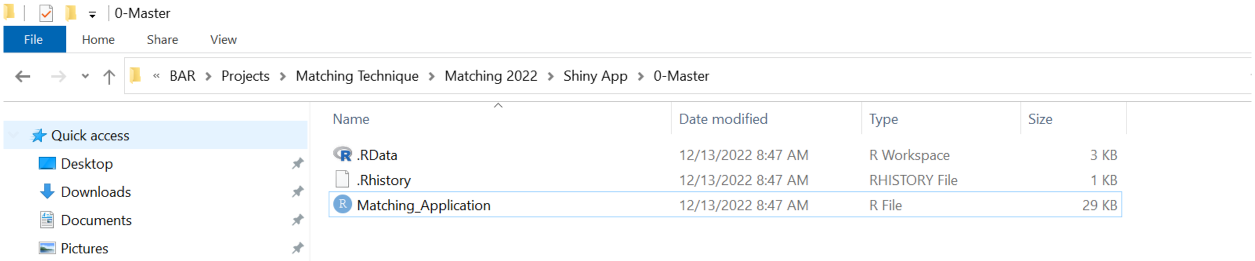
# Guide for Using Matching Tool for Assessment

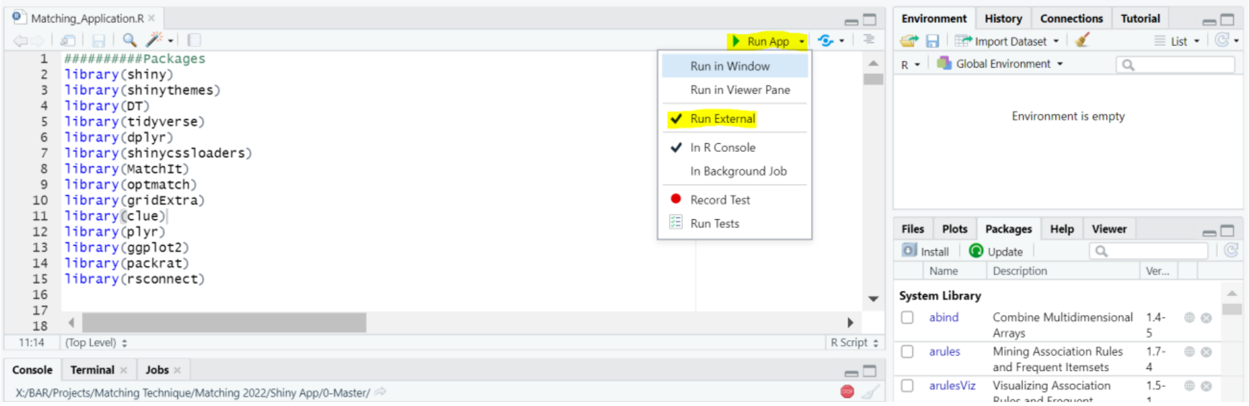
1. Double Click on the “Matching_Application.R” file in file explorer to open the application in R Studio.



« BAR › Projects › Matching Technique › Matching 2022 › Shiny App › 0-Master

| Name | Date modified | Type | Size |
|---|---|---|---|
| .RData | 12/13/2022 8:47 AM | R Workspace | 3 KB |
| .Rhistory | 12/13/2022 8:47 AM | RHISTORY File | 1 KB |
| Matching_Application | 12/13/2022 8:47 AM | R File | 29 KB |

This should open the program in R Studio (you should see something like the below).



```
##########Packages
library(shiny)
library(shinythemes)
library(DT)
library(tidyverse)
library(dplyr)
library(shinycssloaders)
library(MatchIt)
library(optmatch)
library(gridExtra)
library(clue)
library(plyr)
library(ggplot2)
library(packrat)
library(rsconnect)
```

