Andrew Choi(dooby@bu.edu), Sangjoon Lee (sj0726@bu.edu), Ruojia Li(ruojiali@bu.edu), Edona Mujaj(edona@bu.edu)

Team 2: Deliverable 4

- **Project goal**
  - The goal of this project is to take patents that exist in the US and in China or Korea and to see if there was a difference in the accuracy of gender prediction based on the names of the inventors. The data we started off with had the US patents, the romanized names of the patent inventors, and the genders of the inventors based on their names. What we did in this project was that we got the corresponding Korean or Chinese patent and performed gender analysis based on the names of inventors in their respective languages, meaning their names in Korean characters or in Chinese characters. We then compared the accuracy of determining the gender of the names written in Korean or Chinese characters with the gender makeup that was given in our starter data. This was performed to see if there was a difference in how well the gender of inventors was predicted based on whether the name was romanized or not.
- **How we got our data**
  - We used the previous team's collected data, and the World Gender Name Dictionary(WGND). We also used web scraping and Google Patents to extract the exact Korean and Chinese names. We then compiled a csv containing all of the patent holder names and used this as when running our analyses.
- **Code**
  - sort_data.ipynb
    - Code we used for taking last semester's team's data and sorting and parsing them into Chinese/Korean inventor information. Also took data from World Gender Name Dictionary (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/MSEGSJ) and parsed them into a reference data frame that we used for determining the likelihood of a Korean/Chinese name being either female or male.
  - googlePatents.ipynb
    - Used pandas, numpy, json, BeautifulSoup, requests packages to extract the names of the inventors from Google Patents website through web scraping by taking patent numbers as input.
  - KR_gendermatch.ipynb
    - Testing ground for finalize_kr_data.ipynb and finalize_cn_data.ipynb. These were initial codes made to take Korean/Chinese character names and determine the likely gender. The details are explained below.
  - finalize_kr_data.ipynb
    - Used pandas, numpy, and matplotlib.
    - Loaded in the text file for all the Korean patents and made it into a dataframe. Each line of the text file contained a single patent and all the inventors for that patent, so when it was turned into a dataframe, it was

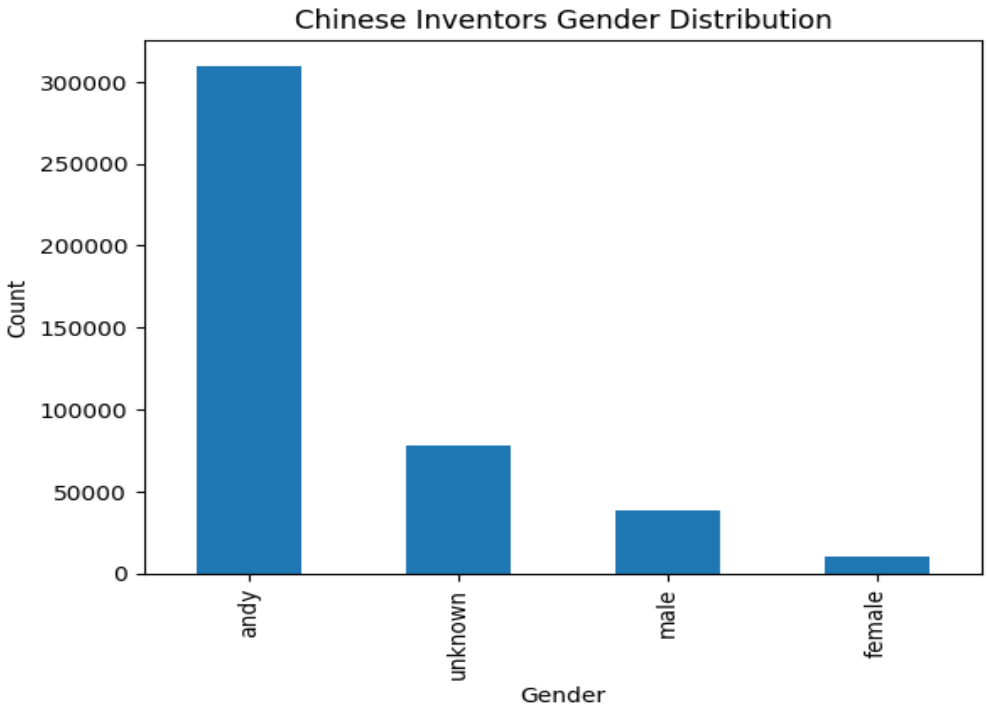
split so that each row contained one name and the corresponding US patent number. The last name was then removed from all of the inventor names, we sampled the first 50,000 names (which we saved to a csv), and gender matched the names with the data received from the World Gender name dictionary. At the end, we produced a bar plot for the different genders.

