Assignment Report

# Top literary award nominees over the years in science fiction and fantasy


## Abstract

*Science fiction and fantasy have been one of the most popular genres in the bookstore. There are multiple awards that have been found to honour prestigious authors in this field. To be nominated and win an award is such a milestone in the life of an author. Now is a good chance to look at the nominations and winning titles to give credits to how well the authors perform over the years. The visualisation highlighted many excellent authors such as Ray Bradbury, John. W. Campbell Jr., Robert Silverberg, Ursula K. Le Guin, Stephen King and Gardner Dozois. Furthermore, most of the nominations were from the Locus Award, and overall science fiction awards have slightly more popularity over fantasy and horror. Moreover, in no surprise, in the description of those nominated titles, some typical words that reflect the genres like science, magic, world, secret, death, war were prominent.*


## Data Collection

My data is provided in text format by The Internet Speculative Fiction Database (ISFDB) and Goodreads.com.

In my opinion, the three aspects of big data are present in my data in the following ways.

Firstly, it has velocity: the site provides downloads to their backup, which is updated weekly.

Moreover, the backup is stored in MySQL and contains 56 tables, presenting various information of bibliographic data of authors, award listings, user verification of submitted data, publications, titles, magazine content listings, anthology and collection content listings, etc. Along with the fact the I also used extra data by scraping from Goodreads.com to get descriptions from the titles that won an award in the ISFDB, this fulfils the variety aspect.

Also, the ISFDB dataset takes over 1.2GB of storage of text and numbers. From the data I collected, I mostly used the information from three tables: *authors*, *titles*, and *awards*. There are a total of 201171 rows, 1704904 rows and 58025 rows respectively in these three.

# Data Exploration, Processing, Cleaning and/or Integration

## Data Cleaning and Processing

*What did you need to do to prepare the dataset(s) to create your graph/chart?*

After collecting the data, I did some more filtering to get only the interested data.

### Author queries

I decided to only explore with the data of the authors who were born after 1900, which narrows down the *authors* table into 29280 rows as most of the awards data are from after 1940. The gathered information are *name, birthplace, date_of_birth, language, debut_year.*

### Bibliography queries

I made queries to get data about their entire bibliography in literacy work (excluding magazines and essays) using the *titles* table. The gathered information are *title, year, type* .

### Award queries

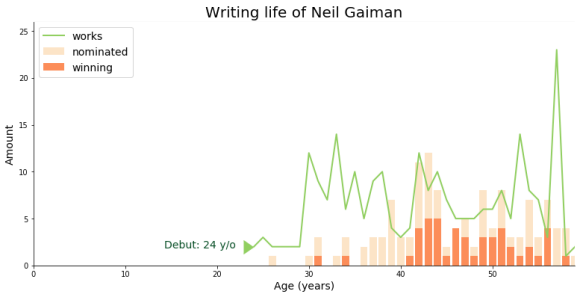
The authors were linked to the *awards* to get data about their award listings. The gathered information are: *year, title, award_name, result_place* (first place is the winner)*, is_pool* (if the award result is based on a pool) and *category* (novel, novelette, novella, short story, anthology)

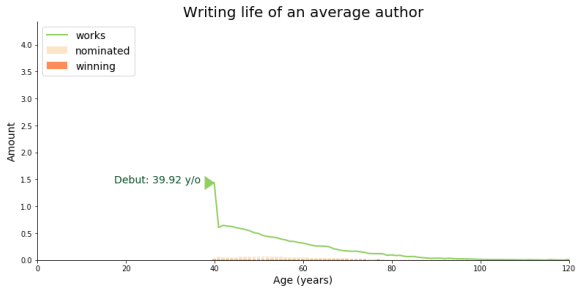
Calculations were made to have each author a profile of:

- NAME
- BIRTH_YEAR
- COUNTRY (needed to be cleaned through OpenRefine)
- DEBUT YEAR
- WORKS_YEAR_XXXX (number of books written in year XXXX)
- NOMINATIONS_YEAR_XXXX (number of awards won and nominated in year XXXX)
- AWARDS_YEAR_XXXX (number of awards won in year XXXX)

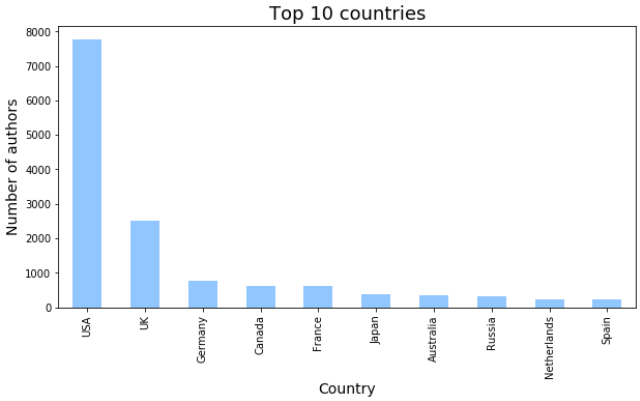
## Data Exploration

The charts below are not the final visualisations. I find it easier for me to get to know the data by visual cues; that is the reason why I included these graphs.