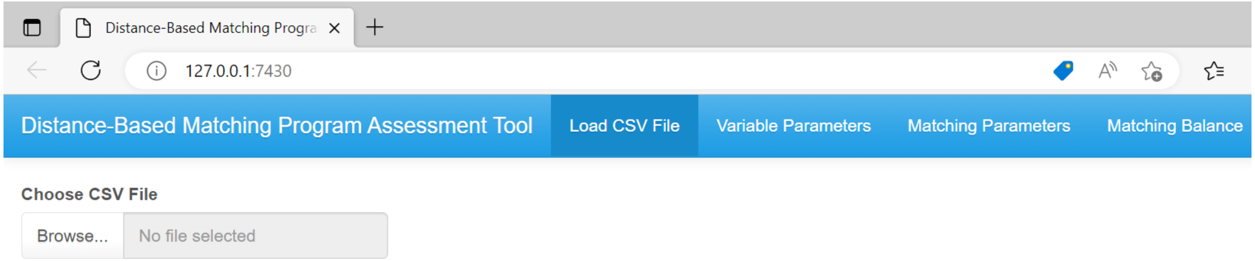
X:/BAR/Projects/Matching Technique/Matching 2022/Shiny App/0-Master/

| | Name | Description | Ver... |
|---|---|---|---|
| □ | abind | Combine Multidimensional Arrays | 1.4-5 |
| □ | arules | Mining Association Rules and Frequent Itemsets | 1.7-4 |
| □ | arulesViz | Visualizing Association Rules and Frequent | 1.5-1 |

2. Find the Green “Run App” Arrow at the top right of the R Script (top left quadrant and highlighted in yellow above) and click on the drop-down carrot to the right of the icon. Then, select “Run External.” This will set the default option to open the application in a separate browser, and the application will work best with this default setting.

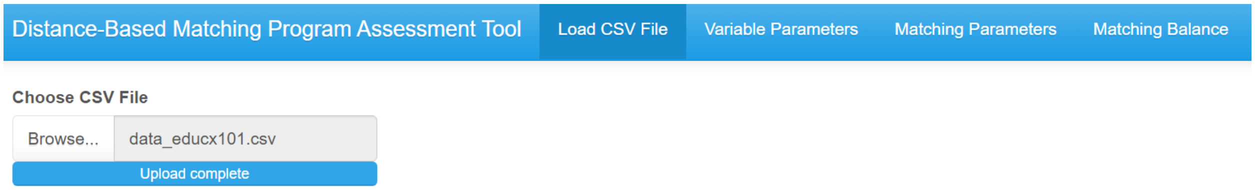
3. After changing the default settings of the run application step as outlined in step 3, click the green “Run App” Arrow to run the application. After doing so, the application should appear in a new browser window like displayed below.

   - Note: if running this for the first time on your machine, you may need to install some packages prior to running the app. If so, a pop up should appear at the top of the script mentioning that you need to install packages. Click “Install” to install the packages. After doing this, proceed with running the application or clicking on the “Run App” green arrow.



