the cleaned dataset contained 430 unique school names. Furthermore, there were 443 unique school names in the raw dataset of the 2023 issue year.Following data preprocessing, a matching process was conducted to identify schools that shared the exact same name across both datasets. The results showed that 346 school names had already matched without requiring further action.

# STATISTICAL METHOD

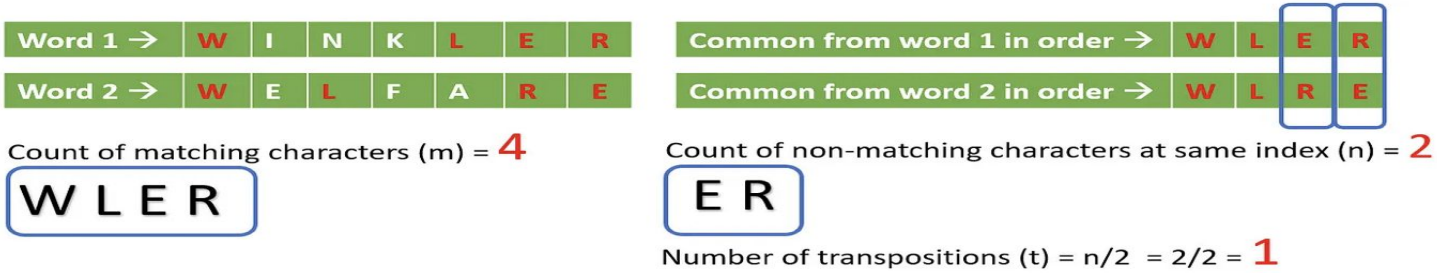
***Round 2:  Implement fuzzy matching method by Jaro-Winkler distance***

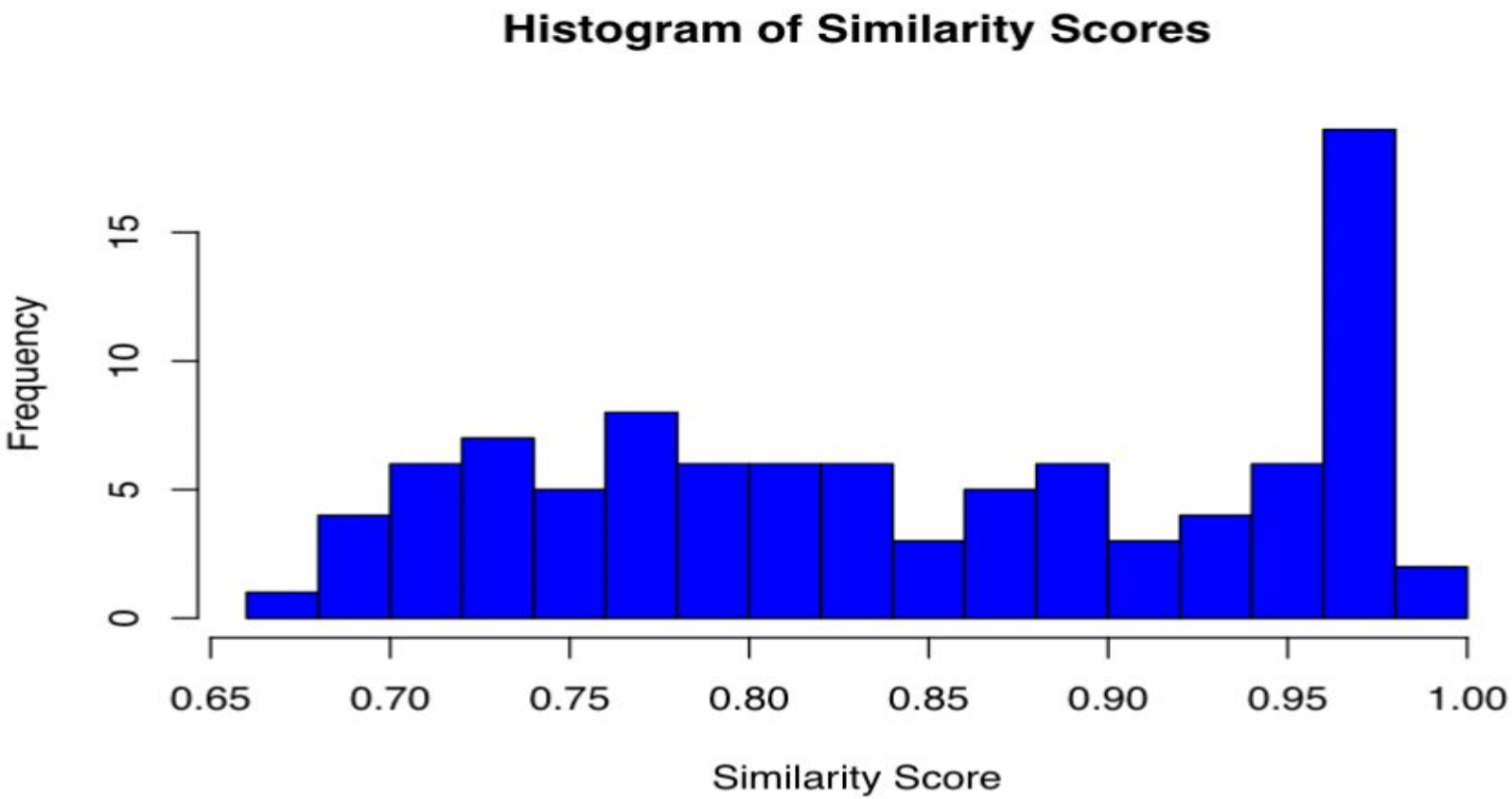
Now after we compared the exactly school names between two datasets, we used fuzzy matching to further match the school name.

