Publication Data Cleaning in Excel for Author Name Gender Analysis

Stephen W Gabrielson[1], Rose L Turner[1]

[1]Health Sciences Library System, University of Pittsburgh

Oct 11, 2022

2 *Works for me* Share

dx.doi.org/10.17504/protocols.io.261genxjwg47/v1

Health Sciences Library System, University of Pittsburgh

Stephen W Gabrielson

ABSTRACT

This protocol provides step-by-step instructions for downloading publication data from Dimensions, cleaning the data to identify first and last authors, and uploading the data into the Namsor name analysis tool to help determine the gender of authors. In these directions, publications from the journal Critical Care Medicine are used as an example.

The data cleaning instructions are based on the use of Microsoft Excel for Microsoft 365 MSO (Version 2202 Build 16.0.14931.20704) 64-bit on a Windows PC.

DOI

dx.doi.org/10.17504/protocols.io.261genxjwg47/v1

PROTOCOL CITATION

Stephen W Gabrielson, Rose L Turner 2022. Publication Data Cleaning in Excel for Author Name Gender Analysis . **protocols.io** https://dx.doi.org/10.17504/protocols.io.261genxjwg47/v1

KEYWORDS

data cleaning, gender name analysis, namsor, dimensions, excel

LICENSE

CREATED

May 04, 2022

LAST MODIFIED

Oct 11, 2022

PROTOCOL INTEGER ID

61978

Getting Publication Data

1

This section shows how to export publication data out of Dimensions as a CSV file. Dimensions is used as the data source since its article records contain more author full names that go farther back than PubMed or Web of Science.

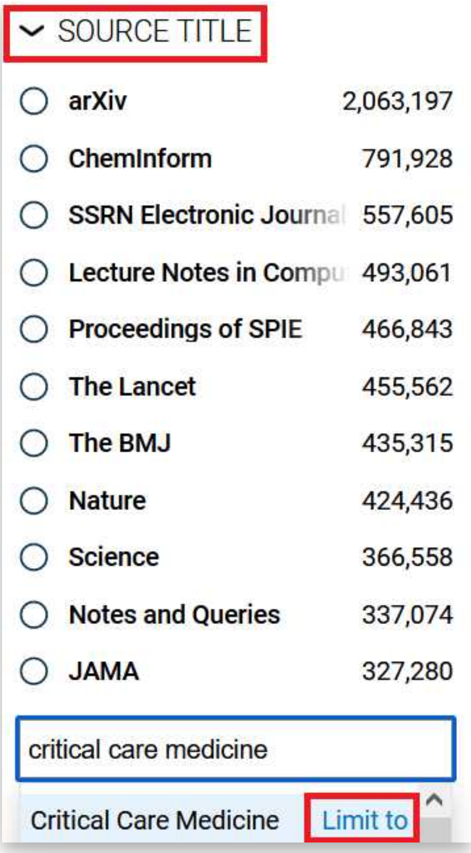
NOTE: Dimensions includes several different publication types. Refer to their page on Which publication types are available in Dimensions? to find out what's included and how they define a publication type. For example, their article filter will retrieve news and editorial content from both scientific journals and trade magazines, alongside research articles.

From the Dimensions homepage, click on **Access Free Web App**.

2 A personal account is needed to export articles from Dimensions. **Log in** or click **Register** in the upper right to create an account.

3 To limit the results to a certain journal on the Free Web App landing page, expand the **Source Title** filter in the left column. Click **More** at the bottom, type in the full name of the journal in the search box, and click **Limit to**. Title abbreviations should not be used (for example, use "Critical Care Medicine" and not "Crit Care Med").

