Team 1: Deliverable 3

Aditya Agrawal : adityaai@bu.edu
Melissa Zhen : mzhen001@bu.edu
Liangzhuo Zhang : zhanglz@bu.edu
Cindy Lu : cindy19@bu.edu

Project Goal

Our goal is to collect, process, and analyze patent data of Korean and Chinese authors from patents filed in the US within the year of 2017 to 2021, so that we can then analyze gender distribution among these patent authors by what country the authors are from. We also want to see if there is a difference in gender analysis when a patent author's original character name is used in comparison to when it's translated to English.

Patents & Names Extraction Method

We used the previous team's (specifically Spring 22 Team 1) data for our analysis. The previous team's data contained information for all patents filed in the USPTO (United States Patent and Trademark Office) from 2005 to 2021. We choose patents for five years from 2017 to 2021, extracting Chinese and Korean patents from the original datasets.

The previous team had labeled the inventor's country of origin for each patent. We filtered the patents that had "CN" or "KR" in their list of inventor countries.

All patent numbers were the US Patent numbers. We use PyPI's google_patent_scraper to get the scraped Google Patents webpage for each US patent. From there we filtered worldwide applications to get the Chinese and Korean patents numbers and extract the original Chinese and Korean names of inventors from them. If a patent had both Chinese & Korean patents, we used the inventor country as a reference to select the appropriate patent.

We collected 38452 for 2017, 38768 for 2018, 50960 for 2019, 55183 for 2020 and 48597 for 2021.

Analysis of Patents & Names Extracted

After consulting with our client, we've decided to perform a gender analysis for ~8,000 Chinese names and ~8,000 Korean names for each of the years from 2017-2021. This currently gives us a sample size of ~40,000 names to work with for Chinese and Korean each.

Methods Used for Gender Prediction

English

For English names, we utilized the methods from the previous year's teams which was using Harvard Dataverse's WGND (World Gender Name Dictionary) to determine if the name was male or female. If the name was not present in the dictionary, we use Python's library "gender-guesser 0.4.0." to guess the gender, which returns {male, female, androgynous, unknown, mostly male, mostly female}.
The previous team noted, and we concur, that this might not be the best method for names that were translated from their Chinese or Korean originals, as it would often classify their genders as "Unknown."

