WMU PROGRAM OF STUDY

Presented By:
Samiksha Chopade – 271188199
Shrawani Palande – 813202795
Mohana Alapati – 344223110

RECAP

Tools for webscraping

Data Preprocessing

Analysis of Transcript and Catalog

Data Comparison, Course Matching

Trying the method on different transcripts

Generating Outputs Seperately

Generating the Latex Template

Final Program of study

CHANGES IN COURSE_DF

```
Core:
STAT 6620 - Applied Linear Models Credits: 3 hours
STAT 5860 - Computer Based Data Analysis Credits: 3 hours
STAT 5870 - Big Data Analysis Using Python Credits: 3 hours
STAT 6800 - SAS Programming Credits: 3 hours
CS 5430 - Database Systems Credits: 3 hours
CS 5610 - Advanced R Programming for Data Science Credits: 4 hours
CS 5821 - Machine Learning Credits: 3 hours
CS 6100 - Advanced Storage, Retrieval and Processing of Big Data Credits: 3 hours

Electives:
STAT 6040 - Fundamentals of Epidemiology and Clinical Trials
STAT 6500 - Statistical Theory I
STAT 6600 - Statistical Theory II
STAT 6640 - Applied Mixed Models
CS 6030 - Studies in Computer Science
CS 6260 - Advanced Parallel Computations
CS 6310 - Advanced Design and Analysis of Algorithms
CS 6430 - Database Management System Implementation
CS 6530 - Data Mining
CS 6820 - Advanced Artificial Intelligence
CS 6821 - Information Retrieval

Masters Project:
STAT 6970 - Data Science Masters Project
CS 6970 - Master’s Project
```

THE OUTPUT

```
In [38]: merge_df
```

Out[38]:

	Course Code	Course Name	Category	Credits	Grade	Semester	Year
0	STAT 6620	Applied Linear Models	Core	3 hours	A	Fall	2023
1	STAT 5860	Computer Based Data Analysis	Core	3 hours	A	Spring	2023
2	STAT 5870	Big Data Analysis Using Python	Core	3 hours	A	Summer1	2023
3	STAT 6800	SAS Programming	Core	3 hours			
4	CS 5430	Database Systems	Core	3 hours	A	Fall	2023
5	CS 5610	Advanced R Programming for Data Science	Core	4 hours	A	Spring	2023
6	CS 5821	Machine Learning	Core	3 hours		Spring	2024
7	CS 6100	Advanced Storage, Retrieval and Processing of ...	Core	3 hours	A	Fall	2023
8	STAT 6040	Fundamentals of Epidemiology and Clinical Trials	Electives	3 hours			
9	STAT 6500	Statistical Theory I	Electives	3 hours			
10	STAT 6600	Statistical Theory II	Electives	3 hours			
11	STAT 6640	Applied Mixed Models	Electives	3 hours			
12	CS 6030	Studies in Computer Science	Electives	3 hours			
13	CS 6260	Advanced Parallel Computations	Electives	3 hours			
14	CS 6310	Advanced Design and Analysis of Algorithms	Electives	3 hours			
15	CS 6430	Database Management System Implementation	Electives	3 hours			
16	CS 6530	Data Mining	Electives	3 hours			
17	CS 6820	Advanced Artificial Intelligence	Electives	3 hours		Spring	2024
18	CS 6821	Information Retrieval	Electives	3 hours			
19	STAT 6970	Data Science Masters Project	Masters Project	4 hours		Spring	2024
20	CS 6970	Master's Project	Masters Project	4 hours			