- finalize_cn_data.ipynb
  - The code used is the same as the code from finalize_kr_data.ipynb, but we did not sample the Chinese inventor names. Due to some project limitations, we were only able to gather 8,000 Chinese inventor names, so we used all 8,000. We also saved all the Chinese inventor names and corresponding patents to a csv.

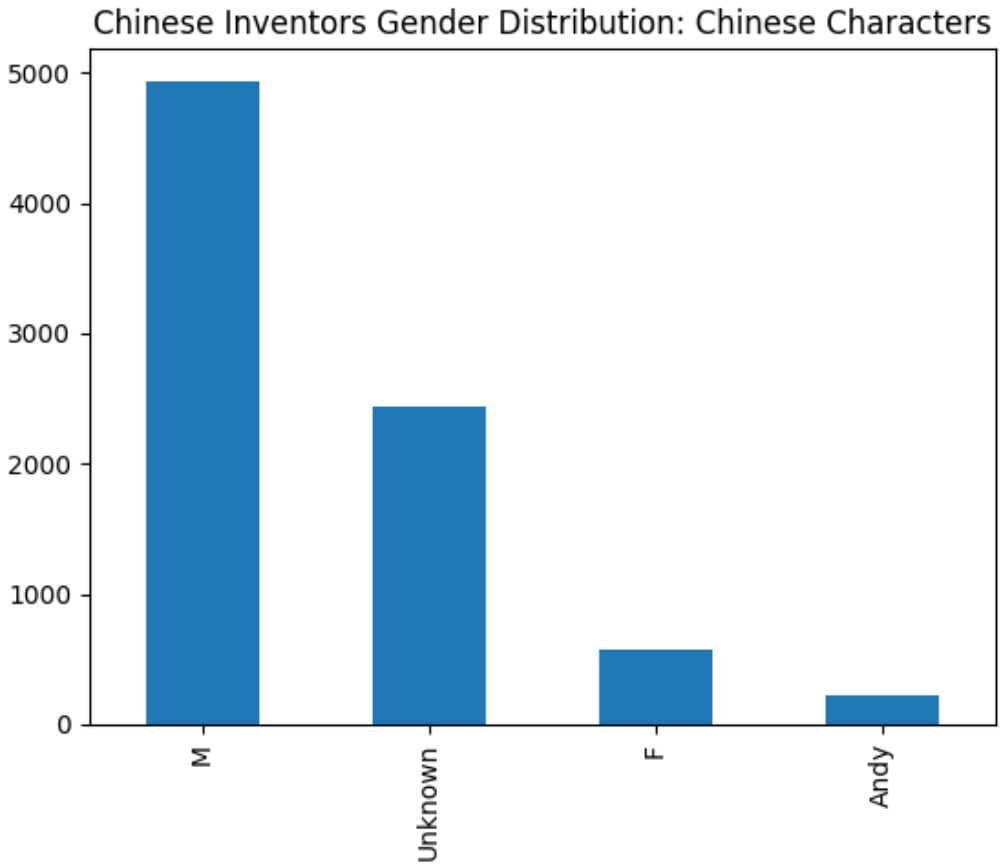
- **CSVs**
  - inventor_data with patent number and claims number.csv
    - Original patent information from the previous team. Was used to create inventors_filtered_by_cn_kr.csv
  - inventors_filtered_by_cn_kr.csv
    - The inventor information that was filtered so that it only contained patent information for patents that have Korean or Chinese origin.
  - wgnd_2_0_code-langcode.csv
    - CSV from the World Gender Name Dictionary 2.0 that contains the language codes and country codes.
  - wgnd_2_0_name-gender-code.csv
    - CSV from the World Gender Name Dictionary 2.0 that contains names, genders, country codes, and weights for the distribution of the name based on gender. This was used to create zhko_with_weight.csv
  - wgnd_2_0_name-gender-langcode.csv
    - CSV from the World Gender Name Dictionary 2.0 that contains names, genders, and language codes. This was used to create zhko.csv
  - zhko.csv
    - Data from wgnd_2_0_name-gender-langcode.csv that is filtered so that it only contains rows with names from China and Korea.
  - zhko_with_weight.csv
    - Data from wgnd_2_0_name-gender-code.csv that is filtered so that it only contains rows with names from China and Korea.
  - KR_50k.csv
    - The data produced from filtered_kr_data.ipynb. This is the data where we removed the last names of the Korean names obtained from web scraping Google patents. We sampled it to the first 50k inventors for time's sake.
  - CN_8k.csv
    - The data produced from filtered_cn_data.ipynb. This is the data where we removed the last names of the Chinese names obtained from web scraping Google patents. The number of Chinese patents were drastically