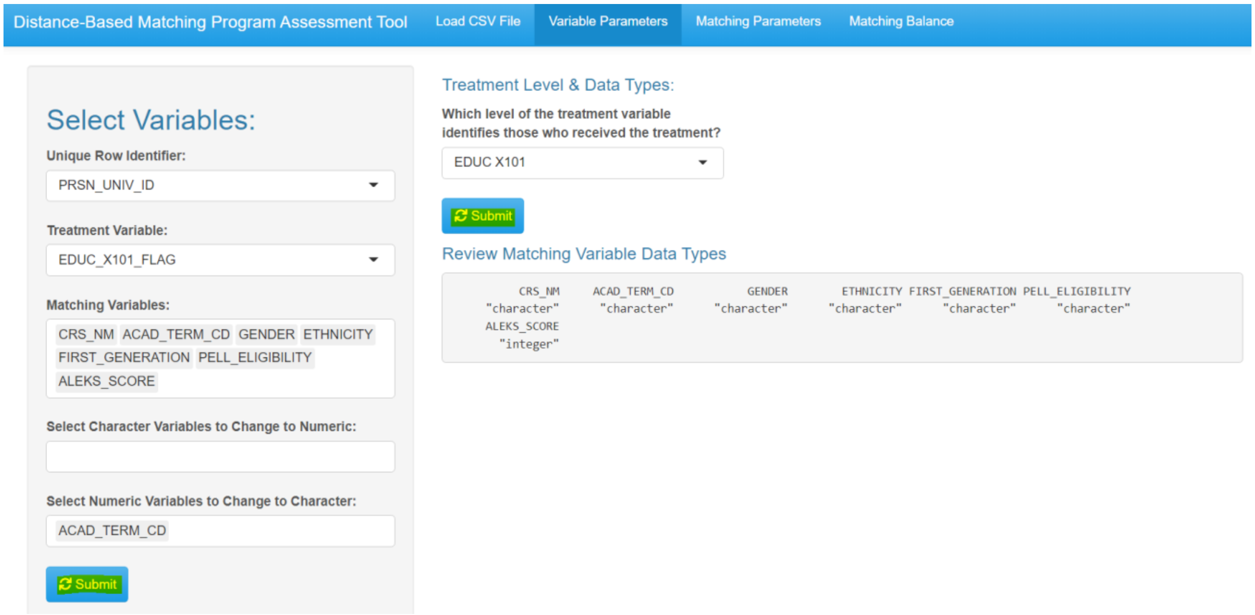
You will use the application by following the menus from left to right, starting with the "Load CSV File" menu.

4. In the "Load CSV File" menu, click on the "Browse…" button. This will allow you to use file explorer to navigate to a CSV file that contains the assessment data for your analysis. After selecting the CSV file through the file explorer, the application should give you an "Upload complete" message.



- Specifications for CSV file:
  - IMPORTANT: FILE GRANULARITY - This CSV file should contain ONLY ROW PER UNIT OF ANALYSIS (i.e. student). There should be an identifier column (i.e. ID column) that identifies a unique row in the dataset. You should ensure there is only one record per the unique row identifier in the dataset, and if not, you should apply deduplication logic on your CSV file to choose one record per each unit of analysis prior to using the application. Otherwise, the application may not produce expected results or produce duplicate matches.
  - Columns with attributes to be used for matching (demographic variables, performance variables, outcome variables, etc.) should be provided as columns in the CSV file.

5. Navigate to the "Variable Parameters" tab and use the sub menus to select the applicable variables for your dataset. Once you navigate and make selections, click the submit buttons to populate options in the "Treatment Level and Data Types" section and to load the selections in the matching program. The below provides an example and steps as a guide.



Select Variables:

Unique Row Identifier: PRSN_UNIV_ID

Treatment Variable: EDUC_X101_FLAG

Matching Variables: CRS_NM ACAD_TERM_CD GENDER ETHNICITY FIRST_GENERATION PELL_ELIGIBILITY ALEKS_SCORE

Select Character Variables to Change to Numeric:

Select Numeric Variables to Change to Character: ACAD_TERM_CD

Treatment Level & Data Types:

Which level of the treatment variable identifies those who received the treatment? EDUC X101

Review Matching Variable Data Types

```
      CRS_NM   ACAD_TERM_CD         GENDER      ETHNICITY FIRST_GENERATION PELL_ELIGIBILITY
 "character"    "character"    "character"    "character"      "character"      "character"
 ALEKS_SCORE
   "integer"
```

a. Research Question: Does a co-curricular math support class (EDUC-X101) help students' outcomes in finite mathematics?