Chinese

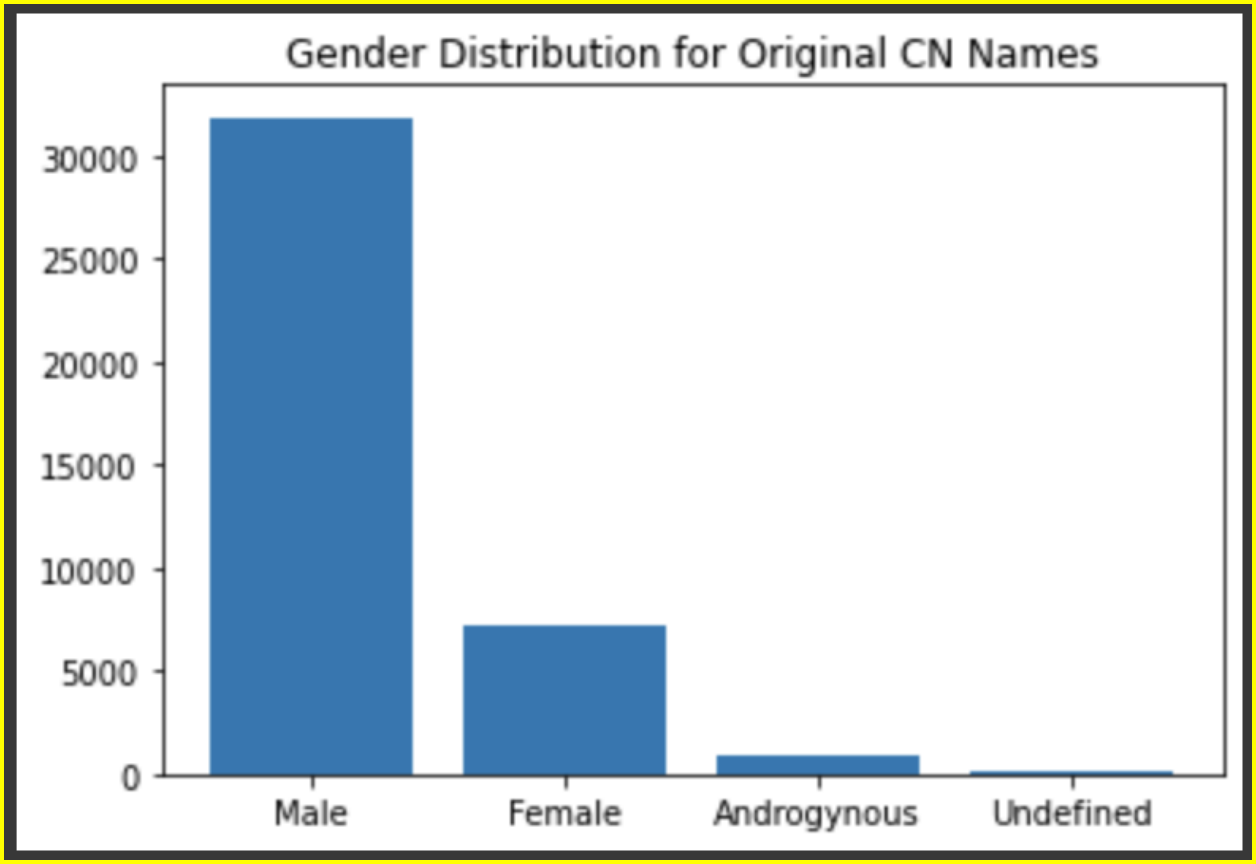
For Chinese names, we used an algorithm developed by Zhengxiang Wang (jaaack-wang/gender-predicator (Github)). It is a Naive Bayes algorithm that was trained on 3.65 million Chinese names to determine the gender of names. It also gives the probability of the predicted gender being correct.

Korean

For Korean names, we used an online, global name checking technology called Namsor. The Namsor API is able to extract specific information by processing and classifying names. We specifically used its "genderize" function which accepts a name as a parameter and returns the most likely gender and a calibrated probability as an accuracy score.