After excluding the matched schools, we are left with 97 unmatched schools in the 2023 dataset. Then we find the third quantile of the similarity scores vec, we conclude that we can match 25 school names by using ***Jaro-Winkler distance.***

# WHAT IS JARO - WINKLER SIMILARITY ?

***Jaro-Winkler similarity*** measures the number of characters between two strings that need to be edited for the strings to match, but Jaro-Winkler gives more weight to the strings that match in earlier characters (i.e., to the first half of the string).




Histogram of Similarity Scores
Frequency
0
5
10
15
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Similarity Score

# STATISTICAL METHOD

***Round 3: Implement fuzzy matching method by Levenshtein distance***

Firstly, we make an new table include school names that we can not match after round 2, then we find there are 72 school names can not matched. We compare certain name of from 72 school names to the dataset which is finding the ***Levenshtein distance***,it is is a string metric for measuring the difference between two sequences. Then find the similarity scores and use max function to find the best score which mean the best match. In the end, we have 5 more school names matched.

# WHAT IS LEVENSHTEIN DISTANCE ?

***Levenshtein distance*** is a string metric for measuring the difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other.

Edits can be replace, insert OR delete a character

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Source | W | I | N | K | L | E | R |
| Destination | W | E | L | F | A | R | E |

| Edits | W<br>No Change | I → E<br>Replace | N → L<br>Replace | K → F<br>Replace | L → A<br>Replace | E → R<br>Replace | R → E<br>Replace |
|---|---|---|---|---|---|---|---|

Levenshtein Distance (LD) = Number of edits = 6

Similarity Score = 1 – LD/MAX_LENGTH( SRC_STRING, DEST_STRING)
= 1 – 6/7
= 1 – 0.85
= 0.15

# STATISTICAL METHOD

***Round 4: Further analysis and processing of data***

According to our analysis, when we truncated the number of school names, we found that there were 67 unmatched school names. Subsequently, we searched for new school names in the National University of 2023 list and identified 62 schools that were not National Universities in the previous year. After truncating the number of school names again, we obtained 11 school names. Further, we observed that in some particular cases, removing common words and location parts resulted in the successful matching of these school names.

# STATISTICAL METHOD

***Round 5: Combine regular expression and fuzzy matching method by Jaro-Winkler distance again.***

We remove common words as “at”, “the”, “College ”and “University”. We repeated the fuzzy matching with Jaro-Winkler method. At this time, we aslo set a higher threshold .

