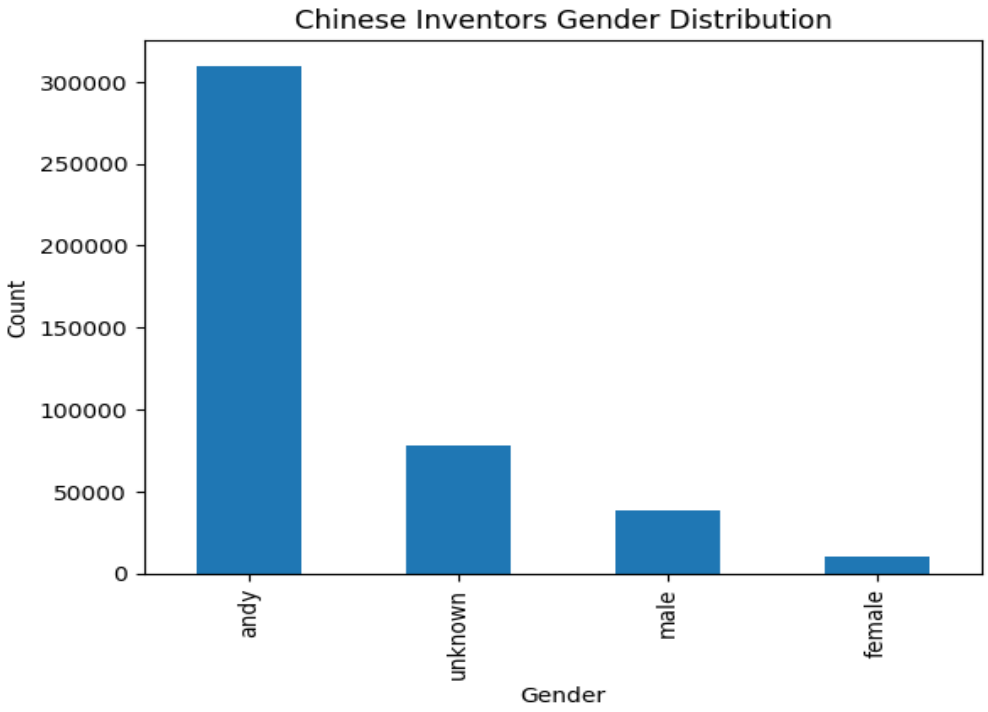
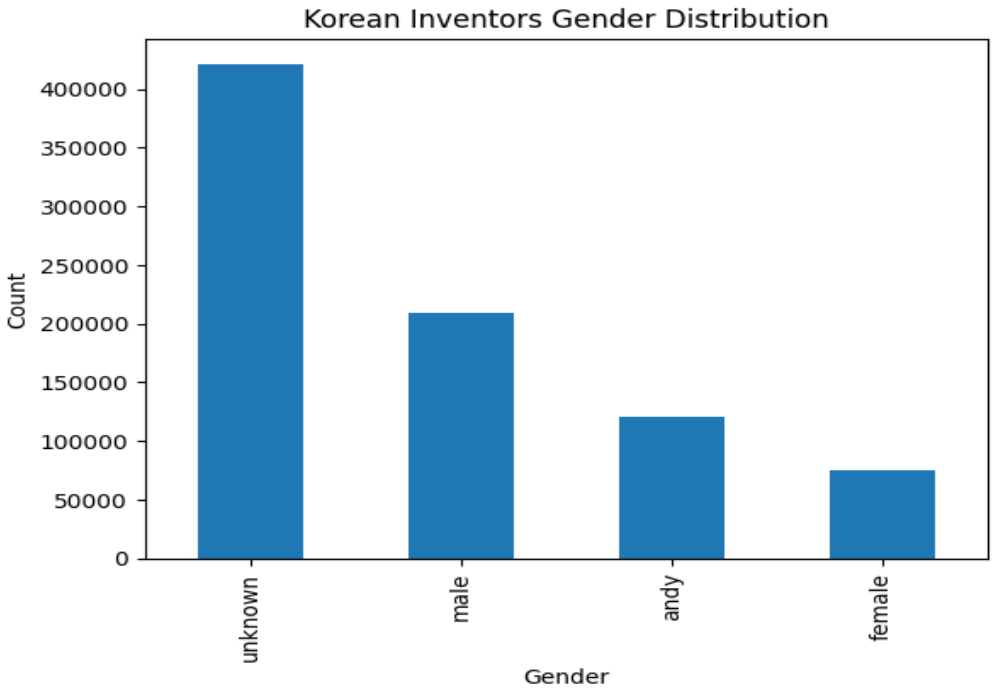
Andrew Choi(dooby@bu.edu), Sangjoon Lee (sj0726@bu.edu), Ruojia Li(ruojiali@bu.edu), Edona Mujaj(edona@bu.edu)

Team 2: Deliverable 3

- **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.
- **Data analysis**
  - 

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.

- **Method for gender prediction**
  - 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**

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

- 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, so we are still working on finding a proper solution to fully answer our two key questions: What is the difference in gender analysis when analyzing characters instead of US translated names? What is the gender makeup of inventors on Chinese patents and Korean patents?

.

- 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.

- **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.