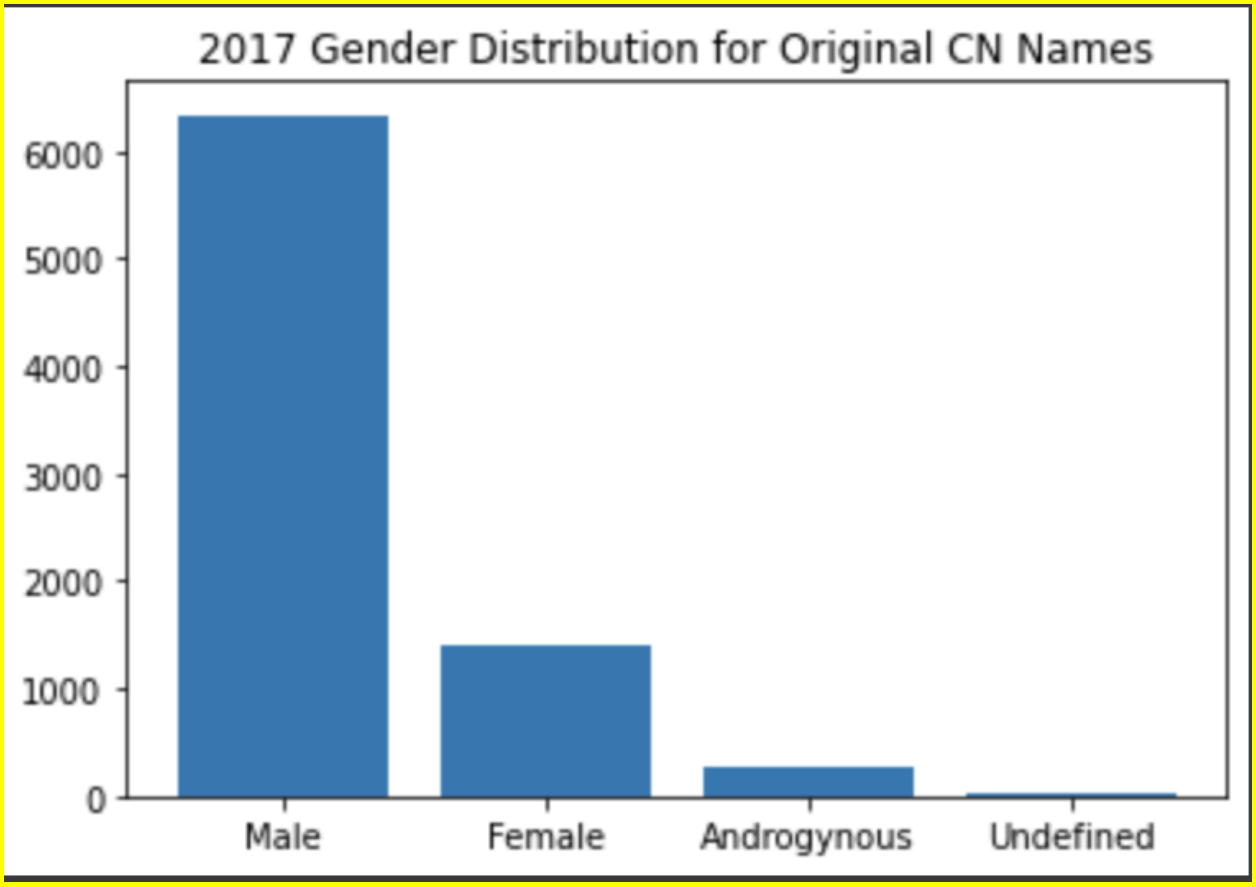
Gender Results Analysis (so far)