Citation: Stephen W Gabrielson, Rose L Turner Publication Data Cleaning in Excel for Author Name Gender Analysis https://dx.doi.org/10.17504/protocols.io.261genxjwg47/v1



SOURCE TITLE

○ arXiv	2,063,197
○ ChemInform	791,928
○ SSRN Electronic Journal	557,605
○ Lecture Notes in Compu	493,061
○ Proceedings of SPIE	466,843
○ The Lancet	455,562
○ The BMJ	435,315
○ Nature	424,436
○ Science	366,558
○ Notes and Queries	337,074
○ JAMA	327,280

critical care medicine

Critical Care Medicine Limit to

4 The export limit in Dimensions varies depending on your subscription. If you are using the free web app, the export limit is 2500 records at a time. Considering this, you may have to export results in smaller sections if you are working with many results. The following step demonstrates how to do this by **Publication Year**.

4.1 From the results page, use the **Publication Year** filter on the left and select a year. Click **Limit to** at the bottom of the filters column.

5 To export records, click **Save / Export** to the right of the search box. Then choose **Export results**.

5.1 Use the export option **for bibliometric mapping** and click **Export**. This option allows you to export up to 2500 results in the free web app. If you need to export more records at once, try using other filters to reduce results. Otherwise, the subscription-based Dimensions allows for exports up to 50,000.

6 To download the CSV, click on your **account profile name** in the upper right corner and select **Export center**. Depending on the file size, it may take a few minutes for it to appear in the Export center.

Citation: Stephen W Gabrielson, Rose L Turner Publication Data Cleaning in Excel for Author Name Gender Analysis https://dx.doi.org/10.17504/protocols.io.261genxjwg47/v1

7 Find your export request and click **Download** to the right. This will save a ZIP file to your computer that contains a CSV file with the publication data.

NOTE: If you need to export several files and would like to combine them into one file for data cleaning and uploading into Namsor, make sure to delete the first two rows in the subsequent export files before moving the data into your main CSV file.

Preparing for Data Cleaning

8

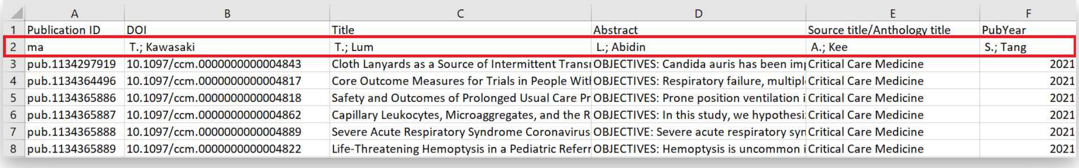
The following sections show how to clean the data before uploading the file into a gender name analysis tool.

Open the CSV file in Excel and delete the first row that starts with "about this data". **Save the original spreadsheet** with a descriptive name (e.g., dimensions_original_year) and refer to later as needed.

NOTE: the spreadsheets you save can be in CSV or Excel file formats.

9 Review your file for data that appears erroneous. Due to an Excel cell character limit of 32,767 characters, data could potentially be reformatted and displayed incorrectly. This might happen in instances when a publication has a very long author list in a single cell.

For example, row 2 in the screenshot below does not contain the correct values for Publication ID, DOI, Title, etc.



	A	B	C	D	E	F
1	Publication ID	DOI	Title	Abstract	Source title/Anthology title	PubYear
2	ma	T.; Kawasaki	T.; Lum	L.; Abidin	A.; Kee	S.; Tang
3	pub.1134297919	10.1097/ccm.0000000000004843	Cloth Lanyards as a Source of Intermittent Transı	OBJECTIVES: Candida auris has been im	Critical Care Medicine	2021
4	pub.1134364496	10.1097/ccm.0000000000004817	Core Outcome Measures for Trials in People Witl	OBJECTIVES: Respiratory failure, multipl	Critical Care Medicine	2021
5	pub.1134365886	10.1097/ccm.0000000000004818	Safety and Outcomes of Prolonged Usual Care Pr	OBJECTIVES: Prone position ventilation i	Critical Care Medicine	2021
6	pub.1134365887	10.1097/ccm.0000000000004862	Capillary Leukocytes, Microaggregates, and the R	OBJECTIVES: In this study, we hypothesi	Critical Care Medicine	2021
7	pub.1134365888	10.1097/ccm.0000000000004889	Severe Acute Respiratory Syndrome Coronavirus	OBJECTIVE: Severe acute respiratory syn	Critical Care Medicine	2021
8	pub.1134365889	10.1097/ccm.0000000000004822	Life-Threatening Hemoptysis in a Pediatric Referr	OBJECTIVES: Hemoptysis is uncommon i	Critical Care Medicine	2021