# SUMMARY RESULT

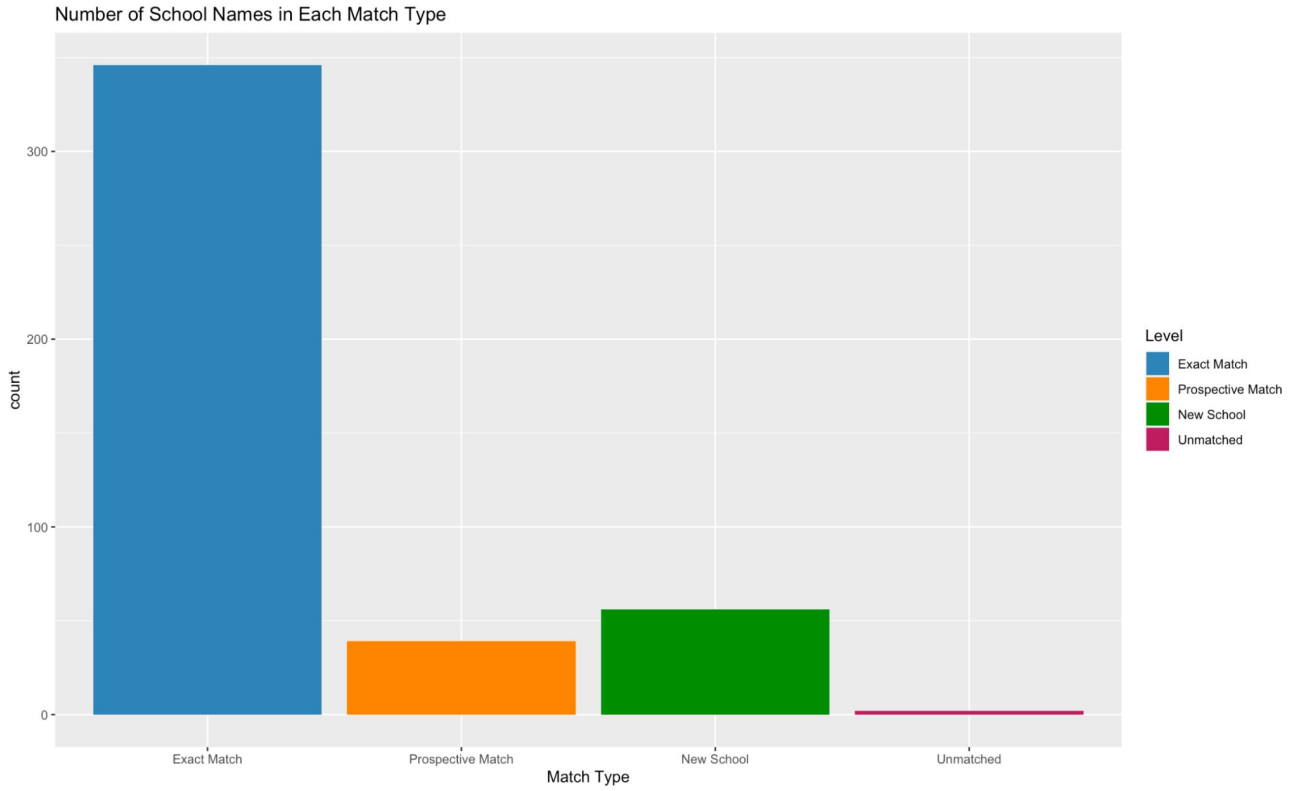
We have introduced a categorical variable that helps researchers determine the match status of a school using our approach. This variable consists of four levels: exact match, prospective match, new school, and unmatched. An exact match signifies schools that have not undergone any changes when compared to the clean dataset. Prospective match pertains to schools that have undergone minor name alterations or have entirely different names. New school is assigned to those that are new to the issue in the current year, which means they are included in the national university ranking for the first time. Lastly, unmatched denotes schools that could not be matched, and further scrutiny and expert analysis are necessary to determine if they are also new or not. If the school genuinely lacks any recorded historical data, our method is unable to match it.

# SUMMARY RESULT

- Percentage of school matching types

| | |
|---|---|
| Exact Match | 78.0% (346) |
| Prospective Match | 8.89 % (39) |
| New School | 13.1 % (56) |
| Unmatched | 0.004% (2) |

# SUMMARY RESULT



# SUMMARY RESULT

- Two unmatched schools are

    1. University of Massachusetts Global
    2. Loma Linda University

    => Further investigation confirms that these two are new schools, with data only for 2023 issue year.