Original Chinese Names

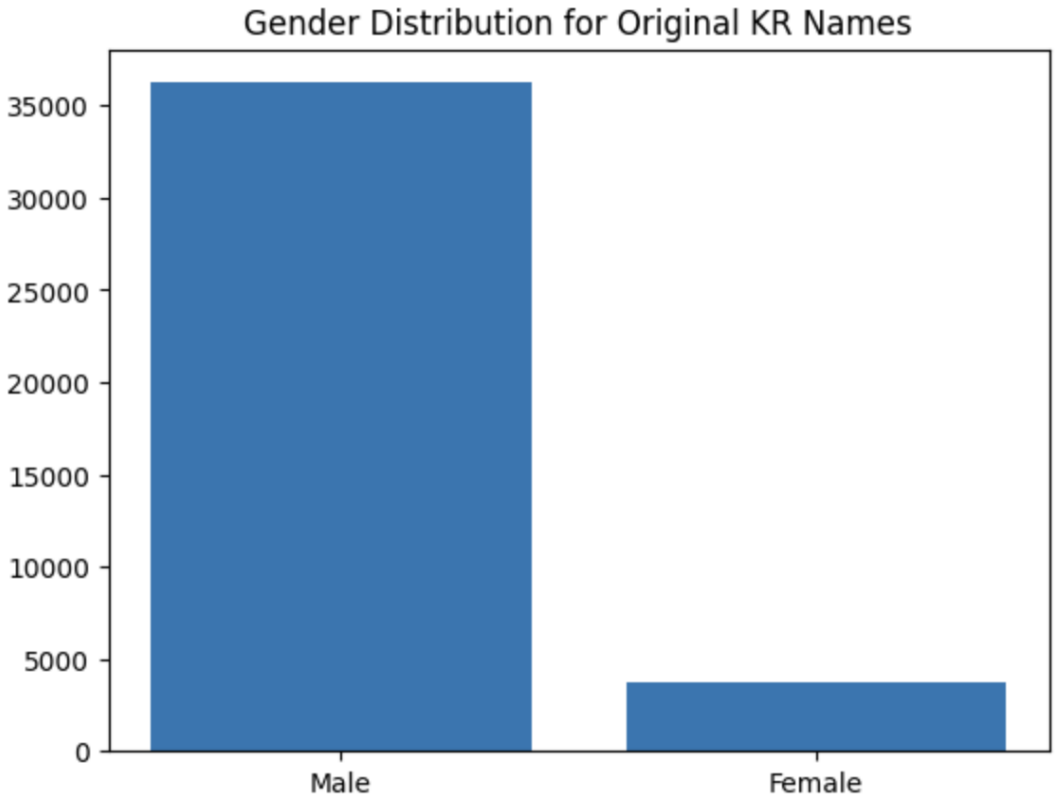


2017: 6335 Male, 1395 Female, 262 Angrogynous, 11 Undefined
2018: 6411 Male, 1406 Female, 179 Androgynous, 9 Undefined
2019: 6385 Male, 1394 Female, 214 Androgynous, 10 Undefined
2020: 6387 Male, 1475 Female, 128 Androgynous, 11 Undefined
2021: 6348 Male, 1538 Female, 104 Androgynous, 15 Undefined
Overall 2017-2021: 31866 Male, 7208 Female, 887 Androgynous, 56 Undefined

Translated CN Names in English



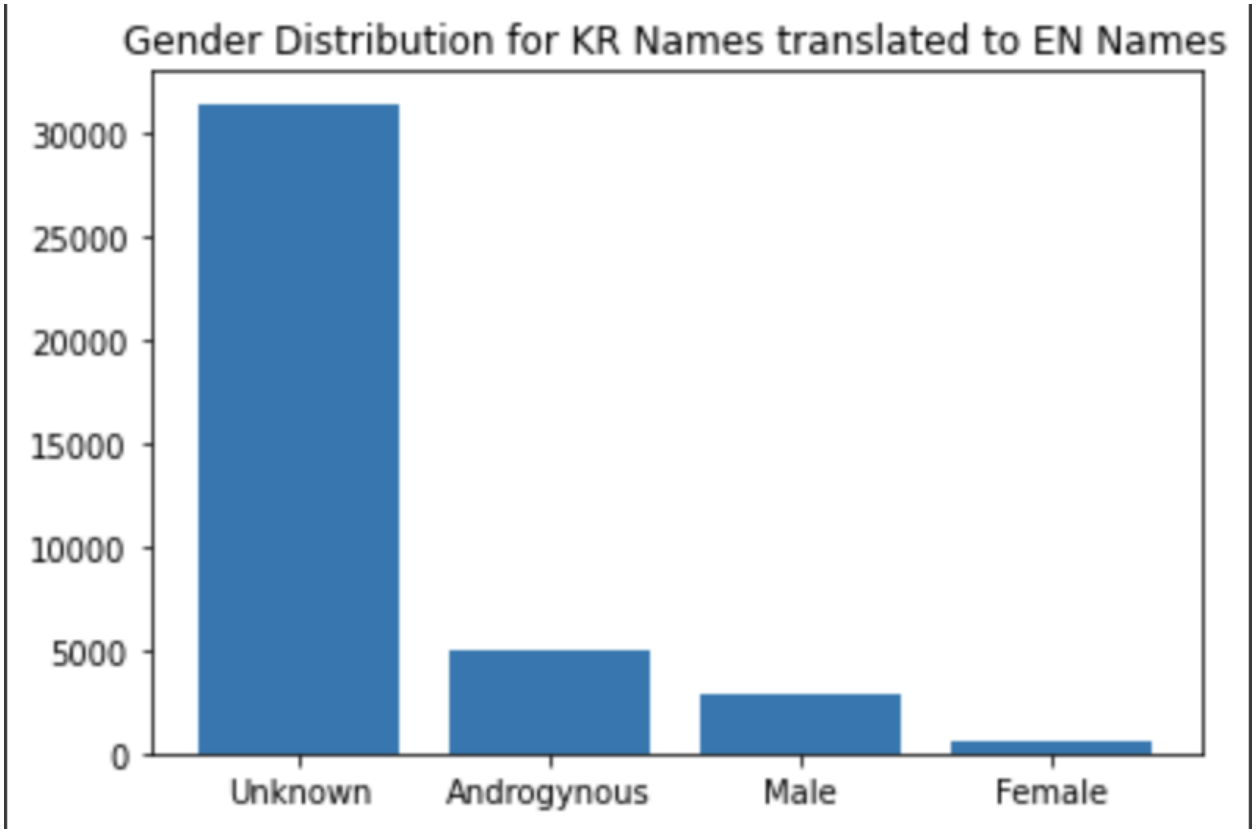
2017: 4800 Unknown, 2618 Androgenous, 493 Male, 92 Female
2018: 4969 Unknown, 2453 Androgenous, 482 Male, 101 Female
2019: 4937 Unknown, 2441 Androgenous, 496 Male, 129 Female
2020: 5000 Unknown, 2474 Androgenous, 417 Male, 110 Female
2021: 4977 Unknown, 2485 Androgenous, 428 Male, 115 Female
Overall 2017-2021: 24683 Unknown, 12471 Androgenous, 2316 Male, 547 Female