b. CSV file: The CSV file contains one record per student (PRSN_UNIV_ID) and contains some of the following information as attributes:

    i. CRS_NM = The specific finite mathematics course that the student is taking at IU Bloomington (MATH-M118, MATH-D116, MATH-D117)
    ii. ACAD_TERM_CD = The term the student is taking finite mathematics at IU Bloomington (4168 – Fall 2016, etc.)
    iii. GENDER = Student's gender
    iv. ETHNICITY = Student's ethnicity
    v. FIRST_GENERATION = Student's first-generation status
    vi. PELL_ELIGIBILITY = Student's status of receiving a Pell grant
    vii. ALEKS_SCORE = Student's score on the Math placement exam
    viii. EDUC_X101_FLAG = Field with two levels (EDUC X101 and NON EDUC X101) identifier which indicates whether the student enrolled in a EDUC-X101 course section for additional instruction in finite math

c. Unique Row Identifier = field in CSV file that identifies the unique unit of analysis (PRSN_UNIV_ID)

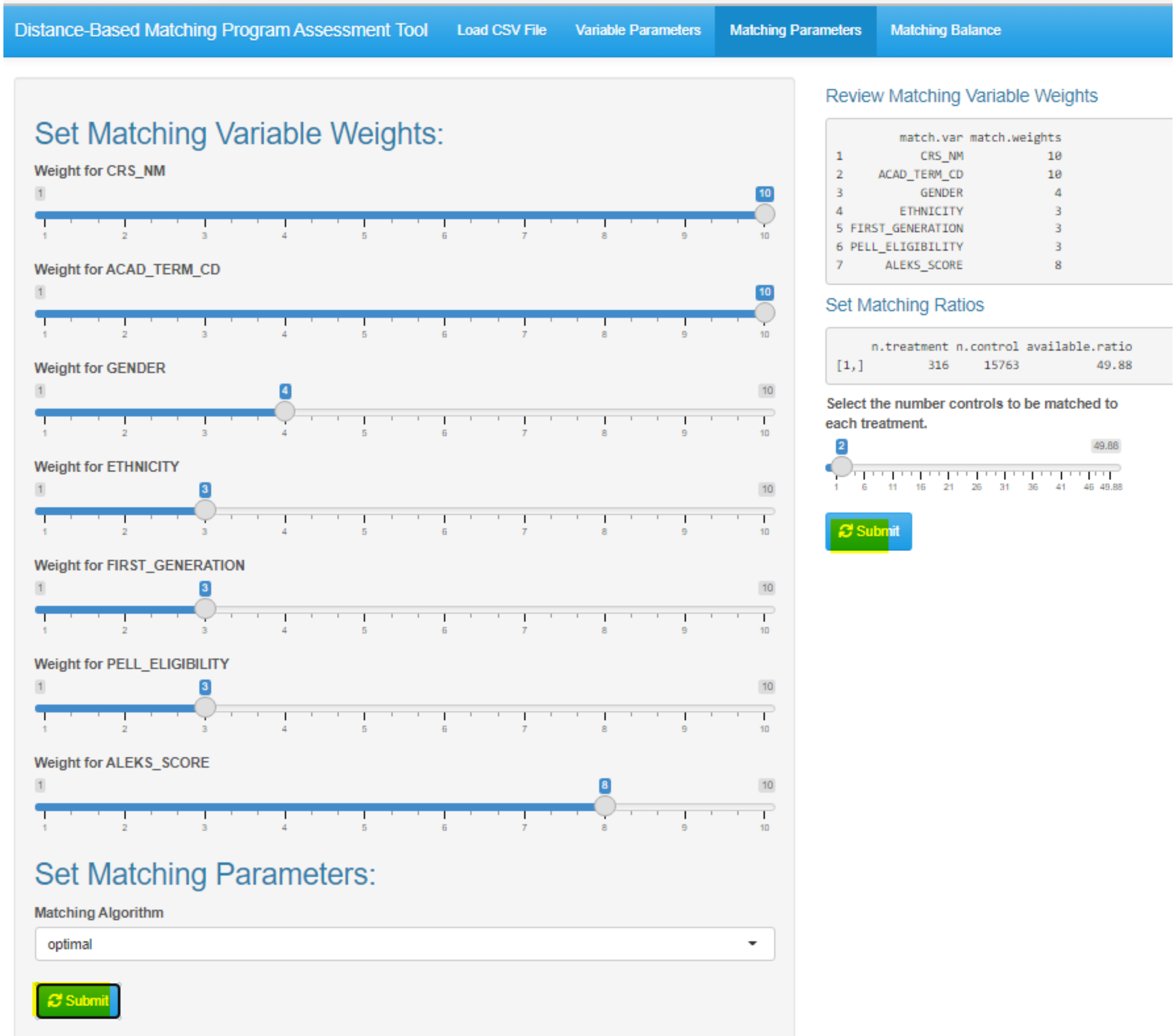
d. Treatment Variable = field in CSV file that can be used to identify who received the treatment, participated in the program, or participated in the intervention that is the subject of the analysis (EDUC_X101_FLAG)