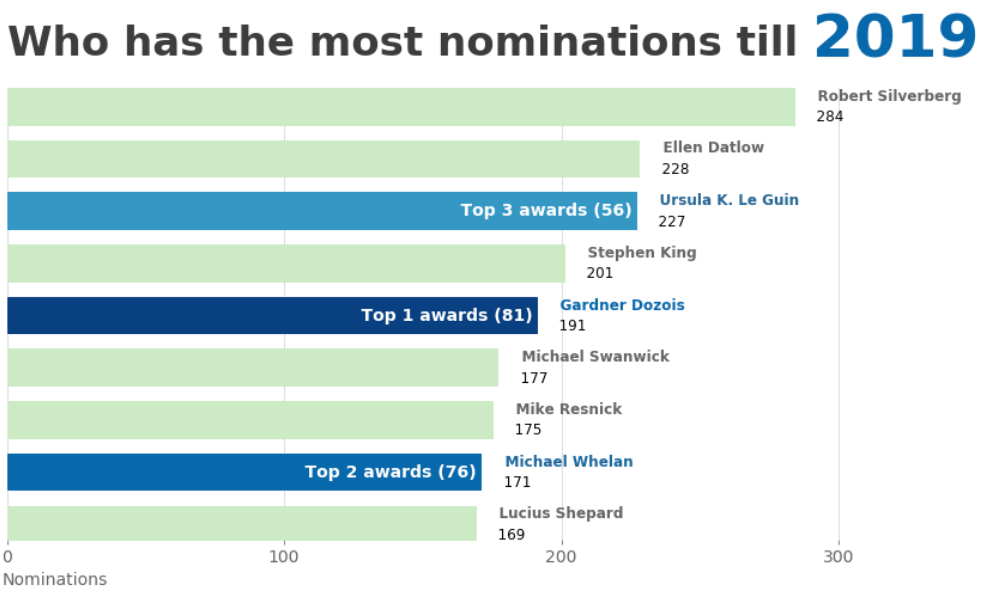

I first explore the data by drawing a sketch graph of the life of my favourite author. It was interesting to me because he won most of the awards he was nominated, especially during his late years. I expected to have something interesting going on when I calculate the mean of every author in the dataset and found the results disappointing. It turns out that the majority of the authors did not win any awards at all.


The life of the normal authors are not much
interesting to explore anymore, that was why I narrowed it down into nominated authors.
Choosing only awarded authors would leave the data too small.
I also wanted to view some relationship between countries and authors but the data was too skewed to the USA so I skipped this.
I was left with the options of analysing the data on the nominations over the years.

## Visualisation

The first one is an animated chart describing the ranking of nominations that authors get over the years.




## Choice of chart types

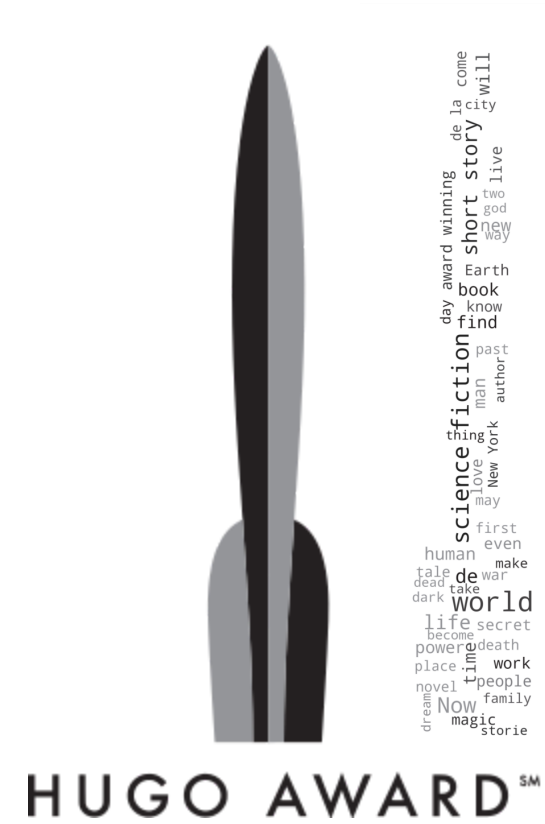
I want to make comparisons between different authors over the years. Authors is a nominal variable, top rankings is ordinal while the number of nominations, awards, books are numeric. Bar charts are simple yet effective for comparisons for these types of data. Horizontal bars allow nuance between the authors, which gives more impact on the difference of 1 nomination. The order of top rankings can be expressed by using colours.