Original Korean Names
2017: 7223 Male, 777 Female
2018: 7283 Male, 717 Female
2019: 7275 Male, 726 Female
2020: 7184 Male, 817 Female
2021: 7267 Male, 738 Female

Overall 2017-2021: 36232 Male, 3775 Female

Translated KR Names to English



2017: 6178 Unknown, 1080 Androgynous, 601 Male, 141 Female
2018: 6317 Unknown, 1026 Androgynous, 543 Male, 115 Female
2019: 6307 Unknown, 991 Androgynous,583 Male, 120 Female
2020: 6264 Unknown, 989 Androgynous, 621 Male, 127 Female
2021: 6419 Unknown, 947 Androgynous, 506 Male, 133 Female
Overall 2017-2021: 31485 Unknown, 5033 Androgynous, 2853 Male, 636 Female

Conclusion

In response to the first question, we found that the main difference in analyzing characters instead of US translated names is that much of the resulting translations are not found within the dictionary. Therefore most names are classified as androgynous or unknown. Gender predictors designed for English names do not handle translated names well.

In response to the second question, we found that there is an overwhelming bias towards males in the gender distribution of both Chinese inventors and Korean inventors. For Chinese inventors, about 80% of analyzed inventors were categorized as male, and this trend did not change throughout the analyzed years. For Korean inventors, we see the same trend occurring also. About 90% of the inventors were categorized as male throughout the years 2017-2021. It

is worthwhile to note, however, that for both the English translated Korean and Chinese names, there are more names that are classified as androgynous than male and female combined.

Limitations

The data we collected from Google Patents does not contain the gender of the inventors. So we can not compare our predictions with the real gender of inventors. We can only compare the most likely accuracy of the generated genders.

Also, because we wanted to find patents in the USPTO dataset, as well as the equivalent Chinese and Korean patent offices, we utilized the patent numbers to cross-match patents across the multiple datasets. However, we soon realized that not all Chinese and Korean patents in the USPTO dataset have a corresponding Chinese or Korean version on Google Patents. Therefore, not all the data we collected is not representative of the entire Chinese and Korean patent dataset.

Additionally, because scraping and processing each year's dataset was taking an overwhelming large amount of time, we were given approval by our client to sample a subset of each year's data from the most recent 5 years–2017 to 2021. There could be unseen trends from previous years that we did not sample from.

Contributions

Aditya: Extracted patent names from 2019 and 2020. Worked on the Korean Gender Predictor and Chinese Gender predictors.
Cindy: Extracted names from 2018 patents, graphs + visualizations for Korean names, writing most of deliverable 3 report
Leon: Extracted patent names from 2021. Explored possible methods for gender prediction on names. Worked on three deliverables. Kept contact with the client and confirmed project details and problems we met with clients.
Melissa: Extracted original CN & KR names from the year 2017, worked on generating the visualizations for the gender distribution of Korean and Chinese names, summarized the gender distributions for each year under each graph, kept contact with our client to update and ask questions whenever necessary.