Code

Data Visualization - The Path to Becoming an Astronaut

Lucas Whitmire

21 July 2020

1 Introduction

The goal of this analysis is to develop an effective graphic which communicates the best (most common) path to becoming a NASA astronaut. The target audience for this graphic is the general public (especially aspiring Astronauts!). Data used to develop this visualization was found using Google's Datasetsearch and was originally published by the National Aeronautics and Space Administration (NASA) under the CC0 (Creative Commons 0) license which dedicates the dataset to the public domain with no copyright.

2 Methods

2.1 Visualization Selection

In order to visualize the 'path' to becoming an Astronaut, multiple categories and their relative sizes need to be displayed for several variables. One option to accomplish this could be the use of small multiples of barplots arranged chronologically. This would allow the viewer to see the most common categories for each variable. The disadvantage of this plot is that it does not show relationships between each of the variables. For example, if everyone who joined the military went on to study Physics/Maths, this information would be lost by using the small multiples plot. The sankey plot allows for this information to be captured in the 'links' between the categories of subsequent variables. It also displays all of information that can be represented with the small multiples of barplots. For this reason, and the ability to highlight a literal 'path' through the variables, the sankey plot was selected to visualize this data.

2.2 Data Exploration and Preparation

The dataset contains background information on 357 Astronauts from 1959 through 2013. Below is a subset of the dataset with just the relevant columns selected. The variables of interest in determining the path are `Gender`, `Undergraduate.Major`, `Graduate.Major`, `Birth.Place`, and `Military.Branch`. `Gender` has two levels and no missing values so no further data preparation will be required for this variable. The other variables are addressed in the subsequent sections.

Code

Table 2.1: Subset of Astronaut Dataset

Name	Birth.Place	Gender	Undergraduate.Major	Graduate.Major	Military.Branch
Joseph M. Acaba	Inglewood, CA	Male	Geology	Geology	NA
Loren W. Acton	Lewiston, MT	Male	Engineering Physics	Solar Physics	NA
James C. Adamson	Warsaw, NY	Male	Engineering	Aerospace Engineering	US Army (Retired)
Thomas D. Akers	St. Louis, MO	Male	Applied Mathematics	Applied Mathematics	US Air Force (Retired)
Buzz Aldrin	Montclair, NJ	Male	Mechanical Engineering	Astronautics	US Air Force (Retired)
Andrew M. Allen	Philadelphia, PA	Male	Mechanical Engineering	Business Administration	US Marine Corps (Retired)
Joseph P. Allen	Crawsfordsville, IN	Male	Mathematics & Physics	Physics	NA
Scott D. Altman	Lincoln, IL	Male	Aeronautical & Astronautical Engineering	Aeronautical Engineering	US Navy (Retired)

2.2.1 *Birth Place*

The `Birth.Place` variable has 272 unique values which is far more than can be displayed on the sankey plot. The data is binned into the categories `USBorn` and `BornAbroad`. All of the United States birthplaces have a 2 character state code given. This was used to bin the birthplaces into the two categories.

Code

2.2.2 *Undergraduate/Graduate Major*

The `Undergraduate.Major` variable contains some missing data. A further look into this missing data (using Wikipedia) revealed that the majority of missing Bachelor's Degrees were for Astronauts that had attended Military Academies. These Academies did not provide named Bachelor's degrees in the earlier years which is the reason the fields were blank. To distinguish these missing values from 'real' missing values, they were populated with the degree "Military Bachelors". Interestingly, there is 1 Astronaut with no official Bachelor's Degree. Thomas J. Hennen only attended 2 years of University before joining the military and never earned a degree.