e. Matching Variables = set of fields in the CSV file for which you want to match or create a combined distance score to identify students who did not receive the treatment but who are as similar as possible to the students who received the treatment (CRS_NM, ACAD_TERM_CD, GENDER, ETHNICITY, FIRST_GENERATION, PELL_ELIGIBILITY, ALEKS_SCORE). Do NOT include outcome measures as matching variables.

f. Select Character Variables to Change to Numeric = this loads a list of variables that the application read as a character variable. Select variables here that need changed to numeric variables prior to matching.
g. Select Numeric Variables to Change to Character = this loads a list of variables that the application read as a numeric variable. Select variables here that need changed to character variables prior to matching. Here, we select ACAD_TERM_CD here because we will be matching on that the variable, and the value signifies a distinct academic term and not a number.
h. Click Submit to populate the information in the “Treatment Level and Data Types” section.
i. Which level of the treatment variable identifies those who received the treatment? = select the level that identifies those who received the treatment, participated in the program, participated in the intervention, etc. Here, EDUC X101 identifies students who participated in supplemental finite mathematics instruction, so we set this parameter on this level.
j. Click Submit so the matching level selected in “I” is loaded in the matching program.
k. Review the data types for the selected matching variables to ensure those variables have correct data types before proceeding to the “Matching Parameters” menu. If they are not correct, use the “Select Character Variables to Change to Numeric” and “Select Numeric Variables to Change to Character” drop downs to change data types.

6. Navigate to the “Matching Parameters” tab. Use this this tab to select matching weights for each variable (if applicable, see matching type), select the matching algorithm, and select the number of controls to be matched to each treatment. Once you navigate and make selections, click the submit buttons to review selected weights in the “Review Matching Variable Weights” section and to load the selections in the matching program. The below provides information, examples, and steps as a guide.

a. Set Matching Variable Weights = This section populates a slider for each matching variable selected in the “Variable Parameters” menu. **Note that this section will only impact the matching process if using distance-based matching. If you are deciding to use propensity score matching (matching through defining distance by the probability of being in the treatment group via an underlying logistic regression model), you can skip this section and move to b – Matching Type. If you still set weights here and then select propensity score matching as the Matching Type, these matching variable weights will have no impact on the propensity score matching algorithm.** For distance-based matching, these weights are used to factor in each variable’s importance in the combined distance measure calculation. There’s no science to selecting weights, but the below provides some description and best practices to guide how to select weights as a starting point. The distance measure calculation and how the weights fit into the calculation is also described a bit more below.