If this occurs, you can either delete these rows, understanding that this data loss is due to a limitation of using Excel, or use a different program that does not have Excel's cell character limits, such as a text editor or Google Sheets.

10 Copy and paste the Authors column **into a new spreadsheet** and give it a descriptive name (e.g., authors_year).

Citation: Stephen W Gabrielson, Rose L Turner Publication Data Cleaning in Excel for Author Name Gender Analysis https://dx.doi.org/10.17504/protocols.io.261genxjwg47/v1

Removing Blank Rows

11 In the newly created spreadsheet with only author data, remove the blank rows, as these were from article records that did not list any authors. The following steps will demonstrate how to remove these blank rows.

11.1 Select the **Authors** column.

11.2 Go to the **Data** tab and select **Text to Columns** in the **Data Tools** section.

11.3 Select **Delimited** and click **Finish**.

> **NOTE:**
> 1. You may not see anything happen after clicking **Finish** and that is normal.
> 2. If you get a message that asks you to overwrite any data, cancel and reopen your file to try again.

11.4 Select the **Authors** column again.

11.5 Go to the **Home** tab and select **Find & Select** in the **Editing** section.

11.6 Select **Go To Special...**

11.7 Select **Blanks** and click **OK**. All blank cells from the Authors column should be highlighted.

11.8 Go to the **Home** tab. In the **Cells section**, click on the **Delete drop down menu**, select **Delete Cells**, then select **Entire row**.

Citation: Stephen W Gabrielson, Rose L Turner Publication Data Cleaning in Excel for Author Name Gender Analysis https://dx.doi.org/10.17504/protocols.io.261genxjwg47/v1

Separating Authors

12 If a publication has multiple authors, they will be listed in a single cell, separated by semicolons. Separate the authors so that they are each in their own column.

12.1 Select the **Authors** column.

12.2 Go to the **Data** tab and select **Text to Columns** in the **Data Tools** section.

12.3 Select **Delimited** and then click **Next**.

12.4 Uncheck the box for **Tab** and select **Semicolon** instead. Click **Next**.

12.5 Leave **General** selected on the next screen and click **Finish**.

Creating First and Last Authors Column

13 Rename the first author column to **First Authors**. These are the first authors.

14 Create an **empty column** called **Last Authors** at the end of your spreadsheet. Depending on your data, you may have to scroll very far to the right in order to locate the first empty column in your spreadsheet. It should be adjacent to a column that has an author.

15 The following Excel formula identifies the last authors. It works by finding the last non-blank cell in the row and puts that value into the **Last Authors** column. When the second column of your spreadsheet is blank, that indicates that the article only has one author, so the formula will not put anything into the Last Authors column.

Copy and paste this formula into row 2 of the Last Authors column, inserting the relevant cell

Citation: Stephen W Gabrielson, Rose L Turner Publication Data Cleaning in Excel for Author Name Gender Analysis https://dx.doi.org/10.17504/protocols.io.261genxjwg47/v1

location for your spreadsheet (the column to the left of your empty Last Authors column). The B2 and A2 cell locations should be the same for your data.