Both `Undergraduate.Major` and `Graduate.Major` each have far too many unique values to be displayed in a meaningful way. The top 10 most frequent degrees for both Undergrad and Graduate levels are shown below in Tables 2.2 and 2.3 . These more common majors were used to determine the new categories. First, degrees containing the terms "Physics" or "Math" were binned into one

category `Physics/Maths`. Because Aerospace and Aeronautical Engineering were such a large portion of both the Undergraduate and Graduate degrees, they were kept separate from the other engineering degrees. Aerospace is also an broad field which encompasses Aeronautics (within atmosphere) and Astronautics (in space) so degrees containing these words were binned together with Aerospace. The new `Sciences` category is the broadest as it contains all degrees which contained the word “science” or one of the common suffixes for a field of study (i.e. ‘-ology’,‘-onomy’,‘-graphy’, etc.). The `Engineering` bin contains all degrees with ‘engineer’ in the name with the exception of those Aerospace/Aeronautical Engineering degrees which were previously binned into the `Aerospace` category. Degrees that did not fall into one of these categories were classified as `Other`. Because of the high number of Medical degrees in the Graduate Majors, it was given its own category. All other groupings are the same as the Undergraduate Majors.

Code

Table 2.2: Most Frequent Undergraduate Majors

Undergraduate.Major	Count
Physics	35
Aerospace Engineering	33
Mechanical Engineering	30
Aeronautical Engineering	28
Electrical Engineering	23
Military Bachelors	16
Engineering Science	13
Engineering	12
Mathematics	11
Chemistry	10

Code

Table 2.3: Most Frequent Graduate Majors

Graduate.Major	Count
Aeronautical Engineering	27
Aerospace Engineering	21
Medicine	17
Physics	15
Mechanical Engineering	13
Electrical Engineering	8
Aeronautics & Astronautics	7
Astronomy	6
Aviation Systems	6
Astronautics	5

Code

2.2.3 *Military Branch*

The unique values for the `Military.Branch` variable can be seen in Table 2.4 below. Since it is not important whether or not the Astronaut is “retired” or in the Reserves, this part of the string was removed. An initial attempt was made to generate the sankey plot using all the branches of the military

as categories, but this yielded a plot that would have confused the viewer. This plot made it appear that the majority of Astronauts had not joined the military, which was not the case. An additional attempt which added another layer after `Military` with a breakout of each of the military branches also made the sankey plot much harder to understand and was abandoned. In determining the 'path' it was determined that the question to be answered for this variable would be "Should I join the military or not?" rather than determining which branch to join. For this reason, the variable was binned into `Military` and `Civilian`.

Code

Table 2.4: Unique Labels in Military.Branch

Military.Branch
US Air Force
US Air Force (Retired)
US Air Force Reserves
US Air Force Reserves (Retired)
US Army
US Army (Retired)
US Coast Guard (Retired)
US Marine Corps
US Marine Corps (Retired)
US Marine Corps Reserves
US Naval Reserves
US Naval Reserves (Retired)
US Navy
US Navy (Retired)

Code

2.2.4 *Preparation for Plotting*

The `sankeyNetwork` function in the `networkD3` package requires two dataframes to generate the plot: one containing nodes, and one containing the links between those nodes and the values of the links. The "nodes" data frame was generated by just concatenating the `levels()` output for each of the desired variables. The "links" dataframe was more difficult to create as the combinations of each of the categories in adjacent variables (in the Sankey) had to be created and then the value of the link determined by grouping the original data for each of the categories to get the number of rows in the result. For an added level of complexity, the "links" dataframe uses indices of the nodes from the "nodes" dataframe rather than the names of the categories. Because R uses "1-indexing" and D3 uses "zero-indexing", the indices of the "links" dataframe are "zero-indexed", while the "nodes" dataframe is "1-indexed". Subsets of the "links" and "nodes" dataframes are given below.

