UBC Salaries: Exploratory Analysis of Gender

Jade Bouchard

Table of contents

Aim

This document explores The University of British Columbia (UBC) faculty salaries based on guessed gender.

Data

Salary data was sourced from (University of British Columbia 2020). To access individual financial reports, click on the yearly links under the header `Statement of Financial Information (SOFI)`.

Gender data was inferred using first names of staff members. In order to guess gender, I used baby name datasets Sharma (n.d.).

Methods

The Python programming language (Van Rossum and Drake 2009) was used to perform this analysis.

Data collection

As mentioned earlier, for UBC salary data, I used the PDF salary information that UBC releases every year (University of British Columbia 2020). The following steps were taken to collect the data.

- Use the `requests` package to accesses the UBC Financial Reports webpage.
- Extract links to annual salary PDFs by locating the "Statement of Financial Information (SOFI)" section on the webpage and parsing the links.
- If there are any links (and associated PDFs) for which we have not already collected salary data, extract the text from the PDF using the package `pypdf`.
- Open the stored salary data dictionary
- Add the new salary text data to the salary data dictionary

The code does not necissarily follow the order of steps described above.

An exerpt of the raw salary data is below.

```
Kolhatkar, Ashra  79,036   550  Kolhatkar, Varada  88,579   -  Kolind, Shannon H  108,817
```

Data cleaning

In this section, the following steps are taken to clean the salary data:

- Remove special charachters from text. For example, §.
- Removes unnecessary text content. For example, the "[Auditor's] Qualified Opinion".
- Uses regex to process the raw text data into a structured DataFrame with columns: `Name`, `Remuneration`, `Expenses`.
- Splits the `Name` column into first and last names.
- Converts salary values to a numeric data type.
- Shortens first and last name to allow for easier name matching between years. For example, someone's name in 2020 could be "Bob M Sherbert" and in 2021 their name could be "Bob-M Sherbert". This name would be shortened to Bob Sherbert to avoid mismatching.
- Concatenate dataframes from all years together.

Table 1 shows an expert of the cleaned salary data.

Table 1: Clean UBC Salary Data

	Last_Name	First_Name	Remuneration	Expenses	Year
0	Aamodt	Tor	193153	5597.0	2023

Table 1: Clean UBC Salary Data

	Last_Name	First_Name	Remuneration	Expenses	Year
1	Abanto	Arleni	107723	393.0	2023
2	Abbassi	Arash	109136	82.0	2023
3	Abdalkhani	Arman	101829	NaN	2023
4	Abdi	Ali	238203	2981.0	2023

Gender Prediction

Babyname Corpus

In order to predict gender, I used datasets with babynames and assigned genders. In order to have a somewhat diverse set of baby names, I used babynames from canadian, american, and indian sources Sharma (n.d.).

For each UBC staff name, I found whether that name was more common among girls or boys in the babyname dataset. Then, I guessed the gender that was most common. Any staff names that had less than an 80% gender majority were given a null gender.

Table 2: Babyname Data

	index	Sex_at_birth	First_Name	Accuracy
95660	24125.0	Female	Irin	0.5

In Table 2, the `Accuracy` column shows the percentage of gender majority from the babyname dataset. For example, if 95% of babys named George in the dataset were male, the `Accuracy` column value would be 0.95. We can see the name Irin has about the same number of boy and girl names. So, since its not above the 80% threshold, anyone with that name would have a gender value of `None` assigned.

NLTK

For UBC staff names that were not found in the babyname datasets, I used a natural language processing model (package: `nltk`) to predict gender.

Below are some examples of names not found in the babyname dataset.

```
0    Fatawu
1    Tamiza
2    Purang
3     Ninan
4      Reto
Name: First_Name, dtype: object
```

In order to train and evaluate the model, the babyname dataset was split into a trianing and test set. The accuracy on the test set was `0.85`. However, the accuracy on the actual data is likely lower than `0.85` since the actual data contains more unique, unusual names that were not found in our babyname dataset. The two features the classifier found most useful were the last 3 letters and last four letters of a name.

Cleaning

Results

References

"Kaggle." n.d. https://www.kaggle.com/datasets/kaggle/us-baby-names/data.

Sharma, Anany. n.d. "US Baby Names." https://www.kaggle.com/datasets/ananysharma/indian-names-dataset.

Statistics Canada. n.d. "Table 17-10-0147-01 First Names at Birth by Sex at Birth, Selected Indicators (Number)." https://doi.org/https://doi.org/10.25318/1710014701-eng.

University of British Columbia. 2020. "UBC Statement of Financial Information." Open Government - Open Data. https://finance.ubc.ca/reporting-planning-analysis/reports-and-disclosures.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.