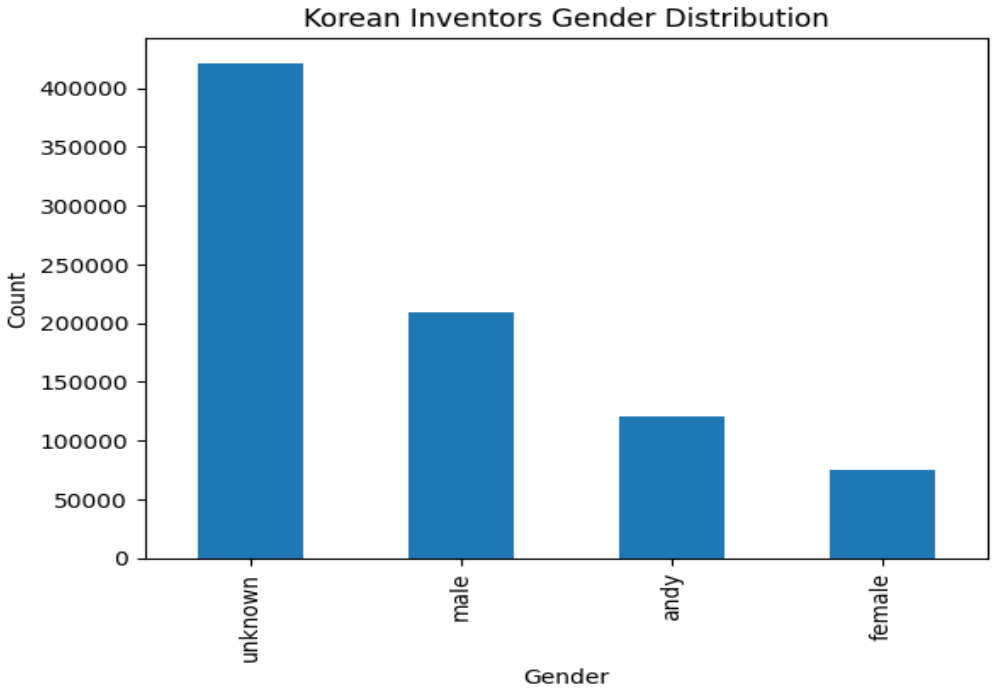
Andrew Choi(dooby@bu.edu), Sangjoon Lee (sj0726@bu.edu), Ruojia Li(ruojiali@bu.edu), Edona Mujaj(edona@bu.edu)

smaller than that of the Korean patents, so we did not decide to sample the data.