EXAMPLE TRANSCRIPTS

Output Generation Code

In [57]:

```
import pandas as pd
from tabulate import tabulate
import subprocess

# Sample data
data = {
    "Course Code": ['STAT 6620', 'STAT 5860', 'STAT 5870', 'STAT 6800', 'CS 5430'],
    "Course Name": ['Applied Linear Models', 'Computer Based Data Analysis', 'Big Data Analysis Using Python', 'SAS Programm
    "Credits": [3, 3, 3, 3, 3],
    "Grade": ['A', 'A', '', '', 'A'],
    "Semester": ['Fall', 'Spring', '', '', 'Fall'],
    "Year": [2023, 2023, '', '', 2023]
}

# Create a DataFrame
merged_df = pd.DataFrame(data)

# Convert DataFrame to LaTeX table with borders
latex_table = tabulate(merged_df, headers='keys', tablefmt='latex_raw')

# Add LaTeX document preamble and tabular environment
latex_content = r"""
\begin{tabular}{lllllll}
\hline
    &   Course Code &                                 Course Name &   Credits &   Grade &   Semester & Year   \\
\hline
 0  & STAT 6620     & Applied Linear Models                       & 3         & A       & Fall       & 2023   \\
 1  & STAT 5860     & Computer Based Data Analysis                & 3         & A       & Spring     & 2023   \\
 2  & STAT 5870     & Big Data Analysis Using Python              & 3         &         &            &        \\
 3  & STAT 6800     & SAS Programming                             & 3         &         &            &        \\
 4  & CS 5430       & Database Systems                            & 3         & A       & Fall       & 2023   \\
 5  & CS 5610       & Advanced R Programming for Data Science     & 4         &         &            &        \\
 6  & CS 5821       & Machine Learning                            & 3         &         & Spring     & 2024   \\
 7  & CS 6100       & Advanced Storage, Retrieval and Processing of Big Data & 3         &         &            &        \\
 8  & STAT 6040     & Fundamentals of Epidemiology and Clinical Trials & 3         &         &            &        \\
 9  & STAT 6500     & Statistical Theory I                        & 3         &         &            &        \\
 10 & STAT 6600     & Statistical Theory II                       & 3         &         &            &        \\
 11 & STAT 6640     & Applied Mixed Models                        & 3         &         &            &        \\
 12 & CS 6030       & Studies In Computer Science                 & 3         &         &            &        \\
 13 & CS 6260       & Advanced Parallel Computations              & 3         &         &            &        \\
 14 & CS 6310       & Advanced Design and Analysis of Algorithms  & 3         &         &            &        \\
 15 & CS 6430       & Database Management System Implementation   & 3         &         &            &        \\
 16 & CS 6530       & Data Mining                                 & 3         &         &            &        \\
 17 & CS 6820       & Advanced Artificial Intelligence            & 3         &         & Spring     & 2024   \\
 18 & CS 6821       & Information Retrieval                       & 3         &         &            &        \\
 19 & STAT 6970     & Data Science Masters Project                & 4         &         & Spring     & 2024   \\
 20 & CS 6970       & Master's Project                            & 4         &         &            &        \\
 21 & CS 5610       & Advanced R for data Science                 & 4         & A       & Spring     & 2023   \\
 22 & STAT 5870     & Big Data Analysis Using python              & 3         & A       & Summer1    & 2023   \\
 23 & CS 6100       & Advance Storage, Retrieval and Processing of Big Data & 3         & A       & Fall       & 2023   \\
\hline
\end{tabular}
"""

latex_content += latex_table

# Add LaTeX document end
latex_content += r"""
\end{tabular}
\end{document}
"""

# Write LaTeX content to a .tex file
with open("table_with_borders.tex", "w") as f:
    f.write(latex_content)

# Compile LaTeX file into a PDF using pdflatex
subprocess.run(['pdflatex', 'table_with_borders.tex'])

print("PDF generated successfully.")
```

```
PDF generated successfully.
```

Generated pdf

Course Code	Course Name	Credits	Grade	Semester	Year
STAT 6620	Applied Linear Models	3	A	Fall	2023
STAT 5860	Computer Based Data Analysis	3	A	Spring	2023
STAT 5870	Big Data Analysis Using Python	3			
STAT 6800	SAS Programming	3			
CS 5430	Database Systems	3	A	Fall	2023