# ABOUT FINAL PRODUCT

We modified the codes and turned into a function called '**name_cleanning_function**'

**INPUT:** Three parameters

- raw.data - a excel file
- clean.data - a csv file
- year - the issue year that the user wants to standardize the school names and metrics

# ABOUT FINAL PRODUCT

**OUTPUT:** A list with two attributes: school_name_df and metric_df.

1. **`school_name_df`** attribute is a table with four columns:
   - ❏ `Raw_name` : unstandardized school names as in raw data set.
   - ❏ `Suggested_Name`: recommended school names after matching.
   - ❏ `Match_Type`: how the school name was matched.
   - ❏ `Similarity_Score`: numerical variable indicating similarity score.

Number of rows in **`school_name_df`** table equals the number of unique school names for the given year.

# ABOUT FINAL PRODUCT

**OUTPUT**: A list with two attributes, including **`school_name_df`** and **`metric_df`.**

2. **`metric_df`** attribute is a table containing three columns and number of row equal to number of unique metrics used in the given year.

- **`Raw_Metric` :** is the unstandardized metrics as in raw data set.
- **`Metric_Similarity_Score`** contains a numerical variable indicating the similarity score between the original metric name and the recommended metric name.
- **`Suggested_Metric`** contains recommended metric names after matching.

# APPEARANCE OF SCHOOL NAME ATTRIBUTE

| Raw_Name | Suggested_Name | Math_Type | Similarity_Score |
|---|---|---|---|
| Abilene Christian University | Abilene Christian University | New School | 0.0000000 |
| Adelphi University | Adelphi University | Exact Match | 1.0000000 |
| Alabama State University | Alabama State University | New School | 0.0000000 |
| Albizu University–Miami | Albizu University–Miami | New School | 0.0000000 |
| Alliant International University | Alliant International University | Exact Match | 1.0000000 |
| Alvernia University | Alvernia University | New School | 0.0000000 |
| American International College | American International College | Exact Match | 1.0000000 |
| American University | American University | Exact Match | 1.0000000 |
| Andrews University | Andrews University | Exact Match | 1.0000000 |
| Arizona State University | Arizona State University | Exact Match | 1.0000000 |
| Arkansas State University | Arkansas State University | Exact Match | 1.0000000 |
| Auburn University | Auburn University | Exact Match | 1.0000000 |
| Augusta University | Augusta University | Exact Match | 1.0000000 |
| Aurora University | Aurora University | Exact Match | 1.0000000 |
| Azusa Pacific University | Azusa Pacific University | Exact Match | 1.0000000 |
| Baker University | Baker University | Exact Match | 1.0000000 |
| Ball State University | Ball State University | Exact Match | 1.0000000 |
| Barry University | Barry University | Exact Match | 1.0000000 |
| Baylor University | Baylor University | Exact Match | 1.0000000 |
| Belhaven University | Belhaven University | New School | 0.0000000 |
| Bellarmine University | Bellarmine University | Exact Match | 1.0000000 |
| Belmont University | Belmont University | Exact Match | 1.0000000 |
| Bethel University (MN) | Bethel University (MN) | Exact Match | 1.0000000 |
| Binghamton University–SUNY | Binghamton University–SUNY | Exact Match | 1.0000000 |
| Biola University | Biola University | Exact Match | 1.0000000 |
| Boise State University | Boise State University | Exact Match | 1.0000000 |
| Boston College | Boston College | Exact Match | 1.0000000 |
| Boston University | Boston University | Exact Match | 1.0000000 |
| Bowling Green State University | Bowling Green State University | Exact Match | 1.0000000 |
| Bradley University | Bradley University | New School | 0.0000000 |
| Brandeis University | Brandeis University | Exact Match | 1.0000000 |