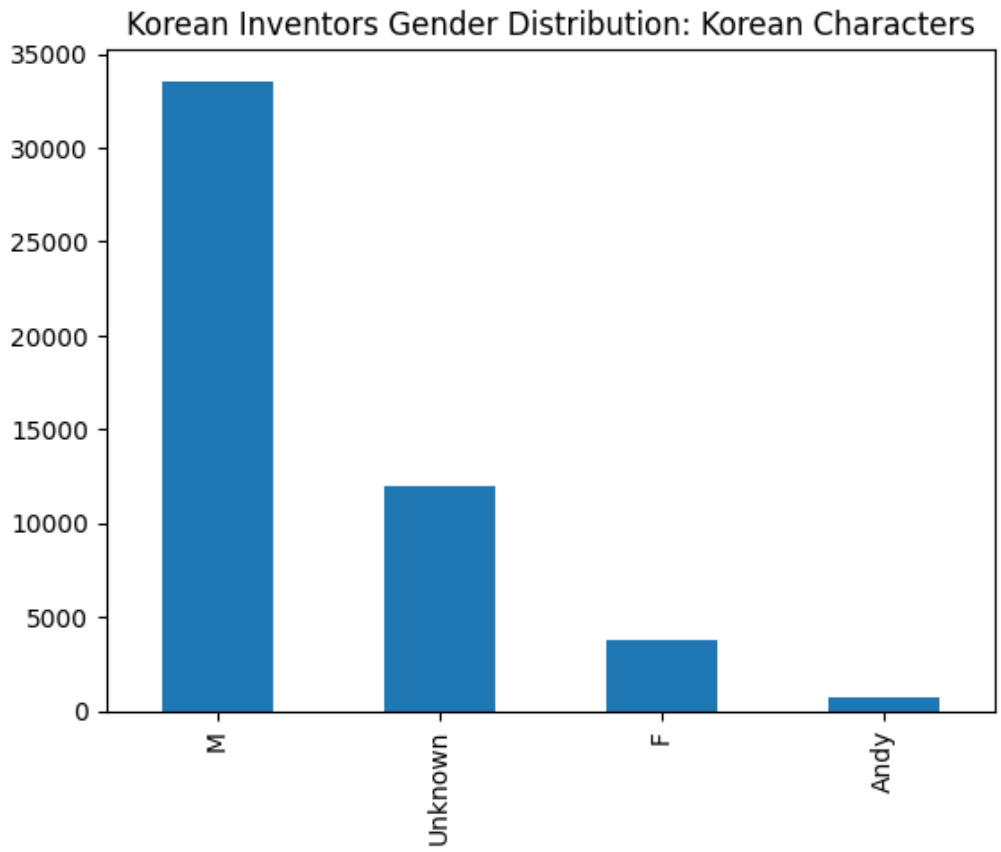
- **Data analysis**
  - 

Andrew Choi(dooby@bu.edu), Sangjoon Lee (sj0726@bu.edu), Ruojia Li(ruojiali@bu.edu), Edona Mujaj(edona@bu.edu)





Andrew Choi(dooby@bu.edu), Sangjoon Lee (sj0726@bu.edu), Ruojia Li(ruojiali@bu.edu), Edona Mujaj(edona@bu.edu)



-
- The graphs above show the gender distribution of the data from the previous team. This is the data of the genders of the inventors who had their genders predicted based on their romanized names.
- We attempted to compare the names we extracted and obtained with the data we have on a Korean/Chinese name's gender distribution. However, because of language encoding issues, the data we obtained wasn't able to match the names in Chinese/Korean characters, which blocked us from making further analysis. With that being said, we will retry to download the data (~400k pages of web scraping) again with proper language encoding this time to ensure smooth workflow and proper analysis. The work is currently in progress and we hope to show it by the next deliverable.
- When we analyzed the names in their native characters, we saw a higher turnover in our analysis (less andy/unknown names) for both Korean and Chinese datasets. Additionally, we saw a clear deviation between Male and Female inventors in both Korea and China.
- Despite the disparity in number of data entries between Chinese and Korean patents, the general pattern is the same, as shown in the graphs, which tells us that the sampling of the data was sufficient to derive a plausible pattern in analysis.

- **Method for gender prediction**

Andrew Choi(dooby@bu.edu), Sangjoon Lee (sj0726@bu.edu), Ruojia Li(ruojiali@bu.edu), Edona Mujaj(edona@bu.edu)

    - To determine the gender of the inventor names, we matched the names in the World Gender Name Dictionary(WGND) provided by Harvard. The WGND has names with their associated gender. The WGND was created to be applied to intellectual property unit-record data naming. The data contains names in different languages and from different countries. To get the data that is used in our prediction, we had to filter the WGND. As the WGND has country and language codes associated with the names, we were able to get the names that were only written with Chinese and Korean characters. Along with the inventor name data that we collected through our web scraping, we matched the names from the collected data with the data from the WGND and assigned the genders. Then we created a dataset and compared it to the results of the previous team's work to see the difference.
- **Conclusion**
    - One thing that we've recognized throughout our work is that when analyzing characters in Korean and Chinese characters is that there are more androgynous names. The process of directly translating these names doesn't work very well initially, so we switched to just analyzing them in their native characters.
    - The difference between gender analysis when analyzing characters in Korean and Chinese is that there are a more androgynous names. Gendering names through just translating them does not work very well.
    - It was interesting seeing the gender distribution and how few female inventors there are. If we had more time and resources, ideally we would run more names to see what the final distribution is. We would also try different APIs or dictionaries to compare the accuracies.
- **Project limitations**
    - There are unisex names for both languages, which leads to difficulty in confidently determining the gender of the names. In terms of data collection, not all US patents were in the Korean/Chinese databases. Another limitation was in hardware. The way that our web scraping worked requires a lot of time to collect the data from the patents. With a better computer, the work could be done faster, but in our case this was not possible. Also similarly to the romanized inventor name data, some the names in their respective languages had no conclusive gender, which led to their gender being either androgynous or being left empty. Another limitation is seen in the Chinese names of inventors. Some of the characters of the Chinese names were not recognizable, which led to them not being able to have their gender predicted.
    - The web scraping was also difficult due to the sheer number of patents. Google patents timed out at around 30,000 patents for Korean, and 2,600 patents for the Chinese one.
    - There were a lot of patents in the US that did not have a corresponding patent in China
    - In our Korean name dataset contained names that werent Korean, which skewed with our analysis

Andrew Choi(dooby@bu.edu), Sangjoon Lee (sj0726@bu.edu), Ruojia Li(ruojiali@bu.edu), Edona Mujaj(edona@bu.edu)

- A lot of Chinese names were also not recognized, and the names we got were not in full Chinese. Some had English letters mixed in, and we had to further parse them to clean up the usable data.