Latex Code Using Python

In [26]:

```
import pandas as pd
from tabulate import tabulate

merged_df = merged_df.fillna('')

latex_table = tabulate(merged_df, headers='keys', tablefmt='latex', colalign=('l' * len(merged_df.columns)))
print(latex_table)
```

```
\begin{tabular}{lllllll}
\hline
    &   Course Code &                                          Course Name &   Credits &   Grade &   Semester & Year   \\
\hline
 0  & STAT 6620     & Applied Linear Models                                & 3         & A       & Fall       & 2023   \\
 1  & STAT 5860     & Computer Based Data Analysis                         & 3         & A       & Spring     & 2023   \\
 2  & STAT 5870     & Big Data Analysis Using Python                       & 3         &         &            &        \\
 3  & STAT 6800     & SAS Programming                                      & 3         &         &            &        \\
 4  & CS 5430       & Database Systems                                     & 3         & A       & Fall       & 2023   \\
 5  & CS 5610       & Advanced R Programming for Data Science              & 4         &         &            &        \\
 6  & CS 5821       & Machine Learning                                     & 3         &         & Spring     & 2024   \\
 7  & CS 6100       & Advanced Storage, Retrieval and Processing of Big Data & 3         &         &            &        \\
 8  & STAT 6040     & Fundamentals of Epidemiology and Clinical Trials     & 3         &         &            &        \\
 9  & STAT 6500     & Statistical Theory I                                 & 3         &         &            &        \\
 10 & STAT 6600     & Statistical Theory II                                & 3         &         &            &        \\
 11 & STAT 6640     & Applied Mixed Models                                 & 3         &         &            &        \\
 12 & CS 6030       & Studies in Computer Science                          & 3         &         &            &        \\
 13 & CS 6260       & Advanced Parallel Computations                       & 3         &         &            &        \\
 14 & CS 6310       & Advanced Design and Analysis of Algorithms           & 3         &         &            &        \\
 15 & CS 6430       & Database Management System Implementation            & 3         &         &            &        \\
 16 & CS 6530       & Data Mining                                          & 3         &         &            &        \\
 17 & CS 6820       & Advanced Artificial Intelligence                     & 3         &         & Spring     & 2024   \\
 18 & CS 6821       & Information Retrieval                                & 3         &         &            &        \\
 19 & STAT 6970     & Data Science Masters Project                         & 4         &         & Spring     & 2024   \\
 20 & CS 6970       & Master’s Project                                     & 4         &         &            &        \\
 21 & CS 5610       & Advanced R for data Science                          & 4         & A       & Spring     & 2023   \\
 22 & STAT 5870     & Big Data Analysis Using python                       & 3         & A       & Summer1    & 2023   \\
 23 & CS 6100       & Advance Storage, Retrieval and Processing of Big Data  & 3         & A       & Fall       & 2023   \\
\hline
\end{tabular}
```

Latex Code

final latex1.tex - TeXworks

File Edit Search Format Typeset Scripts Window Help

pdfLaTeX+MakeIndex+BibTeX

```
\documentclass[10pt]{article}
\usepackage[margin=1in]{geometry}
\usepackage{booktabs} % For professional looking tables

\begin{document}

\title{Data Science Program of Study}
\author{Dr. Kevin Lee}
\date{\today}
\maketitle

\section*{Personal Details}
\begin{tabular}{ll}
   \textbf{Name:} & \\
   \textbf{WIN ID:} & \\
   \textbf{College:} & \\
   \textbf{Major and Department:} & \\
\end{tabular}

\section*{Courses}
\begin{tabular}{llllll}
\toprule
Course Code & Course Name & Category & Credits & Grade & Semester & Year \\
\midrule
STAT 6620 &                    Applied Linear Models &          Core &  3 hours &     A &     Fall &  2023 \\
\hline
STAT 5860 &                Computer Based Data Analysis &          Core &  3 hours &     A &   Spring &  2023 \\
\hline
STAT 5870 &                Big Data Analysis Using Python &          Core &  3 hours &     A &  Summer1 &  2023 \\
\hline
STAT 6800 &                             SAS Programming &          Core &  3 hours &       &          &       \\
\hline
CS 5430 &                          Database Systems &          Core &  3 hours &     A &     Fall &  2023 \\
\hline
 CS 5610 &       Advanced R Programming for Data Science &          Core &  4 hours &     A &   Spring &  2023 \\
```