# APPEARANCE OF METRIC ATTRIBUTE

| Raw_Metric | Suggested_Metric | Metric_Similarity_Score |
|---|---|---|
| % of classes with 50 or more students | % of classes with 50 or more students | 1.0000000 |
| % of classes with fewer than 20 students | % of classes with fewer than 20 students | 1.0000000 |
| % of faculty who are full-time | % of faculty who are full-time | 1.0000000 |
| % of first year student in top 10 percent of high school class | % of first year student in top 10 percent of high school class | 1.0000000 |
| % of full-time faculty with Ph.D. or terminal degree | % of full-time faculty with Ph.D. or terminal degree | 1.0000000 |
| 6-year graduation rate of students who did not receive a Pell grant or subsidized Stafford loan | 6-year graduation rate of students who did not receive a Pell grant or subsidized Stafford loan | 1.0000000 |
| 6-year graduation rate of students who received a subsidized Stafford loan | 6-year graduation rate of students who received a subsidized Stafford loan | 1.0000000 |
| Acceptance rate | Acceptance rate | 1.0000000 |
| ACT Composite average score | ACT Composite average score | 1.0000000 |
| Average 6-year graduation rate | Average 6-year graduation rate | 1.0000000 |
| Average alumni giving rate | Average alumni giving rate | 1.0000000 |
| Average first year student retention rate | Average first year student retention rate | 1.0000000 |
| Average freshman retention rank | Average freshman retention rank | 1.0000000 |
| Difference between six-year graduation rates of Pell and non-Pell Grant recipients | Difference between six-year graduation rates of Pell and non-Pell Grant recipients | 1.0000000 |
| Educational expenditures per student | Educational expenditures per student | 1.0000000 |
| Graduate Indebtedness rank | Graduate Indebtedness rank | 1.0000000 |
| Graduation and retention rank | Graduation and retention rank | 1.0000000 |
| Overall Rank | Overall Rank | 1.0000000 |
| Overall score | Overall score | 1.0000000 |
| Peer assessment score | Peer assessment score | 1.0000000 |
| SAT Math average score | SAT Math average score | 1.0000000 |
| Student/faculty ratio | Student/faculty ratio | 1.0000000 |
| SAT Evidence-Based Reading & Writing average score | SAT Evidence-Based Reading & Writing average score | 1.0000000 |
| 6-year graduation rate of students who did not receive a Pell Grant | 6-year graduation rate of students who did not receive a Pell Grant | 0.9970588 |
| Pell grant comparative graduation rate rank* | Pell grant comparative graduation rate rank | 0.9954545 |
| Pell grant graduation rate rank* | Pell grant graduation rate rank | 0.9937500 |
| Class Size Index rank* | Class Size Index rank | 0.9909091 |
| Alumni giving rank* | Alumni giving rank | 0.9894737 |
| Average Faculty Compensation (CY) | Average Faculty Compensation (CY) | 0.9757576 |

# UTILIZING THE FUNCTION : A QUICK GUIDE

- To use the name_cleaning_function in R, the user should first save the function file in the same directory as the raw dataset and cleaned dataset files. Next, the user needs to add this line of code **source("name_cleaning_function.R")** in order to source the function.
- *Side notes:* The default dataset format for this function is an excel file. If the raw dataset is not in an excel file format, the user may need to make slight modifications to the function code or convert the raw dataset into an excel file format before using the function.
- Use $ in order to access the each attribute of output.

# LIMITATIONS

- Multiple rounds of pairing and filtering limited the efficiency and speed of the function
- Determining whether a school is new or not still requires manual work and expertise in the field.
- These limitations highlight the need for further research and development in this area.

# Future Study

- Incorporate techniques such as Natural Language Processing (NLP) or Web Crawler to enhance accuracy in identifying and filtering of relevant data.
- Consider using other variables such as school.ID and IPEDS.ID to enhance the project's functionality and improve its processing speed and efficiency.

# GROUP CHALLENGES

- The limited knowledge and experience in text mining proved challenging at the outset, which led to delays in starting the project.
- Various challenges, such as stemming and normalizing text, identifying relevant data.

# ACKNOWLEDGEMENTS

We would like to convey our immense appreciation to Dr. Esfandiari and Ms. Huang for their exceptional teaching and unwavering support, which have been invaluable to us throughout this project.

Last but not least, we extend our sincere thanks to Dr. Sugano and Dr. Zhang for their invaluable advice and thoughtful guidance, which have been instrumental in shaping our project.

# SUPPLEMENTARY SOURCES

Article about Jaro Winkler vs Levenshtein Distance

Video explaining Jaro Distance Winkler

Video explaining Levenshtein Distance

# REFERENCES

Nam, N. (2022, Dec) Approximate String Matching in R using Jaro-Winkler Similarity

*https://blog.devgenius.io/approximate-string-matching-in-r-using-jaro-winkler-similarity-a93436ecf38f*

Wijffels, J. (2020, March) word2vec in R

*http://www.bnosac.be/index.php/blog/100-word2vec-in-R?fbclid=IwAR0iE_89XPuGRmLfa9oRQ98qj2-TBczFZj_RmhlqFzy-wtVOYayBXzrn0r8*

# *THANK YOU*