Set Matching Variable Weights:

Weight for CRS_NM: 10

Weight for ACAD_TERM_CD: 10

Weight for GENDER: 4

Weight for ETHNICITY: 3

Weight for FIRST_GENERATION: 3

Weight for PELL_ELIGIBILITY: 3

Weight for ALEKS_SCORE: 8

Set Matching Parameters:

Matching Algorithm: optimal

Review Matching Variable Weights

```
          match.var match.weights
1            CRS_NM            10
2      ACAD_TERM_CD            10
3            GENDER             4
4         ETHNICITY             3
5  FIRST_GENERATION             3
6  PELL_ELIGIBILITY             3
7        ALEKS_SCORE            8
```

Set Matching Ratios

```
     n.treatment n.control available.ratio
[1,]         316     15763           49.88
```

Select the number controls to be matched to each treatment.: 2

i. Selecting Weights

1. Some of the most important factors that might influence selecting higher weights for certain variables are 1) the desire to create a nested structure in the data to replicate the educational context for the analysis and 2) the desire to match on key characteristics tied to a program or intervention's mission.

   a. For example, we really want to make sure that each student who participated in the EDUC X101 intervention is matched with a student who did not participate in the EDUC X101 intervention but who took the same version of finite mathematics and took finite mathematics in the same term. Thus, to nest students within academic term and the type of finite math course taken, it motivates weighting ACAD_TERM_CD and CRS_NM the highest in the matching process (weight = 10).

   b. The mission of the co-curricular math support class (EDUC X101) is to provide a boost for students with lower math skills. Thus, we weight math preparation/math skill as measured through the Aleks Score fairly high in the matching process (weight = 8).

c. Program directors indicated women tend to self-select in the EDUC X101 co-curricular support class at higher rates more than men, so we weight gender slightly higher than other demographic variables (weight =4 for gender versus weight of 3 for other variables).

2. Aside from creating a nested structure within the data and weighting variables based on their relation to the program's mission being studied, another good general rule of thumb is to weight numerical variables higher than categorical variables in the matching process, whenever possible.

ii. Distance Measure

1. The distance measure used for distance-based matching in this application is largely based on the Gower distance formula (D'Orazio 2021), which allows for flexibility in terms of handling missing data and allowing for matching on mixed data types (numerical and categorical data).

2. The final distance between any two units (in this case, any two pairs of students) is calculated as a combined weighted average of the distances between the two units across all variables, where the weights for the weighted average represent the weights that are set in the application.

a. The distance for numerical variables is the absolute value of the difference between the two units on the variable.

b. The distance for categorical variables is calculated as 0 if the units match on the variable and 1 if they do not match.

c. Numerical variables are scaled between 0 and 1 before computing distances to prevent the weighted average distance measure from being skewed toward variables on a higher scale.

d. If there is a missing value on a variable on either of the two units for which you are calculating distance, the distance for that variable is automatically set to 1. So, it deals with missing data automatically.

b. Matching Type = This section allows for the specification of using distance-based matching, which matches units by leveraging the weights as set in 6ai to calculate a distance measure as described in 6aii, or the specification of propensity score matching which does not require the specification of weights, but instead builds a logistic regression model to estimate the probability of a unit being the in the treatment and then matches units who have similar predicted probabilities.

i. Note that specifying distance-based matching in this application will automatically handle missing values using the distance calculation philosophy as described in 6aii.

ii. Note that specifying propensity score matching in this application will NOT automatically handle missing values. If there are missing values in the underlying matching variables, the application will throw out any treatment or control units with any missing values on the matching covariates. If you do not want this listwise deletion, you will need to handle the missing data in the data set itself (i.e. imputation).

    iii. Some research suggests that distance-based matching generally produces better results because it focuses on minimizing the distance between the treatment and control units across all matching variables opposed to just matching units who have a similar probability for being in the treatment (propensity score matching) which can result in matching units that have similar probabilities but different underlying matching covariate values. However, for larger data sets, distance-based matching and propensity score matching should yield similar matching performance.