=IF(B2<>"",LOOKUP(2,1/(A2:*YOURCOLUMNHERE*2<>""),A2:*YOURCOLUMNHERE*2),"")

For example, if your Last Authors column is BU, you would edit the YOURCOLUMNHERE placeholder to BT, so that the formula analyzes all available authors in that row. That would look like: **=IF(B2<>"",LOOKUP(2,1/(A2:BT2<>""),A2:BT2),"")**

> **NOTE:** If you are copying the formula from this protocol, paste it into a text editor first and remove any leading or trailing whitespace beforehand.

16 To copy the formula into the remaining rows, use the **Fill** command.

16.1 Select the cell with the formula and the remaining cells in the Last Authors column.

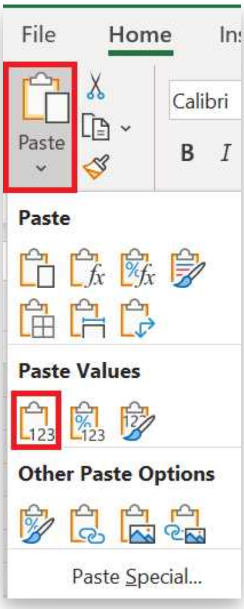
16.2 Go to the **Home** tab and click **Fill** in the **Editing** section. Select the **Down** option.

17 The Last Authors column should now list all the last authors. Before deleting the middle authors, **replace the formula with its values** in the Last Authors column, so that formula's output doesn't change.

17.1 Select the **Last Authors column**.

17.2 Copy the cells: **Control + C**

17.3 Go to the **Home** tab. From the **Paste menu** select **Paste values**.



17.4 Now you can delete the middle author columns without affecting the formula's output.

18 There should now be two columns.

18.1 Copy and paste the **First Authors** column, including the header row, into its own spreadsheet and save it as **first_authors_year**.

18.2 Copy and paste the **Last Authors** column, including the header row, into its own spreadsheet and save it as **last_authors_year**.

Cleaning Author Names

19

The following steps explain how to separate first and last names, as well as middle initials. Follow the steps for both the **first_authors_year** and **last_authors_year** spreadsheets created in the previous section.

20 Separate first and last names.

20.1 Select the column containing author names.

20.2 Go to the **Data** tab and select **Text to Columns** in the **Data Tools** section.

20.3 Select **Delimited** then click **Next**.

20.4 Change the default settings and **select Comma**. Click **Next**.

20.5 Leave **General** selected on the next screen and click **Finish**.

20.6 Rename columns to **Last Name** and **First Name**.

21 You may wish to remove middle initials from names, as this may decrease the name-gender probability match.

NOTE: However, the following directions will also affect first names that have a space, such as "Mary Lou" or "Jun Yeun". Leaving in these full first names increases name-gender probability matches.

21.1 Select the **First Name** column on the right.

21.2 Go to the **Data tab** and select **Text to Columns** in the **Data Tools** section.

Citation: Stephen W Gabrielson, Rose L Turner Publication Data Cleaning in Excel for Author Name Gender Analysis https://dx.doi.org/10.17504/protocols.io.261genxjwg47/v1

21.3 Select **Delimited** then click **Next**.

21.4 Change the default settings and **select Space**. Click **Next**.

21.5 Leave **General** selected on the next screen and click **Finish**.

This removes leading whitespace and separates out the middle initials into their own column. Remove the empty whitespace column and rename the First Name and Last Name columns if necessary.

> **NOTE:** before removing the middle initials column, verify that no first names are included. This may happen depending on how the author's name was formatted in the article record (i.e., M. Elizabeth). Move these names back to the first name column as needed.

22 Review the names for any special characters or punctuation marks that Excel added when data was exported from Dimensions, such as, but not limited to: **§, Ã**, **©**, and **?**. These should be removed and the name spelling should be corrected. Search for the name in your original Dimensions spreadsheet to find the article that the name appears in. Then look up the article in Dimensions to find the correct spelling.

Use the **Replace** option in the **Find & Select dropdown** menu to make corrections based on the special characters and incorrect punctuation marks that you identified.