Code

Table 2.5: 'Nodes' Dataframe

nodes	**index**
BornAbroad	1
USBorn	2

nodes	index
Female	3
Male	4
Civilian	5

Code

Table 2.6: 'Links' Dataframe

source	target	values
0	2	2
0	3	20
1	2	48
1	3	287
2	4	37

2.3 Sankey Plot Visualization Aspects

2.3.1 *Colors*

The most important color selection for this plot was for the selected 'path'. In order to highlight the selected path, red was chosen and a light grey was used for all other paths. The colors for the Male and Female categories were specified to be consistent with the colors used in the Group Report for this project. Remaining colors were pulled from the `RColorBrewer` palettes 'Dark2' and 'Set3'.

2.3.2 *Order of Variables*

The variables were laid out chronologically from left to right. The order of the first two variables, `Gender` and `Birth.Place`, was arbitrary as these events would occur simultaneously. Military service and undergraduate major could also have been placed in either order as these typically would happen at the same time in the US. The final order was chosen to keep the Undergraduate and Graduate degrees together.

2.3.3 *Placement and Size of Nodes*

The height of the node rectangles show the relative sizes of the categories for each variable. This is an attribute of the sankey plot and is one of the reasons this type of plot was selected. Node widths is an attribute that can be set globally for all nodes, and the width was selected mostly based on aesthetics. It was noted, however, that thinner nodes allow for more of the plot to show the links, which are the more important information in this case.

The `sankeyNetwork` function does not provide much control over the placement of the categories for each variable. There is an `iterations` argument to the function that is meant to set the number of calculations of the diagram layout. The results even at high numbers of iterations were not particularly impressive. The package provides the ability to move the nodes in the y-axis manually after the plot is generated. Two improvements were made in this way. The first was to align the matching degrees in the Undergraduate and Graduate degree variables. This 'straightens' the selected path, and also decreases some of the congestion in the plot by reducing the overlapping of the wider paths. Second, the algorithm does not properly draw the link between the `USBorn` and `Military` categories so this is remedied by moving both the `Civilian` and `Military` categories up slightly. Note that the

`sankeyNetwork` command also does not provide a way to specify the horizontal distance between nodes. Its possible this can be accomplished utilizing javascript, but it would require manually editting the positions of the nodes and their associated links in the HTML. This was not performed for this graphic.

2.3.4 *Further Improvement Steps*

One strange aspect of the `sankeyNetwork` function is that it draws connections between nodes even if the 'value' of that link is 0. To remove these connections, as well as remove other relatively small connections, links with values below 4 were removed from the dataframe before plotting. It was recognized that removing some of these smaller links could lead to misrepresentations of the data. For instance, in the final sankey plot (Figure 3.1), it appears that there are no female astronauts who were born abroad when in fact there were two. Ultimately, the clarity of the sankey plot was determined to be more important, and the filter for small link values was retained for the final plot.
The graphic could have benefited from headings on each of the variables, but this proved too difficult with my limited HTML/javascript knowledge and will be left as a future improvement to the plot.

Code

3 Results

The final sankey plot showing the path to becoming an Astronaut is shown in Figure 3.1 below. It is clear from the graphic that the most common path to becoming Astronaut is to:
1. Be born in the USA
2. Be born Male
3. Join the Military
4. Study Aerospace for the Bachelor's Degree
5. Continue studying Aerospace for the Master's Degree