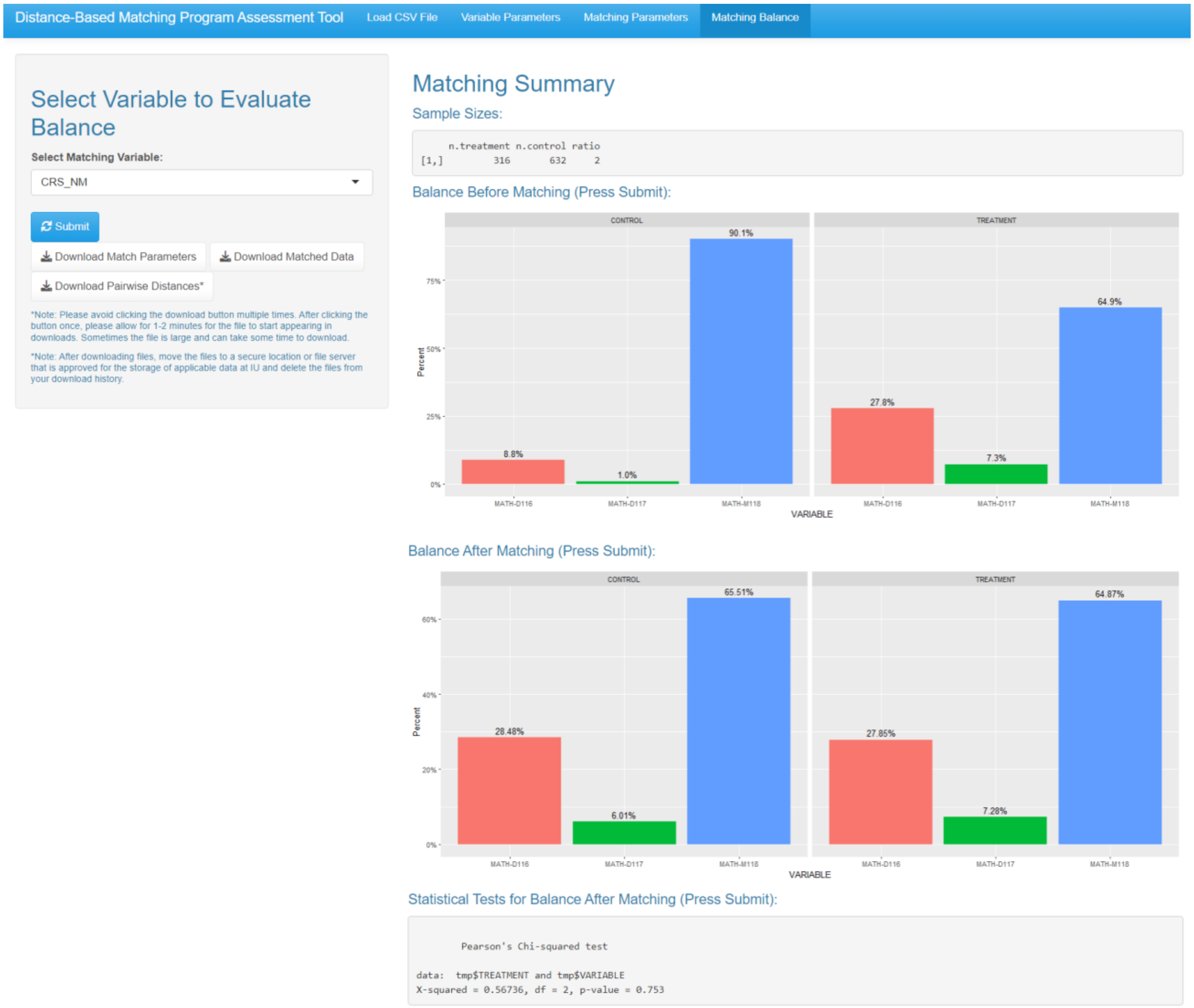
c. Matching Algorithm = This section allows for the user to match based on the nearest neighbors algorithm or an optimal match algorithm. Nearest neighbors is the more traditionally-used method which moves down the treatment list in a sequential manner and, each time, chooses to pair the treated unit with the closest control unit among the pool of remaining control units. As nearest neighbors pairs units in a sequential manner, it is sensitive to order and may produce higher quality matches at the start of the process and lower quality matches toward the end of the process as one moves further down the list and the pool of remaining controls declines. The optimal match algorithm is less sensitive to the order and tries to find optimal pairs across all possible pairs of treatment and controls such that a global distance is minimized. Studies have shown that these methods generally produce similar matched lists, but the optimal match is guaranteed to find the best combination of matched pairs that minimizes a global distance between controls and treatments. The optimal match is more computationally intensive method to run. As a general rule of thumb, try running the optimal match to start but default back to nearest neighbors if the program is taking too long to run on a large data set.