23 Remove any blank rows from your **last_authors_year** file. Publications that did not have a last author account for these blank rows. If there are any blank rows in your file, Namsor will include them when processing the credit charge.

Refer to **Step 11** of the **Removing Blank Rows** section for instructions.

Uploading Data to Namsor

24

> Namsor is a tool that analyzes names to determine gender, origin, and ethnicity. While the service is free for up to 5,000 rows of data per month, there are other pricing plans depending on how much data you need to analyze.
>
> The following directions describe how to upload your CSV or Excel file into Namsor and

Citation: Stephen W Gabrielson, Rose L Turner Publication Data Cleaning in Excel for Author Name Gender Analysis https://dx.doi.org/10.17504/protocols.io.261genxjwg47/v1

download the results.

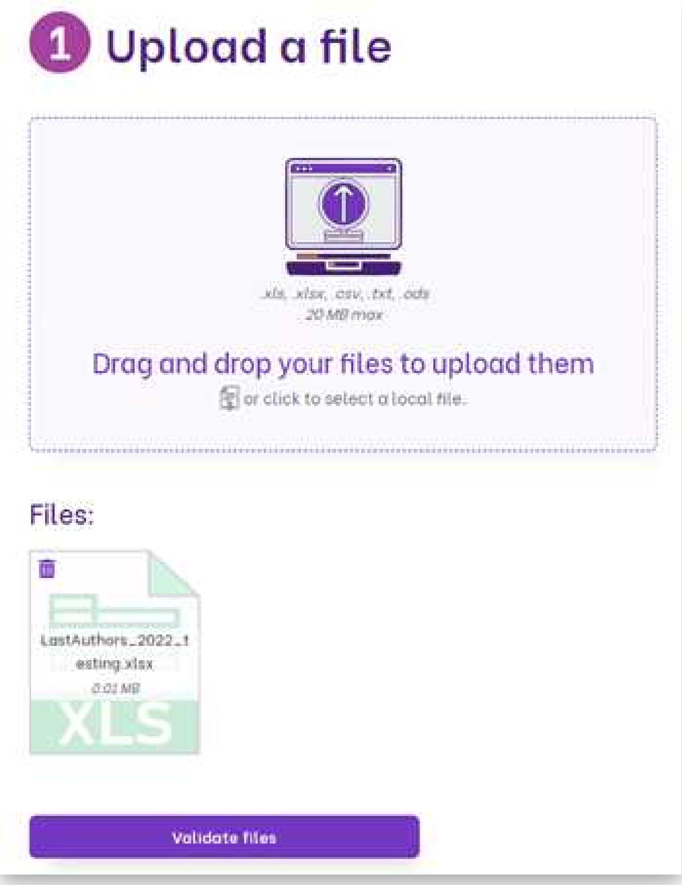
Register for a free account or log in to an existing one. Then click **Excel/CSV** at the top of Namsor's homepage to upload your file. Either drag and drop your file or select a local file on your computer.

NOTE: Review these directions from Namsor regarding file type and size limits:

"Documents must be either Excel (.xls, .xlsx file extension), CSV (.csv file extension), text (.txt file extension) or OpenDocument Spreadsheet (.ods file extension) files.

If your computer has limited power (processor, RAM), favor small or medium-sized files (less than 3MB) even if it means processing several times. This is to avoid saturating your computer's RAM and crashing the browser."