Code

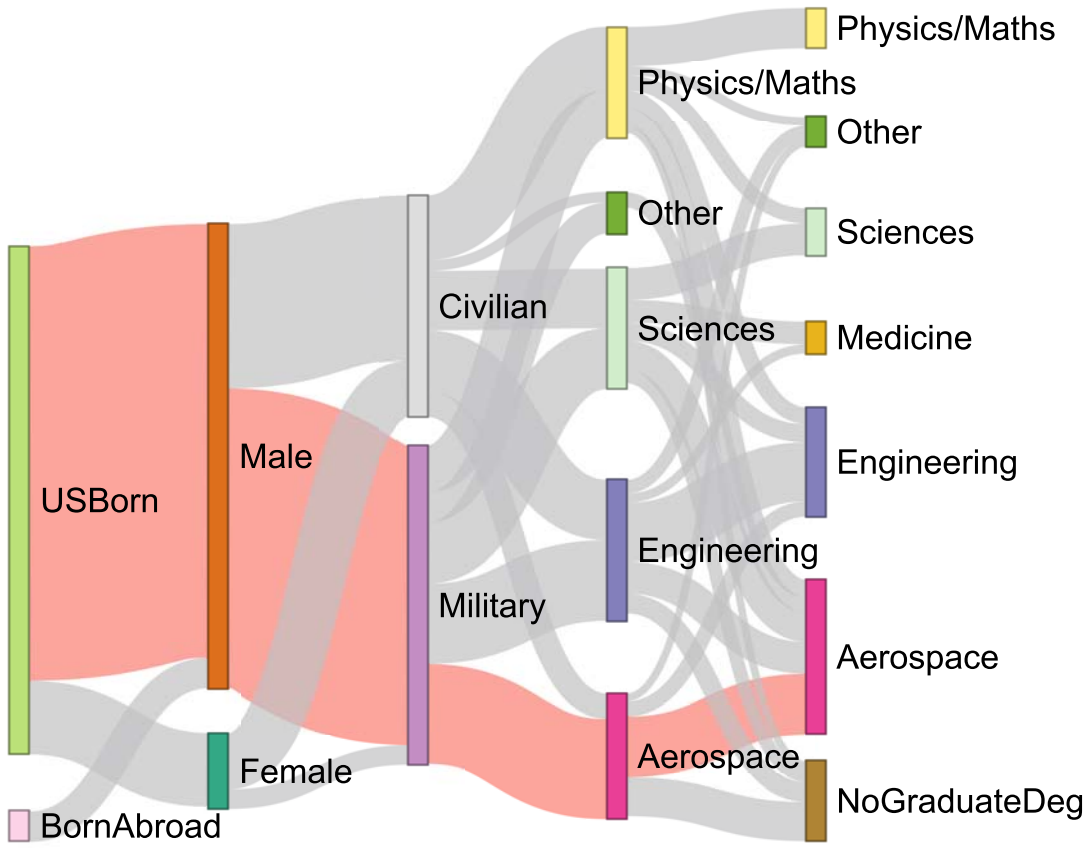

Figure 3.1: Final Sankey Plot

4 Conclusions

The Sankey Plot is an effective method of presenting this type of 'path' data. The plot clearly communicates the most common path to becoming an Astronaut. Although the plot is relatively simple upon first glance, the interactive features allow for a lot of detailed information to be conveyed if the viewer desires to drill into the data.

There are some disadvantages to this implementation of the Sankey Plot. First, extensive data preprocessing is required to achieve a format that the package can use to generate the plot. This makes iterations and modifications to the plot quite difficult. Second, the package doesn't provide much in the way of customization options by default. Specifying the color or making even simple changes to the way the text is displayed require some knowledge of D3.js and an import of the `htmlwidgets` package. Lastly, the algorithm which the package uses to determine the location of nodes was not particularly effective. In the interest of clarity, I would expect the algorithm to try and orient wide 'paths' horizontally by placing their respective nodes at the same elevation in the plot. This would reduce large overlaps in the plot. That being said, the package does allow for manual 'dragging' of the nodes to place them where desired after generating the plot. This feature is a necessity in the absence of options to natively set the location of nodes.

References

- Wikipedia contributors. (2020, April 16). Thomas J. Hennen. In Wikipedia, The Free Encyclopedia. Retrieved 06:55, July 19, 2020, from https://en.wikipedia.org/w/index.php?title=Thomas_J._Hennen&oldid=951313024 (https://en.wikipedia.org/w/index.php?title=Thomas_J._Hennen&oldid=951313024)
- NASA Astronauts Dataset, 1959-Present. https://www.kaggle.com/nasa/astronaut-yearbook (https://www.kaggle.com/nasa/astronaut-yearbook)