d. Click “Submit” after setting the variable weights and matching parameters

e. Select the Number of Controls to Be Matched to Each Treatment = This allows for setting a ratio for the number of controls to find per treatment as part of the matching process. The default value in the application and the most common ratio is one, or one-to-one matching, which matches one treated individual with exactly one control individual. It may be desired to match more than one control with each treatment to increase the sample size of the study. However, as one does that, the quality of matches in terms of the overlap in attribute characteristics between the treatment and control may decline. The literature generally suggests that a ratio of more than five generally yields poor match balance. To strategically set the matching ratio, you can toggle back and forth between “Matching Parameter” and “Matching Balance” menus by selecting different ratios and checking balance.

7. Navigate to the “Matching Balance” tab. Moving blue bar icons will display on the app while the matching process is running. Depending on the number of matching variables, size of the matching pool, and the selected matching algorithm, the matching process could take several

minutes to run. After the matching process is finished, a sample size summary will appear. And you can select a matching variable from the drop-down menu and click “Submit” to check matching balance for each variable individually. Finally, you can use download buttons to download the matched data set for analysis, a record of the matching parameters that were set in the application, etc. More details about matching balance are provided in the example below.


Distance-Based Matching Program Assessment Tool
Load CSV File
Variable Parameters
Matching Parameters
Matching Balance

Select Variable to Evaluate Balance

Select Matching Variable: CRS_NM

Submit
Download Match Parameters
Download Matched Data
Download Pairwise Distances*

*Note: Please avoid clicking the download button multiple times. After clicking the button once, please allow for 1-2 minutes for the file to start appearing in downloads. Sometimes the file is large and can take some time to download.

*Note: After downloading files, move the files to a secure location or file server that is approved for the storage of applicable data at IU and delete the files from your download history.

Matching Summary

Sample Sizes:

```
     n.treatment n.control ratio
[1,]         316       632     2
```

Balance Before Matching (Press Submit):

Balance After Matching (Press Submit):

Statistical Tests for Balance After Matching (Press Submit):

```

	Pearson's Chi-squared test

data:  tmp$TREATMENT and tmp$VARIABLE
X-squared = 0.56736, df = 2, p-value = 0.753
```

i. Matching Balance

   a. The goal of any matched study is to simulate a randomized control experiment as closely as possible by balancing characteristics between a treatment and control group. So, after the matching process, we want to ensure that the control group looks similar, if not nearly identical, to the treatment group with respect to each of the matching characteristics.

      i. Here, for the CRS_NM variable, we can see that prior to matching, 90% of students who did not participate in the EDUC X101 supplemental instruction course took the MATH-M118 version of finite compared to only 65% of the students who did participate in the EDUC X101 supplemental instruction course. However, after matching, the finite math course taking profile distribution is nearly identical between the treatment and control groups. This is confirmed with the formal chi-squared test results (p-value greater than the traditional significance levels of 0.01, 0.05, or 0.10). Here, we would say that the treatment and control groups are balanced on CRS_NM after the matching process.

ii. Keep in mind that statistical tests like the chi-squared test and t-test are sensitive to sample sizes and have a tendency to show significant results as sample size increases. Thus, always compare the statistical test results within the context of effect size or the typical differences displayed in the summary statistic charts.