25 Once your file is uploaded, click **Validate files**.

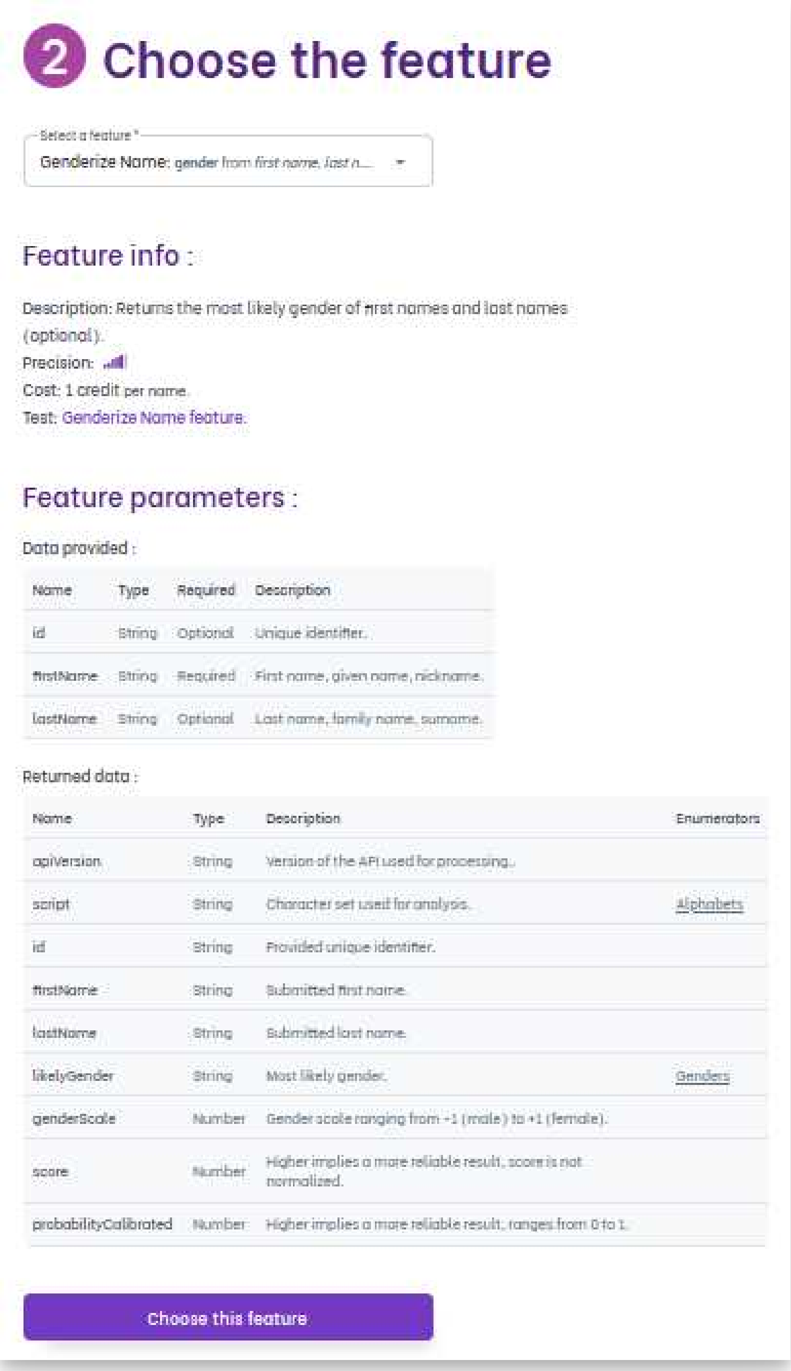

1 Upload a file
.xls, .xlsx, .csv, .txt, .ods
20 MB max
Drag and drop your files to upload them
or click to select a local file.
Files:
LastAuthors_2022_testing.xlsx
0.01 MB
XLS
Validate files

26 Choose the feature. Click the **select a feature** dropdown menu and select **Genderize Name:**

gender from first name, last name (optional). Once you've selected a feature you can view the feature info and feature parameters.

The cost for this analysis is 1 credit per name.

27 Click **Choose this feature** below the Feature info and Feature parameters sections to continue.



Choose the feature

Select a feature *
Genderize Name: gender from *first name, last n...*

Feature info :

Description: Returns the most likely gender of first names and last names (optional).
Precision: ▂▄▆
Cost: 1 credit per name.
Test: Genderize Name feature.