## Design choices

I started from this tutorial: Bar Chart Race in Matplotlib and made the following changes:

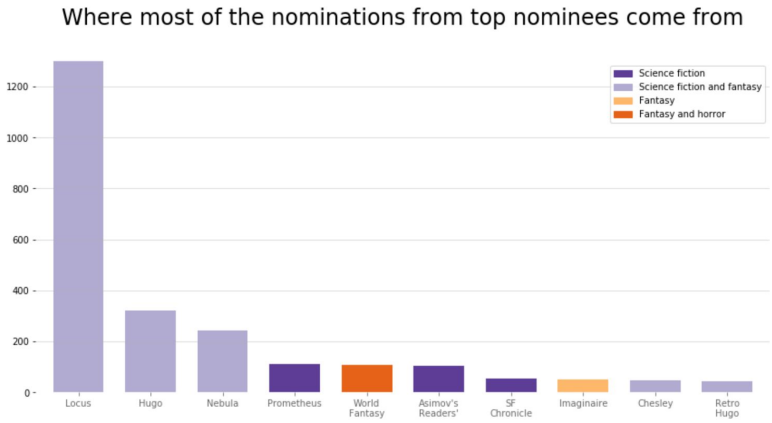
- I changed the rainbow colours into 2 main colours: dark blue and light green. The variety of colours does not convey any extra information in my dataset. Originally I chose the colour for each author according to their country; but as explained before, most of them are in the USA so the chart looked plain.
- I want to highlight the authors that won the most awards so their colours must be distinguishable. The dark colour against the light colour makes a good contrast. Top 1, top 2 and top 3 authors are also given different shades where darker means the higher rank.
- I moved the x-axis down to its normal position because it would create a cluster of words around the title area.
- The most works are put in smoothing the animation so that it would not look jumpy when a new author enter the chart or the positions change. I find the original one hard to follow when applied directly to my data because there are so many changes in the ranking and highlighting.

- I also highlighted the year in the title and moved the names outside of the bars for readability.




## Animation choice

I have the data over the years from 1900 to 2019. Although the awards started around the 1940s, I still have almost 80 years to deal with. To show all the information over the years, the animation is useful.




The second visualisation involving the awards that the top 10 authors have won till 2019. The awards are categorical and the nominations are numeric so bar chart is again simple yet effective for this case. The colours were chosen so that “Science fiction” (only) and “Fantasy and Horror” were two extreme of the scale and the others are a blend in between.

The third one is a word cloud of the description of all nominated titles that I collected from Goodreads.com. The shape was based on the shape of a rocket of the Hugo logo.

## Tools and libraries used

MySQL.connector was used so that I can make queries in Python codes. Python was used to do all the cleaning and processing except for the name cleaning by OpenRefine.

Python, *matplotlib* was used to make the animation and the bar chart. I also used the library *wordcloud* to make the word cloud.

## Conclusion

The biggest problem that I had with the data was in the cleaning and processing. There were many unverified and missing data. The result charts are simple.  I spent most of the time for the animation everything else was rushed through.

For the bar chart race, I wanted to add another colour to distinguish between the top rankings in books (e.g Top 3 titles) and in awards won (Top 3 awards) without it making the chart distracting for having too much information at the same time. One more thing that could be improved is adding an interactive filter for the type of titles that were written, nominated and won (short stories, novels, novelettes, novellas, etc.). One author who wrote mainly in short stories is more likely to have more titles than the one specializing in novels.

For the second chart, I wanted to choose another chart which is not a bar chart so that it is not repetitive to the first visualisation. However, although obtain the data involving processing and cleaning, the final data are so simple and I do not think there is a need to make it complicated and difficult to read. The point I wanted to make is that the Locus Award give most of the nominations to top authors.

The third chart suffered from being hard to read. The shape is narrow on the sides so the words are made to be smaller. Also the colors of the logo were not clearly shown in the word cloud. If I had another award's logo file which is of a simpler shape, the visualisation would be more effective.

## References


The Internet Speculative Fiction Database

Goodreads

For the bar chart race, I started from this and then did the interpolation between frames by myself
https://towardsdatascience.com/bar-chart-race-in-python-with-matplotlib-8e687a5c8a41

The word cloud was guided by this tutorial:
https://www.datacamp.com/community/tutorials/wordcloud-python