final latex1.tex* - TeXworks

File Edit Search Format Typeset Scripts Window Help

pdfLaTeX+MakeIndex+BibTeX

```
\hline
CS 5821 &                    Machine Learning &          Core &  3 hours &       &   Spring &  2024 \\
\hline
 CS 6100 &  Advanced Storage, Retrieval and Processing of ... &          Core &  3 hours &     A &     Fall &  2023 \\
\hline
STAT 6040 &  Fundamentals of Epidemiology and Clinical Trials &      Electives &  3 hours &       &          &       \\
\hline
STAT 6500 &                    Statistical Theory I &      Electives &  3 hours &       &          &       \\
\hline
 STAT 6600 &                   Statistical Theory II &      Electives &  3 hours &       &          &       \\
\hline
 STAT 6640 &                    Applied Mixed Models &      Electives &  3 hours &       &          &       \\
\hline
 CS 6030 &                 Studies in Computer Science &      Electives &  3 hours &       &          &       \\
\hline
CS 6260 &               Advanced Parallel Computations &      Electives &  3 hours &       &          &       \\
\hline
CS 6310 &      Advanced Design and Analysis of Algorithms &      Electives &  3 hours &       &          &       \\
\hline
CS 6430 &      Database Management System Implementation &      Electives &  3 hours &       &          &       \\
\hline
CS 6530 &                          Data Mining &      Electives &  3 hours &       &          &       \\
\hline
CS 6820 &            Advanced Artificial Intelligence &      Electives &  3 hours &       &   Spring &  2024 \\
\hline
CS 6821 &                 Information Retrieval &      Electives &  3 hours &       &          &       \\
\hline
STAT 6970 &              Data Science Masters Project &  Masters Project &  4 hours &       &   Spring &  2024 \\
\hline
CS 6970 &                    Master's Project &  Masters Project &  4 hours &       &          &       \\


\bottomrule
\end{tabular}
\end{document}
```

Latex Output Format

Data Science Program of Study

Dr. Kevin Lee

April 4, 2024

Personal Details

Name:
WIN ID:
College:
Major and Department:

Courses

Course Code	Course Name	Category	Credits	Grade	Semester	Year
STAT 6620	Applied Linear Models	Core	3 hours	A	Fall	2023
STAT 5860	Computer Based Data Analysis	Core	3 hours	A	Spring	2023
STAT 5870	Big Data Analysis Using Python	Core	3 hours	A	Summer1	2023
STAT 6800	SAS Programming	Core	3 hours			
CS 5430	Database Systems	Core	3 hours	A	Fall	2023
CS 5610	Advanced R Programming for Data Science	Core	4 hours	A	Spring	2023
CS 5821	Machine Learning	Core	3 hours		Spring	2024
CS 6100	Advanced Storage, Retrieval and Processing of ...	Core	3 hours	A	Fall	2023
STAT 6040	Fundamentals of Epidemiology and Clinical Trials	Electives	3 hours			
STAT 6500	Statistical Theory I	Electives	3 hours			
STAT 6600	Statistical Theory II	Electives	3 hours			
STAT 6640	Applied Mixed Models	Electives	3 hours			
CS 6030	Studies in Computer Science	Electives	3 hours			
CS 6260	Advanced Parallel Computations	Electives	3 hours			
CS 6310	Advanced Design and Analysis of Algorithms	Electives	3 hours			
CS 6430	Database Management System Implementation	Electives	3 hours			
CS 6530	Data Mining	Electives	3 hours			
CS 6820	Advanced Artificial Intelligence	Electives	3 hours		Spring	2024
CS 6821	Information Retrieval	Electives	3 hours			
STAT 6970	Data Science Masters Project	Masters Project	4 hours		Spring	2024
CS 6970	Master's Project	Masters Project	4 hours			

Content For Final Presentation

- The Whole Coding Process
- Final Latex Templates
- Challenges and Solutions


Thank
You