Feature parameters :

Data provided :

Name	Type	Required	Description
id	String	Optional	Unique identifier.
firstName	String	Required	First name, given name, nickname.
lastName	String	Optional	Last name, family name, surname.

Returned data :

Name	Type	Description	Enumerators
apiVersion	String	Version of the API used for processing..	
script	String	Character set used for analysis.	Alphabets
id	String	Provided unique identifier.	
firstName	String	Submitted first name.	
lastName	String	Submitted last name.	
likelyGender	String	Most likely gender.	Genders
genderScale	Number	Gender scale ranging from -1 (male) to +1 (female).	
score	Number	Higher implies a more reliable result, score is not normalized.	
probabilityCalibrated	Number	Higher implies a more reliable result, ranges from 0 to 1.	

28 Choose your settings in these next steps.

28.1 Select **Keep existing columns** under **Global settings**.

28.2 Select **The file has a header** and confirm that **First row = 2**.

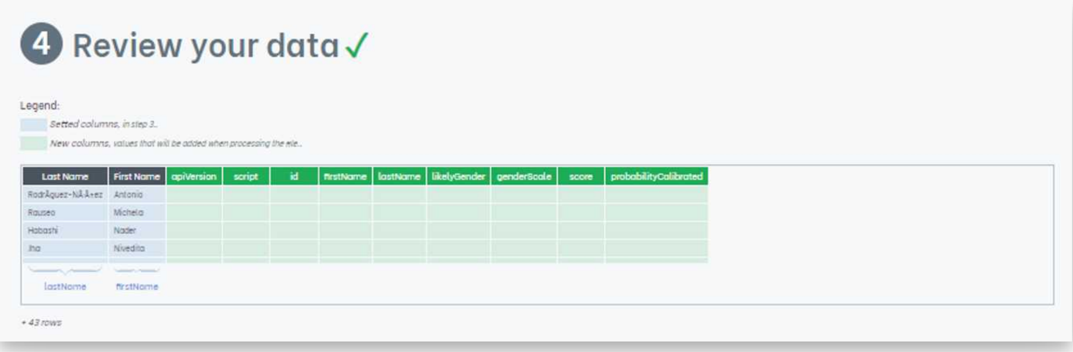
28.3 Identify the **first name column** and **last name column** under **Column settings**. This is based upon the data in your spreadsheet.

28.4 Click **Validate settings**.

29 Review your data before processing your file.

29.1 Check that all the settings made in the previous steps are correct.

29.2 Review the summary chart and click **The data looks good**.



4 Review your data ✓

Legend:

- 🟦 *Setted columns, in step 3..*
- 🟩 *New columns, values that will be added when processing the file..*

Last Name	First Name	apiVersion	script	id	firstName	lastName	likelyGender	genderScale	score	probabilityCalibrated
RodrÃguez-NÃºÃ±ez	Antonio									
Rauseo	Michela									
Habashi	Nader									
Jha	Nivedita									
lastName	firstName									

+ 43 rows

30 Review the summary and credit cost on the next screen.

Check your credit balance.

Citation: Stephen W Gabrielson, Rose L Turner Publication Data Cleaning in Excel for Author Name Gender Analysis https://dx.doi.org/10.17504/protocols.io.261genxjwg47/v1

30.1 **NOTE** from Namsor:

"View the credit cost of processing your files and your balance. In the event that you do not have enough credits, you will be offered to acquire additional credits or process the files to the extent of your remaining credits. Our tool features smart processing and will not charge for analyzing identical data for up to 20 times."

30.2 Click **Process file**.

31 Download the generated file. Look for the column **likelyGender** to review the name-gender matches. To learn more about the data in the remaining columns, such as genderScale, score, and probabilityCalibrated, see Namsor's API Documentation.

Citation: Stephen W Gabrielson, Rose L Turner Publication Data Cleaning in Excel for Author Name Gender Analysis https://dx.doi.org/10.17504/protocols.io.261genxjwg47/v1