Project A17: KAGGLE-NOBEL

Nobel Prize Stats Analysis

Team: Teele Tani, Tuule Tani

Link to project repository (https://github.com/tee1e/Nobel-prize-stats)

Business understanding

The issue that we are interested in investigating is Nobel Prize data. Nobel Prizes are distributed to the most brilliant minds in peace, literature, science and economics. But it is widely known that the Nobel Prize has favoured old white western men for more than 100 years. Therefore, while the general goal of our project is to analyse and visualize Nobel Prize data in a meaningful and interesting way, we also more specifically want to look at the age, gender and country of origin of Nobel Prize laureates and try to find out whether there have been any significant shifts or changes over recent years.

Our project does not have a clear business goal as such, because we cannot think of any organization or group of people (other than us) that will gain any kind of real benefit from the project and its results. We do not expect our project to reveal some type of unexpected, marketable or profitable results. However, we do expect the project to be interesting and enyojable for us and, most importantly, we want the project to help us practice and improve the skills and knowledge gained during the Introduction to Data Science course.

The business success criteria of our project is more easily defined, although it also is of a subjective nature. The finished project has to demonstrate some of the skills and knowledge we have obtained from the lectures and practice sessions of the Introduction to Data Science course. The final presentation of our project has to be informative and engaging. The findings have to be stated clearly and all the visuals used in the final presentation have to be in support of the findings of the project.

Our project does not have any real issues, constraints or risks associated with our business goals. There is also no serious need for a cost-benefit analysis for the project. It would be tricky to measure the cost of the project and it is quite clear that the benefits of the project far outweigh any cost there might be (if indeed there is any). The team of our project is small and the data we use in our project is not only publicly available but also tailor-made for our project. Additionally, the data used in the project is relatively small.

The data-mining technical success criteria aligns with our business success criteria. The main data-mining deliverable is our project report and presentation.

Data understanding

The main data source (‘dataset 1’) of our project is a dataset from Kaggle (https://www.kaggle.com/bahramjannesarr/nobel-prize-from-1901-till-2020). We have also examined other Kaggle Nobel Prize datasets (‘dataset 2’ and ‘dataset 3’) which have different sets of features (https://www.kaggle.com/imdevskp/nobel-prize and https://www.kaggle.com/nobelfoundation/nobel-laureates) and will most likely combine at least two datasets for our project. Thus far, there have not been any difficulties or issues concerning data management and quality, and we do not anticipate to encounter any problems of that kind during our project.

The three datasets referred to above contain information about all the individuals who have been awarded the Nobel Prize since the year 1901. The information of these datasets is trustworthy because their shared data source is the official Nobel Prize API of the Nobel Prize organisation. Therefore, there is no doubt as to the suitability and sufficiency of the data for our data-mining goals. However, while dataset 1 contains information from 1901 to 2020, dataset 2 does not include data from 2020 and dataset 3 does not include data from 2017 to 2020. This is something that needs to be taken into account and to be addressed during the first stages of our project. Also relavant for our project is the fact that the three datasets each have a differing set of features. Dataset 1 contains 14 features, dataset 2 contains 52 features and dataset 3 contains 18 features. For our project, we want to create a unique combination of features from different datasets, possibly generate some new features of our own (based on features already available to us), do statistical analysis, and see whether we can find correlations or contradictions between certain sets of features.

The great thing about Nobel Prize data is that when it comes to the range of values of different features, the data here is quite straightforward and simple to understand. For example, the feature of ‘Prize Category’ has the expected values of ‘Physics’, ‘Chemistry’, ‘Physiology or Medicine’, ‘Literature’, ‘Peace’ and ‘Economic Sciences’. However, we do encounter some small issues even here. Namely, not all the datasets include the category of ‘Economic Sciences’, because it is not one of the original Nobel Prizes established by Alfred Nobel’s will. Another problem that we need to take into account and find an adequate solution to concerns the feature of ‘Country’, because the range of values here includes some countries that no longer exist. Despite these minor issues, data exploration shows that the quality of data is good and there are no missing or incorrect values that cannot be corrected.

Planning your project

The general plan for our project is as follows:

1) Prerequisites: determining the datasets and features we want to use for our project. This step requires a team meeting and it should not take more than a couple of hours to complete.

2) Preprocessing: dealing with missing values, duplicates and other similar issues. This is the most important part of the project and needs several days of work by both team members.

3) General analysis of data: doing statistical analysis with the aim of finding interesting insights. This is the part of the project that is probably the most fun, but also needs several days of exploration and work by both team members.

4) Finalizing the project: choosing the most important insights. This is the part of the project were we look at all the initial analysis and decide what we want to include in our project and what we want to discard. This step requires a team meeting and it should not take more than a couple of hours to complete.

5) Conclusion: writing and perfecting the project report and presentation. This step requires at least a couple of days work by both team members.